

**Divvy Bike Analysis: Q1 FY 2017**

**By: Mark Van Wormer, Jose Coliz, and Jaime Bonilla**

**Purpose**

The purpose of this project is to explore Divvy Bike a relatively new bike sharing company based in Chicago. Our end goals of this project are to obtain a better understanding of how Divvy Bike operated in the beginning quarter (Q1) of 2017 and to gain further insight into how they operate on a day-to-day basis. We also want to explore the different aspects of their daily operation: including number of users, number of stations, location of stations, and etc.

**Background**

**About:**

Divvy Bike is a Bicycle sharing program in the city of Chicago and two of Chicago’s adjacent suburbs which are Evanston and Oak Park. Divvy is operated by a company called Motivate which is a New York City based bike sharing company that operates bike sharing services across the country. Divvy began operation on June 28, 2013 and currently has 580 stations in an area from 87th St to the south, Central St to the north in Evanston, and Harlem Ave to the west in Oak Park. Divvy has 5800 utility bicycles which are bikes that are designed for practical transportation as opposed to recreational purposes. Divvy Bikes are also painted in Chicago Blue to embrace a sense of local pride amongst users.

**Variables:**

On the Divvy Bike website they provide data for analysis by year and quarter. They are also generous enough to provide us with a reference sheet so we can analyze the dataset. The data set is accessed through a CSV (comma separated values) format. They provide you with three different excel files per quarter/year which are the Q1 trips, Q2 trips, and the Stations for Q1/Q2.

Since the files are so large we decided to stick to just Q1 which had 12 variables which are listed below as:

1. Trip\_id: ID attached to trip taken
2. Start\_time: day and time trip started in CST
3. Stop\_time: day and time trip ended in CST
4. Bikeid: ID attached to the bike
5. Tripduration: time of trip in seconds
6. From\_station\_name: name of station trip originated
7. To\_station\_name: name of station trip ended
8. From\_station\_id: ID of station trip originated
9. To\_station\_id: ID of station trip ended
10. Usertype: “Customer” - 24 hour pass; “Subscriber” - annual membership
11. Gender: gender of rider
12. Birthyear: birth year of rider

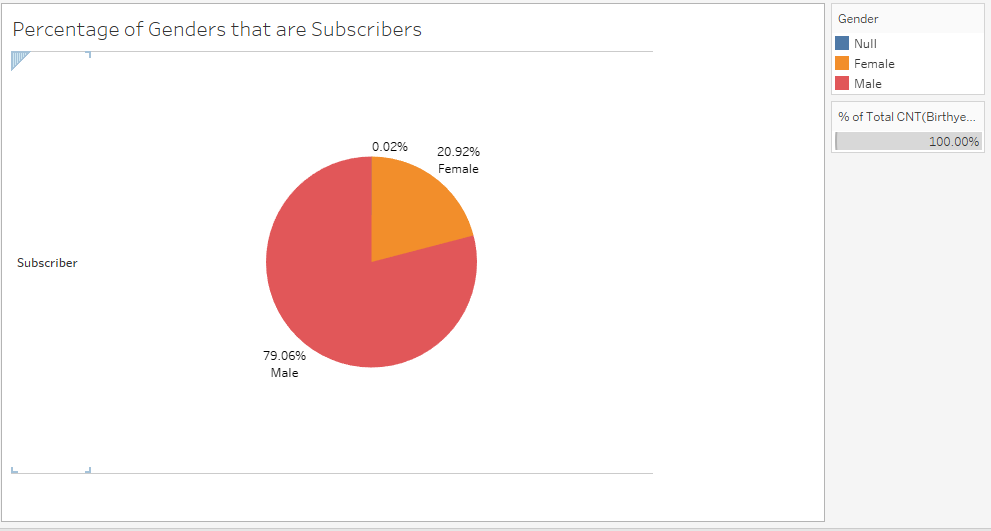
The data also has over 400,000 rows in fact I believe it was 431,691 to be exact. Needless to say it was quite an undertaking for analyzing one dataset so to make sure it is clean for analysis we went divided up the rows amongst the three of us to go through and make sure that the data was ready for the visualizations. Now of course there are some null values in the data set as from discrepancies in Divvy’s data collection The aggregation of the data was determined by the type of visualization also by what we were visualizing, meaning we used a mixture of sum and average to gain further insight into what data was being presented so we could gain valuable information from it. What we noticed in this data set is that there are both discrete and continuous variables which made for an interesting analysis of the data set because it presented new challenges that we had to understand especially when we were cleaning the data preparing it for the creation of our visualizations.

