LP I

Assignment C1

Data Analysis

Title :- Iris Data Analysis

Date of Completion :- 5.11.20

Problem Statement :- Download the Iris Flower Dataset or any other dataset into a DataFrame. Use Python/R and Perform following

1) How many features are there & what are their types?
2) Compute & display summary statistics for each feature available in the dataset.
3) Data Visualization :- Create a histogram for each feature in the dataset to illustrate the feature distributions. Plot each histogram.
4) Create a boxplot for each feature in the dataset. All of the boxplots should be combined into a single plot. Compare distribution & identify outliers

Learning Objective :- 1) Understand dataframes and its features
2) Analyse Iris dataset.

Learning Outcome :- Students will be able to
1) Analyze different datasets.

Software/Hardware Requirement :- OS (Linux), Python, ~~gotot~~ Iris Dataset.

Page:
Date: / /

Theory:-

Libraries used:-

1) pandas
2) Numpy
3) matpotlib
4)

Mathematical Model:-

Let S be the System set:-

$S = \{s; e; X; Y; Fme; DD; NDD; FC; Sc\}$ where Dataset is loaded into the dataframe

s = start state

e = end state ie summary statistics for each feature is captured

X = set of inputs

Y = set of outputs

DD → Deterministic Data

NDD → Non deterministic Data

fc → Failure case

1. Data set is collection of data.
2. Data analysis is a process of inspecting, cleansing, transforming, & modelling data with the goal of discovering useful information, informatn conclusion & decision making.
3. Mean, Standard, Variance, regression, hypothesis are the fundamental data analytics methods.

Page:
Date: / /

Mean

Sum of data entities divided by no. of entities.

Population Mean $\mu = \frac{\sum x}{N}$

Sample Mean $\bar{x} = \frac{\sum x}{n}$

Standard deviation:-

measure variability and consistency of the sample or population.

$$\sigma = \sqrt{\frac{\sum (x-\mu)^2}{N}}$$

Variance:-

averaged squared deviation from the mean.

Dataset Used :- Iris dataset.

- describe()

gives all the parameters like mean, std. deviation, variance

- hist()

creates histogram

- boxplot()

plot a box plot

Test cases

Input	Actual Output	Expected O/P	Remark
describe for column 1.	count 150.0 mean 5.84 std V. 0.82 min 4.30 max 7.90	count 150.0 mean 5.84 std V. 0.82 min 4.30 max 7.9	Passed
plot histogram	plotted	plotted	Passed
Boxplot	plotted	plotted	Passed

Conclusion:- Thus I analyzed iris dataset successfully.

CODE

```
import numpy as np
import pandas as pd
#%matplotlib inline
import matplotlib.pyplot as plt
import seaborn as sns
dat=pd.read_csv('Iris.csv')
dat[0:10]

dat.shape
list(dat.columns)

dat.dtypes

dat['x1'].describe()
dat['x2'].describe()
dat['x3'].describe()
dat['x4'].describe()

dat.mean()

plt.hist(dat['x1'],bins=30)
plt.ylabel('No of times')
plt.show()


plt.hist(dat['x2'],bins=30)
plt.ylabel('No of times')
plt.show()


plt.hist(dat['x3'],bins=30)
plt.ylabel('No of times')
plt.show()


plt.hist(dat['x4'],bins=30)
plt.ylabel('No of times')
plt.show()

sns.boxplot(y=dat['x1'])
sns.boxplot(y=dat['x2'])
sns.boxplot(y=dat['x3'])
sns.boxplot(y=dat['x4'])


dat.max()
dat.min()
sns.boxplot(x=dat['class'],y=dat['x2'])

dat.pstdev()

sns.boxplot(data=dat.ix[:,0:4])

sns.boxplot(x=dat['class'],y=dat['x1'])

sns.boxplot(x=dat['class'],y=dat['x3'])

sns.boxplot(x=dat['class'],y=dat['x4'])
```

