Improving Bike-Sharing Membership numbers with Data Science

Jesutimilehin Onayemi

27th November, 2023

Outline

- Summary
- Introduction
- Methodology
- Conclusion
- Recommendation
- Appendix

Summary

- Summary of methodologies:

  Data Collection

  Data Wrangling

  Exploratory Data Analysis

Summary

- Summary of EDA Insights:

  Riderable_type vs member_casual

  Mean duration vs member_casual

  Day vs member_casual

  Start_station name and end_station_name vs member_casual

Introduction

- Welcome to the Cyclistic bike-share analysis case study! In this case study, I will perform real-world tasks of a junior data scientist. I will work for a fictional company, Cyclistic.

Introduction

- Scenario

I am a junior data scientist working with the marketing analyst team at Cyclistic, a bike-share company in Chicago. The director of marketing believes the company's future success depends on maximizing the number of annual memberships. Therefore, the marketing team wants to understand how casual riders and annual members use Cyclistic bikes differently. From these insights, my team will design a new marketing strategy to convert casual riders into annual members. But first, Cyclistic executives must approve my recommendations, so they must be backed up with compelling data insights and professional data visualizations.

Section 1

Methodology

Methodology

- Data Gathering/Collection
- Data Wrangling
- Exploratory Data Analysis

Data Gathering/Collection

- The data used for this study is made available through:- https://divvy-tripdata.s3.amazonaws.com/index.html
- The data are uploaded each month, so I downloaded the data for the past 12 months which are fairly large in size altogether.
- In this study however, I only used the data of the most recent month, which is October, 2023.

Data Wrangling

- Loaded the csv file into a pandas dataframe
- Checked for the shape of the dataset, it contains 537,113 rows and 13 columns.
- Checked for the datatypes of each of the columns or features, each of them was correct except those for the start and end time of the bike trips. So, I converted them to the datetime datatype.

Exploratory Data Analysis

- Here, I tried to find patterns and trends within the data that could point us to what we need to do to improve the membership numbers of the company.

I did this by examining the features in the dataset, and tried to see what kind of relationship they have with the whether a bike rider is a member or a casual rider.

