AN ANALYSIS AND COMPARISON OF STUDENTS IN 3RD LEVEL EDUCATION
Technical Report

MAY 9, 2017

NATIONAL COLLEGE OF IRELAND

Student Name: Alan Sullivan,
Number: x13114255,
Email address:alansullivan21@gmail.com
Course: BSc (Hons) in Computing
Specialization: Data Analytics

Declaration Cover Sheet for Project Submission

SECTION 1 *Student Details*

Name: Alan Sullivan
Student ID: x13114255
Supervisor: Frances Sheridan

SECTION 2 Confirmation of Authorship

The acceptance of your work is subject to your signature on the following declaration:

I confirm that I have read the College statement on plagiarism (summarised overleaf and printed in full in the Student Handbook) and that the work I have submitted for assessment is entirely my own work.

Signature:_Alan Sullivan_____Date:_5/10/2017_____

Table of Contents

Executive Summary

This report documents the approach the researcher's project will take in order to answer the question posed. The project aims to gather datasets on students and analyse them, in order to see what subset of students is entering 3rd level education and achieving greater success, in terms of dropout and graduation rates, in higher numbers. The two data subsets are secondary school leavers, 18-22 years old, and mature students, 23+ years old.

The document intends to help the reader understand what happened in the project and how the approach taken helped to complete different milestones.

December 2016 marked the mid-point of the project, a stage at which many of the techniques that were needed in order to dive further into the datasets, would only be taught in semester two starting February 2017.

May 2017 marked the completion of the project in its entirety; the following document summarising all of the steps that were taken throughout the duration of this project and the manner of its completion. Using a dataset provided by the National College of Ireland, this project was able to analyse and determine which subset of student was more successful in 3rd level education from historical data (2007-2015), and furthermore, to predict the average age of a dropout student in 10 years' time.

1 Introduction

1.1 Background

The reasoning behind this project stemmed from my own personal experience in 3rd level education and the stresses that came with it. Thinking back to when I was a 19 year old going through the process of applying for colleges which I eventually made the right choice and stopped; I knew I wasn't ready to approach an academic 3rd level challenge at that time due to circumstances that tore my concentration away from school. When I did return as a mature student I carried with me a better outlook on life and how to approach a college environment as a whole, coming from the working world I had a better grasp on how to manage everything that college throws at you.

From early on in college journey at NCI I found myself asking questions of why do young students put themselves through so much pressure and stress when they are not ready for it? Is there a fundamental flaw with our education system that pushes young minds who are not fully developed with a mental fortitude strong enough for college life into something they don't want to do? Why is it the social norm for kids to open their leaving cert results and be overjoyed with the prospect of another minimum 3-4 years of studying? Do they even know the danger of college? Drugs, alcohol, sex, partying, peer pressure on top of all this projects, studying, CAs and for most poor nutrition, this leads many to the inevitable breaking point of dropping out because these young minds haven't got what there mature counter parts have, Life experience. Most, like me have sown their wild oats and experienced the world to a degree where they are purely focused on the school side of college with the ability to block out the other social end of it not worrying what people will think of them if they say 'no I'm not going out tonight'.

I noticed very early on that mature students will have the for lack of a better phrase the 'cop on' to ask for help when they feel the pressure starting to rise, something many of the younger students won't do because they see this as a failure not to themselves but to their peers who they perceive as doing fine without any help and again they end up getting lost in a sea of work they can't keep up with and drop out usually with words of 'oh I just didn't like the course'. How many of these students are then crippled by this, feeling like they have failed and never return to education. Even now in a discussion with friends asking them about their college experiences the general consensus was that they all started college young in a course they didn't enjoy because they felt they had to go to college, but now looking back they knew they weren't ready and as the people they are now could have done so much better in college and would have chosen what they wanted to actually do.

From this an idea sprang, what if I could collect a dataset of students not just in Ireland but from other countries. Comparing the ages of those who have completed 3rd level education stipulating a general 4 year cycle, I could see if there is a trend in what age group would be the best or even what age itself would be the most optimal to enter 3rd level education. I could on ages alone see if there will be a fall or increase in the mean or average age of students entering 3rd level in 10 or 20 years' time, hoping that with this type of prediction my questions have a good basis. Hopefully answering these questions will yield some good answers and will open up the door for a new approach to college recruitment, as research conducted by CDC in 2015 states that among adults aged 18-22 years, the percentages of full-time college students had suicidal thoughts 8.0% or made suicide plans 2.4%.So with suicide and depression amongst kids these days in our world so driven by a fast paced social media environment and everything changing so quickly it might be better to let these young minds know that its ok to breathe and take your time because college is going nowhere and life is for living.

As my final year continued issues arose within some modules which further strengthened my conviction to this project and wanting to find an answer to my question. Seeing a lot of my peers stressed and thinking of quitting made me realize that if this was a majority of younger students having the same issues they might not have the resolve to carry on through it. Seeing how a Mature set of students navigated through a tough semester gave me some insight into the mind frame of an older set of students versus a younger set. Even the students who organized meetings to get us help with our problems were the mature ones showing that younger minds tend to retreat from issues instead of speaking about them.

As stressful as this all is it made me thankful to have a real-life scenario I could pull from.

1.2 Aims

The following is a list of the projects aims and objectives:

Aim 1: The first aim is to find a dataset on students enrolled in 3rd level education this will not be limited in anyway by geographical reference; this is an open study on 3rd level education and its enrolled students as a whole. In order to find the sets that will be required a search, mainly on Government related databases, which host free datasets which will be extremely useful in this study. However, in order to get a more accurate reading the researcher has been in contact with NCI in order to acquire a more comprehensive dataset which will be very advantageous to have in terms of completing this project.

Aim 2: The second aim will be to clean the dataset to pull out information required e.g. age, sex, grades (if possible).

Aim 3: The third aim is to dive deep into the data and compare information across the years and see if there is an optimal age to start 3rd level education based on how many students are enrolled and or graduating.

Aim 4: The fourth aim is to use machine learning algorithms and based on results, see if there is a trend in the way ages have varied over the years and see if there will be a certain majority age in 10 years or 20 years' time.

Aim 5: Create an anonymous survey for students in NCI in order to identify potential correlations between certain age groups and specific stressors, and if certain age groups report a higher average of said specific stressors. Using this as a way of backing up what the study itself may show, while also using figures found on suicide rates among young people.

Aim 6: To complete documentation with views on this study, including ideas on how to help improve college life for students as a whole, whether that be setting up peer support or hosting talks on the unspoken dangers that lie within college life.

1.3 Technologies

R Studios

The project will make use R Studios to construct this project, which is an open-source integrated development environment for R Language which itself is a programming language for statistical computing, graphics and will be the language this project will be written in. R studios will be used in conjunction with excel where the datasets will be store locally and be used to retrieve information as needed such as ages of students.

R Language

R language will be used to build this project and display information it has not been decided which libraries will be utilized within R studios yet.

SPSS

SPSS is one of the most popular statistical packages which can perform highly complex data manipulation and analysis with simple instructions.

The project will make use of SPSS to back up what I'm displaying using it as a more end user friendly way of viewing the information such as that a person with no technical ability will be able

to manipulate what they see if they so wish; although an end user is not a requirement for the project to be a success it will help with testing the information using built in commands using as a testing tool help to solidify the information the project is hoping to produce.

Excel

Excel is a spreadsheet tool with built in statistical and graphing commands that allow a user to manipulate information and data loaded into it.

The project will use excel to hold the information in a more structured format until the cleaning process begins all irrelevant information is removed.

SQL Server

The project will make use of SQL Server which is Microsoft's relational database management system supporting SQL queries to will hold the data on this server using it as a way to easily gain access to it no matter where the admin might be. This way attainability of the data is assured allowing for use on multiple machines.

Tableau

Tableau allows for instantaneous insight by transforming data into visually appealing, interactive visualizations called dashboards.

The project will make use of Tableau to Display all knowledge and interpretations of final figures that are acquired through the datasets so an end user may easily understand the results which are being displayed.

There are more technologies that may be used as the project goes forward provided that they meet the needs of the project, alternatively technologies may be dropped if they are not needed they will be left in this report but an explanation will be provided as too why they have been dropped.

1.4 Research

Before the project could continue, research was required in order to establish that no similar study had been completed on the same topic. After searching through Google Scholar and dataset sites such as Kaggle and UCI depository, it became clear that no project of this nature had previously been conducted.

While interesting studies can be found around students and school life, these pertain more to what happens within classrooms and subsequent teaching methods.

Looking towards reasons why college students might drop out, led to 3 interesting articles. The first of these articles was found on the website journal.ie from 2010; it is entitled "College Dropout Rates Revealed" and provides an account of a survey conducted on some 35,000 full time under-graduates. It revealed a variety of interesting findings across all courses, but found that dropouts in Computer Science appear particularly high, "While the highest rating courses have a progression rate ranging from 98 per cent to 95 per cent, computer science courses have a progression rate of just 73 per cent" ("College Dropout Rates Revealed",2010). The survey also revealed that females are more inclined to progress further in education than males ("College Dropout Rates Revealed",2010). Furthermore, it also found that students who secured an educational grant were more likely to progress further due to the financial security. The financial security factor was also evident in the finding that children of higher skilled professionals were more likely to progress than those of skilled manual workers ("College Dropout Rates Revealed",2010).

This article provided an insightful overview of the characteristics of students who actually progress in third level education. This information was particularly useful in highlighting student factors, such as gender and finances, from which inferences could be drawn to make a tentative hypothesis as to the reason why the trends observed in the data may have been the case. This theme was again explored in another article from 2016 which also covers the topic of dropouts.

Entitled "The Real Reasons College Students Drop Out" ("The Real Reasons College Students Drop Out",2016), the article which can be found on the website fortune.com, also discusses reasons college students might drop out. This article makes reference to an analysis completed on tuition fees for college students attending Washington College. It found that there were two strong factors in dropout rates: financial security and social isolation ("The Real Reasons College Students Drop Out",2016). It states that "students struggling to cover tuition costs – sometimes with excessive borrowing – were more likely to drop out, as were students who felt detached and alone – those, for instance, who were not active in a club, sport, or academic group" ("The Real Reasons College Students Drop Out",2016).

This article provides excellent coverage of similar themes discussed in the 2010 article re the topic of dropouts, whilst introducing the new factor of social isolation. Consideration of this factor led the researcher further down the line of enquiry into related mental health issues, including depression and suicide. This brings us on to the review of the last article found, which discusses suicide in young people.

In 2015, the CDC released a spreadsheet article summarising the results of a survey which investigated suicide rates across genders and age ranges ("Suicide-datasheet-a", 2015). Some

overwhelming figures were revealed from the findings of this survey. It was found that while males are more likely to take their own lives, females were more likely to possess suicidal thoughts ("Suicide-datasheet-a", 2015); revealing that whilst both genders are at risk of developing mental health issues along the suicide continuum, males may potentially present with poorer coping mechanisms than their female counterparts. The survey figures of suicide rates across different age groups, served to highlight that students within the age range of the "school leaver" college student group are more at risk of suicide:

"Suicide is the third leading cause of death among persons aged 10-14, the second among persons aged 15-34 years, the fourth among persons aged 35-44 years, the fifth among persons aged 45-54 years, the eighth among person 55-64 years, and the seventeenth among persons 65 years and older" ("Suicide-datasheet-a", 2015)

The article also provides specific figures for full-time college students between the ages of 18 to 22; it reveals that 8.0% of this population reported suicidal thoughts, and 2.4% of this population reported formulating suicidal plans, when surveyed ("Suicide-datasheet-a", 2015).

This article effectively highlights the dangers that surround younger college students when it comes to suicidal tendencies. Bringing in the male and female groupings shows that there was an excellent breakdown of students and gender which allows for insight into who is most at risk.

The aforementioned research collectively shows that there is a cross-over between dropouts, and in a sense the experience of suicidal tendencies. This will allow for the project to make strong inferences in the conclusion as to why one subset, school leavers versus mature students, experiences more success in third level education in terms of dropouts and graduation. While acknowledging that this is not concrete evidence as to why students drop out, it allows us to make a strong inference that a connection exists between the research discussed here and the project that was undertaken. This will be discussed further in "Conclusions" when the project has been completed, and the results can be properly compared against the findings of the research discussed here.

1.5 Structure

Throughout this projects lifecycle it will adhere to the path process of the KDD and following this path will allow for this project to be completed in full. Below is an image and a brief explanation of what each stage of the KDD is:

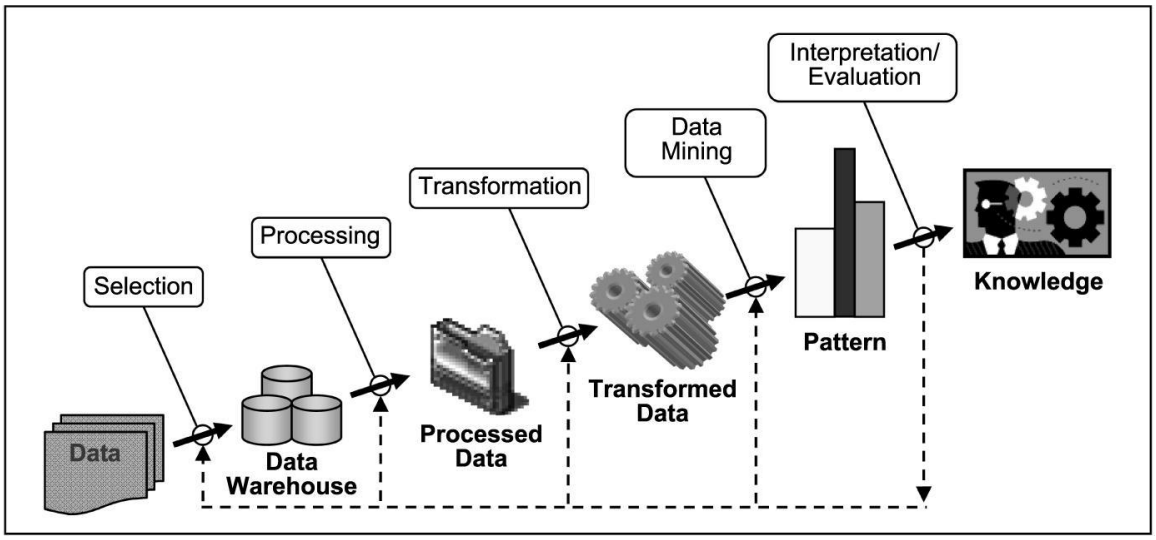


Figure 1 The KDD Methodology

- Selection- This process in the KDD requires a admin to go and search out a dataset that would be relevant to any project work that he/she would wish to complete. Upon finding the appropriate datasets they are then subsequently stored in a data storage waiting for use.

- Processing-This process in the KDD requires the admin to access the data and process it; by cleaning it up this allow the dataset to be reduced to the most optimal dataset that would be needed in order to complete the project.

- Transformation-This process of the KDD involves taking the clean dataset and, generating better data, for the data mining stage methods here include dimension reduction, such as feature selection, and extraction, and record sampling, and attribute transformation such as discretization of numerical attributes and functional transformation.

- Data Mining- This process of the KDD involves applying algorithms to a dataset in order to find valuable information such as any correlation between data or trends that might exist, this will allow for machine learning to take place and predictive analysis to be done.

- Interpretation- This process of the KDD involves taking the mined data and interpreting the results allowing the admin to produce visualizations of the knowledge gained.

1.6 Definitions, Acronyms and Abbreviations

In this section I have placed a list of Definitions, Acronyms and Abbreviations to help the reader better understand what certain parts of the document are referring to.

KDD: Knowledge discovery in databases is the process of discovering useful knowledge from a collection of data. This widely used data mining technique is a process that includes data

preparation and selection, data cleansing, incorporating prior knowledge on data sets and interpreting accurate solutions from the observed results.

Database application: A database application is a Graphical User Interface in which a user who accesses it can create a database to store data sets

Programming application: A Programming application is a User interface that allows a user to use a set of functions or protocols to do such things as create software or manipulate datasets.

Visualization application: A Visualization Application allows a user to visually display and present numerical data, particularly a graphical one. This might include anything from a simple X-Y graph of one dependent variable against one independent variable to a virtual reality which allows you to fly around the data.

Storage: the state of being kept in a place when not being used: the state of being stored somewhere

GitHub: a web-based Git repository hosting service. It offers all of the distributed version control and source code management (SCM) functionality of Git as well as adding its own features.

Google Drive: Google Drive is a personal cloud storage service from Google that lets users store and synchronize digital content across computers, laptops and mobile devices, including Android-powered tablet and smartphone devices

Dropbox: Dropbox is a free cloud storage service for sharing and storing files including photos, documents and videos. ... Files can be shared with others by providing them with a link to your Dropbox folder.

Admin: Short for administrator is a person who controls the use of something

UCI: The UCI Machine Learning Repository is a collection of databases, domain theories, and data generators that are used by the machine learning community for the empirical analysis of machine learning algorithms.

GOV.ie: This is a Government website used to find datasets.

GOV.org: This is a Government website used to find datasets.

HEA.ie: A high performing higher education system is an essential requirement in the development of creative, entrepreneurial people and the creation of new knowledge to support social, cultural and economic development.

X: This is a place marker used as a placeholder for a numeric value

GUI: In computer science, a graphical user interface, is a type of user interface that allows users to interact with electronic devices through graphical icons and visual indicators such as secondary notation, instead of text-based user interfaces, typed command labels or text navigation.

ETL: Extract transform load is the process of extraction, transformation and loading during database use, but particularly during data storage use. It includes the following sub-processes: Retrieving data from external data storage or transmission sources.

Google scholar: Google Scholar provides a simple way to broadly search for scholarly literature. From one place, you can search across many disciplines and sources: articles, theses, books, abstracts and court opinions, from academic publishers, professional societies, online repositories, universities and other web sites.

Kaggle: Kaggle was is a platform for predictive modelling and analytics competitions on which companies and researchers post their data and statisticians and data miners from all over the world compete to produce the best models. It is an open source data collection site hosting thousands of free datasets.

Microsoft Excel: Microsoft Excel is a spreadsheet program included in the Microsoft Office suite of applications. Spreadsheets present tables of values arranged in rows and columns that can be manipulated mathematically using both basic and complex arithmetic operations and functions.

SPSS: Statistical analysis, Data mining, Text analytics, Data collection, Collaboration & Deployment. SPSS Statistics is a software package used for logical batched and non-batched statistical analysis

RStudio: RStudio is a free and open-source integrated development environment (IDE) for R, a programming language for statistical computing and graphics.

RLanguage: R is a programming language and software environment for statistical computing and graphics supported by the R Foundation for Statistical Computing.

MySQL: MySQL is an open source relational database management system. Information in a MySQL database is stored in the form of related tables.

Tableau: Tableau software helps people use data to solve problems. It makes analysing data fast and easy, beautiful and useful.

Algorithm: In mathematics and computer science, an algorithm is a self-contained step-by-step set of operations to be performed. Algorithms perform calculation, data processing, and/or automated reasoning tasks.

Machine Learning: Machine learning is a method of data analysis that automates analytical model building. Using algorithms that iteratively learn from data, machine learning allows computers to find hidden insights without being explicitly programmed where to look.

NCI: National College of Ireland (NCI) or Coláiste Náisiúnta na hÉireann (CNÉ) in Irish is a third-level education college in Dublin. Founded in 1951, it offers full and part-time courses from certificate to degree and postgraduate level in areas related to commerce, industry, and management.

CDC: The Centers for Disease Control and Prevention (CDC) is the leading national public health institute of the United States. The CDC is a federal agency under the Department of Health and Human Services.

Time Series Analysis: is the use of a model to predict future values based on previously observed values.

Map Reduce: is a programming model and an associated implementation for processing and generating big data sets with a parallel, distributed algorithm on a cluster.

Statistical Tests: A procedure whose inputs are samples and whose result is a hypothesis. Region of acceptance. The set of values of the test statistic for which we fail to reject the null hypothesis. Region of rejection / Critical region. The set of values of the test statistic for which the null hypothesis is rejected.

T-Test: A t-test is any statistical hypothesis test in which the test statistic follows a Student's t-distribution under the null hypothesis. It can be used to determine if two sets of data are significantly different from each other.

SEMMA: SEMMA is an acronym that stands for Sample, Explore, Modify, Model, and Assess. It is a list of sequential steps developed by SAS Institute, one of the largest producers of statistics and business intelligence software. It guides the implementation of data mining applications.

CRISP-DM: Cross Industry Standard Process for Data Mining, commonly known by its acronym CRISP-DM, is a data mining process model that describes commonly used approaches that data mining experts use to tackle problems.

Clustering: Clustering is a process of partitioning a set of data (or objects) into a set of meaningful sub-classes, called clusters.

2 System

2.1 Requirements

The following section documents all the Requirements that will be necessary for this project to progress, as the project continues this may grow and changes may happen; if this is to occur an explanation as to why will be provided.

2.1.1 User Requirements Definition

The client requires a study and analysis done on 3rd level students to see how many are entering higher education and if there is a correlation between age and success rates, the objectives will be to acquire datasets on students from 18 years old and on, clean the datasets so we will be left with relevant information such as, ages, schools, degree being studied etc. Datasets attained from free websites such as GOV.org and UCI.edu will be used in this study.

