**Amrita School of Computing**

**Dept of CSE**

**Foundations of Datascience**

**Assignment2**

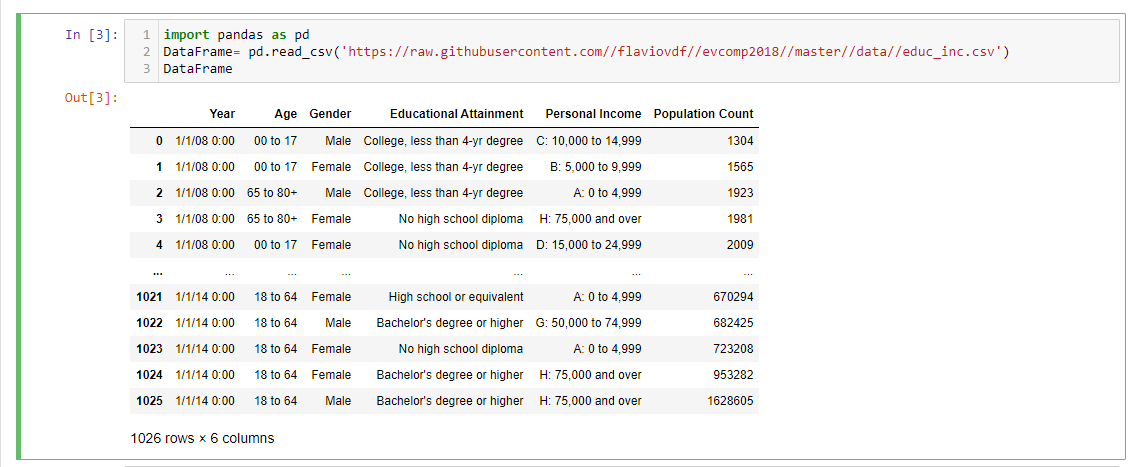
-CB.EN.U4CSE20210

-A.Preetham

1(a) import pandas as pd

DataFrame=pd.read\_csv('https://raw.githubusercontent.com//flaviovdf//evcomp2018//master//data//educ\_inc.csv')

DataFrame



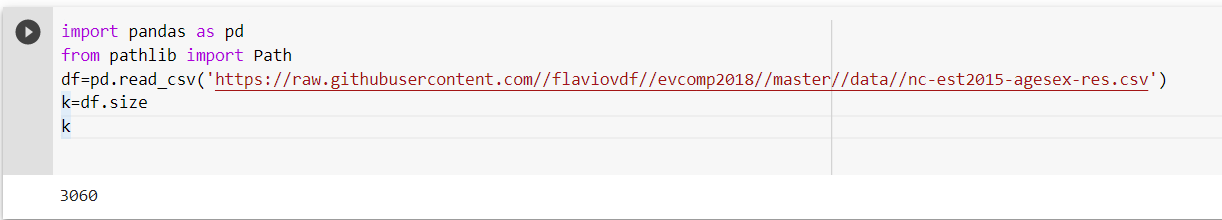
(b) import pandas as pd

from pathlib import Path

df=pd.read\_csv('https://raw.githubusercontent.com//flaviovdf//evcomp2018//master//data//nc-est2015-agesex-res.csv')

k=df.size

k



(c) import pandas as pd

DataFrame=pd.read\_csv('https://raw.githubusercontent.com//flaviovdf//evcomp2018//master//data//educ\_inc.csv')

print(DataFrame.describe())



(d) import pandas as pd

df=pd.read\_csv('https://raw.githubusercontent.com//flaviovdf//evcomp2018//master//data//educ\_inc.csv')

x=pd.DataFrame(df)

x['Year']= x['Year'].str.replace('1/1/08','1/1/2010')

x['Year']= x['Year'].str.replace('1/1/14','1/1/2014')

x

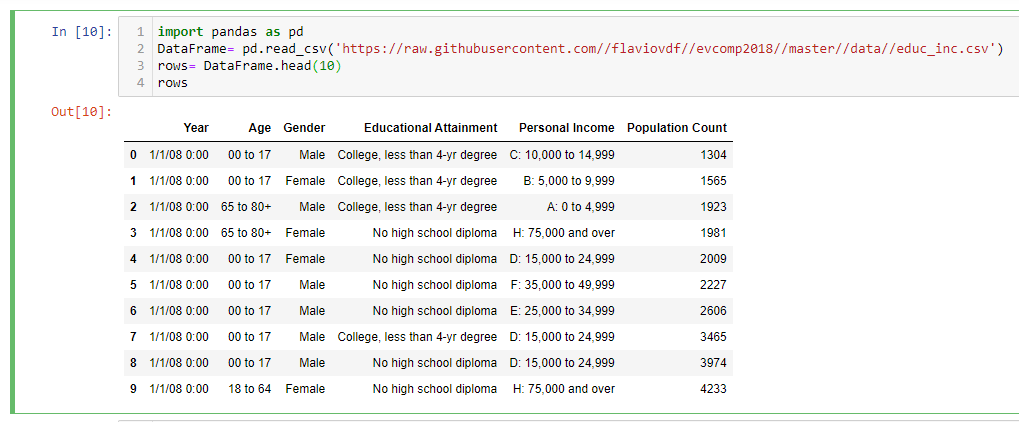


(e) import pandas as pd

DataFrame=pd.read\_csv('https://raw.githubusercontent.com//flaviovdf//evcomp2018//master//data//educ\_inc.csv')

rows= DataFrame.head(10)

