Executive Summary

Exploratory Data Analysis (EDA) of the Cyclistic Bike-Share Project

ISSUE / PROBLEM

Cyclistic is a bike sharing company based in Chicago seaking to increase the number of annual memberships.
To achieve this goal, they want to understand how casual riders and annual members use Cyclistic bikes differently.
To begin, the data needs to be explored, cleaned and structured.

RESPONSE

The data team conducted exploratory data analysis at this stage. The purpose of the exploratory data analysis was to correct issues within the dataset, such as identifying missing and duplicate values, and to identify the variables that would most likely differentiate between Casual and Member customers.

IMPACT

According to the exploratory data analysis, future models will need to account for the higher numer of Member than Casual customers.

Also, there is a high number of duplicate customers IDs. These have been removed from the dataset for further analysis.

Finally, the maximum ride length is 24h. It should be checked whether this is reasonable or the result of a wrong entry.

UNDERSTANDING THE DATA

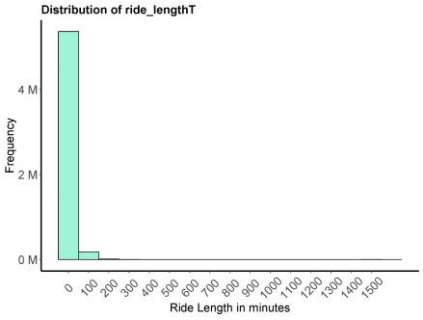
After reviewing the provided dataset, the variable *member_casual* seemed particularly useful, given the client's interest. The following screenshots shows important points required to understand the *member_casual* variable (here renamed to membership).

```
membership_counts = table(totDb_Clean$membership)
print(membership_counts)
```

```
 casual  member
2097141 3325321
```

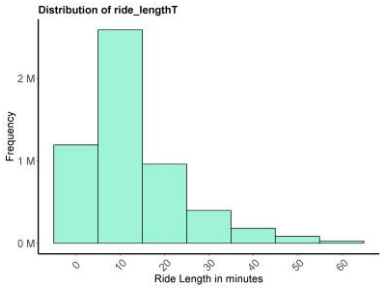
Note: The counts of each rider by membership type are not very balanced. There are 2.850.603 member and 1.846.266 casual riders, circa one million less than members.
The data have been collected between September 2021 and August 2022.
The start_station_name and end_station_name are missing respectively in 884365 and 946303 entries. As this is not relevant for all the analyses, these cases will be left in the dataframe.

The distribution of ride length (in minutes) is extremely skewed. Most values are under 60 minutes. This can be a problem for future analyses.



We suggest analyzing separately the data under and above 60 minutes when using models that can be biased by non-normal distribution.

NOTE: even after selecting data under 60 minutes the data are skewed. We suggest log-transforming this variable to further tackle this issue.



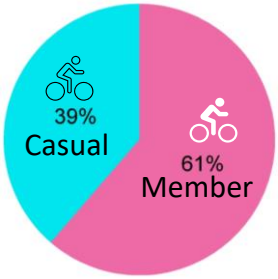
KEY INSIGHTS

- **Null values**

Over 800.000 null values were found for start_station_name. Future models should consider this.

- **Unbalanced groups**

There is an unbalanced number of Member versus Casual. When using models that can be biased by unequal number of cases between groups we have to take preventive measures.



- **Skewed data distribution**

ride_length2 (the ride duration in minutes) seems particulary important. However, this variable is not normally distributed. We recommend square root transforming the data when using this variable in models that are affected by non-normality of the data.

Executive Summary: Statistical Testing Results

Cyclistic Bike-Share Project

Project Overview

The Cyclistic team seeks to understand how casual riders and annual members use Cyclistic bikes differently. In this part of the project, the data team will conduct qualitative analyses and hypothesis testing to investigate the differences between *member and casual riders*.

