bike_share_markdown

Callum Thickett

09/11/2021

Setting up my environment and import the data.

```
Trips_2019_Full <- read.csv("Divvy_trips_2019_combined.csv")
```

Lets take a quick look at the data. (note: should be relatively clean already, some basic steps were taken in SQL when combining the quarterly data sets together.)

```
summary(Trips_2019_Full)
```

```
##     trip_id           start_time          end_time             bikeid    
##  Min.   :21742443   Length:3818004     Length:3818004     Min.   :   1  
##  1st Qu.:22873787   Class :character   Class :character   1st Qu.:1727  
##  Median :23962320   Mode  :character   Mode  :character   Median :3451  
##  Mean   :23915629                                         Mean   :3380  
##  3rd Qu.:24963703                                         3rd Qu.:5046  
##  Max.   :25962904                                         Max.   :6946  
##                                                                          
##   tripduration      from_station_id from_station_name  to_station_id  
##  Min.   :      61   Min.   :  1.0   Length:3818004     Min.   :  1.0  
##  1st Qu.:     411   1st Qu.: 77.0   Class :character   1st Qu.: 77.0  
##  Median :     709   Median :174.0   Mode  :character   Median :174.0  
##  Mean   :    1450   Mean   :201.7                      Mean   :202.6  
##  3rd Qu.:    1283   3rd Qu.:289.0                      3rd Qu.:291.0  
##  Max.   :10628400   Max.   :673.0                      Max.   :673.0  
##                                                                        
##  to_station_name      usertype            gender            birthyear     
##  Length:3818004     Length:3818004     Length:3818004     Min.   :1759    
##  Class :character   Class :character   Class :character   1st Qu.:1979    
##  Mode  :character   Mode  :character   Mode  :character   Median :1987    
##                                                           Mean   :1984    
##                                                           3rd Qu.:1992    
##                                                           Max.   :2014    
##                                                           NA's   :538751
```

potential issues:

- Max trip duration is > 10000000

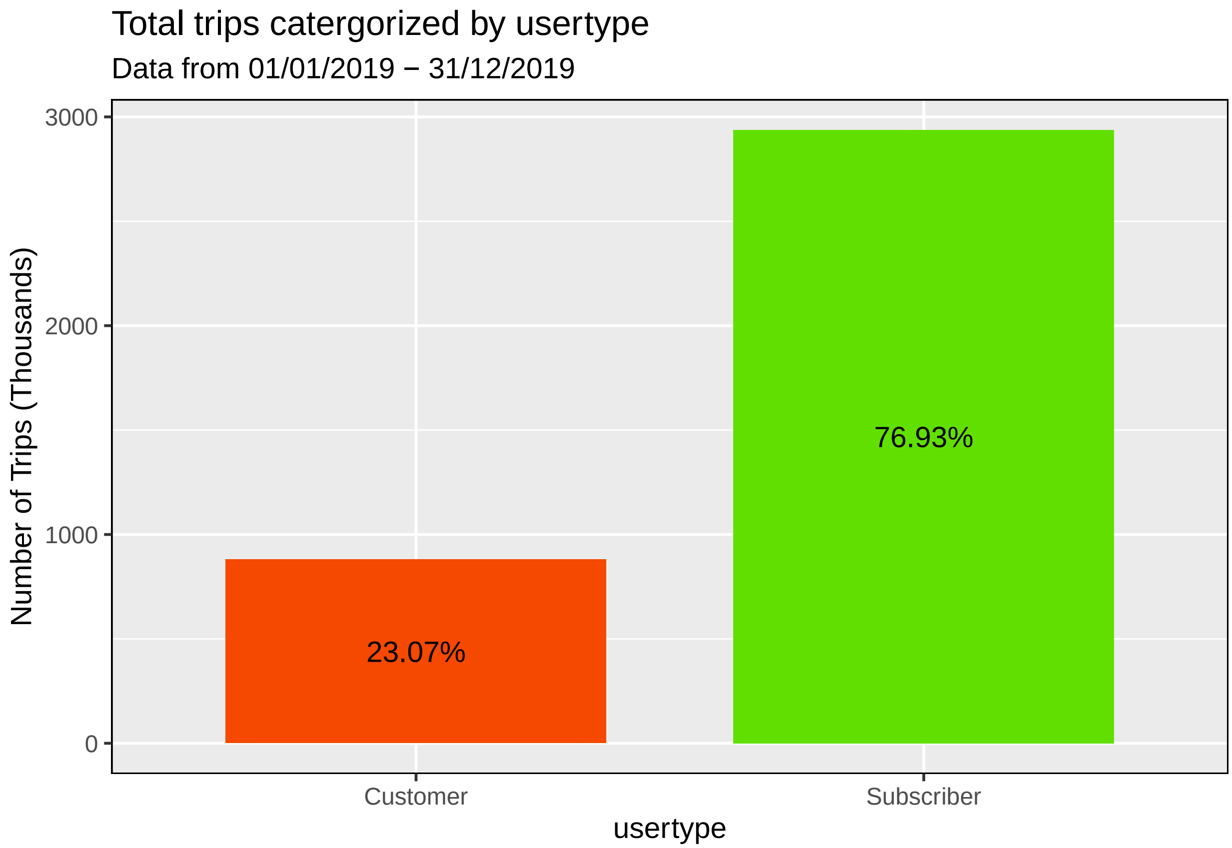
- Min birth year is 1759
- Significant portion of birth year column is empty
  - Until we get to analyzing birth year this wont be an issue(all other trip data is present)
- Significant portion of gender column is empty.
  - Until we get to analyzing gender this wont be an issue(all other trip data is present)

quick overview of the current usertype spread.

