**A STATISTICAL ANALYSIS ON FANTASY FOOTBALL PREMIER LEAGUE (2018/2019) POINTS AWARDING PATTERN**

***UNDERTAKEN***

***BY***

**MICHAEL GBENGA ADEKOYA**

**21147079**

***SUBMITTED***

***TO***

**DR. ATIF AZAD**

**MODULE COORDINATOR**

**APPLIED STATISTICS (CMP7205)**

**BIRMINGHAM CITY UNIVERSITY**

**DECEMBER 13, 2021**

**Table of Contents**

Cover page

Table of contents 1

Executive Summary 2

Section One (1)

1.0 Introduction 3

1.1 Purpose of Report 3

1.2 Problem Domain 3

1.3 Statistical Questions 4

Section Two (2)

2.0 Methodology 7

2.1 Statistical techniques and Justification. 7

Exploratory Data Analysis 7 Correlations Analysis 7

Regression analysis 7

2.2 Justification of Statistical technique for Football Fantasy 7

2.3 Tool (Software) Adopted for Research 8

R programming language 8

Section Three (3)

3.0 Datasets 9

3.1 Fantasy Premier League Dataset Description 9

3.2 Data Preprocessing 10

Section Four (4)

4.0 Result and Discussion 12

4.1 Result 12

Section Five (5)

5.0 Conclusion and Recommendation 24

5.1 Conclusion 24

5.2 Recommendation 24

References

Appendix

**Executive Summary**

This report is an analysis on Fantasy Football Premier League (2018/2019) Season. In this report, three (3) statistical techniques were adopted to critically observed how points is awarded to participants in the Fantasy Football League, Exploratory Data Analysis for observing how data is distributed, Correlation analysis for observing similarity and relationship while Regression analysis was used to predict likely outcome. The report is divided into five (5) sections, with each section having its own purpose. Section one, introduces the domain and problem statement and research questions, while section two, introduces the datasets description and structure, section three, presented the methodology and section four (4) presented the Results and discussion. Section five, contains the conclusion and findings based on analysis and a recommendation. The reference and Appendix came next in the last part of the report.

**Section One (1)**

**1.0 Introduction**

This is an assessment under the module, applied statistics in Big Data Analytics course at Birmingham City University meant to mirror the typical task of data analysis as observed in any organization.

As with most organization, there is always a constant need to make meaning out of available data in other to make critical business decision that is important to weather the growth, revenue, profit margin maximization or customer segmentation depending on the KPI (Key performance indicators or metrics) that drives their business. Hence, this assessment will be in form of a statistical report with necessary information presented as simple as possible for business users and stakeholders of dataset domain.

**1.1 Purpose of Report**

This report is based on modern statistical methods and tools employed in the analysis of different datasets for the purpose of interpretation, explanation and providing or unlocking hidden insights within selected dataset.

In this report, a sample dataset will be extracted from Premiership Fantasy Football league.

In this report, two variations of same datasets will be examined critically. i.e week 0 and week 11 of the dataset.

**1.2 Problem Domain**

There is only one case study of interest in the research work, which will be briefly explained before presenting the research questions.

**Fantasy Football**

Fantasy Football is basically an online game that involves players been allocated points based on how they perform in real life while they are playing football matches on match week basis. In other words, you get to setup a virtual team of Premier League footballers, and if eventually the played well in real-life, then you earn points.

According *to McCambridge Ed (2021)*, in his explanation how points are earned, its reads below:

“Each position on the pitch comes with its own points system. The goalkeepers and defenders are most commonly rewarded for things such as saving a penalty (5 points) or keeping a clean sheet (4 points); midfielders for setting up a goal (3); and strikers for scoring (4). However, players in different positions can be rewarded for different things and you need to check how much each real-life action is worth.

For instance, a defender can also score in real life. It is rarer, though, and this is reflected in the points (6, instead of 4) a defender will earn if his real-life counterpart scores. A midfielder, meanwhile, also plays a part in defending, and will earn points for that too. They aren't quite as important as the defenders in that regard, however, and only get one point if their team keeps their opponents at bay.

It's also worth noting that players can lose points for negative actions in real-life, such as getting sent off (-3 points) or missing a penalty (-2). Your captain each week will get double points. So if your captain is your striker and he scores, he will get eight points for the goal instead of the usual four. Your vice-captain will take over should your captain not play for some reason.”

In summary, “the game is point based system, that has over 8 million players and arguably the biggest fantasy football game in the world as published by Premier League on their official website and also it is free to play and you can win prices.” *Premier League (2021)*

In view, the researcher’s curiosity comes from the place of what metrics can be used to select or build up your virtual team, are there metrics or attributes that when examined critically would make you earn points, can we even predict the likely wood of performance of our virtual team since we can select football players of our choice? Some of these curiosity will be the driving thoughts behind our formulated research questions.

**1.3 Statistical Questions**

Statistical questions are basically questions that can be answered by collecting different information or raw facts that varies and can sufficiently to a high degree provide a satisfying response to questions needing answers.

**Fantasy Football research questions.**

The questions to be asked for the purpose of investigation on the curiosity raised under fantasy football are;

i. What is the distribution of the categorical variables in the two dataset for week 0?

ii. What is the observation of the univariate variables in week 0 and week 11 of the fantasy league result?

iii. What is the relationship between players Threat and Points awarded in week 0?

iv. What is the relationship between players Creativity and Points awarded week 0?

v. What is the relationship between players Influence and Points awarded in week 0?

vi. What is the relationship between players’ ICT, Bonus and Points Scored in week 0 and 11?

vii. What is the degree of relationship between the attributes in week 11?

viii. Using available metrics, can we predict points and bonus score that can be earned by a player in week 11?

**Section Two (2)**

**2.0 Methodology**

**2.1 Statistical techniques and Justification.**

**i. Exploratory Data Analysis:**

This involves the critical process of performing initial investigation by observing pattern, trends, variance etc about a dataset of concern in other to unravel hidden information through presentation of charts, diagrams, tables which may come as a summarized report.

