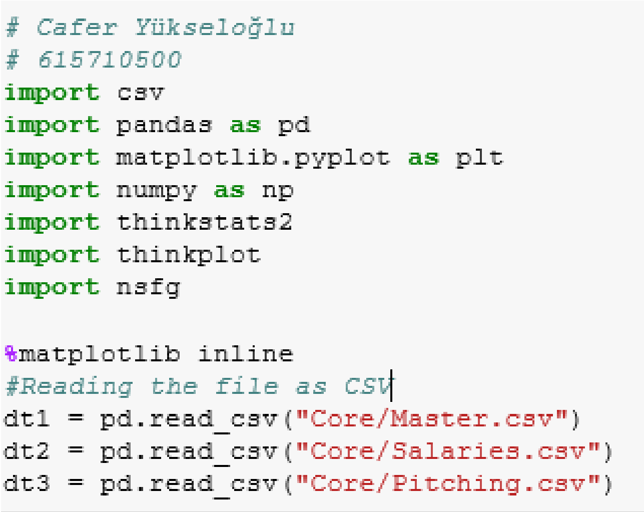
CS 240 Exploratory Data Analysis

Cafer Yükseloğlu
615710500

Question = Are there any colleration between sallary and players played game, wind or lose?

For that I used “Salary.csv” and “Pitching.csv”

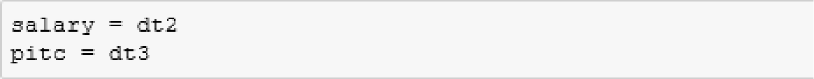
First of I import files Pandas and Numpy also Matplotlib to read data, decribing it and ploting it.



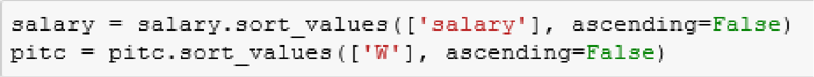
```
# Cafer Yükseloğlu
# 615710500
import csv
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np
import thinkstats2
import thinkplot
import nsfg

%matplotlib inline
#Reading the file as CSV
dt1 = pd.read_csv("Core/Master.csv")
dt2 = pd.read_csv("Core/Salaries.csv")
dt3 = pd.read_csv("Core/Pitching.csv")
```

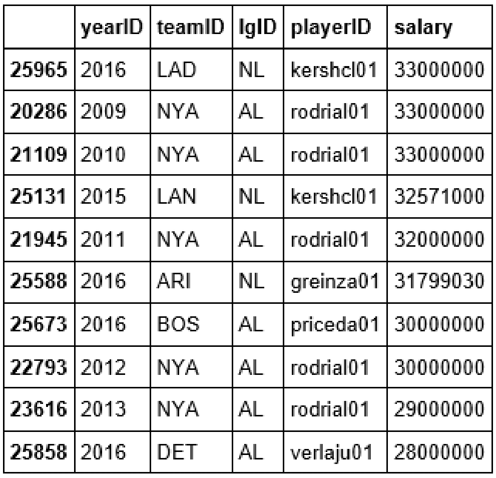
Then I append orginal values to new veriables to change them and sort them for their special columns



```
salary = dt2
pitc = dt3
```

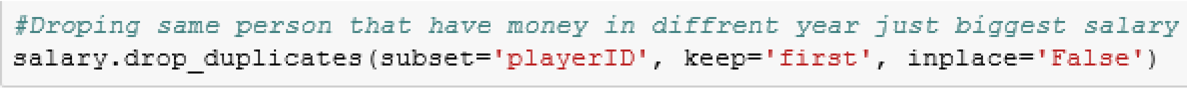


```
salary = salary.sort_values(['salary'], ascending=False)
pitc = pitc.sort_values(['W'], ascending=False)
```



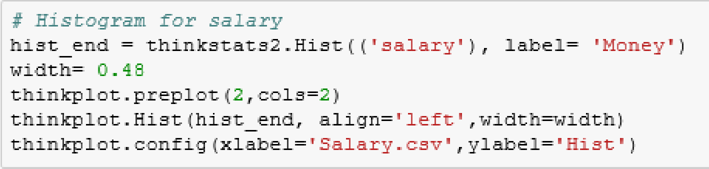
	yearID	teamID	lgID	playerID	salary
25965	2016	LAD	NL	kershcl01	33000000
20286	2009	NYA	AL	rodrial01	33000000
21109	2010	NYA	AL	rodrial01	33000000
25131	2015	LAN	NL	kershcl01	32571000
21945	2011	NYA	AL	rodrial01	32000000
25588	2016	ARI	NL	greinza01	31799030
25673	2016	BOS	AL	priceda01	30000000
22793	2012	NYA	AL	rodrial01	30000000
23616	2013	NYA	AL	rodrial01	29000000
25858	2016	DET	AL	verlaju01	28000000

There was some values that player earning more then 1 that was every year their salary was changing i drop lower prices to see max values they earned.

