Divvy Bikes

Taras The Analyst

2023-02-21

Divvy Bikes - Capstone Project v2

Google Data Analytics Certificate Course @Coursera

The initial dataset is located here. The capstone project required to get data for 2022 - 12 monthly archive files from a data bucket at AWS. The initial files were downloaded, examined, cleaned and transformed for the project, and stored as a separate dataset in Kaggle.

STEP 0 PREPARE THE PLAYGROUND

Install required packages: tidyverse, ggplot2, lubridate.

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.4.0     v purrr   1.0.1
## v tibble  3.1.8     v dplyr   1.1.0
## v tidyr   1.3.0     v stringr 1.5.0
## v readr   2.1.3     v forcats 1.0.0
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
##
## Attaching package: 'lubridate'
##
##
## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union
```

STEP 1: UPLOAD DATA

Upload Divvy datasets (csv files)

```
## [1] "C:/Users/taras.khamardiuk/Documents"
```

```
## Rows: 103770 Columns: 7
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (3): rideable_type, member_casual, date
```

```
## dbl (4): year, month, weekday, trip_length_raw
##
## 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.
```

STEP 2: WRANGLE DATA AND COMBINE INTO A SINGLE FILE

Table names need to be checked before merging into one file. While the names don't have to be in the same order, they DO need to match perfectly before joining them into one file.

```
colnames(rides_2022_01)
```

```
## [1] "rideable_type"   "member_casual"   "year"            "month"
## [5] "weekday"         "date"            "trip_length_raw"
```

```
colnames(rides_2022_02)
```

```
## [1] "rideable_type"   "member_casual"   "year"            "month"
## [5] "weekday"         "date"            "trip_length_raw"
```

```
colnames(rides_2022_03)
```

```
## [1] "rideable_type"   "member_casual"   "year"            "month"
## [5] "weekday"         "date"            "trip_length_raw"
```

```
colnames(rides_2022_04)
```

```
## [1] "rideable_type"   "member_casual"   "year"            "month"
## [5] "weekday"         "date"            "trip_length_raw"
```

```
colnames(rides_2022_05)
```

```
## [1] "rideable_type"   "member_casual"   "year"            "month"
## [5] "weekday"         "date"            "trip_length_raw"
```

```
colnames(rides_2022_06)
```

```
## [1] "rideable_type"   "member_casual"   "year"            "month"
## [5] "weekday"         "date"            "trip_length_raw"
```

```
colnames(rides_2022_07)
```

```
## [1] "rideable_type"   "member_casual"   "year"            "month"
## [5] "weekday"         "date"            "trip_length_raw"
```

```
colnames(rides_2022_08)
```

```
## [1] "rideable_type"   "member_casual"   "year"            "month"
## [5] "weekday"         "date"            "trip_length_raw"
```

```
colnames(rides_2022_09)
```

```
## [1] "rideable_type"   "member_casual"   "year"            "month"
## [5] "weekday"         "date"            "trip_length_raw"
```

```
colnames(rides_2022_10)
```

```
## [1] "rideable_type"   "member_casual"   "year"            "month"
## [5] "weekday"         "date"            "trip_length_raw"
```

```
colnames(rides_2022_11)
```

```
## [1] "rideable_type"   "member_casual"   "year"            "month"
## [5] "weekday"         "date"            "trip_length_raw"
```

```
colnames(rides_2022_12)
```

```
## [1] "rideable_type"   "member_casual"   "year"            "month"
## [5] "weekday"         "date"            "trip_length_raw"
```

Renaming the columns in a bit more user-friendly manner...

```
(rides_2022_01 <- rename(rides_2022_01
                         ,bike_type = rideable_type
                         ,user_type = member_casual))
```

```
## # A tibble: 103,770 x 7
##    bike_type     user_type  year month weekday date     trip_length_raw
##    <chr>         <chr>     <dbl> <dbl>   <dbl> <chr>              <dbl>
##  1 electric_bike casual     2022     1       4 1/4/2022         0.00205
##  2 electric_bike casual     2022     1       1 1/1/2022         0.00302
##  3 classic_bike  member     2022     1       2 1/2/2022         0.00302
##  4 classic_bike  casual     2022     1       2 1/2/2022         0.0104
##  5 classic_bike  member     2022     1       4 1/4/2022         0.00419
##  6 classic_bike  member     2022     1       2 1/2/2022         0.00234
##  7 classic_bike  member     2022     1       7 1/7/2022         0.0115
##  8 classic_bike  member     2022     1       6 1/6/2022         0.00838
##  9 electric_bike member     2022     1       1 1/1/2022         0.0177
## 10 classic_bike  member     2022     1       5 1/5/2022         0.00513
## # ... with 103,760 more rows
```

```
(rides_2022_02 <- rename(rides_2022_02
                         ,bike_type = rideable_type
                         ,user_type = member_casual))
```

```
## # A tibble: 115,609 x 7
##    bike_type     user_type  year month weekday date     trip_length_raw
##    <chr>         <chr>     <dbl> <dbl>   <dbl> <chr>              <dbl>
##  1 classic_bike  member     2022     2       6 2/6/2022         0.0106
```

```
##  2 classic_bike  member     2022     2       7 2/7/2022         0.00308
##  3 classic_bike  member     2022     2       5 2/5/2022         0.00947
##  4 classic_bike  member     2022     2       1 2/1/2022         0.00483
##  5 classic_bike  member     2022     2       3 2/3/2022         0.00201
##  6 classic_bike  member     2022     2       1 2/1/2022         0.0111
##  7 classic_bike  member     2022     2       1 2/1/2022         0.00265
##  8 classic_bike  member     2022     2       2 2/2/2022         0.00605
##  9 electric_bike member     2022     2       5 2/5/2022         0.00391
## 10 classic_bike  member     2022     2       7 2/7/2022         0.00405
## # ... with 115,599 more rows
```

```
(rides_2022_03 <- rename(rides_2022_03
                         ,bike_type = rideable_type
                         ,user_type = member_casual))
```

```
## # A tibble: 284,042 x 7
##    bike_type     user_type  year month weekday date     trip_length_raw
##    <chr>         <chr>     <dbl> <dbl>   <dbl> <chr>              <dbl>
##  1 classic_bike  member     2022     3       1 3/1/2022         0.00436
##  2 electric_bike member     2022     3       3 3/3/2022         0.00438
##  3 classic_bike  member     2022     3       3 3/3/2022         0.00192
##  4 classic_bike  member     2022     3       2 3/2/2022         0.00681
##  5 classic_bike  member     2022     3       1 3/1/2022         0.0293
##  6 classic_bike  member     2022     3       1 3/1/2022         0.00326
##  7 electric_bike member     2022     3       4 3/4/2022         0.00228
##  8 classic_bike  member     2022     3       6 3/6/2022         0.00784
##  9 electric_bike casual     2022     3       4 3/4/2022         0.00744
## 10 classic_bike  member     2022     3       5 3/5/2022         0.00919
## # ... with 284,032 more rows
```

```
(rides_2022_04 <- rename(rides_2022_04
                         ,bike_type = rideable_type
                         ,user_type = member_casual))
```

