Statistics Assignment Submission - (Rishi Kumar Mishra)

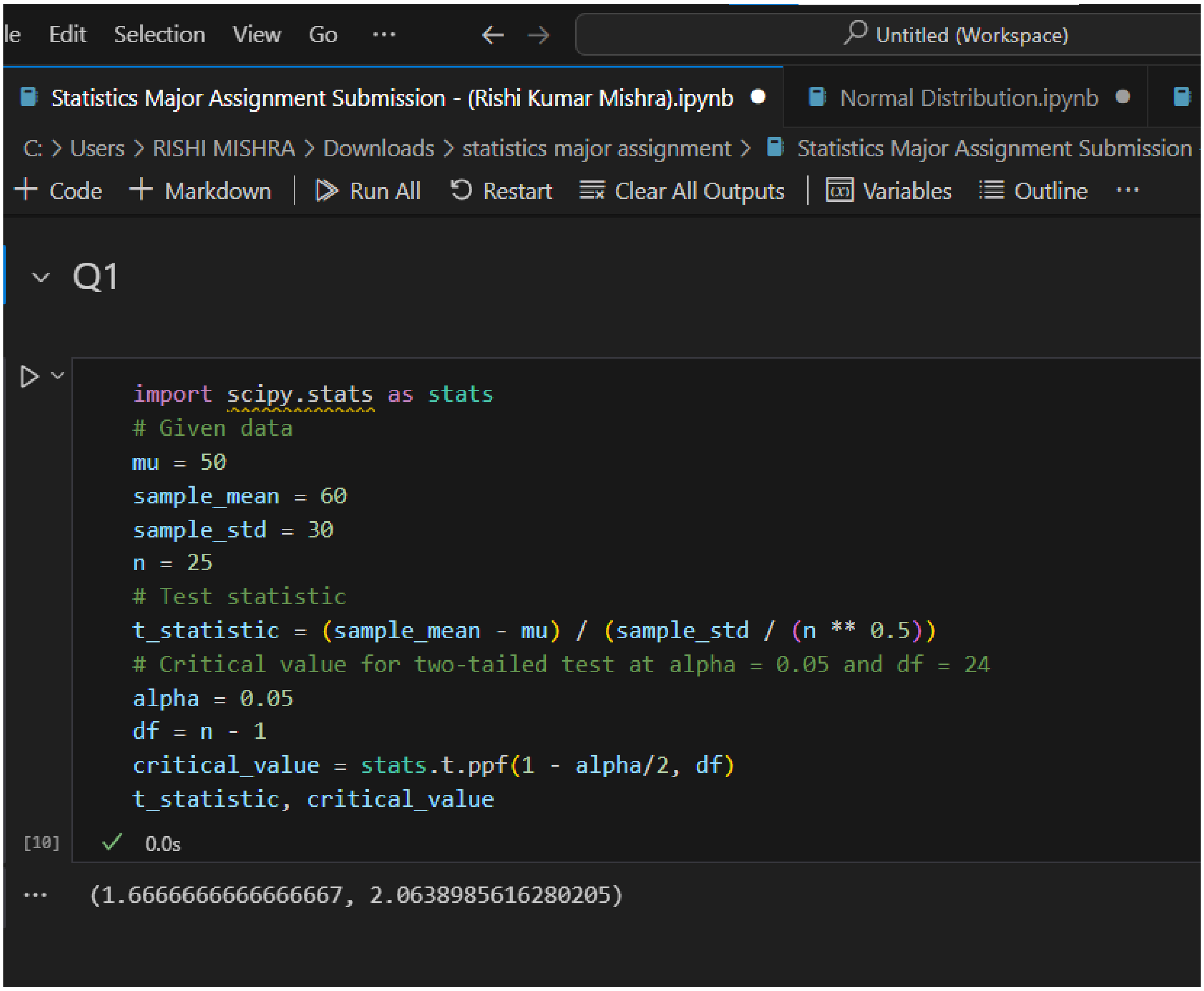
EMAIL **rishimishra089@gmail.com**
Data Science With Python Carrer Program
Mob 9993955483

Q1) According to a study, the daily average time spent by a user on a social media website is 50 minutes. To test the claim of this study, Ramesh, a researcher, takes a sample of 25 website users and finds out that the mean time spent by the sample users is 60 minutes and the sample standard deviation is 30 minutes.
Based on this information, the null and the alternative hypotheses will be:
Ho = The average time spent by the users is 50 minutes
H1 = The average time spent by the users is not 50minutes
Use a 5% significance level to test this hypothesis.

ANS

Given:

Population mean (μ\muμ) = 50 minutes
Sample mean (x¯\bar{x}x¯) = 60 minutes
Sample standard deviation (sss) = 30 minutes
Sample size (nnn) = 25
Significance level (α\alphaα) = 0.05
Hypotheses:
H0H_0H0: The average time spent by the users is 50 minutes (μ=50\mu = 50μ=50)
H1H_1H1: The average time spent by the users is not 50 minutes (μ≠50\mu \neq 50μ□=50)
Test Statistic: t=x¯−μs/nt = \frac{\bar{x} - \mu}{s / \sqrt{n}}t=s/nx¯−μ
Critical Value: For a two-tailed test at α=0.05\alpha = 0.05α=0.05 and df=n−1=24df = n - 1 = 24df=n−1=24, the critical t-value can be found from t-tables or using a statistical software.
Let's calculate this using Python.


le Edit Selection View Go ···
Untitled (Workspace)
Statistics Major Assignment Submission - (Rishi Kumar Mishra).ipynb
Normal Distribution.ipynb
C: > Users > RISHI MISHRA > Downloads > statistics major assignment > Statistics Major Assignment Submission
+ Code + Markdown Run All Restart Clear All Outputs Variables Outline ···
Q1
[10]
0.0s

```python
import scipy.stats as stats
# Given data
mu = 50
sample_mean = 60
sample_std = 30
n = 25
# Test statistic
t_statistic = (sample_mean - mu) / (sample_std / (n ** 0.5))
# Critical value for two-tailed test at alpha = 0.05 and df = 24
alpha = 0.05
df = n - 1
critical_value = stats.t.ppf(1 - alpha/2, df)
t_statistic, critical_value
```

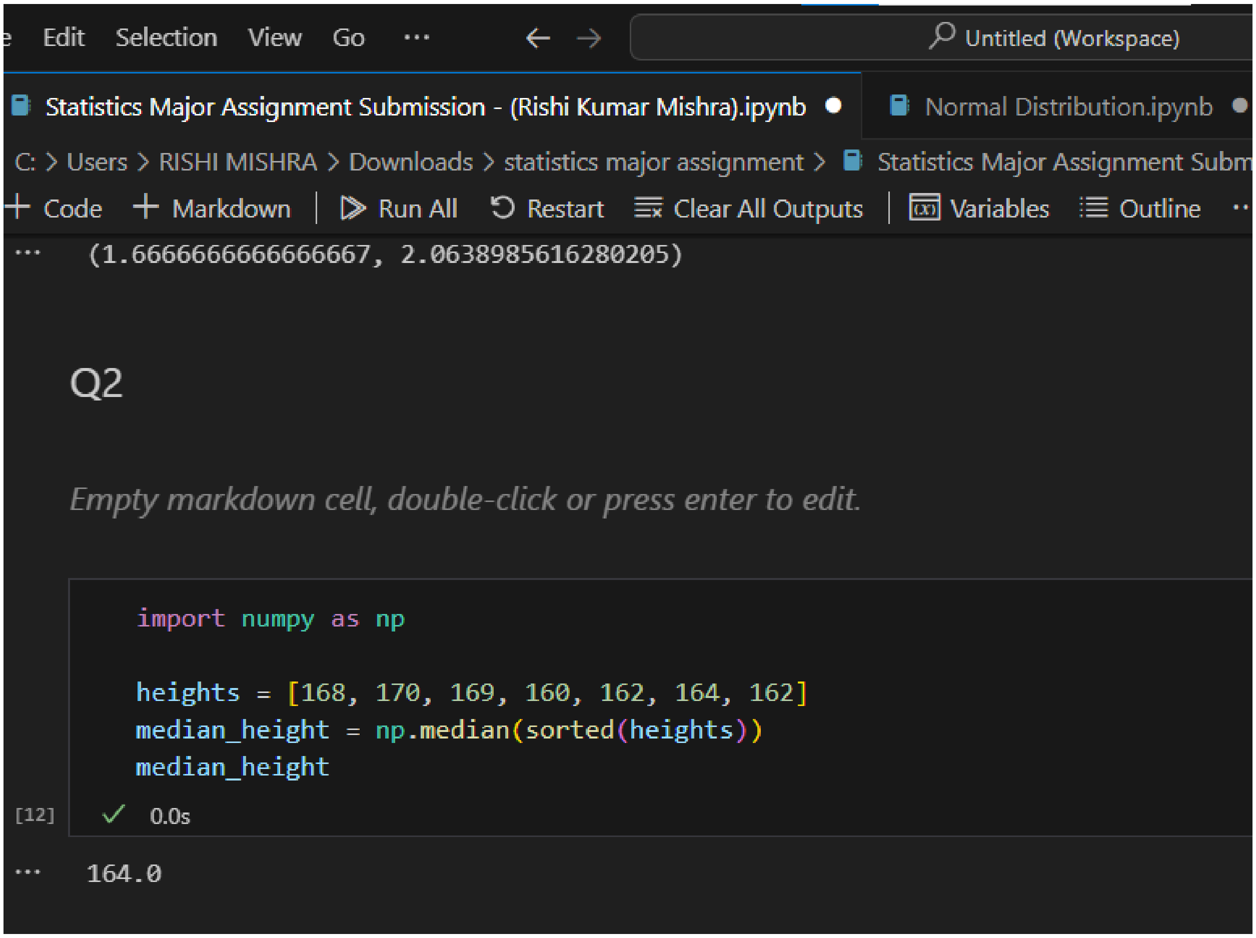
```
(1.6666666666666667, 2.0638985616280205)
```

