Case Study:Cyclistic : Bike-Share Co.

Objective

*The bike sharing company wants to analyze their user data to find the main differences in behaviour between their two types of users, the “casual” who pays for each ride and the annual member who pays a yearly subscription to the service*.

PHASE 1 : ASK

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 and my manager Lily Moreno, the rest of the 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

1. Determine the credibility of the data:

The data is public data from a bike sharing company. It starts from the year 2013 to Jul 2023, 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.

[Visit/Download Dataset here](https://divvy-tripdata.s3.amazonaws.com/index.html)

2. Sort and filter the data :

For this analysis I’m going to focus on the last one year i.e Jul 2022 to Jul 2023 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.

library("tidyverse")

## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──  
## ✔ dplyr 1.1.2 ✔ readr 2.1.4  
## ✔ forcats 1.0.0 ✔ stringr 1.5.0  
## ✔ ggplot2 3.4.3 ✔ tibble 3.2.1  
## ✔ lubridate 1.9.2 ✔ tidyr 1.3.0  
## ✔ purrr 1.0.2   
## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()  
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors

library("ggplot2")  
library("lubridate")  
library("geosphere")

## The legacy packages maptools, rgdal, and rgeos, underpinning the sp package,  
## which was just loaded, will retire in October 2023.  
## Please refer to R-spatial evolution reports for details, especially  
## https://r-spatial.org/r/2023/05/15/evolution4.html.  
## It may be desirable to make the sf package available;  
## package maintainers should consider adding sf to Suggests:.  
## The sp package is now running under evolution status 2  
## (status 2 uses the sf package in place of rgdal)

library("gridExtra")

##   
## Attaching package: 'gridExtra'  
##   
## The following object is masked from 'package:dplyr':  
##   
## combine

library("ggmap")

## ℹ Google's Terms of Service: <https://mapsplatform.google.com>  
## ℹ Please cite ggmap if you use it! Use `citation("ggmap")` for details.

X2022\_07 <- readxl::read\_excel("2022\_07.xlsx")  
X2022\_08 <- readxl::read\_excel("2022\_08.xlsx")  
X2022\_09 <- readxl::read\_excel("2022\_09.xlsx")  
X2022\_10 <- readxl::read\_excel("2022\_10.xlsx")  
X2022\_11 <- readxl::read\_excel("2022\_11.xlsx")  
X2022\_12 <- readxl::read\_excel("2022\_12.xlsx")  
X2023\_01 <- readxl::read\_excel("2023\_01.xlsx")  
X2023\_02 <- readxl::read\_excel("2023\_02.xlsx")  
X2023\_03 <- readxl::read\_excel("2023\_03.xlsx")  
X2023\_04 <- readxl::read\_excel("2023\_04.xlsx")  
X2023\_05 <- readxl::read\_excel("2023\_05.xlsx")  
X2023\_06 <- readxl::read\_excel("2023\_06.xlsx")

combined\_data <-list(X2022\_07,X2022\_08,X2022\_09,X2022\_10,X2022\_11,X2022\_12,  
 X2023\_01,X2023\_02,X2023\_04,X2023\_05,X2023\_06)  
  
combined\_data <- lapply(combined\_data, function(df) {  
 df$end\_station\_id <- as.character(df$end\_station\_id)  
 return(df)   
})  
  
trip\_data <- dplyr::bind\_rows(combined\_data)  
  
#glimpse Last year data combined   
  
dplyr::glimpse(trip\_data)

## Rows: 5,520,766  
## Columns: 13  
## $ ride\_id <chr> "954144C2F67B1932", "292E027607D218B6", "5776585258…  
## $ rideable\_type <chr> "classic\_bike", "classic\_bike", "classic\_bike", "cl…  
## $ started\_at <dttm> 2022-07-05 08:12:47, 2022-07-26 12:53:38, 2022-07-…  
## $ ended\_at <dttm> 2022-07-05 08:24:32, 2022-07-26 12:55:31, 2022-07-…  
## $ start\_station\_name <chr> "Ashland Ave & Blackhawk St", "Buckingham Fountain …  
## $ start\_station\_id <chr> "13224", "15541", "15541", "15541", "TA1307000117",…  
## $ end\_station\_name <chr> "Kingsbury St & Kinzie St", "Michigan Ave & 8th St"…  
## $ end\_station\_id <chr> "KA1503000043", "623", "623", "TA1307000164", "TA13…  
## $ start\_lat <dbl> 41.90707, 41.86962, 41.86962, 41.86962, 41.89147, 4…  
## $ start\_lng <dbl> -87.66725, -87.62398, -87.62398, -87.62398, -87.626…  
## $ end\_lat <dbl> 41.88918, 41.87277, 41.87277, 41.79526, 41.93625, 4…  
## $ end\_lng <dbl> -87.63851, -87.62398, -87.62398, -87.59647, -87.652…  
## $ member\_casual <chr> "member", "casual", "casual", "casual", "member", "…

PHASE 3 : Process

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.