OUTPUT

File Edit Search Source Run Debug Consoles Projects Tools View Help

/home/srushti/BE Sem1/my/LP1/DA/1

Editor - /home/srushti/BE Sem1/my/LP1/DA/1/code.py

c4.py | code.py

```python
import numpy as np
import pandas as pd
#%matplotlib inline
import matplotlib.pyplot as plt
import seaborn as sns


dat=pd.read_csv('Iris.csv')


dat[0:10]


dat.shape
list(dat.columns)

dat.dtypes

dat['x1'].describe()
dat['x2'].describe()
dat['x3'].describe()
dat['x4'].describe()

dat.mean()


plt.hist(dat['x1'],bins=30)            ##############plot histogram
plt.ylabel('No of times')
plt.show()


plt.hist(dat['x2'],bins=30)            ##############plot histogram
plt.ylabel('No of times')
plt.show()


plt.hist(dat['x3'],bins=30)            ##############plot histogram
plt.ylabel('No of times')
plt.show()


plt.hist(dat['x4'],bins=30)            ##############plot histogram
plt.ylabel('No of times')
plt.show()

sns.boxplot(y=dat['x1'])

dat.min()
dat.max()
```

IPython console

Console 1/A

```
   ....
   ...: dat[0:10]
Out[3]:
    x1   x2   x3   x4        class
0  5.1  3.5  1.4  0.2  Iris-setosa
1  4.9  3.0  1.4  0.2  Iris-setosa
2  4.7  3.2  1.3  0.2  Iris-setosa
3  4.6  3.1  1.5  0.2  Iris-setosa
4  5.0  3.6  1.4  0.2  Iris-setosa
5  5.4  3.9  1.7  0.4  Iris-setosa
6  4.6  3.4  1.4  0.3  Iris-setosa
7  5.0  3.4  1.5  0.2  Iris-setosa
8  4.4  2.9  1.4  0.2  Iris-setosa
9  4.9  3.1  1.5  0.1  Iris-setosa

In [4]: dat.shape
Out[4]: (150, 5)

In [5]: list(dat.columns)
Out[5]: ['x1', 'x2', 'x3', 'x4', 'class']

In [6]: dat.dtypes
Out[6]:
x1        float64
x2        float64
x3        float64
x4        float64
class      object
dtype: object

In [7]: dat['x1'].describe()
Out[7]:
count    150.000000
mean       5.843333
std        0.828066
min        4.300000
25%        5.100000
50%        5.800000
75%        6.400000
max        7.900000
Name: x1, dtype: float64

In [8]:
```

IPython console | History log

Permissions: RW

File Edit Search Source Run Debug Consoles Projects Tools View Help

/home/srushti/BE Sem1/my/LP1/DA/1

Editor - /home/srushti/BE Sem1/my/LP1/DA/1/code.py

c4.py | code.py

```python
import numpy as np
import pandas as pd
#%matplotlib inline
import matplotlib.pyplot as plt
import seaborn as sns


dat=pd.read_csv('Iris.csv')


dat[0:10]


dat.shape
list(dat.columns)

dat.dtypes

dat['x1'].describe()
dat['x2'].describe()
dat['x3'].describe()
dat['x4'].describe()

dat.mean()


plt.hist(dat['x1'],bins=30)            ##############plot histogram
plt.ylabel('No of times')
plt.show()


plt.hist(dat['x2'],bins=30)            ##############plot histogram
plt.ylabel('No of times')
plt.show()


plt.hist(dat['x3'],bins=30)            ##############plot histogram
plt.ylabel('No of times')
plt.show()


plt.hist(dat['x4'],bins=30)            ##############plot histogram
plt.ylabel('No of times')
plt.show()

sns.boxplot(y=dat['x1'])

dat.min()
dat.max()
```

