Research Report: R

Valentin Jimenez

02/02/2022

Third Part: Analysis with R

Target question: 1. How do annual members and casual riders use Cyclistic bikes differently?

For this part of the project I decided to use R, because I had some ideas I could develop with python but I realized I could do the same with R, also to show how graphics look like with ggplot2 visualizations.

1. Group by plane distance 12-Months

For the distance traveled we can see from the graphics that patterns for both type of riders are pretty similar with no relevant differences.

```
# In this script we analyze the number of riders, by type, taking into account the
# plane distance traveled.

library(ggplot2)
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v tibble  3.1.6     v dplyr   1.0.7
## v tidyr   1.1.4     v stringr 1.4.0
## v readr   2.1.1     v forcats 0.5.1
## v purrr   0.3.4
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(readxl)


excelFile = read.csv("C:/Users/Valentín/Valasus Dropbox/Valentín Jiménez/PC/Documents/VJDS/Programming/
#View(excelFile)

# We add a Months column to group_by after
excelFile <- excelFile %>%
  mutate(Month = format(as.Date(excelFile$started_at),"%m"))
#View(excelFile)

# we calculate the total by type of member
total_by_memberType <- excelFile %>%
  drop_na(start_lat,end_lat) %>%
  group_by(Month,member_casual) %>%
  summarise(total_rider_type = n())
```

