Analyze

This analysis aims to see the difference in use between 2 types of users: *members*, and *casuals*, using 3 factors:

1. Time: month, weekdays, and hours.
2. Location: ranked stations by use and areas.
3. Electric bike use.

Time: month

To do that, I grouped all months' use and sum the number of rides being taken in this table. Also made a column that will only have the total percentage of members, to help better illustrate the proportional use by the 2 types of users throughout the months

(this process is repited throughout all the time variable).

```
month_data<-member%>%
  group_by(month(started_at),member_casual)%>%
  count()%>%
  bind_rows(casual%>%
       group_by(month(started_at),member_casual)%>%
       count())%>%
  rename("Month"=`month(started_at)`)%>%
  mutate(Month=month.abb[Month])%>%
  ungroup()%>%
  mutate(per_mem=(member%>%count()%>%select(n))/
       ((member%>%count()%>%select(n))+(casual%>%count()%>%select(
       n))))
```

Now that it is done, I will make some graphs to help me better understand what the data is telling me.

First, arrange the months in order

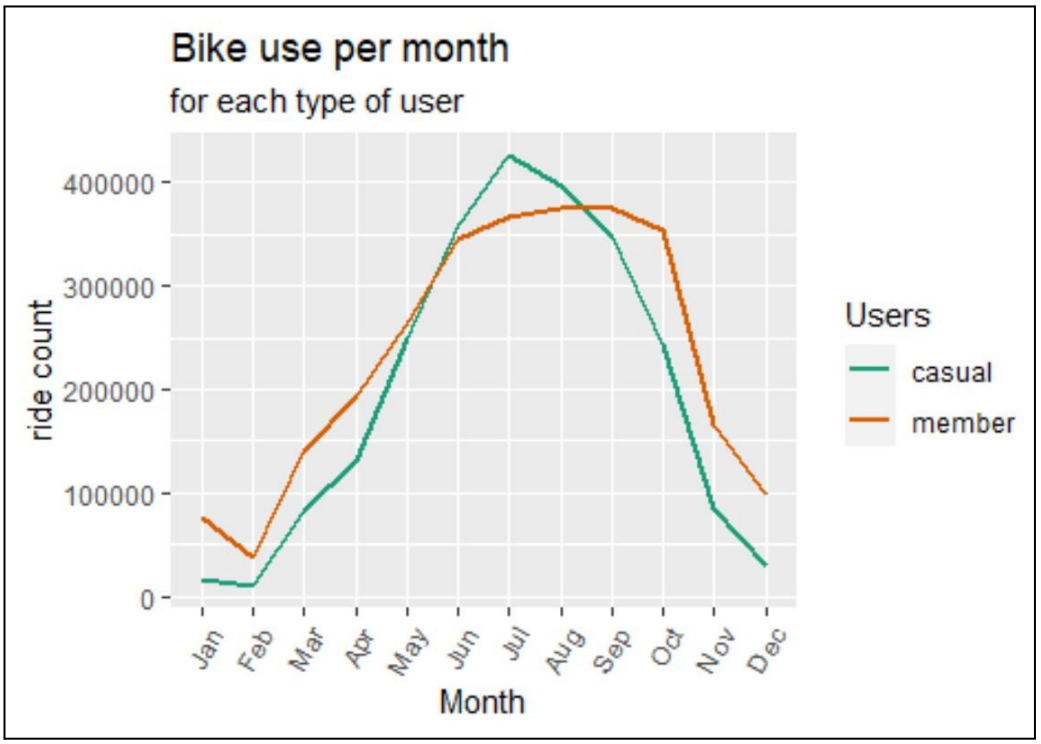
```
month_data$Month<-factor(month_data$Month,
levels=c("Jan","Feb","Mar","Apr","May","Jun","Jul","Aug","Sep",
         "Oct","Nov","Dec"))
```

Then, change the numbers display so the graph doesn't come with scientific notation, which can be useful when showing graphs to a wider audience.

```
options(scipen = 999)
```

Then, I made this graph to present the numbers of rides by each type of rider by month.

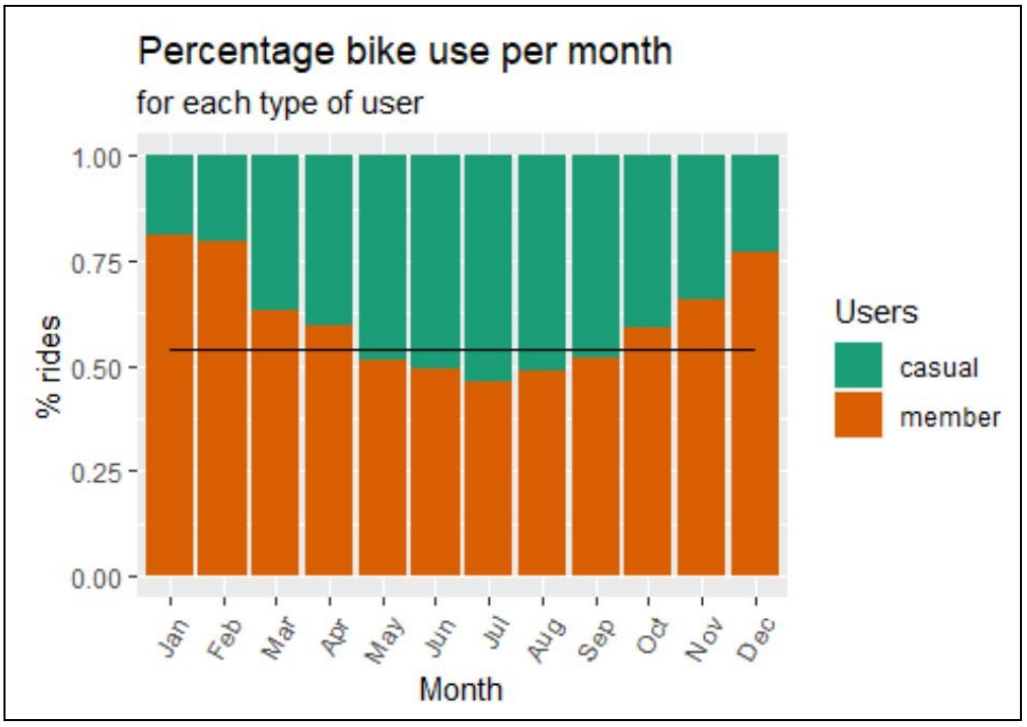
```
month_data%>%
ggplot()+geom_line(aes(Month,n,group=member_casual,color=member_casual),size=1)+
  ylab("ride count")+
  theme(axis.text.x = element_text(angle = 60, hjust = 1))+
  scale_colour_brewer(palette = "Dark2")+
  labs(color='Users',title = "Bike use per month ",
       subtitle = "for each type of user")
```



And finally, this one representing the proportion of riders by month.

```
month_data%>%
 ggplot()+geom_col(aes(Month,n,group=member_casual,fill=member_casual),position="fill")+
```

```
 ylab("% rides")+ theme(axis.text.x = element_text(angle = 60,
 hjust = 1))+
 geom_line(aes(Month,per_mem$n,group=member_casual))+
 scale_fill_brewer(palette = "Dark2")+
 labs(fill='Users',title = "Percentage of bike use per month ",
     subtitle = "for each type of user")
```



■ Time: weekdays

Table

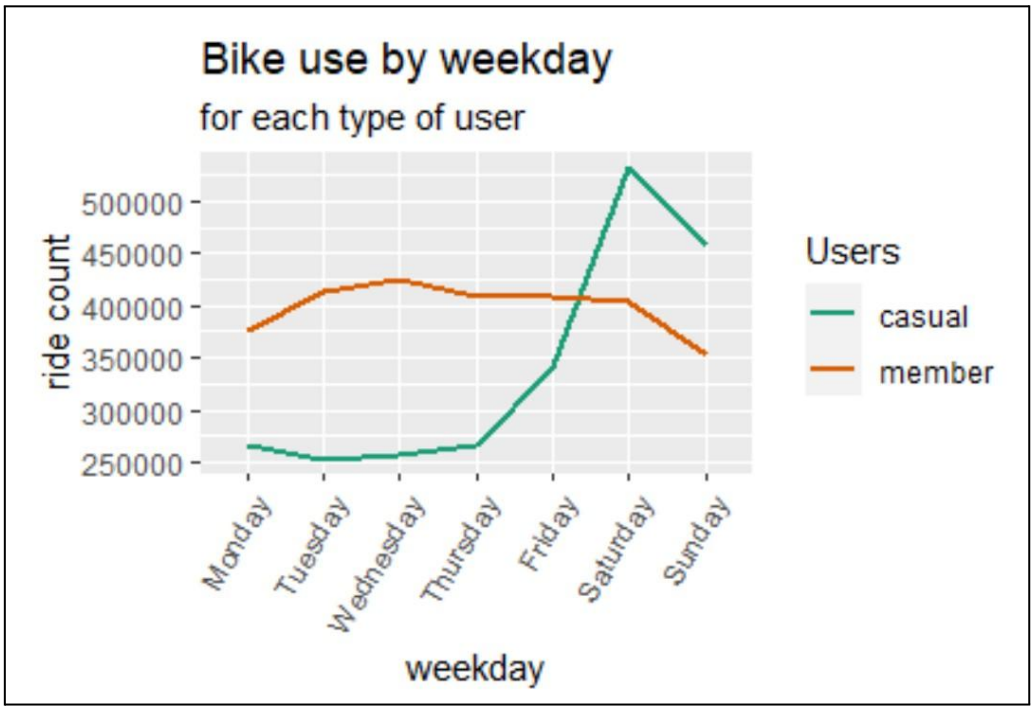
```
weekday_data<-member%>%
 group_by(weekdays(as.POSIXlt(started_at)),member_casual)%>%
 count()%>%
 bind_rows(casual%>%
     group_by(weekdays(as.POSIXlt(started_at)),member_casual)%>%
     count())%>%
 ungroup()%>%
 mutate(per_mem=(member%>%count()%>%select(n))/
     ((member%>%count()%>%select(n))+(casual%>%count()%>%select(n)
     )))
```

```
weekday_data<-weekday_data%>%
rename("weekday"=`weekdays(as.POSIXlt(started_at))`)

weekday_data$weekday<-factor(weekday_data$weekday,
levels =c
  ("Monday","Tuesday","Wednesday","Thursday","Friday","Saturday",
  "Sunday"))
```

Graph weekday by count

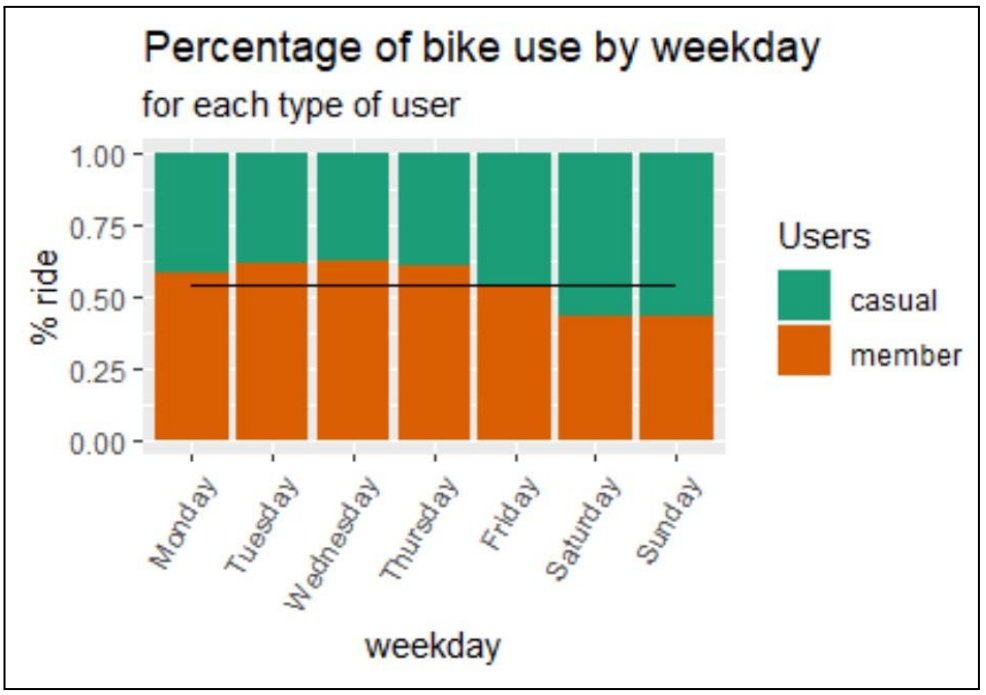
```
weekday_data%>%
  ggplot()+geom_line(aes(weekday,n,group=member_casual,color=memb
  er_casual),size=1)+
  ylab("ride count")+ theme(axis.text.x = element_text(angle =
  60, hjust = 1))+
  scale_color_brewer(palette = "Dark2")+
  labs(color='Users',title = "Bike use by weekday ",
     subtitle = "for each type of user")
```



