Google Analytics Capstone Project - Case Study 1

Kalaichitra Rangasamy

10-19-2023

Introduction

The case study is about "How does a bike-share navigate speedy success?".

For the completion of the project, I will follow all the stages of the data analysis process: Ask, prepare, process, analyze, share, and act.

Business Requirement

- Understand how annual members and casual riders differ and how digital media could influence them.
- Design a new marketing strategy to convert casual riders into annual members.

Ask

Case study Roadmap - Ask

Guiding questions:

- What is the problem you are trying to solve?

  ```
  The main objective is to find a strategy to turn casual bike riders to annual members.
  ```

- How can your insights drive business decisions?

  ```
  The insights will help the business to increase the annual members.
  ```

Deliverable:

A clear statement of the business task

```
Understand how annual members and casual riders differ and how digital media could influence them.
```

Prepare

Case study Roadmap - Prepare

Guiding questions:

- Where is your data located?

  ```
  The data is located in AWS cloud. Downloaded the necessary data and uploaded in posit Cloud for ana
  ```

- How is the data organized?

  ```
  The data is separated as month, quarter. Each on a csv file.
  ```

- Are there issues with bias or credibility in this data? Does your data ROCCC?

    There are no issues with bias in the data and yes, the data is reliable, original, comprehensive, c

- How are you addressing licensing, privacy, security, and accessibility?

      The datasets have a different name because Cyclistic is a fictional company. For the purposes of th

- How did you verify the data's integrity?

      The column names are the same in each csv and it has the correct type of data.

- How does it help you answer your question?

      It has data about the station station, end station and the time taken from start station to end sta

- Are there any problems with the data?

      Some information like city would have helped in understanding the data better.

Deliverable:

A description of all data sources used: I have taken Divvy_Trips_2019_Q4 data for the analysis purpose.

Process

Here, I am going to use R Studio to process the data. R Studio can help us process large datasets and it will be really helpful to view different data visualizations.

Uploaded the csv file in posit. Installed the dependency package tidyverse and load the library.

```
install.packages("tidyverse")
```

```
## Installing package into '/home/kalai/R/x86_64-pc-linux-gnu-library/4.3'
## (as 'lib' is unspecified)
```

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.3     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.4     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.0
## v purrr     1.0.2
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become error
```

Next, load the CSV file in a dataframe.

```
rawdata =  read_csv("~/R/Data/Divvy_Trips_2019_Q4.csv")
```

```
## Rows: 704054 Columns: 12
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (4): from_station_name, to_station_name, usertype, gender
## dbl  (5): trip_id, bikeid, from_station_id, to_station_id, birthyear
## num  (1): tripduration
## dttm (2): start_time, end_time
##
## 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.
```

```
colnames(rawdata)
```

```
##  [1] "trip_id"           "start_time"        "end_time"
##  [4] "bikeid"            "tripduration"      "from_station_id"
##  [7] "from_station_name" "to_station_id"     "to_station_name"
## [10] "usertype"          "gender"            "birthyear"
```

```
summary(rawdata)
```

```
##     trip_id           start_time
##  Min.   :25223640   Min.   :2019-10-01 00:01:39.00
##  1st Qu.:25407380   1st Qu.:2019-10-13 17:38:29.25
##  Median :25590864   Median :2019-10-28 18:04:41.00
##  Mean   :25592222   Mean   :2019-11-05 00:53:59.77
##  3rd Qu.:25777172   3rd Qu.:2019-11-25 16:19:34.00
##  Max.   :25962904   Max.   :2019-12-31 23:57:17.00
##
##     end_time                          bikeid      tripduration
##  Min.   :2019-10-01 00:06:34.00   Min.   :   1   Min.   :     61
##  1st Qu.:2019-10-13 17:59:43.25   1st Qu.:1724   1st Qu.:    354
##  Median :2019-10-28 18:18:41.00   Median :3473   Median :    585
##  Mean   :2019-11-05 01:13:49.22   Mean   :3396   Mean   :   1189
##  3rd Qu.:2019-11-25 16:34:22.75   3rd Qu.:5065   3rd Qu.:   1000
##  Max.   :2020-01-21 13:54:35.00   Max.   :6946   Max.   :8585902
##
##  from_station_id from_station_name  to_station_id   to_station_name
##  Min.   :  2     Length:704054      Min.   :  2.0   Length:704054
##  1st Qu.: 77     Class :character   1st Qu.: 77.0   Class :character
##  Median :174     Mode  :character   Median :174.0   Mode  :character
##  Mean   :204                        Mean   :203.9
##  3rd Qu.:291                        3rd Qu.:291.0
##  Max.   :673                        Max.   :673.0
##
##    usertype            gender            birthyear
##  Length:704054      Length:704054      Min.   :1899
##  Class :character   Class :character   1st Qu.:1978
##  Mode  :character   Mode  :character   Median :1987
##                                        Mean   :1984
##                                        3rd Qu.:1992
##                                        Max.   :2003
##                                        NA's   :61681
```

Check for duplicate rows in the dataframe

