National College of Ireland

TECHNICAL REPORT

Software Project 2016/2017

Daniel Gorman – X13401792

SPORTS ANALYTICS :

ANALYSIS OF THE NATIONAL FOOTBALL LEAGUE

National College of Ireland

BSc in Computing Data analytics

Declaration Cover Sheet for Project Submission

SECTION 1 *Student to complete*

Name: Daniel Gorman
Student ID: X13401792
Supervisor: Manuel Tova-Izquierdo

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 (summarized overleaf and printed in full in the Student Handbook) and that the work I have submitted for assessment is entirely my own work.

Signature: _____

Date:_____

NB. If it is suspected that your assignment contains the work of others false

Table of Contents

Executive Summary

Americas most popular sports organization the National Football League is known for its physicality and chess-like play, although it records the highest revenue of any sporting organization around the globe, it has been slow to adapt analytics likes its competitors due to its complexity of tracking and recording the data (Lindsey, 2017). Evaluating talent within the league has become more of a priority with the new the collective bargaining agreements between the NFL and the NFLPA (Players Association) limiting practice time for players. It has resulted in talent evaluation of rookies becoming less of a priority and more importance on finding undervalued players currently in the league due to their experience and reputation of competing at the professional level. Assistant Professor of Statistics at Skidmore College, Michael Lopez has discussed in his blog 'Approximate value and the NFL draft' about the lack analytical metrics available in the NFL in comparison to the other main sports e.g. Basketball & Baseball (Lopez, 2016). NFL euthenists have not been so successful in creating the same kind of metrics as the other main sports. Instead, analysts have been viewing basic statistics in a very complex statistical sport. The closest metric so far has been approximate value (AV) which as described by Pro-Football-Reference as "putting a single numerical value on any player's season, at any position." (Drinen, 2017). It only gives an overall outlook at a player's season which isn't very definitive when asking the question "Did we use the player's best talents to return a high investment?". Modelling the return on investment in players using the format based on a statistical approach to the player's current season in relevance to their yearly contract value to find if they have been undervalued or overvalued. This method will give General Managers (GM's) a new outlook on true of a player value in the wake of experienced players increased value.

1 Introduction

1.1 Background

The initial plan to implement this research project came from a talk during 3rd-year work placement module where students were invited to hear about the different streams that will be on offer for 4th-year students. Following up the analytics stream with lecturers at the beginning of the semester statistics was the most appealing option with the combination of statistics and sports, it became an fascinating idea for the research project. It was also the perfect opportunity to undertake a project in a fast-growing area of Sports Analytics. Teams in almost every sport have an analytics department now, examples; Bayern Munich (Football), Golden State Warriors (Basketball) and Oakland Athletics (Baseball). National Football League franchises and followers of the sport have been relatively slow integrating useful analytical techniques to the sport. Research of the NFL resulted in numerous research papers discussing the lack of analytical measures to analyse players in the sport pales in comparison to other sports when analytics is involved. In other sports, i.e. Baseball and Basketball it has been easier to quantify and analyze the data in such sports and put a value on player's importance to a team but not so much with American football (Gabler, 2017), with the vast number of variables contained in just one play. The Shane Battier case study (Widjaya, 2015) which former National Basketball Association (NBA) player Shane Battier talks about how big data made him a better player overall by understanding his strengths in different scenarios and positions on the court.

Analytics has not fully taken off yet within the NFL community, with so many more variables and outcomes than other sports. Encapsulating a correct data mining approach to the sport is hard. So, therefore this research report will aim to improve the analytical strategy within American Football

1.2 Aims

The objectives of the project are to provide an exploratory analysis of the return on investment (ROI) in players, ROI will have various factors, e.g. Contract

length/value, snaps played, first downs, yards gained (offensive players). These are just some of the factors that will be used to calculate the ROI of a player to a team. The aim is to build a regression model based on the play-by-play data from the 2014 season.

Other aims of the project include creating a Clustering algorithm; clustering is the act of grouping a set of objects or in this case players based on similar statistics & physical attributes. For example, The No.1 choice player is not available anymore, and the customer wants to see what's the next best player with the most similar attributes to the No.1 choice is.

Due to time constraints, not all aims may be met which is why the project will be following an agile development structure to ensure necessary parts are completed to form the basis of intended aims.

By undertaking this project, the aim is to successfully create various models to achieve the most accurate results which will then be compared with future results to see how accurate the results were.

1.3 Technologies

MySQL – Will be used to store the datasets in an online data warehouse, which will then be imported to RStudio and Tableau for cleaning and visualizations allowing for more efficient use.

GearHost – Will be used to host the MySQL database online for remote access and added security.

R – Is a language and environment for statistical computing and graphics, it will be used to scrap the data from the internet, clean, analyze and create algorithms from the underlying data.

RStudio – RStudio is an IDE for using the programming language 'R.'

Tableau – Is software used to provide data visualization such as diagrams and graphs in an easy to comprehend way. Tableau will be utilized towards the end of the project to show the findings through various graphs that are easier to interpret.

1.4 Structure

The structure of this project will revolve around the Knowledge Discovery in Databases process, often referred to as the KDD process. KDD process is vital to any successful data analytics project. The KDD goes through the process from the gathering the raw data to cleaning, algorithms and the resulting knowledge. The various steps taken below is how the project will be structured to ensure a comprehensive analysis.

Figure 1. Overview of the steps constituting the KDD process

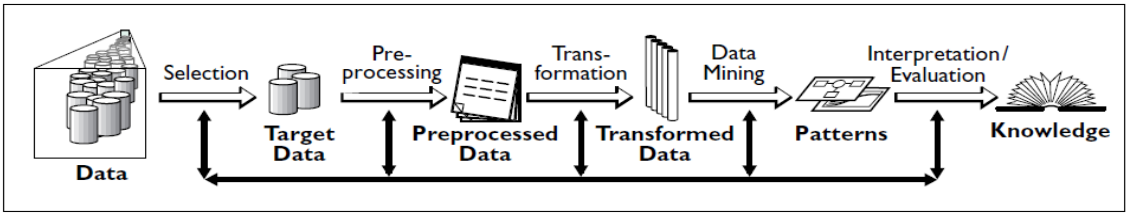


Figure 1: Overview of the KDD Process

1.4.1 Data Selection

The first phase of the KDD process is selecting the datasets needed to carry out the analysis. That is the play by play data sets sourced with the use of the nflscrapR library on R. With this library; we can then move to the next to step to select what is relevant to the project. As the project is based on the 2014 season, Selecting the right data source is a vital part of the project as it will save time moving forward. The datasets obtained for the project will be imported into an online hosted MySQL database for security and reliability purposes for the pre-processing stage. Other datasets such as contracts, snap counts and draft combines will all be scraped manually for a more accurate algorithm.

1.4.2 Cleaning Data

Cleaning and pre-processing the data requires removing any null values that may skew the results to one side or deal with them in a manner that they will not disrupt the analysis later. The play-by-play dataset contains over 45,000 rows with each row describing the play that occurred e.g. Was the ball ran or passed? Was it a complete pass or not? The datasets contain every play that happened in the

season. By cleaning the data, we can remove null values as mentioned, change our datatypes to factors, Integers or numeric types if needed and decide on the strategy to implement on missing data fields. One example of cleaning the data would be to remove each row that contains the work “Kneel.” The reason for this being as stated above it would skew a player’s stats as kneeling in the majority of cases represents when there is little time left on the game clock, and the defensive team is losing thus allowing the offense to snap the ball and kneel to run down the clock.

1.4.3 Data Mining

Data mining is a vital stage in the project; Data mining is the process of finding new information from the datasets that have been examined. During this phase, the datasets will be reviewed to uncover trends about teams well as player’s ROI and tendencies from the play-by-play data. Furthermore, after the data mining stage has been completed, the data can then be pass through several machine learning algorithms to find correlations and value of current players return on investment.

1.4.4 Algorithms Applied

Algorithms will be applied after the data mining phase has been completed to try to find what measurable having the making of a successful NFL player in his position. Doing this all allow will provide insight into to what makes, for example, a defensive player considerably better than others at his position, is its speed, strength, jumping ability? Using algorithms such as regression and clustering will allow the data to be viewed in different ways and uncovering insights to discover answers to finding value in players and their correlation to success.

1.4.5 Knowledge/ Interpretation

The last stage of the KDD process is a crucial stage. With the findings from the stages described above, reports can be wrote based on the findings and what teams should be doing regarding salary negotiations & player signing. Also from the data mining activities and algorithms applied, useful insights can be displayed on graphs to give the user a better understanding of the findings displaying them

on visuals such as Tableau making it clearer to potential customers the benefits of signing “Player A” over “Player B” for example.

1.5 Definitions, Acronyms, and Abbreviations

1.5.1 Definitions

Contract year – Player playing in the final year of his contract with option to leave at the end of the season.

Measurable – Players measurements were taken at the NFL combine e.g. Arm Length, Height, Vertical

Combine – Each player entering the draft is invited to partake in a day of activities measuring players on a series of tests e.g. 40-yard dash, bench pressing, vertical jump.

Database storage – A relational database management for storing datasets.

Programming Application - Is a set of subroutine definitions, protocols, and tools for building software and applications. e.g. RStudio, MySQL Workbench.

Trading - Is transferring a player to another team in exchange for assets in the form of money, players or future draft picks.

Signing – Agreeing on a contract with free agents to join a franchise

Unguaranteed salary – Player salary that includes incentives to perform well and if not met, player does not receive full contract salary

Guaranteed Salary – Salary agreed upon where a player receives a portion of a contract in full without the risk of not meeting objectives set out.

Dropbox - Cloud storage service for sharing and storing files including photos, documents, and videos.

Franchise – Like a football club, the NFL grants permission for a club to build in a certain geographic area.

Collective Bargaining Agreement – The labor agreement sets forward the distribution of league revenues, sets health and safety standards and establishes benefits for NFL players active and retired

NFLPA – NFL Players Association, Union as such operated by active NFL players

Selector Tool - CSS selector generation

1.5.2 Acronyms and Abbreviations

ROI – Return on Investment

AV – Approximate Value

NFL – National Football League

KDD – Knowledge Discovery in Databases

IDE - Integrated development environment

KNN – Knowledge Discovery in Databases

2 System

2.1 Requirements

2.1.1 Project Scope

The extent of the project is to develop an exploratory study into return on investment in players to their respected franchise e.g. “Did we get your value for money with Player D.” The system would be utilized to find an expected value of a player and comparing it with his ROI if traded with the focus on building a winning team with value at a lower cost.

2.1.2 User requirements

An exploratory study into the return on investment in players to their respected franchise when making the decision if players have proved their contract value worth with on-field performances. The system will give information on the statistics of each player in a readable format showing players estimated value, player types; contracts cost, player profile and measurable’s, results will be visible via dashboard for easy viewing.

2.1.3 Data Requirements

The data was sourced from R library nflscrapR. The datasets contain data for the year 2014. Data dictionary is shown in Table 1

Other datasets were scraped from websites that host information related to the player contracts, combine and draft information.

2.1.4 Output requirements

The client requires analysis and visualization of the database data; this includes the following.

- Provide a visualization of results of why the replacement is a viable option and can hold more value than current player
- Descriptive statistics to summarize the data on team by team basis

- Visual: box plot; summary table;

- Predictive analytics, explore the associations within the data and modeling relationships with the different years of data available to create a model of player value is increasing/decreasing

2.1.5 System Evolution

The system could evolve with new metrics for grading players becoming available over time. With newer metrics becoming available could lead to adding new features for spotting trends. If similar data could be obtained for high school and college level, it could lead to a system being developed at all levels to get a more accurate grade of player and rank in the draft system. The system has potential to be very precise if contracts can be dealt with as rookie contracts could skew any possible model building causing experienced players value to fall with rookie contracts disrupting the model build with their low value.

Table 1: Data Dictionary of Play-By-Play Data After cleaning

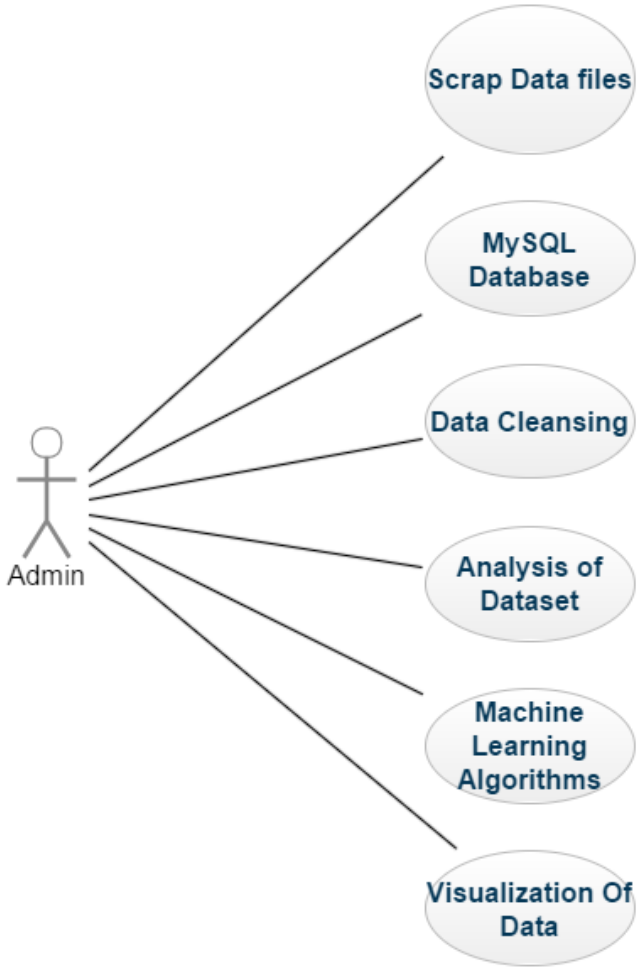
Name	Description
Pbp2014_id	Primary key for database
GameID	Unique ID for each game of the 2014 season
Drive	Equivalent to number of possessions a team has had
Qtr	Current Quarter the game is in, 4Q's each game
Down	Each time has 4 Downs or chances to advance the ball 10 yards until they reach the endzone or the defensive team receives the ball
TimeSecs	Time remaining in the game

SideOfField	Indicates which side of the field the ball is located
YrdIn	Position on the Off/Def Side of the field
YrdLn100	Position on the field
Ydstogo	Yards from the defensive team end zone
Ydsnet	Yards gained on current drive
GoalToGo	1=Down inside opponents 10-yard line, 0= Outside opponents 10-yard line
FirstDown	0=Play did not result in new set of downs, 1= Play did result in new set of downs
DefensiveTeam	Defensive Team on the field
Yards.Gained	Yards gained on the current play
Touchdown	0= No touchdown, 1= Touchdown scored
Safety	0=”No Own Goal”, 1= “Own Goal.”
PlayType	Type of play, e.g. Run, Pass, Punt, Kick
Passer	Name of player passing the ball
PassAttempt	0=No pass attempted, 1=Pass attempted
PassLength	Distance of pass, short, medium,long
PassLocation	Position of pass, left, middle, right
InterceptionThrown	0=No INT thrown, 1= INT Thrown
Rusher	Player who ran the ball if play type occurred
RushAttempt	0=No rush attempt, 1= Rush attempt occurred

RunLocation	Which way the player rushed
RunGap	Run inside or outside
Receiver	Name of player who caught the ball if pass play occurred
Reception	0=Player did not catch the ball, 1=Player caught the ball
Down	Current down of the offensive team, the team had four plays to get gain 10 yards to get a new set of downs.
ScoreDiff	Score difference between teams

2.1.6 Functional requirements

2.1.6.1 Use Case Diagram



2.1.6.2 Requirement 1 <Scraping Data Files>

Description & Priority

Scraping data is the highest importance of all the use cases as, without it, the rest of the use cases cannot function. This use case describes the operation of scraping data sources using RScripts.

Use Case

The admin must have access to RStudio and the corresponding RScripts to scrap and export the new data to CSV files.

