Title: Cyclistic Case Study Documentation
Author: Ruby Smith
Last Updated: 10/10/2024
Output: google_doc
9/24/2024

Ask Phase

Three questions will guide the future marketing program:

1. How do annual members and casual riders use Cyclistic bikes differently?
2. Why would casual riders buy Cyclistic annual memberships?
3. How can Cyclistic use digital media to influence casual riders to become members?

Start creating Cyclistic Report Below

Set up Purpose, Created Questions From What I am supposed to answer, and identify stakeholders.

Prepare Phase

Download Data: Index of bucket "divvy-tripdata"
    12 sets of data: 202309 - 202408 (One Year of Data) .csv Excel docs.
    Importing to BigQuery through Google Cloud.
Listed License in Report and how analysis is shared
File size too large for google sheets.
File size imported from computer to google cloud and kept private.
Imported all 12 datasets to BigQuery to analyze.

Listed under: rs-project-01-429415.cyclistic_bike_data

09/25/2024

Data Organization in Report
Check for data bias in report
Check for data integrity
Check if data can be used to find answers
Check over for possible problems
Complete conclusion for Prepare Phase

Process Phase

GitHub for SQL queries: RubyRene90/Case_Study_Cyclistic: Scenario: non-membership vs membership (github.com)

SQL Part 1: 1_Understanding_DataSet

Start testing data with SQL

Query NULLs in data to understand why NULLs.
Grouped members and nonmembers to check differences in ride count.
Used UNION ALL to combine all 12 datasets without lat and lng NULLs.
Created a new table from unioned datasets.
'rs-project-01-429415.cyclistic_bike_data.annual_202309_202408_tripdata'

SQL queries Part 2: 2_annual_tripdata_cleaning

Checking new 'annual_tripdata' table

Used EXTRACT to make sure all the months are downloaded.
Used COUNT to make sure all results, minus nulls queried and imported.
Checked for duplicates using ride_id. 171 duplicates
Used Rank with ROW_NUMBER and PARTITION, filtered out distinct rows and created an updated table without duplicates.
'Rs-project-01-429415.cyclistic_bike_data.annual_202309_202408_tripdata2'
Used TIMESTAMP_DIFF to calculate trip time for each row.
Created new column 'trip_time_seconds'.
Created 2 new columns 'start_date' and 'end_date' to separate the date from the timestamp in started_at and ended_at.
Checked timestamp_diff between two separate dates was calculated correctly.
Created column 'day_of_ride' to show day of the week with FORMAT_DATE.

Write up Process phase in report
Create list of tools used to process data
Create list of changes made to original data while cleaning data
Create list of updated columns in cleaned table for report
Complete conclusion for Process Phase.

10/01/2024
Analysis Phase

SQL queries Part 3: 3_annual_tripdata_analyzing

Create google sheet document *'202309_202408_CyclisticAnalyzedData'* for aggregated data created.

SQL
Run aggregate functions COUNT, AVG, MIN, MAX while comparing members and casuals.

Negative results in MIN and MAX is over 24 hours.
Verify original date columns started_at and ended_at in case calculations were wrong
Remove rows with time < 0. (344 rows)

Remove rows with time > 24 hours. (389 rows)

Thought, (table with trip time separated, short, average, long)

Re-run aggregate functions again with updated MIN and MAX (annual results)

Aggregate data for members monthly.

Aggregate data for casuals monthly.

SQL queries Part 4: 4_annual_tripdata_analyzing2

Aggregate data for members and casuals on a weekly basis.

Separate ride time counts into time intervals.

SQL queries Part 5: 5_annual_tripdata_analyzing3

Find out the most popular bike starts and ends for members and casuals

Use AVG on _lat and _lng when grouping starts and ends to get latitude and longitude for popular station locations

Since Q2 data had lower results, checking theory about casuals and pulling same data but only for December, January, February.

All SQL queries can be found at:

GitHub Link For All SQL queries: Case_Study_Cyclistic

RubyRene90/Case_Study_Cyclistic: Scenario: non-membership vs membership (github.com)

Part 1: 1_Understanding_DataSet

Part 2: 2_annual_tripdata_cleaning

Part 3: 3_annual_tripdata_analyzing

Part 4: 4_annual_tripdata_analyzing2

Part 5: 5_annual_tripdata_analyzing3

Import data into google worksheet, *'202309_202408_CyclisticAnalyzedData'*

https://docs.google.com/spreadsheets/d/1DlsFo9a9v79fZ12HbLlfPqG-iBwvY2XONY1JVraHPBs/edit?usp=sharing

Sheet 1: Annual

Sheet 2: Monthly

Sheet 3: Weekday

Sheet 4: Quarterly

Sheet 5: Time_Intervals

Sheet 6: Popular_Locations

Sheet 7: Q2_Popular_Locations

Format worksheet so it's more organized and readable.

Secure google sheet file and make it shareable with anyone with the link.

Add comments to google sheets where I thought appropriate

Constantly run functions in google sheets to check query work

Convert all time in seconds to minutes on google sheets tables

List in report locations of SQL queries and aggregated data imported to google sheets.

In the report, listed tables created and column names.

Complete conclusion on analysis phase.

10/9/2024

SHARE PHASE

After organizing and formatting data, I imported my google sheets document, *'202309_202408_CyclisticAnalyzedData'* **https://docs.google.com/spreadsheets/d/1DlsFo9a9v79fZ12HbLIfPqG-iBwvY2XONY1JVraHPBs/edit?usp=sharing**,

to Tableau Public to create visuals for the analyzed data.
https://public.tableau.com/app/profile/ruby.smith/vizzes

4 Vizzes available:
Cyclistic-Location Popularity
Cyclistic Annually, Monthly, and Weekday Dashboards -Tabs
Cyclistic Time Intervals
Cyclistic Quarterly - Tabs
I saved the vizzes as .jpeg files and used them to create a Case Study Slide Show.
Google Slides *'2024_Cyclistic_Case_StudyRS'*
https://docs.google.com/presentation/d/193xSKJ1-WgIpOcSgydBVMSnMlSDIsP1sYla5yYSAQ9o/edit?usp=sharing

10/10/2024
I updated the SHARE part on my Report.
I added a recommendation to my Slides Project.
I completed the ACT part on my Report.