1.In this question 1, we will use attached dataset as FDA.csv.

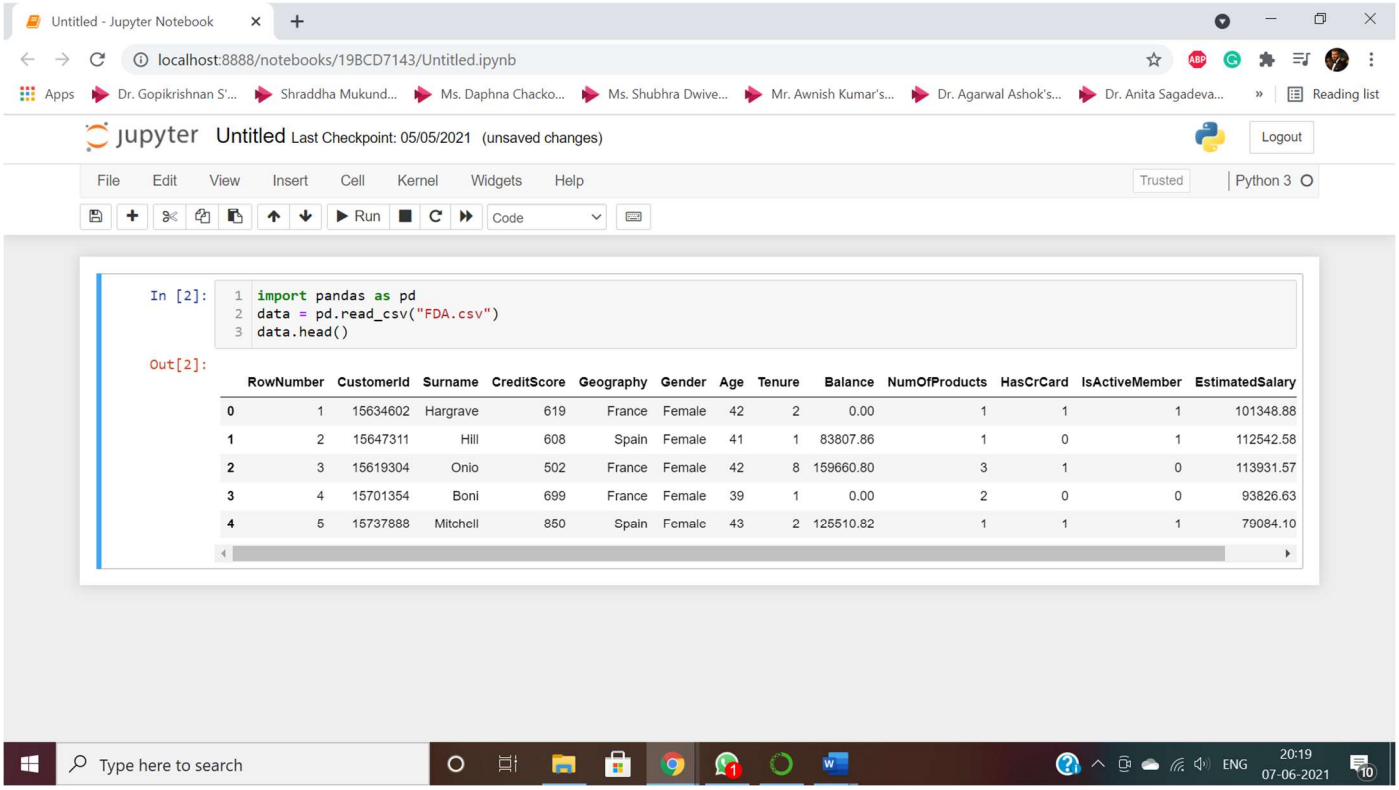
Find the output of the following sections:

i)Write a syntax to load given data

Answer:

```
import pandas as pd
data = pd.read_csv("FDA.csv")
data.head()
```

OUTPUT



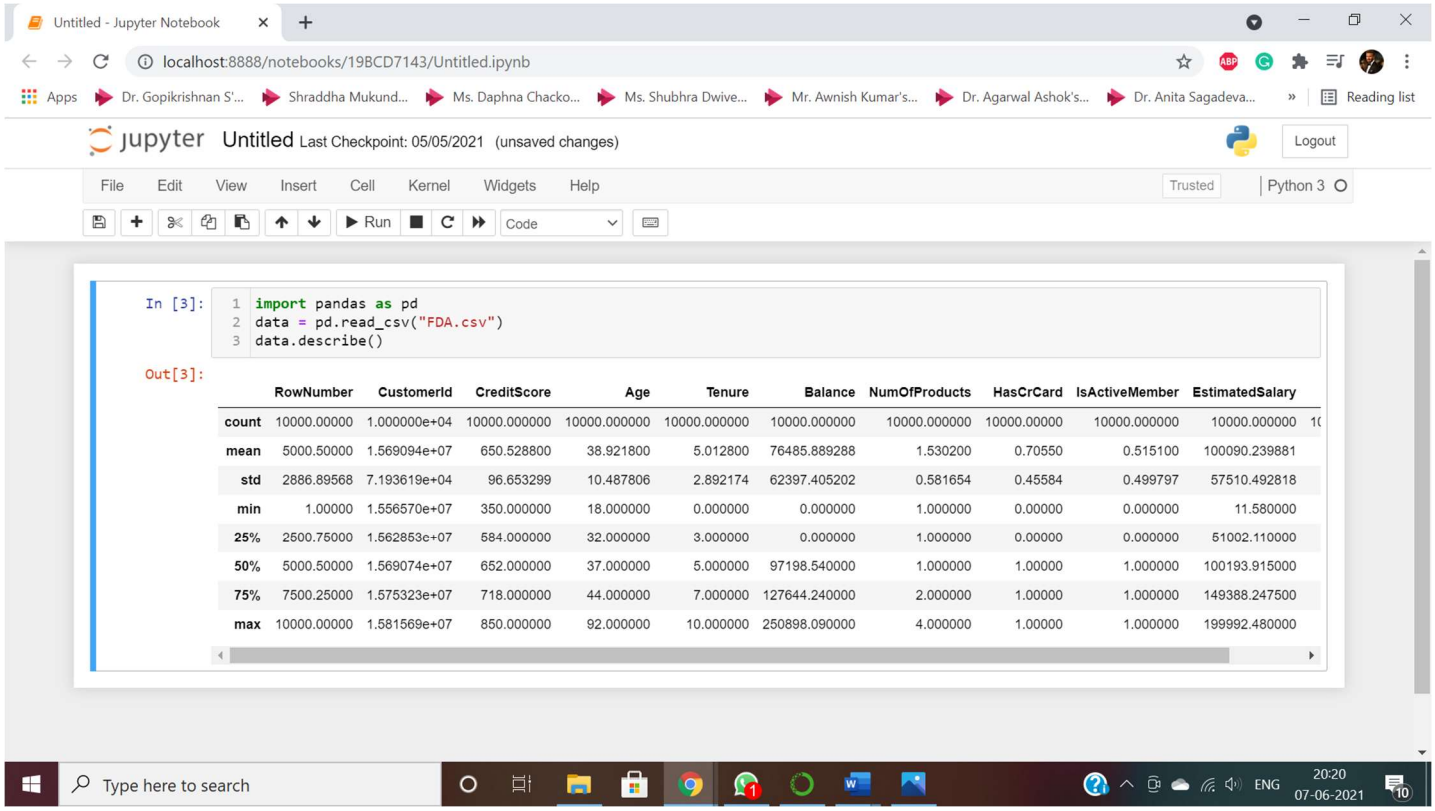
	RowNumber	CustomerId	Surname	CreditScore	Geography	Gender	Age	Tenure	Balance	NumOfProducts	HasCrCard	IsActiveMember	EstimatedSalary
0	1	15634602	Hargrave	619	France	Female	42	2	0.00	1	1	1	101348.88
1	2	15647311	Hill	608	Spain	Female	41	1	83807.86	1	0	1	112542.58
2	3	15619304	Onio	502	France	Female	42	8	159660.80	3	1	0	113931.57
3	4	15701354	Boni	699	France	Female	39	1	0.00	2	0	0	93826.63
4	5	15737888	Mitchell	850	Spain	Female	43	2	125510.82	1	1	1	79084.10

ii) Summarize the data

Answer:

```
import pandas as pd
data = pd.read_csv("FDA.csv")
data.describe()
```

OUTPUT



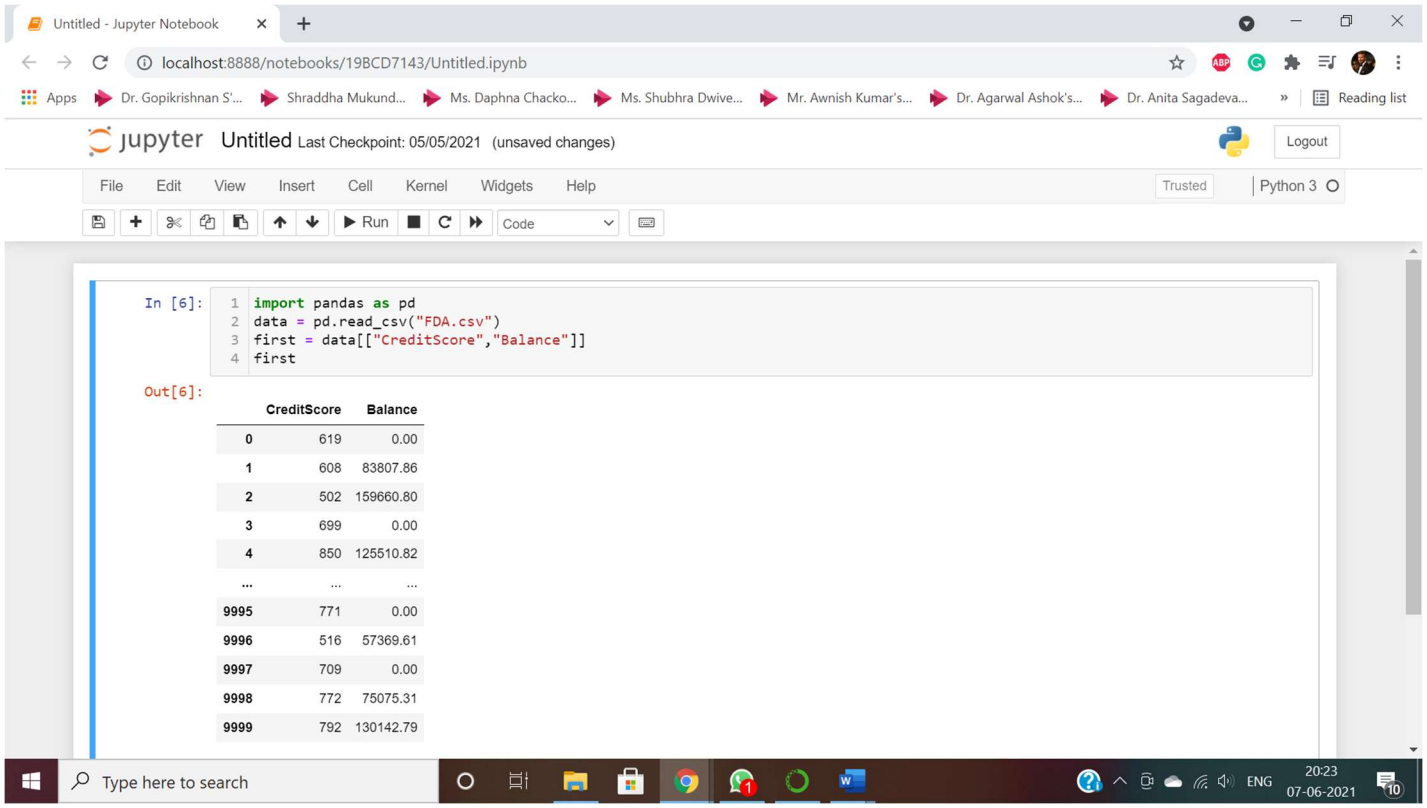
	RowNumber	CustomerId	CreditScore	Age	Tenure	Balance	NumOfProducts	HasCrCard	IsActiveMember	EstimatedSalary
count	10000.00000	1.000000e+04	10000.000000	10000.000000	10000.000000	10000.000000	10000.000000	10000.00000	10000.000000	10000.000000
mean	5000.50000	1.569094e+07	650.528800	38.921800	5.012800	76485.889288	1.530200	0.70550	0.515100	100090.239881
std	2886.89568	7.193619e+04	96.653299	10.487806	2.892174	62397.405202	0.581654	0.45584	0.499797	57510.492818
min	1.00000	1.556570e+07	350.000000	18.000000	0.000000	0.000000	1.000000	0.00000	0.000000	11.580000
25%	2500.75000	1.562853e+07	584.000000	32.000000	3.000000	0.000000	1.000000	0.00000	0.000000	51002.110000
50%	5000.50000	1.569074e+07	652.000000	37.000000	5.000000	97198.540000	1.000000	1.00000	1.000000	100193.915000
75%	7500.25000	1.575323e+07	718.000000	44.000000	7.000000	127644.240000	2.000000	1.00000	1.000000	149388.247500
max	10000.00000	1.581569e+07	850.000000	92.000000	10.000000	250898.090000	4.000000	1.00000	1.000000	199992.480000

iii) To select columns CreditScore and Balance.

Answer:

```
import pandas as pd
data = pd.read_csv("FDA.csv")
first = data[["CreditScore","Balance"]]
first
```

OUTPUT

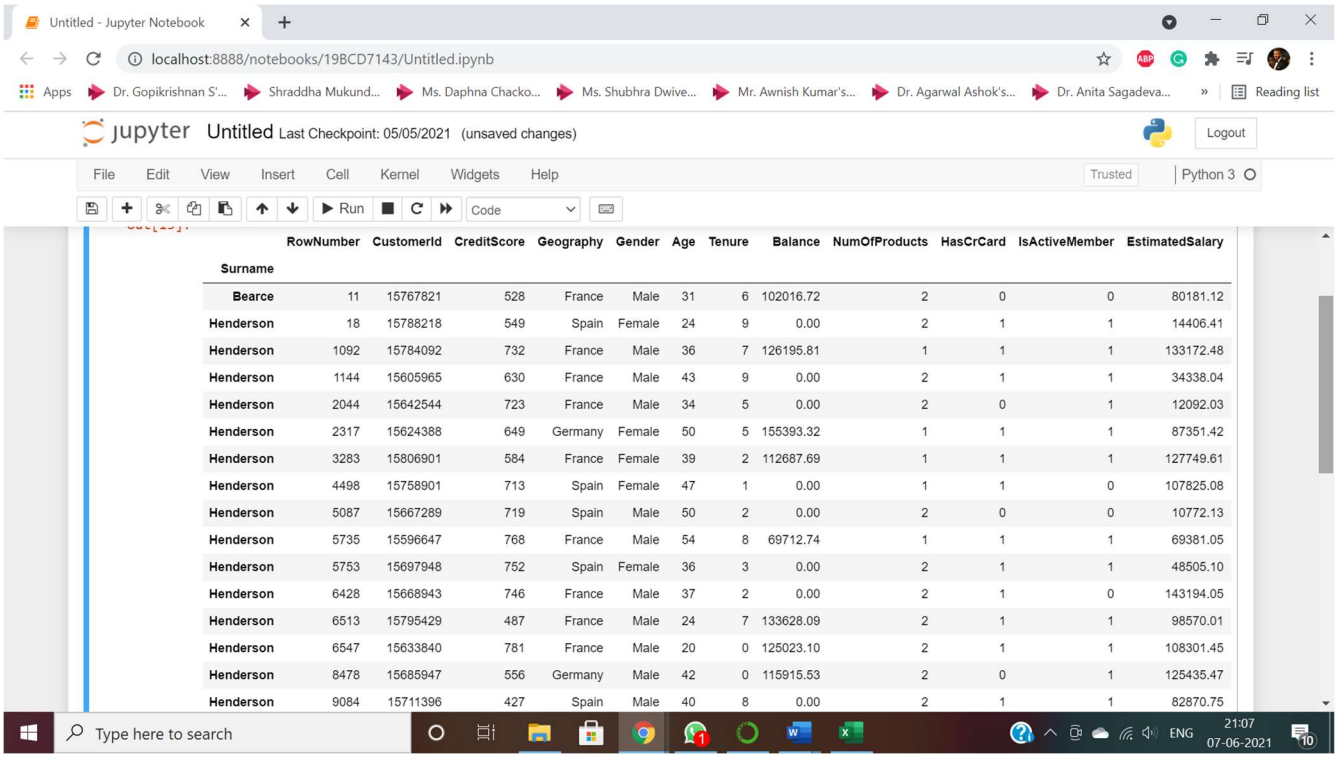


iv)To select row Bearce, Henderson and Lombardo with all columns in a data.

