National College of Ireland
BSc in Computing
2016/2017

Sean Mc Dermott
x13406038
sean.mcdermott@student.ncirl.ie

Future Housing

Technical Report

Table of Contents

Executive Summary

As the economy is starting to gradually improve after the downturn several years ago, the housing market has benefitted and started to rise as a result. As this market has risen, the prices of the houses and apartments have followed suit. Reading an article on https://www.daft.ie/report/ronan-lyons-2016q1-houseprice, Ronan Lyons states:

"Starting in 2013, house price inflation reappeared in the capital, after an absence of six years and within 18 months, prices had risen by over 40% on average – and almost 50% in some districts."

This technical report contains a descriptive analysis and evaluation on the housing prices throughout Dublin. This analysis will be performed to summarize and describe the data and find patterns within the datasets. The second half of this project will introduce machine learning and aim to predict housing prices throughout different areas of Dublin. This analysis will be based on the information contained in the data sets. The prices of the houses/apartments will be the figures that the machine learning algorithm will consume and produce the results from. The results of the machine learning model will be compared to the test data obtained from www.propertypriceregister.ie. The accuracy of the comparisons will determine if the analysis was a success or failure.

Having a model that will accurately predict future housing prices per post code in Dublin will be valuable for home buyers that want an idea of how much they will need to spend if they want to purchase a house or apartment during that time of the year.

Introduction

1.1 Background

Leaving college at the end of 3rd year we were strongly advised to take time over the course of our internship and summer break to think of an idea for our final year project. In my case this was easier said than done because I started my internship a week before the deadline. This meant I was working right through the summer until mid-August.

I decided to pick Data Analytics for my 4th year specialization because I had a keen interest in my Database modules in previous years and wanted to continue to work with data but in a more advanced scale. I researched Data analytics before I chose it to see what type of work was done in it. There's an endless amount of data that you can obtain and numerous analysis' you can do on it.

When we started back in college I still hadn't got a solid idea of what I wanted to do for my project but I did have a vague idea in my head. This idea stemmed from a database project we did in our advanced database module in 3rd year. I came up with the idea in which our team built a database and data warehouse for an online clothing store. This was of course filled with generated mock data which was very useful and accurate. I came up with the idea because an online clothing store has all the essential attributes needed to create a functional database. For my 4th year project I wanted to stay along the lines of clothing and build a project around it. I was toying around with a few ideas but I really didn't know where I was going with them so I decided to contact Michael Bradford who we had in 2nd year for Advanced Databases and see if we could come up with a couple of relevant ideas.

When we met we discussed a couple of ideas that really appealed to me. I left Michaels office with 2 solid ideas and finally picked one to run with. The idea I picked to base my project on is to build a platform that predicts future fashion trends. Because I work part time for a high-end fashion brand I wanted to base my analysis around future trends in clothing products for that brand. I got in contact with the head office to see if they could supply me with a data set. They got back

in touch and forwarded me on to the team that look after that area of the company. After weeks of continuous emails and back and forth I got no reply from them. I sent a couple more emails and still got no response. At this stage I knew I had to change the area in which I wanted to base my project around

After a week of searching around for a new project topic, I came up with the idea to do a predictive analysis on the housing prices in Dublin. I emailed my supervisor with the idea and he thought it was a good choice and pointed out a website that has a large amount of data that is suitable for this type of project. The data sets included the house prices, house addresses and types of houses among other attributes. After spending time playing around with the data I got ideas for different analysis that could be carried out and what information I could retrieve from the data.

Over the course of this project to date, the project proposal required an update due to the unsuccessful retrieval of data for the original idea. The updated project proposal has been implemented and distributed across the relevant sections of this document while the original project proposal is located in the appendix.

1.2 Definitions and Abbreviations

H/A – House/Apartment

R Studio – Open source programming software

R – Programming language

SQL – Communicative database language

Excel – Spreadsheet software

Tableau – Data visualization software

KDD – Data mining process

Algorithm – A process which uses calculations or other problem solving operations to get a result.

Machine Learning – method of analyzing data

1.3 Aims & Objectives

Goal 1: Obtain a data set containing the relevant information for this project.

- Objective 1.1: Browse the internet to find a data set or data sets that has information on housing in Dublin and Ireland.
- Objective 1.2: Download and view the data

Goal 2: Clean the data set

- Objective 2.1: Assess what data will be useful and delete the data that will not be useful.
- Objective 2.2: Structure the data to receive various analysis techniques.

Goal 3: Explore the data set

- Objective 3.1: Study the data set and see what questions can be asked.

Goal 4: Create a database

- Objective 4.1: Create the database and table(s) and save the scripts
- Objective 4.2: Find a suitable database host to connect the database.

Goal 5: Research and learn the fundamentals of machine learning.

- Objective 5.1: Rent books from the library and find resources online to help get an understanding of the different machine learning algorithms.
- Objective 5.2: From this research, decide which algorithms will be used on this project.

Goal 6: Identify what questions this project will answer

- Objective 6.1: what is the objective of this project. Map out questions that will aid a successful analysis on the data.

Goal 7: Create a predictive analysis

- Objective 7.1: Carry out a predictive analysis on the main question this project will try to answer.

Goal 8: Visualize the results

- Objective 8.1: Gather the results of the analysis.
- Objective 8.2: Put the results of the analysis through a visualizing software.
- Objective 8.3: Present the visualized results of the project.

1.4 Research Questions

The main question to be answered in this paper is, do outside factors such as transport locations or proximities to schools etc. have an impact on the price of houses in Dublin. Is there any correlation between the price of a house and these factors that may influence it?

1.5 Technologies

There are various technical approaches implemented when working on this project. Here is a list of several software's, platforms and languages that will be used. Some of these technical approaches will not be implemented until the later stages of the project

R & R Studio
R will be the primary language that will be used throughout this project. R Studio is the open source IDE on which R runs. R will be used to pre-process, transform and mine the data. R Studio contains a number of useful packages for this project such as ggplot. Ggplot is a visualization package which enables the user to plot results via graph.

MySQL
MySQL will be the database management system on which this project will run. The reasons for choosing MySQL are due to its high performance and management ease. MySQL integrates well with R Studio and Excel which will be important when importing tables (from excel to MySQL) and exporting data (from MySQL to R Studio).

Excel
Excel is a spreadsheet software that will be used to store unstructured data and test data offline.

Tableau
Tableau will be used to visually represent the final results. Tableau receives data quite easily and transforms it within the matter of seconds or minutes. Tableau is user friendly and returns professional visualizations.
This software will be implemented towards the end of the project.

1.6 Structure

Knowledge Discovery in Data mining

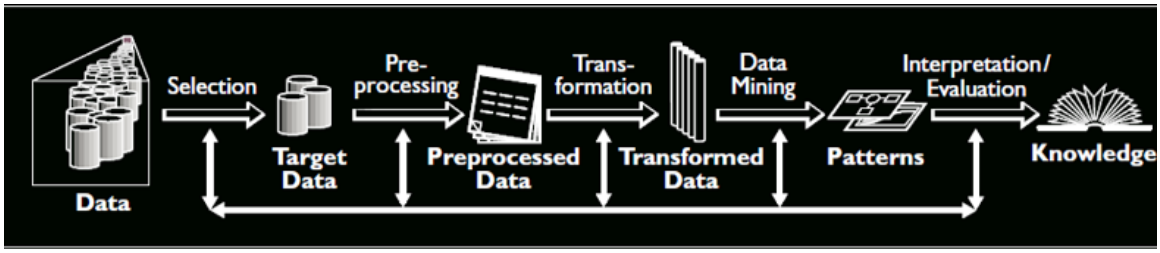


Fig. 1. Steps of the KDD Methodology

This Data Analytics project will be structured from the KDD (Knowledge discovery and data mining) methodology. The KDD methodology will allow the project to be broken up and evaluated by following these 5 steps: Selection, Pre-processing, Transformation, Data Mining and Interpretation/ Evaluation.

Selection:
The data set selected for this project will need to contain relevant and reliable data. That is achieved by knowing what outcome I want from the data. Having the right data is vital because the data mining process will be learning from it to produce the end results.

Pre-processing:
This is the step in which having reliable data is enhanced. Adding data and Removing unreliable data is important if I want accurate results. Strategies will be formed to handle missing data values and outliers. Pre-processing will be continuous for the entirety of the project. New features may also be added throughout the project. R Studio contains very powerful packages that allow for the manipulation of data.

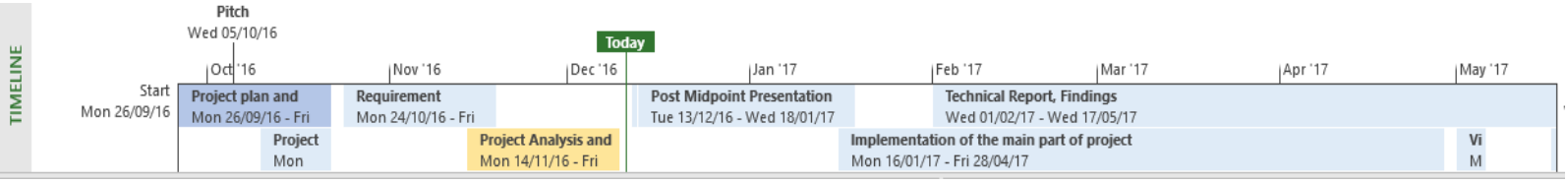
Transformation:
Transforming the data allows the user to tailor so its fit for purpose. This is a crucial part of being able to explore the data in the right way. The data will be organized and necessary attribute types will be converted. Data will be reduced and sampled in this stage to give a small scale understanding of the data set.

Data Mining:
As the goal of my project is to carry out a predictive analysis, the data mining technique that will be used is regression. Other techniques such as clustering and decision trees may also be used if seen fit.

