CGG Graduate Data Geoscientist – 2023 Programme – Technical

Home Assignment

February, 2023

- *Submitted by* -

LAWAL OPEYEMI OLAJIDE

Task 1 Solution

- **Transform the data using whatever method you see fit and map to the appropriate database field.**

  The method I adopted for the data wrangling and data cleaning of this project is **Python tools**. I used Python tools because I feel it has large number of libraries that can be used to clean data from a range of sources and formats. Python is easy to use and it is also reusable. See attached named **Lawal_Opeyemi_Scripts.ipynb** for the scripts used to clean the data.

- **Think about erroneous or data in error. Use whatever approaches you feel appropriate to identify potentially spurious data.**

  Inconsistent Formatting: the depth, porosity, density and permeability column has some values separated by commas instead of decimal points.

  Extreme values: some of the porosity and permeability values are extremely low, e.g., sample ID 483 – 500 has negative porosity, also some other values contain zero porosity, density and permeability which is unlikely to occur in natural reservoirs, and might be a result of measurement error or data entry mistake.

  Duplicates: Sample ID 1508 contains duplicate of the database field and must be removed.

  Outliers: In some wells, the porosity and permeability value are much higher than the other values in the same well, likewise some are much lower than the other values in the same well which could be an outlier or a result of a different measurement technique.

Missing Values: there are many blank fields in some of the column e.g. Pressão (psid), Saturação (%), Método etc. Some fields like Número do testo contains dashes instead of numerical values. All this could be a sign of incomplete data.

One of the approaches I used to identify spurious data is **Visual inspection**. I look at the data in the spreadsheet and checked for any values that seem out of place or inconsistent for example, numbers separated by commas.

- **If you identify errors in the dataset, can you determine the source or cause of the errors?**

  Lots of errors were identified in the data. However, I cannot determine the source or cause of these errors, it is possible that the errors were introduced during data entry, or formatting or may have originated from the equipment used to collect the data.

- **Are there any errors that you can, correct? How would you correct them?**

  Yes

  1. **Inconsistent error:** The comma could be replaced with a dot to ensure consistency in the decimal separator format.
  2. **Missing values:** The missing values could be investigated further, and either filled in or marked as missing data.
  3. **Extreme values:** The porosity and permeability of extremely low and high values could be reviewed for accuracy, and if necessary, corrected or removed from the dataset.
  4. **Outliers:** Wells that have outliers in their porosity and permeability values could be investigated further to determine if it is an outlier or a valid measurement, and if necessary, removed or corrected.

- **Please provide documentation on your methodology**

  See attached named **Lawal_Opeyemi_Scripts.ipynb** for comprehensive documentation of the methods I adopted. I followed the steps below:

  1. I imported Table 1 into a Python environment using the Pandas library, which allows for easy manipulation of data.
  2. I mapped the fields in Table1 with the database fields in Table2.
  3. I corrected any errors in the dataset based on my expertise and understanding. For example, I corrected a typo in depth, porosity, density and permeability column by changing the commas to decimal point, I replaced all negatives and zero values in depth, porosity, density and permeability column with missing data because it is impossible to have zero value for such parameters. I also replaced dashes with missing data.
  4. I replaced missing data and blanks with the value "-99.25".
  5. I confirmed that there were no longer missing observations in the data.
  6. Then, I created a plot of porosity vs depth on each of the wells to help identify potential hydrocarbon reservoirs and identify lithological changes and zones of interest.

Additional questions

- **What is the value of this data?**

  The value of this data lies in its ability to help us understand patterns, trends and relationships which can then be used to extract insights and make informed decisions about exploration (to identify potential areas for exploration).

- **How would it be used?**

  The data can be used to study rock and sediment cores collected from boreholes, to better understand the geology of the subsurface. See attached **Lawal_Opeyemi_Scripts.ipynb** where it shows a plot of porosity vs depth on each well.

  The data can also be used to develop predictive models that can forecast the properties of a subsurface.

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

  Other data types that can be used to validate the values of this data include:

  Well log data: By comparing core data to well log data, geoscientists can validate the accuracy of core measurements. Well log data can help provide continuous measurements of subsurface properties, such as porosity, permeability, and lithology.

  Seismic data: By comparing core data to seismic data, geoscientist can identify the spatial distribution of geological features such as faults or salt domes and validate their interpretation.