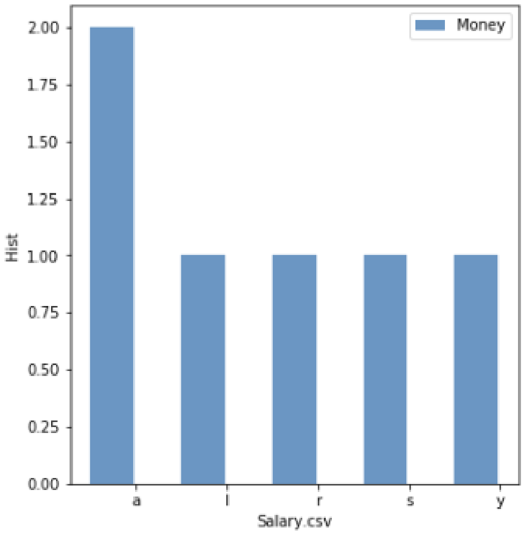


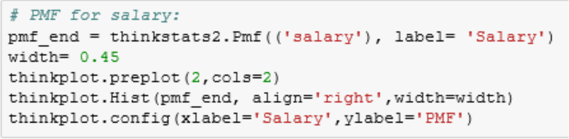
```
#Droping same person that have money in diffrent year just biggest salary
salary.drop_duplicates(subset='playerID', keep='first', inplace='False')
```

Then I showed 3 driffent thinkstart method to graph salary

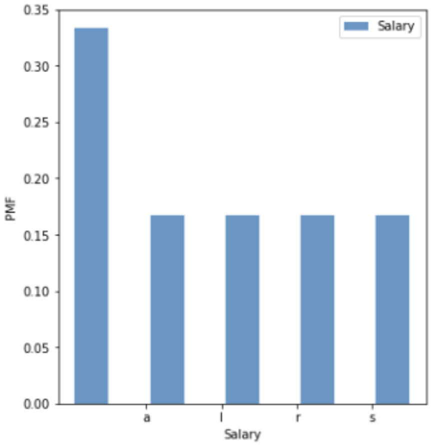


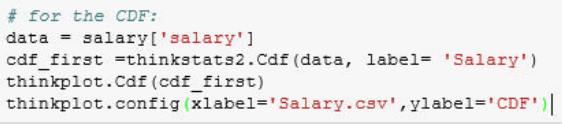
```
# Histogram for salary
hist_end = thinkstats2.Hist(('salary'), label= 'Money')
width= 0.48
thinkplot.preplot(2,cols=2)
thinkplot.Hist(hist_end, align='left',width=width)
thinkplot.config(xlabel='Salary.csv',ylabel='Hist')
```



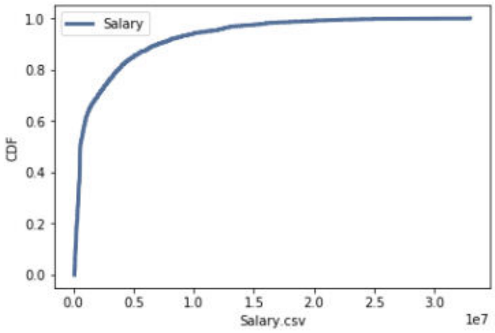


```
# PMF for salary:
pmf_end = thinkstats2.Pmf(('salary'), label= 'Salary')
width= 0.45
thinkplot.preplot(2,cols=2)
thinkplot.Hist(pmf_end, align='right',width=width)
thinkplot.config(xlabel='Salary',ylabel='PMF')
```

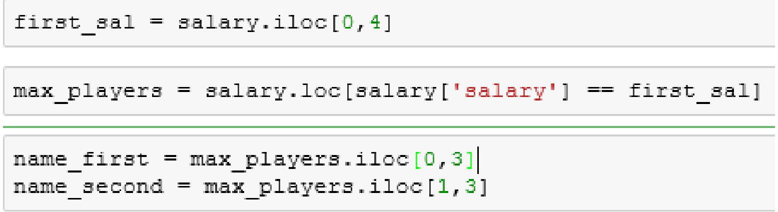




```
# for the CDF:
data = salary['salary']
cdf_first =thinkstats2.Cdf(data, label= 'Salary')
thinkplot.Cdf(cdf_first)
thinkplot.config(xlabel='Salary.csv',ylabel='CDF')
```



