Sheth L.U.J College of Arts & Sir M.V. College of Science and Commerce
Data Science
PRACTICAL NO. 5

Aim: ANOVA (Analysis of Variance)

a) Perform one-way ANOVA to compare means across multiple groups.
b) Conduct post-hoc tests to identify significant differences between group means.

CODE:

- ***Importing libraries***

```
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
from scipy import stats
import statsmodels.api as sm
from statsmodels.formula.api import ols
from statsmodels.stats.multicomp import pairwise_tukeyhsd
```

- ***Load Dataset***

```
df = pd.read_csv("tip.csv")
print(df.head())
```

- ***One-Way ANOVA***

```
days = df['day'].unique()
data = {day: df['total_bill'][df['day'] == day] for day in days}
F_stat, p_value = stats.f_oneway(data['Thur'], data['Fri'], data['Sat'], data['Sun'])
print("F-statistic:", F_stat)
print("p-value:", p_value)
if p_value < 0.05:
   print("Reject null hypothesis → Total bill differs significantly across days.")
else:
   print("Fail to reject null hypothesis → Total bill does not differ significantly across days.")

plt.figure(figsize=(8,5))
sns.boxplot(x='day', y='total_bill', data=df, palette='Set2')
plt.title("Total Bill Across Days")
plt.show()
```

- ***Two-Way ANOVA***

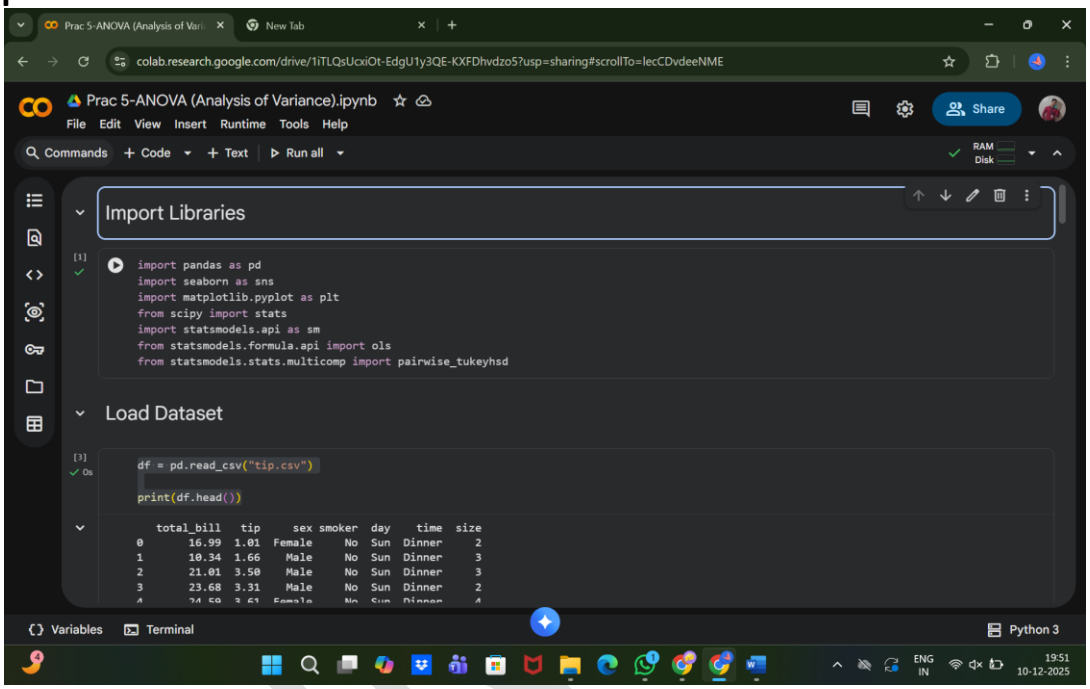
```
# Build the model
model = ols('total_bill ~ C(day) * C(smoker)', data=df).fit()
# ANOVA table
anova_table = sm.stats.anova_lm(model, typ=2)
print(anova_table)

# Interaction plot using seaborn
plt.figure(figsize=(8,5))
sns.pointplot(x='day', y='total_bill', hue='smoker', data=df, palette='Set1', dodge=True, markers=["o", "s"],
capsize=0.1)
plt.title("Interaction Plot: Total Bill by Day and Smoker Status")
plt.show()
```

