2018

Statistical Analysis of Fantasy Premier League

AN IN-DEPTH ANALYSIS OF WEEKS 0-7 OF THE BARCLAYS PRIEMIER LEAGUE

GARRY JAMES MCBRIDE

201867279

MSc Advanced Computer Science

Cs982 Big Data Technologies

Word Count (3256)

Contents

1. Introduction

The English Premier League is a yearly competition where 20 teams play a total of 38 games per club from August to May (of the following year) in England (Morse, Burden and Canterbury, 2017). It was founded on the 20th February 1992 replacing Football League First Division, aka League one (Premierleague.com, 2018). Teams play for points, where a win would achieve 3 points, a draw will achieve 1 and a result of 0 points for a loss (Lee, 1997). England has some of the most expensive players in the world (Coleman, 2018), and a large variety of talent across multiple leagues and youth academies (Calvin, 2017). Teams who finish in the top seeds, progress to the Champions League or Europa League (clubs make more additional money playing in these competitions) where they play teams from other countries who also finished in high positions, and the bottom seeded teams in the league are relegated to a lower division, which is not as prestigious as the top league (Premier League) (Premierleague.com, 2018).

Football is the most popular sport in the world, watched by millions on televisions or attending matches (Murray, 1998). Europe has the biggest clubs in the world and attracts some of the best players across the globe (Murray, 2018). Clubs pay many players a vast sum of money which continues to break records set by themselves (Ellis, 2017). Players are bought and sold for millions of pounds and can sometimes break well past £100 Million (Eurasia Review, 2018).

The English Premier League has the highest percentage of foreign players in the world coming from mostly Central Europe, with a considerable amount from Africa, South America, and Asia (Sky Sports, 2017). It is watched and loved by millions (Facts, 2018) a statistic that continues to rise.

With the amount of money and attention the Premier League attracts, there are key attributes and statistics that are monitored to ensure players/clubs are improving so that they might challenge or dominate their opposition. The dataset has been chosen to find out exactly what aspects players or teams have that ensure their success or cause failure.

1.1 The Dataset and its Variables

The dataset chosen is very detailed as two sheets that are 7 weeks apart have been selected, week '0' and week '7'. Comparing these two sheets will show progress, failure, and statistics that have changed which could possibly affect a Players/Clubs worth or reputation.

There are 542 operations and 21 columns in the dataset. The columns are:

1. **Name**

   Title of each player in the data set, the data set does not give any other personal information on the players. Players in the Data are identified by their surname only.

2. **Team**

   The Dataset has 20 teams, they are identified by abbreviations of the club's title, e.g. Manchester United, Chelsea and Westham United are 'MUN', 'CHE' & 'WHU'. Teams are (in alphabetical order of abbreviations): Arsenal (ARS), Brighton & Hove Albion (BHA), Bournemouth (BOU), Burnley (BUR), Cardiff City (CAR), Chelsea (CHE), Crystal Palace (CRY), Everton (EVE), Fulham (FUL), Huddersfield Town (HUD), Leicester City (LEI), Liverpool (LIV), Manchester City (MCI), Manchester United (MUN), Newcastle United (NEW), Southampton (SOU), Tottenham Hotspur (TOT), Watford (WAT), Westham Untied (WHU), Wolverhampton Wanderers (WOL).

3. **Position**

   The positions are very vague and short, they do not have specific position attributes in Defence, Midfield, or Attack such as: Defence would have Center Back (CB), Left(wing)/Right(wing), and Midfield would consist of Center Defensive Midfield (CDM) & Center Attack Midfield (CAM) to name a few. The positions are labelled as Goal Keeper (GKP), Defence (DEF), Midfield (MID) and Forward (FWD).

4. **Cost**

   The first numeric column, based on the scale of 0.1 Million pounds, e.g. 74 would = £74 Million. Players are bought for much smaller amounts in the Premier League, but this value is based on the players worth from the statistics they hold on performance, they vary from week to week.

5. **Creativity**

   Assesses player performance in terms of producing goalscoring opportunities for others. The ability to create chances and set other players up for goals they might find easier to score.

6. **Influence**

This evaluates the degree to which that player has made an impact on a single match or throughout the season. Boasts team moral and momentum, an experienced player or one that others follow.

7. **Threat**

   A value that examines a player's threat on goal, creates panic or fear in the oppositions side, having to focus or concentrate more on the individual, by marking or defending.

8. **ICT**

   Statistical Index combining the Influence, Creativity and Threat data

9. **Goals_conceded**

   Number of goals conceded while the player was on the field, more focused towards players in a defensive role as their task is to stop the opposition scoring.

10. **Goals_scored**

    Goals scored by the player. Most goals are scored by strikers, but Midfield players or defenders contribute to a large amount of goals, from counter attacks or corners.

11. **Assists**

    Assists provided by the player, an aspect that relates to Creativity, in order to have an assist, a player has created an opportunity by however way in order to give another player a chance on goal with a low amount of effort on goal scorers part.

12. **Own_goals**

    Own goals scored by the player, an attribute no player wants to have positive values in.

13. **Penalties_missed**

    Penalties missed by the player (Usually applies to Goalkeepers). Penalties are common in a game of football, and though it may look like an easy objective, a player has to stay focused and accurate to achieve a goal from the penalty spot against a Goal Keeper.

14. **Penalties_Saved**

    Penalties saved by the player (Usually applies to Goalkeepers). A Goal Keeper has to be quick and, on their feet to save shots from the penalty spot as they can be high/low, fast/slow and players sometimes use methods to distract the Goal Keeper or trick them into diving in the opposite way of the initial shot.

15. **Saves**

    Saves made by the player (Usually applies to Goalkeepers), quick on their feet and can judge were the opposition player is going to place their shot.

16. **Yellow_cards**

Yellow cards are awarded as a warning to foul play, or behaviour on the field. In order to get Yellow carded, or “Booked” a player has tested the patience of the referee, but not drastic enough to be sent off (Krafft, 2018).

17. **Red_cards**

    Red cards are awarded if a player has received two yellow cards and the referee has had enough of the players behaviour. A straight red card can be awarded before a yellow depending on how drastic or dangerous a player is behaving (Krafft, 2018).

