CHRIST (Deemed to be University)

Data Science Department

Bachelors of Science (Data Science)

Name: Siddharth R Bhardwaj Course: BDS471L – Machine Learning

Reg No: 22112028 Exercise No: LAB Exercise – 2

Date: 26 – 01 – 2024

**Topic: Accessing Data**

1. Clearly outline the purpose of accessing the data. Are you trying to gain

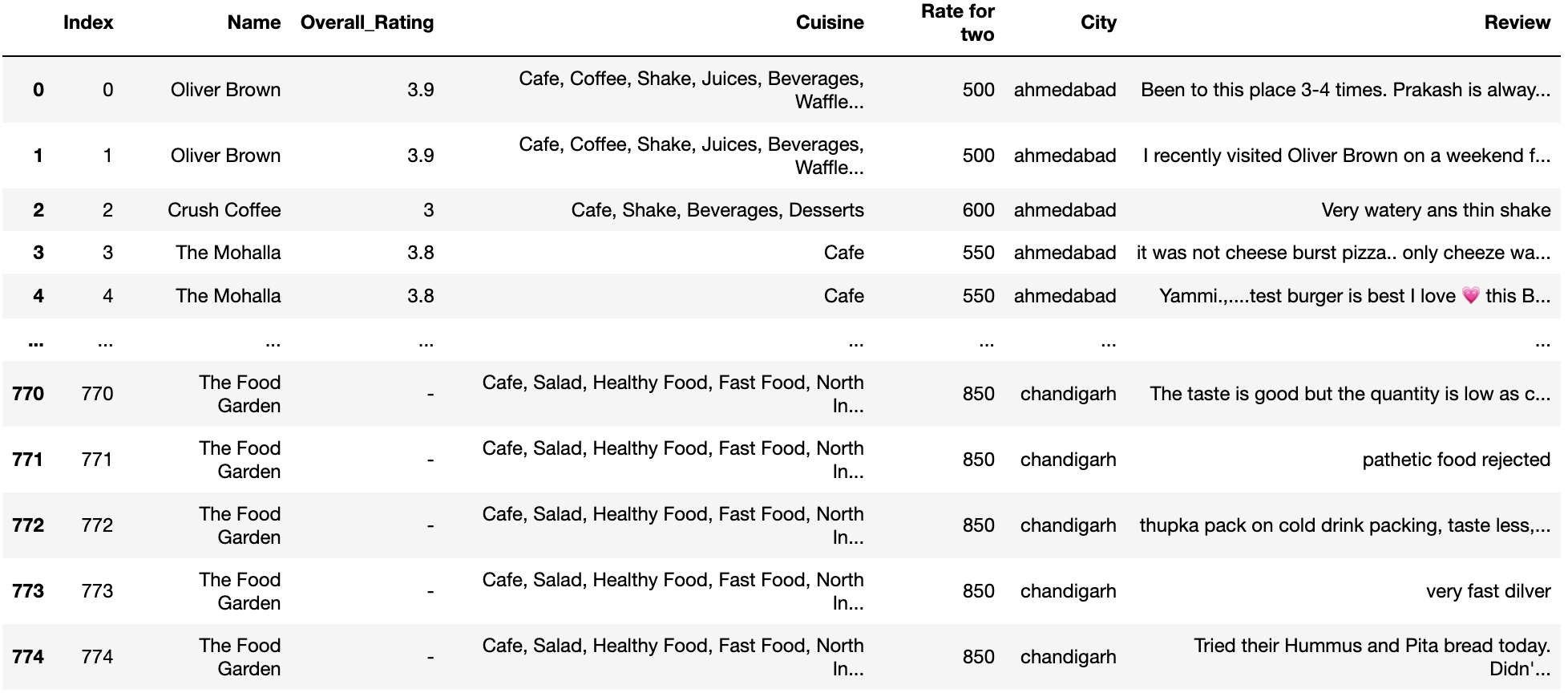
insights, make predictions, identify trends, or something else?

