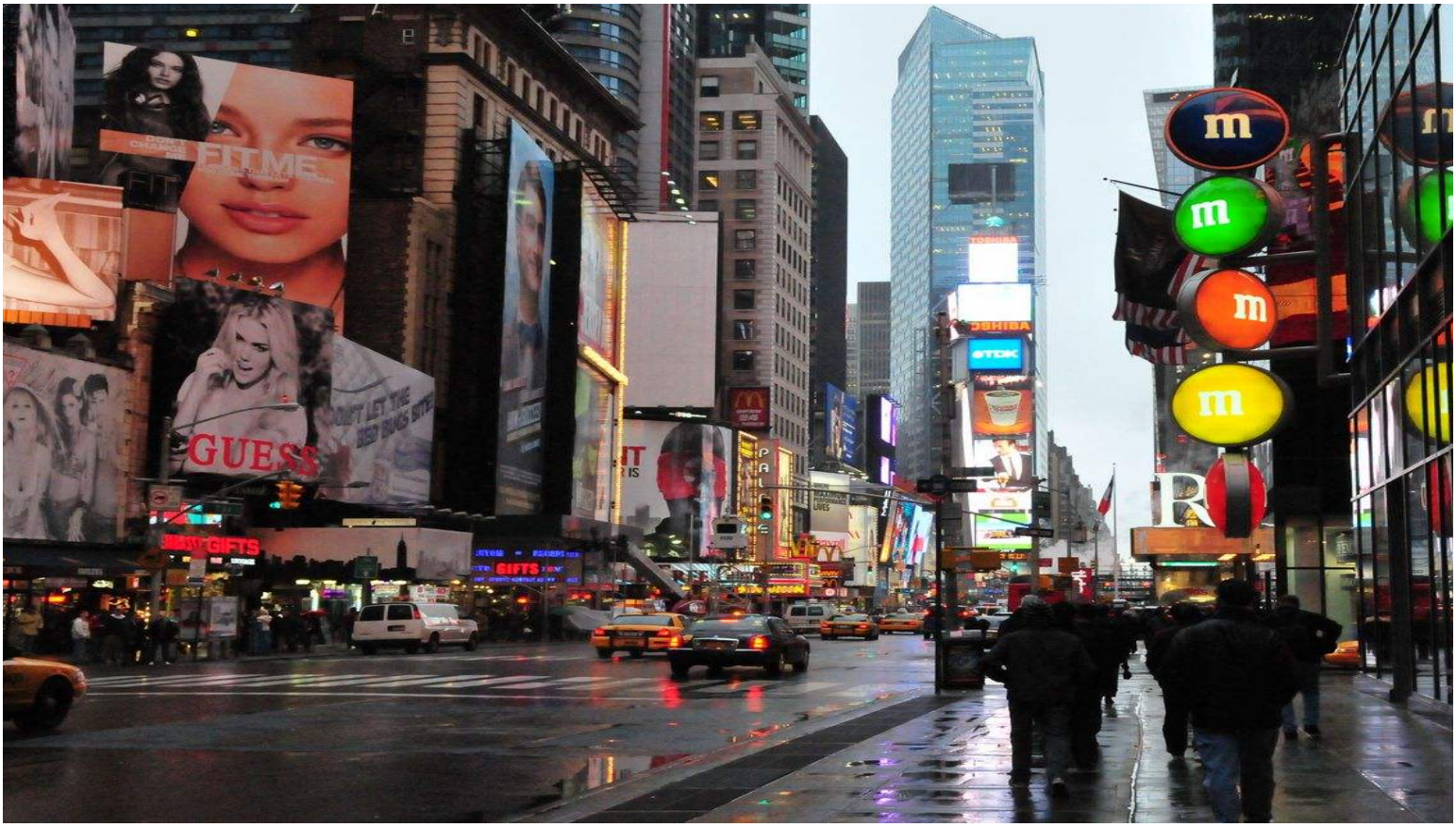


431 Data Visualization Final Group Report

Group Member

Yitian Xia
Rong Zhang(Renee)
Luis Pulgar

Background and Goal

For our final project we decided to use the New York City Taxi & Limousine dataset from the website (http://www.nyc.gov/html/tlc/html/about/trip_record_data.shtml) to see how taxi performed especially with the threat from ride-hailing companies, such as Uber and Lyft. Due to the large amounts of data, we decided to focus on Green Taxi and the data for the month of December from 2013 to 2017.

