DSC 465 – Data Visualization

# Homework 4

PKeener
3-7-2021

## Data Set

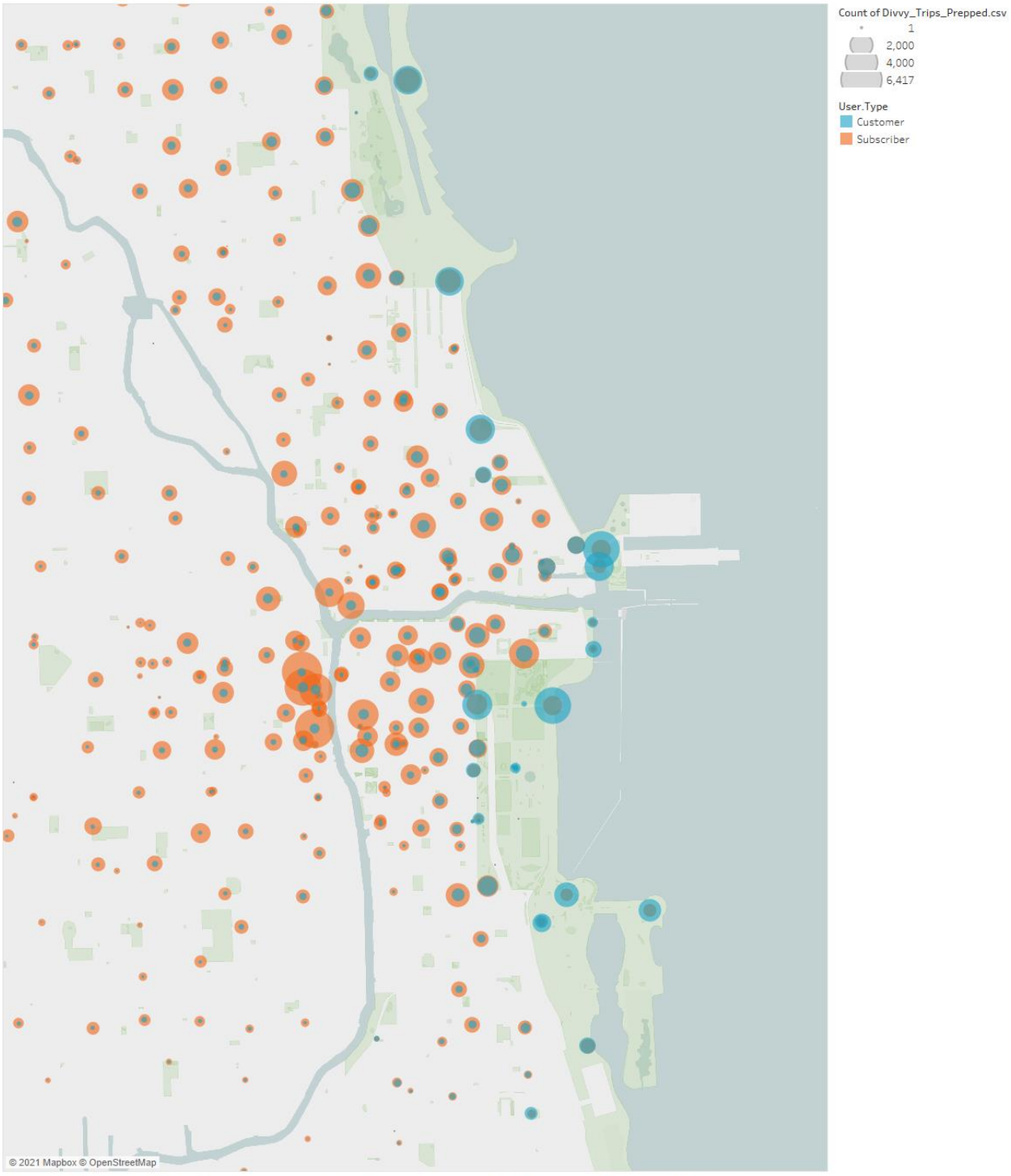
The data set used for this homework is the Divvy Dataset which contains data for Divvy rides spanning several years. To make this data set more manageable with the chosen visualizations, it has been trimmed from 20 million records down to 500,000. There were several new features created: straight-line distance, calculated as the distance between longitude and latitude points, and days, years, and months of the year. With this enrichment, the dataset was then filtered to remove 97 NAs.

The final data set contains 27 variables with 499,903 observations.

## Problem 1

Problem 1 used the project data set (described above) and required the use of any technique from the latter half of the class, including geographical data. For this assignment I opted to use the geographical data.

