Fuel Price Data Analysis with PySpark

Presented by:

- Sushan Kattel
- Rojesh Pradhananga

Mentor:

- Amrit Prasad Phuyal

Table of Contents

Introduction

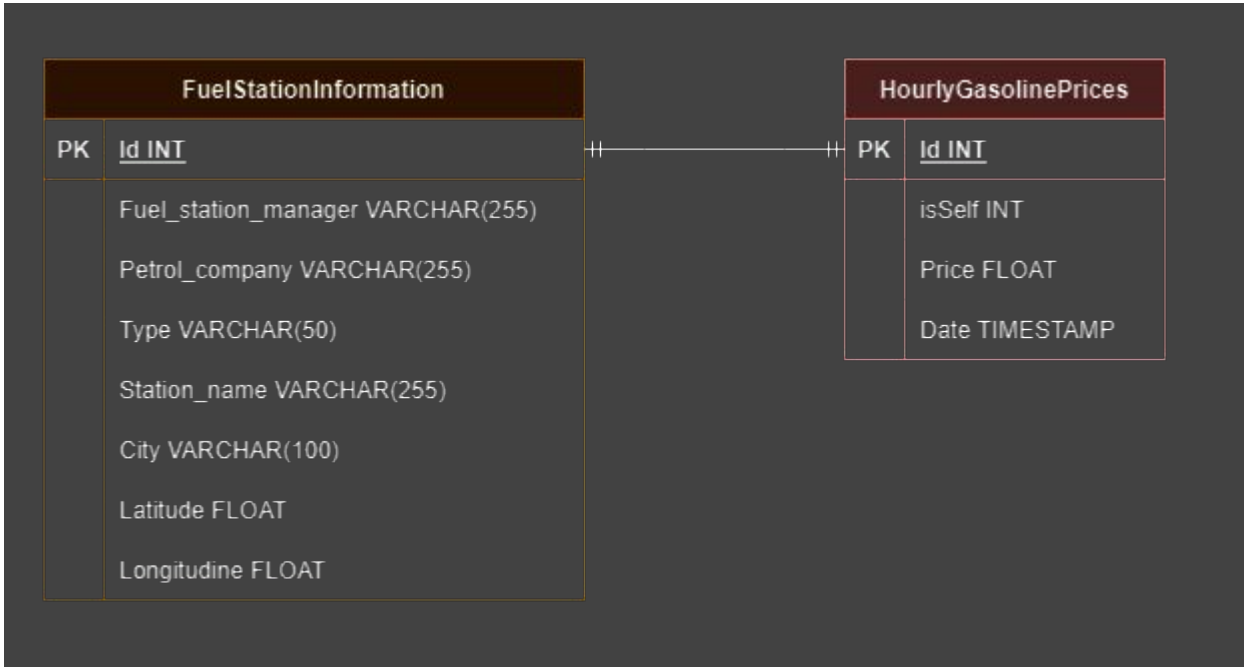
The dataset used in this project:

- 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.

The project aims to analyze and process fuel price and station information data.

Our goal is to provide valuable insights into fuel pricing and station locations.

More About Dataset



Work Division

1) Dataset Searching - In Group

2) Preprocessing - Sushan

3) Researching questions - In Group

4) Individual questions - On own

5) Preparing Presentation - Rojesh

6) Preparing Readme.md file - Sushan

7) Merging Github - Sushan

Data Preprocessing

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.

The 10 Questions !

Question 1: UDF - Currency Conversion

Gasoline prices were successfully converted from USD to Euro using the provided UDF and exchange rate data from the API.

Input:

```
def convert_to_euro(usd_price):
    api_url = "https://cdn.jsdelivr.net/gh/fawazahmed0/currency-api@1/latest/currencies/usd/eur.json"
    response = requests.get(api_url)
    exchange_rate_data = response.json()
    exchange_rate = exchange_rate_data["eur"]
    euro_price = usd_price * exchange_rate
    return euro_price
fuel_prices_df = fuel_prices_df.withColumn("Price_Euro", convert_to_euro(col("Price")))
fuel_prices_df = fuel_prices_df.withColumn("Price_Euro_Rounded", round(col("Price_Euro"), 3))
fuel_prices_df = fuel_prices_df.drop("Price_Euro")
fuel_prices_df.show()
```

Output:

Id	isSelf	Price	Date	Price_Euro_Rounded
5079	1	1.809	2022-01-04 06:39:09	1.676
38752	1	1.804	2022-01-04 07:18:19	1.671
51635	0	1.974	2022-01-04 07:39:23	1.828
6810	0	2.009	2022-01-04 07:50:04	1.861
4983	0	1.758	2022-01-04 07:55:36	1.628
51790	1	1.819	2022-01-04 08:50:47	1.685
42708	0	2.064	2022-01-04 08:51:50	1.912
48545	1	1.799	2022-01-04 09:26:35	1.666
46455	0	1.954	2022-01-04 12:14:28	1.81
47555	0	2.048	2022-01-04 12:58:37	1.897

