# Data Analysis Project #1

## Junior Data Analyst: Zhumabek Nurlybek (R)

## Client/Sponsor: Cyclistic, bike-share company

## Purpose:

*The goal of this project is to study how annual members and casual riders of a Cyclistic, bike-share company differ, and why casual riders would buy annual membership, and how digital media could affect marketing tactics. The project aims to design marketing strategies for converting casual riders to annual members.*

## Deliverables:

*A specific list of things that your project will deliver.*

|  |  |
| --- | --- |
| Deliverable | Description/ Details |
| Clear statement of business task | I was assigned to answer the following question: How do annual members and casual riders use Cyclistic bikes differently? |
| Description of all data sources used | I downloaded Cyclistic’s historical trip data from the Motivate International Inc. database. Data is credible (ROCCC: Reliable, Original, Comprehensive, Current and Cited). |
| Documentation of any cleaning or manipulation of data | Documentation about cleaning or manipulating data is provided below. |
| Summary of my analysis | Summary of analysis is also provided |
| Supporting visualizations and key findings | Visualization is created via Microsoft Excel and provided some analysis |
| My top three recommendations based on my analysis | Conclusion is made by answering the question |

Clear Statement of Business Task

The business problem is how we can maximize the number of annual memberships of the Cyclistic bike share company as defined by manager, Lily moreno. I was assigned to answer the question that states how annual members and casual riders use Cyclistic bikes differently, by a marketing analytics team. My findings regarding the following question will help to drive the decision making process of the executive team for planning for implementation.

Description of all data sources used

The dataset is located at a publicly available HTML webpage with direct links to download the particular data based on its timeframe. The dataset is organized by time, starting from recent to oldest data available. Each data is given as a .csv format file. The current dataset is recognized as credible (ROCCC: Reliable, Original, Comprehensive, Current and Cited). Also the privacy, security and accessibility terms are covered by the [license](https://divvybikes.com/data-license-agreement). Data integrity will be reached by cleaning data from possible errors at the next stage.

Documentation of any cleaning or manipulation of data (R)

Install and load necessary packages

|  |
| --- |
| >install.packages(“tidyverse”) >library(tidyverse)  >install.packages(“janitor”) >library(janitor)  >install.packages(“ggmap”) >library(ggmap)  >install.packages(“geosphere”) >library(geosphere)  >install.packages(“lubridate”) >library(lubridate) |

Uploading dataset of Cyclistic bike-share

|  |
| --- |
| >jan23 <- read\_csv(“202301-divvy-tripdata\_Jan.csv”)  >feb23 <- read\_csv(“202302-divvy-tripdata\_Feb.csv”)  >mar23 <- read\_csv(“202303-divvy-tripdata\_Mar.csv”)  >apr23 <- read\_csv(“202304-divvy-tripdata\_Apr.csv”)  >may23 <- read\_csv(“202305-divvy-tripdata\_May.csv”)  >jun23 <- read\_csv(“202306-divvy-tripdata\_Jun.csv”)  >jul23 <- read\_csv(“202307-divvy-tripdata\_Jul.csv”)  >aug23 <- read\_csv(“202308-divvy-tripdata\_Aug.csv”)  >sep23 <- read\_csv(“202309-divvy-tripdata\_Sep.csv”)  >oct23 <- read\_csv(“202310-divvy-tripdata\_Oct.csv”)  >nov23 <- read\_csv(“202311-divvy-tripdata\_Nov.csv”)  >dec23 <- read\_csv(“202312-divvy-tripdata\_Dec.csv”) |

Check data for consistency and structure

|  |
| --- |
| >colnames(jan23) >str(jan23)  >colnames(feb23) >str(feb23)  >colnames(mar23) >str(mar23)  >colnames(apr23) >str(apr23)  >colnames(may23) >str(may23)  >colnames(jun23) >str(jun23)  >colnames(jul23) >str(jul23)  >colnames(aug23) >str(aug23)  >colnames(sep23) >str(sep23)  >colnames(oct23) >str(oct23)  >colnames(nov23) >str(nov23)  >colnames(dec23) >str(dec23) |

Merging individual monthly datasets into one data frame

|  |
| --- |
| >tripdata23 <- bind\_rows(jan23,feb23,mar23,apr23,may23,jun23,jul23,aug23,sep23,oct23,nov23,dec23) |

Checking merged data frame

|  |
| --- |
| >colnames(tripdata23)  >head(tripdata23)  >str(tripdata23)  >summary(tripdata23) |

Adding date, month, year, day of week columns

|  |
| --- |
| > tripdata23 <- tripdata23 %>%  mutate(year=format(as.Date(started\_at),"%Y")) %>%  mutate(month=format(as.Date(started\_at),"%B")) %>%  mutate(date=format(as.Date(started\_at),"%d")) %>% mutate(day\_of\_week=format(as.Date(started\_at),"%A")) %>% mutate(ride\_length=difftime(ended\_at,started\_at)) %>%  mutate(start\_time=strftime(started\_at, "%H")) |

