Analysis of Divvy Data

Samad Patel

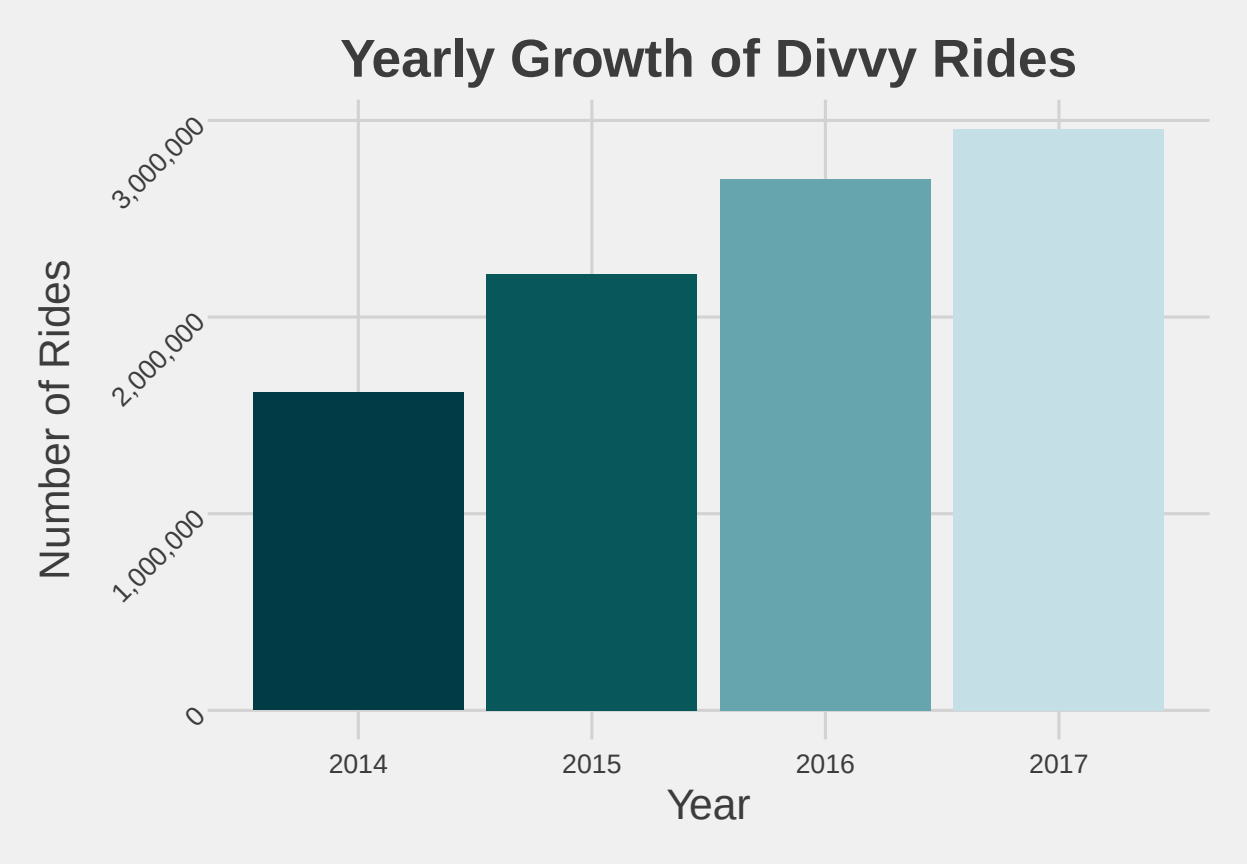
6/29/2018

The purpose of this document is to conduct a exploratory data analysis of Divvy's Chicago Bicyle Data. This will be broken down into basic EDA based on some of the demographic data, and then spatial analysis.

The data is already clean, meaning that there are no missing values, duplicates, or unusually coded classes. We can hop straight into the visualization.

