Name: Mayasa Dablan

ID: 214610755

CS 240 Project's Report

Introduction:

In this project the dataset that I have worked on is teams table, which has multiple columns on the wins of teams, losses, league champion, runs scored, triples, walks by batters, and many other columns which has a good amount of data, on my project from this file's table I have worked on the wins and the runs scored columns, because I guessed that there may be a real relation between the scores of the wins and the number of runs scored. So, my hypothesis is:

If the wins scores of a particular team are high then the runs scored on the same team should be high as well, and if the wins scores are low then the runs scored on that team should also be low.

Which means that the wins are interrelated to the runs scored of the winner team.

The null hypothesis: if a team does have a low wins scores then it should have a low runs scores as well.

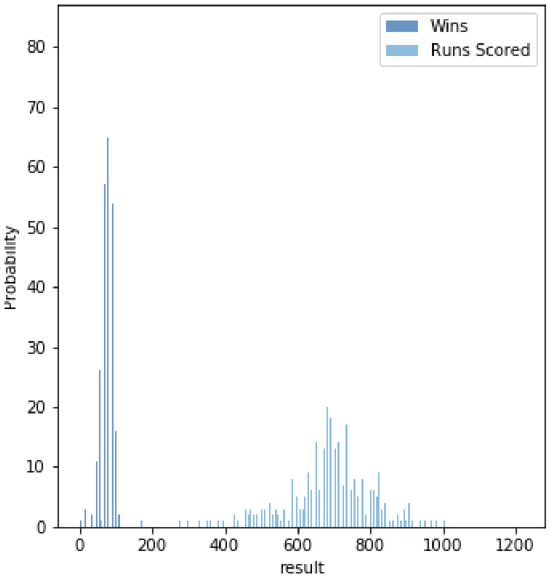
Then we will check whether there is a relatioship between the two scores, through the observation and the calculations that I did in this project.

At first I calculated the Mean and the Standard deviation for both the team wins scores and for the runs scored for the team, and these are the values that I got:

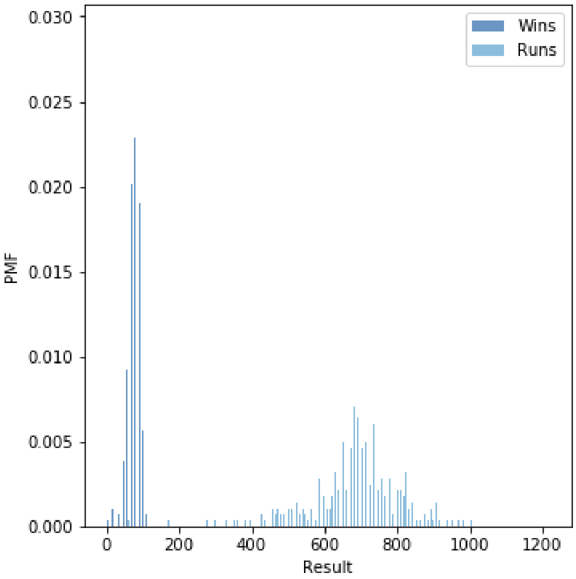
```
74.8141093474
17.5912083846
682.399294533
```

```
135.224393345
```

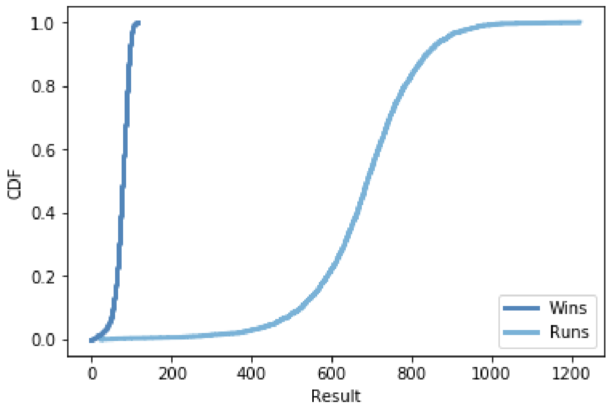
Then I made a histogram based on these wins and runs scores, which looked like:



After that, I made a PMF for these scores which looked like:

