Estudo de caso - Cyclistic

Análise por: Hiêgor Barreto Rodrigues

Data: 20/05/2021

Dados: https://divvy-tripdata.s3.amazonaws.com/index.html

Licença: https://www.divvybikes.com/data-license-agreement

Projeto

O projeto é solicitação da empresa fictícia Cyclistic, localizada em Chicago. A diretora de marketing acredita que o futuro sucesso da empresa depende da maximização do número de membros anuais. Para isso, será realizada uma análise para entender a diferença da utilização das bicicletas da Cyclistic por usuários casuais e os membros anuais.

A partir desses insights, o time de marketing irá desenvolver uma nova estratégia para converter os usuários casuais em membros anuais. Mas, para isso acontecer, os insights devem estar bem apoiados nos dados e visualizações.

Para dar contexto ao projeto, algumas informações se fazem necessárias. Os analistas financeiros da Cyclistic concluíram que os membros anuais são muito mais lucrativos para a empresa do que os usuários casuais, apesar da flexibilidade dos preços atraírem mais clientes. Com isso, a diretora de marketing, Lily Moreno, acredita que existe uma grande chande de converter os usuários casuais em membros, pois eles já conhecem o serviço da Cyclistic e já escolheram a empresa como solução de mobilidade.

Assim, a questão a ser respondida é: **Qual a diferença da utilização das bicicletas da Cyclistic entre os membros anuais e usuários casuais?**

A análise será dividida em 6 etapas:

1. Perguntar
2. Preparar
3. Processar
4. Analisar
5. Compartilhar
6. Agir

1. Perguntar

A Cyclistic necessita dos dados para embasar a decisão para a campanha. Os dados nos ajudarão a entender se é mais eficiente fazer uma campanha geral ou uma campanha direcionada aos usuários casuais, visando ajudar a aumentar a lucratividade da empresa através da maximização do número de membros anuais.

Por isso, buscaremos entender melhor como os membros anuais diferem dos usuários casuais e porque eles são mais lucrativos para a empresa.

Informaremos os resultados da análise à Lily Moreno (diretora de marketing), o time de analista de marketing e time executivo da Cyclistic. É esperado que a análise de dados e os insights gerados, sejam o apoio necessário para lançar a campanha da maneira mais eficiente, possibilitando a tomada de decisão inteligente e baseada em dados.

2. Preparar

Começamos o trabalho no RStudio importando 4 bibliotecas que serão muito úteis para a análise: tidyverse, ggplot, readr e lubridate.

```
library(tidyverse)
library(readr)
library(lubridate)
library(ggplot2)
```

Após isso, importamos os nossos dados para o R, para fazer a limpeza.

```
getwd()
```

```
## [1] "C:/Users/hieeg/Documents/Cursos/Google Data Analyst/Curso 8 - Estudos de caso/Estudo de caso 1 - Cyclisti
c/Dados originais"
```

```
setwd("/Users/hieeg/Documents/Cursos/Google Data Analyst/Curso 8 - Estudos de caso/Estudo de caso 1 - Cyclistic/D
ados originais")
mai_2020 <- read.csv("cyclistic_tripdata_052020.csv")
jun_2020 <- read.csv("cyclistic_tripdata_062020.csv")
jul_2020 <- read.csv("cyclistic_tripdata_072020.csv")
ago_2020 <- read.csv("cyclistic_tripdata_082020.csv")
set_2020 <- read.csv("cyclistic_tripdata_092020.csv")
out_2020 <- read.csv("cyclistic_tripdata_102020.csv")
nov_2020 <- read.csv("cyclistic_tripdata_112020.csv")
dez_2020 <- read.csv("cyclistic_tripdata_122020.csv")
jan_2021 <- read.csv("cyclistic_tripdata_012021.csv")
fev_2021 <- read.csv("cyclistic_tripdata_022021.csv")
mar_2021 <- read.csv("cyclistic_tripdata_032021.csv")
abr_2021 <- read.csv("cyclistic_tripdata_042021.csv")
```

