Fuel Price Data Analysis with PySpark (Extension of Group Project)

Presented by:
- Sushan Kattel

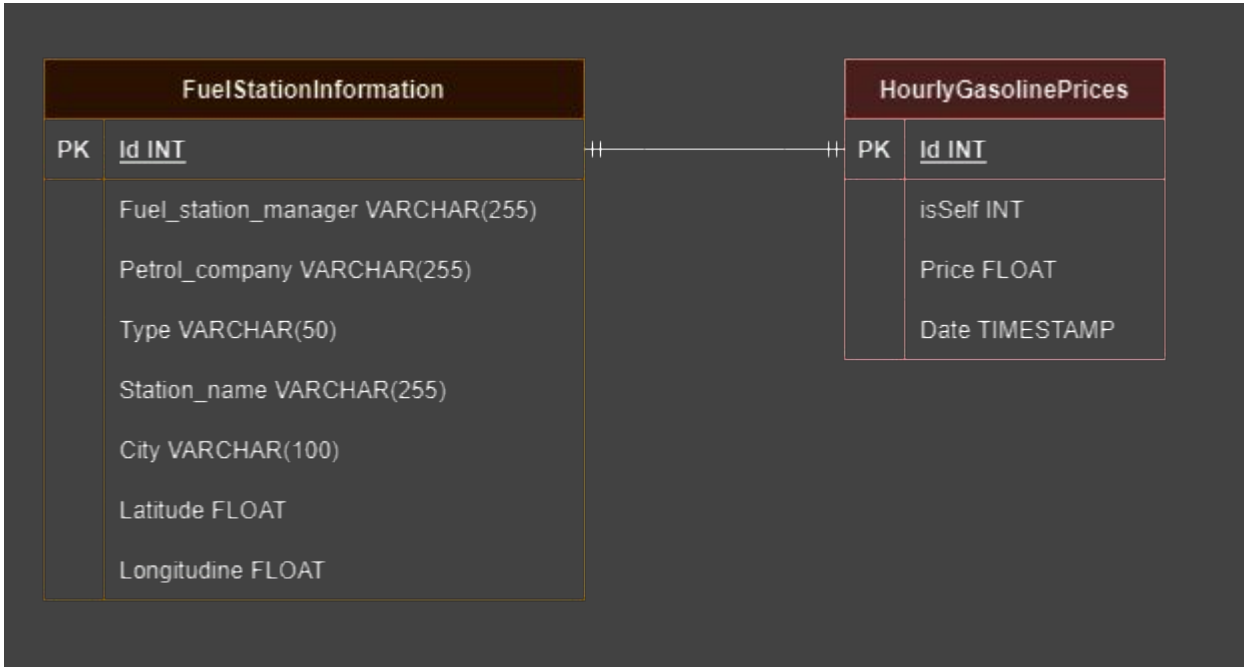
Table of Contents

Introduction

The dataset used in this project:

- https://www.kaggle.com/datasets/alessandrolobello/gasoline-hourly-price-tracker-from-2022?select=Hourly_Gasoline_Prices.csv
- This dataset contains information about fuel stations, petrol companies, types, station names, cities, and coordinates.

