Capstone Project

Predicting Taxi Cab Tip Amounts

How much does a drive get tipped for a ride?

IT DEPENDS

- Is it a weekday? weekend?
- Is it during the day? Night time?
- Is it a holiday?
- Is it raining or snowing?

These are just some of the possible factors that might influence the tip amounts

Why do we care?

Taxi driver benefits:

- Estimate additional income
- Know what the most profitable times / weather conditions to work in are

Ride sharing company benefits:

- Gain insight into customer spending habits
- Adjust rates based on tipping patterns

Data Wrangling

Data is sourced from the following

2016 NYC Yellow Taxi Cab
https://data.cityofnewyork.us/Transportation/2016-Yellow-Taxi-Trip-Data/k67s-dv2t

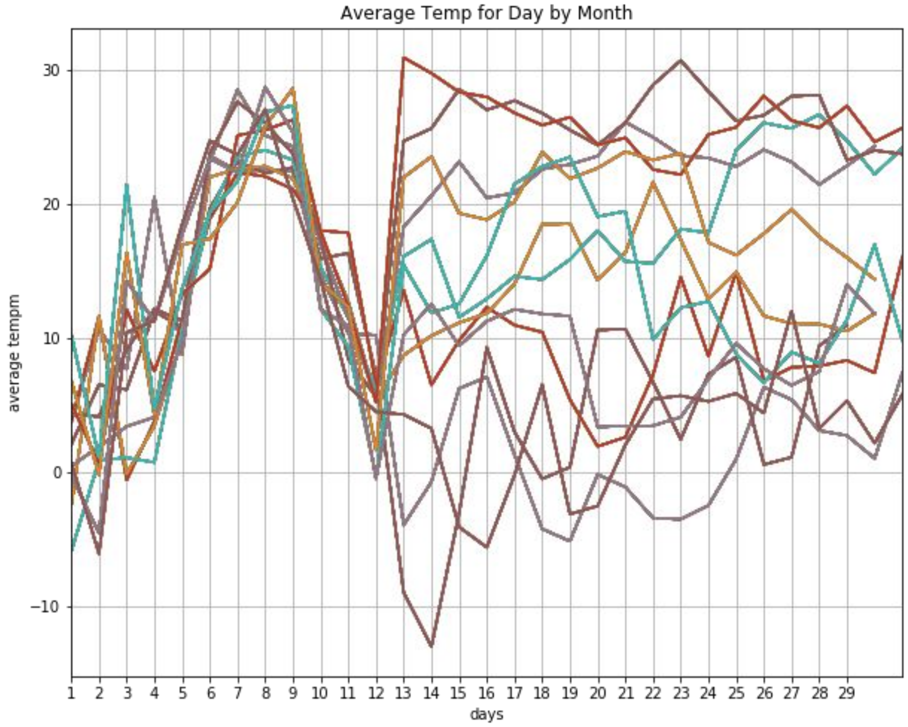
2016 Hourly Weather in New York
https://www.kaggle.com/meinertsen/new-york-city-taxi-trip-hourly-weather-data/download

Let’s take a look at our data

Incorrect Dates

Days and months have been reversed up to day 12 of the month

Uniform weather for first 12 days of all months doesn't make sense.

