Income Analysis Report

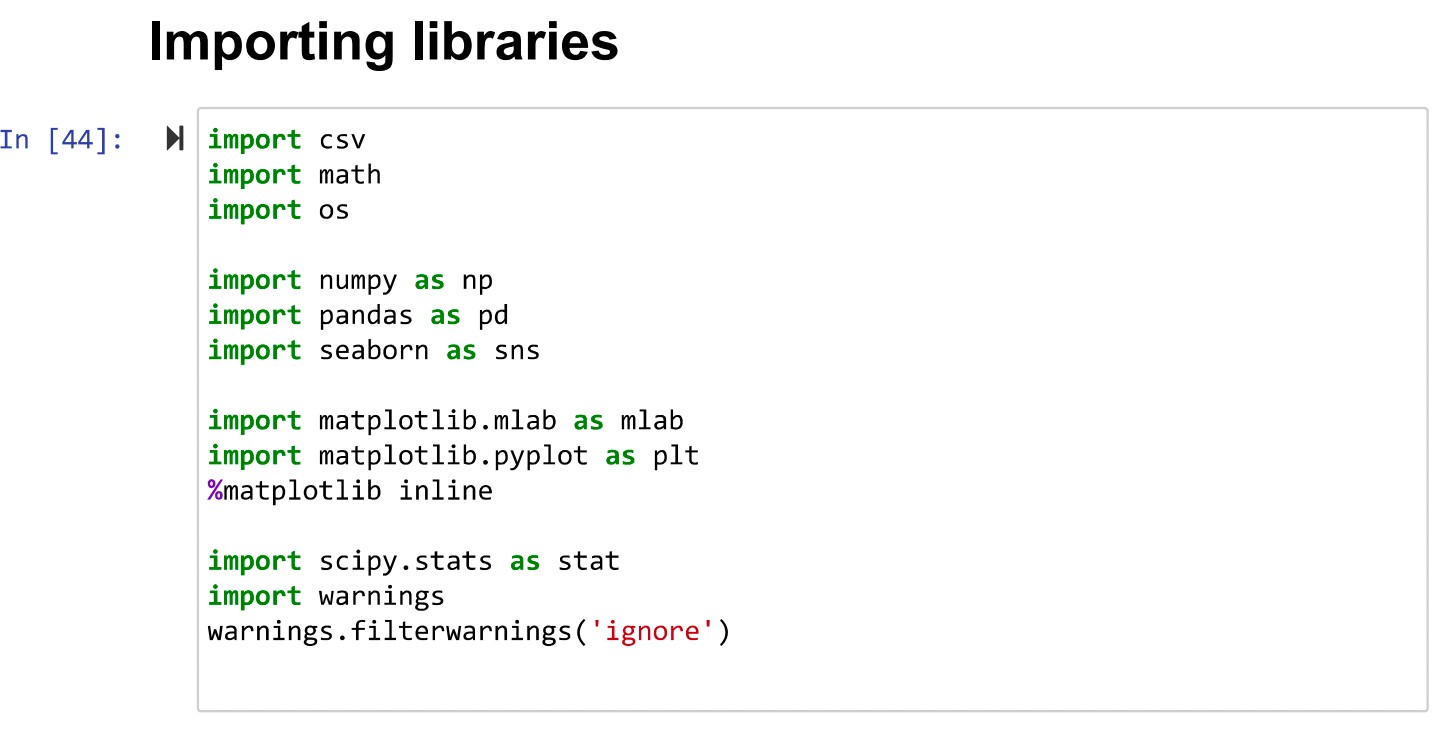
# Abstract

Everybody in society will have the responsibility to manage their living in space. They have varied due to the area can have an income-based on working progress. This progress has an impact on the making lifestyle of that particular area. This will be analyzed based on the income over a particular zone and find the data regarding the working of a person according to the work they do over a period and suggesting to making what will they get in their making income in a manner legal way. This will elaborate on forwarding about the making of the income they get less or more about their work. This will be an aid to the organization regarding the payment of the employer based on their work. This information With a dataset of information from the panel study of income dynamics about the people's household survey, we can measure the hours worked by people.

# Introduction

This project will perform statistical tests on the dataset, remove outliers, and check the correlation with the heat map. In this project, each word is represented by a set of real numbers, and the embeddings exhibit using properties.

Loading needed libraries such as math, seaborn, NumPy, pandas, matplotlib, PyPlot.



**View of the Dataset**

Composing of the data with working, income, age, education, hours of work, kids, marriage status. Nearly 4.8k data are developed. The dataset contains 4855 columns. People’s age group is between 30 to 50, Strong relation found between education & earnings, hours & earnings.

In [74]: df= pd.read\_csv('working.csv')

df

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Out [74] : | Seq No | intnum | persnum | age | educatn | earnings | hours | kids | married |
|  | 0 1 | 4 | 4 | 39 | 12.0 | 77250 | 2940 | 2 | married |
|  | 1 2 | 4 | 6 | 35 | 12.0 | 12000 | 2040 | 2 | divorced |
|  | 3 | 4 | 7 | 33 | 12.0 | 8000 | 693 | 1 | married |
|  | 3 4 | 4 | 173 | 39 | 10.0 | 15000 | 1904 | 2 | married |
|  | 4 5 | 5 | 2 | 47 | 9.0 | 6500 | 1683 | 5 | married |
|  | 4851 4852 | 9297 | 2 | 42 | 2.0 | 3000 | 1040 | 4 | married |
|  | 4852 4853 | 9301 | 2 | 43 | 12.0 | 0 | 0 | 2 | married |
|  | 4853 4854 | 9302 | 1 | 37 | 8.0 | 22045 | 2793 | 98 | divorced |
|  | 4854 4855 | 9300 | 2 | 40 | 6.0 | 134 | 30 | 3 | married |
|  | 4855 4856 | 9306 | 2 | 37 | 17.0 | 33000 | 2423 | 4 | married |