```
## # A tibble: 371,249 x 7
##    bike_type     user_type  year month weekday date     trip_length_raw
##    <chr>         <chr>     <dbl> <dbl>   <dbl> <chr>              <dbl>
##  1 electric_bike member     2022     4       3 4/3/2022         0.00819
##  2 classic_bike  member     2022     4       7 4/7/2022         0.0140
##  3 classic_bike  member     2022     4       3 4/3/2022         0.00426
##  4 classic_bike  casual     2022     4       5 4/5/2022         0.00652
##  5 electric_bike member     2022     4       6 4/6/2022         0.00395
##  6 classic_bike  member     2022     4       4 4/4/2022         0.00299
##  7 classic_bike  member     2022     4       1 4/1/2022         0.00322
##  8 classic_bike  member     2022     4       2 4/2/2022         0.00861
##  9 electric_bike member     2022     4       5 4/5/2022         0.000370
## 10 electric_bike member     2022     4       5 4/5/2022         0.00103
## # ... with 371,239 more rows
```

```
(rides_2022_05 <- rename(rides_2022_05
                         ,bike_type = rideable_type
                         ,user_type = member_casual))
```

```
## # A tibble: 634,858 x 7
##    bike_type     user_type  year month weekday date     trip_length_raw
##    <chr>         <chr>     <dbl> <dbl>   <dbl> <chr>              <dbl>
##  1 classic_bike  member     2022     5       1 5/1/2022         0.0232
##  2 classic_bike  member     2022     5       3 5/3/2022         0.0263
##  3 classic_bike  member     2022     5       4 5/4/2022         0.0152
##  4 classic_bike  member     2022     5       2 5/2/2022         0.00604
##  5 classic_bike  member     2022     5       2 5/2/2022         0.00348
##  6 classic_bike  member     2022     5       3 5/3/2022         0.00497
##  7 classic_bike  member     2022     5       5 5/5/2022         0.00617
##  8 docked_bike   casual     2022     5       7 5/7/2022         0.00845
##  9 classic_bike  member     2022     5       1 5/1/2022         0.0116
## 10 electric_bike member     2022     5       3 5/3/2022         0.00102
## # ... with 634,848 more rows
```

```
(rides_2022_06 <- rename(rides_2022_06
                         ,bike_type = rideable_type
                         ,user_type = member_casual))
```

```
## # A tibble: 769,204 x 7
##    bike_type     user_type  year month weekday date     trip_length_raw
##    <chr>         <chr>     <dbl> <dbl>   <dbl> <chr>              <dbl>
##  1 electric_bike casual     2022     6       4 6/4/2022        0.00512
##  2 electric_bike casual     2022     6       4 6/4/2022        0.00528
##  3 electric_bike casual     2022     6       4 6/4/2022        0.00936
##  4 electric_bike casual     2022     6       4 6/4/2022        0.00299
##  5 electric_bike casual     2022     6       3 6/3/2022        0.00587
##  6 electric_bike casual     2022     6       4 6/4/2022        0.0112
##  7 electric_bike casual     2022     6       4 6/4/2022        0.0180
##  8 electric_bike casual     2022     6       4 6/4/2022        0.0601
##  9 electric_bike casual     2022     6       4 6/4/2022        0.000544
## 10 electric_bike casual     2022     6       4 6/4/2022        0.00237
## # ... with 769,194 more rows
```

```
(rides_2022_07 <- rename(rides_2022_07
                         ,bike_type = rideable_type
                         ,user_type = member_casual))
```

```
## # A tibble: 823,488 x 7
##    bike_type     user_type  year month weekday date     trip_length_raw
##    <chr>         <chr>     <dbl> <dbl>   <dbl> <chr>              <dbl>
##  1 classic_bike  member     2022     7       2 7/2/2022         0.00816
##  2 classic_bike  casual     2022     7       2 7/2/2022         0.00131
##  3 classic_bike  casual     2022     7       7 7/7/2022         0.00536
##  4 classic_bike  casual     2022     7       7 7/7/2022         0.0406
##  5 classic_bike  member     2022     7       3 7/3/2022         0.0183
##  6 electric_bike member     2022     7       5 7/5/2022         0.00605
##  7 classic_bike  member     2022     7       1 7/1/2022         0.00797
##  8 classic_bike  casual     2022     7       4 7/4/2022         0.0214
##  9 classic_bike  member     2022     7       7 7/7/2022         0.00385
## 10 electric_bike member     2022     7       7 7/7/2022         0.00795
## # ... with 823,478 more rows
```

```
(rides_2022_08 <- rename(rides_2022_08
                        ,bike_type = rideable_type
                        ,user_type = member_casual))
```

```
## # A tibble: 785,932 x 7
##    bike_type     user_type  year month weekday date     trip_length_raw
##    <chr>         <chr>     <dbl> <dbl>   <dbl> <chr>              <dbl>
##  1 electric_bike casual     2022     8       7 8/7/2022         0.00522
##  2 electric_bike casual     2022     8       1 8/1/2022         0.00975
##  3 electric_bike casual     2022     8       1 8/1/2022         0.00745
##  4 electric_bike casual     2022     8       1 8/1/2022         0.0105
##  5 electric_bike casual     2022     8       7 8/7/2022         0.00407
##  6 electric_bike casual     2022     8       1 8/1/2022         0.00904
##  7 electric_bike casual     2022     8       1 8/1/2022         0.00620
##  8 electric_bike casual     2022     8       7 8/7/2022         0.0125
##  9 electric_bike casual     2022     8       7 8/7/2022         0.00791
## 10 electric_bike casual     2022     8       7 8/7/2022         0.00774
## # ... with 785,922 more rows
```

```
(rides_2022_09 <- rename(rides_2022_09
                        ,bike_type = rideable_type
                        ,user_type = member_casual))
```

```
## # A tibble: 701,339 x 7
##    bike_type     user_type  year month weekday date     trip_length_raw
##    <chr>         <chr>     <dbl> <dbl>   <dbl> <chr>              <dbl>
##  1 electric_bike casual     2022     9       4 9/4/2022         0.00189
##  2 electric_bike casual     2022     9       4 9/4/2022         0.00227
##  3 electric_bike casual     2022     9       4 9/4/2022         0.000255
##  4 electric_bike casual     2022     9       4 9/4/2022         0.00699
##  5 electric_bike casual     2022     9       4 9/4/2022         0.00168
##  6 electric_bike casual     2022     9       4 9/4/2022         0.0116
##  7 electric_bike casual     2022     9       4 9/4/2022         0.00507
##  8 electric_bike casual     2022     9       4 9/4/2022         0.00436
##  9 electric_bike casual     2022     9       4 9/4/2022         0.00791
## 10 electric_bike casual     2022     9       4 9/4/2022         0.0105
## # ... with 701,329 more rows
```

```
(rides_2022_10 <- rename(rides_2022_10
                        ,bike_type = rideable_type
                        ,user_type = member_casual))
```

```
## # A tibble: 558,685 x 7
##    bike_type     user_type  year month weekday date      trip_length_raw
##    <chr>         <chr>     <dbl> <dbl>   <dbl> <chr>               <dbl>
##  1 classic_bike  member     2022    10       5 10/5/2022         0.00427
##  2 electric_bike casual     2022    10       6 10/6/2022         0.0137
##  3 electric_bike member     2022    10       3 10/3/2022         0.00544
##  4 electric_bike member     2022    10       1 10/1/2022         0.00432
##  5 classic_bike  casual     2022    10       4 10/4/2022         0.0314
##  6 electric_bike casual     2022    10       4 10/4/2022         0.00405
```