IPython console

Console 1/A

```
In [8]: dat['x2'].describe()
Out[8]:
count    150.000000
mean       3.054000
std        0.433594
min        2.000000
25%        2.800000
50%        3.000000
75%        3.300000
max        4.400000
Name: x2, dtype: float64

In [9]: dat['x3'].describe()
Out[9]:
count    150.000000
mean       3.758667
std        1.764420
min        1.000000
25%        1.600000
50%        4.350000
75%        5.100000
max        6.900000
Name: x3, dtype: float64

In [10]: dat['x4'].describe()
Out[10]:
count    150.000000
mean       1.198667
std        0.763161
min        0.100000
25%        0.300000
50%        1.300000
75%        1.800000
max        2.500000
Name: x4, dtype: float64

In [11]: dat.mean()
Out[11]:
x1    5.843333
x2    3.054000
x3    3.758667
x4    1.198667
dtype: float64

In [12]:
```

IPython console | History log

Permissions: R

File Edit Search Source Run Debug Consoles Projects Tools View Help

/home/srushti/BE Sem1/my/LP1/DA/1

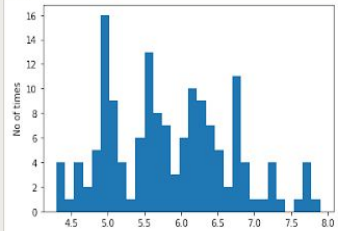
Editor - /home/srushti/BE Sem1/my/LP1/DA/1/code.py

c4.py | code.py

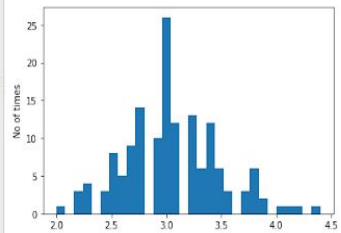
```
dat[0:10]


dat.shape
list(dat.columns)

dat.dtypes

dat['x1'].describe()
dat['x2'].describe()
dat['x3'].describe()
dat['x4'].describe()

dat.mean()


plt.hist(dat['x1'],bins=30)            ##############plot histogram
plt.ylabel('No of times')
plt.show()

plt.hist(dat['x2'],bins=30)            ##############plot histogram
plt.ylabel('No of times')
plt.show()


plt.hist(dat['x3'],bins=30)            ##############plot histogram
plt.ylabel('No of times')
plt.show()


plt.hist(dat['x4'],bins=30)            ##############plot histogram
plt.ylabel('No of times')
plt.show()

sns.boxplot(y=dat['x1'])

dat.min()
dat.max()

sns.boxplot(x=dat['class'],y=dat['x2'])

dat.pstdev()

sns.boxplot(data=dat.ix[:,0:4])

sns.boxplot(x=dat['class'],y=dat['x2'])
```

IPython console

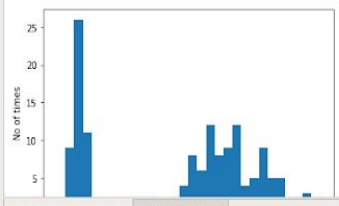
Console 1/A




```
In [13]: plt.hist(dat['x2'],bins=30)          ##############plot histogram
    ...: plt.ylabel('No of times')
    ...: plt.show()
```




```
In [14]: plt.hist(dat['x3'],bins=30)          ##############plot histogram
    ...: plt.ylabel('No of times')
    ...: plt.show()
```



IPython console | History log

Permissions: RW End-of-lines: LF Encoding:

File Edit Search Source Run Debug Consoles Projects Tools View Help

/home/srushti/BE Sem1/my/LP1/DA/1

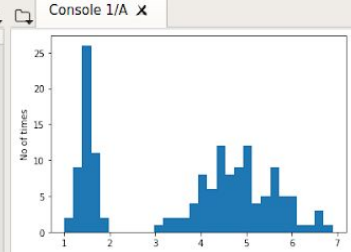
Editor - /home/srushti/BE Sem1/my/LP1/DA/1/code.py

