CS 240: Exploratory Data Analysis Project

Name: Burak Gözütok

Student ID: 214010257

PART 1 – Brainstorm

- What is the relationship between different metrics of Batting Table? Are there any strong relationship between 2 columns?
- Which leagues are have the highest hit scores?
- Is the higher stint (order of appearances within a season) values result in higher hit values?
- What is the relationship between Runs, At Bats and Hits? Are there any strong relationship between them?

My Question:

- What is the relationship between Runs, At Bats and Hits? Are there any strong relationship between them?

My Hypothesis:

- If player has more number of runs, this player will also has more At Bats and Hits value.

PART 2 – Data Organization

I will use Batting Table (Batting.csv) of Baseball Data. I will use columns of R, AB and H. **R** represents number of runs, **AB** represents number of At Bats and **H** represents the value of Hits.

I use pandas and it’s read_csv command to read this csv. I created a variable of data. Pandas module read csv and create a readable Python variable for me.

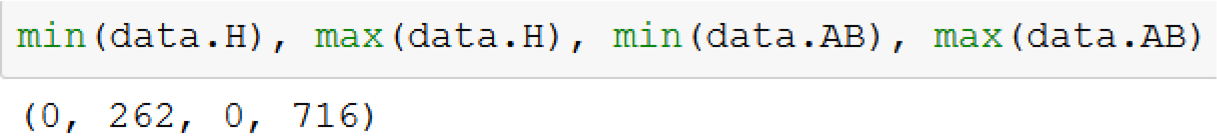
```
data = pandas.read_csv('Batting.csv', header='infer')
```

Then I got the **columns** I need from this data:

```
atbats = data.AB

runs = data.R

hits = data.H
```

PART 3 – Statistics and Graphs

When I checked the values of Hits and At Bats values by printing them I can see values are ranging and there are not much empty values in its range. So, I checked theirs range by getting mininmum and maximum value using min() and max() commands.

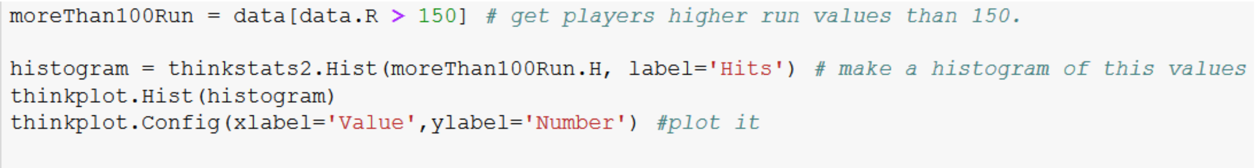


```
min(data.H), max(data.H), min(data.AB), max(data.AB)
```

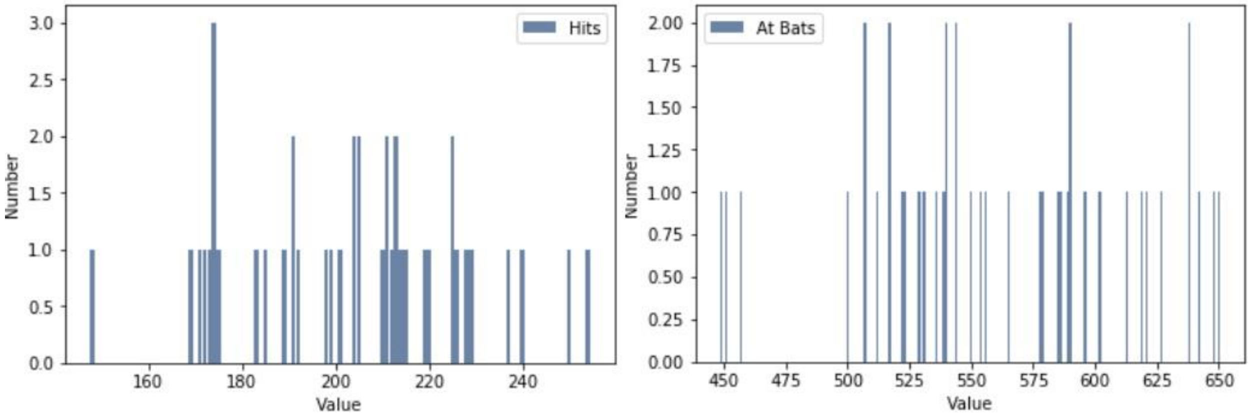
```
(0, 262, 0, 716)
```

I saw that values of Hits are ranging from 0 to 262 and values of At Bats are ranging from 0 to 716.

