Sheth L.U.J College of Arts & Sir M.V. College of Science and Commerce
Data Science
PRACTICAL NO. 4

Aim: Hypothesis Testing -

- **a)** Formulate null and alternative hypotheses for a given problem.
- **b)** Conduct a hypothesis test using appropriate statistical tests (e.g., t-test, chi-square test).
- **c)** Interpret the results and draw conclusions based on the test outcomes.

CODE:

- ***Importing libraries***

```
import numpy as np
import pandas as pd
import scipy.stats as stats
import matplotlib.pyplot as plt
df = pd.read_csv("tip.csv")
df.head()
```

- ***1. One-Sample t-test (Mean Tip Amount)***

```
data = df['tip'].values
print("Data:", data)
H0 = "Average tip = 3 dollars"
H1 = "Average tip ≠ 3 dollars"
t_stat, p_value = stats.ttest_1samp(data, 3)
print("Test Statistic:", t_stat)
print("P-value:", p_value)
if p_value < 0.05:
    print("Reject H0 → Mean tip is significantly different from 3.")
else:
    print("Fail to Reject H0 → No evidence tip is different from 3.")

plt.figure(figsize=(6,5))
plt.hist(data, bins=20)
plt.axvline(3, color='red', linestyle='--', label='Hypothesized Mean = 3')
plt.title("One-Sample t-test: Distribution of Tips")
plt.xlabel("Tip Amount")
plt.ylabel("Frequency")
plt.legend()
plt.show()
```

- ***2. Independent Two-Sample t-test (Male vs Female Tip Amount)***

```
from scipy.stats import ttest_ind
male = df[df['sex']=="Male"]['tip']
female = df[df['sex']=="Female"]['tip']
print("Male mean:", np.mean(male))
print("Female mean:", np.mean(female))
ttest, pval = ttest_ind(male, female)
print("p-value:", pval)
if pval < 0.05:
    print("Reject H0 → Males and Females give different tips.")
else:
    print("Accept H0 → No difference in tip amounts by gender.")

plt.figure(figsize=(6,5))
plt.boxplot([male, female], labels=['Male', 'Female'])
```

```
plt.title("Independent t-test: Tips by Gender")
plt.ylabel("Tip Amount")
plt.show()
```

- ***3. Paired Sample t-test***

```
from scipy import stats
ttest, pval = stats.ttest_rel(df['total_bill'], df['tip'])
print("p-value:", pval)
if pval < 0.05:
   print("Reject H0 → Significant difference between total bill & tip.")
else:
   print("Accept H0 → No significant difference.")

plt.figure(figsize=(6,5))
plt.scatter(df['total_bill'], df['tip'])
plt.title("Paired t-test: Total Bill vs Tip")
plt.xlabel("Total Bill")
plt.ylabel("Tip")
plt.show()
```

- ***4. One-Sample Z-test (Using Total Bill)***

```
from statsmodels.stats import weightstats as stests
ztest, pval = stests.ztest(df['total_bill'], value=20)
print("p-value:", float(pval))
if pval < 0.05:
   print("Reject H0 → Mean total bill ≠ 20.")
else:
   print("Accept H0 → Evidence supports mean = 20.")

plt.figure(figsize=(6,5))
plt.hist(df['total_bill'], bins=20)
plt.axvline(20, color='red', linestyle='--', label='Hypothesized Mean = 20')
plt.title("Z-Test: Distribution of Total Bill")
plt.xlabel("Total Bill")
plt.ylabel("Frequency")
plt.legend()
plt.show()
```

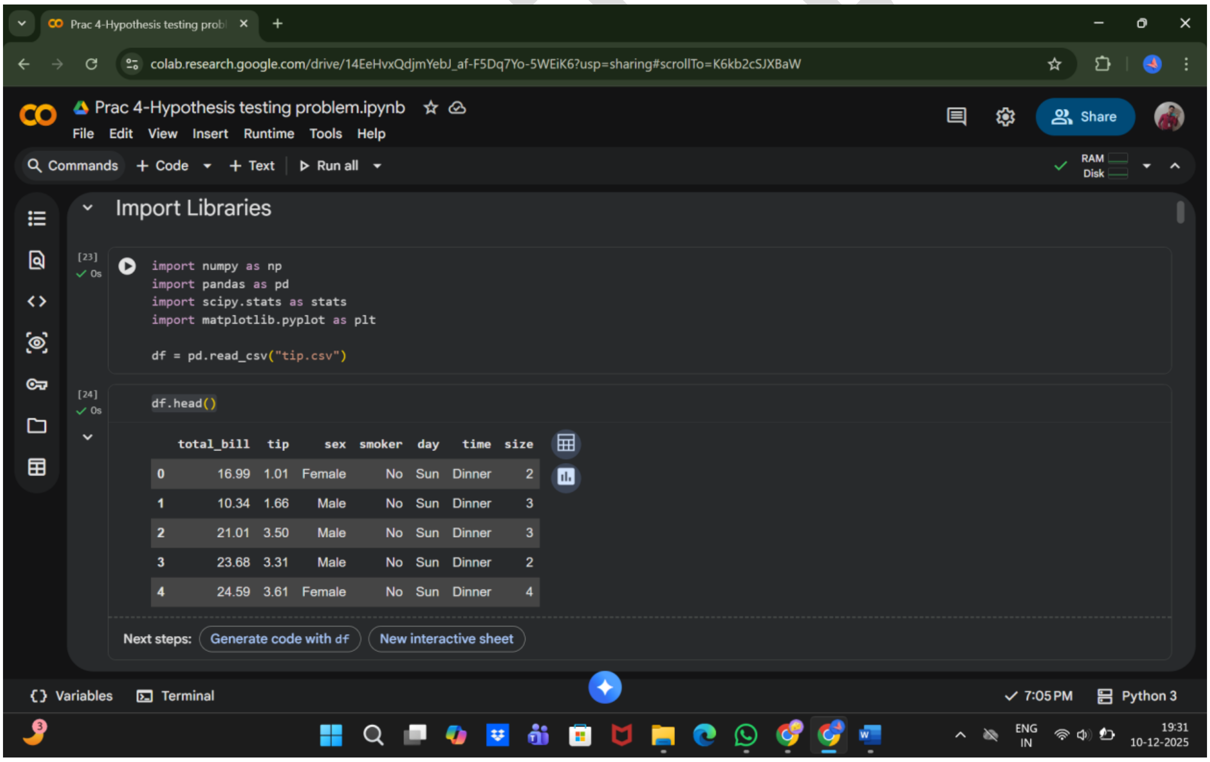
- ***5. Two-Sample Z-test (Smoker vs Non-Smoker Total Bill)***

