DATA SCIENCE LAB MANUAL

1. **Consider the following data of three cricket players in 10 innings T20 Match**

Player	1	2	3	4	5	6	7	8	9	10
Cricketer1	25	10	55	45	55	78	55	0	49	10
Cricketer2	47	62	78	45	100	20	100	0	80	10
Cricketer3	80	17	7	10	45	79	75	75	80	42

a) Find Whose average is better.
b) What is the middlemost value of each player?
c) Whose most frequent value is good.
d) Draw a simple plot to show performance of players.

Solution:

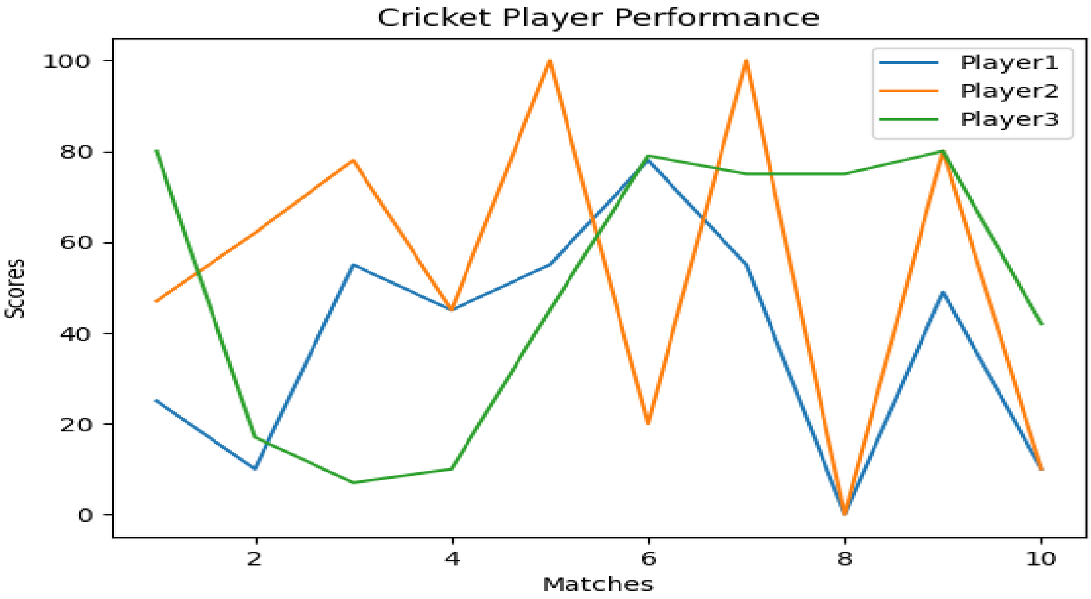
```
#Cricket Player Performance Analysis
import statistics as st
import matplotlib.pyplot as pt
import tabulate
Matches=[1,2,3,4,5,6,7,8,9,10]
Player1=[25,10,55,45,55,78,55,0,49,10]
Player2=[47,62,78,45,100,20,100,0,80,10]
Player3=[80,17,7,10,45,79,75,75,80,42]
#Player1 Summary
print("Player1 Mean = ",st.mean(Player1))
print("Player1 Median = ",st.median(Player1))
print("Player1 Mode = ",st.mode(Player1))
#Player2 Summary
print("Player2 Mean = ",st.mean(Player2))
print("Player2 Median = ",st.median(Player2))
print("Player2 Mode = ",st.mode(Player2))
#Player3 Summary
print("Player3 Mean = ",st.mean(Player3))
print("Player3 Median = ",st.median(Player3))
print("Player3 Mode = ",st.mode(Player3))
#Performance plot
pt.plot(Matches,Player1)
pt.plot(Matches,Player2)
pt.plot(Matches,Player3)
pt.title("Cricket Player Performance")
pt.xlabel("Matches")
pt.ylabel("Scores")
pt.legend(["Player1","Player2","Player3"])
pt.show()
```

OUTPUT:

```
Player1 Mean =  38.2

Player1 Median =  47.0

Player1 Mode =  55

Player2 Mean =  54.2

Player2 Median =  54.5

Player2 Mode =  100

Player3 Mean =  51

Player3 Median =  60.0

Player3 Mode =  80
```

Analysis

a) Player 2 average is better.
b) Player1 Median = 47.0, Player2 Median = 54.5, Player3 Median = 60.0
c) Player2
d) Draw a simple plot to show performance of players.



