Exploratory Data Analysis

To better understand this huge dataset and the features it contains, I first performed EDA. The main purpose of this is to see how the prediction variable behaves with various features and how I can clean and leverage these features in our models to achieve best results.

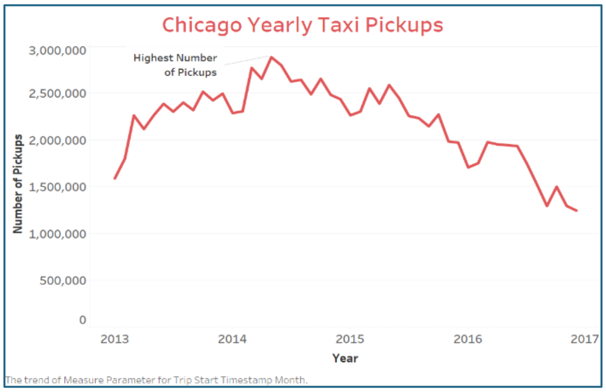

Figure 1

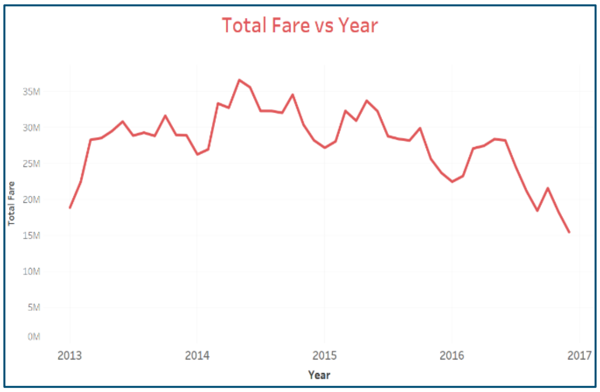

Figure 2

First, I wanted to visualize how much business has the Chicago cabs lost due to Uber and Lyft. It can be clearly seen from Figure 1 that there has been a dramatic decrease in the number of pickups, with an annual rate of 35%, making the Chicago cabs to lose around 45% of their business by the May 2015.

With number of rides decreasing one can easily deduce that the total revenue followed the same trend, total revenue for the Chicago taxi was reduced from $35 million to $18 million. This has caused huge economic burden on the cabbies as they aren't generating enough fares to keep up with their loan payments and meet their expenses.

More than 350 foreclosure notices or foreclosure lawsuits have been initiated against medallion owners in the year 2017, compared to 266 in 2016 and 59 in 2015. Since October, lenders have filed lawsuits against at least 107 medallion owners who have fallen behind on loan payments, according to the union's count. The major reason behind this financial distress is that since the emergence of Uber, Cabbies face an uneven playing field with the ride-share companies, who typically don't face the same permitting and fee rules. **[1]**

After digging a little deeper, I realized that the city of Chicago has enforced set of rules for taxi and ride sharing industries and made it impossible for cabbies to compete with Uber and Lyft. Transportation companies compete for customers, and ultimately it is the consumer who makes the choice. Consumers always look for cheaper options and sadly this has been the reason for the decline in number of rides for the Chicago's Taxis.

Figure 1 shows that there is correlation between number of rides and year. Hence every component of time be it year, week, day or hour should have a role to play in estimating number of rides on any specific day. I wanted to understand the pattern exhibited by pick up density over the course of an entire year. Hence, I plotted a graph (Figure 3) between number of rides vs number of week. This graph led us to one insightful conclusion. Almost every week the number of rides remained constant, except for week 10, 48 and 52.

Week 10 contains the most important holiday celebrated widely in Chicago City i.e. St. Patrick's Day. Week 48 and week 52 contains Thanksgiving and Christmas holiday. Impact of holidays are analyzed later in the report.

