Exploration and Preprocessing of data

Introduction

The purpose of this report is to present an Exploratory Data Analysis (EDA) performed on a dataset containing information about service stations and fuel prices in Spain. For this analysis, the tools used were Pandas, Matplotlib, and AutoViz, which allowed for data visualization, cleaning, and transformation in order to draw relevant conclusions.

This report outlines:

- The structure and characteristics of the dataset.
- The transformations applied to the data.
- The analysis of the main variables.
- The visualizations obtained and the interpretation of the results.

Dataset Description

The provided dataset contains 11,865 records and 10 columns, each representing a specific attribute related to service stations and fuel prices.

Data Transformations Performed

To improve the quality of the analysis, the following data cleaning and transformation steps were carried out:

1. **Conversion of Prices to Numeric Values**

   The values for *Gasoline 95 E5 Price* and *Diesel A Price* were initially formatted as text using the European decimal format (e.g., "1,529" instead of "1.529"). These values were converted into a proper numeric format.

2. **Handling of Missing Values**

   Missing values were identified in the fuel price fields and in the Service Type column. Null values were filled with the mean for numerical data or with the label "Unknown" for categorical variables.

3. **Removal of Duplicates**

   A total of 122 duplicate rows were found in the dataset and were removed to avoid bias in the analysis.

Variable Exploration and Statistical Analysis

Histograms were generated to visualize the distribution of fuel prices, along with key statistical measures such as mean, median, and range.

Correlation Analysis Between Variables

To identify potential relationships between the numerical variables in the dataset, a correlation matrix was generated using Pearson's correlation coefficient. Since the dataset contains only two numerical variables (*Gasoline 95 E5 Price* and *Diesel A Price*), their correlation was analyzed to evaluate whether there is any dependency between them.

Interpretation of Results

Price Distribution

The analysis performed with Matplotlib made it possible to visualize the price distribution of *Gasoline 95 E5* and *Diesel A* in the province of Madrid using histograms. Both variables show a unimodal and relatively symmetrical distribution, with a clear peak in the range between €1.50 and €1.60. This indicates that most service stations set prices around these values. Such concentration suggests market standardization, possibly due to regulations or competitive pressure among brands.
Although most prices fall within a narrow range, a few outliers were detected at both ends

of the distribution: lower prices at automated or low-cost stations, and higher prices at conventional service stations, possibly due to additional services or strategic locations.

Correlation Between Fuels

The statistical analysis carried out with AutoViz revealed a significant positive correlation between the price of *Gasoline 95 E5* and *Diesel A*. This means that when the price of one rises or falls, the other tends to follow the same trend. This behavior is expected, as both fuels are influenced by the same structural factors: crude oil costs, taxes, logistics and refining expenses, as well as general economic policies.

Impact of Station Type

AutoViz also enabled the analysis of how categorical variables influence fuel prices, particularly the type of sale and type of service. It was observed that stations with private sales (P) and automated service (A) offer more competitive prices, likely due to lower operational costs. In contrast, stations with assisted service or operated by major brands tend to have slightly higher prices, reflecting a more traditional, customer-focused business model.
This differentiation suggests that management models and automation play a key role in pricing policies, allowing automated stations to capture a significant share of the market thanks to more affordable rates.

Data Quality

During preprocessing, missing values and duplicate rows were identified and properly handled before analysis. Converting prices to a standard numeric format (changing commas to decimal points) was a key step to ensure the integrity of the visualizations and subsequent statistical calculations.

Conclusions

This study has carried out a comprehensive analysis of fuel price data from service stations using Exploratory Data Analysis (EDA) techniques. Various transformations were applied to clean and normalize the information, ensuring data quality and allowing for a more accurate exploration of price trends.

The results show that gasoline and diesel prices present a relatively homogeneous distribution, with limited variation between stations within the same province. However, some stations were identified with significantly higher or lower prices, suggesting the possible influence of factors such as:

• Local competition among service stations.

• Strategic location of the station.

• Differences in pricing policies between commercial brands.

The correlation analysis between *Gasoline 95* and *Diesel A* prices indicates a strong relationship between the two, meaning that any variation in one fuel is likely to be reflected in the other. This behavior is expected, as both fuels are influenced by the same macroeconomic factors, such as:

• Crude oil prices in international markets.

• Refining and transportation costs.

• Government taxes and regulations.

The use of the AutoViz tool was essential in this analysis, as it allowed for the rapid and efficient generation of a detailed report on the structure and quality of the data. Thanks to this tool, it was possible to:

• Detect issues such as missing values and duplicate records.
• Identify the need to convert categorical variables into numeric format.
• Graphically explore price distributions and relationships between variables.

In conclusion, this study provides an initial approach to fuel price analysis in Spain, using a reproducible methodology that can be applied to other regions or expanded with temporal data to identify changing patterns over time.

Future improvements may include:

• Predictive models to anticipate fuel price changes.

• Geospatial analysis to understand station distribution and its impact on price competition.

• Price comparisons between different provinces to detect regional differences in pricing policies.

• Incorporation of demand and consumption data to evaluate how these factors influence pricing decisions.

This analysis represents a key starting point for future research in the energy sector and fuel market dynamics.