Google Data Analytics - Capstone Project (Cylicstic Bike-Share Analysis)

Amiya Sur

14/08/2021

Case Study: How Does a Bike-Share Company Navigate Speedy Success?

About the Company:

In 2016, Cyclistic launched a successful bike-share offering. Since then, the program has grown to a fleet of 5,824 bicycles that are geotracked and locked into a network of 692 stations across Chicago. The bikes can be unlocked from one station and returned to any other station in the system anytime. Until now, Cyclistic's marketing strategy relied on building general awareness and appealing to broad consumer segments. One approach that helped make these things possible was the flexibility of its pricing plans: single-ride passes, full-day passes, and annual memberships. Customers who purchase single-ride or full-day passes are referred to as casual riders. Customers who purchase annual memberships are Cyclistic members. Cyclistic's finance analysts have concluded that annual members are much more profitable than casual riders. Although the pricing flexibility helps Cyclistic attract more customers, Moreno believes that maximizing the number of annual members will be key to future growth. Rather than creating a marketing campaign that targets all-new customers, Moreno believes there is a very good chance to convert casual riders into members. She notes that casual riders are already aware of the Cyclistic program and have chosen Cyclistic for their mobility needs.

Moreno has set a clear goal: Design marketing strategies aimed at converting casual riders into annual members. In order to do that, however, the marketing analyst team needs to better understand how annual members and casual riders differ, why casual riders would buy a membership, and how digital media could affect their marketing tactics. Moreno and her team are interested in analyzing the Cyclistic historical bike trip data to identify trends.

Scenario :

You are a junior data analyst working in the marketing analyst team at Cyclistic, a bike-share company in Chicago. The director of marketing believes the company's future success depends on maximizing the number of annual memberships. Therefore, your team wants to understand how casual riders and annual members use Cyclistic bikes differently. From these insights, your team will design a new marketing strategy to convert casual riders into annual members.

But first, Cyclistic executives must approve your recommendations, so they must be backed up with compelling data insights and professional data visualizations.

Introduction :

This Version of the Google Data Analytics Capstone Project - Cylicstic Bike-Share Analysis has been developed by Amiya Sur. Please refer to the Google Data Analytics Capstone: Complete a Case Study for the complete documentaion of the said problem.

The following steps will be considered for the successful completion of the Project:

- Foremost the problem will be broken down to six Data Analysis phases
  - Ask
  - Prepare
  - Process
  - Analyze
  - Share
  - Act
- We will further cover the following steps in each of the six phases:
  - Guiding Questions
  - Key Tasks
  - Deliverable

ASK

Data Analyst work with six basic problem types namely:

1. Making Predictions
2. categorizing things
3. Spotting something unusual
4. Identifying themes
5. Discovering connections
6. Finding patterns

However to analyze and draw conclusion regarding the various types of problems listed, first we need to ask **SMART** questions in order to build the foundation for the following process of preparing the data.
SMART stands for Specific, Measurable, Action-oriented, Relevant and Time-bound Questions.

For the time being we already have pre-defined question for the above problem which are as follows:

1. *How do annual members and casual riders use Cyclistic bikes differently?*
2. *Why would casual riders buy Cyclistic annual memberships?*

3. *How can Cyclistic use digital media to influence casual riders to become members?*

Although,Lily Moreno director of marketing and my manager at the same time has asked me to focus mainly on the **Question 1 - How do annual members and casual riders use Cyclistic bikes differently?**

Guiding Question

- What is the problem you are trying to solve?
  The main objective of this problem is to identity the differences between causal and Annual members usage pattern and to recommend marketing strategies for turning maximum casual riders to annual members.

- How many your insights drive business decisions?
  The insights will help the marketing team to increase annual members and maximize company revenue.

Key tasks

- ☒ Identify the business task
- ☒ Consider Key stakeholders

Deliverable

- ☒ A clear statement of the business task

Find the keys differences between casual and members riders and how digital media could influence them.

PREPARE

In this project we will be using the Cyclistic's historic trip data for last 12 months to analyze and identify trends. The data set has been provided by Google and can be downloaded from the following link.

Guiding Questions

- Where is your data located?
  The data was formerly located in server Google provided which has been later downloaded in my local machine.

- How is your data organised?
  The data is organised on the basis of month and each month has its own data. For this analysis I have downloaded last 6 months data from June 21 to July 20.

- Are there issues with bias or credibility in this data? Does your data ROCCC?
  The population of the dataset belongs to Cyclistic's own clients data, therefore bias wont be a problem. We can further conclude that the data is credible at

the same time and finally the data is ROCCC because its Reliable, Original, Comprehensive, Current and Cited.

- How are you addressing licensing, privacy, security, and accessibility?
  The data has been made available by Motivate International Inc. under this license. This is public data that can be used to explore how different customer types are using Cyclistic bikes. However data-privacy issues warned us from using riders' personally identifiable information but as a matter of fact this dataset doesn't contain any personal information about riders such as bank details, credit card numbers etc.

- How did you verify the data's integrity?
  All the data files are in .csv format , files have consistent data type and columns have correct type of data. The following information has been verified with the help of a Spreadsheet tool.

- How does it help you answer your question?
  The data has various sort of information regarding the bike rides, their duration and even the type of rides used by causal and annual members, these data can provide us with key insights on further analysis.

- Are there any problems with the data?
  As of now I didn't find any problem with the dataset, yes there are some limitations, more information about the rider type would have been useful.

Key tasks

- ☒ Download data and store it properly
- ☒ Identify how its organised
- ☒ Sort and filter the data
- ☒ Determine the credibility of data

Deliverable

- ☒ A description of the data source used
  This is historical data provided by the Cyclistic. The data ranges from June 2021 to last year July 2020 covering a span of 12 months.The data has been made available by Motivate International Inc and this is a public open source data that can be used for further **Exploratory Data Analysis**.

PROCESS

All the below steps will prepare and process the data for analysing and finally gaining insights which can be used to create compelling visuals to share insights with the stakeholders.The following approach is also known as data driven decision mmaking.

To begin the processing phase of our data, we need to load the following libraries first.

The following steps will load the .csv files into 12 separate dataframes

```
#import the 12 .csv files first
df1 <- read.csv("data/202007-divvy-tripdata.csv")
df2 <- read.csv("data/202008-divvy-tripdata.csv")
df3 <- read.csv("data/202009-divvy-tripdata.csv")
df4 <- read.csv("data/202010-divvy-tripdata.csv")

df5 <- read.csv("data/202011-divvy-tripdata.csv")
df6 <- read.csv("data/202012-divvy-tripdata.csv")
df7 <- read.csv("data/202101-divvy-tripdata.csv")
df8 <- read.csv("data/202102-divvy-tripdata.csv")
df9 <- read.csv("data/202103-divvy-tripdata.csv")
df10 <- read.csv("data/202104-divvy-tripdata.csv")
df11 <- read.csv("data/202105-divvy-tripdata.csv")
df12 <- read.csv("data/202106-divvy-tripdata.csv")
```

Considering the 12 dataframes each time will be problematic and thus we need to concatenate or combine them in one dataframe.

```
bike_rides <- rbind(df1,df2,df3,df4,df5,df6,df7,df8,df9,df10,df11,df12)
head(bike_rides)
```

```
##            ride_id rideable_type          started_at            ended_at
## 1 762198876D69004D   docked_bike 2020-07-09 15:22:02 2020-07-09 15:25:52
## 2 BEC9C9FBA0D4CF1B   docked_bike 2020-07-24 23:56:30 2020-07-25 00:20:17
## 3 D2FD8EA432C77EC1   docked_bike 2020-07-08 19:49:07 2020-07-08 19:56:22
## 4 54AE594E20B35881   docked_bike 2020-07-17 19:06:42 2020-07-17 19:27:38
## 5 54025FDC7440B56F   docked_bike 2020-07-04 10:39:57 2020-07-04 10:45:05
## 6 65636B619E24257F   docked_bike 2020-07-28 16:33:03 2020-07-28 16:49:10
##              start_station_name start_station_id          end_station_name
## 1         Ritchie Ct & Banks St              180  Wells St & Evergreen Ave
## 2        Halsted St & Roscoe St              299      Broadway & Ridge Ave
## 3 Lake Shore Dr & Diversey Pkwy              329 Clark St & Wellington Ave
## 4      LaSalle St & Illinois St              181   Clark St & Armitage Ave
## 5    Lake Shore Dr & North Blvd              268    Clark St & Schiller St
## 6    Fairbanks St & Superior St              635     Wells St & Concord Ln
##   end_station_id start_lat start_lng  end_lat   end_lng member_casual
## 1            291  41.90687 -87.62622 41.90672 -87.63483        member
## 2            461  41.94367 -87.64895 41.98404 -87.66027        member
## 3            156  41.93259 -87.63643 41.93650 -87.64754        casual
## 4             94  41.89076 -87.63170 41.91831 -87.63628        casual
## 5            301  41.91172 -87.62680 41.90799 -87.63150        member
## 6            289  41.89575 -87.62010 41.91213 -87.63466        casual
```

We have created a new dataframe named `bike_rides` combining all the data from the 12 monthly files and did a quick check to our new dataframe using the **head()** function.