Graph weekday by proportion

```
weekday_data%>%
  ggplot()+geom_col(aes(weekday,n,group=member_casual,fill=member
  _casual),position="fill")+
```

```
  ylab("% ride")+ theme(axis.text.x = element_text(angle = 60,
  hjust = 1))+
  geom_line(aes(weekday,per_mem$n,group=member_casual))+
  scale_fill_brewer(palette = "Dark2")+
  labs(fill='Users',title = "Percentage of bike use by weekday ",
      subtitle = "for each type of user")
```



■ Time: hour

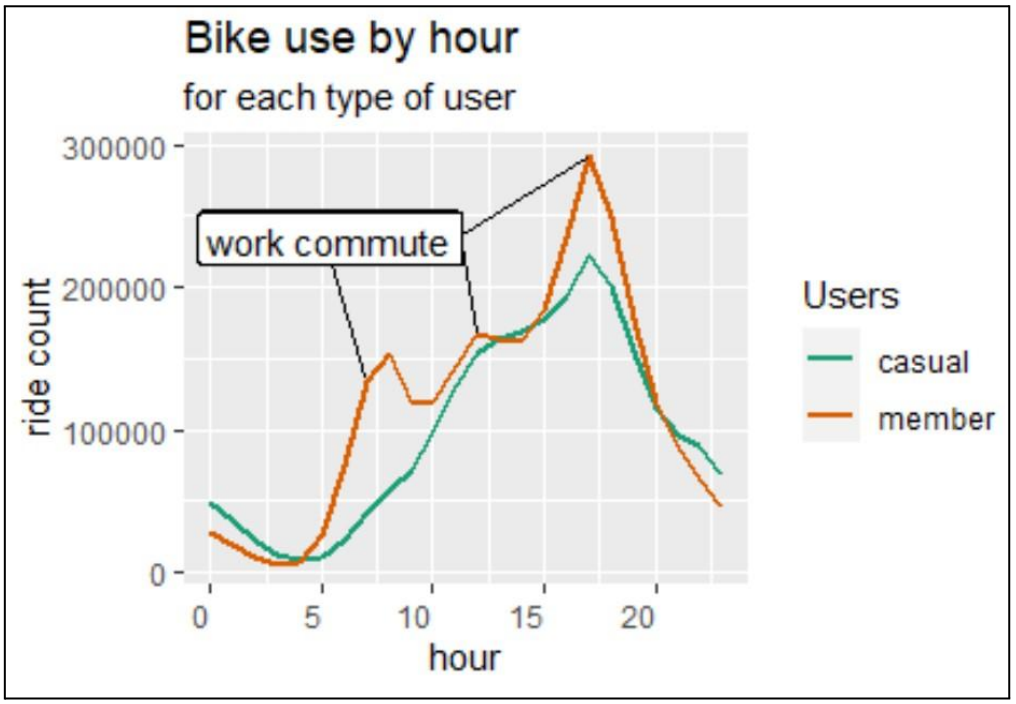
Table

```
hour_data<-member%>%
  group_by(hour(started_at),member_casual)%>%
  count()%>%
  bind_rows(casual%>%
      group_by(hour(started_at),member_casual)%>%
      count())%>%
  rename("hour"=`hour(started_at)`)%>%
  ungroup()%>%
  mutate(per_mem=(member%>%count()%>%select(n))/
      ((member%>%count()%>%select(n))+(casual%>%count()%>%select(n
      ))))
```

```
hour$hour<-factor(hour$hour,
levels=c("0","1","2","3","4","5","6","7","8","9","10","11","12"
       ,"13","14","15","16","17","18","19","20","21","22","23"))
```

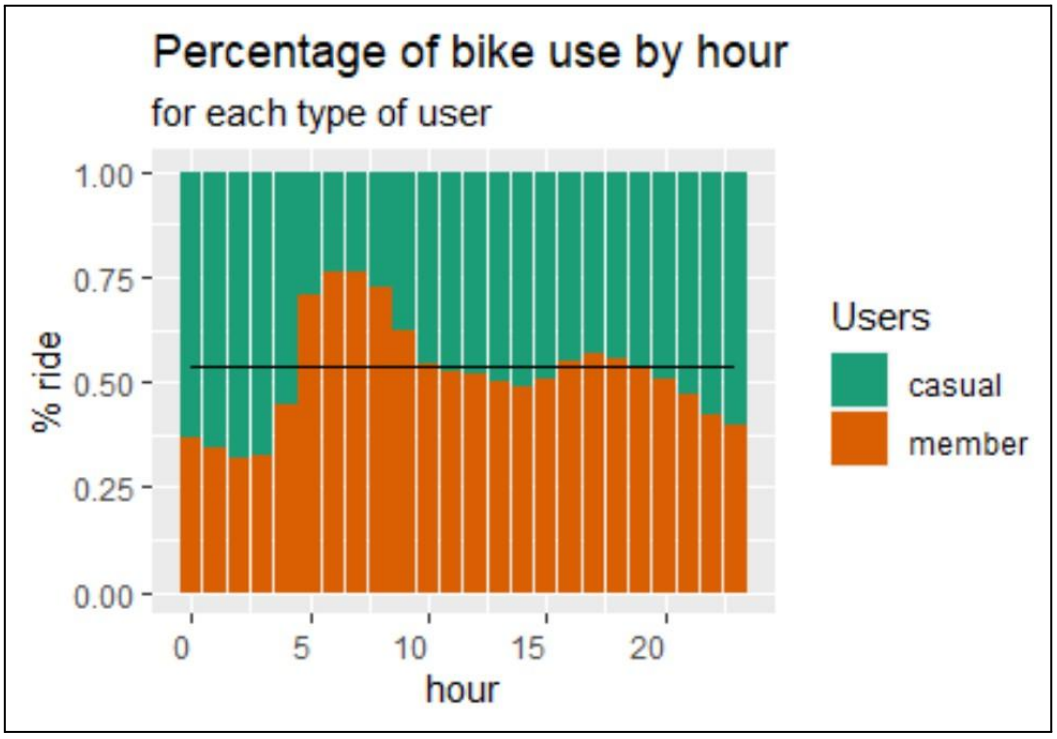
Graph hour by count. Here the hour relative graphs, also added a few labels for future presentation.

```
hour_data%>%
 ggplot(aes(hour,n))+geom_line(aes(group=member_casual,color=memb
 er_casual),size=1)+
 ylab("ride count")+ theme(axis.text.x = element_text(angle = 0,
 hjust = 1))+
 scale_color_brewer(palette = "Dark2")+
 labs(color='Users',title = "Bike use by hour ",
   subtitle = "for each type of user")+
 geom_label_repel(data = hour_data%>%filter(hour==7,
 member_casual=="member"),
       label="work commute",nudge_x = -4,nudge_y = 100000)+
 geom_label_repel(data = hour_data%>%filter(hour==12,
 member_casual=="member"),
       label="work commute",nudge_x = -9,nudge_y = 68000)+
 geom_label_repel(data = hour_data%>%filter(hour==17,
 member_casual=="member"),
       label="work commute",nudge_x = -14,nudge_y = -59000)
```



Graph hour by proportion

```
hour_data%>%
   ggplot()+geom_col(aes(hour,n,group=member_casual,fill=member_c
   asual),position="fill")+
   ylab("% ride")+ theme(axis.text.x = element_text(angle = 0,
   hjust = 1))+
   geom_line(aes(hour,per_mem$n,group=member_casual))+
   scale_fill_brewer(palette = "Dark2")+
   labs(fill='Users',title = "Percentage of bike use by hour ",
      subtitle = "for each type of user")
```



■ Location: ranked stations by use

First we need to define how we would rank the stations, I define them by proportion of use if all station would had the same proportion of use they would have 0.149%, with this in mind, I ranked them this way:

1. rank 1 (5x or more the average use, 0.742%)
2. rank 2 (4x or more the average use, 0.594%)
3. rank 3 (3x or more the average use, 0.446%)
4. rank 4 (2x or more the average use, 0.297%)
5. rank 5 (1x or more the average use, 0.149%)
6. rank 6 (less then average use)

This ranking is made for each type of user so we can see their preference.

First, I rank the *member station* use

```
member_station_count<-member%>%
  group_by(start_station_name)%>%
  count()
member_station_count$ran_station[(member_station_count$n/2789691)
 *100<(100/673)]<-"rank 6"
member_station_count$ran_station[(member_station_count$n/2789691)
 *100>(100/673)]<-"rank 5"
member_station_count$ran_station[(member_station_count$n/2789691)
 *100>(100/673*2)]<-"rank 4"
member_station_count$ran_station[(member_station_count$n/2789691)
 *100>(100/673*3)]<-"rank 3"
member_station_count$ran_station[(member_station_count$n/2789691)
 *100>(100/673*4)]<-"rank 2"
member_station_count$ran_station[(member_station_count$n/2789691)
 *100>(100/673*5)]<-"rank 1"
```

Then, I rank the *casual station* use

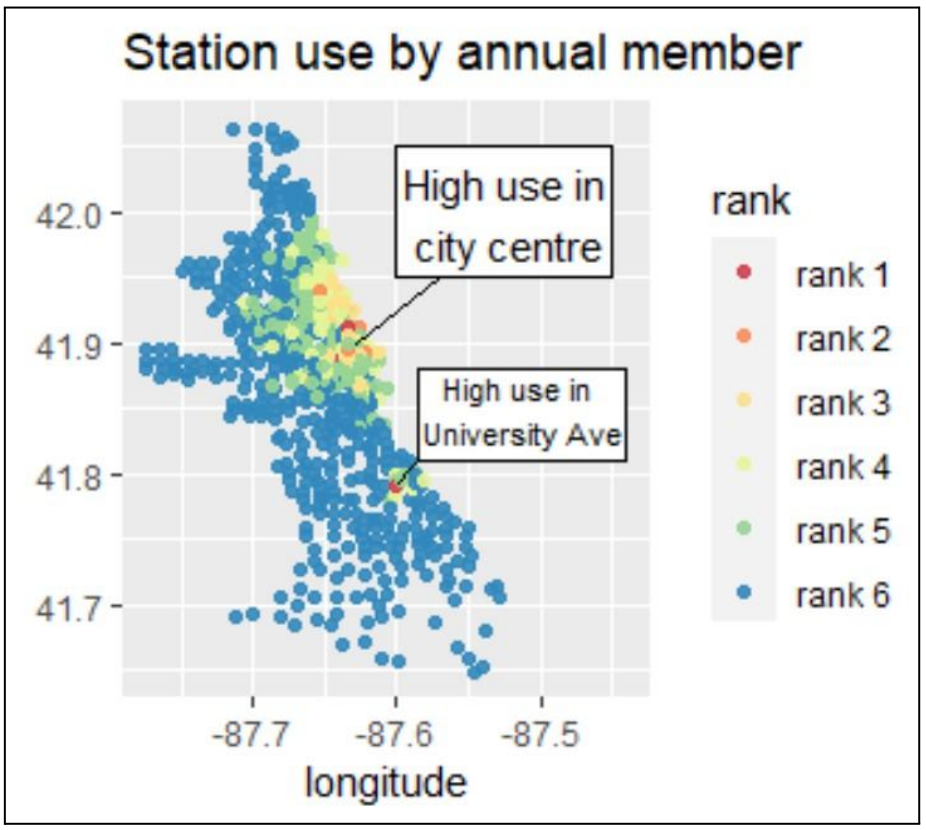
```
casual_station_count<-casual%>%
  group_by(start_station_name)%>%
  count()
casual_station_count$ran_station[(casual_station_count$n/2382677)
 *100<(100/673)]<-"rank 6"
casual_station_count$ran_station[(casual_station_count$n/2382677)
 *100>(100/673)]<-"rank 5"
casual_station_count$ran_station[(casual_station_count$n/2382677)
 *100>(100/673*2)]<-"rank 4"
casual_station_count$ran_station[(casual_station_count$n/2382677)
 *100>(100/673*3)]<-"rank 3"
```

```
casual_station_count$ran_station[(casual_station_count$n/2382677)
 *100>(100/673*4)]<-"rank 2"
casual_station_count$ran_station[(casual_station_count$n/2382677)
 *100>(100/673*5)]<-"rank 1"
```

Now that the ranking is done, I will make some graphs to better illustrate the areas that these stations occupy, along with descript annotation.

