Analysis of TIPS Data to find relationship between variables -

INTRODUCTION

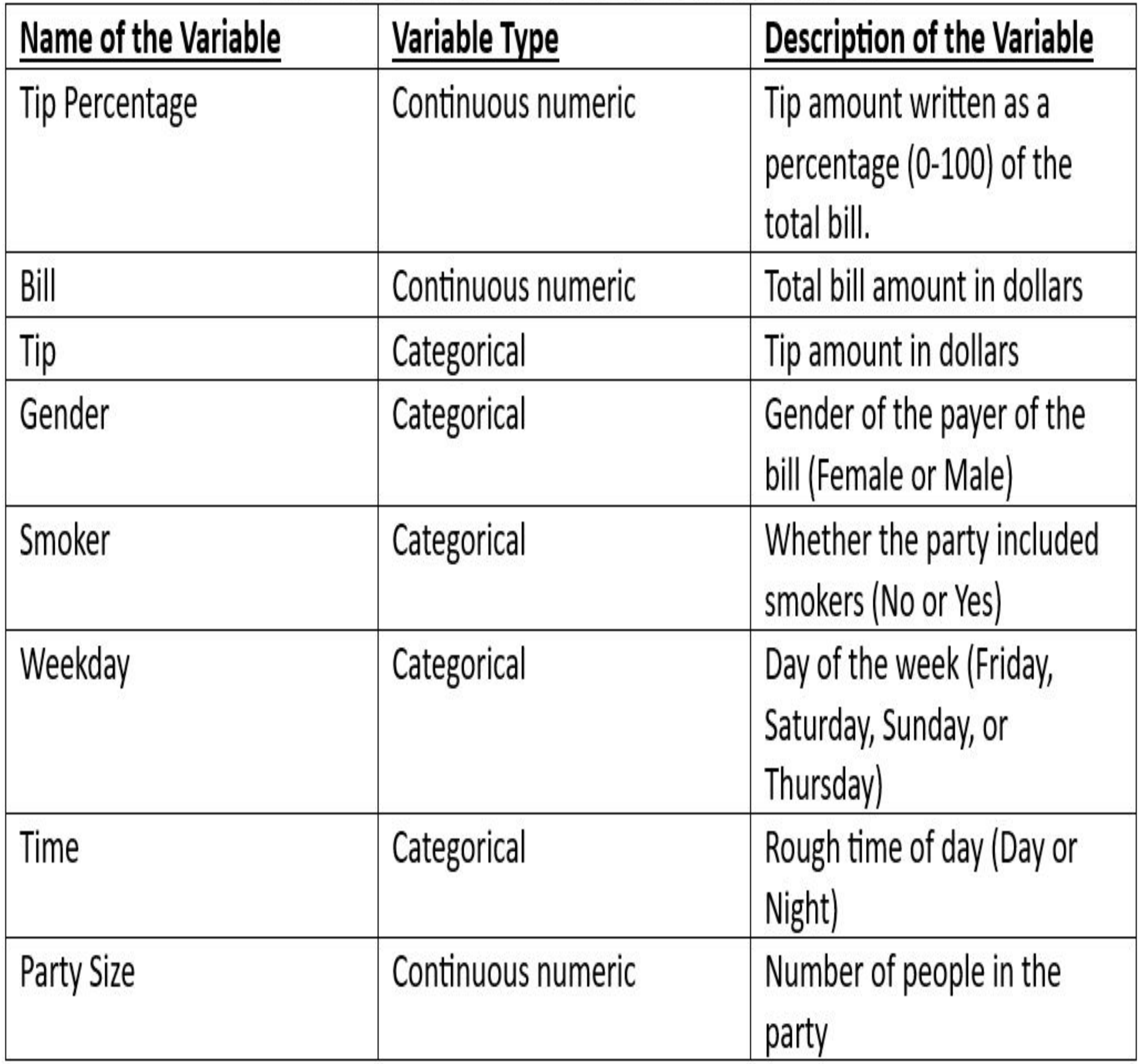
This project is an analysis of relationships between different variables in TIPS data. However, the focus will be on the impacts of gender, weekday, time, smoker, and party size on the tip percentage and the bill amount. Firstly, we are analyzing the variables individually and then examining them in combination to understand their impact on the tip percentage and bill as a whole. We have analyzed the association between categorical variables (gender, weekday, time, and smoker) and numerical variables (tip percentage and bill). The association between two numerical variables is being analyzed with party size and bill. Our goal for this project is to determine which aspect of each variable contributes the most to the waiter's earnings.