Interpretation:
This will be the final part of the project. The results of the previous steps are tied together and visually presented in a way in which a user will be able to understand. The goal of the project is looked back on and reviewed in comparison with the final result

1.7 Project Plan (Gantt Chart)

The project plan illustrates the timeline the project will adhere to



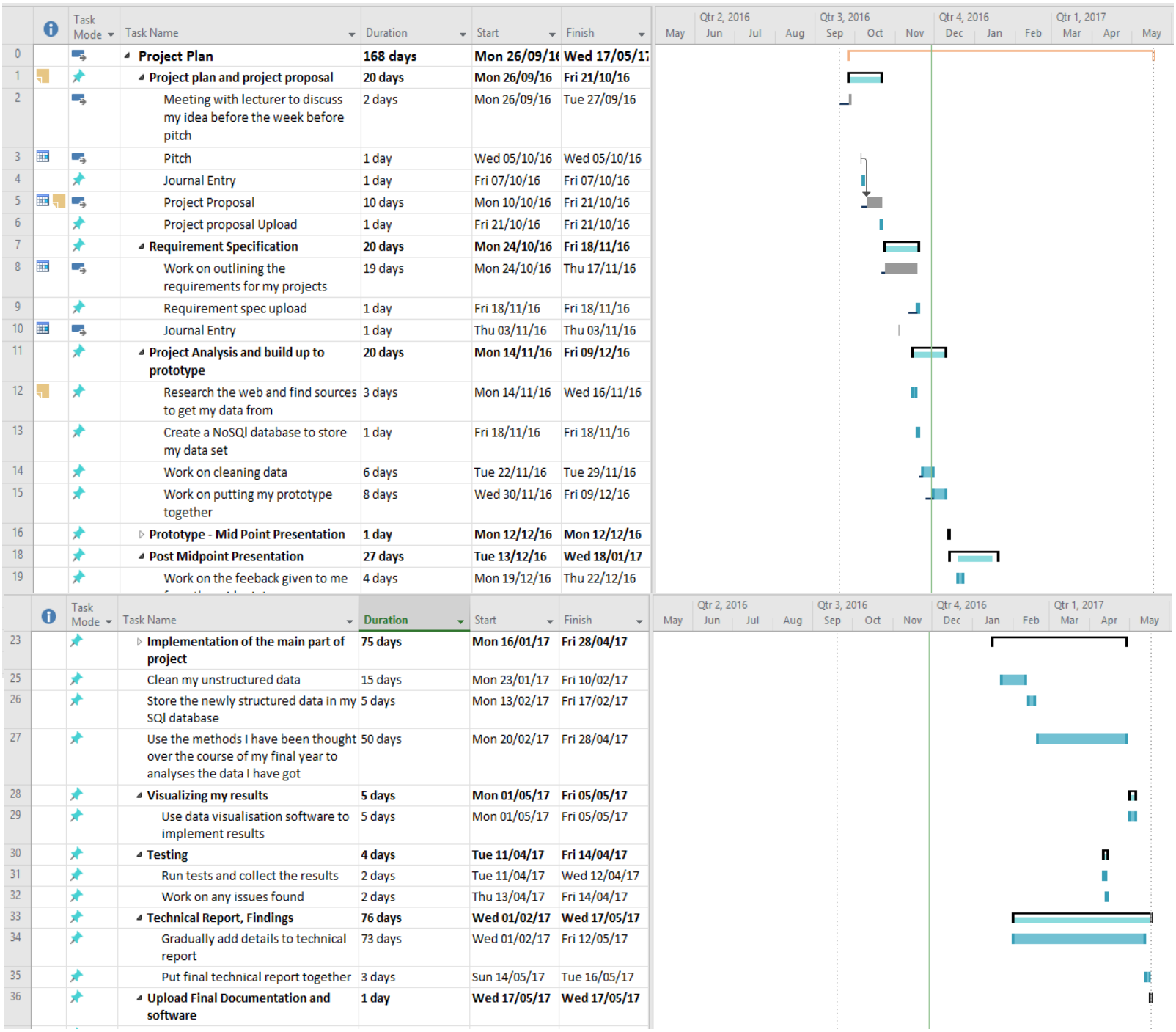


Fig. 2. Gantt chart representing the timeline the project will follow

2 System

2.1 Requirements

2.1.1 Functional Requirements

The functional requirements will include a level of importance. There are two levels of importance that will be assigned to each requirement. Level 1 implies that the requirement is a vital part of the project and level 2 implies that the requirement is very important but not critical and there could be a plan B if that requirement were to fail.

2.1.1.1 Requirement 1 <Obtaining and Cleaning Data>

Description & Priority

Obtaining the correct data is essential to building a successful project. **(Level 1)**. The obtained data must have a necessary amount of information that will allow for the exploration and mining process of the KDD structure. Cleaning the data is vital as there will be many features that are useless as well as defect in the data that will be needed.

Use Case

Scope: The scope of this use case is to show how the admin gets and cleans the data set.

Description: The admin accesses the internet to browse for a relevant dataset. Once this dataset is found its downloaded. The data is open and cleaned in a programming application.

Use Case Diagram

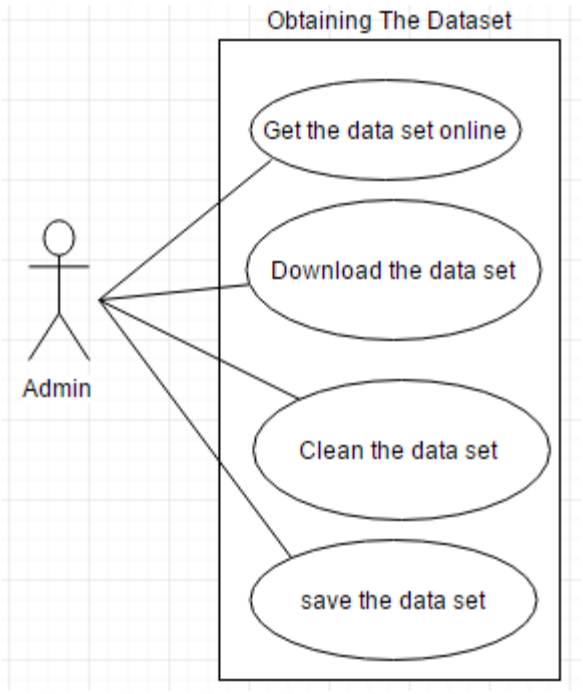


Fig. 3. Obtaining the data set use case

Flow Description

Precondition: The admin must have internet access to browse the internet in search for a dataset.

Activation: This use case starts when the admin gets a dataset

Main flow

1. The admin accesses the internet
2. The admin browses the internet
3. The admin finds the dataset
4. The admin downloads the dataset
5. The admin cleans the dataset
6. The admin saves the dataset

Post condition: he data is downloaded, cleaned and stored.

2.1.1.2 Requirement 2 <Creating The Database>

Description & Priority

This use case is high importance **(level 2)**. The database will be created and managed through an open source relational database management system. It will

be hosted through a free online hosting service. Storing the data in a database enables the admin to access the data quickly and securely from the programming IDE.

Use Case

Scope: The scope of this use case is to show how the database is created and how the data is imported.

Description: This use case describes how the admin creates the database. The database needs to be hosted in order for it to be created. Once there is a host, the database can be created. Finally, the data can be imported from the local files.

Use Case Diagram

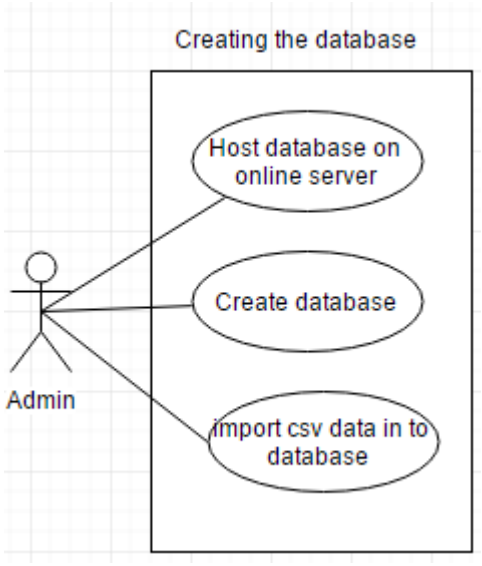


Fig. 4. Creating the database use case

Flow Description

Precondition: The admin must connect to a database hosting platform.

Activation: This use case starts when the admin creates a host.

Main flow

1. The admin accesses hosting platform
2. The admin creates a host for the database
3. The admin opens the database software and creates the database

4. The admin imports the data to the database using an add-on in the spreadsheet software.

Post condition: The data is ready to be exported from the database

2.1.1.3 Requirement 3 <Analysing The Data>

Description & Priority

Analysing the data is the main aspect of a data analytics project **(Level 1)**. It is also the most fun and exciting part. The analysis being carried out for this project will allow the admin to find out valuable information and patterns that are present within the data.

Use Case

Scope: The scope of this use case is to show how the admin perform an analysis on the data they are working with.

Description: This use case describes how the admin can analyse the data using a programming language and produce different results. The scripts containing code used in the analysis is saved in the database

Use Case Diagram

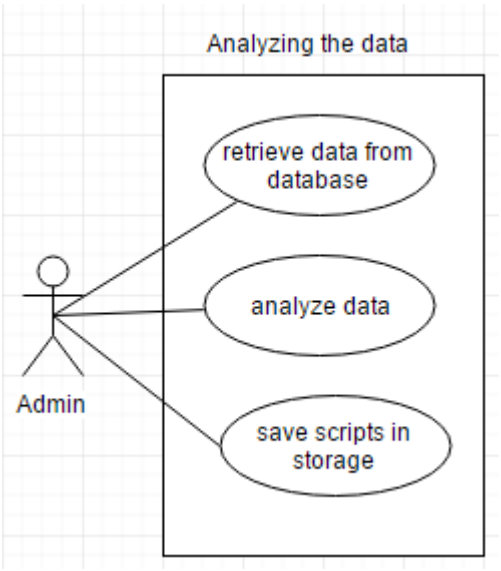


Fig. 5. Analysing the data use case

Flow Description

Precondition: The admin must have a connection to the database

Activation: This use case starts when the admin imports the data from the database into the programming application.

Main flow

1. The admin connects to the database via programming platform
2. The admin retrieves the data from the database
3. The admin performs analysis
4. The admin saves scripts from programming platform in storage

Post condition: The admin has a broader understanding of the data.

2.1.1.4 Requirement 4 <Mining the Data>

Description & Priority

Data mining is the computing process is this project is based around. **(level 1).** The main data mining technique which will be implemented is regression. Various models will be tested to find the most accurate model.

Use Case

Scope: The scope of this use case is to show the steps involved in implementing regression analysis to the data.