Q2) Height of 7 students (in cm) is given below. What is the median?

168 170 169 160 162 164 162.

ANS

Given: Heights: 168, 170, 169, 160, 162, 164, 162
Steps:
Sort the data.
Find the middle value.


e Edit Selection View Go ···
Untitled (Workspace)
Statistics Major Assignment Submission - (Rishi Kumar Mishra).ipynb
Normal Distribution.ipynb
C: > Users > RISHI MISHRA > Downloads > statistics major assignment > Statistics Major Assignment Subm
+ Code + Markdown Run All Restart Clear All Outputs Variables Outline ···

```
(1.6666666666666667, 2.0638985616280205)
```

Q2

Empty markdown cell, double-click or press enter to edit.

```
import numpy as np

heights = [168, 170, 169, 160, 162, 164, 162]
median_height = np.median(sorted(heights))
median_height
```

[12] ✓ 0.0s

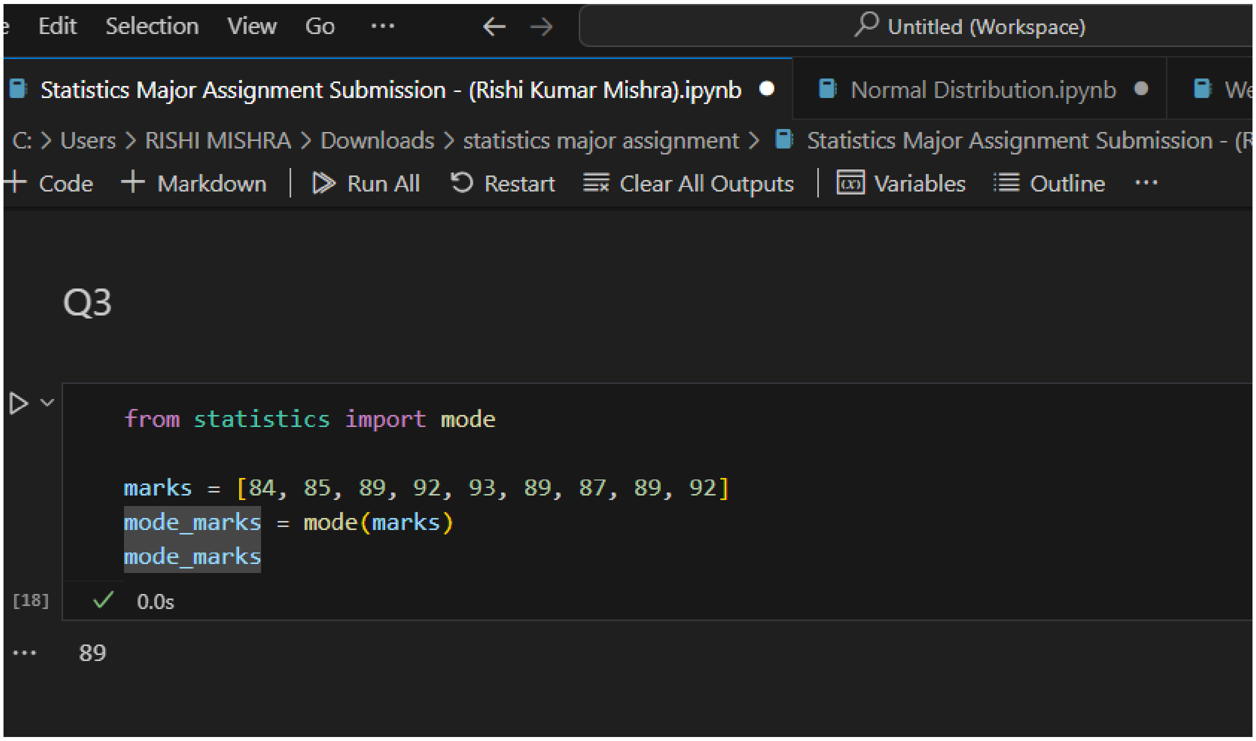
```
164.0
```

Q3) Below are the observations of the marks of a student. Find thevalue of mode.

84 85 89 92 93 89 87 89 92

ANS

Given: Marks: 84, 85, 89, 92, 93, 89, 87, 89, 92
Steps:
Find the most frequent value.


Edit
Selection
View
Go
Untitled (Workspace)
Statistics Major Assignment Submission - (Rishi Kumar Mishra).ipynb
Normal Distribution.ipynb
We
C: > Users > RISHI MISHRA > Downloads > statistics major assignment > Statistics Major Assignment Submission - (R
Code
Markdown
Run All
Restart
Clear All Outputs
Variables
Outline
[18]
0.0s

Q3

```
from statistics import mode

marks = [84, 85, 89, 92, 93, 89, 87, 89, 92]
mode_marks = mode(marks)
mode_marks
```

```
89
```

Q4) From the table given below, what is the mean of marks obtainedby **20 students**?

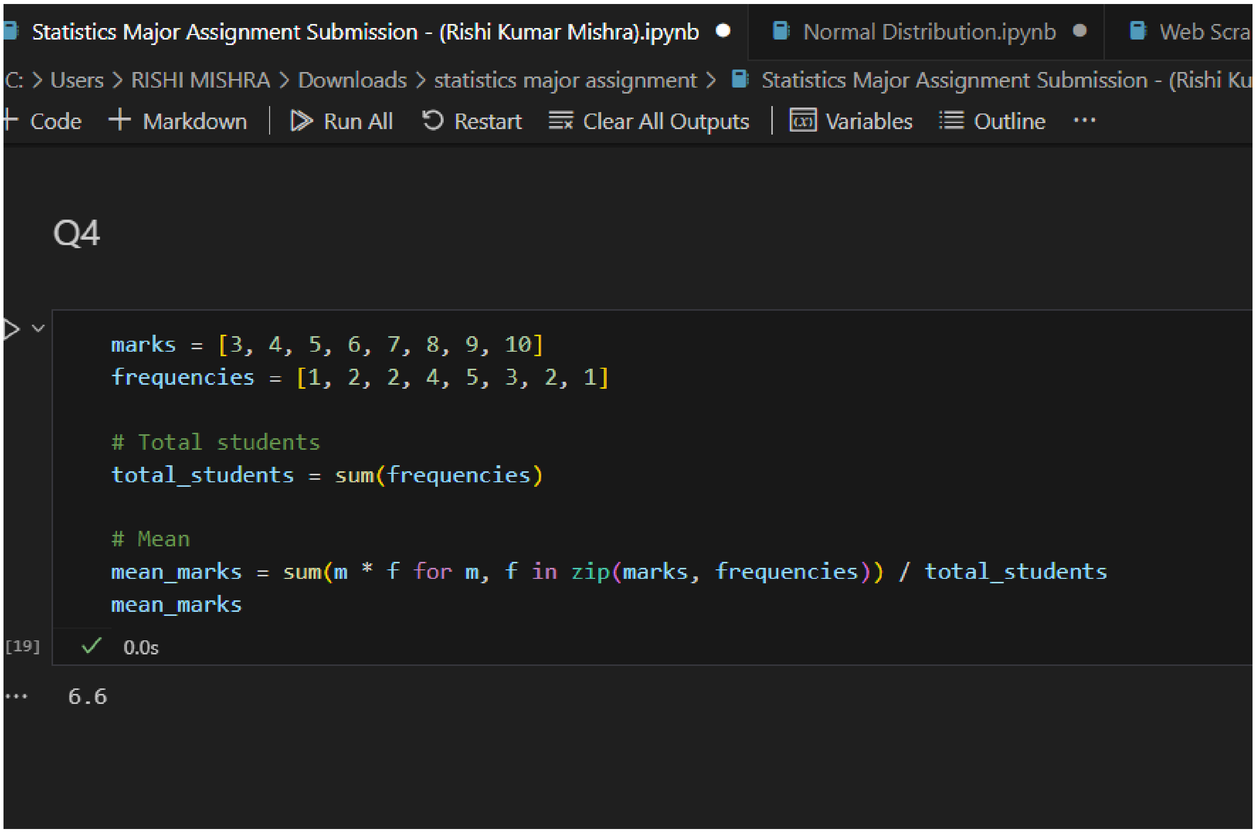
Marks Xi	No. of studentsfi
3	1
4	2
5	2
6	4
7	5
8	3
9	2
10	1
Total	20