2.1.2 Data requirements

The three main data requirements that will be utilised in order to complete this project will be as follows:

2.1.2.1 The dataset on Alcohol consumption in students which is an open source dataset and has been retrieved from the website (http://archive.ics.uci.edu/ml/). This shows students from 17-22 and the rate of which they consume alcohol. The following is a list of attributes provided with the dataset and what they mean:

- school - student's school (binary: "GP" - Gabriel Pereira or "MS" - Mousinho da Silveira)
- sex - student's sex (binary: "F" - female or "M" - male)
- age - student's age (numeric: from 15 to 22) address - student's home address type (binary: "U" - urban or "R" - rural)
- famsize - family size (binary: "LE3" - less or equal to 3 or "GT3" - greater than 3)
- Pstatus - parent's cohabitation status (binary: "T" - living together or "A" - apart)
- Medu - mother's education (numeric: 0 - none, 1 - primary education (4th grade), 2 â€“ 5th to 9th grade, 3 â€“ secondary education or 4 higher education)
- Fedu - father's education (numeric: 0 - none, 1 - primary education (4th grade), 2 5th to 9th grade, 3 secondary education or 4“higher education)
- Mjob - mother's job (nominal: "teacher", "health" care related, civil "services" (e.g. administrative or police), "at_home" or "other")

- Fjob - father's job (nominal: "teacher", "health" care related, civil "services" (e.g. administrative or police), "at_home" or "other")
- reason - reason to choose this school (nominal: close to "home", school "reputation", "course" preference or "other")
- guardian - student's guardian (nominal: "mother", "father" or "other")
- traveltime - home to school travel time (numeric: 1 - <15 min., 2 - 15 to 30 min., 3 - 30 min. to 1 hour, or 4 - >1 hour)
- studytime - weekly study time (numeric: 1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours)
- failures - number of past class failures (numeric: n if 1<=n<3, else 4)
- schoolsup - extra educational support (binary: yes or no)
- famsup - family educational support (binary: yes or no)
- paid - extra paid classes within the course subject (Math or Portuguese) (binary: yes or no)
- activities - extra-curricular activities (binary: yes or no)
- nursery - attended nursery school (binary: yes or no)
- higher - wants to take higher education (binary: yes or no)
- internet - Internet access at home (binary: yes or no)
- romantic - with a romantic relationship (binary: yes or no)
- famrel - quality of family relationships (numeric: from 1 - very bad to 5 - excellent)
- freetime - free time after school (numeric: from 1 - very low to 5 - very high)
- goout - going out with friends (numeric: from 1 - very low to 5 - very high)
- Dalc - workday alcohol consumption (numeric: from 1 - very low to 5 - very high)
- Walc - weekend alcohol consumption (numeric: from 1 - very low to 5 - very high)
- health - current health status (numeric: from 1 - very bad to 5 - very good)
- absences - number of school absences (numeric: from 0 to 93)

2.1.2.2 A data set on the ages of all college student's male and female that are in colleges across Ireland, this has been provided by HEA and can be found at (http://www.hea.ie/). The following is the list of attributes on this dataset:

- Full-time Enrolments by Age -how many students are enrolled in each age group

- M- How many are male
- F- How many are female
- T- How many there is total

2.1.2.3 The final dataset is in relation to students in NCI college it shows both dropouts and graduates and was deemed the most relevant dataset and therefore was the one this project focused on.

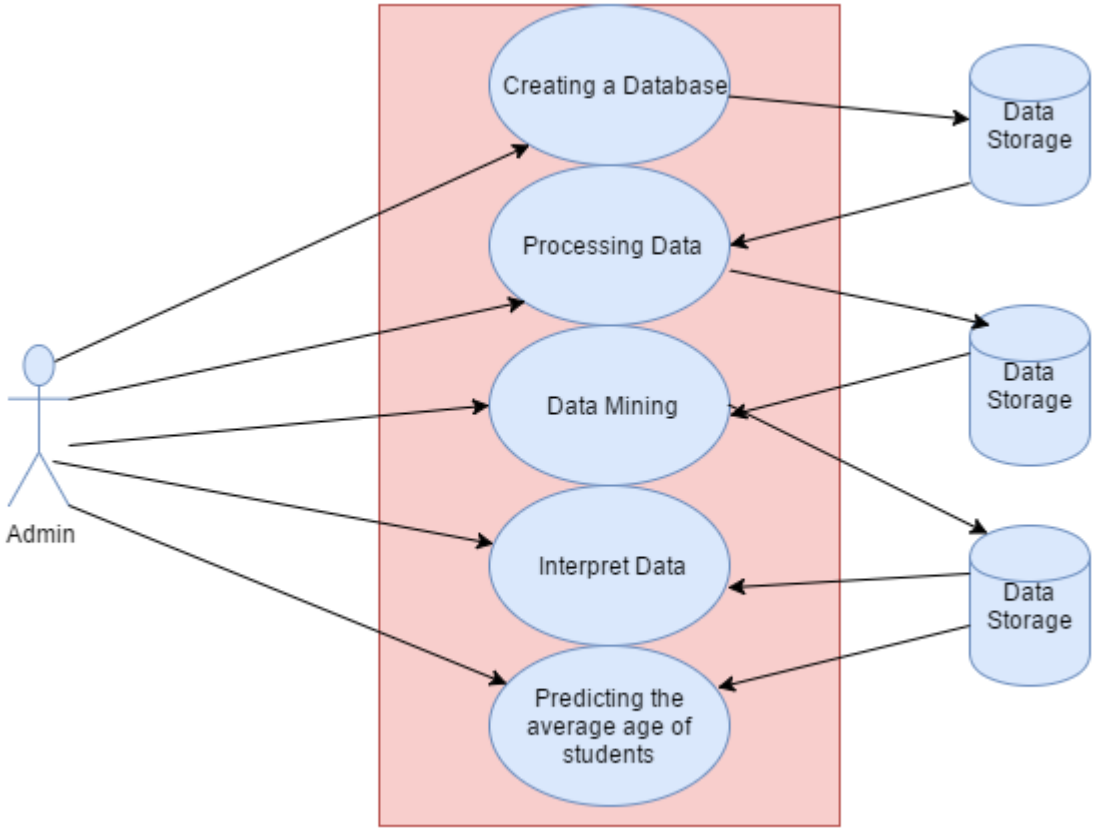
- Reference number- Random Number
- Gender- Male/Female
- Date of Birth- As provided by the student on registration. If null, a default of 01011900 is used. This should not present in this dataset.
- Course Duration- The duration of the course in years.
- Level of Course- Course level of 8 on the National Framework of Qualifications.
- Mode of Study- Full time courses only included
- Course Year- Year one of the programme
- Year Started- Academic Year e.g. 2015 = academic year 2015/16
  Start date of data set is 2007/8, end date of dataset is 2014/15
- Status at the end of first year- One of 3 status: Registered/ Registered and Withdrawn (RW)/Registered and Deferred (RD)
  All 3 status assume that the student started classes
- Withdrawal Reason- If a student has a status of RW, a reason for withdrawal should be present. This may be null particularly in older data
- Date of Withdrawal- This is the date that the student withdrew. If prior to the end of October, of the 1st year, usually discounted as a registration for statistical purposes
- Year Graduated- The academic year that the student successfully completed an award. If null, the student did not graduate. This dataset is restricted to those who did not graduate, or those who graduated on a course at the same level on the NFQ as the course commenced. Year graduated may be later than expected (as calculated from the course duration). This allows the average time of completion to be analysed and take into account students who may defer or have to repeat years, but will still graduate.
- Student Code- This is a data item set by the HEA for data returns. NE is defined as those new to the Higher Education System on entry. By restricting the dataset

to those with a code of NE, repeat students or those that transferred into the course from other courses or colleges are excluded.

2.1.3 Functional requirements

The functional requirements listed are how a user will interact with the system in order to complete this study as I am the main user, my functional requirements will describe how I as the user will interact with the different systems in order to complete this study e.g. The user access programming application, pulls in a dataset from Storage, cleans the dataset by dropping attributes that will be not needed, sends the clean dataset back to Storage and exits application the functional Requirements will be listed in order of importance to the study i.e. 1 being the most important, 2 being next and so on till the functional requirements have been laid out.

2.1.3.1 Use Case Diagram



2.1.3.2 Requirement 1 <Creating a Database>

Description & Priority

This requirement would be considered a level 1 priority and be the first thing that will happen, creating a database is the most important requirement in the study without it there is no place to import the datasets to and from meaning no analysis can take place.

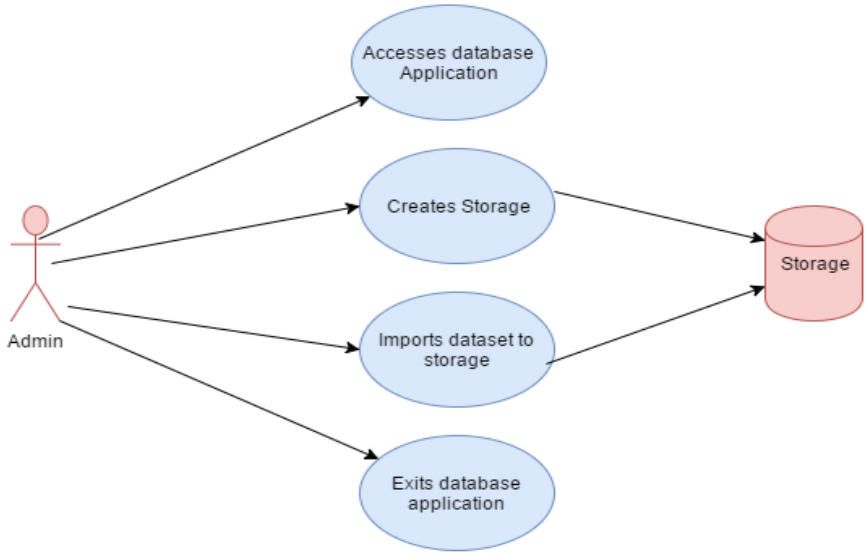
Use Case

The administrator accesses application in order to create a Storage which will house all relevant datasets for the duration of this project.

Scope

The scope of this use case is to create a Storage using a database application; in order to have a location from which I can access all datasets acquired.

Use Case Diagram



Flow Description

Precondition

Database must be accessible if the Storage is to be created by the admin

Activation

This use case starts when the administrator access MySQL opening the application in order to create the Storage,

Main flow

1. The <Admin> opens database application.
2. The <Admin> creates the Storage.
3. The <Admin> imports dataset from local directory to Storage using Programming application
4. The <Admin> exits database application.
5. The <Admin>exits programming application

Exceptional flow

E1 <error has occurred with dataset>

1. The system states and error has occurred and that the dataset is corrupt in some fashion.
2. The <Admin> checks the dataset saved in local directory and finds and corrects the problem.
3. The use case continues at position 3 of the main flow

Termination

The Storage is created within database with the dataset imported; the use case is terminated.

Post condition

The Storage is setup with for datasets imported waiting to be used

2.1.3.3 Requirement 2 <Processing Data>

Description & Priority

This requirement would be considered a level 2 priority, cleaning the data is essential in order to get most accurate figures and statistics.

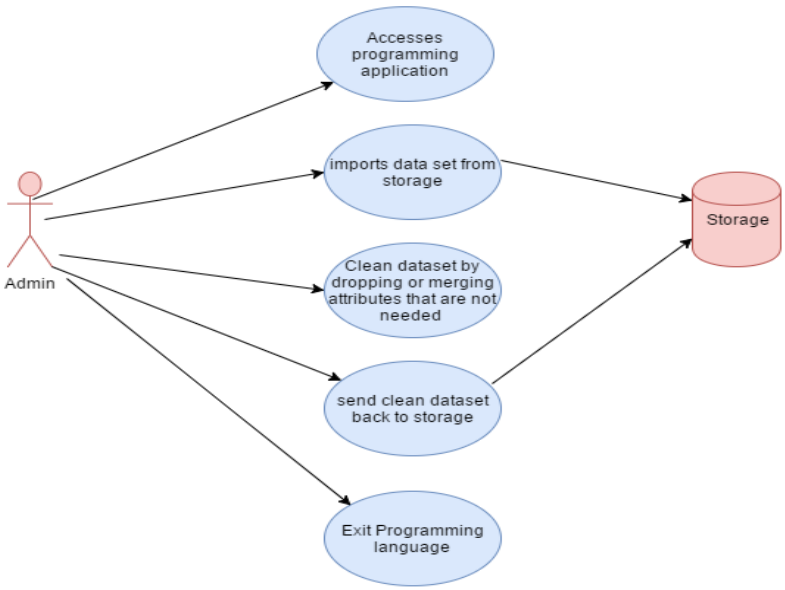
Use Case

The administrator accesses the stored data and begins to strip away all of the irrelevant information.

Scope

The scope of this use case is to clean the dataset in order to throw away all irrelevant information or attributes within the dataset that won't relate to this project.

Use Case Diagram



Flow Description

Precondition

The dataset is in a wait state waiting to be retrieved

Activation

This use case starts when an <Admin>Access the Programming Application.

Main flow

6. The <Admin> accesses Programming.
7. The <Admin> calls in dataset from MySQL.
8. The<Admin> cleans the data creating a new data set with only attributes that are needed.
9. The<Admin> sends new dataset back to the Storage.
10. The<Admin> exits Programming.

Exceptional flow

<Unable to retrieve dataset>

4. The system is unable to retrieve the dataset.
5. The <Admin> checks all file paths are correct.
6. The use case continues at position 7 of the main flow.

Termination

The cleaning has been performed on the dataset the use case is now complete.

Post condition

The new dataset is awaiting retrieval from the Storage.

2.1.3.4 Requirement 3 <Data Mining>

Description & Priority

This requirement is would be considered a level 1 priority, Mining the data while is required for the whole project to succeed and is essential in order to find patterns within the datasets that have been gathered.

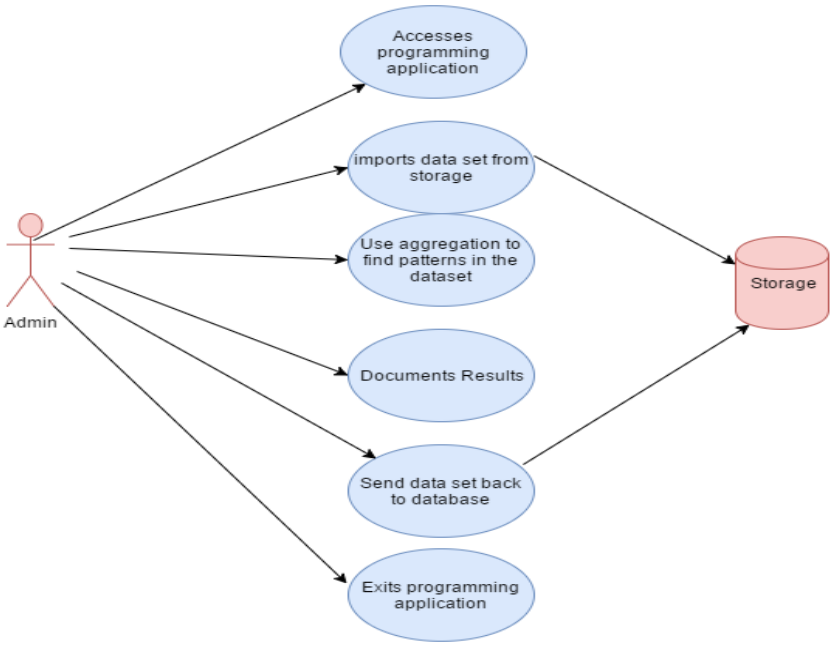
Use Case

The administrator calls in the dataset and uses aggregation to find correlating patterns within the dataset and documents results.

Scope

The scope of this use case is to is to mine deep into the data in order to find patterns within the dataset

Use Case Diagram



Flow Description

Precondition

The dataset is in a waiting state within the storage.

Activation

This use case starts when an <Admin> accesses the Programming application in order to call in the dataset.

Main flow

11. The <Admin> calls in dataset from Database Storage
12. The <Admin>uses aggregation on the dataset in order to find patterns
13. The <Admin> documents results and sends dataset back to the Storage
14. The<Admin> Exits Programming Application.

Exceptional flow

<Programming application error>

7. Programming application can't display graphs and information on datasets.
8. The <Admin> closes programming application.
9. The use case continues at the activation stage before the main flow.

Termination

The dataset is mined and results have been documented the use case is now complete.

Post condition

The dataset is back in a restful state within the Storage ready to be accessed.

2.1.3.5 Requirement 4 <Interpretation>

Description & Priority

This requirement is would be considered a level 3 priority, interpreting the data comes at the end and is the explanation given on the results found. Interpretation is vital in order to present the findings in an easy to understand method.

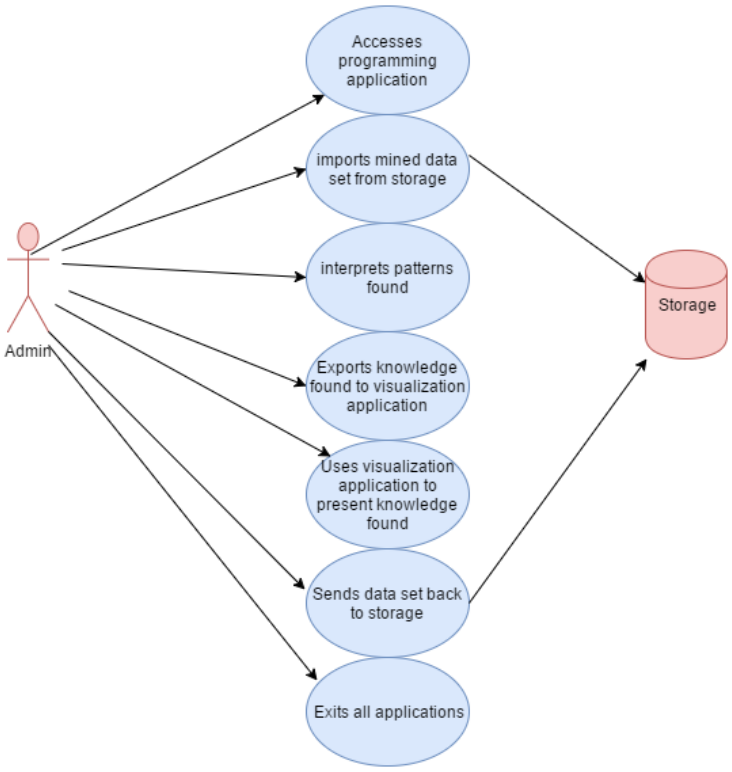
Use Case

The administrator accesses the stored data and begins to interpret the data visually

Scope

The scope of this use case is to interpret and visually display the knowledge gained from the mined datasets gathered

Use Case Diagram



Flow Description

Precondition

The dataset is awaiting retrieval from storage

Activation

This use case starts when an <Admin> Accesses the Programming Application and calls in the mined dataset from storage

Main flow.

15. The <Admin> calls in the mined dataset.
16. The <Admin> Interprets patterns found.
17. The <Admin> exports knowledge to visualization application.
18. The<Admin> uses visualization application to present knowledge found.
19. The<Admin> sends dataset back to storage

Exceptional flow

<Unable to export data>

10. The application is unable to export data.
11. The Admin check data and export destination and fix the problem that has occurred.
12. The use case continues at position 3 of the main flow.

Termination

The dataset has been successful exported and demonstrated the use case is complete.

Post condition

The mined dataset is in a restful state in the Storage.

2.1.3.6 Requirement 5 <Predicting the average age students>

Description & Priority

This requirement would be considered a level 3 priority, implementing algorithms for machine learning to predict what the average age of a student will be in X years' time

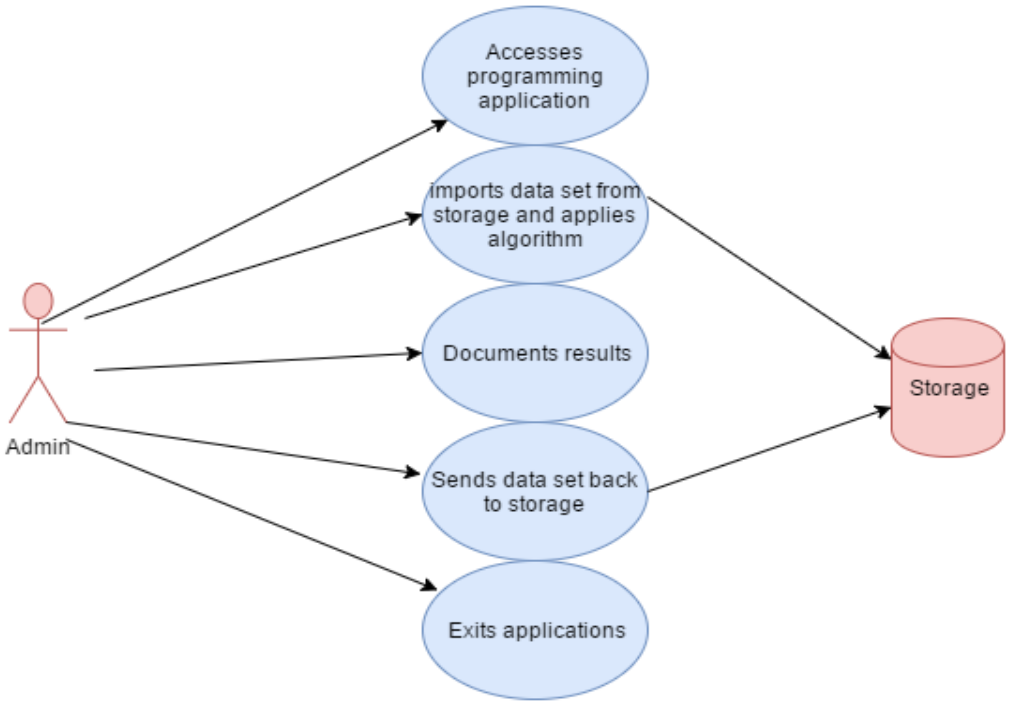
Use Case

The administrator accesses the stored data and applies the predictive algorithm.