Description: This use case describes how the admin inputs the data into the programming software and applies various regression models to the data to retrieve the most accurate result.

Use Case

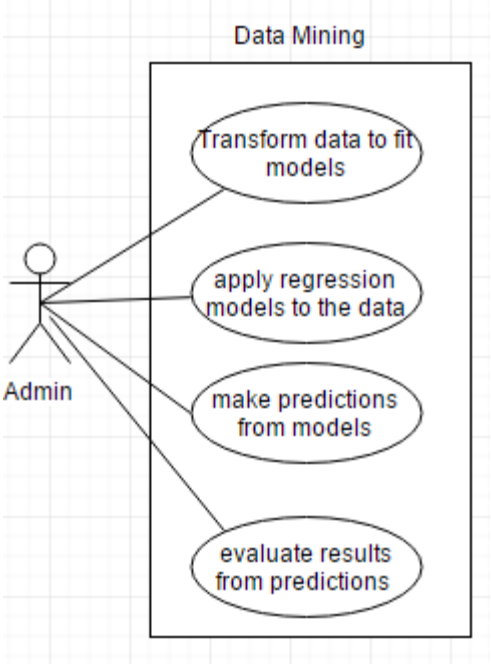


Fig. 6. Data mining use case

Flow Description

Precondition: The data must be transformed so it fits the regression models.

Activation: This use case starts when the data begins to be transformed.

Main flow

1. The admin transforms the data to the appropriate structure.
2. Various models are fitted to the data
3. Predictions are made based of the models.
4. The results are then evaluated.

Post condition: Results from the prediction of the model are outputted and ready to be analyzed.

2.1.1.5 Requirement 5 <Visualizing the Data>

Description & Priority

Visualizing the data allows the admin to clearly communicate the results of the project via statistics, plots, graphs, charts etc. **(level 2).**

Use Case

Scope: The scope of this use case is to illustrate how the results of the analysis will be viewed.

Description: This use case describes how the data uses an online visualization tool called tableau to create visual representations of the analysis conducted in the last use case

Use Case Diagram

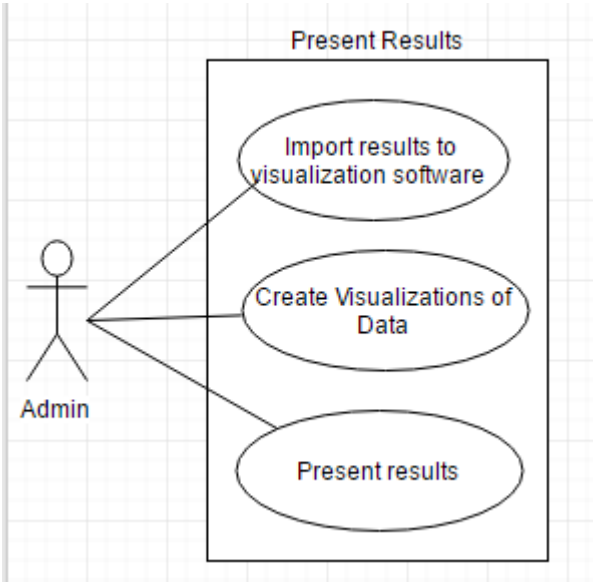


Fig. 7. Present results use case

Flow Description

Precondition: The admin must have their data ready to import into tableau

Activation: This use case starts when the data is imported to tableau

Main flow

1. The admin imports data into visualization software
2. The admin creates visual representations of data
3. The admin presents the results for viewing.

Post Condition: The visualized data is available view.

2.1.2 Non Functional Requirements

2.1.2.1 Environmental requirements

It would be difficult for environmental requirements to affect this specific project as the source of the data is run through the government and must be live for the public to access and view.

2.1.2.2 Recovery requirements

If a hardware failure did occur, there are up to date backups constantly being saved to cloud software.

2.1.2.3 Portability requirements

The database is created through a web based community package in MySQL which enables it to be run from anywhere. The downside is there must be an internet connection to connect to the database.

2.1.2.4 Security Requirements

There was no security measure needed as the project was not accessible to the public. Although, a password was required to access the database.

2.1.3 Data requirements

2.1.3.1 Data Source

Through looking at the scope and purpose of this project, obtaining data sets with specific information in it was key to fulfill the set objectives. Through searching for data that was specific to the purpose of this project, It was advised to look on https://www.propertypriceregister.ie/website/npsra/pprweb.nsf/PPRDownloads?OpenForm. After browsing the datasets on this website, it seemed there was a sufficient amount of relevant data contained in the data sets. This data source has a download criteria that allows the user to generate data for every county in Ireland for each month of the year from 2010 to 2017. The data sets are frequently added and updated by the Property Services Regulatory Authority (PSRA) pursuant to section 86 of the Property Services (Regulation) Act 2011.

By progressing through the second semester in college and gaining more knowledge about data mining it was apparent that the data obtained from propertypriceregister.ie had significantly insufficient attributes within the data and any type of data mining would be meaningless. This is when it was decided to find outside resources that would influence the price of a home and would be able to be fitted to the data. One of the outside factors that was advised was the incorporation of public transport. The details of how this was incorporated will be explained further on. This information for these data sets was obtained from https://www.transportforireland.ie/transitData/PT_Data.html. The data contains the names and geolocations for every LUAS stop, train station and bus stop throughout Dublin.

A second outside factor that is incorporated into the project is education. The data from http://www.education.ie/en/find-a-school contains names and geolocations for every primary and secondary school in Ireland. A description of how this was incorporated will be explained further on.

2.1.3.2 Dataset Description

The main data being utilized for the project comes from the property price register. The data sourced from this website contains the price of every house or apartment that was sold in Ireland during that specific month. For the purpose of this project it was best to use a subset of this data. The data for January in 2015 was selected to be the main data set. The reason for this was because as the functions were implemented, the more data the functions had to go process, the longer it took. Some functions were taking up 60 minutes on the full data set whereas it only took up to 5 seconds on the subset data.

Newdata Data

The following table represents the features that were used from this data set:

Table. 1. Newdata data set description

Address	The full address of the house.
Postal.Code	The area of Dublin the house is located
Price	The total amount paid for the house

The following table represents the features that were added to the data set:

Table. 2. Description of features that were added to newdata data set.

Lat	The latitude of the house
Lon	The longitude of the house
Dist.to.nearest.luas.stop	The distance in km to the nearest luas stop from each house
Dist.to.nearest.train.station	The distance in km to the nearest train station from each house
Dist.to.city.center	The distance in km to the city center for each house
No.of.sec.schools.within.2km	The number of secondary schools within 2km from each house
No.of.prim.schools.within.2km	The number of primary schools within 2km from each house

The data in the table 2 was introduced to the main data set as it had ineffective features that could fit a regression model. These new parameters will be the basis of what the regression models will be created. The process taken to implement some of the features will be explained in the implementation section of this document.

Luas Data

Table. 3. Luas data dataset description

Name	The name of the luas stops
Lat	The latitude of the luas stops
Lon	The longitude of the luas stops

These are the only necessary features in the Luas data set. The name of the Luas stop shows up on the interactive map when a user clicks on the point. The latitude and longitude are essential in being able to find the nearest Luas stops for each house.

Train Data

Table. 4. Train data dataset description

Name	The name of the train stations
Lat	The latitude of the train stations
Lon	The longitude of the train stations

This is effectively the same description as the Luas data above. The features all have same purpose.

Primary Schools Data

Table. 5. Primary schools data dataset description

Lat	The latitude of the primary schools
Lon	The longitude of the primary schools

The latitude and longitude are the only important features of the primary schools' data. These two features are used in conjunction with the house latitude and longitude to find out how many primary schools are within 2km.

Secondary Schools Data

Table. 6. Secondary schools data dataset description

Lat	The latitude of the secondary schools
Lon	The longitude of the secondary schools

The description for the secondary schools' data is the same as the primary schools data as the features serve the same purpose.

2.2 Design and Architecture

2.2.1 System Architecture

The system architecture diagram illustrates a basic overview of the structure and behaviour of this project. The different aspects of the project are presented in the system architecture diagram such as the file format of the source data, the database, the data mining engine and the visualisation tool. These different processes will be broken down further as the project progresses.

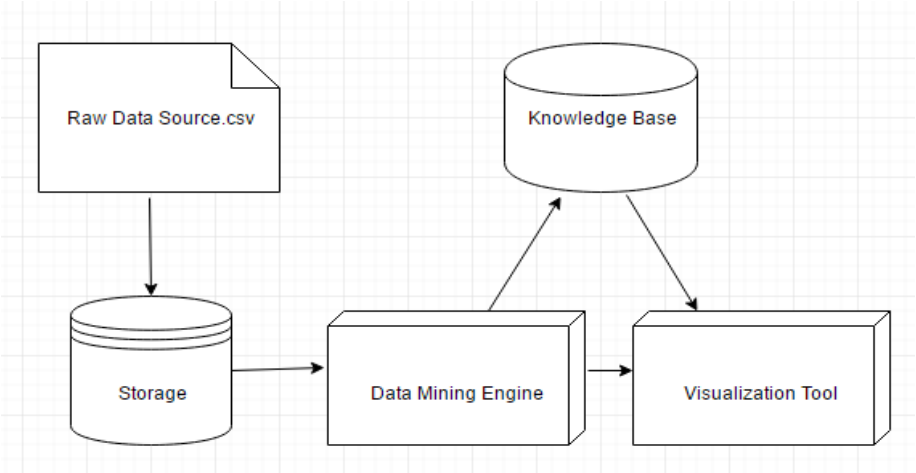


Fig. 8. Data architecture of the entire system

Fig 8 broadly interoperates the high-level architecture of the project. The storage system that is applied to the project is a MySQL relational database that is hosted on an online hosting platform. The reason to include a storage system is because of security and accessibility. Although security is not a major concern for this project, it's imperative that the data is safe in case of any hardware failures or malfunctions. Accessibility is the main reason for the inclusion of a database. The speed in which a data file can be accessed is key along with the ability to write data from R Studio to the tables in the database. The breakdown of the data mining aspect is explained below in figure 9. The visualization tool is important for the final stage of the project. This will tie the project together by visually interoperating the results found throughout the project.

