Group Assignment 1
September 10, 2025

MS 491 - Special topics in Management: Marketing Analytics
Prof. Marcos Inacio Severo de Almeida

Indian Institute of Technology Gandhinagar
Palaj, Gandhinagar - 382355

Marketing Analytics for Data-Rich Environments

Submitted by

Dhruvi Sisodiya (22110075)
Chemical Engineering
Arul Singh (22110036)
Materials Engineering
Darshan Zala (22110297)
Electrical Engineering

Contents

Chapter 1

Introduction

Understanding customer behavior is a cornerstone of marketing analytics, providing businesses with the insights needed to optimize service and strategy. In the restaurant industry, tipping is a critical component of employee compensation and a direct reflection of customer satisfaction and behavior. This report presents a comprehensive statistical analysis of tipping behavior at a restaurant, utilizing a dataset containing variables such as total bill, tip amount, payment method, group size, day of the week, and server. The objective is to identify the key determinants of tipping patterns and provide data-driven insights.

The primary variable of interest is the tip percentage (`PctTip`), as it normalizes for the bill amount and provides a clearer measure of a customer's tipping propensity. This study explores whether the `PctTip` is significantly affected by several operational variables: the attending **server**, the **day** of the week, the number of **guests** in a party, and the **payment method** (cash or credit). The analysis employs a dual approach, using both Gretl and Python to perform descriptive analysis, generate robust visualizations, and conduct inferential statistics, specifically the Analysis of Variance (ANOVA), to test for significant differences in tipping behavior.

Bill	Tip	PctTip	Credit	Guests	Day	Server
23.7	10	42.2	n	2	f	A
36.11	7	19.4	n	3	f	B
31.99	5.01	15.7	y	2	f	A
17.39	3.61	20.8	y	2	f	B
15.41	3	19.5	n	2	f	B
18.62	2.5	13.4	n	2	f	A
21.56	3.44	16	n	2	f	B
19.58	2.42	12.4	n	2	f	A
23.59	3	12.7	n	2	f	A
18.67	2	10.7	n	2	f	B
15.19	1.75	11.5	n	1	f	B
10.17	2.1	20.6	n	1	f	A
8.08	1	12.4	n	1	f	B
8.83	1.25	14.2	n	2	f	B

Figure 1.1: Restaurant Tips Dataset as used

Chapter 2

Gretl Analysis

The dataset was analyzed using Gretl to generate descriptive and inferential statistics. The findings from histograms, boxplots, scatter plots, and ANOVA are summarized below.

2.0.1 Descriptive Statistics & Distributions

An initial exploratory analysis was conducted to understand the distribution of the primary variables: Bill, Tip, and PctTip.

2.0.1.1 Frequency Distributions (Histograms)

The histograms reveal that none of the key financial variables are normally distributed.

Bill: The distribution of bills has a peak around \$20 but is slightly skewed to the right due to some extremely high values. The Chi-square test rejects normality (p=0.0000).

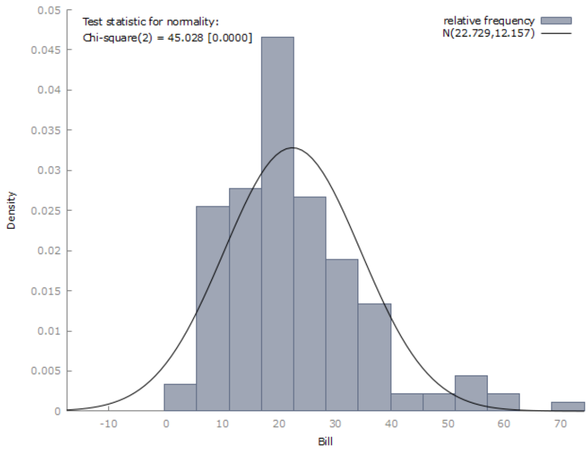

Figure 2.1: Frequency Distribution (Histogram) for Bill.

Tip: The distribution of tip amounts is strongly right-skewed, with most tips clustered between \$2 and \$5. A few large tips create a long tail, which is why the normality test rejects a normal distribution (p=0.0000).

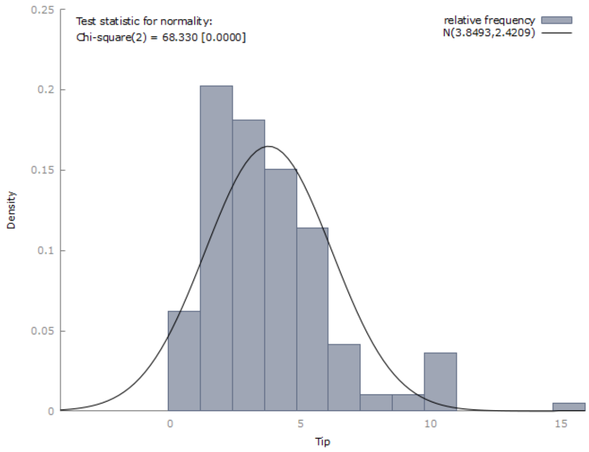

Figure 2.2: Frequency Distribution (Histogram) for Tip.

PctTip: The tip percentage distribution is roughly bell-shaped and centered around 16%. However, it has heavier tails than a normal distribution, with extreme values on both the low and high ends. This deviation led to the rejection of normality via the Chi-square test (p=0.0000)

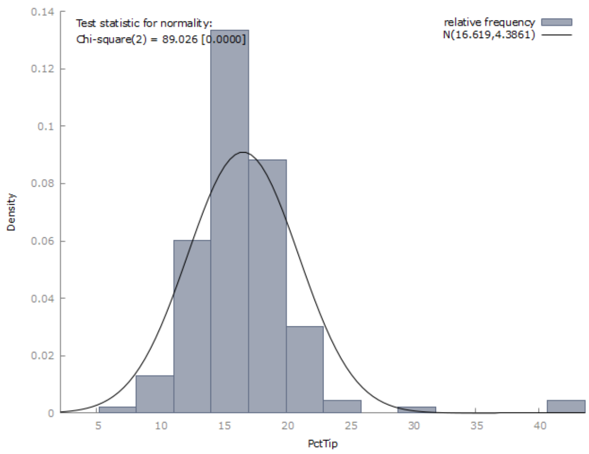

Figure 2.3: Frequency Distribution (Histogram) for PctTip.

