Bike Sharing Capstone

Thang Huynh

2022-11-12

I. Ask

##Guiding Question 1.What is is problem you are trying to solve? Help company convert casual riders to annual members. 2.How can your insights drive business decision?
Insights backed by data to indicate how different casual riders to annual members, which helps launch marketing campaign to convert more casual riders to annual members.

##Key Task 1. Identify business task 2. Consider key stakeholders

##Deliverable A clear statement of the business task: Find the difference between casual riders and annual members and how marketing can help increase numbers of annual members

II. Prepare

##Guiding questions ● Where is your data located? On my laptop. This is my very first project in R and just working on it as a practise to wrap up what I have studied so far in Google Data Analytics, so I will got publish it. ● How is the data organized? In R ● Are there issues with bias or credibility in this data? Does your data ROCCC? There are no issues with this dataset since it comes from a reliable source. Data is ROCCC because it is reliable, original, comprehensive, current and cited ● How are you addressing licensing, privacy, security, and accessibility? Company has their own license over the dataset and this dataset does not contain any personal or sensitive information about clients. ● How did you verify the data's integrity? All files appear to be consistent with values in each column with correct data type. ● How does it help you answer your question? It may help indicate the difference about each group of riders. ● Are there any problems with the data More information about the riders could be useful

##Key tasks 1. Download data and store it appropriately. -> Done 2. Identify how it's organized.-> Done 3. Sort and filter the data.-> Done 4. Determine the credibility of the data.-> Done ##Deliverable A description of all data sources used

III. Process

Guiding questions

● What tools are you choosing and why? I choose R since I want to work on this programming skill ● Have you ensured your data's integrity? Done ● What steps have you taken to ensure that your data is clean? Remove duplicate, remove empty rows, add column to extract started date and started hour ● How can you verify that your data is clean and ready to analyze? ● Have you documented your cleaning process so you can review and share those results? Yes, I did note down any steps taken to clean data for future reference. ## 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 ### Documentation of any cleaning or manipulation of data

```
#tinytex::install_tinytex()
```

```
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 4.2.2
```

```
## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.2 ──
## ✓ ggplot2 3.4.0      ✓ purrr   0.3.5
## ✓ tibble  3.1.8      ✓ dplyr   1.0.10
## ✓ tidyr   1.2.1      ✓ stringr 1.4.1
## ✓ readr   2.1.3      ✓ forcats 0.5.2
```

```
## Warning: package 'ggplot2' was built under R version 4.2.2
```

```
## Warning: package 'readr' was built under R version 4.2.2
```

```
## Warning: package 'purrr' was built under R version 4.2.2
```

```
## Warning: package 'dplyr' was built under R version 4.2.2
```

```
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✕ dplyr::filter() masks stats::filter()
## ✕ dplyr::lag()    masks stats::lag()
```

```
#install.packages("dplyr")
library(dplyr)
```

```
#install.packages("readr")
library(readr)
```

Combining 12 csv files into 1 single file

```
original_csv <- list.files(path = "C:/Users/thang/OneDrive/Desktop/Data Science/Google Data Anal
ytics/Capstone/Bike Dataset/Data 12 months", recursive = TRUE, full.names = TRUE)
merged_bikedata <- do.call(rbind, lapply(original_csv, read.csv))
```

```
#head(merged_bikedata)
```

IV. Data Cleaning

1.Remove Duplicates. The idea here is to remove any duplicate rows for column with unique values (other columns with non-unique values can have duplicates)

```
nodup_bikedata <- merged_bikedata
  #unique(merged_bikedata)
```

###There seems to be no duplicate for ride_id because there are the same observations in nodup_bikedata ## 2. Remove any empty rows

```
noempty_bikedata <- nodup_bikedata[rowSums(is.na(nodup_bikedata)) != ncol(nodup_bikedata), ]
```

There is no empty rows since it shows the same obs

3. To replace blank with NA data

```
noblank_bikedata <- noempty_bikedata
noblank_bikedata[noblank_bikedata == ""] <- "NA"
```

4. Add a new column to calculate time used bike in minutes

```
noblank_bikedata <- noblank_bikedata %>%
mutate(noblank_bikedata, time_used = as.numeric(difftime(ended_at, started_at, units = "mins")))
```

5. Remove any rows with time_used <0 (Not make sense if time used is a negative value). It also means that any other values in other columns that have invalid value will also be removed.

```
noblank_bikedata<- filter(noblank_bikedata, time_used >= 0)
```

