

Capital Bikeshare Membership Analysis

*JaeHo Bahng, Naomi Yamaguchi, Samantha Moon,
Jacky Zhang, Bella Shi*

*DSAN-5100 Final Project
Georgetown University
December 06, 2023*

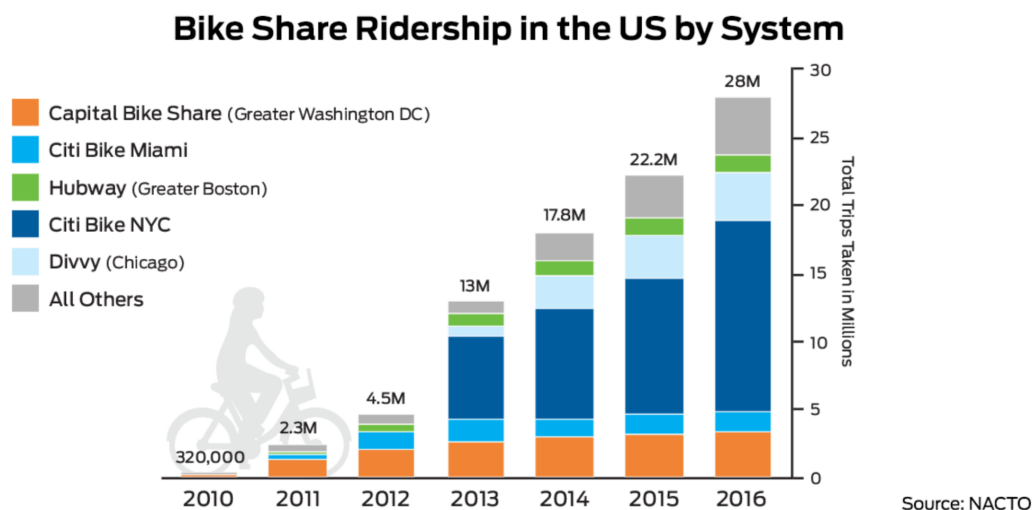
Table of Contents

I.	Introduction.....	3-5
II.	Data.....	6
III.	Data Cleaning.....	7-8
IV.	Exploratory Analysis.....	9-16
V.	Hypothesis Testing.....	16-23
VI.	Classification.....	24-32
VII.	Conclusion.....	33
VIII.	Works Cited.....	34

I. Introduction

In August 2008, the District of Columbia was the first city in North America to launch a bike sharing system. Capital Bikeshare, which was then referred to as SmartBike DC, offered 120 bikes at 10 stations in Washington DC, and 1600 people joined the platform during the first two years of its operation. The name Capital Bikeshare came about during 2010 as the result of a joint project between Arlington, Virginia, Montgomery County, Maryland and the District of Columbia Department of Transportation (DDOT). Until 2020, Capital bikeshare continued to grow in terms of ridership and geographic coverage and introduced new features such as mobile app integration, e-bikes, and improved payment systems for better customer experience. With the start of COVID in 2020, many bike sharing programs, including capital bikeshare had to cope with negative effects due to lockdowns and social distancing measures. However, the industry saw a resurgence as people sought alternative, socially distanced transportation options, as capital bikeshare has currently surpassed the number of total rides compared to its pre-covid peak during 2019.


Figure 1: Bike Share Market Growth



Despite the global growth of bike sharing companies, Capital Bikeshare is facing many challenges regarding revenue shortfalls. In 2014, Capital Bikeshare earned enough revenue to cover only 70% of its expenses, and in 2016 they were able to cover 77% of expenses. Capital Bikeshare constantly relies on local governments to cover its deficits. One of the main factors that puts capital bikeshare to a disadvantage is its docking stations. Newer bike share systems are dockless allowing riders to walk less from docking stations, and provide far cheaper services where apps can just indicate the bikes' locations. Furthermore Capital Bikeshare's revenue shortfalls that were calculated above only include operating costs and excludes the capital costs. In other words, the costs of capital bikeshare constantly expanding docks and bikes every year is not taken into consideration. With the continuing coverage expansion of capital bikeshare and the thriving of dockless bike share companies, capital bikeshare may face a serious issue as time progresses.

With a disadvantage with the system being a docked bikeshare system, the company will have to turn to memberships, and rental fees to make up for its various costs. How is the current membership managed? The membership is distinguished between two groups: members, and casual users. The members pay an annual fee of \$95 to get a free bike unlock with 45 free minutes on a classic bike, then paying 5 cents for every minute, and pay 10 cents every minute for e-bikes with no free minutes. On the other hand, casual users can either buy a single pass or a day pass where they pay \$8 to experience the same benefits of a member for a day, or pay unlock fees with no free minutes as just a single ride user.

Figure 2: Membership fees for Capital Bikeshare

	Single ride \$0.05/min Get the app →	Day pass \$8/day Get a day pass →	Capital Bikeshare \$95/year Join →
Bike unlocks	\$1	Free	Free
Classic bike prices	\$.05/min	45 mins free, then \$.05/min	45 mins free, then \$.05/min
Ebike prices	\$0.15/min	\$0.10/min	\$0.10/min
Bike Angels			

In this project, we plan to investigate the differences in Capital Bikeshare usage between members and casual users in order to lay the foundations of a renewal or modification of memberships. We can modify membership fees, divide memberships into various segments, implement marketing promotions, but all under the circumstance that members and casual users make use of capital bikeshare differently to begin with. We plan on conducting exploratory analysis, hypothesis testing, and classification models to establish a significant difference for further progress of membership renewals.

II. Data

The data was retrieved from the “System Data” tab of the official capital bikeshare website and files from the year 2019 to 2022 were used for analysis. All files were downloaded as an individual monthly csv file and the format of files was slightly modified after April 2020. The columns of the changed datasets are as below. In order to perform analysis on the dataset as a whole, we will clean and merge datasets in the next step of data cleaning.

Columns (2019.01 - 2020.03)	
Attributes	Definition
Duration	Duration of rent in seconds
Start date	Rent session start date
End date	Rent session end date
Start station number	Start station ID
Start station	Start station name
End station number	End station ID
End station	End station name
Bike number	Bike number
Member type	Member/Casual

Columns (2019.01 - 2020.03)	
Attributes	Definition
Ride_id	Ride id
Rideable_type	ebike/regular
Started_at	Session start date
Ended_at	Session end date
Start_station_name	Start station name
Start_station_id	Start station id
End_station_name	End station name
End_station_id	End station id
Start_lat	Start station latitude
Start_lng	Start station longitude
End_lat	End station latitude
End_lng	End station longitude
Member_casual	member

III. Data Cleaning

No major cleaning was necessary within each csv file that carried the data for each month, but minor changes had to be made to merge the two types of datasets and fill in missing values that were only on one format of the data.

Firstly, it is easy to notice that the column names are different on the two different formats of datasets, and it is essential for us to unify column names to merge datasets together. The column names in the datasets after 2020.04 seem to be the more sophisticated column names with the under scores and detail in names, and therefore we will change the column names of the earlier format into the later format. Furthermore, we will eliminate the columns that we will not use for analysis. “Bike number” from the earlier format and “rideable_type” from the second format will be dropped.