Key Insights

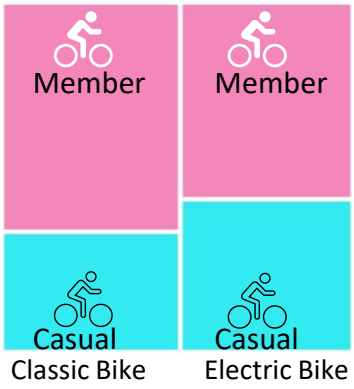
- The analysis shows that there is a difference in number of riders that use a Classical Bike and an Electric Bike for Casual and Member subscription.
- As a result, these findings suggest there might be fundamental behavioral differences between these two types of subscriptions: Casual and Member.
- It would be interesting to investigate the root cause of this behavioral difference. For example, consider:
    - Casual riders do longer rides?
    - Or, do they use the bikes for different purposes with respect to Member riders (e.g., leisure vs commuting)?

Details

The Cyclistic data team considered the membership and bikeType. One approach conducted was to examine percentage of riders that used each bike type by membership.

	Casual	Member
Classic Bike	17.27%	21.4%
Electric Bike	21.4%	27.75%

These findings show that casual riders use more electric bikes than Members. To follow this up, we compared the distributions with a Chi-squared test. Aligned with preliminary observations from the sum values, this statistical analysis shows that the observed difference between subscription types is due to an actual statistical difference in the corresponding distributions (Chi-squared = 52368; $P < 0.001$, see mosaic plot).



Next Steps

The team suggests moving forward and testing whether there is a difference in ride length between Casual and Member riders.

A t-test for ride_length by member can help investigate eventual differences in this behavior between subscription types.

Executive Summary: Statistical Testing Results

Cyclistic Bike-Share Project

Project Overview

In this part of the project, the Cyclistic data team analyzes the ride length and station used to uncover further differences in how Casual and Member riders use Cyclistic's bikes.

Key Insights

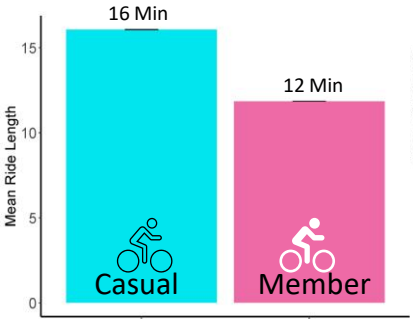
- The analysis shows that on average Casual riders use the bikes for longer rides (Mean ~16m) compared to Member (Mean ~12m).
- Visual inspection suggests that this difference is true for each day of the week.
- It would be interesting to investigate the cause of this difference.
    - Do casual riders use cyclistic bikes for recreational purposes? Or do they use for commuting?
- Investigating differences in the days riders use Cyclistic's bikes as well as the geographical areas could provide an insight to answer this question.

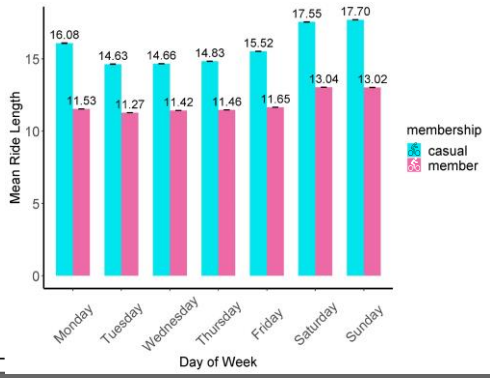
Details

The variable *ride_length* (minutes) was selected to investigate behavioral differences between subscription types.
A first approach was to inspect the mean ride length between groups.

	Mean	Standard Deviation
Casual	00:16:04	00:00:16
Member	00:11:51	00:00:39

On average, riders with Casual subscription use the bikes for longer rides (~16 minutes) versus Member (~12 minutes). A t-test shows that the difference is likely due to a real difference in the population, t (3975071) = 401.69, p < 0.001, 95% CI [4.19 11.84]. This indicates that the true mean difference between subscription types likely falls between 4.19 and 11.84 minutes longer rides for Member as compared to Casual. Visual inspection of the ride length by day of the week suggests that this difference is true for every day.





