
Tipping Behavior in Restaurants: A Data-Driven Exploration Using the Tips Dataset

Rinzin Norsang Sherpa

M.S. Candidate in Data Science and Analytics

Thomas Edison State University (TESU)

Master's in Data Science and Analytics

Course: *DSI-5050 Programming 1: Python*

Date: *October 27, 2025*

Introduction

This project explores tipping behavior in restaurants using the tips.csv dataset. The goal is to understand how various factors—such as gender, smoking status, meal time, day of the week, and party size influence tip amounts. The analysis is conducted in two phases: one using the full dataset (including outliers), and another using a cleaned version with outliers removed. By comparing both versions, we aim to identify patterns in tipping behavior and assess how extreme values affect statistical results and visual interpretations. A regression model is also used to determine which variables significantly predict tip amounts.

Data Filtering and Preparation

The original dataset contained 244 rows and 7 columns, including numerical and categorical variables. A preliminary data check revealed one duplicate row, which was removed to ensure data integrity, leaving 243 unique observations for analysis. No missing values were found, and all columns were correctly typed for numerical or categorical analysis.

	total_bill	tip	sex	smoker	day	time	size
202	13.0	2.0	Female	Yes	Thur	Lunch	2

To better understand tipping behavior across different customer profiles, I applied filters based on gender, smoking status, time of day, and group size. These filters helped me focus the analysis and compare tip rates more accurately. I also created a new column called total_cost, which adds the tip to the total bill. This allowed me to compare the full cost of a meal, not just the bill alone. By calculating tip_pct and bill_pct, I was able to see how much of the total cost came from the tip and how much from the bill, giving a more complete picture of customer spending.

[319]:

	total_bill	tip	sex	smoker	day	time	size	tip_pct	tip_pct_num	total_cost	bill_pct
0	16.99	1.01	Female	No	Sun	Dinner	2	5.94%	5.94	18.00	94.388889
1	10.34	1.66	Male	No	Sun	Dinner	3	16.05%	16.05	12.00	86.166667
2	21.01	3.50	Male	No	Sun	Dinner	3	16.66%	16.66	24.51	85.720114
3	23.68	3.31	Male	No	Sun	Dinner	2	13.98%	13.98	26.99	87.736199
4	24.59	3.61	Female	No	Sun	Dinner	4	14.68%	14.68	28.20	87.198582

To prepare for targeted comparisons, the dataset was filtered into relevant subgroups:

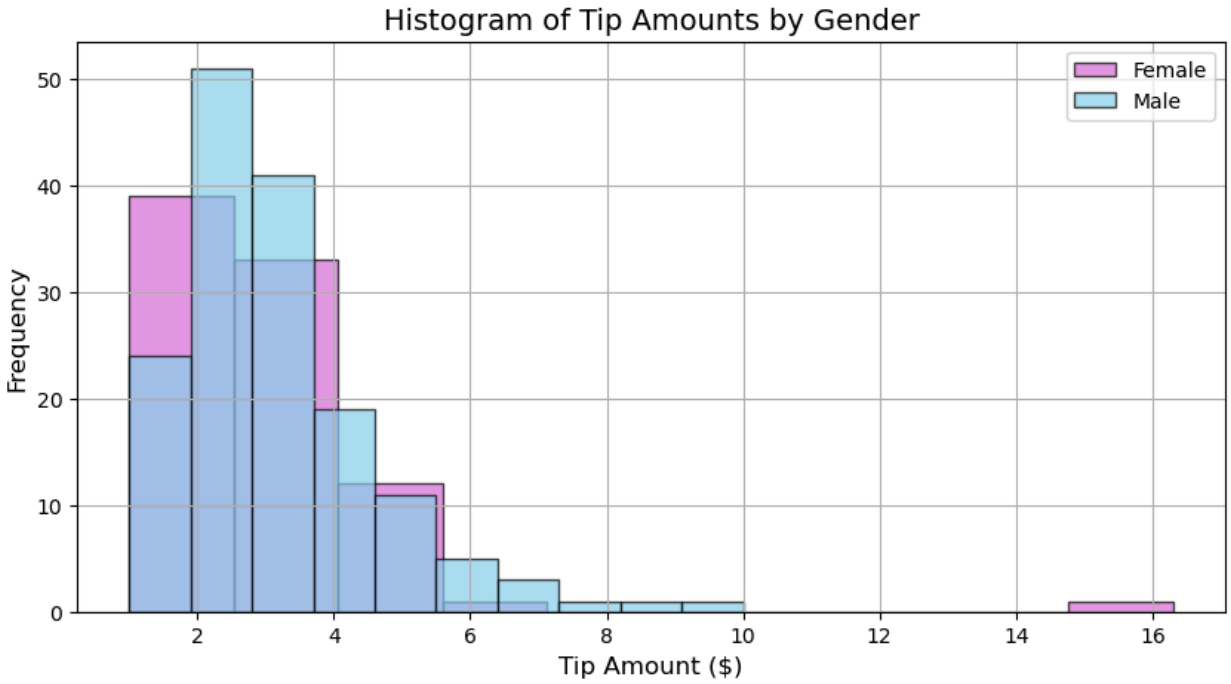
- **Gender:** Male vs Female
- **Smoking Status:** Smoker vs Non-Smoker
- **Meal Time:** Lunch vs Dinner
- **Group Size:** Small (≤ 2) vs Large (> 2)

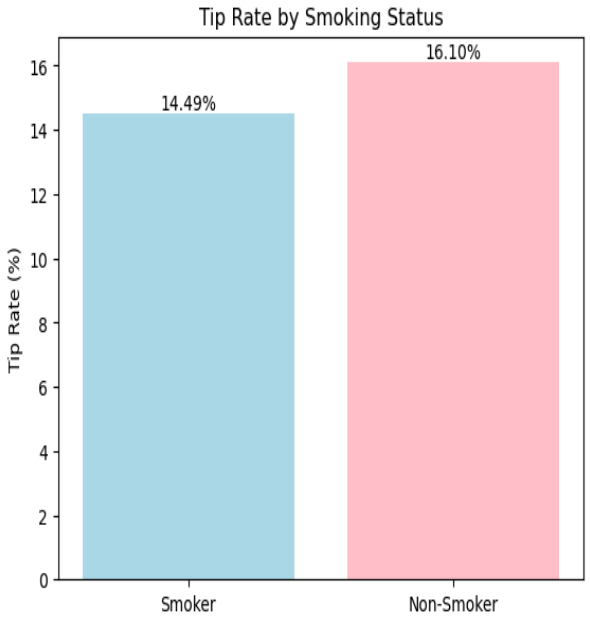
Gender : Male VS Female Tip Percentage

To examine tipping behavior across genders, the dataset was filtered into two groups: male and female customers. The tip rate was calculated as the percentage of total tips relative to total bill amounts for each group.

