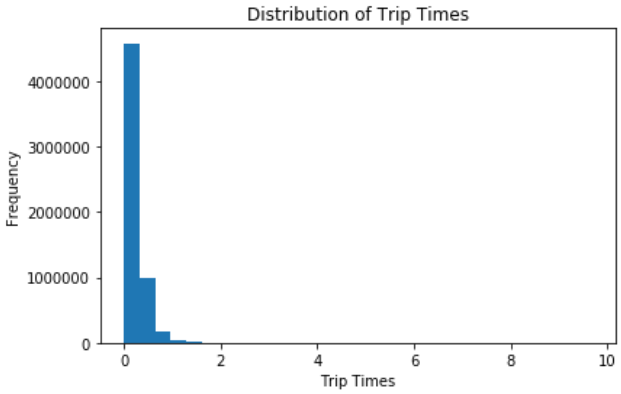
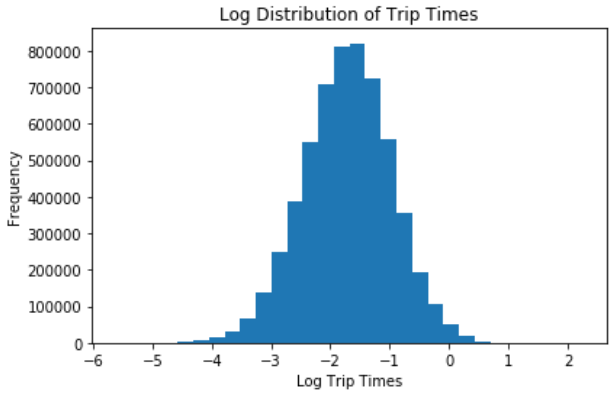
Capstone Project 1 Data Story

We would like to explore the relationships between various attributes of a taxi ride and the tip amount that it receives. This includes features like the trip time, trip distance, average speed, temperature, and visibility.

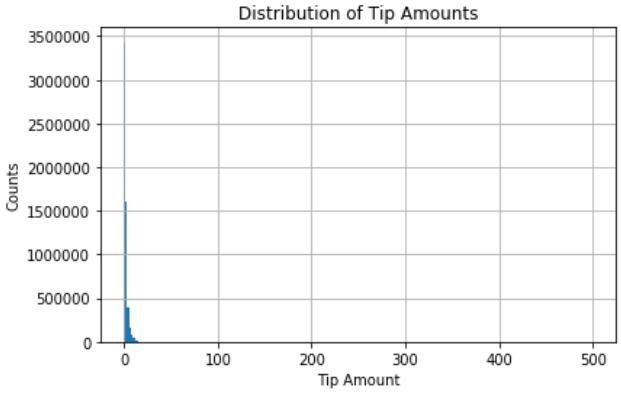
The first thing we do is create a correlation heat map to see if there are maybe some areas we should focus on and explore. From this heatmap we can see that there is a strong correlation between the trip time and the tip amount. It doesn't seem apparent that there is a strong relationship between any of the weather attributes and the tip amount though.

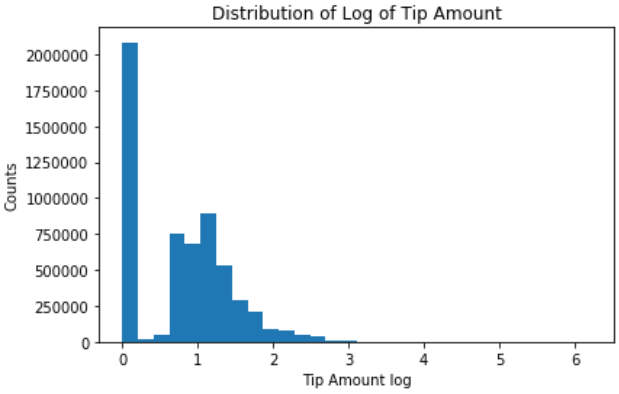
Looking at the trip times and distances, we can see that they are both right skewed in their distributions. Using a log transformation makes them look close to normal.



