Capstone

Wala Faris

2025-01-12

Google Data Analytics Capstone: Case Study- Cyclistic, Sharing Bike

This repository showcases the final project undertaken as part of the Google Data Analytics Professional Certificate program, I enrolled for the 1st course in Aug 10th, 2024 but the actual work was completed between Oct 6th, 2024 and Jan 4th, 2025. The project delves into an analysis of Cyclistic bike-sharing data with the aim of identifying key trends and actionable insights that can inform business decisions for the fictional company.

The analysis encompasses several crucial stages:

- Data Cleaning: Ensuring data accuracy and consistency through thorough cleaning and preparation.
- Exploratory Data Analysis (EDA): Conducting in-depth exploration of the dataset to uncover patterns, trends, and relationships within the data.
- Data Visualization: Creating meaningful and insightful visualizations to effectively communicate findings and key trends to stakeholders.
- Actionable Insights Development: Translating data analysis findings into actionable recommendations for Cyclistic to enhance its business operations and improve customer experience.
- Recommendations: Depends of the insight I developed, I recommended some steps and procedures should the company take to convince the casual riders to convert to an annual member with company.

Introduction

The case study on “Cyclistic, sharing bike” a fictional bikeshare firm located in Chicago. The company owns and operates over 5000 bicycles that are distributed and locked into a network of over 600 stations across Chicago. The company serves two types of customers, who purchase single-ride passes or full-day passes (The casual riders), and who purchase annual memberships (The annual members). The company aims to increase profitability and they found that the annual members are much more profitable than casual rider, so the marketing director was interested in maximizing the number of annual members by creating marketing strategies that aid in the conversion of more casual riders to annual members.to achieve this, I analyzed the historical bike trip data to understand the differences between annual members and casual riders.

My responsibility was to make data-driven recommendations for the marketing campaign by highlighting the differ between the way that the two types of customers behave and the goal is to identify factors that influence membership decisions and develop targeted

marketing strategies leveraging digital media to encourage casual riders to become annual members.

Statement of the Business Task

To highlight the difference between the annual members and casual riders while using the Cyclistic sharing bikes.

The Dataset

I used the "Divvy Bikeshare Dataset," which is owned by the "Divvy" bike sharing company. The data is a third-party dataset made public by Motivate International Inc., the firm that runs the Divvy bike-sharing service. The link to the dataset: https://divvy-tripdata.s3.amazonaws.com/index.html Data is available from the Apr 2020 till Nov 2024. The dataset is structured in the form of spreadsheet. **Issues with the dataset:** There is missing value in the data and some of the column names are inconsistent.

Installing and Loading Packages

I used various packages in R Studio, such as, **Tidyverse, Lubridate** (for datetime functions), **Tidyr** (for data-cleaning), **ggplot2** (for creating visualizations) and many others.

```
options(repos = c(CRAN = "https://cloud.r-project.org"))
install.packages("tidyverse")
```

```
## Installing package into 'C:/Users/yolai/AppData/Local/R/win-library/4.4'
## (as 'lib' is unspecified)
```

```
## package 'tidyverse' successfully unpacked and MD5 sums checked
##
## The downloaded binary packages are in
##  C:\Users\yolai\AppData\Local\Temp\Rtmpc1bsOM\downloaded_packages
```

```
library(tidyverse)
```

```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse
2.0.0 ──
## ✓ dplyr     1.1.4     ✓ readr     2.1.5
## ✓ forcats   1.0.0     ✓ stringr   1.5.1
## ✓ ggplot2   3.5.1     ✓ tibble    3.2.1
## ✓ lubridate 1.9.4     ✓ tidyr     1.3.1
## ✓ 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(readr)
library(dplyr)
library(tidyr)
library(lubridate)
library(ggplot2)
```

Getting my Data Ready in R Studio.

