Context:

The Uber dataset contains 31 million rows of data from September 2014 to August 2015, including information about trip origin, destination, pickup and drop-off dates/times, trip distance and duration. The data allows for estimating Uber's revenue per trip in NYC, but the pickup and drop-off locations are anonymized and grouped as taxi zones for privacy. The dataset is 1.4 GB in size and although it can be worked on a laptop with 16 GB RAM, some transformations may require efficient handling.

Here, I did an EDA on the uber dataset, which consisted of data from rides booked in NYC over the course of one year, from September 2014 to August 2015, and eventually computed the income of Uber in one year as well as the revenue trend during the year. This dataset was limited to around 31 million rows and columns of origin, destination, pickup date and time, travel distance and duration.

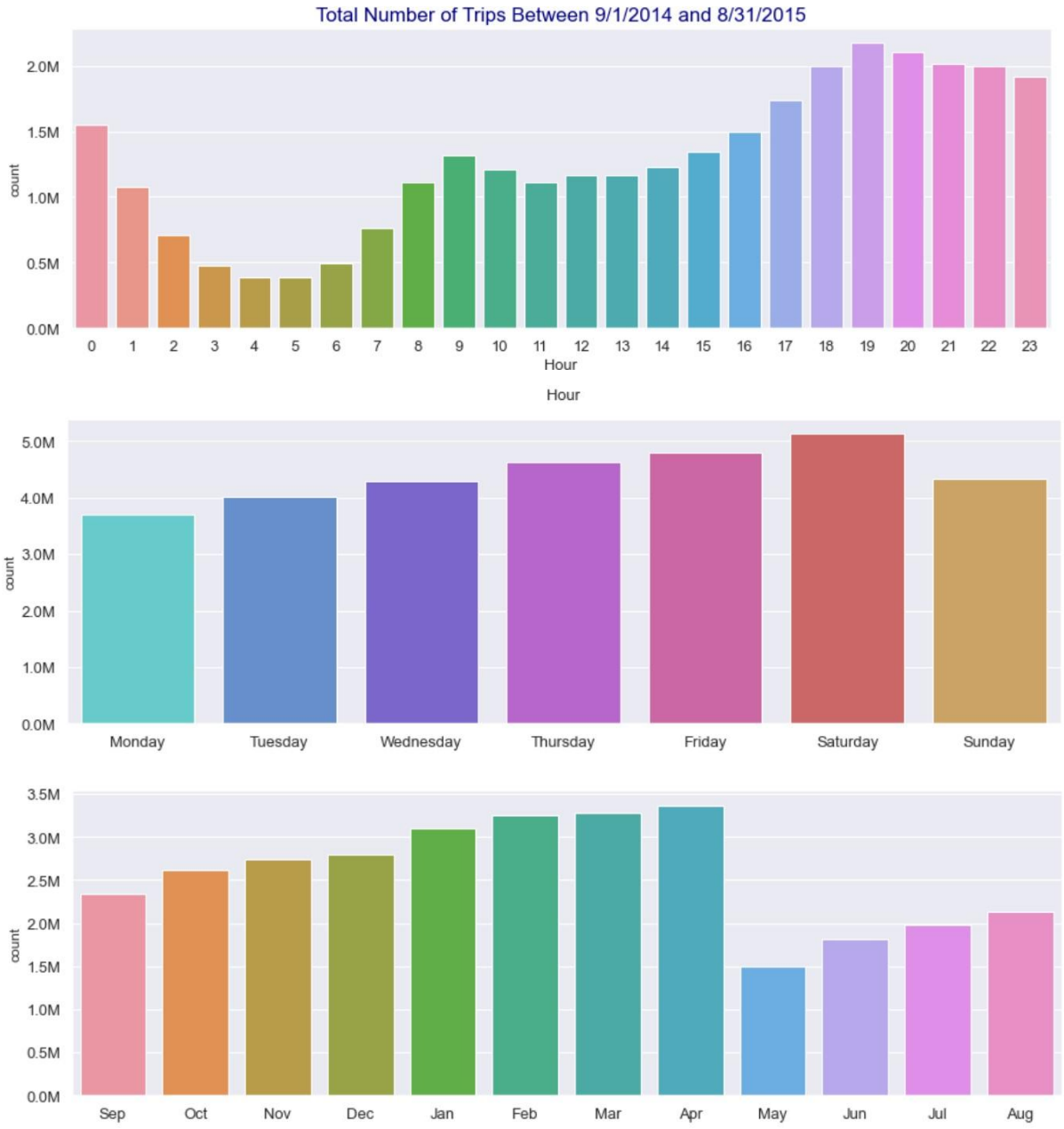
Used Pandas data frame to check for the presence of null values in the dataset and discovered that the destination column had 1.3 million missing data but that the corresponding data in other columns such as origin, trip duration, trip distance, pickup date and time data were present for these missing data, implying that we cannot remove these missing rows from our dataset because they have a significant impact on Uber's total revenue. Because the destination column was categorical, I utilized the mode imputation approach to address missing values. There were some missing values in the trip length and trip distance columns as well, so I used the mean imputation approach to address these.