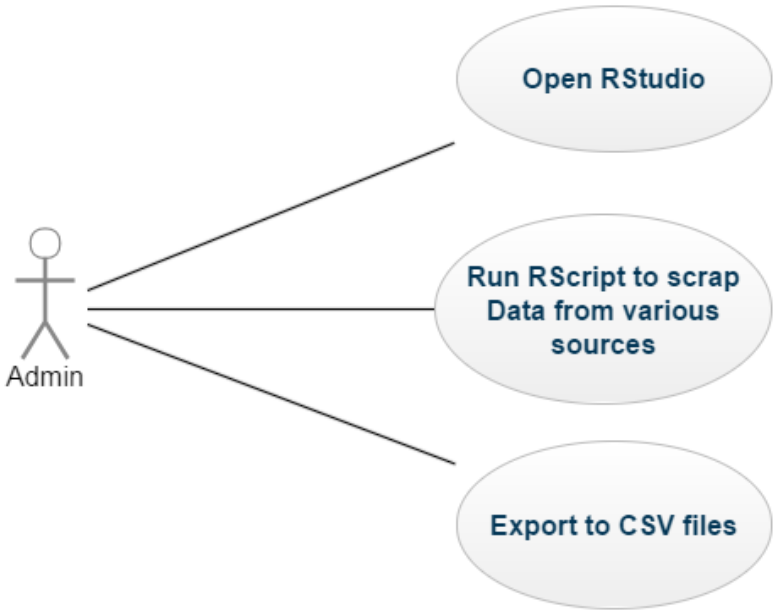
Scope

The extent of this use case is to show how the admin scraps data with the use of RScripts to create new CSV files for analysis.

Description

This use case describes the admin opening R IDE RStudio and executing the RScript to pull data from the various sources and then storing them in new CSV files.

Use Case Diagram



Flow Description

Precondition

The admin must have access to RStudio and the RScripts.

Activation

This use case starts when an <Admin> opens RStudio.

Main flow

1. The <Admin> Opens RStudio
2. The <Admin> Opens & Runs RScript to scrap data
3. The <Admin> Exports new data to CSV files

Termination

The data has been successfully exported CSV files.

Post condition

The system goes into a wait state.

2.1.6.3 Requirement 2 <Database creation and import>

Description & Priority

This use case holds high importance to the completion of the project, without it, the rest of the use cases will struggle to run at an efficient level. This use case is for the creation of the MySQL database and the import/export to and from programming application.

Use Case

The admin must of access to the database application to begin the process of creating a database storage.

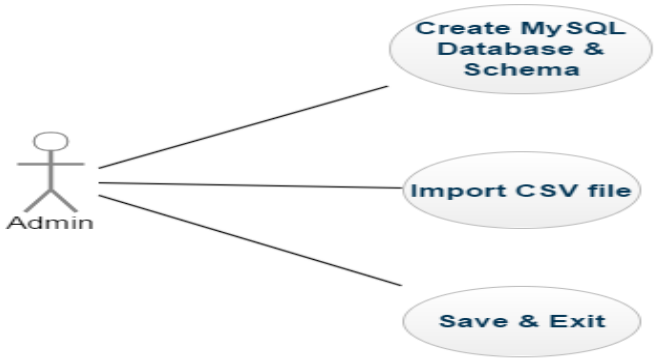
Scope

The extent of this use case is to show how the admin creates a MySQL database and tables.

Description

This use case describes the admin creating the database and tables, then importing the CSV files containing the raw data into the new database.

Use Case Diagram



Flow Description

Precondition

The admin must have access to the database and CSV files.

Activation

This use case starts when an <Admin> creates a database and passes through the correct CSV files

Main flow

4. The <Admin> creates database and tables
5. The <Admin> Imports the right CSV files into the database
6. The <Admin> checks database now holds the imported data.

Termination

The database has been successfully created and populated.

Post condition

The system goes into a wait state

2.1.6.4 Requirement 3 <Cleaning data>

Description & Priority

This requirement would be considered level 3 importance as cleaning the datasets is of high importance to analysis the datasets without encountering problems later in the analyzation and having outliers or Null contaminating the datasets.

Use Case

The admin must have access to the database created to begin the process of cleaning the data.

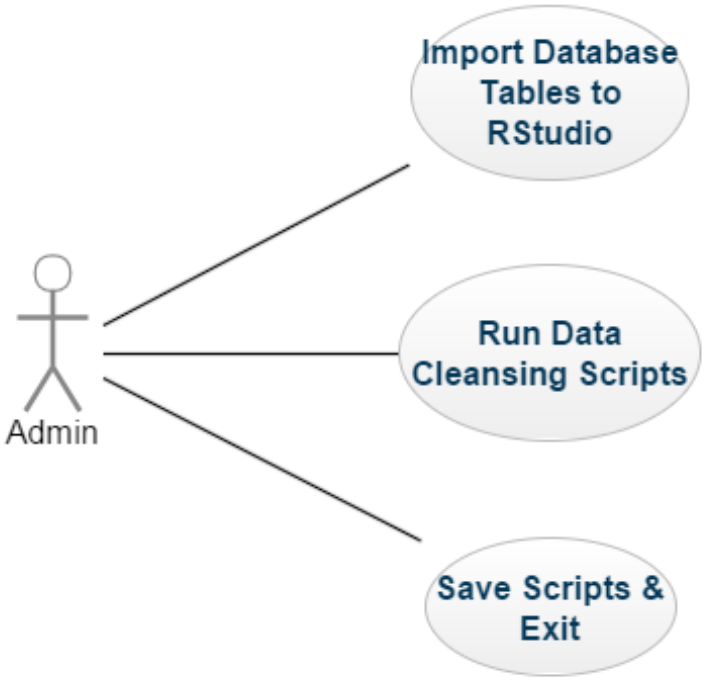
Scope

The extent of this use case is to show how the admin imports the datasets from the database to RStudio and runs RScripts to clean the data.

Description

This use case describes the admin importing the datasets from the database storage into RStudio, using the application to clean the datasets before usage.

Use Case Diagram



Flow Description

Precondition

The admin must access to the database, RStudio, and RScripts for importation.

Activation

This use case starts when an <Admin> Opens RStudio to import the database files

Main flow

1. The <Admin> opens RStudio and connects to the database using RScripts created
2. The <Admin> Imports the correct CSV files into the RStudio.
3. <Admin> proceeds to clean datasets.
4. The <Admin> Saves Cleaning Scripts

Termination

The <Admin> exits RStudio.

2.1.6.5 Requirement 4 <Data Analysis>

Description & Priority

This requirement would be considered level 2 importance, analyzing the data is important to the end goal of finding the results we are looking seeking.

Use Case

Admin must have access to the database to begin the process.

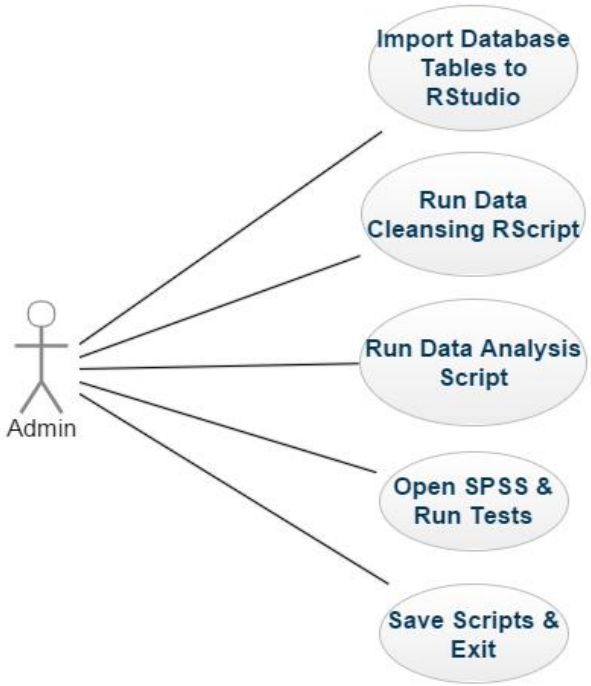
Scope

The extent of this use case is to show how the admin uses RScripts to do analysis on the data sets to find patterns and results.

Description

This use case describes the admin importing the dataset from the database storage, next it shows the admin using RScripts cleaning the data and performing exploratory data analysis using a host of libraries available, the next step shows the admin reviewing the results in SPSS and finally saving the new RScripts for reuse later.

Use Case Diagram



Flow Description

Precondition

The admin must access to the database and RStudio for importing and analyzing the data.

Activation

This use case starts when an <Admin> opens RStudio and imports the datasets

Main flow

1. The <Admin> Opens programming application and connects to the database
2. The <Admin> Imports the required datasets from the database
3. The <Admin> Run Cleaning scripts before analysis
4. The <Admin> Runs exploratory data analysis on data
5. The <Admin> Review results in SPSS
6. The <Admin> Saves results for reuse

Termination

The <Admin> closes applications

2.1.6.6 Requirement 5 <Machine Learning>

Description & Priority

This requirement would be considered level 5 importance. It is the 2nd the last step in project shows the admin importing the data, cleaning and running various machine learning techniques

Use Case

The admin must of access to the database created to begin the process of applying machine learning algorithms.

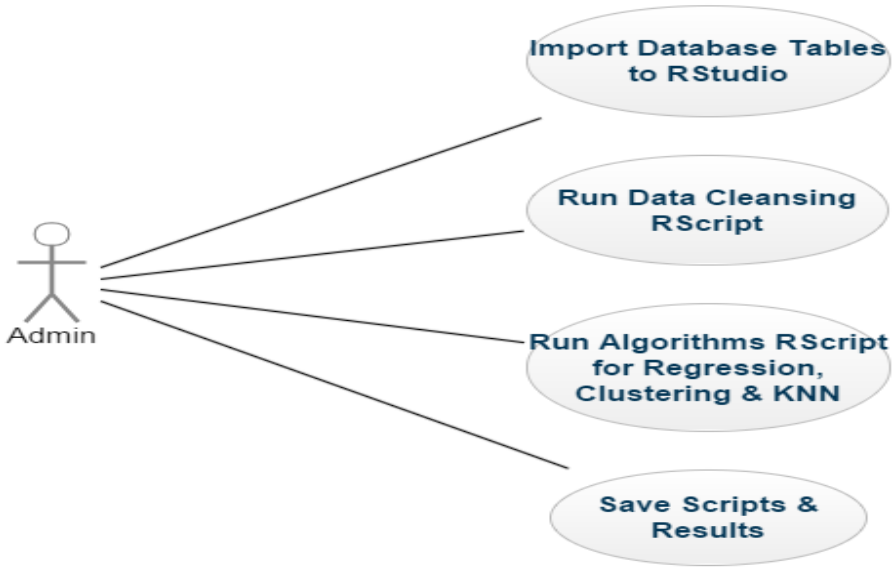
Scope

The extent of this use case is to show how the <Admin> uses machine learning algorithm(s) to create a model for predictive analysis on a player's future ROI to show the actual value of a player also using clustering and classification techniques.

Description

This use case describes the admin performing machine learning techniques on the data sets acquired to create a predictive model on player's ROI to assess the future of a player's career and risk involved signing him.

Use Case Diagram



Flow Description

Precondition

The admin must access to the database and RStudio Scripts for importing data and running machine learning algorithms.

Activation

This use case starts when an <Admin> Opens RStudio and initiates a connection with the database.

Main flow

1. The <Admin> Opens RStudio and connects to the database
2. The <Admin> Imports the required datasets from the database
3. The <Admin> Use RStudio and algorithms to create machine learning models
4. The <Admin> Save Scripts & Results for future use.
5. The <Admin> Exit program.

Termination

The <Admin> closes applications

2.1.6.7 Requirement 6 <Data Visualisation>

Description & Priority

This requirement would be considered level 5 importance. Also, Data visualization has a high importance as being able to convey the result to customers is important as analyzing the data.

Use Case

The admin must of access to the database created to import the files into Tableau for visuals

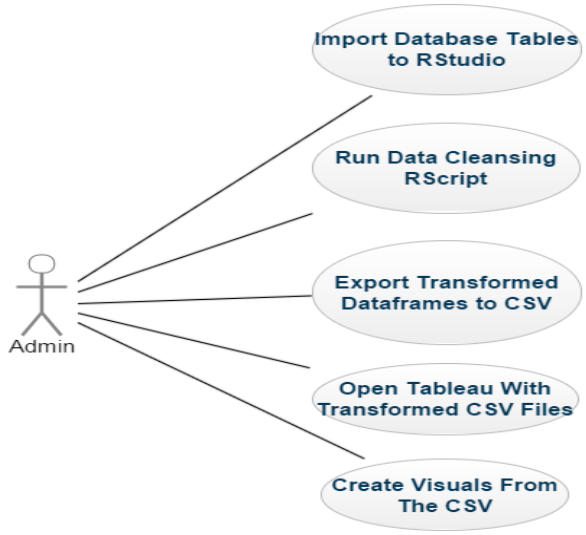
Scope

The extent of this use case is to show how the <Admin> creates data visualizations from the datasets.

Description

This use case describes the admin opening RStudio and importing the datasets from the database created, then the Admin running RScripts to clean the data and export CSV files for use with Tableau to create visual diagrams to represent the data in an easier view for reports and customers.

Use Case Diagram



Flow Description

Precondition

The admin must access to database, RStudio, RScripts & Tableau for achieving results.

Activation

This use case starts when an <Admin> Opens RStudio and initiates a connection with the database.

Main flow

1. The <Admin> Opens RStudio and connects to the database
2. The <Admin> Imports the required datasets from the database
3. The <Admin> Runs RScripts to clean and export data to CSV files
4. The <Admin> Uses Tableau to visualizations diagrams and graphs
5. The <Admin> exports results (graphs and diagrams)

Termination

The Admin closes the connection with the database and closes the program.

2.1.7 Non-Functional Requirements

2.1.7.1 Performance/Response time requirement

The volume of data is approx 45,000 rows for the season of data. Response time is varied on what metrics the system is being asked to process; A specific player may be quicker to process than searching for a broader player type with fewer metrics involved.

2.1.7.2 Availability requirement

The system will be available before the project deadline with all material and code stored for reusability on GitHub.

2.1.7.3 Recover requirement

All data files and scripts will have backups locally and in the cloud with the use of Dropbox. MySQL Database will be hosted online with Gearhost in the case of hardware failure of a PC/Laptop which will allow for easier recovery and the ability to work from home and college with database login.

2.1.7.4 Security requirement

The database will be kept secure with passwords to prevent the misuse and leaking of machine learning algorithms.

2.1.7.5 Reliability requirement

The data has gathered by Carnegie Mellon University statistical researchers, which use an API to scrap and parse data from the official NFL website.

2.1.7.6 Extendibility requirement

Plans to continue the study, using the material and knowledge learned to develop the project into the future with the idea of branding it as a service.

2.1.7.7 Integrity requirement

The integrity of the data is vital to obtaining conclusive results into the study of the topic. Accuracy and consistency throughout the data will lead to more conclusive results after the study has concluded.

2.2 Design and Architecture

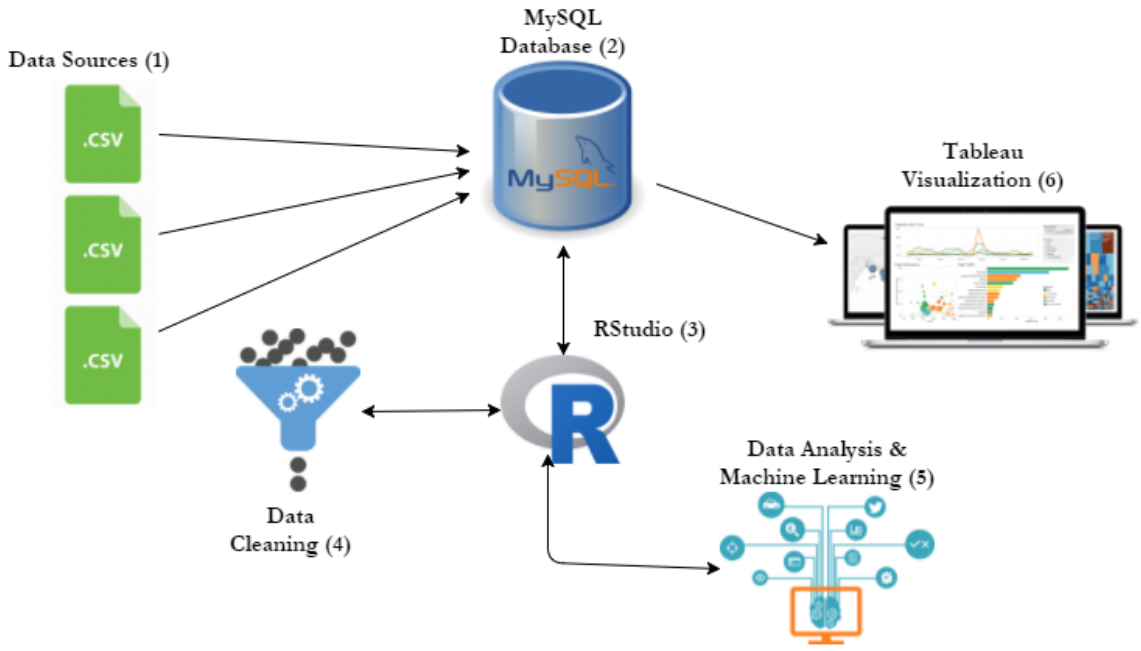


Figure 2: System Architecture

The Architecture of the system begins with the data sources being imported into the MySQL database created. From there the data will be imported into RStudio

with RScripts, from here the imported data sources from the database will be cleaned in preparation for analysis. The next stage involves analyzing the datasets and then applying machine learning algorithms (Regression, Clustering) to find patterns and answer the questions put forward at the start of the project. Finally, the cleaned data and results from the machine learning stage can be exported back to the MySQL database ahead of the data visualizations in Tableau.

