Lisa Taylor
Capstone 1: Milestone Report

This report documents the progress I have made analyzing the Chicago Rideshare dataset.

Problem Statement

For this project, I have taken the perspective of a City Transportation Agency applying the Rideshare dataset to gain insight on travel patterns within the City. Understanding how rideshare customers are utilizing the service provides opportunity to target improvements in public transit offerings, such as making enhancements to bus/train routes and schedules. Insight into rideshare patterns also provides information that can be applied to infrastructure planning and construction projects.

The primary questions that I am trying to answer with the rideshare dataset are:

- Can ride demand be predicted by census tract for a given day of week and time of day?
- Are there clusters of ride patterns that could be potentially better served by a targeted transit option, such as a shuttle bus?

In addition to the City Transportation Agency, other potential clients for this type of analysis may include drivers, who could apply demand predictions to position themselves in areas where they are most likely to pick up rides at a given time. In addition, environmental agencies and citizen environmental groups concerned with minimizing air emissions and traffic impacts associated with excess driving, such as when rideshare drivers circulate while waiting for fares, or when large numbers of riders choose to travel by rideshare in the absence of other options.

Description of the Dataset

Since November 2018, rideshare providers (Lyft, Uber, etc.) operating in Chicago are required by ordinance to submit periodic data reports with basic rideshare information. The main 'Trips' dataset includes information about individual rides. Additional data is available with details on the drivers and their vehicles. The Trips and driver data are anonymized, so trip starting and ending locations are generalized to the nearest census tract, and drivers cannot be linked to particular rides they provided. The driver data has not been applied to this analysis.

Data Preparation

The Chicago Trips dataset was downloaded as a csv file from the City of Chicago Data Portal. Since the data file is very large (12 GB), I decided to perform the initial exploratory data analysis from a sub-sample of the entire dataset. This was accomplished by setting the "chunksize" option of the pandas read_csv function to create an iterator of the csv file that reads in 20,000-line increments, then sampling 5% of each chunk into sample dataframe.

Some additional wrangling steps I performed on the data included:

- Creating datetime fields with the ride start and end datetimes rounded to the nearest hour
- Removing rides with zero distance or zero fare
- Creating additional fields for day of week, month, year, and hour
- Aggregating the dataset by hour and census tract and calculating average fare, average distance, average duration, and frequency counts

This rideshare dataset contains basic information about each rideshare trip: starting and ending census tract, starting and ending Community Area[1], starting and ending time, whether it was a pooled trip, and information about fares and tips. Times are rounded to the nearest 15 minutes. Fares are rounded to the nearest $2.50 and tips are rounded to the nearest $1.00. The census tract data were not available for certain rides (census tracts outside the city boundary or with low frequency of usage) and those ride records were dropped from the dataset.

To enrich the rideshare dataset, I merged in additional data resources having the potential to enhance prediction of trip demand and fares. First, I downloaded and merged in a weather dataset from NOAA containing hourly and daily aggregated measurements of temperature and precipitation. To add information on population characteristics, I used the Census API to download median income and population density per census tract. I also downloaded the census tract and Community Area boundary geojson files from the Chicago Data Portal and loaded them into a geodataframes using the geopandas python library. Analysis results may now be mapped by census tract or Community Area.

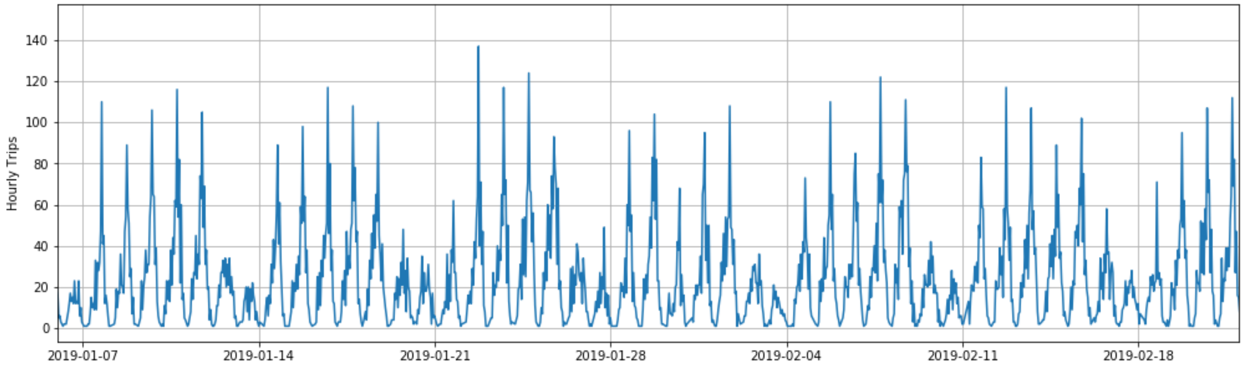
By joining the census geodataframe to the aggregated rideshare and census data I was able to and derive additional fields including population density and distance from downtown Chicago.