DATA

Intro and description of data

The dataset used in my study includes eight variables related to tipping behavior and various factors that may influence it. The variables are categorized as either categorical or continuous numeric variables.

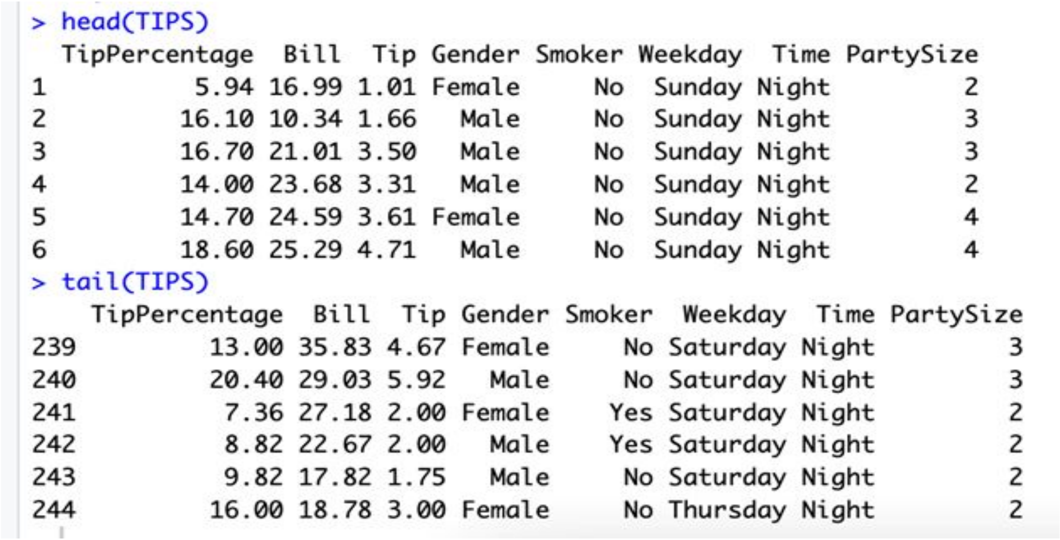
Table 1: Description of all Variables in the Data set



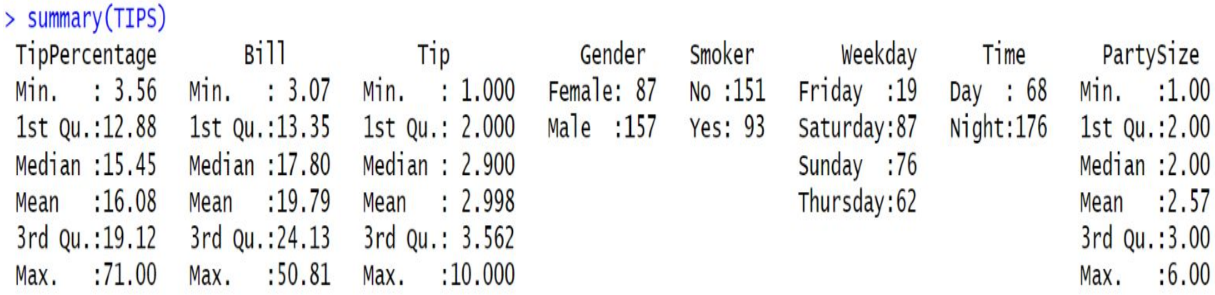
Name of the Variable	Variable Type	Description of the Variable
Tip Percentage	Continuous numeric	Tip amount written as a percentage (0-100) of the total bill.
Bill	Continuous numeric	Total bill amount in dollars
Tip	Categorical	Tip amount in dollars
Gender	Categorical	Gender of the payer of the bill (Female or Male)
Smoker	Categorical	Whether the party included smokers (No or Yes)
Weekday	Categorical	Day of the week (Friday, Saturday, Sunday, or Thursday)
Time	Categorical	Rough time of day (Day or Night)
Party Size	Continuous numeric	Number of people in the party

