Table of content and links of interest:

[**Scenario 3**](#_ofpb79hl0zj7)

[**About the company 3**](#_imkgvxsv7zz3)

[**Assigned Question 4**](#_cm3zlmatfe5t)

[**Deliverables 4**](#_8ywol7auafsj)

[**Guidelines & Steps 5**](#_lib792j3oadb)

[**ASK 5**](#_vno2fkkvbq8t)

[**PREPARE 6**](#_bv45tiirh2pl)

[**ANALYZE 8**](#_50nqevdwvbkz)

[Cleaning and filtering! 12](#_hw4o00l6bb3l)

[Analysis 18](#_9qt5lq23pi39)

## 

## 

## 

## 

## 

## 

## 

## 

## 

## 

## 

## 

## 

## 

## 

## 

## 

# Scenario

You are a junior data analyst working on 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.

## About the company

In 2016, Cyclistic launched a successful bike-share offering. Since then, the program has grown to a fleet of 5,824 bicycles that are geotracked and locked into a network of 692 stations

across Chicago. The bikes can be unlocked from one station and returned to any other station in the system anytime.

Until now, Cyclistic’s marketing strategy relied on building general awareness and appealing to broad consumer segments. One approach that helped make these things possible was the flexibility of its pricing plans: single-ride passes, full-day passes, and annual memberships. Customers who purchase single-ride or full-day passes are referred to as casual riders. Customers who purchase annual memberships are Cyclistic members.

Cyclistic’s finance analysts have concluded that annual members are much more profitable than casual riders. Although the pricing flexibility helps Cyclistic attract more customers, Moreno believes that maximizing the number of annual members will be key to future growth. Rather than creating a marketing campaign that targets all-new customers, Moreno believes there is a 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.

# Assigned Question

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

## **Deliverables**

You will produce a report with the following **deliverables (for presentation and documentation):**

1. A clear statement of the business task

2. A description of all data sources used

3. Documentation of any cleaning or manipulation of data

4. A summary of your analysis

5. Supporting visualizations and key findings

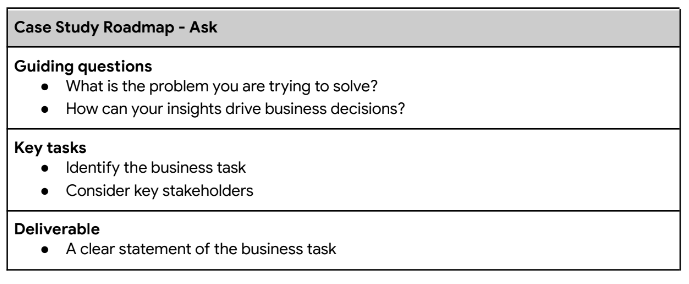
6. Your top three recommendations based on your analysis

* [Presentation link](https://www.canva.com/design/DAF7YnQIero/L4e2uDgFQOzwLcN_KgehfA/view?utm_content=DAF7YnQIero&utm_campaign=designshare&utm_medium=link&utm_source=editor)
* Query and cleaning script
* R programming or Tableau visualizations codes

Note: Completing this case study within a week is a reasonable goal.

# Guidelines & Steps

## **ASK**



**Identify the business task:**

Design marketing strategies aimed at converting casual riders into annual members. We can do this by understanding “How do annual members and casual riders use Cyclistic bikes differently?”

Application:

Making predictions: What data can provide Cyclistic provide

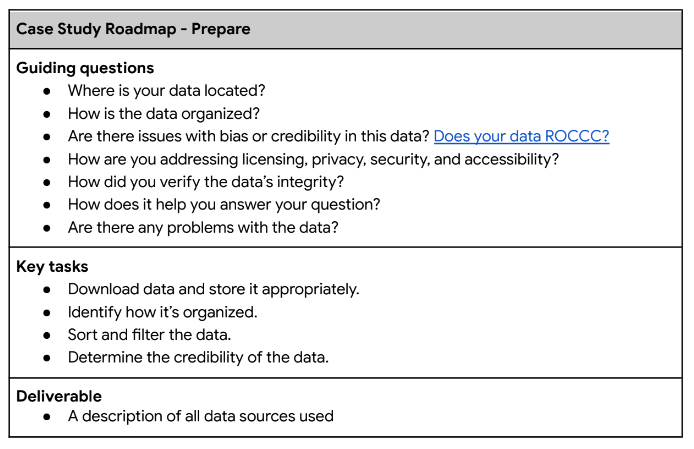
**Consider Key stakeholders:**

* Cyclistic Executives
* Cyclistic Marketing team
* Cyclistic finance analyst team
* Moreno

**Business task:** Design a marketing strategy to convert casual riders into annual members via understanding how casual riders and annual riders differ

## **PREPARE**

(Data usage report)



[ROCCC](https://www.coursera.org/learn/data-preparation/lecture/lHirM/what-is-bad-data)

The data I will be using is a dataset of 2023 customers data provided by Cyclistic rider data from google’s data analytic.

