Trip_Bike_Data

Michael

2023-06-30

Bike Trip Data

Data Source: Coursera

Install packages

- install.packages("tidyverse")
- install.packages("dplyr")
- install.packages("readr")
- install.packages("skimr")

Install libraries

```
library(tidyverse)
```

```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
## ✓ dplyr     1.1.2     ✓ readr     2.1.4
## ✓ forcats   1.0.0     ✓ stringr   1.5.0
## ✓ ggplot2   3.4.2     ✓ tibble    3.2.1
## ✓ lubridate 1.9.2     ✓ tidyr     1.3.0
## ✓ purrr     1.0.1
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to be
come errors
```

```
library(readr)
library(dplyr)
library(skimr)
```

Read all csv files into data single frame

```
year_bike_data <- list.files(pattern = "*.csv") %>%
  map_df(~read_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.
## 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.
## 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.
## 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.
## 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.
## 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.
## 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.
## 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.
## Rows: 181806 Columns: 13
## ── Column specification ────────────────────────────────────────────────────────
## Delimiter: ","
## chr  (7): ride_id, rideable_type, start_station_name, start_station_id, end_...
## dbl  (4): start_lat, start_lng, end_lat, end_lng
## dttm (2): started_at, ended_at
##
## ℹ Use `spec()` to retrieve the full column specification for this data.
## ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
## Rows: 190301 Columns: 13
## ── Column specification ────────────────────────────────────────────────────────
## Delimiter: ","
## chr  (7): ride_id, rideable_type, start_station_name, start_station_id, end_...
## dbl  (4): start_lat, start_lng, end_lat, end_lng
## dttm (2): started_at, ended_at
##
## ℹ Use `spec()` to retrieve the full column specification for this data.
## ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
## Rows: 190445 Columns: 13
## ── Column specification ────────────────────────────────────────────────────────
## Delimiter: ","
## chr  (7): ride_id, rideable_type, start_station_name, start_station_id, end_...
## dbl  (4): start_lat, start_lng, end_lat, end_lng
## dttm (2): started_at, ended_at
##
## ℹ Use `spec()` to retrieve the full column specification for this data.
## ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
## Rows: 258678 Columns: 13
## ── Column specification ────────────────────────────────────────────────────────
## Delimiter: ","
```

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

Preview the data by column names

```
colnames(year_bike_data)
```

```
##  [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"
```

Preview the data using Glimpse function

```
glimpse(year_bike_data)
```

```
## Rows: 5,803,720
## Columns: 13
## $ ride_id            <chr> "3564070EEFD12711", "0B820C7FCF22F489", "89EEEE3229…
## $ rideable_type      <chr> "electric_bike", "classic_bike", "classic_bike", "c…
## $ started_at         <dttm> 2022-04-06 17:42:48, 2022-04-24 19:23:07, 2022-04-…
## $ ended_at           <dttm> 2022-04-06 17:54:36, 2022-04-24 19:43:17, 2022-04-…
## $ start_station_name <chr> "Paulina St & Howard St", "Wentworth Ave & Cermak R…
## $ start_station_id   <chr> "515", "13075", "TA1307000121", "13075", "TA1307000…
## $ end_station_name   <chr> "University Library (NU)", "Green St & Madison St",…
## $ end_station_id     <chr> "605", "TA1307000120", "TA1307000120", "KA170600500…
## $ start_lat          <dbl> 42.01913, 41.85308, 41.87184, 41.85308, 41.87181, 4…
## $ start_lng          <dbl> -87.67353, -87.63193, -87.64664, -87.63193, -87.646…
## $ end_lat            <dbl> 42.05294, 41.88189, 41.88189, 41.86749, 41.88224, 4…
## $ end_lng            <dbl> -87.67345, -87.64879, -87.64879, -87.63219, -87.641…
## $ member_casual      <chr> "member", "member", "member", "casual", "member", "…
```

