**Introduction –**

This is a capstone project as a part of my [Google Data Analytics Professional Certificate](https://www.coursera.org/professional-certificates/google-data-analytics) course. For the analysis, I will be analyzing a public dataset for a fictional company provided by the course I solving this project by using R programming language and [RStudio](https://www.kaggleusercontent.com/kf/87411416/eyJhbGciOiJkaXIiLCJlbmMiOiJBMTI4Q0JDLUhTMjU2In0..4N5P_aTAa_cYoomrZMtpPQ.4cIwo91216an0YGOce82qBeDMf-XtIApysWWyE71f9LKuOE328uAZO6RDavLNHjwm9WsdMBs97L2eCAhkzzDXqhaiXnJMTDMYHQxC8tXbpPuFK7PHIVvrohl9m8qKLDulfgdgqfNEC2d4O12ia1emNXSsP2a64LrLJBpW5Exx69zsXYtxaCf3gFKE-dzt5RK5NyMkraTQf-1mBdcqzMUw9P1f8xHwV_D1iuz7-6q0G_Ua9aa4hQN4RWkfjV5bj6byJHnYv6wItke2NNY4h7XxPcdEQJ65fiVMr3QFBFv-SpzYNizK-M99eFaOwIccY1iFNg3OS4UMqIbBJaTo7IeC1wWwnpVzYELPYFxk6nnsKBVfCkqxhaoAFqXclTnotuZ72-lOsn1CB5i95iDAlbe5tedcLgjOJnALcsFdIjoBeIP3N5QoAJU9IWOhMCUxlp8aoG-t5YzLtIhcitUrraudq0IIXYhrKHIRq5Hnv0l97n-3ANnEskK0N6Xg_5EKUiUKbi_LHsB8non0t-1Rv8cOxF9IzfvR-VOQuP3OvEp_3OwWGXtZw0JH2T-q5S1sPOgm_oH4nl7XMhGbsFrAg4Fd4a3mpYqIo4Db-k-Am5oC7wYhiWcSIhxzK93_Exs2jIzr8etVOXS1FvzPyb-fVcmRxHh4Z-o5WzqgixOupcTEiR89W8LfwWabnRhiLJ19hPp.UgM8bPilYI2bAnhQWv-qYg/www.rstudio.com) IDE for it's easy statistical analysis tools and data visualizations.

Following the steps of the data analysis process:

1. **Ask,**
2. **Prepare,**
3. **Process,**
4. **Analyze,**
5. **Share,**
6. **Act.**

**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.

1.**Ask**

Three questions will guide the future marketing program:

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?

Lily Moreno( The Director of Marketing Department and Manger) has assigned you the first question to answer: How do annual members and casual riders use Cyclistic bikes differently?

**A clear statement of the business task.**

The business task is to identify the differences between casual and annual riders in order to come up with a proper message to campaign to the casual users convincing them to switch to annual subscription.

**Key Stakeholders**

* Cyclistic executive team
* My manager, Lily Moreno
* Director of marketing
* The rest of the marketing analytics team

**2.Prepare**

**Data Preparation**

**Source of Data**

* Data was collected from [Motivate International Inc](https://divvy-tripdata.s3.amazonaws.com/index.html) under this [license](https://ride.divvybikes.com/data-license-agreement). Hence, it is public data.
* Since Cyclistic is a fictional company, the datasets have different names.

**Data organization**

* 12 csv. Files
* 13 variables

**Data credibility**

* The data appears to be in good condition, and it is first-hand information, with a large number of entries and a lot of important information.

**ROCCC Analysis**

* Reliability - I think it is reliable.
* Original - Yes
* Comprehensive - I think the data is not 100% comprehensive since some information is missing although its nominal.
* Current - 12 months old
* Cited - Yes

**Sort and filter the Data** The data which i used was from 2021-2022. This is so because it is the most relevant period to the aforementioned **business task**. In addition to that, it is the period when the data is complete whereby all geo-location coordinates and bike types are included.

