Google Data Analytics Capstone Project

Simona Casini

2022/11

Case Study: How Does a Bike-Share Navigate Speedy Success?

Scenario

You are a junior data analyst working in the marketing analyst team at Cyclistic, a bike-share company in Chicago. The director of marketing believes the company's future success depends on maximizing the number of annual memberships. Therefore, your team wants to understand how casual riders and annual members use Cyclistic bikes differently. From these insights, your team will design a new marketing strategy to convert casual riders into annual members. But first, Cyclistic executives must approve your recommendations, so they must be backed up with compelling data insights and professional data visualizations.

About the company

In 2016, Cyclistic launched a successful bike-share offering. Since then, the program has grown to a fleet of 5,824 bicycles that are geotracked and locked into a network of 692 stations across Chicago. The bikes can be unlocked from one station and returned to any other station in the system anytime. Until now, Cyclistic's marketing strategy relied on building general awareness and appealing to broad consumer segments. One approach that helped make these things possible was the flexibility of its pricing plans: single-ride passes, full-day passes, and annual memberships. Customers who purchase single-ride or full-day passes are referred to as casual riders. Customers who purchase annual memberships are Cyclistic members. Cyclistic's finance analysts have concluded that annual members are much more profitable than casual riders. Although the pricing flexibility helps Cyclistic attract more customers, Moreno believes that maximizing the number of annual members will be key to future growth. Rather than creating a marketing campaign that targets all-new customers, Moreno believes there is a very good chance to convert casual riders into members. She notes that casual riders are already aware of the Cyclistic program and have chosen Cyclistic for their mobility needs. Lily Moreno (the directory of marketing and your manager) has set a clear goal: Design marketing strategies aimed at converting casual riders into annual members. In order to do that, however, the marketing analyst team needs to better understand how annual members and casual riders differ, why casual riders would buy a membership, and how digital media could affect their marketing tactics. Moreno and her team are interested in analyzing the Cyclistic historical bike trip data to identify trends.

Assignement for you the first question to answer: How do annual members and casual riders use Cyclistic bikes differently?

Ask phase

Guiding questions:

- What is the problem you are trying to solve?
- How can your insights drive business decisions?

Design marketing strategies aimed at converting casual riders into annual members. In order to do that, the marketing analyst team needs to better understand how annual members and casual riders differ, why casual riders would buy a membership, and how digital media could affect their marketing tactics. **My analysis will be focused on find relations between annual members and causal riders bike usage in order to support marketing analyst new strategies**. Deep understanding on Users habits can help to find the correct communication register, platform and type of investment that can help to reach the goal and affiliate more annual members reducing casual riders.

Prepare phase

Guiding questions:

- Where is your data located?
- How is the data organized?
- Are there issues with bias or credibility in this data? Does your data ROCCC?
- How are you addressing licensing, privacy, security, and accessibility?
- How did you verify the data's integrity?
- How does it help you answer your question?
- Are there any problems with the data?

In order to reach the goal, about 12 months of data are available (from 2021/10 to 2022/10). These records have been collected by **Divvy**, the Chicagoland's bike share system across Chicago and Evanston. Divvy is a program of the Chicago Department of Transportation (CDOT), which owns the city's bikes, stations and vehicles. Initial funding for the program came from federal grants for projects that promote economic recovery, reduce traffic congestion and improve air quality, as well as additional funds from the City's Tax Increment Financing program. In 2016, Divvy expanded to the neighboring suburb of Evanston with a grant from the State of Illinois.

This data is provided according to the Divvy Data License Agreement (https://ride.divvybikes.com/data-license-agreement) and released on a monthly schedule.

Data has been declared as processed through and anonymization process in order to remove all potential link between User and related trips.

As visible in following table, data has been collected through 13 columns that comprehend:

- trip start day and time

- trip end day and time
- trip start station (latitude and longitude)
- trip end station (latitude and longitude)
- rider type (Member, Single Ride, and Day Pass)

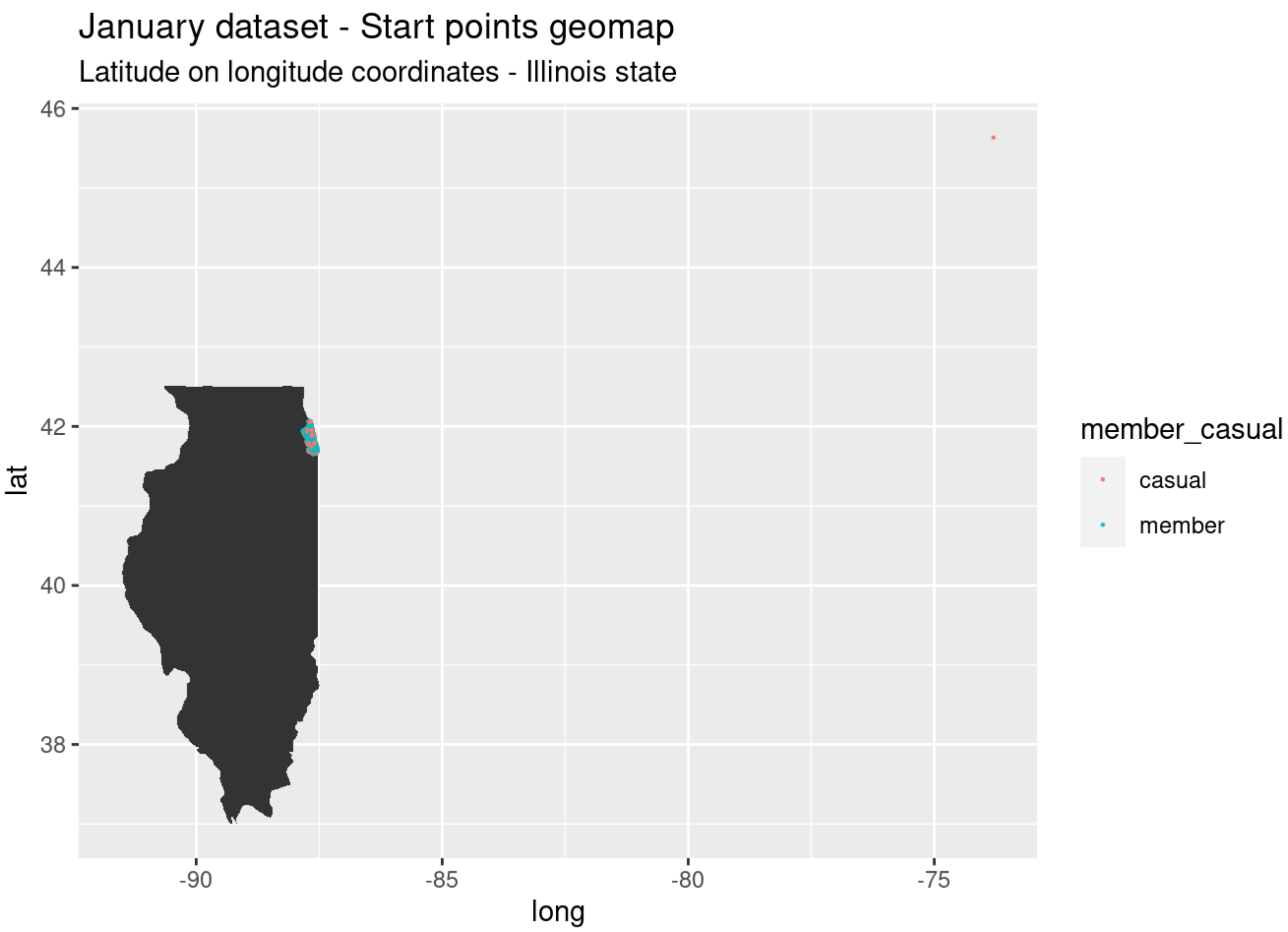
```
colnames(df)
```

```
##  [1] "ride_id"            "rideable_type"      "started_at"
##  [4] "ended_at"           "start_station_name" "start_station_id"
##  [7] "end_station_name"   "end_station_id"     "start_lat"
## [10] "start_lng"          "end_lat"            "end_lng"
## [13] "member_casual"
```