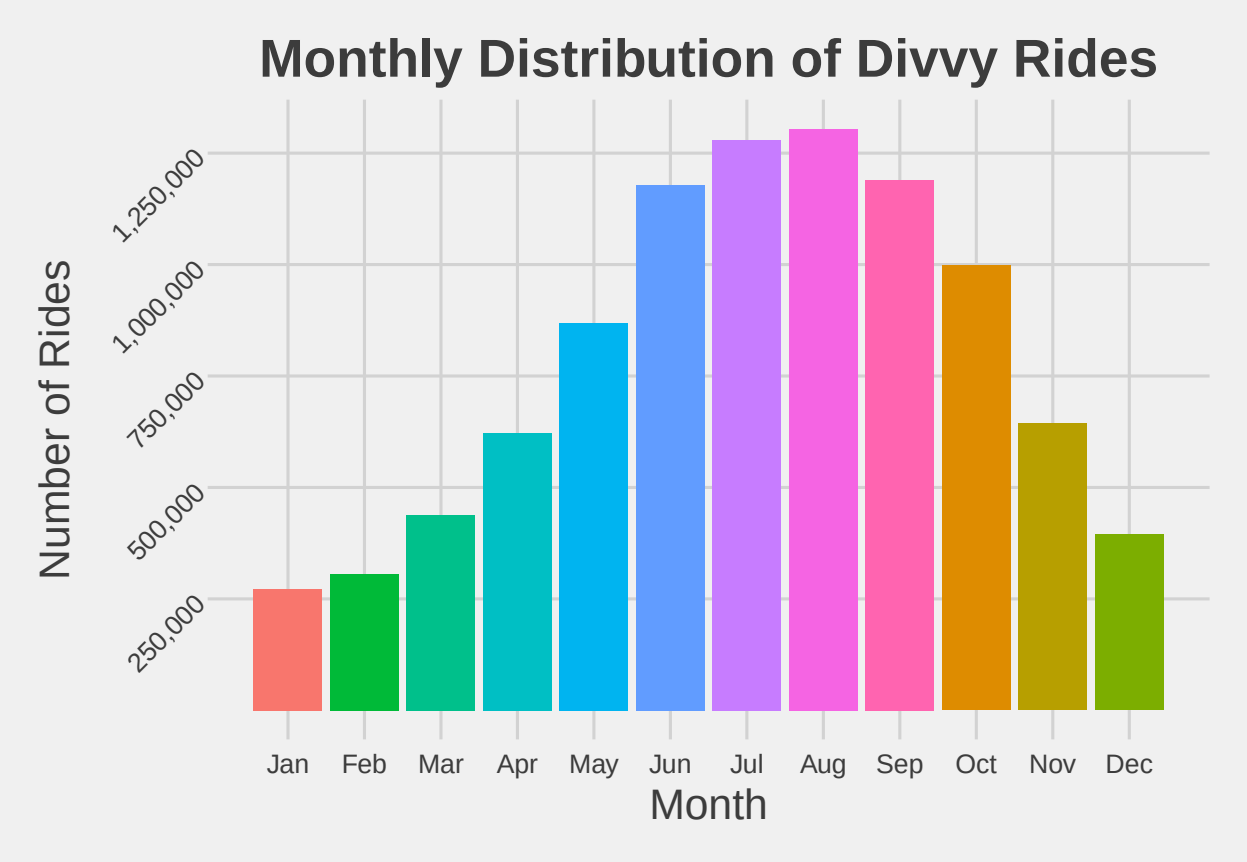
Basic EDA

Year



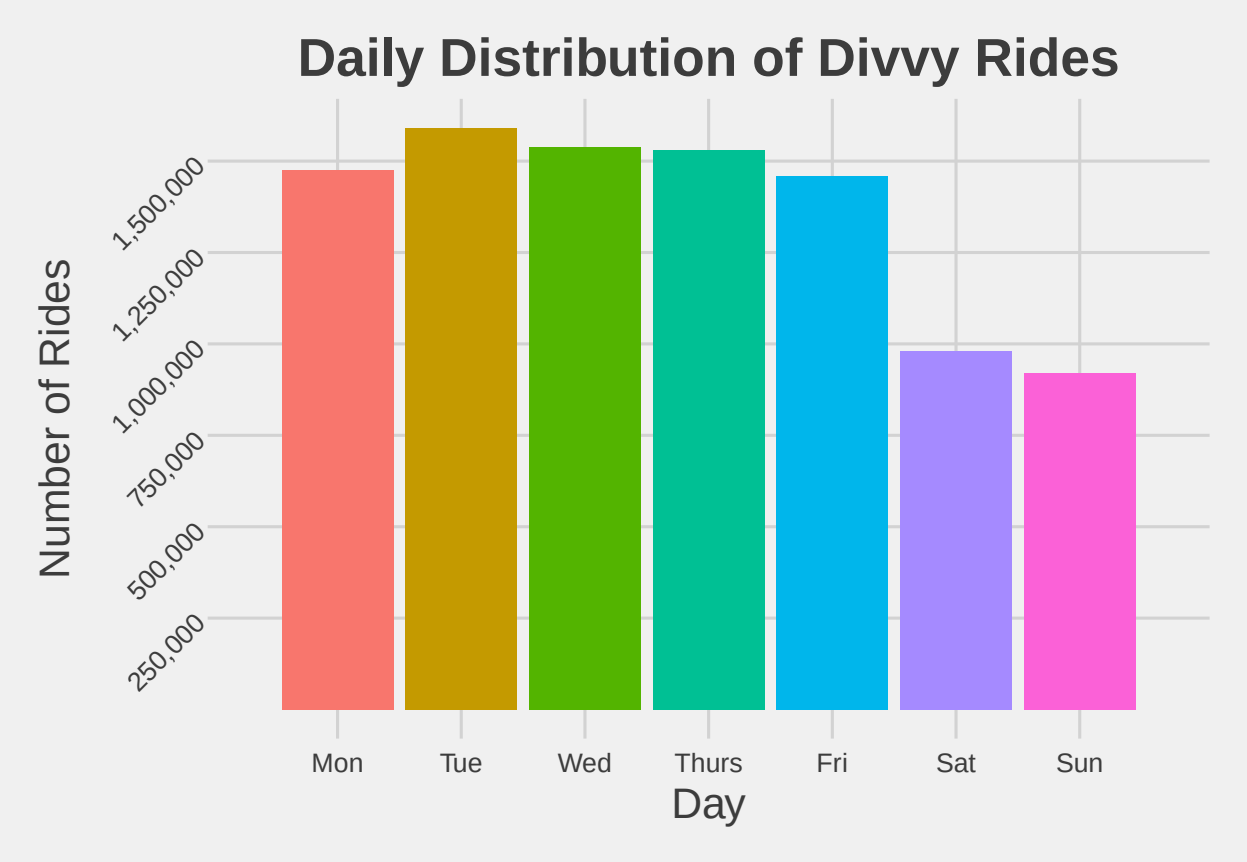
We observe that Divvy is improving year by year in number of customers.

Month



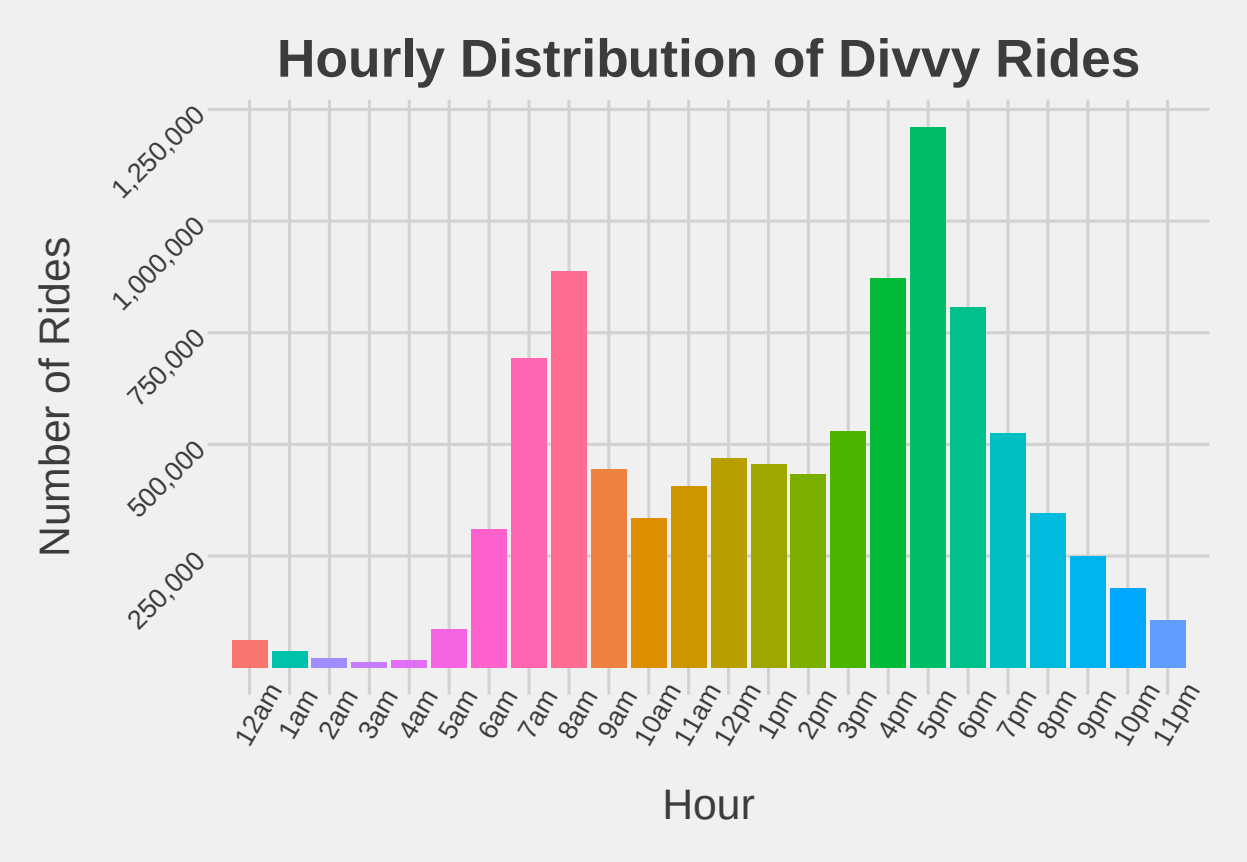
As expected, more individuals are riding Divvy's bikes in the summer, and the fewest are riding in the Winter. There is a gradual increase in riders from spring to summer, and a gradual decrease from fall through winter.

