

Google Data Analytics Capstone: Cyclistic Case Study

Introduction:

Cyclistic: A bike-share program that features more than 5,800 bicycles and 692 docking stations in Chicago.

The bikes are geotracked and can be unlocked from one station and can be returned to system at any station.

There are two types of cyclistic users:

1. Members who purchase an annual membership
2. Casual riders who purchase single ride passes and full day passes

Business Task:

Cyclistic's finance analysts have concluded that annual members are much more profitable than casual riders. So, it is believed that maximizing the number of annual members will be key to future growth. Rather than creating a marketing campaign that targets all-new customers, Company's manager believes that there is a very good chance to convert casual riders into members.

Company's goal is to design marketing strategies aimed at converting casual riders into annual members. In order to do that, the marketing analyst team needs to better understand how annual members and casual riders use cyclistic bikes differently and is interested in analyzing the Cyclistic historical bike trip data to identify trends.

As a Junior Data Analyst, the task is to answer the above-mentioned question:

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

Action:

1. Cyclistic data for year 2021 has been used for analysis and visualization purposes. [Cyclistic Trip Data](#).
2. The data has been collected, combined, explored, cleaned and manipulated using SQL in Big Query to create a final target table.
3. Finally, the dataset contained in the target table has been used in Tableau: A Business Intelligence Platform to create visualizations and identify trends.

All the steps which have been used in SQL are explained in this document below.

1. Data Collection

Link to SQL Queries: [Data Collection](#).

Monthly trip data is contained in individual excel file. The data for each month for the year 2021 has been downloaded and exported into Big Query. Furthermore, all the 12 tables were combined in a single table to contain 2021 data using SQL in Big Query.

2-Data Exploration

Link to SQL Queries: [Data Exploration](#).

Counting the total number of rows

The combined data for year 2021 contains 5595063 rows.

Row	total_rides
1	5595063

Checking the Data Types for all the columns

column_name	data_type
ride_id	STRING
rideable_type	STRING
started_at	TIMESTAMP
ended_at	TIMESTAMP
start_station_name	STRING
start_station_id	STRING
end_station_name	STRING
end_station_id	STRING
start_lat	FLOAT64
start_lng	FLOAT64
end_lat	FLOAT64
end_lng	FLOAT64
member_casual	STRING

Checking the number of NULL values in each column

ride_id	rideable_type	started_at	ended_at	start_station_name	start_station_id	end_station_name	end_station_id	start_lat	start_lng	end_lat	end_lng	member_casual
0	0	0	0	690809	690806	739170	739170	0	0	4771	4771	0

Columns Start_station_name, start_station_id, End_station_name, End_station_id, End_lat, end_lng contain NULL values which will be removed in the Data Cleaning stage.

Checking duplicate rows

There are no duplicate rows in the dataset.

Row	duplicate_rows
1	0

Checking ride ids length for consistency

All the ride_ids have 16 characters in length which show consistency in the data.

Row	length_ride_id	no_of_rows
1	16	5595063

Checking trip count by user type

Row	member_casual	trip_count
1	member	3066058
2	casual	2529005

Checking trip count by bike type

Row	rideable_type	no_of_trips
1	electric_bike	2031692
2	classic_bike	3251028
3	docked_bike	312343

Checking for trip count greater than 19 hours

Trips whose duration is greater than 19 hrs and less than 5 minutes will be removed in the data cleaning stage.

Row	longer_than_19_hour
1	4957

Checking for trip count less than 5 minutes in duration

Row	less_than_5_minutes
1	853621

3-Data Cleaning and Manipulation

Links to SQL Queries: [Data Cleaning and Manipulation](#) & [Data Exploration after cleaning](#)

- In this stage rows with NULL values have been removed.
- Column name ride_id has been replaced with trip_id.
- Column name rideable_type has been replaced with bike_type.
- Column name member_casual has been replaced with user_type.
- Trip_duration column has been converted into an INT64 data type.
- Trip Duration, Day of Week and Month column has been added in the data using different SQL functions.
- A final table has been produced with the relevant columns using a JOIN.

The schema and data types of the final table is:

column_name	data_type
trip_id	STRING
bike_type	STRING
started_at	TIMESTAMP
ended_at	TIMESTAMP
trip_duration	INT64
day_of_week	STRING
month	STRING
start_station_name	STRING
end_station_name	STRING
start_lat	FLOAT64
start_lng	FLOAT64
end_lat	FLOAT64
end_lng	FLOAT64
user_type	STRING

Checking NULL count in each column after cleaning the data

trip_id	bike_type	started_at	ended_at	start_station_name	end_station_name	start_lat	start_lng	end_lat	end_lng	user_type
0	0	0	0	0	0	0	0	0	0	0

The final dataset does not contain any NULL values.

Checking number of rows for the final table

The final dataset has been reduced to 3507716 rows after removing NULL values and irrelevant data.

Row	trip_count
1	3507716