Phase 2 Task 4: Visuals

Team 2: Tsyr Rau Chen, Arailym Duisengali, Sheikh Noohery, Lo Ying Wu, Kuan Rong Yang

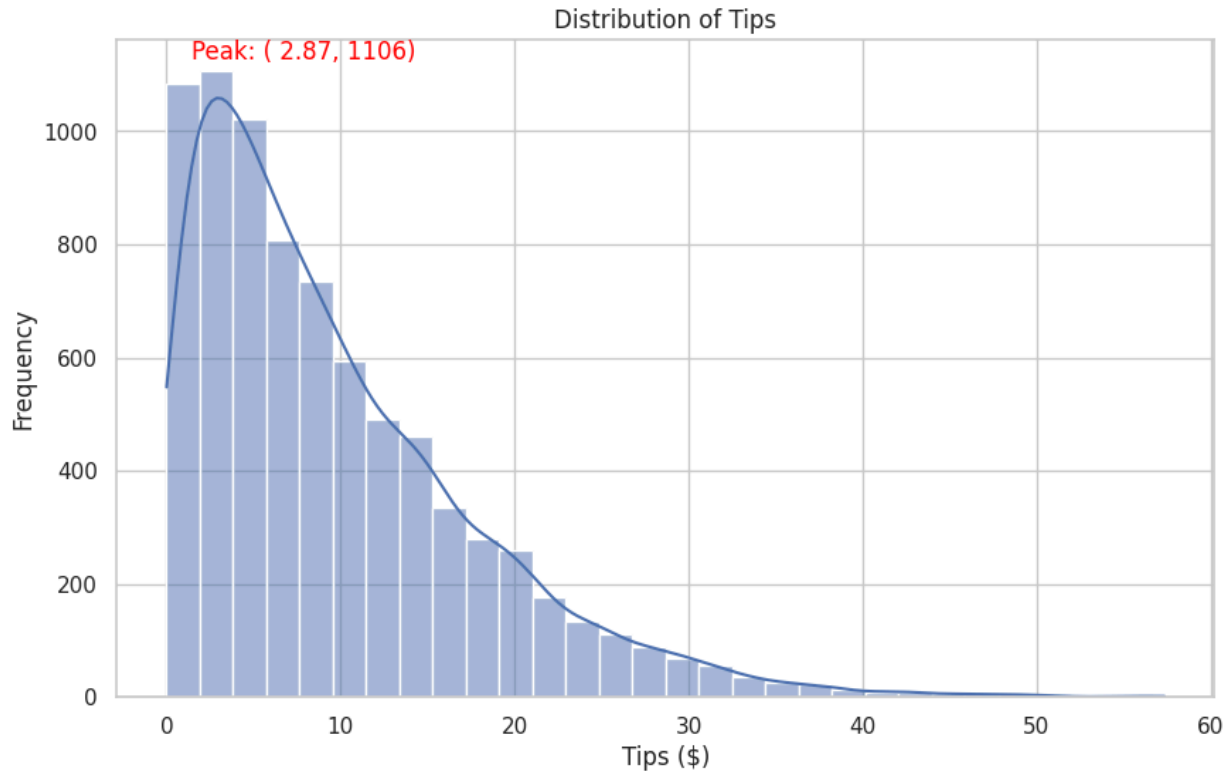

Figure 1: Distribution of Tips with Density Curve

Figure 1 is a histogram that displays the distribution of tip amounts with a density curve. The tip amount distribution is skewed right with most of the tips being concentrated between ~$0 to ~$5. The frequency peaks at $2.87 with 1,106 counts. This indicates that tipping behavior is strongly skewed towards smaller amounts. Larger tips, above $20, are rare, suggesting occasional generous tipping.

Figure 2 is a box plot that shows the distribution of tip amounts in each city of our database: Austin, New York, San Francisco, Chicago, Seattle, and Boston. Each city has a similar distribution with median tip amounts being under $10. All cities have outliers in the $30 - $50 tip amount range, showing that occasional large tips are common across all locations in the dataset. No boxplot stands out dramatically, meaning tipping behavior is broadly consistent across cities in our dataset. Based on this, we can assume that the city variable alone is not a strong predictor of tip amount.

