

Touchdown – A Predictive and Detailed Analysis of the National Football League

Technical Report

Tristan Balita

X15589937

X15589937@student.ncirl.ie

BSc (Hons) in Technology Management

Specialisation – Data Analytics

14/12/2018

1 Contents

2	Executive Summary.....	6
3	Introduction	6
3.1	Background & History	7
3.2	Rules and Procedures.....	9
3.3	Project Scope	13
3.4	Methodology.....	14
3.5	Technical Approach.....	16
3.6	Technical Details	16
3.7	Technical Hardware	17
3.8	Project Plan	17
4	Preliminary Analysis.....	18
4.1	Special resources required.....	18
5	System Requirements.....	18
5.1	Functional requirements.....	18
5.1.1	Use Case Diagram	18
5.1.2	Requirement 1 Interacting with the dataset in Tableau.....	19
5.1.3	Data requirements	21
5.1.4	Performance/Response time requirement.....	21
5.1.5	Availability requirement	21
5.1.6	Recover requirement.....	21
5.1.7	Security requirement	21
5.1.8	Reliability requirement	21
5.1.9	Maintainability requirement.....	21
5.1.10	Extendibility requirement	22
5.1.11	Resource utilization requirement	22
6	System Architecture.....	22
6.1	Implementation	24
6.1.1	Data Selection	24
6.1.2	Data Pre-processing/Data Cleaning	27
6.1.3	Data Transformation.....	29
6.1.4	Data Mining.....	31

6.1.5	Interpretation/Evaluation	32
7	Interface Requirements	34
7.1	Graphical User Interface (GUI).....	35
7.2	Statistical Testing	39
8	References	43
9	Appendix	44
9.1	Definitions, Acronyms, and Abbreviations.....	44

Declaration Cover Sheet for Project Submission

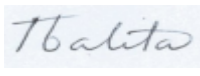
SECTION 1 *Student to complete*

Name: Tristan Balita
Student ID: X15589937
Supervisor: Eugene O'Loughlin

SECTION 2 Confirmation of Authorship

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

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

Signature: 

Date: 5/5/19

2 Executive Summary

The NFL is a league of a sport called American Football originated in United States of America. It was founded in 1920 and called the American Professional Football Association, back then the league was played with only eleven teams. Today, the NFL consists of 32 teams split over two conferences: the American Football Conference (AFC) and the National Football conference (NFC) and both conferences consists of four division of four teams.

Some of the problems I will tackle in the project are finding the right dataset to perform the planned analysis, data cleansing is a big part of data analytics and methodology to be used.

In this technical report, I will set out the objectives, requirements and the end goal for this project. To be able to meet some necessary software would have to be used. Some example of such is R Studio with the programming language of R for retrieving and analysing of datas, Tableau software for a development of an interactive visualisation/dashboard for displaying data in a friendly-user manner, Microsoft Excel for data cleansing, SPSS for interactive statistical analysis including T-test, ANOVA and many more.

3 Introduction

In this project an analysis of multiple dataset from NFL will be examined and analyzed to give meaningful insights, investigate trends, discovering links and patterns and lastly make predictions using Machine Learning using different algorithms.

The plan is to analyze the dataset gathered by performing multiple algorithms to get results, prediction, answers and generate questions, along with performing multiple statistical analysis. This document will show the steps, procedures and software/applications that is involved and needed to carry out the project.

Different data visualization tools will be used to display the analyzed data in a meaningful and clear way for the user to interpret from it. An example of this showing results in Tableau dashboard.

In recent years, the data analysis of NFL has been steadily growing popular to the NFL teams. Specifically, 'Player tracking data' and it is the next analytics arms race in the NFL. With this it allows coaches, General Managers and more to make insights, devise strategy, maximizes

player's skills and strengthen their chances of winning a game. They are able to collect data per player from a RFID tracking device that is attached in player's shoulder pads.

The intended customer for this dataset analysis is NFL fans.

3.1 Background & History

The National Football League in short 'NFL' is an American football organization in the world originating in America back in the 1910s. Way before 1910 the first sign of professional football goes back in the 1892 where William "Pudge" Heffelfinger, a player who was paid 500\$ to play football for a club called 'Pittsburgh'. A little while after that, college football starts to emerge first in the East Coast while professional football in the Midwest.

Before the NFL was created, it was called the American Professional Football Association which was founded in Canton, Ohio back in 1920. It's first president was a legendary athlete Jim Thorpe. The league at first was created as an agreement to not let teams to steal each other's player. This will eventually lead to the formation of the National Football League in 1921 where the APFA started to release official standings of the league.

The NFL started competing in a collegiate football towards the end of the World War 2. One of the reasons NFL gained a large attraction was the creation of the T – formation which created a much faster paced offense that ultimately leads to more touchdowns in a game, this made the NFL more excited to watch. In 1942 the first team in the NFL to become a franchise was the Los Angeles Rams. Around 1950s is where the first game of NFL was aired on television.

Up until the 1960s, baseball was America's favorite spectator sport until NFL surpassed them for the top spot in this decade. Ever since then, American Football was growing at such pace that the NFL couldn't keep up and the creation of the American Football League was created to compete with the NFL. This allows fans to have more opportunities to watch the sport with now that the two league is competing. Then in 1966, both the NFL and AFL decided to merge for the 1970 season and thus creating the two conferences, the National Football Conference (NFC) and the American Football Conference (AFC) to create the "Super Bowl". The Super Bowl is the championship game in NFL based on the winner from the NFC and AFC.

The NFL is made up of a total of 32 teams that is divided into two conferences: The National Football Conference also called (NFC) and the American Football Conference. Each of the conferences is made up of 16 teams to a total of 32 and is divided into four divisions – East, West, North and South which is made up of four teams each:

National Football Conference

Below is the list of the four divisions in the NFC:

- East division: Philadelphia Eagles, Dallas Cowboys, Washington Redskins, New York Giants
- West division: Los Angeles Rams, Seattle Seahawks, Arizona Cardinals, San Francisco 49ers
- North division: Minnesota Vikings, Detroit Lions, Green Bay Packers, Chicago Bears
- South Division: New Orleans Saints, Carolina Panthers, Atlanta Falcons, Tampa Bay Buccaneers

American Football Conference

Below is the list of the four regional divisions in the AFC:

- East division: Miami Dolphins, Buffalo Bills, New England Patriots, New York Jets
- West division: Kansas City Chiefs, Oakland Raiders, Denver Broncos, Los Angeles Chargers
- North division: Pittsburgh Steelers, Baltimore Ravens, Cincinnati Bengals, Cleveland Browns
- South division: Jacksonville Jaguars, Tennessee Titans, Indianapolis Colts, Houston Texas

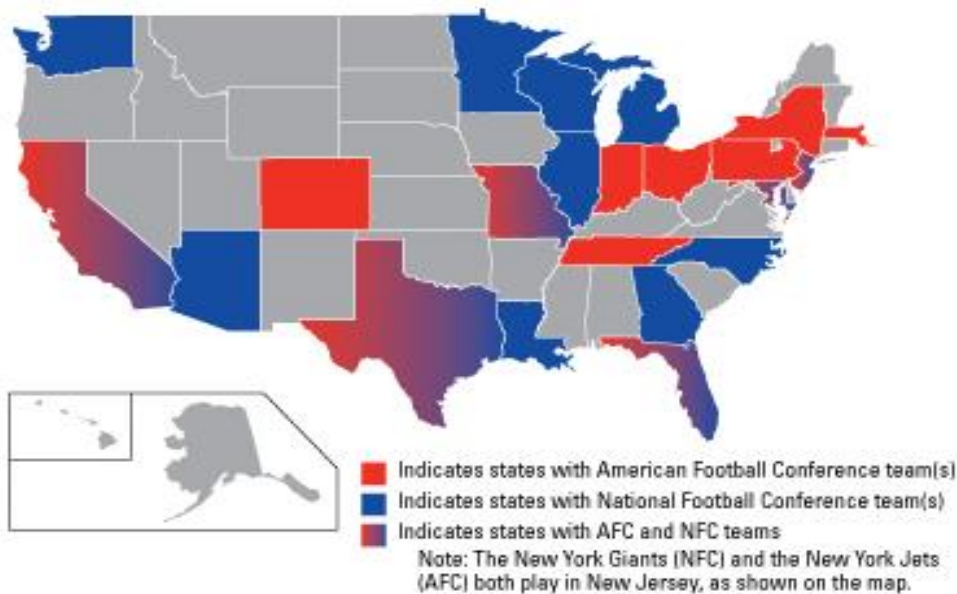


Figure 1: The National Football League Conferences.

Available at: <https://www.dummies.com/sports/football/the-national-football-league-conferences/> [Accessed December 3 2018].

The teams in the AFC today were once part of the old American Football League (AFL) which operated for ten seasons from 1960 until 1969, when it merged with the older League the NFL.

3.2 Rules and Procedures

Basically, American Football is all about gaining territory as much as possible at every play. The two teams in the field battle it out for every yard they can take from the other team. The more yards the team on the offensive takes from the defending team the better. Both the team must defend their side of the field and prevent whichever team is on the offense to gain yards. The ultimate goal for the teams is to gain enough ground on the field to be able to score a touchdown or a field goal.

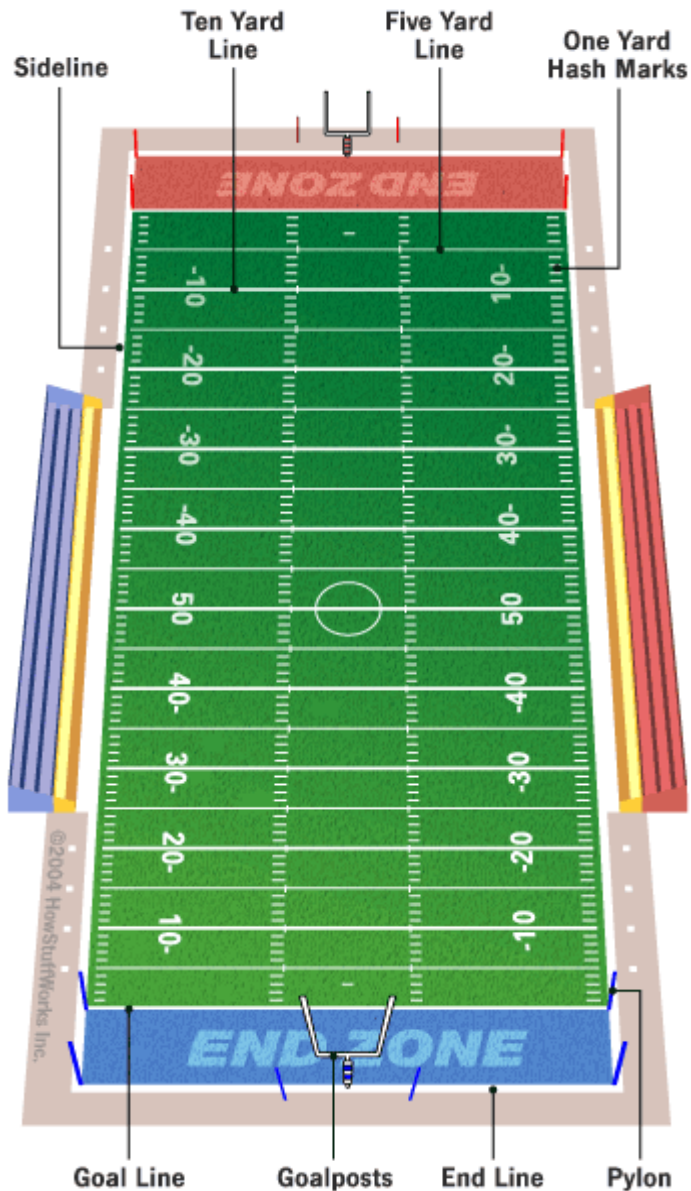


Figure 2: American Football stadium

American Football is a game of gaining yards played on the field by two teams. The playing field for an official NFL field is a rectangle that is 110 metres long (120 yards) and 49 metres wide (53 yards). The fields are covered in grass and is set in an outdoor stadium.

Below will be an example of an American football stadium:

These are the standard features of a professional American Football field:

- **End line** – an end line is the white line towards the end of the football field that connects both two parallel side lines from each side of the football field composing a

rectangular shape of the field as shown in the diagram. They are 6 feet wide and usually coloured white.