18. **TSB**

    Percentage of teams in which the player has been selected.

19. **Minutes**

    Minutes played by the player, a game in the Premier League has 45 minutes in each half, with an additional injury time added to the clock depending on the referee’s decision on how much they have had to stop game play in the halves.

20. **Bonus**

    Bonus points awarded to players on top of standard points given based on their individual performance.

21. **Points**

    Each Player receives points based on their performance and rating of statistics.

2. Challenge

A team will only win a game working together, but sometimes it is one player who makes a difference (whatever position they may play in), or a combination of players in different roles. Relying on one player to be the difference between a win or lose is not a reliable strategy and will potentially not be effective for long if the player becomes injured or is sold to another club. The same can be said where one player is the reason for a loss based on their attitude or behaviour on the park.

The following attributes are problems with the data that need to be taken into consideration:

- Large data set
- A lot of variables and contributors
- Two data sets with 7 weeks in between them
- Too many possibilities for outcomes (must narrow for most detail on certain outcome)
- Positions are not as in-depth as they only categorise the players in 4 sections

3. General descriptive analysis of the variables

This section is a general analysis of the dataset. To start with, the non-numerical will be analysed based on the fact they are the main variables that the results are categorized under. A general analysis on these variables will implement a starting point for then further in-depth findings. The following objectives are based on the differences between weeks 0, and weeks 7. The reason for using two datasets that are 7 weeks apart is for an understanding on variable values increasing or decreasing. These statistics are important to find where players or teams need to improve on certain factors based on their results from the start of the season.

3.1 Players in Teams

Firstly, the number of players in each team in week 0 must be assessed. Teams can drop players or bring players from the reserves to the main squad if they are improving or must replace another player.

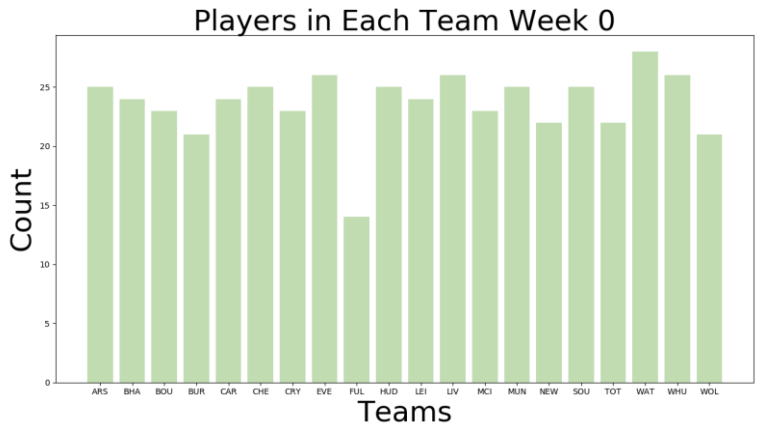

Figure 1.1 - Players in Number of Teams for Week 0 (Bar Chart)

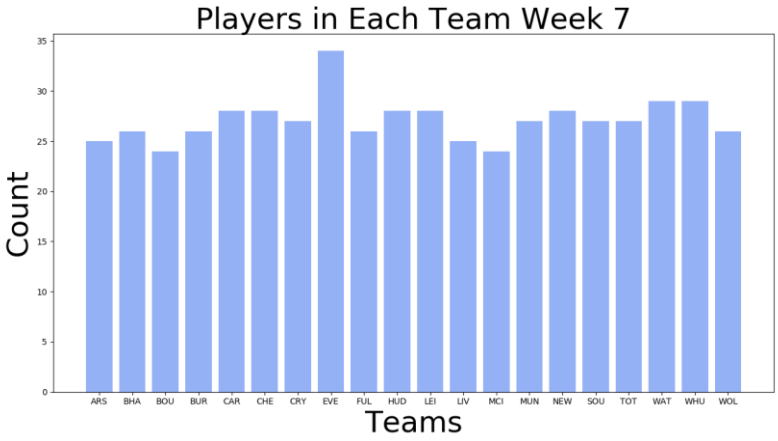

Figure 1.2 - Players in Number of Teams for Week 7 (Bar Chart)

The bar charts (**Figures 1.1** & **1.2**) highlight an increase and decrease with certain teams that have dropped or added players in week 7. Fulham and Everton have increased by a large amount, considering Fulham were nearly promoted to the Premier league from the league below (The Championship), it is surprizing they started the league with fewer players they would be new and not know which players would be effective so early in week 0.

Each team has a similar number of players in their squad (excluding the previous mentioned teams). The reserve bench has no limit of how many players are on it at one time, the only limit is the 11 players on the field.

3.2 Positions in Each Team

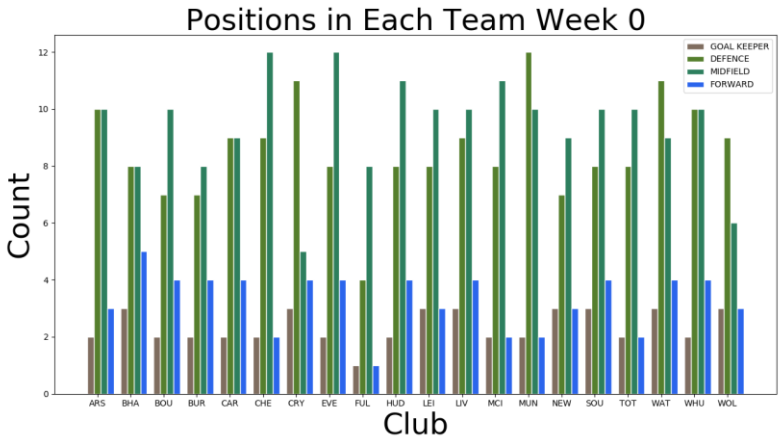

Figure 2.1 - Position Count in Each Team Week 0 (Grouped Bar Chart)

