Cyclistic bike-share ridership analysis

Understanding rider usage patterns to optimize marketing strategies

Paidraig 'Patrick' O'Ceallaigh

2024-04-24

Contents

Executive summary

Business Question: How do annual members and casual riders use Cyclistic bikes differently?

Key Findings:

- Annual members predominantly use the Cyclistic system, accounting for 66% of total rides. Their usage is consistent throughout the year and peaks during weekday commuting times. They prefer shorter, consistent rides and slightly favor classic bikes.
- Casual riders, while fewer, still significantly contribute to the ridership, especially during warmer months and weekends. They prefer longer, leisurely rides and show a strong preference for electric bikes.

Recommendations:

1. Increase maximum ride duration for members during off-peak times: This could encourage longer rides, enhance user experience, and attract casual riders looking for extended rides.
2. Increase the number of electric bikes: To meet the demand of casual riders who prefer electric bikes, expanding the fleet could encourage more casual riders to convert to annual memberships.

3. Implement a loyalty program: A point-based system rewarding casual riders for frequent use and offering incentives to become members could make the transition more appealing to casual riders. Exclusive perks for members could further demonstrate the added value and benefits of membership. These recommendations aim to optimize Cyclistic's bike-share system and encourage casual riders to convert to annual memberships. By catering to the diverse needs and preferences of both rider types, Cyclistic can increase ridership and revenue.

Introduction

In the evolving landscape of urban mobility, bike-share programs have become a cornerstone for convenient and sustainable transportation. Cyclistic, a prominent bike-share company based in Chicago, has carved out a niche in this dynamic market since its inception in 2016. With a diverse fleet of over 5,800 bicycles, including specialized models like reclining bikes and hand tricycles, Cyclistic ensures inclusivity and accessibility for a broad user base. The program's infrastructure is robust, with 692 docking stations across Chicago that support geotracking and seamless bike exchanges. While the company's diverse fleet and flexible pricing plans have been instrumental in attracting a wide range of users, the majority of rides are for leisure, with approximately 30% serving commuters.

Background

Under the guidance of Lily Moreno, the Director of Marketing, Cyclistic has achieved significant growth through strategic marketing and customer engagement initiatives. Historically, the company's marketing efforts have focused on building brand awareness and appealing to a broad consumer base, utilizing channels like email, social media, and direct promotions. Cyclistic offers a range of pricing options from single-ride passes to annual memberships, with the latter being significantly more profitable. This insight has prompted a strategic pivot aimed at increasing the proportion of annual memberships, which are deemed crucial for the company's long-term financial health and market dominance.

Objectives

The primary objective of this analysis is to delineate the usage patterns of annual members versus casual riders. By understanding these differences, the team aims to uncover the underlying factors that influence a rider's decision to opt for casual passes or commit to an annual membership. This investigation will focus on various aspects such as duration of rides, preferred bike types, peak usage times, and the purposes of trips (leisure versus commuting).

The insights derived from this analysis will serve as the foundation for developing targeted marketing strategies aimed at converting casual riders into annual members. This strategic shift is not only expected to enhance customer retention but also to maximize profitability.

Research question

> How do annual members and casual riders use Cyclisitc bikes differently?

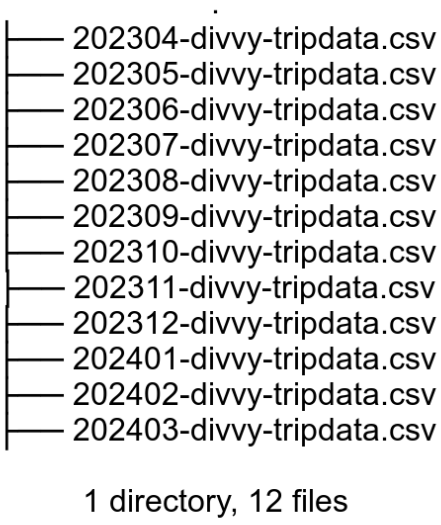
This research question encapsulates the core objective of the analysis, focusing on the distinct usage patterns and preferences of annual members and casual riders within the Cyclistic bike-share program. By delving into this question, the analysis aims to uncover actionable insights that can inform strategic marketing decisions and drive the conversion of casual riders into annual members.

Data collection

Data sources

For the comprehensive analysis of usage patterns between annual members and casual riders, the primary data source comprises the most recent twelve months of historical trip data from Cyclistic's bike-share program. This dataset is publicly accessible and consists of twelve individual files, each representing a month's worth of data, formatted in CSV. These files encapsulate extensive records of all bike trips taken during this period,

capturing a variety of attributes related to each trip. The data has been made available under the terms specified in the provided license and can be downloaded here.



Data description

The raw dataset for this analysis encompasses a substantial volume of 5,750,177 individual observations, organized into 13 distinct variables. Each observation corresponds to a unique bike trip, with the characteristics of each trip delineated by the respective variables and their types. The dataset's structure is designed to capture essential details about the rides, including:

Table 1: Variables

Variable	Type	Description
‘ride_id‘	<chr>	A unique id for each bike ride
‘rideable_type‘	<chr>	Type of bicyle used for the ride
‘started_at‘	<dttm>	Timestamp for the begining of the ride
‘ended_at‘	<dttm>	Timestamp for the end of the ride
‘start_station_name‘	<chr>	Name of the station from which the ride started
‘start_station_id‘	<chr>	ID of the staion from which the ride started
‘end_station_name‘	<chr>	Name of the station at which the ride ended
‘end_station_id‘	<chr>	ID of the station at which the ride ended
‘start_lat‘	<dbl>	Latitude at which the ride started
‘start_lng‘	<dbl>	Longitude at which the ride started
‘end_lat‘	<dbl>	Latitude at which the ride ended
‘end_lng‘	<dbl>	Longitude at which the ride ended
‘member_casual‘	<chr>	Type of rider

This detailed and structured dataset provides a solid foundation for analyzing and comparing the different usage patterns of Cyclistic's bike-share system by annual members and casual riders. Through this analysis, the marketing analytics team aims to derive insights that can influence strategic marketing decisions, ultimately converting more casual riders into annual members.

Challenges and limitations

While the dataset is rich in information and provides a comprehensive view of bike trips taken over the past year, there are several challenges and limitations that need to be considered:

1. *Large Dataset*: The dataset is extensive, comprising over 5,000,000 bike trip observations across 13 variables. This vast volume of data can pose computational challenges and may require sophisticated data processing techniques to handle efficiently.
2. *Data Privacy*: Due to data privacy regulations and ethical considerations, the use of riders' personally identifiable information is prohibited. This limitation prevents the ability to link pass purchases to specific credit card numbers, which restricts the analysis in certain ways. For instance, it is not possible to ascertain whether casual riders reside within the Cyclistic service area or if they have purchased multiple single passes.
3. *Missing Geospatial Data*: Some of the geospatial data, including station names and station IDs, is missing from the dataset. This absence of data limits the scope of investigation into these aspects of the dataset. For example, it might restrict the analysis of usage patterns related to specific stations or geographical areas.
4. *Ride Distance Calculations*: The calculation of ride distances is based on the latitude and longitude data provided in the dataset. These calculations yield "as the crow flies" distances, which are straight-line measurements between the start and end points of each ride. This approach does not account for the actual path taken by the rider, including detours, road layouts, or other factors that might affect the real distance traveled. Consequently, the calculated distances may not accurately reflect the actual distances covered during each ride.