- **Side line** – together with the end line, the side line also measures at 6 feet wide that runs from the end line of one side to the other and covers both side of the field.
- **Goal line** – the goal line measures 8-inch-wide that runs across before the end zone. At the end of both sides of the line lies a two orange coloured pylons flank.
- **End zone** – these are the areas at the end of the field between the end line and goal line. They measure 10-yard-wide and the goal for each team to reach. The end zone behind where a team is playing is their end zone and the one ahead is the opponent's.
- **Goal posts** – the goal posts are positioned centre on the end line at the back of the end zone. They are a pole that measures 10 feet high and horizontal cross bar that's attached on top of the pole. At both ends of the horizontal cross bar lies two 18-foot-high, 6-inch cross bar in an up-right post that measures a height of 30 feet above the ground. On top both the up-right post is a tied ribbon.
- **Yard lines and hash marks** – the yard and hash lines run between the two goal lines, both side of the field and hash marks measures 1 yard each. The yard lines on the other hand runs a white line across the field every 5 yards and for every tenth yard they are numbered on the field (10, 20, 30).

The most important and essential equipment in the football game is the ball. The official NFL footballs that is used in professional American football are made from Wilson Sporting Goods Co. The dimension of a football is 11 inches long and a length circumference of about 28.5 inches and wide circumference of about 21.5 inches from the middle of the ball.

An NFL game is divided into a total of four quarters with each quarter being 15 minutes long and for the halftime break which is between the 2nd quarter and 3rd, an extended break is applied. When by the end of the four quarters the score is tied, there will be an overtime quarter of 15 minutes where the first team to score whether a touchdown or from a field goal wins.

Football Offense Team

The NFL allows a roster to have 53 players on a team, no more than that. Each of the team at one time only is allowed to put 11 players on the field whether this is offense, defense or the special team.

An NFL offensive team consists of:

- **Quarterbacks** – these are the player who throws the ball to the receiver or hands it to a running back on his team.

- Offensive linemen – these are the players that blocks the defending team trying to tackle the quarterback to get disrupt the play.
- Receivers – the receivers run down the field to try to get as much yards as they can as well as try to catch a ball for a first down. There are two types, a wide receiver or tight ends.
- Running backs – are the players who takes the ball from the quarterback to try and run up the field to get a first down.

Defensive team

When a team doesn't have possession of the ball, the defense team is put on the field to try and stop the other team from scoring a touchdown or get a first down. They must stop the offensive team from advancing.

Below are the defensive positions:

- Defensive linemen – these are the players who tries to tackle the quarterback or put pressure on them to try to spoil the play or to tackle him before he releases the ball. They are also responsible for stopping the running backs getting through their defense.
- Linebackers – they are the players who back up their linemen when running backs escapes and they also defend on receivers.
- Cornerbacks – they prevent wide receivers from catching the ball.
- Safeties – these are the players that plays deep in the field of the defensive team. They are the last line of defense once an offensive player gets past the linemen and cornerbacks.

Special team

When a team kicks the ball from a kick-off or a field goal attempt, they send out their special teams unit. These includes a team kicker, offensive line and players who run down the field to try to regain possession of the ball or tackle the returner.

- **Placekicker** – they are the players who try to score by kicking the ball through the goalposts and kicks the ball to the other team to start the game after a score.
- **Punter** – these are the players who kicks the ball to the other side of the field if the team cannot advance after the 3rd down.
- **Returner** – During a kickoff or punt from another team the returner will try to catch the ball and advance to the other side of the field as much as they can.

Possession

On any team in the offense, they will have start at first down to try to get the ball over the 10-yard line to get the back at first down. Any offensive team will have 3 tries to get over the 10-yard line and after that they have to kick the ball to the other side of the field.

Below is a complete list of how points are scored and how many are awarded for each of them.

Method	Description	Points
Touchdown (TD)	A ball is carried into an opponent's end zone or caught in the end zone.	6 points
Extra point	A ball is kicked through the uprights of the opponent's goalpost after a touchdown.	1 points
2-point conversion	A ball is carried into an opponent's end zone or caught in the end zone.	2 points
Field goal	A ball is kicked through the uprights of the opponent's goalpost.	3 points
Safety	A player tackles an opposing player in the opposing player's own end zone.	2 points

After the team on the offense successfully score a touchdown or field goal, they must kick back the ball to the other side of the team at the start of the next play. The only exception to this is if a team scored a safety. Whoever scores the safety will get the ball back on a free kick.

3.3 Project Scope

The scope of the project is to develop a data analysis project based on the dataset of the National Football League that will involve identifying trends, links, patterns and prediction upon studying and analyzing the dataset. This will be done and displayed using multiple applications and algorithms i.e. R Studio, R Shiny, SPSS, Microsoft Excel, Machine learning, Tableau.

The challenges encountered in the project were:

- Finding a dataset to do statistical analysis, test and comparisons.
- Time constraints including two weeks delay from choosing the Data Analytics stream.
- Learning programming from scratch at the start of the semester due to lack of programming module in the Technology Management course and was only available from first semester of year one.
- Most of the project contents will be done during semester 2 due to Data mining module will start at semester 2.
- Data selection including web scraping and manual cleaning.
- Data cleansing a raw dataset

Below are the important deadlines for the project:

- Project Pitch video – Oct 7, 2018
- Project Proposal Document – Oct 28, 2018
- Requirements Specification Document – Nov 11, 2018
- Midpoint presentation and prototype presentation – Dec 17-21, 2018
- Project final documentation and code – May 6, 2019
- Project presentations – May 13-20, 2019
- Project showcase – May 23, 2019

3.4 Methodology

The chosen methodology that will be used for this project is KDD also known as Knowledge Discovery in Databases. The KDD model for data analysis is an interactive model that has a total of nine steps. Mainly it refers to the processes of finding knowledge in data and performing a high level of specific data mining methods.

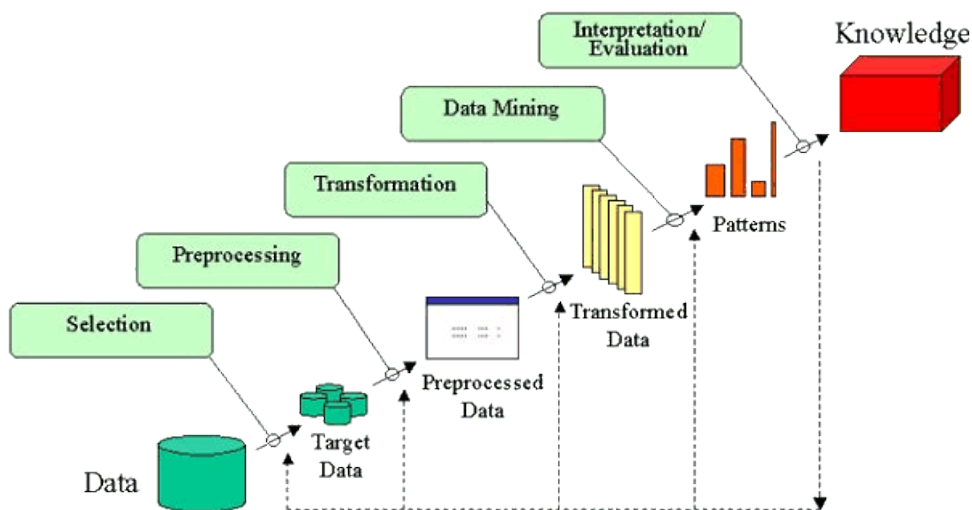


Figure 3: KDD methodology

The main users of the KDD model mostly are researchers in the following area:

1. Machine Learning
2. Databases
3. Artificial Intelligence
4. Data visualization
5. Statistics

Each one of them has a significant use for KDD model. For example, when doing a statistical analysis on a data, it is important that an objective/goals are defined before

Below are the steps involved in performing the KDD model:

1. Identifying goals for the project

In this stage, brainstorming session were performed to narrow down the choices for the type of dataset which will be used, along with how I am interested on exploring insights on the chosen dataset.

2. Selection of dataset

During this stage, I explored a lot of websites containing raw datasets. I ultimately found the NFL dataset in the website Kaggle which contains dataset in different format. I narrow down the filter to only show files that are in CSV format

3. Brainstorming

Brainstorming is a popular method of generating new ideas and solutions for a specific interest and it is usually done in groups but can be also done by yourself. A couple of brainstorming session have been done to enable to generate ideas for the project such as questions to ask in the dataset, algorithms to be use, applications to do the analysis,

4. Data Cleansing/Pre-processing

In this stage the dataset will be accessed for data cleansing. It is the process of detecting corrupt and inaccurate records to be corrected or deleted in the dataset, together with identifying incorrect, irrelevant, incomplete parts of the data and then modifying, replacing or deleting the coarse data. The dataset will be cleaned in this stage to eliminate the irrelevant data e.g. there were a lot of data columns where an example of irrelevant data was present in this case the column was gameId and description. These two will not be needed for analysis as they have no possible connections with other columns.

5. Transforming the dataset

In this stage the transformation of the dataset will be done using R studio in R programming language, where I will be conducting the analysis. The dataset will be loaded into R studio in the form of a csv file.

6. Data mining

Is the practice of examining data sets to generate new information involving methods in machine learning, statistics and database system. In this stage the data gathered from the web

will be used to analyze in R studio to discover patterns in the dataset, predict future trends and solving problems.

7. Interpretation/Evaluation

Lastly in this stage the dataset will then be interpreted to make sense of the results gathered from analysing the dataset in previous stages. The sole purpose for this is to search for meaning behind the results and try to come up with possible solutions to the problems that can occur. This will be reviewed using dashboards, graphs and Microsoft excel.

3.5 Technical Approach

Upon choosing Data Analytics as my specialisation in my final year, I quickly gathered ideas, research and planned on how the project will be tackled in the early stages, such as which website provides raw datasets that can be used for analysis. A quick google search directed me to multiple websites i.e. Data.gov.uk, Opendata.socrata.com, Kaggle.com and many more.

- Research topic ideas for the final year project
- Find a decent size dataset to be used for the project.
- Data cleansing of the dataset
- Project Proposal to be done before the due date and to continuously update it until the final submission.
- Prepare requirements and ideas to be included in the mid-point presentation in December. This will include a Powerpoint presentation along with some statistics done/conducted, diagrams, some preliminary analysis done.
- Create a Microsoft Powerpoint presentation slides for the Mid-point presentation in December, to show progress during semester one.

3.6 Technical Details

Implementation language and principal libraries

- R Studio to analyse my dataset together to generate charts and plots.
- Microsoft Word to document my workload weekly and my projects' documentation
- Microsoft Project to carefully follow the charts respectively and to help me keep in track of my progress.
- Microsoft Excel for creating subset of the raw datasets. Data cleansing will be performed here and some statistical analysis for the mid-point report.

- SPSS to further perform multiple statistical analysis in a quick way.
- Generate data visualisation and interactive visualisation using R language, Shiny and Tableau.
- Machine learning for semester 2 to generate predictive analysis.