Scope

The scope of this use case is to predict the average age of a student in X amount of years' time.

Use Case Diagram



Flow Description

Precondition

The dataset is awaiting retrieval from storage

Activation

This use case starts when an <Admin> Accesses the Programming Application and calls in the dataset from storage

Main flow.

20. The <Admin> calls in the dataset.
21. The <Admin> applies algorithm.
22. The <Admin> documents the result.
23. The<Admin> sends dataset back to storage

Exceptional flow

<algorithm error>

13. The programming application is unable to execute the algorithm.
14. The Admin reviews the algorithm
15. The use case continues at position 21 of the main flow.

Termination

The algorithm has been successfully applied to the data set with the results documented and the data set sent back to storage the use case ends.

Post condition

The mined dataset is in a restful state in the Storage.

2.1.4 Non-Functional Requirements

2.1.4.1 Performance/Response time requirement

Although high performance and response times are seen as a high priority in most system architectures and would be a nice to have implemented with this large dataset it is not fully required for this project. The scope of this project allows the user to analyse the data provided in their own time.

2.1.4.2 Recover requirement

Recoverability is a high priority in this project making sure that all data is fully recoverable in event of hardware failure or server errors. Cloud storage such as Dropbox, GitHub and Google drive will be utilized in order to back up all data relating to this project.

2.1.4.3 Robustness requirement

Robustness is not a requirement for this project. (See 3.2.3) Recover Requirement.

2.1.4.4 Security requirement

The raw data retrieved for this project comes from open source websites such as UCI, GOV.com and GOV.ie and therefore does not have a high security protection. The project which will be developed can only be fully accessed from a personal Laptop that is fully password protected with extra levels of protection coming from database programmes I will be using which are again fully password protected.

2.1.4.5 Reliability requirement

The data sets are composed by government and private assets and are annually updated and maintained. Upon new data been made available the system shall need to be updated to insure reliable and accurate figures.

2.1.4.6 Maintainability requirement

The system is a once off design and needs no maintainability once created.

2.1.4.7 Availability requirement

Data shall remain available to the system throughout the project scope.

2.1.4.8 Extendibility requirement

The project could be very easily extended in the future depending on the question that is to be answered but as of now there is no plans for extension.

2.1.4.9 Resource utilization requirement

Hardware such as a laptop will have to be provided along with internet access and backup storage devices, programming and visualization programs will also be used.

2.2 Design and Architecture

The following diagram shows the Architecture at a high-level view that has been utilized in this project; which is made up of programming applications which will allow for the manipulation of

da stacked-on top of a database used to hold and retrieve data sets, with visualization programs used to display knowledge visually.

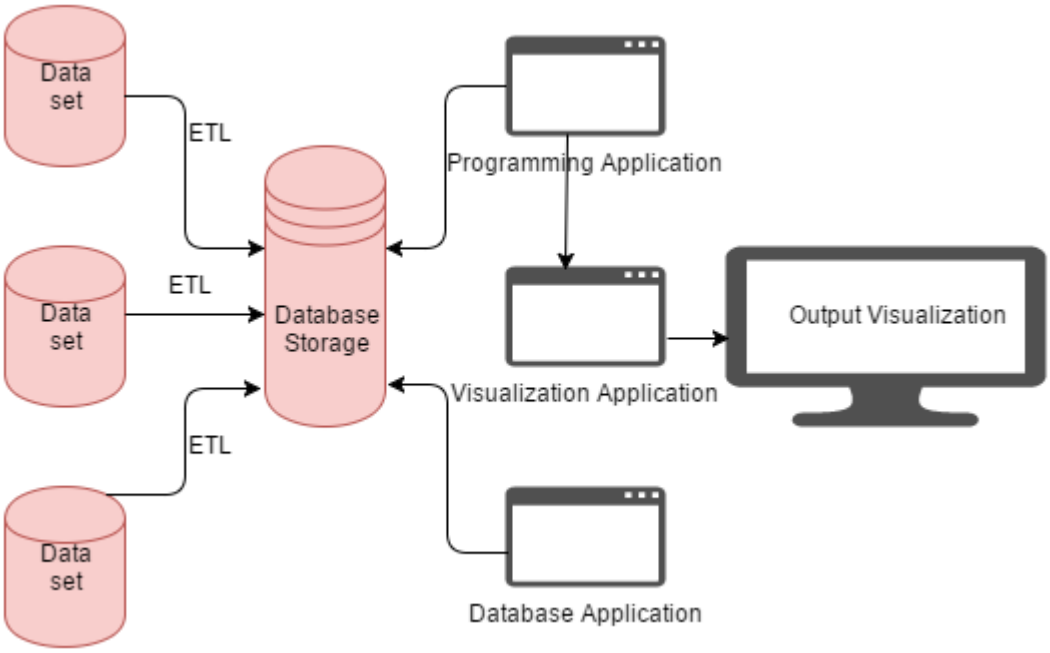


Figure 2 Project Architecture

2.3 Implementation

For the early stages of this project some R functions were utilised in order to get some interesting facts about one of the datasets and to create the database that would be needed for the entirety of this project the following is a few snippets of functions running showing the importing and exporting of the data to a newly created database along with a graph showing the weekend alcohol consumption by age.

The later stages of this project shows how the rest of the KDD was met. The few snippets of show how the data was mined and visually represented.

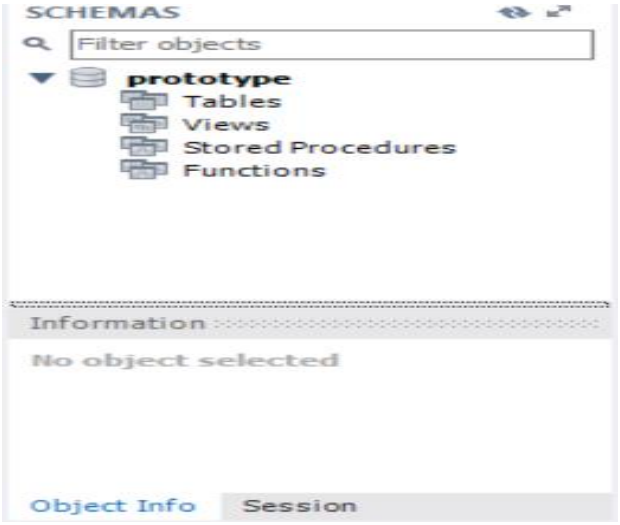


Figure 3: New database

Figure 3: New database with nothing added

```
#set the directory for  orginal data
setwd("D:\\Year4\\Project\\Prototype")

#reading in the original data
data <-read.csv("student-math.csv", sep=";", header= TRUE, stringsAsFactors=F)

#changing the data to a frame structure
data.frame(data)

#Connecting to my hosted database
con<- dbConnect(MySQL(),user="prototype",password= "alan123*",dbname="prototype",host="mysql3.gear.host")


#exporting the orginal data to database in MySQL
dbWriteTable(con, name='test', value=data)

#creating a back up of the orginal data in Mysql
dbWriteTable(con, name='test2', value=data)
```

```
Console D:/Year4/Project/Prototype/

setwd("D:\\Year4\\Project\\Prototype")
data <-read.csv("student-math.csv", sep=";", header= TRUE, stringsAsFactors=F)
data.frame(data)
con<- dbConnect(MySQL(),user="prototype",password= "alan123*",dbname="prototype",host="mysql3.gear.host")
dbWriteTable(con, name='test', value=data)
dbWriteTable(con, name='test2', value=data)
```

Figure 4. Database Connection

Figure 4. Shows the connection to the database and creating new tables within the database.

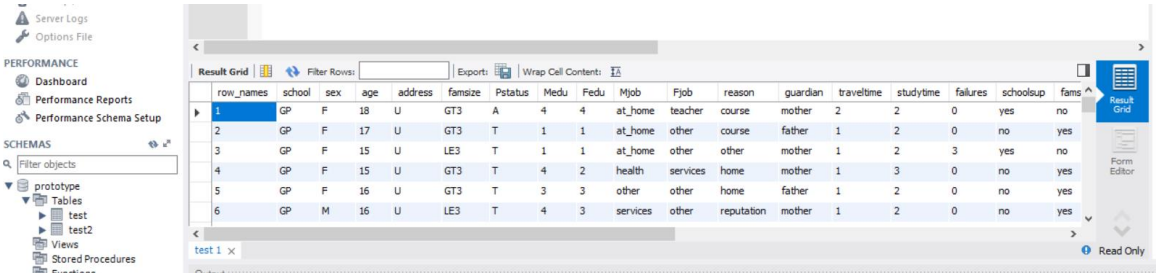


Figure 5: Database Contents

Figure 5: Shows the newly added tables and the contents

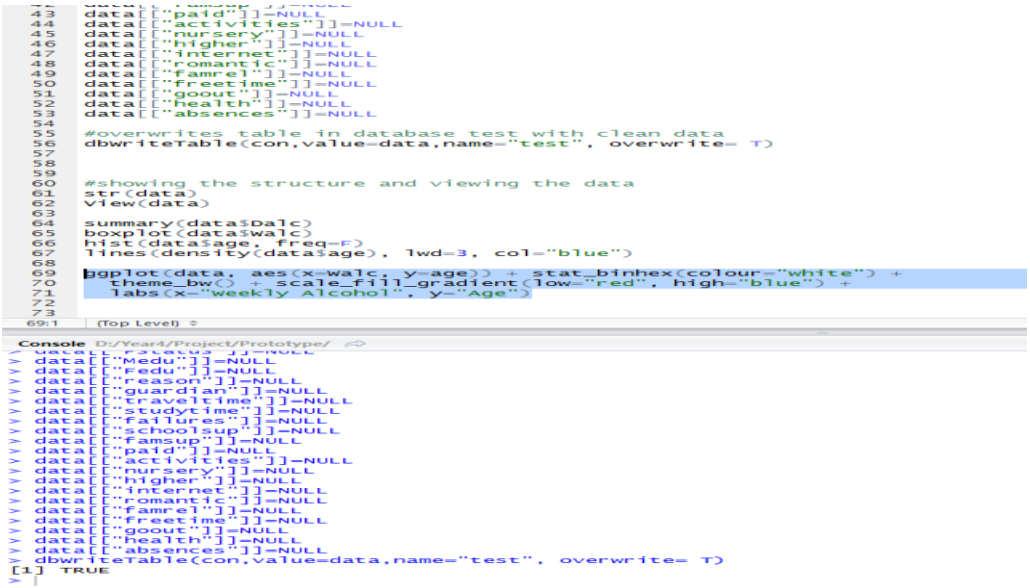


Figure 6 Displaying prototype data

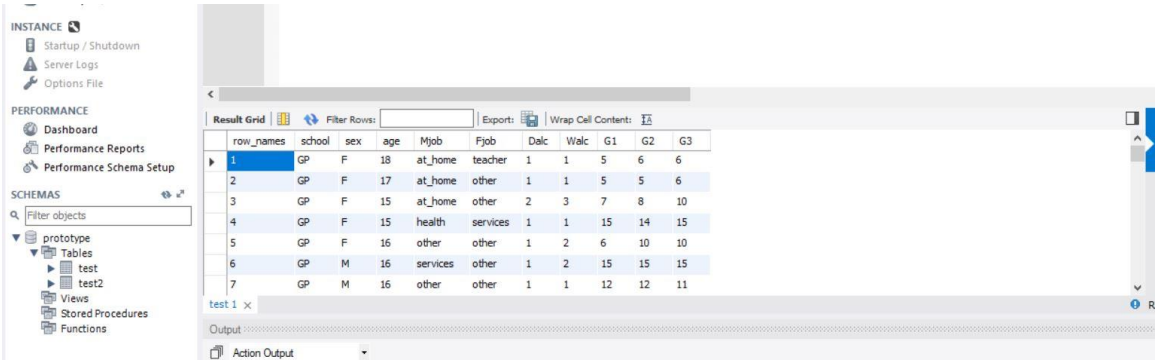


Figure 7 Showing prototype database with new records.

Figures 6&7: Show the cleaning of the dataset and its state in the database after the cleaning

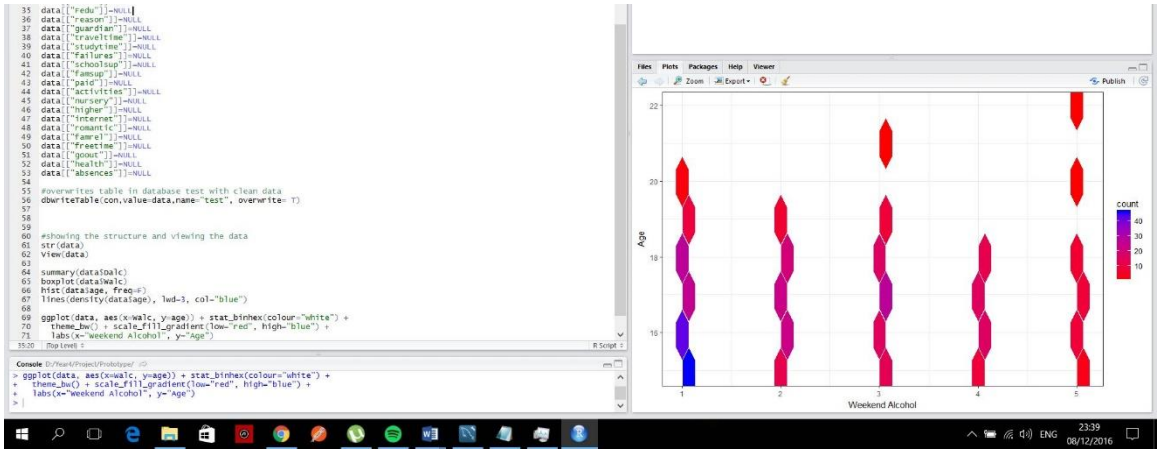


Figure 8 plotting visually prototype data

Figure 8: Shows the amount students consuming alcohol at the weekend by Age with 1 being very low and 5 being very high.

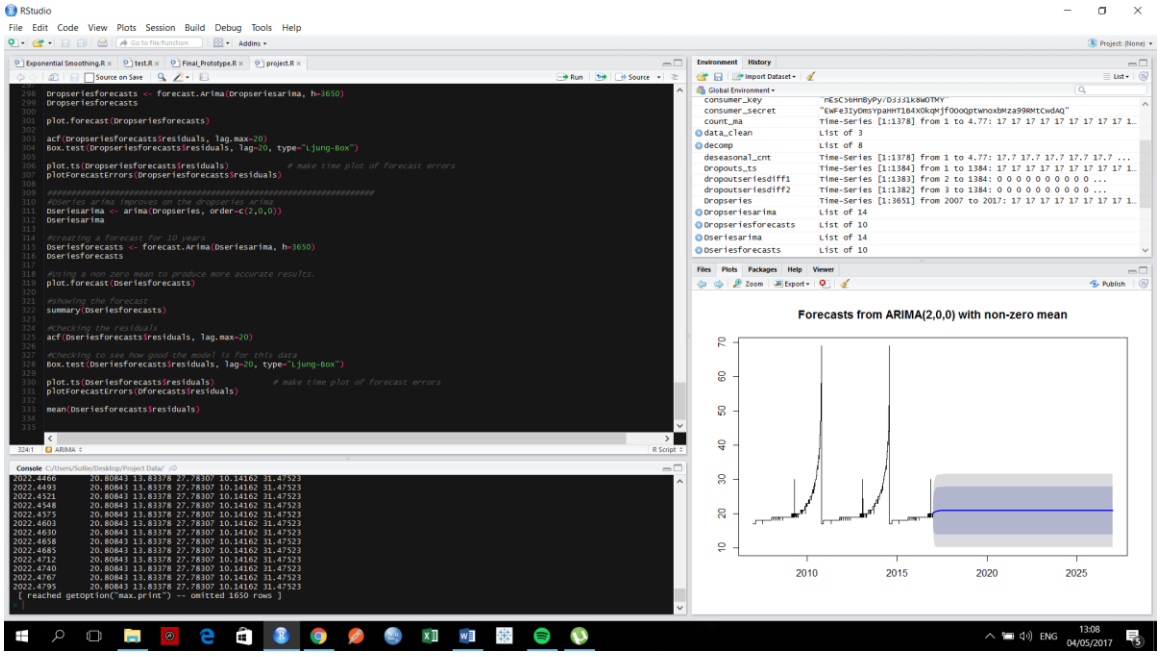


Figure 9 Time series data

Figure 9 Shows the transformation of data into a time series and the data mining process involved with time series showing a predicted average age of student dropouts for 10 years.

The above figures show a representation of how the KDD was implemented in this project. (See 1.5 structure) for a full description of each implementation phase.

2.4 Graphical User Interface (GUI) Layout

No graphical user interface will be implemented into the project at this time. There is scope however for a GUI to be added in later stages of the project provided there is need for one; but as it currently stands this project will be a descriptive analysis producing outputs on datasets that have been acquired.

2.5 Testing

These test carried out are an indicator of how good or valid the data is. As this data anaylsis project rely heavly on the data being valid each test case checks if the data behind the function is correct allowing the user to know that the data here are of good quality.

2.5.1 GG Plots

GGplots are a way of visually describing data within R studios

2.5.1.1 Test Case 1: Male vs Female

Success: In order for this to be a success this GGplot must return how many males gradutated vs how many females graduated.

Failure: This test is a failure if the GGplot fails to return the Expected Result. Check Parameters of the data and test again until success.

Expected Result: The expected result for this plot is the number of males vs the number of females are successfully and coherently displayed in a graph.

Observed Result: The observed result gave back the total number of graduates overall and total number of graduates per year. Test has been successful.

2.5.1.2 Test Case 2: Dropout vs Graduates(Mature vs School Leaver)

Success: In order for this to be a success this GGplot must return how many Dropouts vs Graduates there are and if they are Mature or School Leaver.

Failure: This test is a failure if the GGplot fails to return the Expected Result. Check Parameters of the data and test again until success.

Expected Result: The expected result for this plot is the number of Dropouts vs Graduates both Mature and School Leaver are successfully and coherently displayed in a graph.

Observed Result: The observed result gave back the numbers of how many Dropouts vs Graduates there has been. Test has been successful.

2.5.1.3 Test Case 3: Plots use correct data (plotting dropouts vs graduations)

Success: For this test to be successful the test code should run un interrupted till completion

Failure: This test is a failure if the "testthat" package returns a test failure warning message. Check Parameters of the data and test again until success.

Expected Result: The expected result for this test case is that the code will run without an error message

Observed Result 1: The observed result returned an error that the data did not match the plot.

Observed Result 2: The observed result was that the code ran without any interuption. Test Successful.

2.5.1.4 Test Case 4: Plots use correct data (plotting Males vs Females)

Success: For this test to be successful the test code should run un interrupted till completion

Failure: This test is a failure if the "test that" package returns a test failure warning message. Check Parameters of the data and test again until success.

Expected Result: The expected result for this test case is that the code will run without an error message

Observed Result 1: The observed result returned an error that the data did not match the plot.

Observed Result 2: The observed result was that the code ran without any interruption. Test Successful.

2.5.2 Predictive Models

Predictive models are used to forecast or predict values whether they be future values or what class a certain object falls into e.g. Oranges fall into the class of Fruit.

2.5.2.1 Test Case 1: Holts-Winter Exponential Smoothing

Success: In order for this to be a success the model must run from start to finish and return a a forecasted value within a 95% confidence.

Failure: This test is a failure if the model fails to return a forecasted value. Check Model and parameters of data and test again until success.

Expected Result: A forecasted 10 year Age withing the 95% confidence interval.

Observed Result: This test returned a 10 year Age forecast within the 95% confidence interval. Test is Successful.

2.5.2.2 Test Case 2: Arima Forecast

Success: In order for this to be a success the model must run from start to finish and return a a forecasted value within a 95% confidence.

Failure: This test is a failure if the model fails to return a forecasted value. Check Model and parameters of data and test again until success.

Expected Result: A forecasted 10 year Age withing the 95% confidence interval.

Observed Result: This test returned a 10 year Age forecast within the 95% confidence interval. Test is Successful.

2.5.3 Statistical Tests

Statistical tests are a way of showing that there is in fact a difference between population means within a test group.

2.5.3.1 Test Case 1: T-Test

Success: In order for this test to be a success the T statistic or Significance must be the same in Excel, SPSS and R.

Failure: This test is a failure if one or more results fail to match that of the others e.g(if Excell produces a different T Stat to R).

Expected Result: The expected result is that all three will yield the same result.

Observed Result: The results proved the T stat and significance value are the same in all three cases.

See (7.4 Testing: Visual Results) For the output for each of these test represented visually.

2.6 Evaluation

In this section, the project will be reviewed or more so the method applied to the project will be reviewed.

The KDD (knowledge discovery in databases) allows a user to select a data source and easily navigate through a structured path in order to come to a conclusion about the data.

There are other methods that can be used to produce a data analytics project such as SEMMA or CRISP-DM and while both offer a way to substantially garner information and results from database the KDD was chosen for its linear path and ability to clearly outline each step along the way.

The KDD methodology worked extremely well with the datasets used within the environment of this project. It gave a structure that allowed for a definitive start point and end point to the project meaning that there was very little room for the project to lose its way. Each phase of the KDD method allowed for goals or milestones to be implemented meaning that a very clear progression could be seen through the project from choosing a data source and cleaning it to mining and displaying the knowledge gained. Overall, the KDD methodology suited this type of project which can be evidently seen in the results acquired through its use.

