Cyclist\_data

Ajay

### Loading tidyverse and gt packages

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 3.4.2 ✔ tibble 3.2.1  
✔ lubridate 1.9.2 ✔ tidyr 1.3.0  
✔ purrr 1.0.1   
── 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(gt)

### Loading data of previous 12 months

trpdata\_july\_2022<-read\_csv("F:/Data\_Sci/Cap\_Stone\_Project/Cyclist\_trip\_data/202207-divvy-tripdata/202207-divvy-tripdata.csv")  
  
trpdata\_aug\_2022 <- read\_csv("F:/Data\_Sci/Cap\_Stone\_Project/Cyclist\_trip\_data/202208-divvy-tripdata/202208-divvy-tripdata.csv")  
  
trpdata\_sept\_2022<- read\_csv("F:/Data\_Sci/Cap\_Stone\_Project/Cyclist\_trip\_data/202209-divvy-tripdata/202209-divvy-publictripdata.csv")  
  
trpdata\_oct\_2022<- read\_csv("F:/Data\_Sci/Cap\_Stone\_Project/Cyclist\_trip\_data/202210-divvy-tripdata/202210-divvy-tripdata\_raw.csv")  
  
trpdata\_nov\_2022<- read\_csv("F:/Data\_Sci/Cap\_Stone\_Project/Cyclist\_trip\_data/202211-divvy-tripdata/202211-divvy-tripdata.csv")  
  
trpdata\_dec\_2022 <- read\_csv("F:/Data\_Sci/Cap\_Stone\_Project/Cyclist\_trip\_data/202212-divvy-tripdata/202212-divvy-tripdata.csv")  
  
trpdata\_jan\_2023 <- read\_csv("F:/Data\_Sci/Cap\_Stone\_Project/Cyclist\_trip\_data/202301-divvy-tripdata/202301-divvy-tripdata.csv")  
  
trpdata\_feb\_2023 <- read\_csv("F:/Data\_Sci/Cap\_Stone\_Project/Cyclist\_trip\_data/202302-divvy-tripdata/202302-divvy-tripdata.csv")  
  
trpdata\_mar\_2023 <- read\_csv("F:/Data\_Sci/Cap\_Stone\_Project/Cyclist\_trip\_data/202303-divvy-tripdata/202303-divvy-tripdata.csv")  
  
trpdata\_apr\_2023 <- read\_csv("F:/Data\_Sci/Cap\_Stone\_Project/Cyclist\_trip\_data/202304-divvy-tripdata/202304-divvy-tripdata.csv")  
  
trpdata\_may\_2023 <- read\_csv("F:/Data\_Sci/Cap\_Stone\_Project/Cyclist\_trip\_data/202305-divvy-tripdata/202305-divvy-tripdata.csv")  
  
trpdata\_june\_2023 <- read\_csv("F:/Data\_Sci/Cap\_Stone\_Project/Cyclist\_trip\_data/202306-divvy-tripdata/202306-divvy-tripdata.csv")

### Combining all the monthly data to one previous year data(data\_prev\_year).

data\_prev\_year <- rbind(trpdata\_july\_2022, trpdata\_aug\_2022,  
 trpdata\_sept\_2022, trpdata\_oct\_2022,  
 trpdata\_nov\_2022, trpdata\_dec\_2022,  
 trpdata\_jan\_2023, trpdata\_feb\_2023,  
 trpdata\_mar\_2023, trpdata\_apr\_2023,  
 trpdata\_may\_2023, trpdata\_june\_2023)  
  
glimpse(data\_prev\_year)

Rows: 5,779,444  
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", "…

* Checking and counting “NA” in each column of the dataframe.

na\_in\_cols <- data\_prev\_year %>% map(is.na) %>% map(sum) %>% unlist()  
  
na\_in\_cols

ride\_id rideable\_type started\_at ended\_at   
 0 0 0 0   
start\_station\_name start\_station\_id end\_station\_name end\_station\_id   
 857860 857992 915655 915796   
 start\_lat start\_lng end\_lat end\_lng   
 0 0 5795 5795   
 member\_casual   
 0