Day



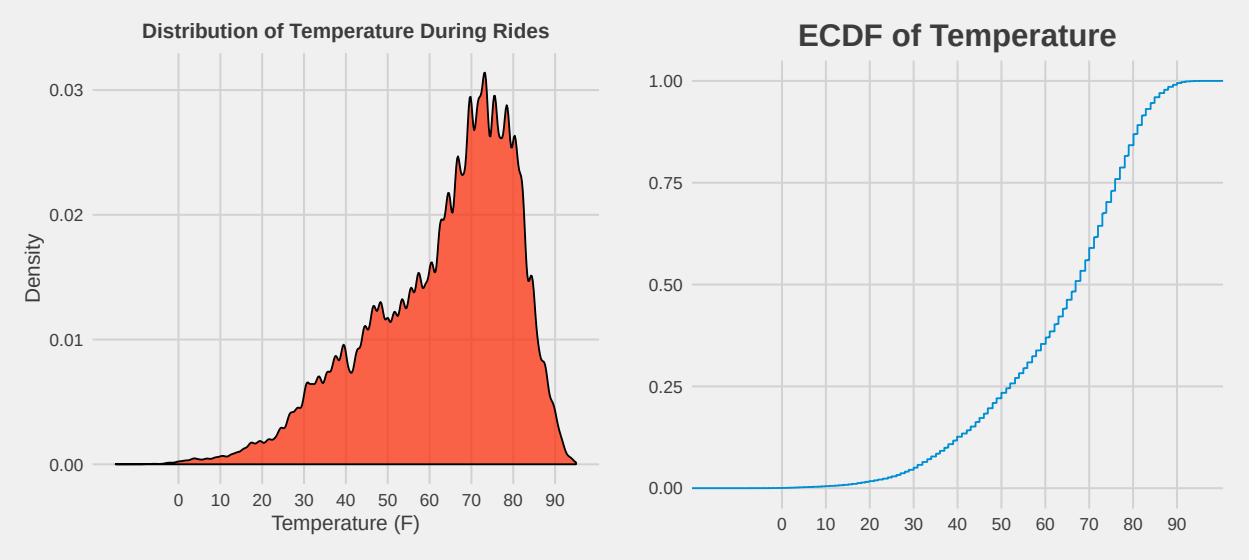
We can observe that Divvy has more traffic during the weekdays more than the weekends. This lends evidence toward the idea that many of the users are riding to go to and from work or school. Monday and Friday have slightly fewer riders than the middle of the week, which can be explained by the fact that a decent number of people work 4-day work weeks.

Hour



We can observe that most of the users are riding in the hours leading up to work (6am-9am), and after work (4pm-7pm). We observe spikes consistent with those who work the typical 9-5. The first large spike is at 8am, when they're aiming to clock in at 9am, and the second large spike is at 5pm, when they're clocking out.

Temperature



As one would expect, warmer weather is more comfortable to ride in. Over 50% of all rides are taken over 65 degrees. There appears to be a drop-off in users over 85 degrees, so 65-85 appears to be the sweet spot.

Snowy Chicago has hardly any riders - around 5% of rides occur when the weather is 30 degrees or below. That's to be expected, as that weather is extremely uncomfortable to ride in.

Spatial Analysis

Busiest Stations

Divvy works by allowing users to ride between stations. The code below confirms that there are 656 unique stations.

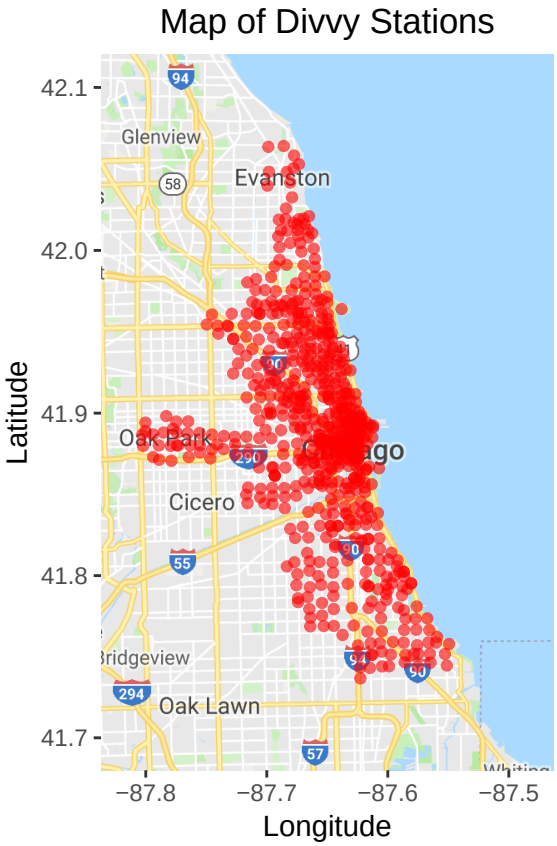
```
length(unique(bike$from_station_name))
```

```
## [1] 656
```

```
length(unique(bike$to_station_name))
```

