Case study: How does a bike-share navigate speedy success?

The marketing department of Cyclistic is looking for ways to grow the company and its services, there are three fundamental questions:

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?**

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 geo-tracked 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 solid opportunity 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 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.

As a Junior Data analyst from the Marketing department, I have conducted the data analysis to answer these questions, based on the Google Data Analysis Process, which consists of 6 phases:

1) Ask
2) Prepare
3) Process
4) Analyze
5) Share
6) Act

This document presents the analysis process and the results, going through each of the phases.

Phase 1 ASK

In this phase, the main objective is to **define the problem or question you want to address using data**. It's crucial to establish clear goals for your analysis right from the start. This ensures you're collecting the right data and setting yourself up for success in the later stages.

These 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?

Summary Table – Ask
Guiding questions • **What is the problem you are trying to solve?** Conduct Data Analysis with the information provided by the company to understand the differences between casual riders and member riders and determine the feasibility and strategy to convert casual riders into members • **How can your insights drive business decisions?** The company will use the information from this report and its recommendations to decide if they will pursue this initiative and proceed to an implementation plan, defining resources, time and funding.
Key tasks • **Identify the business task** **Definition of the Objective of the Analysis, these Critical Questions:** 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? **Definition of the Final Deliverables:** 1. Statement of the business task 2. Description of all data sources used 3. Documentation of cleaning and manipulation of data 4. Summary of your analysis 5. Supporting visualizations and key findings 6. Top three recommendations based on your analysis • **Consider key stakeholders** **Cyclistic executive team: D**etail-oriented executive team. They will decide

whether to approve the recommended marketing program.

Lily Moreno: Director of Marketing responsible for the development of campaigns and initiatives to promote the bike-share program. These may include email, social media, and other channels.

Cyclistic marketing analytics team: A team of data analysts who are responsible for collecting, analyzing, and reporting data that helps guide Cyclistic marketing strategy.

Deliverable

- A clear statement of the business task

PHASE 2 Prepare

In this phase, we **gathered and organized the data** necessary to answer the question defined in the "ask" phase. The steps followed were: Data Collection from it's source, completeness check, and consistent formatting.

The Cyclistic's historical trip data was used for Analysis and Trend identification. The Information is located in:
previous 12 months of Cyclistic trip data here.
https://divvy-tripdata.s3.amazonaws.com/index.html

Note: The datasets have a different name because Cyclistic is a fictional company. For the purposes of this case study, the datasets enabled answering the business questions.

The data has been made available by Motivate International Inc. under this license.

But note that data-privacy issues prohibit using riders' personally identifiable information. Such as connecting pass purchases to credit card numbers to determine if casual riders live in the Cyclistic service area or if they have purchased multiple single passes.

Steps followed to download the information:

1) Copied the list of files from the website into an excel spreadsheet
2) Manually downloaded each of the zip files
3) Extracted the zip files to a sub folder using PowerShell:

```
# Create the subfolder
New-Item -ItemType Directory -Path .\unzipped_files

# Get all .zip files in the current directory
Get-ChildItem -Filter *.zip | ForEach-Object {
    # Expand-Archive extracts the contents of the zip file to the specified
directory
    Expand-Archive $_.FullName -DestinationPath .\unzipped_files
}
```

4) Used PowerShell to generate a text file with the list of downloaded zip files:
   Dir > filelist.txt

5) In excel I used the following formulas to clean up the data:
   Extracting the file name: =RIGHT(A16,LEN(A16)-FIND(".zip",A16)+22)
   Copying and pasting as text values
   Trimming the file names to eliminate blank characters: =TRIM(C9)
   Trim again to eliminate the file size: =TRIM(RIGHT(F6,LEN(F6)-SEARCH(" ",F6)))
   Sort if alphabetically

6) Compared the column of the downloaded files with the column of files from the website, used a formula in the next column ie =D4=C4 which gave me true if the filenames were exactly the same and I could easily verify this for all the files.

Downloaded files	File List on website	
202004-divvy-tripdata.zip	202004-divvy-tripdata.zip	TRUE
202005-divvy-tripdata.zip	202005-divvy-tripdata.zip	TRUE
202006-divvy-tripdata.zip	202006-divvy-tripdata.zip	TRUE
202007-divvy-tripdata.zip	202007-divvy-tripdata.zip	TRUE
202008-divvy-tripdata.zip	202008-divvy-tripdata.zip	TRUE
202009-divvy-tripdata.zip	202009-divvy-tripdata.zip	TRUE
202010-divvy-tripdata.zip	202010-divvy-tripdata.zip	TRUE
202011-divvy-tripdata.zip	202011-divvy-tripdata.zip	TRUE
202012-divvy-tripdata.zip	202012-divvy-tripdata.zip	TRUE
202101-divvy-tripdata.zip	202101-divvy-tripdata.zip	TRUE
202102-divvy-tripdata.zip	202102-divvy-tripdata.zip	TRUE
202103-divvy-tripdata.zip	202103-divvy-tripdata.zip	TRUE
202104-divvy-tripdata.zip	202104-divvy-tripdata.zip	TRUE
202105-divvy-tripdata.zip	202105-divvy-tripdata.zip	TRUE
202106-divvy-tripdata.zip	202106-divvy-tripdata.zip	TRUE
202107-divvy-tripdata.zip	202107-divvy-tripdata.zip	TRUE
202108-divvy-tripdata.zip	202108-divvy-tripdata.zip	TRUE
202109-divvy-tripdata.zip	202109-divvy-tripdata.zip	TRUE
202110-divvy-tripdata.zip	202110-divvy-tripdata.zip	TRUE
202111-divvy-tripdata.zip	202111-divvy-tripdata.zip	TRUE
202112-divvy-tripdata.zip	202112-divvy-tripdata.zip	TRUE

