Análise de compartilhamento de bicicletas Cyclistic

Jander

14-04-2024

Estudo de caso Cyclistic

Cenário Cyclistic: Um programa de compartilhamento de bicicletas que conta com mais de 5.800 bicicletas e 600 estações de compartilhamento. A Cyclistic se diferencia por também oferecer bicicletas reclináveis, triciclos manuais e bicicletas de carga, tornando o compartilhamento de mais inclusivo para pessoas com deficiência e ciclistas que não podem usar uma bicicleta padrão de duas rodas. A maioria dos ciclistas opta por bicicletas tradicionais; cerca de 8% dos motociclistas usam as opções assistivas. Os usuários da Cyclistic são mais propensos a pedalar por lazer, mas cerca de 30% utilizam as bicicletas para se deslocarem ao trabalho diariamente.

Os clientes que adquirem passes de viagem única ou de dia inteiro são chamados de passageiros casuais.

Lily Moreno(Diretora de Marketing) estabeleceu um objetivo claro: criar estratégias de marketing destinadas a converter passageiros casuais em membros anuais. Para fazer isso, no entanto, a equipe de analistas de marketing precisa entender melhor como os membros anuais e os passageiros casuais diferem, por que os passageiros casuais iriam querer adquirir um plano e como a mídia digital poderia afetar suas táticas de marketing.

Perguntar:

Partes Interessadas Lily Moreno: Diretora de marketing e gerente. Responsável pelo desenvolvimento de campanhas e iniciativas de promoção do programa de compartilhamento de bicicletas.

Equipe de análise de marketing da Cyclistic: Uma equipe de analistas de dados responsáveis por coletar, analisar e relatar dados que ajudam a orientar a estratégia de marketing da Cyclistic.

Equipe executiva da Cyclistic: A equipe executiva notoriamente detalhista decidirá se aprova o programa de marketing recomendado.

Tarefa de negócios

Entender como os membros anuais e casuais se diferem.

Qual problema estou tentando resolver?

Por que os passageiros casuais iriam querer adquirir um plano e como a mídia digital poderia afetar táticas de marketing.

Preparar: Onde os dados estão localizados? Os arquivos estão sãos em um servidor remoto (aqui) e fornecido por Motivate International Inc.

Foram feitos os downloads dos arquivos Cyclistic em formato csv e armazenados em meu desktop organizados por data. Não foram identificados viéses de credibilidade.

Licença para uso dos dados:

https://ride.divvybikes.com/data-license-agreement

Processar: A ferramenta que escolhi no processamento dos dados foi R para colocar em prática o conteúdo ensinado no curso. Os arquivos contém dados do ano de 2022.

Carregando e instalando pacotes necessários

Antes de unir os arquivos em um único dataframe comparei o nome das colunas de todos os arquivos.

```
trips01 <- read.csv("202201-divvy-tripdata.csv")
trips02 <- read.csv("202202-divvy-tripdata.csv")
trips03 <- read.csv("202203-divvy-tripdata.csv")
trips04 <- read.csv("202204-divvy-tripdata.csv")
trips05 <- read.csv("202205-divvy-tripdata.csv")
trips06 <- read.csv("202206-divvy-tripdata.csv")
trips07 <- read.csv("202207-divvy-tripdata.csv")
trips08 <- read.csv("202208-divvy-tripdata.csv")
trips09 <- read.csv("202209-divvy-tripdata.csv")
trips10 <- read.csv("202210-divvy-tripdata.csv")
trips11 <- read.csv("202211-divvy-tripdata.csv")
trips12 <- read.csv("202212-divvy-tripdata.csv")


colnames(trips01)
```

```
##  [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(trips02)
```

```
##  [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(trips03)
```

```
##  [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"
```

```r
colnames(trips04)
```

```
##  [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"
```

```r
colnames(trips05)
```

```
##  [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"
```

```r
colnames(trips06)
```

```
##  [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"
```

```r
colnames(trips07)
```

```
##  [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"
```

```r
colnames(trips08)
```

```
##  [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"
```

```r
colnames(trips09)
```

```
##  [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(trips10)
```

```
##  [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(trips11)
```

```
##  [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(trips12)
```

```
##  [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"
```