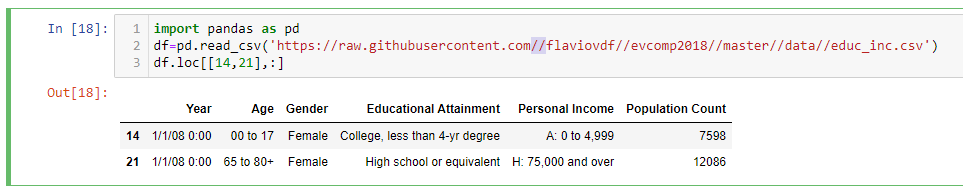
rows



(f) import pandas as pd

df=pd.read\_csv('https://raw.githubusercontent.com//flaviovdf//evcomp2018//master//data//educ\_inc.csv')

df.loc[[14,21],:]



(g) import pandas as pd

df=pd.read\_csv('https://raw.githubusercontent.com//flaviovdf//evcomp2018//master//data//educ\_inc.csv')

df["Change"]= ""

df["Percentage Change"]=""

df.to\_csv('https://raw.githubusercontent.com//flaviovdf//evcomp2018//master//data//educ\_inc.csv', index=False)

x=pd.DataFrame(df)

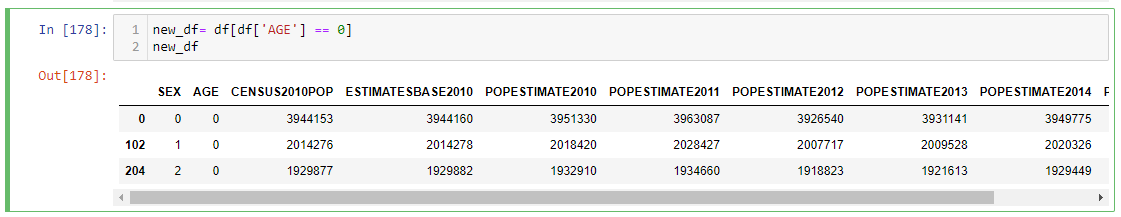
x



(h)

1. new\_df= df[df['AGE'] == 0]

new\_df

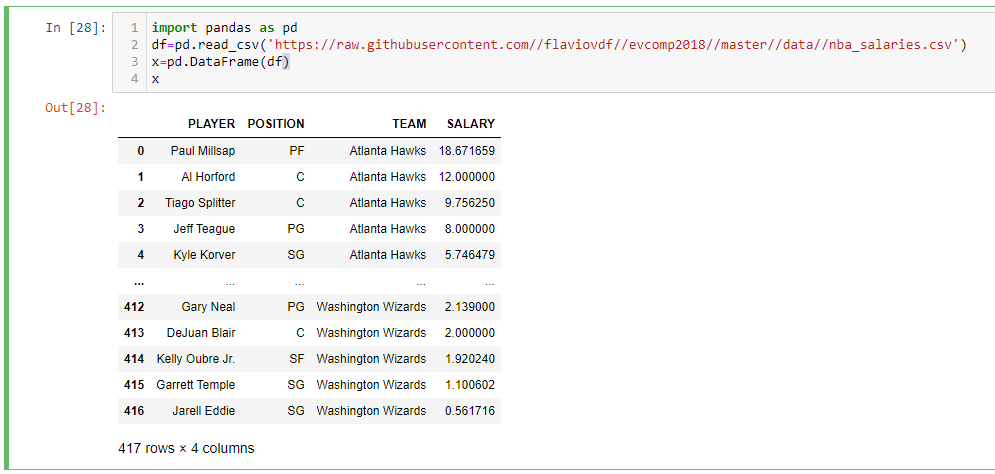


2 (a) import pandas as pd

df=pd.read\_csv('https://raw.githubusercontent.com//flaviovdf//evcomp2018//master//data//nba\_salaries.csv')

x=pd.DataFrame(df)

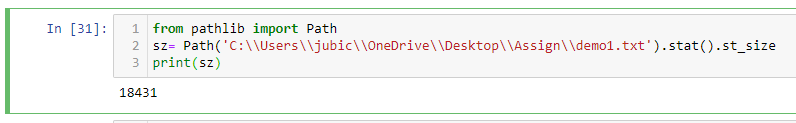
x



(b) from pathlib import Path

sz= Path('C:\\Users\\jubic\\OneDrive\\Desktop\\Assign\\demo1.txt').stat().st\_size

print(sz)