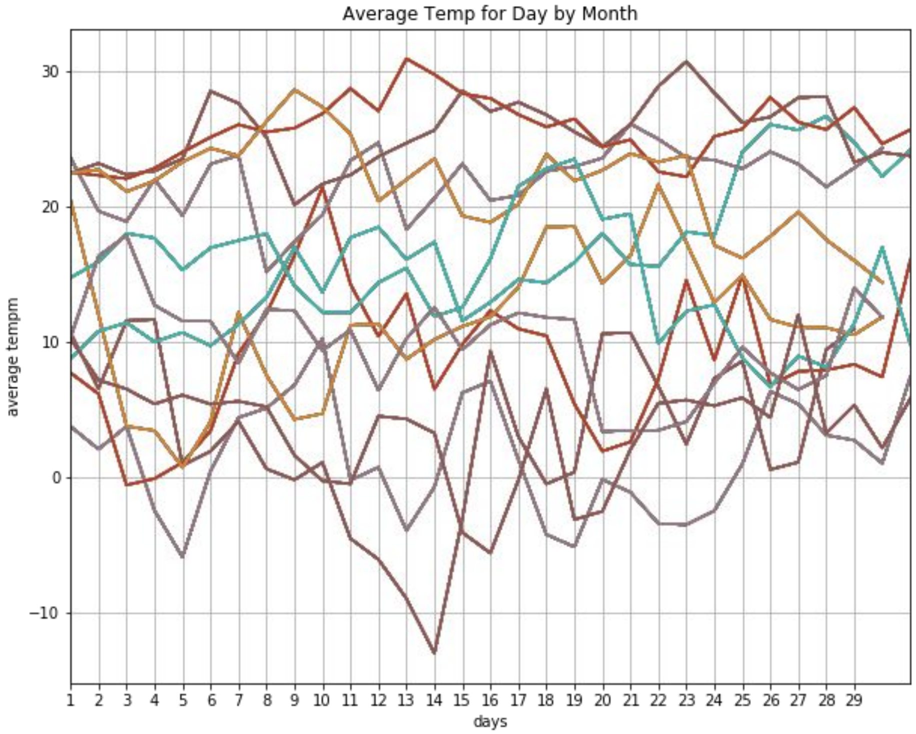


Fixed Dates

Switch the day and the month if the day is less than 12.

Avg temperatures now spread out for different months

This makes more sense.



Merge, Sample, Clean

Merge weather data with taxi data by closest datetime

Impractical to use all 60M rows of data

Take a stratified sample of 10% (6M rows)

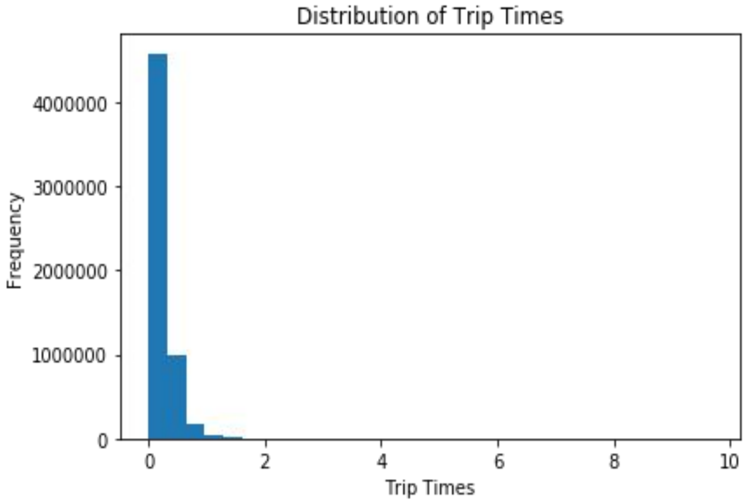
- Group by weather condition: rain, snow, clear

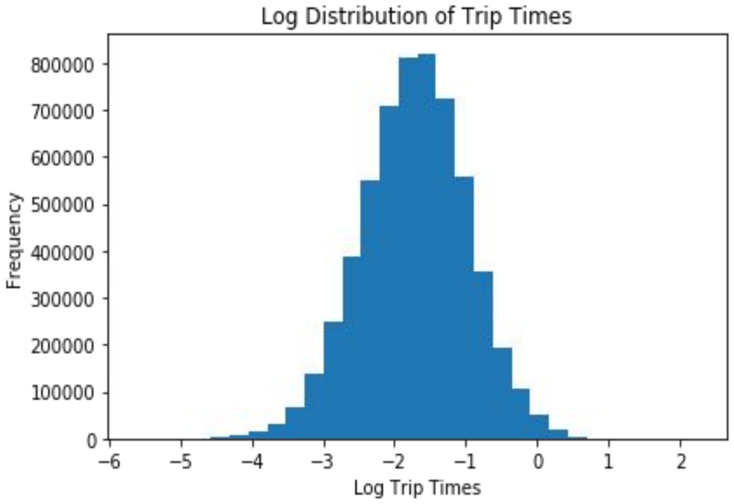
Exclude entries where the trip time or distance is 0