Question 2: Highest Price and Average

ID with Highest Price: 54771

Maximum Average Price: 1.835294686861716

Input:

```
max_price = fuel_prices_df.agg(max("Price")).collect()[0][0]
highest_price_id = fuel_prices_df.filter(col("Price") == max_price).select("Id").first()[0]
sorted_prices_df = fuel_prices_df.orderBy(col("Price").desc())
average_price = fuel_prices_df.agg(avg("Price")).collect()[0][0]
print("ID with Highest Price:", highest_price_id)
print("Maximum Average Price:", average_price)
```

Output:

```
ID with Highest Price: 54771
Maximum Average Price: 1.835294686861716
```

Question 3: Day of the Week Analysis

The dataset now includes a new column "Day_of_Week" with corresponding day names. The days are ranked based on sales counts.

Input:

```
fuel_prices_df = fuel_prices_df.withColumn("Day_of_Week", date_format(col("Date"), "E"))
day_sales_df = fuel_prices_df.groupBy("Day_of_Week").agg(count("*").alias("Sales_Count"))
window_spec = Window.orderBy(col("Sales_Count").desc())
day_sales_ranked_df = day_sales_df.withColumn("Rank", dense_rank().over(window_spec))
day_sales_ranked_df.show()
```

Output:

```
+-----------+-----------+----+
|Day_of_Week|Sales_Count|Rank|
+-----------+-----------+----+
|        Mon|     410910|   1|
|        Thu|     385288|   2|
|        Wed|     375737|   3|
|        Tue|     363014|   4|
|        Fri|     362752|   5|
|        Sat|     308091|   6|
|        Sun|     238545|   7|
+-----------+-----------+----+
```

Question 4: Unique Cities and Nearby Fuel Stations

A list of unique cities was obtained. For each city, the number of other fuel stations present within a 100km radius was determined.

Input:

```
def calculate_distance(lat1, lon1, lat2, lon2):
    return haversine((lat1, lon1), (lat2, lon2), unit=Unit.KILOMETERS)
result_df = df_joined.withColumn("Distance_between", calculate_distance(
    col("Latitude"), col("Longitudine"), col("Latitude_d2"), col("Longitudine_d2")
))
result_df = result_df.select("Id", "Petrol_company", "Type", "Station_name", "City"
filtered_df = result_df.filter(result_df["Distance_between"] < 100)
filtered_df.show(truncate=False)
```

Output:

Type_d2	Station_name_d2	City_d2	Distance_between
Stradale	AYMAVILLES-F.NE VILLETOS S.R. 47	AYMAVILLES	0.0
Stradale	TotalErg	CHIERI	88.12894
Stradale	GIODA AGOSTINO SRL -G.P.L.	VENARIA REALE	72.07052
Stradale	EGI-2GO VILLADOSSOLA	VILLADOSSOLA	89.19637
Stradale	I P	CRESCENTINO	85.81389
Stradale	ip pinet jimmy	VERRE'S	34.593204
Stradale	CENTRO CALOR BRICHERASIO	BRICHERASIO	97.49551
Stradale	IVREA-VIA TORINO 216	IVREA	57.027435
Stradale	ip	RIVOLI	72.784096

Question 5: Pivot Table and Day Counts

A pivot table was created to display the count of sales for each day of the week.

Input:

```
fuel_prices_df = fuel_prices_df.withColumn("Day_of_Week", date_format(col("Date"), "E"))
pivot_df = fuel_prices_df.groupBy().pivot("Day_of_Week").agg(count("*"))
pivot_df.show()
```

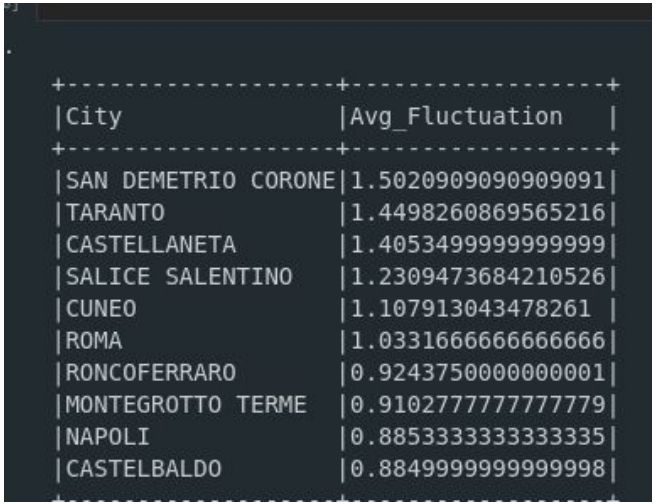
Output:

```
+------+------+------+------+------+------+------+
|   Fri|   Mon|   Sat|   Sun|   Thu|   Tue|   Wed|
+------+------+------+------+------+------+------+
|362752|410910|308091|238545|385288|363014|375737|
+------+------+------+------+------+------+------+
```

