Divvy Bike Data Story

Introduction to Divvy and project aim

Divvy is Chicago's bike sharing system which is owned by the Chicago Department of Transportation with the aim to "promote economic recovery, reduce traffic congestion and improve air quality". With over 6,000 bikesand 580 docking stations in the Chicagoland / Evanston area that are available to the public 24/7 the higher the accuracy rate Divvy's forecasting models can achieve translates into more happy customers looking for a sustainable and fun way to get around the city. Therefore, the aim of this capstone project is to develop a forecasting model for one of Divvy's most popular bike stations located at Lake Shore Dr. & Monroe.

The Data

This data set has a total of 13,821,994 observations and 22 columns including:

```
> variable.names(Dataframe_Divvy)
 [1] "1..TRIP.ID"          "START.TIME"          "STOP.TIME"           "BIKE.ID"             "TRIP.DURATION"
 [6] "FROM.STATION.ID"     "FROM.STATION.NAME"   "TO.STATION.ID"       "TO.STATION.NAME"     "USER.TYPE"
[11] "GENDER"              "BIRTH.YEAR"          "FROM.LATITUDE"       "FROM.LONGITUDE"      "FROM.LOCATION"
[16] "TO.LATITUDE"         "TO.LONGITUDE"        "TO.LOCATION"         "Boundaries...ZIP.Codes" "Zip.Codes"
[21] "Community.Areas"     "Wards"
```

Data collection began when Divvy first started its operations in June 2013 and stopped in December 2017. It is also important to note that since a person can choose to reveal their gender and date of birth or not that data is not always provided.

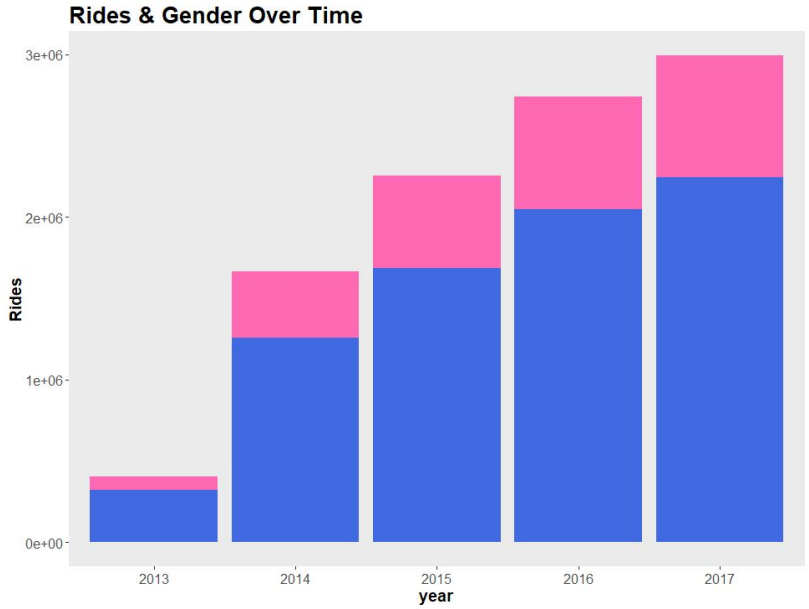
For the purpose of my time series I made 2 subsets of the data. The first subset being the **bikes taken to (203,014 observations)** Lake Shore Dr & Monroe and the 2nd bieng **bikes taken from (210,255 observations)** the Lake Shore Dr. and Monroe station. For the purposes of my forecasting model the "stop.time" column is important for bikes taken to the station and the "start.time" column is important for bikes taken to the station.

Exploratory Data Analysis

In addition to the time series modelling in this capstone I also wanted to take a deeper dive into the overall data set to refine some insights on the Divvy brand and customer. In order to do this I concentrated on the Gender, Birth Year and Trip Duration columns.