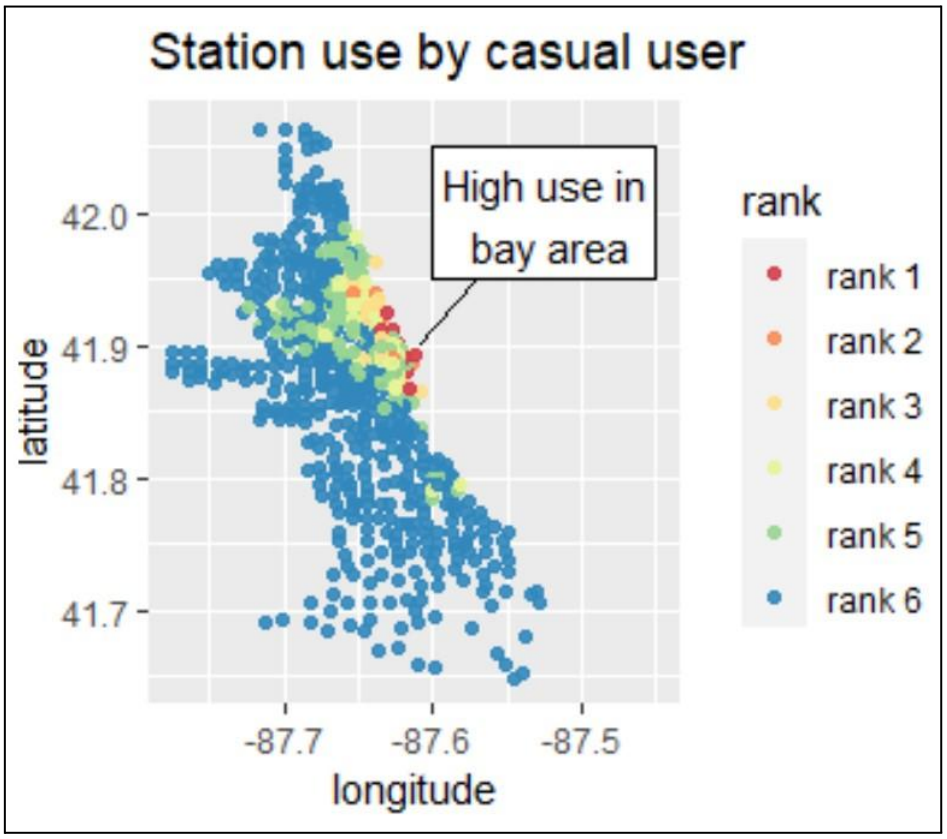
First with *members*

```
member_station_count%>%
 left_join(member%>%
     select(start_station_name,start_lat,start_lng)%>%
     distinct(),by="start_station_name")%>%
 rename("rank"=ran_station)%>%
 ggplot()+geom_point(aes(start_lng,start_lat,color=rank),alpha=.9
 )+
 scale_color_brewer(palette = "Spectral")+
 annotate("rect",xmin=-87.6,xmax = -87.45,ymin =
 41.95,ymax=42.05,
     fill="white",color="black")+
 annotate("text", x = -87.525, y =42, label = " High use in \n
 city centre")+
 annotate("segment",x = -87.553, xend=-87.6, y =41.85,
 yend=41.79)+
 xlab("longitude")+ylab("latitude")+
 annotate("rect",xmin=-87.585,xmax = -87.44,ymin =
 41.81,ymax=41.88,
     fill="white",color="black")+
 annotate("text", x = -87.515, y =41.85, label = " High use in \n
 University Ave",size=3)+
 annotate("segment",x = -87.57, xend=-87.629, y =41.95,
 yend=41.9)+
 labs(title = "Station use by annual member")+
 theme(plot.margin = margin(0,2.2,0.1,0, "cm"))
```



Then with *casuals*

```
casual_station_count%>%
  left_join(casual%>%
        select(start_station_name,start_lat,start_lng)%>%
        distinct(),by="start_station_name")%>%
  rename("rank"=ran_station)%>%
  ggplot()+geom_point(aes(start_lng,start_lat,color=rank),alpha=.
  9)+
  scale_color_brewer(palette = "Spectral")+
  annotate("rect",xmin=-87.6,xmax = -87.45,ymin =
  41.95,ymax=42.05,
        fill="white",color="black")+
  annotate("text",  x = -87.525, y =42, label = " High use in \n
  bay area")+
  annotate("segment",x = -87.57, xend=-87.61, y =41.95,
  yend=41.9)+
  xlab("longitude")+ylab("latitude")+
  labs(title = "Station use by casual user")+
  theme(plot.margin = margin(0,2.2,0.1,0, "cm"))
```



To better understand the use of these stations by the different users, I made a table that groups stations by rank, to better represent their value in overall use.

First by *members*

```
member_station_sum<-member_station_count%>%
 ungroup()%>%
 group_by(ran_station)%>%
 summarise(n=sum(n),per_rides= paste(round(
 (n/2789691)*100,2),"%"))%>%
 left_join(member_station_count%>%group_by(ran_station)%>%
    count()%>%rename("num_station"=n),by="ran_station")%>%
 rename("rank"=ran_station)
```

Then by *casuals*

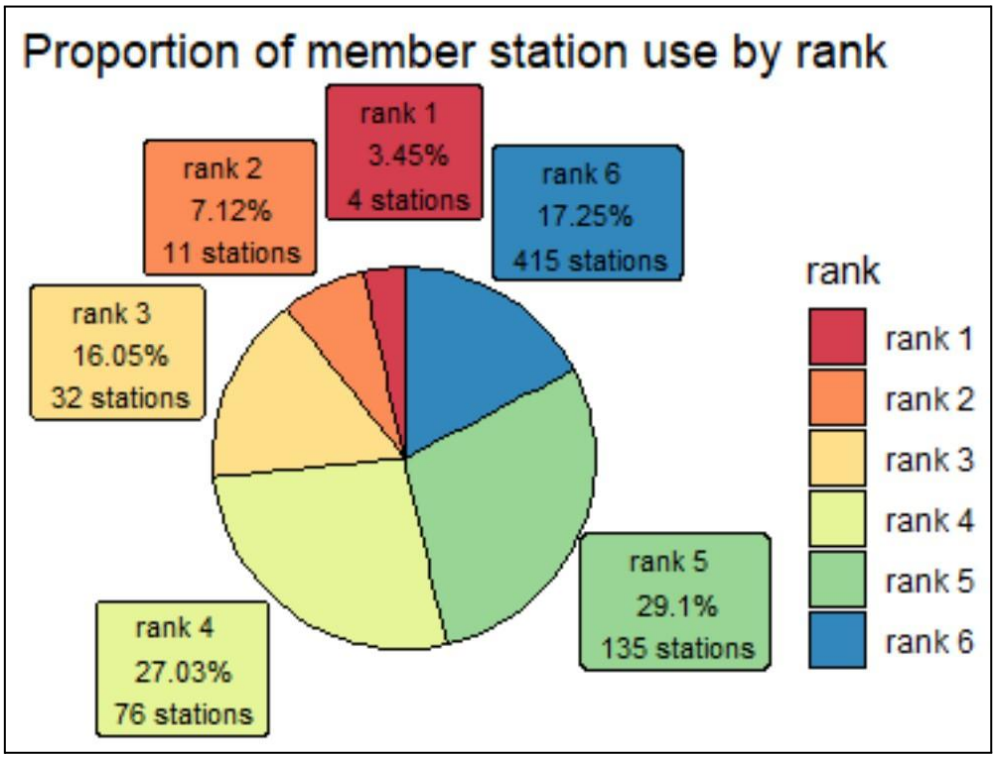
```
casual_station_sum<-casual_station_count%>%
 ungroup()%>%
 group_by(ran_station)%>%
 summarise(n=sum(n),per_rides= paste(round(
 (n/2382677)*100,2),"%"))%>%
 left_join(casual_station_count%>%group_by(ran_station)%>%
```

```
   count()%>%rename("num_station"=n),by="ran_station")%>%
 rename("rank"=ran_station)
```

Now that I have the tables, I can better represent them in a pie chart.

Members

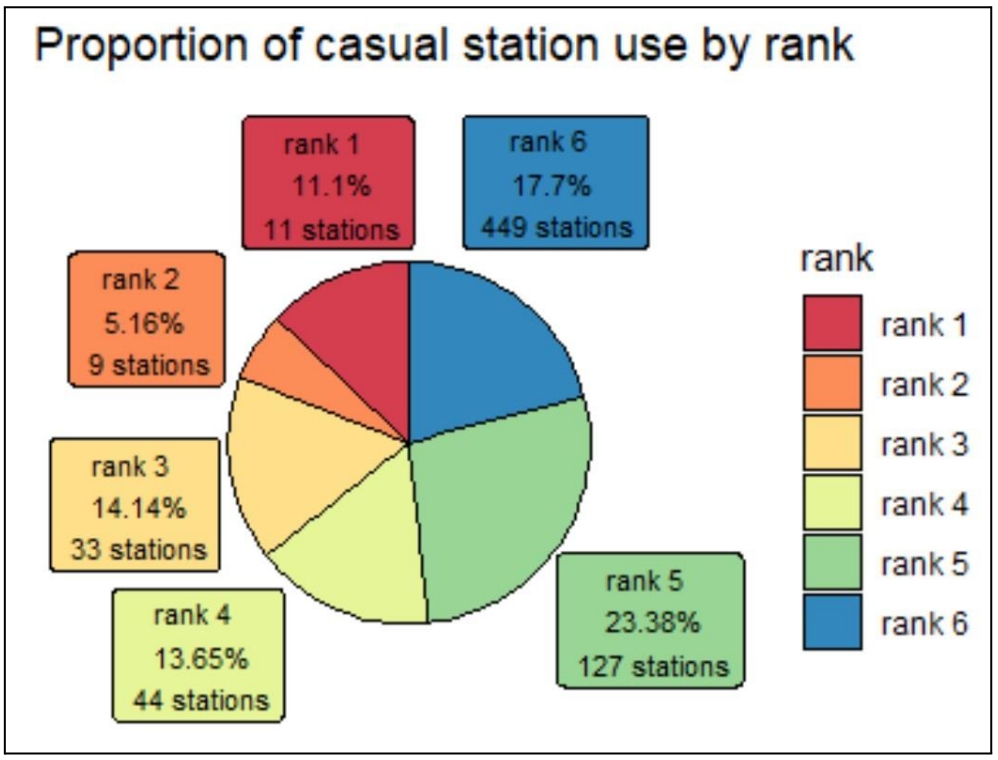
```
member_station_sum%>%
 ggplot(aes(x="",y=n,fill=rank))+
 geom_bar(stat="identity",width=1,color="black")+
 coord_polar("y",start=0)+theme_void()+
 geom_label(data=member_station_sum%>%filter(rank=="rank 1"),
           label="rank 1 \n 3.45%\n 4 stations",nudge_x
 =1.1,nudge_y = -100000,
           show.legend = "",size=3)+
 geom_label(data=member_station_sum%>%filter(rank=="rank 2"),
            label="rank 2 \n 7.12%\n 11 stations",nudge_x
 =1.1,nudge_y = 2319000,
            show.legend = "",size=3)+
 geom_label(data=member_station_sum%>%filter(rank=="rank 3"),
            label="rank 3 \n 16.05%\n 32 stations",nudge_x
 =1.1,nudge_y =1800000,
            show.legend = "",size=3)+
 geom_label(data=member_station_sum%>%filter(rank=="rank 4"),
            label="rank 4 \n 27.03%\n 76 stations",nudge_x
 =1.1,nudge_y = 1000000,
            show.legend = "",size=3)+
 geom_label(data=member_station_sum%>%filter(rank=="rank 5"),
            label="rank 5 \n 29.1%\n 135 stations",nudge_x
 =1.1,nudge_y =100000,
            show.legend ="", size=3 )+
 geom_label(data=member_station_sum%>%filter(rank=="rank 6"),
            label="rank 6 \n 17.25% \n 415 stations",nudge_x
 =1.1,nudge_y = -200000,
            show.legend = "",size=3)+
 scale_fill_brewer(palette = "Spectral")+
 labs(title = "Proportion of member station use by rank")
```