Boxplots

The boxplots provide a clearer view of the median, spread, and outliers for each variable, confirming the skewness observed in the histograms.

Bill: The median bill is near \$20, with most bills falling between \$15 and \$30. The plot shows numerous outliers beyond \$50, with one extreme case near \$70, which explains the right-skewed shape of the distribution.

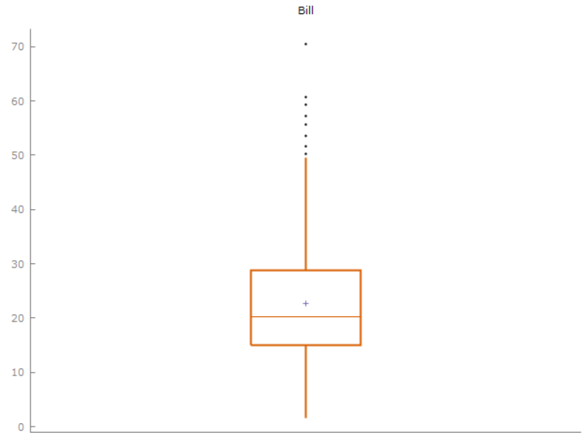

Figure 2.4: Boxplot for Bill.

Tip: The median tip is close to \$3, and the majority of tips are in the \$2-\$5 range. Several outliers appear in the \$10-\$15 range, visually confirming the right skew.

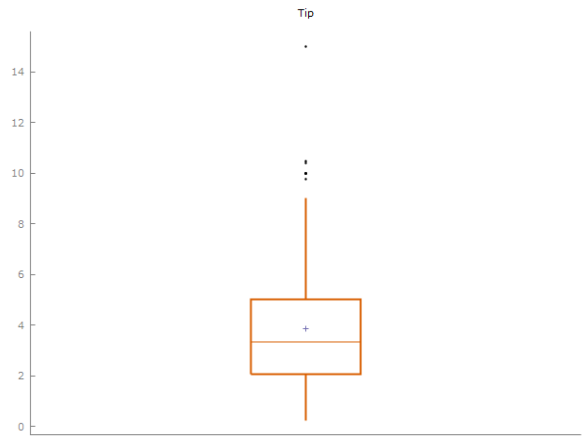

Figure 2.5: Boxplot for Tip.

PctTip: The boxplot for `PctTip` shows a stable central tendency, with a median around 16% and an interquartile range between 14% and 18%. The whiskers and several outliers, with some tips exceeding 30% and 40%, highlight the occasional instances of very high or low tipping that cause the distribution to deviate from normality.

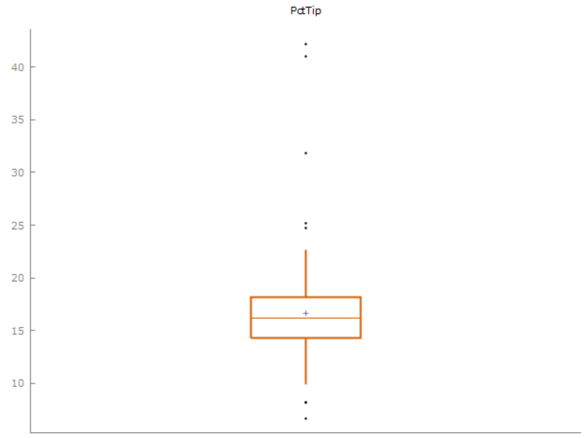

Figure 2.6: Boxplot for PctTip.

2.0.2 Inferential Statistics: ANOVA

An Analysis of Variance (ANOVA) was performed to test whether the mean tip percentage (`PctTip`) differs significantly across various groups. In all cases, the results indicated no statistically significant differences.

Server vs. PctTip: The analysis showed no significant difference in tip percentages based on the server. The F-statistic was F(2,154)=2.19 with a p-value of p=0.115. This indicates customers tip consistently regardless of their server.

```
Analysis of Variance, response = PctTip, treatment = Server:

                     Sum of squares       df      Mean square

  Treatment                 83.1368        2          41.5684
  Residual                  2917.93      154          18.9476
  Total                     3001.06      156          19.2376

  F(2, 154) = 41.5684 / 18.9476 = 2.19386 [p-value 0.1150]

  Level          n        mean      std. dev

  1             60     17.5433        5.5040
  2             65     16.0169        3.4852
  3             32     16.1094        3.3756

  Grand mean = 16.6191
```

Figure 2.7: ANOVA output for Server vs. PctTip.

Guests vs. PctTip: The number of guests at a table has no significant impact on the tipping percentage. Mean tip percentages remained stable around 16-17% regardless of group size. The ANOVA result was highly non-significant (F(6,150)=0.16, p=0.986).

```
Analysis of Variance, response = PctTip, treatment = Guests:

                     Sum of squares       df      Mean square

  Treatment                 19.4265        6          3.23774
  Residual                  2981.64      150          19.8776
  Total                     3001.06      156          19.2376

  F(6, 150) = 3.23774 / 19.8776 = 0.162884 [p-value 0.9861]

  Level          n        mean     std. dev

  1             30     16.1167       3.7321
  2             99     16.6071       4.9370
  3             20      17.045       2.8333
  4              3        17.4       4.0596
  5              2        17.4      0.84853
  6              2        18.1       1.6971
  7              1        17.5           NA

  Grand mean = 16.6191
```

Figure 2.8: ANOVA output for Guests vs. PctTip.

Credit vs. PctTip: The payment method does not significantly influence the tip percentage. The mean tip for cash (16.4%) was very similar to that for credit cards (17.1%). The difference was found to be statistically insignificant (F(1,155)=0.91, p=0.342).

```
Analysis of Variance, response = PctTip, treatment = Credit:

                     Sum of squares       df      Mean square

  Treatment                 17.4686        1          17.4686
  Residual                  2983.59      155           19.249
  Total                     3001.06      156          19.2376

  F(1, 155) = 17.4686 / 19.249 = 0.907508 [p-value 0.3423]

  Level          n        mean     std. dev

  1            106     16.3877       5.0505
  2             51        17.1       2.4710

  Grand mean = 16.6191
```

Figure 2.9: ANOVA output for Credit vs. PctTip.