Next Steps

The team suggests moving forward and investigating whether Casual and Member riders use Cyclistic's bikes on different days of the week.
Furthermore, it should be investigated whether they also differ by geographical area of use.

Executive Summary: Statistical Testing Results

Cyclistic Bike-Share Project

Project Overview

In this part of the project, the Cyclistic data team analyzes in which days of the week and month of the year Casual riders use Cyclistic bikes compared to Members.

Key Insights

- Member riders use Cyclistic bikes the most on Wednesday whereas Casual riders use it the most on Saturday.
- In general, Members use the bikes more during weekdays whereas Casual riders use it them the most on weekends.
- Moreover, both casual and member riders use the bikes more during warmer months (May to August)

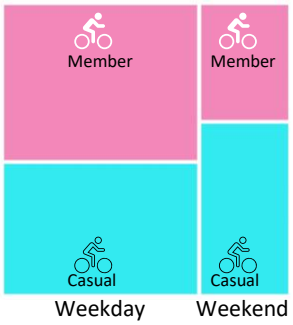
Details

For this analysis, we used the randomly selected dataset which contains an equal number of Casual and Member riders. First, the mode for each group shows that the day Members use the bikes the most it's Wednesday, whereas for Casual riders is Saturday.

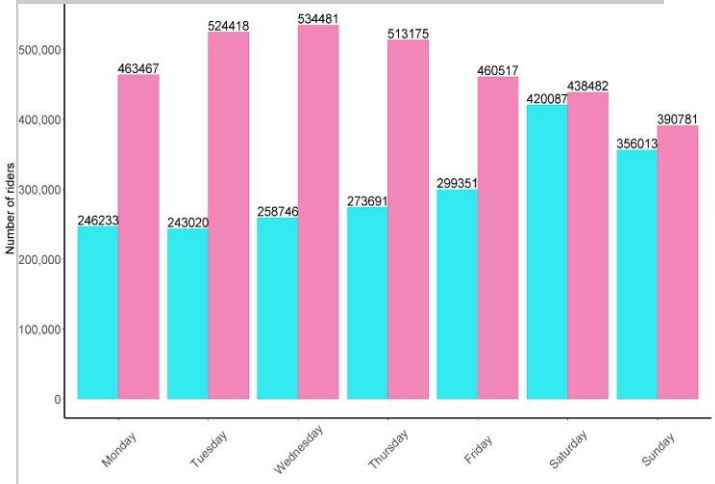
	Mode
Casual	Saturday
Member	Wednesday

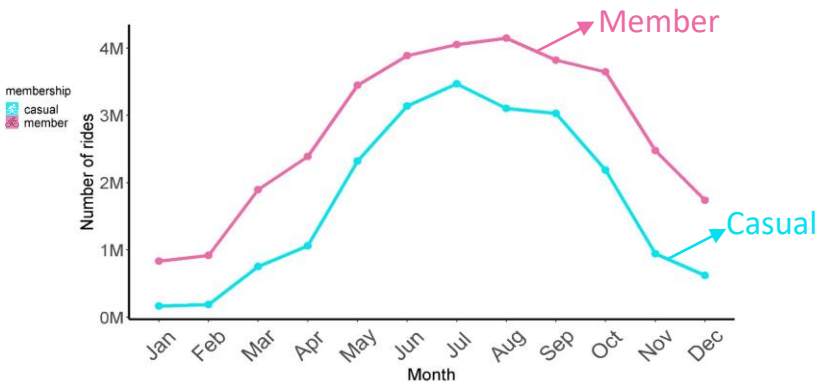
Next, we grouped the days from Monday to Friday as Weekday and Saturday and Sunday as Weekend.