3 Results

Each of the results that have been garnered from this study will be split into their own individual section; this will allow the reader to make sense from each experiment as it took place, rather than trying to make sense of it as a whole documented result. Henceforth each result will be referred to as experiment in the title, for example, “Experiment 1 GGplots”.

3.1 Experiment 1 GGplots:

The GGplots are a way of visually representing the data. Whilst it only appears to be a few simple lines of code to produce a result, the majority of work occurs behind the function. In order to actually produce a worthy graph, the data has to go through the data cleansing stage of the KDD and be broken down into more understandable data frames, which is where the months of work came in. Once all of this had been completed, the results yielded some quite interesting figures.

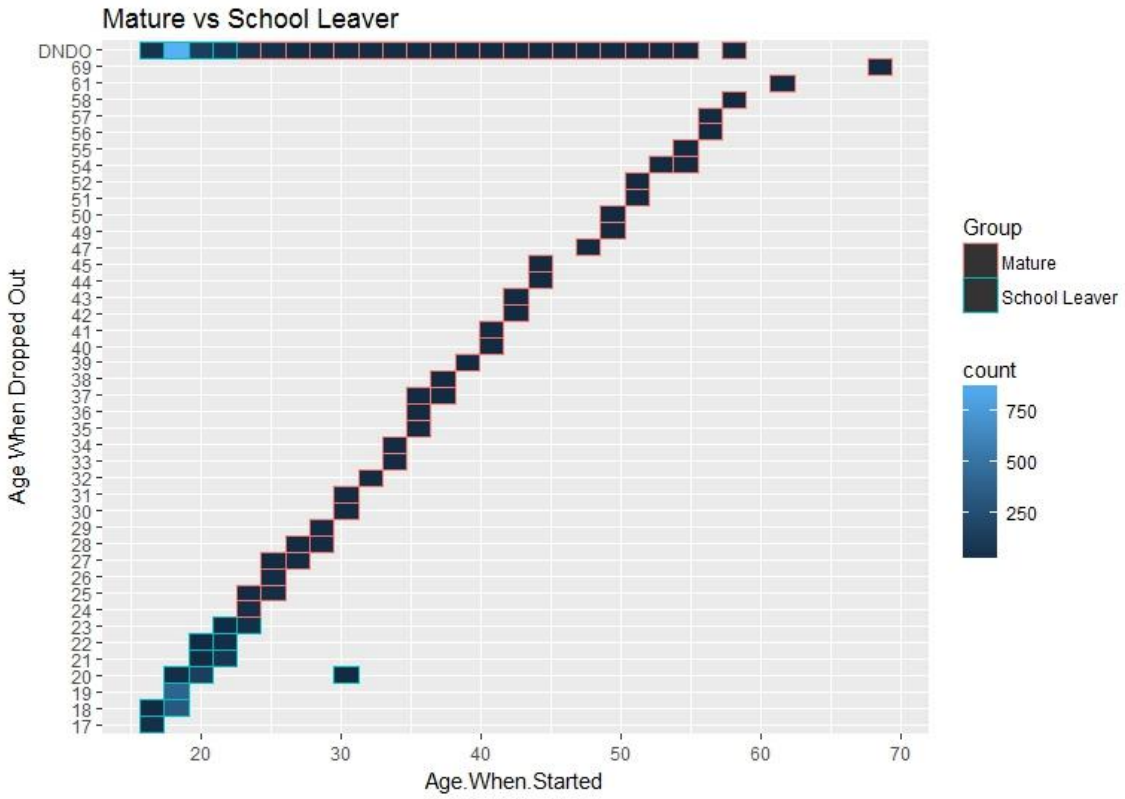


Figure 10 Ages of Students Dropping Out

Figure 10 shows the data on the geom scale, which allows for a multiple factor visualisation of the data, meaning that the age, group, and number of students can successfully be displayed. This visual clearly illustrates that dropouts definitely favour school leavers. These students have an average

age of 18 and a much higher dropout count of 750+ than that of mature students, whose combined dropouts would come in at roughly over 250.

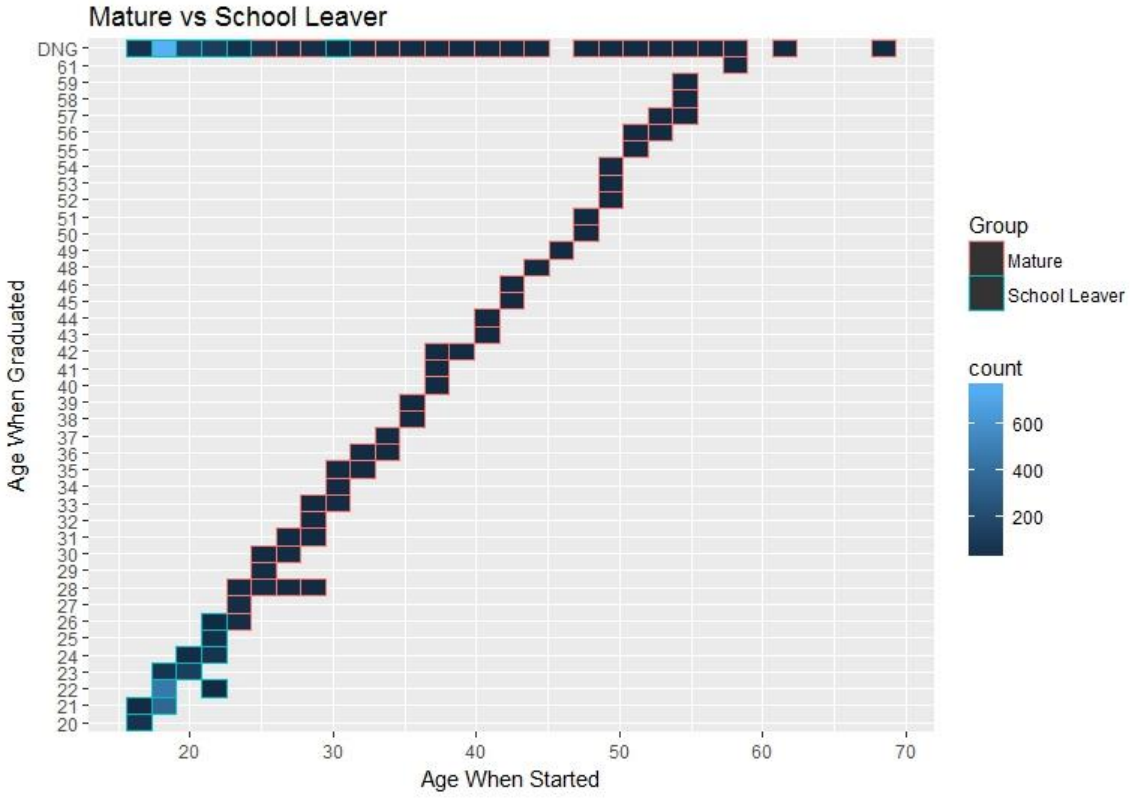


Figure 11 Ages of Students Graduating

Figure 11 illustrates the number of students who have graduated. Whilst this again favours the school leavers, it shows that the average age of graduating is 22-this graph is measured on the x axis by what age the students were when starting college. This indicates that the average starting age of school leavers who successfully graduate is 19; an early indicator that even a marginally more mature school leaver has a higher chance of graduating.

Figures 10 and 11 show the number of students whom have bout dropped out and graduated, their respective ages, and what group they fall into. From this we can deduce that while younger students graduate in higher numbers, they also dropout in higher numbers, and at an alarming figure with 600 hundred being in the region of 18 years of age. These graphs succeeded in answering the overall question of this study-which group set, school leavers or mature students, would achieve better results or graduate in higher numbers. Although it may look like the physical number is better for school leavers, the mature students in fact have a lower entry to dropout ratio. Furthermore, because the sample size of this population is quite large, it can be inferred that if the sample of

mature students was to increase to mirror that of school leavers, there wouldn't be a noticeable change in the amount of dropouts.

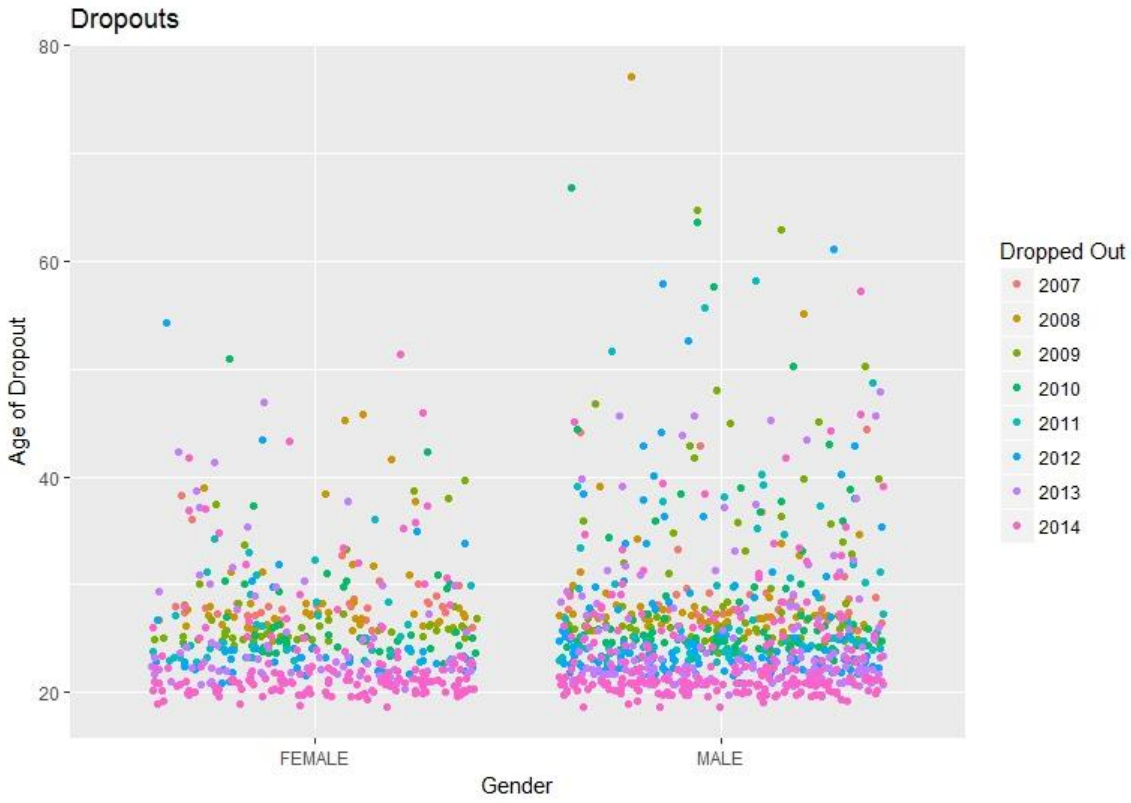


Figure 12 Dropout Male vs Female

Having found an answer to the main question driving this project, it was time to look at another sub category; that of genders Figure 12 shows the dropout ages of males vs females across the 8 years. From this visual, it can be seen that males are at a higher risk of dropping out across all age groups.

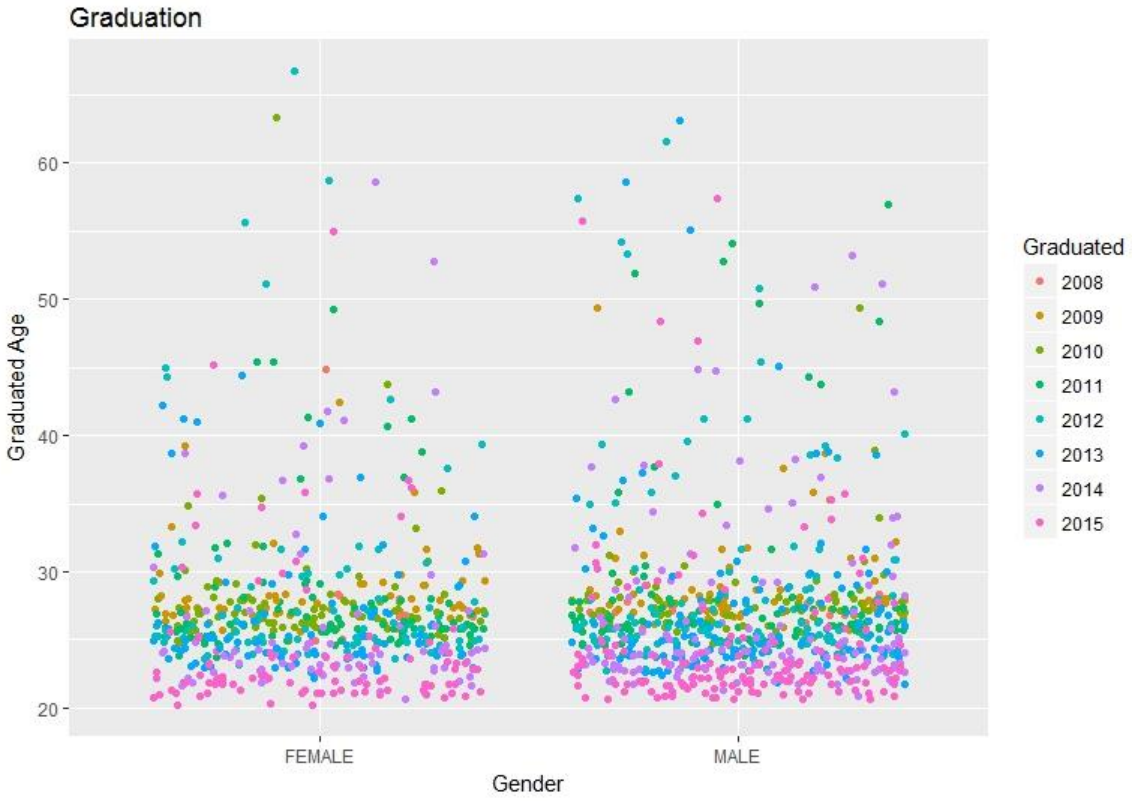


Figure 13 Males vs Females Graduating

Figure 13 illustrates that the number of females and males graduating is a much more even amount. This indicates that although there are more males as a whole in the investigated student population, in numbers females tend to be less likely to dropout.

Figure 12 and 13 show a dispersion of figures, both dropout rates and graduation rates for males and females. The most significant result here, is that as the number of females graduating rises, so too does the male figure. This is mirrored in the dropouts; as dropouts in females rise so too do the dropout levels in males. This could indicate that not only are females less likely to drop out as a whole, but as more progress towards graduation so too do their male counter parts.

Figure 14 reaffirms this, providing a side-by-side comparison of graduation and dropout numbers with each colour being the year, while showing the fraction of the total students who have dropped out or graduated.

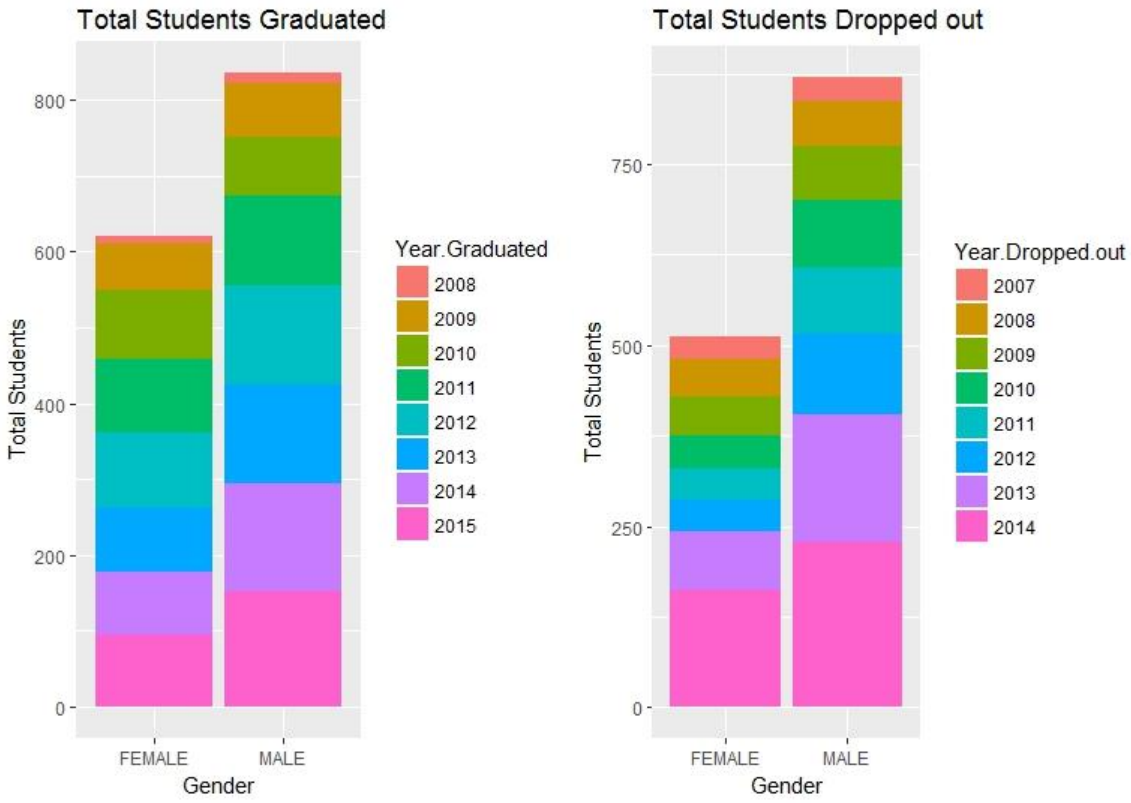


Figure 14 male and female graduation and dropouts

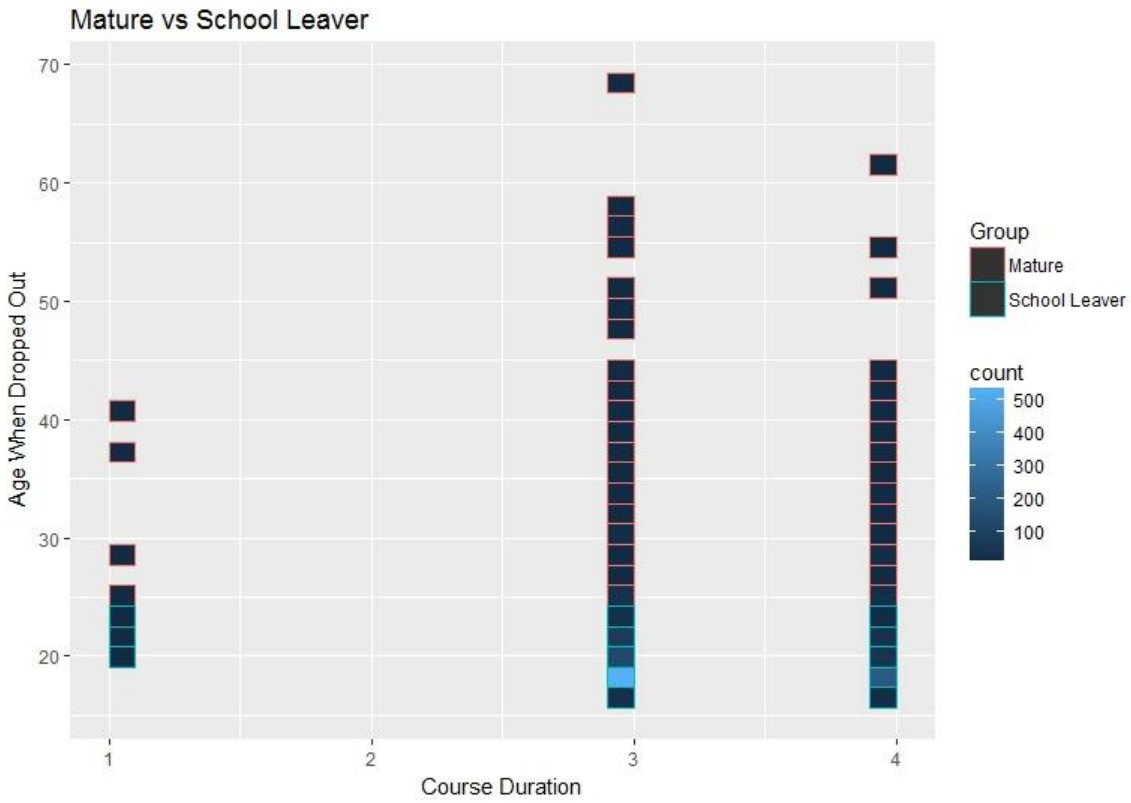


Figure 15 Course Duration Dropouts

Moving forward from these findings, there was still more that could be learned. Consequently, attention was directed towards establishing if the data indicated that the length of a course has influence on dropout rates. This investigation yielded some interesting figures. Figure 15 shows the dropout rates of mature and school leavers by course duration. It shows that courses that last 3 years have a bigger dropout tendency than a 4 year course, with figures of over 500 dropping out from these courses. This accounts for both school leavers and mature students.

