Group 10 | Data Visualization

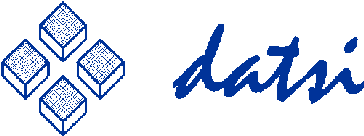
**González Ruiz Alberto**

**Retes Corada Adrián**

**Torrelles Rodríguez Diego Rafael**

DATA SCIENCE MASTER | ETS INGENIEROS INFORMÁTICOS (UPM)

Design of a new interactive data analysis tool



Índice

[Introduction 2](#_Toc148294240)

[Problem characterization in the application domain 2](#_Toc148294241)

[Selection of Data Set: 2](#_Toc148294242)

[Formulated Questions: 3](#_Toc148294243)

[Data and task abstractions 3](#_Toc148294244)

[Data abstractions 3](#_Toc148294245)

[Dataset type: 3](#_Toc148294246)

[Attribute types: 3](#_Toc148294247)

[Cardinality 4](#_Toc148294248)

[Task abstractions 5](#_Toc148294249)

# Introduction

Data visualization has become an essential tool for analyzing and communicating information in a wide range of fields. In this project, we tackle the task of visualizing data related to gas stations in Spain, leveraging a detailed and continually updated dataset obtained from the official portal of the Government of Spain. The methodology followed (Figure 1) in this project is based on a structured approach that spans from dataset selection to the implementation of an interactive application using Shiny, a powerful data analysis tool in R.

Interfaz de usuario gráfica

Descripción generada automáticamente

Figure 1. Design abstraction levels.

In the initial step of our methodology, a careful selection of the dataset was made. From this dataset, a set of key questions were formulated, which will be addressed in this work. To achieve effective answers to these questions, a data and task abstraction approach will be applied, allowing for the proper structuring of data and the definition of specific visualization tasks. Subsequently, the design of a visualization tool will be undertaken, incorporating appropriate visual and interactive elements to effectively address these tasks.

Finally, the implementation of this tool will be carried out in Shiny, enabling analysts and users to interactively explore information about gas stations in Spain, providing a rich and effective data analysis experience.

Throughout this work, we will rigorously follow this methodology to achieve precise, informative data visualization tailored for decision-making, contributing to the understanding and optimization of gas station prices and locations in the Spanish context.

# Problem characterization in the application domain

In this abstraction level we describe specific issues of the application domain and end users involved, such as the problem to solve, user demands and datasets.

## Selection of Data Set:

For this data visualization task, a dataset containing comprehensive information about all the gas stations in Spain has been chosen. This dataset was obtained from the official website of the Government of Spain, specifically at the link "<https://geoportalgasolineras.es/geoportal-instalaciones/DescargarFicheros>". The notable advantage of this file lies in its frequent price updates, with records being refreshed every 30 minutes.

## Formulated Questions:

The following questions have been formulated with the aim of exploring and analyzing the information contained in the dataset:

1. **What is the geographic variation in fuel prices in Spain for each type of fuel?** This question seeks to uncover differences in fuel prices across different regions of Spain, enabling an understanding of geographic trends in gasoline and diesel prices, among others.
2. **What is the average price of different types of fuel in Spain?** This question aims to calculate the average price of various types of fuel throughout the country, providing an overall view of average costs for consumers.
3. **What is the relationship between the prices of gas stations and their location in cities or on highways?** With this question, the objective is to analyze the connection between the location of gas stations, whether in urban settings or on highways, and the prices they offer. This could shed light on how location influences price setting.

These questions will establish a solid framework for the development of data visualizations in the Shiny application, allowing analysts to better explore and understand the dynamics of fuel prices in Spain.

# Data and task abstractions

Goal: translate domain-specific language regarding data to generic

terms:

1. Identify dataset type

2. Identify attribute types

3. Identify cardinality:

• Number of items

• Number of levels of categorical attributes

• Range of quantitative attributes

4. Consider if data transformations are needed/useful:

• Derive, discretize, etc.

## Data abstractions

### Dataset type:

The selected dataset is structured, tabular data in spreadsheet format, often referred to as a "CSV" (Comma-Separated Values) and Excel format. This type of dataset is commonly used for tabular data storage, with rows and columns, making it suitable for structured data analysis.

### Attribute types:

1. **Geographic Information:**

* **Province**: Categorical - Represents the province where the gas station is located.
* **Municipality**: Categorical - Identifies the municipality of the gas station.
* **Locality**: Categorical - Describes the exact locality of the gas station.
* **Postal Code**: Ordinal - Represents the postal code of the gas station's location, which can be considered as an ordinal attribute if postal codes are used to reflect a hierarchy or implicit order based on geographic location.
* **Address**: Categorical - The physical address of the gas station.
* **Longitude and Latitude**: Quantitative - The precise geographic coordinates of the gas station.

1. **Fuel Prices:**

The following attributes represent the prices of different types of fuels at each gas station:

* **Fuel Prices**: Quantitative - Represent the cost of fuel at each gas station.
* **% bioalcohol and % methyl ester**: Quantitative - Percentage of components in the fuel.
* **Prices of liquefied gases, compressed natural gas, liquefied natural gas, and hydrogen**: Quantitative - Prices of alternative fuels.

1. **Gas Station Information:**

• **Sign**: Categorical - The name or sign of the gas station.

**• Sale Type**: Categorical - Describes the type of sale at the gas station.

**• Remarks**: Categorical - Contains additional observations or notes about the gas station.

• **Schedule**: Categorical - The operating hours of the gas station.

**• Service Type**: Categorical - Describes the type of service offered by the gas station.

### Cardinality

* **Province**: Cardinality equal to 52.
* **Municipality**: Cardinality equal to 3432.
* **Locality**: Cardinality equal to 4244.
* **Postal Code**: Cardinality equal to 4544.
* **Address**: High cardinality (11911). Each physical address is unique.
* **Margin**: Low cardinality (3). Different margins are represented by letters such as "D," "I," "N."
* **Longitude and Latitude**: High cardinality (11911). They have unique values.
* **Data Collection**: Cardinality equal to 2480, multiple data collections carried out simultaneously.
* **Fuel Prices**: Cardinality depends on the fuel; there are 437 different prices for gasoline and 481 for diesel.
* **Sign**: Cardinality equal to 4072. Multiple gas station names are identical.
* **Sale Type**: Cardinality equal to 2. Different types of sales, such as "P" (public) or "R" (restricted).
* **Remarks**: Cardinality equal to 2. Different observations or additional notes.
* **Schedule**: Cardinality equal to 1334. Different operating hours.
* **Service Type**: Cardinality equal to 1712. Different types of services offered.

## Task abstractions