Data Analysis Report: Ridership Differences at Divvy

Prepared For: Divvy Executive Leadership

Prepared By: Alec Ciapara

Introduction

This portfolio showcases my capstone project, a key component of the Google Data Analytics Professional Certificate, which provided comprehensive training in data analysis through hands-on assessments and practical exercises. This project demonstrates my ability to apply the complete data analysis process, from understanding business needs to delivering data-driven insights. Throughout this project, I utilized the Google analytics process: asking relevant questions, preparing and processing data, employing various analysis techniques, sharing findings through compelling visualizations, and acting upon the insights to provide actionable recommendations. The hard skills applied in this analysis include proficiency in SQL (using Google's BigQuery), and data visualization with Tableau.

Abstract

As a Junior Data Analyst at Divvy, Chicago's bike-share program operated by Lyft, the objective of this project was to identify differences in bikeshare usage patterns between casual riders and annual members to inform strategies for converting casual riders into members. Utilizing publicly available Divvy trip data in Google BigQuery with SQL, I prepared, cleaned, explored, and analyzed ride data. The analysis indicates that annual members tend to take more frequent but shorter trips consistently throughout the year, suggesting a higher likelihood of using the service for commuting. In contrast, casual riders exhibit longer but less frequent rides, primarily concentrated during the spring and summer months, indicating a preference for leisure or exercise. Based on these findings, potential marketing strategies could include targeted campaigns during the casual riders' peak season with incentives for annual memberships or the development of new membership plans tailored to their usage patterns.

Scenario

Divvy is Chicago's bike share program operated by Lyft. The data is real and available here through a license. **We assume the position of a Jr. Data Analyst for Divvy, which offers both electric and classic bikes through a bike share app.** Divvy has a fleet of over 6000 bikes which can be retrieved or returned to docking stations or otherwise left in designated areas. Divvy services Members, who buy annual membership plans for reduced fares, and Casuals, who do not have a membership and use the service a-la carte.

Ask

Business Problem: The Director of Marketing at Divvy tasked the marketing analytics team with identifying strategies to maximize the number of annual memberships. To contribute to this goal, I was assigned the initial question: '**How do annual members and casual riders use Divvy bikes differently?'** The insights gained from this analysis would provide a foundation for understanding why a casual user might consider purchasing an annual membership and inform the development of targeted marketing campaigns.

Prepare

Data Source and Preparation: To address the business question, I utilized Divvy's historical bike trip data for the period of February 2022 through January 2023. This data, stored in monthly CSV files on Amazon AWS servers, is publicly available under a free license and is considered a credible source for understanding user behavior. While the data lacks personally identifiable information, allowing for the analysis of overall trends between member and casual rider groups, it prevents deeper analysis at the individual user level. Initial exploration of the dataset confirmed its structure and identified key columns relevant to the analysis.

Process

Data Processing and Cleaning: Given the dataset's size (over 5.7 million rows), SQL and Google BigQuery were selected as the primary tools for data processing, exploration, and cleaning. The process involved uploading twelve monthly CSV files to a Google Cloud bucket, then merging them into 5,754,248 rows and 13 columns. Initial exploration using SQL queries (code available on **GitHub**) provided insights into the data structure and identified areas for cleaning. Key cleaning steps included: standardizing bike type categories ('docked bike' to 'classic bike'), removing potential error rides based on duration, ensuring consistency in station names, addressing missing end station data for classic bikes using location coordinates, and creating analytical columns for ride length and day of the week. These steps were essential to ensure data accuracy and prepare it for meaningful analysis. The complete SQL code for the data processing and cleaning steps can be reviewed on **GitHub**.

Field name	Type
ride_id	STRING
rideable_type	STRING
started_at	TIMESTAMP
ended_at	TIMESTAMP
start_station_name	STRING
start_station_id	STRING
end_station_name	STRING
end_station_id	STRING
start_lat	FLOAT
start_lng	FLOAT
end_lat	FLOAT
end_lng	FLOAT
member_casual	STRING

Schema for Divvy data.

After wrangling the data, I began exploring it. **A copy of the SQL code I used to wrangle and explore the data can be found on GitHub.**

```
C5.start_station_name & C6.start_station_id & C7.end_station_name & C8.end_station_name
-Check for leading/trailing/double spaces.
-Verify naming consistency.
*/

SELECT DISTINCT(start_station_name)
FROM ride_data
GROUP BY start_station_name
ORDER BY start_station_name;

SELECT DISTINCT(end_station_name)
FROM ride_data
GROUP BY end_station_name
ORDER BY end_station_name;

SELECT start_station_name
FROM ride_data
WHERE start_station_name LIKE '%   %';

SELECT COUNT(start_station_id)
FROM ride_data;

SELECT start_station_id
FROM ride_data
WHERE start_station_id LIKE '%  %';

SELECT *
FROM ride_data
WHERE start_station_name IS NULL OR end_station_name IS NULL AND
  rideable_type = 'classic_bike';
```

Sample of SQL code used.