2. Consider Insurance Dataset and analyze following
    a) Count Number of Male and Female
    b) What is average age of peoples.
    c) Display simple bar plot Gender wise

Solution:

```
import pandas as pd
import openpyxl
import statistics as st
import matplotlib.pyplot as pt
data = pd.read_csv("E:\Data Science with Python\DataSet\insurance.csv")
print(data)
#Analysis genderwise
ls=data['sex'].tolist()
y1=ls.count('female')
y2=ls.count('male')
print("female Count = ",y1)
print("male Count = ",y2)

#Aveage age of customers
avgage=data['age'].tolist()
print("Average Age=  %.2f " % st.mean(avgage))

#Display Histogram genderwise
x=["FEMALE","MALE"]
```

```
y=[y1,y2]
pt.bar(x,y)
pt.title("Genderwise Insurance Data")
pt.xlabel("Gender")
pt.ylabel("Count")
pt.show()
```

Analysis:

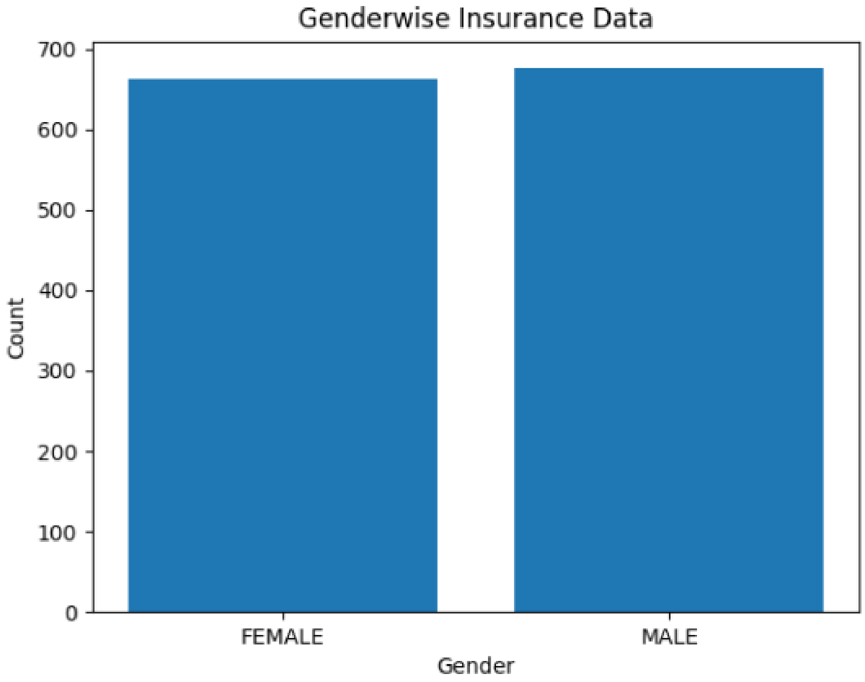
a)

female Count = 662

male Count = 676

b) Average Age= 39.21

c)



3. **Consider Insurance Dataset and analyze data region wise. Also display simple bar chart region wise.**

