Author: **ADAM ZANUB**

Title:

Reporting of the exploration, preparing, cleaning and processing data.

Project Name:

Google Capstone / Cyclistic bike-share

Tool I have used:

R version 4.3.1

The business task I have worked on it is:

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

Data Source:

- I used Cyclistic's historical trip data to analyze and identify trends.

- I downloaded the data from the following link:

  URL data source

- The data has been made available by Motivate International Inc. under this license

- The data were collected during the year 2023, for the first quarter of the year.

Exploration & preparation phase (1):

I have checked out about the following points:

1. Number of variables and observation that we have.

2. The names of each columns.

3. Understanding the structure of the dataset.

4. Discovering the head and the tail of the data.

5. Looking at NA / missing values

6. Distinct values for each column

7. Duplicated data

* I considered closely on each column individually to emphasize and warp up this phase.

Cleaning & Processing Phase (2):

1. Col 2 (rideable_type) & col 13(member_casual) I converted data type to factor

2. Start station:

Creating a data frame which include missing values with two columns (latitude &longitude for purpose of checking the number of digits after the comma)

3. End station:

* I did the same thing I did it at start station, However, through (checK_lat_end & check_lng_end) I have revealed a total of 426 missing values completely.

Station and coordinate information are completely missing.

* We need coordinates out to at least five (5) decimal places to be usable in terms of locating, However, in this data we only have (2) decimal places at the best.

It may be caused by the company's GPS system failing which resulted in the inability to determine the station name and station ID.

4. Extracting the address for each start_station_name, end_station_name based on latitude &longitude:

- I have created the project from version control for this purpose:

  Clona Git repository → the link that I used [URL repository]

5. Calculating ride time by creating a new column named ride_time

I found just one value it looks inconsistent (start point more than end point) then I round the value to just minutes.

*I've created data frame contains the missing values with trip durations

*casual max duration equal to (33604 minutes) it's a unique value, while member max duration equal to (1560 minutes) and there are multiple value though. However, all these max values include missing values at end station name, end lat, end lng.

1) member trip duration min range [1500 - 1560]
2) casual trip duration min range [2 - 33604]

- I noticed that end_station_name with end_lat & end_lng in this piece of data has the biggest number of ride length and the casual customer takes the biggest proportion.

- There's something wrong: end_station_name, latitude , longitude all of them are missing.

- I believe that's in this piece of data the bicycles might have been stolen or been involved in an accident which resulted in not registered on the system so I decided to delete this piece of data (426 NAs) for two reasons: that's not a huge number in addition, there is doubt about accuracy of this data

6. Extracting time from date (for both start and end point):

* I have taken the time of demand from date column and then convert the data type to hms format.

* I considered about the most and least time frequent in the data, however I created a dataframe focusing the times frequent just to scope on the times and get some insights

7. Extracting day from date (foe both start and end point):

* I have taken the day of demand from date column and in a standard format.

* I verified that all days of week are exist.

* I checked out the most and least day is common in data.

8. Study how each type of customer interacts with other variables we're interested in?!

* I have created a data set for only casual including variables I have interested in to observe how they behave with it. So I have some insights about:

1) popular & unpopular start and end day .

2) busiest start and end day & time.

3) using bikes based on type of it.

4) top start and end streets used by them.

(The same things I did it with member)