```
rawdata_no_dups <- rawdata[!duplicated(rawdata$trip_id), ]

print(nrow(rawdata_no_dups))
```

```
## [1] 704054
```

```
print(paste("Removed", nrow(rawdata) - nrow(rawdata_no_dups), "duplicated rows"))
```

```
## [1] "Removed 0 duplicated rows"
```

There were no duplicates found in the dataset.

Next step is to add a column “ride time in minutes”

```
rawdata <- rawdata %>%
  mutate(ride_time_m = as.numeric(rawdata$end_time - rawdata$start_time) / 60)
min(rawdata$ride_time_m)
```

```
## [1] -0.9394444
```

After executing min(rawdata$ride_time_m), found that the dataset contains some negative values. Excluded those rows for further analysis.

```
rawdata_clean <- rawdata %>%
  filter(ride_time_m >= 0.00)
```

For further analysis, rawdata_clean dataset is used.

Added the columns(year_month,weekday,start_hour) for performing analysis.

```
rawdata_clean <- rawdata_clean %>%
  mutate(year_month = paste(strftime(rawdata_clean$start_time, "%Y"),
                            "-",
                            strftime(rawdata_clean$start_time, "%m"),
                            paste("(",strftime(rawdata_clean$start_time, "%b"), ")", sep="")))
unique(rawdata_clean$year_month)
```

```
## [1] "2019 - 09 (Sep)" "2019 - 10 (Oct)" "2019 - 11 (Nov)" "2019 - 12 (Dec)"
```

```
rawdata_clean <- rawdata_clean %>%
  mutate(weekday = paste(strftime(rawdata_clean$end_time, "%u"), "-", strftime(rawdata_clean$end_time,

unique(rawdata_clean$weekday)
```

```
## [1] "1 - Mon" "2 - Tue" "4 - Thu" "3 - Wed" "7 - Sun" "5 - Fri" "6 - Sat"
```

```
rawdata_clean <- rawdata_clean %>%
  mutate(start_hour = strftime(rawdata_clean$end_time, "%H"))

unique(rawdata_clean$start_hour)
```

```
##  [1] "19" "21" "20" "03" "04" "23" "22" "00" "02" "01" "12" "10" "06" "11" "05"
## [16] "08" "18" "07" "09" "15" "16" "13" "14" "17"
```

Case study Roadmap - Process

Guiding questions:

- What tools are you choosing and why?

  ```
  I chose to use R for my quarterly analysis and excel for monthly analysis. R is used for large data
  ```

- Have you ensured your data’s integrity?

  ```
  The data remains consistent for each column.
  ```

- What steps have you taken to ensure that your data is clean?

  ```
  Checked for the duplicates based on rider_id column. No duplicates found.
  ```

- How can you verify that your data is clean and ready to analyze?

  ```
  The dataset does not have any null values and it doesn't have any duplicates so the data is clean a
  ```

- Have you documented your cleaning process so you can review and share those results?