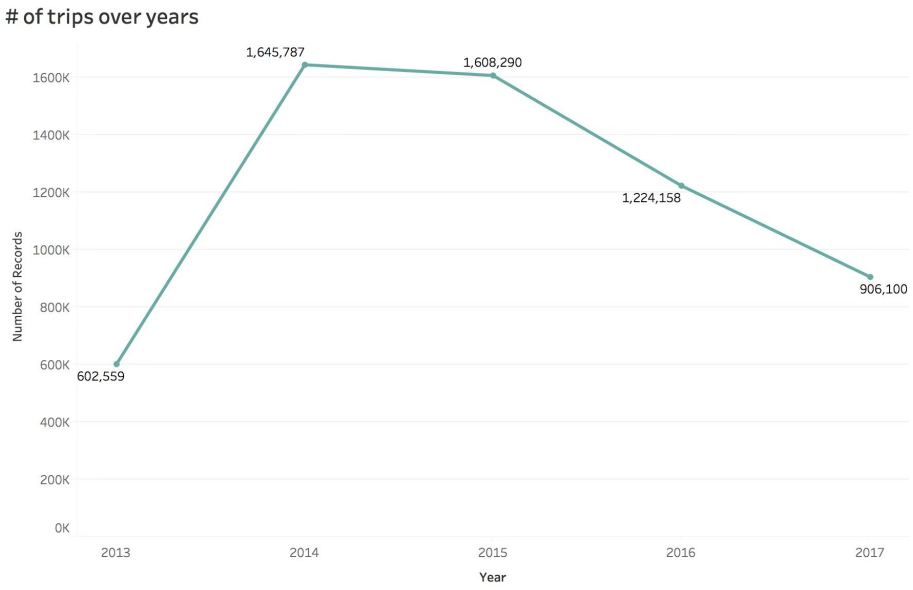
The goal of our project is to provide data-driven insights for the management of the Green Taxi company, including passenger's behavior and how the market is developing. We believe our findings could help the Green Taxi company increase their revenue and improve service quality by adjusting their taxi service.

By using Tableau and Kepler, we explore data and get business insights from following aspects as proposed: historical trip volume changes, traffic change within a day and even passenger behavior such as the trip distance, group size, tip amount, destinations and so on. We will show our findings in detail next.

Data visualization and insights

- **Trip Volume Change over 2013 to 2017**

  In the chart below, the horizontal axis represents the years and the vertical axis represents the number of pick ups during December. As we can see, in 2014 and 2015, the Green Taxi company has the heaviest volume, which means it developed very fast during that period. But after 2016, the number of pick ups drops drastically year by year (2014-2017, 44% drop in three years). We believe ride-hailing companies is the biggest reason. Based on our analysis, the number of cars in NYC that registered under ride-hailing companies, such as Uber and Lyft, increased dramatically since 2015 and the number of registered Uber drivers largely outnumbered the traditional taxi drivers. Although, the average time that an Uber driver is able to work may not be as long as the traditional taxi driver, Uber and other ride-hailing companies have already posed a serious challenge to the traditional taxi industry, not only in New York City but in almost every city where these companies have entered. From our perspective, this is the evolution of an industry, and we are witnessing how a traditional line of business is being disrupted by technology companies which offer easier ways for passengers to commute. From the chart we got, we predict that the number of trips for the Green Taxi company will continue to decrease in the foreseeable future. The question the company should answer is, where and when is it going to stop? At this pace, it can be almost completely out of business in 5 to 10 years. So it is the time for the company to think about the strategy: how to make a change to win the competition?

