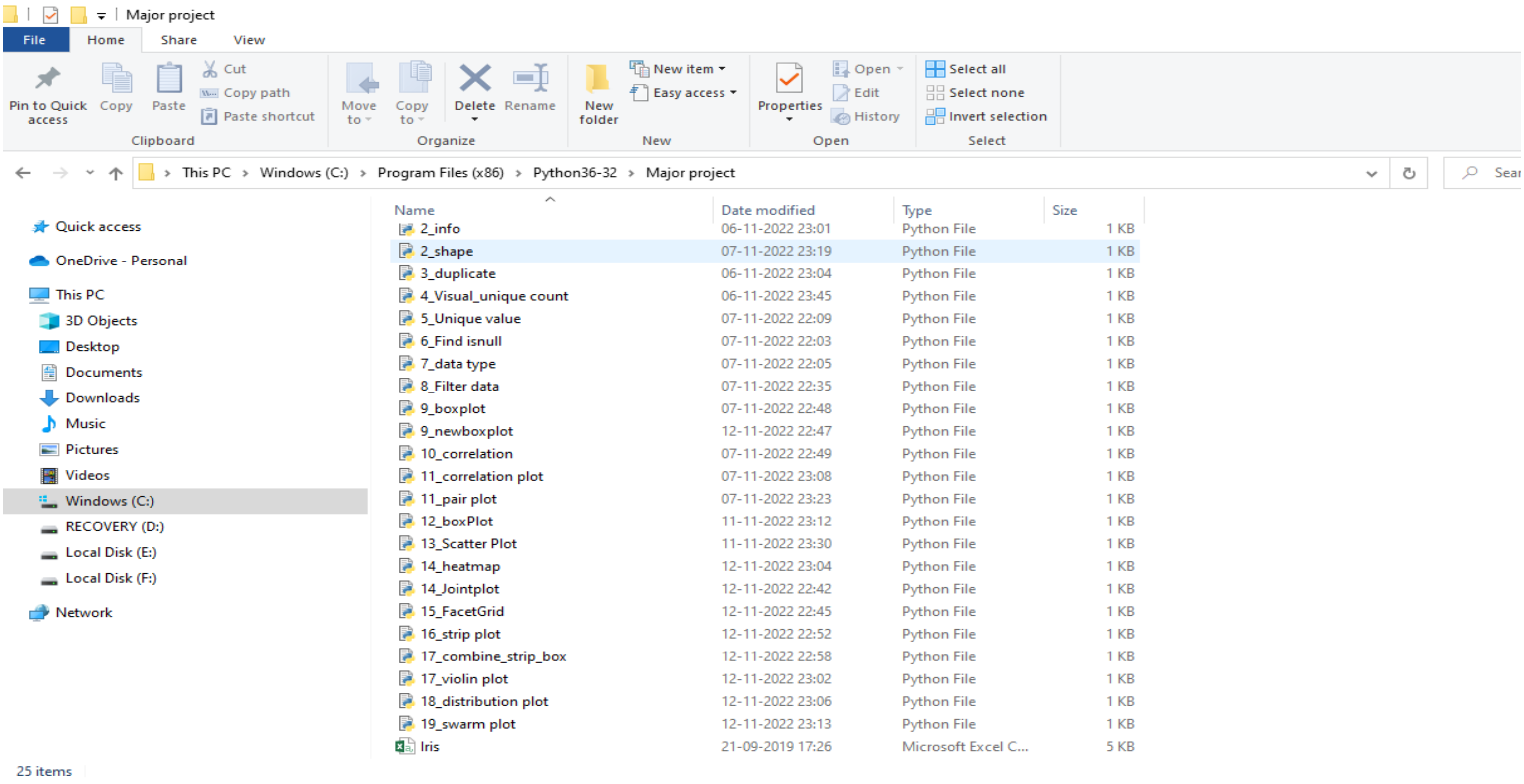
Name –ARYA BHATTACHARJEE
Batch- September
Course-Datascience with python

Take any Dataset of your choice ,perform EDA(Exploratory Data Analysis) and apply a suitable Classifier,Regressor or Clusterer and calculate the accuracy of the model.

Exploratory Data Analysis - EDA

EDA is applied to investigate the data and summarize the key insights.
It will give you the basic understanding of your data, it's distribution, null values and much more.
You can either explore data using graphs or through some python functions.
There will be two type of analysis. Univariate and Bivariate. In the univariate, you will be analyzing a single attribute. But in the bivariate, you will be analyzing an attribute with the target attribute.
In the non-graphical approach, you will be using functions such as shape, summary, describe, isnull, info, datatypes and more.
In the graphical approach, you will be using plots such as scatter, box, bar, density and correlation plots.