Section 2

Insights from the Exploratory Data Analysis

Exploratory Data Analysis

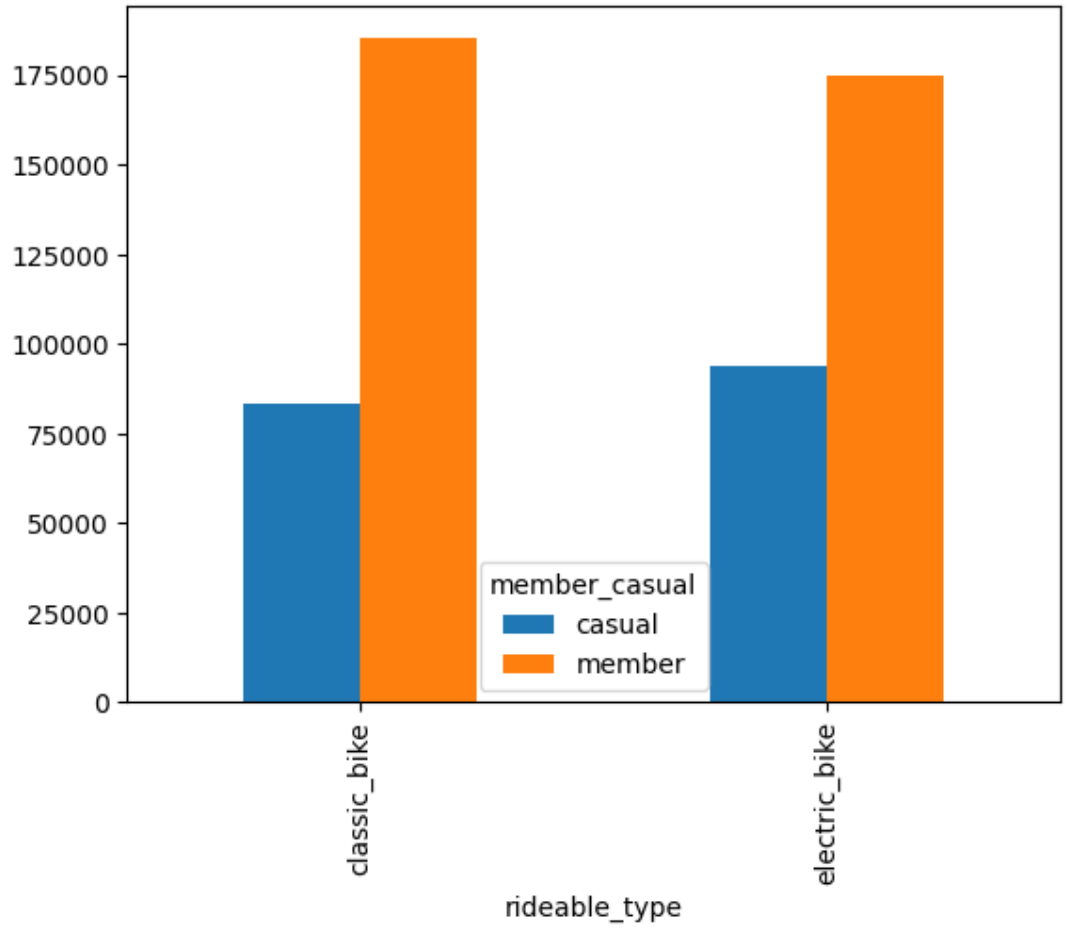
- Riderable_type vs member_casual

The first feature I examined was the type of bike the riders used for each of the trips. This is a categorical variable, and can either be electric or classic. So, in order to see and visualize the frequency of occurrences of each of the bike types with respect to whether it was a casual rider or a member that used them, I used the crosstab function. Cross tabulation (or crosstab) is an important tool for analyzing two categorical variables in a dataset.

Exploratory Data Analysis

- Riderable_type vs member_casual

The picture shows the result of plotting the crosstabulation of the riderable_type and the member_casual columns. As we can see in the bar chart, there are no significant preferences for either type of bikes by both types of rider, they both use the two types of bikes equally. So, I was able to conclude that bike type has no significant effect on whether the rider is casual or a member.



Exploratory Data Analysis

- Start and end time

The next features were the start and the end times of each trip. Further exploration of these features together, provided two things:

1.The duration of each of the bike trips

2. The day of the week when each of the trips were started.

So I subtracted the start time of the trip from the end time, and stored the result in a column, I then converted this difference to seconds which gave me the exact duration of the bike trips in seconds.

Exploratory Data Analysis

- Start and end time vs member_casual

I created a cross tab for the duration in seconds vs the member_casual column, and found out that some of the trip durations were negative in value, I wrote some code to check a cell that falls under this category, and found that the start times were actually further ahead than the end times, this is not possible. So, I did some further data cleaning and deleted the section of the data with unrealistic duration in seconds.