```
Yes, I have documented the cleaning process here.
```

Analyze

Analyze part will have a summary of customer and subscribers difference and when the bike is being used the most.

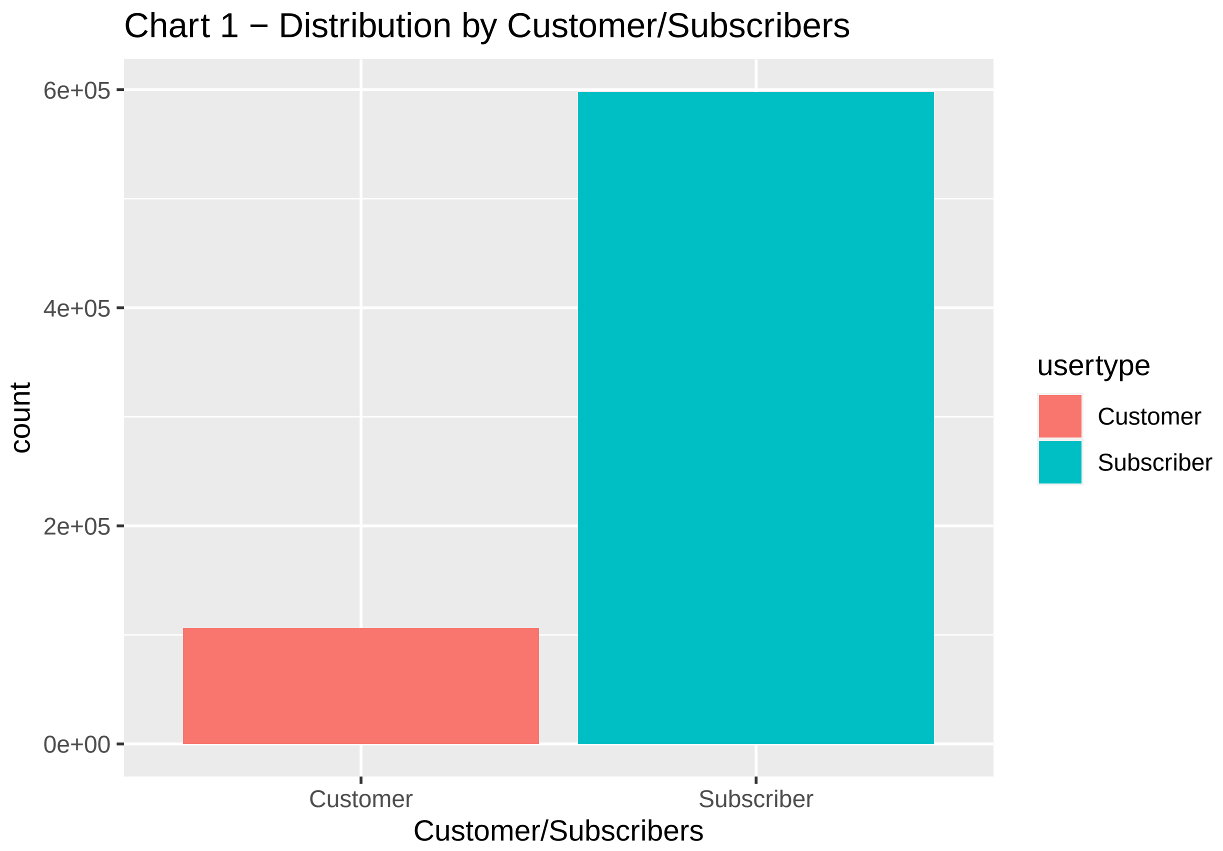
```
tripDataQ42019 <- rawdata_clean
cyclistic_summary <- tripDataQ42019 %>%
  group_by(usertype) %>%
  summarise(count = length(trip_id))

print(cyclistic_summary)
```

Summary and plot of Customer/Subscribers distribution

```
## # A tibble: 2 x 2
##   usertype    count
##   <chr>       <int>
## 1 Customer   106188
## 2 Subscriber 597853
```

