Experiment Number: 2

Batch: FDS-2 **Roll Number: 16010422234** **Name: Chandana Ramesh Galgali**

Aim of the Experiment: 1. Find out measures of central tendency for single and multi-attributes data using Statistical analysis

2. Find out measures of variability of data using statistical analysis

---

Program/ Steps:

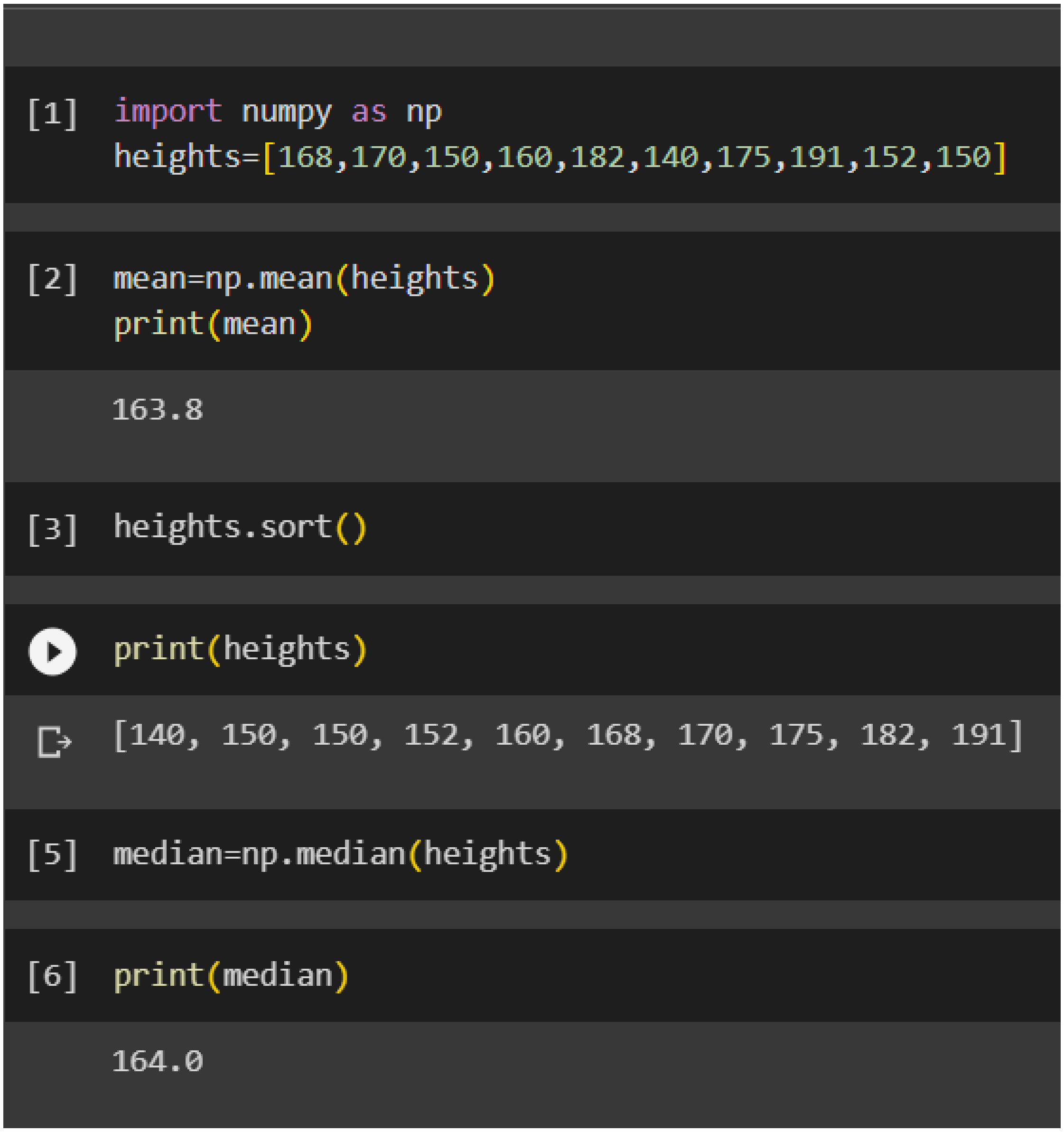
1. Analyze the data to find out frequency, mean, mode and median, standard deviation, variance, interquartile of data?

2. Write a programming code with a procedure to compute mean mode and median, variance, standard deviation, interquartile of a given sample data without using readymade function?