Summary Table – Prepare

Guiding questions

- Where is your data located?
  The data is located in Amazon AWS and is available for download

- How is the data organized?
  From April 2020 to September 2024 - The data is organized in files by month of the year
  Format of the files: YYYYMM-divvy-tripdata.zip

  From 2013 to March 2020 – the data is organized in files by quarter of the year
  Format of the files: Divvy_Trips_YYYY_QXQX.zip or Divvy_Trips_YYYY_QX.zip

- Are there issues with bias or credibility in this data? Does your data ROCCC?
  The data complies with the recommended rules ROCCC:

  Reliable: files passed integrity check and content check
  Original: original source of data provided
  Comprehensive: twelve months of previous rides which is sufficient for the scope of the analysis
  Current: 12 previous months
  Cited: data has been made available by Motivate International Inc. under license

  Identify Potential Sources of Bias: Consider how the data was collected, who collected it, and if any groups or perspectives might be excluded.
  Data has been collected electronically there is no human bias involved in data collection

  Look for Inconsistencies: Analyze the data for any contradictions, outliers, or patterns that seem unusual or unexpected. Out of 5,854,527 records, a total of 25 outlier values were found in latitude and longitude values for initial and end trip coordinates which can be caused

by hardware failure or lost/stolen equipment. The outlier values will not be included in the Analysis and they represent less that **0.0003%**

A Statistical analysis was made on each of the files to obtain Min, Max, Mean, Median, Mode and Standard Deviation of the initial and final coordinates of the trip. Once the few outlier values were removed, the analysis revealed that the values of the mean and the median are very similar, revealing a symmetrical distribution. The standard deviation value for the trip coordinates is les than 0.5, which means that the data is clustered tightly around the mean

Evaluate the Sample Size and Representation: A small or non-representative sample can lead to skewed results. Ensure the data represents the population you're studying.
The data being analyzed is the total rides for the last 12 month, this adds confidence in the analysis and the findings the only values excluded were the outliers.

Cross-Reference with Other Sources: Don't rely on a single source. Compare the data with other reputable sources to confirm its accuracy.
The maximum and minimum values for the coordinates in the data set were compared against coordinates from google maps, to confirm geographical accuracy of the data

Consider the Data's Age: Outdated information might not reflect current realities. Make sure the data is recent and relevant to your analysis.
The analysis will consider the last 12 months of rides, which makes the information recent and relevant for analysis of trends

- How are you addressing licensing, privacy, security, and accessibility?
  The data has been made available by Motivate International Inc. under this license.
  The data in the files does not include personally identifiable information.

- How did you verify the data's integrity? The data was checked for integrity by reviewing each of the 12 csv files to check for: Structural Issues, Validate Data Types and Ranges and Count of Rows and Columns. An integrity Check summary will be included in the data

- How does it help you answer your question?
  The validation process warrantees that we are using the right data for the analysis and that our result will be accurate

- Are there any problems with the data? Empty values in rows

Key tasks

- Download data and store it appropriately. – **Done** Downloaded to personal computer
- Identify how it's organized. **Done**
- Sort and filter the data. **Done**
- Determine the credibility of the data. **Done**

Deliverable

- A description of all data sources used: Summary Table added to deliverables

PHASE 3 Process

In this Phase, the files were merged into a single file for c**leaning and transforming** the data you gathered in the prepare phase. This ensures your data is ready for accurate analysis.

Summary Table - Process

Guiding questions

- What tools are you choosing and why? R because of the volume of information and also because of the available libraries have functions to clean, process data and generate visuals.

- Have you ensured your data's integrity?

  Validation during Collection: Implement checks at the point of data entry to prevent incorrect or inconsistent data from being recorded in the first place. This could involve using dropdown menus for specific fields, setting data type restrictions, or requiring mandatory fields. Data type restrictions were set for ride start and end time
  Data Cleaning: Even with validation, errors can occur. Thoroughly clean your data before analysis. This includes handling missing values, removing duplicates, and addressing outliers. There were missing values for ride start and end date, those values were discarded since it was not possible to obtain ride length or day of the ride. The size of the dataset went from 5,854,520 to 3,528,667, 60.27% of the total records. The size of the sample is 60% and the error appeared randomly through out the files. Sampling was random and unbiased.

