**Data Analysis Case Study #1 Bike-Share**

As part of the Google Data Analytics Certification, it is recommended to try a case study. A case is provided by the course, in analyzing and providing a data report for a fictional company Cyclistic to help them in their marketing and maximizing their memberships. I will follow the six steps provided in the Data Analysis Process to make and report my findings.

**ASK:**Context: Cyclistic is a bike-share company in Chicago who’s director seeks to maximize annual memberships. I am a junior analyst working on the marketing analyst team working on how to market casual riders into members.   
  
**Business Question:**   
How can Cyclistic improve its marketing and engage its customers to joining the annual membership? The Director has asked to find out how to maximize the annual memberships and how to convince its customer’s to join.

* Best ways to entice customers to joining the annual membership
* How best to market the membership to them and give compelling reasoning
* What would a casual and membership rider need for their activities?

**Our Stakeholders Include:  
  
Lily Moreno –** Director of Cyclistic Marketing who makes the new campaigns and initiatives.

**Cyclistic Marketing –** The team of data analysts who record, organize and analyze data for the company to make changes and strategies. I am a new part of the group and help them.

**Cyclistic Executive –** The higher ups of the company who will commit to the new marketing plan provided the evidence of the report supports their actions. They are the main stakeholders and the people we report to and will change the strategy with compelling reports and visualizations.

**Cyclistic Company –** A bike-sharing company centered in Chigago first launched in 2016. Now currently operating 5,824 bicycles and 692 stations across the city in a connected system. Cyclistic offers a wide variety of bikes, with most customers riding for leisure but about 30% using it for work. Marketing has been recently broad, marketing to the common person with flexible pricing plans for daily or single rides. However the financial analysis finds annual membership to be more profitable.  
  
**Report Requirements –**

1. A clear summary of the business task

2. A description of all data sources used

3. Documentation of any cleaning or manipulation of data

4. A summary of your analysis

5. Supporting visualizations and key findings

6. Your top high-level content recommendations based on your analysis  
  
**Relevant Questions:**

* How do casual and annual riders differ? What do either of them want from their experience?
* How could Cyclistic convince casual riders to join an annual membership on a large scale?
* What is the best way to market annual membership to become more enticing and take advantage of multiple media platforms?
* Would there be any unexpected outcomes from said marketing?

**PREPARE:**  
The data is provided by Cyclistic past archived data. The historical data will be used to analyze trends and patterns to make conclusions going forward. Since Cyclistic is a fictional company the data is assumed to be accurate. The data has been linked to Motivate International Inc. as a public dataset so anyone has access to it. Motivate International is a real and official company in New York that like Cyclistic is a bike-sharing company so the data should translate perfectly into the scenario. According to the linked License Agreement it is stated

*License. Bikeshare hereby grants to you a non-exclusive, royalty-free, limited, perpetual license to access, reproduce, analyze, copy, modify, distribute in your product or service and use the Data for any lawful purpose (“License”).*

***Prohibited Conduct. The License does not authorize you to do, and you will not do or assist others in doing, any of the following***

*Use the Data in any unlawful manner or for any unlawful purpose;*

*Host, stream, publish, distribute, sublicense, or sell the Data as a stand-alone dataset; provided, however, you may include the Data as source material, as applicable, in analyses, reports, or studies published or distributed for non-commercial purposes;*

*Access the Data by means other than the interface Bikeshare provides or authorizes for that purpose;*

*Circumvent any access restrictions relating to the Data;*

*Use data mining or other extraction methods in connection with Bikeshare's website or the Data;*

*Attempt to correlate the Data with names, addresses, or other information of customers or Members of Bikeshare; and*

*State or imply that you are affiliated, approved, endorsed, or sponsored by Bikeshare.*

*Use or authorize others to use, without the written permission of the applicable owners, the trademarks or trade names of Lyft Bikes and Scooters, LLC, the City of Chicago or any sponsor of the Divvy service. These marks include, but are not limited to DIVVY, and the DIVVY logo, which are owned by the City of Chicago.*

Since this Case Study is for a fictional company using the data just for analysis purposes for a Case Study it does not violate the license agreement.

