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
```

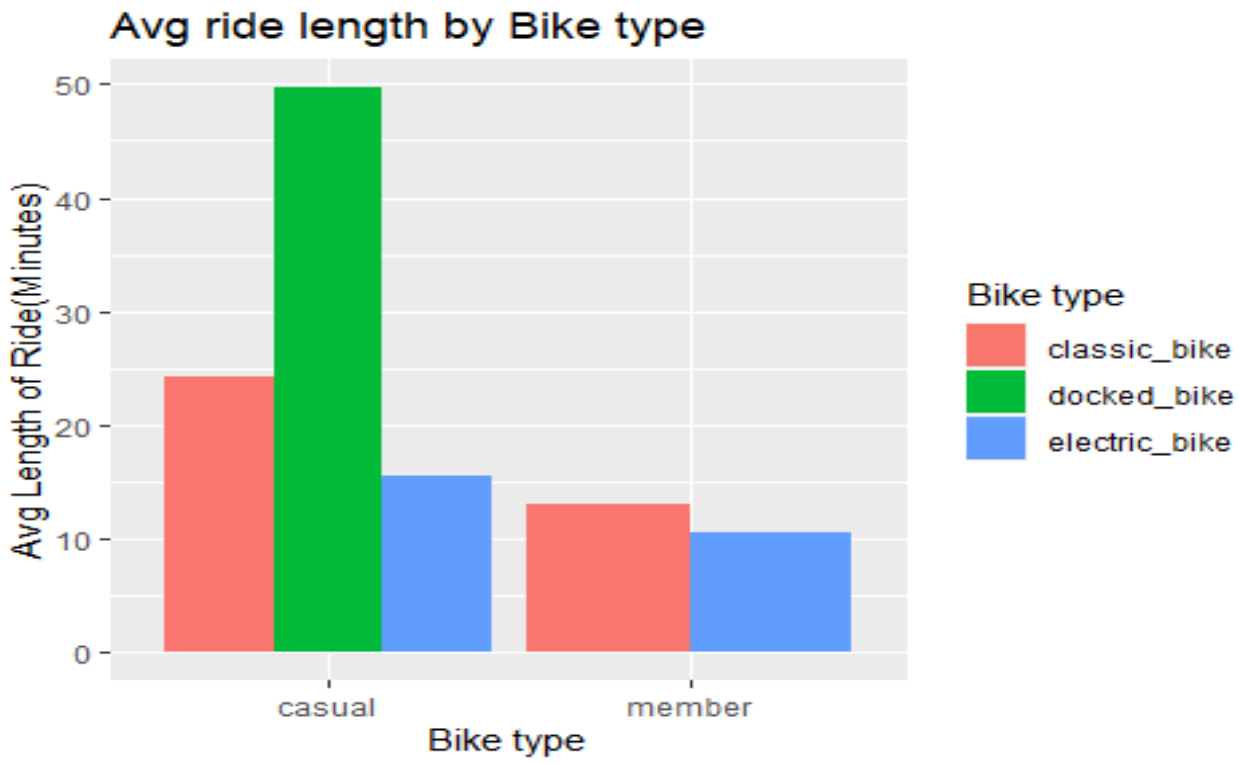
Table 1: Average length of Rides

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
```

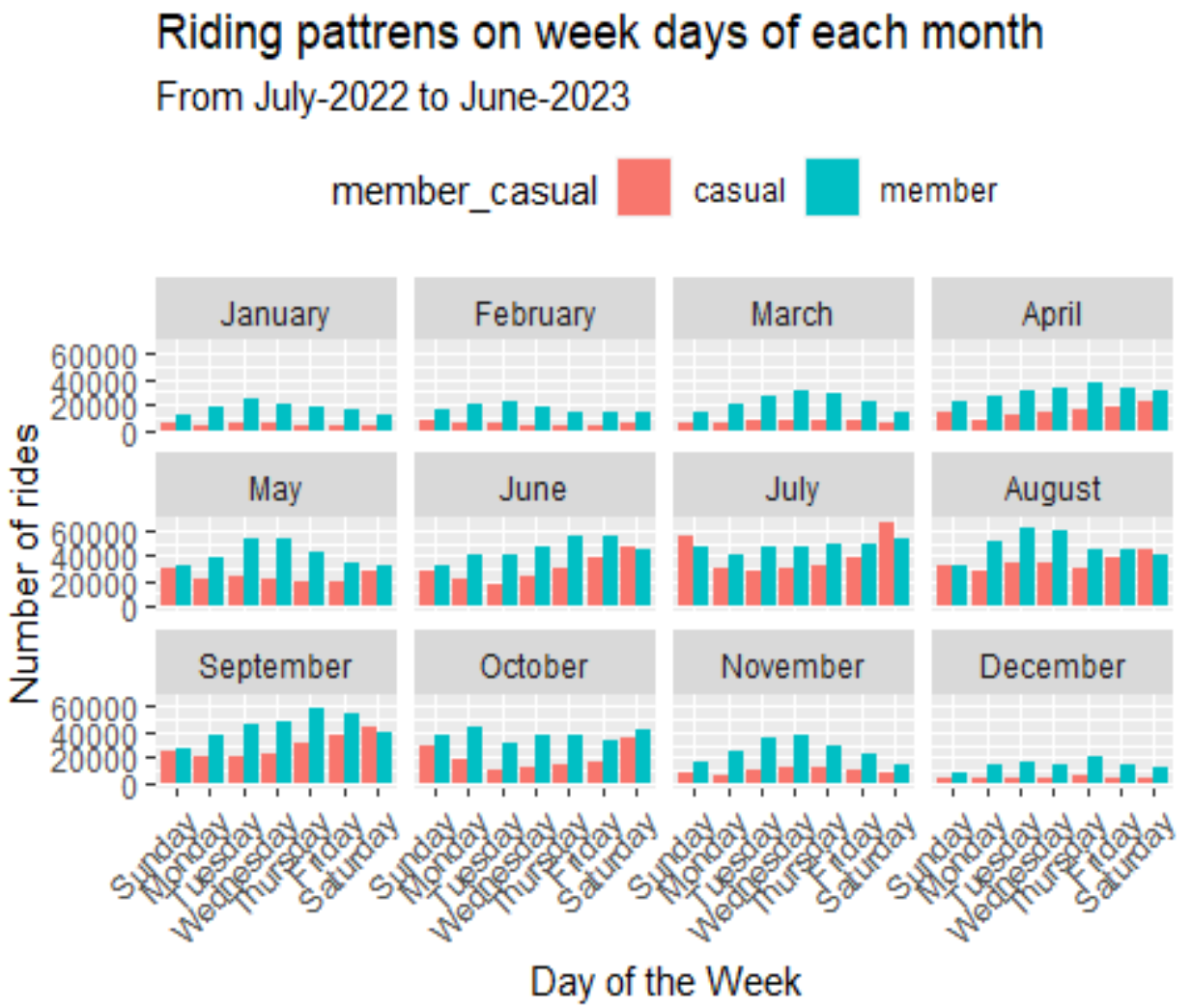


- Calculating and visualizing ride patterns in a week.

```
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,
…
```

- 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,
…
$ 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
```