Delete unused columns

Trip Time Distribution

Most of the taxi rides are short trips, right skewed distribution

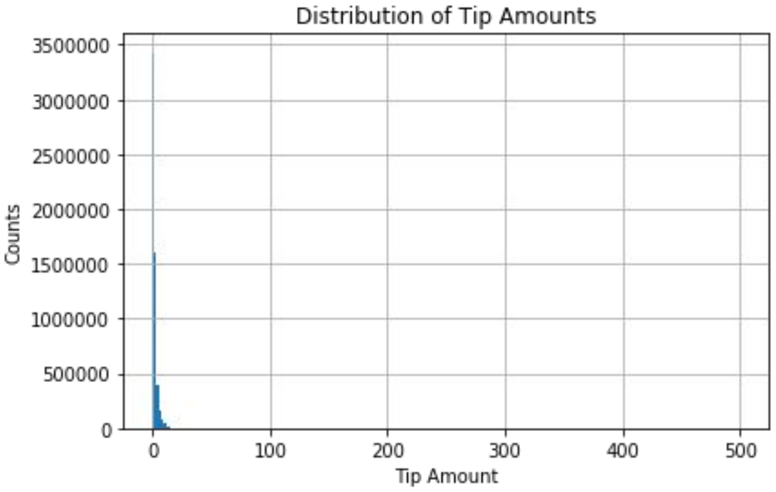


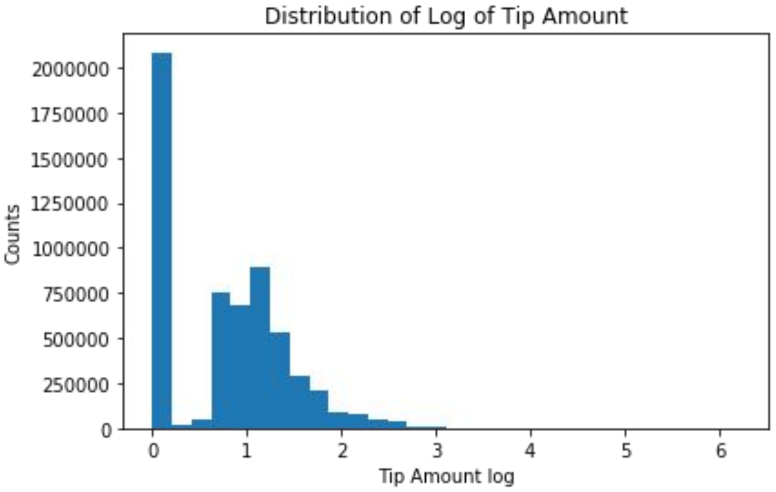


Tip Amount Distribution

Tip amounts are extremely right skewed

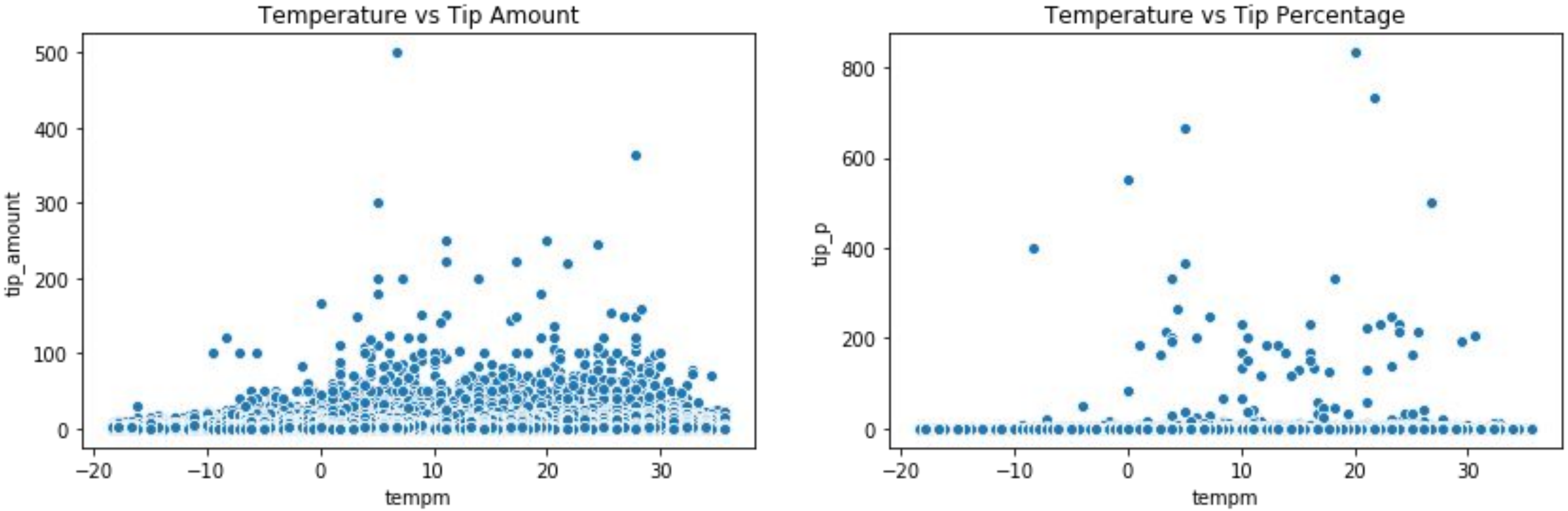
Log distribution shows a clearer picture. Looks bimodal