#drop all the NA null values   
  
tripdata <- tidyr::drop\_na(trip\_data)  
  
glimpse(tripdata)

## Rows: 3,916,736  
## Columns: 13  
## $ ride\_id <chr> "954144C2F67B1932", "292E027607D218B6", "5776585258…  
## $ rideable\_type <chr> "classic\_bike", "classic\_bike", "classic\_bike", "cl…  
## $ started\_at <dttm> 2022-07-05 08:12:47, 2022-07-26 12:53:38, 2022-07-…  
## $ ended\_at <dttm> 2022-07-05 08:24:32, 2022-07-26 12:55:31, 2022-07-…  
## $ start\_station\_name <chr> "Ashland Ave & Blackhawk St", "Buckingham Fountain …  
## $ start\_station\_id <chr> "13224", "15541", "15541", "15541", "TA1307000117",…  
## $ end\_station\_name <chr> "Kingsbury St & Kinzie St", "Michigan Ave & 8th St"…  
## $ end\_station\_id <chr> "KA1503000043", "623", "623", "TA1307000164", "TA13…  
## $ start\_lat <dbl> 41.90707, 41.86962, 41.86962, 41.86962, 41.89147, 4…  
## $ start\_lng <dbl> -87.66725, -87.62398, -87.62398, -87.62398, -87.626…  
## $ end\_lat <dbl> 41.88918, 41.87277, 41.87277, 41.79526, 41.93625, 4…  
## $ end\_lng <dbl> -87.63851, -87.62398, -87.62398, -87.59647, -87.652…  
## $ member\_casual <chr> "member", "casual", "casual", "casual", "member", "…

#Then lets create some new columns.  
#First lets separate the dates into month, day, year and day of the week:  
  
  
tripdata$date <- as.Date(tripdata$started\_at)  
tripdata$month <- format(as.Date(tripdata$date),"%m")  
tripdata$day <- format(as.Date(tripdata$date),"%d")  
tripdata$year <-format(as.Date(tripdata$date),"%y")  
tripdata$week <-format(as.Date(tripdata$date),"%A")  
  
#Then lets make some useful new columns with the duration of the ride, distance traveled, and speed:  
  
tripdata$ride\_duration <- difftime(tripdata$ended\_at,tripdata$started\_at)  
  
#Then the ride distance traveled in km  
  
tripdata$ride\_distance <- distGeo(matrix(c(tripdata$start\_lng,tripdata$start\_lat),ncol = 2),   
 matrix(c(tripdata$end\_lng,tripdata$end\_lat), ncol = 2))  
  
tripdata$ride\_distance <- tripdata$ride\_distance / 1000  
  
#Speed of Ride :   
  
tripdata$ride\_speed = c(tripdata$ride\_distance)/as.numeric(c(tripdata$ride\_duration), units="hours")  
  
# The dataframe includes a few hundred entries when bikes were taken out of docks  
#checked for quality by team or ride\_length was negative  
  
#filtering out the data   
  
tripdata <- tripdata[!(tripdata$start\_station\_name == "HQ QR" | tripdata$ride\_duration<0),]

PHASE 4 : Analyze

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.