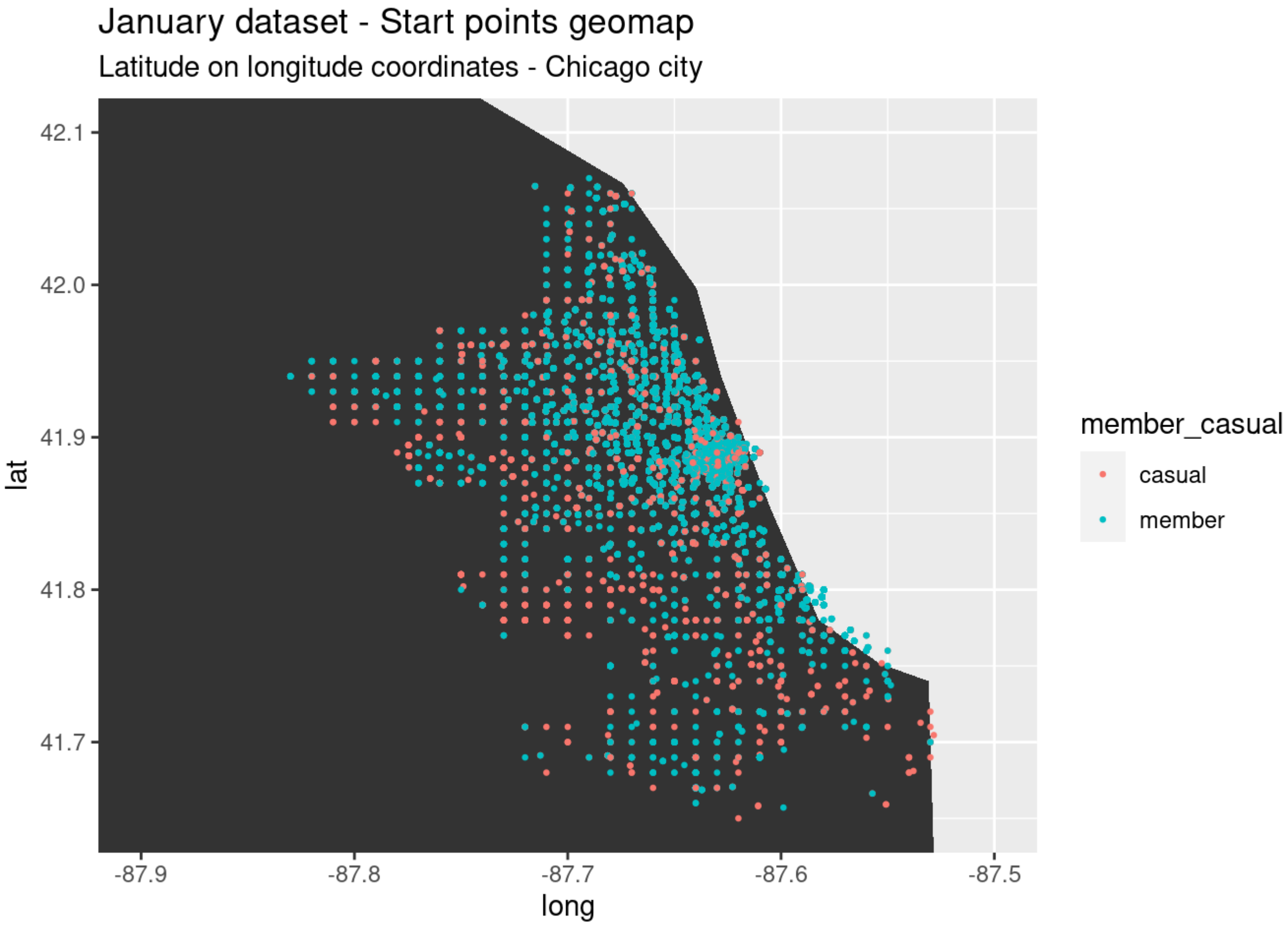
Plooting data through a map (to improve readability only january data will be visualize), is visible that there are some records that must be cleaned (following section) because latitude and longitude are corrupted but at first seems there is integrity though records. In following second figure only Chicago city will be shown. Colour groups respect to member or casual users will be defined. Also looking at this subset of data, no integrity problems have been found.

```
map_world <- map_data("county", "illinois")
```

```
ggplot(data = df) +
  geom_polygon(data = map_world, aes(x = long, y = lat, group = group)) +
  geom_point(mapping = aes(x = start_lng,
             y = start_lat,
             color = member_casual),
             size = 0.1) +
  labs(title = "January dataset - Start points geomap",
       subtitle = "Latitude on longitude coordinates - Illinois state")
```

