

**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
```

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

## Data Science

### PRACTICAL NO. 5

```
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:

The screenshot shows a Google Colab notebook titled "Prac 5-ANOVA (Analysis of Variance).ipynb". The code cell [1] imports pandas, seaborn, matplotlib.pyplot, and other statistical modules. The code cell [3] reads the "tip.csv" dataset and prints its head. The output shows the first few rows of the dataset.

```
[1]: 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

[3]: df = pd.read_csv("tip.csv")
print(df.head())

   total_bill  tip  sex smoker  day    time  size
0       16.99  1.01  Female     No  Sun Dinner     2
1       10.34  1.66    Male     No  Sun Dinner     3
2       21.01  3.50    Male     No  Sun Dinner     3
3       23.68  3.31    Male     No  Sun Dinner     2
4       24.00  3.61  Female     No  Sun Dinner     4
```

The screenshot shows a Google Colab notebook titled "Prac 5-ANOVA (Analysis of Variance).ipynb". The code cell [4] performs a one-way ANOVA test on the "total\_bill" column across different days of the week. The output shows the F-statistic, p-value, and a comparison of the results against the null hypothesis.

```
[4]: days = df['day'].unique()
data = {day: df[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)

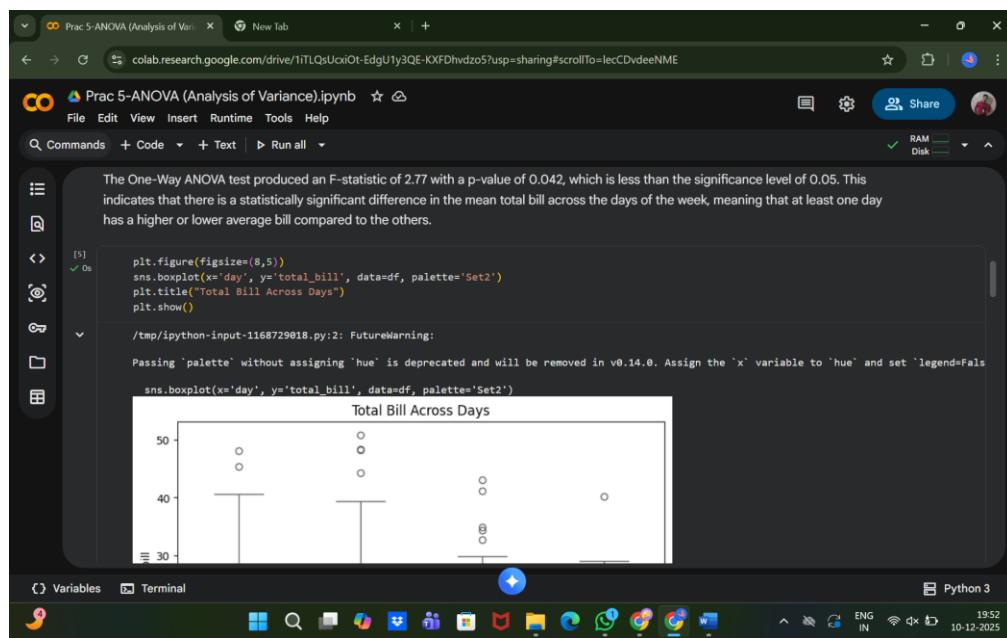
if p_value < 0.05:
    print("Reject null hypothesis → Total bill differs significantly across days.")
else:
    print("Fail to reject null hypothesis → Total bill does not differ significantly across days.")

...
F-statistic: 2.767479443286335
p-value: 0.04245383328952047
```

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

## Data Science

### PRACTICAL NO. 5



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.

