Data Mining Project

Anthony Dubreuil

April 30, 2024

1 Introduction

With the expansion of globalization and the rise of the Internet in the past decades, learning new languages has been made both more important and more accessible, with the help of specialized applications, such as Duolingo.

With this project, my goal was to find a correlation between a country's various attributes and the languages that are learnt the most by the inhabitants of this country.

Are humans living in similar but different countries prone to learn the same language? What are the most important attributes of a country that could explain why its people want to learn a specific language? My work aimed to find an answer to these questions.

The first step of this work was to find a good dataset to study. I decided to combine 2 datasets I found on `kaggle.com`:

The first dataset [3] contains information about every country in the world, including but not limited to geographical (e.g. land area, latitude, longitude, ...), economical (e.g. minimum wage, CPI, ...) or even linguistic (official language) data. It contains 195 observations of 35 variables.

The second dataset [2] contains information about the 2 most learned languages in every country on the Duolingo application, for the years 2020 to 2023. It contains 193 observations of 10 variables. In my study, I focused on the year 2023, so only 2 of the variables were used.

The entirety of my code [1] and the link to both datasets can be found in the References section.

I created the used dataset myself, by adding the ISO 3166-1 alpha-3 code of each country to both datasets by hand. By doing that, I was able to link both datasets using this code to get the rows corresponding to each country to be merged in my new dataset. I retrieved all columns of the world dataset, and the 2 columns related to 2023 from the Duolingo dataset. Two rows were dropped because the corresponding countries were not present in the Duolingo dataset.

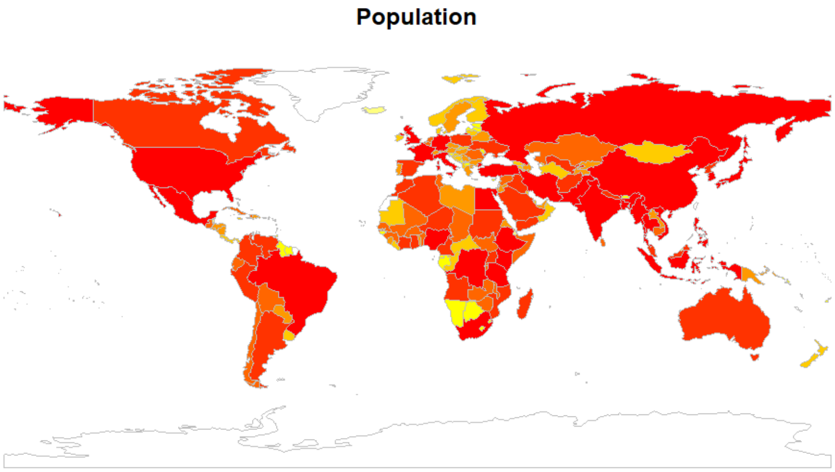


Figure 1: Map plotted using the Population column of the first dataset

2 Data Visualization

After creating the dataset, I proceeded to visualize the data.

As an example, a representation on a world map of one of the columns of the dataset, Population, is shown in Figure 1. This column of the dataset gives information about the population in each country. On this plot, the redder the country is represented, the more people live in the country.

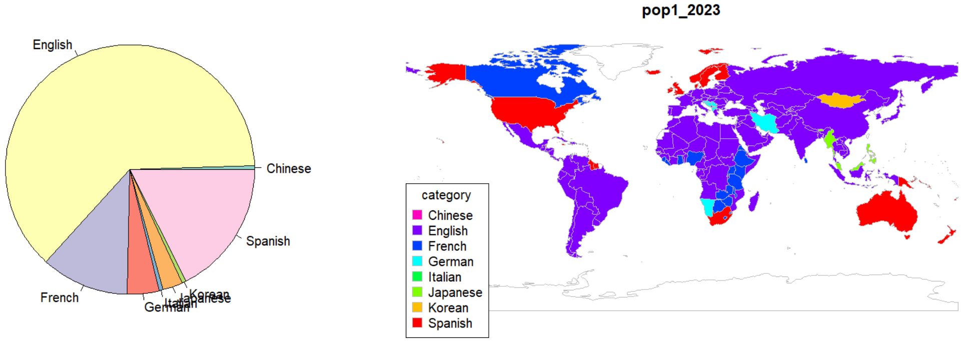
Additionally, I decided to look at the different languages present among the most learned languages in a country in the Duolingo dataset for 2023. A pie chart showcasing the data is presented in Figure 2, along with a plotted map showing which countries learned which languages the most.

