## Introduction:

Soccer is a team sport played between two teams. It is a well known and loved sport, with approximately 3.5 billion fans around the world. But as much as people enjoy sitting in the crowd, at their favorite bars or at home and watching their favorite team plays, there is another aspect to this sport which makes the games even more interesting for their viewers – and that’s betting.

In general, sports betting refers to attempts at predicting sports results, while placing a wager on the outcome. In soccer, there are many aspects of betting – one can bet on simple outcomes in the game, such as the result of the game (winning team), or the exact result (number of points per team). However, betting can be much more complicated as one can also try and predict many other outcomes, such as number of “offsides”, number of “corners”, first team to achieve a “corner”, etc.

The possibilities for the gamblers are endless, but the one that gains the most, statistically, is the betting company. There are many soccer betting companies out there, aiming to increase their income. The main way in increasing the company’s income is by setting the betting odds in the optimal way.

The betting odds can be seen as a function of the likelihood of the outcome – each outcome receives a number, which would increase as the likelihood for that outcome decreases. For example, in a simple case, assume the value for team A’s win was set to and a gambler placed a bet of dollars on that outcome. If this outcome happened, the gambler would gain (after “losing” the they bet) dollars. Therefore, if the company set to be very low, they would lose less money to that gambler.

The questions we aim to solve:

So, the question is, how does the company choose the optimal betting odds for a game? Many things can affect a game’s outcome, such as the teams’ levels, the players’ mood, the weather, etc.

To understand how to answer the question above, we try to solve the following questions:  **How does the outcome of a match affect the team’s outcome in their following match?  
How does the outcome of a match affect the betting odds in the team’s following match?**

In this project we will use the different methods we learned in the course and calculate the ATE. By using matching and learners, and some more statistical hypotheses testing, we plan to answer this question.

## The Data:

The data we use in this project is taken from Kaggle. It is a database, composed of multiple tables, but the tables we will be using are:

* “Match” – a table of almost 26,000 records of matches, including the participating teams, dates, outcomes, and betting odds given to each team by multiple betting companies.
* “Player Attributes” – A table that describes each player, including some statistics and ratings.
* “Team” – A table with information about different teams.

The full database can be found [here](https://www.kaggle.com/datasets/hugomathien/soccer?select=database.sqlite).

## Data Processing:

Now we wish to process the data to get a usable data for our research.

We start by creating two datasets:

### Match dataset

Using “Match” table in the database, we create a dataset that contains information about each match. The information can be split to two groups:

1. General information: season, date, match id and participating teams' ids.
2. Match information:
   1. Number of goals per team.
   2. Goal difference (“home” team goals minus “away” team goals).
   3. Average betting odds given to each outcome (home team wins, away team wins or draw).
   4. Details about player rating per team (changes between different matches):  
      Minimum, maximum, average, std and median.

### Season dataset

This dataset contains seasonal information about all teams, which can also be split into two groups:

1. General information: season and team id.
2. Team information:
   1. Number of wins.
   2. Number of draws.
   3. Total league points (3 point per win and 1 point per draw).
   4. Number of goals in the entire season.

### Player Attributes dataset

This dataset contains information about the team players and many attributes. From this table took the players rating.

We take these three datasets and merge them, so for each match we have the match’s information and the team’s seasonal information.

Our goal is to find whether a match’s result affects the next match’s result or the next match’s betting odds. Therefore, for each team, we need to information about two subsequent games in which the team played.

Therefore, we create the final dataset, which holds for each match all the information we discussed earlier and holds the same information for the next match of the team.

This way every record in the data represents a certain team. Each record includes information about the team that played the game (notated as the "**row team**"), their rival team in the previous game (notated as "**previous rival**", and their rival team in their following game (notated as the "**next rival**").

## Defining the Treatment, Outcome and Covariates:

### Treatment

We defined the if and only if a team lost the previous game with a goal difference that is equal to one. And if and only if a team won the previous game with a goal difference that is equal to one. We decided to only look at goal difference equal to one to hold the SUTVA assumption (more on that later), and to observe cases where the teams were almost matched but one team managed to get the win.