Day vs. PctTip: The day of the week does not significantly affect tipping habits. The ANOVA test confirmed that minor variations in mean tips across days were not statistically significant (F(4,152)=0.53, p=0.717).

```
Analysis of Variance, response = PctTip, treatment = Day:

                     Sum of squares       df      Mean square

  Treatment                 40.9194        4          10.2299
  Residual                  2960.14      152          19.4746
  Total                     3001.06      156          19.2376

  F(4, 152) = 10.2299 / 19.4746 = 0.525291 [p-value 0.7173]

  Level          n        mean     std. dev

  1             26     16.2577       6.1691
  2             36     16.8694       3.3947
  3             62     16.5516       3.4304
  4             13     18.0231       7.5965
  5             20      15.935       3.2042

  Grand mean = 16.6191
```

Figure 2.10: ANOVA output for Day vs. PctTip.

2.0.3 Scatter Plot Observations

Scatter plots were used to visualize the relationships between variables and visually corroborate the ANOVA findings.

Server vs. Bill, Tip, and PctTip: The scatter plot shows that while bills and tips vary widely across the different servers, the tip percentages remain clustered in a narrow and heavily overlapping range. This visual evidence supports the ANOVA result.

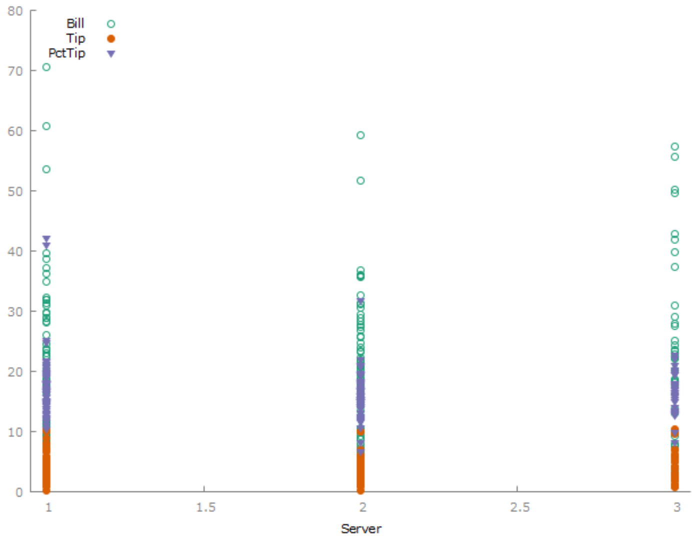

Figure 2.11: Scatter plot matrix of Bill, Tip, and PctTip vs. Server.

Day vs. Bill, Tip, and PctTip: The scatter plot reveals no distinct pattern associated with the day of the week. Bills and tips are scattered evenly across all days, and the percentage of tips stays consistently in the 15-20% range with strong overlap, aligning with the ANOVA conclusion.

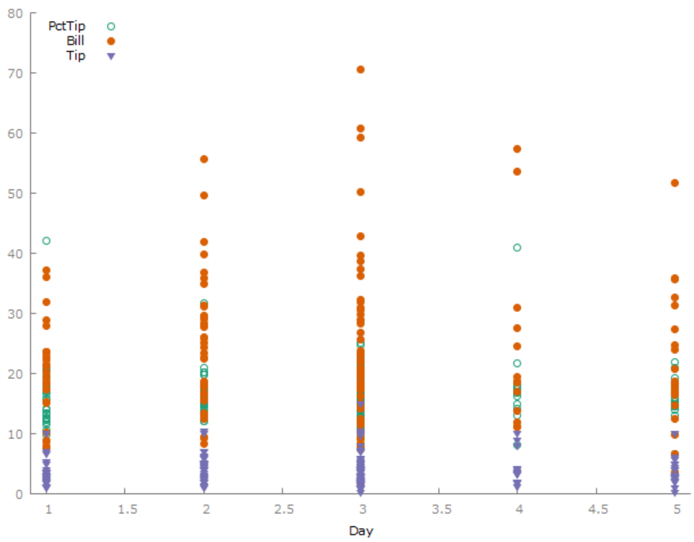

Figure 2.12: Scatter plot matrix of Bill, Tip, and PctTip vs. Day.

Number of Guests vs. Tip: These plots show a visible positive trend where the absolute tip amount increases as the number of guests rises. However, the `PctTip` remains steady across group sizes, confirming the ANOVA result that guest count has no significant effect on tipping percentage.

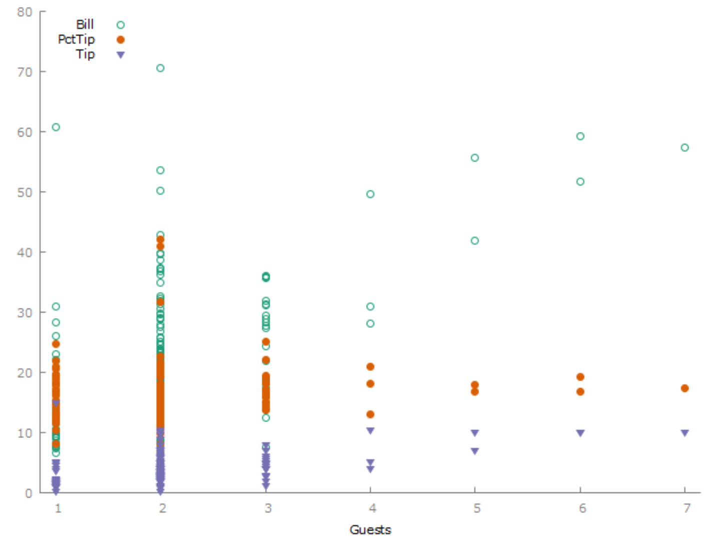

Figure 2.13: Scatter plot matrix of Bill, Tip, and PctTip vs. Guests.