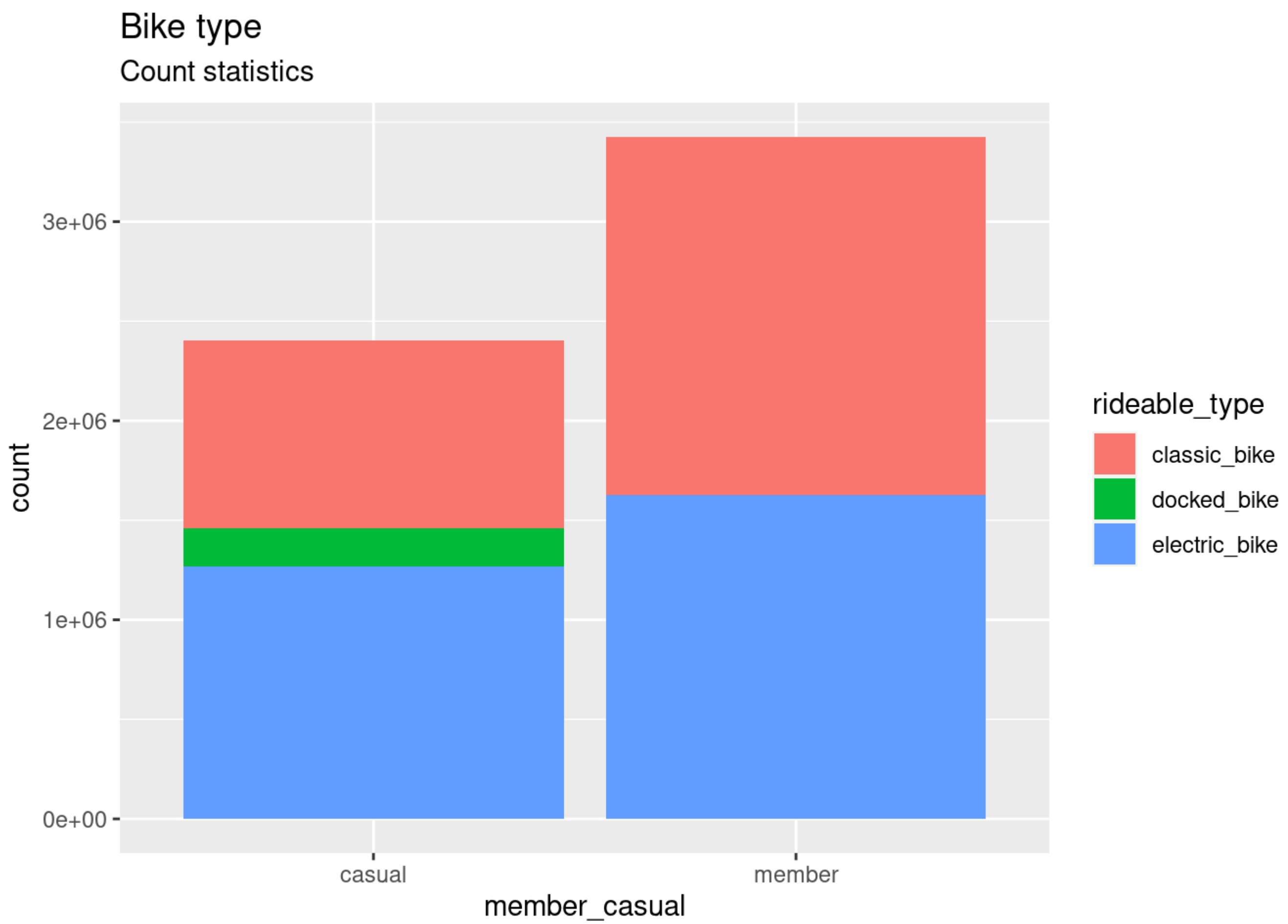

January dataset - Start points geomap
Latitude on longitude coordinates - Illinois state
46
44
42
40
38
lat
-90
-85
-80
-75
long
member_casual
casual
member

```
ggplot(data = df) +
  geom_polygon(data = map_world, aes(x = long, y = lat, group = group)) +
  coord_cartesian(xlim = c(-87.9, -87.50), ylim = c(41.65, 42.1)) +
  geom_point(mapping = aes(x = start_lng,
             y = start_lat,
             color = member_casual),
             size = 0.5) +
  labs(title = "January dataset - Start points geomap",
       subtitle = "Latitude on longitude coordinates - Chicago city")
```


January dataset - Start points geomap
Latitude on longitude coordinates - Chicago city
42.1
42.0
41.9
41.8
41.7
lat
-87.9
-87.8
-87.7
-87.6
-87.5
long
member_casual
casual
member

Another important aspect that could be taken into account during this analysis is the bike type and stations used. Considering related columns, in order to verify ROCCC, following figure explores dataset records (whole dataset - 12 months)

```
ggplot(data = df_12, aes(member_casual)) +
  geom_bar() +
  geom_bar(aes(fill = rideable_type)) +
  labs(title = "Bike type", subtitle = "Count statistics") #nolint
```



Data ROCCC analysis

Data integrity, credibility and accessibility has been analysed thorough ROCCC system.

Reliability: These data are reliable. 12 months of data have been provided and anonymized. There is sufficient information to perform the analysis. Future in depth investigations could be considered *older* data since in 2020/2021 *COVID-19* pandemy could have affect data in these specific amount of months. A particular attention must be however considered, since User habits in more than 2 years could be changed.

Originality: These are the original dataset collected directly by **Divvy**. Only anonymization has been provided for safety and privacy reasons.

Comprehensiveness: These data are comprehensive. There are sufficient information about Users and bikes types (for our purpose). A filter process, in order to erase data that could be potentially corrupeted will be provided (ex. trip duration outplier). A note on stations name: a lot of records have been collected with corrupted data (on latitude and longitude or on station name). A clean process must be designed in order to use only correct data.

Current: Data are not oudated (currenty year collection process).

Cited: **Divvy** could be considered as credible source.

Data integrity and credibility is thus sufficient to provide reliable and comprehensive insights.

Process phase

Guiding questions:

- What tools are you choosing and why?
- Have you ensured your data's integrity?
- What steps have you taken to ensure that your data is clean?
- How can you verify that your data is clean and ready to analyze?
- Have you documented your cleaning process so you can review and share those results?

Note: al the project will be developed in **R**. Some figures have been designed in **Tableau**.

Note: not corrupted record, in this case, means that the single record has a valid value. For example a valid station name or a valid latitude and longitude coordinates.

During this phase, whole datset has been cleaned in order to collect and use only data that are not corrupted. In particular following steps will be designed:

- check station name integrity;
- check member attribute integrity;
- check latitude and longitude coordinates.

Moreover, three columns will be added in order to add information that can be useful during the analysis phase:

- ride lenght;
- day of the week (from date attribute);
- month of the year (from date attribute);
- year (since last 12 month data are both from 2021 and 2022, check on years could be useful).

First step of cleaning process has been designed using a pipeline. In particular strings about station ID, name and type of User have been checked. Using *Spreadsheet* instrument has been infact verified that 5 chars are the minimum number though which these string attributes can be considered as valid. Following code shows filter definitions considering this boundary.

```
min_string_chars <- 5

df_sub <- df_12 %>%
  subset(nchar(as.character(start_station_id)) > min_string_chars &
         nchar(as.character(start_station_name)) > min_string_chars) %>%
  subset(nchar(as.character(end_station_id)) > min_string_chars &
         nchar(as.character(end_station_name)) > min_string_chars) %>%
  subset(nchar(as.character(member_casual)) > min_string_chars)
```

New attributes have been added using date and hours information.

```
df_sub$ride_length <- df_sub$ended_at - df_sub$started_at
df_sub$DoW <- weekdays(as.Date(df_sub$started_at))
df_sub$Month <- month(df_sub$started_at)
df_sub$Year <- year(df_sub$started_at)
df_sub$Hour <- format(as.POSIXct(df_sub$started_at), format = "%H")

head(select(df_sub, ride_length, DoW, started_at, ended_at), 10)
```

```
##     ride_length       DoW          started_at            ended_at
##  1:    163 secs Wednesday 2021-10-06 13:55:33 2021-10-06 13:58:16
##  2:    125 secs  Saturday 2021-10-23 23:33:22 2021-10-23 23:35:27
##  3:   2210 secs    Friday 2021-10-01 13:47:06 2021-10-01 14:23:56
##  4:   1422 secs   Tuesday 2021-10-05 16:01:42 2021-10-05 16:25:24
##  5:   7646 secs  Saturday 2021-10-23 10:32:04 2021-10-23 12:39:30
##  6:    205 secs Wednesday 2021-10-27 12:55:55 2021-10-27 12:59:20
##  7:    255 secs Wednesday 2021-10-27 12:33:54 2021-10-27 12:38:09
##  8:   1243 secs  Thursday 2021-10-21 10:37:08 2021-10-21 10:57:51
##  9:    628 secs   Tuesday 2021-10-26 17:25:05 2021-10-26 17:35:33
## 10:    702 secs Wednesday 2021-10-13 16:45:07 2021-10-13 16:56:49
```