2.2.2 Data Mining Architecture

The data mining technique being utilized in this project is regression. Regression measures the relationship between one variable (dependent variable) and a series of other variables (independent variables). The two main types of regression are linear regression and multiple liner regression. Linear regression measures the relationship between the dependent variable and only one independent variable to try to predict the outcome of the dependent variable. Multiple linear regression uses two or more independent variables to predict the outcome of the dependent variable.

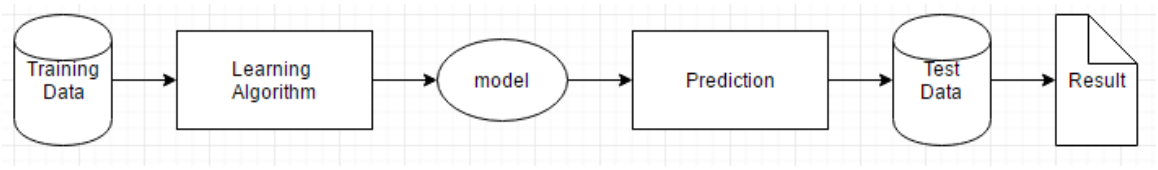


Fig. 9. Data mining architecture

Figure 9 illustrates an overview of how the regression model is applied. The process is split up into two steps; step 1 is the training phase where the learning algorithm is applied to the data which generates a regression model. Step 2 is the testing process which compares the results from the learned model to the test data and obtains the accuracy of the learning algorithm. Once the result is evaluated, the model can be alternated to attempt to achieve a more accurate model.

2.3 Implementation

2.3.1 Database connection and extraction of tables

Connecting to the database is the first step that needs to be executed before the user wants to analyse data. Having a secure connection is vital because the data is stored within the database. The following code snippet shows the connection to from R Studio to the database.

```
#Connects to MySQL database
con <- dbConnect(MySQL(),
                 user="root", password="Password",
                 dbname="housing", host="127.0.0.1")

#View the tables in the database
dbListTables(con)
```

Fig. 10. connection to a MySQL database from R Studio

The first three lines of the code snippet below are creating a connection to the database by consuming the database credentials. These are generated from the database host. Having the database connected to R Studio allows the admin to export tables. The next line of code allows the admin to view what tables are contained in the database.

```
#Selecting all the data from housing table
rs = dbSendQuery(con, "select * from new_jan_2015")
newdata = fetch(rs, n=-1)
```

Fig. 11. pulling data from a table in the database and inputting it into a data frame called newdata.

Figure 8 represents the process of pulling a table from the database. The first line uses the **dbsendQuery** function in the **RMySQL package** in R Studio. **con** is then called as it's the connection to the database and from there the SQL query selects all data from the new_jan_2015 table and is stored in **rs**. In the second line **rs** is then instantiated into a data frame called **newdata**.

2.3.2 Number of houses sold per postcode

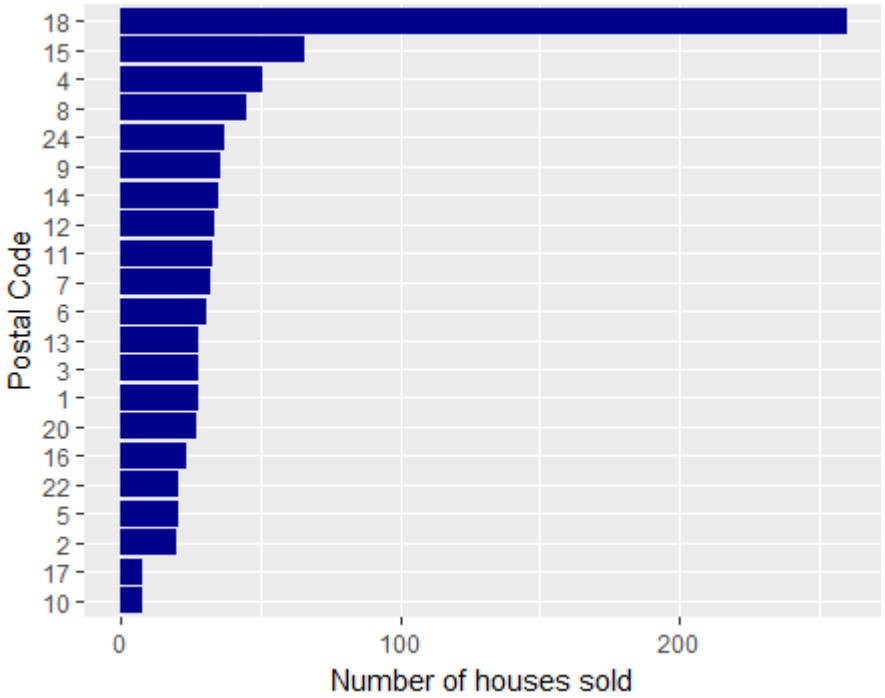


Fig. 12. Number of houses sold per post code.

Figure 12 is a sideways histogram which clearly shows the most number of houses that were sold, was in Dublin 18. Valuable information such as this gives the analyst a platform to diverge into deeper analysis on the post codes that had the most sales. The plot is created from the **ggplot** package in R Studio. Simple plots

such as this unlock patterns and give valuable information about the underlying data.

2.3.3 House prices per post code

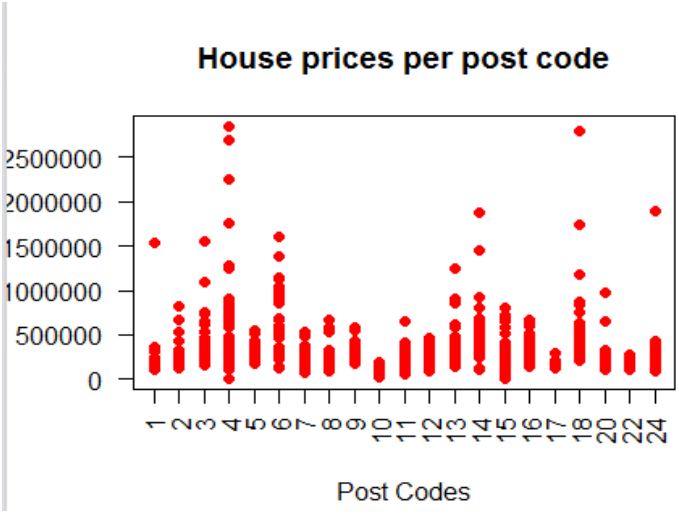


Fig. 13. The price of houses in relation to post codes

The **stripchart** in figure 13 Gives a visual representation of the price ranges for each post code. The most expensive houses were sold in Dublin 4, 14, 18 and 24 although there are only between 1 to 3 houses sold at the higher end of the spectrum. This is important information for the analyst as it presents any potential outliers for further analysis or data mining that may be carried out. The graph also gives a strong indication that the data may vary due to the fluctuation of prices between the post codes. Visual Information such as this is a lot easier to consume rather than looking at figures.

2.3.4 Interactive map

An interactive map was implemented using the **leaflet** package in R Studio. The purpose of adding this map was to give the analyst a visual representation of where various data points such as, the houses or Luas stops where located.

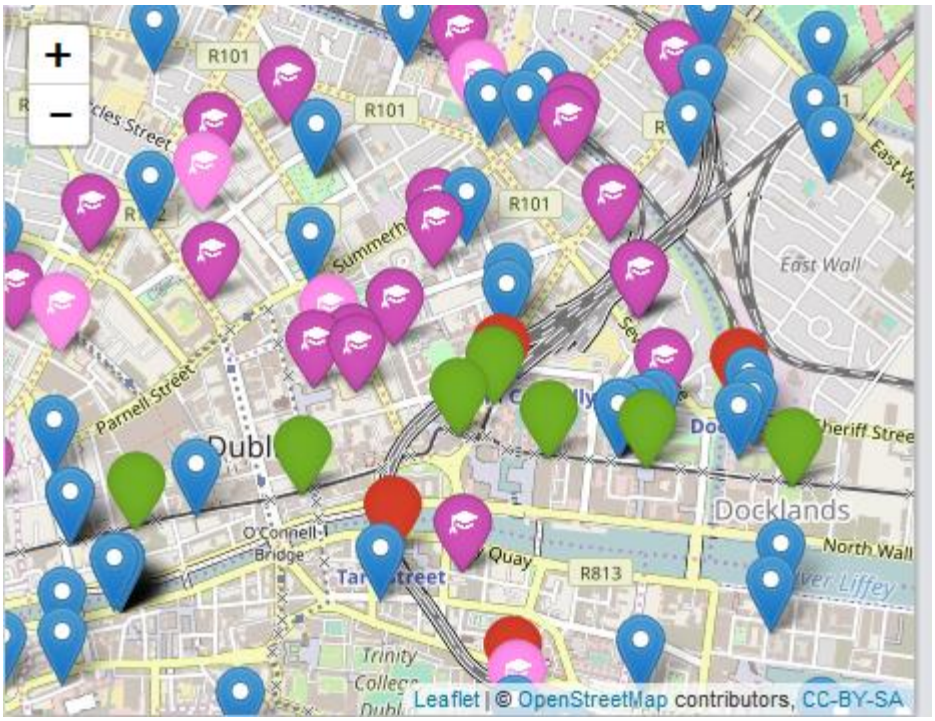


Fig. 14. Interactive map that contains locations of houses

The main idea of implementing the map was to get an idea of where everything was located. The map has five different color markers that represent the five main features that are utilized in the project. The blue marker represents each house in the dataset. When the blue markers are clicked, the price and address of the house popup. The green marker represents the Luas stops. When the Luas markers are clicked the name of the station pops up. The red markers are the train station markers. When they are clicked the name of the station also pops up. The pink and purple markers represent the education features that were added to the data. The purple is the primary schools and when that's clicked the name of the school pops up. Finally, the pink marker is the secondary schools' marker and when that's clicked the name of the secondary schools pops up. This graph is useful not only as a visual to where everything is located, It's can also be used as a rough gauge to test the nearest distance function from a house to a Luas stop. By looking at the map and finding certain points its quite effective to make a judgement if the function has produced the correct distance.