Com os dados importados para o R, conferimos se todas as tabelas possuem os mesmos nomes nas colunas e os mesmos tipos de variável para não ter incompatibilidade na união.

```
colnames(mai_2020)
```

```
##  [1] "ride_id"            "rideable_type"      "started_at"
##  [4] "ended_at"           "start_station_name" "start_station_id"
##  [7] "end_station_name"   "end_station_id"     "start_lat"
## [10] "start_lng"          "end_lat"            "end_lng"
## [13] "member_casual"
```

```
colnames(jun_2020)
```

```
##  [1] "ride_id"            "rideable_type"      "started_at"
##  [4] "ended_at"           "start_station_name" "start_station_id"
##  [7] "end_station_name"   "end_station_id"     "start_lat"
## [10] "start_lng"          "end_lat"            "end_lng"
## [13] "member_casual"
```

```
colnames(jul_2020)
```

```
##  [1] "ride_id"            "rideable_type"      "started_at"
##  [4] "ended_at"           "start_station_name" "start_station_id"
##  [7] "end_station_name"   "end_station_id"     "start_lat"
## [10] "start_lng"          "end_lat"            "end_lng"
## [13] "member_casual"
```

```
colnames(ago_2020)
```

```
##  [1] "ride_id"            "rideable_type"      "started_at"
##  [4] "ended_at"           "start_station_name" "start_station_id"
##  [7] "end_station_name"   "end_station_id"     "start_lat"
## [10] "start_lng"          "end_lat"            "end_lng"
## [13] "member_casual"
```

```
colnames(set_2020)
```

```
##  [1] "ride_id"            "rideable_type"      "started_at"
##  [4] "ended_at"           "start_station_name" "start_station_id"
##  [7] "end_station_name"   "end_station_id"     "start_lat"
## [10] "start_lng"          "end_lat"            "end_lng"
## [13] "member_casual"
```

```
colnames(out_2020)
```

```
##  [1] "ride_id"            "rideable_type"      "started_at"
##  [4] "ended_at"           "start_station_name" "start_station_id"
##  [7] "end_station_name"   "end_station_id"     "start_lat"
## [10] "start_lng"          "end_lat"            "end_lng"
## [13] "member_casual"
```

```
colnames(nov_2020)
```

```
##  [1] "ride_id"            "rideable_type"      "started_at"
##  [4] "ended_at"           "start_station_name" "start_station_id"
##  [7] "end_station_name"   "end_station_id"     "start_lat"
## [10] "start_lng"          "end_lat"            "end_lng"
## [13] "member_casual"
```

```
colnames(dez_2020)
```

```
##  [1] "ride_id"            "rideable_type"      "started_at"
##  [4] "ended_at"           "start_station_name" "start_station_id"
##  [7] "end_station_name"   "end_station_id"     "start_lat"
## [10] "start_lng"          "end_lat"            "end_lng"
## [13] "member_casual"
```

```
colnames(jan_2021)
```

```
##  [1] "ride_id"            "rideable_type"      "started_at"
##  [4] "ended_at"           "start_station_name" "start_station_id"
##  [7] "end_station_name"   "end_station_id"     "start_lat"
## [10] "start_lng"          "end_lat"            "end_lng"
## [13] "member_casual"
```

```
colnames(fev_2021)
```

```
##  [1] "ride_id"            "rideable_type"      "started_at"
##  [4] "ended_at"           "start_station_name" "start_station_id"
##  [7] "end_station_name"   "end_station_id"     "start_lat"
## [10] "start_lng"          "end_lat"            "end_lng"
## [13] "member_casual"
```

```
colnames(mar_2021)
```

```
##  [1] "ride_id"            "rideable_type"      "started_at"
##  [4] "ended_at"           "start_station_name" "start_station_id"
##  [7] "end_station_name"   "end_station_id"     "start_lat"
## [10] "start_lng"          "end_lat"            "end_lng"
## [13] "member_casual"
```

```
colnames(abr_2021)
```

```
##  [1] "ride_id"            "rideable_type"      "started_at"
##  [4] "ended_at"           "start_station_name" "start_station_id"
##  [7] "end_station_name"   "end_station_id"     "start_lat"
## [10] "start_lng"          "end_lat"            "end_lng"
## [13] "member_casual"
```

```
str(mai_2020)
```

```
## 'data.frame':    200274 obs. of  13 variables:
##  $ ride_id           : chr  "02668AD35674B983" "7A50CCAF1EDDB28F" "2FFCDFDB91FE9A52" "58991CF1DB75BA84" ...
##  $ rideable_type     : chr  "docked_bike" "docked_bike" "docked_bike" "docked_bike" ...
##  $ started_at        : chr  "2020-05-27 10:03:52" "2020-05-25 10:47:11" "2020-05-02 14:11:03" "2020-05-02 16:2
5:36" ...
##  $ ended_at          : chr  "2020-05-27 10:16:49" "2020-05-25 11:05:40" "2020-05-02 15:48:21" "2020-05-02 16:3
9:28" ...
##  $ start_station_name: chr  "Franklin St & Jackson Blvd" "Clark St & Wrightwood Ave" "Kedzie Ave & Milwaukee A
ve" "Clarendon Ave & Leland Ave" ...
##  $ start_station_id  : int  36 340 260 251 261 206 261 180 331 219 ...
##  $ end_station_name  : chr  "Wabash Ave & Grand Ave" "Clark St & Leland Ave" "Kedzie Ave & Milwaukee Ave" "Lak
e Shore Dr & Wellington Ave" ...
##  $ end_station_id    : int  199 326 260 157 206 22 261 180 300 305 ...
##  $ start_lat         : num  41.9 41.9 41.9 42 41.9 ...
##  $ start_lng         : num  -87.6 -87.6 -87.7 -87.7 -87.7 ...
##  $ end_lat           : num  41.9 42 41.9 41.9 41.8 ...
##  $ end_lng           : num  -87.6 -87.7 -87.7 -87.6 -87.6 ...
##  $ member_casual     : chr  "member" "casual" "casual" "casual" ...
```

```
str(jun_2020)
```

```
## 'data.frame':    343005 obs. of  13 variables:
##  $ ride_id           : chr  "8CD5DE2C2B6C4CFC" "9A191EB2C751D85D" "F37D14B0B5659BCF" "C41237B506E85FA1" ...
##  $ rideable_type     : chr  "docked_bike" "docked_bike" "docked_bike" "docked_bike" ...
##  $ started_at        : chr  "2020-06-13 23:24:48" "2020-06-26 07:26:10" "2020-06-23 17:12:41" "2020-06-20 01:0
9:35" ...
##  $ ended_at          : chr  "2020-06-13 23:36:55" "2020-06-26 07:31:58" "2020-06-23 17:21:14" "2020-06-20 01:2
8:24" ...
##  $ start_station_name: chr  "Wilton Ave & Belmont Ave" "Federal St & Polk St" "Daley Center Plaza" "Broadway &
Cornelia Ave" ...
##  $ start_station_id  : int  117 41 81 303 327 327 41 115 338 84 ...
##  $ end_station_name  : chr  "Damen Ave & Clybourn Ave" "Daley Center Plaza" "State St & Harrison St" "Broadway
& Berwyn Ave" ...
##  $ end_station_id    : int  163 81 5 294 117 117 81 303 164 53 ...
##  $ start_lat         : num  41.9 41.9 41.9 41.9 41.9 ...
##  $ start_lng         : num  -87.7 -87.6 -87.6 -87.6 -87.7 ...
##  $ end_lat           : num  41.9 41.9 41.9 42 41.9 ...
##  $ end_lng           : num  -87.7 -87.6 -87.6 -87.7 -87.7 ...
##  $ member_casual     : chr  "casual" "member" "member" "casual" ...
```

```
str(jul_2020)
```

```
## 'data.frame':    551480 obs. of  13 variables:
##  $ ride_id           : chr  "762198876D69004D" "BEC9C9FBA0D4CF1B" "D2FD8EA432C77EC1" "54AE594E20B35881" ...
##  $ rideable_type     : chr  "docked_bike" "docked_bike" "docked_bike" "docked_bike" ...
##  $ started_at        : chr  "2020-07-09 15:22:02" "2020-07-24 23:56:30" "2020-07-08 19:49:07" "2020-07-17 19:0
6:42" ...
##  $ ended_at          : chr  "2020-07-09 15:25:52" "2020-07-25 00:20:17" "2020-07-08 19:56:22" "2020-07-17 19:2
7:38" ...
##  $ start_station_name: chr  "Ritchie Ct & Banks St" "Halsted St & Roscoe St" "Lake Shore Dr & Diversey Pkwy"
"LaSalle St & Illinois St" ...
##  $ start_station_id  : int  180 299 329 181 268 635 113 211 176 31 ...
##  $ end_station_name  : chr  "Wells St & Evergreen Ave" "Broadway & Ridge Ave" "Clark St & Wellington Ave" "Cla
rk St & Armitage Ave" ...
##  $ end_station_id    : int  291 461 156 94 301 289 140 31 191 142 ...
##  $ start_lat         : num  41.9 41.9 41.9 41.9 41.9 ...
##  $ start_lng         : num  -87.6 -87.6 -87.6 -87.6 -87.6 ...
##  $ end_lat           : num  41.9 42 41.9 41.9 41.9 ...
##  $ end_lng           : num  -87.6 -87.7 -87.6 -87.6 -87.6 ...
##  $ member_casual     : chr  "member" "member" "casual" "casual" ...
```

```
str(ago_2020)
```

```
## 'data.frame':    622361 obs. of  13 variables:
##  $ ride_id           : chr  "322BD23D287743ED" "2A3AEF1AB9054D8B" "67DC1D133E8B5816" "C79FBBD412E578A7" ...
##  $ rideable_type     : chr  "docked_bike" "electric_bike" "electric_bike" "electric_bike" ...
##  $ started_at        : chr  "2020-08-20 18:08:14" "2020-08-27 18:46:04" "2020-08-26 19:44:14" "2020-08-27 12:0
5:41" ...
##  $ ended_at          : chr  "2020-08-20 18:17:51" "2020-08-27 19:54:51" "2020-08-26 21:53:07" "2020-08-27 12:5
3:45" ...
##  $ start_station_name: chr  "Lake Shore Dr & Diversey Pkwy" "Michigan Ave & 14th St" "Columbus Dr & Randolph S
t" "Daley Center Plaza" ...
##  $ start_station_id  : int  329 168 195 81 658 658 196 67 153 177 ...
##  $ end_station_name  : chr  "Clark St & Lincoln Ave" "Michigan Ave & 14th St" "State St & Randolph St" "State
St & Kinzie St" ...
##  $ end_station_id    : int  141 168 44 47 658 658 49 229 225 305 ...
##  $ start_lat         : num  41.9 41.9 41.9 41.9 41.9 ...
##  $ start_lng         : num  -87.6 -87.6 -87.6 -87.6 -87.7 ...
##  $ end_lat           : num  41.9 41.9 41.9 41.9 41.9 ...
##  $ end_lng           : num  -87.6 -87.6 -87.6 -87.6 -87.7 ...
##  $ member_casual     : chr  "member" "casual" "casual" "casual" ...
```

```
str(set_2020)
```

```
## 'data.frame':    532958 obs. of  13 variables:
##  $ ride_id           : chr  "2B22BD5F95FB2629" "A7FB70B4AFC6CAF2" "86057FA01BAC778E" "57F6DC9A153DB98C" ...
##  $ rideable_type     : chr  "electric_bike" "electric_bike" "electric_bike" "electric_bike" ...
##  $ started_at        : chr  "2020-09-17 14:27:11" "2020-09-17 15:07:31" "2020-09-17 15:09:04" "2020-09-17 18:1
0:46" ...