ANS

Given: ∑fi=20\sum f_i = 20∑fi=20 ∑fiXi\sum f_i X_i∑fiXi

Steps:

1. Calculate the sum of fiXif_i X_ifiXi.
2. Divide by the total number of students.


Statistics Major Assignment Submission - (Rishi Kumar Mishra).ipynb
Normal Distribution.ipynb
Web Scra
C: > Users > RISHI MISHRA > Downloads > statistics major assignment > Statistics Major Assignment Submission - (Rishi Ku
+ Code
+ Markdown
Run All
Restart
Clear All Outputs
Variables
Outline
···
[19]
0.0s

Q4

```
marks = [3, 4, 5, 6, 7, 8, 9, 10]
frequencies = [1, 2, 2, 4, 5, 3, 2, 1]

# Total students
total_students = sum(frequencies)

# Mean
mean_marks = sum(m * f for m, f in zip(marks, frequencies)) / total_students
mean_marks
```

```
6.6
```

Q5 For a certain type of computer, the length of time between charges of the battery is normally distributed with a mean of **50 hours** and a standard deviation of **15 hours.** John owns one of these computers and wants to know the probability that the length of timewill be between **50** and **70 hours.**

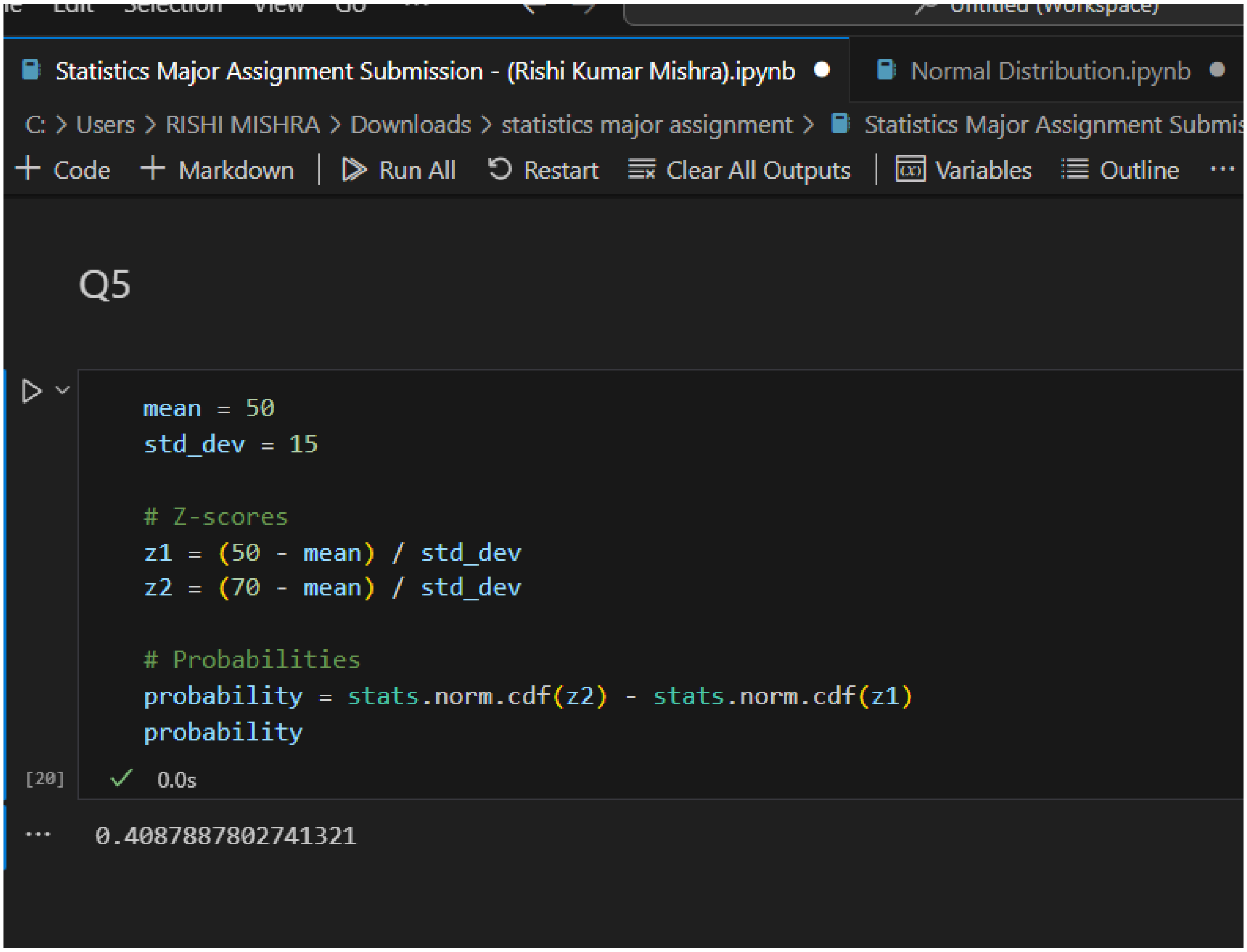
ANS

Given:

- Mean (μ\muμ) = 50 hours
- Standard deviation (σ\sigmaσ) = 15 hours
- P(50≤X≤70)P(50 \leq X \leq 70)P(50≤X≤70)

Steps:

1. Standardize the values.
2. Use the cumulative distribution function (CDF).


Statistics Major Assignment Submission - (Rishi Kumar Mishra).ipynb
Normal Distribution.ipynb
C: > Users > RISHI MISHRA > Downloads > statistics major assignment > Statistics Major Assignment Submis
+ Code
+ Markdown
Run All
Restart
Clear All Outputs
Variables
Outline
[20]
✓ 0.0s

Q5

```
mean = 50
std_dev = 15

# Z-scores
z1 = (50 - mean) / std_dev
z2 = (70 - mean) / std_dev

# Probabilities
probability = stats.norm.cdf(z2) - stats.norm.cdf(z1)
probability
```

```
0.4087887802741321
```

Q6) Find the range of the following.

