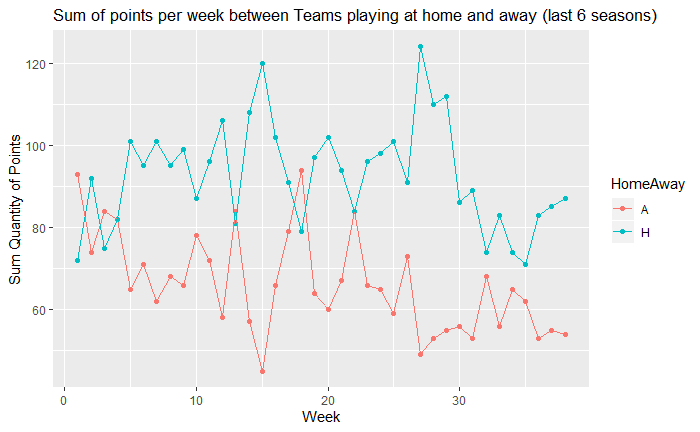
*Marcel Socorro*  
*Big Data Analytics Capstone*

*Exploratory Data Analysis*

*St. Thomas University*

*EPL Prediction*

1st Graph



In the graph above is plotted the sum of points obtained by teams playing at Home and Away every round for the last 6 seasons. The pattern is that teams playing at home are more likely to get more points that teams playing away except the first 4 round where this result is more chaotic.

# 2nd Graph

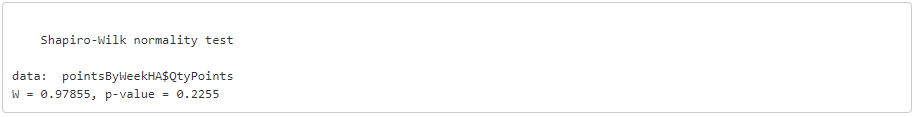
# 

This graph is used to have break down of the variables plotted in the first graph separated by each of the last 6 seasons. Same pattern could be generally observed, first round are more chaotic, but Home teams tend to get more points than teams playing away.

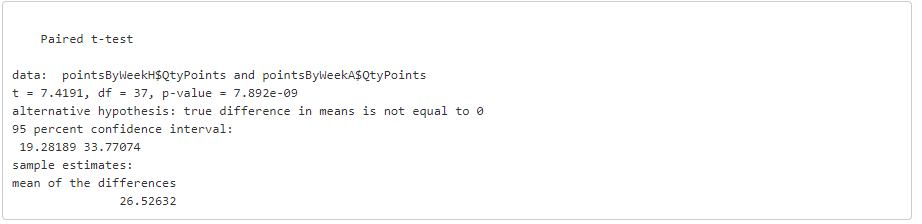
# 3rd Graph

# 

This is a Histogram for the variable Sum of Quantity of points for the last 6 season separated by each round and playing home/away team.In this graph is observed the two curves displaced from the general mean, the mean of the curve Home team is displaced to the right and the mean of the curve Away team is displaced to the left.



Shapiro test is run for the variable Sum of Quantity of points by week and pvalue is higher than 0.05 which means that null hypothesis (The variable is normally distributed) can’t be rejected.



Once it is proved that the variable is normally distributed, it is confirmed that there is a significant difference between the Sum of quantity of points obtained by teams playing at home and away each week. Pvalue is less than 0.05, which means that there is enough evidence to reject null hypothesis (No significant differences between both variables). It could be confirmed that playing at home or away, it’s a factor impacting the outcome of the game. The test was done setting the parameter paired=TRUE because both variables are not independent of each other.