These limitations should be taken into account when interpreting the results of the analysis. Despite these challenges, the dataset still offers valuable insights into the usage patterns of Cyclistic's bike-share system by different rider types.

Data cleaning

Data inspection

The data inspection process serves as a preliminary check to ensure the quality and structure of the dataset before proceeding with detailed analysis. This stage is crucial for identifying potential issues that could affect the validity of the study's conclusions. The following steps were undertaken to inspect the dataset comprehensively:

Examine structure of data To begin, we examined the overall structure and type of the data contained within the dataset. This was accomplished using the glimpse() function from the dplyr package, which provides a compact display of the types of each column along with a quick view of the first few entries in each.

```
Rows: 5,750,177
Columns: 13
$ ride_id            <chr> "8FE8F7D9C10E88C7", "34E4ED3ADF1D821B", "5296BF07A2F77CB5", "40759916B76D5D52",
$ rideable_type      <chr> "electric_bike", "electric_bike", "electric_bike", "electric_bike", "electric_b
$ started_at         <dttm> 2023-04-02 08:37:28, 2023-04-19 11:29:02, 2023-04-19 08:41:22, 2023-04-19 13:3
$ ended_at           <dttm> 2023-04-02 08:41:37, 2023-04-19 11:52:12, 2023-04-19 08:43:22, 2023-04-19 13:3
$ start_station_name <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,
$ start_station_id   <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,
$ end_station_name   <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,
$ end_station_id     <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,
$ start_lat          <dbl> 41.80, 41.87, 41.93, 41.92, 41.91, 41.91, 41.93, 42.00, 41.99, 41.88, 41.87, 41
$ start_lng          <dbl> -87.60, -87.65, -87.66, -87.65, -87.65, -87.63, -87.66, -87.66, -87.66, -87.65,
$ end_lat            <dbl> 41.79, 41.93, 41.93, 41.91, 41.91, 41.92, 41.91, 41.99, 42.00, 41.88, 41.93, 41
$ end_lng            <dbl> -87.60, -87.68, -87.66, -87.65, -87.63, -87.65, -87.65, -87.66, -87.66, -87.65,
$ member_casual      <chr> "member", "member", "member", "member", "member", "member", "member", "member",
```

Check duplicate observations Identifying and eliminating duplicate records is essential to ensure the integrity of the analysis. Given that each record in the dataset contains a unique identifier called `ride_id`, we anticipated minimal duplication. To verify this, the dataset was processed using the `dplyr` package, which is part of the comprehensive tidyverse suite of tools for data manipulation in R. We grouped the data by all columns to account for complete record duplication, filtered out any duplicates, and counted the remaining rows.

Identifying and eliminating duplicate records is essential to ensure the integrity of the analysis. Given that each record in the dataset contains a unique identifier called `ride_id`, we anticipated minimal duplication. To verify this, the dataset was processed using the `dplyr` package, which is part of the comprehensive tidyverse suite of tools for data manipulation in R. We grouped the data by all columns to account for complete record duplication, filtered out any duplicates, and counted the remaining rows:

```
duplicate_obs <- data_raw %>%
  group_by_all() %>%
  filter(n() > 1) %>%
  ungroup() %>%
  nrow_lazy_dt()
```

After this thorough check, no duplicate records were found in the dataset, confirming that each `ride_id` indeed represents a unique trip event. This finding solidifies the dataset's reliability for subsequent analyses, ensuring that each entry uniquely contributes to the insights generated.

Check missing values Handling missing data is crucial as it can significantly impact the analysis. To assess and quantify missing values within the dataset, we utilized the `dplyr` and `tidyr` packages from the tidyverse collection, which are robust tools for data manipulation in R. The procedure involved summarizing the missing values across all columns and transforming the results for a clearer presentation.

```
missing_obs <- data_raw %>%
  summarise(across(everything(), ~ sum(is.na(.), na.rm = TRUE))) %>%
  pivot_longer(
    everything(),
    names_to = "variable",
    values_to = "missing_count"
  ) %>%
  filter(missing_count > 0) %>%
  as_tibble()
```

Upon inspection, it was discovered that certain variables related to station identification and location data exhibited notable gaps.

Table 2: Missing values

Variable	Missing values
‘start_station_name‘	874450
‘start_station_id‘	874450
‘end_station_name‘	929226
‘end_station_id‘	929226
‘end_lat‘	7566
‘end_lng‘	7566

These findings indicate significant missing data in station identifiers and geolocation details, which could affect spatial analysis and ride linkage accuracy. Such gaps will necessitate careful consideration in subsequent stages of data cleaning and analysis to ensure that conclusions drawn are based on comprehensive and accurate information.

Check categorical variables Analyzing the consistency and accuracy of categorical variables is essential for ensuring data integrity. In this analysis, the `dplyr` package was employed to count and evaluate the distinct categories present in the `rideable_type` and `member_casual` variables.

Rideable_type variable The `rideable_type` variable categorizes the types of bikes used in the trips. While only `classic_bike` and `electric_bike` were expected based on company records, an additional category, `docked bike`, was identified, which was not anticipated:

```
count_rideable_type <- data_raw %>%
  count_categorical(rideable_type)
```

The presence of `docked bike` raises questions about its meaning, as it could potentially refer to bikes that are either not clearly classified between the two expected types or perhaps represent bikes that have been temporarily removed from active service for maintenance. This ambiguity requires further investigation to accurately interpret the data involving this category.

Member_casual variable For the `member_casual` variable, which distinguishes between members (annual subscribers) and casual riders (non-subscribers), the results were as expected. Only the two anticipated values, "member" and "casual," were found, confirming the accuracy of data entries for this variable:

```
count_member_casual <- data_raw %>%
  count_categorical(member_casual)
```

This assessment underscores the importance of validating categorical data against expected outcomes to ensure the reliability of analyses based on these classifications. For the `rideable_type` variable, particularly, the unexpected category will be noted for potential adjustments in the analysis framework or for clarification from data management teams.

Data transformation

The data transformation process involves several steps aimed at refining the dataset to enhance the quality and relevance of the analysis. These transformations are crucial for ensuring that the data is clean, properly structured, and ready for detailed analytical tasks. The following transformations were performed on the Cyclistic dataset:

Remove unnecessary variables Initially, the dataset included variables related to station identification, specifically the names and IDs of start and end stations (`start_station_name`, `start_station_id`, `end_station_name`, `end_station_id`). These variables comprised the majority of missing data within the dataset, posing potential challenges for complete and accurate analysis. While station identification information could offer valuable insights for future studies, particularly in understanding spatial patterns and usage trends, it was not deemed pertinent to the current analysis' objectives. Consequently, to streamline the dataset and focus on the most relevant variables for examining rider behavior and bike usage patterns, these station identification variables were removed. This step was executed using the dplyr package's select() function to exclude the specified columns.

```
irrelevant_vars <- c(
  "start_station_name",
  "start_station_id",
  "end_station_name",
  "end_station_id"
)

data_processed <- data_raw %>%
  select(-all_of(irrelevant_vars))
```

By removing unnecessary variables, we reduced the complexity of the dataset, making it more manageable and aligned with the specific objectives of the analysis. This focused approach ensures that the subsequent analyses are conducted efficiently and effectively, emphasizing the key variables that drive insights into rider behavior and bike usage patterns.

