**Google Data Analytics Certificate – Capstone Project**

Case Study: How Does a Bike-Share Navigate Speedy Success?

**Introduction**

This is the culmination of my Google Data Analytics Certificate journey. In this case study I use the framework taught in this course to produce a variety of deliverables similar to a real world project. The deliverables include:

1. A clear statement of the business task - Ask
2. A description of the data sources used - Prepare
3. Documentation of any cleaning or manipulation of the data - Process
4. A summary of the analysis - Analyze
5. Supporting visualizations and key findings - Share
6. The top 3 recommendations based on the analysis - Act

**Case Study Summary**

The premise of this case study is that a fictional bike share company, Cyclistic, is trying to convert casual users of their service into annual members. Their marketing director has asked me to analyze the past years’ worth of user data to see how casual users and members use their service differently. I used the 6 steps of the data analysis process presented in this course to complete this project and organized this report in a similar manner.

**Ask**

The business task for this case study is to analyze the last years’ worth of data to understand how casual users and annual members use Cyclistic differently. After analyzing the data, the other business task is to offer reasons why casual users may want to purchase an annual membership.

**Prepare**

The data used in this case study is from: <https://divvy-tripdata.s3.amazonaws.com/index.html>. It is used under this license: <https://ride.divvybikes.com/data-license-agreement>.

The data is composed of the last 12 months of user data, specifically June 2022 through May 2023. It is organized into separate files by month. Each file contains the following columns: 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, and member\_casual.

Using excel, I browsed through the data using filtering and conditional formatting to make sure the data was consistent and not missing information. During this analysis I found that the start\_station\_name, start\_station\_id, end\_station\_name, and end\_station\_id columns were the only ones that had nulls. Other than these columns, the data is very consistent and complete. Combined, there are over 5.8 million rows of data.

**Process**

After making sure each of the files were organized in the same manner and were consistent and complete, I had to decide which tools I would use to perform the main analysis of the project. Due to the large amount of data I was working with I decided to use BigQuery rather than using a spreadsheet program. Using BigQuery, however, complicated the process slightly due to the limitations on upload size for free users.

Due to the large file size of each of the months’ worth of data, I had to eliminate certain columns that did not support the analysis I was trying to complete. Normally this is not something that would be done but was a work around to the upload limits.

I removed the start\_station\_name, start\_station\_id, end\_station\_name, end\_station\_id, start\_lat, start\_lng, end\_lat, and end\_lng columns as I decided to focus on the number of rentals and the length of rentals in my analysis and these columns did not contribute to that type of analysis. Though, in reality, this information could be used to see where casual users and annual members begin and end their trips or the distance traveled during a rental period which could also be used in the marketing strategy.

**Analyze**

To get a sense of what kind of information I might be able to pull from this data I first looked at the data from June 2022 on its own in Google Sheets.

I began by creating a pivot table that counted the number of rentals by user type (member or casual).

I then added two new columns of calculated data: day\_of\_week and length\_of\_trip. day\_of\_week was calculated from the time stamp in the started\_at column. length\_of\_trip was calculated using the ended\_at and started\_at columns. While looking at the length\_of\_trip column I found 8 negative values, this is where the ended\_at time came before the started\_at time for some reason (some kind of entry error perhapse).

I then created a pivot table that compared the average trip length, max trip length, and minimum trip length between casual users and annual members. This began to show some very interesting differences between the user types. I also compared the average trip lengths with and without the negative values included. Because they were such a small percentage of the total trips, they did not affect the average. As such I decided not to exclude them for ease of analysis.

Lastly, I created a pivot table separating out the number and average length of trips per day of the week by user type.

Spreadsheet here: https://docs.google.com/spreadsheets/d/1Kpw5NSu3SsnhQ\_SoOhPMe2gJjuyJgoO7nkzwoY3TjjM/edit?usp=sharing

With the results of these pivot tables guiding my analysis, I now knew what kind of queries I would have to write in BigQuery and SQL. After much trial and error, I was able to formulate 2 separate queries that provided similar results to the pivot tables I created in the spreadsheet. At this time, I also realized that further breaking down the data by month was also important and created a third query to analyze that aspect of the data.

The main findings of my analysis are:

1. Members rent their bicycles for a much shorter amount of time.
2. Members’ average rental period is consistent throughout the week, while casual users rent for a very wide range of times.
3. Casual users mostly rent on the weekends while members rent mostly during the middle of the week.
4. Peak months for rentals are during the summer.
5. Members have a much higher use during the winter months than casual users.

**Share**

With this information I had to decide how to best present these findings in an easy-to-understand, visual way. I decided to use Tableau to make these visualizations.

A graph of green and orange bars

Description automatically generated

A graph of green and orange bars

Description automatically generated

A graph of green and orange bars

Description automatically generated

A graph of a bar chart

Description automatically generated

After producing these visualizations, I created a PowerPoint to fully present all of my findings and my final recommendations.

**Act**

To fulfil the business task my top 3 recommendations for converting casual users to annual members are:

1. For casual users that are renting for extremely long periods: show them how much money could be saved with a membership and using it for short trips more often rather than renting for one long period.
2. For casual users who are renting throughout the middle of the week: show how convenient it is to grab a bike and go as needed with a membership.
3. Compare casual users that rent for short periods of time but often to annual members and show how much more convenient and cheaper a membership is to having to pay for each individual ride.

**Conclusion**

This was a very fun project to work on. It really tested all the skills that I have been building during the Google Data Analytics Certificate. I also learned a lot. There is a big difference between learning about all these different programs and actually using them. Although there were many challenges and I made many mistakes along the way, I learned from them and have a much stronger sense of how to use the programs and much more respect for those who do this work effortlessly. I am excited to work on more projects like this one as well as other types that will bring new and exciting challenges to overcome.