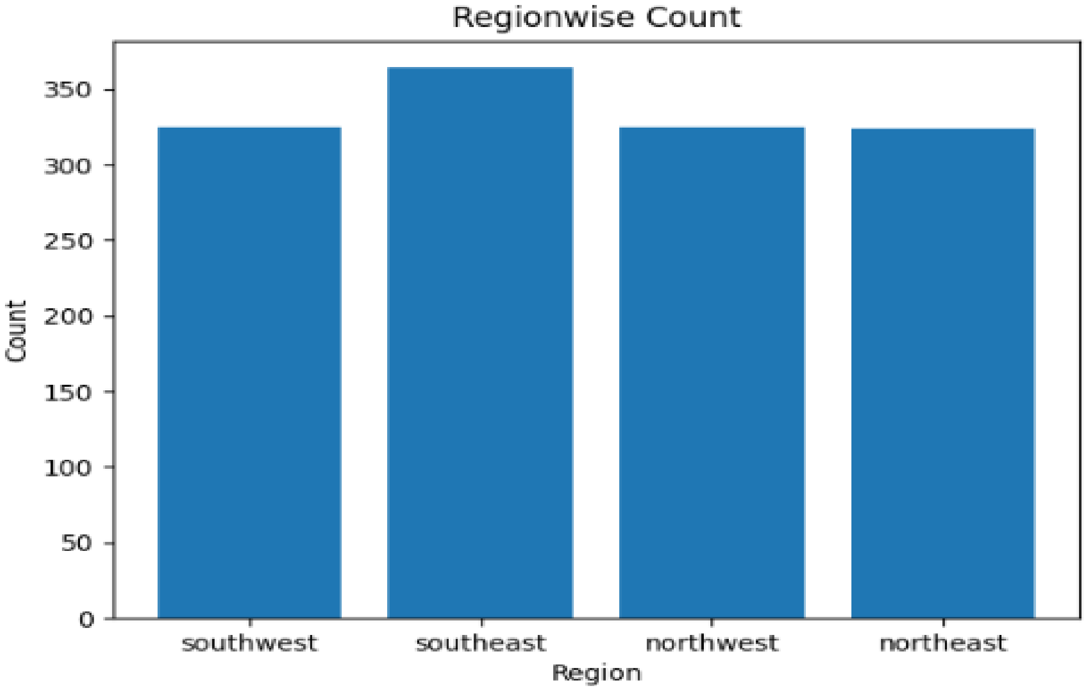
Solution:

```
import pandas as pd
import openpyxl
import matplotlib.pyplot as pt
data = pd.read_csv("E:\Data Science with Python\DataSet\insurance.csv")
print(data)

#Regionwise count
region=data['region'].tolist()
output=[]
for x in region:
    if x not in output:
        output.append(x)
print(output)
y1=region.count('southwest')
y2=region.count('southeast')
```

```
y3=region.count('northwest')
y4=region.count('northeast')
print("Southwest count= ",y1)
print("southeast count= ",y2)
print("northwest count= ",y3)
print("northeast count= ",y4)
pt.title("Regionwise Count")
pt.xlabel("Region")
pt.ylabel("Count")
y=[y1,y2,y3,y4]
pt.bar(output,y)
pt.show()
```

Analysis:

```
Southwest count=  325
southeast count=  364
northwest count=  325
northeast count=  324
```



4. Consider temperature dataset and analyze average of minimum and maximum temperature, minimum temperature, maximum temperature month wise.

   Solution:

```
import pandas as pd
import openpyxl
import numpy as np
data=pd.read_excel("E:\\Data Science with
Python\\DataSet\\belgavitemp2022.xlsx")
print(data)
df1 = (data.groupby(["Year",
"Month"],sort=False).agg(Avg_of_Max_Temp=("Max", 'mean'),
   Max_temp=("Max",'max'),Avg_of_Min_Temp=("Min",
'mean'),Min_temp=("Min",'min')))
print(df1)
```

Analysis:

```
               Avg_of_Max_Temp  Max_temp  Avg_of_Min_Temp  Min_temp
Year Month
2022 January         29.290323        33        14.838710        11
         February    32.535714        35        16.928571        14
         March       35.451613        39        20.322581        17
        April        36.666667        39        22.300000        19
         May         33.838710        38        21.612903        19
        June         31.533333        36        21.033333        20
        July         28.225806        33        20.451613        19
       August        28.419355        32        20.258065        19
       September     29.533333        32        19.833333        18
       October       29.741935        32        18.677419        14
       November      30.433333        32        16.433333        11
       December      29.870968        33        17.967742        14
```

5.Consider following data and calculate Descriptive statistics using formules.

22,26,14,30,18,1135,41,12,32

Solution:

```
import numpy as np
import pandas as pd
data=[22,26,14,30,18,11,35,41,12,32]
print("Mean = %.2f"% np.mean(data))
print("Median = ",np.median(data))
print("Max = ",np.max(data))
print("Min = ",np.min(data))
print("First Quartile =",np.quantile(data,0.25))
print("Second Quartile = ",np.quantile(data,0.50))
print("Third Quartile = ",np.quantile(data,0.75))
print("20 th Percentilee = ",np.percentile(data,20))
print("99 th Percentilee = ",np.percentile(data,99))
print("Standard deviation = %.2f" % np.std(data))
print("Variance  = ",np.var(data))
```

OUTPUT:

```
Mean = 24.10
Median =  24.0
Max =  41
Min =  11
First Quartile = 15.0
Second Quartile =  24.0
Third Quartile =  31.5
20 th Percentilee =  13.6
99 th Percentilee =  40.46
Standard deviation = 9.83
Variance  =  96.69
```

6. Find the Quartiles for the following Students Score data and visualize graphically.
   50,50,47,97,49,3,53,42,26,74,82,62,37,15,70,27,36,35,48,52,63,64.