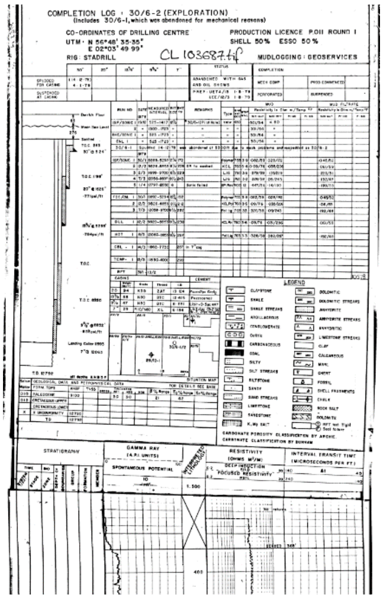
  Geochemical data: By comparing core data to geochemical data, geoscientist can better understand the origin and evolution of rock formation and validate their interpretations.

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

  The meta-data that I feel is missing from the provided data is the **location of the wells** and the **formation names**. The location of the wells is important for understanding the regional geology and comparing the properties of different formations. The formation names can also provide additional context for the samples and help us identify different geological units.

Task 2: Image Classification task Solution

- **Can you identify the data/document type for each image?**



COMPLETION LOG · 30/6-2 (EXPLORATION)
(Includes 30/6-1, which was abandoned for mechanical reasons)

CO-ORDINATES OF DRILLING CENTRE
UTM · N 56°48' 35·35"
E 02°03' 49·99"
RIG: STADRILL

CL 103687.tif

PRODUCTION LICENCE P.011 ROUND I
SHELL 50% ESSO 50%
MUDLOGGING: GEOSERVICES

Image_1

Image_1 is a completion log that is created after a well has been drilled and completed.

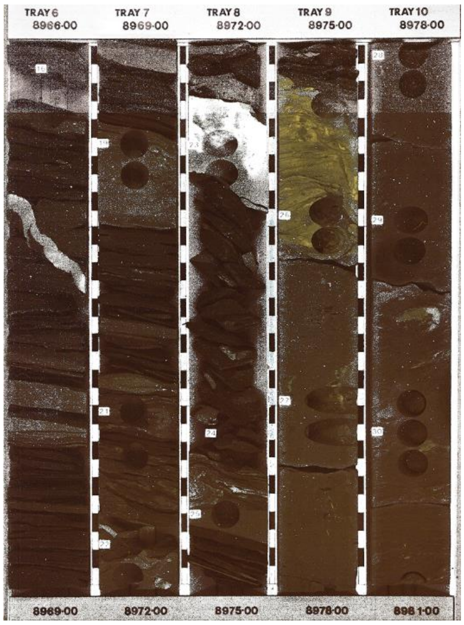


Image_2

Image_2 is an image of cores viewed under ultraviolet light showing uniform bright gold fluorescence in tray 9 (possibly oil)

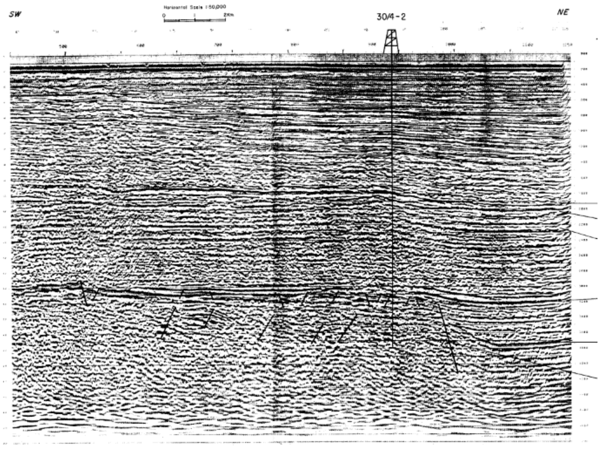


Image_3

Image_3 is a seismic section showing a well passing through a fractured zone.

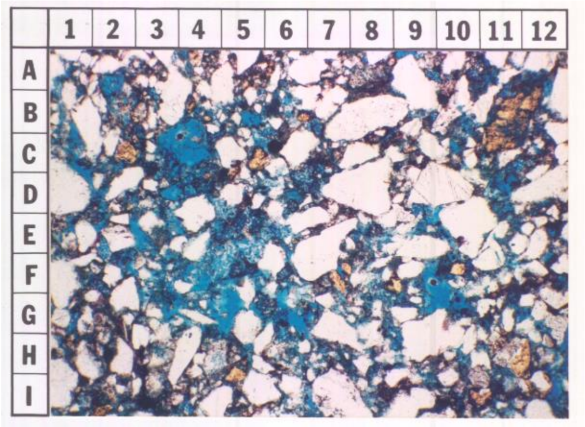


Image_4

Image_4 is a thin section photomicrograph showing oil stains.