Then I checked If players with larger Run values will have large Hit and At Bats values as well. To check that, I got the players have more than run values of 150 and I drew their histograms for Hits and At bats.

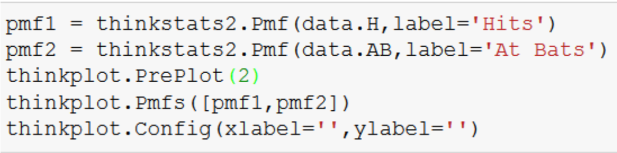


```
moreThan100Run = data[data.R > 150] # get players higher run values than 150.

histogram = thinkstats2.Hist(moreThan100Run.H, label='Hits') # make a histogram of this values
thinkplot.Hist(histogram)
thinkplot.Config(xlabel='Value',ylabel='Number') #plot it
```

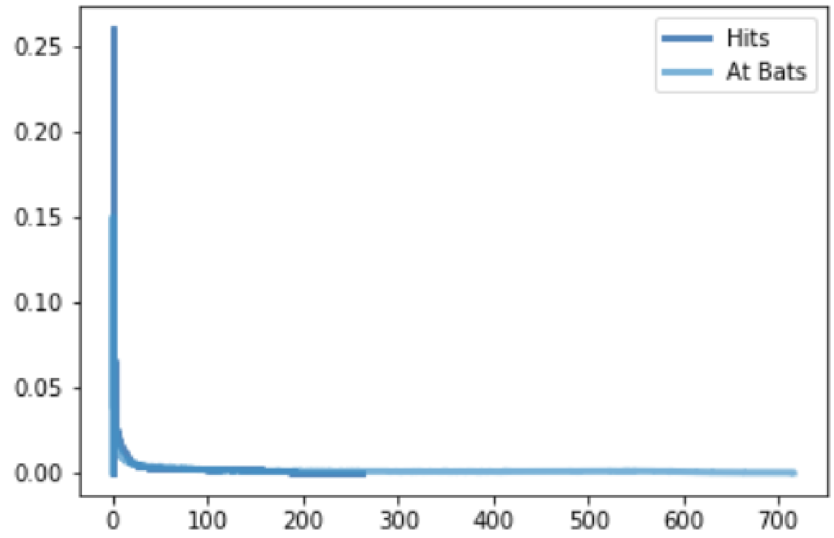


From histogram, I saw that we can't see lower values of Hits and At Bats here. So I thought that the relationship may exist but I need more tests to be sure.

For another test, I found their Probability Mass Functions and I sketch PMF graphs of both hits and at bats.

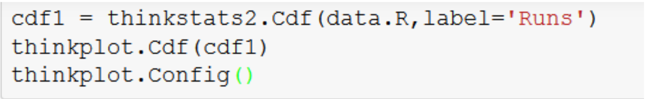


```
pmf1 = thinkstats2.Pmf(data.H,label='Hits')
pmf2 = thinkstats2.Pmf(data.AB,label='At Bats')
thinkplot.PrePlot(2)
thinkplot.Pmfs([pmf1,pmf2])
thinkplot.Config(xlabel='',ylabel='')
```