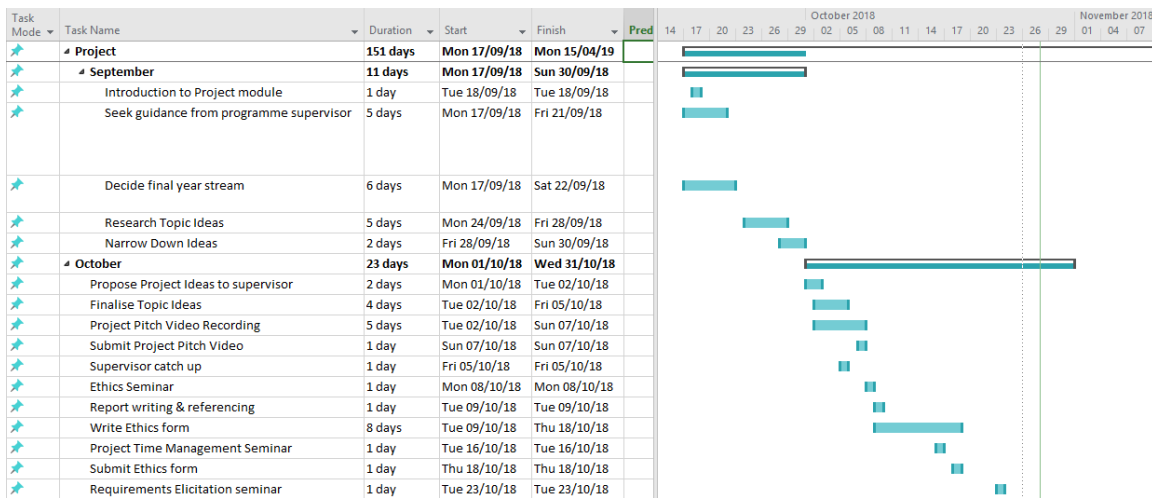
3.7 Technical Hardware

The computer used in this project is 14-inch Lenovo yoga 530 2 in 1 convertible laptop. The specifications are listed below:

- **CPU – AMD Ryzen 5 2500u with integrated Graphics card**
with 4 cores and 8 threads, this CPU allows for heavy multitasking with multiple applications running in the background and tabs open in Google Chrome. allows faster and efficient time usage.
- **Video Card – AMD Radeon Vega 8**
- **Storage – 256 Solid State Drive.**
Allows for faster access of files stored in the computer.
Faster boot up of applications such as R studio, SPSS etc.
Decreases the amount of loading times of applications
- **RAM – 8gb of random-access memory.**
Good for multitasking

3.8 Project Plan

Gantt chart using Microsoft Project with details on implementation steps and timelines



★	November	22 days	Thu 01/11/18	Fri 30/11/18
★	Mid-term break	3 days	Thu 01/11/18	Sun 04/11/18
★	Supervisor catch up	1 day	Fri 09/11/18	Fri 09/11/18
★	Requirement specification write up	8 days	Thu 01/11/18	Sat 10/11/18
★	Requirements specification submission	1 day	Sun 11/11/18	Sun 11/11/18
★	Supervisor catch up	1 day	Fri 30/11/18	Fri 30/11/18
★	Pre mid-point presentation mock	13 days	Thu 01/11/18	Sun 18/11/18
★?	Generate questions			
★?	Data Cleansing			
★?	Pick 3 Data visualisation diagrams			
★?	Setup Github			
★?	Preliminary analysis ideas			
★	December	22 days	Sat 01/12/18	Mon 31/12/18
★?	Some analysis done			
★?	Few conducted statistics			
★	Supervisor catch up	1 day	Fri 07/12/18	Fri 07/12/18
★	Mid-point final check up	2 days	Fri 14/12/18	Sun 16/12/18
★	Mid-point presentation	5 days	Mon 17/12/18	Fri 21/12/18

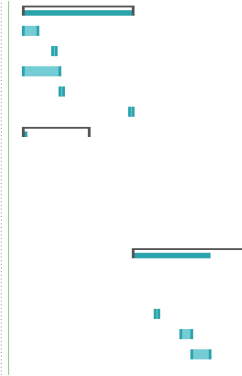


Figure 4: Project Plan

4 Preliminary Analysis

4.1 Special resources required

If applicable, e.g., books, hardware, etc.

- The use of Microsoft Project to create my Project Plan for semester 1 and semester 2. As well as Youtube to further tutorials how to maximize the functions in creating Gantt charts
- YouTube and research on Machine learning for semester 2, as it will be included in my project during semester 2.
- R Cookbook by Paul Teetor for R studio.
- A laptop for college with enough specifications to run programs such R studio and to be able to run multiple programs and apps at the same time and to have all my work in one computer instead of using the college desktops

5 System Requirements

5.1 Functional requirements

N/A

5.1.1 Use Case Diagram

The Use Case Diagram provides an overview of all functional requirements.

5.1.2 Requirement 1 Interacting with the dataset in Tableau

5.1.2.1 Description & Priority

This is the only use case that would be used in the project therefore this would be the highest priority to be able for users to interact with Tableau.

5.1.2.2 Use Case

Scope

The scope of this use case is to retrieve the chosen dashboard from Tableau to the user and to successfully display and interact with the dashboard. By using either the Tableau software in my laptop or an access from Tableau's website

Description

This use case describes the process of retrieving the chosen dashboard from the user.

Use Case Diagram

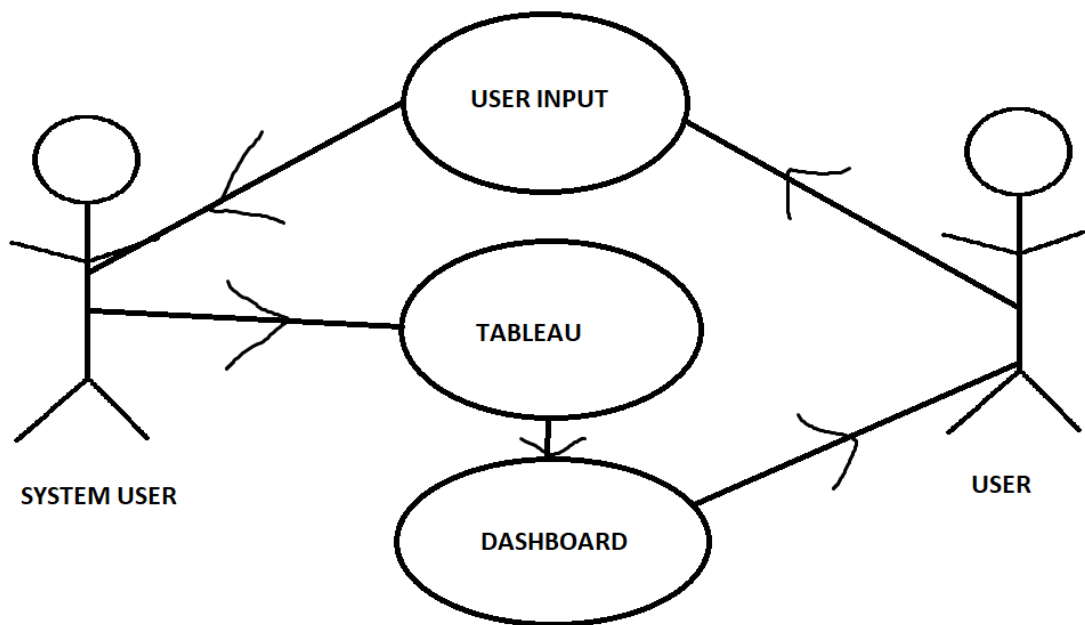


Figure 5: use case for Tableau

Flow Description

Precondition

The system is in initialisation mode is when Tableau is open and running.

Activation

This use case starts when a user inputs the chosen Tableau analysis to be displayed in a dashboard

Main flow

1. The system identifies the input of the user
2. Tableau recognizes the input and retrieve the correct dashboard
3. Tableau displays the interactive dashboard
4. User interacts with the dashboard and learns from it.

Alternate flow

A1: No internet connection

1. The system doesn't identify the input of the user
2. The User is presented with blank page.
3. Tableau doesn't display the interactive dashboard chosen.

Termination

The system presents the correct Tableau dashboards in the computer to the user.

Post condition

Back to home page where Tableau dashboard will be chosen.

5.1.3 Data requirements

5.1.4 Performance/Response time requirement

The response and performance time requirement for the project will be the data that will be used to perform multiple analysis in software's such as R studio, Tableau and SPSS. The response time will be dependent on the software that will be used, and the performance will be based on the specification of the computer used, in this case my laptop.

5.1.5 Availability requirement

The availability requirement of the project depends on when, where and how the dataset is collected. Before any statistical analysis is performed on the dataset, a small data cleansing will be necessary.

5.1.6 Recover requirement

In case of a lost or broken computer the recovery requirement for the project is a weekly backup of the files in a 64gb USB. This will include everything from Excel files to R Scripts. Secondly the files such as Microsoft Word and Excel will be backed up in OneDrive for every session.

5.1.7 Security requirement

The security requirement for the project will include a fingerprint identification laptop to go with a pin number and for the files backed up in OneDrive, the password is made up of six characters, 3 numbers and one special characters for a better secured password.

5.1.8 Reliability requirement

5.1.9 Maintainability requirement

The project will be done throughout both the semester in 4th year which has a deadline in May of 2019. I intend to continue updating and working on my project as I am a big fan of the NFL.

5.1.10 Extendibility requirement

In the future I am hoping to develop more skills and statistical analysis methods in able to continue the project and maybe move to another dataset with similar category such as the National Basketball Association (NBA).

5.1.11 Resource utilization requirement

N/A

6 System Architecture

The main programming language that will be used for the project is R Studio. This is where most of the analysis will be performed, from reading from a csv file to displaying visualized data.

R

is a programming language and a free software that is used for statistical computation to perform complex data analysis and to display it in a variety of visual graphics. It is widely used by data miners and statisticians.

R Studio

Is an open source IDE (integrated development environment for the programming language R. Mainly for statistical computing and graphics. it contains tools for plotting graphs, scatter plots, workspace management and debugging.

Tableau

Tableau is a software that produces interactive data visualization in a form of a dashboard that focuses on business intelligence. It can connect multiple data sources and allows fast insight on them through presentation in the form of dashboards that is user friendly and interactive.

Excel

Is a spreadsheet application that is used to create grids of text, numbers and formula. They are extremely valuable for many businesses for recording expenditures, budgets, income plans, to

chart data and perform formula calculations. This will be used mainly for data cleansing for the dataset.

SPSS

Is a software platform that offers a wide variety of library for machine learning algorithms, advanced statistical analysis, text analysis and integration with big data. It is a world’s leading software for solving business problems by performing hypothesis testing and predictive analysis. Organizations use it to help them understand data, forecast, plan and analyze trends.

Below is a simple diagram of the system architecture of the project.

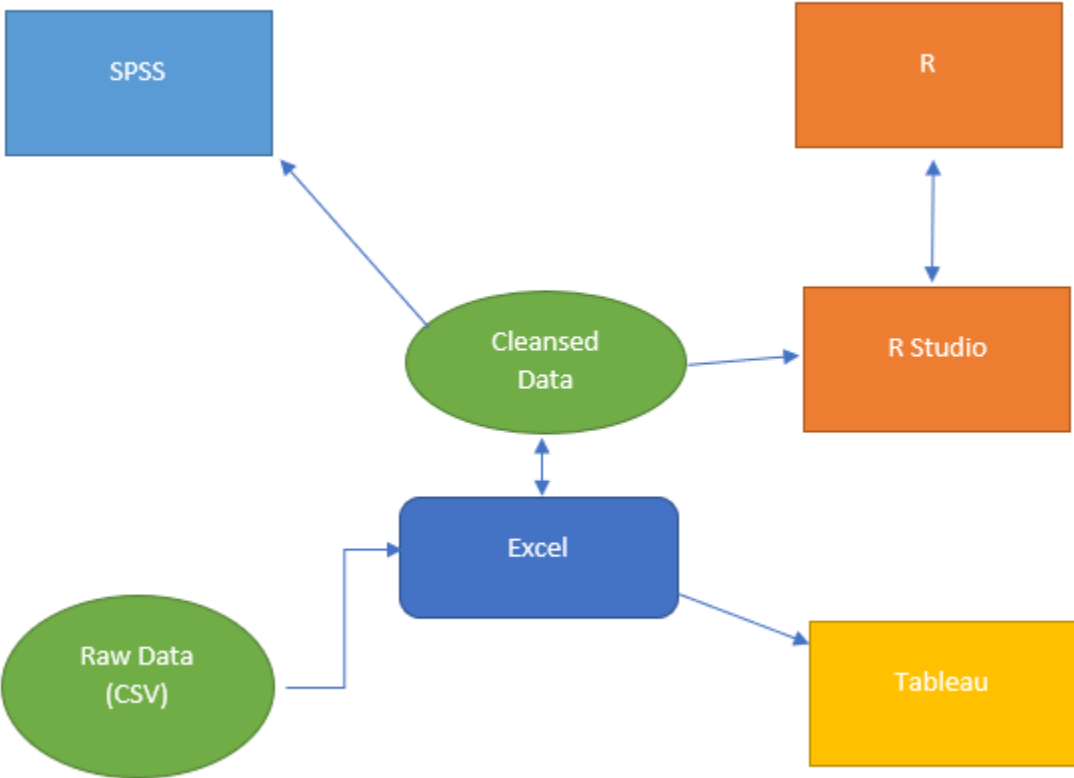


Figure 6: system architecture

What is happening here is that in the diagram multiple raw data are pulled in from multiple websites in the form of a CSV file. This will then be opened using Excel to study and understand the data together with identify goals. Secondly data cleansing will be performed prior to the goals identified for the data, this will include removing irrelevant columns, outliers and strategies for handling missing data fields.

Then the cleansed data will then be ready to be analyzed using multiple software applications such as SPSS, Tableau and R studio. In R studio, this is where the CSV file will be loaded to begin the analysis of the cleaned data. As well as showing graphs and plots to back up insights that are identified.

