Divy Case Study

Divvy 2021 Year Analysis

PHASE 1 : ASK

Key objectives:

1.Identify the business task:

- The company wants to improve their earnings reaching out to their “casual” riders, and for that they have to analyze in what aspects the “casual” and the annual customers differ, to be able to create a focused and successful marketing message to the “casual” customers that makes them change to the annual subscription.

2.Consider key stakeholders:

- The main stakeholders here are the director of marketing,marketing analytics team, and the Cyclistic executive team.

3.The business task:

Given these facts, the business task is defined as searching for differences in the two identified kinds of users in order to make a focused marketing campaign to the “casual” users in order for them to change to the annual subscription, or resumed in a question:

What could motivate the “casual” users to change to an annual subscription based on their behavior?

PHASE 2 : Prepare

Key objectives:

1.Determine the credibility of the data:

- The data is public data from a bike sharing company. It starts from the year 2013 until 2022, there isn’t much of a naming convention as the files are sometimes organized by quarter, or month, or the whole year and their names vary a lot. The naming of the columns also changes and there are some columns added and deleted over the years. Nevertheless the data seems to be in good condition and its first hand data collected by the company itself with lots of entries and with lots of useful data.

2.Sort and filter the data:

- For this analysis I’m going to focus on the 2021-2022 period as it’s the more relevant period to the business task and it has the more complete data with geo-location coordinates, and types of bike used.

```r
#First I add all the libraries necessary to my analysis


library("tidyverse")
library("lubridate")
library("ggplot2")
library("geosphere")
library("gridExtra")
library("ggmap")
library("readr")
```

PHASE 3 : Process

Key objectives:

1.Clean the data, and prepare the data for analysis:

- Now that we have all the data in one place we can start to clean the data of possible errors like NA. Also we will make some changes to the data adding useful new columns based on calculations of already existing columns in order to facilitate our analysis and arrive at more insightful conclusions.

```
## [1] " ##### Glimpse #####"
```

```
## Rows: 5,595,063
## Columns: 13
## $ ride_id            <chr> "E19E6F1B8D4C42ED", "DC88F20C2C55F27F", "EC45C94683~
## $ rideable_type      <chr> "electric_bike", "electric_bike", "electric_bike", ~
## $ started_at         <dttm> 2021-01-23 16:14:19, 2021-01-27 18:43:08, 2021-01-~
## $ ended_at           <dttm> 2021-01-23 16:24:44, 2021-01-27 18:47:12, 2021-01-~
## $ start_station_name <chr> "California Ave & Cortez St", "California Ave & Cor~
## $ start_station_id   <chr> "17660", "17660", "17660", "17660", "17660", "17660~
## $ end_station_name   <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, "Wood St & Augu~
## $ end_station_id     <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, "657", "13258",~
## $ start_lat          <dbl> 41.90034, 41.90033, 41.90031, 41.90040, 41.90033, 4~
## $ start_lng          <dbl> -87.69674, -87.69671, -87.69664, -87.69666, -87.696~
## $ end_lat            <dbl> 41.89000, 41.90000, 41.90000, 41.92000, 41.90000, 4~
## $ end_lng            <dbl> -87.72000, -87.69000, -87.70000, -87.69000, -87.700~
## $ member_casual      <chr> "member", "member", "member", "member", "casual", "~
```

```
## [1] "####### SUMMARY #######"
```

```
##    ride_id          rideable_type        started_at
##  Length:5595063     Length:5595063     Min.   :2021-01-01 00:02:05
##  Class :character   Class :character   1st Qu.:2021-06-06 23:52:40
##  Mode  :character   Mode  :character   Median :2021-08-01 01:52:11
##                                        Mean   :2021-07-29 07:41:02
##                                        3rd Qu.:2021-09-24 16:36:16
##                                        Max.   :2021-12-31 23:59:48
##
##     ended_at                   start_station_name start_station_id
##  Min.   :2021-01-01 00:08:39   Length:5595063     Length:5595063
##  1st Qu.:2021-06-07 00:44:21   Class :character   Class :character
##  Median :2021-08-01 02:21:55   Mode  :character   Mode  :character
##  Mean   :2021-07-29 08:02:58
##  3rd Qu.:2021-09-24 16:54:05
##  Max.   :2022-01-03 17:32:18
##
##  end_station_name   end_station_id       start_lat       start_lng
##  Length:5595063     Length:5595063     Min.   :41.64   Min.   :-87.84
##  Class :character   Class :character   1st Qu.:41.88   1st Qu.:-87.66
##  Mode  :character   Mode  :character   Median :41.90   Median :-87.64
##                                        Mean   :41.90   Mean   :-87.65
##                                        3rd Qu.:41.93   3rd Qu.:-87.63
##                                        Max.   :42.07   Max.   :-87.52
##
##     end_lat         end_lng       member_casual
##  Min.   :41.39   Min.   :-88.97   Length:5595063
##  1st Qu.:41.88   1st Qu.:-87.66   Class :character
##  Median :41.90   Median :-87.64   Mode  :character
##  Mean   :41.90   Mean   :-87.65
##  3rd Qu.:41.93   3rd Qu.:-87.63
##  Max.   :42.17   Max.   :-87.49
##  NA’s   :4771    NA’s   :4771
```

```
#Cleaning the Data:
#Now lets clean the data to be able to properly work with it:
#Fist we make a copy of data :
df_copy <- data.frame(df)
#Check if the copy has been created in the different memory:
tracemem(df_copy)==tracemem(df)
```

```
## [1] FALSE
```

```
#Now we drop all NA:
df_copy <- drop_na(df_copy)
```

```
## tracemem[0x0000000026004fc0 -> 0x00000000238f3b50]: vec_detect_complete drop_na.data.frame drop_na e
```

```
#Preparing the Data:
#Then lets create some new columns:
#First lets separate the dates into month, day, year and day of the week:
df_copy$started_at <- as.POSIXct(df_copy$started_at, format="%m/%d/%Y %H:%M")
df_copy$ended_at <- as.POSIXct(df_copy$ended_at, format="%m/%d/%Y %H:%M")
df_copy$date <- as.Date(df_copy$started_at)
df_copy$month <- format(as.Date(df_copy$date),"%m")
df_copy$day <- format(as.Date(df_copy$date),"%d")
df_copy$year <- format(as.Date(df_copy$date),"%Y")
df_copy$day_of_week <- format(as.Date(df_copy$date),"%A")


#Then lets make some useful new columns with the duration of the ride, distance traveled, and speed:
```

```
#First the ride length in seconds:

df_copy$ride_length <-difftime(df_copy$ended_at,df_copy$started_at)

#Then the ride distance traveled in km
df_copy$ride_distance <- distGeo(matrix(c(df_copy$start_lng, df_copy$start_lat),ncol=2),matrix( c(df_cop

# Takes lot of computing power, check why?
#df_copy$ride_distance <- distVincentyEllipsoid(matrix(c(df_copy$start_lng, df_copy$start_lat),ncol=2),

df_copy$ride_distance <- df_copy$ride_distance/1000

#At last the speed in Km/h
df_copy$ride_speed <- c(df_copy$ride_distance)/as.numeric(c(df_copy$ride_length),units="hours")

# The dataframe includes a few hundred entries when bikes were taken out of docks and checked for quali
df_copy <- df_copy[!(df_copy$start_station_name == "HQ QR" | df_copy$ride_length<0),]
```

