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

*Submitted by **Bianca Aspin** in partial fulfillment of the requirements for the*
Google Data Analytics Professional Certificate.

Background

In this scenario, I am a junior data analyst working for the marketing analysis team at Cyclistic, a fictional bike-share company based in Chicago.

Note:

Cyclistic is based on the real-life company of Divvy, a program run by the Chicago Department of Transportation in partnership with ride-share company Lyft. I'll be using public data from Divvy for this analysis.

Cyclistic has seen great growth since its launch in 2016, having expanded its fleet to **5,824 geotracked bicycles** over a network of **692 stations** throughout Chicago and neighboring Evanston. The company is well-known and appreciated by users for its flexibility in bike offerings (classic and electric bikes) and pricing plans. Currently, casual riders have the option of purchasing a single ride pass for $3.30/30-minute trip (with an additional charge of $0.15/minute over that 30-minute limit) or a day pass for $15/day. Regular riders can purchase an annual membership for $9/month which includes unlimited 45-minute rides (with an additional charge of $0.15/minute over that 45-minute limit).

Note:

The case study documentation mentions that Cyclistic also offers disability-friendly bike options that are utilized 8% of the time. However, the data used for this project does not include any information on accessible options, so I won't cover this option during this analysis.

Cyclistic's finance team has recently concluded that annual members are more profitable than casual riders. This insight has inspired the Director of Marketing (my manager, in this scenario) to launch a campaign to convert casual riders into members. As such, she's tasked my team with analyzing our historical bike trip data to identify trends and present our recommendations to the executive team.

I've been tasked to answer the question, **"How do annual members and casual riders use Cyclistic bikes differently?"** My colleagues will be digging into more specifics about why casual riders would buy an annual membership and how the company could use digital media to influence casual riders to become members.

Business Task

Cyclistic has identified an opportunity to increase profits and grow as a company by maximizing its membership numbers. To realize this opportunity, this project will analyze data about current Cyclistic users to explore how casual riders and annual members differ in their use of the company's bike-sharing services. This research will inform future marketing campaigns focusing on converting Cyclistic's casual riders into members.

Research Questions

After considering the overarching question of how members and casual riders differ in their use of our bikes, as well as performing an initial review of the data my manager tasked me to use, I settled on five areas to investigate further:

1. **Stations:** Where are casual riders most commonly starting and ending their trips?
2. **Time of Year:** What does ridership look like for casual riders during a given month?
3. **Day of Week:** What does ridership look like for casual riders during a given day of the week?
4. **Time of Day:** What does ridership look like for casual riders during a given hour of the day?
5. **Ride Length:** How long do casual riders' trips typically last?

Note:

Initially, I also wanted to investigate an additional area - the distance traveled by each user type - and performed various transformations on the data for this purpose. However, I later abandoned this line of inquiry because there wasn't enough information to create a meaningful and reliable metric. Numerous trips started and ended from the same location, and even when the start and end differed, there was no way of knowing which route a rider might have taken to get there.

Methodology

Data Description

Source and Licensing

This project utilizes publicly available, anonymized trip data provided by Divvy, a bikeshare program operated by Lyft with oversight from the Chicago Department of Transportation. The data provided by Divvy is original data, coming directly from its service rather than a second or third party, which makes it a reliable and credible source. It can be used for this project thanks to the license agreement Divvy has established in conjunction with the City of Chicago. This agreement allows users to download and use the data as source material for analyses, reports, and visualizations. However, among other stipulations, the agreement prohibits the user from connecting this data with the personally identifiable information of the bikeshare customers. This measure has been put in place to preserve data privacy.

Privacy

This data was anonymized before its upload to Divvy's public repository, so no additional measures need to be taken to secure it for this project. However, using this anonymized data means we will have no way of determining whether a new or returning customer initiated each trip. We also won't be able to gauge how many trips each casual rider might make or glean the current rate at which casual riders upgrade to annual memberships.

Data Storage

Data from the entire 2021 calendar year (the most current data available) will be downloaded from Divvy's website and stored locally for processing within Microsoft Excel and PostgreSQL for the duration of this project. In addition, summary data will be stored in the cloud for processing within Tableau Public.

Data Structure

The data for each month comes packaged from Divvy as a zipped file folder; each folder contains a CSV file with the corresponding month's data. Each dataset has 13 attributes, which are detailed in the data dictionary in the appendix.

Data Integrity and Issues Found

I performed an initial review of each of the 12 datasets by creating pivot tables to perform counts of all records to check for missing values, as well as checking the filter view for each column to check the consistency of the categories associated with each attribute. I found a few issues, which I will detail below. None of these issues were major enough to abandon this data source. However, they led me to remove **4,771** records and revise another **311,705** records to ensure the accuracy and consistency of the data being analyzed.

Missing Values

An initial overview revealed that each dataset was reliably missing data in the start station name/ID attributes, end station name/ID attributes, and end latitude/longitude attributes. A breakdown of the missing values in each month's dataset can be found in the appendix.