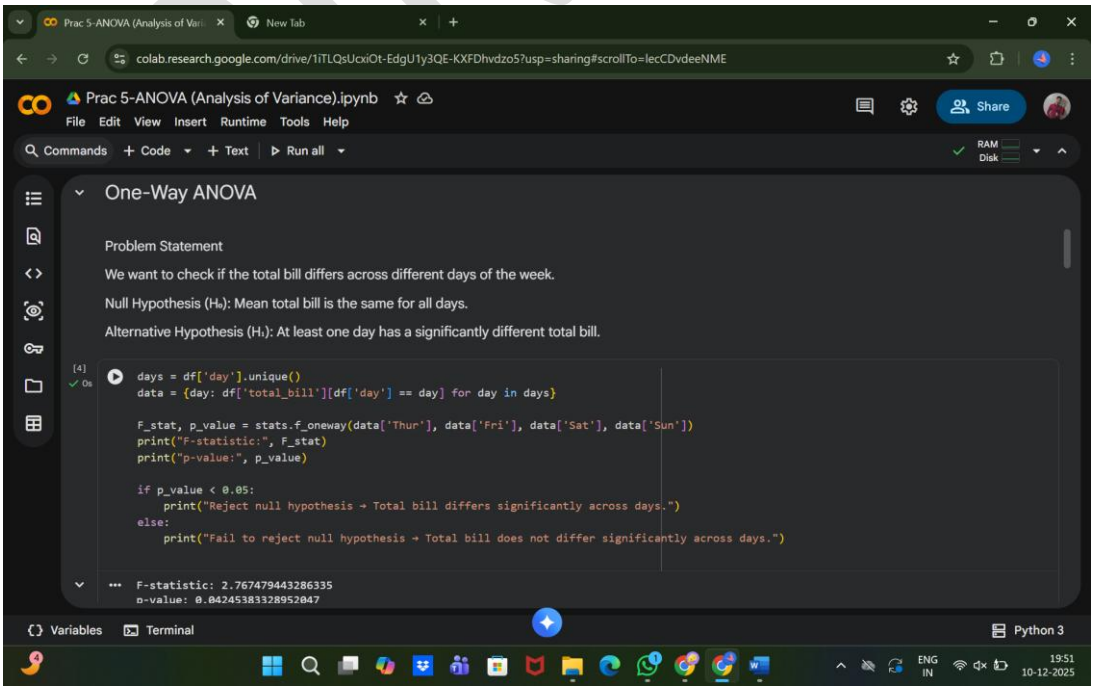
- ***Post-Hoc Test: Tukey HSD***

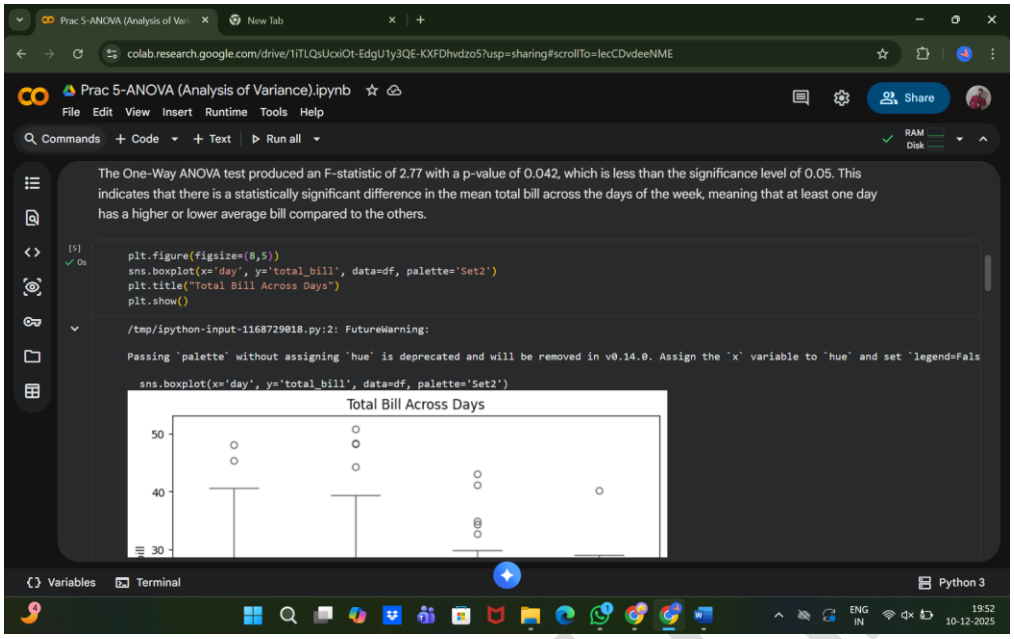
```
# Tukey HSD test for pairwise comparison
```

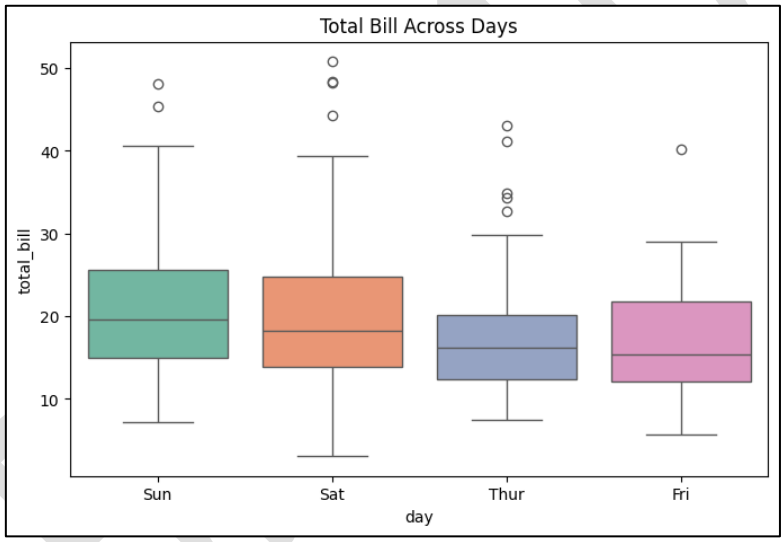
```
tukey = pairwise_tukeyhsd(endog=df['total_bill'],
                          groups=df['day'],
                          alpha=0.05)
print(tukey)

# Plotting Tukey results
tukey.plot_simultaneous(figsize=(8,5))
plt.title("Tukey HSD: Total Bill Comparison Across Days")
plt.show()
```