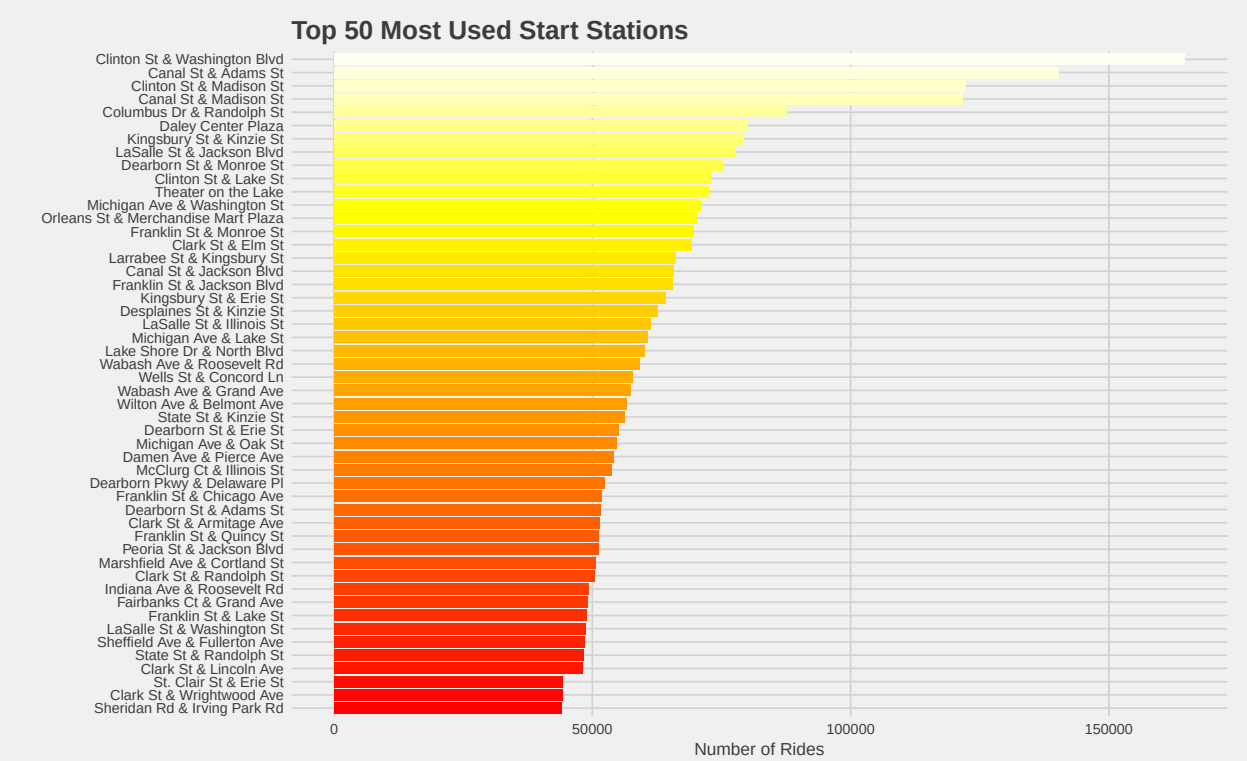
```
## [1] 656
```

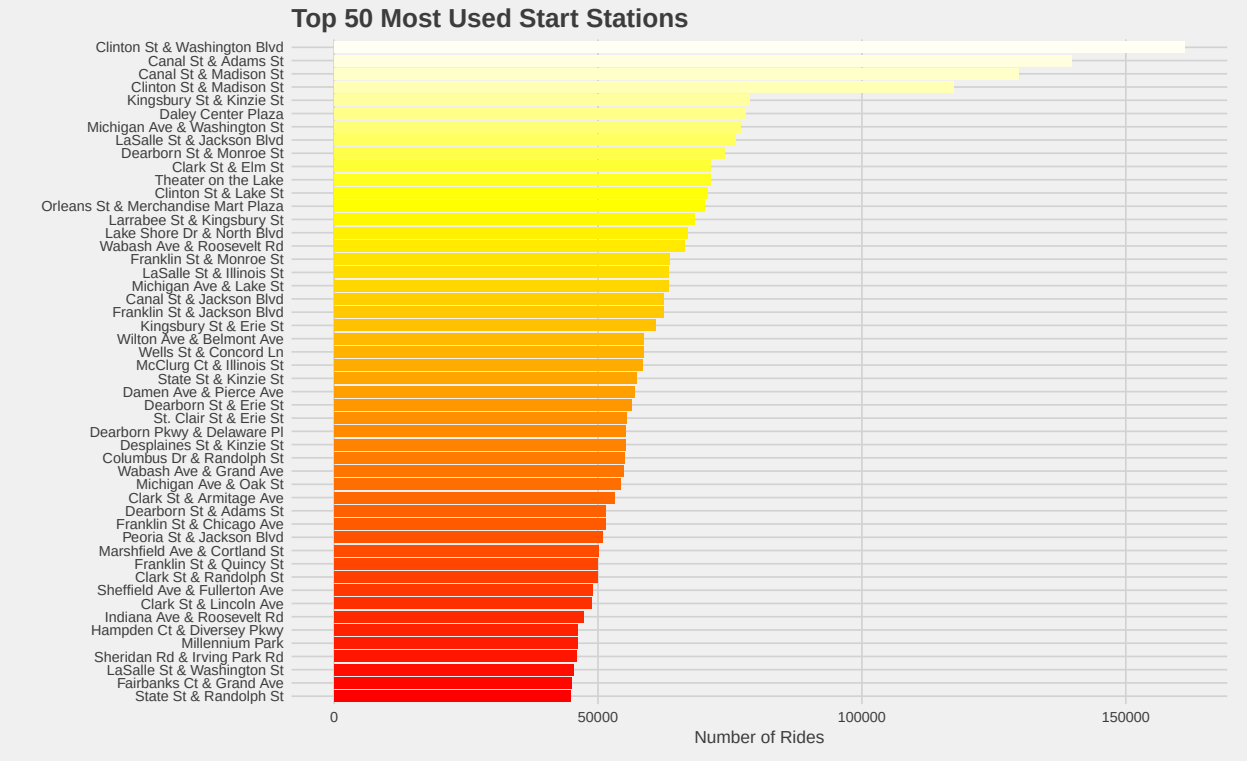
We can visualize where, on a map, these stations are scattered.



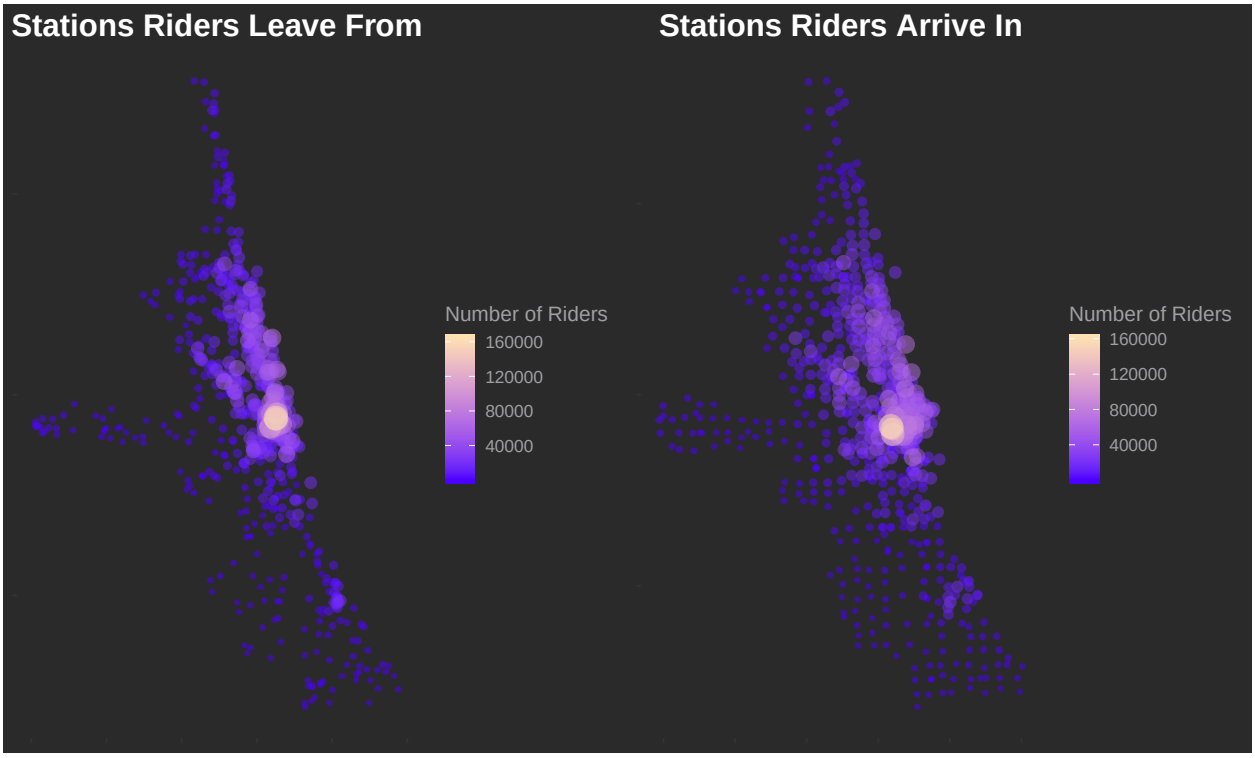
There appears to be a higher concentration of Divvy stations Downtown. This must mean that most of Divvy's users are riding Downtown. Divvy has the following pricing model: single rides are $3 for every 30 min, and subscriptions can be purchased at $15/day or $99/year for different features. Depending on the usage nature of the rides in Downtown, there could potentially be a better pricing model available.