2.3.5 Finding the closest Luas/Train station to each house

As mentioned earlier in the document, due to the lack of effectiveness of the original features from the main data set, other parameters for the regression model needed to be established. The snippet below shows the output from a function that finds the closest distance from *point **A*** in *data set **A*** to *point **B*** in *data set **B.** The code for this function was sourced from:*

http://stackoverflow.com/questions/22121742/calculate-the-distance-between-two-points-of-two-datasets-nearest-neighbor

dist.to.nearest.luas.stop	dist.to.nearest.train.station	dist.to.city.center
2.25	1.95	19.91
2.19	0.91	19.79
2.03	0.89	19.48

Fig. 15. Output of function that finds the distance of the closest Luas and train station's

The three outputs seen in figure 15 have been run from the same function. These are the first of five parameters that were computed in order to create a successful model further on in the project. The function worked by starting off with the first house point in the **newdata** data set and iterating through the Luas data set and finding the shortest distance and so on. (Part of code submission) Having these three data points are a realistic outside factor that can influence a house price and will be important parameters for the regression models. The same function that worked for the public transport distances worked for the city centre distance because it was only finding the distance to one geolocation and whether that was the shortest or longest distance, it was always going to output the same result.

2.3.6 Number of primary and secondary schools within certain radius

Distances to public transport stations and to the city centre are important outside factors when selling a house. An even bigger influence is having schools within a reasonable distance from a family's house.

no.of.primary.schools.within.2km	no.of.secondary.schools.within.2km
5	1
7	4
6	3

Fig. 16. Output of function to retrieve number of primary & secondary schools within 2km of each house

Figure 16 shows the output from a function that retrieved, firstly, the number of primary schools within 2km from each house, then done the same for the secondary schools. These are two very strong features that can influence the price of a house.

2.3.7 Building Regression Models

The final implementation of the project is the introduction of the regression models. There will be numerous models created to find the model with the most accuracy. The reason for multiple models is because there are seven parameters that can be fed into the model and some of these parameters will add significant value and some may add no value to the model. The seven main variables to be tested in the models are **Price** (dependent), **dist.to.city.centre, dist.to.nearest.luas.stop, dist.to.nearest.train.station,Postal.Code,no.of.primary.schools.within.2km** and **no.of.sec.schools.within.2km** (all independent)**.**

Before building any model the first step is to split the data into training and test data sets. The training data is what the model is learned from and the testing data is what the prediction is made on. The split for this project was 70% training and 30% testing.

The code for building the regression models was sourced from:

https://rpubs.com/chocka314/251613

2.3.7.1 Model 1: General Linear Model

The first regression model that has been built is a generalized linear model called **lm**. This model is trained by having Price as the dependent variable and the remaining variables in the training data as the independent variables. The first step in this model is to check the coefficients of each variable for skewness. The

results of the skewness check saw that the dependent variable Price, skewed way right. This meant a log transformation had to be added to the price in the next model to restore its symmetry. Moving on to the results of the prediction on the model, The RMSE (Root Mean Squared Error) was 5.56 and the R2 (R Squared) value was 0.10. The RMSE represents how concentrated the data is around the line of best fit. The lower the RMSE value, the better the fit of the model. R2 represents the improvement of prediction of the regression model with 0 indicating the model does not improve prediction and 1 indicating perfect prediction.

2.3.7.2 Model 2: General Linear Model with Log added to Price

The second model was created identical to the first model but a log was added to the dependent variable Price. The purpose of adding a log is to restore symmetry and normally distribute the variable. When the log was added to the dependent variable in the second model the RMSE was the same at 5.56 but the R2 value dropped to 0.098. Although it was a very slight drop in the R2 value, it still decreased the prediction power of the model therefore, it was decided that the log would be removed for further models.

2.3.7.3 Checking the p-value for each feature

In pursuance of the most accurate model a hypothesis test is carried out to find out which of the independent variables variation cannot explain the variation in the dependent variable (**Price**)**.** The null hypothesis states that the dependent variable cannot be explained by anything other than randomization. This is tested before model 3 is created because if there are features that have a p-value not significant ($p < 0.05$), they can be removed from the model as they may not be contributing to the model.

Table. 7. Checking the p-values

```
Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                         473034      38332  12.341  < 2e-16
dist.to.nearest.luas.stop            13133      13224   0.993    0.321
dist.to.nearest.train.station         4776      11956   0.399    0.690
dist.to.city.center                 -24704      20985  -1.177    0.240
no.of.primary.schools.within.2km    -78988      13847  -5.704 1.84e-08
no.of.secondary.schools.within.2km   22174      13674   1.622    0.105
Postal.Code                         -11300       2846  -3.971 8.04e-05
```

It can be seen that Postal.Code and no.of.primary.schools.within.2km both have p-values that are less than 0.05, therefore are not significant and can be removed from model 3. Removing the features does not insure a more accurate model but may insure a model fits the data better.

2.3.7.4 Model 3: Linear Model with Features Removed

From reviewing the results of the hypothesis test, it was decided to remove Postal.Code and no.of.primary.schools.within.2km from the third model. When the two features were removed from the mode it yielded an RMSE value of 3.62 which was reduced from 5.56 from the two previous models indicating the model was a better fit. Although the model was a better fit to the data, it returned an R2 value of 0.02 which is a significant decrease in prediction power compared to the previous two models. This suggests that just because the model is a better fit to the data doesn't mean the prediction will be more accurate. The loss of two independent variables has had a major impact on the results of the third model.

2.3.7.5 Model 4: Random Forest Model

The final model created was a Random Forest regression model. A Random Forest model was included because its generally a more accurate model than a linear model. After unconvincing results in the third model it was decided to add the two independent variables back in to the Random Forest model. The results of this model were, as expected, greater than the linear models. The model produced an RMSE value of 3.23 whereas the lowest RMSE value in the liner models was 3.62. It also produced a higher R2 value of 0.19, which is higher than any of the linear

models. Figure 17 plots the actual vs predicted values using the Random Forest model

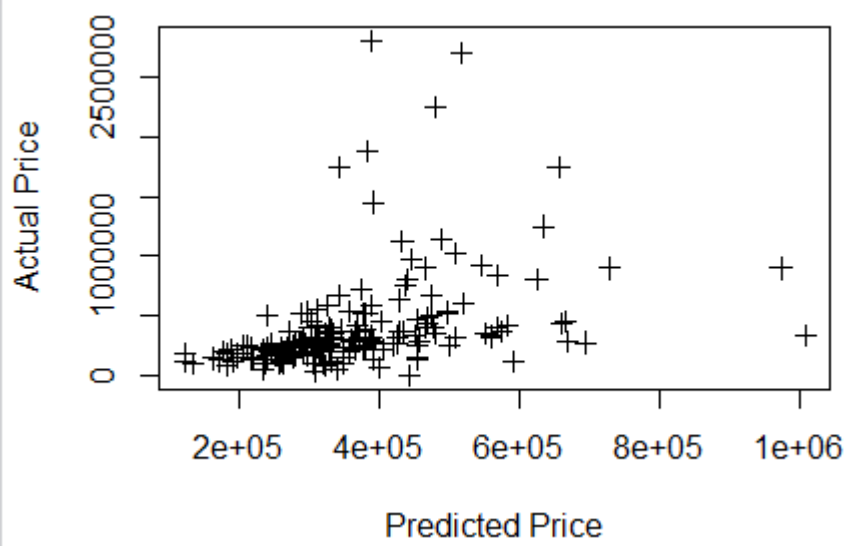


Fig. 17. Plot representing the actual vs predicted values using the Random Forest model

As seen in the plot, the points are plotted on a slightly positive incline but it suggests that the actual values and predicted values are quite different from each, hence the low R2 value. In figure 18 below, there are seven examples of actual values and predicted values.

	Price	predicted.values
56	295000.0	482746.8
57	280000.0	374318.3
58	140000.0	527674.7
59	550000.0	335810.1
60	450500.0	478828.7
61	360000.0	395671.4
62	234000.0	381579.1

Fig. 18. Actual vs Predicted values from the random forest model.

The results from figure 18 clearly show an inconsistency in prediction. Some of the predictions are quite close to the actual prices but for the most part they are very different. A more detailed figure that includes the independent variables that contributed to the predicted values can be found in Appendix A.

2.3.7.6 Regression Conclusion

Regression was the most important aspect of this entire project. The main scope of the project revolved around generating a predicted value based of a set of variables. For this section, four regression models were tested in order to find the most accurate. The first linear model used price as the dependent variable and every other variable from the training data as the independent variables. This model had a high RMSE when compared to the other liner models but also had the highest R2 value which is the more important of the two. This meant when compared to the second and third linear models, it had the highest prediction accuracy.

Moving on to the Random Forest model, this surpassed the linear models by having a lower RMSE value and a higher R2 value that the three linear models. As this was clearly the best model it was decided to analyse a little deeper and produce the actual vs predicted values via plot and insert them into the data frame.

2.4 Testing

2.4.1 Black Box Testing

Table. 8. Database connection test

Black Box Test #1			
Name	Database Connection	**Date of Test**	01/02/2017`
Test ID	BBT#1	**Iteration ID**	2.0

Purpose Of Test	**To Guarantee that:** • **A secure connection to the database from R Studio is constantly available**

Test Steps	**Within RStudio the tester should:** • **Enter valid database connection credentials and run the connection** • **Run a query to ensure the database is connected**
Expected Result	**The database will be connected as the query will output a valid result**
Actual Result	**The query outputted the list of tables in the database**
Suggested Action	**N/A**
Resolution	**N/A**

Table. 9. Nearest Luas stop function test

Black Box Test #2			
Name	Nearest Luas function	**Date of Test**	20/04/2017`
Test ID	BBT#2	**Iteration ID**	2.0

Purpose Of Test	**To Guarantee that:** • **The nearest luas stop from the house is generated**
Test Steps	**Within RStudio the tester should:** • **Run the entire NearestL function**
Expected Result	**The nearest Luas stop should output in kilometers**
Actual Result	**The nearest Luas stop outputted in kilometers.**

Suggested Action	**N/A**
Resolution	**N/A**

Table. 10. Number of schools within radius function test

Black Box Test #2			
Name	No. of schools within radius	**Date of Test**	01/05/2017`
Test ID	BBT#3	**Iteration ID**	2.0

Purpose Of Test	**To Guarantee that:** • **The correct amount of schools are calculated within a 2 km radium**
Test Steps	**Within RStudio the tester should:** • **Run the entire function**
Expected Result	**The correct number of schools within 2km should be added to the data frame.**
Actual Result	**The correct number of schools within 2 miles was added to the data frame**
Suggested Action	**Add a calculation to the formula that converts miles to kilometers.**
Resolution	**The calculation successfully converted miles to radius**

2.4.2 Testing Graphs

Testing is a very important step when building a successful project. Unit testing ensures individual functions perform accordingly. The following test was carried out using the *testthat* package in R Studio. A strip chart was tested to find out if the values within the chart are correctly distributed. Figure 19 displays the original untested chart. See screenshots of the code used to generate this test below in figure 21.

