Case Study 1: How Does a Bike-Share Navigate Speedy Success?

Jorge Blanco

November 20, 2023

Abstract

This report explores the analysis of data from a fictional bike-rental company provided by the Google Data Analytics Certificate course. The goal is to formulate effective strategies to attract casual customers and convert them into annual members. Through a comprehensive examination of usage behavior, patterns that guide our recommendations were identified. The proposal consists of implementing targeted offers and discounts for casual users, strategically focused on locations and schedules where they are more prevalent than members, particularly during summer months (June to September) on weekends between 12:00 and 19:00. Last but not least, a dashboard with a heatmap of the geographical distribution of the data is provided, this dashboard allows users to filter the data in the visualization to specific days of the week and months, which results in an useful tool to facilitate data driven decision-making.

1 Introduction

Cyclistic is a fictional bike rental company that offers that has grown to a fleet of 5,824 bicycles that are geotracked and locked into a network of 692 stations across Chicago. The bikes can be unlocked from one station and returned to any other station in the system anytime. Cyclistic's marketing strategy relied on its pricing plans: single-ride passes, full-day passes, and annual memberships. Customers who purchase single-ride or full-day passes are referred to as *casual riders*. Customers who purchase annual memberships are *Cyclistic members*.

Cyclistic's finance analysts have concluded that annual members are much more profitable than casual riders. Although the pricing flexibility helps Cyclistic attract more customers, the director of marketing believes that maximizing the number of annual members will be key to future growth. Rather than creating a marketing campaign that targets all-new customers, she believes there is a very good chance to convert casual riders into members since casual riders are already aware of the Cyclistic program and have chosen Cyclistic for their mobility needs.

The manager has set a clear goal: Design marketing strategies aimed at converting casual riders into annual members. In order to do that, however, the marketing analyst team needs to better understand how annual members and casual riders differ, why casual riders would buy a membership, and how digital media could affect their marketing tactics. Three questions will guide the future marketing program:

1. How do annual members and casual riders use Cyclistic bikes differently?
2. Why would casual riders buy Cyclistic annual memberships?
3. How can Cyclistic use digital media to influence casual riders to become members?

The task for this report is to answer the first question.

2 Methodology

The methodology of the taught in the Google Data Anlatytics certificate is divided in six steps:

- Ask
- Prepare
- Process
- Analyze
- Share
- Act

However, the Analize step is divided in two sections last two steps are the recommendations

2.1 Ask

The question to be answered was already given:

- How do annual members and casual riders use Cyclistic bikes differently?

But we can divide this question into more specific ones taking advantage of the temporal and geographical data given.

- Is there any preferred time for casual riders that is different from that of members?
- What's the season of the year with more casual riders?
- What's the preferred day of the week for riders and members?
- Are there any preferred stations or places for casual riders?

These are questions that can be answered by this data given by Divvy - Chicago from Motivate International under a license agreement.

2.2 Prepare

In order to get information of a long enough period of time to answer the questions, data from the last year will be used. The data files for each moth were all downloaded in the same folder, afterwards, the `.csv` files were all joined together into a single dataframe which was processed using R. The R markdown notebook used to make the cleaning and analysis is show in the following links The data had the following columns:

- ride_id: The ID of the ride, unique for each row.
- rideable_type: The kind of bike used, that can be "electric bike" or "classical bike", however there are also "doked bike", but seems to be irrelevant.
- started_at: The time at which the ride started.
- ended_at:The time at which the ride ended.
- start_station_name: The name of the station at which the ride started.
- start_station_id: The ID of the station at which the ride started.
- end_station_name: The name of the station at which the ride ended.
- end_station_id: The ID of the station at which the ride ended.
- start_lat: The latitude at which the ride started.

- start_lng: The longitude at which the ride started.
- end_lat: The latitude at which the ride ended.
- end_lng: The longitude at which the ride ended.
- member_casual: Can be "member" or "casual", describes wheather the user has an annual membership or not.

It is important to mention that not all rides have valid station names, there are a lot of them that are empty. Also, not all positions (latitude and longitude) are valid since we know that most of them should be in Chicago and there are some of them that show 0 as a value, which makes no sense.