6. Remove any irrelevant columns (any columns with

longitude and lattitude information)

```
noblank_bikedata <-noblank_bikedata %>%
  select(-(start_lat:end_lng))
```

7. Rename column for easier readability (member_casual to member_type)

```
colnames(noblank_bikedata)[9] = "member_type"
```

V. Analysing Data

Guiding questions ● How should you organize your data to perform analysis on it? I removed any unrelated columns and extract information from started_at column to add new columns which show separate and meaningful values for analyzing data ● Has your data been properly formatted? Yes, I already cleaned data in previous step and now it is ready for analyzing ● What surprises did you discover in the data? It is surprising to see that there are more annual members than casual members but the average time spent from casual members are double those from annual members. ● What trends or relationships did you find in the data? - In terms of month, summer is the peak time for both group of users but while annual members are pretty consistent throughout the year, the number of casual members surge dramatically when it gets closer to the summer months. - In terms of weekday, annual members are also consistent throughout the week but casual members increase significantly and surpass number of annual members for 2 days in the weekend. - In terms of hour in a day, annual members start using bike earlier than casual members but both groups share a similar patterns reaching their own peak at around 5pm. ● How will these insights help answer your business questions? - This would help understand the habits of each group and useful for marketing campaign to convert more casual members to annual members. 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 A summary of your analysis

1. Extracting Weekdate from datetime value

```
library(lubridate)
```

```
## Loading required package: timechange
```

```
## Warning: package 'timechange' was built under R version 4.2.2
```

```
##
## Attaching package: 'lubridate'
```

```
## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union
```

Add column “started_weekday” to extract date from started_at column. This new column will be used later to detect the pattern of casual and annual members.

https://stackoverflow.com/questions/37545082/extract-weekdays-from-a-set-of-dates-in-r (https://stackoverflow.com/questions/37545082/extract-weekdays-from-a-set-of-dates-in-r)

```
noblank_bikedata$weekday <- wday(noblank_bikedata$started_at, label = TRUE)
```

Add column “started_hour” to extract hour from started_at column. This new column will be used to detect the pattern of casual and annual members

```
noblank_bikedata <- noblank_bikedata %>%
  mutate(noblank_bikedata, started_hours = hour(started_at))
```

Add column “started_month” to extract month from started_at column. This new column will be used to detect the pattern of casual and annual members

```
noblank_bikedata <- noblank_bikedata %>%
  mutate(noblank_bikedata, started_month= month(started_at))
```

Format month number to month string

```
noblank_bikedata$started_month<- month.abb[noblank_bikedata$started_month]
```

Saving clean dataset

```
noblank_bikedata %>%
  write.csv("clean_bikedata.csv")
```

Change dataset name for easier use

```
bikedata <- noblank_bikedata
```

Having an overall look at dataset

```
summary(bikedata)
```

```
##    ride_id          rideable_type       started_at          ended_at
##  Length:5594916     Length:5594916     Length:5594916     Length:5594916
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##  start_station_name start_station_id   end_station_name   end_station_id
##  Length:5594916     Length:5594916     Length:5594916     Length:5594916
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##  member_type          time_used        weekday      started_hours
##  Length:5594916     Min.   :    0.00   Sun:857221   Min.   : 0.00
##  Class :character   1st Qu.:    6.75   Mon:702577   1st Qu.:11.00
##  Mode  :character   Median :   12.00   Tue:739897   Median :15.00
##                     Mean   :   21.94   Wed:756104   Mean   :14.24
##                     3rd Qu.:   21.78   Thu:737584   3rd Qu.:18.00
##                     Max.   :55944.15   Fri:810498   Max.   :23.00
##                                        Sat:991035
##  started_month
##  Length:5594916
##  Class :character
##  Mode  :character
##
##
##
##
```