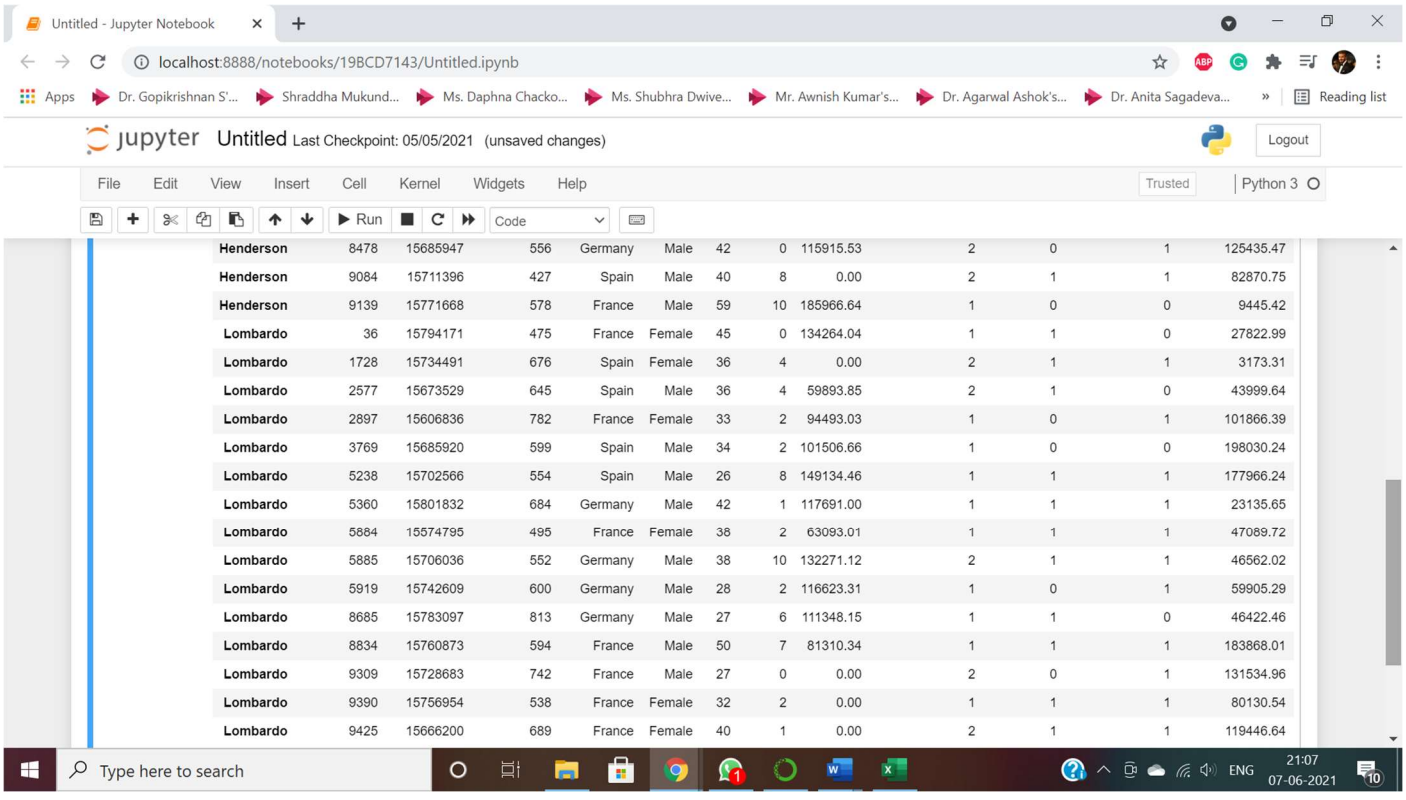
Answer:

```
import pandas as pd
data = pd.read_csv("FDA.csv",index_col ="Surname")
first = data.loc[["Bearce","Henderson","Lombardo"]]
first
```

OUTPUT



Surname	RowNumber	CustomerId	CreditScore	Geography	Gender	Age	Tenure	Balance	NumOfProducts	HasCrCard	IsActiveMember	EstimatedSalary
Bearce	11	15767821	528	France	Male	31	6	102016.72	2	0	0	80181.12
Henderson	18	15788218	549	Spain	Female	24	9	0.00	2	1	1	14406.41
Henderson	1092	15784092	732	France	Male	36	7	126195.81	1	1	1	133172.48
Henderson	1144	15605965	630	France	Male	43	9	0.00	2	1	1	34338.04
Henderson	2044	15642544	723	France	Male	34	5	0.00	2	0	1	12092.03
Henderson	2317	15624388	649	Germany	Female	50	5	155393.32	1	1	1	87351.42
Henderson	3283	15806901	584	France	Female	39	2	112687.69	1	1	1	127749.61
Henderson	4498	15758901	713	Spain	Female	47	1	0.00	1	1	0	107825.08
Henderson	5087	15667289	719	Spain	Male	50	2	0.00	2	0	0	10772.13
Henderson	5735	15596647	768	France	Male	54	8	69712.74	1	1	1	69381.05
Henderson	5753	15697948	752	Spain	Female	36	3	0.00	2	1	1	48505.10
Henderson	6428	15668943	746	France	Male	37	2	0.00	2	1	0	143194.05
Henderson	6513	15795429	487	France	Male	24	7	133628.09	2	1	1	98570.01
Henderson	6547	15633840	781	France	Male	20	0	125023.10	2	1	1	108301.45
Henderson	8478	15685947	556	Germany	Male	42	0	115915.53	2	0	1	125435.47
Henderson	9084	15711396	427	Spain	Male	40	8	0.00	2	1	1	82870.75



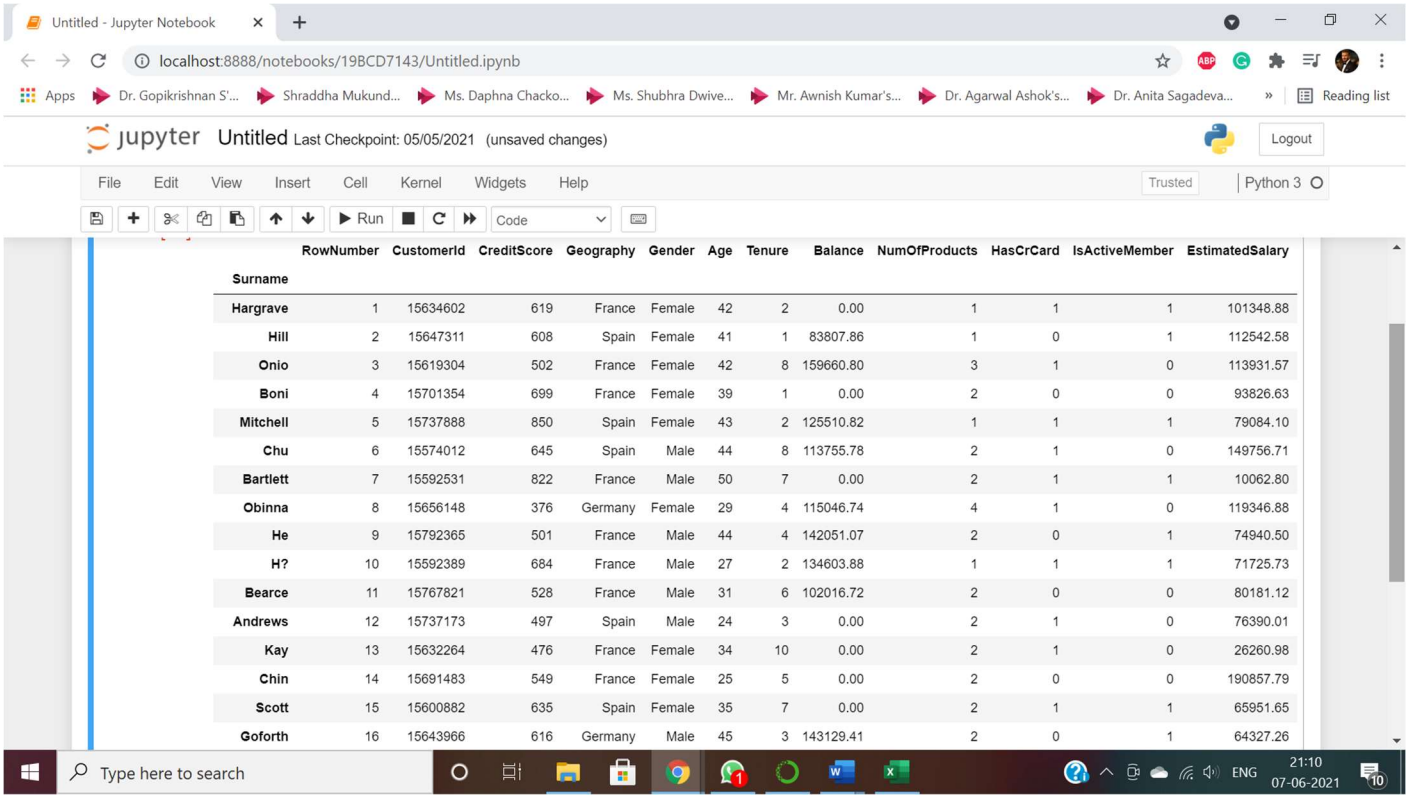
Henderson	8478	15685947	556	Germany	Male	42	0	115915.53	2	0	1	125435.47
Henderson	9084	15711396	427	Spain	Male	40	8	0.00	2	1	1	82870.75
Henderson	9139	15771668	578	France	Male	59	10	185966.64	1	0	0	9445.42
Lombardo	36	15794171	475	France	Female	45	0	134264.04	1	1	0	27822.99
Lombardo	1728	15734491	676	Spain	Female	36	4	0.00	2	1	1	3173.31
Lombardo	2577	15673529	645	Spain	Male	36	4	59893.85	2	1	0	43999.64
Lombardo	2897	15606836	782	France	Female	33	2	94493.03	1	0	1	101866.39
Lombardo	3769	15685920	599	Spain	Male	34	2	101506.66	1	0	0	198030.24
Lombardo	5238	15702566	554	Spain	Male	26	8	149134.46	1	1	1	177966.24
Lombardo	5360	15801832	684	Germany	Male	42	1	117691.00	1	1	1	23135.65
Lombardo	5884	15574795	495	France	Female	38	2	63093.01	1	1	1	47089.72
Lombardo	5885	15706036	552	Germany	Male	38	10	132271.12	2	1	1	46562.02
Lombardo	5919	15742609	600	Germany	Male	28	2	116623.31	1	0	1	59905.29
Lombardo	8685	15783097	813	Germany	Male	27	6	111348.15	1	1	0	46422.46
Lombardo	8834	15760873	594	France	Male	50	7	81310.34	1	1	1	183868.01
Lombardo	9309	15728683	742	France	Male	27	0	0.00	2	0	1	131534.96
Lombardo	9390	15756954	538	France	Female	32	2	0.00	1	1	1	80130.54
Lombardo	9425	15666200	689	France	Female	40	1	0.00	2	1	1	119446.64

v)print the first 20 rows of data

Answer:

```
import pandas as pd
data = pd.read_csv("FDA.csv",index_col ="Surname")
data.head(20)
```