Unindo arquivos em um único dafa frame

```
all_trips <- list.files(path='C:/Users/jande/Documentos/Estudo de Caso/datasets/trip_files/divvy_tripda
  lapply(read_csv) %>%
  bind_rows
```

Data Cleaning e preparação para análise

```
# Nome de colunas
colnames(all_trips)
```

```
##  [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"
```

```
# Número de linhas
nrow(all_trips)
```

```
## [1] 5667717
```

```
# Dimensoes do df
dim(all_trips)
```

```
## [1] 5667717      13
```

```r
# 6 primeiras linhas do df
head(all_trips)
```

```
## # A tibble: 6 x 13
##   ride_id          rideable_type started_at          ended_at
##   <chr>            <chr>         <dttm>              <dttm>
## 1 C2F7DD78E82EC875 electric_bike 2022-01-13 11:59:47 2022-01-13 12:02:44
## 2 A6CF8980A652D272 electric_bike 2022-01-10 08:41:56 2022-01-10 08:46:17
## 3 BD0F91DFF741C66D classic_bike  2022-01-25 04:53:40 2022-01-25 04:58:01
## 4 CBB80ED419105406 classic_bike  2022-01-04 00:18:04 2022-01-04 00:33:00
## 5 DDC963BFDDA51EEA classic_bike  2022-01-20 01:31:10 2022-01-20 01:37:12
## 6 A39C6F6CC0586C0B classic_bike  2022-01-11 18:48:09 2022-01-11 18:51:31
## # 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>
```

```r
# Lista de colunas e tipos de dado
str(all_trips)
```

```
## spc_tbl_ [5,667,717 x 13] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ ride_id           : chr [1:5667717] "C2F7DD78E82EC875" "A6CF8980A652D272" "BD0F91DFF741C66D" "CBB8
##  $ rideable_type     : chr [1:5667717] "electric_bike" "electric_bike" "classic_bike" "classic_bike"
##  $ started_at        : POSIXct[1:5667717], format: "2022-01-13 11:59:47" "2022-01-10 08:41:56" ...
##  $ ended_at          : POSIXct[1:5667717], format: "2022-01-13 12:02:44" "2022-01-10 08:46:17" ...
##  $ start_station_name: chr [1:5667717] "Glenwood Ave & Touhy Ave" "Glenwood Ave & Touhy Ave" "Sheffi
##  $ start_station_id  : chr [1:5667717] "525" "525" "TA1306000016" "KA1504000151" ...
##  $ end_station_name  : chr [1:5667717] "Clark St & Touhy Ave" "Clark St & Touhy Ave" "Greenview Ave &
##  $ end_station_id    : chr [1:5667717] "RP-007" "RP-007" "TA1307000001" "TA1309000021" ...
##  $ start_lat         : num [1:5667717] 42 42 41.9 42 41.9 ...
##  $ start_lng         : num [1:5667717] -87.7 -87.7 -87.7 -87.7 -87.6 ...
##  $ end_lat           : num [1:5667717] 42 42 41.9 42 41.9 ...
##  $ end_lng           : num [1:5667717] -87.7 -87.7 -87.7 -87.7 -87.6 ...
##  $ member_casual     : chr [1:5667717] "casual" "casual" "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_character(),
##   ..   end_station_name = col_character(),
##   ..   end_station_id = col_character(),
##   ..   start_lat = col_double(),
##   ..   start_lng = col_double(),
##   ..   end_lat = col_double(),
##   ..   end_lng = col_double(),
##   ..   member_casual = col_character()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