Missing Start/End Station Data

One of the factors that likely contributed to the missing start and end station data was that electric bikes do not need to be docked at a station. For an extra fee, Divvy's ebike users can end their ride by locking the bike to a public bike rack, light pole, signpost, or retired parking meter (as long as it's still within the service area). Relatedly, as ebike usage increased over the year, the count of missing station data increased, too. Rider error and technical issues could also have contributed to this information being missing, particularly for those classic bikes that are missing docking data. For instance, it is possible that bikes were returned to an offline dock and were not correctly tracked.

I determined that it would be permissible to leave these missing station values as is, as long as latitude and longitude data were present, as it would still be helpful for compiling summary statistics of station usage.

Missing End Latitude/Longitude Data

It is unclear what might have contributed to the missing end latitude/longitude data other than rider error, a technical error, or perhaps the utilization of the valet service at certain popular stations during peak times. Valet users can hand their bike off to a Divvy staff member who will park it for them as dock space becomes available.

Given that a) the cause of this missing data was unclear and threw the accuracy of the corresponding end times into question as well, and b) latitude and longitude would be critical for making calculations on distance (for the line of inquiry I later abandoned), I determined it would be best to remove the records missing end latitude and longitude values.

Unclear Categories

While only two types of bikes (or 'rideables' as they are referred to in the datasets) are used by Divvy, three categories consistently appeared in the rideable_type column: classic_bike, electric_bike, and docked_bike. Upon further research and review of historical data, I determined that 'docked_bike' was most likely a legacy classifier for 'classic_bike', as all of Divvy's bikes were categorized as 'docked_bike' in the historical data before they introduced electric bikes in 2020. To ensure consistency, I replaced all instances of 'docked_bike' with 'classic_bike' in the datasets. A count of the records replaced in each month's dataset can be found in the appendix.

Time Entry Glitches

End Time Earlier Than Start Time

A small number of entries each month experienced a glitch causing the end time (ended_at) to be earlier than the start time (started_at). 11/7/21 was a particularly problematic date due to the daylight savings switch, which created multiple errors. I removed these records. A count of the records removed from each month's dataset can be found in the appendix.

End Time Greater Than 24 hours After Start Time

After removing the above entries, I discovered another issue related to a small number of time entries, wherein the 'ended_at' time was greater than one day after the 'started_at' time. This error may or may not be the product of another glitch with time encoding; it could be that these users returned their bikes late. However, given that this does not represent the typical use case for these bikes (the longest a user can officially rent a bike is 24 hours for a full-day pass), I determined that it would be best to remove these 811 records to ensure accuracy of the analysis.

Data Preparation

All steps outlined below were performed on the CSV datasets for each month before uploading them to the master table (bikes_2021) for analysis with SQL. A summary of the columns added during the data preparation process can be found in the data dictionary in the appendix.

Calculate Ride Length

I created a new column, 'ride_length', to calculate the length of each trip by subtracting the ended_at column from the started_at column for each record. I then updated the formatting of this column to HH:MM:SS using Excel's number formatting options for time-related data.

Calculate Day of Week

I created a new column, 'day_of_week', to calculate which day of the week each trip started utilizing Excel's WEEKDAY function. The result was initially formatted as a general number representing each weekday (1 = Sunday, 7 = Saturday). However, I later replaced these numbers with the actual name for each day to improve the readability of outputs generated during the analysis.

Round Geographical Coordinates

The four columns containing geographical coordinates ('start_lat', 'start_lng', 'end_lat', 'end_lng') contained decimals of varying length from month to month. To ensure consistency, I rounded these columns to 6 decimal places in all datasets, as I determined this would be a sufficient level of precision for this analysis.

Calculate Distance Traveled

I created a new column, 'distance_traveled', that would calculate the distance in miles between the starting and ending coordinates.

It turns out that there is no built-in function in Excel for calculating the distance between coordinates. However, I managed to find a reliable formula via Stack Overflow, which used trigonometry to convert the four coordinates into the distance in kilometers. I then wrapped this in a function to convert that distance into miles.

Note:

This effort was undertaken as part of the distance-related line of inquiry I later abandoned, but I used this column to establish another column, 'trip_type', which I frequently used to filter 'canceled' trips from later queries.

Import to PostgreSQL database

After performing the above steps, I imported all 12 individual datasets into the bikes_2021 master table for further analysis with SQL. This process involved ensuring that the columns in bikes_2021 were configured correctly for each data type and matched the columns in the CSV files being ingested.

Calculate Trip Type

After importing the datasets into the bikes_2021 table, I created a new column, 'trip_type', to categorize trips into three buckets.

Note:

Again, this effort was undertaken as part of the distance-related line of inquiry I later abandoned, but I used this column to filter out 'canceled' trips from later queries.

Trip Type Categories

- **Canceled:** Trips where there was no difference between the starting and ending coordinates and the ride length was less than 1 minute.
- **Out-and-Back:** Trips that started and ended at the same station/coordinates.
- **Point-to-Point:** Trips that started and ended at different stations/coordinates.

Analysis and Findings

All steps outlined in the following sections were performed in PostgreSQL 14. The output for each query was exported as a CSV file for additional exploration within Excel and visualization within Tableau.

Note:

I opted for PostgreSQL over Google BigQuery (the platform we utilized throughout the Certificate program). The immense size of the files to be analyzed made using a local platform preferable to a cloud-based platform.

Stations: Where are casual riders most commonly starting and ending their trips?

