Cyclistic

Genaro Angeloni

03/01/2022

Introduction

Cyclistic is a bike-share company based in Chicago, that features more than 5,800 bikes and 600 docking stations. This program makes bike-sharing more inclusive to people with disabilities and riders who can't use a standard two-wheeled bikes, also offering reclining bikes, hand tricycles and cargo bikes, representing these options the 8% of the riders. The company offers three flexible pricing plans: single-ride passes, full-day passes, and annual memberships. Customers who purchase *single-ride* or *full-day passes* are referred to as **casual riders**, and the customers who purchase *annual memberships* are **Cyclistic members**.

Statement of the business task

Finance analysts have concluded that annual members are much more profitable than casual riders. Although the pricing flexibility helps to attract more customers, we believe that *maximizing the number of annual members will be key to future growth.* Rather than creating a marketing campaign that targets all new customers, we believe that there is a very good chance to **convert casual riders into members**. The goal is **to design marketing strategies aimed at converting casual riders into annual members**. In order to do that, we need to better understand how annual members and casual riders differ, which casual riders would buy a membership, and how digital media could affect their marketing tactics. In this analysis we are going to address the first question:

How do annual members and casual riders use Cyclistic bikes differently?

We hope that answering this first question will help us to decipher what differential characteristics of the product offered by the company are those that attract casual customers, and how to enhance these characteristics in order to convert these customers to members of the company. To do this, we propose to answer the following questions:

- What is the average travel duration for each type of customer?
- Are there any seasonal trend in customer demand for bycicles?
- Is there a weekly trend in customer demand for bicycles?

- Are there peak times in the demand for bicycles by customers?
- What are the preferred stations for casual riders?
- Do casual riders prefer electric or regular bikes?

Considering key stakeholders

For this project, we're considering the next key stakeholders:

- **Manager**: Lily Moreno is the director of marketing, responsible for the campaigns and initiatives to promote the bike-sharing program. She will be interested in the insights I'd get from the analysis to develop a data-driven strategy to attract casual riders to the annual membership.
- **Marketing Analytics Team**: we are responsible for collecting, analyzing and reporting the data. My teammates will be interested in the data cleaning process, the tools that I use for cleaning and analyzing the data, and how I will share them through data visualization tools.
- **Executive team**: they are responsible for decide whether to approve the recommended marketing program. They are going to be interested in the insights that support the marketing strategy proposed.

Prepairing data: description of sources used

The dataset that we will use corresponds to **twelve months of Cyclistic's trips**. This kind of data is named *first-party data*, collected by the company using their own resources (This is a fictional case study, and the company Ciclystic doesn't exists. The data has been made available by Motivate International Inc. under this license). The data is located in the cloud of the company, so it can be easily accessed. All trips for a particular month are stored in a single data frame, identifying each trip with an ID. In addition to the type of customer and bicycle, the day and time, station, latitude and longitude of both the start and end of the trip are stored for each of them. As the data was collected by the company, we could say that it is reliable and original. On the other hand, being the most recent data available, we have a current data set. In addition, we believe that this data will allow us to characterize each trip, and together, to be able to build a profile of casual client and member, which will be key to meeting the objective of the analysis. Finally, the privacy that the company imparts on these data means that is not cited by third parties. However, considering the origin of the data, we can say that this should not be a problem or question its reliability. An essential point in the data set is the concern for customer privacy, avoiding collecting rider's personally identifiable information. This means that with this data we will not be able to connect trips with credit card numbers to access sensitive information about our clients. The information available is the minimum necessary to be able to characterize each bicycle trip.

Now we are going to start the preparing phase of the project. Here, we'll check the data integrity, and we are going to determine if there are any problems with the data. The first step will be loading the data:

```
library(tidyverse)

data1 <- read_csv("data/202101-divvy-tripdata.csv")
data2 <- read_csv("data/202102-divvy-tripdata.csv")
data3 <- read_csv("data/202103-divvy-tripdata.csv")
data4 <- read_csv("data/202104-divvy-tripdata.csv")
data5 <- read_csv("data/202105-divvy-tripdata.csv")
data6 <- read_csv("data/202106-divvy-tripdata.csv")
data7 <- read_csv("data/202107-divvy-tripdata.csv")
data8 <- read_csv("data/202108-divvy-tripdata.csv")
data9 <- read_csv("data/202109-divvy-tripdata.csv")
data10 <- read_csv("data/202110-divvy-tripdata.csv")
data11 <- read_csv("data/202111-divvy-tripdata.csv")
data12 <- read_csv("data/202012-divvy-tripdata.csv")
```