The data itself is comprised of about 90 .CSV files in long format. It covers monthly trip data from April 2020 to April 2024. It also includes Quarterly reports in a different format from 2013 to 2019. The ones from 2013-2019 have a largely different format and information. With that and that Cyclistic would want the most relevant trends from the past few years of relevant transit and digital marketing strategy, information from 2020-2024 will be considered. Especially after the 2020 Global Pandemic has drastically changed the way people commute and interact. The main focus of the Business task is to prioritize analysing the data from the last 12 months.

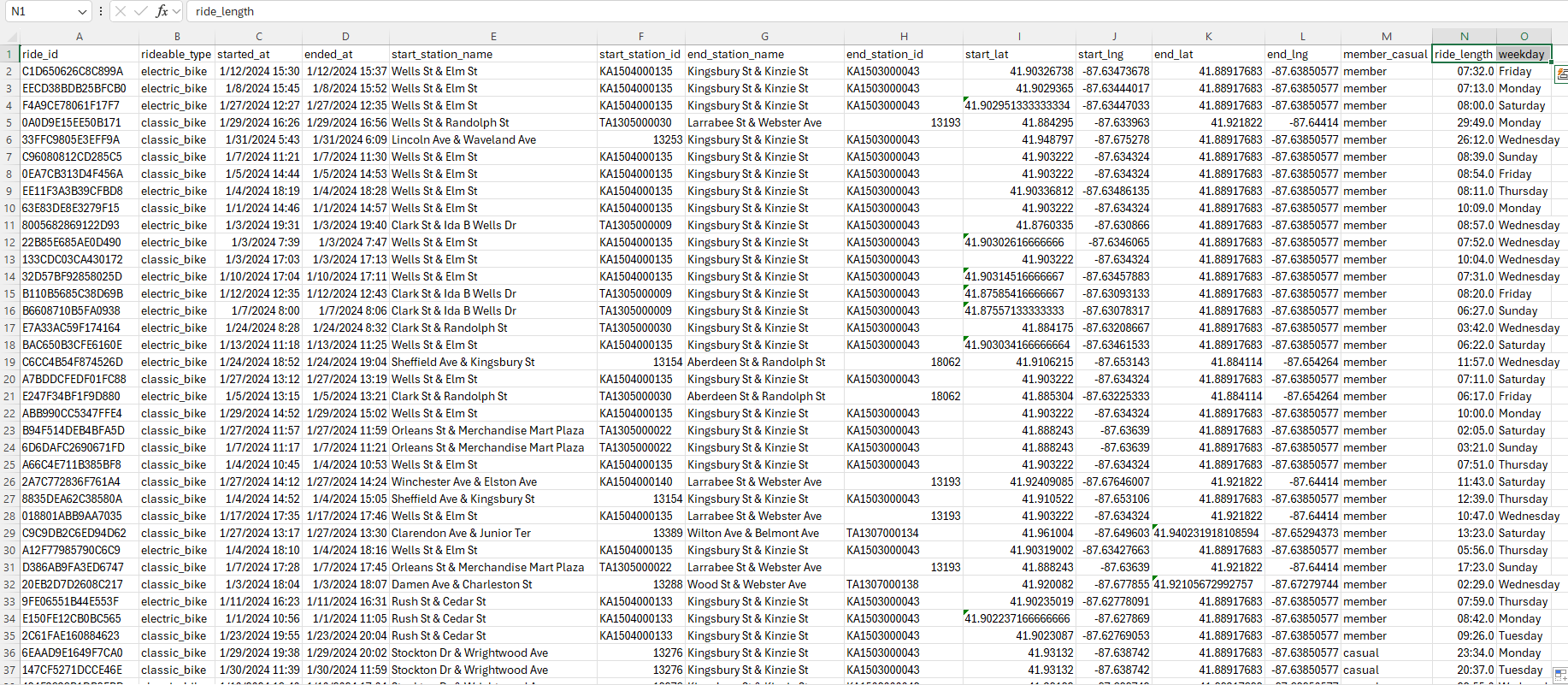
Information and columns are divided into