Cleaning the data

In the following few steps we will be cleaning, wrangling and modifying the data for the final analysis.

- Removing the empty rows and columns

```
bike_rides <- janitor :: remove_empty(bike_rides, which = c("cols"))
bike_rides <- janitor :: remove_empty(bike_rides, which = c("rows"))
```

- Using a `str()` function, we can come to the conclusion that many of our columns are in 'chr' format which we need to covert into 'factors' for further requirement in the `ANALYSE` phase.

- converting categorical columns into factors

```
bike_rides[, c("ride_id", "rideable_type", "started_at", "ended_at",
"start_station_name", "end_station_name", "member_casual")] <-
  lapply(bike_rides[, c("ride_id", "rideable_type", "started_at", "ended_at",
"start_station_name", "end_station_name", "member_casual")], factor)
```

- Lets do a quick `str()` to check the result:

```
str(bike_rides)
```

```
## 'data.frame':    4460151 obs. of  13 variables:
##  $ ride_id           : Factor w/ 4459942 levels "000001004784CD35",..:
2055916 3323367 3676528 1473596 1461917 1764200 607390 3516483 4388319
2405548 ...
##  $ rideable_type     : Factor w/ 3 levels "classic_bike",..: 2 2 2 2 2 2 2
2 2 2 ...
##  $ started_at        : Factor w/ 3813396 levels "2020-07-01 00:00:14",..:
123022 355167 114644 246083 48091 413950 438954 189410 438314 86496 ...
##  $ ended_at          : Factor w/ 3801221 levels "2020-07-01 00:03:01",..:
122578 354581 114058 245800 47706 413288 438371 189045 437712 86018 ...
##  $ start_station_name: Factor w/ 713 levels "","2112 W Peterson Ave",..:
535 309 368 388 370 260 53 588 141 269 ...
##  $ start_station_id  : chr  "180" "299" "329" "181" ...
##  $ end_station_name  : Factor w/ 714 levels "","2112 W Peterson Ave",..:
664 64 157 135 154 662 215 270 103 436 ...
##  $ end_station_id    : chr  "291" "461" "156" "94" ...
##  $ start_lat         : num  41.9 41.9 41.9 41.9 41.9 ...
##  $ start_lng         : num  -87.6 -87.6 -87.6 -87.6 -87.6 ...
##  $ end_lat           : num  41.9 42 41.9 41.9 41.9 ...
##  $ end_lng           : num  -87.6 -87.7 -87.6 -87.6 -87.6 ...
##  $ member_casual     : Factor w/ 2 levels "casual","member": 2 2 1 1 2 1 2
2 2 2 ...
```

- Checking for rows with NA values

Our dataframe has `4460151` rows distributed across 13 variables as of now. Performing a quick summation of rows with NA values.

```
sum(is.na(bike_rides))
```

```
## [1] 191371
```

Clearly,the output tell us that there are **191371** rows with NA values that can skew our analysis.

Discussing with our team members, I came to the conclusion that there must be some issues with the data which for our analysis can be ignored and thus we will omit them as of now.

```
bike_rides <- na.omit(bike_rides)
```

NOTE: Our new dataframe consist of 4324369 rows.

- Further data cleaning for duplicate rows, if any!

```
bike_rides_no_dups <- bike_rides[!duplicated(bike_rides$ride_id), ]
print(paste("Removed", nrow(bike_rides) - nrow(bike_rides_no_dups),
"duplicated rows"))
```

```
## [1] "Removed 208 duplicated rows"
```

For our convenience, we will store the `bike_rides_no_dups` back to our convenient dataframe `bike_rides`.

```
bike_rides <- bike_rides_no_dups
```

Data manipulation and formatting

```
bike_rides$Ymd <- as.Date(bike_rides$started_at)
bike_rides$started_at <-
  lubridate :: ymd_hms(bike_rides$started_at)
bike_rides$ended_at <-
  lubridate :: ymd_hms(bike_rides$ended_at)
bike_rides$start_hour <-
  lubridate :: hour(bike_rides$started_at)
bike_rides$end_hour <-
  lubridate::hour(bike_rides$ended_at)
```

- Creating computational columns that will lay the stepping stone in the following phases before we reach any conclusion.

Extracting duration time in minutes

The duration time will be an important factor for finding the difference in usage patterns b/w member and casual riders.

```
bike_rides$duration_time <-
  as.numeric((bike_rides$ended_at - bike_rides$started_at)/60)
```

Extracting weekday

This will extract the weekday from the date column `Ymd`

```
bike_rides <- bike_rides %>%
  mutate(weekday = weekdays(Ymd))
```

Finally, We will be creating one last column that will extract the `year_month` from `Ymd` column.

```
bike_rides <- bike_rides %>%
  mutate(year_month = paste(format(as.Date(Ymd),"%Y-%m"),"(",
months(Ymd),")"))
```

Saving the result as .CSV file

Guiding Questions

- What tools are you choosing and why?
  I'm using R for this project, for two main reasons: Because handling large dataset is way more convenient in R as also to gather experience with the language.

- Have you ensured your data's integrity?
  Yes, the data is consistent throughout the columns.

- What steps have you taken to ensure that your data is clean?
  I have taken the following steps to ensure this:

  - **Empty rows and columns have been removed.**

  - **Rows with NA values has been omitted in final analysis.**

  - **We have further checked for duplicate rows as well.**

- How can you verify that your data is clean and ready to analyze?
  It can be verified from the above steps and also we have formatted few categorical columns to factors for our analyzing purpose.

- Have you documented your cleaning process so you can review and share those results?
  Yes, it's all documented in this .rmd file.

Keys tasks

- ☒ Check the data for errors.
- ☒ Choose your tools.
- ☒ Transform the data so you can work with it efectively.
- ☒ Document the cleaning process.

Deliverable

- ☒ Documentation of any cleaning or manipulation of data.

ANALYSE

I will take the EDA approach for analyzing the problem and come up with a data-driven decision keeping in mind the the question that was being asked atbthe first point. Exploratory data Analysis (EDA) is the process of analyzing and visualizing the data to get a better understanding and gain insights from it.

Question How do annual members and casual riders use Cyclistic bikes differently?.

To quick start, lets generate the summary for our dataset `bike_rides`.

```
summary(bike_rides)
```

```
##              ride_id                     rideable_type
##  000001004784CD35:      1   classic_bike :1276607
##  000002EBE159AE82:      1   docked_bike  :2046699
##  00001A81D056B01B:      1   electric_bike:1000855
##  00001DCF2BC423F4:      1
##  00001E17DEF40948:      1
##  00002279D7D315A5:      1
##  (Other)         :4324155
##    started_at                     ended_at
##  Min.   :2020-07-01 00:00:14   Min.   :2020-07-01 00:03:01
##  1st Qu.:2020-08-28 15:59:40   1st Qu.:2020-08-28 16:23:29
##  Median :2020-11-19 07:50:03   Median :2020-11-19 08:01:37
##  Mean   :2020-12-27 12:07:16   Mean   :2020-12-27 12:32:00
##  3rd Qu.:2021-05-14 08:46:33   3rd Qu.:2021-05-14 09:01:53
##  Max.   :2021-06-30 23:59:59   Max.   :2021-07-13 22:51:35
##
##                   start_station_name  start_station_id
##                             : 199111   Length:4324161
##  Streeter Dr & Grand Ave    :  57914   Class :character
##  Lake Shore Dr & Monroe St  :  43544   Mode  :character
##  Theater on the Lake        :  39583
##  Clark St & Elm St          :  38369
##  Lake Shore Dr & North Blvd:  38029
##  (Other)                    :3907611
##                     end_station_name   end_station_id       start_lat
##                             : 216323   Length:4324161     Min.   :41.64
##  Streeter Dr & Grand Ave    :  60572   Class :character   1st Qu.:41.88
##  Lake Shore Dr & Monroe St  :  42503   Mode  :character   Median :41.90
##  Theater on the Lake        :  41262                      Mean   :41.90
##  Lake Shore Dr & North Blvd:  40527                      3rd Qu.:41.93
##  Clark St & Elm St          :  38106                      Max.   :42.07
##  (Other)                    :3884868
##    start_lng         end_lat         end_lng       member_casual
##  Min.   :-87.78   Min.   :41.51   Min.   :-88.07   casual:1868140
##  1st Qu.:-87.66   1st Qu.:41.88   1st Qu.:-87.66   member:2456021
##  Median :-87.64   Median :41.90   Median :-87.64
##  Mean   :-87.64   Mean   :41.90   Mean   :-87.64
```

```
##  3rd Qu.:-87.63   3rd Qu.:41.93   3rd Qu.:-87.63
##  Max.   :-87.52   Max.   :42.15   Max.   :-87.49
##
##       Ymd               start_hour       end_hour     duration_time
##  Min.   :2020-07-01   Min.   : 0.00   Min.   : 0.00   Min.   :-29049.97
##  1st Qu.:2020-08-28   1st Qu.:11.00   1st Qu.:12.00   1st Qu.:     7.57
##  Median :2020-11-19   Median :15.00   Median :15.00   Median :    13.68
##  Mean   :2020-12-26   Mean   :14.35   Mean   :14.53   Mean   :    24.74
##  3rd Qu.:2021-05-14   3rd Qu.:18.00   3rd Qu.:18.00   3rd Qu.:    25.17
##  Max.   :2021-06-30   Max.   :23.00   Max.   :23.00   Max.   : 55944.15
##
##    weekday           year_month
##  Length:4324161     Length:4324161
##  Class :character   Class :character
##  Mode  :character   Mode  :character
##
##
##
##
```

One thing that immediately catches the attention is duration_time. This variable has negative values, and the biggest value is `55944.15`, which is 38.85 days. This field will be explored further in the document.