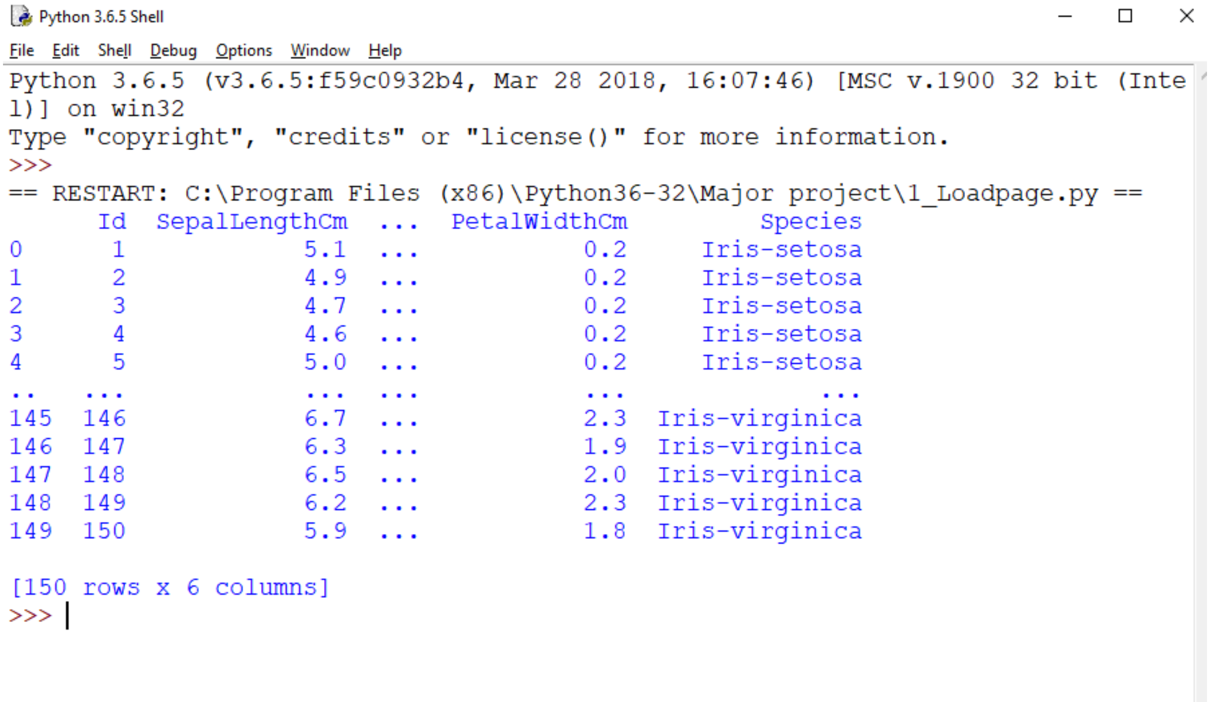


Name	Date modified	Type	Size
2_info	06-11-2022 23:01	Python File	1 KB
2_shape	07-11-2022 23:19	Python File	1 KB
3_duplicate	06-11-2022 23:04	Python File	1 KB
4_Visual_unique count	06-11-2022 23:45	Python File	1 KB
5_Unique value	07-11-2022 22:09	Python File	1 KB
6_Find isnull	07-11-2022 22:03	Python File	1 KB
7_data type	07-11-2022 22:05	Python File	1 KB
8_Filter data	07-11-2022 22:35	Python File	1 KB
9_boxplot	07-11-2022 22:48	Python File	1 KB
9_newboxplot	12-11-2022 22:47	Python File	1 KB
10_correlation	07-11-2022 22:49	Python File	1 KB
11_correlation plot	07-11-2022 23:08	Python File	1 KB
11_pair plot	07-11-2022 23:23	Python File	1 KB
12_boxPlot	11-11-2022 23:12	Python File	1 KB
13_Scatter Plot	11-11-2022 23:30	Python File	1 KB
14_heatmap	12-11-2022 23:04	Python File	1 KB
14_Jointplot	12-11-2022 22:42	Python File	1 KB
15_FacetGrid	12-11-2022 22:45	Python File	1 KB
16_strip plot	12-11-2022 22:52	Python File	1 KB
17_combine_strip_box	12-11-2022 22:58	Python File	1 KB
17_violin plot	12-11-2022 23:02	Python File	1 KB
18_distribution plot	12-11-2022 23:06	Python File	1 KB
19_swarm plot	12-11-2022 23:13	Python File	1 KB
Iris	21-09-2019 17:26	Microsoft Excel C...	5 KB

Load the Data

first things first. We will load the titanic dataset into python to perform EDA.

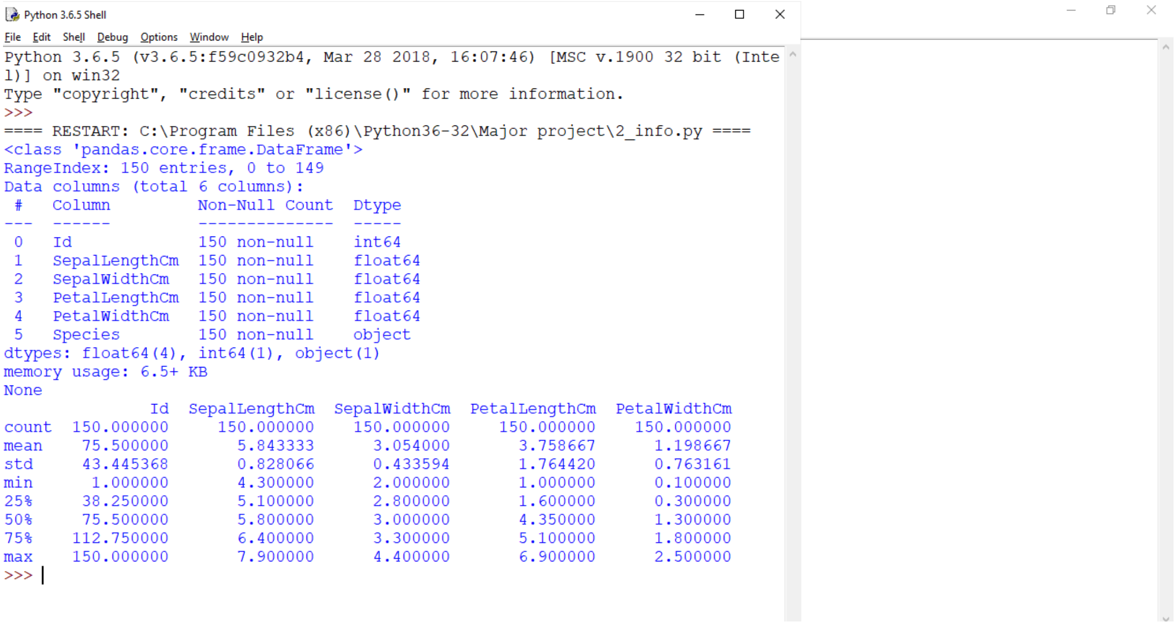
```
import pandas as pd
import numpy as np
import seaborn as sns
df=pd.read_csv("Iris.csv")
df.head()
print(df)
```



```
Python 3.6.5 (v3.6.5:f59c0932b4, Mar 28 2018, 16:07:46) [MSC v.1900 32 bit (Inte
l)] on win32
Type "copyright", "credits" or "license()" for more information.
>>>
== RESTART: C:\Program Files (x86)\Python36-32\Major project\1_Loadpage.py ==
      Id  SepalLengthCm  ...  PetalWidthCm         Species
0      1            5.1  ...           0.2     Iris-setosa
1      2            4.9  ...           0.2     Iris-setosa
2      3            4.7  ...           0.2     Iris-setosa
3      4            4.6  ...           0.2     Iris-setosa
4      5            5.0  ...           0.2     Iris-setosa
..   ...            ...  ...           ...             ...
145  146            6.7  ...           2.3  Iris-virginica
146  147            6.3  ...           1.9  Iris-virginica
147  148            6.5  ...           2.0  Iris-virginica
148  149            6.2  ...           2.3  Iris-virginica
149  150            5.9  ...           1.8  Iris-virginica

[150 rows x 6 columns]
>>> 
```

