Russian Federation: A Discriminant Analysis on the Paleoclimatology in Russia

Julia Cuellar

Data Science, Bellevue University

DSC 680: Applied Data Science

Professor Catherine Williams

July 25, 2021

Russian Federation: A Discriminant Analysis on the Paleoclimatology in Russia

ABSTRACT

The Russian Federation is a Eurasian country that is rich in history as well as paleoclimatology (see Appendix A). Russia also has two significant lakes (Lake Baikal and Lake Elgygytgyn) with collected biogenic silica and sediments during the Pleistocene epoch or Quaternary period (see Appendix B). A discriminant analysis will be performed on three different collections of data based off of the paleoclimatology in Russia.

BACKGROUND

Lake Baikal is a rift lake located in Russia, situated in southern Siberia (Wikipedia). It is the deepest, oldest, and largest freshwater lake by volume in the world (Lakepedia). While Lake Elgygytgyn is an impact crater lake located in northeast Siberia (Wikipedia). Compared to Lake Baikal, Lake Elgygytgyn's purpose is for a multinational drilling campaign (Dosecc).

BUSINESS UNDERSTANDING

Both Lake Baikal and Lake Elgygytgyn have a unique establishment of when these bodies of water were formed in the country of Russia. Since both Russian lakes were formed ages ago, biogenic silica as well as sediments are buried under layers of deposits within both lakes. Hence, a discriminant analysis based off a common column variable from three different file formatted data of the two Russian lakes will be performed to understand how separate data can be collaborated into a single dataset which in then transfers into a database.

DATA UNDERSTANDING

Table 1

Paleoclimatology – Baikal.txt dataset

A dataframe with 3,669 observations and 7 variables.

```
##        Hole Core-Sec Int Depth Biosil  Age        X
## 1: BDP-96-2     GC-1  11  0.11   13.1 11.4 876.7123
## 2: BDP-96-2     GC-1  13  0.13   15.7 11.9 876.7123
## 3: BDP-96-2     GC-1  15  0.15   13.5 12.3 876.7123
## 4: BDP-96-2     GC-1  17  0.17    7.7 12.8 876.7123
## 5: BDP-96-2     GC-1  21  0.21    8.8 13.7 876.7123
## 6: BDP-96-2     GC-1  23  0.23    4.5 14.2 876.7123
```

Table 2

Paleoclimatology – Lake.json dataset

A dataframe with 91 observations and 9 variables.

```
##        Age MTWM-degC MTWM-deg-C MTWM-deg+C  PANN-mm PANN--mm PANN-
+mm
## 1 0.000000  8.800000   8.800000   8.800000 255.0000 255.0000 255.0000
## 2 2.596983  8.888037   8.792730   8.992730 253.3380 253.3380 253.3380
## 3 2.929250  9.151988   9.072456   9.272456 259.9207 259.9207 259.9207
## 4 3.479885  8.898796   8.850015   9.050015 258.6479 258.6479 258.6479
```

```
## 5 4.090889  8.927818   8.850605   9.050605 258.6614 258.6614 258.6614

## 6 4.461754  8.896577   8.800574   9.000574 253.3380 253.3380 253.3380

##   Trees & Shrubs Picea

## 1          63.0     0

## 2          74.9     0

## 3          67.2     0

## 4          61.6     0

## 5          71.1     0

## 6          73.8     0
```

Table 3

Paleoclimatology – Lake2.xls dataset

A dataframe with 99 observations and 19 variables.

```
##    Age Picea.sect..Eupicea P..s.g.Haplo.T Larix Betula Alnus Salix Poaceae

## 1 1.50                 0          8.1     0   33.2  36.8   0.8     9.2

## 2 2.27                 0          7.2     0   25.2  37.2   0.6    12.9

## 3 2.62                 0         16.1     0   12.4  38.7   0.6    13.3

## 4 3.14                 0         12.6     0   21.0  38.9   1.0    10.1

## 5 3.46                 0          7.7     0   24.6  43.8   0.8     7.9
```

