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.character(end\_station\_id))  
  
may\_data\_2020 <- may\_data\_2020 %>% mutate(start\_station\_id=as.character(start\_station\_id),end\_station\_id=as.character(end\_station\_id))  
  
jun\_data\_2020 <- jun\_data\_2020 %>%  
mutate(start\_station\_id=as.character(start\_station\_id),end\_station\_id=as.character(end\_station\_id))  
  
jul\_data\_2020 <- jul\_data\_2020 %>% mutate(start\_station\_id=as.character(start\_station\_id),end\_station\_id=as.character(end\_station\_id))  
  
aug\_data\_2020 <- aug\_data\_2020 %>%  
mutate(start\_station\_id=as.character(start\_station\_id),end\_station\_id=as.character(end\_station\_id))  
  
sep\_data\_2020 <- sep\_data\_2020 %>% mutate(start\_station\_id=as.character(start\_station\_id),end\_station\_id=as.character(end\_station\_id))  
  
oct\_data\_2020 <- oct\_data\_2020 %>% mutate(start\_station\_id=as.character(start\_station\_id),end\_station\_id=as.character(end\_station\_id))  
  
nov\_data\_2020 <- nov\_data\_2020 %>% mutate(start\_station\_id=as.character(start\_station\_id),end\_station\_id=as.character(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

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)

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

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

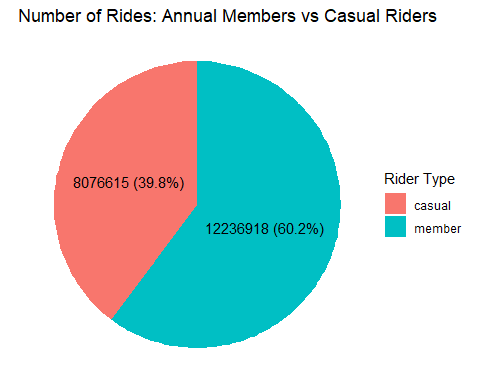
Here I’ve calculated the most common stations that the casual riders are ending their ride to, also I limited for the top 5 end stations

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

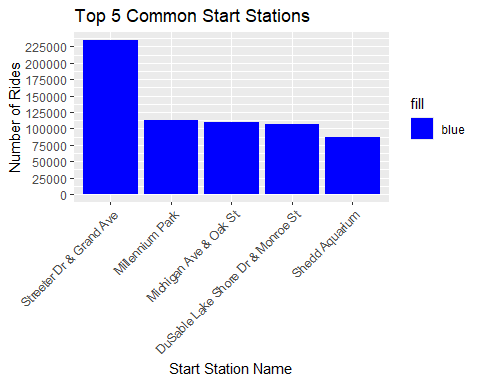
### Create the Visualizations

#### Pie chart



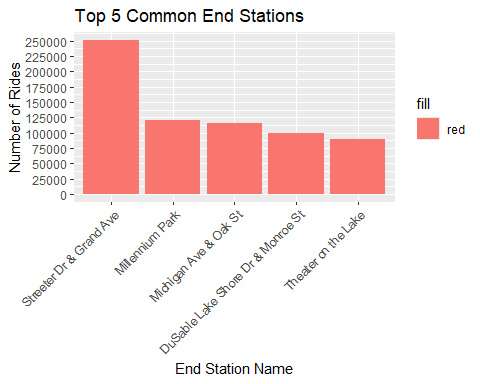
This pie chart I created to explore the Number of Rides for the Annual Members vs Casual Riders

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



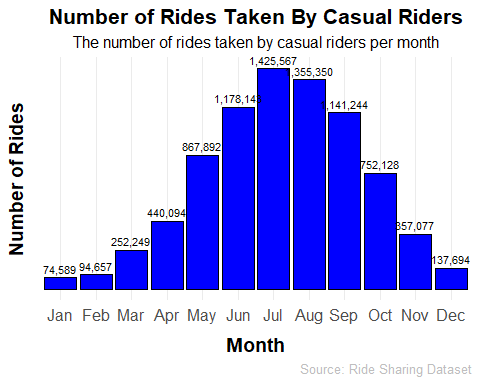
This bar chart I created to explore the Top 5 common Stations the Casual Rider are staring their rides from to use these stations for the marketing campaign later

#### 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 Recommondations

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.