Figure 3: columns on two types of datasets

```
Columns for dataset 201901 - 202003
Index(['Duration', 'Start date', 'End date', 'Start station number',
      'Start station', 'End station number', 'End station', 'Bike number',
      'Member type'],
      dtype='object')

Columns for dataset 202004 - 202212
Index(['ride_id', 'rideable_type', 'started_at', 'ended_at',
      'start_station_name', 'start_station_id', 'end_station_name',
      'end_station_id', 'start_lat', 'start_lng', 'end_lat', 'end_lng',
      'member_casual'],
      dtype='object')
```

When the datasets are merged, blank data points with NaN values will inevitably be present because some columns are only on one of the data formats. [Duration, start_lat, start_lng, end_lat, end_lng] are columns with NaN values that must be filled in. Duration is simple. We can subtract the “started_at” column from the “ended_at” column to retrieve the number of seconds

the bike rental session occurred. The latitude and longitude of the starting and ending stations took a little more extra steps. First, I made a dictionary of the station IDs with its corresponding longitude and latitudes from the datasets after 2020.04. Then I filled in latitude and longitudes of both the starting stations and the ending stations. After filling them in, I realized that there were 2 stations that still had NaN values and noticed that these stations were probably withdrawn before 2020.04, and were worth about 1000 rows of data from my entire merged dataset which was 11 million rows. Since 1000 rows was not a significant amount, I eliminated these rows for convenience of future analysis.

Figure 4: NaN values of merged dataset

Duration	started_at	ended_at	start_station_id	start_station_name	end_station_id	end_station_name	member_casual	start_lat	start_lng	end_lat	end_lng
230.0	2019-01-01 00:04:48	2019-01-01 00:08:39	31203.0	14th & Rhode Island Ave NW	31200.0	Massachusetts Ave & Dupont Circle NW	Member	NaN	NaN	NaN	NaN
1549.0	2019-01-01 00:06:37	2019-01-01 00:32:27	31321.0	15th St & Constitution Ave NW	31114.0	18th St & Wyoming Ave NW	Casual	NaN	NaN	NaN	NaN
177.0	2019-01-01 00:08:46	2019-01-01 00:11:44	31104.0	Adams Mill & Columbia Rd NW	31323.0	Woodley Park Metro / Calvert St & Connecticut ...	Casual	NaN	NaN	NaN	NaN
228.0	2019-01-01 00:08:47	2019-01-01 00:12:35	31281.0	8th & O St NW	31280.0	11th & S St NW	Member	NaN	NaN	NaN	NaN
1300.0	2019-01-01 00:12:29	2019-01-01 00:34:10	31014.0	Lynn & 19th St North	31923.0	Columbia Pike & S Taylor St	Member	NaN	NaN	NaN	NaN
...
NaN	2022-12-29 11:50:13	2022-12-29 12:00:30	31600.0	5th & K St NW	31655.0	New Jersey Ave & F St NW	member	38.903040	-77.019027	38.897108	-77.011616
NaN	2022-12-05 19:14:05	2022-12-05 19:22:10	31600.0	5th & K St NW	31655.0	New Jersey Ave & F St NW	casual	38.903040	-77.019027	38.897108	-77.011616
NaN	2022-12-05 12:51:38	2022-12-05 12:56:16	31600.0	5th & K St NW	31655.0	New Jersey Ave & F St NW	casual	38.903068	-77.018793	38.897108	-77.011616

No extensive feature extraction was done, but a few features were added for convenience. ‘Year’, ‘Month’, ‘Day’, ‘Hour’ columns were added, which were extracted from the ‘started_at’ column and a ‘distance’ column was added to calculate the distance of the rental session by converting the coordinates of stations to distance in miles. Now we are ready to perform exploratory analysis on our merged dataset with 11 million rows and 16 columns.

IV. Exploratory Analysis

We conducted some exploratory data analysis to obtain a better understanding of the Capital Bikeshare data. After the datasets were cleaned and merged, we were left with a large set of ride data to work with and visualize. Our goal was to summarize the dataset using various plots to understand the differences between member and casual rider activity and the respective usages of bikeshare. We demonstrated a group of visualizations, including histograms, box plot, pie and line charts to perform this analysis.

We looked at key features of the cleaned dataset, which include: ride duration, start and end times, date information, member and casual classification. We initially plotted the frequencies of the spread of ride durations, distinguishing between member and casual riders.

Figure 5: Ride Duration Histogram

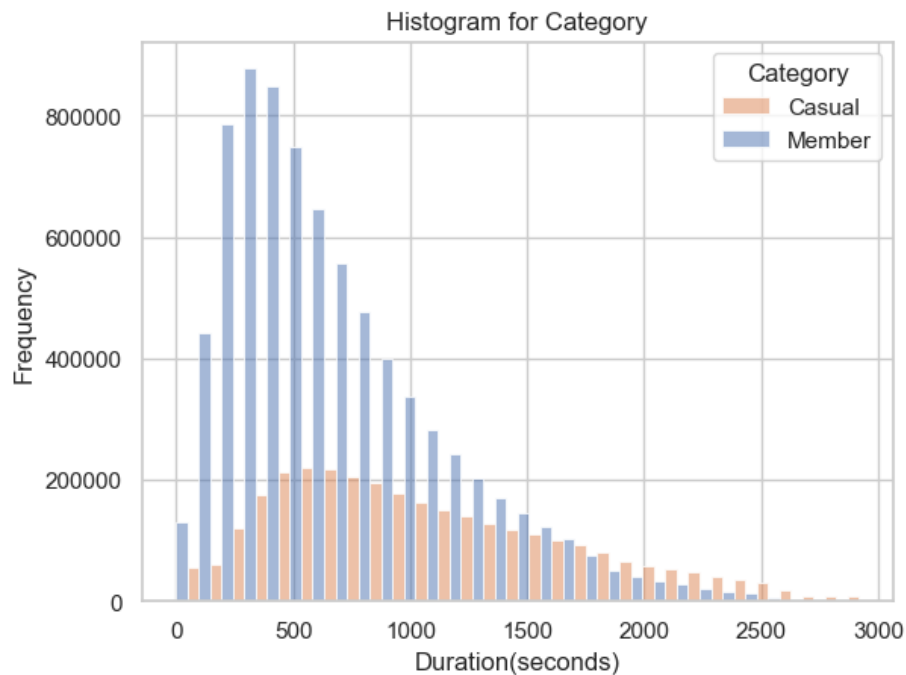
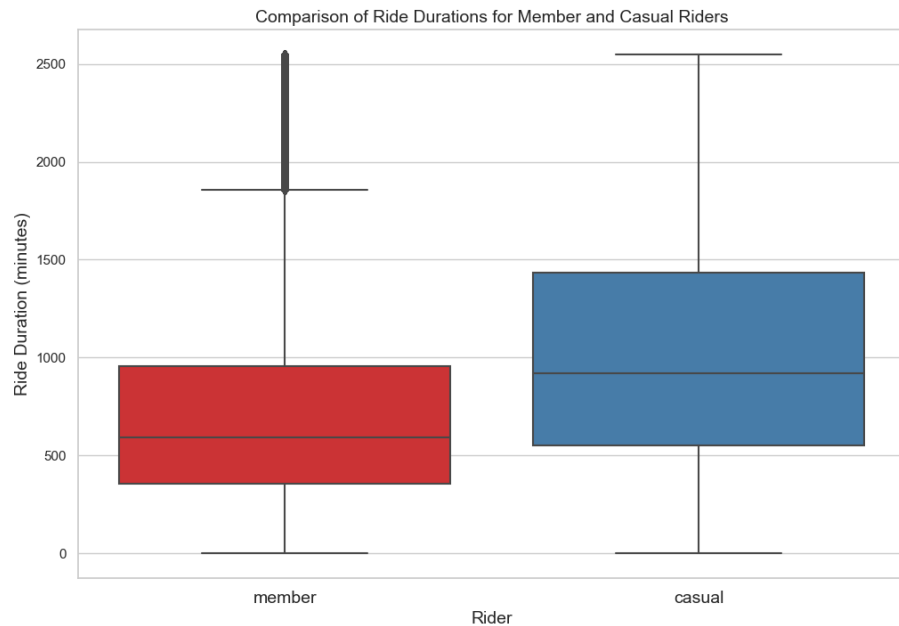