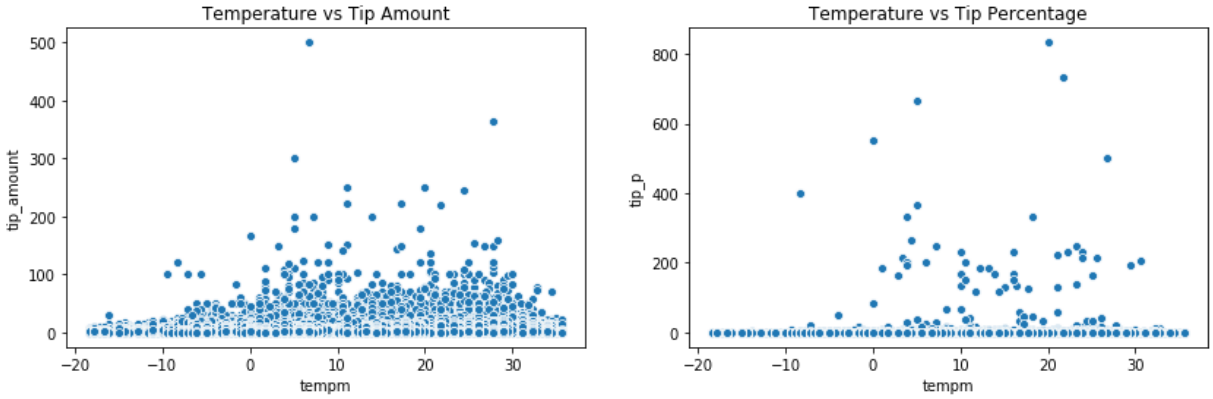


The trip amounts appear to be bimodal, with a large spike at 0 and a right skewed distributions. This intuitively makes sense since there would be a lot of trips with 0 tips, and the trips that do tip would have a range of amounts.