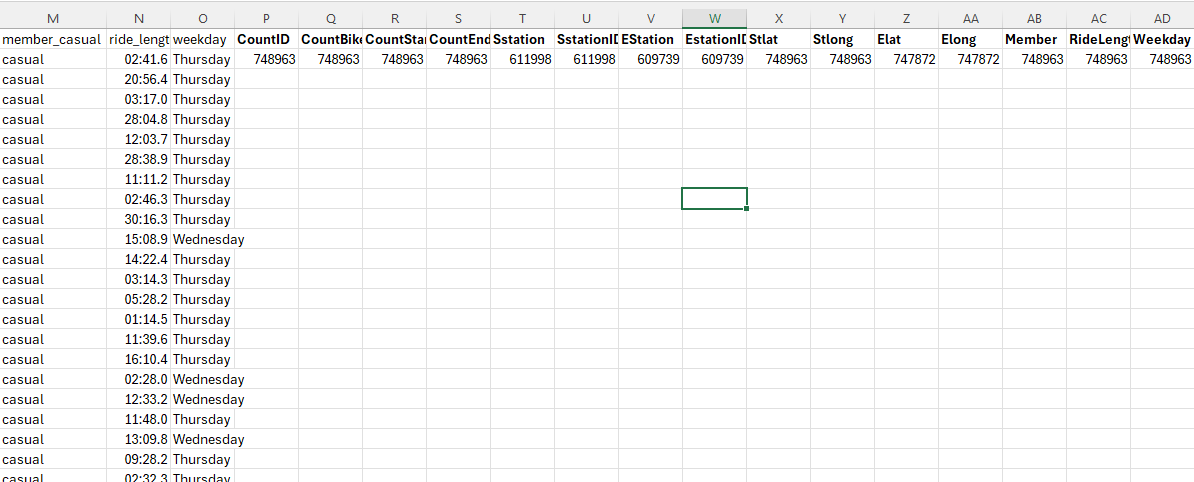
* Rider\_id – The numeric ID of the bicyclist
* Ridable\_type – The type of bike the rider used
* Started At / Ended At – The time it took for the rider to start and end their bike ride
* Start Station / End Station – The station that the rider started at and the station they ended with.
* Start Station / End Station Id – The Id number of the station a rider starts and ends with
* Start Lat / Start Lon – The Latitude and Longitude Global coordinates the rider started at.
* End Lat / End Lon – The Latitude and Longitude Global coordinates the rider ended at.
* Member / Casual – States if the Rider is identified as a casual rider or member.

The given tables provide information on the monthly interactions and distance of the thousands of riders that engage in Cyclistic’s services. The most important information is describing if a member is casual or a member as its needed to separate the behaviors and statistics between the two groups. The Started At / Ended at provides the length of time between the rides and get an idea of how long a rider was using the service and Start/End latitude and longitude for the distance they traveled. The listed stations could be used to find which is the most popular stations and where marketing should be focused on, and ridable type could show the preference between bicycles available.  
  
**PROCESS:**  
I will be using Spreadsheets to visualize and process the dataset. But the data is too large for complex analysis so I will also be using SQL and BigQuery.   
  
Before going to the Analysis I will need to clean and process the data to ensure its accuracy.

The Data Tables being used are divvy tripdata 202408 to 20230408, or 12 tables. Going through them on Spreadsheets, it seems that chunks of the data are missing which seems like a problem with Excel loosing data on such a massive file, but going on a .CSV editor shows that the data in these entries aren’t filled at all. These entries are in the Starting Station and Ending station where it seems the riders never had archived start and stops. Though it is largely irrelevant as the start and end times are intact.