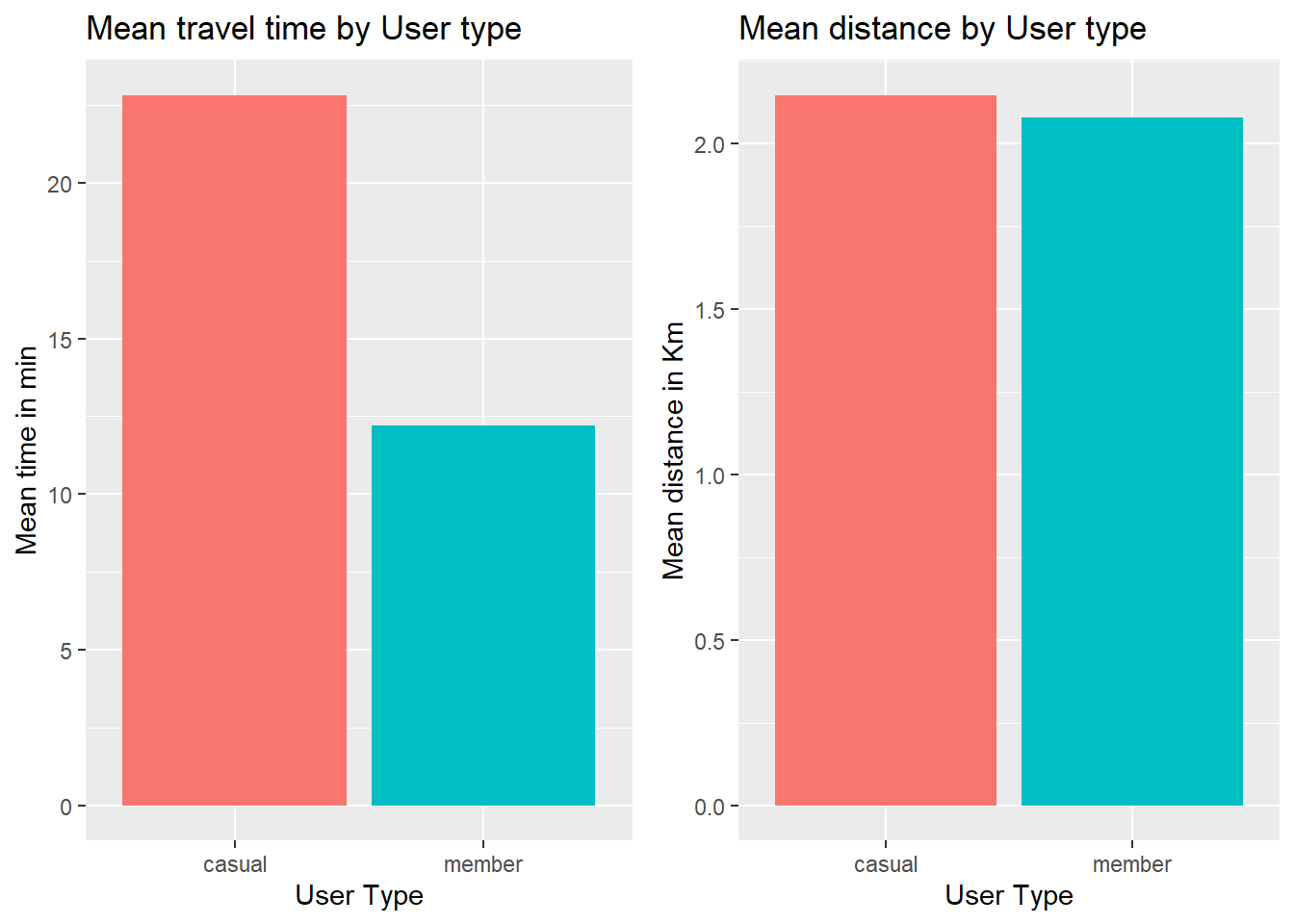
#Fist we calculate the average distance, distance for both the casual and member type users:  
  
summary(tripdata$ride\_duration)