Exploratory analysis usually deals with using visualizations to tell a statistical or business story that can be understood by business users and stake holders.

**ii. Correlations Analysis**

The usage of correlation analysis or regression analysis depends on your data set and the objective of the study. Correlation analysis is used to quantify the degree to which two variables are related. Through the correlation analysis, you evaluate correlation coefficient that tells you how much one variable changes when the other one does. Correlation analysis provides you with a linear relationship between two variables. Editage Insights (2018)

**iii. Regression analysis** is a related technique to assess the relationship between an outcome variable and one or more risk factors or confounding variables. The outcome variable is also called the response or dependent variable and the risk factors and confounders are called the predictors, or independent variables. Regression analysis is useful when you have to identify the impact of a unit change in the known variable (x) on the estimated variable (y). Editage Insights (2018)

**2.2 Justification of Statistical technique for Football Fantasy**

Considering the available dataset, i.e. fantasy football, the outcome variable is points earned.

Points are of numeric datatype, and other attributes that will aid in finding the relationship between attributes as raised in the research questions are also of numerical type, hence a lot of comparism will be done with the aim to establish a relationship between the attributes if any exist, hence correlation analysis.

Regression on the other hand, will also be employed to find out the outcome of points earned while considering the effects of other attributes that can influence the points earn during the cause of our prediction.

**2.3 Tool (Software) Adopted for Research**

**R programming language**

R is a programming language and free software environment for statistical computing and graphics. It is supported by the R Core Team and the R Foundation for Statistical Computing. It is widely used among statisticians and data miners for developing statistical software and data analysis. Wikipedia (2021).

Furthermore, R is the language in which the module was thought. Although any other modern tool could be used, but the researcher has chosen R.

**Section Three (3)**

**3.0 Datasets**

**3.1 Fantasy Premier League Dataset Description**

The dataset is made up of 21 attributes, while only 3 are categorical datatypes, the remaining are numeric data types.

***Column description***

Name: Shows the name of the selected player in the buildup of virtual team.

Team: Shows the team the selected player belongs.

Position: Shows the position that a selected football player plays.

Cost: This shows the cost of buying a player, unit is given in 0.1 Million pounds.

Creativity: Creativity assesses player performance in terms of producing goal scoring opportunities for others. It can be used as a guide to identify the players most likely to supply assists.

Threat: This is a value that examines a player's threat on goal. It gauges the individuals most likely to score goals.

Influence: Influence evaluates the degree to which a player has made an impact on a single match or throughout the season.

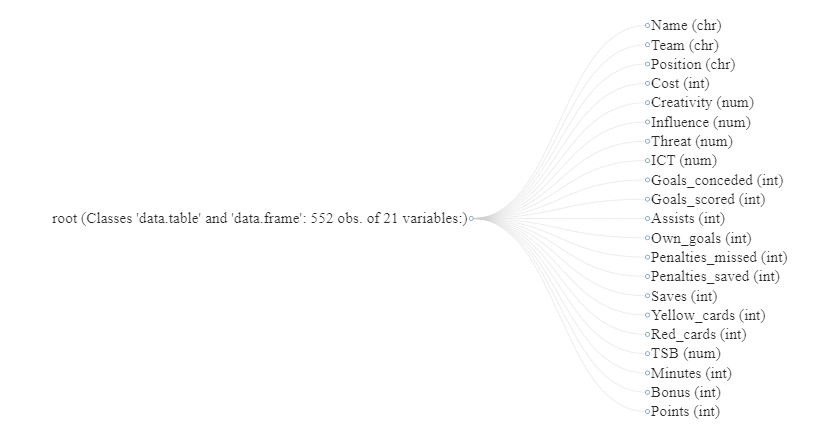
ICT: All three of these scores (Influence, threat, creativity) are combined to create an overall ICT Index score.

Point: This is the value of the point earned at the end of the current week.

Bonus: This is the value assigned when there’s an exceptional performance from a player.

Other columns show all independent attributes that also measures players key performance metrics, before a point or bonus is earned for a certain play week.

In summary, a player’s Creativity, Threat, Influence and ICT will be considered as input variables while Points and Bonus will be the output variable. All research questions presented will be answered by analyzing all input variables and other selected attributes that will aid the result.



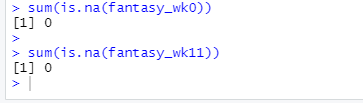
***Figure 3.1: Columns description and properties***

**3.2 Data Preprocessing**

The dataset is a clean dataset which contains a total of 473 players’ performance examined.

The researcher critically observed the following:

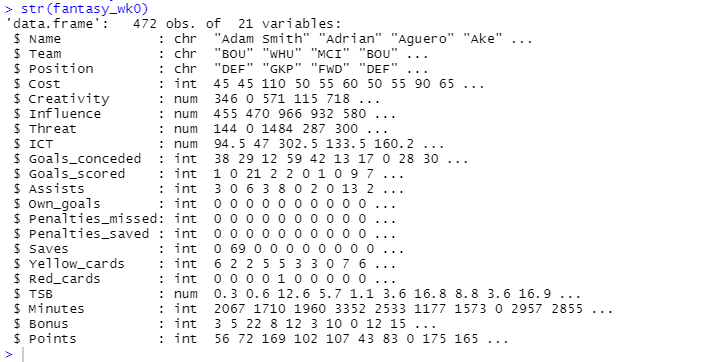
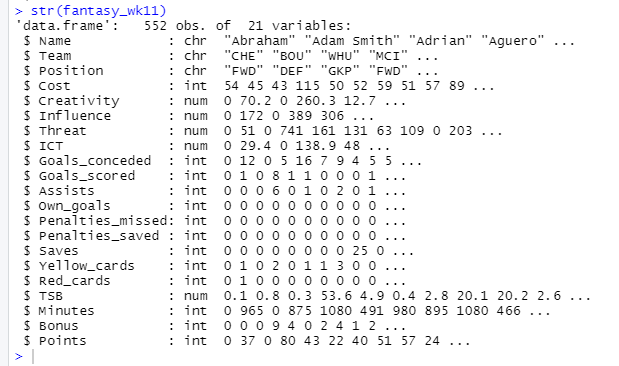
**Test for Missing Values (na)**



***Figure 3.2: Testing for Missing Values***

No missing values was observed in both datasets are seen above in figure 3.1

**Datasets structure for verification**

***Figure 3.3: Snapshot of both datasets***

The snapshot above shows the structure of dataset when loaded into the R studio, with 21 variables each as earlier mentioned and their datatypes. The two datasets were checked against their datatype to ensure they matched.