DATA PREPROCESSING

drop columns:- Here, we are dropping the unrequired columns for the analysis

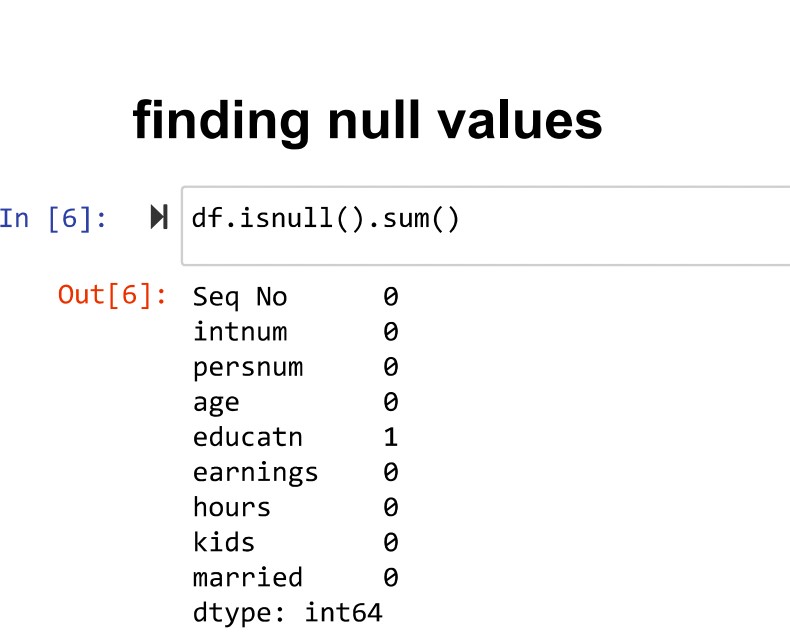
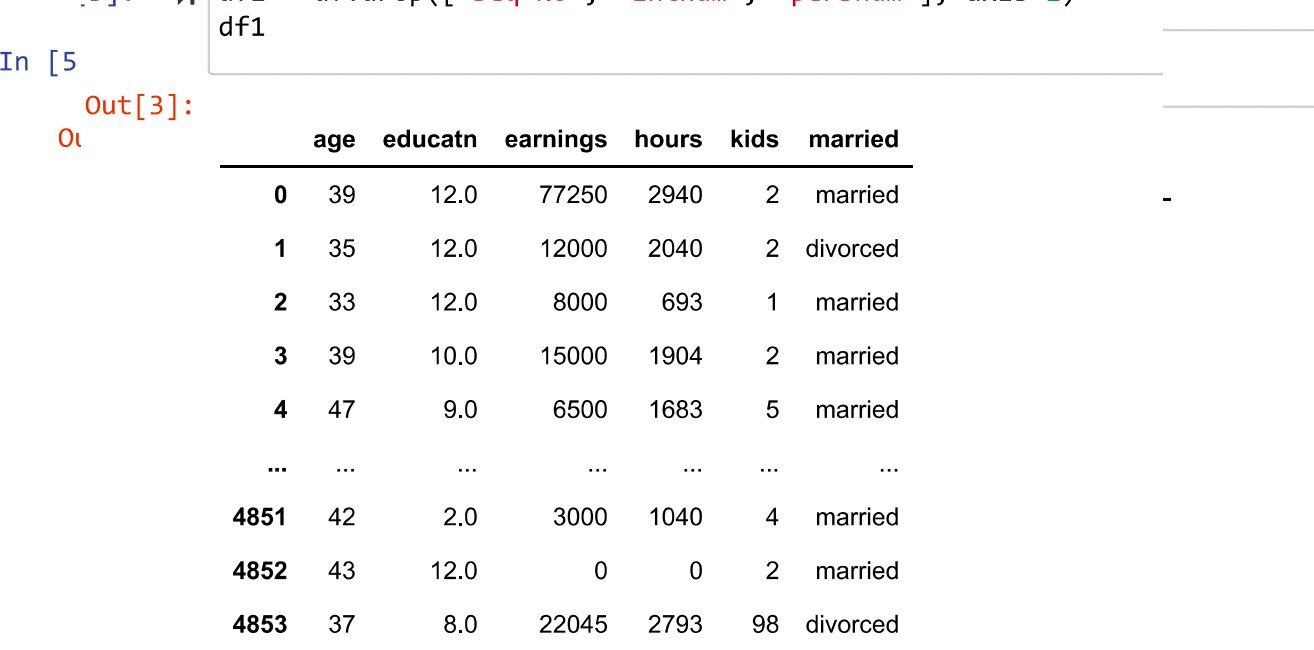
3] : @ df1 = df.drop(["Seq No”, ”intnum“, “persnum"], axis=1) df1

