CS 240 – EXPLORATORY DATA ANALYSIS

PROJECT REPORT

MERVE DİKMEN

213012387

PROJECT REPORT

SECTION 1

Questions: What is the relationship between played minutes and gained points? Is the relationship positive or negative? How these variables affect each other?

My Hypothesis: As the minutes played in matches increases, the points that the all-star players gained increases.

- When I take a brief look at the data tables, I noticed that it seems like there is a relationship between minutes and points. Therefore, I chose this question to show the relationship and to prove the hypothesis.

SECTION 2: Datasets And Data Reading

- For my hypothesis, I preferred to use *basketball_player_allstar.csv* file among other data files in the basketball dataset.
- In this data set, there are lots of columns, but I decided to work on *Minutes* and *Points* column.
- I read the csv file with importing pandas as pd and then I applied read_csv. After reading the file, I needed to take the columns that I am interested in. For this reason, I took the columns *Minutes* and *Points* among all other columns. Reading the data, taking the columns and some part of the columns are shown in the figure below:

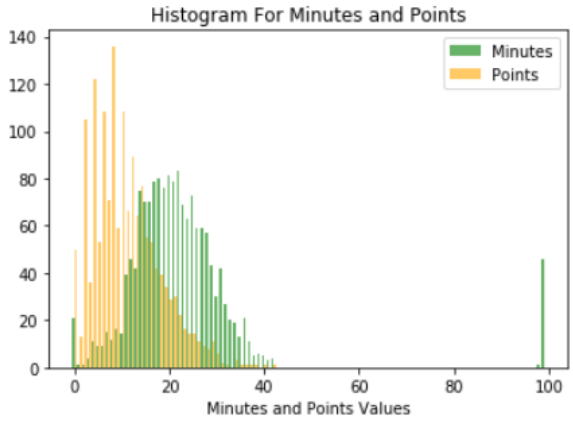
```
In [42]: data = pd.read_csv('basketball_player_allstar.csv',encoding="ISO-8859-1", delimiter=";") #reading database

         minutes = data['minutes'] #minutes played
         points = data['points']#points gathered

         print (minutes)
         print (points)
```

```
0       28
1       18
2       13
3       14
4       27
5       32
6       23
7       37
8       32
9       23
10      30
```

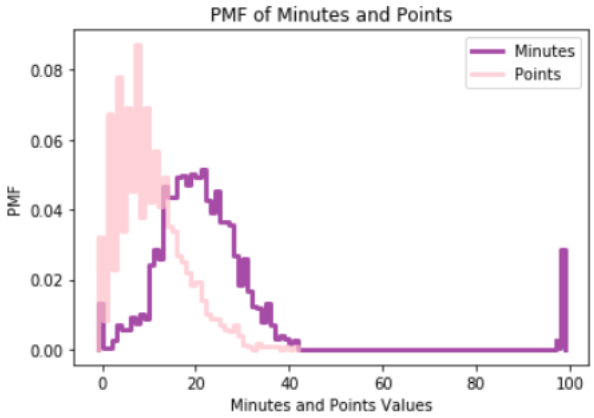
SECTION 3: Descriptive Statistics, Histograms, PMF & CDF

- The descriptive statistics that I used in the project is maximum, minimum, standard deviation, mean, mode and median for both of the columns. The values are as shown:

	MINUTES	POINTS
Maximum	99	42
Minimum	0	0
Standard Deviation	15.221598498267607	6.962658929759844
Mean	23.10316967060286	10.749679897567221
Median	21.0	10.0
Mode	0 22	0 8.0

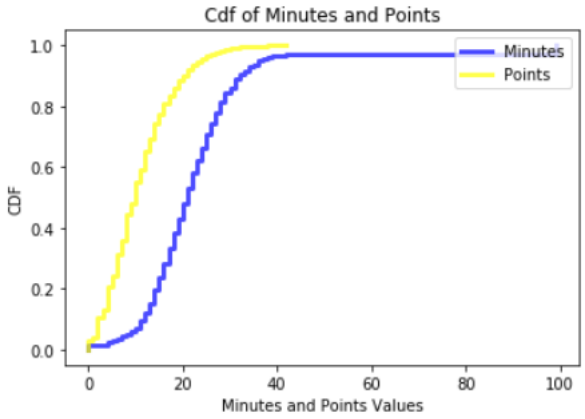


This is the histogram of minutes and points values. I plotted this to represent the distribution of the data. Histogram's x-axis shows the minutes and points values. As it is shown, the mass of the data is concentrated between 0 and almost 40.