Not all stations are created equal, so we can confirm whether or not Downtown truly is the most busy. Let's observe which stations users leave from the most, and which stations users arrive at the most.

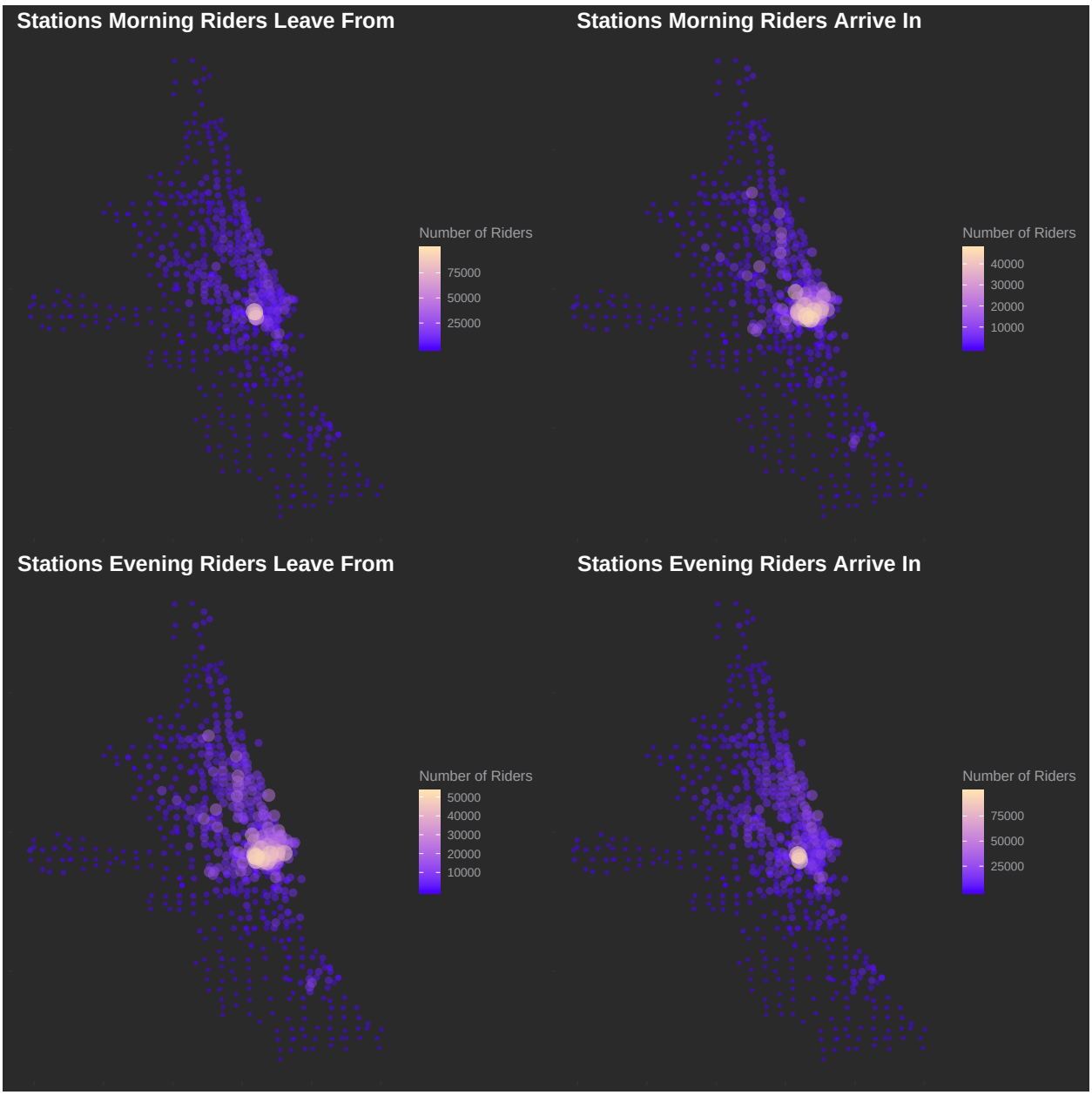




Here's a map to visualize the same information:

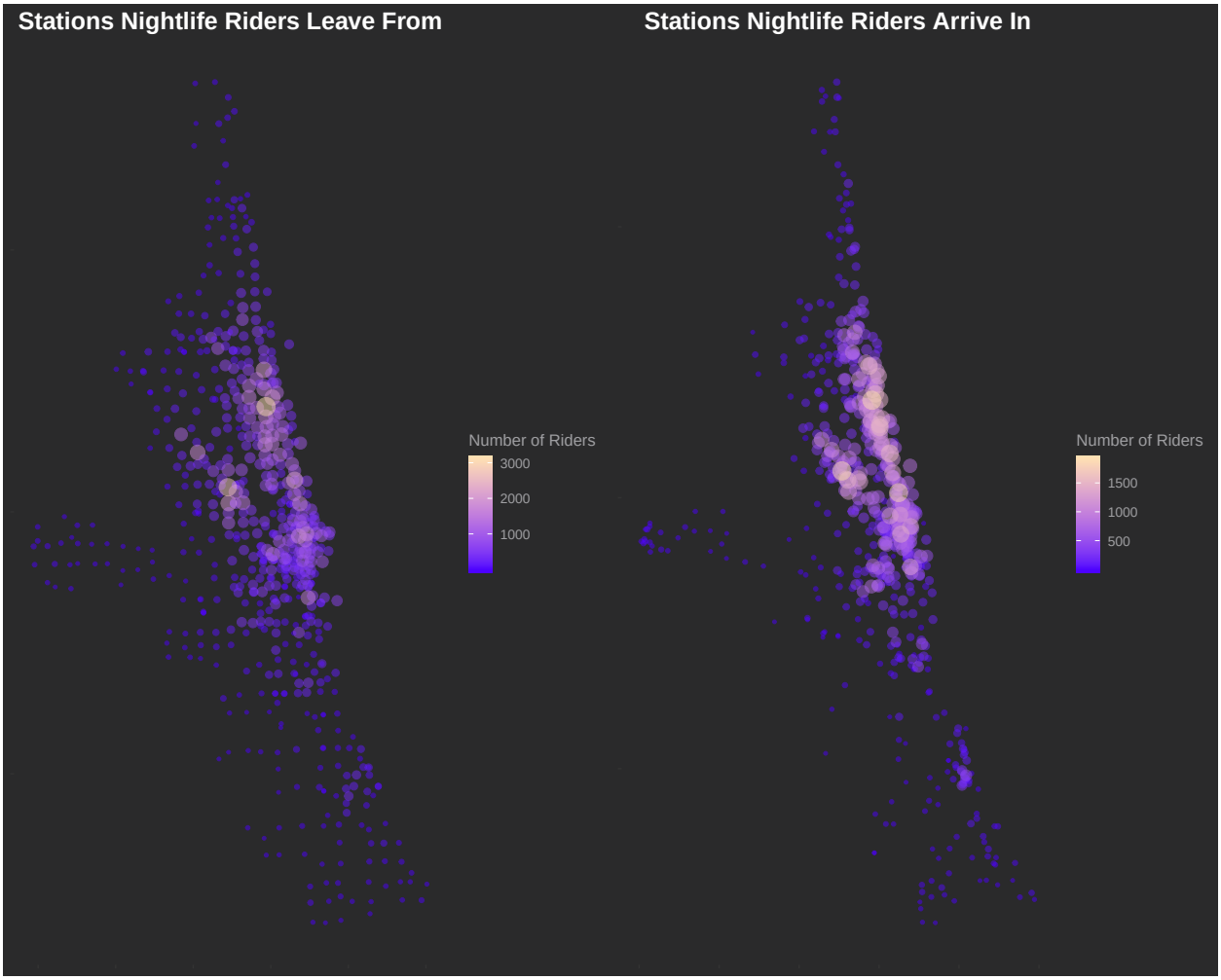


We can see commuting patterns by splitting up the information above into night in day. In other words, we'll observe the number of riders who commute in the morning (5am - 10am) and the evening (4pm - 8pm) between Monday and Friday.



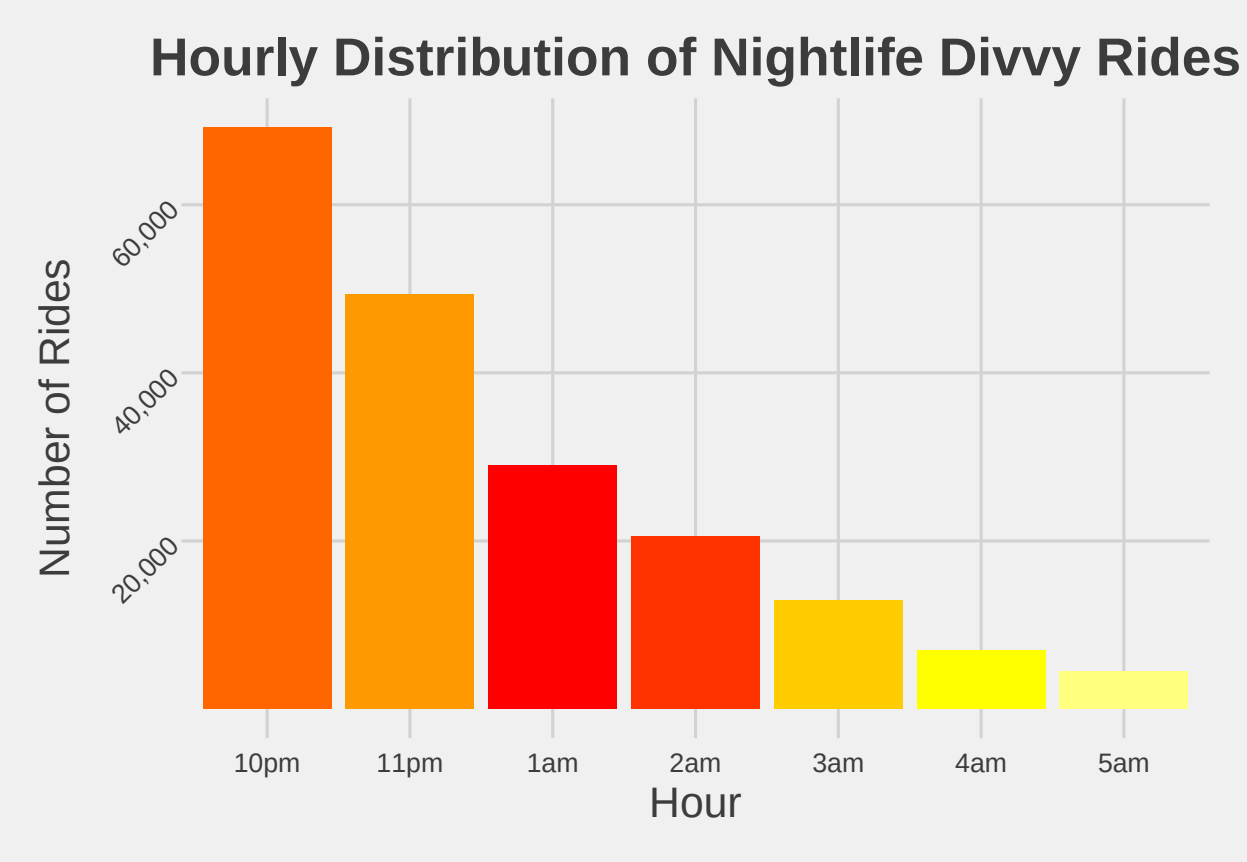
We can certainly see a pattern of riders leaving and returning home (or at least, in nearby neighborhoods) between morning and evening. What's more is that we actually see riders spread out around Chicago from centralized areas Downtown when they leave in the mornings. There is no discernable pattern outside this centroid.

A natural question to ask now is if Divvy has broken into the nightlife scene at all. We can potentially see whether or not this is true based on the riding patterns when people tend to go out to bars or clubs, on Friday and Saturday nights. More specifically, that's Friday (10pm-12am), Saturday (12am-4am, 10pm-12am), and Sunday (12am-4am).



We can see that there is a consolidated band of riders arriving Downtown. We also see that there are many riders leaving far from Downtown, yet not nearly as many riders arriving in those same areas (in the areas outside the centroid). It is very possible that many users are riding Divvy bikes during their nights out - this graphic certainly backs up that idea.

I find it highly unlikely that many users are riding after drinking, however, and are more likely to be riding early on in the night. The following visualization can demonstrate when the riders are riding.



Therefore we can conclude that the majority of these rides are happening early in the night rather than later.

