Group 10 | Data Visualization

**González Ruiz Alberto**

**Retes Corada Adrián**

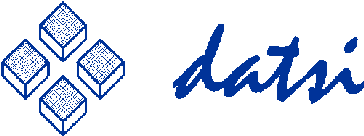
**Torrelles Rodríguez Diego Rafael**

DATA SCIENCE MASTER | ETS INGENIEROS INFORMÁTICOS (UPM)

Date: January 8th, 2023

Design of a new interactive data analysis tool

Una caricatura de una persona

Descripción generada automáticamente con confianza media

Index

[Figure Index 1](#_Toc151687853)

[Table Index 1](#_Toc151687854)

[1. Introduction 1](#_Toc151687855)

[2. Problem characterization in the application domain 2](#_Toc151687856)

[2.1. Selection of Data Set 2](#_Toc151687857)

[2.2. Formulated Questions 3](#_Toc151687858)

[3. Data and task abstractions 4](#_Toc151687859)

[3.1. Data abstractions 4](#_Toc151687860)

[3.1.1. Dataset type: 4](#_Toc151687861)

[3.1.2. Attribute types: 4](#_Toc151687862)

[3.1.3. Attributes cardinality 7](#_Toc151687863)

[3.2. Task abstractions 8](#_Toc151687864)

[3.2.1. Visualization 1 (Price per regions) 8](#_Toc151687865)

[3.2.2. Visualization 2 (Price statistics) 9](#_Toc151687866)

[3.2.3. Visualization 3 (Price per road) 11](#_Toc151687867)

[4. Interaction and visual encoding 12](#_Toc151687868)

[References 12](#_Toc151687869)

# Figure Index

[Figure 1. Design visualization nested levels. 2](#_Toc151687870)

[Figure 2. Fuel prices at a popular station in Spain [1]. 3](#_Toc151687871)

[Figure 3. Spain Provinces Map [2]. 3](#_Toc151687872)

[Figure 4. General fuel types in Spain **¡Error! No se encuentra el origen de la referencia.**. 3](#_Toc151687873)

[Figure 5. Gas station on Spanish highway signal [4]. 4](#_Toc151687874)

[Figure 6. Abstract task, abstract data and views and methods schema. 8](#_Toc151687875)

# Table Index

[Table 1. Summarize of task abstraction for visualization 1. 9](#_Toc151687849)

[Table 2. Summarize of task abstraction for visualization 2. 10](#_Toc151687850)

[Table 2. Summarize of task abstraction for visualization 2. 11](#_Toc151687851)

# 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 visualization nested 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 prices in Spain (Figure 2) 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.



Figure 2. Fuel prices at a popular station in Spain [1].

## 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?**

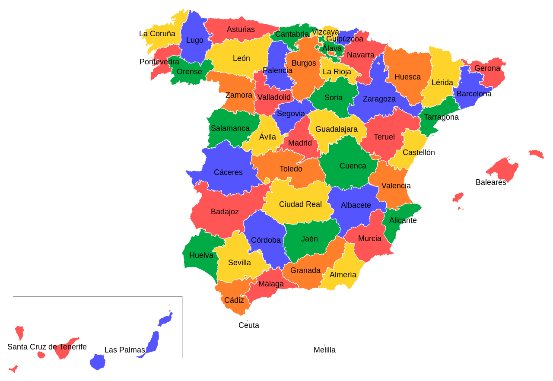


Figure 3. Spain Provinces Map [2].

1. **What is the average price of different types of fuel in Spain?**



Figure 4. General fuel types in Spain **¡Error! No se encuentra el origen de la referencia.**.

1. **What is the relationship between the prices of gas stations and their location in cities or on highways?**



Figure 5. Gas station on Spanish highway signal [4].

Question 1 is important because it 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. As for question 2, this question is the most open; it aims to calculate the statistics to determine the bulk of the price distribution of various types of fuel throughout the country, providing an overall view of average costs for analytics. Finally, the objective of question 3 is to analyze the connection between the road location of gas stations, whether in urban settings or on highways, and the prices they offer. This could shed light on how road location, not geographical location as in the first question, influences price setting.

These questions will establish a solid framework for the development of data visualizations in the Shiny application, allowing analysts to find and answer the problems while exploring and understanding the dynamics of fuel prices in Spain.

