Google Data Analysis Capstone Track 1

DATA SOURCES

INTRODUCTION

For this data analysis project, Divvy's bike sharing service data was the main data source, providing data-sets on bike trips, and stations, accompanied by the data-sets on supercentenarians provided by the Gerontology Research Group - GRG, the table below gives access to further information on the data providers and the data itself.

	Divvy's	GRG
About	https://divvybikes.com/about	https://grg.org/
Data	https://divvy-tripdata.s3.amazonaws.com/index.html	https://grg.org/WSRL/TableE.aspx
License	https://divvybikes.com/data-license-agreement	Free for public use with citation

PREPARING DATA

The prepare step in data analysis comes before data processing, and is a way to ensure data accuracy and ease of use, through addressing a series of characteristics of the given data, as follows.

Data Sources

The data coming from Divvy is provided in .csv format, structured as a table, the data-set on trips present on each row information around ride, station and user, the portion about the ride presents an id for the trip and it's date and time of beginning and ending, related to stations it has name and id for both departure and arrival, and for users it disposes of name, gender, customer type, and year of birth.

Still on data coming from Divvy, the data-set on stations, is also provided in .csv format, the data is structured as a table, and the columns of interest have data on station name, id, and coordinates, it also dispose of data on when the station went ‘online’ and their capacity.

This capstone project revolves around the data provided by Divvy, while the next data to be presented, provided by the Gerontology Research Group - GRG, is ‘collateral’ and used only in the cleaning process, as a way to check the validity of some age entries.

The supercentenarians data-set, is hosted by the Gerontology Research Group - GRG on their site (HTML) it comprises 3 different tables on supercentenarians, divided by their status on the date of their last research, the status are: proven to be alive; proven to be deceased; and "Limbo" (unknown) for supercentenarians that they lost track of. The data was scraped from their site using google sheets formula IMPORTHTML(), and written permission for use was given by the group.

Some of the most important data entries that the data-set provides for the Cyclistic business goal are age, sex, residence and birthplace, since these entries can be correlated to the entries present in the Divvy dataset, that acts for the purpose of this capstone project, as data related to the fictional company called Cyclistic.

Issues and Bias

Issues with credibility may involve, rides that were started but shortly after aborted by the user, leading to inconsistencies on trip length metrics, entries of birth years state that certain users are of advanced age, with some being more than 110 years of age, the common sense points for it to be a wrongly entered value, but to evade bias a data-set will be presented to validate or invalidate these entries.

Some customers have entries for age and gender, even though the metadata on the Divvy trips data-set states that there shouldn't be any, those issues will be addressed during the cleaning process to secure credibility.

Licensing, Privacy, Security and Accessibility

Privacy is already secured by the data nature, no names, payment info, or any other way to pinpoint informations to a single user, when it comes to license Divvy allows to use their data for educational purposes, and GRG data can be used for public project and citation, when it comes to security, the data doesn't pose any threat to the security of anyone that had their data of service use collected, the data being used is from 2016, and insights from it aren't of use for possible business competitors, when it comes to accessibility, all the cleaning and analyzing processes will be documented and shared on kaggle, high contrast colors will be used on graphs to facilitate visualization for all users.

Data Integrity

When it comes to data integrity all data-sets will be checked using R programming language and manual verification, scripts will be made to ensure that the data follows what is depicted on their provided metadata, all entries will be verified to be compatible on their expected type and format, nulls will be discarded, and duplicated values normalized, through manual checking paired with scripts, entries will be grouped, sorted and checked to spot possible misspellings, all processes and resulted data will be hosted on kaggle and google drive.

How Does Data Helps To Answer The Business Goal

How does the data help me answer my question: it allows me to identify different patterns of use between types of user when it comes to user preference on period of the year and of the week, ride trajectory and duration, it also permits to identify proportions between the frequency of use of each user type, and to construct and age and gender profile of the subscribers.

Problems With The Data

There are no actual problems, what appears to be a problem will be worked through data cleaning, but it would be very useful for the business goal to have some qualitative and quantitative research, to further understand Cyclistic's users, and backup what is being analyzed on this dataset.