## Length Class Mode   
## 3916664 difftime numeric

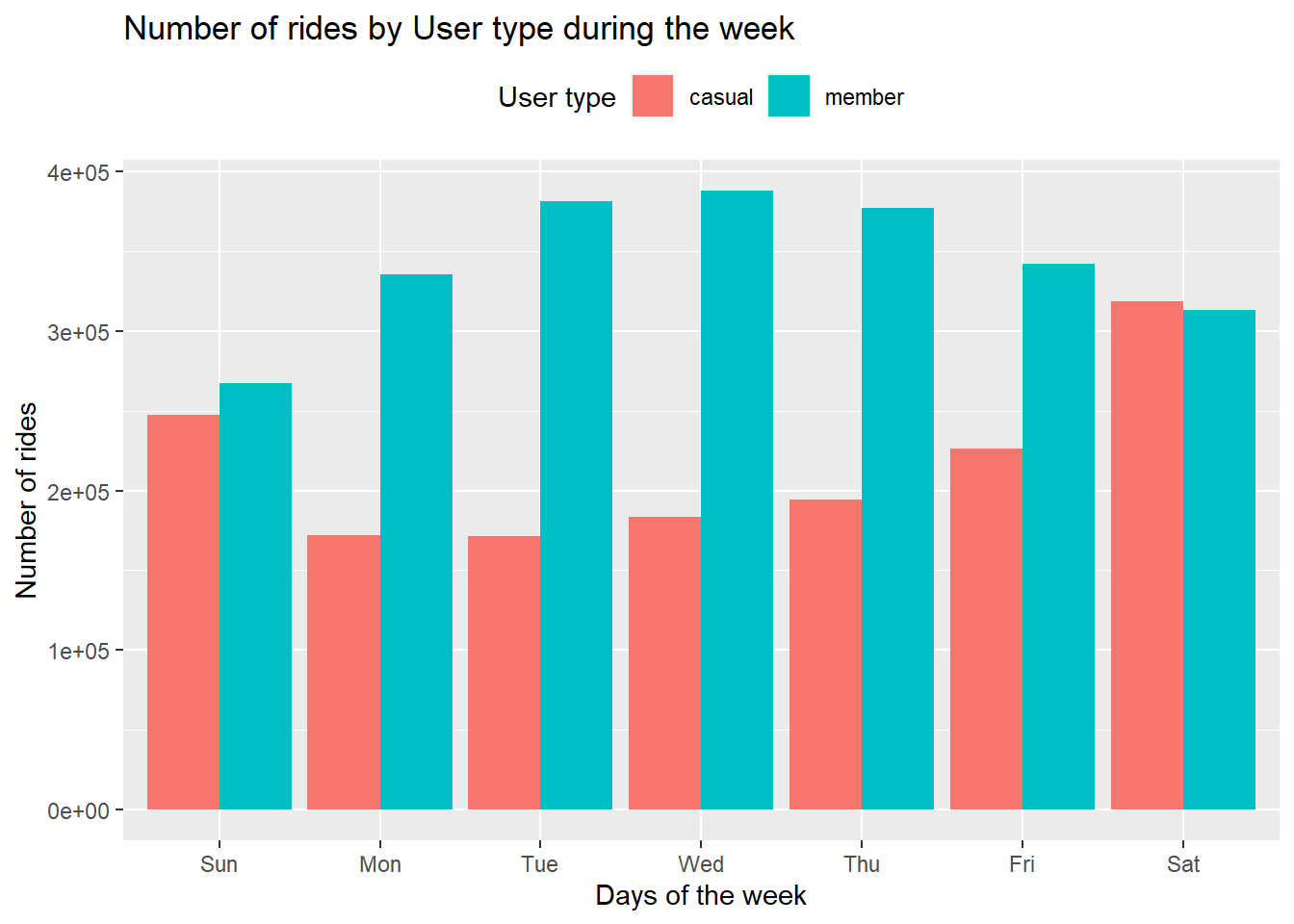
# Convert ride\_duration from seconds to minutes  
tripdata$ride\_duration\_minutes <- as.numeric(tripdata$ride\_duration / 60)  
  
  
userType\_means <- tripdata %>% group\_by(member\_casual) %>% summarise(mean\_time = mean(ride\_duration\_minutes),mean\_distance = mean(ride\_distance))  
  
head(userType\_means) #mean distance and time

## # A tibble: 2 × 3  
## member\_casual mean\_time mean\_distance  
## <chr> <dbl> <dbl>  
## 1 casual 22.8 2.15  
## 2 member 12.2 2.08

membervstime <- ggplot(userType\_means) +   
 geom\_col(mapping=aes(x=member\_casual,y=mean\_time,fill=member\_casual), show.legend = FALSE)+  
 labs(title = "Mean travel time by User type",x="User Type",y="Mean time in min")  
  
membervsdistance <- ggplot(userType\_means) +  
 geom\_col(mapping = aes(x=member\_casual,y=mean\_distance,fill=member\_casual),show.legend = FALSE)+  
 labs(title = "Mean distance by User type",x="User Type",y="Mean distance in Km")  
  
grid.arrange(membervstime, membervsdistance, ncol = 2)