OUTPUT



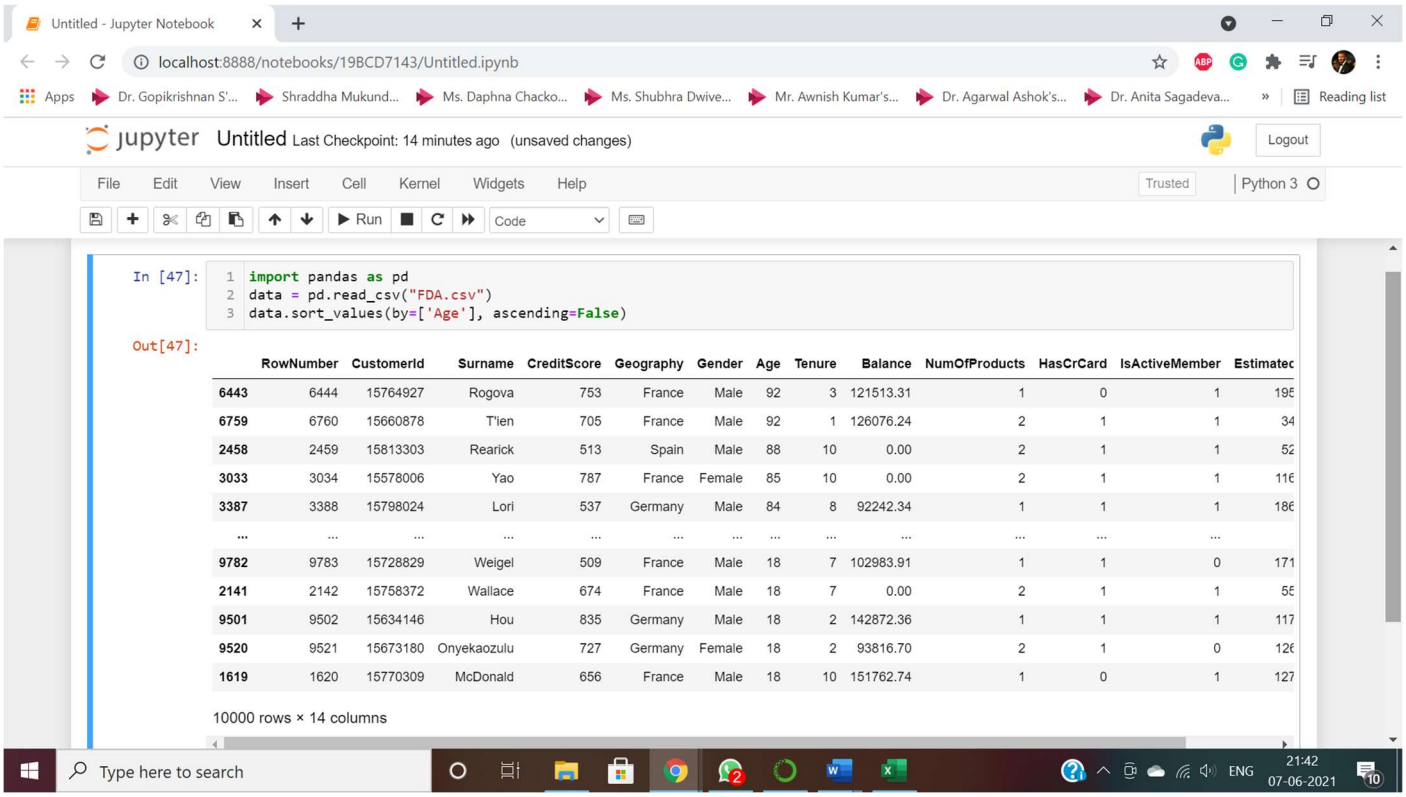
Surname	RowNumber	CustomerId	CreditScore	Geography	Gender	Age	Tenure	Balance	NumOfProducts	HasCrCard	IsActiveMember	EstimatedSalary
Hargrave	1	15634602	619	France	Female	42	2	0.00	1	1	1	101348.88
Hill	2	15647311	608	Spain	Female	41	1	83807.86	1	0	1	112542.58
Onio	3	15619304	502	France	Female	42	8	159660.80	3	1	0	113931.57
Boni	4	15701354	699	France	Female	39	1	0.00	2	0	0	93826.63
Mitchell	5	15737888	850	Spain	Female	43	2	125510.82	1	1	1	79084.10
Chu	6	15574012	645	Spain	Male	44	8	113755.78	2	1	0	149756.71
Bartlett	7	15592531	822	France	Male	50	7	0.00	2	1	1	10062.80
Obinna	8	15656148	376	Germany	Female	29	4	115046.74	4	1	0	119346.88
He	9	15792365	501	France	Male	44	4	142051.07	2	0	1	74940.50
H?	10	15592389	684	France	Male	27	2	134603.88	1	1	1	71725.73
Bearce	11	15767821	528	France	Male	31	6	102016.72	2	0	0	80181.12
Andrews	12	15737173	497	Spain	Male	24	3	0.00	2	1	0	76390.01
Kay	13	15632264	476	France	Female	34	10	0.00	2	1	0	26260.98
Chin	14	15691483	549	France	Female	25	5	0.00	2	0	0	190857.79
Scott	15	15600882	635	Spain	Female	35	7	0.00	2	1	1	65951.65
Goforth	16	15643966	616	Germany	Male	45	3	143129.41	2	0	1	64327.26

vi)Sort the data in descending order of Age.

Answer:

```
import pandas as pd

data = pd.read_csv("FDA.csv")

data.sort_values(by=['Age'], ascending=False)
```

OUTPUT



	RowNumber	CustomerId	Surname	CreditScore	Geography	Gender	Age	Tenure	Balance	NumOfProducts	HasCrCard	IsActiveMember	Estimatec
6443	6444	15764927	Rogova	753	France	Male	92	3	121513.31	1	0	1	195
6759	6760	15660878	T'ien	705	France	Male	92	1	126076.24	2	1	1	34
2458	2459	15813303	Rearick	513	Spain	Male	88	10	0.00	2	1	1	52
3033	3034	15578006	Yao	787	France	Female	85	10	0.00	2	1	1	116
3387	3388	15798024	Lori	537	Germany	Male	84	8	92242.34	1	1	1	186
...	...	...	...	...	...	...	...	...	...	...	...	...	
9782	9783	15728829	Weigel	509	France	Male	18	7	102983.91	1	1	0	171
2141	2142	15758372	Wallace	674	France	Male	18	7	0.00	2	1	1	55
9501	9502	15634146	Hou	835	Germany	Male	18	2	142872.36	1	1	1	117
9520	9521	15673180	Onyekaozulu	727	Germany	Female	18	2	93816.70	2	1	0	126
1619	1620	15770309	McDonald	656	France	Male	18	10	151762.74	1	0	1	127

10000 rows × 14 columns

2.

a)Draw the number of customers in each country in given data based on bar plot.

b) In the bar plot adjust x limit and y limit in the range of 0 to 7000 by gap 1000.

c) Draw a notched box plot of the CreditScore.

Answer:

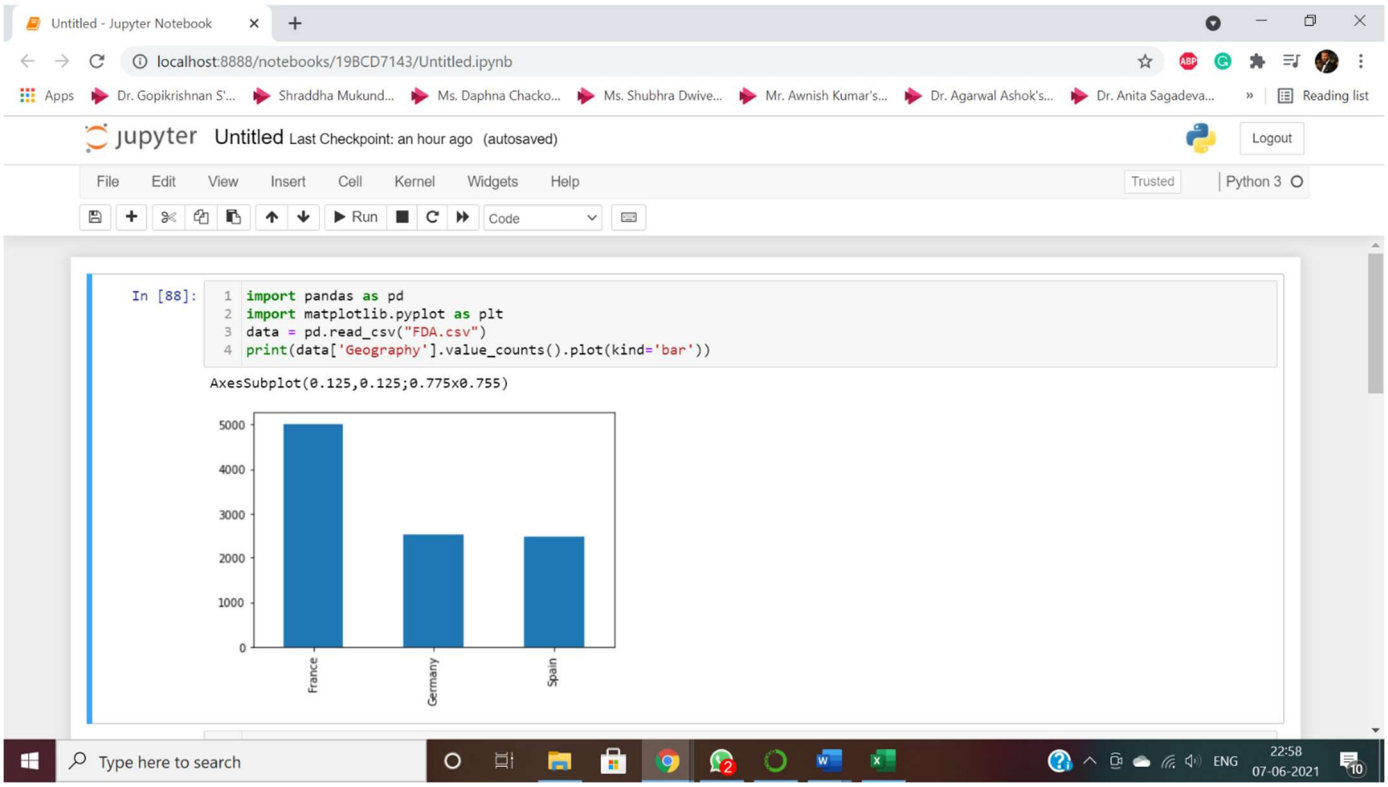
a) import pandas as pd

import matplotlib.pyplot as plt

data = pd.read_csv("FDA.csv")

print(data['Geography'].value_counts().plot(kind='bar'))

OUTPUT



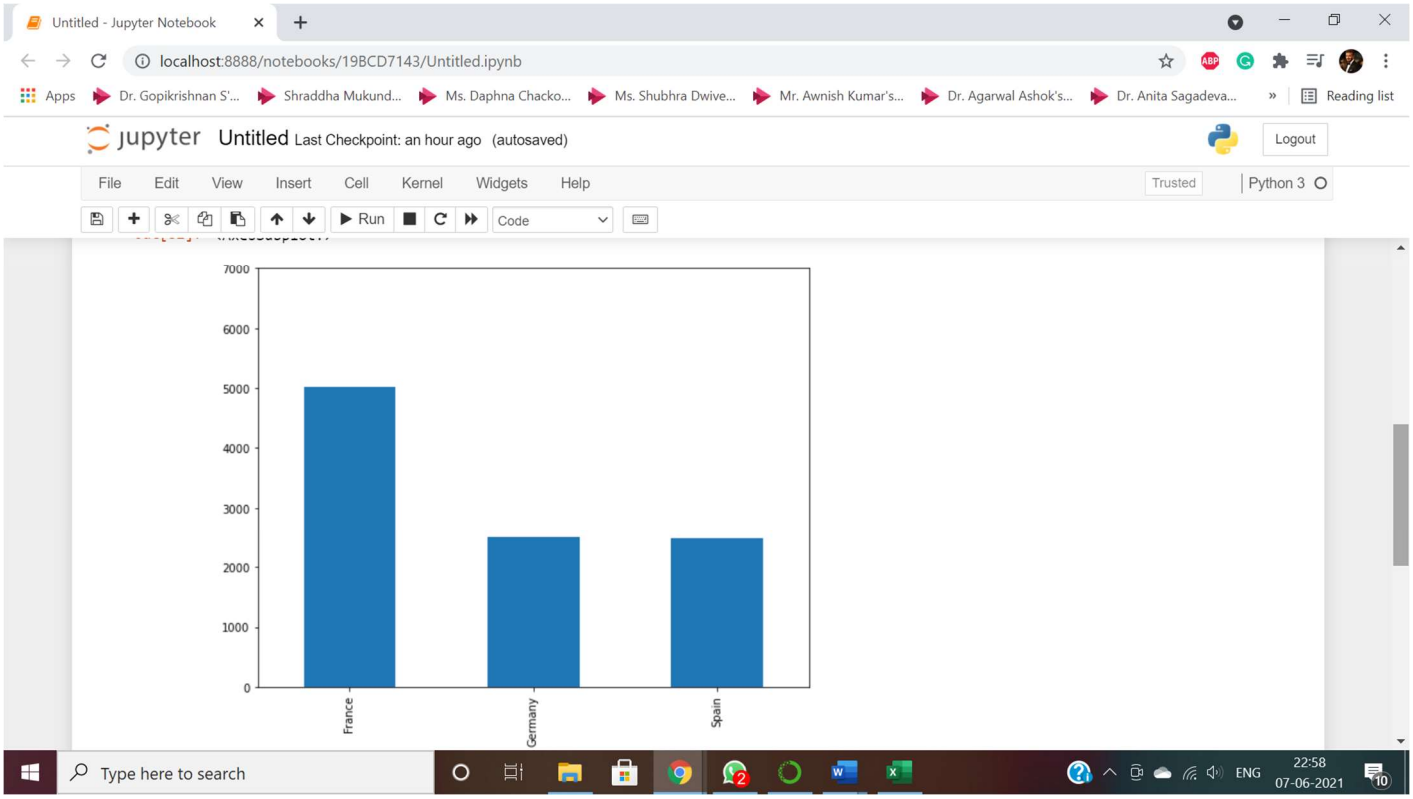
b) import pandas as pd

import matplotlib.pyplot as plt

data = pd.read_csv("FDA.csv")

data['Geography'].value_counts().plot(kind='bar',ylim=(0,7000),figsize=(9,7))

OUTPUT



c) import pandas as pd

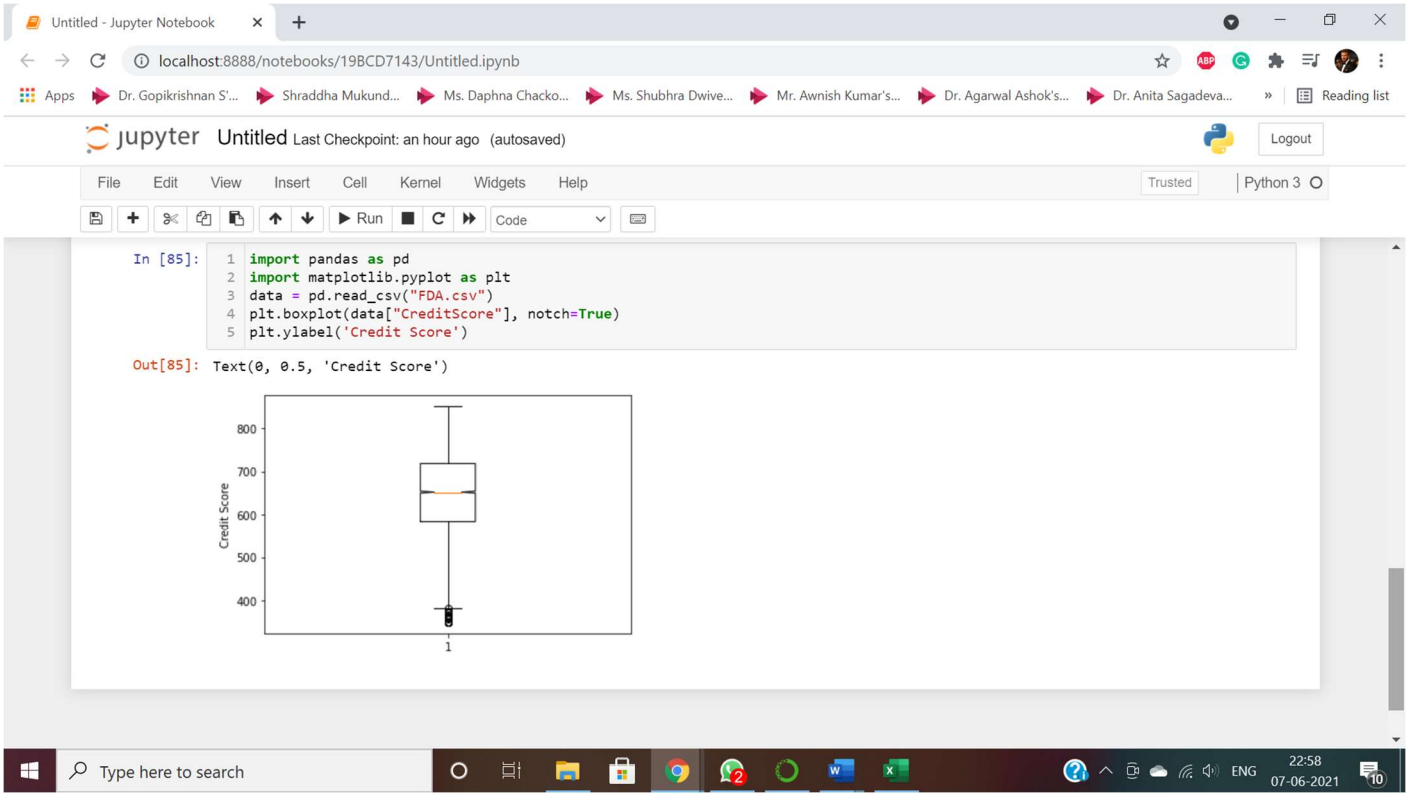
import matplotlib.pyplot as plt

data = pd.read_csv("FDA.csv")

plt.boxplot(data["CreditScore"], notch=True)

plt.ylabel('Credit Score')