##  $ ended_at          : chr  "2020-09-17 14:44:24" "2020-09-17 15:07:45" "2020-09-17 15:09:35" "2020-09-17 18:3
5:49" ...
##  $ start_station_name: chr  "Michigan Ave & Lake St" "W Oakdale Ave & N Broadway" "W Oakdale Ave & N Broadway"
"Ashland Ave & Belle Plaine Ave" ...
##  $ start_station_id  : int  52 NA NA 246 24 94 291 NA NA NA ...
##  $ end_station_name  : chr  "Green St & Randolph St" "W Oakdale Ave & N Broadway" "W Oakdale Ave & N Broadway"
"Montrose Harbor" ...
##  $ end_station_id    : int  112 NA NA 249 24 NA 256 NA NA NA ...
##  $ start_lat         : num  41.9 41.9 41.9 42 41.9 ...
##  $ start_lng         : num  -87.6 -87.6 -87.6 -87.7 -87.6 ...
##  $ end_lat           : num  41.9 41.9 41.9 42 41.9 ...
##  $ end_lng           : num  -87.6 -87.6 -87.6 -87.6 -87.6 ...
##  $ member_casual     : chr  "casual" "casual" "casual" "casual" ...
```

```
str(out_2020)
```

```
## 'data.frame':    388653 obs. of  13 variables:
##  $ ride_id           : chr  "ACB6B40CF5B9044C" "DF450C72FD109C01" "B6396B54A15AC0DF" "44A4AEE261B9E854" ...
##  $ rideable_type     : chr  "electric_bike" "electric_bike" "electric_bike" "electric_bike" ...
##  $ started_at        : chr  "2020-10-31 19:39:43" "2020-10-31 23:50:08" "2020-10-31 23:00:01" "2020-10-31 22:1
6:43" ...
##  $ ended_at          : chr  "2020-10-31 19:57:12" "2020-11-01 00:04:16" "2020-10-31 23:08:22" "2020-10-31 22:1
9:35" ...
##  $ start_station_name: chr  "Lakeview Ave & Fullerton Pkwy" "Southport Ave & Waveland Ave" "Stony Island Ave &
67th St" "Clark St & Grace St" ...
##  $ start_station_id  : int  313 227 102 165 190 359 313 125 NA 174 ...
##  $ end_station_name  : chr  "Rush St & Hubbard St" "Kedzie Ave & Milwaukee Ave" "University Ave & 57th St" "Br
oadway & Sheridan Rd" ...
##  $ end_station_id    : int  125 260 423 256 185 53 125 313 199 635 ...
##  $ start_lat         : num  41.9 41.9 41.8 42 41.9 ...
##  $ start_lng         : num  -87.6 -87.7 -87.6 -87.7 -87.7 ...
##  $ end_lat           : num  41.9 41.9 41.8 42 41.9 ...
##  $ end_lng           : num  -87.6 -87.7 -87.6 -87.7 -87.7 ...
##  $ member_casual     : chr  "casual" "casual" "casual" "casual" ...
```

```
str(nov_2020)
```

```
## 'data.frame':    259716 obs. of  13 variables:
##  $ ride_id           : chr  "BD0A6FF6FFF9B921" "96A7A7A4BDE4F82D" "C61526D06582BDC5" "E533E89C32080B9E" ...
##  $ rideable_type     : chr  "electric_bike" "electric_bike" "electric_bike" "electric_bike" ...
##  $ started_at        : chr  "2020-11-01 13:36:00" "2020-11-01 10:03:26" "2020-11-01 00:34:05" "2020-11-01 00:4
5:16" ...
##  $ ended_at          : chr  "2020-11-01 13:45:40" "2020-11-01 10:14:45" "2020-11-01 01:03:06" "2020-11-01 00:5
4:31" ...
##  $ start_station_name: chr  "Dearborn St & Erie St" "Franklin St & Illinois St" "Lake Shore Dr & Monroe St" "L
eavitt St & Chicago Ave" ...
##  $ start_station_id  : int  110 672 76 659 2 72 76 NA 58 394 ...
##  $ end_station_name  : chr  "St. Clair St & Erie St" "Noble St & Milwaukee Ave" "Federal St & Polk St" "Stave
St & Armitage Ave" ...
##  $ end_station_id    : int  211 29 41 185 2 76 72 NA 288 273 ...
##  $ start_lat         : num  41.9 41.9 41.9 41.9 41.9 ...
##  $ start_lng         : num  -87.6 -87.6 -87.6 -87.7 -87.6 ...
##  $ end_lat           : num  41.9 41.9 41.9 41.9 41.9 ...
##  $ end_lng           : num  -87.6 -87.7 -87.6 -87.7 -87.6 ...
##  $ member_casual     : chr  "casual" "casual" "casual" "casual" ...
```

```
str(dez_2020)
```

```
## 'data.frame':    131573 obs. of  13 variables:
##  $ ride_id           : chr  "70B6A9A437D4C30D" "158A465D4E74C54A" "5262016E0F1F2F9A" "BE119628E44F871E" ...
##  $ rideable_type     : chr  "classic_bike" "electric_bike" "electric_bike" "electric_bike" ...
##  $ started_at        : chr  "2020-12-27 12:44:29" "2020-12-18 17:37:15" "2020-12-15 15:04:33" "2020-12-15 15:5
4:18" ...
##  $ ended_at          : chr  "2020-12-27 12:55:06" "2020-12-18 17:44:19" "2020-12-15 15:11:28" "2020-12-15 16:0
0:11" ...
##  $ start_station_name: chr  "Aberdeen St & Jackson Blvd" "" "" "" ...
##  $ start_station_id  : chr  "13157" "" "" "" ...
##  $ end_station_name  : chr  "Desplaines St & Kinzie St" "" "" "" ...
##  $ end_station_id    : chr  "TA1306000003" "" "" "" ...
##  $ start_lat         : num  41.9 41.9 41.9 41.9 41.8 ...
##  $ start_lng         : num  -87.7 -87.7 -87.7 -87.7 -87.6 ...
##  $ end_lat           : num  41.9 41.9 41.9 41.9 41.8 ...
##  $ end_lng           : num  -87.6 -87.7 -87.7 -87.7 -87.6 ...
##  $ member_casual     : chr  "member" "member" "member" "member" ...
```

```
str(jan_2021)
```

```
## 'data.frame':    96834 obs. of  13 variables:
##  $ ride_id           : chr  "E19E6F1B8D4C42ED" "DC88F20C2C55F27F" "EC45C94683FE3F27" "4FA453A75AE377DB" ...
##  $ rideable_type     : chr  "electric_bike" "electric_bike" "electric_bike" "electric_bike" ...
##  $ started_at        : chr  "2021-01-23 16:14:19" "2021-01-27 18:43:08" "2021-01-21 22:35:54" "2021-01-07 13:3
1:13" ...
##  $ ended_at          : chr  "2021-01-23 16:24:44" "2021-01-27 18:47:12" "2021-01-21 22:37:14" "2021-01-07 13:4
2:55" ...
##  $ start_station_name: chr  "California Ave & Cortez St" "California Ave & Cortez St" "California Ave & Cortez
St" "California Ave & Cortez St" ...
##  $ start_station_id  : chr  "17660" "17660" "17660" "17660" ...
##  $ end_station_name  : chr  "" "" "" "" ...
##  $ end_station_id    : chr  "" "" "" "" ...
##  $ start_lat         : num  41.9 41.9 41.9 41.9 41.9 ...
##  $ start_lng         : num  -87.7 -87.7 -87.7 -87.7 -87.7 ...
##  $ end_lat           : num  41.9 41.9 41.9 41.9 41.9 ...
##  $ end_lng           : num  -87.7 -87.7 -87.7 -87.7 -87.7 ...
##  $ member_casual     : chr  "member" "member" "member" "member" ...
```

```
str(fev_2021)
```

```
## 'data.frame':    49622 obs. of  13 variables:
##  $ ride_id           : chr  "89E7AA6C29227EFF" "0FEFDE2603568365" "E6159D746B2DBB91" "B32D3199F1C2E75B" ...
##  $ rideable_type     : chr  "classic_bike" "classic_bike" "electric_bike" "classic_bike" ...
##  $ started_at        : chr  "2021-02-12 16:14:56" "2021-02-14 17:52:38" "2021-02-09 19:10:18" "2021-02-02 17:4
9:41" ...
##  $ ended_at          : chr  "2021-02-12 16:21:43" "2021-02-14 18:12:09" "2021-02-09 19:19:10" "2021-02-02 17:5
4:06" ...
##  $ start_station_name: chr  "Glenwood Ave & Touhy Ave" "Glenwood Ave & Touhy Ave" "Clark St & Lake St" "Wood S
t & Chicago Ave" ...
##  $ start_station_id  : chr  "525" "525" "KA1503000012" "637" ...
##  $ end_station_name  : chr  "Sheridan Rd & Columbia Ave" "Bosworth Ave & Howard St" "State St & Randolph St"
"Honore St & Division St" ...
##  $ end_station_id    : chr  "660" "16806" "TA1305000029" "TA1305000034" ...
##  $ start_lat         : num  42 42 41.9 41.9 41.8 ...
##  $ start_lng         : num  -87.7 -87.7 -87.6 -87.7 -87.6 ...
##  $ end_lat           : num  42 42 41.9 41.9 41.8 ...
##  $ end_lng           : num  -87.7 -87.7 -87.6 -87.7 -87.6 ...
##  $ member_casual     : chr  "member" "casual" "member" "member" ...
```

```
str(mar_2021)
```

```
## 'data.frame':    228496 obs. of  13 variables:
##  $ ride_id           : chr  "CFA86D4455AA1030" "30D9DC61227D1AF3" "846D87A15682A284" "994D05AA75A168F2" ...
##  $ rideable_type     : chr  "classic_bike" "classic_bike" "classic_bike" "classic_bike" ...
##  $ started_at        : chr  "2021-03-16 08:32:30" "2021-03-28 01:26:28" "2021-03-11 21:17:29" "2021-03-11 13:2
6:42" ...
##  $ ended_at          : chr  "2021-03-16 08:36:34" "2021-03-28 01:36:55" "2021-03-11 21:33:53" "2021-03-11 13:5
5:41" ...
##  $ start_station_name: chr  "Humboldt Blvd & Armitage Ave" "Humboldt Blvd & Armitage Ave" "Shields Ave & 28th
Pl" "Winthrop Ave & Lawrence Ave" ...
##  $ start_station_id  : chr  "15651" "15651" "15443" "TA1308000021" ...
##  $ end_station_name  : chr  "Stave St & Armitage Ave" "Central Park Ave & Bloomingdale Ave" "Halsted St & 35th
St" "Broadway & Sheridan Rd" ...
##  $ end_station_id    : chr  "13266" "18017" "TA1308000043" "13323" ...
##  $ start_lat         : num  41.9 41.9 41.8 42 42 ...
##  $ start_lng         : num  -87.7 -87.7 -87.6 -87.7 -87.7 ...
##  $ end_lat           : num  41.9 41.9 41.8 42 42.1 ...
##  $ end_lng           : num  -87.7 -87.7 -87.6 -87.6 -87.7 ...
##  $ member_casual     : chr  "casual" "casual" "casual" "casual" ...
```

```
str(abr_2021)
```

```
## 'data.frame':    337230 obs. of  13 variables:
##  $ ride_id           : chr  "6C992BD37A98A63F" "1E0145613A209000" "E498E15508A80BAD" "1887262AD101C604" ...
##  $ rideable_type     : chr  "classic_bike" "docked_bike" "docked_bike" "classic_bike" ...
##  $ started_at        : chr  "2021-04-12 18:25:36" "2021-04-27 17:27:11" "2021-04-03 12:42:45" "2021-04-17 09:1
7:42" ...
##  $ ended_at          : chr  "2021-04-12 18:56:55" "2021-04-27 18:31:29" "2021-04-07 11:40:24" "2021-04-17 09:4
2:48" ...
##  $ start_station_name: chr  "State St & Pearson St" "Dorchester Ave & 49th St" "Loomis Blvd & 84th St" "Honore
St & Division St" ...
##  $ start_station_id  : chr  "TA1307000061" "KA1503000069" "20121" "TA1305000034" ...
##  $ end_station_name  : chr  "Southport Ave & Waveland Ave" "Dorchester Ave & 49th St" "Loomis Blvd & 84th St"
"Southport Ave & Waveland Ave" ...
##  $ end_station_id    : chr  "13235" "KA1503000069" "20121" "13235" ...
##  $ start_lat         : num  41.9 41.8 41.7 41.9 41.7 ...
##  $ start_lng         : num  -87.6 -87.6 -87.7 -87.7 -87.7 ...
##  $ end_lat           : num  41.9 41.8 41.7 41.9 41.7 ...
##  $ end_lng           : num  -87.7 -87.6 -87.7 -87.7 -87.7 ...
##  $ member_casual     : chr  "member" "casual" "casual" "member" ...
```