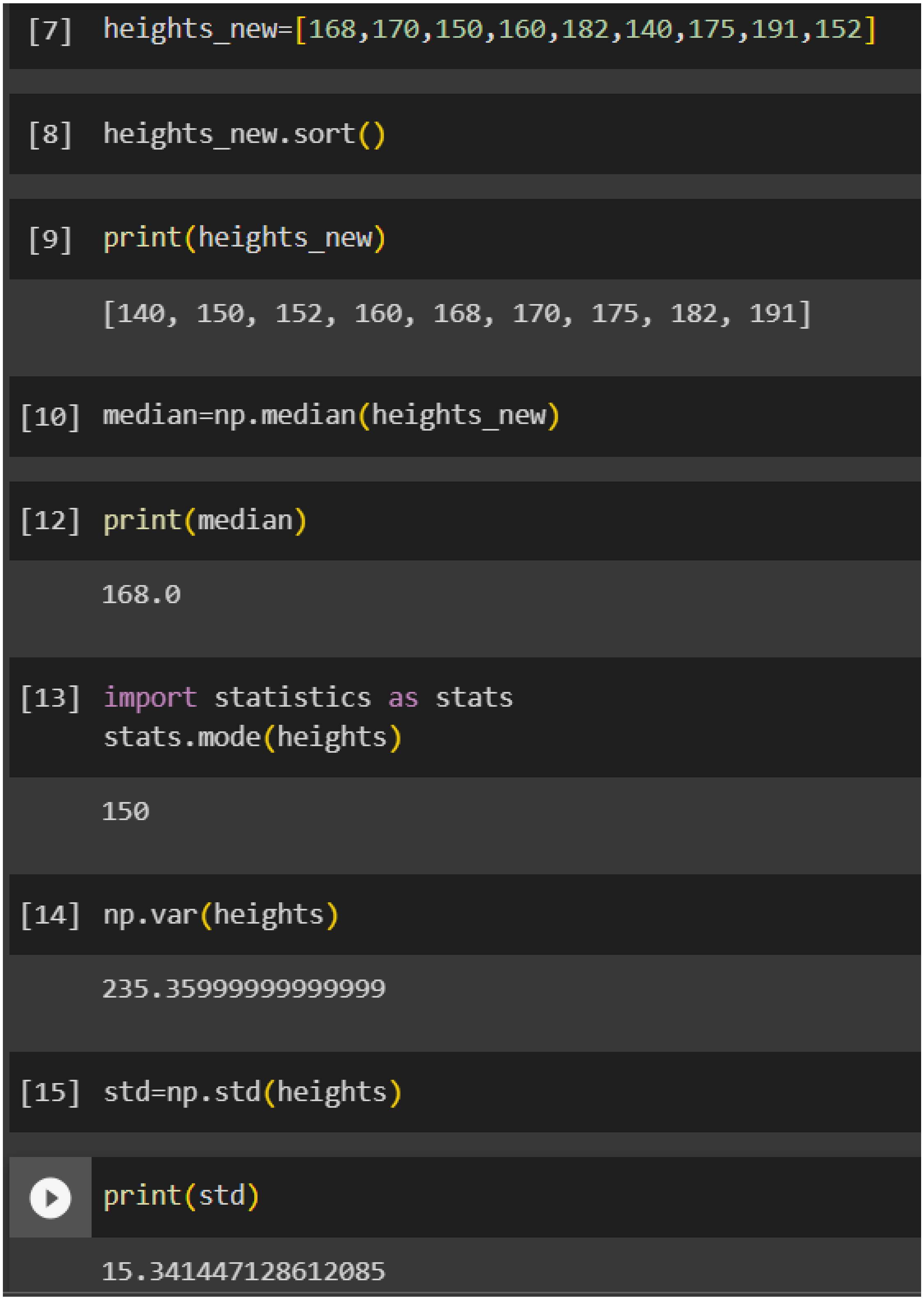
3. Compute and analyze the group data to find out frequency, mean, mode and median, standard deviation, variance, interquartile of data?

Code with Output/Result:

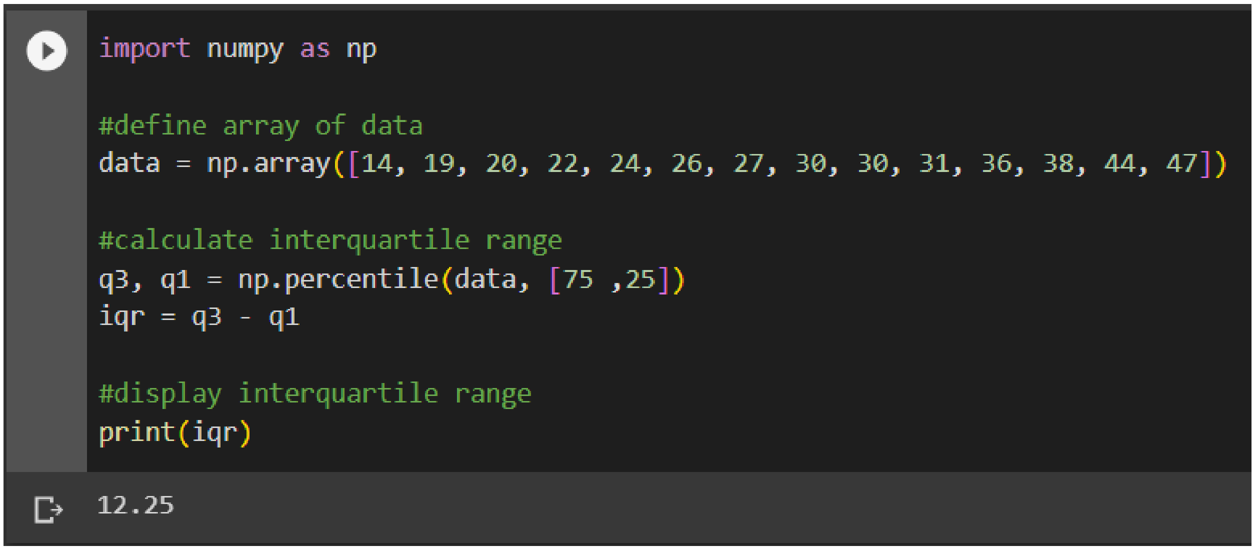
1)



```
[1]  import numpy as np
     heights=[168,170,150,160,182,140,175,191,152,150]

[2]  mean=np.mean(heights)
     print(mean)

     163.8

[3]  heights.sort()

     print(heights)

     [140, 150, 150, 152, 160, 168, 170, 175, 182, 191]

[5]  median=np.median(heights)

[6]  print(median)

     164.0
```