Also, feature engineering was used to construct several columns that can help with analysis and expand the area of study. As a result, I made distinct columns for the year, month, and day of the trip. These columns were then utilized to determine the data's monthly trend. Also included a column for trip revenue, which was used to determine total revenue and average revenue. The revenue consisted of various components like:

Base fare = 2.55
Per minute = 0.35
Per mile = 1.75
Minimum fare = 8

Findings:

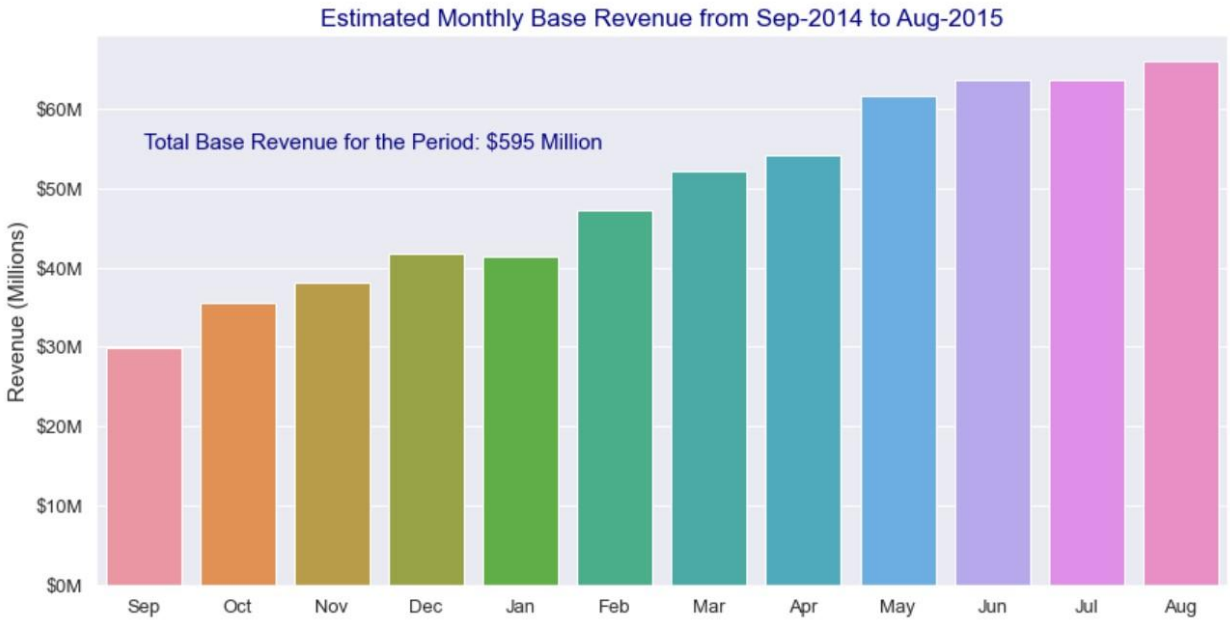
1. **The effect of time on demand for Uber rides: distribution per hour, weekday, and month.**



- In the bar charts above, we can see that the **demand for Uber is higher** from 4 PM until around midnight.
- In the second chart demand is gradually increasing from monday to saturday and **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.