Figure 3 is a representation of the member and casual rider duration spread across the entire dataset. It's evident that the ride frequencies of member riders far exceed that of the casual rider, by observing the blue bars as member rides. If we observe the duration of these two groups, we notice the casual rider distribution is more steady-sloped, displaying a wider spread than the member rider. Member rides peak from 400-500 seconds, which is approximately 6.6-8.3 minutes per ride. The member distribution of rides is heavily skewed to the right whereas casual riders are still skewed, but not as distinct. We can interpret this in several ways. It looks as though casual rides are more likely to be longer in duration than member rides, while less consistent. Member rides are consistent and consistently shorter in duration. This is possibly due to the nature of the member vs casual rides. It's likely that a large bulk of member rides are dedicated to consistent commutes, whether that be for work, grocery shopping, etc. The explanation for the wider distribution of casual rides might be that there is a larger variety of reasons for using bikeshare amongst casual riders.

To observe duration measures further, we created a box plot to take a closer look at the spread of the data. Furthering the observation that member rides are shorter in duration, a box plot will allow us to observe central tendencies of the data and the range in which much of the data lies.

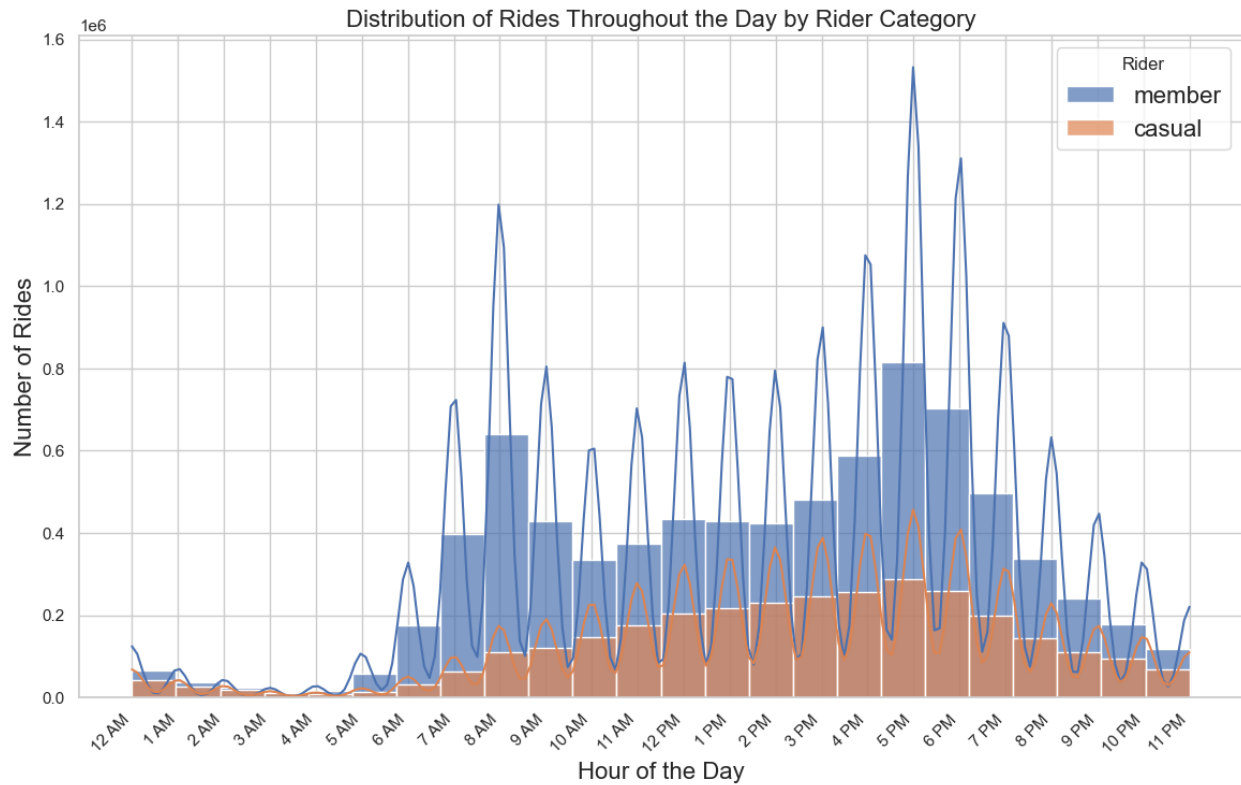
Figure 6: Ride Duration Box Plot



From Figure 4, we can see that the median of the member ride durations lies lower than the casual ride duration median. We also observe that the range of the casual ride durations is larger, with a higher maximum duration. Fifty percent of the data for casual rides are more spread out than that of the member rides, as member ride durations are more concentrated.

After studying the ride durations between members and casual riders, we move on to explore the distribution of rides throughout the day. This will allow us to get a better idea of whether or not there are trends in daily ride activity, so that we might be able to interpret the usage patterns between members and casual riders. We initialize this exploration by adding a new column called 'Hour' to our dataframe, extracting the hour component from the 'started-at' column. This will represent the hour of the day when each bike ride started. We then plotted a histogram to display the daily ride distribution between the two categories by plotting the hour of the day by the number of rides per hour.

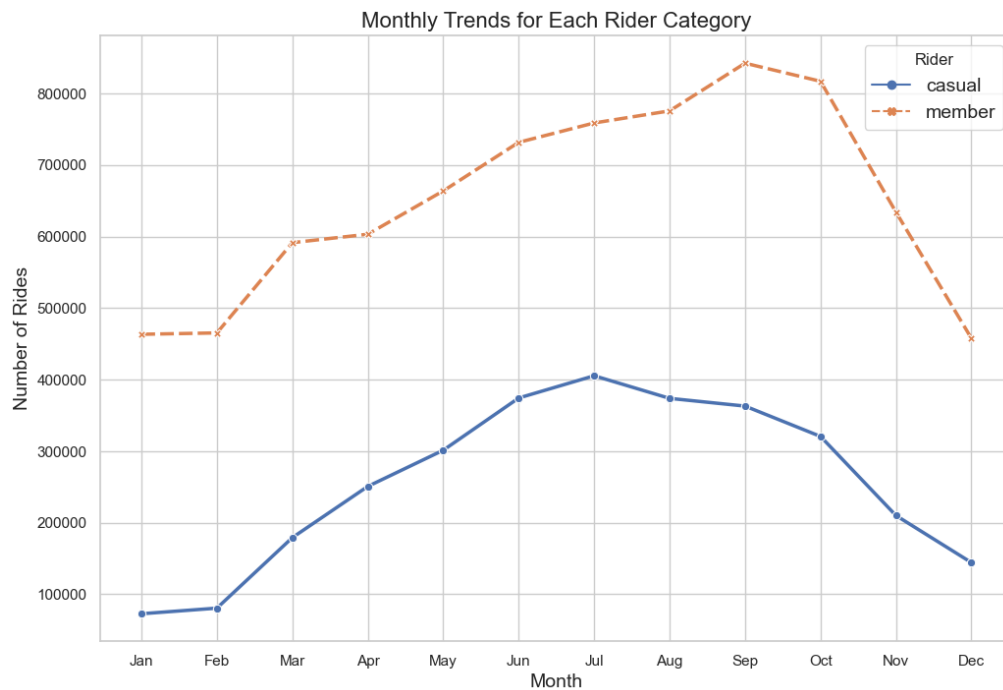
Figure 7: Daily Ride Distribution Histogram



