GDA-CASE STUDY-1: CYCLISTIC BIKE-SHARE ANALYSIS

BY: RAHAT, A.D.

DATE: 14th AUGUST 2023

Table of Contant

Introduction

Welcome to the Cyclistic bike-share analysis case study! In this case study, I performed many real-world tasks of a junior data analyst. I worked for a fictional company, Cyclistic, and met different characters and team members. In order to answer the critical business questions, I followed the steps of the data analysis process: **ask, prepare, process, analyze, share,** and **act**.

Scenario

I am a junior data analyst working in 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, my 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 your recommendations, so they must be backed up with compelling data insights and professional data visualizations. Characters and teams:

❖ **Cyclistic:** A bike-share program that features more than 5,800 bicycles and 600 docking stations. Cyclistic sets itself apart by also offering reclining bikes, hand tricycles, and cargo bikes, making bike-share more inclusive to people with disabilities and riders who can't use a standard two-wheeled bike. The majority of riders opt for traditional bikes; about 8% of riders use assistive options. Cyclistic users are more likely to ride for leisure, but about 30% use them to commute to work each day.

❖ **Lily Moreno:** The director of marketing and my manager. Moreno is responsible for the development of campaigns and initiatives to promote the bike-share program. These may include email, social media, and other channels.

❖ **Cyclistic marketing analytics team:** A team of data analysts who are responsible for collecting, analyzing, and reporting data that helps guide Cyclistic marketing strategy. I joined this team six months ago and have been busy learning about Cyclistic's mission and business goals — as well as how I, as a junior data analyst, can help Cyclistic achieve them.

- **Cyclistic executive team:** The notoriously detail-oriented executive team will decide whether to approve the recommended marketing program.

About the company

In 2016, Cyclistic launched a successful bike-share offering. Since then, the program has grown to a fleet of 5,824 bicycles that are geo-tracked 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 at any time.

Until now, Cyclistic's marketing strategy relied on building general awareness and appealing to broad consumer segments. One approach that helped make these things possible was the flexibility of 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, Moreno 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, Moreno believes there is a very good chance to convert casual riders into members. She notes that casual riders are already aware of the Cyclistic program and have chosen Cyclistic for their mobility needs.

Moreno 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. Moreno and her team are interested in analyzing the Cyclistic historical bike trip data to identify trends.

6 Stages of the Analysis

1. **Ask:** In this stage, the critical business questions will be identified along with the stakeholders.
2. **Prepare:** Collect the data, identify how it's organized, and determine the credibility of the data.
3. **Process:** Select the tool for data cleaning, check for errors, and document the cleaning process.
4. **Analyze:** Organize and format the data, aggregate the data so that it's useful, perform calculations, and identify trends and relationships.
5. **Share:** Use design thinking principles and a data-driven storytelling approach, to present the findings with effective visualization. Ensure the analysis has answered the business task.
6. **Act:** Share the final conclusion and the recommendations.

PHASE-1: ASK

★ **Task:** Identify the types of membership and analyze them to figure out their modes of use and prepare a marketing strategy to convert casual riders into annual members.

★ **Stakeholders:**
  - **Lily Moreno:** Director of marketing and manager of Cyclistic Executive Team.
  - **Cyclistic Executive Team:** A team that will decide whether to approve the recommended marketing program or not.
  - **Cyclistic marketing analytics Team:** A team of data analysts responsible for collecting, analyzing, and reporting data.

PHASE-2: PREPARE

The key points of this phase are:-

1. Data Source

I used Cyclistic's historical trip data to analyze and identify the trends. I downloaded the data from https://divvy-tripdata.s3.amazonaws.com/index.html.

2. Data Organization

In the mentioned link all the data files are uploaded monthly. As I only need 1 year of data I downloaded the 12 data files for the year 2022. I uploaded all the files to "Google Cloud Storage" and using "big query" merge them all together and make a single file.

3. Checking Data Quality

To Measure the quality of data. I follow the following steps:

1. The data is generated by the company itself, meaning it is first-party data. As a result, there was no bias and credibility as well as its maintained integrity.
2. The data maintain ROCCC as the data source is reliable, original, comprehensive, current, and cited.

4. Security, Privacy, and Accessibility

Security, privacy, and accessibility weren't a problem as the dataset is public and available for everyone to use.

PHASE-3: PROCESS

To process the data I have followed the steps:

1. ## Tools

   First, I used **"Google Cloud Storage"** and **"big-query Console"** to merge the data set into a single file. Then I used **R** to clean, process, and analyze the data.

2. ## Removing Null Entries

   To remove the null entries I took **2 steps**.

   a. In the first phase, I checked the null entries in the dataset and found there were some columns that had null values. The columns are start_station_name, start_station_id, end_station_name, end_station_id, start_lat, start_lng, end_lat, end_lng. Then I remove the rows containing null values.

   b. In the second phase, I re-checked the null entries and found there are **NO null values** in the data but still have some **empty string instances**. I found some of the columns still have some empty strings. The columns are start_station_name, start_station_id, end_station_name, end_station_id. As we have the latitude and longitude for these stations, we didn't remove them.

   Lastly, I calculated the percentage of loss after the null value removal and I found o**nly 0.1034%** of the total data were lost.

3. ## Calculating trip duration

   I have followed some steps to calculate the trip duration of each ride.

   Step-1: Convert the start time and end time into the date-time format, as they are stored as strings.

   Step-2: Calculate the difference between these two columns.

   Step-3: Analyze the highest and lowest time difference. While analyzing the difference of the time I found some of the trip duration is negative which is impossible. So, I **remove**

those columns as there are only **100** rows (**0.0018%** of total data points) containing these error ride duration data.

4. Rename the Membership Type Data

In the data set, there are 2 unique membership types data those are casual, and member. As the types are confusing to understand. I renamed the causal members as **one-time users** and annual members as **subscribers**.

5. Process the Weekday Data

In order to get the day of the week, I convert the started time to date-time format and use the **weekday()** function to get the day name.

6. Process the Weekend/Weekdays Data

To get the weekend/weekday data I used the weekday column. The new column contains only two categorical data weekend and weekdays.

7. Derive the Month of Ride

As I need to analyze the monthly ride pattern, I need the month of the ride. So, I use the started_at feature to derive the month of each ride.

8. Formulate the Part of the Day for Each Ride

To get the busiest time of the day for riding I need to set some custom chunks. I divided a day into 4 parts:

i.) 6 AM to before 12 PM is **morning** time.

ii.) 12 PM to before 5 PM is the **afternoon**.

iii.) 5 PM to before 9 PM is the **evening** and

iv.) 9 PM to before 6 AM is the **nighttime**.

For setting the chunks, I convert the started_at feature to date-time format and then set the specific chunk's name.

9. Determine the Trip Type

While randomly checking the start and end station names, I observed some of them have the same start and end station. Which indicates some of the rides were **round-tripped**. So, I added one more feature to the data set where the trips are categorized into two types: **round trips** and **one-way trips**.