  Data Transformation: Transforming data in a consistent and documented manner is key. This might involve standardizing formats, converting units, or creating new variables. Be sure to document these transformations for clarity and reproducibility.
  All the transformations were documented and saved as R scripts and are part of the deliverables
  Summary of transformations:

  Merging Data:

```
merged_data <- lapply(csv_files, read.csv) %>%  bind_rows()
```

  Formatting columns as date fields:

```
 merged_data <- merged_data %>%
mutate(started_at = as.POSIXct(started_at, format = "%Y-%m-%d %H:%M"),
    ended_at = as.POSIXct(ended_at, format = "%Y-%m-%d %H:%M"))
```

  Adding a calculated field for trip length:

```
# Calculate the difference
diff <- as.numeric(difftime(merged_data$ended_at, merged_data$started_at, units = "secs"))

# Convert the difference to hours, minutes, and seconds
merged_data$ride_length <- sprintf("%02d:%02d:%02d", diff %/% 3600, (diff %% 3600) %/%
60, diff %% 60)
```

Data Backup and Recovery: Regularly back up your data to prevent loss due to errors or technical issues. Having a robust recovery plan ensures you can restore your data to a previous state if needed.
Data for the analysis is saved to the cloud to prevent loss from local hard drive

Documentation and Version Control: Maintain clear documentation of your data sources, cleaning procedures, and transformations. Version control systems can help track changes to your data and code, ensuring you're always working with the most accurate and up-to-date information.
Accurate documentation and version control is being kept for this analysis and this document along with the scripts are part of the deliverables.

- steps to ensure that your data is clean
  Data verification by histogram of ride length and bar char of ride length per day of week

- How can you verify that your data is clean and ready to analyze?
  The visualization of the histogram ride length values and ride of length per week helped verify that the data is clean and ready to analyze

- Documentation of cleaning process to review and share those results?
  The cleaning process was documented in the R scripts and the scripts and visuals will be included as part of the deliverables

Key tasks

- Check the data for errors. Done fixed missing values and data errors Done
- Choose your tools. Done
- Transform the data so you can work with it effectively. Done
- Document the cleaning process. Done

Deliverable

- Documentation of any cleaning or manipulation of data Added to Deliverables report

Steps followed:

1. Downloaded the previous 12 months of trip data.
Note: If you are planning on using Posit's RStudio, use the Divvy 2019 Q1 and Divvy 2020 Q1 datasets. Choosing other data might lead to errors because the data exceeds the memory available in the free plan.
2. Unzip the files.
3. Create a folder on your desktop or Drive to house the files. Use appropriate file-naming conventions.
4. Create subfolders for the .csv file and the .xls or Sheets file so that you have a copy of the original data. Move the downloaded files to the appropriate subfolder.
5. Follow these instructions for either Excel (a) or Google Sheets (b):
    a. Launch Excel, open each file, and choose to Save As an Excel Workbook file. Put

it in the subfolder you created for .xls files.

b. Open each .csv file in Google Sheets and save it to the appropriate subfolder.

6. Open your spreadsheet and create a column called **ride_length**. Calculate the length of each ride by subtracting the column **started_at** from the column **ended_at** (for example, =D2-C2) and format as HH:MM:SS using Format > Cells > Time > 37:30:55.
7. Create a column called **day_of_week**, and calculate the day of the week that each ride started using the **WEEKDAY** command (for example, **=WEEKDAY(C2,1)**) in each file. Format as General or as a number with no decimals, noting that 1 = Sunday and 7 = Saturday.
8. Proceed to the analyze step.

The following analysis and visuals helped better familiarize, and identify new approaches to answering the business questions.

1. 01 Script to join the CSV files.R

```
# Load necessary libraries
library(dplyr)
library(readr)
library(lubridate)

# Set the directory containing the CSV files
setwd("C:/clean")

# List all CSV files in the directory
csv_files <- list.files(pattern = "*.csv")

# Read and merge all CSV files into a single data frame
merged_data <- lapply(csv_files, read.csv) %>%
  bind_rows()

View(merged_data)

# Export the merged data frame to a new CSV file
write.csv(merged_data, "merged_data.csv", row.names = TRUE)
```

2. 02 Script to add trip length column.R

```
#Script to Calculate additional fields
# necessary libraries
library(dplyr)
library(readr)
library(lubridate)

setwd("C:/clean")

# Load the CSV file into a data frame
merged_data <- read_csv("merged_data.csv")
```

```
# Remove the 'ride_length' column
# merged_data <- merged_data %>% select(-ride_length) this wasn't used
in the end

# Convert started_at and ended_at columns to Date  datetime format
merged_data <- merged_data %>%
  mutate(started_at = as.POSIXct(started_at, format = "%Y-%m-%d
%H:%M"),
         ended_at = as.POSIXct(ended_at, format = "%Y-%m-%d %H:%M"))
head(merged_data)
str(merged_data)
View(merged_data)

# Assuming you already have a data frame 'df' with columns 'date1' and
'date2'

# Calculate the difference
diff <- as.numeric(difftime(merged_data$ended_at,
merged_data$started_at, units = "secs"))

# Convert the difference to hours, minutes, and seconds
merged_data$ride_length <- sprintf("%02d:%02d:%02d", diff %/% 3600,
(diff %% 3600) %/% 60, diff %% 60)


# Print the updated data frame
print(df)

write.csv(merged_data, "merged_data_trip_length.csv", row.names = TRUE)


# View the first few rows of the data frame
head(merged_data)
View(merged_data)
```

3. 03 Script to add Day of Week column.R

```
#Libraries
library(dplyr)
library(readr)
library(lubridate)

#Open file
setwd("C:/clean")
# Load the CSV file into a data frame
merged_data <- read_csv("merged_data_trip_length.csv")

problems(merged_data)
```