2.3 Process

The data cleaning process started with a detailed examination of missing values in geographical coordinates (latitude and longitude). Utilizing a subset of the data that excluded rows with missing location information, the code replaced zero values in latitude and longitude columns with NA. Subsequently, rows with any remaining NA values were filtered out, confirming that no incomplete entries persisted in the dataset. The reason for deleting rows with 0 values in latitude or longitude was that this value was nonsense considering that the location of the bikes must be on the state of Illinois.

Following this, attention shifted to refining data types. All columns were recognized with their correct data type except for latitude and longitude, which were initially characterized as characters. To rectify this, the code converted the started_at and ended_at columns to date-time format for enhanced analytical capabilities.

A critical aspect of the cleaning process involved addressing empty strings in station names. The code replaced these empty strings with NA values, revealing that around 25% of the dataset contained rows with missing station names. Considering the potential information retained in these rows, a decision was made to retain them rather than discard them entirely.

After that, the member_casual column, pivotal for distinguishing between user populations, was examined. A subset was created, filtering out for values 'member' and 'casual'. As this subset yielded zero rows, the absence of unexpected values was confirmed and no further adjustments were needed.

To assess the feasibility of mapping stations using geographical coordinates, a list of distinct stations and their average coordinates was created. However, given the variability in coordinates for a single station and the large number of stations (over 1500), it was determined that the mapping may be difficult and would not increase the information of the dataset as much.

In the subsequent sections, additional columns were created to enhance the dataset's quality. Displacement calculations, derived from latitude and longitude coordinates, were added to facilitate population differentiation. The Haversine formula was used to calculate the displacement:

$$d = 2R \arcsin\left(\sqrt{\sin^2\left(\frac{\varphi_2 - \varphi_1}{2}\right) + \cos\varphi_1 \cos\varphi_2 \sin^2\left(\frac{\lambda_2 - \lambda_1}{2}\right)}\right)$$

Were φ refers to latitude and *lambda* to longitude. Additionally, the duration of rides was computed. It is important to mention that rides with durations less than a minute were identified as potential errors and therefore were removed from the dataset.

Finally, a cleaned dataset was obtained and saved in a file named "cleaned_data.csv" for further analysis and exploration.

2.4 Analysis

2.5 Displacement

Analizing the displacements of the riders, we can see if there's any difference betweeen casual riders and annual members. Figure 1 shows the distributions of the displacements of the riders, it is hard to spot any difference between casuals and members in terms of the displacement. Table 1 shows the some basics statistics of the displacement variable, the median is always lower than the average due to the skewness of the data. All of the variables show in this table are reasonable considering the margin of error of the coordinates.

Table 1: Summary statistics of the displacement distribution among riders

Membership	**Distance [km]**		
	Average	**Median**	**Maximum**
Casual	1.63	1.11	34.47
Member	1.56	1.02	29.44

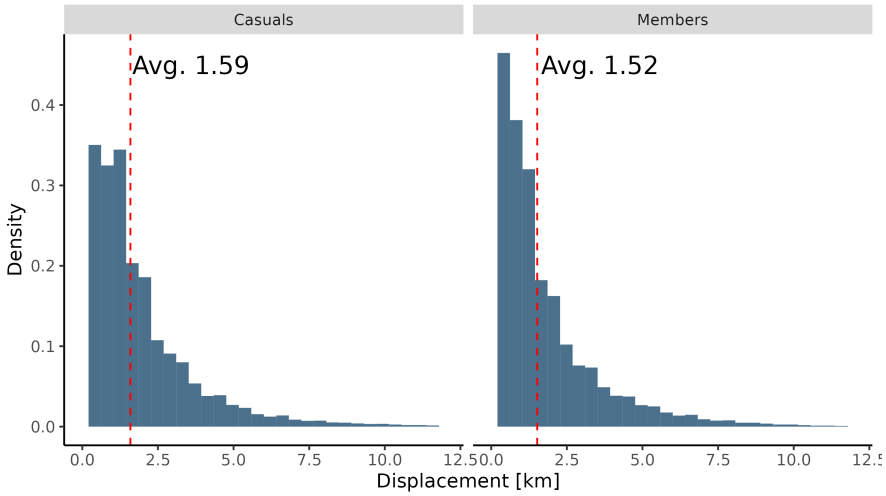

Figure 1: Displacement made by casual riders (left) and annual members (right). Both distributions are quite similar.