as can be seen, ~76% of the total **trips** are taken by subscribers. this is **not** to say that 76% of the userbase are subscribed. Of course, a subscriber is much more likely to use the service multiple times compared to a customer.

need more data i.e user_Ids to get an actual accurate idea of the spread of usertype.

```
PCT_total_trips <- Trips_2019_Full %>%
  group_by(usertype) %>%
  summarise(number_of_trips =n())


Total_count <-PCT_total_trips[[1,2]]+PCT_total_trips[[2,2]]
Total_Subs <- (PCT_total_trips[[2,2]] / Total_count)*100
Total_Customers <- (PCT_total_trips[[1,2]] / Total_count)*100

Total_count <-PCT_total_trips[[1,2]]+PCT_total_trips[[2,2]]
PCT_Subs <- (PCT_total_trips[[2,2]] / Total_count)*100
PCT_Customers <- (PCT_total_trips[[1,2]] / Total_count)*100
PCT_subs_aes <- paste(round(PCT_Subs,2),"%", sep = "")
PCT_customers_aes <- paste(round(PCT_Customers,2),"%", sep = "")
y_subs <-PCT_total_trips[[2,2]] /2
y_Customers <- PCT_total_trips[[1,2]] /2

ggplot(data=Trips_2019_Full, aes(x=usertype,fill=usertype)) +geom_bar(width=0.75) +

theme(legend.position = "none",
      panel.border = element_rect(colour="Black",
      fill=NA) ) +
scale_y_continuous(name = "Number of Trips (Thousands)",
                   labels = function(y) y/1000) +
scale_fill_manual(values = c("#F54800","#61E000")) +
  annotate("text", label =PCT_subs_aes, x=2,y=y_subs)+
  annotate("text", label =PCT_customers_aes, x=1,y=y_Customers) +
  labs(title = "Total trips catergorized by usertype", subtitle = "Data from 01/01/2019 - 31/12/2019")
```

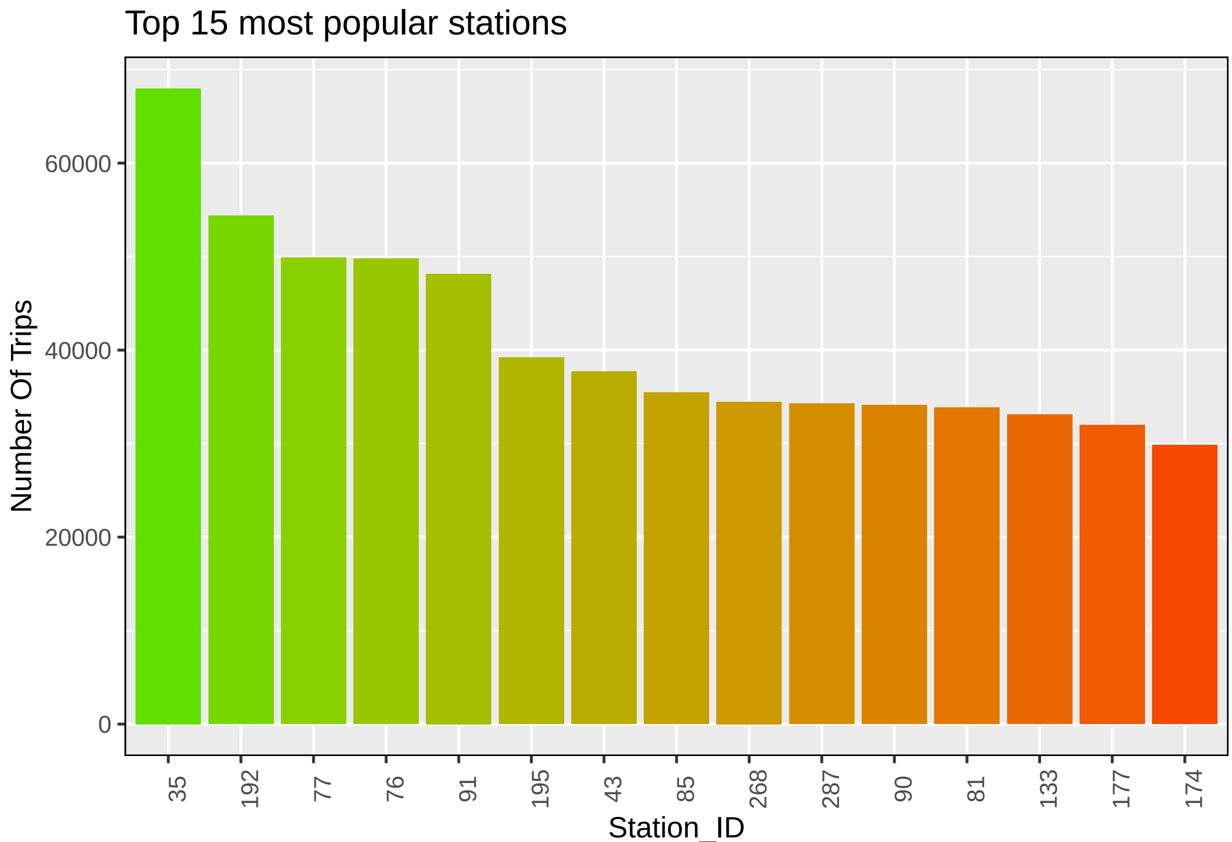


A look at the most popular stations

Next lets look at how subscriber and customer trips differ, i.e which stations and journeys are most popular by both groups. First, an overall look at the most popular stations.

```
Station_Slice_Cust <- 15
Top_ten_station <-Trips_2019_Full %>%
  group_by(from_station_id) %>%
  summarise(station_count=n()) %>%
  slice_max(station_count, n=Station_Slice_Cust)


Trips_top_ten_stations <-Trips_2019_Full %>%
  filter(from_station_id %in% Top_ten_station$from_station_id)

#make a colour scale to use in scale_fill_manual

cc <- scales::seq_gradient_pal("green2", "red", "Lab")(seq(0.1,0.9,length.out=Station_Slice_Cust))

ggplot(data=Trips_top_ten_stations, aes(x=fct_infreq(as.character(from_station_id)))) +geom_bar(aes(fil
  theme(legend.position = "none", axis.text.x =element_text(angle = 90) ) +
  labs(x="Station_ID", y="Number Of Trips") +
 scale_fill_manual(values = cc) +
  labs(title=paste("Top",Station_Slice_Cust, "most popular stations" )) +
```