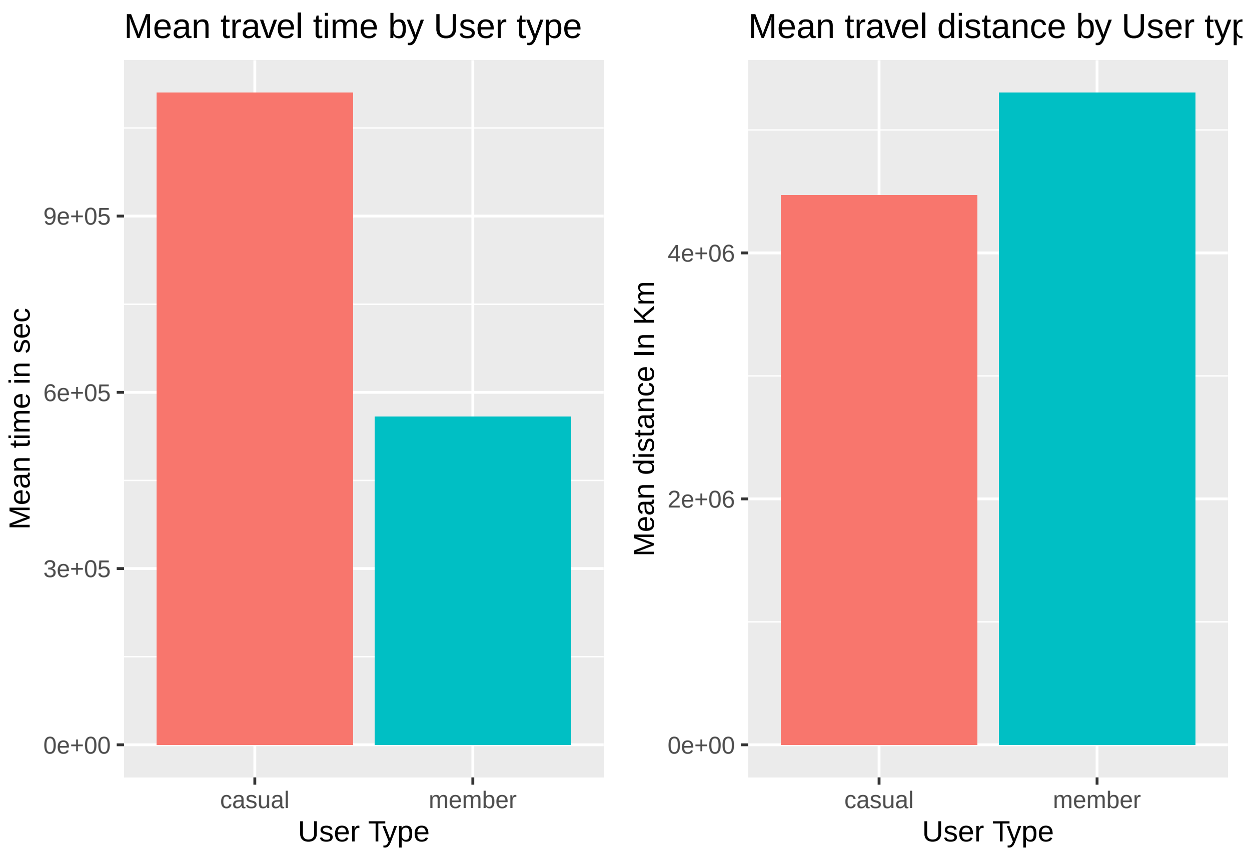
PHASE 4 : Analyze

Key objectives:

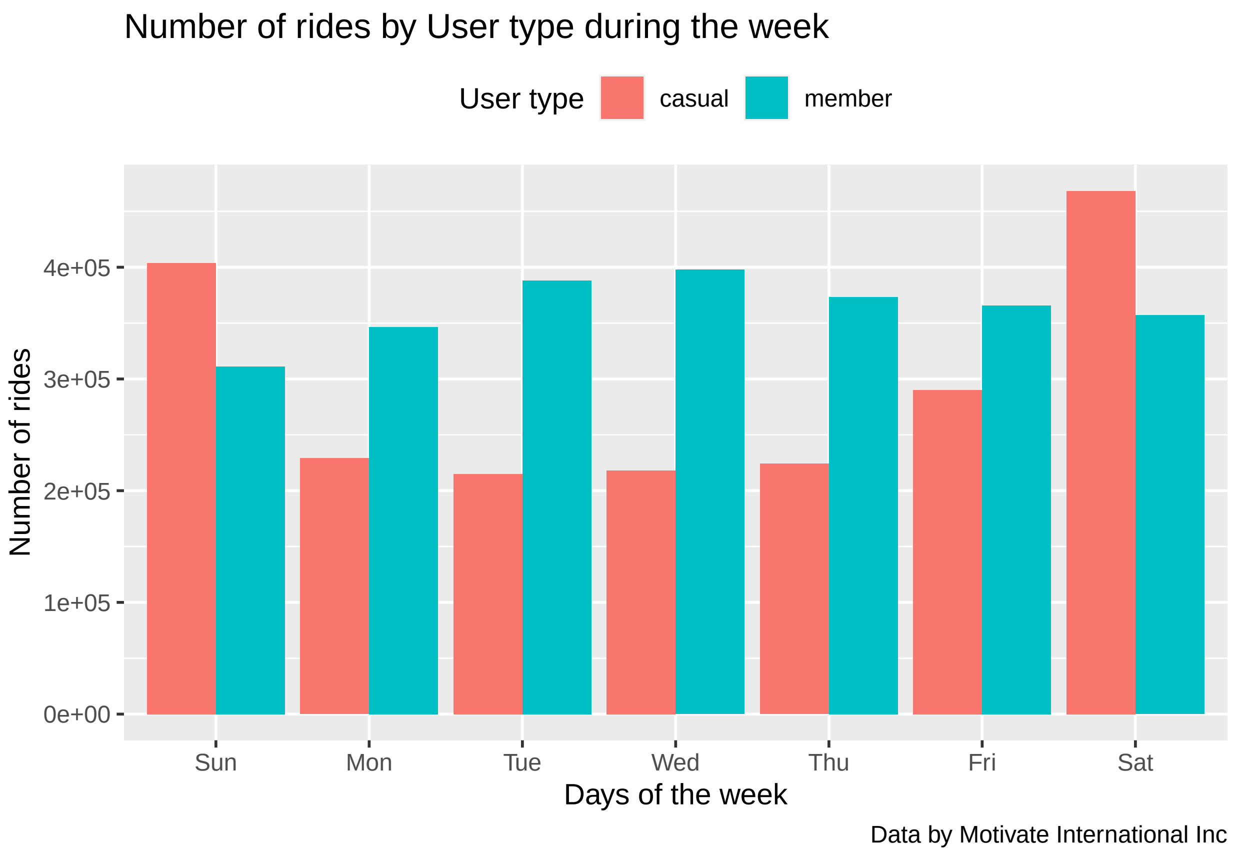
1.Identify trends and relationships.:

- We have now a complete data frame with all the info we need to identify the differences in behaviour between the casual and the member users.

```
#Fist we calculate the average distance, distance for both the casual and member type users:

userTypeMean <- df_copy %>% group_by(member_casual) %>% summarise(mean_time =mean(ride_length), mean_di

userTypeMean$mean_time <- as.numeric(userTypeMean$mean_time, unit="hours")
membervstime <-  ggplot(userTypeMean) + geom_col(mapping=aes(x=member_casual,y=mean_time,fill=member_ca
  labs(title = "Mean travel time by User type",x="User Type",y="Mean time in sec")

membervsdistance <- ggplot(userTypeMean) +
  geom_col(mapping=aes(x=member_casual,y=mean_distance,fill=member_casual), show.legend = FALSE)+
  labs(title = "Mean travel distance by User type",x="User Type",y="Mean distance In Km")

grid.arrange(membervstime, membervsdistance, ncol = 2)
```



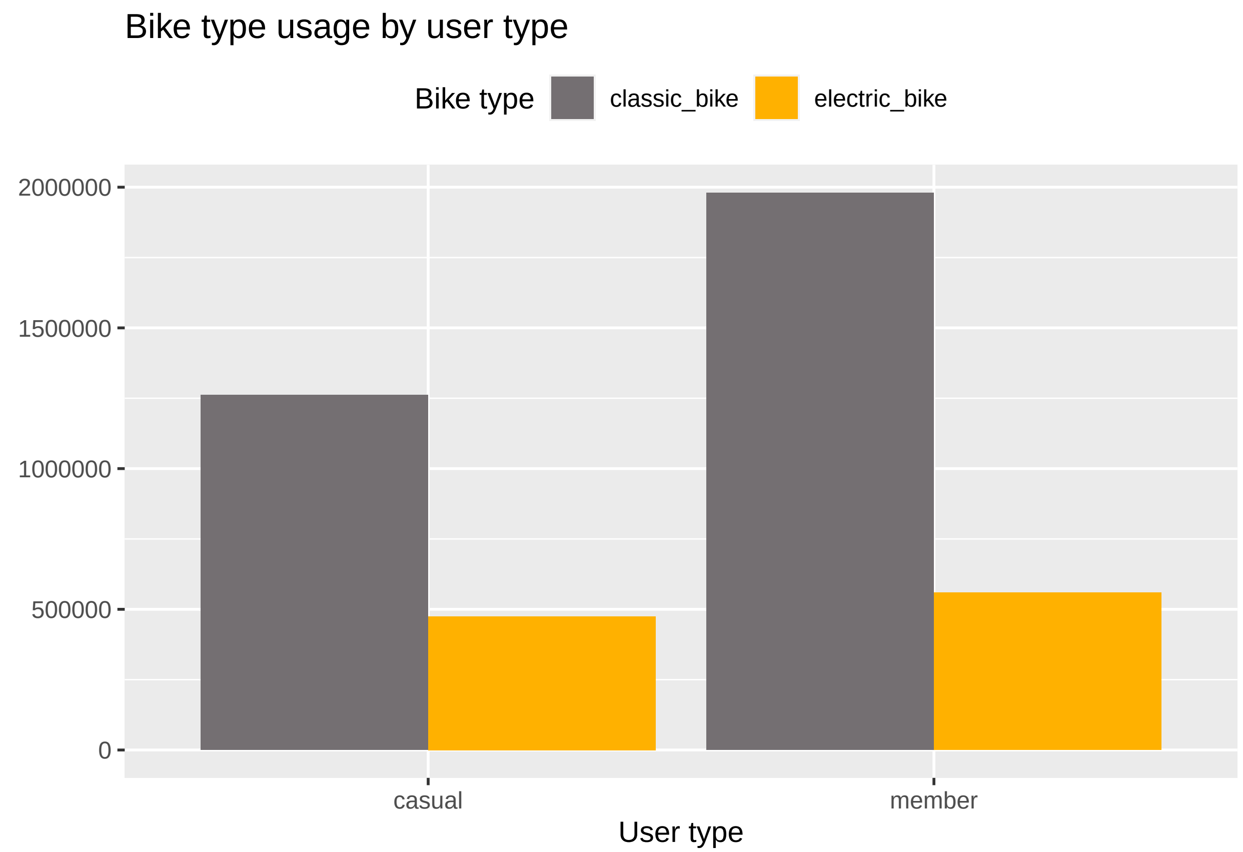
```
#Then we check the number of rides differences by weekday:

df_copy   %>% mutate(weekday = wday(started_at, label = TRUE)) %>%
  group_by(member_casual, weekday) %>%
  summarise(number_of_rides = n()
           ,average_duration = mean(ride_length),.groups = 'drop') %>%
  arrange(member_casual, weekday)  %>%
  ggplot(aes(x = weekday, y = number_of_rides, fill = member_casual)) +
  geom_col(position = "dodge") +
  labs(title = "Number of rides by User type during the week",x="Days of the week",y="Number of rides",
  theme(legend.position="top")
```



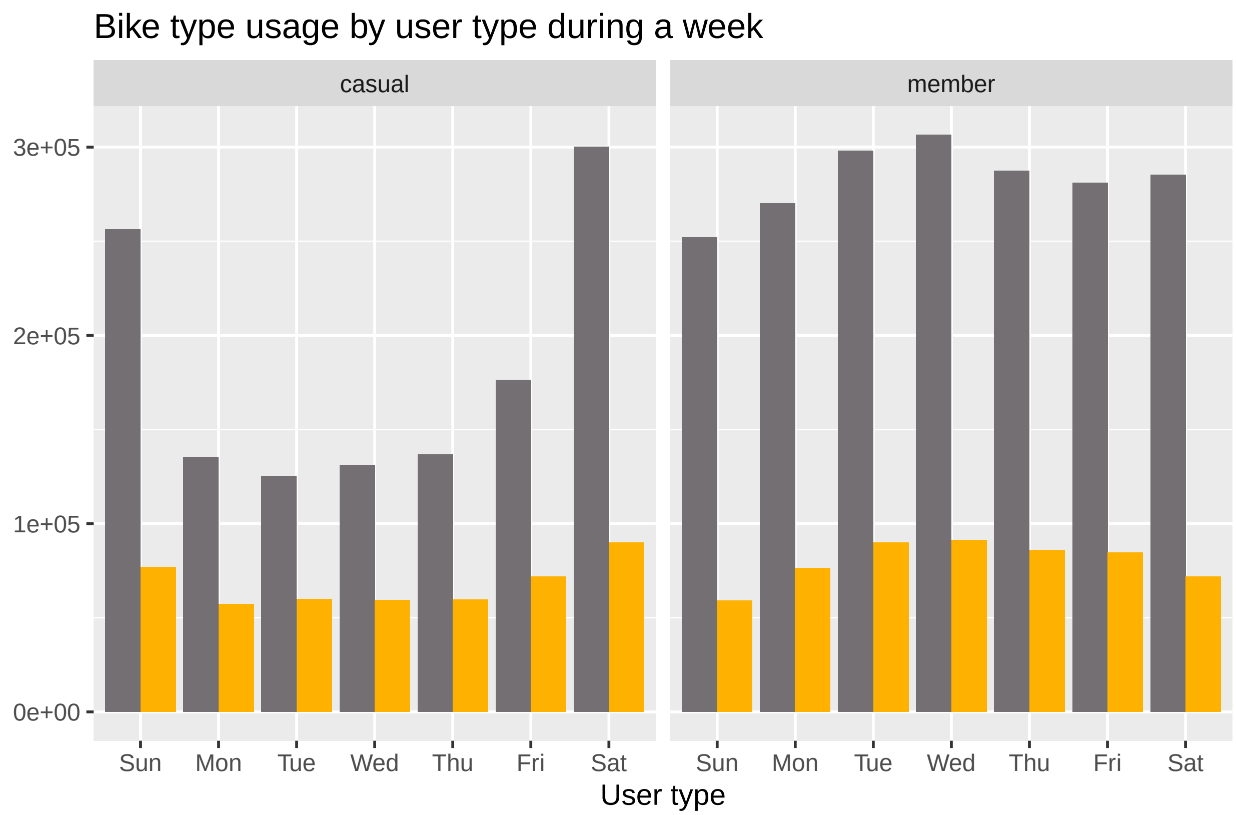
Analysis:

- It seems that the casual users travel the same average distance than the member users, but they have much longer rides, that would indicate a more leisure oriented usage vs a more “public transport” or pragmatic use of the bikes by the annual members.
- This idea is reinforced by the fact that annual users have a very stable use of the service during the week, but the casual users are more of a weekend user.

```
#Create a new data frame with only the rows with info in the "bike type" column:

with_bike_type <- df_copy %>% filter(rideable_type=="classic_bike" |rideable_type=="electric_bike")

#Then lets check the bike type usage by user type:

with_bike_type %>%
  group_by(member_casual,rideable_type) %>%
  summarise(totals=n(), .groups="drop")  %>%
  ggplot()+
  geom_col(aes(x=member_casual,y=totals,fill=rideable_type), position = "dodge") +
  labs(title = "Bike type usage by user type",x="User type",y=NULL, fill="Bike type") +
  scale_fill_manual(values = c("classic_bike" = "#746F72","electric_bike" = "#FFB100")) +
  theme(legend.position="top")
```



```
#And their usage by both user types during a week:

with_bike_type %>%
  mutate(weekday = wday(started_at, label = TRUE)) %>%
  group_by(member_casual,rideable_type,weekday) %>%
  summarise(totals=n(), .groups="drop") %>%

  ggplot(aes(x=weekday,y=totals, fill=rideable_type)) +
  geom_col( position = "dodge") +
  facet_wrap(~member_casual) +
  labs(title = "Bike type usage by user type during a week",x="User type",y=NULL,caption = "Data by Mot
  scale_fill_manual(values = c("classic_bike" = "#746F72","electric_bike" = "#FFB100")) +
  theme(legend.position="none")
```