### Outcome

We were interested in two questions and therefore we defined two sets of outcomes (each outcome was checked independently). The first, being the goal difference of the subsequent game of the team. The second, being the betting mean bet odds of the subsequent game.

### Covariates

* Prev home - Indicating whether the "**row team**" played at home at the previous game
* Next home - Indicating whether the "**row team**" played at home at the next game
* Time between games – number of days between the two games

For each of the "**row team**", "**previous rival**", "**next rival**" we calculated the

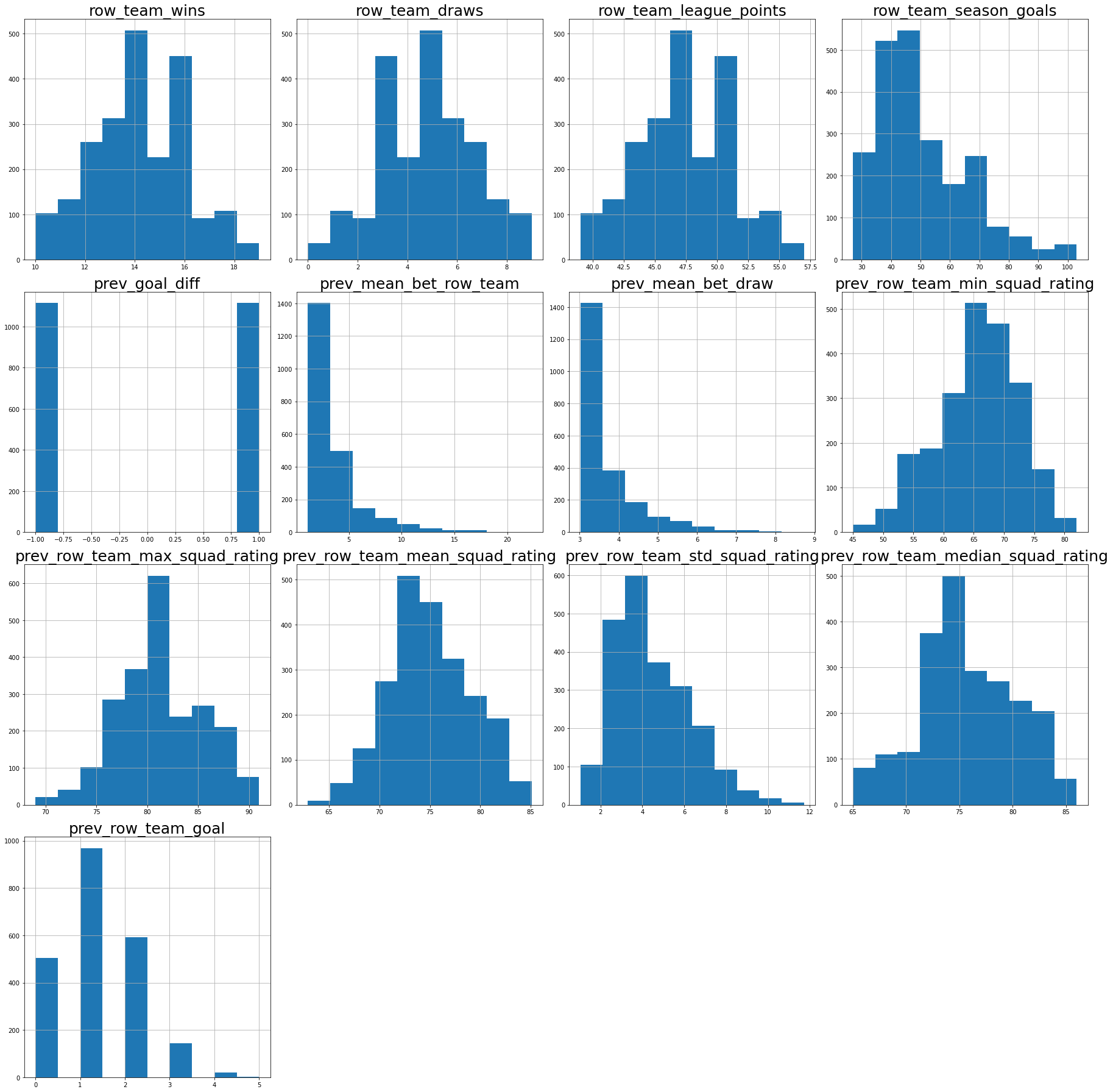
* Team wins - number of wins the team had in the season
* Team draws - number of draws the team had in the season
* Team league points - score of the team in the season (3 points for a win, 1 for a draw, 0 for a loss)
* Team goals – number of goals scored in the season.
* Squad rating measures – the min, max, std and median of the team's players rating taken from FIFA

We hoped that these covariates would explain the quality of the teams in a match, and other factors that could influence the goal difference like playing in the how court and the time between games.