First we are going to make sure that the data frames are consistent with each other, in order to make the final merge. We will review column names and formats:

```
glimpse(data1)
```

```
## Rows: 96,834
## Columns: 13
## $ ride_id            <chr> "E19E6F1B8D4C42ED", "DC88F20C2C55F27F", "EC45C94683~
## $ rideable_type      <chr> "electric_bike", "electric_bike", "electric_bike", ~
## $ started_at         <dttm> 2021-01-23 16:14:19, 2021-01-27 18:43:08, 2021-01-~
## $ ended_at           <dttm> 2021-01-23 16:24:44, 2021-01-27 18:47:12, 2021-01-~
## $ start_station_name <chr> "California Ave & Cortez St", "California Ave & Cor~
## $ start_station_id   <chr> "17660", "17660", "17660", "17660", "17660", "17660~
## $ end_station_name   <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, "Wood St & Augu~
## $ end_station_id     <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, "657", "13258",~
## $ start_lat          <dbl> 41.90034, 41.90033, 41.90031, 41.90040, 41.90033, 4~
## $ start_lng          <dbl> -87.69674, -87.69671, -87.69664, -87.69666, -87.696~
## $ end_lat            <dbl> 41.89000, 41.90000, 41.90000, 41.92000, 41.90000, 4~
## $ end_lng            <dbl> -87.72000, -87.69000, -87.70000, -87.69000, -87.700~
## $ member_casual      <chr> "member", "member", "member", "member", "casual", "~
```

```
glimpse(data2)
```

```
## Rows: 49,622
```

```
## Columns: 13
## $ ride_id            <chr> "89E7AA6C29227EFF", "0FEFDE2603568365", "E6159D746B~
## $ rideable_type      <chr> "classic_bike", "classic_bike", "electric_bike", "c~
## $ started_at         <dttm> 2021-02-12 16:14:56, 2021-02-14 17:52:38, 2021-02-~
## $ ended_at           <dttm> 2021-02-12 16:21:43, 2021-02-14 18:12:09, 2021-02-~
## $ start_station_name <chr> "Glenwood Ave & Touhy Ave", "Glenwood Ave & Touhy A~
## $ start_station_id   <chr> "525", "525", "KA1503000012", "637", "13216", "1800~
## $ end_station_name   <chr> "Sheridan Rd & Columbia Ave", "Bosworth Ave & Howar~
## $ end_station_id     <chr> "660", "16806", "TA1305000029", "TA1305000034", "TA~
## $ start_lat          <dbl> 42.01270, 42.01270, 41.88579, 41.89563, 41.83473, 4~
## $ start_lng          <dbl> -87.66606, -87.66606, -87.63110, -87.67207, -87.625~
## $ end_lat            <dbl> 42.00458, 42.01954, 41.88487, 41.90312, 41.83816, 4~
## $ end_lng            <dbl> -87.66141, -87.66956, -87.62750, -87.67394, -87.645~
## $ member_casual      <chr> "member", "casual", "member", "member", "member", "~
```

```
glimpse(data3)
```

```
## Rows: 228,496
## Columns: 13
## $ ride_id            <chr> "CFA86D4455AA1030", "30D9DC61227D1AF3", "846D87A156~
## $ rideable_type      <chr> "classic_bike", "classic_bike", "classic_bike", "cl~
## $ started_at         <dttm> 2021-03-16 08:32:30, 2021-03-28 01:26:28, 2021-03-~
## $ ended_at           <dttm> 2021-03-16 08:36:34, 2021-03-28 01:36:55, 2021-03-~
## $ start_station_name <chr> "Humboldt Blvd & Armitage Ave", "Humboldt Blvd & Ar~
## $ start_station_id   <chr> "15651", "15651", "15443", "TA1308000021", "525", "~
## $ end_station_name   <chr> "Stave St & Armitage Ave", "Central Park Ave & Bloo~
## $ end_station_id     <chr> "13266", "18017", "TA1308000043", "13323", "E008", ~
## $ start_lat          <dbl> 41.91751, 41.91751, 41.84273, 41.96881, 42.01270, 4~
## $ start_lng          <dbl> -87.70181, -87.70181, -87.63549, -87.65766, -87.666~
## $ end_lat            <dbl> 41.91774, 41.91417, 41.83066, 41.95283, 42.05049, 4~
## $ end_lng            <dbl> -87.69139, -87.71676, -87.64717, -87.64999, -87.677~
## $ member_casual      <chr> "casual", "casual", "casual", "casual", "casual", "~
```

```
glimpse(data4)
```

```
## Rows: 337,230
## Columns: 13
## $ ride_id            <chr> "6C992BD37A98A63F", "1E0145613A209000", "E498E15508~
## $ rideable_type      <chr> "classic_bike", "docked_bike", "docked_bike", "clas~
## $ started_at         <dttm> 2021-04-12 18:25:36, 2021-04-27 17:27:11, 2021-04-~
## $ ended_at           <dttm> 2021-04-12 18:56:55, 2021-04-27 18:31:29, 2021-04-~
## $ start_station_name <chr> "State St & Pearson St", "Dorchester Ave & 49th St"~
## $ start_station_id   <chr> "TA1307000061", "KA1503000069", "20121", "TA1305000~
```

```
## $ end_station_name   <chr> "Southport Ave & Waveland Ave", "Dorchester Ave & 4~
## $ end_station_id     <chr> "13235", "KA1503000069", "20121", "13235", "20121",~
## $ start_lat          <dbl> 41.89745, 41.80577, 41.74149, 41.90312, 41.74149, 4~
## $ start_lng          <dbl> -87.62872, -87.59246, -87.65841, -87.67394, -87.658~
## $ end_lat            <dbl> 41.94815, 41.80577, 41.74149, 41.94815, 41.74149, 4~
## $ end_lng            <dbl> -87.66394, -87.59246, -87.65841, -87.66394, -87.658~
## $ member_casual      <chr> "member", "casual", "casual", "member", "casual", "~
```

```
glimpse(data5)
```

```
## Rows: 531,633
## Columns: 13
## $ ride_id            <chr> "C809ED75D6160B2A", "DD59FDCE0ACACAF3", "0AB83CB88C~
## $ rideable_type      <chr> "electric_bike", "electric_bike", "electric_bike", ~
## $ started_at         <dttm> 2021-05-30 11:58:15, 2021-05-30 11:29:14, 2021-05-~
## $ ended_at           <dttm> 2021-05-30 12:10:39, 2021-05-30 12:14:09, 2021-05-~
## $ start_station_name <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,~
## $ start_station_id   <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,~
## $ end_station_name   <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,~
## $ end_station_id     <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,~
## $ start_lat          <dbl> 41.90000, 41.88000, 41.92000, 41.92000, 41.94000, 4~
## $ start_lng          <dbl> -87.63000, -87.62000, -87.70000, -87.70000, -87.690~
## $ end_lat            <dbl> 41.89000, 41.79000, 41.92000, 41.94000, 41.94000, 4~
## $ end_lng            <dbl> -87.61000, -87.58000, -87.70000, -87.69000, -87.700~
## $ member_casual      <chr> "casual", "casual", "casual", "casual", "casual", "~
```

```
glimpse(data6)
```

```
## Rows: 729,595
## Columns: 13
## $ ride_id            <chr> "99FEC93BA843FB20", "06048DCFC8520CAF", "9598066F68~
## $ rideable_type      <chr> "electric_bike", "electric_bike", "electric_bike", ~
## $ started_at         <dttm> 2021-06-13 14:31:28, 2021-06-04 11:18:02, 2021-06-~
## $ ended_at           <dttm> 2021-06-13 14:34:11, 2021-06-04 11:24:19, 2021-06-~
## $ start_station_name <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,~
## $ start_station_id   <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,~
## $ end_station_name   <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, "Michigan Ave &~
## $ end_station_id     <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, "13042", NA, NA~
## $ start_lat          <dbl> 41.80, 41.79, 41.80, 41.78, 41.80, 41.78, 41.79, 41~
## $ start_lng          <dbl> -87.59, -87.59, -87.60, -87.58, -87.59, -87.58, -87~
## $ end_lat            <dbl> 41.80000, 41.80000, 41.79000, 41.80000, 41.79000, 4~
## $ end_lng            <dbl> -87.6000, -87.6000, -87.5900, -87.6000, -87.5900, -~
## $ member_casual      <chr> "member", "member", "member", "member", "member", "~
```

```
glimpse(data7)
```

```
## Rows: 822,410
## Columns: 13
## $ ride_id            <chr> "0A1B623926EF4E16", "B2D5583A5A5E76EE", "6F264597DD~
## $ rideable_type      <chr> "docked_bike", "classic_bike", "classic_bike", "cla~
## $ started_at         <dttm> 2021-07-02 14:44:36, 2021-07-07 16:57:42, 2021-07-~
## $ ended_at           <dttm> 2021-07-02 15:19:58, 2021-07-07 17:16:09, 2021-07-~
## $ start_station_name <chr> "Michigan Ave & Washington St", "California Ave & C~
## $ start_station_id   <chr> "13001", "17660", "SL-012", "17660", "17660", "1766~
## $ end_station_name   <chr> "Halsted St & North Branch St", "Wood St & Hubbard ~
## $ end_station_id     <chr> "KA1504000117", "13432", "KA1503000044", "13196", "~
## $ start_lat          <dbl> 41.88398, 41.90036, 41.86038, 41.90036, 41.90035, 4~
## $ start_lng          <dbl> -87.62468, -87.69670, -87.62581, -87.69670, -87.696~
## $ end_lat            <dbl> 41.89937, 41.88990, 41.89017, 41.89456, 41.88659, 4~
## $ end_lng            <dbl> -87.64848, -87.67147, -87.62619, -87.65345, -87.658~
## $ member_casual      <chr> "casual", "casual", "member", "member", "casual", "~
```

```
glimpse(data8)
```

```
## Rows: 804,352
## Columns: 13
## $ ride_id            <chr> "99103BB87CC6C1BB", "EAFCCCFB0A3FC5A1", "9EF4F46C57~
## $ rideable_type      <chr> "electric_bike", "electric_bike", "electric_bike", ~
## $ started_at         <dttm> 2021-08-10 17:15:49, 2021-08-10 17:23:14, 2021-08-~
## $ ended_at           <dttm> 2021-08-10 17:22:44, 2021-08-10 17:39:24, 2021-08-~
## $ start_station_name <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,~
## $ start_station_id   <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,~
## $ end_station_name   <chr> NA, NA, NA, NA, NA, NA, NA, "Clark St & Grace St", ~
## $ end_station_id     <chr> NA, NA, NA, NA, NA, NA, NA, "TA1307000127", NA, NA,~
## $ start_lat          <dbl> 41.77000, 41.77000, 41.95000, 41.97000, 41.79000, 4~
## $ start_lng          <dbl> -87.68000, -87.68000, -87.65000, -87.67000, -87.600~
## $ end_lat            <dbl> 41.77000, 41.77000, 41.97000, 41.95000, 41.77000, 4~
## $ end_lng            <dbl> -87.68000, -87.63000, -87.66000, -87.65000, -87.620~
## $ member_casual      <chr> "member", "member", "member", "member", "member", "~
```

```
glimpse(data9)
```

```
## Rows: 756,147
## Columns: 13
## $ ride_id            <chr> "9DC7B962304CBFD8", "F930E2C6872D6B32", "6EF7213790~
## $ rideable_type      <chr> "electric_bike", "electric_bike", "electric_bike", ~
```

```
## $ started_at         <dttm> 2021-09-28 16:07:10, 2021-09-28 14:24:51, 2021-09-~
## $ ended_at           <dttm> 2021-09-28 16:09:54, 2021-09-28 14:40:05, 2021-09-~
## $ start_station_name <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, "Clark St &~
## $ start_station_id   <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, "TA13070001~
## $ end_station_name   <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,~
## $ end_station_id     <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,~
## $ start_lat          <dbl> 41.89000, 41.94000, 41.81000, 41.80000, 41.88000, 4~
## $ start_lng          <dbl> -87.68000, -87.64000, -87.72000, -87.72000, -87.740~
## $ end_lat            <dbl> 41.89, 41.98, 41.80, 41.81, 41.88, 41.88, 41.74, 41~
## $ end_lng            <dbl> -87.67, -87.67, -87.72, -87.72, -87.71, -87.74, -87~
## $ member_casual      <chr> "casual", "casual", "casual", "casual", "casual", "~
```

```
glimpse(data10)
```

```
## Rows: 631,226
## Columns: 13
## $ ride_id            <chr> "620BC6107255BF4C", "4471C70731AB2E45", "26CA69D43D~
## $ rideable_type      <chr> "electric_bike", "electric_bike", "electric_bike", ~
## $ started_at         <dttm> 2021-10-22 12:46:42, 2021-10-21 09:12:37, 2021-10-~
## $ ended_at           <dttm> 2021-10-22 12:49:50, 2021-10-21 09:14:14, 2021-10-~
## $ start_station_name <chr> "Kingsbury St & Kinzie St", NA, NA, NA, NA, NA, NA,~
## $ start_station_id   <chr> "KA1503000043", NA, NA, NA, NA, NA, NA, NA, NA, NA,~
## $ end_station_name   <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,~
## $ end_station_id     <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,~
## $ start_lat          <dbl> 41.88919, 41.93000, 41.92000, 41.92000, 41.89000, 4~
## $ start_lng          <dbl> -87.63850, -87.70000, -87.70000, -87.69000, -87.710~
## $ end_lat            <dbl> 41.89000, 41.93000, 41.94000, 41.92000, 41.89000, 4~
## $ end_lng            <dbl> -87.63000, -87.71000, -87.72000, -87.69000, -87.690~
## $ member_casual      <chr> "member", "member", "member", "member", "member", "~
```

```
glimpse(data11)
```

```
## Rows: 359,978
## Columns: 13
## $ ride_id            <chr> "7C00A93E10556E47", "90854840DFD508BA", "0A7D10CDD1~
## $ rideable_type      <chr> "electric_bike", "electric_bike", "electric_bike", ~
## $ started_at         <dttm> 2021-11-27 13:27:38, 2021-11-27 13:38:25, 2021-11-~
## $ ended_at           <dttm> 2021-11-27 13:46:38, 2021-11-27 13:56:10, 2021-11-~
## $ start_station_name <chr> NA, NA, NA, NA, NA, "Michigan Ave & Oak St", NA, NA~
## $ start_station_id   <chr> NA, NA, NA, NA, NA, "13042", NA, NA, NA, NA, NA, NA~
## $ end_station_name   <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,~
## $ end_station_id     <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,~
## $ start_lat          <dbl> 41.93000, 41.96000, 41.96000, 41.94000, 41.90000, 4~
```

```
## $ start_lng          <dbl> -87.72000, -87.70000, -87.70000, -87.79000, -87.630~
## $ end_lat            <dbl> 41.96, 41.92, 41.96, 41.93, 41.88, 41.90, 41.80, 41~
## $ end_lng            <dbl> -87.73, -87.70, -87.70, -87.79, -87.62, -87.63, -87~
## $ member_casual      <chr> "casual", "casual", "casual", "casual", "casual", "~
```

```
glimpse(data12)
```

```
## Rows: 131,573
## Columns: 13
## $ ride_id            <chr> "70B6A9A437D4C30D", "158A465D4E74C54A", "5262016E0F~
## $ rideable_type      <chr> "classic_bike", "electric_bike", "electric_bike", "~
## $ started_at         <dttm> 2020-12-27 12:44:29, 2020-12-18 17:37:15, 2020-12-~
## $ ended_at           <dttm> 2020-12-27 12:55:06, 2020-12-18 17:44:19, 2020-12-~
## $ start_station_name <chr> "Aberdeen St & Jackson Blvd", NA, NA, NA, NA, NA, N~
## $ start_station_id   <chr> "13157", NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA~
## $ end_station_name   <chr> "Desplaines St & Kinzie St", NA, NA, NA, NA, NA, NA~
## $ end_station_id     <chr> "TA1306000003", NA, NA, NA, NA, NA, NA, NA, NA, NA,~
## $ start_lat          <dbl> 41.87773, 41.93000, 41.91000, 41.92000, 41.80000, 4~
## $ start_lng          <dbl> -87.65479, -87.70000, -87.69000, -87.70000, -87.590~
## $ end_lat            <dbl> 41.88872, 41.91000, 41.93000, 41.91000, 41.80000, 4~
## $ end_lng            <dbl> -87.64445, -87.70000, -87.70000, -87.70000, -87.590~
## $ member_casual      <chr> "member", "member", "member", "member", "member", "~
```