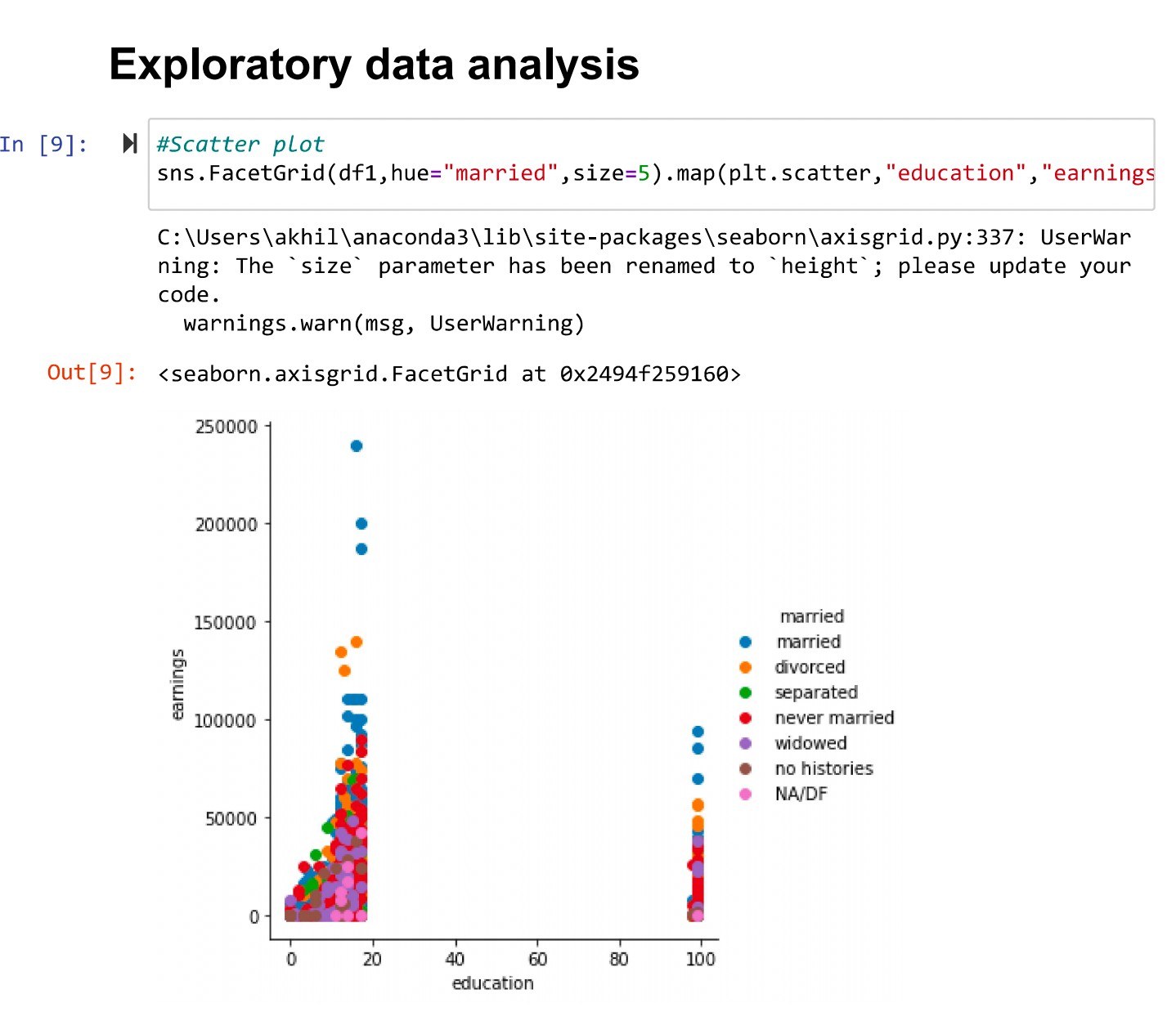
|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Out[3]: |  | | | | | | |
|  |  | age | educatn | earnings | hours | kids | married |
|  | 0 | 39 | 12.0 | 77250 | 2940 | 2 | married |
|  | 1 | 35 | 12.0 | 12000 | 2040 | 2 | divorced |
|  | 2 | 33 | 12.0 | 8000 | 693 | 1 | married |
|  | 3 | 39 | 10.0 | 15000 | 1904 | 2 | married |
|  | 4 | 47 | 9.0 | 6500 | 1683 | 5 | married |
