## Case Study Roadmap - Ask

Guiding questions

● What is the problem you are trying to solve?

● How can your insights drive business decisions?

Key tasks

1. Identify the business task  
Analyze the Cyclistic’ s historical data in order to understand How do annual members and casual riders use Cyclistic bikes differently? This will help the marketing director to design better strategies aimed at converting casual riders into annual members.

2. Consider key stakeholders: The Marketing Director, Lily Moreno.

Deliverable: A clear statement of the business task

Analyze the Cyclistic’ s historical data in order to understand How do annual members and casual riders use Cyclistic bikes differently? This will help the marketing director to design better strategies aimed at converting casual riders into annual members.

## Case Study Roadmap - Prepare

Guiding questions

● Where is your data located?

● How is the data organized?

● Are there issues with bias or credibility in this data? Does your data ROCCC?

● How are you addressing licensing, privacy, security, and accessibility?

● How did you verify the data’s integrity?

● How does it help you answer your question?

● Are there any problems with the data?

Key tasks

1. Download data and store it appropriately.

2. Identify how it’s organized.

3. Sort and filter the data.

4. Determine the credibility of the data.

Deliverable: A description of all data sources used  
  
Data is located in a public shared repository ( <https://divvy-tripdata.s3.amazonaws.com/index.html> ) and organized in a monthly and quarterly .zip files containing .csv files with the data for corresponding period compressed.

I selected the most recent twelve months of data (Aug 2022 – Jul 2023) for this analysis. Each file consists of 13 columns containing information related to ride id, ridership type, ride time, start location and end location and geographic coordinates and others. I merged the files using python in Jupyter notebooks.

The information collected came from Motivate, Inc., which is the company that runs the Cyclistic Bike Share program for the City of Chicago. It must be specified that all personal information has been stripped from the data set, therefore ensuring the privacy of the users. This limits the analysis as we can’t determine if casual riders are repeat-riders or if casual riders are residents of the Chicago area. Information is released under this [license](https://ride.divvybikes.com/data-license-agreement) for public analysis.

To be able to answer the business question and understand the differences between annual members and casual riders, the available data provides good enough details. The main difference would be that casual riders pay for individual rides where regular members prefer to buy the annual subscription.

Out of the 5,723,606 rows available in the combined data set multiple rows had no details on the starting or ending stations, making it difficult to understand key preferences in user’s stations for pickup/drop-off. Remove rows with missing values and new count is 4,854,834 rows, making a total new file of 1.6 Gb. The data set was still difficult to manage as it took too much to process, so I decided to remove non-used data columns coming down to a new data file export of 519 KB instead. From this point on I’m moving into Tableau for additional data exploration and dashboard building.

## Case Study Roadmap - Process

Guiding questions

● What tools are you choosing and why?

● Have you ensured your data’s integrity?

● What steps have you taken to ensure that your data is clean?

● How can you verify that your data is clean and ready to analyze?

● Have you documented your cleaning process so you can review and share those results?

Key tasks

1. Check the data for errors.

2. Choose your tools.

3. Transform the data so you can work with it effectively.

4. Document the cleaning process.

Deliverable: Documentation of any cleaning or manipulation of data

I have chosen Jupyter notebook to initiate the data manipulation part of the analysis.

1. Started by loading, unzipping and merging all the files into 1 single data set.
2. Setting a date/time format to the startedat and enddat data columns
3. Created a new column to capture the clean values of year + month
4. Create a new column with the time difference totals for each start / end times of the trips
5. Created a new column to capture the day of the week.
6. Created a new column to capture the hour of the day.
7. Then created a heatmap as initial data exploration to get an idea on overall distribution in preferences for riders by the type of Rideable Type.
8. Remove rows where no values were provided for end starting or ending stations.
9. Selected only those columns that I will be using for the analysis and created a new data set object with these values only.
10. Exported the data set into my local system for additional exploration on Tableau.

## *Case Study Roadmap - Analyze*

*Guiding questions*

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

*● Has your data been properly formatted?*

*● What surprises did you discover in the data?*

*● What trends or relationships did you found in the data?*

*● How will these insights help answer your business questions?*

*Key tasks*

*1. Aggregate your data so it’s useful and accessible.*

*2. Organize and format your data.*

*3. Perform calculations.*

*4. Identify trends and relationships.*

Deliverable: A summary of your analysis

I have used R as advised by the guide on the data analysis for processing the information downloaded.