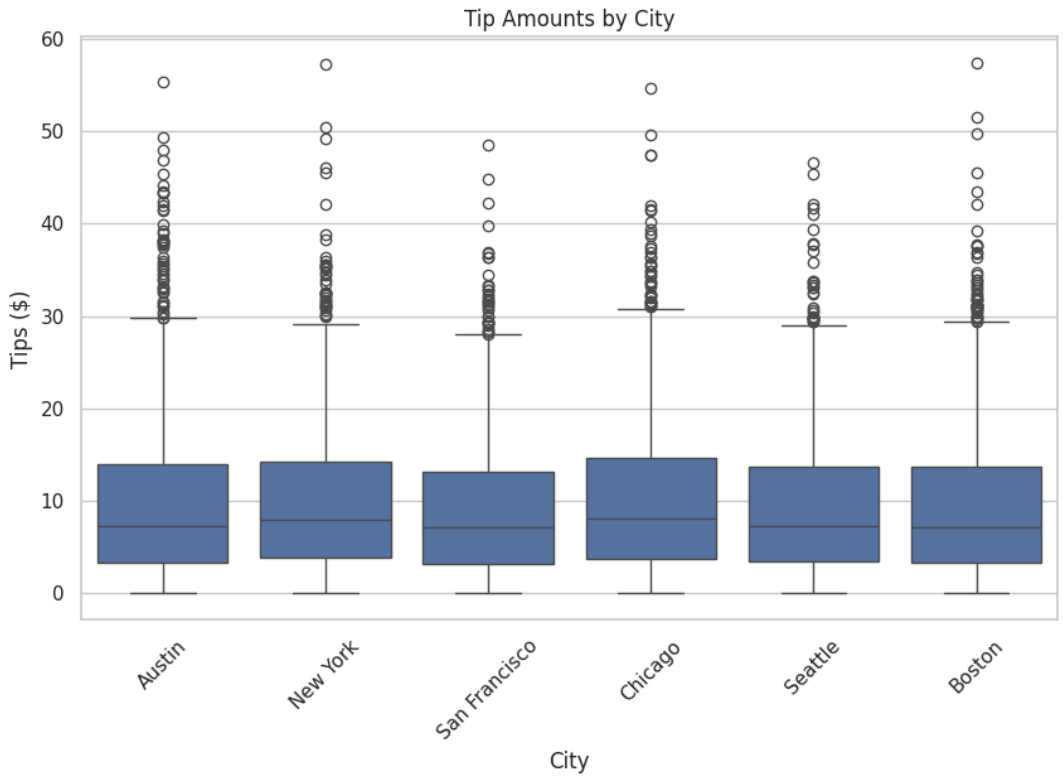

Figure 2: Distribution of Tips in Each City

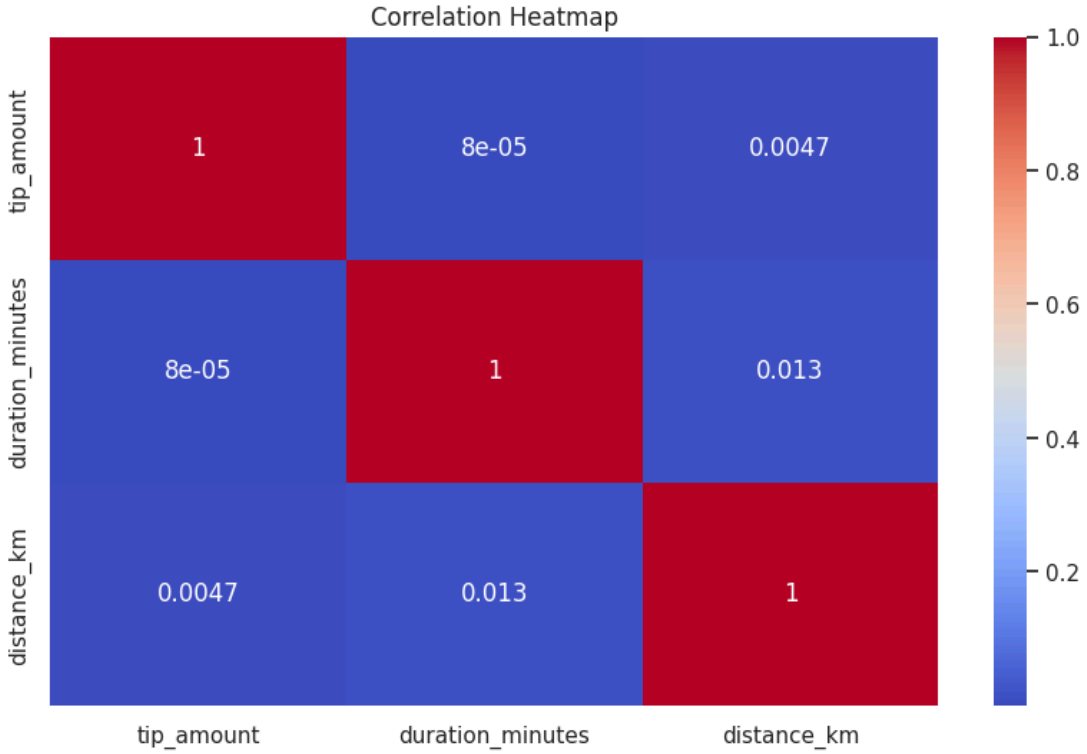

Figure 3: Correlation Heatmap of Numeric Variables

Figure 3 is a correlation heatmap of the numeric variables used in our machine learning models: tip_amount, duration_minutes, and distance_km. The correlation values for tip_amount vs duration_minutes is 0.00008, tip_amount vs distance_km is 0.0047, and duration_minutes vs distance_km is 0.013. The correlation values are extremely low for all combinations of the variables. That means there is no multicollinearity concern.