Question 6 : Top 10 cities with the highest price fluctuations

Total revenue generated by each petrol company for self-service fuel stations in 5 cities, categorized by fuel station types, was calculated.

Output:



City	Avg_Fluctuation
SAN DEMETRIO CORONE	1.5020909090909091
TARANTO	1.4498260869565216
CASTELLANETA	1.4053499999999999
SALICE SALENTINO	1.2309473684210526
CUNEO	1.107913043478261
ROMA	1.0331666666666666
RONCOFERRARO	0.9243750000000001
MONTEGROTTO TERME	0.9102777777777779
NAPOLI	0.8853333333333335
CASTELBALDO	0.8849999999999998

Question 7: Month with lowest number of hourly records.

It gives the month with lowest number of hourly records for any given city

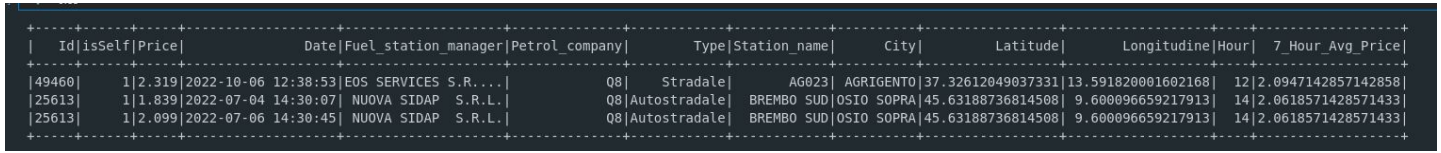
Output:

```
The month with the lowest number of hourly records for 'Stradale' fuel stations in 'SERRAVALLE SCRIVIA' is: February
The number of records for that month is: 9
```

Question 8: Rolling 7-Hour Average Gasoline Price

It gives Rolling 7-Hour price average for given petro company.

Output:



Id	isSelf	Price	Date	Fuel_station_manager	Petrol_company	Type	Station_name	City	Latitude	Longitudine	Hour	7_Hour_Avg_Price
49460	1	2.319	2022-10-06 12:38:53	EOS SERVICES S.R....	Q8	Stradale	AG023	AGRIGENTO	37.32612049037331	13.591820001602168	12	2.0947142857142858
25613	1	1.839	2022-07-04 14:30:07	NUOVA SIDAP S.R.L.	Q8	Autostradale	BREMBO SUD	OSIO SOPRA	45.63188736814508	9.600096659217913	14	2.0618571428571433
25613	1	2.099	2022-07-06 14:30:45	NUOVA SIDAP S.R.L.	Q8	Autostradale	BREMBO SUD	OSIO SOPRA	45.63188736814508	9.600096659217913	14	2.0618571428571433

Question 9: Average Gasoline Prices by Petrol Company

It gives us insight about average gasoline prices by petrol_company.

Output:

Petrol_company	AveragePrice
Edison metano	2.0391612903225806
RPetroli	2.0345555555555555
europetroli	2.0015
Sarni Oil	1.9728359697386528
messina carburanti	1.958948717948718
STOM	1.957798165137615
COLAGROSSI CARBUR...	1.954388888888889
CarbonOil	1.9518230088495576
Economysrl	1.9483506666666666
null	1.9309999999999998
Adamo Idrocarburi	1.926859402460457
FUELPP	1.9195670103092783
Italiana Carburanti	1.9189697286012524
VULCANGAS	1.9122450980392158
MOVE 2	1.9107894736842106
Energy Rete	1.9088771929824562
""CANTINA SAN L...	1.9030729166666667
GAN	1.901095303867403
78	1.8975000000000002
Fuel99	1.892888888888889

only showing top 20 rows

Question 10: Highest Average Hourly Gasoline Price

It gives the output about highest average hourly gasoline price.

Output:

```
The combination with the highest average hourly gasoline price is 'Type: Stradale', 'City: SAN DEMETRIO CORONE' with an average price of 2.89.
Maximum price was 4.00 on 2022-10-17 09:58:42
```

Thank You !