|  | 4851 | 42 | 2.0 | 3000 | 1040 | 4 | married |
|  | 4852 | 43 | 12.0 | 0 | 0 | 2 | married |
|  | 4853 | 37 | 8.0 | 22045 | 2793 | 98 | divorced |
|  | 4854 | 40 | 6.0 | 134 | 30 | 3 | married |
|  | 4855 | 37 | 17.0 | 33000 | 2422 | 4 | married |

4856 rows • 6 column

**Describing**

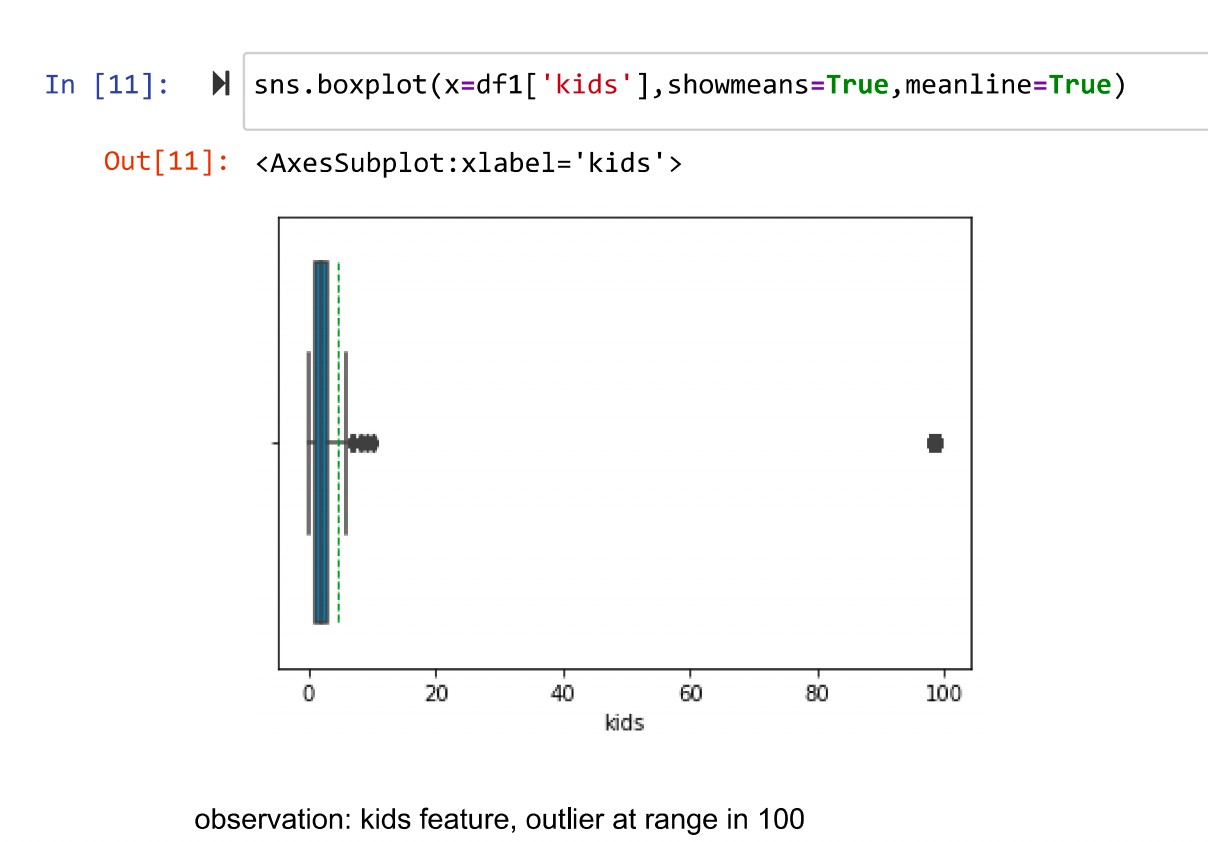
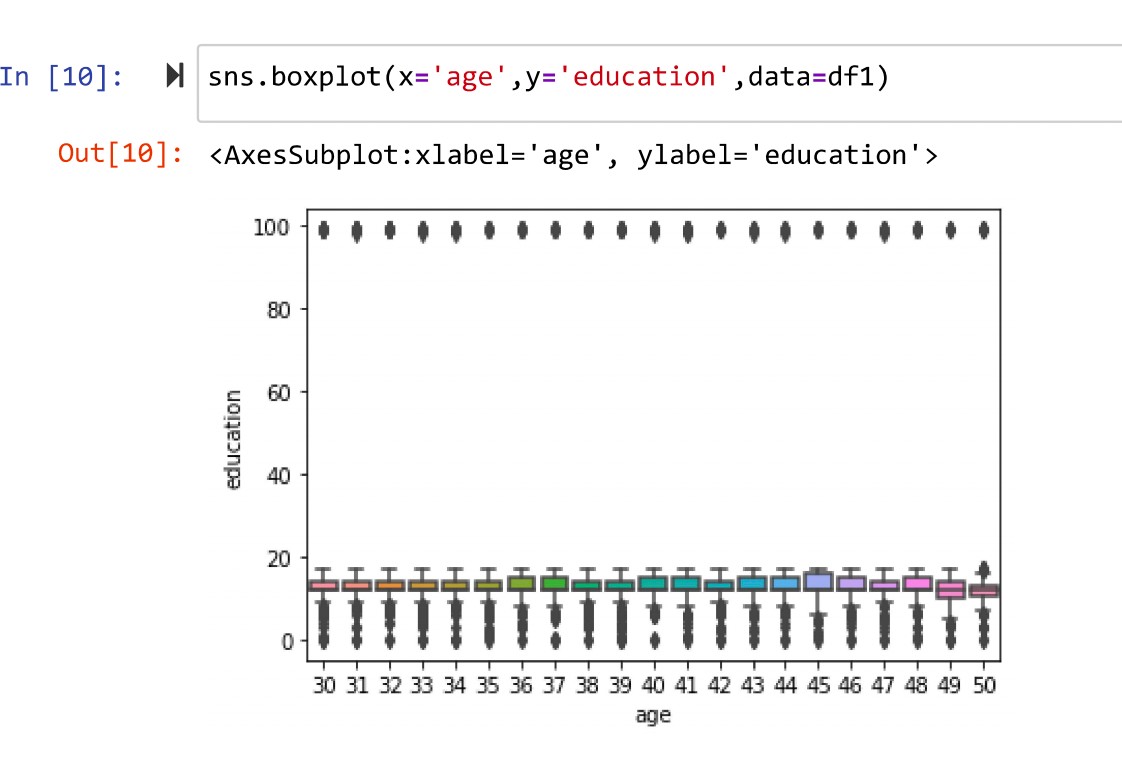
Each age group has been taken into consideration with their kids too.