Summary of the Variable

Example and Summary



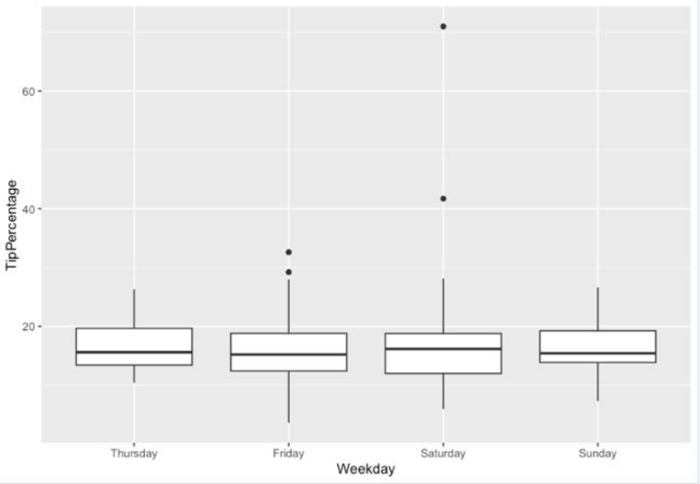
```
> head(TIPS)
  TipPercentage  Bill  Tip Gender Smoker Weekday  Time PartySize
1          5.94 16.99 1.01 Female     No  Sunday Night         2
2         16.10 10.34 1.66   Male     No  Sunday Night         3
3         16.70 21.01 3.50   Male     No  Sunday Night         3
4         14.00 23.68 3.31   Male     No  Sunday Night         2
5         14.70 24.59 3.61 Female     No  Sunday Night         4
6         18.60 25.29 4.71   Male     No  Sunday Night         4
> tail(TIPS)
    TipPercentage  Bill  Tip Gender Smoker  Weekday  Time PartySize
239         13.00 35.83 4.67 Female     No Saturday Night         3
240         20.40 29.03 5.92   Male     No Saturday Night         3
241          7.36 27.18 2.00 Female    Yes Saturday Night         2
242          8.82 22.67 2.00   Male    Yes Saturday Night         2
243          9.82 17.82 1.75   Male     No Saturday Night         2
244         16.00 18.78 3.00 Female     No Thursday Night         2
```

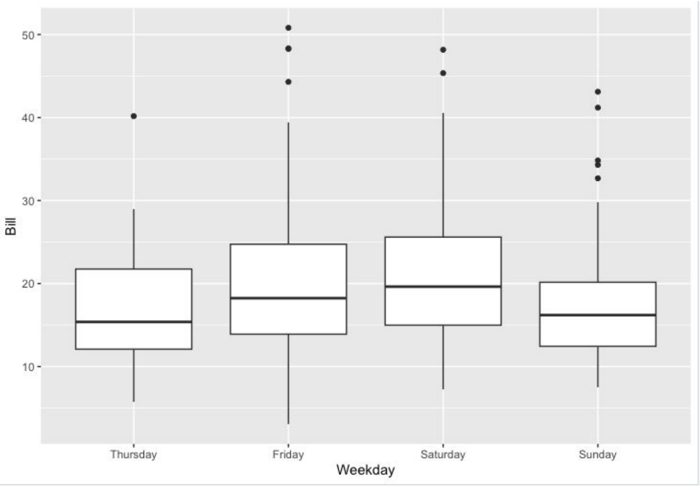


```
> summary(TIPS)
 TipPercentage        Bill            Tip          Gender     Smoker       Weekday        Time       PartySize  
 Min.   : 3.56   Min.   : 3.07   Min.   : 1.000   Female: 87   No :151   Friday  :19   Day  : 68   Min.   :1.00  
 1st Qu.:12.88   1st Qu.:13.35   1st Qu.: 2.000   Male  :157   Yes: 93   Saturday:87   Night:176   1st Qu.:2.00  
 Median :15.45   Median :17.80   Median : 2.900                          Sunday  :76               Median :2.00  
 Mean   :16.08   Mean   :19.79   Mean   : 2.998                          Thursday:62               Mean   :2.57  
 3rd Qu.:19.12   3rd Qu.:24.13   3rd Qu.: 3.562                                                    3rd Qu.:3.00  
 Max.   :71.00   Max.   :50.81   Max.   :10.000                                                    Max.   :6.00  
```