Temperature vs Tip Amount

Higher tips looks like they occur at warmer temperatures



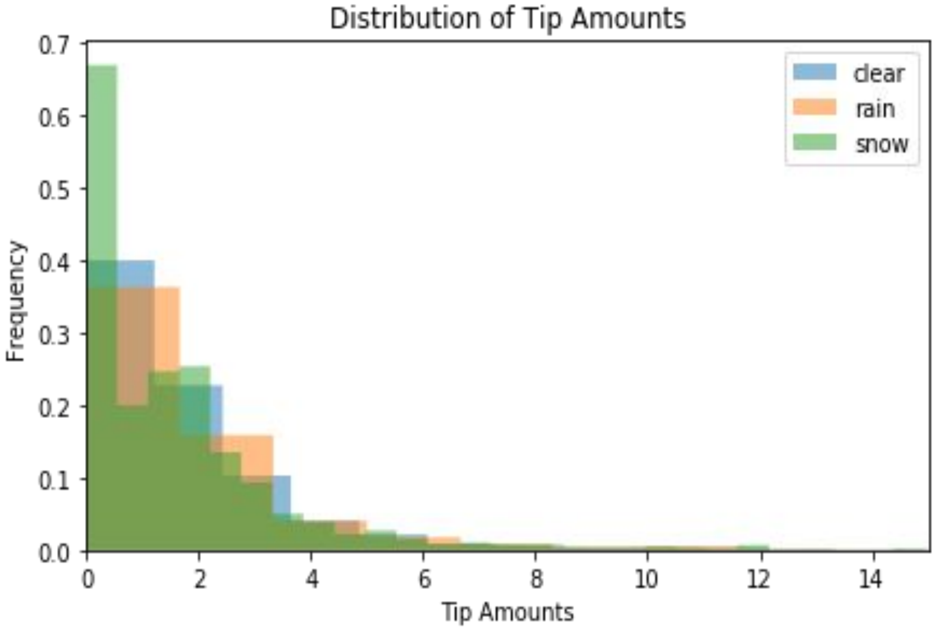
Tip Amount Distributions by Weather

Distributions do not seem to differ drastically

Mean Tip Amounts by Weather

- clear: 1.69
- rain: 1.68
- snow: 1.60

Average tips lower when it's snowing

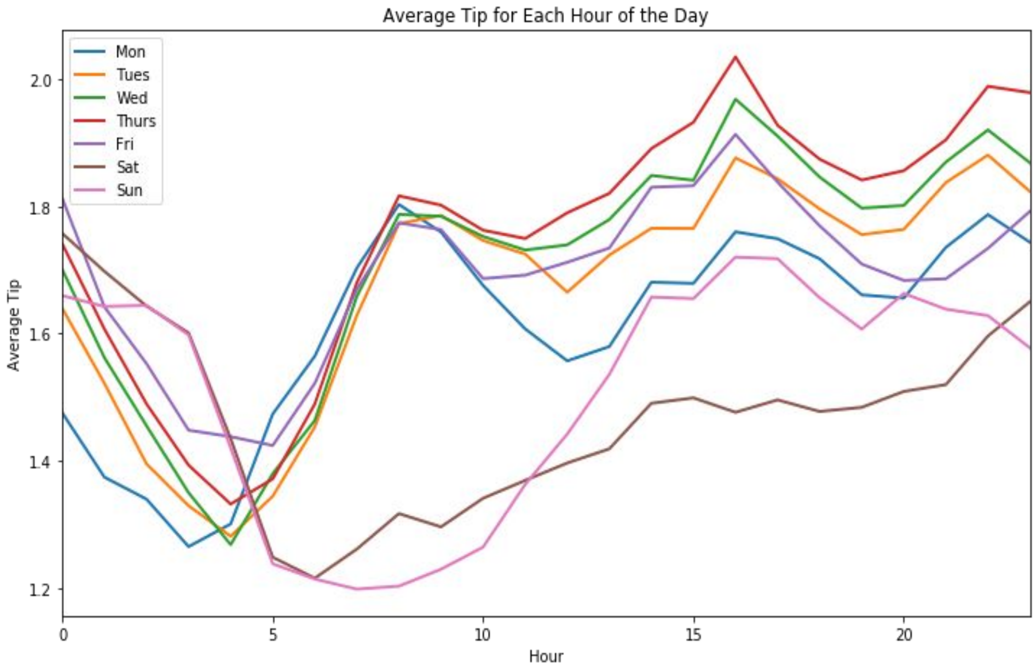