- Male Tip Rate: 14.89%
- Female Tip Rate: 16.64%

This indicates that, on average, female customers tipped more generously than male customers relative to their bill size. The difference suggests a potential behavioral trend in tipping habits, with women contributing a higher percentage of their bill as a tip.



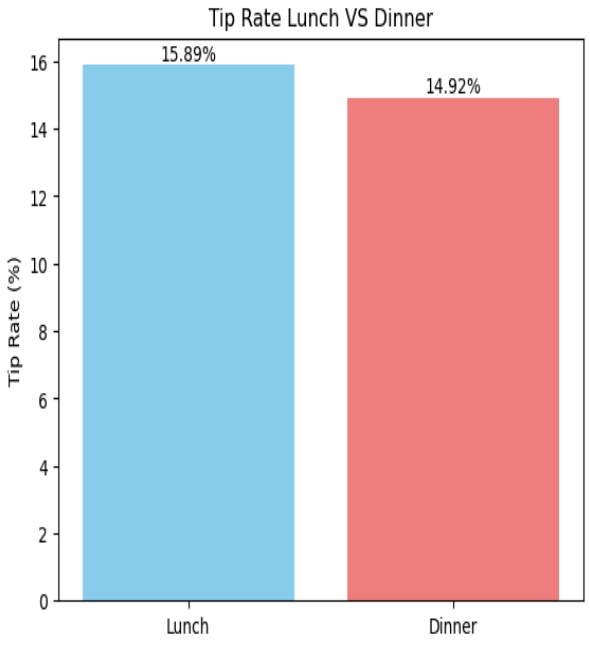


Smoking Status: Smoker vs Non-Smoker

The dataset was analyzed to compare tipping behavior between smokers and non-smokers. Tip rate was calculated as the percentage of total tips relative to total bill amounts for each group.

- **Smoker Tip Rate: 14.49%**
- **Non-Smoker Tip Rate: 16.10%**

This indicates that non-smokers, on average, tipped a higher percentage of their bill compared to smokers. The difference suggests a modest but consistent trend in tipping behavior based on smoking status.



Meal Time: Lunch vs Dinner

Tip behavior was compared across meal times using the raw dataset:

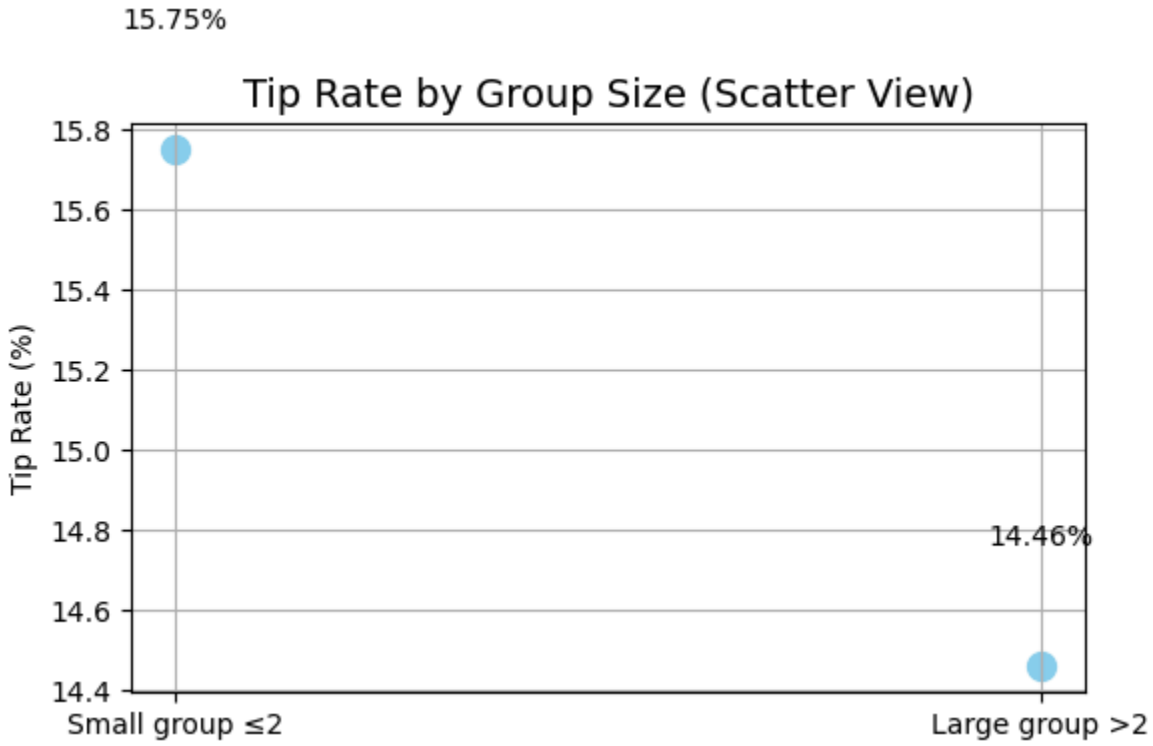
- **Lunch Tip Rate: 15.89%**
- **Dinner Tip Rate: 14.92%**