OUTPUT



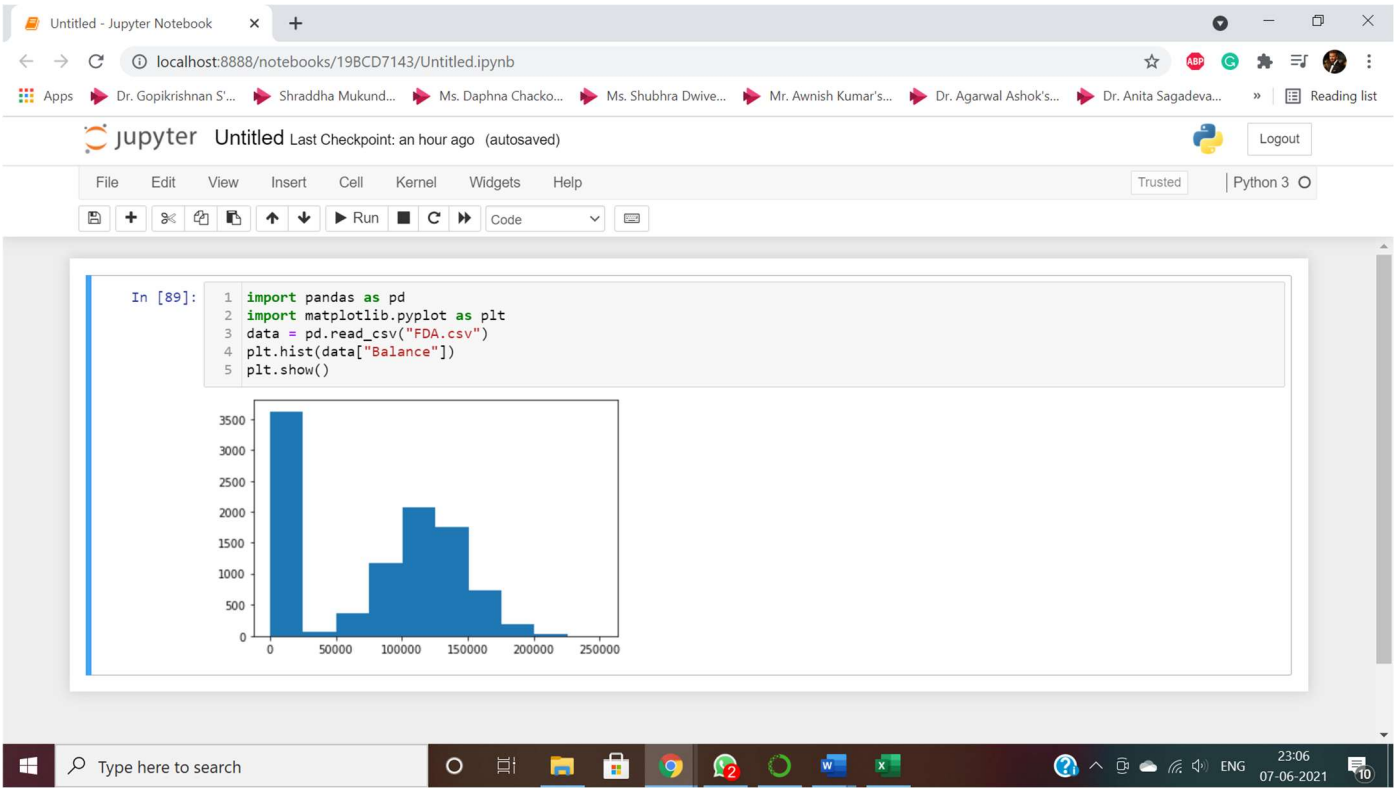
3.

a)Draw the histogram of customer Balance attribute in given data.

Answer:

```
import pandas as pd
import matplotlib.pyplot as plt
data = pd.read_csv("FDA.csv")
plt.hist(data["Balance"])
plt.show()
```

OUTPUT

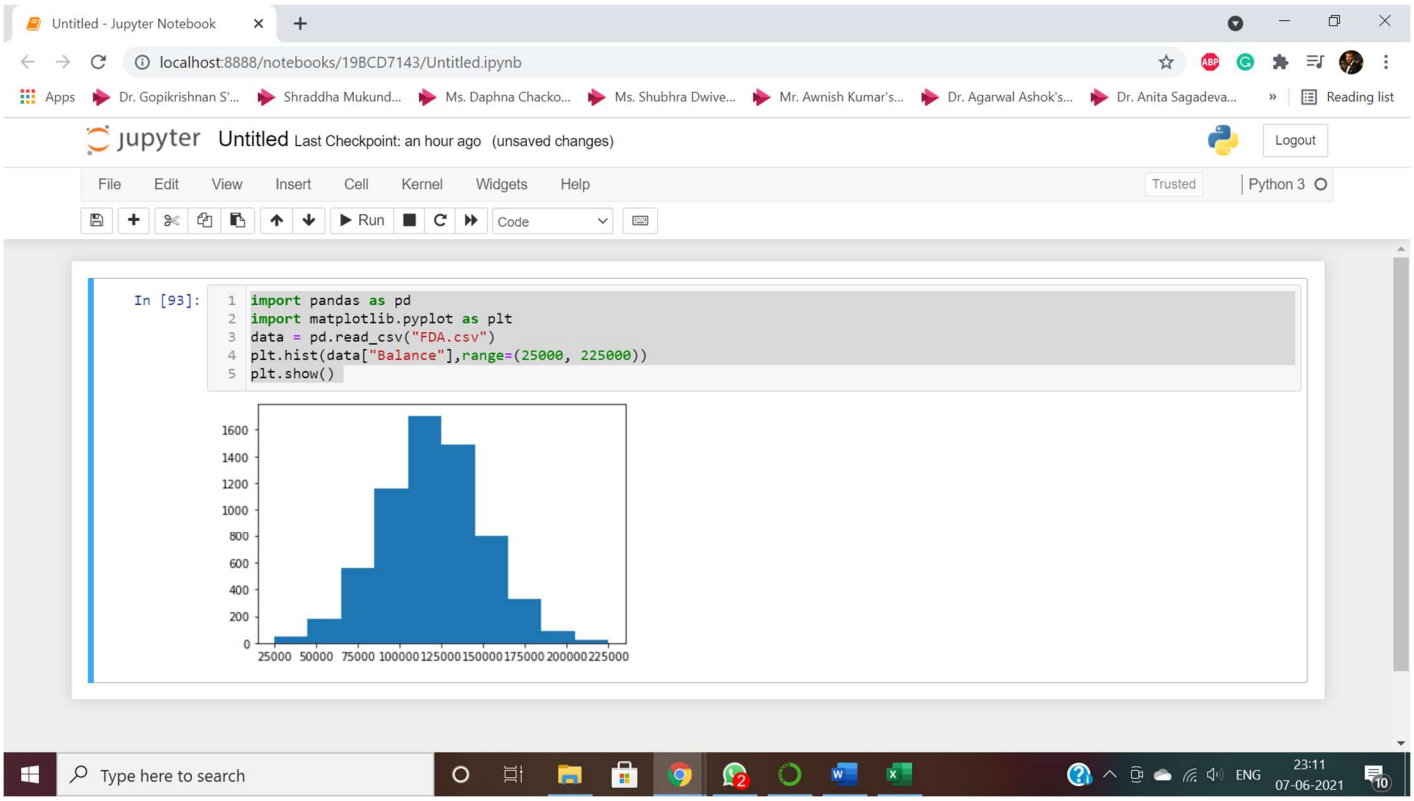


b)Draw the Customizing histogram of customer on range = (25000, 225000) by Balance attribute in given data.

Answer:

```
import pandas as pd

import matplotlib.pyplot as plt

data = pd.read_csv("FDA.csv")

plt.hist(data["Balance"],range=(25000, 225000))

plt.show()
```

OUTPUT



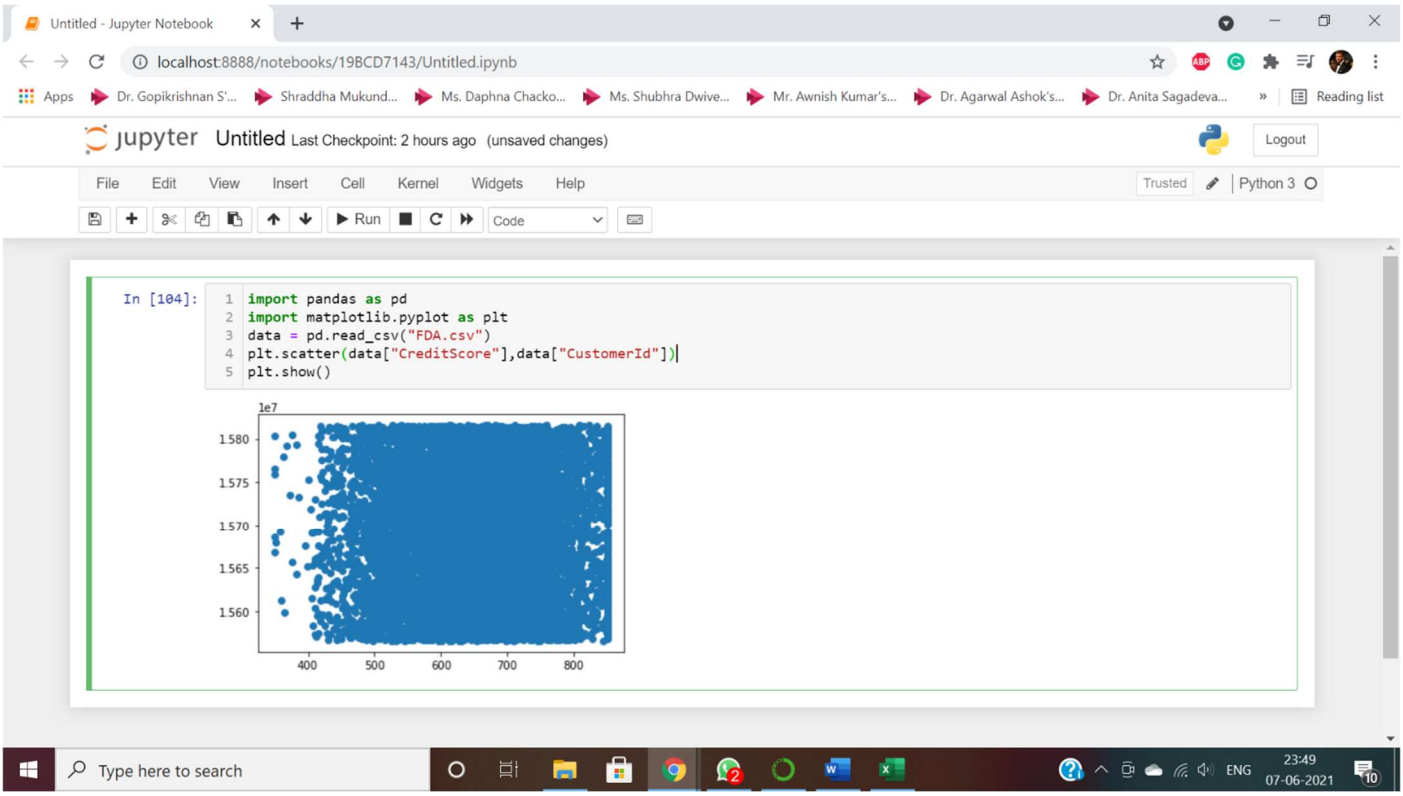
4.

a)Draw scatter plot of customer CreditScore attribute in given data.

Answer:

```
import pandas as pd
import matplotlib.pyplot as plt
data = pd.read_csv("FDA.csv")
plt.scatter(data["CreditScore"],data["CustomerId"])
plt.show()
```

OUTPUT

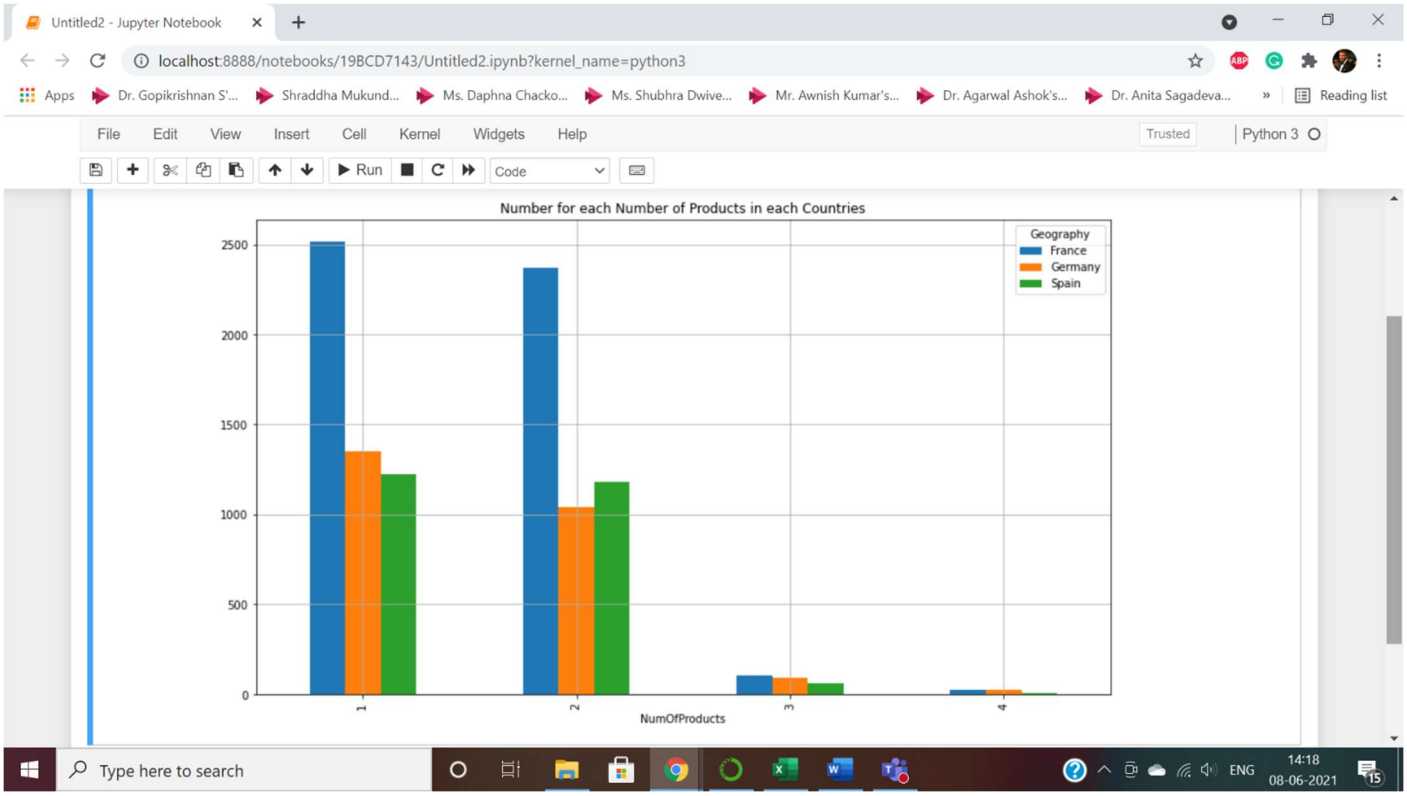


b)Draw grid of 2 columns and add bar plots attribute Countries and Number of Products.