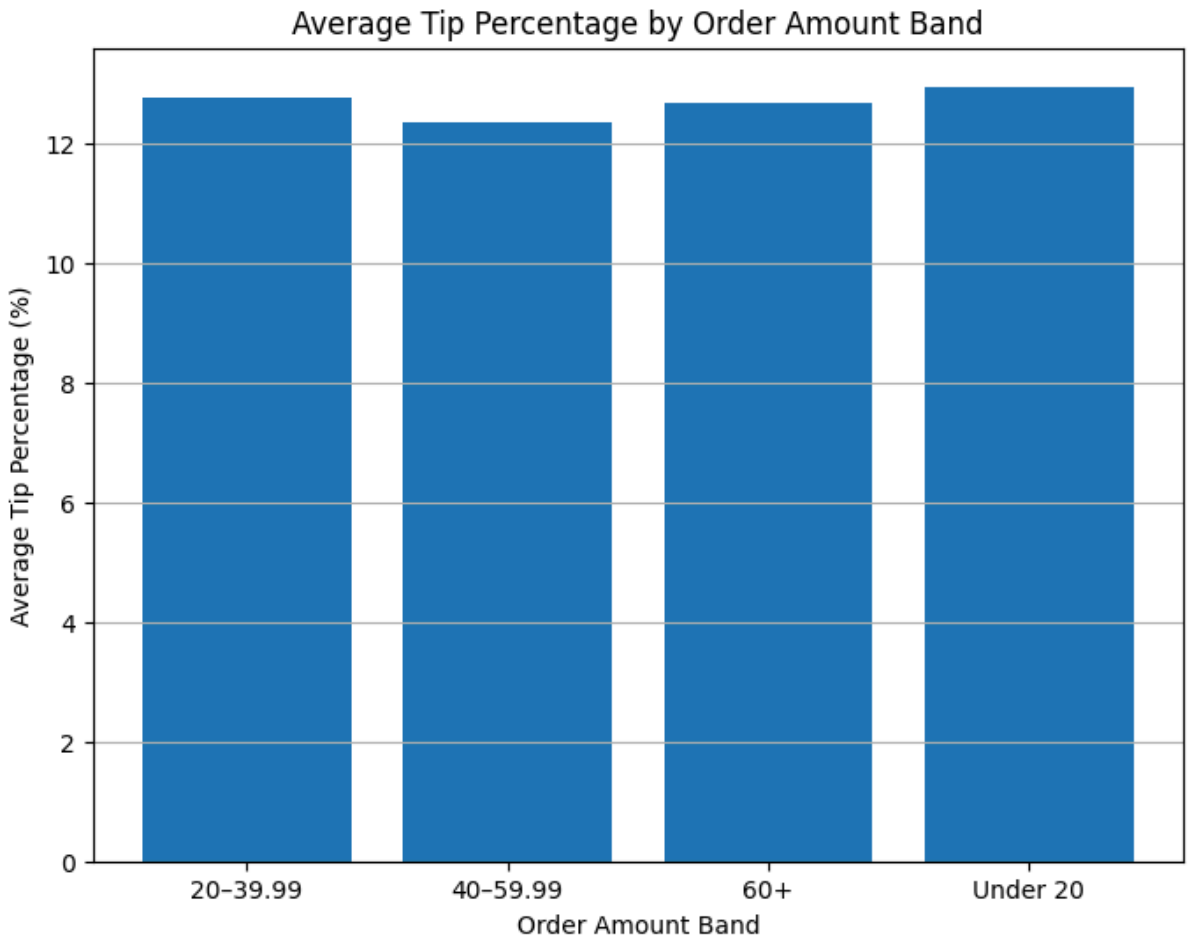

Figure 4: Relationship between average tip percentages and order amount

Figure 4 shows that the average tip percentage is fairly consistent across all order amount bands. Customers tend to tip around 12–13% regardless of whether their order total is under 20, between 20–39.99, 40–59.99, or above 60. This suggests that tip behavior does not strongly depend on how much the customer spends. Even higher-value orders do not lead to significantly higher or lower tip percentages, indicating that customers typically follow a standard tipping habit rather than adjusting the tip based on the order amount.