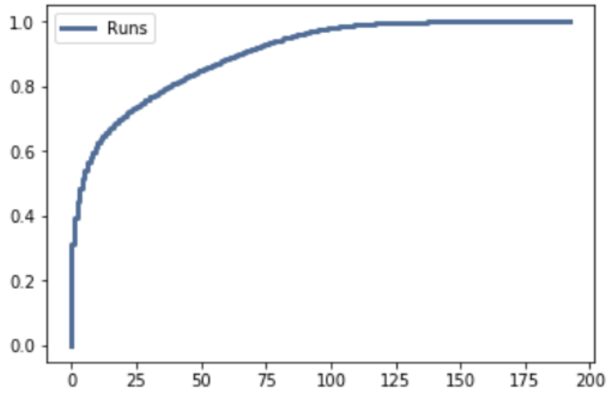


From PMF graphs I can see that Hits and At Bats value probabilities are too close to each other. They have higher probabilities in lower values and both have lower probabilities in higher values. It is another mark for my hypothesis that shows me I am on correct way.

For this part, I finally drew Cumulative Distrubiton Function of Run number of players. From this graph I see in 0 values there is leap. And Cdf value become 0.8 in very low values almost 60. Most of the players has lower run values like I see for hits and at bats in pmf graphs.

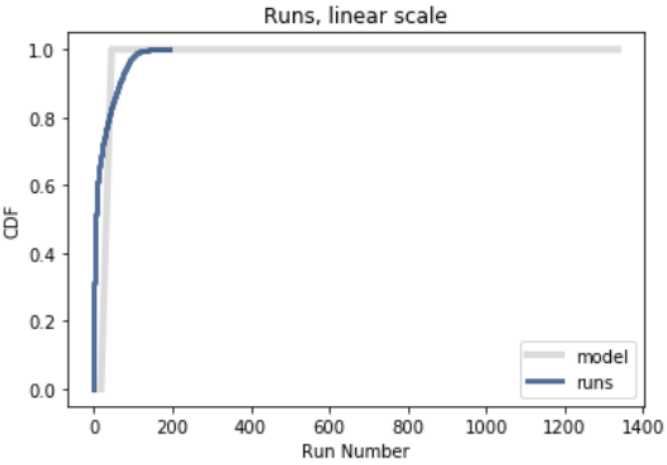


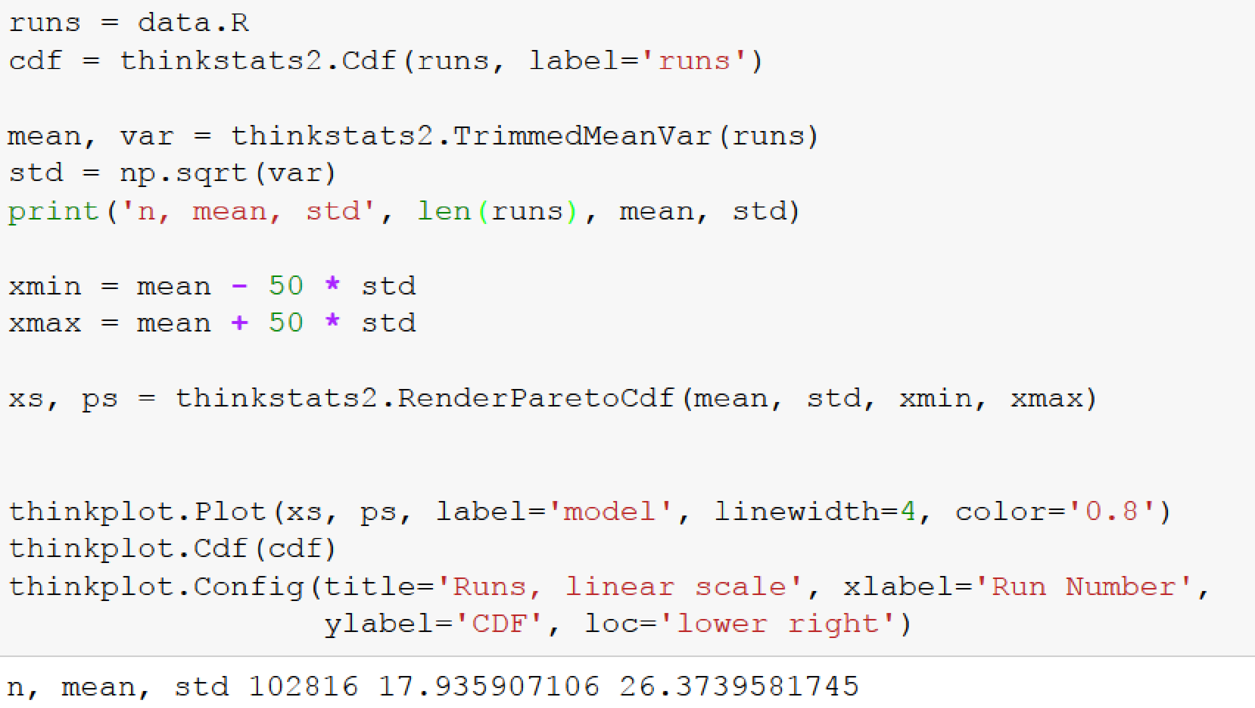
```
cdf1 = thinkstats2.Cdf(data.R,label='Runs')
thinkplot.Cdf(cdf1)
thinkplot.Config()
```