Group Size

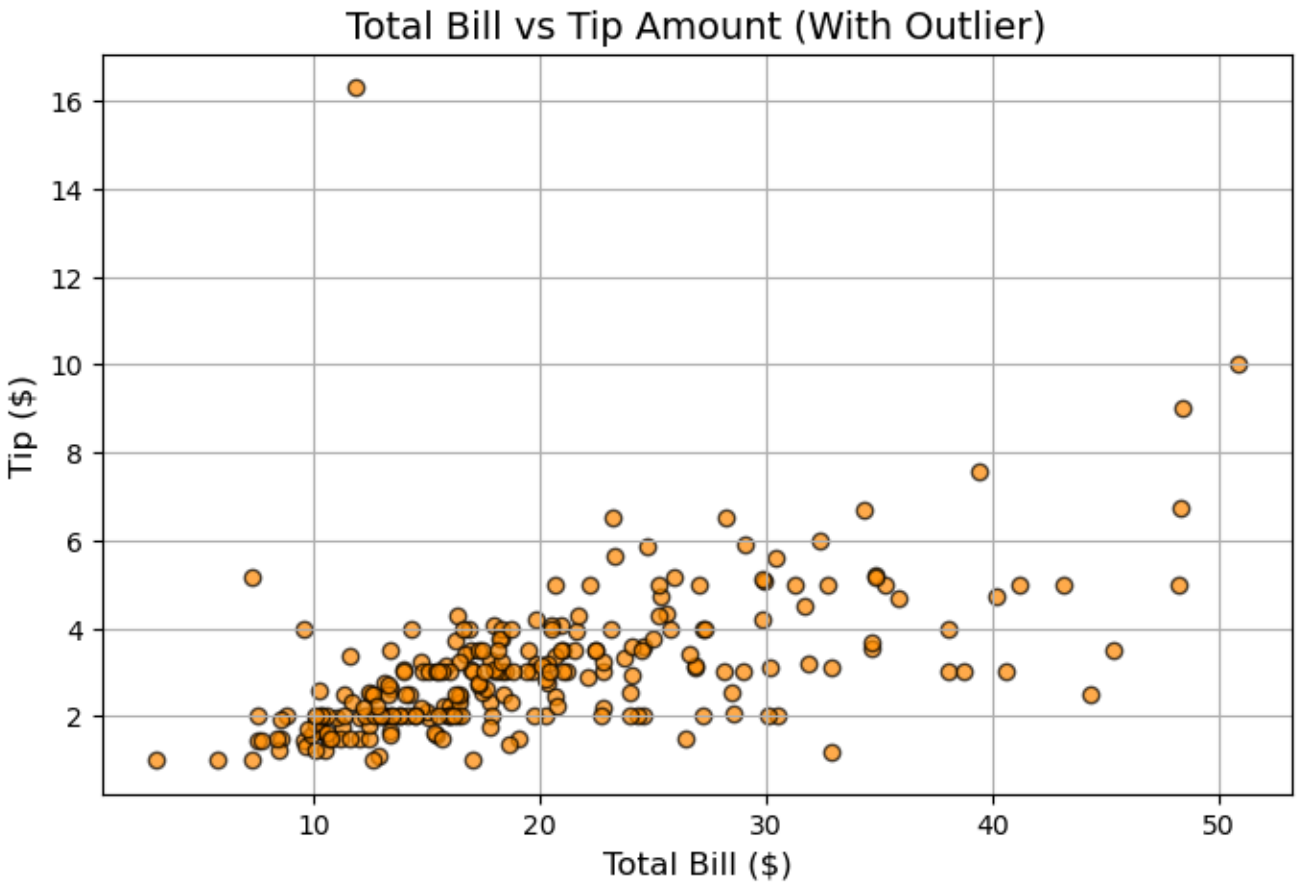
This analysis explores how restaurant party size affects tipping behavior using the tips.csv dataset. The data was filtered into two categories: **small groups** (party size ≤ 2) and **large groups** (party size > 2). Tip percentages were calculated for each group to assess generosity relative to bill size.

- **Small group tip rate:** 15.75%
- **Large group tip rate:** 14.46%

While large groups contributed more in total tip dollars due to higher bills, their tip percentage was slightly lower. This suggests that smaller parties tend to tip more generously in proportion to their spending.

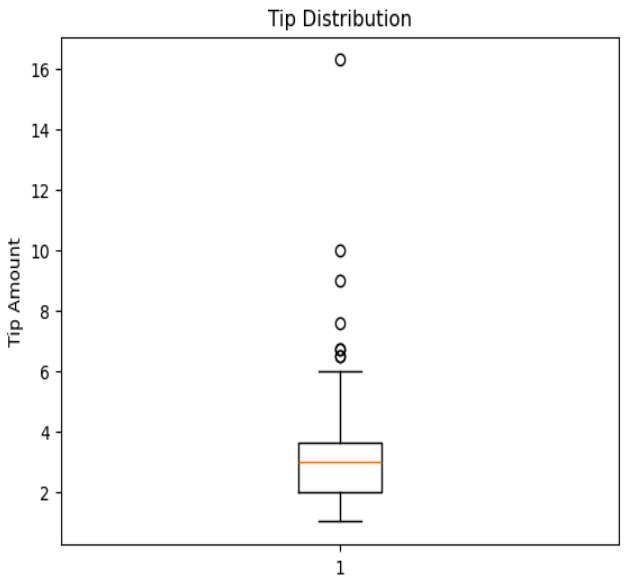


This chart shows how tips change based on the total bill. Each orange dot is a real example from the data. Most of the dots are close together, meaning people usually give small tips for small bills. But there's one dot far away from the others, it shows a very big tip for a high bill. That dot is an outlier, which means it's not normal and can affect the results. Including it helps show the full picture, but it also makes the graph look stretched.



Outliers

An outlier was spotted in the dataset, a tip amount of $16.30 linked to a total bill of only $11.87. This tip was unusually high and didn't match the typical tipping behavior seen in the rest of the data. I removed this outlier because it was affecting the overall analysis by inflating the average tip rate and stretching the visualizations, especially the boxplot and scatter plots. Keeping it in would make it look like people tip more than they actually do, and it could hide the real patterns in how most customers tip. Removing it helped me get a clearer and more accurate view of the data.



Outlier values in 'tip'