The code for the test function was sourced from:

http://stackoverflow.com/questions/30246789/how-to-test-graphical-output-of-functions

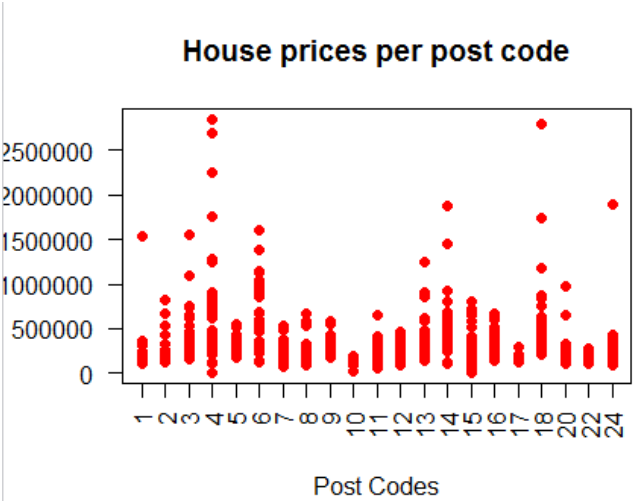


Fig. 19. Untested strip chart

Figure 20 displays the results from the tested strip chart.

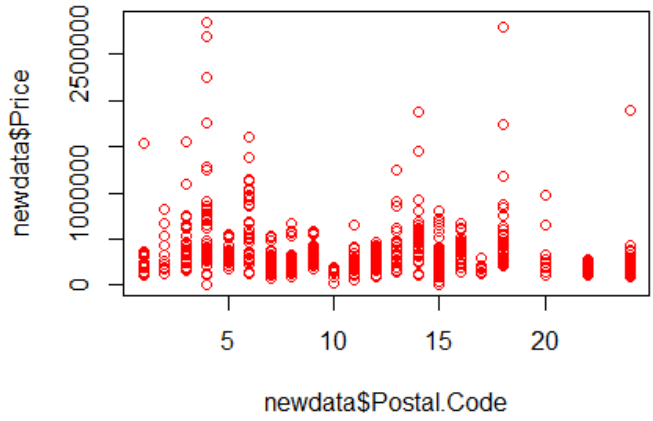


Fig. 20. Tested strip chart.

When comparing the two charts its clear to see that the expected output is identical to the actual output indicating an accurate measure. The tested strip chart revealed no errors when the code in fig 21 was run in R Studio.

```
img <- function(){
  plot(newdata$Price ~ newdata$Postal.Code)
}
## example for a succesful test
expect_identical(plot(newdata$Price ~ newdata$Postal.Code, col ="red"), img())
```

Fig. 21. Code snippet of test

2.4.3 Hypothesis Testing

Hypothesis testing is useful when trying to make an inference if a certain condition is true for an entire population. For this type of project, hypothesis testing wasn't really a prominent feature although one hypothesis test was carried out when implementing a regression model. For this test, the null hypothesis stated that the dependent variable cannot be explained by anything other than randomization. The idea of the test was to find any variables with a p-value less than 0.05 and remove them from the model. Running this test resulted in the output of table 7 In section 2.3.7.3. The p-values of two independent variables were below 0.05 and got removed from the following model.

2.5 Evaluation

The results of the main analysis throughout the project are documented and evaluated in section 2.3. This section will detail the importance of using the KDD structure as a guide through the entirety of the project. As the KDD methodology has such definitive steps, it allowed a clear and decisive structure to be followed. Following the structure included obtaining sufficient data that would suit the needs of the analyst. The next steps included pre-processing and transforming the data to tailor it to suit the analysts' requirements. The final two steps of the methodology required the analyst to mine for patterns through regression and clearly interpret these patterns concisely. The reason for following the KDD methodology before following another data mining methodology is because this structure allows an

analyst to start a project without any knowledge of a topic and through thoroughly following each step, the analyst can come out the other end with an abundance of information surrounding that topic.

3 Conclusion

Overall, the project produced some interesting hurdles and results. The first hurdle encountered was how little information was contained within the main data set. Most of the information in important features such as the size of the property and number of bedrooms were missing. The agreed solution was to think outside of the box and search for outside factors that may influence a house price as the primary factors were missing from the data set. Sourcing the outside factors and implementing functions that generated new features for the data was a real sense of satisfaction as the task was extremely difficult. By introducing these new features, they were used as parameters for the regression models that would predict the house prices. There is plenty of outside factors that can give this project the opportunity to add more in and create a variation of regression models but time limits have restricted the amount in this project. To conclude, the project has been quite successful. Although the regression models have not been as accurate as first hoped. The extra features added to the project and the numerous regression model tested make this project unique and give it the platform to expand and improve.

4 Further development or research

Further development on this project is vital if the aim is to improve the accuracy of the regression models. The reality of the project is that a high accuracy rate will never be achieved because the internal factors such as number of bedrooms or square foot of the property is not available within the data. The introduction of many more outside factors such as crime rates, leisure/recreation locations, park locations could all be brought in and added to a model.

5 References

Cise.ufl.edu. (2016). *Knowledge Discovery in Databases*. [online] Available at: https://www.cise.ufl.edu/~ddd/cap6635/Fall-97/Short-papers/KDD3.htm [Accessed 29 Nov. 2016].

Anon, (2016). [online] Available at: https://www.propertypriceregister.ie/website/npsra/pprweb.nsf/PPRDownloads?OpenForm [Accessed 29 Nov. 2016].

ADVENTURES IN CRE. (2016). *Auto-Populate Latitude and Longitude in Excel*. [online] Available at: https://www.adventuresincre.com/auto-populate-latitude-longitude-excel/ [Accessed 2 Dec. 2016].

Khan Academy. (2016). *Algorithms | Computer science |Khan Academy*. [online] Available at: https://www.khanacademy.org/computing/computer-science/algorithms [Accessed 08 Dec. 2016].

Tableau Software. (2016). *About Us*. [online] Available at: http://www.tableau.com/about [Accessed 01 Dec. 2016].

WhatIs.com. (2016). *What is macro? - Definition from WhatIs.com*. [online] Available at: http://whatis.techtarget.com/definition/macro [Accessed 05 Dec. 2016].

"Find A School - Department Of Education And Skills". *Education.ie*. N.p., 2017. Web. 21 Apr. 2017.

functions?, How. "How To Test Graphical Output Of Functions?". *Stackoverflow.com*. N.p., 2017. Web. 8 May 2017.

neighbor), Calculate. "Calculate The Distance Between Two Points Of Two Datasets (Nearest Neighbor)". *Stackoverflow.com*. N.p., 2017. Web. 4 Apr. 2017.

"Public Transport Data Download Page". *Transportforireland.ie*. N.p., 2017. Web. 24 Mar. 2017.

"Rpubs - Boston Housing Price Prediction". *Rpubs.com*. N.p., 2017. Web. 1 May 2017.

Svoboda, Ali. "Applied Regression Analysis: Project 2". *Rstudio-pubs-static.s3.amazonaws.com*. N.p., 2017. Web. 1 May 2017.

6 Appendix

6.1 Appendix A

Table. 11. Parameters used in the regression models plus the predicted values they generated

Postal.Code	Price	dist.to.nearest.luas.stop	dist.to.nearest.train.station	dist.to.city.center	no.of.primary.schools.within.2km
14	595000.0	4.48	5.37	6.73	11
14	454125.0	1.48	3.32	6.55	6
14	261578.0	4.08	5.38	6.44	10
14	617500.0	1.14	3.70	6.27	6
6	505000.0	4.28	5.28	7.26	11
14	1875000.0	3.62	5.98	8.95	11
14	620000.0	2.09	5.37	5.69	10

no.of.secondary.schools.within.2km	predicted.values
6	532434.7
5	393357.5
7	423848.1
6	367566.8
5	456615.8
5	527829.7
4	458244.4

6.2 Project Proposal (Original)

6.2.1 Goals and Objectives

Goal 1: To research and learn the fundamentals of predictive analysis.

- Objective 1.1: To get books from the library and resources online and study predictive analysis.
- Objective 1.2: To work out how to apply predictive analysis to my project.

Goal 2: Find out what websites and social medial pages are best to pull relevant data from to create my data sets.

- Objective 2.1: Browse fashion/clothing forums, websites, social media pages, e-zines and find the necessary information and data that I will need.

Goal 3: Find out how to extract the data from these pages.

- Objective 3.1: Find out which programming techniques are effective to achieve this.
- Objective 3.2: Find out how to implement these programming techniques to grab the data from the websites.

Goal 4: Store the unstructured data pulled from the internet into a database.

- Objective 4.1: Evaluate which type of database I will be using to store the data.

Goal 5: Structure the data and find out how to analyze it.

- Objective 5.1: Structure the data either manually or with a tool.
- Objective 5.2: Use programming techniques to analyse the structured data.

Goal 6: Use my analyzed data set to predict future trends.

- Objective 6.1: Thoroughly go through my analysis and use the different analysis models to predict future trends.

Goal 7: Display my analysis

- Objective 7.1: Use a tool such as Tableau to display my analysis.

6.2.2 Background

Leaving college at the end of 3rd year the main piece of advice thrown at us was, make sure that you think of your final year project idea over the course of your internship and summer. In my case this was easier said than done because I started my internship a week before the deadline. This meant I was working right through the summer until mid-August. Every time I thought about college the fact I had to come up with an idea was in the back of my head and I kept putting it off. I decided to pick Data Analytics for my 4th year specialization because I had a keen interest in my Database modules in previous years and wanted to continue to work with data but in a more advanced scale. I researched Data analytics before I chose it to see what type of work was done in it. There's an endless amount of data that you can obtain and numerous analysis' you can do on it.

