Case Study: Cyclistic

Ernesto González (@bluezef)

2023-10-31

Cyclistic

How Does a Bike-Share Navigate Speedy Success?

Business task: Investigate the differences between Cyclistic annual members and casual rider in relation to the service use.

Importing Libraries

Importing libraries tidyverse,dplyr,readr,ggplot2 and lubricate for data analysis

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.3     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.4     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.0
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(dplyr)
library(readr)
library(ggplot2)
library(lubridate)
```

Merging CSV files into one data frame

Collecting data from 12 .csv files

```
divvy_tripdata <- list.files(path='~/Case Study Cyclistic', pattern="*.csv") %>%
  lapply(read_csv) %>%
  bind_rows
```

```
## Rows: 558685 Columns: 13
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (7): ride_id, rideable_type, start_station_name, start_station_id, end_...
## dbl  (4): start_lat, start_lng, end_lat, end_lng
## dttm (2): started_at, ended_at
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
## Rows: 337735 Columns: 13
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (7): ride_id, rideable_type, start_station_name, start_station_id, end_...
## dbl  (4): start_lat, start_lng, end_lat, end_lng
## dttm (2): started_at, ended_at
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
## Rows: 181806 Columns: 13
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (7): ride_id, rideable_type, start_station_name, start_station_id, end_...
## dbl  (4): start_lat, start_lng, end_lat, end_lng
## dttm (2): started_at, ended_at
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
## Rows: 190301 Columns: 13
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (7): ride_id, rideable_type, start_station_name, start_station_id, end_...
## dbl  (4): start_lat, start_lng, end_lat, end_lng
## dttm (2): started_at, ended_at
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
## Rows: 190445 Columns: 13
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (7): ride_id, rideable_type, start_station_name, start_station_id, end_...
## dbl  (4): start_lat, start_lng, end_lat, end_lng
## dttm (2): started_at, ended_at
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
## Rows: 258678 Columns: 13
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (7): ride_id, rideable_type, start_station_name, start_station_id, end_...
## dbl  (4): start_lat, start_lng, end_lat, end_lng
## dttm (2): started_at, ended_at
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
## Rows: 426590 Columns: 13
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (7): ride_id, rideable_type, start_station_name, start_station_id, end_...
## dbl  (4): start_lat, start_lng, end_lat, end_lng
## dttm (2): started_at, ended_at
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
## Rows: 604827 Columns: 13
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (7): ride_id, rideable_type, start_station_name, start_station_id, end_...
## dbl  (4): start_lat, start_lng, end_lat, end_lng
## dttm (2): started_at, ended_at
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
## Rows: 719618 Columns: 13
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (7): ride_id, rideable_type, start_station_name, start_station_id, end_...
## dbl  (4): start_lat, start_lng, end_lat, end_lng
## dttm (2): started_at, ended_at
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
## Rows: 767650 Columns: 13
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (7): ride_id, rideable_type, start_station_name, start_station_id, end_...
## dbl  (4): start_lat, start_lng, end_lat, end_lng
## dttm (2): started_at, ended_at
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
## Rows: 771693 Columns: 13
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (7): ride_id, rideable_type, start_station_name, start_station_id, end_...
## dbl  (4): start_lat, start_lng, end_lat, end_lng
## dttm (2): started_at, ended_at
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
## Rows: 666371 Columns: 13
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (7): ride_id, rideable_type, start_station_name, start_station_id, end_...
## dbl  (4): start_lat, start_lng, end_lat, end_lng
## dttm (2): started_at, ended_at
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```r
glimpse(divvy_tripdata)
```