Assumptions:

* The 2023 data accounts for potential risk such as pandemic wide viruses
* The usage of only January and March data due to the limitation of allowable megabytes in bigquery google cloud console.
* Data Points with null variable and ride\_length of + or - 3 standard deviation are invalid data points and to be removed.

The data provided a series of columns that showcased

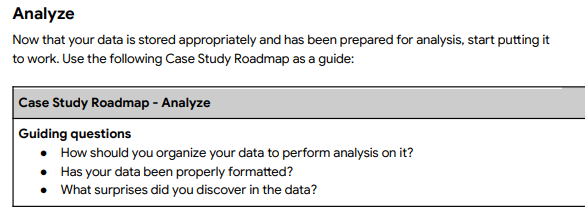
* Ride\_id: the ID of the ride used
* Rideable\_Type: shows one of 3 types of bike, Electric, Docked, Classic
* Started\_at: shows the start time of the ride
* Ended\_at: shows the end time of the ride
* Starting\_station\_name: Shows the start station of the ride
* Starting\_station\_id: provide the start station ID
* End\_station\_name: shows the end station name of the ride
* End\_station\_id: provide the end station ID
* Member\_Casual: shows if the ride was ridden by a casual rider or an annual member.

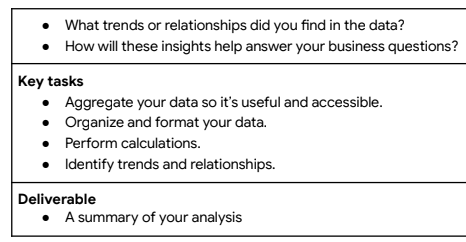
**Credibility**: The data I used is provided by Motivate International Inc, provided under the license of Lyft Bikes and Scooters LLC.

Cleaning steps taken:

| * Select all via top left blank cell between 1 and A   > Data > Sort > Member\_Casual and Rideable\_Type  as sorting variable   * I selected column E > right click > insert: to create a new column and I named it “ride\_length” * In the ride\_length column I subtracted ended\_at with started\_at to get data on how long the ride was. * But it gave me a weird number with subtracted dates as well. So I selected the entire column > right click > format cells> time > HOUR:MINUTES:SECONDS. * Next I created a column next to the previous step and named it “day\_of\_week” and I used the =WEEKDAY function. * In that column I formatted it in CUSTOM > dddd. Then for the first cell I wrote (=WEEKDAY(C2, 1)) and this gave me monday for the first cell. The 1 at the end of the function dictates that Sunday is considered 1 and Saturday is considered 7. * I then did this for all the remaining sheets I have |
| --- |

## ANALYZE





**Follow these steps for using SQL**

Open your SQL tool of choice, then complete the following steps:

1. Import your data.

2. Explore your data, perhaps looking at the total number of rows, distinct values,

maximum, minimum, or mean values.

3. Where relevant, use JOIN statements to combine your relevant data into one table.

4. Create summary statistics.

5. Investigate interesting trends and save that information to a table.

Due to the upload limitations of the free version of bigquery, this following practice will only accommodate 2 files which are the 202301-divvy-tripdata and the 202303-divvy-tripdata CSV files as they’re the smallest in which I can practice on.

Note to self:

| 1. Creating a dataset and table in big query    1. <https://console.cloud.google.com/>    2. Top left there will be a 3 hexagon tab next to the project name    3. Either create a new project or use an existing one and click “Run a query in Bigquery”     d. The 3 dashes on top left and click on Big Query to start a query.  E. Once in the console you can start importing dataset and CSV file by Create Dataset > create table |
| --- |