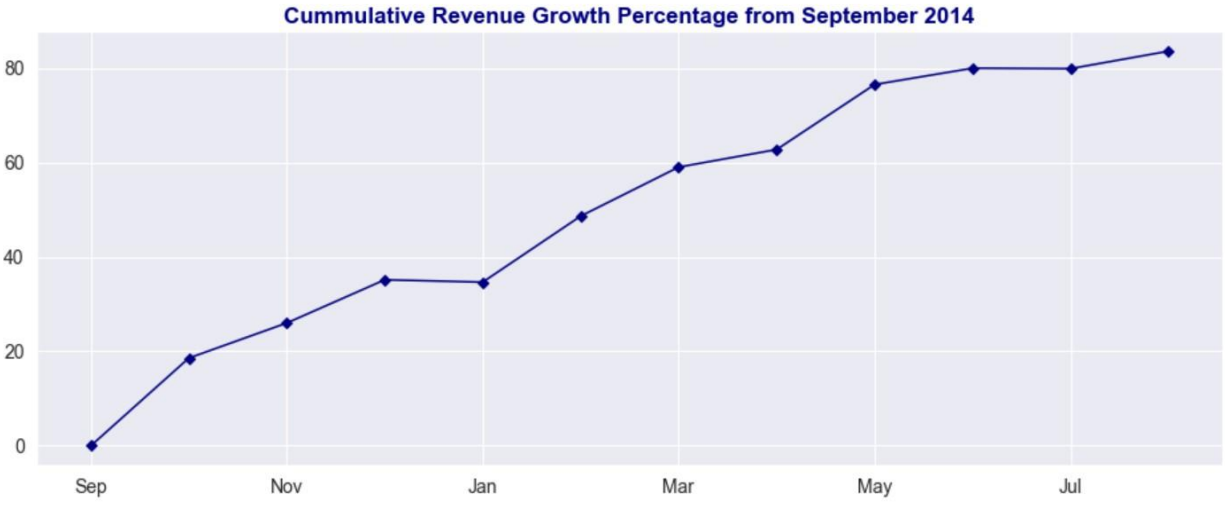
2. **Estimated Monthly Base Revenue: how much was the NYC market worth in the period?**

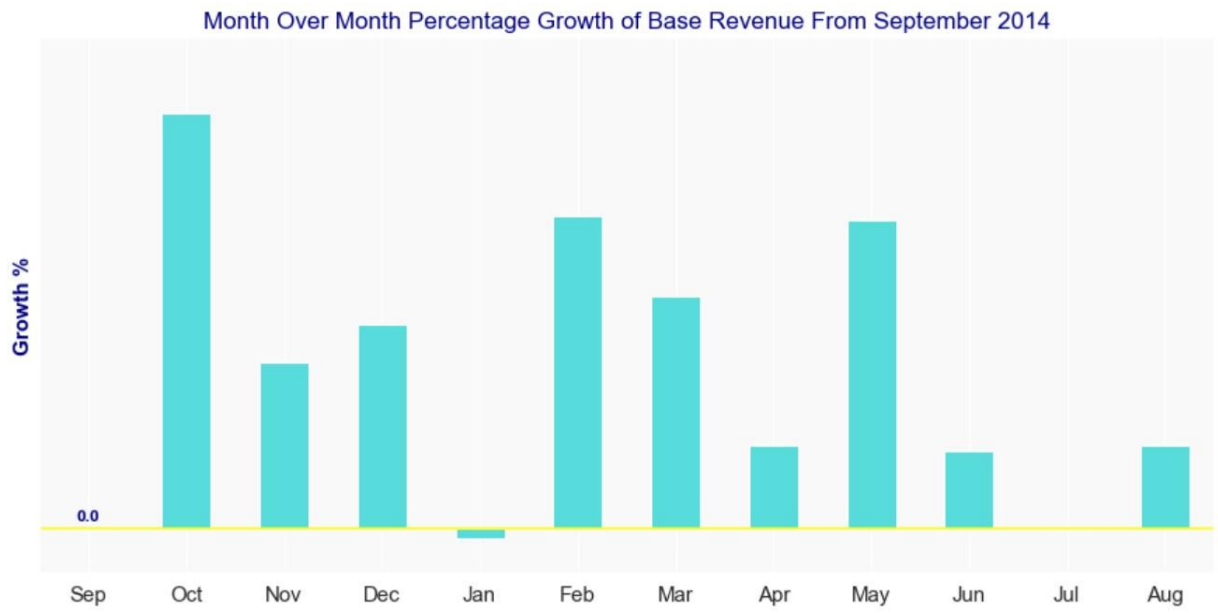
Mean fare from revenue estimate for the period: $19.24



- It's important to note that from the gross estimated revenue, Uber's share is about 25% of the total. Therefore, we can conservatively estimate that Uber's gross margin in NYC from September 2014 to August 2015 was in the order of **150 million dollars**. The estimated gross margin, considering instead the 27 average fare previously mentioned, was of the order of **210 million dollars**.
- **Total revenue** of Uber in 1 Year is **595M and Gross margin of 149M**.
- In the bar plot we observed that revenue per month generally increases month by month in that period except the month of January and June. This may happen because of 2 Federal holidays in the month of January and summer vacation starting from july.

