Ahsen AKPINAR 214051471

Name & Surname : Ahsen Akpınar

Cs 240 Term Project Report

In this project, my data file is basketball_teams.csv. The data is about basketball and it includes years, league Id, franchID, name of teams and many other columns. Moreover it includes home won, home lost and away won, away lost points in years between 1937-2011. There are 3 questions that i think about this projects: " if the attandance is full then the teams won number is high totally?", "ıs which team played higher minutes is successful?". The columns that i used in my project is homeWon, homeLost, awayWon, awayLost. The title that i worked is, " Is the number of home won always bigger than the number of away won? And also compare the whole datas about homeWon, homeLost, awayWon, awayLost."

Fırst of all i uploaded my file to jupyter and put the data file "basketball_teams.csv" same file with jupyter notebook. Then show the excel table. I read the csv file then used .describe( ) function. Describe( ) function calculates mean, standard deviation, min., max. %25, %50, %75 values automatically. There are the values, variables , datasets from jupyter notebook below (Figure1 and Figure2).

	year	rank	confRank	o_fgm	o_fga	o_ftm	o_fta	o_3pm	o_3pa	o_oreb	...	confWon
count	1536.000000	1536.000000	1536.000000	1536.000000	1536.000000	1536.000000	1536.000000	1536.000000	1536.000000	1536.000000	...	1536.000000
mean	1981.472656	3.346354	4.397135	2920.631510	6363.819010	1572.903646	2093.155599	194.324870	563.151042	741.756510	...	16.682292
std	20.123185	1.773812	4.555967	901.436961	2111.758438	476.075297	670.675615	212.856804	585.410835	520.660237	...	14.477500
min	1937.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...	0.000000
25%	1968.000000	2.000000	0.000000	2839.750000	6410.500000	1455.000000	1955.000000	0.000000	0.000000	0.000000	...	0.000000
50%	1984.000000	3.000000	3.000000	3169.500000	6894.500000	1639.000000	2191.500000	106.000000	352.500000	962.000000	...	19.000000
75%	1999.000000	5.000000	8.000000	3504.000000	7471.000000	1861.500000	2472.000000	363.000000	1060.000000	1138.250000	...	29.000000
max	2011.000000	9.000000	15.000000	4059.000000	9295.000000	2607.000000	3434.000000	841.000000	2283.000000	1845.000000	...	48.000000

8 rows × 51 columns

Figure 1

confLoss	divWon	divLoss	pace	won	lost	games	min	attendance
1536.000000	1536.000000	1536.000000	1536.000000	1536.000000	1536.000000	1536.000000	1322.000000	1536.000000
16.682292	11.761068	11.763672	5.529297	37.552734	37.557943	70.618490	19148.793495	25710.521484
14.473945	7.765761	7.742737	22.181582	14.166431	14.074785	23.965236	2020.620665	13469.406103
0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	2640.000000	0.000000
0.000000	7.000000	7.000000	0.000000	28.000000	28.000000	75.000000	19730.000000	32767.000000
19.000000	12.000000	12.000000	0.000000	39.000000	38.000000	82.000000	19780.000000	32767.000000
28.000000	17.000000	17.000000	0.000000	48.000000	47.000000	82.000000	19855.000000	32767.000000
52.000000	44.000000	41.000000	102.000000	72.000000	73.000000	84.000000	20460.000000	32767.000000

Figure 2

In addition, i demonstrate that number of winnig at home by using thinkstats2.Hist() function. First, i wrote the code for the homeWon column. Then wrote the same function for awayWon. The numbers are shown in ordered which is below.