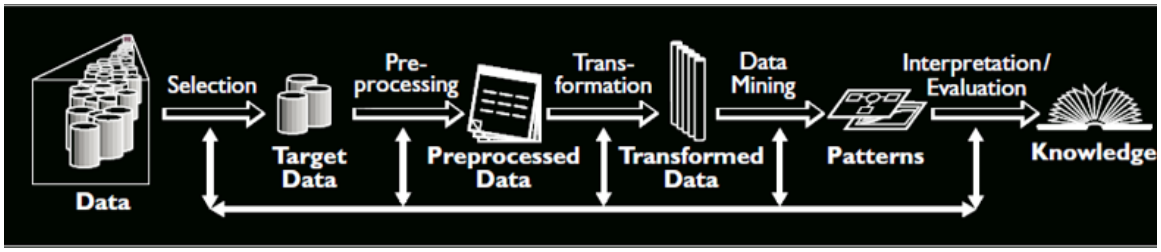
Eventually when we started back in college I still hadn't got a solid idea of what I wanted to do for my project but I did have a vague idea in my head. This idea stemmed from a database project we did in our advanced database module in 3rd year. I came up with the idea that build a database and data warehouse for an online clothing store and fill it up with mock data. I came up with the idea because something like that fits perfect with the number of tables you would need, plus the foreign keys would match up correctly. For my 4th year project I wanted to keep the clothing idea afloat and build a

project around it. I was toying around with a few ideas but I really didn't know where I was going with them so I decided to contact Michael Bradford who we had in 2nd year for Advanced Databases and see if we could come up with a couple of ideas in that field.

When we met we discussed a couple of ideas that really appealed to me. I left Michaels office with 2 solid ideas and finally picked one to run with. The idea I picked to base my project on is to build a platform that predicts future trends. Because I work part time for a high end fashion brand I wanted to base my analysis around future trends in clothing products. In the meeting that Michael and myself had he discussed that the data set that I would be using would be from social media, online magazines, blogs, forums etc. and that I would have to pull this data from these sources. The project that I am creating is a framework that should enable someone else to pick a different topic other than clothing and predict future trends about that.

When the project is complete it will be useful for various fashion retailers. For example, this project should tell you what types of jeans or what types of suits etc. will be trending in the future. This will be useful for them when they are designing the next few seasons' clothing range. In turn, getting a head start on the opposition with the type of clothes that will be in fashion will return a lot more sales and bigger profit.

6.2.3 Technical Approach



As my project is based around my specialization which is Data Analytics, I will be building my project on the KDD (Knowledge discovery and data mining) methodology. There are 5 sections to this methodology which are, Selection, Pre-processing, Transformation, Data Mining and Interpretation/ Evaluation.

Selection:
I will be pulling my data from various different sources on the internet. An example of this data is what types of clothes are in fashion at the moment. This data will be obtained from social media fashion pages, clothing blogs/ forums, online news articles. When I obtain this data this will be the mainframe for my dataset.

Pre-processing:
When I am pulling this data from these sources there will be a lot of data that I won't need. This is where I will have to try use tools to set up restrictions. When pre-processing data, it is easy to lose a few data fields. I will find strategies for dealing with these missing data fields

Transformation:
I will be looking to clean up the unstructured data that has be pre-processed. This will take up a large amount of time because it must be done manually. I will need to find invariant representations for the data.

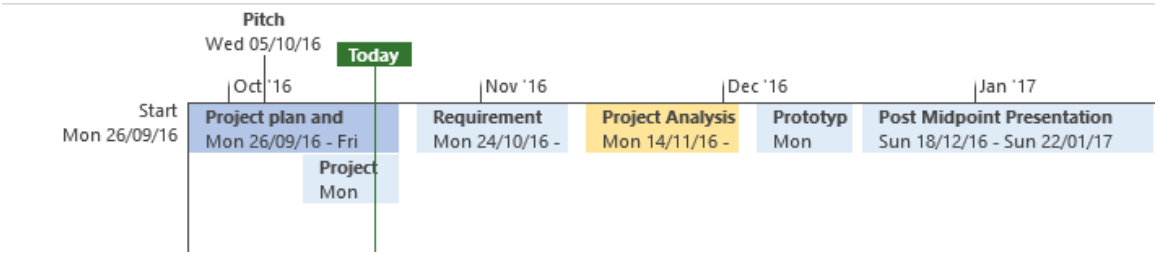
Data Mining:
I will be looking to find the best process to mine the data. From mining the data using techniques from fields such as statistics, machine learning, pattern recognition and databases I will be able to move on to the final stage off the KDD methodology.

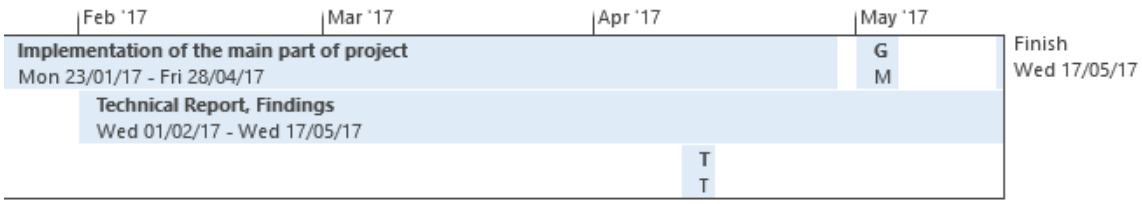
Interpretation:
This will be the final part of my main project. This is the part where I tie all of my analysis together and present it. This will be a dashboard for all of my work.

6.2.4 Special Resources Required

The main resources I feel I will need is books on predictive analysis. I will be renting books out of the library so I can read on up the main aspect of my project.
I will be mainly using my laptop over the course of the project so it is important that I download all necessary software that I need and make sure I have an adequate amount of data free.

6.2.5 Project Plan





Task Mode	Task Name	Duration	Start	Finish	Predecessors
Manually Scheduled	**Project plan and project proposal**	**20 days**	**Mon 26/09/16**	**Fri 21/10/16**	
Auto Scheduled	Meeting with lecturer to discuss my idea before the week before pitch	2 days	Mon 26/09/16	Tue 27/09/16	
Auto Scheduled	Pitch	1 day	Wed 05/10/16	Wed 05/10/16	
Manually Scheduled	Journal Entry	1 day	Fri 07/10/16	Fri 07/10/16	
Auto Scheduled	Project Proposal	10 days	Mon 10/10/16	Fri 21/10/16	3
Manually Scheduled	Project proposal Upload	1 day	Fri 21/10/16	Fri 21/10/16	
Manually Scheduled	**Requirement Specification**	**15 days**	**Mon 24/10/16**	**Fri 11/11/16**	
Auto Scheduled	Work on outlining the requirements for my projects	14 days	Mon 24/10/16	Thu 10/11/16	
Manually Scheduled	Requirement spec upload	1 day	Fri 11/11/16	Fri 11/11/16	
Auto Scheduled	Journal Entry	1 day	Thu 03/11/16	Thu 03/11/16	
Manually Scheduled	**Project Analysis and build up to prototype**	**15 days**	**Mon 14/11/16**	**Fri 02/12/16**	

Manually Scheduled	Research the web and find sources to get my data from	5 days	Mon 14/11/16	Fri 18/11/16	
Manually Scheduled	Create a MySQl database to store my data set	2 days	Fri 18/11/16	Mon 21/11/16	
Manually Scheduled	Work on extracting data from the websites I have found and put the data into my database	6 days	Tue 22/11/16	Tue 29/11/16	
Manually Scheduled	Work on putting my prototype together	2 days	Wed 23/11/16	Thu 24/11/16	
Manually Scheduled	Present my prototype	1 day	Fri 02/12/16	Fri 02/12/16	
Manually Scheduled	**Prototype - Mid Point Presentation**	**10 days**	**Mon 05/12/16**	**Fri 16/12/16**	
Manually Scheduled	Build on prototype	9 days	Mon 05/12/16	Thu 15/12/16	
Manually Scheduled	**Post Midpoint Presentation**	**27 days**	**Sun 18/12/16**	**Sun 22/01/17**	
Manually Scheduled	Work on the feedback given to me from the midpoint	4 days	Mon 19/12/16	Thu 22/12/16	
Manually Scheduled	Take break for Christmas	5 days	Mon 26/12/16	Fri 30/12/16	
Manually Scheduled	journal Entry	1 day	Wed 04/01/17	Wed 04/01/17	
Manually Scheduled	Study Break for exams	17 days	Wed 28/12/16	Thu 19/01/17	

Manually Scheduled	**Implementation of the main part of project**	**70 days**	**Mon 23/01/17**	**Fri 28/04/17**	
Manually Scheduled	Clean my unstructured data	15 days	Mon 23/01/17	Fri 10/02/17	
Manually Scheduled	Store the newly structured data in my SQL database	5 days	Mon 13/02/17	Fri 17/02/17	
Manually Scheduled	Use the methods I have been thought over the course of my final year to analyses the data I have got	30 days	Mon 20/02/17	Fri 31/03/17	
Manually Scheduled	From my analysis, bring the concept of my project into practice and start predicting future clothing trends	20 days	Mon 03/04/17	Fri 28/04/17	
Manually Scheduled	**Going over the project**	**5 days**	**Mon 01/05/17**	**Fri 05/05/17**	
Manually Scheduled	Clean up the code and structure of project	5 days	Mon 01/05/17	Fri 05/05/17	
Manually Scheduled	**Testing**	**4 days**	**Tue 11/04/17**	**Fri 14/04/17**	
Manually Scheduled	Run tests and collect the results	2 days	Tue 11/04/17	Wed 12/04/17	
Manually Scheduled	Work on any issues found	2 days	Thu 13/04/17	Fri 14/04/17	
Manually Scheduled	**Technical Report, Findings**	**76 days**	**Wed 01/02/17**	**Wed 17/05/17**	

Manually Scheduled	Review	1 day	Wed 17/05/17	Wed 17/05/17	
Manually Scheduled	**Upload Final Documentation and software**	**1 day**	**Wed 17/05/17**	**Wed 17/05/17**	
Manually Scheduled					

6.2.6 Technical Details

There are various technical approaches I will be taking when working on my project. Here I will list several software's, platforms and languages I will be using.

R
R will be the main language that I will be using throughout my project. This will obviously be run through **RStudio.** I will be using R to pull data from Twitter and from some webpages. I will need Rest APIs to pull data from Twitter such as **JSON** and **JavaScript.** Most of the work I will be using R for I have not learned yet so I will update when I am more comfortable with the language.