```
## Rows: 5,674,399
## Columns: 13
## $ ride_id            <chr> "A50255C1E17942AB", "DB692A70BD2DD4E3", "3C02727AAF~
## $ rideable_type      <chr> "classic_bike", "electric_bike", "electric_bike", "~
## $ started_at         <dttm> 2022-10-14 17:13:30, 2022-10-01 16:29:26, 2022-10-~
## $ ended_at           <dttm> 2022-10-14 17:19:39, 2022-10-01 16:49:06, 2022-10-~
## $ start_station_name <chr> "Noble St & Milwaukee Ave", "Damen Ave & Charleston~
## $ start_station_id   <chr> "13290", "13288", "655", "KA1504000133", "13028", "~
## $ end_station_name   <chr> "Larrabee St & Division St", "Damen Ave & Cullerton~
## $ end_station_id     <chr> "KA1504000079", "13089", "TA1307000140", "620", "13~
## $ start_lat          <dbl> 41.90068, 41.92004, 41.97988, 41.90227, 41.87475, 4~
## $ start_lng          <dbl> -87.66260, -87.67794, -87.68190, -87.62769, -87.649~
## $ end_lat            <dbl> 41.90349, 41.85497, 41.96640, 41.89820, 41.86610, 4~
## $ end_lng            <dbl> -87.64335, -87.67570, -87.68870, -87.63754, -87.607~
## $ member_casual      <chr> "member", "casual", "member", "member", "casual", "~
```

Cleaning data

Removing null values and duplicated

```r
divvy_tripdata_cleaned <- divvy_tripdata %>%
  na.omit() %>%
  distinct()
```

Previewing data

Data Frame Structure

```r
str(divvy_tripdata_cleaned)
```

```
## tibble [4,291,451 x 13] (S3: tbl_df/tbl/data.frame)
##  $ ride_id           : chr [1:4291451] "A50255C1E17942AB" "DB692A70BD2DD4E3" "3C02727AAF60F873" "47E
##  $ rideable_type     : chr [1:4291451] "classic_bike" "electric_bike" "electric_bike" "electric_bike
##  $ started_at        : POSIXct[1:4291451], format: "2022-10-14 17:13:30" "2022-10-01 16:29:26" ...
##  $ ended_at          : POSIXct[1:4291451], format: "2022-10-14 17:19:39" "2022-10-01 16:49:06" ...
##  $ start_station_name: chr [1:4291451] "Noble St & Milwaukee Ave" "Damen Ave & Charleston St" "Hoyne
##  $ start_station_id  : chr [1:4291451] "13290" "13288" "655" "KA1504000133" ...
##  $ end_station_name  : chr [1:4291451] "Larrabee St & Division St" "Damen Ave & Cullerton St" "Weste
##  $ end_station_id    : chr [1:4291451] "KA1504000079" "13089" "TA1307000140" "620" ...
##  $ start_lat         : num [1:4291451] 41.9 41.9 42 41.9 41.9 ...
##  $ start_lng         : num [1:4291451] -87.7 -87.7 -87.7 -87.6 -87.6 ...
##  $ end_lat           : num [1:4291451] 41.9 41.9 42 41.9 41.9 ...
##  $ end_lng           : num [1:4291451] -87.6 -87.7 -87.7 -87.6 -87.6 ...
##  $ member_casual     : chr [1:4291451] "member" "casual" "member" "member" ...
##  - attr(*, "na.action")= 'omit' Named int [1:1382948] 2848 2849 2850 2851 2852 2854 2856 4359 4360 4
##   ..- attr(*, "names")= chr [1:1382948] "2848" "2849" "2850" "2851" ...
```

Calculating travel times, day of the week and month of trip

Adding columns for trip length, day of the week in which the trip was done, and month of the trip

```
divvy_tripdata_cleaned <- divvy_tripdata_cleaned %>%
  mutate(trip_length=ended_at-started_at) %>%
  mutate(day_of_week=weekdays(started_at)) %>%
  mutate(month_trip=month(started_at))
str(divvy_tripdata_cleaned)
```