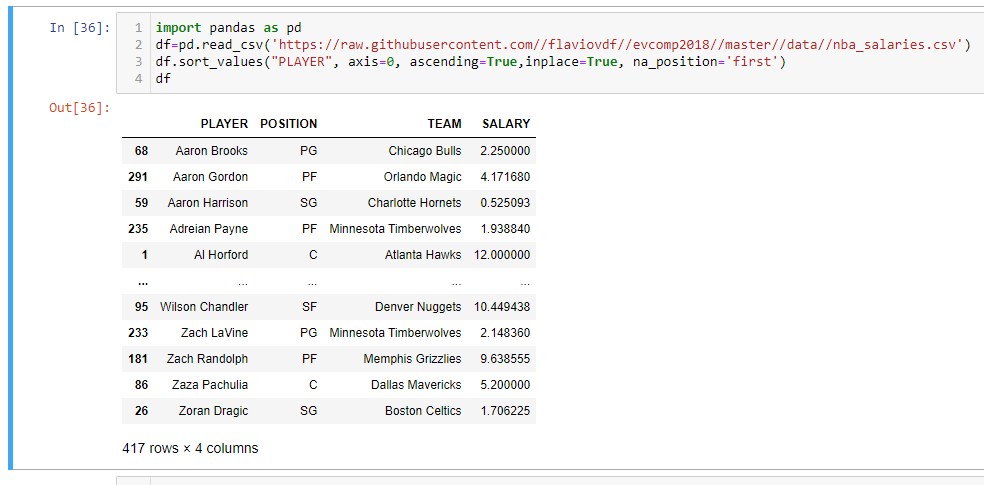


(c) import pandas as pd

df=pd.read\_csv('https://raw.githubusercontent.com//flaviovdf//evcomp2018//master//data//nba\_salaries.csv')

df.sort\_values("PLAYER", axis=0, ascending=True,inplace=True, na\_position='first')

df



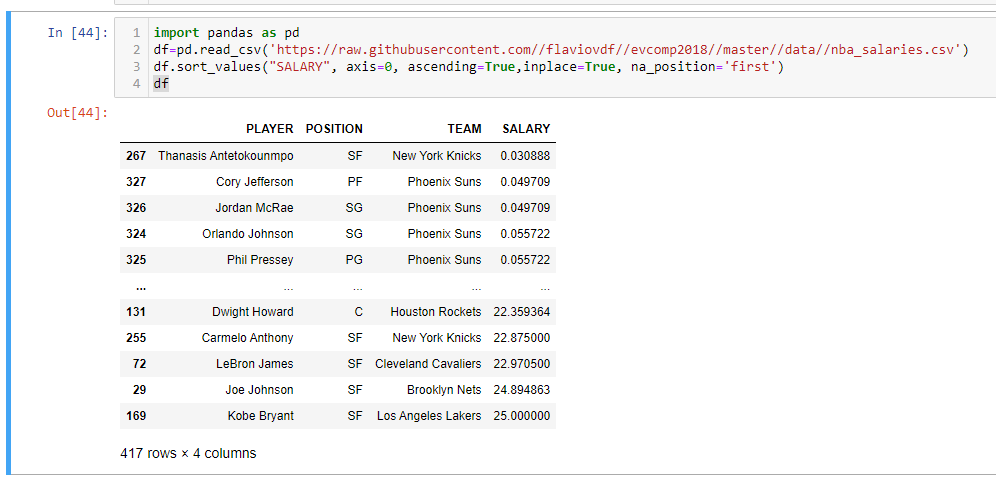
(d) Increasing:

import pandas as pd

df=pd.read\_csv('https://raw.githubusercontent.com//flaviovdf//evcomp2018//master//data//nba\_salaries.csv')

df.sort\_values("SALARY", axis=0, ascending=True,inplace=True, na\_position='first')

df



Decreasing:

df.sort\_values("SALARY", axis=0, ascending=False,inplace=True, na\_position='first')