Therefore i got first 2 “playerID” thats earned more than others

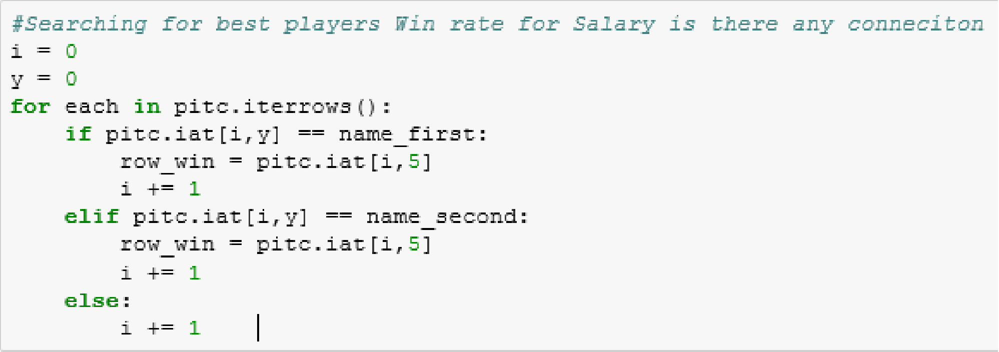


```
first_sal = salary.iloc[0,4]
```

```
max_players = salary.loc[salary['salary'] == first_sal]
```

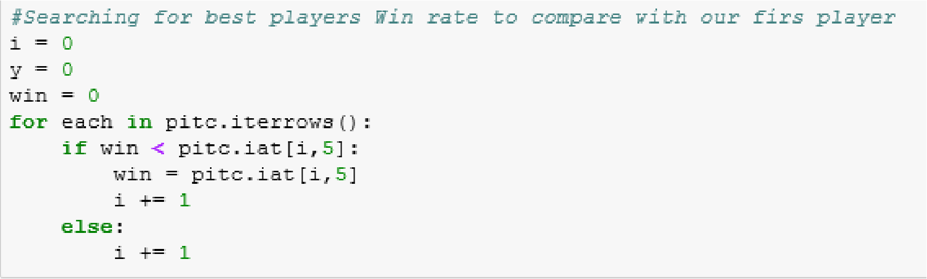
```
name_first = max_players.iloc[0,3]
name_second = max_players.iloc[1,3]
```

After then that, i add “pitching.csv” as “pich” and compared “playerID” inside of it if there any person called as our name_first to show his win and game values

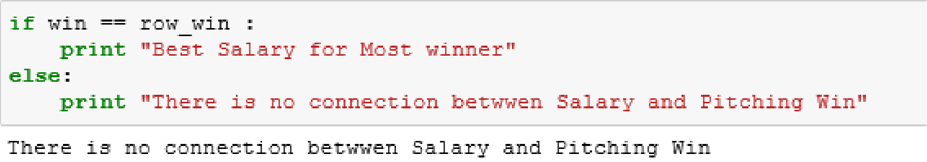


```
#Searching for best players Win rate for Salary is there any conneciton
i = 0
y = 0
for each in pitc.iterrows():
    if pitc.iat[i,y] == name_first:
        row_win = pitc.iat[i,5]
        i += 1
    elif pitc.iat[i,y] == name_second:
        row_win = pitc.iat[i,5]
        i += 1
    else:
        i += 1
```

Then took best player as game and win with that I compared 2 of them

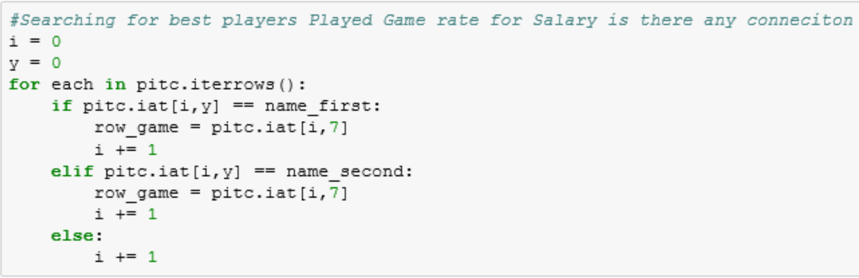


```
#Searching for best players Win rate to compare with our firs player
i = 0
y = 0
win = 0
for each in pitc.iterrows():
    if win < pitc.iat[i,5]:
        win = pitc.iat[i,5]
        i += 1
    else:
        i += 1
```

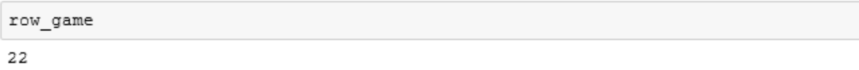


```
if win == row_win :
    print "Best Salary for Most winner"
else:
    print "There is no connection betwwen Salary and Pitching Win"
```

```
There is no connection betwwen Salary and Pitching Win
```