The cleaned data will be used in Tableau to display in an easily understandable format in the form of an interactive dashboard and for easy data analysis of the data.

SPSS will be used for additional statistical analysis of the data with easy, fast and a more user-friendly way of getting results.

6.1 Implementation

6.1.1 Data Selection

The first stage of the project was to gather the required dataset to machine learning models. The site espn.com and nfl.com provided me of the dataset to be used. Multiple dataset was gathered manually by extracting the URL into excel and creating a table from there. The steps taken for the retrieval of the data's are listed below.

1. Select Data -> From Web

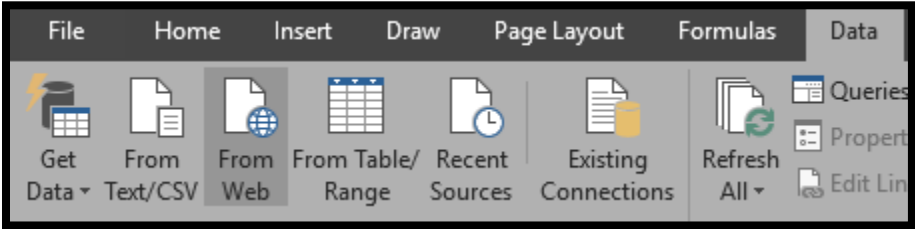


Figure 7: Data cleaning pt1

2. Paste link in the pop-up box, press ok.

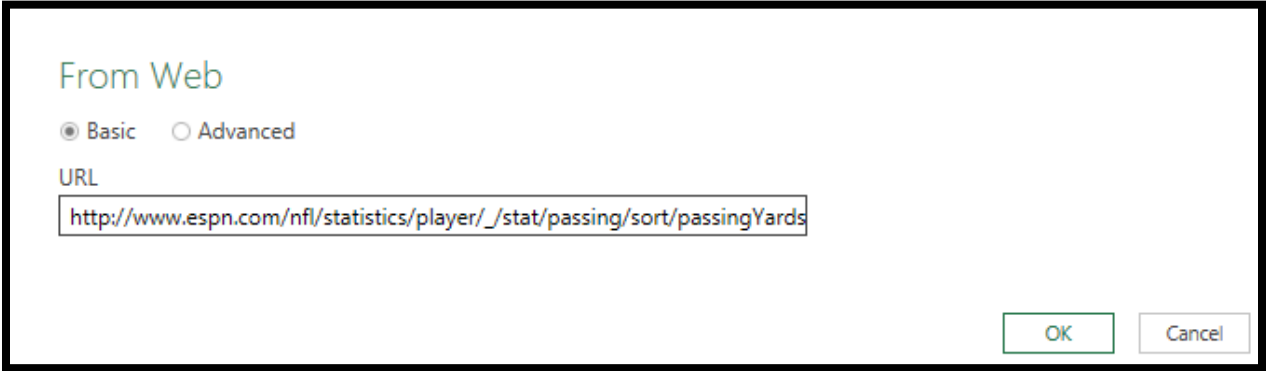


Figure 8: data cleaning part 2

3. Click "Table 0" once and press Load.

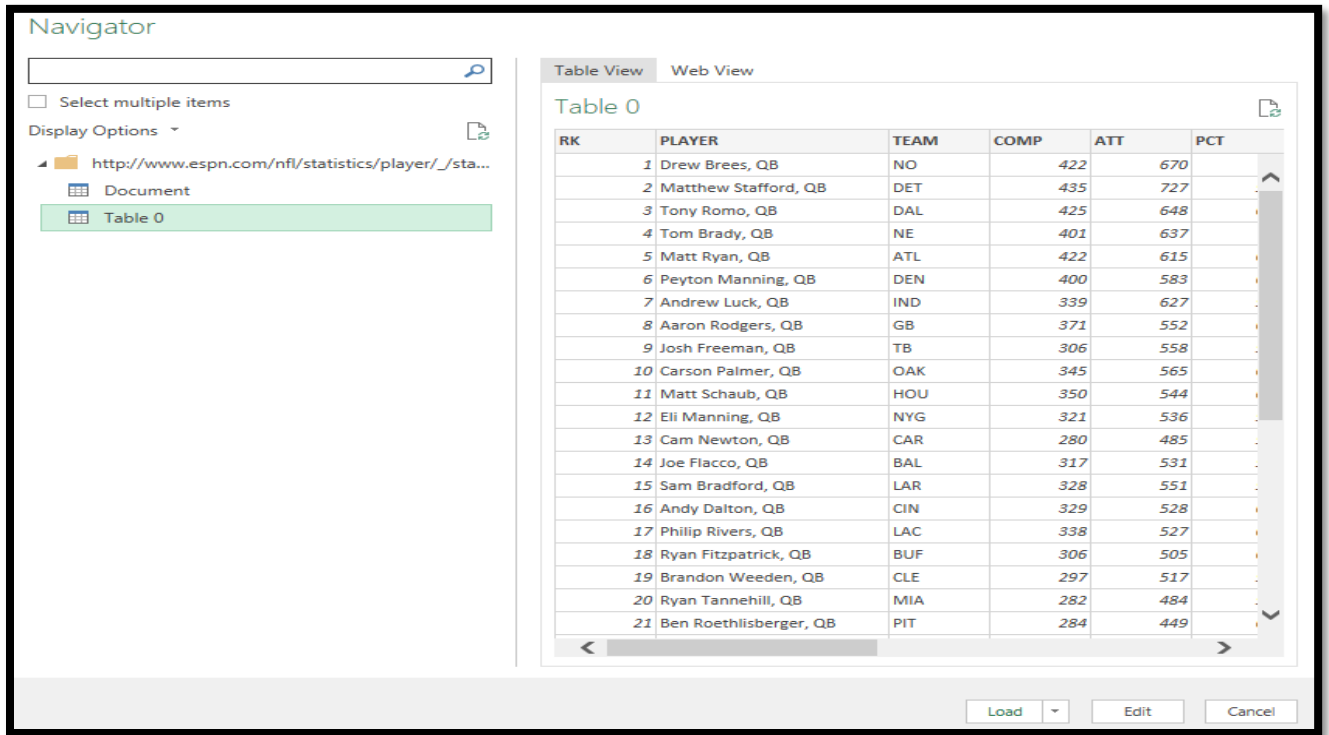


Figure 9: Data Cleaning part 3

4. Once loaded they will be separated into different sheets. As the website only displays a maximum of 40 players each webpage, steps 1 to 3 were run several times to complete one season of the quarterback's performance stats.



Figure 10: Data Cleaning part 4

5. Manual copy and paste were done to combine sheet separated players into one season. This was done for the 10 seasons.

41	40	Charlie Batch, QB	PIT	45	70	64.3	475	6.79	43	1	4	3	64.9	238
42														
43														
44														

41	40 Charlie Batch, QB	PIT	45	70	64.3	475	6.79	43	1	4	3	64.9	238
42	41 Kirk Cousins, QB	WSH	33	48	68.8	466	9.71	77	4	3	3	101.6	155
43	42 Brian Hoyer, QB	ARI/PIT	30	53	56.6	330	6.23	53	1	2	4	65.8	165
44	43 Byron Leftwich, QB	PIT	25	53	47.2	272	5.13	37	0	1	3	54.9	136
45	44 Jason Campbell, QB	CHI	32	51	62.7	265	5.2	45	2	2	6	72.8	44
46	45 Greg McElroy, QB	NYJ	19	31	61.3	214	6.9	30	1	1	11	79.2	107
47	46 Thaddeus Lewis, QB	CLE	22	32	68.8	204	6.38	23	1	1	3	83.3	204
48	47 Tyrod Taylor, QB	BAL	17	29	58.6	179	6.17	25	0	1	3	62.3	26
49	48 Shaun Hill, QB	DET	10	13	76.9	172	13.23	46	2	0	0	157.9	172
50	49 Terrelle Pryor, WR	OAK	14	30	46.7	155	5.17	38	2	1	0	70.8	52
51	50 Matt Moore, QB	MIA	11	19	57.9	131	6.9	37	1	0	2	96.6	66

Figure 11: Data Cleaning part 5

The dataset contains quarterback's statistical performance for one season. The data was collected on 150 quarterbacks who played a 1-10 seasons during 2008 to 2017. The csv files are all separated so it needs to be combined by using rbind seen in figure 12.

```
qb2017 <- read.csv("QB2017.csv", header = T, sep = ",")
qb2016 <- read.csv("QB2016.csv", header = T, sep = ",")
qb2015 <- read.csv("QB2015.csv", header = T, sep = ",")
qb2014 <- read.csv("QB2014.csv", header = T, sep = ",")
qb2013 <- read.csv("QB2013.csv", header = T, sep = ",")
qb2012 <- read.csv("QB2012.csv", header = T, sep = ",")
qb2011 <- read.csv("QB2011.csv", header = T, sep = ",")
qb2010 <- read.csv("QB2010.csv", header = T, sep = ",")
qb2009 <- read.csv("QB2009.csv", header = T, sep = ",")
qb2008 <- read.csv("QB2008.csv", header = T, sep = ",")
```

Figure 12: Data Cleaning part 6

```
combine.qb <- rbind(qb2017, qb2016, qb2015, qb2014, qb2013, qb2012,
qb2011, qb2010, qb2009, qb2008)
```

Figure 13: combining dataset

6.1.2 Data Pre-processing/Data Cleaning

In this stage, the dataset will be cleaned to carefully carry out analysis. The combined datasets were missing a column that distinguish their year, so a column was added seen in figure 14.

```
# add column year
qb2017$YEAR <- "2017"
qb2016$YEAR <- "2016"
qb2015$YEAR <- "2015"
qb2014$YEAR <- "2014"
qb2013$YEAR <- "2013"
qb2012$YEAR <- "2012"
qb2011$YEAR <- "2011"
qb2010$YEAR <- "2010"
qb2009$YEAR <- "2009"
qb2008$YEAR <- "2008"
```

Figure 15: fill year

Some rows from the PLAYER's column includes players that are not Quarterback, see figure 16. This is a result of a tricky play by a team where another player from another position throws the ball.

951	91	Patrick Crayton	WR	DAL	0	1	0.0	0	0.00	0	0	0	0	39.6	0
952	92	Eddie Royal	WR	DEN	0	1	0.0	0	0.00	0	0	0	0	39.6	0
953	93	Dante Hall	WR	LAR	0	1	0.0	0	0.00	0	0	0	0	39.6	0
954	94	Donnie Jones	P	LAR	0	1	0.0	0	0.00	0	0	0	0	39.6	0
955	95	Sidney Rice	WR	MIN	0	1	0.0	0	0.00	0	0	0	0	39.6	0
956	96	Lance Moore	WR	NO	0	1	0.0	0	0.00	0	0	1	0	0.0	0
957	97	DeSean Jackson	WR	PHI	0	1	0.0	0	0.00	0	0	1	0	0.0	0
958	98	J.J. Arrington	RB	ARI	0	1	0.0	0	0.00	0	0	0	0	39.6	0
959	99	Isaac Bruce	WR	SF	0	1	0.0	0	0.00	0	0	0	0	39.6	0
960	100	Luke McCown	QB	TB	0	1	0.0	0	0.00	0	0	0	0	39.6	0
961	101	Cleo Lemor	QB	JAX	0	2	0.0	0	0.00	0	0	0	0	39.6	0
962	102	Matt Turk	P	HOU	0	1	0.0	0	0.00	0	0	0	0	39.6	0
963	103	Kevin Faulk	RB	NE	1	1	100.0	-2	-2.00	-2	0	0	0	79.2	0

Figure 16

The following code were executed to remove these rows containing false quarterback players, see figure 17.

```
combine.qb <- combine.qb[!grepl("WR", combine.qb$PLAYER),]
combine.qb <- combine.qb[!grepl("P", combine.qb$PLAYER),]
combine.qb <- combine.qb[!grepl("TE", combine.qb$PLAYER),]
combine.qb <- combine.qb[!grepl("S", combine.qb$PLAYER),]
combine.qb <- combine.qb[!grepl("RB", combine.qb$PLAYER),]
combine.qb <- combine.qb[!grepl("LB", combine.qb$PLAYER),]
combine.qb <- combine.qb[!grepl("FB", combine.qb$PLAYER),]
```

Figure 17:

For validation purpose the validation set approach was followed, see in figure 18. The data was split into 75:25 ratio, for the training data = 75% of the data and for the test data it had 25%. The chosen dependent variable was YDS.G.

```

set.seed(101) #always run this to get the same result
smp_size <- floor(0.75 * nrow(combine.qb))
smp_size

trainqb2 <- sample(seq_len(nrow(combine.qb)), size = smp_size)
traindata <- combine.qb[trainqb2, ]
testdata <- combine.qb[-trainqb2, ]

```

Figure 18:

6.1.3 Data Transformation

In this stage, potential variable reduction and feature selection will be carried out and data exploration in hopes of finding insights, examine visuals and to better understand the data.