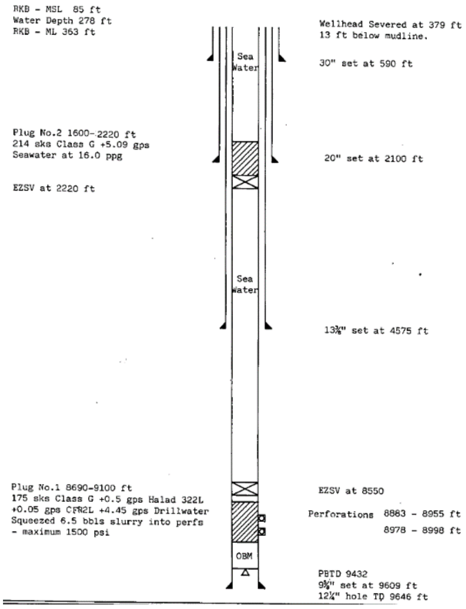


Image_5

Image_5 is a cross section of drill hole to test variations in rock composition with depth

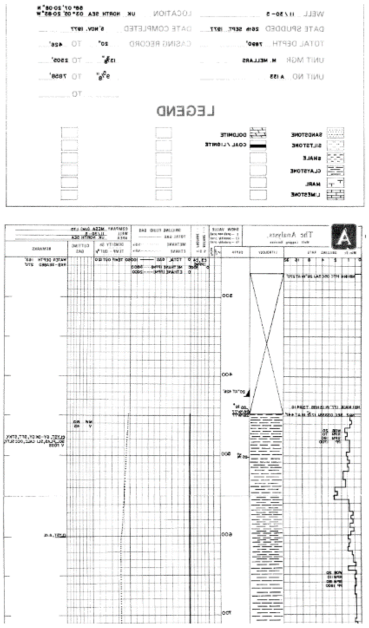


Image_6

Image_6 is a borehole log that can be used to correlate different sections of a well.

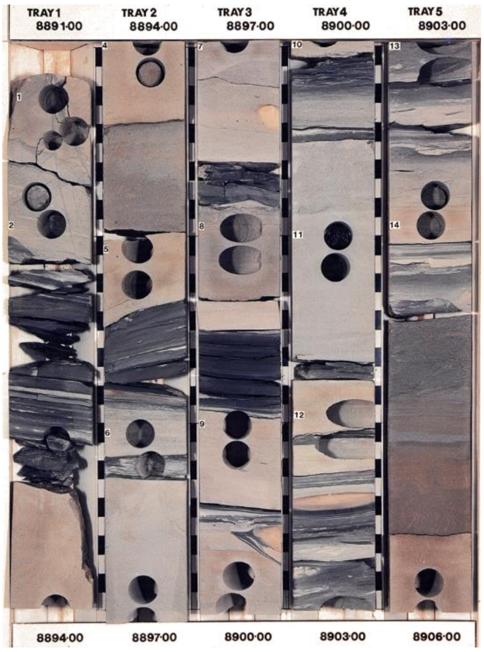


Image_7

Image_7 is a high-resolution imagery of microfacies similar to Image_2 but viewed under white light.

- **What metadata do you feel is most important for each image?**

  I feel the most important metadata for each of this image is the **Sample Location**. Sample Location can provide valuable information about the subsurface geology and help geoscientists understand the characteristics of the reservoir.

- **What additional data could be extracted from the images to enrich subsurface data sets?**

  Some additional data that can be extracted from these images to enrich subsurface data sets include:

  <u>Mineralogy</u>: Microfacies in thin sections can provide information about the mineralogy of the rocks, which can be used to identify the lithology and diagenetic history of the rocks. This information can be used to build a petrophysical model that can help in estimating the reservoir properties such as porosity and permeability.

Facies data: Based on the features observed in the thin sections, such as grain size, shape, sorting, and sedimentary structures, the depositional environment and facies can be interpreted. This information can be used to identify the distribution of different lithofacies and understand their depositional history.

Structural data: By analyzing the seismic sections, the structure and geometry of the subsurface can be interpreted, including faults, folds, and unconformities. This information can be used to identify structural traps and migration pathways for hydrocarbons.

Stratigraphic data: Seismic sections can provide information about the thickness, geometry, and lateral extent of the different stratigraphic units. This information can be used to build a sequence stratigraphic framework that can help in identifying the potential reservoir targets.

Rock properties data: Borehole logs can provide information about the physical and mechanical properties of the rocks and sediments encountered. This information can be used to determine the strength, permeability and other properties of the subsurface.