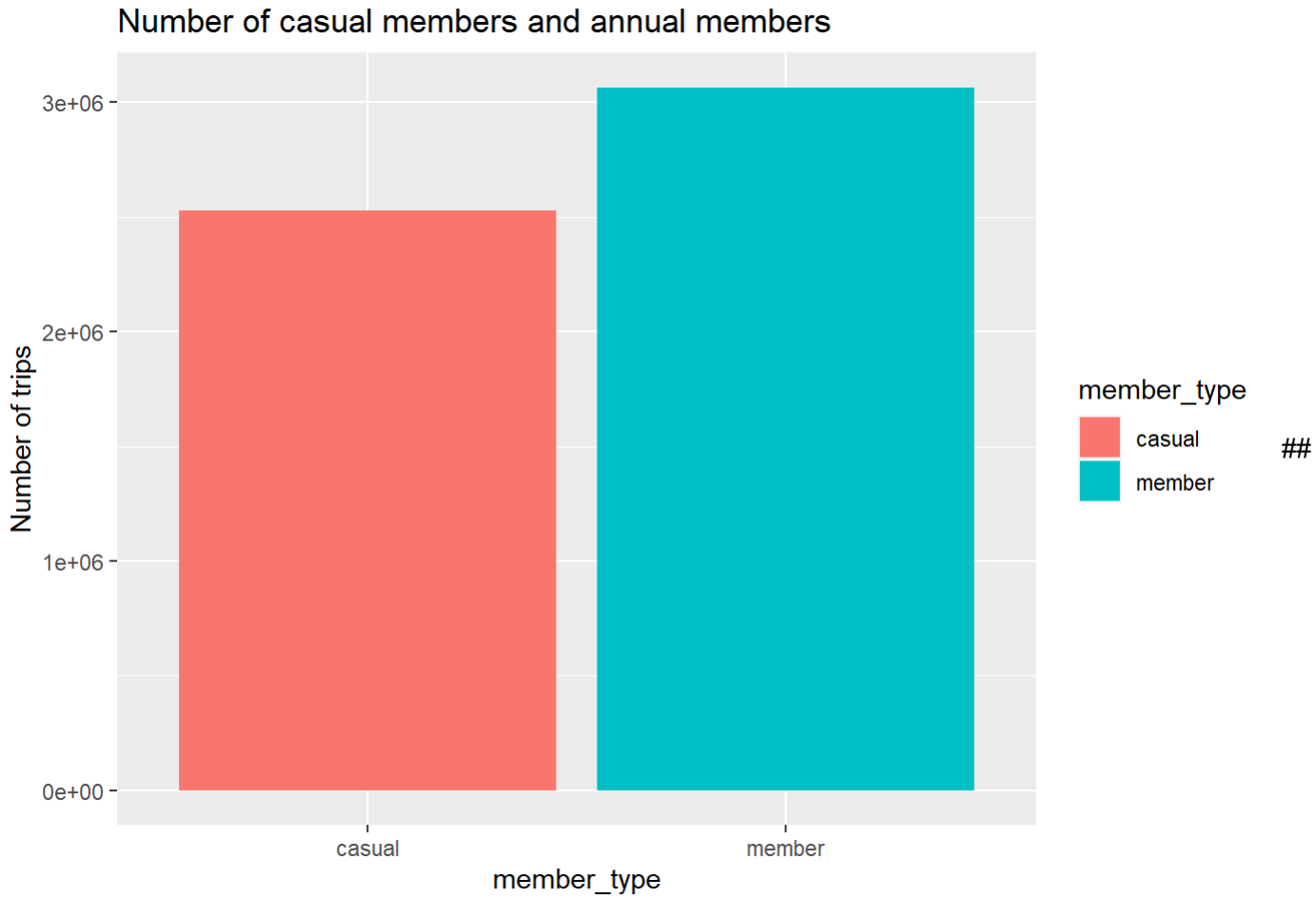
```
library(ggplot2)
```

Comparing total number of trips and their percentage classified by casual members and annual members

```
bikedata %>%
  group_by(member_type) %>%
  summarise(number_of_trips=n()) %>%
  mutate(percent = (number_of_trips/sum(number_of_trips)*100))
```

```
## # A tibble: 2 × 3
##   member_type number_of_trips percent
##   <chr>                 <int>   <dbl>
## 1 casual              2528946    45.2
## 2 member              3065970    54.8
```

```
ggplot(data = bikedata) +
  geom_bar(mapping = aes(x = member_type, fill= member_type)) +
  labs(title ="Number of casual members and annual members", y= "Number of trips")
```



We can see from that graph that there are more annual members and annual members takes up around 10 percent more than casual members.