```r
# resumo estatistico dos dados
summary(all_trips)
```

```
##    ride_id          rideable_type        started_at
##  Length:5667717     Length:5667717     Min.   :2022-01-01 00:00:05.00
##  Class :character   Class :character   1st Qu.:2022-05-28 19:21:05.00
##  Mode  :character   Mode  :character   Median :2022-07-22 15:03:59.00
##                                        Mean   :2022-07-20 07:21:18.74
##                                        3rd Qu.:2022-09-16 07:21:29.00
##                                        Max.   :2022-12-31 23:59:26.00
##
##     ended_at                      start_station_name start_station_id
##  Min.   :2022-01-01 00:01:48.00   Length:5667717     Length:5667717
##  1st Qu.:2022-05-28 19:43:07.00   Class :character   Class :character
##  Median :2022-07-22 15:24:44.00   Mode  :character   Mode  :character
##  Mean   :2022-07-20 07:40:45.33
##  3rd Qu.:2022-09-16 07:39:03.00
##  Max.   :2023-01-02 04:56:45.00
##
##  end_station_name   end_station_id       start_lat       start_lng
##  Length:5667717     Length:5667717     Min.   :41.64   Min.   :-87.84
##  Class :character   Class :character   1st Qu.:41.88   1st Qu.:-87.66
##  Mode  :character   Mode  :character   Median :41.90   Median :-87.64
##                                        Mean   :41.90   Mean   :-87.65
##                                        3rd Qu.:41.93   3rd Qu.:-87.63
##                                        Max.   :45.64   Max.   :-73.80
##
##     end_lat         end_lng       member_casual
##  Min.   : 0.00   Min.   :-88.14   Length:5667717
##  1st Qu.:41.88   1st Qu.:-87.66   Class :character
##  Median :41.90   Median :-87.64   Mode  :character
##  Mean   :41.90   Mean   :-87.65
##  3rd Qu.:41.93   3rd Qu.:-87.63
##  Max.   :42.37   Max.   :  0.00
##  NA's   :5858    NA's   :5858
```

```r
# Vizualisado colunas
glimpse(all_trips)
```

```
## Rows: 5,667,717
## Columns: 13
## $ ride_id            <chr> "C2F7DD78E82EC875", "A6CF8980A652D272", "BD0F91DFF7~
## $ rideable_type      <chr> "electric_bike", "electric_bike", "classic_bike", "~
## $ started_at         <dttm> 2022-01-13 11:59:47, 2022-01-10 08:41:56, 2022-01-~
## $ ended_at           <dttm> 2022-01-13 12:02:44, 2022-01-10 08:46:17, 2022-01-~
## $ start_station_name <chr> "Glenwood Ave & Touhy Ave", "Glenwood Ave & Touhy A~
## $ start_station_id   <chr> "525", "525", "TA1306000016", "KA1504000151", "TA13~
## $ end_station_name   <chr> "Clark St & Touhy Ave", "Clark St & Touhy Ave", "Gr~
## $ end_station_id     <chr> "RP-007", "RP-007", "TA1307000001", "TA1309000021",~
## $ start_lat          <dbl> 42.01280, 42.01276, 41.92560, 41.98359, 41.87785, 4~
## $ start_lng          <dbl> -87.66591, -87.66597, -87.65371, -87.66915, -87.624~
## $ end_lat            <dbl> 42.01256, 42.01256, 41.92533, 41.96151, 41.88462, 4~
## $ end_lng            <dbl> -87.67437, -87.67437, -87.66580, -87.67139, -87.627~
```

```
## $ member_casual      <chr> "casual", "casual", "member", "casual", "member", "~
```

Buscando possíveis erros de digitação

```
table(all_trips$member_casual)
```

```
##
##  casual  member
## 2322032 3345685
```

```
table(all_trips$rideable_type)
```

```
##
##  classic_bike   docked_bike electric_bike
##       2601214        177474       2889029
```

Adicionando colunas para agregações futuras

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

Verificando e removendo valores nulos

```
sum(is.na(all_trips))
```

```
## [1] 3463328
```

```
all_trips_v2 <- all_trips %>% drop_na()
sum(is.na(all_trips_v2))
```

```
## [1] 0
```

Adicionando duração do passeio em segundos

```
all_trips_v2$duracao_passeio <- difftime(all_trips_v2$ended_at, all_trips_v2$started_at)
```

Convertendo duracao_passeio para numérico

```
all_trips_v2$duracao_passeio <- as.numeric(as.character(all_trips_v2$duracao_passeio))
is.numeric(all_trips_v2$duracao_passeio)
```

```
## [1] TRUE
```

Removendo duracao_passeio negativa

```r
all_trips_v2 <- all_trips_v2[!(all_trips_v2$start_station_name == "HQ QR" | all_trips_v2$duracao_passeio
```

Análise Descritiva