```
ggplot(data = tripDataQ42019,mapping = aes(x= usertype,fill = usertype)) + geom_bar() +
  labs(x="Customer/Subscribers", title="Chart 1 - Distribution by Customer/Subscribers")
```



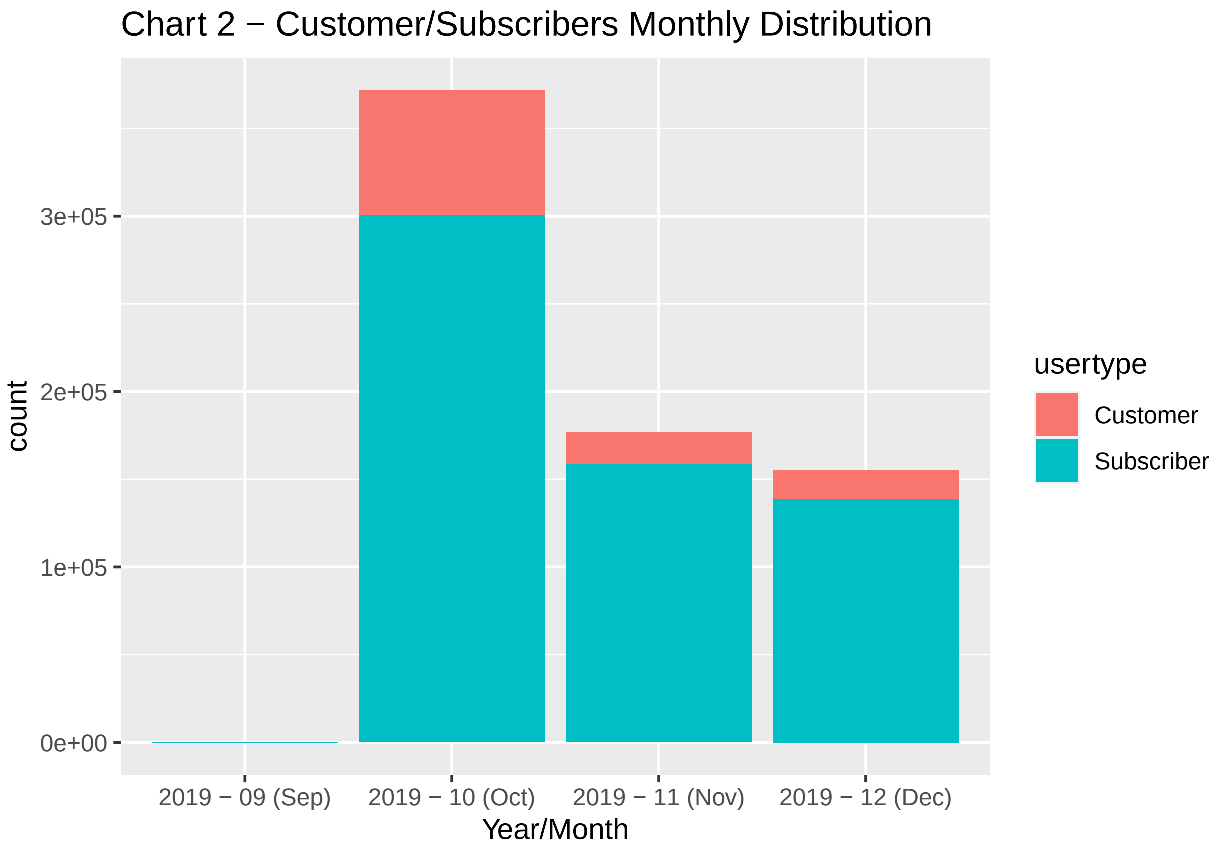
```
cyclistic_month = tripDataQ42019 %>%
  group_by(year_month) %>%
  summarise(count = length(trip_id))

print(cyclistic_month)
```

Monthly distribution - summary

```
## # A tibble: 4 x 2
##   year_month       count
##   <chr>            <int>
## 1 2019 - 09 (Sep)    196
## 2 2019 - 10 (Oct) 371672
## 3 2019 - 11 (Nov) 177173
## 4 2019 - 12 (Dec) 155000
```