We observe two distinct peaks in the member distribution. The tallest peak around 5-6pm and the second tallest peak around 8-9am. Casual rides do not reflect these peaks, as the data is again, more spread out. The interpretation of this plot is similar to the interpretation of the duration plots, and it furthers the idea that a large portion of Capital Bikeshare memberships are used for consistent transportation, such as work commutes. This visualization provides heavy evidence for this theory. We can also interpret the wide spread of the casual rides similar to the way we interpreted it before, adding the fact that the daily ride distribution leads us to believe that casual rides are more often for leisure in nature, or overall less consistent than member rides.

We move on to analyze the monthly ride frequencies of the data to observe any patterns. Since we have monthly data, this was an easy task to accomplish. We plotted the months by the frequency of rides per month, and distinguished between member and casual categories.

Figure 8: Monthly Ride Distributions Line Plot

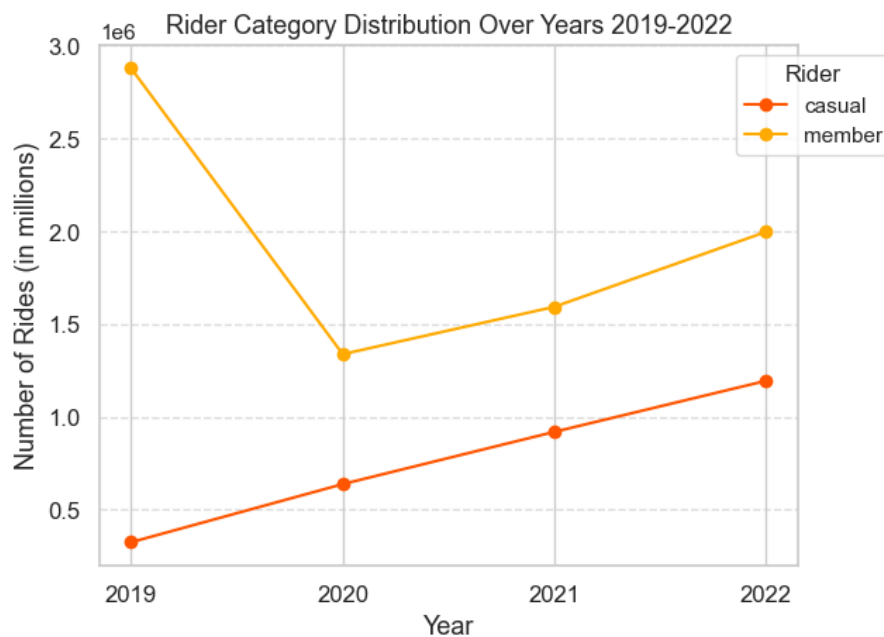


The trends between member and casual rides resemble each other pretty closely, but we can observe different peak months for both. Member rides peak in the month of September, with October as a close runner-up. There is no significant interpretation of this discovery, but we can still gain valuable insight from this plot through studying the trend of the casual rider. Casual rides peak in July, and are most concentrated amongst the mid-late summer months. This leads us to believe that the usage of bikeshare amongst casual riders is majorly determined by the convenience provided by warmer weather. It also leads us to draw conclusions around the nature

of bikeshare usage amongst casual riders, which are most likely leisure activities, or just heavily dependent on the weather. It's no surprise that the lowest frequencies in rides occur in colder months, as the weather might deter riders from committing to riding bikes. Overall, we gain exclusive insight through this model to the monthly patterns of Capital Bikeshare usage, which can be utilized in the future in maximizing profits in peak seasons.

Since we've analyzed the duration, hourly, and monthly distribution of rides between members and casual riders, it would aid us to visualize the differences in these two groups through a wider lens. We have data ranging from years 2019-2022, so we'll observe the ride distributions between members and casual riders over the four years to detect any trends in the usage of bikeshare.

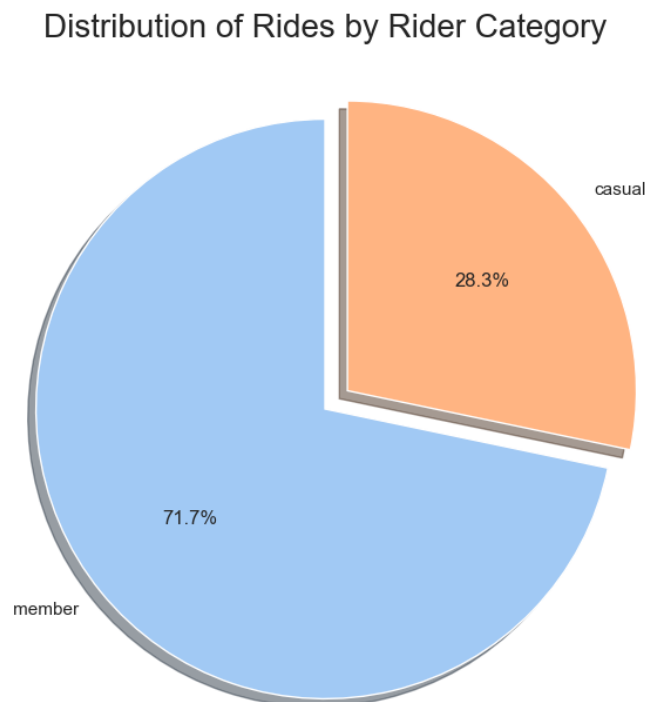
Figure 9: Member vs. Casual Rides Over Years 2019-2022



From this plot we can see a significant decline in the frequency of member rides between years 2019 and 2020. This is most likely due to the initial impact of the COVID-19 pandemic. Member rides picked back up after 2020 and have been increasing at a favorable rate. Casual rides however, have been increasing at a consistent rate since 2019. This information could aid in analyzing patterns of growth for Capital Bikeshare, as it is clear that bike usage is on the rise. This poses an opportunity for marketing, possibly for more casual to membership conversion, or more profitability through raising ride fares.

Taking another glimpse at the distribution of member rides and casual rides in the entirety of the dataset, we'll create a simple pie chart to visualize the frequencies of each category.

Figure 10: Distribution of Rides Pie Chart



We observe that 71.7% of the ride data belongs to member rides, whereas 28.3% of the data belongs to casual rides. This is a pretty large difference, and it suggests that there is a significant difference in the usage of Capital Bikeshare by members and casual riders. To study this significance, we will perform hypothesis testing.

V. Hypothesis Testing

The study explores the extensive dataset provided by the Capital Bikeshare system to uncover the distinct usage patterns of its two primary user categories: members and casual riders. Through a series of hypothesis tests, particularly T-tests, this research aims to dissect the variances in ride durations between these groups over a span of several years. The intention is to provide a nuanced understanding of how different users engage with the bike-sharing service, highlighting potential variations in their riding habits.