3. Processar

Na verificação foi notado que existiam 4 colunas que apresentavam tipos que não condiziam com os dados nas tabelas de maio até novembro de 2020. Assim, transformamos as colunas *start_station_id* e *end_station_id* em *character*, e as colunas *started_at* e *ended_at* em *datetime*. Nas tabelas de dezembro de 2020 até abril de 2021, apenas as colunas *started_at e ended_at foram alteradas para datetime, já que as outras estavam no tipo correto.

```
mai_2020 <- mutate(mai_2020,
                   start_station_id = as.character(start_station_id),
                   end_station_id = as.character(end_station_id),
                   started_at = as_datetime(started_at),
                   ended_at = as_datetime(ended_at))
```

```
jun_2020 <- mutate(jun_2020,
                   start_station_id = as.character(start_station_id),
                   end_station_id = as.character(end_station_id),
                   started_at = as_datetime(started_at),
                   ended_at = as_datetime(ended_at))

jul_2020 <- mutate(jul_2020,
                   start_station_id = as.character(start_station_id),
                   end_station_id = as.character(end_station_id),
                   started_at = as_datetime(started_at),
                   ended_at = as_datetime(ended_at))

ago_2020 <- mutate(ago_2020,
                   start_station_id = as.character(start_station_id),
                   end_station_id = as.character(end_station_id),
                   started_at = as_datetime(started_at),
                   ended_at = as_datetime(ended_at))

set_2020 <- mutate(set_2020,
                   start_station_id = as.character(start_station_id),
                   end_station_id = as.character(end_station_id),
                   started_at = as_datetime(started_at),
                   ended_at = as_datetime(ended_at))

out_2020 <- mutate(out_2020,
                   start_station_id = as.character(start_station_id),
                   end_station_id = as.character(end_station_id),
                   started_at = as_datetime(started_at),
                   ended_at = as_datetime(ended_at))

nov_2020 <- mutate(nov_2020,
                   start_station_id = as.character(start_station_id),
                   end_station_id = as.character(end_station_id),
                   started_at = as_datetime(started_at),
                   ended_at = as_datetime(ended_at))

dez_2020 <- mutate(dez_2020,
                   started_at = as_datetime(started_at),
                   ended_at = as_datetime(ended_at))

jan_2021 <- mutate(jan_2021,
                   started_at = as_datetime(started_at),
                   ended_at = as_datetime(ended_at))

fev_2021 <- mutate(fev_2021,
                   started_at = as_datetime(started_at),
                   ended_at = as_datetime(ended_at))

mar_2021 <- mutate(mar_2021,
                   started_at = as_datetime(started_at),
                   ended_at = as_datetime(ended_at))

abr_2021 <- mutate(abr_2021,
                   started_at = as_datetime(started_at),
                   ended_at = as_datetime(ended_at))
```

Com a compatibilidade de todas as tabelas, unimos todos os dados em uma única tabela chamada *all_trips*.

```
all_trips <- bind_rows(mai_2020,
                       jun_2020,
                       jul_2020,
                       ago_2020,
                       set_2020,
                       out_2020,
                       nov_2020,
                       dez_2020,
                       jan_2021,
                       fev_2021,
                       mar_2021,
                       abr_2021)
```

Na tabela *all_trips*, retiramos as colunas que não contribuem para a análise desse projeto, que são as colunas *start_station_id*, *end_station_id*, *start_lat*, *start_lng*, *end_lat*, *end_lng*. Posto que estamos interessados em entender o comportamento dos usuários, os números de identificação das estações assim como a latitude e longitude iniciais e finais não são necessárias à nossa análise.

```
all_trips <- all_trips %>%
  select(-c(start_station_id,
            end_station_id,
            start_lat,
            start_lng,
            end_lat,
            end_lng))
```