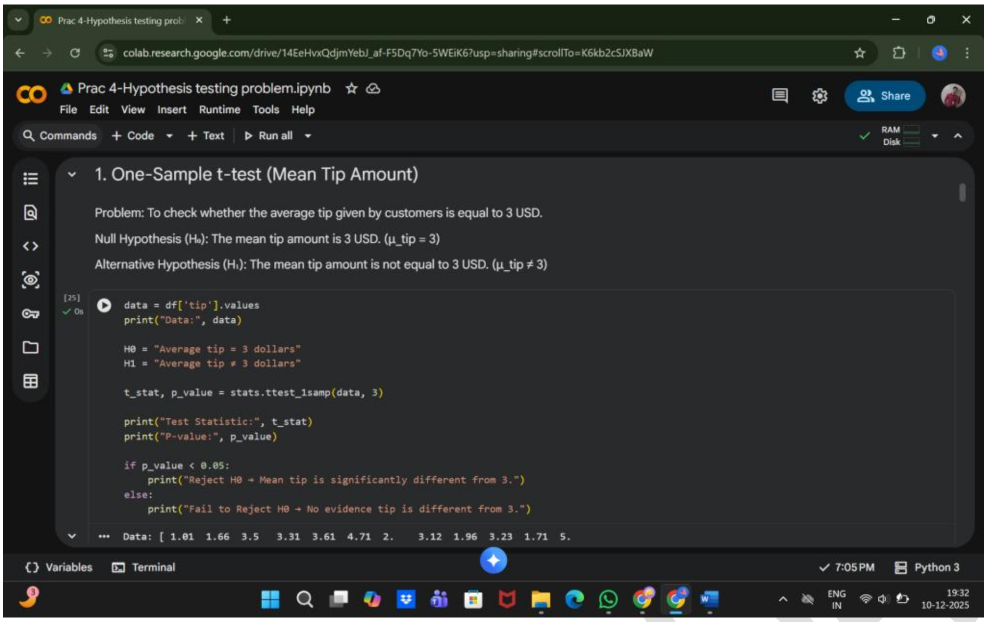
```
smoker = df[df['smoker']=="Yes"]['total_bill']
non_smoker = df[df['smoker']=="No"]['total_bill']
ztest, pval = stests.ztest(smoker, x2=non_smoker, value=0)
print("p-value:", float(pval))
if pval < 0.05:
   print("Reject H0 → Smokers & Non-smokers differ in spending.")
else:
   print("Accept H0 → No difference in spending.")

plt.figure(figsize=(6,5))
plt.boxplot([smoker, non_smoker], labels=['Smoker', 'Non-Smoker'])
plt.title("Two-Sample Z-Test: Total Bill Comparison")
plt.ylabel("Total Bill")
plt.show()
```

➢ *6. Chi-Square Test (Gender vs Smoker Status)*

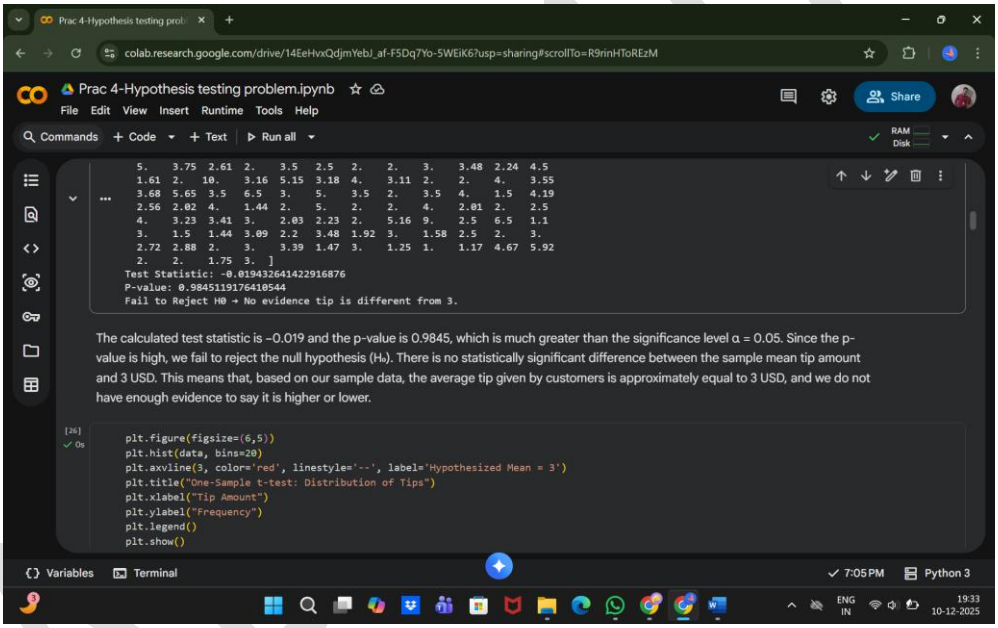
```
from scipy.stats import chi2_contingency
table = pd.crosstab(df['sex'], df['smoker'])
print("Contingency Table:\n", table)
chi2, pval, dof, expected = chi2_contingency(table)
print("Chi-square:", chi2)
print("p-value:", pval)
print("Degrees of freedom:", dof)
print("Expected frequencies:\n", expected)
if pval < 0.05:
   print("Reject H0 → Gender and Smoking are dependent.")
else:
   print("Accept H0 → Gender and Smoking are independent.")

plt.figure(figsize=(6,5))
table.plot(kind='bar', figsize=(6,5))
plt.title("Chi-Square Test: Gender vs Smoker Status")
plt.xlabel("Gender")
plt.ylabel("Count")
plt.show()
```