As a result, no difference can be found in the displacement of both populations.

2.6 Trip duration

Similarly as with displacement, the trip duration can be analized. Figure 2 shows the distribution of ride duration by casual users and members. Again, a skewed distribution can be observed, however, for riders it is flatter, which shows that there are more casual riders that take long trips than members. This effect can be also apreciated in table 2, which shows that, on average, casual riders make loger trips than members. One can also spot an error in the maximum duration of the trips, their excessive length may be an indicator of errors that can occur at the moment of delivering back the bicycle.

Table 2: Summary statistics of the trip duration distribution among riders

Membership	**Duration [min]**		
	Average	**Median**	**Maximum**
Casual	20.81	12.17	12136.3
Member	12.37	8.75	1499.933

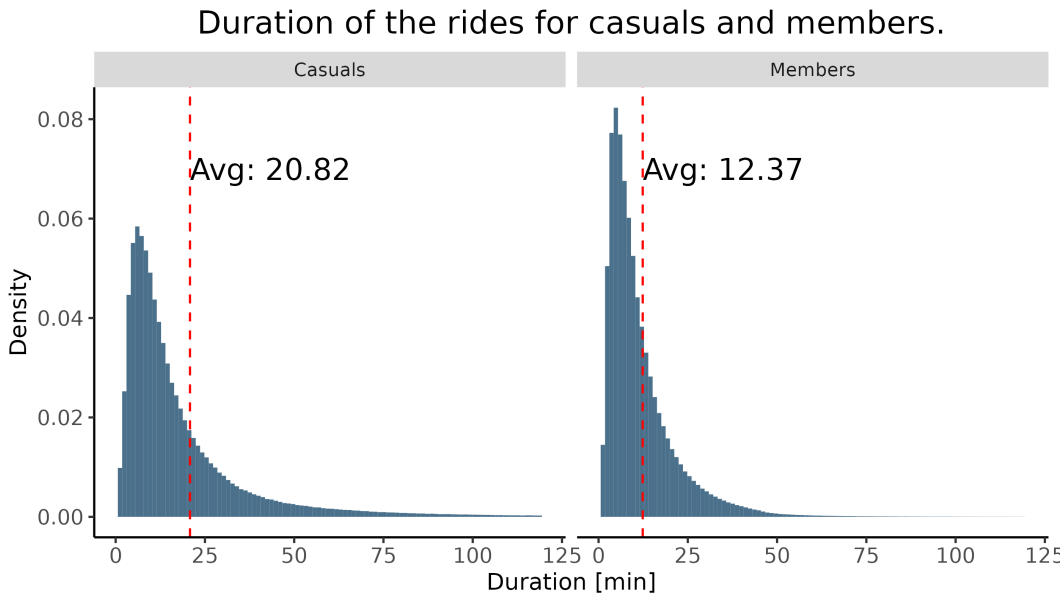

Figure 2: Trip duration by casual riders (left) and annual members (right). It is easy to notice that there are more riders which are willing to make longer trips in the group of casuals.

Casual riders taking longer trips than members may be an expected indicator of the motivation of the riders to rent a bicycle. Since members use their bicycles to commute to work (in general, for daily activities), their rides should take less time. On the other hand, casual riders may use the bicycles for recreation, which would take them much more time.

2.7 Temporal cycle patterns

It is important see the weekly pattern of rides by membership. Figure **??** shows that annual members rent bicycles more frequently in between weekdays, this supports the argument of members renting bicycles for commuting to work or for routine activities. On contrast, there is an increase in casual rides on weekends, reinforcing the hypothesis of recreational rides.

