DS1_C5_S1_PRACTICE

In [27]:

```
#Import the required liberary
import pandas as pd
import warnings
warnings.filterwarnings('ignore')
import matplotlib.pyplot as plt
import statistics as st
```

In [28]:

```
student = pd.read_csv('DS1_C4_S5_Students_Scores_Data_Practice.csv')
student
```

Out[28]:

	Name	Statistics	Python	Tableau
0	David	62	89	56
1	James	47	87	86
2	Robert	55	67	77
3	Thomas	74	55	45
4	Steven	31	47	73
5	Paul	77	72	62
6	Gary	85	76	74
7	Justin	63	79	89
8	Patrick	42	44	67
9	Tyler	32	99	67
10	Peter	71	99	97
11	Bryan	63	69	68

Task 1

In [29]:

```
sd = student.iloc[:,-3:]
sd
```

Out[29]:

	Statistics	Python	Tableau
0	62	89	56
1	47	87	86
2	55	67	77
3	74	55	45
4	31	47	73
5	77	72	62
6	85	76	74
7	63	79	89
8	42	44	67
9	32	99	67
10	71	99	97
11	63	69	68

In [30]:

```
sd.isnull().sum()
```

Out[30]:

```
Statistics    0
Python        0
Tableau       0
dtype: int64
```

In [31]:

```
for x in sd:
    print (x)
```

```
Statistics
Python
Tableau
```

In [33]:

```
#declare 3 list of mean ,median ,mode
mean =[]
mode =[]
median =[]

for col in sd:
    mean.append(st.mean(sd[col]))
    mode.append(st.mode(sd[col]))
    median.append(st.median(sd[col]))

row_head = ['mean', 'mode', 'median']
col_name = ['Statistics', 'Python', 'Tableau']

# create dataframe of mean , median ,mode
d_data = pd.DataFrame ([mean, mode, median],  columns = col_name)
d_data

# insert column
d_data.insert(0,"Measures", row_head)
d_data
```

Out[33]:

```
   Measures  Statistics     Python  Tableau
0      mean        58.5  73.583333    71.75
1      mode        63.0  99.000000    67.00
2    median        62.5  74.000000    70.50
```

Task 2

In [34]:

```
mean= []
SD =[]
CV=[]

# iterate  each column
for col in sd:
    col_mean= sd[col].mean()   #creating mean of each column
    mean.append(col_mean)              #storing the calculated mean in mean named folder
    col_std= sd[col].std()     #calculating standard deviation of each column
    SD.append(col_std)                 #storing the calculated SDin SD name folder
    CV.append(col_std/col_mean*100)

row_head = ['mean', 'SD', 'CV']
col_name = ['Statistics', 'Python', 'Tableau']

# create dataframe of mean , median ,mode
d_data1 = pd.DataFrame ([mean, SD, CV],  columns = col_name)
d_data1


# insert column
d_data1.insert(0,"Measures", row_head)
d_data1
```

Out[34]:

	Measures	Statistics	Python	Tableau
0	mean	58.500000	73.583333	71.750000
1	SD	17.500649	18.436418	14.429295
2	CV	29.915640	25.055155	20.110515

Task 3

In [45]:

```
for subject in sd:
    LO =[]
    UO =[]
    marks1 = pd.Series(sd[subject])
    Min = min(marks1)
    Max = max(marks1)
    Range = Max-Min

# calculate IQR
    Q1 = marks1.quantile(0.25)
    Q3 = marks1.quantile(0.75)
    IQR= Q3-Q1
    UF = Q3+1.5*IQR                    # Upper Fence
    LF = Q1-1.5*IQR                    # lower fence

#To check outlier and store in empty folder
    for marks2 in sd[subject]:
        if(marks2 < LF):
            LO.append(marks2)
        if(marks2 > UF):
            UO.append(marks2)

# Storing all information in folder
        if(subject == 'Statistics'):
            Statistics=['Statistics', Min, Max, Range, Q1, Q3, IQR, UF, LF , [LO,UO]]
        elif(subject == 'Python'):
            Python=['Python', Min, Max, Range, Q1, Q3, IQR, UF, LF , [LO,UO]]
        else :
            Tableau=['Tableau', Min, Max, Range, Q1, Q3, IQR, UF, LF , [LO,UO]]

#
col_names =[ 'Subject', 'Min','Max',' Range', 'Q1', 'Q3',' IQR', 'Lower Fence', 'Upper Fence' , 'Outlier']

d_data7 = pd.DataFrame([Statistics,Python,Tableau], columns = col_names)
d_data7
```