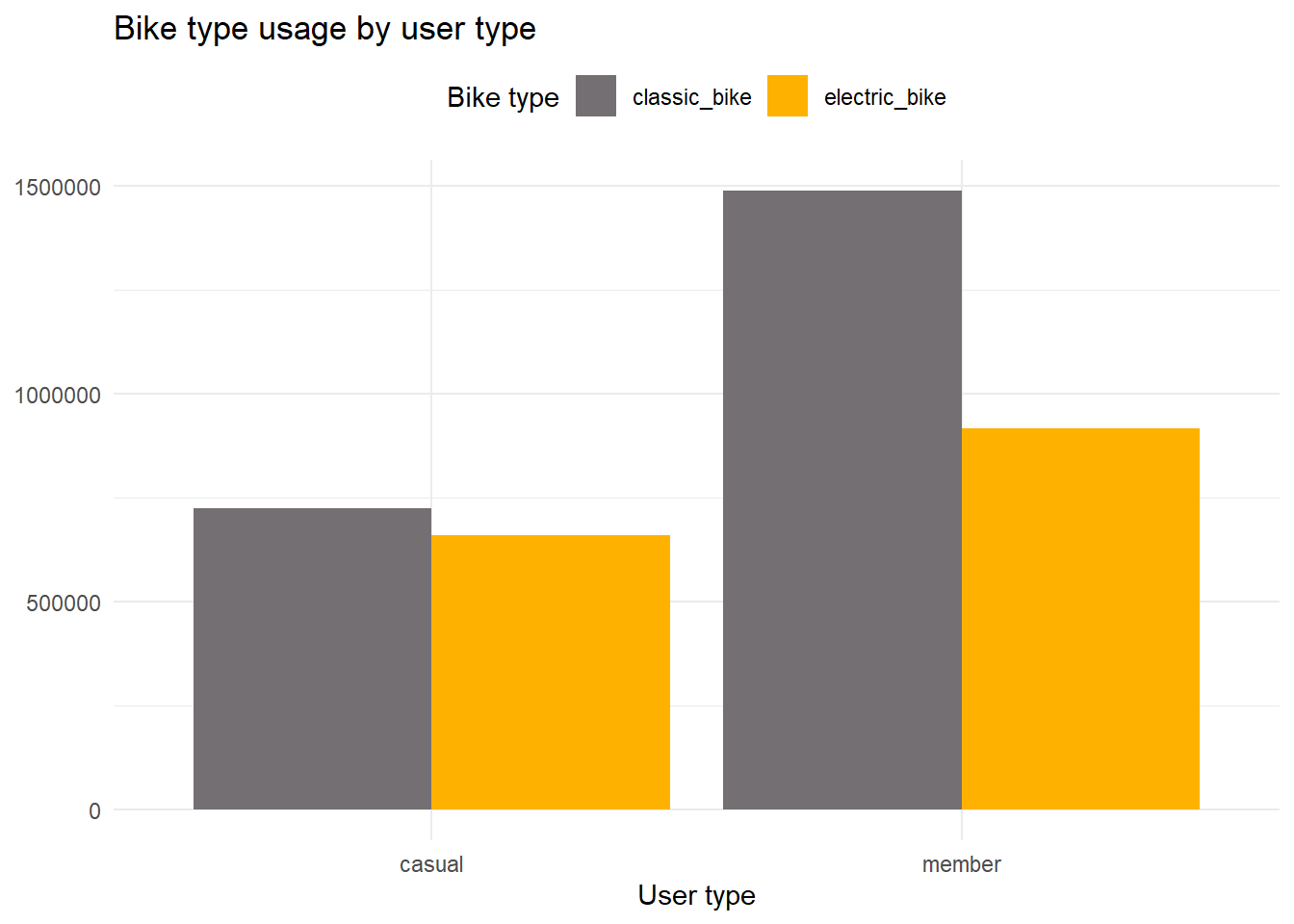
#number of rides by weekdays   
  
tripdata %>% mutate(weekday = wday(started\_at, label = TRUE)) %>%   
 group\_by(member\_casual, weekday) %>%   
 summarise(number\_of\_rides = n()  
 ,average\_duration = mean(ride\_duration\_minutes),.groups = 'drop') %>%   
 arrange(member\_casual, weekday) %>%   
 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", fill="User type") +  
 theme(legend.position="top")



