experion
technologies

Case Study

D&A Graduates Training

Contents

Problem statement

You are a foodie who wants to do some exploratory data analysis and find trends and insights on the eating habits of people across the country, restaurant ratings and value for money from a sample restaurant data set provided by one of the leading food delivery companies in the country.

About the dataset

- The dataset provided (in the Appendix section) is in a csv format with multiple files for each city in India.
- All the files have the same structure.
- The delimiter used is '|'
- The dataset provides information such as restaurant name, location, type of cuisine, average cost for two people, ratings, votes, URL etc.

What you must do

- Spin up a database either locally in your machine or as a manged instance in AWS or Azure.
- Create the necessary staging and reporting schemas and tables as required.
- Build an end-to-end ETL pipeline using python that will iterate through all the folders and load the csv data into the database.
- Keep in mind these points while building your pipeline – error handling and logging, restartability, performance optimization, data integrity.
- Build dashboards in PowerBI/Tableau to visually represent the analysis you have done.
- Use ML algorithms/models to do predictive analytics.

What you need to find out

This is an exploratory data analysis exercise. Hence what you can find out from the data depends on how well you understand the data. But here are few pointers to get you started.

- Which locality lists the maximum number of hotels.
- What cuisine is served in the highest rated hotels in the country.
- Is there any relation between the cost of dining and the hotel rating?
- Plot on a map the highest rated hotels in your city.
- Build a dashboard to show the distribution of restaurants across cities.
- Which cuisine dominates the Indian taste buds and is there any relation to where the hotel is located?

Data Enrichment

- The dataset provided is just a sample set and has limited data points.
- To improve your analysis this data can be enriched with information that is available from other sources.
  HINT: To plot the location of a hotel on a dashboard, the latitude and longitude co-ordinates are required.
- The hotel URL opens a webpage that has many other data points that can be easily scrapped to enrich your data and provide better analysis.

How we plan to do this

- Understand the problem statement.
- Discuss on the requirements and possible approaches.
- Analyse the datasets provided.
- Design the high level ETL framework.
- Design the database and objects.
- Build the ETL pipeline.
- Design the reporting dashboards.
- Build the dashboards
- Design and build your ML models
- Finally, present your work and submit your case study.

Appendix

- Dataset link – Zomato_Dataset.zip
- Softwares/Applications required
  - Python
  - PostgreSQL/MSQL database
  - Dbeaver
  - AWS/AZURE account
  - Tableau/Power BI desktop version