After downloading the all data files I start importing them by using the **read_csv()** function from the **readr** package.

```
apr_data_2020 <- read_csv("C:/Users/yolai/Downloads/202004-divvy-tripdata.csv")
```

```
## Rows: 84776 Columns: 13
## ── Column specification ────────────────────────────────────────────────────────
## Delimiter: ","
## chr  (5): ride_id, rideable_type, start_station_name, end_station_name, memb...
## dbl  (6): start_station_id, end_station_id, start_lat, start_lng, end_lat, e...
## 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.
```

```
may_data_2020 <- read_csv("C:/Users/yolai/Downloads/202005-divvy-tripdata.csv")
```

```
## Rows: 200274 Columns: 13
## ── Column specification ────────────────────────────────────────────────────────
## Delimiter: ","
## chr  (5): ride_id, rideable_type, start_station_name, end_station_name, memb...
## dbl  (6): start_station_id, end_station_id, start_lat, start_lng, end_lat, e...
## 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.
```

```
jun_data_2020 <- read_csv("C:/Users/yolai/Downloads/202006-divvy-tripdata.csv")
```

```
## Rows: 343005 Columns: 13
## ── Column specification ────────────────────────────────────────────────────────
```

```
## Delimiter: ","
## chr  (5): ride_id, rideable_type, start_station_name, end_station_name,
memb...
## dbl  (6): start_station_id, end_station_id, start_lat, start_lng, end_lat,
e...
## 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.
```

```
jul_data_2020 <- read_csv("C:/Users/yolai/Downloads/202007-divvy-
tripdata.csv")
```

```
## Rows: 551480 Columns: 13
## -- Column specification
--------------------------------------------------------
## Delimiter: ","
## chr  (5): ride_id, rideable_type, start_station_name, end_station_name,
memb...
## dbl  (6): start_station_id, end_station_id, start_lat, start_lng, end_lat,
e...
## 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.
```

```
aug_data_2020 <- read_csv("C:/Users/yolai/Downloads/202008-divvy-
tripdata.csv")
```

```
## Rows: 622361 Columns: 13
## -- Column specification
--------------------------------------------------------
## Delimiter: ","
## chr  (5): ride_id, rideable_type, start_station_name, end_station_name,
memb...
## dbl  (6): start_station_id, end_station_id, start_lat, start_lng, end_lat,
e...
## 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.
```

```
sep_data_2020 <- read_csv("C:/Users/yolai/Downloads/202009-divvy-
tripdata.csv")
```

```
## Rows: 532958 Columns: 13
## -- Column specification
--------------------------------------------------------
```

```
## Delimiter: ","
## chr  (5): ride_id, rideable_type, start_station_name, end_station_name,
memb...
## dbl  (6): start_station_id, end_station_id, start_lat, start_lng, end_lat,
e...
## 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.
```

```
oct_data_2020 <- read_csv("C:/Users/yolai/Downloads/202010-divvy-
tripdata.csv")
```

```
## Rows: 388653 Columns: 13
## -- Column specification
--------------------------------------------------------
## Delimiter: ","
## chr  (5): ride_id, rideable_type, start_station_name, end_station_name,
memb...
## dbl  (6): start_station_id, end_station_id, start_lat, start_lng, end_lat,
e...
## 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.
```

```
nov_data_2020 <- read_csv("C:/Users/yolai/Downloads/202011-divvy-
tripdata.csv")
```

```
## Rows: 259716 Columns: 13
## -- Column specification
--------------------------------------------------------
## Delimiter: ","
## chr  (5): ride_id, rideable_type, start_station_name, end_station_name,
memb...
## dbl  (6): start_station_id, end_station_id, start_lat, start_lng, end_lat,
e...
## 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.
```

```
dec_data_2020 <- read_csv("C:/Users/yolai/Downloads/202012-divvy-
tripdata.csv")
```

```
## Rows: 131573 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.
```

```
jan_data_2021 <- read_csv("C:/Users/yolai/Downloads/202101-divvy-
tripdata.csv")
```

```
## Rows: 96834 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.
```

```
feb_data_2021 <- read_csv("C:/Users/yolai/Downloads/202102-divvy-
tripdata.csv")
```

```
## Rows: 49622 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.
```

```
mar_data_2021 <- read_csv("C:/Users/yolai/Downloads/202103-divvy-
tripdata.csv")
```

```
## Rows: 228496 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.
```

```
apr_data_2021 <- read_csv("C:/Users/yolai/Downloads/202104-divvy-
tripdata.csv")
```

```
## Rows: 337230 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.
```

```
may_data_2021 <- read_csv("C:/Users/yolai/Downloads/202105-divvy-
tripdata.csv")
```

```
## Rows: 531633 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.
```

```
jun_data_2021 <- read_csv("C:/Users/yolai/Downloads/202106-divvy-
tripdata.csv")
```

```
## Rows: 729595 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.

jul_data_2021 <- read_csv("C:/Users/yolai/Downloads/202107-divvy-
tripdata.csv")

## Rows: 822410 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.

aug_data_2021 <- read_csv("C:/Users/yolai/Downloads/202108-divvy-
tripdata.csv")

## Rows: 804352 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.

sep_data_2021 <- read_csv("C:/Users/yolai/Downloads/202109-divvy-
tripdata.csv")

## Rows: 756147 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.
```

```
oct_data_2021 <- read_csv("C:/Users/yolai/Downloads/202110-divvy-
tripdata.csv")
```

```
## Rows: 631226 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.
```

```
nov_data_2021 <- read_csv("C:/Users/yolai/Downloads/202111-divvy-
tripdata.csv")
```

```
## Rows: 359978 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.
```

```
dec_data_2021 <- read_csv("C:/Users/yolai/Downloads/202112-divvy-
tripdata.csv")
```

```
## Rows: 247540 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.
```

```
jan_data_2022 <- read_csv("C:/Users/yolai/Downloads/202201-divvy-
tripdata.csv")
```

```
## Rows: 103770 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.
```

```
feb_data_2022 <- read_csv("C:/Users/yolai/Downloads/202202-divvy-
tripdata.csv")
```

```
## Rows: 115609 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.
```

```
mar_data_2022 <- read_csv("C:/Users/yolai/Downloads/202203-divvy-
tripdata.csv")
```

```
## Rows: 284042 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.
```

```
apr_data_2022 <- read_csv("C:/Users/yolai/Downloads/202204-divvy-
tripdata.csv")
```

```
## Rows: 371249 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.
```

```
may_data_2022 <- read_csv("C:/Users/yolai/Downloads/202205-divvy-
tripdata.csv")
```

```
## Rows: 634858 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.
```

```
jun_data_2022 <- read_csv("C:/Users/yolai/Downloads/202206-divvy-
tripdata.csv")
```

```
## Rows: 769204 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.
```

```
jul_data_2022 <- read_csv("C:/Users/yolai/Downloads/202207-divvy-
tripdata.csv")
```

```
## Rows: 823488 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.
```

```
aug_data_2022 <- read_csv("C:/Users/yolai/Downloads/202208-divvy-
tripdata.csv")
```

```
## Rows: 785932 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.
```

```
sep_data_2022 <- read_csv("C:/Users/yolai/Downloads/202209-divvy-
tripdata.csv")
```

```
## Rows: 701339 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.
```

```
oct_data_2022 <- read_csv("C:/Users/yolai/Downloads/202210-divvy-
tripdata.csv")
```

```
## 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.
```

```
nov_data_2022 <- read_csv("C:/Users/yolai/Downloads/202211-divvy-
tripdata.csv")
```

```
## 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.
```

```
dec_data_2022 <- read_csv("C:/Users/yolai/Downloads/202212-divvy-
tripdata.csv")
```

```
## 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.
```

```
jan_data_2023 <- read_csv("C:/Users/yolai/Downloads/202301-divvy-
tripdata.csv")
```

```
## 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.
```

```
feb_data_2023 <- read_csv("C:/Users/yolai/Downloads/202302-divvy-
tripdata.csv")
```

```
## 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.
```

```
mar_data_2023 <- read_csv("C:/Users/yolai/Downloads/202303-divvy-
tripdata.csv")
```

```
## 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.
```

```
apr_data_2023 <- read_csv("C:/Users/yolai/Downloads/202304-divvy-
tripdata.csv")
```

```
## 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.
```

```
may_data_2023 <- read_csv("C:/Users/yolai/Downloads/202305-divvy-
tripdata.csv")
```

```
## 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.
```

```
jun_data_2023 <- read_csv("C:/Users/yolai/Downloads/202306-divvy-
tripdata.csv")
```

```
## 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.
```

```
jul_data_2023 <- read_csv("C:/Users/yolai/Downloads/202307-divvy-
tripdata.csv")
```

```
## 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.
```

```
aug_data_2023 <- read_csv("C:/Users/yolai/Downloads/202308-divvy-
tripdata.csv")
```

```
## 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.
```

```
sep_data_2023 <- read_csv("C:/Users/yolai/Downloads/202309-divvy-
tripdata.csv")
```

```
## 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.
```

```
oct_data_2023 <- read_csv("C:/Users/yolai/Downloads/202310-divvy-
tripdata.csv")
```

```
## Rows: 537113 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.
```

```
nov_data_2023 <- read_csv("C:/Users/yolai/Downloads/202311-divvy-
tripdata.csv")
```

```
## Rows: 362518 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.
```

```
dec_data_2023 <- read_csv("C:/Users/yolai/Downloads/202312-divvy-
tripdata.csv")
```

```
## Rows: 224073 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.
```

```
jan_data_2024 <- read_csv("C:/Users/yolai/Downloads/202401-divvy-
tripdata.csv")
```

```
## Rows: 144873 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.
```

```
feb_data_2024 <- read_csv("C:/Users/yolai/Downloads/202402-divvy-
tripdata.csv")
```

```
## Rows: 223164 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.

mar_data_2024 <- read_csv("C:/Users/yolai/Downloads/202403-divvy-
tripdata.csv")

## Rows: 301687 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.

apr_data_2024 <- read_csv("C:/Users/yolai/Downloads/202404-divvy-
tripdata.csv")

## Rows: 415025 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.

may_data_2024 <- read_csv("C:/Users/yolai/Downloads/202405-divvy-
tripdata.csv")

## Rows: 609493 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.
```

```
jun_data_2024 <- read_csv("C:/Users/yolai/Downloads/202406-divvy-
tripdata.csv")

## Rows: 710721 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.

jul_data_2024 <- read_csv("C:/Users/yolai/Downloads/202407-divvy-
tripdata.csv")

## Rows: 748962 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.

aug_data_2024 <- read_csv("C:/Users/yolai/Downloads/202408-divvy-
tripdata.csv")

## Rows: 755639 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.

sep_data_2024 <- read_csv("C:/Users/yolai/Downloads/202409-divvy-
tripdata.csv")
```

```
## Rows: 821276 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.
```

```
oct_data_2024 <- read_csv("C:/Users/yolai/Downloads/202410-divvy-
tripdata.csv")
```

```
## Rows: 616281 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.
```

```
nov_data_2024 <- read_csv("C:/Users/yolai/Downloads/202411-divvy-
tripdata.csv")
```

```
## Rows: 335075 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.
```