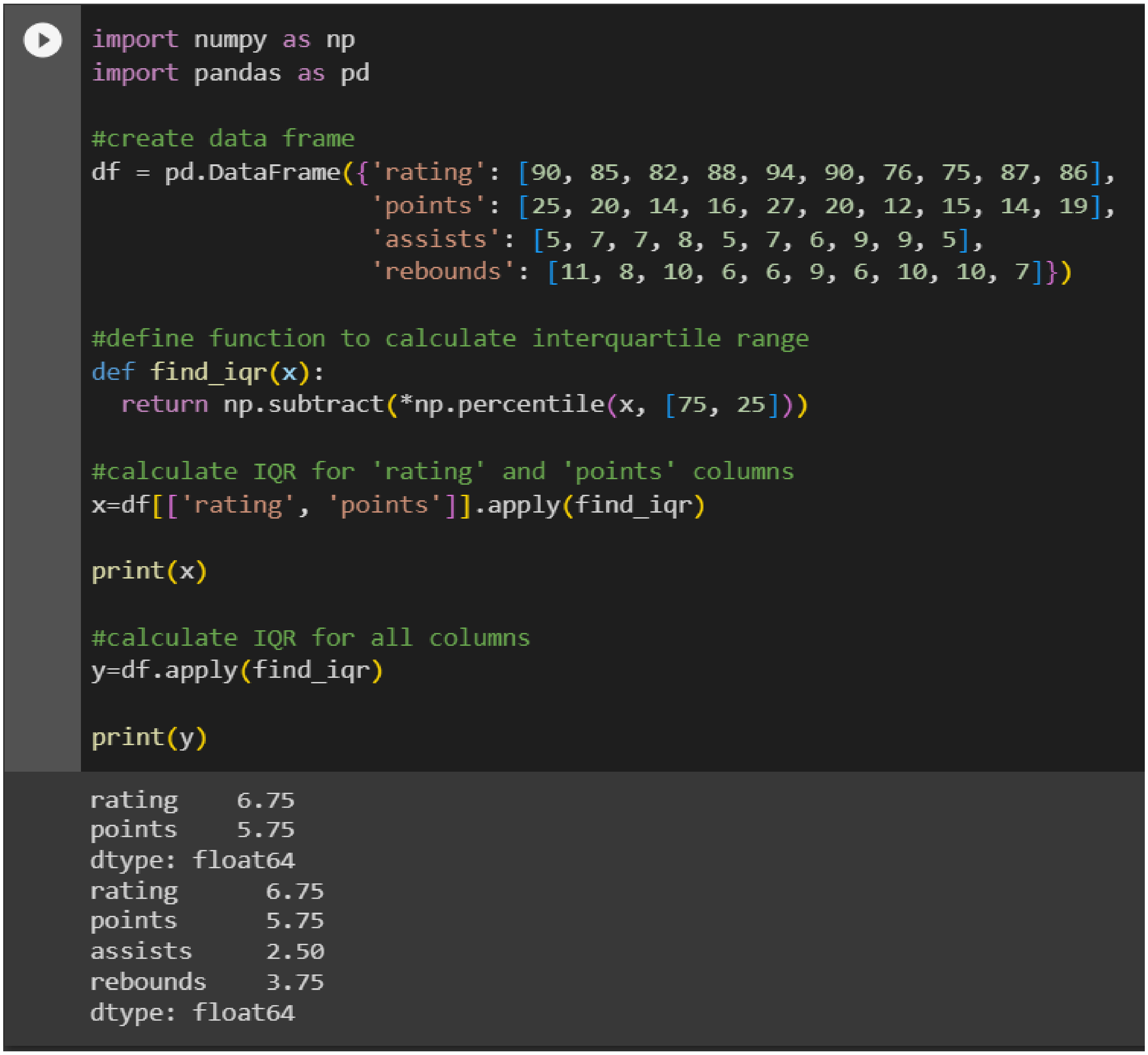


```
[7]  heights_new=[168,170,150,160,182,140,175,191,152]

[8]  heights_new.sort()

[9]  print(heights_new)

     [140, 150, 152, 160, 168, 170, 175, 182, 191]

[10] median=np.median(heights_new)

[12] print(median)

     168.0

[13] import statistics as stats
     stats.mode(heights)

     150

[14] np.var(heights)

     235.35999999999999

[15] std=np.std(heights)

     print(std)

     15.341447128612085
```



```
import numpy as np

#define array of data
data = np.array([14, 19, 20, 22, 24, 26, 27, 30, 30, 31, 36, 38, 44, 47])

#calculate interquartile range
q3, q1 = np.percentile(data, [75 ,25])
iqr = q3 - q1

#display interquartile range
print(iqr)
```