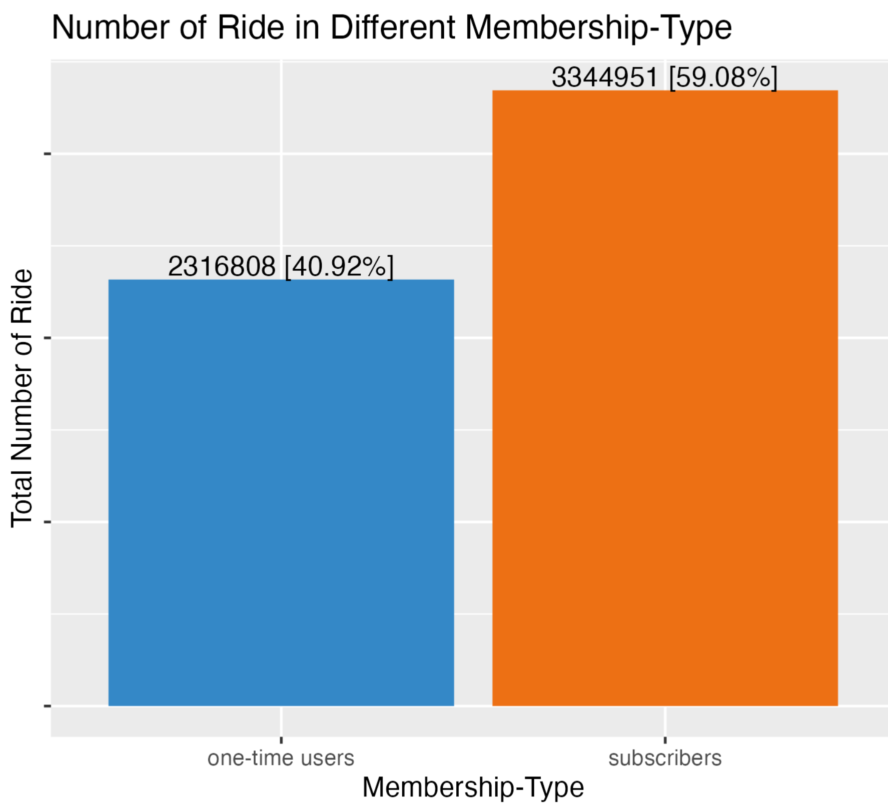
PHASE-4: ANALYZE

1. Analyze Ride Counts

At first, I did an analysis of the total number of rides. The insights are:-

i. The total number of unique rides is **5661759 ~ 5.66 M.**

ii. Among the user base, ride counts are distributed across two main categories: '**One-time users**' accounted for approximately **2.32 million** rides, representing **40.92%** of the total. Meanwhile, **'Subscribers'** exhibited robust engagement with **3.34 million** rides, constituting **59.08%** of the total.



iii. The visualization clearly indicates that 'Subscribers' have significantly heightened ride engagement, with approximately 1,028,143 rides more than 'One-time users'—a substantial difference of 1 million. This means there are 1.44 times more rides for 'Subscribers,' clearly showing how much they prefer the services compared to 'One-time users.'

2. Analyze Ride Length

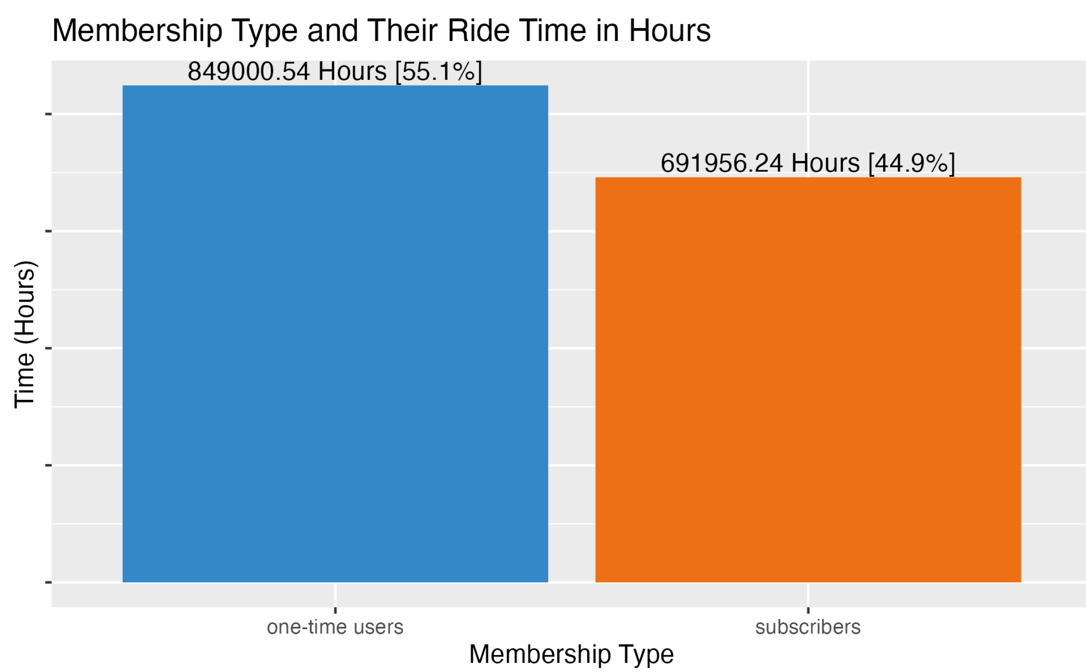
In this section of the analysis, I have analyzed the total trip time of each ride. And the key points are:-

i. The data points are distributed from the **1st of January 20**22 to the **31st of December 2022**. So, I have exactly one year of data to analyze.

ii. Total ride duration is **1,540,956 Hours 46 Minutes and 40 Seconds**. In short, the total ride time is more than **1.5 M Hours**.

iii. On average each ride durate approximately **16 Minutes and 20 Seconds**.

iv. One-time user ride **849,000.5 hours** which is **55.1%** of total rides and the subscribers ride a total of **691,956.2 hours** which is **44.9%** of the total trip duration.



v. Although the **subscriber** has a **1.44 times** higher ride number than the one-time user, the **one-time user** has a **1.23 times** higher trip duration than the subscriber. So, I investigate the average trip duration of each type.

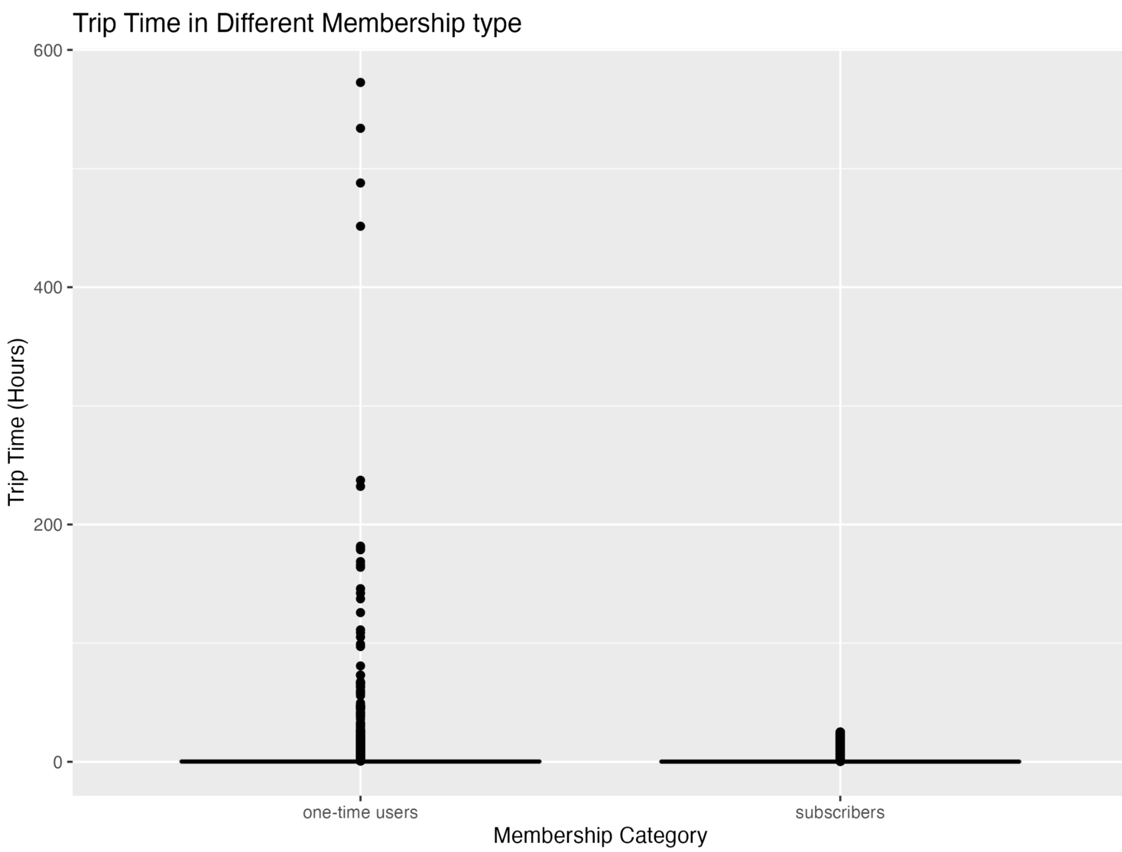
vi. The average ride time for **one-time users** is **21.98 minutes**, contrasting with **12.41 minutes** for **subscribers**, aligning with my expectations.