**Install the required packages:**

**library**("tidyverse")

## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.1 ──

## ✓ ggplot2 3.3.5 ✓ purrr 0.3.4

## ✓ tibble 3.1.6 ✓ dplyr 1.0.8

## ✓ tidyr 1.2.0 ✓ stringr 1.4.0

## ✓ readr 2.1.2 ✓ forcats 0.5.1

## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──

## x dplyr::filter() masks stats::filter()

## x dplyr::lag() masks stats::lag()

**library**("ggplot2")

**library**("lubridate")

##

## Attaching package: 'lubridate'

## The following objects are masked from 'package:base':

##

## date, intersect, setdiff, union

**library**("knitr")

**library**("dplyr")

**Loading Data From 06/2022 to 05/2023**

X202206\_divvy\_tripdata <- read\_csv("C:/Users/shukl/OneDrive/Desktop/cyclistic bike data 2022-2023/csv files/202206-divvy-tripdata.csv")

X202207\_divvy\_tripdata <- read\_csv("C:/Users/shukl/OneDrive/Desktop/cyclistic bike data 2022-2023/csv files/202207-divvy-tripdata.csv")

X202208\_divvy\_tripdata <- read\_csv("C:/Users/shukl/OneDrive/Desktop/cyclistic bike data 2022-2023/csv files/202208-divvy-tripdata.csv")

X202206\_divvy\_publictripdata <- read\_csv("C:/Users/shukl/OneDrive/Desktop/cyclistic bike data 2022-2023/csv files/202209-divvy-publictripdata.csv")

X202210\_divvy\_tripdata <- read\_csv("C:/Users/shukl/OneDrive/Desktop/cyclistic bike data 2022-2023/csv files/202210-divvy-tripdata.csv")

X202211\_divvy\_tripdata <- read\_csv("C:/Users/shukl/OneDrive/Desktop/cyclistic bike data 2022-2023/csv files/202211-divvy-tripdata.csv")

X202212\_divvy\_tripdata <- read\_csv("C:/Users/shukl/OneDrive/Desktop/cyclistic bike data 2022-2023/csv files/202212-divvy-tripdata.csv")

X202301\_divvy\_tripdata <- read\_csv("C:/Users/shukl/OneDrive/Desktop/cyclistic bike data 2022-2023/csv files/202301-divvy-tripdata.csv")

X202302\_divvy\_tripdata <- read\_csv("C:/Users/shukl/OneDrive/Desktop/cyclistic bike data 2022-2023/csv files/202302-divvy-tripdata.csv")

X202303\_divvy\_tripdata <- read\_csv("C:/Users/shukl/OneDrive/Desktop/cyclistic bike data 2022-2023/csv files/202303-divvy-tripdata.csv")

X202304\_divvy\_tripdata <- read\_csv("C:/Users/shukl/OneDrive/Desktop/cyclistic bike data 2022-2023/csv files/202304-divvy-tripdata.csv”)

X202305\_divvy\_tripdata <- read\_csv("C:/Users/shukl/OneDrive/Desktop/cyclistic bike data 2022-2023/csv files/202305-divvy-tripdata.csv”)

**Combining all the data**

tripdata<-bind\_rows(X202206\_divvy\_tripdata,X202207\_divvy\_tripdata, X202208\_divvy\_tripdata,X202209\_divvy\_publictripdata, X202210\_divvy\_tripdata,X202211\_divvy\_tripdata, X202212\_divvy\_tripdata,X202301\_divvy\_tripdata, X202302\_divvy\_tripdata,X202303\_divvy\_tripdata, X202304\_divvy\_tripdata,X202305\_divvy\_tripdata).

**3.Process**

Cleaning and preparing for data analysis

**Key tasks**

1. Check the data for errors.
2. Choose your tools.
3. Transform the data so you can work with it effectively.
4. Document the cleaning process.