PART 4 – Modeling Distrubition

I use Pareto distribution to model my data. I used it because my Cdf graphs look like the Pareto distribution. They fit well. I used RenderParetoCdf command of thinkstats for that model. I plot it as model and I also plot my run values cdf. Then I see their graphs are looks similar. I set min and max values by substracting or adding 50 times of std to mean. I found this 50 values by trying. I see with higher values my model become more close to my data.



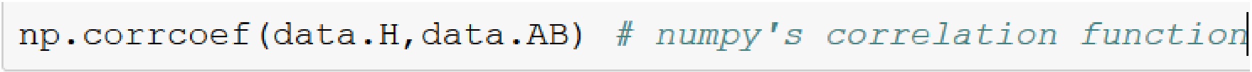


```
runs = data.R
cdf = thinkstats2.Cdf(runs, label='runs')

mean, var = thinkstats2.TrimmedMeanVar(runs)
std = np.sqrt(var)
print('n, mean, std', len(runs), mean, std)

xmin = mean - 50 * std
xmax = mean + 50 * std

xs, ps = thinkstats2.RenderParetoCdf(mean, std, xmin, xmax)


thinkplot.Plot(xs, ps, label='model', linewidth=4, color='0.8')
thinkplot.Cdf(cdf)
thinkplot.Config(title='Runs, linear scale', xlabel='Run Number',
                 ylabel='CDF', loc='lower right')
```

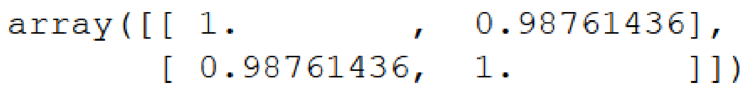
```
n, mean, std 102816 17.935907106 26.3739581745
```

PART 5 – Building a Relationship

I chose At Bats (AB) and Hits (H) columns for this part. First of all, I checked their relation by using numpy's corrcoef function. This function gives the correlation coefficient of 2 values.



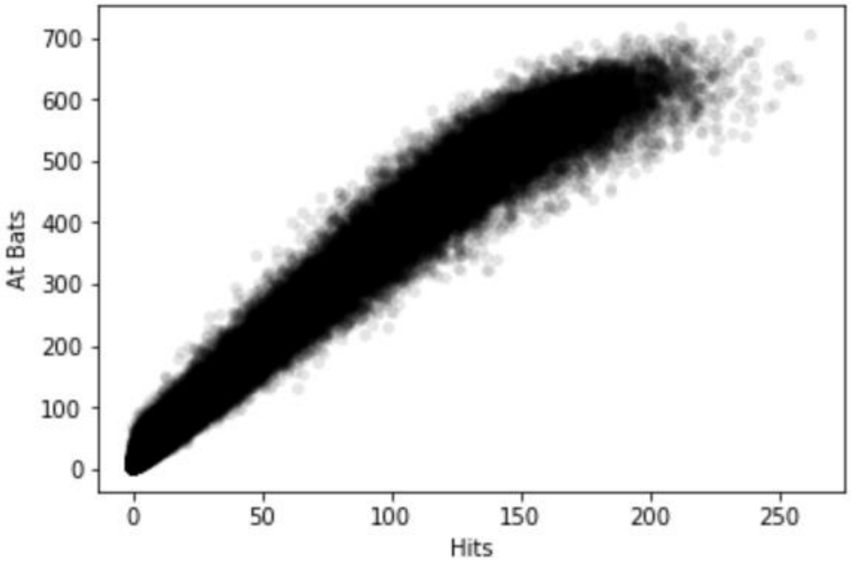
```
np.corrcoef(data.H,data.AB) # numpy's correlation function
```



```
array([[ 1.        ,  0.98761436],
       [ 0.98761436,  1.        ]])
```