### Cleaning and filtering!

| We want to go through a checklist when it comes to cleaning in SQL, google’s analyst teachers have taught me to use the following:   * **Sources of errors**: Did you use the right tools and functions to find the source of the errors in your dataset? * **Null data**: Did you search for NULLs using conditional formatting and filters? (IS NULL) * **Misspelled words**: Did you locate all misspellings? (Distinct…) * **Mistyped numbers**: Did you double-check that your numeric data has been entered correctly? (Standard Deviation…) * **Extra spaces and characters**: Did you remove any extra spaces or characters using the **TRIM** function? * **Duplicates**: Did you remove duplicates in spreadsheets using the **Remove Duplicates** function or **DISTINCT** in SQL? * **Mismatched data types**: Did you check that numeric, date, and string data are typecast correctly? * **Messy (inconsistent) strings**: Did you make sure that all of your strings are consistent and meaningful? * **Messy (inconsistent) date formats**: Did you format the dates consistently throughout your dataset? * **Misleading variable labels (columns)**: Did you name your columns meaningfully? * **Truncated data:** Did you check for truncated or missing data that needs correction? * **Business Logic**: Did you check that the data makes sense given your knowledge of the business?   So for starters I use the ‘DISTINCT’ to filter out any outliers such as: misspelling, abnormally long ride\_lengths, extra spaces and characters.  Then I used ‘IS NULL’ functions to find NULL values within the dataset:   | SELECT  DISTINCT rideable\_type  FROM `capstone-guided-project.2023\_Cyclistic\_Data.202301`  In which null irregularities was found:    SELECT  DISTINCT day\_of\_week  FROM `capstone-guided-project.2023\_Cyclistic\_Data.202301`  In which null irregularities was found:    SELECT  DISTINCT member\_casual  FROM `capstone-guided-project.2023\_Cyclistic\_Data.202301`  In which no irregularities was found:    SELECT  \*  FROM `capstone-guided-project.2023\_Cyclistic\_Data.202301`  WHERE  start\_station\_name IS NULL  OR  start\_station\_id IS NULL  OR  end\_station\_name IS NULL  OR  end\_station\_id IS NULL  OR  start\_lat IS NULL  OR  start\_lng IS NULL  OR  end\_lat IS NULL  OR  end\_lng IS NULL  OR  member\_casual IS NULL  In which 300,695 data points were found to have null in at least one of the aforementioned column in the WHERE clause amongst 448,979 data points.    Next we want to find outliers in ride\_length via basic statistical methods. But since ride\_length is under the TIME data type, Bigquery won’t allow me to use AVG, or STDDEV scripts on it. A work around this is to convert it to seconds which will turn it into an INTEGER data type.  CREATE TEMP TABLE `TimeInSeconds` AS (  SELECT  \*,  TIME\_DIFF(ride\_length, TIME '00:00:00', SECOND) AS ridel\_seconds  FROM  `capstone-guided-project.2023\_Cyclistic\_Data.2023013\_v1` );    --This temp table is made to convert the data type of ride\_length to seconds. This is because the TIME datatype cannot be used in AVG or STDDEV\_SAMP. It is also an opportunity for me to practice temp tables!--  CREATE TEMP TABLE `Average` AS(  SELECT  AVG(ridel\_seconds) AS AverageSeconds,  STDDEV\_SAMP(ridel\_seconds) AS STDDEV\_Seconds,  FROM  `TimeInSeconds` );  CREATE TEMP TABLE `Outlier` AS(  SELECT  AverageSeconds + (3\*STDDEV\_Seconds) AS Outlier\_high,  AverageSeconds - (3\*STDDEV\_Seconds) AS Outlier\_low  FROM `Average`  );  --These 2 temp table is made to find the outlier in ride length using average and standard deviations. This roundabout method is used since this is a capstone project guided online project with no real business context behind it.--      The outlier value in seconds would be 5663.390 seconds in the upper curve  And the outlier value in seconds would be -4348.4559 in the lower curve  Since it doesn’t make sense for a ride to be in the negative seconds we decided to filter out any rides that are below 3 minutes.  With this we can finally filter out all the nulls, trim any missing spaces, and exclude potential system error in the form of outliers.  SELECT  \*  FROM  `capstone-guided-project.\_8d1056725aed3cd3e17d2686f9a1e14779d07ee3.\_fea9022a\_51bd\_43bc\_a096\_6a5ad313afbb\_TimeInSeconds`  WHERE  start\_station\_name IS NOT NULL  AND  start\_station\_id IS NOT NULL  AND  end\_station\_name IS NOT NULL  AND  end\_station\_id IS NOT NULL  AND  start\_lat IS NOT NULL  AND  start\_lng IS NOT NULL  AND  end\_lat IS NOT NULL  AND  end\_lng IS NOT NULL  AND  member\_casual IS NOT NULL  AND  ridel\_seconds < 5663.390  AND  ridel\_seconds > 180  The final clean table we got from the query will result in 130,841 datapoint   Analysis The next step is to create a summary statistics table and create a table that excludes any unnecessary columns that might not further the investigation on how to convert casual riders to members.  SELECT  AVG (ridel\_seconds) AS Average\_seconds,  MIN(ridel\_seconds) AS Min\_seconds,  MAX (ridel\_seconds) AS Max\_seconds,  STDDEV (ridel\_seconds) AS Standard\_Deviation,  VARIANCE(ridel\_seconds) AS Var\_seconds,  FROM `capstone-guided-project.2023\_Cyclistic\_Data.2023013\_v2`  Our summary statistics results:    SELECT  rideable\_type,  ridel\_seconds,  day\_of\_week,  member\_casual,  FROM `capstone-guided-project.2023\_Cyclistic\_Data.2023013\_v2`  ORDER BY day\_of\_week  --These are the columns we decided to go along with to identify trends in a visual--  The graph type we decided to go with is bar graph  X-Axis: day of the week  Y-Axis: ride length (seconds)  Color:  Member - Electric: Dark Blue with dots (#003f5c)  Member - Classic: Medium Blue with stripes (#2f4b7c)  Member - Docked: Light Blue with diagonal lines (#665191)  Casual - Electric: Dark Orange with dots (#ff8c00)  Casual - Classic: Medium Orange with stripes (#ffa500)  Casual - Docked: Light Orange with diagonal lines (#ffd700)  This will be divided into 2 different graphs one for members only and one for casuals only to avoid data cluttering. | | --- | |
| --- | --- |

We’re done! The presentation slides will be included amongst other files that highlight my steps taken for this project and findings found during queries phase.