**Deliverable**

1. Documentation of any cleaning or manipulation of data

Following code chunks will be used for this phase.

head(trip\_data) #see the first 6 rows of the data frame

nrow(trip\_data) #how many rows are in the data frame

colnames(trip\_data) #list of column names

dim(trip\_data) #dimensions of the data frame

summary(trip\_data) #statistical summary of data, mainly for numerics

str(trip\_data) #see list of columns and data types

**Adding columns for date, month, year, day of the week into the data frame.**

tripdata$month <- format(as.Date(tripdata$date), "%m")

tripdata$year <- format(as.Date(tripdata$date),"%y")

tripdata$day <- format(as.Date(tripdata$date), "%d")

tripdata$day\_of\_the\_week <- format(as.Date(tripdata$date), "%A")

**Add a ride\_length to tripdata**

tripdata$ride\_length <- difftime(tripdata$ended\_at, tripdata$started\_at)

str(tripdata)

spc\_tbl\_ [5,829,030 × 20] (S3: spec\_tbl\_df/tbl\_df/tbl/data.frame)

$ ride\_id : chr [1:5829030] "600CFD130D0FD2A4" "F5E6B5C1682C6464" "B6EB6D27BAD771D2" "C9C320375DE1D5C6" ...

$ rideable\_type : chr [1:5829030] "electric\_bike" "electric\_bike" "electric\_bike" "electric\_bike" ...

$ started\_at : POSIXct[1:5829030], format: "2022-06-30 17:27:53" "2022-06-30 18:39:52" ...

$ ended\_at : POSIXct[1:5829030], format: "2022-06-30 17:35:15" "2022-06-30 18:47:28" ...

$ start\_station\_name: chr [1:5829030] NA NA NA NA ...

$ start\_station\_id : chr [1:5829030] NA NA NA NA ...

$ end\_station\_name : chr [1:5829030] NA NA NA NA ...

$ end\_station\_id : chr [1:5829030] NA NA NA NA ...

$ start\_lat : num [1:5829030] 41.9 41.9 41.9 41.8 41.9 ...

$ start\_lng : num [1:5829030] -87.6 -87.6 -87.7 -87.7 -87.6 ...

$ end\_lat : num [1:5829030] 41.9 41.9 41.9 41.8 41.9 ...

$ end\_lng : num [1:5829030] -87.6 -87.6 -87.6 -87.7 -87.6 ...

$ member\_casual : chr [1:5829030] "casual" "casual" "casual" "casual" ...

$ date : Date[1:5829030], format: "2022-06-30" "2022-06-30" ...

$ month : chr [1:5829030] "06" "06" "06" "06" ...

$ year : chr [1:5829030] "22" "22" "22" "22" ...

$ day : chr [1:5829030] "30" "30" "30" "30" ...

$ day\_of\_the\_week : chr [1:5829030] "Thursday" "Thursday" "Thursday" "Thursday" ...

$ ride\_length : 'difftime' num [1:5829030] 442 456 809 258 ...

..- attr(\*, "units")= chr "secs"

$ ride\_distance : num [1:5829030] 2.221 2.371 3.994 0.831 2.371 ...

- attr(\*, "spec")=

.. cols(

.. ride\_id = col\_character(),

.. rideable\_type = col\_character(),

.. started\_at = col\_datetime(format = ""),

.. ended\_at = col\_datetime(format = ""),

.. start\_station\_name = col\_character(),

.. start\_station\_id = col\_character(),

.. end\_station\_name = col\_character(),

.. end\_station\_id = col\_character(),

.. start\_lat = col\_double(),

.. start\_lng = col\_double(),

.. end\_lat = col\_double(),

.. end\_lng = col\_double(),

.. member\_casual = col\_character()

.. )

- attr(\*, "problems")=<externalptr>

Convert ride\_length from Factor to Numeric in order to run calculations

tripdata$ride\_length <- as.numeric(as.character(tripdata$ride\_length))

is.numeric(tripdata$ride\_length)

Add ride\_distance calculation to the tripdata

tripdata$ride\_distance <- distGeo(matrix(c(tripdata$start\_lng, tripdata$start\_lat), ncol = 2), matrix(c(tripdata$end\_lng, tripdata$end\_lat), ncol = 2))

tripdata$ride\_distance <- tripdata$ride\_distance/1000

Remove bad data

tripdata\_clean <- tripdata[!(tripdata$ride\_length <=0),]

glimpse(tripdata\_clean)

4. **Analyze**

Now that your data is stored appropriately and has been prepared for analysis, start putting it to work.

Key tasks

1.Aggregate your data so it’s useful and accessible.