- Ride ID- This column contains a specific identifier connecting it to one specific instance of bike use, from start to finish. **I verified each Ride ID as distinct and unique, and confirmed each was a string of only 16 characters.**
- Rideable Type- This column names the type of bike used in the ride, classic or electric. I confirmed there were no empty rows, but docked bike is present. **Docked bike is an outdated term for classic bikes, I corrected 'docked bike' entries to 'classic bike' to ensure consistency in bike type categorization.**

- Started At & Ended At- These columns represent timestamps of a bike ride being initiated and then completed. As they are two timestamps, I verified them together. The time is created as Chicago local time, but appears in the data as UTC, this will be corrected. **Using a TimeStamp_Diff function we can subtract the starting time from the ending time to verify the ride duration. There are 5,390 rides over a day, and 229,452 rides less than a minute or null. Removed rides with durations less than one minute or greater than 24 hours, as these were likely indicative of system or user errors or data anomalies.**
- Start Station Name/End Station Name/Start Station ID/End Station ID- As these are all strings meant to identify where a ride started and where it ended, I will verify them together. I standardized station name formatting by trimming unnecessary spaces and correcting capitalization for consistency. **More problematically, over 800,000 Classic bike rides, which must end in a dock, have no ending station logged, and so we will correct this in cleaning using latitude and longitude data available to us later in the schema. Start and End station IDs have no value to us and are part of Divvy's interior statistics, as such we will remove these columns during the cleaning process.**
- Start Lat/Start Lng/End Lat/End Lng- These columns all contain floats which give us coordinate data on where a ride started and ended. **I verified 5,899 rows contain null values for at least one of these values, these will have to be purged from our final data so that we can map the coordinates. An end station at Green St. & Madison shows 8 rows of null coordinate data, but as we have the data in other instances of this station, we can correct this. Finally, we will use this coordinate data to correct the missing station names as stated in the bullet point above.**
- Member/Casual- This column simply lists the membership status of the account associated with the bike ride. **I confirmed there were only 2 possible values present and no null values. No further cleaning necessary.**

Having completed my exploration, I used SQL to clean the data. **A copy of the SQL code used to clean the data can be found here. Cleaning the data was performed by creating 2 temporary tables, then joining them in a third table to be queried for the analysis.**

```
47    cleaned_data AS (
48      SELECT *
49      FROM
50        (SELECT ride_id,
51          CASE
52            WHEN rideable_type = 'docked_bike' THEN 'classic_bike'
53            ELSE rideable_type
54            END AS bike_type,
55            started_at, ended_at, TRIM(INITCAP(start_station_name)) AS start_station_name_c,
56            TRIM(INITCAP(end_station_name)) AS end_station_name_c, start_lat, start_lng,
57          CASE
58            WHEN end_lat = 0.0 THEN 41.881827376571351
59            ELSE end_lat
60            END AS end_lat_c,
61          CASE
62            WHEN end_lng = 0.0 THEN -87.648831903934479
63            ELSE end_lng
64            END AS end_lng_c,
65            member_casual
66        FROM total_data
67        )
68      WHERE start_lat IS NOT NULL AND
69        start_lng IS NOT NULL AND
70        end_lat_c IS NOT NULL AND
71        end_lng_c IS NOT NULL
72    ),
```

Query cleaning existing data.

1. The first temporary table (line 47) corrected the ending latitude and longitude for Green Street and Madison AVE., replace instances of docked_bike with classic_bike, and removes unwanted columns start_station_id and end_station_id. I also cleaned text strings for station names by trimming unnecessary spaces and correcting capitalizations and removed rides missing any coordinate information.

2. The second temporary table creates columns for analysis (line 78 on GitHub). I extract the day of the week, day of month, month, year, and ride duration from data present in the data set, making sure to use Chicago as time zone to correct the instances of UTC. I also dropped rides with less than a minute and more than 24-hour ride duration.