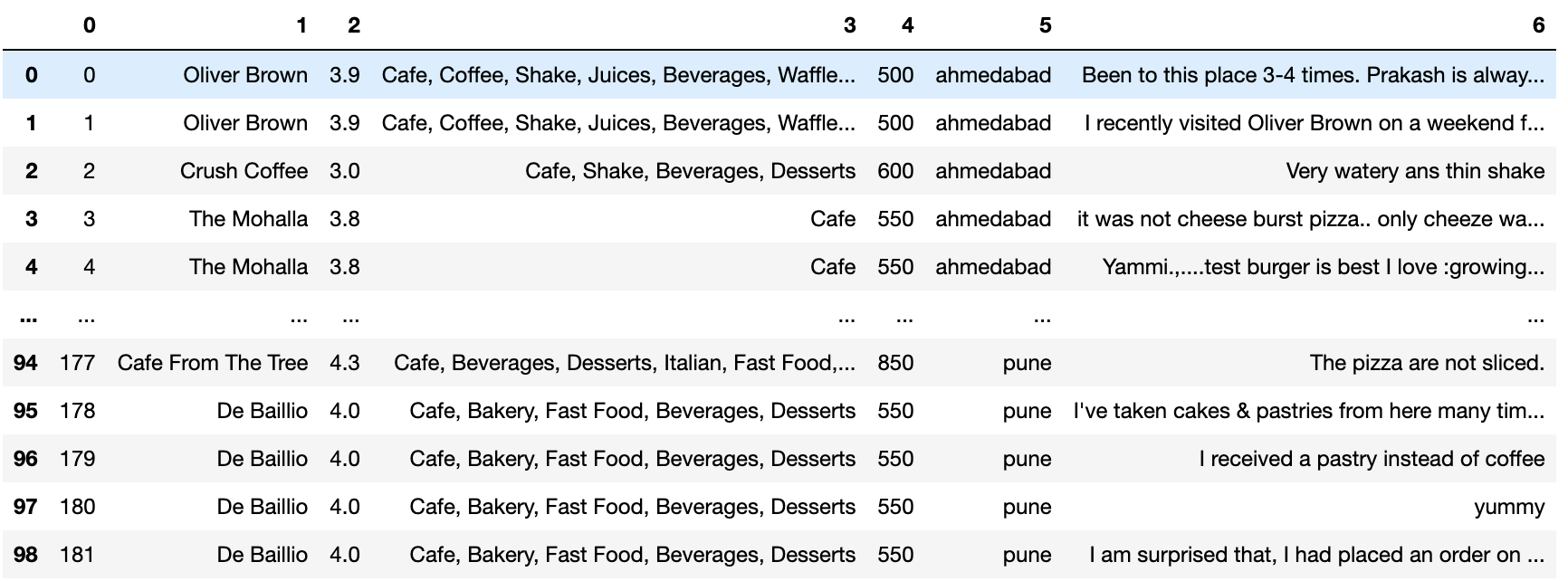
- It aims to gain insights into cafe culture, customer preferences, and city-specific dining trends. Through culinary research, we can explore diverse cuisines and flavors.

- Sentiment analysis can provide us with a deeper understanding of customer experiences, identifying praised aspects and areas for improvement.

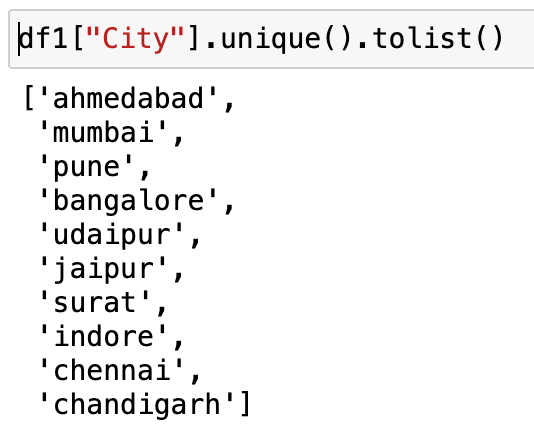
- Urban exploration enables the observation of trends across different cities.

Overall, CafeCritic dataset’s purpose is to gain insights into cafe culture, customer preferences, and city-specific dining trends.