One census tract appeared to have an unusually high population density. To check if this was an error, I located the tract in google maps and confirmed that it corresponded to a census tract covering a very small area but housing a series of 40-story residential towers. Given this construction, the population density seems reasonable for this tract.

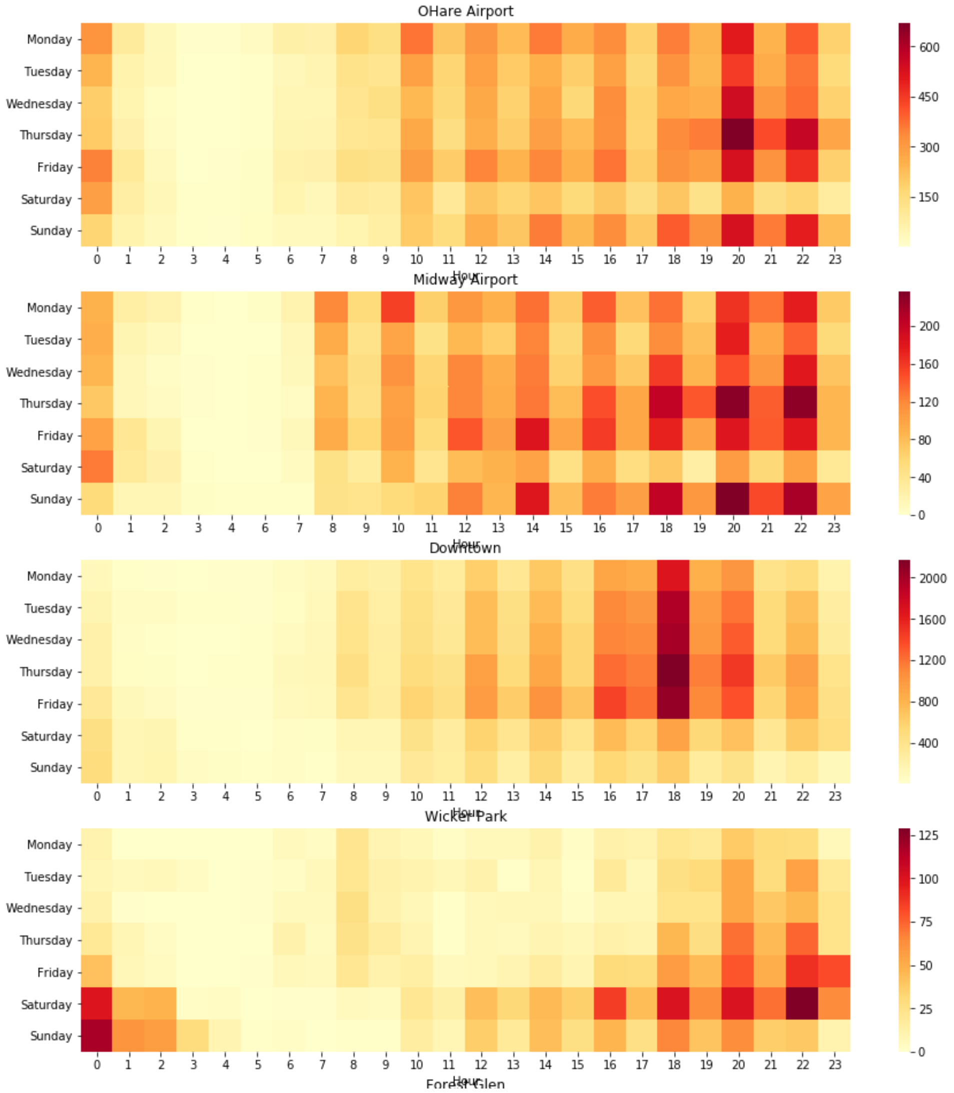
Major Findings

By plotting and mapping the enhanced rideshare dataset I was able to make some preliminary findings:

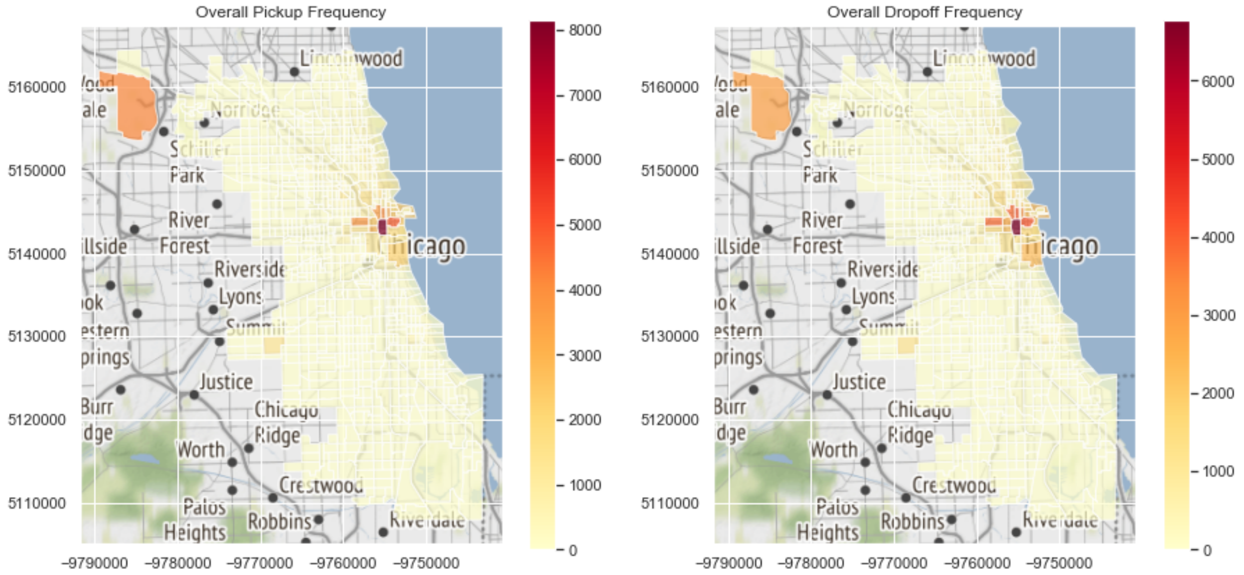
1) Rideshare usage is cyclical, with a usage pattern demonstrating both daily and hourly components. The following plot shows systemwide usage for several weeks in January and February 2019, with vertical gridlines indicating midnight on Sundays.



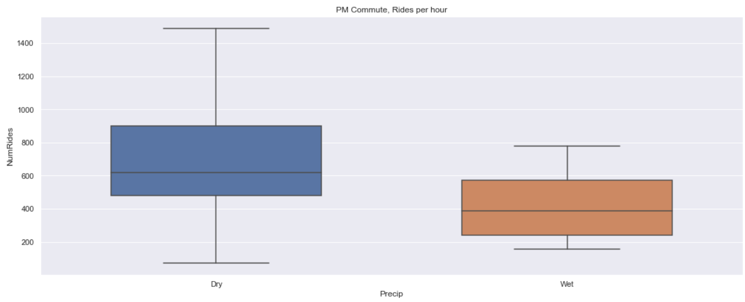
2) Rideshare usage patterns vary across the city. The data demonstrates distinct patterns for downtown, airports, residential neighborhoods, etc. The following heatmaps illustrate usage patterns for a few areas within the city, and show distinct variations between airports, downtown and a selected neighborhood.

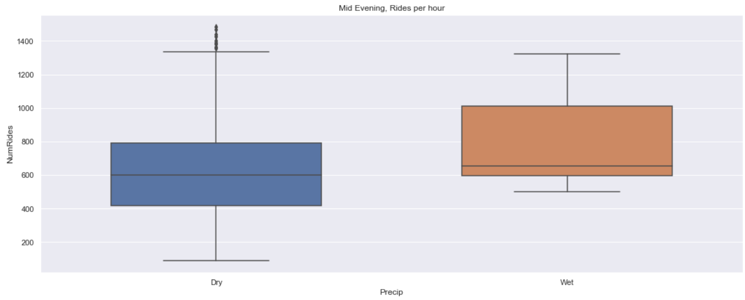


3) The highest rideshare usage occurs within the downtown core and to/from airports, and drops off quickly away from these areas. The following chloropleth map, which displays the the sum of total pickups and dropoffs by census tract, illustrates this pattern clearly:

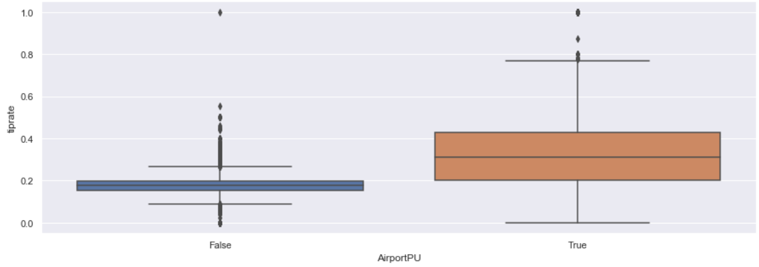


4) The relationship between rideshare usage patterns and weather is not obvious. When viewed at the daily level, there was no clear pattern illustrating a link between rideshare utilization and either temperature or precipitation. Zooming in to the hourly level, there appears to be no statistical relationship between average ride duration and precipitation, and only a small statistically significant effect (decrease) in average ride distance. If we isolate a more specific time window (weekday evening commute, post-commute hours) we see more specific patterns, with utilization declining during the commute period, and then increasing in the post-commute period. [Future work - look at impact on usage by route, not tract]

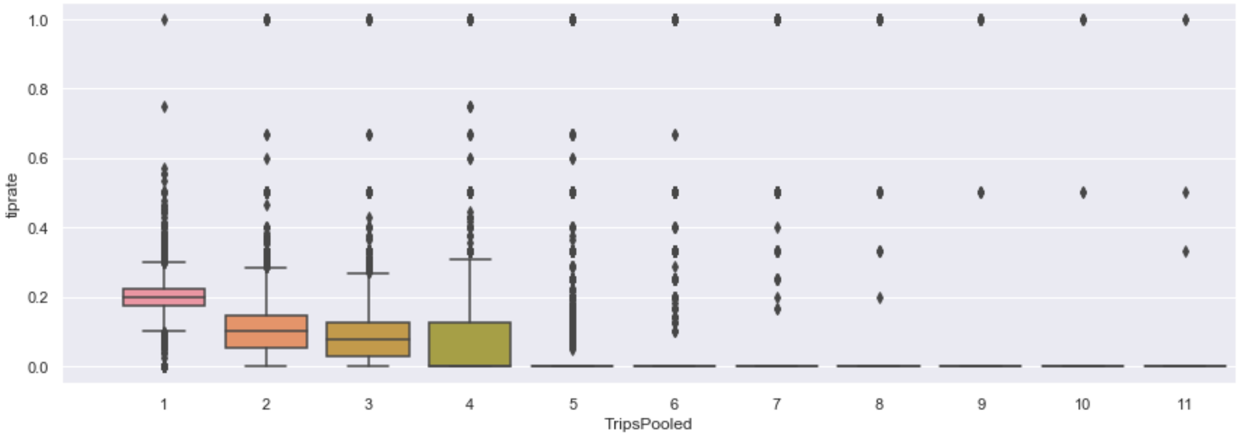




5) Riders generally tip the driver on approximately 20% of rides. Factors influencing the likelihood of a tip include whether the ride originates at the airport, and whether the ride is pooled. A two-sample z-test of the proportion of tipped rides between airport/non-airport rides demonstrated a statistically significant increase in the odds of tipping for airport pickups. The following box plot, which shows the percentage of tipped rides after the data was sampled into hourly bins, illustrates this effect.



The following boxplot illustrates the distribution of tipping rates for pooled and unpooled rides, with the number of riders in the pool increasing towards the right (TripsPooled=1 indicates the ride was not pooled, though the rider may have authorized pooling). As we move to the right we see the average tipping rate (percent of rides with tips) declining. This suggests price-sensitive customers that are willing to take on a pooled ride are less likely to tip.



6) Given the nature of the Trips dataset, there exist strong relationships between many of the parameters. For instance, the Pearson correlation coefficients between both trip distance and duration (0.80) and trip distance and fare (0.84) indicate strong correlations between these variables. We also see evidence of an inverse correlation (-0.53) between distance from downtown and population density, indicating a decrease in population density as you move away from downtown.

Summary of Findings

This evaluation of the Chicago RIdeshare dataset has uncovered many interesting patterns in the data. Future work will involve developing predictive models that can learn from these patterns to predict rideshare demand, and developing clustering models to gain additional insight into how rideshare consumers are using these services.

[1] Community Areas are groupings of neighborhoods used for planning purposes by the City. They are larger than Census Tracts, approximately 3 square miles.