The library package Boruta was used for feature selection for the combined quarterback's data and to see which variable are important and unimportant, this will then help us reduce variables that are not important with the dependent variable. In figure below we can see the code for running the feature selection.

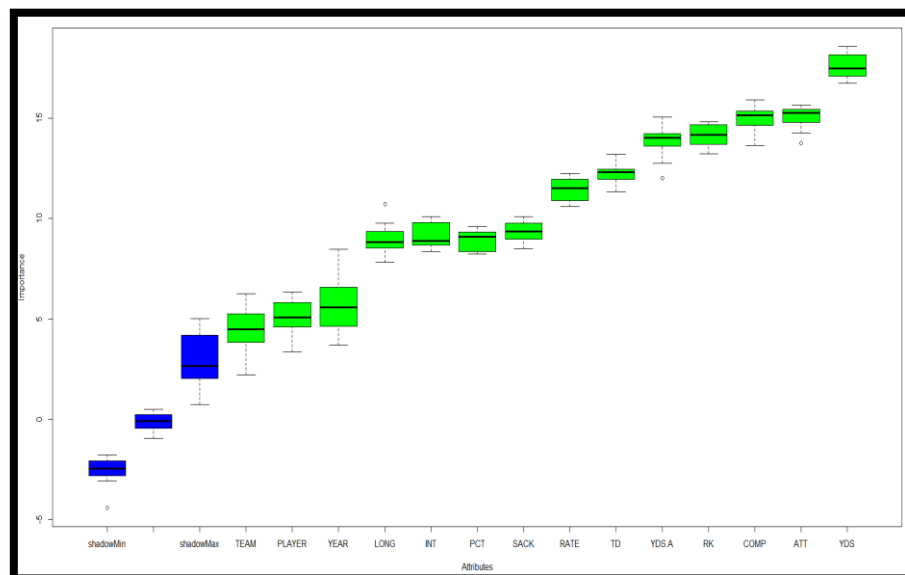
```

boruta_variables <- Boruta(YDS.G ~., data = combine.qb, doTrace = 2)

```

It resulted in 14 important attributes which is all the variables excluding the dependent variable YDS.G and no variable will be removed.

RK	Confirmed
PLAYER	Confirmed
TEAM	Confirmed
COMP	Confirmed
ATT	Confirmed
PCT	Confirmed
YDS	Confirmed
YDS.A	Confirmed
LONG	Confirmed
TD	Confirmed
INT	Confirmed
SACK	Confirmed
RATE	Confirmed
YEAR	Confirmed



```

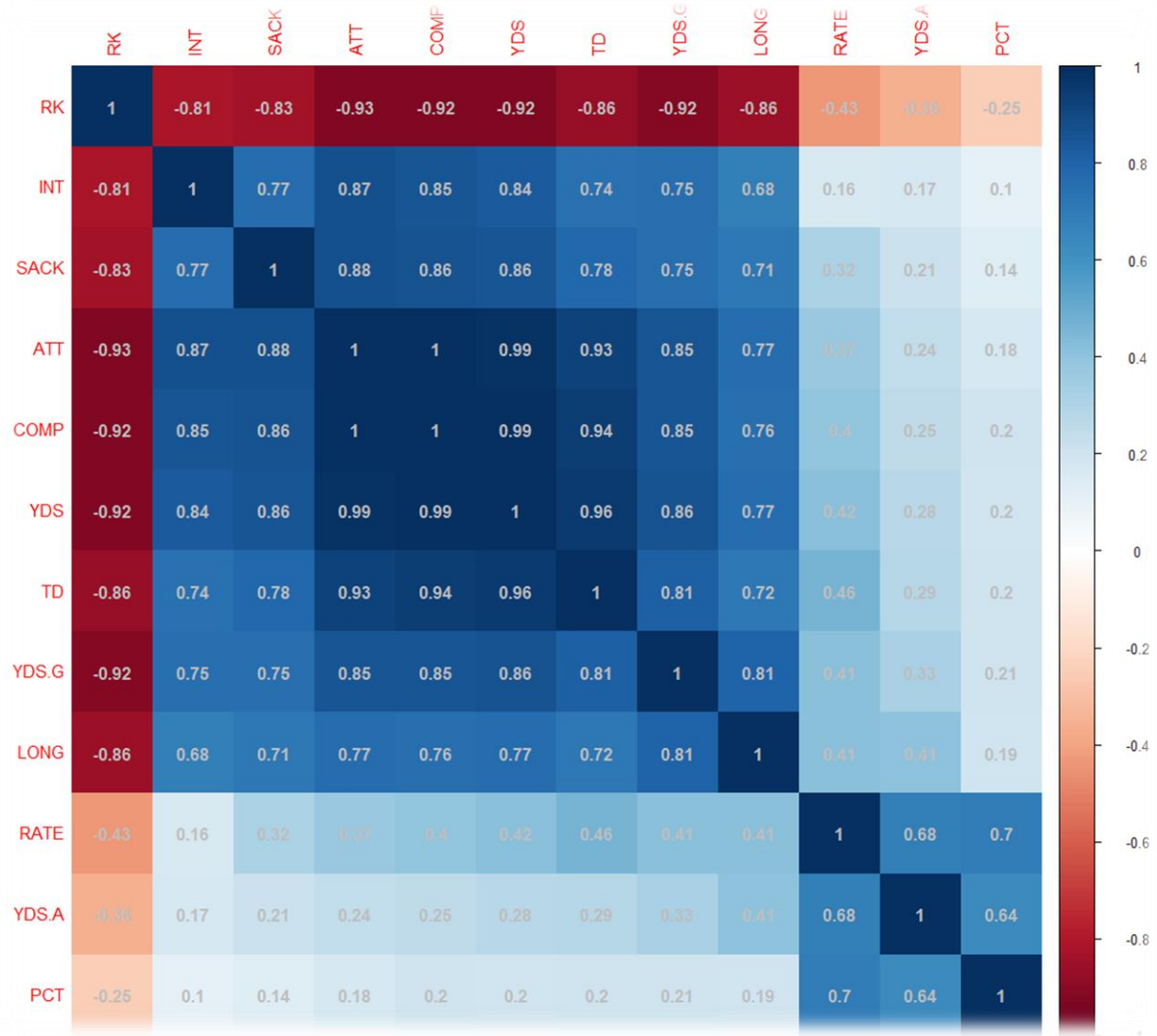
Boruta performed 22 iterations in 4.573107 secs.
14 attributes confirmed important: ATT, COMP, INT, LONG, PCT and 9 more;
No attributes deemed unimportant.

```

A correlation matrix and correlation plot is shown in figure below containing the 12 important variables that were chosen upon running the Boruta package feature selection excluding the categorical variables. The dependent variable here is the YDS.G. It is where the total of yards completed by a quarterback per game and the other 11 variables contributes to it. In the figure shown we can see which has the highest relationship with YDS.G. A total of 7 variables are above 0.8 which shows a strong correlation with the dependent variable, these are: INT, SACK, ATT, COMP, YDS, TD and LONG. This is quite unusual as the variable YDS.A (yards attempt) and PCT (percentage) are both very low, we would normally assume that if a quarterback attempted a lot of pass it should equal to more yards gained per game and same goes with completed pass percentage.

We can see a strong correlation between interception and sack, this can suggest that when every time a quarterback is taken to the ground there is a high chance of a lose ball and may end up in the hands of the other team which will result in an interception. Again, interception is highly correlated with pass attempts, completed pass attempts, touchdowns and long passes. The higher times the QB passes the more chance it can get intercepted, especially long passes which allows the ball in the air for a long time giving the cornerbacks enough time to position themselves for an interception.

Another strong correlation is the touchdown variable with ATT, COMP and YDS that are above .90, this suggest that with every pass attempts and completed pass will eventually result in a touchdown for the team. Same goes with the yards, if the yards are high it means that a lot of the passes are completed and attempted.



6.1.4 Data Mining

Support Vector Regression

the library “e1071” was used for the machine learning algorithm Support Vector Machine type = regression. Support Vector Regression produces 4 kernels in which we can run side by side to see which kernel produces the best result in terms of accuracy and low error rate. There are four types of kernels used in the regression analysis: linear, polynomial, radial and sigmoid.

SVR difference with simple regression model is that in a simple regression model the task is to minimise the error rate so that the outcome of the model will have a better accuracy than before but in SVR these errors are within a threshold known as the margin inside the boundary line.

The points inside the boundary lines are points which has least error rates and thus by doing only taking the points that are near the hyperplane and within the boundary lines, it will give us a better fitting model. The code used to run the SVM function is displayed in [figure](#) , here we are changing the type of the model from classification to regression by putting the type = "eps-regression". To use different kernel the code kernel = "linear" was used for the four types of kernel to be run.

```
svm1 <- svm(YDS.G ~ ., data = traindata, type = "eps-regression", kernel = "linear") #
svm2 <- svm(YDS.G ~ ., data = traindata, type = "eps-regression", kernel = "polynomial") #
svm3 <- svm(YDS.G ~ ., data = traindata, type = "eps-regression", kernel = "radial") #
svm4 <- svm(YDS.G ~ ., data = traindata, type = "eps-regression", kernel = "sigmoid") #
```

6.1.5 Interpretation/Evaluation

Secondly, we are to predict the yards per game for the 565 quarterback players (duplicate players are allowed in different years) from 2008 to 2017 by running the function predict see [figure](#) . this will allow us to produce a result and compare the SVR train data model to the untouched test data.

```
# predictions of the four kernels
predlinear <- predict(svm1, testdata)
predpoly <- predict(svm2, testdata)
predradial <- predict(svm3, testdata)
predsigmoid <- predict(svm4, testdata)
```

The following result are gathered upon running the predict function in [figure](#) . Below are the 4 kernels with there prediction result. As of right now, we cannot determine the accuracy of each of them. We will need a code for assessing the R2 squared, RMSE (Root mean squared error) and MAE (Mean Absolute Error). To get this value we will run a code in [figure](#)

```
4      15      25      26      34      39      50      61      68      72      75      76
302.423212 217.289007 178.008818 204.583570 180.811491 144.338863 128.336977 55.627742 -15.954102 35.160368 -18.659926 29.789755
89      93      98      99      103      106      110      115      121      124      126      127
-36.617532 -21.251368 328.740669 274.465879 269.042399 255.299017 219.041501 228.418004 212.275368 166.863363 194.530043 181.460977
135      142      147      152      163      167      198      201      210      216      223      236
128.508225 91.915411 106.324209 72.281144 43.484001 34.700233 279.281245 302.400022 263.149832 200.599329 192.103956 127.573480
237      239      244      250      257      259      285      292      295      302      306      313
145.069468 160.648386 113.322188 88.436626 92.745432 59.241980 283.518455 243.108063 236.527670 218.543217 237.838507 184.887878
316      322      325      326      331      337      341      342      346      354      395      398
146.772920 95.118539 145.905031 124.343616 83.203881 83.741387 44.174105 33.883464 52.579389 10.557471 238.067576 245.949712
399      404      406      413      416      419      420      421      427      429      430      435
228.473488 197.546511 226.915798 193.187879 175.385376 156.428405 159.710871 157.292202 150.637677 106.069919 125.005254 92.680450
438      471      486      487      497      501      504      506      508      511      512      514
73.220652 274.328606 246.495543 222.115740 159.670572 184.736212 151.373267 176.648111 165.333536 104.182820 101.542230 94.082632
523      524      537      542      554      560      562      566      570      572      584      600
62.794651 40.819432 41.258927 -14.151432 309.579548 266.031784 272.699866 263.758850 237.988081 222.736423 180.512386 119.771405
602      606      608      613      615      616      618      629      652      677      687      689
90.973741 131.448842 110.284194 55.946363 66.027512 66.404920 32.839621 -8.009598 304.827718 194.245164 156.984800 92.603352
695      696      698      700      709      728      756      758      774      779      782      784
119.124420 59.642347 85.745664 74.118318 79.312159 12.005137 281.393939 291.134263 191.773282 159.361897 175.558973 147.451375
787      792      794      796      805      808      832      834      837      838      861      862
103.270890 136.201877 137.862041 49.867591 61.382582 87.319234 -33.408489 -42.592036 -11.913015 -21.516789 311.513281 288.439815
870      873      874      877      884      899      909      917      919      929      960
222.589254 234.886335 239.298317 192.654838 192.263309 127.946436 82.277387 128.241108 55.542625 -69.450998
```