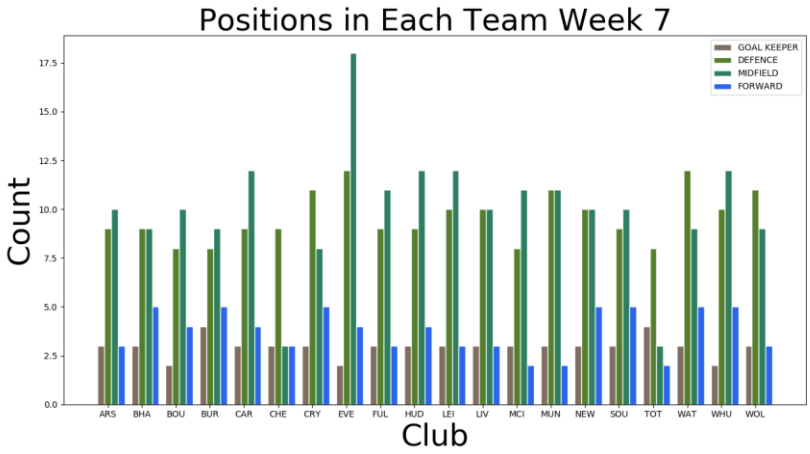

Figure 2.2 - Position Count in Each Team Week 7 (Grouped Bar Chart)

Analysed now is the number of positions that each player is assigned to. The Figures **2.1** & **2.2** show that there is an increase in Goal Keepers added to the squad, and the positions most associated with each team are Defence and Midfield. Everton have a much larger number of Midfielders than any other. As for defenders, Everton and Watford & Crystal Palace have the largest quantity of Defenders. With a small number of Forward players, each seem are relying on their Defence and Midfield quality.

3.3 Players in Certain Positions

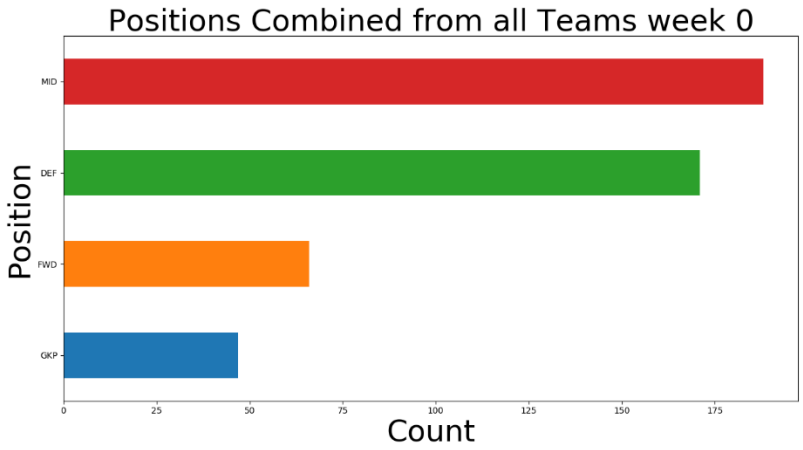

Figure 3.1 - Position Count in Combined Teams' Week 0 (Grouped Bar Chart)

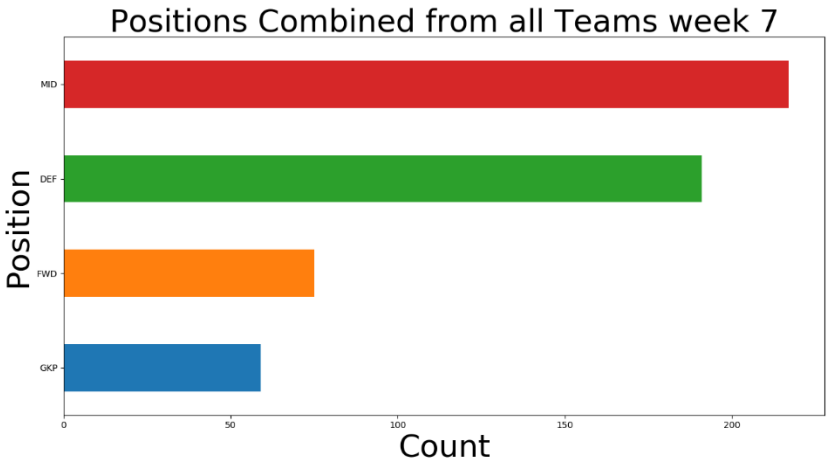

Figure 3.2 - Position Count in Combined Teams' Week 7 (Grouped Bar Chart)

The figures **3.1** & **3.2** look at positions from teams combined. This gives a more accurate analysis of what positions have increased or decreased from weeks 0 to 7, the reason for this section is to look deeper into the interesting findings from the section before as it highlighted results that must be queried.

The increase in midfield is the most drastic increase, proven from the **figures 2.1** & **2.2** about and the section before. Numbers of Goal Keepers and Forward positioned players hasn't changed by a large amount, all teams seem to be focusing on Defence and Midfield, a trend is present in the Premier League.

3.4 Relation

The sections before proved there is a trending factor amongst the teams. Next the relationship between variables must be analysed to find out more from these trends. Doing this will highlight the key aspects that influence the choice that squads are making on which players and where their place is.

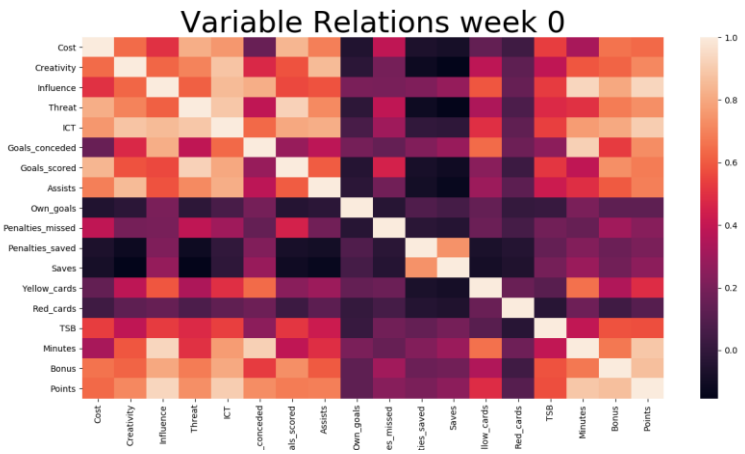

Figure 4.1 - Variable Relation Week 0 (Heat Map)