Data Distribution

We will answer the most basic questions and analyze how our data is distributed across various factors.

Casual rides vs Annual members

```
bike_rides %>%
  group_by(member_casual) %>%
  summarise(count = length(ride_id),
            '%' = (length(ride_id) / nrow(bike_rides)) * 100)
```

```
## # A tibble: 2 x 3
##   member_casual   count   `%`
##   <fct>           <int> <dbl>
## 1 casual        1868140  43.2
## 2 member        2456021  56.8
```

The summary tells us among the two categories of riders, Cyclistic has already 56% member which is ~13 % more than casual riders. We will conclude this with a clear visual in our next phase.

Distribution by Month

Lets look at the trends and usage of bikes across month.

```
bike_rides %>%
  group_by(year_month) %>%
```

```
  summarise(count = length(ride_id),
            '%' = (length(ride_id) / nrow(bike_rides)) * 100,
            'members_p' = (sum(member_casual == "member") / length(ride_id))
* 100,
            'casual_p' = (sum(member_casual == "casual") / length(ride_id)) *
100,
            '% diff b/w member vs causal' = members_p - casual_p)

## # A tibble: 12 x 6
##    year_month           count   `%` members_p casual_p `% diff b/w member
vs ca~
##    <chr>                <int> <dbl>     <dbl>    <dbl>
<dbl>
##  1 2020-07 ( July )    550425 12.7       51.2     48.8
2.35
##  2 2020-08 ( August )  608510 14.1       53.5     46.5
6.98
##  3 2020-09 ( Septembe~ 500390 11.6       57.0     43.0
13.9
##  4 2020-10 ( October ) 339303  7.85      63.8     36.2
27.6
##  5 2020-11 ( November~ 222789  5.15      67.2     32.8
34.4
##  6 2020-12 ( December~ 131254  3.04      77.1     22.9
54.3
##  7 2021-01 ( January )  96731  2.24      81.3     18.7
62.6
##  8 2021-02 ( February~  49408  1.14      79.6     20.4
59.2
##  9 2021-03 ( March )   228329  5.28      63.2     36.8
26.5
## 10 2021-04 ( April )   336963  7.79      59.5     40.5
19.0
## 11 2021-05 ( May )     531181 12.3       51.7     48.3
3.39
## 12 2021-06 ( June )    728878 16.9       49.2     50.8
-1.57
```

Seeing the patterns in the data, we can take some considerations at this moment.

- The month with the biggest count of data points was June with almost 16% of the dataset.

- The difference of proportion b/w member and casual riders appears to shrink between May to August each year.

We will further analyse the above scenario with visuals, as we know sometime data can confusing and in such cases visuals can make things clearer.

Distribution by Weekday

```
bike_rides %>%
  group_by(weekday) %>%
  summarise(count = length(ride_id),
            '%' = (length(ride_id) / nrow(bike_rides)) * 100,
            'members_p' = (sum(member_casual == "member") / length(ride_id))
* 100,
            'casual_p' = (sum(member_casual == "casual") / length(ride_id)) *
100,
            '% diff b/w member_casual category' = members_p - casual_p)

## # A tibble: 7 x 6
##   weekday    count   `%` members_p casual_p `% diff b/w member_casual
category`
##   <chr>      <int> <dbl>     <dbl>    <dbl>
<dbl>
## 1 Friday    634881  14.7      57.2     42.8
14.5
## 2 Monday    530434  12.3      61.8     38.2
23.6
## 3 Saturday  791608  18.3      46.1     53.9
-7.86
## 4 Sunday    670202  15.5      46.8     53.2
-6.46
## 5 Thursday  557409  12.9      63.3     36.7
26.6
## 6 Tuesday   552502  12.8      64.2     35.8
28.4
## 7 Wednesday 587125  13.6      64.5     35.5
29.1
```

Some key points to be noted:

- This summarized tables says that number of rides are much higher during weekends than weekdays (mainly on Saturday).
- Weekends have the biggest volume of casuals, starting on Friday with a rise of ~(6 - 19)%.

Hour of the Day

How data is distributed by hours of the day

```
bike_rides %>%
  group_by(start_hour) %>%
  summarise(count = length(ride_id),
            '%' = (length(ride_id) / nrow(bike_rides)) * 100,
            'members_p' = (sum(member_casual == "member") / length(ride_id))
* 100,
            'casual_p' = (sum(member_casual == "casual") / length(ride_id)) *
```

```
100,
            '% diff b/w member_casual category' = members_p - casual_p)

## # A tibble: 24 x 6
##    start_hour  count   `%` members_p casual_p `% diff b/w member_casual
categor~
##         <int>  <int> <dbl>     <dbl>    <dbl>
<dbl>
##  1          0  54587 1.26       34.8     65.2
-30.4
##  2          1  35196 0.814      32.4     67.6
-35.2
##  3          2  19915 0.461      30.5     69.5
-39.0
##  4          3  10771 0.249      33.4     66.6
-33.2
##  5          4  10279 0.238      47.2     52.8
-5.61
##  6          5  29517 0.683      75.1     24.9
50.1
##  7          6  83323 1.93       80.4     19.6
60.7
##  8          7 145598 3.37       80.0     20.0
59.9
##  9          8 168475 3.90       75.7     24.3
51.3
## 10          9 156337 3.62       66.3     33.7
32.7
## # ... with 14 more rows
```

Some Key analysis:

- There is larger volume of rides during afternoon

With the help of a bar graph we will explore this analysis in depth and try to connect the dots hidden in data to make up a logical story inspired by our data and visuals.

Rideable type

Finally, we will also see how data is distributed across the available `rideable_type`.

```
bike_rides %>%
  group_by(rideable_type) %>%
  summarise(count = length(ride_id),
            '%' = (length(ride_id) / nrow(bike_rides)) * 100,
            'members_p' = (sum(member_casual == "member") / length(ride_id))
* 100,
            'casual_p' = (sum(member_casual == "casual") / length(ride_id)) *
100,
            '% diff b/w member_casual category' = members_p - casual_p)
```

```
## # A tibble: 3 x 6
##   rideable_type   count   `%` members_p casual_p `% diff b/w member_casual
cate~
##   <fct>           <int> <dbl>     <dbl>    <dbl>
<dbl>
## 1 classic_bike  1276607  29.5      64.6     35.4
29.1
## 2 docked_bike   2046699  47.3      52.6     47.4
5.26
## 3 electric_bike 1000855  23.1      55.4     44.6
10.9
```

Key points from the above analysis:

- Docked bike has the highest demand.
- Although there seems to be a preference for classic bikes among members.

Three Key question:

- Mean of ride_duration

```
mean_duration_time <- mean(bike_rides$duration_time)
mean_duration_time
```

```
## [1] 24.73723
```

- Max of ride_duration

```
max_duration_time <- max(bike_rides$duration_time)
max_duration_time
```

```
## [1] 55944.15
```

- Frequency of weekday

```
bike_rides$weekday <- as.factor(bike_rides$weekday)
table(bike_rides$weekday)
```

```
##
##    Friday    Monday  Saturday    Sunday  Thursday   Tuesday Wednesday
##    634881    530434    791608    670202    557409    552502    587125
```

As, concluded previously weekends shows huge rise in the number of rides.

Considering outliers in data

Lets, summarize the column `duration_time`

```
summary(bike_rides$duration_time)
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
## -29049.97      7.57     13.68     24.74     25.17  55944.15
```

The summary() function on the column `duration_time` showed us some interesting statistics. Definitely the negative min() value is due to bad data, but the max() value is also

surprisingly skimmed. A max value of `55944.15` mins means 38 days. Probably, there are some stations that returns bad data.