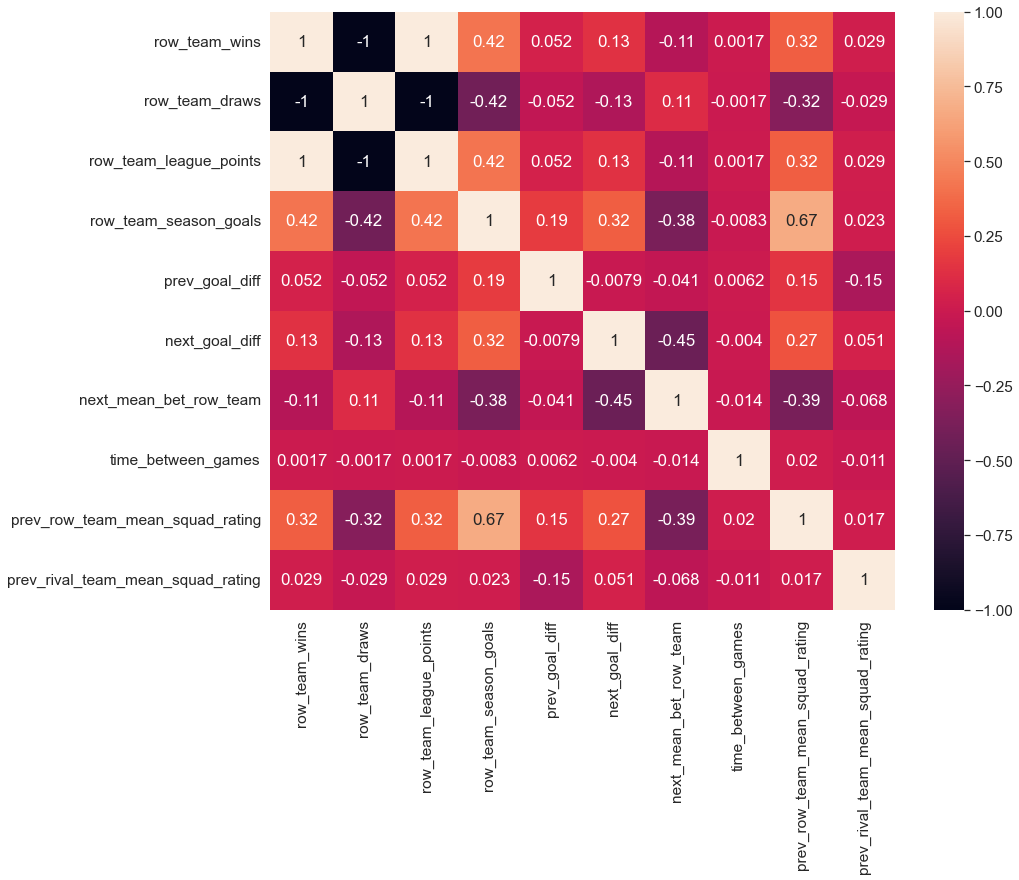
## Exploratory Data Analysis:

To get a better understanding of our data we performed an EDA on our data.

First, to see the distribution of the features we plotted the histograms of a few key features.



Then, to understand the relationship between the features we plotted a correlation matrix.



From this correlation matrix we see a high correlation between the mean squad rating and the season goals. We see a very weak correlation between the goal difference in the previous game and the in next game. There is also a weak correlation between the goal difference in the previous game and the mean betting odds in the next game.

