# Case study: Cyclistic

## Case Background

In my role as a junior data analyst on Cyclistic's marketing team, I am responsible for comprehending the distinctions in how Cyclistic bikes are utilized by casual riders and annual members. Casual riders refer to customers who purchase single-ride or full-day passes, while annual members subscribe to yearly plans for unlimited biking access. The marketing director posits that the company's future success hinges on increasing the number of yearly memberships by converting casual riders into loyal annual members. Subject to executive approval, my team will develop a fresh marketing strategy to pursue this concept.

The primary objective of this project is to unearth and communicate actionable insights that will guide Cyclistic's decision-making in crafting their new marketing approach

## Scope of Work

| Deliverables | Activities |
|---|---|
| Define the project | • Gather business information • Understaned the situation presented and the product(s) involved • Define goals |
| Prepare the data for analysis | • Select, extract and import data • Determine the quality of the data based on several parameters, besides verifying its integrity • Initial operations with data - filtering and sorting |
| Data processing (pre-analysis) | • Check the dataset for dirty and inconsistent data • Identify potential problems and/or bias in the data |
| Data transformation/cleaning | • Remove inconsistencies in the data • Check for bias and representativeness of data • Apply data manipulation tools |
| (Descriptive) Analysis, visualization and identifying relationships within the data | • Organize and aggregate data • Identify trends and relationships in the data • Perform calculations in the data • Provide insights on the analysed material • build visualizations for the data |
| Share results with stakeholders | • Create a comprehensive presentation that emcompasses key findings • Communicate results |
| Act | • Provide recommendations for the marketing campaign |

### Business Information

- Main revenue source and product: bike-sharing service across Chicago with geotracking and network-locked bikes features.
  - Product details and metrics:
    - 5,824 bicycles and 692 docking stations
    - More than 50% of riders select traditional bikes
    - 8% of riders use the assistive bike options
    - 30% of users use bikes to commute to work daily
    - Casual riders commonly choose Cyclistic for mobility conditions
    - Leisure is the main reason for biking among users
- Customer types: members and casual riders
  - Also acts as a revenue diversification strategy.
- Competitive advantages:
  - Pricing flexibility
  - Different types of customers

#### Stakeholders and Goals

| **Stakeholders** | **Expectations** | **Project and business Goals** |
|---|---|---|
| Cyclistic Executive Team | Relevant insights to inform data-driven decision-making | Implementing strategic initiatives to promote business growth |
| Lily Moreno (Director of | Evidence to support her theory and | Convert casual riders into annual members |

| | | |
|---|---|---|
| Marketing) | recommendations on marketing | |
| Marketing analytics team | Define the differences between casual riders and annual members | Generate data-driven and actionable results to enable more efficient business decisions |

### The Problem

In the pursuit of company growth, Cyclistic faces an uncertain future, necessitating a departure from its reliance on traditional marketing strategies that aimed to raise general awareness and cater to various customer needs. To capitalize on the lucrative profit margins of annual subscribers, the director of marketing proposes targeting existing casual customers and encouraging their transition into yearly subscribers. A well-executed marketing campaign based on this strategy holds the potential to generate more sustainable long-term revenue. To determine its plausibility, we must conduct an in-depth analysis of how and why Cyclistic casual bikers and members differ. This analysis will provide valuable insights to weigh the evidence, identify opportunities, and address any potential barriers for future marketing endeavors.

## Preparing the data for analysis and assessing it

### Data source

The dataset used for this analysis was made available by Motivate International Inc. under the following license.

The data is made up of a repository made up mostly of quantitative measurements collected over periods of time. Each data point represents a single bike trip from one docking station to the next. Furthermore, it is valid to note that this data provides information about what different customer types may do differently, but not the reasons.

Finally, the data analyzed encompasses the operational period of 2022.

### Data Quality

| Aspect considered | Issues detected? | Details |
|---|---|---|
| Reliability | Yes | • Some inconsistent dates • 1-to-1 relationship between stations and ids sometimes violated • Duplicates |
| Originality | No | First-party data |
| Comprehensiveness | No | --- |
| Current | No | The data refers to a time period close enough to the date of this analysis |
| Vetted | No | In general, data records are accurate and free of major errors. |

### Ethical aspects

The data ethics is guaranteed by several factors, including:

- The data is managed by employees and secured by a trusted global cloud services provider
- The dataset is open to the public
- Licensing is properly maintained
- The data comes from a credible organization

### Data Integrity

Some inconsistencies exist and were already reported, but they do not compromise the overall quality and usefulness of the data.

Based on that, this study assumes the integrity and unbiased nature of the data from now on.