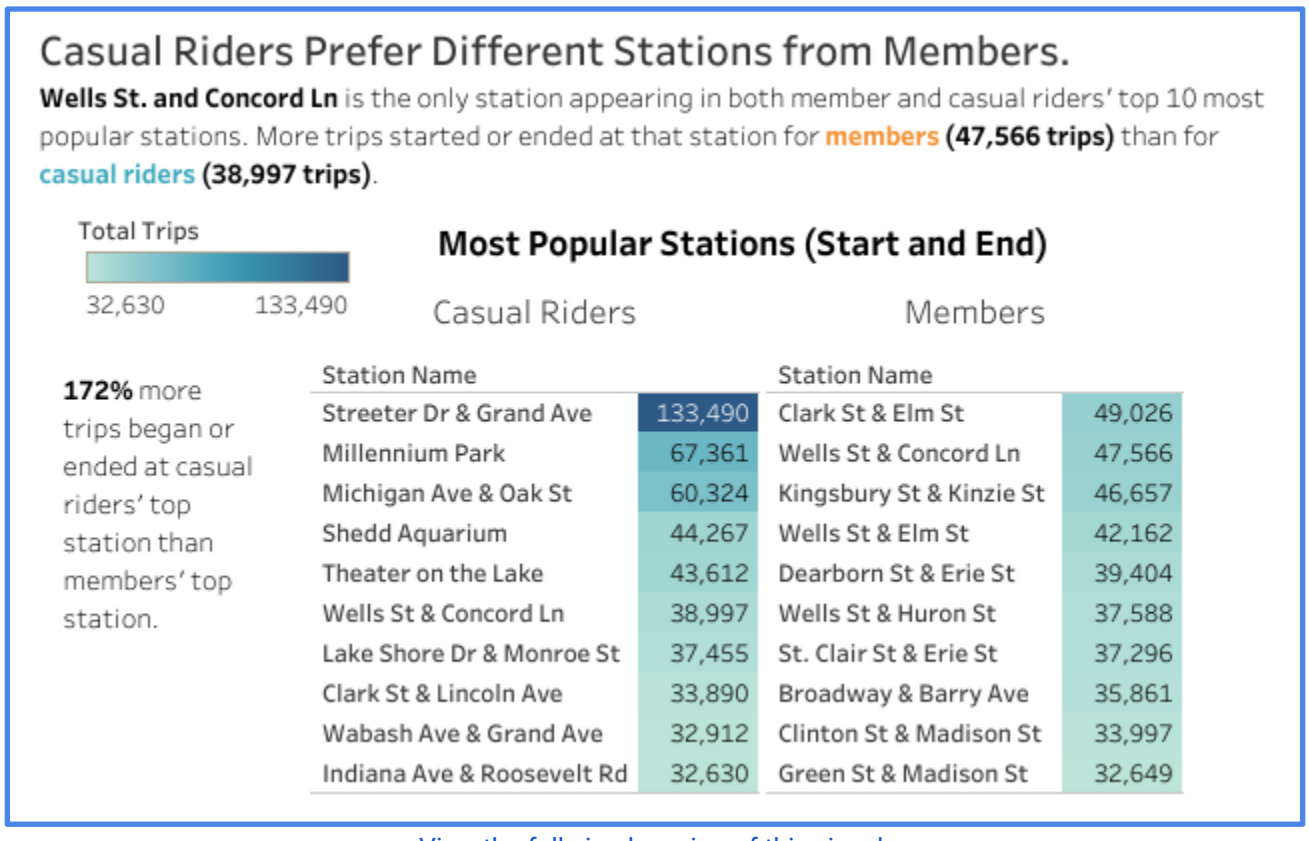
Analysis

To determine which stations were most popular for each rider type, I wrote queries that selected the top 10 most popular start stations and top 10 most popular end stations for each group. Each query counted the number of unique ride ids present, grouped by station name and filtering out all false starts and canceled trips. The output was sorted in descending order, and only the first ten records for each query were retrieved.

The results showed that there was not much difference between the most popular start and end stations within each rider type. Eight of the top ten start stations for casual riders also appeared in their top ten end stations. Likewise, nine of members' top ten start stations also appeared in their top ten end stations.

This consistency led me to combine start and end trip totals to determine the overall top stations for each rider type rather than examine their preferred start and end stations separately. I queried SQL again to get a master list of the ride totals for all start stations; I rewrote my previous queries to add rider type as its own column and then grouped the results first by station name and then by rider type. Next, I ran a similar query to get a master list of end stations. I combined and cleaned the output for each of these lists in Excel to get a master list of the total trips (start AND end) for each rider type at each station suitable to create a side-by-side comparison in Tableau.

Findings



View the full-sized version of this visual.

Casual riders tended to prefer different stations from members. Reviewing the top ten most popular stations for each group, we can see little overlap between casual riders' and members' preferences. **Wells St and Concord Ln** is the only station that appears on both lists, ranked 6th most popular for casual riders (38,997 trips) and 2nd most popular for members (47,566 trips).

There is a marked difference between casual riders' and members' top stations. **172%** more trips began or ended at casual riders' top station (Streeter Dr and Grand Ave - 133,490 trips) than members' top station (Clark St and Elm St - 49,026 trips). This difference could be because the Streeter Dr and Grand Ave station is nearby several tourist attractions, including Navy Pier.