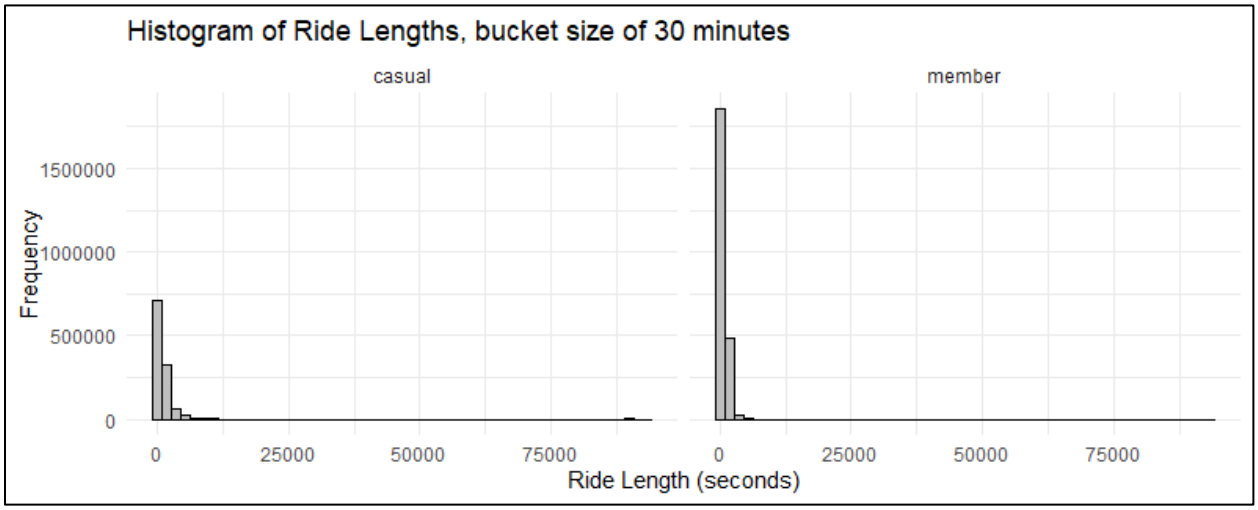
```
#Add Column

merged_data$ride_weekday <- weekdays(merged_data$started_at)

head(merged_data)
View(merged_data)

write.csv(merged_data, "merged_data_trip_length_day_of_week.csv",
row.names = TRUE)
```

