COMP 551 - Applied Machine Learning

Fall 2020

Mini-Project 1

Analyzing COVID-19 Search Trends and Hospitalization

Group 19

Group Members	Student IDs
Alain Daccache	260714615
Shayan Sheikh	260738247
Bera Sogut	260788386

Project Submitted To:

Professor Mohsen Ravanbask

McGill University

20th October, 2020

Table of Contents

1. Abstract **3**

2. Introduction **3**

3. Datasets **4**

3.1. Search Trends Dataset 4

3.2. Hospitalizations Dataset 6

4. Results **7**

4.1. Search-Trends Visualization 7

4.2. K-Means Clustering Visualization 8

4.3. Regression Comparison 10

4.3.1. KNN with 5-fold cross validation 10

4.3.2. Decision Trees 11

4.3.3. Comparison 13

4.4. A Third Approach: Time Series Analysis 13

5. Discussion and Conclusion **14**

6. Statement of Contributions **14**

7. Appendix **15**

1. Abstract

The purpose of this project was to analyze, investigate, and compare the performance of k-nearest neighbours and decision trees, two of the most important regression models in machine learning, by using them to predict COVID-19 hospitalization cases from related symptoms search. Such analysis is vital, as properly predicting a surge of COVID-19 cases in a particular region at a particular time would help authorities to better manage their resources (i.e. testing material, hospital beds, etc.) ahead of time, to avoid overwhelming the healthcare system. K-nearest neighbour regression outperformed decision trees regression in terms of accuracy as it reported a MSE value of **~5800**, in comparison to **~6800** MSE value reported by decision trees.

2. Introduction

The task of this project was to gain a better idea of the pipeline behind approaching typical Machine Learning problems, within the real-world context of COVID-19. Specifically, this task was divided as follows:

- **Data Collection & Preprocessing:** Downloading two datasets: one relating to search trends of COVID-19 symptoms, and the second relating hospitalization COVID-19 cases, deaths, tests, hospitalizations etc. They are loaded into useful Python data structures, and cleaned based on fill-factor, normalization, and correlation.
- **Data Visualization & Clustering:** Visualizing this data across dates and regions, then further reducing its dimensionality to better observe trends and clusters.
- **Machine Learning Frameworks:** Evaluating two supervised learning algorithms, namely KNN and Decision Trees, to predict the hospitalization cases given the search trends data.

Other academic papers examined those algorithms among others (such as Support Vector Machine, Artificial Neural Networks, Random Forest), by using more exogenous variables such as daily temperature [1], predicting other variables such as mortality rate in patients with COVID-19 [2], and evaluating their accuracy with a similar approach (i.e. confusion matrix) [2]

3. Datasets

3.1. Search Trends Dataset

The first dataset shows trends in search patterns for health symptoms (along with signs and other conditions) over time (daily) and across geographic regions, such as fever, difficulty breathing, stress etc., based on the volume of corresponding Google searches. To process the data, we:

1. Made it compatible with the Python environment by parsing dates into DateTime objects and setting MultiIndexes for date and region.
2. Removed features for which there was more than a certain threshold of missing entries (i.e. 30%)
3. Used max-min normalization per region to undo the unknown region-specific scaling factor used by the dataset providers to divide the raw values of search trends. To do so, we normalize our data across regions to a different metric in order to compare data across different regions. If we normalize each symptom separately (instead of symptom per region), we wrongly assume the mean of all symptoms are the same, since some symptoms are probably more popular than others and therefore have a higher mean of search trends data. This resulted in each symptom being scaled between the value of 0 and 1 for each region.
4. Removed correlated features above a certain threshold (80%), as they provide little to no extra information (keep one per pair).
5. Removed regions where we observe negative values for the number of new hospitalized Covid cases as we figured that they could provide wrong information and reduce our accuracy.
6. Removed the data before March 1st. The dataset had a lot of 0 hospitalizations before so we figured that the states may have not started registering the number of hospitalizations before that date.