df



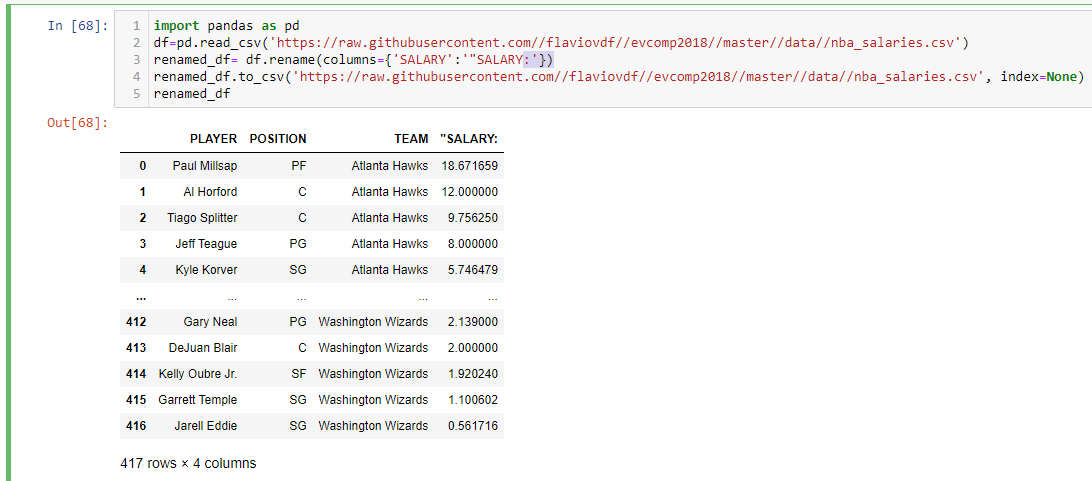
(e) import pandas as pd

df=pd.read\_csv('https://raw.githubusercontent.com//flaviovdf//evcomp2018//master//data//nba\_salaries.csv')

renamed\_df= df.rename(columns={'SALARY':'"SALARY:'})

renamed\_df.to\_csv('https://raw.githubusercontent.com//flaviovdf//evcomp2018//master//data//nba\_salaries.csv', index=None)

renamed\_df



(f) import pandas as pd

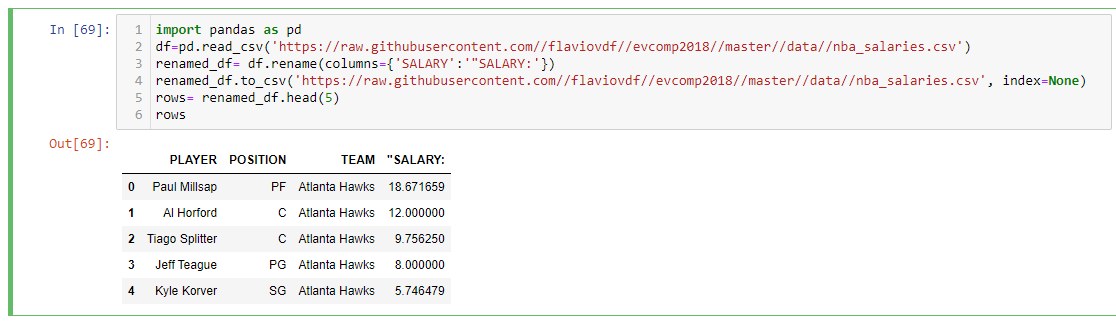
df=pd.read\_csv('https://raw.githubusercontent.com//flaviovdf//evcomp2018//master//data//nba\_salaries.csv')

renamed\_df= df.rename(columns={'SALARY':'"SALARY:'})

renamed\_df.to\_csv('https://raw.githubusercontent.com//flaviovdf//evcomp2018//master//data//nba\_salaries.csv', index=None)

rows= renamed\_df.head(5)

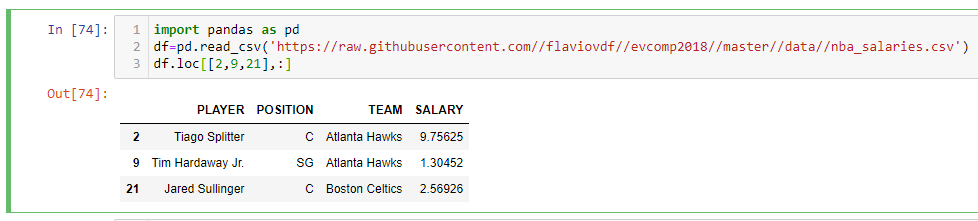
rows



(g) import pandas as pd

df=pd.read\_csv('https://raw.githubusercontent.com//flaviovdf//evcomp2018//master//data//nba\_salaries.csv')

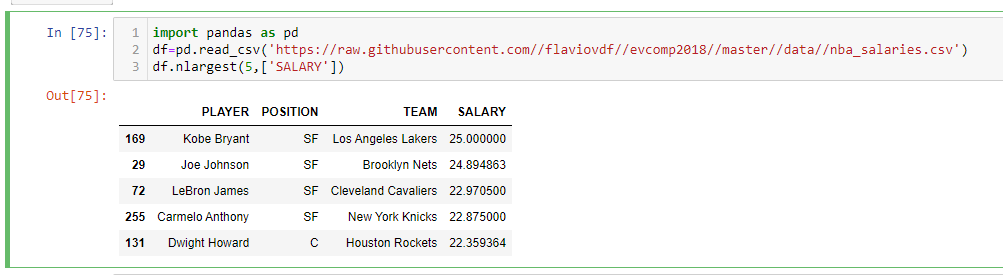
df.loc[[2,9,21],:]



(h) import pandas as pd

df=pd.read\_csv('https://raw.githubusercontent.com//flaviovdf//evcomp2018//master//data//nba\_salaries.csv')

