Hive Exercise 2

Realtor Data - Use realtor-data.csv

1. Create a new database called exercise
2. In this newly created database, create an empty table named housing_price
3. Load data from the local CSV file into the Hive Table
4. Display the first 5 rows of the table
5. Get distinct zip codes from the dataset
6. List the total number of records in the table
7. Get the Total number of beds
8. Get number of properties where state = "Massachusetts"
9. Create a new table based on the previous table where city "San Juan"
10. Write a query such that if there is an entry called Warwick print "My City" else "Not My City"

India Pin/Zip Code (City, Area, District, State) - use India_pincode.csv

Exercise: Exploring India Pincode Data

In this exercise, you have a dataset called "India_pincode.csv" containing information about cities, areas, pincodes, districts, and states in India. You will perform the following tasks:

Task 1: Load the Data

Load the "India_pincode.csv" dataset into a Hive table named "india_pincode."

Task 2: Data Exploration

1. Find the total number of unique cities in the dataset.
2. Find the total number of unique states in the dataset.
3. Find the number of unique pincodes in the dataset.

Task 3: Data Filtering

Filter the data to retrieve information about cities in a specific state, e.g., "Maharashtra."

Certainly! Here are a few more exercises involving Hive Query Language (HiveQL) that you can use to practice data manipulation and analysis:

Task 4: Find the Most Populated Districts in a State

In this exercise, you have the "India_pincode.csv" dataset, and you want to find the top N most populated districts in a specific state. Assume you are interested in the state of "Karnataka."

Task 5: Calculate the Average Pincode Length

In this exercise, you will calculate the average length of pincodes (number of digits) in the dataset. This can provide insights into the structure of pincodes in India.

Task 6: Identify Duplicate Pincodes

In this exercise, you will identify and list any duplicate pincodes in the dataset. This is useful for data quality checking.

Task 7: Filter Data by Area Name

In this exercise, you will filter the data to retrieve information about a specific area name (e.g., "Baner") and list the cities in that area.

Task 8: Calculate the Number of Cities per State

In this exercise, you will calculate the number of unique cities in each state and present the results in descending order of city count.

Task 9: Find the State with the Highest Number of Pincodes

In this exercise, you will find and display the state with the highest number of unique pincodes.

Task 10: Data Aggregation by District

In this exercise, you will calculate the total number of pincodes and areas per district and order the results by district name.
