**MUHAMMED FURKAN GÜLSEREN**

**214050466**

**CS 240 – PROJECT**

**Introduction:**

In this Project,i am going to take datas for abseball and i am going to analyze salaries for every leagues. This date starts from 1985 to 2016. I will use pandas, Numpy, Thinkplot and lastly Thinkstats2 for analyzing datas.

**Part1:**

In this part, i am going to generate some questions about the baseball datas.

1-)Is there any relationship between leagues for salaries?

2-)Is there any relationship between hits and wins?

3-)Is there any relationship between hits and loses?

My question is 1. I am going to try to analyze this question.

**Part2:**

from \_\_future\_\_ import print\_function, division

%matplotlib inline

import numpy as np

import pandas as pd

import thinkstats2

import thinkplot

// I imported this module because i used it in my project.

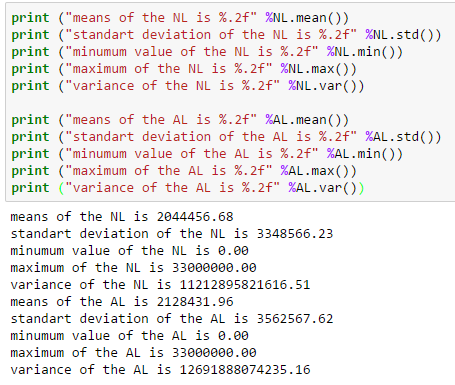
df = pd.read\_csv(‘Salaries.csv’) // importing the Salaries.csv.

NL = df.salary[df.lgID] == “NL”].dropna()

AL = df.salary[df.lgID] == “AL”].dropna()

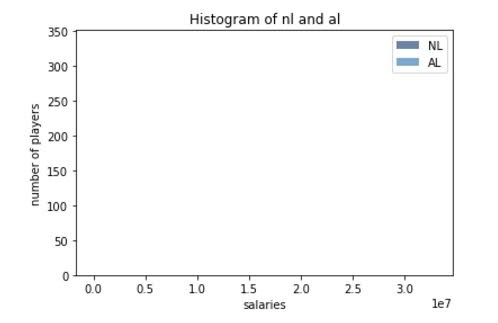
// after importing part I took the columns which are NL and AL. With the help of dropna, i cleaned the empty cells.

**Part3:**



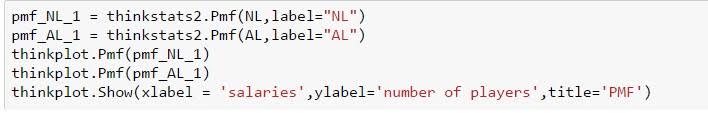
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

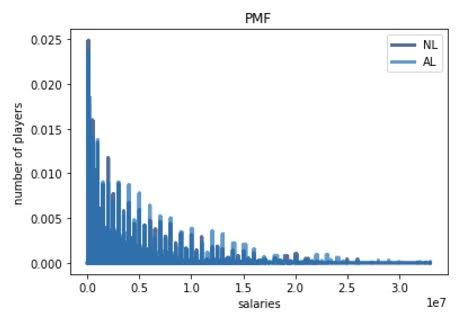


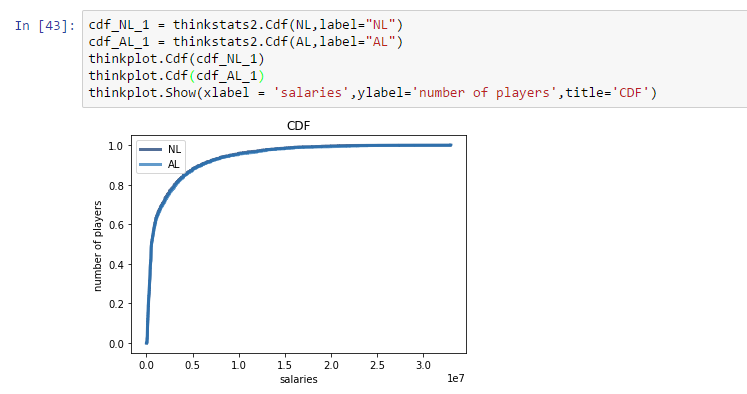


In histogram part, I had defined two different variable for nl(national league) and al(American league) then I used the ‘Hist’ to take histogram of them. Lastly, I showed them in the graphic. As we can see, the income of the both American Leagues and National Leagues are similar in terms of salaries.