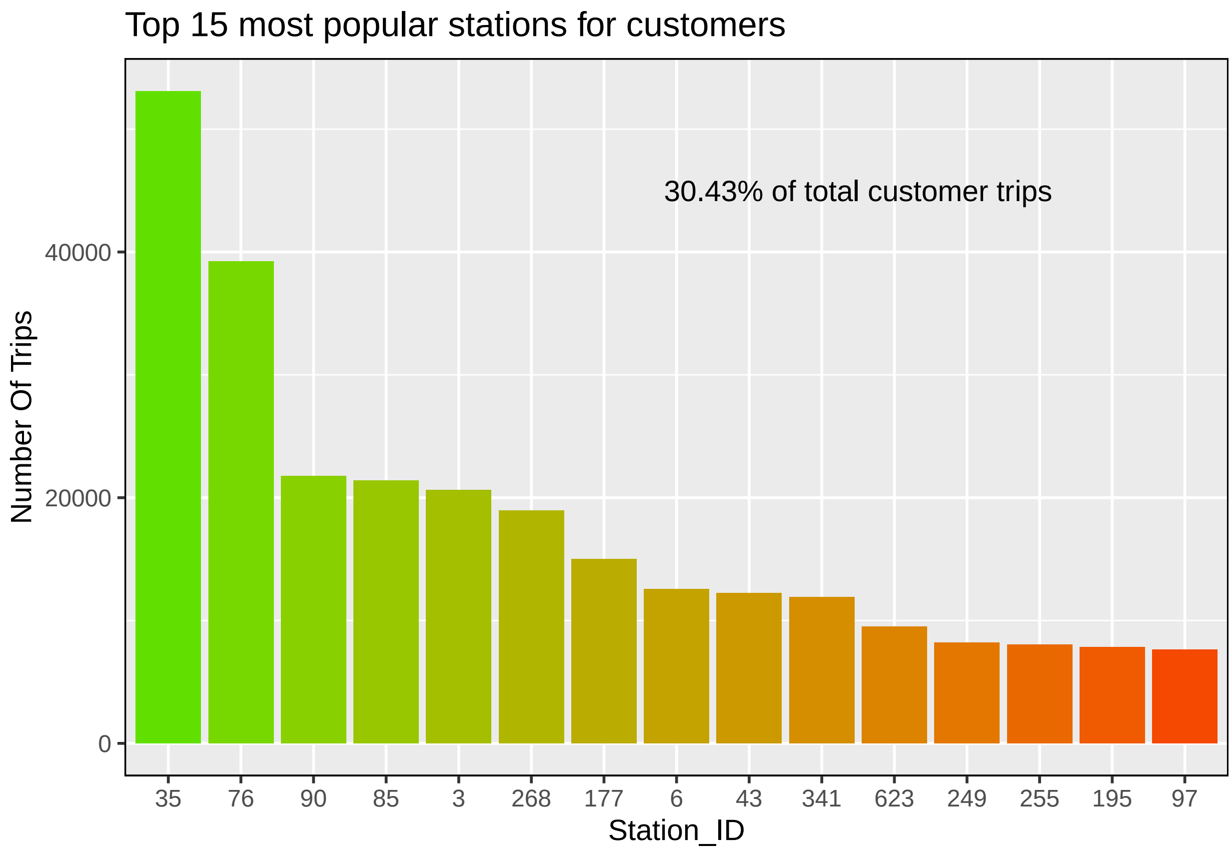
```
theme( panel.border = element_rect(colour="Black",
      fill=NA) )
```



Lets see the same thing but only include customers.

```
top_stations_Cust_filter <- Trips_2019_Full %>%
  filter(usertype=="Customer") %>%
  group_by(from_station_id) %>%
  summarise(Station_Count = n()) %>%
  slice_max(Station_Count, n =Station_Slice_Cust)

top_stations_Sub_filter <- Trips_2019_Full %>%
  filter(usertype=="Subscriber") %>%
  group_by(from_station_id) %>%
  summarise(Station_Count = n()) %>%
  slice_max(Station_Count, n =Station_Slice_Cust)

top_stations_Cust <- Trips_2019_Full %>%
  filter(from_station_id %in% top_stations_Cust_filter$from_station_id)
#pct values are wrong, need to change its including all trips not just customer!
   Pct_Value_Stations_Cust<- paste(round((top_stations_Cust %>% filter(usertype=="Customer") %>%
  summarise(n())) /
  (Trips_2019_Full %>%
```

```
    filter(usertype=="Customer") %>%
  summarise(n()))*100, 2),"%", sep = "")

   ggplot(data=top_stations_Cust %>% filter(usertype=="Customer"),aes(x=fct_infreq(as.character(from_sta
  labs(title = paste("Top",Station_Slice_Cust, "most popular stations for customers")) +
  labs(x="Station_ID", y="Number Of Trips") +
  annotate("text",label = paste(Pct_Value_Stations_Cust, "of total customer trips"), x=10.5,y=45000) +
theme( panel.border = element_rect(colour="Black",
      fill=NA) ) +
    scale_fill_manual(values = cc)
```



Despite customer trips making up only ~24% of total trips, they make up the vast majority of trips at some of the most popular stations!(**35**, and **76**).