To solve this problem, which will cause error while plotting some values we will take a sample data from our whole population, that represent the population.

```
tiles = quantile(bike_rides$duration_time, seq(0, 1, by=0.05))
tiles
```

```
##            0%            5%           10%           15%           20%
## -29049.966667      3.033333      4.383333      5.500000      6.533333
##           25%           30%           35%           40%           45%
##      7.566667      8.633333      9.750000     10.933333     12.233333
##           50%           55%           60%           65%           70%
##     13.683333     15.300000     17.183333     19.366667     21.983333
##           75%           80%           85%           90%           95%
##     25.166667     29.083333     34.650000     43.883333     68.400000
##          100%
##  55944.150000
```

We can see, that the difference between 5% and 95% is 65.37 and so we will take a subset from our `bike_rides` dataframe comprising only the 95% data.

```
bike_rides_without_outliers <- bike_rides %>%
    filter(duration_time > as.numeric(tiles['5%'])) %>%
    filter(duration_time < as.numeric(tiles['95%']))

print(paste("Removed", nrow(bike_rides) - nrow(bike_rides_without_outliers),
"rows as outliers" ))
```

```
## [1] "Removed 434312 rows as outliers"
```

Riding_time vs member_type

```
bike_rides_without_outliers %>%
    group_by(member_casual) %>%
    summarise(mean = mean(duration_time),
              'first_quarter' = as.numeric(quantile(duration_time, .25)),
              'median' = median(duration_time),
              'third_quarter' = as.numeric(quantile(duration_time, .75)),
              'IR' = third_quarter - first_quarter)
```

```
## # A tibble: 2 x 6
##   member_casual  mean first_quarter median third_quarter    IR
##   <fct>         <dbl>         <dbl>  <dbl>         <dbl> <dbl>
## 1 casual         22.0         10.5    17.7          29.4  18.8
## 2 member         14.7          6.98   11.5          19.3  12.3
```

Key Point to note from the above analysis:

- Average riding time for casual riders is greater than members.

Finally, We will conclude our Analysis phase. We definitely gained some key insights summarizing and analyzing the data which will be further plotted into visuals to identity patters or trends, spot differences and even to unfold any hidden stories.

Guiding Questions

- How should you organize your data to perform analysis on it?
  The data has been organized into a single CSV combining all the files from the dataset.

- Has your data been properly formatted?
  Yes, all the columns have their correct data type.

- What surprises did you discover in the data?
  One of the main surprises is how members differ from casuals when analysed by weekdays. Also that members have less riding time than casual.

- What trends or relationships did you find in the data?

  - **There are more members than casuals in the dataset.**
  - **There are more number of rides in the 2nd and 3rd Quarter of an year.**
  - **There are more of a difference between the flow of members/casual from midweek to weekends.**
  - **Members have less riding time.**
  - **Riders prefer docked bikes with members having a fascination for classic and Casual riders for Docked bike type.**

Key tasks

- ☒ Aggregate your data so it's useful and accessible.

- ☒ Organize and format your data.

- ☒ Perform calculations.

- ☒ Identify trends and relationships.

Deliverable

- ☒ A summary of your analysis.

SHARE

The problem scenario has been analysed properly, while analyzing we have came across certain key points which was decided to be further taken care of through some visuals.

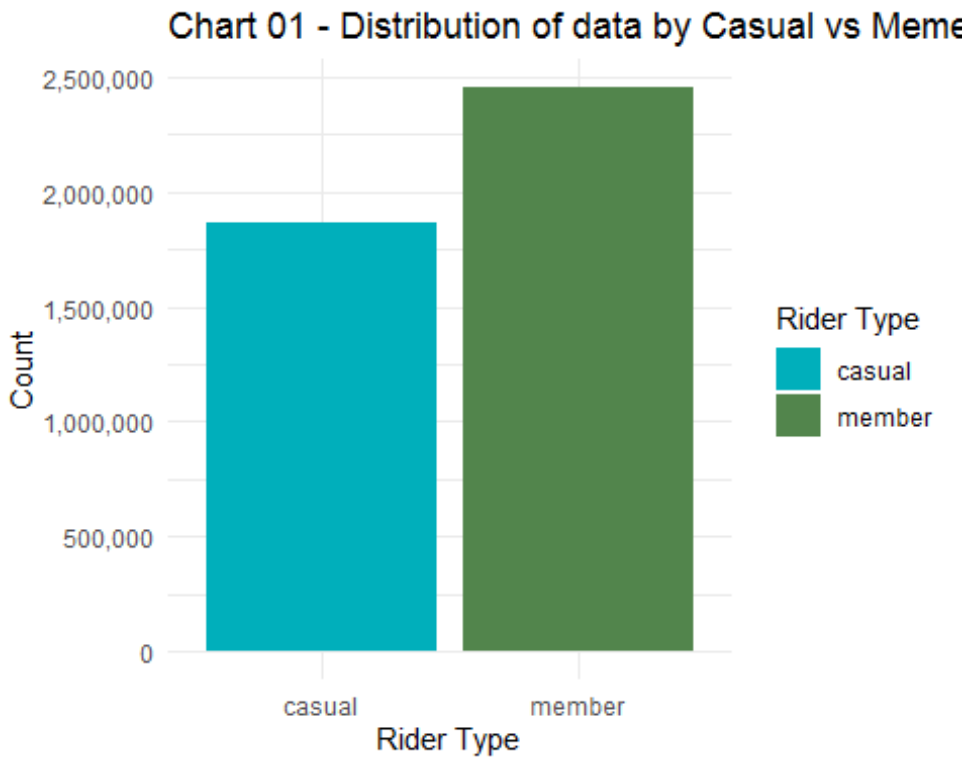
I will break down the whole analysis into sections and each case from our analyse phase will be considered and visualized which in turn will help us draw conclusions and key points for our marketing team.

Before we begin visualizing, we will load couple more libraries which will be required in plotting the graphs.

```
library(scales)

##
## Attaching package: 'scales'

## The following object is masked from 'package:purrr':
##
##     discard

## The following object is masked from 'package:readr':
##
##     col_factor

library(hrbrthemes)

## NOTE: Either Arial Narrow or Roboto Condensed fonts are required to use
these themes.

##       Please use hrbrthemes::import_roboto_condensed() to install Roboto
Condensed and

##       if Arial Narrow is not on your system, please see
https://bit.ly/arialnarrow

library(viridis)

## Loading required package: viridisLite

##
## Attaching package: 'viridis'

## The following object is masked from 'package:scales':
##
##     viridis_pal
```

% of Casual Riders vs Annual Members

```
ggplot(data = bike_rides)+
  geom_bar(mapping = aes(x = bike_rides$member_casual, fill =
bike_rides$member_casual ))+
  labs(title = "Chart 01 - Distribution of data by Casual vs Memeber" ,x =
"Rider Type", y = "Count", fill = "Rider Type")+
  scale_y_continuous(labels = comma) +
  scale_fill_manual(values = c("#00AFBB", "#52854C")) + theme_minimal()
```



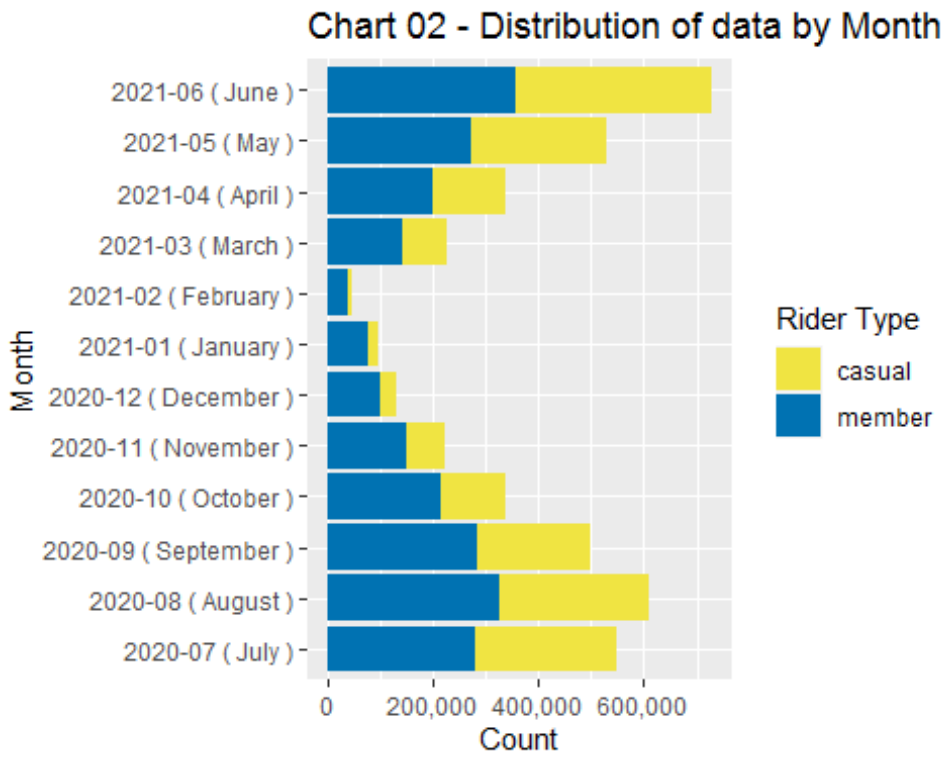
Key Conclusions to be made:

- We can see on the member vs casual bar graph, members have a bigger proportion of the dataset, Now, combining the visual with our previous summary we can conclude composing annual members comprise ~56% of total ride type of Cyclistic, ~13% bigger than the count of casual riders.