Basic information about data - EDA

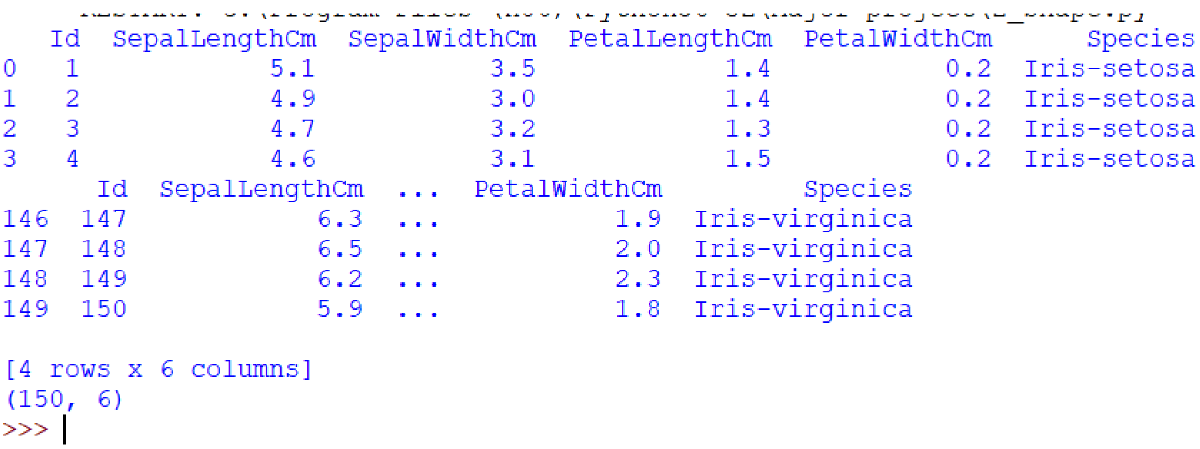
```
import pandas as pd
import numpy as np
import seaborn as sns
df=pd.read_csv("Iris.csv")
print(df.info())
print(df.describe())
```



```
Python 3.6.5 (v3.6.5:f59c0932b4, Mar 28 2018, 16:07:46) [MSC v.1900 32 bit (Inte
l)] on win32
Type "copyright", "credits" or "license()" for more information.
>>>
==== RESTART: C:\Program Files (x86)\Python36-32\Major project\2_info.py ====
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 150 entries, 0 to 149
Data columns (total 6 columns):
 #   Column         Non-Null Count  Dtype  
---  ------         --------------  -----  
 0   Id             150 non-null    int64  
 1   SepalLengthCm  150 non-null    float64
 2   SepalWidthCm   150 non-null    float64
 3   PetalLengthCm  150 non-null    float64
 4   PetalWidthCm   150 non-null    float64
 5   Species        150 non-null    object 
dtypes: float64(4), int64(1), object(1)
memory usage: 6.5+ KB
None
               Id  SepalLengthCm  SepalWidthCm  PetalLengthCm  PetalWidthCm
count  150.000000     150.000000    150.000000     150.000000    150.000000
mean    75.500000       5.843333      3.054000       3.758667      1.198667
std     43.445368       0.828066      0.433594       1.764420      0.763161
min      1.000000       4.300000      2.000000       1.000000      0.100000
25%     38.250000       5.100000      2.800000       1.600000      0.300000
50%     75.500000       5.800000      3.000000       4.350000      1.300000
75%    112.750000       6.400000      3.300000       5.100000      1.800000
max    150.000000       7.900000      4.400000       6.900000      2.500000
>>> 
```

Shape

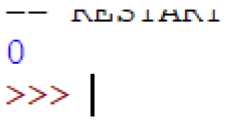
```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
df=pd.read_csv("Iris.csv")
print(df.head(4))
print(df.tail(4))
print(df.shape)
```



```
  RESTART: C:\Program Files (x86)\Python37-32\major project\2_shape.py]
   Id  SepalLengthCm  SepalWidthCm  PetalLengthCm  PetalWidthCm      Species
0   1            5.1           3.5            1.4           0.2  Iris-setosa
1   2            4.9           3.0            1.4           0.2  Iris-setosa
2   3            4.7           3.2            1.3           0.2  Iris-setosa
3   4            4.6           3.1            1.5           0.2  Iris-setosa
      Id  SepalLengthCm  ...  PetalWidthCm         Species
146  147            6.3  ...           1.9  Iris-virginica
147  148            6.5  ...           2.0  Iris-virginica
148  149            6.2  ...           2.3  Iris-virginica
149  150            5.9  ...           1.8  Iris-virginica

[4 rows x 6 columns]
(150, 6)
>>> 
```