Casuals

```
casual_station_sum%>%
 ggplot(aes(x="",y=n,fill=rank))+
 geom_bar(stat="identity",width=1,color="black")+
 coord_polar("y",start=0)+theme_void()+
 geom_label(data=casual_station_sum%>%filter(rank=="rank1"),
       label="rank1\n11.1%\n11stations",nudge_x=1,nudge_y=1950000,
       show.legend="",size=3)+
 geom_label(data=casual_station_sum%>%filter(rank=="rank2"),
       label="rank2\n5.16%\n9stations",nudge_x=1.1,nudge_y=1800000
       ,show.legend="",size=3)+
 geom_label(data=casual_station_sum%>%filter(rank=="rank 3"),
       label="rank3\n14.14%\n33stations",nudge_x=1.05,nudge_y=1300
       000,
       show.legend="",size=3)+
 geom_label(data=casual_station_sum%>%filter(rank=="rank 4"),
       label="rank4\n13.65%\n44stations",nudge_x=1.15,nudge_y=1100
       000,
       show.legend="",size=3)+
 geom_label(data=casual_station_sum%>%filter(rank=="rank 5"),
       label="rank5\n23.38%\n127stations",nudge_x=1.15,nudge_y=180
       000,
```

```
      show.legend ="", size=3 )+
geom_label(data=casual_station_sum%>%filter(rank=="rank 6"),
      label="rank 6 \n 17.7% \n 449
      stations",nudge_x=1.15,nudge_y=-300000,
      show.legend="",size=3)+
scale_fill_brewer(palette="Spectral")+
labs(title="Proportion of casual station use by rank")
```



■ Location: ranked stations by area

In order to see some patterns in use by geographical position, it is important to group stations by geographical proximity. To do that, the easiest way would be by rounding coordinates, and grouping all stations to those rounded coordinates.

I rounded the coordinates to one decimal point, giving me 13 areas, a small number good for pattern recognition. This is a good method that will provide good results for the most part, however, this can (and in this case will) result in some areas being too small.

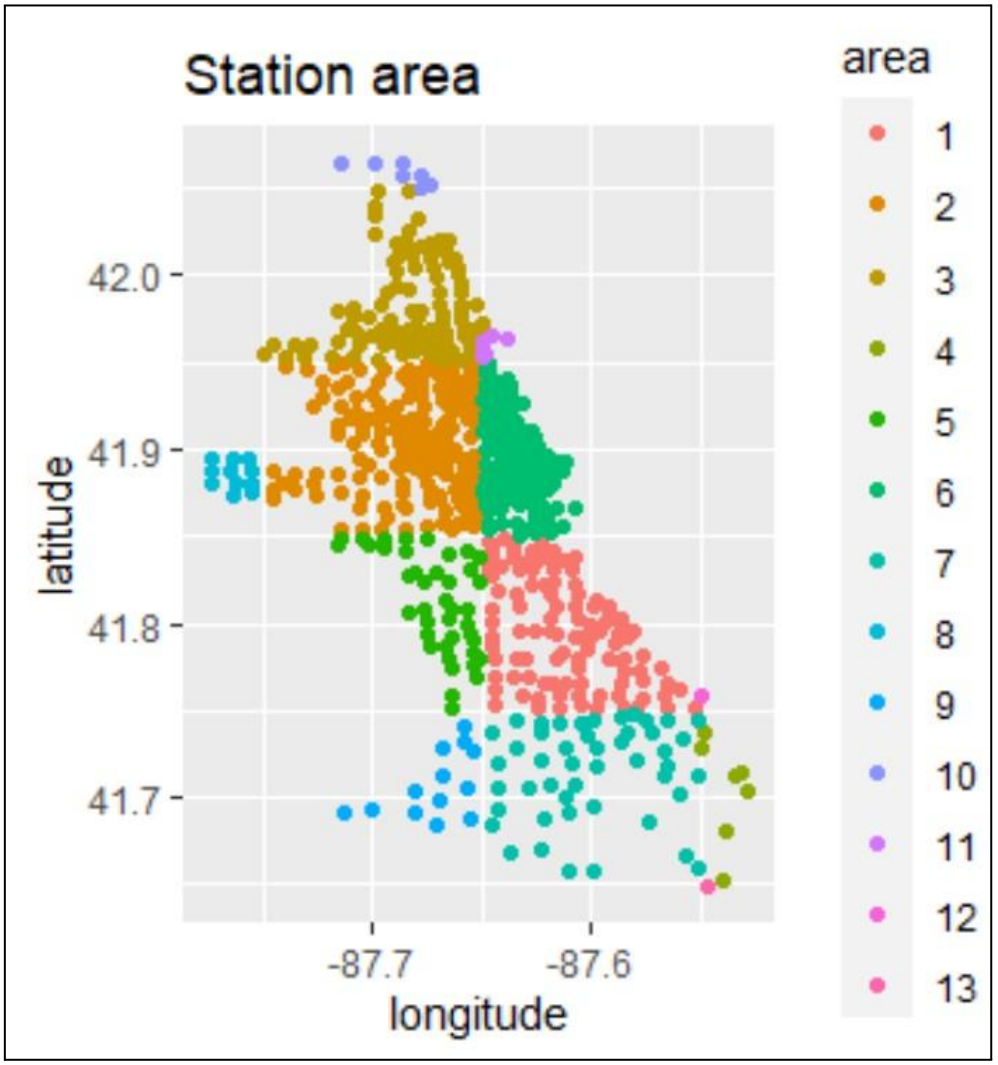
I used mean_loc, the table with all the mean coordinates of all stations to do that, like this:

```
area_station<-mean_loc%>%
mutate(round_lat=round(mean_lat,1),round_lng=round(mean_lng,1))

round_coord<-area_station%>%
select(round_lat,round_lng)%>%
distinct()

round_coord$area<-as.character(seq.int(nrow(round_coord)))

area_station<-area_station%>%
left_join(round_coord,by=c("round_lat","round_lng"))
```



As said before, there are some areas that are very small, like one station small. However, for the most used areas, they are big enough, giving me useful data. To see the overall use of areas I did it like following.

First, I grabbed the end and start station from *members*, labelled the area to them, and accounted for it's use:

```
member_area_ride_count<-bind_cols(member%>%
  left_join(area_station%>%
       select(start_station_name,area)%>%
       rename("start_area"=area),by="start_station_name")%>%
 group_by(start_area)%>%
  count()%>%
  rename("start_area_n"=n),
 member%>%
  left_join(area_station%>%
       select(start_station_name,area)%>%
       rename("end_area"=area),
       by=c("end_station_name"="start_station_name"))%>%
  group_by(end_area)%>%
  count()%>%
  rename("end_area_n"=n))
```

I do again the same to *casuals*:

```
casual_area_ride_count<-bind_cols(casual%>%
  left_join(area_station%>%
       select(start_station_name,area)%>%
       rename("start_area"=area),by="start_station_name")%>%
 group_by(start_area)%>%
  count()%>%
  rename("start_area_n"=n),
 casual%>%
  left_join(area_station%>%
       select(start_station_name,area)%>%
       rename("end_area"=area),
       by=c("end_station_name"="start_station_name"))%>%
  group_by(end_area)%>%
  count()%>%
  rename("end_area_n"=n))
```

Then, I bind then together to see the compared use of the stations by type of user:

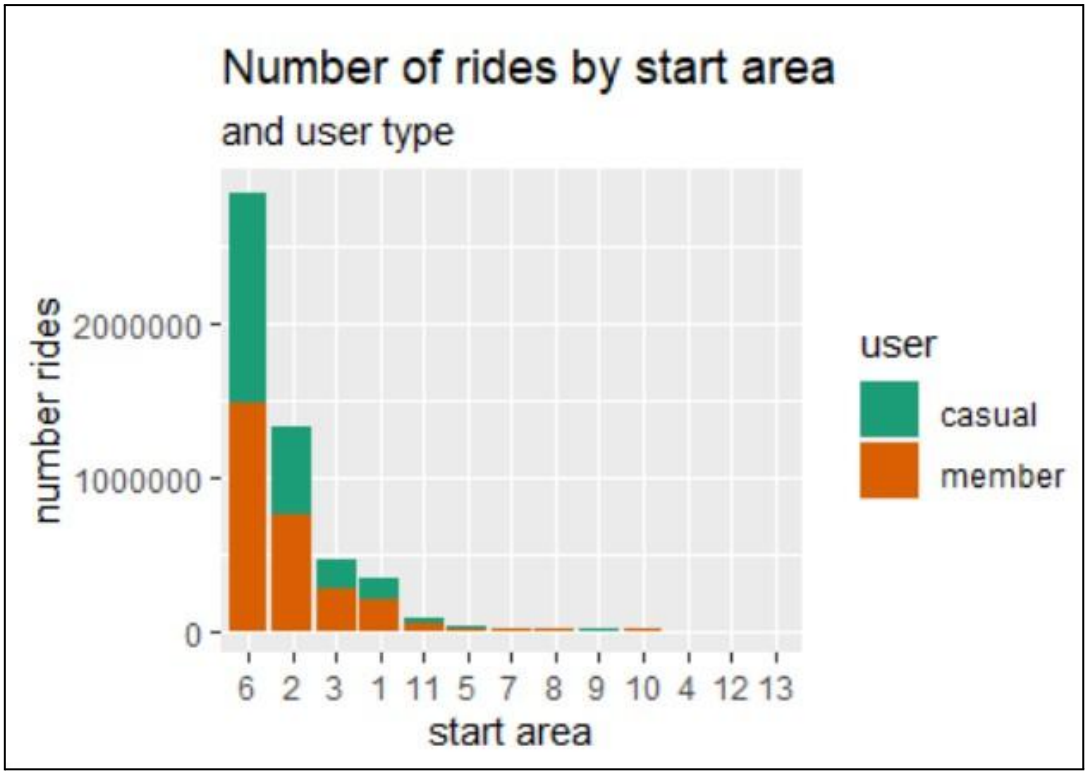
```
area_ride_count<-bind_rows(
    member_area_ride_count%>%mutate(user="member"),
    casual_area_ride_count%>%mutate(user="casual"))
```

Then, I create a new column to give me the overall proportion of types of users:

```
area_ride_count<-area_ride_count%>%
  ungroup()%>%
  group_by(user)%>%
  mutate(per_user=sum(start_area_n)/5172368)
```