Total rides according to month

We already know, the maximum number of rides have been registered during the month of August, lets find further insights from the visuals.

```
bike_rides %>%
  ggplot(aes(x = year_month, fill = member_casual)) +
  geom_bar()+
  labs(title = "Chart 02 - Distribution of data by Month" ,x = "Month", y =
"Count", fill = "Rider Type")+
  scale_y_continuous(labels = comma)+
  scale_fill_manual(values = c("#F0E442", "#0072B2")) +
  coord_flip() + theme_gray()
```



Some important consideration can be taken from this visual:

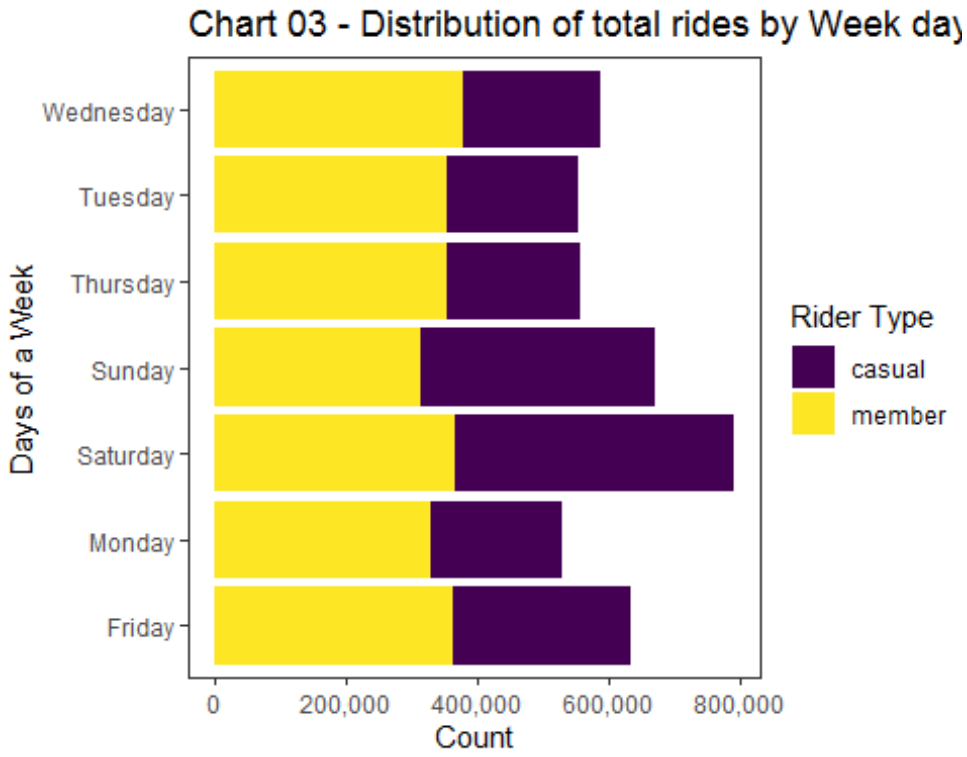
- There's more data points during the 2nd and 3rd Quarter of an year mainly during favorable climatic condition.
- The month with the biggest count of data points was June 2021.
- In most of the months we have more members' rides than casual rides.

Further to the above conclusion, the trends in data tells us that temperature and climate can be a important and considerable factor, this can be key information for our marketting team and thus should be concluded at the end.

Distribution of bike usage according to week days

Previously we have seen data saying that weekend records the maximum number of rides for our company, lets see whether we can prove that with visuals.

```
ggplot(data = bike_rides)+
  geom_bar(mapping = aes(bike_rides$weekday, fill =
bike_rides$member_casual))+
  labs(title = "Chart 03 - Distribution of total rides by Week days" ,x =
"Days of a Week", y = "Count", fill = "Rider Type")+
scale_y_continuous(labels = comma) +
  coord_flip() + scale_fill_viridis(discrete = TRUE) + theme_test()
```



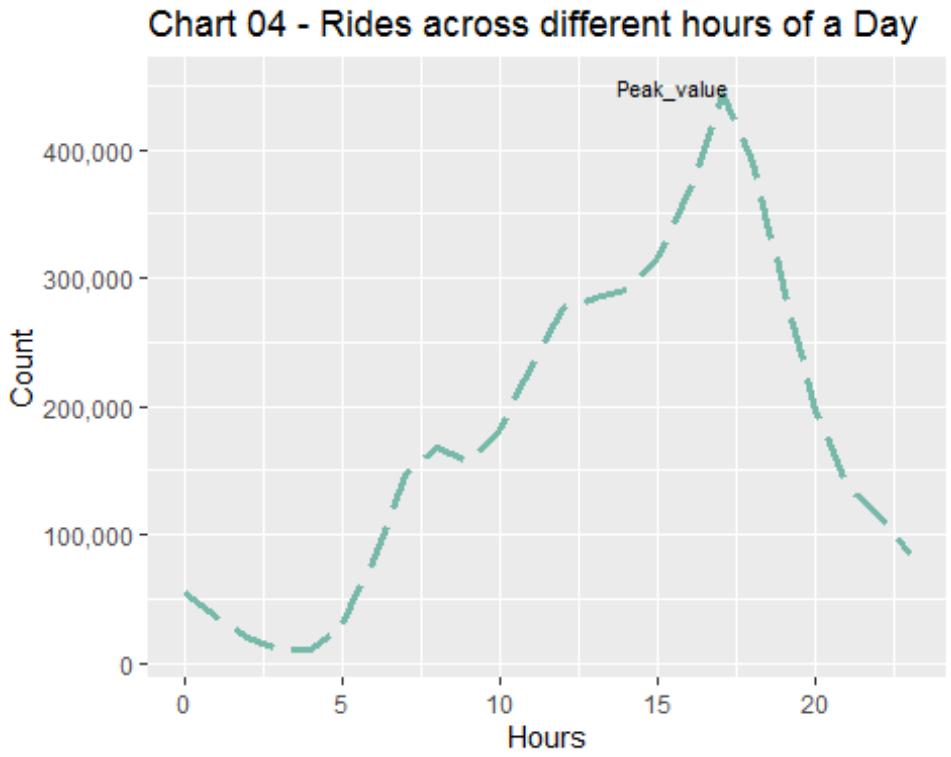
Key conclusions that can be drawn:

- The largest recorded data is during the weekend.
- Our company witness the maximum demand on Saturday.
- Casual rides shows active readings during Saturday and Sunday, which starts to rise from Friday, ~20% percent overall increases from weekdays.

Bike usage during different hours of a day

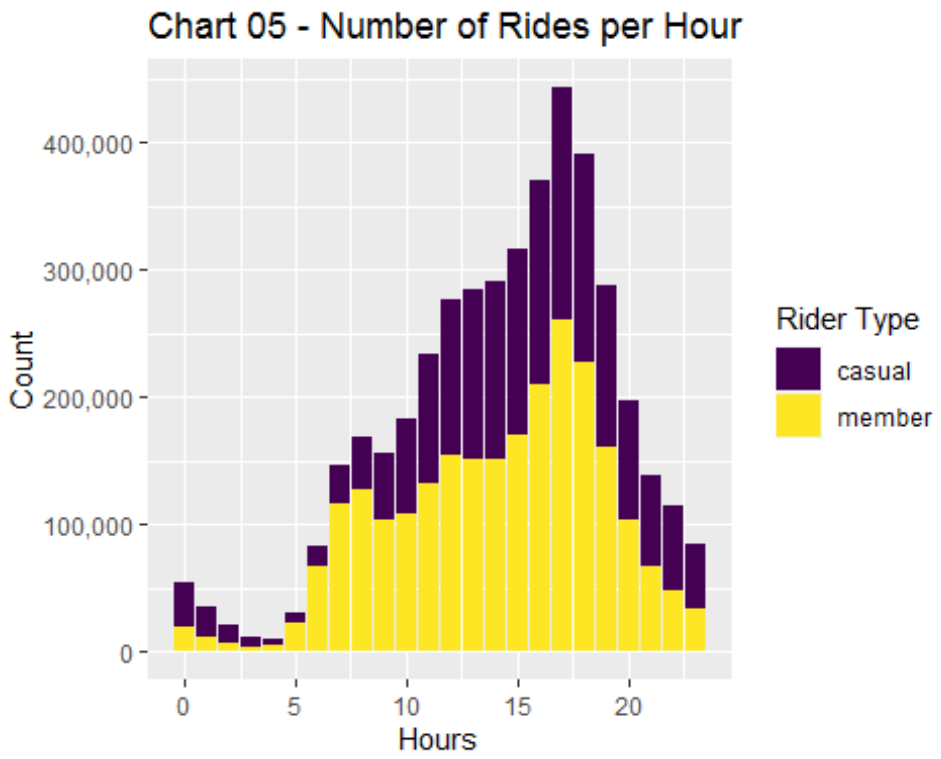
we decided to the explore this part of our analysis through some bar and line graphs, mainly to see how the data distribution has been recorded. Accordingly we can share the findings so that hourly usage may be considered while planning to maximize annual members.

```
bike_rides %>%
  select(start_hour,member_casual) %>%
  count(start_hour, sort = T) %>%
  ggplot()+
  geom_line(aes(x = start_hour, y = n), alpha = 0.9,linetype = 5, color =
"#69b3a2", lwd = 1.25) +
  labs(title = "Chart 04 - Rides across different hours of a Day",x  =
"Hours", y = "Count") +
  annotate("text",x = 15.5, y  = 450000, label = "Peak_value", size = 3) +
  scale_y_continuous(labels = comma)
```