Output:

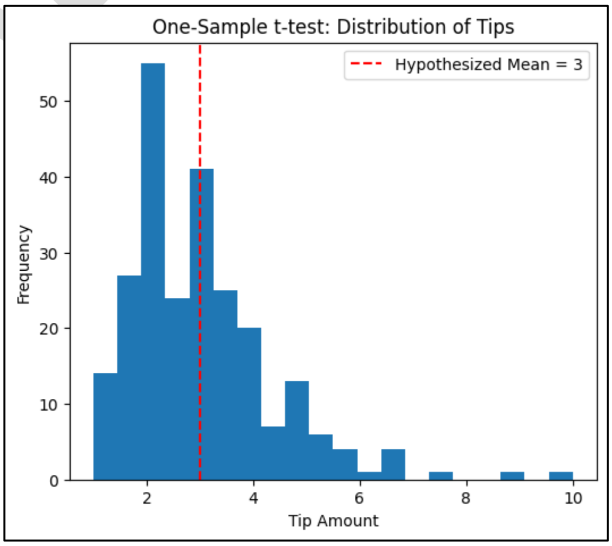




Prac 4-Hypothesis testing problem.ipynb
1. One-Sample t-test (Mean Tip Amount)
Problem: To check whether the average tip given by customers is equal to 3 USD.
Null Hypothesis (H₀): The mean tip amount is 3 USD. (μ_tip = 3)
Alternative Hypothesis (H₁): The mean tip amount is not equal to 3 USD. (μ_tip ≠ 3)
data = df['tip'].values
print("Data:", data)
H0 = "Average tip = 3 dollars"
H1 = "Average tip ≠ 3 dollars"
t_stat, p_value = stats.ttest_1samp(data, 3)
print("Test Statistic:", t_stat)
print("P-value:", p_value)
if p_value < 0.05:
print("Reject H0 → Mean tip is significantly different from 3.")
else:
print("Fail to Reject H0 → No evidence tip is different from 3.")
Data: [1.01 1.66 3.5 3.31 3.61 4.71 2. 3.12 1.96 3.23 1.71 5.



Prac 4-Hypothesis testing problem.ipynb
5. 3.75 2.61 2. 3.5 2.5 2. 2. 3. 3.48 2.24 4.5
1.61 2. 10. 3.16 5.15 3.18 4. 3.11 2. 2. 4. 3.55
3.68 5.65 3.5 6.5 3. 5. 3.5 2. 3.5 4. 1.5 4.19
2.56 2.02 4. 1.44 2. 5. 2. 2. 4. 2.01 2. 2.5
4. 3.23 3.41 3. 2.03 2.23 2. 5.16 9. 2.5 6.5 1.1
3. 1.5 1.44 3.09 2.2 3.48 1.92 3. 1.58 2.5 2. 3.
2.72 2.88 2. 3. 3.39 1.47 3. 1.25 1. 1.17 4.67 5.92
2. 2. 1.75 3.]
Test Statistic: -0.019432641422916876
P-value: 0.9845119176410544
Fail to Reject H0 → No evidence tip is different from 3.
The calculated test statistic is -0.019 and the p-value is 0.9845, which is much greater than the significance level α = 0.05. Since the p-value is high, we fail to reject the null hypothesis (H₀). There is no statistically significant difference between the sample mean tip amount and 3 USD. This means that, based on our sample data, the average tip given by customers is approximately equal to 3 USD, and we do not have enough evidence to say it is higher or lower.
plt.figure(figsize=(6,5))
plt.hist(data, bins=20)
plt.axvline(3, color='red', linestyle='--', label='Hypothesized Mean = 3')
plt.title("One-Sample t-test: Distribution of Tips")
plt.xlabel("Tip Amount")
plt.ylabel("Frequency")
plt.legend()
plt.show()