* Finding the length of rides taken by riders by making a new column ride\_length in minutes. Eliminating stations where station names and longitude and latitude co-ordinates are not present.

data\_prev\_year <- data\_prev\_year %>%   
 mutate(ride\_length = difftime(ended\_at, started\_at,  
 units = "min")) %>%  
 mutate(ride\_length = as.numeric(ride\_length)) %>%   
 mutate(ride\_length = if\_else(ride\_length < 0, 0, ride\_length)) %>%   
 filter(start\_station\_name != "" & end\_station\_name != "" &   
 !is.na(start\_lat) & !is.na(start\_lng) &  
 !is.na(end\_lat) & !is.na(end\_lng)) %>% arrange(ride\_length)  
  
  
glimpse(data\_prev\_year)

Rows: 4,409,335  
Columns: 14  
$ ride\_id <chr> "86CD09DA24761714", "27024CD08288BD45", "029D853B5C…  
$ rideable\_type <chr> "electric\_bike", "electric\_bike", "classic\_bike", "…  
$ started\_at <dttm> 2022-07-20 16:21:48, 2022-07-30 23:42:46, 2022-07-…  
$ ended\_at <dttm> 2022-07-20 16:21:48, 2022-07-30 23:42:46, 2022-07-…  
$ start\_station\_name <chr> "Racine Ave & Fullerton Ave", "Albany Ave & 26th St…  
$ start\_station\_id <chr> "TA1306000026", "15691", "chargingstx5", "chargings…  
$ end\_station\_name <chr> "Racine Ave & Fullerton Ave", "Albany Ave & 26th St…  
$ end\_station\_id <chr> "TA1306000026", "15691", "chargingstx5", "chargings…  
$ start\_lat <dbl> 41.92556, 41.84452, 41.94335, 41.94335, 41.94335, 4…  
$ start\_lng <dbl> -87.65859, -87.70209, -87.67067, -87.67067, -87.670…  
$ end\_lat <dbl> 41.92556, 41.84448, 41.94335, 41.94335, 41.94335, 4…  
$ end\_lng <dbl> -87.65840, -87.70201, -87.67067, -87.67067, -87.670…  
$ member\_casual <chr> "member", "casual", "member", "member", "casual", "…  
$ ride\_length <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, …

* A total of sum(data\_prev\_year$ride\_length) minutes were ridden by both casual and membership holders.
* Aggregating data to see **“Average minutes per ride”** grouped by “bike type” and “rider type” after removing rides less than 2 minutes (As rides less than 2 minutes tend to have the same start and stop stations.).

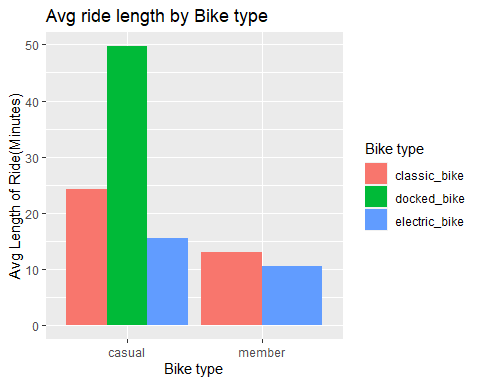
data\_prev\_year\_aggregate <- data\_prev\_year%>%   
 select(ride\_id, rideable\_type, member\_casual, started\_at, ended\_at,  
 ride\_length, everything()) %>%  
 filter(ride\_length >= 2) %>%   
 summarise("Number of Rides" = n(),   
 "Ride Length" = sum(ride\_length, na.rm = TRUE),  
 "Avg Ride Length in Minutes" = mean(ride\_length),  
 .by = c(member\_casual, rideable\_type)) %>%   
 arrange(desc("Avg Ride Length in Minutes")) %>%   
 gt() %>% tab\_header(title = "Average length of Rides") %>%   
 cols\_label(member\_casual = "Rider type",  
 rideable\_type = "Bike type")  
  