**Section Four (4)**

**4.0 Result and Discussion**

**4.1 Result**

All the statistical questions will be presented and appropriate statistical findings provided.

***Research Question One***

i. What is the distribution of the categorical variables in the dataset for week 0?

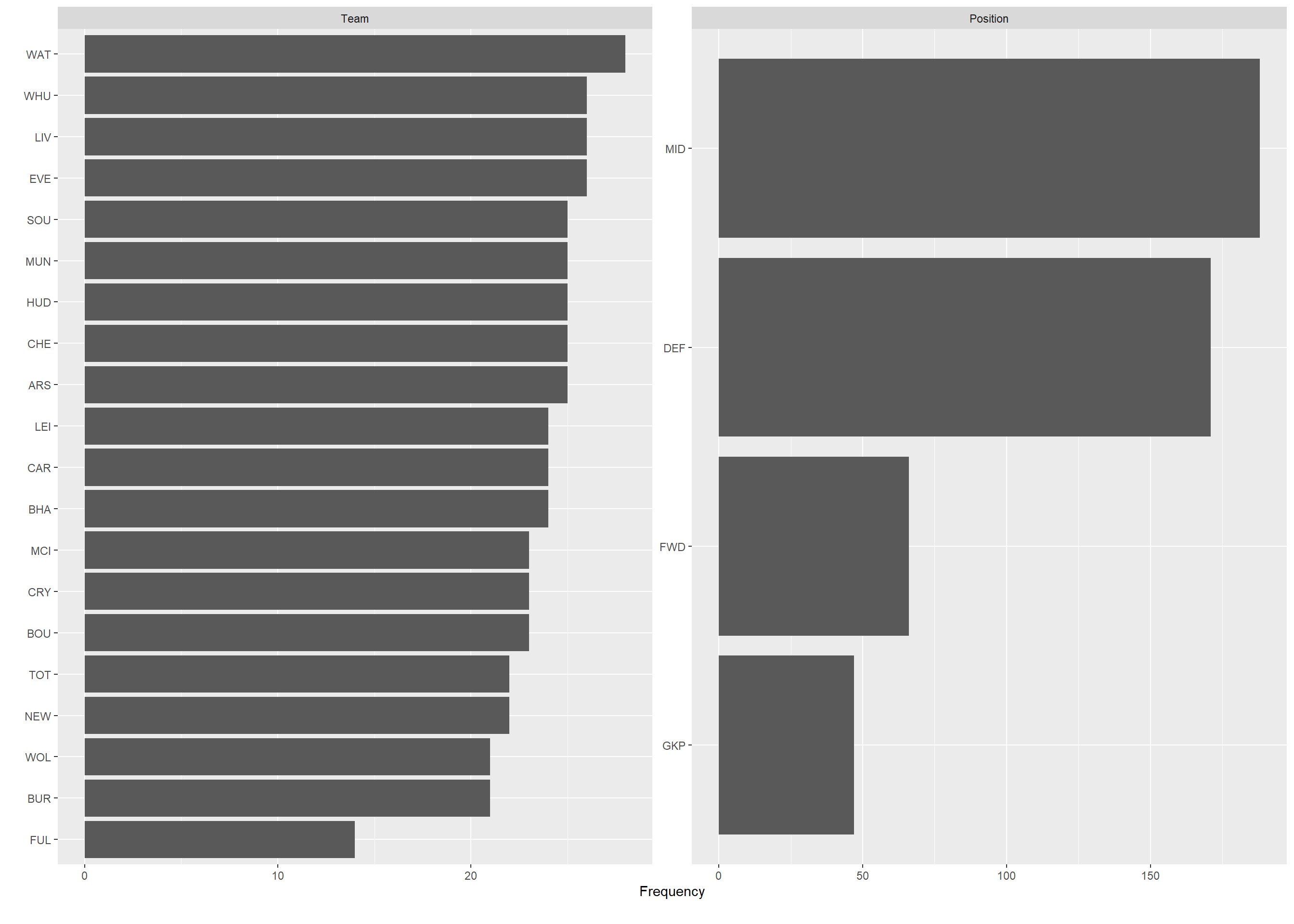


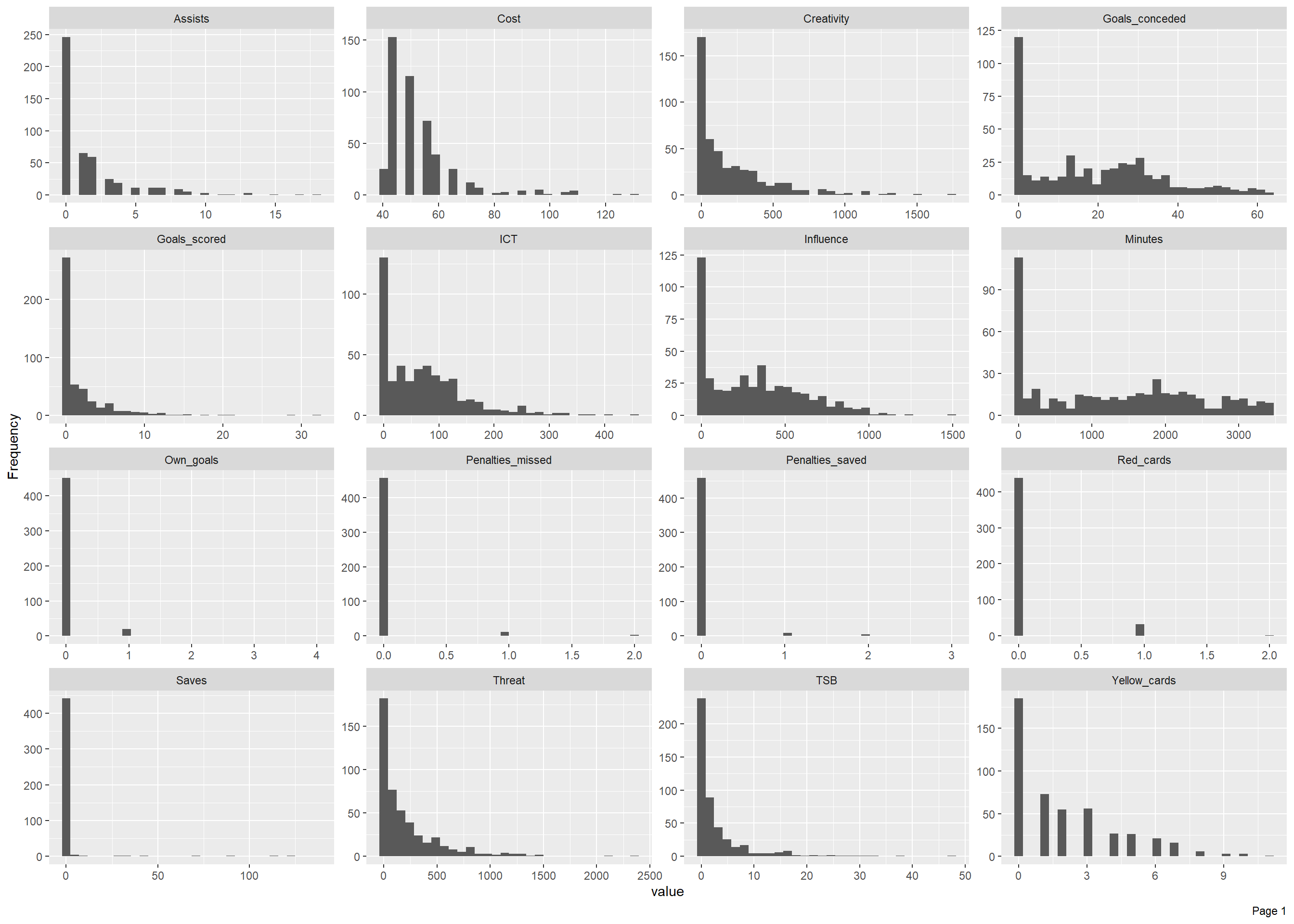
Figure 4.1: Teams and Position Barchart

In figure 4.1 above, the observations in the first chart for teams, indicates that WAT – Watford team has highest frequency in the dataset while Fulham has the lowest occurrence. From the distribution, it is observed that it is multimodal showing different peak of the teams.

On the second chart for players’ position, its indicates MID- Midfield players have the highest occurrence, next is DEF – Defenders, then FWD – Forward and lastly GKP – Goalkeeper in descending order.