4	15	25	26	34	39	50	61	68	72	75	76	89	93
183.8233	183.7459	183.4899	183.5120	183.4356	183.4319	183.2986	183.1211	182.5391	182.8382	182.2267	182.8729	180.4682	180.9932
98	99	103	106	110	115	121	124	126	127	135	142	147	152
184.1668	184.2661	184.0927	183.9078	183.8239	183.6341	183.5781	183.4568	183.4348	183.4269	183.4065	183.3615	183.2739	182.9525
163	167	198	201	210	216	223	236	237	239	244	250	257	259
182.6938	181.7133	184.7044	184.0215	183.7344	183.4709	183.4495	183.3547	183.3968	183.3671	183.2365	183.2559	183.1603	183.1137
285	292	295	302	306	313	316	322	325	326	331	337	341	342
184.3120	183.8226	183.7022	183.6046	183.4783	183.4418	183.4345	183.3725	183.4014	183.3299	183.2178	183.2231	183.1404	182.8632
346	354	395	398	399	404	406	413	416	419	420	421	427	429
183.1363	182.9293	183.8112	183.6133	183.5842	183.5637	183.4776	183.4357	183.4364	183.4343	183.4293	183.4323	183.3496	183.2496
430	435	438	471	486	487	497	501	504	506	508	511	512	514
183.2423	183.0804	183.1161	184.2486	183.6313	183.6054	183.4352	183.4372	183.4302	183.3469	183.3856	183.3464	183.3275	183.2341
523	524	537	542	554	560	562	566	570	572	584	600	602	606
183.1667	182.7573	182.9276	182.8418	184.8917	183.9620	183.9865	183.6633	183.5771	183.5794	183.4361	183.3733	183.2830	183.1665
608	613	615	616	618	629	652	677	687	689	695	696	698	700
183.2863	183.1525	183.2373	182.9602	183.1464	182.2328	184.7157	183.4771	183.4269	183.4112	183.3743	183.3490	183.3908	183.3532
709	728	756	758	774	779	782	784	787	792	794	796	805	808
183.1542	182.9224	184.2746	184.1298	183.5209	183.4359	183.4427	183.4544	183.4272	183.3702	183.4123	183.3662	183.2923	183.1478
832	834	837	838	861	862	870	873	874	877	884	899	909	917
183.0366	182.0697	182.8170	181.6579	184.4682	184.2628	183.6656	183.5523	183.7077	183.5476	183.4429	183.4186	183.2093	183.0627
929	960												
182.1773	180.8682												

4	15	25	26	34	39	50	61	68	72	75	76		
254.070258	242.661294	192.301674	210.259933	162.098123	136.040008	84.705010	48.618640	41.976552	27.637147	19.702187	22.489223		
89	93	98	99	103	106	110	115	121	124	126	127		
-5.029878	-2.890856	277.656293	272.386956	261.478869	252.072442	241.755392	228.895383	217.934974	186.927516	151.393003	147.986823		
135	142	147	152	163	167	198	198	201	210	216	223		
128.160622	96.084803	73.907963	51.216302	28.557352	19.251521	283.664860	267.651903	246.590542	197.401891	195.696588	101.232535		
237	239	244	250	257	259	285	285	292	295	302	306		
114.592157	103.588621	73.688173	82.582423	52.652380	42.629782	280.184432	254.120253	236.409523	231.029594	203.798750	165.932096		
316	322	325	326	331	337	341	342	346	354	395	398		
144.497023	103.268678	115.547695	92.249730	72.202923	66.475315	59.073112	41.866097	59.524189	29.728150	244.290935	226.273244		
399	404	406	413	416	419	420	421	427	429	430	435		
222.865116	222.330662	207.359210	168.100503	172.201713	149.623906	142.695277	137.391770	113.465783	96.647718	76.436386	73.349263		
438	471	486	487	497	501	504	506	508	511	512	514		
55.246878	277.612796	233.184800	226.040626	159.848057	173.398868	137.712536	106.170357	125.733270	94.081058	80.331939	68.525361		
523	524	537	542	554	560	562	566	570	572	584	600		
67.147077	50.467574	31.503776	23.205438	296.496624	267.157197	260.441094	245.384588	219.314806	216.097688	163.938875	105.040237		
602	606	608	613	615	616	618	629	652	677	687	689		
87.531690	69.924601	76.787626	63.948348	63.182372	49.647543	48.253150	19.330828	289.910182	199.970715	134.495335	117.735860		
695	696	698	700	709	728	756	758	774	779	782	784		
107.135424	94.522471	106.070029	92.462868	58.911839	22.930269	268.391448	271.139582	211.881185	173.393004	182.521806	167.062437		
787	792	794	796	805	808	832	834	837	838	861	862		
138.593345	121.682102	126.383470	105.242645	74.202616	67.005151	27.960245	5.361290	4.332054	4.057178	293.854232	280.646887		
870	873	874	877	884	899	909	917	929	960				
227.661880	230.691242	236.559637	217.576721	192.570475	121.926853	77.082746	55.998353	32.029481	-10.192140				

4	15	25	26	34	39	50	61	68	72	75			
246.0590375	241.0507964	191.4804919	204.9089717	158.1090087	132.7768044	84.3742577	51.9053901	45.8494061	33.7086984	24.8536501			
76	89	93	98	99	103	106	110	115	121	124			
29.3322945	0.2707123	1.6373884	275.9930621	272.9706171	262.2548186	255.0927686	241.1384729	224.5390952	215.4397994	184.3605610			
126	127	135	142	147	152	163	167	198	201	210			
144.9486241	143.7322108	127.8750352	99.3034736	76.5539821	53.1972973	33.7222353	23.4091894	294.6329277	263.8514121	243.0509812			
216	223	236	237	239	244	250	257	259	285	292			
190.2609316	188.3182665	97.7044419	111.6966799	100.9340814	79.2775591	82.4326564	57.2198581	49.2981159	279.7850461	249.3469201			
295	302	306	313	316	322	325	326	331	337	341			
234.7242247	226.5343693	196.8117717	164.3283436	137.6056697	103.0446332	112.8128537	89.6072725	70.8889325	67.7999615	61.4162807			
342	346	354	395	398	399	404	406	413	416	419			
43.8966624	62.1161219	36.1873723	244.5802492	222.6831810	218.8446523	218.9849768	200.4092891	161.0982097	167.1129272	144.8141525			
420	421	427	429	430	435	438	471	486	487	497			
139.0673877	133.0299130	110.1936603	93.2756262	75.3726204	85.6252747	57.1878628	278.5542653	229.7021800	225.2640850	155.4277816			
501	504	506	508	511	512	514	523	524	537	542			
168.3005992	134.5621192	102.6228216	127.1334313	93.2283705	83.2554028	70.0137924	69.3262880	74.1945757	37.0809154	43.5804877			
554	560	562	566	570	572	584	600	602	606	608			
304.6643831	263.0232239	263.2228850	239.8201218	215.2949274	215.5053028	158.2040746	105.9567182	85.8342368	69.8504586	81.1759523			
613	615	616	618	629	652	677	687	689	695	696			
66.7010382	68.6448327	51.7697798	55.5581603	24.2380379	295.7886489	194.5323125	130.6297310	116.3639981	105.4739321	94.6225708			
698	700	709	728	756	758	774	779	782	784	787			
106.6390022	92.5491830	62.0463732	30.8947477	274.2406390	266.9476482	208.7189011	166.8053297	179.6913277	167.6999317	137.5323853			
792	794	796	805	808	832	834	837	838	861	862			
118.4695587	125.4232918	103.2262967	78.2965576	67.5728719	38.3620402	12.5224249	18.4854649	10.2468065	291.9562105	279.0592495			
870	873	874	877	884	899	909	917	929	960				
225.6778320	221.2566978	234.7431689	210.5828994	182.7702353	118.5381295	75.7654823	57.0945066	34.6332848	-4.4885345				

```

val_linear <- data.frame( R2 = R2(predlinear, testdata$YDS.G),
                        RMSE = RMSE(predlinear, testdata$YDS.G),
                        MAE = MAE(predlinear, testdata$YDS.G))

val_poly <- data.frame( R2 = R2(predpoly, testdata$YDS.G),
                      RMSE = RMSE(predpoly, testdata$YDS.G),
                      MAE = MAE(predpoly, testdata$YDS.G))

val_radial <- data.frame( R2 = R2(predradial, testdata$YDS.G),
                        RMSE = RMSE(predradial, testdata$YDS.G),
                        MAE = MAE(predradial, testdata$YDS.G))

val_sigmoid <- data.frame( R2 = R2(predsigmoid, testdata$YDS.G),
                          RMSE = RMSE(predsigmoid, testdata$YDS.G),
                          MAE = MAE(predsigmoid, testdata$YDS.G))

```

Upon running the code in figure below, the linear kernel produced the best R2 squared value of .82 which translates to 82% of the model were successfully predicted with the Support Vector Regression kernel type linear. This prediction was compared to the test data of 25% (142 rows) of the total number of rows from the original dataset. It also gave the lowest RMSE of 40.5 and MAE value of 32, see figure below. The first kernel is linear and the last is sigmoid.

```

> val_linear #gives the most accurate model from the four kernel and lowest rmse
      R2      RMSE      MAE
1 0.8274123 40.51164 31.91519
> val_poly
      R2      RMSE      MAE
1 0.5842866 102.2077 84.40637
> val_radial
      R2      RMSE      MAE
1 0.8194728 42.28456 30.26959
> val_sigmoid
      R2      RMSE      MAE
1 0.8047312 43.97816 31.83158

```

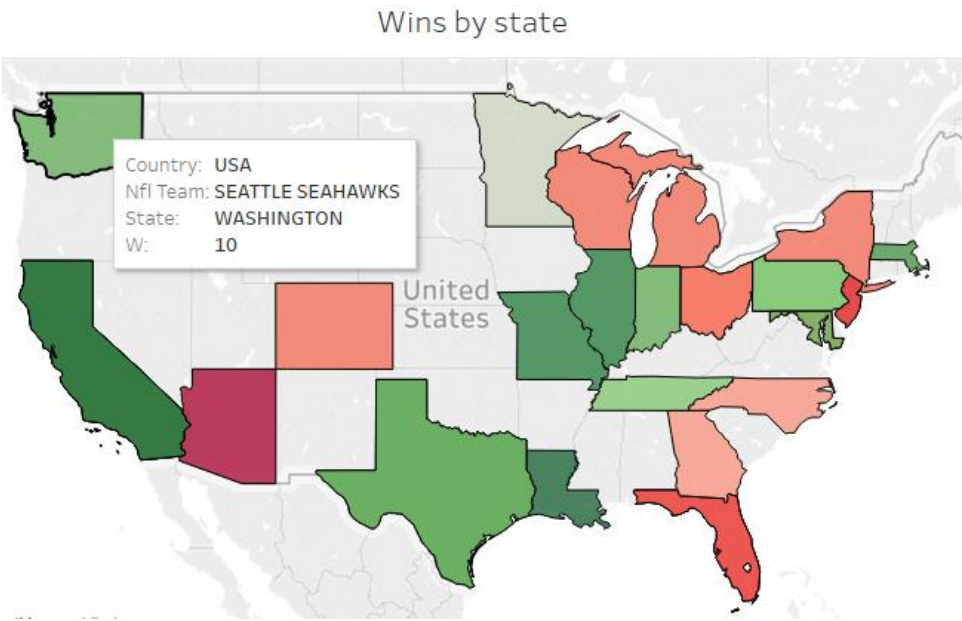
7 Interface Requirements

This section will describe how the software interfaces with other software used in the project and the input and output of users interacting with it. This will include interfaces such as data files, data visualization tools, programming software, data streams, graphical user interface and so forth.

