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

# Summary

This project is a case study of “Cyclistic”, a fictional bike-share company. The goal behind this case study is to use the data analysis process as a tool to answer a business question.

The context is as follows. According to Cyclistic’s finance analysts have concluded that annual members are much more profitable than casual riders. For that reason, the director of marketing believes the company’s future success depends on maximizing the number of annual memberships. Therefore, your team wants to understand how casual riders and annual members use Cyclistic bikes differently. From these insights, your team will design a new marketing strategy to convert casual riders into annual members. The director of marketing has assigned you the following question to answer: “**How do annual members and casual riders use Cyclistic bikes differently?**”

# Step 1: Ask

Being the first step of the data analysis process, the Ask step is all about framing the project, in order to make sure that our work is concise and straight to the point.

First, we must **identify the stakeholders** of our project, which means the people with a vested interest in our project. The key stakeholders are the company’s executive team since they’re the ones that need to make a decision based on the data we give them. As for secondary stakeholders are the marketing director and the rest of the marketing analytics team that we’ll be working with to form recommendations for the marketing strategy to be adopted.

After that, we move to **identifying the business task**. In other words we need to define the questions we want answered through our analysis to meet the needs of our stakeholders. In this case, our stakeholders need to convert casual customers to annual members. To do that, we must first understand how both of these segments behave and interact with the product, and this involves finding answers to a number of questions. The question we were tasked to answers is **“How do annual members and casual riders use Cyclistic bikes differently?”**

Then we decide what type of data we need to answer it.

# Step 2: Prepare

In this step, we are more concerned with the collection and handling of data.

First, we are beginning with the **collection**. According to the context of the case, since the data we need is related to customer use, that means we can just get the data using the resources of the company itself and no need for second or third party intervention. This method of collection is the best possible since it supplies us with data that’s more reliable, the original source is easily accessible and verifiable, comprehensive, and current.

In reality, the actual raw data can be obtained through the following link: <https://divvy-tripdata.s3.amazonaws.com/index.html>

The handling of the data covers multiple aspects. Starting with the storage, which will be handled using a MySQL local server, while also keeping a copy of the raw data to act as a backup. Initially the data is organized into smaller datasets divided by month, and we made sure that each of them has a similar structure.

As for the security, the access to the server is using a password. And on top of that all personally identifiable data like names, addresses, and billing information… was scrubbed from the dataset.

The license for this dataset, we are allowed non-exclusive, royalty-free, limited, perpetual permission to access, reproduce, analyze, copy, modify, distribute in your product or service and use the Data for any lawful purpose.

# Step 3: Process

This is the step where we clean the data from all anomalies that may influence the accuracy of our data and insure its integrity and usefulness. For this, we will use a combination of Microsoft Excel and **MySQL** for different steps of the process. We will do the cleaning process for each file, using the exact same procedures.

The types of dirty data we screen for are:

* Data missing important fields
* Any fields containing the wrong values
* Any cells with extra spaces
* Columns with the wrong data type
* Duplicate rows
* Any data that uses different formats to represent the same thing

## SQL

When it comes to working with large amounts of data, SQL is the tool for the job. It offers multiple tools that can help streamline the data cleaning process. For this project, we will use MySQL to upload the data into our database and interact with it. The steps that were taken to clean the data are as follows:

* Check the different columns for NULLs and empty fields
* Check for duplicates in both some specific columns and also for entire rows
* Check for wrong data based on the data type and the specifics of each column.

For actually the SQL query we used for this procedure, check the “cleaning process” SQL file.

Note that the cleaning process revealed that the data related to the different stations is too dirty to be cleaned effectively or used in its original state. Some of the problems noticed in the data are:

* Inconsistent naming leading to different variations of the same location
* Locations IDs that are associated with different places at once

Under normal circumstances this could be addressed with management to figure out a possible solution, but for this analysis, since we have no alternatives, we decided to remove the names, IDs, and longitude and latitude values related to stations and just focus on the other elements in our analysis.

# Step 4: Analyze

After cleaning the data, now it’s time for analysis. In this step, we try to detect trends within the data. For this purpose, we use the pivot tables feature on the spreadsheets program Google sheets.

After exporting the clean data from our database, we use certain sheets functions and formulas to produce the following calculated fields

* Duration: describes the duration of each trip by subtracting the ended\_at column from startrd\_at
* Start\_date: we used the LEFT function to split the started\_at field and obtain the date part. This would be useful to group trips by date.
* Day of the week: obtained using the WEEKDAY to helps determine levels of activity for each day of the week.

Next comes the pivot tables.

* The number of daily trips for each customer segment
* The number of daily trips for each bike type
* The number of trips per day of the week for each customer segment
* The number of trips per day of the week for each bike type
* The number of trips per bike type for each customer segment
* Average trip durations per bike type for each customer segment
* Median trip durations per bike type for each customer segment
* Max trip durations per bike type for each customer segment
* Percentage of trips over 24 hours.

These results are grouped in the extracted data.xlsx file.

# Step 5: Visualization

In this step, we use tableau to make a dashboard out of the data we extracted making it easy to spot the trends and determine the differences between casual users and members.

The dashboard we created is composed of the following elements:

* The total number of trips per category
* A line chart representing the number of trips per month for each customer segment.
* A heat map Level of activity for each day of the week
* Line charts both of the average and median of trip durations

The dashboard can be accessed using the following link: <https://public.tableau.com/app/profile/hamza.belhaj/viz/Bike-shareapp/Dashboard1?publish=yes>

# Insights

From the data we got using pivot tables and also the visualizations we built using tableau, we can draw the following conclusions:

* Judging by the change in the volume of activity throughout the months, can say that we have a seasonal activity, where it drops to its lowest during the winter then slowly picks up to reach its peak during summer. The probable cause for this variation is the weather, the warmer and clearer the weather, the more likely people are to use bikes. This correlation can be verified in more details by checking the daily usage data coupled with the weather data for that same time period.
* Generally, electric bikes are slightly more preferred than classic bikes, while docked bikes are significantly less used than both.
* Over all, casual users surpass members in terms of average, median and max when it comes to the durations of trips. But for the total number of uses, members use the service more than casuals.
* The heat map reveals that uses of the service are more spread out through the months for members (except for winter), while casual uses are more concentrated in the summer. Moreover, members use the bikes all week long, while casual users prefer the weekend.
* Docked bikes are only used by casual users, and sometimes for extended periods of times (continuously over days). The number of these outliers is significant enough to throw off the average trip durations, which explains the irregular distribution that we notice on the line chart.

# Recommendations

Based on all of the above, we recommend the following:

* Since the variation in the frequency of usage throughout the year indicates that bikes aren’t used all year around. So one possible reason for why people prefer to use the service casually over being a member is that they consider the yearlong subscription too much of a commitment, and they won’t get all the value out of what they’re paying. One possible solution to that is to diversify the offer and add a monthly/ per trimester subscription modes. Summer is also known as a month when people travel and visit new places, so another option that can be considered is a traveler pass can be useful for people visiting the city but won’t be using the service.
* Using more accurate location data, we could also determine which stations are frequented by which type of customers to help with bike availability and also find out the stations mostly frequented by casual users in order to deploy advertisement for the membership programs into.
* We can also use weather data to determine the correlation between usage and weather conditions, which will help us predict the volume of uses for each bike type based on weather forecast, which would help minimize the costs since we won’t have to deploy more bikes than we need to.