```
## tibble [4,291,451 x 16] (S3: tbl_df/tbl/data.frame)
##  $ ride_id           : chr [1:4291451] "A50255C1E17942AB" "DB692A70BD2DD4E3" "3C02727AAF60F873" "47E
##  $ rideable_type     : chr [1:4291451] "classic_bike" "electric_bike" "electric_bike" "electric_bike
##  $ started_at        : POSIXct[1:4291451], format: "2022-10-14 17:13:30" "2022-10-01 16:29:26" ...
##  $ ended_at          : POSIXct[1:4291451], format: "2022-10-14 17:19:39" "2022-10-01 16:49:06" ...
##  $ start_station_name: chr [1:4291451] "Noble St & Milwaukee Ave" "Damen Ave & Charleston St" "Hoyne
##  $ start_station_id  : chr [1:4291451] "13290" "13288" "655" "KA1504000133" ...
##  $ end_station_name  : chr [1:4291451] "Larrabee St & Division St" "Damen Ave & Cullerton St" "Weste
##  $ end_station_id    : chr [1:4291451] "KA1504000079" "13089" "TA1307000140" "620" ...
##  $ start_lat         : num [1:4291451] 41.9 41.9 42 41.9 41.9 ...
##  $ start_lng         : num [1:4291451] -87.7 -87.7 -87.7 -87.6 -87.6 ...
##  $ end_lat           : num [1:4291451] 41.9 41.9 42 41.9 41.9 ...
##  $ end_lng           : num [1:4291451] -87.6 -87.7 -87.7 -87.6 -87.6 ...
##  $ member_casual     : chr [1:4291451] "member" "casual" "member" "member" ...
##  $ trip_length       : 'difftime' num [1:4291451] 369 1180 470 373 ...
##   ..- attr(*, "units")= chr "secs"
##  $ day_of_week       : chr [1:4291451] "viernes" "sábado" "miércoles" "lunes" ...
##  $ month_trip        : num [1:4291451] 10 10 10 10 10 10 10 10 10 10 ...
##  - attr(*, "na.action")= 'omit' Named int [1:1382948] 2848 2849 2850 2851 2852 2854 2856 4359 4360 43
##   ..- attr(*, "names")= chr [1:1382948] "2848" "2849" "2850" "2851" ...
```

Min,Max and Mean

Minimum, Maximum and Mean values for current dataframe

```
min(divvy_tripdata_cleaned$trip_length)
```

```
## Time difference of -10122 secs
```

```
max(divvy_tripdata_cleaned$trip_length)
```

```
## Time difference of 728178 secs
```

```
mean(divvy_tripdata_cleaned$trip_length)
```

```
## Time difference of 957.9625 secs
```

Max and Min far from Mean

Sorting

Looking for outliers sorting data ascending and descending by trip length

```
select(arrange(divvy_tripdata_cleaned,trip_length),trip_length)
```

```
## # A tibble: 4,291,451 x 1
##    trip_length
##    <drtn>
##  1 -10122 secs
##  2  -3479 secs
##  3  -3422 secs
##  4  -3314 secs
##  5  -3302 secs
##  6  -3164 secs
##  7  -3090 secs
##  8  -3078 secs
##  9  -3020 secs
## 10  -2964 secs
## # i 4,291,441 more rows
```