```
ggplot(tripDataQ42019, aes(year_month, fill=usertype)) +
  geom_bar() +
  labs(x="Year/Month", title="Chart 2 - Customer/Subscribers Monthly Distribution")
```



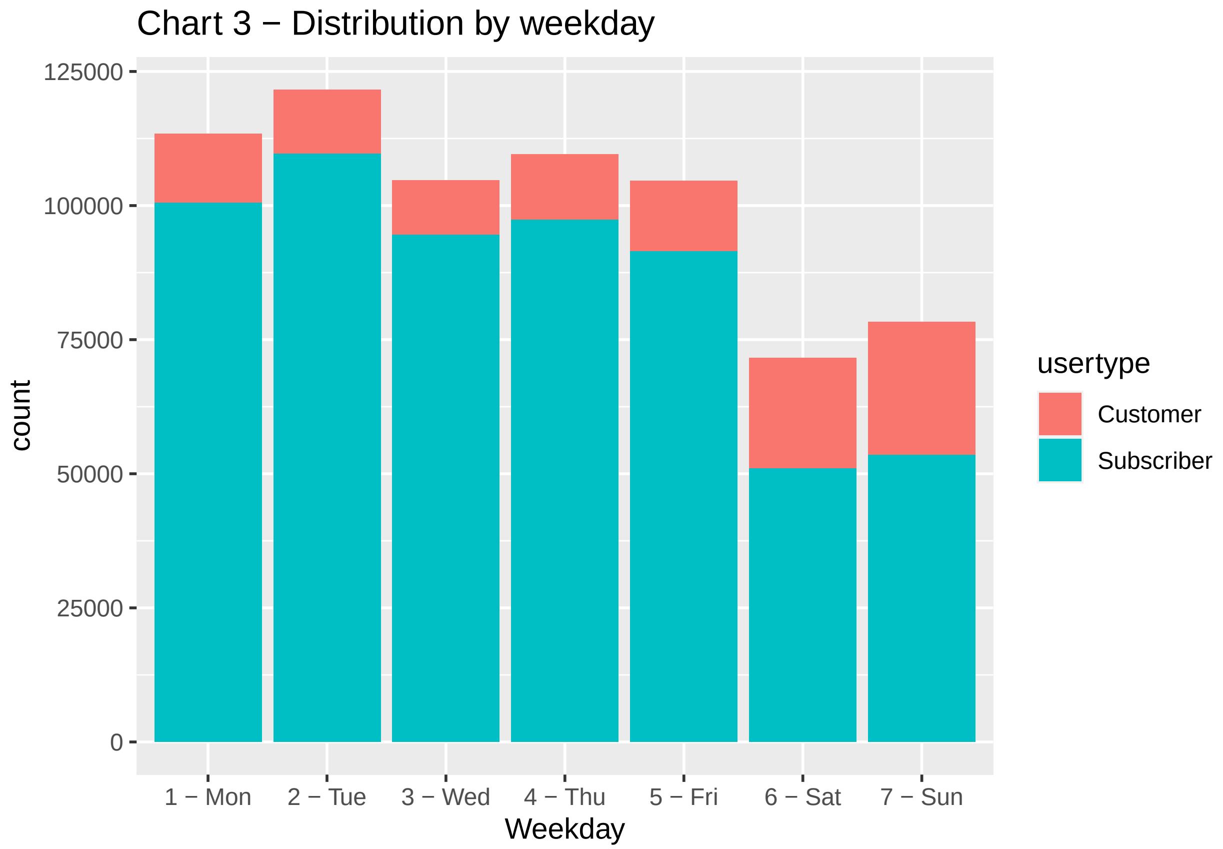
```
cyclistic_weekday = tripDataQ42019 %>%
  group_by(weekday) %>%
  summarise(count = length(trip_id))

print(cyclistic_weekday)
```

Weekday Summary

```
## # A tibble: 7 x 2
##   weekday  count
##   <chr>    <int>
## 1 1 - Mon 113436
## 2 2 - Tue 121618
## 3 3 - Wed 104717
## 4 4 - Thu 109610
## 5 5 - Fri 104653
## 6 6 - Sat  71623
## 7 7 - Sun  78384
```

