Uber - NYC Data Analysis

Dost Arora -1RV16IS015
Anmol Gaba - 1RV16IS008

Problem Statement

Early in 2017, the NYC Taxi and Limousine Commission (TLC) released a dataset about Uber's ridership between September 2014 and August 2015.
This project aims to:

- visualize Uber's ridership growth in NYC during the period
- characterize the demand based on identified patterns in the time series
- estimate the value of the NYC market for Uber, and its revenue growth
- other insights about the usage of the service

Data Set

The data comprises one complete year of trips, with a total of about 31 million entries. The uncompressed file itself is 1.4 GB, which is still fine to work on a laptop with 16 GB of RAM. However, some objects will be large enough to require better reasoning about how to efficiently apply transformations to them, from date-time parsing to arithmetic functions.

Data Quality and Consistency

There were very few clearly erroneous entries in the dataset and a small proportion of suspicious cases or anomalies that warrant further internal analysis. These cases are, for example, those with very long distance traveled, but destination still recorded within New York City, or those with average speed slower than walking, but very long duration (beyond a reasonable assumption for the amount of time necessary to get out of some really bad traffic gridlock, or the unlikely situation of a driver left waiting).

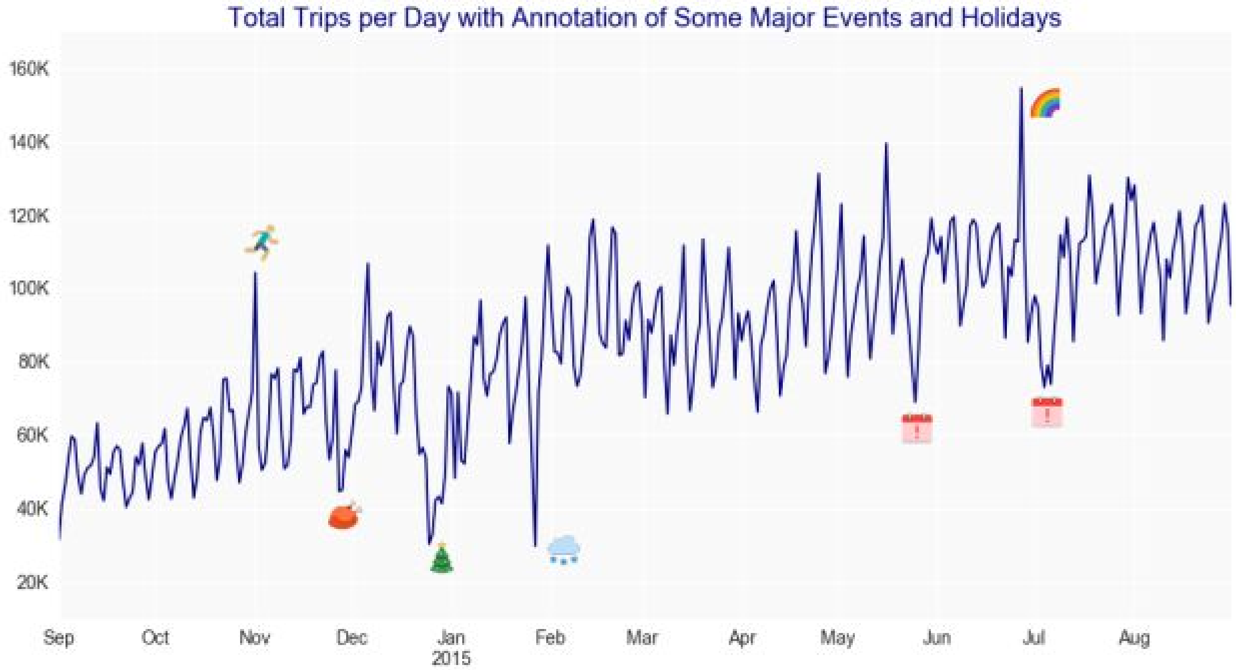
In addition, there was a small proportion of cases with distance and duration equal to zero. Do they represent canceled trips? A small subset actually shows distinct origin and destination zones, indicating that some distance was driven but not recorded. In other cases, the recorded distance was zero, but the trip duration was more than that, even beyond 5 minutes in rarer cases. Are these system errors, fraud?

Finally, about 4% of the destination data were missing, and an extremely small number of cases had missing trip distance and destination. The imputation method chosen for the latter set was the mean distance and duration of their respective origin-destination pair. The entries with missing destination were left unchanged, although the information from the vast number of complete cases could potentially be used to determine the most probable destination.

