[**MIS 6380.002 – Data Visualization Project**](https://elearning.utdallas.edu/webapps/blackboard/execute/launcher?type=Course&id=_119109_1&url=)

**Betting in English Premier League**

**Project Report Content**

**Page No.**

1. Location of Source Data Files………….……………………………………………………………….…….1
2. Data Cleaning and Preparation………………………………………………………….…………………..1
3. Topic of Analysis………………………………………………………………………………………………..…2
4. Final Insights from Analysis…………………………………………………………………………………..2
5. Index of each Visualization and the Data encoding used………...………………………..….…..3
6. Why did we choose the specific data encoding for the Visualizations?...............................4
7. Discussion of the Storyline Chosen……………………………...…………………………………….…...6
8. Visual Data Narrative Process…………………………………………………………………………...…..7
9. Complete Description of Insights…………………...………………………………………………………8

**Location of Source Data Files:**

Primary dataset (payout rates data) - https://www.indatabet.com/ft-free-2in1.html

Secondary dataset (match results data) - http://www.football-data.co.uk/englandm.php

**Data Cleaning and Preparation:**

Primary dataset contained betting data from all football leagues around the world. For purpose of our analysis, we kept English Premier League betting data and filtered out the rest of the data in csv file itself. After that, we removed the data prior to the year 2009 as our analysis was going to be based on the betting data from year 2009 to 2018. Similarly, we modified our match result dataset to be in sync with the primary data set.

Once we had our csv files ready, we used python to manipulate data in match results file to get files suitable for Tableau analysis. Below are the changes done in the file using Python,

* Column Rename
* Three new columns (‘Winning\_Team’, ‘Losing\_Team’, ‘Payout\_Winning Team’) were added in match results file
* For draw result, ‘Winning\_Team’ and ‘Losing\_Team’ columns include both the names of teams participating in the game.
* Values (‘H’, ‘A’, ‘D’) were replaced with (‘Home’, ‘Away’, ‘Draw’) respectively from columns ‘Final\_R’, ‘SecondHalf\_R’ and ’FirstHalf\_R’

After importing the file from jupyter notebook, we added a new column ‘Match\_ID’ in both the betting and match results csv files. With the help of this ‘Match\_ID’ column, we created a join in Tableau to get 3345 records of data which was sufficient for our analysis purpose.

**Topic of Analysis**:

How can we bet accurately in English Premier League so that our probability of winning the bet is high?

**Final Insights from Analysis:**

**Betting based on location and match fixtures**

1. Teams like Manchester City, Manchester United, Chelsea, Arsenal and Tottenham have good records in their home ground when playing against specific opponents (*we have listed all the opponents in the end of the end of document*). Therefore, betting on these team should be preferred over other teams.
2. Teams like Manchester City, Arsenal and Tottenham have good records in away grounds when playing against specific opponents (*we have listed all the opponents in the end of the end of document*). Therefore, betting on these team should be preferred over other teams.
3. Teams like Stoke City and Tottenham are involved in most draws (*we have listed all the match fixtures in the end of the end of document*). Therefore, betting for a draw result is recommended.

**Betting based on teams involved in under and over 2.5 goals scored in a match**

1. Matches involving Liverpool, West Brom, Everton and Stoke City leads to less than 2.5 total goals in a match. Thus, betting for under 2.5 goals should be favored for matches with these clubs.
2. Matches involving Liverpool, Arsenal, Tottenham and Everton leads to more than 2.5 total goals in a match. Thus, betting for over 2.5 goals should be favored for matches with these clubs.