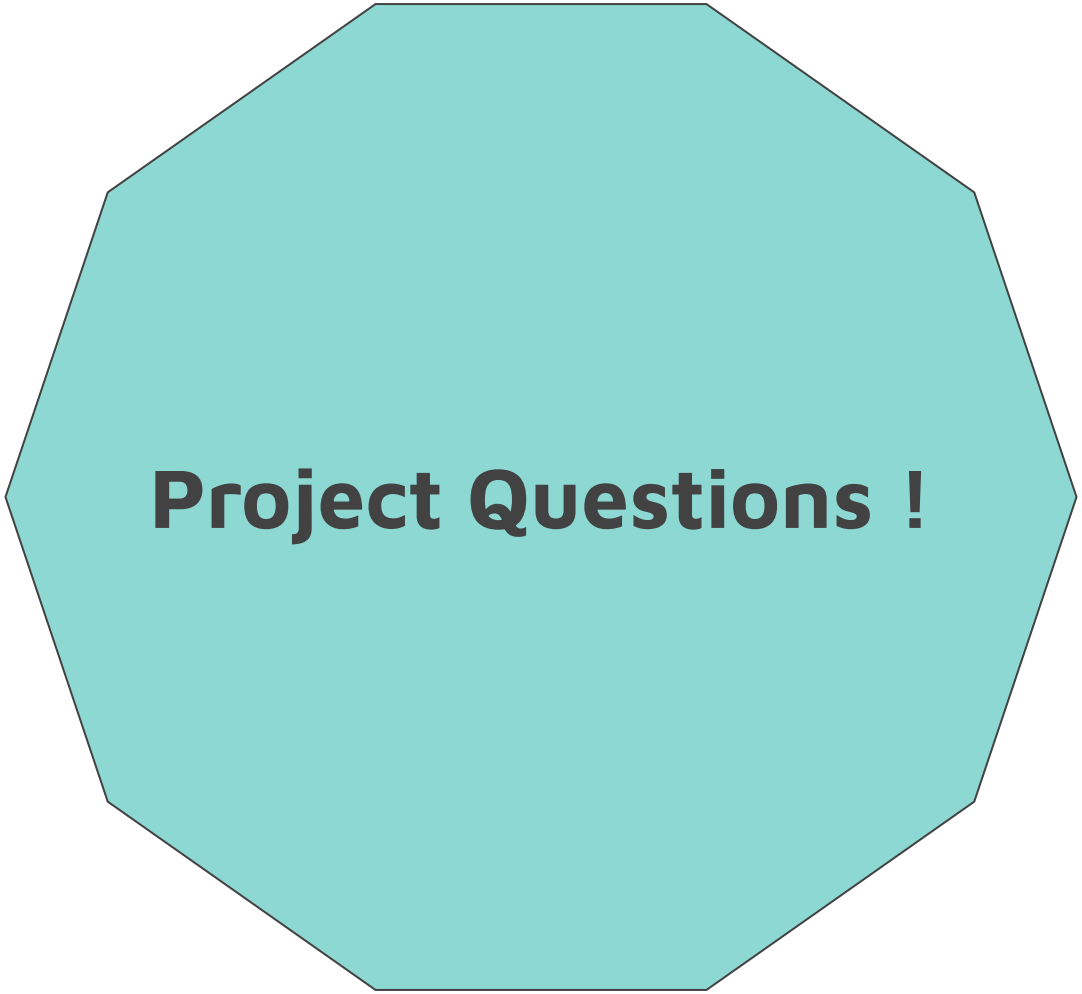
More About Dataset



Data Cleaning

To complete the task:

1. Data Loading: Load the csv data files
2. Duplicate Removal: Duplicates from both data were removed for data integrity.
3. Handling Null Values: Rows containing null values were dropped from both datasets.
4. Column Transformation: The "Date" column in the "Hourly Gasoline Prices" dataset was transformed into a timestamp format for easier analysis of time-series data.
5. Data Storage: Cleaned datasets were saved as Parquet files for efficient storage and retrieval in future analysis. Then those files were written to the database.

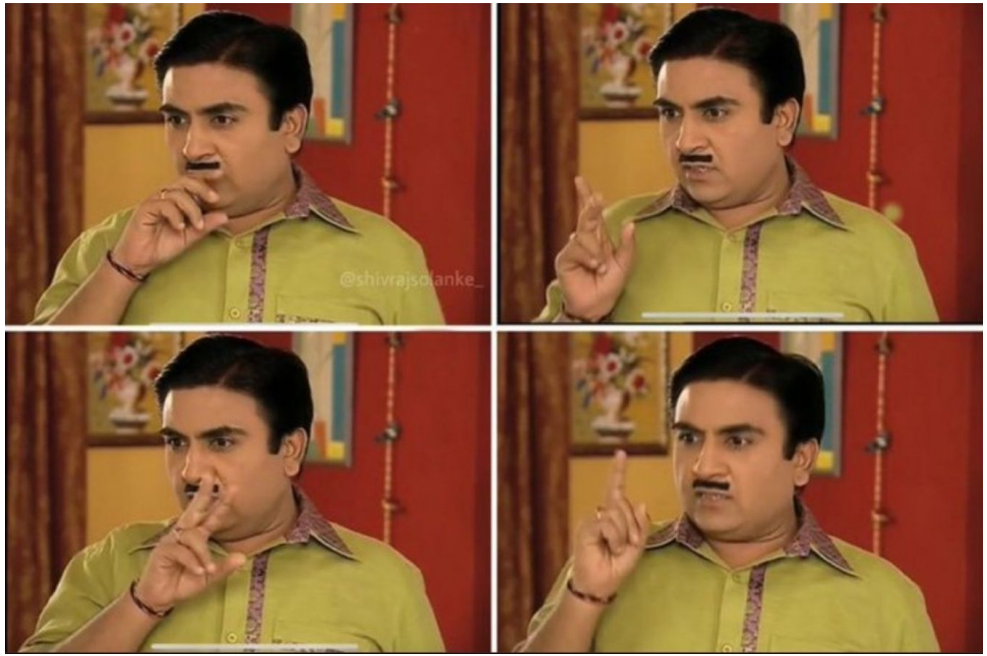

Project Questions !

Q1. Find the Most Stable Company and Most Volatile Company based on Average Daily Price Variation.

To complete the task:

- Calculated daily price changes for each company
- Ranked them by their average variation.
- The company with the lowest average was the most stable, while the highest was the most volatile.
- This analysis helps identify pricing behaviors among different companies.

Q2. Find 9 nearest stations to a certain reference station and calculate the price difference between the fuel station and its nearest competitor



To complete the task:

- Found the 9 nearest fuel stations to a reference station based on latitude and longitude data
- Calculated the price difference between each station and its closest competitor.

Lets move to code !

Thank You !