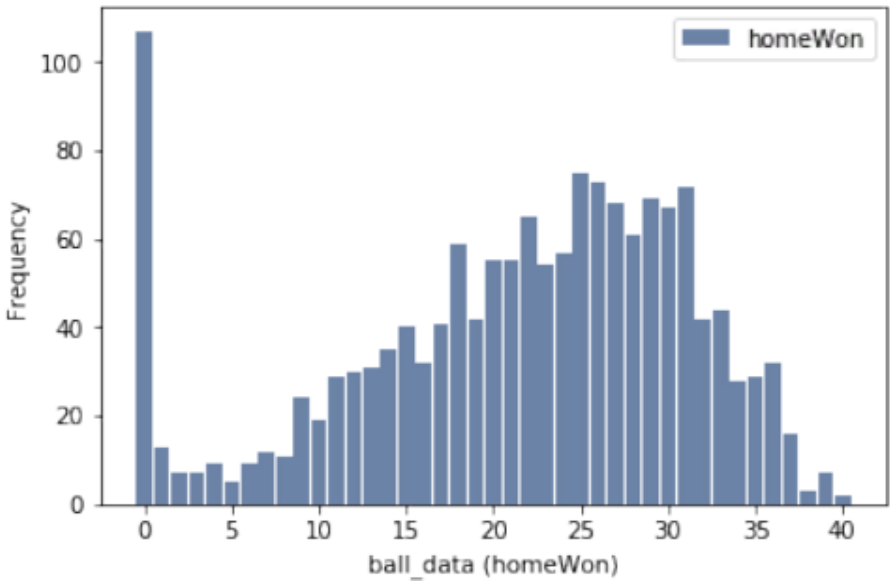
```
Hist({0: 107, 25: 75, 26: 73, 31: 72, 29: 69, 27: 68, 30: 67, 22: 65, 28: 61, 18: 59, 24: 57, 21: 55, 20: 55, 23: 54, 33: 44, 3
2: 42, 19: 42, 17: 41, 15: 40, 14: 35, 36: 32, 16: 32, 13: 31, 12: 30, 35: 29, 11: 29, 34: 28, 9: 24, 10: 19, 37: 16, 1: 13, 7:
12, 8: 11, 6: 9, 4: 9, 39: 7, 3: 7, 2: 7, 5: 5, 38: 3, 40: 2})
```

Number of homeWon

```
Hist({0: 129, 12: 94, 15: 82, 8: 78, 9: 76, 10: 76, 14: 74, 18: 70, 7: 66, 13: 63, 17: 63, 11: 62, 19: 58, 6: 56, 16: 55, 20: 5
5, 21: 49, 4: 44, 5: 37, 26: 31, 3: 29, 24: 28, 27: 27, 23: 25, 22: 24, 25: 22, 2: 20, 1: 16, 28: 9, 31: 6, 29: 6, 32: 3, 30:
2, 33: 1})
```

Number of awayWon

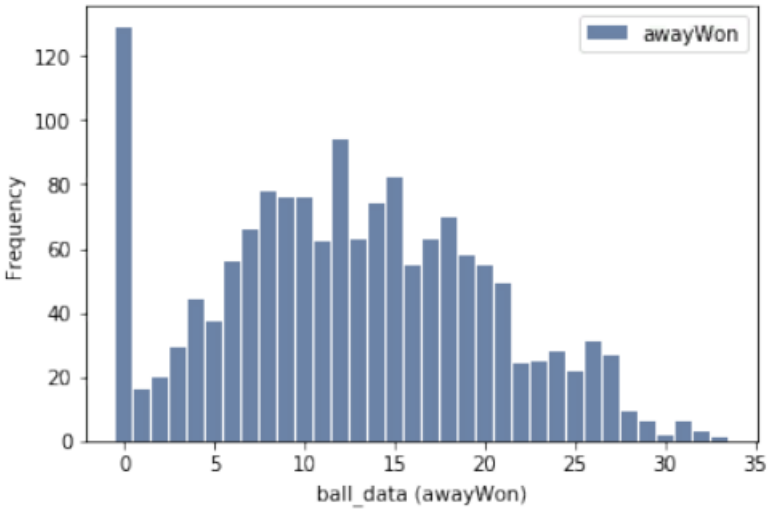
Then created histogram graph for both homeWon and awayWon since to get comparative answer . In general term, histogram is certain graphical representation of the distribution of numerical data. It also illustrates frequency of data items. Hıstogram graph created by using thinkplot.Hist() function.