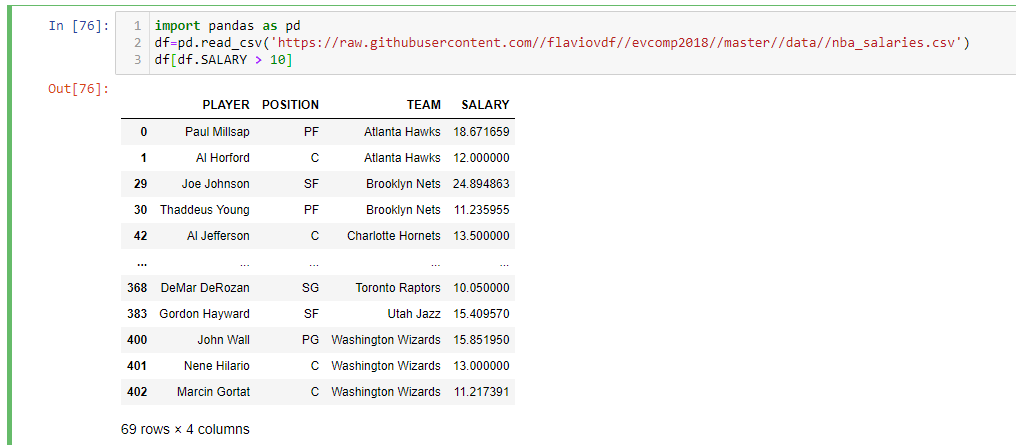
df.nlargest(5,['SALARY'])



1. import pandas as pd

df=pd.read\_csv('https://raw.githubusercontent.com//flaviovdf//evcomp2018//master//data//nba\_salaries.csv')

df[df.SALARY > 10]



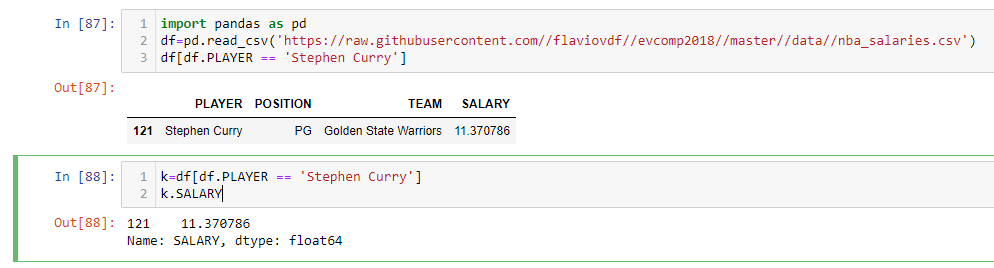
(j) import pandas as pd

df=pd.read\_csv('https://raw.githubusercontent.com//flaviovdf//evcomp2018//master//data//nba\_salaries.csv')

df[df.PLAYER == 'Stephen Curry']

k=df[df.PLAYER == 'Stephen Curry']

k.SALARY

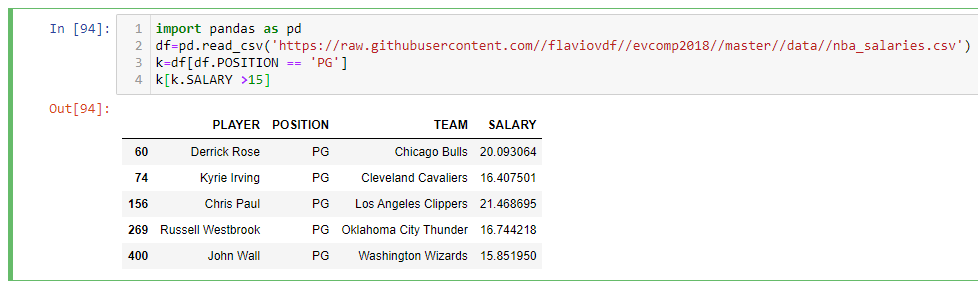


(k) import pandas as pd

df=pd.read\_csv('https://raw.githubusercontent.com//flaviovdf//evcomp2018//master//data//nba\_salaries.csv')

k=df[df.POSITION == 'PG']

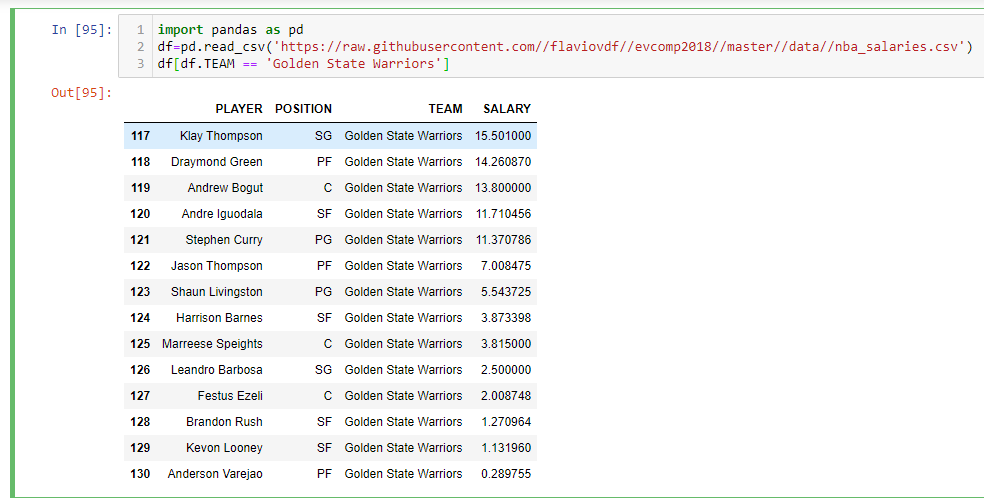
k[k.SALARY >15]