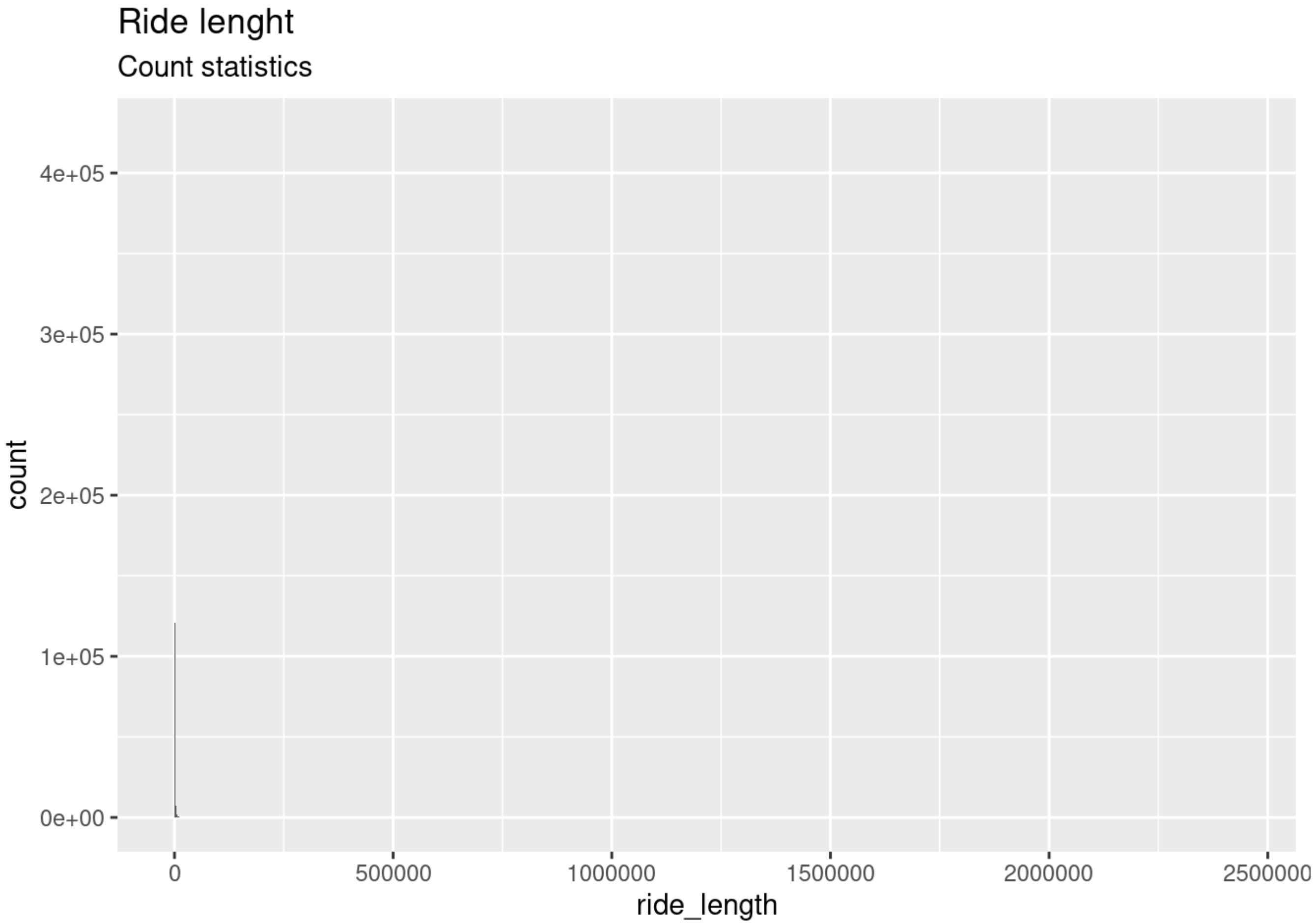
Adding new columns, dataset is then composed by following attributes:

```
colnames(df_sub)
```

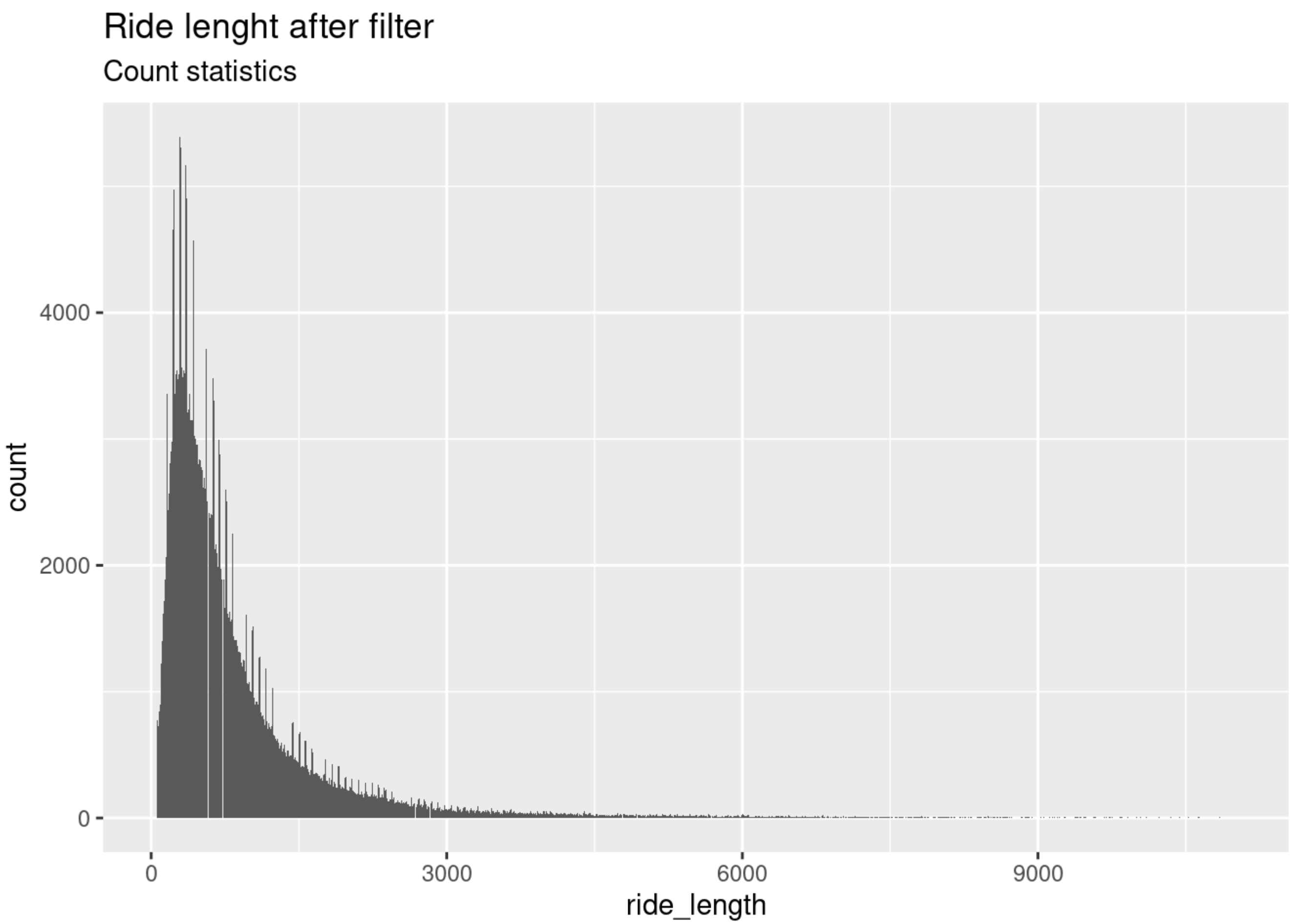
```
##  [1] "ride_id"            "rideable_type"      "started_at"        
##  [4] "ended_at"           "start_station_name" "start_station_id"  
##  [7] "end_station_name"   "end_station_id"     "start_lat"         
## [10] "start_lng"          "end_lat"            "end_lng"           
## [13] "member_casual"      "ride_length"        "DoW"               
## [16] "Month"              "Year"               "Hour"
```

The new ride lenght column has been used to check also the integrity of data collected in *start* and *end* ride data column and to verify the presence of potential outliers. Following figure (histogram distribution) shows that there are multiple records that are clearly outliers and that represent rides lenght that are not reliable (11000 secs, about 3 hours is the time evaluated as the best boundary). Also in this case, the cleaning process provides as output only data that are not corrupted.

```
ggplot(data = df_sub, aes(x = ride_length)) +
  geom_histogram(bins = 9000) +
  labs(title = "Ride lenght", subtitle = "Count statistics")
```



```
df_sub <- df_sub %>%
  subset(ride_length < 11000 & ride_length > 60)

ggplot(data = df_sub, aes(x = ride_length)) +
  geom_histogram(bins = 5000) +
  labs(title = "Ride lenght after filter", subtitle = "Count statistics")
```



Dataset obtained from previous operations will be saved in a new *csv* file. Percentages of corrupted data will be print below.

```
## Original dataset lenght: 5828235
```

```
## Managed dataset lenght: 1437108
```

```
## Percentages of corrupted data: 24.65769
```

Analyze and share phase

Guiding questions:

- How should you organize your data to perform analysis on it?
- Has your data been properly formatted?
- What surprises did you discover in the data?
- What trends or relationships did you find in the data?
- How will these insights help answer your business questions?
- Were you able to answer the question of how annual members and casual riders use Cyclistic bikes differently?
- What story does your data tell?
- How do your findings relate to your original question?
- Who is your audience? What is the best way to communicate with them?
- Can data visualization help you share your findings?
- Is your presentation accessible to your audience?