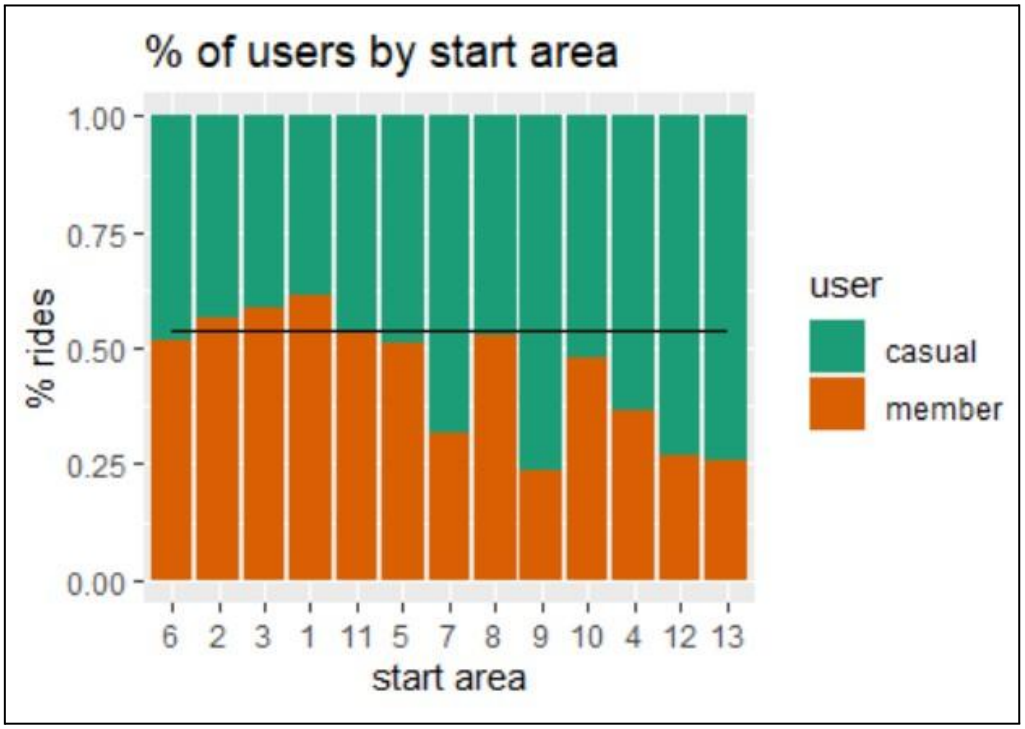
From these tables, I create a bar chart from start area use by type of user, like so:

```
area_ride_count%>%
 ggplot()+geom_col(aes(start_area,start_area_n,group=start_area,
      fill=user))+
 labs(x="start area",y="number rides", title="Number of rides by
      start area",
   subtitle="and user type")+
 scale_fill_brewer(palette="Dark2")
```



And a proportional chart, with a line indicating the overall member use, to better visualise in what areas members are overrepresented:

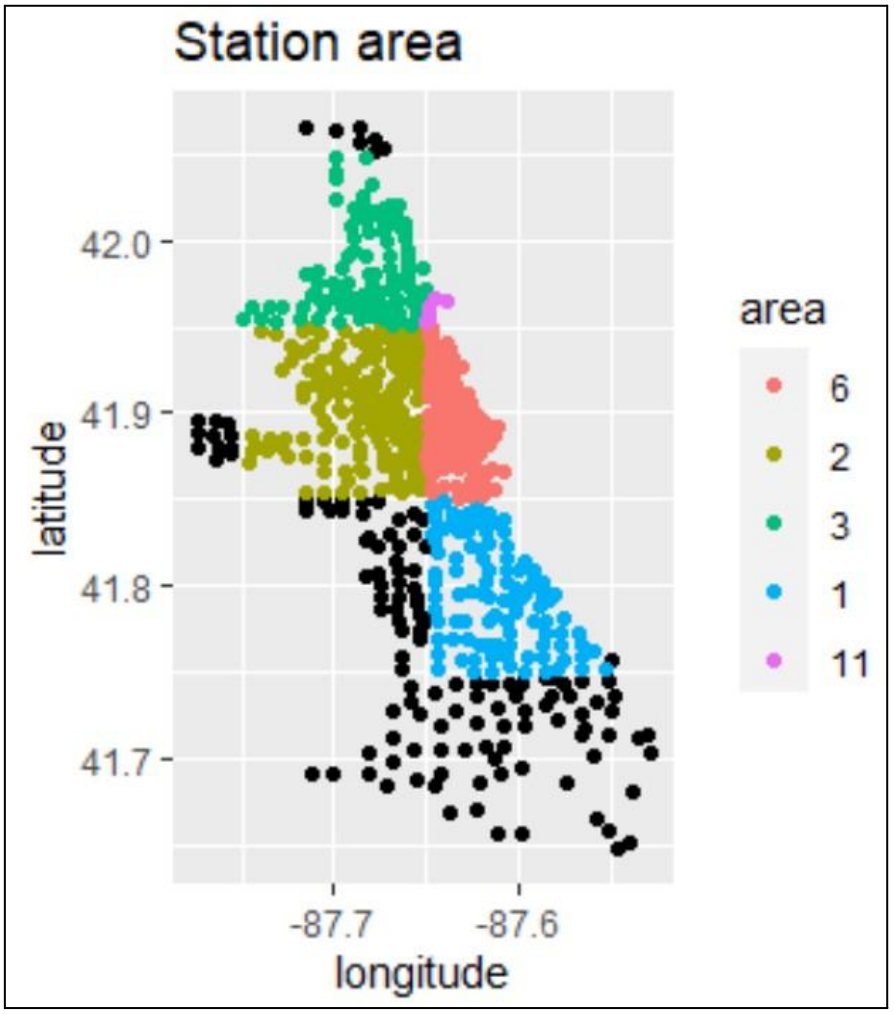
```
area_ride_count%>%
 ggplot()+geom_col(aes(start_area,start_area_n,group=start_area,
           fill=user),position="fill")+
 labs(x="start area",y="% rides", title="% of users by start
      area")+
geom_line(data=area_ride_count%>%filter(user=="member"),
   aes(start_area,per_user,group=user))+
 scale_fill_brewer(palette="Dark2")
```



From these graphs, now having an understanding on how rides are distributed throughout the areas, I made a graph to explain where I am going to focus on my analyses to avoid outliers, when looking at *electric bike use*.

```
area_station%>%
  ggplot()+geom_point(aes(start_lng,start_lat))+
  geom_point(data= area_station%>%
      filter(area%in%c(1,2,3,6,11)),aes(start_lng,start_lat,color=
            area))+
```

```
  xlab("longitude")+ylab("latitude")+labs(title="Station area")
```



And a pie chart to better explain why that is the case:

First, I will make the table needed for that pie chart:

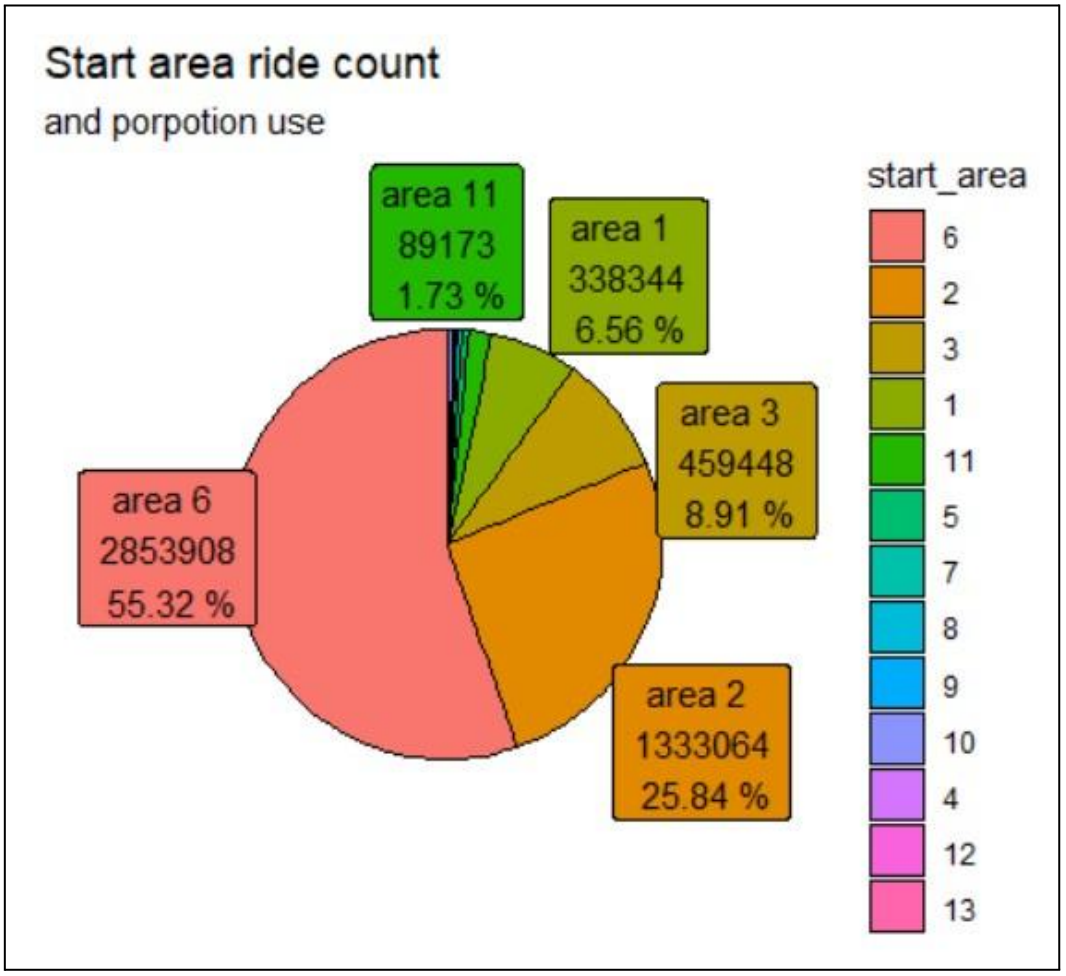
```
full_per_ride<-area_ride_count%>%
  ungroup()%>%
  group_by(start_area)%>%
  mutate(full_ride=sum(start_area_n))%>%
  select(start_area,full_ride)%>%
  distinct()%>%
  ungroup()%>%
  mutate(full_sum_ride=full_ride/sum(full_ride)*100)
```

And here I make that pie chart:

```
full_per_ride%>%
  ggplot(aes(x="",y=full_ride,fill=start_area))+
```

```
  geom_bar(stat="identity",width=1,color="black")+
  coord_polar("y",start=0)+theme_void()+
  geom_label(data=full_per_ride%>%filter(start_area=="6"),
     label=paste("area6\n",
          full_per_ride%>%filter(start_area=="6")%>%select(full_ri
          de),
          "\n",full_per_ride%>%
          filter(start_area=="6")%>%
          select(full_sum_ride)%>%
          mutate(full_sum_ride=round(full_sum_ride,2)),"%"),
     nudge_x=0.8,nudge_y=1000000,show.legend="")+
geom_label(data=full_per_ride%>%filter(start_area=="2"),
     label=paste("area2\n",
          full_per_ride%>%filter(start_area=="2")%>%select(full_ri
          de),
          "\n",full_per_ride%>%
          filter(start_area=="2")%>%
          select(full_sum_ride)%>%
          mutate(full_sum_ride=round(full_sum_ride,2)),"%"),
     nudge_x=1,nudge_y=500000,show.legend="")+
geom_label(data=full_per_ride%>%filter(start_area=="3"),
     label= paste("area3\n",
          full_per_ride%>%filter(start_area=="3")%>%select(full_ri
          de),
          "\n",full_per_ride%>%
          filter(start_area=="3")%>%
          select(full_sum_ride)%>%
          mutate(full_sum_ride=round(full_sum_ride,2)),"%"),
     nudge_x=0.9,nudge_y=600000,show.legend="")+
geom_label(data=full_per_ride%>%filter(start_area=="1"),
     label=paste("area1\n",
         full_per_ride%>%filter(start_area=="1")%>%select(full_rid
          e),
          "\n",full_per_ride%>%
          filter(start_area=="1")%>%
          select(full_sum_ride)%>%
          mutate(full_sum_ride=round(full_sum_ride,2)),"%"),
     nudge_x=1,nudge_y=150000,show.legend="")+
geom_label(data=full_per_ride%>%filter(start_area=="11"),
     label=paste("area11\n",
```

```
            full_per_ride%>%filter(start_area=="11")%>%select(full_r
            ide),
            "\n",full_per_ride%>%
            filter(start_area=="11")%>%
            select(full_sum_ride)%>%
            mutate(full_sum_ride=round(full_sum_ride,2)),"%"),
      nudge_x=.9,nudge_y=-90000,show.legend="")+
labs(title="Start area ride count",subtitle="and porpotion use")
```