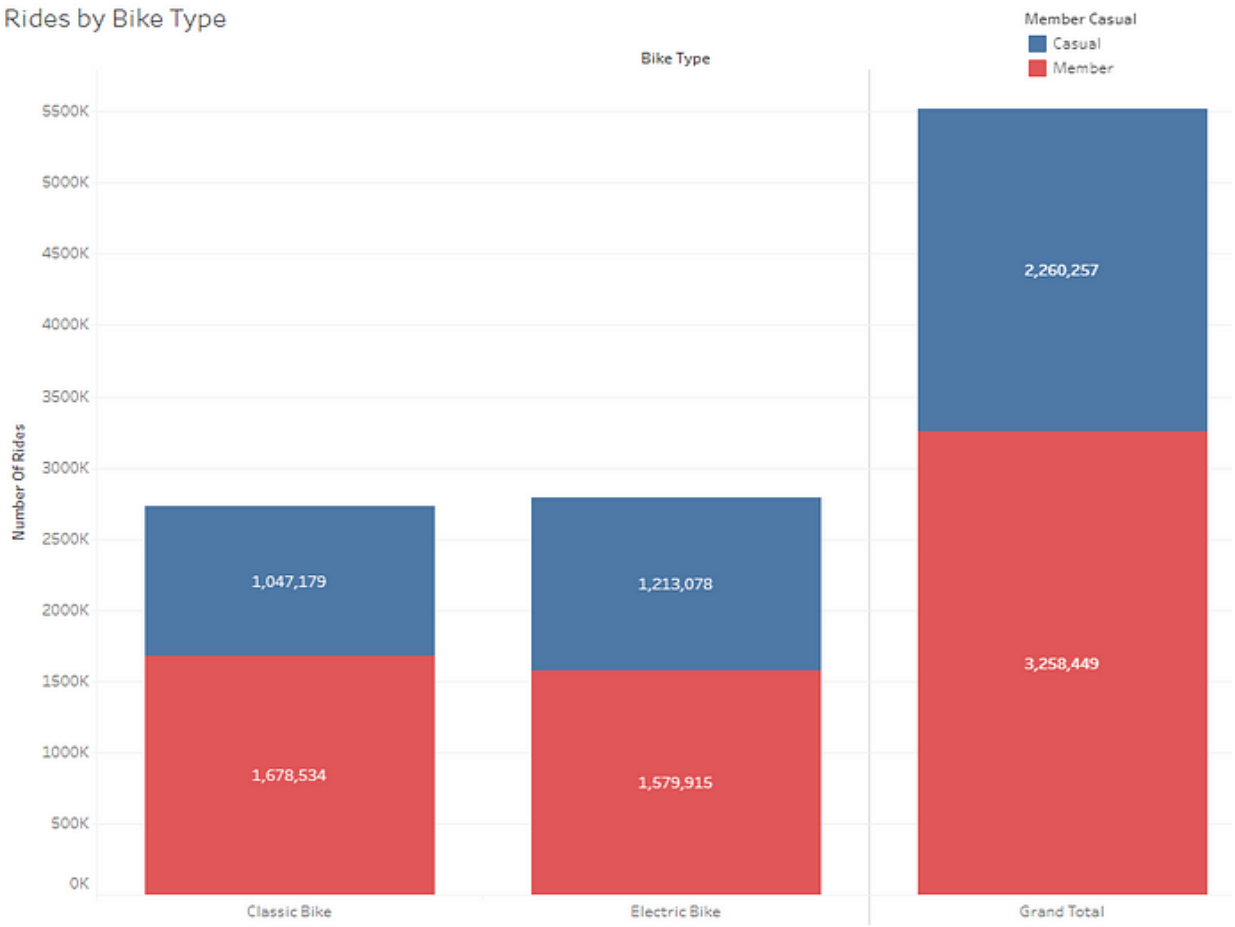
3. The third table (line 118) uses an INNER JOIN function to create a singular table from the other two. This table, ride_data, will be used for our analysis.

Analyze

Analysis: To understand the differences in how annual members and casual riders utilize Divvy bikes, I conducted a thorough analysis of the ride_data using SQL in Google BigQuery. **The analysis queries can be found here, starting on line 126.** . The primary focus was to identify patterns in ride frequency and duration across various temporal dimensions, including hour of the day, day of the week, and month of the year, as well as to examine popular start and end locations for both user groups. The SQL queries used for this analysis can be reviewed on **GitHub**. The key findings from this analysis are presented visually below.