Visual data analysis of variables

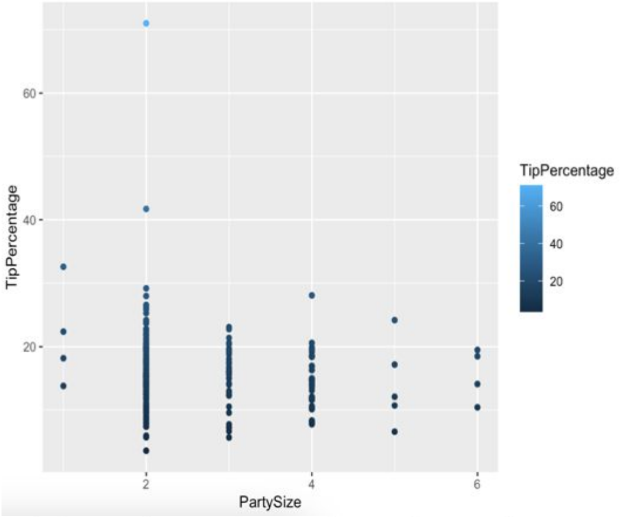
Categorical > Numeric

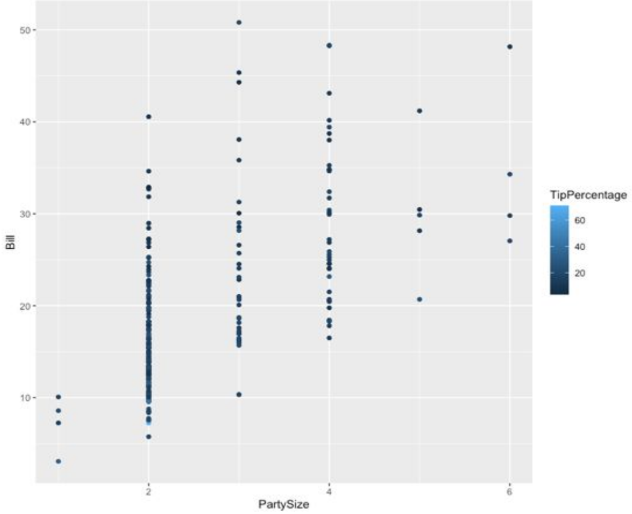




I made some visuals to look for associations between different variables. I created a box plot to examine associations between all categorical variables and the numeric variables Tip Percentage and Bill. The assumption about association was based on the median in the plot. For associations between numeric and numeric variables, I utilized scatter plots, examining linearity as an indicator of association. On the sides are some examples of all the graphs I created.