In order to understand Causual User habits respect to member one, some statistic analysis and graphical distributions have been analysed and provided.

Considering the global number of trips in 12 months, 37.8% of rides have been done by causual User while 62% has been done by Divvy members.

```
## Number of riders - 12 months 1437108
```

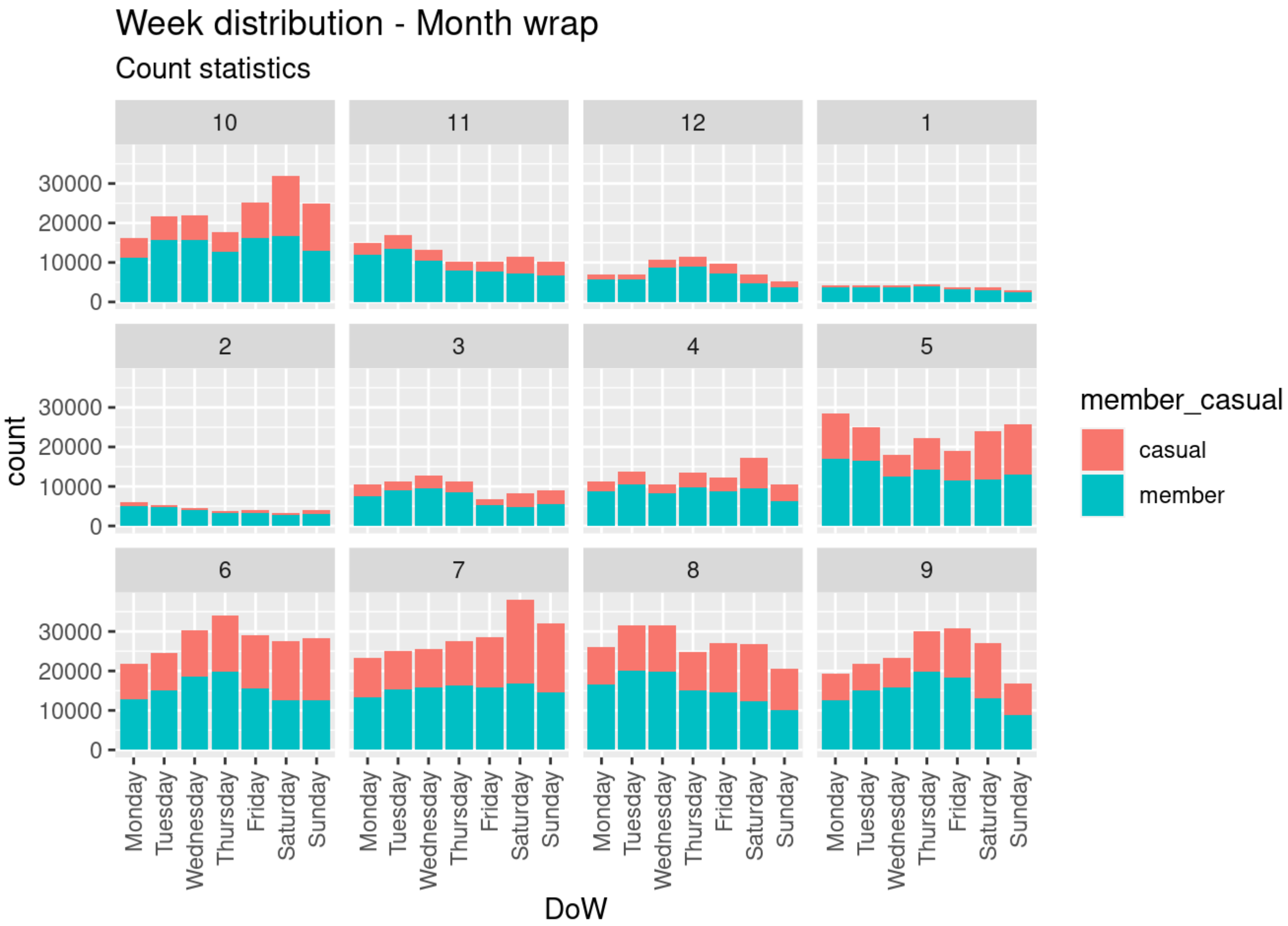
```
## Number of casual Users - 12 months 540511
```

```
## Number of member Users - 12 months 896597
```

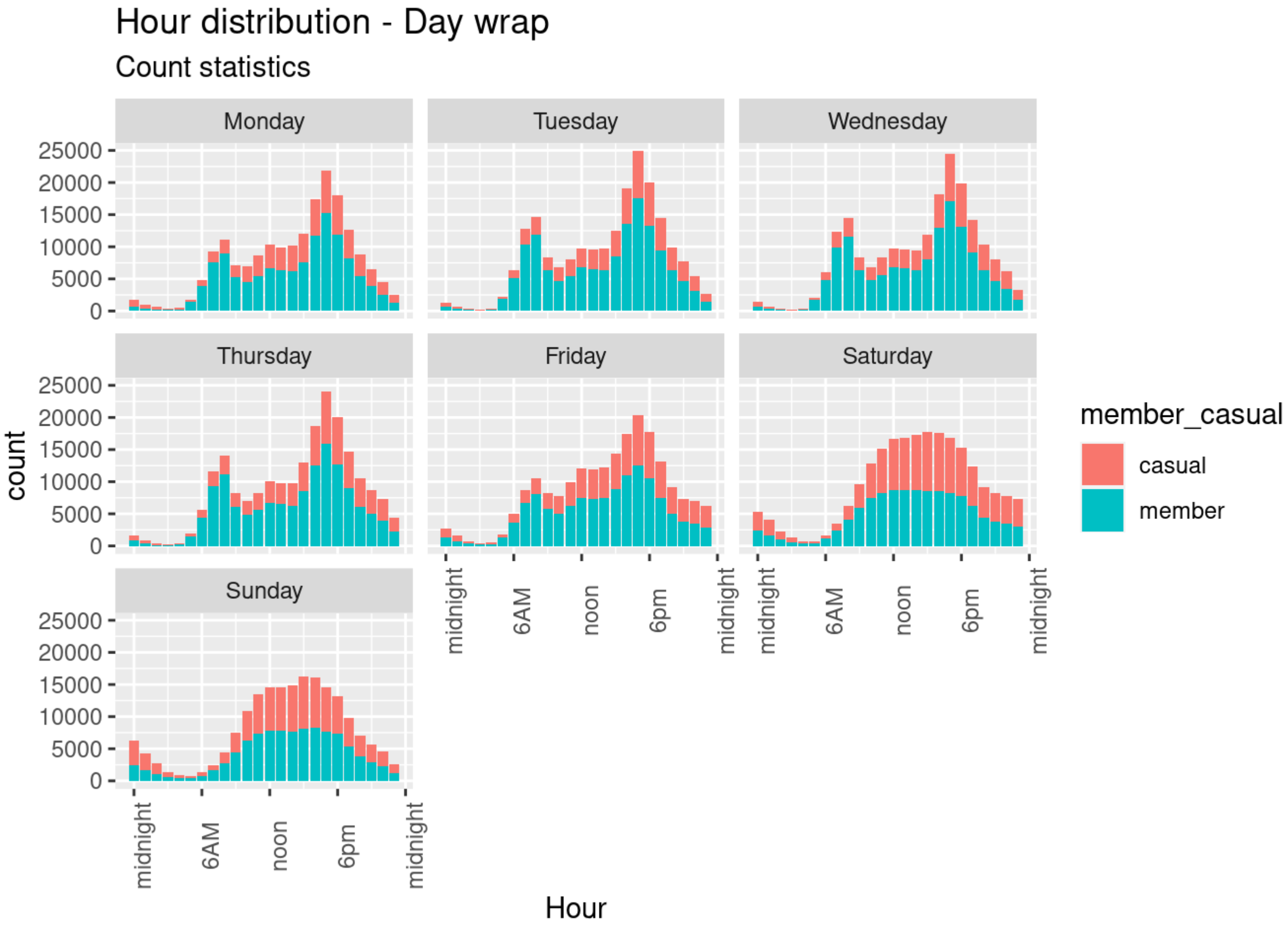
There is no possibility to investigate further the strict number of members instead of the number of ride because the anonymization provides different ride ID each time and there is no rider ID attribute. However, in order to analyse User habits this is not so fundamental.