data\_prev\_year\_aggregate

Table : Average length of Rides

| Rider type | Bike type | Number of Rides | Ride Length | Avg Ride Length in Minutes |
| --- | --- | --- | --- | --- |
| member | classic\_bike | 1630991 | 21996488 | 13.48658 |
| casual | classic\_bike | 781530 | 19383358 | 24.80181 |
| casual | electric\_bike | 709649 | 11372659 | 16.02575 |
| member | electric\_bike | 984688 | 10968684 | 11.13925 |
| casual | docked\_bike | 136794 | 6899998 | 50.44079 |

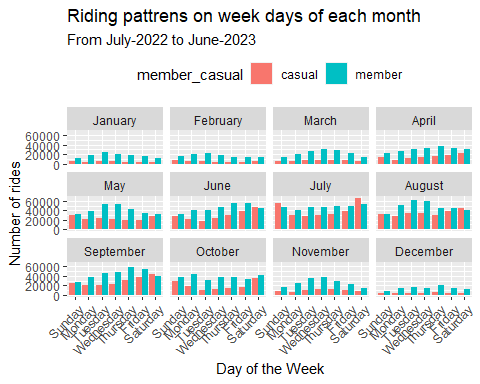
* Calculating and visualizing “Average ride length” by “Rider type”.

average\_ride\_by\_rideable\_type <- data\_prev\_year %>%  
 rename("Rider type" = member\_casual, "Bike type" = rideable\_type) %>%   
 summarise(ride\_length = sum(ride\_length, na.rm = TRUE),  
 ride\_count = n(),  
 avg\_ride\_length = ride\_length/ride\_count,  
 .by = c(`Rider type`, `Bike type`)) %>%   
 ggplot(aes(`Rider type`, avg\_ride\_length)) +   
 geom\_col(aes(fill = `Bike type`), position = "dodge") +   
 labs(x = "Bike type", y = "Avg Length of Ride(Minutes)",  
 title = "Avg ride length by Bike type")  
  
average\_ride\_by\_rideable\_type



* Calculating and visualizing ride patterns in a week.

rideable\_order <- c("classic\_bike", "electric\_bike", "docked\_bike")  
  
rides\_on\_days <- data\_prev\_year %>%  
 filter(rideable\_type != "docked\_bike") %>%  
 mutate(month = month(started\_at, label = TRUE, abbr = FALSE)) %>%   
 mutate(rideable\_type = factor(rideable\_type, levels = rideable\_order)) %>% ggplot(aes(wday(started\_at, label = TRUE, abbr = FALSE))) +   
 geom\_bar(aes(fill = member\_casual), position = "dodge") +  
 facet\_wrap(~month, nrow = 3) +   
 labs(x = "Day of the Week", y = "Number of rides",  
 title = "Riding pattrens on week days of each month",  
 subtitle = "From July-2022 to June-2023") +  
 theme(legend.position = "top",  
 axis.text.x = element\_text(angle = 45, hjust = 1))  
  
rides\_on\_days



Removing “NA” from the dataframe and blanks.

data\_prev\_year <- data\_prev\_year %>%  
 drop\_na(start\_station\_name) %>%   
 drop\_na(end\_station\_name) %>%   
 filter(start\_station\_name != "" & end\_station\_name != "",  
 started\_at != ended\_at)   
  
glimpse(data\_prev\_year)