Tips by Hour of the Day

Saturday and Sunday morning provide lowest average tips

Higher during the 9-5 workday

Lower late at night / early morning

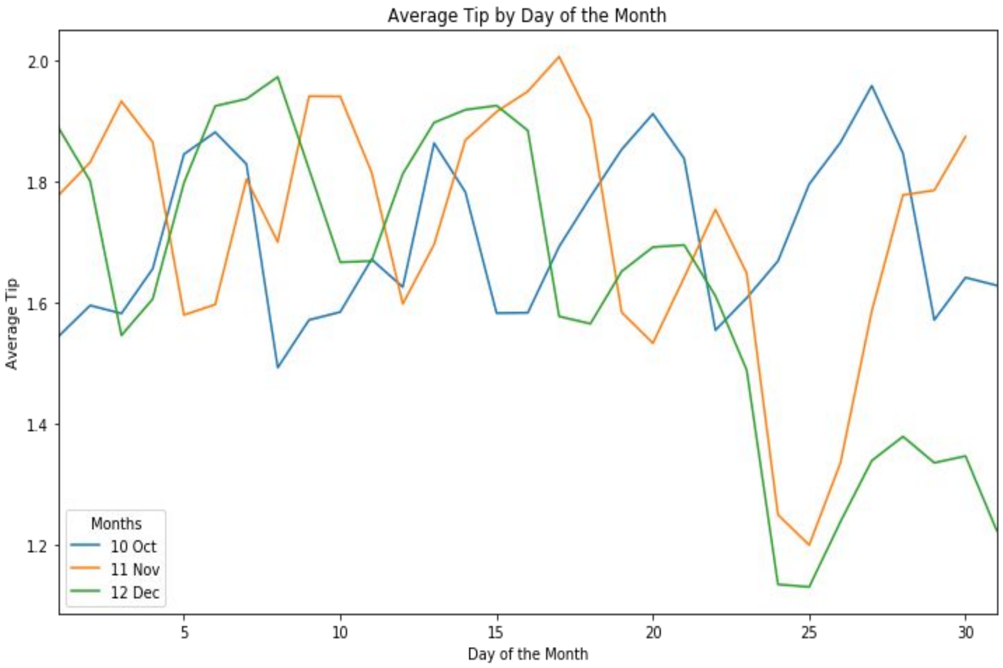


Tips by Day of the Month

Avg tips lower during the major American holidays

For example,

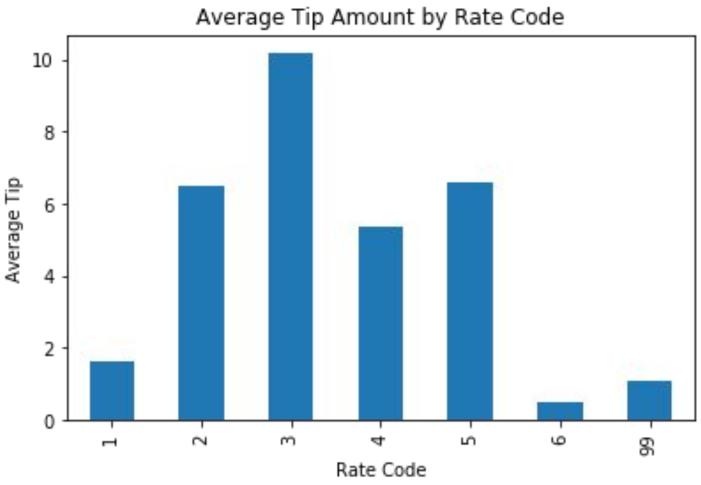
- Nov 24, Thanksgiving
- Dec 25, Christmas