We can see from this chart that both type of members use bike sharing service pretty consistently throughout weekdays, except for the weekend when we witness a rise in number of riders. It is also interesting to see that 2 days on weekend are the only days when there are more casual members

because throughout weekdays, there are more annual members. It makes sense since casual members may dont usually use bike sharing on a daily basis and they tend to ride bikes for a particular purpose on the weekend(either exercise or recreational…)

##https://stackoverflow.com/questions/36020146/how-can-i-order-the-months-chronologically-in-ggplot2-short-of-writing-the-month (https://stackoverflow.com/questions/36020146/how-can-i-order-the-months-chronologically-in-ggplot2-short-of-writing-the-month) (Getting this source for line scale_x_discrete to sort months in order)

```
ggplot(data = bikedata) +
  geom_bar(mapping = aes(x=started_month, fill=member_type)) +
  scale_x_discrete(limits = month.abb) +
  labs(title = "Casual members vs annual members during months", y="Number of trips")
```

This bar charts shows us that the period from June to early October is the peak time for the number of users. It is also within this period that we can see casual members soar significanly and somehow exceed those of annual members. For the rest of the months, the amound of annual members are significanly higher than casual members. It is interesting to see that the closer to summer, the higher proportion of casual members compared to

annual members. It makes sense since casual members may tend to use bike sharing more in the summer when they have more free time with their friends and family while annual members engage in the service pretty consistenly throughout the year.

Creating this tibble to see the distribution of trips in terms of weekday and member type:

```
library(scales)
```

```
##
## Attaching package: 'scales'
```

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

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

```
bikedata_1 <- bikedata %>%
  group_by(weekday,member_type) %>%
  summarise(total_trips=n()) %>%
  mutate(percent=total_trips/sum(total_trips))
```

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

```
bikedata_1$percent <- percent(bikedata_1$percent, accuracy = 1)
bikedata_1
```

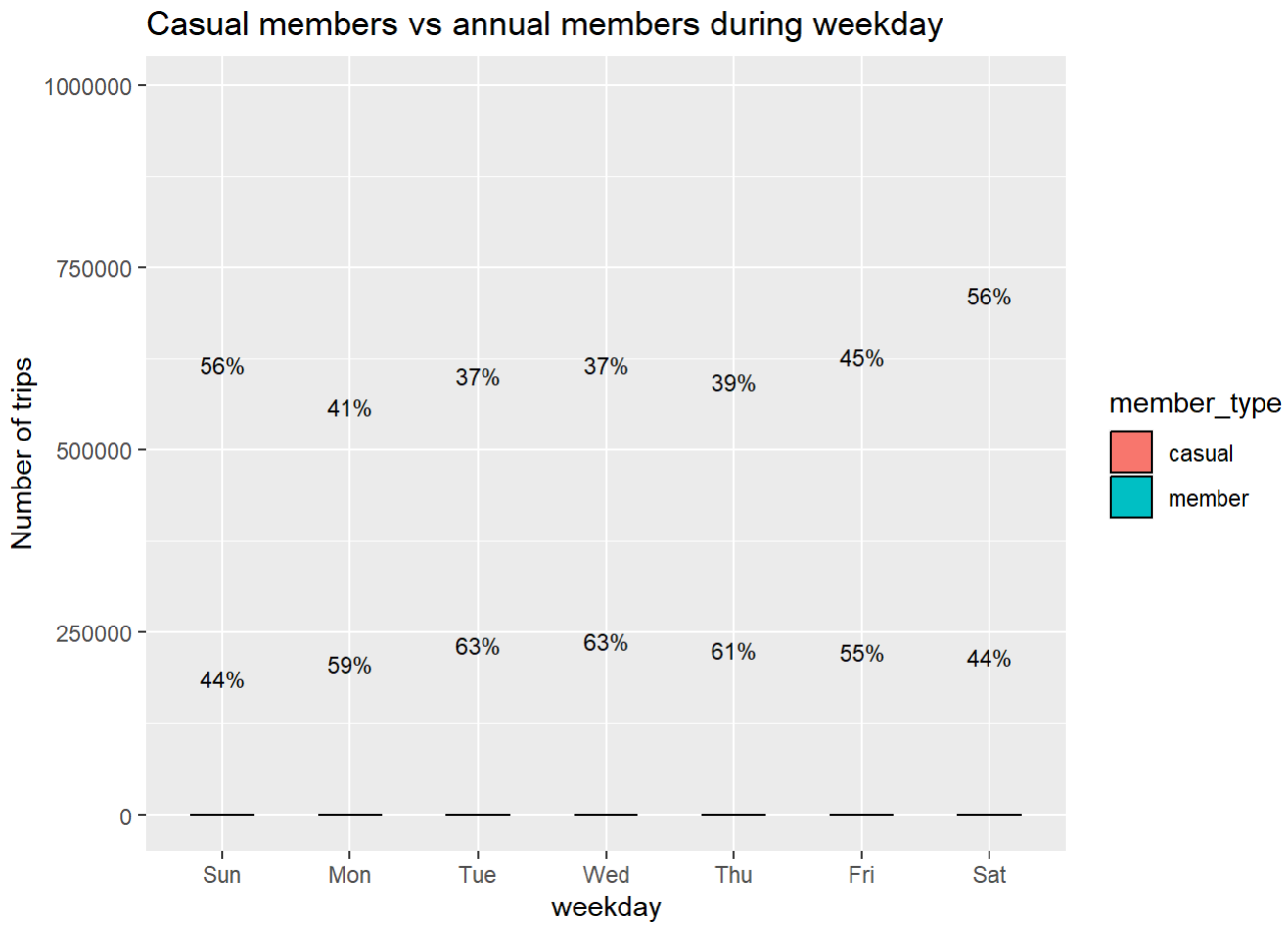
```
## # A tibble: 14 × 4
## # Groups:   weekday [7]
##    weekday member_type total_trips percent
##    <ord>   <chr>             <int> <chr>
##  1 Sun     casual           481104 56%
##  2 Sun     member           376117 44%
##  3 Mon     casual           286373 41%
##  4 Mon     member           416204 59%
##  5 Tue     casual           274388 37%
##  6 Tue     member           465509 63%
##  7 Wed     casual           278948 37%
##  8 Wed     member           477156 63%
##  9 Thu     casual           286064 39%
## 10 Thu     member           451520 61%
## 11 Fri     casual           364075 45%
## 12 Fri     member           446423 55%
## 13 Sat     casual           557994 56%
## 14 Sat     member           433041 44%
```