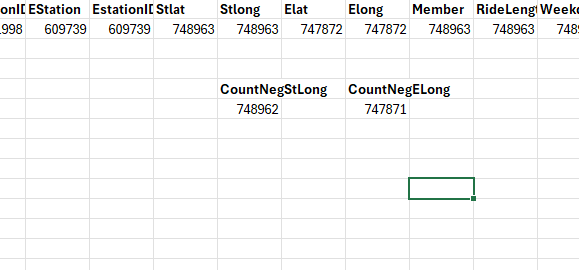
We will need to calculate the ride-time and the weekday of the ride. We calculate the Ride-Time by subtracting the End Time with the Start time and formatting it to a Minute-Second-Millisecond format. Weekday is created using the Weekday formula in Spreadsheet to calculate with the given date and a custom if-statement to change the numbers to the corresponding dates

  
This is done to all 12 of the CSV. Files we will be using to update all the data we need. We also need to check the accuracy of each of the tables. So using a CountA function we can check each column to return the count of items in each cell and compare it to the total row number. An example in one of the sheets includes:

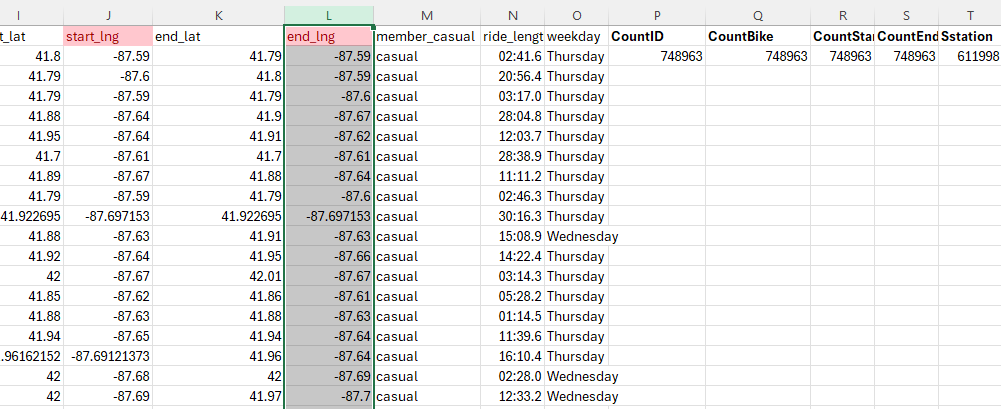


As the results showed most of the data entries are consistent and fully take up the rows and columns. The only ones that miss some are the Start and Ending stations and some of the Ending Lat/Long as previously stated. As they are not too relevant to the business task and removing the rows they contain would remove important information, the document does not need to change.

I would like to be thorough, and check if any of the negative numbers in the Longitude Columns might've been inputted incorrectly, since it is a negative number. So I made a function to count every row in the Longitude Columns that contain a negative number using =COUNTIF(J:J, "<0").



From the results we see that there is one row without a negative longitude. We use conditional formatting to find the rows that have the wrong information.

  
But we see it counted the Header row as the non-negative row. This is a very easy mistake to make and its nice to clarify it early on but it will clear up confusion going forward.

Doing the same to the rest of the tables will provide similar results, with the number of rows changing due to the different numbers of riders and customers per month. But the proportions are still always consistent so there are no missing pieces of important data in the tables.

**ANALYZE:**

Now that the tables have been processed it's time to go to the analysis step. Now that the tables have been cleaned we can upload them to BigQuery to do more detailed analysis in SQL. Due to BigQuery's formatting rules, some of the date and time columns have to be reformatted, such as changing start\_time and end\_time to a YYYY/MM/DD format and ride\_length to an HH/MM/SS format.



A screenshot of a computer

Description automatically generated  
  
The Business Task is to compare the data between casual and member riders of Cyclistic and make conclusions by it. So we need to come up with statistical questions to search for and compile our findings.