23	7.58
59	6.73
141	6.70
147	**16.30**
170	10.00
183	6.50
212	9.00
214	6.50

Regression analysis with raw data df

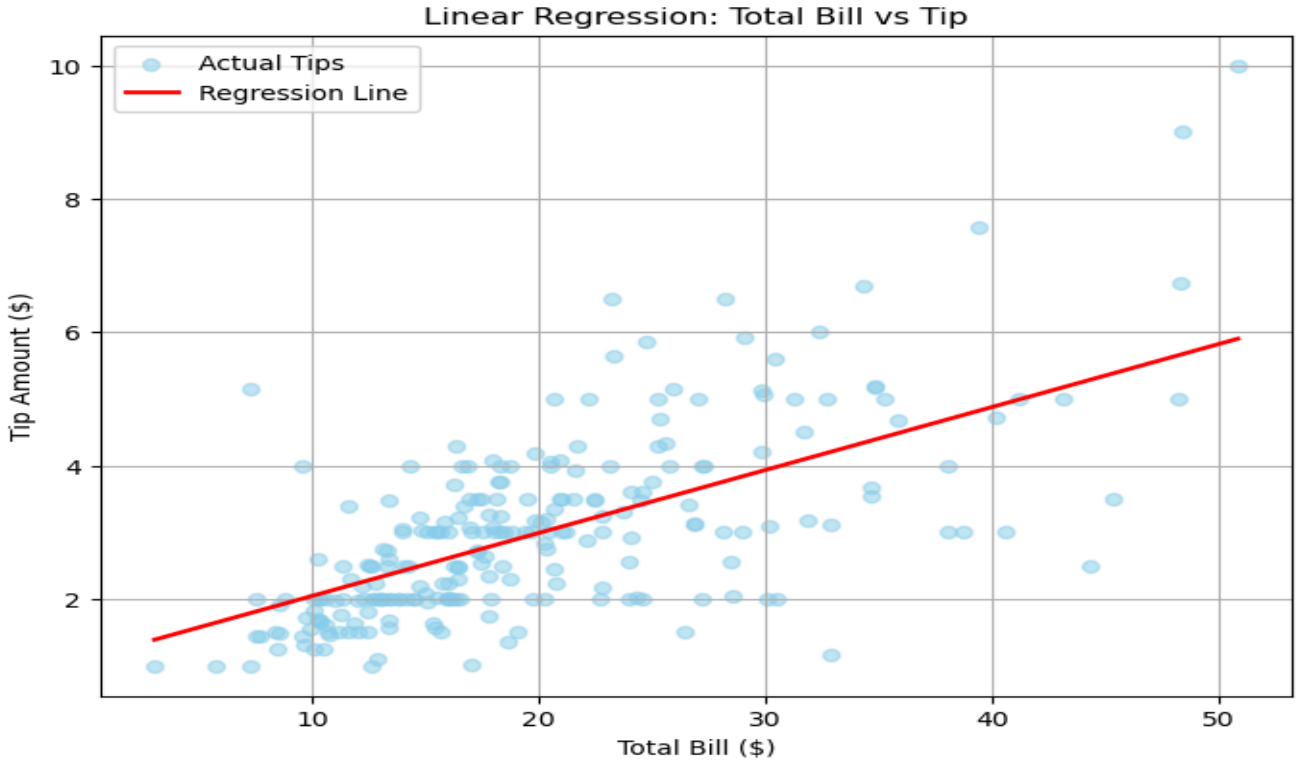
A linear regression model was used to understand which factors influence tip amount. The model included variables such as total bill, party size, meal time, day of the week, gender, and smoking status.
The regression results indicate that:

The results show that total_bill is the only statistically significant predictor of tip amount, with a p-value less than 0.001. This means that as the bill increases, the tip also increases—about $0.092 extra tip for every $1 increase in the bill.

Other variables like meal time (Lunch), day (Saturday, Sunday, Thursday), gender (Male), smoking status (Yes), and party size did not show significant effects (all p-values > 0.05). Their influence on tipping is weak or uncertain in this dataset.

The intercept of the model is approximately $0.99, which represents the baseline tip when all other variables are at their reference levels.

In summary, bill size is the main driver of tip amount, while other factors like gender, smoking, and time of day do not have a strong impact in this raw dataset



The line in the above image confirms what the regression model showed: total bill is the strongest predictor of tip amount. Other variables like time of day, gender, or smoking status didn't have statistically significant effects (based on their p-values), but total bill had a strong, consistent influence

Establishing df_clean: A Refined Foundation for Tipping Analysis

After identifying and removing the extreme outlier, I created a filtered dataset called df_clean to ensure more accurate and reliable analysis. This cleaned version excludes the unusually high tip that had been distorting averages and visualizations. With df_clean, I was able to re-run summary statistics, generate clearer plots, and build regression models that better reflect typical tipping behavior. This dataset serves as the foundation for all subsequent analysis, allowing me to focus on meaningful patterns without the influence of outlier noise

```
:  df_clean = df[
       (df['tip'] >= lower_tip) & (df['tip'] <= upper_tip) &
       (df['tip_pct_num'] >= lower_pct) & (df['tip_pct_num'] <= upper_pct) &
       (df['total_bill'] >= lower_bill) & (df['total_bill'] <= upper_bill) &
       (df['bill_pct'] >= lower_bill_pct) & (df['bill_pct'] <= upper_bill_pct)
   ]
```

Data Filtering and Preparation

The cleaned dataset, df_clean contains 219 rows after applying gender-specific outlier filtering across key tipping and billing metrics. During this process, I identified 1 duplicated row, which will be addressed to maintain data integrity. Additionally, a check for missing values confirmed that no rows contain null values, ensuring that the dataset is complete and ready for reliable analysis. With outliers removed, duplicates flagged, and no missing data, provides a stable foundation for exploring tipping behavior with confidence

```
[188]: df_clean[df_clean.duplicated()]
```

[188]:

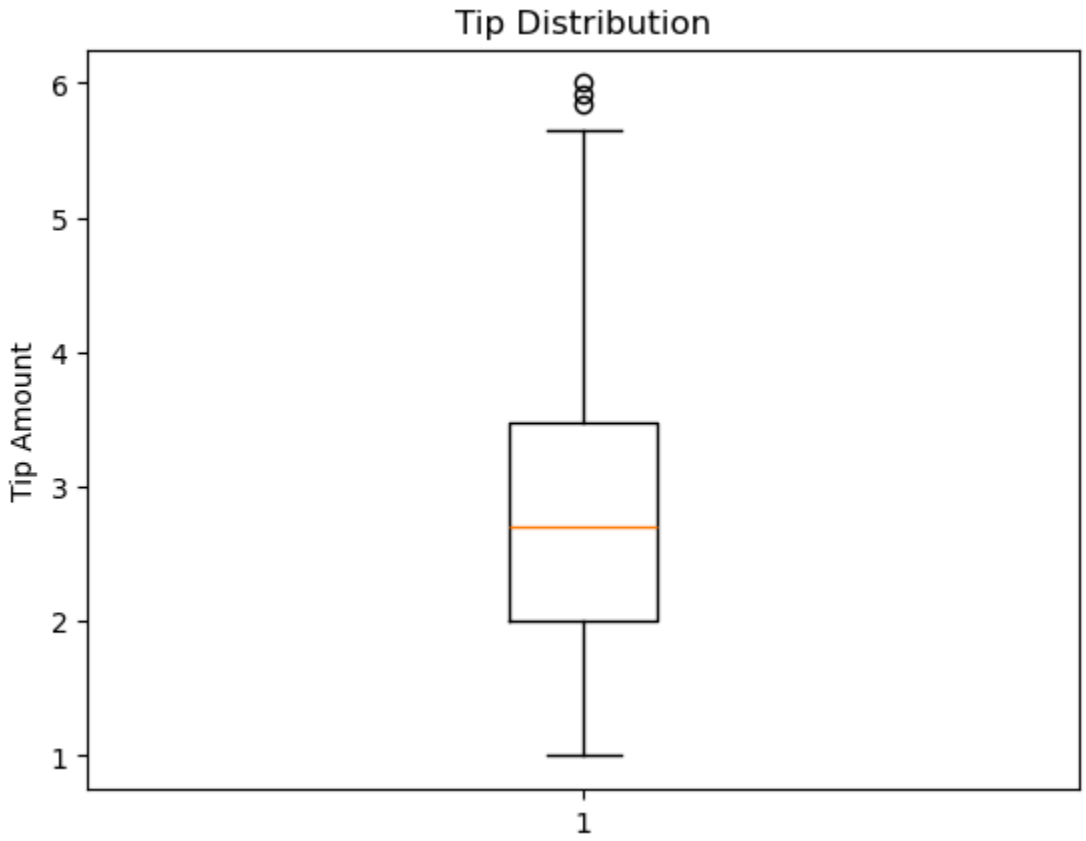
	total_bill	tip	sex	smoker	day	time	size	tip_pct	tip_pct_num	total_cost	bill_pct
202	13.0	2.0	Female	Yes	Thur	Lunch	2	15.38%	15.38	15.0	86.666667

```
print(f"Original rows: {len(df)}")
print(f"Cleaned rows: {len(df_clean)}")
print(f"Rows removed: {len(df) - len(df_clean)}")
```

```
Original rows: 244
Cleaned rows: 224
Rows removed: 20
```