2.3 Implementation

2.3.1 Domain Knowledge

The first part of the Knowledge Discovery in Databases involves understanding the domain of the project data. With prior knowledge of the NFL, it was an added benefit when research began for an appropriate dataset, it must contain the all the necessary variables that represented the play. Knowing the domain of the data is crucial when deciding what data mining technique to implement. As discussed earlier data mining is the application of exploiting patterns in datasets, to be successful, the researcher must understand what data mining approaches can work with the data and what can't

2.3.2 Data Selection

Beginning the project started with research of how data could be scraped from online sources, after some time searching the internet It was evident the R library called ‘nflscrapR’ on GitHub was the best fit for the primary data set. NflscrapR is an R package developed by Carnegie Mellon University statistical researchers, which use an API to scrap and parse data from the official NFL website. With this library, It could extract the official play-by-play data from any season I wished. Implementing this is shown below:

It began by installing ‘devtools’ package, devtools is an R package that allows users to install packages that were not located on CRAN (Comprehensive R Archive Network) CRAN is the storage center for all things R packages hosting all

the latest R versions of code along with the documentation. The following is the usage of the devtools package:

```
#Install Package
install.packages("devtools")
#Enable Library
library(devtools)
```

This code installs the devtools package to the local computer, and the package is enabled through the library command.

```
#Allows Me To Install the nflscrapR library from Github Repo
install_github(repo = "maksimhorowitz/nflscrapR")
#Importing & Storing 2014 play-by-play data in a data frame
pbp2014 <- season_play_by_play(2014)
```

With devtools enabled the nflscrapR package from Github could be installed to scrap the 2014 NFL Play-By-Play data, saving it in a data frame called pbp2014 which were then used later for analysis and creating machine learning algorithms.

Next on the list of scrapings was scraping the SnapCount from football outsiders, SnapCount is a term employed in American Football to count how many plays a player appeared. As there are unlimited substitutions in American Football, This was included in the model building to fully explain a player's value to the team as it goes behind the general statistic of games played by an individual.

```
#Web Scraping package
library(rvest)
#Scraping NFL SnapCount from Football Outsiders
# http://stackoverflow.com/questions/38257579/using-r-to-navigate-and-
scrape-a-webpage-with-drop-down-html-forms
snap <-html_session("http://www.footballoutsiders.com/stats/snapcounts")
  FootOutside<-html_form(snap)[[3]]
  filled_form <-set_values(FootOutside,
                           "team" = "ALL",
                           "week" = "ALL",
                           "pos"  = "ALL",
                           "year" = "2014"
   )
  d <- submit_form(session=snap, form=filled_form)
  SnapCountQB <- d %>%
    html_nodes("table") %>%
    .[[2]] %>%
    html_table(header=TRUE)
```

The above snippet is the 2nd scraping method performed, using the html_session R is told to simulate a session in an HTML browser with the given URL, html_form is then used to tell R that this web page contained a form that needs to be filled

correctly with right variables required otherwise it would return the default data. Submit_form simulated pressing the form button storing it in R so that all the information could be transformed to the SnapCountQB data frame.

```
#Contract Values
#http://www.michaeljgrogan.com/rvest-web-scraping-using-r/
#http://selectorgadget.com/
html <-read_html("http://www.spotrac.com/nfl/rankings/2014/contract-value/quarterback/")
player <- html_nodes(html, ".team-name , .tablesorter-headerUnSorted tablesorter-header-inner")
cash <- html_nodes(html, ".noborderright:nth-child(4) :nth-child(1)")
player <- html_text(player)
cash<- html_text(cash)
Value = data.frame(player, cash)
```

Above snippet from R was the final scraping method used to collect contracts details from Spotrac.com with the use of Google Chrome extension SelectorGadget, SelectorGadget is an open source tool that makes CSS selector generation and discovery on newer more complicated websites easy. As Spotrac website was using a slightly more complicated drop-down menu, it was acceptable to use SelectorGadget in this instance. The implementation was more manual than the previous methods as it involved filling in the drop-down form manually selecting the fields one column at a time that applied to goals of the project before writing the code to extract the information. Reading the website URL was executed with read_html(), storing it as an object in R. html_nodes allow users to extract pieces of HTML using CSS selectors which are where SelectorGadget is used to select first the player names and store them as an object ‘player’ followed by contract value being stored as object ‘cash’. html_text() converts the stored nodes into readable text. Since both player and contract values were being stored separately, it was possible to use the data.frame() to merge the objects together into a data.frame which will be used extensively later. The newly scraped data were stored as CSV files.

2.3.3 Database

After successfully gathering the data sets and doing some pre-cleaning needed to proceed with this project the next step was to store the data in an online hosted database, this was a way to protect against the accidental damaging of a CSV file

and protect against any hardware failure whilst easily allowing easy access to the files with just a username and password from any computer. Working on the project from both home and college effectively would require the database to hosted online with GearHost, GearHost has been defined as a Platform as a Service (PaaS) built for developers to manage applications and databases in the cloud. After creating a database on GearHost website and successfully connecting the database with MySQL Workbench, the next phase was to set up a table that would be able to handle the incoming data. Using the RPackage RMySQL to initiate a connection with the database I could then create tables by exporting the data to the MySQL database using the WriteTable Command.

```
#WriteTable To Database
dbWriteTable(con, "pbp1", pbp2014f[1:45502, ], append = TRUE, row names
= FALSE)
```

With the database set up and the data loaded into the tables, R could be used with MySQL to load in tables from MySQL to R as data frames for cleaning and analysis.

2.3.4 Cleaning & Reduction

Cleaning is a massive part of having a successful project, many parts of algorithms and standard functions will not work if the data is not cleaned and setup in the correct format that users intend to use. As the main dataset of the research is the play by play of the 2014 NFL season depending on the play, only certain variables will record data. i.e. If it is pass play, columns like Rusher, RushAttempt, RunGap won't apply as they would only be used if the ball is rushed thus leaving rows of data with NA values which need to be carefully dealt with for an effective analysis. Before the data could be inputted into the newly created database from eyeballing the multiple datasets, there was a glaring problem that player's names were being spelt differently in each dataset which meant implemening some pre-cleaning through the use of Excel to save time when merging datasets, later on, the rest of the cleaning will need to be completed after the database creation stage in preparation for exploratory analysis. Throughout the cleaning phase, the powerful R package dplyr will be used, with this the data can be manipulated in tons of

different ways. Data reduction is major operation within data mining as picking the right variables is critical to creating an accurate model. Data reduction is discussed in this section on how the pbp2014 dataset was transformed and reduced to display valuable features.

Before uploading any CSV files to a database, some house cleaning was necessary. Through the various datasets, the datasets spelled players names differently which would cause a problem when merging the data later.

Table 2: Player Name Problem

Dataset	Player_Name	Problem
Contract	**Aaron Rodgers** QB – Green Bay	**First & Last name spelled normally**
Draft	**A. Rodgers** QB – Green Bay	**The first name abbreviated & the Last name are spelled fully**

From the above example, we can tell that this player is the same, but the way his name has been represented is different. Since humans automatically know this and computer machines cannot it can cause a problem when performing tasks as the computer applications will behave as if the player is two different instances. Solving this problem was done by abbreviating the first name of players was completed through Excel functions i.e.

Table 3:Solution To Player Name Problem

Contract Dataset	Player_Name	Transform
Old	**Aaron Rodgers**	**Abbreviate the first name to initial of the first name followed by. Last Name**
New	**A.Rodgers**	**Compatible with other datasets**

Solution was achieved through Excel function LEFT, LEN, RIGHT

Table 4: Excel Function To Parse Out Strings

Excel Function
=LEFT(B2,1) &"."&RIGHT(B2, LEN(B2)-SEARCH(" ",B2))

With this function, it was possible to parse out substrings from strings of data. This function works in a way where ‘B2’ is the excel cell of the player, LEFT(B2,1) &”.” would remove all but the first character in player’s name followed by “.”. RIGHT(B2, LEN(B2) searches from right to left with LEN telling Excel to retain the length of string and finally -SEARCH(“ “,B2). Tells Excel to return the string minus the first space it finds, When the Excel functions are combined, it will create a new column to allow the merge of different datasets together based on the name.

This Excel function was a small but necessary process before importing the datasets into a database.

After setting up a connection with MySQL for importing the database, When the data was imported some variables came through in the wrong format due to NA values. Below is an efficient way of dealing with a large dataset that requires changes of the same nature.

```
#Changing Columns from Chara To INTs
cols <- c(2:8,10:15,18:21,25:26,30,34,37,41,48:49,52:53,55,62:63)
pbp2014[cols] <- lapply(pbp2014[cols], as.integer)
```

With this, it was possible to combine all the columns needed by their position in the dataset and store them in an object called ‘Cols’. The lapply function is part of the apply family which allow a user to apply functions over a range of different objects. lapply allows users to apply a function to a list of vectors; Cols is a list of vectors created in this instance. With as.integer, we are telling lapply to change every column in the vectors data type to an integer allowing the user to run statistical tests which would not have been able to if the columns remained as characters.

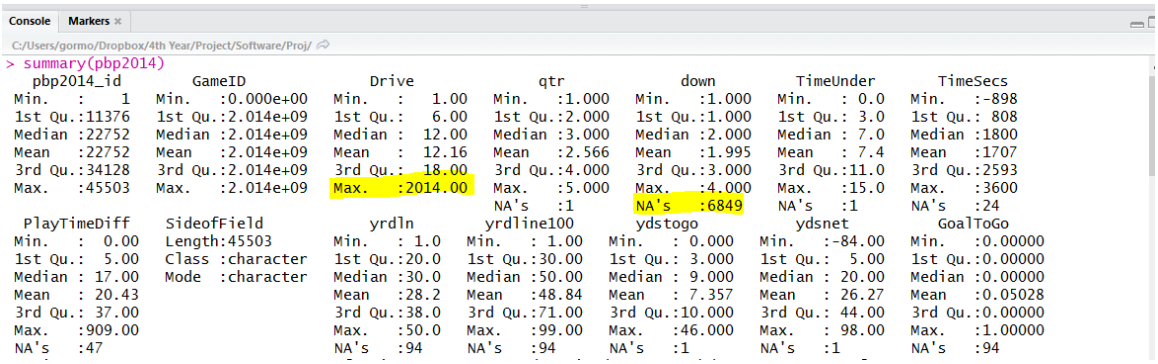


Figure 3: Summary() command executed to explore the data frame

Figure 3 above is an overview of the underlying dataset pbp2014; it allows a user to view the dataset and its records for some common statistics such as Mean, Median & Max. It is also fascinating when a user first come across a dataset to find if there is any outliers or corrupt values hidden in the data. Two instances of this have been highlighted in Figure 3. The first is 'Drive,' Drive has a max of 2014. Drives, as mentioned before, is the total amount of possessions a team had during one game. To a spectator of the sport, it is noticeable 2014 drives in one game cannot be right but to the untrained eye, a quick look at the other statistics such as Mean, first and the third Quartile it becomes obvious that 2014 drives are there by accident. To rectify the mistake, the user can view the dataset to confirm the theory that the specific row is there by accident, after confirming the assumption we can remove the row, so it will not cause any problems when running models or creating graphs. The following is a snippet showing the removal

```
#Remove row W/Outlier
pbp2014 <- pbp2014[-22801,]
```

Table 5: Before & After With outlier removed

Old Down	**Min. 1.00**	**1st Qu. 6.00**	**Median 12.00**	**Mean 12.16**	**3rd Qu. 18.00**	**Max. 2014.00**
New Down	**Min. 1.00**	**1st Qu. 6.00**	**Median 12.00**	**Mean 12.12**	**3rd Qu. 18.00**	**Max. 33.00**

We simply filter the Drives column to find the outlier and then remove it by telling R we want the specific row to be deleted and the data frame saved as in this case as pbp2014.

Secondly, the 2nd highlighted row in figure 3 tells us that the in the down column there contains 6849 empty rows with no information. Usually we would remove NA as they can cause problems creating graphs among other things, but in this case, after examining some empty rows NA rows in the Down column signal a stoppage in play such as a timeout or kickoff where no play occurred.

Next stage in cleaning takes place when removing unneeded columns to make viewing the data frame easier saving time as a result. It can be accomplished with the code snippet below, where the data frame is stated, the column to be removed

and finally telling R to make it Null or in other words “Remove DefTwoPoint from data frame pbp2014 it is not needed.”

```
#Removing Columns Not needed
pbp2014[["ExPointResult"]] = NULL
pbp2014[["TwoPointConv"]] = NULL
pbp2014[["DefTwoPoint"]] = NULL
```

Common practices of cleaning also include changing datatypes to factors; factors are as categorical variables, it is an indication to R that is variable is nominal e.g. its only purpose is to serve as a name, the snippet below is an example of converting a variable to a factor.

```
#Changing To Factors/Numeric/INT
draft$position = as.factor(draft$position)
draft$college = as.factor(draft$college)
```

In other datasets, similar methods of cleaning are required i.e. changing of column names, grouping several player positions together as one group to merge more efficiently with another dataset. This was the case when I tried merging two datasets together but found the positions on the two datasets were name differently, my solution was to group the players positions into more general positions i.e. Defensive backs instead of specific positions i.e. Cornerback, Free Safety, Strong Safety

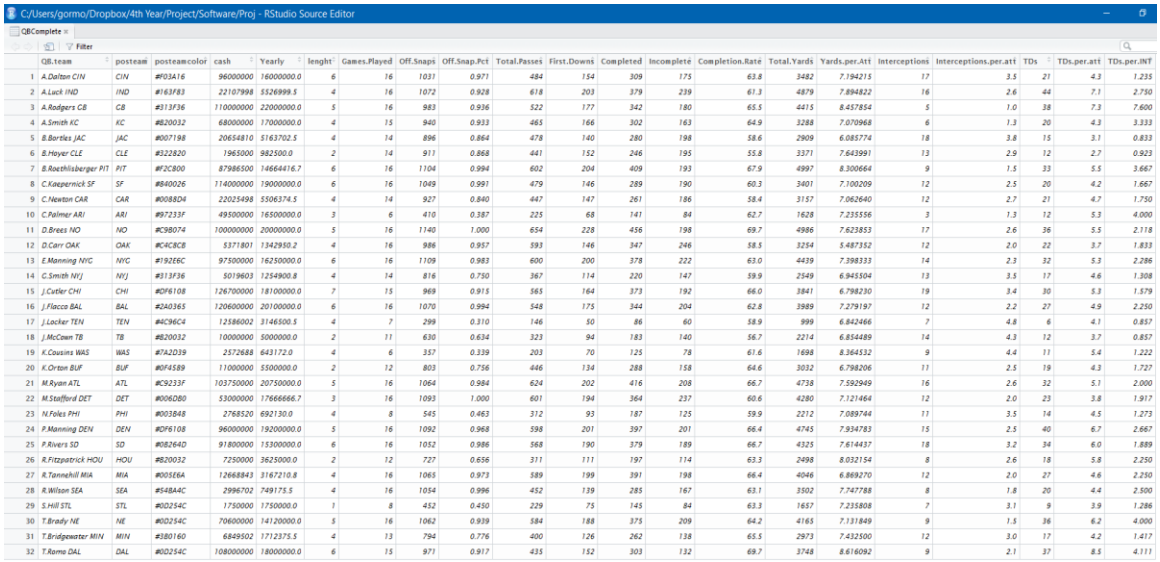
```
#Recoding Factors for Merging
combine$position <- combineLevels(combine$position,levs = c("CB", "FS",
"SS"), newLabel = c("DB") )
combine$position <- combineLevels(combine$position,levs = c("ILB",
"OLB"), newLabel = c("LB"))
combine$position <- combineLevels(combine$position,levs = c("C","OC"),
newLabel = c("C"))
```

it is a small part of cleaning but makes the project go a lot smoother when the data is assessed and cleaned right.