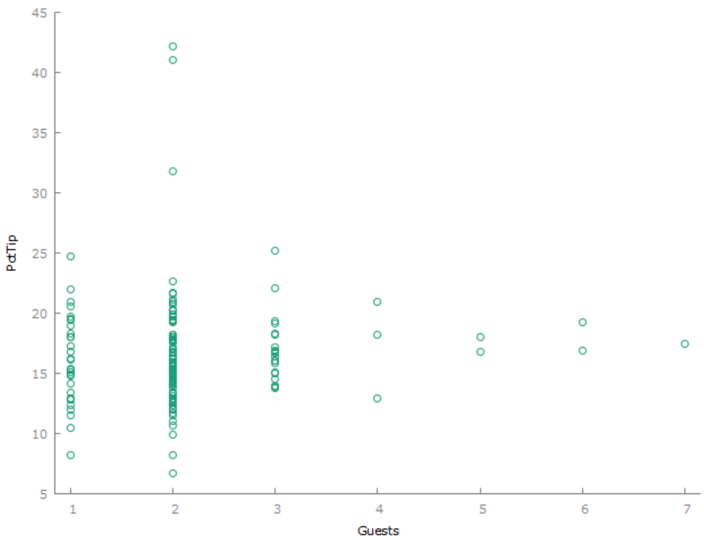

Figure 2.14: Scatter plot of Tip vs. Number of Guests.

Payment Method (Credit) vs. Tip: When factored by server and by day, these plots show strong overlapping spreads in tip amounts for both cash and credit payments. No interaction effect is visible, and tipping patterns remain consistent, confirming the ANOVA findings.

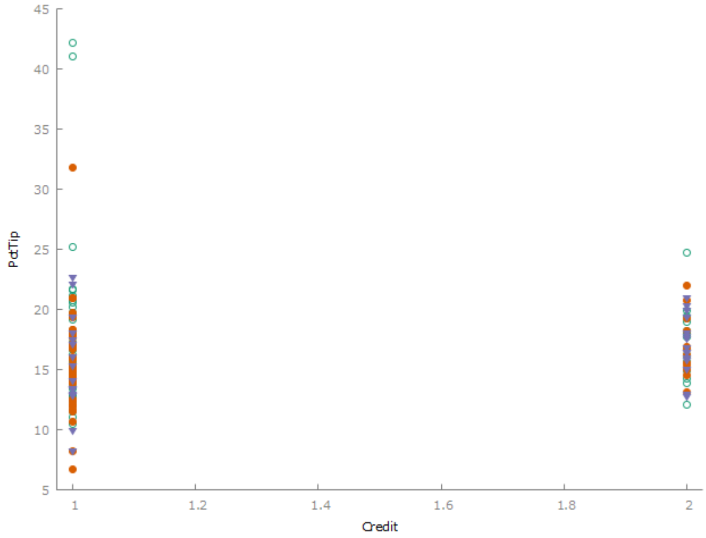

Figure 2.15: Scatter plot of Tip vs. Credit, factored by Server.

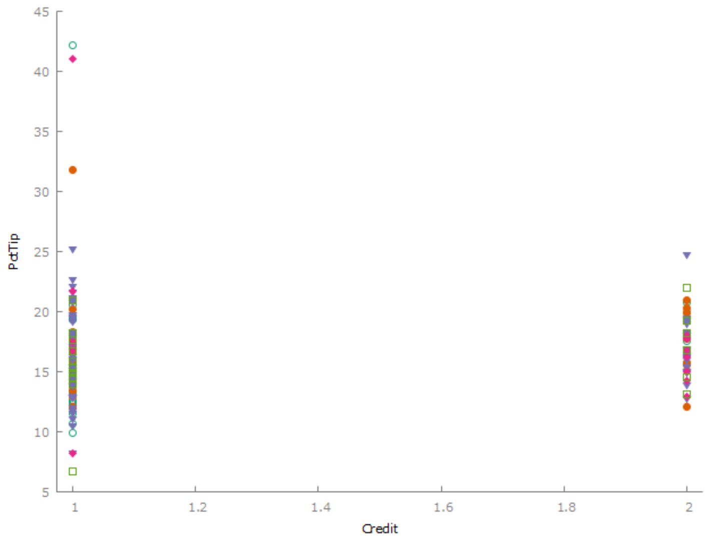

Figure 2.16: Scatter plot of Tip vs. Credit, factored by Day.

Chapter 3

Python Analysis

To validate and expand upon the initial findings, the dataset was further analyzed using Python with its powerful data science libraries, including `pandas` for data manipulation, `matplotlib` and `seaborn` for visualization, and `statsmodels` for statistical testing. This analysis mirrored the steps taken in Gretl and confirmed the previous conclusions through a deeper visual exploration.

3.0.1 Data Exploration and Summary Statistics

The dataset was loaded into a `pandas` DataFrame for initial exploration. The summary statistics revealed that for both `Bill` and `Tip`, the mean is higher than the median, confirming the right-skewed nature of their distributions. An analysis of categorical variables showed that most dining parties consist of two guests.

3.0.2 Univariate and Bivariate Visualization

Visualizations were created to understand the distribution of individual variables and the relationships between pairs of variables.
Distributions: Histograms and box plots reaffirmed the right-skewed distributions for `Bill` and `Tip` and the presence of outliers.

```
      Bill    Tip  Guests
0    23.70  10.00       2
1    36.11   7.00       3
2    31.99   5.01       2
3    17.39   3.61       2
4    15.41   3.00       2
..     ...    ...     ...
152  31.30   5.70       3
153  12.57   2.00       3
154  14.87   3.13       2
155  51.68  10.00       6
156  17.12   2.50       2

[157 rows x 3 columns]



IQR For Total Bill :  13.65
IQR For Tip :  2.9
IQR For Total Bill & Tip :  13.094999999999999
```

Figure 3.1: Interquartile Range (IQR) for key variables.