We can clearly visualize, maximum bike rides has occurred during the afternoon hours from 3 PM - 4 PM. Lets further analyze this to see the distribution among rider type.

```
bike_rides %>%
  ggplot()+
  geom_bar(aes(x = start_hour,fill = member_casual)) +
  labs(title = "Chart 05 - Number of Rides per Hour",x  = "Hours", y =
"Count", fill = "Rider Type")+
  scale_fill_viridis(discrete = TRUE) + scale_y_continuous(labels = comma)
```

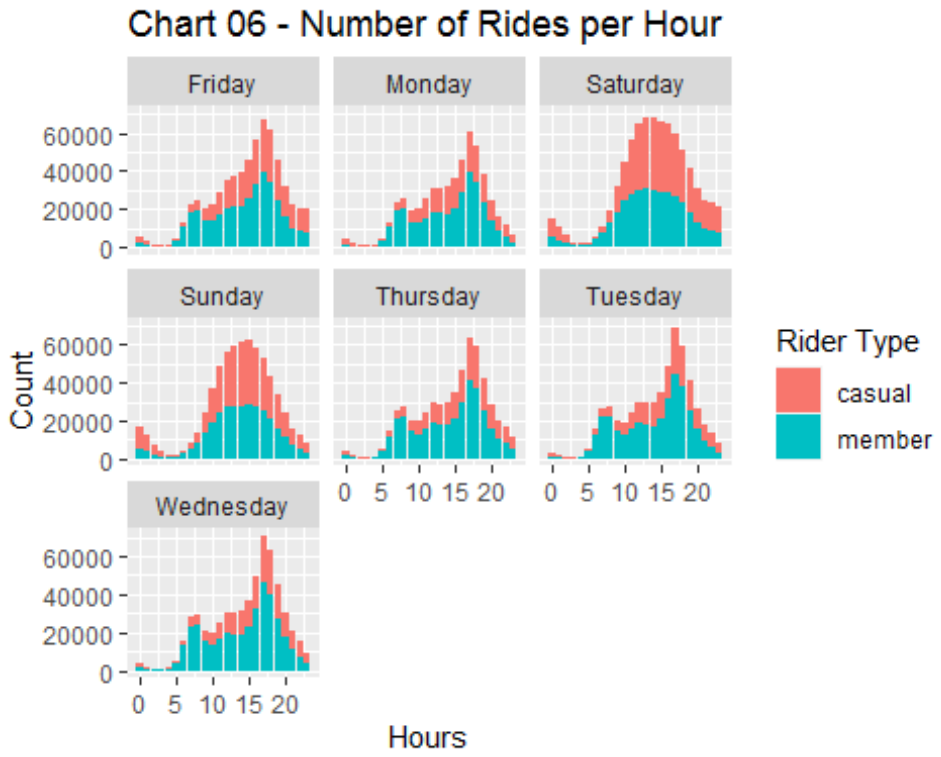


Points to be noted at this point:

- There's a bigger volume of bikers in the afternoon.
- One interesting fact is how number of members shows larger data points mainly during 5AM - 9 AM, which tends to fall,after that stayed consistent till it rises from afternoon to 6 PM. This observation can help us predict purpose of bike usage.
- More number of casuals between 11 AM - 4 PM.

We will drill down the hourly analysis according to various days of a week and to try to find similar trends in data.

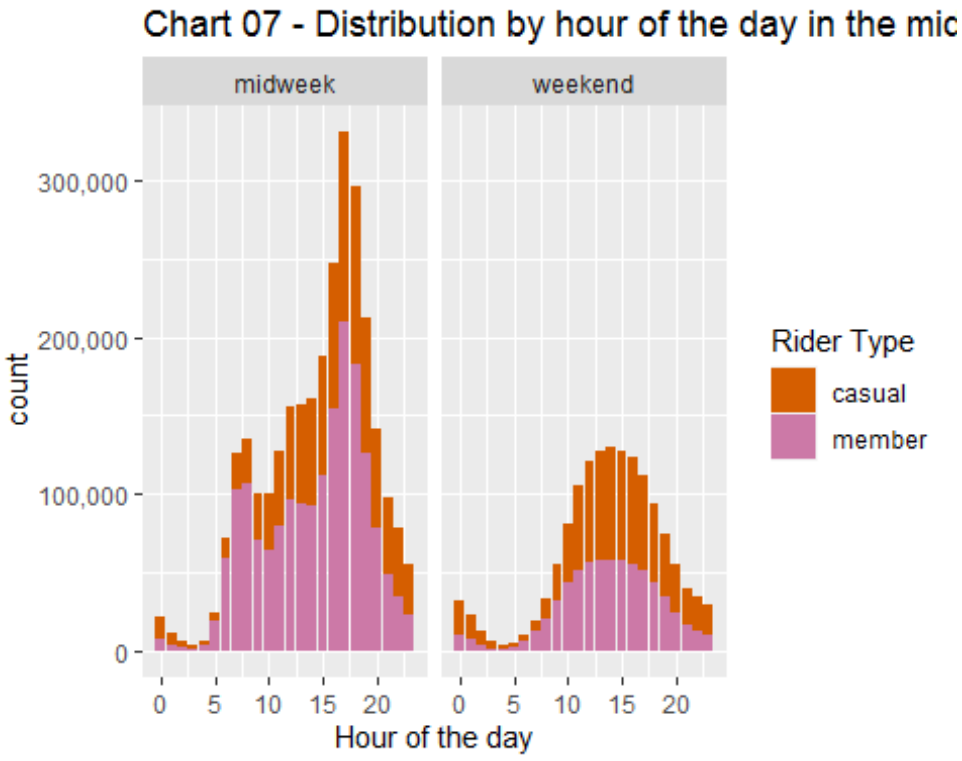
```
bike_rides %>%
  ggplot()+
  geom_bar(aes(x = start_hour,fill = member_casual)) +
  labs(title = "Chart 06 - Number of Rides per Hour",x  = "Hours", y =
"Count", fill = "Rider Type")+
  facet_wrap(~weekday)
```



we can spot some patterns, mainly between weekdays and weekends.

Lets compare the data on the basis of weekday and weekends:

```
bike_rides %>%
  mutate(type_of_weekday = ifelse(weekday == 'Saturday' | weekday ==
'Sunday',
                                  'weekend',
                                  'midweek')) %>%
  ggplot(aes(start_hour, fill=member_casual)) +
  labs(x="Hour of the day", title="Chart 07 - Distribution by hour of the day
in the midweek", fill  = "Rider Type") +
  geom_bar() +
  facet_wrap(~ type_of_weekday) +
  scale_fill_manual(values = c("#D55E00","#CC79A7")) +
scale_y_continuous(labels = comma)
```



The above two plots differ in the following ways:

- While the weekends have a smooth flow of data points, the midweek have a more steep flow of data.
- During weekdays, the bike usage tends to rise from 6 AM - 9 AM which shows a a slight fall and gradually rises reaching the maximum during 4 PM - 6 PM.
- During the weekend we have a bigger flow of casuals between 11am to 6pm.

**A important breakthrough, On carefully observing the patterns and considering a previous plot we can see during weekdays from (6AM - 9AM) and (3 PM to 6 PM) a large amount of member's data has been spotted.

Why do they require bikes particularly during those hours?