```
select(arrange(divvy_tripdata_cleaned,desc(trip_length)),trip_length)
```

```
## # A tibble: 4,291,451 x 1
##    trip_length
##    <drtn>
##  1 728178 secs
##  2 669136 secs
##  3 494628 secs
##  4 413473 secs
##  5 322740 secs
##  6 294187 secs
##  7 229104 secs
##  8 198050 secs
##  9 194747 secs
## 10 147471 secs
## # i 4,291,441 more rows
```

Filtering due to outliers

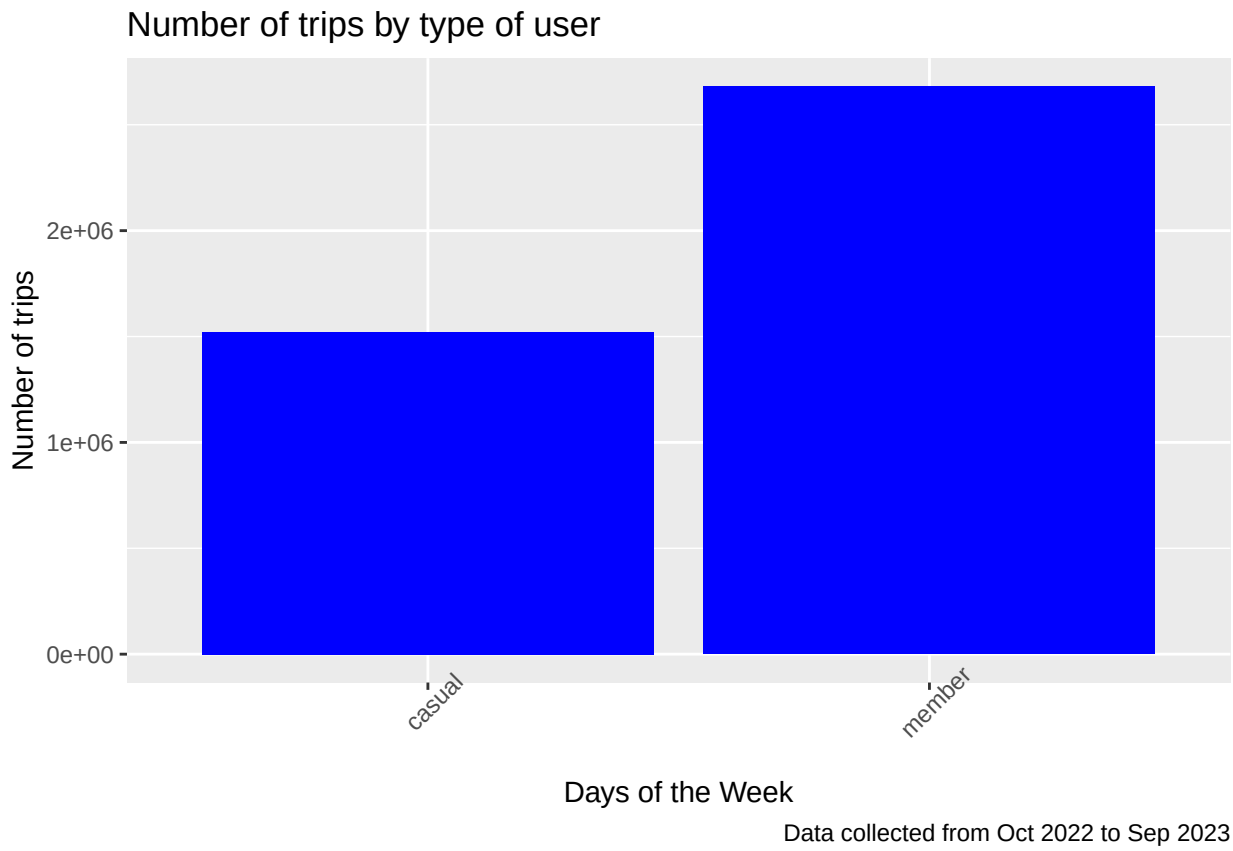
Removing the outliers