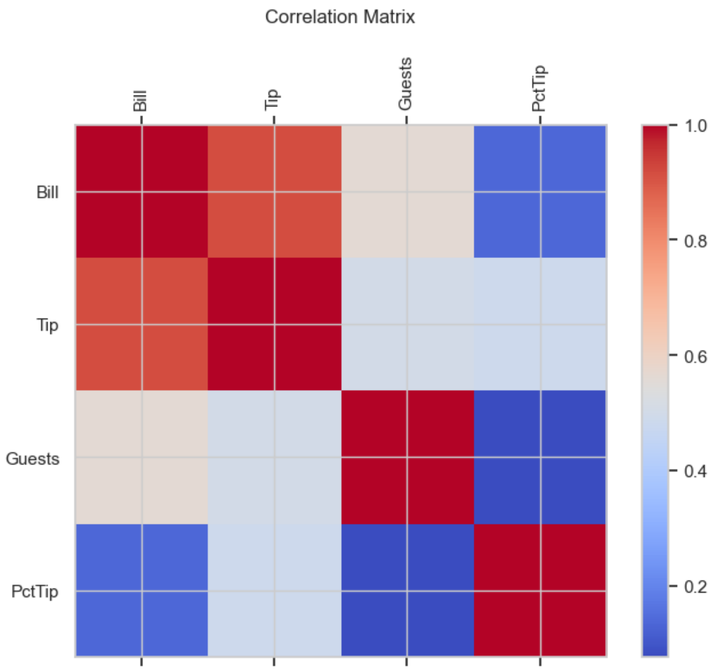

Figure 3.2: Correlation matrix heatmap.

This correlation matrix heatmap visually shows the relationships among bill amount, tip amount, number of guests, and tip percentage (PctTip). As expected, the bill and tip amounts are strongly positively correlated (deep red), meaning larger bills naturally result in higher absolute tips. Similarly, the number of guests correlates positively with bill size and tip amount, reflecting that bigger groups tend to spend more and tip more in absolute terms. In contrast, PctTip shows weak or even negative correlations with bill, tip, and guests (blue shades), indicating that while total spending grows with group size, the percentage of tipping remains relatively constant and independent of these factors. This reinforces the earlier ANOVA findings that operational variables like group size or bill size do not significantly affect tipping percentage.

Pairwise Relationships: A `seaborn` pairplot and a correlation heatmap show a strong positive linear relationship between `Bill` and `Tip`.

This pairplot displays the distributions and scatterplots of bill amount, tip amount, number of guests, and tip percentage (PctTip), making it easier to observe relationships between variables. The diagonal histograms show that most bills and tips are concentrated at lower values, while guest counts cluster around smaller group sizes (2–4). The scatterplots reveal a strong linear relationship between bill and tip, confirming that larger bills lead to proportionally larger tips. Guest count shows a positive link with bill size and tip amount, though with more spread, since group spending varies. In contrast, PctTip does not display a clear upward trend with bill, tip, or guests, appearing scattered across values—this supports the finding that while spending and absolute tipping grow with group size, tipping percentage remains largely independent and stable.

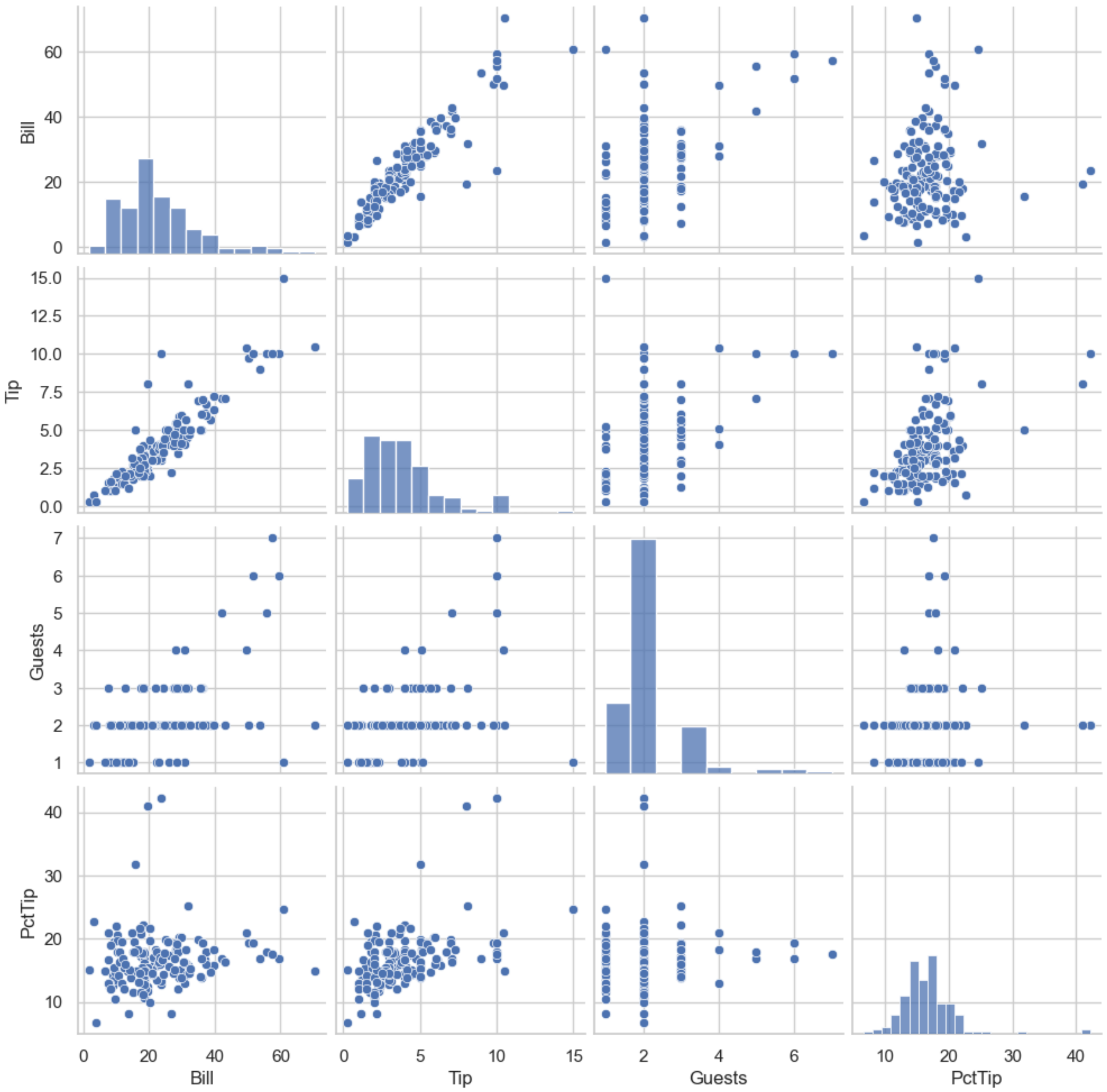

Figure 3.3: Pairplot of numerical variables.