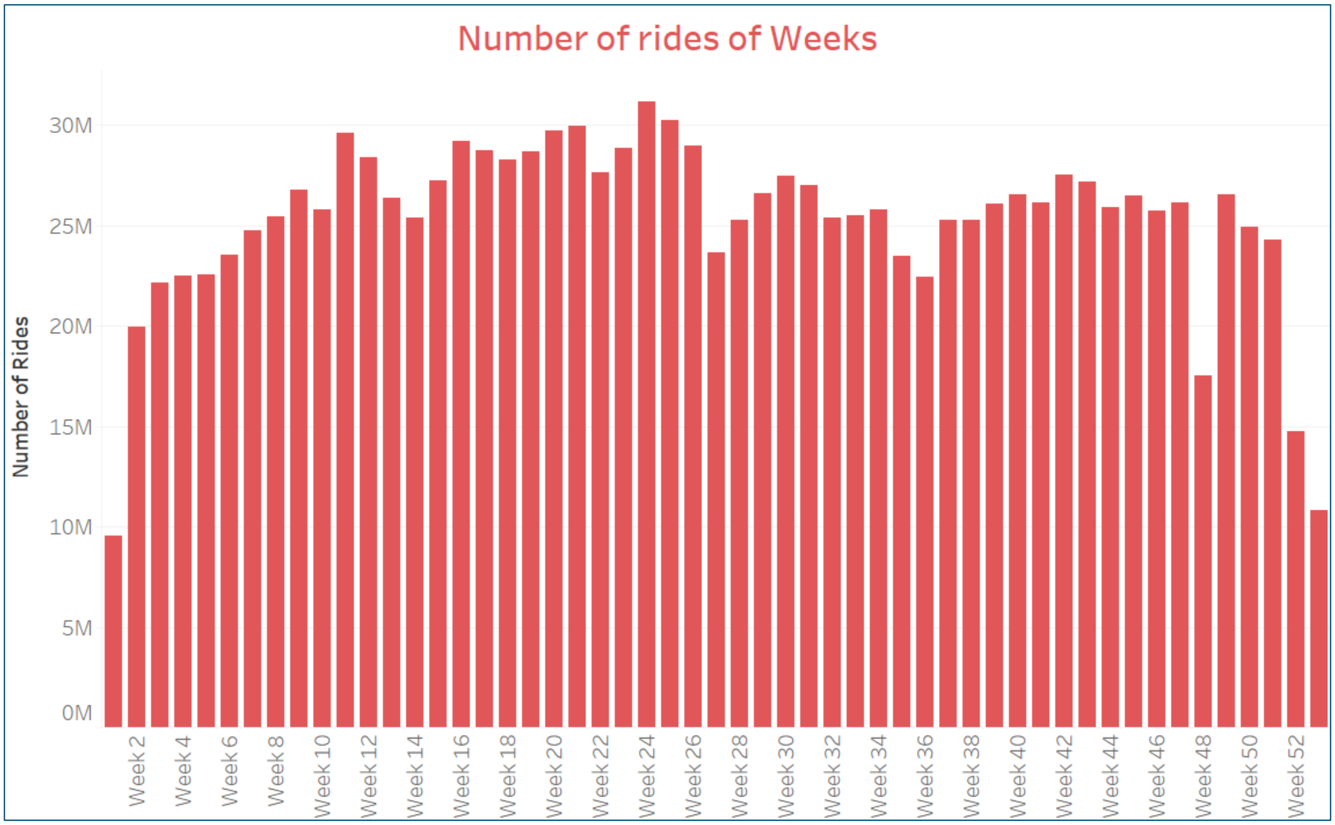

Figure 3

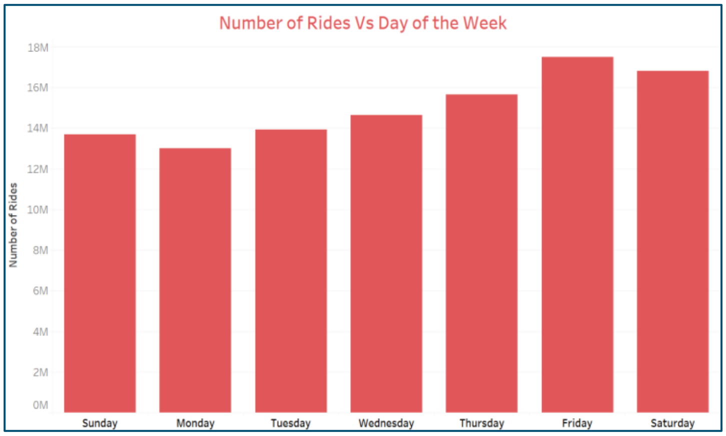

Figure 4

After analyzing the weekly trend, we wanted to see how number of rides varied with day of the week. The trend was pretty obvious, while most of the days number of rides were constant except on weekend i.e. on Friday, which had maximum number of rides. Total trip amount followed the same trend as well.

For more granularity I analyzed the number of pickups occurring by hour over the course of the day.