Answer:

```
import pandas as pd

import matplotlib.pyplot as plt

data = pd.read_csv("FDA.csv")


ax =
data.groupby(["Geography","NumOfProducts"])["NumOfProducts"].count().unstack(0).plot.bar(title="Number for each Number of Products in each Countries", figsize=(14,8))

plt.grid()
```

OUTPUT

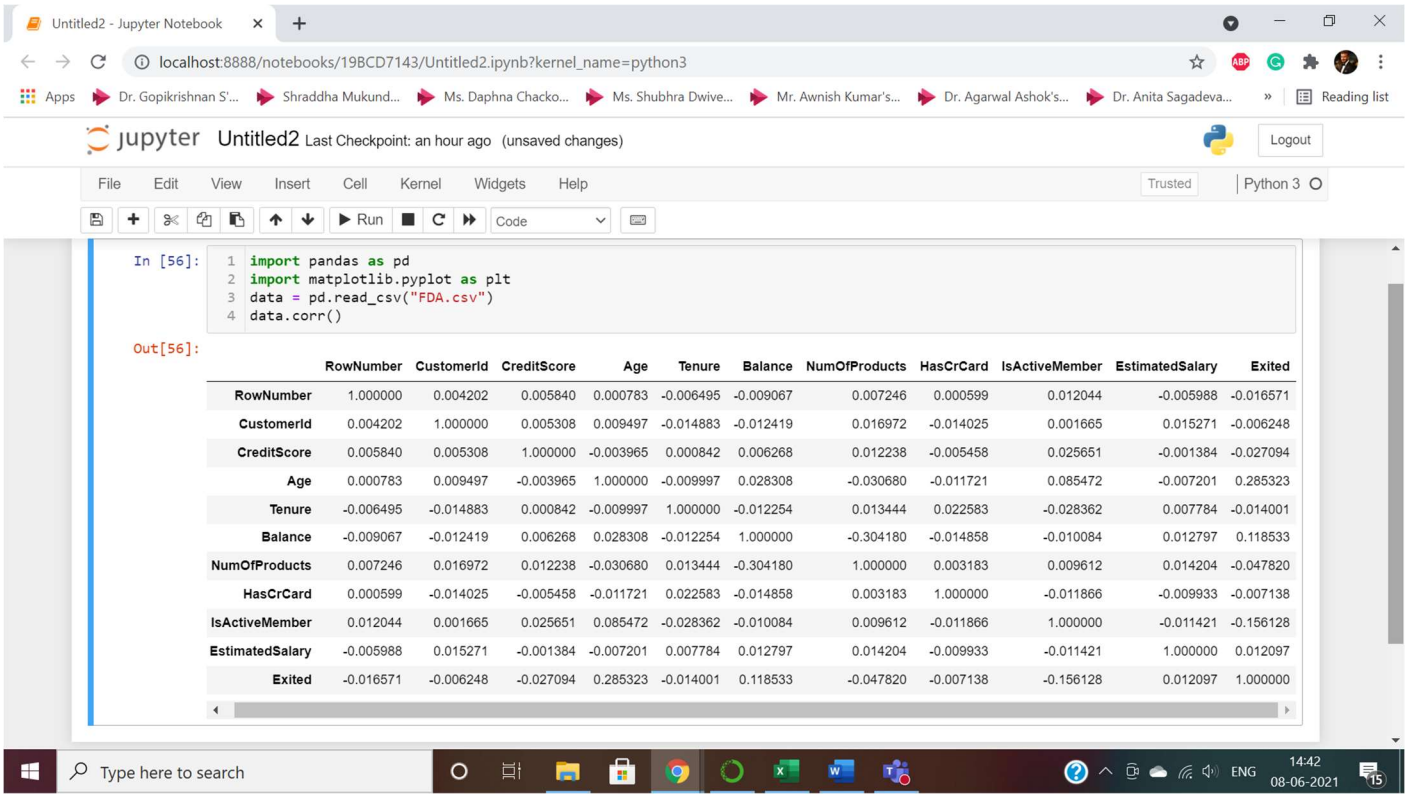


c)Draw Correlation Matrices.

Answer:

```
import pandas as pd
import matplotlib.pyplot as plt
data = pd.read_csv("FDA.csv")
data.corr()
```

OUTPUT



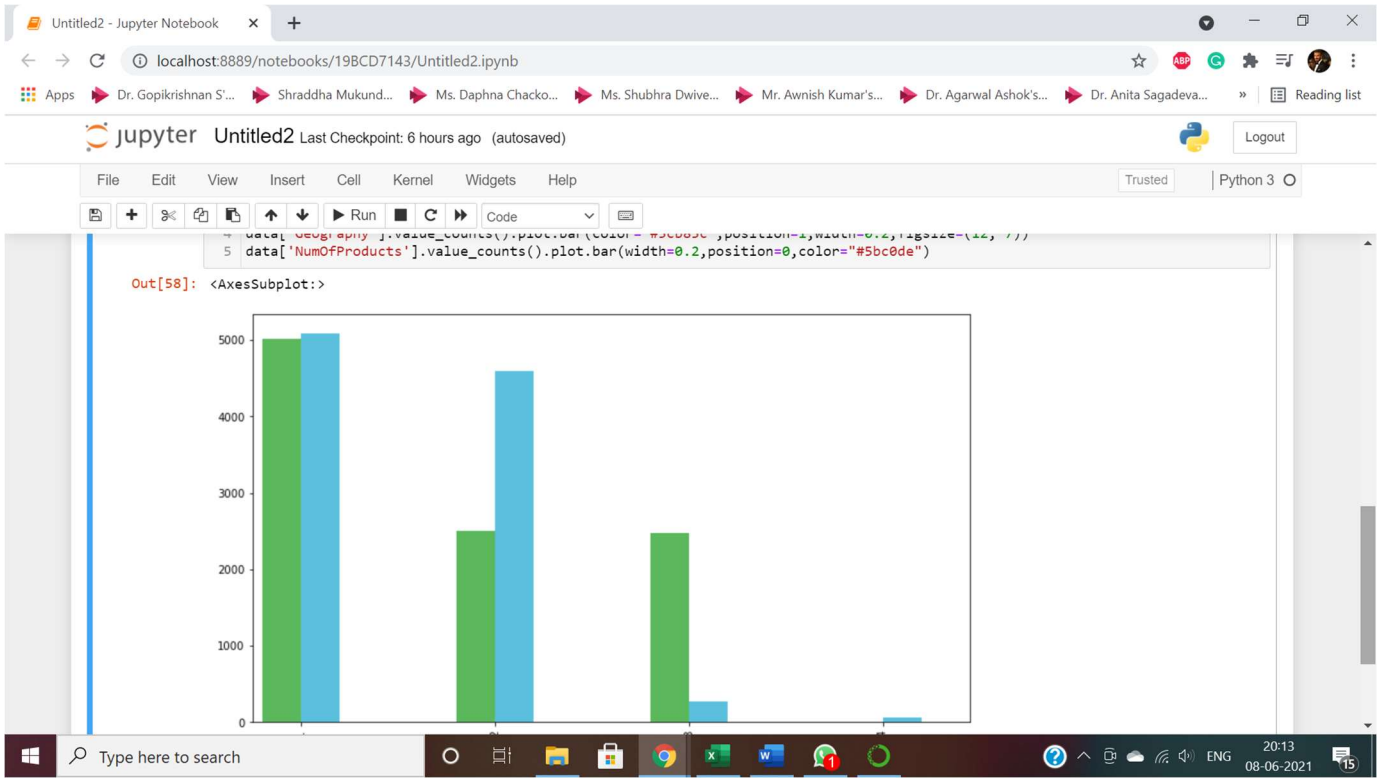
	RowNumber	CustomerId	CreditScore	Age	Tenure	Balance	NumOfProducts	HasCrCard	IsActiveMember	EstimatedSalary	Exited
RowNumber	1.000000	0.004202	0.005840	0.000783	-0.006495	-0.009067	0.007246	0.000599	0.012044	-0.005988	-0.016571
CustomerId	0.004202	1.000000	0.005308	0.009497	-0.014883	-0.012419	0.016972	-0.014025	0.001665	0.015271	-0.006248
CreditScore	0.005840	0.005308	1.000000	-0.003965	0.000842	0.006268	0.012238	-0.005458	0.025651	-0.001384	-0.027094
Age	0.000783	0.009497	-0.003965	1.000000	-0.009997	0.028308	-0.030680	-0.011721	0.085472	-0.007201	0.285323
Tenure	-0.006495	-0.014883	0.000842	-0.009997	1.000000	-0.012254	0.013444	0.022583	-0.028362	0.007784	-0.014001
Balance	-0.009067	-0.012419	0.006268	0.028308	-0.012254	1.000000	-0.304180	-0.014858	-0.010084	0.012797	0.118533
NumOfProducts	0.007246	0.016972	0.012238	-0.030680	0.013444	-0.304180	1.000000	0.003183	0.009612	0.014204	-0.047820
HasCrCard	0.000599	-0.014025	-0.005458	-0.011721	0.022583	-0.014858	0.003183	1.000000	-0.011866	-0.009933	-0.007138
IsActiveMember	0.012044	0.001665	0.025651	0.085472	-0.028362	-0.010084	0.009612	-0.011866	1.000000	-0.011421	-0.156128
EstimatedSalary	-0.005988	0.015271	-0.001384	-0.007201	0.007784	0.012797	0.014204	-0.009933	-0.011421	1.000000	0.012097
Exited	-0.016571	-0.006248	-0.027094	0.285323	-0.014001	0.118533	-0.047820	-0.007138	-0.156128	0.012097	1.000000

d)Use color #5cb85c for the Countries bar plot, color #5bc0de for the Number of Products bar plot.

Answer:

```
import pandas as pd

import matplotlib.pyplot as plt

data = pd.read_csv("FDA.csv")

data['Geography'].value_counts().plot.bar(color="#5cb85c",position=1,width=0.2,figsize=(12, 7))

data['NumOfProducts'].value_counts().plot.bar(width=0.2,position=0,color="#5bc0de")
```

OUTPUT



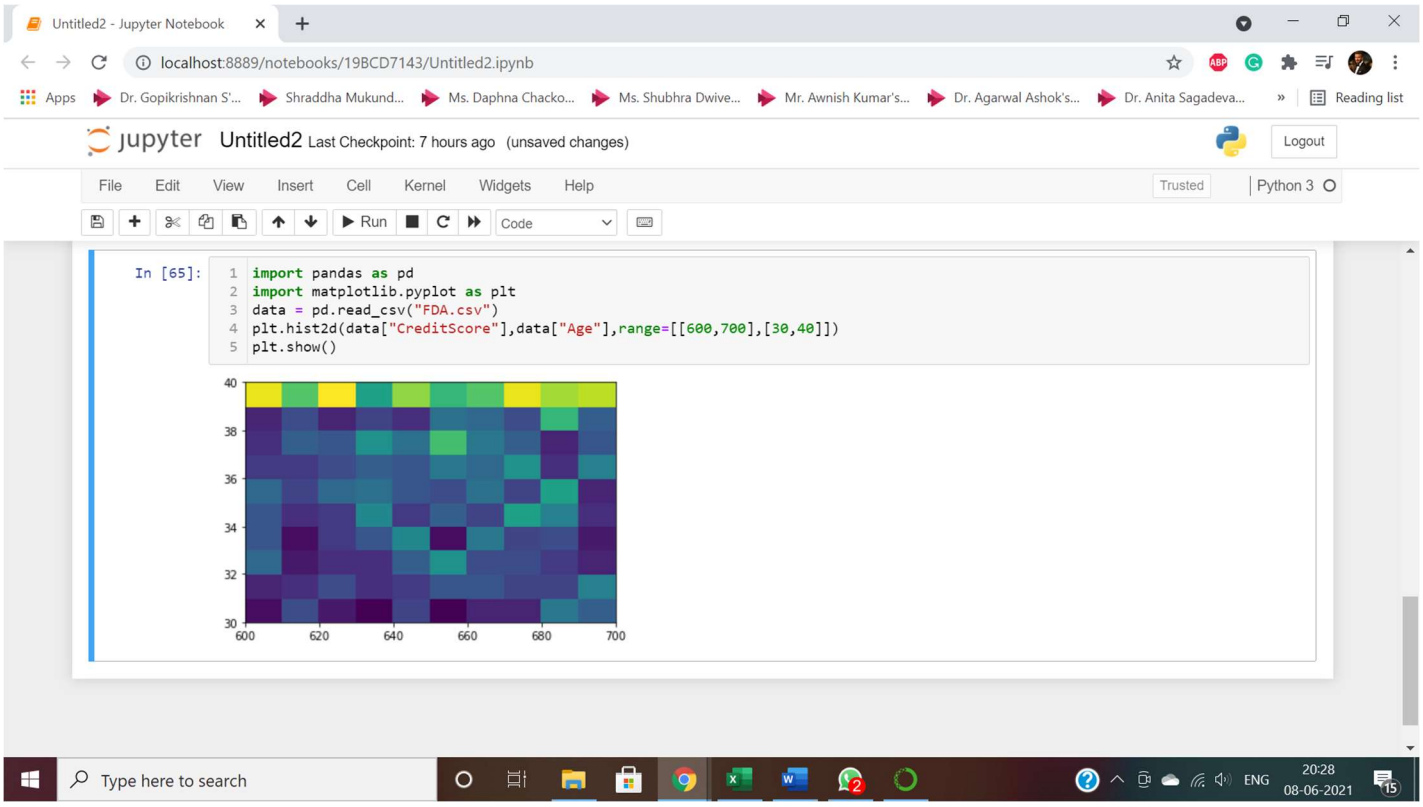
5.

a)Draw the Customizing 2 D histogram plot of CreditScore and Age attribute in given data based on between ages 30 and 40 and have credit scores between 600 and 700.

