1. **Introduction**

In this document I identify the possible strategies to analyze the datasets.

I propose strategies and indicate what information can be gained from doing each particular analysis.

1. **Plotted Question**
2. **How do annual members and casual riders use Cyclistic bikes differently?**
3. **Description of the datasets**

Column names and explanation:

1. **ride\_id**: identification number of the ride.
2. **rideable\_type**: type of bicycle used.
3. **started\_at**: date and time of the start of the ride.
4. **ended\_at**: date and time of the end of the ride.
5. **start\_station\_name**: station name of the start of the ride.
6. **start\_station\_id**: id of the station of the start of the ride.
7. **end\_station\_name**: station name of the end of the ride.
8. **end\_station\_id**: id of the station of the end of the ride.
9. **start\_lat**: latitude coordinate of the start of the ride.
10. **start\_lng**: longitude coordinate of the start of the ride.
11. **end\_lat**: latitude coordinate of the end of the ride.
12. **end\_lng**: longitude coordinate of the end of the ride.
13. **member\_casual**: type of membership.
14. **Initial observations about the dataset**

The **ride\_id** is a unique association of a number to a ride, this has nothing to do with knowing if to the same person corresponds many **ride\_id** values, in this sense this column will not give us valuable information by itself. In other words, these ids could be replaced with cardinal numbers starting at 1 for simplicity.

The **start\_station\_name** and **start\_station\_id** give us a way to identify the same station, so in a way these two columns give us redundant information, we could merge them to create a unique one, we will call it **start\_station\_id\_complete**. Same thing applies to columns **end\_station\_name** and **end\_station\_id,** we will call this column **end\_station\_id\_complete.**

The columns **start\_lat, start\_lng, end\_lat, end\_lng** are the coordinates of the start station and the end station respectively, and can be used to calculate distance, velocity and we can locate nearby businesses, which can attract the users.

The columns that contain more information are: **rideable\_type**, **started\_at**, **ended\_at** and **member\_casual**.

1. **Plotting strategies**

**Explore datasets**

First, we will explore the 12 excel workbooks containing the datasets for each month of the year 2021, we will create some new columns, make some calculations and create some pivot tables to make a basic exploration of the data and to decide what would be a good and optimal way to make a deeper analysis.

**Beginning deeper analysis**

Now that we know better our datasets, we can plot our strategie. We will do this by combining different columns of our dataset and checking what kind of information can we get from that combination.

Given the analysis in the last section, we can reduce the columns to the following ones:

1. **rideable\_type**
2. **started\_at**
3. **ended\_at**
4. **start\_station\_id\_complete**
5. **end\_station\_id\_complete**
6. **location coordinates**
7. **member\_casual**

**First strategy: member\_casual + rideable\_type**

By analyzing these columns, we will know if there is a preference of the type of bike for each group of riders, member or casual riders.

**Second strategy: member\_casual + start\_station\_id\_complete or + end\_station\_id\_complete**

We will know if member or casual riders belong to a certain location or if many of them go to a certain place, for example, they could belong to a certain company located near one of our stations, or maybe many of them are students.

**Third strategy: member\_casual + started\_at and + ended\_at**

With this information we can know if there is a preferred season of the year, preferred month, preferred day or preferred hour at which member or casual riders like or have to use our service. This will be maybe useful to split the **started\_at** and **ended\_at** columns in date and hour columns, so we can work better with this info.

We also will know if there is a certain location where there is more demand of bicycles during the morning, afternoon or if many bicycles end up at a certain location at night or maybe during weekends.

**Fourth strategy: member\_casual + location coordinates + started\_at and + ended\_at**

With this information we may know if member or casual riders prefer or have to use our service for traveling long or short distances, for example, they could prefer traveling longer distances or weekends if they want to go camping or to a park.

1. **Deciding what tools we will use for each strategy**

**First strategy: member\_casual + rideable\_type**

For this strategy we need the twelve months merged into a single dataset, since this dataset will be quite large, we will use SQL to join the datasets and then use R to create the visualizations.

**Second strategy: member\_casual + start\_station\_id\_complete or + end\_station\_id\_complete**

Same situation as the first strategy, so we will use SQL and R.

**Third strategy: member\_casual + started\_at and + ended\_at**

Since here we will consider information by month it will be convenient to make some calculations in each of the twelve workbooks of excel and then use SQL to merge the new tables created and then use R to create the visualizations.

**Fourth strategy: member\_casual + location coordinates + started\_at and + ended\_at**

This is a situation similar to the first and second strategies, so here we will use SQL and R.

1. **Order of Action**

Given the complexity of our strategies, the most convenient way to do the research is as follows.

1. **First strategy.**
2. **Second strategy.**
3. **Forth strategy.**
4. **Third strategy.**
5. Create a Dashboard

We will use Tableau and Markdown to make the research report and dashboards.