It is clear that English is by far the most represented language in the list. As it is learnt all around the world for various reasons and no matter the country's situation, I decided to work on data not taking into account the English language, as explained in the Preprocessing section.

3 Preprocessing the data

After visualizing the data, it is important to modify it so that it can be used properly.

As mentioned in the Visualization section, the first decision that I made was to create a new column for the next parts of this study, that would not consider the English language among the learnt languages on Duolingo. The method was



(a)

(b)

Figure 2: Pie chart and map plot showing what language was the most learnt on Duolingo for each country

fairly simple: if the most learnt language in a country was the English language in 2023, I would take into consideration the second most learnt language from this year in the country instead. The new representations of the data can be seen in Figure 3.

In addition to this modification, I decided to drop several columns that were not relevant to this study. Namely, the non numerical columns, except the column indicating the official language of each country. I made this decision because the names of cities or currencies are difficult to compare between countries, and do not seem to be relevant enough to follow through these difficulties.

As such, the dataset used in my study contained 193 observations of 33 variables.

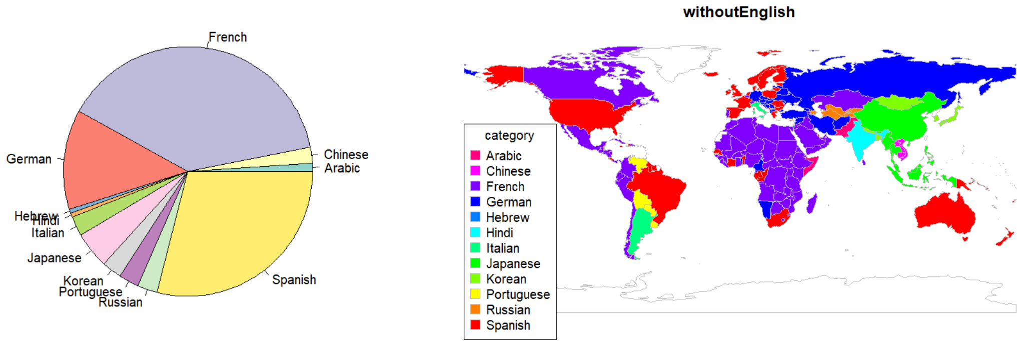
After deciding on the relevant columns, I needed to transform them into numbers, because most of them were classified as characters, because of commas in the middle of the numbers, percent signs and dollar signs.

Additionally, for any missing value in a column, it was replaced by the column's mean, to ensure the mean does not change, which is important for the Transaction section.

After all of this was done, I started working on the data.

I plotted a correlation matrix on all variables, but it is hardly readable and most of the results are not directly linked to languages. Proper correlation will be elaborated upon in the Correlation section.

Additionally, I realised a Principal Components Analysis (PCA) on my dataset, but I decided to only briefly mention it in this report, as the results are difficult to read and I did not do further research using them. The code and



(a) (b)

Figure 3: Pie chart and map plot showing what language was the most learnt on Duolingo for each country, barring the English language

plot of this PCA can be found in the provided code.

4 Transaction rules - Apriori

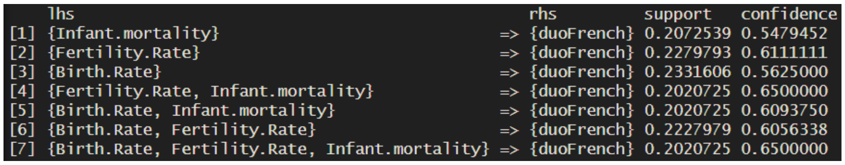
The main focus of my work has been to try running an Apriori algorithm on my data. However, to do that, I needed transactions.

To create these transactions, I first decided to transform all my remaining character columns, all linked to languages, by one-hot encoding them. This led to the creation of a lot more columns, this is why I decided, for the official language column, to only keep the information if the language was part of the list of languages seen in the most learnt languages (and also English). This means that if a country has German as its official language, I kept this information, but I did not if it was Armenian, for example.

For the most learnt language column, it was a simple one-hot encoding keeping all the information. These 2 columns were replaced by 12 columns each (+1 for the official language to add English), each corresponding to a language. The value is 1 if this is the corresponding language, 0 else.