Answer:

```
import pandas as pd

import matplotlib.pyplot as plt

data = pd.read_csv("FDA.csv")

plt.hist2d(data["CreditScore"],data["Age"],range=[[600,700],[30,40]])

plt.show()
```

OUTPUT

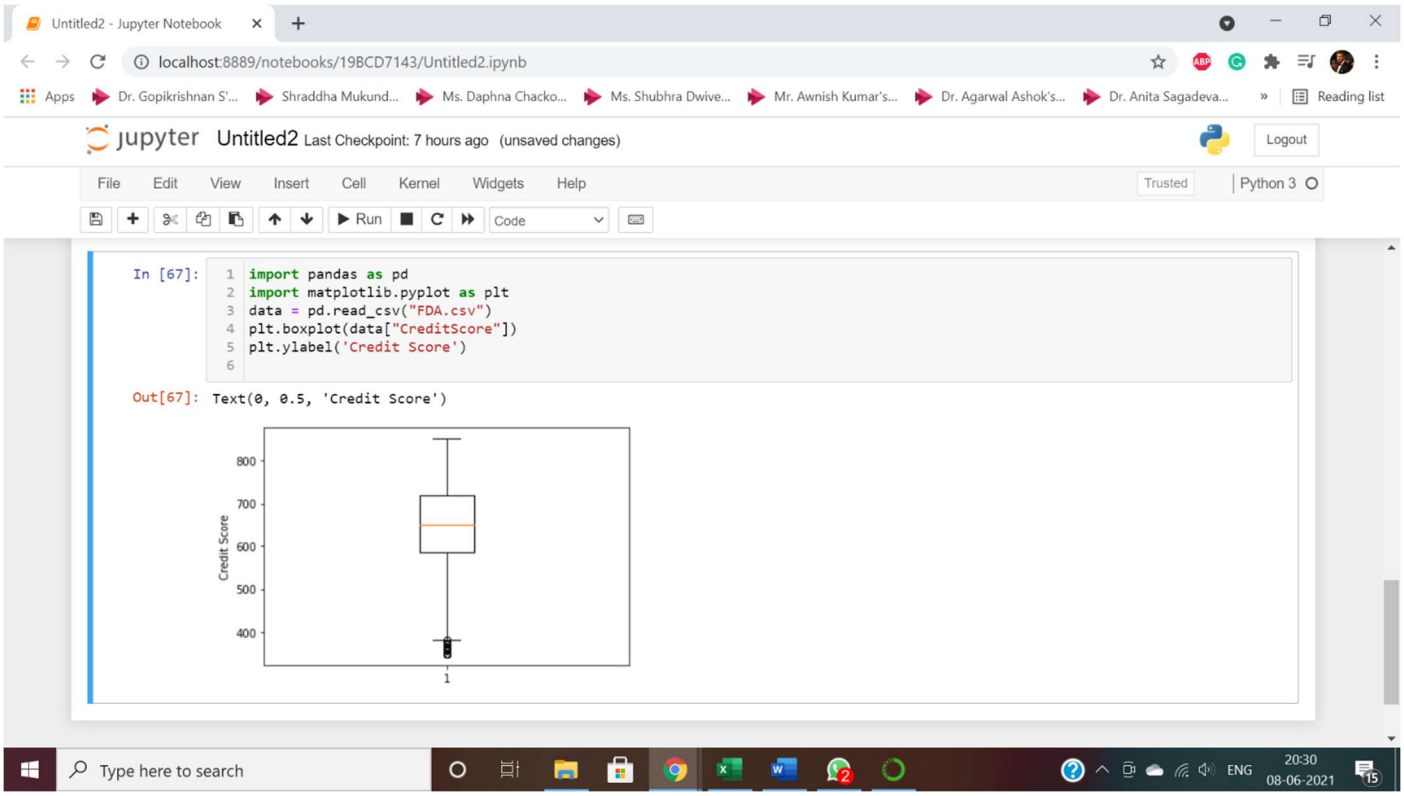


b)Draw the box plot of CreditScore attribute in given dataset.

Answer:

```
import pandas as pd
import matplotlib.pyplot as plt
data = pd.read_csv("FDA.csv")
plt.boxplot(data["CreditScore"])
plt.ylabel('Credit Score')
```

OUTPUT



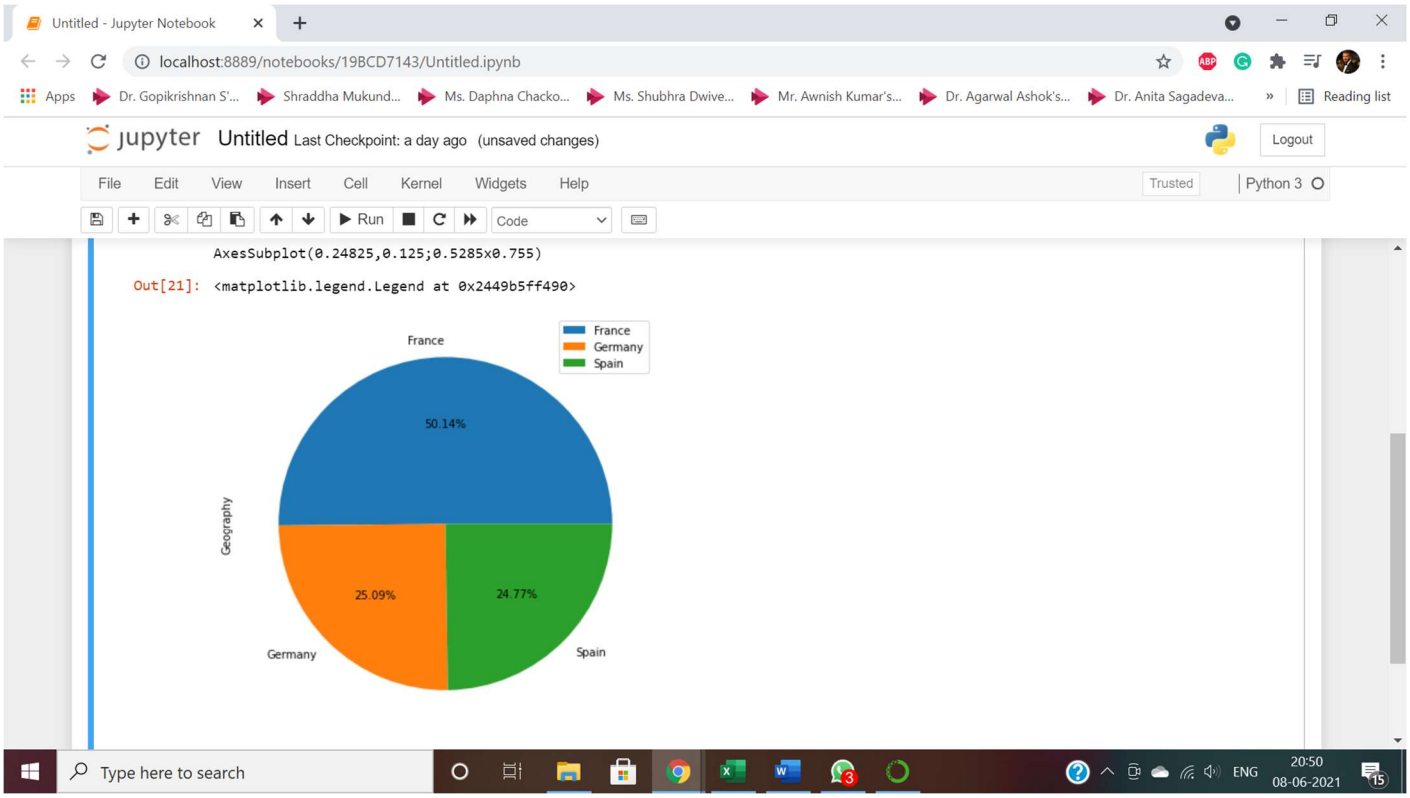
c)Draw Pie Chart with Slice Percentages and Chart Legend of Geography attribute.

Answer:

```
import pandas as pd

import matplotlib.pyplot as plt

import seaborn as sns

data = pd.read_csv("FDA.csv")


print(data['Geography'].value_counts().plot(kind='pie',autopct='%1.2f%%',figsize =(10, 7)))
```

plt.legend()

OUTPUT



6. In this question 6, we will use attached dataset as iris.csv. Find the output of the following sections:

i)Write a syntax to load given data

Answer:

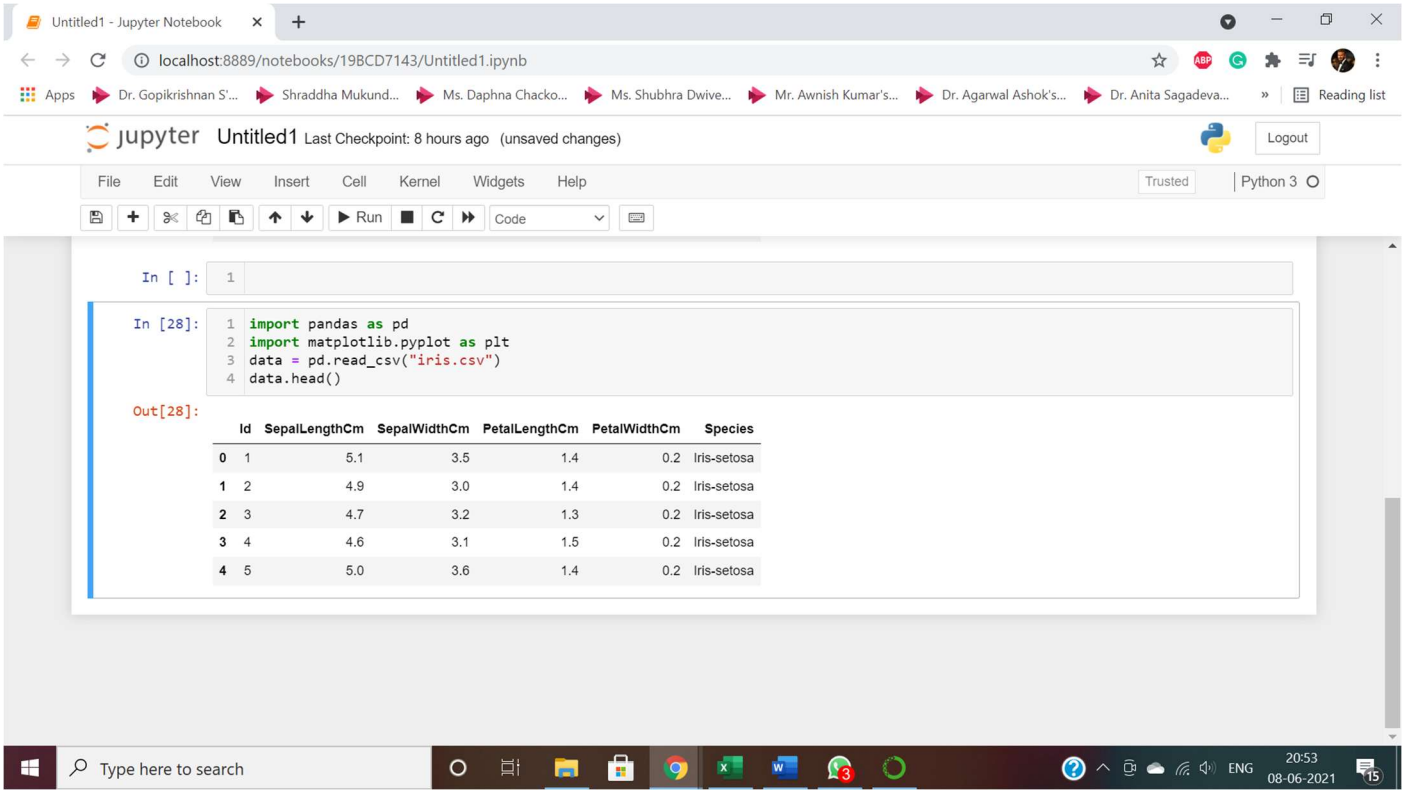
import pandas as pd

import matplotlib.pyplot as plt

data = pd.read_csv("iris.csv")

data.head()

OUTPUT

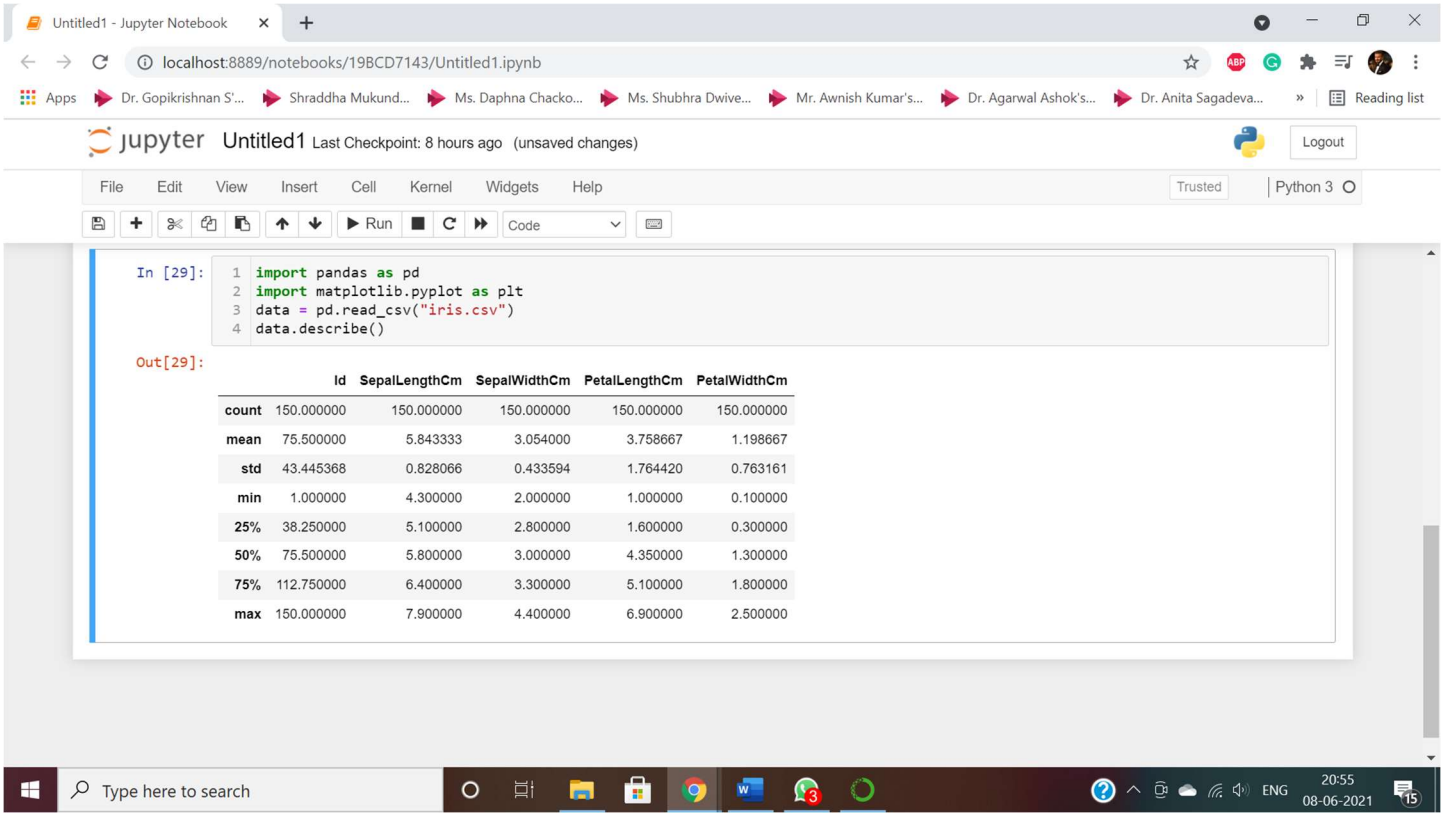


ii)Summarize the data

Answer:

```
import pandas as pd
data = pd.read_csv("iris.csv")
data.describe()
```

