EXPLORATORY DATA ANALYSIS
(CS 240)
PROJECT

Nursena KARAKAŞ

213862883

Part 1 : Brainstorming

Questions:

1. What is the relationship between Won and homeWon?
2. Could it be any strong negative or positive relationship between Won and homeWon?
3. What are the characteristics of wons in datasets?

My Hypothesis:

If homewons will increase, number of wins increase.

Part 2 : Data Analysis

I used basketball_teams.csv of Basketball Data. I chose the won and homeWon columns. Won represents the number of won match and homewon means number of match that won at own home.

I used pandas and read_csv for read the csv file.

```
df = pd.read_csv('basketball_teams.csv') #read database
w = df.won #number of won
homewon = df.homeWon
```

Part 3 : Histogram , PMF, CDF

To find some statistics, I used min(), max(), mean(), var(), and std() functions and I defined these values.

For wons max value is 72, and for homewons 40. Both of them have 0 for min value.

```
(0, 72, 0, 40)
```

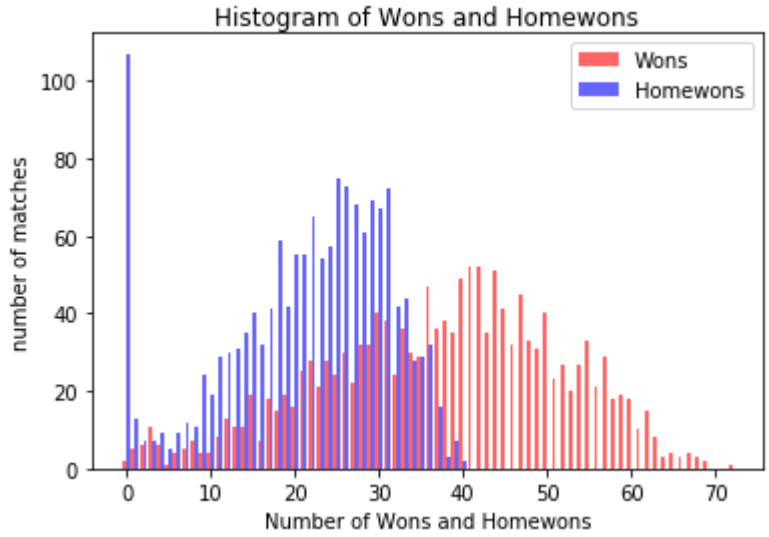
Mean of Homewons is lower than Mean of Wons.

Wons variance is 200 and it is approximately 2 times of Wons variance.

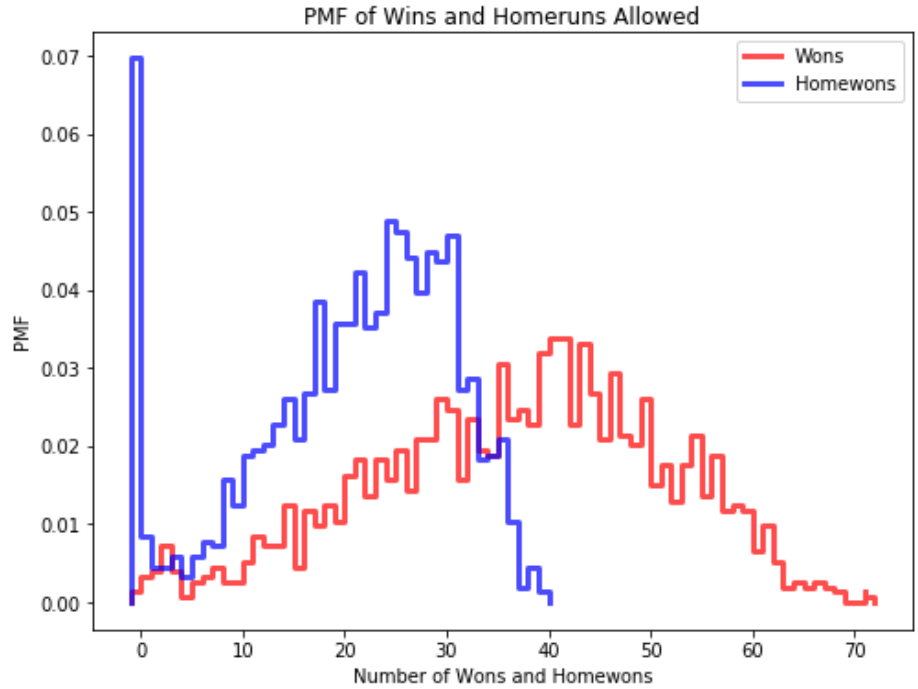
Standar deviation of Wons is bigger than standard deviation of Homewons.