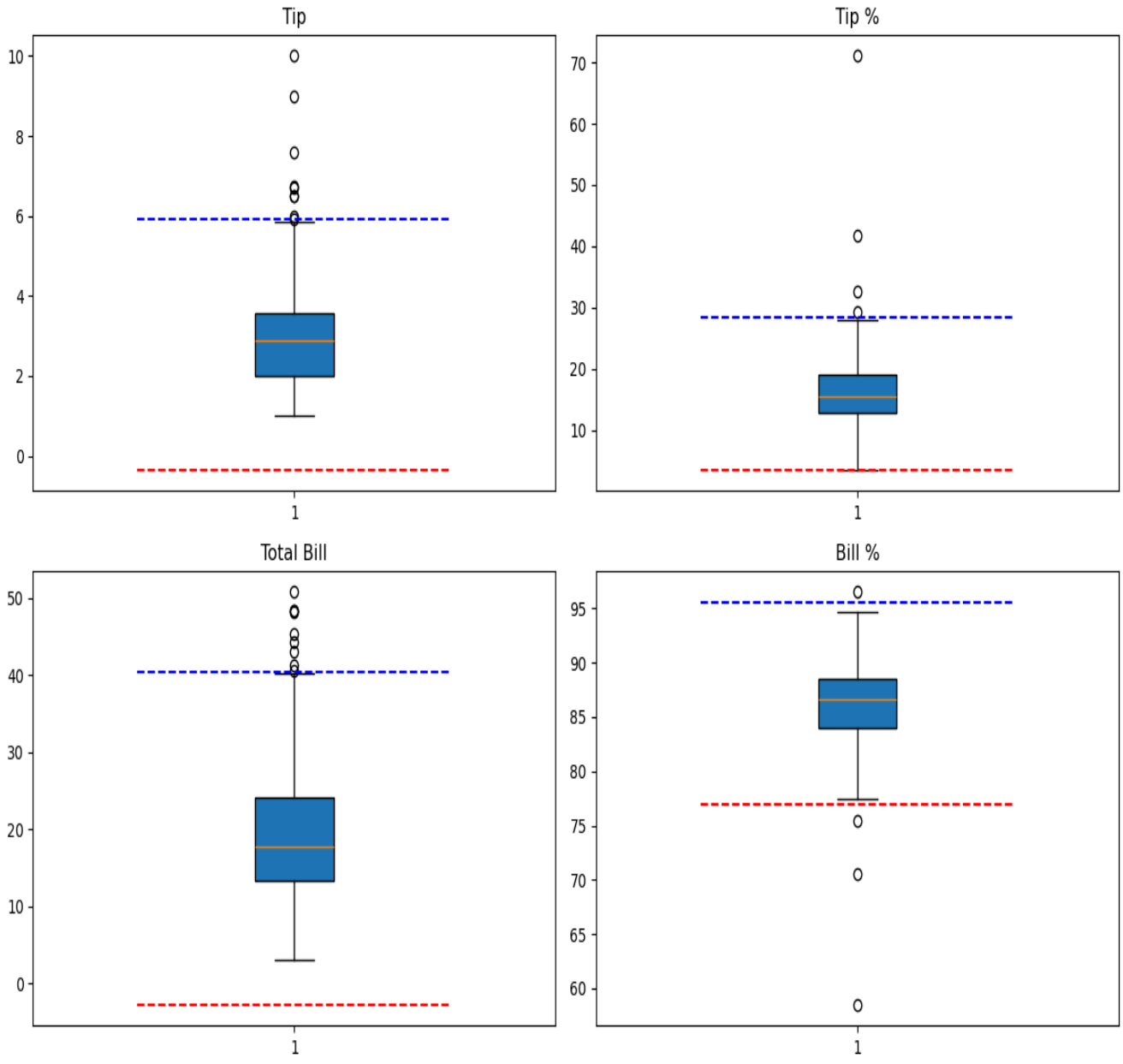
How One Outlier Skewed the Story and What Changed After Its Removal

After removing the extreme outlier (tip = $16.30), the boxplots for tip amount, tip percentage, total bill, and bill percentage became much clearer and more balanced. The distributions now show tighter interquartile ranges and more accurate medians, without being stretched by unrealistic values. This cleanup helped highlight the true patterns in the data, for example, typical tip percentages now fall within a more reasonable range, and the total bill distribution reflects actual spending behavior. By removing the outlier, I was able to focus on the majority of transactions and get a more reliable view of how customers tip across different situations



<u>From Distortion to Clarity: Boxplot Insights</u>

The boxplots provide a clear summary of the distributions for tip amount, tip percentage, total bill, and bill percentage after cleaning the dataset. Each plot shows the median, interquartile range, and any remaining outliers. The tip and tip percentage plots reveal that most tips fall within a modest range, with a few higher values still present but no longer distorting the scale. The total bill distribution is more balanced, and the bill percentage plot shows that the majority of the meal cost comes from the bill itself, not the tip. These visualizations confirm that the data is now more stable and representative, allowing for more accurate analysis of customer spending and tipping behavior.