Remove missing values The presence of missing values can hinder the accuracy and reliability of the analysis. To address this issue, the dataset was filtered to exclude rows with missing values in the latitude and longitude columns for both the start and end stations. This step ensures that only complete and valid data entries are retained for further analysis, enhancing the dataset's quality and integrity. The removal of missing values was carried out using the `filter()` function from the `dplyr` package.

```
data_processed <- data_processed %>%
  filter(
    !is.na(end_lat),
    !is.na(end_lng)
  )
```

Exclude invalid categories During the initial data inspection phase, it was noted that all categorical variables conformed to expected categories except for one unexpected entry in the `rideable_type` variable. The `rideable_type` included an unforeseen category labeled `docked_bike`, which was not anticipated based on the operational definitions provided for the dataset. The presence of this unexpected category raised concerns about its definition, as it could potentially represent a classification error, bikes taken out of active circulation for maintenance, or another unspecified category.

Given the uncertainty surrounding the "docked bike" category and its deviation from the expected "classic bike" and "electric bike" dichotomy, a decision was made to exclude this category from the current analysis. This approach ensures that the analysis remains focused on well-defined and understood categories, thus maintaining the integrity of the analytical outcomes.

The exclusion of the docked_bike category was implemented using the `filter()` function from the `dplyr` package.

```
data_processed <- data_processed %>%
  filter(!rideable_type %in% "docked_bike")
```

Removing the `docked_bike` category ensures that the dataset is consistent with the expected bike types and that the subsequent analyses are based on accurate and reliable data. This step is essential for maintaining the quality and validity of the analysis results. For future analyses, further clarification on the definition and implications of the docked_bike category may be necessary to incorporate this data effectively.

Convert data types In R, it's often helpful to convert categorical variables to factors. This conversion facilitates efficient handling of categorical data in functions used for statistical modeling and plotting. Factors help define a set of unique values that R recognizes, optimizing both memory use and processing.

Converting `rideable_type` and `member_casual` from characters to factors ensures that these variables are appropriately processed during analysis, preventing potential errors and improving performance. This conversion was carried out using the `mutate()` function from the `dplyr` package.

```
data_processed <- data_processed %>%
  mutate(
    rideable_type = as.factor(rideable_type),
    member_casual = as.factor(member_casual)
  )
```

This transformation standardizes the data types within the dataset, ensuring that categorical variables are correctly represented as factors for consistent and accurate analysis.

Table 3: Additional variables

Variable	Description
'ride_duration'	Represents the duration of the ride in minutes. This is calculated from the difference between the start and end timestamps, providing a quantitative measure of ride length.
'ride_month'	Indicates the month in which the ride started. This variable can help analyze monthly usage patterns and seasonal variations in bike usage.
'ride_week'	Identifies the week number relative to the beginning of the dataset, allowing for analysis of weekly trends or changes over time.
'ride_day_of_week'	Specifies the day of the week the ride began. This categorical variable is useful for understanding daily usage patterns, such as distinguishing between weekday and weekend trends.
'ride_start_hour'	Captures the hour of the day the ride started, which can be crucial for identifying peak usage times during the day.
'day_type'	Categorizes each ride as either occurring on a 'Weekday' or 'Weekend.' This differentiation can help in analyzing differences in riding behavior between workdays and leisure days.
'ride_distance'	Measures the straight-line distance of the ride, calculated using the Haversine formula to determine the 'as the crow flies' distance between the start and end coordinates. This provides an estimate of the spatial extent of each ride.

Rename variables The dataset's variable names were initially based on the original data source and may not be intuitive or aligned with the analysis objectives. To this end, renaming variables to more descriptive and concise names can enhance clarity and facilitate easier interpretation of the data. This renaming process was carried out using the `rename()` function from the `dplyr` package.

```
data_processed <- data_processed %>%
  rename(
    bike_type = rideable_type,
    rider_type = member_casual,
    start_timestamp = started_at,
    end_timestamp = ended_at,
    start_latitude = start_lat,
    start_longitude = start_lng,
    end_latitude = end_lat,
    end_longitude = end_lng
  )
```

Standardizing variable names enhances clarity and coherence within the dataset, making it easier to interpret and analyze the data effectively. This step is essential for maintaining consistency and ensuring that all variables are clearly labeled and identifiable throughout the analysis process.

Create additional variables To enhance the dataset with additional information that could provide deeper insights into riding patterns, several new variables were created. Each variable is derived from existing data but is designed to facilitate more specific analyses. The new variables include `ride_duration`, `ride_month`, `ride_week`, `ride_day_of_week`, `ride_start_hour`, `day_type`, and `ride_distance`. These variables capture essential aspects of bike rides, such as duration, temporal patterns, and spatial characteristics, enabling more detailed and nuanced analyses.

These new variables offer valuable insights into ride duration, temporal and spatial patterns, and day-of-week trends, providing a comprehensive view of rider behavior and bike usage patterns. It is expected that these variables will play a crucial role in the subsequent analyses, enabling a more detailed exploration of the

differences between annual members and casual riders. This variable creation process was executed using the `mutate()` function from the `dplyr` package.

```
data_processed <- data_processed %>%
  mutate(
    ride_duration = as.numeric(difftime(
      end_timestamp,
      start_timestamp,
      units = "mins"
    )),
    ride_month = as.Date(floor_date(start_timestamp, "month")),
    ride_week = floor_date(start_timestamp, "week"),
    ride_day_of_week = factor(wday(start_timestamp, label = TRUE)),
    ride_start_hour = factor(hour(start_timestamp)),
    day_type = factor(ifelse(
      ride_day_of_week %in% c("Sat", "Sun"),
      "Weekend",
      "Weekday"
    )),
    ride_distance = distHaversine(
      cbind(start_longitude, start_latitude),
      cbind(end_longitude, end_latitude)
    )
  )
```

These transformations enhance the dataset and prepare it for in-depth exploratory analyses, ensuring that the data is well-suited for comprehensive and advanced analytical tasks.

Remove outliers Removing outliers is crucial to maintaining the integrity of the data analysis. Outliers can skew results, distort patterns, and lead to inaccurate conclusions. In this analysis, we focused on eliminating atypical values in ride duration and distance that could potentially distort the results. By setting appropriate thresholds based on operational insights and statistical considerations, we identified and removed anomalous observations to ensure the dataset's reliability and accuracy.