Gender

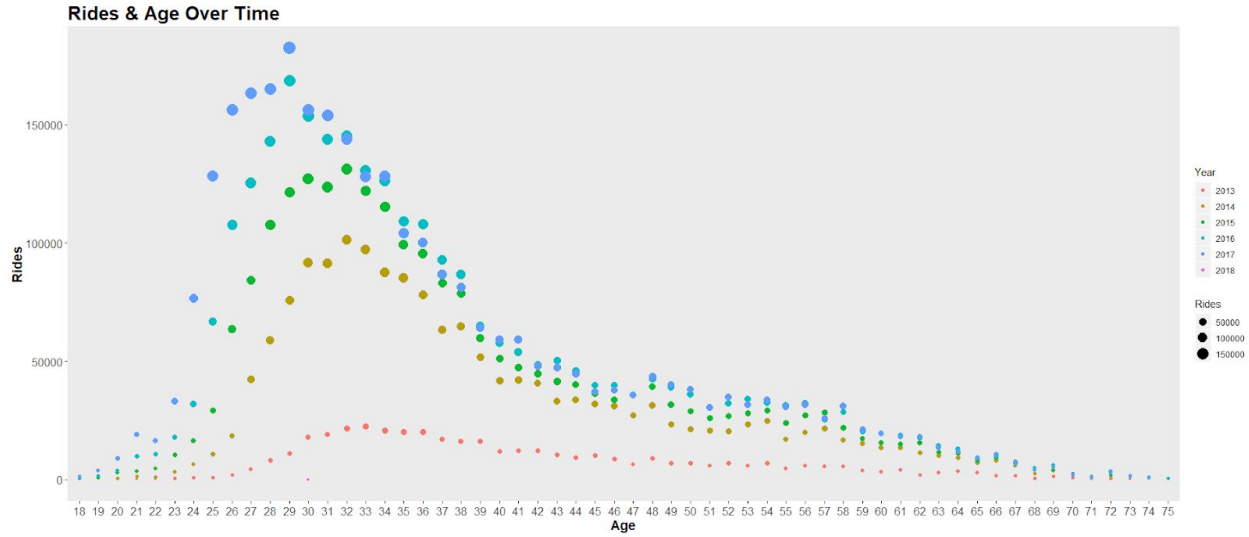
The overall male/female ratio of Divvy trips was 7,552,990 / 2,495,777 as depicted in the graph on the following page we can see that although Divvy has significantly increased its female riders since 2013 from 21% to 25% in 2017 this gender ratio is not reflective of the overall Chicago gender ratio of Chicago which cites the Chicago population as 51% female as reported in US census data.



Male / Female Trip Ratio:
7,552,990 / 2,495,777

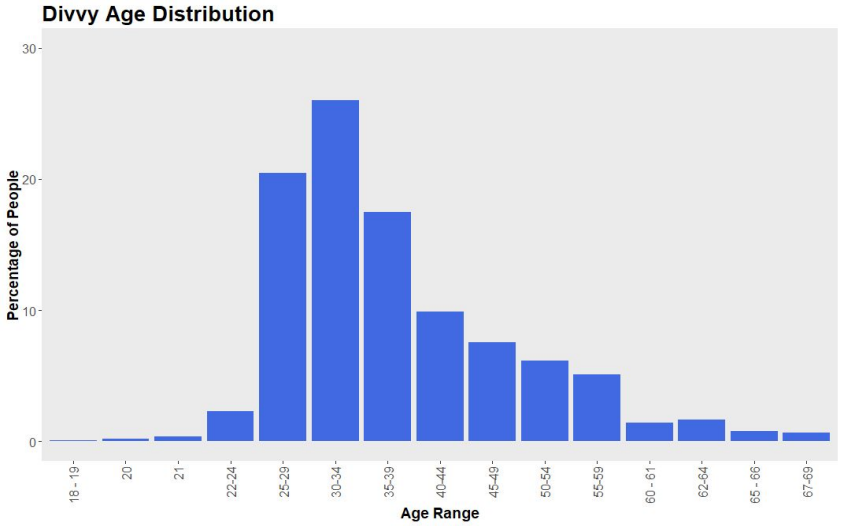
increase in female riders trips from only 21% in 2013 to 25% in 2017