**What we need to look for are**

* The longest ride length times for Casual and Member Riders
* The average ride length for Casual and Member Riders
* See if Members or Casual Riders would have longer ride lengths
* The most popular bike types for Casual and Member Riders
* The most popular stations for Casual and Member Riders
* What day of the week is most popular for riders
* How many rides are done each day of the week

We will use Spreadsheets and BigQuery to find the answers and data for these questions. For later reports and visualizations we will also store our findings on a spreadsheet.

A white sheet with lines and text

Description automatically generated with medium confidence

To calculate the Longest Ride, we use BigQuery to return the MAX of the entire 2024-08 table for both the casual and member riders. We denote it by *SELECT MAX(ride\_length)* and adding the *WHERE member\_casual = ‘casual’ or ‘member’*We use a subquery to find the two datasets at once to save time.

A screenshot of a computer

Description automatically generated  
Though it seems that BigQuery has trouble with the ride\_length as a TIME value and comparing it. We’ll try using it in Spreadsheets instead. We’ll re-format the ride\_length to be in [H]-M-S format to better deal with large outliers. And we’ll combine the IF and MAX functions to find the MAX of the row if the member\_casual value is “member” or “casual”. Doing so gives us.

A screenshot of a computer

Description automatically generated  
We can even confirm this by going to the “Sort and Filter” ribbon above and sorting the column from Largest to Smallest.

A screenshot of a spreadsheet

Description automatically generated

The sorted data matches the formula. So they’ll get added to the Main spreadsheet.

With the data we now insert it into the main spreadsheet to save it and do a similar function next but instead finding the average ride\_time. SQL has trouble finding the Average of a Time function so we’ll use Spreadsheets and the AVERAGE\_IF function to find the average ride\_length if member\_casual = ‘casual’ or ‘member’

*A screenshot of a computer

Description automatically generated*

Moving on we need to find the most popular bike sold by Cyclist that month. And using SQL we can use COUNT (\*) to count every value and sort it by the rideable type in descending order to get this.

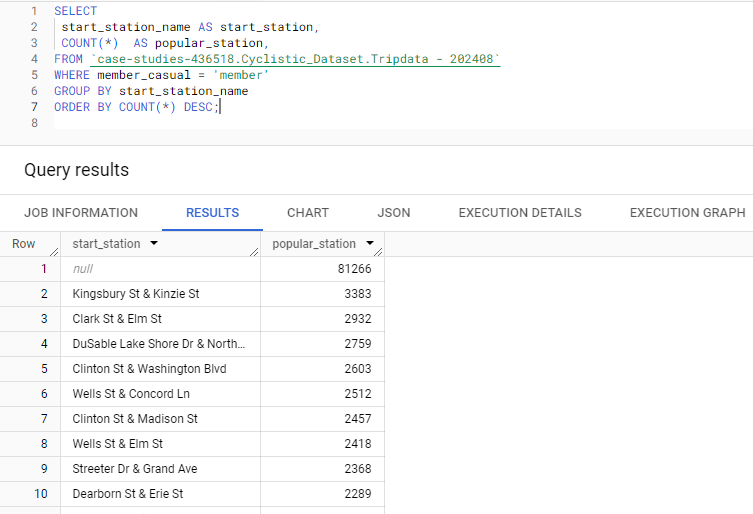
A screenshot of a computer

Description automatically generated  
We cannot use a subquery to find both the member and casual value in the same query so we would have to do another query for member values

A screenshot of a computer

Description automatically generated

We use a similar query format to find the most popular start and end station of the month. Just changing the values of rideable\_type to start\_station\_name



And doing the same to end\_station\_name

A screenshot of a computer

Description automatically generated

We find the busiest weekday in a similar format, as we change the value to analyze the weekday in the same vein.

A screenshot of a computer

Description automatically generated

Now we can compile all the relevant information in the main spreadsheet to fill in the first two rows. It now looks something like this.

A screenshot of a spreadsheet

Description automatically generated  
Now we must do the exact same steps for every other table in the dataset, only needing to change where the data is being taken from. Once completed our full sheet looks like this.