Creating a bar chart to show the distribution of each type of biker througout days in a week

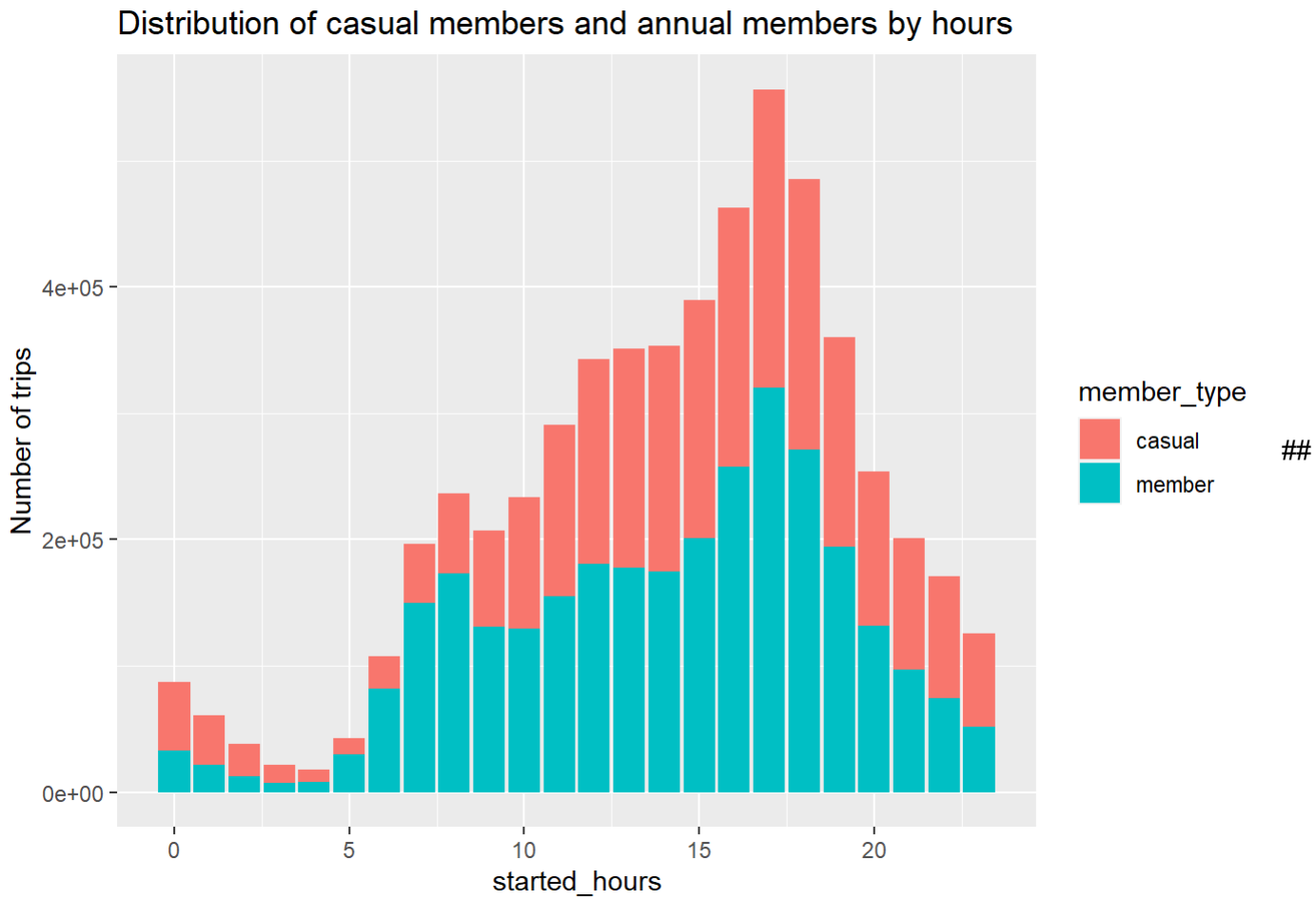
```
ggplot(data = bikedata) +
  geom_bar(mapping = aes(x=weekday, fill=member_type)) +
  labs(title = "Casual members vs annual members during weekday", y="Number of trips")
```