Considering number of ride by month, following figure shows that during winter there are more difficulties in bike usage. Most probable reason has been found in Chicago temperature, precipitation and wind distribution during different months. Looking at graphs provided by Weather-US (https://www.weather-us.com/en/illinois-usa/chicago-climate) website, during winter months there are lower and prohibited temperature, higher wind velocities and higher precipitations probability. Bike Users habits found reflect these climate characteristics. However a lower bike usage during november, december and january is true for both casual and memeber Users.

Following graphs show instead week and hours distributions of rides. Is visible that during work days of autumn, winter and fall seasons, member Users doing the higher number of rides. During summer and weekend (friday, saturday and sunday - whole year), number of causual User rides increase. This because maybe there is the influence of turists and free trime to spend outside. However, there is always a good number of casual User rides that can be used as starting point to convert al least a percentages from casual to member Users.

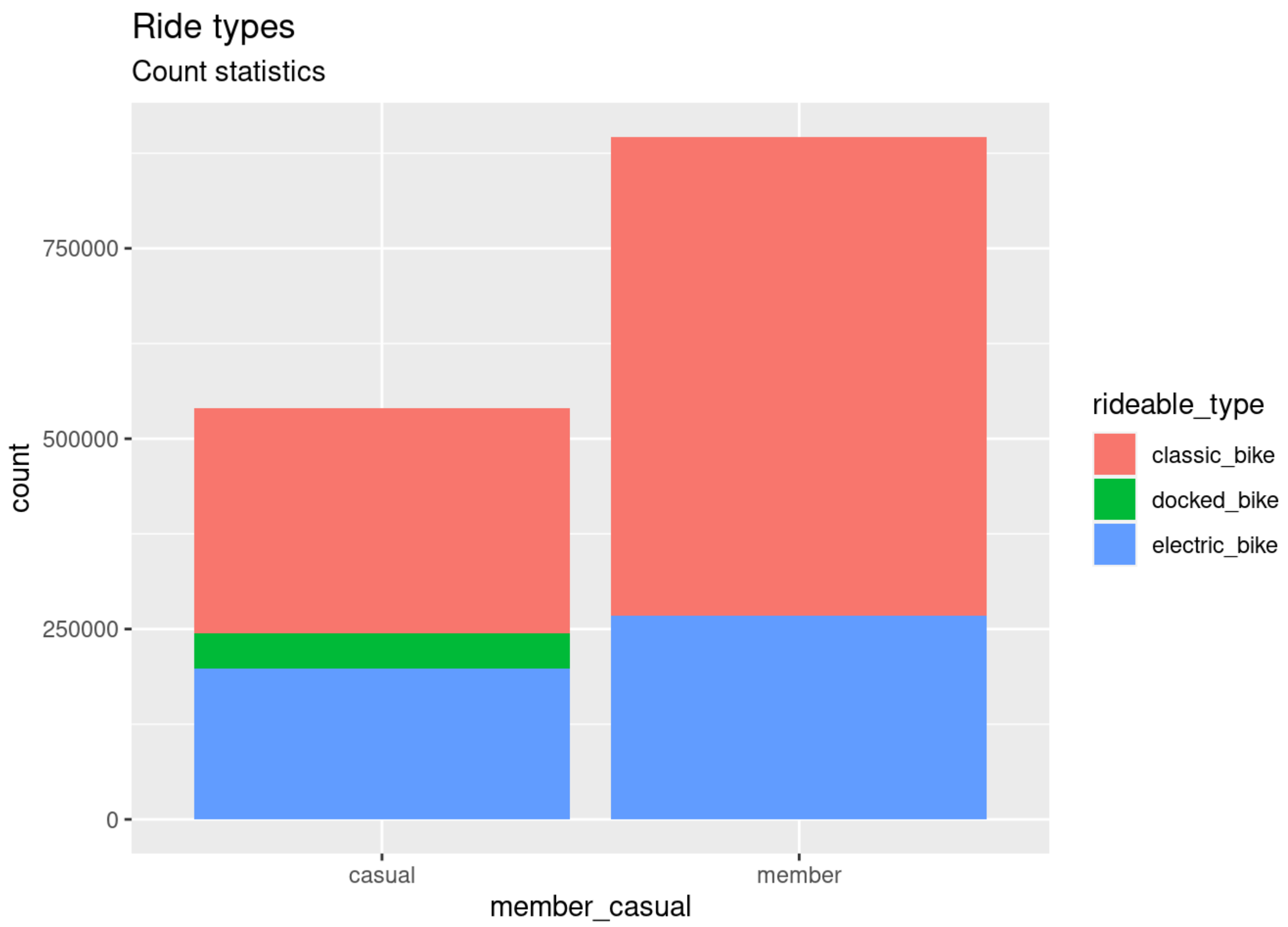


Considering instead hours of the day reaspect to days of the week, also in this case is visible that there is a trend in which causual User rides increase during mid-day in general and during weekend in particular. Also in this graph is visible that on saturday and sunday the number of casual User rides is bigger that member User one and is confirmed that there are however a base line of casual Users during all day.



Interesting behaviour differences between casual and memeber User consider bike types used. Following figure shows that casual riders chose electric bike more times (realtive percentages) respect to member riders, that insted use more frequently classic bike. To in depth understand this difference an in-depth investigation must be performed but data available during this analysis are not sufficient to find an answer.

```
ggplot(data = df_sub, aes(member_casual)) +
  geom_bar() +
  geom_bar(aes(fill = rideable_type)) +
  labs(title = "Ride types", subtitle = "Count statistics") #nolint
```



However, looking directly parking conditions on Divvy (https://divvybikes.com/how-it-works/parking) website, seems that electric bike, from 2022, can be locked in a bigger area respect to the classic one. Dataset confirms this hypothesis. Both casual and member Users favourite start/end station is *West Peterson avenue* that is located in one of the area reserved to ebike parking only.

```
## [1] "Start/end main station - ebike only"
```

```
##                  Var1 Freq
## 1 2112 W Peterson Ave  283
##                  Var1 Freq
## 1 2112 W Peterson Ave  315
```

Additional analysis have been done looking at statistics on ride lenght mean value, respect to bike types subgroups for both casual and memeber Users. Is visible that mean value of ride lenght for casual Users is higher respect to memebers one.

```
print("Ebike ride length vs Day Of the Week")
```

```
## [1] "Ebike ride length vs Day Of the Week"
```

```
aggregate(df_sub_e$ride_length,
  by = list(df_sub_e$member_casual),
  FUN = mean)
```

```
##   Group.1             x
## 1  casual 974.2495 secs
## 2  member 631.0728 secs
```

```
print("Classic bike ride length vs Day Of the Week")
```

```
## [1] "Classic bike ride length vs Day Of the Week"
```

```
aggregate(df_sub_c$ride_length,
  by = list(df_sub_c$member_casual),
  FUN = mean)
```

```
##   Group.1              x
## 1  casual 1265.8571 secs
## 2  member  721.4173 secs
```

This could be and interesting information if we considering rental price. With a price of 1$ to unlock and 0.16$ a minutes, a casual member pay more than 5$ per each single ride, while a member has a significative reduction of price. Without particular member discount, longer trip means higher price.