**Betting based on payout rates**

1. Betting on home team should be preferred when the payout rate for home team win is approximately 2.1
2. Betting on away team should be preferred when the payout rate for away team win is approximately 2.1 or 3.5
3. Betting for a draw should be preferred when the payout rate is approximately 2.1

**Index of each Visualization and the Data encoding used:**

|  |  |  |
| --- | --- | --- |
| **S.no.** | **Visualization** | **Data Encoding used** |
| 1. | Average payout rates for home team win, away team win, draw game, total goals in a match over or under 2.5 | Line Chart |
| 2. | Comprehensive final match result for Premier League | Pie Chart |
| 3. | Team wise distribution of matches won on home ground, away ground or resulted in a draw | Bubble Chart |
| 4. | Distribution of matches with over and under 2.5 total goals | Pie Chart |
| 5. | Most home wins against an opponent | Tree Map |
| 6. | Most away wins against an opponent | Tree Map |
| 7. | Match Fixtures frequently resulting in a draw | Tree Map |
| 8. | Teams involved in fixtures where average of total number of goals is under 2.5 | Stacked Bar Chart |
| 9. | Teams involved in fixtures where average of total number of goals is over 2.5 | Stacked Bar Chart |
| 10. | Betting analysis based on payout rates | R – Integration /Cluster Analysis |

**Why did we choose the specific data encoding for the Visualizations?**

1. Average payout rates for home team win, away team win, draw Game, total Goals in a match over or under 2.5 visualized **using *Line chart:***

We wanted to represent payout rates over the seasons for different categories: home team win, away team win, draw game, total goals in a match over or under 2.5 over the span of 9 seasons. Line charts are used to show time series data and continuous variables and to compare changes (in our case the payout rate) over the same period (9 seasons) for more than one group (home team win, away team win etc.). Accordingly, we represented the years on X axis and the payout rate (measurement) on Y axis.

1. Comprehensive final match result for Premier League visualized **using *pie Chart***

A pie chart is best used when we are trying to work out the composition of something. Since, we had categorical data, using a pie chart worked well as each slice represented a different category and showed its contribution: Percentage of matches in which home team won, away team won or ended up in a draw.

1. Team wise distribution of matches won on home ground, away ground or resulted in a draw visualized **using *Bubble chart***

When our data has three data series that each contain a set of values it is more eye pleasing and easy to interpret if we use a bubble chart over a scatter plot. A bubble chart is a variation of a scatter chart in which the data points are replaced with bubbles, and an additional dimension of the data is represented in the size of the bubbles. In addition to the x values and y values that are plotted in a scatter chart, a bubble chart plots x values, y values, and z (size) values. In our chart we have 3 different colored bubbles representing: home team wins, away team wins and draws. Each of the bubbles represents a team with their number of wins at home or away along with match fixtures that resulted in draw. The sizes of these bubbles determine the matches in the respective categories.

1. Distribution of matches with over and under 2.5 total goals visualized **using *Pie chart***

A pie chart is best used when we are trying to work out the composition of something. Since, we had categorical data then using a pie chart worked well as each slice represented a different category and showed its contribution: percentage of matches in which goals are scored are under or over 2.5.

1. Most home wins against an opponent, most away against an opponent win, most match fixtures resulting in a draw visualized **using *Tree Map***

The tree map chart is used for representing hierarchical data in a tree-like structure. Data, organized as branches and sub-branches, is represented using rectangles, the dimensions and plot colors of which are calculated w.r.t the quantitative variables associated with each rectangle—each rectangle represents two numerical values. This makes the at-a-glance distinguishing between categories and data values easy. Thus, in our visualizations, average payout rate of the winning team is represented by the block sizes and the color encoding helps to visually understand number of matches won. In addition, match fixtures are reflected by labels making it easy for the audience to know the team involved in most number of home win, away win and draw result.

1. Teams involved in fixtures where average of total number of goals is under 2.5 / over 2.5 visualized **using *Stacked Bar Chart***

Bar charts are a type of graph that are used to display and compare the number, frequency or other measure (e.g. mean) for different discrete categories of data. In our visualization we are showing different fixtures where the match on an average resulted over/under 2.5 goals along with average goals contribution from each team. Bar Charts makes understanding this easy for the reader.