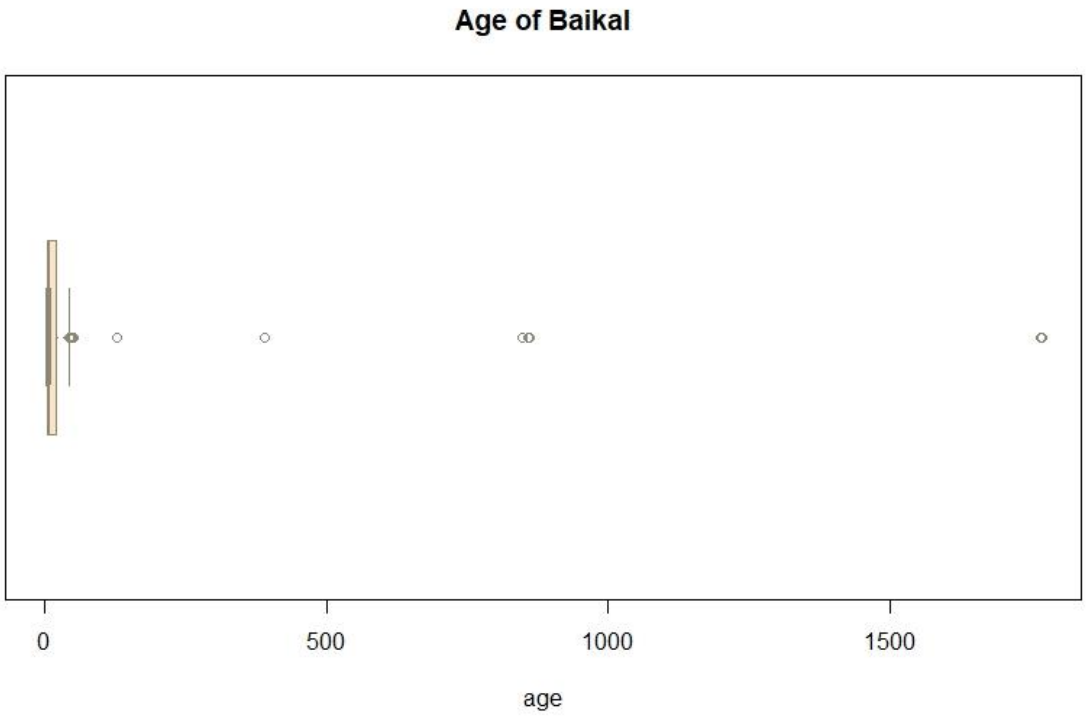
```
## 6 3.65              0        11.5    0  12.4  35.1  0.3   24.2

##   Cyperaceae Artemisia Ericales Cphae Papae Ranae Tha Sax Lyc.a.T Sel.r Sph

## 1        4.9       3.2     0.8   0.8   0.8   0.3 0.0 0.0     0.3   1.9 2.7

## 2        7.2       2.1     3.0   0.0   0.3   0.3 0.0 0.0     0.0   2.4 1.8

## 3        2.3       5.6     1.1   2.8   2.0   0.0 0.3 0.3     0.8   4.5 2.8

## 4        8.9       2.7     1.7   0.0   0.5   0.2 0.0 0.7     0.0   1.7 1.5

## 5        5.8       4.2     1.5   0.2   1.0   0.2 0.0 0.0     0.2   0.8 1.7

## 6        3.4       3.1     0.3   0.9   2.5   0.0 0.3 0.0     1.5   6.2 8.7
```

DATA PREPARATION

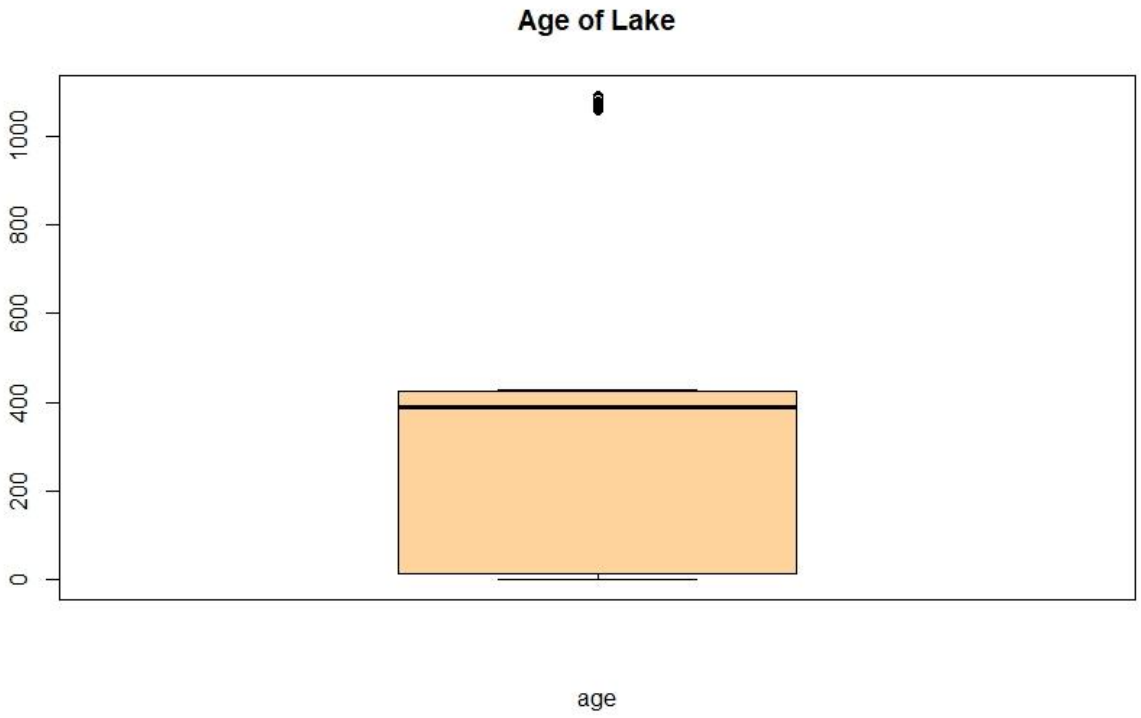
Due to the fact that these datasets are supplied from data.gov, there is no Nulls, but there is NA's, thus the only data manipulation will be replacing those NA's as well as renaming column variables.