```
##  7 electric_bike member     2022    10       4 10/4/2022         0.00260
##  8 classic_bike  member     2022    10       3 10/3/2022         0.00541
##  9 classic_bike  casual     2022    10       6 10/6/2022         0.00678
## 10 electric_bike member     2022    10       1 10/1/2022         0.00612
## # ... with 558,675 more rows
```

```
(rides_2022_11 <- rename(rides_2022_11
                         ,bike_type = rideable_type
                         ,user_type = member_casual))
```

```
## # A tibble: 337,735 x 7
##    bike_type     user_type  year month weekday date      trip_length_raw
##    <chr>         <chr>     <dbl> <dbl>   <dbl> <chr>               <dbl>
##  1 electric_bike member     2022    11       4 11/4/2022         0.00662
##  2 classic_bike  member     2022    11       5 11/5/2022         0.0101
##  3 classic_bike  member     2022    11       1 11/1/2022         0.00980
##  4 classic_bike  member     2022    11       5 11/5/2022         0.0109
##  5 classic_bike  member     2022    11       2 11/2/2022         0.0128
##  6 classic_bike  member     2022    11       5 11/5/2022         0.00819
##  7 classic_bike  member     2022    11       7 11/7/2022         0.00661
##  8 classic_bike  member     2022    11       2 11/2/2022         0.00612
##  9 electric_bike member     2022    11       7 11/7/2022         0.00661
## 10 classic_bike  member     2022    11       2 11/2/2022         0.00443
## # ... with 337,725 more rows
```

```
(rides_2022_12 <- rename(rides_2022_12
                         ,bike_type = rideable_type
                         ,user_type = member_casual))
```

```
## # A tibble: 181,806 x 7
##    bike_type     user_type  year month weekday date      trip_length_raw
##    <chr>         <chr>     <dbl> <dbl>   <dbl> <chr>               <dbl>
##  1 electric_bike member     2022    12       1 12/1/2022         0.00644
##  2 classic_bike  casual     2022    12       7 12/7/2022         0.0182
##  3 electric_bike member     2022    12       2 12/2/2022         0.00840
##  4 classic_bike  member     2022    12       2 12/2/2022         0.0202
##  5 classic_bike  casual     2022    12       3 12/3/2022         0.00985
##  6 electric_bike member     2022    12       5 12/5/2022         0.00656
##  7 classic_bike  member     2022    12       2 12/2/2022         0.0124
##  8 classic_bike  member     2022    12       2 12/2/2022         0.00601
##  9 classic_bike  member     2022    12       2 12/2/2022         0.0131
## 10 classic_bike  member     2022    12       3 12/3/2022         0.00922
## # ... with 181,796 more rows
```

Some inspection of names made is made.

```
# inspect the renamed columns
str(rides_2022_01)
```

```
## spc_tbl_ [103,770 x 7] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ bike_type      : chr [1:103770] "electric_bike" "electric_bike" "classic_bike" "classic_bike" ...
##  $ user_type      : chr [1:103770] "casual" "casual" "member" "casual" ...
```

```
##  $ year           : num [1:103770] 2022 2022 2022 2022 2022 ...
##  $ month          : num [1:103770] 1 1 1 1 1 1 1 1 1 1 ...
##  $ weekday        : num [1:103770] 4 1 2 2 4 2 7 6 1 5 ...
##  $ date           : chr [1:103770] "1/4/2022" "1/1/2022" "1/2/2022" "1/2/2022" ...
##  $ trip_length_raw: num [1:103770] 0.00205 0.00302 0.00302 0.01037 0.00419 ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   rideable_type = col_character(),
##   ..   member_casual = col_character(),
##   ..   year = col_double(),
##   ..   month = col_double(),
##   ..   weekday = col_double(),
##   ..   date = col_character(),
##   ..   trip_length_raw = col_double()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

```
str(rides_2022_02)
```

```
## spc_tbl_ [115,609 x 7] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ bike_type      : chr [1:115609] "classic_bike" "classic_bike" "classic_bike" "classic_bike" ...
##  $ user_type      : chr [1:115609] "member" "member" "member" "member" ...
##  $ year           : num [1:115609] 2022 2022 2022 2022 2022 ...
##  $ month          : num [1:115609] 2 2 2 2 2 2 2 2 2 2 ...
##  $ weekday        : num [1:115609] 6 7 5 1 3 1 1 2 5 7 ...
##  $ date           : chr [1:115609] "2/6/2022" "2/7/2022" "2/5/2022" "2/1/2022" ...
##  $ trip_length_raw: num [1:115609] 0.01059 0.00308 0.00947 0.00483 0.00201 ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   rideable_type = col_character(),
##   ..   member_casual = col_character(),
##   ..   year = col_double(),
##   ..   month = col_double(),
##   ..   weekday = col_double(),
##   ..   date = col_character(),
##   ..   trip_length_raw = col_double()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

```
str(rides_2022_08)
```

```
## spc_tbl_ [785,932 x 7] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ bike_type      : chr [1:785932] "electric_bike" "electric_bike" "electric_bike" "electric_bike" .
##  $ user_type      : chr [1:785932] "casual" "casual" "casual" "casual" ...
##  $ year           : num [1:785932] 2022 2022 2022 2022 2022 ...
##  $ month          : num [1:785932] 8 8 8 8 8 8 8 8 8 8 ...
##  $ weekday        : num [1:785932] 7 1 1 1 7 1 1 7 7 7 ...
##  $ date           : chr [1:785932] "8/7/2022" "8/1/2022" "8/1/2022" "8/1/2022" ...
##  $ trip_length_raw: num [1:785932] 0.00522 0.00975 0.00745 0.01045 0.00407 ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   rideable_type = col_character(),
##   ..   member_casual = col_character(),
```

```
##   ..   year = col_double(),
##   ..   month = col_double(),
##   ..   weekday = col_double(),
##   ..   date = col_character(),
##   ..   trip_length_raw = col_double()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

Then, the files are merged into a united dataset.

```
# Stack individual quarter's data frames into one big data frame
all_rides <- bind_rows(rides_2022_01,rides_2022_02, rides_2022_03,rides_2022_04,
                       rides_2022_05,rides_2022_06,rides_2022_07,rides_2022_08,
                       rides_2022_09,rides_2022_10,rides_2022_11,rides_2022_12)
```

And the whole picture of a merged table is like this.

```
# get the whole picture on a merged table
str(all_rides)
```

```
## spc_tbl_ [5,667,717 x 7] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ bike_type      : chr [1:5667717] "electric_bike" "electric_bike" "classic_bike" "classic_bike" ..
##  $ user_type      : chr [1:5667717] "casual" "casual" "member" "casual" ...
##  $ year           : num [1:5667717] 2022 2022 2022 2022 2022 ...
##  $ month          : num [1:5667717] 1 1 1 1 1 1 1 1 1 1 ...
##  $ weekday        : num [1:5667717] 4 1 2 2 4 2 7 6 1 5 ...
##  $ date           : chr [1:5667717] "1/4/2022" "1/1/2022" "1/2/2022" "1/2/2022" ...
##  $ trip_length_raw: num [1:5667717] 0.00205 0.00302 0.00302 0.01037 0.00419 ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   rideable_type = col_character(),
##   ..   member_casual = col_character(),
##   ..   year = col_double(),
##   ..   month = col_double(),
##   ..   weekday = col_double(),
##   ..   date = col_character(),
##   ..   trip_length_raw = col_double()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

The tables needs a small shape-up for further manipulations.