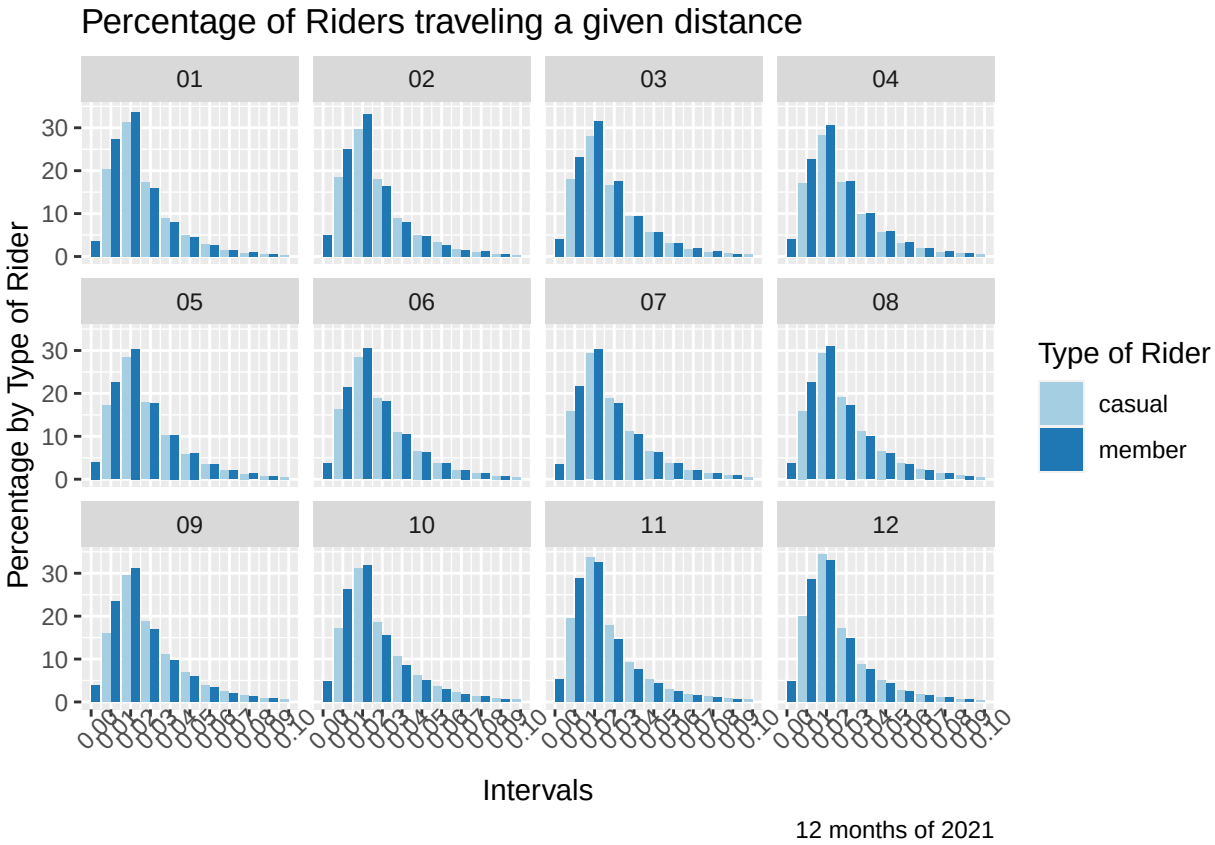
```
## `summarise()` has grouped output by 'Month'. You can override using the `.groups` argument.
```

```
# transformation for graph
proof2 <-
  excelFile %>%
  select(rideable_type,started_at,start_lat,start_lng,end_lat,end_lng,
         member_casual,ride_duration_hours,Month) %>%
  drop_na(start_lat,end_lat) %>%
  mutate(plane_distance = sqrt((start_lat-end_lat)^2+(start_lng-end_lng)^2)) %>%
  select(rideable_type, started_at, member_casual, ride_duration_hours,plane_distance,Month) %>%
  mutate( Intervals = case_when
         ( plane_distance == 0 & ride_duration_hours == 0 ~ -1,
           plane_distance == 0 & ride_duration_hours > 0 ~ 0,
           plane_distance > 0 & plane_distance <= 0.01 ~ 0.01,
           plane_distance > 0.01 & plane_distance <= 0.02 ~ 0.02,
           plane_distance > 0.02 & plane_distance <= 0.03 ~ 0.03,
           plane_distance > 0.03 & plane_distance <= 0.04 ~ 0.04,
           plane_distance > 0.04 & plane_distance <= 0.05 ~ 0.05,
           plane_distance > 0.05 & plane_distance <= 0.06 ~ 0.06,
           plane_distance > 0.06 & plane_distance <= 0.07 ~ 0.07,
           plane_distance > 0.07 & plane_distance <= 0.08 ~ 0.08,
           plane_distance > 0.08 & plane_distance <= 0.09 ~ 0.09,
           plane_distance > 0.09 & plane_distance <= 0.1 ~ 0.1,
           TRUE ~ -1
         )
        ) %>%
  group_by(Month,Intervals, member_casual) %>%
  summarise(total_by_interval = n()) %>%
  group_by(Month, Intervals,member_casual) %>%