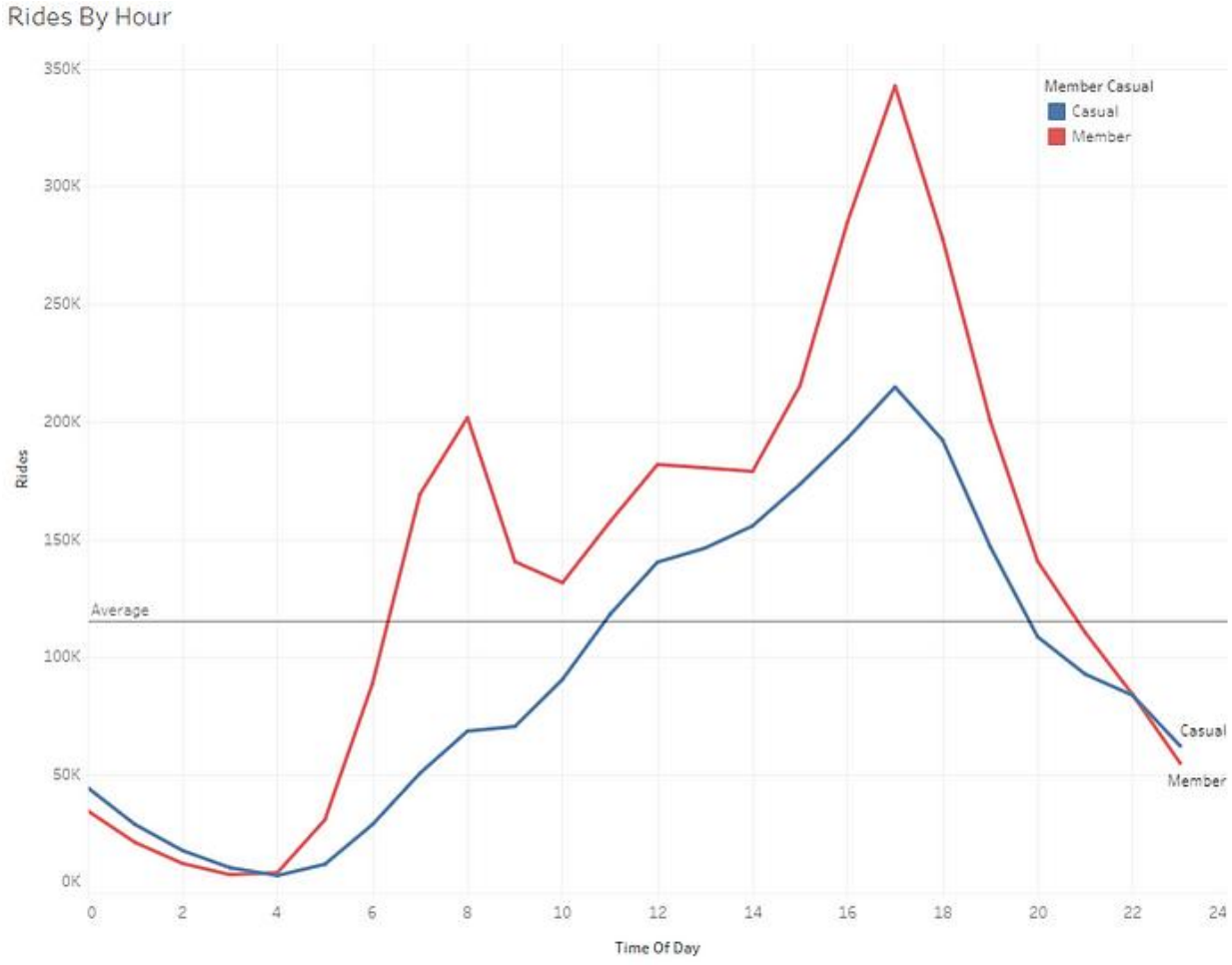
Visualize

The visual representation of the analysis can be found here, on Tableau Public. The scope of this project : "H**ow do annual members and casual riders use Divvy bikes differently?**"