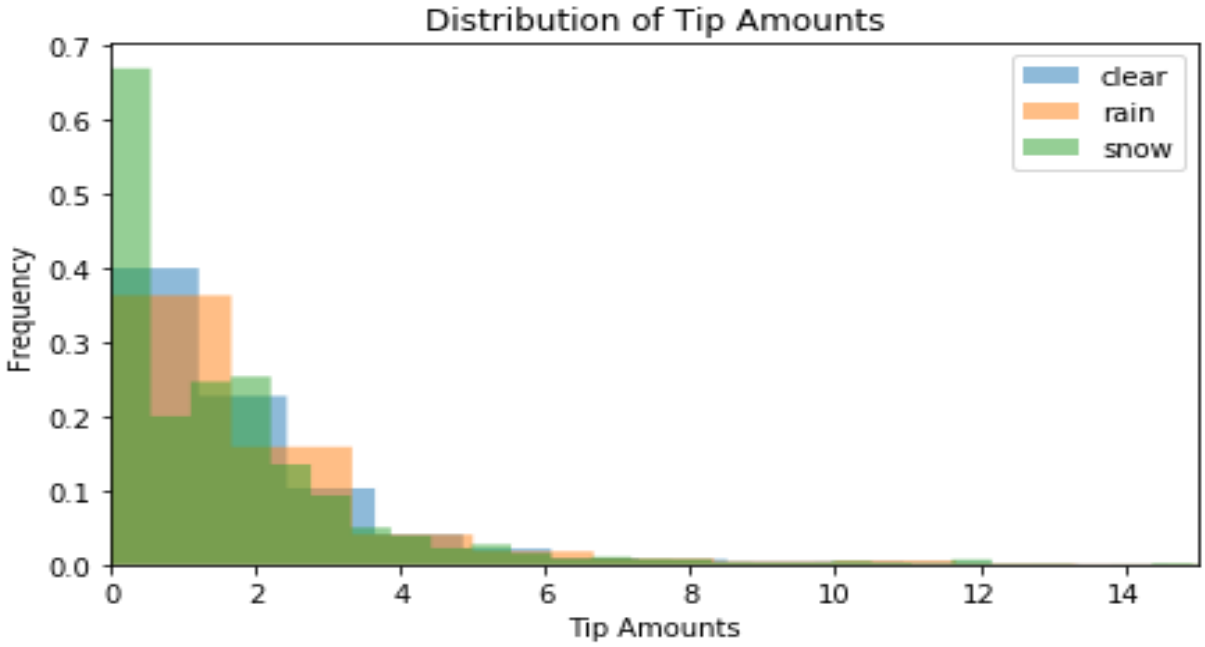




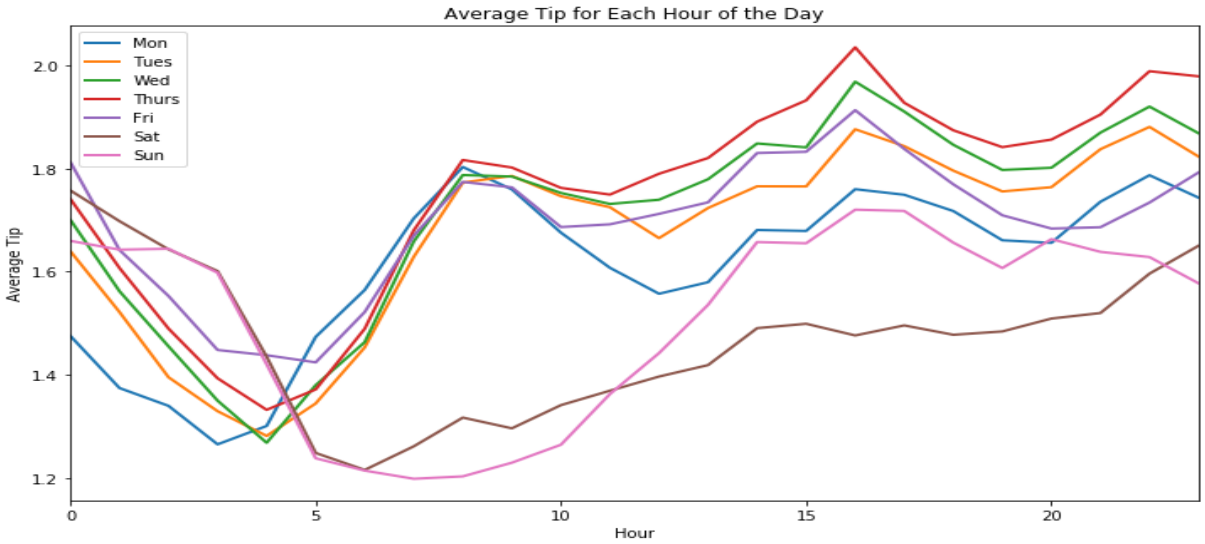
Next we can explore the various weather attributes in relation to the tip amount. For example we can plot the tip amounts vs the temperature and see that it higher tip amounts tend to happen during the warmer days.



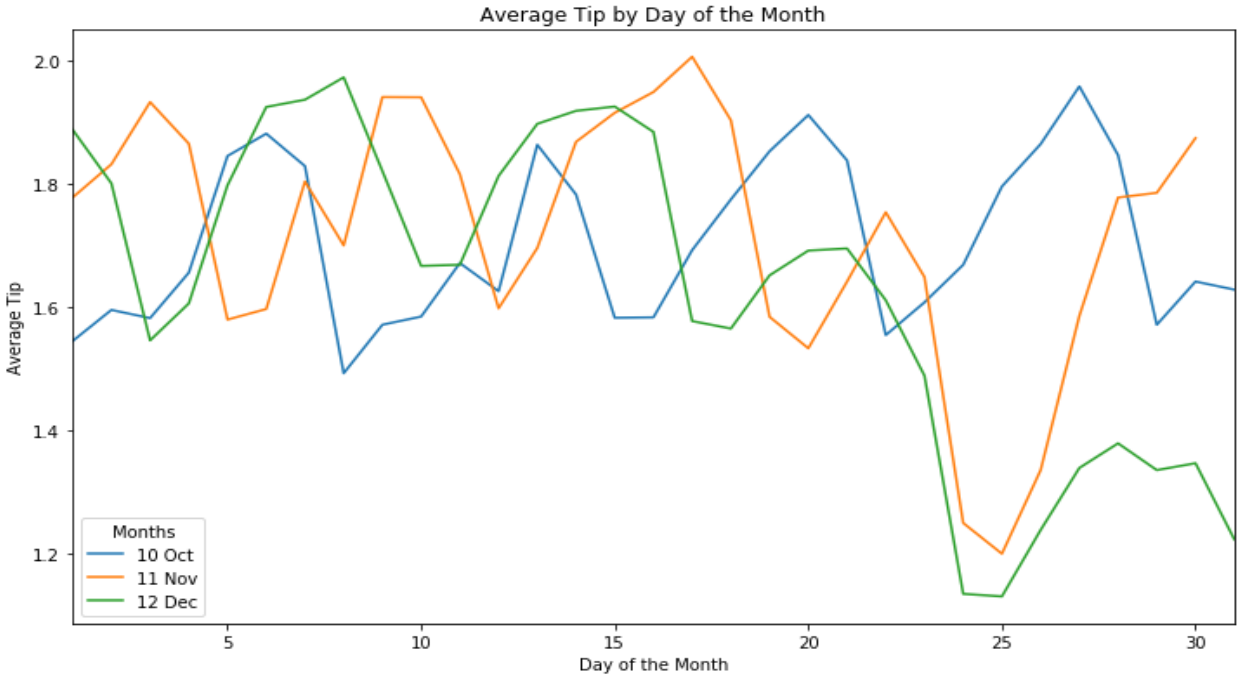
Plotting the distributions of tip amounts for rain, snow, and clear days suggests that the weather does change the distributions. The snow days seem to have the lowest mean tip amounts, and the clear days the highest.