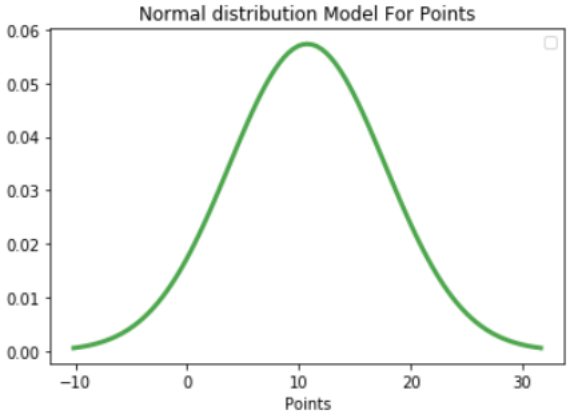
The graph is the plot of the probability mass function. While the histogram is showing the values, the PMF is showing the probability of each value for minutes and points. In the

PMF, we see that both of minutes and points showing almost the same distribution. However, the probabilities for minutes are less than points. While the probability for minutes is almost maximum 0.53 and minimum 0, the probabilities for points are almost maximum 0.92 and minimum 0.



This third graph is showing the cumulative distribution function for minutes and points. It shows the values of percentile ranks.

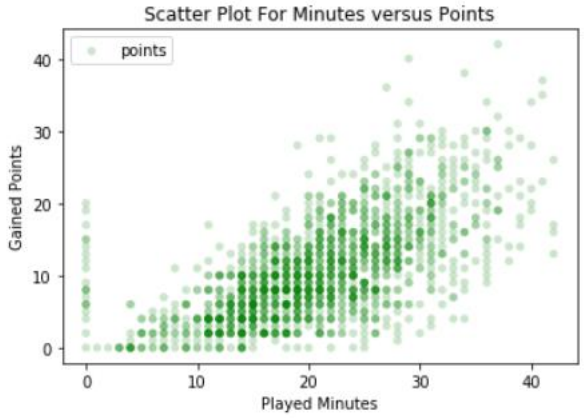
SECTION 4: Modelling The Data



- In my project, I used normal probability density function distribution for my data with mean 10.749679897567221 and standard deviation 6.962658929759844.

SECTION 5: Correlation Between Minutes And Points

- One of the best ways to show the relationship between two variables is scatter plots. I applied Scatter plot points versus minutes. As it is shown below in my scatter plot, when the minutes played increases, the points that are gained also increases. Higher minutes has the higher points.



SECTION 6: Hypothesis Testing For Minutes And Points

- I used hypothesis testing to test my hypothesis. For testing, I used 3 steps and then I found the p-value. Firstly, I found the test statistics with `TestStatistcs` function to measure the size of the effect. The test statistic in my hypothesis is the absolute value of the difference in the means of minutes and points. The next code which I used is `MakeModel` and this helps me to group the values as self.pool. In the next code, `RunModel` shuffles the pool and simulates the null hypothesis. After that, I created my null hypothesis. Finally, I concluded my hypothesis test with finding the p-value.
- In my hypothesis, the test statistic is the difference between means of values. Additionally, the null hypothesis is that no relation exists between minutes and points.
- When the test is completed, the p-value is founded as 0.0. this result indicates that the relationship between minutes and points are statistically significant.

SECTION 7: Conclusion

Taking all these computations and analysis into consideration, I verified that there is a significant relationship between the minutes that all-star players played in the match and the points that they gained. Most importantly, p-value gives this result. Because of the fact that it is 0.0, we can say that the relation between minutes and points are statistically significant. Additionally, the histograms show that the distributions of the minutes and points data are similar to each other. Therefore, it is without doubt that as the played minutes increases, the points that the players gain increases at the same time.