3.0.3 Multivariate Visualization

To investigate the influence of categorical variables, multivariate plots were generated.
Scatter Plots with Hue: Scatter plots of `Tip` vs. `Bill` colored by `Server` and `Day` show that categories are thoroughly intermingled, suggesting no distinct tipping patterns.

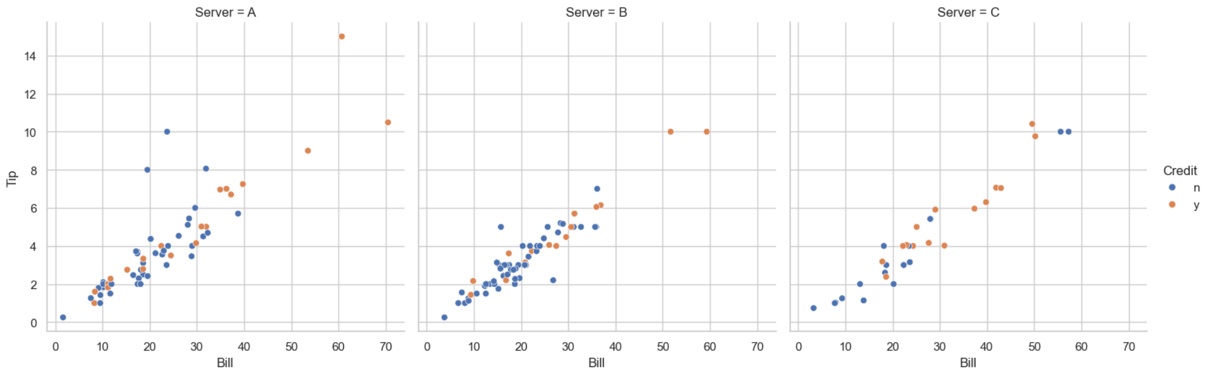

Figure 3.4: Tip vs. Bill, colored by Server.

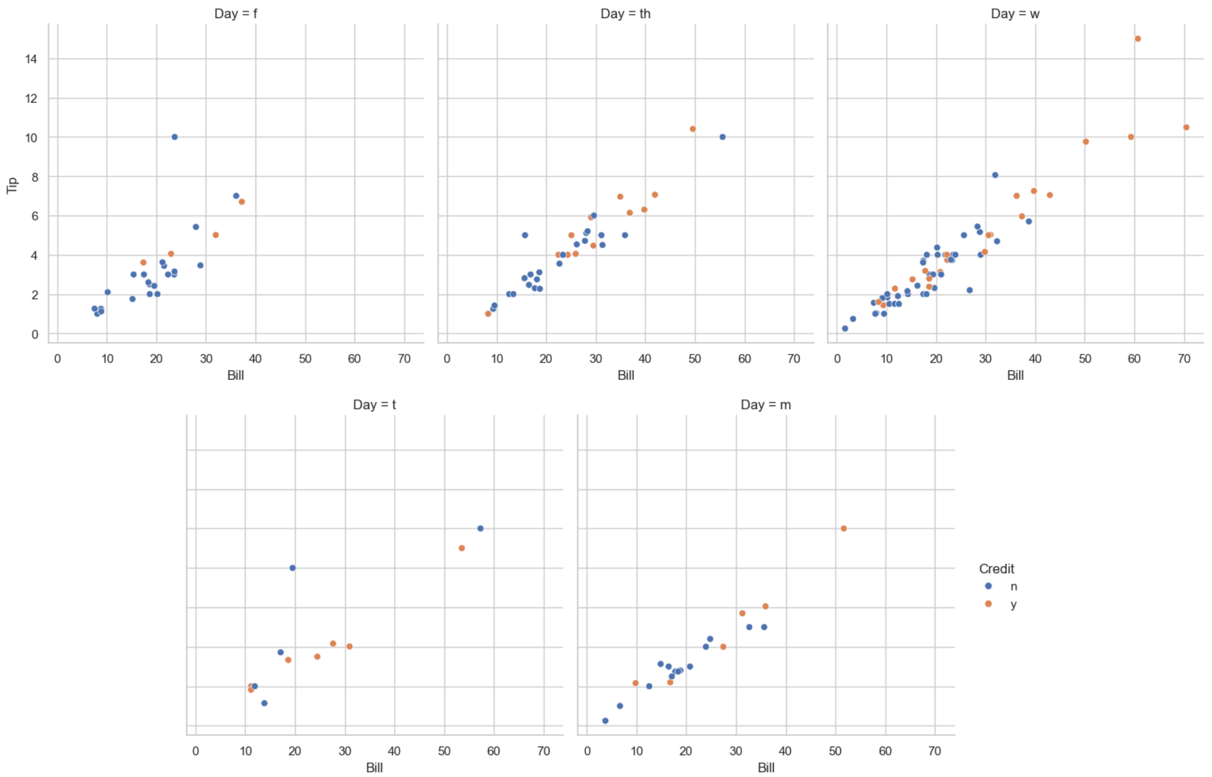

Figure 3.5: Tip vs. Bill, colored by Day.

This ScatterpLots also highlights the presence of outliers and disparities among groups. A few extreme points appear in both bill and tip values, showing unusually high spending and tipping that stand apart from the main cluster of data. Similarly, some PctTip values exceed 30–40%, which are clear outliers compared to the majority tipping norm of 15–20%. Disparity is also visible in group sizes: most transactions are concentrated around 2–4 guests, while larger groups are rare, leading to uneven data distribution across categories. These patterns reinforce that while the overall trend is consistent, individual anomalies and unequal group representation must be acknowledged when interpreting the results.