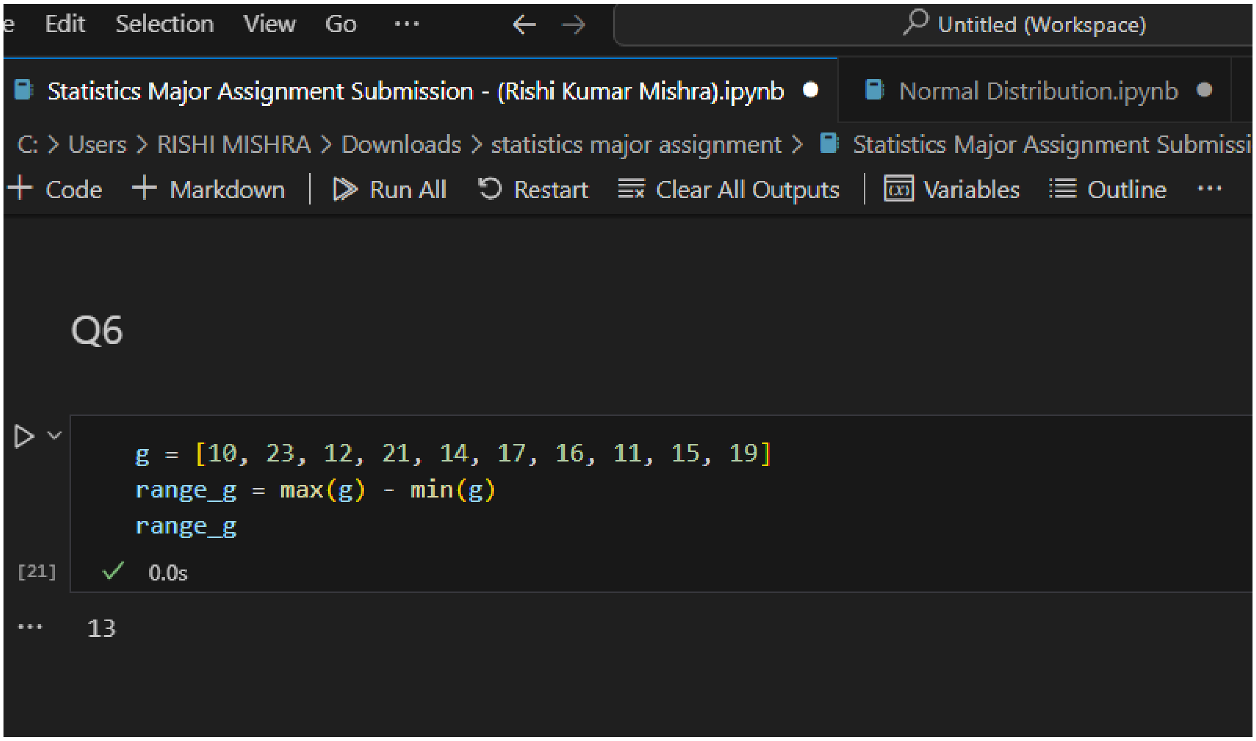
g = [10, 23, 12, 21, 14, 17, 16, 11, 15, 19]

ANS

Given: g=[10,23,12,21,14,17,16,11,15,19]

Steps:

1. Find the difference between the maximum and minimum values.


e Edit Selection View Go ···
Untitled (Workspace)
Statistics Major Assignment Submission - (Rishi Kumar Mishra).ipynb
Normal Distribution.ipynb
C: > Users > RISHI MISHRA > Downloads > statistics major assignment > Statistics Major Assignment Submissi
+ Code + Markdown Run All Restart Clear All Outputs Variables Outline ···

Q6

```
g = [10, 23, 12, 21, 14, 17, 16, 11, 15, 19]
range_g = max(g) - min(g)
range_g
```

[21] ✓ 0.0s

```
13
```

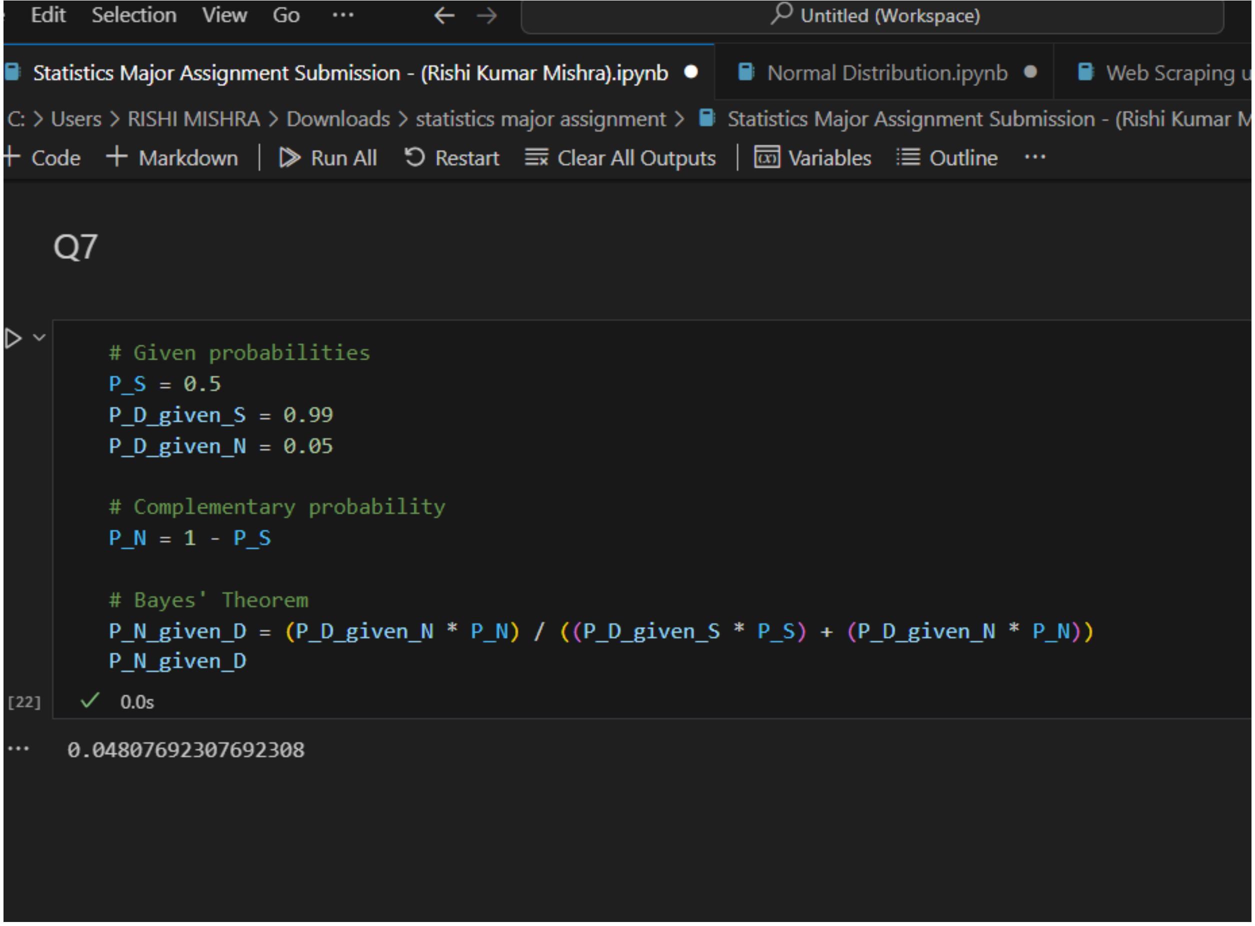
Q7) It is estimated that **50%** of emails are spam emails. Some softwarehas been applied to filter these spam emails before they reach yourinbox. A certain brand of software claims that it can detect **99%** of spam emails, and the probability of a false positive (a non-spam email detected as spam) is **5%**. Now if an email is detected as **spam,** thenwhat is the probability that it is in fact a **non-spam email**?

ANS

Given:

- Probability of spam (P(S)P(S)P(S)) = 0.5
- Probability of detecting spam given it is spam (P(D|S)P(D|S)P(D|S)) = 0.99
- Probability of false positive (P(D|N)P(D|N)P(D|N)) = 0.05

Use Bayes' Theorem


Edit
Selection
View
Go
···
Untitled (Workspace)
Statistics Major Assignment Submission - (Rishi Kumar Mishra).ipynb ●
Normal Distribution.ipynb ●
Web Scraping u
C: > Users > RISHI MISHRA > Downloads > statistics major assignment > Statistics Major Assignment Submission - (Rishi Kumar M
+ Code
+ Markdown
Run All
Restart
Clear All Outputs
Variables
Outline
···
[22]
✓ 0.0s

Q7

```
# Given probabilities
P_S = 0.5
P_D_given_S = 0.99
P_D_given_N = 0.05

# Complementary probability
P_N = 1 - P_S

# Bayes' Theorem
P_N_given_D = (P_D_given_N * P_N) / ((P_D_given_S * P_S) + (P_D_given_N * P_N))
P_N_given_D
```

```
0.04807692307692308
```

Q8) Given the following distribution of returns, determine the **lowerquartile**:

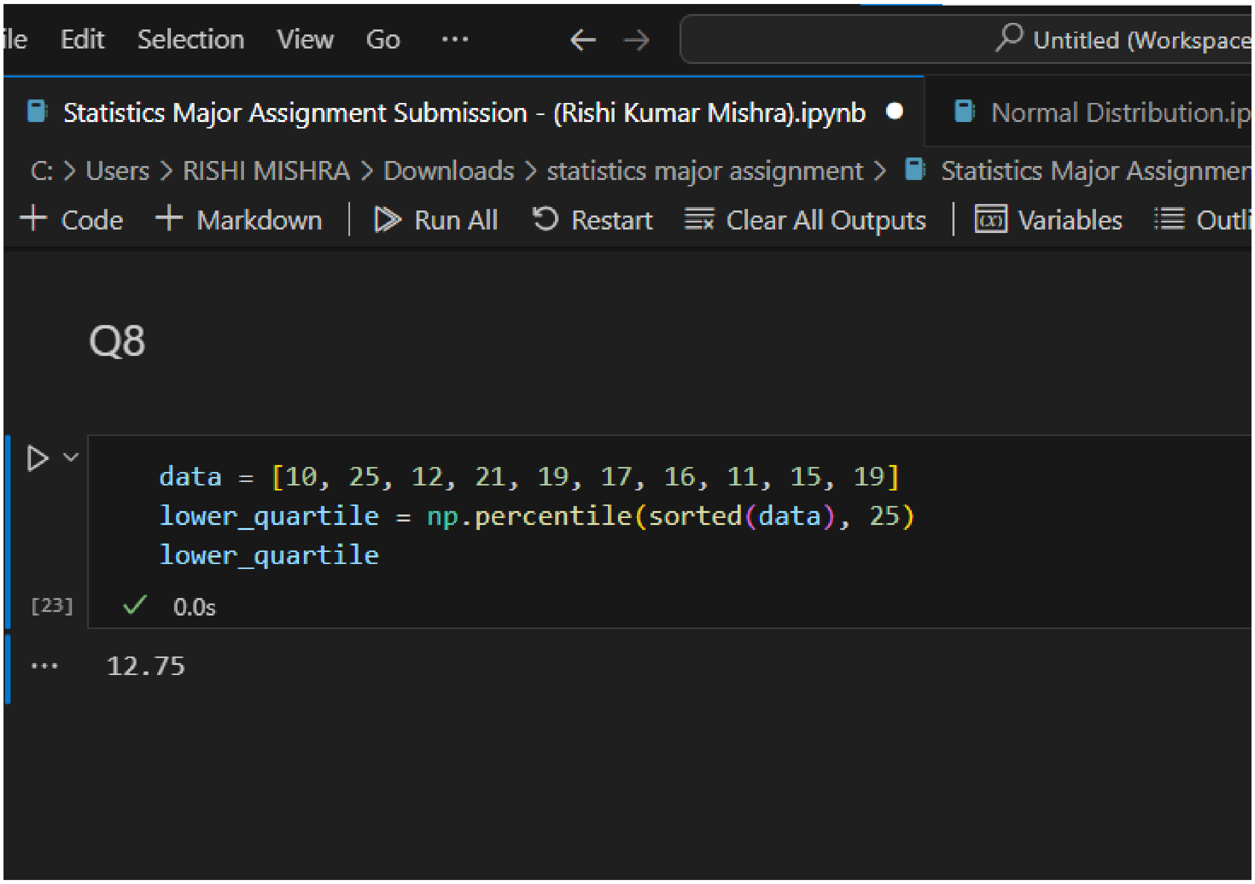
{10 25 12 21 19 17 16 11 15 19}

ANS

Given: Data=[10,25,12,21,19,17,16,11,15,19]\text{Data} = [10, 25, 12, 21, 19, 17, 16, 11, 15, 19]Data=[10,25,12,21,19,17,16,11,15,19]

Steps:

1. Sort the data.
2. Find the 25th percentile.


ile
Edit
Selection
View
Go
···
Untitled (Workspace
Statistics Major Assignment Submission - (Rishi Kumar Mishra).ipynb
Normal Distribution.ip
C: > Users > RISHI MISHRA > Downloads > statistics major assignment > Statistics Major Assignmen
+ Code
+ Markdown
Run All
Restart
Clear All Outputs
Variables
Outli
[23]
✓ 0.0s

Q8

```
data = [10, 25, 12, 21, 19, 17, 16, 11, 15, 19]
lower_quartile = np.percentile(sorted(data), 25)
lower_quartile
```

```
12.75
```

Q9) For a Binomial distribution, the number of trials(n) is **25**, and theprobability of success is **0.3.** What’s the variability of the distribution?

ANS

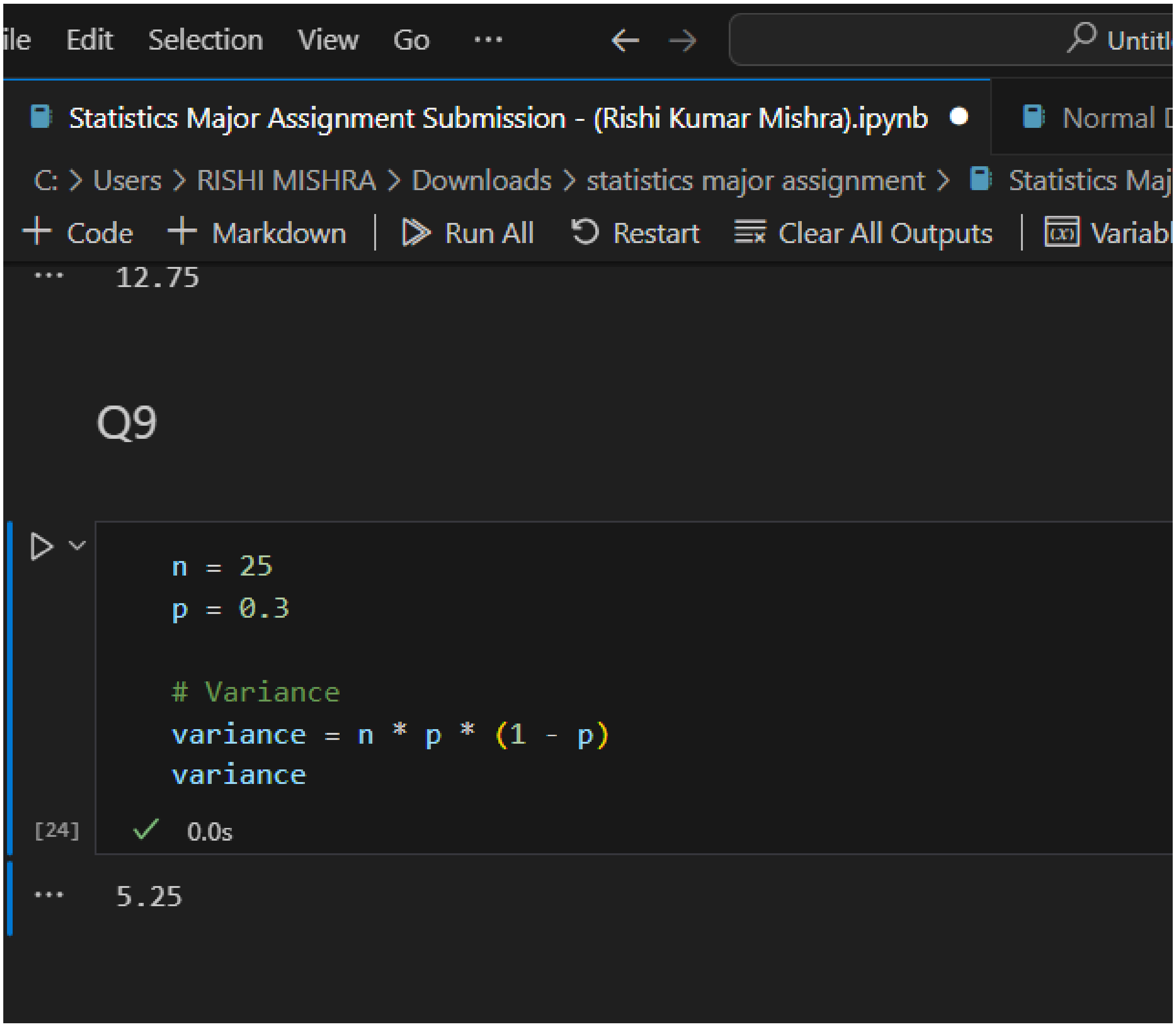
Given:

- Number of trials (nnn) = 25
- Probability of success (ppp) = 0.3

Steps:

1. Calculate the variance.

σ2=n·p·(1−p)\sigma^2 = n \cdot p \cdot (1 - p)σ2=n·p·(1−p)


ile
Edit
Selection
View
Go
···
Untitl
Statistics Major Assignment Submission - (Rishi Kumar Mishra).ipynb ●
Normal I
C: > Users > RISHI MISHRA > Downloads > statistics major assignment > Statistics Maj
+ Code
+ Markdown
Run All
Restart
Clear All Outputs
Variabl

```
12.75
```

Q9

```python
n = 25
p = 0.3

# Variance
variance = n * p * (1 - p)
variance
```

[24] ✓ 0.0s

```
5.25
```

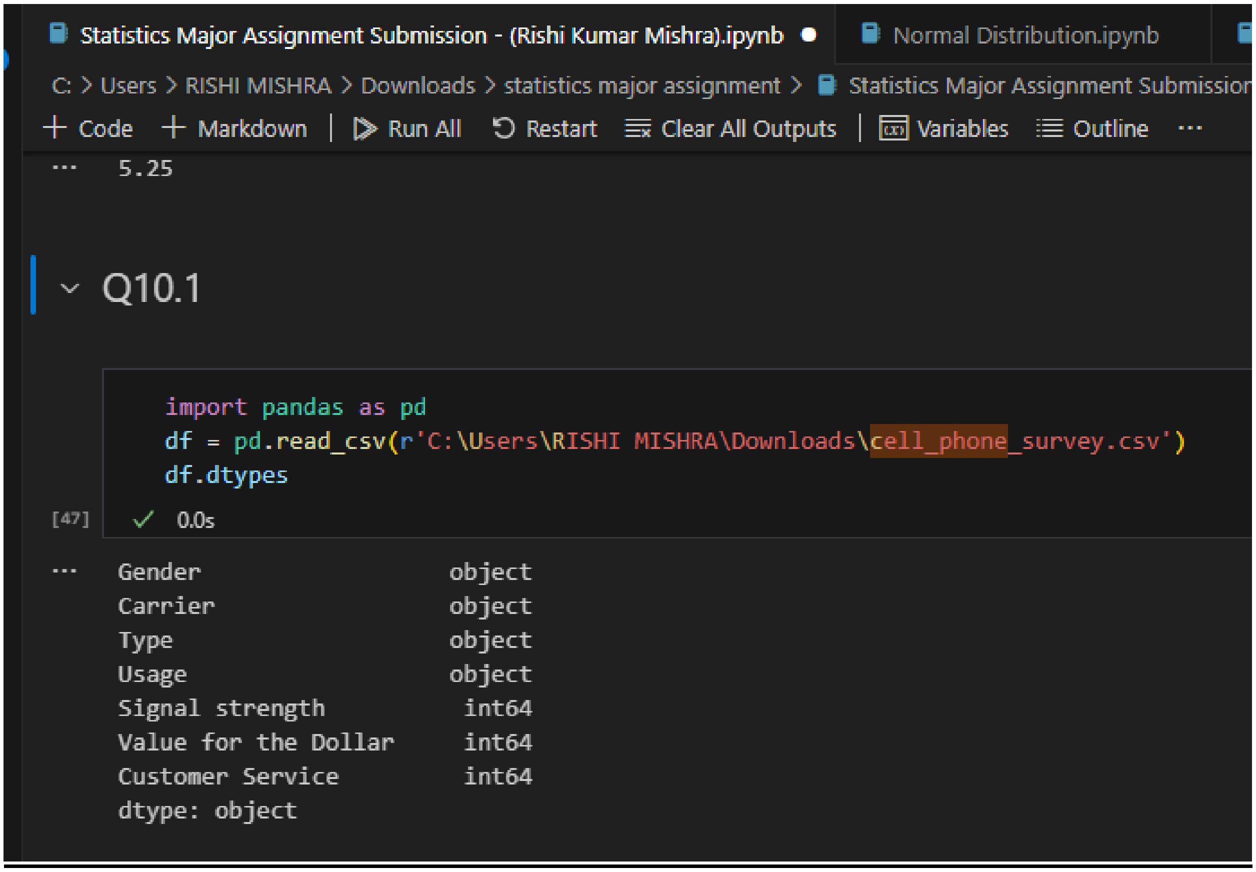
Q10) Download the **Cell Phone Survey Dataset** and perform the below-mentioned operations on the dataset:-

- Checking **datatypes** of each column in the dataset.
- Find the **Mean** of the Signal strength column using the Pandas and Statistics library.
- Find the **Median** of Customer Service column using Pandas and Statistics library.
- Find the **Mode** of Signal strength column using Pandas andStatistics library.
- Find the **Standard deviation** of the Customer Service column using **Pandas** and **Statistics** library.
- Find the **Variance** of Customer Service column using **Pandas** and **Statistics** library.

- Calculate **Percentiles** of Value for the Dollar column usingNumPy.
- Calculate the **Range** of Value for the Dollar column using Pandas.
- Calculate **IQR** of Value for the Dollar column using Pandas.
- Hypothesis Testing - Using the data in the Cell Phone Survey dataset, apply **ANOVA** to determine if the **mean** response for Value for dollar is the same for different types of cell phones.

ANS

10.1 Checking datatypes:



Statistics Major Assignment Submission - (Rishi Kumar Mishra).ipynb
Normal Distribution.ipynb
C: > Users > RISHI MISHRA > Downloads > statistics major assignment > Statistics Major Assignment Submissior
+ Code
+ Markdown
Run All
Restart
Clear All Outputs
Variables
Outline
[47]
0.0s

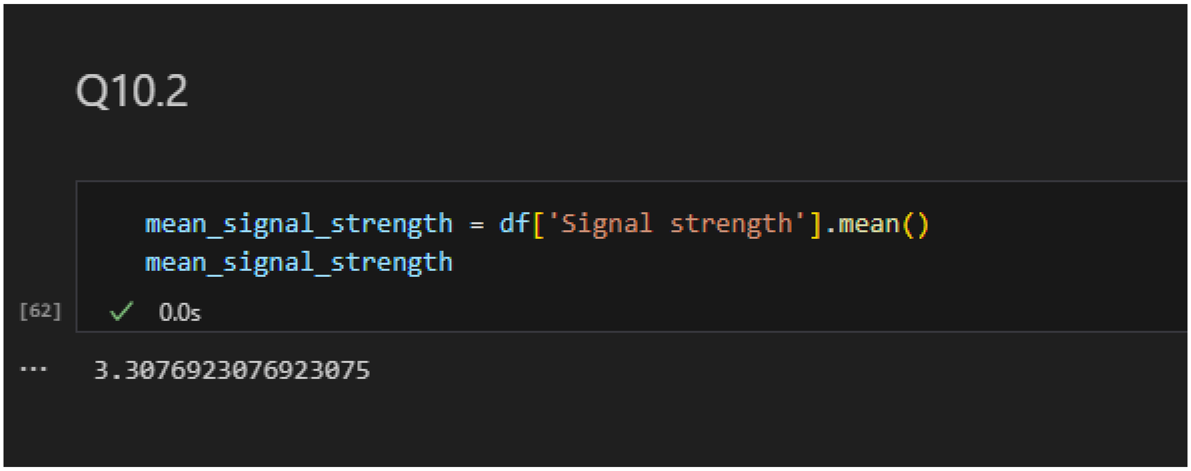
```
5.25
```

Q10.1

```
import pandas as pd
df = pd.read_csv(r'C:\Users\RISHI MISHRA\Downloads\cell_phone_survey.csv')
df.dtypes
```

```
Gender                  object
Carrier                 object
Type                    object
Usage                   object
Signal strength          int64
Value for the Dollar     int64
Customer Service         int64
dtype: object
```