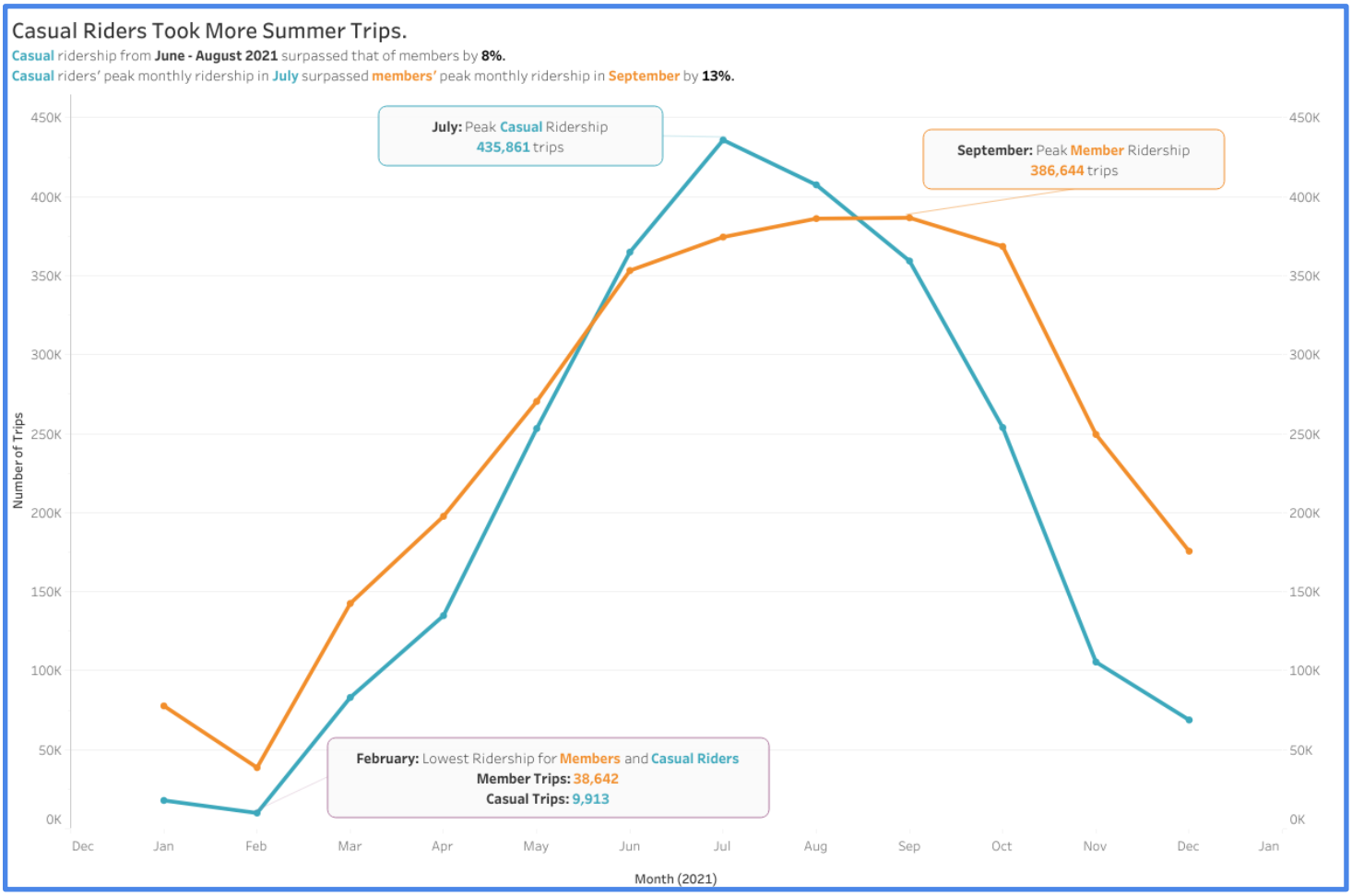
Time of Year: What does ridership look like for casual riders during a given month?

Analysis

I initially wanted to look more granularly at how many daily trips each rider type took throughout the year and how this fluctuated over time. With this goal in mind, I queried the master table in SQL to count the number of unique ride ids present for each date (extracted from the starting timestamp) in 2021, filtering by rider type (executed for members first, and then re-executed for casual riders) and filtering out all false starts and canceled trips. The output for each rider type was grouped by trip date. The query also included a column for each date's weekday.

I combined the results for members and casual riders into one sheet in Excel for further analysis there and in Tableau. While I didn't end up exploring the day-to-day counts as I originally planned (this would have been 'information overload' for the question I was trying to answer), having the data broken down by the individual dates and the associated day of the week meant that I could use this same dataset to explore trip totals for each rider type by month *and* by day of the week utilizing pivot tables in Excel and visualization in Tableau.

Findings



View the full-sized version of this visual.

The overall trend for 2021 is that ridership for all groups increased steadily as the season changed from winter into spring and summer, peaked around midsummer (for casual riders) and early fall (for members), and then decreased again from fall into winter. Based on this pattern, users seem more likely to start trips in warmer weather.