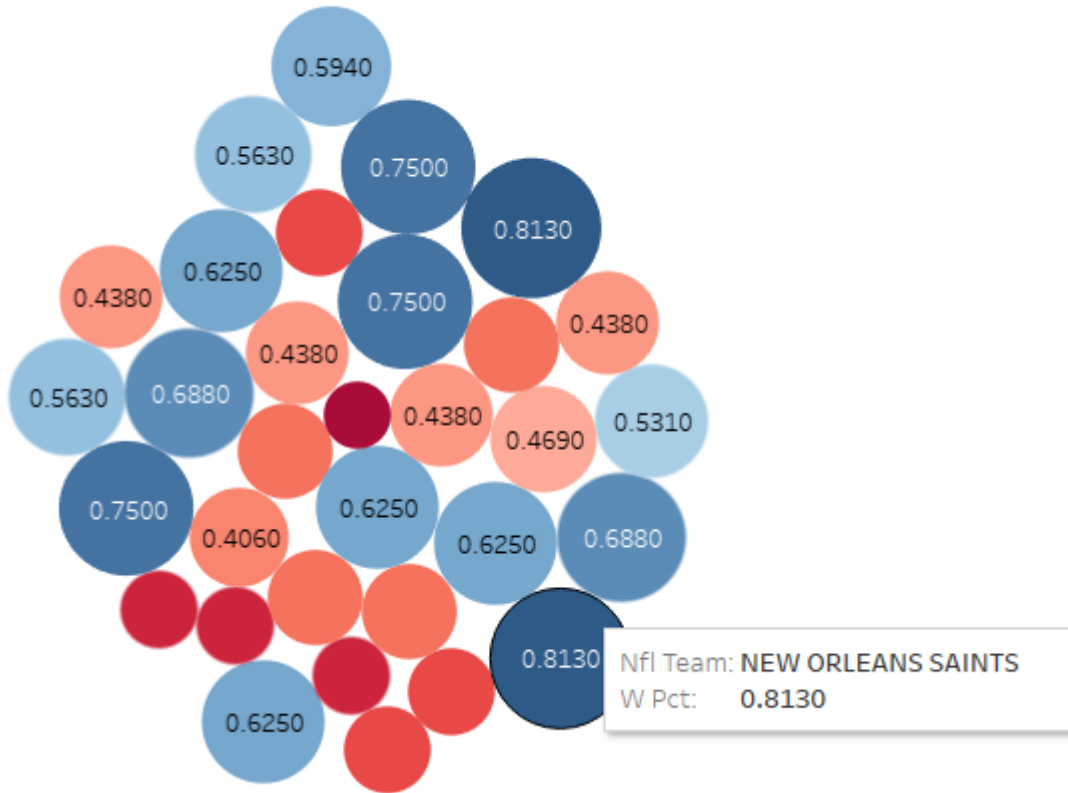
7.1 Graphical User Interface (GUI)

Tableau is a software application that is mainly used for fast analytical business intelligence. The Tableau Desktop is a data visualization application which lets user analyse structured data, combine them and to freely manipulate their columns produce a highly interactive dashboards, diagrams and reports in a quickly manner. Below is an example of a dashboard in Tableau displaying the win percentage of each team, their wins according to their region and their Net points by team.

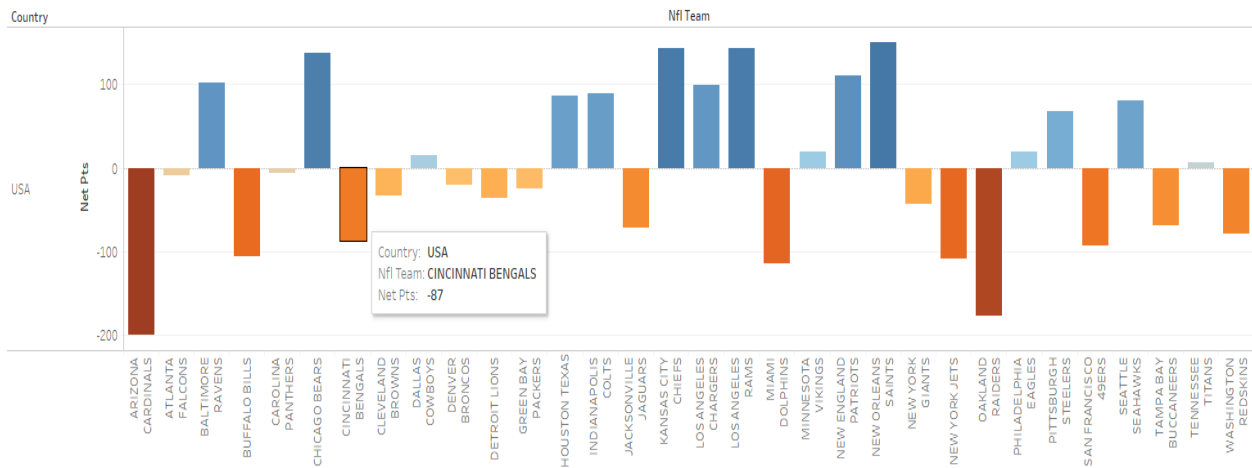
By using the Tableau software, it offered a user-friendly interactive dashboard that are easy and simple to use.



Win percentage by team

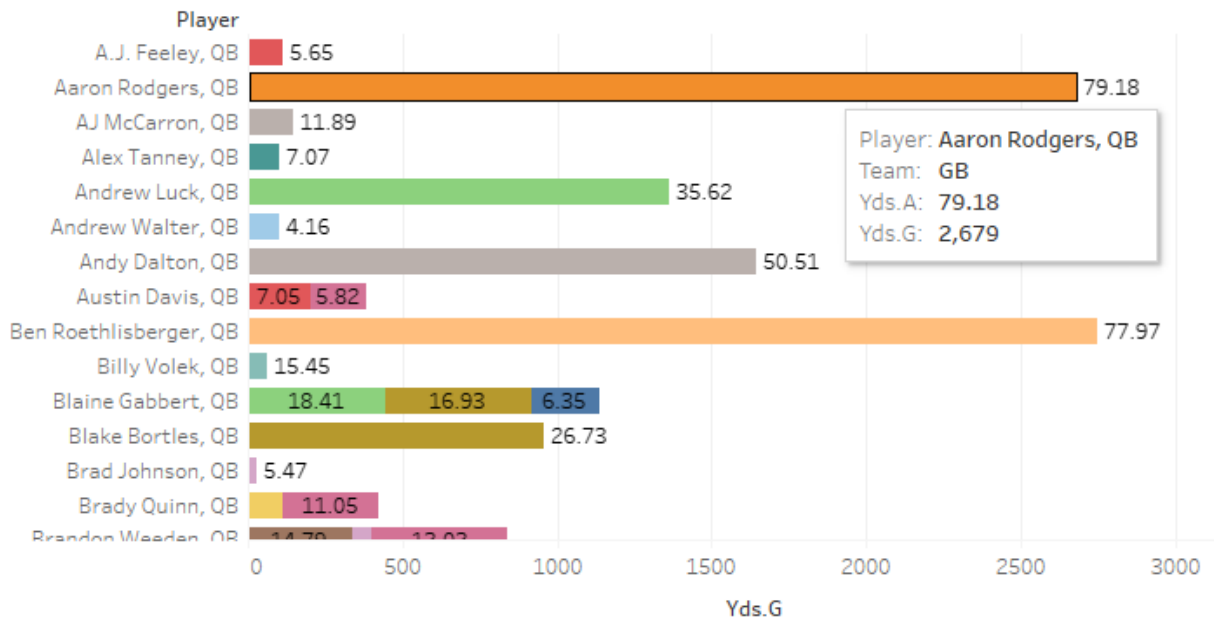


Net Points per team



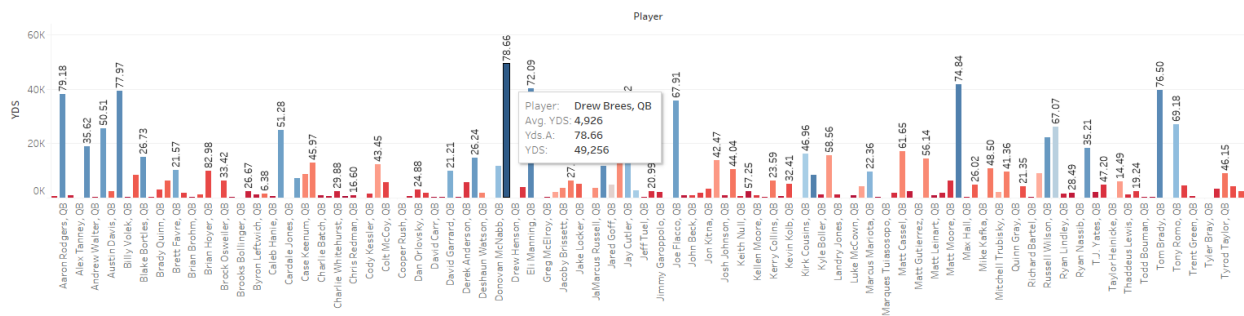
Another tableau dashboard was developed for the Quarterback's performance in a 10 years timeframe from 2008 to 2017. In **figure**, it displays the players yards per game with teams and yards per pass attempt. The different colours in one bar chart is an indication that the following player played on multiple teams, hence the different numbers inside a single bar chart.

Yards Per Game with Teams and Yards Per Pass Attempt



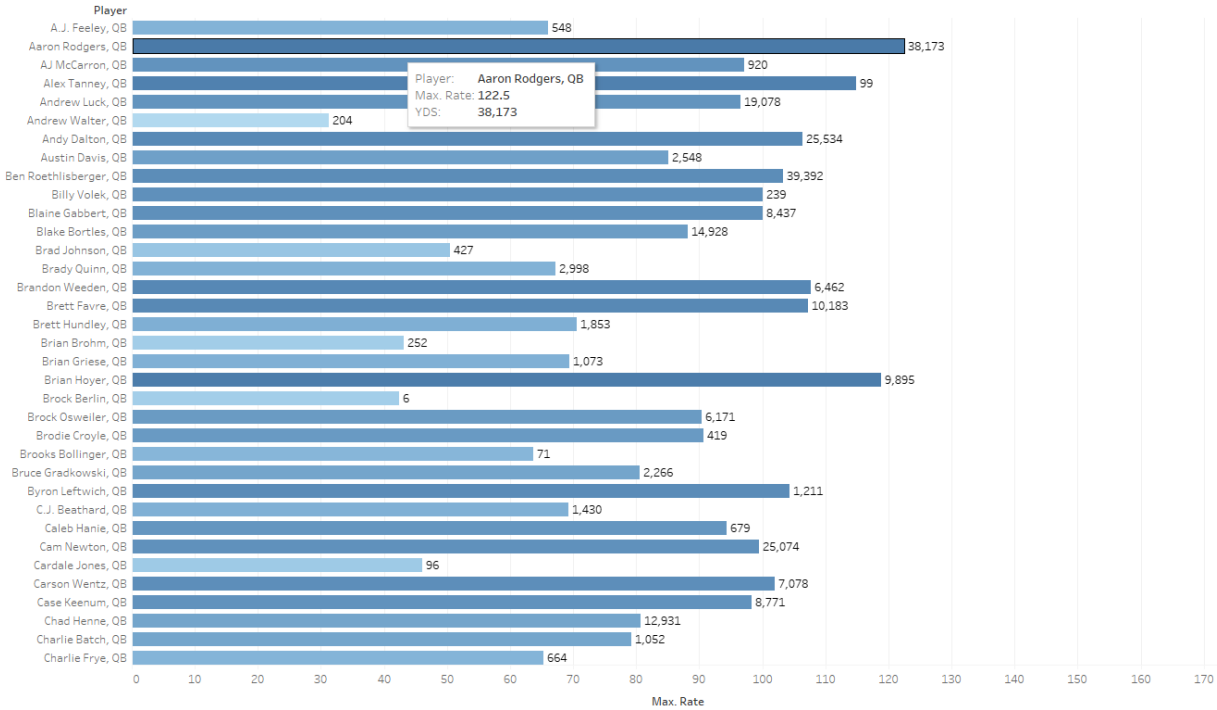
In **figure**, the dashboard is about average yards with pass attempts and total yards. The darker the blue colour suggest a strong/high average passing yards per player and the darker the red colour suggest a very low and weak average yards. It shows that the quarterbacks that has a high amount of passing yards achieves a high average of pass yards.