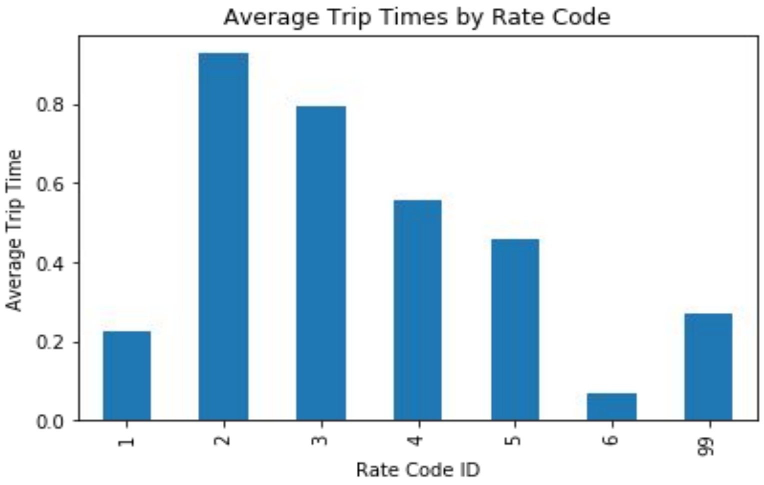


Tip Amounts by Rate Code

1= Standard rate, 2=JFK, 3=Newark, 4=Nassau or Westchester, 5=Negotiated fare, 6=Group ride, 99=Unknown

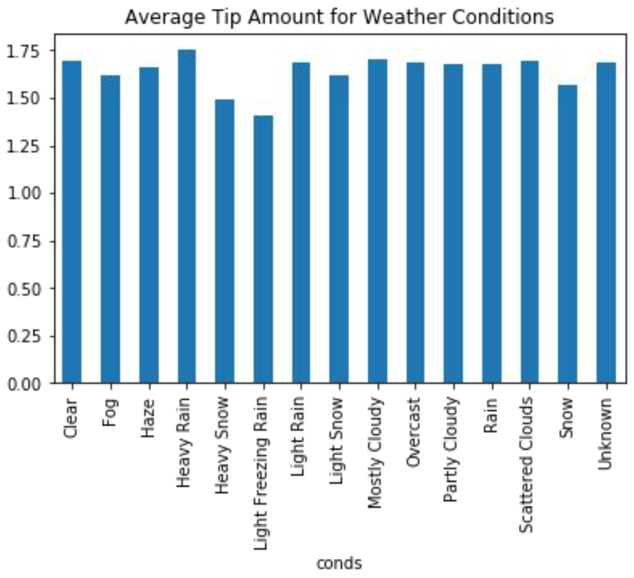
Trips with rate code 1 generated lower average tips and trip times

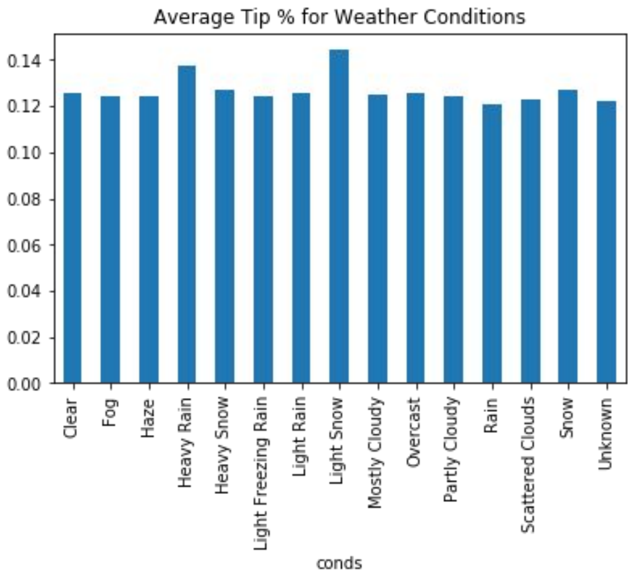




Average Tip Across Weather Conditions

Lowest average tip during Light Freezing Rain





Statistical Analysis

Perform Kruskal-Wallis test to compare distributions across categorical variables

- Tip amount distributions for Ratecode 3 and 4 are similar, as well as 6 and 99
- Distributions for payment types 3 and 4 are similar
- Rain / snow / clear tip amount distributions are different