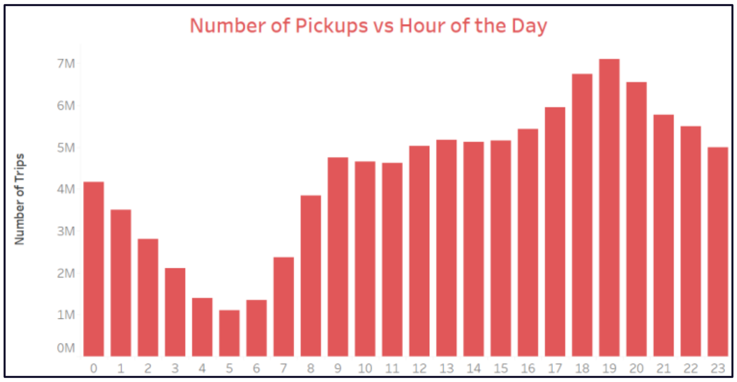

Figure 5

The morning and evening rush hours are clearly visible. Chicago is an all-day and all-night city as evidenced by the number of pickups throughout the hours of the day and night. It can be seen that more taxi trips begin an hour before the midnight than there are in the morning.

Notice the dip in the number of trips at the 5th-6th hour of the day. The number gets reduced from 2000k to 1400k, I analyzed how fare changed in this interval.

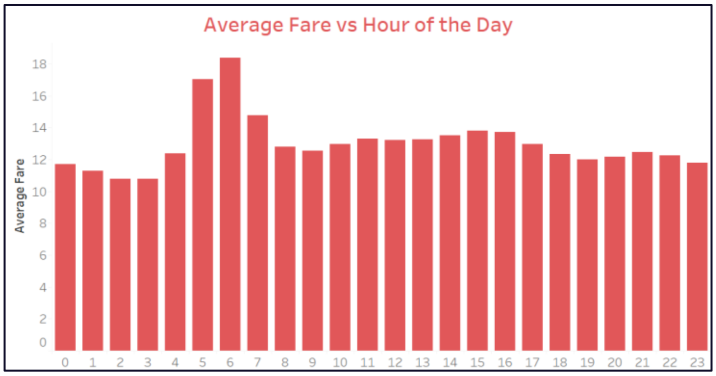

Figure 6

Results were quite interesting as contrary to our hypothesis; average fare i.e. fare/ride has peaked during 5th-6th hour of the day. Although graph for total fare vs hour followed the same trend as number of pickups.

After some iterations I figured out that trip distance has been influencing the average fare, it has the same bar chart as well.

I plotted the distribution of trip miles (figure 7) and deduced why trip miles were skyrocketing during that interval. It turns out that number of drop offs at O'Hare International Airport (Second busiest airport worldwide) were maximum during the 6th hour of the day i.e. many Chicagoans uses taxi to board their early morning flights. Since O'Hare is located on the far Northwest Side of Chicago, Illinois, 14 miles northwest of Chicago's Loop business district (community responsible for highest number of pick-ups), trip miles and total fare are usually higher **[2]**. Although the number of drops is greater in the evening as well, but the average remains low due to large number of rides.

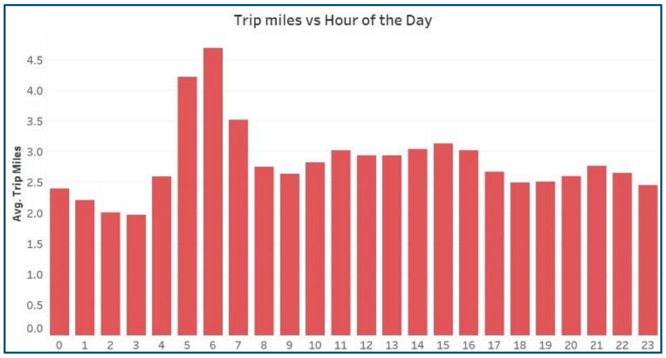

Figure 7

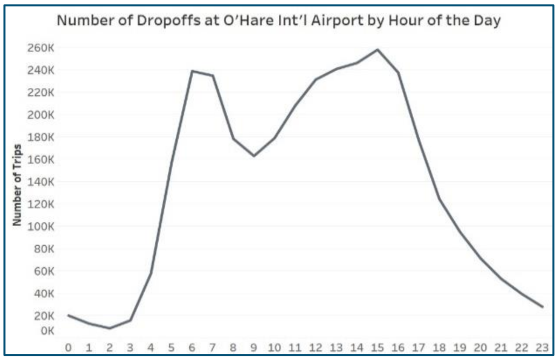

Figure 8

Weekly and hourly analysis made us curious and I wanted to see the combine effect of day of the week and hour of the day on number of rides. It would provide the exact the day and hour at which Chicago Transport Authority (CTA) should have maximum number of active taxis.