Distance

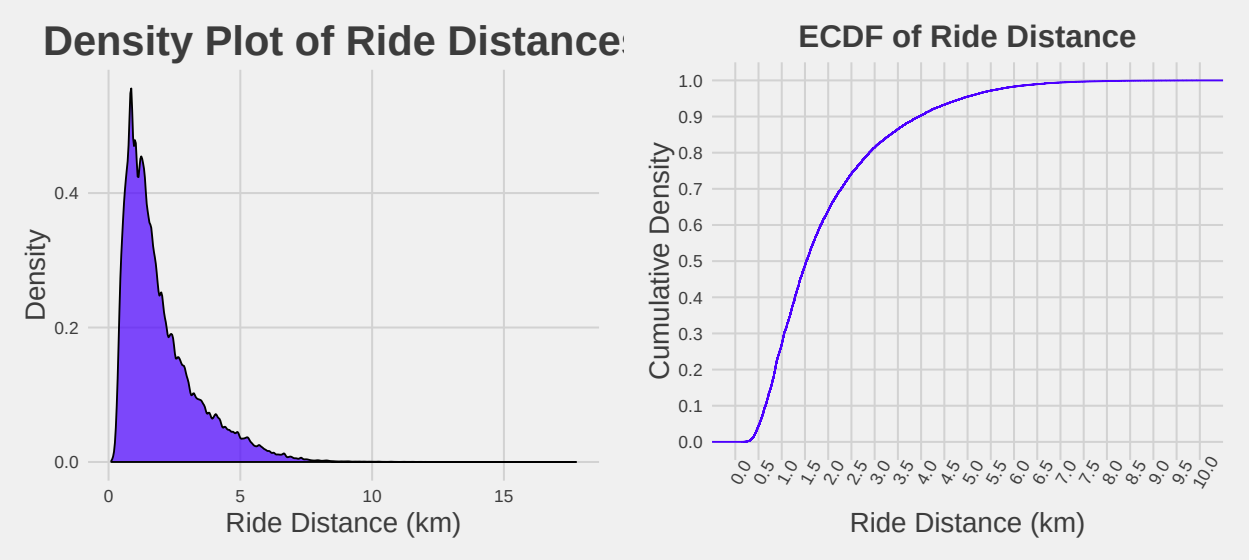
If the pricing is based on time, and distance and time are inherently linked, it is valuable to also look into the distance travelled between stations.

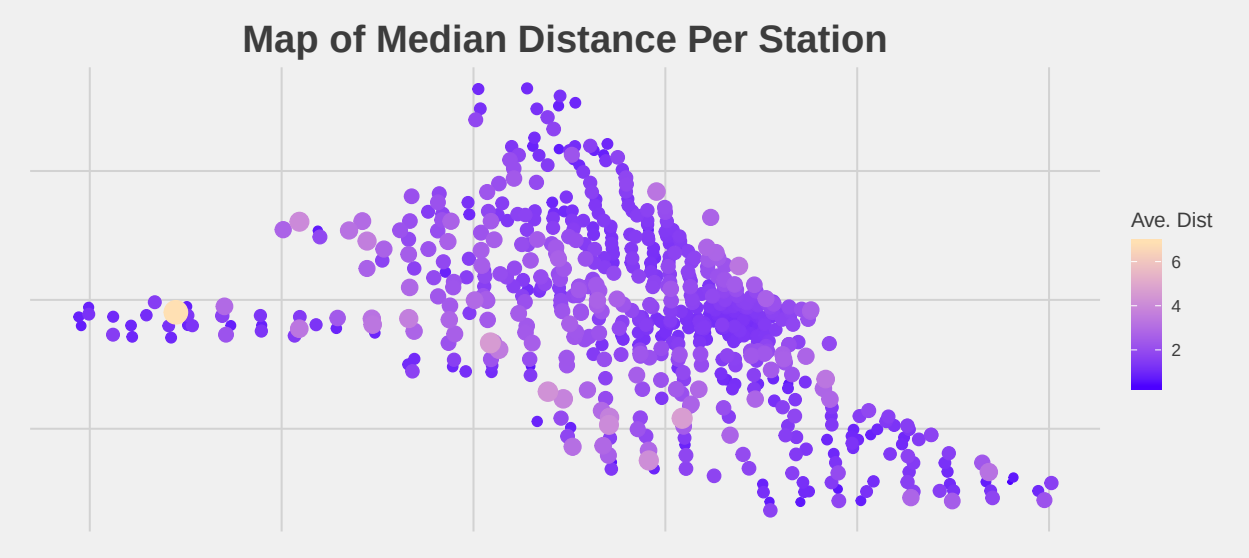
First, we can use the latitude and longitude of the start and stop points to determine how long the trips are.

```
# Function to get distance per row
mydist <- function(row){
   start <- matrix(as.numeric(row[2:1]), ncol = 2)
   end <- matrix(as.numeric(row[4:3]), ncol = 2)
   distance <- spDistsN1(pts = start, pt = end, longlat = T)
   return(distance)
}
# Filter relevant rows to speed up procesing a bit
distDf <- select(bike, 16:17, 21:22)
# Pre-allocate space to vector x
x <- numeric(nrow(distDf))
# Apply function to distDf to get distance
x <- apply(distDf, 1, mydist)
# Add to bike dataframe
bike$ride_distance <- x
# There appear to be rides of 0 km. Let's remove these values.
condition <- bike$ride_distance!=0 # pre-allocate condition
```

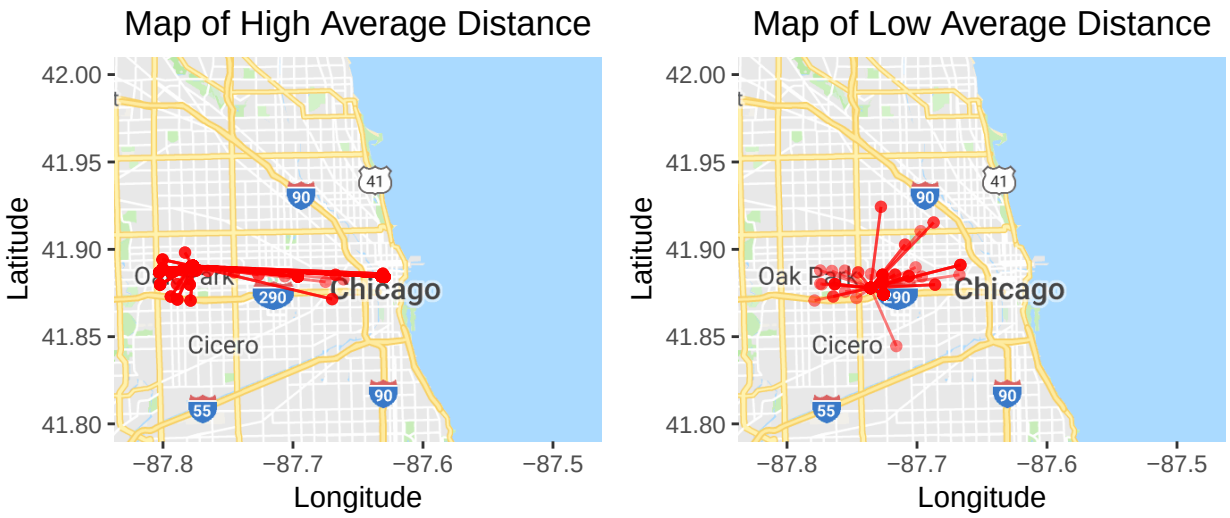
```
bike <- filter(bike, condition)
```

We can view the distribution of the ride distances, as well as the average distance per station.





We can also visualize the actual rides taken between a station with a high median distance, and one with a low median distance.