vii. Notably, there exists a time disparity of nearly **9.5** minutes between the **average ride duration** of these **two membership types**.

viii. Looking more closely, **one-time users** tend to ride for a median duration of about **12.97 minutes**, while **subscribers** ride for around **8.33 minutes**.

ix. This **difference** in **median** ride times for each type of member is about **4.64 minutes** while it's **half** of the difference seen in **average** ride times (**9.5 minutes**).

x. The discernible **variance** between the **mean** and **median** t**rip duration** for **one-time users**, suggesting approximately **9 minutes**, points toward the potential existence of **outliers** within this group.

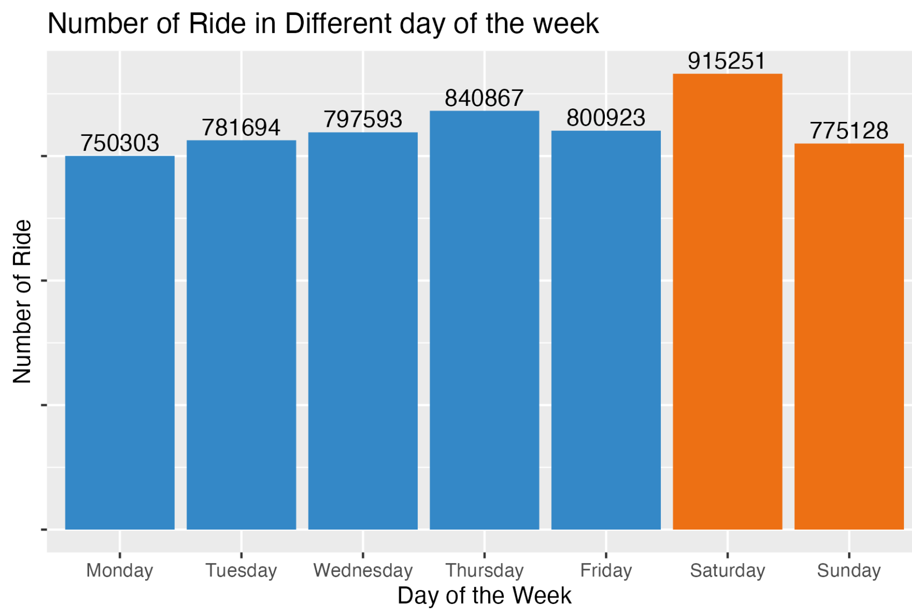


xi. Based on the plot, it's evident that **Subscribers consistently** use bikes in a **steady manner**. On the other hand, **one-time users**, who tend to have **irregular patterns** of usage, sometimes have **longer ride durations** in certain instances.

3. Evaluate the Riding Trends Throughout Different Weekdays

In this segment, I conduct an analysis of trip patterns across different days of the week. To analyze this feature comprehensively, I have meticulously followed the outlined steps below:

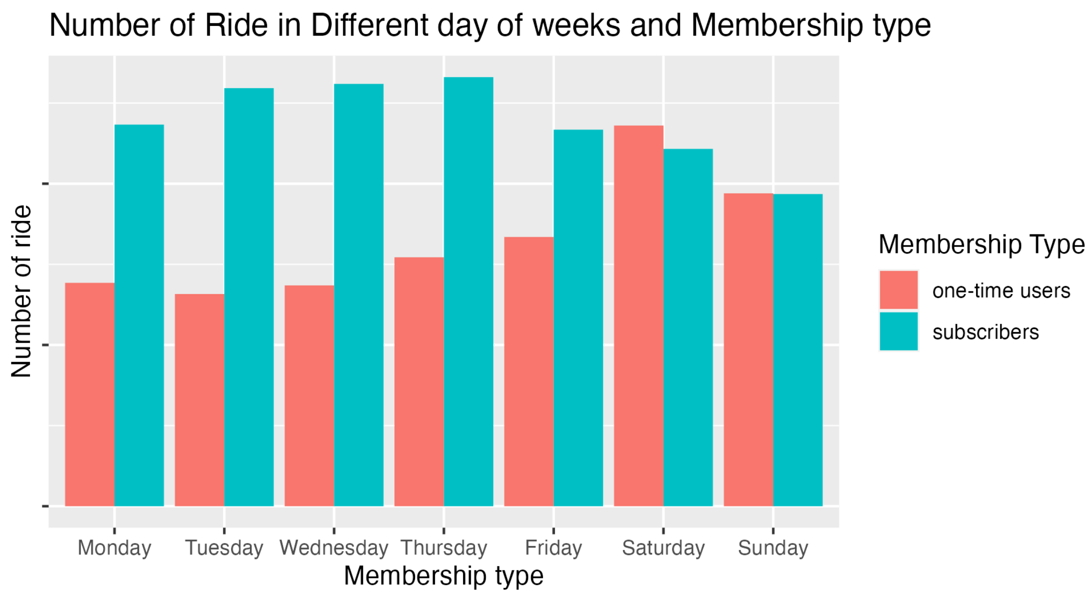
I. From the attached figure, ride distribution is **well-balanced** throughout the days of the week. Notably, **Saturday** experiences the **zenith** of **ride activity**, while **Monday** records the **most modest** ride count.



II. The average number of ride on **weekdays** is **79,4276** and on **weekend** is **84,5189.5**, which clearly indicate that users **favor weekends over weekdays**.

