CGG – Data Geoscientist Graduate Position. Task 1 24.02.2023

What is the value of this data?

The value of this data is in its ability to provide insights into the characteristics of different geological formations. The data could be used to map out an area with an emphasis on its porosity, density and permeability. It could be used to map out , specifically related to porosity, density, permeability, and other relevant factors. This information is important for oil and gas exploration, as well as other types of geological studies.

How would it be used?

The data can be used by geoscientist toe analyse the properties of a specific geological formations. This could be used to survey an area and make informed decisions about resource extraction or other applications.

There are several examples of where it could be used:

Petrophysical analysis: This data contains various petrophysical properties such as porosity, permeability, density, and saturation, which are crucial in analysing the physical properties of rocks and predicting their behaviour under different conditions.

Reservoir modelling: This data could be used to build a 3D geological model of the reservoir, which could be used for reservoir simulation studies.

Well completion and stimulation design: The permeability, porosity, and saturation data could be used to design an optimal completion or stimulation strategy for the well.

Formation evaluation: The data could be used to evaluate the properties of the formation and determine the lithology, cementation, and other properties. This could be used by industry for digging or to correlate different formations, identify potential reservoirs, and evaluate the potential for hydrocarbon production or metals production.

What other datatypes could be used to cross validate the values that you see here?

1. geological maps and surveys
2. seismic data
3. rock strength (RQD)
4. drilling log data
5. production data
6. well completion

All of these example when compounded with the current information will give a better understand of the geological subsurface.

<u>What meta-data is missing that if present could provide extra context, accuracy and value to the data points?</u>

The main additional metadata which would be most useful is the specific location and orientation of the datapoints. This would allow for the information to be mapped out to a location, being able to build a 3D model of each well-location. Additionally, the purpose of each well would be ideal so that the data analysis could focus on the information needed to maximise the output of the intended purpose.

Notes and thought processes through transformation:

Process of looking at the data:

First I looked at the data visually. These are the problems that I noticed.

1. Many of the information columns are empty – Possible place to decrease size of file
2. The language of the file is in Portuguese. Could lead to issues with character encoding when writing/reading the file
3. There are gaps in the data.
   a. Some of these gaps are filled with –, 0, or blank. Information voids should be standardly filled
   b. Where applicable, an mean average can be applied to blank spaces to fill in data without introducing too much error
4. Decimals are indicated by commas instead of decimal points in some cases. The documents will be standardised into using commas.
5. The data is broken up vertically with the headings repeated on row 1410. This row also has a sample ID number of 1508 meaning that the sample ID is given in this document and not anywhere else.
6. There is duplicate data as shown by:

1479	6037915	VH1	-	4711	12	2,67	0,251	4100
1509	6037914	VH1	-	4711	12	2,67	0,251	4100

   a. It seems that some "file_id" contain the same information. Therefore the sample_id and file_id was left out of the drop_duplication function.

Statistical Analysis:

To perform a statistical analysis, I turned all the values of the columns to float values from string. Then I created a box-plot separated by the well-name. Here I was able to see if there were any out-liers. Once I saw there were, I went into the data and found out why. This is what I found:

In profundidade, I was able to see that there are numbers > 100,000 which could mean a multitude of options:

1. Human error on input and should be void
2. Decimal place is missing and needs to be divided by 100
3. Rows were scrambled during process

I found that the problem is most like #2 so I created a function to add a decimal if the number was too absurdly high.

For well Q it is too difficult to see whether the porosity should be divided or if the inputs are just wrong. Maybe there as something wrong with the instruments. For this reason I have decided to void the data for well Q However after dividing by 100 there is less data. This could be a place to look in the future to reduce error.