1. Betting Analysis based on payout rate **using *Cluster Analysis***

In Clustering analysis, a set of objects are grouped in such a way that objects in the same group (called a cluster) are more similar (in some sense or another) to each other than to those in other groups (clusters). In our analysis, clusters were created using K-means algorithm in R programming to find set of payout rates for a given match outcome i.e. Home team wins, Away team or Draw. Clusters formed helped us to learn on what match result to bet, whether on home team win, away team win or a draw in case of specific combinations(clusters) of payout rates (Home team payout rate, away team payout rate, match resulting in draw payout rate, goals under/over 2.5 payout rate). Looking at the payouts and match results in each cluster, a person can bet on a match result with better chance of winning the bet.

**Discussion of the Storyline Chosen:**

We tried to extend our learning from the lectures in a pragmatic way. We incorporated the Strategic Story Telling (Shaw, Brown, Bromiley Harvard Business Review May 1998) discussed during lectures in our story line.

**We divided our story into the following 3 broad categories:**

1. Set the stage

Our goal is to come up with strategies/ trends observed in English Premier League matches and betting so that we can help the viewer make more accurate and profitable bets. We started with by giving a brief explanation about English Premier League and how the teams and games are structured in a season. We introduced some of key terminologies used in betting in EPL to get the viewers to understand betting better.

1. Introducing Drama Conflict

Here, we raised the few questions,

Why betting based on the higher payout rates is not a wise decision?

Why betting on home team is not always recommended even though the past match results accounts for 46% of home team win which is a significant portion of all the match results?

How can we bet based on number of goals scored in a game? As matches have ended in under or over 2.5 total goals equal number of times, making it difficult to choose one outcome for betting.

We presented a brief overview of the payout rates trends over the years followed by actual match results distribution for the period starting from August 2009 to March 2018. For the same, we introduced visualizations “Average payout rates for home team win, away team win, draw Game, total Goals in a match over or under 2.5”, “Comprehensive final match result for Premier League”, “Team wise distribution of matches won on home ground, away ground or resulted in a draw”, “Distribution of matches with over and under 2.5 total goals” to raise the question/tension whether betting on a few generalized trends will result in accurate betting. We raised the issue that betting on a few trends observed over a period might not be a best idea.

1. Resolution

We came up with logical solutions to the tension we created and answered the questions that instead of basing our bets on generalized trends, we had to be granular in our approach. We should be concentrating more on the combined factors like: match location, home team, away team, past payout rates, goals scored in previous matches; before placing a bet. We brought attention to these using the visualizations “Most home wins against an opponent”, “Most away against an opponent win”, “most match fixtures resulting in a draw”, “Teams involved in fixtures where average of total number of goals is under 2.5 / over 2.5”, “Betting Analysis based on payout rate”.

**Visual Data Narrative Process:**

1. Audience

We engaged the audience: viewers who are aware/ not aware of EPL, viewers who are frequently bet/ not bet frequently. We engaged all audience by highlighting about EPL games and betting trends. Insights were revealed from the data which can allow the audience bet accurately without losing out their money.

1. Evidence

We presented appropriate visualizations and data to support the claims. We demonstrated the knowledge we gained through the visualization in an incremental way. We layered the information to unfold the sort in an interesting way and revealing important trends and facts. We achieved this by compounding builds in visualization and by sequencing different types of visualizations, drilling deeper into a single visualization.

1. Rhetorical Argument

Logical arguments are taken into consideration for betting rather than placing bets purely based on just on a few general trends. We recommended that it is smarter to base our betting based on specific teams, match location and payout rates.

1. Narrative Structure

As mentioned above, we incorporated: setting the scene by giving a brief introduction on EPL and betting. Challenges were defined, if we bet based on generalizations as each fixture is unique. Finally, we recommended ways to resolve the tension by providing visualizations which discussed outcomes of specific fixtures and stressed on the importance of carefully studying a fixture before placing a bet.