Data Exploration to Assuring the Datatypes Consistency

I checked the files one by one using the **str()** function to verify if there is un consistency in the dataset, So I found that the columns **start_station_id** and **end_station_id** are double

datatype in the datasets starting from Apr 2020 - Nov 2020 while these fields are character datatype in the rest of data from Dec 2020 - Nov 2024.

```
str(apr_data_2020)
```

```
## spc_tbl_ [84,776 × 13] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ ride_id           : chr [1:84776] "A847FADBBC638E45" "5405B80E996FF60D"
"5DD24A79A4E006F4" "2A59BBDF5CDBA725" ...
##  $ rideable_type     : chr [1:84776] "docked_bike" "docked_bike"
"docked_bike" "docked_bike" ...
##  $ started_at        : POSIXct[1:84776], format: "2020-04-26 17:45:14"
"2020-04-17 17:08:54" ...
##  $ ended_at          : POSIXct[1:84776], format: "2020-04-26 18:12:03"
"2020-04-17 17:17:03" ...
##  $ start_station_name: chr [1:84776] "Eckhart Park" "Drake Ave & Fullerton
Ave" "McClurg Ct & Erie St" "California Ave & Division St" ...
##  $ start_station_id  : num [1:84776] 86 503 142 216 125 173 35 434 627 377
...
##  $ end_station_name  : chr [1:84776] "Lincoln Ave & Diversey Pkwy"
"Kosciuszko Park" "Indiana Ave & Roosevelt Rd" "Wood St & Augusta Blvd" ...
##  $ end_station_id    : num [1:84776] 152 499 255 657 323 35 635 382 359
508 ...
##  $ start_lat         : num [1:84776] 41.9 41.9 41.9 41.9 41.9 ...
##  $ start_lng         : num [1:84776] -87.7 -87.7 -87.6 -87.7 -87.6 ...
##  $ end_lat           : num [1:84776] 41.9 41.9 41.9 41.9 42 ...
##  $ end_lng           : num [1:84776] -87.7 -87.7 -87.6 -87.7 -87.7 ...
##  $ member_casual     : chr [1:84776] "member" "member" "member" "member"
...
##  - attr(*, "spec")=
##   .. cols(
##   ..   ride_id = col_character(),
##   ..   rideable_type = col_character(),
##   ..   started_at = col_datetime(format = ""),
##   ..   ended_at = col_datetime(format = ""),
##   ..   start_station_name = col_character(),
##   ..   start_station_id = col_double(),
##   ..   end_station_name = col_character(),
##   ..   end_station_id = col_double(),
##   ..   start_lat = col_double(),
##   ..   start_lng = col_double(),
##   ..   end_lat = col_double(),
##   ..   end_lng = col_double(),
##   ..   member_casual = col_character()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

```
str(may_data_2020)
```

```
## spc_tbl_ [200,274 × 13] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ ride_id           : chr [1:200274] "02668AD35674B983"
"7A50CCAF1EDDB28F" "2FFCDFDB91FE9A52" "58991CF1DB75BA84" ...
```