c4.py | code.py

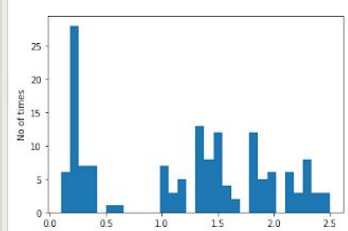
```
dat[0:10]


dat.shape
list(dat.columns)

dat.dtypes

dat['x1'].describe()
dat['x2'].describe()
dat['x3'].describe()
dat['x4'].describe()

dat.mean()


plt.hist(dat['x1'],bins=30)            ##############plot histogram
plt.ylabel('No of times')
plt.show()


plt.hist(dat['x2'],bins=30)            ##############plot histogram
plt.ylabel('No of times')
plt.show()


plt.hist(dat['x3'],bins=30)            ##############plot histogram
plt.ylabel('No of times')
plt.show()


plt.hist(dat['x4'],bins=30)            ##############plot histogram
plt.ylabel('No of times')
plt.show()

sns.boxplot(y=dat['x1'])

dat.min()
dat.max()

sns.boxplot(x=dat['class'],y=dat['x2'])

dat.pstdev()

sns.boxplot(data=dat.ix[:,0:4])

sns.boxplot(x=dat['class'],y=dat['x2'])
```

IPython console

Console 1/A




```
In [15]: plt.hist(dat['x4'],bins=30)          ##############plot histogram
    ...: plt.ylabel('No of times')
    ...: plt.show()
```




```
In [16]:
```

IPython console | History log

Permissions: RW End-of-lines: LF Encoding:

File Edit Search Source Run Debug Consoles Projects Tools View Help

/home/srushti/BE Sem1/my/LP1/DA/1

Editor - /home/srushti/BE Sem1/my/LP1/DA/1/code.py

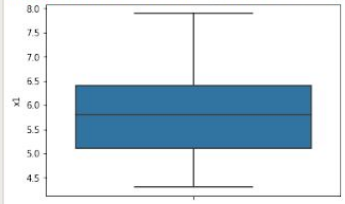
c4.py | code.py*

```
dat[0:10]


dat.shape
list(dat.columns)

dat.dtypes

dat['x1'].describe()
dat['x2'].describe()
dat['x3'].describe()
dat['x4'].describe()

dat.mean()


plt.hist(dat['x1'],bins=30)            #############plot histogram
plt.ylabel('No of times')
plt.show()


plt.hist(dat['x2'],bins=30)            #############plot histogram
plt.ylabel('No of times')
plt.show()


plt.hist(dat['x3'],bins=30)            #############plot histogram
plt.ylabel('No of times')
plt.show()


plt.hist(dat['x4'],bins=30)            #############plot histogram
plt.ylabel('No of times')
plt.show()

sns.boxplot(y=dat['x1'])
sns.boxplot(y=dat['x2'])
sns.boxplot(y=dat['x3'])
sns.boxplot(y=dat['x4'])

dat.min()
dat.max()

sns.boxplot(x=dat['class'],y=dat['x2'])

dat.pstdev()

sns.boxplot(data=dat.ix[:,0:4])
```

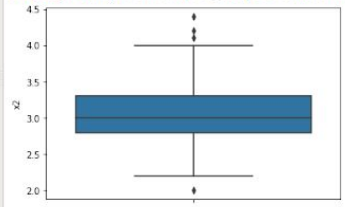
IPython console

Console 1/A

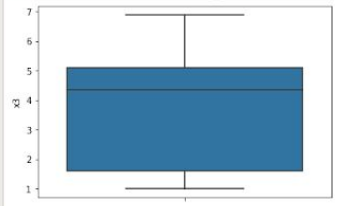
```
In [16]: sns.boxplot(y=dat['x1'])
Out[16]: <matplotlib.axes._subplots.AxesSubplot at 0x7fc0e5fdff10>
```




```
In [17]: sns.boxplot(y=dat['x2'])
Out[17]: <matplotlib.axes._subplots.AxesSubplot at 0x7fc0e600c250>
```




```
In [18]: sns.boxplot(y=dat['x3'])
Out[18]: <matplotlib.axes._subplots.AxesSubplot at 0x7fc0e61459d0>
```