Uber's Growth

Uber launched in NYC in May of 2011, the first city outside of its San Francisco headquarters. NYC is probably the largest and most lucrative rideshare market in the world, with a total demand (for taxis and for-hire vehicles) in 2017 of more than 240 million trips per year.

The number of Uber trips per day in NYC is still growing significantly. In 2017 so far, this number has often surpassed 200,000, but the plot below shows that by mid-2015 it was hovering around 120,000.



Another interesting insight from the plot above is the effect of major events on the number of trips. For the period of time analyzed, negative impacts are related to Thanksgiving, Christmas, Memorial Day, and Independence Day. A lingering (two consecutive days) drop in activity is seen for all these holidays but Memorial Day. It turns out that the July 4th holiday was observed on Friday in 2015.

In addition, an apparently odd and very significant drop in the number of trips is shown on January 27th. This was a result of a curfew imposed by the NYC's mayor in preparation for a blizzard.

Trends in the Demand for Rides in the City

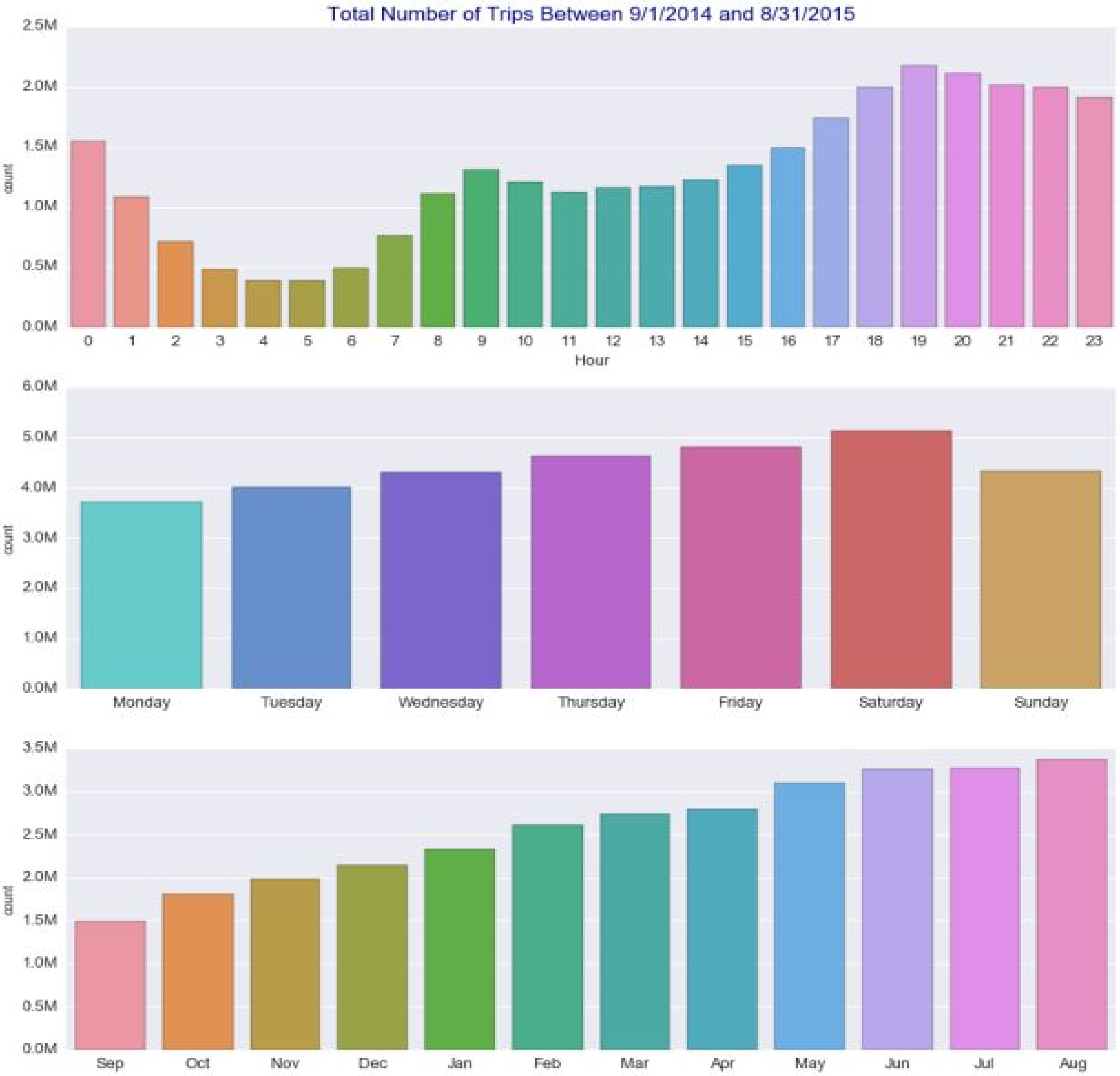
The data also allows us to visualize other interesting trends over time. In the bar charts below, we can see that the demand for Uber is higher from 4 PM until around midnight. Saturday has the highest demand. Interestingly, Sunday shows a level of demand similar to Wednesday, which is higher than Monday or Tuesday. When looking at the total demand per month along the period of time analyzed, seasonal effects are masked by the consistent month-to-month growth.

