Capstone

Hungdinh

2022-12-30

Steps taken to analyse Cyclist Capstone Project

(This analysis is based on the Divvy case study "'Sophisticated, Clear, and Polished': Divvy and Data Visualization" written by Kevin Hartman)

1. Collect data

Dowload datasets from: https://divvy-tripdata.s3.amazonaws.com/index.html Save the dataset into "C:/Users/Admin/Desktop/Divvy-Datatrip-12-months"

Import Datasets (12months) into RStudio as CSV files by names x202112, x202201, x202202, x202203, x202204, x202205, x202206, x202207, x202208, x202209, x202210, x202211

```
library(readr)
X202112 <- read_csv("C:/Users/Admin/Desktop/Divvy-Datatrip-12-months/202112-divvy-tripdata.csv")
```

```
## Rows: 247540 Columns: 13
## ── Column specification ────────────────────────────────────────────────────────
## Delimiter: ","
## chr  (7): ride_id, rideable_type, start_station_name, start_station_id, end_...
## dbl  (4): start_lat, start_lng, end_lat, end_lng
## dttm (2): started_at, ended_at
##
## ℹ Use `spec()` to retrieve the full column specification for this data.
## ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
X202201 <- read_csv("C:/Users/Admin/Desktop/Divvy-Datatrip-12-months/202201-divvy-tripdata.csv")
```

```
## Rows: 103770 Columns: 13
## ── Column specification ────────────────────────────────────────────────────────
## Delimiter: ","
## chr  (7): ride_id, rideable_type, start_station_name, start_station_id, end_...
## dbl  (4): start_lat, start_lng, end_lat, end_lng
## dttm (2): started_at, ended_at
##
## ℹ Use `spec()` to retrieve the full column specification for this data.
## ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
X202202 <- read_csv("C:/Users/Admin/Desktop/Divvy-Datatrip-12-months/202202-divvy-tripdata.csv")
```

```
## Rows: 115609 Columns: 13
## ── Column specification ────────────────────────────────────────────────────────
## Delimiter: ","
## chr  (7): ride_id, rideable_type, start_station_name, start_station_id, end_...
## dbl  (4): start_lat, start_lng, end_lat, end_lng
## dttm (2): started_at, ended_at
##
## ℹ Use `spec()` to retrieve the full column specification for this data.
## ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
X202203 <- read_csv("C:/Users/Admin/Desktop/Divvy-Datatrip-12-months/202203-divvy-tripdata.csv")
```

```
## Rows: 284042 Columns: 13
## ── Column specification ────────────────────────────────────────────────────────
## Delimiter: ","
## chr  (7): ride_id, rideable_type, start_station_name, start_station_id, end_...
## dbl  (4): start_lat, start_lng, end_lat, end_lng
## dttm (2): started_at, ended_at
##
## ℹ Use `spec()` to retrieve the full column specification for this data.
## ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
X202204 <- read_csv("C:/Users/Admin/Desktop/Divvy-Datatrip-12-months/202204-divvy-tripdata.csv")
```

```
## Rows: 371249 Columns: 13
## ── Column specification ────────────────────────────────────────────────────────
## Delimiter: ","
## chr  (7): ride_id, rideable_type, start_station_name, start_station_id, end_...
## dbl  (4): start_lat, start_lng, end_lat, end_lng
## dttm (2): started_at, ended_at
##
## ℹ Use `spec()` to retrieve the full column specification for this data.
## ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
X202205 <- read_csv("C:/Users/Admin/Desktop/Divvy-Datatrip-12-months/202205-divvy-tripdata.csv")
```

```
## Rows: 634858 Columns: 13
## ── Column specification ────────────────────────────────────────────────────────
## Delimiter: ","
## chr  (7): ride_id, rideable_type, start_station_name, start_station_id, end_...
## dbl  (4): start_lat, start_lng, end_lat, end_lng
## dttm (2): started_at, ended_at
##
## ℹ Use `spec()` to retrieve the full column specification for this data.
## ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
X202206 <- read_csv("C:/Users/Admin/Desktop/Divvy-Datatrip-12-months/202206-divvy-tripdata.csv")