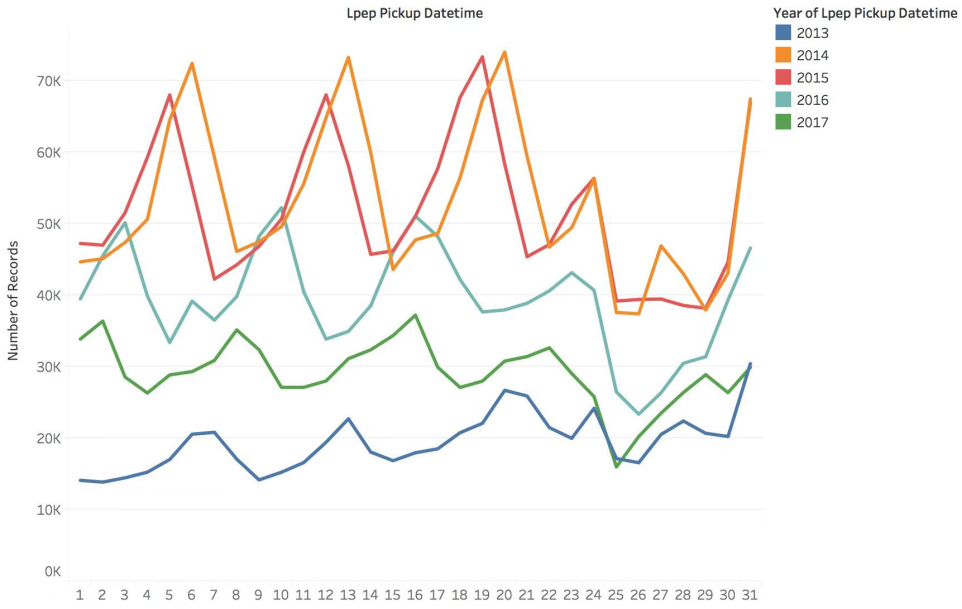


The trend of sum of Number of Records for Lpep Pickup Datetime Year. The data is filtered on Lpep Pickup Datetime Year, which keeps 2013, 2014, 2015, 2016 and 2017.

- **Daily Trips Change in December**

We created the chart below to see the trend of number of picks ups per day on December. The different line colors represent different years. The most obvious finding is on December 25th, the number of pick ups on that day drops dramatically. We believe it is because that date is the Christmas holiday and the majority of people stay home, following a high level on Dec 24th when people usually are with their families and friends. Another finding we identified is that there is a periodic variation. And the periodic change is 7 days. Take the red line (2015) as an example: The first peak is on the 5th, the second one is on the 12th and the third one is on the 19th. These peak days are Saturday and it's the same for 2014. For 2013, 2016 and 2017, most of those peak days are Friday or Saturday. So far, we can draw a safe conclusion that Fridays and Saturdays have the highest volume of pick ups during a week. We can suggest that is mainly because of people getting out of work and relaxing outside of their home with friends and family.