Skim the data

```
skim_without_charts(year_bike_data)
```

Data summary

Name	year_bike_data
Number of rows	5803720

Number of columns	13
_____	
Column type frequency:	
character	7
numeric	4
POSIXct	2
_____	
Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
ride_id	0	1.00	16	16	0	5803720	0
rideable_type	0	1.00	11	13	0	3	0
start_station_name	839082	0.86	7	64	0	1699	0
start_station_id	839214	0.86	3	36	0	1315	0
end_station_name	896319	0.85	9	64	0	1723	0
end_station_id	896460	0.85	3	36	0	1320	0
member_casual	0	1.00	6	6	0	2	0

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100
start_lat	0	1	41.90	0.05	41.64	41.88	41.90	41.93	42.07
start_lng	0	1	-87.65	0.03	-87.84	-87.66	-87.64	-87.63	-87.52
end_lat	5855	1	41.90	0.07	0.00	41.88	41.90	41.93	42.37
end_lng	5855	1	-87.65	0.11	-88.14	-87.66	-87.64	-87.63	0.00

Variable type: POSIXct

skim_variable	n_missing	complete_rate	min	max	median	n_unique
started_at	0	1	2022-04-01 00:01:48	2023-03-31 23:59:28	2022-08-13 11:37:32	4874281
ended_at	0	1	2022-04-01 00:02:15	2023-04-03 11:41:11	2022-08-13 12:00:07	4887969

Get a summary of the data

```
summary(year_bike_data)
```

```
##    ride_id          rideable_type        started_at
##  Length:5803720     Length:5803720     Min.   :2022-04-01 00:01:48.00
##  Class :character   Class :character   1st Qu.:2022-06-18 23:27:00.25
##  Mode  :character   Mode  :character   Median :2022-08-13 11:37:32.00
##                                        Mean   :2022-08-25 07:04:55.95
##                                        3rd Qu.:2022-10-14 18:04:21.00
##                                        Max.   :2023-03-31 23:59:28.00
##
##     ended_at                      start_station_name start_station_id
##  Min.   :2022-04-01 00:02:15.00   Length:5803720     Length:5803720
##  1st Qu.:2022-06-18 23:51:55.75   Class :character   Class :character
##  Median :2022-08-13 12:00:07.50   Mode  :character   Mode  :character
##  Mean   :2022-08-25 07:23:54.70
##  3rd Qu.:2022-10-14 18:19:10.25
##  Max.   :2023-04-03 11:41:11.00
##
##  end_station_name   end_station_id       start_lat       start_lng
##  Length:5803720     Length:5803720     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.   :42.07   Max.   :-87.52
##
##     end_lat         end_lng       member_casual
##  Min.   : 0.00   Min.   :-88.14   Length:5803720
##  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   :5855    NA's   :5855
```

Preview rows with missing values

```
colSums(is.na(year_bike_data))
```

```
##            ride_id      rideable_type         started_at           ended_at
##                  0                  0                  0                  0
## start_station_name   start_station_id   end_station_name     end_station_id
##             839082             839214             896319             896460
##          start_lat          start_lng            end_lat            end_lng
##                  0                  0               5855               5855
##      member_casual
##                  0
```

Clean the data

Remove rows with missing values

```
year_bike_data_cleaned <- year_bike_data[complete.cases(year_bike_data), ]
```

Filter out data where started_at is greater than ended_at

```
year_bike_data_cleaned <- year_bike_data_cleaned %>%
  filter(year_bike_data_cleaned$started_at < year_bike_data_cleaned$ended_at)
```

Process the data

Add 4 new columns to indicate the following

Day the bike was hired as day_bike_used

Month the bike was hired as month_bike_used

Quarter the bike was hired as quarter_bike_used

Duration of the Trip as ride_length

```
year_bike_data_cleaned <- year_bike_data_cleaned %>%
  mutate(day_bike_used = wday(year_bike_data_cleaned$started_at)) %>%
  mutate(month_bike_used = month(year_bike_data_cleaned$started_at)) %>%
  mutate(quarter_bike_used =  quarter(year_bike_data_cleaned$started_at)) %>%
  mutate(ride_length = year_bike_data_cleaned$ended_at - year_bike_data_cleaned$started_at)
```