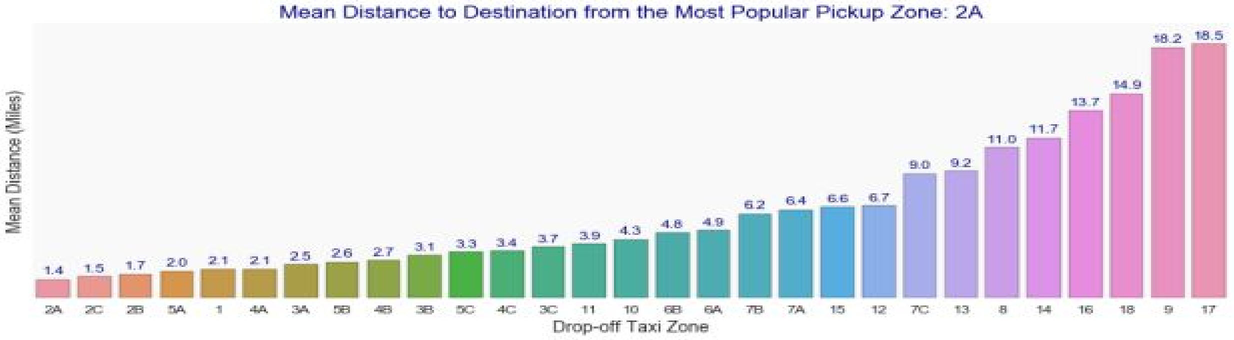
It's well-known that **Manhattan dominates the demand for taxis and rideshare services**. The TLC states that 92% of all trips by yellow cabs start there, whereas this number is about 70% for all FHV app-based companies.

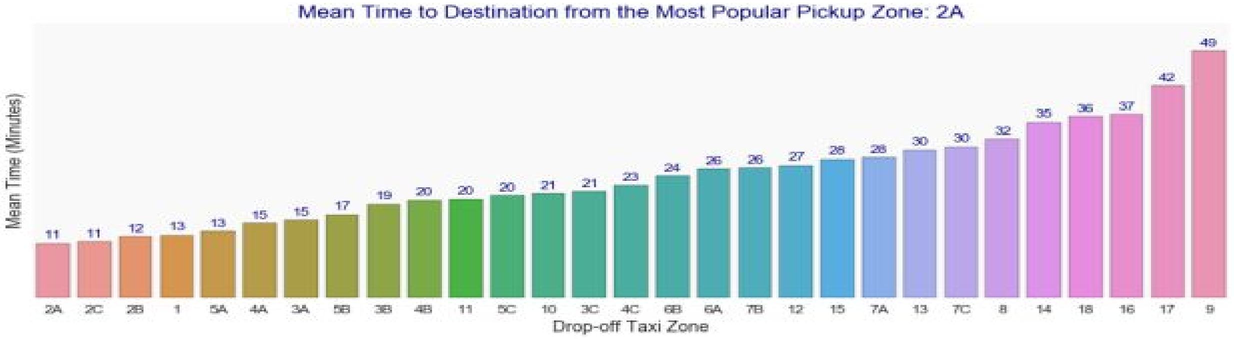
In the dataset, the locations have been anonymized, but it's reasonable to assume that the top origin codes are probably based in Manhattan. In this case, the top destination codes are also based in Manhattan, because they overlap, as can be seen in the plot below.

It's well-known that Manhattan dominates the demand for taxis and rideshare services. The TLC states that 92% of all trips by yellow cabs start there, whereas this number is about 70% for all FHV app-based companies.

In the dataset, the locations have been anonymized, but it's reasonable to assume that the top origin codes are probably based in Manhattan. In this case, the top destination codes are also based in Manhattan, because they overlap, as can be seen in the plot below.


