Europe's Language Families

Justin Peter
Data Science
Bowling Green State University
Ohio, USA
jgpeter@bgsu.edu

Donghyun Jeon
Data Science
Bowling Green State University
Ohio, USA
djeon@bgsu.edu

Daniel Felbah
Data Science
Bowling Green State University
Ohio, USA
dfelbah@bgsu.edu

***Abstract*—The purpose of this project is to showcase the different cluster categories we made in Europe and then make another model to classify which cluster stands out. This could be based on the number of cases, deaths, tests, etc. The idea behind this is because the media tends to talk a lot about the spike in covid cases and deaths during the first wave especially on the Romance countries. Romance countries are naturally known for physical contact which usually causes the spike in covid cases. We will analyze more on Germanic and Slavic countries and see whether our analysis says anything different about our data.**

***Index Terms*—time series, clustering, classification, python, allnighters**

I. Introduction

Back in the hackathon, we presented our analysis based on continents. However, due to the immense amount of data that we have, we decided that it would be best to work on all countries in one continent. From our hackathon, we thought that South America would have been a better continent to analyze because of the amount of poverty that exists. However, this is incorrect because for example, Africa has huge amounts of poverty but we don't see huge spikes in covid cases over there. So poverty and hunger has nothing to do healthwise to covid cases. In addition, unless the country in South America is Brazil, most of the other countries are not really talked about. The media tends to talk more about the covid cases that would occur in Europe. It's interesting because the USA is the focal point of all covid-19 discussion, but unfortunately our data doesn't provide information on the states.

II. Background

So our dataset is very huge and it updates based on the date. So for every date in the data, we will see the total number of cases, deaths, tests, and many more columns for each country. Apart from that, we have more than 40 columns in the data and we ended up selecting certain columns for feature selection that would assist us in the clustering aspects. Other columns were easily used and manipulated for visualization purposes. For more information, see references.

This dataset belongs to Our World in Data. Most of the dataset has been worked on and updated by many researchers around the world. Their motto is: Research and data to make progress against the world's largest problems. They have a Github page which is linked in the references that displays various information about the data and the columns that will update each day. Right now, you will see 60,862 rows and 50 columns which will eventually be filtered.

III. Methodology

A. Data Processing

In almost every dataset, you are going to see missing data. It would be so easy to replace data with a random number, but with covid cases, not so much. So we decided to define a function that would take the data and implement the simple imputer to replace the numbers.

B. Data Manipulation

Luckily, we did not have data tidying since all the date values belong in one row. When manipulating the data, we adjusted one column that displayed the exact month, day and year. For some of the visualizations especially with those from the Hackathon, we adjusted our visualizations based on the trends of the various line plots that moved based on month. Also, we used a correlation heatmap that was able to display the relationship among the accumulated increase in tests, cases, and deaths.

C. Algorithms

For algorithms, we used KMeans clustering at first, but unfortunately it falied. Please see part A in the results section. We eventually formed our own algorithm by forming our own clusters. We did this in three different manners. The first way was by language families, so for example, Poland, Russia, and Ukraine speak slavic languages that are located in the North. Similarly, we did the same for germanic and romance languages. The second way was dividing the location of each country by region so quite literally, the countries are clustered. The third way was through a travel competiveness index that you can find below.

- Romance : Spain, France, Italy, Portugal
- Slavic : Russia, Ukraine, Poland, Croatia
- Germanic : Ireland, Norway, Sweden, Germany
- Other : Hungary, Albania
- West : Spain, France, Italy, Portugal
- East : Russia, Ukraine, Poland, Slovakia
- North : Finland, Norway, Sweden, Germany
- South : Hungary, Albania, Greece, Macedonia

The first item displays the language families and the second item displays the regions. We repeated results such as that for

the different indexes, however the values that reached a certain threshold would get the label high if the value was higher than the threshold and viceversa if the value was lower.

IV. Results

A. Regression Models

We ran some regression models that would display the value behind each column and the impact it would bring. We group the country and aggregate the following columns: human development index, gdp per capita, population, life expectancy, extreme poverty, and stringency index primarily by their maximum values. We obtained a dataframe and then replaced missing values with the SimpleImputer from SkLearn by most frequent. After, we preprocessed the data with PCA and obtained the optimal number of clusters to be 7 when the number of countries resemble close to 40. We were able to scatterplot however the values for each cluster were not effective in terms of placing a sure connection.