Glimpse new data set

```
glimpse(year_bike_data_cleaned)
```

```
## Rows: 4,482,044
## Columns: 17
## $ ride_id            <chr> "3564070EEFD12711", "0B820C7FCF22F489", "89EEEE3229…
## $ rideable_type      <chr> "electric_bike", "classic_bike", "classic_bike", "c…
## $ started_at         <dttm> 2022-04-06 17:42:48, 2022-04-24 19:23:07, 2022-04-…
## $ ended_at           <dttm> 2022-04-06 17:54:36, 2022-04-24 19:43:17, 2022-04-…
## $ start_station_name <chr> "Paulina St & Howard St", "Wentworth Ave & Cermak R…
## $ start_station_id   <chr> "515", "13075", "TA1307000121", "13075", "TA1307000…
## $ end_station_name   <chr> "University Library (NU)", "Green St & Madison St",…
## $ end_station_id     <chr> "605", "TA1307000120", "TA1307000120", "KA170600500…
## $ start_lat          <dbl> 42.01913, 41.85308, 41.87184, 41.85308, 41.87181, 4…
## $ start_lng          <dbl> -87.67353, -87.63193, -87.64664, -87.63193, -87.646…
## $ end_lat            <dbl> 42.05294, 41.88189, 41.88189, 41.86749, 41.88224, 4…
## $ end_lng            <dbl> -87.67345, -87.64879, -87.64879, -87.63219, -87.641…
## $ member_casual      <chr> "member", "member", "member", "casual", "member", "…
## $ day_bike_used      <dbl> 4, 1, 4, 6, 7, 5, 2, 3, 6, 6, 7, 4, 4, 7, 4, 2, 2, …
## $ month_bike_used    <dbl> 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, …
## $ quarter_bike_used  <int> 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, …
## $ ride_length        <drtn> 708 secs, 1210 secs, 368 secs, 563 secs, 341 secs,…
```

Split data set for members and casual users

```
for (variable in unique(year_bike_data_cleaned$member_casual)) {
  assign( variable, year_bike_data_cleaned %>% filter (member_casual == variable), envir = .G
lobalEnv)
}
```

Glimpse the member data set

```
glimpse(member)
```

```
## Rows: 2,709,717
## Columns: 17
## $ ride_id            <chr> "3564070EEFD12711", "0B820C7FCF22F489", "89EEEE3229…
## $ rideable_type      <chr> "electric_bike", "classic_bike", "classic_bike", "e…
## $ started_at         <dttm> 2022-04-06 17:42:48, 2022-04-24 19:23:07, 2022-04-…
## $ ended_at           <dttm> 2022-04-06 17:54:36, 2022-04-24 19:43:17, 2022-04-…
## $ start_station_name <chr> "Paulina St & Howard St", "Wentworth Ave & Cermak R…
## $ start_station_id   <chr> "515", "13075", "TA1307000121", "TA1307000121", "15…
## $ end_station_name   <chr> "University Library (NU)", "Green St & Madison St",…
## $ end_station_id     <chr> "605", "TA1307000120", "TA1307000120", "TA130500003…
## $ start_lat          <dbl> 42.01913, 41.85308, 41.87184, 41.87181, 41.88462, 4…
## $ start_lng          <dbl> -87.67353, -87.63193, -87.64664, -87.64657, -87.644…
## $ end_lat            <dbl> 42.05294, 41.88189, 41.88189, 41.88224, 41.87926, 4…
## $ end_lng            <dbl> -87.67345, -87.64879, -87.64879, -87.64107, -87.639…
## $ member_casual      <chr> "member", "member", "member", "member", "member", "…
## $ day_bike_used      <dbl> 4, 1, 4, 7, 5, 2, 3, 6, 6, 4, 4, 2, 2, 5, 7, 3, 3, …
## $ month_bike_used    <dbl> 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, …
## $ quarter_bike_used  <int> 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, …
## $ ride_length        <drtn> 708 secs, 1210 secs, 368 secs, 341 secs, 258 secs,…
```