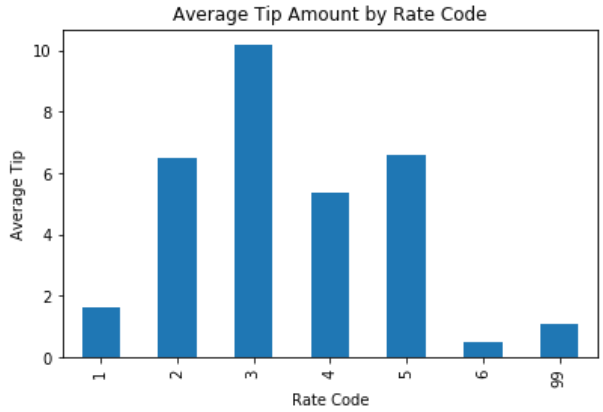
The day of the week seems to also have a clear effect on the amount of tips a taxi ride will get.

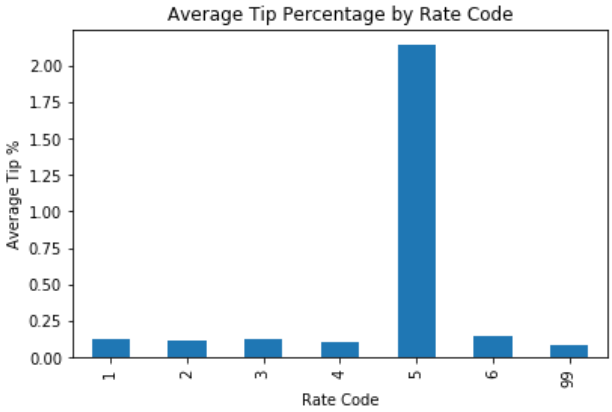


The day of the month seems to affect the average tip amounts as well. From the previous plot we can see that the weekends generate the lowest average tips. The weekend dips can be seen in the plots of the average tip by the day of the month. This also shows us that holidays affect the tip amounts as well. For example in the following plot, we can see that during thanksgiving (around nov 24) and christmas (dec 25) the tip amounts see a significant drop.

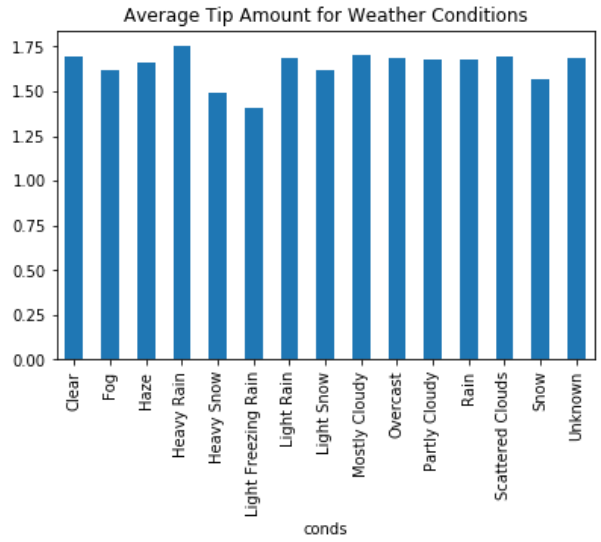


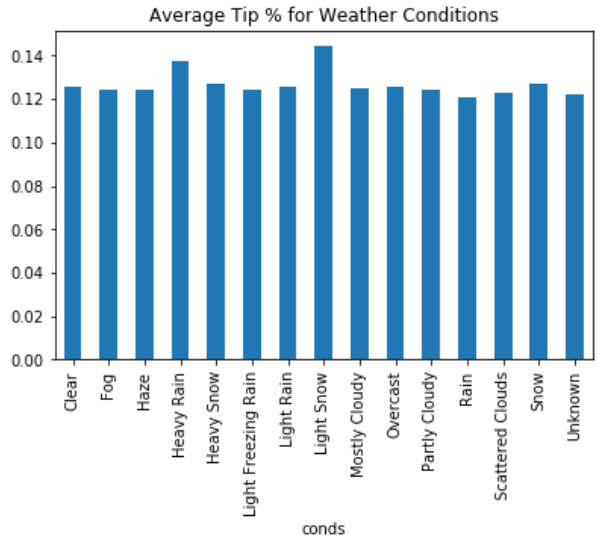
Lastly, we can see what the average tip amounts are across various categorical variables. For example, we can see what the average tip amount is for the different rate codes, as well as across all the weather conditions.





We can see that in general rate code 1 offers the lowest average tips, and rate code 3 the most. It's interesting to note that rate code 5 offers the highest tip by percentage.





Here we can see that rides during light freezing rain have the lowest average tip, followed by heavy snow.

From exploring this data, it's not entirely obvious which features have a large impact on the tip amounts. The weather does seem to play a role though, as seen in the temperature vs tips scatter plot, as well as the average tips across weather conditions. It's also seems clear that the date as well as the hour of the day has an impact. We can see from our exploration that the weekends and holidays have a lower average tip amount. This seems counterintuitive, but one possible theory could be that maybe people tip more when they are in a hurry during the work day and don't think about the amount too much.