Group 10 | Data Visualization

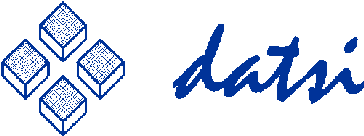
**González Ruiz Alberto**

**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

Diagrama

Descripción generada automáticamente

### Visualization 1

**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 explore and discover geographical patterns in prices, customize the visualization by selecting the desired type of fuel, and obtain additional information through pop-ups displaying the name of the province and the average price.

* Consume (Present): The visualization is used to present information to drivers and fuel distributors. A map of Spain is displayed, with each province colored according to the average price of the selected type of fuel. Visual presentations are valuable in clearly communicating differences in fuel prices across various provinces. This is relevant for helping drivers make informed travel decisions and aiding fuel distributors in understanding price trends that may impact their business.
* Consume (Discover): In the "consume" category, the visualization is employed for drivers and fuel distributors to discover patterns and trends in fuel prices in different provinces. Users can explore the map to find information about geographical variations in prices and uncover differences that might not be evident in raw data tables. This is especially relevant for helping drivers find regions with more economical prices and enabling fuel distributors to make informed decisions about pricing and distribution strategies.
* Produce (Annotate, Record): In the project, users have the capability to perform two "produce" actions:
  + Annotate: Users can interact with the visualization to obtain additional information about the provinces. Clicking on a province on the map displays a pop-up that includes the name of the province and the average price of the selected type of fuel. This equates to "annotating" the visualization with additional information, allowing users to obtain specific details about prices and the locations of provinces.
  + Record: Users can use a selector to choose the type of fuel they want to visualize on the map. This is akin to "recording" their choice of fuel to customize the visualization according to their needs and preferences.

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

* Search: Users perform a search action to find specific information within the visualization. For example, they may search for the average price of a particular type of fuel in a specific province.
* Lookup (Detailed Search):
  + Specific Users: The "Lookup" action is particularly relevant to users with a specific need for obtaining detailed information about a particular province or region. This could include drivers who want to know fuel prices in the province where they live or plan to travel, as well as fuel distributors who need precise data on prices in their operational areas.
  + Precise Details: By using "Lookup," users can select a specific province or region and obtain precise details on the average prices of different types of fuel in that specific location. This allows them to make informed decisions based on concrete data and tailor their actions based on the detailed information they acquire.

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

* 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.

**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?**

1. Encode (Map): Decide how you will represent the average fuel price on the map.
2. Encode (Define Visual marks and channels): Select colors as the "visual marks" and the use of colors as the "channel" to communicate information about fuel prices.
3. Encode (Use): Load the Spain metadata map onto which will be applied colors to indicate price ranges for conveying information about fuel prices.
4. Manipulate (Select): The user select the fuel which he/she is looking for.
5. Reduce (Filter): Filter the dataset to retain only the relevant records containing the prices of the selected fuel from gas stations and their associated province information.
6. Reduce (Aggregate) : Once the data is filtered, perform a calculation that allows to obtain the average fuel price for each province. This can be developed by joining the data from the gas stations with the data from a created provinces map library and then calculating the price averages for each province.
7. Facet (Superimpose): overlay additional elements, incorporating a pop-up feature, that displays the province's name and its average price, and the legend for the colors.
8. Reduce (Embed): Display the generated visualization within the Shiny window to incorporate the map into the overall presentation. This allows for a seamless integration of the map with other content, providing a comprehensive view of the data and its geographical context.

|  |  |  |
| --- | --- | --- |
| **What?\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_** | **Why?\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_** | **How?\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_** |
| * Choropleth map | * Actions * Present/Discover * Lookup * Summarize * Targets: * Average fuel price * Map region. | * Map * Visual Mark and channels * Use * Select * Filter * Aggregate * Superimpose * Embed |

### Visualization 2

**Why is visualization being used?**

This visualization is used to get an effective approach to provide a clear, concise, and insightful answer to the question about the average price of different types of fuel in Spain. It allows for easy comparison, highlights trends, engages the audience, supports decision-making and other things.

* Consume Present)

Visualization involves introducing the topic, presenting the data visually, providing context and analysis, this could help in all kinds of decision-making and planning.

* Consume(Discover)

This visualization provides the foundation for creating effective visualizations that will tell a structure data and allow us to communicate the average prices of all fuels in Spain in the clearest way possible, ensuring that the information is processed by users and third parties.

* Produce(Derive)

In the context of our inquiry into the average prices of various fuel types in Spain, this phase involves the generation of new data from existing sources, which could mean modifying attributes or datasets. It prompts us to consider whether we should view the data in its current state or apply alterations for better insights.

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

The "Lookup" action is particularly relevant to users with a specific need for obtaining detailed information about the average prices of different kind of fuels, being able to use the visualization for planning and making different decisions about many things.

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

A summarization query aims to provide an overview or summary of all targets in the dataset. It offers a high-level understanding of the information, which in this case is what we want to achive by obtaining the average price of the fuels in spain.

**What are the different task targets?**

* Average fuel price: This is calculated as the mean of all prices within a specific region.

**How is going to be performed?**

1. Encode(align): The visualization will contain an align graph to specify the average off fuels.
2. Reduce(filter): This visualization employs mathematical operations to condense the dataset's information, resulting in a reduction of its original size, obtaining one value for each column on the dataset.

### Visualization 3

**Why is visualization being used?**   
The use of visualization is primarily aligned with “Consume” level, more specifically with the "Discover" level, which includes the options to "Generate" and "Verify."

"Generate" refers to using visualization to find new knowledge, which matches the goal of uncovering the relationship between gas station prices and their location in cities or on highways. "Verify" is also relevant in this case because, as part of the discovery process, visualization can be used to verify existing hypotheses or test assumptions regarding the relationship between gas station prices and location.

“Present” and “Enjoy” aren’t the primary actions because these visualizations don’t have the purpose of presenting information to a third party nor for the enjoyment of its users.

Since "Produce" refers to the creation of visual content or artifacts as the end goal, for this case, the primary objective is not simply to produce visual content for its own sake, but rather to use visualization as a tool to gain new knowledge and insights. If, for example, the end goal of the visualization process was to create reports or documents for a wider audience, this action would become more relevant.

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

In this case, since the visualization is a simple bar char where you will have the different types of fuels and their prices in and out of a city, almost all the users will be performing a “Lookout” search mainly because of the simplicity of the visualization even though they may be “Browse” searches since a user may not know a specific type of fuel. Since the options in and out of city are straightforward, there will be almost any “Locate” and “Explore” types of searches.

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

The kind of query that is made is a “Compare” query since, as mentioned before, the main idea of these visualization is a bar char with different fuels and their price in and out of cities meaning that for every fuel the users search for, they will be able to compare the prices in and out of the city.