The initial phase of the analysis involved applying a T-test to assess the average duration of rides taken by members compared to casual users. This test revealed a marked disparity in the riding behavior of the two groups. On average, members recorded ride durations of about 710.14 minutes, while casual riders exhibited a higher average duration of approximately 1037.69 minutes. The substantial T-statistic value of -948.30, coupled with a negligible P-value, strongly suggests that these differences are statistically significant and not merely random variations.

Figure 11: T.test Duration|Member

```
Member Mean Duration : 710.1391102552708
Casual Mean Duration : 1037.6919172908017
T-statistic: -948.3039531458
P-value: 0.000000000000000000000000000000
```


The research extended to an annual examination of ride durations from 2019 through 2022, employing T-tests for each year. This longitudinal analysis consistently demonstrated that casual users generally spend more time per ride compared to members. In 2019, the average ride duration for members was 706.81 minutes, in contrast to 1151.27 minutes for casual users, indicated by a T-statistic of -522.51. This pattern persisted in the subsequent years. The year 2020 saw members with an average of 756.29 minutes and casual users at 1155.99 minutes, resulting in a T-statistic of -469.80. The trend slightly modified in 2021, with members at 708.49 minutes and casuals at 1018.85 minutes, and a T-statistic of -452.05. The 2022 data aligned with previous years, showing members with an average of 685.07 minutes and casuals at 956.49 minutes, with a T-statistic of -450.12.

Figure 12: T.test Duration|Member Yearly

Duration testing 2019
Member Mean Duration : 706.8051580869256
Casual Mean Duration : 1151.2655089326556
T-statistic: -522.5065999620
P-value: 0.000000000000000000000000000000

Duration testing 2020
Member Mean Duration : 756.2917174932938
Casual Mean Duration : 1155.9983777548127
T-statistic: -469.8007640537
P-value: 0.000000000000000000000000000000

Duration testing 2021
Member Mean Duration : 708.4943655510165
Casual Mean Duration : 1018.8490588423092
T-statistic: -452.0485809585
P-value: 0.000000000000000000000000000000

Duration testing 2022
Member Mean Duration : 685.0668944502081
Casual Mean Duration : 956.4866722290163
T-statistic: -450.1210059029
P-value: 0.000000000000000000000000000000

The consistency in these findings across multiple years highlights a distinct and persistent difference in how members and casual riders use the Capital Bikeshare system. The shorter ride durations of members might suggest their use of the service for more routine or

commute-focused activities, while longer durations for casual riders may indicate usage for leisure or exploration. The extremely low P-values observed in each test reinforce the validity of these results, suggesting that these patterns in ride durations are reflective of intrinsic differences in the behaviors of these two user cohorts.

In continuing our analytical exploration of the Capital Bikeshare system, we turn our attention to a Chi-squared test to examine the relationship between the year of usage and the type of user – whether they are members or casual users. This statistical test offers insight into how user behavior may have varied over the years and whether these changes are significant.

The Chi-squared test was performed on a contingency table comparing the counts of casual and member users across different years, from 2019 to 2022. The observed frequencies in the table were as follows:

Figure 13: Chi-Squared Test Year|Member

```

Year|Member Table
member_casual  casual  member
Year
2019           310434  2860673
2020           665744  1346899
2021           927156  1595542
2022          1188946  1994291

Chi-square statistic: 773802.371502223
P-value: 0.0
Expected frequencies:
[[ 900480.66164999 2270626.33835001]
 [ 571518.4319877  1441124.5680123 ]
 [ 716355.7597341  1806342.2402659 ]
 [ 903925.14662821 2279311.85337179]]

```

The results of the Chi-squared test were striking. The Chi-square statistic was calculated to be 773,802.371, and the P-value was 0.0, indicating that the differences observed in the table were extremely statistically significant. The expected frequencies, which represent what the numbers would be if there were no relationship between the year and the type of user, were also calculated.

The results of the Chi-squared test were striking. The Chi-square statistic was calculated to be 773,802.371, and the P-value was 0.0, indicating that the differences observed in the table were extremely statistically significant. The results of expected frequencies highlight the disparity between the actual and the expected user counts, further emphasizing the significant relationship between the year and the type of user.

The Chi-squared test's findings suggest a notable shift in the user composition of Capital Bikeshare over the years. The increase in casual users relative to members, especially noticeable in the later years, could reflect changing patterns in how people engage with bike-sharing services. This shift might be influenced by various factors, such as changes in city tourism, commuting patterns, or even the broader adoption of bike-sharing as a leisure activity.

The exploration of Capital Bikeshare's extensive dataset continues with an in-depth analysis focusing on the hourly usage patterns of members and casual users. This segment of the study applies T-tests to compare the mean hours of usage between these two categories of riders, providing a nuanced understanding of how the service is utilized throughout different times of the day and across various years.

The initial step in this analysis involved conducting a T-test on the combined data set to compare the average hourly usage of members and casual users. The results revealed a noticeable difference in the usage patterns of the two groups. Members had an average hourly usage of about 13.89 hours, while casual users had a slightly higher average of 14.49 hours. The T-statistic was calculated at -184.77, accompanied by a P-value so small it approaches zero, signifying a statistically significant difference between the two groups.

Figure 14: T.test Member|Hour

```

Member Mean Hours : 13.894269952631676
Casual Mean Hours : 14.493881537247598
T-statistic: -184.7718997571
P-value: 0.000000000000000000000000000000

```

To gain a year-by-year perspective, T-tests were conducted for each year from 2019 to 2022, examining the hourly differences in usage between members and casual users.

Figure 15: T.test Member|Hour Yearly

```

Hour Difference Testing 2019
Member Mean Duration : 13.646487382514534
Casual Mean Duration : 14.528869260454718
T-statistic: -95.6752329605
P-value: 0.000000000000000000000000000000

Hour Difference Testing 2020
Member Mean Duration : 13.945640318984571
Casual Mean Duration : 14.672127724771084
T-statistic: -105.7669613119
P-value: 0.000000000000000000000000000000

Hour Difference Testing 2021
Member Mean Duration : 14.071411470208869
Casual Mean Duration : 14.38851606417906
T-statistic: -50.0514408261
P-value: 0.000000000000000000000000000000

Hour Difference Testing 2022
Member Mean Duration : 14.073279676837533
Casual Mean Duration : 14.467103636329993
T-statistic: -69.3911786482
P-value: 0.000000000000000000000000000000

```

This consistent pattern over four years highlights subtle yet important differences in how members and casual users engage with the bikeshare system. The higher average hourly usage among casual users across all years suggests that they might be using the bikes for more extended periods per session, possibly for leisure or touring purposes. In contrast, members, with slightly lower average hours, seem to use the service more for shorter, possibly routine trips.

In the final segment of our study on the Capital Bikeshare system, we turn our focus to the distances traveled by different user groups. Utilizing the T-test, a statistical method, we examined the distances covered by members versus casual users. This analysis was conducted

both on a combined dataset and on a yearly basis, providing a detailed view of travel patterns over time.

The combined data analysis for distance traveled revealed noteworthy differences between members and casual users. On average, members traveled for 1.1316 hours, while casual users recorded a slightly higher mean of 1.2027 hours. The T-statistic for this comparison was -41.5667, indicating a significant statistical difference, supported by an extremely low P-value.