```r
mean(all_trips_v2$duracao_passeio)
```

```
## [1] 1025.75
```

```r
median(all_trips_v2$duracao_passeio)
```

```
## [1] 636
```

```r
max(all_trips_v2$duracao_passeio)
```

```
## [1] 2061244
```

```r
min(all_trips_v2$duracao_passeio)
```

```
## [1] 0
```

```r
# As linhas acima podem ser resumidas utilizando "summary"
summary(all_trips_v2$duracao_passeio)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       0     363     636    1026    1141 2061244
```

Comparando membros e usuários casuais

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

```
##   all_trips_v2$member_casual all_trips_v2$duracao_passeio
## 1                     casual                     1439.586
## 2                     member                      747.104
```

```r
aggregate(all_trips_v2$duracao_passeio ~ all_trips_v2$member_casual, FUN = median)
```

```
##   all_trips_v2$member_casual all_trips_v2$duracao_passeio
## 1                     casual                          831
## 2                     member                          539
```

```r
aggregate(all_trips_v2$duracao_passeio ~ all_trips_v2$member_casual, FUN = max)
```

```
##   all_trips_v2$member_casual all_trips_v2$duracao_passeio
## 1                     casual                     2061244
## 2                     member                       89594
```

```r
aggregate(all_trips_v2$duracao_passeio ~ all_trips_v2$member_casual, FUN = min)
```

```
##   all_trips_v2$member_casual all_trips_v2$duracao_passeio
## 1                     casual                            0
## 2                     member                            0
```

Média de duração de passeios por dia membros vs usuários casuais

```r
# Ordenando dias da semana
all_trips_v2$day_of_week <- ordered(all_trips_v2$day_of_week, levels=c("domingo", "segunda-feira", "ter
```

```r
aggregate(all_trips_v2$duracao_passeio ~ all_trips_v2$member_casual + all_trips_v2$day_of_week, FUN = m
```

```
##    all_trips_v2$member_casual all_trips_v2$day_of_week
## 1                      casual                  domingo
## 2                      member                  domingo
## 3                      casual            segunda-feira
## 4                      member            segunda-feira
## 5                      casual              terça-feira
## 6                      member              terça-feira
## 7                      casual             quarta-feira
## 8                      member             quarta-feira
## 9                      casual             quinta-feira
## 10                     member             quinta-feira
## 11                     casual              sexta-feira
## 12                     member              sexta-feira
## 13                     casual                   sábado
## 14                     member                   sábado
##    all_trips_v2$duracao_passeio
## 1                     1633.6491
## 2                      831.0463
## 3                     1490.0468
## 4                      721.9727
## 5                     1286.5190
## 6                      707.4686
## 7                     1243.0783
## 8                      710.8143
## 9                     1284.2069
## 10                     721.9113
## 11                    1341.4274
## 12                     733.6164
## 13                    1605.9639
## 14                     838.8955
```

Média duração de passeio por tipo de bicicleta e dia da semana

```r
aggregate(all_trips_v2$duracao_passeio ~ all_trips_v2$rideable_type + all_trips_v2$day_of_week, FUN = m
```