February was the month of the lowest ridership for members (38,642 trips) and casual riders (9,913 trips). This severe low is not surprising, given that February was the coldest month for the Chicago area. According to National Weather Service data, the month's average temperature of 20.2 degrees Fahrenheit was 8.6 degrees lower than the normal temperature for that month.

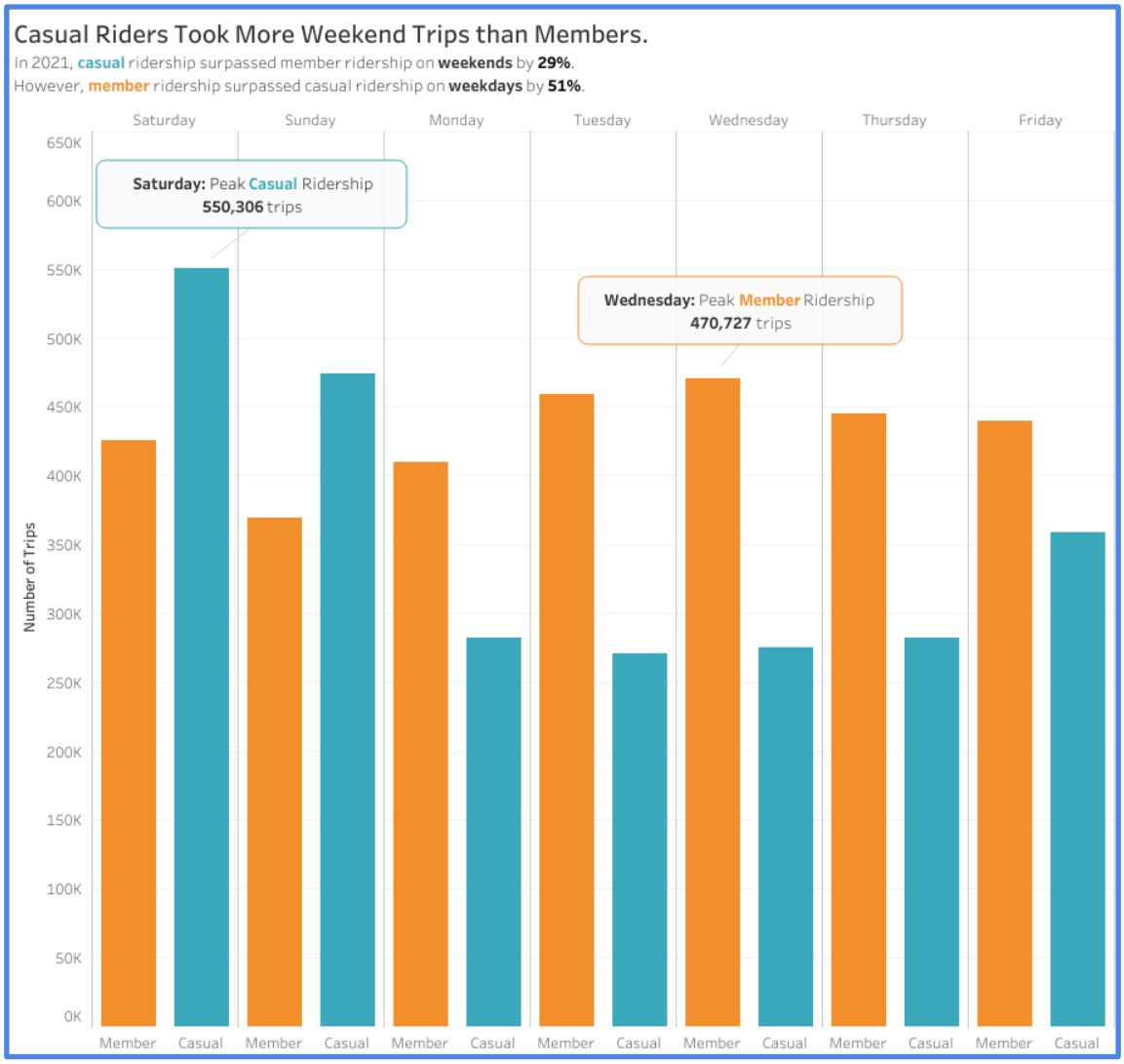
Member ridership surpassed casual ridership for most of the year, with the difference being most pronounced during the fall and winter. In January 2021, over triple as many trips were started by members than casual riders. However, the difference became less pronounced in the spring, and then the balance shifted in the summer months. From June - August 2021, casual ridership surpassed that of members by 8%. This boom in casual ridership is likely related to summer tourism in the area. Casual riders' peak monthly ridership in July (435,861 trips) surpassed members' peak monthly ridership in September (386,644 trips) by **13%.**

Day of Week: What does ridership look like for casual riders during a given day of the week?

Analysis

Please see the previous section.

Findings



View the full-sized version of this visual.

Member ridership in 2021 remained relatively stable regardless of the day of the week, which tracks with the idea that many members use Cyclistic bikes to commute to work. The total trips completed by this group were consistently within the 350K-475K range for each day of the week, with the most popular day for member trips being Wednesday (470,727 trips).