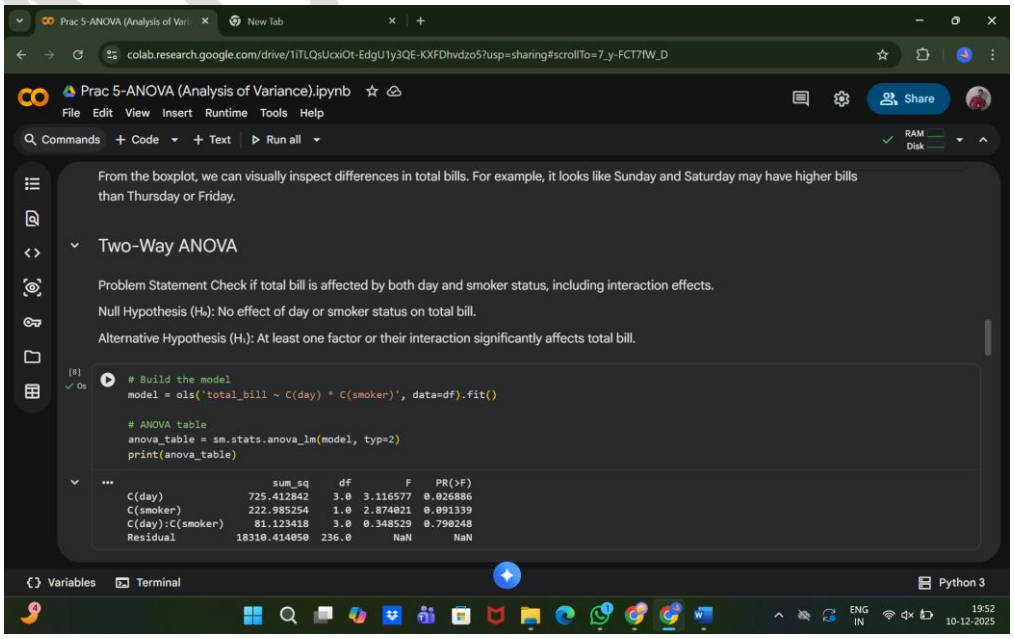
Output:

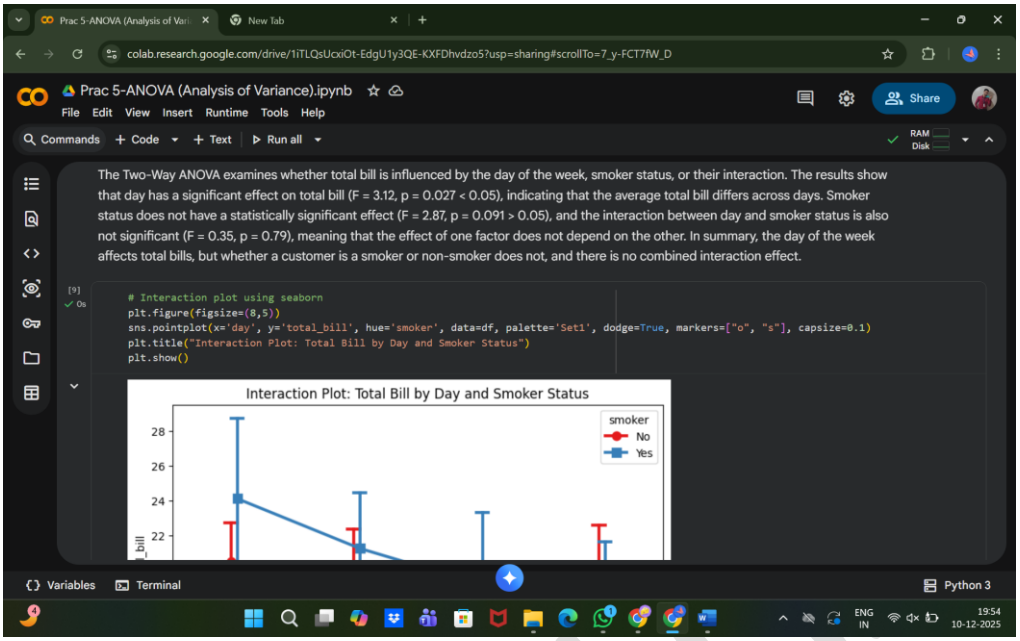


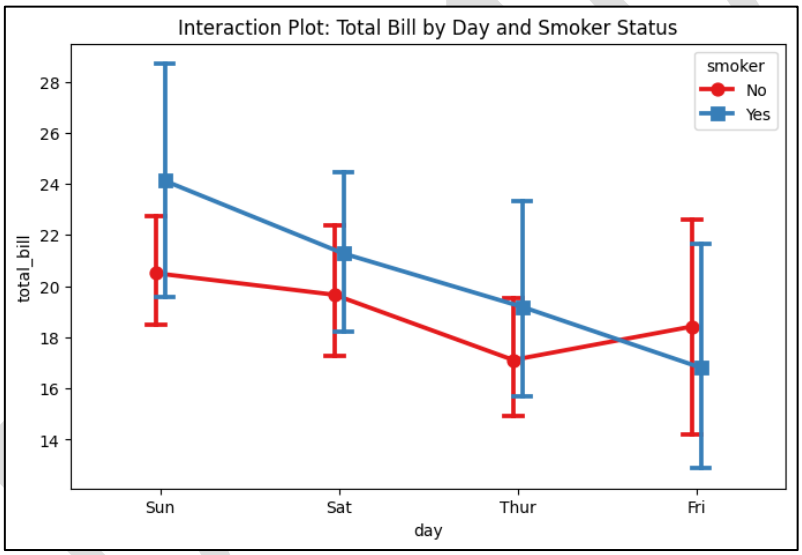


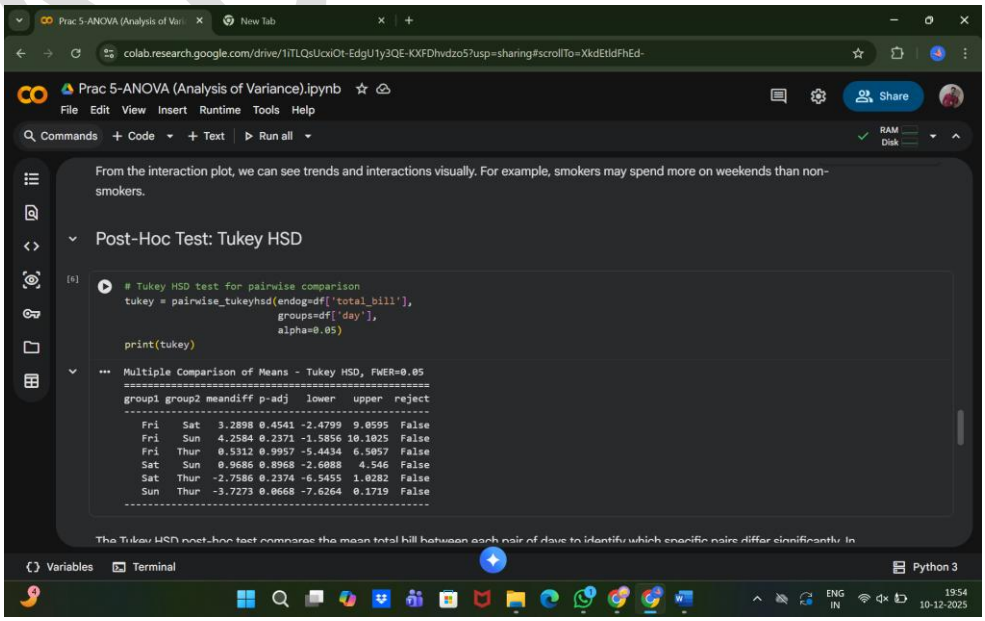

Prac 5-ANOVA (Analysis of Variance).ipynb
File Edit View Insert Runtime Tools Help
Share
Commands + Code + Text Run all
The One-Way ANOVA test produced an F-statistic of 2.77 with a p-value of 0.042, which is less than the significance level of 0.05. This indicates that there is a statistically significant difference in the mean total bill across the days of the week, meaning that at least one day has a higher or lower average bill compared to the others.
plt.figure(figsize=(8,5))
sns.boxplot(x='day', y='total_bill', data=df, palette='Set2')
plt.title("Total Bill Across Days")
plt.show()
/tmp/ipython-input-1168729018.py:2: FutureWarning:
Passing `palette` without assigning `hue` is deprecated and will be removed in v0.14.0. Assign the `x` variable to `hue` and set `legend=Fals
sns.boxplot(x='day', y='total_bill', data=df, palette='Set2')
Total Bill Across Days
50
40
30
Variables
Terminal
Python 3