The data frames look consistent, so let's merge them as the raw version of the data:

```
data <- rbind(data1,data2,data3,data4,data5,data6,data7,data8,data9,data10,data11,data12
```

To get an idea about the stored data, let's look at the unique values, and the NAs in each column:

```
for (col in colnames(data)){
  val <- data[col] %>% unlist() %>% unique() %>% length()

  print(paste0(col,": ",val))
}
```

```
## [1] "ride_id: 5479096"
## [1] "rideable_type: 3"
## [1] "started_at: 4574825"
## [1] "ended_at: 4568129"
## [1] "start_station_name: 846"
## [1] "start_station_id: 832"
## [1] "end_station_name: 840"
```

```
## [1] "end_station_id: 827"
## [1] "start_lat: 377732"
## [1] "start_lng: 358178"
## [1] "end_lat: 443642"
## [1] "end_lng: 403743"
## [1] "member_casual: 2"
```

```
for (col in colnames(data)){
  val <- data[col] %>% unlist() %>% is.na() %>% sum()

  print(paste0(col,": ",val))
}
```

```
## [1] "ride_id: 0"
## [1] "rideable_type: 0"
## [1] "started_at: 0"
## [1] "ended_at: 0"
## [1] "start_station_name: 651445"
## [1] "start_station_id: 651442"
## [1] "end_station_name: 698909"
## [1] "end_station_id: 698909"
## [1] "start_lat: 0"
## [1] "start_lng: 0"
## [1] "end_lat: 4738"
## [1] "end_lng: 4738"
## [1] "member_casual: 0"
```

So, as we see, we have some inconsistencies to clean. Before that, we can't be sure about the integrity of the dataset.

Data Cleaning

First of all, we are going to remove some columns that are not necessary for the analysis that we'll carry out, like the station and ride ID and the latitude and longitude of the stations. We are going to call this new dataset **data_v2**. Also, we are going to rename some columns:

```
data_v2 <- data.frame(
  "customer" = data$member_casual,
  "bike" = data$rideable_type,
  "started_at" = data$started_at,
  "ended_at" = data$ended_at,
  "start_station" = data$start_station_name,
```

```
  "end_station" = data$end_station_name
 )

glimpse(data_v2)
```

```
## Rows: 5,479,096
## Columns: 6
## $ customer      <chr> "member", "member", "member", "member", "casual", "casua~
## $ bike          <chr> "electric_bike", "electric_bike", "electric_bike", "elec~
## $ started_at    <dttm> 2021-01-23 16:14:19, 2021-01-27 18:43:08, 2021-01-21 22~
## $ ended_at      <dttm> 2021-01-23 16:24:44, 2021-01-27 18:47:12, 2021-01-21 22~
## $ start_station <chr> "California Ave & Cortez St", "California Ave & Cortez S~
## $ end_station   <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, "Wood St & Augusta B~
```

We can see that the data type of each column is the appropiate one. The only modification we will make is to convert customer to factor:

```
data_v2["customer"] <- as.factor(data_v2$customer)
```

Let's check the second column, the bike type:

```
unique(data_v2$bike)
```

```
## [1] "electric_bike" "classic_bike"  "docked_bike"
```

This is an inconsistency, as we can say that docked bike is the same as classic bike, so we are going to fix this:

```
data_v2["bike"] <- replace(data_v2$bike, data_v2$bike=="docked_bike","classic_bike")
```

Now let's take a look to the stations. Let's start replacing the NA's values on the stations columns to the value "Unknown":

```
data_v2["start_station"] <- replace_na(data_v2$start_station,"Unknown")

data_v2["end_station"] <- replace_na(data_v2$end_station,"Unknown")
```

Diving into the unique values of the stations, we can note some inconsistencies:

- Some stations have a *(Temp)* or *(NEXT Apts)*, like Halsted St & 18th St, at the end of the name.

- There are a "DIVVY CASSETTE REPAIR MOBILE STATION" and a "HUBBARD ST BIKE CHECKING (LBS-WH-TEST)" station, that seems to be some company maintenance stations.

  So, in order to fix the inconsistencies, let's remove the company maintenance stations and let's erase the *(Temp)* and *(NEXT Apts)*.