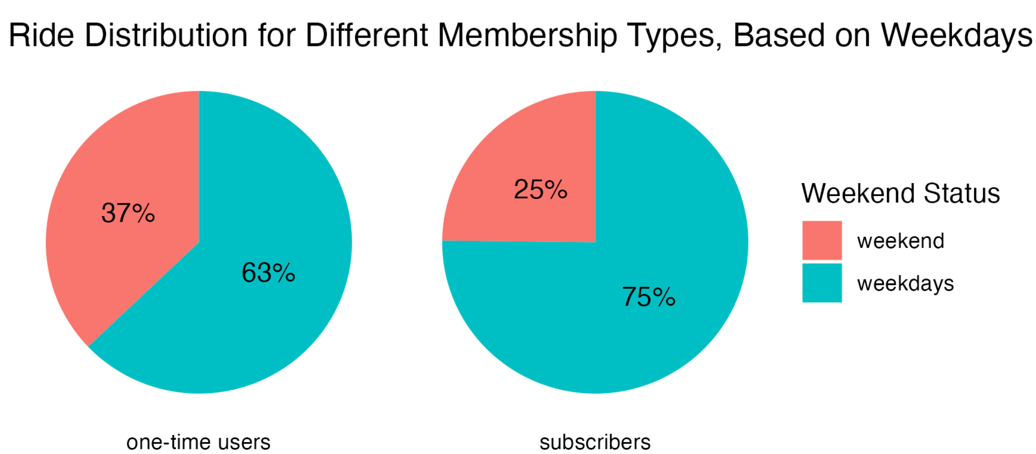
III. From the below figure, During **weekdays**, there is a consistent trend of higher ride usage among **subscribers** compared to **one-time users**. Over **weekends**, the disparity diminishes, yet there remains a slight elevation in ride frequency among one-time users.

IV. From the below visualization, it's clear that **one-time users** exhibit heightened ride activity during **weekends**, in contrast to **subscribers** who demonstrate an **opposing trend**. Upon observing this distinct pattern, it is reasonable to postulate that **one-time users**, possibly comprising **visitors**, **favor weekend rides**, while subscribers, conceivably consisting of **students** or **professionals**, predominantly engage in rides during **weekdays**.



V. **One-time users** exhibit **heightened** ride activity **during weekends**, in contrast to **subscribers** who demonstrate an **opposing trend**.

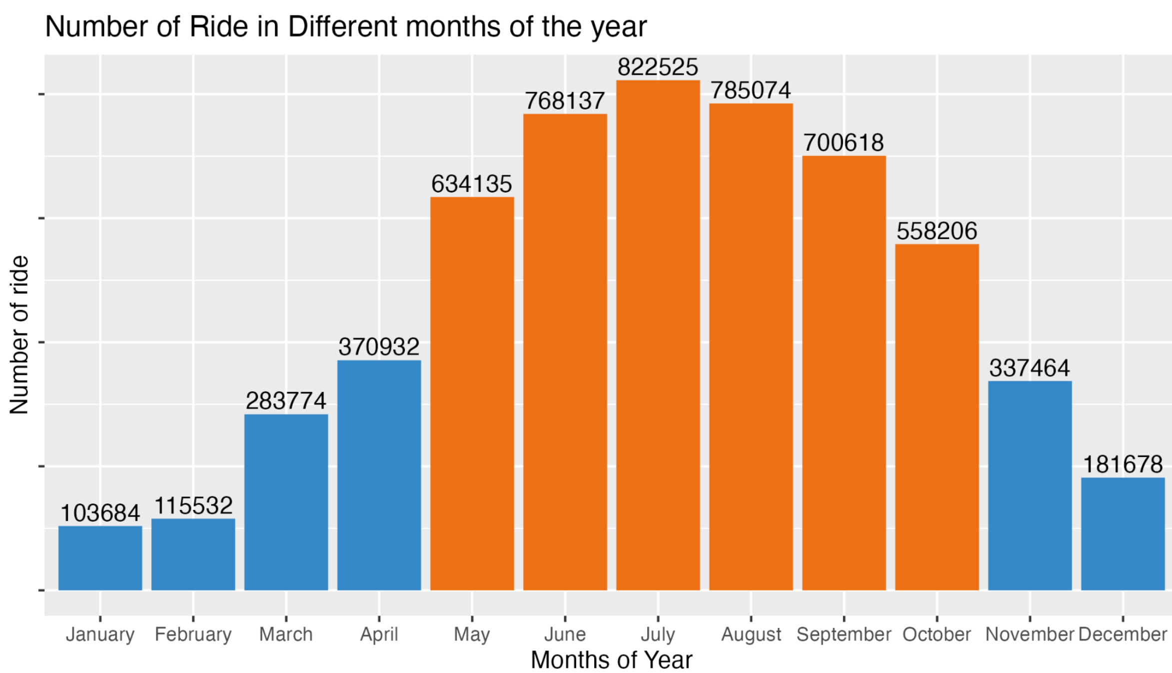
VI. According to the following Pie-Chart, the **majority** of **subscribers** (**75%**) opt for rides on **weekdays**, while a **minority** (**25%**) prefer **weekends**. For **one-time users**, **37%** of their rides occur on **weekends**. This highlights that **one-time users** exhibit a **stronger** preference for **weekend** rides compared to the more **weekday-centric** pattern of **subscribers**. Upon observing this distinct pattern, it is reasonable to postulate that **one-time users**, possibly comprising **visitors**, favor **weekend** rides, while **subscribers**, conceivably consisting of **students** or **professionals**, predominantly engage in rides during **weekdays**.



4. Analysis of Bike Ride Patterns Throughout Different Months of the Year

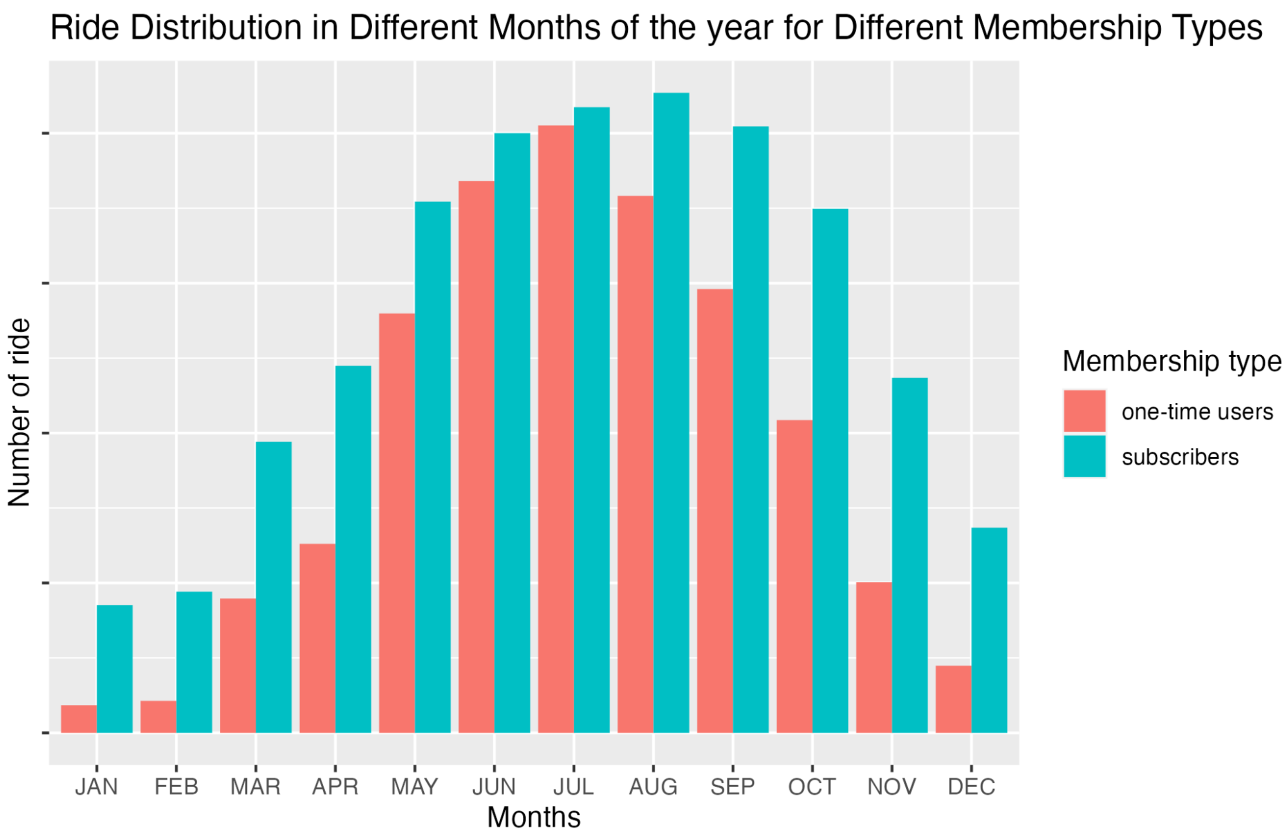
In this portion of my analysis, I explain the relationship of ride attitude throughout the **months of the year**. To analyze these attributes, I followed the below steps:

I. From the following visualization, it seems that the data is **normally distributed**. The **highest** number of rides captured in **July** which is **822,525** and the **lowest** count recorded in **January** which is **103,684**.

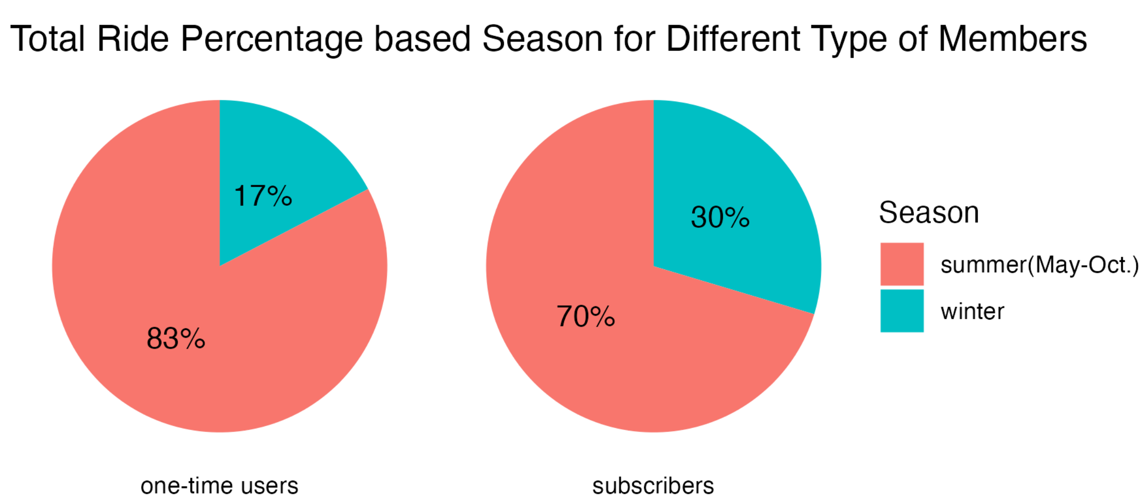


II. In the above visual, **May to October** months are **summer** which are marked as **blue** and the number of rides dramatically increases in the **summer** season (May-October). Surprisingly **75.4%** of the **total ride** (**means 3 out of 4 rides**) were recorded in **summer**.

III. Throughout the year, subscribers use more rides than one-time users. The **increasing trends** of **summer** clearly presented in the below plot, but if we consider the **membership** type then the **one-time users** have a much more **dramatic upward** trend in the **summer season** with respect to the **subscribers**.



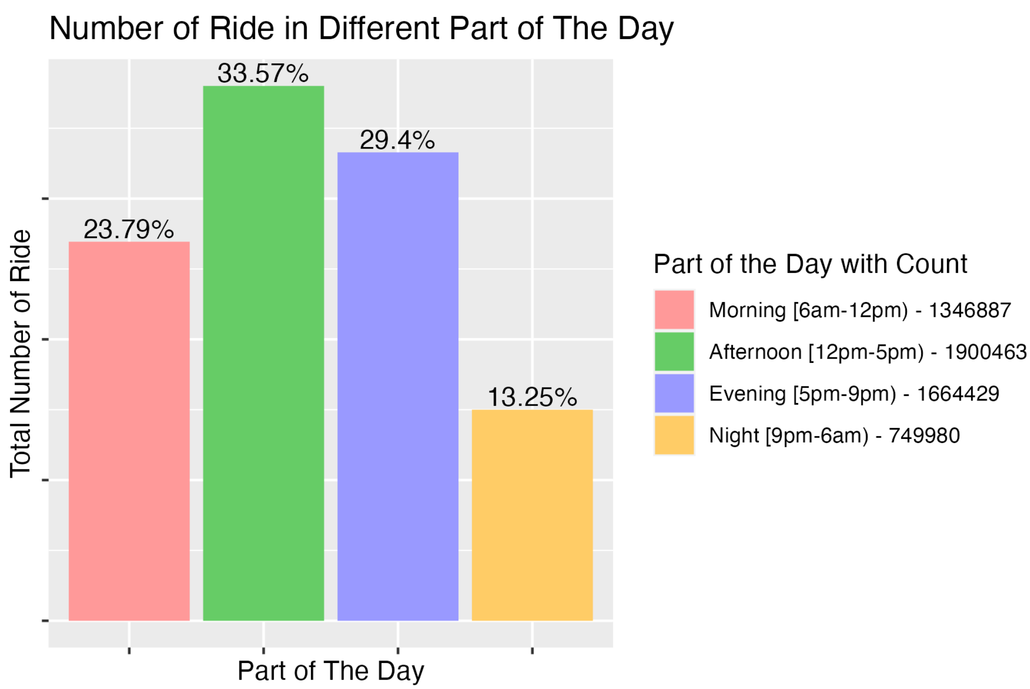
IV. Further, I checked the percentage of rides used based on different membership types and seasons. The pie chart shows, how the users ride in different seasons. Approximately **83%** of total **one-time** users ride in **summer** while **70%** of **subscribers** ride in **summer**. So, the one-time users exhibit a strong affinity for the summer season.



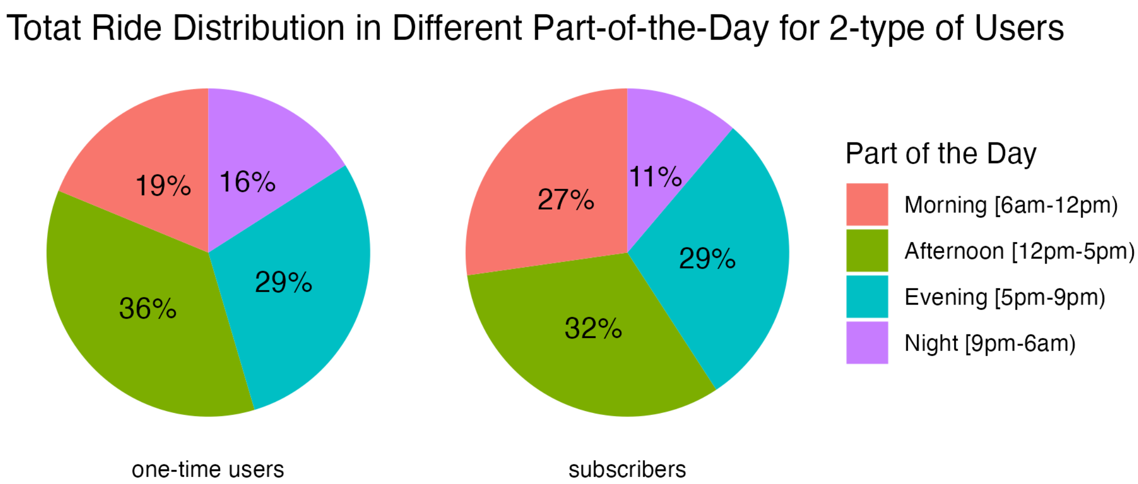
5. Analyze the Ride Pattern Across Different Parts of the Day

In this part of the report, I have analyzed the ride behavior in **different parts of the day**, the **busiest** and **modest** part of the day, and how the **membership type affects** the **counts**.