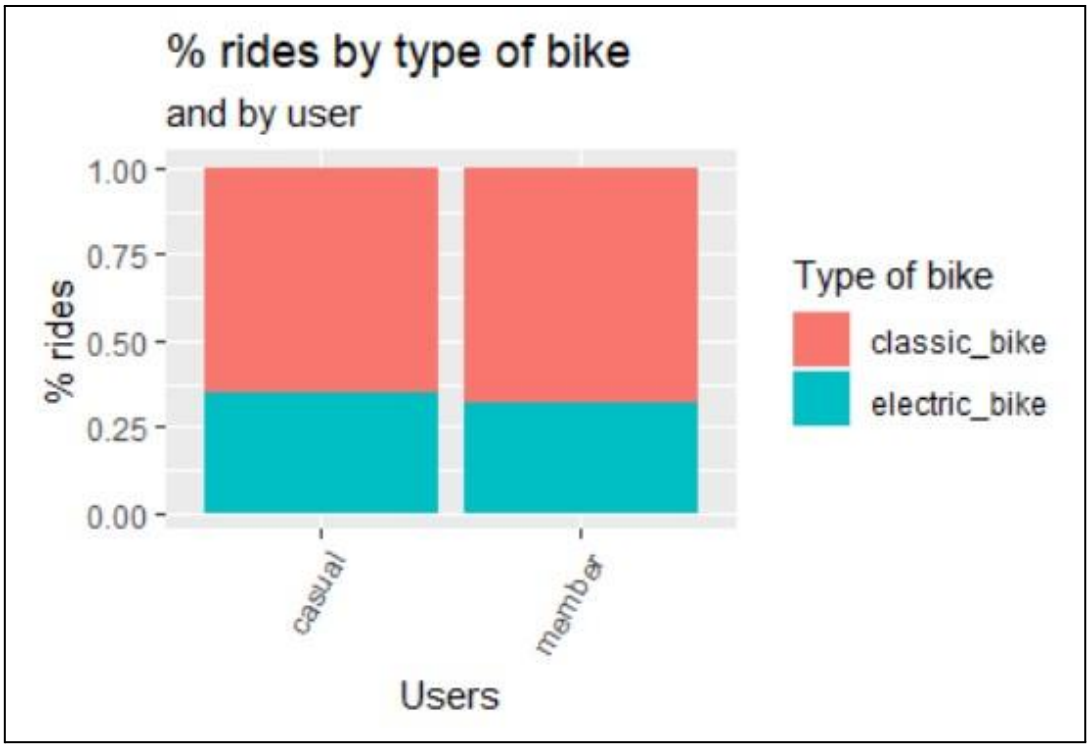


■ Electric bike use

First, let's check how overall use differ between users

```
member%>%
 bind_rows(casual)%>%
 group_by(member_casual,rideable_type)%>%
 count()%>%
```

```
ggplot()+geom_col(aes(member_casual,n,fill=rideable_type),
      position="fill")
```



As you can see, the overall use is not much different between users, however, *casuals* are overrepresented in use. And though I am not sure if electric bikes are overrepresented, as the project does not mention electric bikes numbers, only mentioning "assistive options", when talking about reclining bikes, hand tricycles and cargo bikes. One could assume that an electric bike is an "assistive option" and, if that was the case, electric bikes would be overrepresented in use, as "assistive options" represent only 8% of bikes. Regardless of the case, I hypothesize that electric bikes are overrepresented, as no different cost between electric to classic bikes is mentioned, and I personally would rather use an electric bike.

Now, let's see how that use as change throughout quarter between users.
First, make a table:

```
electric_ride_quarter<-member%>%
```

```
 group_by(member_casual,rideable_type,quarter(started_at))%>%
 count()%>%
 bind_rows(casual%>%
      group_by(member_casual,rideable_type,quarter(started_at))%>%
      count())%>%
 rename("quarter"=`quarter(started_at)`)%>%
 ungroup()%>%
 group_by(member_casual,quarter)%>%
 mutate(per_ride_RT=n/sum(n))
```

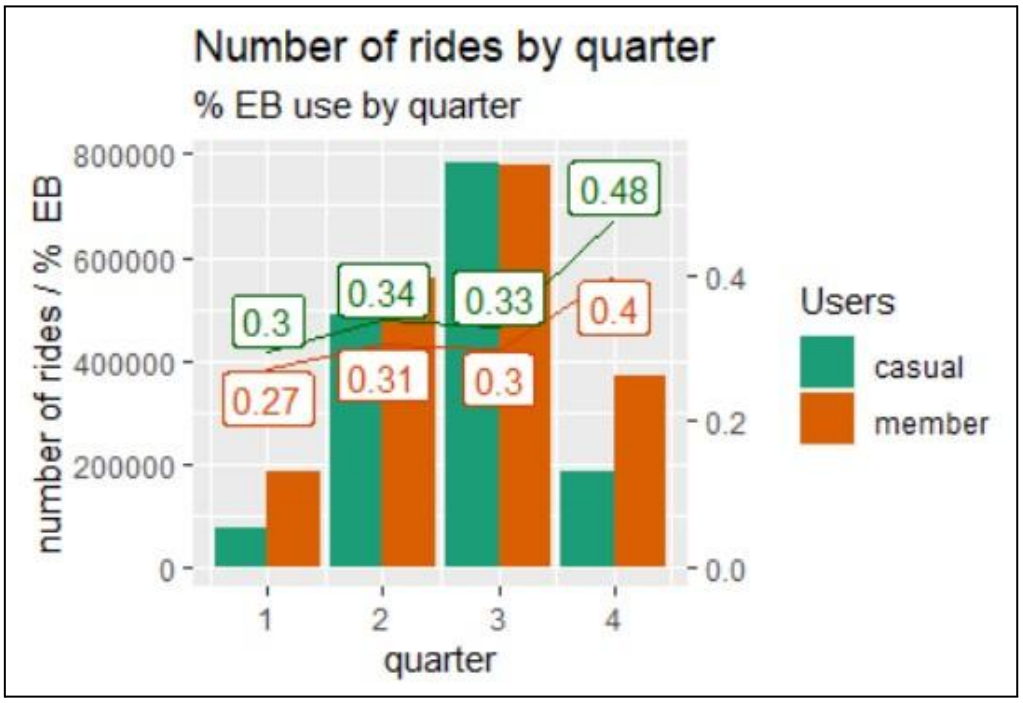
Then, create a coefficient for the secondary y axis:

```
coef<-max(electric_ride_quarter$n)/max(electric_ride_quarter$per_
ride_RT)
```

Then, a graph:

```
electric_ride_quarter%>%
 ggplot(aes(member_casual,n))+
 geom_col(aes(quarter,n,group=member_casual,fill=member_casual),p
 osition="dodge")+
 geom_line(data=electric_ride_quarter%>%filter(member_casual=="ca
 sual",
    rideable_type=="electric_bike"),
    aes(quarter,per_ride_RT*coef,group=rideable_type),color="dark
    green")+
 geom_line(data=electric_ride_quarter%>%filter(member_casual=="me
 mber",
    rideable_type=="electric_bike"),
    aes(quarter,per_ride_RT*coef,group=rideable_type),color="#D840
    00")+
 scale_y_continuous(sec.axis=sec_axis(trans=~./(coef)))+
 geom_label(data=electric_ride_quarter%>%filter(member_casual=="m
 ember",
    rideable_type=="electric_bike"),
    aes(quarter,y=coef*per_ride_RT-60000,
      label=round(per_ride_RT,2)),color="#D84000")+
 geom_label(data=electric_ride_quarter%>%filter(member_casual=="c
 asual",
    rideable_type=="electric_bike"),
    aes(quarter,y=coef*per_ride_RT+60000,
      label=round(per_ride_RT,2)),color="dark green")+
 scale_fill_brewer(palette="Dark2")+
 scale_color_brewer(palette="Dark2")+
```

```
labs(x="quarter",y="number of rides/% EB",
  title="Number of rides by quarter",
  subtitle="%EB use by quarter",fill="Users")
```



With the previous graph, we can see that throughout all the year, *casual* have had higher use of electric bikes in their rides. Let's check for hour use.

First the table:

```
hour_EB<-member%>%
  group_by(hour(started_at))%>%
  count()%>%
  bind_rows(casual%>%
      group_by(hour(started_at))%>%
      count())%>%
  rename("hour"=`hour(started_at)`)%>%
  summarise(n=sum(n))%>%
  ungroup()%>%
  mutate(per_mem=(member%>%group_by(hour(started_at))%>%count()%>
  %ungroup()%>%select(n))/
     ((member%>%group_by(hour(started_at))%>%count()%>%ungroup()%>
     %select(n))+
```

```
 (casual%>%group_by(hour(started_at))%>%count()%>%ungroup()%>%
 select(n))),
per_cas=1-per_mem,
num_EB=(casual%>%filter(rideable_type=="electric_bike")%>%
   group_by(hour(started_at))%>%count()%>%ungroup()%>%select(n
   ))+
(member%>%filter(rideable_type=="electric_bike")%>%
group_by(hour(started_at))%>%count()%>%ungroup()%>%select(n)),
per_cas_EB=(casual%>%filter(rideable_type=="electric_bike")%>%
      group_by(hour(started_at))%>%
          count()%>%ungroup()%>%select(n))/
      (casual%>%group_by(hour(started_at))%>%count()%>%ungroup
      ()%>%
      select(n)),
per_mem_EB=(member%>%filter(rideable_type=="electric_bike")%>%
      group_by(hour(started_at))%>%
      count()%>%ungroup()%>%select(n))/
(member%>%group_by(hour(started_at))%>%count()%>%ungroup()%>%s
elect(n)))
```

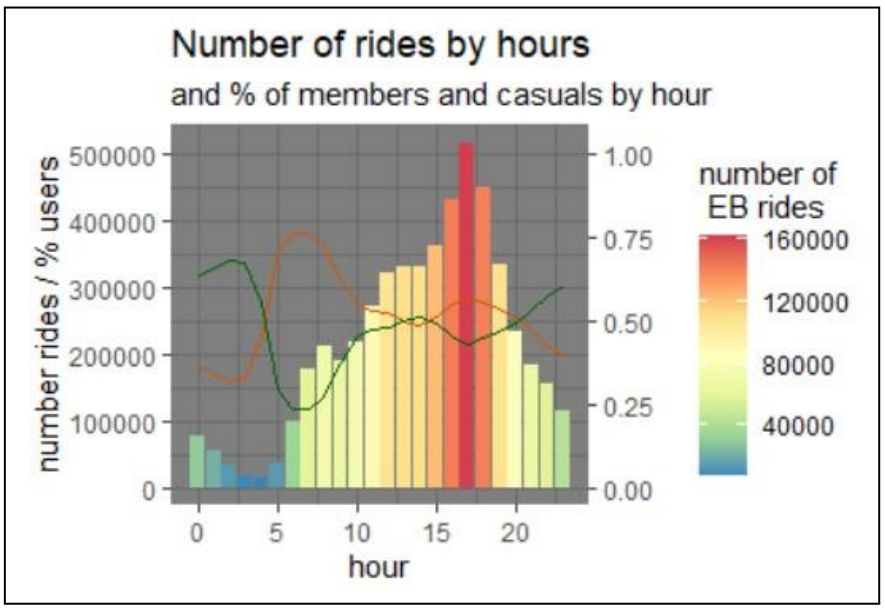
Then the graph, to show the percentage of each users and the electric bike number of use:

First define a coefficient:

```
c<-max(hour_EB$n)
```

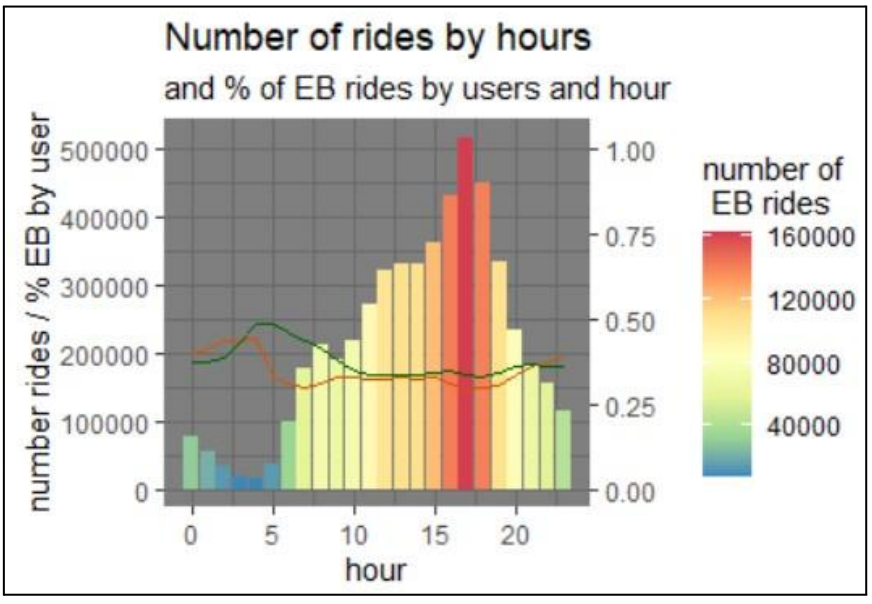
Then the graph:

```
hour_EB%>%
 ggplot()+geom_col(aes(hour,n,fill=hour_EB$num_EB$n))+
 geom_line(aes(hour,hour_EB$per_mem$n*c),color="#D85500")+
 geom_line(aes(hour,hour_EB$per_cas$n*c),color="dark green")+
 scale_y_continuous(sec.axis =sec_axis(trans=~./(c)))+
 labs(y="number rides/% users", fill="number of\n EB rides",
   title="Number of rides by hours",
   subtitle="and % of members and casuals by hour")+
 scale_fill_distiller(palette = "Spectral")+theme_dark()
```



Now that we can see the percentage of use between users, let's check their percentage of electric bike use:

```
hour_EB%>%
 ggplot()+geom_col(aes(hour,n,fill=hour_EB$num_EB$n))+
 geom_line(aes(hour,hour_EB$per_mem_EB$n*c),color="#D85500")+
 geom_line(aes(hour,hour_EB$per_cas_EB$n*c),color="dark green")+
 scale_y_continuous(sec.axis = sec_axis(trans=~./(c)))+
 labs(y="number rides/%EB by user",fill="number of\n EB rides",
   title="Number of rides by hours",
   subtitle="and % of EB rides by users and hour")+
 scale_fill_distiller(palette="Spectral")+theme_dark()
```



Interesting thing happens at morning commute hours.