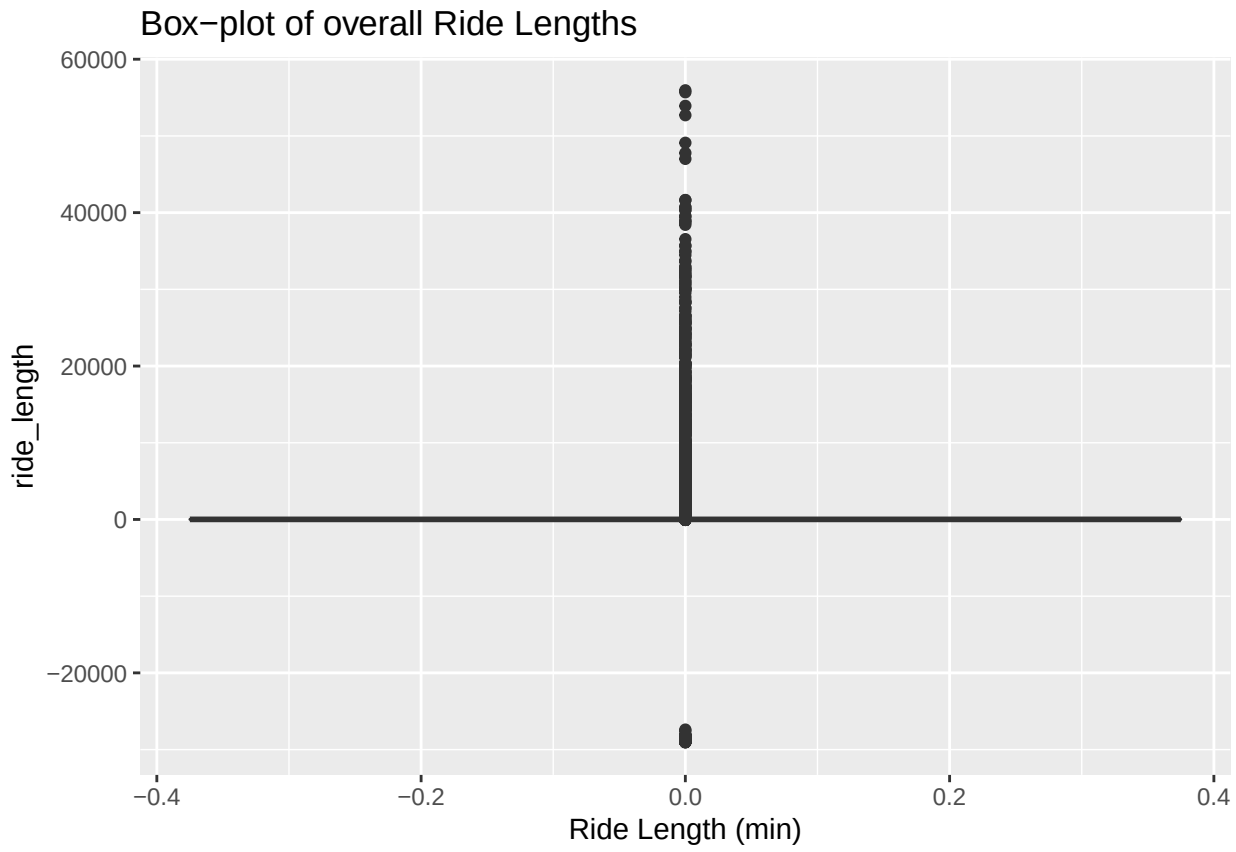
```
data_v2$start_station <- str_replace_all(data_v2$start_station, c("Temp"="","NEXT Apts"=
data_v2$start_station <- str_replace(data_v2$start_station, "\\(\\)", "")

data_v2 <- dplyr::filter(data_v2,start_station!="DIVVY CASSETTE REPAIR MOBILE STATION")
data_v2 <- dplyr::filter(data_v2,start_station!="HUBBARD ST BIKE CHECKING (LBS-WH-TEST)"
data_v2 <- dplyr::filter(data_v2,end_station!="DIVVY CASSETTE REPAIR MOBILE STATION")
data_v2 <- dplyr::filter(data_v2,end_station!="HUBBARD ST BIKE CHECKING (LBS-WH-TEST)")
```

Now let's add a column "ride_length" substracting the start and ending time. It will be expressed in minutes:

```
data_v2["ride_length"] <- as.integer(difftime(data_v2$ended_at,data_v2$started_at,units
```

Let's explore this new data:

```
ggplot(data_v2, aes(y=ride_length))+
  geom_boxplot()+
  labs(x="Ride Length (min)", title="Box-plot of overall Ride Lengths")
```



Quickly, in the last graph we can see that there are some issues with this data, so we are going to check how many outliers we have, and if we'll be able to ignore them

```
sum(data_v2$ride_length<0)
```

```
## [1] 455
```

We have 455 rows that we should ignore: are just a few negative values that represents a very small portion of data that could skew future analysis. Let's take a look to some of them before remove it:

```
head(data_v2[data_v2$ride_length<0,])
```

```
##         customer          bike          started_at            ended_at
## 6593      member  classic_bike 2021-01-09 15:42:45 2021-01-09 15:41:02
## 35753     member electric_bike 2021-01-06 18:33:12 2021-01-06 18:31:07
## 406797    member electric_bike 2021-04-27 17:13:44 2021-04-27 17:11:32
## 1313176   casual  classic_bike 2021-06-15 20:58:03 2021-06-15 20:54:51
## 1493185   member electric_bike 2021-06-02 17:52:32 2021-06-02 17:47:26
## 2900696   casual electric_bike 2021-08-10 09:32:21 2021-08-10 09:31:11
```

```
##                            start_station                      end_station
## 6593    Monticello Ave & Irving Park Rd                          Unknown
## 35753                Daley Center Plaza                          Unknown
## 406797                          Unknown                          Unknown
## 1313176          Broadway & Sheridan Rd          Broadway & Sheridan Rd
## 1493185                         Unknown Southport Ave & Wrightwood Ave
## 2900696           Rush St & Superior St                          Unknown
##         ride_length
## 6593             -1
## 35753            -2
## 406797           -2
## 1313176          -3
## 1493185          -5
## 2900696          -1
```

As we see, there are some errors with the starting and ending time of trip, so we are going to keep only positive values. At this point, is going to be useful to keep the raw data obtained so far, so we are starting a new version of the dataset:

```
data_v3 <- data_v2[data_v2$ride_length>0,]
```

Now lets take a look to the outliers of the dataset:

```
q_m <- quantile(data_v3$ride_length[data_v3$customer=="member"], probs=c(0.25, 0.75))

q_c <- quantile(data_v3$ride_length[data_v3$customer=="casual"], probs=c(0.25, 0.75))

iqr_m <- q_m[2]-q_m[1]
iqr_c <- q_c[2]-q_c[1]

upper_m <- q_m[2]+(1.5*iqr_m)
lower_m <- q_m[1]-(1.5*iqr_m)
upper_c <- q_c[2]+(1.5*iqr_c)
lower_c <- q_c[1]-(1.5*iqr_c)

print(paste("Casual outliers above:", sum(data_v3$ride_length[data_v3$customer=="casual"
```

```
## [1] "Casual outliers above: 223285"
```

```
print(paste("Casual outliers below:", sum(data_v3$ride_length[data_v3$customer=="casual"
```

```
## [1] "Casual outliers below: 0"
```

```
print(paste("Members outliers above:", sum(data_v3$ride_length[data_v3$customer=="member
```

```
## [1] "Members outliers above: 175789"
```

```
print(paste("Members outliers below:", sum(data_v3$ride_length[data_v3$customer=="member
```

```
## [1] "Members outliers below: 0"
```

Let's move the outliers to another data frame:

```
outliers <- rbind(
  dplyr::filter(data_v3, customer=="casual" & ride_length>upper_c),
  dplyr::filter(data_v3, customer=="member" & ride_length>upper_m)
)

data_v3 <- rbind(
  dplyr::filter(data_v3, customer=="casual" & ride_length <= upper_c),
  dplyr::filter(data_v3, customer=="member" & ride_length <= upper_m)
)
```

Now it's time to take a closer look to the lower values.

```
head(data_v3[data_v3$ride_length<2,])
```

```
##     customer          bike          started_at            ended_at
## 319   casual  classic_bike 2021-01-27 23:30:43 2021-01-27 23:32:28
## 463   casual  classic_bike 2021-01-09 23:49:27 2021-01-09 23:51:07
## 792   casual  classic_bike 2021-01-11 07:07:21 2021-01-11 07:08:29
## 806   casual  classic_bike 2021-01-09 12:02:35 2021-01-09 12:03:54
## 807   casual  classic_bike 2021-01-09 12:00:34 2021-01-09 12:01:47
## 872   casual electric_bike 2021-01-25 09:27:13 2021-01-25 09:28:42
##                     start_station                   end_station ride_length
## 319 Sheffield Ave & Fullerton Ave Sheffield Ave & Wrightwood Ave           1
## 463   Milwaukee Ave & Rockwell St California Ave & Milwaukee Ave           1
## 792    Indiana Ave & Roosevelt Rd     Indiana Ave & Roosevelt Rd           1
## 806       Damen Ave & Cortland St        Damen Ave & Cortland St           1
## 807       Damen Ave & Cortland St        Damen Ave & Cortland St           1
## 872 Financial Pl & Ida B Wells Dr  Financial Pl & Ida B Wells Dr           1
```