Numeric > Numeric





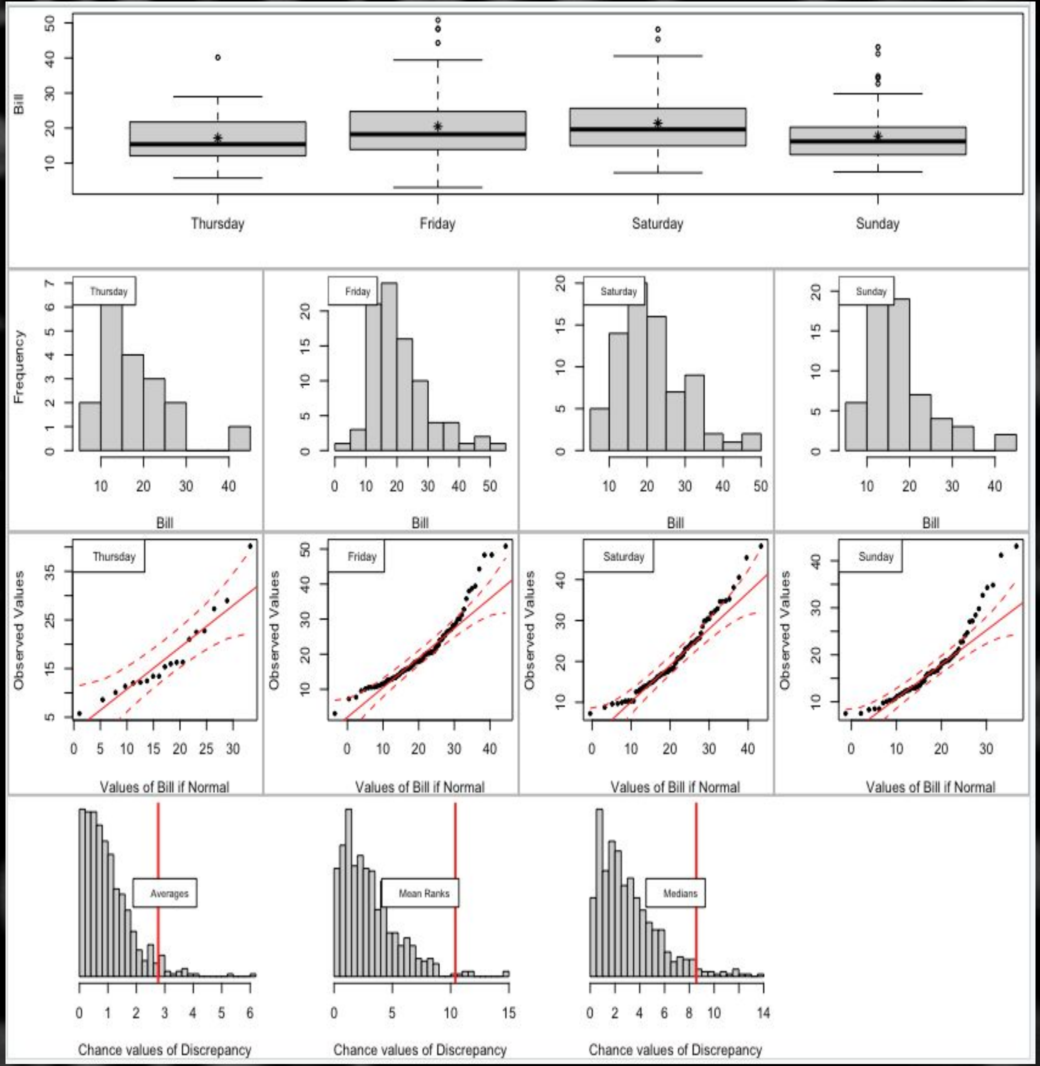
METHODS

Association between a Categorical variable and Numeric variable	Association between Numeric and Numeric	Linear Regression
I investigated the influence of gender, weekday, time, and smoker on the bill amount to identify statistical significance. I considered the association to be statistically significant if the p-values obtained were lower than the significance level of 95%. In cases where the distributions are normal, I used average values to determine statistical significance; otherwise, I relied on the median. values will be used.	I determined the association between Bill and PartySize. While the overall trajectory appears monotonic in the scatterplots, there is noticeable heteroscedasticity and outliers. Consequently, I opted to use Spearman's rank correlation to assess if the associations are statistically significant. The association will be deemed statistically significant if the p-value is less.	I conducted a simple linear regression to analyze the relationship between the explanatory variable, PartySize, and the response variable, Bill. The aim was to quantify if there is a significant linear association between the two variables. By regressing Bill against PartySize, I identified the best-fitting line that minimizes the difference between the dependent and independent variables' data points and predicted values. To assess the statistical significance of the simple linear regression, I tested the significance of the coefficients of the model. If the p-value of the slope coefficient is below 0.05, the model is considered statistically significant. With 95% confidence, I generated an interval for the slope and intercept. If the 95% confidence interval for the true slope does not include 0, the relationship is deemed statistically significant, indicating a connection between the predictor variable and its corresponding response. variable.

What were the Findings?

Significant or Insignificant?