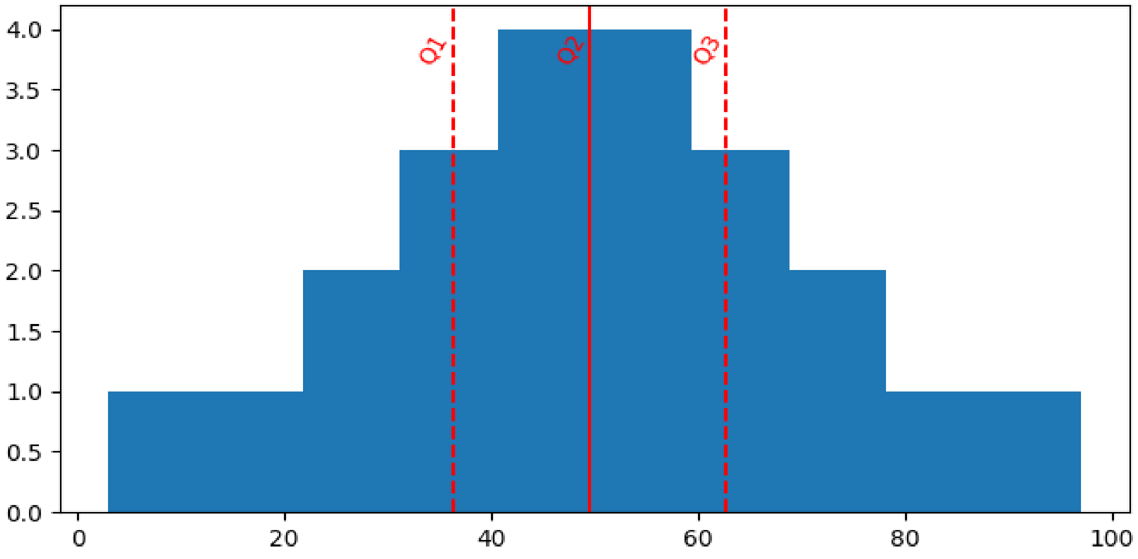
Solution:

```
import numpy as np
import matplotlib.pyplot as pt
import numpy as np
import pandas as pd
```

```
data=[50,50,47,97,49,3,53,42,26,74,82,62,37,15,70,27,36,35,48,52,63,64]
print(data)
print("Quartile 1 = %.2f"%np.quantile(data,0.25))
print("Quartile 2 = %.2f"%np.quantile(data,0.50))
print("Quartile 3 = %.2f"%np.quantile(data,0.75))
pt.figure(figsize=(8,4))
pt.hist(data)
# Vertical lines for each percentile of interest
pt.axvline(np.quantile(data, 0.25), linestyle='--', color='red')
pt.text(np.quantile(data, 0.25), 4, 'Q1', color='r', ha='right', va='top',
rotation=60)
pt.axvline(np.quantile(data, 0.50), linestyle='-',  color='red')
pt.text(np.quantile(data, 0.50), 4, 'Q2', color='r', ha='right', va='top',
rotation=60)
pt.axvline(np.quantile(data, 0.75), linestyle='--', color='red')
pt.text(np.quantile(data, 0.75), 4, 'Q3', color='r', ha='right', va='top',
rotation=60)
pt.show()
```