Após a retirada, adicionaremos colunas complementares para melhorar a nossa análise. A primeira coluna adicionada será a *ride_length*, que nos fornecerá o tempo de cada corrida, utilizando as colunas *started_at* e *ended_at*.

```
all_trips <- mutate(all_trips,
                    ride_length = ended_at - started_at)
```

Adicionaremos também mais 5 colunas de datas, para que possamos fazer análises utilizando diferentes medidas de tempo, visando entender com profundidade o comportamento dos diferentes tipos de usuários. Adicionaremos as colunas *date*, *day*, *day_of_week*, *month* e *year*, que nos fornecem a data, o dia, o dia da semana, o mês e o ano, respectivamente.

```
all_trips$date        <- as.Date(all_trips$started_at)
all_trips$day         <- format(as.Date(all_trips$date), "%d")
all_trips$day_of_week <- format(as.Date(all_trips$date), "%A")
all_trips$month       <- format(as.Date(all_trips$date), "%m")
all_trips$year        <- format(as.Date(all_trips$date), "%y")
```

A primeira coluna adicionada necessita de verificação. Na coluna *ride_length*, nós estamos lidando com uma unidade de tempo, portanto valores negativos não fazem sentido. Assim, retiraremos dos nossos dados todas as linhas em que a duração da corrida é negativa, que são 10506 linhas. Como estamos retirando dados, vamos criar outra tabela chamada *all_trips_v2*.

```
all_trips_v2 <- all_trips[!(all_trips$ride_length<0),]
```

4. Analisar

Na etapa de análise, faremos 8 análises para buscar entender a diferença de comportamento entre os usuários casuais e os membros anuais.

I. Quantidade de corridas dos tipos de usuário

Verificamos que os membros anuais possuem um número **42,3%** superior de corridas comparado aos usuários casuais.

```
all_trips_v2 %>%
  group_by(member_casual) %>%
  summarize(number_of_rides = n()) %>%
  arrange(member_casual)
```

```
## # A tibble: 2 x 2
##   member_casual number_of_rides
##   <chr>                  <int>
## 1 casual                 1540112
## 2 member                 2191584
```

II. Duração média da corrida dos tipos de usuário

Verificamos que a média de duração de corrida dos usuários casuais é **177%** maior que a média dos membros anuais.

```
aggregate(all_trips_v2$ride_length ~ all_trips_v2$member_casual, FUN = mean)
```

```
##   all_trips_v2$member_casual all_trips_v2$ride_length
## 1                     casual           2635.3052 secs
## 2                     member            950.0996 secs
```

III. Média de tempo de corrida x tipos de usuário + dias da semana

Ambos os usuários tem o domingo como o dia de maior média.

A variação das médias de durações durante a semana é semelhante para ambos os usuários: de segunda a sexta possuem uma média de duração inferior aos finais de semana.

```
all_trips_v2$day_of_week <- ordered(all_trips_v2$day_of_week,
                                    levels=c("domingo",
                                             "segunda-feira",
                                             "terça-feira",
                                             "quarta-feira",
                                             "quinta-feira",
                                             "sexta-feira",
                                             "sábado"))
aggregate(all_trips_v2$ride_length ~ all_trips_v2$member_casual + all_trips_v2$day_of_week, FUN = mean)
```

```
##    all_trips_v2$member_casual all_trips_v2$day_of_week all_trips_v2$ride_length
## 1                      casual                  domingo           2987.7343 secs
## 2                      member                  domingo           1064.1544 secs
## 3                      casual            segunda-feira           2636.1478 secs
## 4                      member            segunda-feira            910.8418 secs
## 5                      casual              terça-feira           2362.4766 secs
## 6                      member              terça-feira            891.9936 secs
## 7                      casual             quarta-feira           2394.9069 secs
## 8                      member             quarta-feira            908.7622 secs
## 9                      casual             quinta-feira           2476.0688 secs
## 10                     member             quinta-feira            893.9189 secs
## 11                     casual              sexta-feira           2511.7763 secs
## 12                     member              sexta-feira            934.8350 secs
## 13                     casual                   sábado           2752.1568 secs
## 14                     member                   sábado           1048.9659 secs
```

IV. Dia da semana x tipos de usuário (quantidade de corrida e média de duração)

Já na quantidade de corridas, notamos uma diferença significativa na variação entre os dias da semana.

Para ambos os tipos de usuários, sábado é o dia que tem mais corridas.

Nas corridas dos usuários casuais nota-se uma grande diferença dos dias de semana, que possuem uma média de 180079 corridas, e os finais de semana, que possuem uma média de 319875 corridas: média **77,6%** maior. Se comparamos isoladamente o dia com maior quantidade de corridas (sábado) com o dia de menor número (terça-feira), essa diferença é ainda mais significativa: **121,45%**.

Nas corridas dos membros anuais, a diferença é menos acentuada, com um detalhe. Apesar de sábado ser o dia com maior quantidade de corridas, os membros anuais possuem uma média de corridas superior nos dias da semana se comparado aos finais de semana, com uma diferença sutil de **1,40%**. Se compararmos isoladamente o dia com maior quantidade de corridas (sábado) com o dia de menor número (domingo), essa diferença é de **22,19%**.

```
all_trips_v2 %>%
  group_by(member_casual, day_of_week) %>%
  summarize(number_of_rides = n(),
            avg_duration = mean(ride_length)) %>%
  arrange(member_casual, day_of_week)
```

```
## # A tibble: 14 x 4
## # Groups:   member_casual [2]
##    member_casual day_of_week   number_of_rides avg_duration
##    <chr>         <ord>                  <int> <drtn>
##  1 casual        domingo               280994 2987.7343 secs
##  2 casual        segunda-feira         164524 2636.1478 secs
##  3 casual        terça-feira           161990 2362.4766 secs
##  4 casual        quarta-feira          168598 2394.9069 secs
##  5 casual        quinta-feira          176313 2476.0688 secs
##  6 casual        sexta-feira           228972 2511.7763 secs
##  7 casual        sábado                358721 2752.1568 secs
##  8 member        domingo               279022 1064.1544 secs
##  9 member        segunda-feira         287008  910.8418 secs
## 10 member        terça-feira           307072  891.9936 secs
## 11 member        quarta-feira          323538  908.7622 secs
## 12 member        quinta-feira          319275  893.9189 secs
## 13 member        sexta-feira           334732  934.8350 secs
## 14 member        sábado                340937 1048.9659 secs
```

V. Mês x tipos de usuário (quantidade de corrida e média de duração)

Analisando os usuários casuais, a quantidade de corridas nos meses de inverno em Chicago (dezembro, janeiro e fevereiro) é de **19415**. No verão, outono e primavera, porém, a quantidade de corridas cresce significativamente: **1122%**, **694%** e **428%** respectivamente. Já quando analisamos a média de duração das corridas, os meses do outono apresentam a menor média: **2005,88 segundos**. No verão, primavera e inverno, as durações aumentam em **56%**, **27%** e **2%**, respectivamente.

Os membros anuais apresentam a menor quantidade também nos meses de inverno, com **73116** corridas. Nos outros meses há um crescimento, porém menos acentuado que os usuários casuais, com aumentos de **265%**, **226%** e **109%** para os meses de verão, outono e primavera, respectivamente. Analisando a média de duração das corridas, os meses do outono apresentam a menor média: **863,56 segundos**. No verão, primavera e inverno, as durações aumentam em **24%**, **12%** e **1%**, respectivamente.

```
all_trips_v2 %>%
  group_by(member_casual, month) %>%
  summarize(number_of_rides = n(),
            avg_duration = mean(ride_length)) %>%
  arrange(month, member_casual)
```

```
## # A tibble: 24 x 4
## # Groups:   member_casual [2]
##    member_casual month number_of_rides avg_duration
##    <chr>         <chr>           <int> <drtn>
##  1 casual        01              18117 1541.0754 secs
##  2 member        01              78715  772.3387 secs
##  3 casual        02              10131 2962.3937 secs
##  4 member        02              39491 1081.3251 secs
##  5 casual        03              84032 2289.5511 secs
##  6 member        03             144462  838.2031 secs
##  7 casual        04             136601 2281.3794 secs
##  8 member        04             200624  881.3527 secs
##  9 casual        05              86844 3073.0522 secs
## 10 member        05             113258 1186.3409 secs
## # ... with 14 more rows
```