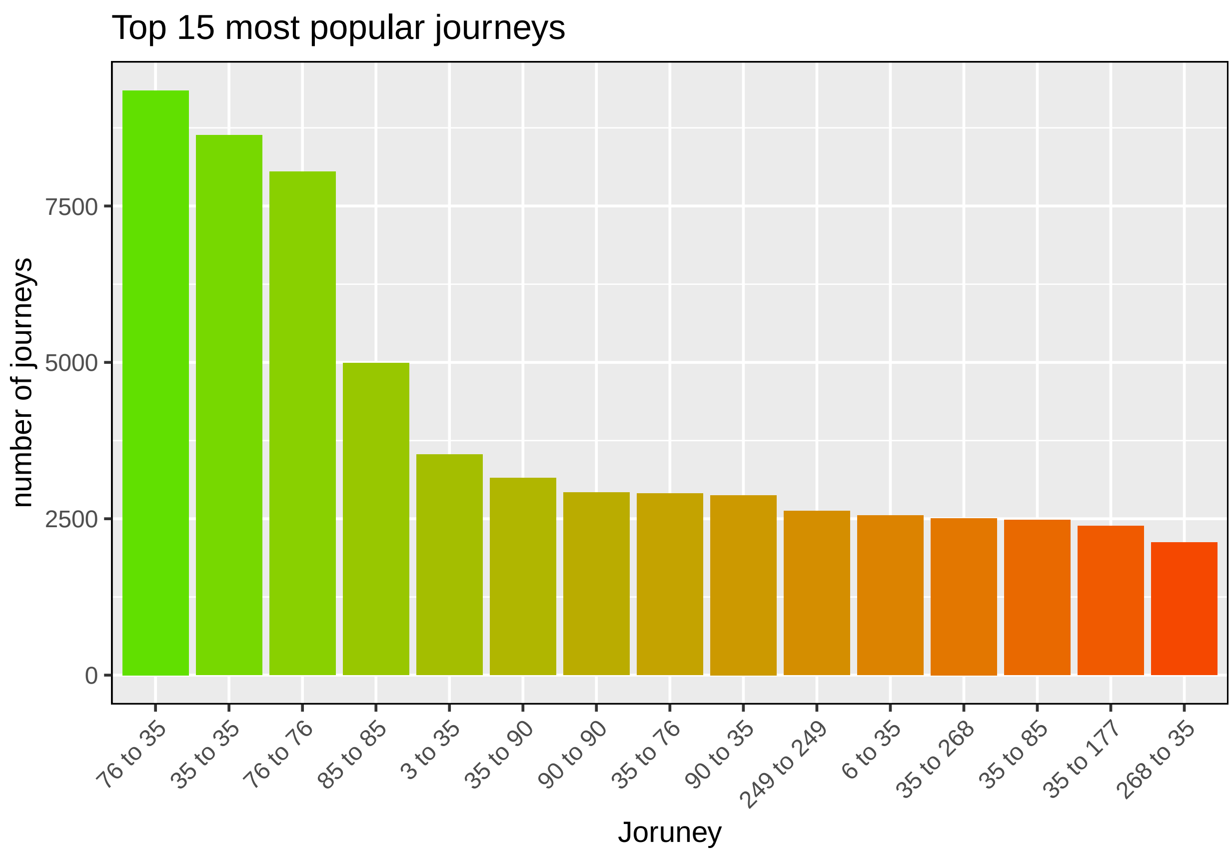
Also, despite there being 640 stations, 30.43% of ALL customer trips start at one of these 15 stations.

We could look at the same information for destination stations, but a better metric for popularity is to simply look at *journeys* as a whole. This gives us an overall idea of what stations are used.

To see *journeys* we need to combine the from_station_id and to_station_id columns, and then filter the resulting column.

```
Trips_2019_Full_Journeys <-Trips_2019_Full %>%
  mutate(journeys = paste(from_station_id, "to", to_station_id))

Journeys_filter <- Trips_2019_Full_Journeys %>%
  filter(usertype=="Customer") %>%
```

```
  group_by(journeys) %>%
  summarise(journey_count=n()) %>%
  slice_max(journey_count,n=Station_Slice_Cust)

Top_Journeys <- Trips_2019_Full_Journeys %>%
  filter(journeys %in% Journeys_filter$journeys)

ggplot(data=Top_Journeys %>% filter(usertype=="Customer"),aes(x=fct_infreq(as.character(journeys)))) +ge
  theme(axis.text.x = element_text(angle=45,hjust = 1), panel.border = element_rect(colour="black",fill=
  labs(title=paste("Top", Station_Slice_Cust, "most popular journeys"),
       x="Joruney",y="number of journeys") +
  scale_fill_manual(values = cc)
```



clearly, the majority of trips involve stations 76 or 35.