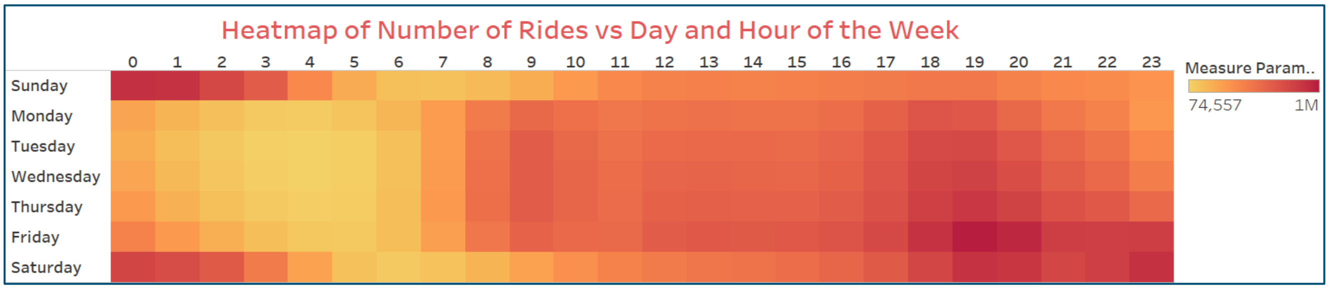

Figure 9

This heatmap (figure 9) gives a clear picture about dependence of both the parameters on number of rides. From 0-6 on Monday through Friday there are less number of rides, hence less number of taxis required to balance the demand. Chicagoans generally like to relax on Friday and Saturday, traveling a lot within the city, drinking which gives rise to more number of rides. Also, talking about allocation of resources CTA should have large number of taxis from 0th hour to 6th hour on Saturday and Sunday. *Heatmap for Total trip is similar to the number of rides.*

As mentioned previously, heatmap is entirely different for average fare. If some of the cabbies are looking to make more money per ride, following heatmap is very useful for them. It paints a profitable picture where there is high return on investment, if a cabby is willing to serve from 4AM-7AM. The traffic density at this time is less which aides higher fuel efficiency.

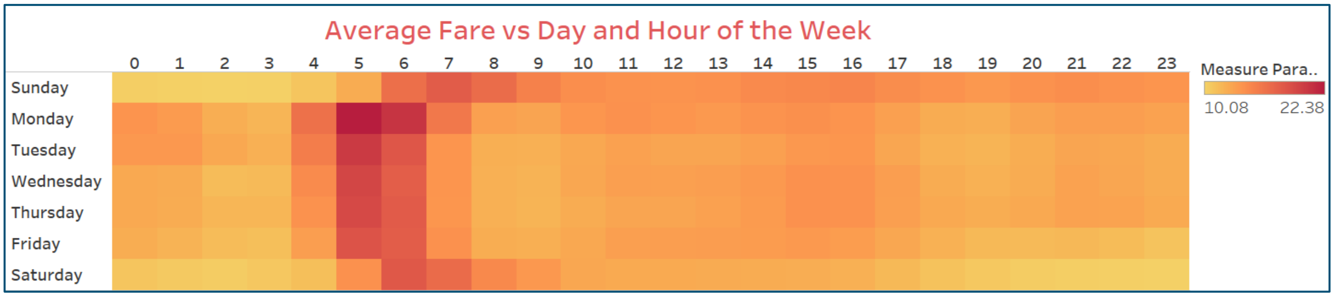

Figure 10

Also keeping in mind that since these rides are lengthier time and distance wise, from our own experiences, I have always tipped more in such situations. Hence, I plotted the similar heatmap for average tips, our intuition was right as many people appreciated the work done by cabbies and that too at 5AM or 6AM. Average tips were fairly high in the night time as well.