For the numerical columns, I decided to put 1 in rows with a value greater than the mean of the column, and 0 in others. With this method, I then had a full dataset composed only of 1 and 0, which I transformed into a list of vectors: each row of the dataset generated a corresponding vector which contained the columns for which the row had a value of 1.

This allowed me to run the Apriori algorithm on this list, with a minsup of 0.2 and a minconf of 0.5, and to check the association rules that had the learning of a language on Duolingo as the consequence. However, no matter what values I tried, the only language that produced significant results is the French language,



```
    lhs                                           rhs         support   confidence
[1] {Infant.mortality}                         => {duoFrench} 0.2072539 0.5479452
[2] {Fertility.Rate}                           => {duoFrench} 0.2279793 0.6111111
[3] {Birth.Rate}                               => {duoFrench} 0.2331606 0.5625000
[4] {Fertility.Rate, Infant.mortality}         => {duoFrench} 0.2020725 0.6500000
[5] {Birth.Rate, Infant.mortality}             => {duoFrench} 0.2020725 0.6093750
[6] {Birth.Rate, Fertility.Rate}               => {duoFrench} 0.2227979 0.6056338
[7] {Birth.Rate, Fertility.Rate, Infant.mortality} => {duoFrench} 0.2020725 0.6500000
```

Figure 4: Relevant association rules found by the Apriori algorithm

meaning that the apriori method was not sufficient to find relationships as I had hoped. The association rules I found for the French language can be seen on Figure 4.

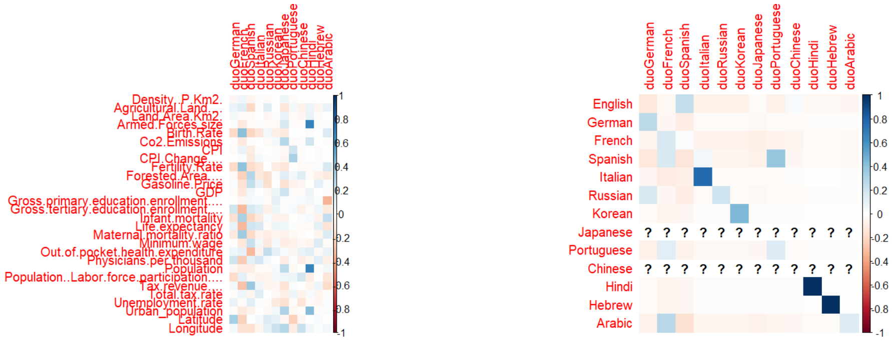
5 Correlation

To better understand the relationships between my data and the learnt languages, I decided to plot the correlation matrix again, with the one-hot encoding done. I divided it into 2 plots: one showing the correlation between the numerical values and the learnt languages, and the other showing the correlation between the official languages and the learnt languages.

Both of these plots can be found on Figure 5.

These plots show us that little to no correlation exists between our data and our target languages, except on some rare instances, like the ones found by the Apriori algorithm for the French language. Interestingly, the algorithm did not return any association rule for the Hindi language, even if it seems to be the most correlated language with other data on our results. This is probably due to its rarity among the learnt languages, however.

Aside from the numerical values, the only notable correlations are found in countries that mostly learn their own official language, namely Italian, Korean, Hindi and Hebrew, and hispanophonic countries learning Portuguese. As a side note, we can see that no country seems to have Japanese or Chinese as their official language in our dataset.



(a) Numerical Values

(b) Official Languages

Figure 5: Coefficient mattrices between our data and the learnt languages on Duolingo in 2023

6 Conclusion

As this project showed us, there does not seem to be any obvious data which would be heavily correlated to the languages learnt in a country. As such, it can be interpreted that similar countries (according to our dataset) do not necessarily learn the same languages, and that way more context (e.g. historical or religious context) would be needed to understand these tendencies. However, while the results I found were not numerous, their existence proves that these tendencies exist and can be studied, to better understand why people want to learn a language, and maybe adapt the way a language is taught as a result.

References

[1] Github repository containing the code for the project. https://github.com/Nythan1409/DM_Project.

[2] Duolingo and Cindy Blanco. Global language learning popularity [2020-2023]. https://www.kaggle.com/datasets/surakarthikeya/global-language-learning-popularity-2020-2023, 2023.

[3] NIDULA ELGIRIYEWITHANA. Global country information dataset 2023. https://www.kaggle.com/datasets/nelgiriyewithana/countries-of-the-world-2023, 2023.