  mutate(percentage = case_when(
      member_casual == "casual" ~
        as.numeric((total_by_interval*100)/total_by_memberType[as.numeric(Month)*2-1,3]),
      member_casual == "member" ~
        as.numeric((total_by_interval*100)/total_by_memberType[as.numeric(Month)*2,3])
    )
    )
```

```
## `summarise()` has grouped output by 'Month', 'Intervals'. You can override using the `.groups` argume
```

```
#View(proof2)

# Here we separate the graph from the dataFrame which we name "proof2"
proof2 %>%
  filter(Intervals != -1) %>%
  group_by(Month) %>%
  ggplot(aes(fill=member_casual, y = percentage , x = Intervals)) +
  geom_bar(position="dodge", stat="identity")+
  theme(axis.text.x = element_text(angle = 45))+
  ggtitle("Percentage of Riders traveling a given distance")+
  labs(y = "Percentage by Type of Rider",caption = "12 months of 2021",
       fill = "Type of Rider")+
  scale_x_continuous(breaks = seq(0,0.1,0.01),limits = c(0,0.1))+
  scale_fill_brewer(palette = "Paired" )+
  facet_wrap(~Month)
```

```
## Warning: Removed 24 rows containing missing values (geom_bar).
```



2. Group by street annual

In this part, first I considered the exact location of each station; however, I got so many values that I did not obtain relevant information, but if instead we group the stations by street name we get more useful information: I used the initial name of a street to classify the stations. In the final results I show only names of stations with more or equal to 2% of users.