```
ggplot(tripDataQ42019, aes(weekday, fill=usertype)) +
  geom_bar() +
  labs(x="Weekday", title="Chart 3 - Distribution by weekday")
```



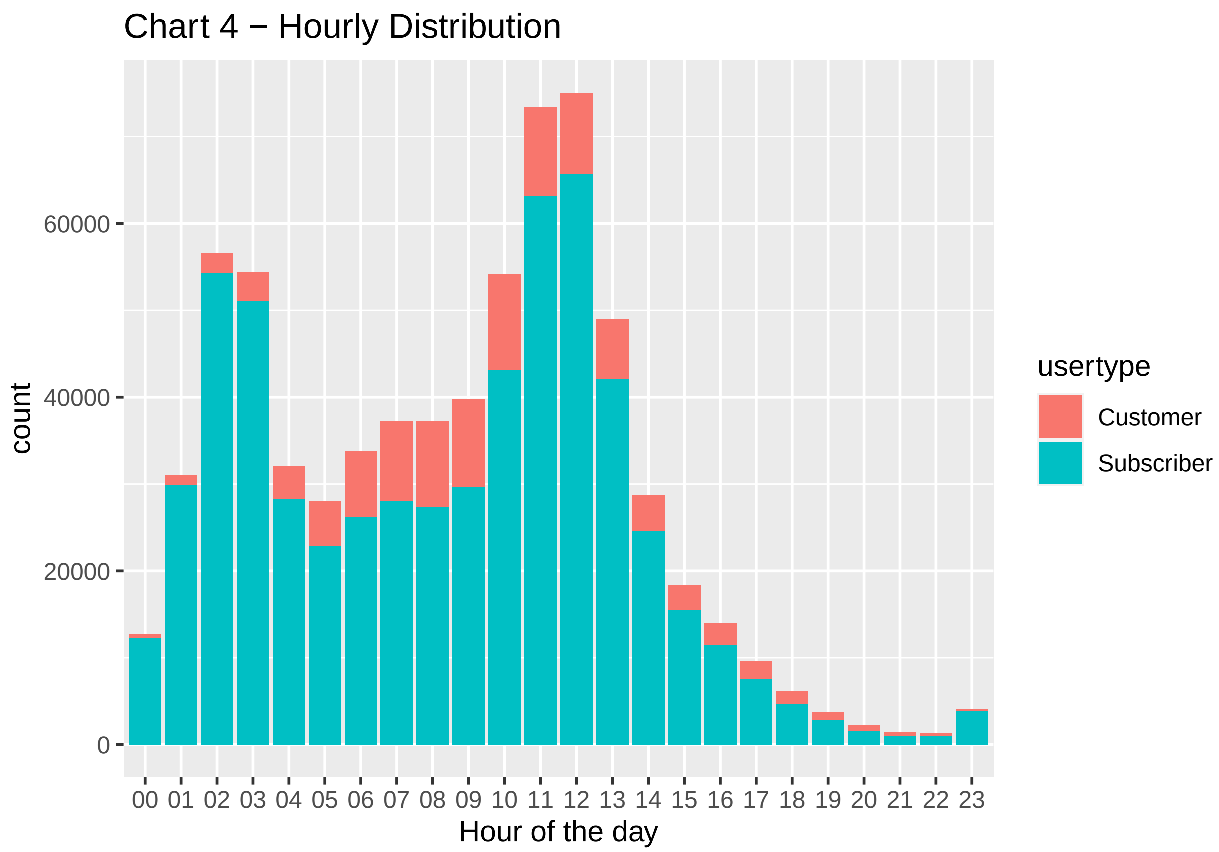
```
cyclistic_hour <- tripDataQ42019 %>%
  group_by(start_hour) %>%
  summarise(count = length(trip_id))

print(cyclistic_hour)
```

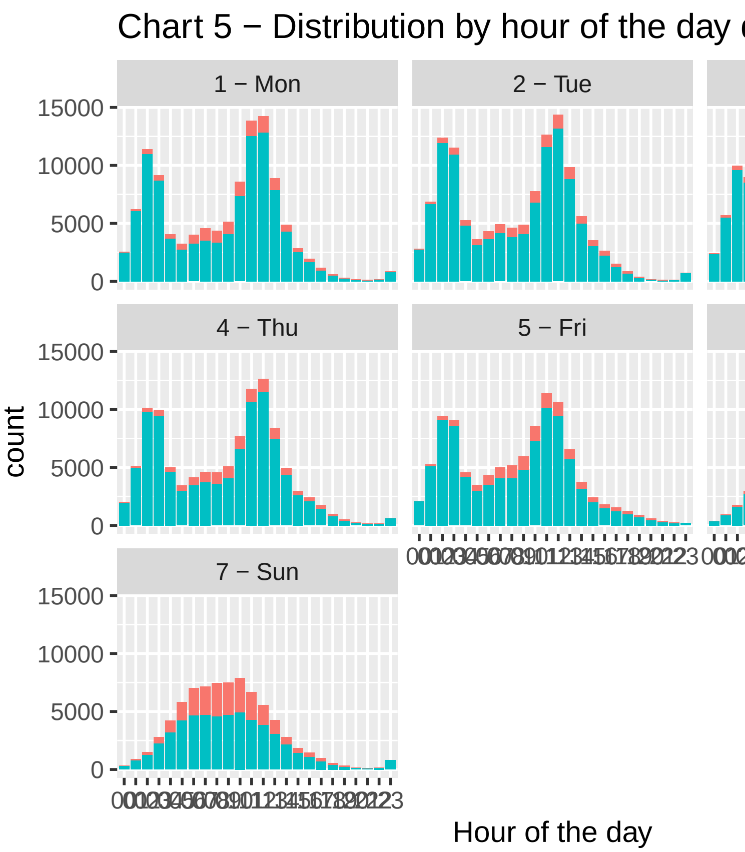
Hour of the day summary