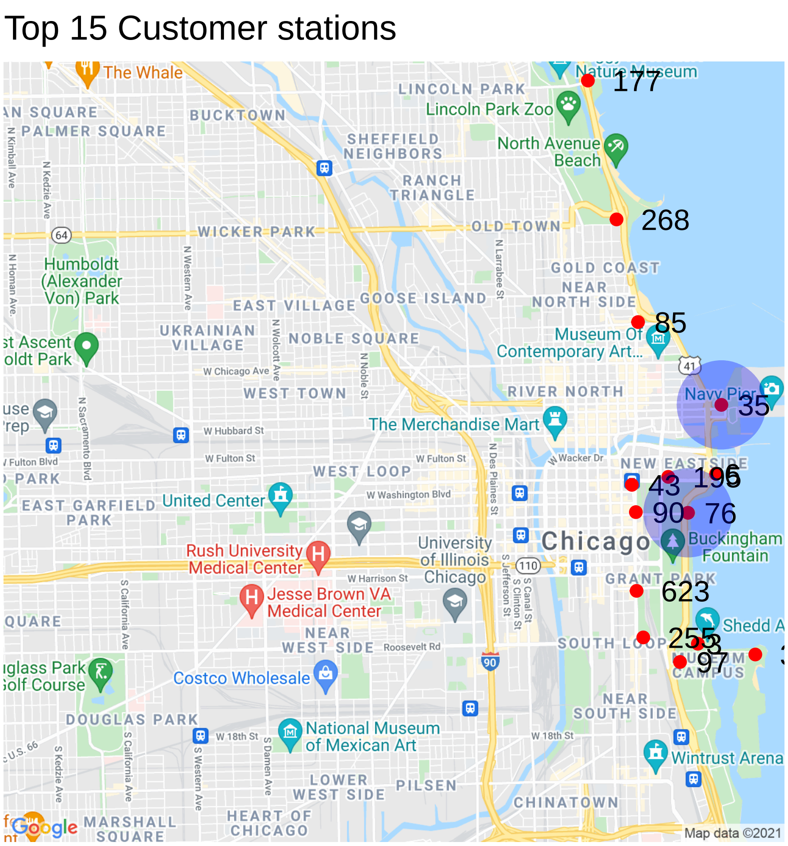
```
#need to import the stations data set to get long/lat for the stations.
station_locations <- read.csv("Divvy_Stations_2017_Q3Q4.csv")

#getting a subset of the data so we only have the most popular customer stations.
station_locations_subset <- station_locations %>%
  filter(id %in% top_stations_Cust_filter$from_station_id)
Cust_map <- ggmap(get_googlemap(center =c(lon=-87.6582,lat=41.88742),
```

```
                                    zoom=13,
                                    maptype = "road",
                                    color = "color")) +
  geom_point(data=station_locations_subset, aes(x=longitude,y=latitude),colour="red", size=1.75) +
  geom_text(data=station_locations_subset,aes(x=longitude,y=latitude),label=station_locations_subset$id
  geom_point(data=station_locations_subset, aes(x=-87.612043,y=41.892278),size=15, alpha=0.025,colour="
  geom_point(data=station_locations_subset, aes(x=-87.616743,y=41.880958),size=15, alpha=0.025,colour="
  theme(axis.text.x =element_blank(),
        axis.text.y=element_blank(),
        axis.ticks.x = element_blank(),
        axis.ticks.y = element_blank(),
        axis.title.x = element_blank(),
        axis.title.y = element_blank()) +
  labs(title=paste("Top",Station_Slice_Cust,"Customer stations"))
```

Lets look at some geographical data using the long/lat of the stations to see why this could be.

```
## Source : https://maps.googleapis.com/maps/api/staticmap?center=41.88742,-87.6582&zoom=13&size=640x64
```

```
Cust_map
```



As can be seen from the above map, both stations 35 and 76 are in key tourism spots, one on the pier front, and the other in between the waterfront and Grant Park.

Lets compare this to the top 15 Subscriber visited stations.

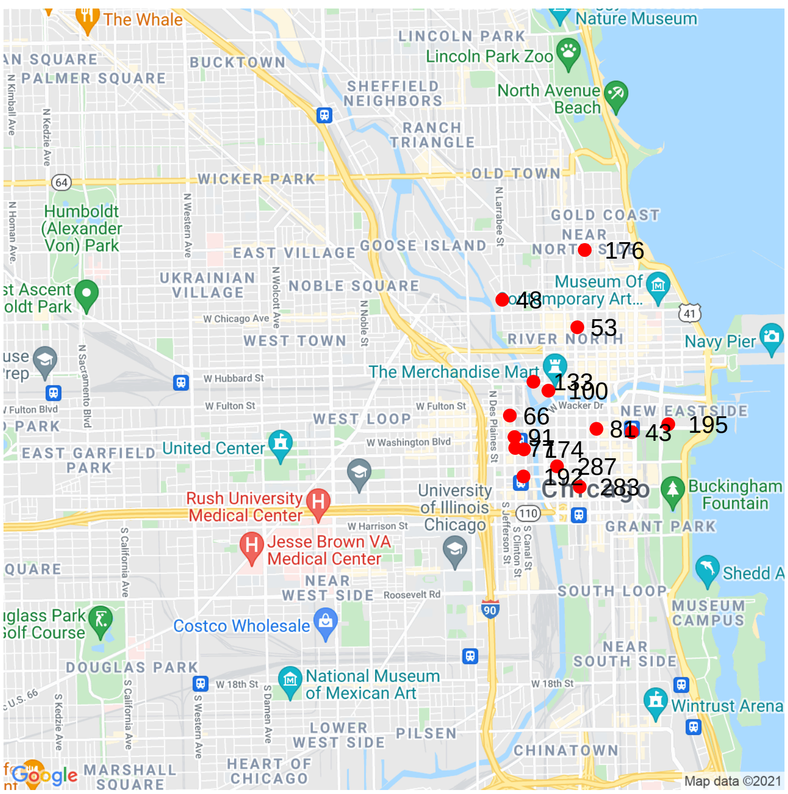
As we can see the most popular stations used by Subscribers are much more city based.

```
station_locations_subset_subs <- station_locations %>%
  filter(id %in% top_stations_Sub_filter$from_station_id)
Subs_map <- ggmap(get_googlemap(center =c(lon=-87.6582,lat=41.88742),
                                zoom=13,
                                maptype = "road",
                                color = "color")) +
  geom_point(data=station_locations_subset_subs, aes(x=longitude,y=latitude),colour="red", size=1.75) +
  geom_text(data=station_locations_subset_subs,aes(x=longitude,y=latitude),label=station_locations_subs
  labs(title=paste("Top",Station_Slice_Cust,"Subscriber stations"))+
  theme(axis.text.x =element_blank(),
        axis.text.y=element_blank(),
        axis.ticks.x = element_blank(),
        axis.ticks.y = element_blank(),
        axis.title.x = element_blank(),
        axis.title.y = element_blank())
```

```
## Source : https://maps.googleapis.com/maps/api/staticmap?center=41.88742,-87.6582&zoom=13&size=640x64
```

```
Subs_map
```