(l) import pandas as pd

df=pd.read\_csv('https://raw.githubusercontent.com//flaviovdf//evcomp2018//master//data//nba\_salaries.csv')

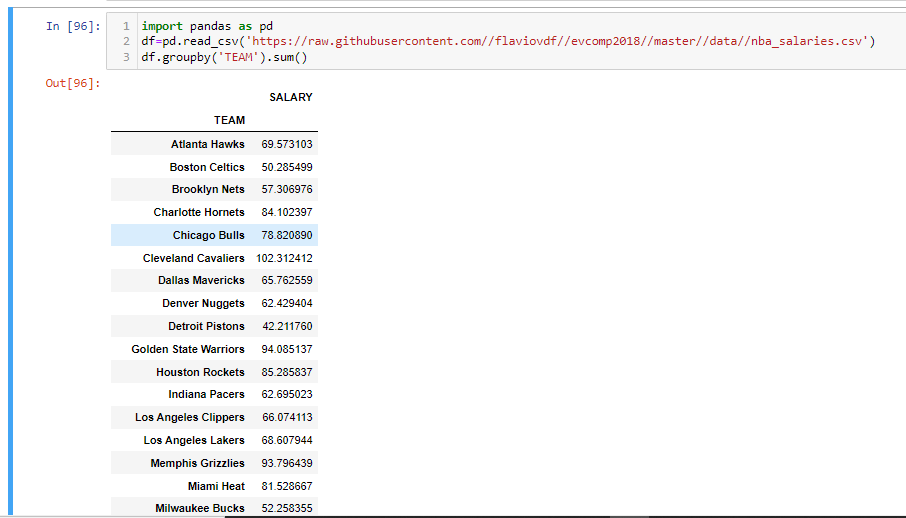
df[df.TEAM == 'Golden State Warriors']

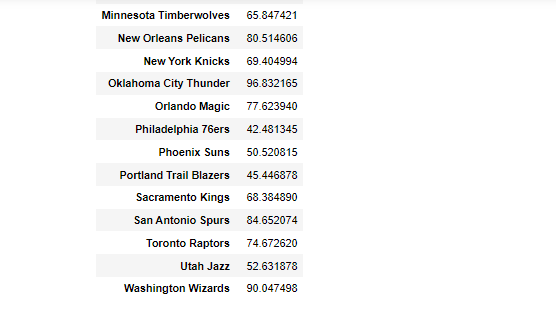


(m) import pandas as pd

df=pd.read\_csv('https://raw.githubusercontent.com//flaviovdf//evcomp2018//master//data//nba\_salaries.csv')

df.groupby('TEAM').sum()

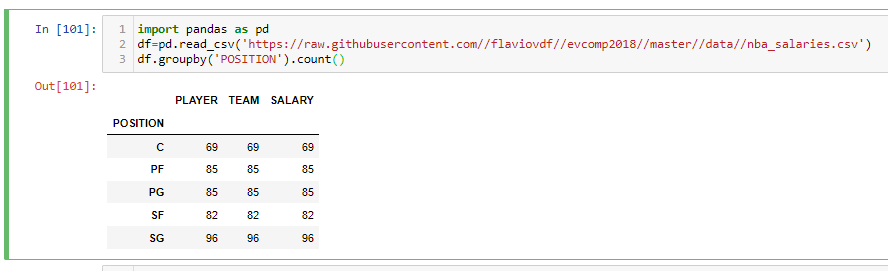




(n) import pandas as pd

df=pd.read\_csv('https://raw.githubusercontent.com//flaviovdf//evcomp2018//master//data//nba\_salaries.csv')

df.groupby('POSITION').count()



(o) import pandas as pd

df=pd.read\_csv('https://raw.githubusercontent.com//flaviovdf//evcomp2018//master//data//nba\_salaries.csv')

df.groupby('POSITION').mean()



1. For each year, the table records the Population Count of Californians in many different combinations of age, gender, educational attainment, and personal income. We will study only the data for the year 2014.