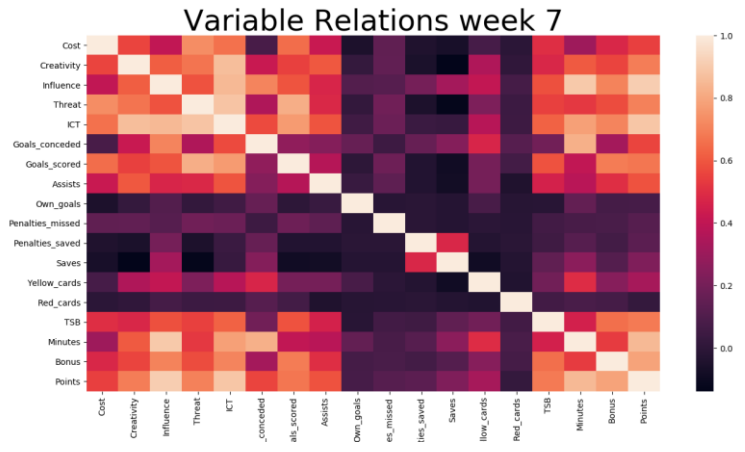

Figure 4.2 - Variables Relation Week 7 (Heat Map)

The heat maps shown above highlight a decrease in tempeture in most relationships between variables. **Own_goals**, **Penalties_missed**, **Penatlies_saved**, **Saves**, **Yellow_cards** and **Red_Cards** have a large relationship with most of all aspects. Variables such as **Creativity**, **Influence** and **Threat** are high in relation to these factors.

Moving away from the first factors mentioned in the paragraph above, as **Creativity**, **Influence** and **Threat** are very broad variables in the dataset, they can measure any other variable based on their objective. A more in-depth analysis into these factors could potentially prove to be interesting.

4. In-Depth Analysis of Variables

This section is an in-depth analysis of important findings from the section before. Firstly, this section will look at the Variables: Creativity, Influence, Threat. These factors had interesting findings before and looking into them more in-depth should prove to find factors that when tied or compared with will have interesting outcomes. The report will then use Unsupervised and Supervised methods applied to factors that must be developed further.

4.1 Teams Based on Creativity

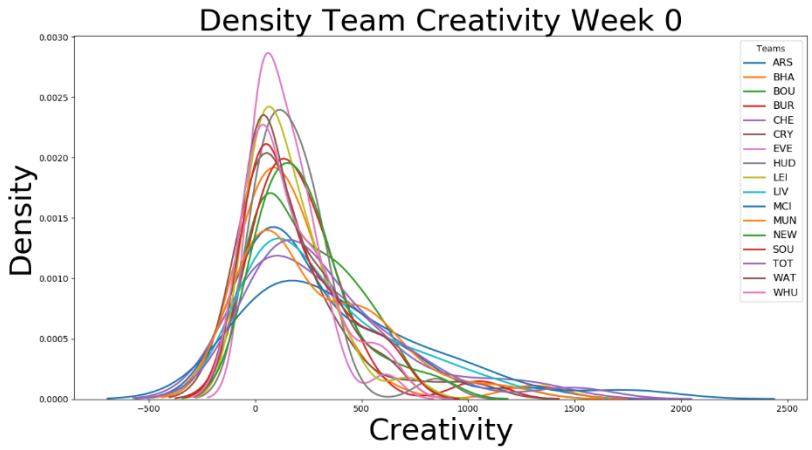

Figure 5.1 - Squad Creativity Week 0 (Density Diagram Teams Individually)

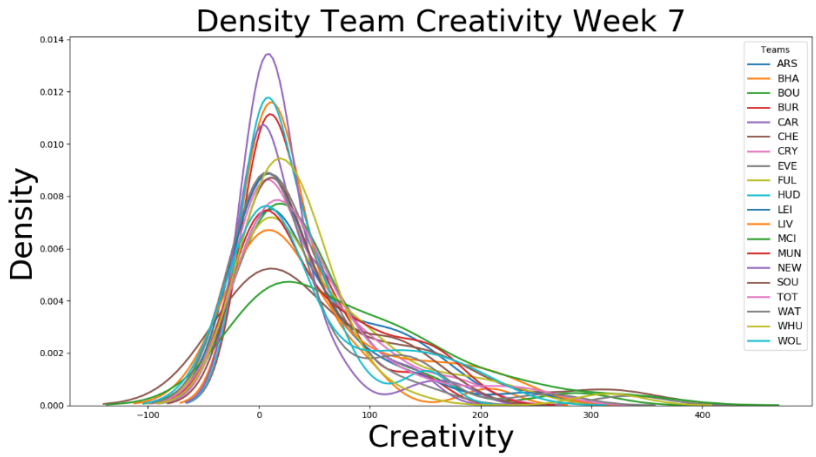

Figure 5.2 - Squad Creativity Week 7 (Density Diagram Teams Individually)

Figure 5.1 above show in week 0's creativity levels were almost halved in relation to creativity levels being higher in more players in a team and then levels being lower in other teams throughout the squad. Manchester City, Chelsea and Arsenal's creativity levels are not as dense as the rest of the league. Their creativity is spread out through the squad and reaches high levels with a fair number of players holding this aspect.