3. **Month over Month Base Revenue Growth: how fast has Uber grown in the period?**

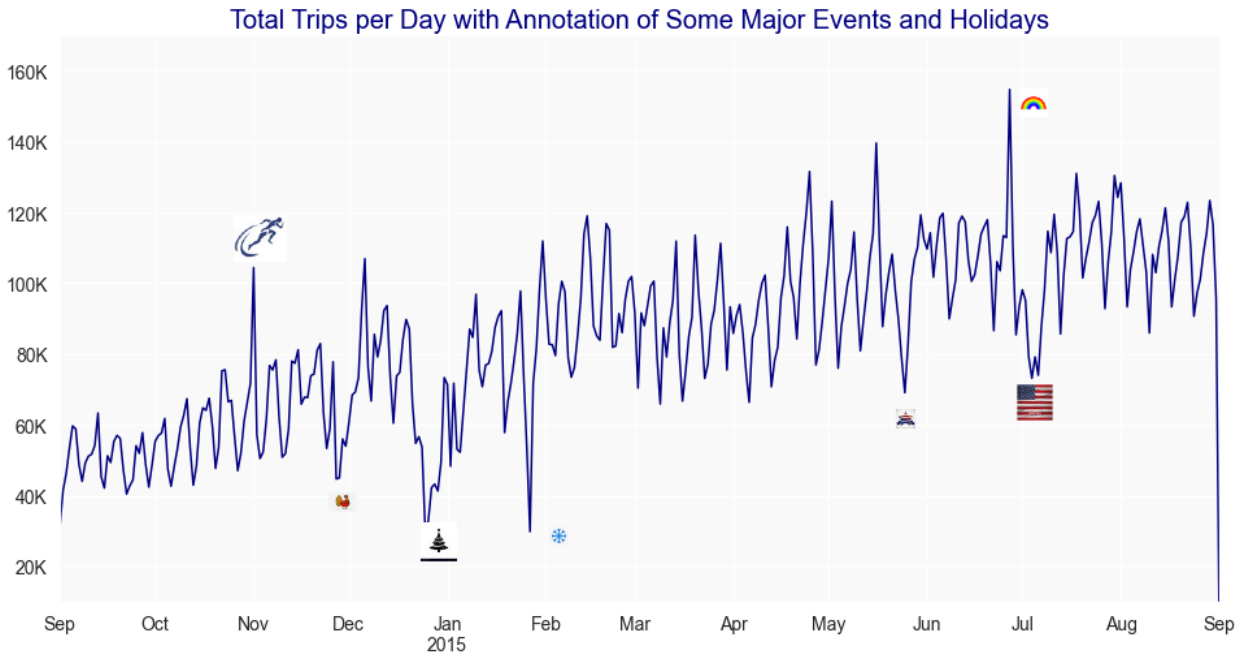




- Sep Revenue = $29,967,741 Growth % = 0.0%
- Oct Revenue = $35,531,001 Growth % = 18.6%
- Nov Revenue = $38,170,687 Growth % = 7.4%
- Dec Revenue = $41,661,569 Growth % = 9.1%
- Jan Revenue = $41,457,151 Growth % = -0.5%
- Feb Revenue = $47,252,852 Growth % = 14.0%
- Mar Revenue = $52,154,385 Growth % = 10.4%
- Apr Revenue = $54,095,066 Growth % = 3.7%
- May Revenue = $61,539,912 Growth % = 13.8%
- Jun Revenue = $63,667,666 Growth % = 3.5%
- Jul Revenue = $63,607,348 Growth % = -0.1%
- Aug Revenue = $65,961,099 Growth % = 3.7%