After implementing different techniques of cleaning the dataset, it is now time for transformation of the data with the use of the dplyr package which allow users to manipulate the data into different forms.

```
#Changing "-" in date to "/" So I can convert date
pbp2014$Date <- gsub("-", "/", pbp2014$Date)
pbp2014$Date<- as.Date(pbp2014$Date, format = "%Y/%m/%d")
```

After importing from the created database, the issue of dates is then encountered. The issue is when the database is imported RStudio see the dates column as a character due to the '-' in between the dates and not the '/' it is designed to recognize. The solution to this was to use the gsub function in the dplyr package to remove the '-' and replace it with the '/' so RStudio knows we want the column in the American-style date format.



	QB.team	posteam	posteamcolor	cash	Yearly	lenght	Games.Played	Off.Snaps	Off.Snap.Pct	Total.Passes	First.Downs	Completed	Incomplete	Completion.Rate	Total.Yards	Yards.per.Att	Interceptions	Interceptions.per.att	TDs	TDs.per.att	TDs.per.INT
1	A.Dalton CIN	CIN	#F03A16	96000000	16000000.0	6	16	1031	0.971	484	154	309	175	63.8	3482	7.194215	17	3.5	21	4.3	1.235
2	A.Luck IND	IND	#163F83	22107998	5526999.5	4	16	1072	0.928	618	203	379	239	61.3	4879	7.894822	16	2.6	44	7.1	2.750
3	A.Rodgers GB	GB	#313F36	110000000	22000000.0	5	16	983	0.936	522	177	342	180	65.5	4415	8.457854	5	1.0	38	7.3	7.600
4	A.Smith KC	KC	#B20032	68000000	17000000.0	4	15	940	0.933	465	166	302	163	64.9	3288	7.070968	6	1.3	20	4.3	3.333
5	B.Bortles JAC	JAC	#007198	20654810	5163702.5	4	14	896	0.864	478	140	280	198	58.6	2909	6.085774	18	3.8	15	3.1	0.833
6	B.Hoyer CLE	CLE	#322820	1965000	982500.0	2	14	911	0.868	441	152	246	195	55.8	3371	7.643991	13	2.9	12	2.7	0.923
7	B.Roethlisberger PIT	PIT	#FFC800	87986500	14664416.7	6	16	1104	0.994	602	204	409	193	67.9	4997	8.300664	9	1.5	33	5.5	3.667
8	C.Kaepernick SF	SF	#840026	114000000	19000000.0	6	16	1049	0.991	479	146	289	190	60.3	3401	7.100209	12	2.5	20	4.2	1.667
9	C.Newton CAR	CAR	#0088D4	22025498	5506374.5	4	14	927	0.840	447	147	261	186	58.4	3157	7.062640	12	2.7	21	4.7	1.750
10	C.Palmer ARI	ARI	#97233F	49500000	16500000.0	3	6	410	0.387	225	68	141	84	62.7	1628	7.235556	3	1.3	12	5.3	4.000
11	D.Brees NO	NO	#C98074	100000000	20000000.0	5	16	1140	1.000	654	228	456	198	69.7	4986	7.623853	17	2.6	36	5.5	2.118
12	D.Carr OAK	OAK	#C4C8CB	5371801	1342950.2	4	16	986	0.957	593	146	347	246	58.5	3254	5.487352	12	2.0	22	3.7	1.833
13	E.Manning NYG	NYG	#192E6C	97500000	16250000.0	6	16	1109	0.983	600	200	378	222	63.0	4439	7.398333	14	2.3	32	5.3	2.286
14	G.Smith NYJ	NYJ	#313F36	5019603	1254900.8	4	14	816	0.750	367	114	220	147	59.9	2549	6.945504	13	3.5	17	4.6	1.308
15	J.Cutler CHI	CHI	#DF6108	126700000	18100000.0	7	15	969	0.915	565	164	373	192	66.0	3841	6.798230	19	3.4	30	5.3	1.579
16	J.Flacco BAL	BAL	#240365	120600000	20100000.0	6	16	1070	0.994	548	175	344	204	62.8	3989	7.279197	12	2.2	27	4.9	2.250
17	J.Locker TEN	TEN	#4C96C4	12586002	3146500.5	4	7	299	0.310	146	50	86	60	58.9	999	6.842466	7	4.8	6	4.1	0.857
18	J.McCown TB	TB	#B20032	10000000	5000000.0	2	11	630	0.634	323	94	183	140	56.7	2214	6.854489	14	4.3	12	3.7	0.857
19	K.Cousins WAS	WAS	#7A2D39	2572688	643172.0	4	6	357	0.339	203	70	125	78	61.6	1698	8.364532	9	4.4	11	5.4	1.222
20	K.Orton BUF	BUF	#0F4589	11000000	5500000.0	2	12	803	0.756	446	134	288	158	64.6	3032	6.798206	11	2.5	19	4.3	1.727
21	M.Ryan ATL	ATL	#C9233F	103750000	20750000.0	5	16	1064	0.984	624	202	416	208	66.7	4738	7.592949	16	2.6	32	5.1	2.000
22	M.Stafford DET	DET	#006DB0	53000000	17666666.7	3	16	1093	1.000	601	194	364	237	60.6	4280	7.121464	12	2.0	23	3.8	1.917
23	N.Foles PHI	PHI	#003B48	2768520	692130.0	4	8	545	0.463	312	93	187	125	59.9	2212	7.089744	11	3.5	14	4.5	1.273
24	P.Manning DEN	DEN	#DF6108	96000000	19200000.0	5	16	1092	0.968	598	201	397	201	66.4	4745	7.934783	15	2.5	40	6.7	2.667
25	P.Rivers SD	SD	#08264D	91800000	15300000.0	6	16	1052	0.986	568	190	379	189	66.7	4325	7.614437	18	3.2	34	6.0	1.889
26	R.Fitzpatrick HOU	HOU	#B20032	7250000	3625000.0	2	12	727	0.656	311	111	197	114	63.3	2498	8.032154	8	2.6	18	5.8	2.250
27	R.Tannehill MIA	MIA	#005E6A	12668843	3167210.8	4	16	1065	0.973	589	199	391	198	66.4	4046	6.869270	12	2.0	27	4.6	2.250
28	R.Wilson SEA	SEA	#548A4C	2996702	749175.5	4	16	1054	0.996	452	139	285	167	63.1	3502	7.747788	8	1.8	20	4.4	2.500
29	S.Hill STL	STL	#0D254C	1750000	1750000.0	1	8	452	0.430	229	75	145	84	63.3	1657	7.235808	7	3.1	9	3.9	1.286
30	T.Brady NE	NE	#0D254C	70600000	14120000.0	5	16	1062	0.939	584	188	375	209	64.2	4165	7.131849	9	1.5	36	6.2	4.000
31	T.Bridgewater MIN	MIN	#380160	6849502	1712375.5	4	13	794	0.776	400	126	262	138	65.5	2973	7.432500	12	3.0	17	4.2	1.417
32	T.Romo DAL	DAL	#0D254C	108000000	18000000.0	6	15	971	0.917	435	152	303	132	69.7	3748	8.616092	9	2.1	37	8.5	4.111

Figure 4: Output of pbp2014 data with data manipulation

Figure 4 above is an example of how great dplyr can be; it was made by using the pbp2014 data frame which contained over 45000 row and has been transformed into a data frame with less than 32 rows by filter(), mutate() & Summarise() functions in dplyr. To create this output, started with gathering the list of starting quarterbacks for the 2014 season turning it into a vector for R to later filter out players that might have only thrown the ball once or twice the entire season. Next, the team tags of each team were added to each player for graphing purposes.

```
#Taking all guys who have thrown a ball 2014 and add their team tag to
their names for identification
passers <-
as.character(unique(paste(subset(pbp2014,PlayType=="Pass")$Passer,
subset(pbp2014,PlayType=="Pass")$posteam,sep=" ")))
```

Next was filtering out the real starting Quarterbacks based on the vector I created earlier called 'Starters'. It will then return 32 Passers with their team abbreviation attached.

```
#identify, all the starting QBs within the pass-throwers, Using the
above list of players to identify QB's
starters.team <- passers[grep(paste(starters,collapse="|"),passers)]
```

Colour schemes of each team were then added to make more visual graphs later, Colour schemes were bounded with the newly created team tags which extracted the 32 unique teams from the possession team column in the pbp2014 dataset this was done so I could then merge pbp2014 with colour schemes of each team as colour schemes on their own had no identity to any team it just represented by code i.e. #97233F.

The creation of the table output was completed by using the various functions dplyr offers.

```
#Creating a data frame of seasonal stats for QB's
QB.Rating <- pbp2014 %>%
  filter(PlayType=="Pass") %>%
  mutate(QB.team = paste(Passer,posteam,sep=" ")) %>%
  group_by(QB.team, posteam, posteamcolour) %>%
  summarise(Games.Played = n_distinct(GameID),
            Total.Passes = n(),
            First.Downs = sum(FirstDown),
            Completed = length(PassOutcome[PassOutcome=="Complete"]),
            Incomplete = length(PassOutcome[PassOutcome=="Incomplete
Pass"]),
            Completion.Rate = round(Completed/Total.Passes,3)*100,
            Total.Yards = sum(Yards.Gained),
            Yards.per.Att = sum(Yards.Gained)/Total.Passes,
            Interceptions = sum(InterceptionThrown),
            Interceptions.per.att = round(Total.Passes/Interceptions),
            TDs = sum(Touchdown),
            TDs.per.att = round(Total.Passes/TDs),
            TDs.per.INT = round(TDs/Interceptions,3),
            Fumbles = sum(Fumble))%>%
  filter(QB.team %in% starters.team) %>%
  arrange(posteam,-Completion.Rate, -Total.Passes)
```

This code represents the beginning of creating figure 4, it takes the data from the pbp2014 dataframe and combines it to make a summary of the Quarterbacks season which I can then use for machine learning algorithms later on. It starts by creating a new dataframe called QB.Rating then moves on to filtering that data by play type = Pass, in other words means its told to only look at rows were PlayType contains the word ‘Pass’. Mutate is a function that allows users to create new variables based on present variables while keeping the existing variables, I decided to add Quarterbacks names with their respected team name so that I could

use the object I had created earlier called starters.team which if you recall combined starters and their respected teams which could be then used to further filter the output to show a more meaningful result to customers and viewers alike. This was achieved after creating the variables QB.team.

```
filter(QB.team %in% starters.team)
```

The summarise function is the function that allows us to compute the seasonal statistics of each Quarterback from the pbp2014 dataset.
An example of this being used is;

```
summarise(Games.Played = n_distinct(GameID),
```

The above line of code creates a variable called Game.Played by using the n_distinct command which checks how many times an individual name appeared beside a unique GameID and returning the games played by the player.

To create a complete seasonal summary of each Quarterback it was necessary to add the contract details of each player contract details and SnapCount of each to see if it could be helpful in building a model later. After importing both newly scraped files, some necessary cleaning steps were taken in both Excel and R. As described above the problem of player's names not matching happened on this occurrence which was quickly resolved to allow the continuation with the merge. Merging is then completed by finding a common variable in the two underlying datasets the user intends to put together, as for these two datasets the common variables were the newly repaired player names in the format of 'A.Rodgers'.

```
#Merging two datasets together to form a new data frame
premerge <- merge(Contracts, SnapCount, by.x = "player", by.y =
"Player", all = F)
```

Merging is straightforward and efficient once the underlying datasets have been duly prepared. The newly merged data frame was then given new columns through the mutate function so that they could then merge 'premerge' with 'QB.Rating' on the QB.team variable. After successfully merging the two data frames together it was possible to conduct a find & replace on the variable 'contract length' to remove 'years' from all the rows so in return the contract length variable could be changed

to a numeric value for use in an analysis later. Finally, the data was reordered to make viewing easier on the eye when searching.

```
#Reordering
QBComplete <- QBComplete[,c(1:3,17,22,18,4,20,21,5:16)]
```

This line of code is simply rearranging variables from their original position to a position more suited towards the information they provide. The above code snippet is the final part of cleaning as the code was reused for other potential rich datasets to help in the process of building effective algorithms.

2.3.5 Analysis

With the data transformed into a more meaningful display with all the relevant variables shown in one data frame the raw data could now be viewed in different ways to produce graphs that have a specific meaning behind them and explain to potential customers the value of replacing ‘player A’ with ‘player C'. Also, with the data prepared its possible to perform statistical tests as the players are be grouped based on features. The analysis was carried out to check if the data meet the required criteria for the use of data mining techniques intended.

2.3.6 Data Mining Approach

During this phase, we can use the newly cleaned and transformed data to perform data mining tasks to retrieve previously unknown information from the datasets. The data mining approaches undertaken for this project are clustering and regression. Clustering is the process of making a group of abstract objects into classes of similar objects (Anon, 2017), the idea behind this is objects can be grouped together into the different tiers & types of quarterbacks using an array of variables allowing users and customers to find similar individuals to the player they are searching. The latter data mining approach explored was a linear regression model. Regression is using independent variables (e.g. Sleep, Study, Diet) to predict one dependent variable (e.g. Test score). For regression to be successful the variables must pass a series of tests/assumptions, the series of trials were carried out in SPSS/R to test if there was a relationship between the independent

variables and the dependent variable and to test if there is multicollinearity between the model building independent variables. Multiple iterations of multiple linear regression were carried out to find the best model by combing datasets together to test if external variables such as physical attributes increased the accuracy of the model. The goal of the project was to predict the yearly salary of a player to assess the return on investment in players to an organisation, the methodology behind the idea was that yearly salary could be used as the dependent variable and season statistics would be utilized as the independent variables to predict if the return on investment was positive or negative. The implementation of multiple linear regression started by filtering the data into a sample, the sample taken was the Quarterback (QB's) performance over the season as shown earlier the seasonal statistics of the QB's were compiled through vigorous cleaning and transformation giving I and users alike an overview of the season the QB had. After the merging of contracts and seasonal stats had been complete, it was time for testing to take place on the underlying dataset to fit the right fit for the model building process. The assumptions the chosen variables must pass are:

- Linearity of residuals: Data must follow a linear pattern
- Independence of Errors: If the data points are showing a pattern, it would indicate that the errors are influencing each other i.e. Linear regression cannot be run if the data does not have a random error.
- Normal distribution of residuals: Residuals must not be skewed heavily
- Multicollinearity: independent variables cannot be highly correlated with each other (<±0.8)

Once the assumptions have been met the building of the model can commence through R. Checking for linearity involved creating multiple scatterplots in SPSS to find if they did follow a linear pattern followed by with the use of Linear Regression Scatterplots function. The first model was tested and built using all the available features; it failed due to the presence of multicollinearity between many of the independent variables. Multicollinearity tells the user once high correlation is

present between variables we must remove one of the independent variables that have shown multicollinearity, in this case, the correlation between many of the variables meant there were very limited variables left to choose. The 2nd attempt at building the model included outside data from Quarterback rushing stats, as some Quarterbacks are more athletic than others they can use their speed and quickness to gain yards when a passing opportunity is not available at the expense of becoming prone to hits from defensive players thus limiting their career span and long-term value to an organisation with an average career span for Quarterbacks being 6.5 Years, Almost 1 year above the mean of other positions.

```
#Finding average career length of QB's
qbdraft <- draft%>%
  filter(Position.Standard == "QB")
qbdraft <-as.data.frame( qbdraft$To - qbdraft$Year)
qbdraft <-qbdraft  %>%
filter(`qbdraft$To - qbdraft$Year` > 0)
#Summary of QB Career Span
summary(qbdraft)
```

It is vital that Quarterbacks stay protected by their offensive line and not venture into the open field although it can pay dividends at times, the opposite type of Quarterbacks known as 'Pocket Passing Quarterbacks' have a strong history of success (Discussed later during clustering). With outside factors such as Rush yards per game now included in the model building process, we can try the 2nd iteration of the model build. The data passed the assumptions but seen the new outside factors as negative coefficients meaning there is statistical evidence of a negative relationship between the variables thus devaluing quarterbacks who tend to rush for yardage, depending on the contract length this could be a positive or negative result for the basis of this project I'm investigating the short-term return on investment and not looking at the long-term contract effect. This model build can be considered a 2nd failed attempt at building a regression model with inclusion of outside variables.