Use Chi-squared test on categorical variables

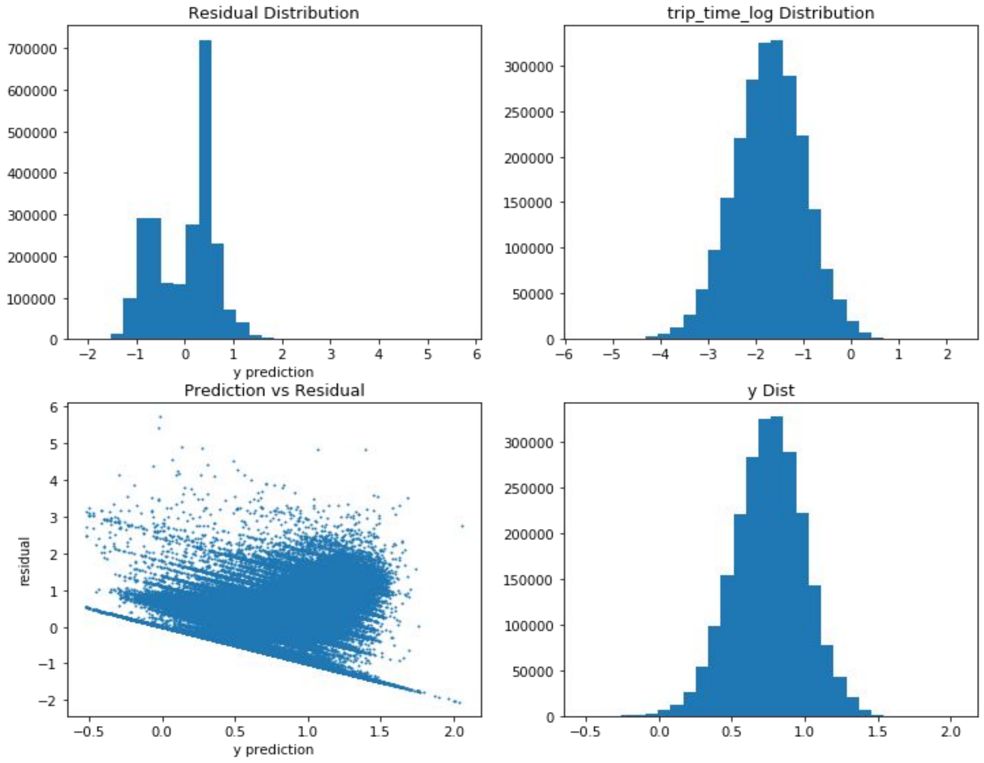
- Statistically significant relationship between all 3 variables
- To avoid multicollinearity, avoid using all 3 features in our predictor model

Linear Regression

Test out some simple linear regression models using only 1 feature variable

Bottom left plot has a clear pattern with a line at the bottom.

Indicates heteroscedasticity



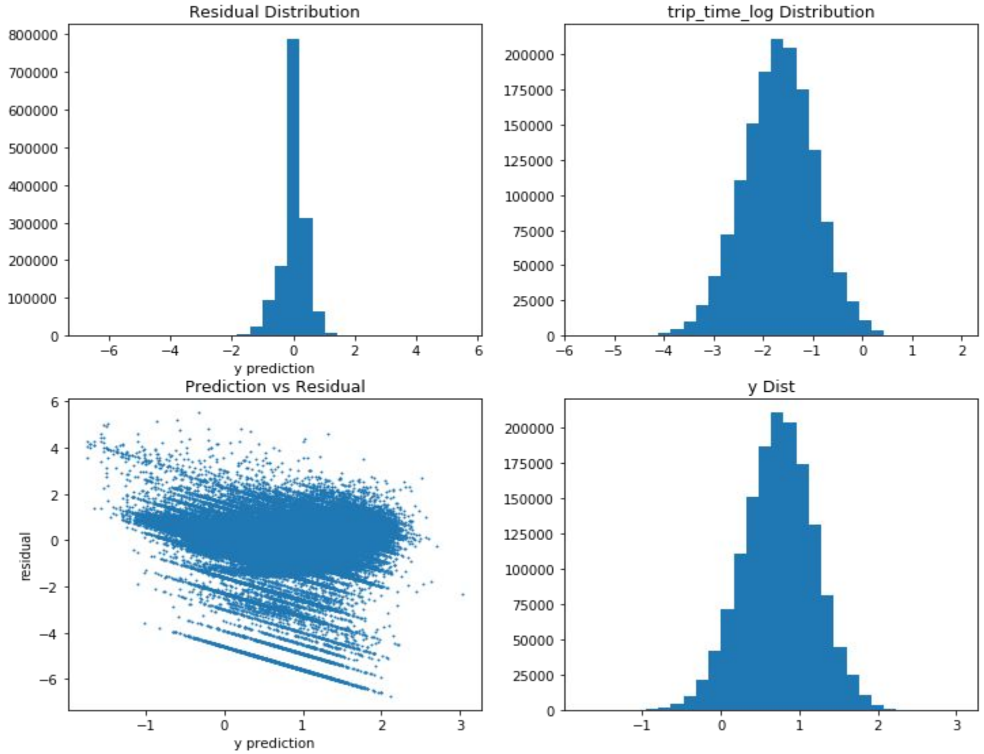
Linear Regression Take 2

Try again, but exclude samples where tip_amount = 0

Line at the bottom of the lower left plot is not as pronounced, but heteroscedasticity still there

Maybe valid linear regression model not feasible.

