**Capstone Project - Data Wrangling**

My capstone project is a machine learning model for revenue prediction for individual rideshare drivers who chose to work at a given time of day in the city of Chicago. The Chicago Data Portal has a dataset of 45.3 million rideshare rides that occurred between November 1, 2018 and April 30, 2019. Each row of the data is a trip. An initial sample of 100,000 trips was downloaded into a data frame for initial investigation of the data cleaning actions required. A two-step approach was applied to this dataset. First, the NaN values in each column were identified, assessed and rows were deleted as required. Secondly, the clean dataset was assessed to identify rows that had outliers or unrealistic values.

Prior to discussing the cleaning steps, the following three pieces of information from the data portal should be noted:

1. It was stated on the data portal that trips that start outside of Chicago have NaN values for their pickup locations and trips that end outside of Chicago will have NaN values for their drop off locations.
2. A spreadsheet called “CensusTractsTIGER2010” was obtained from the data portal. This sheet maps 800 census tracts to 76 community areas that have clear boundaries shown on a city map. This sheet also maps centroid locations to the census tracts.
3. The time stamp values are rounded to the nearest 15 minutes for both start and end time stamps. The trip duration indicated in the trip\_seconds column is the exact trip time.

Using .info() and .isna().sum() functions, the number of NaN values in each column were readily counted. These functions were used multiple times throughout the analysis of the columns so that the decrease in NaN values could be readily observed as operations were performed on the data.

The three time columns, ‘trip\_start\_timestamp’, ‘trip\_end\_timestamp’ and ‘trip\_seconds’, were analyzed first. There were NaN values in the trip\_seconds column. It was observed that the end time occurred before the start time for all of the trip\_seconds values that were NaN. All start and end time stamps occurred on November 4 between midnight and 2:00 AM. As these errors only affected less than 30 of 100,000 trips and occurred in a very small time range, a decision was made to delete these rows from the main dataset data frame. The timestamp columns were converted to proper datetime datatype and trip\_seconds was converted to minutes and renamed as trip\_minutes to be better aligned with the time stamps. The dataframe was also indexed based on ascending trip\_start\_timestamp so that the trips are in chronological order.

The ten location columns,‘dropoff\_census\_tract', 'dropoff\_community\_area', 'dropoff\_centroid\_latitude', 'dropoff\_centroid\_longitude', 'dropoff\_centroid\_location', 'pickup\_census\_tract', 'pickup\_community\_area', 'pickup\_centroid\_latitude', 'pickup\_centroid\_longitude', 'pickup\_centroid\_location', were analyzed next.

A data frame of the five drop off columns was created which contained all the trips for which all five drop off columns were NaN. These trips were subtracted from the main dataset dataframe. After this subtraction occurred, a dataframe of the five pickup columns was created which contained all the trips for which all five pick up columns were NaN. These trips were subtracted from the main dataset dataframe. It was then observed that all six centroid columns no longer had any NaN values. These two operations got rid of all trips that end outside of Chicago and all trips that start outside of Chicago. Those trips have to be omitted from the revenue model as the trip occurring after a trip out of Chicago or the trip occurring before a trip into Chicago for a given driver can not be predicted because the driver’s location is not known.

It was then observed that the pickup\_community\_area and dropoff\_community\_area both had a small number of NaN values and that the pickup\_census\_tract and dropoff\_census\_tract columns both contained in excess of 10,000 NaN values. The next step was to determine if the community area columns that were NaN could be assigned to a community area using data from the census\_tract or centroid columns. It turned out that the census tract columns contained data for all rows where the pick up or drop off community area was NaN. However, in all such cases, the corresponding census tract number was 17031770700. This number isn’t mapped to a community area in the spreadsheet. The centroid latitude and longitudes in rows with a census tract of 17031770700 were also not found in the spreadsheet. Thus, all of these trips were deleted as they could not be mapped to a community area.

Next, the columns for dropoff\_census\_tract and pickup\_census\_tract were deleted completely. The reason for this deletion is that the community area columns for both pickup and drop off contain data to show where trip started and ended. It will be the 76 community areas used for locations in the model rather than 800 census tracts as the number of community areas is less by a factor of 10 and is more reasonable to use for such a model. The boundaries of the community areas and where they are in relation to one another is also shown on a geographic map in the data portal.

There now only exists three NaN values in the main dataset data frame. There is one NaN value in each of the fare, additional\_charges and trip\_total columns. It is suspected that all three NaN values are associated with a single trip. The trip with fare value of NaN was identified and its additional charges and trip\_total were found to be NaN. This trip was deleted. There are now no more NaNs in the dataset. The number of trips has been reduced by 14% to 85,970.

The second step was to assess values in the “NaN-free” data frame to determine if there were any outliers or inconsistencies in the data. The following observations were made:

* The date range of the 100,000 trips dataset covers from November 1, 2018 to December 31, 2018 whereas the total dataset online ranges from November 1, 2018 to April 30, 2019.
* Some trips were found to have a value of zero reported in the time\_seconds column, trip\_miles column and fare column.
* Most trips have a 15 minute difference between start time stamp and end time stamp. The mean of the time stamp differences is 15:55. The 25th, 50th and 75th percentile values are all 15 minutes.
* 99.55% of the trips have a timestamp difference of 1 hour or less.

How to treat the zero value fields and which trip durations should possibly be deleted from the dataset will be determined when exploratory data analysis is performed on a full dataset and trends in the data are known and understood. No rows/trips were deleted or modified based on the second step of the analysis.