Russian Federation: A Discriminant Analysis on the Paleoclimatology in Russia

Julia Cuellar

Summer 2021

https://github.com/BVUjac8/Russian-Federation.git

Any surprises from your domain from these data?

A dataframe with 91 observations and 9 variables.

A dataframe with 99 observations and 19 variables.

A dataframe with 3,669 observations and 6 variables.

Undoubtedly yes. Working in JSON file formats in the programming language, R, is quite difficult compared to wielding it in Python code. While the original files were all in txt file formats to where I just changed one to a JSON and the other to a xls, the JSON one was the more difficult to work with. However, through extensive trial and error, I was able to read the JSON file and transform it into a dataframe. Other than that, the domain from these datasets are just in tune for the\is discriminant analysis

The dataset is what you thought it was?

https://www.ncei.noaa.gov/pub/data/paleo/paleolimnology/asia/russia/elgygytgyn2012.txt

https://www.ncei.noaa.gov/pub/data/paleo/paleolimnology/asia/russia/baikal2006.txt

These datasets were almost exactly what I expected. While performing cleaning of these datasets, I once again noticed that were no nulls, but just NA's. However, renaming column names was a crucial function to consider. My research question is determining what the common column variable is to merge all datasets into a database, thus, renaming a column to an akin name is vital to go through this problem.

Have you had to adjust your approach or research questions?

- What is the common column variable that can be made from these 3 different file formats (txt, JSON, & xls) to merge into a database?

No, I have not. The biggest hurdle was reading different file formats in a programming language I am trying to get better at. As I mentioned, JSON file formats are not sympatico with R programming as compared to Python code (in which I am at least have more than intermediate knowledge about), hence, the struggle to read the JSON and then change it into a dataframe. After that though, the rest of the EDA was easy to perform such like summary statistics, determining if there is nulls or NA's, and then renaming column name(s).

Is your method working?

Discriminant analysis is a classification technique that utilizes variable measurements on different groups of items to underline points that distinguish the groups. Referencing discriminant analysis from the prior sentence, the common column variable was known as **Age**. After plotting the age column variable from each data set, there were some outliers, yet the range of the age of the two different lakes in the Russian

Federation displays how old or how newly formed the lakes are. Although, for merging the datasets into a database is where the most distress occurred.

What challenges are you having?

As I mentioned previously, putting the three different file formats into a single dataset based off of a common column variable (i.e. Age) and thus into a database was quite difficult. This is my second time working with SQLite, and I am still having strain. Yet, I was finally able to get the expected dimensions for rows and columns in the conjoined dataset.

I am doing all this analysis in R programming; I am familiar with R programming, but is not as efficient with it compared to Python, hence, still need to learn the various packages from the programming language. Ergo, data.table, tidyverse, rjson, readxl, dplyr, and RSQLite for all the various steps I performed on this discriminant analysis.