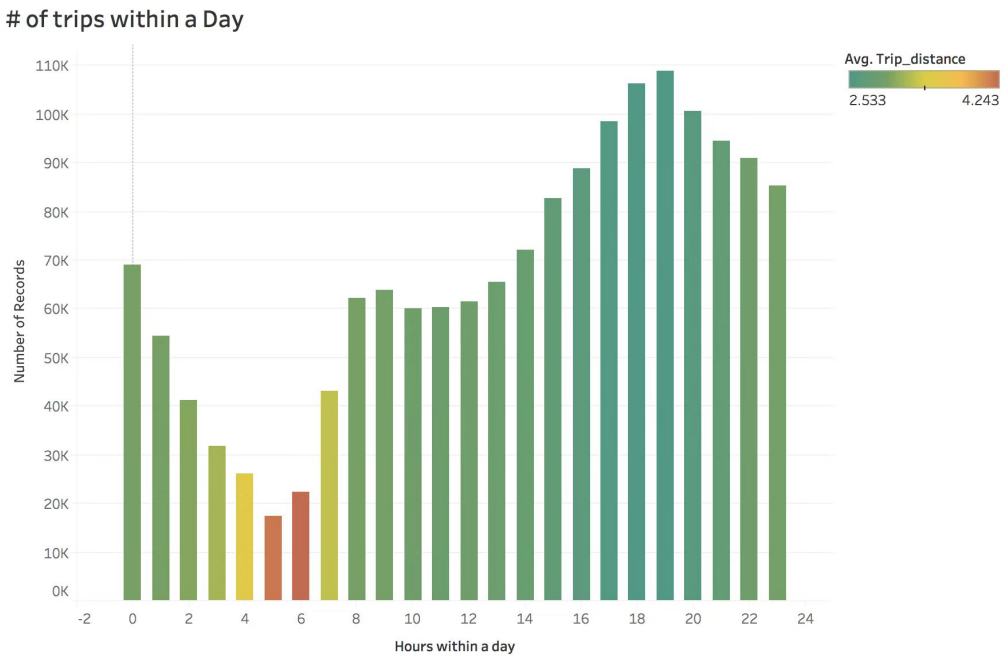
Ridership in December



The trend of sum of Number of Records for Lpep Pickup Datetime Day. Color shows details about Lpep Pickup Datetime Year. The view is filtered on Lpep Pickup Datetime Day and Exclusions (DAY(Lpep Pickup Datetime),YEAR(Lpep Pickup Datetime)). The Lpep Pickup Datetime Day filter excludes Null. The Exclusions (DAY(Lpep Pickup Datetime),YEAR(Lpep Pickup Datetime)) filter keeps 155 members.

- **Volume of pick ups within a day**

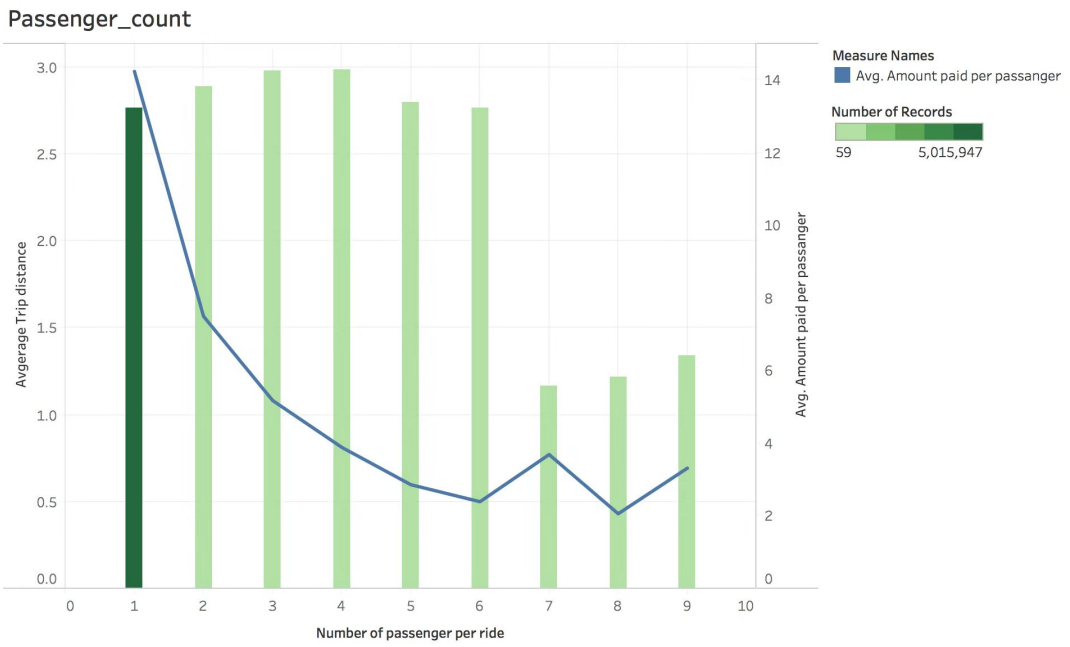
We created the chart to see how the number of pick ups is different throughout the day in a 24 hour span. The darkness of the color represents the average trip distance per pick up. We also set the year of pick up as the filter, so that the reader can check the variation in the different years. The chart tells us that 5am to 6am is the slowest time of the day for Green Taxi company, but the average trip distance is the longest. 6pm to 8pm is when the highest volume of pick ups occurs but to relatively short distances since most of them are for commute only. We suggest that the company should deploy more taxis during the evening, and less taxis early in the morning, and further investigate where people come from in the morning from such long distances so there are available taxis for the customers. This further exploration will be covered in later part when using Kepler.