When we made a new correlation heatmap to display the relationship among those specific columns, we knew that the columns were an excellent example of feature selection, however, it did not display the use in identifying why KMeans chose the countries' clusters the way it was. Using StatsModels and LinearRegression, we fit the model on training and testiing data by dropping the cluster. Once we fit the model, all predicted coefficients were zero, and the intercept was negative.

We then ran some OLS Regression Results. In all the results displayed that the condition number was really large. Adjusting the residuals did not really affect the R-squared value or F-statistic. In fact, the F-statistic increased as the number of residuals went up. The R-squared value maintained at least a 0.5 throughout all the results but the final model displayed the least R-squared value as the number of columns used were decreased. It seemed as if there was a connection among the gdp per capita, population, and the stringency index.

B. TSLearn

After making our own algorithmic clusters, we saved our data from Excel into a csv file. We then imported TimeSeriesKMeans and a utility tool that would transition our original dataset that failed from the KMeans more into a time series dataset. We chose four clusters and a dtw metric as opposed to a Euclidean metric from our model. When we ran a confusion matrix, we found the following:

- Germanic countries have a high quality of life
- Slavic countries are twice as likely to have a low quality of live
- Romantic or Other countries could go either way

From there, we make strip plot that places the language family to the region in respect to the quality of live. For the most part, the majority of the clusters make sense, however the slavic countries display clusters with low and high qualities of life, so we should take that into consideration.

After that, we reduced our dataframe to only the rows in the data that represented both Slavic and Germanic countries. We then displayed an excellent for loop that is displayed like this:

- Form a list that takes in cases, deaths, icu patients, hospital patients in the millions along with new tests, positive rate, and the imputed stringency index
- In respect to the location and date, we adjust the date to columns and index our countries
- Apply the TimeSeriesKMeans model with 3 clusters in the dataset and predict
- Each cluster is labeled based on language family, quality of life, and healthcare
- In addition to the line plot trend is a confusion matrix that places each label to its prediction
- Repeat for every column in the list

Here are interesting points for the Slavic countries:

- Cases will accumulate more or less each day
- Deaths will increase more like a spike
- Most likely to not have patients in the ICU
- Slavic and Germanic will have less patients in the hospital
- Tests might be consistent
- More Slavic countries will test positive

Here are interesting points for the Germanic countries:

- Cases are very low but may increase a little
- Deaths are a fluctuation so results remain hidden regarding increase or decrease
- The change in ICU patients are either going to be very little or a small spike
- Slavic and Germanic will have less patients in the hospital
- Less tests may take place
- Germanic countries will test positive remains a doubt

V. Conclusion

A. Summary

Throughout the paper, we displayed a background of where we obtained the data, why we decided to use the data, what information the data consists of, and the manipulation we made on the data before performing analysis.

Apart from our visualizations, we made KMeans clustering and time series clustering models that displayed different results but we were able to establish a connection among Slavic and Germanic countries regarding high or low quality of life.

Few points for accumulated cases:

- East Slavic countries stand out the most
- UK and Germany represent the majority of cases in the Germanic family

Few points for accumulated cases in the millions

- South Slavic countries accumulated by the millions as opposed to East Slavic
- Germany's neighbors such as Austria and Netherlands, form a different result

B. Potential

There is huge potential with this data in terms of other questions we can address such as:

- Which country will reach 5 million cases first?
- Which country will be the first to have 1 million deaths?
- Which country will reach at least 5 million tests first?
- Do language families or branches have anything to do with Covid cases?

As you can see, most of these questions are based on classification. From our data and our work, you can see that regression algorithms are not that effective when a majority of the data is referring to various countries and continents along with their statistics. We hope that we can uncover much more information and potential with time series clustering when it comes to the questions above.

REFERENCES

[1] Max Roser and Esteban Ortiz-Ospina (2019) – "Global Rise of Education". Published online at OurWorldInData.org. Retrieved from: 'https://ourworldindata.org/global-rise-of-education' [Online Resource]

[2] Hasell, J., Mathieu, E., Beltekian, D. et al. A cross-country database of COVID-19 testing. Sci Data 7, 345 (2020). 'https://doi.org/10.1038/s41597-020-00688-8' [Online Resource]

[3] Edouard Matheiu - owid-covid-codebook.csv(2020). 'https://github.com/owid/covid-19-data/blob/master/public/data/owid-covid-codebook.csv' [Columns].

[4] World Economic Forum- Travel Competiveness Index by Year 'https://www.weforum.org/reports' [Countries and Values].