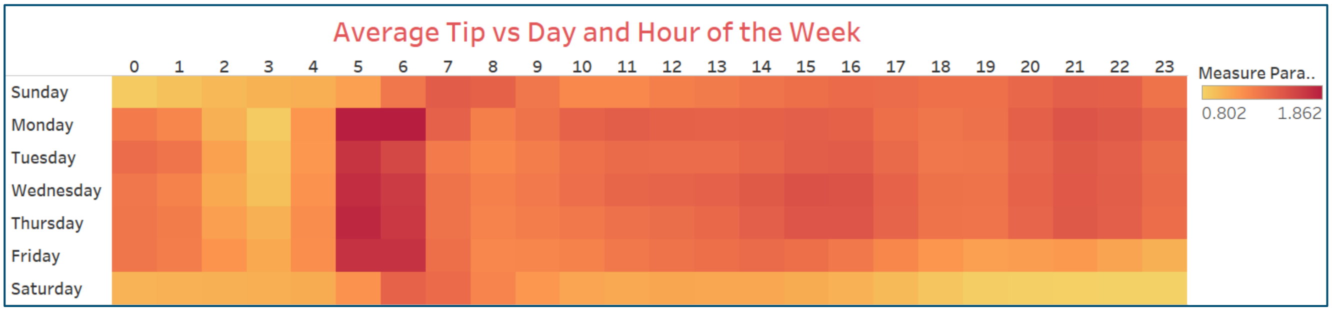

Figure 11

So far, I understood the importance of time features. Next, I wanted to understand the importance of pick up and drop off location.

Chicago's taxi pickup declines are not evenly distributed among the city's 77 community areas. For example, the Loop, Chicago's central business district, shows a 23% annual decline, while Logan Square on the northwest side shows a 50% annual decline. In general, the areas located closest to the central business district show smaller declines in taxi activity.

I defined 5 particular community areas—the Loop, Near North Side, Near West Side, Near South Side, and O'Hare Airport—as the "core", then compared pickups inside and outside of the core. (Figure 12) As of November 2016, pickups inside the core shows a 27% annual decline compared to a 42% annual decline outside of the core. On a cumulative basis, core pickups have declined 39% since June 2014, while non-core pickups have declined a whopping 65%. **[3]**