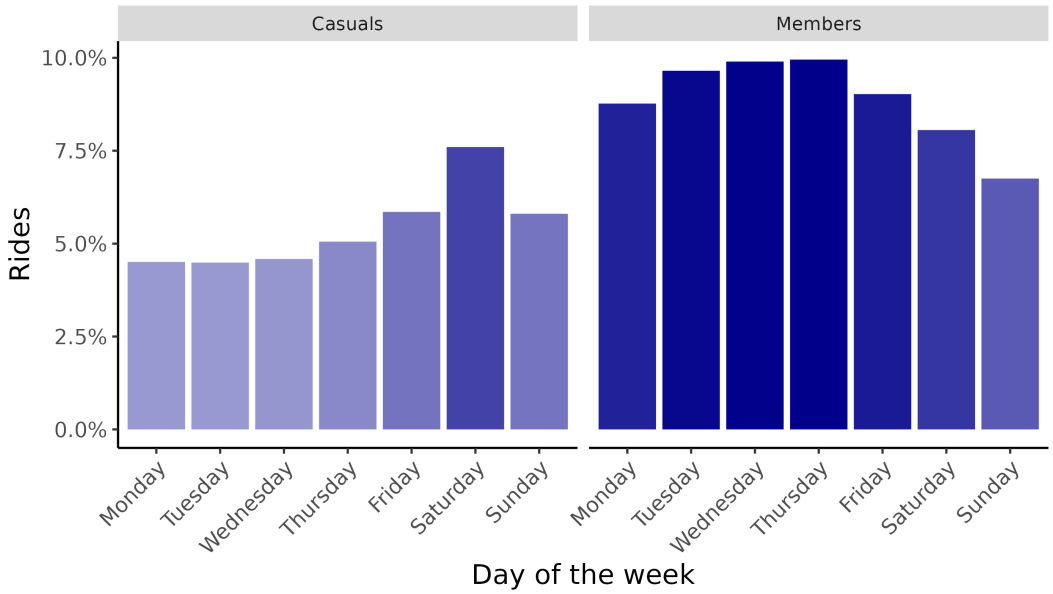

Figure 3: Weekly bike trips distribution among casuals and members

Going deeper into this pattern, a histogram of the rides made per hour shows that there are mainly two distributions, that o f weekdays and another one for weekends. On weekdays two peaks can be noticed in the members distribution, the first peak is around 8 to 9 am , associated with people commuting to work and the second one goes from 4 to 6 pm, related to people going back home or taking a ride after work. In casuals, the first peak is reduced to a shoulder, which indicates that there are still some casual riders that use their bicycles to commute to work, but the population size is reduced. Meanwhile, the second peak still appears on the casual distribution, this can be related with people taking a ride after work or transporting by bicycle for shopping as it happens for members.

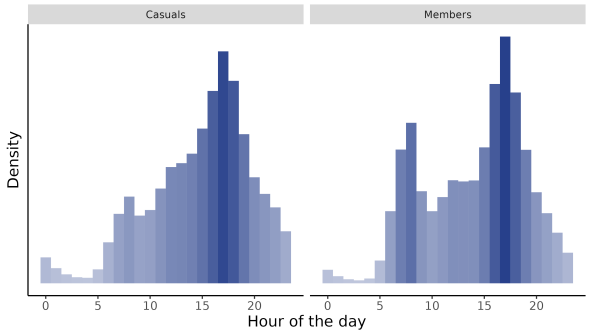

Figure 4: Weekday daily ride distribution

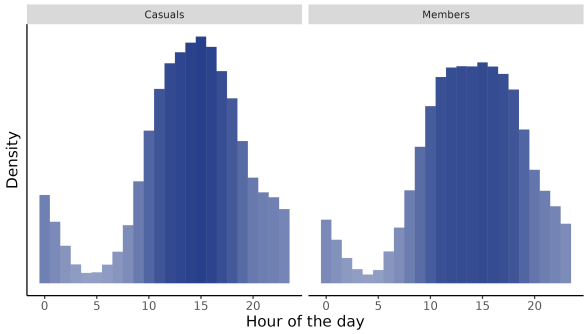

Figure 5: Weekday daily ride distribution

Finally the monthly distribution of rides can also be analyzed, this is interesting because is related to the climatic season. However, it is important to mention that there's an uncertainty related to studying monthly rides with data from only the last year. Figure 6 shows that there is an seasonal variation in rides, they increase from April to November and reach the minimum value from December to March. The amplitude of this variation is much grater for casual riders than for members, the highest amount of casual rides (August) is almost 10 times more than the lowest (January). For annual members the variation is around a factor of three, which is much less than for casual riders.