```
##    all_trips_v2$rideable_type all_trips_v2$day_of_week
## 1                classic_bike                  domingo
## 2                 docked_bike                  domingo
```

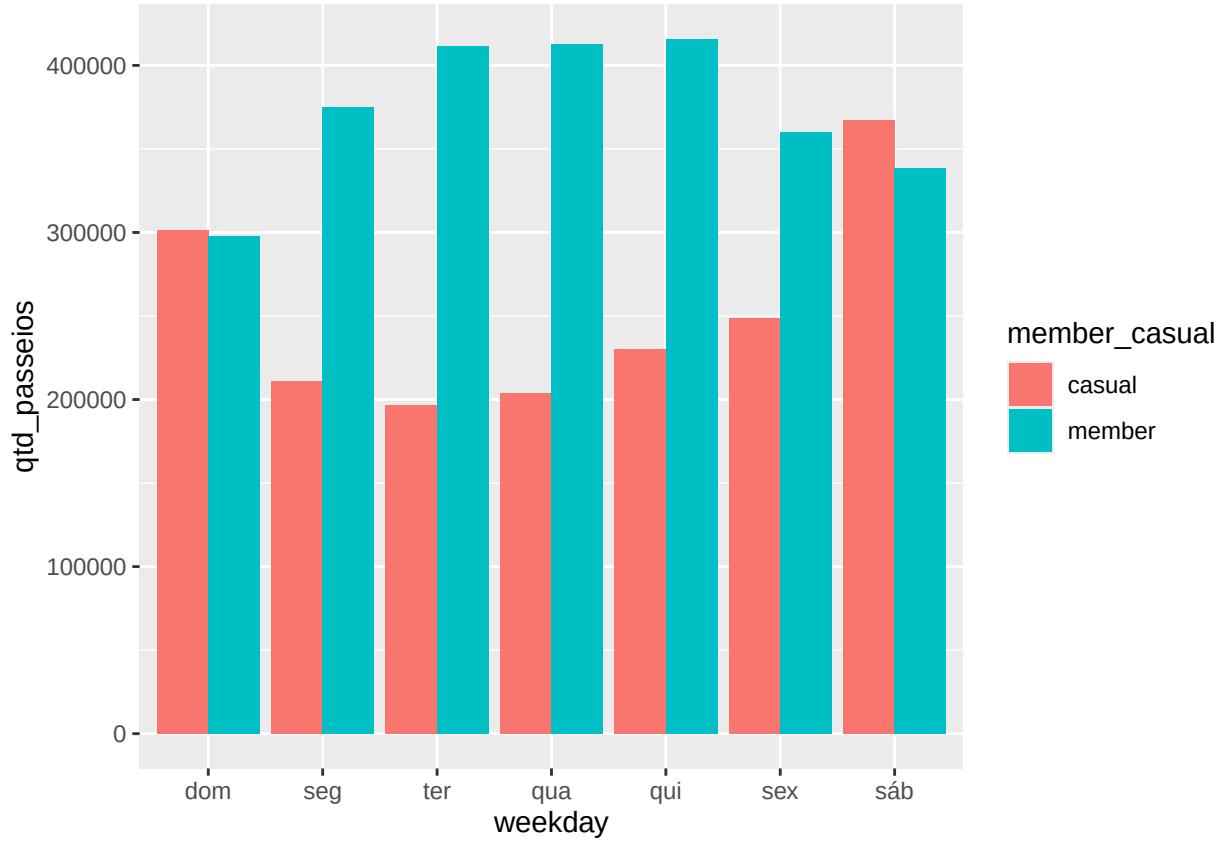
```
## 3               electric_bike               domingo
## 4                classic_bike          segunda-feira
## 5                 docked_bike          segunda-feira
## 6               electric_bike          segunda-feira
## 7                classic_bike            terça-feira
## 8                 docked_bike            terça-feira
## 9               electric_bike            terça-feira
## 10               classic_bike           quarta-feira
## 11                docked_bike           quarta-feira
## 12              electric_bike           quarta-feira
## 13               classic_bike           quinta-feira
## 14                docked_bike           quinta-feira
## 15              electric_bike           quinta-feira
## 16               classic_bike            sexta-feira
## 17                docked_bike            sexta-feira
## 18              electric_bike            sexta-feira
## 19               classic_bike                 sábado
## 20                docked_bike                 sábado
## 21              electric_bike                 sábado
##    all_trips_v2$duracao_passeio
## 1                     1208.8029
## 2                     3141.7058
## 3                      952.5210
## 4                      983.3772
## 5                     3303.8398
## 6                      781.9968
## 7                      910.0892
## 8                     2854.0780
## 9                      717.9874
## 10                     901.0409
## 11                    2868.9331
## 12                     720.0832
## 13                     935.6433
## 14                    2971.4789
## 15                     738.6677
## 16                     987.4033
## 17                    2785.5459
## 18                     794.3174
## 19                    1216.0147
## 20                    3148.2605
## 21                     957.9425
```

Construindo vizualizações com ggplot2

Média de passeios por tipo de membro e dia da semana