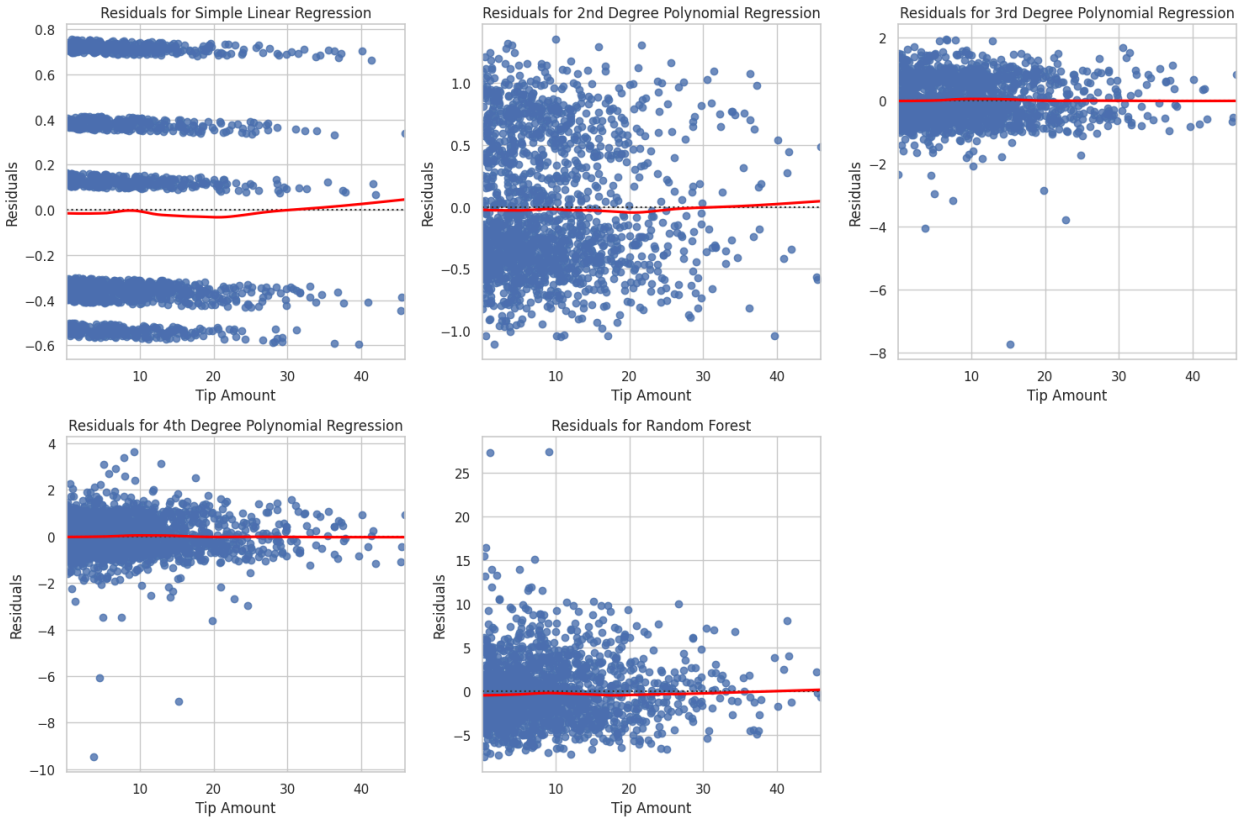

Figure 5: Residual Plots for Machine Learning Models

Figure 5 has residual plots that show how each model performed when predicting tip amounts. The simple linear regression residual plot has a clear pattern, which indicates underfitting. Polynomial models (2nd - 4th degree) are improvements but the higher degrees have extreme outliers, which is an indicator for overfitting. The random forest model also has extreme outliers in its residuals.

Figure 6 shows the relationship between how recently customers placed an order (recency) and how often they order overall (frequency). Most customers who order frequently tend to have low recency values, meaning they come back regularly and are highly engaged. As recency increases, the number of total orders generally decreases, indicating that customers who haven't ordered in a long time tend to be lower-frequency or at risk of churning. Overall, the plot highlights a clear pattern: loyal customers order both often and recently, while inactive customers show higher recency and fewer total orders.