**Complete description of insights:**

**Visualization 1** - Average Payouts for Home team win, Away team win, Draw game, Total Goals Over and Under 2.5 (August 2009 to March 2018)

* Payouts for home team win, away team win and draw result are higher in year 2017 and 2018 as compared to the payouts in previous years, while payouts for over and under 2.5 goals has not changed much for the past 9 years

**Visualization 2 -** Overall Actual Match Result Statistics

* Manchester City, Manchester United and Chelsea are teams winning the most number of games both at home and away grounds. We should look out for these teams before placing a bet on their matches.

**Visualization 3 -** Most Home wins against an opponent

* Betting on the below mentioned home teams is recommended when playing against an opponent:

|  |  |  |  |
| --- | --- | --- | --- |
| **Home team** | **Away team** | **Avg. Payout Winning Team** | **Number of matches** |
| Swansea | West Brom | 2.243 | 6.000 |
| West Brom | Sunderland | 2.118 | 6.000 |
| Liverpool | Stoke City | 1.495 | 6.000 |
| Manchester City | 2.583 | 6.000 |
| Tottenham | Sunderland | 1.427 | 7.000 |
| Everton | 1.835 | 6.000 |
| Arsenal | 2.428 | 6.000 |
| Swansea | 1.478 | 6.000 |
| Everton | Sunderland | 1.638 | 6.000 |
| Manchester United | Sunderland | 1.273 | 6.000 |
| Stoke City | 1.319 | 7.000 |
| Everton | 1.545 | 6.000 |
| Arsenal | 1.938 | 6.000 |
| West Ham | 1.247 | 6.000 |
| Aston Villa | 1.270 | 6.000 |
| Tottenham | 1.690 | 7.000 |
| Liverpool | 1.867 | 7.000 |
| Chelsea | Sunderland | 1.258 | 6.000 |
| Stoke City | 1.233 | 7.000 |
| West Brom | 1.227 | 6.000 |
| Everton | 1.478 | 6.000 |
| Arsenal | 1.834 | 7.000 |
| Swansea | 1.252 | 6.000 |
| Manchester United | 2.092 | 6.000 |
| Manchester City | Sunderland | 1.317 | 6.000 |
| Stoke City | 1.287 | 7.000 |
| West Brom | 1.248 | 8.000 |
| West Ham | 1.283 | 7.000 |
| Swansea | 1.250 | 6.000 |
| Aston Villa | 1.373 | 7.000 |
| Tottenham | 1.648 | 6.000 |
| Chelsea | 2.475 | 6.000 |
| Newcastle Utd | 1.270 | 6.000 |
| Arsenal | Stoke City | 1.369 | 8.000 |
| West Brom | 1.327 | 7.000 |
| Everton | 1.493 | 6.000 |
| West Ham | 1.327 | 6.000 |

**Visualization 4 -** Most Away wins against an opponent

* Betting on the below mentioned away teams is recommended when playing against an opponent:

|  |  |  |  |
| --- | --- | --- | --- |
| **Away team** | **Home team** | **Avg. Payout Winning Team** | **Number of matches** |
| Arsenal | Aston Villa | 1.740 | 5.000 |
| Sunderland | 1.748 | 5.000 |
| West Ham | 1.782 | 5.000 |
| Everton | 2.194 | 5.000 |
| Manchester City | West Brom | 1.560 | 7.000 |
| Manchester United | 2.898 | 5.000 |
| Newcastle Utd | 1.748 | 5.000 |
| Tottenham | Aston Villa | 2.038 | 5.000 |
| Stoke City | 2.075 | 6.000 |
| Swansea | 2.026 | 5.000 |
| Chelsea | Sunderland | 1.618 | 6.000 |
| Stoke City | 1.606 | 5.000 |
| Liverpool | Aston Villa | 2.012 | 6.000 |
| West Ham | 1.658 | 5.000 |
| Manchester United | Sunderland | 1.474 | 5.000 |
| West Brom | 1.584 | 5.000 |

