Title: Data Wrangling Summary

Data Aggregation: Data was collected from the NYC Taxi and Limousine Commission for the year of 2017. The data was divided by month, with each month containing ~9.5 million records. To aggregate my data, here are the steps I took:

1. For each month, I filtered the records to include regular cab rides only (not airport rides or group rides for example).
2. I sorted each month's data by pickup date/time and then chose every 400th element to generate a representative sample.
3. I concatenated the records for all months together into a single DataFrame and saved the data as a CSV file: ~265 K records.

Data Wrangling: The aggregated dataset was complete (no missing values), but there was a lot of data that needed to be cleaned/removed. Here are the measures I took.

1. Explored the dataset thoroughly first.
2. Removed unnecessary columns (pickup and dropoff location, passenger count).
3. Placed logical and quantitative thresholds for key features such as trip duration (2 minutes to 2 hours), trip distance (0.2 miles to 60 miles), avg speed (2mph to 60 mph), etc. Records outside these thresholds were deleted.
4. Cleaned negative values and outliers, by removing them or inverting their signs.
5. Compared quantitative features against each other (distance, duration, fare amount) and ensured they made sense relative to each other. Extreme outliers were handled either by using thresholds, or by removing the most extreme 0.25% of records at each end of the data.
6. Created new features from the existing dataset such as hour (numerical) and weekday (categorical).

Here are the Jupyter Notebooks for Data Aggregation as well as for the NYC Fare Prediction Project respectively.