Meanwhile, casual ridership fluctuated throughout the week. The total trips completed by this group were consistently within the 250K-300K range for most weekdays (Monday through Thursday). However, they were markedly higher for

Friday (359,016 trips) and even higher for Saturday (550,306 trips - the most popular day for casual riders) and Sunday (474,302 trips). This tracks with the idea that many casual riders use Cyclistic bikes for recreation.

Overall, on weekdays, member ridership surpassed casual ridership by **51%**. On weekends, casual ridership surpassed member ridership by **29%**.

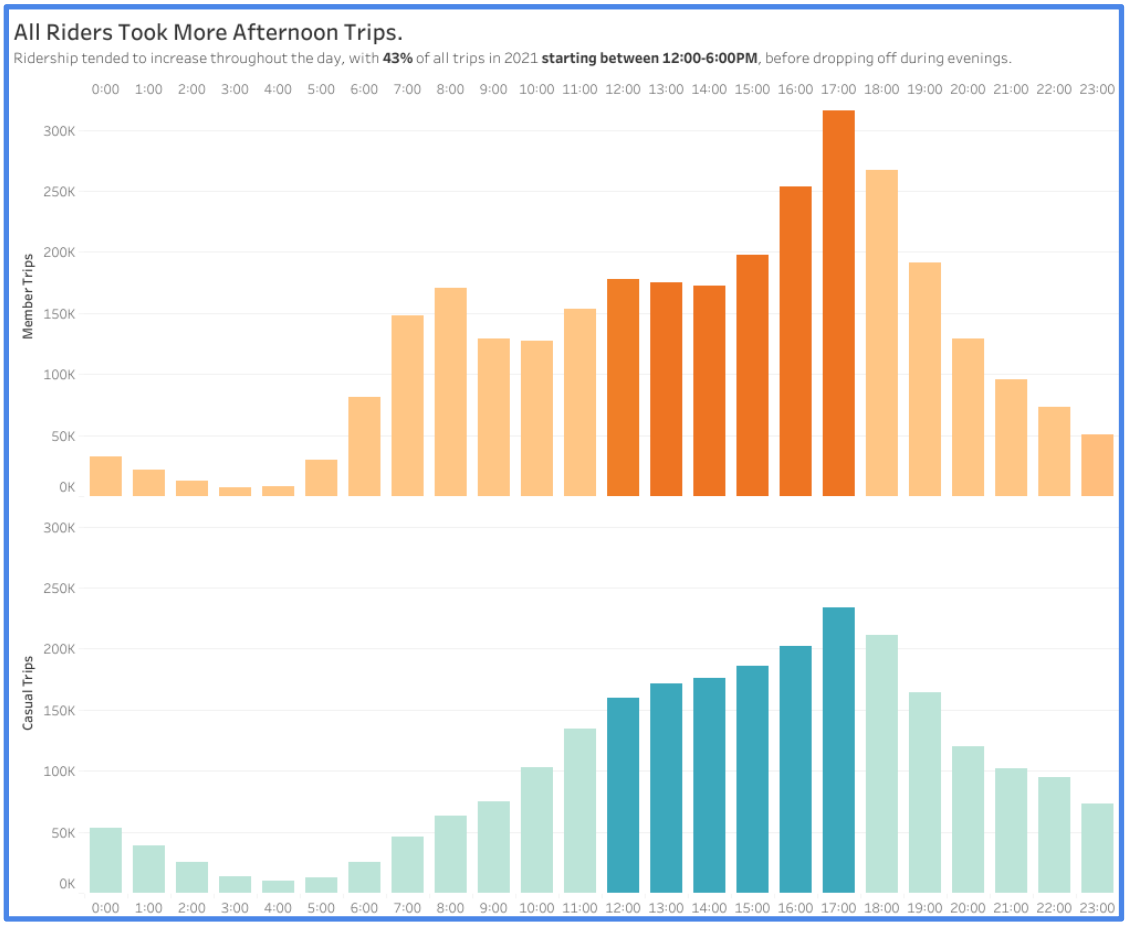
Time of Day: What does ridership look like for casual riders during a given hour of the day?