```
divvy_tripdata_cleaned <- divvy_tripdata_cleaned %>%
  filter(trip_length>60 & trip_length<90000)
```

Trips per type of user

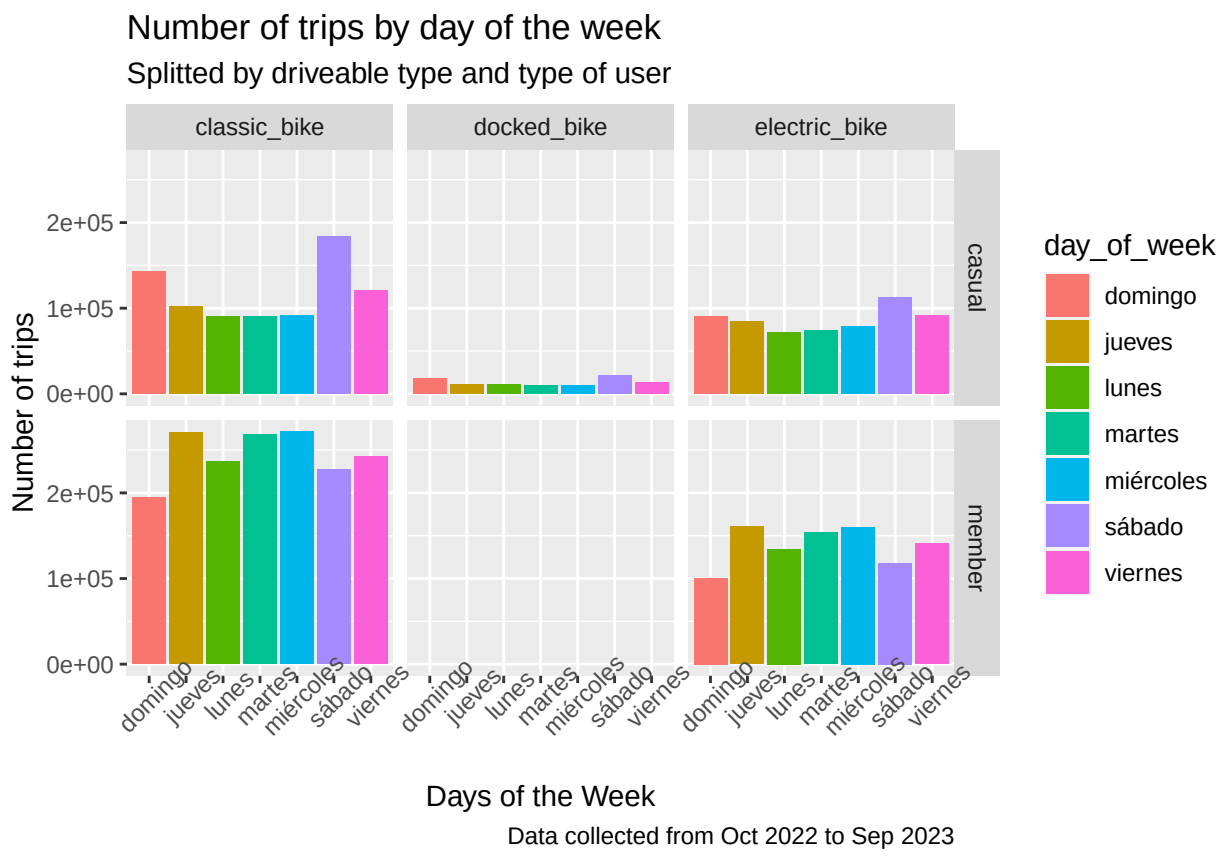
```
ggplot(divvy_tripdata_cleaned)+
  geom_bar(aes(member_casual),fill="blue")+
```

```
  theme(axis.text.x = element_text(angle = 45))+
  labs(title="Number of trips by type of user",
       x="Days of the Week", y="Number of trips",
       caption="Data collected from Oct 2022 to Sep 2023")
```



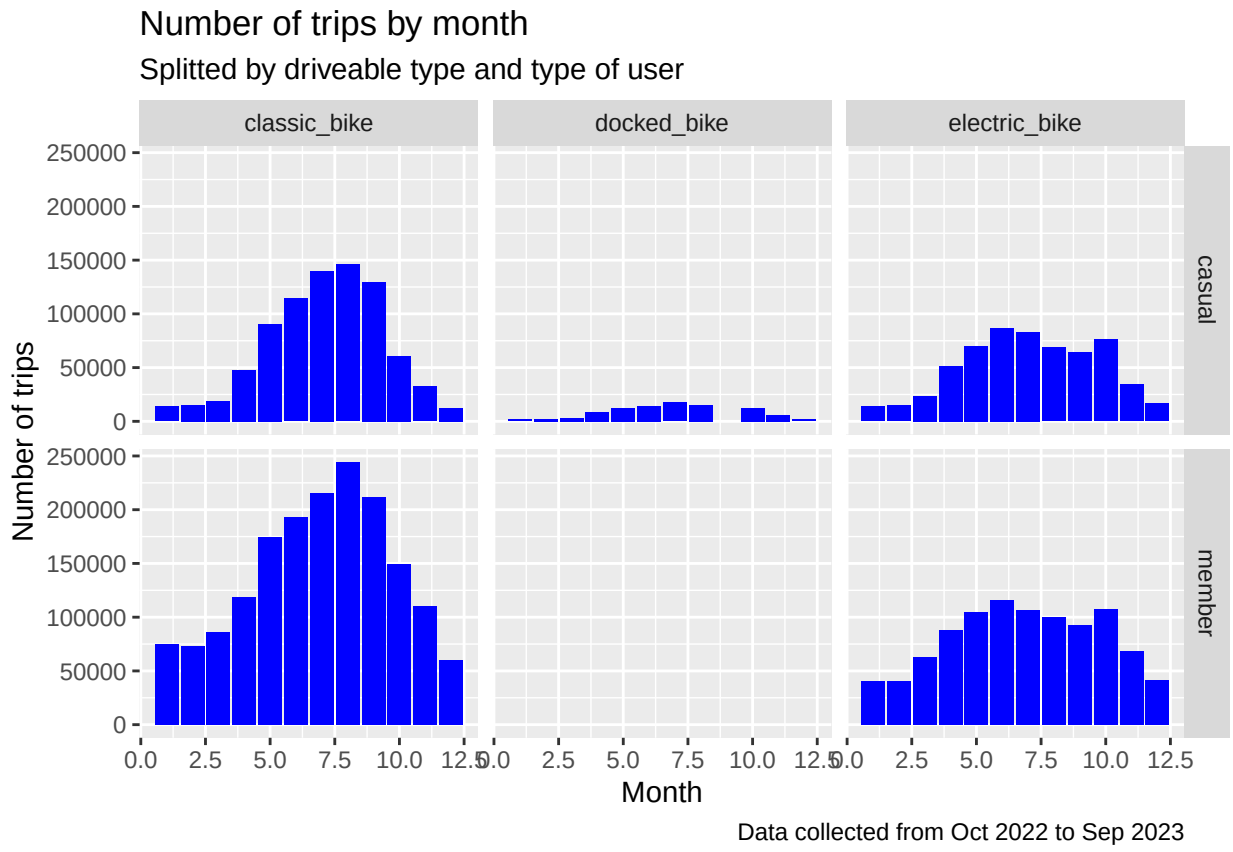
Trips per day of the week