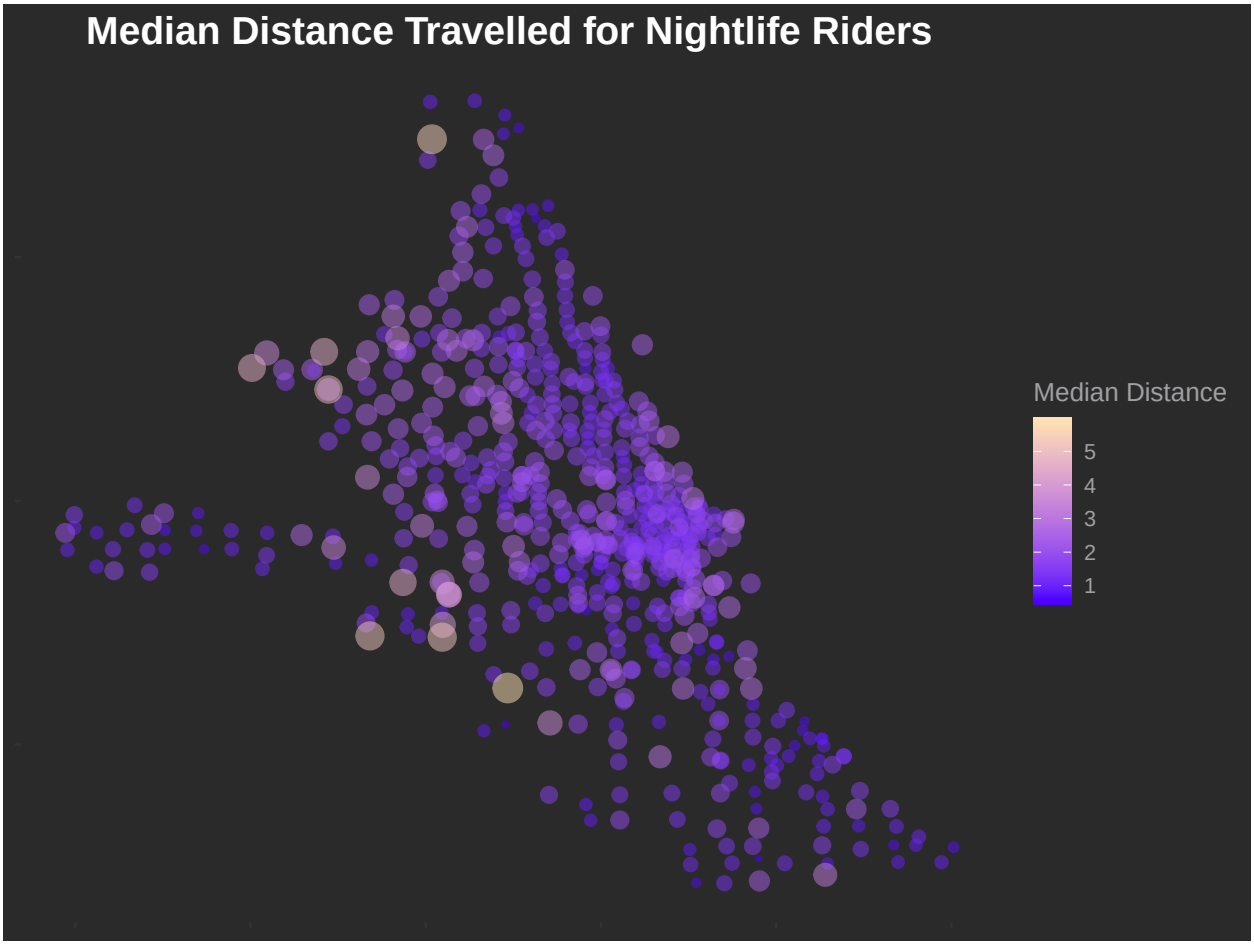


Average distance can clearly become skewed by a handful of rides, so it's not an especially reliable metric to determine whether or not riders truly ride more from certain stations.

We can create similar graphics in the above section re: commuting and nightlife to see if there's any pattern for median distance.

Note that the color on the scales is different, but the size is comparable. It appears as though average distance tends to increase in the Downtown area, but stays approximately the same elsewhere.

Below is a plot of the average distance for users who ride during Friday and Saturday nights.

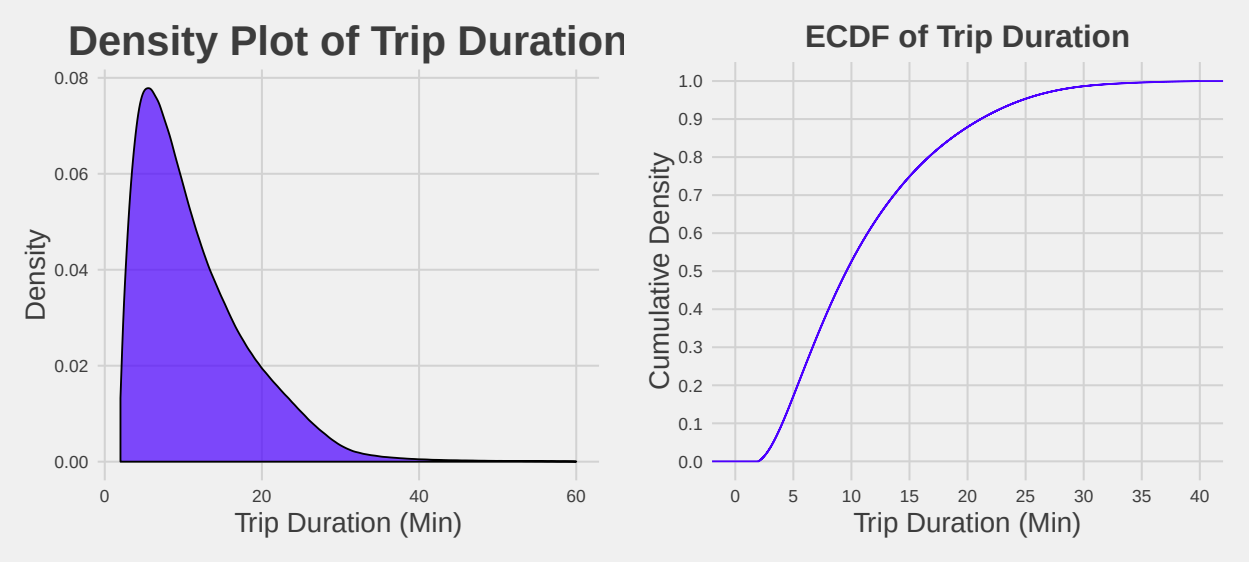


The above plot suggests that riders further from Downtown ride longer distances. This loosely backs up the earlier discussion that Divvy has broken into the Chicago nightlife scene.

Time

As noted earlier, Divvy's single-ride pricing model utilized only the ride-duration, so it's worth creating some visualizations for this variable as well.

Density and CDF:



Conclusion

About 99% of all ride are shorter than 30 minutes, meaning only 1% of rides will yield an extra $3 from renewing the bike. However, 10% of rides are over 4 km and 25% are over 2.5 km.

The correlation between ride distance and the duration of a trip is 0.8254783, so much of the variation in ride-distance is accounted for in trip duration alone. However, creating a model for circumstances in which there is a low trip duration but high distance travelled or vice-versa might enable greater profit. Further investigation into this idea would require financial and marketing information.