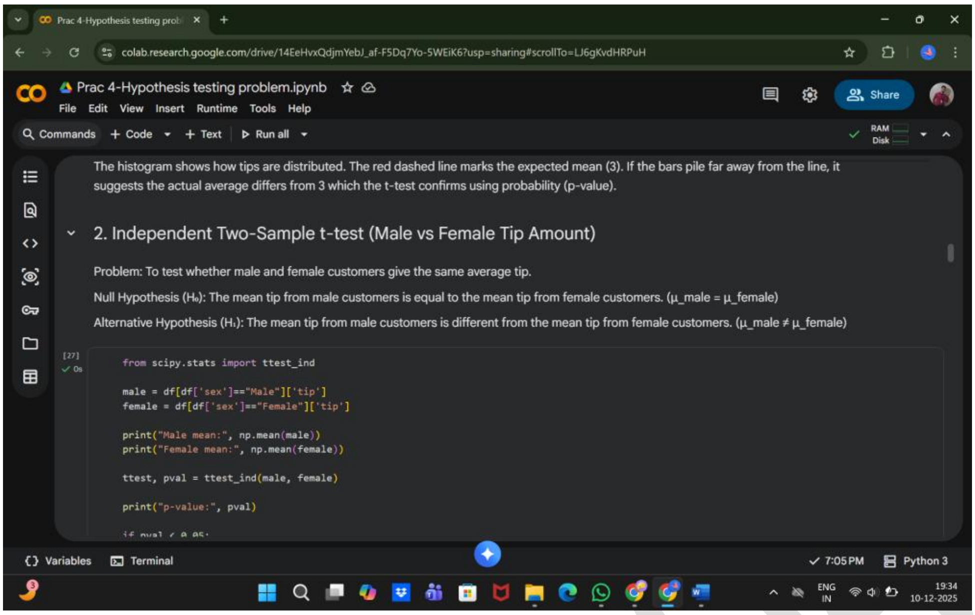


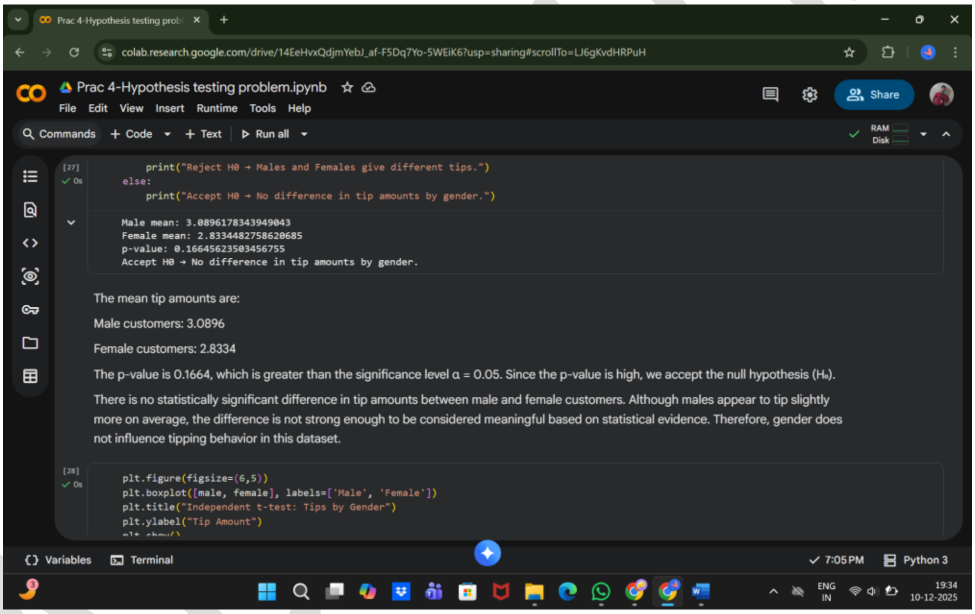
One-Sample t-test: Distribution of Tips
Hypothesized Mean = 3
Frequency
Tip Amount
0
10
20
30
40
50
2
4
6
8
10

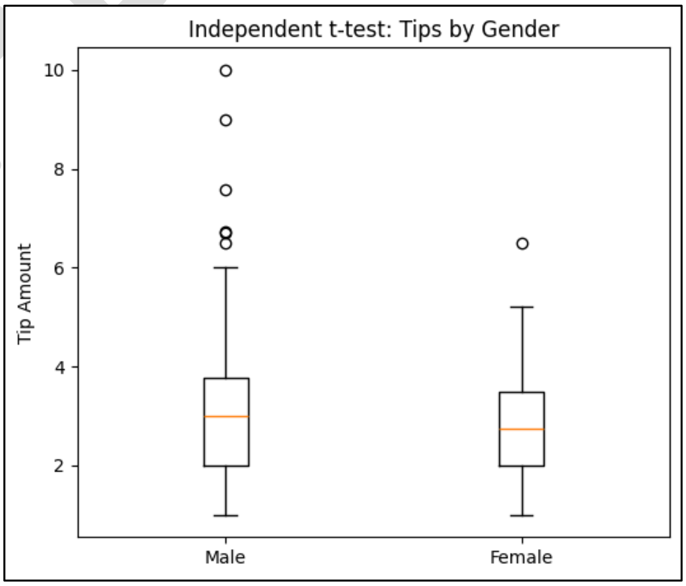
Sheth L.U.J College of Arts & Sir M.V. College of Science and Commerce