Rows: 4,409,072  
Columns: 14  
$ ride\_id <chr> "029D853B5C38426E", "C1D6D749139CB6C0", "D3E7C0B68E…  
$ rideable\_type <chr> "classic\_bike", "classic\_bike", "classic\_bike", "cl…  
$ started\_at <dttm> 2022-07-26 20:07:33, 2022-07-26 20:08:04, 2022-07-…  
$ ended\_at <dttm> 2022-07-26 19:59:34, 2022-07-26 19:59:34, 2022-07-…  
$ start\_station\_name <chr> "Lincoln Ave & Roscoe St\*", "Lincoln Ave & Roscoe S…  
$ start\_station\_id <chr> "chargingstx5", "chargingstx5", "chargingstx5", "ch…  
$ end\_station\_name <chr> "Lincoln Ave & Roscoe St\*", "Lincoln Ave & Roscoe S…  
$ end\_station\_id <chr> "chargingstx5", "chargingstx5", "chargingstx5", "ch…  
$ start\_lat <dbl> 41.94335, 41.94335, 41.94335, 41.94335, 41.93945, 4…  
$ start\_lng <dbl> -87.67067, -87.67067, -87.67067, -87.67067, -87.663…  
$ end\_lat <dbl> 41.94335, 41.94335, 41.94335, 41.94335, 41.93948, 4…  
$ end\_lng <dbl> -87.67067, -87.67067, -87.67067, -87.67067, -87.663…  
$ member\_casual <chr> "member", "member", "casual", "casual", "member", "…  
$ ride\_length <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, …

* Making a new column to identify travelled stations.

data\_prev\_year <- data\_prev\_year %>%   
 mutate(stations\_travelled = paste(start\_station\_name,   
 "-", end\_station\_name))  
  
glimpse(data\_prev\_year)

Rows: 4,409,072  
Columns: 15  
$ ride\_id <chr> "029D853B5C38426E", "C1D6D749139CB6C0", "D3E7C0B68E…  
$ rideable\_type <chr> "classic\_bike", "classic\_bike", "classic\_bike", "cl…  
$ started\_at <dttm> 2022-07-26 20:07:33, 2022-07-26 20:08:04, 2022-07-…  
$ ended\_at <dttm> 2022-07-26 19:59:34, 2022-07-26 19:59:34, 2022-07-…  
$ start\_station\_name <chr> "Lincoln Ave & Roscoe St\*", "Lincoln Ave & Roscoe S…  
$ start\_station\_id <chr> "chargingstx5", "chargingstx5", "chargingstx5", "ch…  
$ end\_station\_name <chr> "Lincoln Ave & Roscoe St\*", "Lincoln Ave & Roscoe S…  
$ end\_station\_id <chr> "chargingstx5", "chargingstx5", "chargingstx5", "ch…  
$ start\_lat <dbl> 41.94335, 41.94335, 41.94335, 41.94335, 41.93945, 4…  
$ start\_lng <dbl> -87.67067, -87.67067, -87.67067, -87.67067, -87.663…  
$ end\_lat <dbl> 41.94335, 41.94335, 41.94335, 41.94335, 41.93948, 4…  
$ end\_lng <dbl> -87.67067, -87.67067, -87.67067, -87.67067, -87.663…  
$ member\_casual <chr> "member", "member", "casual", "casual", "member", "…  
$ ride\_length <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, …  
$ stations\_travelled <chr> "Lincoln Ave & Roscoe St\* - Lincoln Ave & Roscoe St…

* Finding which route is traveled most by **casual riders**.

most\_travelled\_routes\_casual <- data\_prev\_year %>%  
 filter(member\_casual == "casual") %>%   
 summarise(ride\_count = n(),  
 avg\_ride\_length = round(mean(ride\_length), 2),  
 .by = c(stations\_travelled)) %>%  
 arrange(desc(ride\_count))  
  
head(most\_travelled\_routes\_casual)

# A tibble: 6 × 3  
 stations\_travelled ride\_count avg\_ride\_length  
 <chr> <int> <dbl>  
1 Streeter Dr & Grand Ave - Streeter Dr & Grand Ave 9698 39.6  
2 DuSable Lake Shore Dr & Monroe St - DuSable Lake S… 6584 33.4  
3 DuSable Lake Shore Dr & Monroe St - Streeter Dr & … 4840 27.1  
4 Michigan Ave & Oak St - Michigan Ave & Oak St 4292 44.6  
5 Millennium Park - Millennium Park 3884 37.4  
6 Montrose Harbor - Montrose Harbor 2711 48.3

