**Google Data Analytics Case Study #1:**

**Guided Questions**

**Part 2: *Prepare***

**1. Where is your data located?**

- The data is located in a large folder on the Amazon Web Services and is provided by Lyft and Divvy. Each dataset for the 2022 year is broken up by month of the year and are saved as CSV files.

**2. How is the data organized?**

- The data is organized into 12 CSV files based on each month of the 2022 year. Each dataset consists of columns named: ride\_id, rideable\_type, started\_at, ended\_at, start\_station\_name, start\_station\_id, end\_station\_name, end\_station\_id, start\_lat, start\_long, end\_lat, end\_long, and member\_casual. These columns provide information regarding the locations of all bike rides (beginning and end), the id assigned to the user, the user type (member or casual), and the type of bike used for transport.

**3. Are there issues with bias or credibility in this data?**

- This data is collected directly via the company and come from trackers installed within each bike. In a general overview of the data, there seems to be no areas that would be able to have bias included within. As this data is product data, the company uses the data internally to analyze the use of their products to gain insights towards developing future approaches to their products.

**4. How are you addressing licensing, privacy, security, and accessibility?**

- The licensing for this data is provided via the Google Data Analytics Case Study 1 project guide. The data is owned by Lyft Bikes and Scooters, LLC, as the Cyclistic company is a fictional company being used for this project. The link for the data license can be found below.

**5. How did you verify the data’s integrity?**

- The data's integrity is verified through the source in which it is collected. As each bike has a tracker attached to the frame that collects data as the bike is used, this data is immediately available and directly input into the database without the needs of user input. This data is strictly observational based on the activation of the tracker when a bike is unlocked for use.

**6. How does it help you answer your question?**

- The data contains plenty of useful information that can help us compare and contrast the uses of bikes via each user type. There is a column dedicated to identifying the user type of the individual who uses a bike, which allows us to make direct comparisons of how each user type uses the system. Through the use of the member\_casual column, we can create our own columns that calculate different components for comparison, such as number of rides by day, weekday, and month, or by ride lengths per category by using the started\_at and ended\_at columns to produce a ride time for each record. These can be easily reviewed and visualized to get a good view on how members and causal riders use the bikes in different ways.

**7. Are there any problems with the data?**

- After checking for missing, not a number, and null results in each column, it was found that there were zero missing values in all columns of the dataset. This helps confirm the data integrity, as all columns have a value for each record. We also insured that the ride\_id column did not contain any duplicate values. This is important because it ensures that each record is a unique entry and is being recorded in the database properly. On the other hand, there are some changes to the data that can be made to set up our analysis. We can use the started\_at and ended\_at columns to extract a ride time and a date (including the day name, day number, and month name) that allows us to analyze how each user type differs in ride time, daily use, weekly use, and monthly use.

**Part 3: *Process***

**1. What tools are you choosing and why?**

- For this data analysis, I am choosing to use Python and its associated libraries (plotly, numpy, seaborn, matplotlib, datetime, and pandas). I chose to use Python due to its accessibility to manipulate and analyze large datasets. Since the entire year’s worth of data will be formulated into a single entity, Python will work well with the size of the dataset. Python also allows for easy data manipulation, calculation, and visualization, thus making it a "one-stop shop" for all my needs for this analysis.

**2. Have you ensured your data's integrity?**

- The data's integrity has been shown through its collection and pre-viewing. The data contains no null, not a number, or blank values within the dataset, along with no duplicate values within the ride\_id column. The data is from a first-hand source, and is directly taken from collection into a database without the need of input from users.

**3. What steps have you taken to ensure your data is clean?**

- For this dataset, I started by checking for null, not a number, and blank values within the combined dataset. This result showed none of these values, ensuring that each column contained a value for each record. I then checked for duplicate values within the ride\_id column and found no duplicates within the column. This shows that each record in the dataset is unique. At this point, I chose to drop the columns that I deemed unneccesary for my analysis (start\_station\_name, start\_station\_id, end\_station\_name, end\_station\_id, start\_lat, start\_lng, end\_lat, end\_lng). From here, I needed to create new columns for ride time, weekday, day of month, month, and year. To do this, I started by changing the value type of the started\_at and ended\_at columns from objects to datetime64[ns]. This allows me to use the datetime library to manipulate these columns to create colums for ride time, weekday, day of month, month, and year. First, I use the pandas timedelta function in conjunction with subtracting the ended\_at column time from the started\_at column time. I then transferred the datavalue from datetime64[ns] to int32 and created the column ride\_time. Next, I checked to ensure that all values from this calculation greater than zero. I found that 74 records had a negative value for ride\_time. I chose to drop those records by setting the dataset equal to all ride\_time records that were greater than zero, and I reset and dropped the index column. After this, I checked for any values less than zero and the result was zero columns in the dataset with negative values. I then created a day\_num, day\_of\_week, month, and year columns by using the dt.day, dt.day\_name, dt.month\_name, and dt.year on the started\_at column to return values for each record in the dataset. I confirmed the code worked correctly by looking at a head of the dataset and ensuring that the values were correct. From here, I ensured that each column in my dataset had the correct amount of unique values, specifically the rideable\_type, member\_casual, day\_num, day\_of\_week, month, and year columns, by using the .nunique() function. This returned the expected number of results for each column. Lastly, I double checked that there were no null, not a number, and blank values, along with checking the ride\_id column for duplicate values. Both tests reuturned zero results. From here, I concluded that my dataset was clean and ready for analysis. I saved the new dataset as a CSV and went into the analysis portion of this project.