Exploratory Data Analysis

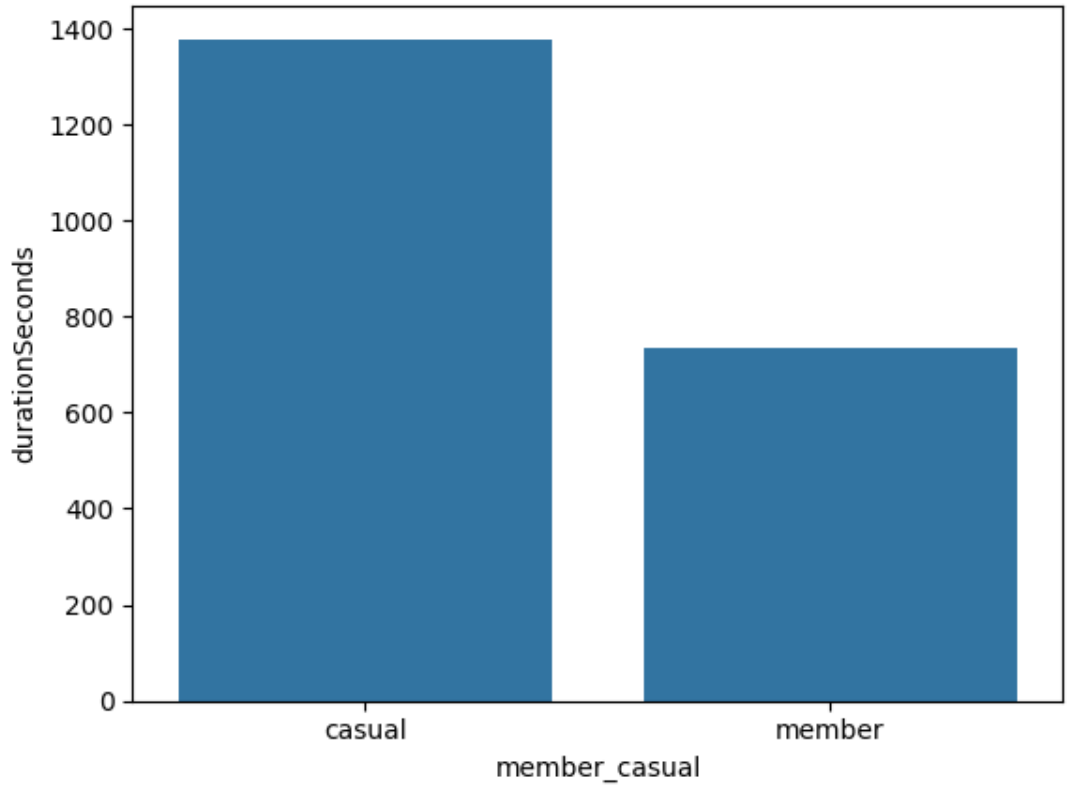
- Mean duration vs member_casual

For the sake of finding hidden trends and patterns in the data, I grouped the data by the member_casual column and applied the mean () function to the duration column, this was to find the mean of the trips duration for both the members and the casual riders.

Exploratory Data Analysis

- Mean duration vs member_casual

The bar chart to the right shows the mean value of the bike trips duration for both the casual riders and the members. From the picture, we can deduce that casual riders ride for about 600 seconds(10 minutes) longer than members



Exploratory Data Analysis

- Day vs member_casual

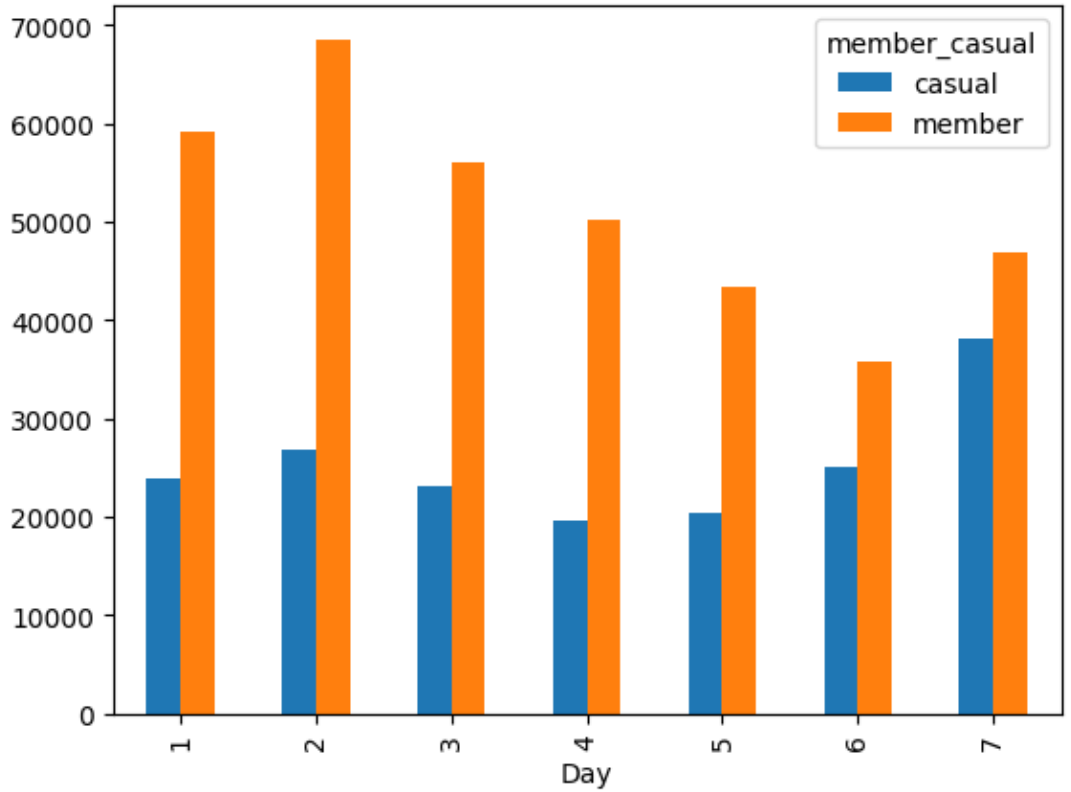
The next feature I examined was the day the each trip was taken. I was able to get the days by applying the .dt.isocalendar().day function to the start_time column, the results were numbers from 1-7 with 1 indicating it was a Monday, while 7 indicated it was a Sunday. I went ahead to see how the days in which these trips were taken differed for the two types of riders. I used the crosstab function to inspect the relationship, and then plotted a bar chart to visualize it.