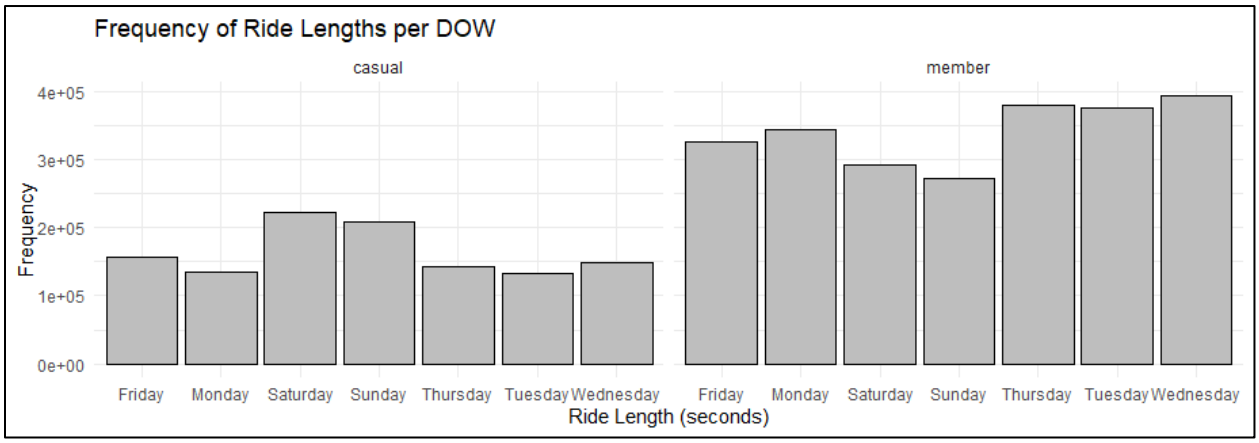
4. 04 Data Readiness Verification.R

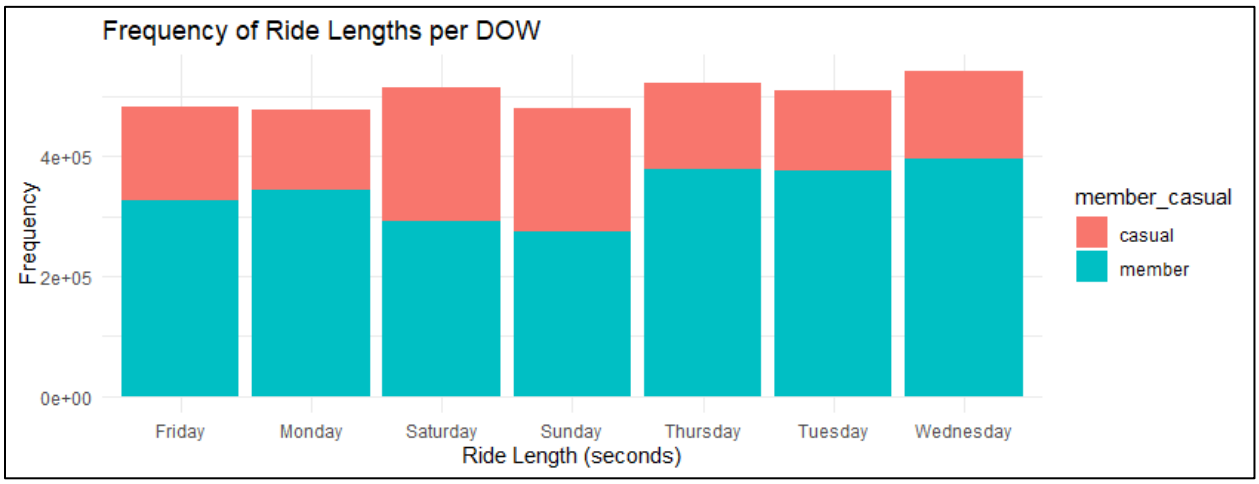
```
library(dplyr)
library(readr)
library(lubridate)
library(ggplot2)

#Open file
setwd("C:/clean")
# Load the CSV file into a data frame
merged_data <- read_csv("merged_data_trip_length_day_of_week.csv")

View(merged_data)
str(merged_data)

# Convert ride_length to numeric (seconds)
merged_data <- merged_data %>% mutate(ride_length_sec =
as.numeric(ride_length))



# 1 Create a histogram of the ride_length column
ggplot(merged_data, aes(x = ride_length_sec)) +
  geom_histogram(binwidth = 1800, fill = "grey", color = "black") +
  theme_minimal() +
  labs(title = "Histogram of Ride Lengths, bucket size of 30 minutes",
x = "Ride Length (seconds)", y = "Frequency") +
  facet_grid(~member_casual)

# 2 Frequency of Ride Lengths per DOW
ggplot(merged_data, aes(x = ride_weekday)) +
  geom_bar(binwidth = 1800, fill = "grey", color = "black") +
  theme_minimal() +
  labs(title = "Frequency of Ride Lengths per DOW", x = "Ride Length
(seconds)", y = "Frequency") +
  facet_grid(~member_casual)
```

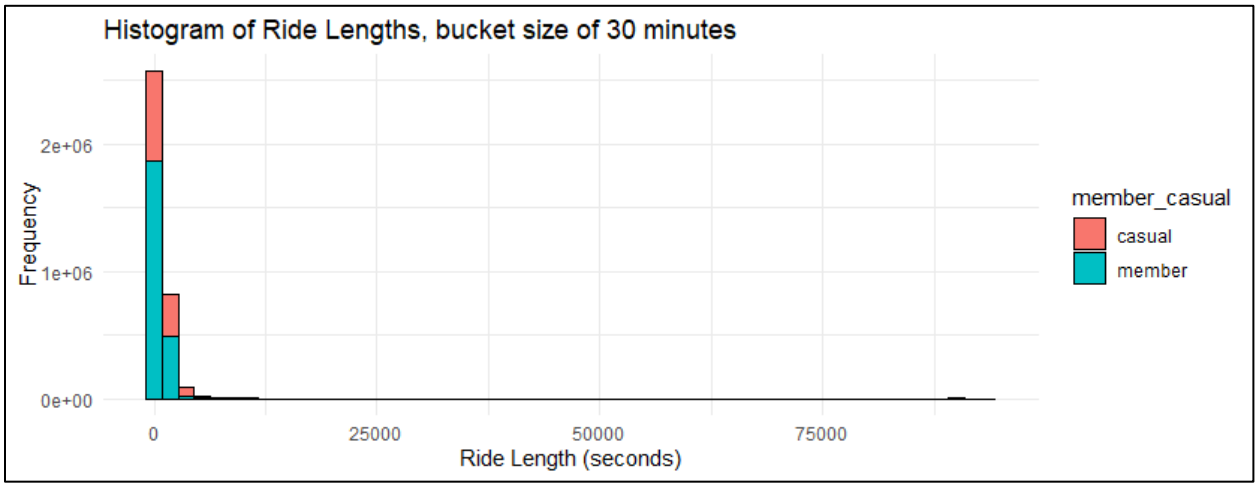
```
# 3
ggplot(merged_data, aes(x = ride_weekday, fill = member_casual)) +
  geom_bar(binwidth = 1800) +
  theme_minimal() +
  labs(title = "Frequency of Ride Lengths per DOW", x = "Ride Length
(seconds)", y = "Frequency")

# 4 Create a histogram of the ride_length column
ggplot(merged_data, aes(x = ride_length_sec, fill = member_casual)) +
  geom_histogram(binwidth = 1800,color = "black") +
  theme_minimal() +
  labs(title = "Histogram of Ride Lengths, bucket size of 30 minutes",
x = "Ride Length (seconds)", y = "Frequency")


# 5 Create a histogram of the ride_length column
ggplot(merged_data, aes(x = ride_length_sec, fill = ride_weekday)) +
  geom_histogram(binwidth = 1800, color = "black") +
  theme_minimal() +
  labs(title = "Histogram of Ride Lengths, bucket size of 30 minutes",
x = "Ride Length (seconds)", y = "Frequency") +
  facet_grid(~member_casual)



# sort datagram
# Sort the dataframe by the "Score" column in descending order

# Sort the dataframe by the "Score" column in descending order using
dplyr
merged_data_sorted <- merged_data %>% arrange(desc(ride_length_sec))
View(merged_data_sorted)

write.csv(merged_data,
"merged_data_trip_length_day_of_week_seconds_numeric.csv", row.names =
TRUE)
#Save the new dataframe with new column for seconds
```