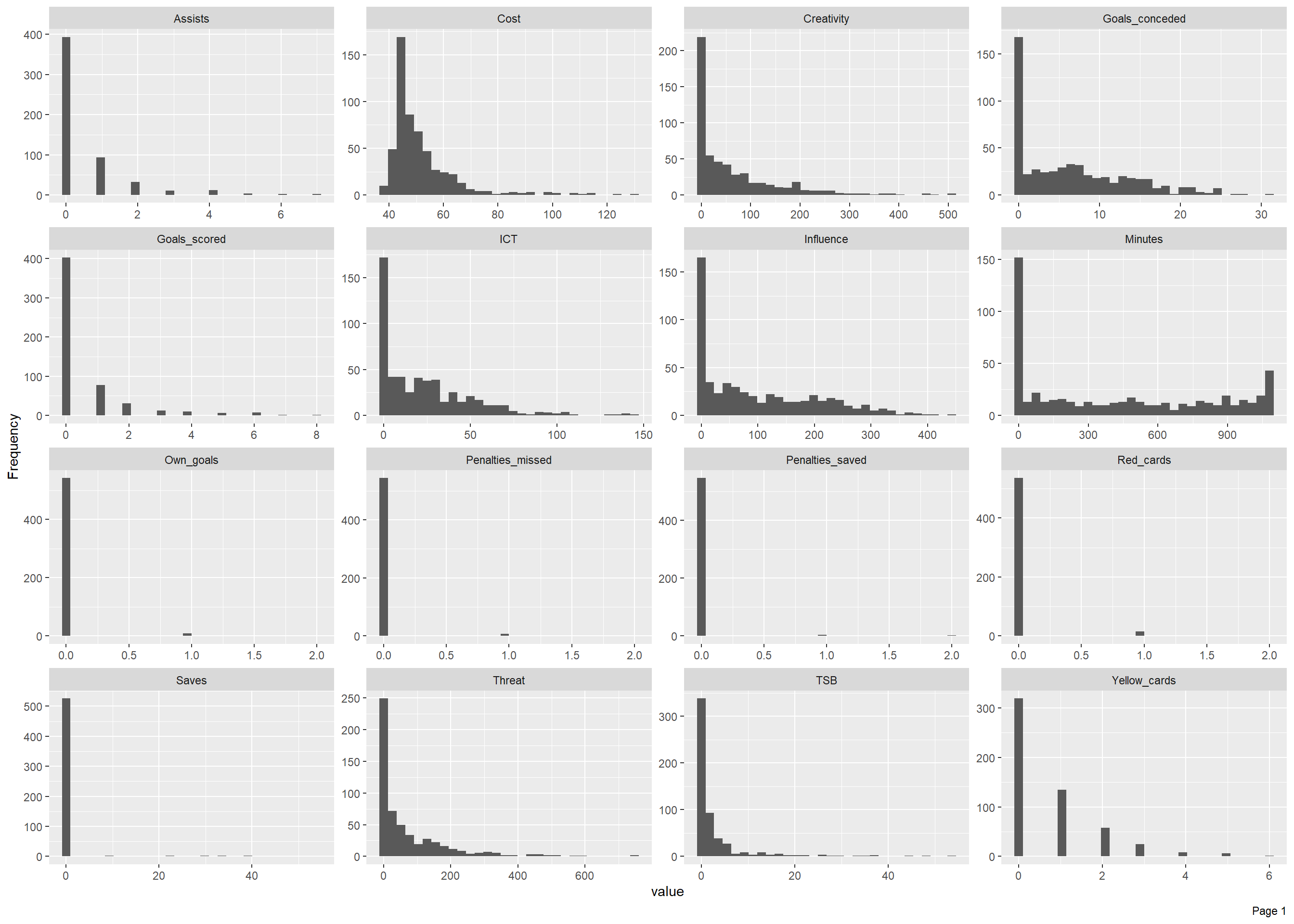
**Research question Two (2)**

ii. What is the observation of the univariate variables in week 0 and week 11 of the fantasy league result?



***Figure 4.2: Observation of Univariate Variables in Week 0.***

The observation from figure 4.2 indicates that ICT, Creativity, Influence and Threat are all left skewed, meaning they all positively skewed and their values of the skewedness is expected to be greater than 0. It further tells us that there are more lower values in the distribution.

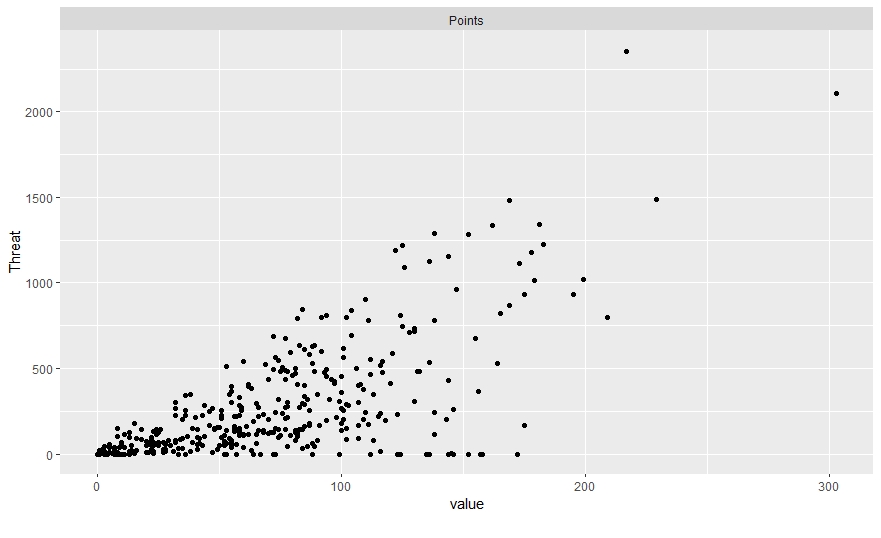


***Figure 4.3: Observation of Univariate Variables in week 11.***

In Figure 4.3 above, similar to the occurrence in week 0, ICT, Creativity, Threat and Influence maintains their positive skewedness. In other words, there is consistence in the data distribution between week 0 and week 11 of the fantasy football league.

**Research Question three (3)**

iii. What is the relationship between players Threat and Points awarded in week 0?

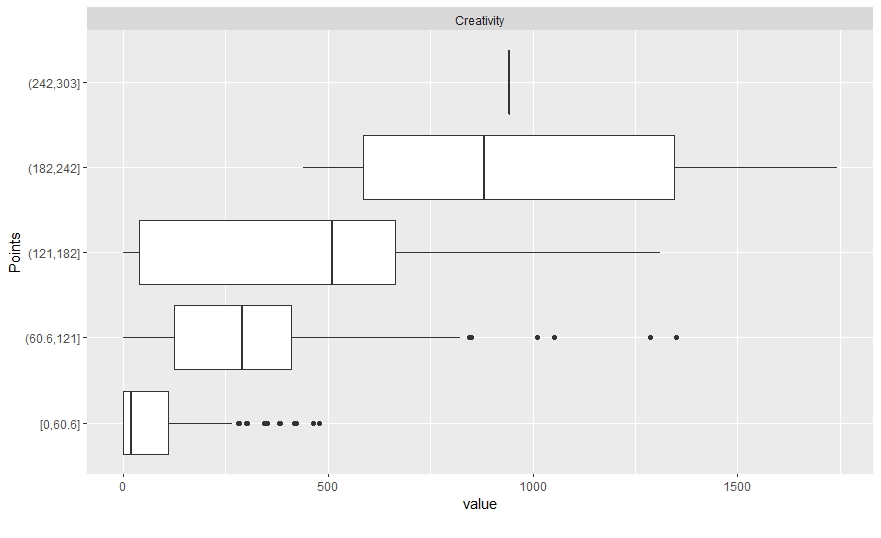
****

**Figure 4.4: Scatterplot Bivariate distribution of Point by Threats**

In figure 4.4 above, the scatter plot indicates that there is a strong positive linear relationship between Threat and Points awarded. Although, there are few observed outliers.

**Research question Four (4)**

What is the relationship between players Creativity and Points awarded week 0?