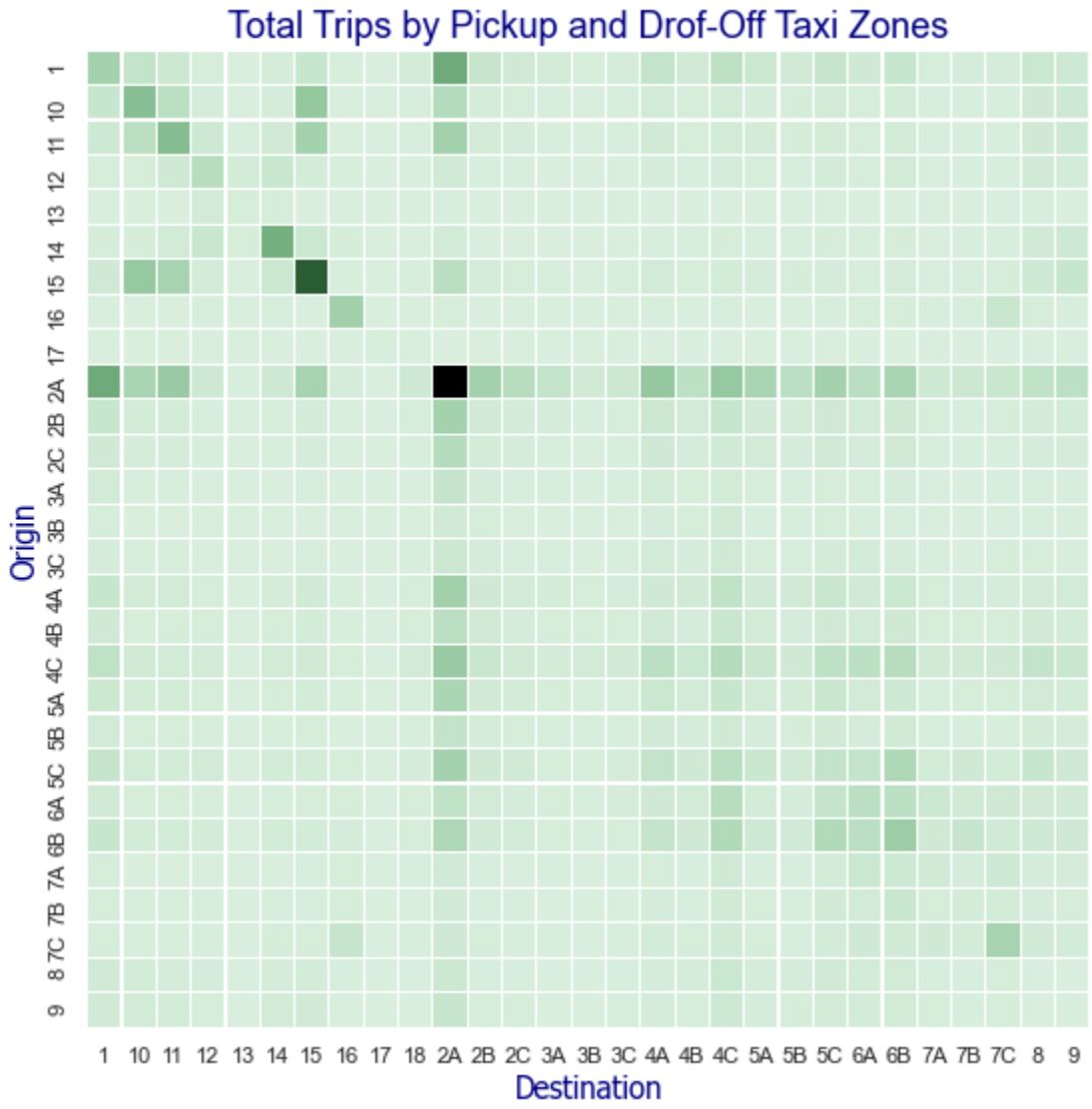
Cumulative % Growth Over Period: 83.54793827524823

4. **A plot with the total number of trips per day, highlighting some changepoints associated with major holidays and other weather and touristic/cultural events.**



- **The effect of major events on the number of trips**.
    - **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.
    - 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 NYC's mayor in preparation for a blizzard.
    - The plot also highlights which events have positively impacted the number of trips that year, with the International Marathon and the Gay Pride Week standing out as the **strongest contributors**.

5. **Visualizing the most popular pick up and drop off location pairs.**



2A and 2A code have the **highest correlation** and 15 and 15 after that is one of most pickup and drop locations.