import numpy as np

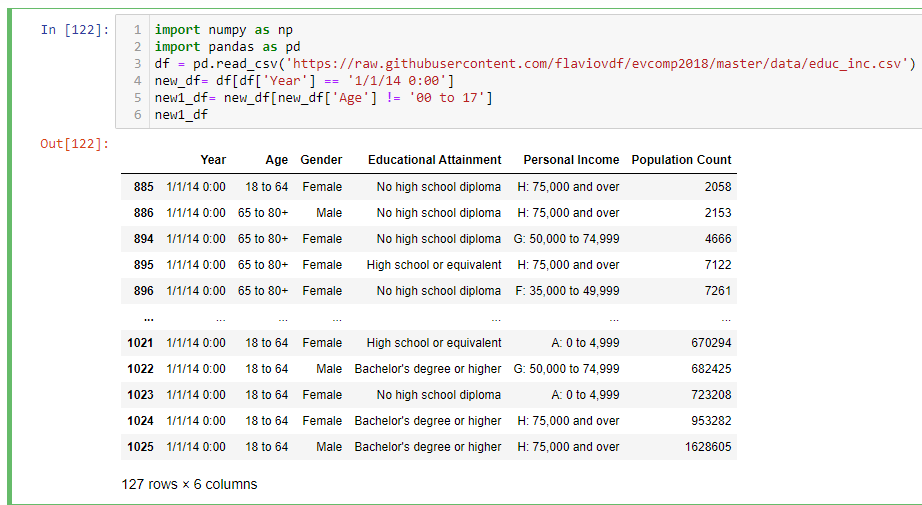
import pandas as pd

df = pd.read\_csv('https://raw.githubusercontent.com/flaviovdf/evcomp2018/master/data/educ\_inc.csv')

new\_df= df[df['Year'] == '1/1/14 0:00']

new1\_df= new\_df[new\_df['Age'] != '00 to 17']

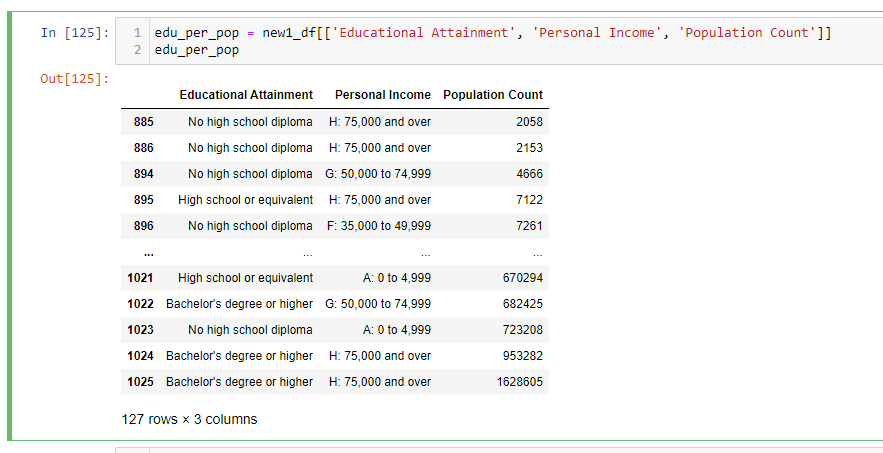
new1\_df



Educational attainment and personal income.

edu\_per\_pop = new1\_df[['Educational Attainment', 'Personal Income', 'Population Count']]

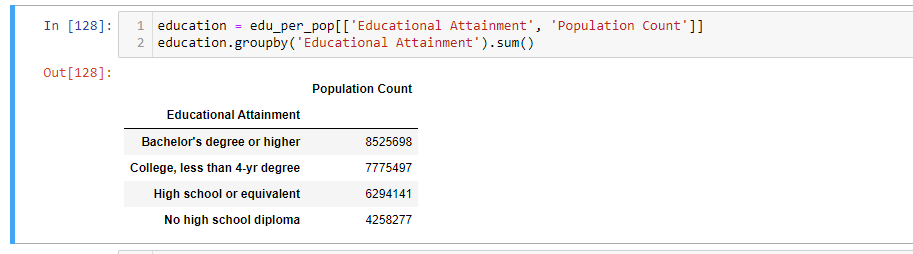
edu\_per\_pop



1 Group the table by Educational Attainment and sum the Population Count in each category

education = edu\_per\_pop[['Educational Attainment', 'Population Count']]

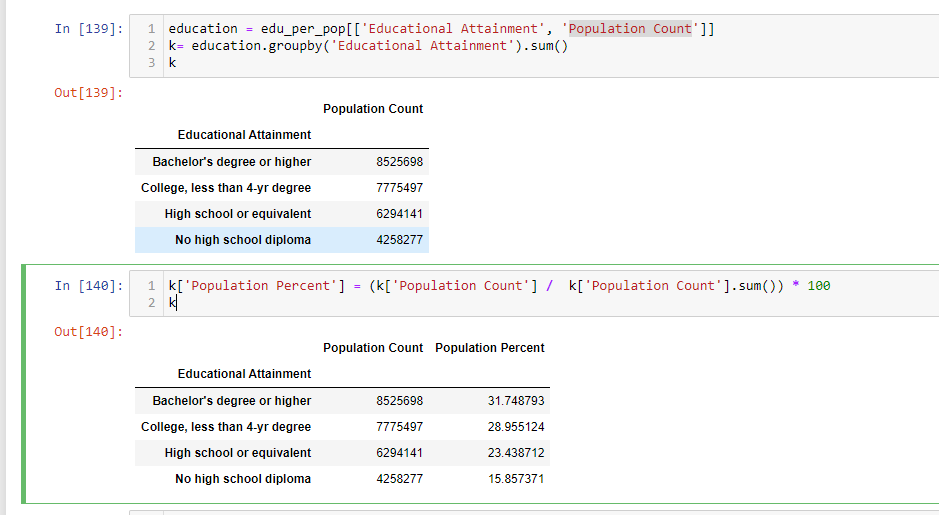
education.groupby('Educational Attainment').sum()



