Sinead O'Connor

CAPSTONE 1:
MILESTONE REPORT
PREDICTING MLB OUTCOMES

The Problem:
What is probability of the home team winning a regular-season game in Major League Baseball?

The Client:
Billions of dollars are bet on sports each year in America. Given this and the fact that are 2,430 regular season Major League Baseball games each year, it could be very useful to sports betting companies to have an accurate model to predict the outcomes of baseball games. They can use the results to set betting odds.

The Data:
I will be using Game Log data sets from Retrosheet.org. The game logs contain information on the starting players of each team as well as defensive, offensive, and pitching statistics for both the home team and the away team. It also includes information such as date, day of the week, time of day (day/night), and attendance. The game-log files are imported from the website using urllib request. I gathered the game logs for the 2009-2018 regular seasons.
https://www.retrosheet.org/gamelogs/

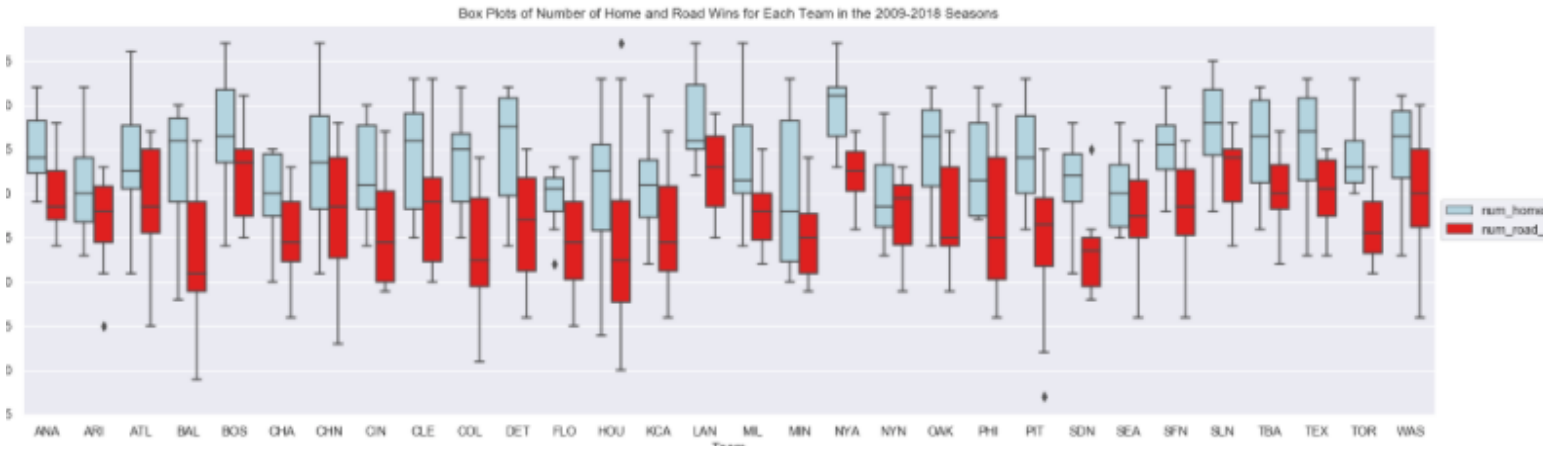
Since starting pitching is considered to be a very important part of the game, I will also use starting pitching logs from baseballmusings.com. The pitching logs have information about the starting pitchers for each game such as Innings Pitched, and number of walks, strikeouts, and earned runs. The starting pitcher logs are scraped from the web using BeautifulSoup.
https://www.baseballmusings.com/ChooseStarterTeam.html

I cleaned the Retrosheet gamelogs and the Baseball Musings starting pitcher logs individually. Once they were cleaned, I merged them. The resulting dataframe has 24,298 rows, each representing a regular-season game, and 205 columns, each giving a different piece of information about the events of the game.

Exploratory Data Analysis – Data Storytelling:
In the first part of Exploratory Data Analysis, Data Storytelling, I explored many questions.

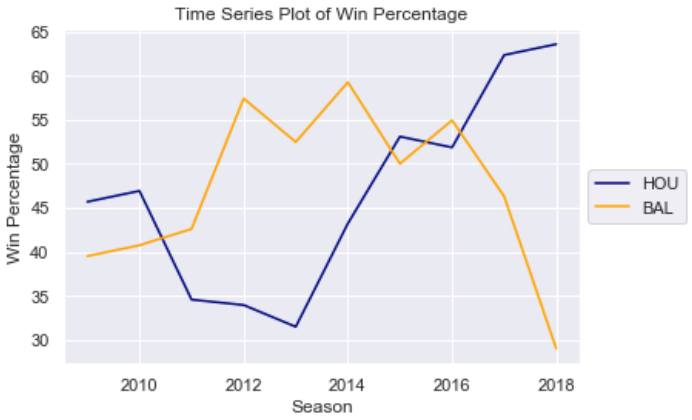
I found that in the 2009-2018, the home team won 53.7% of games. In each individual season, the home teams have won a higher percentage of games. In addition, each team higher percentage of their home games than their road games.