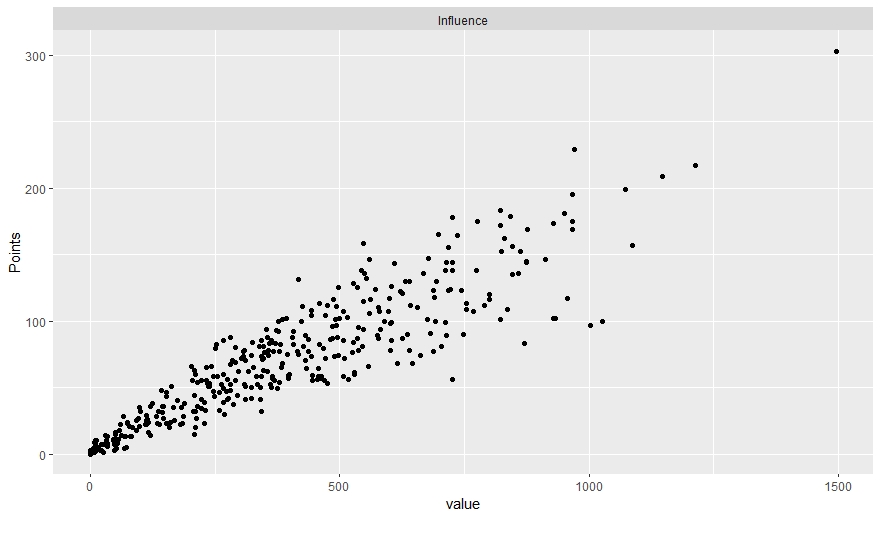
***Figure 4.5: Boxplot Bivariate distribution of Point by Creativity***

In figure 4.5 above, the boxplot gives a good indication of how players Threat values are spread. It shows the five numbers summary of the minimum, first quartile (Q1), Median, Third Quarter and the Maximum. It also indicates that the creativity dispersion is not constant all through the distribution. From the point of (0,60.6) the creativity is skewed to the left and contains outliers while at (60.6, 121) there are outliers and mean slightly tilt towards Q3. At (121,182) the mean moves farther to the Q3 as observed earlier. Finally, at Point (182, 242) the mean is closer to the Q1 and the creativity is left skewed.

This indicates that, the relationship cannot be established, this could be as a result of the outlier.

**Research question Five (5)**

v. What is the relationship between players Influence and Points awarded in week 0?



***Figure 4.6: Scatter plot Bivariate distribution of Point by Influence***

In figure 4.6 above, the scatter plot indicates a strong relationship between players’ influence value and points awarded. The association between both variable is linear.

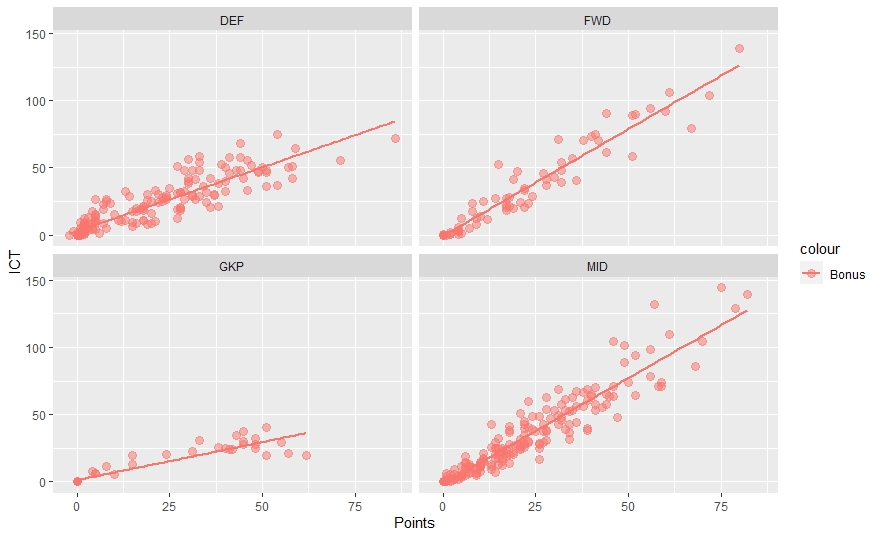
In summary, it means the higher the influence rating of a player, the likely his contribution to be awarded more points.

**Research question Six (6)**

vi. What is the relationship between players’ ICT, Bonus and Points Scored in week 0 and 11?



*Figure 4.7: ICT, Point and Bonus Distribution in Week 0*



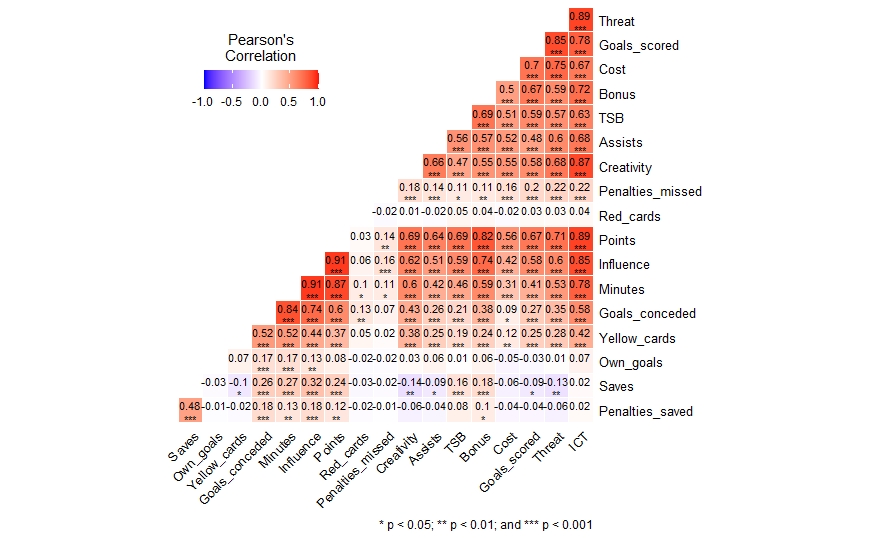
***Figure 4.8: ICT, Point and Bonus Distribution in Week 11***

In both Figure 4.7 and 4.8, it indicates that there is a consistent positive and strong linear relationship between players’ ICT, Bonus and Points awards at different position played by a player.

In view, yes, there exist a relationship between the three variables.