Ride duration For ride duration, we established thresholds based on realistic expectations and specific operational insights provided by the bike-share program. The minimum duration was set at 1 minute to account for "false starts" or instances where users might re-dock their bikes shortly after undocking to ensure the bike was securely locked. According to the dataset guidelines provided by the company, any trip recorded under 1 minute is likely not a genuine trip but rather an operational anomaly such as this. The maximum allowable duration was set at 180 minutes (3 hours). This upper limit is designed to exclude rides that unusually extend beyond normal usage patterns, which might indicate situations where bikes were removed for maintenance but their operational status was not updated correctly in the system. Such errors can occur and lead to falsely prolonged ride times being recorded.

```
# Check for anomalous ride duration
min_duration <- 1
max_duration <- 180

# Count anomalous durations
anomalous_duration <- data_processed %>%
  count_anomalies(ride_duration, min_duration, max_duration)

# Remove rides with anomalous duration
data_processed <- data_processed %>%
  remove_anomalies(ride_duration, min_duration, max_duration)
```

Rides less than 1 minute accounted for a small proportion of the remaining dataset (2.51%), while rides exceeding 3 hours represented an even smaller proportion of the dataset (0.22%). It was determined that removing these outliers would not significantly impact the overall dataset's integrity, as they were considered operational anomalies rather than genuine rides. By setting these thresholds and removing the identified outliers, the dataset was refined to exclude extreme values that could distort the analysis results.

Ride distance For ride distance, outliers were addressed by setting statistical thresholds based on the distribution of the data. The maximum distance (`max_distance`) was calculated as three times the interquartile range (IQR) above the 75th percentile. Three times the IQR was chosen to account for potential variability in ride distances while excluding extreme values that could indicate errors or anomalies. The minimum distance (`min_distance`) was set at the 1st percentile to filter out rides that were improbably short, potentially indicative of errors like GPS discrepancies at the start or end of a ride.

```
ride_distance_stats <- data_processed %>%
  calculate_summary_stats(ride_distance)

# Calulate IQR
iqr_distance <- ride_distance_stats$q75 - ride_distance_stats$q25

# Calculate maximum threshold
max_distance <- ride_distance_stats$q75 + 3 * iqr_distance

# Calculate 99th percentile for the minimum threshold
min_distance <- data_processed %>%
  filter(ride_distance > 0) %>%
  summarise(min_dist = quantile(ride_distance, 0.01)) %>%
  pull()

# Count anomalous distances
anomalous_distance <- data_processed %>%
  count_anomalies(ride_distance, min_distance, max_distance)

# Remove rides with anomalous distance
data_processed <- data_processed %>%
  remove_anomalies(ride_distance, min_distance, max_distance)

obs_nonanomalous <- data_processed %>%
  nrow_lazy_dt()
```

The maximum distance threshold was calculated to be 8441.47 meters (5.25 miles), while the minimum distance threshold was set at 55.79 meters (183.03 feet). After removing the 557,483 outliers, the dataset retained 5,192,694 non-anomalous observations, ensuring that the data was free of extreme values that could distort the analysis results, thereby enhancing the dataset's reliability and accuracy.

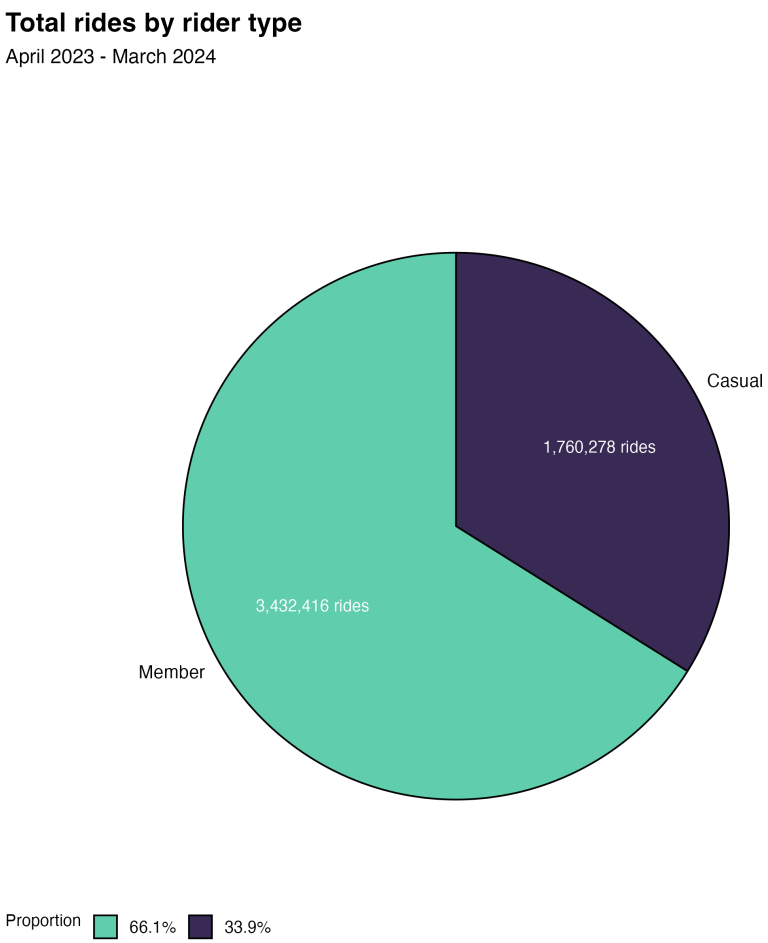
Data analysis

After cleaning and transforming the dataset to ensure its quality and relevance, we moved to the data analysis phase, which was designed to uncover actionable insights on how Cyclistic's annual members and casual riders utilize the bike-share system differently. Through a combination of statistical methods and graphical techniques, the analysis aims to highlight distinct usage patterns and preferences between the two rider types. Here are the key steps and findings from this in-depth analysis.