Total Number of Trips Between 9/1/2014 and 8/31/2015


Mean Distance to Destination from the Most Popular Pickup Zone: 2A


Mean Time to Destination from the Most Popular Pickup Zone: 2A

The most popular pickup and drop-off locations are 2A. In fact, 29% of all Uber trips during the analyzed period have either started or ended in this zone. The charts below show the mean distance and time to destination for a trip originating at 2A.

he data has 28 unique origin codes and 29 unique destination codes. Assuming that code 18, the extra destination code, represents the Newark Airport (EWR), a relevant destination outside of New York City, then we can infer that 2A is in Midtown, based on the mean time and distance to arrive at location 18.

Given the concentrated demand within Manhattan, about 68% of all Uber trips have a driven distance of 5 miles or less. However, as noted earlier, FHV app-based companies (as well as Green Cabs, by design) tend to serve more the outer boroughs than Yellow Cabs.

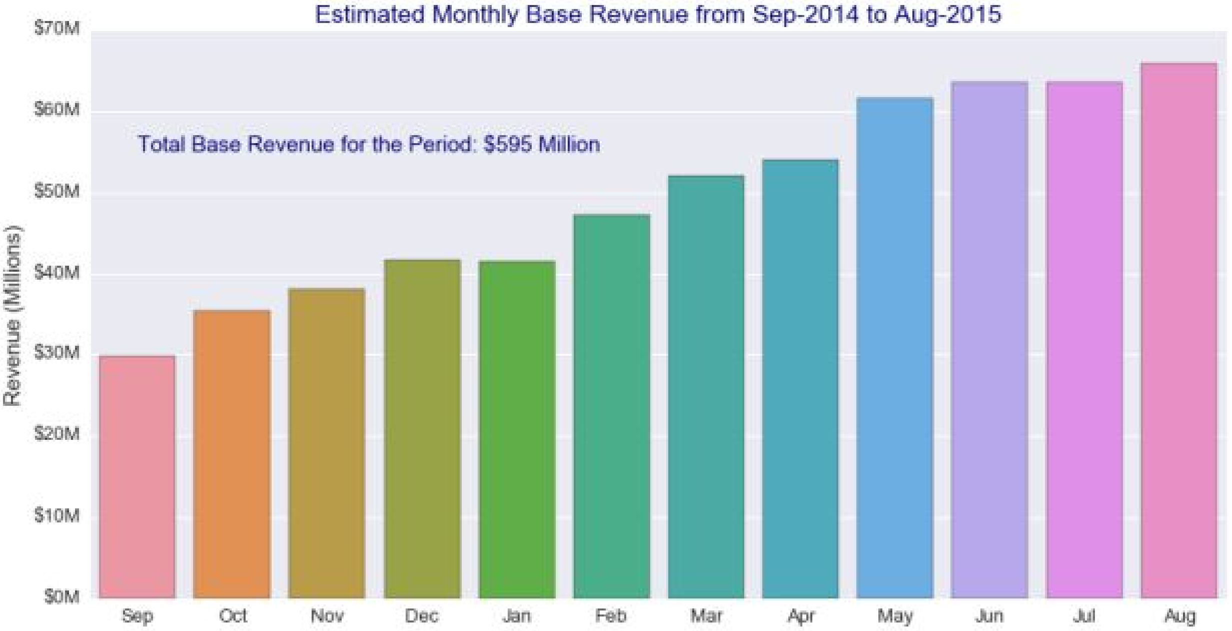
Uber's Revenue Figures and Growth

Uber offers different types of services with distinct prices, namely Uber X, Uber XL, Uber Black, Uber SUV, and Uber Pool. Except for the latter, all other services carry a higher fare than Uber X. Moreover, Uber practices "price surging", which affects the revenue positively.

We chose to use Uber X published fares to calculate the revenue as this is probably the most popular product. Therefore, the base revenue is a conservative estimate of the actual revenue.

Indeed, the mean revenue per trip between September 2014 and August 2015, calculated from the data by assuming they were all Uber X, was $19. Comparatively, Uber has published that the average NYC Uber X fare was $27 in September 2014.

The chart below shows the estimated base revenue growth for each month:



Based on other data shared by Uber, it's possible to roughly estimate the revenue associated with Uber

However, the impact of Uber Pool in the first 9 months since launching seems to not have been significant, considering that there were more than 25 million trips during this period of time. Despite the apparent "slow" growth in the first months, the Uber Pool product is important because it attracts new riders. The average fare has dropped overall, but the number of users has increased. Lower fares mean less attractive pay for the drivers, who operate as contractors. Thus, increasing the number of drivers (or decreasing turnover) at the same pace of the business growth has become a bigger