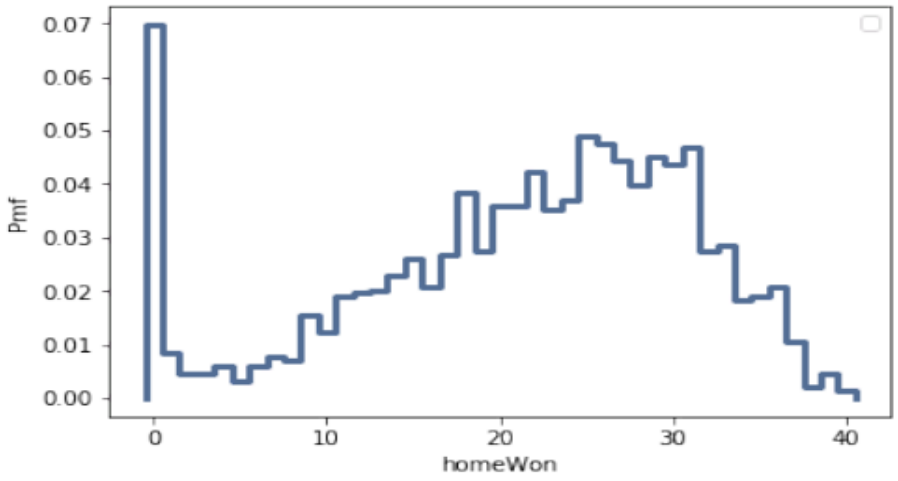
homeWon

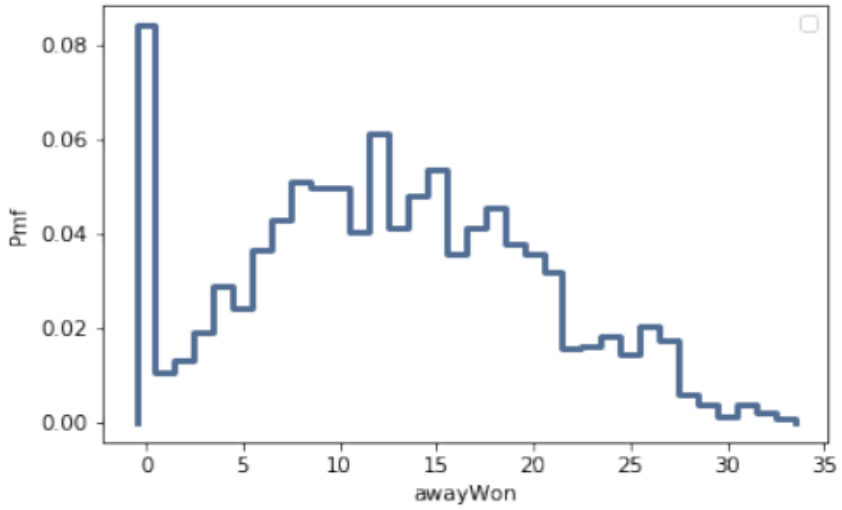
Ahsen AKPINAR 214051471

Histogram graph of awayWon is below.



Moreover, in this part i calculated Probability Mass Functions (PMF). I calculated the all values that in each sample by using PMF. Histograms and PMFs are useful while you are exploring data and trying to identify patterns and relationships. Also, Pmf is really good at analyzing and discussing the datas which you have. homeWon and awayWon PMF graphs are shown below.