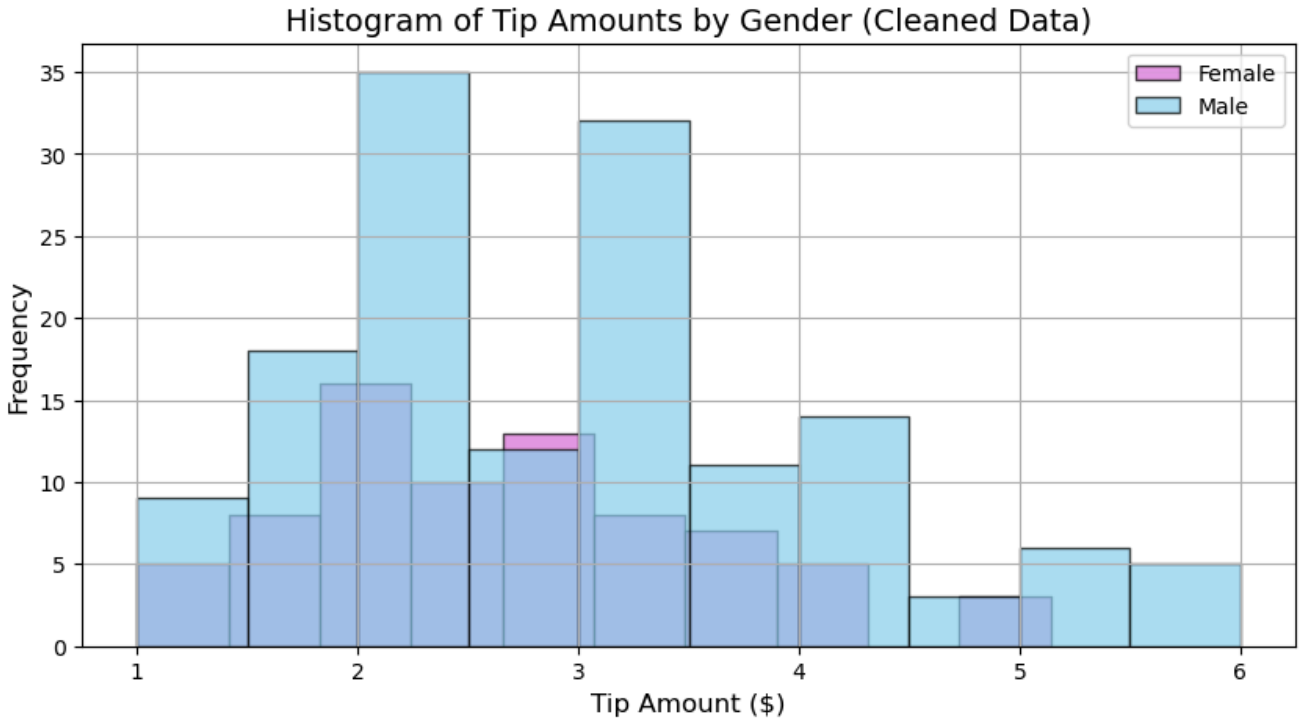


- **Gender:** Male vs Female

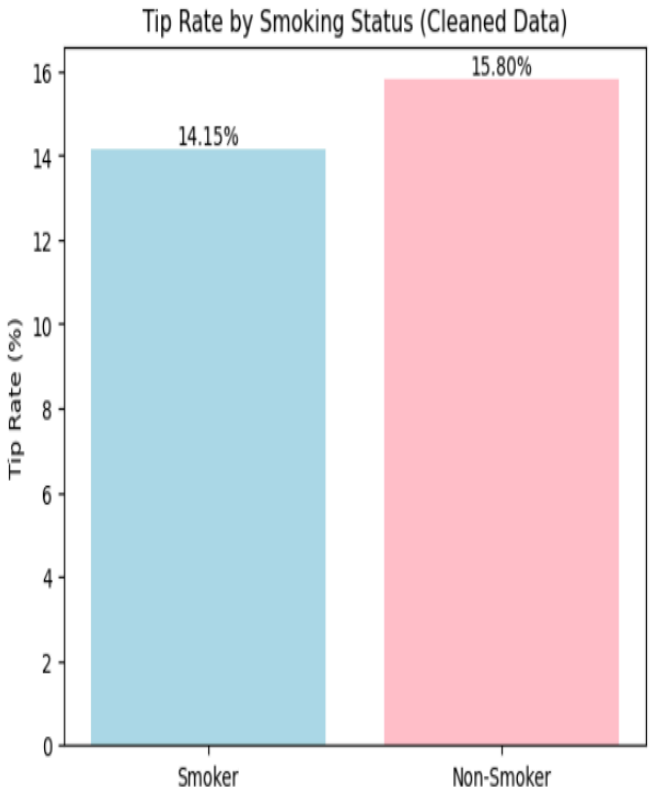
- **Smoking Status:** Smoker vs Non-Smoker
- **Meal Time:** Lunch vs Dinner
- **Group Size:** Small (≤ 2) vs Large (> 2)

Gender Male VS Female

This chart shows how tip amounts are spread out for men and women. Most tips are around $2, and men tend to give tips across a wider range of amounts. Women's tips are more often in the lower range. The chart helps us see how tipping habits can be different between genders.



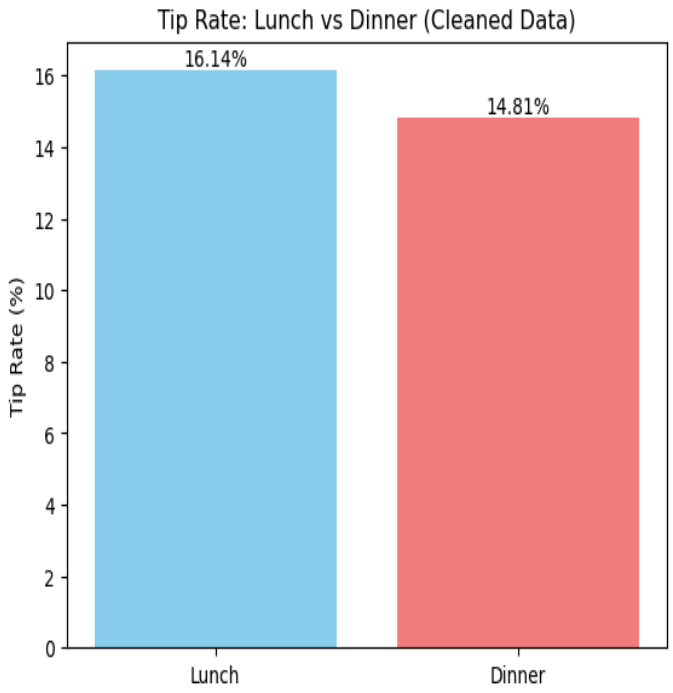
- **Smoking Status:** Smoker vs Non-Smoker
- **Meal Time:** Lunch vs Dinner



This chart shows how much smokers and non-smokers tip, based on cleaned data.

