Dylan Seligman

5/27/23

Google Capstone Case Study

Over the past few months, I have been working on completing the Google Data Analytics certification on Coursera. While I have a data analytics background through my college experience at Baruch College, it is always nice to refresh your skills so that you do not lose them.

This report will be sharing the approach I used towards this case study using Google BigQuery and Tableau. Through this program, I learned how to utilize the analysis process which is: ***Ask, Prepare, Process, Analyze, Share & Act***.

**Scenario:**

You are a junior data analyst working in the marketing analyst team at Cyclistic, a bike-share company in Chicago. The director of marketing Lily Moreno believes the company’s future success depends on maximizing the number of annual memberships.

Our goal is to help identify trends/ gaps seen between casual riders and our annual members to help create targeted marketing strategies tailored towards converting casual riders into annual members. To do this, we need to understand how each group uses Cyclistic differently.

**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 very good chance 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.

**Ask**

After careful deliberation, three questions have been identified to 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?

**Prepare**

To better understand where Cyclistics currently stands today, it is important to understand how they have been performing, at least in the short-term. To do this, we will be using the last 12 months of Cyclistics data (located [here](https://divvy-tripdata.s3.amazonaws.com/index.html)) from April 2022 to April 2023 (202204-divvy-tripdata.zip -> 202304-divvy-tripdata.zip). We will then store these items as csv’s in a folder I called “Divvy CSV” to understand context and separate our raw data from our transformed data which will come later. We will also rename each file in order to improve readability in this type of format seen below:

(202204-divvy-tripdata.csv -> divvy\_tripdata\_april\_2022.csv)

**\*\*Note: The datasets have a different name because Cyclistic is a fictional company. For the purposes of this case study, the datasets are appropriate and will enable you to answer the business questions. The data has been made available by Motivate International Inc. under** [**this**](https://www.divvybikes.com/data-license-agreement) **license).\*\***

**Prepare**

**In Excel:**

To ensure that our data is cleaned and ready to be uploaded to BigQuery, I went into each file and removed any duplicate records I could find. To do this, I went to find and replace -> remove duplicates on the rows entitled “row\_id”, “ride\_start\_time”,”ride\_end\_time.” I also ensured that the record formatting was uniform throughout.

The total amount of data for this exercise was nearly 6 million records and using excel spreadsheets alone to analyze this dataset would quite literally crash it (I know from personal experience). In this case, I went ahead and uploaded the CSV’s individually into BigQuery. See the process below.

**Step 1: Ensure you are logged into the Google Cloud platform and scroll over to the “Storage” section. Once there, click “Cloud Storage”.**

**A screenshot of a computer

Description automatically generated with low confidence**

**Step 2: Once clicked, create your bucket and make sure to name it as this is where we will be storing our data in Google Cloud.**

\*\*Note, to use buckets for your data you must have unlimited access to Google Cloud which is a paid product outside of the free sandbox version. If you do not currently own this, Google offers a free $300 credit for you to try before committing. \*\*\*

* Once we create our bucket, select ‘Region — lowest latency within a single region’ for the cheapest storing option and pick the one closest to you. In my case, South Carolina was the closest to me and storage for my CSVs was quoted at $.0020 per month.

**Step 3: Upload the CSVs individually into your bucket**

**A screenshot of a computer

Description automatically generated with medium confidence**

**Step 4: Accessing your CSVs on BigQuery**

* Head to the BigQuery dashboard and click the blue plus sign. When prompted, select create table from the Google Cloud Storage option

A picture containing text, screenshot, businesscard, font

Description automatically generated

A screenshot of a computer

Description automatically generated with medium confidence

Now that we have imported our files successfully, it is time to move onto the **Process** phase

**Process**

Firstly, as our data is imported, ensure that the following match up.

* The Schemas
* The naming and formatting of each record type was uniform throughout.

A screenshot of a computer

Description automatically generated with medium confidence

A screenshot of a computer

Description automatically generated with medium confidence

Once we confirm that the schemas and records are the same, it is integral that we combine all the files into one table.

To do this, our options are between using JOIN or Union in SQL. The difference between the two is that using JOIN combines data horizontally based on their relationship while Union combines data vertically and eliminates duplicated items by default. As such, for this exercise, it is more efficient to use Union to combine the tables

A screenshot of a computer

Description automatically generated with medium confidence

I then saved this query into a new table which I called ‘df\_merged’.