**Visualization 5 -** Match Fixtures frequently resulting in Draw

* Betting for a draw should be favored for the below match fixtures:

|  |  |  |  |
| --- | --- | --- | --- |
| **Home team** | **Away team** | **Avg. Payout Winning Team** | **Number of matches** |
| Stoke City | West Brom | 3.250 | 4.000 |
| Everton | 3.290 | 5.000 |
| Manchester City | 3.600 | 5.000 |
| Manchester United | 3.820 | 4.000 |
| Arsenal | Chelsea | 3.348 | 4.000 |
| Manchester City | 3.390 | 5.000 |
| Liverpool | 3.473 | 4.000 |
| Tottenham | West Brom | 4.470 | 5.000 |
| Chelsea | 3.288 | 4.000 |
| Manchester United | 3.338 | 4.000 |
| Everton | Aston Villa | 3.595 | 4.000 |
| Liverpool | 3.300 | 4.000 |
| Tottenham | 3.375 | 4.000 |
| Liverpool | Everton | 3.850 | 4.000 |
| Chelsea | 3.398 | 4.000 |
| Sunderland | 4.458 | 4.000 |
| Sunderland | West Brom | 3.238 | 4.000 |
| Everton | 3.338 | 4.000 |
| West Ham | 3.225 | 4.000 |
| West Ham | West Brom | 3.370 | 5.000 |
| Stoke City | 3.375 | 4.000 |
| West Brom | Aston Villa | 3.475 | 4.000 |
| Tottenham | 3.375 | 4.000 |
| Aston Villa | Sunderland | 3.410 | 5.000 |
| Manchester City | Everton | 4.525 | 4.000 |
| Swansea | Crystal Palace | 3.425 | 4.000 |

**Visualization 6 -** Teams Involved in Fixtures where Average Total Goal scored is Under 2.5

* Average total goal scored of 1.75 is the lowest when:

Manchester United plays Manchester City in their home ground

Manchester United plays Sunderland in their home ground

Manchester City plays Sunderland in their home ground

* Betting for total goal scored Under 2.5 should be preferred for the below match fixtures:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Away team** | **Home team** | **Average Away Team Goals** | **Average Home Team Goals** | **Average Total Goals** |
| Liverpool | Chelsea | 1.375 | 1.125 | 2.500 |
| Everton | 1.375 | 1.000 | 2.375 |
| Sunderland | 1.250 | 0.750 | 2.000 |
| Manchester United | Manchester City | 0.750 | 1.000 | 1.750 |
| Sunderland | 1.250 | 0.500 | 1.750 |
| Arsenal | Sunderland | 1.625 | 0.625 | 2.250 |
| West Ham | Manchester United | 0.375 | 2.125 | 2.500 |
| West Brom | Liverpool | 1.000 | 1.375 | 2.375 |
| Tottenham | 0.875 | 1.375 | 2.250 |
| Everton | 0.875 | 1.125 | 2.000 |
| Manchester City | Stoke City | 1.500 | 1.000 | 2.500 |
| Sunderland | 1.000 | 0.750 | 1.750 |
| Tottenham | West Ham | 1.125 | 1.250 | 2.375 |
| Sunderland | 1.250 | 1.000 | 2.250 |
| Everton | Stoke City | 0.875 | 1.000 | 1.875 |
| West Brom | 1.125 | 0.875 | 2.000 |
| Sunderland | 1.125 | 1.125 | 2.250 |
| Stoke City | Liverpool | 0.250 | 2.000 | 2.250 |
| West Brom | 1.125 | 0.750 | 1.875 |
| Sunderland | 0.625 | 1.750 | 2.375 |
| Sunderland | Arsenal | 0.375 | 1.625 | 2.000 |
| Stoke City | 0.625 | 1.250 | 1.875 |