OUTPUT

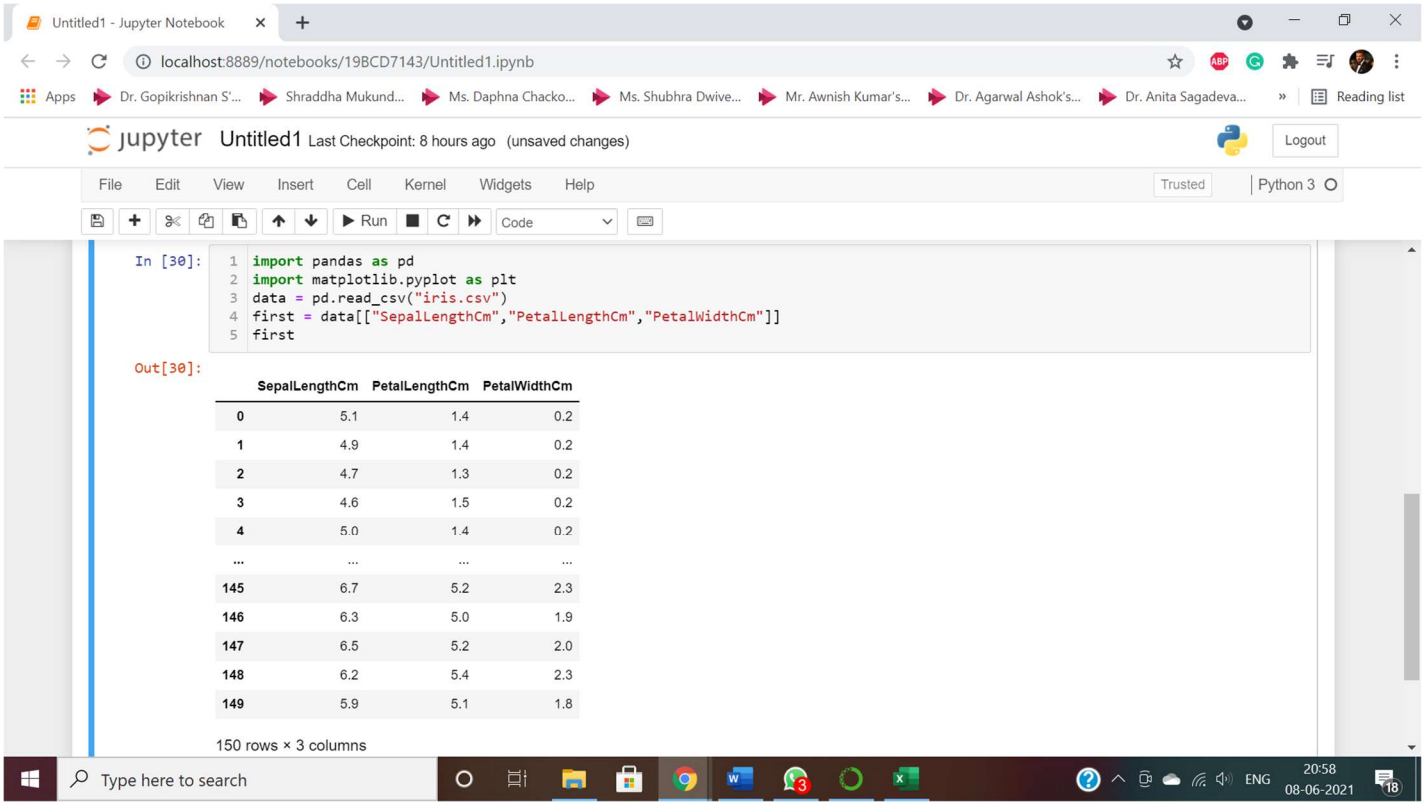


iii)To select columns 'sepal_length', 'petal_length', 'petal_width.

Answer:

```
import pandas as pd
import matplotlib.pyplot as plt
data = pd.read_csv("iris.csv")
first = data[["SepalLengthCm","PetalLengthCm","PetalWidthCm"]]
first
```

OUTPUT

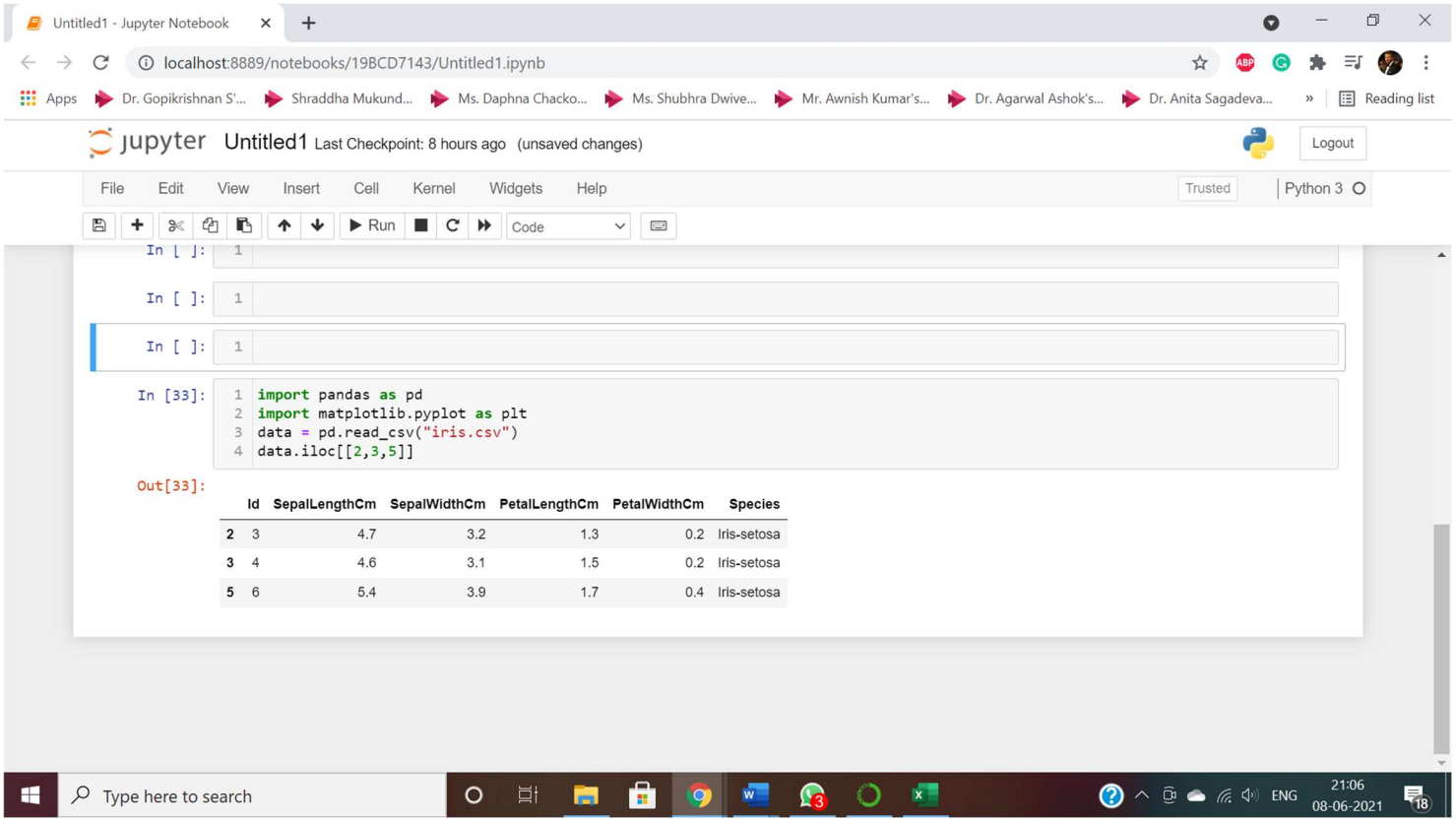


iv)To select row 2, 3 and 5 with all columns in a data.

Answer:

```
import pandas as pd
import matplotlib.pyplot as plt
data = pd.read_csv("iris.csv")
data.iloc[[2,3,5]]
```

OUTPUT

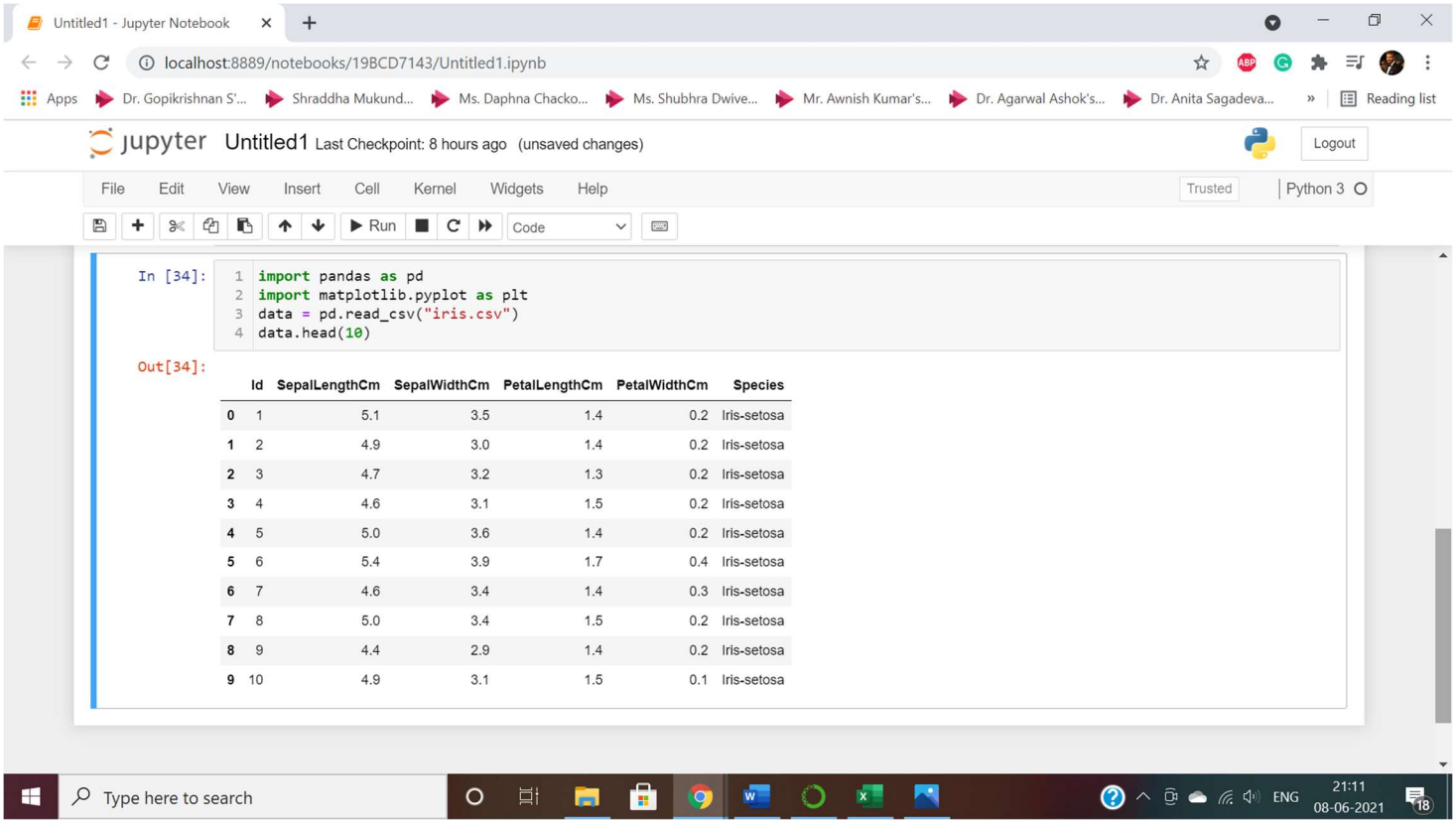


v)print the first 10 rows of data

Answer:

```
import pandas as pd
import matplotlib.pyplot as plt
data = pd.read_csv("iris.csv")
data.head(10)
```

OUTPUT

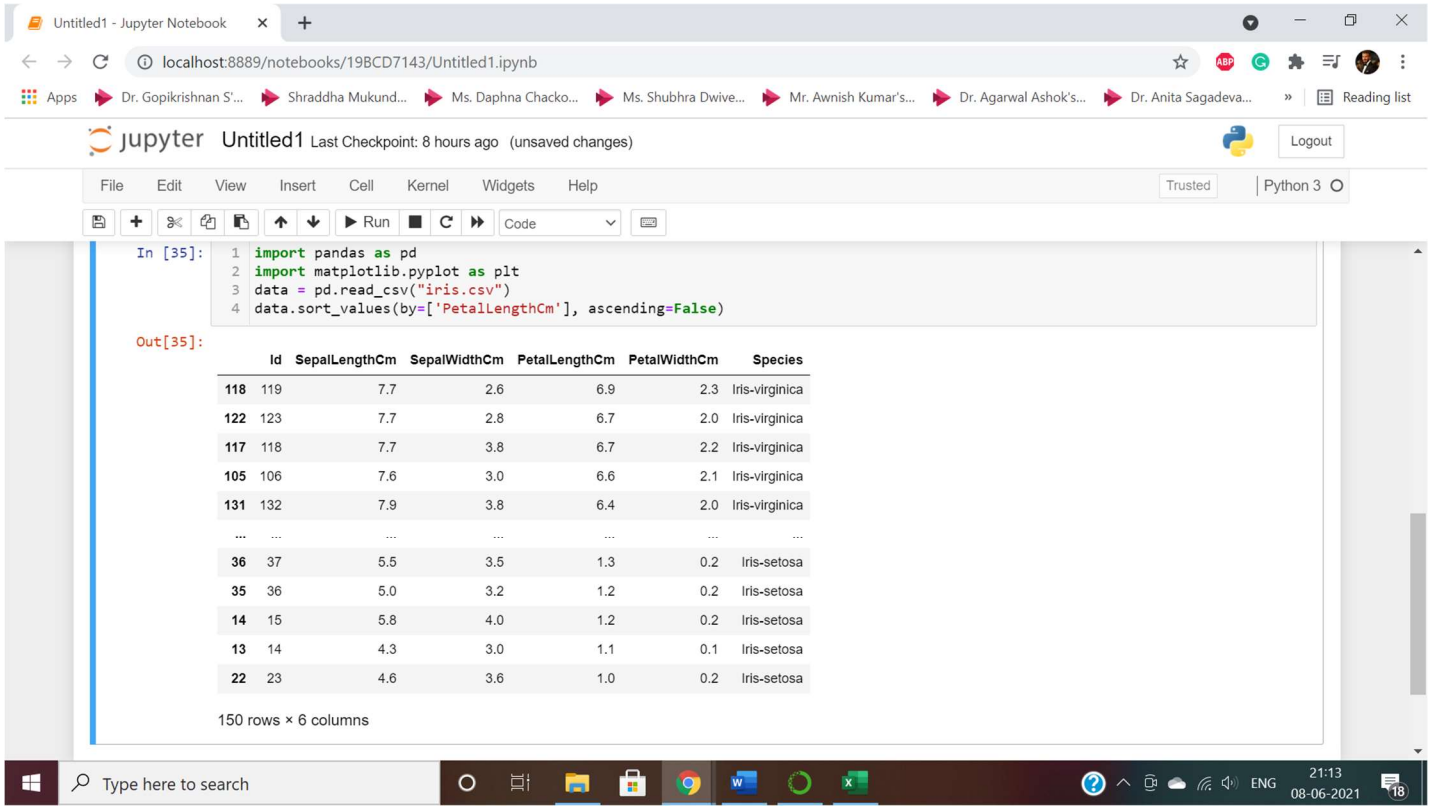


vi)Sort the data in descending order of PetalLength.