There is something wrong with this chart(still figuring out)

```
ggplot(bikedata_1, aes(x=weekday, y=total_trips, fill=member_type)) +
  geom_bar(position = "fill", stat = "identity",color='black',width=0.5) +
  geom_text(aes(label = paste0(percent)),
            position = position_stack(vjust = 0.5), size = 3) +
  labs(title = "Casual members vs annual members during weekday", y="Number of trips")
```




```
ggplot(data = bikedata) +
  geom_bar(mapping = aes(x=started_hours, fill=member_type)) +
  labs(title = "Distribution of casual members and annual members by hours", y="Number of trip
s")
```



From this chart, we can see that in general, there are more annual members than casual members within a day and it is consistent with what we found ealier when there are more annual members than casual members.A closer look reveals that annual members starts using bike sharing earlier since there is a high jump of users at round 7am and the number stay pretty consistently until reaching their peak at 5pm. On the other hand, casual members start using bike sharing later in a day when the number starts to rise at 10am and pretty much follow the similar pattern from annual members.

Fitler only casual member. This will be useful to launch marketing campaign for only this group.

```
casual_bikedata <- bikedata %>%
  filter(member_type=="casual")
```

Arrange top start_station

```
top_start_station <- casual_bikedata %>%
  group_by(start_station_name) %>%
  summarise(total=n())
  arrange(top_start_station, desc(total))
```

```
## # A tibble: 842 × 2
##    start_station_name         total
##    <chr>                      <int>
##  1 NA                        317679
##  2 Streeter Dr & Grand Ave    66359
##  3 Millennium Park            33590
##  4 Michigan Ave & Oak St      29780
##  5 Shedd Aquarium             23251
##  6 Theater on the Lake        21351
##  7 Wells St & Concord Ln      19891
##  8 Lake Shore Dr & Monroe St  19617
##  9 Clark St & Lincoln Ave     17033
## 10 Wells St & Elm St          16668
## # … with 832 more rows
```

Arrange top end_station

```
top_end_station <- casual_bikedata %>%
  group_by(end_station_name) %>%
  summarise(total=n())
arrange(top_end_station, desc(total))
```

```
## # A tibble: 841 × 2
##    end_station_name            total
##    <chr>                       <int>
##  1 NA                         365076
##  2 Streeter Dr & Grand Ave     68676
##  3 Millennium Park             34565
##  4 Michigan Ave & Oak St       31192
##  5 Theater on the Lake         22760
##  6 Shedd Aquarium              21564
##  7 Wells St & Concord Ln       19457
##  8 Lake Shore Dr & Monroe St   18264
##  9 Lake Shore Dr & North Blvd  17439
## 10 Clark St & Lincoln Ave      17180
## # … with 831 more rows
```

We can see from these two top stations the list of stations that are used most by casual members. These routes can be used as targeting locations for digital marketing campaigns. Launching promotion

for annual membership package will help encourage casual members to convert to annual members to save money.