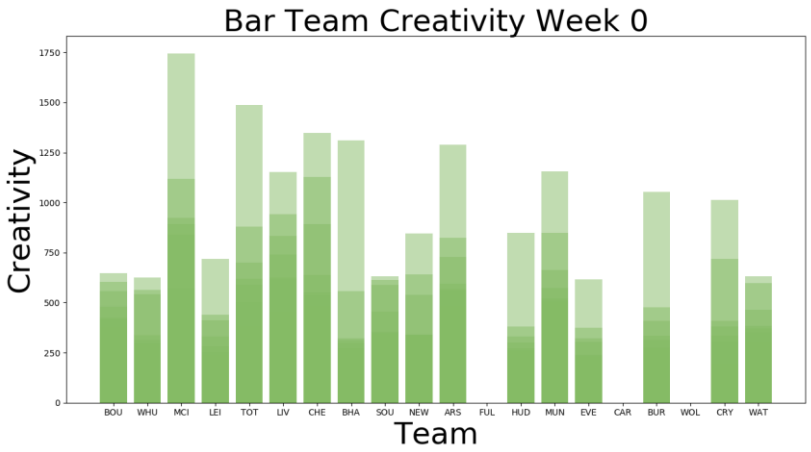

Figure 5.3 - Squads Creativity Week 0 (Bar Chart Teams Individually)

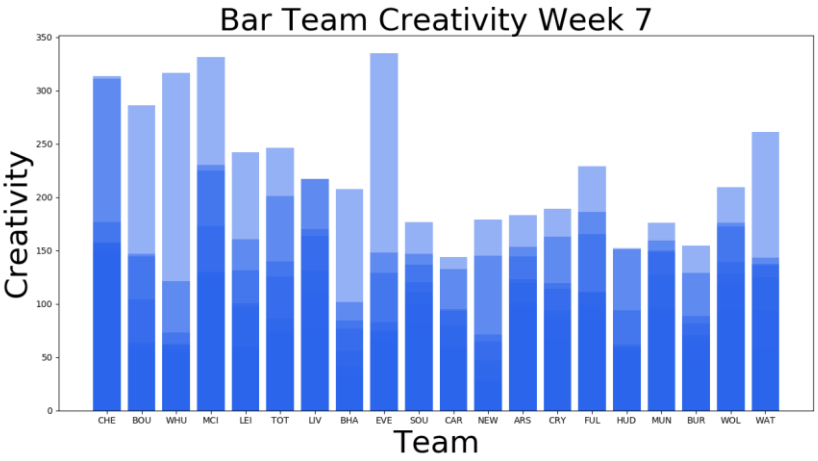

Figure 5.4 - Squads Creativity Week 7 (Bar Chart Teams Individually)

From the **Figures 5.3** & **5.4** (above), each team has a rise in creativity, mainly from one individual player, apart from chelsea who have two players drastically rising. Most of each player creativity levels are densely populated together at a very lower level than certain individuals' level is a large value more.

From week 0 to week 7, even though creativity levels have increased in teams, the variable has decreased drastically for every team and every player. It's understandable that creativity would drop from week 0 to week 7, as the first week of the season's statistics are based on pre-season analysis, not the factual statistics found from the teams playing over the course of 7 weeks.

4.2 Teams based on Influence

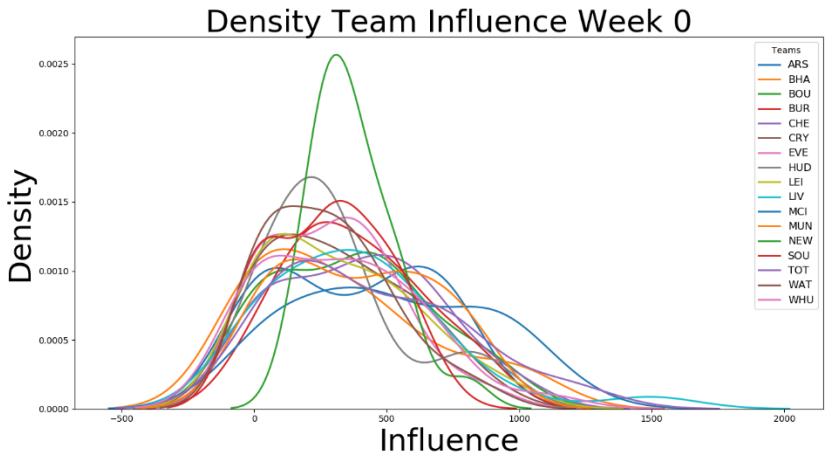

Figure 6.1 - Squads Influence Week 0 (Density Diagram Teams Individually)

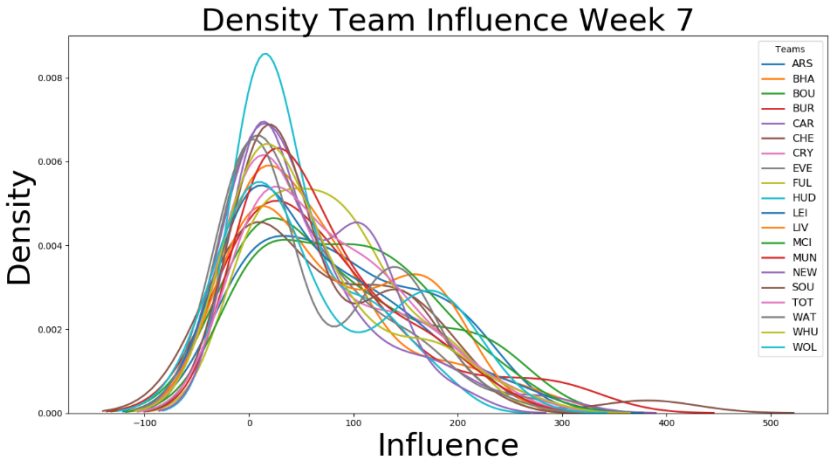

Figure 6.2 - Squads Influence Week 7 (Density Diagram Teams Individually)

Newcastle United Influence in week 0 is most dense from a higher level than all other teams (see **Figure 6.1)**. It stands out from them in figure, shown in green. The lines shape is similar to other teams, but the position is higher on the influence axis, proving that the teams influence level is not varied, is has a an above average rating with a large number of players holding this factor, but no players seem to stand out with influence away from the rest of the squad. In week 7, players in Newcastle United whose Influence level is higher than the rest of the squad, has decreased and joined the average levels the players have (see **Figure 6.2)**.