Ridership patterns

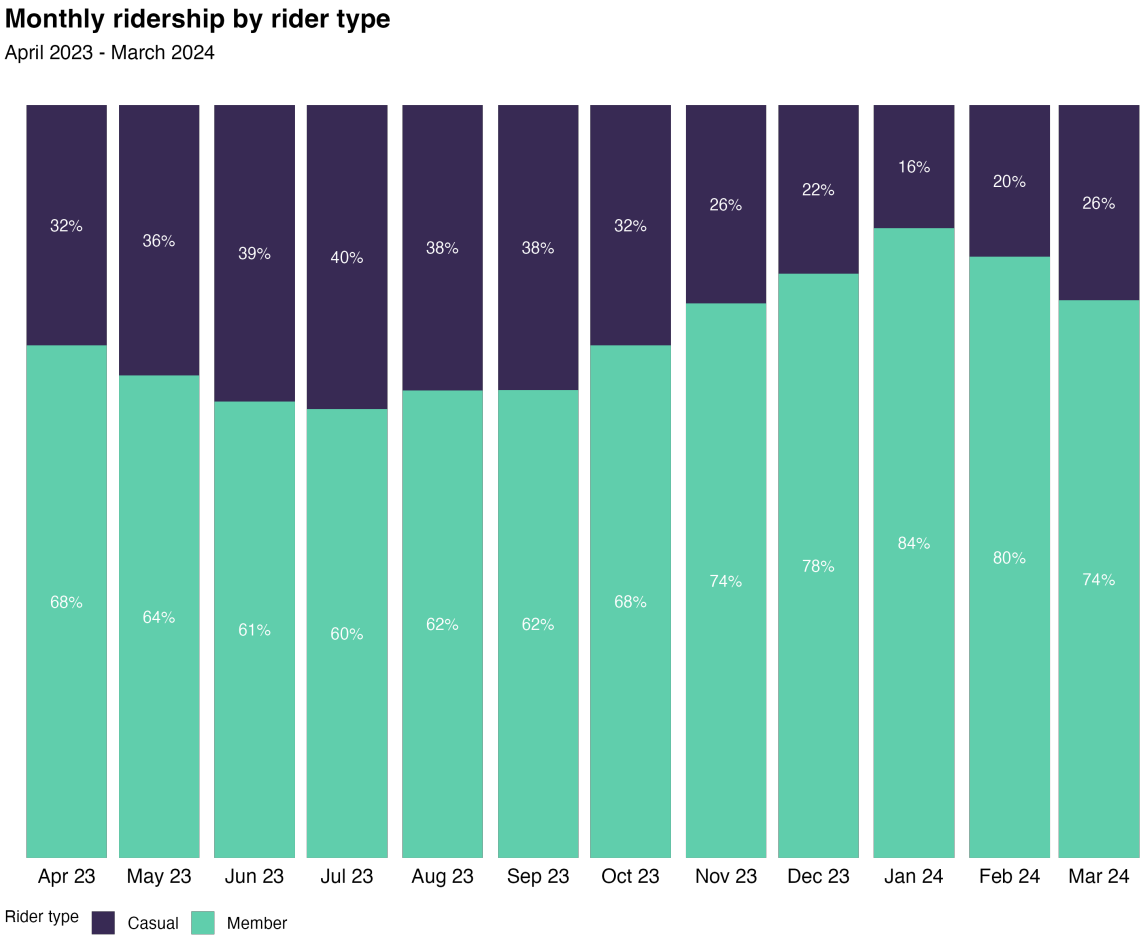
The analysis began by examining the ridership patterns of annual members and casual riders to identify differences in their usage behaviors. This involved exploring the distribution of rides between the two rider types across different time frames, including annual, monthly, daily, and hourly ridership trends.

Annual ridership To get a general understanding of the patterns of Cyclistic's bike-share riders, we first examined the proportion of rides each type of rider took during the twelve-month period under study. This analysis was visualized using a pie chart to illustrate the distribution of rides between annual members and casual riders.



The pie chart illustrates that annual members accounted for the majority of rides, representing 66% of the total rides for the twelve-month period. This distribution highlights the predominance of annual members in utilizing the Cyclistic system, indicating a higher level of engagement and consistent usage among this rider group. Although casual riders comprised a smaller proportion of rides at 34%, their presence is still significant, reflecting a substantial user base that contributes to the overall ridership.

Monthly ridership While the annual ridership distribution provides an overview of the ridership patterns, analyzing the monthly ridership trends can offer more detailed insights into how the two rider types engage with the bike-share system over time. By visualizing the monthly ridership for annual members and casual riders, we can identify seasonal variations and usage patterns that may influence marketing strategies and operational decisions.



The bar plot shows that from April 2023 to March 2024, members consistently outnumber casual riders, though the proportions vary by month. Throughout most of the warmer months, from April to October, there is a significant presence of casual riders, suggesting a higher activity level or preference for the bike-share system among this group during favorable weather conditions. Casual riders often accounted for 32% to 40% of the total rides during these months.

Conversely, from November to February, there is a noticeable decline in the proportion of rides taken by casual riders. Casual riders only accounted for 16% to 26% of the total rides during these months. This change highlights a substantial seasonal effect, with casual rider usage visibly reducing during the colder months.

By contrast, annual members consistently maintained a higher proportion of rides throughout the year, ranging from 60% to 84% of the total rides. This stable pattern suggests that annual members exhibit more consistent usage behavior, with less variability across different months compared to casual riders. The monthly ridership analysis provides valuable insights into the seasonal trends and preferences of Cyclistic's riders, offering a nuanced view of how different rider types engage with the bike-share system over time.

Daily ridership The monthly ridership analysis provided insights into seasonal variations in ridership patterns. To further explore the differences in usage behaviors between annual members and casual riders, we delved into the daily ridership trends. By examining the average number of rides taken by each rider type on different days of the week, we set out to identify the days with the highest average ridership for annual members and casual riders. This analysis helps pinpoint the days when each rider type is most active, providing valuable information for targeted marketing strategies and service planning.

To this end, we calculated the average daily ridership for annual members and casual riders and identified

the days of the week with the highest average number of rides for each rider group. The results are presented in the table below:

Day with the highest average number of rides

Mode

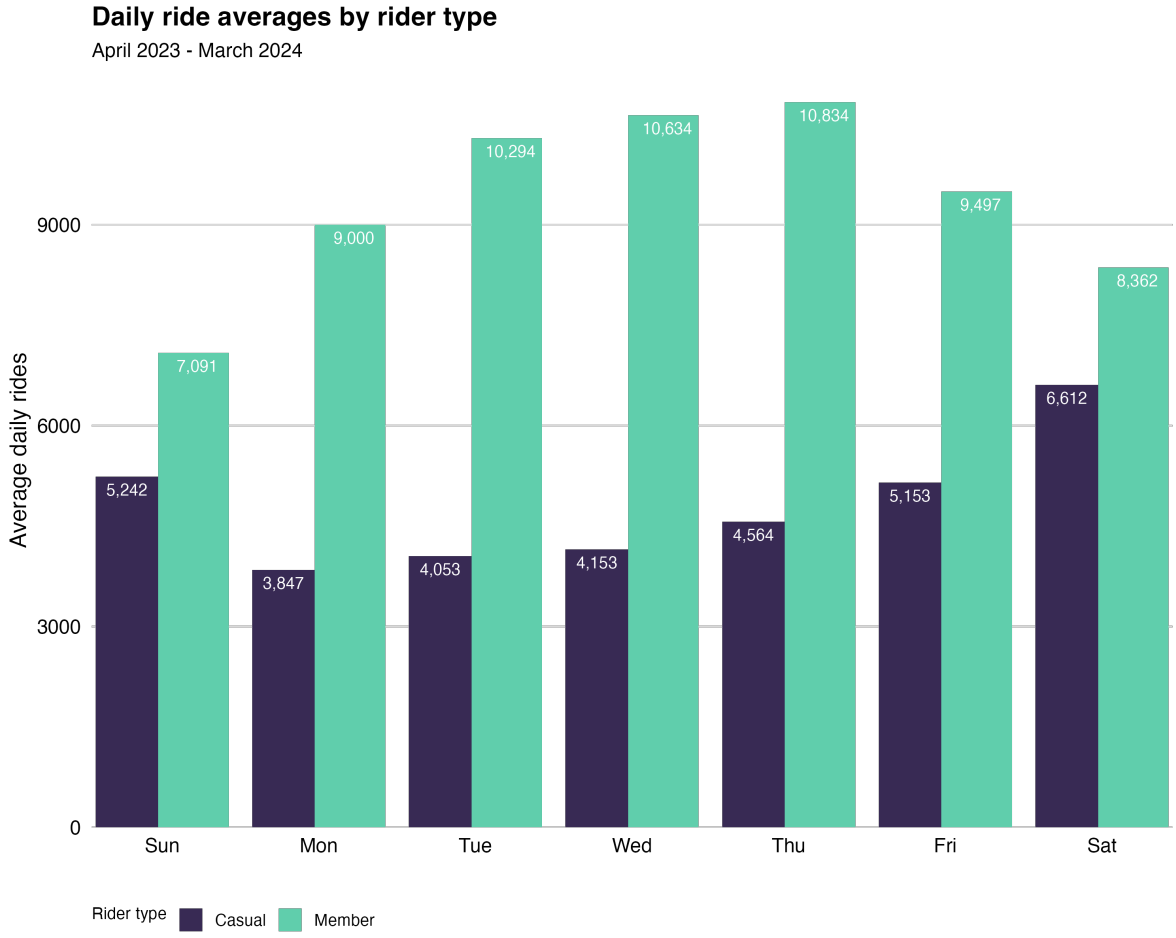
Annual members

Thu

Casual riders

Sat

Thursday was the day with the highest average number of rides for annual members, while Saturday was the peak day for casual riders. These findings indicate distinct usage patterns between the two rider types, with annual members showing a preference for weekday rides and casual riders favoring weekends for bike trips. To further visualize these daily ridership trends, we created a bar plot that illustrates the average daily rides for annual members and casual riders across each day of the week.



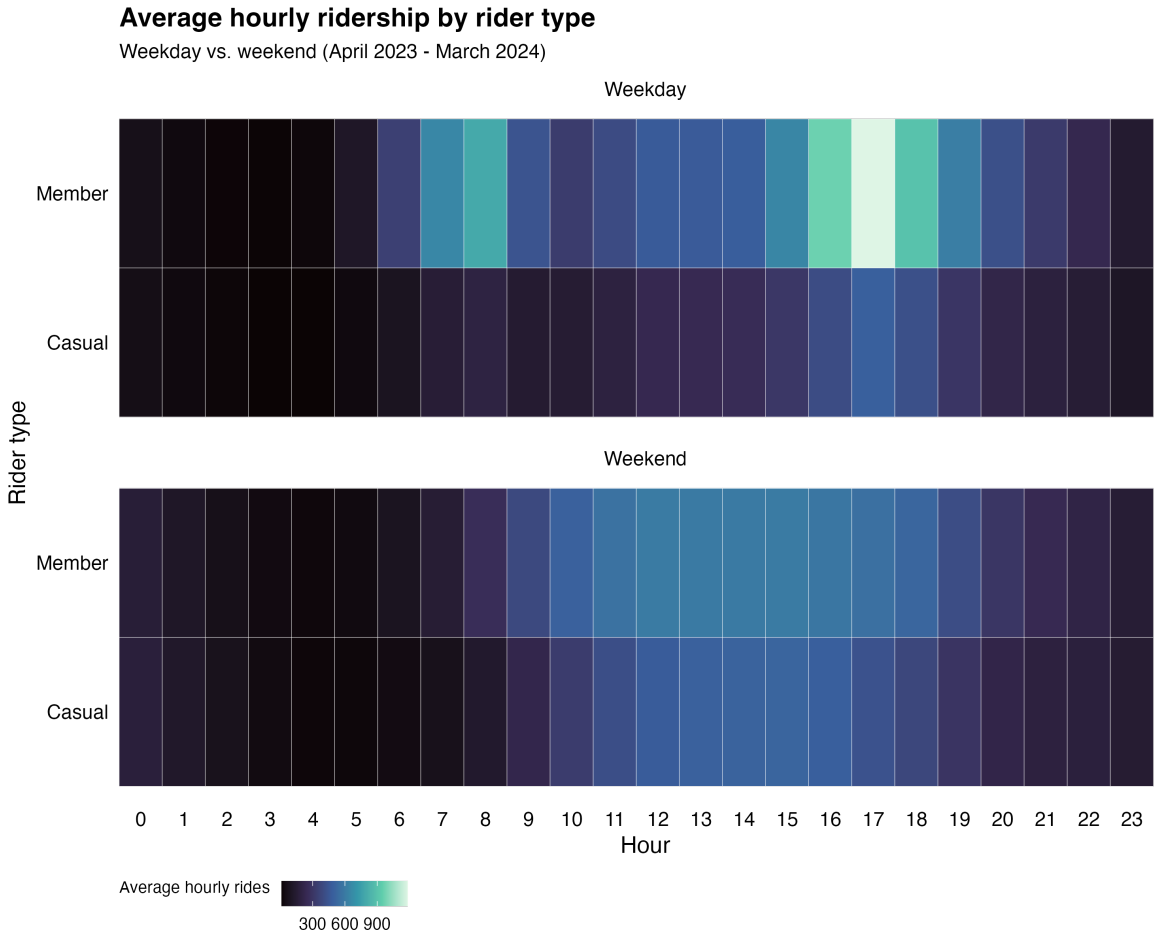
The bar plot revealed distinct ridership patterns between annual members and casual riders across different days of the week. Annual member usage remains robust throughout the week with slight variability but generally high levels. They averaged between 9,000 and 10,834 rides per day during the week, with a notable peak on Thursday. In contrast, casual riders exhibit a marked increase in rides over the weekend, particularly on Saturday. Casual ridership peaks on Saturdays with 6,612 rides, with high activity also observed on Sundays (5,242 rides) and Fridays (5,153 rides). Monday to Thursday sees comparatively lower activity,

ranging from 3,847 to 4,564 rides. This pattern suggests that casual riders prefer weekends for bike trips, likely for leisure or recreational purposes while annual members maintain consistent usage throughout the week, likely driven by commuting habits.

On weekdays, members averaged almost 2.5 times the number of rides compared to casual riders. This gap narrowed on weekends, with members still maintaining a higher ridership level but with a smaller margin. The distinct ridership patterns between members and casual riders suggest varying usage behaviors and preferences that can inform targeted marketing strategies and service planning initiatives.

Hourly ridership The daily ridership analysis provided insights into the average number of rides taken by annual members and casual riders on different days of the week. To further explore the temporal patterns of bike usage, we investigated the hourly ridership trends for each rider type. By examining the hourly distribution of rides throughout the day, we aimed to identify peak usage hours and differences in riding behaviors between annual members and casual riders. This analysis can help Cyclistic optimize service offerings and operational strategies to meet the distinct needs of each rider group.

To accomplish this, we calculated the average number of rides taken by annual members and casual riders during each hour of the day. The results were visualized using heatmaps to illustrate the hourly ridership patterns for weekdays and weekends, highlighting the intensity of bike usage across different hours. The heatmaps provide a clear visualization of the peak usage times for each rider type, enabling a detailed examination of the temporal dynamics of bike-sharing activity.