OUTPUT:

```
Quartile 1 = 36.25

Quartile 2 = 49.50

Quartile 3 = 62.75
```



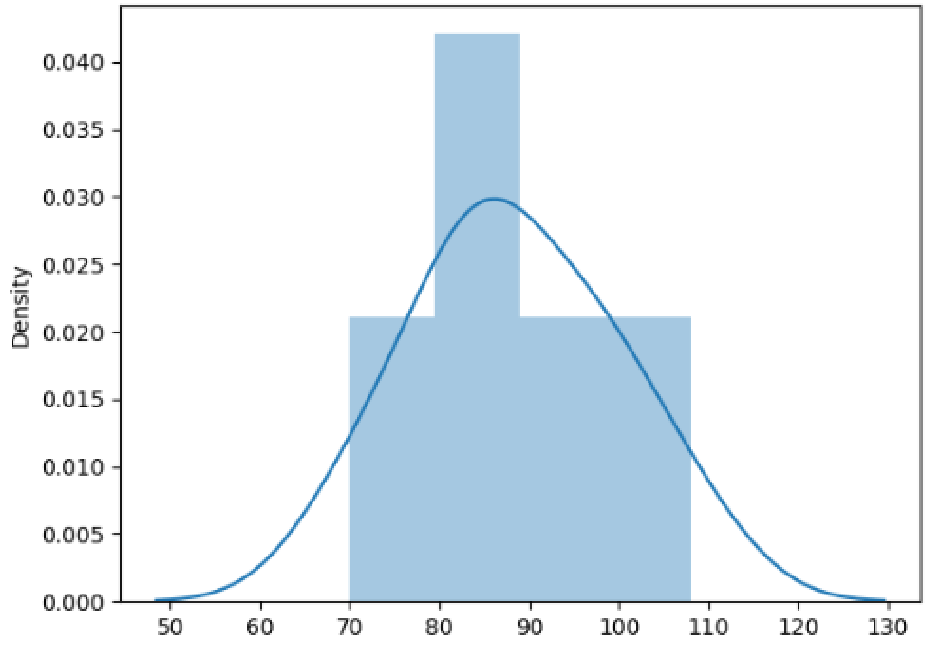
7. Calculate the skewness for the following data also conclude skewness 85,96,76,108,84,100,86,70,95,84

Solution

```
# Importing library
import matplotlib.pyplot as pt
import statistics as st
import seaborn as sns
# Creating a dataset
dataset =[85,96,76,108,84,100,86,70,95,84]
meandata=st.mean(dataset)
print("Mean =  %.2f"%meandata)
modedata=st.mode(dataset)
print("Mode =  %.2f"%modedata)
meddata=st.median(dataset)
print("Median =  %.2f"%meddata)
# Calculate the skewness
stddata=st.stdev(dataset)
print("Standard Deviation =%.2f" % stddata)
sk=(meandata-modedata)/stddata
print("Skewness= %.2f" % sk)
sns.distplot(dataset)
pt.show()
```

OUTPUT:

```
Mean =  88.40
Mode =  84.00
Median =  85.50
```



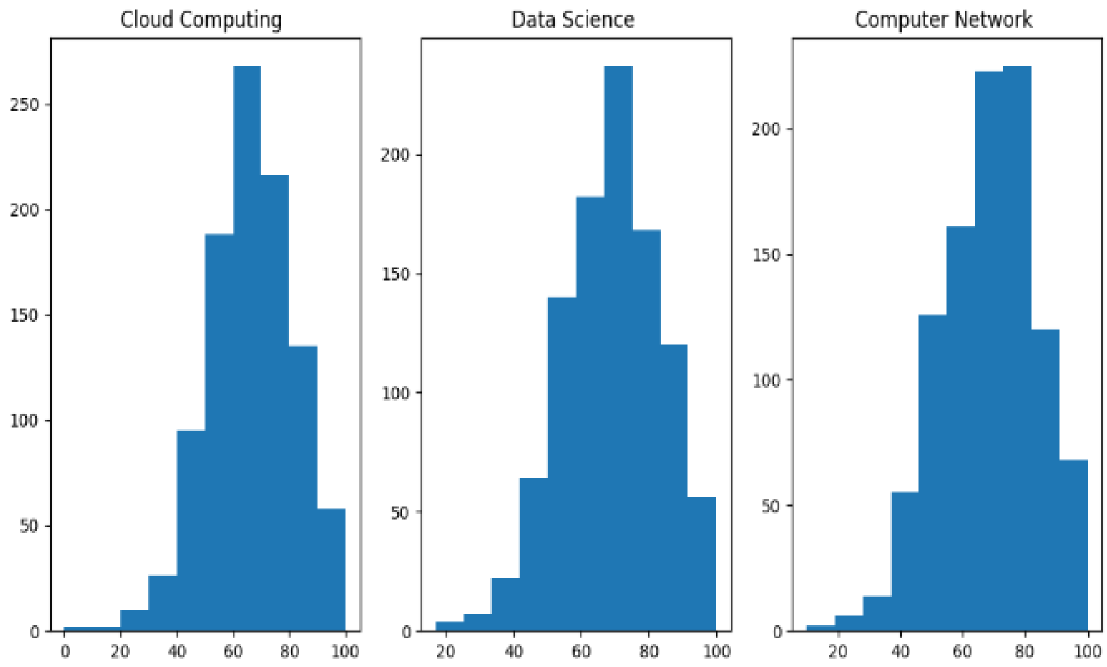
Analysis: Distribution is Positively Skewed.