Converting “ride\_length” to numeric data type for calculation

|  |
| --- |
| > tripdata23 <- tripdata23 %>% mutate(ride\_length=as.numeric(ride\_length))  > is.numeric(tripdata23$ride\_length) recheck |

Adding ride distance in km

|  |
| --- |
| > tripdata23$ride\_distance <- distGeo(matrix(c(tripdata23$start\_lng,tripdata23$start\_lat),ncol = 2),matrix(c(tripdata23$end\_lng,tripdata23$end\_lat),ncol = 2))  > tripdata23$ride\_distance <- tripdata23$ride\_distance/1000 |

Data cleaning procedures

|  |
| --- |
| > colSums(is.na(tripdata23)) - check    > tripdata23\_cln1 <- tripdata23[!is.na(tripdata23$end\_lat),] – remove null values in “eng\_lat”  > colSums(is.na(tripdata23\_cln1))    Checking for blanks as “ “  > sum(tripdata23\_cln1$start\_station\_name == '')  [1] 875716  > sum(tripdata23\_cln1$start\_station\_id == '')  [1] 875848  > sum(tripdata23\_cln1$end\_station\_id == '')  [1] 922469  > sum(tripdata23\_cln1$end\_station\_name == '')  [1] 922328  Delete those blanks “ “ by creating another new dataframes (versions)  > tripdata23\_cln2 <- tripdata23\_cln1[trimws(tripdata23\_cln1$start\_station\_name) !='',]  > tripdata23\_cln3 <- tripdata23\_cln2[trimws(tripdata23\_cln2$start\_station\_id) !='',]  > tripdata23\_cln4 <- tripdata23\_cln3[trimws(tripdata23\_cln3$end\_station\_name) !='',]  > tripdata23\_cln5 <- tripdata23\_cln4[trimws(tripdata23\_cln4$end\_station\_id) !='',]  Removing those versions  > rm(tripdata23\_cln4)  > rm(tripdata23\_cln3)  > rm(tripdata23\_cln2)  > rm(tripdata23\_cln1) |

A summary of an analysis (R)

Conduct descriptive analysis

|  |
| --- |
| **Determining mean, max, min, median**  > tripdata23\_cln5 %>% summarise(average\_ride\_length=mean(ride\_length),median\_length=median(ride\_length),max\_ride\_length=max(ride\_length),min\_ride\_length=min(ride\_length))    **Members vs casual riders difference depending on total rides taken**  > tripdata23\_cln5 %>% group\_by(member\_casual) %>% summarise(ride\_count=length(ride\_id),ride\_percentage=(length(ride\_id)/nrow(tripdata23\_cln5))\*100)    **Casual vs member riders’ distribution**  > ggplot(tripdata23\_cln5, aes(x=member\_casual,fill = member\_casual)) + geom\_bar() + labs(x="Casuals vs Members",y="Number of Rides",title = "Casuals vs Members distribution")    **Comparison between Members Causal riders depending on ride length (mean, median, minimum, maximum)**  > tripdata23\_cln5 %>% group\_by(member\_casual) %>% summarise(average\_ride\_length=mean(ride\_length),median\_length=median(ride\_length),max\_ride\_length=max(ride\_length),min\_ride\_length=min(ride\_length))  A computer code with black text  Description automatically generated with medium confidence |

Supporting visualizations and key findings (R)