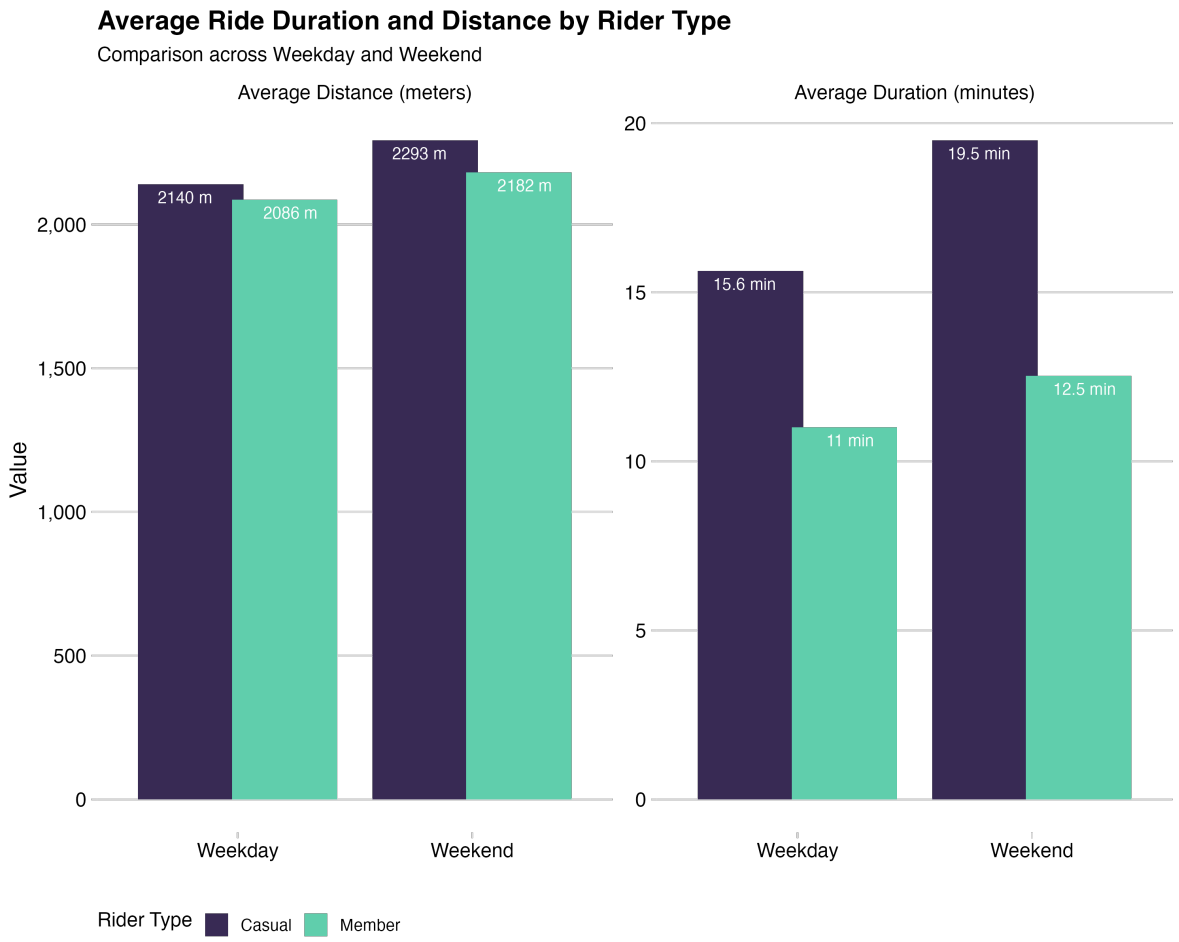
The heatmaps reveal distinct hourly ridership patterns between annual members and casual riders on weekdays and weekends. The top heatmap, representing weekday ridership, shows that annual members have consistent bike usage throughout the day, with two distinct peaks observed during typical commuting times in the

morning (06:00 to 10:00) and evening (15:00 to 20:00). This pattern aligns with the expectation that annual members primarily use the bike-share system for commuting purposes, resulting in higher ridership during weekday mornings and evenings. In contrast, casual riders exhibit much less ride activity during weekday mornings and afternoons. While they do display some activity, their peak usage hours are concentrated in the late afternoon and early evening (16:00 to 19:00), and it is significantly lower than that of annual members. This pattern suggests that casual riders are more likely to use the bike-share system for leisure or recreational purposes during the late afternoon and early evening hours. Both rider groups show a decline in activity during late night and early morning hours, indicating minimal bike usage during these times.

The bottom heatmap, representing weekend ridership, displays a different usage pattern compared to weekdays. Annual members maintain consistent bike usage throughout the day on weekends, from 08:00 until 22:00, with the most rides occurring between 11:00 and 18:00. This pattern suggests that annual members engage in more leisurely rides on weekends, possibly for recreational purposes or errands. Casual riders also exhibit steady bike usage throughout the day on weekends, albeit with a slightly shorter active period from 09:00 to 21:00, with peak usage between 12:00 and 17:00. The heatmaps provide a detailed view of the hourly ridership patterns for annual members and casual riders, highlighting the temporal dynamics of bike-sharing activity and the distinct usage behaviors between the two rider types. Again, there was very little activity in the late night and early morning for both groups, but the period of lower activity was shorter on weekends (01:00 to 07:00) compared to weekdays (00:00 to 05:00).

Ride characteristics

Ride distance and duration We next examined how ride distances and durations vary between Cyclistic's casual and member riders. By comparing the average ride distances and durations for each rider type, we aimed to uncover insights into the typical ride lengths and durations preferred by members and casual riders.