	Casual	Member
Weekday	24.36%	46.03%
Weekend	14.31%	15.29%



A Chi-squared test revealed that the distribution of the percentage of riders between Weekday and Weekend did not differ between Members, (Chi-squared = 1.6579; P = 0.19). Finally, we investigated the number of rides by month of the year. This shows that both types of riders use the bikes more during warmer months (May to August). The pattern is similar between groups.





Next Steps

The team suggests moving forward and testing investigating whether Casual and Member riders use Cyclistic's bikes in different geographical areas.

Executive Summary: Statistical Testing Results

Cyclistic Bike-Share Project

Project Overview

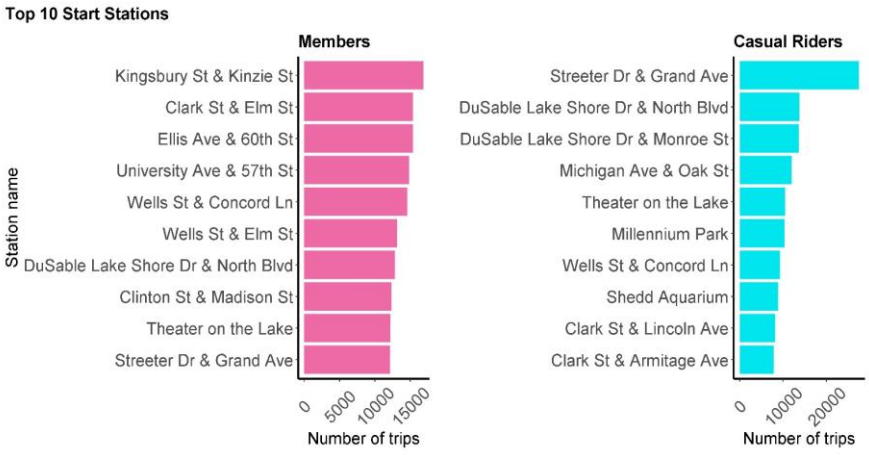
In this part of the project, the Cyclistic data team analyzes whether there are geographical differences in how members and casual riders use Cyclistic's bikes.