Total Bill Across Days
total_bill
50
40
30
20
10
Sun
Sat
Thur
Fri
day


Prac 5-ANOVA (Analysis of Variance).ipynb
File Edit View Insert Runtime Tools Help
Share
Commands + Code + Text Run all
From the boxplot, we can visually inspect differences in total bills. For example, it looks like Sunday and Saturday may have higher bills than Thursday or Friday.
Two-Way ANOVA
Problem Statement Check if total bill is affected by both day and smoker status, including interaction effects.
Null Hypothesis (H₀): No effect of day or smoker status on total bill.
Alternative Hypothesis (H₁): At least one factor or their interaction significantly affects total bill.
# Build the model
model = ols('total_bill ~ C(day) * C(smoker)', data=df).fit()
# ANOVA table
anova_table = sm.stats.anova_lm(model, typ=2)
print(anova_table)
sum_sq df F PR(>F)
C(day) 725.412842 3.0 3.116577 0.026886
C(smoker) 222.985254 1.0 2.874021 0.091339
C(day):C(smoker) 81.123418 3.0 0.348529 0.790248
Residual 18310.414050 236.0 NaN NaN
Variables
Terminal
Python 3


Prac 5-ANOVA (Analysis of Variance).ipynb
File Edit View Insert Runtime Tools Help
Share
Commands + Code + Text Run all
The Two-Way ANOVA examines whether total bill is influenced by the day of the week, smoker status, or their interaction. The results show that day has a significant effect on total bill (F = 3.12, p = 0.027 < 0.05), indicating that the average total bill differs across days. Smoker status does not have a statistically significant effect (F = 2.87, p = 0.091 > 0.05), and the interaction between day and smoker status is also not significant (F = 0.35, p = 0.79), meaning that the effect of one factor does not depend on the other. In summary, the day of the week affects total bills, but whether a customer is a smoker or non-smoker does not, and there is no combined interaction effect.
# Interaction plot using seaborn
plt.figure(figsize=(8,5))
sns.pointplot(x='day', y='total_bill', hue='smoker', data=df, palette='Set1', dodge=True, markers=["o", "s"], capsize=0.1)
plt.title("Interaction Plot: Total Bill by Day and Smoker Status")
plt.show()
Interaction Plot: Total Bill by Day and Smoker Status
smoker
No
Yes
Variables Terminal
Python 3


Interaction Plot: Total Bill by Day and Smoker Status
smoker
No
Yes
total_bill
day
Sun
Sat
Thur
Fri


Prac 5-ANOVA (Analysis of Variance).ipynb
File Edit View Insert Runtime Tools Help
Share
From the interaction plot, we can see trends and interactions visually. For example, smokers may spend more on weekends than non-smokers.
Post-Hoc Test: Tukey HSD
# Tukey HSD test for pairwise comparison
tukey = pairwise_tukeyhsd(endog=df['total_bill'],
groups=df['day'],
alpha=0.05)
print(tukey)
Multiple Comparison of Means - Tukey HSD, FWER=0.05
group1 group2 meandiff p-adj lower upper reject
Fri Sat 3.2898 0.4541 -2.4799 9.0595 False
Fri Sun 4.2584 0.2371 -1.5856 10.1025 False
Fri Thur 0.5312 0.9957 -5.4434 6.5057 False
Sat Sun 0.9686 0.8968 -2.6088 4.546 False
Sat Thur -2.7586 0.2374 -6.5455 1.0282 False
Sun Thur -3.7273 0.0668 -7.6264 0.1719 False
Variables Terminal
Python 3

Sheth L.U.J College of Arts & Sir M.V. College of Science and Commerce
Data Science
PRACTICAL NO. 5