Comparing casual members vs annual members in terms of the total riding time and percentage.

```
bikedata %>%
  group_by(member_type) %>%
  summarise(ridetime= sum(time_used)) %>%
  mutate(percent = (ridetime/sum(ridetime)*100))
```

```
## # A tibble: 2 × 3
##   member_type  ridetime percent
##   <chr>           <dbl>   <dbl>
## 1 casual      80934024.    65.9
## 2 member      41802332.    34.1
```

Comparing average riding time of casual members and annual members.

```
bikedata %>%
  group_by(member_type) %>%
  summarise(average=mean(time_used)) %>%
  mutate(percent = (average/sum(average)*100))
```

```
## # A tibble: 2 × 3
##   member_type average percent
##   <chr>         <dbl>   <dbl>
## 1 casual         32.0    70.1
## 2 member         13.6    29.9
```

Ridetime number indicates that ride time of casual members are double those of annual members. This is interesting because we found earlier that there are more annual members than casual members. Looking at the average number, it shows that on average casual members also spend more than 2

times the amount compared to time spent from annual members.It also means that casual members drive longer distances than annual members.

This finding leaves lot of place for the conversion since it shows that although there are more annual members, the amount each casual member spent for bike sharing service are much more. Marketing team can take advantage of this statistics point to convincing casual members that by upgrading to annual members, they are actually saving more money.

Comparing proportion of three types of bike used by customers.

```
bikedata %>%
  group_by(rideable_type) %>%
  summarise(total=n())  %>%
  mutate(percent=(total/sum(total)*100))
```

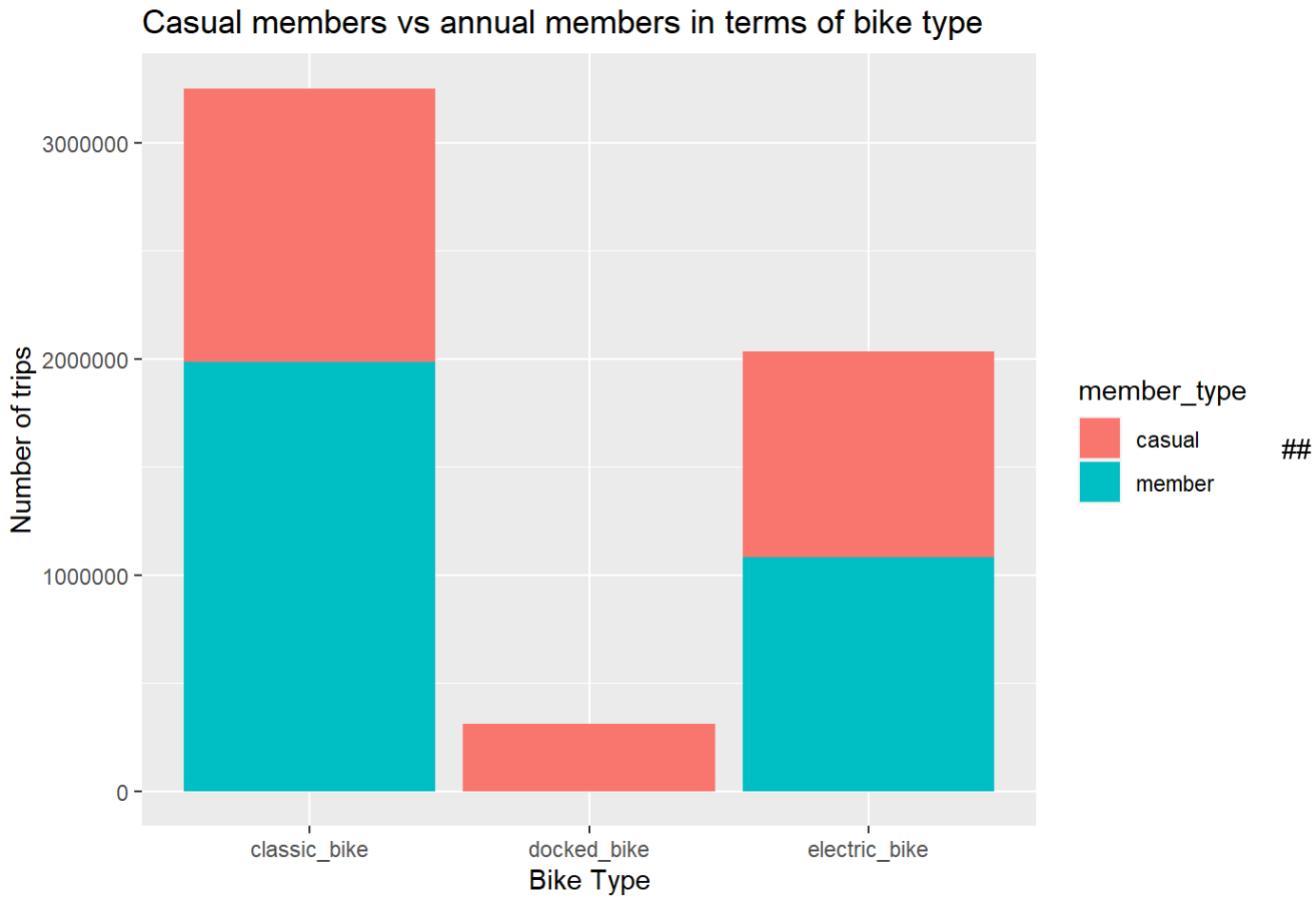
```
## # A tibble: 3 × 3
##   rideable_type   total percent
##   <chr>           <int>   <dbl>
## 1 classic_bike  3250943   58.1
## 2 docked_bike    312338    5.58
## 3 electric_bike 2031635   36.3
```