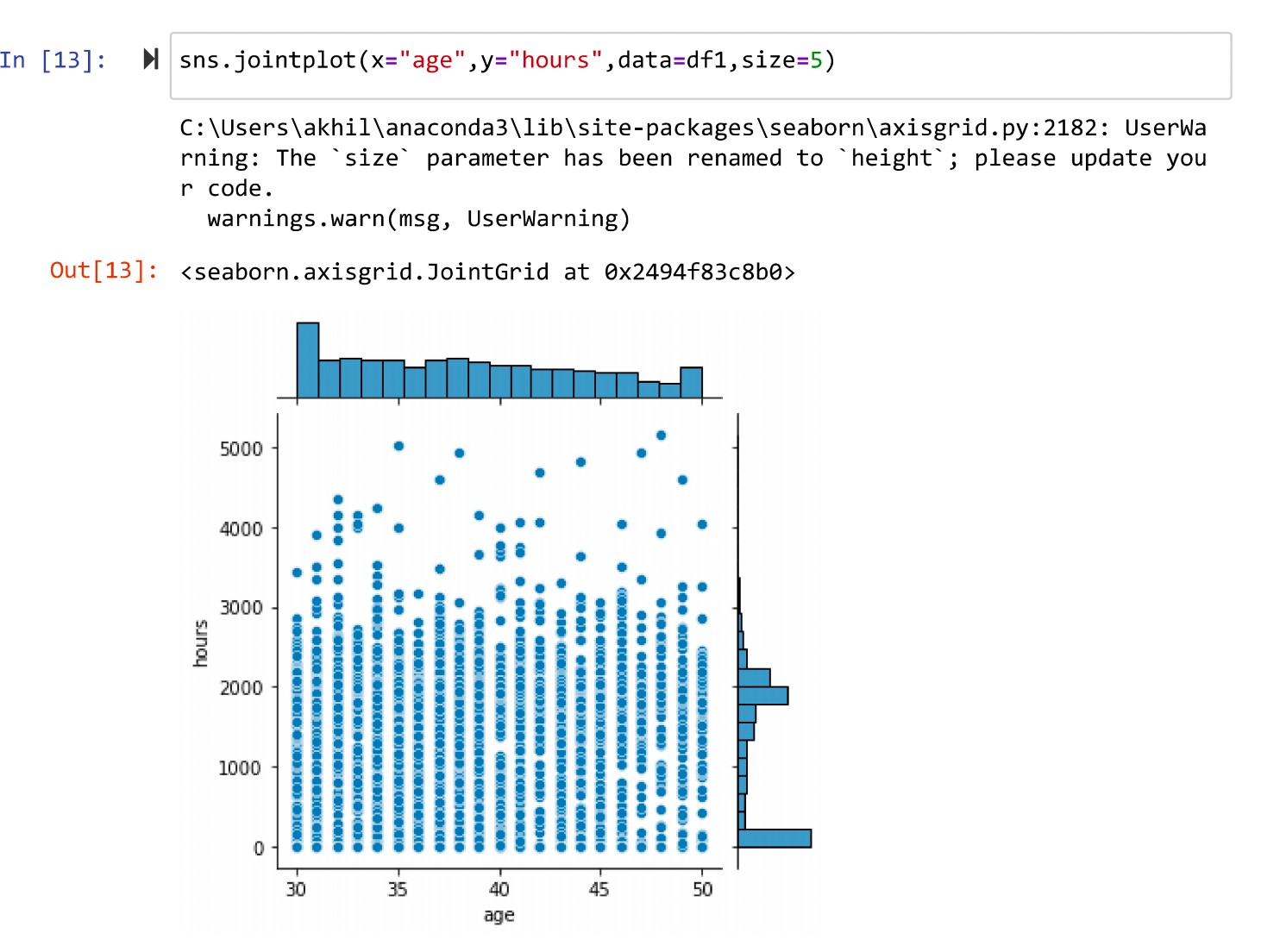
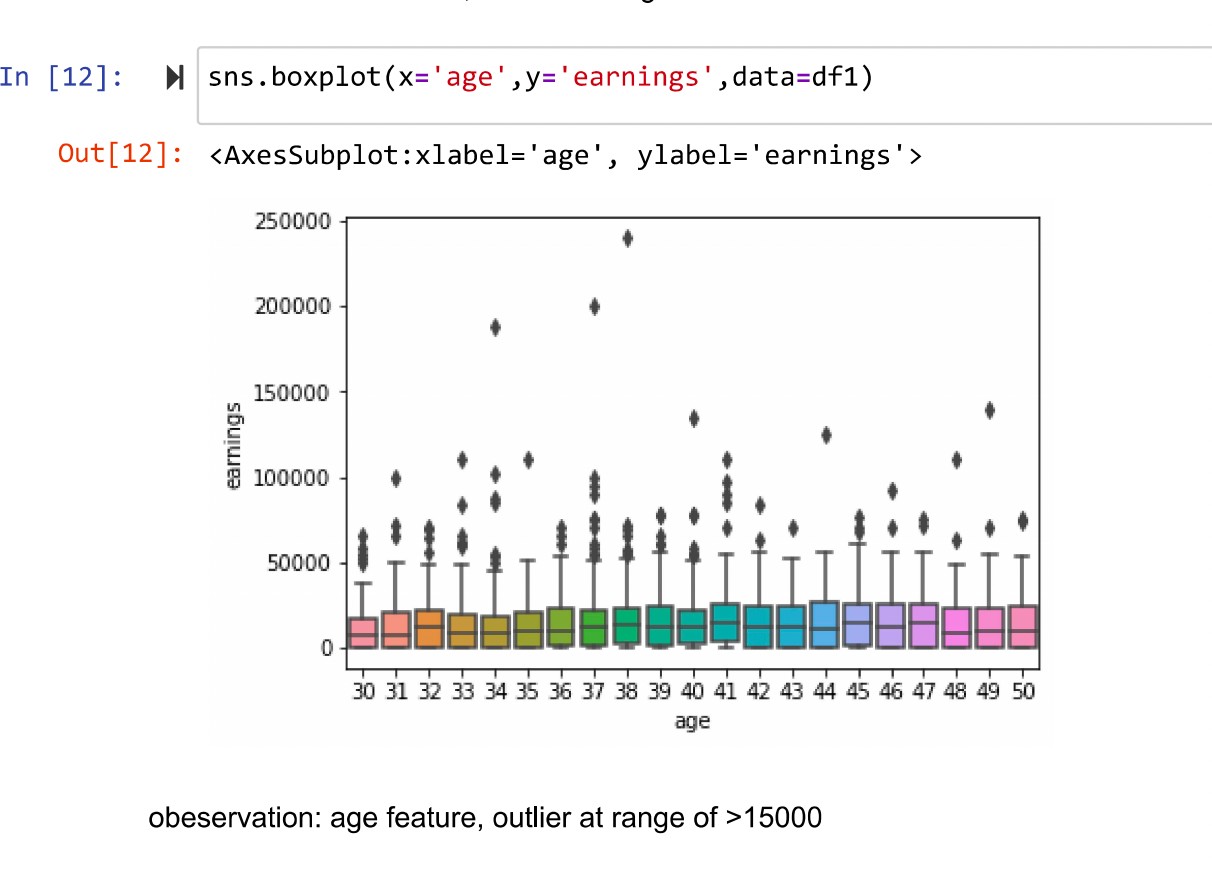
Null values are in the unfilled data position in order to derive good and eﬃcient data processes