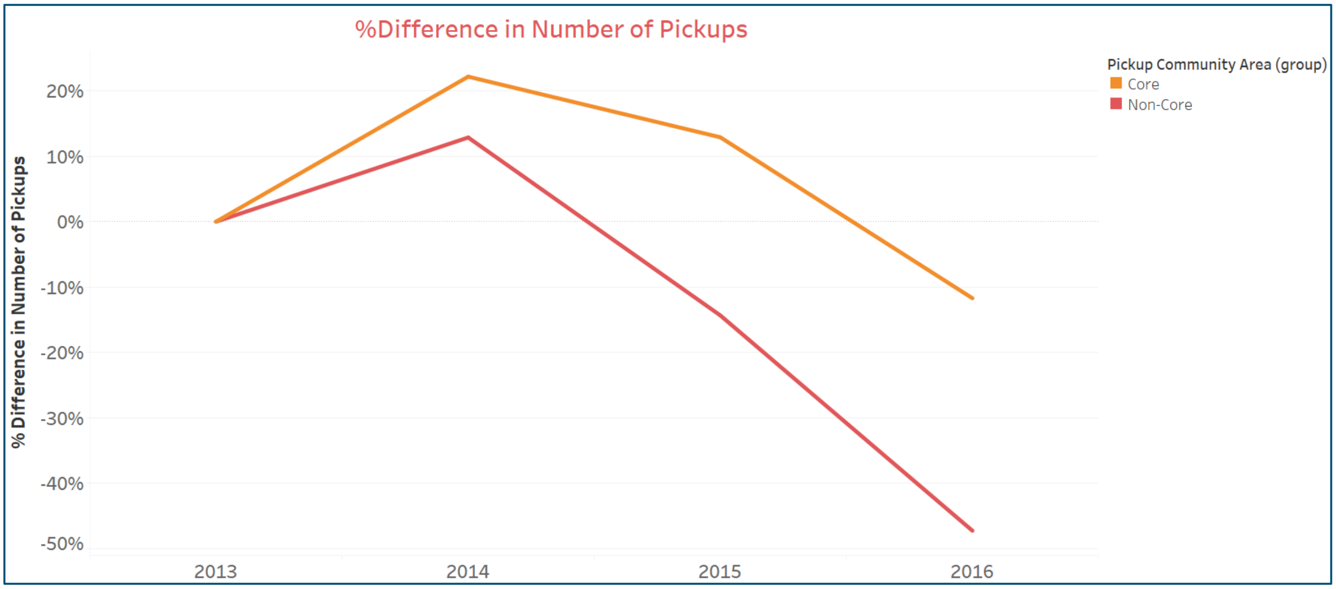

Figure 12

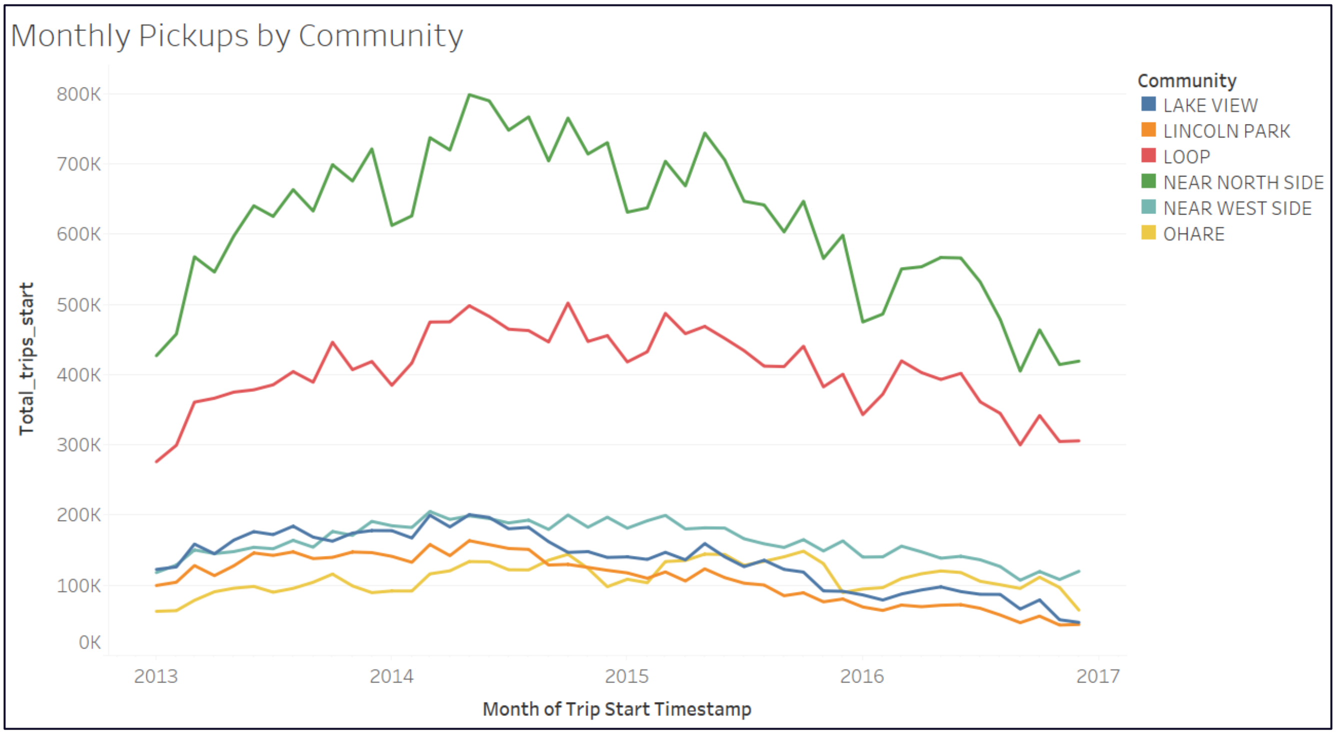

Figure 14

The above figure depicts the distribution of number of rides for six different areas. It again emphasizes the importance of those 'Core' areas I talked about earlier.