Top 15 Subscriber stations



If we look at the average trip duration times for these stations compared to the average time for all stations, we see it is significantly higher.

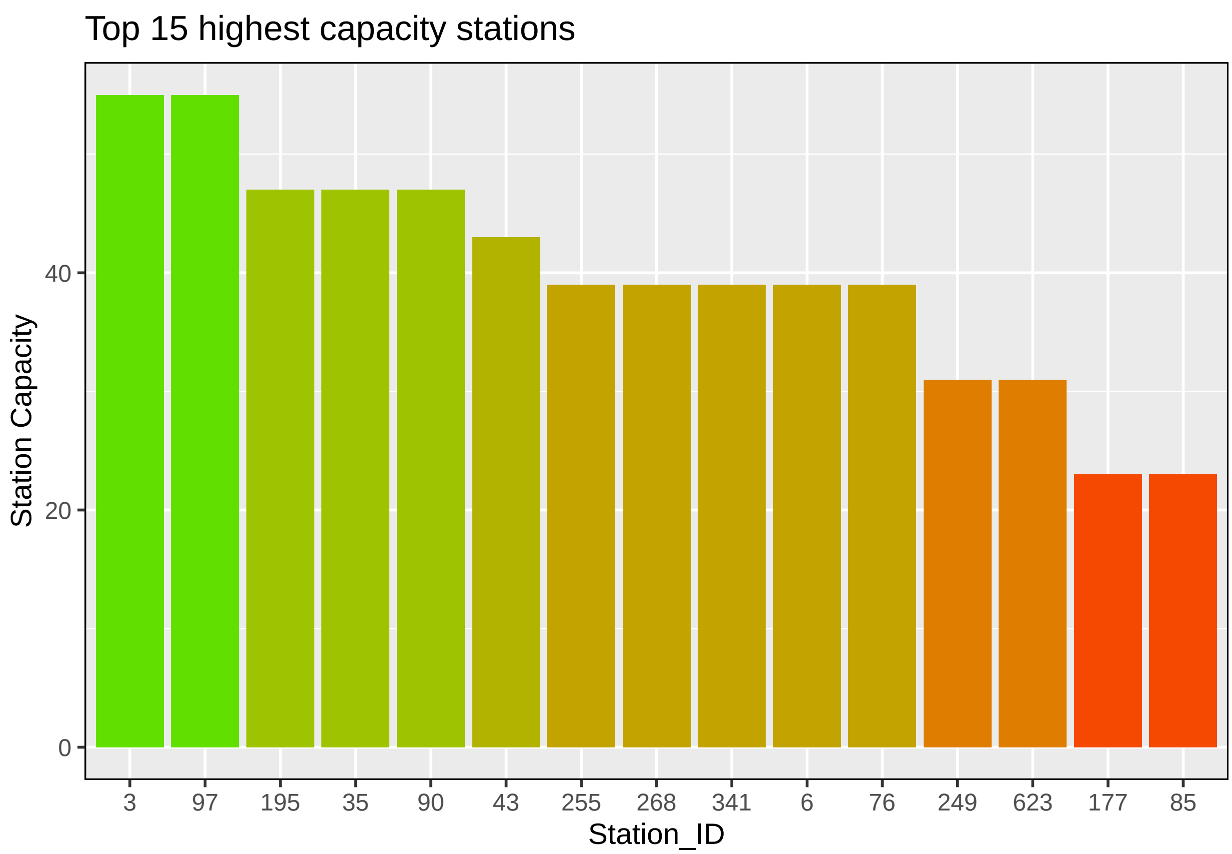
This supports the idea the the majority of users here are tourists, probably hiring out the bikes for fun rather than function. on the other hand, the total average time is much shorter, suggesting the service as a whole is more so being used by commuters and the likes travelling shorter distances within the city. Of course this is just speculation and would need further analysis to conclude anything.

```
Trimmed_mean_35<-Trips_2019_Full %>%
  filter(from_station_id==35) %>%
  summarise(Trimmed_mean=mean(tripduration,trim=0.10))

Trimmed_mean_76<-Trips_2019_Full %>%
  filter(from_station_id==76) %>%
  summarise(Trimmed_mean=mean(tripduration,trim=0.10))

Trimmed_mean_90<-Trips_2019_Full %>%
  filter(from_station_id==90) %>%
  summarise(Trimmed_mean=mean(tripduration,trim=0.10))


Trimmed_mean_all<-Trips_2019_Full %>%
  summarise(Trimmed_mean=mean(tripduration,trim=0.10))

Trimmed_mean_values <- c(Trimmed_mean_35[[1,1]],Trimmed_mean_76[[1,1]],Trimmed_mean_90[[1,1]],Trimmed_me


Top_Station_IDS <-c(top_stations_Cust_filter$from_station_id)


Trimmed_means_df <- data.frame(Station_Ids =c(Top_Station_IDS[1:3],"Overall average"),mean_trip_duratio
Trimmed_means_df
```

```
##        Station_Ids mean_trip_duration
## 1               35          1864.0118
## 2               76          1871.4839
## 3               90          1731.6775
## 4 Overall average           838.7208
```

We can also have a quick look at capacity sizes for these stations to see if that has any effect.