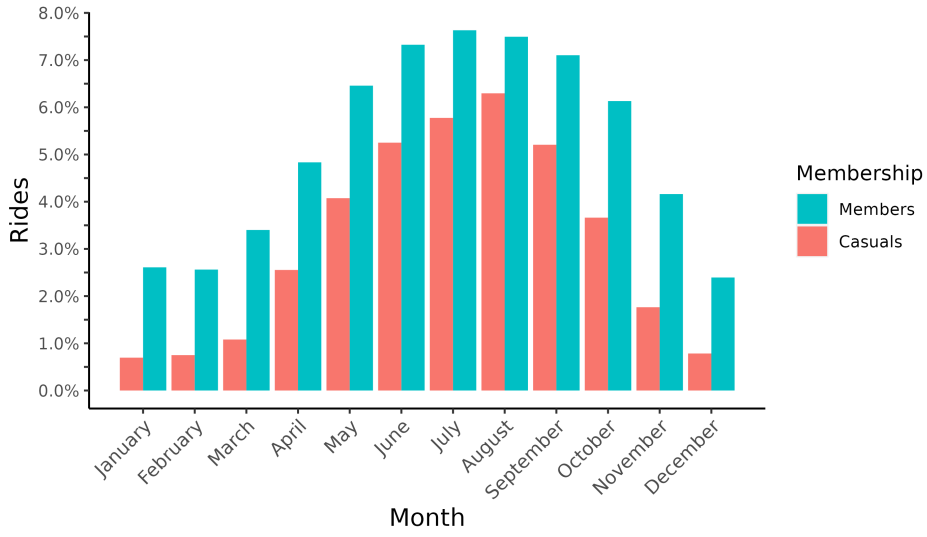

Figure 6: Monthly rides made by casual riders and annual members. The summer rides increase is much greater for casuals than for members

2.8 Bike Type

The dataset also included the type of bike used for the ride, so that can be useful to see if there's any preference on it by the type of rider.

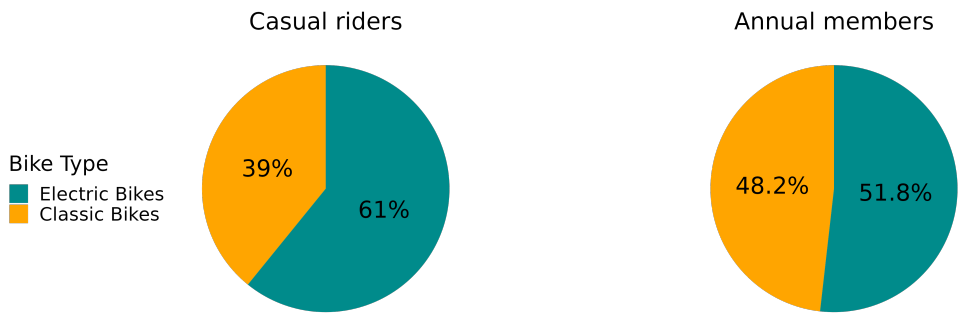

Figure 7: Bicycle type usage by membership

Figure 9 shows that, on average, annual members do not prefer any type of bike, while casual riders use more the electric bike, the difference is not overwhelming, but can be an aspect to consider.

2.9 Stations

It is important to use also the information available to evaluate which are the stations where casual users start their trips the most. For that purpose, the data was grouped by stations, counting the members and casuals for each. Then, a list of the stations with higher casual to member ratio was made for those stations with more than 1000 users (Table 3), because focusing on stations with less than approximately 3 users per day could be a waste of resources

Table 3: Top 10 Stations with higher casuals to member ratio and more than 1000 users