Out[45]:

	Subject	Min	Max	Range	Q1	Q3	IQR	Lower Fence	Upper Fence	Outlier
0	Statistics	31	85	54	45.75	71.75	26.0	110.75	6.75	[[], []]

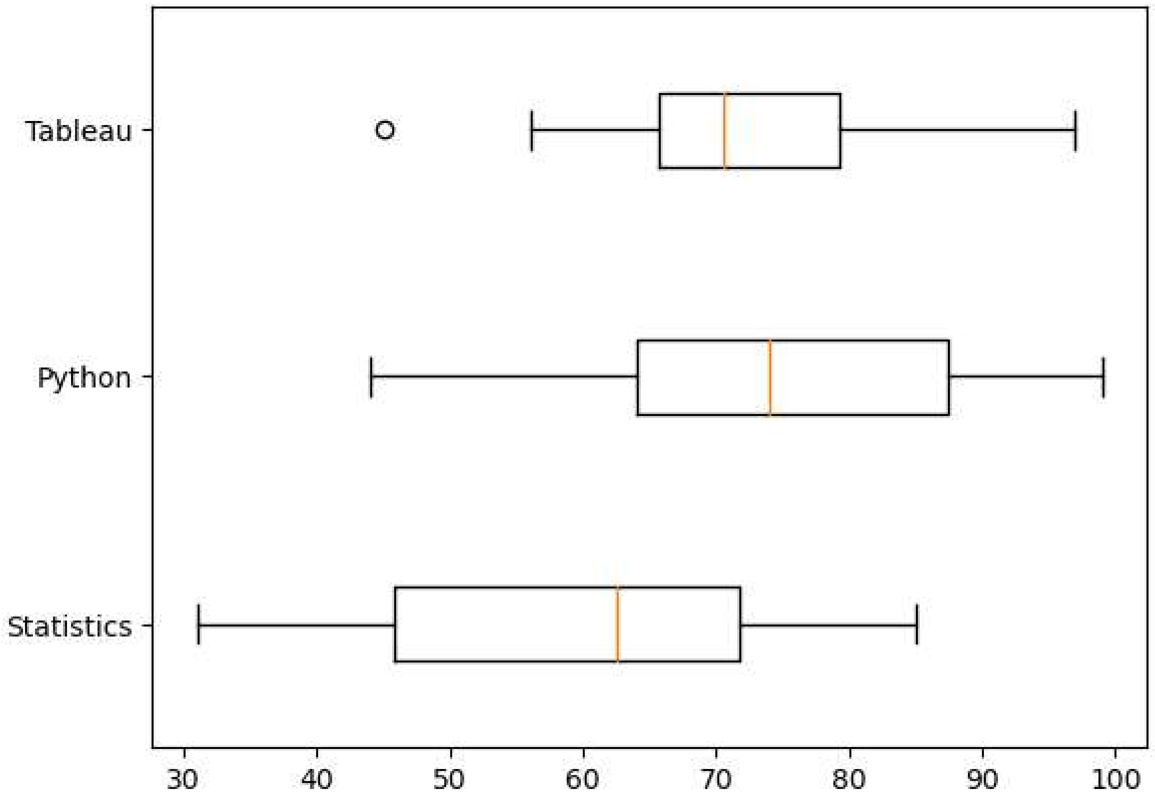
	Subject	Min	Max	Range	Q1	Q3	IQR	Lower Fence	Upper Fence	Outlier
1	Python	44	99	55	64.00	87.50	23.5	122.75	28.75	[[], []]
2	Tableau	45	97	52	65.75	79.25	13.5	99.50	45.50	[[45], []]

Task 4

In [47]:
```
Statistics_data = sd['Statistics'].tolist()
Python_data = sd['Python'].tolist()
Tableau_data = sd['Tableau'].tolist()
```

In [48]:
```
plt.boxplot([Statistics_data, Python_data, Tableau_data], vert=0)
plt.yticks([1,2,3],['Statistics', 'Python', 'Tableau'])
```

Out[48]:
```
([<matplotlib.axis.YTick at 0x2580f8edf40>,
  <matplotlib.axis.YTick at 0x2580f8edd90>,
  <matplotlib.axis.YTick at 0x2580f939880>],
 [Text(0, 1, 'Statistics'), Text(0, 2, 'Python'), Text(0, 3, 'Tableau')])
```



