Cyclistic bike-share analysis case study

Author: Dawid Sosinski

Last Updated: 19/12/2021

# Problem Statement

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**.

In order to do that the marketing analyst team needs to better **understand how annual members and casual riders differ, why casual riders would buy a membership, and how digital media could affect their marketing tactics**. Moreno and her team are interested in analyzing the Cyclistic historical bike trip data to identify trends.

# Background

Until now, Cyclistic’s marketing strategy relied on building general awareness and appealing to broad consumer segments. One approach that helped make these things possible was the flexibility of its pricing plans: single-ride passes, full-day passes, and annual memberships. Customers who purchase **single-ride or full-day passes** are referred to as **casual riders**. Customers who purchase **annual memberships** are **Cyclistic members**.

Moreno believes that maximizing the number of annual members will be key to future growth. Rather than creating a marketing campaign that targets all-new customers, Moreno believes there is a very good chance to **convert casual riders into members**. She notes that casual riders are already aware of the Cyclistic program and have chosen Cyclistic for their mobility needs.

# Goal

**Design marketing strategy** aimed at converting casual riders into annual members.

An **annual member profile** will be created that will describe customers who decided to purchase annual membership. The profile will include:

* Customer gander
* Customer age
* Customer Area of residence (where did they start the journey)
* When (which week day of week, what time in a day) most likely customer take their trip

Then a comparison will be executed to check opportunity to find similar profile within casual customers

# Key Results

50% Increase of conversion from casual customers to annual members

# Deliverables

A **report** that will answer the following question “**How do annual members and casual riders use Cyclistic bikes differently?**”

Report will include:

1. A clear statement of the business task - Done
2. A description of all data sources used - Done
3. Documentation of any cleaning or manipulation of data - Done
4. A summary of your analysis
5. Supporting visualizations and key findings
6. Your top three recommendations based on your analysis

# Stakeholders

* **Cyclistic executive team**: The executive team will decide whether to approve the recommended marketing program.
  + Project sponsor
  + Final presentation addressee
* **Lily Moreno**: The director of marketing and your manager
  + Status update once a week
  + Highlight potential obstacles and risks
* **Cyclistic marketing analytics team**: A team of data analysts who are responsible for collecting, analyzing, and reporting data that helps guide Cyclistic marketing strategy
  + Daily updates (Morning stand-ups)
  + Ad Hoc meeting

# Timeline - Milestones

Log available in [here](https://docs.google.com/document/d/1bluH4PuHX8GEH0lwOfUJu38w2ptrOrX3Q9dnglqSiQg/edit#heading=h.usr6nsk3mn9n)

| **Milestone** | **Status** | **Last Updated Date** |
| --- | --- | --- |
| Project Start | Completed | 19/12/2021 |
| Project requirements | Completed | 27/12/2021 |
| Data Preparation | Completed | 05/01/2021 |
| Data Analysis | Completed | 12/01/2022 |
| Final Report |  |  |

# Data Processing

Cyclistic’s historical trip data to analyze and identify trends. [Download here](https://divvy-tripdata.s3.amazonaws.com/index.html).

*Note: The datasets have a different name because Cyclistic is a fictional company. For the purposes of this case study, the datasets are appropriate and will enable you to answer the business questions. The data has been made available by Motivate International Inc. under this license. This is public data that you can use to explore how different customer types are using Cyclistic bikes. But note that data-privacy issues prohibit you from using riders’ personally identifiable information. This means that you won’t be able to connect pass purchases to credit card numbers to determine if casual riders live in the Cyclistic service area or if they have purchased multiple single passes.*

## Data Assessment

Data comes from the website DivvyBikes.com. However, URL listed in the README files does not exist [http://DivvyBikes.com/data](http://divvybikes.com/data). Email address [data@DivvyBikes.co](mailto:data@DivvyBikes.co) listed in the README file seems to be valid. It was decided to verify the data by emailing [data@DivvyBikes.com](mailto:data@DivvyBikes.com) to clarify the data origin and source.

### Data Format

Date comes with **two distinct formats**:

#### Format 1 “Trips and Stations”

* Stored in divy\_stations\_trips folder
* Divided into Station Details file and Trip Details File
* Mostly CSV. One XLSX file
* Date range from 2013 until 2020
* 2.7 GB

Pros:

* Good amount of data
* Can calculate trip date and duration
* Can identify start and finish station
* Can distinct between casual and annual customer
* Can analyse customer gender and age

Cons:

* Inconsistent column names
* A lot of missing data in Gender and usertype
* Will require some effort to load the data due to inconsistent column naming

[Format 1 Trips and Stations](https://docs.google.com/document/d/1VZ5LVvLedHyr-f1F0LI8Uj2xMlms6XSyS0QfVOtRGoA/edit)

#### Format 2 “Trips Combined”

* Stored in divvy\_tripdata\_csv folder
* Single file format
* CSV only
* Date range from 2020 until 2021
* 1.7 GB

Pros:

* Good amount of data but less comparing to Format 1
* Can calculate trip date and duration
* Can identify start and finish station
* Can distinct between casual and annual customer

Cons:

* No data regarding customer gender and age
* Some missing data regained station and trp duration

[Format 2 Trips Combined](https://docs.google.com/document/d/1RznX2md1Ga0Tl9TDEAU3yHZ95PfbBzESXwp_5QRTdYE/edit)

### Conclusion:

As the main goal of this project is to design a new digital marketing campaign it is crucial to identify as much as possible about our customers. Any information about their gender and age will be a great value to marketing campaigns via digital media such as Google and Facebook.