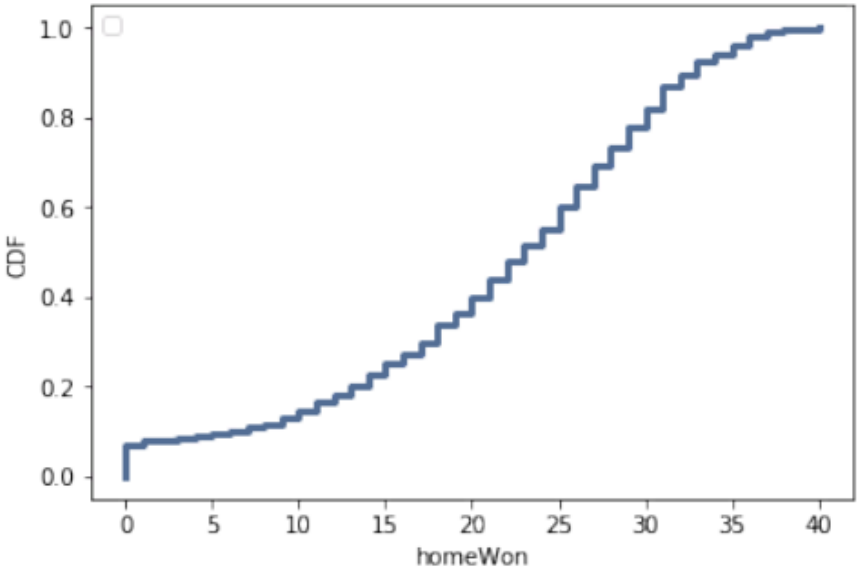




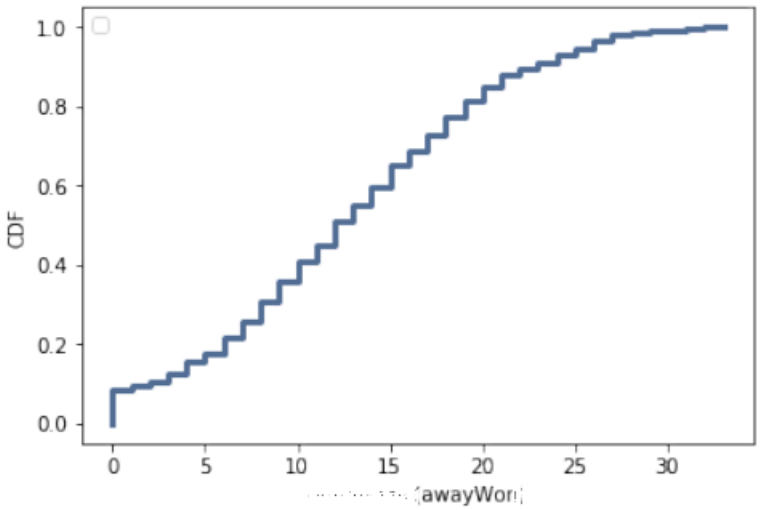
Ahsen AKPINAR 214051471

For homeWon CDF graph

CDF is Cumulative Distribution Function.



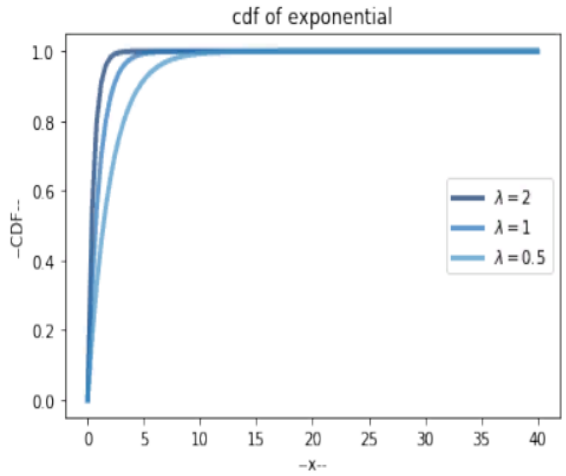
For awayWon CDF graph

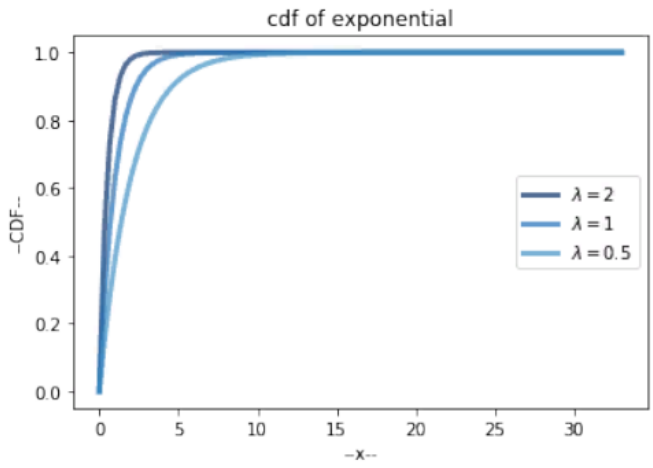


The CDF is the function that maps from a value to its percentile rank. We can apply that in any value in any datas. CDF is also good at analyzing the datas.

Ahsen AKPINAR 214051471

I choose exponential distribution as a model distribution. I illustrate that I took the distribution based on lambdas.





In addition, ı calculated correlation value. The main consequences of correlation is called by name of “correlation coefficient”. Or we can say “r”. It ranges from -1.0 to +1.0. The closer r is to +1 or -1, the more closely the two variables are related. If r is close to 0, which means there is no relation between the variables that i compare. If r is positive, it means that as one variable gets larger the other gets larger. If r is negative it means that as one gets larger, the other gets smaller (often called an "inverse" correlation). These are also about dependency or ındependency.

Taking everything into consideration, i applied hypothesis test then , calculated p-value by using some codes that we covered in class. Hypothesis tests are utilized to test the legitimacy of a claim that is made about the things. This claim that is on trial, basically, is known as the null hypothesis. A small p-value (typically ≤ 0.05) indicates strong evidence against the null hypothesis, so you reject the null hypothesis. A large p-value (> 0.05) indicates not strong evidence against the null hypothesis, so you fail to reject the null hypothesis. P-values are too close to the cutoff (0.05) are considered to be marginal. Always report the p-value so your readers can draw their own conclusions.

The p-value that i found is nearly zero. I guess real p-value is too close the 0. And python rounded it to 0.0. That’ s why the effects are significant. As you see i can easily compare the teams homewon and awaywon values by using histogram ,pmf and cdf graphs model.