Let's try Random Forest Regressor



Random Forest Regressor

Create test run using all of the available features

- In general would take too long to train for a useful model

Rank the features by importance

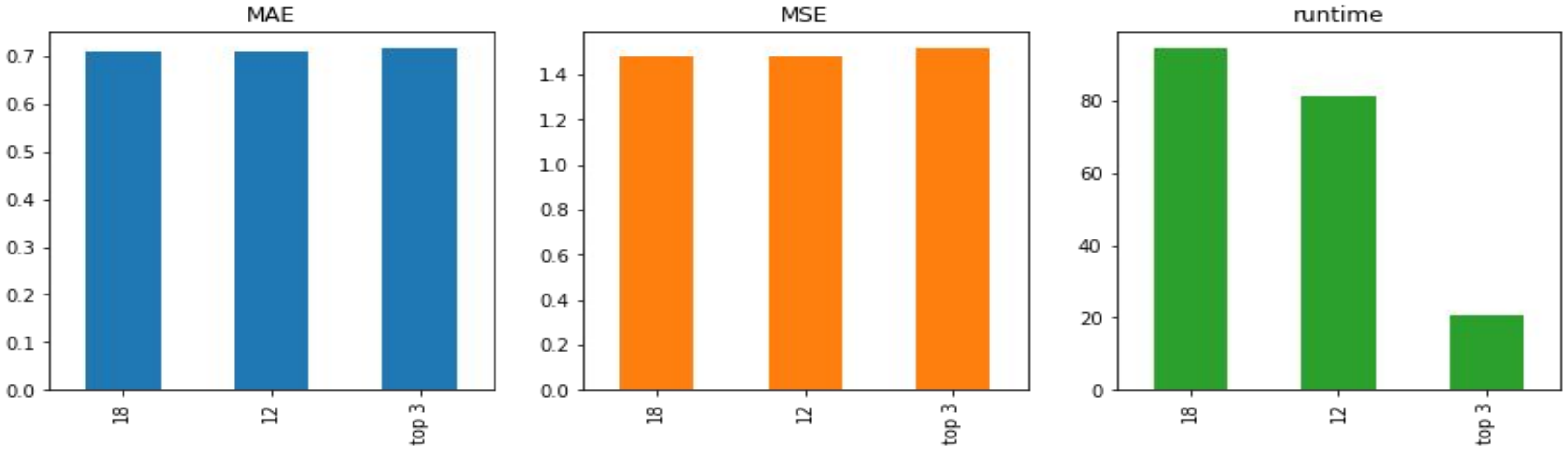
Top 3 important features

- fare_amount_log = 0.57
- avg_speed_log = 0.07
- trip_time_log = 0.07

Feature Selection

Test 3 different models using various combinations of the top features

MAE, MSE are close. Model using top 3 features has quickest runtime.



Metrics

Use Mean Absolute Error, and Mean Squared Error as our main metrics

MAE 0.716

MSE 1.516

Data appears to consist of a wide variety of tip amounts and outliers

No reason for large errors to be penalized extra

- Mainly focus on MAE

Parameter Tuning

Use GridSearchCV to test out combinations of parameters

Let's use 'neg_mean_absolute_error' for scoring

Best parameters:

- N_estimators = 100
- Max_features = log2

Using max_features = sqrt gives us a close score, but longer run time.

Go with max_features = log2

Random Forest Regressor New and Improved

Run random forest regressor with new parameters:

Old Model	**New Model**
MAE 0.716	MAE 0.712
MSE 1.516	MSE 1.522
20min 52s	11min 26s

New model is better with a lower MAE, and quicker run time

Conclusion

Used Random Forest Regression to create a model to predict the tip amount

Insights:

- Weather had minimal impact in predicting tip amounts
- Date and time also minimal importance
- Fare amount, trip time, and average speed are the most important features

Ways to improve model:

- Include pickup and drop off location data
- Whether or not the taxi was hailed on the street or called for