# Introduction

This data analysis is the capstone project for Grow with Google’s Data Analytics course on Coursera. The purpose of this report is to demonstrate the knowledge and skills acquired through the course by documenting the steps of the data analysis process. This report will follow the data analytics process: Ask, Prepare, Process, Analyze, Share, and Act. Each section will cover the tasks of the case study project, what steps were taken to complete each step and document the code for reference. This report and all files, images, and spreadsheets used will be available on my portfolio via Linkedin and Github.

## Case Study

This case study is titled, “How Does a Bike-Share Navigate Speedy Success?”. The scenario involves being a junior data analyst working for 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.

## Methodology

In this case study the data analysis steps will involve using the program language R using the application R Studio. To complete this case study multiple R packages were used to process and analyze the data. The most notable package used was Tidyverse, a collection of R packages. Below is a list and brief description of the primary packages used in Tidyverse as well as other R packages to complete the case study.

* Dplyr – Enables data manipulation of data allowing datasets to be filtered and enable datasets to be summarized and calculated.
* Readr – Package that enabled CSV files from the source to be integrated into R for data analysis.
* Tibble – Package that allowed datasets to be easily read and understood while identifying what type of data is.
* Ggplot2 – Package that enables the creation of visualizations from datasets such as bar charts.
* Lubridate – Package that simplifies and enables the calculation using date and time.

I’d like to mention that the original goal of this case study was to include the use of Excel and T-SQL using MS Access. However, due to the size of the dataset that involved over 5 million observations the ability to clean and combine the data for data analysis was too large. Excel was not designed to calculate or clean that much data. Using MS Access involved the same challenges, while it was possible to import the data, the 2GB file size limit made creating queries where the data was cleaned exceed the maximum file size. Thus, my decision to use R for the entire case study.

# Ask

To create a marketing strategy to convert casual riders into annual members, Lily Moreno wants to understand how annual members and casual riders differ, why casual riders would buy a membership, and how digital media could affect their marketing tactics. Out of these three questions, I have been assigned the following question, “How annual members and casual riders use Cyclistic bikes differently?”.

## Stakeholders

The primary stakeholders for this case study are as follows.

* Lily Moreno – Director of Marketing at Cyclistic who believes the future success of Cyclistic is to maximize the number of annual memberships. Lily Moreno creates the assignment for the case study to determine if the future of Cyclistic is in annual memberships.
* Cyclistic Executive Team – The executive team is also a stakeholder as approval of any marketing programs require their approval. A presentation explaining the results of the data analysis as well as recommendations on how to increase annual memberships will need to be given to the executive team.

# Prepare

To answer the question “How do annual members and casual riders use Cyclistic bikes differently?”. I will be using Cyclistic’s historical trip data analyzing and identifying trends.

## The Data

The historical data I will be using for the capstone come from the website Divvy since Cyclistic is a fictional company. The data is located at <https://divvy-tripdata.s3.amazonaws.com/index.html>. I can use this data under the Data License Agreement at <https://ride.divvybikes.com/data-license-agreement>. The data is anonymous by identifying use of bikes by Ride ID so there is no identifiable information. The data is organized separating the ride data monthly. For this report, the data I used ranges from November 2020 to October 2021.

The data involves the following column names.

* Ride\_id
* Rideable\_type
* Started\_at
* Ended\_at
* Start\_station\_name
* Start\_station\_id
* End\_station\_name
* End\_station\_id
* Start\_lat
* Start\_lng
* End\_lat
* End\_lng
* Member\_casual

As a result, the data used is reliable, original, comprehensive, current, and cited.

# Process

Before importing the data into R Studio, I needed to examine if the data was dirty or clean. While the website Divvy stated that the data was processed removing trips taking by staff and trips that were below 60 seconds in length. The reality was that the data upon examining the data in Excel using the filter tool showed that the data was dirty in several ways.

## Dirty Data

The data was dirty by having blank entries, rides that were under 60 seconds and over 24 hours, trips where the start time was after the end time, and station names that were identified as test stations. Any observations that had these criteria would skew the analysis.

It was also at this time that I also identified data that would be irrelevant to the analysis answering the question, “How annual members and casual riders use Cyclistic bikes differently?”. I decided that removing the columns start\_lat, start\_lng, end\_lat, and end\_lng needed to be done as the observations did not help answer how casual riders and annual members differ.

## Cleaning Data

Now that the dirty data was identified, it was time to import the data into R Studio.

With the trip data imported, the first step was to combine every month’s data into a single dataset which I labeled all\_trips that had a total of 5,378,834 observations

-Insert code chunk bind rows

Next, I removed the columns start\_lat, start\_lng, end\_lat, and end\_lng as they were not needed for the analysis phase.

-inset code chunk removing columns

With the columns removed it was time to start cleaning the data starting with removing observations that had blank values. To ensure the integrity of the data, each step cleaning data I created a new data set version of all\_trips creating all\_tripsv2. The result of removing all missing values reduced the number of observations from 5,378,834 to 4,492,727.