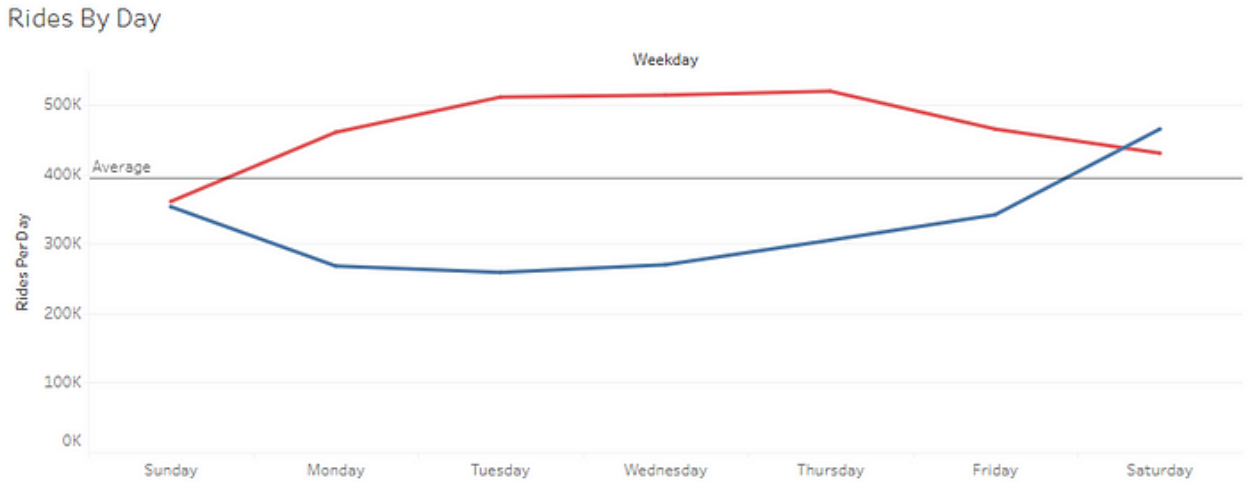
Total Rides by Bike Type

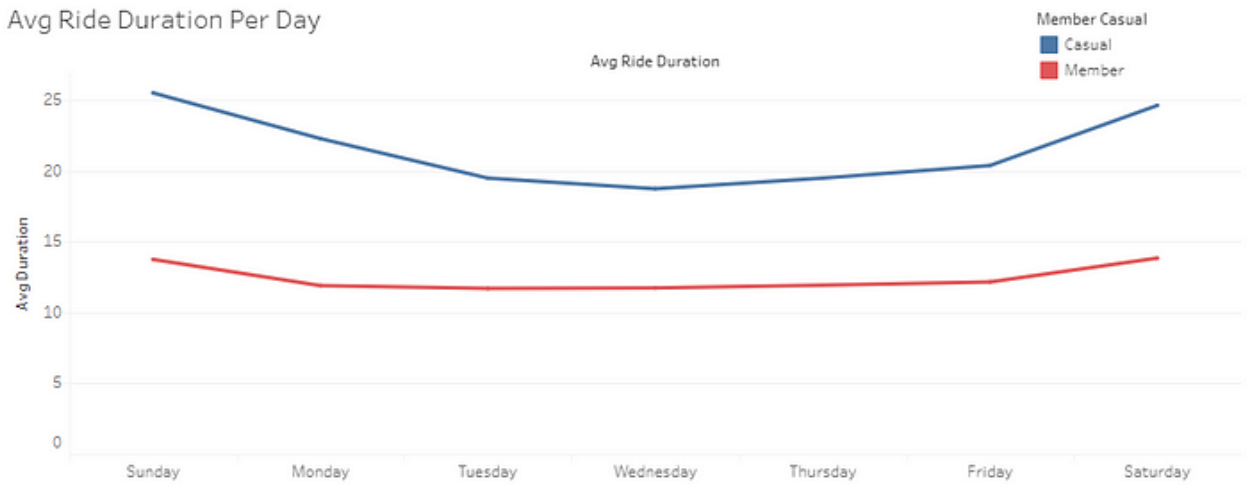
As we can see, during the year February 2022 through January 2023, Members made up 59% of all rides. Neither group has a large preference for either the Classic (manual) bike, or the Electric bike. Members choose Classic bikes for 51.51% of their rides, while Casual riders opt for Electric bikes 53.67% of their rides. **However, while the number of rides by members is greater than that of casual users, lacking any personally identifiable information (PII), we simply cannot conclude that anyone using the rideshare would be more likely to be a member.** It may simply be the case each member takes advantage of the rideshare in greater numbers throughout the year.



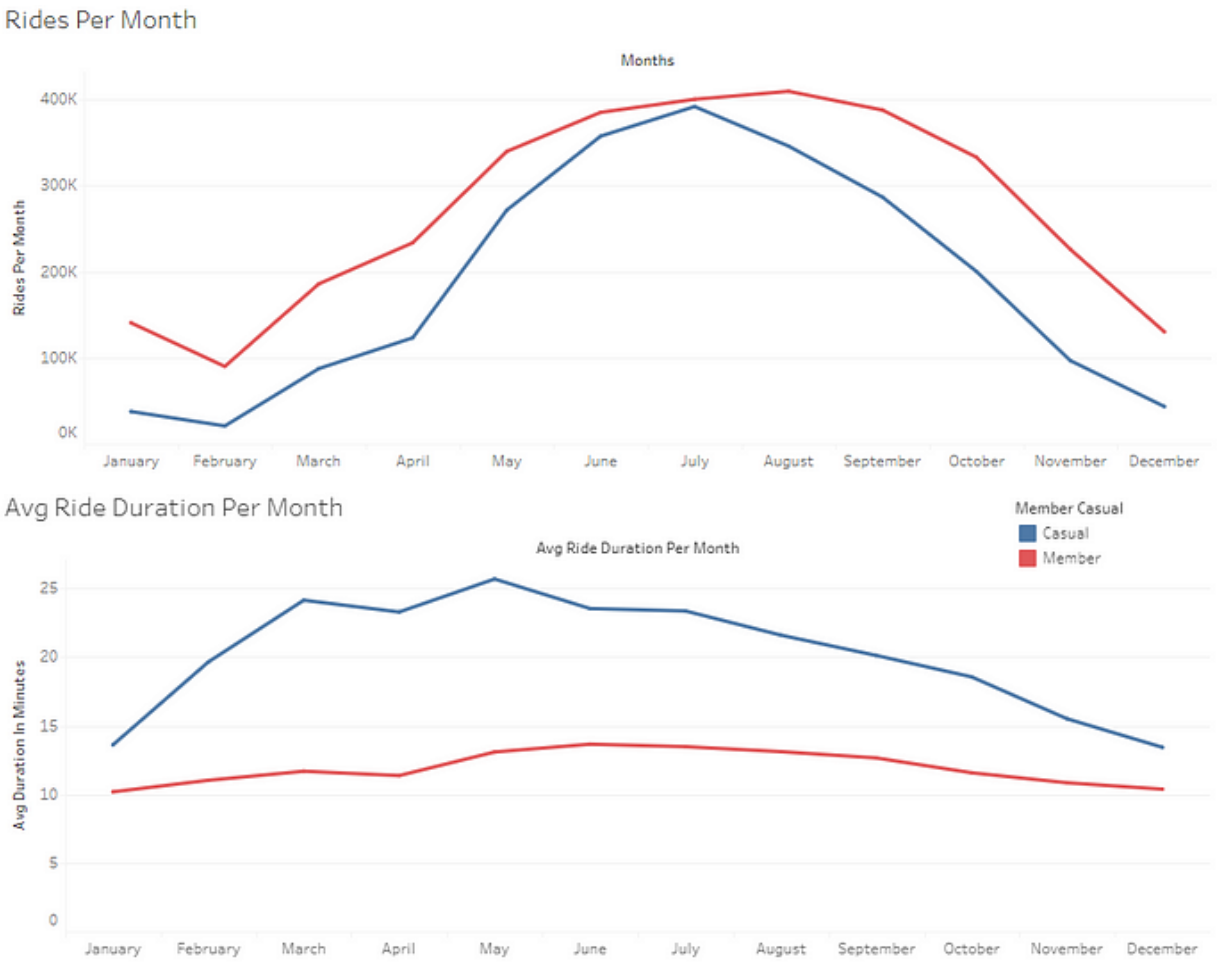
Next, we look at usage patterns through out the day. We can see, from the chart above, the number of rides by members is higher throughout most of the day than that of casuals. **People who are members are likely to take advantage of their membership frequently. Another pattern stands out here, usage by members soars above the overall median during commutes to and from work, 6–9 A.M. and 3–8 P.M., and during conventional lunch hours, 11 A.M. to 1 P.M..** Casual riders only see a ride from about noon to 7 P.M.. **A story we can see developing is members may be high frequency users**