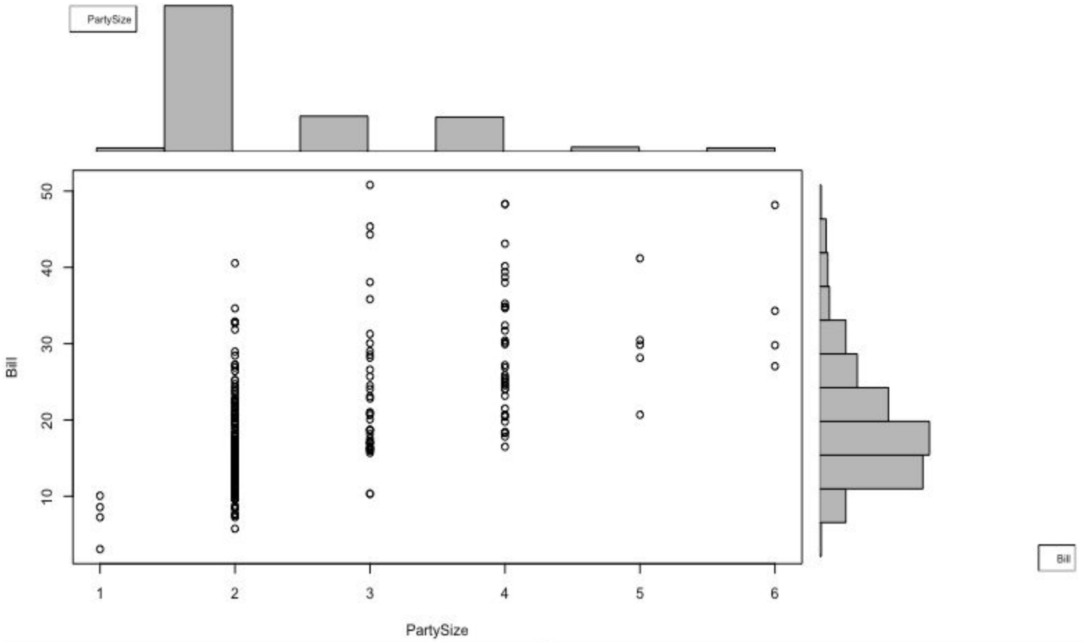
1.Association between a Categorical variable and Numeric variable

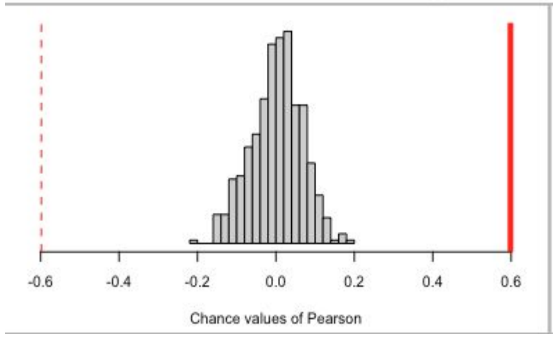


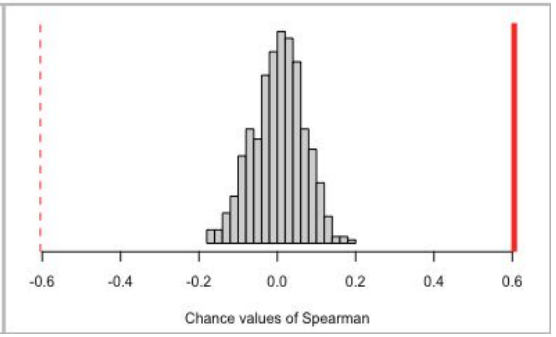
The distribution_Bill values based on each weekday as seen in the figure is skewed, we will use the median test to evaluate if the association is statistically significant. With the use of the Permutation Test (default permutations=500) we see the 95% Confidence interval of the p-value between 0.02 and 0.054. Since 0.05 is within this p-value range, the test is inconclusive.
After increasing the permutations to 2000, the p-value is between 0.029 and 0.046. Now, the p-values range is less than 0.05 so the association between Bill and Weekday is statistically significant.

2. Association between a Numeric variable and Numeric variable

The relationship between PartySize and Bill as seen on the scatterplot shows a slight positive linear growth, but there is more concentration on the PartySize 2 and less for a larger party size, so it is not consistent. Hence the Spearman's Rank correlation would be a better test to use for this.
The 95% confidence interval of the p-value is between 0 and 0.007, which is less than 0.05 and hence this relationship has statistical significance.