IPython console | History log

Permissions: RW End-of-lines: LF E

File Edit Search Source Run Debug Consoles Projects Tools View Help

/home/srushti/BE Sem1/my/LP1/DA/1

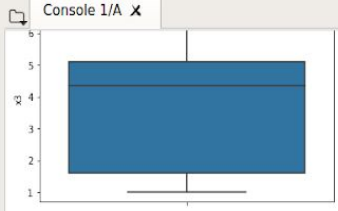
Editor - /home/srushti/BE Sem1/my/LP1/DA/1/code.py

c4.py | code.py*

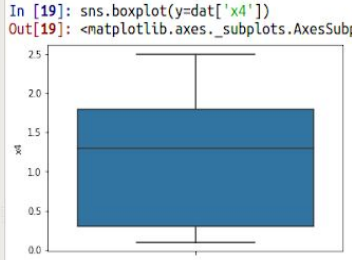
```
dat.shape
list(dat.columns)

dat.dtypes

dat['x1'].describe()
dat['x2'].describe()
dat['x3'].describe()
dat['x4'].describe()

dat.mean()


plt.hist(dat['x1'],bins=30)            #############plot histogram
plt.ylabel('No of times')
plt.show()


plt.hist(dat['x2'],bins=30)            #############plot histogram
plt.ylabel('No of times')
plt.show()


plt.hist(dat['x3'],bins=30)            #############plot histogram
plt.ylabel('No of times')
plt.show()


plt.hist(dat['x4'],bins=30)            #############plot histogram
plt.ylabel('No of times')
plt.show()

sns.boxplot(y=dat['x1'])
sns.boxplot(y=dat['x2'])
sns.boxplot(y=dat['x3'])
sns.boxplot(y=dat['x4'])


dat.max()
dat.min()
sns.boxplot(x=dat['class'],y=dat['x2'])

dat.pstdev()

sns.boxplot(data=dat.ix[:,0:4])

sns.boxplot(x=dat['class'],y=dat['x2'])

```

IPython console

Console 1/A




```
In [19]: sns.boxplot(y=dat['x4'])
Out[19]: <matplotlib.axes._subplots.AxesSubplot at 0x7fc0e65f6950>
```




```
In [20]: dat.min()
    ...: dat.max()
Out[20]:
x1                 7.9
x2                 4.4
x3                 6.9
x4                 2.5
class    Iris-virginica
dtype: object

In [21]: dat.min()
Out[21]:
x1               4.3
x2                 2
x3                 1
x4               0.1
class    Iris-setosa
dtype: object

In [22]: sns.boxplot(x=dat['class'],y=dat['x2'])
```

IPython console | History log

Permissions: RW End-of-lines: LF Enc

File Edit Search Source Run Debug Consoles Projects Tools View Help

/home/srushti/BE Sem1/my/LP1/DA/1

Editor - /home/srushti/BE Sem1/my/LP1/DA/1/code.py

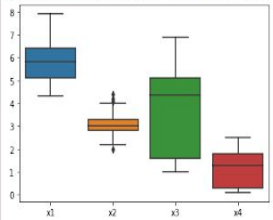
c4.py | code.py*

```
dat.shape
list(dat.columns)

dat.dtypes

dat['x1'].describe()
dat['x2'].describe()
dat['x3'].describe()
dat['x4'].describe()

dat.mean()


plt.hist(dat['x1'],bins=30)            #############plot histogram
plt.ylabel('No of times')
plt.show()


plt.hist(dat['x2'],bins=30)            #############plot histogram
plt.ylabel('No of times')
plt.show()


plt.hist(dat['x3'],bins=30)            #############plot histogram
plt.ylabel('No of times')
plt.show()


plt.hist(dat['x4'],bins=30)            #############plot histogram
plt.ylabel('No of times')
plt.show()

sns.boxplot(y=dat['x1'])
sns.boxplot(y=dat['x2'])
sns.boxplot(y=dat['x3'])
sns.boxplot(y=dat['x4'])


dat.max()
dat.min()
sns.boxplot(x=dat['class'],y=dat['x2'])

dat.pstdev()

sns.boxplot(data=dat.ix[:,0:4])

sns.boxplot(x=dat['class'],y=dat['x2'])
```

