Introduction:

This case study is focused around a fictional bike sharing company called Cyclistic. The company runs a bike-share program that features over 5,800 bicycles and 600 docking stations. They have both casual riders as well as membership plans and offer various types of bicycles. The goal of the case study is to use Cyclistic’s rider data to analyze how casual riders differ from member riders in order to think of marketing strategies to increase membership sales.

Questions:

* How do annual members and causal riders use Cyclistic Bikes Differently?
* Why would causal riders buy Cyclistic annual memberships?
* How can Cyclistic use digital media to influence riders to become members?

Business Task: Analyze trip data from the past year to identify differences between casual riders and members in order to market towards causal riders to increase annual memberships.

Data Sources:

The data sources used for this case study consisted of twelve csv files, all containing rider data for each month for the past year. Each data set represents one month and all contain the same attributes. These consist of:

* Ride\_id: alphanumeric code representing a single ride
* Rideable\_type: type of bike used for the ride
* Started\_at, ended\_at: date and time which the ride began and ended
* start\_station\_name, start\_station\_id: name and alphanumeric id for starting station
* end\_station\_name, end\_station\_id: name and alphanumeric id for ending station
* (start/end)\_lat, (start/end)\_lng: starting and ending latitude and longitudes
* Member\_casual: type of rider

Each of these datasets consisted of thousands of rides, but differed in size depending on month as some months contained significantly more rides than others. The data also contained various types of errors and null values throughout which must be accounted for.

For my analysis, I decided to look at how the proportion of rideable types differed between member and causal riders, as well as the total number of each, and how each used the bikes differently through the ride times of each. I also wanted to look at how each type used the bikes differently throughout the year as looking at the use trends could give insight into when to market more heavily.

Before starting the process of data cleaning, I looked at the datasets to see how complete they were in order to decide which attributes would be most useful in my analysis. Looking at the data, each ride was mostly complete except for the start/end station name and id. These fields were often null and provided little context for the analysis I would be performing so I decided not to include them in the data I would be working with. Along with these, I also decided to exclude the starting and ending longitude and latitude. These would prove much more useful if I were to be doing a geographic analysis but since I was not, I removed these columns as well. The final column I removed was the ride\_id. This was a unique alphanumeric code which would not be useful in the final dataset as the id would be different for each ride.

Data Cleaning:

All of the data cleaning for this case study was done using excels built in functions. The first step in the data cleaning process was to identify and delete null values. The first column I did this for was the member\_casual column. I saw this column as the most important as it identified what to compare and was the basis for this analysis. Because of this, I decided to remove any rows which contained a null value for this column. This was done by filtering and deleting all of the null values. Next, I needed to ensure that the values, “member” and “causal” were the only two for this column so used excels built in filter function to check. Thankfully the data was already complete in this regard. These two steps could be replicated for the rideable type column as it was a column with only three possible values. The next step of the data cleaning phase was to look at the ride lengths. In order to clean these columns, I first did the same steps as before to delete null and error values from each column. After this, I decided to add another column which consisted of the actual length of the ride. Once I created this column, I then looked for value errors within it. There were many occurrences of the starting time being after the ending time which were then deleted as this would not be possible. Along with deleting these columns, I sorted this column by length to see how long the rides could potentially last. When doing this, I found rides that were logged at being ridiculous lengths like 36 hours long. I assumed that this error was likely either a system error where the end was logged incorrectly, or a rider that forgot to end the ride and decided that they were not representative of the population. I decided that any ride over 8 hours long would be considered an error. The steps I took were for every csv file used.

Preparation:

After cleaning the datasets, I loaded them all into power bi to prepare for my analysis. I created a single table that combine the data from the twelve datasets. This would allow me to perform operations on all of the data at once rather than each of the twelve on their own. I then had to create three new columns; one column extracting the day of the week from the starting date and saving it as a number 1-7, one column extracting just the date from the starting date, and one column converting the ride time from being a decimal value representing the proportion of the day, to the actual total minutes of each ride. The commands used were:

Date = DateValue([started\_at])

Ride Length minutes = [ride\_length] \* 60 \* 24

Next I needed to create the secondary tables which summarized the data to be visualized. The first table I made was the daily ride lengths table, which consisted of the original data table grouped by each day, in order to give daily statistics. This would include the date, member type, average ride length, member count, and counts for each type of bike. The command used was:

Daily Ride Lengths = SUMMARIZECOLUMNS('Cyclistic-trip-data-year'[Date], 'Cyclistic-trip-data-year'[member\_casual], "Average Ride Length", AVERAGE('Cyclistic-trip-data-year'[Ride Length minutes]), "rider count", COUNT('Cyclistic-trip-data-year'[member\_casual]), "Electric Count", COUNTX(FILTER('Cyclistic-trip-data-year', 'Cyclistic-trip-data-year'[rideable\_type] = "electric\_bike"), 'Cyclistic-trip-data-year'[rideable\_type]), "Classic Count",  COUNTX(FILTER('Cyclistic-trip-data-year', 'Cyclistic-trip-data-year'[rideable\_type] = "classic\_bike"), 'Cyclistic-trip-data-year'[rideable\_type]), "Docked Count", COUNTX(FILTER('Cyclistic-trip-data-year','Cyclistic-trip-data-year'[rideable\_type] = "docked\_bike"), 'Cyclistic-trip-data-year'[rideable\_type]))