-Insert code chunk using complete.cases

The next step cleaning involved removing observations where the ended\_at date/time occurred before the started\_at date/time.

-Insert code chunk using subset

To remove rides that were less than 60 seconds and more than 24 hours, I needed to create a new column called ride\_length using the difftime function between the started\_at and ended\_at date/time. Then I used the subset function to create all\_tripsv4 removing any observations less than 60 seconds and greater than 86400 seconds (24 hours).

* Insert code chunk.

## Preparations

To prepare the data for analysis I chose to separate the time and dates into their own columns that will allow me to aggregate the dataset. This involved creating the following columns.

* Date
* Year
* Month
* Day
* Day of week (Monday, Tuesday, etc.)

With the dataset cleaned from all erroneous data and columns that will let me aggregate the data I can now move onto analyzing the data.

# Analyze

To analyze the data and to answer the question “How do annual members and casual riders use Cyclistic bikes differently?”. There were four questions I wanted to learn from the data available.

1. How many rides have casual riders take compared to annual members?
2. What kind of bikes do casual riders use compared to annual members?
3. How many rides do casual riders take each day of the week compared to annual members?
4. How long do casual riders rent out bikes compared to annual members?

## Rides Taken

The first piece of data I wanted to analyze is to count the number of rides casual riders take compared to annual members.

-Insert picture of table and code chunk

It turns out that in terms of bike rentals, annual members use Cyclist bikes more than casual users. With annual members already being the primary user of Cyclist I’ll need to find how casual riders use Cyclist bikes more than annual members.

When I break down the number of rides taken by what day of the week the first difference between casual riders and annual members appears.

-insert picture of rides taken per day table and code chunk

Casual riders use Cyclist bikes more during the weekends while annual members use Cyclist bikes during the weekdays.

## Bikes Preference

Another aspect I thought would tell me the difference between casual riders and annual members was the type of bike they rented out. There are three types of bikes that Cyclist rents out: Classic bikes, docked bikes, and Electric bikes.

-Insert table of bike type and code chunk

While classic bikes are the preferred choice between both casual riders and annual members, casual riders are more flexible in the type of bikes they use. Compared to annual members, they prefer using classic bikes and electric bikes more than docked bikes.

## Average Trip Duration

The last difference I wanted to find out was how long do both casual riders and annual members use Cyclist bikes each day of the week.

-Insert picture of duration table

The surprising difference is that despite annual members taking more trips than casual riders, annual riders use Cyclist bikes for a shorter period compared to casual riders.

Based on the number of rides taken, how many rides are taken each day, the type of bike used, and the duration the analysis shows some stark differences between annual members and casual riders. We know casual riders now use Cyclist bikes for longer periods of time and primary use them on weekends and much less on weekdays. These are habits that can help convert casual riders into annual members.

# Share

With the analysis complete, I needed to create visualizations based on the data but also create a presentation for the executive team answering the question and providing recommendations to convert casual riders to annual members.

## Creating visualizations

Thanks to ggplot2 package, I was able to convert the results of my analysis into visualizations in R studio. To make the visualizations presentable for the executive team, I used the classic theme and added labels to the x-axis, y-axis, legend, title, subtitle, and caption citing the source of data. I chose to present the visualizations through bar charts due as it was the easiest way to convey the massive number of rides taken over the course of the year and provide a direct comparison between casual riders and annual members.

* Insert picture and code

## Presentation

With the visualizations completed, I opted to create a slideshow presentation for the executive team using PowerPoint. <https://docs.google.com/presentation/d/1o-HOgjYE81pqacblnLpN0ypVDsuCxZNX5dv9b81S8Ws/edit?usp=sharing>

This presentation shows the executive team the problem, the result of the analysis and recommendations that not only answer the question of what makes casual riders different then annual members but provide solutions based on the data.

# Act

Based on the data analysis that show the differences between annual members and casual riders. Here are my recommendations that Cyclistic can do to convert casual riders into annual members.

1. Target casual riders who use Cyclistic bikes on weekdays.
2. Offer incentives for casual riders to use electric bikes
3. Offer weekend incentives such as longer rental times to convert casual riders to become annual members.

By expanding the annual member benefits towards how casual riders use Cyclistic bikes create the best opportunity for casual riders to convert towards annual members. Allowing longer rental times over the weekend can be marketed as a cost saving measure to casual riders who primary use Cyclistic bikes over the weekend. While classic bikes are still dominant, the growth of electric bikes is something that can convince casual riders to convert as well. Finally, there is a sizeable number of casual riders who use Cyclistic bikes on the weekdays which we can hypothesize that they also use Cyclistic bikes to get to work like annual members. Targeting those casual riders can also lead to annual member conversions and secure the future of Cyclistic.