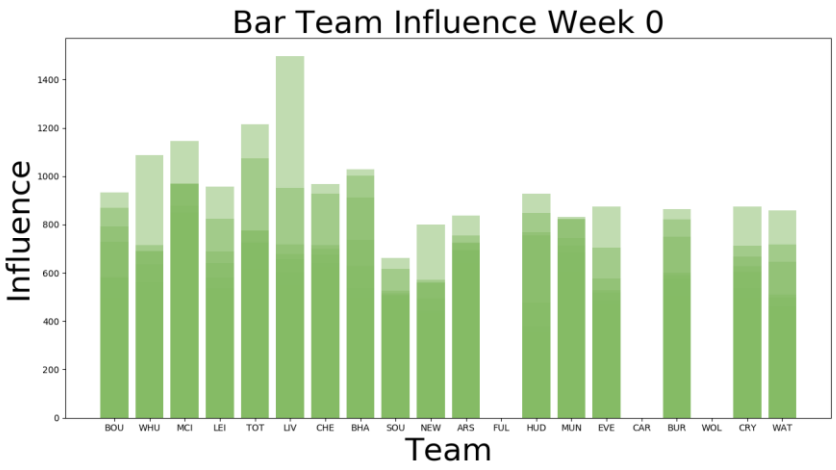

Figure 6.3 - Squads Influence Week 7 (Bar Chart Teams Individually)

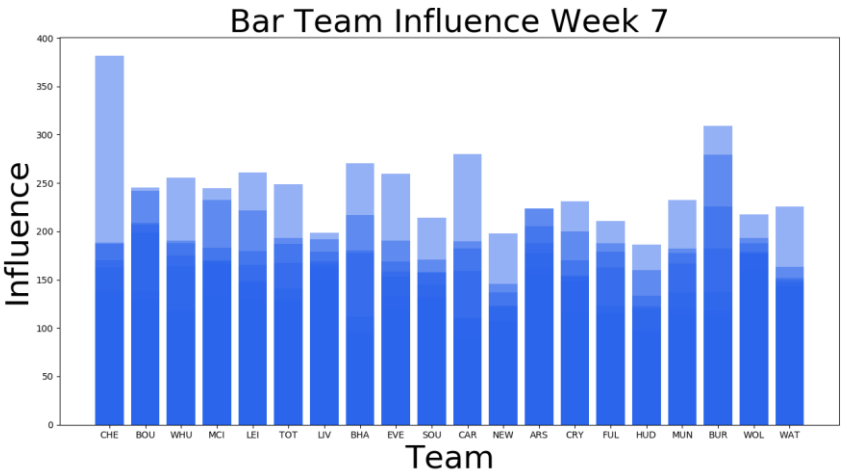

Figure 6.4 - Squads Influence Week 7 (Bar Chart Teams Individually)

Figures 6.3 & **6.4**above follow the same trend as Creativity and Influence from week 0 to week 7, a drastic drop in levels, remember this is because week 0 is based on pre-season statistics. Chelsea's Influence is the highest, but again only because of one player, the teams influence is relativity average in comparison to other squads. This seems to be a recurring trend throughout, as each team have an individual adding a large amount of Influence to the over all rating.

4.3 Teams based on Threat

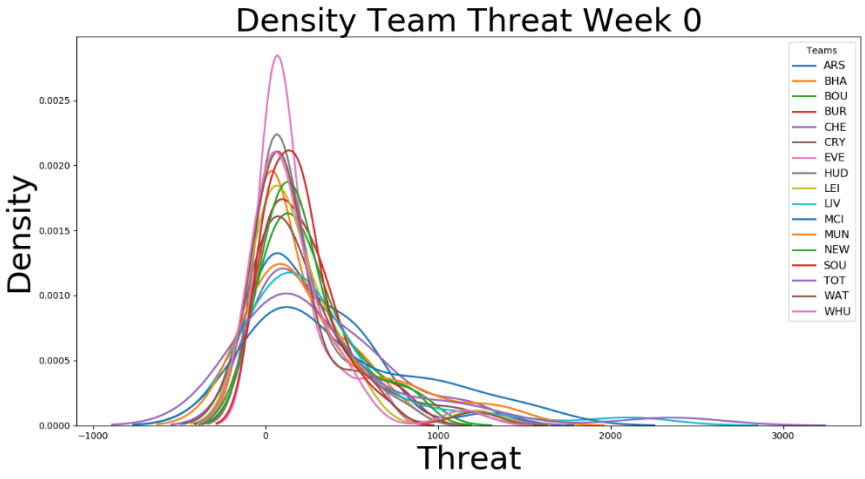

Figure 7.1 - Squads Threat Week 7 (Density Diagram Teams Individually)

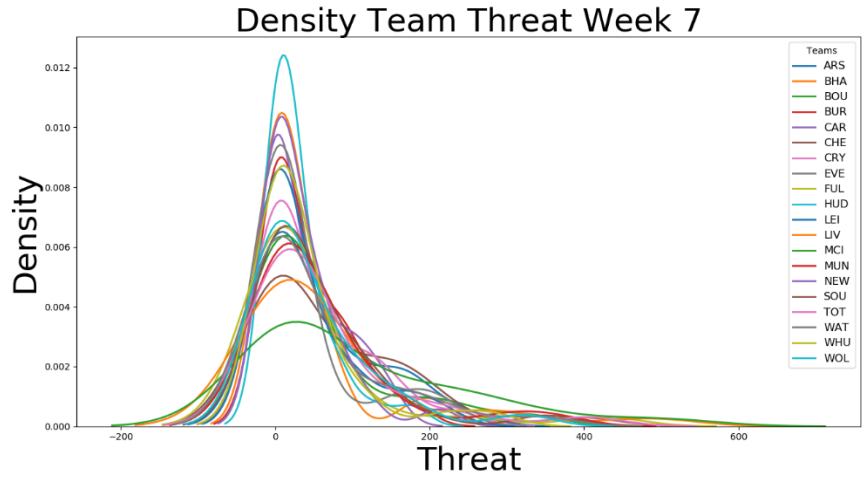

Figure 7.2 - Squads Threat Week 0 (Density Diagram Teams Individually)

From the above diagrams (see **Figures 7.1** & **7.2)**, Manchester City has the biggest threat levels throughout their squad in week 7 (**Figure 7.2**), the same is said for week 0(**Figures 7.1**). Threat levels have stayed relatively the same when comparing statistics from both weeks, the only difference being a slight increase in more players seen as threat to the opposition.