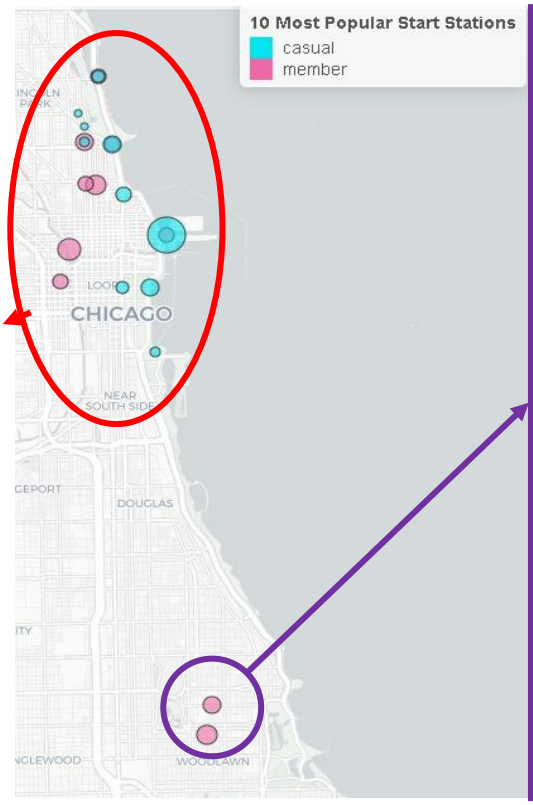
Key Insights

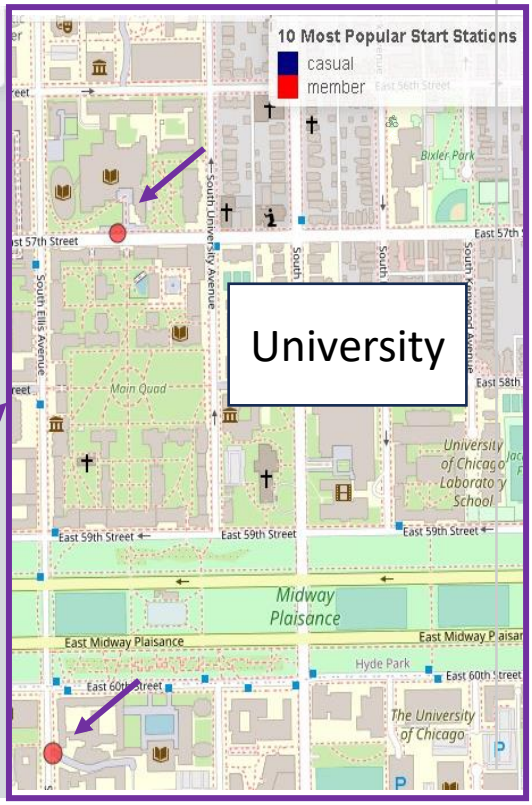
- Most Casual riders start their rides from areas close to parks and the waterfront.
- Most members start their rides from more urban areas such as the University area.
- It is possible to speculate that most Casual Riders use Cyclistic's bikes for leisure and Members for commuting.

Details

The variable *start_station_name* was selected to investigate the geographical locations most used by Members and Casual riders. The maps below show the most popular locations. The open street map helps visualizing patterns such as closeness to parks or urban areas of interest such as the University.







Executive Summary: Statistical Testing Results

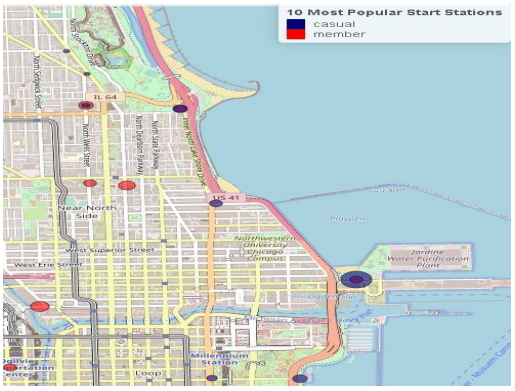
Cyclistic Bike-Share Project

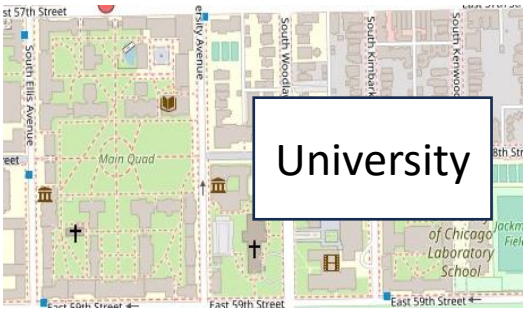
Project Summary

Here we summarise the results of the analyses and provide some actionable insights based on these findings.

Actionable Insights

1. **Geographical location**
   - The main difference between Casual and Member riders is that the most popular locations for Casual riders are close to the waterfront (Addams memorial park) and Lincoln park, compared to members which start their rides from urban location (noticeably, the University). This suggests that Casual riders use the bikes mostly for leisure and Members for commuting.





2. **Rides duration**
   - A second important behavioural difference is that Casual riders tend to use the bikes for longer rides compared to Members. In line with the previous finding, this might also indicate that Casual riders use the bikes for leisure.

	Mean	Standard Deviation
Casual	00:16:04	00:00:16
Member	00:11:51	00:00:39

3. **Time of usage**
   - Finally, casual use Cyclistic's bikes mostly on weekends.

Recommended actions

Based on the previous findings, we recommend Cyclistic's advertisement team to 1) particularly target the Park areas (especially Addams memorial park and Lincoln park; and 2) advertise using the bikes not only for leisure but also for other daily needs, such as commuting and going to buy groceries.