Here we see that we could have rides like the one in the 3rd row: the trip lasted one minute, and the starting and ending station is the same. This could be because a customer took a bike, regretted his decision, and ended up returning it without using it. These rows can be isolated and dropped looking for trips where the starting and ending stations are the same, and the duration of the trip is less than about 5 minutes:

```
data_v3 %>%
  dplyr::filter(
    start_station == end_station &
    ride_length < 5
  ) %>%
  dim()
```

```
## [1] 121436      7
```

We have 121,436 rides that have these characteristics, and we are going to drop that rows to another data frame:

```
short_trips <- data_v3 %>%
  dplyr::filter(start_station==end_station & ride_length<5)

data_v3 <- data_v3 %>%
  dplyr::filter(!(start_station==end_station & ride_length<5))
```

Finally, let's add two columns for the month, the day of the week, and the ride hour, and save the dataset as the cleaned data:

```
library(lubridate)
```

```
## Warning: package 'lubridate' was built under R version 4.0.5
```

```
Sys.setlocale("LC_TIME", "C")
```

```
## [1] "C"
```

```
data_v3["ride_month"] <- factor(format(data_v3$started_at,"%b"),levels=substr(month.name

data_v3["ride_day"] <- factor(substr(weekdays(data_v3$started_at),1,3), levels=c("Mon","

data_v3["ride_hour"] <- data_v3$started_at %>%
  format(format="%H:%M") %>%
  as.POSIXct(format="%H:%M")

#write.csv(data_v3, "data_v3.csv", row.names=FALSE)
```

Now we have all we need to the analyze phase:

```
glimpse(data_v3)
```

```
## Rows: 4,875,175
## Columns: 10
## $ customer      <fct> casual, casual, casual, casual, casual, casual, casual, ~
## $ bike          <chr> "electric_bike", "electric_bike", "electric_bike", "clas~
## $ started_at    <dttm> 2021-01-09 14:24:07, 2021-01-09 15:28:04, 2021-01-09 15~
## $ ended_at      <dttm> 2021-01-09 15:17:54, 2021-01-09 15:37:51, 2021-01-09 15~
## $ start_station <chr> "California Ave & Cortez St", "California Ave & Cortez S~
## $ end_station   <chr> "Unknown", "Wood St & Augusta Blvd", "Wood St & Augusta ~
## $ ride_length   <int> 53, 9, 8, 10, 7, 7, 7, 8, 10, 14, 17, 19, 20, 27, 19, 24~
## $ ride_month    <fct> Jan, Jan, Jan, Jan, Jan, Jan, Jan, Jan, Jan, Jan, Jan, J~
## $ ride_day      <fct> Sat, Sat, Sat, Sun, Sun, Mon, Thu, Mon, Fri, Sat, Wed, F~
## $ ride_hour     <dttm> 2022-01-31 14:24:00, 2022-01-31 15:28:00, 2022-01-31 15~
```

Analizing the data

In this section, we are going to try to characterize each type of customer, looking for differences in their respective ride trips.

What is the average travel duration for each type of customer?

The average ride length for each customer type is:

```
print(paste("Mean for Casuals: ", round(mean(data_v3$ride_length[data_v3$customer=="casu
```

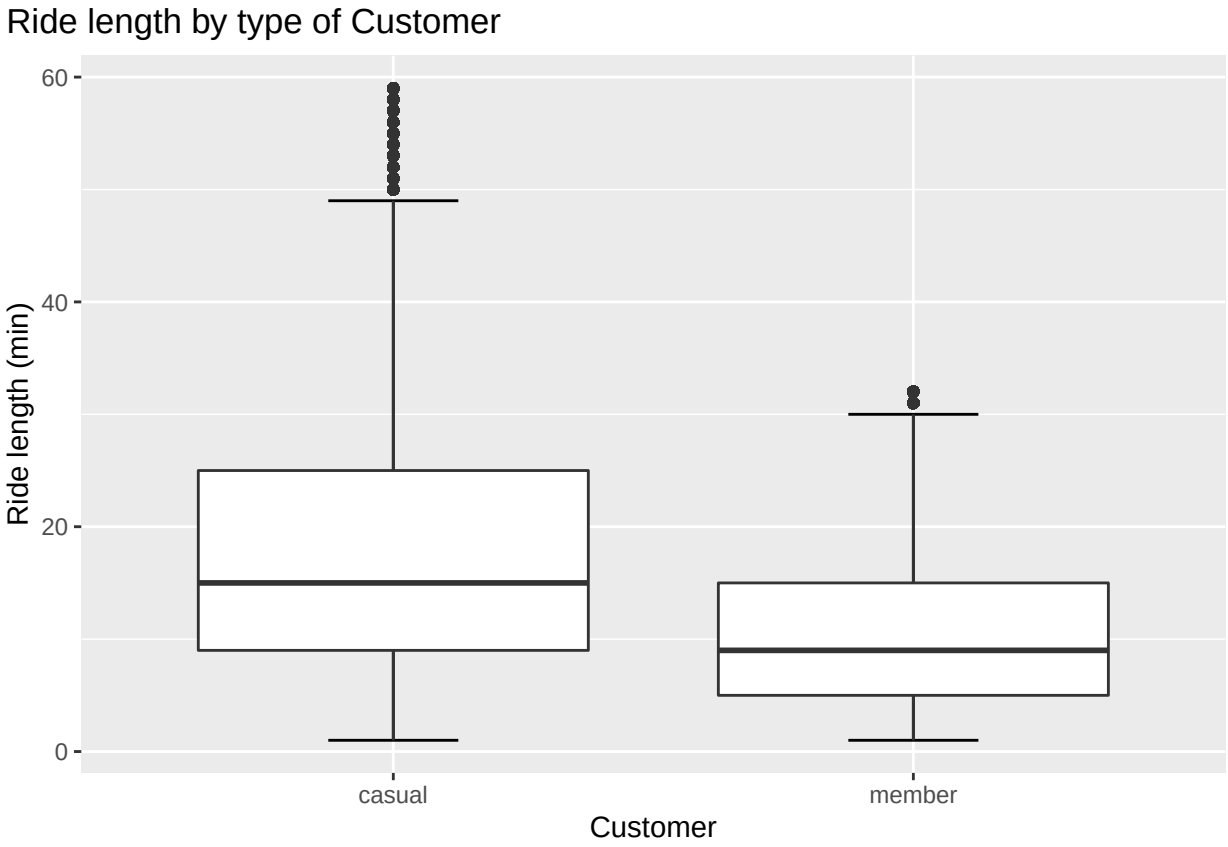
```
## [1] "Mean for Casuals:  18  min"
```

```
print(paste("Mean for Members: ", round(mean(data_v3$ride_length[data_v3$customer=="memb
```

```
## [1] "Mean for Members:  11  min"
```

The first impression with these results is that **casuals riders tend to take longer trips**. This could be explained by claiming that casual riders could use bicycles to travel around the city as tourism, while members could use them to more precise things, like go to work. Let's plot this data:

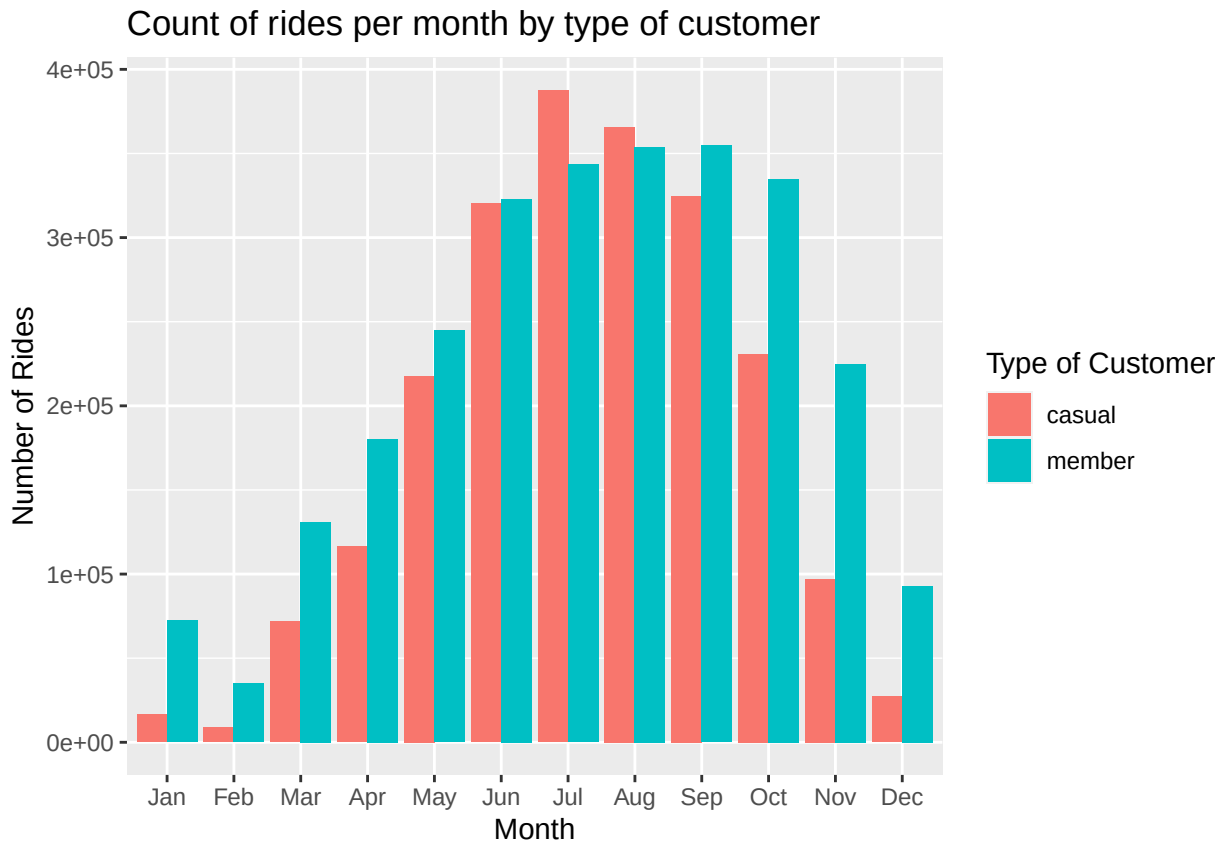
```
ggplot(data=data_v3, aes(x=customer,y=ride_length)) +
  stat_boxplot(geom="errorbar",width=0.25) +
  geom_boxplot() +
  labs(x="Customer", y="Ride length (min)", title="Ride length by type of Customer") +
  theme(plot.title.position = "plot")
```



Are there any seasonal trend in customer demand for bycicles?

Let's check if there are some seasonal trend in the number of rides:

```
ggplot(data=data_v3) +
  geom_bar(mapping=aes(x=ride_month,fill=customer), position="dodge") +
  labs(x="Month", y="Number of Rides", title="Count of rides per month by type of custom
```

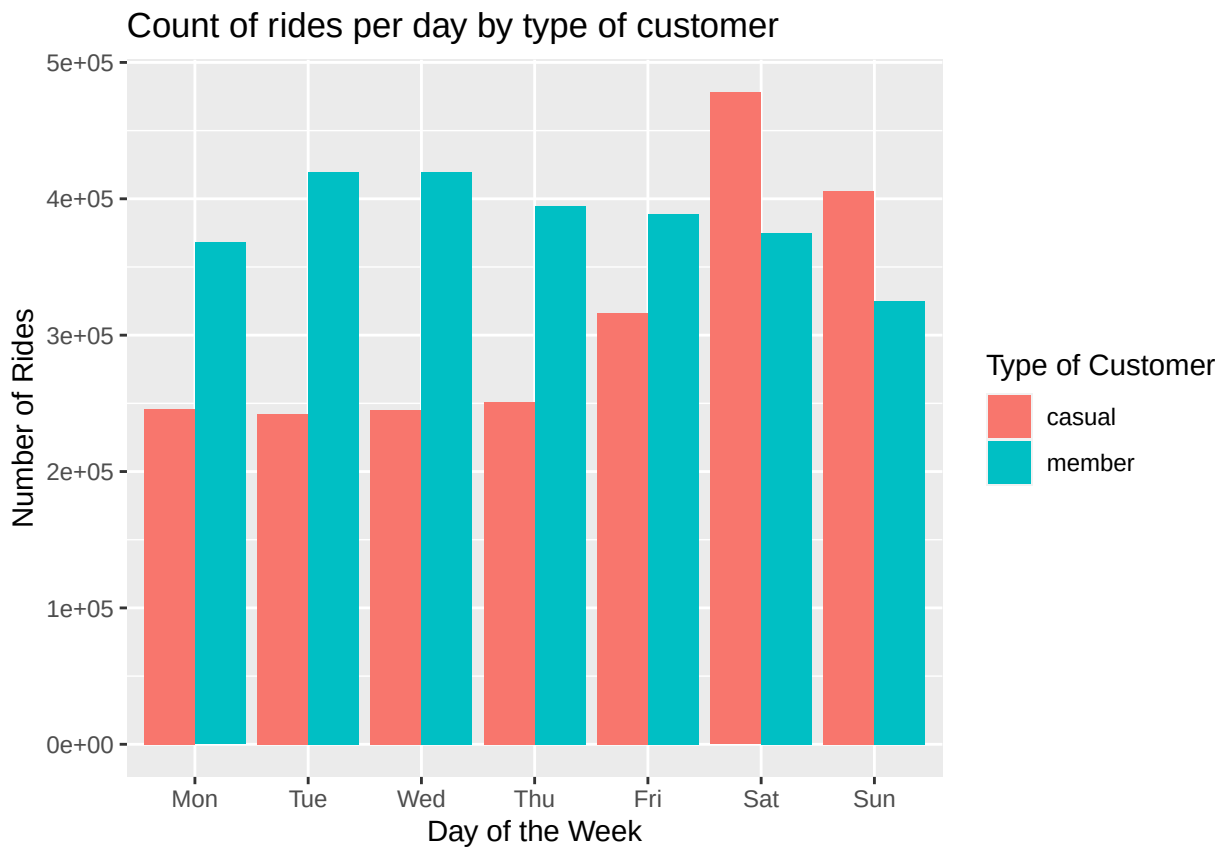


In the graph we can see a general trend towards **fewer trips in the winter months**. This can be attributed to the slightly more adverse environmental climatic conditions of these months, which may make customers prefer other means of transport. In relation to casual riders, we can see that the peak of demand is between the months of June and September, something that may be related to summer vacations, a phenomenon that can attract more tourists to use bicycles to take walks around the city. town. For members, the peak is not so marked and includes a longer period of time, being able to consider the months between May and November. Here, the greater and lesser demand for bicycles could be attributed only to a question of seasons that favor or adversely affect outdoor transport.

Is there a weekly trend in customer demand for bicycles?

Now, let's apply the same question concept to the week:

```
ggplot(data=data_v3) +
  geom_bar(mapping=aes(x=ride_day,fill=customer), position="dodge") +
  labs(x="Day of the Week", y="Number of Rides", title="Count of rides per day by type o
```