Task 5

In [52]:

```
from scipy.stats import norm
```

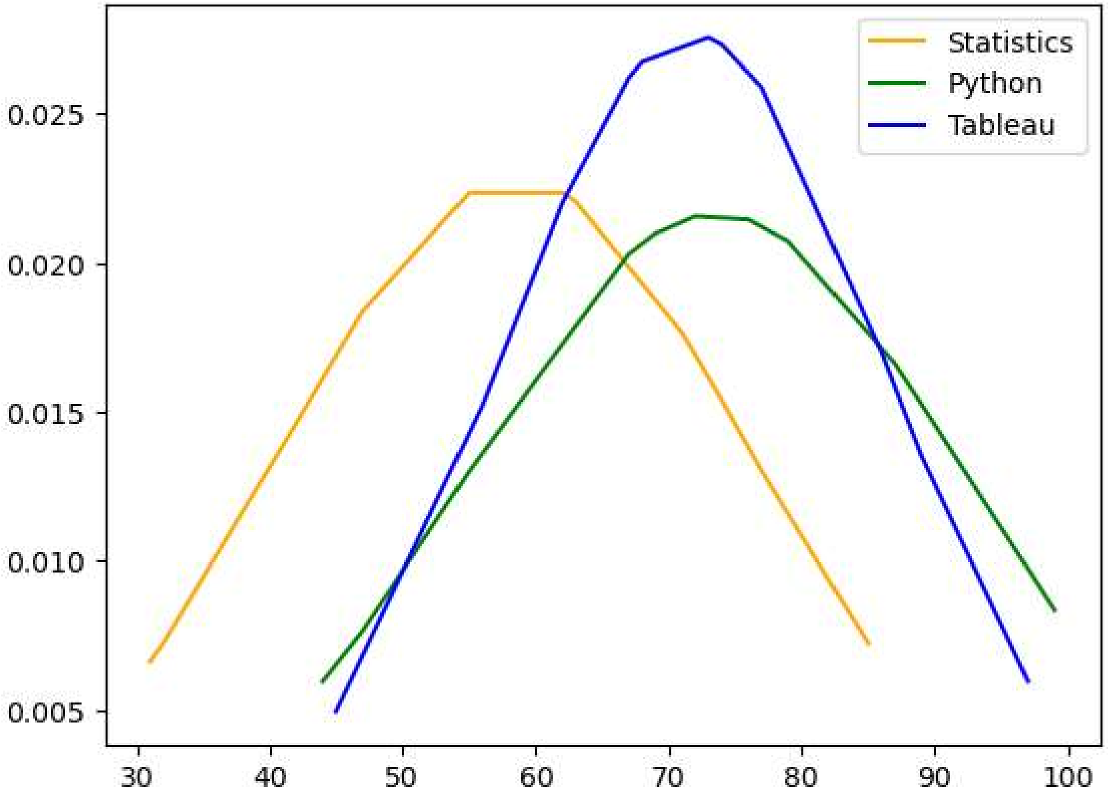
In [53]:
```
Statistics_data = sorted(Statistics_data)
Python_data=sorted(Python_data)
Tableau_data = sorted(Tableau_data)

I_mean = st.mean(Statistics_data)
I_sd = st.stdev(Statistics_data)
M_mean = st.mean(Python_data)
M_sd = st.stdev(Python_data)
A_mean = st.mean(Tableau_data)
A_sd = st.stdev(Tableau_data)

plt.plot(Statistics_data, norm.pdf(Statistics_data, I_mean, I_sd), color = 'orange', label = 'Statistics')
plt.plot(Python_data, norm.pdf(Python_data, M_mean, M_sd), color = 'green', label = 'Python')
plt.plot(Tableau_data, norm.pdf(Tableau_data, A_mean, A_sd), color = 'blue', label = 'Tableau')

plt.legend()
plt.show
```

Out[53]:
```
<function matplotlib.pyplot.show(close=None, block=None)>
```



In [54]:
```
from scipy.stats import kurtosis

print(kurtosis(Statistics_data))
print(kurtosis(Python_data))
print(kurtosis(Tableau_data))
```

```
-1.0438440476747421
-1.0159472252820538
-0.4356912494591376
```

Task 6

In [ ]:

```
"""In the data set only Tableau has only one outliers and Statistics has the large spread of data as the value
      of  standard deviation is greater than the tableau and python ,
      and from box plot we can see large spread of data in statistics .
      Tableau shows more curve nature .

"""
```