The 3rd and final regression model was built using the following Dependent/Independent variables:

Table 6:Variables Used In Model Building

Dependent Variable	Explanation
Yearly	Overall Salary of an individual player for that year, Divided total contract by number of years

Independent Variables	Explanation
Pass Completion Percentage	Higher % the better
First Down Per Game	Higher value is better, indicates QB can keep his team on the field
Touchdown To Interception Ratio	Higher value the better, Touchdown is offensive team scoring, Interception is throwing the ball to a defensive team player

These variables were chosen as they didn't penalize players who didn't play a full season due to injury, if the total sum of these variables were chosen the model would be compromised due to the fact the difference between passing leader and the player who had only played a few games that season would skew the model results leaving an inaccurate model. Testing in SPSS was conducted with the Correlation table that tells us how correlated the variables are to each to each other.

Table 7: Strengths Of Correlations

Correlations	Strength
±0.0 To 0.2	Weak or no relationship
±0.2 To 0.4	Weak relationship
±0.4 To 0.6	Moderate relationship
±0.6 To 0.8	Strong relationship
±0.8 To 1.0	Very strong relationship

As no variables had a correlation relationship of >±0.8 we can rule out Multicollinearity among the variables. ANOVA table was set up to find if there was statistical significant difference between the dependent and independent variable, other important statistical results included Coefficients which will be used in the

multiple linear regression (MLR) formula to predict the dependent variable yearly salary, the formula for MLR is:

$$y = a + bx1 + bx2\ldots$$

Where;

- x1 = value (Coefficients) of 1st independent variable
- x2 = value of 2nd independent variable
- b = the slope of the regression line
- a = the y-intercept of the regression line

Lastly, histograms and scatter plots were implemented to check for skewness and linearity about the inputted variables.

Once the variables had passed the assumptions set out it was time to setup the model environment in R and test the model. First thing was to rename the row numbers to the player's name so we can compare the predicted and actual values. Next was to remove columns not needed for this model build. Once data frame model was setup we can then use the set.seed function in R, this feature allows users to reproduce the same results anytime they please. In order to test the model being produced we need actual results to test it against, common practice in building machine learning models is to split the data into two parts 80% for training the model and 20% to test it later on.

```
#Create Training and Test data -
set.seed(3) # setting seed to reproduce results of random sampling
#row indices for training data
trainingRowIndex <- sample(1:nrow(QBModel), 0.8*nrow(QBModel))
trainingData <- QBModel[trainingRowIndex, ] #Model training data
testData  <- QBModel[-trainingRowIndex, ] #Test data
```

Next was to create the model using R linear model function lm(), this is where we input our dependent and independent variables and then use the summary(model) to call the model and can view the information about the model similar to the statistics SPSS can provide us with.

```
#Build Model to predict Yearly salary based on training data
```

```
model <- lm(Yearly ~ ., data = trainingData)
#summary model
summary(model)
```

Now the model has been built and can be viewed in various plot to get an indication of the model quality. Now comes the time to test the model's accuracy, we've trained the model with 80% of the 32 observations leaving 20% to test the accuracy, test accuracy was conducted by using the predict() function and asking R to use our model on the testData created earlier, data frame was then created to display the correlations between the actuals and predicted values which are a form of accuracy, a high accuracy percentage suggests that the actuals and predicted values have similar directional movement.

```
#Predict our model using the 20% Test set
Pred <- predict(model, testData)
actuals_preds <- data.frame(cbind(actuals=testData$Yearly,
predicteds=Pred))  # make actuals_predicteds data frame.
```

With the model created and setup to predict a test set its time to validate the model accuracy and if the model provides a good fit. For validation purposes, this project used K-Folds Cross-validation. When the data is split into a train and test set the train set loses some element of training as it can't use the test set to make a stronger model but with K-Folds this weaknesses is taken care of by dividing the full dataset into selected amount of K sets hence K-Folds, for example when the entire dataset contains 32 data points with K-Fold you split the 32 data points into selected number of bins e.g. 8 with each bin now contains 4 data points which will act as a train and test set each iteration the K-Fold validation is computed giving the model more chance to be successful with iterating train sets.

```
#K-Fold Cross Validation
install.packages("DAAG")
library(DAAG)
cvResults <- suppressWarnings(CVlm(data = QBModel, form.lm = Yearly ~ .,
m=8, dots=FALSE, seed=3, legend.pos="topleft",  printit=FALSE))
attr(cvResults, 'ms')
```

The 2nd data mining approach developed over the course of this project was an unsupervised clustering algorithm. Unsupervised learning is type of machine learning where you have a dataset with inputted data with no response label e.g. We have a dataset that contains the information about cars but the data set doesn't tell us the brand of the car, we can you use clustering to group together similar kinds of cars without knowing the response variable (Car name). For this project, a type of clustering called Hierarchical Clustering was used to assign each observation its own cluster and then computing the distance (similarity) between each of the clusters and joining the two most similar clusters. The steps are then repeated until there is only a single cluster left. The objective of this clustering technique is to find the most similar players based on annual statistics but also their physical attributes to truly find the most similar players around the league. The implementation of the clustering began by using an annual summary of players to group them into different cluster segments, within the NFL there are four different types of quarterbacks:

Table 8: Quarterback Types

QB Types	Explanation
Pocket Passers	**Smart breed, like to stay in the pocket and read defenses to find weaknesses**
Dual-threat	**Athletic QB's that can throw and run at good/elite levels can run like Running Backs but also throw like top tier QB's**
Athletic QB's	**Like pocket passers but can operate at an average level and keep defenses guessing in their play calling**
Run First	**Built like running backs and have inferior passing qualities**

With this knowledge, we can tell our clustering algorithm we would like 4 clusters to be created to understand which quarterbacks are grouped closely together. On

the first iteration, the clustering algorithm was provided with just the annual passing statistics to create a view of 4 types of quality in quarterbacks. This iteration gave a high-level overview of quarterbacks passing ability considering total yards, touchdowns, and interceptions among others. On the 2nd iteration, players seasonal rushing statistics were added with passing statistics to create improved clustering algorithm which could give customers an insight into what types of players were like each other. Upon the 2nd iteration creation, it was decided to take variables from the combine dataset which contained the physical attributes of players recorded before they are drafted into the NFL. The reasoning for this was that variables like height, weight, 40 Yard dash, the vertical jump could be used as a deciding factor if players were alike. Before creating the 3rd iteration players were removed that had played less than half the season as they could skew the clusters and the outlook of the cluster plots because the clustering algorithm was considered the sum of all yards gained which differ greatly from players that had played close to a full season. After running the algorithm and deciding to rerun a final iteration this time dealing with the NA values in a different manner. As discussed earlier all NA values present in the datasets were changed to 0 where appropriate, but with the combine dataset we're dealing with players physical attributes, it made little sense to change the missing NA to zeros, one idea was to take the average of the variable and use that as replacement metric of missing NA values, but again this could skew the data with some players being more athletic than others it could cause an 'unathletic' player to be forced into a cluster with athletic players. The idea was to use another machine learning algorithm called K-Nearest Neighbours known as KNN. KNN is a classifying algorithm whereby it takes all labeled observations to classify an unknown observation which will be done to replace the NA values. An observation is classified by a majority vote of its neighbors, a distance function called Euclidean Distance is used to compute the difference between the observation being classified and the available case.

$$d = \sqrt{\sum_{i=1} n\,(X2^2 - X1^2)}$$

- X2 & X1 are the examples to be compared
- Each has N attributes
- X2 is the value of the $i_T h$ attribute of instance X2
- X1 is the value of the $i_T h$ attribute of instance X1
- Compare the values of each attribute

With KNN we set a K value which tells the algorithm how many neighbors to consider when making the decision of the classifying the unlabeled case. If K=1 we would assign the unlabelled case to the nearest neighbor. For the implementation of KNN, we will set K=3 meaning the closest neighbors will classify the unlabelled value. 2 cases of unlabelled data appeared in the forty-yard dash variable using Height and vertical as indicators we can estimate the speed of the unknown cases while for vertical there were 7 cases of missing values using Height and forty-yard dash time as indicators to classify the missing values.

```
#Install packages
install.packages("VIM")
library(VIM)
#KNN Algorithm k=3, use dataset=cluster4
cluster4<-kNN(cluster4 ,k=3,variable = c("vertical", "fortyyd"),
dist_var = c("fortyyd", "vertical", "HeightInCM"))
```

When the missing values are taken care, the clustering algorithm could know complete the 4th and final iteration plotting many dendrograms plots to show the similarities between players. A function was used to an array of colors to the dendrograms plots to help make an interference about the players; the colors are assigned to the players based on the 4 clusters segments set up earlier to show the different types of quarterbacks available in the league. The tree was then pruned using various R functions to improve the appearance of the dendrograms. Pruning was the last part of implementation before the interpretation and visualization of the data through Tableau and RStudio.

2.3.7 Visualization

At this stage of the implementation, it was about showing how the raw data of over 45,000 rows was transformed into meaningful data to gain insights and discover new knowledge previously unknown. Tableau was the main source of visualization

with different types of plots and visuals used to represent the transformed data and algorithms. Since all the datasets are being hosted on a MySQL database Tableau allows the connection to hosted databases with the use of username and password. After selecting the database tables to import there were some changes needed to be done with the data type i.e. Changing any numerical data from strings to numeric and converting the necessary variables to measures, to produce the visuals. The created visuals are displayed in the results section below.

2.4 Testing

Throughout the project, testing was completed to ensure the code was executing working without errors or bugs. The testing methodologies used were:

White Box Testing: Testing technique that helps developers reason carefully with implementation, White box testing examines and delivers test data from the program code

Black Box Testing: Quite the opposite of white box testing, it is method testing that inspects the functionality of an application based on its specifications. The tester should only know how the system will function not how the actual program code. Test scripts will be formatted in a table structure as follows:

Table 9: Test Case Format

Heading	Description
Name	**Name of application under test**
Date of Test	**Date test case was conducted**
Test ID	**unique number to identify each test case**
Purpose of Test	**Short description explaining what the test case is covering**
Test steps	**How the test was performed**
Expected results	**Brief explanation of the desired result**

Actual results	**Brief explanation of the real results of the code being executed**
Action Required	**If tests fail, what action should be taken**
Resolution	**How was the failed test resolved**

White Box Test #1			
Name	Data Cleaning	**Date of Test**	28/01/2017
Test ID	WB#1	**Iteration ID**	1.0

Purpose Of Test	**To Guarantee that:** • **The database connection is open** • **Confirm that database tables can be populated into R as data frame**
Test Steps	**Within RStudio the tester should:** • **Ensure correct database details are entered** • **Run script containing MySQL import function**
Expected Result	**Dataframe named "Final" will be shown in the global environment in RStudio with 47,000+ rows**
Actual Result	**Dataframe named 'Final' was imported from the database with all columns and rows present**
Suggested Action	**N/A**
Resolution	**N/A**

White Box Test #2			
Name	Data Algorithms	**Date of Test**	20/03/2017
Test ID	WB#2	**Iteration ID**	1.0

Purpose Of Test	**To Guarantee that:** • **That correct data frames are available** • **Ensure Regression algorithm runs without error**
Test Steps	**Within RStudio the tester should:** • **Ensure correct database details are entered** • **Run full script to import and create data frame required** • **Run script containing Regression model**
Expected Result	**Dataframe 'Final' will be imported and the corresponding data frame will be created. As a result, then regression model will run returning predicted values from test set**
Actual Result	**Dataframe named 'Final' was created from database import, resulting data frame has been set up to run the regression model. Regression model ran and returned predictions with test set**
Suggested Action	**N/A**
Resolution	**N/A**

Black Box Test #1			
Name	Data Export	**Date of Test**	24/03/2017`
Test ID	BBT#1	**Iteration ID**	2.0

Purpose Of Test	**To Guarantee that:** • **That the system exports CSV files**
Test Steps	**Within RStudio the tester should:** • **Ensure data frame is loaded in the Global Environment** • **Run full export script** • **Check folder for new CSV file**
Expected Result	**Dataframe will be exported to corresponding working directory as a CSV file**
Actual Result	**Dataframe was exported to the wrong folder**
Suggested Action	**Change working directory within RStudio to our desired location and run Script again**
Resolution	**CSV file successfully saved to the desired location**

Black Box Test #2			
Name	Data Cleaning	**Date of Test**	15/02/2017`
Test ID	BBT#2	**Iteration ID**	2.0

Purpose Of Test	**To Guarantee that:** • **Data Cleansing occurred** • **Check corresponding data frame was successfully populated with data from the previous dataset**
Test Steps	**Within RStudio the tester should:** • **Ensure data frame is loaded in the Global environment with the use of Data Import script** • **Run Data Cleaning script and check global environment**
Expected Result	**Dataframe will be imported, and NA values will be dealt with and new data frame QB. Rating will be created from the old data frame with summary statistics**
Actual Result	**Dataframe was imported, and scripts ran error free, QB. Rating now available with populated data**
Suggested Action	**N/A**
Resolution	**N/A**

2.5 Customer Testing

Customer testing involved examining the from a customer's perspective that the product/service works as expected. Testing was conducted with Tableau producing the correct graphs without an error occurring before or in the resulting graphs.

Usability Test #1			
Name	Tableau Visual	**Date of Test**	01/05/2017`
Test ID	UT#1	**Iteration ID**	3.0

Purpose Of Test	**To Guarantee that:** • **Exported data is represented correctly in Tableau**
Test Steps	**Within RStudio the tester should:** • **Run data export script** • **Load exported CSV file into Tableau** • **Select Bar Chart option with corresponding variables**
Expected Result	**CSV file will be exported to the folder and Tableau will produce Bar Chart**
Actual Result	**CSV successfully exported but Tableau failed to create bar chart as all variables were imported as characters**
Suggested Action	
Resolution	**N/A**

3 Results

3.1 Experiment 1: Visualizations

Visualizations were completed in Tableau due to their resolution and superior quality to that of RStudio ggplot package.

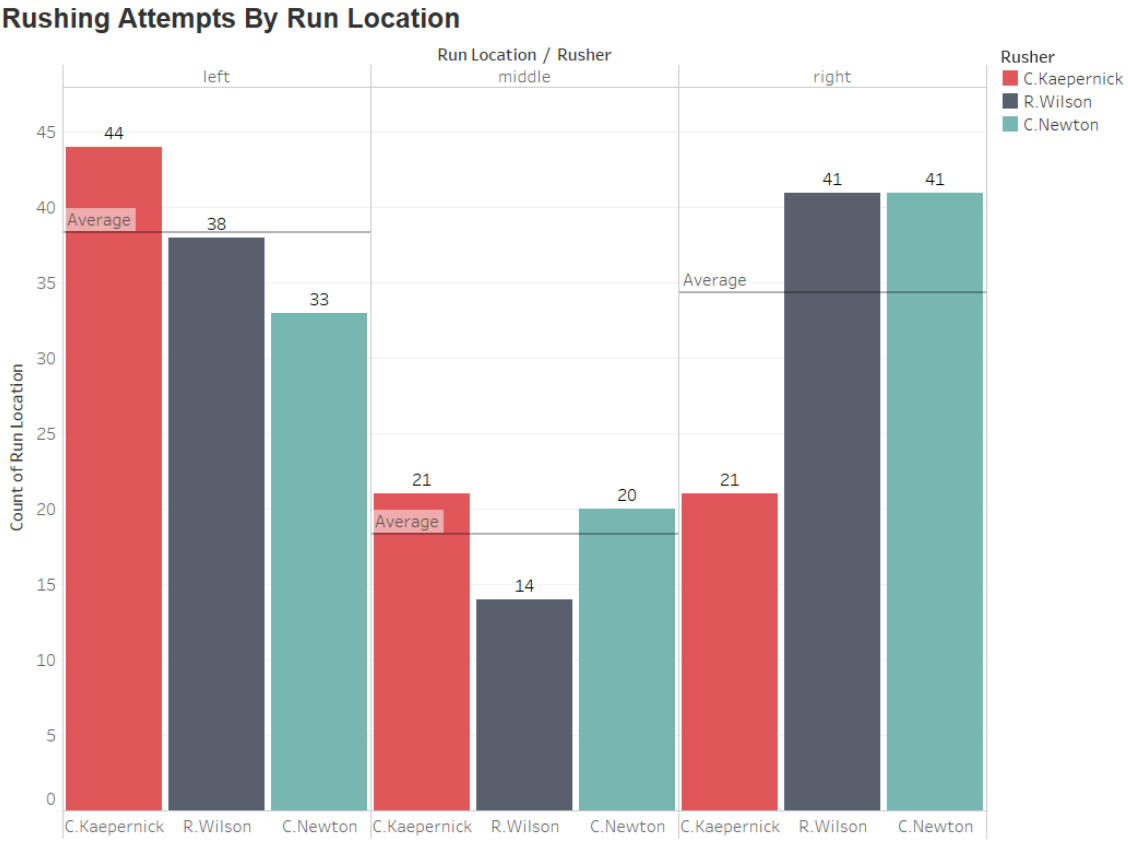


Figure 5: Rushing Attempts By Run Location

The first visual represents the running location of three different Quarterbacks who will be discussed in the clustering experiment later in the results section. It is a simple visual to introduce the user and explain what they represent before moving onto more advanced visuals. From the above bar chart, we can see the running location tendencies of Dual-threat Quarterbacks. One inference to be made is the direction C. Kaepernick likes to go in, Kaepernick primaries runs to the left side of the defense and unlike the other two players. The 2nd conclusion to take away from the bar chart is the lack of runs completed through the middle of the field, running up the middle is a weakness that could be exploited by these players as defenses

would not expect the most precious player on the team to run through traffic for a potential small gain.