```
##  $ rideable_type     : chr [1:200274] "docked_bike" "docked_bike"
"docked_bike" "docked_bike" ...
##  $ started_at        : POSIXct[1:200274], format: "2020-05-27 10:03:52"
"2020-05-25 10:47:11" ...
##  $ ended_at          : POSIXct[1:200274], format: "2020-05-27 10:16:49"
"2020-05-25 11:05:40" ...
##  $ start_station_name: chr [1:200274] "Franklin St & Jackson Blvd" "Clark
St & Wrightwood Ave" "Kedzie Ave & Milwaukee Ave" "Clarendon Ave & Leland
Ave" ...
##  $ start_station_id  : num [1:200274] 36 340 260 251 261 206 261 180 331
219 ...
##  $ end_station_name  : chr [1:200274] "Wabash Ave & Grand Ave" "Clark St &
Leland Ave" "Kedzie Ave & Milwaukee Ave" "Lake Shore Dr & Wellington Ave" ...
##  $ end_station_id    : num [1:200274] 199 326 260 157 206 22 261 180 300
305 ...
##  $ start_lat         : num [1:200274] 41.9 41.9 41.9 42 41.9 ...
##  $ start_lng         : num [1:200274] -87.6 -87.6 -87.7 -87.7 -87.7 ...
##  $ end_lat           : num [1:200274] 41.9 42 41.9 41.9 41.8 ...
##  $ end_lng           : num [1:200274] -87.6 -87.7 -87.7 -87.6 -87.6 ...
##  $ member_casual     : chr [1:200274] "member" "casual" "casual" "casual"
...
##  - attr(*, "spec")=
##   .. cols(
##   ..   ride_id = col_character(),
##   ..   rideable_type = col_character(),
##   ..   started_at = col_datetime(format = ""),
##   ..   ended_at = col_datetime(format = ""),
##   ..   start_station_name = col_character(),
##   ..   start_station_id = col_double(),
##   ..   end_station_name = col_character(),
##   ..   end_station_id = col_double(),
##   ..   start_lat = col_double(),
##   ..   start_lng = col_double(),
##   ..   end_lat = col_double(),
##   ..   end_lng = col_double(),
##   ..   member_casual = col_character()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

```
str(jun_data_2020)
```

```
## spc_tbl_ [343,005 × 13] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ ride_id           : chr [1:343005] "8CD5DE2C2B6C4CFC"
"9A191EB2C751D85D" "F37D14B0B5659BCF" "C41237B506E85FA1" ...
##  $ rideable_type     : chr [1:343005] "docked_bike" "docked_bike"
"docked_bike" "docked_bike" ...
##  $ started_at        : POSIXct[1:343005], format: "2020-06-13 23:24:48"
"2020-06-26 07:26:10" ...
##  $ ended_at          : POSIXct[1:343005], format: "2020-06-13 23:36:55"
"2020-06-26 07:31:58" ...
```

```
##  $ start_station_name: chr [1:343005] "Wilton Ave & Belmont Ave" "Federal
St & Polk St" "Daley Center Plaza" "Broadway & Cornelia Ave" ...
##  $ start_station_id  : num [1:343005] 117 41 81 303 327 327 41 115 338 84
...
##  $ end_station_name  : chr [1:343005] "Damen Ave & Clybourn Ave" "Daley
Center Plaza" "State St & Harrison St" "Broadway & Berwyn Ave" ...
##  $ end_station_id    : num [1:343005] 163 81 5 294 117 117 81 303 164 53
...
##  $ start_lat         : num [1:343005] 41.9 41.9 41.9 41.9 41.9 ...
##  $ start_lng         : num [1:343005] -87.7 -87.6 -87.6 -87.6 -87.7 ...
##  $ end_lat           : num [1:343005] 41.9 41.9 41.9 42 41.9 ...
##  $ end_lng           : num [1:343005] -87.7 -87.6 -87.6 -87.7 -87.7 ...
##  $ member_casual     : chr [1:343005] "casual" "member" "member" "casual"
...
##  - attr(*, "spec")=
##   .. cols(
##   ..   ride_id = col_character(),
##   ..   rideable_type = col_character(),
##   ..   started_at = col_datetime(format = ""),
##   ..   ended_at = col_datetime(format = ""),
##   ..   start_station_name = col_character(),
##   ..   start_station_id = col_double(),
##   ..   end_station_name = col_character(),
##   ..   end_station_id = col_double(),
##   ..   start_lat = col_double(),
##   ..   start_lng = col_double(),
##   ..   end_lat = col_double(),
##   ..   end_lng = col_double(),
##   ..   member_casual = col_character()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

I repeated it for all the data files.

Solving Data Consistency Issue

To ensure datatype consistency in the datasets, I converted the columns **start_station_id** and **end_station_id** to character datatype for the data files from Apr 2020 - Nov 2020 so now the two fields are character datatype in all the datasets.

```
apr_data_2020 <- apr_data_2020 %>%
mutate(start_station_id=as.character(start_station_id),end_station_id=as.char
acter(end_station_id))

may_data_2020 <- may_data_2020 %>%
mutate(start_station_id=as.character(start_station_id),end_station_id=as.char
acter(end_station_id))

jun_data_2020 <- jun_data_2020 %>%
mutate(start_station_id=as.character(start_station_id),end_station_id=as.char
```

```
acter(end_station_id))

jul_data_2020 <- jul_data_2020 %>%
mutate(start_station_id=as.character(start_station_id),end_station_id=as.char
acter(end_station_id))

aug_data_2020 <- aug_data_2020 %>%
mutate(start_station_id=as.character(start_station_id),end_station_id=as.char
acter(end_station_id))

sep_data_2020 <- sep_data_2020 %>%
mutate(start_station_id=as.character(start_station_id),end_station_id=as.char
acter(end_station_id))

oct_data_2020 <- oct_data_2020 %>%
mutate(start_station_id=as.character(start_station_id),end_station_id=as.char
acter(end_station_id))

nov_data_2020 <- nov_data_2020 %>%
mutate(start_station_id=as.character(start_station_id),end_station_id=as.char
acter(end_station_id))
```

Combining the Data

I combined all the data files in one dataset “data_combined” using the rbind () function

```
data_combined <- rbind(
  apr_data_2020, may_data_2020, jun_data_2020, jul_data_2020, aug_data_2020,
sep_data_2020, oct_data_2020, nov_data_2020, dec_data_2020, jan_data_2021,
feb_data_2021, mar_data_2021, apr_data_2021, may_data_2021, jun_data_2021,
jul_data_2021, aug_data_2021, sep_data_2021, oct_data_2021, nov_data_2021,
dec_data_2021, jan_data_2022, feb_data_2022, mar_data_2022, apr_data_2022,
may_data_2022, jun_data_2022, jul_data_2022, aug_data_2022, sep_data_2022,
oct_data_2022, nov_data_2022, dec_data_2022, jan_data_2023, feb_data_2023,
mar_data_2023, apr_data_2023, may_data_2023, jun_data_2023, jul_data_2023,
aug_data_2023, sep_data_2023, oct_data_2023, nov_data_2023, dec_data_2023,
jan_data_2024, feb_data_2024, mar_data_2024, apr_data_2024, may_data_2024,
jun_data_2024, jul_data_2024, aug_data_2024, sep_data_2024, oct_data_2024,
nov_data_2024
)
str(data_combined)