**Visualizations**

This project called for nine visualizations of which six had to be different from the rest. Divvy’s data set was large enough to where it was possible to come up with a variety of different visualizations to explore this data set. We decided to use: Pie Chart, Geographic Map, Heat Maps, Bubble Charts, Line Charts, and etc. to show the different aspects of the Divvy Data.

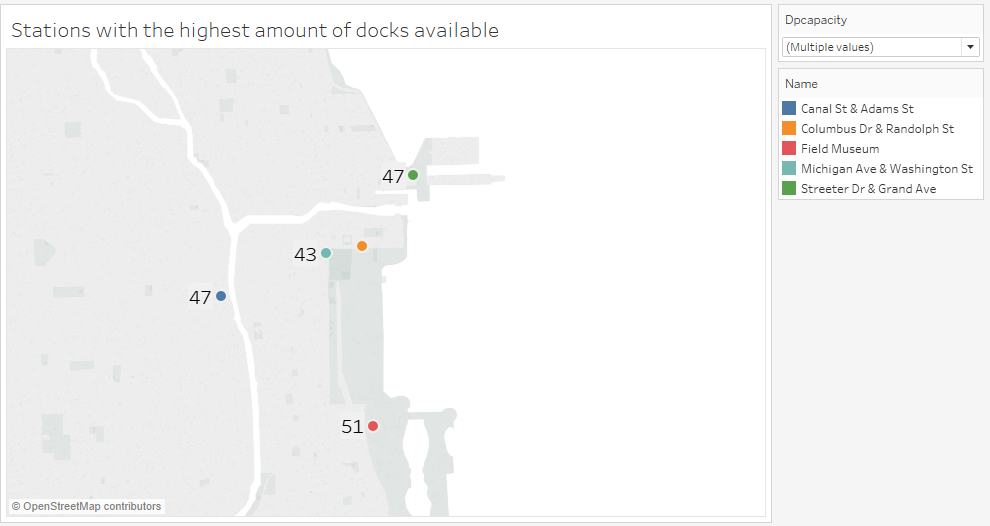
We used different charts in a way that we saw best fitting of the data presented to us from the data file Divvy provided on their company website. Each graph displays topical information that we believe could be helpful to Divvy as a whole and will better its chances of gaining new insights and new user bases. Reasons why we chose the visualizations we did was from a fitness standpoint as we wanted to determine which visualization would best tell a story with the data.

**Pie Chart: User Genders**

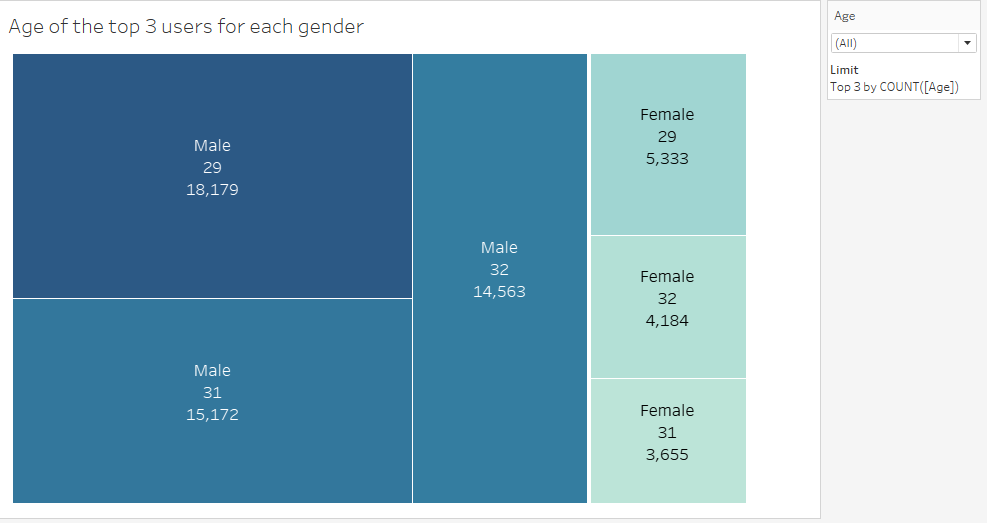
****

This is what the Divvy Bike user makeup looks like for Q1 in 2017. We notice that a majority (79.06% of users are male while only 20.92% are female). We also notice that 0.02% of users have not disclosed their gender. We decided to use a pie chart so we could see the demographic of their users as a whole and which is useful from a marketing to gain an insight into possible new customers but to also help retain current customers.