Answer:

```
import pandas as pd
import matplotlib.pyplot as plt
data = pd.read_csv("iris.csv")
data.sort_values(by=['PetalLengthCm'], ascending=False)
```

OUTPUT



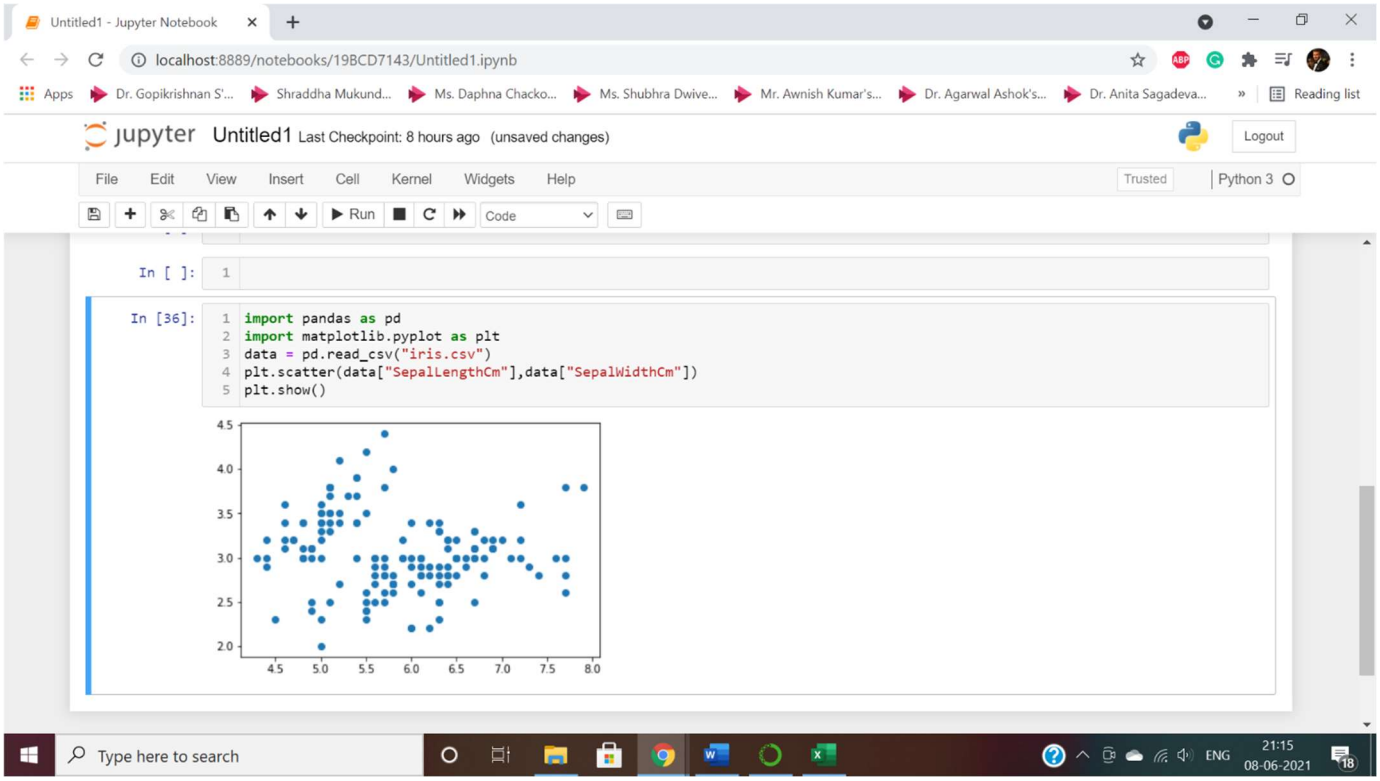
7.

a)Draw scatter between sepal_length and sepal_width.

Answer:

```
import pandas as pd
import matplotlib.pyplot as plt
data = pd.read_csv("iris.csv")
plt.scatter(data["SepalLengthCm"],data["SepalWidthCm"])
plt.show()
```

OUTPUT



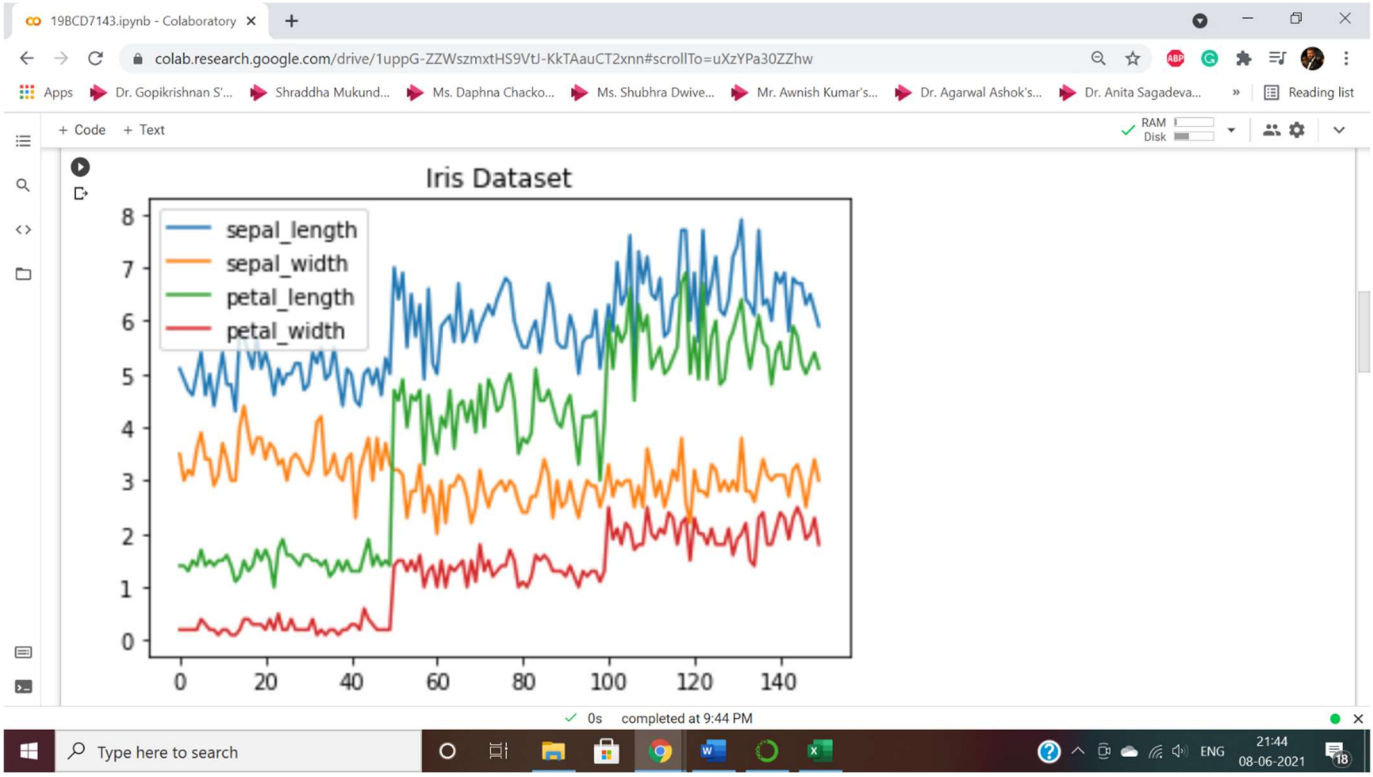
b)Draw line graph of all class of iris data.

Answer:

```
import pandas as pd

iris = pd.read_csv('iris.csv', names=['sepal_length', 'sepal_width',
'petal_length', 'petal_width', 'class'])

columns = iris.columns.drop(['class'])

x_data = range(0, iris.shape[0])

fig, ax = plt.subplots()

for column in columns:
```

```
    ax.plot(x_data, iris[column])


ax.set_title('Iris Dataset')

ax.legend()
```

OUTPUT



c)Compute iris.plot.hist(subplots=True, layout=(2,2), figsize=(10, 10), bins=20).

Answer:

```
import pandas as pd

import matplotlib.pyplot as plt

data = pd.read_csv("iris.csv")


data.plot.hist(subplots=True, layout=(5,5), figsize=(10, 10), bins=20)
```

OUTPUT

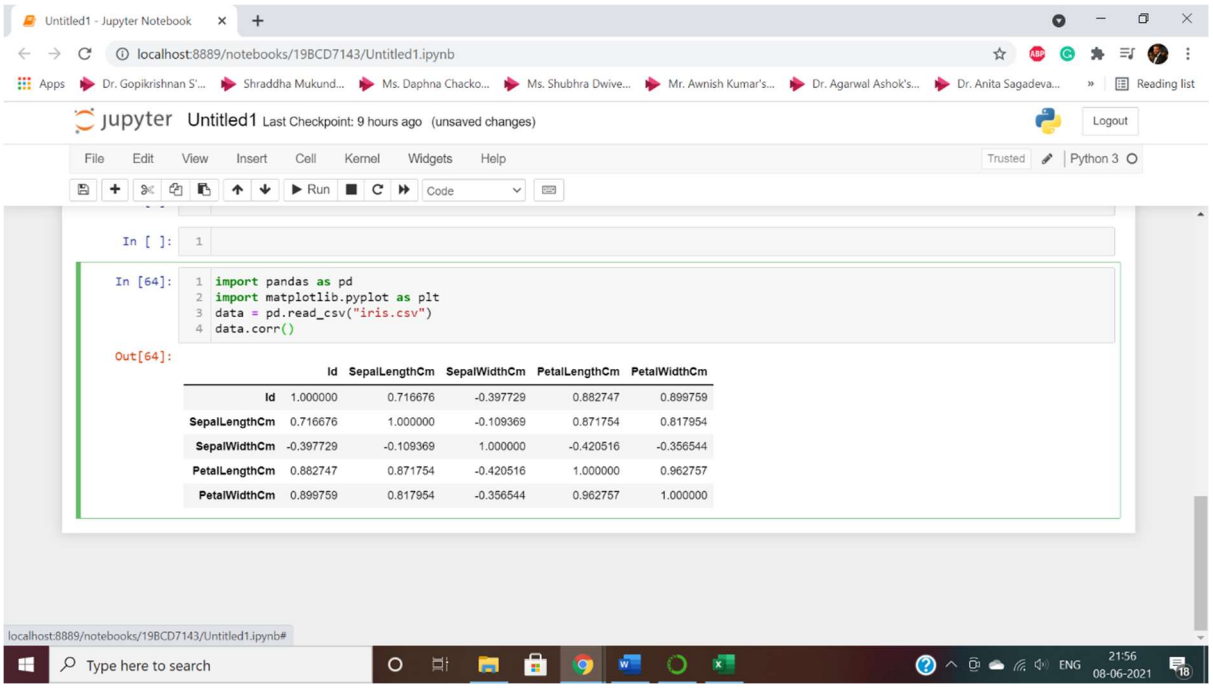
d)Draw correlation matrix between all class.

Answer:

```
import pandas as pd
import matplotlib.pyplot as plt
data = pd.read_csv("iris.csv")
data.corr()
```

```
corr=data.corr()
fig, ax = plt.subplots()

# create heatmap
im = ax.imshow(corr.values)


# set labels
ax.set_xticks(np.arange(len(corr.columns)))
ax.set_yticks(np.arange(len(corr.columns)))
ax.set_xticklabels(corr.columns)
ax.set_yticklabels(corr.columns)


# Rotate the tick labels and set their alignment.
plt.setp(ax.get_xticklabels(), rotation=45, ha="right",
         rotation_mode="anchor")

# Loop over data dimensions and create text annotations.
for i in range(len(corr.columns)):
    for j in range(len(corr.columns)):
        text = ax.text(j, i, np.around(corr.iloc[i, j], decimals=2),
ha="center", va="center", color="black")
```

OUTPUT


Untitled1 - Jupyter Notebook
localhost:8889/notebooks/19BCD7143/Untitled1.ipynb
jupyter Untitled1 Last Checkpoint: 9 hours ago (unsaved changes)
Logout
File Edit View Insert Cell Kernel Widgets Help
Trusted Python 3
Run Code
In [ ]: 1
In [64]:
1 import pandas as pd
2 import matplotlib.pyplot as plt
3 data = pd.read_csv("iris.csv")
4 data.corr()
Out[64]:
Id SepalLengthCm SepalWidthCm PetalLengthCm PetalWidthCm
Id 1.000000 0.716676 -0.397729 0.882747 0.899759
SepalLengthCm 0.716676 1.000000 -0.109369 0.871754 0.817954
SepalWidthCm -0.397729 -0.109369 1.000000 -0.420516 -0.356544
PetalLengthCm 0.882747 0.871754 -0.420516 1.000000 0.962757
PetalWidthCm 0.899759 0.817954 -0.356544 0.962757 1.000000
localhost:8889/notebooks/19BCD7143/Untitled1.ipynb#
Type here to search
ENG
21:56
08-06-2021


19BCD7143.ipynb - Colaboratory
colab.research.google.com/drive/1uppG-ZZWszmxtHS9VU-KkTAauCT2xnn#scrollTo=uXzYPa30ZZhw
+ Code + Text
RAM
Disk
sepal_length 1.0 -0.11 0.87 0.82
sepal_width -0.11 1.0 -0.42 -0.36
petal_length 0.87 -0.42 1.0 0.96
petal_width 0.82 -0.36 0.96 1.0
sepal_length
sepal_width
petal_length
petal_width
0s completed at 10:00 PM
Type here to search
ENG
22:00
08-06-2021