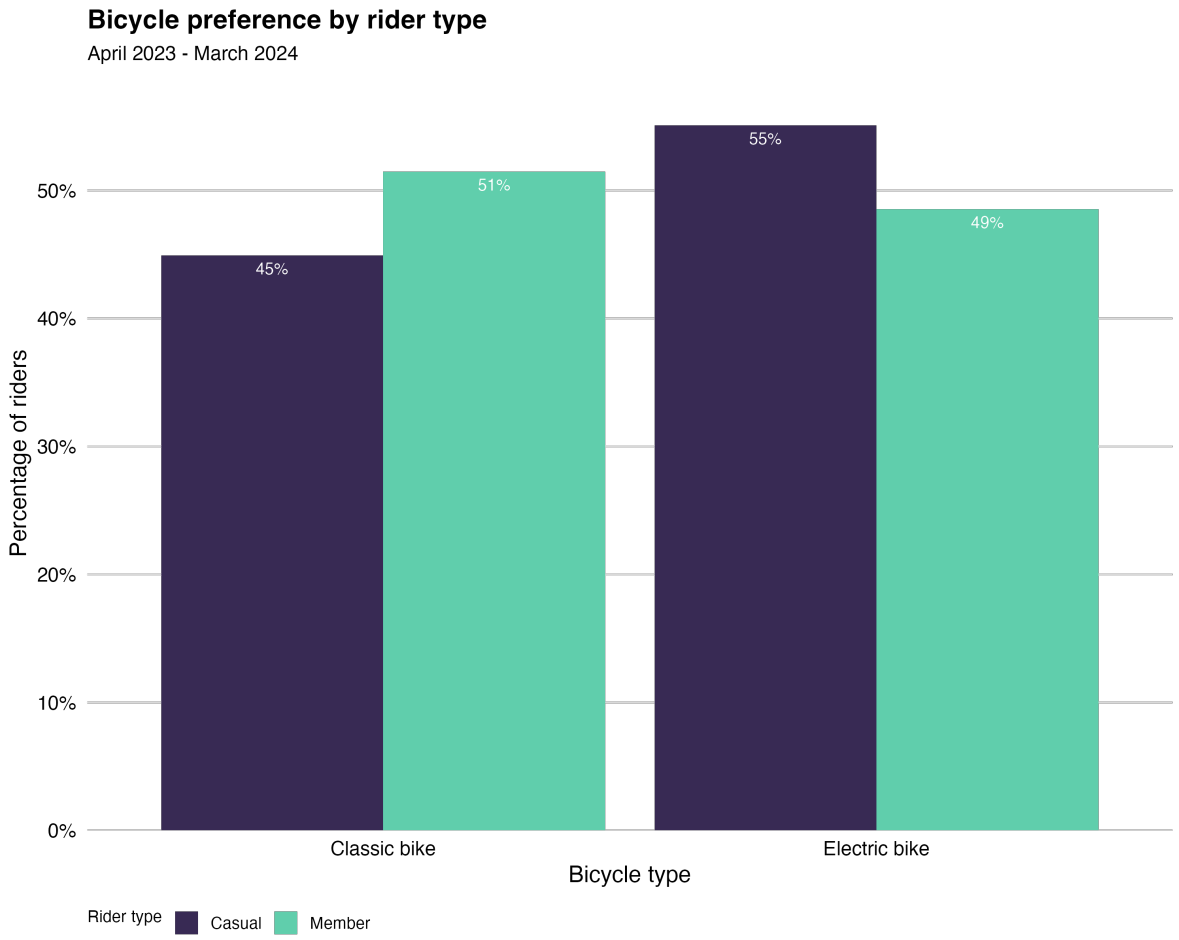
Ride distance The ride distance plot illustrates that casual riders tend to travel longer distances than annual members, with an average ride distance of approximately 2,140 meters on weekdays and 2,290 meters on weekends. In contrast, annual members have slightly shorter average ride distances, with about 2,086 meters on weekdays and 2,182 meters on weekends. This pattern suggests that casual riders consistently travel longer distances than annual members, indicating a preference for more extended rides, which might reflect their more leisure-based use of the bike-share system and a possible preference for longer exploratory trips.

This analysis provides a nuanced view of how ride distances vary among different rider types and days, offering strategic insights that can help Cyclistic tailor its service offerings and marketing approaches to better cater to the distinct preferences of casual and member riders.

Ride duration The bar plot for ride duration shows that casual riders spend significantly more time riding than member riders. On weekdays, casual riders spend about 42% more time riding than members (15.6 minutes vs. 11 minutes). This difference is even more pronounced on weekends, with casual riders spending about 56% more time riding than members (19.5 minutes vs. 12.5 minutes). Both rider types spend more time riding on weekends compared to weekdays, which could be due to more leisure time available. The increase in ride duration from weekdays to weekends is more pronounced for casual riders (25% increase) compared to member riders (14% increase). This suggests that casual riders are more likely to take longer rides than members, especially on weekends, indicating a preference for extended bike trips for leisure or recreational purposes.

Rider preferences

Bike type preference We examined the preferences for bike types among Cyclistic's casual and member riders to better understand how different types of bikes are used across these rider segments.



The bar plot illustrates that casual riders show a notable preference for electric bikes, with 55% opting for electric bikes compared to 45% for classic bikes. This indicates that casual riders favor the convenience and ease of use that electric bikes offer, especially since they tend to take longer rides than members. In contrast, members display a more balanced preference between bike types, slightly favoring classic bikes (51%) over electric bikes (49%). This suggests that members are comfortable using both types of bikes for their varied needs, including commuting and leisure.

Findings and Recommendations

Summary of findings

Annual members of Cyclistic demonstrate a consistent and engagement-driven usage pattern. They account for 66% of the total rides over a twelve-month period, highlighting their predominant role in the system. Their ridership remains stable throughout the year, ranging from 60% to 84% of monthly rides, with minimal seasonal variation. Members show a steady daily ridership, peaking on Thursdays and maintaining high levels on weekdays, averaging between 9,000 and 10,834 rides per day. This indicates a primary use for commuting, reflected in their peak hours during typical morning and evening commuting times (06:00-10:00 and 15:00-20:00). Their average ride distance is slightly shorter, around 2,086 to 2,182 meters, and their ride duration is consistent, at approximately 11 to 12.5 minutes. Members have a balanced preference for bike types, slightly favoring classic bikes (51%) over electric bikes (49%).

In contrast, casual riders exhibit more variable and seasonally influenced usage patterns. They make up 34% of the total rides, with a significant presence during the warmer months, where they account for 32% to 40% of the monthly rides. Their ridership drops to 16% to 26% during colder months. Casual riders prefer weekends, with peak ridership on Saturdays, averaging 6,612 rides, and showing higher activity on Sundays and Fridays. Their peak usage hours on weekdays are in the late afternoon and early evening (16:00-19:00), indicating a leisure-based use. Casual riders travel longer distances, averaging 2,140 to 2,290 meters, and spend more time riding, with durations of 15.6 minutes on weekdays and 19.5 minutes on weekends. They show a notable preference for electric bikes, with 55% opting for them, highlighting their preference for ease of use and convenience during longer, recreational rides.
Comparing the two groups, annual members are more consistent and commuting-oriented, whereas casual riders are influenced by season and leisure activities, preferring longer and more leisurely rides. This contrast in behavior and preferences offers valuable insights for tailoring marketing strategies and operational plans to meet the distinct needs of each rider segment.

Recommendations

Based on the findings from the data analysis, the following recommendations are proposed to optimize Cyclistic's bike-share system and encourage casual riders to convert to annual memberships:

1. Increase maximum ride duration for members during off-peak times:

- The current maximum ride duration for members is 45 minutes, compared to 180 minutes for casual riders. Given the difference in the way the two groups appear to use the system, increasing the maximum ride duration limit for members during non-peak hours could encourage longer rides and accommodate leisurely trips. This adjustment could enhance the user experience for members and potentially attract casual riders looking for extended rides. By offering more flexibility in ride duration, Cyclistic can cater to the diverse needs of riders and encourage casual riders to consider becoming annual members.

2. Increase the number of electric bikes:

- Casual riders, who tend to take longer rides and prefer weekends for bike trips, show a strong preference for electric bikes. To meet this demand and enhance the user experience, Cyclistic should consider expanding its fleet of electric bikes. By increasing the availability of electric bikes, the system can better accommodate casual riders' preferences for convenience and ease of use, potentially encouraging more casual riders to convert to annual memberships. This strategic investment could lead to increased ridership and revenue for Cyclistic.

3. Implement a loyalty program:

- Develop a loyalty program that rewards casual riders for frequent use and offers incentives to become members. Implement a point-based system where casual riders earn points for each ride, which can be redeemed for discounts on memberships or extended ride times. Additionally, offer exclusive perks to members, such as priority access to new bikes, special events, or partnerships with local businesses for discounts and deals. Demonstrating the added value and exclusive benefits of membership can make the transition more appealing to casual riders who regularly use the bike-share system.

These recommendations, grounded in the insights from the data analysis, aim to enhance the user experience, meet the distinct needs of Cyclistic's rider segments, and drive the conversion of casual riders into loyal annual members.