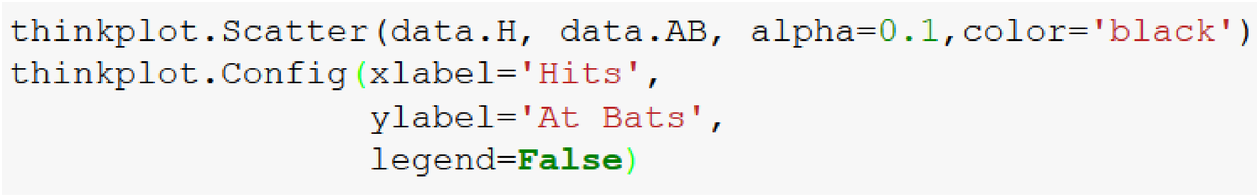
I see the correlation value of this 2 is 0.98761436 which is almost 1. It means this 2 value are too much related.

To visualize this correlation, I use scatter plots. I drew Scatter plot of this 2 columns by using thinkplot. First time I try to plot, I encounter with a graph that looks ugly and it is difficult to differentiate and understand. To make it more understandable and clear, I decreased alpha value to 0.1. In this way the spread of values became more clear.



From this scatter plot I see that high correlation value. Hits and at bats values are increasing together they have linear like shape. I can say there is thick line looks like y=x line in plot. They are too much related.

To plot this, I wrote the code below:



```
thinkplot.Scatter(data.H, data.AB, alpha=0.1,color='black')
thinkplot.Config(xlabel='Hits',
                 ylabel='At Bats',
                 legend=False)
```

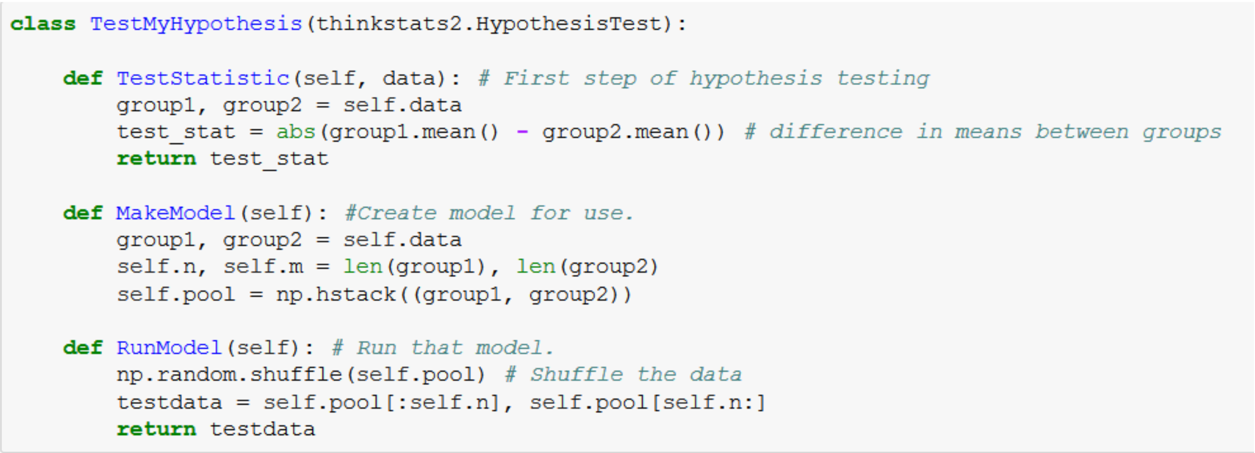
PART 6 – Hypothesis Testing

I tested my hypothesis which is “If player has more number of runs, this player will also has more At Bats and Hits value.”