Python
I will also be using python at some stage along my project. I have read that Python is very useful for data manipulation and analysis. It's a similar situation with R in the sense I will be learning Python throughout the year.

MySQL
MySQL is the database platform that I am choosing to put my data in. MySQL is practical in the sense that CRUD is easily implemented when working with this type of database.

Tableau
I will be using Tableau to visually transform my data into interactive dashboards. Tableau is a simple software to work on and has a drag and drop interface. This Is an important piece of software for me because my main analysis will be shown through this.

Excel
Excel will be a place where I can store unstructured data and test data offline.

SPSS
This is going to be a good tool for me to test data and view various diagrams. SPSS can perform complex data manipulation. This is also an easy and effective tool to use.

6.2.7 Evaluation

Describe how you will evaluate the system with real technical data using system tests, integration tests etc. In addition, where possible describe how you will evaluate the system with an end user.

I will be running various tests such as API testing, Conversion testing and storage testing.

When the project is complete my main target of testers will be fashion retailers. This will be useful for fashion retailers to get a head start on their competition.

of the work that will be carried out this year

6.3 Monthly Journals

6.3.1 Month 1: September

Student name: Sean Mc Dermott

Programme: BSc in Computing

Month: September

My Achievements

For my first month, I was successfully able to come up with a project idea for my pitch. I was playing around with a very vague idea that could have had a lot of different paths take with it. I organized a meeting with one of the lecturers to discuss my vague idea and try find out how to whittle it down and get an idea from it. After about 20 minutes I had a couple of ideas on the page. This meeting enabled me to choose my idea and prepare for the pitch. I have put a small bit of

research into my project and realize how ambitious it is, seen as I'm only learning most of the content for my project this year.

My Reflection

During the summer and during the first couple of weeks of the semester I was confused on what way this project would be structured and how was I going to think of an idea then try implement it. I'm still concerned with the scale of the project and how little I know about it. I'm sure when my supervisor is assigned and the brief etc. is out I will be a lot clearer about what I have to do.

I feel the idea of bringing the project vetting in was very successful and very smart. If I had an idea that a lecturer didn't think would work I feel it's a lot better to find out now and not 3 months' time when I'm trying to put code together or get an impossible data set.

Intended Changes

For next month, I hope to talk to my supervisor on a couple of occasions and get a few goals set out. I will be majorly researching the main aspects of my project and working on the programming languages that I will using.

Supervisor Meetings

Date of Meeting:N/A

Items discussed:

Action Items:

6.3.2 Month 2: October

Student name: Sean Mc Dermott

Programme: BSc in Computing – Data Analytics

Month: 2 (October)

My Achievements

For my second month, I have not achieved as much as I was hoping to. The main thing I got done was my project proposal. This took a bit of time and effort which enabled me to get an idea of the workflow required to create a successful project.

My Reflection

I am still searching for the data set I need. I am waiting for a place to get back to me regarding this. I met with my supervisor for the first time and we discussed my idea. It worried me at how broad my project idea was and how difficult it would be to obtain a data set for my idea. This means I am going to have to think of a plan B topic for my idea but still use the same concept to build my project. For example, I want to build a 'fashion' trend predictor through predictive analysis but I might have to replace the idea of fashion with something else that with a more accessible data set.

Intended Changes

Next moth I will work very hard on obtaining a data set either for my current topic or a different topic. I must be quick about this because I haven't got much time until the end of the semester. This month I will also complete my requirements spec. I realize for me to complete this I will need to get my data set sorted ASAP.

Supervisor Meetings

Date of Meeting :28/10/2016

Items discussed: The main thing we discussed was how broad my project idea was. He told me that I really need to narrow this down to a specific type of fashion and that I should have a plan B if I can't obtain a dataset.

Action Items: Narrow my project idea down and try find a relative dataset. Also, come up with a couple of other topics in case my idea doesn't work out.

6.3.3 Month 3: November

Student name: Sean Mc Dermott

Programme: BSc in Computing – Data Analytics

Month: 3 (November)

My Achievements

For the first part of my third month I was fully concentrated on putting my requirements specification together. This took a lot longer than I was hoping because of the workload that piled up from other modules. From mapping out my requirements I gained a better understanding of set out in creating the project. I also start playing around with my data and got a greater understanding for the programming language I will be using for this project.

I have completely changed my idea as I couldn't obtain a dataset for my original idea. The project idea remains the same but the data has changed.

My Reflection

The main reflection I have from this month is the fact I have changed the data on which I want to create my project on. I feel the time was right to search for a plan B as I had to create my requirements specification. Looking back on my original idea I feel it was kind of a saving grace as I didn't really know what data I needed or what I was going to achieve if I did get data. My new idea is about predicting property prices in Dublin. The data for this idea is plentiful and readily available online.

Intended Changes

The main thing I need to change this month is my Project Proposal because it contains the details of my old idea. The reason I need to change it is because parts of it needs to be integrated into my midpoint technical report. I also need to change parts of my requirements spec to fit into my midpoint tech report.

Supervisor Meetings

Date of Meeting:04/11/2016

Items discussed: I went into this meeting with a new idea but still hadn't found an informative data set. Funnily enough my supervisor was analysing data on the

same idea I wanted to base my project around. He gave me a source to get data that he thought would be effective for my project.

Action Items: My supervisor asked me to browse through the data and write a mini report for the next meeting on what I want to do what the data sets. He said to include the questions I wanted to answer from this project by using these data sets.

Date of Meeting: 11/11/2016

Items discussed: We discussed the report that I created for my supervisor. Going through this report we agreed that I still hadn't got a clear idea of what I wanted to achieve with this data set. I also went through my requirements document. He had a quick read through and pointed out what areas I needed to improve on.

Action Items: My supervisor advised that I clearly specify the main aim of the project. He also gave me a few tips to add to the requirements specification.

Date of Meeting: 25/11/2016

Items discussed: In this meeting I wanted to have a chat about where I was at that point. I was confused about how the project was going to come together for the prototype and the second part of the project next semester. My supervisor showed me some books which would give me a better understanding for the machine learning part of the project.

Action items: If I had any spare time, to go to the library and rent the books out.

6.3.4 Month 4: December

Student name: Sean Mc Dermott

Programme: BSc in Computing – Data Analytics

Month: December

My Achievements

December was a busy month working on this project. The main achievements for the month was completing the midpoint technical report and the prototype. A lot of effort went into the tech report and it was a relief when I finally uploaded it.

My Reflection

My reflection on the past month has been a slight waste of time and frustration due to my midpoint presentation. I went into the presentation with a functional prototype which was fine but when I explained to the two lecturers what I planned on doing for my second stage of the project I was told the idea I was using would not work whatsoever. I was really appreciative of this advice because it opened my eyes to what I actually need to do but as I have never done this type of analysis before I never knew the way I was going to do it wouldn't work. Not having a clear idea of how to implement the next stage of the project had a major impact on the technical report as the vast majority of what I put in to the document is wrong. A bit more guidance that my way of implementing the next part of the project wouldn't work, would have been nice.

Intended Changes

I have major changes to make from this point on. I need to sit down with my supervisor and discuss what way he thinks I should proceed from the point I am currently at. I have a module in my next semester that covers the main aspect of my project which is predictive analytics. This is a module I will need to put all my effort into to try gain an understanding of what needs to be done.

Supervisor Meetings

Date of Meeting:02/12/2016

Items discussed: The focus of this meeting was to see what sections in the midpoint tech report my supervisor felt I should fill out and what sections I could leave for the final tech document. I also wanted to find out some of the sections meant as I was a little confused.

Action Items: My supervisor gave me some advice on what sections I needed to complete and gave me a brief description of sections I wasn't too sure about.

6.3.5 Month 5: January

Student name: Sean Mc Dermott

Programme: BSc in Computing – Data Analytics

Month: January

My Achievements

During this month, I haven't really done much work on my project due to the fact I was fully concentrated on my exams and also wanted to take a week break from any work before I got back into it.

My Reflection

I haven't reflected much on January as I haven't done much work. I plan to meet my supervisor a number of times in February to see what direction to take with my project.

Intended Changes

I have no intended changes for January.

Supervisor Meetings

No meetings in January

6.3.6 Month 6: February

Student name: Sean Mc Dermott

Programme: BSc in Computing – Data Analytics

Month: February

My Achievements

February was a month in which I started to get back into the swing of things with my project. I met with my supervisor for the first time since before Christmas and discussed my project.

I obtained a number of datasets that will aid me in my creating a model for my project.

I also successfully manipulated my dataset to a state that will enhance my analysis. I done this by writing a piece of code that moves partial data in one column to another.

My Reflection

Overall my reflection from February was positive as I gained some valuable ground on where I needed to be after a slow January. I feel I can now make some progress understanding what model I need to use to finish this project

Intended Changes

Next month I plan on finding a suitable model to use for the regression part of my project. This will involve doing a lot of research and speaking to my supervisor some more.

Supervisor Meetings

Date of Meeting:09/02/2017

Items discussed: We discussed the fact that my main dataset hasn't got many attributes to build an effective regression model from and that I should try find external data to work into this model

Action Items: I successfully found these external datasets and feel these will be a massive help in this project

6.3.7 Month 7: March

Student name: Sean Mc Dermott

Programme: BSc in Computing – Data Analytics

Month: 2 (October)

My Achievements

This month has been my most successful yet. Having obtained external data resources (luas & train station locations), I successfully generated the shortest distance from every house in my data set to a train and luas stop. This distance

will be used a parameter in my prediction model. I also discovered other ideas for parameters that will allow me to create multiple models.

My Reflection

I am happy with the progress I have made during March. I fell things will slow down a bit in April due to the end of semester submissions and exams. I will work to find time for my project and build on what I have achieved in March

Intended Changes

At this stage of the project there aren't so many changes as there are improvements. These improvements are discussed in the action items section of my supervisor meeting.

Supervisor Meetings

Date of Meeting:30/03/2017

Items discussed: My supervisor and I discussed the progress I had made with the new parameters and pushed me to expand on these parameters and try find various others that were of similar context

Action Items: I will be focusing on building an accurate model as well as locating more resources for various other models which will enhance the results of the project.