```
# now let's reshape a bit the table existing
# creating ordered day names and factor attributes for weekday parameter
all_rides$day <- ifelse(all_rides$weekday == 1, "Monday",
              ifelse(all_rides$weekday == 2, "Tuesday",
              ifelse(all_rides$weekday == 3, "Wednesday",
              ifelse(all_rides$weekday == 4, "Thursday",
              ifelse(all_rides$weekday == 5, "Friday",
              ifelse(all_rides$weekday == 6, "Saturday",
              ifelse(all_rides$weekday == 7, "Sunday", "N/A")))))))
all_rides$day <- factor(all_rides$day)
```

```
all_rides$day <- ordered(all_rides$day, c("Monday", "Tuesday", "Wednesday", "Thursday",
                                          "Friday", "Saturday", "Sunday"))
str(all_rides)
```

```
## spc_tbl_ [5,667,717 x 8] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ bike_type      : chr [1:5667717] "electric_bike" "electric_bike" "classic_bike" "classic_bike" ..
##  $ user_type      : chr [1:5667717] "casual" "casual" "member" "casual" ...
##  $ year           : num [1:5667717] 2022 2022 2022 2022 2022 ...
##  $ month          : num [1:5667717] 1 1 1 1 1 1 1 1 1 1 ...
##  $ weekday        : num [1:5667717] 4 1 2 2 4 2 7 6 1 5 ...
##  $ date           : chr [1:5667717] "1/4/2022" "1/1/2022" "1/2/2022" "1/2/2022" ...
##  $ trip_length_raw: num [1:5667717] 0.00205 0.00302 0.00302 0.01037 0.00419 ...
##  $ day            : Ord.factor w/ 7 levels "Monday"<"Tuesday"<..: 4 1 2 2 4 2 7 6 1 5 ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   rideable_type = col_character(),
##   ..   member_casual = col_character(),
##   ..   year = col_double(),
##   ..   month = col_double(),
##   ..   weekday = col_double(),
##   ..   date = col_character(),
##   ..   trip_length_raw = col_double()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

```
# creating ordered month names and factor attributes for month parameter
all_rides$month_name <- ifelse(all_rides$month == 1, "January",
                  ifelse(all_rides$month == 2, "February",
                  ifelse(all_rides$month == 3, "March",
                  ifelse(all_rides$month == 4, "April",
                  ifelse(all_rides$month == 5, "May",
                  ifelse(all_rides$month == 6, "June",
                  ifelse(all_rides$month == 7, "July",
                  ifelse(all_rides$month == 8, "August",
                  ifelse(all_rides$month == 9, "September",
                  ifelse(all_rides$month == 10, "October",
                  ifelse(all_rides$month == 11, "November",
                  ifelse(all_rides$month == 12, "December", "N/A")))))))))))
all_rides$month_name <- factor(all_rides$month_name)
all_rides$month_name <- ordered(all_rides$month_name, c("January", "February", "March", "April",
                                    "May", "June", "July", "August",
                                    "September", "October", "November", "December"))
#all_rides$month

str(all_rides)
```

```
## spc_tbl_ [5,667,717 x 9] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ bike_type      : chr [1:5667717] "electric_bike" "electric_bike" "classic_bike" "classic_bike" ..
##  $ user_type      : chr [1:5667717] "casual" "casual" "member" "casual" ...
##  $ year           : num [1:5667717] 2022 2022 2022 2022 2022 ...
##  $ month          : num [1:5667717] 1 1 1 1 1 1 1 1 1 1 ...
##  $ weekday        : num [1:5667717] 4 1 2 2 4 2 7 6 1 5 ...
##  $ date           : chr [1:5667717] "1/4/2022" "1/1/2022" "1/2/2022" "1/2/2022" ...
```

```
##  $ trip_length_raw: num [1:5667717] 0.00205 0.00302 0.00302 0.01037 0.00419 ...
##  $ day            : Ord.factor w/ 7 levels "Monday"<"Tuesday"<..: 4 1 2 2 4 2 7 6 1 5 ...
##  $ month_name     : Ord.factor w/ 12 levels "January"<"February"<..: 1 1 1 1 1 1 1 1 1 1 ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   rideable_type = col_character(),
##   ..   member_casual = col_character(),
##   ..   year = col_double(),
##   ..   month = col_double(),
##   ..   weekday = col_double(),
##   ..   date = col_character(),
##   ..   trip_length_raw = col_double()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

```
# let's make some changes in table structure getting some factors
all_rides$bike_type <- factor(all_rides$bike_type)
str(all_rides$bike_type)
```

```
##  Factor w/ 3 levels "classic_bike",..: 3 3 1 1 1 1 1 1 3 1 ...
```

```
head(all_rides$bike_type)
```

```
## [1] electric_bike electric_bike classic_bike  classic_bike  classic_bike
## [6] classic_bike
## Levels: classic_bike docked_bike electric_bike
```

```
all_rides$user_type <- factor(all_rides$user_type)
str(all_rides$user_type)
```

```
##  Factor w/ 2 levels "casual","member": 1 1 2 1 2 2 2 2 2 2 ...
```

```
head(all_rides$user_type)
```

```
## [1] casual casual member casual member member
## Levels: casual member
```

STEP 3: CONDUCT DESCRIPTIVE ANALYSIS

3.1 Basic descriptive analytics

Now, let's proceed with some basic calculations.

```
# Descriptive analysis on weekday (all figures in days)

mean(all_rides$weekday) #straight average (exemplar calculations)
```

```
## [1] 4.061312
```

```
median(all_rides$weekday) #straight median (exemplar calculations)
```

```
## [1] 4
```

```
max(all_rides$weekday) #straight max (exemplar calculations)
```

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

```
min(all_rides$weekday) #straight min (exemplar calculations)
```

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

```
# this is just a matter of example, so to get the basics on how things work
```

```
# Condensed four lines above to one line using summary() on the weekday attribute
summary(all_rides$weekday) #straight summary (exemplar calculations)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   2.000   4.000   4.061   6.000   7.000
```

```
# Compare members and casual users (user_type attribute) - exemplar calculations
aggregate(all_rides$weekday ~ all_rides$user_type, FUN = mean)
```

```
##   all_rides$user_type all_rides$weekday
## 1              casual          4.350261
## 2              member          3.860770
```

```
aggregate(all_rides$weekday ~ all_rides$user_type, FUN = median)
```

```
##   all_rides$user_type all_rides$weekday
## 1              casual                 5
## 2              member                 4
```

```
aggregate(all_rides$weekday ~ all_rides$user_type, FUN = max)
```

```
##   all_rides$user_type all_rides$weekday
## 1              casual                 7
## 2              member                 7
```

```
aggregate(all_rides$weekday ~ all_rides$user_type, FUN = min)
```

```
##   all_rides$user_type all_rides$weekday
## 1              casual                 1
## 2              member                 1
```