VI. Tipo de bicicleta x tipos de usuário (quantidade de corrida)

Analisando os dados foi notado que a grande maioria das bicicletas utilizadas são do tipo *docked_bike,* seguida do tipo *eletric_bike*, e depois a *classic_bike*, com a preferência de **67%**, **19%** e **14%**, respectivamente.

```
all_trips_v2 %>%
  group_by(rideable_type, member_casual) %>%
  summarize(number_of_rides = n()) %>%
  arrange(rideable_type, member_casual)
```

```
## # A tibble: 6 x 3
## # Groups:   rideable_type [3]
##   rideable_type member_casual number_of_rides
##   <chr>         <chr>                  <int>
## 1 classic_bike  casual                141576
## 2 classic_bike  member                392911
## 3 docked_bike   casual               1114512
## 4 docked_bike   member               1373605
## 5 electric_bike casual                284024
## 6 electric_bike member                425068
```

VII. Estação de corrida (início) x tipos de usuário (quantidade de corrida)

As 3 estações com maiores números de saídas são Streeter Dr & Grand Ave, Lake Shore Dr & Monroe St e Millenium Park, todas por usuários casuais.

```
all_trips_v2 %>%
  group_by(start_station_name, member_casual) %>%
  summarize(number_of_rides = n()) %>%
  arrange(desc(number_of_rides), member_casual)
```

```
## # A tibble: 1,403 x 3
## # Groups:   start_station_name [712]
##    start_station_name          member_casual number_of_rides
##    <chr>                       <chr>                  <int>
##  1 ""                          member                 91265
##  2 ""                          casual                 56918
##  3 "Streeter Dr & Grand Ave"   casual                 28488
##  4 "Lake Shore Dr & Monroe St" casual                 23538
##  5 "Millennium Park"           casual                 21507
##  6 "Clark St & Elm St"         member                 21396
##  7 "Broadway & Barry Ave"      member                 16581
##  8 "Wells St & Concord Ln"     member                 16283
##  9 "Dearborn St & Erie St"     member                 16267
## 10 "Theater on the Lake"       casual                 16243
## # ... with 1,393 more rows
```

VIII. Estação de corrida (final) x tipos de usuário (quantidade de corrida)

As 3 estações com os maiores números de chegadas são Streeter Dr & Grand Ave, Lake Shore Dr & Monroe St e Millenium Park, todas por usuários casuais.

```
all_trips_v2 %>%
  group_by(end_station_name, member_casual) %>%
  summarize(number_of_rides = n()) %>%
  arrange(desc(number_of_rides), member_casual)
```

```
## `summarise()` has grouped output by 'end_station_name'. You can override using the `.groups` argument.
```

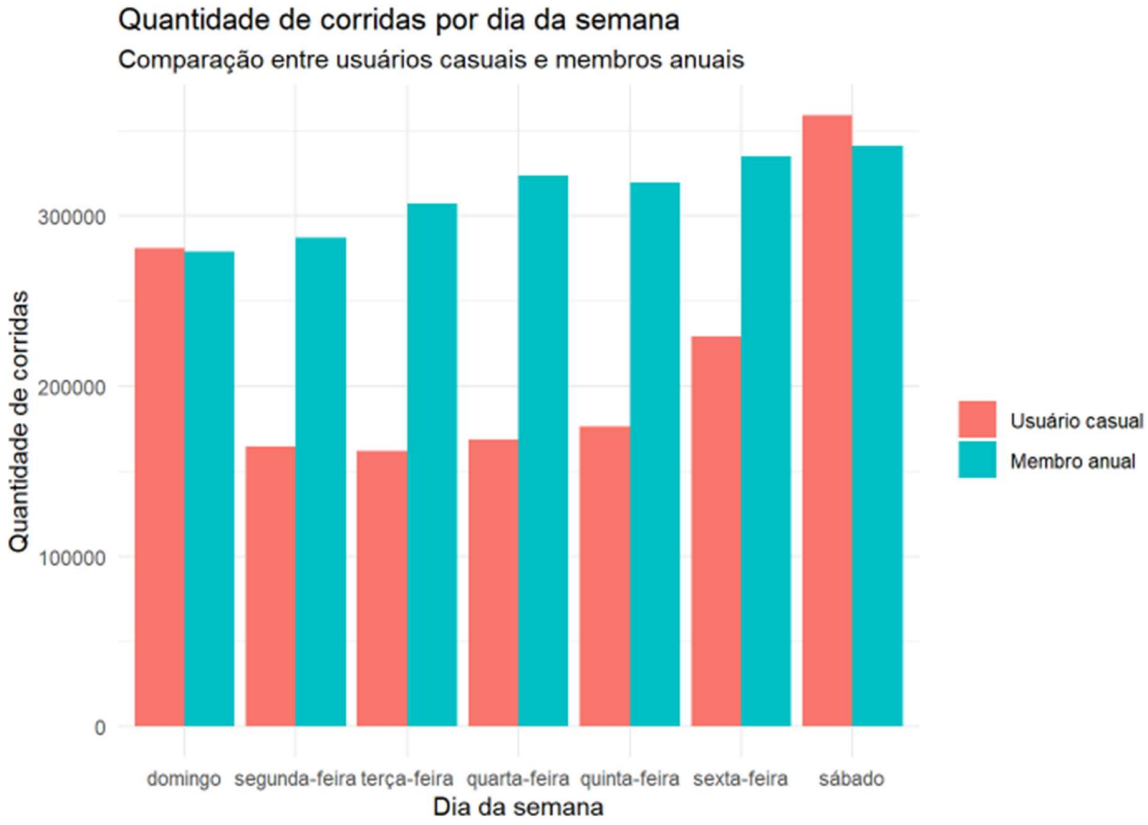
```
## # A tibble: 1,414 x 3
## # Groups:   end_station_name [713]
##    end_station_name            member_casual number_of_rides
##    <chr>                       <chr>                  <int>
##  1 ""                          member                 99902
##  2 ""                          casual                 71330
##  3 "Streeter Dr & Grand Ave"   casual                 30718
##  4 "Lake Shore Dr & Monroe St" casual                 22769
##  5 "Millennium Park"           casual                 22330
##  6 "Clark St & Elm St"         member                 21835
##  7 "Theater on the Lake"       casual                 18212
##  8 "St. Clair St & Erie St"    member                 17287
##  9 "Dearborn St & Erie St"     member                 16850
## 10 "Broadway & Barry Ave"      member                 16777
## # ... with 1,404 more rows
```

5. Compartilhar

Com as análises feitas, criaremos as visualizações de dados para tornar a nossa análise mais intuitiva e acessível para os stakeholders.