The plot of count of Lpep Pickup Datetime for Lpep Pickup Datetime Hour. Color shows average of Trip_distance. The data is filtered on Lpep Pickup Datetime Year, which keeps 2015.

- **Relationship between Number of Passengers and Trip Distances**

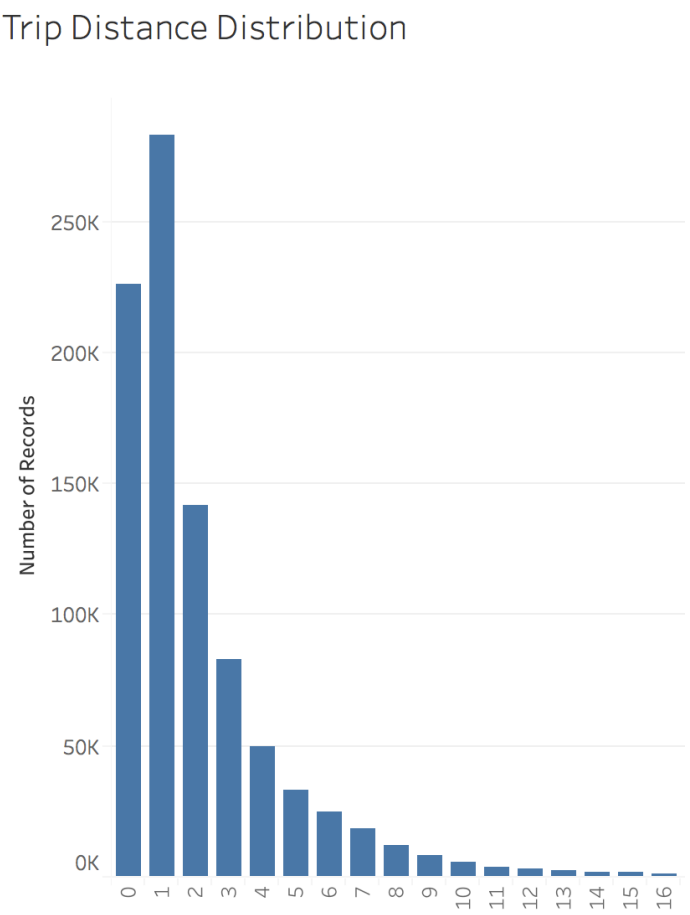
We created this chart to see how the number of passengers would influence the trip. The bar represents the average distance per trip, the darkness of color represents the number of pick ups for each number of passengers, and the line represents the average amount paid per passenger. From the darkness of the color, you can see that single passengers are the majority group size of trips and their trip distances are relatively long. From the bar chart, we can infer that when the number of passengers is below 7, passengers would choose to take a taxi when the trip distance is long. On the other hand, if the number of passengers is 7 or more, the trip distance becomes shorter, as passengers think the amount each person shared won't be much. Based on our analysis, we suggest the taxi company should put more small cars into the market, since single passenger makes most of requests.