**Geographic Map: Largest Stations**

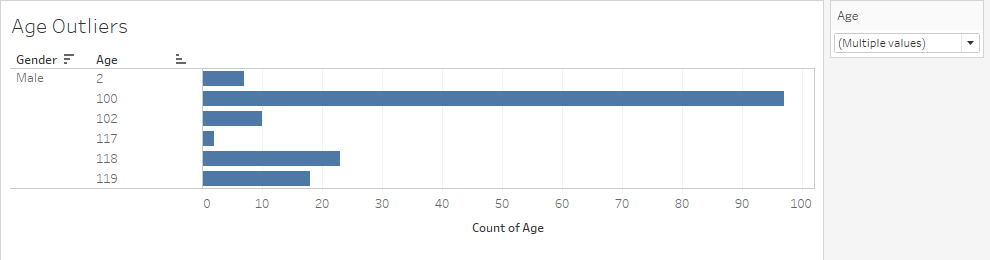
****This is a geographic map that shows the stations with the highest amount of docks available. For example at the field museum there are 51 docks or spaces where users can park their rented bikes. We chose a geographic map because it gives a good representation of where the biggest stations are in proximity to one another and it also gives an insight into further expansion of new stations or upgrading existing stations.

**Heat Map: Top 3 Age of Users both Male and Female**



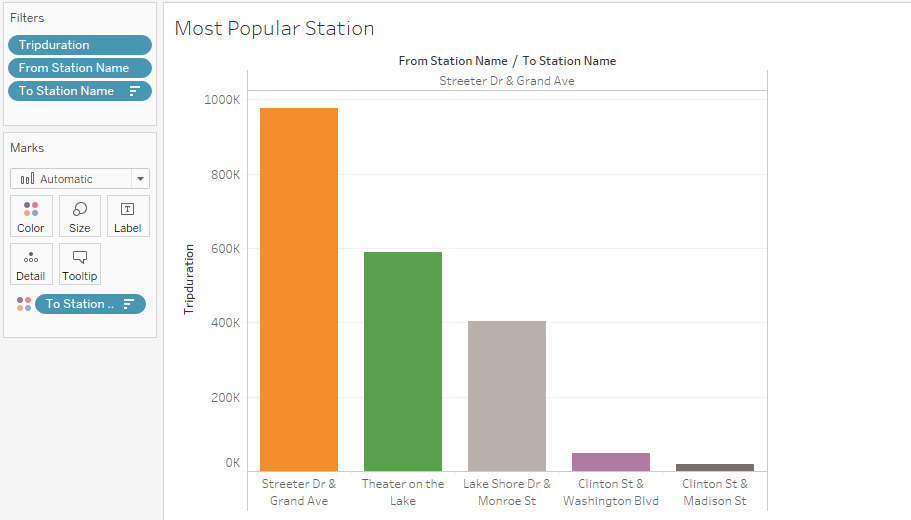
Similar what we did for gender with the pie chart we wanted to find what ages were the users who were riding Divvy bikes based on the data we found that Divvy users were generally in their upper 20s and low 30s. We chose a Heat Map to represent this data because Heat Maps are generally good for rank and showing quantity by space (how big the sector is). This shows us that most of the users are in fact males aged 29 and we also see that females are smaller since they make up less than 21% of Divvy users.

