MOYINOLUWA JAWO

215991040

CS 240 – EXPLORATORY DATA ANALYSIS Project

SECTION 1

1. What is the relationship between the amount of games won whenever teams play at home and the amount of games lost whenever teams play at the opponents place, away from home?
2. Is there a positive or negative correlation between the variable ‘homeWon’ and the variable ‘awayLost’?
3. Is there a difference in the means of the variable ‘homeWon’ and the variable ‘awayLost’?

Null hypothesis: There is no difference between the means of variable ‘homeWon’ and variable ‘awayLost’. The distributions are the same

Alternative hypothesis: There is a difference between the means of variable ‘homeWon’ and variable ‘awayLost’. The distributions are not the same

SECTION 2

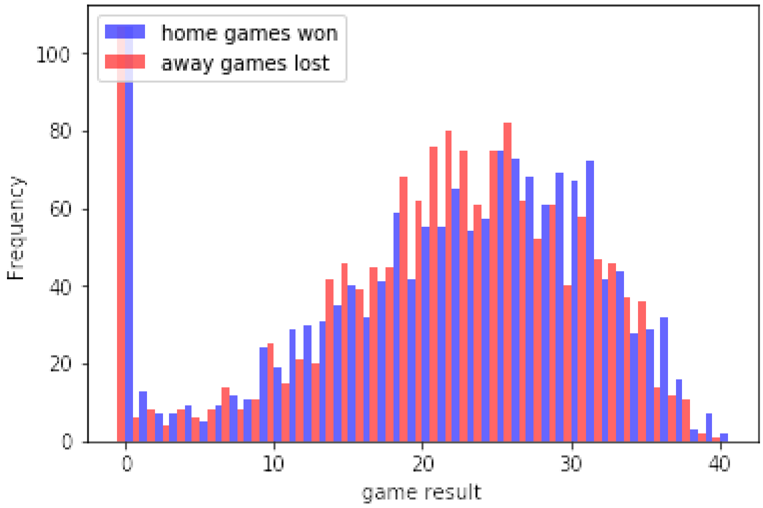
A dataframe, df, was created from the file ‘basketball-teams.csv’. In this project, the variables, ‘homeWon’ and ‘awayLost’ were used. A summary of the values of these variable shows a high number of zeros, 0, recorded which represents the amount of times a team either wins or loses 0 games in total. This project is basically comparing the games won at home and games lost away of each of the teams. The values in each variable ranges from 0-40 with different frequencies.

SECTION 3

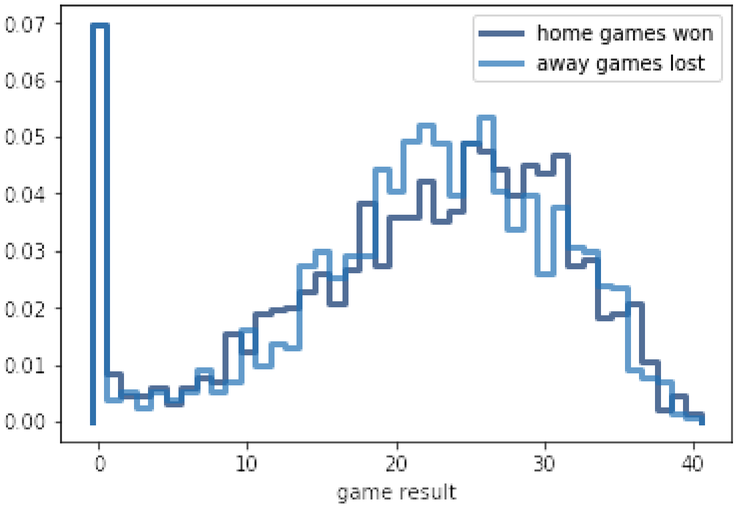
Below we can see some descriptive statistics for both variable ‘homeWon’ and variable ‘awayLost’:

‘homeWon’	‘awayLost’
count 1536.000000	count 1536.000000
mean 21.361328	mean 21.363281
std 9.846326	std 9.429432
min 0.000000	min 0.000000
25% 15.000000	25% 16.000000
50% 23.000000	50% 23.000000
75% 29.000000	75% 28.000000
max 40.000000	max 40.000000