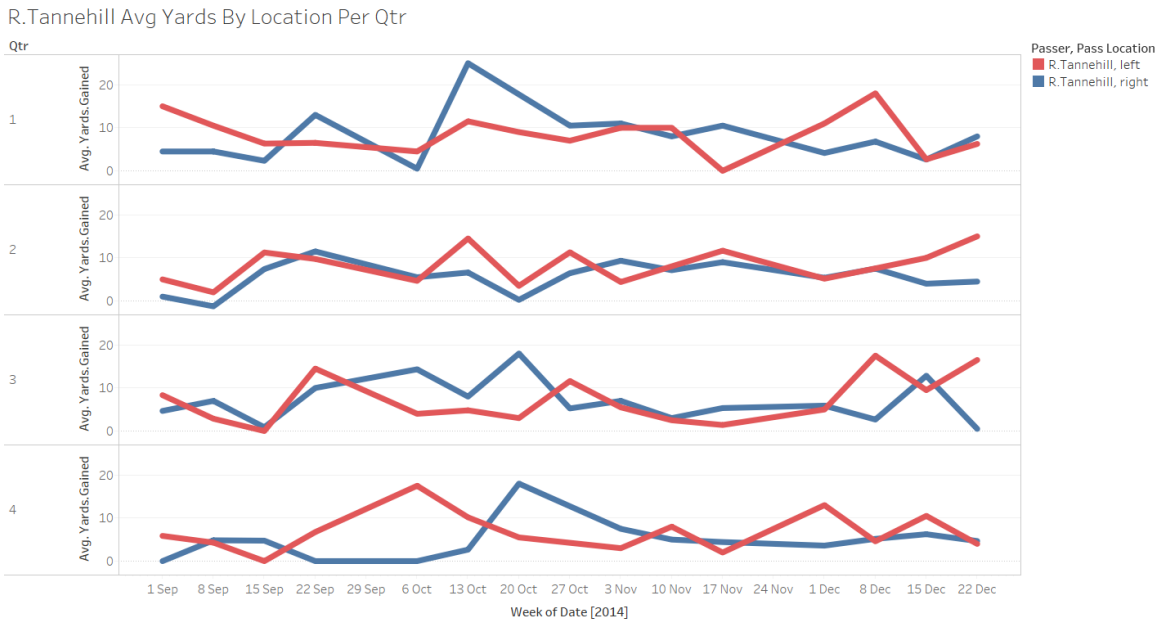


Figure 6: R.Tannehill Average Yards By Location Per Quarter

This trend line graph exhibits the use of features within the dataset to explore capabilities of the visualizations, here we had Ryan Tannehill passing averages when throwing right (Blue Color) and left (Red Color) split by the 4 Quarters an NFL games is played in. The X-Axis (Bottom axis) is the game weeks Tannehill played. On examination, we can tell that the distributions are even except the high averages exhibited in the middle to late October where passing right achieved a significantly greater average in the 1st & 4th Quarter than passing to the left-hand side of the field. Ryan Tannehill will be discussed in the regression section following.

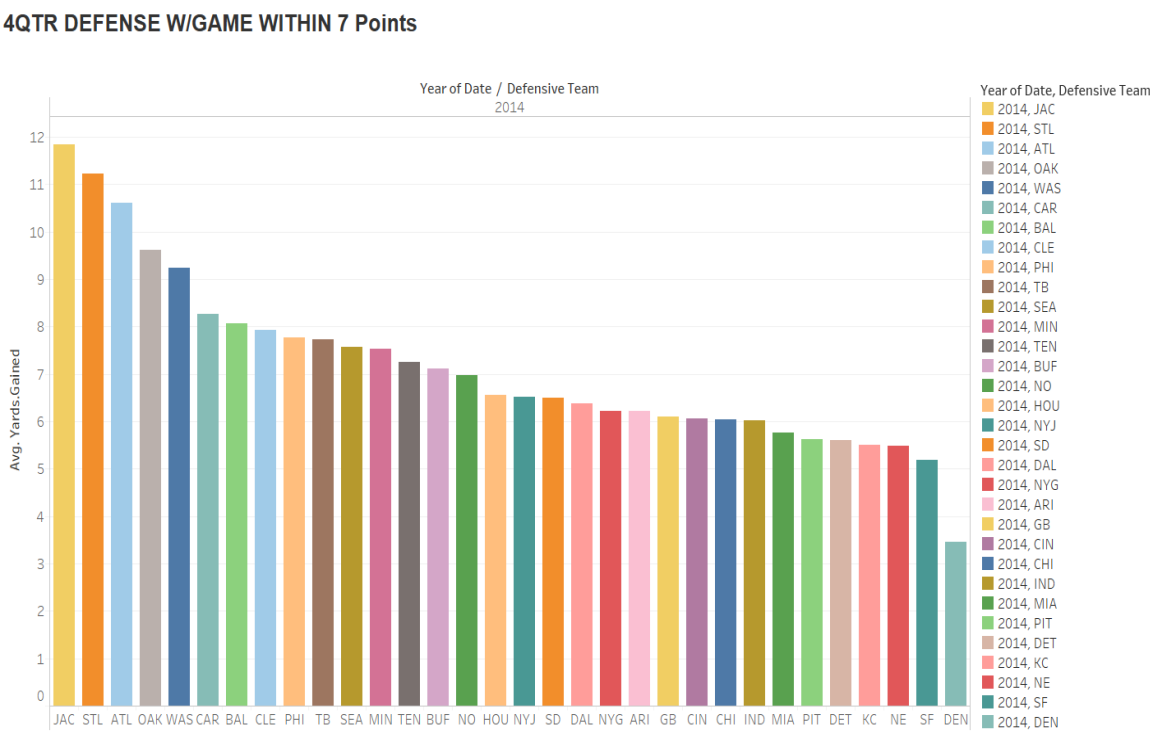


Figure 7: 4th Quarter Defense With The Game Within In One Score

The final graph is another example of the use of the features in the dataset and how they can be conveyed to a user. It describes the resilient defenses in NFL when the game is on the line. 3rd best defense and eventual Super Bowl champions of 2014 New England Patriots (NE) show why they deserved to be the champions with a strong defensive showing when the game mattered most. Their Super Bowl finalist opponents Seattle Seahawks (SEA) sit just left of the center which is ironic in hindsight of the 10 point 4th Quarter comeback led by Tom Brady & Patriots in the Super Bowl.

3.2 Experiment 2: Regression

In this following section, the results of the project are discussed regarding the material covered in the implementation and the resulting work; this section will also set forward the hypotheses used to build and test models. The results in the following section will show the final model created by the machine learning algorithms with output from both SPSS and RStudio

The regression model set up with the use of SPSS and RStudio. Using our reduced dataset, it was possible to run the CSV file through SPSS to gain some basic information about the underlying data.

Table 10: Descriptive Statistics SPSS

Descriptive Statistics

	Mean	Std. Deviation	N
Yearly	10325439.846354168	7865259.0497866980	32
First.Downs.PerGame	10.798613	1.6157245	32
Completion.Rate	63.022	3.5947	32
TDs.per.INT	2.22984	1.354524	32

The descriptive graph above tells the user a little about the variables that will be used in the model. It gives the mean which is the average of each variable, the Standard deviation of each variable. Standard Deviation is a measure of how spread out the numbers in the data are; it tells us what data is considered small, normal or large by ± the Std.Deviation value from the mean to the actual individual data point e.g. Aaron Rodgers Yearly contract is $22Millon, With the average at just over $10Million we can say that Rodgers Is within two standard deviations of the mean which is quite large. N is the number of observations contained within the dataset in this case the 32 NFL team starting Quarterbacks.

Hypothesis was set out to test if there was relationship between the independent variables and dependent variables

H0: p = 0

There is no relationship between the independent variables (First Downs.PerGame, Completion.Rate, TDs.per.INT) and the dependent variable (Yearly)

H1: p ≠ 0

There is a relationship between the independent variables (First Downs.PerGame, Completion.Rate, TDs.per.INT) and the dependent variable (Yearly)

The hypothesis was set out to test if there was a relationship between the independent variables and dependent variables, the test is conducted at an alpha value of 0.05 meaning this test has an accuracy level of 95%.

The Correlation graph explains the relationship between Variable A & B; a positive correlation says that as one variable increase so do the other variables and a negative correlation says as one variable decreases the other variables increase. The correlation matrix is also crucial for the reason of spotting Multicollinearity between variables; Multicollinearity is a term used if two of our variables are highly correlated (r>±0.8) it can skew the results of the model build. If Multicollinearity was to be found its recommended to remove one of the variables from the model.

Table 11: Correlation Matrix SPSS

Correlations

		Yearly	First.Downs.PerGame	Completion.Rate	TDs.per.INT
Pearson Correlation	Yearly	1.000	.471	.559	.520
	First.Downs.PerGame	.471	1.000	.514	.301
	Completion.Rate	.559	.514	1.000	.477
	TDs.per.INT	.520	.301	.477	1.000

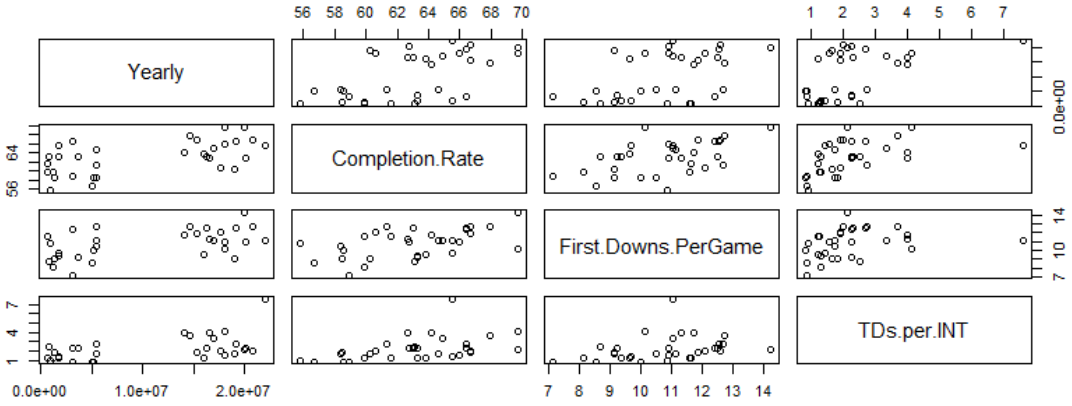


Figure 8: Correlation Matrix Visual

The test conducted was to determine whether there is a relationship between the independent variables (First Downs.PerGame, Completion.Rate, TDs.per.INT) and the dependent variable (Yearly). The test was carried at an alpha value of 0.05 which means the test has a 95% accuracy rate. The test was conducted

primarily in R with SPSS being as a reference point for similar results and tables. The results from the correlation matrix show that there was a medium correlation between Frist Downs.PerGame & Completion.Rate (0.514), the low correlation found between Completion.Rate & TDs.Per.INT (0.477), the low correlation also observed between First Downs.PerGame & TDs.per.INT (0.301). Because of the low to medium correlations, it has been confirmed that no multicollinearity has been found allowing all variables inputted to be used in this iteration of the model build. Turing our intentions to the R output of the Regression model.

The summary is that of the linear model built; it provides information that can be used to assert if there is a relationship between the independent and dependent variables. Focusing on the yellow highlighted column, we can see that:

Table 12: Linear Regression Model Summary

```
Residuals:
     Min       1Q   Median       3Q      Max
-8921483 -2859318   292864  2375190  9075647

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)         -60621431   17757324  -3.414  0.00261 **
Completion.Rate        684435     316820   2.160  0.04246 *
First.Downs.PerGame   2293672     716812   3.200  0.00431 **
TDs.per.INT           1459305     744665   1.960  0.06344 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4805000 on 21 degrees of freedom
Multiple R-squared:  0.6619,   Adjusted R-squared:  0.6136
F-statistic:  13.7 on 3 and 21 DF,  p-value: 3.572e-05
```

- Completion.Rate P-Value is < than 0.05, Indicates that there is a relationship between Completion.Rate and the dependent variable
- First Downs.PerGame P-Value is < than 0.05, Indicates that there is a relationship between First Downs.PerGame and the dependent variable
- TD.per.INT P-Value is > than 0.05, Although the value is higher than the P-Value, it is very close to the alpha value (0.05), and when it was taken out the Adjusted R-Squared value was lowered, this can be considered

a tradeoff and TD.per.INT will be kept as an independent variable as a result.

The above results are enough evidence to reject our null hypothesis that there is no relationship between the independent variables and dependent variables in favor of the alternative hypothesis that states there is a relationship between the independent variables and dependent variables.

R's Summary of the model also provides other useful information on the quality of the model, The summary() function returns the Residuals, residuals are the difference between the actual and predicted values, it assess how well the model was fitted to the data although Median of 0 is considered a perfect fit for the model Rookie contracts played a big part as they're considerably lower than experienced players so as a result, it influenced the high residual values. The model also returns the coefficients which are used in the Multiple linear regression models later with Intercept being 'a' Value and the corresponding independent variables being the 'x' values

$$y = a + bx1 + bx2\ldots$$

The summary() function provides with the Adjusted R-squared value which gives an indication of how well the model explains the values of the dependent variable; it helps measure the linear relationship between the predictor variables and the response variable. With an R-Squared value of 0.6136, the model has said that 61.4% of the variance found in the response variable can be explained by the predictor variables inputted. The reported value of 0.614 can be considered a good R-Squared value with the field of study the project is focused on as it is hard to define what influences the Quarterback value.

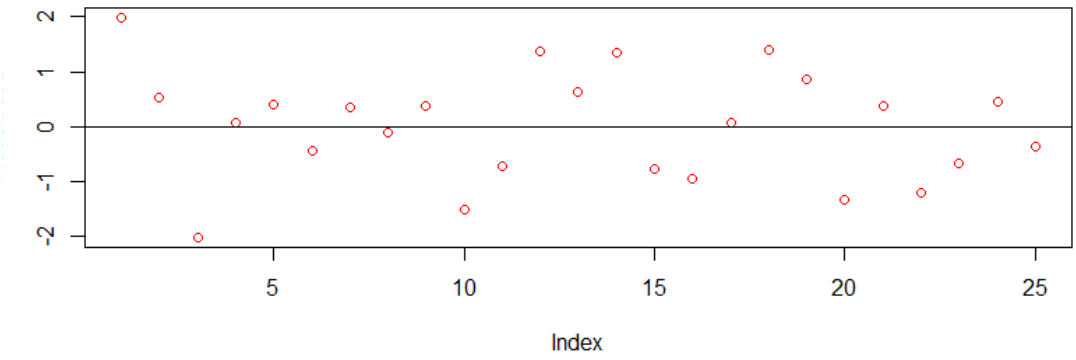


Figure 9: Residual Plot

This above diagram is known as a residual plot; residuals are the difference between the observed value of the dependent value and predicted value. It is a good indicator of the accuracy of the model as 100% of the data points lies within two standard deviations of the mean.

The predicted results vs. the actuals shown in Figure 10 below were quite surprising with the test set, players showed signs of being undervalued and overvalued as expected, when I cross referenced the test data players with the draft profiles of players it was evident that the linear model showed preference in valuing players on rookie contracts, a big reason for this was the vast majority of players in the training set would be experienced pros who have a higher salary therefore the test data worked very well on evaluating the ROI if a player was on a rookie contract in the instance they are valued as experienced pros in other words teams were getting their return on investment, hence the high difference for Ryan Tannehill who showed a difference of +$13Million giving a ROI of 425%. The model showed a tendency to scrutinized overvalued players with Colin Kaepernick being overvalued by more than $15Milion

	actuals	predicteds	ROI.Pct
B.Roethlisberger PIT	14664417	20447291	39.4
C.Kaepernick SF	19000000	4012415	-78.9
J.Flacco BAL	20100000	10731557	-46.6
J.Locker TEN	3146501	-2674217	-185.0
K.Cousins WAS	643172	10082538	1467.6
N.Foles PHI	692130	8897854	1185.6
R.Tannehill MIA	3167211	16636031	425.3

Figure 10: Test Set Predictions Using Final Model W/ROI as a percentage

Testing the final model was done by using K-Fold Cross-validation, K-Fold will split the data into eight separate training and test sets so that all the information can be used to train and test the model over the eight iterations.