**Horizontal Bar Chart: User Age Outliers**



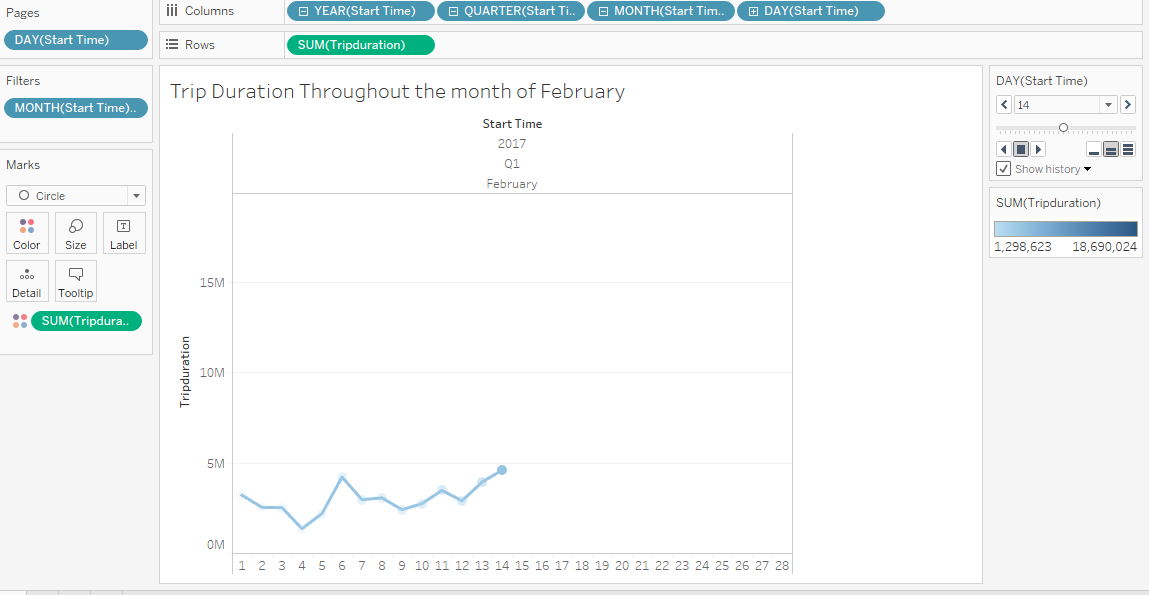
What we noticed while scrolling through the data set is that many people input or indicated a false birth year while signing up for an account. I don’t know of many two year olds or people in their hundreds riding bikes through the streets of chicago. This Horizontal Bar Chart not only helped us find the outliers since there was such a small grouping of them but it also counted how many times that age popped up given registration of users and you notice that close to 100 people are 100 years old. This can help Divvy change a policy to make it mandatory to put in a “correct” birth year because two year olds can ride bicycles.

**Vertical Bar Chart: Most Popular Bike Station**



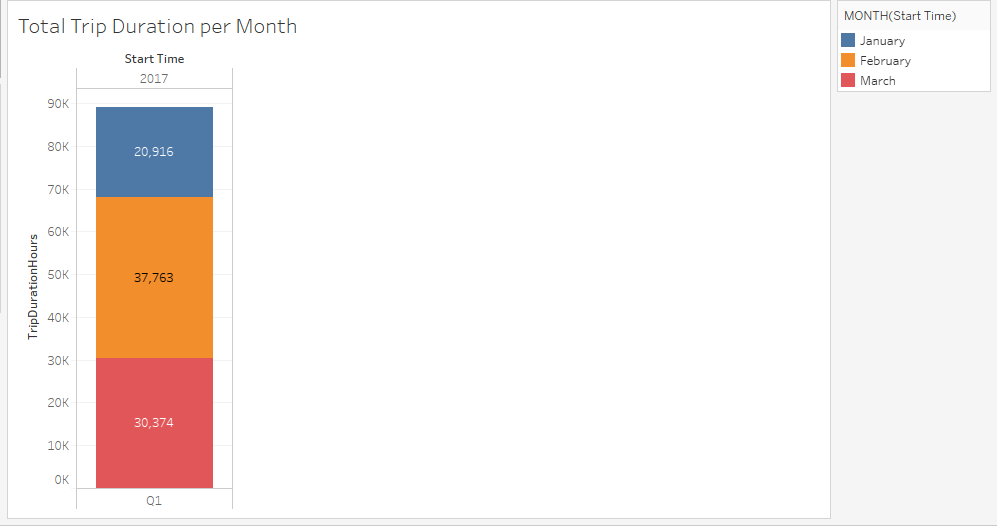
With this vertical bar chart we wanted to find out which station was the most popular through the trip duration. What we found out was that Streeter Dr & Grand ave was the most popular station with almost one million seconds in trip duration. We decided to use a bar chart because they are good in comparing things between different groups. In our case we compared the trip duration from the starting station to the ending station.