```
# See the average weekday time by each month for members vs casual users
aggregate(all_rides$weekday ~ all_rides$user_type + all_rides$month, FUN = mean)
```

```
##    all_rides$user_type all_rides$month all_rides$weekday
## 1               casual               1          4.168575
## 2               member               1          3.763930
## 3               casual               2          3.978521
## 4               member               2          3.618910
## 5               casual               3          4.096282
## 6               member               3          3.654275
## 7               casual               4          4.568974
## 8               member               4          3.891452
## 9               casual               5          4.240016
## 10              member               5          3.797587
## 11              casual               6          4.397352
## 12              member               6          3.961900
## 13              casual               7          4.551760
## 14              member               7          4.123404
## 15              casual               8          4.145206
## 16              member               8          3.703823
## 17              casual               9          4.362218
## 18              member               9          3.925240
## 19              casual              10          4.625794
## 20              member              10          4.026137
## 21              casual              11          3.955067
## 22              member              11          3.570541
## 23              casual              12          4.138972
## 24              member              12          3.771576
```

```
# this is just a matter of example, so to get the basics on how things work
```

3.2 Basic tabular analytics

Table counts of total bike type records and bike users

```
# table counts of total bike type records and bike users

table(all_rides$user_type)
```

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

```
table(all_rides$bike_type)
```

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

Categorical table calculations of bike types by users and visa versa

```
# categorical table calculations of bike types by users and visa versa

table(all_rides$user_type, all_rides$bike_type)
```

```
##
##          classic_bike docked_bike electric_bike
##   casual       891459      177474       1253099
##   member      1709755           0       1635930
```

```
table(all_rides$bike_type, all_rides$user_type)
```

```
##
##                  casual  member
##   classic_bike   891459 1709755
##   docked_bike    177474       0
##   electric_bike 1253099 1635930
```

Bike type preferences by day of week

```
# tabular exploratory analysis
# bike type preferences by day of week
table(all_rides$bike_type, all_rides$day)
```

```
##
##                 Monday Tuesday Wednesday Thursday Friday Saturday Sunday
##   classic_bike  351659  364491    364664   381292 354935   424502 359671
##   docked_bike    22535   17756     17335    19774  23387    40958  35729
##   electric_bike 376820  400125    416224   440525 423465   451011 380859
```

User type preferences by day of week

```
# user type preferences by day of week
table(all_rides$user_type, all_rides$day)
```

```
##
##          Monday Tuesday Wednesday Thursday Friday Saturday Sunday
##   casual 277675  263746    274354   309330 334701   473190 389036
##   member 473339  518626    523869   532261 467086   443281 387223
```

Bike type preferences vs users by the day of week

```
# bike type preferences vs users by the day of week

table(all_rides$bike_type, all_rides$user_type, all_rides$day)
```

```
## , ,  = Monday
##
##
##                 casual member
##   classic_bike  104257 247402
```

```
##   docked_bike    22535      0
##   electric_bike 150883 225937
##
## , ,  = Tuesday
##
##
##                 casual member
##   classic_bike   96125 268366
##   docked_bike    17756      0
##   electric_bike 149865 250260
##
## , ,  = Wednesday
##
##
##                 casual member
##   classic_bike   98363 266301
##   docked_bike    17335      0
##   electric_bike 158656 257568
##
## , ,  = Thursday
##
##
##                 casual member
##   classic_bike  113837 267455
##   docked_bike    19774      0
##   electric_bike 175719 264806
##
## , ,  = Friday
##
##
##                 casual member
##   classic_bike  123126 231809
##   docked_bike    23387      0
##   electric_bike 188188 235277
##
## , ,  = Saturday
##
##
##                 casual member
##   classic_bike  197170 227332
##   docked_bike    40958      0
##   electric_bike 235062 215949
##
## , ,  = Sunday
##
##
##                 casual member
##   classic_bike  158581 201090
##   docked_bike    35729      0
##   electric_bike 194726 186133
```

User's bike type preferences by the day of week

```
# user's bike type preferences by the day of week

table(all_rides$user_type, all_rides$bike_type, all_rides$day)
```

```
## , ,  = Monday
##
##
##          classic_bike docked_bike electric_bike
##   casual       104257       22535        150883
##   member       247402           0        225937
##
## , ,  = Tuesday
##
##
##          classic_bike docked_bike electric_bike
##   casual        96125       17756        149865
##   member       268366           0        250260
##
## , ,  = Wednesday
##
##
##          classic_bike docked_bike electric_bike
##   casual        98363       17335        158656
##   member       266301           0        257568
##
## , ,  = Thursday
##
##
##          classic_bike docked_bike electric_bike
##   casual       113837       19774        175719
##   member       267455           0        264806
##
## , ,  = Friday
##
##
##          classic_bike docked_bike electric_bike
##   casual       123126       23387        188188
##   member       231809           0        235277
##
## , ,  = Saturday
##
##
##          classic_bike docked_bike electric_bike
##   casual       197170       40958        235062
##   member       227332           0        215949
##
## , ,  = Sunday
##
##
##          classic_bike docked_bike electric_bike
##   casual       158581       35729        194726
##   member       201090           0        186133
```

And here are some trip duration calculations in raw format.

```
# exploratory calculations on trip duration by user type

tapply(all_rides$trip_length_raw, all_rides$user_type, sum)
```

```
##   casual   member
## 46995.82 29531.29
```

```
# exploratory calculations on trip duration by bike type

tapply(all_rides$trip_length_raw, all_rides$bike_type, sum)
```

```
##  classic_bike   docked_bike electric_bike
##      34315.64      15123.69      27087.78
```

```
# exploratory calculations on trip duration: user vs bike type

tapply(all_rides$trip_length_raw, list(all_rides$user_type, all_rides$bike_type), sum)
```

```
##        classic_bike docked_bike electric_bike
## casual     17798.80    15123.69      14073.33
## member     16516.84          NA      13014.45
```

Now, let's add some context to the general calculations...

```
# adding some context to the general calculations
addmargins(table(all_rides$user_type, all_rides$bike_type))
```

```
##
##          classic_bike docked_bike electric_bike     Sum
##   casual       891459      177474       1253099 2322032
##   member      1709755           0       1635930 3345685
##   Sum         2601214      177474       2889029 5667717
```

```
addmargins(table(all_rides$bike_type, all_rides$user_type))
```

```
##
##                  casual  member     Sum
##   classic_bike   891459 1709755 2601214
##   docked_bike    177474       0  177474
##   electric_bike 1253099 1635930 2889029
##   Sum           2322032 3345685 5667717
```

```
# adding some context to the general calculations

addmargins(table(all_rides$user_type, all_rides$bike_type))
```

```
##
##          classic_bike docked_bike electric_bike     Sum
##   casual       891459      177474       1253099 2322032
##   member      1709755           0       1635930 3345685
##   Sum         2601214      177474       2889029 5667717
```

```
addmargins(table(all_rides$bike_type, all_rides$user_type))
```

```
##
##                  casual  member     Sum
##   classic_bike   891459 1709755 2601214
##   docked_bike    177474       0  177474
##   electric_bike 1253099 1635930 2889029
##   Sum           2322032 3345685 5667717
```