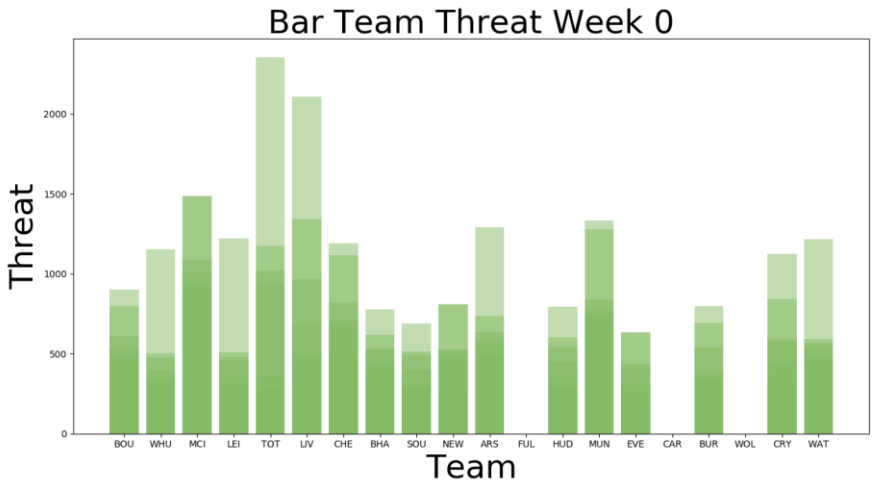

Figure 7.3 - Squads Threat Week 0 (Bar Chart Teams Individually)

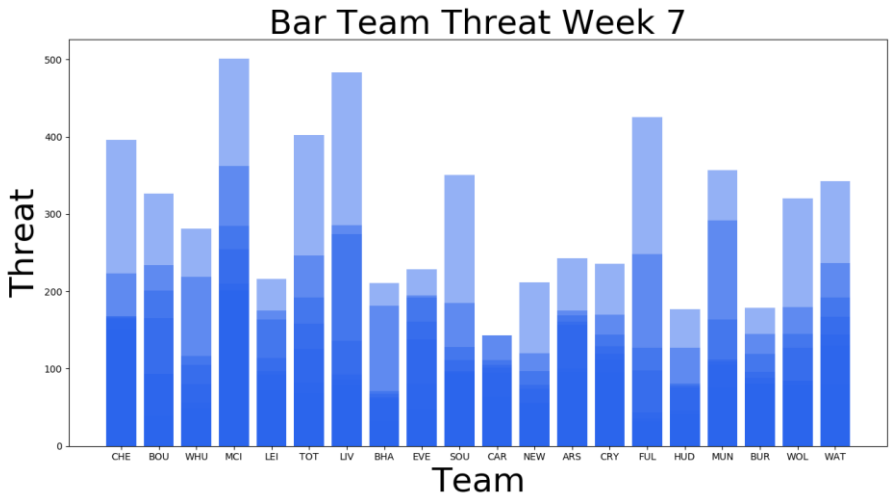

Figure 7.4 - Squads Threat Week 7 (Bar Chart Teams Individually)

The same trend that has followed this dataset through out this section, although with threat, an interesting factor presents itself. Not only does each team have one individual with a higher level than the rest of their squad, it seems that each teams threat level is varied (see **figures 7.3** & **7.4**). Liverpool and Manchester City's squads prove to be more threating than others by a large amount. Cardiff, Huddersfield, Newcastle, Brighton & Hove Albion have the lowest levels of threat, although Newcastle have a player who proves to be effective and increases the teams rating.

4.4 Unsupervised

Now that we have spotted trends and found interesting results from the dataset, it is time to use clustering for more in-depth results. Cost is a factor not yet used, interesting outcomes might be found using clustering on this factor.

Hierarchical Clustering is an unsupervised method used to discover patterns/relationships in data. Not labelled, only input variables are given to this method, this will help with understanding in a different way. Searching for regions of space where the data is most dense, distances are used to evaluate different clusters.

Common distance measures

- Euclidean
- Hamming
- Manhattan
- Cosine

Using cost as the main variable, dendrogram shows clusters related to cost with the variables of Creativity, Threat and Influence.

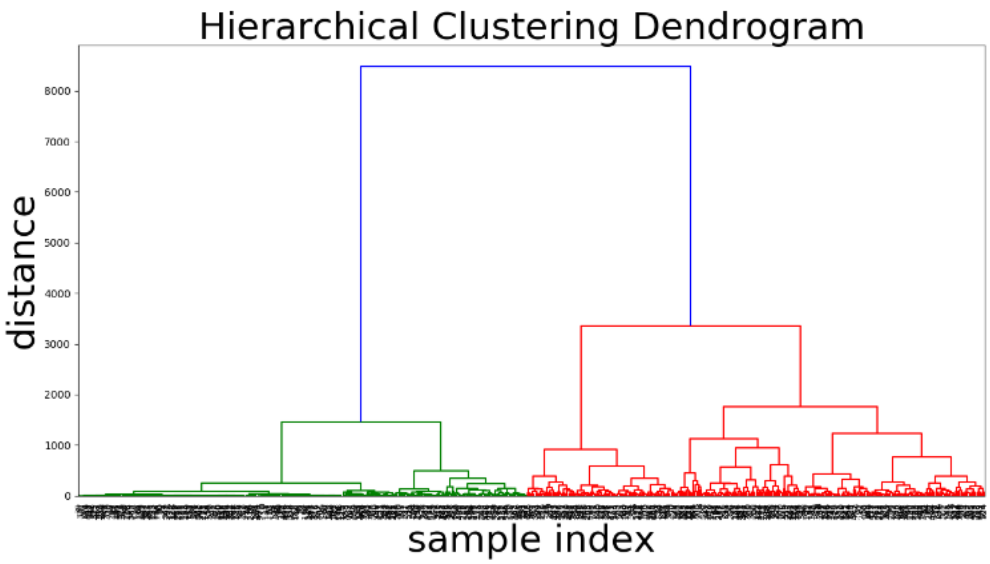

Figure 8 Hierarchical Clustering (Dendrogram)

From the Dendrogram (**Figure 8**), it is clear there is many clusters. There is a lot of similarity for the outcome. The correct number of clusters can't be seen from the Dendrogram, but there are a lot of correct numbers of clusters.