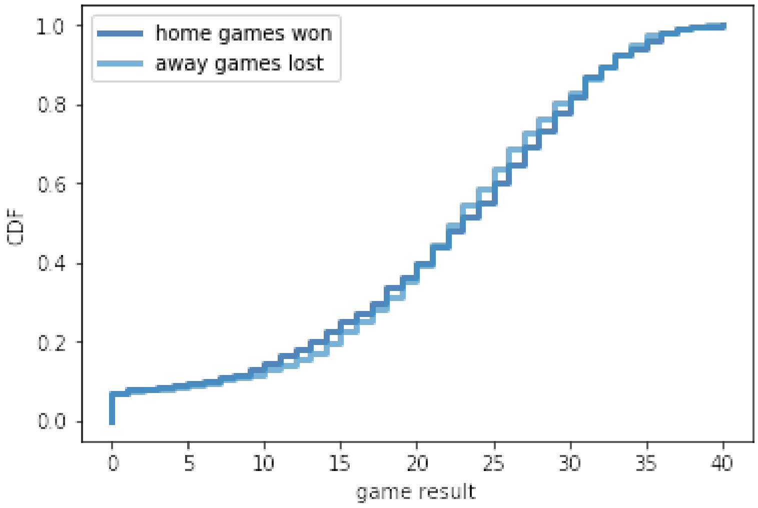
The count is the same as were observing data from 1536 different teams. The mean and standard deviation of both variables are shown above. The histogram, PMF and CDF of these variables are shown below.



The histogram above shows the values of the two variables 'homeWon', color blue, and 'awayLost', color red, plotted sided-by-side. Aside for the spike at value zero, both variables seem to have the shape of a Gaussian (normal) distribution.



According to the PMF, the value zero seems to have the higher probability for both variables. And there is a higher probability of less games lost away than more games won.



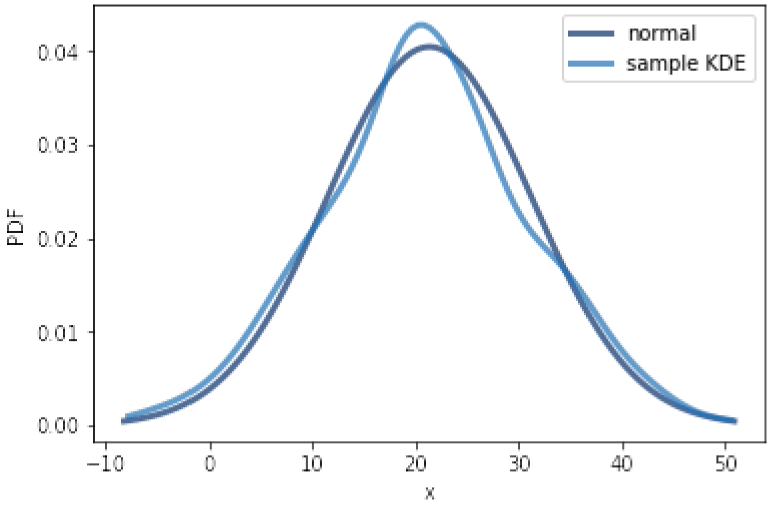
The CDF for both variables are highly similar.

SECTION 4

Normal distribution was used to model the data. Both variable ’homeWon’ and variable ’awayLost’ were modelled below.

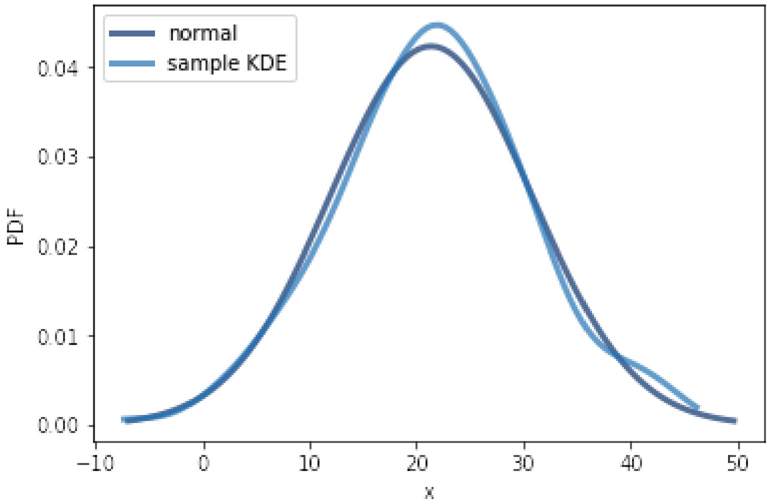
‘homeWon’:

The figure below shows the normal density function and a Kernel Density Estimate of the variable ‘homeWon’ based on a sample of 500 random game winnings. The estimate is a good match for the original distribution.