Figure 16: T.test Distance|Member

```
Member Mean Hours : 1.131606281526995
Casual Mean Hours : 1.2026891000323163
T-statistic: -41.5667142940
P-value: 0.000000000000000000000000000000
```

A more granular analysis was conducted for the years 2019 through 2022, offering insights into how travel patterns evolved over time.

Figure 17: T.test Distance|Member Yearly

```
Distance Difference Testing 2019
Member Mean Duration : 1.0910735693959963
Casual Mean Duration : 1.0519490604310209
T-statistic: 28.6286413251
P-value: 0.000000000000000000000000000000
```

```
Distance Difference Testing 2020
Member Mean Duration : 1.1473874721844302
Casual Mean Duration : 1.2631473130770035
T-statistic: -19.3117739909
P-value: 0.000000000000000000000000000000
```

```
Distance Difference Testing 2021
Member Mean Duration : 1.1302872447974477
Casual Mean Duration : 1.2025651504160397
T-statistic: -64.9833722488
P-value: 0.000000000000000000000000000000
```

```
Distance Difference Testing 2022
Member Mean Duration : 1.1106918189808965
Casual Mean Duration : 1.1658051189052236
T-statistic: -58.9321792040
P-value: 0.000000000000000000000000000000
```

The results indicate a persistent pattern where casual users tend to travel longer distances compared to members. This could suggest that casual users are more likely to use the bikeshare for leisure or exploration, possibly involving longer routes. The significant T-statistics and near-zero P-values across the analyses underline the robustness of these findings. The year 2019 stands out as an anomaly, where members traveled slightly more than casual users. This deviation could be attributed to specific events or changes in user behavior during that year, warranting further investigation.

The application of T-tests to the Capital Bikeshare dataset has provided significant insights into the differing usage patterns of members and casual riders. Over several years, the tests consistently revealed a notable disparity in ride durations between these two user groups. Members generally exhibited shorter ride durations, suggesting more routine or commute-driven use, while casual riders tended to use the service for longer periods, likely for leisure or exploration. The statistical significance of these findings, evidenced by extremely low P-values, indicates that these patterns are not random but are inherent to the behavior of the two distinct user groups.

While the T-test offered valuable insights into the ride durations, another dimension of analysis comes from the Chi-Squared test, which could be used to examine categorical variables in the dataset. For instance, examining the relationship between the type of user (member or casual) and other categorical variables like station locations, time of day, or seasonality could reveal patterns in user preferences and behaviors. A significant Chi-Squared statistic in this context would indicate a non-random association between these categorical variables, further enriching our understanding of how different user groups engage with the Capital Bikeshare system.

The combined use of T-tests and Chi-Squared tests in analyzing the Capital Bikeshare data not only enhances our understanding of current user behaviors but also aids in predicting future trends and user needs. These insights are vital for strategic planning, marketing, and improving the overall user experience. Knowing the specific preferences and usage patterns of members and casual riders can guide the development of targeted initiatives, such as promotional campaigns or station placement strategies.

Building on the foundation laid by these statistical analyses, the next logical step in our exploration of the Capital Bikeshare data is to delve into classification methods. Classification algorithms in machine learning can be employed to predict user behavior based on historical data. For example, we can use these methods to predict whether a new rider will be a member or a casual user based on their riding patterns, or to identify which features most significantly influence ride duration and user type. This predictive modeling not only deepens our understanding of the data but also opens up avenues for proactive decision-making and service optimization in the realm of urban bike-sharing systems.

VI. Classification

Prior to initiating the development of our classification model, it is imperative to first determine the specific features within our dataset that are of interest and to strategize their effective implementation. This process commences with the composition of an elementary data cleaning and Exploratory Data Analysis (EDA). The purpose is to facilitate a comprehensive understanding of the dataset, thereby enabling us to identify features that will prove instrumental in the subsequent model training phase.

We implemented this idea by writing a function by utilizing the glob library. This function is designed to efficiently import multiple files from the specified directory, as indicated by the path “../data/2020*.csv”. When written this way, all of the data sets that are under 2020 will be seamlessly integrated into a singular data frame. On top of that, we created a function by incorporating a 'for' loop based on this concept so we can read every month of 2020 in one take. Moreover, in the function, we also used some data-cleaning techniques like creating a sub-data frame that only includes the column we need and dropping rows that contain 'NaN' values so we won't run into issues later in the modeling phase. Another critical aspect of our data preparation process involves ensuring data type consistency. This was achieved through the application of the 'Astype' function to columns such as 'start_lat', 'start_lng', 'end_lat', and 'end_lng', converting them to the 'float64' data type. This conversion is crucial for accurately calculating distances in miles, thereby avoiding data type-related errors. In the final steps of the data preparation, we employed binary encoding on the 'member_casual' column, transforming it into a binary classification of 0s and 1s. This step is vital for the model to classify between the two categories effectively.

Subsequent to the data preparation, we utilized a library called 'Goepy'. This library takes inputs such as 'start_lat', 'start_lng', 'end_lat', and 'end_lng' and then converts them into starting coordinates and ending coordinates. Utilizing these coordinates, 'Goepy' is capable of calculating the distance between two points in miles. This calculated distance was then stored in a newly created column within the data frame named 'distance_mile.'

After applying the cleaning algorithm and the distance calculation function provided by the Goepy library, the data spanning from 2019 to 2022 is transformed as follows:

Figure 18: Data

	Duration	distance	member_casual	start_lat	start_lng	end_lat	end_lng
0	231.0	0.658793	0	38.908600	-77.032300	38.910100	-77.044400
1	1550.0	1.889525	1	38.892244	-77.033234	38.918809	-77.041571
2	178.0	0.498394	1	38.922925	-77.042581	38.923389	-77.051833
3	228.0	0.421363	0	38.908640	-77.022770	38.913761	-77.027025
4	1301.0	3.207022	0	38.897315	-77.070993	38.858524	-77.103728

After getting the output from the two functions that clean and create the 'distance' column, our attention shifted towards the application of a suitable model. But a challenge presented itself before doing that; we realized that we had an issue with using an extensive data set for the model, as it was anticipated to prolong the model's execution and output delivery time. We wanted to know if there was a way to optimize the feature selection process and the amalgamation of datasets from different years.

So, to ensure there would not be an issue during the processing, features were selectively combined and analyzed on a yearly basis; then, we see if there were errors in each year, and in the end we combined them together. This method, although not enhancing the overall performance efficiency, had the advantage of reducing the volume of data processed at any given time. Nonetheless, this resulted in an average processing duration of approximately two minutes

per annual dataset. In the future, we may want to try to use Cuda from the RAPIDS suite for handling extensive datasets, leveraging GPU's parallel processing capabilities for improved efficiency.

After the cleaning and feature selection part, the focus shifted to model selection. We picked Decision Trees and Random Forests because of how those two models work. First, let's talk about Decision Tree; the reason that we picked Decision Tree is because this is a general algorithm that can handle different types of data without any problems. So this model is ideal as a baseline model to see if we can find any patterns within the data set. Moreover, it resonates with human problem-solving logic. However, there are limitations with the Decision Tree model, because of its simplicity, sometimes it will not capture the whole picture of the data set. To address this, the Random Forest Classifier model was selected as a complementary model. This model excels in managing non-linear data and enhances accuracy by averaging multiple decision trees' outputs, thereby yielding more robust and reliable results.