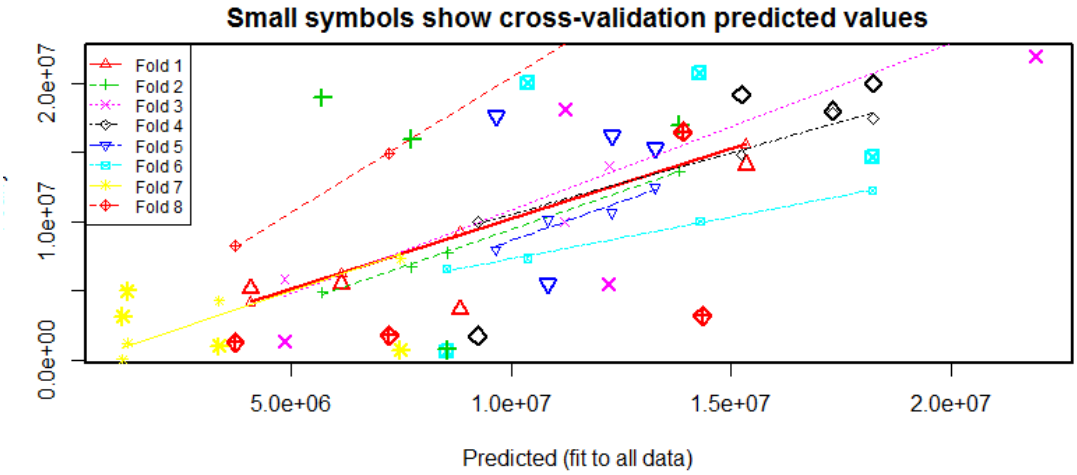


Figure 11: K-Fold Cross Validation

The K-Folds Cross Validation plot is a visual indication of the accuracy, with 32 observations in the full dataset I split the data into 8 Bins (4 Observations in each) hence the eight different dashed lines. The larger symbols represent the actual values while, the smaller symbols represent the predicted value of each fold created. The indication of a good model is when the dashed lines are parallel to each other, and the larger symbols are closer to the line. Although the plot doesn't exhibit a perfect model it must be noted that it's a complex build for regression with the mixture of players on rookie contracts outperforming experienced players which would have an impact on the model, in my view this doesn't show a poor

model as it has also demonstrated the return on investment in players is clear to see in Figure 7 with Ryan Tannehill having a positive impact on his team with a relatively low risk, it's also to be noted that Ryan Tannehill signed a new contract in 2017 (2 Seasons later) with a base salary of $17,975,000 and Colin Kaepernick has since left his franchise and is unable to find a new franchise to take a risk on him indicating a decline in the player which the model also predicted many pundits and analysts described the contract as robbery since he has left. In summary, the model did not prove poor or inaccurate as the level of ROI was measured using the player's statistics as an evaluation metric.

3.3 Experiment 3: Hierarchical Clustering

Clustering by definition is an unsupervised learning algorithm; testing does not usually occur in clustering as there is nothing to test, the goal is to cluster data points based on similarity. With a multitude of datasets, using data sourced over the duration of the project they were combined to achieve a representation of similar player types. The supervised learning algorithm KNN was used within this algorithm to replace missing players combine stats as some players do not take part in all activities. The number of missing values were small, so it did not have much effect on the outcome of the clustering algorithm.

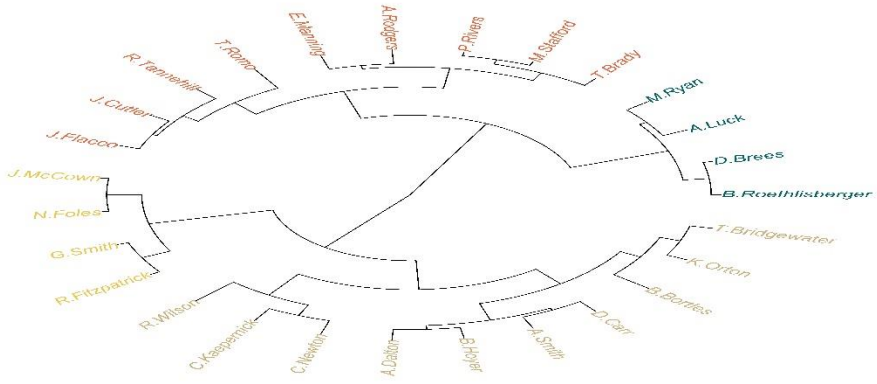


Figure 12: Dendrogram Type: Fan

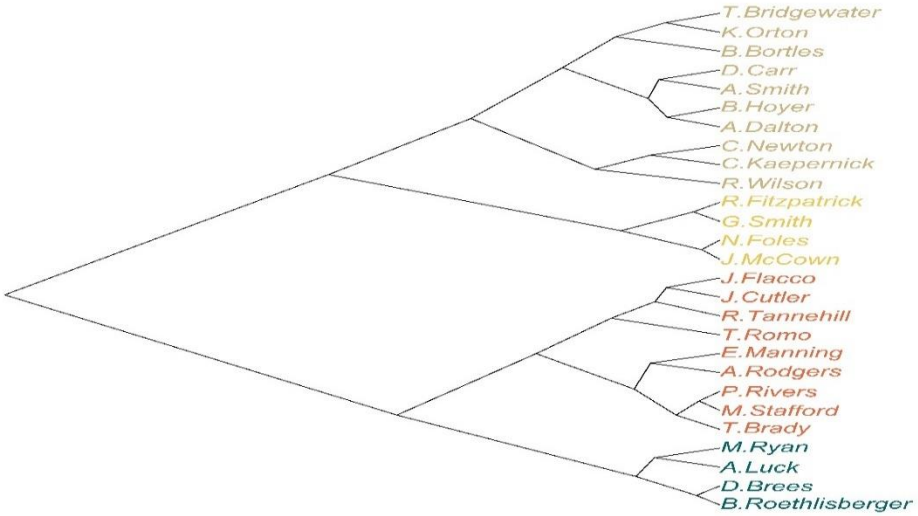


Figure 13: Dendrogram

The above figure is known as a dendrogram, it groups data based on similarity, with a combination of statistics from a player season & their physical attributes it was produced to show what players are alike in the NFL, as mentioned there is no evaluation of the algorithm as it just provides insights for the end user. Using the four quarterback types discussed in the implementation earlier, there were four colors assigned for the visual aid of customers and end users. An unofficial performance measure would be an eyeball test from prior knowledge of the NFL players.

One way of interrupting the above graph is to focus on the sub-clusters, focusing on the cluster trio of R. Wilson, C. Kaepernick & C. Newton (Upper Half)

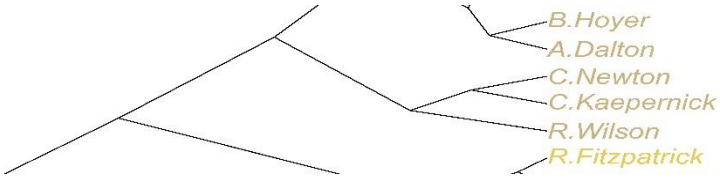


Figure 14: Zoomed in version of Dendrogram

We know that the graph does not tell us anything more than these three players are similar in many aspects of the sport otherwise they would not have been so excluded from the other groups around them. Using a Google search as an indicator of the model similarity measure results in numerous articles about the trio and how they have revolutionized the Quarterback position through their Dual-Threat running ability keeping defenses flustered on whether they will throw, hand the ball off or run for yardage themselves. The following is an extract from ESPN sports journalist Jeffri Chadiha:

“Looking back, so many unpredictable forces aligned to create the hype that surrounded Wilson, Kaepernick, Griffin and Newton. Each player was blessed with a head coach who saw the unique opportunities his QB's mobility presented. Each player also had great timing. Defenses were not familiar with the read-option, and each of these quarterbacks used it to his advantage.

Newton, the No. 1 pick in the 2011 draft, gained more than 700 rushing yards in each of his first two seasons. Griffin produced 815 rushing yards in 2012 as a rookie who sometimes even ran the triple-option. Kaepernick gashed Green Bay for 181 rushing yards in a January 2013 playoff game, while Wilson has been Seattle's second-leading rusher in all three of his seasons. They all provided great highlight” (Chadiha, 2014)

The article is a good indication of the clustering model results as journalists have articles about the trio ability and how they have redefined the position. The mention of the 4th player Robert Griffin who is not in the cluster or results of any nature was due to his season-ending injury early into the season who was replaced with Kirk Cousins his replacement for the analysis.

4 Conclusion

Analytics in the NFL remains some pace off that of other major sports due to its complexity; it has come a long way with the data being made available to the public and being encouraged to find patterns and useful information within the data. With the annual MIT Sloan Sports Analytics Conference showing an increase in NFL related research papers and products in recent yeas although it faces the challenge of recording the data with the playing field so large and so much happening on a single play compared to a basketball possession where there it is five v five on a confined court. Analysis showed that NFL players career only last on average 5 Years equivalent to 1 long term contract meaning finding the player that contribute at the right price is important as the career expectancy is at such a low level. The results of finding the return on investment in players when putting into perspective argued that long-term high-value contracts are a burden on organizations as the physicality of the sport increase the chances that the player the franchise had one year ago will not return the next year to the same standard. In this way, return on investment can be seen as having a major influence on team success with the salary capped at a certain level each season finding players undervalued is crucial more than ever.

4.1 Evaluation

The KDD gave a map like structure to follow which allowed this project to be successful through implementation and results. With the KDD providing the map to follow when conducting a data mining project as insightful as this one the right plan was needed to make sure the project stayed on course and the objectives were achievable. With the previous experience using the KDD to significant effect, it was the number one choice of methodologies to follow over similar data mining approaches such as SEMMA and CRISP-DM. The steps the KDD provided explain in detail small but vital steps to a successful data mining project with steps like picking the right dataset with data types, the importance of dealing with missing values as well as it can affect your machine learning models in many ways which

were shown by using KNN to solve the missing values problem with clustering algorithm.

4.2 Future Work

With the small sample size used for the prediction of return on investment in players. naturally, the next step would be to expand the study to other positions and sports. A future study into the area could be boosted by Zebra Technologies RFID player tracking system coming available to the public via the NFL. RFID tracking would allow future models to improve by looking players throwing speed, route running speed and other on-field metrics that can't be measured by a human eye.

5 References

[1] Widjaya, I. (2016). *How Big Data Analytics Make Shane Battier a Better NBA Player (Kobe Bryant Case Study) - Biz Epic*. [online] Biz Epic. Available at: http://www.bizepic.com/2015/02/26/how-big-data-analytics-make-shane-battier-a-better-nba-player-kobe-bryant-case-study/ [Accessed 18 Jun. 2016].

[2] Footballoutsiders.com. (2016). *FOOTBALL OUTSIDERS: Innovative Statistics, Intelligent Analysis | CONTACT THE OUTSIDERS*. [online] Available at: http://www.footballoutsiders.com/contact [Accessed 17 Nov. 2016].

[3] SearchSoftwareQuality. (2016). *What is application program? - Definition from WhatIs.com*. [online] Available at: http://searchsoftwarequality.techtarget.com/definition/application-program [Accessed 17 Nov. 2016].

[4] Sports-Reference.com - Sports Stats, a. and Sports Stats, a. (2016). *Approximate Value - SR Blog*. [online] Sports-Reference.com. Available at: http://www.sports-reference.com/blog/approximate-value/ [Accessed 17 Nov. 2016].

[5] Docs.oracle.com. (2016). *Classification*. [online] Available at: https://docs.oracle.com/cd/B28359_01/datamine.111/b28129/classify.htm#i1005746 [Accessed 1 Dec. 2016].

[6] Harris, D. (2016). *Gigamp | Here's more evidence that sports are a goldmine for machine learning*. [online] Gigaom.com. Available at: https://gigaom.com/2015/02/12/heres-more-evidence-that-sports-is-a-goldmine-for-machine-learning/ [Accessed 7 Dec. 2016].

[7] Lopez, M. (2016). *StatsbyLopez*. [online] StatsbyLopez. Available at: https://statsbylopez.com/ [Accessed 23 Sep. 2016].

[8] GitHub. (2017). *maksimhorowitz/nflscrapR*. [online] Available at: https://github.com/maksimhorowitz/nflscrapR [Accessed 3 Nov. 2017].

[9] Guru99.com. (2017). *Cite a Website - Cite This For Me*. [online] Available at: http://www.guru99.com/black-box-testing.html [Accessed 6 May 2017].

[10] Hocking, D. (2017). *High Resolution Figures in R*. [online] Daniel J. Hocking. Available at: https://danieljhocking.wordpress.com/2013/03/12/high-resolution-figures-in-r/ [Accessed 8 May 2017].

[11] Institute, D. (2017). *K Means Clustering Algorithm: Explained – DnI Institute*. [online] Dni-institute.in. Available at: http://dni-institute.in/blogs/k-means-clustering-algorithm-explained/ [Accessed 19 Apr. 2017].

[12] Kaggle.com. (2017). *Detailed NFL Play-by-Play Data 2015 | Kaggle*. [online] Available at: https://www.kaggle.com/maxhorowitz/nflplaybyplay2015 [Accessed 29 Nov. 2016].

[13] Moodle.ncirl.ie. (2017). *Sign in to your account*. [online] Available at: https://moodle.ncirl.ie/mod/resource/view.php?id=30111 [Accessed 8 Apr. 2017].

[14] MySQL Tutorial. (2017). *Import CSV File Into MySQL Table*. [online] Available at: http://www.mysqltutorial.org/import-csv-file-mysql-table/ [Accessed 9 Dec. 2017].

[15] Nflsavant.com. (2017). *Cite a Website - Cite This For Me*. [online] Available at: http://www.nflsavant.com/dump/combine.csv?year=2015 [Accessed 3 Feb. 2017].

[16] R Statistics.Net - Main. (2017). *Advanced Linear Regression: A Case study*. [online] Available at: http://rstatistics.net/linear-regression-advanced-modelling-algorithm-example-with-r/ [Accessed 8 Apr. 2017].

[17] Rego, F. (2017). *Quick Guide: Interpreting Simple Linear Model Output in R*. [online] Feliperego.github.io. Available at: http://feliperego.github.io/blog/2015/10/23/Interpreting-Model-Output-In-R [Accessed 1 May 2017].

[18] Rpubs.com. (2017). *RPubs - Visualizing Dendrograms in R*. [online] Available at: https://rpubs.com/gaston/dendrograms [Accessed 27 Apr. 2017].

[19] Selectorgadget.com. (2017). *SelectorGadget: point and click CSS selectors*. [online] Available at: http://selectorgadget.com/ [Accessed 5 Mar. 2017].

[20] Spotrac.com. (2017). *NFL Rankings*. [online] Available at: http://www.spotrac.com/nfl/rankings/2014/contract-value/quarterback/ [Accessed 7 Mar. 2017].

[21] Stattrek.com. (2017). *Residual Analysis in Regression*. [online] Available at: http://stattrek.com/regression/residual-analysis.aspx?Tutorial=AP [Accessed 2 May 2017].

[22] YouTube. (2017). *K-Fold Cross Validation - Intro to Machine Learning*. [online] Available at: https://www.youtube.com/watch?v=TIgfjmp-4BA&t=64s [Accessed 31 Apr. 2017].

[23] YouTube. (2017). *Missing Value - kNN imputation in R*. [online] Available at: https://www.youtube.com/watch?v=u8XvfhBdbMw [Accessed 2 May 2017].

[24] Chadiha, J. (2017). *What's wrong with RG III, Newton, Kaepernick?*. [online] ESPN.com. Available at: http://www.espn.com/nfl/story/_/page/hotread141204/russell-wilson-robert-griffin-iii-colin-kaepernick-cam-newton-face-adversity-scrutiny [Accessed 8 May 2017].

[25] Anon, (2017). [online] Available at: https://home.deib.polimi.it/matteucc/Clustering/tutorial_html/ [Accessed 9 May 2017].

[26] Drinen, D. (2017). *Approximate Value » Pro-football-reference.com blog*. [online] Pro-football-reference.com. Available at: http://www.pro-football-reference.com/blog/index37a8.html [Accessed 13 Oct. 2016].