**Visualization 7 -** Teams Involved in Fixtures where Average Total Goal scored is Over 2.5

* Average total goal scored of 3.889 is the highest when Manchester City plays Manchester United in their home ground.
* Betting for total goal scored Over 2.5 should be preferred for the below match fixtures:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Away team** | **Home team** | **Average Home Team Goals** | **Average Away Team Goals** | **Average Total Goals** |
| Liverpool | Manchester City | 2.222 | 1.000 | 3.222 |
| Manchester United | 2.000 | 1.222 | 3.222 |
| Arsenal | 1.778 | 1.444 | 3.222 |
| Tottenham | 1.667 | 1.444 | 3.111 |
| Manchester United | Chelsea | 2.000 | 1.000 | 3.000 |
| Everton | 1.444 | 1.222 | 2.667 |
| Arsenal | Manchester City | 2.111 | 1.667 | 3.778 |
| Chelsea | 2.444 | 0.778 | 3.222 |
| Liverpool | 2.222 | 1.556 | 3.778 |
| Tottenham | 1.778 | 1.111 | 2.889 |
| Everton | 1.333 | 2.222 | 3.556 |
| Manchester City | Chelsea | 1.222 | 1.333 | 2.556 |
| Manchester United | 1.556 | 2.333 | 3.889 |
| Liverpool | 2.333 | 1.222 | 3.556 |
| Tottenham | Manchester City | 2.556 | 1.111 | 3.667 |
| Manchester United | 1.889 | 0.667 | 2.556 |
| Arsenal | 2.333 | 1.111 | 3.444 |
| Liverpool | 1.889 | 1.000 | 2.889 |
| Chelsea | Liverpool | 1.333 | 1.333 | 2.667 |
| Tottenham | 1.667 | 1.444 | 3.111 |
| Everton | 1.444 | 1.444 | 2.889 |
| Stoke City | 0.778 | 1.889 | 2.667 |
| Everton | Chelsea | 2.333 | 1.000 | 3.333 |
| Manchester United | 2.000 | 0.778 | 2.778 |
| Arsenal | 2.000 | 0.778 | 2.778 |
| Liverpool | 2.111 | 0.556 | 2.667 |
| Tottenham | 1.889 | 0.778 | 2.667 |
| Stoke City | Manchester City | 2.556 | 0.333 | 2.889 |

**Visualization 8 -** Betting based on payout rates (Cluster Analysis)

We will be focusing particularly on Cluster 3, 4 and 5 because,

Cluster 3 consists of the most number of home wins

Cluster 3 and 5 consists of the most number of away wins

Cluster 4 consists of the most number of draw

1) As most of the home team wins are present in Cluster 3, betting on home team should be preferred when the payout rate is approximately 2.1

From Cluster 3,

|  |  |
| --- | --- |
| **Expected match results** | **Median Payout rates** |
| Away team win | 3.4 |
| Home team win | 2.1 |
| Draw game | 3.3 |
| Goals over 2.5 | 2 |
| Goals under 2.5 | 1.8 |

2) As most of the away team wins are present in Cluster 3 and 5, betting on away team should be preferred when the payout rate is approximately 2.1 or 3.5

From Cluster 3,

|  |  |
| --- | --- |
| **Expected match results** | **Median Payout rates** |
| Away team win | 2.1 |
| Home team win | 3.4 |
| Draw game | 3.4 |
| Goals over 2.5 | 1.9 |
| Goals under 2.5 | 1.9 |

From Cluster 5,

|  |  |
| --- | --- |
| **Expected match results** | **Median Payout rates** |
| Away team win | 3.5 |
| Home team win | 3.4 |
| Draw game | 2.1 |
| Goals over 2.5 | 2 |
| Goals under 2.5 | 1.8 |

2) As most of the draw games are present in Cluster 3, betting for a draw should be preferred when the payout rate is approximately 2.1

From Cluster 4,

|  |  |
| --- | --- |
| **Expected match results** | **Median Payout rates** |
| Away team win | 3.4 |
| Home team win | 3.39 |
| Draw game | 2.1 |
| Goals over 2.5 | 2 |
| Goals under 2.5 | 1.8 |