A screenshot of a computer

Description automatically generated  
  
From the table we can already see some notable patterns and trends.

* Casual Riders have consistently higher times in longest and average rides compared to member riders
* Electric bikes have become a much more common bike type in recent months compared to last year
* Casual members consistently start and end at Steeter Dr & Grand Ave implying full cycle trips and Members are more varied but Kingsbury & Kinzie St has become more popular recently
* Casual members consistently have the most rides on Saturdays and member riders consistently ride the most on weekdays.

To better study the weekday trends between riders, I also used the same steps to gather the Weekday column to make a separate table to compare the most active days for the 12 months for the two riders. Where Green is Casual and Blue is Members.

A screenshot of a calendar

Description automatically generated  
Some trends are better seen here and went into more detail.

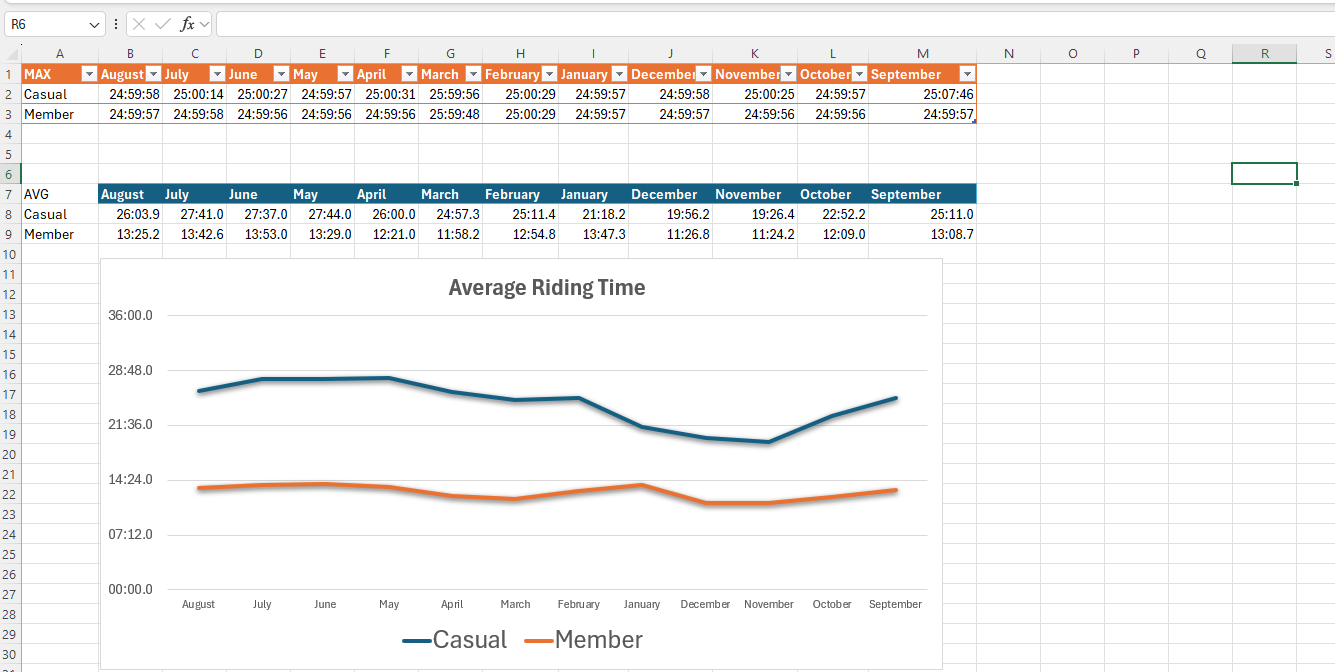
* Casual Riders largely prefer weekend rides, with rides consistently being the most popular on Fridays, Sundays, and especially Saturdays. With the exception of January and December.
* Casual Riders ride the least during weekdays largely Tuesdays and Thursday
* Member Riders consistently prefer Weekday rides, preferring Tuesday and Thursday rides.
* Member Riders usually ride the least during Sunday and Saturday

This gives us a much better picture of the trends between riders and why most riders pick the ride type they do. It implies that Casual riders go more on weekdays for recreational use and potentially tourism, as Member riders seem to largely use it to go to work. This would also match the parallels to the Max and Average columns on the main Spreadsheet where Member riders have less ride time than Casual Riders.   
  