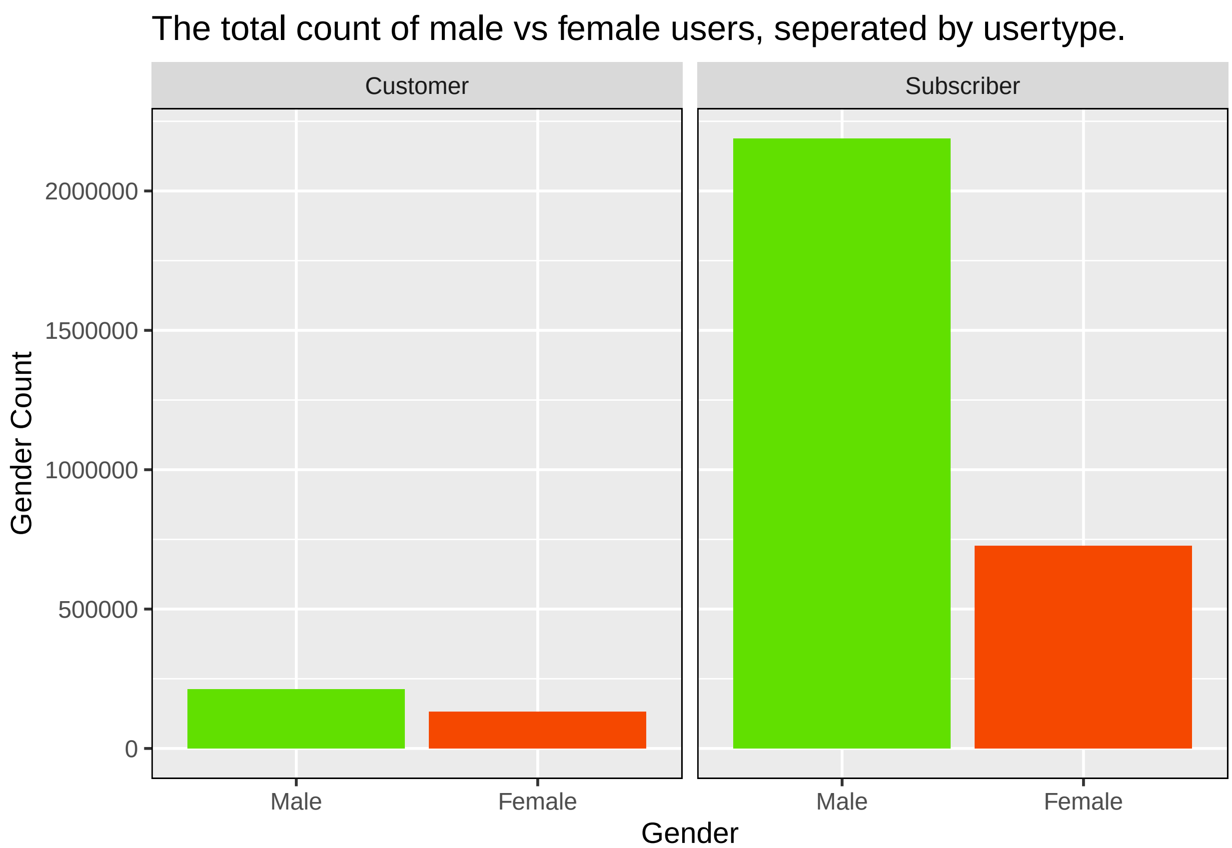
```
cc2 <-scales::seq_gradient_pal("green2", "red", "Lab")(seq(0.1,0.9,length.out=6))

ggplot(data=station_locations_subset,aes(x=reorder(as.character(id),-dpcapacity),y=dpcapacity))+geom_ba
  scale_fill_gradient(low ="#F54800",high = "#61E000" ) +
  theme(panel.border = element_rect(colour="black",fill = FALSE)) +
  labs(title = paste("Top", Station_Slice_Cust, "highest capacity stations"),x="Station_ID", y="Station
```



It would appear that at first glance there isnt much correlation between capactity size and usertype.

Look at gender and age contribution to usertype.

```
Trips_2019_Full %>%
  filter(gender!="") %>%
ggplot(aes(x=fct_infreq(gender))) +geom_bar(aes(fill=gender),show.legend = FALSE) +
  facet_wrap(~usertype) +
  scale_fill_manual(values = c("#F54800","#61E000")) +
  theme(panel.border = element_rect(colour="black",fill=FALSE))+
  labs(x="Gender",y="Gender Count",title = "The total count of male vs female users, seperated by usert
```




```
#calculate % of women users for customer vs subs.

Female_Cnt_Subs_df <-Trips_2019_Full %>%
  filter(usertype=="Subscriber",gender=="Female") %>%
  summarise(row_count=n())

Male_Cnt_Subs_df <-Trips_2019_Full %>%
  filter(usertype=="Subscriber",gender=="Male") %>%
  summarise(row_count=n())

Female_Cnt_Cust_df <-Trips_2019_Full %>%
  filter(usertype=="Customer",gender=="Female") %>%
  summarise(row_count=n())

Male_Cnt_Cust_df <-Trips_2019_Full %>%
  filter(usertype=="Customer",gender=="Male") %>%
  summarise(row_count=n())
Total_Cust_Count <- Female_Cnt_Cust_df[[1,1]] + Male_Cnt_Cust_df[[1,1]]
Total_Subs_Count <- Male_Cnt_Subs_df[[1,1]] + Female_Cnt_Subs_df[[1,1]]


Pct_Female_Subs <- paste((Female_Cnt_Subs_df[[1,1]]/Total_Subs_Count)*100,"%")

Pct_Female_Cust <- paste((Female_Cnt_Cust_df[[1,1]]/Total_Cust_Count)*100,"%")
Pct_Female_Cust
```

```
## [1] "38.1888070846238 %"
```

```
Pct_Female_Subs
```

```
## [1] "24.9274346946562 %"
```

