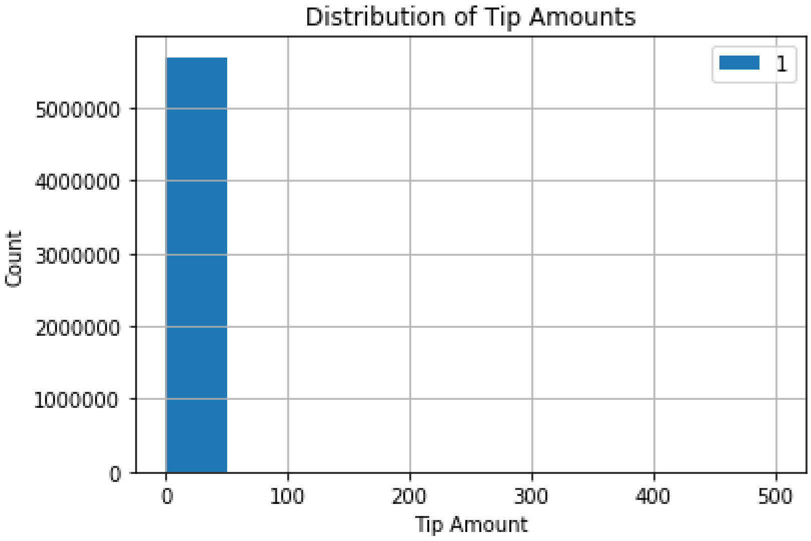
Capstone Project 1 Statistical Data Analysis

From our exploratory data analysis, we can see that there are several categorical variables across which we can perform some statistical analysis.

Looking at the different Rate Codes, there are 6 categories of codes. After plotting the distributions of the tip amount across the different rate codes, we can see that they are right skewed and not normal (as shown below), so we can't use a typical t test.



We can however use the Kruskal-Wallis H test to compare the distributions of various samples. Using an alpha of 0.05 and a null hypothesis that the distributions are equal, we can see from the kruskal test that the distribution of tip amounts for rate codes 3 and 4 are similar, and the distribution of tip amounts for rate codes 6 and 99 are also similar.

We can repeat the kruskal test for the various payment types, as well as the rain/snow conditions. From these tests we can see that the payment types 1 and 2 have a p value below the alpha of 0.05 and satisfy the null hypothesis. Payment types 3 and 4 have a $p>0.05$ and so we can conclude that their distributions are similar. The kruskal tests also shows us that the distributions for the rain/snow/clear conditions are not the same.

Looking at our categorical variables, we should perform a chi-square test to see if any of them are highly correlated. We should check for multicollinearity between our variables to help us in the following regression phase. Performing a chi2_contingency test between our categorical variables returns a $p < 0.05$ between all 3 pairs, suggesting that there is a statistically significant relationship between all 3 of our categorical variables. We should avoid using all of them in our model at the same time going forward.