Start Station	Members	Casuals	Casuals/Members
Buckingham Fountain	1068	4287	4.01
Field Museum	1986	7943	4.00
Shedd Aquarium	4750	18257	3.84
DuSable Lake Shore Dr & Monroe St	9141	29488	3.23
Streeter Dr & Grand Ave	16732	48800	2.92
Dusable Harbor	5188	13598	2.62
Adler Planetarium	4555	11130	2.44
Millennium Park	9537	21851	2.29
McCormick Place	2273	4333	1.91
Kedzie Ave & 48th Pl	390	716	1.84

Here it can be noticed that stations with higher casual-to-member ratio have not necessarily higher casual rides. Table 4 shows the top stations on the absolute value of casual members

Table 4: Top 10 Stations with higher amount of casual users 1000 users

Start Station	Members	Casuals	Casual/Member
Streeter Dr & Grand Ave	16732	48800	2.92
DuSable Lake Shore Dr & Monroe St	9141	29488	3.23
Michigan Ave & Oak St	14391	22652	1.57
Millennium Park	9537	21851	2.29
DuSable Lake Shore Dr & North Blvd	15170	20474	1.35
Shedd Aquarium	4750	18257	3.84
Theater on the Lake	13693	16402	1.20
Wells St & Concord Ln	21047	13827	0.66
Dusable Harbor	5188	13598	2.62
Indiana Ave & Roosevelt Rd	13858	12036	0.87

2.10 Location

With the start coordinates of each ride, a density map was made in Tableau (shown here). It was made as a Dashboard to allow the user to play with the filters by month or day of the week.

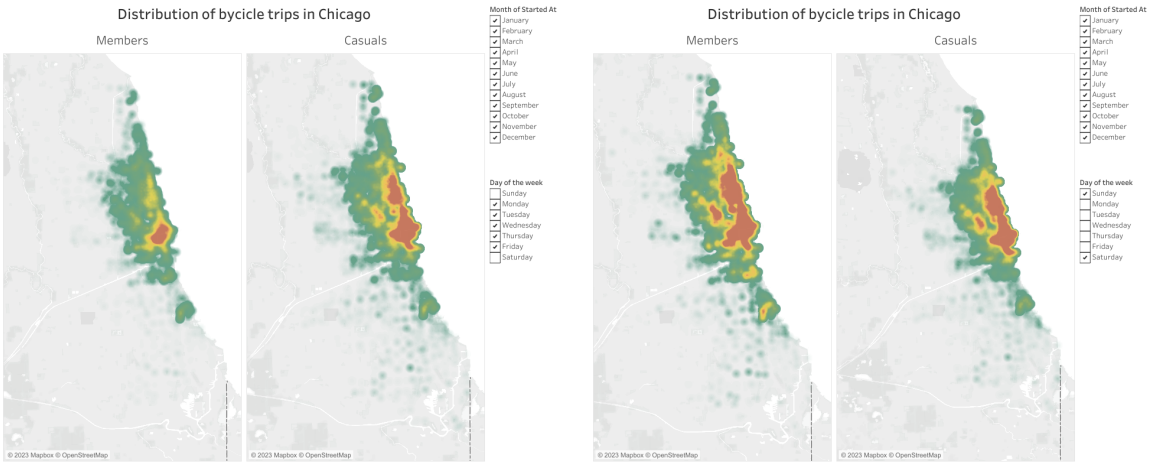

Figure 8: Geographical distributions of rides on weekdays (left) and weekends (right) for all months of the year