Can see the education with the age.

Can the box plot showing till 15 kids



Remove outliers ( kids >=20 and education >= 20 and earnings >= 150000 and Hours=0)

As of kids and education more than 20 is not valid.

In [17]: H

Out [ I7 ] :

df1 = df1.drop(df1[df1.education > 20].index) df1['education'].plot(kind='hist', bins=50) pit.xlabel('Education Level')

Text: (B . 5, 6, ' Edu cat:1on Level )

1750

1500

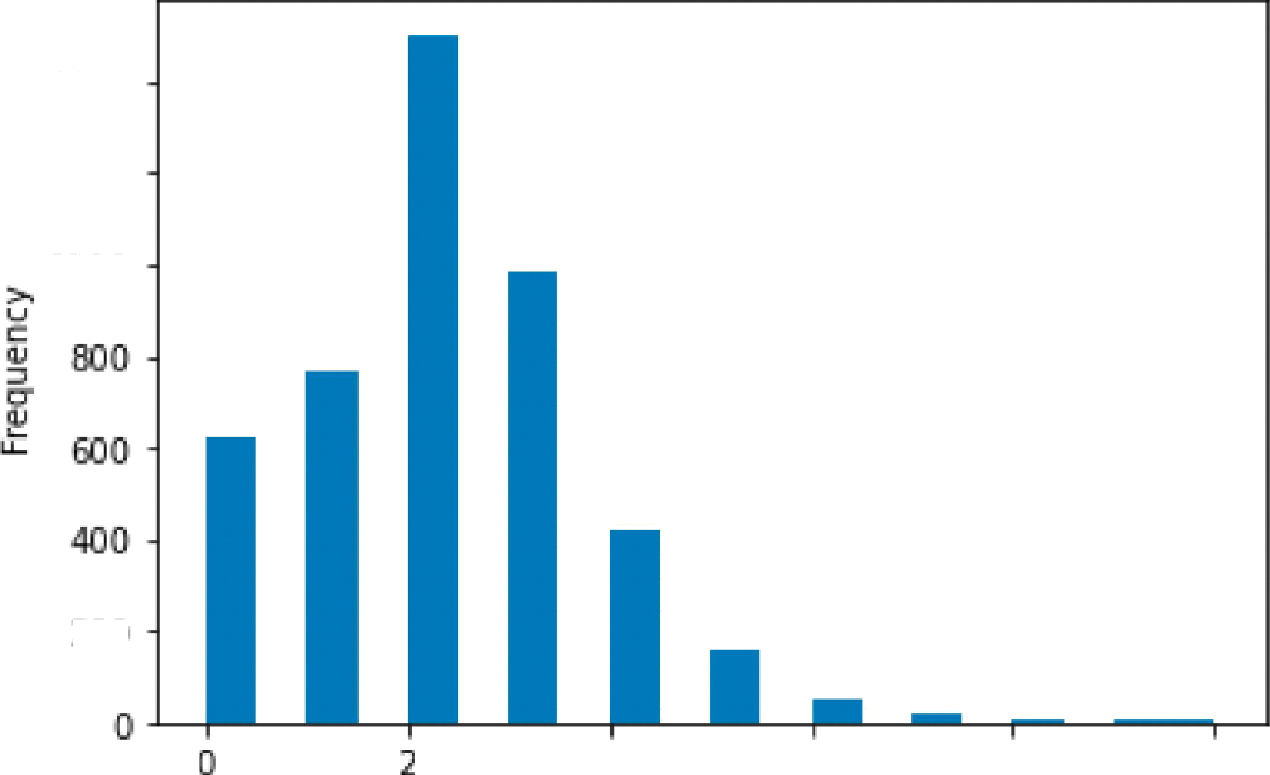
1250

1000

500

In [18]: 8 df1 = dfl.drop(df1[df1.kids > 20].index) df1[' kids'].plot(kind='hist', bins=20) plt.xlabel('Number of Kids')

Out[18]: Text(0.5, 0, 'Number of Kids')



140 0

120 0

100 0

200

4

6

B

10

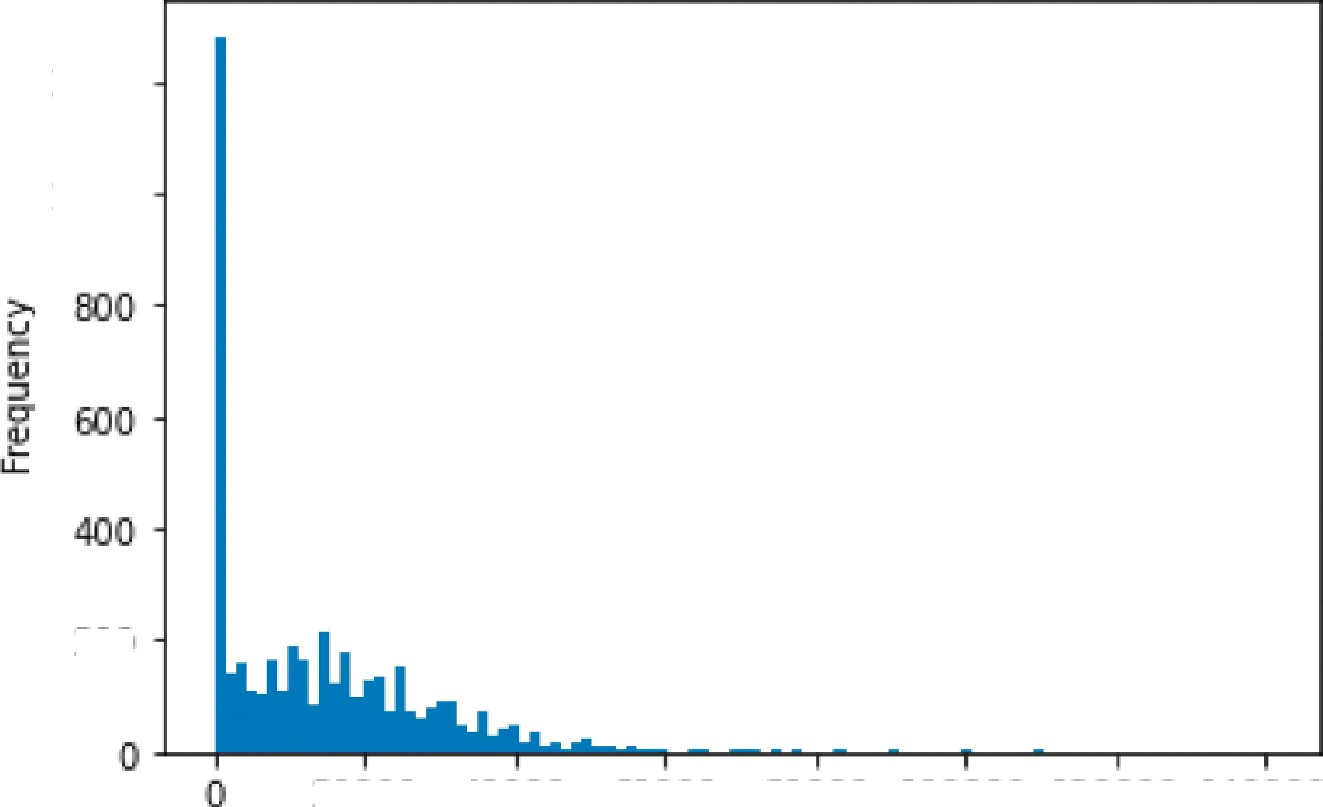
Number of Kids

In [20]: H df1 = df1.drop(df1[df1.earnings › 150000].index)

df1['earnings'].plot(kind='hist', bins=100)

plt.xlabel('earnings')

Out[20]: Text(0.5, 0, 'earnings')