### Code for the preparation stage

```
{r}
#imported Dataframes: D1,D2,D3,D4

#load libraries
library(tidyverse)
library(dplyr)
library(lubridate)
library(ggplot2)
```

```
#import data
D1 <- read_csv("Divvy_Trips_2019_Q2.csv")
D2 <- read_csv("Divvy_Trips_2019_Q3.csv")
D3 <- read_csv("Divvy_Trips_2019_Q4.csv")
D4 <- read_csv("Divvy_Trips_2020_Q1.csv")

#merge complementary  dfs
merged_df <- rbind(D1, D3, D4)

#extract basic information from database in order to asses content and usability of fields
head(merged_df)

colnames(merged_df)

glimpse(merged_df)

nrow(merged_df)

initial_metrics<- summary(merged_df)
View(initial_metrics)

head(D2)

glimpse(D2)

#determine main dfs` size
size_bytes <- object.size(merged_df)
print(size_bytes)
```

# Data processing (pre-analysis)

## Tools

The size output in the preparation step is 553.60 MB, which can be easily handled by RAM and the R language (which was used for processing and analysis) while preserving processing speed. If more performance-intensive approaches are needed, SQL is reserved as an option.

## Data processing/identifying issues

```
{r}
#check data type
# Use sapply() with class() to get the data types of each column
column_types <- sapply(merged_df, class)

# Print the result
print(column_types)
```

```
It is possible to assess that some fields could have better data types. In a future moment, we may want to make some typecasting.
```

```
Then, it is possible to deal with duplicates by analysing the number of unique entries in each column:
```

```
{r}
#get how many unique entries per column
unique_counts <- sapply(merged_df, function(x) n_distinct(x))
# Print the result
print(unique_counts)
```

It is noticeable that:

- "ride_id" has duplicates since the number of unique values does not conform to the data frame's size. This can be problematic since this field may be an intended primary key for the table.
- There are more station ids than station names, so there are station names that are associated with multiple station IDs

Furthermore, the following code helps us identify the nature of the correlation between stations and duplicated IDs:

```
{r}
duplicate_ids <- merged_df[duplicated(merged_df$to_station_id) | duplicated(merged_df$to_station_id, fromLast = TRUE), ]
```

```
print(duplicate_ids)
```

From the results and similarities between instances of stations with the same ID, it seems this problem comes mostly from the series used (stations changing place and streets have different names on the dataset, for instance). Therefore, without much harm to the quality of the analysis, it is possible to consider (at least mostly) of the entries as valid.

Additionally, for trips, we have:

```
{r}
# checking if there are trips with the same id
records_with_same_id <- merged_df %>%
  group_by(trip_id) %>%
  filter(n() > 1)
View(records_with_same_id)
```

The results tibble in empty, which ends the case for duplicate trips.

We can also target records that contain inconsistent dates (that is, the end time is impossible, considering the start time)

```
{r}
#check if there are date inconsistencies:
inconsistent_dates <- merged_df %>%
  filter(end_time< start_time)
View(inconsistent_dates)
```

The inconsistent records found will be addressed soon.

# Data transformation and cleaning

Starting by the duplicates and based on our discussing on their nature, the following code solves the problems with duplicates:

```
{r}
duplicates <- merged_df %>%
  group_by_all() %>%
  mutate(duplicate_count = n()) %>%
  filter(duplicate_count > 1)

# Print the duplicates along with their counts
print(duplicates)

# Eliminate duplicates from the original table and store it in a new data frame
merged_df_unique <- merged_df %>%
  distinct()

# Print the modified table without duplicates
print(unique_table)
```

Afterward, it is possible to address the records that contain inconsistent dates:

```
{r}
# Perform the anti-join to subtract inconsistent_dates dataframe from our main dataframe using all columns as keys
result_df <- anti_join(merged_df_unique, inconsistent_dates, by = ".")
```

Next, we can check and transform data related to the ids (bike_id, and station ids). The following code handles the task of checking for ids that do not have a numeric data type and converts them into the correct type (to secure the integrity of the datatype). By analyzing samples of the data, it is also possible to conclude that this operation is possible and handles the nature of the data (that is, the conversion works because the non-numeric fields "look like" numeric values):

```
{r}
#fix ids data types
inconsistent_ids <- result_df[!grepl("^\\d+$", as.character(result_df$trip_id)), ]

# Step 2: Stop and display any inconsistent ID records found
if (nrow(inconsistent_ids) > 0) {
  print("Inconsistent ID data types found:")
  View(inconsistent_ids)
}
```

```
# Step 3: Convert the ID column to numeric
result_df$trip_id <- as.numeric(as.character(result_df$trip_id))
```

The procedure above was repeated with start and end station ids.

