Assignment 3

Data Visualization and Pre-processing

Description:-

```
Predicting the age of abalone from physical measurements. The age of abalone is determined
```

Building a Regression Model

Double-click (or enter) to edit

1. Perform Below Visualizations.

Univariate Analysis

1. Summary Statistics

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import statsmodels.api as sm
```

```
file_data = pd.read_csv('C:/KavinKumar/abalone.csv')
file_data
```

	Sex	Length	Diameter	Height	Whole weight	Shucked weight	Visc wei
0	M	0.455	0.365	0.095	0.5140	0.2245	0.1
1	M	0.350	0.265	0.090	0.2255	0.0995	0.0
2	F	0.530	0.420	0.135	0.6770	0.2565	0.1
3	M	0.440	0.365	0.125	0.5160	0.2155	0.1
4	I	0.330	0.255	0.080	0.2050	0.0895	0.0

Add a Age column in a dataset

```
file_data['Age']=''
file_data.head()
```

	Sex	Length	Diameter	Height	Whole weight	Shucked weight	Viscera weight
0	M	0.455	0.365	0.095	0.5140	0.2245	0.1010
1	M	0.350	0.265	0.090	0.2255	0.0995	0.0485
2	F	0.530	0.420	0.135	0.6770	0.2565	0.1415
3	M	0.440	0.365	0.125	0.5160	0.2155	0.1140
4	I	0.330	0.255	0.080	0.2050	0.0895	0.0395

```
file_data['Age']=file_data['Rings']+1.5
file_data.head()
```

	Sex	Length	Diameter	Height	Whole weight	Shucked weight	Viscera weight
0	M	0.455	0.365	0.095	0.5140	0.2245	0.1010
1	M	0.350	0.265	0.090	0.2255	0.0995	0.0485
2	F	0.530	0.420	0.135	0.6770	0.2565	0.1415
3	M	0.440	0.365	0.125	0.5160	0.2155	0.1140
4	I	0.330	0.255	0.080	0.2050	0.0895	0.0395

Drop the Rings Column

```
file_data = file_data.drop(columns=['Rings'],axis=1)
```

```
file_data
```

	Sex	Length	Diameter	Height	Whole weight	Shucked weight	Viscera wei
0	M	0.455	0.365	0.095	0.5140	0.2245	0.1
1	M	0.350	0.265	0.090	0.2255	0.0995	0.0
2	F	0.530	0.420	0.135	0.6770	0.2565	0.1
3	M	0.440	0.365	0.125	0.5160	0.2155	0.1
4	I	0.330	0.255	0.080	0.2050	0.0895	0.0
...	...	...	...	...	...	...	
4172	F	0.565	0.450	0.165	0.8870	0.3700	0.2
4173	M	0.590	0.440	0.135	0.9660	0.4390	0.2
4174	M	0.600	0.475	0.205	1.1760	0.5255	0.2
4175	F	0.625	0.485	0.150	1.0945	0.5310	0.2
4176	M	0.710	0.555	0.195	1.9485	0.9455	0.3

4177 rows × 9 columns

```
file_data['Height'].mean()
```

```
0.1395163993296614
```

```
file_data['Height'].median()
```

```
0.14
```

```
file_data['Height'].std()
```

```
0.04182705660725703
```

2. Frequency Table

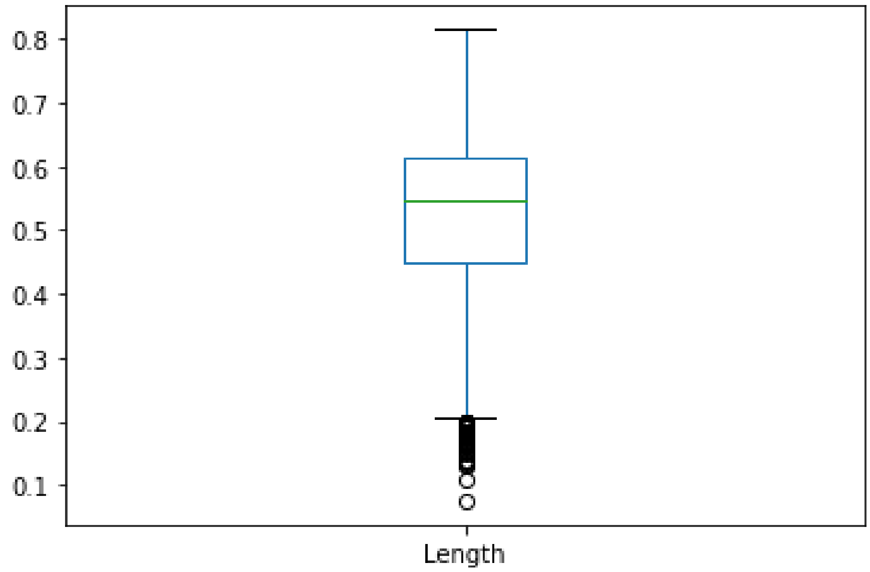
```
file_data['Sex'].value_counts()
```