There are 4 steps of hypothesis testing and I apply them all:

1- Choosing test statistic
2- Defining null hypothesis
3- Computing p-value
4- Interpreting the result

To test the hypothesis I used the thinkstats2 class of HypothesisTest(). There are some methods that class requires I defined them. In TestStatistic() method first step of hypothesis testing is doing. I chose test statistic that difference between columns of H, AB and R. I compare them by 2. In MakeModel() function my data variables are creating and in RunModel() it shuffles the data to see occurrence probability by chance. It is second step of hypothesis testing because it simulates null hypothesis.

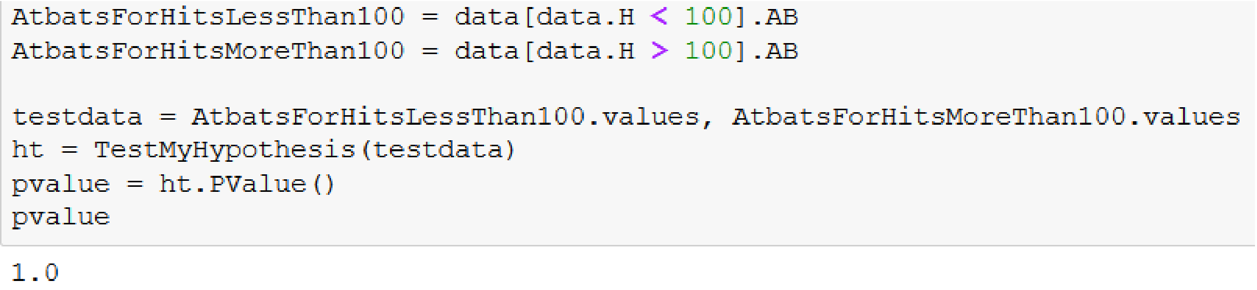


```
class TestMyHypothesis(thinkstats2.HypothesisTest):

    def TestStatistic(self, data): # First step of hypothesis testing
        group1, group2 = self.data
        test_stat = abs(group1.mean() - group2.mean()) # difference in means between groups
        return test_stat

    def MakeModel(self): #Create model for use.
        group1, group2 = self.data
        self.n, self.m = len(group1), len(group2)
        self.pool = np.hstack((group1, group2))

    def RunModel(self): # Run that model.
        np.random.shuffle(self.pool) # Shuffle the data
        testdata = self.pool[:self.n], self.pool[self.n:]
        return testdata
```

After these, I go on third step of hypothesis testing which is computing the p value. Thinkstats’ HypothesisTest class has a function for it which is PValue(). I use it and found p values.

Because of my hypothesis about the finding the relationships of H, AB and R columns, I split data to 2 parts. In first, I splitted by Hits (H) values with greater than 100 and smaller than 100. Then I run my test using their At bats value. By this way, I will see the relationship of hits and at bats. When I run it I see the p value of 1.0. Which is the maximum of it can be. So, my p value is very high. In fourth step of