```
all_trips_v2 %>%
  mutate(weekday = wday(started_at, label = TRUE)) %>%
  group_by(member_casual, weekday) %>%
  summarise(qtd_passeios = n()
            ,average_duration = mean(duracao_passeio)) %>%
  arrange(member_casual, weekday)  %>%
  ggplot(aes(x = weekday, y = qtd_passeios, fill = member_casual)) +
  geom_col(position = "dodge")
```

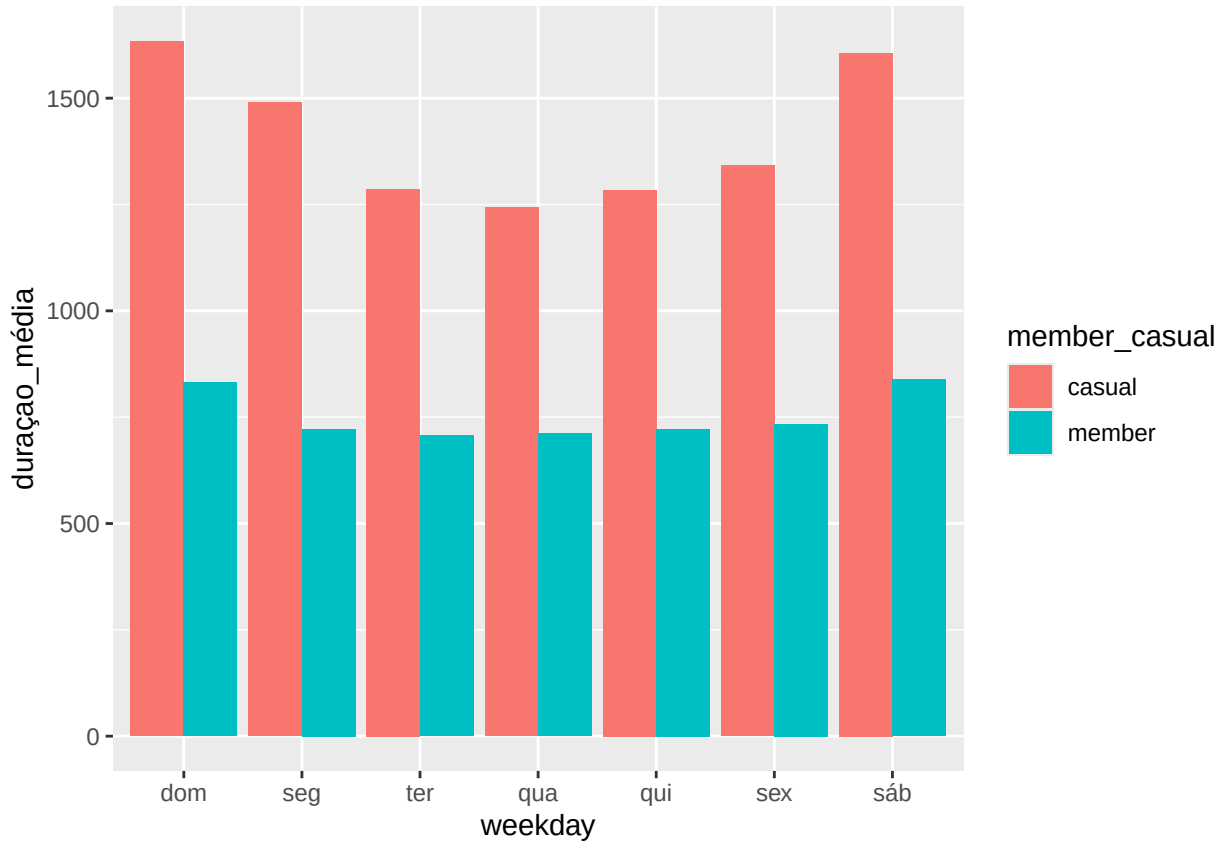
```
## `summarise()` has grouped output by 'member_casual'. You can override using the
## `.groups` argument.
```



Média de duração de passeios por dia da semana