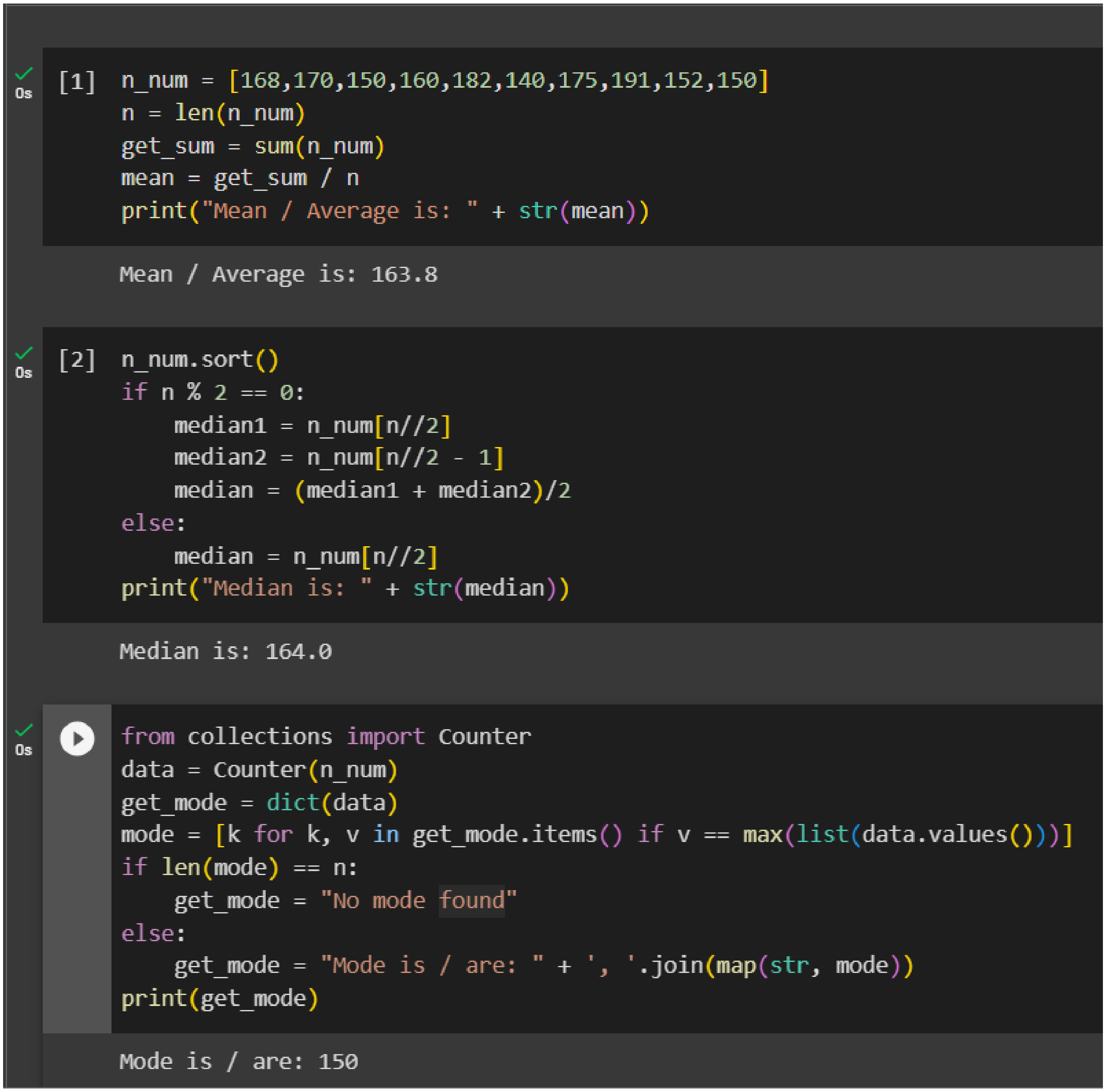
```
12.25
```



```
import numpy as np
import pandas as pd

#create data frame
df = pd.DataFrame({'rating': [90, 85, 82, 88, 94, 90, 76, 75, 87, 86],
                   'points': [25, 20, 14, 16, 27, 20, 12, 15, 14, 19],
                   'assists': [5, 7, 7, 8, 5, 7, 6, 9, 9, 5],
                   'rebounds': [11, 8, 10, 6, 6, 9, 6, 10, 10, 7]})

#define function to calculate interquartile range
def find_iqr(x):
  return np.subtract(*np.percentile(x, [75, 25]))

#calculate IQR for 'rating' and 'points' columns
x=df[['rating', 'points']].apply(find_iqr)

print(x)

#calculate IQR for all columns
y=df.apply(find_iqr)

print(y)
```

```
rating    6.75
points    5.75
dtype: float64
rating      6.75
points      5.75
assists     2.50
rebounds    3.75
dtype: float64
```

2)



```
n_num = [168,170,150,160,182,140,175,191,152,150]
n = len(n_num)
get_sum = sum(n_num)
mean = get_sum / n
print("Mean / Average is: " + str(mean))
```