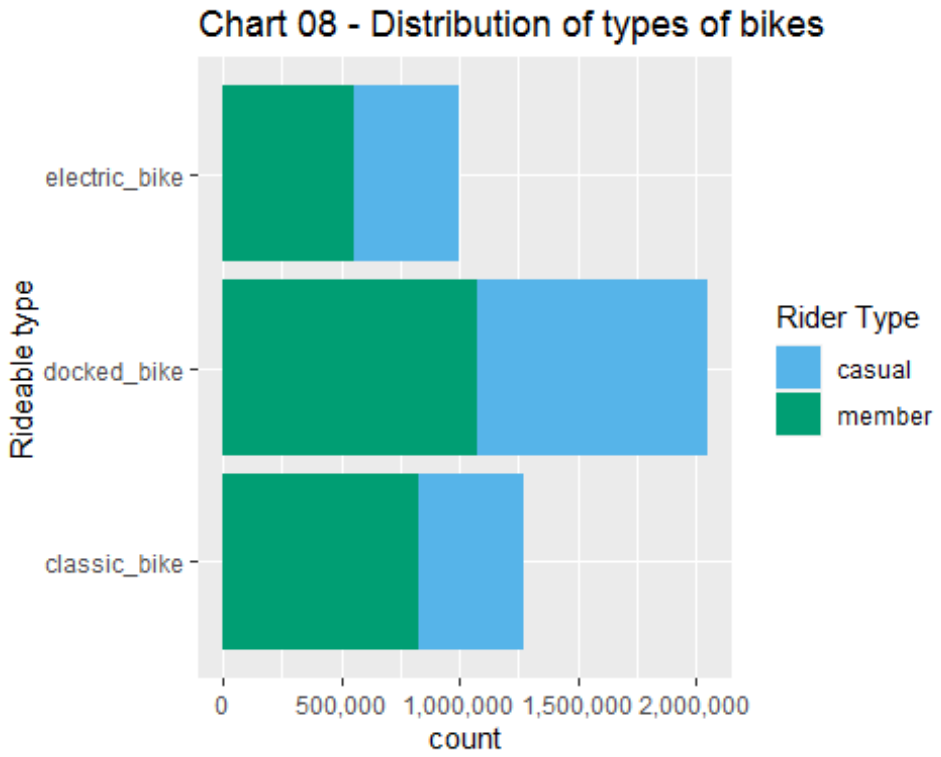
The answer can be, they use bikes for office commute purposes (data points between 6am to 9am in midweek), go back from work (data points between 3pm to 6pm).**

Most popular type of Bikes

Previously, docked bike has been identified as the most favorite, lets create a plot to see the same.

```
ggplot(bike_rides, aes(rideable_type, fill=member_casual)) +
  labs(x="Rideable type", title="Chart 08 - Distribution of types of bikes",
fill = "Rider Type") +
```

```
  geom_bar() + scale_y_continuous(labels = comma)+
  coord_flip() +
  scale_fill_manual(values = c("#56B4E9","#009E73"))
```



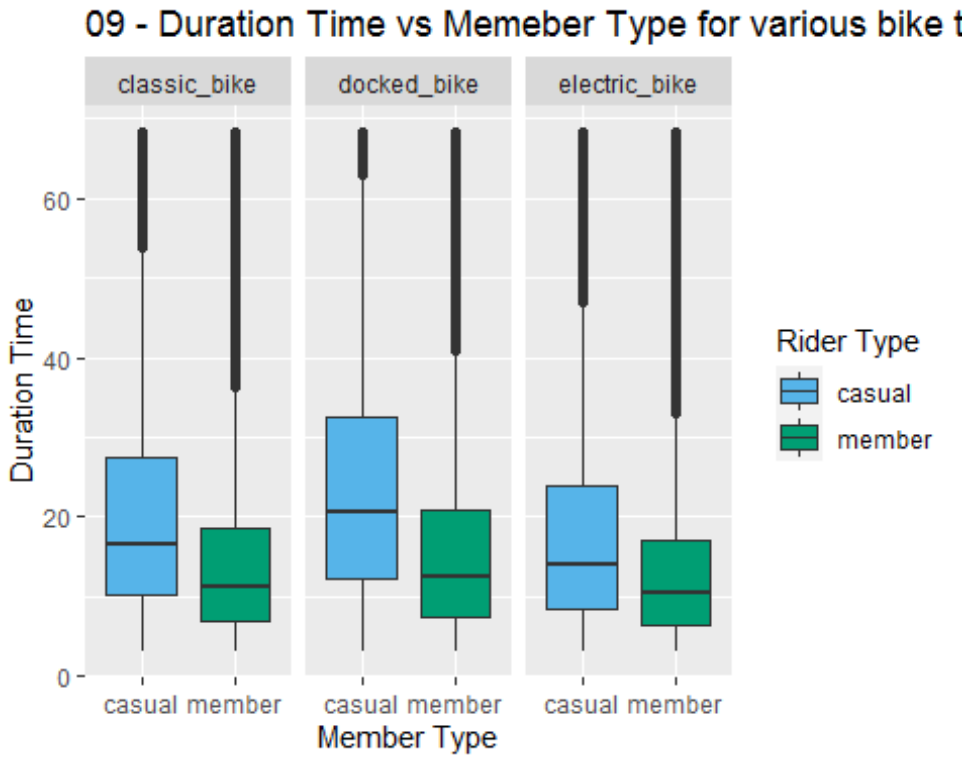
Important point to notes:

- Docked bikes have the biggest share of preference among users.
- In classic bikes, members comprise a great chunk of share to casuals.

Duration time

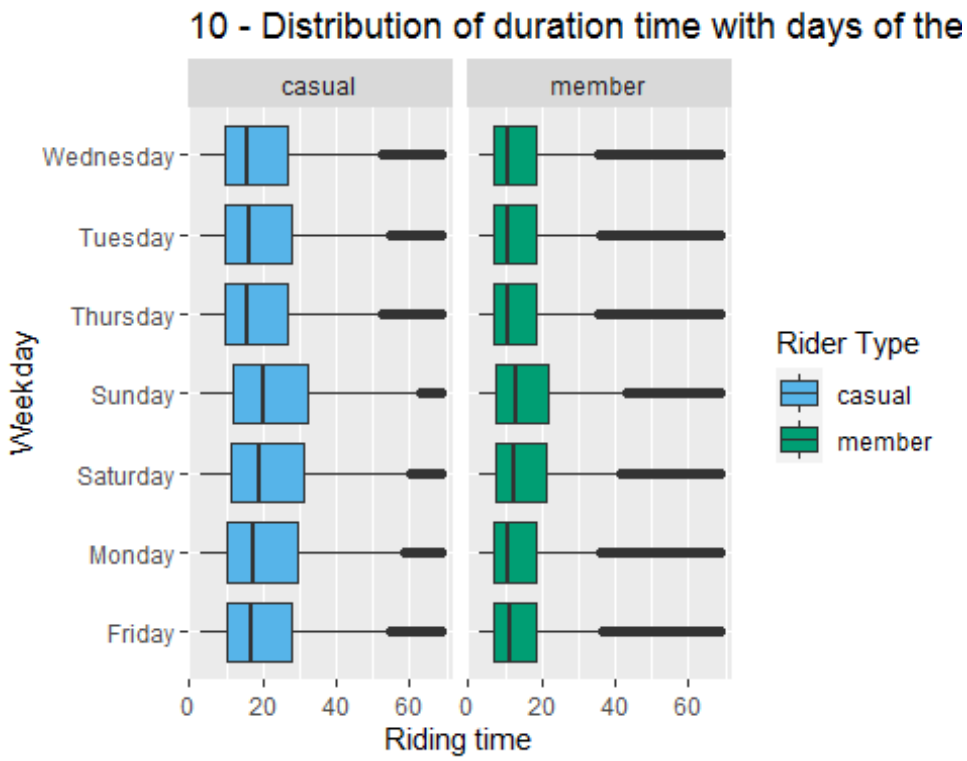
Finally we will focus on the duration time factor, we will work with our no outliers dataframe for avoiding skewed graphs.

```
ggplot(data = bike_rides_without_outliers, aes(x =
bike_rides_without_outliers$member_casual, y =
bike_rides_without_outliers$duration_time, fill  =
bike_rides_without_outliers$member_casual))+
  facet_wrap(~bike_rides_without_outliers$rideable_type) +
  labs(title = "09 - Duration Time vs Memeber Type for various bike types", x
= "Member Type", y  = "Duration Time", fill = "Rider Type") +
  geom_boxplot() + scale_fill_manual(values = c("#56B4E9","#009E73"))
```



We can further see how the duration time can be distributed according to weekdays by both Casual riders and Annual members.

```
ggplot(bike_rides_without_outliers, aes(x=weekday, y=duration_time,
fill=member_casual)) +
  geom_boxplot() +
  facet_wrap(~ member_casual) +
  labs(x="Weekday", y="Riding time", title="10 - Distribution of duration
time with days of the week", fill = "Rider Type") +
  scale_fill_manual(values = c("#56B4E9","#009E73")) +
  coord_flip()
```



Some key conclusion drawn from the above two visual:

- Duration time per ride for members seems to remain same through weekdays, but increases during weekends.
- Casuals too follow the same pattern but data is smoothly distributed.
- Electric bike has less ride time compared to the other categories.
- Among members, docked bikes are mostly preferred.