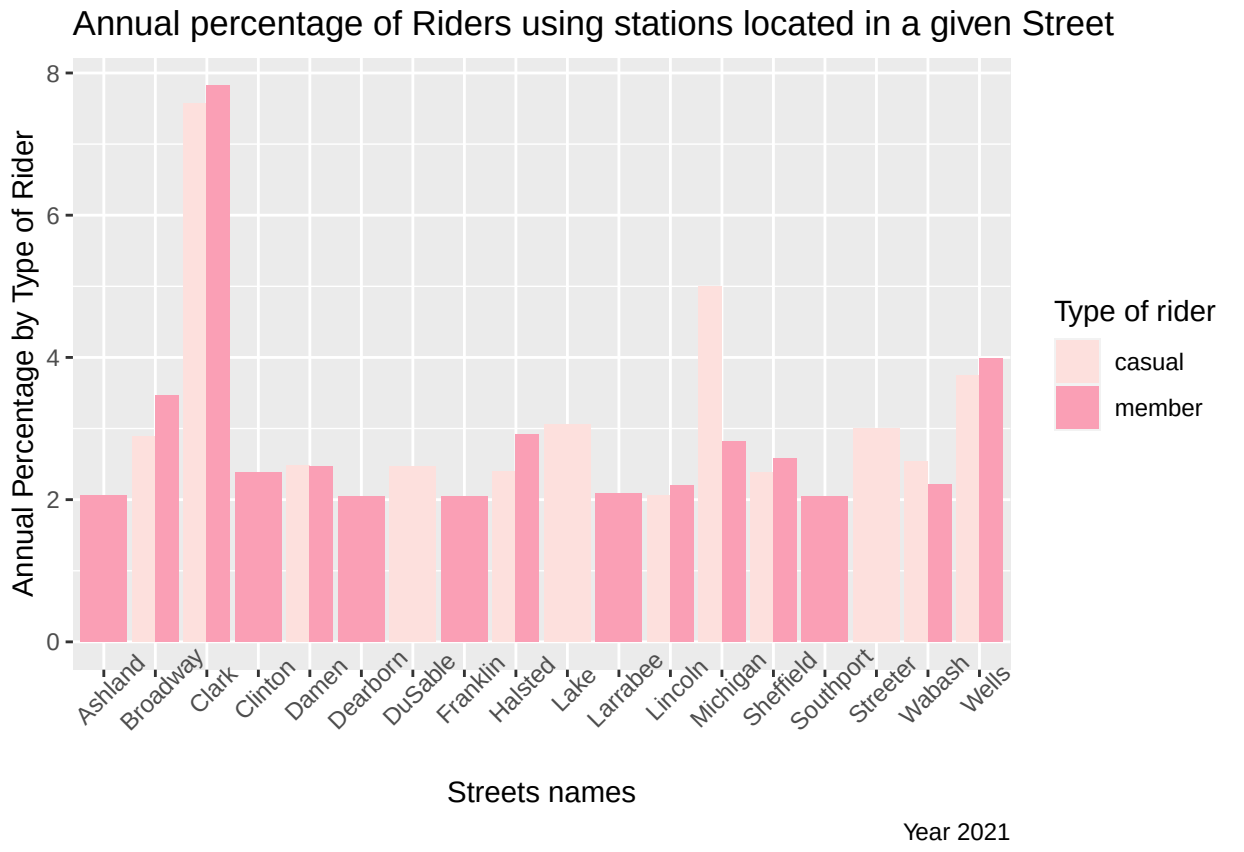
We see that annually, stations located at **Clark** street are by far the most used station for both type of riders. For casual riders there are specific streets used only by them, these are **Ashland, DuSable, Lake, Streeter**. Streets used by both riders but with casual riders being the most ones are **Michigan and Wells**.

```
# In this script we analyze the annual percentage of riders, by type, using stations
# located at a given street.

library(ggplot2)
library(tidyverse)
library(readxl)
library(stringr)

#excelFile = read.csv("C:/Users/Valentín/Valasus Dropbox/Valentín Jiménez/PC/Documents/VJDS/Programming
#View(excelFile)

# Here we use regex to form a column with the beginning of the name street so
# we can calssify better
excelFile <- excelFile %>%
  mutate(street = str_extract(string = excelFile$start_station_id_complete,
                              pattern = "^\\S+"))
```

```r
# we count the number of total riders by type using each street
excelFile <- excelFile %>%
  drop_na(street) %>%
  group_by(member_casual, street) %>%
  summarise(n=n())
```

```
## `summarise()` has grouped output by 'member_casual'. You can override using the `.groups` argument.
```

```r
# we calculate the total by type of member and by street
totals_by_riderType <- excelFile %>%
  group_by(member_casual) %>%
  summarise(sum(n))

# we add the percentage column
excelFile <- excelFile %>%
  group_by(member_casual,street) %>%
  mutate(percentage = case_when(
    member_casual == "casual" ~ as.numeric((n*100)/totals_by_riderType[1,2]),
    member_casual == "member" ~ as.numeric((n*100)/totals_by_riderType[2,2])
    ))
#View(excelFile)

# Now we can visualize properly
excelFile %>%
  filter(percentage >=2) %>%
  ggplot(aes(fill=member_casual, y = percentage , x = street)) +
  geom_bar(position="dodge", stat="identity")+
  ggtitle("Annual percentage of Riders using stations located in a given Street")+
  labs(x = "Streets names", caption = "Year 2021", y = "Annual Percentage by Type of Rider",
       fill = "Type of rider")+
  scale_fill_brewer(palette = "RdPu")+
  theme(axis.text.x = element_text(angle = 45))
```