- 38% of all customer trips are female, (assuming these numbers are consistent for those who did not provide a gender, although this is an area that needs to be further looked into)
- 24% of all subscriber trips were female.
- this means a bigger percentage of overall customer trips are from females. i.e males seem more likely to subscribe than females.

looking at age now, so need to clean the data accordingly.

```
  min(Trips_2019_Full$birthyear, na.rm = TRUE)
```

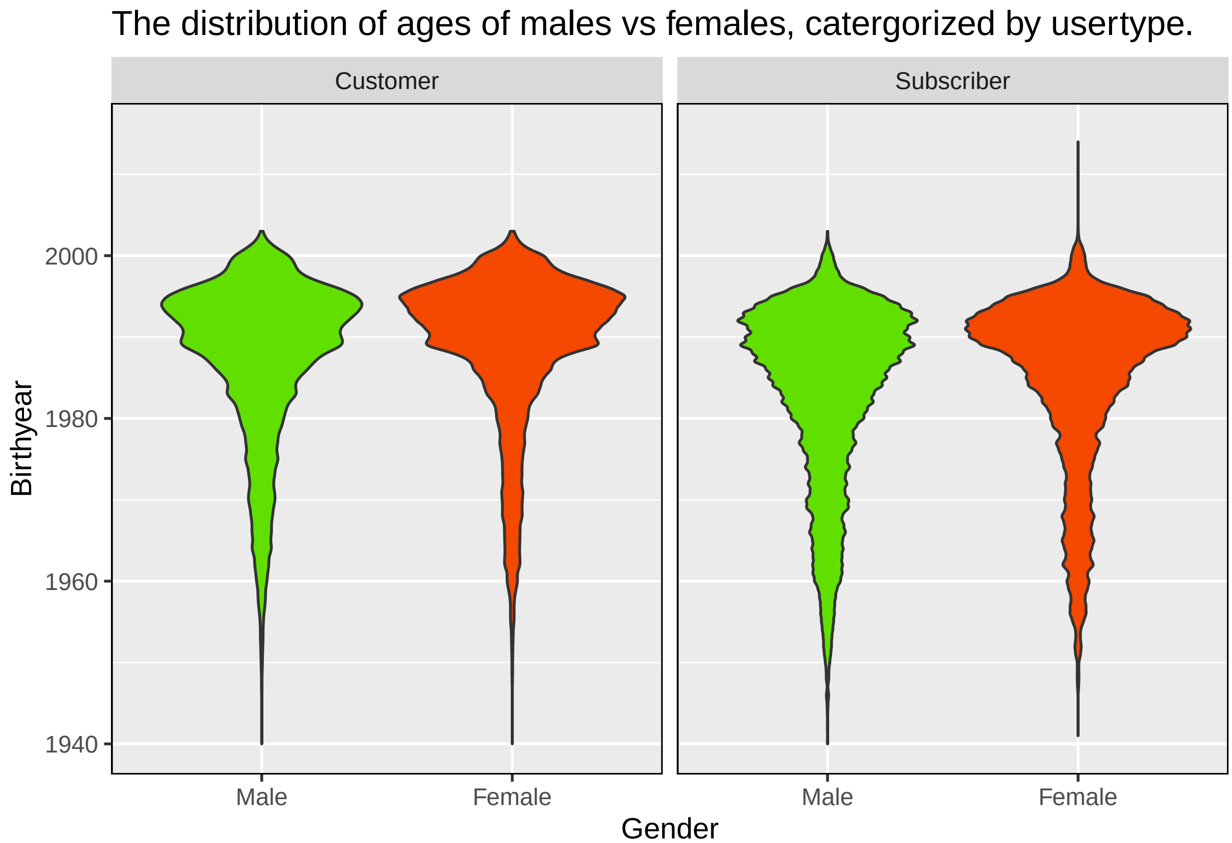
```
## [1] 1759
```

```
  max(Trips_2019_Full$birthyear, na.rm = TRUE)
```

```
## [1] 2014
```

clearly people were not born in the 1700s

```
Trips_2019_Full %>%
  filter(birthyear>1921, gender != "") %>%
  ggplot(aes(x=fct_infreq(gender),y=birthyear)) +geom_violin(aes(fill=gender),show.legend = FALSE) +
  facet_wrap(~usertype) +
  ylim(1940,2015)+
  scale_fill_manual(values = c("#F54800","#61E000")) +
  theme(panel.border = element_rect(colour="black",fill=FALSE)) +
  labs(x="Gender",y="Birthyear",title = "The distribution of ages of males vs females, catergorized by u
```



Subscriber base in general is a little older, females have less even distribution, which can be seen more clearly in the following density plot

```
Trips_2019_Full %>%
  filter(birthyear>1921,gender!="") %>%
    ggplot(aes(x=birthyear)) +geom_density(aes(fill=gender,colour=gender),alpha=0.1)+
  facet_wrap(~usertype) +
  scale_fill_manual(values = c("#F54800","#61E000")) +
  scale_colour_manual(values = c("#F54800","#61E000")) +
  theme(panel.border = element_rect(colour="black",fill=FALSE)) +
  labs(x="Birthyear",y="Density",title = "A density plot further showing the distribution of ages of mal
```

A density plot further showing the distribution of ages of males vs females,