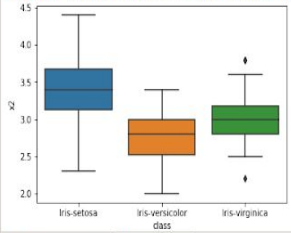
IPython console

Console 1/A

```
In [24]: sns.boxplot(data=dat.ix[:,0:4])
__main__:1: FutureWarning:
.ix is deprecated. Please use
.loc for label based indexing or
.iloc for positional indexing

See the documentation here:
http://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#ix-indexer-is-deprecated
/home/srushti/anaconda3/lib/python3.7/site-packages/pandas/core/indexing.py:822: FutureWarning:
.ix is deprecated. Please use
.loc for label based indexing or
.iloc for positional indexing

See the documentation here:
http://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#ix-indexer-is-deprecated
  retval = getattr(retval, self.name)._getitem_axis(key, axis=i)
Out[24]: <matplotlib.axes._subplots.AxesSubplot at 0x7fc0e5ef32d0>
```




```
In [25]: sns.boxplot(x=dat['class'],y=dat['x2'])
Out[25]: <matplotlib.axes._subplots.AxesSubplot at 0x7fc0e5e98b50>
```



IPython console | History log

Permissions: RW End-of-lines: LF Encoding: UTF-8 Line: 61

File Edit Search Source Run Debug Consoles Projects Tools View Help

/home/srushti/BE Sem1/my/LP1/DA/1

Editor - /home/srushti/BE Sem1/my/LP1/DA/1/code.py

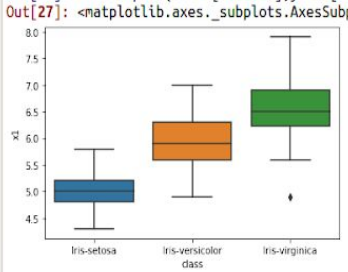
c4.py | code.py*

```
dat.dtypes

dat['x1'].describe()
dat['x2'].describe()
dat['x3'].describe()
dat['x4'].describe()

dat.mean()


plt.hist(dat['x1'],bins=30)            #############plot histogram
plt.ylabel('No of times')
plt.show()


plt.hist(dat['x2'],bins=30)            #############plot histogram
plt.ylabel('No of times')
plt.show()


plt.hist(dat['x3'],bins=30)            #############plot histogram
plt.ylabel('No of times')
plt.show()


plt.hist(dat['x4'],bins=30)            #############plot histogram
plt.ylabel('No of times')
plt.show()

sns.boxplot(y=dat['x1'])
sns.boxplot(y=dat['x2'])
sns.boxplot(y=dat['x3'])
sns.boxplot(y=dat['x4'])


dat.max()
dat.min()
sns.boxplot(x=dat['class'],y=dat['x2'])

dat.pstdev()

sns.boxplot(data=dat.ix[:,0:4])

sns.boxplot(x=dat['class'],y=dat['x1'])

sns.boxplot(x=dat['class'],y=dat['x3'])

sns.boxplot(x=dat['class'],y=dat['x4'])
```

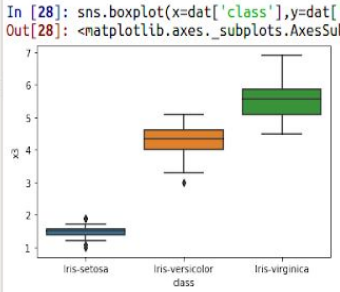
IPython console

Console 1/A

```
In [27]: sns.boxplot(x=dat['class'],y=dat['x1'])
Out[27]: <matplotlib.axes._subplots.AxesSubplot at 0x7fc0e5d4b210>
```




```
In [28]: sns.boxplot(x=dat['class'],y=dat['x3'])
Out[28]: <matplotlib.axes._subplots.AxesSubplot at 0x7fc0e5cd5450>
```




```
In [29]:
```

IPython console | History log

Permissions: RW End-of-lines: LF Enc