Average Yards with Yards Attempt & Total Yards



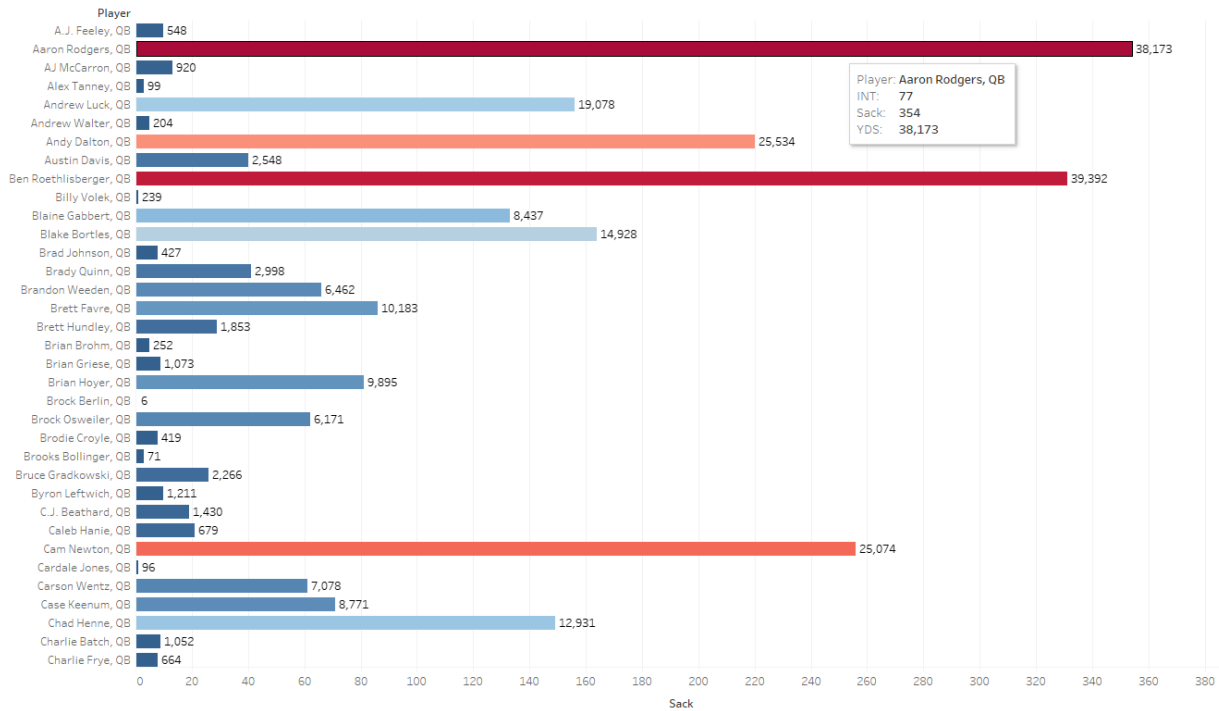
In figure , it displays the quarterback's maximum rating during the a maximum of 10 years in playing in the league together with their total yards achieved, note that not everyone of the quarterback listed has played for 10 years.

Maximum Rate with Total Yards



The last figure, shows the recorded Sacks and Interception of the quarterbacks with their total yards.

Sacks and Interception



By using Tableau, it will give the user a quick information about the player and leagues performance.

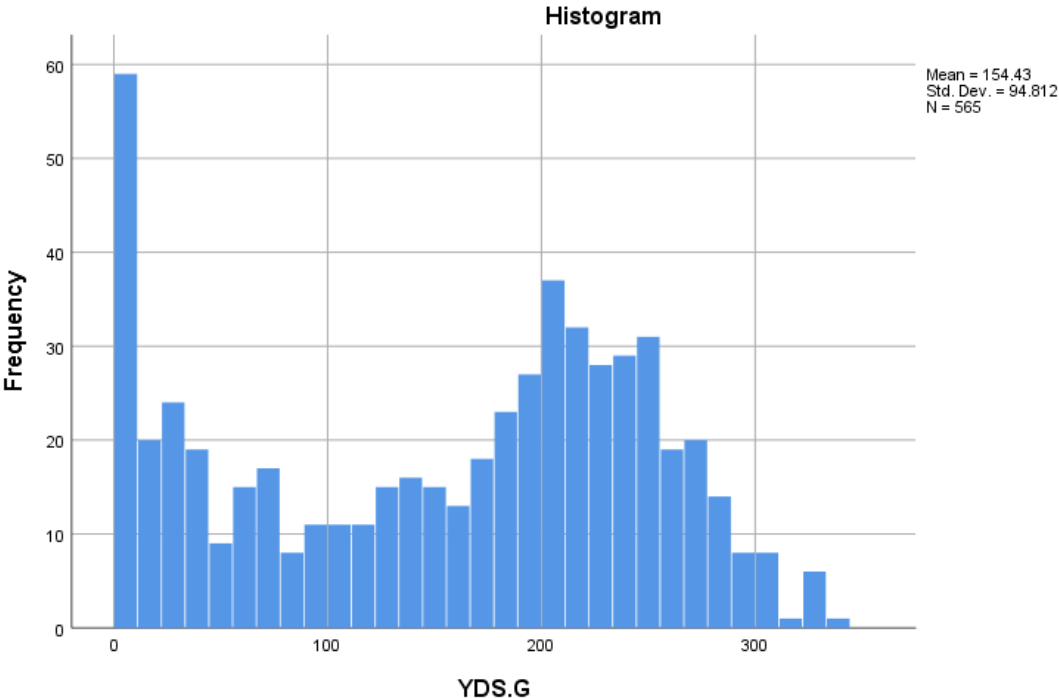
7.2 Statistical Testing

The dependent variable YDS.G was tested using the Shapiro-Wilk test in SPSS to determine its normality. The following alpha of 0.05% was used for this experiment.

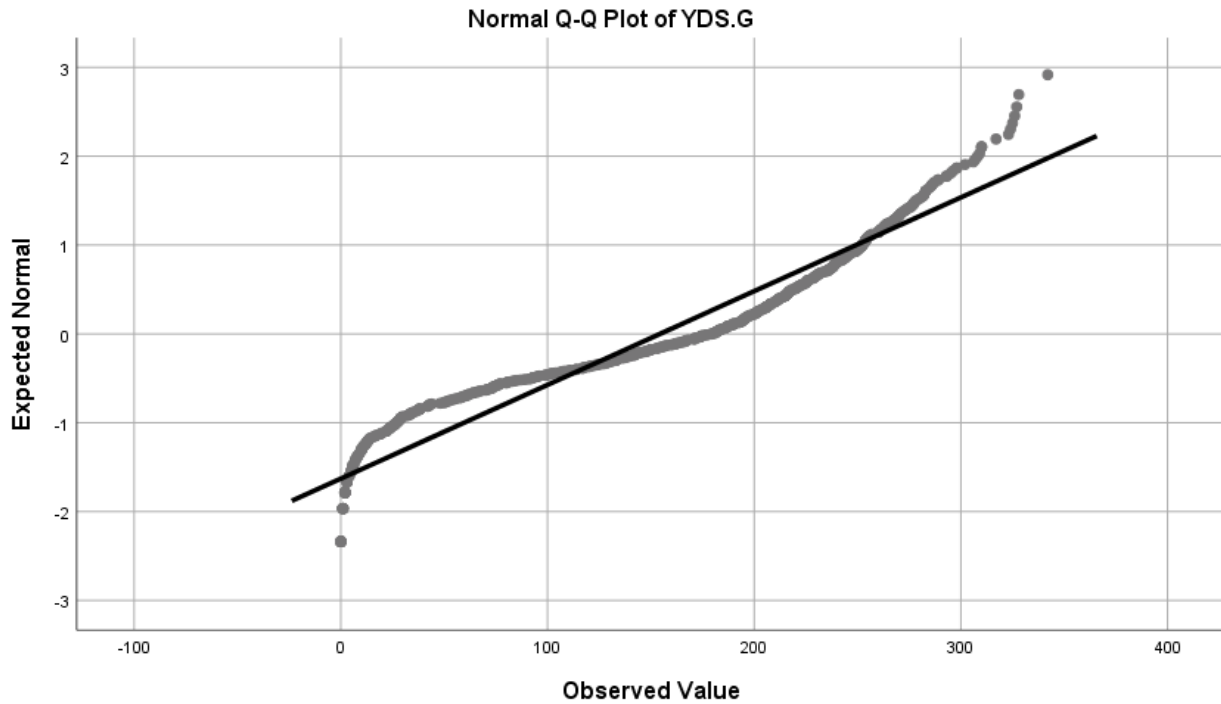
Null hypothesis: is that the dependent variable YDS.G is normally distributed.

Alt hypothesis: is that the dependent variable YDS.G is not normally distributed.

In this figure we can see that the distribution is slightly positive skewed or slightly right skewed. This is a first proof of a non-normal distribution.



Another proof of a non-normal distribution is the QQ plot. The dots doesn't follow the line hence it will not be a normal distribution.



A Shapiro-Wilk test was conducted to test if the variable YDS.G is a normal dataset.

The alpha value used for this test is 0.05, which means that I will allow a 5% risk of a failure to occur, in other words having a type 1 error that states that we are rejecting the null hypothesis when in fact the null hypothesis is true.

In this test we are rejecting the null hypothesis that states that the data is normally distributed. The p is below 0.05 which means we can reject the null hypothesis.

Tests of Normality

	Kolmogorov-Smirnov ^a			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
YDS.G	.108	565	.000	.931	565	.000

a. Lilliefors Significance Correction

8 References

- Sporting Charts (2018) *National Football League – NFL*. Available at: <https://www.sportingcharts.com/dictionary/nfl/national-football-league-nfl.aspx> [Accessed 10 December 2018]
- Dummies *The National Football League Conferences*. Available at: <https://www.dummies.com/sports/football/the-national-football-league-conferences/> [Accessed 10 December 2018]
- History of the NFL. *Background Information of the NFL*. Available at: <https://sites.google.com/site/barthfamilysportspicks/background-information> [Accessed 11 December 2018]
- Dataset - <https://www.kaggle.com/jerrinv/nfl-offense-cleaned-2017to2007>
- Howstuffworks (2018) *How American Football Works* Available at: <https://entertainment.howstuffworks.com/football.htm> [Accessed 9 December 2018]
- NFL (2018) *Brady to have season-ending knee surgery, will be placed on IR*. Available at: <http://www.nfl.com/news/story/09000d5d80a95089/article/brady-to-have-seasonending-knee-surgery-will-be-placed-on-ir> [Accessed December 11 2018]
- <https://www.kaggle.com/jerrinv/nfl-offense-cleaned-2017to2007> – nfl quarterback offense
- ESPN (2017) *NFL Player Passing Statistics – 2017*. Available at: http://www.espn.com/nfl/statistics/player/_/stat/passing/sort/passingYards/year/2017 [Accessed 6 May 2019]
- Coinmonks (2018) *Support Vector Regression*. Available at: <https://medium.com/coinmonks/support-vector-regression-or-svr-8eb3acf6d0ff> [accessed at 5 May 2019]
- Stack Overflow (2018) *R, remove row if there is a certain character [duplicate]*. Available at: <https://stackoverflow.com/questions/34007025/r-remove-row-if-there-is-a-certain-character?noredirect=1> [accessed 5 may 2019]

9 Appendix

9.1 Definitions, Acronyms, and Abbreviations

NFL - National Football League. A league in the sports of American Football. It consists of a total of 32 teams split into two conferences, the NFC and AFC which consists of 16 teams each conference.

SVM – Support Vector Machine are supervised learning models with associated learning algorithms to analyse data used for clarification and regression analysis

GUI – Graphical User Interface such as tableau, shiny and this will allow me to display my analysis results in a dashboard.

SPSS – Statistical Package for Social Science. SPSS Is a software that is mainly used for interactive statistical analysis.

ANOVA – Analysis of variance a statistical model for testing out if there is a significant different between means.

KDD – Knowledge Discovery and Data mining. It is a methodology or processes used in data mining to finding knowledge within a data by using data mining algorithms.

SEMMA – is an acronym for Sample, Explore, Modify, Model and Assess.

APFA – American Professional Football Association. Before the NFL was created there was the APFA for the sports of American Football

NFC - National Football Conference.

AFC – American Football Conference

YPA – Yards per attempt

AI – Artificial Intelligence

CSV – a file format that is used to store data in the form of spreadsheet or database.

COMP – Completions

ATT – Passing Attempts

PCT – Completion percentage

YDS – Passing yards

YDS/A – Yards per pass attempt

LONG – Longest Pass

TD – Passing Touchdowns

QB – Quarterback

INT – Interceptions

SACK – Sacks

Rate – Passer rating

YDS/G – Passing yards per game