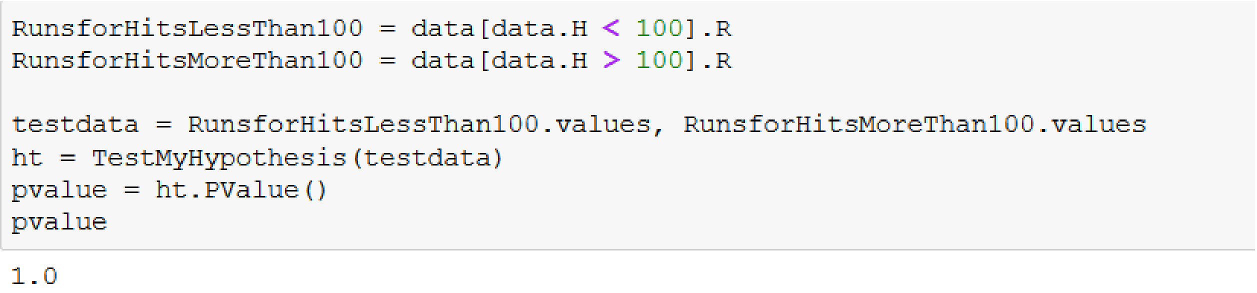
Hypothesis Testing I evaluated it, I can say because of high p value it is not statistically significant. So it is more likely to occurred by chance.



```
AtbatsForHitsLessThan100 = data[data.H < 100].AB
AtbatsForHitsMoreThan100 = data[data.H > 100].AB

testdata = AtbatsForHitsLessThan100.values, AtbatsForHitsMoreThan100.values
ht = TestMyHypothesis(testdata)
pvalue = ht.PValue()
pvalue
```

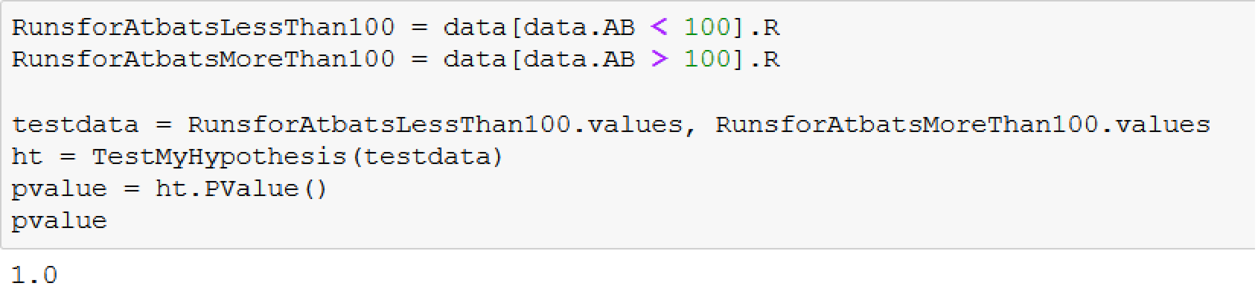
```
1.0
```

I make this tests between Hits and Runs as well and I see again 1.0 value which means it is more likely occurring by chance. Finally I test between At bats values and Run values but still I got 1.0 p value. All of them occurs by chance more likely.



```
RunsforHitsLessThan100 = data[data.H < 100].R
RunsforHitsMoreThan100 = data[data.H > 100].R

testdata = RunsforHitsLessThan100.values, RunsforHitsMoreThan100.values
ht = TestMyHypothesis(testdata)
pvalue = ht.PValue()
pvalue
```

```
1.0
```



```
RunsforAtbatsLessThan100 = data[data.AB < 100].R
RunsforAtbatsMoreThan100 = data[data.AB > 100].R

testdata = RunsforAtbatsLessThan100.values, RunsforAtbatsMoreThan100.values
ht = TestMyHypothesis(testdata)
pvalue = ht.PValue()
pvalue
```

```
1.0
```

PART 7 – Conclusion

In conclusion, from the baseball data, I chose H, AB and R columns which are Hits, At bats and Runs and I make research about their relationships. I drew their histograms, cdfs and pmfs. I found their modeling distribution, I drew cdf graph with Pareto model. All of these, let me understand the data more and they let me see how data changed how they are related etc. Then I tried to find their relation and I found correlation values by using numpy. This correlation values are too high which shows me they have huge relationship. I drew scatter plot as well, I see straight line on it which also shows the relationship. All of these are not enough and I made the final by Hypothesis Testing. There were 4 steps of them and I apply these. I checked p values. I see p values of 1.0 in all time which shows me they are occurring by chance more likely. The relationship I found looks like occurring by chance.