Non-smokers tip about 15.80% of their bill, while smokers tip around 14.15%.

That means non-smokers tend to give slightly higher tips than smokers.



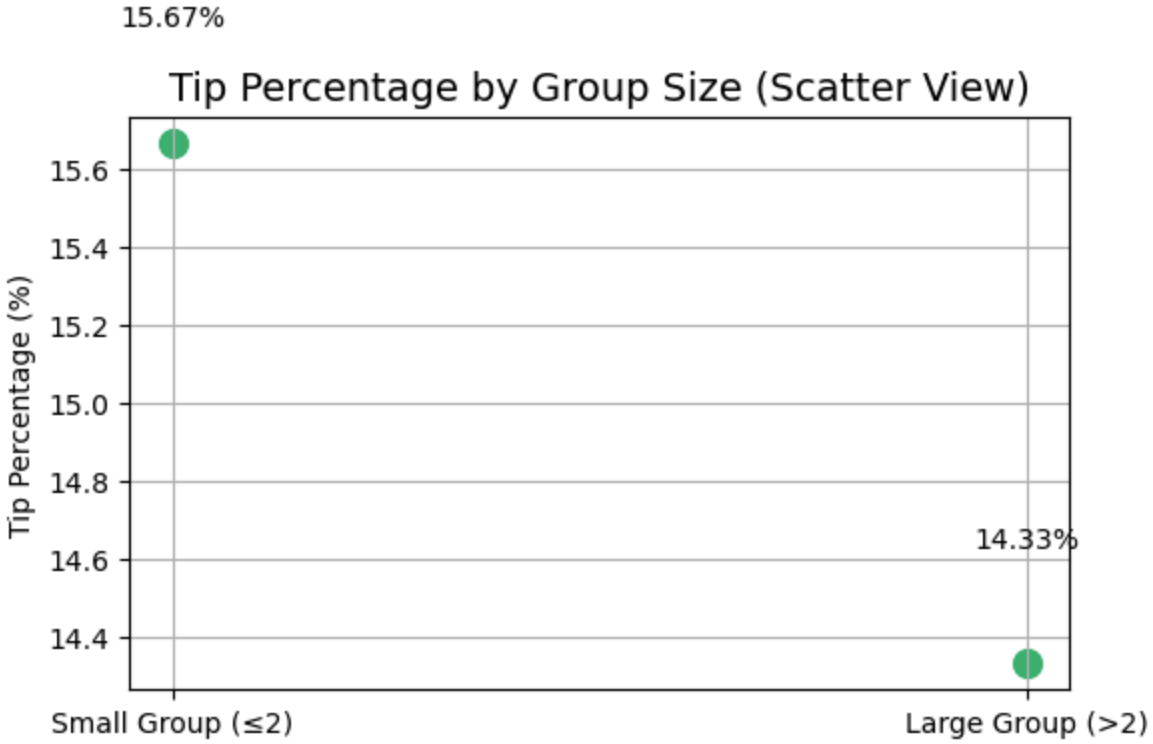
This chart compares how much people tip during lunch and dinner.

It shows that lunch tips are higher, with an average of 16.14%, while dinner tips are around 14.81%.

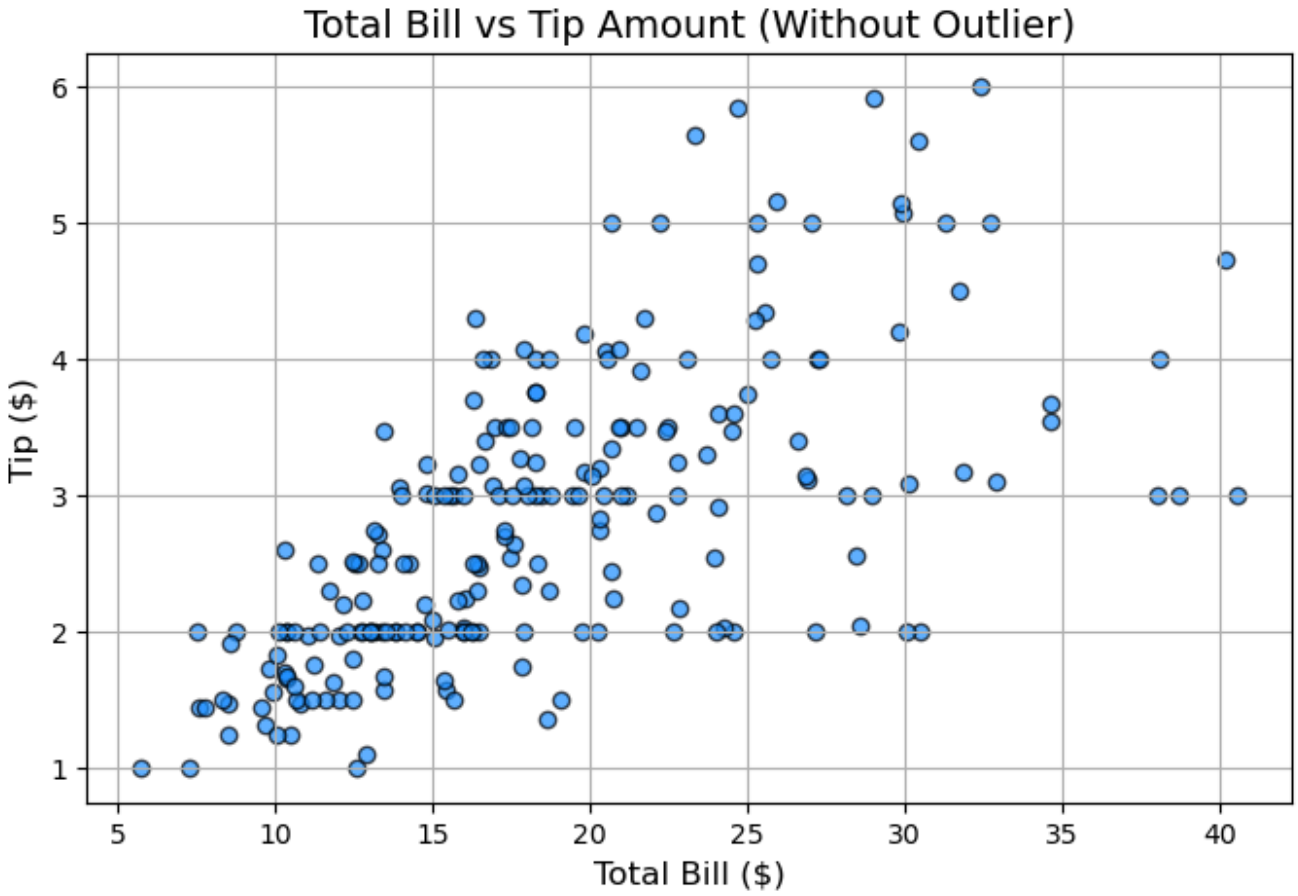
That means people tend to be a little more generous when tipping at lunch than at dinner.