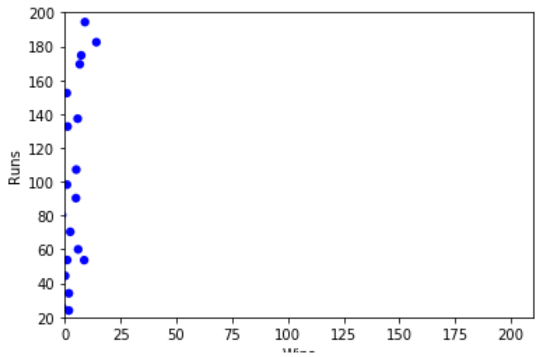


And lastly I made a CDF for them:



After that, we were supposed to model the distributions, and there are multiple ways to model distributions, the type of modeling that I have used is the exponential distribution, and I have chosen it because it is simpler.

Then, I created the jitter plots after creating the covariance function, the jitter plots are very useful because some information may have been lost in the scatter plot for that we use jittering data in order to minimize the effect on the scatter plot, and we do the jittering by adding random noise to reverse the effect of rounding off, and it looked like this:



After that, I have computed the values of the covariance after writing down the function of it, and I also have calculated the value of the correlation, here I have used both the Pearson's correlation function and the Spearman's correlation.

The Pearson's correlation was:

Corr(wins_sample,runs_sample)

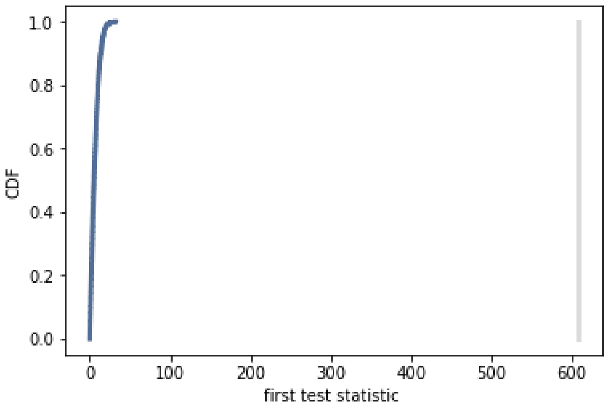
```
0.67714666815157321
```

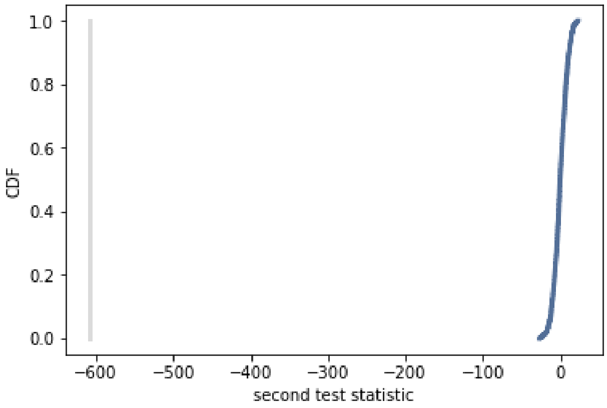
And the spearman's correlation was:

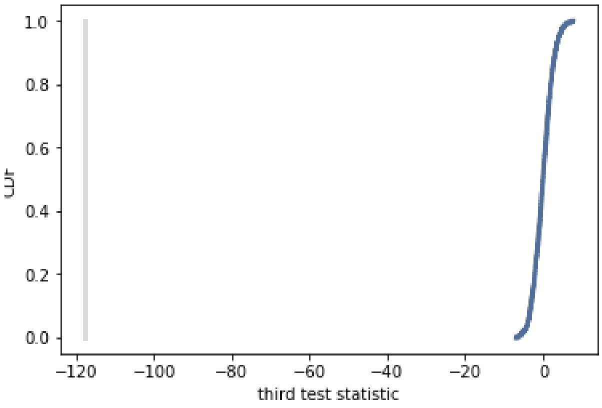
SpearmanCorr(wins_sample, runs_sample)

```
0.57517562492207863
```

Then I defined the class for the hypothesis testing and in that class, it will compute the value of the P-value and test statistics, just as explained in the book. Lastly I have plotted the cdf of that, for the first, second, and third test statistic:







Conclusion:

In conclusion, after plotting multiple plots and making various computations and calculations we can say that the main hypothesis that we have provided at the beginning is correct. And we were able to prove that through the P-value, which has shown the relationship importance between the wins of a particular team and the runs scored of that same team.