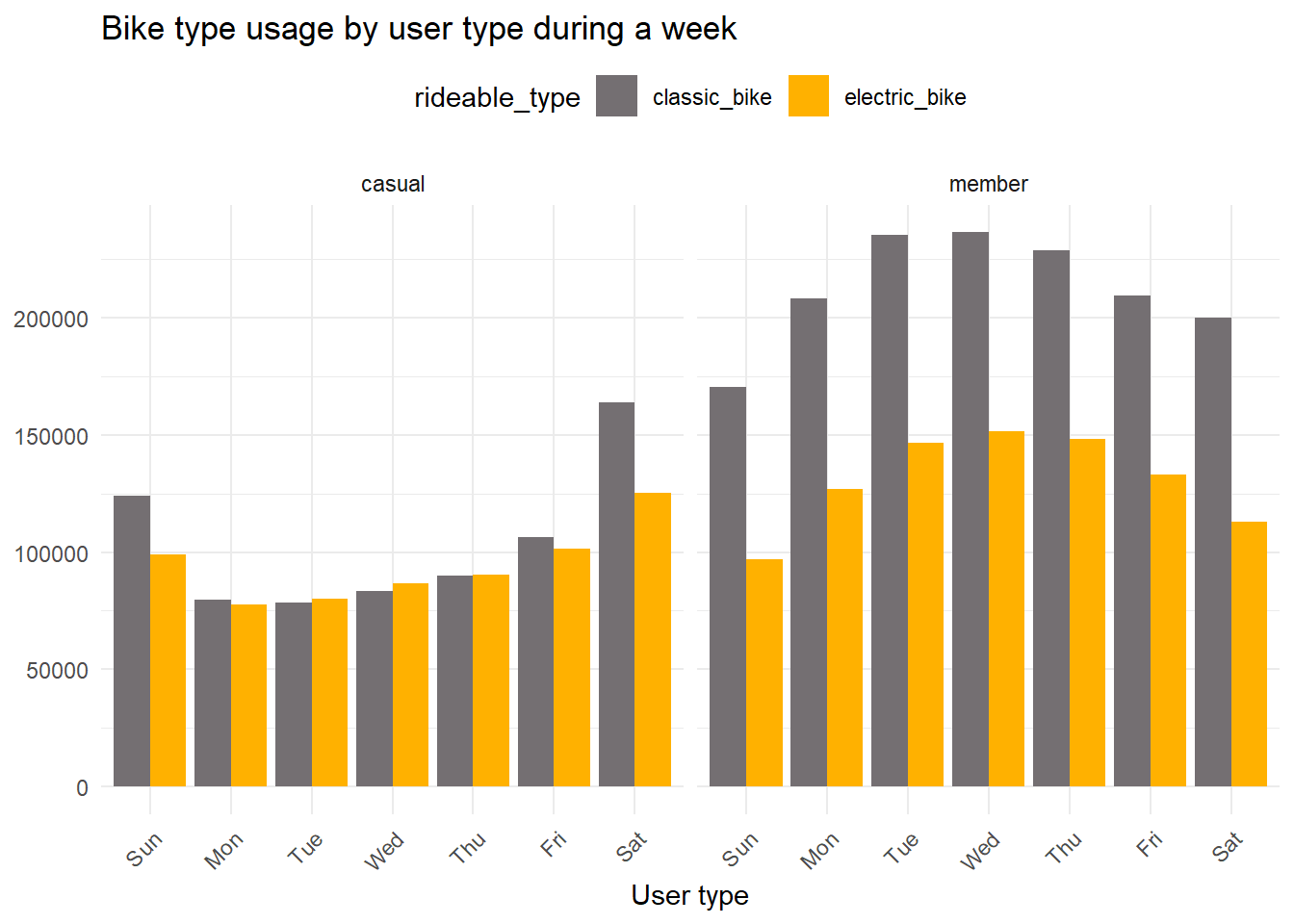
Analysis :

* It seems that the casual users travel the almost 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:  
  
bike\_type <- tripdata %>%   
 filter(rideable\_type=="classic\_bike" | rideable\_type=="electric\_bike")  
  
#Then lets check the bike type usage by user type:  
  
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\_minimal() +  
 theme(legend.position="top")



#And their Usage by both user types during a week :  
  
bike\_type %>%  
 mutate(weekday = wday(started\_at, label = TRUE)) %>%   
 group\_by(member\_casual,rideable\_type,weekday) %>%  
 summarise(totals=n(), .groups="drop") %>%   
 ggplot()+  
 geom\_col(aes(x=weekday,y=totals,fill=rideable\_type),position = "dodge")+  
 facet\_wrap(~member\_casual) +  
 labs(title = "Bike type usage by user type during a week",x="User type",y=NULL) +  
 scale\_fill\_manual(values = c("classic\_bike" = "#746F72","electric\_bike" = "#FFB100")) +  
 theme\_minimal()+  
 theme(axis.text.x = element\_text(angle = 45, hjust = 1))+  
 theme(legend.position="top")



Analysis:

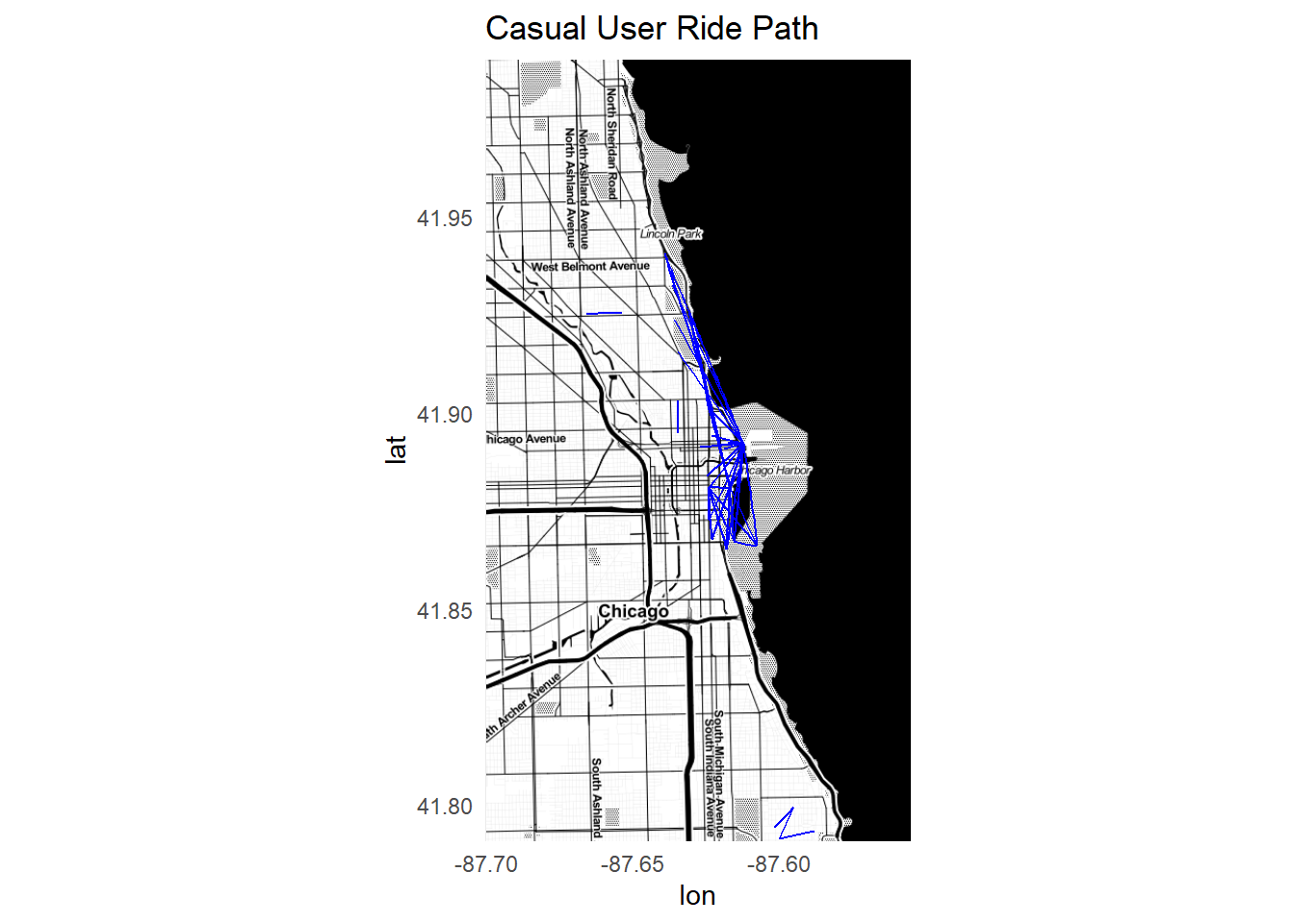
* Here we can see that the casual members use both types of bikes for their rides, but the member users show a higher preference for the Classic bikes.
* On a weekly basis we can see that for the casual members there is a small difference of usage in weekdays where as on weekend diffrence is higer and prefered to use classic bike more. Also on weekend there is high usages compare to weekdays.
* For the member users we see the usage on weedays is more and on weekends it much less. Prefrence to choose bike remains same in all days of week.

#Lets check now 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 <- tripdata %>%   
 filter(start\_lat != end\_lat & start\_lng != end\_lng) %>%   
 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")  
  
# Set the bounding box coordinates  
chi\_bb <- c(  
 left = -87.700424,  
 bottom = 41.790769,  
 right = -87.554855,  
 top = 41.990119  
)  
  