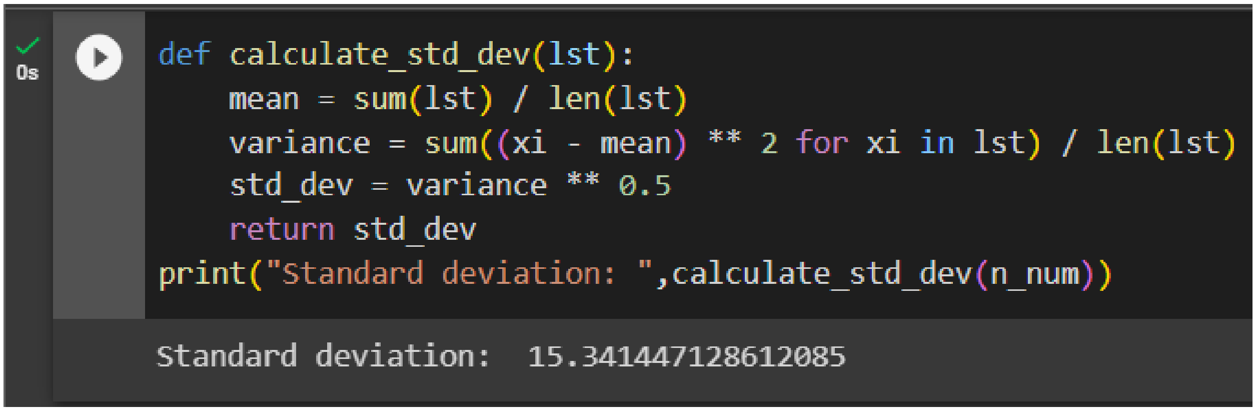
```
Mean / Average is: 163.8
```

```
n_num.sort()
if n % 2 == 0:
    median1 = n_num[n//2]
    median2 = n_num[n//2 - 1]
    median = (median1 + median2)/2
else:
    median = n_num[n//2]
print("Median is: " + str(median))
```

```
Median is: 164.0
```

```
from collections import Counter
data = Counter(n_num)
get_mode = dict(data)
mode = [k for k, v in get_mode.items() if v == max(list(data.values()))]
if len(mode) == n:
    get_mode = "No mode found"
else:
    get_mode = "Mode is / are: " + ', '.join(map(str, mode))
print(get_mode)
```

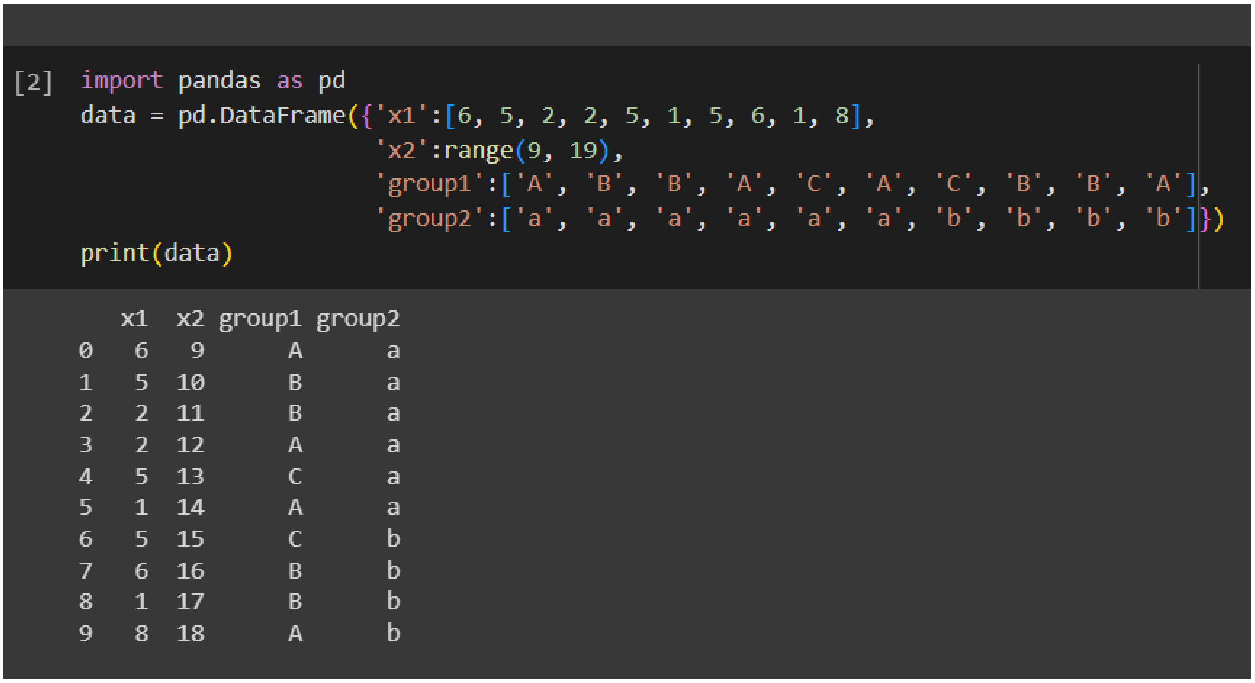
```
Mode is / are: 150
```



```
def calculate_std_dev(lst):
    mean = sum(lst) / len(lst)
    variance = sum((xi - mean) ** 2 for xi in lst) / len(lst)
    std_dev = variance ** 0.5
    return std_dev
print("Standard deviation: ",calculate_std_dev(n_num))
```