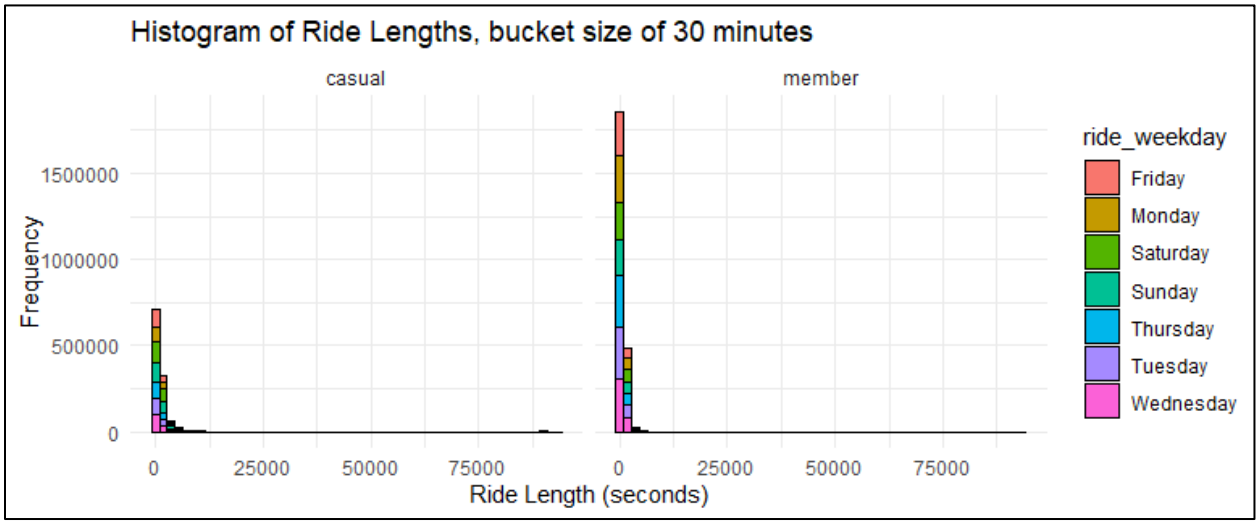
Data Processing - Visuals generated from R code:











Phase 4 Analyze

The objective here is to identify trends, patterns, and insights that can help answer your initial question. This often involves using various data analysis techniques and tools to understand the relationships within your data.

Now that your data is stored appropriately and has been prepared for analysis, start putting it to work. Use the following Case Study Roadmap as a guide:

Summary Table - Analyze
Guiding questions • Organization of the data to perform the analysis:

The records were organized as per the following process:

1) Consolidation of all records in single file
2) Removal or rows with empty values
3) Creation of new columns with calculated trip distance and breakdown of trip date by month and day
4) Removal of values where trip distance was negative or equal to zero, a few hundred entries when bikes were taken out of docks and checked for quality by the company or ride_length was negative
5) Sorting of the data by date

- Has your data been properly formatted?
  Yes, the format of ride length in seconds allows arithmetic and statistical calculations

- What surprises did you discover in the data?
  a) Empty values the number of records for analysis went from 5,854,520 to 3,528,667 and then cleaning incorrect values for distance reduced the dataset to 2,586,054
  b) Outliers in time and distance: there were trips where the duration was of several days and (this could be explained by customer keeping the bike several days and then returning it) The analysis of minimum and maximum distances from the trips revealed a large area over 150 miles. Need to dig deeper into trips long distance outlier cases to find the answer (out of the scope of this analysis)
  c) The fact that even though casuals perform less trip than members (expected), the duration of the trip for members is always less than the duration of the trip for casuals

- What trends or relationships did you find in the data?
  The following trends were discovered during the analysis:

  - Mean Ride Length for Casual Riders is 51.83% longer than Casual Ride Length
  - Average Ride Length for Casuals is at least 70% longer than Members on any given day.
  - On Sundays the ride length of Casuals is 100.72% higher than Members (highest difference)

  - MEAN Value of Ride Length for Members is 48.19% shorter than Casual riders
  - MEDIAN Value of Ride Length for Members is 30.76% shorter than Casual riders
  - MAXIMUM Value of Ride Length for Members and Casual Riders is virtually the same
  - MINIMUM Value of Ride Length for Members and Casual Riders is the same

  - **Casual Riders Trips last 51.83% longer**
  - **On weekends, casual rides are 20% less than members but the difference grows on weekdays, where the difference oscillates between 53% and 66%**

- How will these insights help answer your business questions?
  The analysis explain how the casual and member riders are different in duration and frequency over the whole week, this allows us to think on ways to engage the casual riders

Key tasks
• Aggregate your data so it's useful and accessible. Done • Organize and format your data. Done • Perform calculations. Done • Identify trends and relationships. Done
Deliverable A summary of your analysis has been included to deliverables: **Deliverable 04 A Analysis Summary** **Delivery 04 B Detailed Analysis**

Deliverable 04 Analysis Summary

Below is summary of steps and values obtained, the complete scripts and results are included in appendix 1 – Data Analysis with R **Delivery 04 B Detailed Analysis**