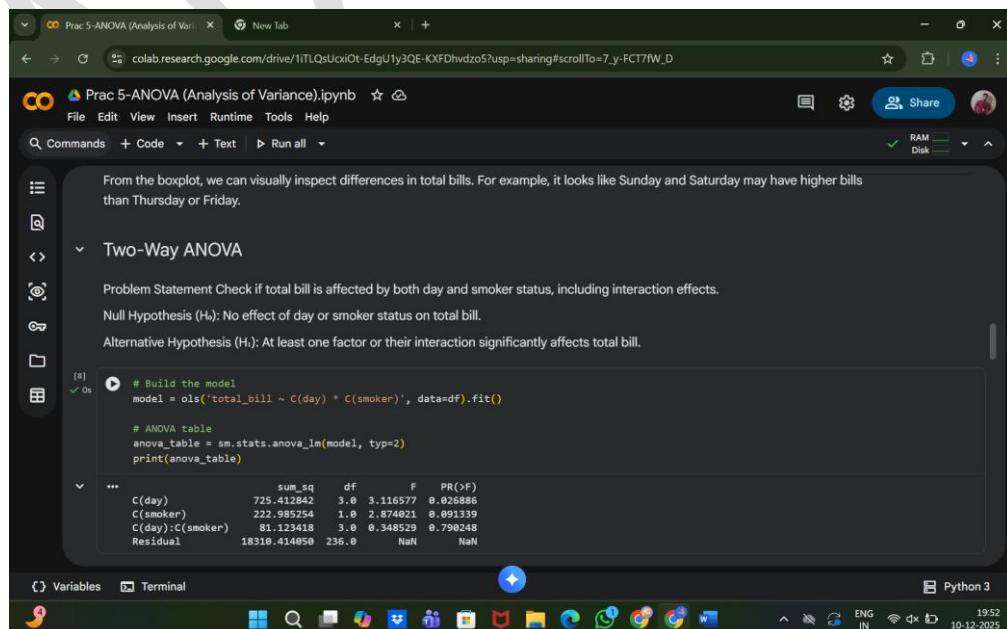
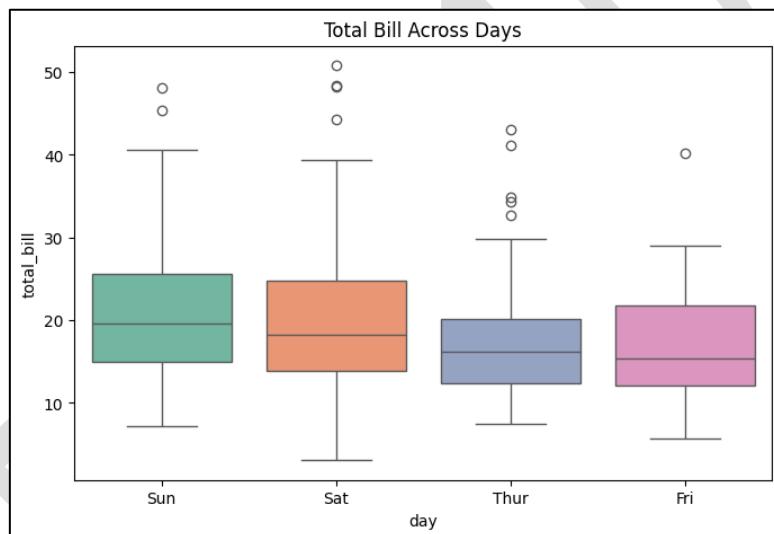
```
[5]: plt.figure(figsize=(8,5))
sns.boxplot(x='day', y='total_bill', data=df, palette='Set2')
plt.title("Total Bill Across Days")
plt.show()

/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=False'.
sns.boxplot(x='day', y='total_bill', data=df, palette='Set2')
```

Total Bill Across Days

Variables Terminal Python 3

10:12:2025



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 ( $H_0$ ): No effect of day or smoker status on total bill.