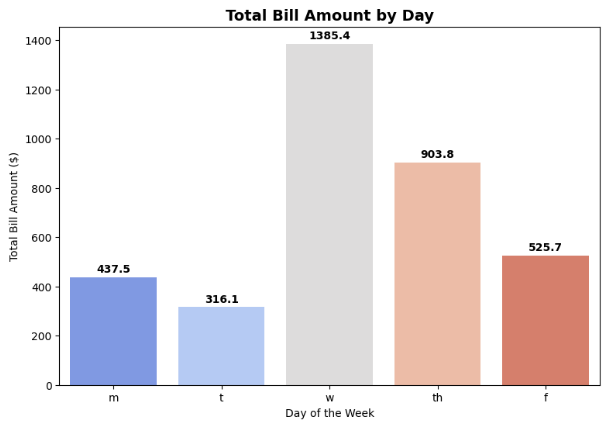

Figure 3.6: Total Bill amount for each Day

The observed pattern, with Wednesday showing the highest total bill amount and Monday–Tuesday recording the lowest, can be explained by customer behavior and dining habits. Mid-week often serves as a natural break from work routines, making Wednesday a popular choice for socializing or dining out, sometimes further boosted by restaurant promotions or special offers. In contrast, the start of the week usually sees lower customer turnout as people focus on work, budgeting after weekend spending, or preferring home-cooked meals. By Friday, spending is slightly lower than the mid-week peak as many customers defer their dining plans to the weekend, leading to the distinctive mid-week surge in restaurant bills.
Hue-Differentiated Pairplots: Full pairplots colored by `Day` and `Credit` show no clear separation of data points, visually reinforcing the ANOVA conclusions.

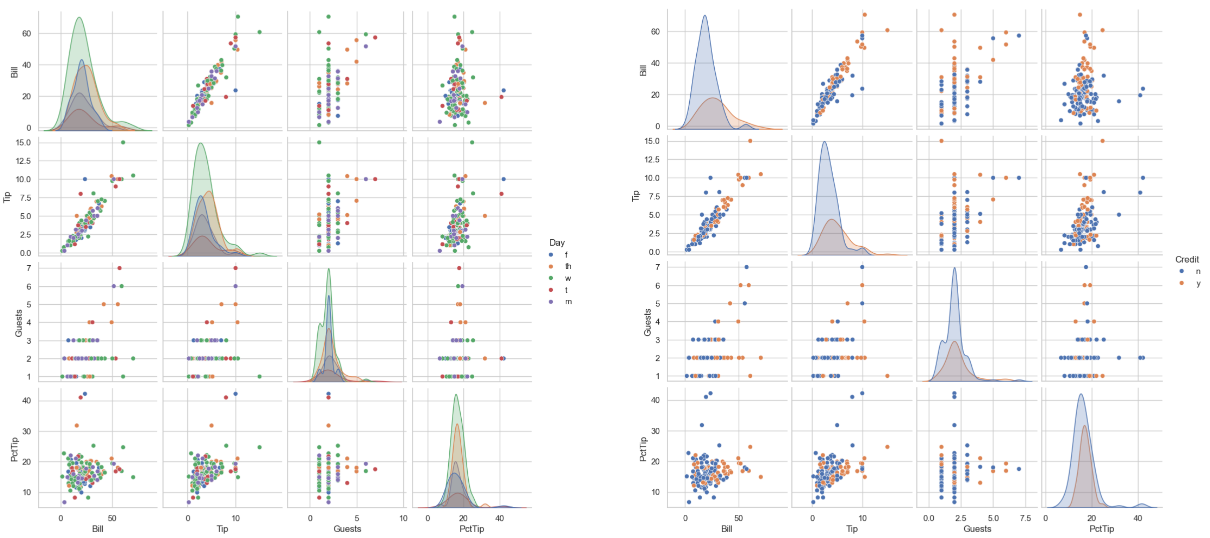

Figure 3.7: Python Pairplot with data points colored by Day and Credit Payment respectively.

3.0.4 Statistical Testing: ANOVA

Formal ANOVA tests, replicated in Python, confirmed with high statistical confidence that none of the tested variables have a significant effect on the tip percentage.

	df	sum_sq	mean_sq	F	PR(>F)
C(Day)	4.0	40.919409	10.229852	0.525291	0.717291
Residual	152.0	2960.143266	19.474627	NaN	NaN

	df	sum_sq	mean_sq	F	PR(>F)
C(Credit)	1.0	17.468619	17.468619	0.907508	0.34226
Residual	155.0	2983.594057	19.248994	NaN	NaN

Figure 3.8: Python ANOVA results for Day and Credit vs. PctTip.

	df	sum_sq	mean_sq	F	PR(>F)
C(Server)	2.0	83.136770	41.568385	2.193864	0.114958
Residual	154.0	2917.925905	18.947571	NaN	NaN

	df	sum_sq	mean_sq	F	PR(>F)
C(Guests)	6.0	19.426458	3.237743	0.162884	0.986098
Residual	150.0	2981.636217	19.877575	NaN	NaN

Figure 3.9: Python ANOVA results for Server and Guests vs. PctTip.

Chapter 4

SQL Analysis

Guests	Credit	Transactions	Avg_Tip_Percentage
1	n	23	15.16
1	y	7	19.26
2	n	66	16.51
2	y	33	16.79
3	n	14	17.49
3	y	6	16.02
4	n	1	18.2
4	y	2	17
5	n	1	18
5	y	1	16.8
6	y	2	18.1
7	n	1	17.5

Figure 4.1: SQL Query output for guest size behavior with Credit card usage

This table highlights the relationship between group size, payment mode, and tipping behavior. We can observe that for smaller groups (1–2 guests), cash payments dominate, but as the number of guests increases—and with it, the total bill amount—the proportion of credit card transactions rises noticeably. For example, at 2 guests, nearly one-third of the payments are already on credit, and for groups of 4–6, credit card usage becomes more common despite fewer transactions overall. This suggests that customers tend to switch to credit cards for larger bills, likely for convenience and higher spending capacity, while smaller bills are comfortably paid in cash. Interestingly, the average tip percentage remains within a steady band (15–19%) across all segments, reinforcing the idea that while payment method shifts with bill size, tipping norms remain socially consistent.