10.2 Mean of Signal strength



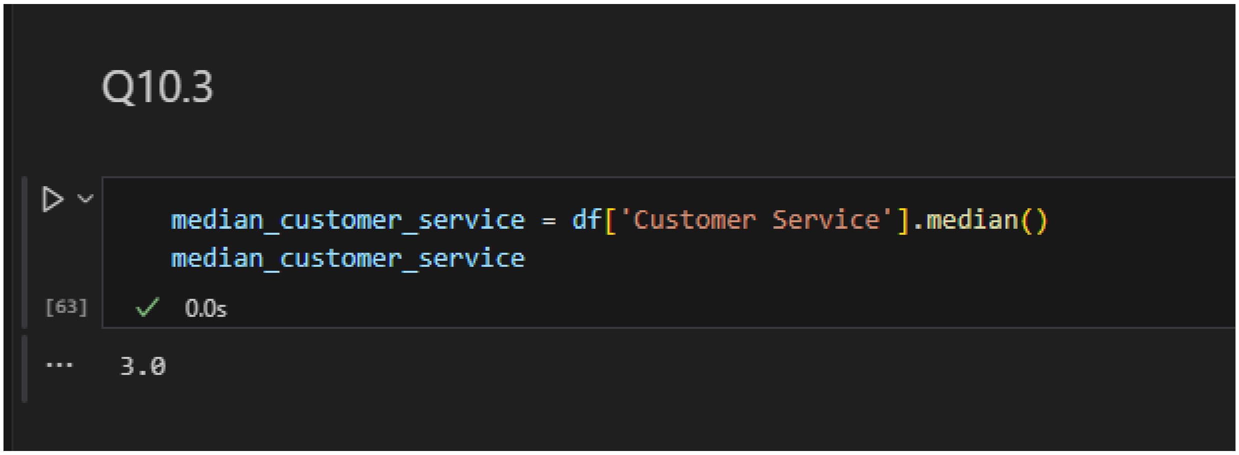
Q10.2

```
mean_signal_strength = df['Signal strength'].mean()
mean_signal_strength
```

```
[62] ✓ 0.0s
```

```
... 3.3076923076923075
```

10.3 Median of Customer Service



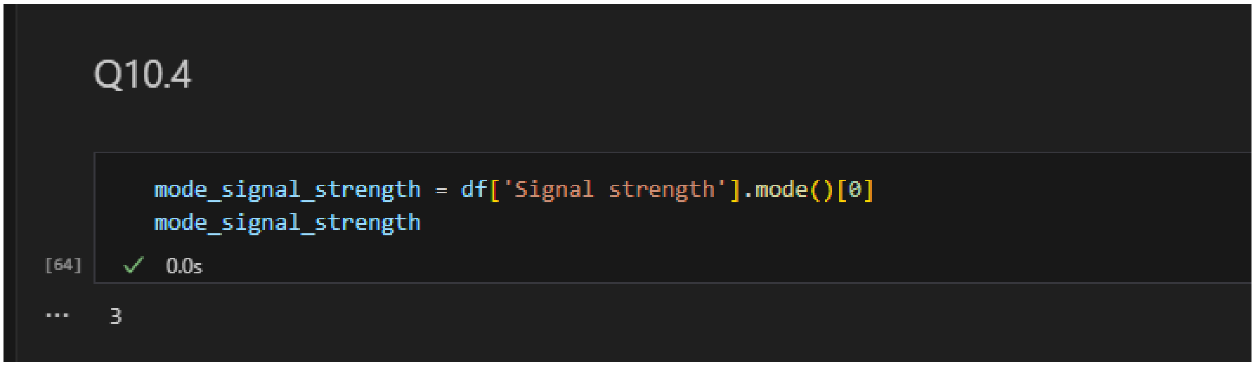
Q10.3

```
median_customer_service = df['Customer Service'].median()
median_customer_service
```

```
[63] ✓ 0.0s
```

```
... 3.0
```

10.4 Mode of Signal strength



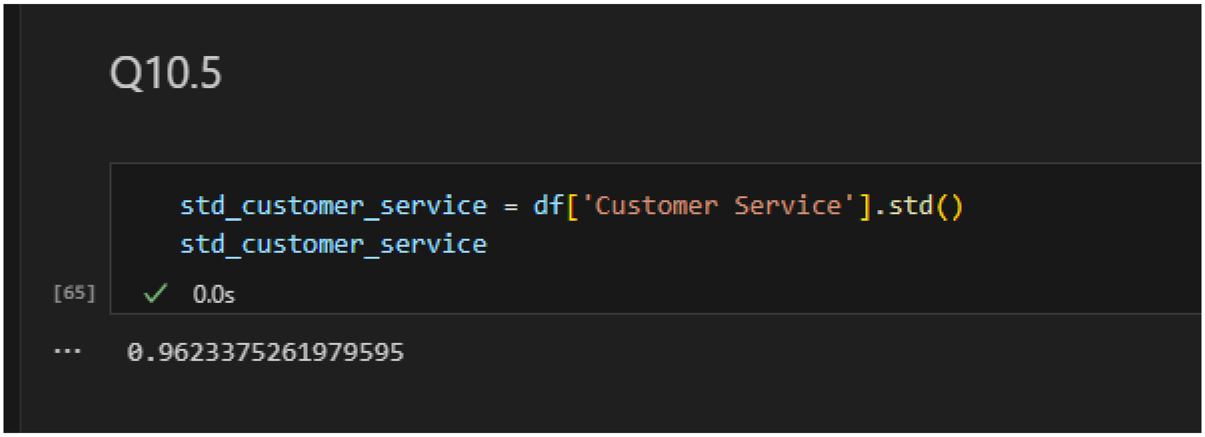
Q10.4

```
mode_signal_strength = df['Signal strength'].mode()[0]
mode_signal_strength
```

```
[64] ✓ 0.0s
```

```
... 3
```

10.5 Standard deviation of Customer Service

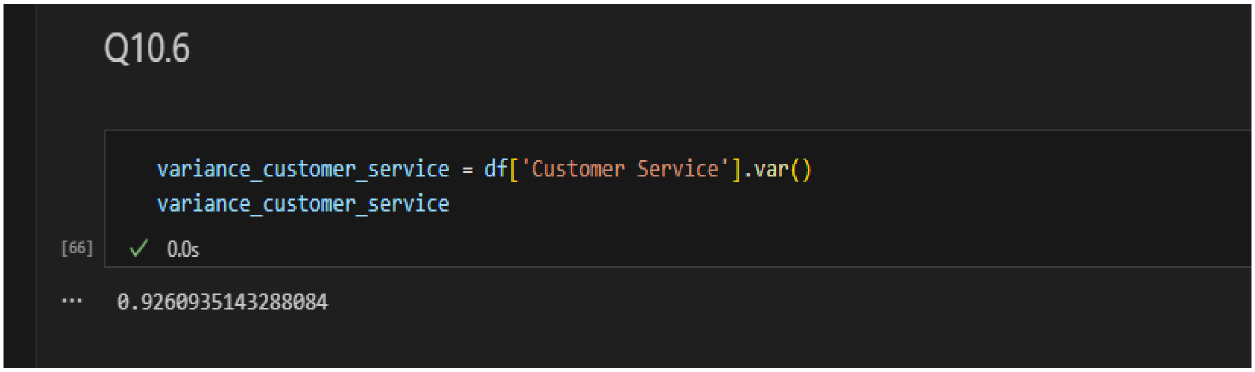


Q10.5

```
std_customer_service = df['Customer Service'].std()
std_customer_service
```

```
0.9623375261979595
```

10.6 Variance of Customer Service

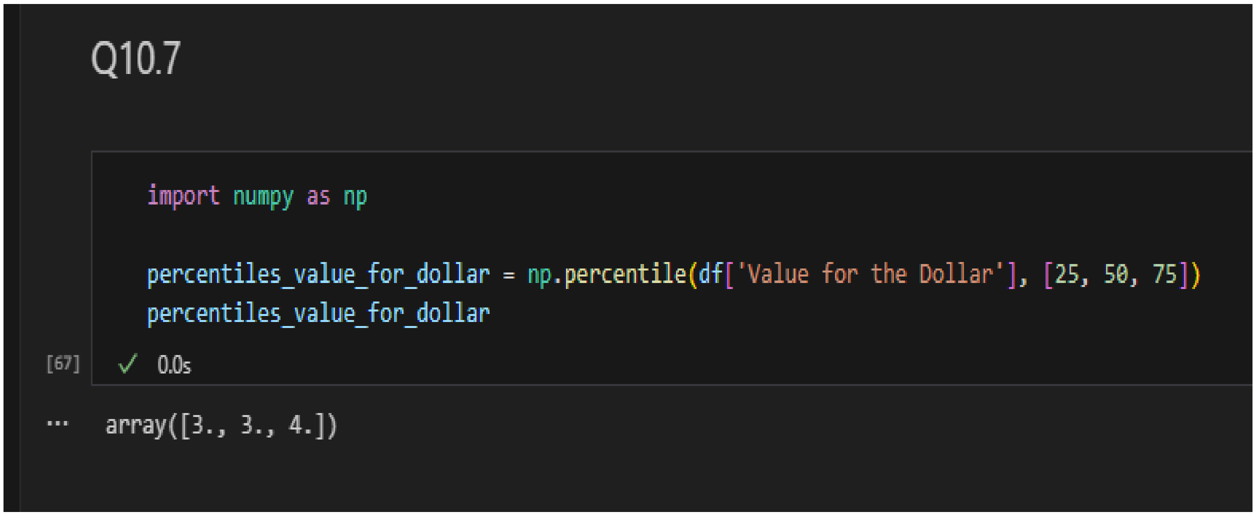


Q10.6

```
variance_customer_service = df['Customer Service'].var()
variance_customer_service
```

```
0.9260935143288084
```