Act phase

Guiding questions:

- What is your final conclusion based on your analysis?
- How could your team and business apply your insights?
- What next steps would you or your stakeholders take based on your findings?
- Is there additional data you could use to expand on your findings?

Following main User information obtained from previous analysis could be also as strategies to convert casual User into member.

- presence of casual Users during all year, months and day time (this confirm the possibility to expand member area with new Users);
- presence of particular peacks during summer month or during weekend;
- ebike additional parking possibilities;
- mean ride lenght value;

Electric bike Users seem indeed to be the suitable target for this purpose. Dedicated member packages that comprehends for example 3 days offers or single month offers, merged with possibility to extend the area of bike usage could be main ideas of the acquisition strategy, in addition to price comparison campaign in order to have visual feedback on savings. Moreover, incentives on summer seasons could be useful in order to convert also Users that during other months use different vehicle.

Considering the best *strong* hypothesis that is converting all ebike casual Users to members, percentages of new acquisition (respect to actual number of affiliate Users) will be about the 22%.

This is obviously not a realistic target but only a minor conversion percentages could be however a good result for the company.

```
## Number of casual ebike rides 197671
```

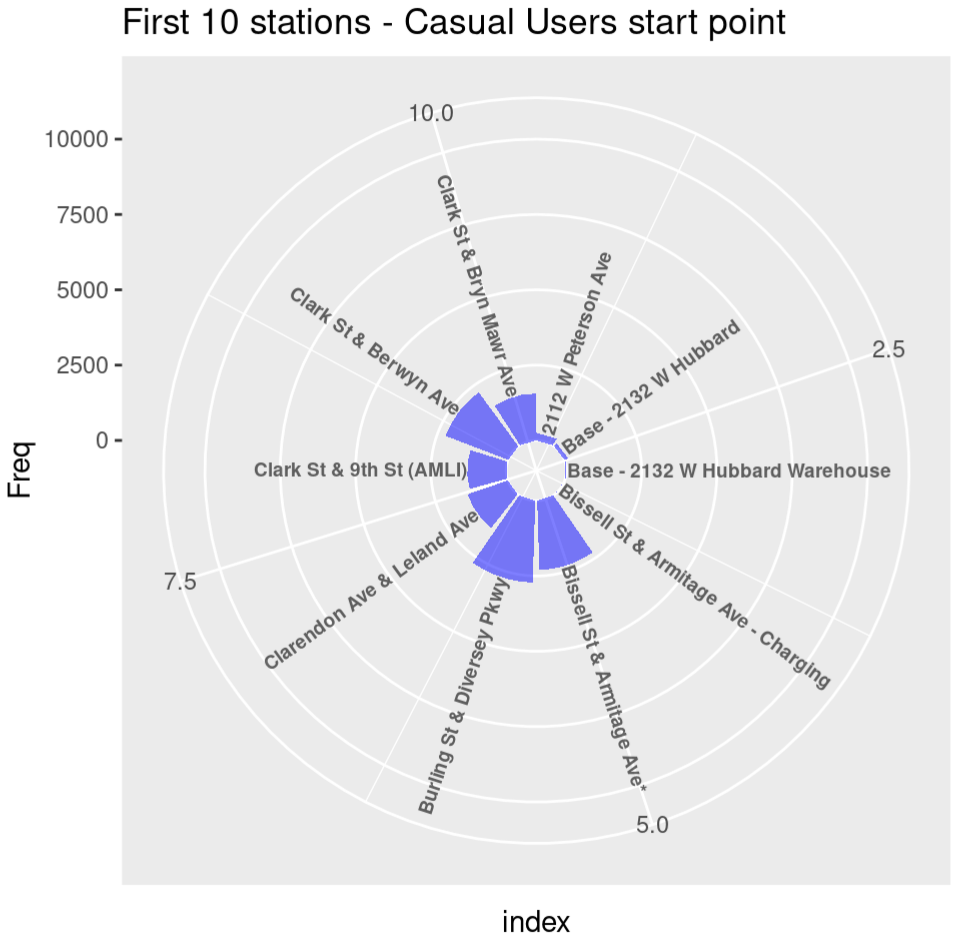
```
## Number of potential new members - 12 months 22.04681
```

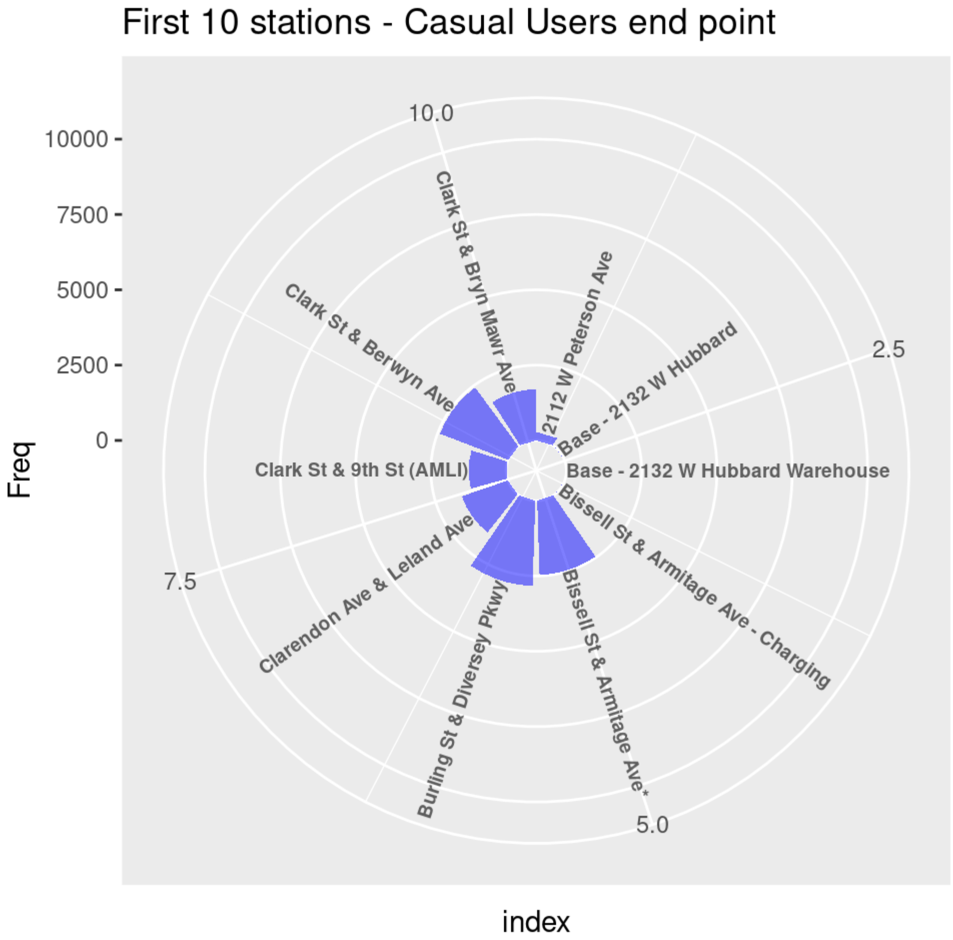
```
main_station_start <-
  as.data.frame(head(table(df_sub_casual$start_station_name), 10))
main_station_end <-
  as.data.frame(head(table(df_sub_casual$end_station_name), 10))
```

Additional suggestion in order to maximise the attention of casual User on new campaigns, is to consider the possibility to advertise Divvy affiliation in those stations in which there are the max number of bickers.

Following figures show the 10 main start/end stations for causual Users.