Business Question

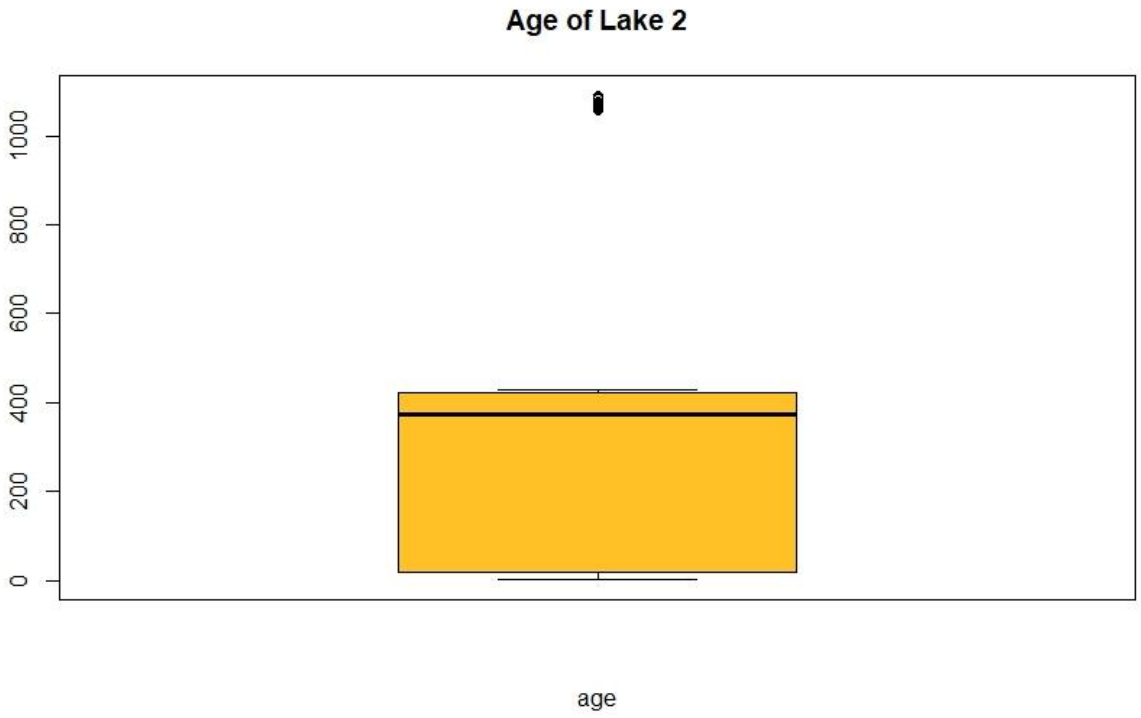
1) What is the common column variable that can be made from these 3 different file formats (txt, JSON, & xls) to merge into a database?



A boxplot was made for the age variable based off the Baikal data set; there is outliers which will still be included (i.e., not removed from the original dataset).



A boxplot was made for the age variable based off the Lake data set; there is outliers which will still be included (i.e., not removed from the original dataset).



A boxplot was made for the age variable based off the Lake2 data set; there is outliers which will still be included (i.e., not removed from the original dataset).