**4. How can you verify that your data is clean and ready to analyze?**

- I verified that my data was clean through various checks on the data. I started with ensuring that each column had the expected number of unique values, specifically the rideable\_type, member\_casual, day\_num, day\_of\_week, month, and year columns, by using the .nunique() function. I then checked the dataset for null, not a number and blank values by using the isna() function in conjunction with the .sum() function. Lastly, I ensured that the ride\_id column had no dupelicate values that would skew my analysis by using the .duplicated() function in conjunction with the .sum() function.

**5. Have you documented your cleaning process so you can review and share those results?**

- I used the comment style text within each coding cell to ensure that each step of the process was documented.

**Part 4: *Analyze***

**1. How should you organize your data to perform analysis on it?**

- I will organize my data into smaller data frames that focus on the key aspects that I want to analyze. I will start by viewing the percentages of members to users throughout 2022. Then I will look at the average ride times by user type, then the average ride types by weekday, day of month, and month per user type. From here, I will look at the total number of rides per user type, then I will look at the total number of rides by weekday, day of month, and month per user type. Lastly, I will look at the total number of rides by bike per user type. These facets should give an in-depth comparison on how each user type uses the bike-share program differently.

**2. Has your data been properly formatted?**

- I have creating smaller data frames for each aspect of the analysis to allow for quicker visualization and an easier to read format. Each data frame provides a simple table-like view of the data and allows stakeholders to easily compare the differences between each user type. I also kept the analysis in a order that makes sense, in terms of flow and groupings. This will allow my visualizations to follow the same pattern.

**3. What surprises did you discover in the data?**

- In this analysis, I found it very interesting that there were a greater number of members versus casual users for the entire 2022 year. This, combined with the greater number of rides, allowed me to make a key guess into what the data is trying to show. I believe that the data shows this because the members of the bike-share program are individuals who live within Chicago and use this program as a primary means of transportation across the city. This is backed by the lower ride times, in comparison to the casual users.

**4. What trends or relationships did you find in the data?**

- Similar to the answer above, the data shows that there were more member users than casual users over the 2022 year. The member users also trump the casual users in total number of rides throughout the year. The casual users, however, have a much larger average ride time than members in all categories (weekday, day of month, month). Lastly, the data shows that classic bikes and electric bikes are favored by members and docked bikes are only used by casual users.

**5. How will these insights help answer your business questions?**

- These insights help paint a story on how casual users and members of the bike-share program use the system in different ways. For example, the data shown for average ride time between casual users and members idicates a greater average ride time for casual members by week day, day of month, and month. This is likely due to the time limit that is set for casual members. In my experience, I have been a casual user of a bike-share program in Tulsa, Oklahoma and I had a four hour time limit for bike use. I wanted to make sure that I got my full value out of the money I spent, so I made sure to use as much time as I could to ride the bike throughout the city. On top of this, the data shows that only casual members used the docked bike types. This makes sense, as these bikes are commonly locked up at various access points across the city and are not available for long-term renting. Members are able to take out both classic and electric bikes, likely with no added charge, so they will prefer to use these bikes for their daily travels. Knowing this information will be key to provide potential insights for the stakeholders that can be used to boost the number of members for the bike-share program.

**Part 5: Share**

**1. Were you able to answer the question of how annual members and casual riders use Cyclistic bikes differently?**

- Through the analytical and visual analysis, I am able to answer the business task of this project. The data shows relationships that reflect the likely story being told by this dataset.

**2. What story does your data tell?**

- The story of this dataset references the ways that casual users and members of the bike-share program use it differently. To start, the data shows that there are a greater number of members in the bike-share program than casual riders. This shows that the program is doing a good job of getting users to become members of the program, but only compromises 59 percent of activity in the program. This number is good, but the company would likely want to see a larger margin between the members and casual users. Next, the majority of bike rides are taken by members, yet they average a much smaller average ride time than casual users. This gives an insight that most members of the bike-share program likely live within the city limits of Chicago, and likely use the bike-share program as a primary means of transportation within the city. On the side of the casual users, the data shows some similarity to my personal experiences with using a bike-share program in my home state. I have been a casual user for that program, and it is a service that allows to rent docked bikes throughout the city for a given amount of time, based on how much an individual is paying for the service. When I use the program, I make sure that I get the most value out of my money, and make sure to ride the bike as much as possible within the allotted time for riding. This is shown by the data, as casual members boast a far greater average ride time versus members. Casual users only have a limited time that they can ride the bikes, so they are likely wanting to spend as much time riding as possible. Also, the data shows that only casual users used the docked bikes, while members have a greater usage of the electric and classic bike types. Docked bikes are available at many access points across the city, and are not available for long-term rentals, and likely come at the lowest price. The greater usage of electric and classic bikes among members likely suggest that these come at no greater cost to the members of the program and are available for long-term rental. As these bike options are likely of no extra charge to members, it would make sense that members will use these bikes more frequently. On the other hand, casual users are likely visitors to the city and want a quick way to get around the city, so using the docked bikes is a perfect option.