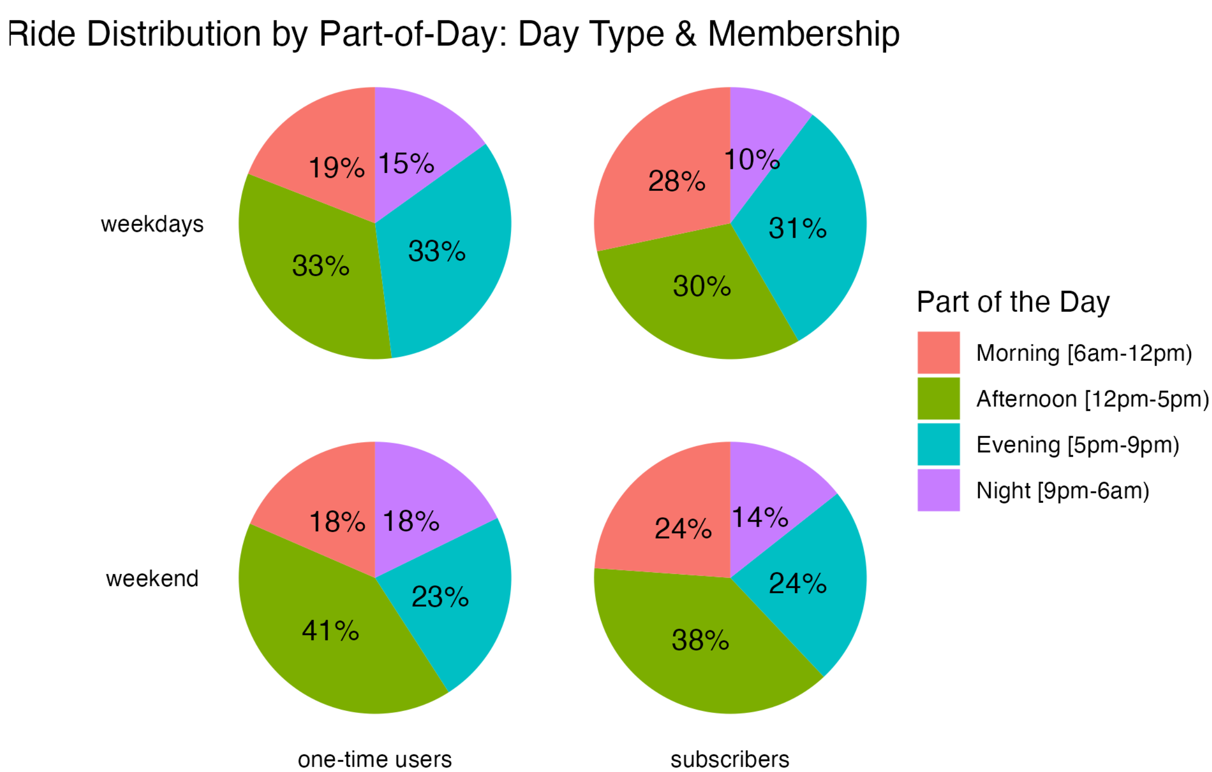
I. The below figure represents the total counts and respective percentage of rides. The **busiest** part of the day is **afternoon** time [**12 PM to 5 PM**). A total of **33.57%** of rides were recorded on those **5 hours**. The **second busiest** time frame is the **evening** [**5 PM to 9 PM**) time. **29.4%** of total rides were **recorded** in that **period**. If we consider **these two time frames** then **62.97%** of **rides encounter** in these **9 hours** only.



II. The below visual shows how the two **membership types consumed** the rides in **various parts** of the **day**. According to the visual, **Afternoon** time is the **pick time** for **both types** of **members**. Around **36%** of **one-time use**r rides in the **afternoon** while it's **32%** of the **subscribers** do. **29%** of **both types** of riders **prefer evening** time. **27%** of the **subscribers** are **morning** riders while the percentage is a bit lower and **19%** for **one-time users**. In the **nighttime**, **16%** of the **one-time users** ride while only **11%** of **subscribers** prefer night rides.



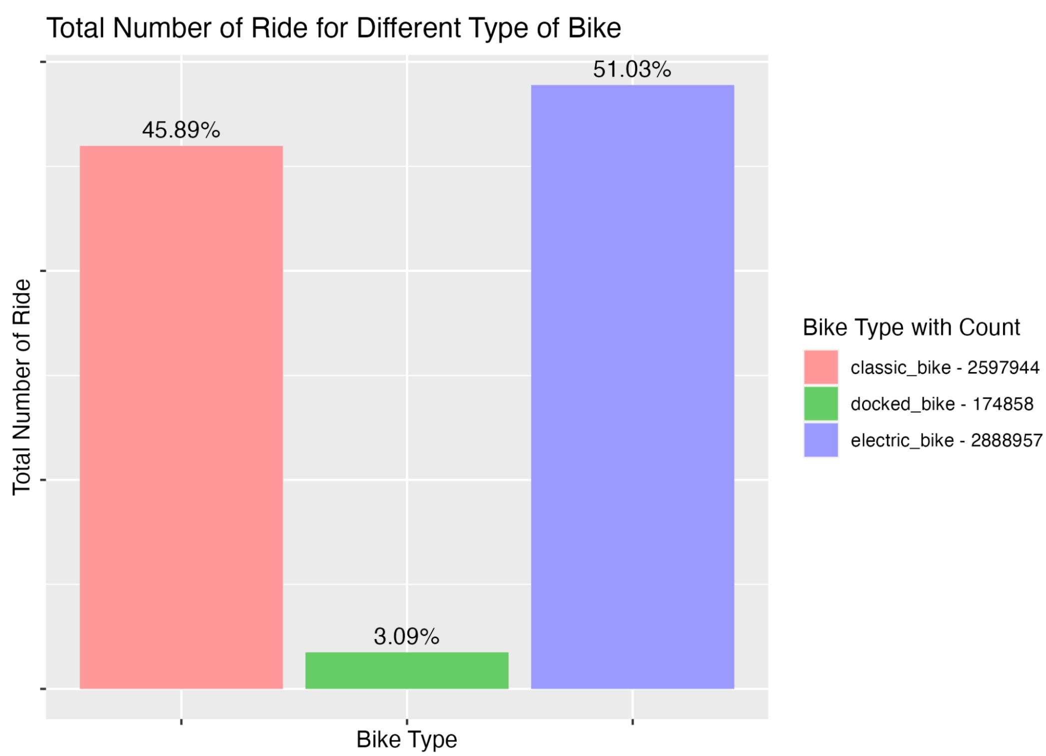
III. The below **visualization advocates** that, the **percentage** of **morning** and **nighttime** rides are **almost similar** (with slight changes) on **weekends** and **weekdays** from **both types** of **users**. But there is a **sharp increase** in the **percentage** of **Afternoon** rides for **both types** of **users during** the **weekends** and to **adjust** the increase in the **afternoon** there are **similar drops** in the **evening rides** which is **9-10%** for **one-time users** and 7-8**%** for **subscribers**. This means that on **weekends** the **afternoon** ride **percentage increases** and on **weekdays** the **evening** ride **percentage increases** and **vice-versa.**