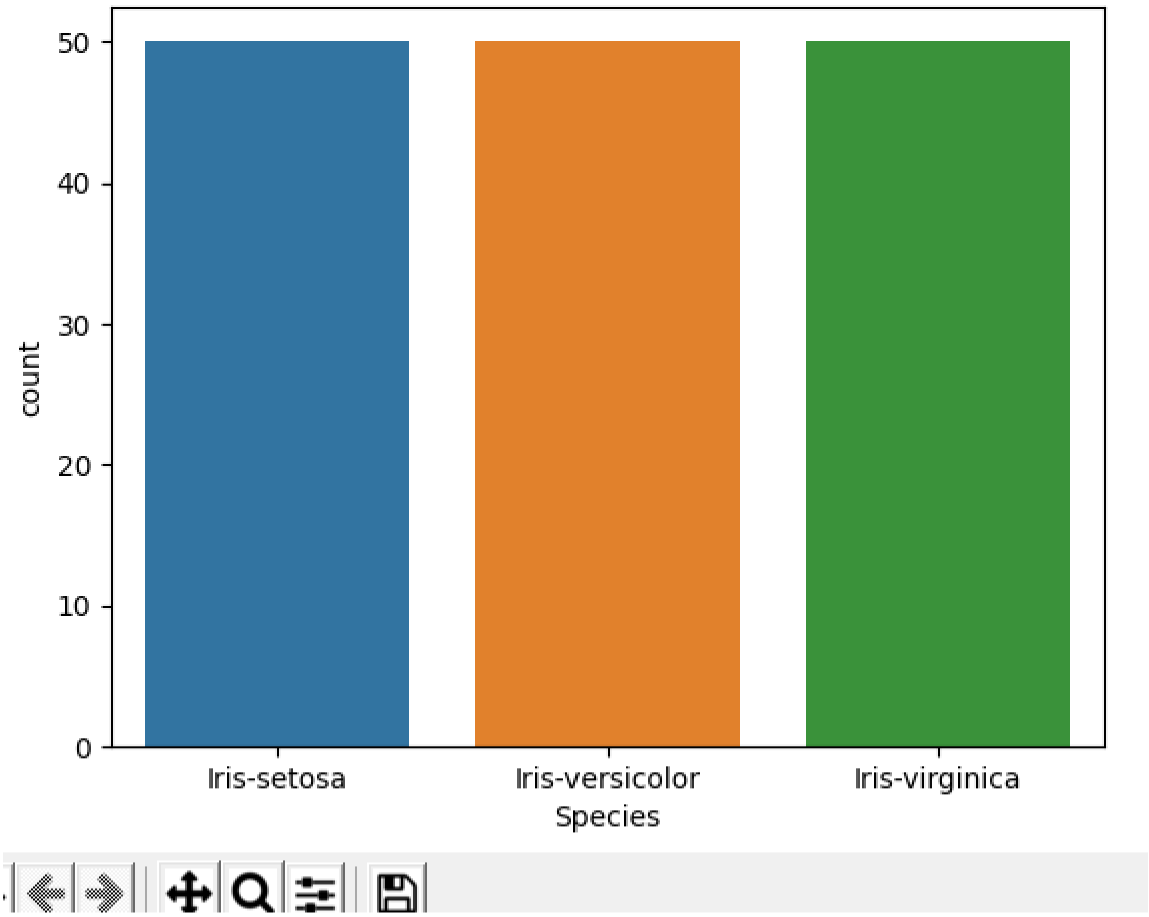
Duplicate

```
import pandas as pd
import numpy as np
import seaborn as sns
df=pd.read_csv("Iris.csv")
print(df.duplicated().sum())
```

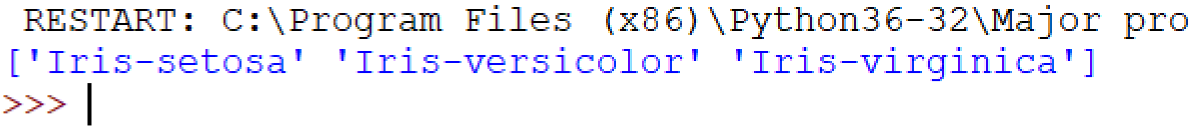


```
== RESTART
0
>>> 
```

```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
df=pd.read_csv("Iris.csv")

sns.countplot(x='Species',data=df)
plt.show()
#print(df.sum())
```

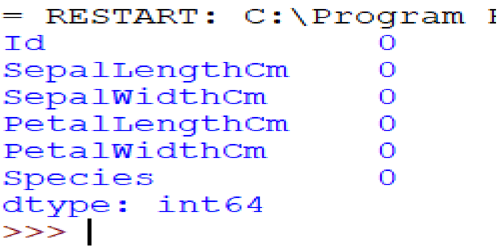


```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
df=pd.read_csv("Iris.csv")
print(df['Species'].unique())
```



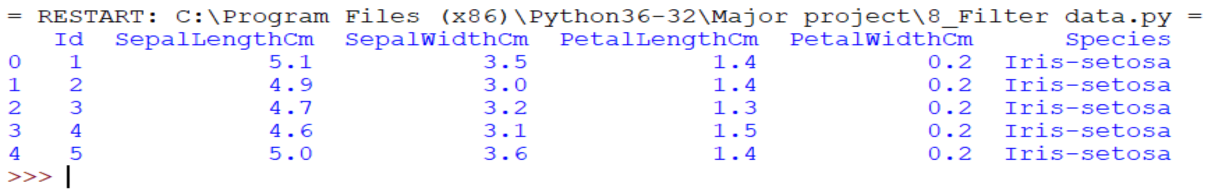
```
 RESTART: C:\Program Files (x86)\Python36-32\Major pro
['Iris-setosa' 'Iris-versicolor' 'Iris-virginica']
>>> 
```

```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
df=pd.read_csv("Iris.csv")
print(df.isnull().sum())
```