When the Random Forest Classifier can generate much better results than a single decision tree, however, because we are working with multiple decision trees, it comes with an increased demand for computational resources, as the simultaneous operation of multiple decision trees extends the model's runtime. To combat this, we implemented an efficient optimization technique from the Sklearn library, it is encapsulated in a single yet powerful line of code within the `RandomForestClassifier()` function: With this line of code, we can cut down the render time from 1 minute 34 seconds to 14.3 seconds, a whopping 84.79% increase in performance. The underlying mechanism facilitating this enhancement is the '`n_jobs`' function from the Joblib library, which leverages the multi-core capabilities of modern CPUs. When `n_jobs=-1`, the process of the Random Forest gets divided into 10 separate jobs (corresponding to

10 cores CPU for example), and each job is fed into a core, and they process the data independently, which will increase the efficiency with parallel processing greatly compared to scenarios where only a single core is utilized.

After the setup, we are moving on using the classification model. The first model that we are using is the Decision Tree Classifier. Before we delve into the decision tree classification, we discuss how the decision tree actually works.

Decision tree operate on the principle of partitioning the dataset into subsets using a series of binary questions based on the features. This process starts at the root node, which is the initial node, and with each question posed, the dataset bifurcates, leading to the formation of new nodes. These subsequent nodes represent the outcomes of the questions posed at each stage. As the tree grows, the dataset is increasingly segmented. This segmentation continues until the process reaches a point where either no further rules can be applied or the dataset cannot be divided further. The final nodes at which the segmentation process stops are called leaf nodes. In this framework, one branch of the tree signifies the subset of data that corresponds to a 'Yes' response to the rule proposed at the preceding node. Conversely, the other branch contains the subset of data that represents 'No'. This method of division reduces the feature space with each split, ensuring that each data point is confined to a unique region in this space. The objective of a decision tree is to refine the feature space through successive splits and the application of rules, culminating in a result where no further subdivision is possible or necessary. This approach aims to isolate data points into distinct classes, each represented by a leaf node.

Then we picked the features to be Duration, distance, start_lat, start_lng, end_lat, and end_lng columns; we are setting them to be X. For the y, which is the target, we are going to use the member_casual column, which has been converted to 0 and 1. Next, we will use

`train_test_split` from the Sklearn library to split the data into 80% train and 20% test with a random state of 50 to choose the test and train sets randomly. After completion of the model's execution, we delve into a thorough evaluation of its performance. This was achieved through a confusion matrix and a classification report, both of which provide invaluable insights into the effectiveness and precision of the Decision Tree Classifier model.

Figure 19: Decision Tree Classifier Confusion Matrix

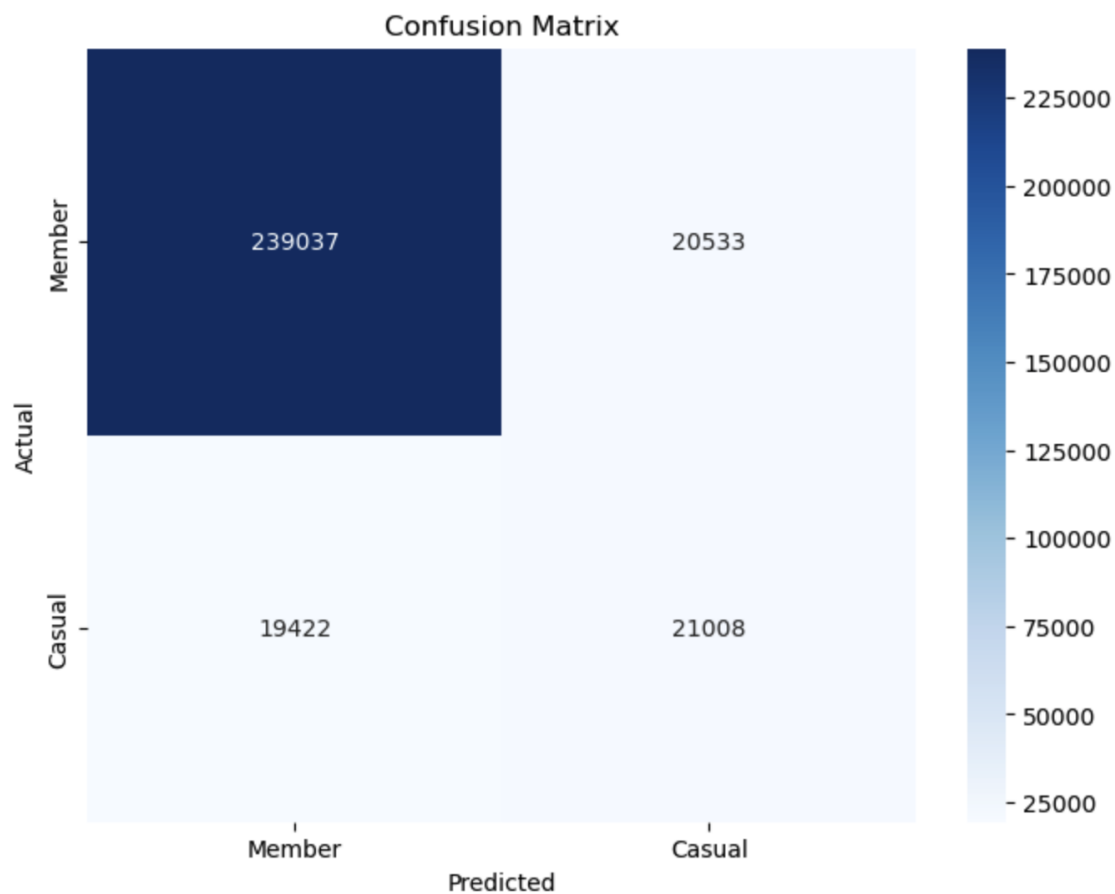


Figure 20: Decision Tree Classifier Classification Report (0 for member 1 for casual)

	precision	recall	f1-score	support
0	0.92	0.92	0.92	259570
1	0.51	0.52	0.51	40430
accuracy			0.87	300000
macro avg	0.72	0.72	0.72	300000
weighted avg	0.87	0.87	0.87	300000

From the table above, we can see that the precision metric for class '0', representing 'Member' users, stands high at 0.92. This indicates a strong accuracy in the model's ability to identify 'Member' users correctly. However, the precision for class '1', denoting 'Casual' users, is not that great, at 0.51. This means the classifier is only correct about half the time. On the side of recall, the member user is scored at 0.92, which is just as high as the precision, but for the casual user, the recall falls to 0.52, highlighting a deficiency in the model's ability to capture all actual 'Casual' users within the predicted class. The f1-score is really similar to precision and recall as it is driven from them, given a 0.92 for member users and 0.51 for casual users. Overall the model's accuracy is around 0.87, but this can be a bit misleading as the precision and recall for the two classes are not very balanced. When looking at the confusion matrix, we can also see this issue, as 239037 of the predicted members are actually true. This number is substantially higher compared to the mere 21,008 correct predictions for 'Casual' users.

Next, let's see how Random Forest can handle the data sets. We will prepare very similarly to the decision tree, using `train_test_split` from the Sklearn library with an 80% training set, 20% testing, and a random state of 50. This is the result.