Alternative Hypothesis ( $H_1$ ): At least one factor or their interaction significantly affects total bill.

```
[8]: # 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	2.874621	0.091339
C(day):(smoker)	81.123418	3.0	0.348529	0.790248
Residual	18310.414850	236.0	NaN	NaN

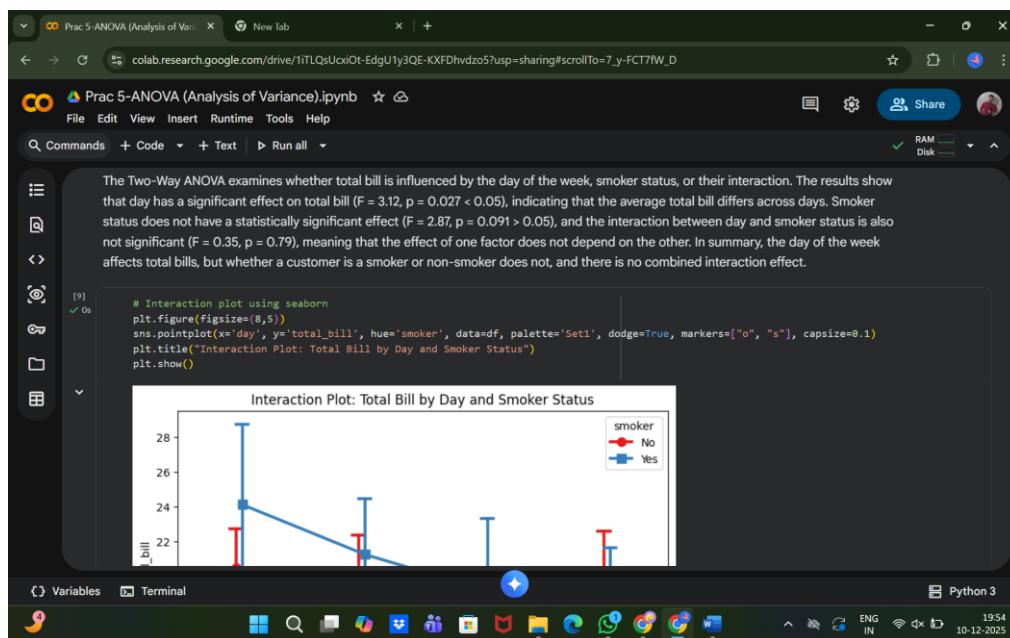
Variables Terminal Python 3

10:12:2025

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

## Data Science

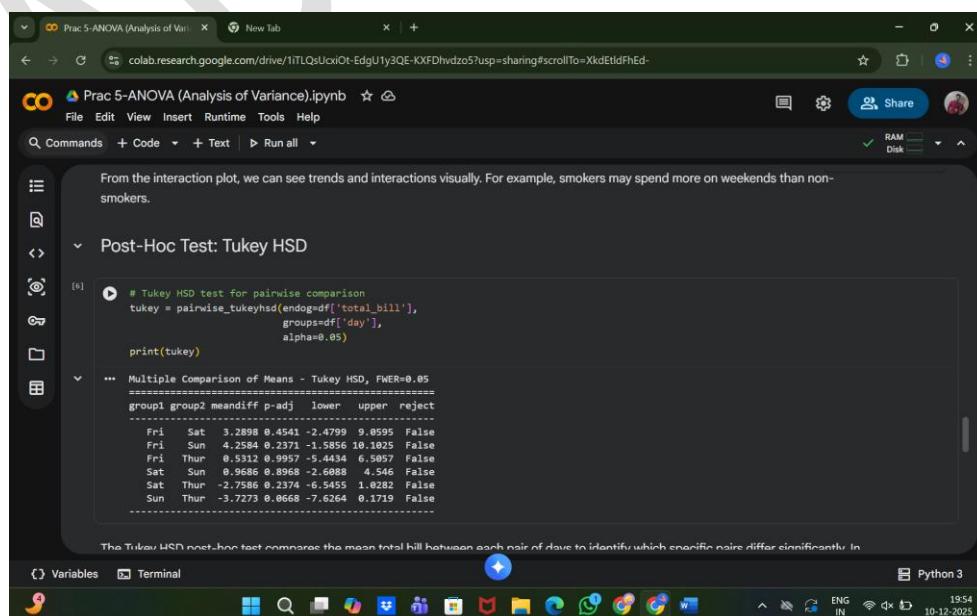
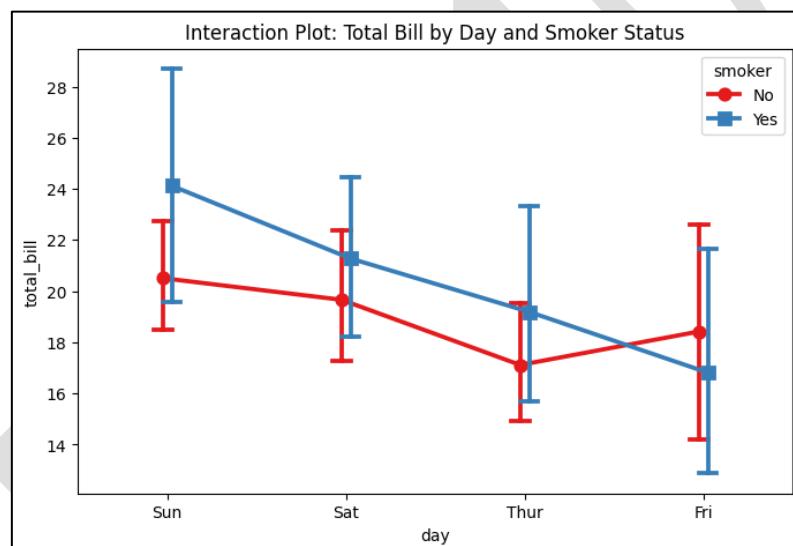
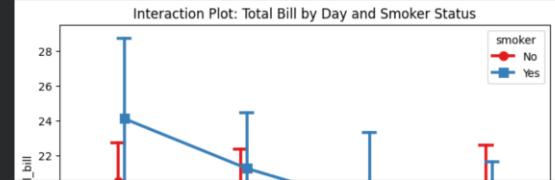
### PRACTICAL NO. 5