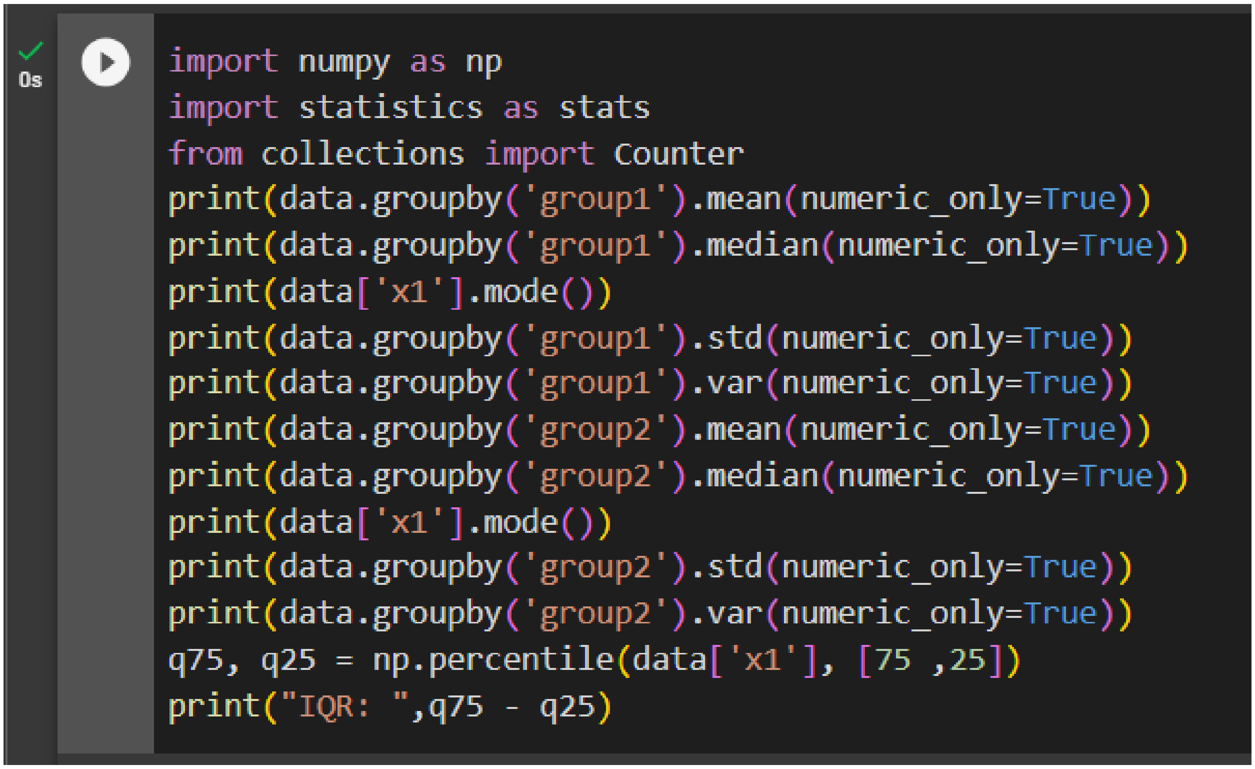
```
Standard deviation:  15.341447128612085
```

3)

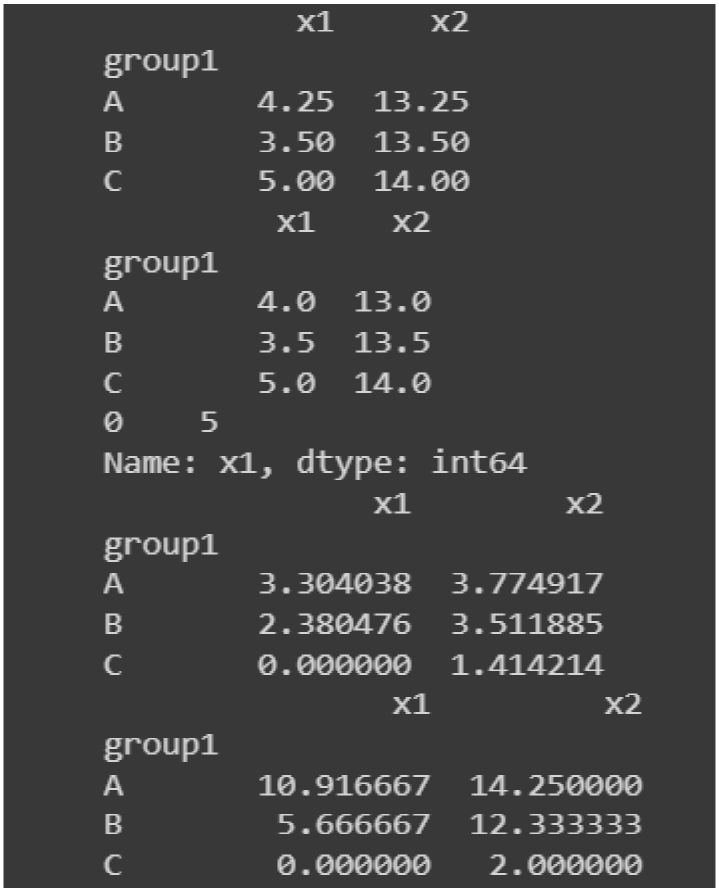


```
[2] import pandas as pd
    data = pd.DataFrame({'x1':[6, 5, 2, 2, 5, 1, 5, 6, 1, 8],
                         'x2':range(9, 19),
                         'group1':['A', 'B', 'B', 'A', 'C', 'A', 'C', 'B', 'B', 'A'],
                         'group2':['a', 'a', 'a', 'a', 'a', 'a', 'b', 'b', 'b', 'b']})
    print(data)
```