10.7 Percentiles of Value for the Dollar



Q10.7

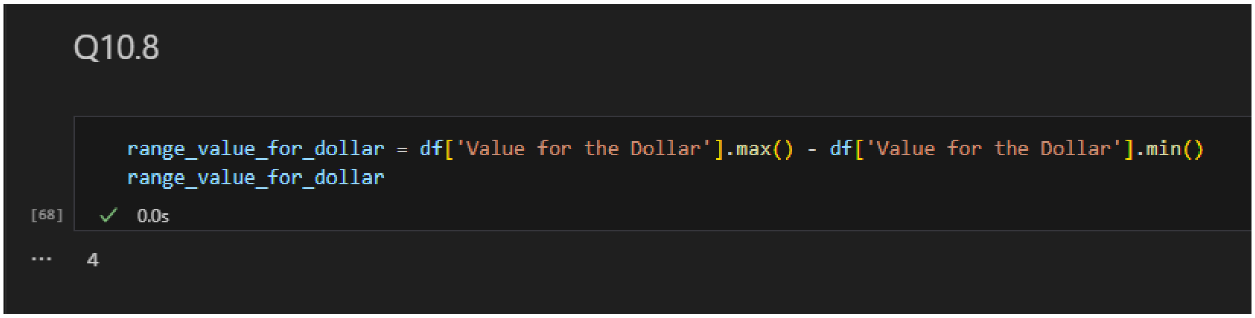
```
import numpy as np

percentiles_value_for_dollar = np.percentile(df['Value for the Dollar'], [25, 50, 75])
percentiles_value_for_dollar
```

[67] ✓ 0.0s

```
··· array([3., 3., 4.])
```

10.8 Range of Value for the Dollar



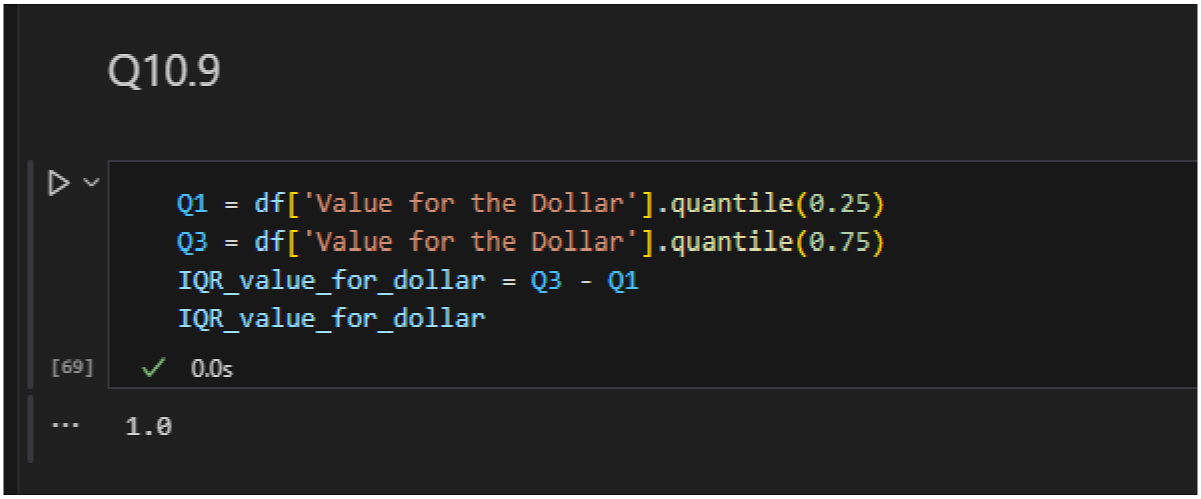
Q10.8

```
range_value_for_dollar = df['Value for the Dollar'].max() - df['Value for the Dollar'].min()
range_value_for_dollar
```

[68] ✓ 0.0s

```
··· 4
```

10.9 IQR of Value for the Dollar



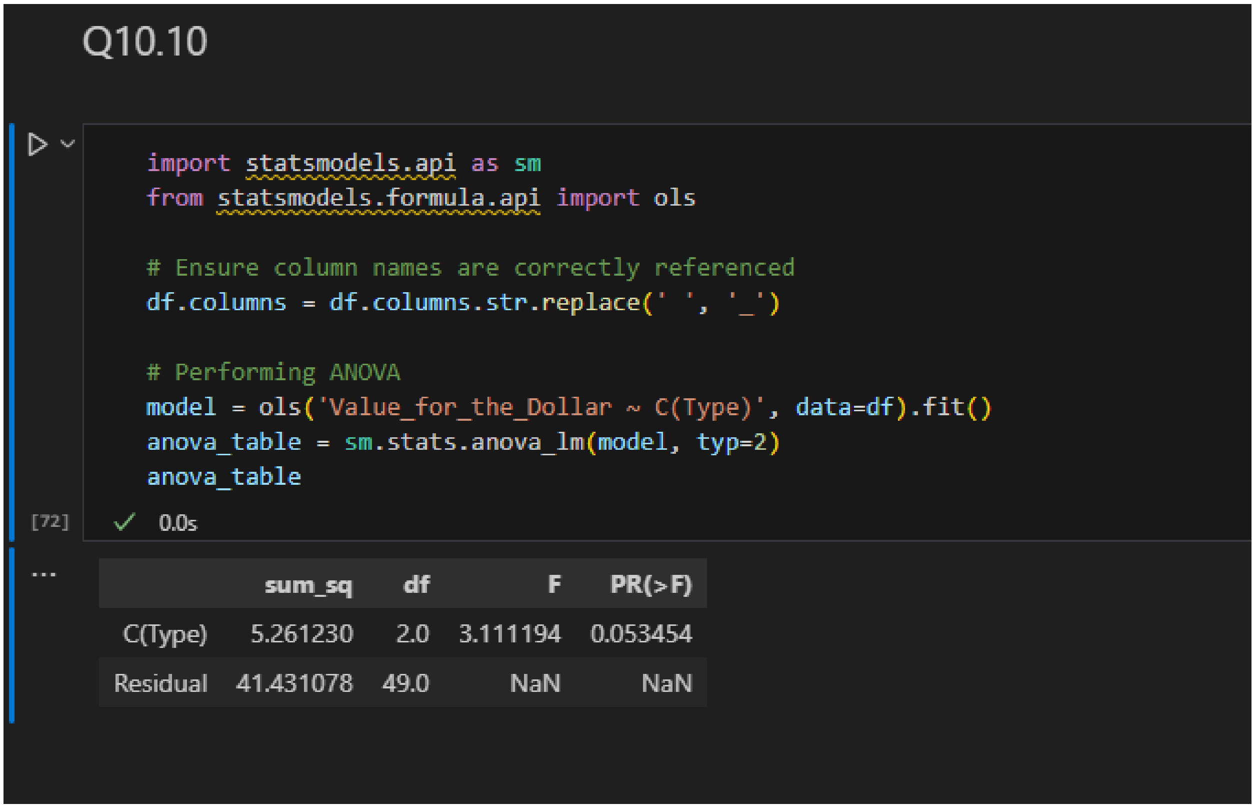
Q10.9

```
Q1 = df['Value for the Dollar'].quantile(0.25)
Q3 = df['Value for the Dollar'].quantile(0.75)
IQR_value_for_dollar = Q3 - Q1
IQR_value_for_dollar
```

[69] ✓ 0.0s

```
··· 1.0
```

10.10 Hypothesis Testing – ANOVA



Q10.10

```python
import statsmodels.api as sm
from statsmodels.formula.api import ols

# Ensure column names are correctly referenced
df.columns = df.columns.str.replace(' ', '_')

# Performing ANOVA
model = ols('Value_for_the_Dollar ~ C(Type)', data=df).fit()
anova_table = sm.stats.anova_lm(model, typ=2)
anova_table
```

```
             sum_sq    df         F    PR(>F)
C(Type)    5.261230   2.0  3.111194  0.053454
Residual  41.431078  49.0       NaN       NaN
```

THANK YOU

RISHI KUMAR MISHRA