3. Group by street 12-Months

This part follows the ideas of the point **2.** of this part of the project, but now we look for street names at each Month of 2021 and with a monthly percentage of more or equal to 2.5%.

We confirm that for each month **Clark** is still the most used one for both type of riders. We can see that the streets **Broadway, Lake, Wells**, are used by casual members through out each month of the year.

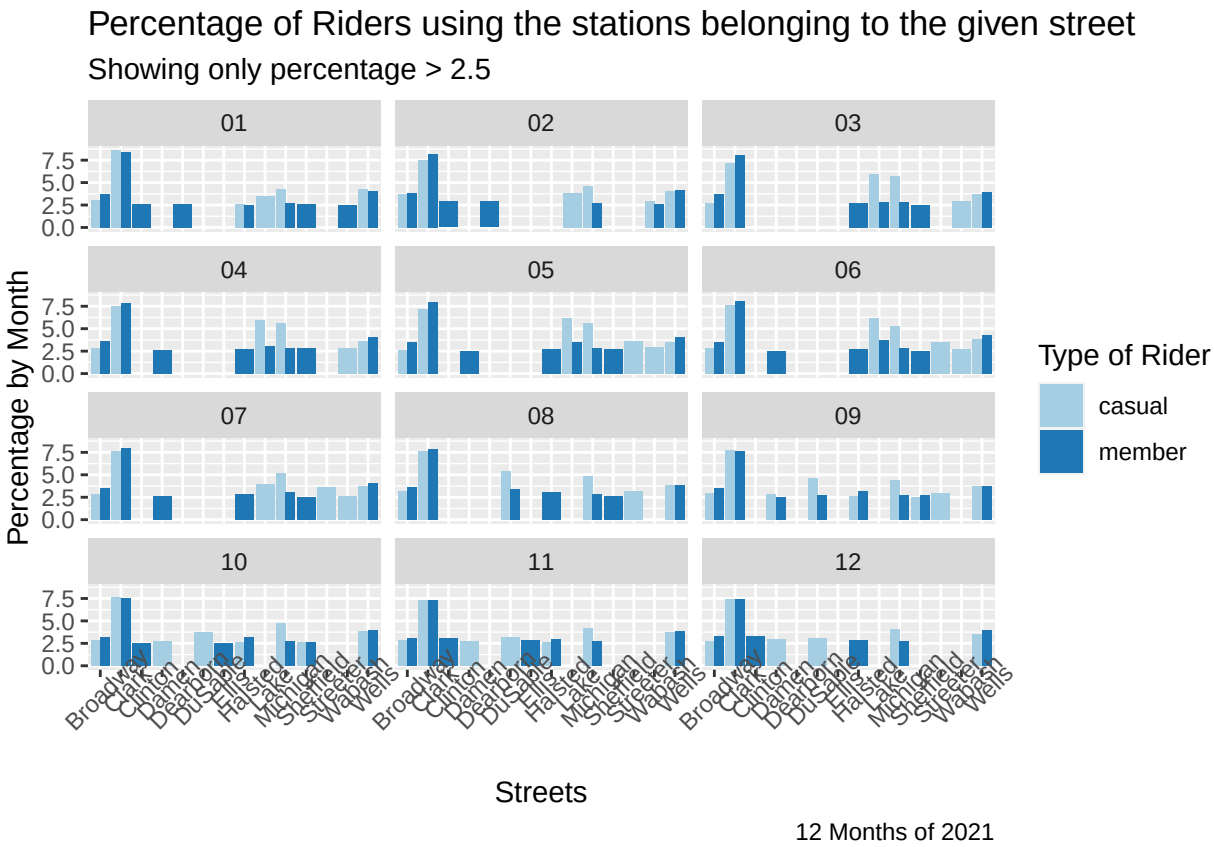
```
# In this script we analyze the number of riders, using the stations
# located at a certain street.

library(ggplot2)
library(tidyverse)
library(readxl)
library(stringr)

excelFile = read.csv("C:/Users/Valentín/Valasus Dropbox/Valentín Jiménez/PC/Documents/VJDS/Programming/
#View(excelFile)

# Here we use regex to form a column with the beginning of the name street so
# we can calssify better
excelFile <- excelFile %>%
  mutate(street = str_extract(string = excelFile$start_station_id_complete,
                              pattern = "^\\S+")) %>%
  mutate(Month = format(as.Date(excelFile$started_at),"%m")) %>%
  drop_na(street) %>%
  group_by(Month,member_casual,street) %>%
  summarise(n=n())
```

```
## `summarise()` has grouped output by 'Month', 'member_casual'. You can override using the `.groups` a
```

```
#View(excelFile)

# we calculate the total by type of rider and by month
totals_by_Month <- excelFile %>%
  group_by(Month,member_casual) %>%
  summarise(sum_Month = sum(n))
```

```
## `summarise()` has grouped output by 'Month'. You can override using the `.groups` argument.
```

```
#View(totals_by_Month)

# we add the percentage column
excelFile <- excelFile %>%
  group_by(Month, member_casual,street) %>%
  mutate(percentage = case_when(
    member_casual == "casual" ~ as.numeric((n*100)/totals_by_Month[as.numeric(Month)*2-1,3]),
    member_casual == "member" ~ as.numeric((n*100)/totals_by_Month[as.numeric(Month)*2,3])
    ))
#View(excelFile)

# Now we can visualize properly
  excelFile %>%
  #select(Month, member_casual,street,percentage) %>%
  filter(percentage >2.5) %>%
  ggplot(aes(fill=member_casual, y = percentage , x = street)) +
  geom_bar(position="dodge", stat="identity")+
  theme(axis.text.x = element_text(angle = 45))+
  ggtitle("Percentage of Riders using the stations belonging to the given street")+
  labs(subtitle = "Showing only percentage > 2.5", y = "Percentage by Month", x = "Streets",
       fill = "Type of Rider",caption = "12 Months of 2021")+
  scale_fill_brewer(palette = "Paired")+
  facet_wrap(~Month, ncol = 3)
```


Percentage of Riders using the stations belonging to the given street
Showing only percentage > 2.5
01
02
03
04
05
06
07
08
09
10
11
12
7.5
5.0
2.5
0.0
Percentage by Month
Broadway
Clark
Clinton
Damen
Dearborn
Du Sable
Ellis
Halsted
Lake
Michigan
Sheffield
Streeter
Wabash
Wells
Streets
Type of Rider
casual
member
12 Months of 2021