The below boxplots show the number of home games and road games each team has won in the 2009-2018 seasons.



These boxplots suggest that each team does play better at home than on the road. This seems to be particularly true of teams such as the Los Angeles Angels (ANA), and the New York Yankees (NYA), and San Diego Padres (SDN), and San Francisco Giants, whose box plots for home wins and road wins have very little overlap. Teams such as the New York Mets (NYN), and the Seattle Mariners (SEA), whose boxplots do not seem to have a lot of overlap.

Also, while some teams such as the New York Yankees consistently have winning records, other teams have not been very consistent, such as the Houston Astros and the Baltimore Ravens.



None of the findings in the Data Storytelling were particularly surprising. Some other conclusions are summarized below:

- The American League won 56.8% of games at home when playing against a National League team. The National League team won 50.3% of games at

home when playing American League team, so it seems the American League might be slightly better.

- As average attendance increases, so does the number of games won. It could be that the large crowds help the home team perform better, or it could be that there is higher attendance because the team is winning more, so may be more exciting to watch.

- The number of errors made per team per game is very low, which is a good sign since it is professional baseball, and there does not seem to be a big difference between the number of errors made by the home team and the visiting team.

- Starting pitchers on the winning team do seem to pitch more innings than the starting pitcher on the losing teams. This makes sense because if the starting pitcher is performing poorly, they will be taken out of the game earlier.

- In 52.8% of games in the 2009-2018 seasons the team with the higher on-base percentage won the game. This is not too surprising because in most games the average on-base percentage of the opponents are not very far apart, and there are many factors that go into winning an individual game.

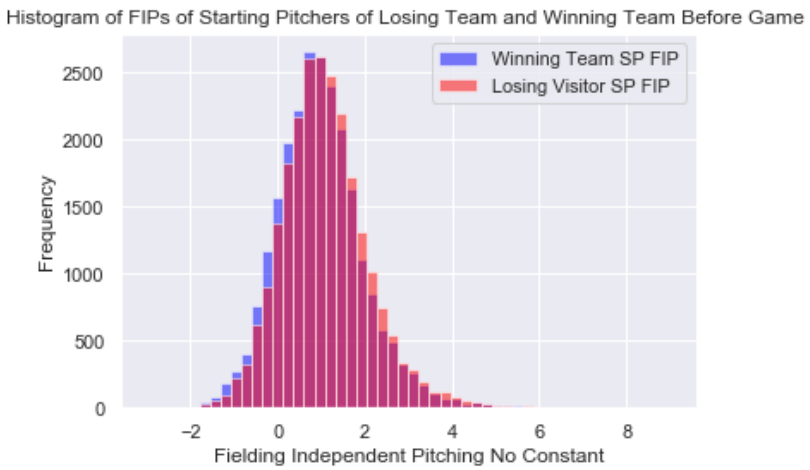
Exploratory Data Analysis – Inferential Statistics:

I did some more data wrangling, computing averages for starting pitchers and team averages. I also added columns for major statistics such as On-Base Percentage, Isolated Power, Earned Run Averages, Fielding Independent Pitching, Batting Average on Balls in Play, and more. I computed the averages per team per season, and used data from the end of the prior season to fill the first 5 games of each season. The averages are computed using only information from before the game takes place, so it can be used to predict the outcome of future games.

I then moved on to Inferential Statistics where I explored many hypotheses.

Here is a summary of the findings:

- The proportion of home games won by the American League and that won by the National League from the 2009-2018 seasons (53.974% and 53.488%, respectively) was not statistically significant.
- There is a statistical significance between the teams On-Base Percentage at home and on the road. It tends to be higher at home. We must keep in mind that this could be due to the very large sample size.
- There is a statistically significant difference between the Fielding Independent Pitching stat prior to the game of the starting pitcher on the team that goes on to win the game and the starting pitcher on the losing game.



I also identified collinear variables by creating a heat map of the correlation matrix using all columns that included averages, as well as the column that indicates whether the home team won or lost.

Of the variables that were highly correlated, most were trivial. For example, earned runs (AvgTER) and individual earned runs(AvgER) as well as the ones that are functions of other variables, for example BIP (balls in play) is a function of K (strikeouts). I will keep in mind to not use highly correlated variables in the analysis.

What's Next?:

I now need to build classification models using information about the teams and starting pitchers to predict the probability of the home team winning a game.

As a next step, I will build a baseline logistic regression model using all of the historical data from the 2009-2018 seasons. I will see how that performs and make adjustments accordingly.