Now let's check the percentage electric bike use by the percentage of the type of users in the 5 top areas.

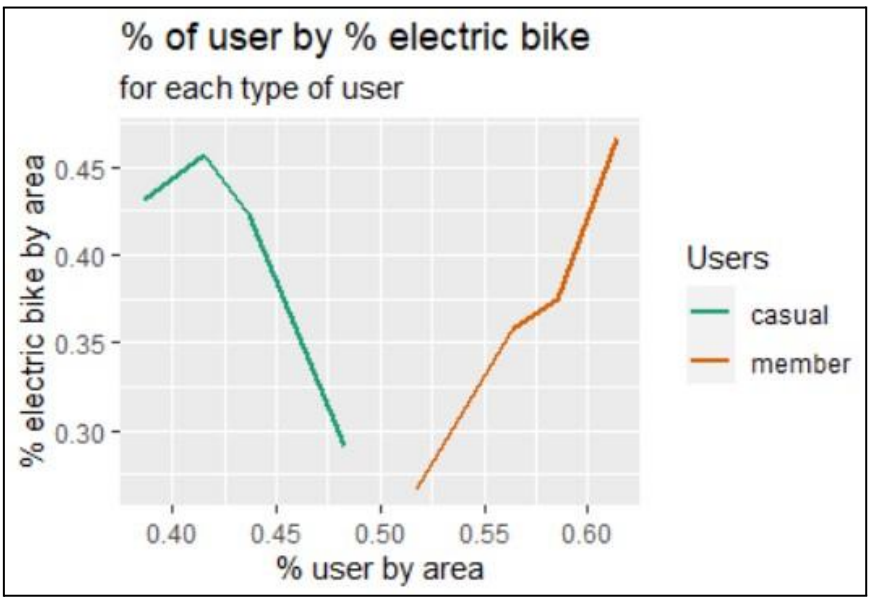
First the table:

```
area_EB<-member%>%
 left_join(area_station,
     by=c("start_station_name"))%>%
 group_by(area,rideable_type,member_casual)%>%
 count()%>%
 bind_rows(casual%>%
     left_join(area_station,
          by=c("start_station_name"))%>%
     group_by(area,rideable_type,member_casual)%>%
     count())%>%
 ungroup()%>%
 group_by(area)%>%
 mutate(count_rides_area=sum(n))%>%
 group_by(area,member_casual)%>%
 mutate(per_user=sum(n)/count_rides_area,count_user_area=sum(n))%
 >%
 group_by(area,member_casual,rideable_type)%>%
 mutate(per_EB_ride=n/count_user_area)

top_area_EB<-area_EB%>%
 mutate(area=paste("area",as.character(area)))%>%
 filter(area %in% c("area1","area2","area3","area6","area 11"))
```

Then, the graph:

```
top_area_EB%>%
 filter(rideable_type=="electric_bike")%>%
 ggplot()+
 geom_line(aes(per_user,per_EB_ride,color=member_casual),size=1)+
 scale_color_brewer(palette="Dark2")+
 labs(color='Users',title="% of user by % electric bike ",
    subtitle="for each type of user")+xlab("% user by area")+
 ylab("% electric bike by area")
```



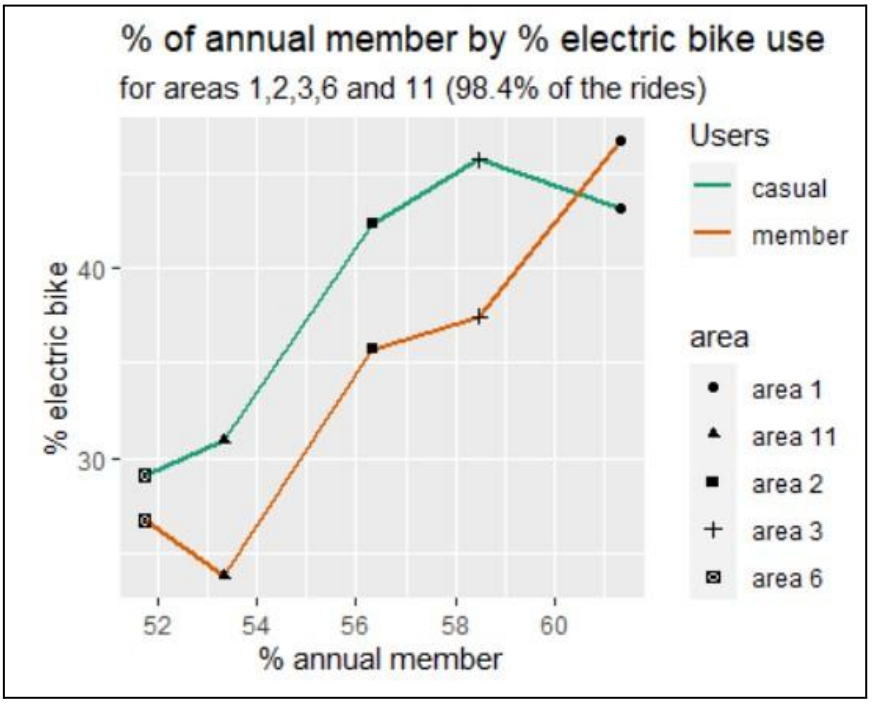
Now, let's just zoom in on the changing percentage of members and see how casual and members change their electric bike use in the top 5 areas.

First the table:

```
user_EB<-bind_rows(top_area_EB%>%
 ungroup()%>%
 filter(rideable_type=="electric_bike",member_casual=="member")%>
 %
 select(area,per_user,per_EB_ride,member_casual)%>%
 mutate(per_user=per_user*100,per_EB_ride=per_EB_ride*100),
 top_area_EB%>%
  ungroup()%>%
  filter(rideable_type=="electric_bike",member_casual=="casual")%
  >%
  select(area,per_EB_ride,member_casual)%>%
  mutate(per_EB_ride=per_EB_ride*100)%>%
  left_join(top_area_EB%>%
       ungroup()%>%
       filter(rideable_type=="electric_bike",member_casual=="membe
       r")%>%
       select(area,per_user)%>%
       mutate(per_user=per_user*100),
      by="area"))
```

Then the graph:

```
user_EB%>%
 ggplot()+
 geom_line(aes(per_user,per_EB_ride,color=member_casual),size=1)+
 scale_color_brewer(palette="Dark2")+
 labs(color='Users',title=% of annual member by % electric bike
 use",
   subtitle="for areas 1,2,3,6 and 11 (98.4% of the
   rides)")+xlab("% annual member")+
 ylab("% electric bike")+
 geom_point(aes(per_user,per_EB_ride,shape=area))
```



Here we can see that, as the percentage of members increase, so does the percentages of use of electric bikes, for both, members and casual. With 2 outliers: “area 11”, where the use of electric bikes decreases among members, and “area 1”, where the use of electric bikes decreases among casual.

Outliers:

For “area 11”, that can be explained by the fact that it is a small area next to the biggest use area, “area 6”. So I would not place much value on that data point. However, “area 1” is not

that small, and also represents more rides. Let's scrutinize that area further.

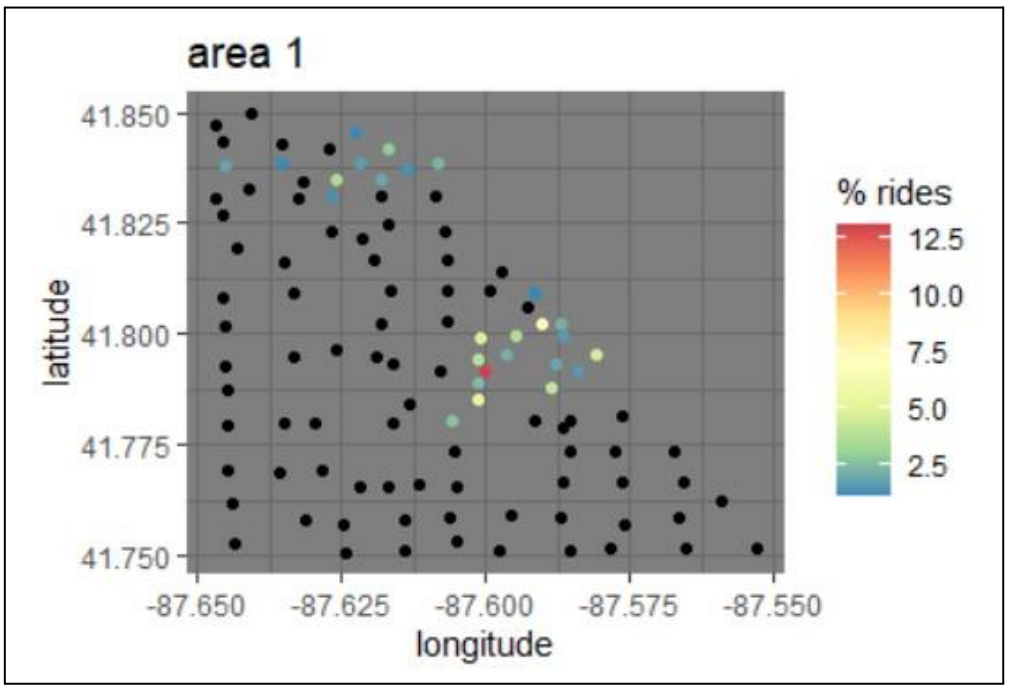
To start, we can check the distribution of rides in “area 1” for each station. This can help determine the type of member that is using the service.

First, let's make a table of number of rides of members in “area 1” stations :

```
area_1<-member%>%
 left_join(area_station%>%select(area,start_station_name),
     by="start_station_name")%>%
 filter(area==1)%>%
 group_by(start_station_name,start_lat,start_lng)%>%
 count()%>%
 arrange(-n)%>%
 ungroup()%>%
 mutate(per_ride=n/sum(n)*100)
```

Then, we make a graph:

```
area_1%>%
 ggplot()+geom_point(aes(start_lng,start_lat))+
 scale_color_distiller(palette="Spectral")+theme_dark()+
 geom_point(data=area_1%>%filter(per_ride>1) ,
     aes(start_lng,start_lat,color=per_ride))+
 labs(y="latitude",x="longitude",color="% rides",
    title="area 1")
```



As we can see, members will use the majority of "area 1" rides near University Ave. Interesting, especially in the area with the biggest membership ratio.

Now, I would like to check the number of rides in the university area in comparison with non-university area by month, and by type of users. Also, would like to check the percentage of electric bikes by users by months, to better understand the outlier.

I will make 2 tables, one in the university area, and another from the non-university area.