The trends of average of Trip_distance and Avg. Amount paid per passanger for Passenger_count. For pane Average of Trip_distance: Color shows sum of Number of Records. For pane Average of Amount paid per passanger: Color shows details about Avg. Amount paid per passanger. The data is filtered on Lpep Pickup Datetime Year and Passenger_count. The Lpep Pickup Datetime Year filter keeps 2013, 2014, 2015, 2016 and 2017. The Passenger_count filter excludes 0.

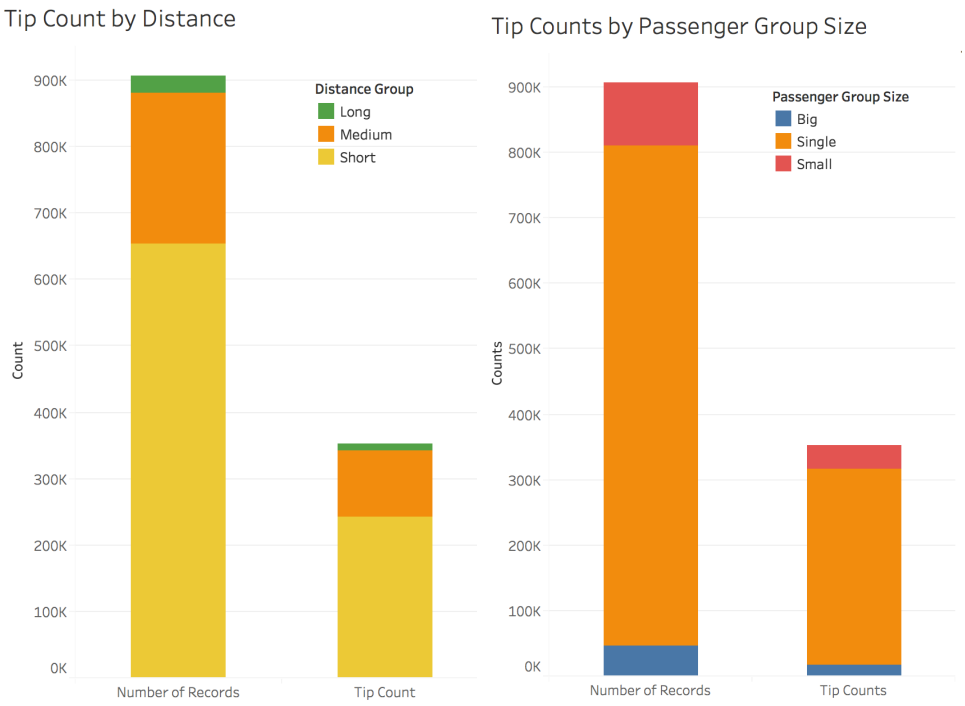
- **Trip Distance Distribution**

As the foundation for our next topic (Passenger's trip behavior), we explored the average trip distance and its volume first. We plotted a histogram and not surprisingly, found the distribution is right skewed, which means most passengers took taxi for short trip (distance less than 3 miles). It'd be interesting to see the locations of these trips, and deploy the right amount of taxis in the busy districts and the peak hours in order to satisfy every customer, which we do some preliminary analysis using Kepler in the later part.



