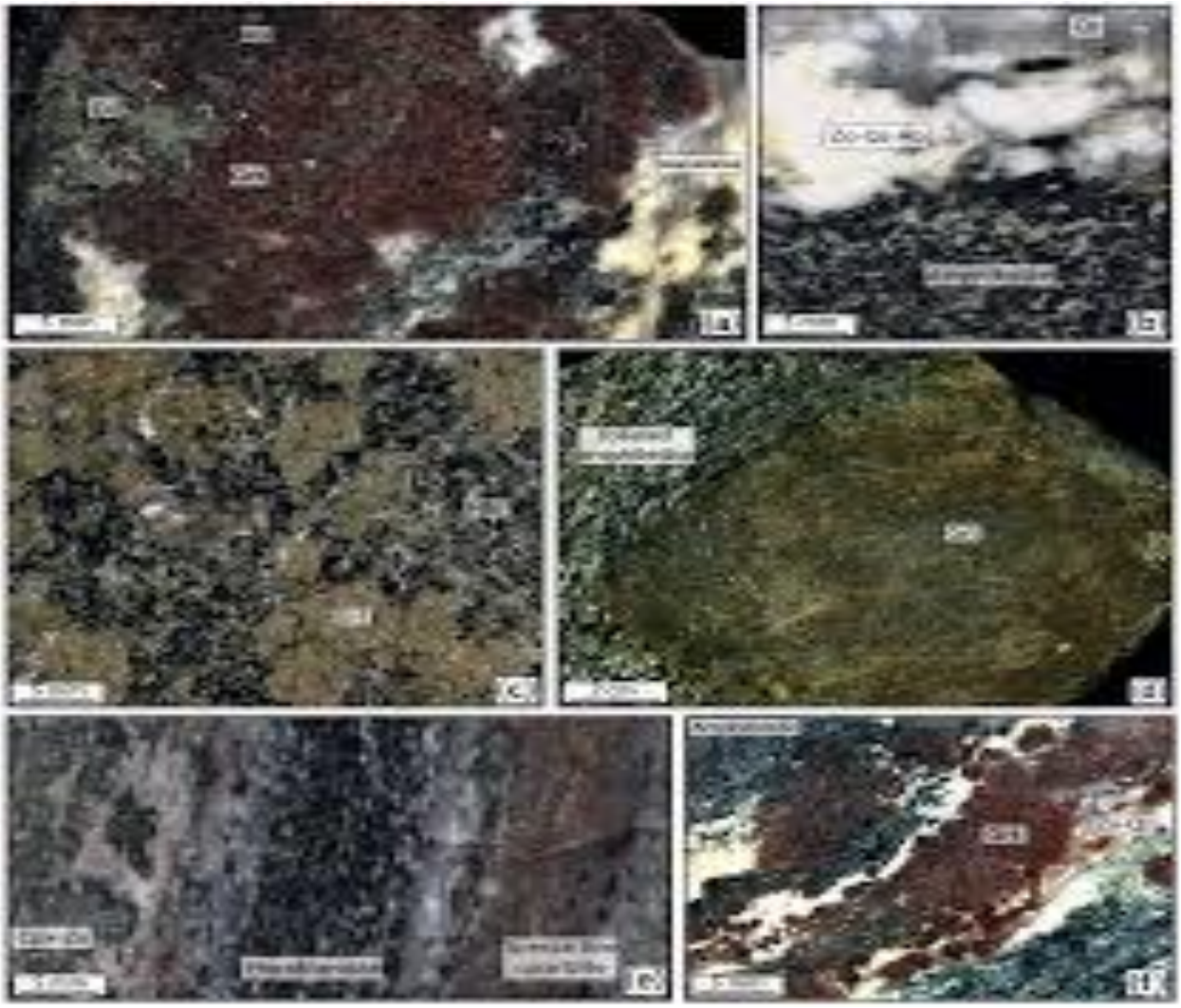
Date : 7/22/2020

Submitted to : Professor Karin Block

Done By: Natasha Arokium

Class : Capstone project (Data Science And Engineering)



Similarity in the chemical composition of the world's ocean floor

EarthChem Project Summary

The objective behind this research project conducted by the National Science Foundation with reference to the collection, identify, and classification of samples of igneous rocks found on the earth's ocean floor, seeks to find a global perspective of the chemical composition of the Earth.

These rock samples are collected from several Geologists, analyzed and characterized, and are kept in the PetDB database. Since these samples are collected by different researchers, at different times, and from different parts of the world's ocean floor, there appear several varying naming conventions used for labelling and identifying these specimens. Our task is to write a script that loops through the entire collection of samples, searching for similarities using certain features to classify them collectively.

One of the proposed methods for approaching a solution to this project, is to first perform data analysis and data cleaning processes after linking and compiling each researcher's report to ascertain a more refined, intuitive database structure. We will also attempt to identify whether specimen names are similar or different by comparing additional features found in the observations/rows (Researchers name, Latitude and Longitude data, feature_of_interest_type_name, feature_of_interest_cv_name, etc).

The Linear regression classification machine Learning algorithm will also be experimented with to predict whether samples collected from all the researchers combined, are similar or whether they are different with respect to their naming conventions.

Another appropriate method would be to apply the KNN (K nearest neighbors) machine learning algorithm approach, which allows for the classifications of similar rock specimens/samples by using the Euclidian distance metric to group samples that are closely similar in characteristics.