One of the parameter that is most important for any taxi company to know is what are some of the most influential pick up areas in a particular city. Generally, these are the areas which have a great revenue potential and by focusing on these areas one can increase a taxi's overall profit. As mentioned earlier there are some of the core areas, the reason I called them as a core area was because of their extreme affluence, typified by the Magnificent Mile, Gold Coast, Navy Pier, and its world-famous skyscrapers. Magnificent mile and Gold Coast are among the top ten richest neighborhoods in America. Navy Pier is the number one tourist destination in Chicago City, drawing nearly nine million visitors annually. **[4]**

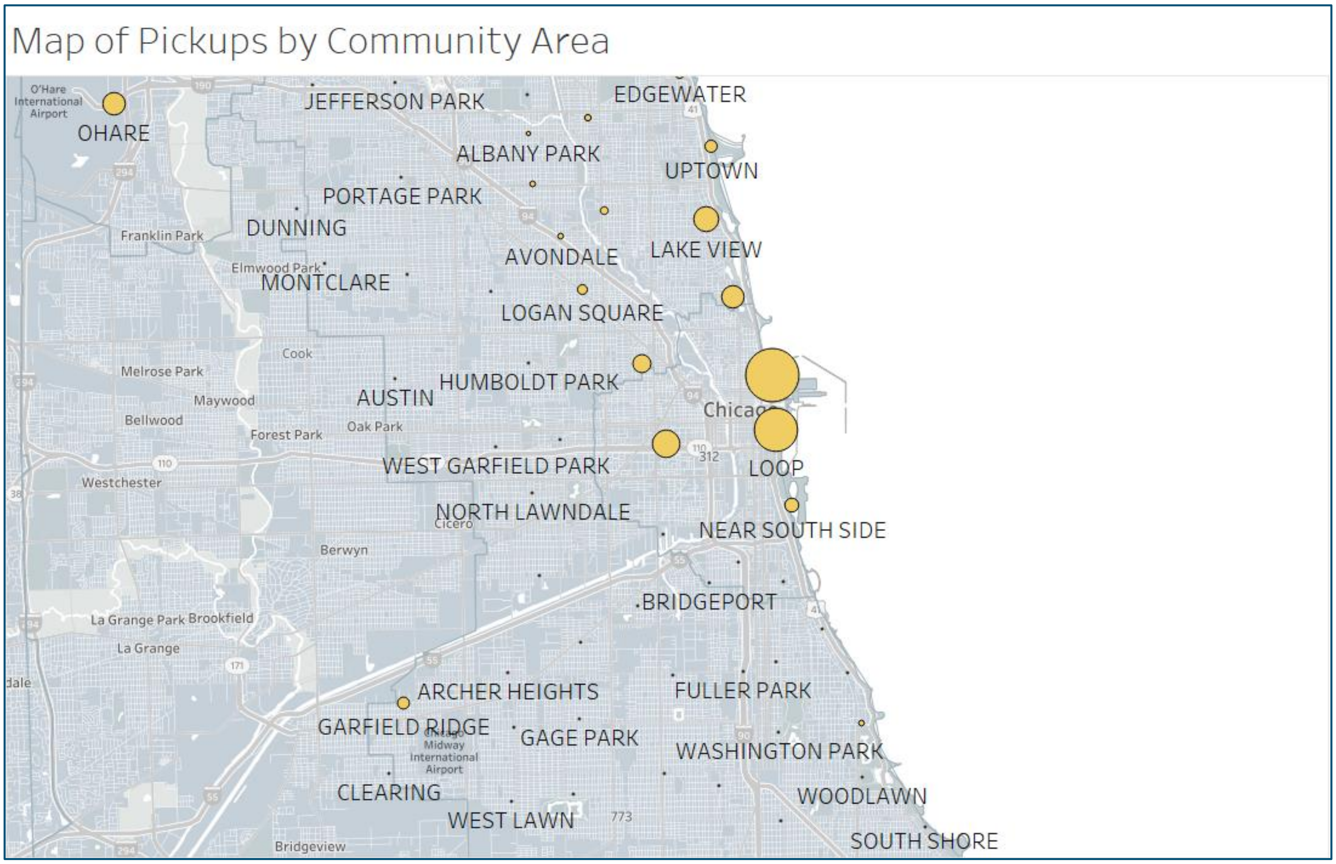

Figure 13

Because of such affluence, business and tourism these areas account for a major portion of revenue for the Chicago Taxis. O'Hare, as mentioned before, is the second-busiest airport in the world by the number of takeoffs and landings and also contributes majorly towards the revenue. **[2]**

So far in this report I have talked about how time have affected taxi revenues, how location plays an important role in determining whether or not return on investment will be higher. Its time to talk about other things which has a huge impact on taxi revenues.

Remember those holidays I talked about, let's discuss about them in detail.