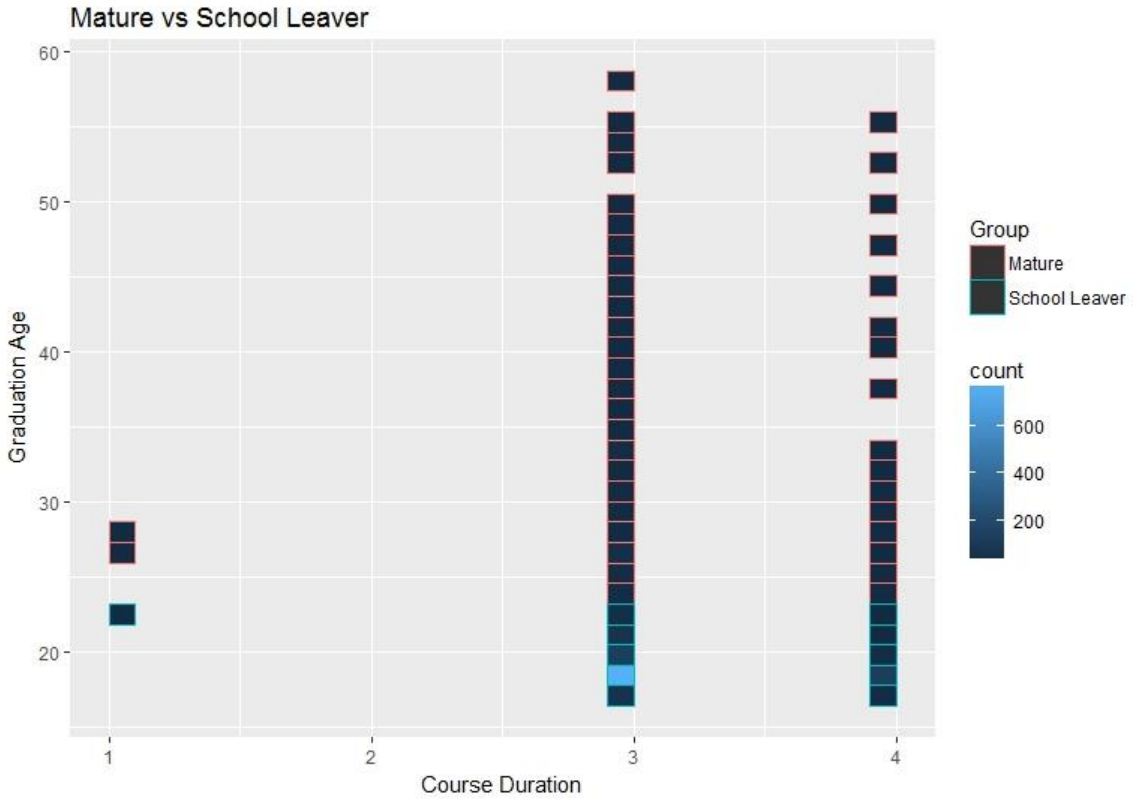


Figure 16 Course Duration Graduation

Figure 16 surprisingly shows that while there is a greater dropout rate in 3 year courses over 4 year there is also a greater graduation rate with over 600+ students graduating across both Group factors.

While both figure 15 and 16 show that graduation and dropout rates are bigger numbers in 3 year courses, this doesn't necessarily mean that 3 year courses are more favourable, as while there may be less graduating in 4 year courses there is also less dropouts.

3.2 Experiment 2 Time Series:

Using two separate time series methods, this project was able to predict the average age of a drop out student to within a 95% confidence level over a ten-year period. The first of these methods was Holts Winters exponential smoothing.

3.2.1 Holts Winters exponential smoothing

Whilst producing a forecast within the confidence bounds, Holts Winters exponential smoothing lacks a certain flexibility that other models have, which will be discussed further on. This lack of flexibility means that only one column in a dataset can be used as the data source for making a prediction, meaning that it's not going to be the most optimal. Isolating the column "Age When Dropped Out" from the newly created dataset named "Dropouts", the exponential model was able to produce a forecast. Utilising the function ACF (Auto Correlation Function) as an evaluation tool which shows variance over time, the accuracy of the model could be measured. The Box-LJung test, a test to determine how well the model fits the data, was conducted to back up the ACF function.

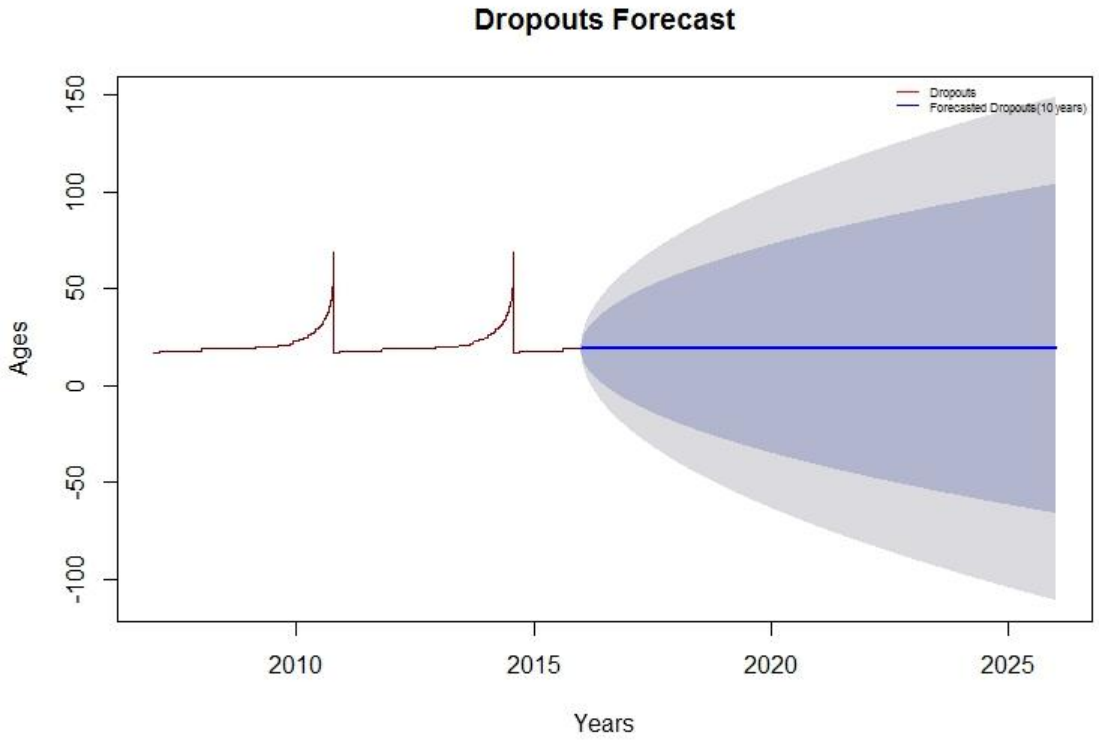


Figure 17 HoltWInters 10 year forecast

Figure 17 shows the predicted forecast for a 10 year period. While it does show a continued trend along the dropout line, the confidence bounds are giving extreme highs within the 95% and lower 80% confidence bounds. This shows that the age over the next ten years of a dropout student will fall within these levels, with it more than likely falling around the trend line.

2021.4740	19	-43.898562780	81.89856	-77.19505063	115.19505
2021.4767	19	-43.914298060	81.91430	-77.21911566	115.21912
2021.4795	19	-43.930029405	81.93003	-77.24317468	115.24317

Figure 18 HoltWinters point forecast

Figure 18 shows the point forecast along with the 95% hi and low Forecast and the 80% hi and low Forecast. From this we can see that 19 is the average age. It gives a +80% range of 81 yrs old and a -80% range of -43, and a +95% range of 115 and a -95% range of -77. The reason for this is because time series must predict a Hi-Low trend.

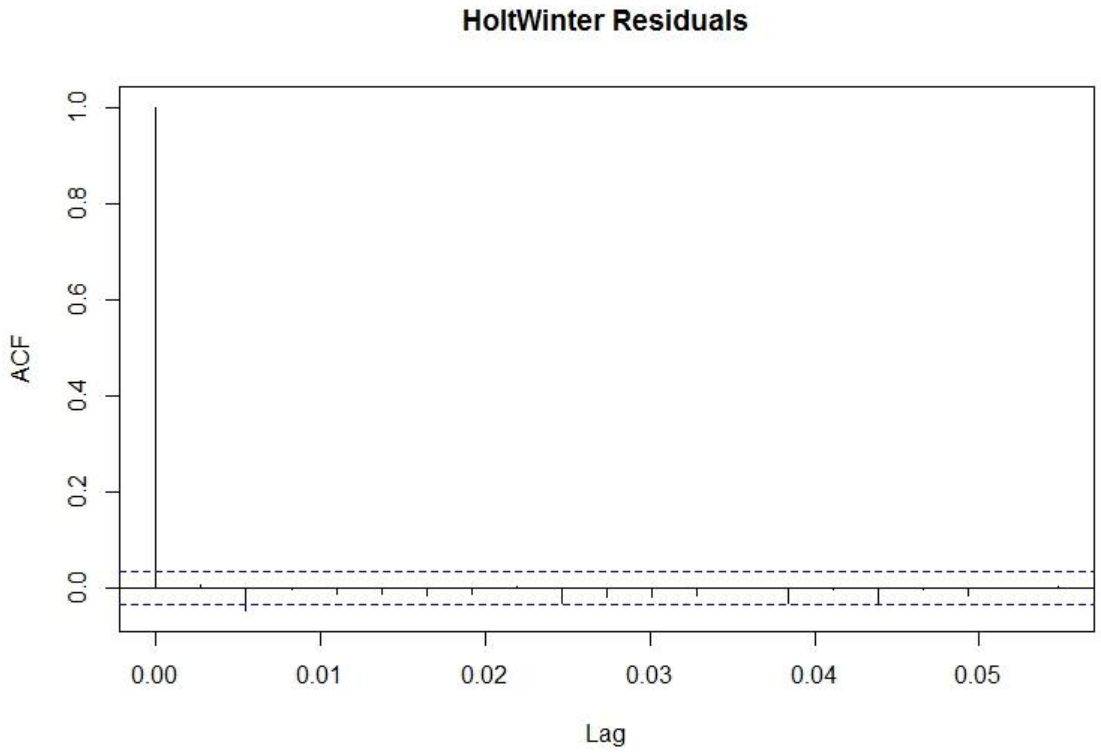


Figure 19 ACF plot for HoltsWinter Residuals

While there is a predicted forecast, just how good that is remains to be seen. Figure 10 shows the auto correlated residuals-it shows that only 2 residuals fall outside the threshold, meaning that this model is a good fit for the data. Supporting this in Figure 19 is the Box- LJung test. This test allows for a more understandable read on how well the model fits the data. First, we set a null hypothesis,

which in this case is that the model is a good fit for the data, and an alternative hypothesis that the models is not a good fit for the data.

```
        Box-Ljung test

data:  Dropoutsseriesforecasts2$residuals
X-squared = 25.36, df = 20, p-value = 0.188
```

Figure 20 HoltsWinter Box-Ljung test

Figure 20 Shows that a P value of 0.19 rounded has been returned. Since the default alpha value is 0.05, there is not sufficient evidence to reject the null hypotheses therefore this model has been proven a good fit.

3.2.2 Arima Model

The Arima model takes into account the entire dataset allowing for a much more accurate forecast. This means that this project can get even closer to a prediction of the 10-year dropout age. The accuracy is again measured using the ACF function, and using the Box Ljung test to back up how accurate the model here is.

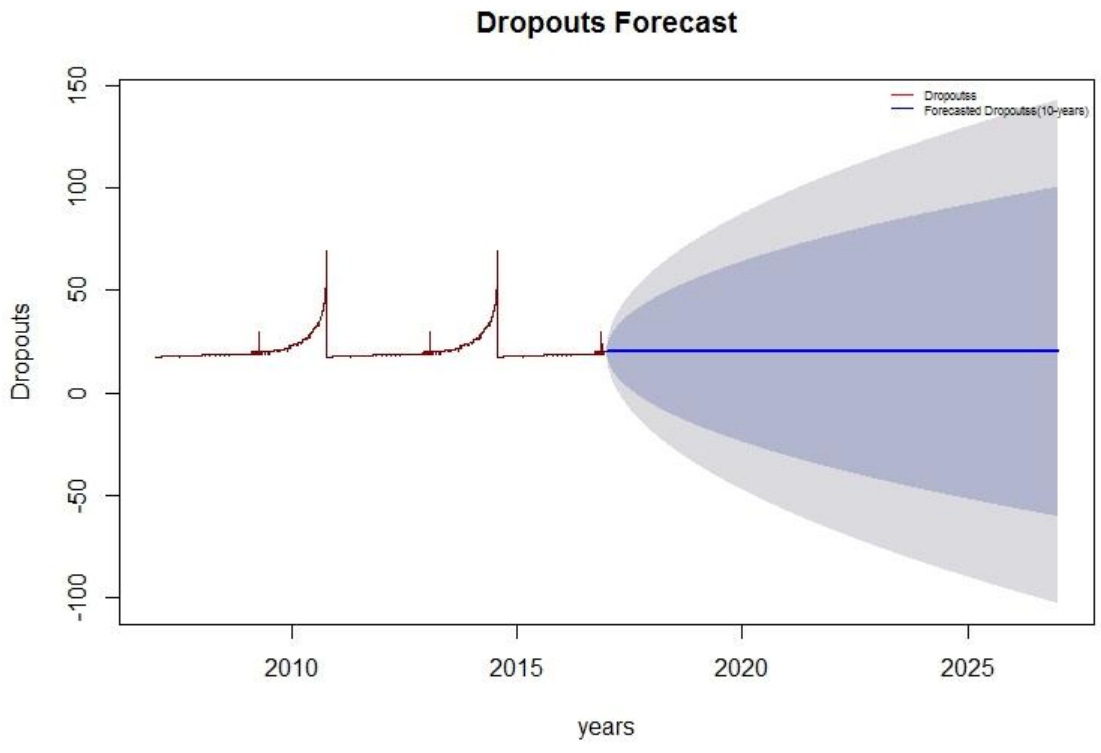


Figure 21 Arima 10 year Forecast

As with the Holts Winter Model, Figure 21 shows the predicted forecast for a 10-year period. Showing a continued trend along the dropout line, the confidence bounds show that the forecast doesn't hit the same extremes within the 95% and lower 80% confidence bounds. While still being quite high and low, it can be seen already that the age is getting more accurate within the boundaries. This is showing that the age over the next ten years of a dropout student will fall within these levels with it more than likely falling around the trend line.

2022.4658	19.99068	-39.31175004	79.29311	-70.70455929	110.68592
2022.4685	19.99068	-39.32660642	79.30796	-70.72728017	110.70864
2022.4712	19.99068	-39.34145908	79.32282	-70.74999536	110.73135
2022.4740	19.99068	-39.35630803	79.33766	-70.77270486	110.75406
2022.4767	19.99068	-39.37115326	79.35251	-70.79540868	110.77677
2022.4795	19.99068	-39.38599478	79.36735	-70.81810683	110.79946

Figure 22 Arima Forecasted Age

Figure 22 shows the point forecast for the Arima, along with the 95% hi and low Forecast and the 80% hi and low Forecast. From this we can see that 19 is again the average age. Although this time the model gives a +80% range of 79 yrs old and a -80% range of -39, and a +95% range of 110 and a -95% range of -70. Already we can see a marked improvement in the model.

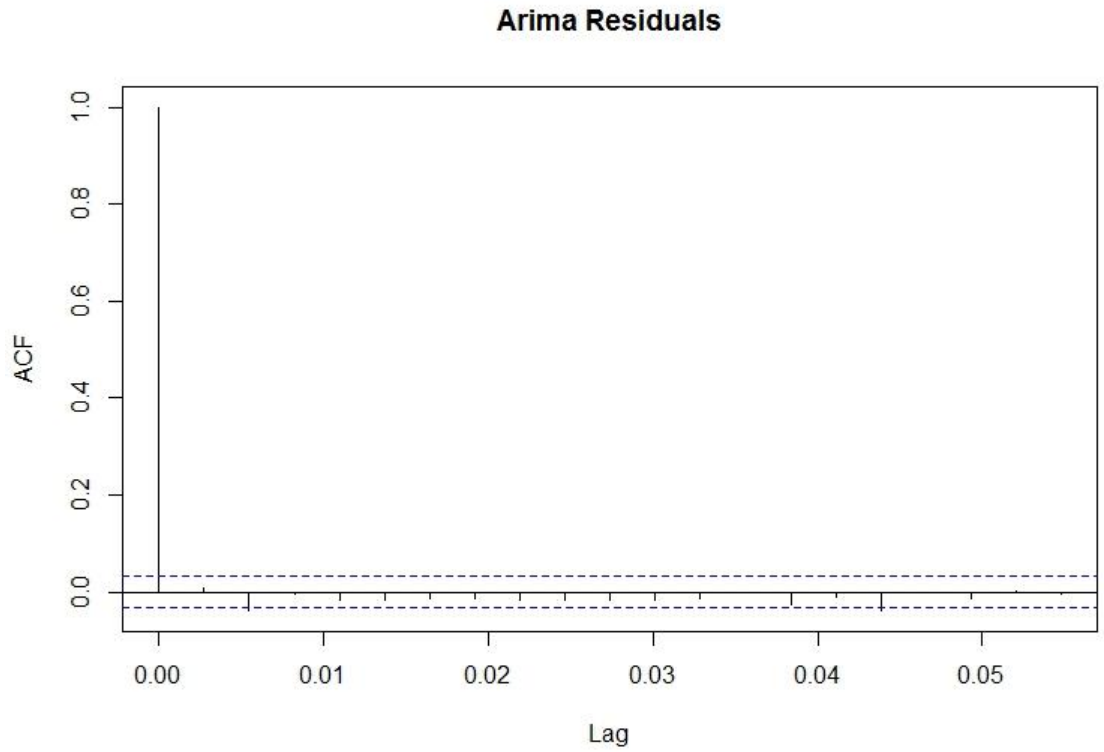


Figure 23 ACF plot for Arima Residuals

Figure 23 shows the auto correlated residuals for the Arima. This shows that again only 2 residuals fall outside the threshold, while the majority of the residuals fall closer to the 0.0 line. This indicates again that the Arima model is a better fit for the data. Figure 13 is the Box-LJung test and will show exactly how much of an improvement has been made. As before, we set the null hypothesis that the model is a good fit for the data, and an alternative hypothesis that the model is not a good fit for the data.

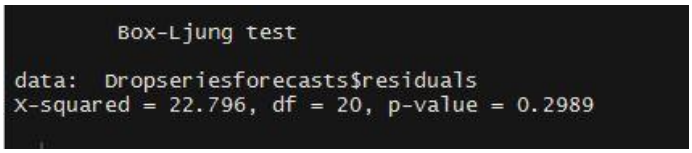


Figure 24 Arima Box-Ljung test

This figure shows a P value of 0.30 rounded has been returned. Since the default alpha value is 0.05, there is not sufficient evidence to reject the null hypotheses, therefore this model has been proven a good fit. While proving that it is a good fit, it shows again that there is a huge improvement in accuracy using the Arima Model over the HoltWinters. With this finding we could stop without the application of any further models. However, not being satisfied with the confidence bounds being so high, there was one more model to introduce in an attempt to bring the extremes down. Another Arima model but measured with non-zero mean.

The non-zero mean Arima is set up identically to the original mean, this time stipulating that the new model should take into account that there is a mean value from the data. The following is what came back from filtering the data through the model.

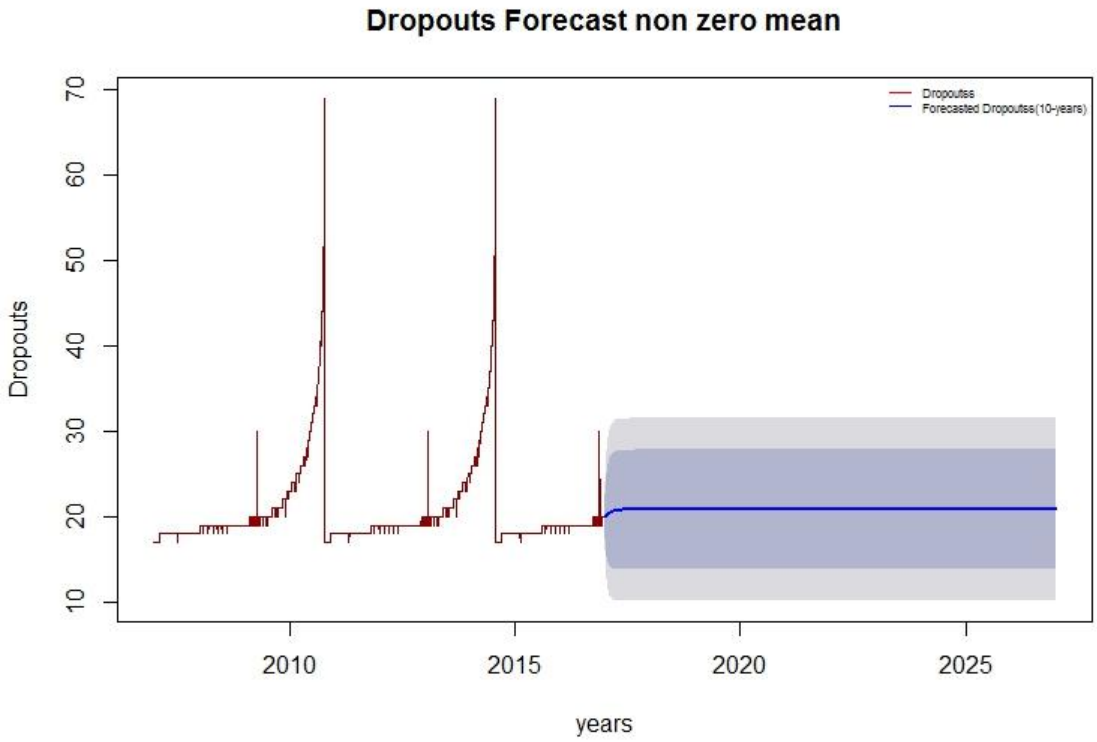


Figure 25 Arima non-zero mean 10 year Forecast

Already a huge difference can be seen in the 10-year forecast. Figure 25 shows that the confidence bounds are greatly decreased from that of the HoltsWinter and Arima models. This shows that the age over the next ten years of a dropout student will fall within these levels, with it again more than likely falling around the trend line.

```
2022.4658       20.80843 13.83378 27.78307 10.14162 31.47523
2022.4685       20.80843 13.83378 27.78307 10.14162 31.47523
2022.4712       20.80843 13.83378 27.78307 10.14162 31.47523
2022.4740       20.80843 13.83378 27.78307 10.14162 31.47523
2022.4767       20.80843 13.83378 27.78307 10.14162 31.47523
2022.4795       20.80843 13.83378 27.78307 10.14162 31.47523
```

Figure 26 Arima non-zero mean forecast

Here we can see a huge improvement in the age that has been forecast. Figure 26 reveals that the average age of a drop out is 20 years, with a +80% range of 27 yrs old and a -80% range of 13, and a +95% range of 31 and a -95% range of 13. The ages here are staring to balance out more, showing how much improvement on the data has been made.