Glimpse the casual data set

```
glimpse(casual)
```

```
## Rows: 1,772,327
## Columns: 17
## $ ride_id            <chr> "84D4751AEB31888D", "F04AF7DB8CE260D1", "B975D67976…
## $ rideable_type      <chr> "classic_bike", "electric_bike", "classic_bike", "e…
## $ started_at         <dttm> 2022-04-22 21:14:06, 2022-04-23 15:13:07, 2022-04-…
## $ ended_at           <dttm> 2022-04-22 21:23:29, 2022-04-23 15:26:53, 2022-04-…
## $ start_station_name <chr> "Wentworth Ave & Cermak Rd", "Wentworth Ave & Cerma…
## $ start_station_id   <chr> "13075", "13075", "TA1307000121", "624", "TA1307000…
## $ end_station_name   <chr> "Delano Ct & Roosevelt Rd", "Calumet Ave & 18th St"…
## $ end_station_id     <chr> "KA1706005007", "13102", "TA1309000001", "13016", "…
## $ start_lat          <dbl> 41.85308, 41.85313, 41.87184, 41.87613, 41.89147, 4…
## $ start_lng          <dbl> -87.63193, -87.63187, -87.64664, -87.62974, -87.626…
## $ end_lat            <dbl> 41.86749, 41.85761, 41.86488, 41.89435, 41.87725, 4…
## $ end_lng            <dbl> -87.63219, -87.61941, -87.64707, -87.62280, -87.639…
## $ member_casual      <chr> "casual", "casual", "casual", "casual", "casual", "…
## $ day_bike_used      <dbl> 6, 7, 4, 7, 7, 3, 1, 6, 1, 1, 2, 2, 2, 6, 3, 3, 4, …
## $ month_bike_used    <dbl> 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, …
## $ quarter_bike_used  <int> 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, …
## $ ride_length        <drtn> 563 secs, 826 secs, 312 secs, 617 secs, 3313 secs,…
```

Analysis

Summarize the stats for member riders

```
member %>%
    summarise(min(member$ride_length), mean(member$ride_length), max(member$ride_length))
```

```
## # A tibble: 1 × 3
##   `min(member$ride_length)` `mean(member$ride_length)` `max(member$ride_length)`
##   <drtn>                    <drtn>                     <drtn>
## 1 1 secs                    736.088 secs               89872 secs
```