2. Define the time frame, geographical area, or any other specific criteria relevant to your analysis.



- This data was collected by JUHI BHOJANI and it was UPDATED 5 MONTHS AGO, using the method of Web-Scraping.  
Sources: Zomato Website: [https://zomato.com](https://zomato.com/" \t "/Users/sid24082/Documents/THOUCENTRIC/REPORTS/x/_blank)

- The analysis is focused on cafe reviews in specific cities: Ahmedabad, Mumbai, Pune, Bangalore, Udaipur, Jaipur, Surat, Indore, Chennai, and Chandigarh.

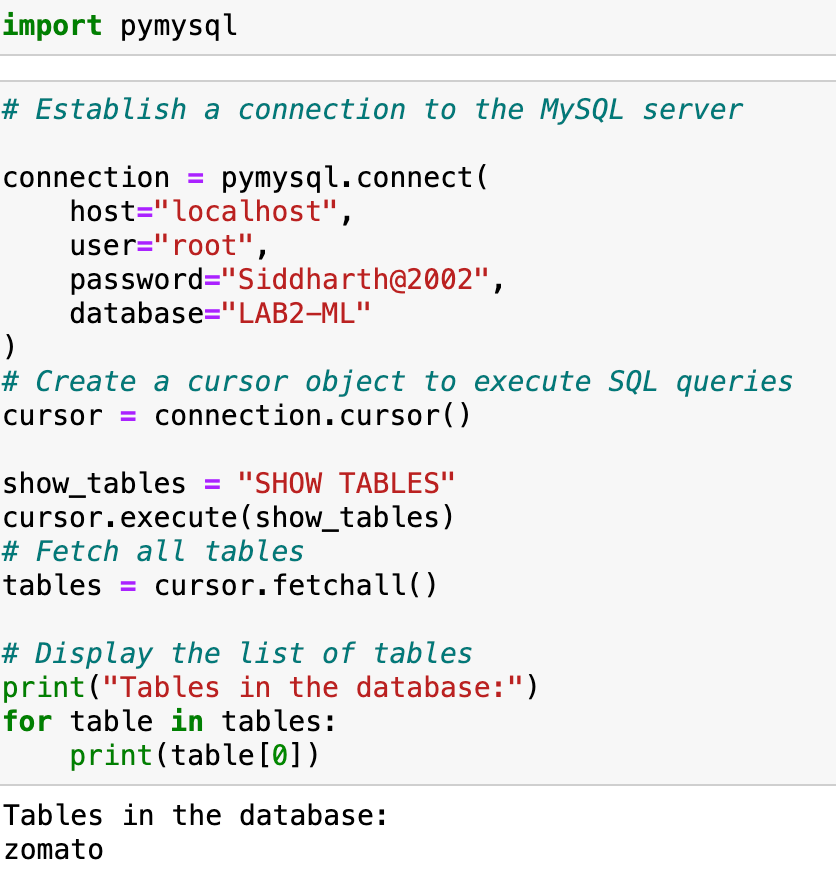
1. Identify and mention the different data sources where the required data is available.

- The similar data can be web scraped from the company’s website and various resources, few sources where the data is available are

* Kaggle
* Github Repositories
* Social Media Platforms
* Amazon AWS Public Datasets
* Google Dataset Search