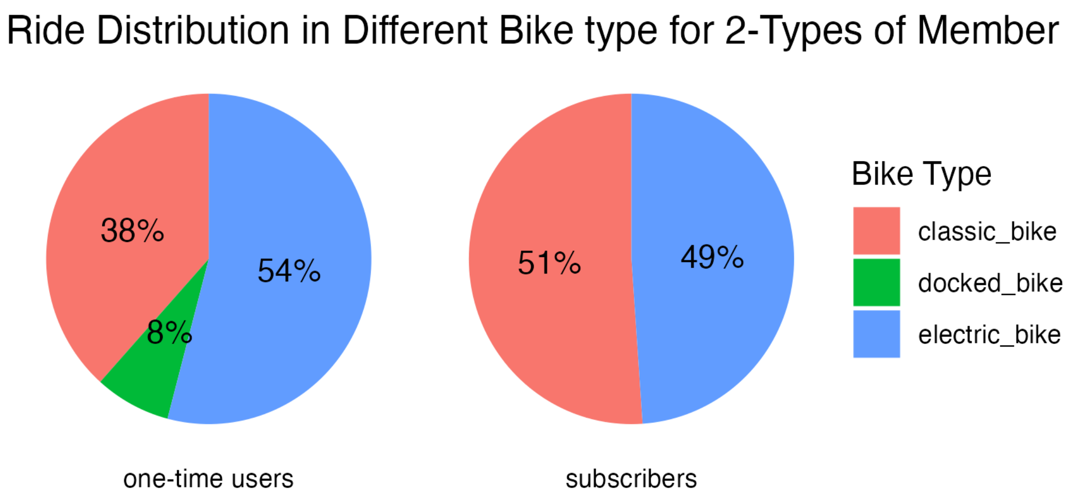
6. Examine the Riding Trends Based on the Type of Rideable Vehicles

In this **segment**, I analyze how the **riding trend** is **distributed** in **different rideable types**. The company has **three** types of **bikes**: **classic**, **docked**, and **electric**. I have analyzed this feature in some specific steps that are:

I. The below visual expressed that, **Electric** bikes and **Classic** bikes are mostly used. More than **51%** of the riders choose **electric** bikes and almost **46%** of the riders **prefer classic** bikes. Only **3%** of the users **prefer docked** bikes.



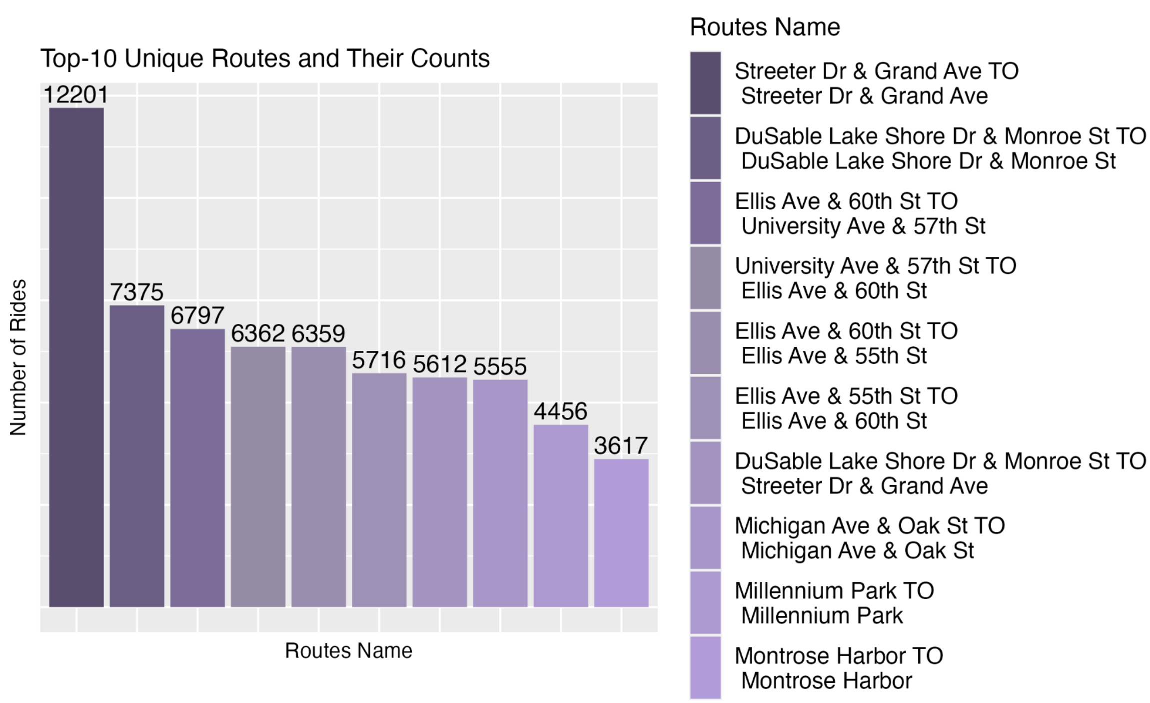
II. After analyzing the total rides of different bike types, I further investigated the **membership-based** ride distribution of three bike types. From the analysis, it is clear that **subscribers** only used **classic** and **electric** bikes and the **ratio** is **almost** the **same** **51-49%**. Within the **one-time user** category, **electric** bikes emerge as the **dominant** choice, constituting a substantial **54%** of the total rides. Further investigation into the **one-time user** group reveals that **38%** opt for **classic** bikes, showcasing a diverse range of preferences. Additionally, a noteworthy **8%** of users in this category opt for **docked** bikes.



7. Location Based-Analysis

Within this **section** of the report, I will find out which **locations** are mostly **favorites** among **various groups** of **users**.

I. There are a total of **1,675 unique start stations** and **1,693 unique end stations** are recorded in the year 2022.

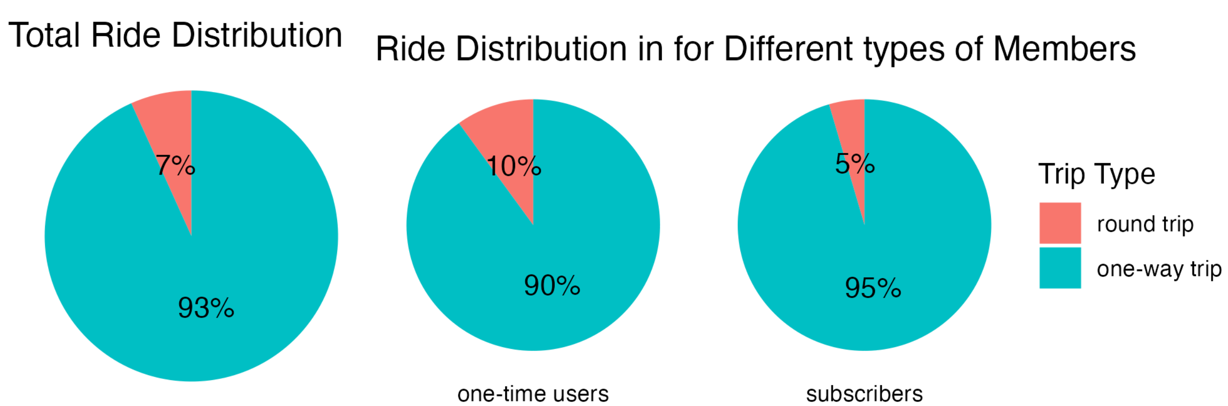


II. The above figure contains the most **commonly** used **routes** of the **users**. The **top** most used **route** is "**Streeter Dr & Grand Ave TO Streeter Dr & Grand Ave**" with t a count of **12,201**. According to the graph, it's clear that **some** of the **routes start** and **end** at the **same station** indicating a **round trip**. While investigating deeper, I found **50%** of the **top 10 routes** are **considering** a **round trip**.