• **Passenger Tip Behavior**

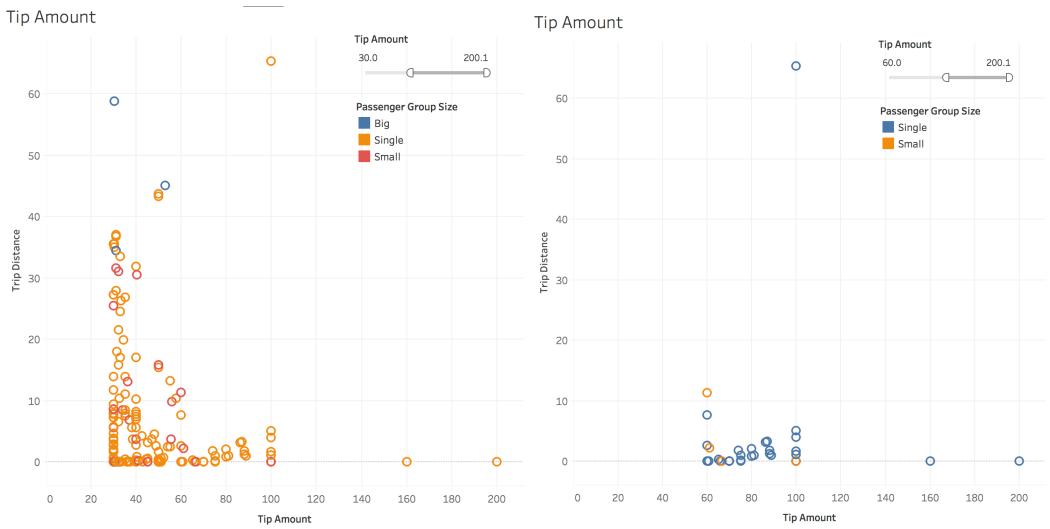
An interesting topic is what factors make passengers be more willing to give generous tips. Before doing any analytics work, our guess is the trip distance and passenger group size are the main reasons because tip amount normally is proportionate to fare amount (higher fare amount means more tip) and just like attendants serve dinners (more dinners of a table means more tip), more passengers served at a time indicate more tips. In the first step, we wanted to see whether distance and passenger group size impact the passengers' willingness to tip. So we created two new calculated fields in Tableau – Distance Group and Passenger Group Size. For Distance Group, we define it as: "Long" for trip distance over 10 miles, "Small" for trip distance no more than 3 miles and "Medium" for the middle range. And for Passenger Group Size, we define it as: "Single" for number of passengers equals to 1, "Small" for number of passengers between 2 and 4 and "Big" for number of passengers more than 4. Since most of the passengers don't tip, to see the proportion of passengers who tip within each group, we created our third new calculated field: Tip Count. If passengers tip, we define Tip Count as 1 and if passengers don't, we define it as 0. After that we imported 2017 December data and drew two bar charts as below:



From these charts, we didn't see apparent proportion difference among different segments, neither for distance nor for passenger group size. It indicates that these two factors don't have much impact on passengers' willingness to tip.

Then we explored the assumption: Passengers in larger group size and taking long distance trip are more likely to give big tips.

Here we plotted scatter plots to see the relationship between tip amount, trip distance and passenger group size. We defined big tip as amount over $30. So we found long trip distance slightly made passengers more willing to give big tips but it would not make people give extremely big tips (see chart on the right when we filter tip amount over $60). As for passenger group size, we didn't see any pattern indicating larger group size made passengers more generous to give big tips. One interesting finding is that for those who give extremely big tips (amount over $60), almost all of them had small group size (single or small) and took not a very long trip (below 10 miles). So it's quite unpredictable, maybe just because passengers are happy that day.

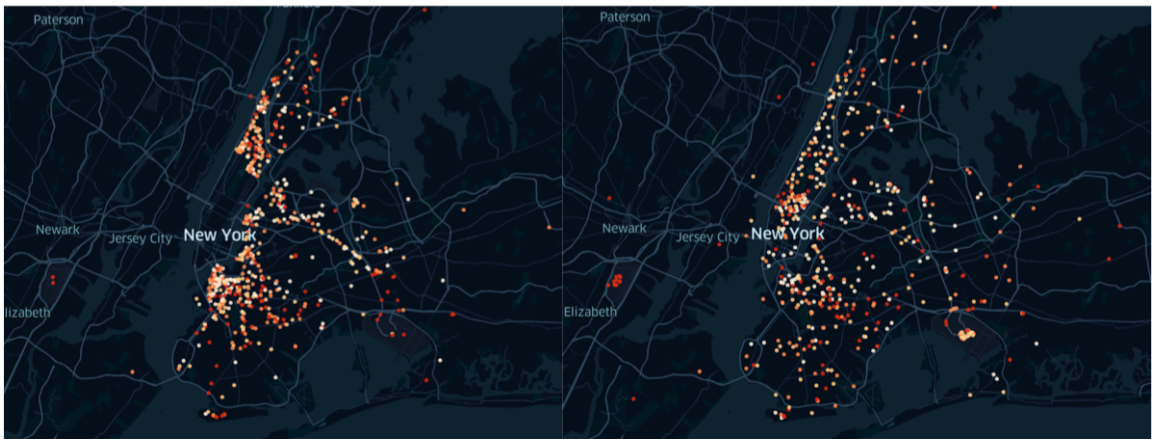


- **Trips within a Day (dynamic)**

Since our dataset included geographic information, we tried to do some cool stuff. We used Kepler to see the change of trips within a day in a more dynamic way. We selected a normal weekday Dec 2nd, 2015 to see how traffic changed. Each map below shows the traffic of a two-hour period so in total we got twelve maps. We used purple as pickup locations and red as drop off locations. As we can see, from 2am to 6am, the traffic volume is relatively low. Maps after 6am don't have many differences. But one thing we noticed is that from 6am to 4pm, people went from outskirt areas to Manhattan and Brooklyn but after 4pm, people went in the opposite direction as we found denser red points in outskirt areas after 4pm. Another interesting finding is taxis were only allowed to drop off but not pick up passengers in certain parts of Manhattan since we only saw red points in that area. To verify our guess, we searched and leant that Green Taxis aer only allowed to pick up passengers within some certain areas while Yellow Taxis have more freedom (good to know). Since Uber drivers can go to these area to pick up passengers, this makes Uber more convenient.

• **Long Distance Trips**

Since taxi fares depend on the trip's distance, taxi drivers can earn more if passengers take long trips. So it is worthwhile to see the routes of the long trips. The map on the left shows the pickup locations of long trips and the map on the right shows the drop off locations and the darker the color, the longer the trip. We can see that people are more likely to take a long trip from downtown areas to outskirt areas as pickup locations cluster while drop off locations are more disperse. Also, from the chart on the right we noticed two clusters, where the New York area airports are located. But on the left chart, we didn't see the same clusters. So we would say people was more likely to take long taxi trip to airports than from airports maybe because they were in a rush to catch flights so they had to take taxi and paid a lot. Another interesting finding is extremely long distance trips (red points) are more likely to start in outskirt areas compared with downtown areas as we can see from the pickup map on the left. So the company can think about launching long-distance trip promotions in the outskirt areas.



Design Choices

The most frequently used charts in our project is bar chart and line chart. Bar chart is suitable for comparison and line chart is used to see the trend. So that is why we chose line chart for number of trips over years, and trends for days of December for you can clearly see the changes over the years and days. We used bar chart to visualize the number of trips over different time within a day and to compare how number of passenger would influence people's behavior. We also used packed bubbles to visualize the number of passenger impact, but it still feels that bar chart is more intuitive and straightforward. The principle of data visualization is delivering business insights clearly. So simple is the best. Our interactive elements in this dashboard is # of trips within a day, where we set the years as the filter. The audience could check the difference between the years by dragging the slider. To explore the relationship between tip, trip distance and passenger group size, we used scatter plots because scatter was the best fit to show relationship. Since we had three elements in our charts, besides x and y axis, we used color to show group size. One creative method we used in our project is we segmented distance and passenger counts and turned these numerical variables into categorical variables (Group Size and Passenger Group Size) and made it possible for us to do in-depth analysis. When seeing the proportion of passengers who tip within each group (Distance Group and Passenger Group Size), we tried to use table but since Passenger Group Size is not a dimension, Tableau could not do what we desired. So we changed to bar charts which turned out work well also.

When using Kepler to show the animation of trip change, first we considered using arc so Kepler could show routes. But after we tried, we found the visualization was messy and people could hardly see how trips went. Then we considered using a heat map but because of the high density of data, heat maps didn't change much as time elapsed. So finally we chose points and used two maps to show the pickup locations and drop off locations.