```
M    1528
I    1342
F    1307
Name: Sex, dtype: int64
```

3. Create Charts

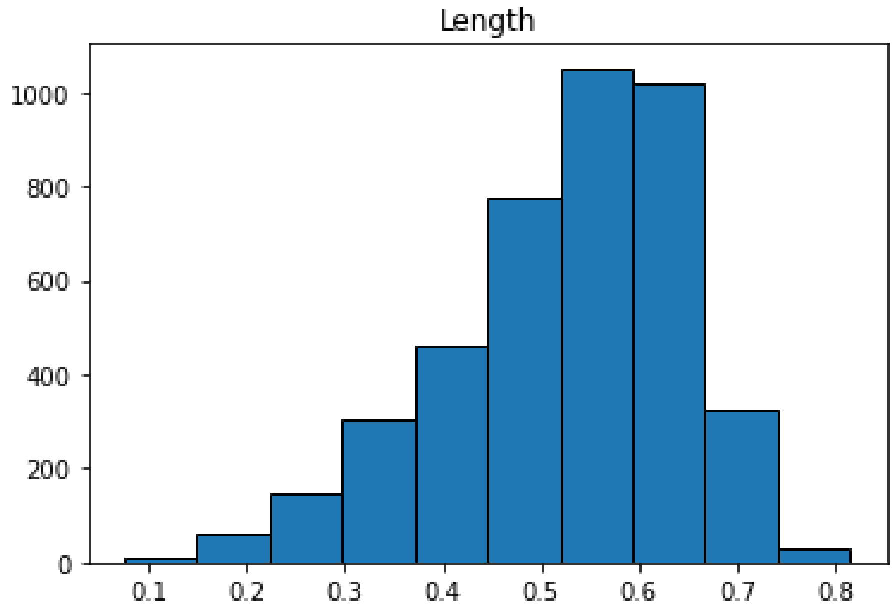
```
file_data.boxplot(column=['Length'], grid=False)
```

```
<AxesSubplot:>
```



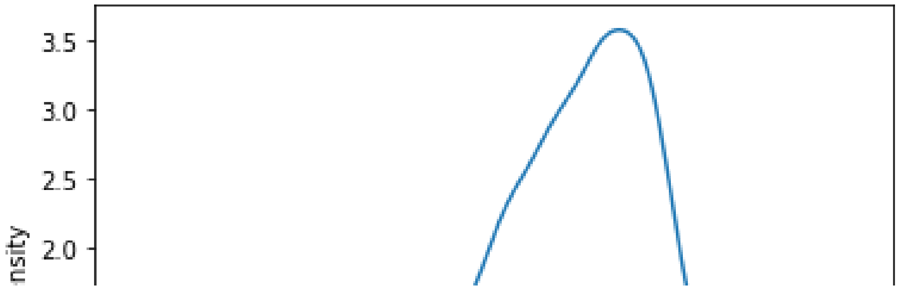
```
file_data.hist(column='Length', grid=False, edgecolor='black')
```

```
array([[<AxesSubplot:title={'center':'Length'}>]], dtype=object)
```



```
sns.kdeplot(file_data['Length'])
```

```
<AxesSubplot:xlabel='Length', ylabel='Density'>
```



Bi - Variate Analysis

[ ] ↳ *8 cells hidden*

Multi - Variate Analysis

[ ] ↳ *1 cell hidden*

4. Perform descriptive statistics on the dataset.

[ ] ↳ *14 cells hidden*

5. Handle the Missing values.

[ ] ↳ *3 cells hidden*

6. Find the outliers and replace the outliers

[ ] ↳ *4 cells hidden*

7. Check for Categorical columns and perform encoding.

[ ] ↳ *3 cells hidden*

8. Split the data into dependent and independent variables.

[ ] ↳ *2 cells hidden*

9. Scale the independent variables

[ ] ↳ *2 cells hidden*

10. Split the data into training and testing

[ ] ↳ *7 cells hidden*

11. Build the Model

[ ] ↳ *3 cells hidden*

12.Train the Model

[ ] ↳ *1 cell hidden*

13.Test the Model

[ ] ↳ *1 cell hidden*

14. Measure the performance using Metrics

[ ] ↳ *1 cell hidden*

Colab paid products - Cancel contracts here