|  |
| --- |
| **See total rides and average ride time by each day for members vs casual riders**  > tripdata23\_cln5$day\_of\_week <- ordered(tripdata23\_cln5$day\_of\_week,levels=c("Sunday","Monday","Tuesday","Wednesday","Thursday","Friday","Saturday")) – **lets fix days of week order**  > tripdata23\_cln5 %>% group\_by(member\_casual,day\_of\_week) %>% summarise(number\_of\_rides = n(),average\_ride\_length=mean(ride\_length),.groups = "drop") %>% arrange(member\_casual,day\_of\_week)    **Visualize total rides data by type and day of week**  > tripdata23\_cln5 %>% group\_by(member\_casual,day\_of\_week) %>% summarise(number\_of\_rides=n(),.groups = "drop") %>% arrange(member\_casual,day\_of\_week) %>% ggplot(aes(x=day\_of\_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 average ride time data by type and day of week**  > tripdata23\_cln5 %>% group\_by(member\_casual,day\_of\_week) %>% summarise(average\_ride\_length=mean(ride\_length),.groups = "drop") %>% ggplot(aes(x=day\_of\_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")    **See total rides and average ride time by each month for members vs casual riders**  > tripdata23\_cln5$month <- ordered(tripdata23\_cln5$month,levels=c("January","February","March","April","May","June","July","August","September","October","November","December")) – lets fix months of year order  > tripdata23\_cln5 %>% group\_by(member\_casual,month) %>% summarise(number\_of\_rides=n(),average\_ride\_length=mean(ride\_length),.groups = "drop") %>% arrange(member\_casual,month)    **Visualize total rides data by type and month**  > tripdata23\_cln5 %>% 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.title.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))  A graph with blue and red bars  **Visualize average ride time data by type and month**  > tripdata23\_cln5 %>% group\_by(member\_casual,month) %>% summarise(average\_ride\_length=mean(ride\_length),.groups = "drop") %>% ggplot(aes(x=month,y=average\_ride\_length,fill = member\_casual))+geom\_col(width = 0.5,position = position\_dodge(width = 0.5)) + labs(title = "Average ride length by Members and Casual riders Vs. Month") + theme(axis.title.x = element\_text(angle = 30))    **Comparison between Members and Casual riders depending on ride distance**  > tripdata23\_cln5 %>% group\_by(member\_casual) %>% 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 travel distance by Members and Casual riders",x="Member and Casual riders",y="Average distance In Km")    Analysis and visualization on cyclistic's bike demand by hour in a day  > tripdata23\_cln5 %>% ggplot(aes(start\_time,fill=member\_casual)) + labs(x="Hour of the day",title = "Cyclistic's Bike demand by hour in a day") + geom\_bar()    Analysis and visualization on cyclistic's bike demand per hour by day of the week  > tripdata23\_cln5 %>% ggplot(aes(start\_time,fill = member\_casual)) + geom\_bar() + labs(x="Hour of the day",title = "Cyclistic's bike demand per hour by day of the week") + facet\_wrap(~ day\_of\_week)    Analysis and visualization of Rideable type Vs. total rides by Members and casual riders  > tripdata23\_cln5 %>% group\_by(rideable\_type) %>% summarise(count = length(ride\_id))    > ggplot(tripdata23\_cln5, aes(x=rideable\_type,fill = member\_casual)) + labs(x="Rideable type", title = "Rideable type Vs. total riders by Members and casual riders") + geom\_bar()    Now analyze and visualize the dataset on coordinate basis  #Lets check the coordinates data of the rides.  #adding a new data frame only for the most popular routes >200 rides  > coordinates\_df <- tripdata23\_cln5 %>% filter(start\_lng !=end\_lng & start\_lat !=end\_lat) %>% group\_by(start\_lng, start\_lat, end\_lng, end\_lat, member\_casual, rideable\_type) %>% summarise(total\_rides = n(),.groups = "drop") %>% filter(total\_rides>200)  # now lets create two different data frames depending on rider type (member\_casual)  > casual\_riders <- coordinates\_df %>% filter(member\_casual=="casual")  > member\_riders <- coordinates\_df %>% filter(member\_casual=="member")  Lets setup ggmap and store map of Chicago (bbox, stamen map)  > chicago <- c(left = -87.700424, bottom = 41.790769, right = -87.554855, top = 41.990119)  Visualization on the map  Registered on website of Stadiamap and received API key:  > register\_stadiamaps(key = "d031d9b2-74eb-4a34-9889-ec5d366984ab",write = FALSE)  > stadiamaps\_key()  [1] "d031d9b2-74eb-4a34-9889-ec5d366984ab"  > has\_stadiamaps\_key()  [1] TRUE  > chicago\_map <- get\_stadiamap(bbox = chicago, zoom = 12, maptype = "stamen\_terrain")  ℹ © Stadia Maps © Stamen Design © OpenMapTiles © OpenStreetMap contributors.  > ggmap(chicago\_map,darken = c(0.1,"white")) + geom\_point(casual\_riders,mapping = aes(x=start\_lng,y=start\_lat,color=rideable\_type),size=2) + coord\_fixed(0.8) + labs(title = "Most used routes by Casual riders",x=NULL,y=NULL) + theme(legend.position = "none")  Coordinate system already present. Adding new coordinate system, which will replace the existing one.  Warning message:  Removed 9 rows containing missing values or values outside the scale range (`geom\_point()`).    > ggmap(chicago\_map,darken = c(0.1,"white")) + geom\_point(member\_riders,mapping = aes(x=start\_lng,y=start\_lat,color=rideable\_type),size=2) + coord\_fixed(0.8) + labs(title = "Most used routes by Member riders",x=NULL,y=NULL) + theme(legend.position = "none")  Coordinate system already present. Adding new coordinate system, which will replace the  existing one.  Warning message:  Removed 46 rows containing missing values or values outside the scale range  (`geom\_point()`).  A map of a city with red dots |

The first visualization shows the overall trend of ride\_length (ride duration) of casual riders and annual members for all months of 2023 year. As you can see, casual riders had a higher figure of ride duration throughout a year than annual members. In addition, casual members had a fluctuating trend over a year and reached its peak in May, June, July, August and September. On the other hand, annual members had a stable trend throughout the year and lower ride\_length (ride duration) than casual ones.

Top three recommendations based on analysis

1. Educate casual riders on the perks of memberships as a lack of knowledge can be the reason for the long usage by casual riders
2. An ad campaign at the popular start stations for casual riders can increase engagement or interest in memberships
3. A campaign to increase the usage by members especially in the less popular months could increase the number of trips. This campaign should provide benefits or rewards for continued usage with a focus on the less popular months.