Bill_Category	Avg_Tip	Avg_Tip_Percentage	Transactions
Large (>50)	10.53	18.44	8
Medium (20-50)	4.84	16.89	71
Small (<20)	2.26	16.19	78

Figure 4.2: SQL Query Output for Bill amount segregation and impact on Tip

This table shows how tipping patterns vary across different bill size categories. As expected, the average tip amount rises significantly with larger bills—moving from □2.26 for small bills to

□10.53 for large bills—reflecting the direct link between bill size and absolute tip. Interestingly, the average tip percentage also increases slightly with larger bills, from about 16% for small bills to over 18% for large bills. This indicates that not only do customers leave higher tips when the bill is larger, but they also tend to be more generous in percentage terms on big-ticket dining experiences, possibly due to social norms or the higher perceived service value.

Server	Bill_Category	Avg_Tip_Percentage	Orders
A	Large (>50)	18.8	3
A	Medium (20-50)	17.88	26
A	Small (<20)	17.14	31
B	Large (>50)	18.1	2
B	Medium (20-50)	16.25	28
B	Small (<20)	15.71	35
C	Large (>50)	18.3	3
C	Medium (20-50)	16.43	17
C	Small (<20)	15.11	12

Figure 4.3: SQL Query Output for Server names Grouped by Bill amount and it's impact on tips

This table compares average tip percentages across servers and bill categories. A clear pattern emerges: for all servers, larger bills (>50) consistently attract higher tip percentages (18–19%), while smaller bills (<20) see noticeably lower averages (15–17%). For example, Server A receives around 18.8% tips on large bills versus 17.1% on small bills, and a similar gap is seen for Servers B and C. This suggests that customer generosity scales both in absolute and percentage terms with bill size, independent of who the server is. The differences between servers are relatively minor, with averages falling within a narrow band, indicating that server identity itself does not strongly influence tipping patterns, but bill size does.

Day	Server	Transactions	Total_Bill_Amount	Total_Tip_Amount
w	C	13	276.11	47.09
w	B	21	437	68.75
w	A	28	672.27	118.03
th	C	9	287.82	51.93
th	B	14	331.79	55.65
th	A	13	284.2	47.7
t	C	4	129.8	19.32
t	A	9	186.27	36.72
m	B	20	437.47	71.69
f	C	6	135.56	20.22
f	B	10	167.61	27.17
f	A	10	222.56	40.07

Figure 4.4: SQL Query Output for Total Bill and Tip collected by each server on all weekdays

Server A consistently handles the highest bill volumes and tips, particularly on Wednesdays, where both bill and tip totals peak at \$672.27 and \$118.03 respectively. Server B follows closely, performing strongly on both Wednesdays and Mondays with high transaction counts and significant contributions to overall revenue. Server C, however, records lower bill and tip totals across all days, reflecting fewer transactions or smaller group sizes served. Notably, Wednesdays emerge as the busiest day for all servers combined, while early-week days like Tuesday and later days such as Friday show reduced totals, highlighting distinct mid-week peaks in customer activity and spending.

Chapter 5

Inference and Final Remarks

The collective evidence from both the Gretl and Python analyses leads to a single, robust inference: the percentage of a bill that a customer leaves as a tip is a remarkably stable metric, governed more by convention than by the immediate operational circumstances of their visit.

While descriptive statistics clearly show that absolute `Bill` and `Tip` amounts are volatile and right-skewed, the `PctTip` remains consistently centered around a median of 16-17%. This stability is the central finding of the report. The formal ANOVA tests, replicated across both software packages, provide definitive statistical confirmation. In every test—comparing tip percentages across servers, days of the week, payment methods, and group sizes—the resulting p-values were high (all > 0.1), indicating a lack of statistical significance. We therefore fail to reject the null hypothesis in all cases and conclude that these variables do not have a meaningful effect on the tipping rate.

This statistical conclusion is strongly supported by the data visualizations. The Gretl scatter plots and the more detailed Python pairplots, which show relationships colored by different categories, reveal a complete intermingling of data points. There is no evidence of clustering or separation that would suggest, for example, that customers of Server A tip differently than customers of Server B, or that tipping behavior changes on a weekend.

The most logical explanation for this consistency is the presence of a powerful social norm. Customers appear to follow a deeply ingrained heuristic for tipping (e.g., 15-20% of the bill) that overrides the other contextual factors examined. The only variable that reliably predicts the absolute tip amount is the total bill itself, as shown by the strong positive correlation. This, however, is not a change in behavior but rather the application of the same percentage rule to a different base amount.

Conclusion

The objective of this report was to dissect a restaurant's transaction data to identify the key drivers of customer tipping percentage. Through a comprehensive analysis utilizing descriptive statistics, extensive data visualization, and formal ANOVA testing in both Gretl and Python, a clear and consistent pattern emerged.

The central finding of this study is that tipping percentage is a highly predictable and stable aspect of customer behavior, and it is not significantly influenced by the operational variables investigated. The analysis conclusively shows that there is no statistically significant difference in the mean tip percentage based on the day of the week, the payment method used, the specific server attending the table, or the number of guests in a party.

While absolute bill and tip amounts can vary widely, the tipping rate itself remains firmly anchored around a median of 16-17%, suggesting behavior is dictated by a strong social convention rather than situational factors. From a business and marketing perspective, this implies that strategies aimed at increasing overall tip revenue should focus on increasing the base bill amount through upselling or promotional deals, rather than expecting certain days or staff members to generate inherently higher tipping rates. The restaurant's tipping culture is, in effect, a constant.

Table 5.1: Summary of Key Conclusions

Factor	**Conclusion**
Day of the Week	No significant difference in tipping percentage across weekdays and weekends.
Payment Method (Cash vs. Credit)	Tipping percentage remains stable regardless of payment type.
Server Identity	Customers tip consistently; server has no significant effect.
Number of Guests	Larger groups increase bill and absolute tip, but percentage tip remains unchanged.
Overall Behavior	Tipping percentage is stable around 16–17%, driven by social convention rather than situational factors.
Business Implication	Strategies to raise tip revenue should focus on increasing bill size (upselling/promotions), not operational variables.