It is also possible to asses the case of missing values across the data:

```
{r}
missing_counts <- colSums(is.na(result_df))
View(missing_counts)
```

There is no relevant presence of missing values, but let`s create a table with no missing values on important fields in case it is needed in the future:

```
{r}
columns_with_missing_values <- c(from_station_id, to_station_id)

# Remove records with missing values on the specified columns
cleaned_table <- my_table[complete.cases(my_table[, columns_with_missing_values]), ]
```

The station ids were considered the "important fields" because the absence of their value could indicate that the corresponding records do not actually represent a real trip.

It is also important to alter the table in order to keep it consistent with the lables used after 2020 by the company. In that sense, we can use the following code:

```
{r}
# Reassign to the desired values (2020 standards)
result_df <-  result_df %>%
  mutate(member_casual = recode(member_casual
                           ,"Subscriber" = "member"
                           ,"Customer" = "casual"))

# Check to ensure if all necessary transformation were made
table(all_trips$member_casual)
```

## Adding additional columns for analysis

In order to increase our capacity of analyzing the data, some additional columns will be created:

- "ride_length_time": length (time) of each ride
    - process:

      ```
      {r}
      result_df$ride_length <- difftime(result_df$end_time,result_df$start_time)#

      #Convert ride_length from Factor to numeric
      is.factor(result_df$ride_length)
      all_trips$ride_length <- as.numeric(as.character(result_df$ride_length))
      is.numeric(result_df$ride_length)
      ```

- "start_day_week": day of the week when the bike was picked up
    - process:

      ```
      {r}
      result_df$day_of_week_start <- format(as.Date(result_df$start_time), "%A")
      ```

- "end_day_week": the day of the week when the bike was dropped off
    - process:

      ```
      {r}
      result_df$day_of_week_end <- format(as.Date(result_df$end_time), "%A")
      ```

- columns for more specific time stamps
    - Allows for a more precise and localized, besides less granular, analysis
    - process:

    ```
    {r}
    # Convert the date_time to a Date format
    result_df$date_time <- as.Date(result_df$date_time)

    # Create new columns for year, month, and day
    result_df$year <- format(result_df$date_time, "%Y")
    result_df$month <- format(result_df$date_time, "%m")
    result_df$day <- format(result_df$date_time, "%d")
    ```

# Descriptive analysis

Now, we`ll perform data aggregation and statistical analysis of our data in order to derive insights.

First, let´s work with ride_lenght:

```
{r}
general_mean <-mean(result_df$ride_lenght)

general_median <-median(result_df$ride_lenght)

max <-max(result_df$ride_lenght) #longest ride

min<-min(result_df$ride_lenght) #shortest ride
```

Althtough the code above can provide us some information, a comparative analysis between the two types of customer is our primary goal, so let´s be more specific:

```
{r}
aggregate(merged_df$ride_length ~ merged_df$usertype, FUN = mean)

aggregate(merged_df$ride_length ~ merged_df$usertype, FUN = median)

aggregate(merged_df$ride_length ~ merged_df$usertype, FUN = max)

aggregate(merged_df$ride_length ~ merged_df$usertype, FUN = min)

#
# casual        60.57387 mins
# member         14.44182 mins
```

Analyzing the results, it is possible to see that, on average, casuals spend 4.2 times more time on each ride. This might suggest bigger tracks, however, when the distances traveled are compared, they are not very different. It is not possible to provide a definitive answer for this, given that we do not have information on what happens during the ride. As casuals usually ride for fun, they might not prioritize speed on their trips, differently from a substancial portion of the members, who commute or perform focused activities (which could explain the shorter trips). Nevertheless, this is only a hypothesis that cannot be properly tested with the present resources and data.

Now, it is suitable to analyse the distributions of riders over the days of the week for each group of customer:

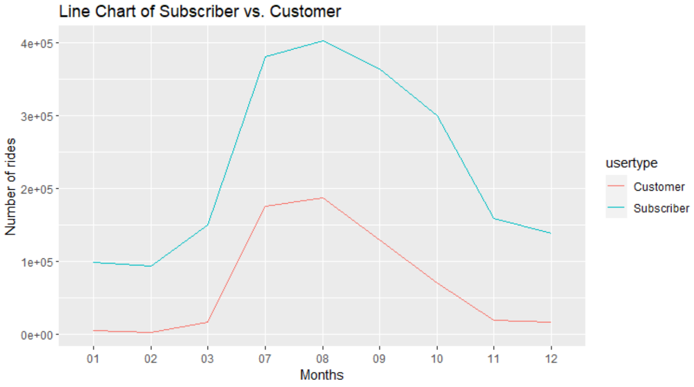
```
{r}
#Get the day of the week for each date

result_df$date_column <- as.Date(result_df$date_column)
result_df$day_of_week <- weekdays(result_df$date_column)
result_df$day_of_week <- ordered(result_df$day_of_week, levels=c("Sunday", "Monday", "Tuesday", "Wednesday", "Thursday", "Friday", "Sa

#process data and group data items by weekday and usertype
result_df <- result_df %>%
  mutate(weekday = wday(start_time, label = TRUE)) %>%
  filter(member_casual == "Customer") %>%
  group_by(member_casual, weekday) %>%
  summarise(number_of_rides = n(), average_duration = mean(ride_length)) %>%
  arrange(member_casual, weekday)

#create visualizations
result_df <- result_df %>%
```

```
    mutate(weekday = wday(start_time, label = TRUE)) %>%
    filter(member_casual == "Customer") %>%
    group_by(member_casual, weekday) %>%
    summarise(number_of_rides = n(), average_duration = mean(ride_length)) %>%
    arrange(member_casual, weekday)  %>%
    ggplot(aes(x = weekday, y = number_of_rides, fill = member_casual)) +
    geom_col(position = "dodge")
```

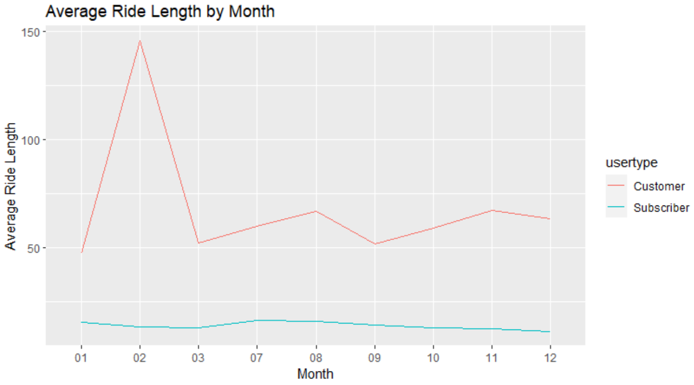
The results indicate that casual riders tend to have greater activity on the weekends, whereas members have a more uniform distribution. The main motivation for the trip for the different groups (leisure for the first, need/commuting for the second) is the most likely explanation.

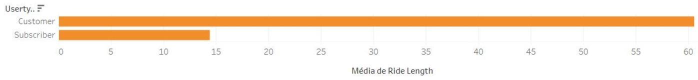
# Visualization and identifying relationships within the data

## Data visualization



- Makes explicit the diminished activities during colder months for both customer types





## Identifying trends and relationships

### Shared by casual riders and members:

- Less activity in winter (probably due to colder wheater)
- Similar distances traveled
- Similar patterns of usage on weekends, with peaks in the afternoon. In that scenario, both customer types may use Cyclistic for similar purposes.

### Casuals:

- Take longer trips, in general throughout the entire year. Due to their leisure motivation, speed is probably not a concern.
- Peak of activity during the Summer.
- Activity diminishes expressively during colder months (Winter)

### Members:

- Have small deviation on their ride lengths (time duration) throughout the year
- Peak of activity during rush hours and encompassing more periods of the day

## key findings

After conducting the analysis and determining some behavioral differences between the two types of members considered, it is not possible two make a definitive conclusion. Although both customers were better defined, what is already a more advanced situation, more data analysis would be necessary to actually reach a description of typical customers.

In that sense, it is possible to connect the main motivation of each type of customer and the data analyzed in order to draw some hypotheses, but many of them cannot be proved with the current resources.

For instance, casual riders, who generally use Cyclistic for leisure, are less active on weekdays, have longer trips, and have less consistent trips on different time series. They probably do not use the bicycles enough to get a membership.

Members, on the other hand, have more consistent and frequent rides, with peaks on rush hours. In that sense, it is economically worth it to buy a membership.

## Recommended actions

- Promote more data collection and analysis efforts
    - In order to more strongly define each type of customer and their characteristics, more data collection regarding a representative sample (to avoid bias) of users. More specifically, these efforts could focus on qualitative data, confirming the customers´ motivations behind their patterns and a description of what, qualitatively, happens during riders.
- Update approaches for conversion
    - Given the distribution of riders in different days of the week, more options for memberships could be considered. For instance, more localized membership fees may convert a bigger portion of current casual members by enabling them to purchase, for the sake of exemplification, weekend or Summer memberships.
    - Given that the less frequent usage of bicycles by casual riders would hardly motivate an upgrade to a membership, a loyalty and/or exclusive benefits program could be implemented. Within this system, for instance, members could have access to more advanced and exclusive vehicles, or the feature to schedule riders beforehand. Provided that such implementation would differ more drastically, a phased rollout strategy could be applied.