Age

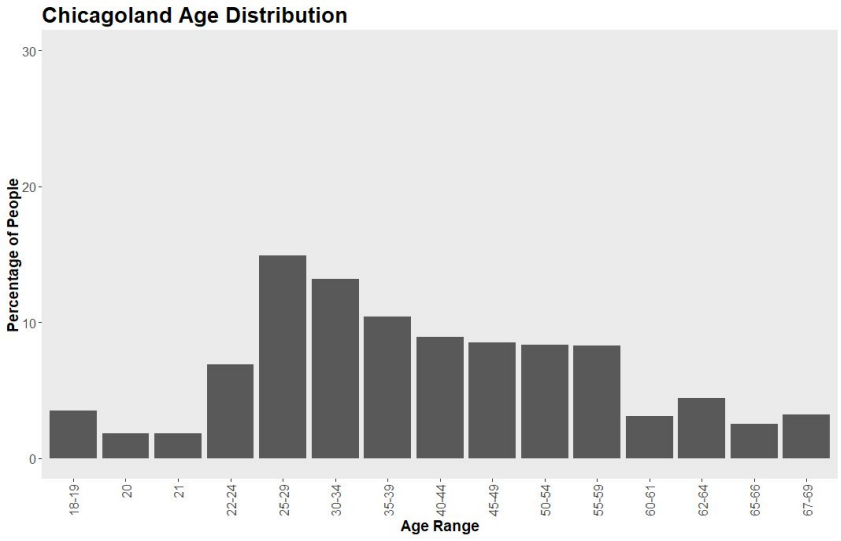


In the graph above I have plotted the age and number of riders. Each year Divvy was in operation is indicated by the different colors of the dots and the larger the dot the higher the number of observations.

From this visualization initially it seems that that over time the “peak” age for riders has shifted from 33 in 2013 to around 28 in 2018 and Divvy seems to capture the late 20s / early 30’s demographic quite well. How well to Divvy’s age segments reflect the overall population of Chicago?



In order to get a better idea of what a "baseline" customer age distribution would look like the visualizations the the left compare the Divvy age segmentation to the overall City of Chicago age segmentation from the U.S. Census. The y-axis is listed as an overall percentage of the Divvy / Chicago population in order to represent these data sets in as much of an unbiased manner as possible.



The Divvy age distribution is skewed to the left starting in the early 20's compared to the overall population in Chicago which demonstrates there is untapped potential in the mid-age market segment as well as the 18-24 segment.

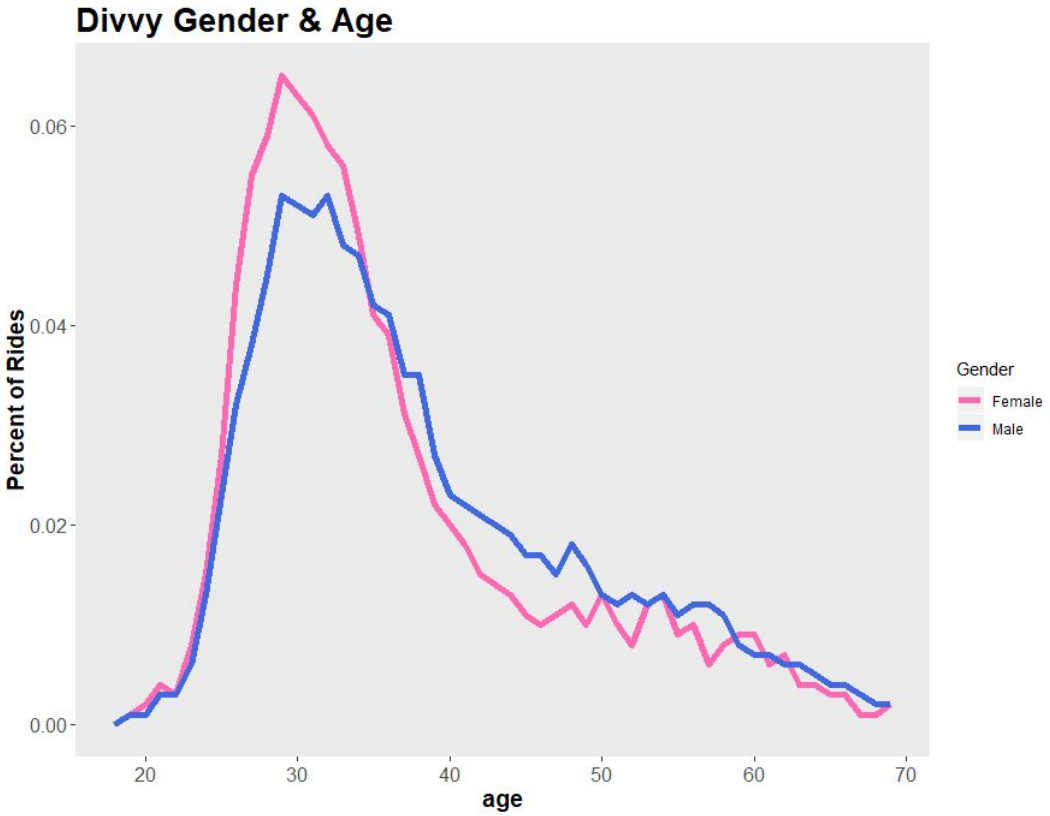
Divvy age standard deviation is 10.4 compared to the Chicago age standard deviation of 14.2.