Data Science

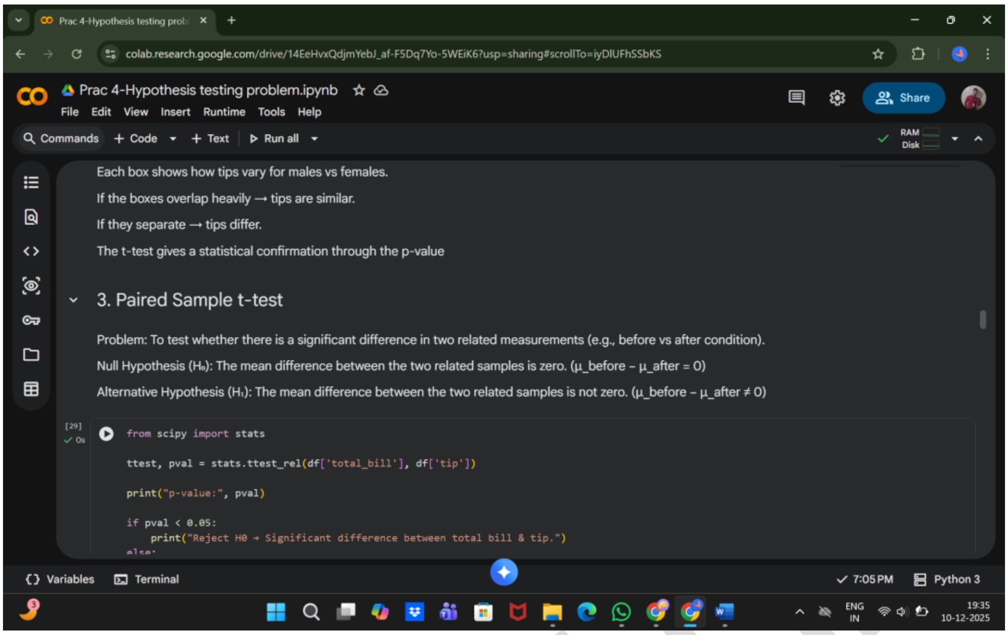
PRACTICAL NO. 4

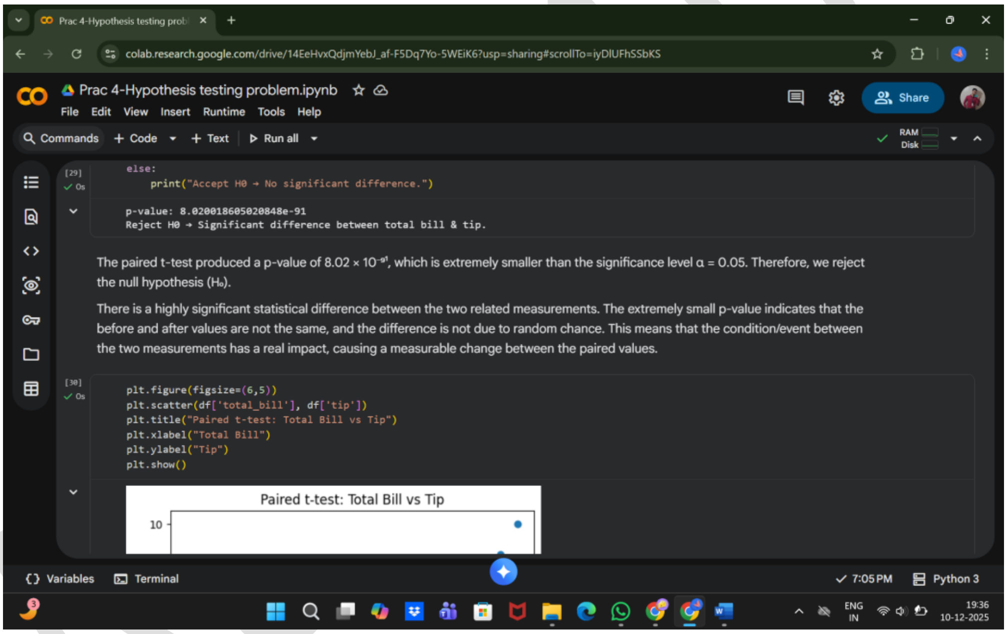


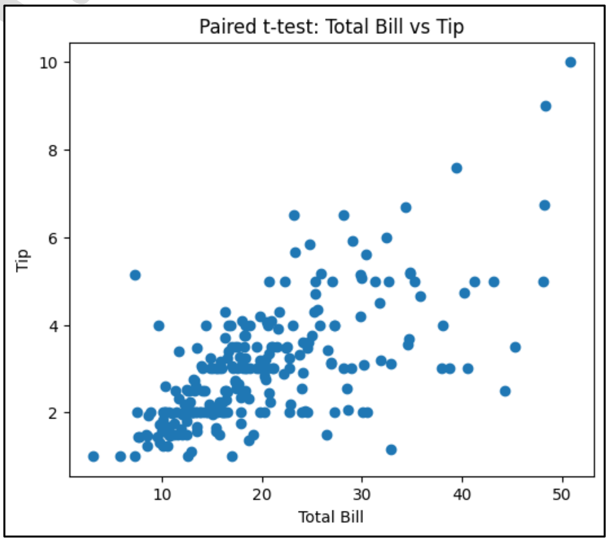




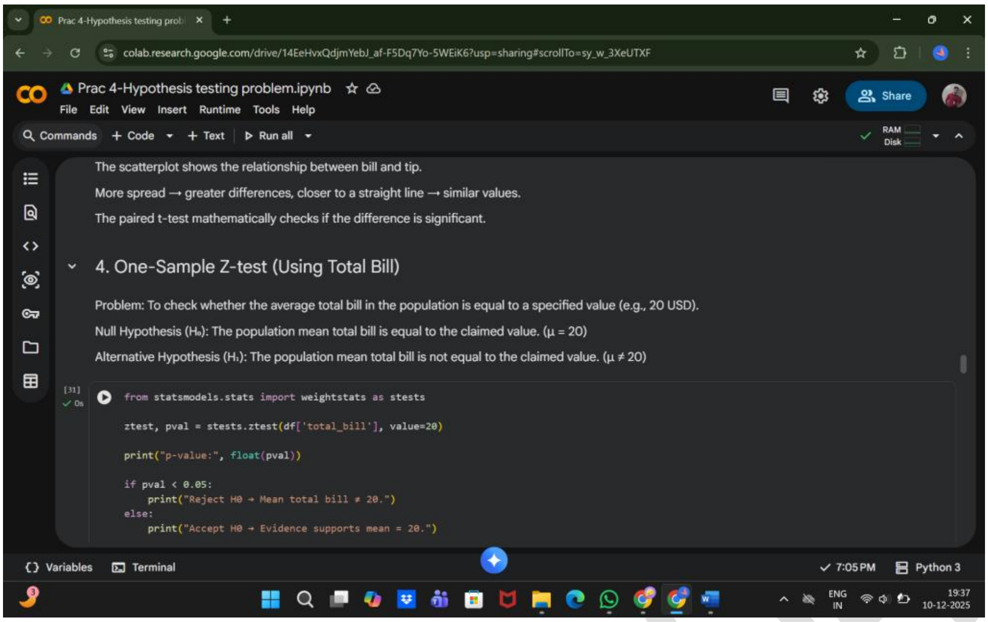
Sheth L.U.J College of Arts & Sir M.V. College of Science and Commerce
Data Science
PRACTICAL NO. 4