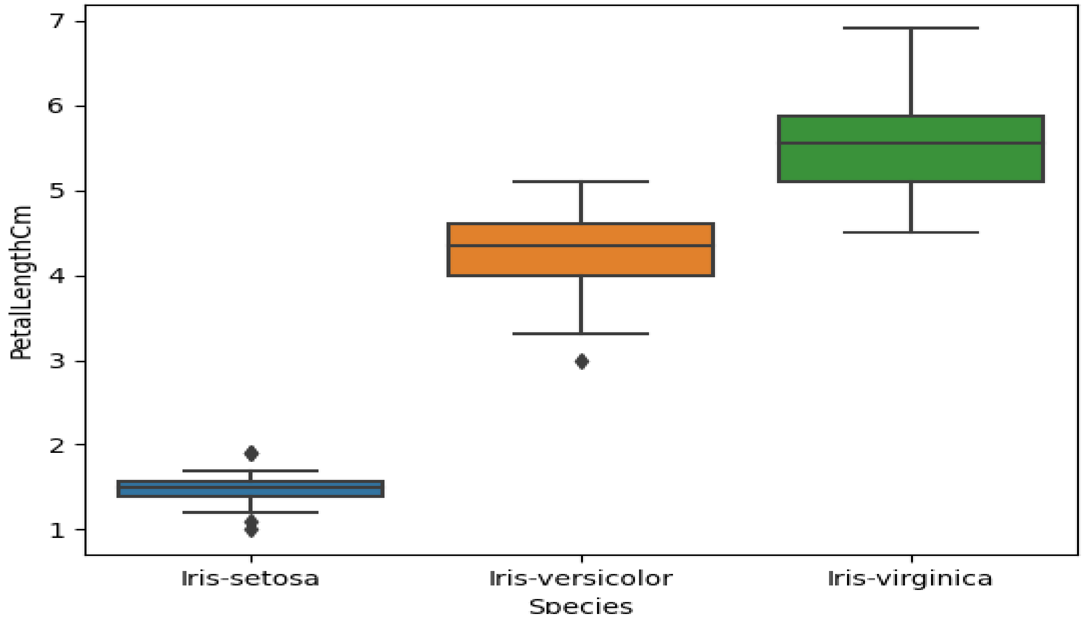
```
= RESTART: C:\Program I
Id               0
SepalLengthCm    0
SepalWidthCm     0
PetalLengthCm    0
PetalWidthCm     0
Species          0
dtype: int64
>>> 
```

```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
df=pd.read_csv("Iris.csv")
print(df[df['Species']=='Iris-setosa'].head())
```

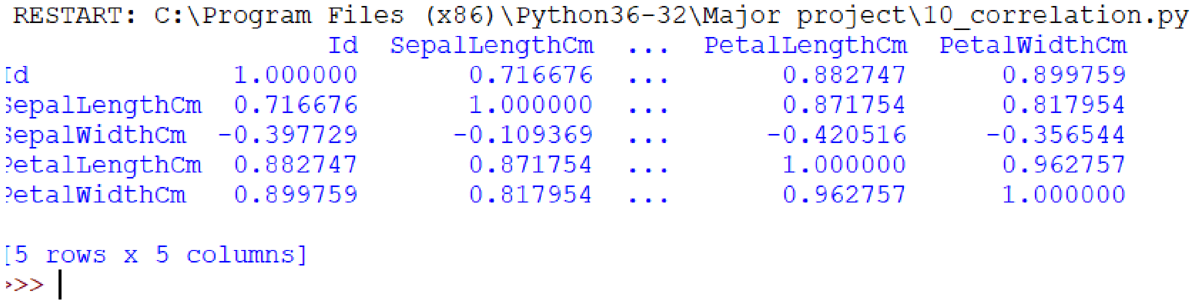


```
= RESTART: C:\Program Files (x86)\Python36-32\Major project\8_Filter data.py =
   Id  SepalLengthCm  SepalWidthCm  PetalLengthCm  PetalWidthCm      Species
0   1            5.1           3.5            1.4           0.2  Iris-setosa
1   2            4.9           3.0            1.4           0.2  Iris-setosa
2   3            4.7           3.2            1.3           0.2  Iris-setosa
3   4            4.6           3.1            1.5           0.2  Iris-setosa
4   5            5.0           3.6            1.4           0.2  Iris-setosa
>>> |
```

```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
df=pd.read_csv("Iris.csv")
print(df.head())
sns.boxplot(x='Species',y='PetalLengthCm',data=df)
plt.show()
```

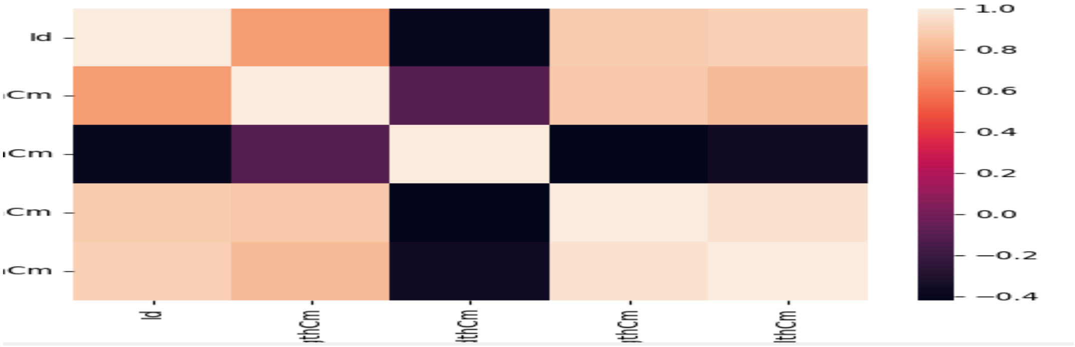


```
import pandas as pd

import numpy as np

import seaborn as sns

import matplotlib.pyplot as plt

df=pd.read_csv("Iris.csv")

print(df.corr())
```



```
 RESTART: C:\Program Files (x86)\Python36-32\Major project\10_correlation.py
                     Id  SepalLengthCm  ...  PetalLengthCm  PetalWidthCm
Id             1.000000       0.716676  ...       0.882747      0.899759
SepalLengthCm  0.716676       1.000000  ...       0.871754      0.817954
SepalWidthCm  -0.397729      -0.109369  ...      -0.420516     -0.356544
PetalLengthCm  0.882747       0.871754  ...       1.000000      0.962757
PetalWidthCm   0.899759       0.817954  ...       0.962757      1.000000

[5 rows x 5 columns]
>>> 
```

```
import pandas as pd

import numpy as np

import seaborn as sns

import matplotlib.pyplot as plt

df=pd.read_csv("Iris.csv")

sns.heatmap(df.corr())

plt.show()
```