## tibble [25,779,649 × 13] (S3: tbl_df/tbl/data.frame)
##  $ ride_id           : chr [1:25779649] "A847FADBBC638E45"
"5405B80E996FF60D" "5DD24A79A4E006F4" "2A59BBDF5CDBA725" ...
##  $ rideable_type     : chr [1:25779649] "docked_bike" "docked_bike"
"docked_bike" "docked_bike" ...
##  $ started_at        : POSIXct[1:25779649], format: "2020-04-26 17:45:14"
"2020-04-17 17:08:54" ...
##  $ ended_at          : POSIXct[1:25779649], format: "2020-04-26 18:12:03"
```

```
"2020-04-17 17:17:03" ...
##  $ start_station_name: chr [1:25779649] "Eckhart Park" "Drake Ave &
Fullerton Ave" "McClurg Ct & Erie St" "California Ave & Division St" ...
##  $ start_station_id  : chr [1:25779649] "86" "503" "142" "216" ...
##  $ end_station_name  : chr [1:25779649] "Lincoln Ave & Diversey Pkwy"
"Kosciuszko Park" "Indiana Ave & Roosevelt Rd" "Wood St & Augusta Blvd" ...
##  $ end_station_id    : chr [1:25779649] "152" "499" "255" "657" ...
##  $ start_lat         : num [1:25779649] 41.9 41.9 41.9 41.9 41.9 ...
##  $ start_lng         : num [1:25779649] -87.7 -87.7 -87.6 -87.7 -87.6 ...
##  $ end_lat           : num [1:25779649] 41.9 41.9 41.9 41.9 42 ...
##  $ end_lng           : num [1:25779649] -87.7 -87.7 -87.6 -87.7 -87.7 ...
##  $ member_casual     : chr [1:25779649] "member" "member" "member"
"member" ...
```

Data Cleaning

Removing null values

I used **drop_na()** function to remove the null values from the dataset.

```
no_null_data <- drop_na(data_combined)
str(no_null_data)
```

```
## tibble [20,329,443 × 13] (S3: tbl_df/tbl/data.frame)
##  $ ride_id           : chr [1:20329443] "A847FADBBC638E45"
"5405B80E996FF60D" "5DD24A79A4E006F4" "2A59BBDF5CDBA725" ...
##  $ rideable_type     : chr [1:20329443] "docked_bike" "docked_bike"
"docked_bike" "docked_bike" ...
##  $ started_at        : POSIXct[1:20329443], format: "2020-04-26 17:45:14"
"2020-04-17 17:08:54" ...
##  $ ended_at          : POSIXct[1:20329443], format: "2020-04-26 18:12:03"
"2020-04-17 17:17:03" ...
##  $ start_station_name: chr [1:20329443] "Eckhart Park" "Drake Ave &
Fullerton Ave" "McClurg Ct & Erie St" "California Ave & Division St" ...
##  $ start_station_id  : chr [1:20329443] "86" "503" "142" "216" ...
##  $ end_station_name  : chr [1:20329443] "Lincoln Ave & Diversey Pkwy"
"Kosciuszko Park" "Indiana Ave & Roosevelt Rd" "Wood St & Augusta Blvd" ...
##  $ end_station_id    : chr [1:20329443] "152" "499" "255" "657" ...
##  $ start_lat         : num [1:20329443] 41.9 41.9 41.9 41.9 41.9 ...
##  $ start_lng         : num [1:20329443] -87.7 -87.7 -87.6 -87.7 -87.6 ...
##  $ end_lat           : num [1:20329443] 41.9 41.9 41.9 41.9 42 ...
##  $ end_lng           : num [1:20329443] -87.7 -87.7 -87.6 -87.7 -87.7 ...
##  $ member_casual     : chr [1:20329443] "member" "member" "member"
"member" ...
```

Check the Data Strucure after Drob the Null Values

```
clean_data <- no_null_data
str(clean_data)
```

```
## tibble [20,329,443 × 13] (S3: tbl_df/tbl/data.frame)
##  $ ride_id           : chr [1:20329443] "A847FADBBC638E45"
```

```
"5405B80E996FF60D" "5DD24A79A4E006F4" "2A59BBDF5CDBA725" ...
##  $ rideable_type     : chr [1:20329443] "docked_bike" "docked_bike"
"docked_bike" "docked_bike" ...
##  $ started_at        : POSIXct[1:20329443], format: "2020-04-26 17:45:14"
"2020-04-17 17:08:54" ...
##  $ ended_at          : POSIXct[1:20329443], format: "2020-04-26 18:12:03"
"2020-04-17 17:17:03" ...
##  $ start_station_name: chr [1:20329443] "Eckhart Park" "Drake Ave &
Fullerton Ave" "McClurg Ct & Erie St" "California Ave & Division St" ...
##  $ start_station_id  : chr [1:20329443] "86" "503" "142" "216" ...
##  $ end_station_name  : chr [1:20329443] "Lincoln Ave & Diversey Pkwy"
"Kosciuszko Park" "Indiana Ave & Roosevelt Rd" "Wood St & Augusta Blvd" ...
##  $ end_station_id    : chr [1:20329443] "152" "499" "255" "657" ...
##  $ start_lat         : num [1:20329443] 41.9 41.9 41.9 41.9 41.9 ...
##  $ start_lng         : num [1:20329443] -87.7 -87.7 -87.6 -87.7 -87.6 ...
##  $ end_lat           : num [1:20329443] 41.9 41.9 41.9 41.9 42 ...
##  $ end_lng           : num [1:20329443] -87.7 -87.7 -87.6 -87.7 -87.7 ...
##  $ member_casual     : chr [1:20329443] "member" "member" "member"
"member" ...
```

Exploring some the Data

```
head(clean_data)
```

```
## # A tibble: 6 × 13
##   ride_id          rideable_type started_at          ended_at
##   <chr>            <chr>         <dttm>              <dttm>
## 1 A847FADBBC638E45 docked_bike   2020-04-26 17:45:14 2020-04-26 18:12:03
## 2 5405B80E996FF60D docked_bike   2020-04-17 17:08:54 2020-04-17 17:17:03
## 3 5DD24A79A4E006F4 docked_bike   2020-04-01 17:54:13 2020-04-01 18:08:36
## 4 2A59BBDF5CDBA725 docked_bike   2020-04-07 12:50:19 2020-04-07 13:02:31
## 5 27AD306C119C6158 docked_bike   2020-04-18 10:22:59 2020-04-18 11:15:54
## 6 356216E875132F61 docked_bike   2020-04-30 17:55:47 2020-04-30 18:01:11
## # ℹ 9 more variables: start_station_name <chr>, start_station_id <chr>,
## #   end_station_name <chr>, end_station_id <chr>, start_lat <dbl>,
## #   start_lng <dbl>, end_lat <dbl>, end_lng <dbl>, member_casual <chr>
```

Exploring the Data more deeper

I Explored the Data by **glimpse()** function