8. Consider Student Performance dataset and find skewness for all subjects.

```
import pandas as pd
import matplotlib.pyplot as plt
import openpyxl
data =pd.read_csv("E:\Data Science with
Python\DataSet\StudentsPerformance.csv")
print(data)
print("Skew of Cloud Computing score:
%.2f"%data['Cloud Computing'].skew())
print("Skew of Data Science: %.2f"%data['Data
Science'].skew())
print("Skew of Computer Networks:
%.2f"%data['Computer Network'].skew())

plt.figure(figsize = (12,6))
plt.subplot(1, 3, 1)
plt.hist(data['Cloud Computing'])
plt.title('Cloud Computing ')

plt.subplot(1, 3, 2)
plt.hist(data['Data Science'])
plt.title('Data Science ')

plt.subplot(1,3,3)
plt.hist(data['Computer Network'])
plt.title('Computer Network ')

plt.show()
```

OUTPUT:

```
Skew of Cloud Computing score: -0.28
Skew of Data Science: -0.26
Skew of Computer Networks: -0.29
```



Analysis:

All subjects Distribution is negatively skewed.

Maximum students score between 60-100.

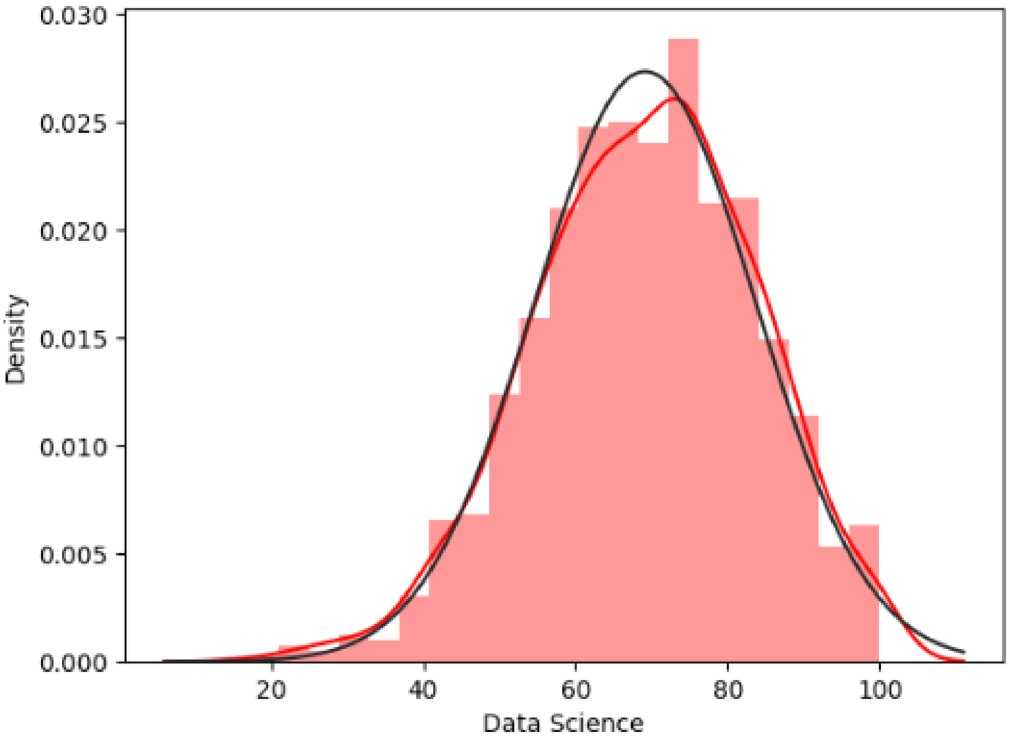
9. Consider Student Performance dataset find basic statistics of data science subject using pandas describe function, calculate skewness also visualize distribution.