We can see from this tibble that while classic bike takes up largest porportion from all bike types with 58%, docked bike only takes a small part of 5.5%

This chunk is to prevent exponential numbers

```
options(scipen=999)
```

```
ggplot(data=bikedata) +
  geom_bar(mapping = aes(x=rideable_type, fill=member_type)) +
  labs(title="Casual members vs annual members in terms of bike type",x= "Bike Type",y= "Number
of trips")
```



From this graph, we can see that in terms of the most popular bike type which is casual bike, annual members takes 1.5 times more than the amount of casual members. In terms of electric bike, annual members also exceed the casual members though the difference is not that significant. It is interesing to see that there seems to be no annual members riding docked bike.

VI. Share

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 am able to compare the two groups based on month, weekday and hours. ● What story does your data tell? - The idea of my story based on data is to point out the different habits of using bike sharing service from two groups of users. ● How do your findings relate to your original question? - The comparison between two groups of bikers strongly answers the original question ” Find the key differences between casual and annual riders”. ● Who is your audience? What is the best way to communicate with them? There are 3 group of audience: - Lily Moreno: director of marketing - Cyclistic marketing analytics team - Cyclistic executive team Best way to communicate with them is through either powerpoint or google slides ● Can data visualization help you share your findings? - Yes it definitely helps to share my finding since there are quite a lot of comparisons with distinctive features to highligh the difference of using bike between

the 2 groups. ● Is your presentation accessible to your audience? - Yes it will be available through google slides Key tasks 1. Determine the best way to share your findings. 2. Create effective data visualizations. 3. Present your findings. 4. Ensure your work is accessible. Deliverable Supporting visualizations and key findings

VII. Act

Now that you have finished creating your visualizations, act on your findings. Prepare the deliverables Morena asked you to create, including the three top recommendations based on your analysis. Use the following Case Study Roadmap as a guide: Case Study Roadmap - Act Guiding questions ● What is your final conclusion based on your analysis? ● How could your team and business apply your insights? ● What next steps would you or your stakeholders take based on your findings? ● Is there additional data you could use to expand on your findings? Key tasks 1. Create your portfolio. 2. Add your case study. 3. Practice presenting your case study to a friend or family member. Deliverable Your top three recommendations based on your analysis Follow these steps: 1. If you do not have one already, create an online portfolio. (Use Creating an Interactive Portfolio with Google Sites or Build a Portfolio with Google Sites.) 2. Consider how you want to feature your case study in your portfolio. 3. Upload or link your case study findings to your portfolio. 4. Write a brief paragraph describing the case study, your process, and your discoveries. 5. Add the paragraph to introduce your case study in your portfolio