```
glimpse(clean_data)
```

```
## Rows: 20,329,443
## Columns: 13
## $ ride_id           <chr> "A847FADBBC638E45", "5405B80E996FF60D",
"5DD24A79A4…
## $ rideable_type     <chr> "docked_bike", "docked_bike", "docked_bike",
"docke…
## $ started_at        <dttm> 2020-04-26 17:45:14, 2020-04-17 17:08:54,
2020-04-…
## $ ended_at          <dttm> 2020-04-26 18:12:03, 2020-04-17 17:17:03,
```

```
2020-04-…
## $ start_station_name <chr> "Eckhart Park", "Drake Ave & Fullerton Ave",
"McClu…
## $ start_station_id   <chr> "86", "503", "142", "216", "125", "173", "35",
"434…
## $ end_station_name   <chr> "Lincoln Ave & Diversey Pkwy", "Kosciuszko
Park", "…
## $ end_station_id     <chr> "152", "499", "255", "657", "323", "35", "635",
"38…
## $ start_lat          <dbl> 41.8964, 41.9244, 41.8945, 41.9030, 41.8902,
41.896…
## $ start_lng          <dbl> -87.6610, -87.7154, -87.6179, -87.6975, -
87.6262, -…
## $ end_lat            <dbl> 41.9322, 41.9306, 41.8679, 41.8992, 41.9695,
41.892…
## $ end_lng            <dbl> -87.6586, -87.7238, -87.6230, -87.6722, -
87.6547, -…
## $ member_casual      <chr> "member", "member", "member", "member",
"casual", "…
```

Remove the unnecessary fields

I removed the latitude and longitude fields from the data

```
clean_data <- clean_data %>%
  select(-c(start_lat, start_lng, end_lat, end_lng))
```

Explore the Data columns

I Explored the columns after removing the unnecessary ones

```
colnames(clean_data)

## [1] "ride_id"            "rideable_type"      "started_at"
## [4] "ended_at"           "start_station_name" "start_station_id"
## [7] "end_station_name"   "end_station_id"     "member_casual"
```

Adding new columns

I added 3 columns to the dataset by abstracting the **date** and **month** from the column started_at, then I calculated new field for **ride_length** from the started_at and ended_at fields.

Create the date, month and ride_length columns

I Abstracted the date and month columns from the **started_at** column

```
clean_data$date <- as.Date(clean_data$started_at)
clean_data$month <- format(as.Date(clean_data$date), "%B")
```

I created the **ride_length** by calculating the different time between the **ended_at** and **started_at** columns

```
clean_data <-clean_data %>%
  mutate (ride_length=difftime(ended_at, started_at, unit="mins"))
```

Again I checked the Data after adding the new fields using the **glimpse()** function

```
glimpse(clean_data)
```

```
## Rows: 20,329,443
## Columns: 12
## $ ride_id            <chr> "A847FADBBC638E45", "5405B80E996FF60D",
"5DD24A79A4…
## $ rideable_type      <chr> "docked_bike", "docked_bike", "docked_bike",
"docke…
## $ started_at         <dttm> 2020-04-26 17:45:14, 2020-04-17 17:08:54,
2020-04-…
## $ ended_at           <dttm> 2020-04-26 18:12:03, 2020-04-17 17:17:03,
2020-04-…
## $ start_station_name <chr> "Eckhart Park", "Drake Ave & Fullerton Ave",
"McClu…
## $ start_station_id   <chr> "86", "503", "142", "216", "125", "173", "35",
"434…
## $ end_station_name   <chr> "Lincoln Ave & Diversey Pkwy", "Kosciuszko
Park", "…
## $ end_station_id     <chr> "152", "499", "255", "657", "323", "35", "635",
"38…
## $ member_casual      <chr> "member", "member", "member", "member",
"casual", "…
## $ date               <date> 2020-04-26, 2020-04-17, 2020-04-01, 2020-04-
07, 20…
## $ month              <chr> "April", "April", "April", "April", "April",
"April…
## $ ride_length        <drtn> 26.816667 mins, 8.150000 mins, 14.383333 mins,
12.…
```

Verifying that the field ride_length containing only correct values

```
clean_data <- clean_data[!(clean_data$ride_length <= 0 |
clean_data$ride_length > 1440),]
```

Change the data frame name to final_data

```
final_data <- clean_data
```

Analysis

Here I will analyse the data to find some patterns and trends in it, so I start by calculating the **mean, median, max and min** for all the dataset to exploring the data.

```
final_data %>%
  summarise(ride_length_mean=mean(ride_length)
            ,ride_length_median=median(ride_length)
```

```
            ,ride_length_min=min(ride_length)
            ,ride_length_max=max(ride_length))
```

```
## # A tibble: 1 × 4
##   ride_length_mean ride_length_median ride_length_min  ride_length_max
##   <drtn>           <drtn>             <drtn>           <drtn>
## 1 18.52197 mins    11.21667 mins      0.001716669 mins 1439.9 mins
```

Calculate the mean, median, max and min by the rider type

I calculated the mean for ride length grouping by the rider type

```
mean_by_rider_type <- final_data %>%
  group_by(member_casual) %>%
  summarize(mean_ride_length = mean(ride_length, na.rm = TRUE))
mean_by_rider_type
```

```
## # A tibble: 2 × 2
##   member_casual mean_ride_length
##   <chr>         <drtn>
## 1 casual        26.8789 mins
## 2 member        13.0062 mins
```

I calculated the median for ride length grouping by the rider type

```
median_by_rider_type <- final_data %>%
  group_by(member_casual) %>%
  summarize(median_ride_length = median(ride_length, na.rm = TRUE))
median_by_rider_type
```

```
## # A tibble: 2 × 2
##   member_casual median_ride_length
##   <chr>         <drtn>
## 1 casual        15.266667 mins
## 2 member         9.366667 mins
```

I calculated the max for ride length grouping by the rider type

```
max_by_rider_type <- final_data %>%
  group_by(member_casual) %>%
  summarize(max_ride_length = max(ride_length, na.rm = TRUE))
max_by_rider_type
```

```
## # A tibble: 2 × 2
##   member_casual max_ride_length
##   <chr>         <drtn>
## 1 casual        1439.900 mins
## 2 member        1439.867 mins
```

I calculated the min for ride length grouping by the rider type