[27] Gabler, N. (2017). *Football is Becoming a Science: the Rise of NFL Analytics - Last Word on Pro Football*. [online] Last Word on Pro Football. Available at: http://lastwordonprofootball.com/2017/04/07/rise-nfl-analytics/ [Accessed 9 May 2017].

[28] LINDSEY, J. (2017). *Cite a Website - Cite This For Me*. [online] Wired.com. Available at: https://www.wired.com/2016/01/the-nfls-impending-data-revolution/ [Accessed 9 May 2017].

[29] Rosenthal, G. (2017). *The CBA in a nutshell*. [online] Profootballtalk.nbcsports.com. Available at: http://profootballtalk.nbcsports.com/2011/07/25/the-cba-in-a-nutshell/ [Accessed 9 May 2017].

6 Appendix

6.1 Project Proposal

6.1.1 Objectives

My objectives of this project are to analyze play-play data for the 2014 NFL season to pull insightful information from and apply machine learning algorithms like regression and clustering to define a player return on investment to a franchise. I will achieve this by using yearly salary as an indicator of investment allowing me to use serval factors to predict the true yearly salary of player thus resulting in a return on investment figure on player's importance to the franchise. As a result, I can make the conclusion of which method holds more value. Return on Investment will play a pivotal role in this project, by assigning ROI to draftee's and players signed it can create a picture of the success front offices in the NFL have had in their return on investment of players.

Success will be based on several factors.

- Offensive efficiency
- Wins
- Value to predicted value

Example:

Person A says: "Player B and Player C are the best combinations for the Eagles!"

Person B says: "Player B and Player X are a better combination for the Eagles!"

Other objectives of the project are to create a clustering model by where I can group players on similarity in return giving customers a visual image of what players are similar in both performance and physical attributes.

6.1.2 Background

My project idea stemmed from my interest in the NFL (National Football League) and statistics. One regular season for a team consists of 16 games with each pass

and run analyzed down to which way direction the player ran, how much help did he have from teammates in gaining/losing yards and which type of defense did the opposing team set up in to allow a run to gain yards. The sport is rich in data because of the way the games are analyzed leaving a lot for data analysts to interpret. By analyzing the datasets obtained, I can then interpret team's tendencies in signing players that prove valuable to the franchise. Using the historical data, I can then create a predictive model based on past signings and drafted players to find value in future negotiations.

When the college decided that the Data analytics stream would be going ahead I felt doing a sports analytics project to show off my skills that I've learned in the four years would be something that I'd enjoy over the course of my 4th year. From watching and researching NFL the past years I know where to start looking for data sets on websites like:

http://www.footballoutsiders.com (Advanced statistics)
http://www.espn.com/nfl/statistics (More basic statistics)
Seeing how sabermetrics *("the search for objective knowledge about baseball."* - Bill James) was applied in baseball, I was always curious would it possible to ever apply the same idea to NFL in some aspect, as baseball is more player vs. player where as American Football is team vs. team in most cases.

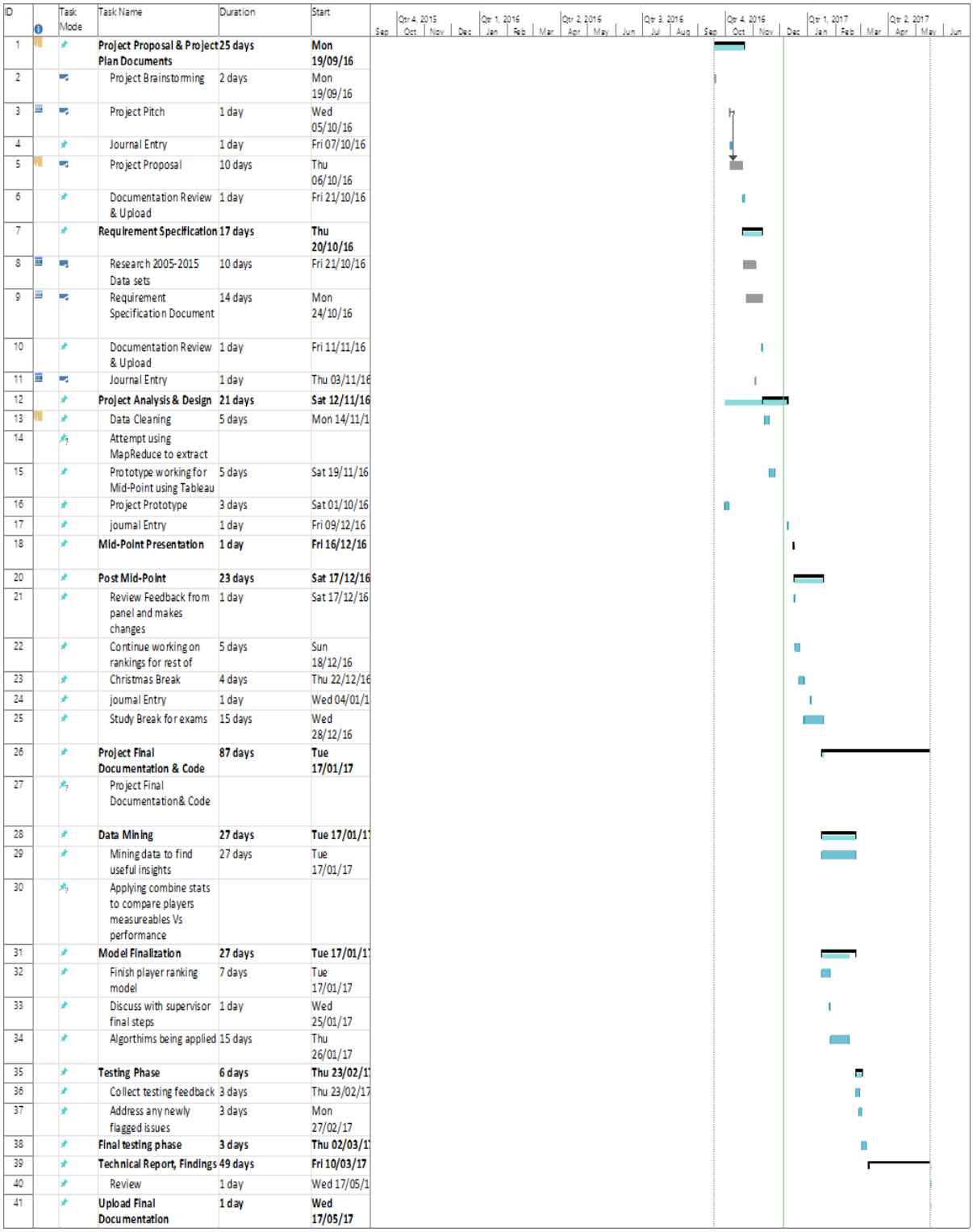
6.1.3 Technical Approach

From previous project experience, I will be following an agile approach during the lifecycle of this deliverable. The reason using this approach is that I want to deliver on the objectives I am asking myself.

Since this is a solo project and can't delegate tasks to other people, I could find myself caught up at any stage of the project. I understand there is a small risk that I will not get my project complete the way I intended and am planning for the situation that I do not and which pieces of the project will be most valuable to the overall project.

By phasing out my project, I can put more important pieces to the forefront while leaving the less valuable pieces till I am happy with the completed segments.

6.1.4 Project Plan



ID	Task Mode	Task Name	Duration	Start
1		**Project Proposal & Project Plan Documents**	**25 days**	**Mon 19/09/16**
2		Project Brainstorming	2 days	Mon 19/09/16
3		Project Pitch	1 day	Wed 05/10/16
4		Journal Entry	1 day	Fri 07/10/16
5		Project Proposal	10 days	Thu 06/10/16
6		Documentation Review & Upload	1 day	Fri 21/10/16
7		**Requirement Specification**	**17 days**	**Thu 20/10/16**
8		Research 2005-2015 Data sets	10 days	Fri 21/10/16
9		Requirement Specification Document	14 days	Mon 24/10/16
10		Documentation Review & Upload	1 day	Fri 11/11/16
11		Journal Entry	1 day	Thu 03/11/16
12		**Project Analysis & Design**	**21 days**	**Sat 12/11/16**
13		Data Cleaning	5 days	Mon 14/11/1
14		Attempt using MapReduce to extract		
15		Prototype working for Mid-Point using Tableau	5 days	Sat 19/11/16
16		Project Prototype	3 days	Sat 01/10/16
17		journal Entry	1 day	Fri 09/12/16
18		**Mid-Point Presentation**	**1 day**	**Fri 16/12/16**
20		**Post Mid-Point**	**23 days**	**Sat 17/12/16**
21		Review Feedback from panel and makes changes	1 day	Sat 17/12/16
22		Continue working on rankings for rest of	5 days	Sun 18/12/16
23		Christmas Break	4 days	Thu 22/12/16
24		journal Entry	1 day	Wed 04/01/1
25		Study Break for exams	15 days	Wed 28/12/16
26		**Project Final Documentation & Code**	**87 days**	**Tue 17/01/17**
27		Project Final Documentation& Code		
28		**Data Mining**	**27 days**	**Tue 17/01/1**
29		Mining data to find useful insights	27 days	Tue 17/01/17
30		Applying combine stats to compare players measureables Vs performance		
31		**Model Finalization**	**27 days**	**Tue 17/01/1**
32		Finish player ranking model	7 days	Tue 17/01/17
33		Discuss with supervisor final steps	1 day	Wed 25/01/17
34		Algorthims being applied	15 days	Thu 26/01/17
35		**Testing Phase**	**6 days**	**Thu 23/02/1**
36		Collect testing feedback	3 days	Thu 23/02/17
37		Address any newly flagged issues	3 days	Mon 27/02/17
38		**Final testing phase**	**3 days**	**Thu 02/03/1**
39		**Technical Report, Findings**	**49 days**	**Fri 10/03/17**
40		Review	1 day	Wed 17/05/1
41		**Upload Final Documentation**	**1 day**	**Wed 17/05/17**

Figure 15: Project Plan

6.1.5 Project Restrictions

Restrictions are recognised at the instigation of the project which incorporates the following.

Time

The duration of the project is confined to pre-agreed dates which have a starting date of September 19th, 2015 and a completion date of 10th May 2016.

Copyright

The data was provided by the Carnegie Mellon University statistical researchers, who used an API to parse the official NFL website and created a package in RStudio allowing users to download and exploit the data. Other data sources were scraped and will be discussed in detail throughout the document

Software Resources

Software needed is provided by the college with SPSS offered free under the student licensing agreement.

Budget

Budget was planned for unforeseen costs that may arise later in the project; this was the case when hosting the database online via GearHost who charged €5 for one month's hosting.

6.2 Monthly Journals

6.2.1.1 Month: September

My Achievements

This month, I decided on my project idea ahead of my project pitch.

My Reflection

I felt, I researched my project idea well and found it will be very interesting to do. Well, look well on my resume as it will show how I applied data analysis to find value and form an analytical view of the future of the sports teams.

Intended Changes

Next month, I intend to gather data sets for all draft classes between 2005-2015 and offseason signings of the same time frame.

Supervisor Meetings

No supervisor assigned.

6.2.1.2 Month: October

My Achievements

This month, I was able to contact ‘football outsiders’ about getting the large data set I required for several years. Have set up MySQL Database in the meantime waiting for the all the data to be sent to me. Have got gained experience using Tableau with sample data sets of a similar nature to my project.

My contributions to the projects included contacting football outsiders and arranging what information I needed and signing off any forms they asked about the use of the data.

My Reflection

However, I was not successful in getting as strong as start in the project I might hope for with so many other CA’s due over the course of the month

Intended Changes

Next month, I will allocate more time to getting my project up to standard for the mid-point presentation

Supervisor Meetings

Date of Meeting: None

Items discussed: couldn’t organize a meeting with my supervisor before the reading week, we have arranged to meet every Thursday at 7 pm from after reading week

Action Items: Organised meeting times

6.2.1.3 Month: November

My Achievements

This month, I worked on the completion of both documents the Requirements specification and the Technical report. Completing both documents took most the time up. I was also able to create a prototype for my midpoint presentation which contains the first two phases of the KDD process which Manuel suffice as a prototype.

My Reflection

I felt I worked hard this month even with several other project deadlines lurking. Although I was hopeful to maybe more done by the Christmas period to showcase the research and work I had put in over the last month. I felt I put the maximum effort into creating and formatting the documents for upload. It helped when creating the prototype; I had more clear idea of what I wanted to do and the process of doing it.

Intended Changes

Next month, I intend to continue working on the KDD process I have in place. Will need to cut back on time spent on the project with Exams starting in early January. Intended changes for next month includes revisiting the technical report to evaluate the structure and make sure they match the goals I have set out. The agile development approach I am following allows me to change certain features of the project that aren't at the core of the project. So, by the end of the month, I will have a clear picture if the goals set out are doable by April and which processes need the most attention.

Supervisor Meetings

Have met with Manuel on few occasions this month, He has helped me on understanding what I want the result to look like and my what to include and to remove from both my documents and overall project.

6.2.1.4 Month: December

My Achievements

This month I completed all the documentation and my prototype due for the mid-point presentation. I also attended my mid-point presentation and received my grade for my efforts over the past few months.

My Reflection

I felt I worked extremely hard for the 1st half of the month making sure all documentation was completed and my prototype was an up to scratch as me and Manual had agreed upon. I didn't find much time over the 2nd half of the month

with other deadlines, Christmas period and studying for January exams taking up much of the time.

Intended Changes

Next month, I intend to continue working on the KDD process with a redefined project plan being used after the feedback I received during the mid-point presentation being taking on aboard and how to approach the 2nd half of the module and ensure the completion of the project.

Supervisor Meetings

Only chance I got to meet with my supervisor Manual was during the mid-point presentation as are meeting slot was cancelled due to Manual being asked to sit in on a mid-point presentations while other marker was out sick.

6.2.1.5 January

My Achievements

This month I spent much of my time researching algorithms to suit my project e.g. Regression. Regression will be a major part of the machine learning process for me with creating metrics to evaluate a player worth to a franchise.

I have received feedback from my supervisor and quickly adjusted the mistakes Manuel pointed out to ensure a more complete project come final deadline

My Reflection

I felt I spent much of time this doing more of the theory side of the project with less emphasis on coding until Ralf (Data & Web Mining lecturer) explains in detail the process behind implementing machine learning algorithms into my project. It was a well needed period of reflection to see where exactly where I am at and how far there is to go and plan accordingly

Intended Changes

Next month, I intend to continue working on the KDD process with a redefined project plan being used after hearing feedback from Manuel on the importance of documentation in my stream.

Supervisor Meetings

I got to meet Manuel once since I've started back and heard some good unbiased feedback and used it as motivation moving forward through the last semester to gauge how well I was performing in certain areas and where to improve in other areas.

My Achievements

Spend much of the month focusing on new packages on RStudio like the NFL Package to dig up new techniques to find the predicated value of a player. I've made minor changes to my documentation throughout the month

Also, gave access to Dropbox and Github to Manuel to keep track of my ongoing work.

Completed my profile ahead of the showcase selection.

My Reflection

I completed a lot of tedious and necessary work that needed review and changes, during the month little was done coding but more on the process of how I intend to finish out the project in the coming months

Intended Changes

Next month, I intend to continue working on the KDD process with a redefined project plan being used importing the use of regression algorithms to gauge a how accurate the algorithm is currently.

Supervisor Meetings

Kept in constant touch with Manuel via email and face to face. Discussed similar areas of the upload and document work mainly.

6.2.1.6 February

My Achievements

Spend much of the month focusing on new packages on RStudio like the NFL Package to dig up new techniques to find the predicated value of a player. I've made minor changes to my documentation throughout the month

Also, gave access to Dropbox and GitHub to Manuel to keep track of my ongoing work.

Completed my profile ahead of the showcase selection.

My Reflection

I completed a lot of tedious and necessary work that needed review and changes, during the month little was done coding but more on the process of how I intend to finish out the project in the coming months

Intended Changes

Next month, I intend to continue working on the KDD process with a redefined project plan being used importing the use of regression algorithms to gauge a how accurate the algorithm is currently.

Supervisor Meetings

Kept in constant touch with Manuel via email and face to face. Discussed similar areas of the upload and document work mainly.

6.2.1.7 March

My Achievements

Created my poster and built my first algorithm of the project, needs some fine tuning but its good start with the deadline a good distance away

My Reflection

Made a big step in the completion of the project with my first of 3 Algorithms being built, have ideas to expand on the model improve it

Intended Changes

Next month on the run in to the deadline I plan to improve the regression model while developing a KNN and clustering algorithm

Supervisor Meetings

Kept in contact over the course of the month via email and face to face to discuss small changes and ideas.