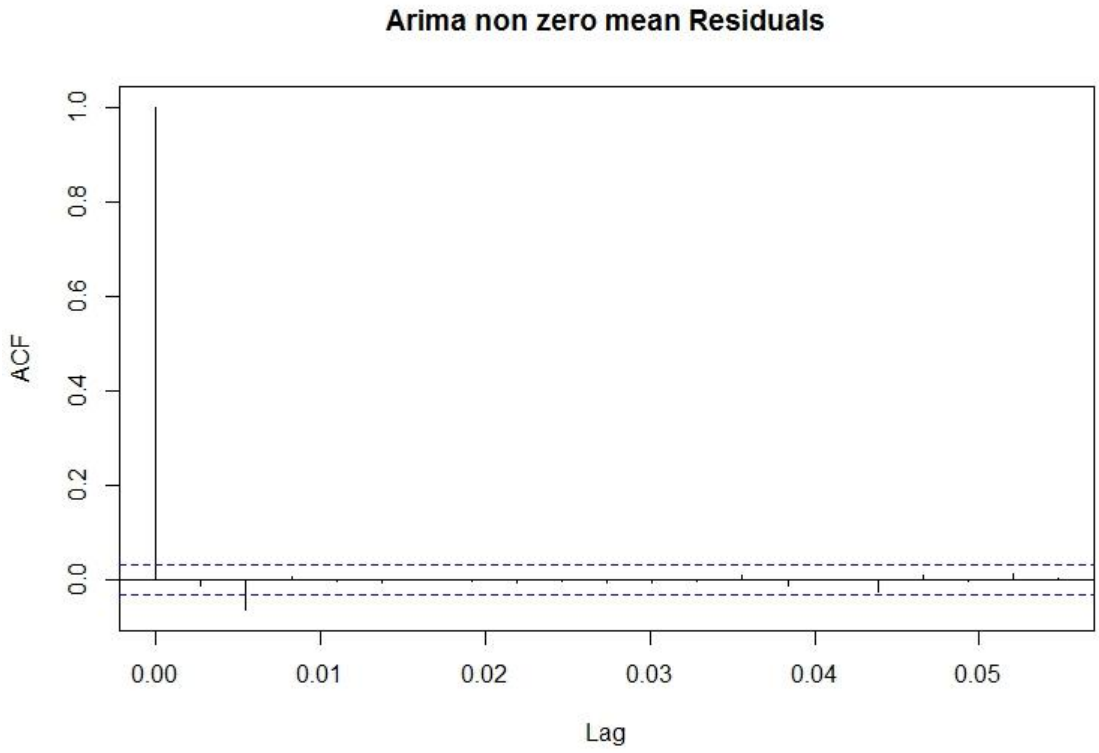


Figure 27 ACF plot for Arima non-zero Residuals

Figure 27 again shows the accuracy of the model through residual representation. It shows that only 1 residual point goes over the threshold bounds while the rest of the residual points stay very close to the 0.0 line. This is a strong indicator that this is the best model for the data. However, before committing to this a Box LJung test is utilised to back up what has been found.

```
        Box-Ljung test

data:  Dseriesforecasts$residuals
X-squared = 20.86, df = 20, p-value = 0.4054
```

Figure 28 Arima non-zero Box-Ljung test

Again, setting the null hypothesis which states that the model is a good fit for the data, and alternative hypothesis which states that the model is not a good fit for the data, the test is conducted. Figure 28 shows the results of this test. We see here that a P value of 0.40 has been returned, which tells us that the non-zero mean Arima model is a much stronger fit than both the HoltsWinter and standard Arima model, making it the best candidate over all when forecasting future dropout ages.

Overall each of the models used could fit the data with an incredible degree of precision, each showing good variance on the ACF residuals and strong P values. However, with the non zero Arima model the age boundary was narrowed down excellently.

Each result shows that if the trend continues, the age of dropouts will stay firmly in the school leaver group, making them more at risk than their mature counterparts. A worrying thought that things like this don't seem to change over time.

3.3 Experiment 3 Map Reduce:

Making use of a Mapper and Reducer, this project took advantage of python to mine the data and pull back the top ten ages from all of the students that dropped out across the 8 years of data. This was done by splitting the data into different years e.g. (07,08,09) and filtering them through the python code created in the notepad++ environment. These results strengthened the hypotheses that there is a difference between mature students and school leavers when it comes to success in college, with the top ten ages of dropouts over the 8 years being 18 years old, as shown in Figure 29.

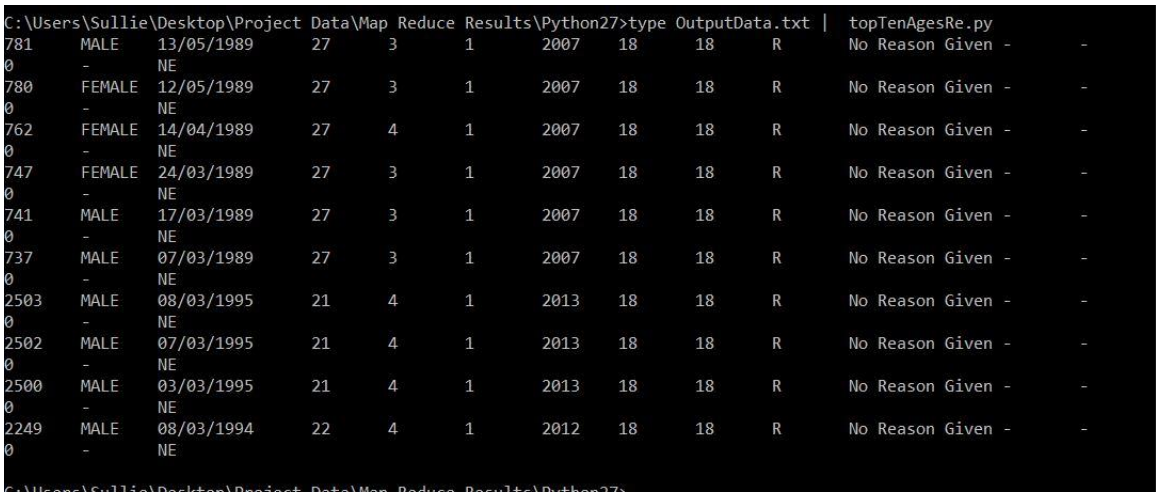


Figure 29 Top ten ages from 8 years

3.4 Experiment 4: Statistical Tests

Statistical tests provide a means of deciphering interesting facts about a dataset. For this project, not many statistical test could be performed because of the nature of the data set. One test known as an un-paired T Test which allows the user to test if there is a significant difference between two groups was utilised. Using a newly created dataset called T-test.csv, four columns were created

taken from the main NCI Completion dataset; the columns were Grad Mature, Dropout Mature, Grad School Leaver and Mature School Leaver. The four columns were then separated into two groups; Grad Mature and Grad Dropout were tested together and Grad School Leaver and Grad Dropout were tested together. The test was performed in excel and found the following.

Setting the null and alternative hypotheses Grad Mature and Dropout Mature:

H0= The null hypothesis states that there is no significant difference between the two groups being tested, Grad Mature and Dropout Mature.

H1= The alternative hypothesis states that there is a significant difference between the two groups being tested, Grad Mature and Dropout Mature.

Setting the null and alternative hypotheses Grad School Leaver and Dropout School Leaver:

H0= The null hypothesis states that there is no significant difference between the two groups being tested, Grad School Leaver and Dropout School Leaver.

H1= The alternative hypothesis states that there is a significant difference between the two groups being tested, Grad School Leaver and Dropout School Leaver.

$\alpha = 0.05$: the alpha value was set at 0.05 which means that there is a 5% chance of committing a type 1 error which is rejecting the null when we should accept- which is deemed an acceptable measure for this type of test.

t-Test: Two-Sample Assuming Unequal Variances		
	Grad Mature	Dropout Mature
Mean	30.94736842	30.24437299
Variance	67.08258451	64.19170211
Observations	247	311
Hypothesized Mean Difference	0	
df	523	
t Stat	1.016813872	
P(T<=t) one-tail	0.154856063	
t Critical one-tail	1.647772343	
P(T<=t) two-tail	0.309712126	
t Critical two-tail	1.964510213	

Figure 30 Grad Mature vs Drop out Mature

Report: (T=1.07, df=523)

Figure 30 shows a mean values age of 30 years for both Grads and Dropouts in the mature group. Tested at an alpha value of 0.05, the T Statistic has been found to be 1.07 from 523 degrees of freedom with a t critical value of 1. Therefore, because the t stat is lower than the t crit, it falls outside the reject region and therefore we do not have sufficient evidence to reject the null hypothesis which states that there is no significant difference between the two groups being test Grad Mature and Dropout Mature.

This means that there is likely no portent age at which a mature student is most at risk to drop out or most likely to succeed and graduate.

t-Test: Two-Sample Assuming Unequal Variances		
	Grad School Leaver	Dropout School Leaver
Mean	18.87014061	19.02516309
Variance	1.136268275	1.201791601
Observations	1209	1073
Hypothesized Mean Difference	0	
df	2232	
t Stat	-3.415660427	
P(T<=t) one-tail	0.00032383	
t Critical one-tail	1.645536604	
P(T<=t) two-tail	0.00064766	
t Critical two-tail	1.961027397	

Figure 31 Grad School Leaver vs Dropout School Leaver

Report: (T=-3.41, df=2232)

Figure 31 shows us a mean values age of 18 years for Grads and 19 years of age for Dropouts in the school leaver group. Tested at an alpha value of 0.05, the T Statistic has been found to be -3.41 from 2232 degrees of freedom with a t critical value of -1.96. Therefore, because this a two tail test and the t critical is both 1.96 and -1.96, it has been found that the t stat is lower than the -t crit and falls inside the reject region. Therefore, we do have sufficient evidence to reject the null hypothesis which states that there is a significant difference between the two groups being tested, Grad School Leaver and Dropout School Leaver.

This could mean that because the ages of the graduates are observed not as they graduate, but rather, as they start college, that even at a slightly more mature starting age of 18 close to 19 years of age, there is a much greater chance that these students will complete the 3 or 4-year cycle. The ages of

dropouts could indicate that any student on this threshold or lower are at higher risk of dropping out.

The statistical tests performed here only further strengthen the overall of this paper, placing school leavers at a much higher risk of dropping out per student over their mature counter parts.

In the following conclusion, the researcher's thoughts on this project will be outlined, and a discussion of the relevant research literature found (see 1.4) will be used to infer a potential reason for the trends drawn from the data set. Finally, the conclusion will offer views on how this project could be improved and suggested scope for future works.

4 Conclusions

In conclusion, based on the dataset provided by NCI, this project has found that there is a definite differentiation in success rates between school leavers and mature students. Although there are more school leavers per year than mature students, meaning there is more graduating school leavers, the disparity between the amount dropping out and graduating between the two factor groups makes all the difference when considering the question of a successful third level education experience. Therefore, the finding that mature students have less dropouts per student than the younger school leaver indicates that this student group are the more successful of the two.

Although no definite reasoning has been investigated as to why this trend is the case, strong inferences of a reason for same can be made based on the reviewed research literature (see section 1.4). To recap, the research indicates that younger students are more likely to feel the strain of college on a financial level more than mature students. It can be inferred that this in turn leads to feelings of social isolation, which could very easily create suicidal tendencies for any individual, let alone a young college student trying to deal with the pressures of college work as well as social, in the absence of the maturity and life experience of their mature counterparts.

Unfortunately, not many mature students would be willing to talk about mental health issues let alone a younger student in the 18-22 years' age group. Failure to discuss and potentially address these issues, is a possible mitigating factor leading to a dropout; the best-case scenario when one considers the other possible negative consequences of such pressure placed on a younger mind.

It was found that not alone do females have less dropouts than males, but as their numbers in graduates rise so too do the numbers of males. The implication of this finding could be huge when considering the reviewed research, which revealed that although females have higher suicidal thoughts, more males act on those thoughts. Considering this finding along with the finding of the current research which indicates greater success for female students, provides an indication that females perhaps deal with the pressure of life, including college, better than males. This information provides a sound rationale for targeting an increase in the number of females in courses within the computer science field, with an aim of reducing dropouts in men and thus dropouts as a whole.

In addition to targeting an increase in the female student body, the findings of this study provide motivation for consideration of other measures that could be put in place with a view of reducing

dropout rates. As it has been repeatedly found that exercise improves concentration and mental happiness, one such measure might include the introduction of a nutrition and exercise class to teach not just younger students, but all students, how to look after the body and mind through proper eating and exercise routines.

Whilst interesting results have been found in the current research, it must be said that this project was limited by the level of access that was given to student records. Whilst the access allowed provided an answer as to the most successful student group, unrelenting access to the entire school database, including grades, would have enabled the researcher to narrow down the reasons why students' dropout and to narrow down an exact age of risk of a dropout, with some coercion of the machine learning models used.

5 Further development or research

Given more time and full unrelenting access to the NCI database, the results of this project could not only show with more accuracy the ages of dropouts and graduates; but also, the reasons why.

There is also scope here for the project to be moved on to a national scale. With the environment being left behind, it is just a matter of filtering the appropriate data into the current study's models to find results.

There is a strong case that the results of this project could lead into collaborative research in the area of mental health which currently dominates public discourse as a key issue. Building on the findings of the current research, the use of a different data analysis method, clustering, in conjunction with the research of other fields (e.g. psychology), may help to open up further discussion of potential mental health issues affecting college students. Future works may serve to strengthen the current irrefutable evidence, which indicates that measures need to be put in place to protect the mental health of students.

This project touched on only one field of study, however, the environment could easily be adapted to find results in other fields of study showing what ages are most at risk of dropping out from each field of study. The implication of this on school resources alone would be huge, allowing them to put into place early detection procedures for students who might be struggling.

The possibilities for further development and research truly are endless in terms of education.

6 References

"About The HEA | Higher Education Authority". *Hea.ie*. N.p., 2016. Web. 17 Nov. 2016.

"Definition Of ADMINISTRATOR". *Merriam-webster.com*. N.p., 2016. Web. 17 Nov. 2016.

"Definition Of STORAGE". *Merriam-webster.com*. N.p., 2016. Web. 17 Nov. 2016.

"Extract Transform Load - Definition". *Techopedia.com*. N.p., 2016. Web. 17 Nov. 2016.

"Github". *En.wikipedia.org*. N.p., 2016. Web. 17 Nov. 2016.

"Graphical User Interface". *En.wikipedia.org*. N.p., 2016. Web. 17 Nov. 2016.

"UCI Machine Learning Repository: About". *Archive.ics.uci.edu*. N.p., 2016. Web. 17 Nov. 2016.

"What Is Dropbox? Webopedia Definition". *Webopedia.com*. N.p., 2016. Web. 17 Nov. 2016.

"What Is Google Drive (Cloud Storage)? Webopedia Definition". *Webopedia.com*. N.p., 2016. Web. 17 Nov. 2016.

"The Definition Of Visualisation". *Dictionary.com*. N.p., 2016. Web. 18 Nov. 2016.

"About Google Scholar". *Scholar.google.com*. N.p., 2016. Web. 9 Dec. 2016.

"Algorithm". *En.wikipedia.org*. N.p., 2016. Web. 9 Dec. 2016.

"Definition: Mysql Database | Motive Glossary". *Motive.co.nz*. N.p., 2016. Web. 9 Dec. 2016.

"Kaggle". *En.wikipedia.org*. N.p., 2016. Web. 9 Dec. 2016.

"Machine Learning: What It Is And Why It Matters". *Sas.com*. N.p., 2016. Web. 9 Dec. 2016.

"Mission". *Tableau Software*. N.p., 2016. Web. 9 Dec. 2016.

"National College Of Ireland". *En.wikipedia.org*. N.p., 2016. Web. 9 Dec. 2016.

"R (Programming Language)". *En.wikipedia.org*. N.p., 2016. Web. 9 Dec. 2016.

"Rstudio". *En.wikipedia.org*. N.p., 2016. Web. 9 Dec. 2016.

"SPSS". *En.wikipedia.org*. N.p., 2016. Web. 9 Dec. 2016.

"Student Alcohol Consumption | Kaggle". *Kaggle.com*. N.p., 2016. Web. 9 Dec. 2016.

"Students In 3Rd Level Education - Google Scholar". *Scholar.google.com*. N.p., 2016. Web. 9 Dec. 2016.

"What Is Knowledge Discovery In Databases (KDD)? - Definition From Techopedia". *Techopedia.com*. N.p., 2016. Web. 8 Dec. 2016.

"What Is Microsoft Excel (Spreadsheet Software)? Webopedia Definition". *Webopedia.com*. N.p., 2016. Web. 9 Dec. 2016.

“Suicide-datasheet-a”N.p., 2016. Web. 11 Dec. 2016.

"Centers For Disease Control And Prevention". En.wikipedia.org. N.p., 2016. Web. 11 Dec. 2016.

A-little-book-of-r-for-time-series.readthedocs.io. (2017). *Using R for Time Series Analysis — Time Series 0.2 documentation*. [online] Available at: http://a-little-book-of-r-for-time-series.readthedocs.io/en/latest/src/timeseries.html [Accessed 8 May 2017].

Anon, (2017). [online] Available at: https://onlinecourses.science.psu.edu/stat510/node/60 [Accessed 8 May 2017].

Anon, (2017). [online] Available at: https://www.cdc.gov/violenceprevention/pdf/suicide-datasheet-a.pdf [Accessed 8 May 2017].

En.wikipedia.org. (2017). *Cross Industry Standard Process for Data Mining*. [online] Available at: https://en.wikipedia.org/wiki/Cross_Industry_Standard_Process_for_Data_Mining [Accessed 8 May 2017].

En.wikipedia.org. (2017). *MapReduce*. [online] Available at: https://en.wikipedia.org/wiki/MapReduce [Accessed 8 May 2017].

En.wikipedia.org. (2017). *SEMMA*. [online] Available at: https://en.wikipedia.org/wiki/SEMMA [Accessed 8 May 2017].

En.wikipedia.org. (2017). *Statistical hypothesis testing*. [online] Available at: https://en.wikipedia.org/wiki/Statistical_hypothesis_testing [Accessed 8 May 2017].

En.wikipedia.org. (2017). *Student's t-test*. [online] Available at: https://en.wikipedia.org/wiki/Student%27s_t-test [Accessed 8 May 2017].

En.wikipedia.org. (2017). *Time series*. [online] Available at: https://en.wikipedia.org/wiki/Time_series [Accessed 8 May 2017].

Wade, S. (2017). *College drop-out rates revealed*. [online] TheJournal.ie. Available at: http://www.thejournal.ie/college-drop-out-rates-revealed-40207-Oct2010/ [Accessed 8 May 2017].

Fortune.com. (2017). *The Real Reasons College Students Drop Out*. [online] Available at: http://fortune.com/2016/03/08/mount-st-marys-firing-simon-newman/ [Accessed 8 May 2017].

Noll, M. (2017). *Writing An Hadoop MapReduce Program In Python - Michael G. Noll*. [online] Michael-noll.com. Available at: http://www.michael-noll.com/tutorials/writing-an-hadoop-mapreduce-program-in-python/ [Accessed 8 May 2017].

Anon, (2017). [online] Available at: http://www.cs.put.poznan.pl/jstefanowski/sed/DM-7clusteringnew.pdf [Accessed 8 May 2017].

7 Appendix

7.1 Project Proposal

7.1.1 Objectives

Objective 1: My first objective is to find a dataset on students enrolled in 3rd level education this will not be limited in anyway by geographical reference this is an open study on 3rd level education and its enrolled students as a whole. In order to find the sets that will be required I will be searching mainly Government related database that host free datasets that will be extremely useful in this study.

Objective 2: My second objective will be to clean the dataset to pull out information I require e.g. age, sex, grades (if possible).

Objective 3: My third objective is to compare information across the years and see if there is an optimal age to start 3rd level education based on how many students are enrolled and or graduating.

Objective 4: My fourth objective using some machine learning based on results try to see if there is a trend in the way ages have varied over the years and see if there will be a certain majority age in 10 years or 20 years' time.

Objective 5: Create an anonymous survey for students in NCI and compare age groups against stressors and use this as a way of backing up what the study itself may show I will also use figures found on suicide rates among young people.