```
# adding some proportions to the general calculation

prop.table(table(all_rides$user_type, all_rides$bike_type))
```

```
##
##          classic_bike docked_bike electric_bike
##   casual   0.15728714  0.03131314    0.22109414
##   member   0.30166556  0.00000000    0.28864003
```

```
prop.table(table(all_rides$bike_type, all_rides$user_type))
```

```
##
##                     casual     member
##   classic_bike  0.15728714 0.30166556
##   docked_bike   0.03131314 0.00000000
##   electric_bike 0.22109414 0.28864003
```

And let's see the proportions of different factors within our calculations

```
# adding more clarity into proportions of calculations
addmargins(prop.table(table(all_rides$bike_type, all_rides$user_type)))
```

```
##
##                     casual     member        Sum
##   classic_bike  0.15728714 0.30166556 0.45895270
##   docked_bike   0.03131314 0.00000000 0.03131314
##   electric_bike 0.22109414 0.28864003 0.50973417
##   Sum           0.40969441 0.59030559 1.00000000
```

```
addmargins(prop.table(table(all_rides$user_type, all_rides$bike_type)))
```

```
##
##          classic_bike docked_bike electric_bike        Sum
##   casual   0.15728714  0.03131314    0.22109414 0.40969441
##   member   0.30166556  0.00000000    0.28864003 0.59030559
##   Sum      0.45895270  0.03131314    0.50973417 1.00000000
```

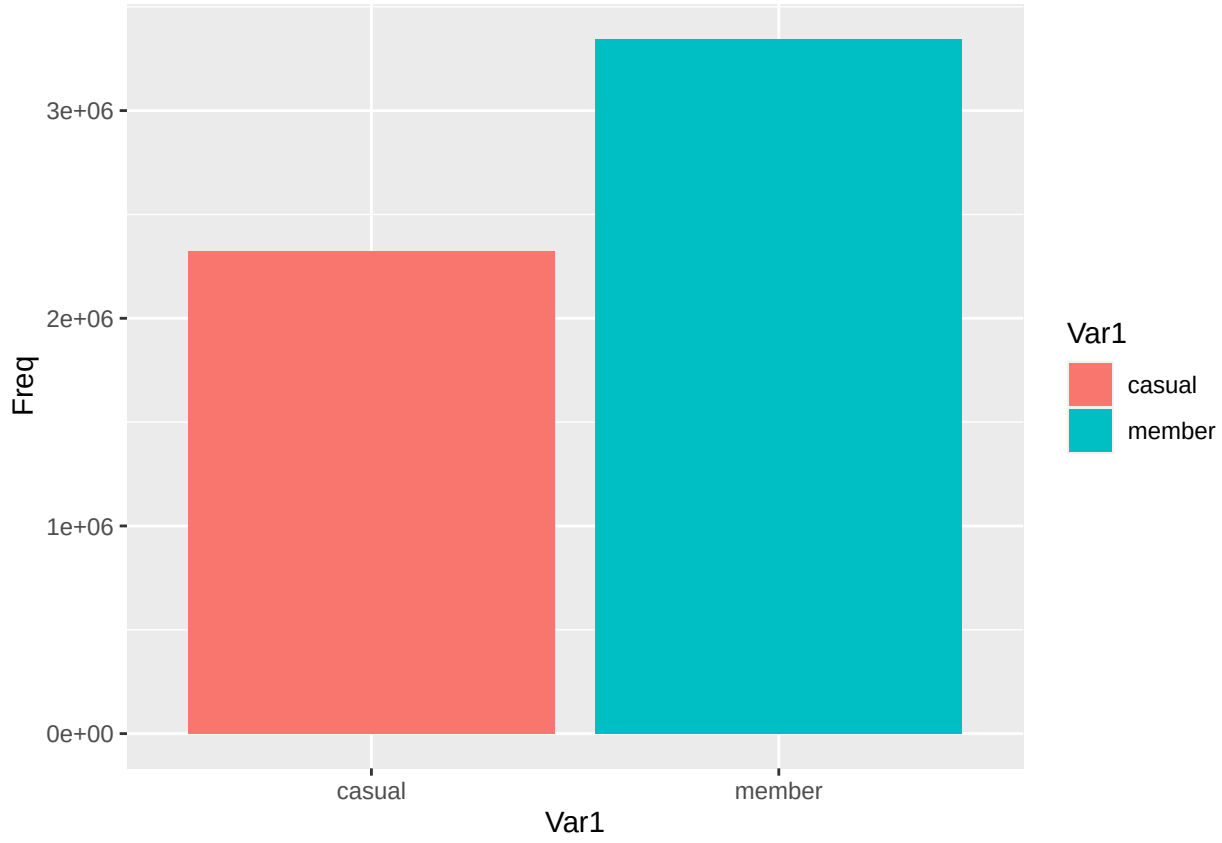
3.3 Basic Visualizations of table calculations

Wrapping it all with some visual presentation of results of calculations.

```r
# quantity of bike users
td <- table(all_rides$user_type)
my_data <- as.data.frame(td)                    # Convert table to data.frame
my_data                                             # Print data frame
```

```
##     Var1    Freq
## 1 casual 2322032
## 2 member 3345685
```

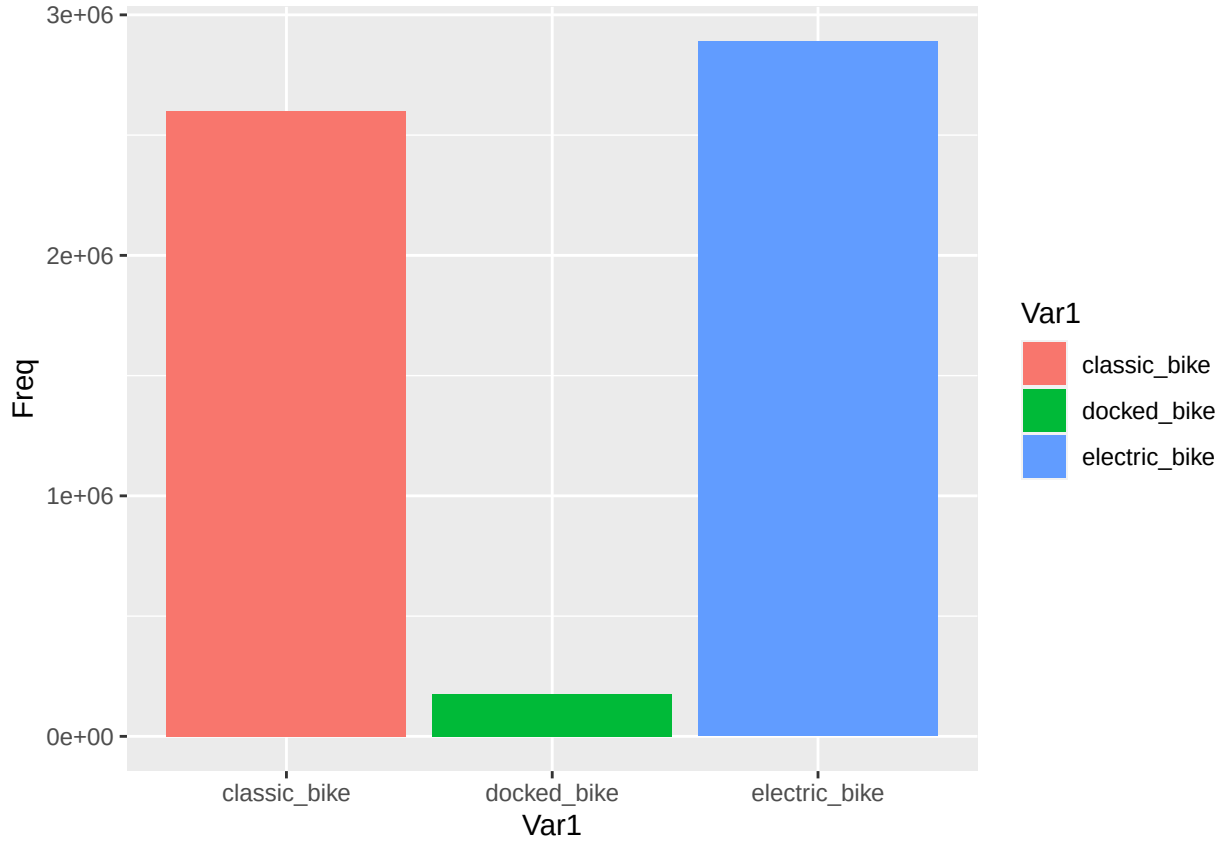
```r
ggplot(my_data,                                     # Draw bar chart of table
       aes(x = Var1,
           y = Freq, fill = Var1)) +
  geom_bar(stat = "identity")
```