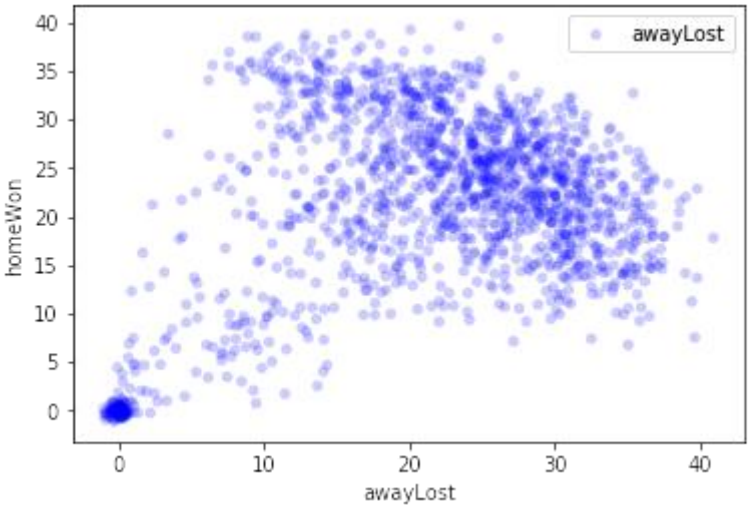
‘awayLost’:

The figure below shows the normal density function and a Kernel Density Estimate of the variable ‘awayLost’ based on a sample of 500 random game loses. The estimate is a good match for the original distribution.

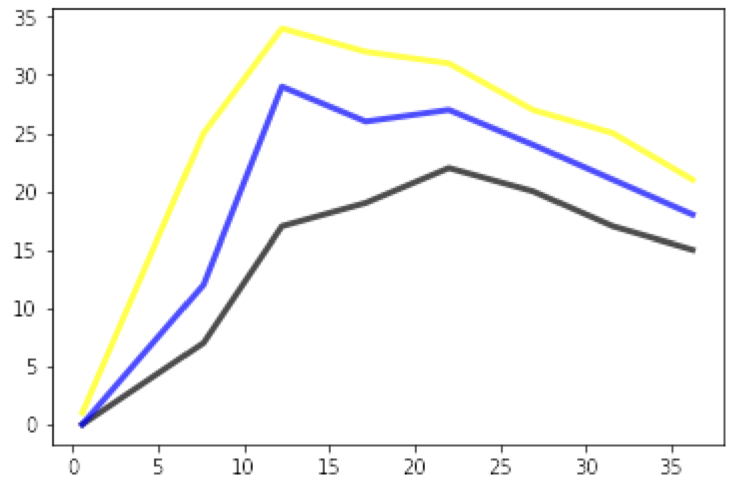


SECTION 5

The scatter plot and percentiles plot. The scatter plot shows a major saturation effect at point zero, which is as a result of some teams having neither won nor scored in the games so the plot is denser in this area. Other than that, there seems to be a relationship between the two variables that does not look linear.



The percentile plot of the values of 'awayLost' for a range of values of 'homeWon' does not appear linear.



SECTION 6

According to the hypothesis given in section 1, we are checking for the difference in the means of the two groups of values from the variables. The class DiffMeansPermute, which was given in class, along with its TestStatistic, MakeModel and RunModel methods were utilized in the hypothesis testing.

Below are the values gotten from the actual value and the maximum test statistic of the test.

```
(8.750651041666666, 1.1842447916666679)
```

SECTION 7

The p-value gotten from testing the hypothesis is 0.0. The effect of a difference in the distributions of both variables is statistically significant. We can conclude that there is a difference in the means of the 'homeWon' and 'awayLost' variables.