With a level based between 0-1, the outcomes have high scores:

Silhouette Score = 0.4197920979327144

With a score rounded to 41%, the silhouette score is from -1 to +1 the silhouette score shows that objects are very similar their own clusters, compared to other clusters.

Completeness Score = 0.7262158753406041

With a score rounded to 73%, the homogeneity score is good which means the data points that are members of a given class are elements of the same cluster.

Homogeneity Score = 0.7747442655299202

With a score rounded to 77%, the homogeneity score is good which means the clusters contain only data points which are members of a single class.

4.5 Supervised

Next a supervised method is implemented to variables in the data set. The data is labelled when implementing Supervised Methods.

Decision tree method has been used using Cost a main variable with Goals_conceded, Goals_scored, Assists, Own_goals, Penalties_Missed, Penalties_saved, Saves, Yellow_cards and Red_cards under question in relation to cost, the data was split.

	Precision	*Recall*	*F1-score*	*Support*
0.0	0.98	0.84	0.91	75
1.0	0.00	0.00	0.00	1
2.0	0.00	0.00	0.00	4
4.0	0.00	0.00	0.00	0
5.0	0.00	0.00	0.00	1

Table 1.1 - Decision Tree Results

With the higher the percentage the more successful the decision tree was. 0.0 in the table has very high scores (**see Table 1.1**), showing that the data was unsuccessful in finding its way through the decision tree method as numbers hit 0 in the next line, and so on.

Micro avg	0.39	0.39	0.39	163
Marco avg	0.01	0.01	0.01	163
Weighted avg	0.45	0.39	0.42	163

Table 1.2 - Decision Tree Averages

5. Reflection & Conclusion

After analysis of the dataset the results show a trend in individual players being the difference between ratings for certain aspects. Teams have similar statistics, but it is these individuals who contribute to the team's success through means of bringing a certain factor and skill to the club.

There are many issues related to these findings, e.g. if a club has spent millions of pounds on a player, and they become injured (which is likely) they will need replacing by a player who can 'fill their boots', a cheaper player just as creative and effective, which would then increase the value of this player for when the club decides to either sell or keep the replacement.

The methods used for analysing the data proved to be accurate on covering aspects from different approaches. The data set was quite complicated (even without an added sheet), the bar charts were able to group positions together and give them values, density plots were useful for teams individually and easy to follow. There are many other ways of visualizing the data, such as pie charts based on each players attribute, if the data set was to be analysed even further.

With further analysis, the data could prove usual to not only those creating the perfect team on fantasy league but would also help clubs work out where they are failing and need to improve. Scouts can look out for certain attributes and factors. Applying the variables more focused on Cost, teams can work out where they can buy players to replace or cover a place in their existing squad, under their budget or cheaper than their existing player. Imagine a team built under budget, that proves to be more effective than their existing, more expensive squad selection.

6. References

Morse, B., Burden, B. and Canterbury, F. (2017). *The Sports Tourists' Guide to the English Premier League*. Silver Way Publishing Inc., p.16.

Premierleague.com. (2018). *Premier League History, Origins & List of Past Champions*. [online] Available at: https://www.premierleague.com/history [Accessed 19 Nov. 2018].

Lee, A. (1997). *Modeling Scores in the Premier League: Is Manchester United Really the Best?*. Taylor & Francis, p.15.

Coleman, J. (2018). *The most expensive players in the Premier League who are desperately out of form*. [online] talkSPORT. Available at: https://talksport.com/football/426410/the-most-expensive-players-premier-league-out-of-form-wilshere-martial-benteke/ [Accessed 19 Nov. 2018].

Calvin, M. (2017). *No hunger in paradise - the players. the journey. the dream*. Random House.

Premierleague.com. (2018). *European qualification for UEFA competitions explained*. [online] Available at: https://www.premierleague.com/european-qualification-explained [Accessed 19 Nov. 2018].

Murray, W. (1998). *The world's game*. Urbana: University of Illinois Press.

Murray, T. (2018). *The 20 most popular rich-list football teams on social media*. [online] Business Insider. Available at: http://uk.businessinsider.com/the-20-most-popular-rich-list-football-teams-on-social-media-2018-8 [Accessed 19 Nov. 2018].

Eurasia Review. (2018). *The Value Of Top Footballers, Bubbles And Pitfalls Of Free Market – Analysis*. [online] Available at: http://www.eurasiareview.com/09102017-the-value-of-top-footballers-bubbles-and-pitfalls-of-free-market-analysis/ [Accessed 19 Nov. 2018].

Ellis, W. (2017). *Footballers are paid too much. The market has got it wrong - Reaction*. [online] Reaction. Available at: https://reaction.life/footballers-are-paid-too-much/ [Accessed 19 Nov. 2018].

Sky Sports. (2017). Premier League has highest percentage of foreign players – UEFA report. [online] Available at: https://www.skysports.com/football/news/11661/10725849/premier-league-has-highest-percentage-of-foreign-players-8211-uefa-report [Accessed 19 Nov. 2018].

Facts, P. (2018). Topic: Premier League. [online] www.statista.com. Available at: https://www.statista.com/topics/1773/premier-league/ [Accessed 19 Nov. 2018].

Krafft, S. (2018). What do those yellow and red cards mean in soccer?. [online] KSAZ. Available at: http://www.fox10phoenix.com/sports/what-do-those-yellow-and-red-cards-mean-in-soccer- [Accessed 19 Nov. 2018].

7. Development Details

Software used for code

PyCharm by Jetbrains

Language

Python3

Notes on code

Change data set link at top for weeks 0 and weeks 7 by removing and replacing '#'. It doesn't matter which graph week you use, as they both are the same, the only difference is the title of the graph says week 0 or 7.

There is some code I haven't used. Unsupervised method is in "Premier League Analysis Code" along with all graphs and diagrams. Supervised method has been sent in a separate file due to complications with packages.

Data set

https://www.kaggle.com/delayedkarma/fantasy-premier-league-20182019

CVS files week 0 and week 7

Packages Used

Pandas, Sklearn, Matplotlib, Seaborn, Scipy**,**