Analysis

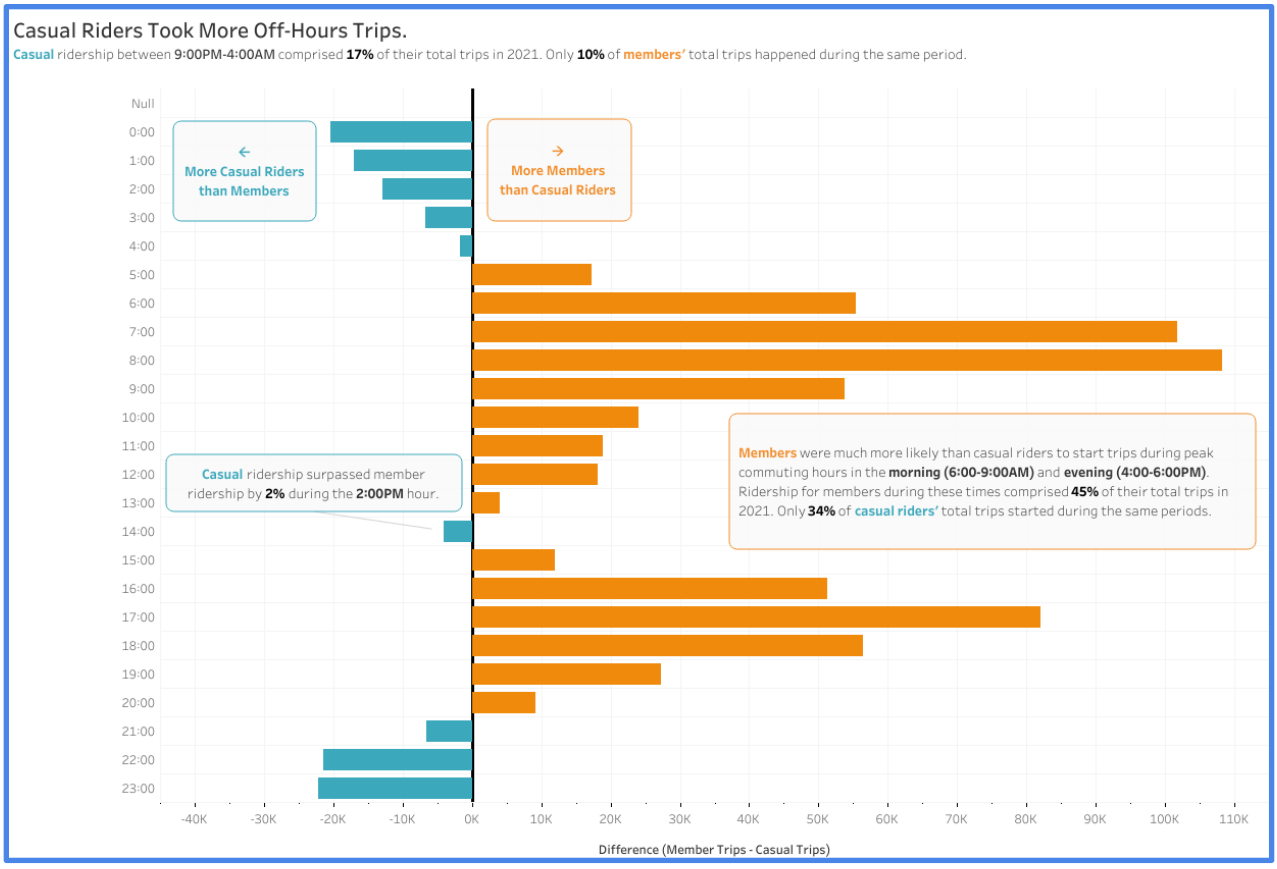
To find the number of trips that each rider type started during a given hour, I queried the master table in SQL to count the number of unique ride ids present for each hour of the day (extracted from the starting timestamp) in 2021, filtering out all false starts and canceled trips. I included a column for rider type but did not filter based on this column, instead grouping the output by starting hour and then by rider type.

I cleaned this output to create a sheet in Excel that showed member, casual, and overall ride totals side by side. Next, I worked with this dataset in Tableau to create another calculated field that tabulated the difference between member and casual totals for each hour, which helped me visualize how much higher one group's total was over the other group's total for a given hour.

Findings



View the full-sized version of this visual.



View the full-sized version of this visual.

In 2021, ridership for all groups tended to increase throughout the day, peak during the evening commute, and drop off during the evenings. **43%** of all trips in 2021 took place between **12:00 - 6:00 PM**.

Members were much more likely than casual riders to start trips during peak commuting hours (morning 6:00 - 9:00 AM, evening 4:00 - 6:00 PM); trips started during those time ranges comprised **45%** of their total trips in 2021. Conversely, only **34%** of casual riders' total trips began during the same periods.

While member ridership surpasses casual ridership throughout much of the day, at times by as much as double (during morning commute hours), casual riders were more likely than members to start trips later in the evening. Their ridership from 9:00 PM - 4:00 AM comprised **17%** of their total trips in 2021. Only **10%** of members' total trips happened during the same period. One other brief period when casual riders were slightly more likely to ride than members was during the 2:00 PM hour when their ridership surpassed that of members by **2%.**

Ride Length: How long do casual riders' trips typically last?

Analysis

I started by querying the master table in SQL for various ride length statistics (average, median, max) grouped by each rider type, filtering out all false starts and canceled trips. Over this query, I also calculated the percentage of the overall total hours the sum of each rider type's time represented.

The results of this query indicated that there was a significant difference between casual riders and members when it came to ride lengths. Not only did casual riders tend to take longer rides (their average and median ride lengths

outstripped members by around 14 and 7 minutes, respectively), but they also rode thousands more hours than members did in 2021. Casual riders' total ride time comprised 62% of the overall total.

Seeing this stark difference between the rider types prompted me to explore whether ride lengths differed for each group depending on the day of the week. I modified the previous query to add a column for the day of the week and grouped the output by day of the week and rider type.

The average times for each rider group skewed longer than the median times because riders can rent a bike on a full-day pass. Because of this bias, I used the median rather than the average to measure the difference between rider types.

Findings

View the full-sized version of this visual.