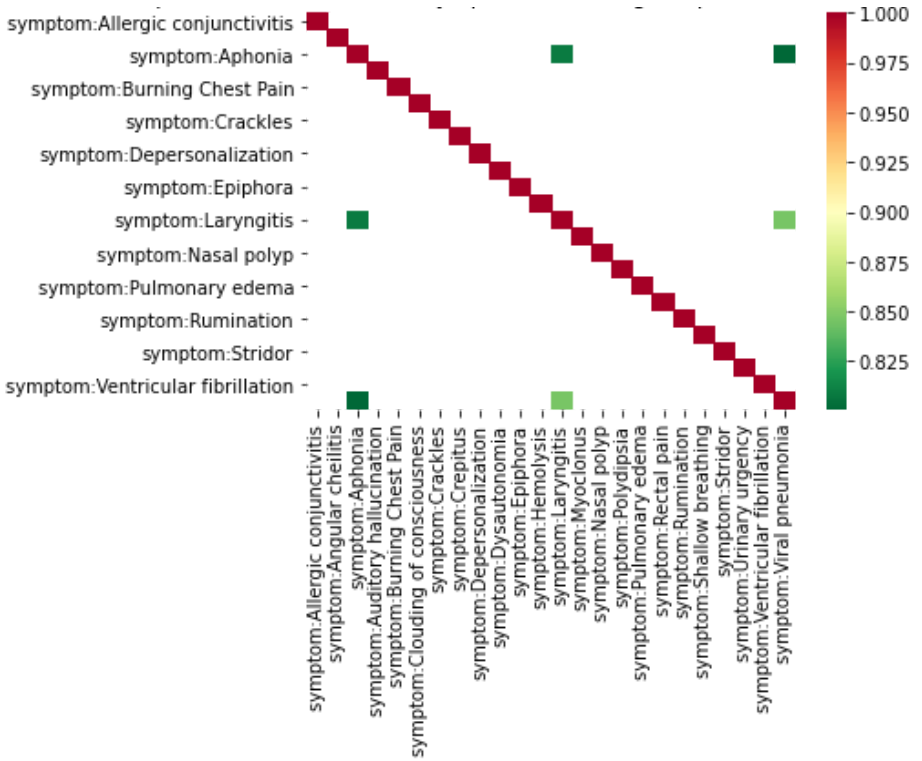


Figure 1: Correlation heatmap between the different symptoms (masking the pairs that are less than 80% correlated)

To observe how the search popularity of those symptoms evolved over time, we produced a barchart for each region. For the purpose of brevity, we are only showing the one for New Hampshire below, and the others can be found in the appendix.

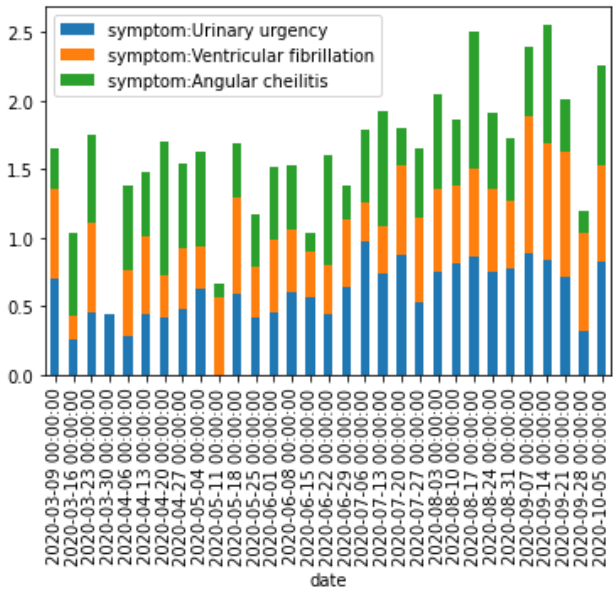


Figure 2: Evolution of the popularity of the Top 3 Symptoms relating to COVID-19 in New Hampshire

3.2. Hospitalizations Dataset

The second dataset aggregates public COVID-19 data sources into a single dataset, and covers features such as COVID-19 cases, deaths, tests, hospitalizations etc. To process the data, we only kept the relevant feature (new hospitalization) that we'll use as a target for our prediction. We merged it with our previous dataset by aggregating the data to weekly (and matching the days by shifting by one), with the key being the `open_covid_region_code`.

To visualize the data from this dataset, we produce a time series describing the changes in hospitalization cases over time. This data will be useful to inspect statistical features such as stationarity, and use for our time series analysis.