```r
# quantity of bike types
tdd <- table(all_rides$bike_type)
my_data <- as.data.frame(tdd)                    # Convert table to data.frame
my_data                                             # Print data frame
```

```
##            Var1    Freq
## 1  classic_bike 2601214
## 2   docked_bike  177474
## 3 electric_bike 2889029
```

```
ggplot(my_data,                                    # Draw bar chart of table
       aes(x = Var1,
           y = Freq, fill = Var1)) +
  geom_bar(stat = "identity")
```




```
# daily spread of rides
table(all_rides$day, all_rides$user_type)
```

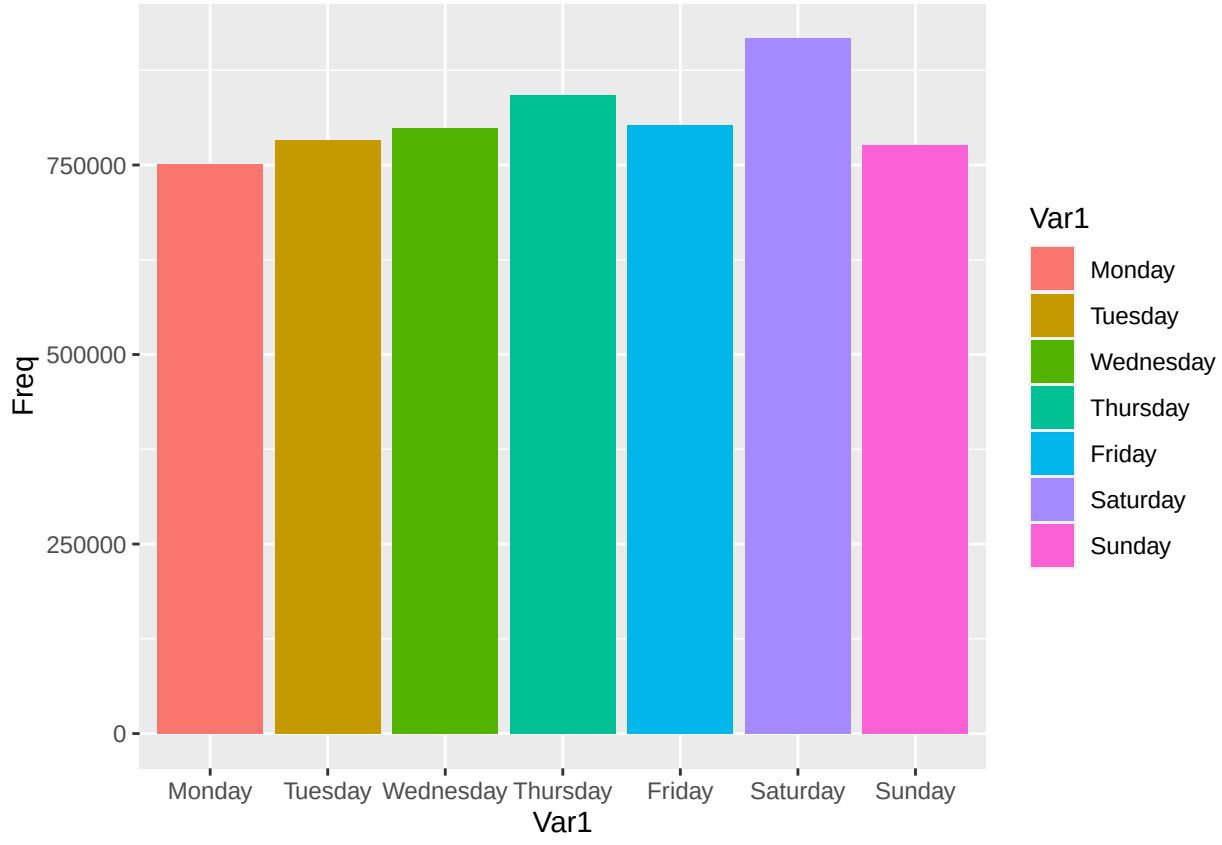
```
##
##             casual member
##   Monday    277675 473339
##   Tuesday   263746 518626
##   Wednesday 274354 523869
##   Thursday  309330 532261
##   Friday    334701 467086
##   Saturday  473190 443281
##   Sunday    389036 387223
```