**Research question Seven (7)**

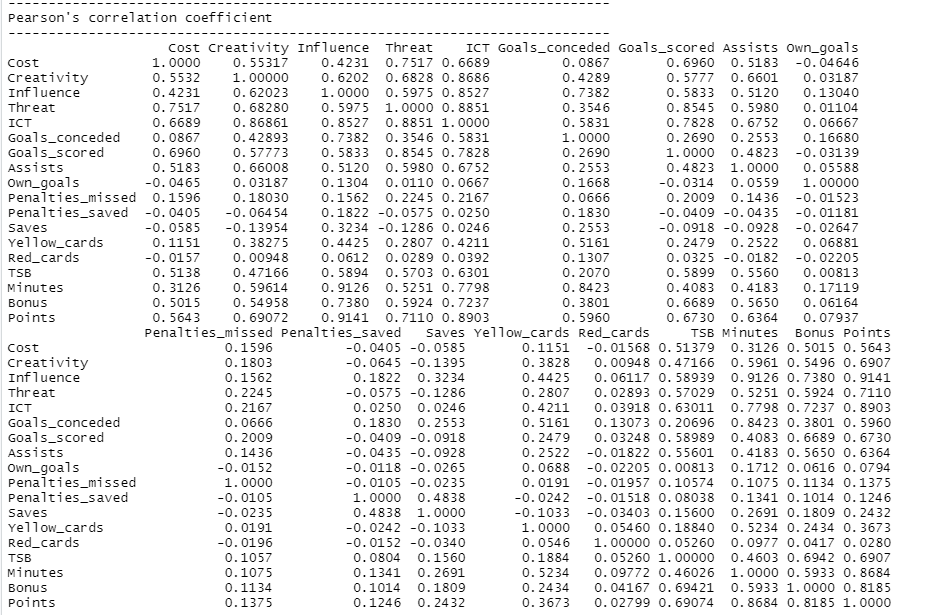
what is the degree of relationship between the attributes in week 11?

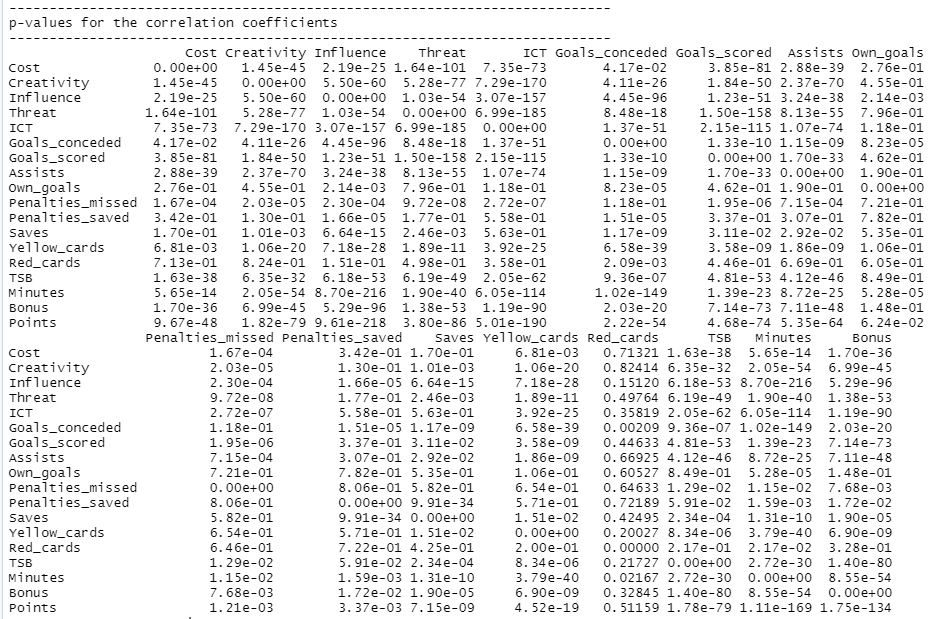


***Figure 4.9: Pearson Correlation Heat map***

In figure 4.9 above, it presents the Pearson’s correlation coefficient which indicate the degree of correlation between the variables. Threat, Bonus, Influence, Creativity shows a strong correlation coefficient of value closer to 1. Note that, negative values on the Heat map indicate negative correlation, 0 value indicates no association between variables and a positive indicates a positive correlation at a confident level of 0.5.

Below is the tabular presentation of the Person’s correlation coefficient.

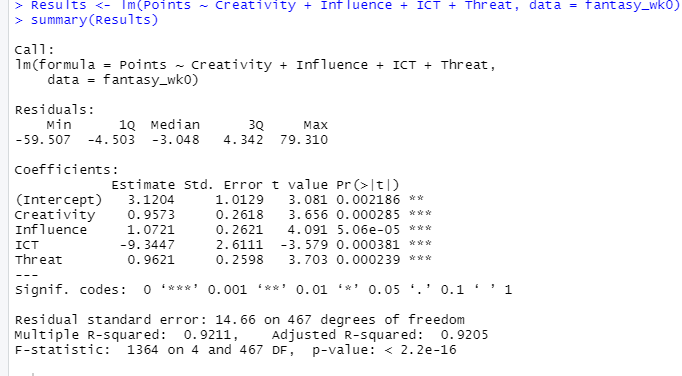




***Figure 4.10: Pearson’s Correlation coefficients and p-values***

**Research question Eight (8)**

Using available metrics, can we predict points and bonus score that can be earned by a player in week 11?



***Figure 4.11: A model to predict Points awarded***

Figure 4.11 presents the output of the Model.

The residuals are basically the difference between observed and predicted values. It also shows the min and max data points alongside first and third quarter, then the median.

The coefficients; Estimates are used to predict the value of the response variable.

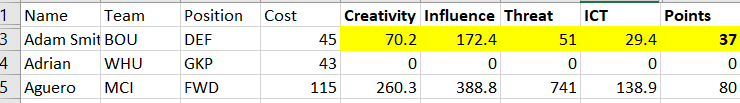
Std. Error: Is the average amount that the estimate varies from the actual value of points awarded, the t-value also shows how many standard values are between the estimated and zero (0).