```
# 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()
```

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: Total Bill by Day and Smoker Status



```
# Tukey HSD test for pairwise comparison
tukey = pairwise_tukeyhsd(ending=df['total_bill'],
                           groups=df['day'],
                           alpha=0.05)
print(tukey)
```

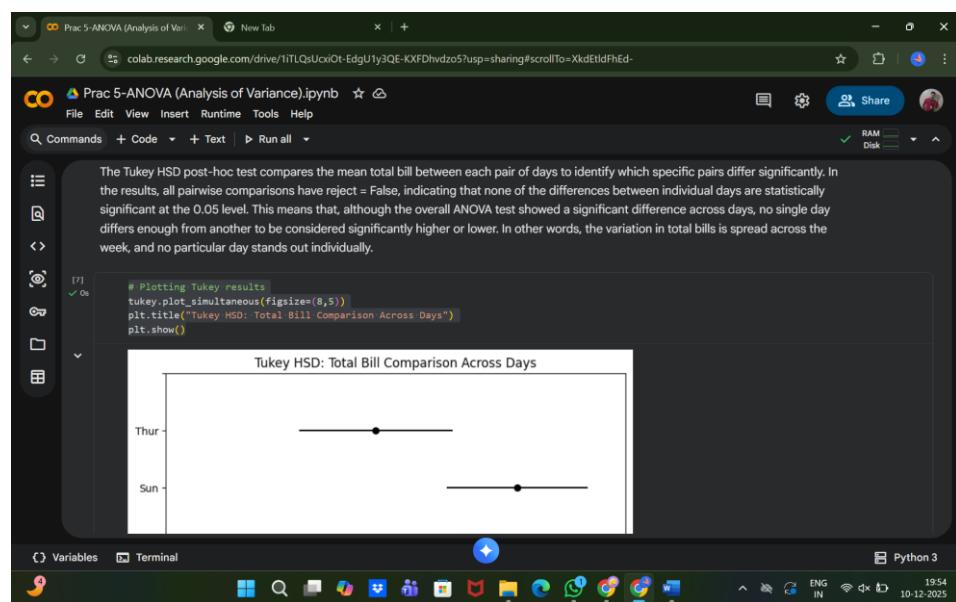
From the interaction plot, we can see trends and interactions visually. For example, smokers may spend more on weekends than non-smokers.

Post-Hoc Test: Tukey HSD

Day 1	Day 2	Mean Difference	p-value	reject
Fri	Sat	3.2898 0.4541 -2.4799 9.8595	False	
Fri	Sun	4.2584 0.2371 -1.5856 10.1825	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 0.8668 -7.6264 0.1719	False	

The Tukey HSD post-hoc test compares the mean total bill between each pair of days to identify which specific pairs differ significantly. In this case, none of the pairs are statistically significant at the 0.05 level.

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



The screenshot shows a Google Colab notebook titled "Prac 5-ANOVA (Analysis of Variance).ipynb". The code cell contains the following Python code:

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

The resulting plot is titled "Tukey HSD: Total Bill Comparison Across Days". The y-axis lists the days of the week: Thur, Sun, Sat, and Fri. The x-axis represents the total bill amount, ranging from 14 to 24. Horizontal error bars with black dots at their centers represent the mean total bill for each day. The error bars for Thur, Sun, and Sat overlap significantly, while the error bar for Fri is positioned lower on the x-axis.