III. Further, my **intention** of **investigation** moves one step deeper and I **checked**- are these **round trips** somehow **co-related** with the **membership types**? And my **hypothesis** has some **numbers** to prove it. I saw both types of members have **different** routing **patterns**. In the top **10** routes of **one-time user**s, **7** of them are considered as a **round trip** which is **0** for the **subscribers.** Also, there are **NO common routes** found in their **individual top 10** lists. Further, I checked the **common routes** from top **individual** routes for **different** types of users and the results are:-

1. **5 out of 50** routes are common.
2. **22 out of 100** routes are common.
3. **177 out of 500** routes are common.
4. **408 out of 1000** routes are common.

IV. From the below visualization, there is a clear indication of how the **one-time users** ride more round trips relative to **subscribers**. **10%** of **one-time users** ride **round trips** which are **5%** for **subscribers** and **7%** for **both** types of users.



PHASE-5: SHARE

In this segment I present the key insight of the project:-

1. During the time frame from **1st January** to **31st December** 2022 approximately **5.66 M** number of ride recorded.
2. Out of these, approximately **2.32 M** rides are carried out by **one-time users**, while **subscribers** account for around **3.4 M** rides. This indicates that subscribers are the predominant user group, with their ride count exceeding that of **one-time users** by a noteworthy **1.44 times**.
3. The cumulative **trip duration** tallies up to around **1.5 million hours**, with an **average** trip lasting approximately **16 minutes**.
4. While subscribers exhibit a considerable ride count, **one-time** users experience **trip durations** that are **1.23 times greater** than those of **subscribers**. Consequently, **one-time users** showcase a higher **average ride duration** (approximately **22 minutes**), surpassing that of **subscribers** (approximately **12 minutes**).
5. Trip duration remains fairly **consistent** across rides for **subscribers**, whereas there is a **degree** of **variability** within the **one-time users** group.
6. Overall, **Saturdays** emerge as the **busiest** day of the week, whereas **Mondays** witness the lowest ride counts. However, when considering ride counts based on **membership type,** a notable disparity becomes apparent. **One-time users** exhibit a preference for **weekends**, with **37%** of their **total rides** occurring during this time, while **subscribers** predominantly ride on **weekdays**, accounting for **75%** of their **rides**. This deviation can be attributed to potential differences in their usage purposes, where **one-time users** may consist largely of **visitors** or **tourists**, while **subscribers** likely comprise **professionals** or **students**.
7. As a general trend, **July** stands out as the month with the **highest** ride count, while **January** records comparatively **modest numbers**. The period from **May to October** is

deemed the **summer** season, accounting for over **75%** of rides during this span. Interestingly, when examining membership types, a prominent pattern emerges: **one-time users** primarily opt for rides during the summer months, constituting **83%** of the total count, while **subscribers** exhibit **slightly fewer** counts at **70%**.

8. The most **hectic period** of the day is the **afternoon** (from 12 PM to 6 PM), accounting for **33.57%** of the total ride count. When focusing on the top two time frames, the **evening slot** also emerges as significant alongside the afternoon hours. Collectively, these **nine hours** encompass nearly **63%** of the **overall** rides. When categorizing by membership type, the **dominant** pattern **persists** in the **top two time frames**. However, a shift occurs when introducing another parameter: weekdays and weekends. During **weekends**, **afternoon** rides observe an **increase** while **evening** rides **decline**, **affecting** both **user types**. The **opposite trend** is observed during **weekdays**.

9. The dataset showcases **three** distinct **bike** types: **Electric bikes (51%), Classic bikes (46%)**, and **Docked bikes (3%)**. Among **one-time users**, **54%** opt for **electric bikes**, **38%** favor **classic bikes**, and the remaining **8%** choose **docked bikes**. In contrast, **51%** of **subscribers** select **classic bikes**, while **49%** opt for **electric bikes**.

10. The dataset encompasses a total of **1,675** unique **starting stations** and **1,693** unique **ending stations**. Among these, the most frequently traveled route is "**Streeter Dr & Grand Ave TO Streeter Dr & Grand Ave,**" with a substantial count of **12,201 rides**. Within the top **10 routes**, **five** are **round trips**. Remarkably, **members** of **both types** exhibit **distinct preferences** in route selection. Interestingly, there is **no overlap** between the **top 10** individual **routes**, while **five** routes are shared among the **top 50 individual routes**. Furthermore, among the **top 10** routes based on **membership type**, **one-time users** have **seven round-trip routes**, while **subscribers** exclusively opt for **one-way trips**. Notably, **one-time users** account for **10%** of **round-trip rides**, while **subscribers** make up only **5%** of such rides.

PHASE-6: ACT

In this section, I've put forward several suggestions that have the potential to enhance Cyclistic's marketing strategy.

1. **Programmatic Advertisement**

   Currently, programmatic advertising has gained significant popularity as a marketing approach. Businesses are structuring segments to effectively reach their desired audiences. This involves leveraging analyzed data to pinpoint the target demographic.

   I. For one-time users, the marketing team employs personalized strategies, delivering ads through social media or email to encourage subscriptions.

   II. Additionally, focusing on specific regions or areas near where frequent one-time users travel can be an effective approach.

   III. Attracting fresh customers involves targeting individuals expressing a recent intent to travel to the city.

2. **Hotspot Based Marketing**

   An alternative strategy could involve on-location marketing at key spots. The fact that 75% of rides were taken during the summer, likely due to numerous festivals or an influx of tourists, suggests potential strategies. For example, setting up a booth at these festivals, tourist hubs, or college/university events could be effective. Another option could involve placing advertising billboards in highly frequented starting or ending stations, which are preferred by one-time users.

3. **Increase Partnership with Delivery/Logistics**

   An alternative approach could involve forming alliances with delivery or logistics service providers. These entities consistently require bicycles year-round, potentially presenting an opportunity to tap into a pool of potential subscribers.

4. **Introduce Half-Yearly Subscription**

   In the analysis, 75% of the users are summer riders. So maybe introducing a half-yearly subscription may influence the one-time users to subscribers.

5. **Discount to Long-time Users**
   Certain occasional users exhibit a longer average ride duration compared to the norm. Identifying these users and extending subscription offers to them could be a strategic move.

Conclusion

The CYCLISTIC bike share analysis project has successfully exemplified the implementation of the data analyst principles: Ask, Prepare, Process, Analyze, Share, and Act. In embracing the complete data analyst lifecycle, this project has not only provided insights into ride patterns but has also showcased the power of data-driven decision-making. By adhering to these principles, I have unlocked the potential of data to enhance operational efficiency and user experiences within the ride-sharing service.

References

Github link: https://github.com/AhmedDiderRahat/ride_history_insights.git
Linkedin Post:
