Title: Google Data Analytics Capstone project  
Subtitle: Cyclistic Case Study  
Author: Ishika Shrivastava  
Date: 08 JUNE 2023

\_ ALL ABOUT

For the capstone project of the Google Data Analytics certificate, I have chosen the Cyclistic bike share data to work on. For the case study, I will perform real-world tasks of a junior data analyst for the marketing team at Cyclistic, a fictional bike-share company in Chicago.  
To answer key business questions, I will follow the six steps of the data analysis process : Ask, Prepare, Process, Analyze, Share and Act.

\_ THE SCENARIO  
The director of marketing of Cyclistic, Lily Moreno, believes that the company’s future growth depends on maximizing the number of annual memberships. Hence, the marketing analyst team wants to understand how casual riders and annual members use Cyclistic bikes differently. From these insights, the analytics team could be able to design a new marketing strategy to convert casual riders into annual members.

Three questions will guide the future marketing campaign:

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

2.Why would casual riders buy Cyclistic annual memberships?

3.How can Cyclitic use digital media to influence casual riders to become members?

1. Ask phase

* A statement of the business task:

Cyclistic has concluded that annual members are much more profitable than casual riders. So, we want to design a marketing strategies and a campaign that helps us converting casual riders into annual members.

* My key stakeholders are:

1-Lily Moreno: The director of marketing and my manager. Moreno has initiated this strategy. The first stakeholder to deliver to.

2-The executive team: For Moreno´s idea to work, the executive team must approve our recommendations, so so they must be backed up with compelling data insights and professional data visualizations.

1. Prepare phase  
   Past 12 month of original bike share data set from 01/01/2021 to 31/12/2021 was extracted as 12 zipped .csv [files]:  
   LINK:<https://divvy-tripdata.s3.amazonaws.com/index.html>  
   The data is made available and licensed by Motivate International Inc under this [

Data Organization & Description:

File naming convention: YYYY\_MM

File Type: csv format

File Content: Each csv file consist of 13 columns which contain information related to ride id, rider type, ride start and end time, start and end location etc. Number of rows varies between 49k to 531k from different excel files.

Data credibility:

The data set is reliable, the data is complete and accurate for the chosen time window.

The data is original, it is a first arty information.

The data is comprehensive, the data set contains all information needed to answer the question.

The data is current, rider data of the last 12 months was used.

The data is cited and vetted by Chicago department of transportation.

Data Security: Riders’ personal identifiable information is hidden through tokenization.

Original files are backed up in a separate folder.

Data Limitations: As riders’ personal identifiable information is hidden, thus will not be able to connect pass purchases to credit cards numbers to determine if casual riders live in the Cyclistic service area or if they have purchased multiple single passes.

Credibility of data:

. The data set is trustworthy, and the data is complete and accurate for the time period selected.

. The information is unique; it is the first of its kind.

. The data is complete; the data set contains all of the information required to answer the question.

. The data is current; rider data from the previous year was used.

. The Chicago Department of Transportation cites and verifies the data.

. Tokenization conceals the personally identifiable information of riders.

. The original files are kept in a separate folder.

Data Restrictions: Because riders' personally identifiable information is hidden, it will be unable to link pass purchases to credit card numbers in order to determine whether casual riders live in the Cyclistic service region.

1. Process Phase

I used R for data verification and cleaning: Reasons:  
The 12 data sets combined will contain more than 5 million rows of data. Excel worksheet limitation is 1,048,576 rows. Moreover, some csv files could not uploaded to BigQuery for file size problems. Thus, R is used to perform all tasks from organizing, cleaning analyzing and visualizing data.

1. Analyze Phase  
   USING R STUDIO FOR THIS PHASE:  
   several libraries that help reading, cleaning, organizing and analyzing the data.  
   library(tidyverse)  
   library(janitor)  
   library(skimr)  
   library(here)  
   library(hablar)  
   library(readxl)  
   library(data.table)  
   library(chron)  
   library(readr)  
   library(lubridate)  
   library(magrittr)  
   library(DescTools)  
   library(metR)
2. Share phase

* Weekends are the busiest days for casual rides (plot 3). There's a good chance they're tourists taking in the views of the city, or they're regular Chicago residents out for a weekend bike ride. The lengthier average ride time for casual riders (plot2), which also peaks during the weekend, demonstrates this point.
* Ride length for members are relatively shorter compared to casual riders. This could clarified as such, that most members use the bikes to commute on workdays. This clarification would also explain the short riding durations of members. They ride from point A to B, namely roughly always the same ride lengths and the same distance

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_THANK YOU\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

Top of Form

Bottom of Form