NROW(most\_travelled\_routes\_casual)

[1] 130660

# A tibble: 6 × 4  
 stations\_travelled ride\_count total\_ride\_length ride\_length  
 <chr> <int> <dbl> <dbl>  
1 Ellis Ave & 60th St - University Ave… 6153 25936. 4.22  
2 University Ave & 57th St - Ellis Ave… 5786 26634. 4.6   
3 Ellis Ave & 60th St - Ellis Ave & 55… 5676 28427. 5.01  
4 Ellis Ave & 55th St - Ellis Ave & 60… 5347 27187. 5.08  
5 State St & 33rd St - Calumet Ave & 3… 4156 18014. 4.33  
6 Calumet Ave & 33rd St - State St & 3… 4027 15887. 3.95

[1] 145104

* Finding which station has most ride starting points and which station has most ending points.

most\_starting\_points <- data\_prev\_year %>%   
 summarise(ride\_count = n(),  
 .by = start\_station\_name) %>%  
 select(start\_station\_name, ride\_count) %>%  
 slice\_max(ride\_count, n = 10)  
  
most\_starting\_points

# A tibble: 10 × 2  
 start\_station\_name ride\_count  
 <chr> <int>  
 1 Streeter Dr & Grand Ave 65892  
 2 DuSable Lake Shore Dr & Monroe St 37939  
 3 Michigan Ave & Oak St 36036  
 4 DuSable Lake Shore Dr & North Blvd 35091  
 5 Wells St & Concord Ln 33250  
 6 Clark St & Elm St 32751  
 7 Kingsbury St & Kinzie St 31876  
 8 Millennium Park 30917  
 9 Theater on the Lake 29600  
10 Wells St & Elm St 28063

most\_starting\_points$ride\_count %>% sum()

[1] 361415

most\_ending\_points <- data\_prev\_year %>%   
 summarise(ride\_count = n(),  
 .by = end\_station\_name) %>%  
 select(end\_station\_name, ride\_count) %>%   
 slice\_max(ride\_count, n = 10)  
  
most\_ending\_points

# A tibble: 10 × 2  
 end\_station\_name ride\_count  
 <chr> <int>  
 1 Streeter Dr & Grand Ave 67536  
 2 DuSable Lake Shore Dr & North Blvd 38026  
 3 Michigan Ave & Oak St 36976  
 4 DuSable Lake Shore Dr & Monroe St 36806  
 5 Wells St & Concord Ln 33814  
 6 Clark St & Elm St 32325  
 7 Millennium Park 32046  
 8 Kingsbury St & Kinzie St 31058  
 9 Theater on the Lake 30214  
10 Wells St & Elm St 28212

most\_ending\_points$ride\_count %>% sum()

[1] 367013

* Finding all the stations and number of total unique stations.

unique\_start\_stations\_name <- data\_prev\_year %>%  
 filter(start\_station\_name != "") %>%   
 distinct(start\_station\_name)  
  
unique\_end\_station\_name <- data\_prev\_year %>%  
 filter(end\_station\_name != "") %>%   
 distinct(end\_station\_name)  
  
unique\_stations <-   
 union(unique\_end\_station\_name$end\_station\_name,  
 unique\_start\_stations\_name$start\_station\_name)  
  
head(unique\_stations)

[1] "Lincoln Ave & Roscoe St\*" "Southport Ave & Belmont Ave"  
[3] "Leavitt St & Chicago Ave" "Clark St & Armitage Ave"   
[5] "Streeter Dr & Grand Ave" "Clark St & Montrose Ave"

NROW(unique\_stations)

[1] 1791