Staion name with maximum footfalls

Having seen how the duration time of each ride relates to our two main categories of rider type. We will further proceed to find some insights on the most popular station in the area.

We will be working with a new dataframe particularly for this section.

Installing `ggwordcloud` for plotting purpose.

```
library(ggwordcloud)
```

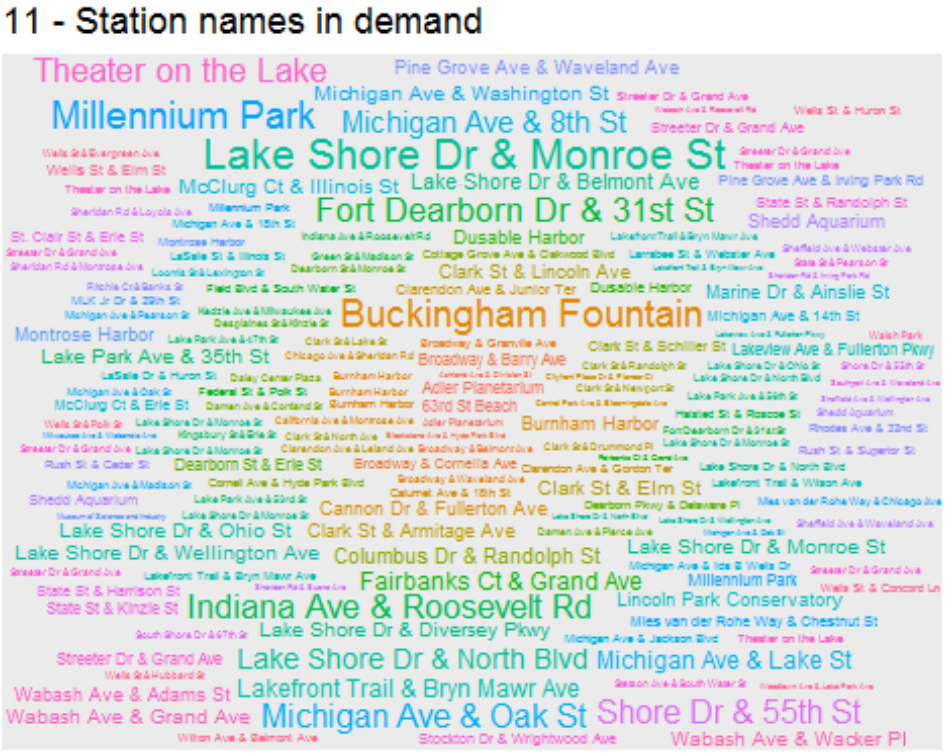
Lets create the dataframe:

```
df_new <- na.omit(rbind(df1,df2))
df_new <- df_new[!duplicated(df_new$ride_id), ]
```

```
station_max_footfall <- df_new %>%
  group_by(start_station_name, end_station_name) %>%
  summarise(count = length(as.factor(start_station_id)), count_n =
length(as.factor(end_station_id)))

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

We have created `station_max_footfall` dataframe from two of our previous dataframe after cleaning and removing bad data. Lets see the most popular places with the help of an aesthetic visual and try to extract some kind of insights from it.

```
station_max_footfall %>%
  filter(count >= 320) %>%
  ggplot(aes(label = start_station_name, size = count, color =
start_station_name)) +
  labs(title="11 - Station names in demand") +
  geom_text_wordcloud(area_corr = TRUE,rm_outside = TRUE, max_steps =
1,grid_size = 1, eccentricity = .9) +
  scale_size_area(max_size = 6.5) +
  theme_get()

## Some words could not fit on page. They have been removed.
```



And we can clearly visualize the most popular places in term of maximum bike rides for both category of riders. Some of them are `Streeter Dr & Grand Ave`, `Lake Shore Dr & Monroe St` etc.

Sometime with visuals we can easily spot important trends or data points.

Time to summarize all the findings to reach at a conclusion.

Key points about the dataset:

- Members have the biggest proportion of the dataset, ~13% bigger than casuals.
- There's more data points during 2nd and 3rd Quarter of an year.
- The month with the biggest count of data points was June with ~16% of the dataset.
- Temperature may have a huge influence on the usage pattern of rides in the month.
- The biggest volume of data is during the the weekend.
- There is a larger volume of bikers in the afternoon.
- Most in-demand stations has been identified.

We have seen the distribution of data through months, one possible question is why more members than casual riders? A probable answer to this can be members are those user who use bike ride for daily purposes such as office commute.

Another remarkable observation was huge increase in demands during weekends. This can be of the fact that both casual riders and annual members use bike for recreational activities or for exercising. This can be backed by the point that maximum bike ride has been recorded during the afternoon period even on weekends.

How members differ from casual rider?

- Members may have the biggest volume of data, besides on saturday. On this weekday, casuals take place as having the most data points.
- Weekends have the biggest volume of casuals, starting on friday, a ~16% increase.
- We have more members during the morning, mainly between 6am and 10am. And more casuals between 11am and 4pm.
- During weekdays, the bike usage tends to rise from 6 AM - 9 AM which shows a a slight fall and gradually rises reaching the maximum during 4 PM - 6 PM.
- During the weekend we have a bigger flow of casuals between 11am to 6pm.
- Casuals have more riding time than members.

- Docked bikes have the biggest share of preference among users.Although members prefer classic bike too.

- Riding time for members increase during weekends.

- Casuals follow a more curve distribution, peaking on Sundays and smoothing on Wednesday/Thursday.

From the above observations and trends in data among the two rider type category we can conclude:
* Members use bike for fixed purposes such as going to office.
* Bikes are used for exercise and recreation during weekends.
* Temperature and season plays a vital role for both category of user.
* Casual riders have overall more ride time than annual members.

Guiding Question

- Were you able to answer the question of how annual members and casual riders use Cyclistic's bikes differently?
  Yes, I have clearly spotted the difference in their usage pattern and documented them in this file.

- What story does your data tell?
  The main story that my data has generated is members use bike for more specific reason and thus spending less time where as casual riders spend more time with bikes.

- How do your findings relate to your original question?
  The findings build a profile for members, relating to *Find the keys differences between casuals and annual riders*, also knowing whey they use the bikes helps to find "How digital media could influence them.

- Who is your audience? What is the best way to communicate with them?
  My audience is Cyclistic's marketing team and of course my manager Lily Moreno. The best way to communicate is by giving a presentation of this whole process of analysis.

- Can data visualization help you share your findings?
  Ofcourse, data can often be confusing, where as visuals make it easier for everyone to understand the findings.

- Is your presentation accessible to your audience? **I hope this visuals are clearly plotted using diff kind of plots, used proper labeling and also filled with colors for easily distinguishable.**

Key taks

- ☒ Determine the best way to share your findings.

- ☒ Create effective data visualizations.
- ☒ Present your findings.
- ☒ Ensure your work is accessible.

Deliverable

- ☒ Supporting visualizations and key findings.

ACT

The main takeaway for this phase will be my top three recommendation on the basis of data analysis.

Guding Questions

- What is your final conclusion based on your analysis?
  Cyclistic's bike are used for different purposes between the two rider type. Members mainly use bikes for specific tasks while casual riders for variety of purposes. We have also identified station names with maximum footfalls.
- How could your team and business apply your insights?
  The insights from my above analysis can be further used and applied by the marketing team for creating ad campaigns that will target converting casual riders to annual members which will turn increase company revenue. They can use the station names showing higher demand to target the casual riders
- What next steps would you or your stakeholders take based on your findings?
  As said, the insights, trends and patterns can further be analysed to curate appropriate advertisements and policies.

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

- One limitation I faced is missing climate data with temperature.
- More information on avg pricing difference between annual and casual subscription could have helped gain further insights and allow us focus on the key problems.
- Distance between places could have also helped us predict why the most number of rides occur to and fro and also its geological and economical aspects.

Key tasks

- ☒ Create your portfolio.
- ☒ Add your case study.

- ☒ Practice presenting your case study to a friend or family member.

Deliverable

- My top three recommendations:
  - Since, we need to turn casual riders to members, marketing should be done showing how bikes can be game changer in case of saving time apart from office commute.
  - Cold months records low bike rides, thus various benefits can be distributed during that period including discounts and others.
  - As bike are also used during weekends for recreational purposes, ads showing how biking can keep your mind fresh and healthy should be focused. This will further add value to the company.

Conclusion

I want to thank Google and Coursera for creating this program for aspirants like me who want to explore the world of data and get familiarized with how data is `prepared`,`processed`,`analyzed` and finally `visualized`. Through the course Google Data Analytics Professional Certificate I have gained in depth knowledge about some of the most demanding programming languages such as `SQL`, `R` and familiarized working with tools like `Excel`, `Google Sheets`, `Big Query`, `R Studio` and finally `Tableau`. The above scenario could have been further analysed and for acquiring the certificate I will limit my findings up-till here.