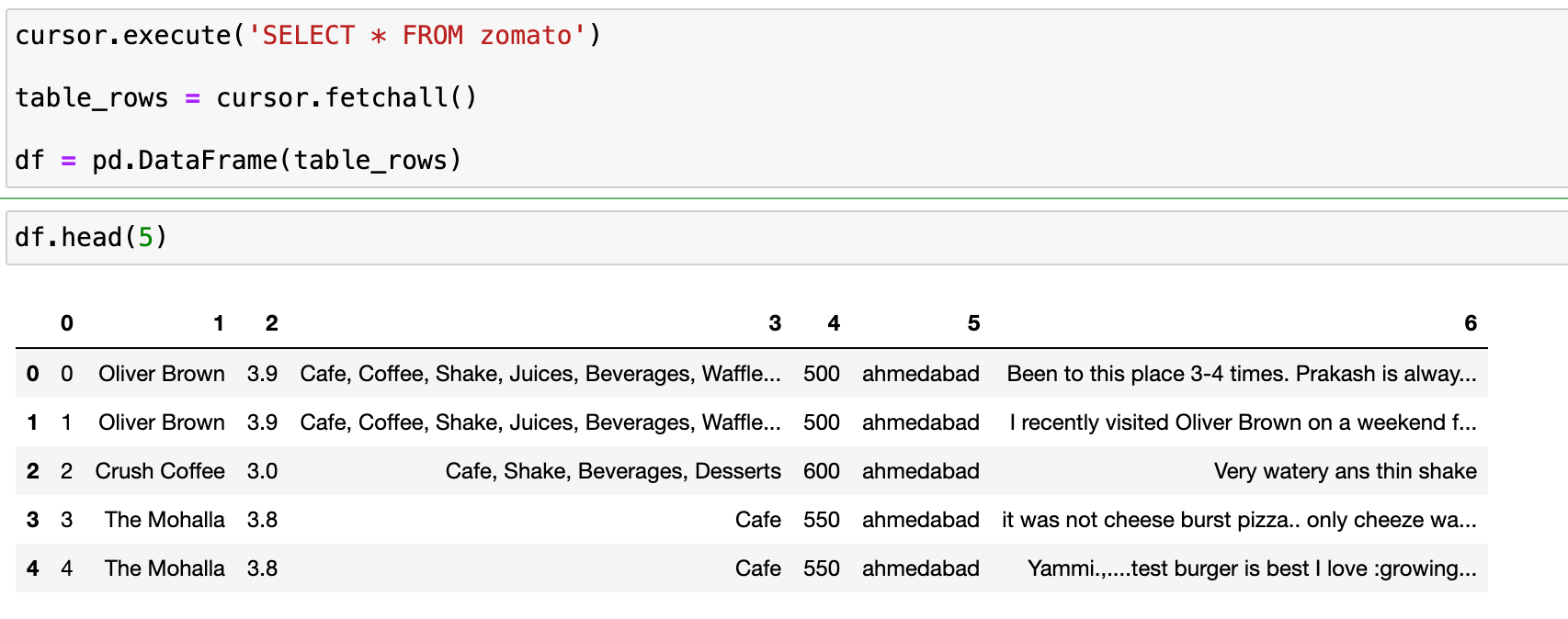
and many more according to our requirement.

1. Perform the below steps under Data Wrangling
2. Asking Questions ( Minimum 5 Questions)
3. What are the top three most popular cuisines across all cities in the dataset?
4. Which city has the highest average overall rating for cafes, and which one has the lowest?
5. Identify any correlation between the average cost for two people dining and the overall rating of cafes?
6. What are the most common positive sentiments mentioned in customer reviews, and are there any recurring negative sentiments?
7. Are there any specific trends or patterns in customer preferences or ratings unique to a particular city?
8. Gathering Data
9. Extract the data from external source store in Mysql Database.