**3. How do your findings relate to your original question?**

- These findings help show potential relationships in the varying uses of the bike-share program between casual users and members. With these findings, I am better able to provide data-driven insights into potential ways to recommend how the company can boost the number of users that become members of the bike-share program. The findings show that members are leading the overall usage of the bike-share program and, combined with the greater number of rides, show how the members use the program differently than casual users. This gives a data-driven insight into what type of audience would more likely be inclined to use the membership of the program.

**4. Who is your audience? What is the best way to communicate with them?**

- The audience for this analysis is Lily Moreno and the marketing department of the company. To communicate these findings with the marketing team, it would be best to schedule a meeting for the presentation of the findings that would include the key players in the development of programs to boost member percentages within the bike-share program

**5. Can data visualization help you share your findings?**

- The visualizations above follow the order of the analysis in the Analyze section. Each visualization is a chart that represents each data frame in the analyze step. The visualizations show the analysis at a quicker glance, and in an easier to read fashion. Each visualization helps paint the picture of what the data is showing and helps create a better understanding of the findings.

**6. Is your presentation accessible to your audience?**

- The PowerPoint presentation will be available through the Google Cloud and will be able to accessed by all stakeholders and members of the marketing team at any time. I have taken the liberty to mark the file as view only to ensure that the original presentation is not accidentally altered in any way.

**Part 6: *Ask***

**1. What is your final conclusion based on your analysis?**

- This analysis provided four major insights into the data from the 2022 year of the Cyclistic bike-share program. First, there are more members than casual users of the program, fifty-nine percent in relation to forty-one percent. Next, members of the program had a greater number of rides per day of week, day of month, and month than casual users. In opposition, casual users had longer average ride times versus members per day of week, day of month, and month. Lastly, the data shows that members of the program used casual and electric bike types more than casual users, while casual users were the only ones to use the docked bike types. These insights helped me generate a theory as to suggest why these trends show in the data. First, members of the program having a greater number of rides and averaging a lower ride time versus casual users suggests that members of the bike-share program may be individuals that live within the Chicago area. In major cities across the United States, many individuals do not have the necessity for a car, but will still need a primary method of transport to get across their city. In this analysis, it is likely that the members of the bike-share program are people who live in Chicago and use this program to obtain an affordable primary means of transportation. This is further supported by the total number of rides per bike type. Members boast a much larger usage of classic and electric bikes versus users. This is likely due to these bike types having an upcharge for casual users, and are likely used for long-term rentals. Members, on the other hand, are likely able to rent these bikes at no upcharge, and are able to rent them for greater amounts of time. On the other hand, Casual users have a much greater average ride time, while have a lower number of total rides. This suggests that casual users are likely visitors to the city, and are needing a one-time method to be able to get across the Chicago area. As these users are on a pay-per-hour style of renting, they likely want to ensure that they spend most of their time riding the bike rental that they paid for.

**2. How could your team and business apply your insights?**

- There are three methods of advancements that I am going to recommend to the audience. First, I would start by getting a better understanding of the members and casual users of the bikeshare program. This can be through an email or mail service survey that would include items such as: reasonings for choosing to become a member, what reasonings the user is using the bike share program for, where the user is from, and other demographical information that can provide more insight into the audience of the bike-share program. Next, I would recommend that a special offer be made to all residents within the Chicago area, such as a percentage off memberships, for becoming a member of the bike-share program. This could be boosted by providing scientific evidence that shows the benefits of using bikes over cars as a primary means of transportation within large cities. Providing a discount, plus reasonings as to why bikes are better for inner-city travel over cars could boost the amount of people that are willing to become members of the bike-share program. Lastly, members of the marketing team could create an automated email to be sent to casual users after their "Xth" number of rides. When causal user is a recurring user of the bike-share program, they may not be aware of the benefits that a membership provides. This could be a great way to provide an "exclusive discount" for casual users to become a member of the program after a given number of rides.

**3. What next steps would you or your stakeholders take based on your findings?**

- For my team, we could take a look at datasets from years prior to 2022 and run an analysis to find the differences between each year. This could be used to back track with previous adjustments to boosting the number of members on the bike-share program to see which methods provided the largest increase in members within a year. These methods could be taken to the stakeholder to generate updates or modifications to previous programs to ensure that a greater number of users are likely to become members of the program.

**4. Is there additional data you could use to expand on your findings?**

- The findings of this analysis could be compared with other bike-share programs from major cities that have relative size and population numbers in comparison to Chicago. These datasets could show if other major cities are seeing similar trends, or if these cities are showing a larger percentage of users that become members of their bike-share programs.