import pandas as pd

import numpy as np

import seaborn as sns

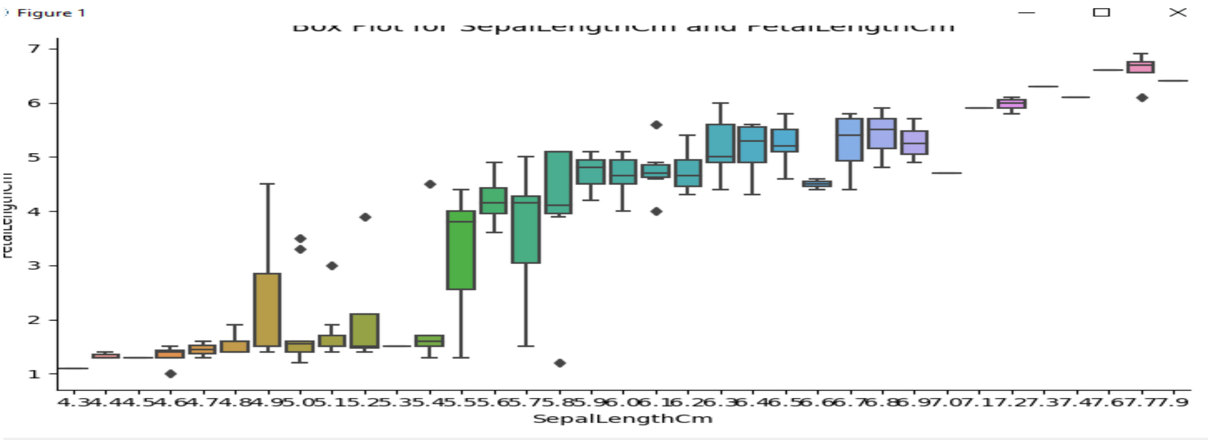
import matplotlib.pyplot as plt

df=pd.read_csv("Iris.csv")

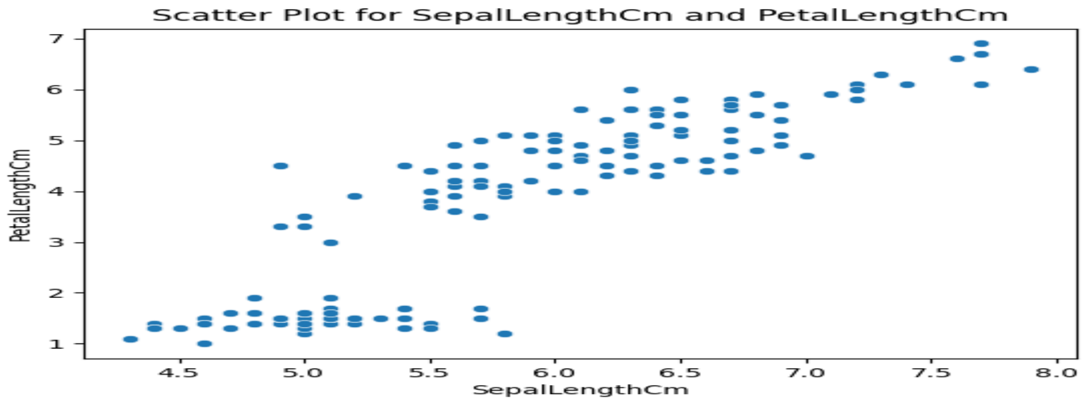
sns.catplot(x="SepalLengthCm",y="PetalLengthCm", data=df, kind="box",aspect=1.5)

plt.title("Box Plot for SepalLengthCm and PetalLengthCm")

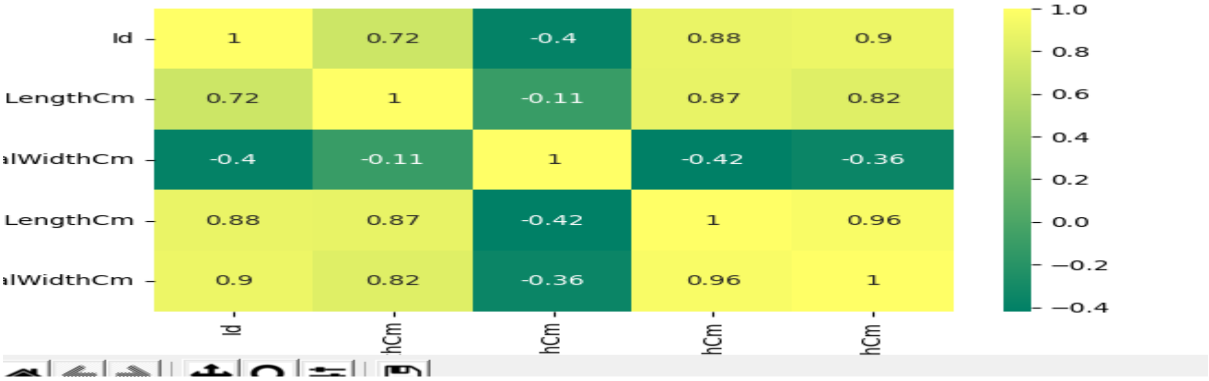
plt.show()



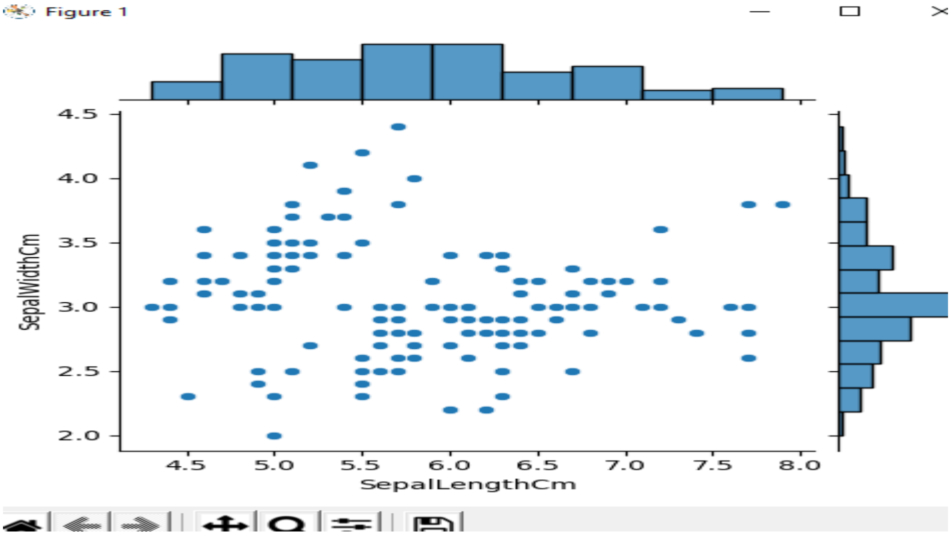
```
import pandas as pd

import numpy as np

import seaborn as sns

import matplotlib.pyplot as plt

df=pd.read_csv("Iris.csv")

sns.scatterplot(x="SepalLengthCm",y="PetalLengthCm", data=df)

plt.title("Scatter Plot for SepalLengthCm and PetalLengthCm")

plt.show()
```