# Data and task abstractions

The primary objective is to translate domain-specific language related to data into generic terms. This involves a structured approach:

* Identification of dataset types.
* Recognition of attribute types within the dataset.
* Determination of cardinality, involving considerations such as the number of items, levels of categorical attributes, and the range of quantitative attributes.
* Evaluation of whether data transformations are necessary or beneficial, including processes like derivation and discretization. This systematic process ensures a comprehensive understanding of the data and paves the way for effective analysis.

## 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. It is composed of 11911 instances, each one described by 31 attributes.

### Attribute types:

|  |  |  |  |
| --- | --- | --- | --- |
| Data group | Field Name | Data Type | Description |
| Geographic Information | Province | Categorical | Represents the province where the gas station is located. |
| Geographic Information | Municipality | Categorical | Identifies the municipality of the gas station. |
| Geographic Information | Locality | Categorical | Describes the exact locality of the gas station. |
| Geographic Information | Postal Code | Ordinal | Represents the postal code of the gas station's location, which can be considered as an ordinal attribute if postal codes reflect a hierarchy or implicit order. |
| Geographic Information | Address | Categorical | The physical address of the gas station. |
| Geographic Information | Margin | Categorical | Identifies the exact side of the road. |
| Geographic Information | Longitude | Quantitative | The precise geographic coordinate representing the east-west position of the gas station. |
| Geographic Information | Latitude | Quantitative | The precise geographic coordinate representing the north-south position of the gas station. |
| Data gathering | Timestamp | Categorical | Date (day/month/year) and hour (hour:minutes) of data gathering. |
| Fuel Prices and components | Precio gasolina 95 E5 | Quantitative | Represents the cost of gasoline with 95 octane and 5% ethanol at each gas station. |
| Fuel Prices and components | Precio gasolina 95 E10 | Quantitative | Represents the cost of gasoline with 95 octane and 10% ethanol at each gas station. |
| Fuel Prices and components | Precio gasolina 95 E5 Premium | Quantitative | Represents the cost of premium gasoline with 95 octane and 5% ethanol at each gas station. |
| Fuel Prices and components | Precio gasolina 98 E5 | Quantitative | Represents the cost of gasoline with 98 octane and 5% ethanol at each gas station. |
| Fuel Prices and components | Precio gasolina 98 E10 | Quantitative | Represents the cost of gasoline with 98 octane and 10% ethanol at each gas station. |
| Fuel Prices and components | Precio gasóleo A | Quantitative | Represents the cost of standard diesel at each gas station. |
| Fuel Prices and components | Precio gasóleo Premium | Quantitative | Represents the cost of premium diesel at each gas station. |
| Fuel Prices and components | Precio gasóleo B | Quantitative | Represents the cost of biodiesel at each gas station. |
| Fuel Prices and components | Precio gasóleo C | Quantitative | Represents the cost of a different type of diesel at each gas station. |
| Fuel Prices and components | Precio bioethanol | Quantitative | Represents the cost of bioethanol at each gas station. |
| Fuel Prices and components | Precio biodiésel | Quantitative | Represents the cost of biodiesel at each gas station. |
| Fuel Prices and components | % bioalcohol | Quantitative | Percentage of bioalcohol in the fuel. |
| Fuel Prices and components | % methyl ester | Quantitative | Percentage of methyl ester in the fuel. |
| Fuel Prices and components | Prices of liquefied gases | Quantitative | Prices of liquefied gases at each gas station. |
| Fuel Prices and components | Prices of compressed natural gas | Quantitative | Prices of compressed natural gas at each gas station. |
| Fuel Prices and components | Prices of liquefied natural gas | Quantitative | Prices of liquefied natural gas at each gas station. |
| Fuel Prices and components | Prices of hydrogen | Quantitative | Prices of hydrogen at each gas station. |
| Gas Station Information | Sign | Categorical | The name or sign of the gas station. |
| Fuel Prices and components | Sale Type | Categorical | Describes the type of sale at the gas station. |
| Fuel Prices and components | Remarks | Categorical | Contains additional observations or notes about the gas station. |
| Fuel Prices and components | Schedule | Categorical | The operating hours of the gas station. |
| Fuel Prices and components | Service Type | Categorical | Describes the type of service offered by the gas station. |

### Attributes 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.

The data was enhanced by incorporating shape data that includes the boundaries of the various provinces in Spain, along with additional information such as their identifier and the corresponding borough. In the subsequent sections, we will elaborate on the distinct transformations applied to the data to implement various idioms.