Then, we wanted to make sure that that the distribution according to treatment was balanced. Not surprisingly we have the same number of records with T=0 and with T=1. That is because every game appears in two records. One record that refers to the home team, and the other to the away team. This means that every game that ended with a goal difference of one, contributed to a record where the previous goal difference was -1 (T=1), and a record where the previous goal difference was 1 (T=0)



We wanted to see the distribution of the next game's goal diff according to the treatment. We created two boxplots to see that, one for records with T=1 and the other with records with T=0. It seams that the distribution is quite similar. The main difference is that for T=1 there were more outliars.

Chart, box and whisker chart

Description automatically generated

Then, we looked at the same boxplot but this time measure the next game's mean bet for the team. Again the distributions look rather similar.

Chart, box and whisker chart

Description automatically generated

## The Course Assumptions:

### Stable Unit Treatment Value Assumption (SUTVA)

This assumption requires two things. First, the potential outcomes for any unit do not vary with the treatments assigned to other units. It seems reasonable that the loss of a certain team does not affect the performance of a different team on the next match.   
Second, for each unit, there are no different forms or versions of each treatment level, which lead to different potential outcomes. This part of the assumption also partially holds because we only looked at games with a goal difference of 1. But it is important to note that there could be two games that ended with the same goal difference, but in one game the teams were very even, and in the other one team dominated the match but managed to score only one goal more than the other team. Also, a goal difference of one includes both games that ended in 0-1 and games that ended in 3-2.

### Consistency

This assumption requires that for a unit that receives treatment T, we observe the corresponding potential outcome This requirement does apply in our case, because we can see the result in the next match of the team.

### Ignorability

This assumption requires that the potential outcomes are independent of treatment assignment, conditioned on observed covaries x. It is very hard for this assumption to fully hold. In our case there are many things that could affect the consecutive games of a team, like personal issues of the players, weather and more. We did try to collect as many covariates as possible that would explain the quality of the opposing teams, and relevant information about the game.

### Common Support

This assumption requires that there is a positive probability for each treatment to be assigned to any of the data points. To validate this assumption, we created a plot showing the propensity score for data points with T=1 and data points with T=0. To calculate the propensity scores, we train a logistic regression model on X with T as the labels. Thus, for each combination of control variables, we get a probability for getting treatment T. Initially, we received the plot shown in *Figure 1* below which doesn't have a good overlap. To get a better overlap trimmed by propensity, removing records with propensity lower than 0.2 or greater than 0.75. After doing that we got the plot shown in *Figure 2* which shows a better overlap.

Chart, histogram

Description automatically generated

Figure 1

Chart, histogram

Description automatically generated

Figure 2

## The Methods:

First, we want to understand the effect of a match’s outcome on the next match outcome, and therefore we set “prev\_win” variable as the treatment (T) and “next\_win” variable as the outcome (Y). All other variables are the control variables (X).

We use the methods taught in class to calculate ATT.

### IPW

At first, we use the propensity scores we explained in the previous section.

We use the propensity score to calculate ATT using IPW.

* IPW formula:   
  When is the propensity score of .

### S-Learner (with and without interactions)

Here we fit a linear regression model on X (along with T) with Y as labels (predict next win given the control parameters). We fit the model on all the data, and then use it to predict the treated data (teams that lost their previous games) and the same data, this time with treatment 0 (predict the Y value in case they weren’t treated). We did this both with and without interactions within the non-categorical features.   
We then calculate the ATT using this formula:

### T-Learner

Here we train two linear regression models: One on the treated data () and another one on the untreated data ). We use only the control variables to train the model (as the variable is the same for all data for each model anyway). We use each model to predict the value for the treated group, and calculate the ATT using this formula:

### Matching

Here we estimate the ATE by taking the mean of the estimated ITE. We tried 1NN matching and 10 NN matching. We will formally explain the procedure for 1NN, but in the case of 10NN we simply take the mean ITE among the 10 nearest records.

For every record we find the most similar record, that got a different treatment. Formally:

Them we estimate the ITE

Finally, we estimate the ATE by taking the mean ITE

## Results

### Question 1 - (TODO fill this)

### Table Description automatically generated

### Question 2 - (TODO fill this)

## 

## Possible Weaknesses- (TODO fill this)

## Discussion - (TODO fill this)

## Bibliography - (TODO fill this)