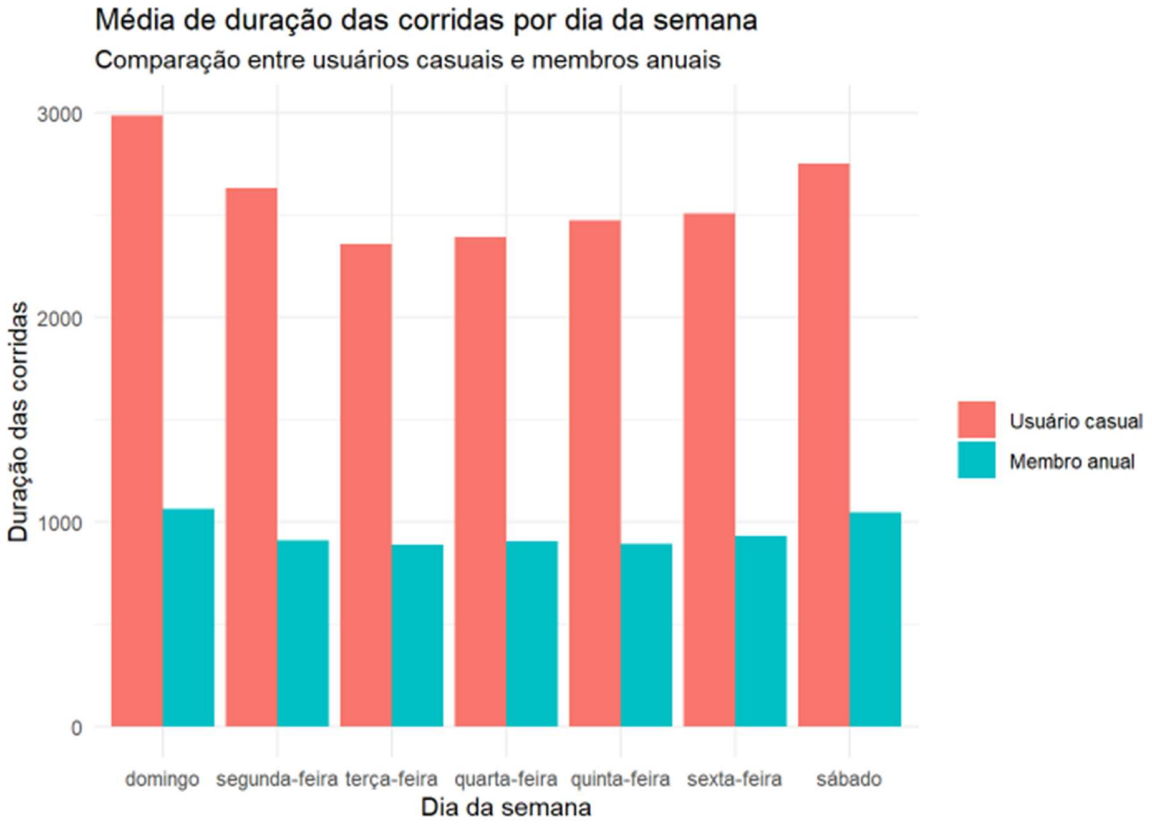
I. Utilização por semana

```
options(scipen = 100)   #comando para desativar a notação científica
all_trips_v2 %>%
  group_by(member_casual, day_of_week) %>%
  summarize(number_of_rides = n(),
            avg_duration = mean(ride_length)) %>%
  arrange(member_casual, day_of_week) %>%
  ggplot(aes(x = day_of_week, y = number_of_rides, fill = member_casual)) +
  geom_col(position = "dodge") +
  labs(title = "Quantidade de corridas por dia da semana",
       subtitle = "Comparação entre usuários casuais e membros anuais",
       x = "Dia da semana",
       y = "Quantidade de corridas") +
  theme_minimal() +
  scale_fill_discrete(name="",
                      labels=c("Usuário casual", "Membro anual")) +
  theme(legend.position="right")
```



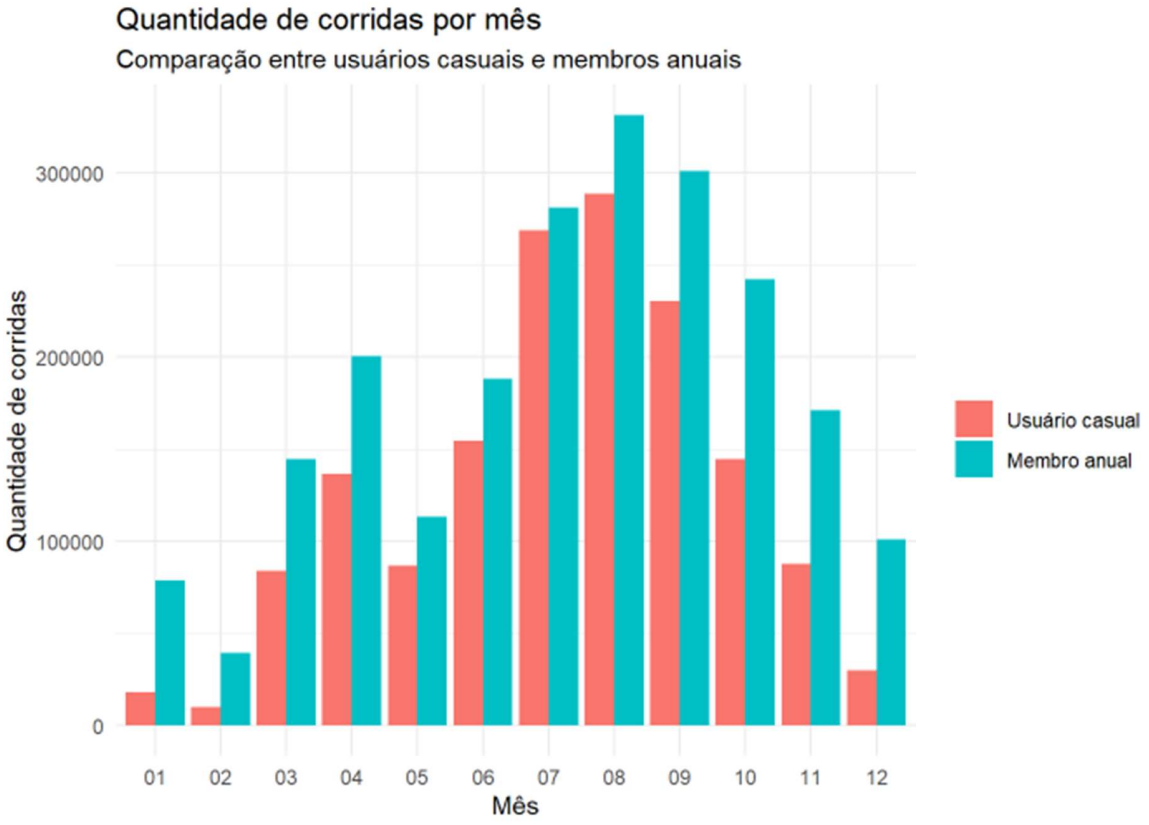
II. Duração de uso por semana

```
all_trips_v2 %>%
  group_by(member_casual, day_of_week) %>%
  summarize(number_of_rides = n(),
            avg_duration = mean(ride_length)) %>%
  arrange(member_casual, day_of_week) %>%
  ggplot(aes(x = day_of_week, y = avg_duration, fill = member_casual)) +
  geom_col(position = "dodge") +
  labs(title = "Média de duração das corridas por dia da semana",
       subtitle = "Comparação entre usuários casuais e membros anuais",
       x = "Dia da semana",
       y = "Duração das corridas") +
  theme_minimal() +
  scale_fill_discrete(name="",
                      labels=c("Usuário casual", "Membro anual")) +
  theme(legend.position="right")
```