```
ggplot(divvy_tripdata_cleaned)+
  geom_bar(aes(day_of_week,fill=day_of_week))+
  facet_grid(~member_casual~rideable_type)+
  theme(axis.text.x = element_text(angle = 45))+
  labs(title="Number of trips by day of the week",
       subtitle="Splitted by driveable type and type of user",
       x="Days of the Week", y="Number of trips",
       caption="Data collected from Oct 2022 to Sep 2023")
```



Trips per month

```
ggplot(divvy_tripdata_cleaned)+
  geom_bar(aes(month_trip),fill="blue")+
  facet_grid(~member_casual~rideable_type)+
  labs(title="Number of trips by month",
       subtitle="Splitted by driveable type and type of user",
       x="Month", y="Number of trips",
       caption="Data collected from Oct 2022 to Sep 2023")
```



Min,Max and Mean for Members Minimum, Maximum and Mean values for Member users

```
min(filter(divvy_tripdata_cleaned,member_casual=="member")$trip_length)
```

```
## Time difference of 61 secs
```

```
max(filter(divvy_tripdata_cleaned,member_casual=="member")$trip_length)
```

```
## Time difference of 89872 secs
```

```
mean(filter(divvy_tripdata_cleaned,member_casual=="member")$trip_length)
```

```
## Time difference of 743.3323 secs
```

Min,Max and Mean for Casual

Minimum, Maximum and Mean values for Casual users

```
min(filter(divvy_tripdata_cleaned,member_casual=="casual")$trip_length)
```

```
## Time difference of 61 secs
```

```r
max(filter(divvy_tripdata_cleaned,member_casual=="casual")$trip_length)
```

```
## Time difference of 89929 secs
```

```r
mean(filter(divvy_tripdata_cleaned,member_casual=="casual")$trip_length)
```

```
## Time difference of 1388.234 secs
```

Findings

- Casual users make around 50% of the trips that the Members do
- Non member users mostly use our service during Saturday, this may be due to the fact of it being weekend, as recreation. While member users use our service throughout the whole week, this due to them using the service constantly to go to work.
- The mean trip length for Members is 743.33 secs while for Casuals is 1388.23 secs, as they may be using the service as recreation, this may be the reason why their trips are longer compared to the member trips who mainly use it to go to work.
- The month of the year in where most trips are made are in August, this may be due to the Summer season.

Possible future studies

- Why do Casuals take longer trips when trips are less than the amount of trips made by Members? Hypothesis: The members mostly use the service to go to work on weekdays, which make the room for longer trips smaller.
- Why do Casuals do much more less trips than Members? Hypothesis: They drive, walk or take a bus to their work, so a membership is not really necessary

Recommendations

- Start campaigns or promotional events for joining new members on August's Saturdays to reach most of the casual users.