```
## # A tibble: 24 x 2
##    start_hour count
##    <chr>      <int>
##  1 00         12656
##  2 01         30968
##  3 02         56628
##  4 03         54460
##  5 04         32032
##  6 05         28068
##  7 06         33792
##  8 07         37205
##  9 08         37251
## 10 09         39768
## # i 14 more rows
```

```r
ggplot(tripDataQ42019, aes(start_hour, fill=usertype)) +
  geom_bar() +
  labs(x="Hour of the day", title="Chart 4 - Hourly Distribution")
```



```r
tripDataQ42019 %>%
  ggplot(aes(start_hour, fill=usertype)) +
  geom_bar() +
  labs(x="Hour of the day", title="Chart 5 - Distribution by hour of the day divided by weekday") +
  facet_wrap(~ weekday)
```



Distribution by hour of the day divided by weekday

Case study Roadmap - Analyze

Guiding questions:

- How should you organize your data to perform analysis on it?

  ```
   I have taken one quarter of a year data(Q4 2019) so the data is ready for analysis.
  ```

- Has your data been properly formatted?

  ```
  Yes, all the columns in the dataset has the correct data.
  ```

- What surprises did you discover in the data?

  ```
  There are were some records where the start duration is greater than the end duration so I have ign
  ```

- What trends or relationships did you find in the data?
  - There are more subscribers than the customers in the dataset.
  - There are more users in the weekdays than weekends.
  - In weekdays, the usage is more in the mornings and evenings whereas in weekends, the usage is similar throughout the day.
  - The subscribers use the bike more than the customers.

- How will these insights help answer your business questions?

  ```
  These insights will help me build the answers for the business questions.
  ```

Share

This R notebook has the necessary summary and graphs for the presentation.

Here are the keypoints:

- The bikes are mostly used by subscribers than customers.
- It is being used mostly on the weekdays mornings and evening.
- In weekends, the usage is consistent throughout the day.

The bikes are mainly used for commute to work or for the physical activities in the weekdays. On weekends, it is mainly for physical activities or recreational activities.

Weather could also be a reason for less usage in the early mornings and late evenings.

Case Study Roadmap - Share

Guiding questions:

- Were you able to answer the question of how annual members and casual riders use Cyclistic bikes differently?

  ```
  Yes, I was able to answer the question of how annual members and casual riders with the data point
  ```

- What story does your data tell?

  ```
  The story that data tells is that the subscribers use bikes more than the customers and the bikes
  ```

- How do your findings relate to your original question?

  ```
  The findings helps me to understand better about how subscribers differ from customers.
  ```

- Who is your audience? What is the best way to communicate with them?

  ```
  The audiences are Lily Moreno- The director of marketing and the manager. And the Cyclistic execut
  ```

- Can data visualization help you share your findings?

  ```
  The data visualizations helps to share the findings in clear and understandable way.
  ```

- Is your presentation accessible to your audience?

  ```
  Yes, the presentation is accessible by the audience.
  ```

Act

The top three recommendations based on the analysis are

- Create ads to show how bikes helps avoid the traffic and it is a physical activity.
- Increase the benefits for subscribers than customers.
- Provide additional benefits for the customers/subscribers who uses bike on weekends so the usage increases during weekends.

Case Study Roadmap - Act

Guiding questions:

- What is your final conclusion based on your analysis?

  ```
  The final conclusion is that the bike is less used during winters and on weekends. Subscribers use
  ```

- How could your team and business apply your insights?

  ```
  The team and business can apply the insights by providing benefits for people who use bike on week
  ```

- What next steps would you or your stakeholders take based on your findings?

  ```
  The next steps would be to come up with ideas to increase the benefits.
  ```

- Is there additional data you could use to expand on your findings?

  Additional data like city, weather will help in understanding the data better.