```
tdd <- table(all_rides$day, all_rides$user_type)
my_data <- as.data.frame(tdd)                  # Convert table to data.frame
my_data                                           # Print data frame
```

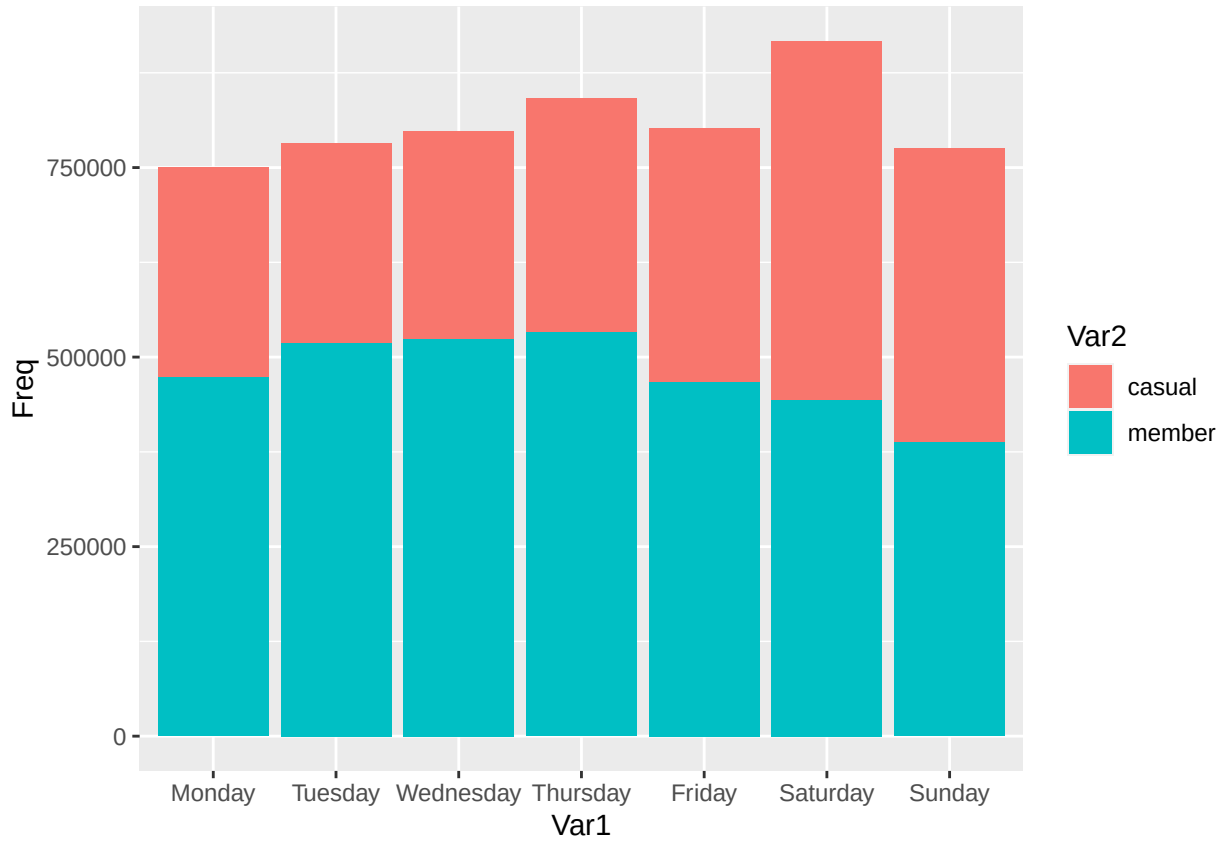
```
##         Var1   Var2   Freq
## 1     Monday casual 277675
```

```
## 2    Tuesday casual 263746
## 3  Wednesday casual 274354
## 4   Thursday casual 309330
## 5     Friday casual 334701
## 6   Saturday casual 473190
## 7     Sunday casual 389036
## 8     Monday member 473339
## 9    Tuesday member 518626
## 10 Wednesday member 523869
## 11  Thursday member 532261
## 12    Friday member 467086
## 13  Saturday member 443281
## 14    Sunday member 387223
```

```
ggplot(my_data,                                    # Draw bar chart of table
       aes(x = Var1,
           y = Freq, fill = Var1)) +
  geom_bar(stat = "identity")
```

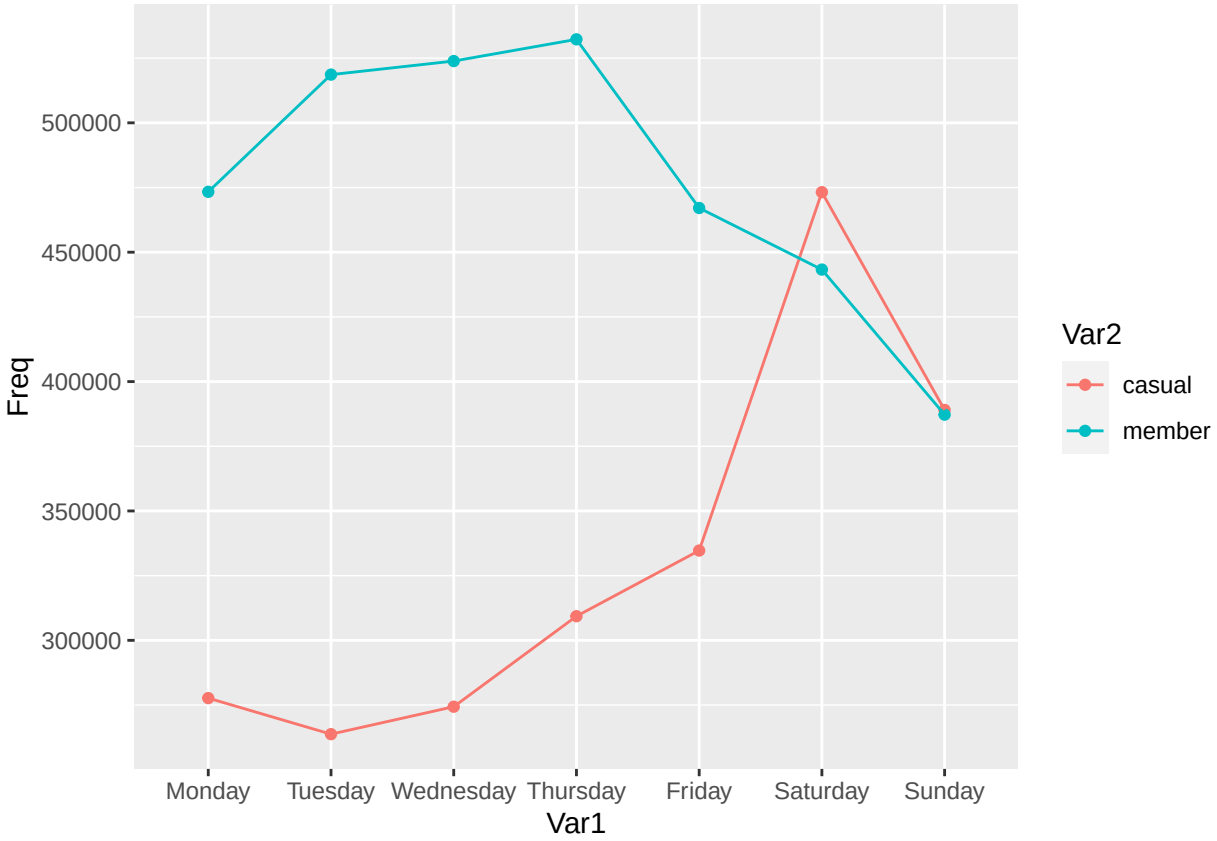



```
ggplot(my_data,                                    # Draw barchart of table
       aes(x = Var1,
           y = Freq, fill = Var2)) +
  geom_bar(stat = "identity")
```




```
# draw a comparison of daily usage proportions

ggplot(my_data, aes(Var1, Freq, group = Var2, colour = Var2)) +
  geom_point() +
  geom_line()
```



STEP 4: EXPORT SUMMARY FILE FOR FURTHER ANALYSIS

```
# Create a csv file that could be further visualized in Excel, Tableau, or other
# presentation software of choice.
# N.B.: This file location is for a  PC, change the file location accordingly:)

# write.csv(all_rides,
#           "C:/Users/YOUR-USER-NAME/Downloads/Divvy Project/all_rides.csv",
#            row.names=TRUE)

# Excavation is done! Congratulations!:)

# P.S.: I've got all_rides.csv of 300+ MB:)
```

Summary¶

First, I did my analysis in different way (Excel, BigQuery, Looker), but later on decided to try R. And this workbook describes the second path. Some examples are not that comprehensive, but this is my first experience with R - so don't judge too strong.

I had here also an attempt of visualizing things, but some un-copable message was thrown, so I skipped trying for now. But I will in the future, I believe. . .)