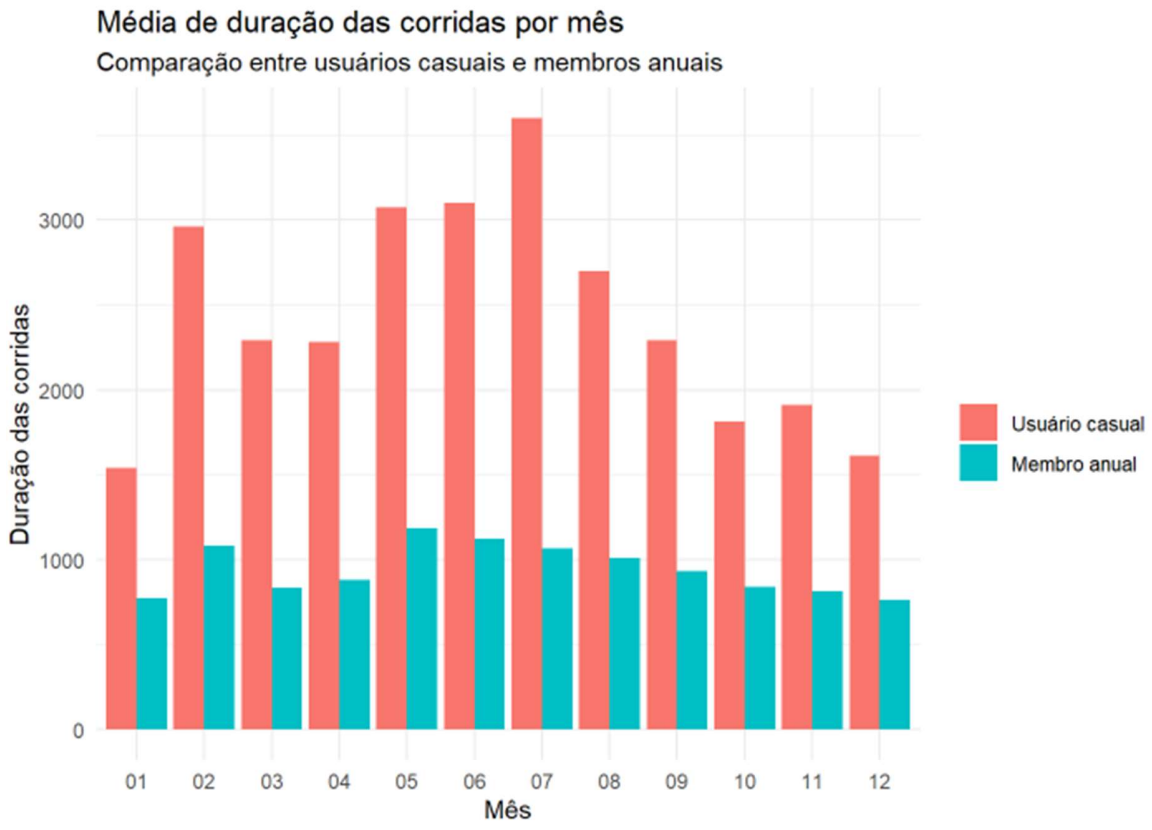
III. Utilização por mês

```
all_trips_v2 %>%
  group_by(member_casual, month) %>%
  summarize(number_of_rides = n(),
            avg_duration = mean(ride_length)) %>%
  arrange(member_casual, month) %>%
  ggplot(aes(x = month, y = number_of_rides, fill = member_casual)) +
  geom_col(position = "dodge") +
  labs(title = "Quantidade de corridas por mês",
       subtitle = "Comparação entre usuários casuais e membros anuais",
       x = "Mês",
       y = "Quantidade de corridas") +
  theme_minimal() +
  scale_fill_discrete(name="",
                      labels=c("Usuário casual", "Membro anual")) +
  theme(legend.position="right")
```



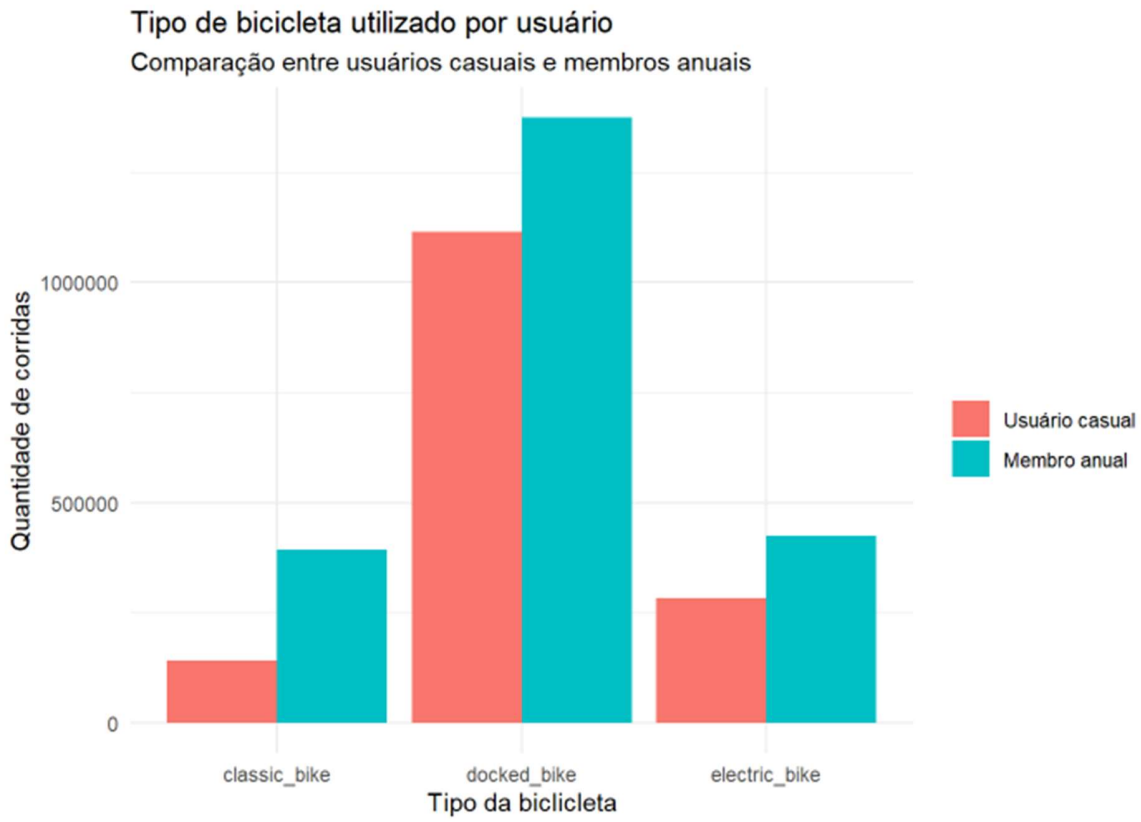
IV. Duração de uso por mês

```
all_trips_v2 %>%
  group_by(member_casual, month) %>%
  summarize(number_of_rides = n(),
            avg_duration = mean(ride_length)) %>%
  arrange(member_casual, month) %>%
  ggplot(aes(x = month, y = avg_duration, fill = member_casual)) +
  geom_col(position = "dodge") +
  labs(title = "Média de duração das corridas por mês",
       subtitle = "Comparação entre usuários casuais e membros anuais",
       x = "Mês",
       y = "Duração das corridas") +
  theme_minimal() +
  scale_fill_discrete(name="",
                      labels=c("Usuário casual", "Membro anual")) +
  theme(legend.position="right")
```



V. Tipo de bicicleta por usuário

```
all_trips_v2 %>%
  group_by(member_casual, rideable_type) %>%
  summarize(number_of_rides = n(),) %>%
  arrange(member_casual, rideable_type) %>%
  ggplot(aes(x = rideable_type, y = number_of_rides, fill = member_casual)) +
  geom_col(position = "dodge") +
  labs(title = "Tipo de bicicleta utilizado por usuário",
       subtitle = "Comparação entre usuários casuais e membros anuais",
       x = "Tipo da biclicleta",
       y = "Quantidade de corridas") +
  theme_minimal() +
  scale_fill_discrete(name="",
                      labels=c("Usuário casual", "Membro anual")) +
  theme(legend.position="right")
```



6. Agir

Com a análise dos dados disponibilizados, criamos perfis para cada tipo de usuário.

O usuário casual utiliza a bicicleta para corridas mais duradouras e com um foco maior nas corridas ao fim de semana, indicando o uso mais voltado ao lazer. Além do destaque para os finais de semana, os meses de verão são um grande sucesso para os usuários casuais, pois o tempo é favorável para o uso da bicicleta como meio de transporte e como lazer.

O membro anual, por sua vez, possuem um perfil de corridas mais rotineiros e repetitivos, pois apresentam variações leves tanto na quantidade quanto na duração ao longo da semana. Existe um destaque também para os meses de verão e outono, pois o clima favorece o uso da bicicleta.

Assim, temos de nos mostrar financeiramente viáveis para os usuários casuais através de planos mais flexíveis ou pacotes de corridas para criar um vínculo mais forte com a Cyclistic, aproximando-os de uma assinatura.

Além disso, considerar a criação de eventos para o público, como passeio por pontos turísticos marcantes da cidade, visando fomentar a utilização da bicicleta como mobilidade e divulgar a nossa solução, além de desenvolver um senso de comunidade entre os usuários da Cyclistic, aumentando a fidelidade dos clientes à marca e o marketing boca-a-boca.

Por fim, é recomendada uma análise mais aprofundada comparando os preços das corridas individuais, do ticket diário com o preço da assinatura anual, para entender como podemos nos tornar mais financeiramente viável para os usuários casuais. Outra informação que pode nos trazer insights é a distância percorrida por corrida, que pode nos ajudar a traçar ainda melhor o perfil dos clientes da Cyclistic.