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

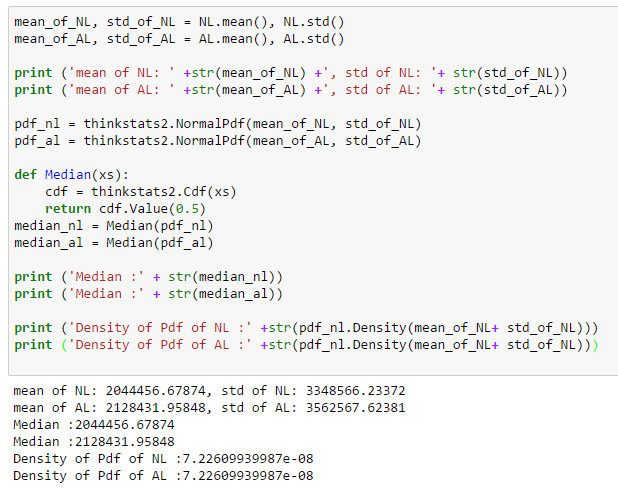




PMF basically means that probability mass function and in this part, there are player from NL and Al and it shows salaries for NL and AL. PMfs of both AL and NL shows the similarity but we can see that mid-income of the players in AL is higher than NL.

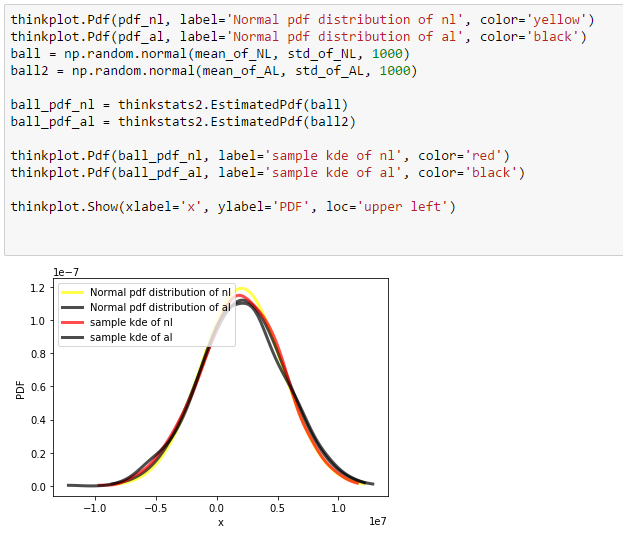
This is a cumulative distribution part which is called CDF. There is a (x,y) and x symbolizes number of players; y symbolizes salaries. Graphic show us percentile of the salaries of both NL and AL. Here as we can see that both NL and AL are almost the same.

**Part4:**

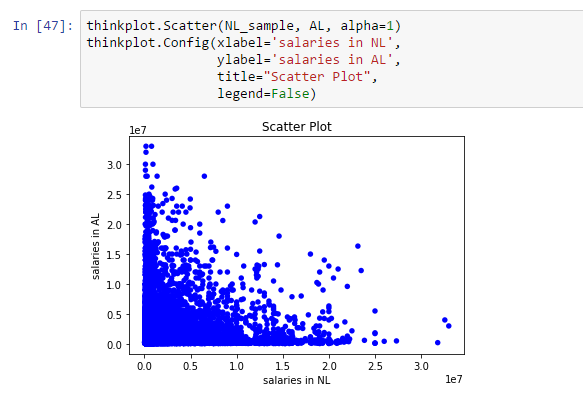


Here by estimating the mean of both leagues, I want to see the average income of both NL and AL. Also, in order to analyze the normal distribution I also found the standard deviation and the median of the leagues.