Casual riders consistently took longer trips than members. Their median ride lengths ranged between **50-73% greater** than members' on any given day of the week in 2021. This contrast could be because casual riders are more likely to ride for

leisure and take more time for sightseeing or to enjoy the ride itself, whereas members are more likely to ride to commute and want to get from point A to B as efficiently as possible.

Casual riders' ride lengths also tended to fluctuate throughout the week, with their longest trips taking place on Saturdays (where their median ride length was 64% longer than that of members), Sundays (71% longer), and Mondays (73% longer). Meanwhile, members' ride lengths stayed relatively consistent; their median ride length hovered around 9 minutes on weekdays and 11 minutes on weekends.

Recommendations

Firstly, Cyclistic should time its next marketing campaign strategically to get the attention of casual riders when they are using its services the most. As we've identified that casual ridership peaks during the summer months, digital and in-person marketing efforts should take place during the general timeframe of June, July, and August. Relatedly, we should target in-person marketing efforts (e.g., membership sign-up drives at Cyclistic stations) to happen when casual riders are most likely to be at stations, which we've identified as weekends and afternoons. Cyclistic staff could be available onsite at a minimum on Fridays, Saturdays, and Sundays between 12:00 - 6:00 PM to assist with sign-ups.

Secondly, Cyclistic should target those in-person marketing efforts at the most popular stations utilized by casual riders, which we've identified to be distinct from those of members. Onsite staffing to assist with the membership drive could be concentrated on casual riders' top 10 stations (especially the Streeter Dr & Grand Ave station, which is the most popular station for casual riders by a wide margin). In addition, promotional materials (such as posters) could be made available at a broader range of casual riders' preferred stations.

Thirdly, messaging for the marketing campaign should emphasize the flexibility of a Cyclistic membership to convince casual riders that membership is a worthwhile investment. Given casual riders tend to take longer rides than members, a membership might be a more cost-effective option for them since the $0.15/minute charge only kicks in after the first 45 minutes of a trip for members (as opposed to 30 minutes with the single ride pass casual riders purchase). In addition, with membership, they would have access to unlimited rides included with the monthly fee; as long as they can get to another station to exchange bikes before their 45 minutes is up, they could avoid the per-minute charge entirely.

Additional Considerations

While it was beyond the scope of the objective I was working on, as well as what could be garnered from the data at my disposal, Cyclistic should conduct additional research into the viability of creating a separate membership tier that might appeal more to current casual riders than the existing annual membership.

Based on my research into how members and casual riders differ, while the existing membership plan makes a lot of sense for the numerous members who use Cyclistic to commute on weekdays, it might not be enticing to casual riders if they're largely riding for leisure on weekends alone. One option that Cyclistic could research further would be a "weekends only" membership tier at a lower price point that would allow users to take advantage of unlimited 45-minute rides on Fridays, Saturdays, and Sundays. This research could be informed by the insights one of my colleagues uncovered as they explored the question, "why would casual riders buy Cyclistic annual memberships?"

Appendix

Data Dictionary

Original Datasets

For more information about the datasets, please see the Data Structure section.

Field Name	Type	Description
rideable_type	STRING	Type of bike used for trip (classic bike, electric bike, docked bike).
started_at	DATETIME	Date and time trip started.
ended_at	DATETIME	Date and time trip ended.
start_station_name	STRING	Station where trip started.
start_station_id	STRING	Identifier for station where trip started.
end_station_name	STRING	Station where trip ended.
end_station_id	STRING	Identifier for station where trip ended.
start_lat	FLOAT	Latitude where trip started.
start_lng	FLOAT	Longitude where trip started.
end_lat	FLOAT	Latitude where trip ended.
end_lng	FLOAT	Longitude where trip ended.
member_casual	STRING	Membership status of rider (casual, member).

Columns Added During Analysis

For more information about how the columns below were created, please see the Data Preparation section.

Field Name	Type	Description
ride_length	INTERVAL	Time elapsed (hh:mm:ss) between started_at and ended_at timestamps.
day_of_week	STRING	Day of the week when the trip started.
distance_traveled	FLOAT	Distance (in miles) between start and end coordinates.
trip_type	STRING	Type of trip (canceled, out-and-back, point-to-point).

Data Integrity Issues

Missing Values - Count by Month

For more information about the missing values below, please see the Data Integrity and Issues Found section.

Month	Missing Start Station Name/ID	Missing End Station Name/ID	Missing End Latitude/Longitude
January	8625	10277	103
February	4046	5358	214
March	14848	16727	167
April	26506	28174	267
May	53744	58194	452
June	80093	86387	717
July	87263	93158	731
August	88458	94115	706
September	93113	99261	595
October	108210	114834	484
November	75290	79187	191
December	51063	53498	144
Total	**691259**	**739170**	**4771**

Unclear Categories - Records Updated by Month

For more information, please see the Data Integrity and Issues Found section.

Month	Records Updated
January	2106
February	1271
March	15657
April	24714
May	43353
June	51716
July	57698
August	45065
September	35337
October	22449
November	7461
December	4878
Total	**311705**

Time Entry Glitches - Records Removed by Month

For more information, please see the Data Integrity and Issues Found section.

Month	Records Removed
January	1
February	0
March	2
April	5
May	2
June	5
July	13
August	29
September	36
October	0
November	53
December	0
Total	**146**