Taking Divvy's product / service into consideration the wider spread / standard deviation of the Chicago standard deviation could in part be explained by the fact older people, which are considered part of the Chicago population, are not interested or physically unable to bike ride thus preventing Divvy bike from capturing this segment. In any case taking standard deviation into consideration if Divvy targets new age segments for which their customer base is currently much lower compared to the Chicago population (18-24 segment and the 45+ segments for example) the Divvy standard deviation will naturally increase towards the Chicago standard deviation.

Age and Gender

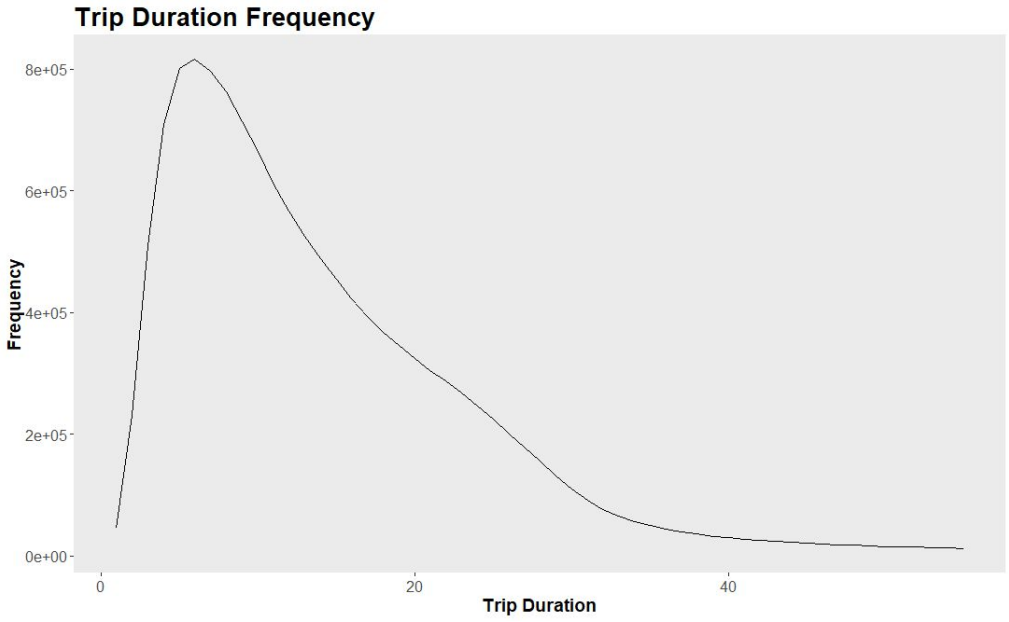
In order to see how the components age and gender interact I created a visualization comparing the percentage of rides for male and female riders and their corresponding age segments on the following page. The male and female data experiences similar peaks and dips with some exceptions. A peak in male an female riders happens around the late 20's - however males experiences another peak in the early 30's. There is also a peak in male riders in their late 40's while female riders experience a dip in that age - a similar instance happens in late 50's. Overall it seems that Divvy male riders come from a more diverse age segment pool compared to

females which peak in their late 20s/early 30s and decrease significantly after that. If Divvy is able to attract more female riders in the future it will be interesting to see if the female segment develops a wider spread as well and is not so dependant on the late 20s/early 30s demographic.



Trip duration

Tidying up trip duration data was the most time consuming. Since the data set only provided a start and stop time I needed to convert those into time objects in R and subtract the stop time from the start time. Trip duration below is recorded in minutes.



As shown in the graph above there is a clear peak between 0-20 followed by a steep drop. This trip length visualization makes sense when taking Divvy's price structure into consideration, keeping in mind that this data set's latest observation was in 2017. Divvy was receiving complaints about their 30 minute ride pricing-structure and in 2018 they decided to change their pricing structure in order to encourage rides longer than 30 minutes. Since the price structure which divvy had set up for the daily and annual offering encouraged rides of 30 minutes or less prior to 2018, naturally that was also reflected in the trip duration visualization since people wanted to avoid any extra cost for exceeding the standard trip duration.

It will be especially interesting to follow overall trip duration on Divvy rides after the announcement of the explorer pass which costs $15 for unlimited rides of up to 3 hours in order to encourage riders interested in longer trips.

Next Steps

For the next steps in my capstone project I will use time-series modelling to create a forecast for the Lake Shore Drive and Monroe Divvy station. This will also include more exploratory data analysis of that station's rides over time and visualizations to see if any patterns of seasonality can be observed.