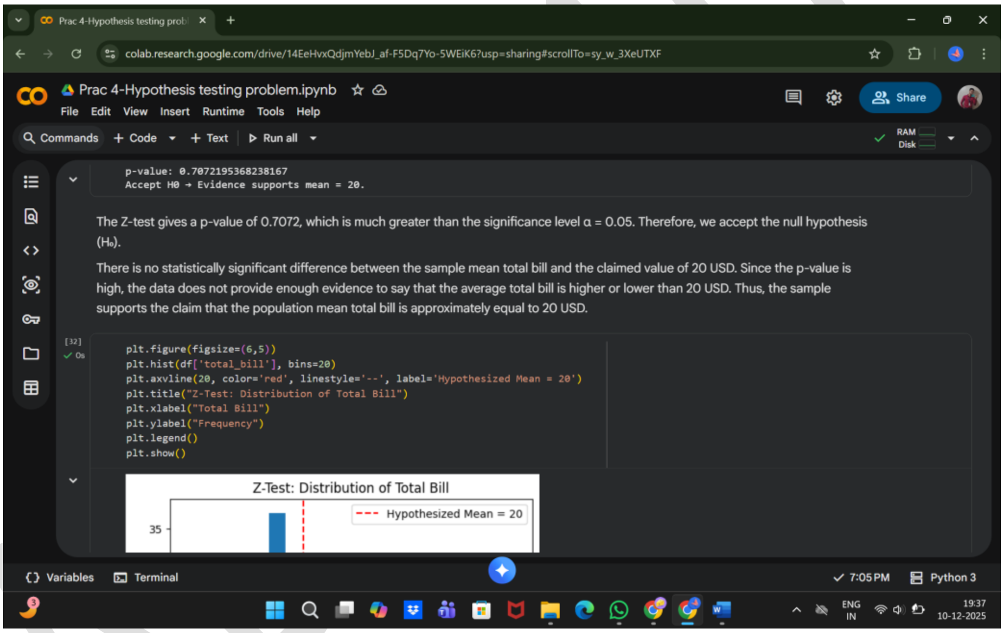


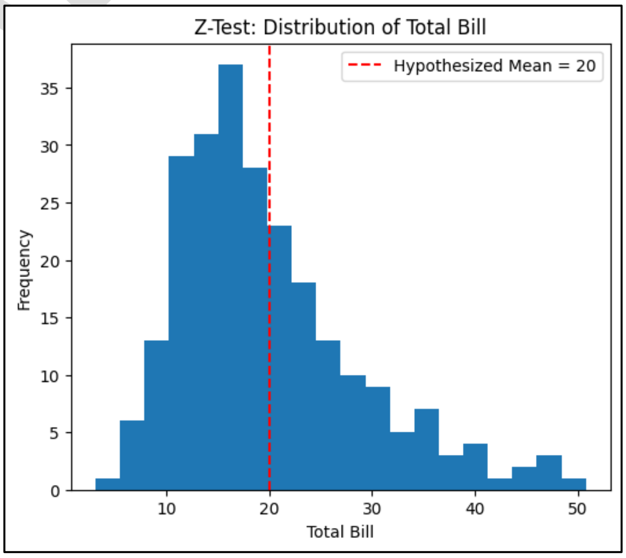




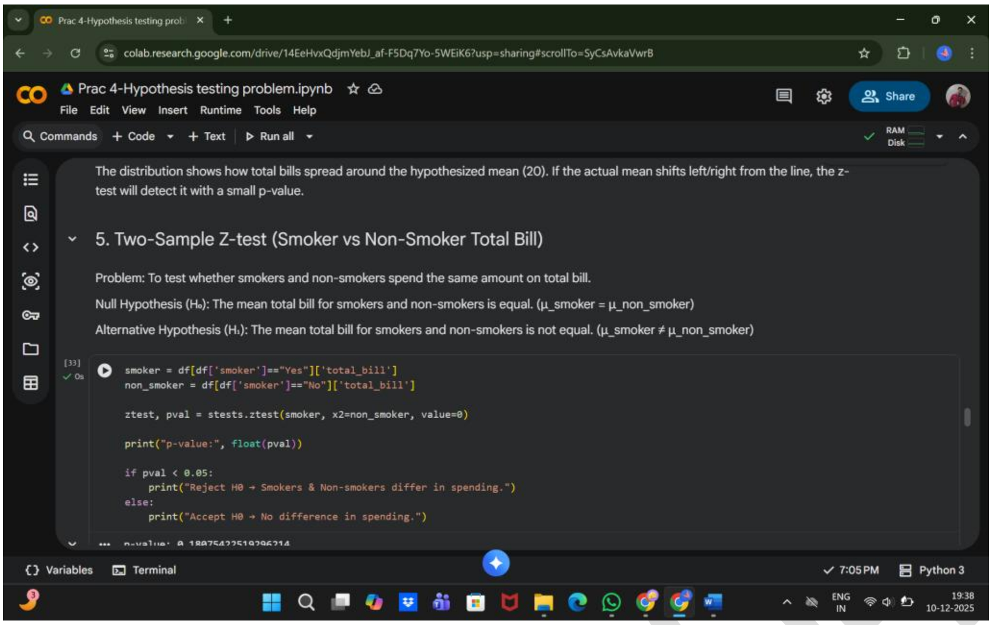
PRACTICAL NO. 4

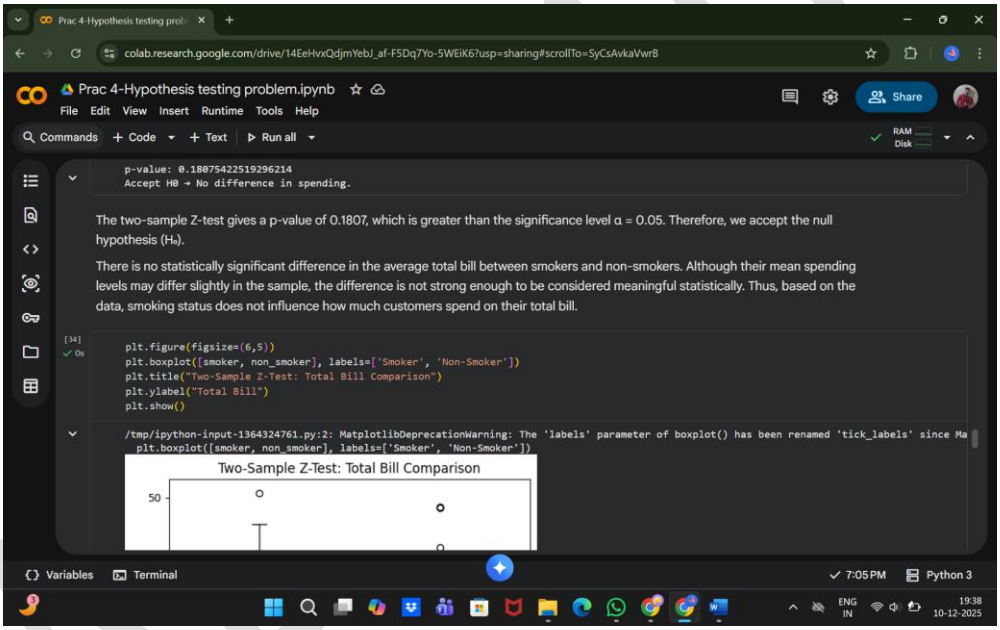


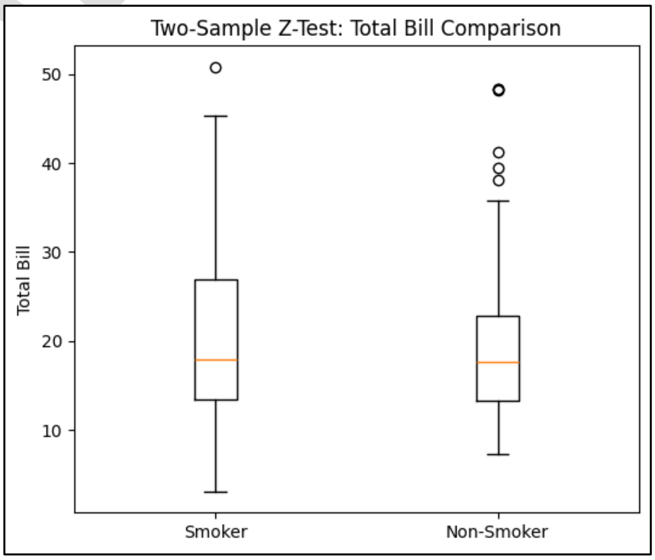


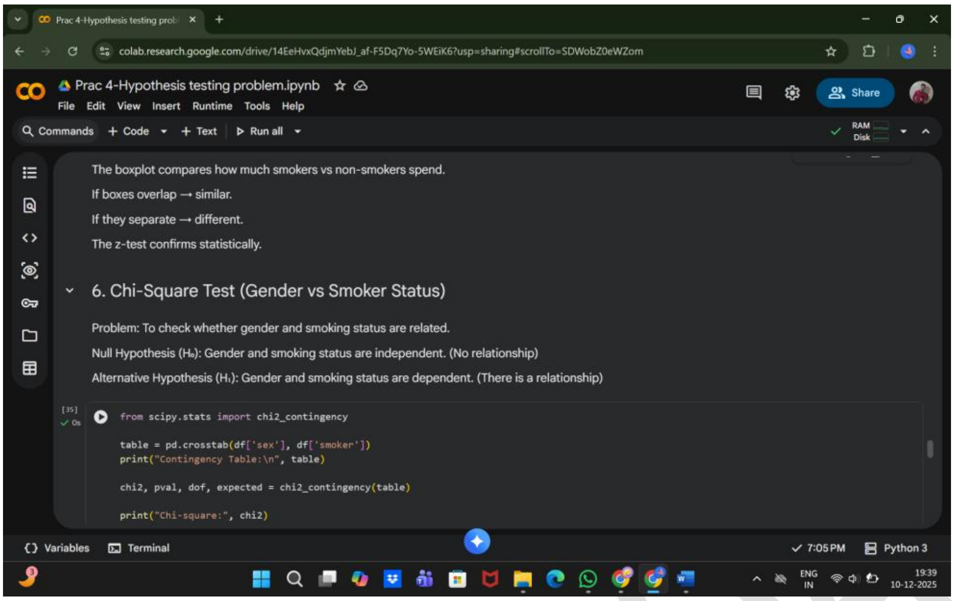


PRACTICAL NO. 4

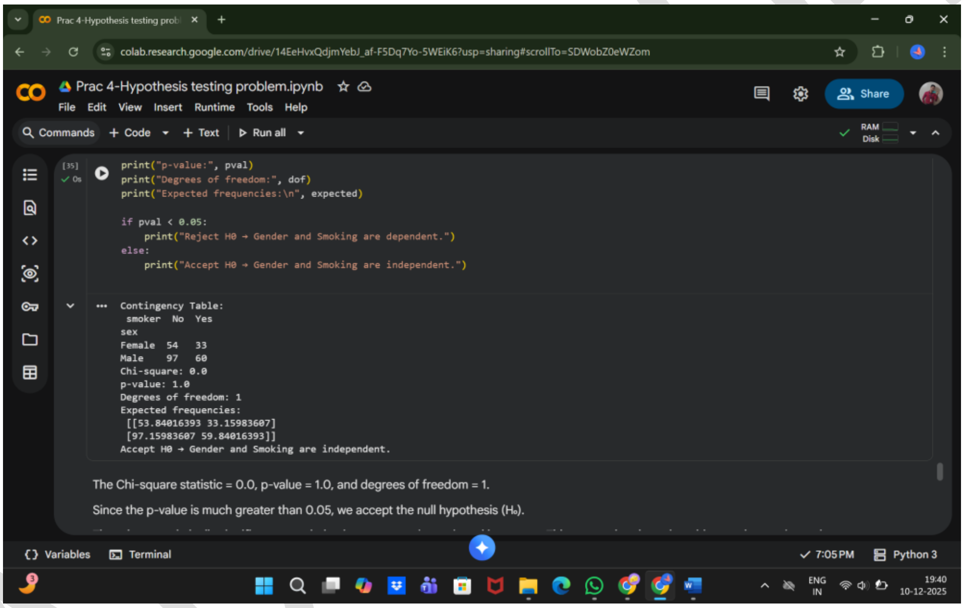




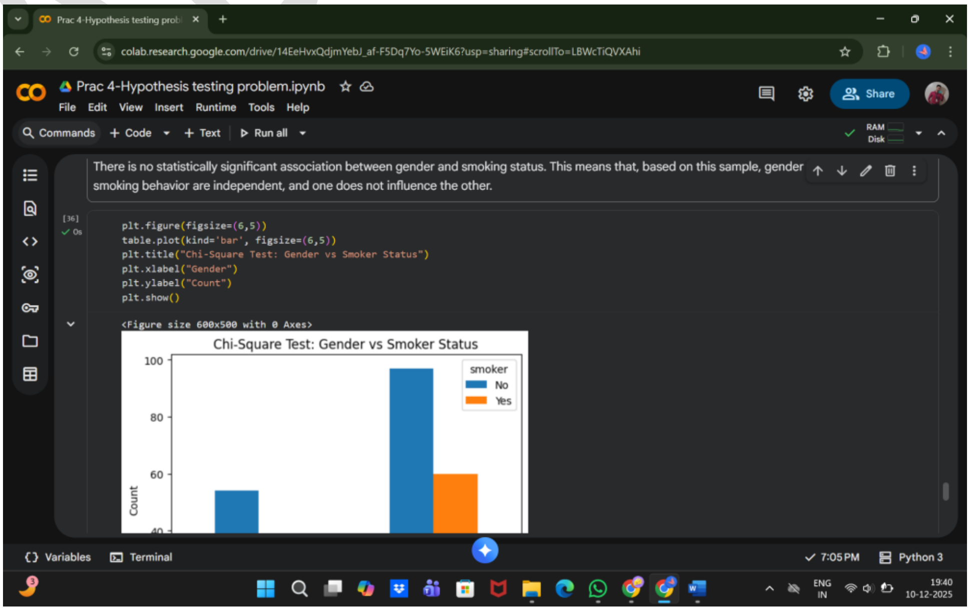




Prac 4-Hypothesis testing problem.ipynb
The boxplot compares how much smokers vs non-smokers spend.
If boxes overlap → similar.
If they separate → different.
The z-test confirms statistically.
6. Chi-Square Test (Gender vs Smoker Status)
Problem: To check whether gender and smoking status are related.
Null Hypothesis (H₀): Gender and smoking status are independent. (No relationship)