2 Analyse the percentage distribution of educational attainment among adult Californians.

k['Population Percent'] = (k['Population Count'] / k['Population Count'].sum()) \* 100

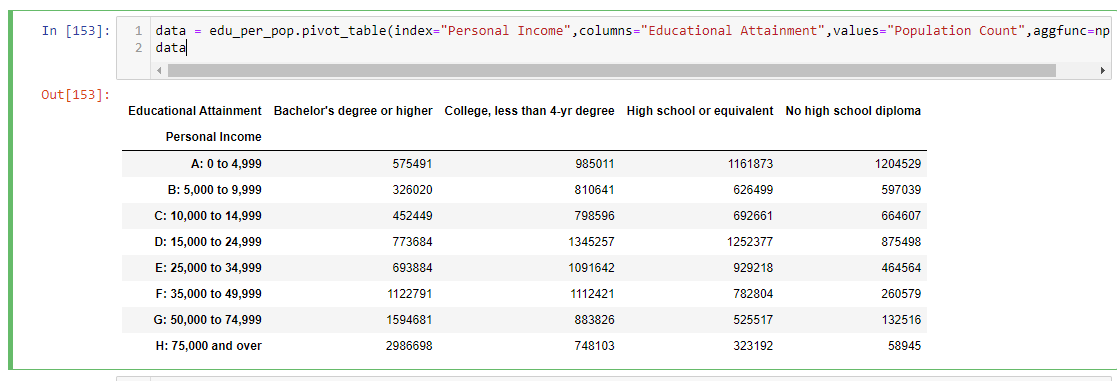
k



3 Using pivot , get a contingency table (a table of counts) of adult Californians cross-classified by Educational Attainment and Personal Income.

data = edu\_per\_pop.pivot\_table(index="Personal Income",columns="Educational Attainment",values="Population Count",aggfunc=np.sum)

data



*Converting the counts to percents allows us to compare the four distributions*

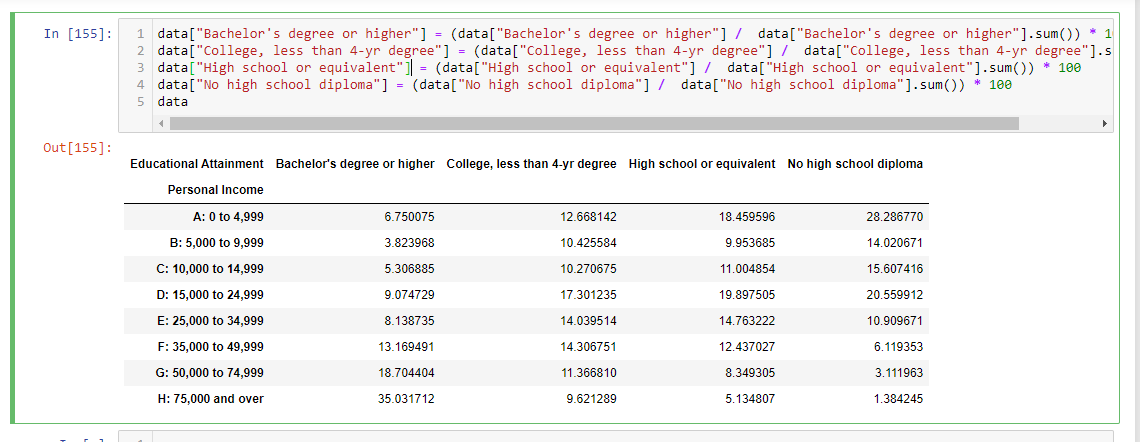
data["Bachelor's degree or higher"] = (data["Bachelor's degree or higher"] / data["Bachelor's degree or higher"].sum()) \* 100

data["College, less than 4-yr degree"] = (data["College, less than 4-yr degree"] / data["College, less than 4-yr degree"].sum()) \* 100

data["High school or equivalent"] = (data["High school or equivalent"] / data["High school or equivalent"].sum()) \* 100

data["No high school diploma"] = (data["No high school diploma"] / data["No high school diploma"].sum()) \* 100

data



4 import matplotlib.pyplot as plt

import pandas as pd

data.plot(kind='barh')