**Motion Chart: Total Trip Duration in February**



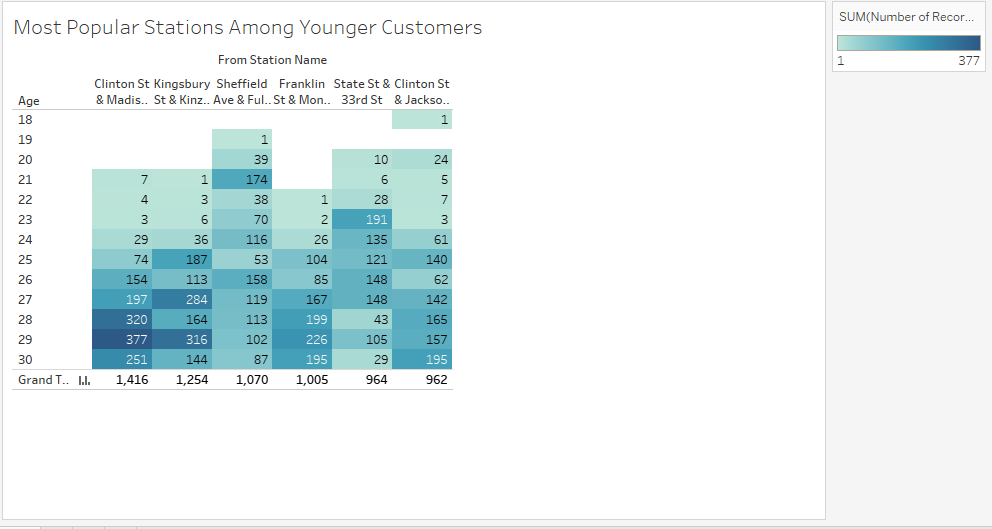
This is a motion chart that shows the trip duration throughout the month of February. What we found out was that during the month of February the highest trip duration was on the 19th. We used a motion chart because it allowed us to see the trip duration of each day for the month of February.

**Stacked Bar Chart: Trip Duration (in hours) per Month**

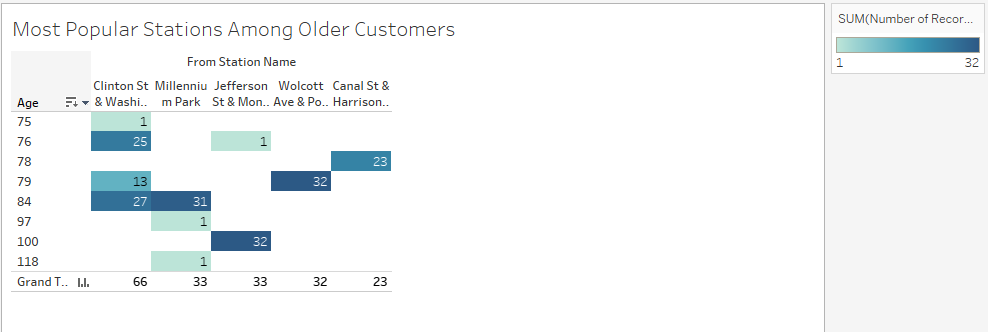
****

After finding the time duration for the month of February for each day we decided to find the total trip duration for each month. For the month of January the total trip duration was 20,916 hours and for March the total trip duration was 30,374 hrs. The month that had the most total trip duration was February with 37,763 hours. We decided to use a stacked bar chart because we wanted to compare the total trip duration of all three month and show each separate month total trip duration.

**Highlighted Table: Most Popular stations for younger customers**

****

With this highlighted table we wanted to find out what the most popular stations for younger customers was. For younger customer the range was from 18 to 30 years old. What we found out was that Clinton St & Madison was the most popular stations for younger customers. On top of that, the customer of age 29 had the most go to Clinton St & Madison stations with 377. We used a highlighted table to see which stations was the most popular but to also see what age group made the majority of that.

**Highlighted Table: Most Popular stations for older customers**

Just like the other highlighted table we also wanted to find out the most popular stations for older customers. The age range we used for old customer was 75-118 years old. The stations that was the most popular for older customers was Clinton St & Washington with 66 customer in total. We decided to use a highlighted table in order to compare the total customer for each stations while seeing the most popular age for older customers. This information is useful to us because maybe we can offer better accommodations to elderly riders if need be.

**Conclusion:**

After a thorough review of the Divvy bike data set we were able to gain new insights into how they function as an organization. After sifting through 431,691 rows of data we found that all 12 variables were crucial into the set up the various visualizations that we needed to create for this project. Challenges were faced in deciding which visualizations would best tell a story with the data that we were given. Even though a lot of the data provided by Divvy was qualitative we were able to take the quantitative data and make useful visualizations from them that showed a range of valuable information from demographics like gender and age, to various geographic information like what distances were travelled, as well as seasonality which is always important for a company that operates bikes in Chicago during the winter season. In the end we learned a lot of new ideas on how Divvy works and operates around Chicago.