Data by Motivate International Inc

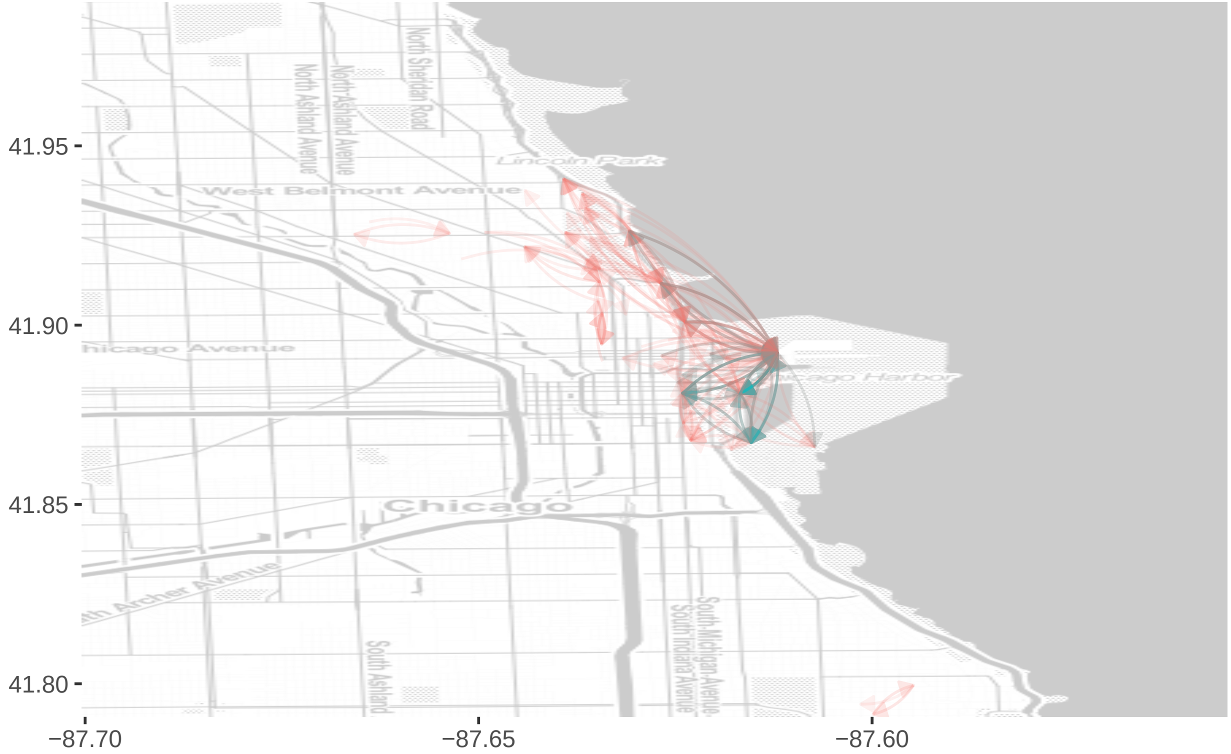
Analysis:

- Here we can see that the annual members and the casual users show a clear preference for the classic bikes, which makes sense given the long duration of their rides.
- On a weekly basis we can see that for the annual members and the casual users we see in general the same pattern of usage from the previous weekly charts, preferring the classic bikes
- Casual users show more weekend usage of the service and prefer classic bikes

```
#Now let's the coordinates data of the rides, to see if is there any interesting pattern:

#First we create a table only for the most popular routes (>250 times)

coordinates_table <- df_copy %>%
  filter(start_lng != end_lng & start_lat != end_lat) %>%
  group_by(start_lng, start_lat, end_lng, end_lat, member_casual, rideable_type) %>%
  summarise(total = n(),.groups="drop") %>%
  filter(total > 250)

#Then we create two sub tables for each user type
casual <- coordinates_table %>% filter(member_casual == "casual")
member <- coordinates_table %>% filter(member_casual == "member")

#Lets store bounding box coordinates for ggmap:
chi_bb <- c(
  left = -87.700424,
  bottom = 41.790769,
  right = -87.554855,
  top = 41.990119
)

#Here we store the stamen map of Chicago
chicago_stamen <- get_stamenmap(
  bbox = chi_bb,
  zoom = 12,
  maptype = "toner"
)
```

```
#Then we plot the data on the map
ggmap(chicago_stamen,darken = c(0.8, "white")) +
  geom_curve(casual, mapping = aes(x = start_lng, y = start_lat, xend = end_lng, yend = end_lat, alpha=
  coord_cartesian() +
  labs(title = "Most popular routes by casual users",x=NULL,y=NULL, color="User type", caption = "Data
  theme(legend.position="none")
```