```
all_trips_v2 %>%
  mutate(weekday = wday(started_at, label = TRUE)) %>%
  group_by(member_casual, weekday) %>%
  summarise(number_of_rides = n()
            ,duraçao_média = mean(duracao_passeio)) %>%
  arrange(member_casual, weekday)  %>%
  ggplot(aes(x = weekday, y = duraçao_média, fill = member_casual)) +
  geom_col(position = "dodge")
```

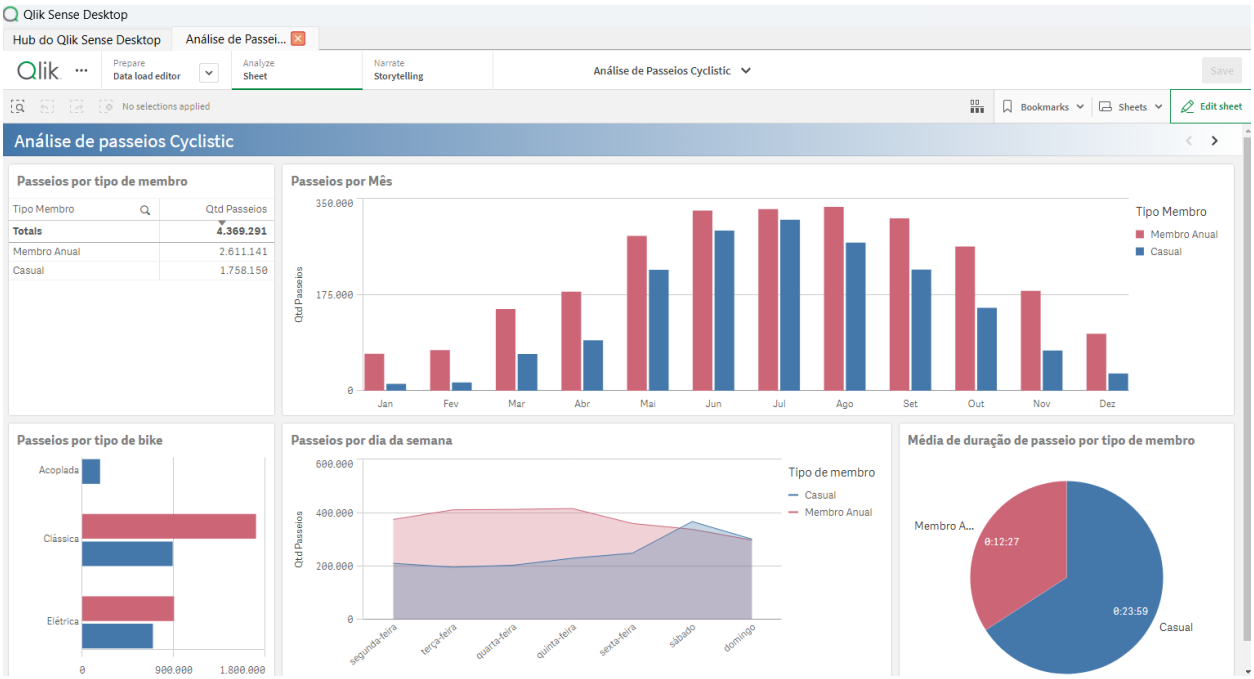
```
## `summarise()` has grouped output by 'member_casual'. You can override using the
## `.groups` argument.
```



Exportando para csv

write.csv(all_trips_v2, "C:\Users\jande\Documentos\estudo_de_caso\divvy_trip_data_2022.csv", row.names=FALSE)

Compartilhar:
A ferramenta que escolhi para compartilhamento foi Qlik Sense Desktop pois já tive contato profissional com o mesmo. O Processo de ETL neste caso não envolveu ligações entre qvds diferentes por se tratar somente de um único arquivo csv.



Agir:

Conclusão

Membros anuais utilizam mais vezes o serviço porém membros casuais têm duração de passeios maior. Suas três principais recomendações com base em sua análise:

- Programa de fidelidade para membros casuais pelo tempo de uso.
- Promoções de início de semana pois o número de corridas diminui.
- Aplicar pesquisas de satisfação com clientes e fazer uma nova comparação com os novos dados.

Todas as recomendações devem ser trabalhadas junto as partes interessadas.