In short, all homewons values are less than wons values.

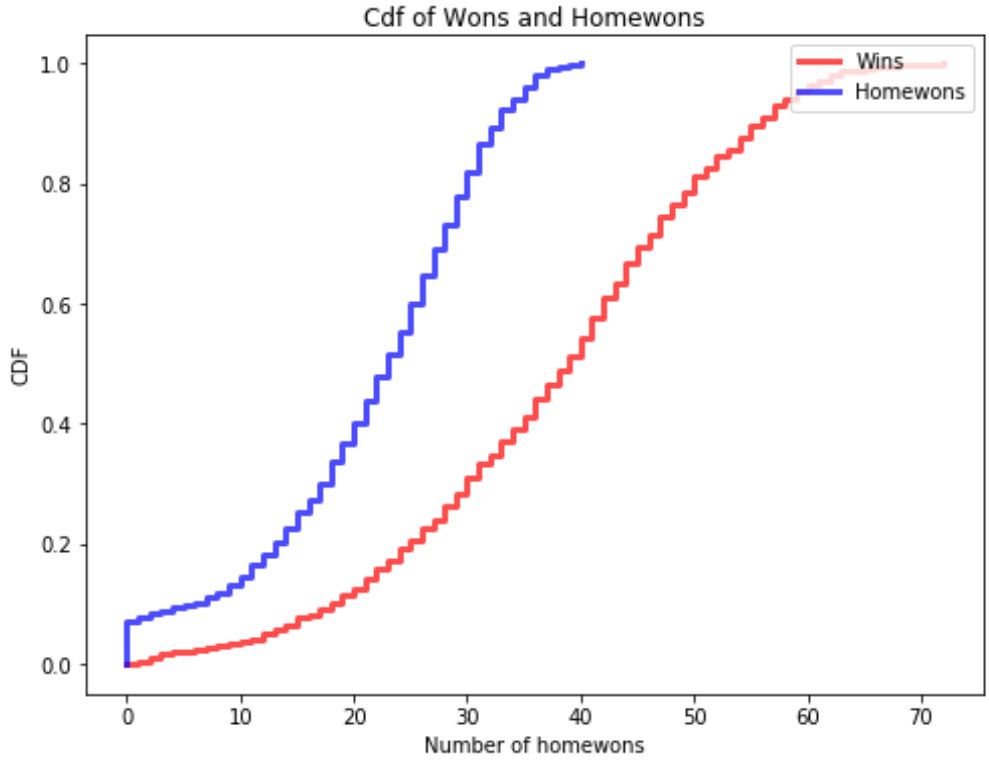
```
Mean of W: 37.552734375
Variance of W: 200.68777102
Standard Deviation of W: 14.1664311321

Mean of Homewon: 21.361328125
Variance of Homewon: 96.950138691
Standard Deviation of Homewon:9.84632615197
```



This histogram shows the number of wons and homewons around. X-axis represents the number of wons and homewons. Y-axis represents the number of matches.



In PMF, it shows probabilites of wons and homewons. As we see on the graph, probabilities are flactuated for both of them. However, for wons values are closer, and max value is about 0.035. When wons value reach the max, its probability will be minimum. Homewons probability has maximum probability at 0 and its probability is equal to 0.07, after that it decrease and approximately at 40, it has minimum probability.