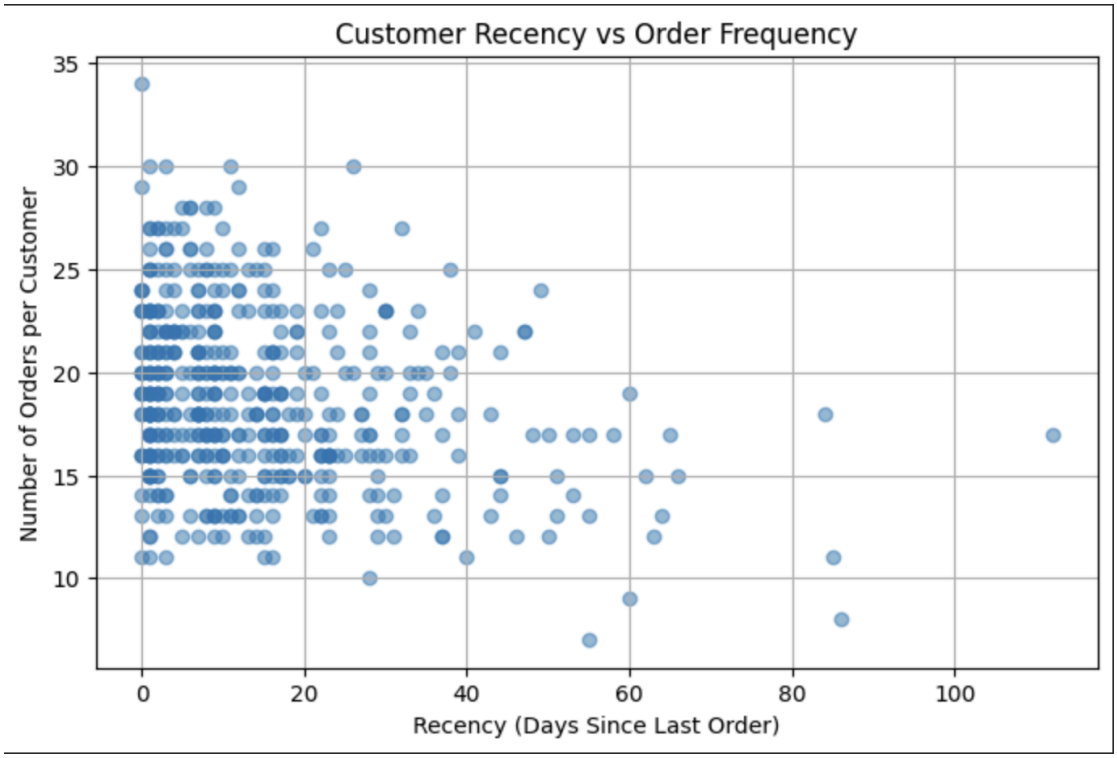

Figure 6: Customer Recency vs Order Frequency

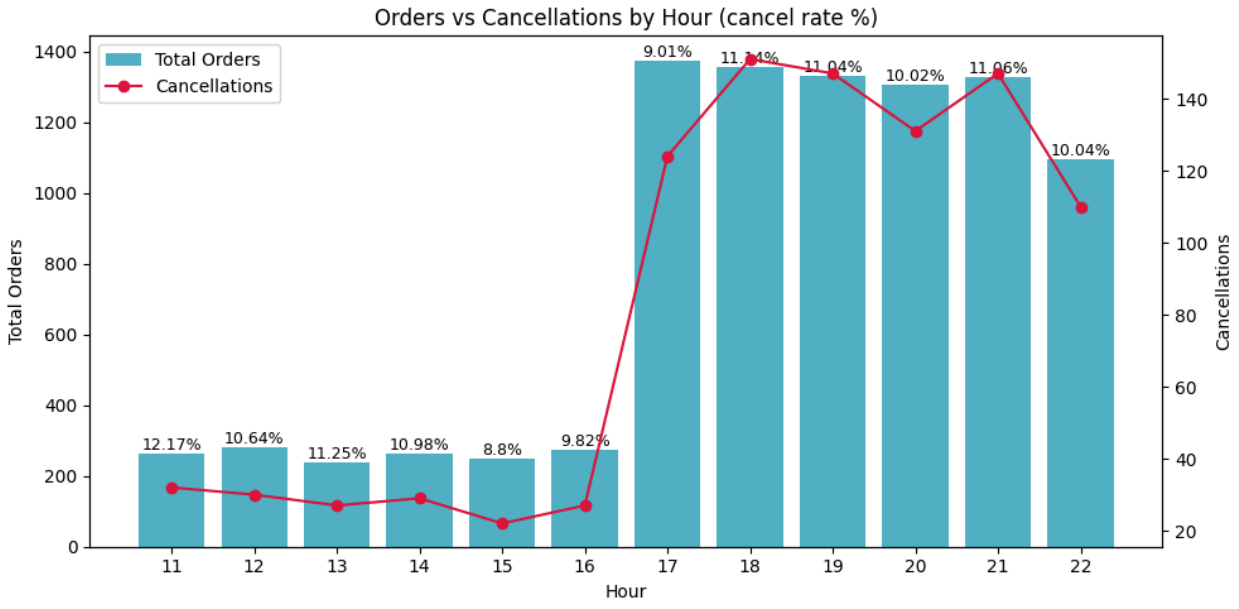

Figure 7: Orders vs Cancellations by Hour (with cancellations rate)

The plot shows how total orders and cancellations vary by hour, along with the cancellation rate. Order volume is relatively low between 11:00 and 16:00, and the cancellation rate during this period stays around 8–12%. Starting at 17:00, total orders increase sharply, reaching their peak between 18:00 and 21:00. Cancellations also rise during these peak hours, but the cancellation rate remains fairly stable around 10–11%, indicating that the higher volume does not disproportionately increase cancellations. After 21:00, both orders and cancellations begin to decline, and the cancellation rate slightly drops as well. Overall, the plot highlights a strong evening peak in demand with consistently moderate cancellation rates throughout the day.