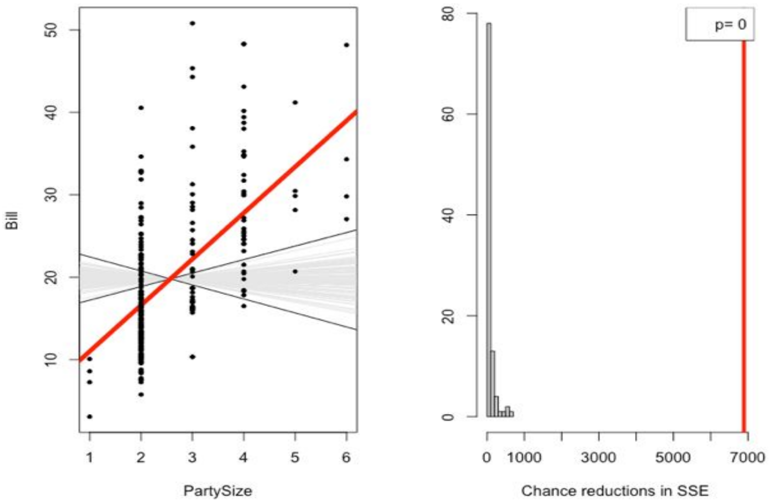


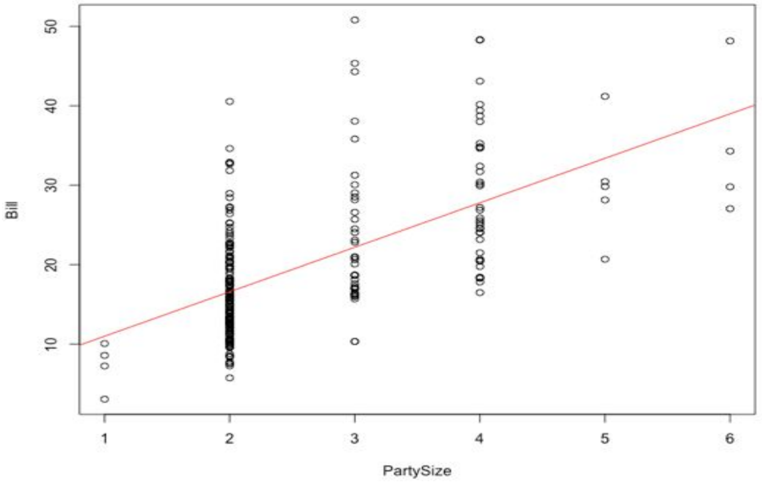


3. Linear Regression

	Estimated value	Standard Error	P-value	95% confidence interval
Intercept	5.3950	1.3207	4.085	0.00006
Weekday	5.6003	0.4821	11.616	<2e-16 ***

The linear regression analysis results indicate a significant positive relationship between the number of people in a party ("PartySize") and the bill amount ("Bills"). The intercept term suggests that when there are no people in the party (PartySize is zero), the expected bill amount is estimated to be $5.3950. For each additional person in the party, the bill amount increases by an average of $5.6003. The coefficient estimate for PartySize is highly statistically significant, with a t-value of 11.616 and a p-value of <2e-16. This indicates strong evidence against the null hypothesis of no relationship between PartySize and Bills. The low standard error of 0.4821 suggests relatively precise estimation of the coefficient. Overall, the results provide compelling evidence that the number of people in a party has a significant positive impact on the bill amount. It is important to note that there are some observations with relatively large positive and negative residuals, indicating potential deviations from the model's predictions.





Summary

The aim of the study was to understand how the several independent variables influenced the dependent variables, TipPercentage and Bill. From our analysis we have arrived at several conclusions:

- A statistically significant association was identified between the numeric vs categorical variables Bill and Weekday.

- For the analysis of numeric vs numeric variables, Bill and PartySize, a statistically significant association was also distinguished. By use of Spearman Rank Correlation test, a moderate correction was identified between Bill and PartySize.

- For the regression analysis, we examined the relationship between Bill and PartySize. This analysis also proved a significant association between the variables. A positive linear relationship was identified from the model.

THANK YOU!