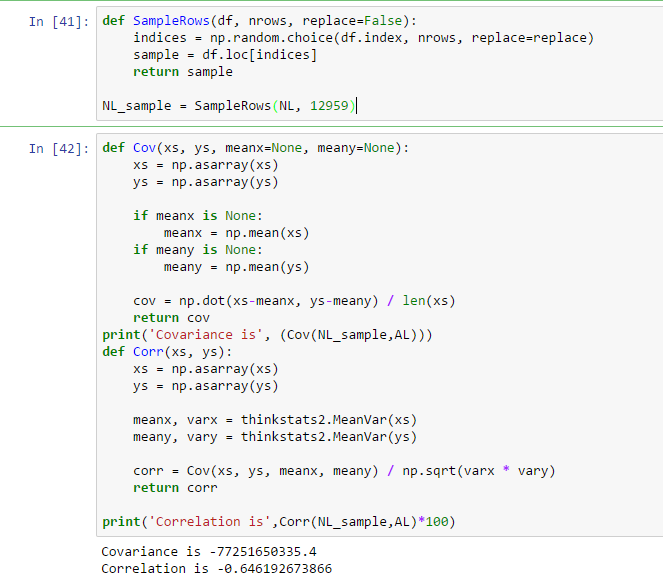
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%



**PART 5 :**



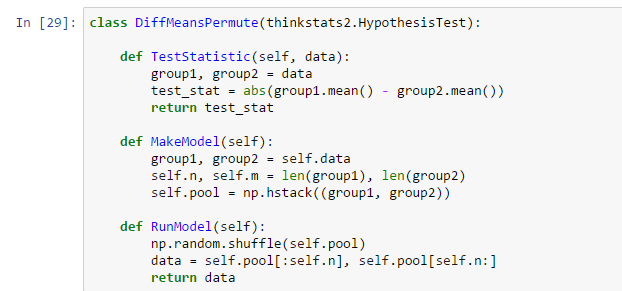
There is a density on the left botton corner. As the salary in NL increase, the salary is in AL decrease.



I added a function because there is a error which is aligned problem for this reason with the help of def SampleRow i chose the Nl that is 12959 which means they were equal between NL(national league) and Al(american league).

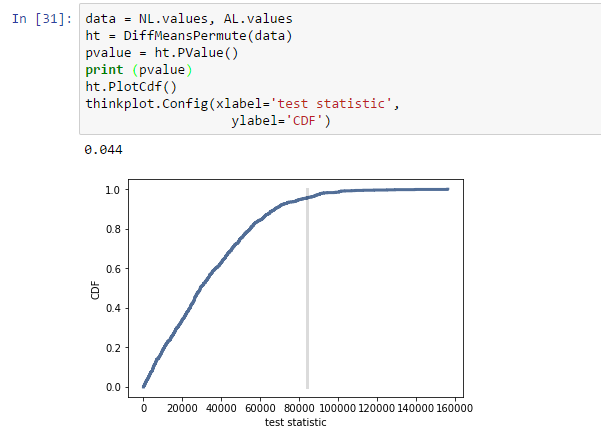
Only Covariance does not mean many things.

Correlation is too small because of negative value.

**PART 6 :**

A **test statistic** is a standardized value that is calculated from sample data during a hypothesis **test.**

**Null hypothesis** = (in a statistical test) the hypothesis that there is no significant difference between specified populations, any observed difference being due to sampling or experimental error.



**NOTE : Pvalues is 0.044. This number shows us that statistically significant.**

**PART 7 :**

To answer the question, we had run some simulations to make sure that our hypthesis is right. So, if say there is a high positive relationship between the salaries of National League and Amrican League. We can conculude that both players of NL and AL gets the almost the same salary.

**REFERENCES :**

Think Stats