1200

1000

ZOO

20000 400 00 L0000 BOX 00 100000 12O 000 140000

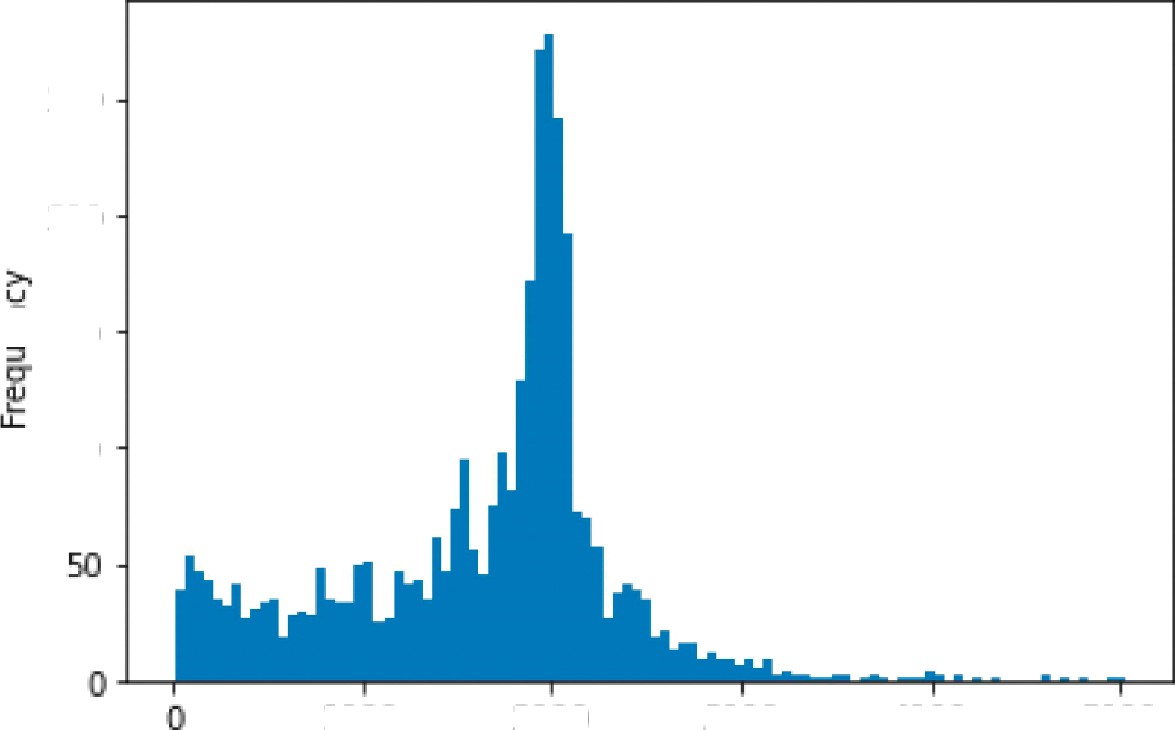
earning s

In [21]: |H

df1 = df1.drop(df1[df1.hours == O].index) df1['hours '].plot(kind='hist', bins=100) plt.x1abel('Hours')

Out[21]: Text(0.5, 0, 'Hours')

Hours



250

700

150

100

J00 0

2000

3000

4000

5000

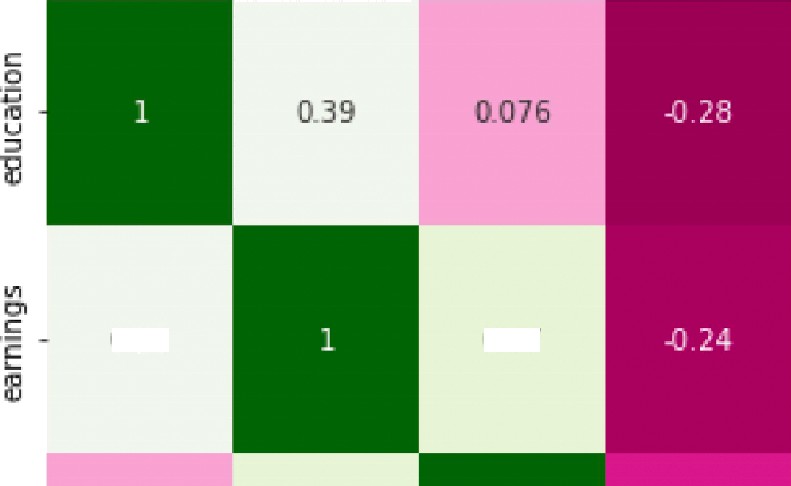
**strong correlation is in** between Education & Earning, Hours & Earning

In [26]: H corrmap = df1.corr(method='pearson')

earningfeatures corrmap.index[abs(corrmap['earnings']) ›= 0.2] print(earningfeatures)

plt.figure(figsize:(6,6))

ax sns.heatmap(df1[earningfeatures].corr(), annot=True, cmap=”PiYG") Index(['education', 'earnings', ' hours', 'kids'], dtype='object')



C.39

- 08

- 06

- 0 4

In [66] : N

In [67] : g

samplingsize = 1500

hypothesisVal = df1.earnings.median() print(hypothesisVal)