Objective 6: Complete documentation with my views on my study and how if any my ideas on to help improve college life for students as a whole whether that be setting up peer support or hosting talks on the unspoken dangers that lie within the college life.

7.1.2 Background

The reason for my idea stemmed from my own personal experience in 3rd level education and the stresses that come with it. Thinking back to when I was a 19-year-old going through the process of applying for colleges which I eventually made the right choice and stopped; I knew I wasn't ready to approach an academic 3rd level challenge at that time due to circumstances that tore my concentration away from school. When I did return as a mature student I carried with me a better outlook on life and how to approach a college environment as a whole, coming from the working world I had a better grasp on how to manage everything that college throws at you.

From early on in college journey at NCI I found myself asking questions of why do young students put themselves through so much pressure and stress when they are not ready for it? Is there a fundamental flaw with our education system that pushes young minds who are not fully developed with a mental fortitude strong enough for college life into something they don't want to do? Why is it the social norm for kids to open their leaving cert results and be overjoyed with the prospect of another minimum 3-4 years of studying? Do they even know the danger of college? Drugs, alcohol, sex, partying, peer pressure on top of all this project, studying, CAs and for most poor nutrition leads many to the inevitable breaking point of dropping out because these young minds haven't got what their mature counter parts have Life experience. Most like me have done their partying and experienced the world to a degree where they are purely focused on the school side of college with the ability to block out the other social end of it not worrying what people will think of them if they say, 'no I'm not going out tonight'.

I noticed very early on that mature students will have the for lack of a better phrase the 'cop on' to ask for help when they feel the pressure starting to rise, something many of the younger students won't do because they see this as a failure not to themselves but to their peers who they perceive as doing fine without any help and again they end up getting lost in a sea of work they can't keep up with and drop out usually with words of 'oh I just didn't like the course'. How many of these students are then crippled by this, feeling like they have failed and never return to education. Even now in a discussion with friends asking them about their college experiences the general consensus was that they all started college young in a course they didn't enjoy because they felt they had to go to college, but now looking back they knew they weren't ready and as the people they are now and could have done so much better in college and would have chosen what they wanted to actually do.

From this and idea sprang, what if I could collect a dataset of students not just in Ireland but from other countries. Comparing the ages of those who have completed 3rd level education stipulating a general 4-year cycle I could see if there is a trend in what age group would be the best or even what age itself would be the most optimal to enter 3rd level education. I could basing on ages alone see if there will be a fall or increase the mean or average age of students entering 3rd level in 10 or 20 years' time, hoping that with this type of prediction my questions have a good basis. Hopefully answering these questions will yield some good answers and will open up the door for a new approach to college recruitment because according to research done by as research conducted by CDC in 2015 states that among adults aged 18-22 years, the percentages of full-time college students had suicidal thoughts 8.0% or made suicide plans 2.4% ("Suicide-datasheet-a", 2015).So

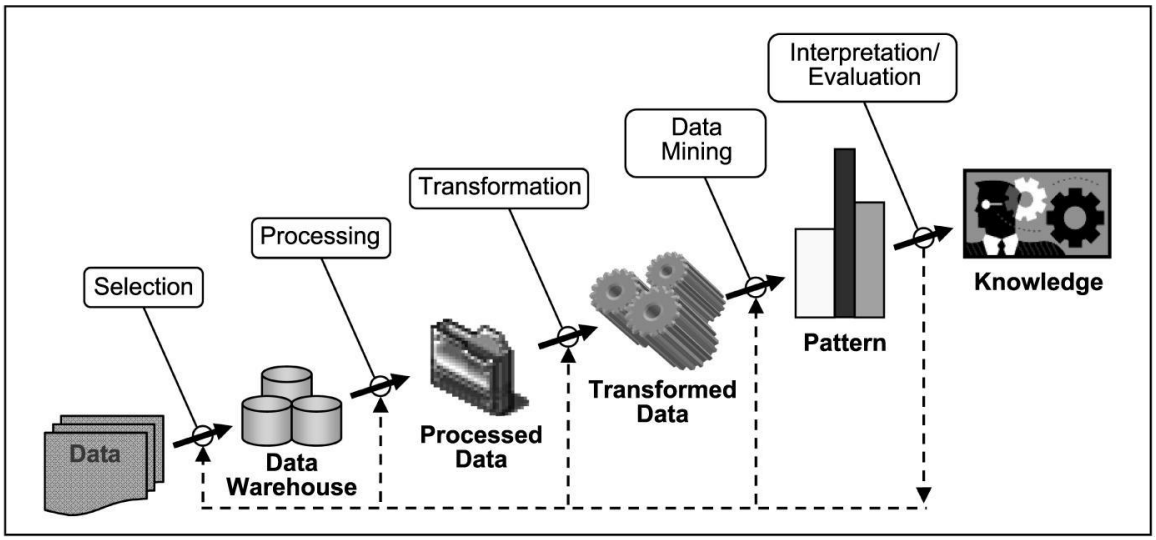
with suicide and depression amongst kids these days in our world so driven by a fast paced social media being so prevalent and everything changing so quickly it might be better to let these young minds know that its ok to breathe and take your time because college is going nowhere and life is also for living.

7.1.3 Technical Approach

In this section I will talk about the Technical Approach to be followed.

KDD

I will approach this project using the KDD (Knowledge Discovery in Databases) approach which has some key steps in order to build a successful data analytics project.



The approach for this is as follows:

- Selection- This involves finding the information and datasets that will be best for my project in terms of what I want my end result to be.

- Processing-This will involve me cleaning the data so that I can get rid of all unwanted data that won't pertain to my project.

- Transformation-This will involve taking the clean dataset and, generating better data, for the data mining stage methods here include dimension reduction, such as feature selection, and extraction, and record sampling, and attribute transformation such as discretization of numerical attributes and functional transformation.

- Data Mining- Here I will use the descriptive data mining technique which includes the unsupervised, and visualization aspects of data mining.

- Interpretation- This is where I will interpret the knowledge gained from the records I have acquired.

7.1.4 Special resources required

As of writing this project proposal I have no required special resources.

7.1.5 Technical Details

R Studios

I will use R Studios to construct this project, which is an open-source integrated development environment for R Language which itself is a programming language for statistical computing, graphics and will be the language I write this project in. R studios will be used in conjunction with excel where I will store me dataset and be used to retrieve information as needed such as ages of students.

R Language

I will use R language to build my project and display information I have not decided which libraries I will use within R studios yet.

SPSS

SPSS is one of the most popular statistical packages which can perform highly complex data manipulation and analysis with simple instructions.

I will use SPSS to back up what I'm displaying using it as a more end user friendly way of viewing the information such as that a person with no technical ability will be able to manipulate the information with built in commands using it as a testing tool will help me to solidify the information I'm looking to produce.

Excel

Excel is a spreadsheet tool with built in statistical and graphing commands that allow a user to manipulate information and data loaded into it.

I will use excel to hold the information in a more structured format until I start to clean it and remove all irrelevant information.

SQL Server

Using SQL Server which is Microsoft's relational database management system supporting SQL queries I will hold my data on this server using it as a way to easily gain access to it no matter where I am. This way I will be able to assure attainability of my data even if I am away from my machine.

Tableau

Tableau allows for instantaneous insight by transforming data into visually appealing, interactive visualizations called dashboards.

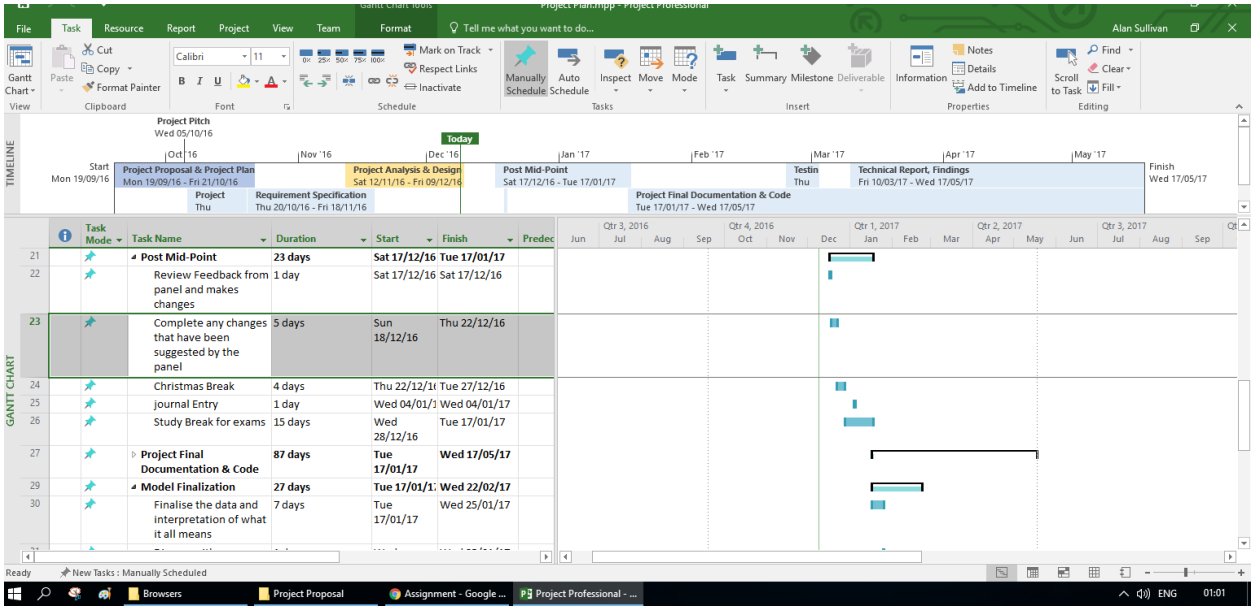
I will use this to Display all knowledge and interpretations of final figures I acquire through my datasets so an end user may easily understand the results which are being displayed.

7.1.6 Evaluation

I will use different environments such as SPSS to test my data accuracy. I will run various other test, such as API testing, Conversion testing, reliability testing and storage testing.

I will test my results by asking a user to navigate through my dashboard, making sure all KPIs are working properly.

7.2 Project Plan



7.3 Monthly Journals

7.3.1 September

Student name: Alan Sullivan

Programme: BSHC in Computing,

Specialization: Data Analytics

Month: 09/2016

My Achievements

This month saw the start of 4th year in the BSHC in computing course in which I hope is a successful one. The first week we arrived back was more of a re familiarize ourselves with NCI week considering I had been away for the last 9 months on a work placement. With that being said there was not a lot to get our teeth into other than getting to know our lecturers for each module and in turn learning what the course work would be like.

The second week was a little more foot on the gas as we started to get our projects for the following weeks given to us. The most immediate one that had to take precedent was a report on "Facebooks system architecture" this was a group project for Out API Development module. The group consists of Myself, Declan Barnes, Sean McDermott and Chris Doran who I've worked with in the past and knew I could rely on them to produce the work. For my part of this project I had to research and talk about Facebooks Architecture which after some sourced information about it, it wasn't too bad. This week also saw the start of both our AI project and report. The report being an individual research report on three techniques used in developing AI for a chess game and the project being develop a chess game in which a player can play against an AI this is the end of semester goal the first part being a player vs player game. I also got the CA project for my Strategic Management class which is to do a report on a cisco case study it is a three-man report with an individual element based around and individual analysis of the case study. Youcef o Connor and Chris are my partners for this project again people I can rely on to get the work done. At the end of week three I had meeting with Michael Bradford a lecturer who specializes in data analytics to discuss my possible project ahead of the planned project pitch, I got a nice dose of reality in this meeting although he liked my idea he said it might need to be restructured in the way I Approached it due to data protection laws, rethinking it was something I spent the weekend doing in hopes of getting around the data set issue.

In week three I had the project pitch which was in front of Eugene O Loughlin, Ron Elliot and Catherine Mulwa, it was my job to convince them that my idea was a feasible study and would have enough substance to be considered a 4th year project. I did a good job selling my idea (thank you Arnotts) and the three where excited to see what I could do while also offering some very good and helpful tips. The rest of the week was uneventful mainly just finishing of any outstanding projects or reports, while getting started on the long-term projects.

My Reflections

There was not a lot to look back on and feel I need to change due to it being such a short month and the projects and reports being so straight forward in saying that I am under no illusion of how tough things are about to get…. fun times ahead.

Intended Changes

Next month I'm going to have to come up with a game plan that can fit in all the study I need to get done as well as time for the gym. An extensive study time table will be needed I think.

Supervisor Meetings

Supervisors have yet to be assigned.

7.3.2 October

Student name: Alan Sullivan

Programme: BSHC in Computing,

Specialization: Data Analytics

Month: 10/2016

My Achievements

The majority of this month was finding a good pace to keep up with the work load that had significantly increased since September. With reports and CAs and projects due this month before the end of first term I really began to feel what it is like to be a 4th year student. The first week was filled with a lot of group meetings with my project groups for both strategic management and web services mainly to see which direction we were heading with our projects and how much time we could allocate without getting to far off the beating road. A good structure was put in place and allowed me to balance the work load that had rested so quickly on the 4th years.

The second and third week of this month was nothing really to report about, a lot of project work for Artificial intelligence, Strategic management and my software project this was coupled with CAs for Data Applications and Business Data Analysis which meant that these two weeks where rinse and repeat. The project proposal for software project was due this month which took more energy then I'd like to admit putting together mainly trying to get my head around that it was a living document and not the final product, which meant I spent too much time trying to get it perfect which at the point meant I was already deviating from my project plan I found this quite Ironic but in the end I was pleased with what I had uploaded even though I know that my project will probably make some significant changes in the coming months I will be happy to look back and see how different the end product is in terms of how it started out.

My final week of this month was just tying of the loose ends of any projects that were due before the midterm which I managed to complete with ample time giving me more time to study over the reading week. I final got to meet my supervisor who was assigned to me this week Frances Sheridan who I only had as a lecturer for a semester in year one but made that particular module very easy to deal with. It was just an overview of what to expect but her calm nature about it made me feel 100 times better about the coming months, she said she's not too far away if we need her for anything and to drop her an email if we want approval of any work we're doing… I feel by the end of this year she will hate me... sorry Frances. The rest of the week was uneventful just a matter of uploading a few projects that I had completed which led me into reading week where finding time to study all of my subjects seemed like an easy task before it but unfortunately, I got weighed down on Web Services trying to fix problems that kept occurring with the tutorials we were given; I did manage to get some work done on my requirements specification but not near the amount of overall work I would have liked to get done. I'll have to make up for it with a few late nights next week, curse of the student life strikes again I suppose

My Reflection

Looking back, I feel that my time spent on each subject need to be shortened I have a tendency to get weighed down for hours on one problem while I should be letting it go and get to work on another subject.

Intended Changes

I intend to focus on more on my software project, it's too easy to say it's not due till next year so I can give it a little breathing space. Hopefully I can start a study group once a week were a few of us will stay back after college and work through any problems we have run into.

Supervisor Meetings

The Supervisor meeting with Frances Sheridan was an informal group meeting just to gauge where everyone is at. One to one sessions are available at request but for the first meeting there was just a few helpful tips about managing time.

7.3.3 November

Student name: Alan Sullivan

Programme: BSHC in Computing,

Specialization: Data Analytics

Month: 11/2016

My Achievements

Back at it; the month of November was a rollercoaster and by far the longest month I've experienced in my entire time at NCI. The reading week fell at the start of this month and was a good chance to get ahead on project work and some extra study but 4th year is where good intentions go to die. The week started out great going over a few things here and there and was progressing smoothly until I opened up Web Services; I'll be the first to say my programming abilities are not the strongest and it's not a module I am particularly interested in but I do want good marks out of the module so I decided to put a real effort into studying it for the CA that was coming up after reading week. It was only after spending hours studying the subject that I realised I knew very little about it, unfortunately for me the way in which I learn made it tough to study. This then took up the whole week and by the end of it I felt even less prepared for the CA then when I started. When the CA did roll around it got rescheduled due to a Citrix problem and then when we eventually sat it well it didn't go so well none of that extra time helped at all. I'm going to need a miracle for the project and exam so my grades don't come crashing down around me.

One thing that I was able to get out of the way was the requirements specification document which allowed me to move on quickly with the technical report.

The second week in November saw a shift in focus away from my main project and onto the CAs that were starting to pile up, the first being in data applications which I ended up having technical

difficulties and had to apply for a resit which was approved and web services which didn't go so well but I learn and move on. Nothing else really happened this week other than touching up my technical document for my main project, I did take the opportunity to meet with Frances and go over the document and ask what needed to be done for the deliverable due in the following week. She was a great help and let me know what I should be looking at getting finished I spent the weekend looking at it.

In week three the first thing I did was meet with Frances again and showed her what I had done which was tweaked and approved for upload. I also discussed how to get a dataset that would help bolster my project from the college she was a great help with that and is still currently doing what she can to secure it; getting this dataset would be the ideal outcome for this project allowing me to go into a great deal of detail, but it is protected making it hard to acquire. This week was the breaking point for many students including myself who just couldn't keep up with the workload but thanks to the efforts of the class reps and the college a lot of deadlines got moved back so the pressure is off for now. Data Application CA was pushed back a week due to the lecturer being sick, the pressure is starting to mount up now.

The final week of this month had 3 main focus points, finishing the prototype and technical report for midpoint presentation, finishing the AI CA4 and Data Applications CA; makes it a long week to say the least. I met with Frances again just briefly to go over what they would like to see in the mid-point presentation and to make sure that the things I left out of the technical document were ok to leave out till next year, I got everything finished and uploaded for that only 2 left. The AI CA was tricky but I done as much as I could with the time was given so hopefully it was enough and finally Data Applications CA which again I had technical difficulties and require a resit oh well who needs sleep anyway.

My Reflection

Looking back at the month I feel there was not much I could have done differently, with 6 modules on this year I am doing my best to get all the work done and dividing my time as best I can. I put a lot of work and hours in and I'm happy about that

Intended Changes

Next month will be the Christmas break; the only changes I will make is to allow for more time on my project work which should be easier with the end of semester fast approaching.

Supervisor Meetings

I had a few meetings with Frances this month and we went over a plan of how I should set up for my presentations as well as how much time I should be giving to each module. The technical document was gone over back to front and both of us were happy with it in the end.

7.3.4 December

Student name: Alan Sullivan

Programme: BSHC in Computing,

Specialization: Data Analytics

Month: 12/2016

My Achievements

December is finally here, this month saw the end of the semester and consisted mainly of just wrapping up projects and going over study material with lecturers for the upcoming exams starting on the 5th of Jan.

The first week I had another meeting with Frances in order to finalise my project mid-point presentation which was marked out of 25% and thanks to her tips and edits I managed to score really high in this which takes a lot of the pressure of moving into semester two. Finalising the presentation and the prototype was my main priority and took up the whole week, no wanting to shift my focus until I was happy with what I was going to show off I made sure to give it the time required. The only other major event this week was going through my project code for the final part of the Chess CA; this was just a short 5 mins with Keith Maycock where I was asked to explain various parts, this went much better than anticipated. So far so good for my modules.

The second week which was the week leading up to Christmas itself, was a busy one. At the start of the week I had to present my prototype which was well received; but at the end of the week had to resit 2 CAs on the same day which didn't give me a lot of opportunity to be prepared as I would have liked but in the end I did very well and was happy with the results. The last few days this week were just for ticking he boxes on 2 projects that were due for Data Applications and Web Services making sure that everything was in order and working properly. Once these were checked and checked again they were uploaded and it was time to take a break for a few days before the madness of study.

The next weeks were Christmas break which meant relaxing and a bit of fun with family and friends; yea right, I should be so lucky, I took a few days off over the Christmas but jumped straight into study mode considering I didn't have as much time to study as I would have liked it was

important stay focused and get my study in, which as of writing this still consumes my life. It's not so bad and I'm nearly there, to think I'll never have to Christmas exams ever again, well until my bright idea to start a Masters but I'll worry about that when and if I get there.

Anyway, there's not much else to report so I should get back to the study even though this entry has been a nice break I can't really extend it any further this month has been rather uneventful considering how the semester went.

My Reflection

Looking back at semester one, I wouldn't have done anything differently, I put the work in and it paid off. My Grades are all good and I have to keep them that way.

Intended Changes

Next month will be the start of semester 2 and I intended to sit down and evaluate what is left to do on my project and make changes where they are needed.

Supervisor Meetings

I only had the one meeting with Frances this month and that was just to go over the mid-point presentation, she gave me pointers on the how the structure should be and looked over my intended prototype. She was happy with it and gave me the green light. I'll sit down with here when I get back and go over what's left to get done.

7.3.5 January

Student name: Alan Sullivan

Programme: BSHC in Computing,

Specialization: Data Analytics

Month: 01/2017

My Achievements

Well it's the start of a new year and the final semester in my four years at NCI, still find it hard to believe I only have 3 Months to go considering I never thought I'd get this far in the first place. January was a bit of an unusual month to say the very least, with exams and transitioning into a new semester my main project took a bit of a back seat.

The first two weeks were all about studying and sitting my exams, which I had four in total and they went well enough. I definitely did enough to make sure I keep my average up and now it sits

just off a 1:1 so hopefully I can make a strong finish to this year and get that first finish. After the exams were over I took a small break for myself to recharge and get my head straight about how I wanted to proceed with my project; this was something that was definitely needed.

With the start of the semester I still didn't know about one of my main datasets that I needed off of NCI in order to really make a proper go of it but on the Monday I received good news off of Frances that my request was approved the dataset I needed would be made available to me. I can't thank Frances enough for her work. I organised a meeting to just finalise a few things with her i.e. what I exactly needed; she invited Jonathan Lynch a member of student support to this meeting because he has good access to the data and was happy to help me which was a great relief because it meant I didn't have to go through it myself. After this I set a meeting with Jonathan so we could go over what structure I wanted the data in.