Pr(>|t|) gives a pvalue for the test to determine if the coefficient is significant.

Multiple R-square gives a measurement of what % of the variance in the response variable can be explained by the regression. Adjusted R-sqaure typically controls each additional predictor to avoid overfitting of the model.

F-statistic basically indicates if the model as a whole is statistically significant. Although with a p-value smaller than zero (0), it makes the model significant.

A test was conducted using the model above and data for week 0 against a likely outcome in week 11, and below is the output.



Week 11 test data



***Figure 4.12: Test and Output from Model.***

In figure 4.12 above, the created model was able to predict a value for points that would be awarded to a player with input variables marked in yellow within Week 11. The model predicted a total point of 29.5, while the observed point in week 11 is 37.

With about (-7) difference in the predicted model, it shows that atleast, there exist some strong relationship between the input variables for the purpose of prediction.

**Section Five (5)**

**5.0 Conclusion and Recommendation**

**5.1 Conclusion**

The key findings of this report can be best summarized as follows:

i. The data distribution observations of the two datasets of week 0 and week 11 maintain a steady pattern in terms of spread, mean, mode and median.

ii. They were established relationship between the players’ threat, influence, Creativity and ICT metrics against total points awarded. Infact, there exist a strong positive linear relationship between the input variables and the target variable.

iii. The statistical techniques used in the analysis was sufficient enough to provide answers based on results observations to the research question.

iv. Lastly, for every player who wants to earn more points in the Fantasy Football premier league, it is advises able to critically observe the players Threat score, Influence, Creativity and ICT as these metrics gives the individual a higher chance of been awarded more points at the end of the game.

**5.2 Recommendation**

The researcher would like to recommend that, a more advanced research can be carried out by merging all datasets from (Week 0-11) to create and train a model for a better predicted values of next fantasy football league session.

**Reference**

Editage Insights (2018) *Q&A Forum, Data Analysis.* Available at: <https://www.editage.com/insights/when-can-i-use-correlation-analysis-as-opposed-to-regression-analysis> (Accessed on 10th December 2021).

McCambridge Ed (2021) *A beginner's guide to Fantasy Premier League - the game your mates don't shut up about every year.* Available at: <https://www.fourfourtwo.com/features/fpl-what-is-fantasy-football-and-how-does-it-work> (Accessed: 9th December 2021).

Premier League (2021) *Fantasy*. Available at: <https://fantasy.premierleague.com/> (Accessed: 9th December 2021).

Wikipedia (2021) *R (programming language)*. Available at: <https://en.wikipedia.org/wiki/R_(programming_language)> (Accessed on 10th December 2021).

**Appendix**

library(tidyverse)

library(stats)

library(dplyr)

install.packages("DataExplorer")

library(DataExplorer)

install.packages("metan")

library(metan)

#Dataset Load and verification

fantasy\_wk0 <- read.csv("FPL\_2018\_19\_Wk0.csv")

fantasy\_wk11 <- read.csv("FPL\_2018\_19\_Wk11.csv")

view(fantasy\_wk0)

str(fantasy\_wk0)

str(fantasy\_wk11)

# Data Preprocessing, testing for missing values

sum(is.na(fantasy\_wk0))

sum(is.na(fantasy\_wk11))

#Carrying out EDA

fantasy\_wk0 %>% glimpse()

fantasy\_wk0 %>%

create\_report(

output\_file = "Fantasy Week 0 Report",

output\_dir = "AppliedStatsAssement/",

y = "Points",

report\_title = "EDA Report for Fantasy League Week 0"

)

fantasy\_wk11 %>%

create\_report(

output\_file = "Fantasy Week 11 Report",

output\_dir = "AppliedStatsAssement/",

y = "Points",

report\_title = "EDA Report for Fantasy League Week 11"

)

fantasy\_wk0 %>% plot\_correlation(maxcat = 5, type = "c")

plot\_correlation(na.omit(fantasy\_wk0), type = "d")

introduce(fantasy\_wk0)

## Reduce data size for demo purpose - Bivariable

threat\_point <- fantasy\_wk0[, c("Threat","Points")]

## Call boxplot function

plot\_boxplot(threat\_point, by = "Points")

plot\_scatterplot(threat\_point, by = "Threat")

## Reduce data for Creativity agiainst Point

Creativity\_point <- fantasy\_wk0[, c("Creativity","Points")]

## Call boxplot function

plot\_boxplot(Creativity\_point, by = "Points")

## Reduce data for Influence agianst Point

Influence\_point <- fantasy\_wk0[, c("Influence","Points")]

## Call boxplot function

plot\_scatterplot(Influence\_point, by = "Points")

## Research question 6 - testing relationship between ICt, Bonus and Points awarded in week 0

fantasy\_wk0 %>%

ggplot(aes(Points,ICT, color = "Bonus")) +

geom\_point(size =3, alpha = 0.5) +

geom\_smooth(method = lm, se=FALSE) +

facet\_wrap(~Position)

labs(title = "ICT, Point & Bonus Distribution") +

theme\_bw()

# testing relationship between ICt, Bonus and Points awarded in week 11

fantasy\_wk11 %>%

ggplot(aes(Points,ICT, color = "Bonus")) +

geom\_point(size =3, alpha = 0.5) +

geom\_smooth(method = lm, se=FALSE) +

facet\_wrap(~Position)

labs(title = "ICT, Point & Bonus Distribution") +

theme\_bw()

getwd()

# Correlation test

corrl2 <- corr\_coef(fantasy\_wk11)

plot(corrl2)

print(corrl2)

sink("corrl2.txt")

print(corrl2)

sink()

#Performing a linear regression

Results <- lm(Points ~ Creativity + Influence + ICT + Threat, data = fantasy\_wk0)

summary(Results)

predict(Results, data.frame(Creativity = 70.2, Influence = 172.4, ICT = 29.4, Threat = 51))