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. **Date Range Validation**: Since we were working with 2017 rides specifically, I checked to see if we had any trips outside the specified range. A few records were found and removed.
2. **Invalid Trip Durations**: Here, I dealt with unrealistic/invalid (negative, very low, very high) trip durations. I removed outliers through manual inspection as well as by placing limits on the min and max allowed durations [2min, 1 hour]. This decision was made quantitatively and intuitively.
3. **Trip Distance and Fare Amount**: Similar to above, I dealt with outliers through visualization, manual inspection, and by placing logical limits on the fare amounts and trip distances.
4. **Categorizing Data**: Most of my remaining data was categorizable, so I inspected/cleaned any errors and casted the data to type category (for operational efficiency).
5. **Adding New Features**: I added extra features such as weekday, hour of day, trip duration, and avg_speed. These were calculated using existing data and will be helpful in the EDA process.
6. **Cleaning Disproportionate Data**: Finally, I checked certain numeric columns against each other (distance vs duration, distance vs fare amount, etc) to see if there was any disproportionate data. I visualized my data through scatter plots and removed outliers manually as well as by placing logical limits on certain values (avg speed for example cannot be > 100 mph).

Here are the Jupyter Notebooks for Data Aggregation and Data Wrangling respectively.