16000.0

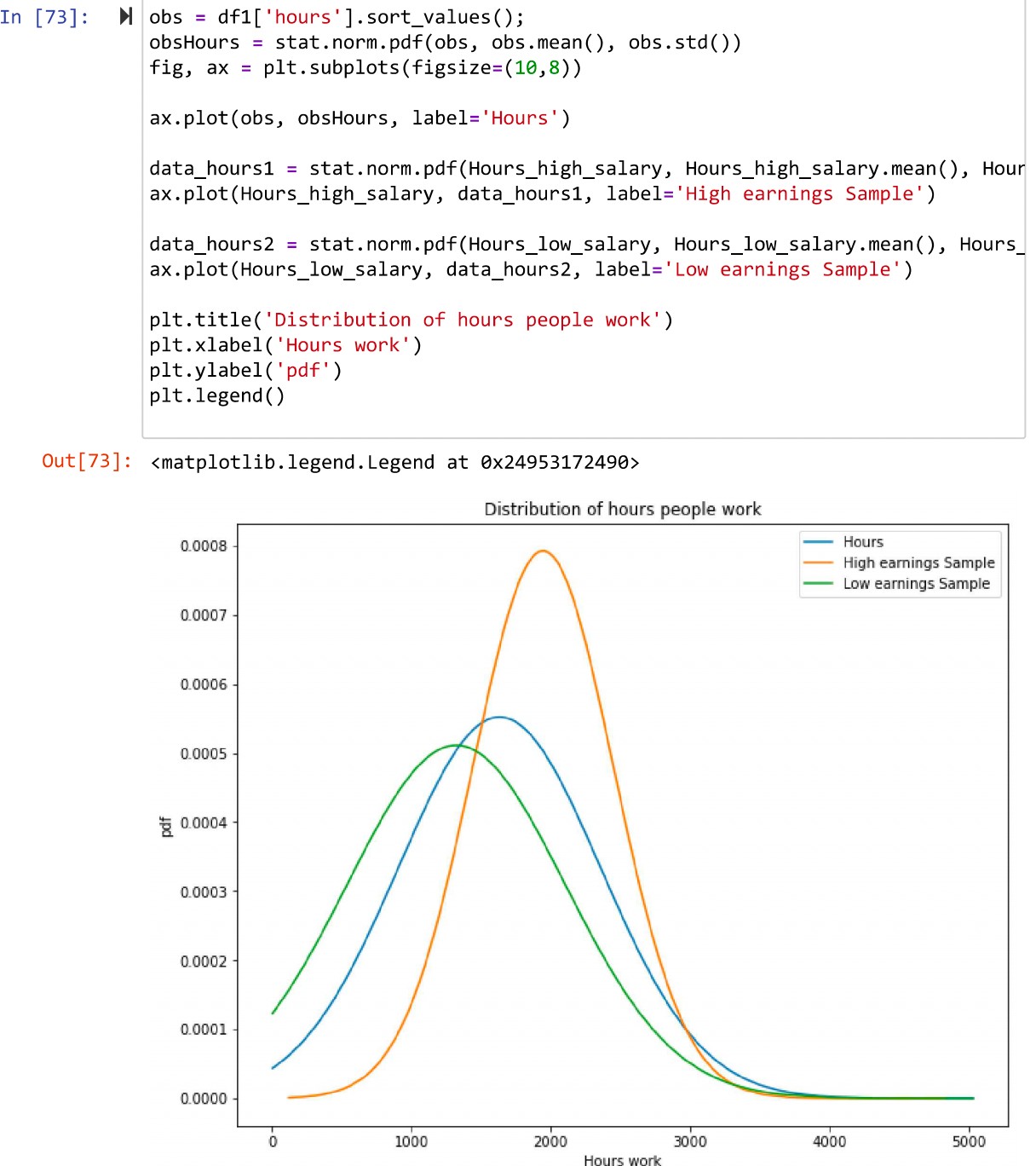
Hours\_high\_salary = df1[df1.earnings› hypothesisVal].hours.sample(samplingsiz Hours\_high\_salary.describe()

|  |  |  |
| --- | --- | --- |
| Out [ 67 : | count | 1500.000000 |
|  | mean | 1944.961333 |
|  | std | 503.459550 |
|  | m1n | **122.000006** |
|  | 25% | 1776.000000 |
|  | 50% | 1960.000000 |
|  | 75% | 2100.000000 |
|  | max | 4821.000000 |
|  | Name: | hours, dtype: float64 |

In [68] : N

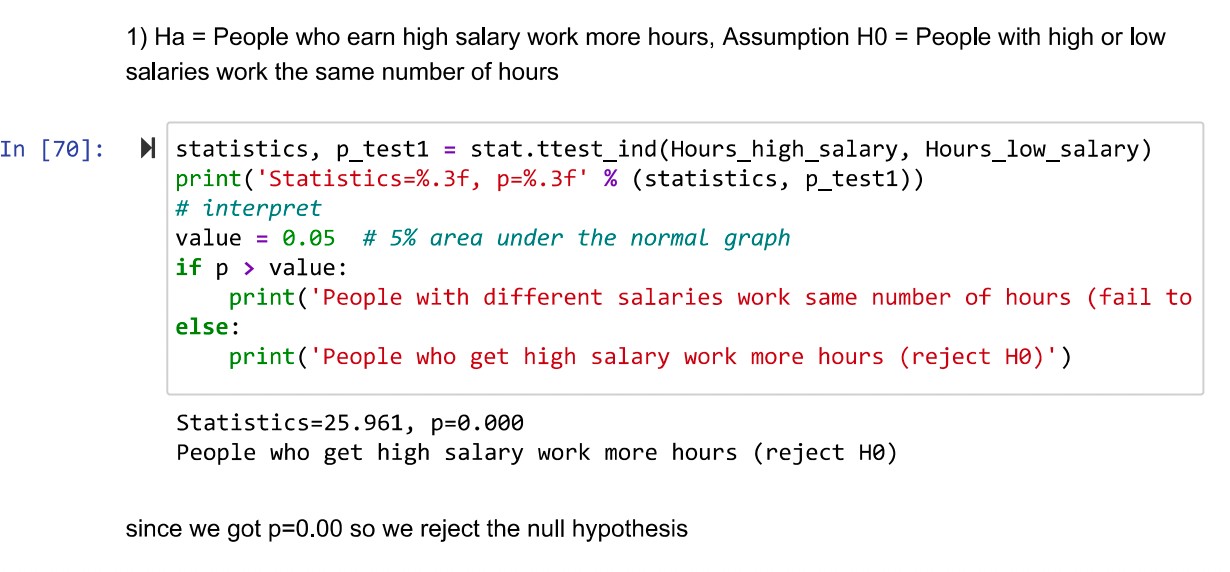
Hours\_low\_sa1ary = df1[df1.earnings<= hypothesisVal].hours.sample(samplingsiz Hours\_low\_sa1ary.describe()

|  |  |  |
| --- | --- | --- |
| Out [ 68 ] : | count | 1500.000000 |
|  | mean | 1322.400000 |
|  | std | 780.491698 |
|  | min | 6.000000 |
|  | 25% | 672.000000 |
|  | 50% | 1440.000000 |
|  | 75% | 1960.000000 |
|  | max | 5025.000000 |
|  | Name: | hours, dtype: float64 |

**\**

**Statistical observation**

Over the phase, we can P-Test hours high low earnings when the graph shows low earnings. Every high-earning sample is based on the hours. One of the most common and important financial statements you'll come across is an income statement. Income statements, also referred as profits and losses (P&L) statements, describe all income and costs for a certain time period, including cumulative impact of income, gain, expenditure, and loss transactions.



# Conclusion

We got the statistics value 25.961, p=0.00 reject the Ho, that is people who get high salary work more hours.

Finally, in this, we have analyzed the values of higher-income people's salaries to the lower- income people.