A screenshot of a computer

Description automatically generated with medium confidence

As this data stands, it was not enough to gleam any insight of use. Therefore, we will have to create new columns. I started by using Extract() and Case() to assign each item, in the “started\_at” column a number from 1-7 and renamed these numbers a day of the week starting from Sunday and ending at Saturday

A screenshot of a computer

Description automatically generated with medium confidence

I then calculated the trip duration by subtracting the ended\_at column from the started\_at column using the TIMESTAMP\_DIFF function

A screenshot of a computer

Description automatically generated with medium confidence

A screenshot of a computer code

Description automatically generated with low confidence

A screenshot of a calendar

Description automatically generated with medium confidence

I then filtered out any rides that were under 1 minute by using the following code

A screenshot of a computer code

Description automatically generated with low confidence

**Share**

Now that the SQL part and calculations are out of the way, we must export the data back into a CSV file and visualize in Tableau.

To do this, I exported my query results into a new table called “divvy\_vf.”

A picture containing text, font, white, screenshot

Description automatically generated

Then, I went to my table and clicked the three dot icon before clicking export. When exporting, make sure you click export to GCS. Clicking any other options will only export a fraction of the data.

A screenshot of a computer

Description automatically generated

Next, I selected the bucket I had created in the very beginning as an export destination and named my file “divvy\_transf.csv.” It is important to include the .csv at the end of the file name, otherwise you will be exporting a text file. Once that was done, I simply downloaded the new CSV file to my laptop

**Share – Tableau**

Here, I made a connection between Tableau and my recently downloaded CSV file. In order to do this, I first converted my CSV to excel and then established a connection between Tableau and Excel.

A screenshot of a computer

Description automatically generated with medium confidence

Before making any sort of visualizations, it is important to always ask yourself of what is being asked. In this instance it is a 3-fold question

1. How do casual cyclists and members use our cycles differently throughout the week?
2. What is the average trip duration difference between casual cyclists and members?
3. How do our trips differ between casual cyclists and members vary by month

A picture containing text, screenshot, plot, diagram

Description automatically generated

Here you can see that while for a majority of the year members tend to ride bikes more frequently than casual cyclists, casual cyclists tend to ride more frequently than members from April until August.

This could be because the weather during the months where annual members tend to out bike casual cyclists tend to be months where biking around may not be ideal and that casual cyclists prefer to bike during periods of historically warmer weather.

A picture containing text, screenshot, diagram, plot

Description automatically generated

Moreover, annual members tend to bike more then casual cyclists on weekdays while casual cyclists tend to bike more than members on the weekends. This could signify that members tend to cycle to and from their workplaces while casual cyclists tend to cycle for their leisure or downtime.

A graph with red and green lines

Description automatically generated with low confidence

Here you can see that the average trip duration for both tends to increase as the year gets warmer, but at a much higher rate for members than casual cyclists. In conjunction with my previous visualization, This could indicate that casual riders tend to ride shorter distances at an increased frequency during warmer months while members tend to ride longer distances at a decreased frequency during these same months.

**Act**

Based on my findings, I wanted to offer my insight as to why the data has reflected what it has

1. I believe that casual riders tend to consist both of people who prefer taking other forms of transportation during the colder months as biking around in the rain or snow is not an ideal way of going to and from work.
2. I also believe that a majority of the annual members are working class adults as the average trip durations indicate that members tend to bike longer distances than casual members, indicating possible times to and from workplaces.
3. There is also a possibility that given how companies have switched from a WFH model to majority time in the office, that this can account for the increased number of members you see biking from Tuesday through Friday as a lot of companies have switched to either 3 days in office, 2 days remote or 4 days in office, 1 remote.

Moving into what strategies we can implement to convert casual riders into members, I would suggest the following:

1. Market heavily in the Spring-Summer months as those are times, we see a huge uptick in the number of casual riders. Offer discounts and incentives to entice casual riders that the membership is worth it. Offer additional benefits to casual customers who have been using our product but have not converted to the membership.
2. Impose a distance/time limit on casual riders limiting the distance/time they can use our bikes. This is because casual members already tend to use our product for shorter times than our members. Limiting the time/distance they can use our bikes while offering unlimited time/distance to our members could mean that a number of casual riders convert to the membership option.