Solution:

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from scipy.stats import skew, skewtest, norm
import openpyxl
data =pd.read_csv("E:\Data Science with
Python\DataSet\StudentsPerformance.csv")
print(data)
print(data['Data Science'].describe())
print("Skewness= %.2f"%data['Data Science'].skew())
sns.distplot(data['Data Science'], fit=norm, color="r")
plt.show()
```

OUTPUT:

```
count    1000.000000
mean       69.169000
std        14.600192
min        17.000000
25%        59.000000
50%        70.000000
75%        79.000000
max       100.000000
Name: Data Science, dtype: float64
Skewness= -0.26
```



0.030
0.025
0.020
0.015
0.010
0.005
0.000
Density
20
40
60
80
100
Data Science

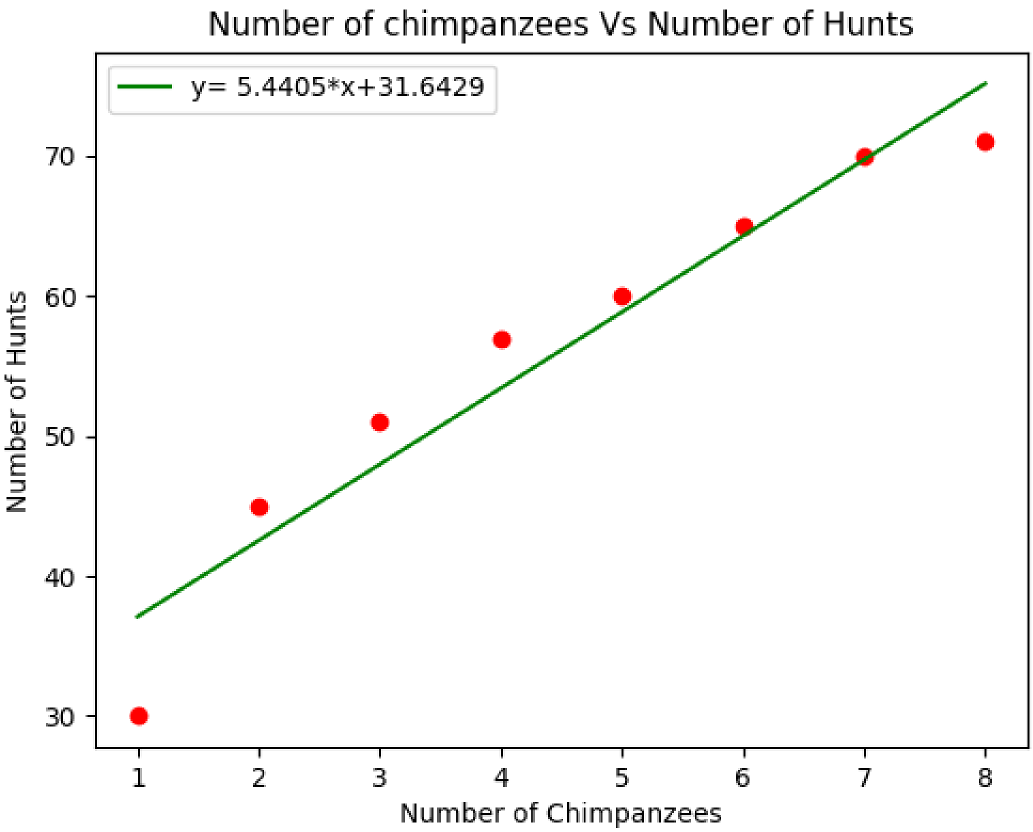
10.Draw Regression Line for the following data. Conclude your analysis.

No. of chimpanzees	1	2	3	4	5	6	7	8
No. of hunting	30	45	51	57	60	65	70	71

```
# Import packages
import numpy as np
import matplotlib.pyplot as plt
x= np.array([1,2,3,4,5,6,7,8])
# Dependent Variable - percent of successful hunts
y = np.array([30,45,51,57,60,65,70,71])
n = np.size(x)
x_mean = np.mean(x)
y_mean = np.mean(y)
b1=n * np.sum(x*y)-np.sum(x)*np.sum(y)
b2=(n * sum(x*x) - (np.sum(x)*np.sum(x)))
b=(b1/b2)
a= y_mean-b*x_mean
print("Line Slope is : %.4f"%b)
print("Line Intercept is: %.4f"%a)
y_pred=b*x+a
plt.scatter(x, y, color = 'red')
plt.plot(x, y_pred, color = 'green',label='y= 5.4405*x+31.6429')
plt.xlabel('Number of Chimpanzees')
plt.ylabel('Number of Hunts')
plt.title("Number of chimpanzees Vs Number of Hunts")
plt.legend()
plt.show()
```

OUTPUT:

Line Slope is : 5.4405
Line Intercept is: 31.6429



Analysis:
Positive Correlation exist between number of chipanzees and number of hunts.

[redacted]

[redacted]