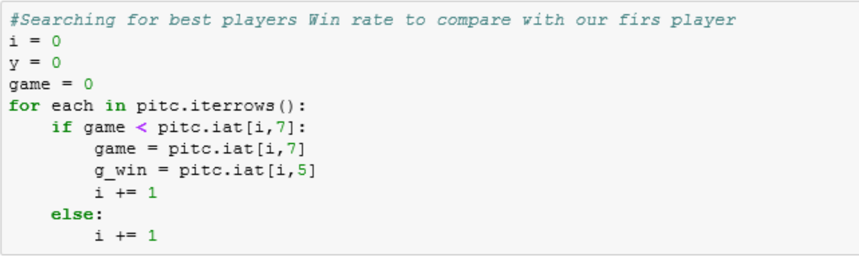


```
#Searching for best players Played Game rate for Salary is there any conneciton
i = 0
y = 0
for each in pitc.iterrows():
    if pitc.iat[i,y] == name_first:
        row_game = pitc.iat[i,7]
        i += 1
    elif pitc.iat[i,y] == name_second:
        row_game = pitc.iat[i,7]
        i += 1
    else:
        i += 1
```

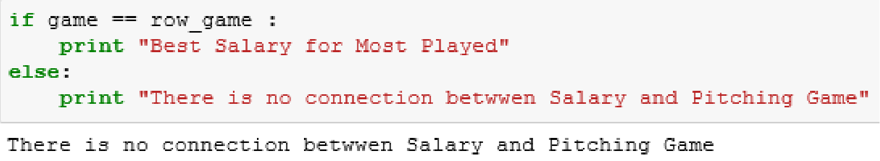


```
row_game
```

```
22
```



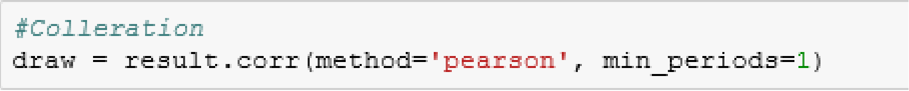
```
#Searching for best players Win rate to compare with our firs player
i = 0
y = 0
game = 0
for each in pitc.iterrows():
    if game < pitc.iat[i,7]:
        game = pitc.iat[i,7]
        g_win = pitc.iat[i,5]
        i += 1
    else:
        i += 1
```



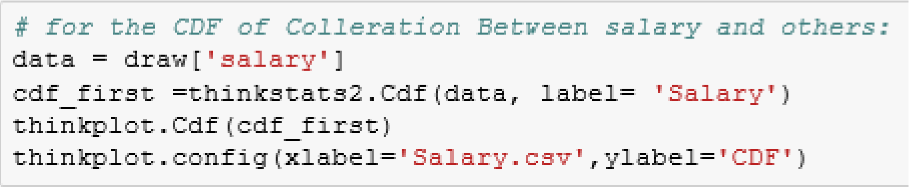
```
if game == row_game :
    print "Best Salary for Most Played"
else:
    print "There is no connection betwwen Salary and Pitching Game"
```

```
There is no connection betwwen Salary and Pitching Game
```

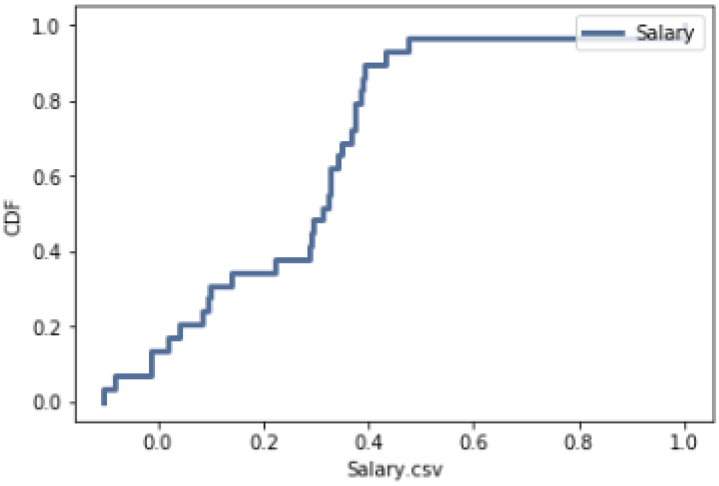
Finaly I merged the “Salary” and “Pitching” table and find coloration for salary



```
#Colleration
draw = result.corr(method='pearson', min_periods=1)
```



```
# for the CDF of Colleration Between salary and others:
data = draw['salary']
cdf_first =thinkstats2.Cdf(data, label= 'Salary')
thinkplot.Cdf(cdf_first)
thinkplot.config(xlabel='Salary.csv',ylabel='CDF')
```



Conclusion : In this exprimentel data I tryed to merge 3 diffrent dataframe to see there is any colleration between games and salary on a player but I coldn’t find much collaration between salary and games but there is some coloration because Good players also get enouhg money for their plays.