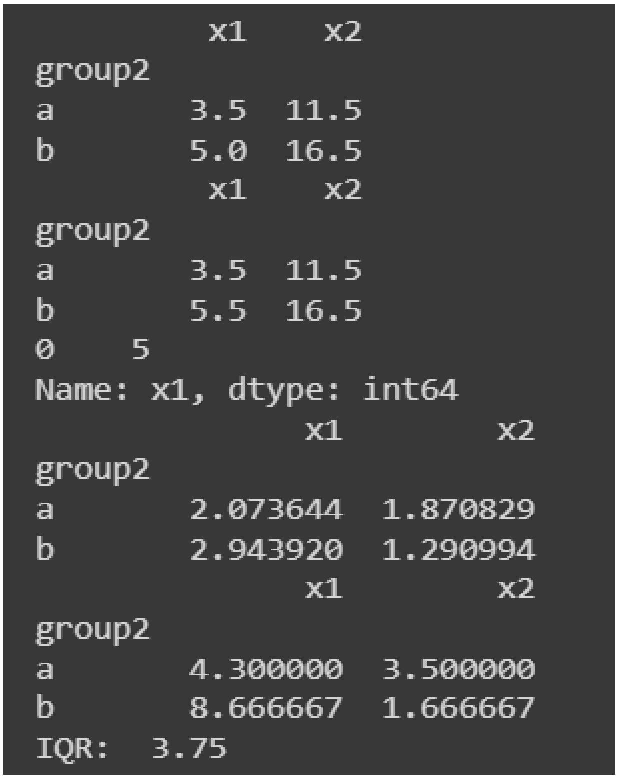
```
   x1  x2 group1 group2
0   6   9      A      a
1   5  10      B      a
2   2  11      B      a
3   2  12      A      a
4   5  13      C      a
5   1  14      A      a
6   5  15      C      b
7   6  16      B      b
8   1  17      B      b
9   8  18      A      b
```



```
import numpy as np
import statistics as stats
from collections import Counter
print(data.groupby('group1').mean(numeric_only=True))
print(data.groupby('group1').median(numeric_only=True))
print(data['x1'].mode())
print(data.groupby('group1').std(numeric_only=True))
print(data.groupby('group1').var(numeric_only=True))
print(data.groupby('group2').mean(numeric_only=True))
print(data.groupby('group2').median(numeric_only=True))
print(data['x1'].mode())
print(data.groupby('group2').std(numeric_only=True))
print(data.groupby('group2').var(numeric_only=True))
q75, q25 = np.percentile(data['x1'], [75 ,25])
print("IQR: ",q75 - q25)
```

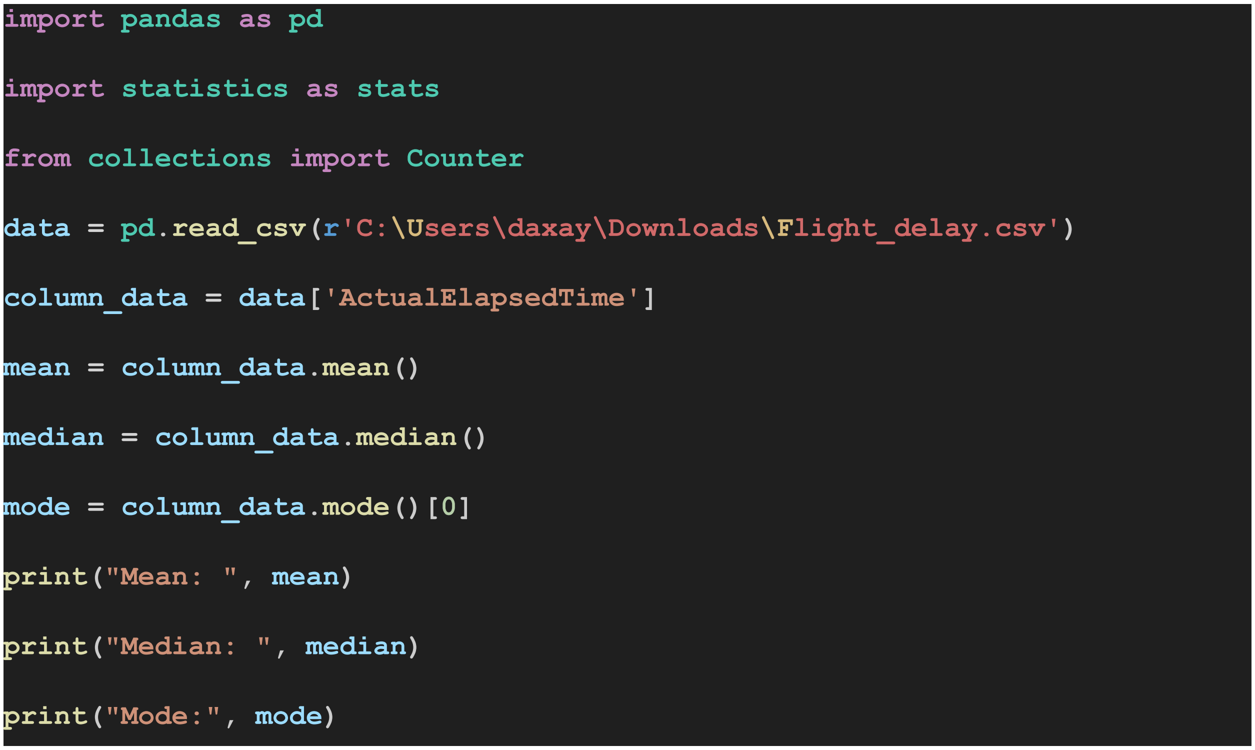


```
              x1     x2
    group1
    A       4.25  13.25
    B       3.50  13.50
    C       5.00  14.00
             x1    x2
    group1
    A       4.0  13.0
    B       3.5  13.5
    C       5.0  14.0
    0    5
    Name: x1, dtype: int64
                  x1         x2
    group1
    A       3.304038  3.774917
    B       2.380476  3.511885
    C       0.000000  1.414214
                   x1          x2
    group1
    A       10.916667  14.250000
    B        5.666667  12.333333
    C        0.000000   2.000000
```