**Format 1 “Trips and Stations” will be selected for further processing** **and analysis** as it includes additional metadata points about our costumes as well as start and finish stations,trip date and duration. It is the most complete data set considering stated requirements.

### Risks

It needs to be said that that data is not 100% complete in terms in Gender and Age. Additionally, inconsistent column naming will require more time to prepare and load the data for analysis.

## Data Cleaning

Executed by script merge\_clean\_csv\_files.R in [GitHub](https://github.com/DSosinski/CyclisticAnalysis)

1. Added file\_name column to specify source data file
2. Converted all dates to YYYY-MM-DD hh:mm:ss
3. Divvy\_Trips\_2014-Q3-07.cs - column name birthyear replaced with birthday
4. Divvy\_Trips\_2017\_Q1.csv - column name start\_time replaced with starttime
5. Divvy\_Trips\_2017\_Q1.csv - column name end\_time replaced with stoptime
6. Divvy\_Trips\_2018\_Q1.csv - header was replaced with standard column names
7. Divvy\_Trips\_2020\_Q1.csv is is Format 2 and has been removed from analysis
8. Changed column name starttime to start\_time
9. Changed column name stoptime to stop\_time
10. Changed column name bikeidto bike\_id
11. Changed column name tripduration to trip\_duration
12. Changed column name usertype to user\_type
13. Removed records without user\_type value
14. Removed records with user\_type equal to Dependent
15. Excluded any age (2021- birthday) older than 85
16. Excluded any age (2021- birthday) older qual less than 5
17. Exclude trips longer than 181 minutes
18. Due to 3.Gb of data, it was decided to use a parquet file (0.3Gb) to transfer data to AWS Athena.

*Note: Due to the large amount of data R is not able to process queries in a satisfactory time. It was decided to use AWS Athena to execute SQL analysis. Data import to Athena will be done by creating a parquet file and uploading it to AWS S3.*

## Data Analysis

Following analysis will be done regarding annual membership customers:

1. Analyze customer gender in relation to number trips.

*Customers identified as Male are the majority of annual member trips.*

1. Analyze customer age in relation to number trips.

*Males between age of 26 and 40 made the most amount of trips*

1. Analyze trip duration in relation to number trips.

*Trip durations between 1 min and 16 mins are the majority of annual member trips.*

1. Create an annual membership customer profile.

***Males between age of 26 and 40 that ride between 1 min and 16 mins*** *will be key factors for our* ***annual membership customer profile****.*

Following analysis will be done regarding casual customers:

1. Analyze the annual membership customer profile using casual rider data only

*There were 56240 trips made in 2019 by casual riders that share the same characteristics with the annual membership customer profile*

1. Create measurable Key Result as a major element for assessing success of the new marketing strategy

*Key Result 1:* ***26%*** *conversion from* ***casual male riders*** *to* ***annual members***

*Key Result 2:* ***16%*** *conversion from* ***all casual riders*** *to* ***annual members***

Log available [here](https://docs.google.com/spreadsheets/d/1E_69rG0A8eqUCOc5f398sVmsP3MRch0RGzmeX2VK-l8/edit#gid=1918858937)

# Tech Details

1. **Merge** multiple CSV trip data files into one main file using R
2. **Cleaning** will be done using R due to large amount of data
3. **Export** will be done via parquet format and uploaded to AWS S3 to be processed by Glue and Athena due to 3.7 Gib of raw data
4. **Analysis** will be executed via Athena SQL and stored in GSheet as spreadsheets
5. **Visualisation** will be created Tableau using exported data (Vis file stored in GitHub)
6. **Presentation** will be created a in Google Slides

Tools used:

* R
* AWS S3 Glue Athena (SQL)
* GSheet
* Tableau
* Google Slides
* GitHub

[GitHub Repository](https://github.com/DSosinski/CyclisticAnalysis)

# Final Document - Design

The final report will contain following slides:

1. Table of content
2. Why do we want to solve this problem
3. How are we going to achieve our goal
4. What do we need to do
5. How do we measure success
6. Next actions

# Related Documents

[Cyclistic bike-share analysis case study Timeline](https://docs.google.com/document/d/1bluH4PuHX8GEH0lwOfUJu38w2ptrOrX3Q9dnglqSiQg/edit#heading=h.usr6nsk3mn9n)

[Format 1 Trips and Stations](https://docs.google.com/document/d/1VZ5LVvLedHyr-f1F0LI8Uj2xMlms6XSyS0QfVOtRGoA/edit)

[Format 2 Trips Combined](https://docs.google.com/document/d/1RznX2md1Ga0Tl9TDEAU3yHZ95PfbBzESXwp_5QRTdYE/edit)

[GitHub Repository](https://github.com/DSosinski/CyclisticAnalysis)