First university area table:

```
area_1_uni<-member%>%
 left_join(area_station%>%select(area,start_station_name),
     by="start_station_name")%>%
 filter(area==1)%>%
 filter(start_lat>=41.775 & start_lat<=41.8125)%>%
 filter(start_lng>=-87.6125 & start_lng<=-87.575)%>%
 group_by(month(started_at),member_casual)%>%
 count()%>%
 rename("Month"=`month(started_at)`)%>%
```

```
 mutate(Month=month.abb[Month])%>%
 left_join(member%>%
     left_join(area_station%>%select(area,start_station_name),
          by="start_station_name")%>%
     filter(area==1 & rideable_type=="electric_bike")%>%
     filter(start_lat>=41.775 & start_lat<=41.8125)%>%
     filter(start_lng>=-87.6125 & start_lng<=-87.575)%>%
     group_by(month(started_at),member_casual)%>%
     count()%>%
     rename("Month"=`month(started_at)`,
          "n_Eb"=n)%>%
     mutate(Month=month.abb[Month])%>%
     ungroup()%>%
     select(Month,n_Eb),by="Month")%>%
 ungroup()%>%
 group_by(Month)%>%
 mutate(per_eb=n_Eb/n*100)%>%
bind_rows(casual%>%
     left_join(area_station%>%select(area,start_station_name),
          by="start_station_name")%>%
     filter(area==1)%>%
     filter(start_lat>=41.775 & start_lat<=41.8125)%>%
     filter(start_lng>=-87.6125 & start_lng<=-87.575)%>%
     group_by(month(started_at),member_casual)%>%
     count()%>%
     rename("Month"=`month(started_at)`)%>%
     mutate(Month=month.abb[Month])%>%
     left_join(casual%>%
          left_join(area_station%>%select(area,start_station_name)
          ,
               by="start_station_name")%>%
          filter(area==1 & rideable_type=="electric_bike")%>%
          filter(start_lat>=41.775 & start_lat<=41.8125)%>%
          filter(start_lng>=-87.6125 & start_lng<=-87.575)%>%
          group_by(month(started_at),member_casual)%>%
          count()%>%
          rename("Month"=`month(started_at)`,
               "n_Eb"=n)%>%
          mutate(Month=month.abb[Month])%>%
          ungroup()%>%
          select(Month,n_Eb),by="Month")%>%
```

```
      ungroup()%>%
      group_by(Month)%>%
      mutate(per_eb=n_Eb/n*100))%>%
  ungroup()%>%
  group_by(Month)%>%
  mutate(per_user=n/sum(n)*100)
```

Then, the non-university area table:

```
area_1_non_uni<-member%>%
  left_join(area_station%>%select(area,start_station_name),
       by="start_station_name")%>%
  filter(area==1)%>%
  filter(start_lat<=41.775 | start_lat>=41.8125)%>%
  filter(start_lng<=-87.6125 | start_lng>=-87.575)%>%
  group_by(month(started_at),member_casual)%>%
  count()%>%
  rename("Month"=`month(started_at)`)%>%
  mutate(Month=month.abb[Month])%>%
  left_join(member%>%
       left_join(area_station%>%select(area,start_station_name),
            by="start_station_name")%>%
       filter(area==1 & rideable_type=="electric_bike")%>%
       filter(start_lat<=41.775 | start_lat>=41.8125)%>%
       filter(start_lng<=-87.6125 | start_lng>=-87.575)%>%
       group_by(month(started_at),member_casual)%>%
       count()%>%
       rename("Month"=`month(started_at)`,
            "n_Eb"=n)%>%
       mutate(Month=month.abb[Month])%>%
       ungroup()%>%
       select(Month,n_Eb),by="Month")%>%
ungroup()%>%
group_by(Month)%>%
mutate(per_eb=n_Eb/n*100)%>%
bind_rows(casual%>%
       left_join(area_station%>%select(area,start_station_name),
            by="start_station_name")%>%
       filter(area==1)%>%
       filter(start_lat<=41.775 | start_lat>=41.8125)%>%
       filter(start_lng<=-87.6125 | start_lng>=-87.575)%>%
       group_by(month(started_at),member_casual)%>%
       count()%>%
```

```
      rename("Month"=`month(started_at)`)%>%
      mutate(Month=month.abb[Month])%>%
     left_join(casual%>%
          left_join(area_station%>%select(area,start_station_name)
,
               by="start_station_name")%>%
          filter(area==1 & rideable_type=="electric_bike")%>%
          filter(start_lat<=41.775 | start_lat>=41.8125)%>%
          filter(start_lng<=-87.6125 | start_lng>=-87.575)%>%
          group_by(month(started_at),member_casual)%>%
          count()%>%
          rename("Month"=`month(started_at)`,
               "n_Eb"=n)%>%
          mutate(Month=month.abb[Month])%>%
          ungroup()%>%
          select(Month,n_Eb),by="Month")%>%
     ungroup()%>%
     group_by(Month)%>%
     mutate(per_eb=n_Eb/n*100))%>%
ungroup()%>%
group_by(Month)%>%
mutate(per_user=n/sum(n)*100)
```

Now, let's see the graph for the university area:

Order the months

```
area_1_uni$Month<-factor(area_1_uni$Month,
      levels =c
("Jan","Feb","Mar","Apr","May","Jun","Jul","Aug","Sep","Oct","Nov
","Dec"))
```

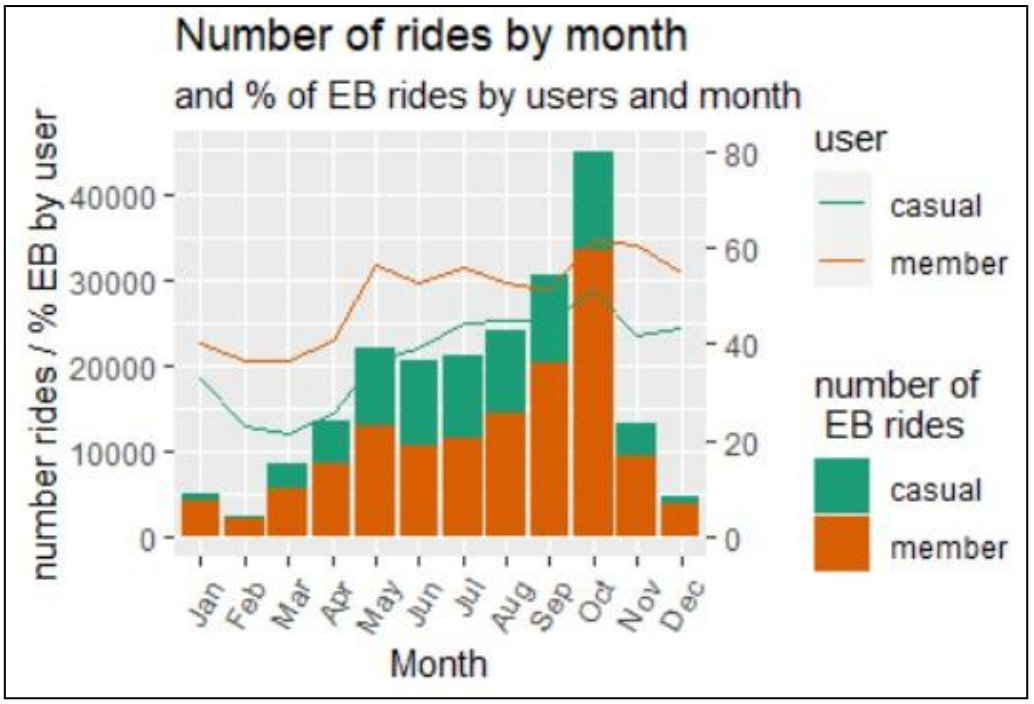
Make a coefficient

```
coof<-max(area_1_uni$n)/max(area_1_uni$per_eb)
```

Make the graph:

```
area_1_uni%>%
 ggplot()+geom_col(aes(Month,n,fill=member_casual))+
 scale_fill_brewer(palette = "Dark2")+
 geom_line(aes(Month,per_eb*coof,group=member_casual,color=member
 _casual))+
 scale_color_brewer(palette = "Dark2")+
 scale_y_continuous(sec.axis = sec_axis(trans=~./(coof)))+
```

```
 labs(y="number rides/% EB by user", fill="number of \n EB
rides",
    color="user",
    title="Number of rides by month",
    subtitle="and % of EB rides by users and month")+
 theme(axis.text.x = element_text(angle=60, hjust=1))
```



It's interesting that the only time that *members* use of electric bike is superior to the use of *casuals*. Also, it is clear to see when university starts: September. Good thing to have in mind that October data is the last month data.

Now, lets check the non-university area:

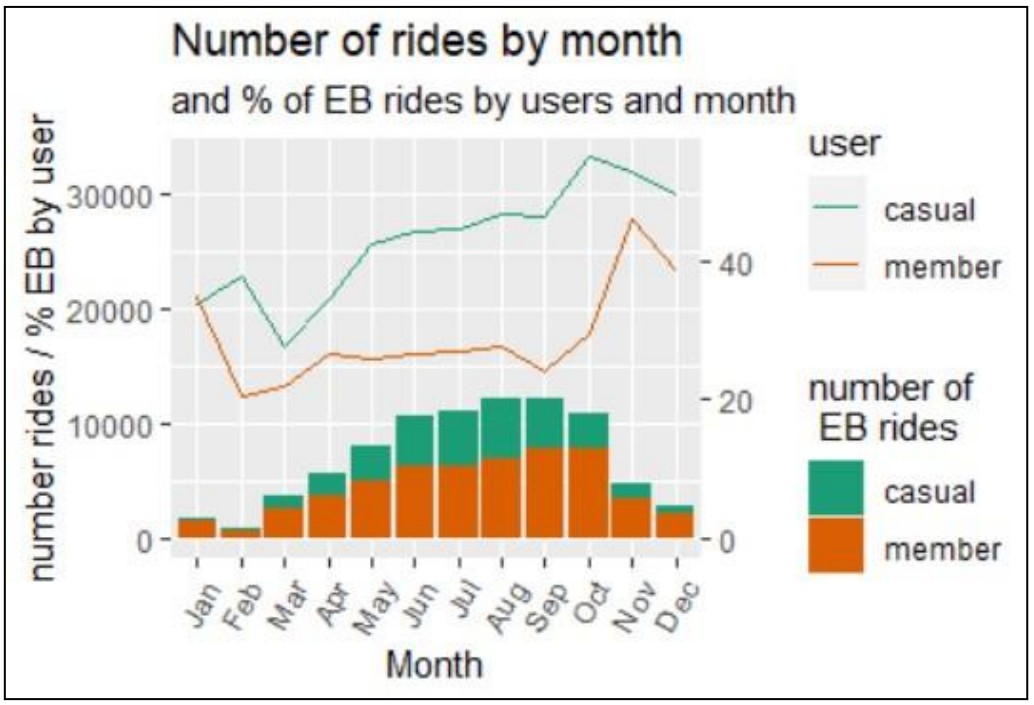
Order the months

```
area_1_non_uni$Month<-factor(area_1_non_uni$Month,
           levels=c
("Jan","Feb","Mar","Apr","May","Jun","Jul","Aug","Sep","Oct","Nov
","Dec"))
```

Get the graphs

```
area_1_non_uni%>%
 ggplot()+geom_col(aes(Month,n,fill=member_casual))+
 scale_fill_brewer(palette="Dark2")+
```

```
geom_line(aes(Month,per_eb*coof,group=member_casual,color=member
_casual))+
scale_color_brewer(palette="Dark2")+
scale_y_continuous(sec.axis=sec_axis(trans=~./(coof)))+
labs(y="number rides/% EB by user", fill="number of \n EB
rides",
  color="user",
  title="Number of rides by month",
  subtitle="and % of EB rides by users and month")+
theme(axis.text.x=element_text(angle=60, hjust=1))
```



Conclusion

- **Time metrics:**
  1. Months: both users are very seasonal, although *casuals* are more seasonal.
  2. Weekdays: *casuals* use the service mostly on the weekend.
  3. Hour: *members* use the service for commutes more than *casuals*.

- **Location:**

  1. Ranked stations: *casuals* use the service disproportionally on stations in the bay area, while *members* use the service less disproportionally. Also, where the service has higher use is in the City Center and University Ave.
  2. Area use: both use 5 areas mostly, and of those areas, only “area 6”, the most used area, is disproportionally used by *casuals*.

- **Electric use:**

  Electric bike use in areas is correlated with higher *member* use.