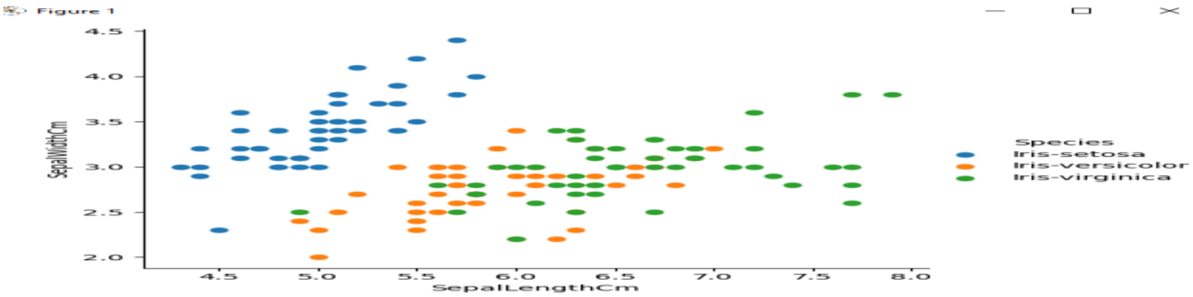
```
import pandas as pd

import numpy as np

import seaborn as sns

import matplotlib.pyplot as plt

df=pd.read_csv("Iris.csv")

print(df.head())

plt.figure(figsize=(7,4))

sns.heatmap(df.corr(),annot=True,cmap='summer')

plt.show()
```



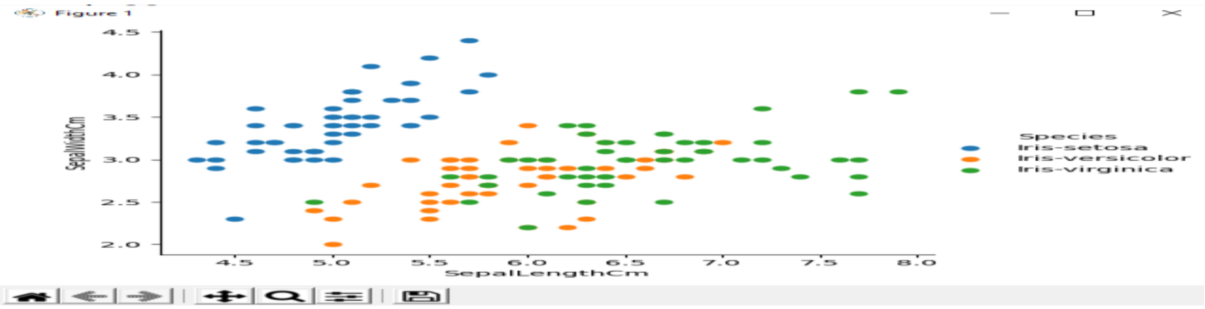
```
import pandas as pd

import numpy as np

import seaborn as sns

import matplotlib.pyplot as plt

df=pd.read_csv("Iris.csv")

print(df.head())

sns.jointplot(x='SepalLengthCm',y='SepalWidthCm',data=df,height=5)

plt.show()
```



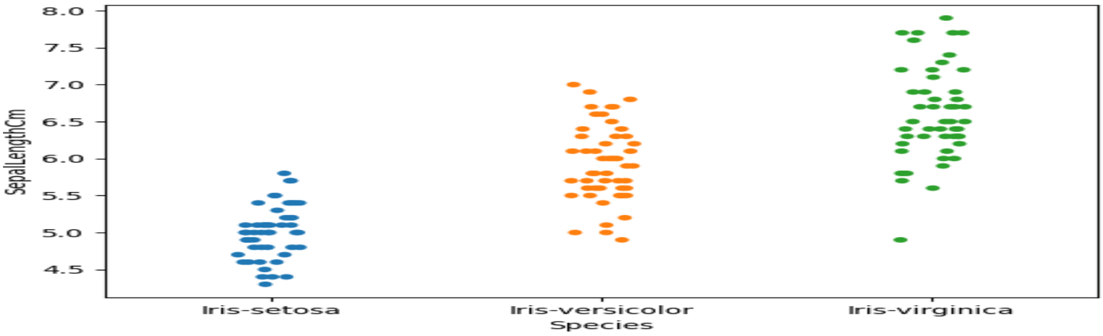
```
import pandas as pd

import numpy as np

import seaborn as sns

import matplotlib.pyplot as plt

df=pd.read_csv("Iris.csv")

print(df.head())

sns.FacetGrid(df,hue='Species',height=5)\

.map(plt.scatter,'SepalLengthCm','SepalWidthCm')\

.add_legend()

plt.show()
```



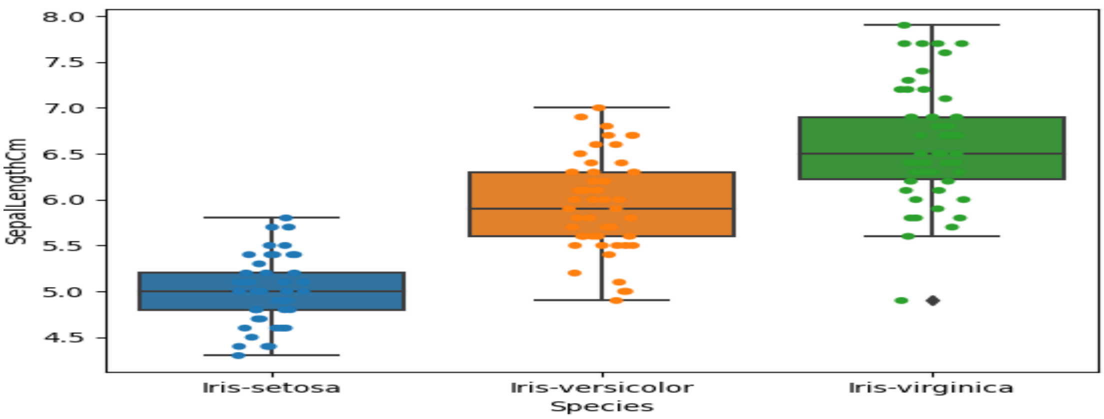
```
import pandas as pd

import numpy as np

import seaborn as sns

import matplotlib.pyplot as plt

df=pd.read_csv("Iris.csv")

print(df.head())

sns.FacetGrid(df,hue='Species',height=5)\

.map(plt.scatter,'SepalLengthCm','SepalWidthCm')\

.add_legend()

plt.show()
```