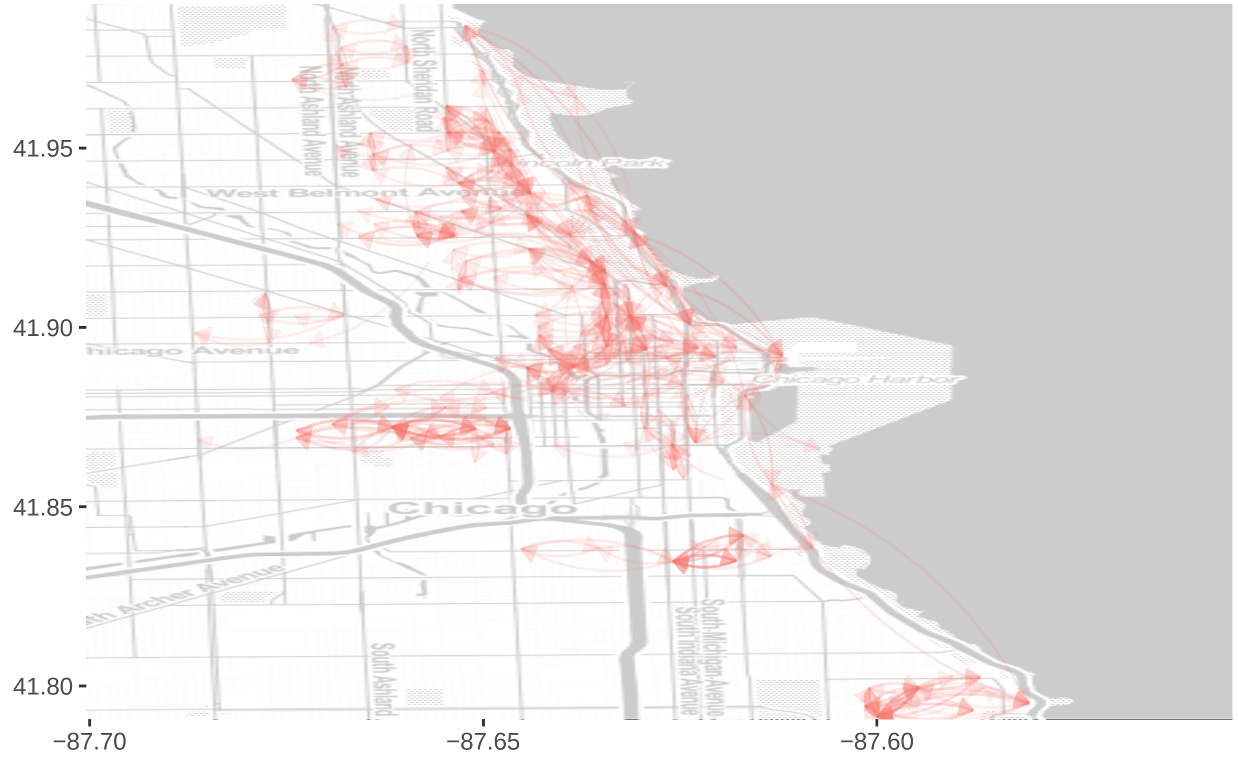
Most popular routes by casual users



Data by Motivate International Inc

```
ggmap(chicago_stamen,darken = c(0.8, "white")) +
  geom_curve(member, mapping = aes(x = start_lng, y = start_lat, xend = end_lng, yend = end_lat, alpha=
  coord_cartesian() +
  labs(title = "Most popular routes by annual members",x=NULL,y=NULL, caption = "Data by Motivate Inter
  theme(legend.position="none")
```

Most popular routes by annual members



Data by Motivate International Inc

Analysis:

- The coordinates data resulted to be very interesting, as we can clearly see the casual is usually located around the center of the town, with all their trips located around that area which makes sense given that they have a more relaxed leisure rides, on weekends probably also tourist or sightseeing related rides, that naturally focus more on the downtown area where most of the interest points are.
- This contrasts heavily with the longer range of the annual users that connect the downtown with the outskirts of the city, that would suggest they are mostly people that live outside the downtown and use the service to commute everyday to their works in the city.

PHASE 5 : Share

Key objectives:

1.Share my conclusions.:

- Taking in consideration both the business task: **¿What could motivate the "casual" users to change to an annual subscription based on their behavior?** and the insights we've learned from the available data we can make some conclusions.

  1)**The Casual users** have **leisure & health**, and **tourism** rides mostly on **weekends** and using **Classic bikes**.

  2)**The Annual users** have **commute** or **pragmatic** rides, during **all week** preferably using both **classic bikes**

- I would share this info, the data and my analysis to the marketing team, and I would suggest that in order to **convert the casual to the annual** users it would be interesting to focus the messages on the **leisure & health** aspect of the service, and maybe offer some kind of **promotion related to weekends and/or classic bikes**.