who use the bikeshare to commute, while casual users use the bikeshare less frequently and possibly as an alternate form of commute.



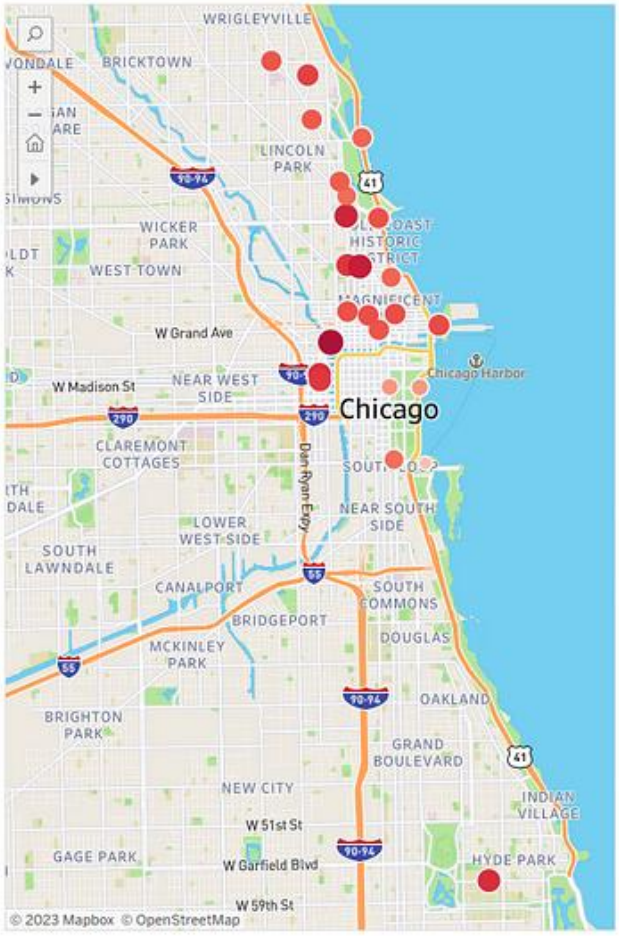


Next, we look at usage patterns by weekday. As we see above, members have a higher than average usage rate throughout the week, but particularly during weekdays. Casual riders are only above the average usage rate on Saturdays, with Sunday being their second highest day. When we look at average duration of ride per day, however, we see casual users, while taking overall lower total number of rides per day, certainly take longer rides, no matter the day. **Members certainly seem to use the service more during weekdays, while casual usage peaks during weekends, but casual rides are possibly being used for leisure, exercise, or extended paths.**