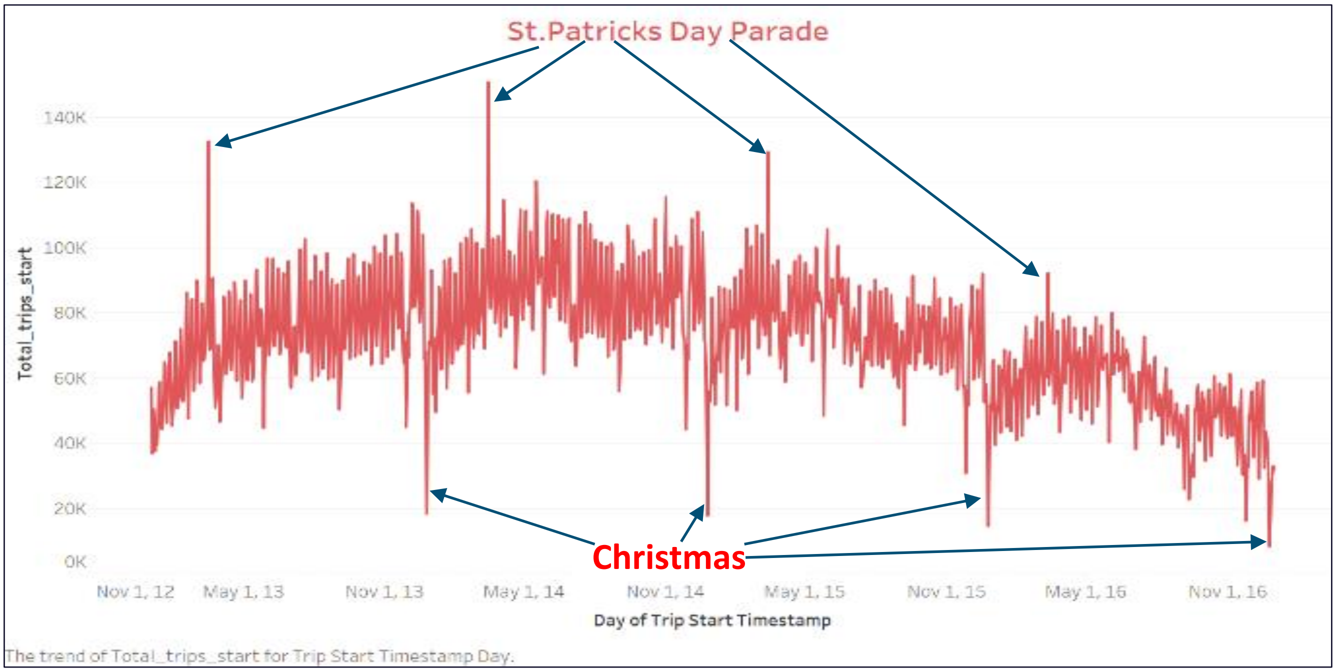

Figure 15

Holidays have always been special to the Chicago City. From raucous pub crawls to lively parades, there is nothing quite like St. Patrick's Day in Chicago. In March, Irish taverns are packed with revelers, jovial crowds jam the city streets and the Chicago River sparkles brilliant shades of emerald green. With so much going on be it the downtown parade or dyeing of the Chicago river, Chicagoans travels a lot and they generally prefer public taxis. **[5]**

It turns out that St. Patrick's Day parade which is held in march of every year accounts for maximum number of taxi trips on a particular day.

One of the important national holiday is Labor Day, created to celebrate the contributions of the American worker. It falls on the first Monday of September, resulting in a dearly coveted three-day weekend. Labor Day is essentially a day off for cabbies and this results in less number of ride on that day.

Memorial Day, Thanksgiving and Christmas of every year has the least number of taxi trips respectively. This is again due to the fact that there are few Taxi's in service on these holidays. This explicit visualization motivated us to incorporate important holidays as parameters in our predictive model.

After all of the Exploratory Data Analysis, it was of utmost importance to look for anomaly in this huge dataset, hence while querying the data most of the outliers were treated by ensuring following things: -

- The ratio of trip total and trip miles should be greater than 2, for instance if trip distance is 2 miles, the charge should be at least $4. Also, to prevent high outliers the same ratio should also be less than 10.
- With the city traffic and speed restrictions, speed i.e. ratio of trip miles and trip hour was kept less than 70. This enforces another restriction that trip seconds should have non-zero values.
- The minimum fare amount for any trip in Chicago is $2.25, hence all the fare values should have non-zero values.