Exploratory Data Analysis

- Day vs member_casual

From the resulting bar chart, we can see that casual riders rode mostly on Sundays, while members rode mostly on Tuesday.

Investigating what could be the cause of this, can help us to know what to do to convert casual riders into members.



Exploratory Data Analysis

- Start_station_name and end_station_name vs member_casual

Here I created two new dataframes out of the original dataframe:

1. One that only contained the information about the casual riders

2. Another that only contained information about the members.

I then used a function to determine the frequency of the start and end stations for these two new dataframes.

Exploratory Data Analysis

- Start_station name and end_station_name vs member_casual

I discovered that certain start station names and end station names were mostly used for the casual riders during their trips, this was also the same case for the members. The stations that were used most frequently by casual riders were distinct from those used most frequently by the members.

Since start station id and end station id are just substitutes for the station names, there is no need to analyse them further

Conclusion

1. The bike type does not determine the class of rider that will use it for their trip, both types were interchangeably used.
2. Casual riders ride for longer on the average, than members during their trips.
3. Casual riders ride mostly on Sundays, while members ride mostly during weekdays.
4. Certain station are more commonly used by the casual riders, and vice versa.

Recommendation

1. Majority of the casual riders use the bike sharing service on Sunday, which is during the weekend, and comparing this to the members who use the bike sharing system mostly during weekdays, we can infer that one of the discrepancies is that the casual riders don't use the sharing service to commute to work, or if they do, they don't use it regularly. The company should put an advert in place targeted at their casual customers, highlighting how the members have benefited from biking to work regularly, and the role biking has to play in reducing fossil fuels emissions. Appeal to their sense of preservation of the environment, and tell them the less cars on the road during workdays, the better. Biking is also a way to exercise and improve overall health, let the digital ads empasize these points. Encourage them to get a membership, to enable them bike regularly to and from their places of work, and play their part in environmental preservation.

Recommendation

2. So, on the average, the casual riders use the bikes for about 10 minutes more than the riders who are members. what could be the reason for this? The casual riders use the bikes for longer times probably in an effort of trying to maximize the amount of time they spend using the bikes because of expensive one-time payments compared to annual payments for members. The company should make a targeted follow up advert to the casual riders, telling them how they can save money in the long run by becoming members

Recommendation

3. The most common start and end station names for the casual riders and the members are distinct enough. Hence the name of the station most used by the casual riders should be taken note of, and the director of marketing can direct marketing physical marketing campaigns towards this regions. It would entail a reiteration of the benefits of biking, and why these set of users should bike more, and eventually subscribe for a membership, to help them save costs in the long run.

Appendix

- Go to the URL address to view the complete code and associated materials for this project:
- https://github.com/Jesutimilehin-Onayemi/Improving-Bike-Sharing-membership-numbers/blob/main/Cyclistic_tripdata_analysis.ipynb


Thank you!
SPEED LIMIT 25
Heartland Bike Share
HEARTLAND