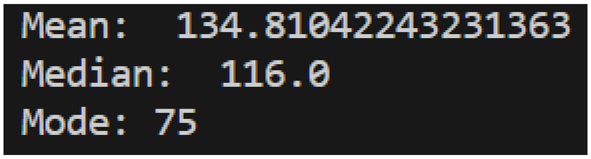
```
         x1    x2
group2
a       3.5  11.5
b       5.0  16.5
         x1    x2
group2
a       3.5  11.5
b       5.5  16.5
0    5
Name: x1, dtype: int64
              x1         x2
group2
a       2.073644  1.870829
b       2.943920  1.290994
              x1         x2
group2
a       4.300000  3.500000
b       8.666667  1.666667
IQR:  3.75
```

Code:



```
import pandas as pd

import statistics as stats

from collections import Counter

data = pd.read_csv(r'C:\Users\daxay\Downloads\Flight_delay.csv')

column_data = data['ActualElapsedTime']

mean = column_data.mean()

median = column_data.median()

mode = column_data.mode()[0]

print("Mean: ", mean)

print("Median: ", median)

print("Mode:", mode)
```

Output:



```
Mean:  134.81042243231363
Median:  116.0
Mode: 75
```

Post Lab Question-Answers:

1. What are the various applications of central tendency and variability of data?

Ans: The measures of central tendency and variability are fundamental statistical concepts used to summarize and analyze data. They have various applications across different fields. Here are some common applications:

1. Descriptive Statistics: Central tendency measures, such as the mean, median, and mode, provide a summary of the typical or central value of a dataset. Variability measures, such as the range, variance, and standard deviation, describe the spread or dispersion of the data. These statistics help in understanding the overall characteristics of a dataset.
2. Data Analysis: Central tendency and variability measures are used extensively in data analysis. They help in identifying patterns, trends, and outliers in the data. For example, in finance, these measures are used to analyze stock market returns or assess risk in investment portfolios.

3. Research and Surveys: In research studies and surveys, central tendency measures are used to summarize and present data. They provide a concise representation of the data, making it easier to interpret and compare different groups or variables. Variability measures help in assessing the consistency or variability of responses.
4. Quality Control: Central tendency and variability measures are used in quality control processes to monitor and improve product or process performance. They help in identifying variations and deviations from desired standards, enabling corrective actions to be taken.
5. Forecasting and Predictive Modeling: Central tendency measures can be used as a basis for forecasting future values. For example, in time series analysis, the moving average is often used to predict future trends. Variability measures help in assessing the uncertainty or reliability of the forecasts.
6. Performance Evaluation: Central tendency and variability measures are used to evaluate performance in various domains. For instance, in sports, these measures are used to compare player statistics and assess team performance. In education, they are used to evaluate student performance and measure learning outcomes.
7. Process Improvement: Central tendency and variability measures are used in process improvement methodologies like Six Sigma. They help in identifying process variations, analyzing root causes, and implementing corrective actions to reduce variability and improve efficiency.

2. What are the various applications of finding Central Tendency of Data?

Ans: Finding the central tendency of data is a fundamental statistical concept that has various applications across different fields. Here are some common applications:

1. Data Summarization: Central tendency measures, such as the mean, median, and mode, provide a summary of the typical or central value of a dataset. These measures help in simplifying complex data by reducing it to a single representative value. This is useful for summarizing large datasets and making them more manageable and interpretable.
2. Data Comparison: Central tendency measures are used to compare different groups or variables. By calculating the central tendency for each group, you can determine if there are any significant differences or similarities between them. This is particularly useful in research studies, social sciences, and business analytics.
3. Forecasting and Prediction: Central tendency measures can be used as a basis for forecasting future values. For example, in time series analysis, the moving average is often used to predict future trends. By identifying the central tendency of historical data, you can make informed predictions about future outcomes.
4. Decision Making: Central tendency measures provide valuable information for decision-making processes. For instance, in business, the mean or median can be used to determine the average sales or customer satisfaction level. This information helps in setting benchmarks, evaluating performance, and making informed decisions.
5. Quality Control: Central tendency measures are used in quality control processes to monitor and improve product or process performance. By calculating the mean or median of a set of measurements, you can assess whether the process is operating within acceptable limits. Deviations from the central tendency may indicate the need for corrective actions.

6. Sampling Techniques: Central tendency measures are used in sampling techniques to estimate population parameters. By calculating the mean or median of a sample, you can make inferences about the population as a whole. This is particularly useful when it is not feasible or practical to collect data from the entire population.
7. Data Visualization: Central tendency measures are often used in data visualization techniques. For example, the mean or median can be represented as a central point in a box plot or as a reference line in a line graph. These visual representations help in understanding the distribution and characteristics of the data.

Outcomes:

Comprehend descriptive and proximity measures of data.

Conclusion (based on the Results and outcomes achieved):

The experiment successfully achieved its aim of finding measures of central tendency and variability for single and multi-attribute data using statistical analysis. The obtained results provide valuable insights into the dataset, facilitating further analysis and interpretation in the relevant field of study or application.

References:

Books/ Journals/ Websites

1. Han, Kamber, "Data Mining Concepts and Techniques", Morgan Kaufmann 3nd Edition

2. S.C. Gupta , V. K. Kapoor Fundamentals of mathematical statistics Sultan Chand and Sons 2014