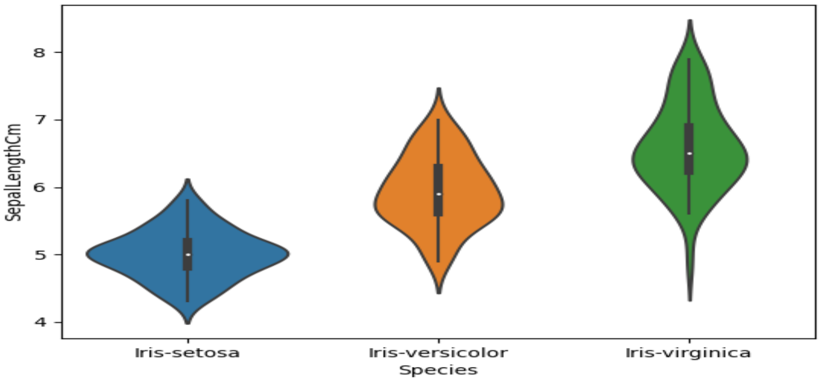
```
import pandas as pd

import numpy as np

import seaborn as sns

import matplotlib.pyplot as plt

df=pd.read_csv("Iris.csv")

print(df.head())

ax=sns.stripplot(x='Species',y='SepalLengthCm',data=df,jitter=True,edgecolor='gray')

plt.show()
```



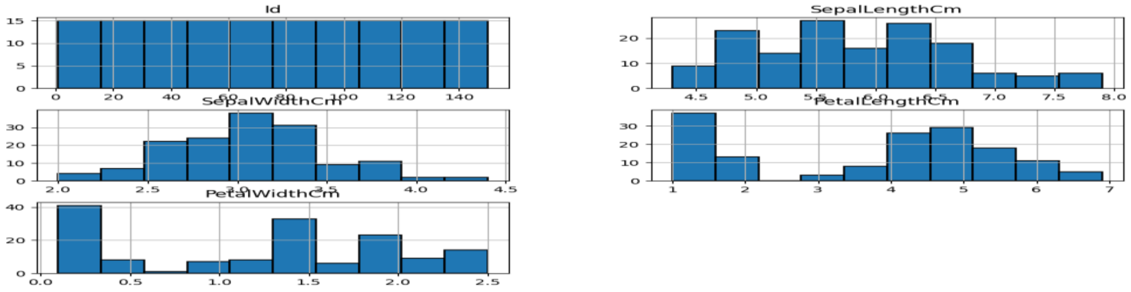
```
import pandas as pd

import numpy as np

import seaborn as sns

import matplotlib.pyplot as plt

df=pd.read_csv("Iris.csv")

print(df.head())

ax=sns.boxplot(x='Species',y='SepalLengthCm',data=df)

ax1=sns.stripplot(x='Species',y='SepalLengthCm',data=df,jitter=True,edgecolor='gray')

plt.show()
```



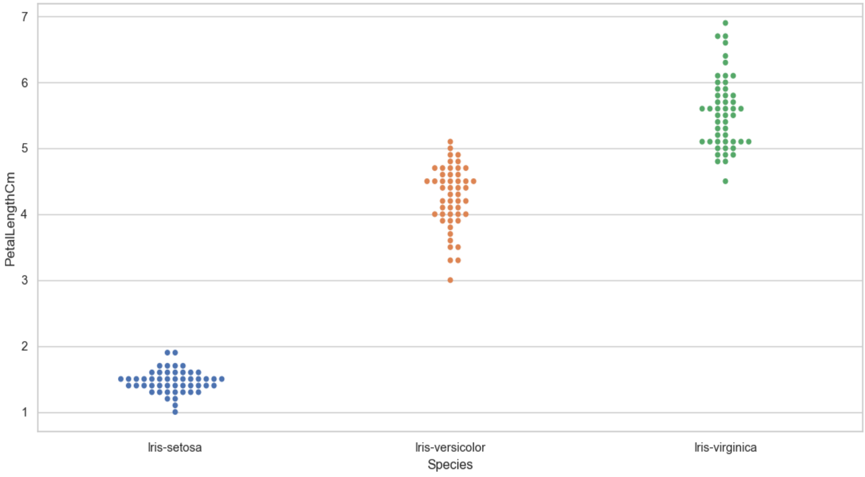
```
import pandas as pd

import numpy as np

import seaborn as sns

import matplotlib.pyplot as plt

df=pd.read_csv("Iris.csv")

print(df.head())

sns.violinplot(x='Species',y='SepalLengthCm',data=df,height=6)

plt.show()
```



```
import pandas as pd

import numpy as np

import seaborn as sns

import matplotlib.pyplot as plt

df=pd.read_csv("Iris.csv")

print(df.head())

df.hist(edgecolor='black', linewidth=1.2)

fig=plt.gcf()

fig.set_size_inches(12,6)

plt.show()
```



```
import pandas as pd

import numpy as np

import seaborn as sns

import matplotlib.pyplot as plt

df=pd.read_csv("Iris.csv")

print(df.head())

sns.set(style="whitegrid")

fig=plt.gcf()

fig.set_size_inches(10,7)

fig = sns.swarmplot(x="Species", y="PetalLengthCm", data=df)

plt.show()
```


PetalLengthCm
Iris-setosa
Iris-versicolor
Iris-virginica
Species