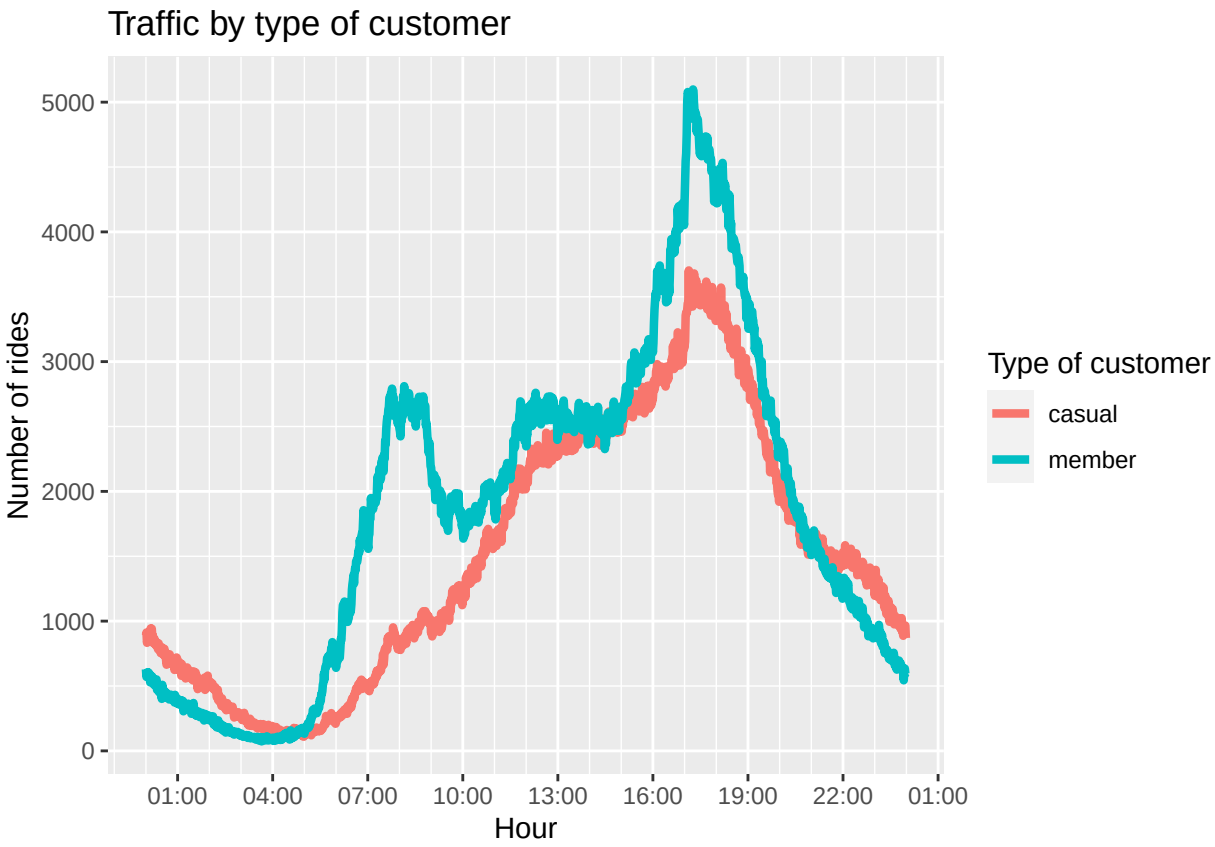
As in the previous case, we see a marked **trend in the demand for casual riders on weekends**, which may gradually confirm the idea of the tourist use that these customers give the product. The contrast of this is the almost constant demand for members, which even seems to decline slightly on the weekends.

Are there peak times in the demand for bicycles by customers?

Let's see what time of the day prefers each customer:

```
times_m <- data.frame(table(data_v3$ride_hour[data_v3$customer=="member"]))
colnames(times_m) <- c("hour","count")
times_m$customer <- rep("member",dim(times_m)[1])

times_c <- data.frame(table(data_v3$ride_hour[data_v3$customer=="casual"]))
colnames(times_c) <- c("hour","count")
times_c$customer <- rep("casual",dim(times_m)[1])

times <- rbind(times_m,times_c)
times$hour <- as.POSIXct(times$hour)

ggplot(data=times) +
```

```
  geom_line(mapping=aes(x=hour,y=count, color=customer, group=customer), size=1.5) +
  scale_x_datetime(date_breaks="3 hours", date_labels = "%H:%M", minor_breaks = "1 hour"
  labs(x="Hour", y="Number of rides", title="Traffic by type of customer", color="Type o
```



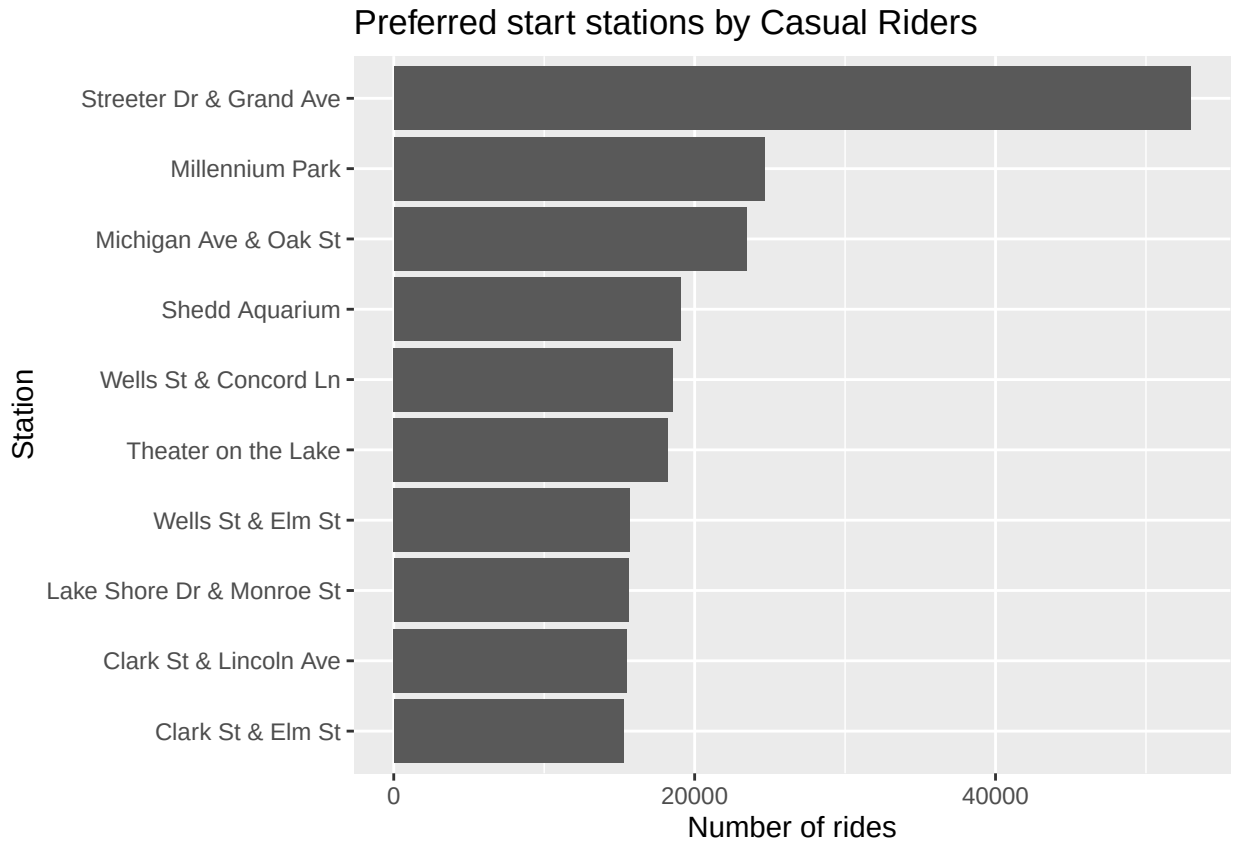
In the graph we can distinguish two key behaviors for each customer. On the one hand, the demand for bicycles by members has two marked peaks: one between 8 and 9am, and another between 5 and 7pm. This behavior can be attributed to a significant demand on the part of these clients in commuting schedules to and from work. On the other hand, the demand from casual customers grows gradually, taking on significant values after noon and reaching its **peak between 5:00 p.m. and 6:00 p.m.** This schedule may be the preferred time for this type of client to go on bike rides around the city.

What are the preferred stations for casual riders?

Now we will try to find the stations with the highest traffic from casual riders:

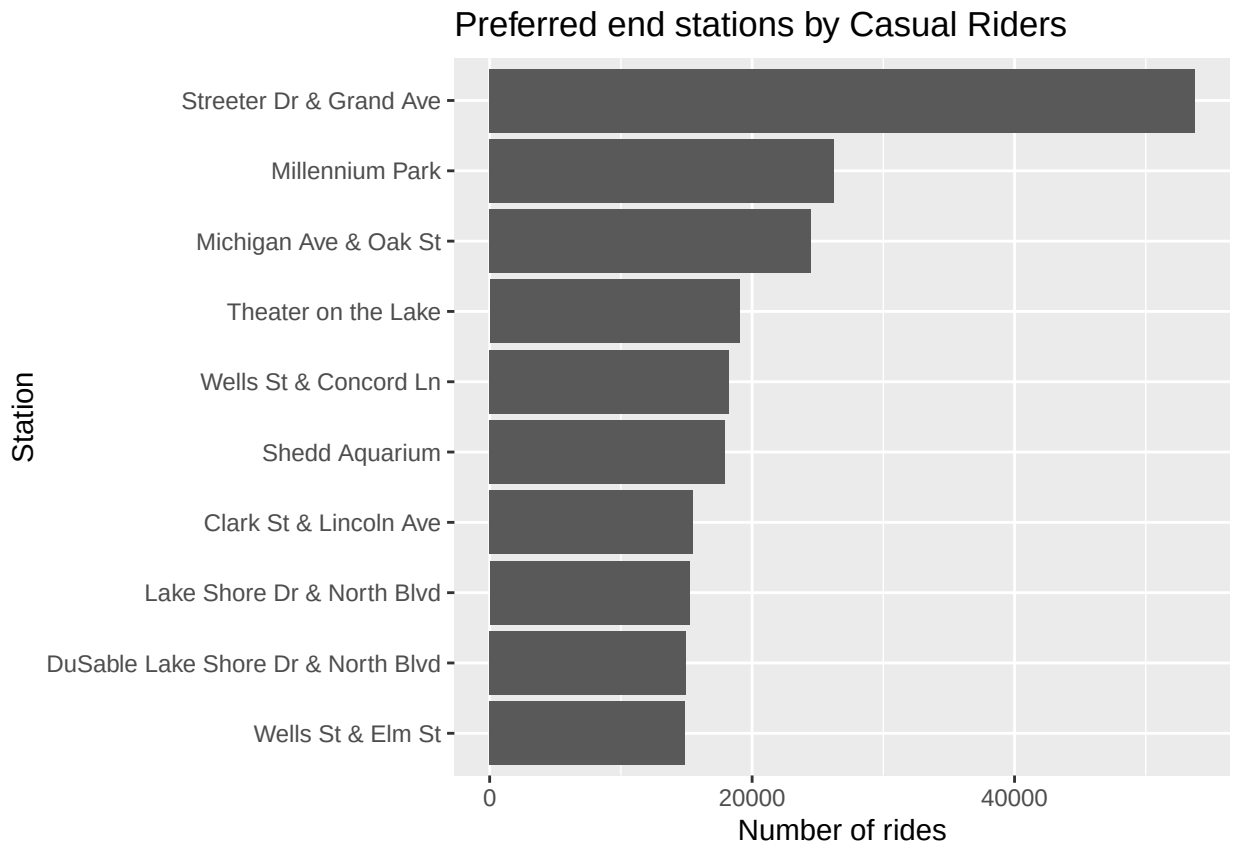
```
start_stations <- table(data_v3$start_station,data_v3$customer) %>%
  data.frame() %>%
  dplyr::filter(Var1!="Unknown" & Var2=="casual") %>%
  arrange(-Freq) %>%
```

```
  head(10)

colnames(start_stations) <- c("station","customer","count")

end_stations <- table(data_v3$end_station,data_v3$customer) %>%
  data.frame() %>%
  dplyr::filter(Var1!="Unknown" & Var2=="casual") %>%
  arrange(-Freq) %>%
  head(10)

colnames(end_stations) <- c("station","customer","count")


ggplot(data=start_stations) +
  geom_col(mapping=aes(x=reorder(station,count),y=count)) +
  labs(x="Station", y="Number of rides", title="Preferred start stations by Casual Rider
  coord_flip()
```




```
ggplot(data=end_stations) +
  geom_col(mapping=aes(x=reorder(station,count),y=count)) +
```

```
  labs(x="Station", y="Number of rides", title="Preferred end stations by Casual Riders"
  coord_flip()
```

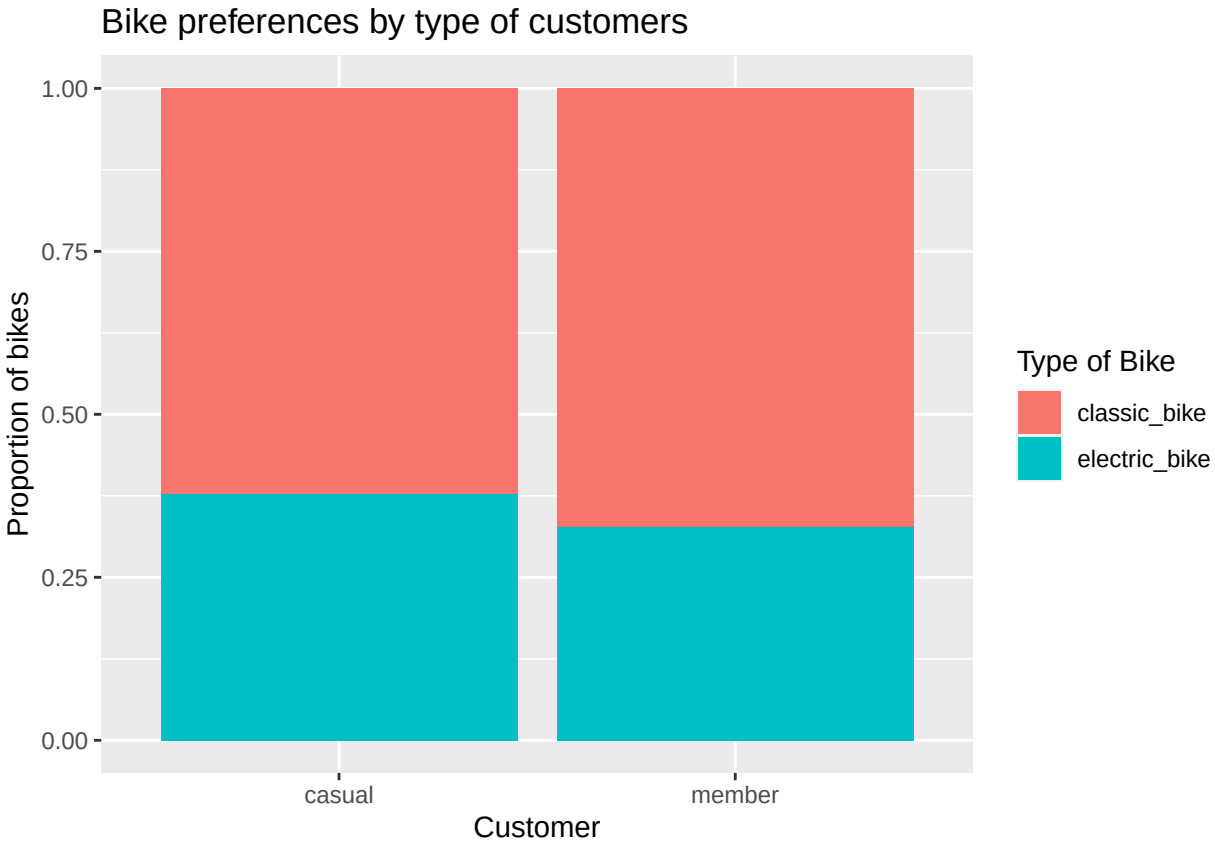


These two graphs allow us to continue supporting the theory of the recreational use of bicycles by casual riders, since **some of the main stations from which trips depart and to which trips arrive are tourist spots in the city**, such as Millennium Park, Theater on the Lake or Shedd Aquarium.

Do casual riders prefer electric or regular bikes?

In the last question, we are going to see if are there any preference on choosing the bike:

```
ggplot(data=data_v3) +
  geom_bar(mapping=aes(x=customer, fill=bike), position="fill")+
  labs(y="Proportion of bikes", x="Customer", fill="Type of Bike", title="Bike preferenc
```



In this case, we can say that there is **no differential behavior in the choice of the type of bicycle** by both clients. In both cases there is a preference for classic bicycles, but there are no significant differences between them.

Conclusions

In this project, the analysis of twelve months of information on bicycle trips from the Cyclistic company was carried out, with the aim of evaluating how annual members and casual riders use bicycles differently. The dataset used is located in the company's cloud, and meets the minimum reliability and utility requirements to answer the initial question posed.

The differences in the user profile of casual riders are centered on the recreational and tourism use they give to bicycles, and the insights found in the different stages of analysis support this idea:

- Casual riders tend to take somewhat longer trips. This can be explained by attributing these trips to tourism for these users, while the annual members could use the bicycles for more precise purposes such as traveling to and from work.
- The demand for casual riders increases on weekends. This is consistent with recreational use of bicycles.

- After noon is the time where the demand for casual riders increases, with a peak at 5pm. In contrast, the members present two peaks marked in the morning and in the afternoon, also attributable to use as transportation to work.

- The main stations preferred by casual riders seem to be tourist spots.

  On the other hand, no differential behaviors were observed in the choice of the type of bicycle, nor in monthly trends. Both customers have roughly the same preference for types of bicycles, and the demand by both customers increases in the months between May and October, probably due to weather issues.

  As the profile of casual riders is aimed at tourism, this is the key point to emphasize when attracting these customers to an annual membership. Some suggestions that could contribute to this transition could be:

- Offer different thematic tours around the city of Chicago at different times of the year, visiting the main tourist centers. Establishing these recreational activities on a regular basis could attract casual riders to sign up for a monthly membership.

- Establish relationships with local food and souvenirs present in major Chicago tourist centers to try to establish discounts for annual members of Cyclistics.

- Offer tours at night through the main bars and restaurants in the city, showing the "other side of Chicago", with free drinks and / or discounts on them.

  These are some of the ideas that we think could attract casual riders to pay an annual membership. While many more ideas may emerge, we believe it is crucial that they focus on **exploiting Chicago city tourism on Cyclistics bikes**.