Alternative Hypothesis (H₁): Gender and smoking status are dependent. (There is a relationship)
[35]
from scipy.stats import chi2_contingency
table = pd.crosstab(df['sex'], df['smoker'])
print("Contingency Table:\n", table)
chi2, pval, dof, expected = chi2_contingency(table)
print("Chi-square:", chi2)



Prac 4-Hypothesis testing problem.ipynb
[35]
print("p-value:", pval)
print("Degrees of freedom:", dof)
print("Expected frequencies:\n", expected)
if pval < 0.05:
print("Reject H0 → Gender and Smoking are dependent.")
else:
print("Accept H0 → Gender and Smoking are independent.")
Contingency Table:
smoker No Yes
sex
Female 54 33
Male 97 60
Chi-square: 0.0
p-value: 1.0
Degrees of freedom: 1
Expected frequencies:
[[53.84016393 33.15983607]
[97.15983607 59.84016393]]
Accept H0 → Gender and Smoking are independent.
The Chi-square statistic = 0.0, p-value = 1.0, and degrees of freedom = 1.
Since the p-value is much greater than 0.05, we accept the null hypothesis (H₀).



Prac 4-Hypothesis testing problem.ipynb
There is no statistically significant association between gender and smoking status. This means that, based on this sample, gender
smoking behavior are independent, and one does not influence the other.
[36]
plt.figure(figsize=(6,5))
table.plot(kind='bar', figsize=(6,5))
plt.title("Chi-Square Test: Gender vs Smoker Status")
plt.xlabel("Gender")
plt.ylabel("Count")
plt.show()
<Figure size 600x500 with 0 Axes>
Chi-Square Test: Gender vs Smoker Status
100
80
60
40
Count
smoker
No
Yes

PRACTICAL NO. 4