Starting this semester, I found myself with a lot more time due to the fact that I only have 2 modules this semester which are Advanced Business Data Analysis which is a continuation of the Business Data Analysis from semester 1 with Eugene O Loughlin and Data and Web Mining with Ralf Bierig both are very interesting so hopefully I can do well in them. The rest of the week was just getting back into the swing of things and getting the layout of how the modules will go.

The Week starting 30th/01/2017 in my meeting with Jonathan he suggested that we meet with Sinead O Sullivan who is the Director of Quality Assurance & Statistical Services at NCI and would be the person to talk to about the data needed and what way we could best structure it. This meeting was set up for the following Monday 6th/2/2017 which I gave Frances a quick email about to keep her in the loop of what was happening and to set up a meeting with her for after.

Friday was Results and I'm happy to report I did quite well in my Exams and very well overall in the semester which solidified my grades. These results took a lot of pressure of the rest of the year.

The following Monday 6th/2/2017 I had met with Sinead who I found to be a no nonsense very direct woman, something I like in people she was very pleasant though and more than happy to assist me and after about 40 minutes we settled on the exact data I would need in order to complete my study, job done everyone was happy; next I met Frances to let her know how it went and to get some feedback on my mid-point presentation as well as ask her a few questions about some concerns I had with the documentation for the final parts of the project. As always she went through everything with me and will get the answer to the question she couldn't answer. The rest of the week was just business as usual, consisting of class ad some tutorials.

All in all, I'm feeling confident but somewhat lacking in motivation, I'm sure that will pass though and I'll come through strong. Anyway, there's only 7 weeks left so its game time; no use in complaining just got to get through it

My Reflection

Looking back at the month I feel like I've made huge strides to completing my project as well as solidifying at least a grade of 2:1. All in all a good month/

Intended Changes

Going into the next month I Intend to review my project plan and make changes to allow for more project time.

Supervisor Meetings

This month's meetings where all very positive, Frances has really gone above and beyond to help me, bringing in the right people in Jonathan Lynch and in turn Sinead O Sullivan. We discussed the documents and where I can hover up the easy marks and she also was very quick to tell me to send her any documents I do so she can ok them before I upload anything.

7.3.6 February

Student name: Alan Sullivan

Programme: BSHC in Computing,

Specialization: Data Analytics

Month: 02/2017

My Achievements

This month was a lot slower than I thought it would be in terms of work load, considering by the second month in semester one I felt that I wouldn't have enough time for anything I have now found myself with more than enough time then I know what to do with. This didn't mean that I had any less to concentrate on it just meant I could shift my attention to my main project.

After receiving my main dataset from the college, I got to work on setting it up so I could run analysis on it this included creating a new dataset by adding more columns to the current one. All in all, this took the guts of two weeks to get done due to the size of dataset and how careful I had to be with it. I met with Frances again to go through the next steps of the project and to let her know

how I was getting on and she was happy enough to let me keep going in the direction I was heading in.

The next two weeks I had to park the project as I had a literature review to do for my data and web mining class and due to the restrictions on the page size it meant that this became one of the trickier CAs I've had to do. Once that was finished and handed in I set up a meeting with Frances so I could show her what I have found so far this will fall in the revision week will be the week following this upload.

The final bit for this month was filling in my project profile for the showcase which Eamon rightly so is putting pressure on everyone to get done. This is a 5% that can bring the grade up significantly.

This semester seems to be going by faster than I can keep up and with only three weeks of lectures left the pressure is starting to mount.

My Reflection

Looking back, it was quite a successful month in terms of project work and organisation for the final upload. Given more time I would have prepared more on my dataset but time is a factor that won't wait for me to catch up.

Intended Changes

With a month left to go I have set my final goals for finishing my project which are more in the way of giving it more time.

Supervisor Meetings

The Supervisor Meetings with Frances as always were very productive. Going over what's been done and what needs to be finished. I'm feeling very confident that I'll be able to produce a project with meaningful results.

7.3.7 March

Student name: Alan Sullivan

Programme: BSHC in Computing,

Specialization: Data Analytics

Month: 03/2017

My Achievements

This journal entry will be the end of my 4-year cycle as a student at NCI. There is not a whole lot to report on considering a whole month has gone by. At this late stage in the year the focus has mainly been on finishing up project which only two major ones remain. One is for data and web mining while the other is the main software project.

There has been no real layout for how I approached these projects other than dividing my time evenly between the two.

My meetings with Frances have been very productive in terms of keeping me on track and it is now with confidence I feel I can answer the questions posed at the start of my overall project. Frances has been very good all year with helping me out and keeping me calm and focused on the task at hand.

My final week saw my last official meeting with here as a current attending student, this was just to go over a few points for the final presentation. After the meeting, I put the software project on hold while I finished off the data and web mining project which is due at midnight Friday. The presentation for this project went quite well which takes a bit of the stress off.

Anyway, there's not much else to report on so I'll leave it here. This has been an eventful year to say the least and I am definitely happy to see the back of it.

My Reflection

Looking back there is not much that could of went differently. I managed my time well this month and got work done on both projects.

Intended Changes

Only intended change is to throw my full concentration at the software project

Supervisor Meetings

The supervisor meetings this entire year have been excellent. With Frances's guidance, I am in a really strong position going into my final presentation.

7.4 Testing Visual Results

7.4.1 GG Plots

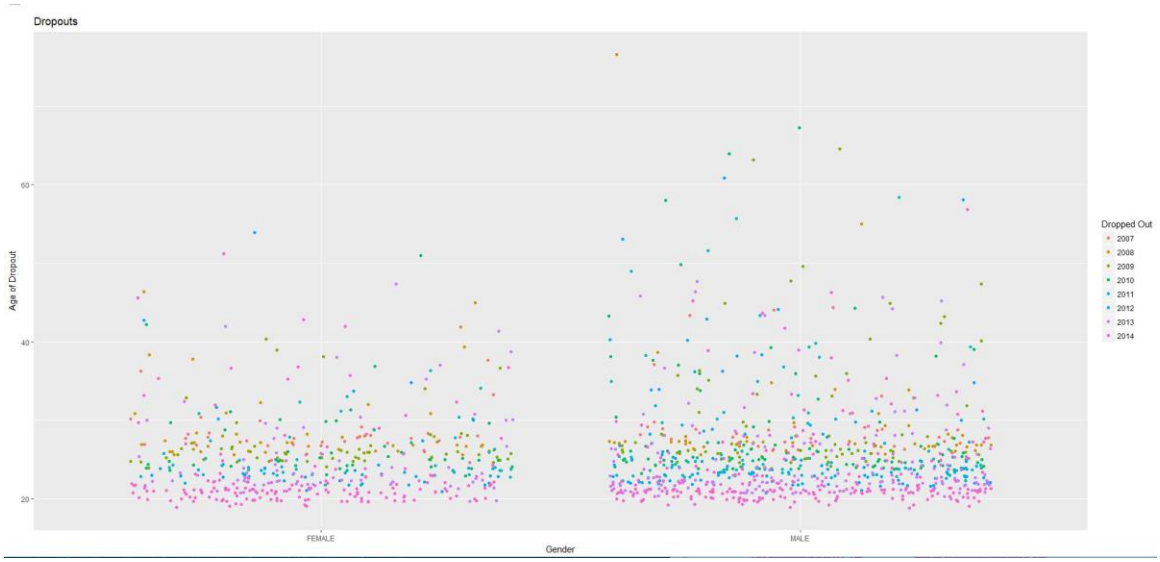


Figure 32 Showing Test Case 1 Success

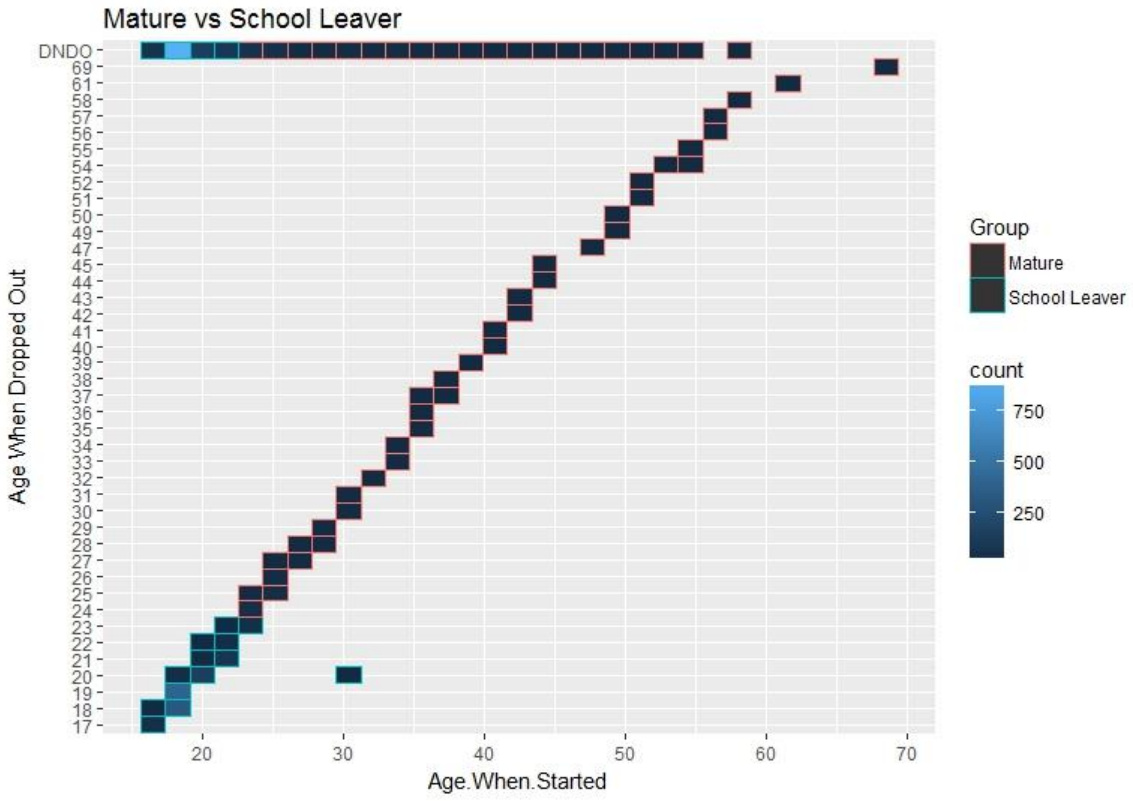


Figure 33 Showing Test Case 2 success

```
Error: Test failed: 'Plot uses correct data'
* lazyLoadDBfetch(c(263556L, 158L), datafile, compressed, envhook) not equal to p$Data.
Modes of target, current: function, NULL
target, current do not match when deparsed
>
```

Figure 34 Showing a Test Case 3 Failure

```
> ##plot is using correct data
> test_that("Plot uses correct data", {
+ u<- a2(Data)
+ expect_that(Data, equals(Data))
+ })
>
```

Figure 35 Showing a Test Case 3 Success

```
Error: Test failed: 'Plot uses correct data'
* object of type 'closure' is not subsettable
1: d2(Data6) at :2
2: grid.arrange(d1, d2, nrow = 1) at :4
3: grid.draw(g)
4: grid.draw.gTree(g)
5: recordGraphics(drawGTree(x), list(x = x), getNamespace("grid"))
6: drawGTree(x)
7: makeContent(x)
8: makeContent.gtable(x)
9: mapply(wrap_gtableChild, x$grobs, children_vps, SIMPLIFY = FALSE)
10: (function (grob, vp)
    {
```

Figure 36 Showing a Test Case 4 Failure

```
> test_that("Plot uses correct data", {
+ g<- c1(Data6)
+ expect_that(Data, equals(Data))
+ })
> TEST SUCCESS
```

Figure 37 Showing a Test Case 4 Success

7.4.2 Predictive Models

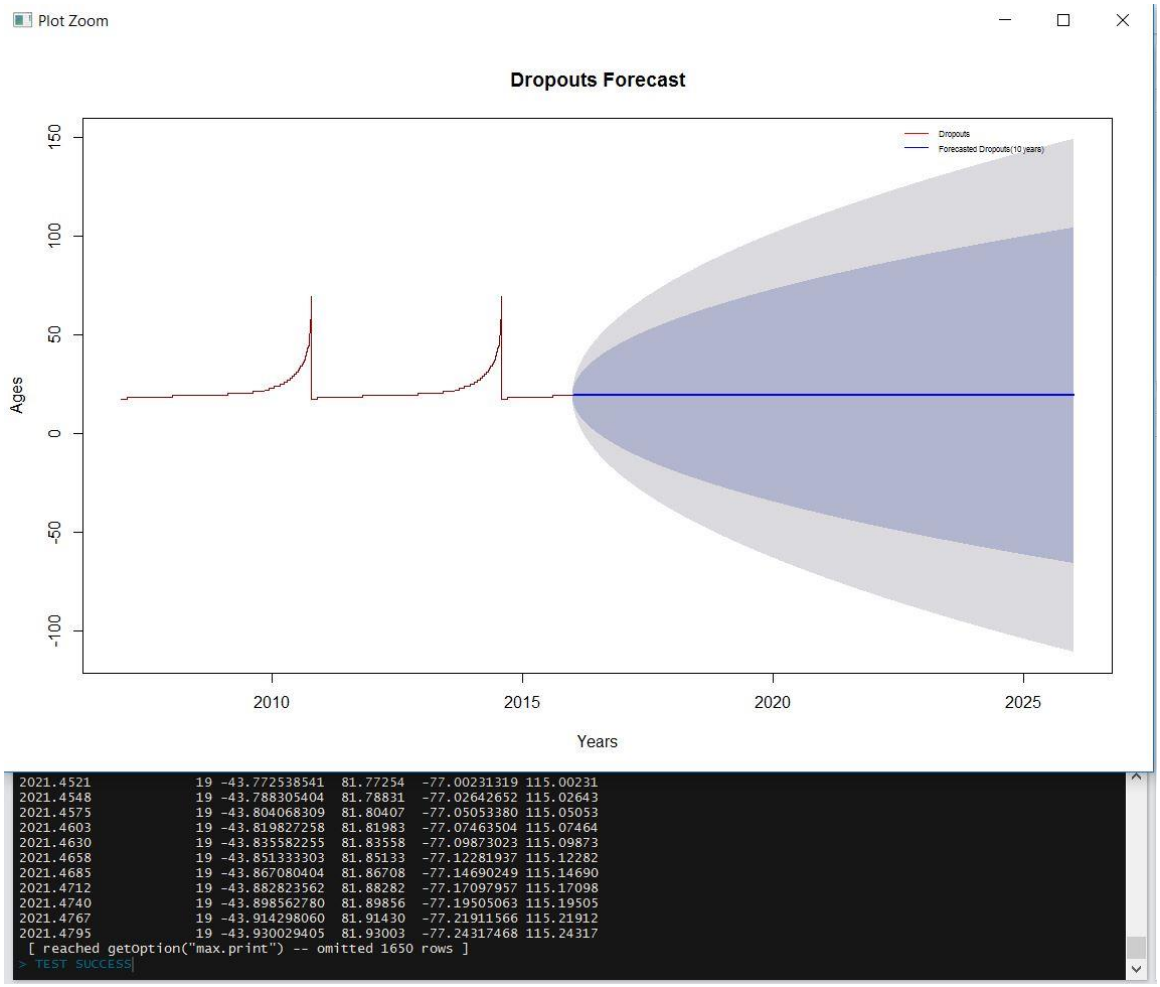


Figure 38 Showing a Test Case 1 Success

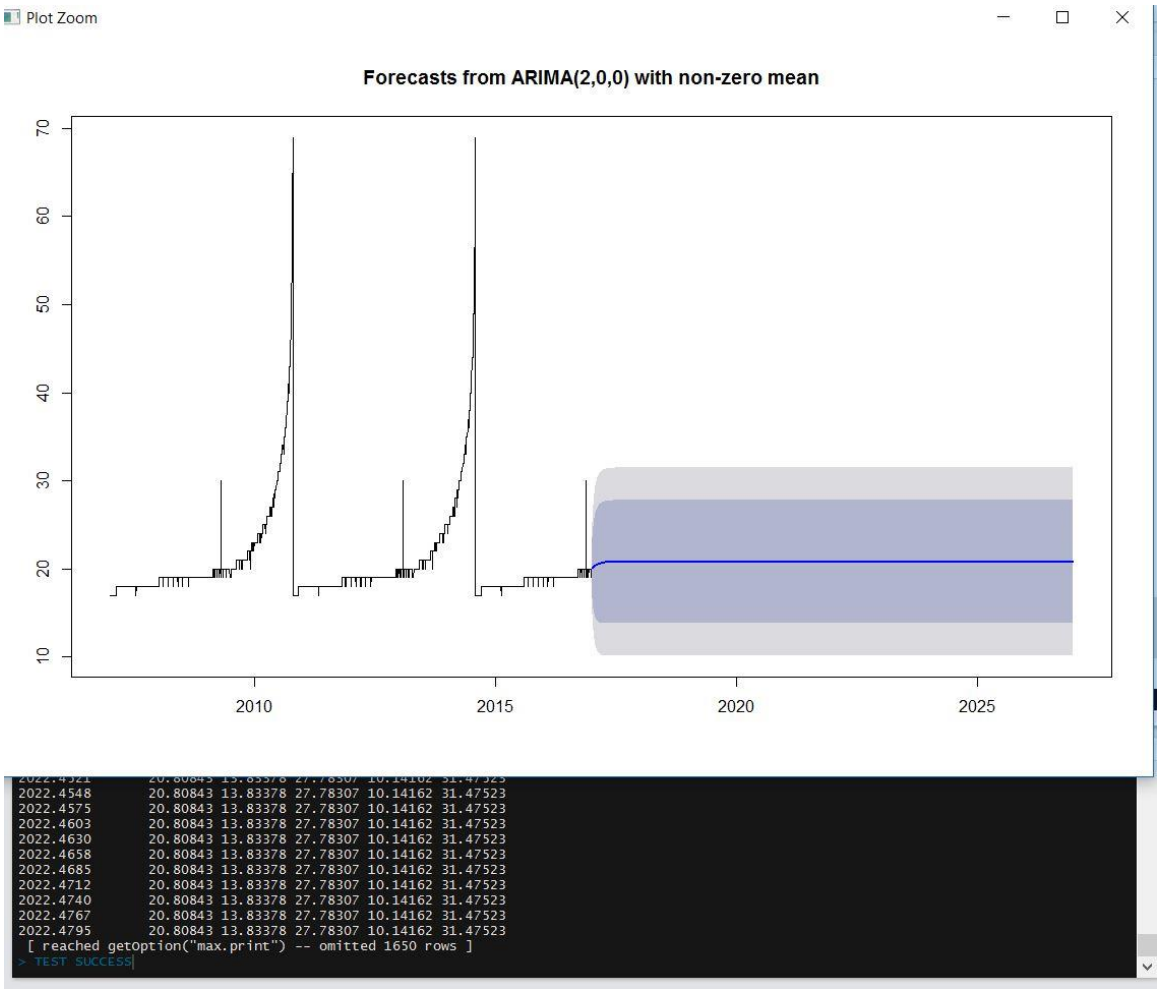


Figure 39 Showing a Test Case 2 Success

7.4.3 Statistical Tests

```
        Welch Two Sample t-test

data:  Data$Grad.Mature and Data$Dropout.Mature
t = 1.0168, df = 522.51, p-value = 0.3097
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.6552125  2.0612034
sample estimates:
mean of x mean of y
 30.94737  30.24437
```

Figure 40 R T-test Mature

```
        Welch Two Sample t-test

data:  Data$Grad.School.Leaver and Data$Dropout.School.Leaver
t = -3.4157, df = 2231.5, p-value = 0.0006477
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.24402529 -0.06601967
sample estimates:
mean of x mean of y
 18.87014  19.02516
```

Figure 41 R T-test School Leaver

Independent Sample

		Levene's Test for Equality of Variances			
		F	Sig.	t	df
Student	Equal variances assumed	.726	.395	1.019	556
	Equal variances not assumed			1.017	522.511

Figure 42 SPSS T-test Mature

Independent Samples Test

		Levene's Test for Equality of Variances				
		F	Sig.	t	df	Sig. (2-tailed)
Student	Equal variances assumed	.145	.704	-3.421	2280	.001
	Equal variances not assumed			-3.416	2231.525	.001

Figure 43 SPSS T-test School Leaver

t-Test: Two-Sample Assuming Unequal Variances		
	Grad School Leaver	Dropout School Leaver
Mean	18.87014061	19.02516309
Variance	1.136268275	1.201791601
Observatic	1209	1073
Hypothesi:	0	
df	2232	
t Stat	-3.415660427	
P(T<=t) on	0.00032383	
t Critical o	1.645536604	
P(T<=t) tw	0.00064766	
t Critical tv	1.961027397	

Figure 44 Excel T-test School Leavers

t-Test: Two-Sample Assuming Unequal Variances		
	Grad Mature	Dropout Mature
Mean	30.94736842	30.24437299
Variance	67.08258451	64.19170211
Observatic	247	311
Hypothesi:	0	
df	523	
t Stat	1.016813872	
P(T<=t) on	0.154856063	
t Critical o	1.647772343	
P(T<=t) tw	0.309712126	
t Critical tv	1.964510213	

Figure 45 Excel T-test Mature

They above six figures show matching t stats in all showing that the test success parameters have been met for Test Case 1.