Group Size: Small (≤ 2) vs Large (> 2)

Small groups tip a higher percentage than large groups. On average, groups of 2 or fewer people tip about 15.67% of their bill, while larger groups tip around 14.33%. This means smaller parties tend to be a little more generous with tips compared to bigger ones.


15.67%
Tip Percentage by Group Size (Scatter View)
15.6
15.4
15.2
15.0
14.8
14.6
14.4
Tip Percentage (%)
14.33%
Small Group (≤2)
Large Group (>2)

This graph shows how tip amounts change as the total bill increases, using cleaned data without any extreme values. Each blue dot is a real example from the dataset. The dots go upward overall, which means that when people spend more money on their meal, they usually give a bigger tip. By removing the outlier, the chart gives a clearer and more accurate picture of normal tipping behavior.



Regression analysis is a method used to understand how different factors influence a specific outcome in this case, how various features like total bill, party size, time of day, and customer traits affect tip amounts. It helps us see which variables have the strongest impact and whether those effects are positive or negative.

The regression results indicate that:

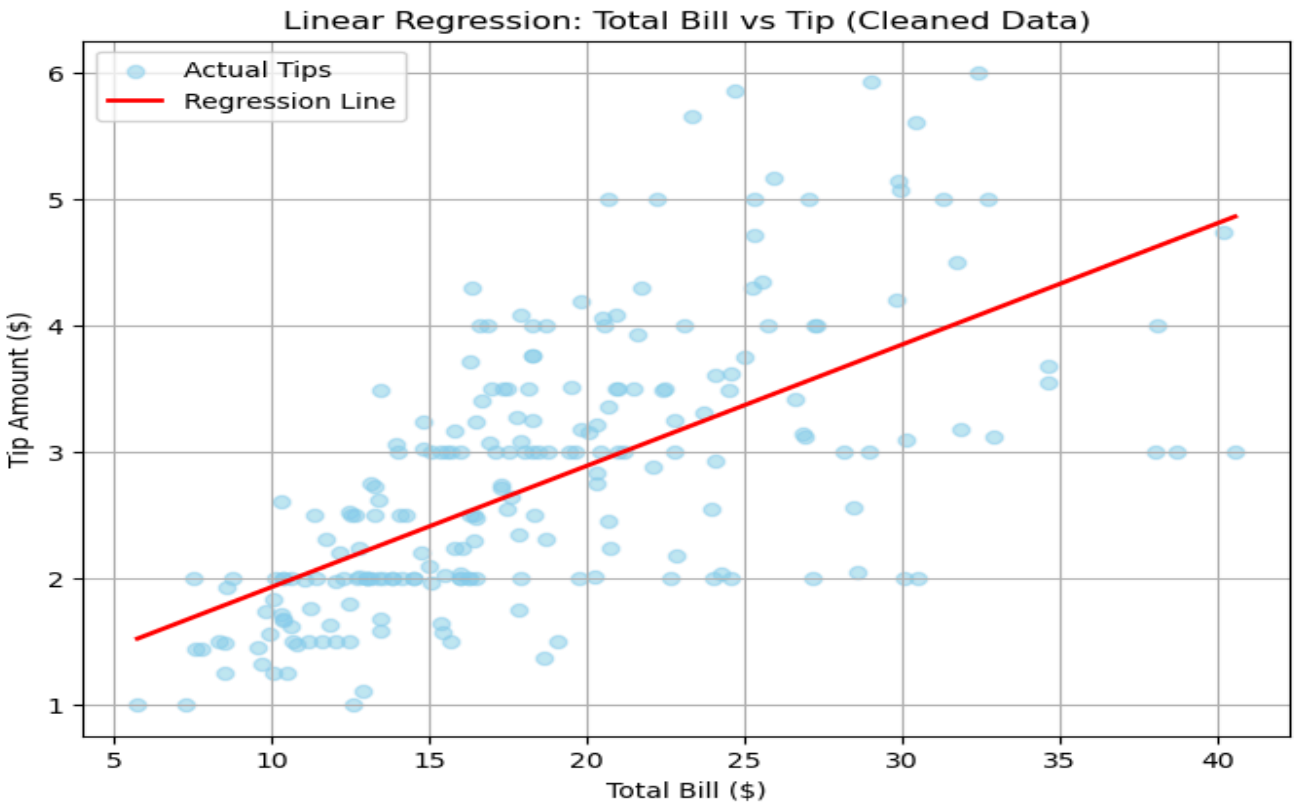
total_bill is the only statistically significant predictor of tip amount ($p < 0.001$). The coefficient suggests that for each additional dollar spent, the tip increases by approximately $0.095.

All other variables time, day,sex,smoker and size did not show statistically significant effects on tip amount ($p > 0.05$).

The intercept value of approximately $0.78 represents the baseline tip when all categorical variables are at their reference levels.

This analysis demonstrates that tip amounts are primarily influenced by the size of the bill, while other contextual and demographic factors have limited predictive value in this dataset.

In the graph, I focused on the relationship between total bill and tip amount, while keeping other factors constant. Each blue dot shows a real tip from the data, and the red line shows the model's prediction. The upward slope of the line tells us that as the total bill increases, the tip amount also tends to go up. This confirms that total bill is the most powerful predictor of tipping behavior in the model.


Linear Regression: Total Bill vs Tip (Cleaned Data)
Actual Tips
Regression Line
Tip Amount ($)
Total Bill ($)
1
2
3
4
5
6
5
10
15
20
25
30
35
40

Conclusion

In this project, I studied how people tip in restaurants and how the results change when I removed unusual data. One tip: $16.30 on an $11.87 bill was much higher than normal and made the charts and averages look strange. After I removed that outlier, the data became easier to understand and showed more realistic tipping habits.

I also used regression analysis to find out what affects tip amounts. The results showed that the **total bill** is the main reason people **tip more**. **When the bill goes up, the tip usually goes up too**. Other things like gender, smoking, meal time, or group size didn't have a strong effect in the model. **So, the biggest reason for a larger tip is simply spending more money.**

While men may give larger tips in absolute dollars occasionally, women tend to tip a **higher proportion** of their bill. This pattern holds even after removing outliers, suggesting it's a stable behavioral difference in the dataset.