# 4th Graph

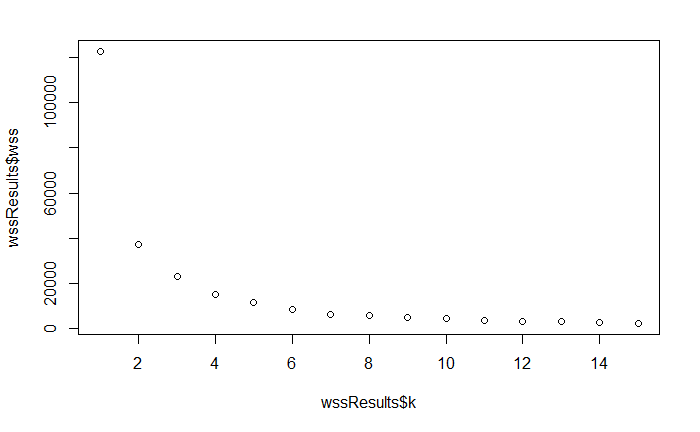
# 

The graph above is a subset of the last round for each of the last 6 seasons. The cumulative variables throughout each season, goal difference and points per team are plotted. According to the graph there is a strong positive correlation between these 2 variables.

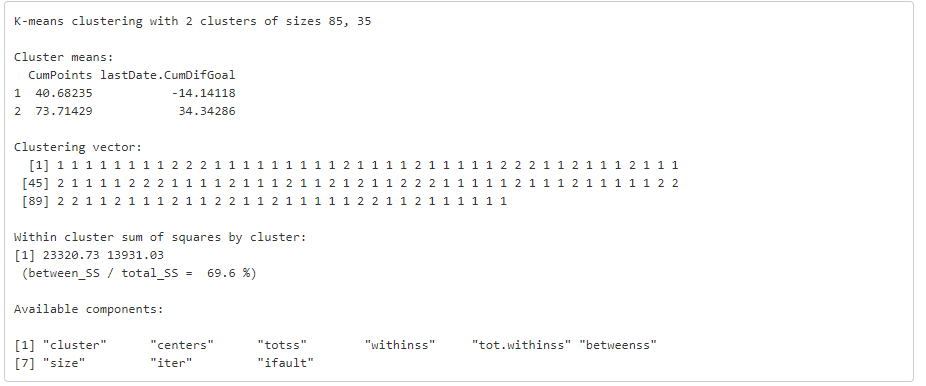
# 5th Graph

## **Getting the optimum cluster**





According to wss vs k analysis, the optimal cluster is chosen for 2 units, the big drop happens in the second cluster, and from the third cluster, the drop in the wss is less than the half of the previous cluster, for instance the wss linked to the 3rd cluster represents the 62% of the wss linked to the 2nd cluster which represents a drop of 38%. This selection is done assuming a big drop for a difference greater than 50%. This criteria could be reevaluated for chosing 4 clusters to test the model.



# 6th Graph

# 

In the graph above, each point (Team) is assigned to a cluster, this could be an element to group teams with similar performance, and then, getting metrics from this

# 7th Graph

# 

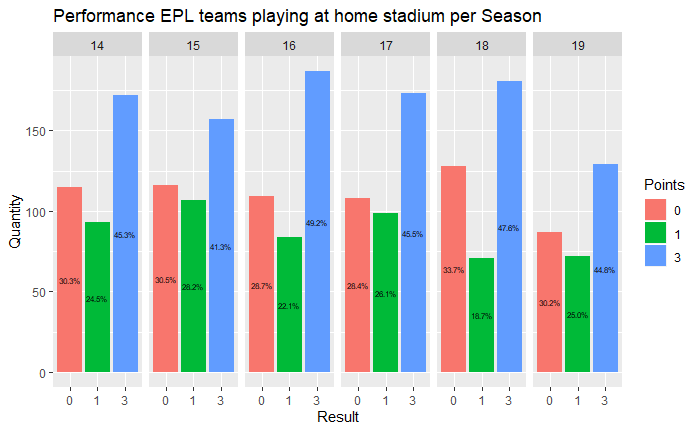
Using the cluster assigned it was compared the goals for and against for each group. The difference is remarkable between them.

# 8th Graph

# 

Analyzing the graph based on general probabilities, could be seen the proportion for the performance of all teams playing at home during the last 6 seasons. We could see that the outcome less likely is the draw, and the outcome most likely is the victory. For the teams playing away will be the same results only switching classes between defeats and victories, the outcome most likely for a team playing away is the defeat.

# 9th graph



This graph is a breakdown of the previous graph, it could be seen that probabilities of a win for a team playing at home oscillated between 41.3% and 49.2% for the last 6 seasons.