```r
min_by_rider_type <- final_data %>%
  group_by(member_casual) %>%
  summarize(min_ride_length = min(ride_length, na.rm = TRUE))
min_by_rider_type
```

```
## # A tibble: 2 × 2
##   member_casual min_ride_length
##   <chr>         <drtn>
## 1 casual        0.001716669 mins
## 2 member        0.002316666 mins
```

I counted the number of rides has taken by different types of customers (Annual member or Casual rider)

```r
num_rides_by_type <- final_data %>%
  group_by(member_casual) %>%
  summarize(num_rides = n_distinct(ride_id, na.rm = TRUE))
num_rides_by_type
```

```
## # A tibble: 2 × 2
##   member_casual num_rides
##   <chr>             <int>
## 1 casual          8076615
## 2 member         12236918
```

Here I've calculated the most common stations that the casual riders are starting from every time the take a ride, I limited for the top 5 start stations

```r
common_start_station <- final_data %>%
  filter(member_casual == "casual") %>%
  group_by(start_station_name) %>%
  summarize(num_rides = n()) %>%
  arrange(desc(num_rides)) %>%
  slice_head(n = 5)
common_start_station
```

```
## # A tibble: 5 × 2
##   start_station_name                num_rides
##   <chr>                                 <int>
## 1 Streeter Dr & Grand Ave              235034
## 2 Millennium Park                      113080
## 3 Michigan Ave & Oak St                109184
## 4 DuSable Lake Shore Dr & Monroe St    106403
## 5 Shedd Aquarium                        87137
```

Here I've calculated the most common stations that the casual riders are ending their ride to, also I limited for the top 5 end stations

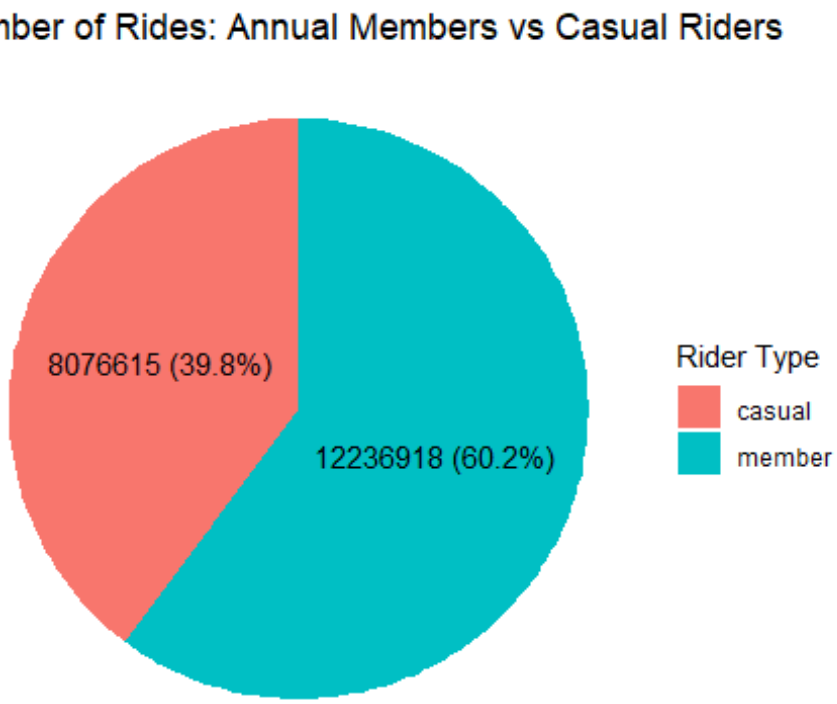
```r
common_end_station <- final_data %>%
  filter(member_casual == "casual") %>%
  group_by(end_station_name) %>%
```

```
  summarize(num_rides = n()) %>%
  arrange(desc(num_rides)) %>%
  slice_head(n = 5)
common_end_station