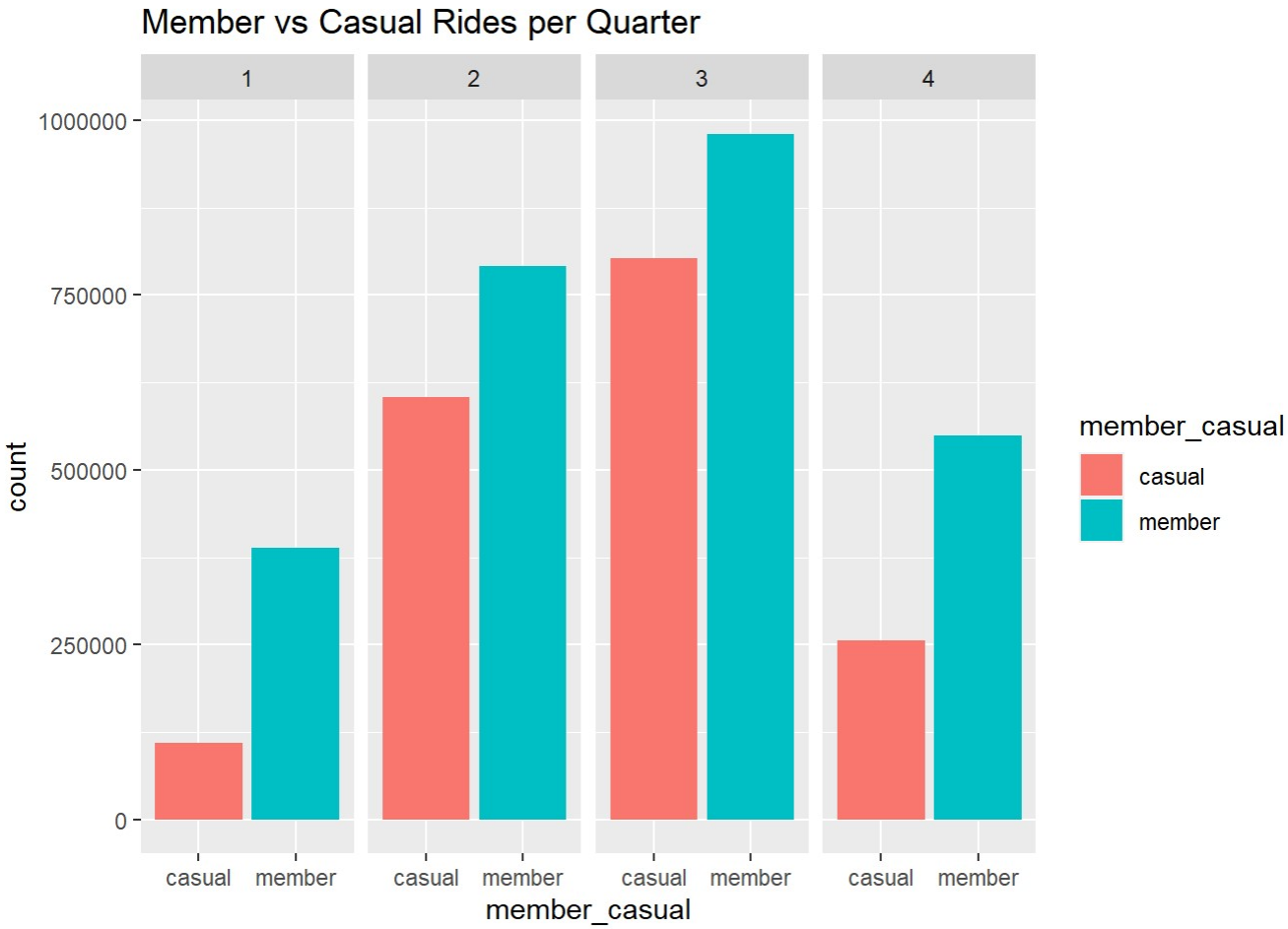
Summarize the stats for casual riders

```
casual %>%
  summarise(min(casual$ride_length), mean(casual$ride_length), max(casual$ride_length))
```

```
## # A tibble: 1 × 3
##   `min(casual$ride_length)` `mean(casual$ride_length)` `max(casual$ride_length)`
##   <drtn>                    <drtn>                     <drtn>
## 1 1 secs                    1400.035 secs              1922127 secs
```