But its best that we take this new data and make graphs and visualizations to best show our findings and trends and to share to stakeholders.

**SHARE:**

Excel provides numerous graphical options for datasets in its spreadsheets so we will use them for ease of access and to format them well in Power Point where we will present out point, with some stylization.

First we try and make a graph for the Max and Average Ride Length. However this is difficult since the Max ride lengths are so similar to each other and have barely any deviation asides from one massive outlier in March. Putting them in a graph does not give much valuable information and can be very confusing to stakeholders. However Average Ride Length is more useful not only to get the average time sample for the analysis, but also that the deviation and comparison is much more obvious to put in a line graph comparing the casual and member ride lengths. We take all of the data from the previous spreadsheet and make a new spread sheet just for it, adding two tables of the Max comparing the Casual and Member riders and the Average comparing the member and Casual Riders. Using a simple graph in Spreadsheets we come up with this:



The graph does make trends more obvious, notably the harsh distinction between the casual and member ride lengths and how the months affect the average. But the important points will be stated in the next step.

Then we find the most popular bike types. We use the extended data we got from the previous Analysis step to get the number of both bikes used in the year and compile them in a separate graph. We organize them by the month and make two rows each of the number of classic and electrical bikes sold for casual and member riders and use a Clustered Column chart to compare the two.   
A screenshot of a graph

Description automatically generated  
Here we see notable trends, largely the dip in sales in the winter seasons but more relevantly, the larger increase in electric bike sales in recent months compared to later months which had dominantly classic bike.   
  
We try and make a graph of the Popular Start and End Stations. However it is hard to do it as we can’t use numerical data for each of the rows. Instead we’d have to compile the most common stations on both the Start and End and make a bar graph of those. With the distinction of the trends of the tables to add more context. Then we get:

A screenshot of a computer

Description automatically generated

We see more notable trends. We know that Streeter Grand Ave is continually the most popular start and end station for Casual Riders with the exception of June and January. The other data is derived from the Members who have more diverse start and end stations but dominantly prefer Clinton St & Washington Road and Kingsbury & Kinzie St.

And we need to make a graph for the Weekdays to show their popularity. We want to add the numeric values to compare the days to each other. So we make a new table and add all the numeric values from the weekdays to make a detailed clustered column chart for each day of the week in each month and split them by casual or member.

A screenshot of a computer

Description automatically generated

We see some obvious trends like the dominance of Saturday on the chart, which is because of Saturday dominating the busiest Casual days with a few exceptions. As other weekdays, namely Wednesday and Thursday are more common for Member riders, as the table and graph show a clear pattern of Casual riders preferring Weekends and Member Riders preferring Weekends.

So we have compiled our data and have made some connections helped with some visuals given by Excel. We’ve helped to answer the business question to find trends and patterns with the Casual and Member riders and hopefully can come up with smart business strategies so that Cyclistic has to take advantage of it. So we can showcase these points and data using the standard industry practice of compiling them in a PowerPoint presentation and present it. We will keep in mind that we are presenting to Cyclistic’s Marketing and Lily Moreno to best show the trends and data and encourage them to make marketing decisions.   
  
**ACT:**

We’ll put together a PowerPoint presentation and order it based on how we progressed our findings. We address the Business question, then give context on how we gathered the data, then present our individual points and end it with conclusions from the data and how we can best apply it to marketing strategies and end with questions. To view the Report, please open the Cyclistic Data Report Powerpoint Presentation to conclude the case study.