```

```
## Rows: 769204 Columns: 13
## ── Column specification ────────────────────────────────────────────────────────
## Delimiter: ","
## chr  (7): ride_id, rideable_type, start_station_name, start_station_id, end_...
## dbl  (4): start_lat, start_lng, end_lat, end_lng
## dttm (2): started_at, ended_at
##
## ℹ Use `spec()` to retrieve the full column specification for this data.
## ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
X202207 <- read_csv("C:/Users/Admin/Desktop/Divvy-Datatrip-12-months/202207-divvy-tripdata.csv")
```

```
## Rows: 823488 Columns: 13
## ── Column specification ────────────────────────────────────────────────────────
## Delimiter: ","
## chr  (7): ride_id, rideable_type, start_station_name, start_station_id, end_...
## dbl  (4): start_lat, start_lng, end_lat, end_lng
## dttm (2): started_at, ended_at
##
## ℹ Use `spec()` to retrieve the full column specification for this data.
## ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
X202208 <- read_csv("C:/Users/Admin/Desktop/Divvy-Datatrip-12-months/202208-divvy-tripdata.csv")
```

```
## Rows: 785932 Columns: 13
## ── Column specification ────────────────────────────────────────────────────────
## Delimiter: ","
## chr  (7): ride_id, rideable_type, start_station_name, start_station_id, end_...
## dbl  (4): start_lat, start_lng, end_lat, end_lng
## dttm (2): started_at, ended_at
##
## ℹ Use `spec()` to retrieve the full column specification for this data.
## ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
X202209 <- read_csv("C:/Users/Admin/Desktop/Divvy-Datatrip-12-months/202209-divvy-publictripdata.csv")
```

```
## Rows: 701339 Columns: 13
## ── Column specification ────────────────────────────────────────────────────────
## Delimiter: ","
## chr  (7): ride_id, rideable_type, start_station_name, start_station_id, end_...
## dbl  (4): start_lat, start_lng, end_lat, end_lng
## dttm (2): started_at, ended_at
##
## ℹ Use `spec()` to retrieve the full column specification for this data.
## ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
X202210 <- read_csv("C:/Users/Admin/Desktop/Divvy-Datatrip-12-months/202210-divvy-tripdata.csv")
```

```
## Rows: 558685 Columns: 13
## ── Column specification ────────────────────────────────────────────────────────
## Delimiter: ","
## chr  (7): ride_id, rideable_type, start_station_name, start_station_id, end_...
## dbl  (4): start_lat, start_lng, end_lat, end_lng
## dttm (2): started_at, ended_at
##
## ℹ Use `spec()` to retrieve the full column specification for this data.
## ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
X202211 <- read_csv("C:/Users/Admin/Desktop/Divvy-Datatrip-12-months/202211-divvy-tripdata.csv")
```

```
## Rows: 337735 Columns: 13
## ── Column specification ────────────────────────────────────────────────────────
## Delimiter: ","
## chr  (7): ride_id, rideable_type, start_station_name, start_station_id, end_...
## dbl  (4): start_lat, start_lng, end_lat, end_lng
## dttm (2): started_at, ended_at
##
## ℹ Use `spec()` to retrieve the full column specification for this data.
## ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

2. Wrangle Data and combine into 1 single data frame

- Check if all columns in each dataset fit together in Environment pane
- Combine by bind_row() – dplyr package

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
data_trip <- bind_rows(X202112,X202201,X202202,X202203,X202204,X202205,X202206,X202207,X202208,X202209,X202210,X2
02211)
```