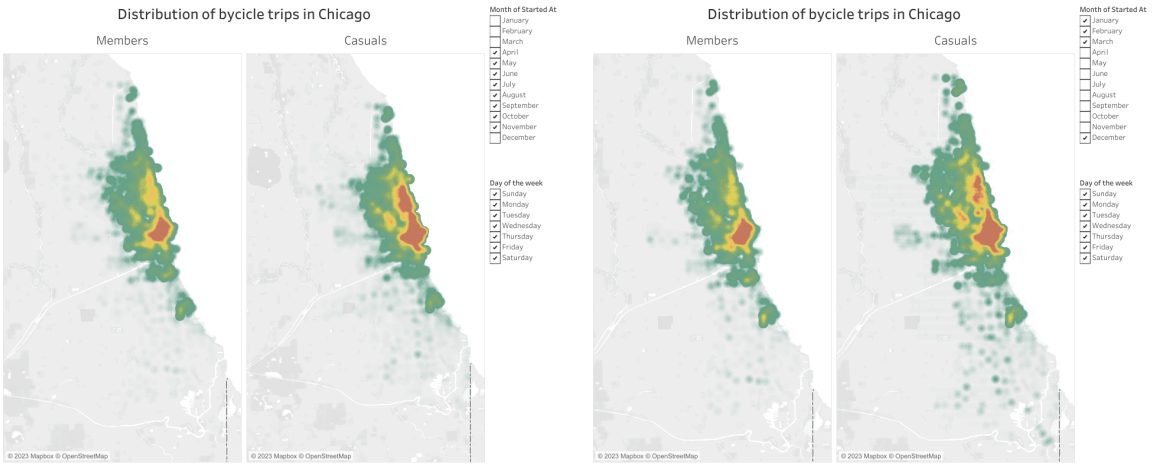

Figure 9: Geographical distributions of rides in summer (left) and winter (right) for all days of the week

It is interesting to notice that the geographical distribution of the casual rides bearly changes with the day of the week or the season. Also, only in weekends the distribution of members changes. Most of the time, annual members appear to be more concentrated in Central Chicago, while casual rides are more spread, but mainly dispersed to the North

3 Conclusions

Considering the question to be answered was how do annual members and casual riders differ from each other, these are the conclusions for differenciating this populations:

- The analysis reveals that casual rides tend to have a longer duration compared to rides taken by annual members.
- A trend can be noticed as annual members reduce their rides on weekends, while casual riders exhibit an increase in rides, particularly on Saturdays.
- Daily riding patterns show a peak among annual members from 8 to 9 am on weekdays. In contrast, casual riders exhibit a less pronounced morning peak and both populations have shared peak in the afternoon, between 16:00 and 18:00.
- The seasonal variation is more pronounced among casual riders. While annual member's' rides can decrease by a factor of 3 with respect to the highest month, casual rides can decrease by a factor of 10.
- Data suggests a slightly higher utilization of electric bikes among casual members.
- stations such as Streeter DR & Grand Ave, DuSable Lake Shore Dr & Monroe St, Millennium Park, and Shedd Aquarium exhibit a higher influx of casual riders compared to annual members, positioning them as top stations for targeting casual rider engagement
- The annual member spatial distribution stays consistent except for weekends. This suggest that annual members are concentrated in the Central area of Chicago and therefore, the marketing strategies should target other places on the north side.

4 Limitations

One limitation is the absence of a rider ID, without this variable one can not distinguish if a certain amount of rides were made by a single person or by different people. Also, having rider IDs would have allowed to make subsets of the population that meet certain conditions and learn more about them. For example, it would be interesting to make a population of the casual riders that frequently use their bicycles on weekdays morning and learn about them to see what would make them buy a membership. Another limitation was the reduced precision on the spatial data, there were a considerable amount of rides without a station name and the coordinates for each station had a lot of variation, making it dificult to map the stations accordingly eith the coordinates, this would have made the analysis on of the stations and locations much more exact.

4.1 Absence of Rider ID:

One limitation is the absence of a unique rider ID in the dataset. The lack of this identifier restricts the ability to differentiate rides made by a single person from those made by different people. Rider IDs would have enabled the creation of subsets based on specific conditions, offering a deeper understanding of user behavior. For instance, identifying and studying casual riders who frequently use bicycles on weekday mornings could be useful in tailoring strategies to potentially convert them into annual members.

4.2 Reduced Precision in Spatial Data:

Another limitation revolves around the reduced precision in spatial data. A considerable number of rides lack a station name, and the coordinates for each station exhibit significant variation. This variability makes difficult mapping stations to their respective coordinates, affecting the precision of analyses related to the stations. A more precise spatial dataset would have facilitated a more accurate examination of station-specific trends and user behaviors tied to particular locations.

These limitations highlights the importance of careful consideration when interpreting and generalizing the findings. While the current analysis provides valuable insights, addressing these limitations

in future data collection or analysis endeavors could enhance the depth and accuracy of conclusions drawn from the dataset.