## Task abstractions

The methodology followed to perform task abstraction encompasses a total of 3 questions (Figure 6):

1. What?: This is related to search and query, involving the identification and definition of key elements.
2. Why?: This is related to questions and targets, aiming to understand the purpose and objectives.
3. How?: This is related to design choices, focusing on the methods and approaches used in the abstraction process

Diagrama

Descripción generada automáticamente

Figure 6. Abstract task, abstract data and views and methods schema.

The idea is to identify *why?* it is needed, *what?* can be achieved, and *how?* it can be implemented with a more general visualization before moving on to a specific case.

In this section, a series of questions related to task abstractions will be presented. For each question, a brief description is provided, and below these questions, the tables corresponding to the abstract tasks are presented.

### Visualization 1 (Price per regions)

**Why is visualization being used?**

The visualization is used to present clear and effective information about fuel prices in different provinces of Spain, targeting drivers and fuel distributors. Users can find new knowledge about geographical patterns and mean values in prices that weren’t presented before in the original dataset. That makes "discover" the main action.

**What kind of search is performed based on whether the target and the location are known or not?**

Users initiate a search action to locate specific information within the visualization. The patterns are predetermined, they are aware that they are seeking a relation with the price level (target), such as finding an expensive, medium, or cheap regions group. However, users are uncertain about the exact location for identifying the pattern (where it will be discovered in the visualization). In this case, the search will involve "locating".

**What kind of query is made based on the results of the previous question?**

The three types of queries are made:

* Summarize: Since users have a full view of the map with all provinces and their respective fuel prices, the ‘summarize’ action fits well. Users can use this action to obtain a general overview of fuel prices in all provinces in a concise and understandable manner. It allows them to grasp the data overview without the need to search for specific details or compare multiple targets.
* Identify: The users can Look for a specific province to know the price.
* Compare: The users can check differences in the prices of the regions.

The primary emphasis will center around comparing the outcomes across various regions. This comparative analysis aims to highlight distinctions in fuel prices, providing users with a valuable understanding of the regional variations. By focusing on the comparison of results, users can discern trends, identify disparities, and draw insights into the factors influencing pricing differences among different provinces. In this case, the query will involve "comparing".

**What are the different task targets?**

The known target that the user is trying to find within an unknown location is the price level relation between provinces (or regions), such as identifying a high price shared by two provinces.

**How is going to be performed?**

The user is required to navigate within the visualization, selecting a particular fuel type, and choosing provinces for the purpose of comparing prices.

**Summarize**

Task abstraction for this question would be "discover, locate, and compare prices mean between province zones" (Table 1).

Table 1. Summarize of task abstraction for visualization 1.

|  |  |  |
| --- | --- | --- |
| **What?\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_** | **Why?\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_** | **How?\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_** |
| * Locate * Compare | * Actions: * Discover * Targets: * Price level relationship between provinces. | * Navigate * Select (Fuel type) . * Select (Province or Provinces). |

**Beneficial transformation**

To effectively implement the concept associated with this visualization, it would be advisable to reduce the dimensionality from the 32 variables to the four primary fuels (Gasolina 95, Gasolina 98, Gasoleo A, and Gasoleo A Plus), along with the fuel station coordinates. Subsequently, performing a join with a mapping library containing the coordinates per province would allow the determination of the province in which each fuel station is located. This step is crucial before calculating the average price per province.

### Visualization 2 (Price statistics)

**Why is visualization being used?**

This visualization serves as an effective means of knowing the bulk of the distribution of fuel prices in Spain. The main objective is to find key statistical metrics, such as the mean, median, quartiles and outliers, providing a clear, concise, and insightful answer to questions about the average price of various types of fuel. The violin plot facilitates comparison, highlights trends, engages the audience, and supports decision making by providing a visual representation of the concentration and distribution of prices in different fuel categories. It is a powerful tool for discerning where the bulk of prices reside and gaining a comprehensive understanding of the overall price distribution picture. This makes the main action "discover".

**What kind of search is performed based on whether the target and the location are known or not?**

In this case, determining the pattern involves considering various metrics, as mentioned in the preceding question. Since the specific target (type of statistic) is not known, the user needs to identify where to locate the statistic. Therefore, employing a comprehensive visualization displaying all relevant statistics would be beneficial to accommodate the unknown nature of the target. This approach allows the user to explore and identify the most relevant statistic for their analysis within the shared visualization. This makes the search type "Browse".