# Retrieve the Stamen map  
chicago\_stamen <- get\_stamenmap(  
 bbox = chi\_bb,  
 zoom = 12,  
 maptype = "toner"  
)

## ℹ Map tiles by Stamen Design, under CC BY 3.0. Data by OpenStreetMap, under ODbL.

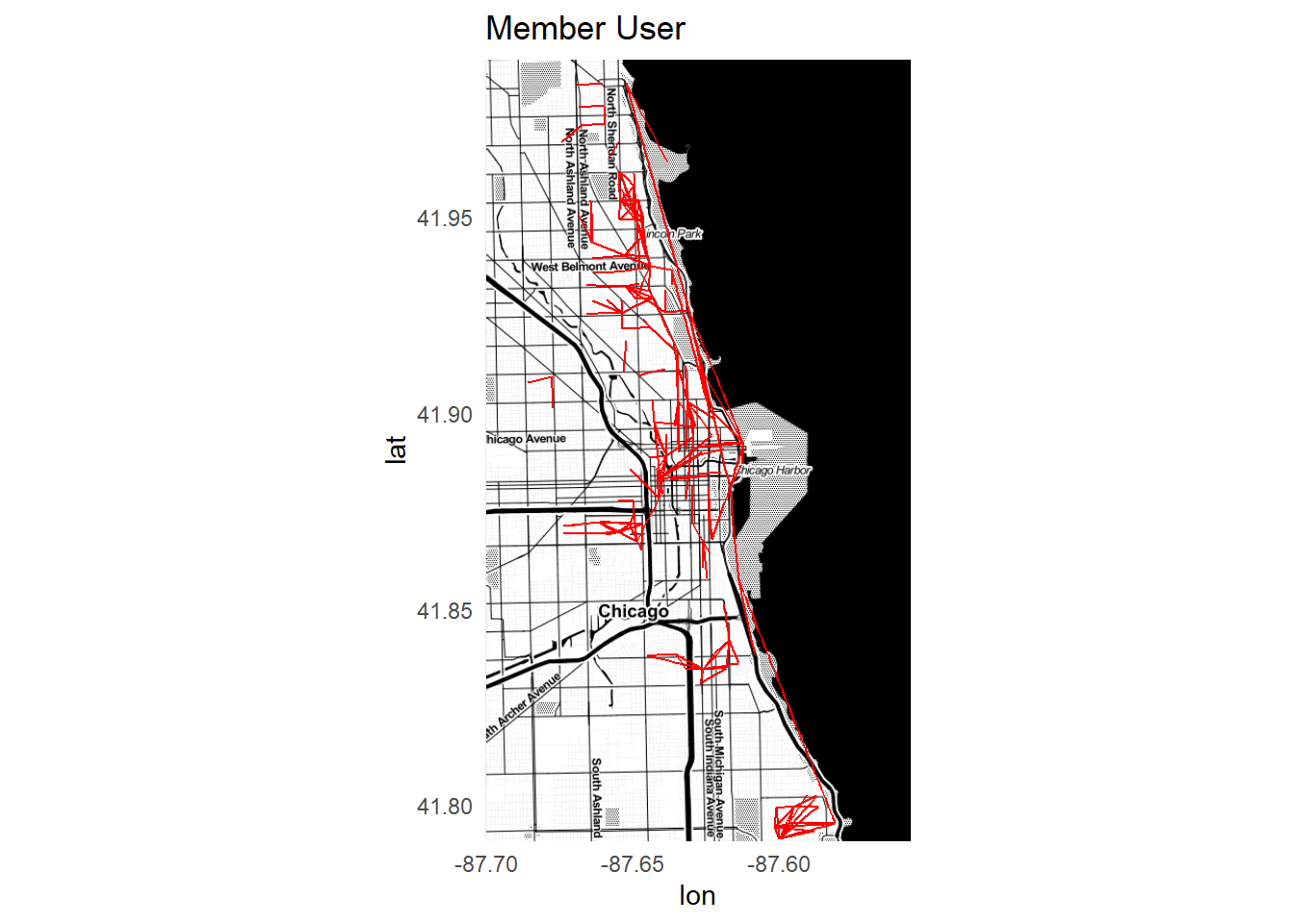
# Create a ggplot object and layer the map and ride paths  
map\_plot <- ggmap(chicago\_stamen) +  
 geom\_segment(data = casual, aes(x = start\_lng, y = start\_lat, xend = end\_lng, yend = end\_lat), color = "blue") +  
 ggtitle("Casual User Ride Path") +  
 theme\_minimal()  
  
map\_plot\_2 <- ggmap(chicago\_stamen) +  
 geom\_segment(data = member, aes(x = start\_lng, y = start\_lat, xend = end\_lng, yend = end\_lat), color = "red") +  
 ggtitle("Member User") +  
 theme\_minimal()  
  
print(map\_plot)

## Warning: Removed 8 rows containing missing values (`geom\_segment()`).



print(map\_plot\_2)

## Warning: Removed 36 rows containing missing values (`geom\_segment()`).



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 member users that connect the downtown with the outskirts of the city as well, that would suggest they are mostly people that live outside the downtown and use the service to commute for their daily work in various part of city.

PHASE 5 : Share

Share 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, and tourism rides mostly on weekends and using both electric and Classic bikes.Usages is more near city centre.

2)The Annual users have commute or pragmatic rides, during all days of week. They prefer classic bikes more then electric bike.

*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 aspect of the service, and maybe offer some kind of promotion related to weekends or to run promotional campaign on specific rout where usage high*