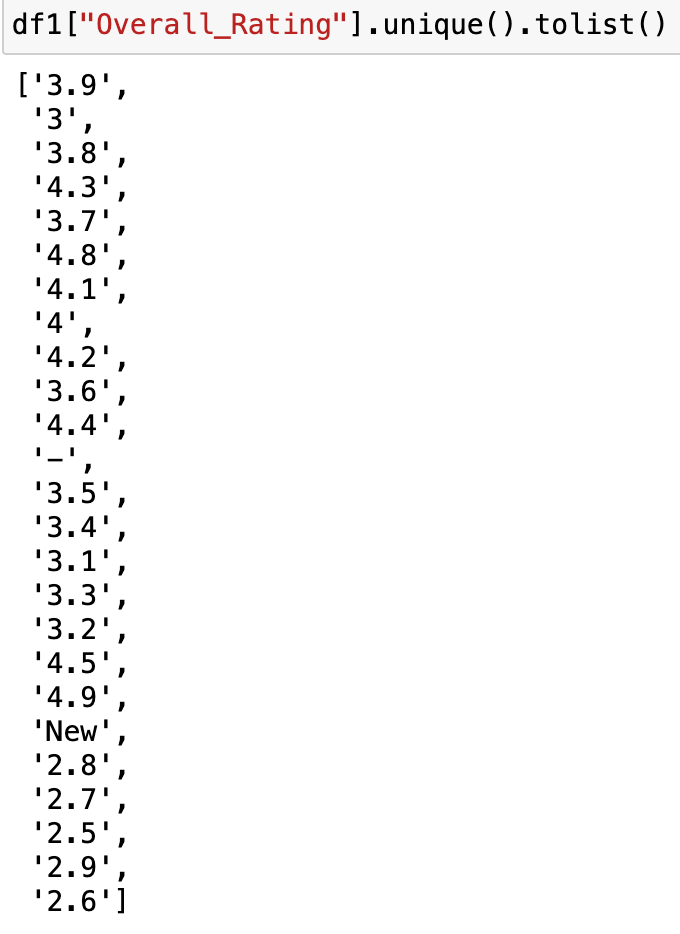


ii. Extract the data from Mysql Database and convert to a

data Frame for analysis.



c. List the all the quality issues in the dataset

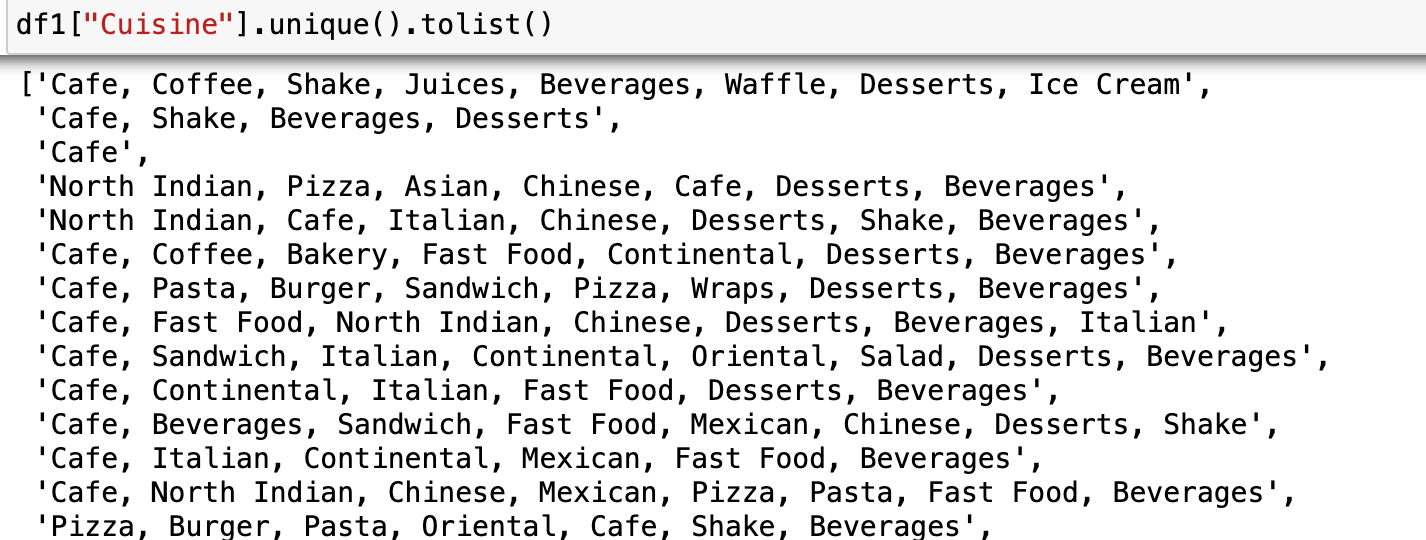


- Inconsistent Rating Format:

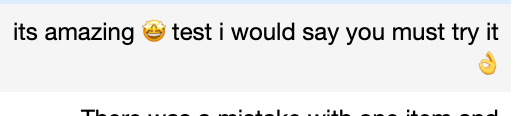
Ratings are expressed as both integers (e.g., 4) and decimals (e.g., 3.8). Ensure consistency in the rating format for easier analysis.

- Inconsistent Cuisine Format:

The "Cuisine" column seems to contain multiple cuisine types separated by commas.



- Review includes emoji, this made the task of uploading the dataset to my database very tough. I removed the emojis to make it utf-8 format for the Mysql-Workbench to be able to recognise it.



- Incorrect Reviews: Some of the reviews have “-” and “New” as the reviews which cannot be used or interpretated. Any analysis done for the cafe based on these reviews will end up providing bad results or inaccurate rating.