**What kind of query is made based on the results of the previous question?**

The two following types of queries are made:

* Identify: The users can Look for a specific province to know the price.
* Compare: The users can check differences in the prices of the regions.

In this case, it will be more useful to choose a specific statistic and understand it than to make comparisons between statistics, which would come later. Therefore, the main query will be "identifying".

**What are the different task targets?**

The targets focus on price statistics, primarily emphasizing factors such as the maximum, minimum, median, and density. Understanding these statistical measures provides valuable insights into the pricing dynamics and distribution, offering a comprehensive perspective for analytical purposes.

**How is going to be performed?**

To carry out the process, it is necessary for the user analyst to identify the appropriate statistic in the corresponding fuel category and retrieve the statistical value following the legend and understanding its functioning.

**Summarize**

Task abstraction for this question would be "discover, browse, and identify price statistics" (Table 2).

Table 2. Summarize of task abstraction for visualization 2.

|  |  |  |
| --- | --- | --- |
| **What?\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_** | **Why?\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_** | **How?\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_** |
| * Browse * Identify | * Actions * Discover * Targets: * Price statistics. | * Find statistic. * Retrieve statistic value. |

**Beneficial transformation**

To perform the idiom corresponding to this visualization, it would be beneficial to reduce the dimensionality from the 32 variables to the four main fuels (Gasolina 95, Gasolina 98, Gasoleo A, and Gasoleo A Plus). Following this, performing the statistical calculations based on each fuel type would be a sound approach.

### Visualization 3 (Price per road)

**Why is visualization being used?**

The visualization is used to explore and understand the relationship between gas station prices and the type of road, whether in urban environments or on highways. It allows for the analysis of patterns and trends that may emerge when examining how fuel prices vary in different locations, but with a focus on two types of roads, unlike the first visualization. By graphically representing this data, the visualization facilitates the identification of previously unknown correlations between the location of gas stations and price levels, providing valuable insights into the price dynamics based on geographic location. Therefore, the main action will be "discover."

**What kind of search is performed based on whether the target and the location are known or not?**

Because the visualization is intended to be simple, with two main types (urban or highway) and four subtypes (fuel types), the user needs to be familiar with the pattern's location, as it will quickly stand out with a value clearly distinguishable from the rest. There are far fewer cases than in the first visualization, and the type of visualization used will be simpler. Additionally, as in the first visualization, the sought pattern is also known (the level of the price mean). Therefore, the suitable search type will be "lookup."

**What kind of query is made based on the results of the previous question?**

The primary aim of these visualizations is to be as simple as possible, incorporating different road types, fuel categories, and their prices to easily identify outlier values. This means that users can compare prices both within and outside the city for every fuel type they search. Consequently, the most suitable query type for this purpose is "compare".

**What are the different task targets?**

* Average fuel price: This is calculated as the mean of all prices within a specific region.
* The region: It serves to define the boundaries of information areas for price data.

**How is going to be performed?**

* To perform the process is necessary to determine the encode, manipulations, facets, and reductions in the process.
* The primary steps in creating a visualization include data derivation, encoding, and annotation.

**Summarize**

Task abstraction for this question would be "discover, lookup, and compare price values between fuel and road types" (Table 3).

Table 3. Summarize of task abstraction for visualization 2.

|  |  |  |
| --- | --- | --- |
| **What?\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_** | **Why?\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_** | **How?\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_** |
| * lookup * Identify | * Actions * Discover * Targets: * Price level relationship between fuel and roads. | * Find fuel and road types. * Retrieve mean values. |

**Beneficial transformation**

To perform the task corresponding to this visualization, it would be beneficial to create a column named "en\_autovía" that will take the value yes if the gas station has the word "autovía" or "autopista" in its address. Subsequently, reduce the dimensionality from the 32 variables to the four main fuels (Gasolina 95, Gasolina 98, Gasoleo A, and Gasoleo A Plus), and the "en\_autovía" variable. Finally, calculate the average prices based on the type of fuel and the presence on the highway or urban location.

# Interaction and visual encoding

# References

1. <https://www.lavanguardia.com/motor/rankings/20201002/33621/son-gasolineras-mas-baratas-espana-ocu.html>
2. <https://commons.wikimedia.org/wiki/File:Provinces_of_Spain.svg>
3. <https://www.malagaweb.com/espanol/blog/wp-content/uploads/cepsa-repsol-etiquetas.jpg>
4. <https://n332.es/wp-content/uploads/2022/03/Finding-Fuel-on-the-Motorway.jpeg>