## # A tibble: 5 × 2
##   end_station_name                  num_rides
##   <chr>                                 <int>
## 1 Streeter Dr & Grand Ave              251551
## 2 Millennium Park                      121556
## 3 Michigan Ave & Oak St                115171
## 4 DuSable Lake Shore Dr & Monroe St     99192
## 5 Theater on the Lake                   90186
```

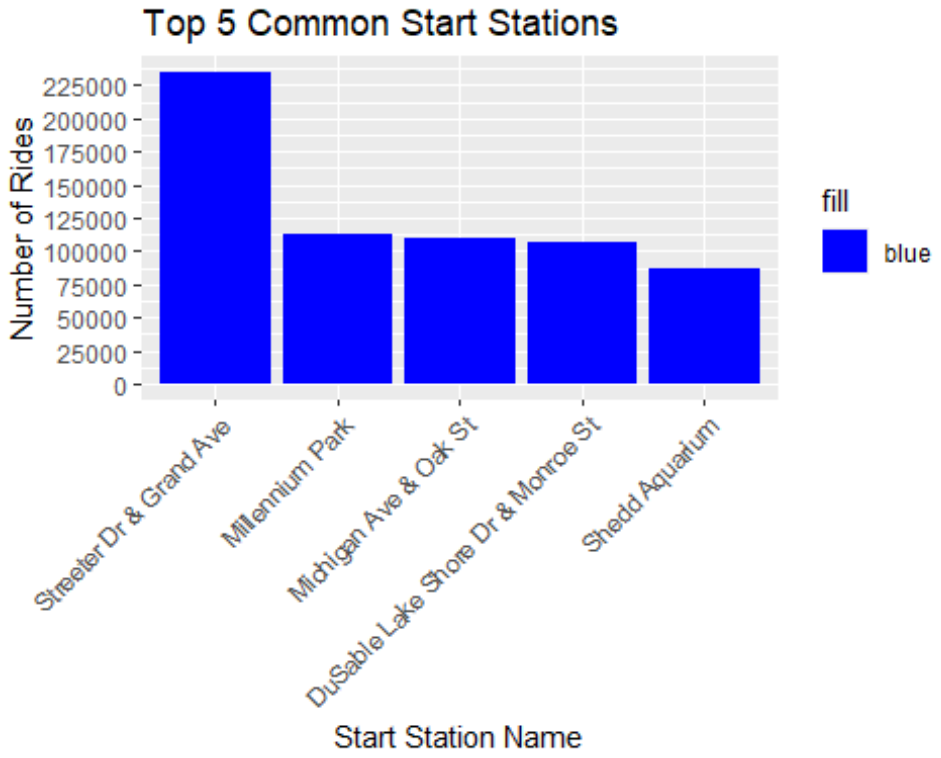
Create the Visualizations

Pie chart



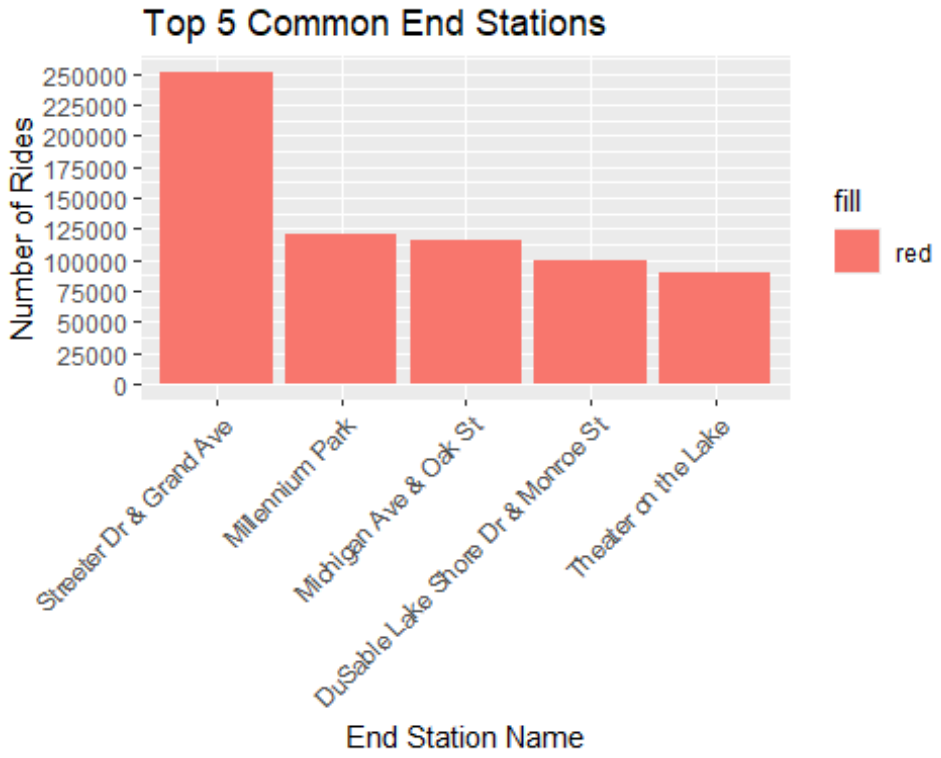
This pie chart I created to explore the Number of Rides for the Annual Members vs Casual Riders

Bar chart to explore the Top 5 common Stations the Casual Rider Start from



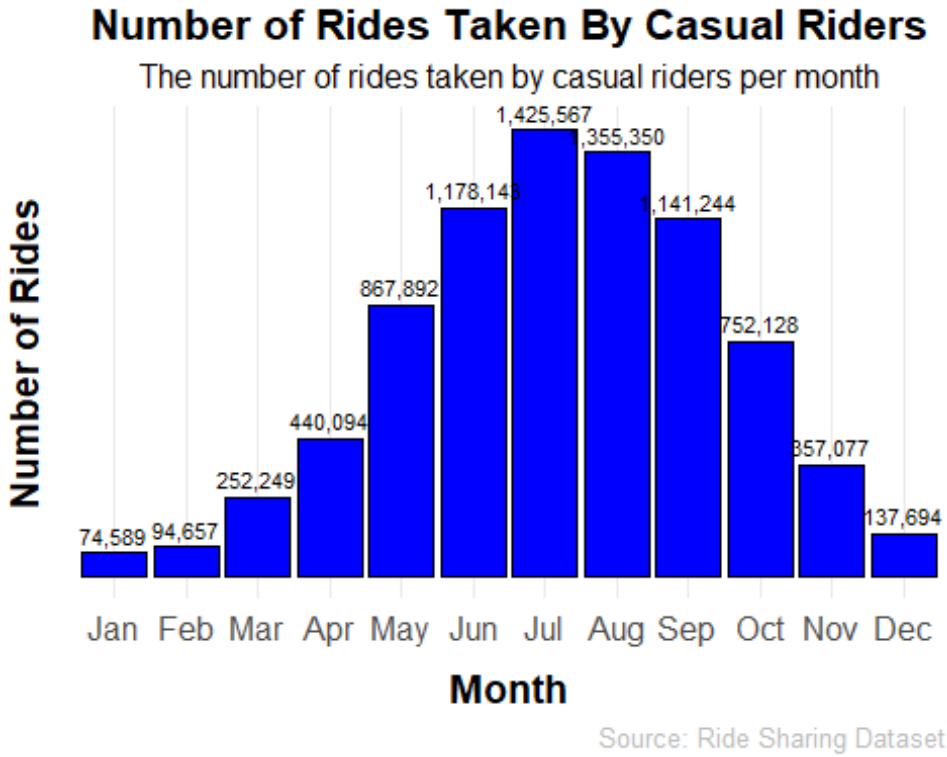
This bar chart I created to explore the Top 5 common Stations the Casual Rider are staring their rides from to use these stations for the marketing campaign later

Bar chart to explore the Top 5 common Stations the Casual Rider End to



This bar chart I created to explore the Top 5 common Stations the Casual Rider end their rides at to use them also for the marketing campaign, and I discovered that the top 4 stations the casual riders start from are the same top 4 stations end to.

Histogram chart to explore the number of rides taken by the casual riders per month



This Histogram chart exploring the number of rides has been taken by the casual riders per month to find the maximum number of rides its been taken in which months

My Recommendations

1. As we explore the data and visuals I recommend to use the top 4 stations **(Streeter Dr & Grand Ave, Millennium Park, Michigan Ave & Oak St, DuSable Lake Shore Dr & Monroe St)** in the marketing campaign by making Billboards, flyers and posters in these stations that convincing the casual riders to convert to annual members
2. It's clearly that the **Streeter Dr & Grand Ave** station is most common station and have very high number of casual rider are passing by in the beginning or ends their rides, so I recommend to make special offer for annual members who passing by it in their 1st year at least that will attract more casual riders.
3. I see also it's helpful to Intensify the marketing campaign in the summer **(June, July, August and September)** because it's the most months that the casual riders are using the bike sharing services.