1) Combining the Data into a single file

Total Rows	Total Columns
3,528,667	20

2) Determine total of observations by casual and member

Rides/Percentage	Total Observations Casual	Total Observations Member	Total
Rides	1,145,263	2,383,404	3,528,667.00
Percentage	**32%**	**68%**	**100%**

3) Add Columns that list the date, month, day and year of each ride

4) Add a "ride_length" calculation to all trips (in seconds)

5) Removing Bad Data and redundant columns

6) Determine total of observations by casual and member

Rides/Percentage	Total Observations Casual	Total Observations Member	Total
Rides	820,689	1,765,365	2,586,054.00
Percentage	**32%**	**68%**	**100%**

7) Data Calculations

Calculation	Value In Seconds	Value In Minutes	Value In Hours

Mean value for ride length	952.88	15.88	0.26
Maximum ride length	90,600.00	1,510.00	25.17
Mode of Day_of_Week	WEDNESDAYS	WEDNESDAYS	WEDNESDAYS

Mean ride length for members and casuals

Member / Casual	Mean Value In Seconds	Mean Value In Minutes
Casual	1420.00	23.67
Member	736.00	12.27

Mean Ride Length for Casual Riders is 51.83% longer than Casual Ride Length
Casual Riders Trips last 51.83% longer

Calculate the average ride_length for members and casuals by day_of_week in minutes

Day	Members AVG Ride Length	Casuals avg Ride Length	% Casual is longer than member
Friday	11.8	22.4	89.83
Monday	11.8	22.9	94.07
Saturday	13.7	26.9	96.35
Sunday	13.9	27.9	100.72
Thursday	11.6	19.8	70.69
Tuesday	11.8	20.3	72.03
Wednesday	12.0	21.2	76.67
Average Ride Length for Casuals is at least 70% longer than Members on any given day, Sunday is the highest value of 100.72%			

Calculate the number of rides for users by day_of_week

Day	Member Total Rides	Casual Total Rides	% Casual perform less trips than members
Friday	237,995	110,606	53.53
Monday	261,675	98,368	62.41
Saturday	207,807	161,568	22.25

Sunday	198,359	152,451	23.14
Thursday	282,269	99,216	64.85
Tuesday	283,112	94,566	66.60
Wednesday	294,148	103,914	64.67
Total Rides	**1,765,365**	**820,689**	**53.51**

On the weekend, the casual rides are 20% less than members but the difference grows on weekdays, where the difference oscillates between 53% and 66%

Mean trip length value in minutes	15.88136 min.
Median trip length value in minutes	10 min.
Max trip length value in hours	25.16667 hrs.
Min trip length value in seconds	60 secs.

```
summary(merged_data$ride_length_sec) RESULT IN SECONDS
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    60.0   360.0   600.0   952.9  1020.0 90600.0

summary(merged_data$ride_length_sec/60) RESULT IN MINUTES
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.00    6.00   10.00   15.88   17.00 1510.00

summary(merged_data$ride_length_sec/3600) RESULT IN HOURS
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##  0.01667  0.10000  0.16667  0.26469  0.28333 25.16667
```

Compare members and casual members

```
aggregate(merged_data$ride_length ~ merged_data$member_casual, FUN = mean)
```

MEAN VALUE FOR RIDE LENGTH	
Rider Type	**Mean Value Ride Length Seconds**
Casual	1,420.15
Member	735.66
MEAN Value of Ride Length for Members is 48.19% shorter than Casual riders	

```
aggregate(merged_data$ride_length ~ merged_data$member_casual, FUN = median)
```

MEDIAN VALUE FOR RIDE LENGTH	
Rider Type	**Median Value Ride Length Seconds**
Casual	780.00
Member	540.00
MEDIAN Value of Ride Length for Members is 30.76% shorter than Casual riders	

```
aggregate(merged_data$ride_length ~ merged_data$member_casual, FUN = max)
```

MAXIMUM VALUE FOR RIDE LENGTH	
Rider Type	**Maximum Value Ride Length Seconds**
Casual	90,600.00
Member	89,880.00
MAXIMUM Value of Ride Length for Members and Casual Riders is virtually the same	

```
aggregate(merged_data$ride_length ~ merged_data$member_casual, FUN = min)
```

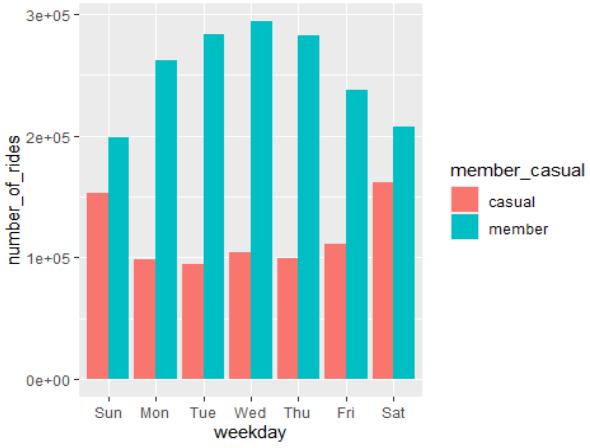
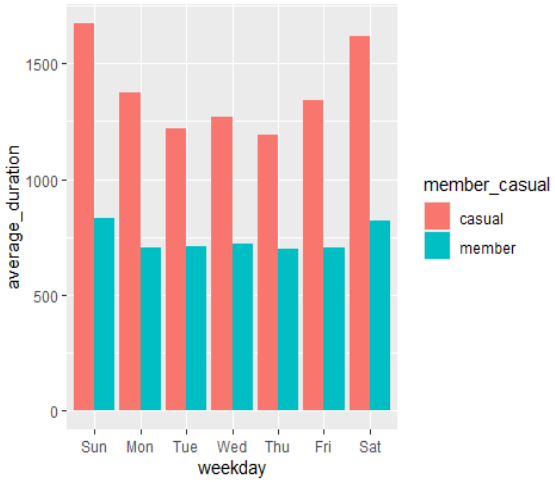
MINIMUM VALUE FOR RIDE LENGTH	
Rider Type	**Minimum Value Ride Length Seconds**
Casual	60.00
Member	60.00
MINIMUM Value of Ride Length for Members and Casual Riders is the same	