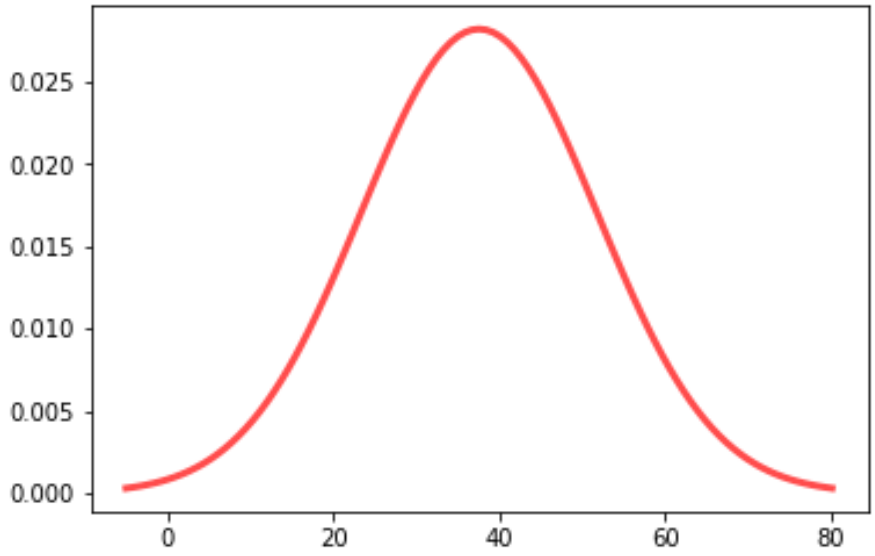


In CDF, Wons and Homewons increase. Increase of homewons value is faster than wons.

PART 4 : Modelling Distribution

At this part, I utilized normal pdf distribution.

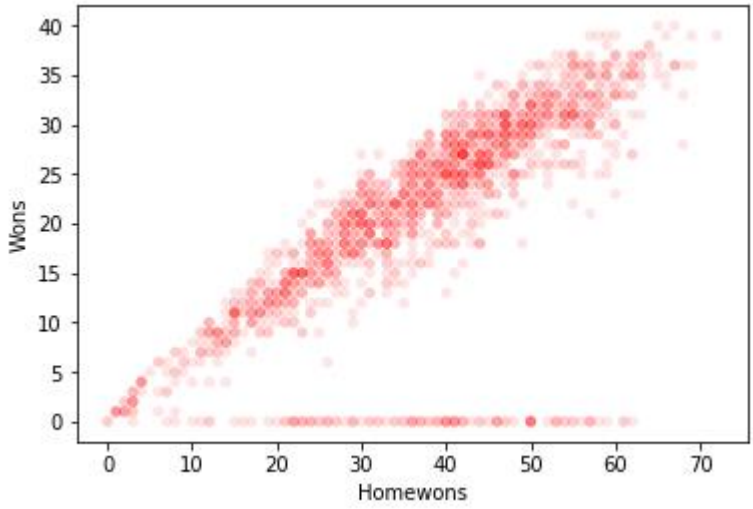
'Normal Pdf Distribution of Wons'



PART 5 : Correlation

I checked the numpy's correlation coefficient of wons and homewons. I obtained 0.73912946 for value. For this, correlation value is high.

```
[[ 1.          0.73912946]
 [ 0.73912946  1.        ]]
```



PART 6 : Hypothesis Testing

I used hypothesis testing and applied 4 steps as we see in class. Firstly, I controlled test statistic, Then I defined the null hypotesis. The next step, I compute the p-value, and I interpreted the result.

I used thinkstats2.HypothesisTest class to make hypothesis testing. I chose test statistic and compare wons and homewons.

Test Statistic : I used means of wons and homewons as test statistics.

Null Hypothesis : There is a no relationship between wons and homewons.

```
p-value: 0.0 .
```
It is below the threshold 0.05. It means statistically significant.

PART 7 : Conclusion

In conclusion, there is a relationship between wons and homewons. When I checked the correlation, similarity between them was high. In hypothesis testing, test statistic was statistically significant.