* First of all, I started by loading all 12 months of data into R-Studio.
* Then confirmed all columns across data sets are the same so no additional transformation before merging is required
* Check each loaded file for inconsistencies – none identified.
* Merged all 12 into 1 single data set object using bind\_rows()
* Removed columns that would not add any value to the analysis
* At this point, we have left 7 columns
  + ride\_id
  + ridable\_type
  + started\_at
  + ended\_at
  + start\_station\_name
  + end\_station\_name
  + member\_casual
* With a total of 5,723,606 records with 7 dimensions/columns
* Distribution of Type of members is
  + Casual: 2,169,555 -- 38%
  + Member: 3,554,051 – 62%
* Added a new calculated column to get the length of the rides
* Transform that last column into a numeric format
* Remove all trips with negative values, aka: errors
* Calculate overall metrics for length/duration of the rides
  + Average 1,101.573
  + Median 576
  + Longest ride (max) 3,087,684
  + Shortest ride (min) 0
* When breaking down by User Type:
  + Member:
    - Average 1,688.6136
    - Median 711
    - Longest ride (max) 3,087,684
    - Shortest ride (min) 0
  + Casual:
    - Average 743.2197
    - Median 511
    - Longest ride (max) 93,580
    - Shortest ride (min) 0
* From the above we can clearly see that members do have the longest usage/duration of usage by far.

Full project is available at Github <https://github.com/alhidalg/GoogleAnalytics_Capstone>

## *Case Study Roadmap - Share*

Guiding questions

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

*● What story does your data tell?*

*● How do your findings relate to your original question?*

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

*● Can data visualization help you share your findings?*

*● Is your presentation accessible to your audience?*

*Key tasks*

*9*

*1. Determine the best way to share your findings.*

*2. Create effective data visualizations.*

*3. Present your findings.*

*4. Ensure your work is accessible.*

Deliverable: Supporting visualizations and key findings

A graph of a graph

Description automatically generated with medium confidenceWhen reviewing the distribution of duration by user type and day of the week we notice that in both groups there is an increase in duration on the weekends, even thought is less marked on the members than we see on the casual members – group in which we do notice a significant increase in comparison for the average duration numbers during the week.

A graph of a heatmap

Description automatically generated

Also, by cross-checking the hour of the day vs the day of the week we learn that overall usage of the different types of ridable is at its highest usage rate during the late afternoon – early evening times, specifically between 4PM and 6 PM.

A graph of a number of different colored squares

Description automatically generated

Looking into the distribution of the type of rides we can realize that both, electric bikes as well as traditional/classic bikes lead the volume of choices across customers.

A graph of a graph of a number of months

Description automatically generated with medium confidence

Also wanted to understand the overall distribution of rides per month of the year, which clearly demonstrates that the highest usage of products is between the month of May through October when evaluation usage for the last 12 months of the 2023-2022 period. Being July and August the peak of utilization preference among the clientele.

## Case Study Roadmap - Act

*Guiding questions*

*● What is your nal conclusion based on your analysis?*

*● How could your team and business apply your insights?*

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

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

*10*

*Key tasks*

*1. Create your portfolio.*

*2. Add your case study.*

*3. Practice presenting your case study to a friend or family member.*

Deliverable: Your top three recommendations based on your analysis

**Top Recommendations**

Based on my findings from analyzing the Cyclistic’s riders data for the 2022-2023 data I recommend the following considerations to be taken into account when planning your marketing strategies:

1. Electric and Classic bikes are the most popular choices across users, focus on this group.
2. Highest usage seems to be over the weekends, a remarkable difference can be noted on the causal members increasing the usage of services during the weekend, taking that preference might be useful.
3. Also do consider the most popular times for bikers/users, which can be noted to be highest during the late afternoon – early evening times, specifically between 4PM and 6 PM. Triggering marketing activities around those times might become effective.
4. Launching the marketing campaigns during the months of May to October seems to be the right time as it’s the highest months in terms of historical demand – focus for July and August.