2. Organize and format your data.

3. Perform calculations.

4. Identify trends and relationships.

**Deliverable**

1.**Summary of my analysis**

Summary(tripdata\_clean)

ride\_id rideable\_type started\_at

Length:5828463 Length:5828463 Min. :2022-06-01 00:00:04.00

Class :character Class :character 1st Qu.:2022-07-27 14:53:43.00

Mode :character Mode :character Median :2022-09-23 07:55:42.00

Mean :2022-10-28 01:02:05.01

3rd Qu.:2023-02-05 15:35:42.00

Max. :2023-05-31 23:59:58.00

ended\_at start\_station\_name start\_station\_id

Min. :2022-06-01 00:02:38.00 Length:5828463 Length:5828463

1st Qu.:2022-07-27 15:14:32.50 Class :character Class :character

Median :2022-09-23 08:13:54.00 Mode :character Mode :character

Mean :2022-10-28 01:20:47.88

3rd Qu.:2023-02-05 15:51:28.50

Max. :2023-06-07 23:04:26.00

end\_station\_name end\_station\_id start\_lat start\_lng

Length:5828463 Length:5828463 Min. :41.64 Min. :-87.87

Class :character Class :character 1st Qu.:41.88 1st Qu.:-87.66

Mode :character Mode :character Median :41.90 Median :-87.64

Mean :41.90 Mean :-87.65

3rd Qu.:41.93 3rd Qu.:-87.63

Max. :42.07 Max. :-87.52

end\_lat end\_lng member\_casual date

Min. : 0.00 Min. :-88.11 Length:5828463 Min. :2022-06-01

1st Qu.:41.88 1st Qu.:-87.66 Class :character 1st Qu.:2022-07-27

Median :41.90 Median :-87.64 Mode :character Median :2022-09-23

Mean :41.90 Mean :-87.65 Mean :2022-10-27

3rd Qu.:41.93 3rd Qu.:-87.63 3rd Qu.:2023-02-05

Max. :42.37 Max. : 0.00 Max. :2023-05-31

NA's :5961 NA's :5961

month year day day\_of\_the\_week

Length:5828463 Length:5828463 Length:5828463 Length:5828463

Class :character Class :character Class :character Class :character

Mode :character Mode :character Mode :character Mode :character

ride\_length ride\_distance

Min. : 1 Min. : 0.000

1st Qu.: 335 1st Qu.: 0.867

Median : 591 Median : 1.555

Mean : 1123 Mean : 2.115

3rd Qu.: 1057 3rd Qu.: 2.757

Max. :2483235 Max. :9817.319

NA's :5961

Descriptive Analysis

* mean = straight average (total ride length / total rides)
* median = midpoint number of ride length
* max = longest ride
* min = shortest ride

tripdata\_clean %>%

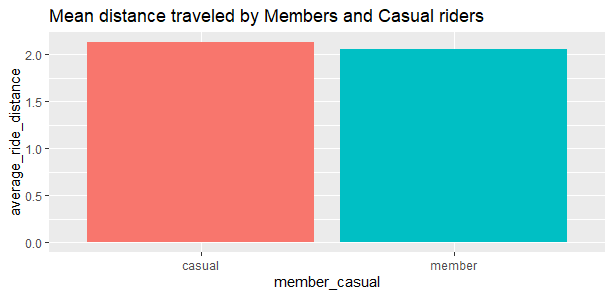
group\_by(member\_casual, day\_of\_the\_week) %>%

summarise(number\_of\_rides = n()

,average\_ride\_length = mean(ride\_length),.groups="drop") %>%

arrange(member\_casual, day\_of\_the\_week)

|  |
| --- |
| # A tibble: 14 × 4  member\_casual day\_of\_the\_week number\_of\_rides average\_ride\_length  *<chr>* *<chr>* *<int>* *<dbl>*  1 casual Friday 346524 1642.  2 casual Monday 257799 1654.  3 casual Saturday 455777 1925.  4 casual Sunday 375582 2003.  5 casual Thursday 313310 1462.  6 casual Tuesday 270936 1517.  7 casual Wednesday 291886 1437.  8 member Friday 501423 741.  9 member Monday 471362 711.  10 member Saturday 452714 835.  11 member Sunday 394506 831.  12 member Thursday 564862 722.  13 member Tuesday 552484 719.  14 member Wednesday 579298 715.  Compare member vs Casual rides  tripdata\_clean %>%  group\_by(member\_casual) %>%  summarise(ride\_count = length(ride\_id))  # A tibble: 2 × 2  member\_casual ride\_count  *<chr>* *<int>*  1 casual 2311814  2 member 3516649  *Visualize*  ggplot(tripdata\_clean, aes(x = member\_casual, fill=member\_casual)) +  geom\_bar() +  labs(x="Casuals vs Members", y="Number Of Rides", title= "Casuals vs Members distribution")    **Conclusion**  From the above graph, we can observe that there are more member riders compared to casual rides based on the ride count. Total rides and average ride time by each day for members vs casual riders tripdata\_clean %>%  + group\_by(member\_casual, day\_of\_the\_week) %>%  + summarise(number\_of\_rides = n(), .groups="drop") %>%  + arrange(member\_casual, day\_of\_the\_week) %>%  + ggplot(aes(x = day\_of\_the\_week, y = number\_of\_rides, fill = member\_casual)) +  + labs(title ="Total rides by Members and Casual riders Vs. Day of the week") +  + geom\_col(width=0.5, position = position\_dodge(width=0.5)) +  + scale\_y\_continuous(labels = function(x) format(x, scientific = FALSE)  *Visualize*    Conclusion:  From the above Graph it is observed that Casual have highest numbers of rides on the Weekends compared to other days while members are quiet consistent but they have lowest number of rides during Weekends.  Visualization of Average ride time by Members and Casual rider’s vs Day of the week  tripdata\_clean %>%  group\_by(member\_casual, day\_of\_the\_week) %>%  summarise(average\_ride\_length = mean(ride\_length), .groups="drop") %>%  ggplot(aes(x = day\_of\_the\_week, y = average\_ride\_length, fill = member\_casual)) +  geom\_col(width=0.5, position = position\_dodge(width=0.5)) +  labs(title ="Average ride time by Members and Casual riders Vs. Day of the week")  *Visualize*    Conclusion:    From the above Graph it is observed that Casual rider have highest ride time on the Weekends as compared to other days and the Members have pretty consistent ride time during the week and highest ride time on Weekends.  **Visualization of Total rides taken by Members and Casuals vs Month**  tripdata\_clean %>%  group\_by(member\_casual, month) %>%  summarise(number\_of\_rides = n(),.groups="drop") %>%  arrange(member\_casual, month) %>%  ggplot(aes(x = month, y = number\_of\_rides, fill = member\_casual)) +  labs(title ="Total rides by Members and Casual riders Vs. Month", x = "Month", y= "Number Of Rides") +  theme(axis.text.x = element\_text(angle = 45)) +  geom\_col(width=0.5, position = position\_dodge(width=0.5)) +  scale\_y\_continuous(labels = function(x) format(x, scientific = FALSE))    **Conclusion**  From the above graph we can observe that members had higher number of rides all through the year except in June, July and August where casual riders took more rides. Let's compare Members and Casual riders depending on ride distance tripdata\_clean %>%  group\_by(member\_casual) %>% drop\_na() %>%  summarise(average\_ride\_distance = mean(ride\_distance)) %>%  ggplot() +  geom\_col(mapping= aes(x= member\_casual,y= average\_ride\_distance,fill=member\_casual), show.legend = FALSE)+  labs(title = "Mean distance traveled by Members and Casual riders")  *Visualize* |
|  |
| |  | | --- | |  | |



From the above graph, we can observe that casual riders went a longer distance compared to members by a few kilometers.

**5.Share my conclusions**

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

1. Casual users have most rides during weekends.
2. Member users have been taking more rides than casual members.

**Recommendations**

Since both rides and duration for casual users are more during weekends than weekdays. I suggest Cycllistic company to add some of the offers related to holiday, tourism or weekend to Weekday promotions**.** By doing so, more casual users will be motivated to convert their membership since there will be a slight or no difference between the two membership types.

**6.ACT**

This phase will be carried out by the executive team, Director of Marketing (Lily Moreno) and the Marketing Analytics team based on my analysis.

**Further Analysis**

For further analysis, i think it will be necessary to know the preferred type of bikes, locations or most popular routes by casual users. This will help the company by to get insights which might lead them to offer the best promotions in those specific routes to convert more casual users to member users.

**Thanks for Reading**