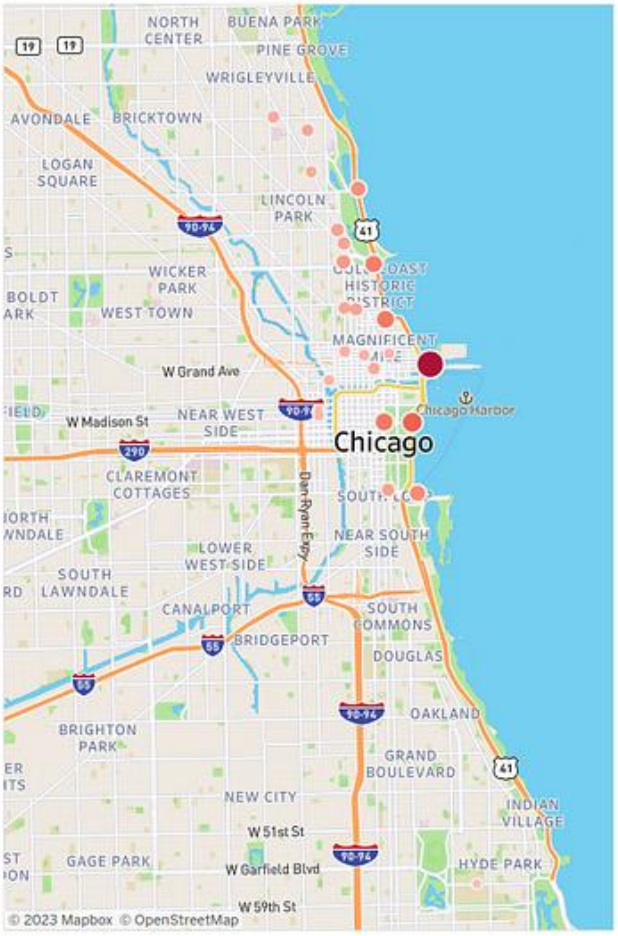


The chart above shows usage patterns per month. Total number of rides by both groups sees a drop during winter months, this is to be expected of Chicago's notoriously cold winters. However, members see a fuller bell curve throughout the remainder of the year, while casual riders have a much steeper rise in usage during late spring and steeper drop mid-autumn. Average ride duration per month is also interesting, as we see a rise in ride duration for casual riders particularly during spring and summer rides, but member ride durations largely remain flat throughout the year. **This further leads to the idea members use the bikeshare as a primary source of transportation, while casual users may well be using it for leisure or exercise.**

Popular Member Starts

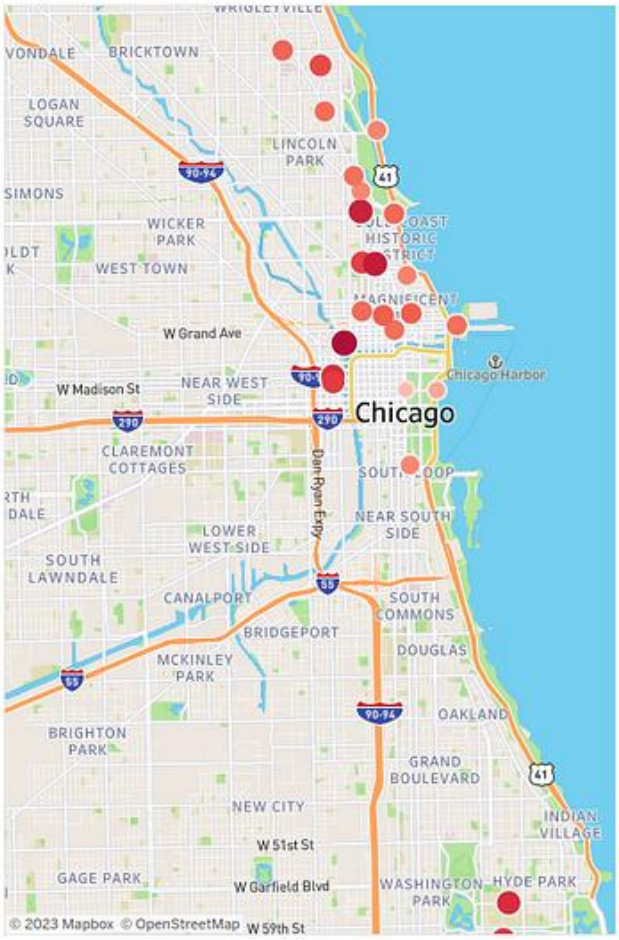


Popular Casual Starts

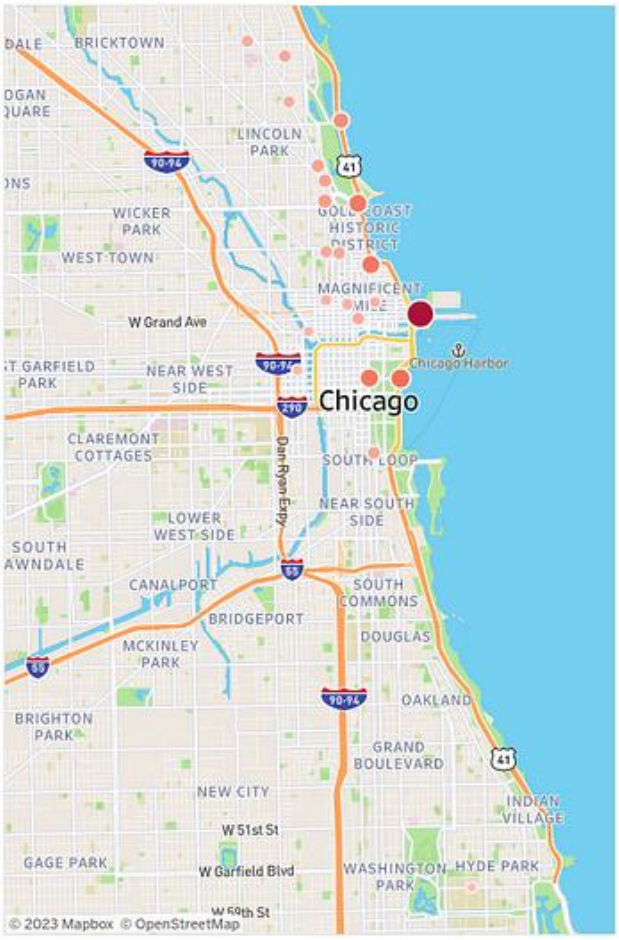


To test for casual users simply taking longer routes, I mapped the starting docked locations for members on the left, and casual users on the right. The top 25 starts are shown above. Members have a higher frequency of starting and ending rides at docks, with many casual members simply taking opportunity of the option to leave a bike locked to a street light or other city property. Below, I also mapped out the popular end docks. **Casual riders, when they do use a dock, have less bias about where they begin or end rides, with the top 50 docking stations being similar in number of uses to the top 25. Members largely use docks in the northeast, University of Chicago, and University of Illinois campuses.**

Popular Member Ends



Popular Casual Ends



Act

“How do annual members and casual riders use Divvy bikes differently?”