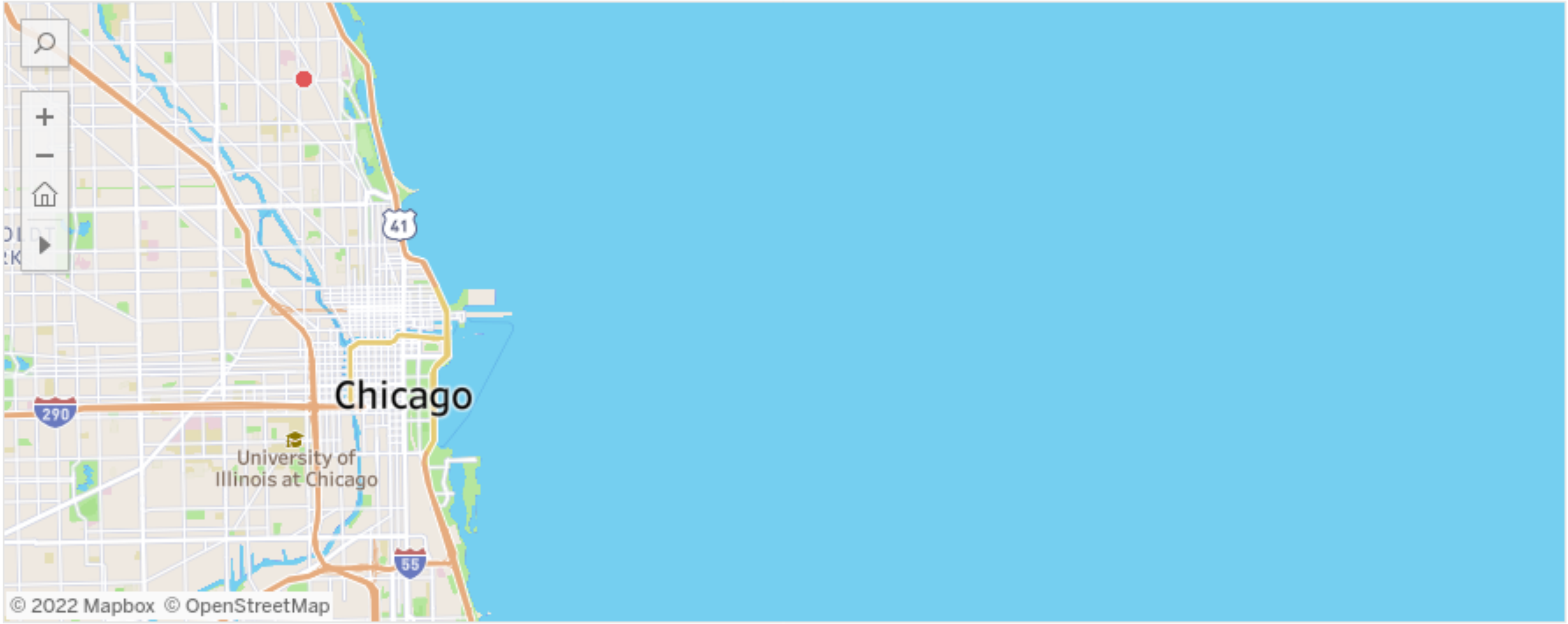
```
data_with_labels <- main_station_start
data_with_labels$index <- seq(1, 10)
number_of_label <- nrow(data_with_labels)
angle <-  90 - 360 * (data_with_labels$index - 0.5) / number_of_label
data_with_labels$hjust <- ifelse(angle < -90, 1, 0)
data_with_labels$angle <- ifelse(angle < -90, angle + 180, angle)

ggplot(data = data_with_labels, aes(x = index, y = Freq)) +
  geom_bar(stat = "identity", fill = alpha("blue", 0.5)) +
  ylim(-1000, 10000) +
  coord_polar(start = 0) +
  geom_text(data = data_with_labels, aes(x = index, y = Freq,
    label = Var1, hjust = hjust),
    color = "black", fontface = "bold",
    alpha = 0.6, size = 2.5,
    angle = data_with_labels$angle,
    inherit.aes = FALSE) +
  labs(title = "First 10 stations - Casual Users start point")
```


First 10 stations - Casual Users start point
Freq
10000
7500
5000
2500
0
10.0
2.5
5.0
7.5
index
Clark St & Bryn Mawr Ave
2112 W Peterson Ave
Base - 2132 W Hubbard
Base - 2132 W Hubbard Warehouse
Bissell St & Armitage Ave - Charging
Bissell St & Armitage Ave*
Burling St & Diversey Pkwy
Clarendon Ave & Leland Ave
Clark St & 9th St (AMLI)
Clark St & Berwyn Ave



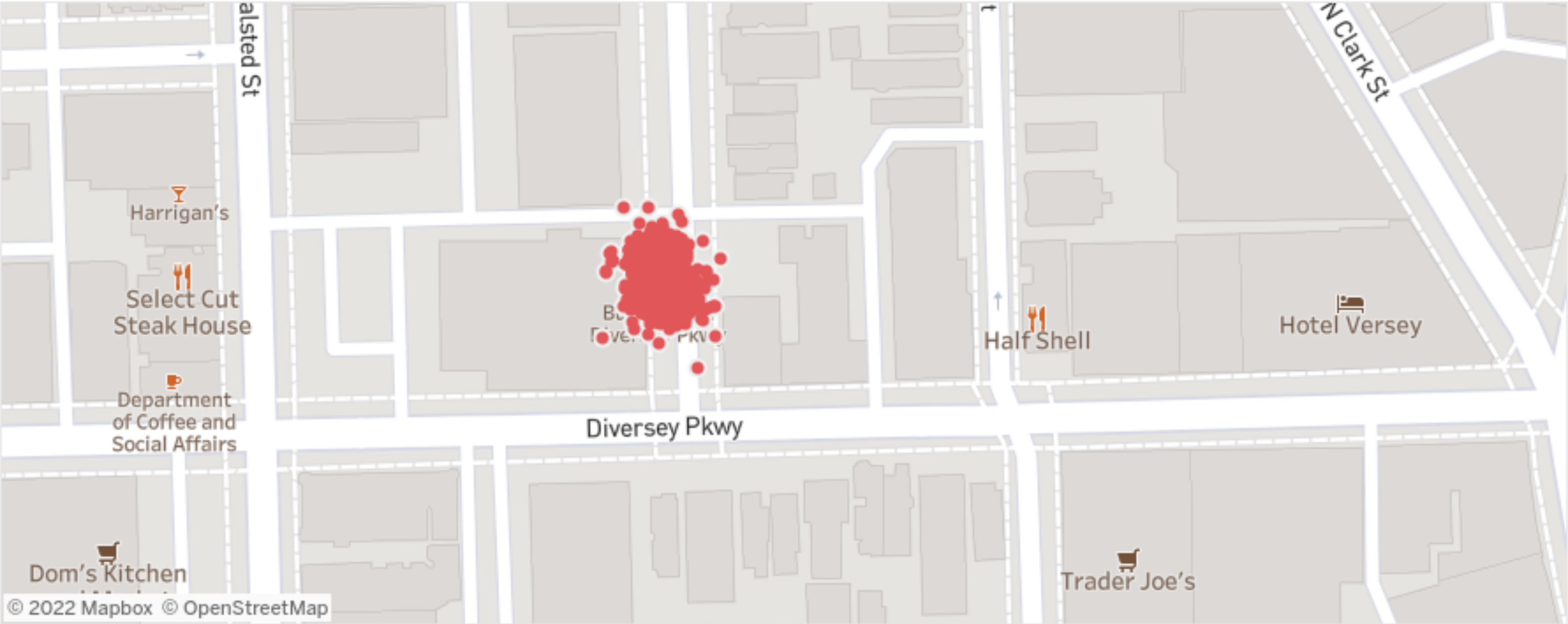
Following figures show maps obtained through **Tableau** that geographically identify main start/end station for casual Users.

Main start/end station casual Users


Chicago
University of Illinois at Chicago
41
290
55
© 2022 Mapbox © OpenStreetMap

Main station position - whole city

Main start/end station casual Users



Main station position - neighborhood zoom