Location Primary Customer Type




Map based on From.Longitude and From.Latitude. Color shows details about User.Type. Size shows count of Divvy_Trips_Prepped.csv. The data is filtered on From.Station.Name, which has multiple members selected. The view is filtered on User.Type, which keeps Customer and Subscriber.

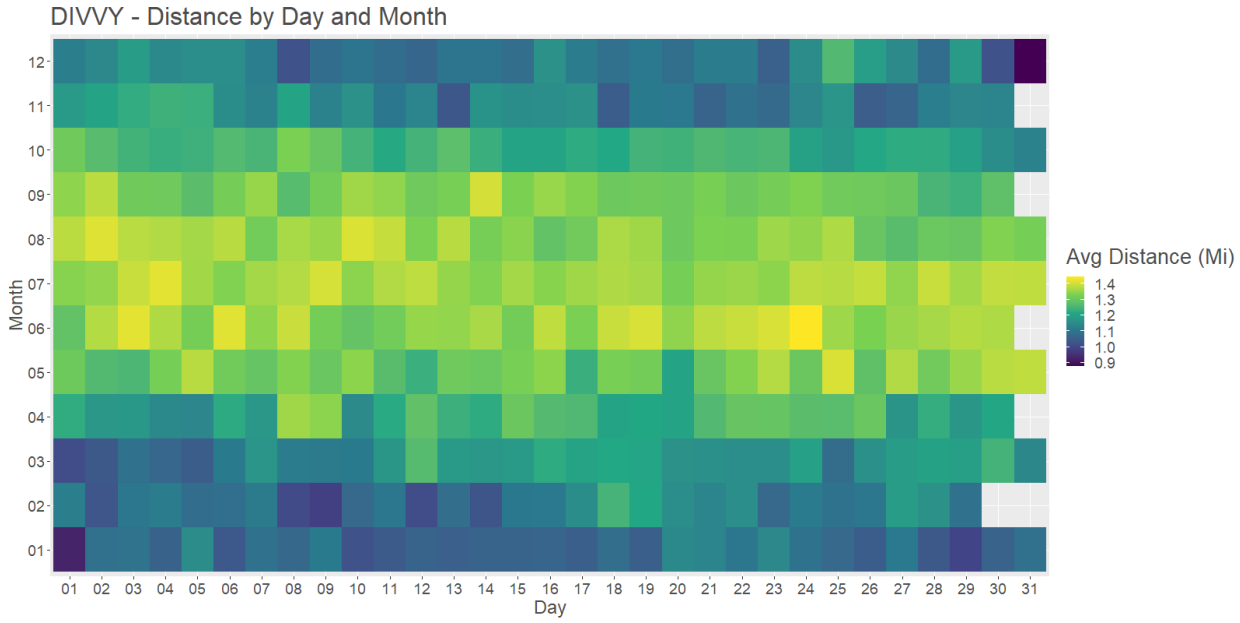
The above chart contains glyphs at the location of divvy stations. The glyphs are divided into two colors: orange for Subscribers and blue for Customers. Subscribers are those who have a monthly or yearly subscription to the service (which allows for collection of additional data like gender and age), while customers are users who purchase a single ride pass (and thus lack the additional data).

One interesting item that this map shows is that customers are the primary users along the lake and Magnificent Mile, which subscribers dominate elsewhere. This may have several reasons: those without subscriptions may not live in the city (tourists), and the areas where customer dominate are the city's tourist areas. Additionally, the large concentration of subscriber hotspots- near where the rivers meet- is next to Ogilvy and Union train stations, indicating heavy use by commuters. Anecdotally, I used to work at a building next to Ogilvy so I would divvy from Lincoln Park to work.

This map was designed in Tableau. The longitude and latitude columns from the data set were used for column/row and represent the station locations. I then set color by user type (which subset the glyphs so two appear at each location). I also set the alpha to 50%, which allows the user to see both glyphs when the additional factor is accounted for (trips from station) and allows easy differentiation between stations with a predominant customer type. The final step is to add the additional factor – to change the size of the glyph based on the number of rides.

## Problem 2

This problem required another visualization to be created, using any graph including week 4. For this purpose, a heat map of a time series was chosen.




This heat map shows the distance by day and month, averaged over each year of data. The colorblind color pallet was chosen because, interestingly, it created excellent distinction that highlighted the differences between the months. It is clear from the graph that distance increases during the months with mild weather (5-10) and reach their apex in months 6-9, which is the summer. It also appears that part of the day of the month does not have any influence on the distance used.

This graph was created in R. The dataset had to be subset and summarized (using 'mean') to create the bins used. After this, geom_til() was used to specify the heat map, and scale_fill_viridis_c() was used to set the colors. The theme was customized to include larger font sizes as well as customized axis labels.