Conclusion and Recommendations:

Based on the analysis of Divvy's bike trip data, we can conclude that:

- **Members: Exhibit usage patterns consistent with daily commuting or frequent, short trips, with higher usage during weekdays and commute hours throughout the year. Possibly University students and local office workers.**
- **Casual Riders: Tend to take longer rides, primarily on weekends and during the spring and summer months, suggesting a greater inclination towards leisure or recreational use. Possibly tourists or leisure locals.**

Recommendations:

1. **Targeted Seasonal Campaigns: Initiate marketing campaigns aimed at casual riders in late winter to early spring, highlighting the benefits of annual memberships as they begin to increase their usage for the spring and summer seasons. Consider offering discounted membership rates for new members during this period to incentivize conversion.** *This recommendation is supported by the finding that casual rider usage peaks significantly during the spring and summer months.*
2. **Weekend or Extended Ride Membership Options: Explore the feasibility of offering a reduced-price membership option for weekend use or a membership tier that provides better rates for longer rides. This could appeal to casual riders who primarily use the service for leisure on weekends or for longer durations.** *This aligns with the observation that casual riders have higher usage on weekends and take longer average rides.*
3. **In-App Engagement for Leisure Riders: Develop in-app features such as suggested scenic routes or points of interest to further cater to casual riders' apparent preference for leisure and exploration. This could increase their engagement with the Divvy app and potentially lead to considering a membership for more frequent access to these features.**

Conclusion

Limitations:

"While this analysis provides valuable insights into the different usage patterns of annual members and casual riders, it's important to acknowledge certain limitations:

- **Lack of Personally Identifiable Information (PII):** The absence of PII in the publicly available dataset meant that the analysis was conducted at an aggregate level for each rider type. We were unable to delve into individual user behavior or demographics beyond their membership status.
- **Single Year of Data:** This analysis is based on one year of operational data (February 2022 - January 2023). Usage patterns may be subject to seasonal variations and could evolve over longer periods. Analyzing data from multiple years would provide a more comprehensive understanding of long-term trends.

- **Assumption of Consistent User Behavior:** The analysis assumes that the behavior of casual riders and members remained relatively consistent throughout the analyzed year. External factors or significant events could have influenced these patterns.
- **Focus on Usage Patterns:** This project primarily focused on *how* members and casual riders use the bikes. Further research would be needed to understand the *reasons* behind these usage patterns and the motivations of casual riders to potentially become members.

Future Work:

Building upon this analysis, several avenues for future work could be explored:

- **Longitudinal Analysis:** Analyzing historical data spanning multiple years would allow for the identification of trends and seasonality patterns with greater confidence and could reveal how user behavior evolves over time.
- **Integration of External Data:** Incorporating external data sources, such as weather information or event calendars, could provide further context for understanding fluctuations in bike-share usage.
- **User Surveys and Qualitative Research:** Conducting surveys or focus groups with casual riders could provide valuable qualitative insights into their motivations, needs, and potential barriers to becoming annual members.
- **Predictive Modeling:** Developing a predictive model to identify casual riders who exhibit characteristics similar to current annual members could enable more targeted and effective marketing campaigns.
- **A/B Testing of Marketing Strategies:** Implementing and tracking the results of different marketing strategies aimed at converting casual riders (such as those suggested in the recommendations) through A/B testing would allow for data-driven optimization of these efforts.

Lessons Learned:

"Through the completion of this capstone project, I gained valuable experience and learned several key lessons:

- **The Importance of Data Exploration:** Thoroughly exploring the dataset, even with a clear business question in mind, revealed nuances and potential data quality issues that needed to be addressed during the cleaning process.

- **Power of SQL for Large Datasets:** Utilizing SQL in Google BigQuery proved to be an efficient and powerful method for wrangling and analyzing a dataset of this size, highlighting the importance of database management skills in data analytics.
- **Connecting Analysis to Business Objectives:** Continuously linking the analytical findings back to the initial business question (maximizing annual memberships) ensured that the insights and recommendations were relevant and actionable.
- **Balancing Breadth and Depth of Analysis:** While the analysis covered various aspects of rider behavior, there is always an opportunity to delve deeper into specific areas. Understanding the trade-offs between breadth and depth is crucial in real-world data analysis scenarios.
- **Effective Communication Through Visualization:** The process of creating visualizations in Tableau Public reinforced the importance of clear and concise visual communication in conveying complex data insights to a non-technical audience.