Order days of the week

```
merged_data$day_of_week <- ordered(merged_data$day_of_week, levels=c("Sunday"
, "Monday", "Tuesday", "Wednesday", "Thursday", "Friday", "Saturday"))
```

average ride time and ride length by each day for members vs casual users

Day	Casual Total Rides	Member Total Rides	Average Ride Time Secs Casual	Average Ride Time Secs Members
Sunday	152,451	198,359	1674.3165	832.022
Monday	98,368	261,675	1376.7453	705.487
Tuesday	94,566	283,112	1219.3249	710.7073
Wednesday	103,914	294,148	1270.0814	720.5954
Thursday	99,216	282,269	1189.1467	698.3024
Friday	110,606	237,995	1341.8147	705.7202
Saturday	161,568	207,807	1616.2734	822.0136
Total/Average	**820,689**	**1,765,365**	**1383.957557**	**742.1211286**

Number of Rides by Rider Type	Average Duration
	

Phase 5 Share

In the **share** phase, the focus is on **communicating your findings** to your audience in a clear and compelling way. This often involves creating visualizations, presentations, or reports that effectively convey the insights you've uncovered during the analysis.

Share

Now that you have performed your analysis and gained some insights into your data, create visualizations to share your findings. Moreno has reminded you that they should be sophisticated and polished in order to effectively communicate to the executive team. Use the following Case Study Roadmap as a guide:

Summary Table - Share
Guiding questions • Were you able to answer the question of how annual members and casual riders use Cyclistic bikes differently? Yes, the data analysis allows us to define the different duration and frequency of rides between members and casuals • What story does your data tell? The opportunity to convert casual riders into members • How do your findings relate to your original question? The findings answered the question on difference in use between casuals and members and also helps us know the best way to engage casuals into becoming members on benefits both economical and health related. This information can be addressed to casual members in a targeted digital media campaign • Who is your audience? What is the best way to communicate with them? Analytics Team and Executive Team, the best way to communicate with them is a Power Point Presentation • Can data visualization help you share your findings? • Is your presentation accessible to your audience?
Key tasks • Determine the best way to share your findings. Done • Create effective data visualizations. Done • Present your findings. Done • Ensure your work is accessible. Done
Deliverable • Supporting visualizations and key finding Delivery 05 Presentation

Follow these steps:

1. Take out a piece of paper and a pen and sketch some ideas for how you will visualize the data.
2. Once you choose a visual form, open your tool of choice to create your visualization.

Use a presentation software, such as PowerPoint or Google Slides, your spreadsheet program, Tableau, or R.
3. Create your data visualization, remembering that contrast should be used to draw your audience's attention to the most important insights. Use artistic principles including size, color, and shape.
4. Ensure clear meaning through the proper use of common elements, such as headlines, subtitles, and labels.
5. Refine your data visualization by applying deep attention to detail.

Phase 6 Act

The **act** phase is all about putting your findings to work! The objective here is to **make informed decisions and take action** based on the insights you've gained from the data. This might involve implementing changes, making recommendations, or setting new goals.

Remember, the data analysis process is iterative. You might find yourself revisiting previous phases as you gain new insights or refine your understanding of the problem.

> 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:

Summary Table - Act
Guiding questions • What is your final conclusion based on your analysis? There is a good opportunity to convert casual to members and based of the numbers, it is feasible with a good campaign • How could your team and business apply your insights? Targeted marketing campaign to casual riders, advertising the savings in cost for becoming a member since they will be able to do additional trips for a better member price and the health benefits of using a bicycle to commute to work, or running other frequent errands or events in their schedule. Offer a discount coupon for waterproof coat to arrive dry to the workplace when upgrading to member Securely Adding personal metrics like current height, weight, body mass, blood pressure to the app to generate a report comparing the rides and their frequency with so that both casual and members can see their health improves as their bike riding exercise increases with the frequency of rides Notify the all casual and members of a donation to the heart and stroke foundation based on the total number of rides made by members, that would encourage the casuals to save money as members and contribute to a noble cause • What next steps would you or your stakeholders take based on your findings? Approval of the project and designation of project manager and resources to calculate time and effort for Marketing Campaign, App Modification and Public Relations • Is there additional data you could use to expand on your findings? No
Key tasks • Create your portfolio. Done • Add your case study. Done • Practice presenting your case study to a friend or family member. Done
Deliverable • Your top three recommendations based on your analysis **Delivery 06 Top three recommendations based on your analysis (included in the final presentation)**

Follow these steps:

1. If you do not have one already, create an online portfolio.
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.