3. Clean up and add data to prepare for analyzing

- Add columns that list the date, month, day, and trip duration of each ride

```
 data_trip <- mutate(data_trip, date = as.Date(data_trip$started_at), weekday = weekdays(data_trip$started_at, FA
LSE), month = months(data_trip$started_at, FALSE), ride_length = as.numeric(data_trip$ended_at - data_trip$starte
d_at))
```

- Remove bad data: By summary(), ride_length show lot of negative results. Check with sum()

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

```
## [1] 100
```

These rows need to be removed.

Doing that need to create new dataframe: data_trip_v2

```
data_trip_v2 <- data_trip[!(data_trip$ride_length <0),]
```

4. Conduct analyses

- Descriptive analysis on ride_length (all figures in seconds)

```
mean(data_trip_v2$ride_length) #straight average (total ride length / rides)
```

```
## [1] 1165.316
```

```
median(data_trip_v2$ride_length) #midpoint number in the ascending array of ride lengths
```

```
## [1] 618
```

```
max(data_trip_v2$ride_length) #longest ride
```

```
## [1] 2483235
```

```
min(data_trip_v2$ride_length) #shortest ride
```

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

- Summary on ride length

```
summary(data_trip_v2$ride_length)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       0     350     618    1165    1110 2483235
```

- Aggregate on ride length between 2 kinds of customer: casual and member

```
aggregate(data_trip_v2$ride_length ~ data_trip_v2$member_casual, FUN = mean)
```

```
##   data_trip_v2$member_casual data_trip_v2$ride_length
## 1                     casual                1746.5364
## 2                     member                 762.5375
```

```
aggregate(data_trip_v2$ride_length ~ data_trip_v2$member_casual, FUN = median)
```

```
##   data_trip_v2$member_casual data_trip_v2$ride_length
## 1                     casual                      783
## 2                     member                      530
```

```
aggregate(data_trip_v2$ride_length ~ data_trip_v2$member_casual, FUN = max)
```

```
##   data_trip_v2$member_casual data_trip_v2$ride_length
## 1                     casual                  2483235
## 2                     member                    93594
```

```
aggregate(data_trip_v2$ride_length ~ data_trip_v2$member_casual, FUN = min)
```

```
##   data_trip_v2$member_casual data_trip_v2$ride_length
## 1                     casual                        0
## 2                     member                        0
```

- See the average ride time by each day for members vs casual users

```
aggregate(data_trip_v2$ride_length ~ data_trip_v2$member_casual + data_trip_v2$weekday, FUN = mean)
```

```
##    data_trip_v2$member_casual data_trip_v2$weekday data_trip_v2$ride_length
## 1                      casual               Friday                1667.3108
## 2                      member               Friday                 750.0025
## 3                      casual               Monday                1753.9351
## 4                      member               Monday                 736.4146
## 5                      casual             Saturday                1951.9865
## 6                      member             Saturday                 849.0422
## 7                      casual               Sunday                2045.0549
## 8                      member               Sunday                 842.8941
## 9                      casual             Thursday                1534.3805
## 10                     member             Thursday                 737.9848
## 11                     casual              Tuesday                1558.6146
## 12                     member              Tuesday                 728.8662
## 13                     casual            Wednesday                1482.0886
## 14                     member            Wednesday                 723.9585
```

** Notice that weekdays are out of order. Let's fix that

```
data_trip_v2$weekday <- ordered(data_trip_v2$weekday, levels=c("Sunday", "Monday", "Tuesday", "Wednesday", "Thurs
day", "Friday", "Saturday"))
```

- Run again the function to see the average ride time by each day for members vs casual users

```
aggregate(data_trip_v2$ride_length ~ data_trip_v2$member_casual + data_trip_v2$weekday, FUN = mean)
```

```
##    data_trip_v2$member_casual data_trip_v2$weekday data_trip_v2$ride_length
## 1                      casual               Sunday                2045.0549
## 2                      member               Sunday                 842.8941
## 3                      casual               Monday                1753.9351
## 4                      member               Monday                 736.4146
## 5                      casual              Tuesday                1558.6146
## 6                      member              Tuesday                 728.8662
## 7                      casual            Wednesday                1482.0886
## 8                      member            Wednesday                 723.9585
## 9                      casual             Thursday                1534.3805
## 10                     member             Thursday                 737.9848
## 11                     casual               Friday                1667.3108
## 12                     member               Friday                 750.0025
## 13                     casual             Saturday                1951.9865
## 14                     member             Saturday                 849.0422
```

*analyze ridership by type and weekday

```
data_trip_v2 %>%
  group_by(member_casual,weekday) %>%
  summarise(number_of_rides = n(), average_duration = mean(ride_length)) %>%
  arrange(member_casual,weekday)
```

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

```
## # A tibble: 14 × 4
## # Groups:   member_casual [2]
##    member_casual weekday   number_of_rides average_duration
##    <chr>         <ord>               <int>            <dbl>
##  1 casual        Sunday             392105            2045.
##  2 casual        Monday             280468            1754.
##  3 casual        Tuesday            264053            1559.
##  4 casual        Wednesday          279380            1482.
##  5 casual        Thursday           313736            1534.
##  6 casual        Friday             340496            1667.
##  7 casual        Saturday           476583            1952.
##  8 member        Sunday             390487             843.
##  9 member        Monday             476931             736.
## 10 member        Tuesday            518657             729.
## 11 member        Wednesday          537743             724.
## 12 member        Thursday           540341             738.
## 13 member        Friday             476905             750.
## 14 member        Saturday           445466             849.
```

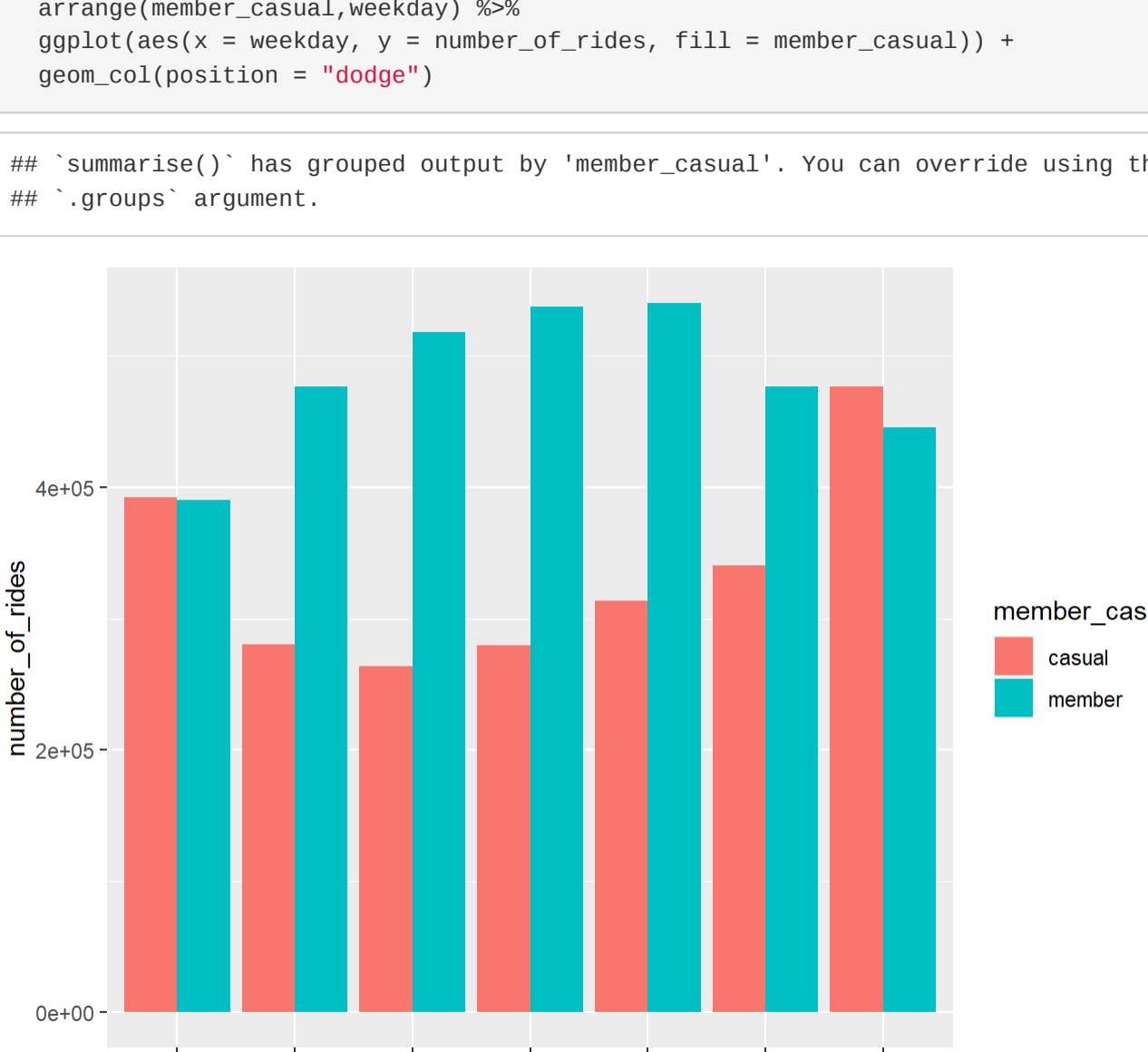
#Visualization

*Number of rides by rider type

```
library(ggplot2)
```

```
data_trip_v2 %>%
  group_by(member_casual,weekday) %>%
  summarise(number_of_rides = n(), average_duration = mean(ride_length)) %>%
  arrange(member_casual,weekday) %>%
  ggplot(aes(x = weekday, y = number_of_rides, fill = member_casual)) +
  geom_col(position = "dodge")
```

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



- Average duration

```
data_trip_v2 %>%
  group_by(member_casual,weekday) %>%
  summarise(number_of_rides = n(), average_duration = mean(ride_length)) %>%
  arrange(member_casual,weekday) %>%
  ggplot(aes(x = weekday, y = average_duration, fill = member_casual)) +
  geom_col(position = "dodge")
```

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