Figure 21: Random Forest Confusion Matrix

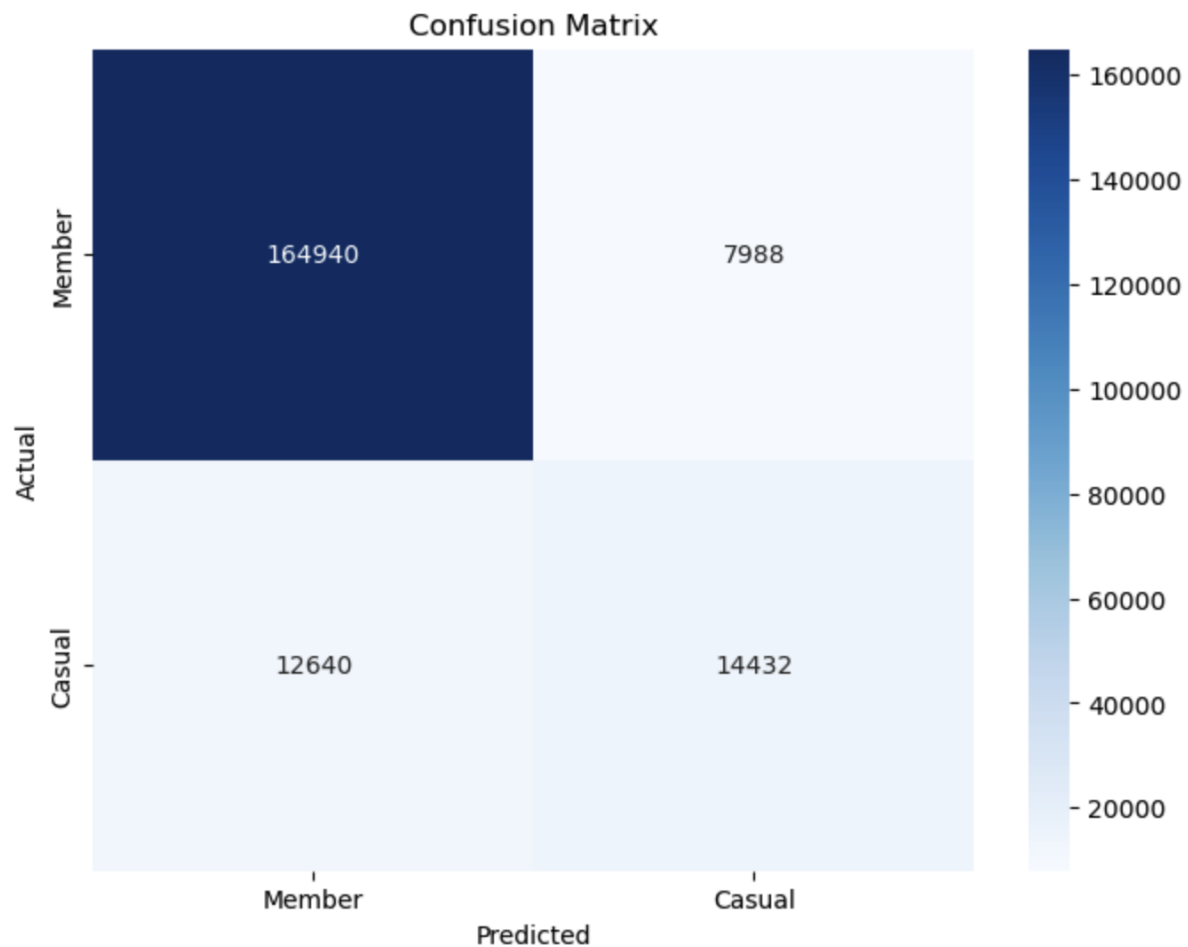


Figure 22: Random Forest Classifier Classification Report (0 for member 1 for casual)

	precision	recall	f1-score	support
0	0.93	0.95	0.94	172928
1	0.64	0.53	0.58	27072
accuracy			0.90	200000
macro avg	0.79	0.74	0.76	200000
weighted avg	0.89	0.90	0.89	200000

From the Random forest classification report above, we can see that the precision for both member and casual users has gone up; for members, it has gone from 0.92 to 0.93; there is a slight difference, but the significant difference lies in the casual user which increased from 0.51 to 0.64, this is a considerable increase. On the side of the recall, the difference is not that significant; for member users, it increased from 0.92 to 0.95, and casual users went from 0.52 to 0.53. This is relatively minor compared to the significant difference in the precision. The same thing can be seen in the confusion matrix, as most members are classified correctly, with 164940 predicted correctly, and on the casual side, 14432 has been predicted correctly.

In the end, in this part of the project, we first chose features from the data sets that are correct for the member users' and casual users' behavior. Then we decided to use classification models like the Decision Tree and the Random Forest. After applying the models we can see that Random Forest can perform better than Decision Tree by signification margin, especially on precision. But overall, the model collectively did much better at predicting member users than casual users. This phenomenon is fascinating, as many more casual users have been wrongfully classified as member users. This means those people are riding the bike, which shows a similar distance, duration, and start/end location to the members and we can target them with a marketing campaign to convert them into members. In addition, we can also investigate the characteristics of the false negatives to understand why members might be misclassified as casual users and adjust the service or communication strategy accordingly. What's more, we can analyze the characteristics of true positives to segment the member base further and create more targeted marketing strategies for each segment. In the future, we can also utilize the 'Start location' and 'End location' data to identify popular areas for casual users and members, tailoring geographical marketing and placement of services like bike stations to these trends. We believe

those interesting facts that we drive from using the classification model are just as useful as correct classification because they can usually tell the big picture on the trend of how people are using the service. Those results can be used by the company's advertising departments for a more targeted delivery, like giving casual users who are wrongfully classified as member users a discount on the membership, so they are more likely to become a member, or sending reminders to those who are member users but wrongfully classified as casual user to ride more.

VII. Conclusion

Our analysis focused on proving that there is a significant difference in the way members and casual users make use of the Capital Bikeshare platform to lay a foundation for structuring new membership strategies for revenue generation. As the hypothesis testing indicates, there is a significant difference in almost every aspect available on the dataset when members and casual users were compared. From duration, to distance, hour of the day, to chi-square test on yearly users of members and casual users, all p-values of the testing indicated that the null hypothesis of members and casual users not being significantly different was rejected.

Limitations to our study include being unable to perform further analysis because there was no column for a unique member ID which is essential for diving into memberships, but with the basis of members and casual users acting differently, we can suggest future actions and research for capital bikeshare to proceed with differentiating its memberships or implement marketing strategies for further revenue generation. Since we can classify members and casual members, capital bikeshare may be able to send personalized messages to casual users who use or almost use capital bikeshare like a member to purchase a membership. Dividing memberships into different segmentations for more casual users to convert to memberships, or increasing the fee for the current membership in the case that members are overusing the service far more than they are paying for could all be possible strategies for generating additional revenue.

VIII. Work Cited

Ink, Social. "Bike Share in the US: 2010-2016." *National Association of City Transportation Officials*, nacto.org/bike-share-statistics-2016/. Accessed 5 Dec. 2023.

Capital Bikeshare. (n.d.). About Capital Bikeshare. Retrieved December 5, 2023, from <https://capitalbikeshare.com/pricing>

Capital Bikeshare. (n.d.). Choose your plan. Retrieved December 5, 2023, from <https://capitalbikeshare.com/about>

Capital Bikeshare. (n.d.). Index of bucket "capital bikeshare-data". Retrieved December 5, 2023, from <https://s3.amazonaws.com/capitalbikeshare-data/index.html>

Decision Tree Classifier. Explained in Real-life. Retrieved December 6, 2023, from <https://towardsdatascience.com/decision-tree-classifier-explained-in-real-life-picking-a-vacation-destination-6226b2b60575>