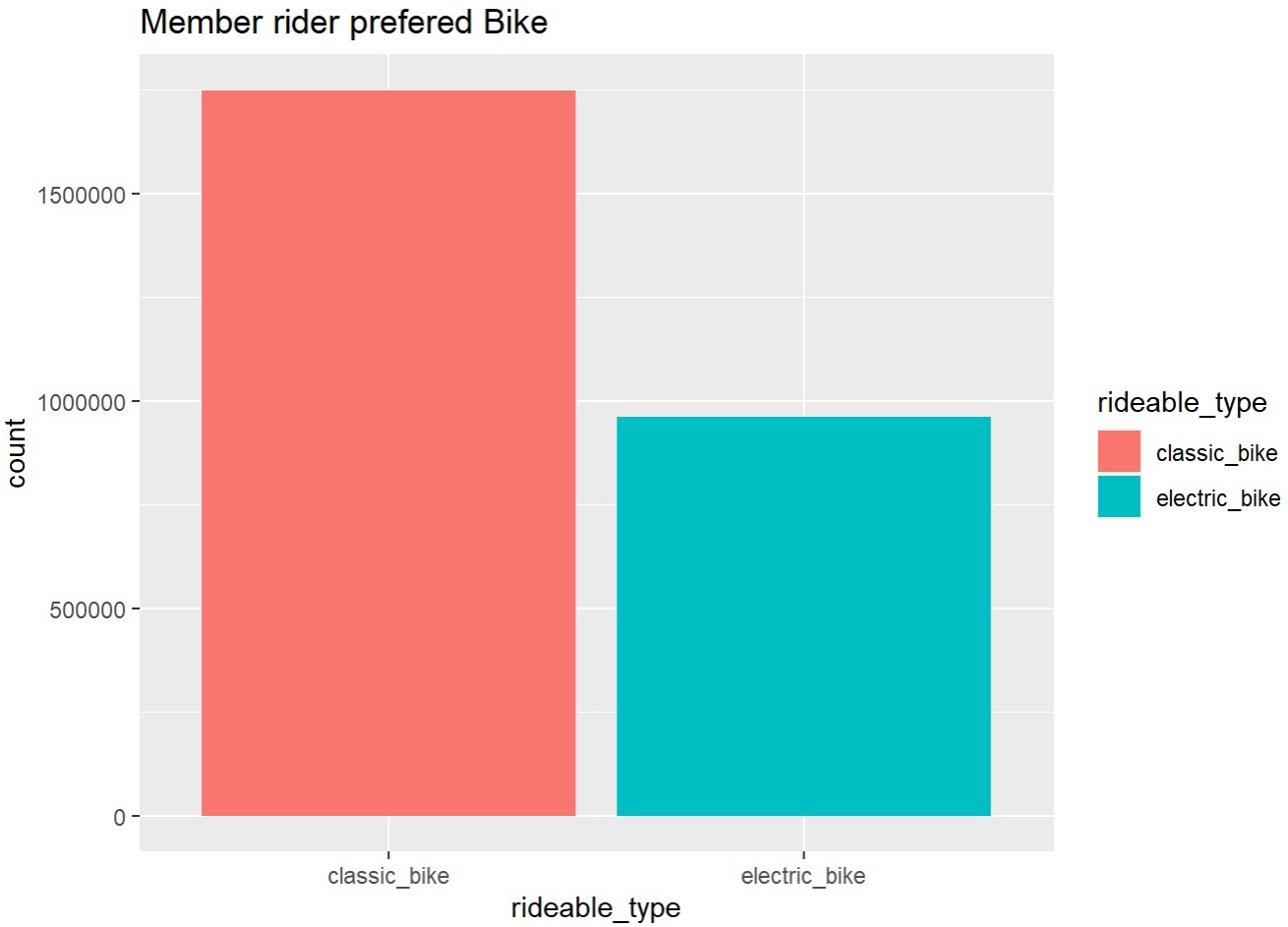
Make quarterly comparison of Member riders vs Casual riders

```
year_bike_data_cleaned %>%
  ggplot(aes(x = member_casual, fill = member_casual)) +
  geom_bar() +
  facet_grid(~quarter_bike_used) +
  labs(title = "Member vs Casual Rides per Quarter")
```



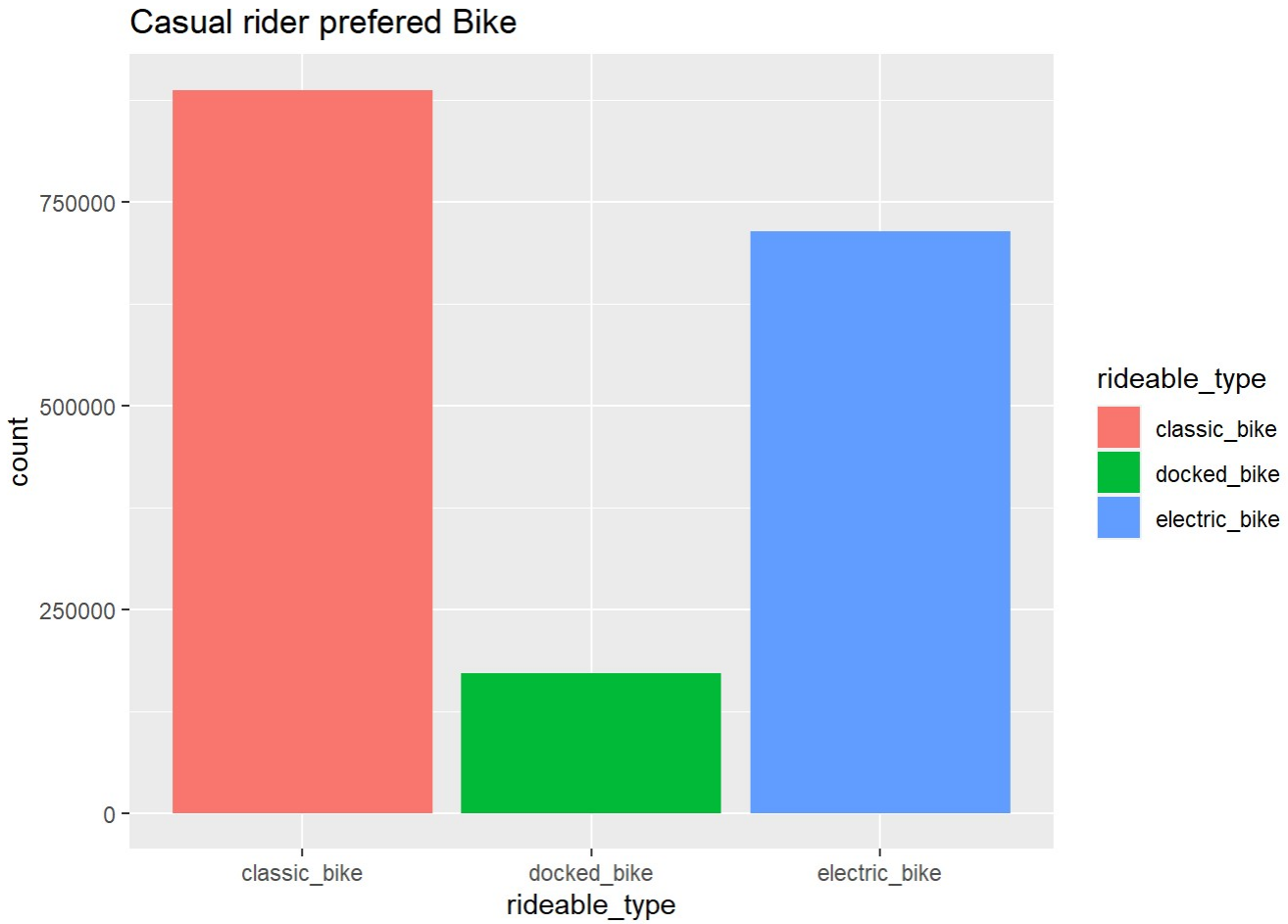
Ascertain which bikes member riders prefer

```
member %>%
  ggplot(aes(x = rideable_type, fill = rideable_type)) +
  geom_bar() +
  labs(title = "Member rider prefered Bike")
```



Ascertain which bikes casual riders prefer

```
casual %>%
  ggplot(aes(x = rideable_type, fill = rideable_type)) +
  geom_bar() +
  labs(title = "Casual rider prefered Bike")
```



Export Cleaned data sets as csv files

```
write.csv(member, "C:\\Users\\M3\\Downloads\\Study_Documents\\Case_study\\Clean_data\\member.
csv")
write.csv(casual, "C:\\Users\\M3\\Downloads\\Study_Documents\\Case_study\\Clean_data\\casual.
csv")
write.csv(year_bike_data_cleaned, "C:\\Users\\M3\\Downloads\\Study_Documents\\Case_study\\Cle
an_data\\year_bike_data_cleaned.csv")
```

Conclusion

About 22% of the data had incomplete information

Incomplete data was removed from the data set

The following observations where made from the complete data

- There are generally more member riders than casual riders in the scheme
- Quarter 1 and quarter 4 are the least busiest likely due to weather
- We can see an uptick in users in quarter 2, climaxing at quarter 3
- The best time to run promotions would then be in quarter 2 and 3 when demand is high

- Also curiously we see docked bikes are only used by casual users, that trend needs further analysis
- The classic bike still remains the most used bike by both casual and member riders