The next table I created was the riding statistics for each day of the week. The command used was:

Weekly Distributions = SUMMARIZECOLUMNS('Cyclistic-trip-data-year'[day\_of\_week] ,'Cyclistic-trip-data-year'[member\_casual], "Number of Riders", COUNT('Cyclistic-trip-data-year'[member\_casual])/52, "ride length", AVERAGE('Cyclistic-trip-data-year'[ride\_length]))

After creating these tables, I was ready to visualize the data and perform my analysis. The first thing I wanted to visualize was the overall distribution of rider types throughout the year. I chose to look at this because it gives insight into the overall proportion of members vs casual riders, as well as how the amount changes throughout the year. I created 4 visualizations using the rider counts data. These included a pie chart showing the overall proportion of each type, a scatter plot showing the amount of each type daily throughout the year, a bar chart showing the monthly rider counts, and a stacked bar chart showing the weekly averages. From the visualizations, we can see that the amount of casual vs member riders are relatively similar. The total amount of each has the amount of members accounting for only around ten percent more at ~59% of the total population. While this may look good, as the goal is to gain more memberships, it also shows that nearly half of all riders are casual, indicating that something needs to be done to draw in more memberships. Looking at how these counts are distributed throughout the year, the two follow similar trends with the lowest amount of each being during the winter months early in the year, with the amount increasing into the spring and summer before falling again during the fall. While the bar and scatter plot both show a similar overall trend in rider activity, the scatterplot shows the amount of casual riders varying much more heavily in the later summer months than the number of member riders. This could indicate that while the amount of casual riders may increase due to better weather, the consistency of this increase is weak, possibly showing many new one time riders occurring but less repeat riders. It also shows that the amount of casual riders could be impacted by some outside factor more than member riders. Looking at the weekly averages, the member riders stayed relatively similar throughout the week, only increasing on Saturday, while the causal riders have higher values on the weekend and lower on the week days. This is expected as the casual riders most likely are riding these bikes during their free time or for some form of short transportation on a weekend rather than a consistent method of transportation.

The next thing I wanted to look at was how the overall ride lengths differed between casual riders and members. To do this I created similar plots to the rider counts visualizations, only replacing the pie chart and stacked bar charts with regular bar charts as we were no longer focused on proportions. From the visualizations, it was clear that the casual riders were riding the bikes for much longer than the members. Over all 4 visualizations, the casual riders consistently rode for much longer on average than the members. This was surprising, as the member ship plan would likely incentive more use of the bikes overall, however this didn’t appear to be the case. The ride lengths were relatively consistent throughout the year however. While there was still an increase in the summer months, the overall trend was much less noticeable than for the rider counts. One thing I did notice was that a similar trend was seen in both the weekly bar graphs for the casual members. The weekends had the highest average ride length with the length getting shorter in the middle of the week, just like the rider counts. The trends seen in the ride length data could indicate that because casual riders might only be using the bikes for specific transportation on occasion, the length is likely greater as casual members would only use the bike to get to locations where walking would take too long. This would explain why the member riders tended to have lower lengths as they would likely be using the bikes more consistently and for a much shorter distance.

The final analysis I performed was to look at the differences in bicycle types between members and causal riders. To do this, I created two plots for each type of rider, a pie chart showing the overall proportion of each type of bike, and a stacked column chart showing the counts throughout the year. The most noticeable thing I saw was that the docked bikes were only being used by casual riders and not members. This proportion was small compared to the other two types of bikes but was only seen in the causal data, indicating that there might be little incentive for people who want to use these kinds of bikes to get annual memberships. Besides this, the proportion of electric bikes to classic bikes differed between casual and member riders as well. The casual riders heavily preferred the electric bikes with them accounting for 54% of the total rides that year, with the classic bike only accounting for 38%. This was not seen in the member riders as the proportion overall was nearly equal with the classic bikes actually being greater than the electric bike by about 2%. This likely strengthens the idea that casual members are using the bikes mainly as a method of transportation. The higher proportion of classic bikes for member riders indicates that people with memberships are not only using the bikes for easy transportation but likely for exercise as well. These proportions were relatively consistent throughout the year, showing the same overall trend as the rider count data with the amount being higher in the summer months.

Recommendations:

From the analysis I performed, my recommendation would be to offer more incentives for casual riders to get memberships. This could include discounts on longer rides or a deal where members could gain “free” ride time based on how long they have been riding. This would incentivize people to buy memberships even if they were only using the bikes for transportation rather than exercise. I also would target people using the docked bikes by creating a discount for the use of these bikes when purchasing a membership. In terms of marketing, the best strategy would be to save the marketing for the later months. The summer months had the most casual riders by far so marketing these deals as something like a “summertime discount” would reach the most customers and incentivize the most people to get an annual membership even if they were only going to ride for the summer. Also, creating some incentive focused around the weekends would draw many new customers in. For example, if Cyclistic was to create a program where you get bonuses based on total ride time, they could increase the points earned on Fridays and Saturdays. Finally, since the members appear to use the bikes for exercise rather than just transportation, you could create marketing campaigns in partnership with fitness apps or devices like Fitbit, where you can connect the Cyclistic ride to whatever technology you were using to incentivize those focused on fitness towards memberships.