Assumptions

I. The common column variable will be the **Age** variable due to varying differences between Lake Baikal and Lake Elgygytgyn in the Russian Federation based off the paleoclimatology in that Eurasian country.

MODELING/METHODS

To perform this discriminant analysis based off the common column variable, **Age**, rudimentary EDA will be executed followed by converging the datasets into a single dataset and implementing them into a database.

DEPLOYMENT/RESULTS

After finding a common column variable (i.e. **Age**) between the different file formatted datasets and placing the conjoined dataset into a database, a performance check should be implemented but has no relevance in this particular discriminant analysis.

SUMMARY & CONCLUSIONS

Since the three different datasets were structurally formatted in a way where there is some data preparation or wrangling that needs to be done, the more legwork will be finding the common column variable and inputting the merged data into a database based off the business question proposed to the business problem of which column variable is the link between the three datasets to form a single dataset and tool it into a database. Following the CRISP-DM process, after discovering the common column variable, a performance check should be administered but this has no significance to this discriminant analysis. For the business problem, the **Age** column variable was the common variable from each dataset. This was due to the fact of the varying in age for the biogenic silica and sediments collected at both Lake Baikal and Lake Elgygytgyn. Lastly, an understanding of what the discriminant analysis of the datasets will be executed. This discriminant analysis concluded with the three different file formatted data (txt, JSON, & xls) being merged into a single dataset by the **Age** variable (the common column variable name) and placed within a database (SQLite). The merged data had a dimension of 3856 rows & 33 columns, a little more from the original datasets, but this database data was joined together by the **Age** column variable. What was significant about the **Age** column in the database data from a simple boxplot indicates that the age of the sediments collected at Lake Elgygytgyn is older compared to the age of biogenic silica buried at Lake Baikal even though this lake is way older compared to the newer lake, Lake Elgygytgyn. In conclusion, to study the paleoclimatology of the biogenic silica in Lake Baikal and the sediments in Lake Elgygytgyn, the age of both

minerals should be considered within the two lakes, thus showing the significance of the **Age** column variable during the Pleistocene epoch/Quaternary period.

References

10 Key Types of Data Analysis Methods and Techniques. (n.d.). Retrieved from 10 Top Types of Data Analysis Methods and Techniques (intellspot.com)

Lake Baikal. (n.d.). Retrieved from Lake Baikal - Wikipedia

Lake Baikal: The Pearl of Siberia. (n.d.). Retrieved from Lake Baikal in Siberia, Russia | Baikal Lake Facts & Map (lakepedia.com)

Lake Elgygytgyn. (n.d.). Retrieved from Lake Elgygytgyn - Wikipedia

NOAA/WDS Paleoclimatology – Lake Baikal Composite BDP-96 Pleistocene Biogenic Silica Data. (n.d.). Retrieved from NOAA/WDS Paleoclimatology - Lake Baikal Composite BDP-96 Pleistocene Biogenic Silica Data - CKAN

NOAA/WDS Paleoclimatology – Lake El'gygytgyn, NE Russia Quaternary Multiproxy Lake Sediment Data. (n.d.). Retrieved from NOAA/WDS Paleoclimatology - Lake El'gygytgyn, NE Russia Quaternary Multiproxy Lake Sediment Data - CKAN

Pleistocene Epoch. (n.d.). Retrieved from Pleistocene Epoch | Characteristics, Plants, Animals, Climate, & Facts | Britannica

Strauss, B. (2020, January 21). *Prehistoric Life During the Pleistocene Epoch*. Prehistoric Life During the Pleistocene Epoch (thoughtco.com)

The Thrill to Drill in the Chill – 3.6 Million Years of Arctic Climate Change from Lake El'gygytgyn, NE Russia. (n.d.). Retrieved from | Lake El'gygytgyn (dosecc.com)

Vermillion, S. (2021, April 16). *Siberia's Lake Baikal Is the World's Oldest and Weirdest*. Siberia's Lake Baikal Is the World's Oldest and Weirdest | HowStuffWorks

What is Paleoclimatology?. (n.d.). Retrieved from What is Paleoclimatology? | National Centers for Environmental Information (NCEI) formerly known as National Climatic Data Center (NCDC) (noaa.gov)

Appendices

Appendix A

What is Paleoclimatology?

The article defines what paleoclimatology is as well as the benefits of the scientific study.

Appendix B

Pleistocene Epoch

This article defines when the Pleistocene epoch occurred, the stratigraphy, and lastly, the chronology and correlation of the epoch.