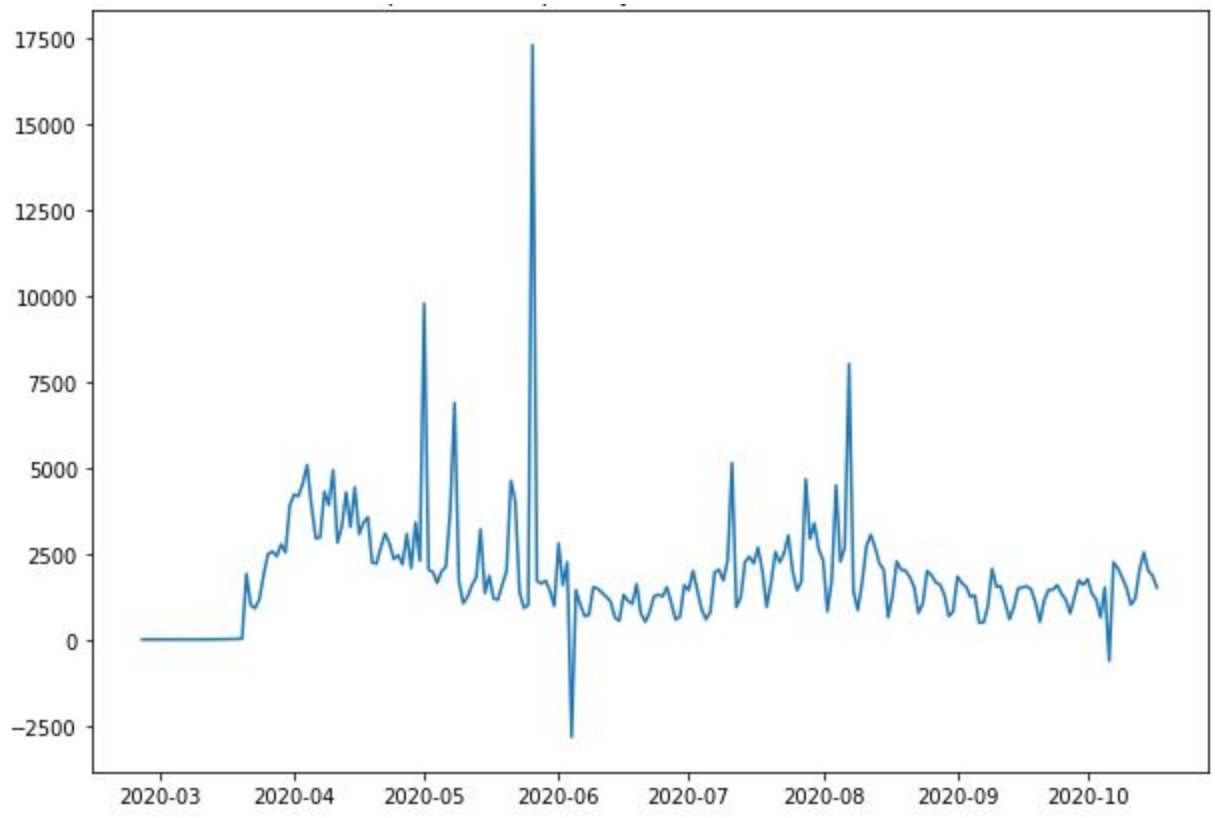


Figure 3: Time series representing new hospitalizations per day related to COVID-19 in the USA

4. Results

In this section, we now describe the results of all the experiments mentioned in Task 2 and 3.

4.1. Search-Trends Visualization

To visualize search trends in a lower dimensional space, we plotted a 2-component (2 features with highest variance) PCA and colored the different states. The plot is shown below:

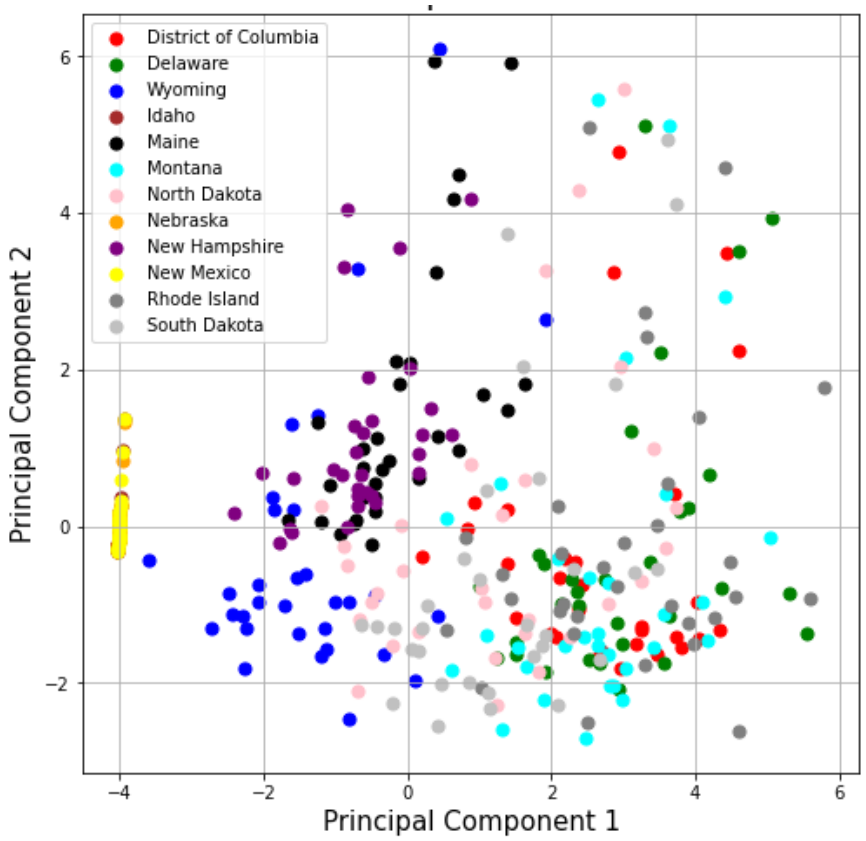


Figure 4: 2 Component PCA

The plot above shows that the generally search trends differ by each state, as can be seen by the colors being spread out in different sections of the plot. These search trends are influenced due to many reasons and hence they differ state by state. However, there is some overlap in some regions, for instance Maine and New Hampshire have their plots very close together. This could be explained by the proximity of the two states as they border each other and have similar conditions/environment. In comparison, New Mexico has its plot very different compared to New Hampshire, and one of the reasons influencing that could be the distance between the two states leading to more difference in the environment and conditions.

4.2. K-Means Clustering Visualization

Using the same 2-PCs with most variance, we plotted k-means clusters to evaluate possible groups in the search trends dataset. To pick the number of clusters, we used the elbow method using distortion. Based on the results of the elbow method, we picked 5 as the number of clusters for our k-means plot for low and high dimensions. The results of elbow method using distortion are given below:

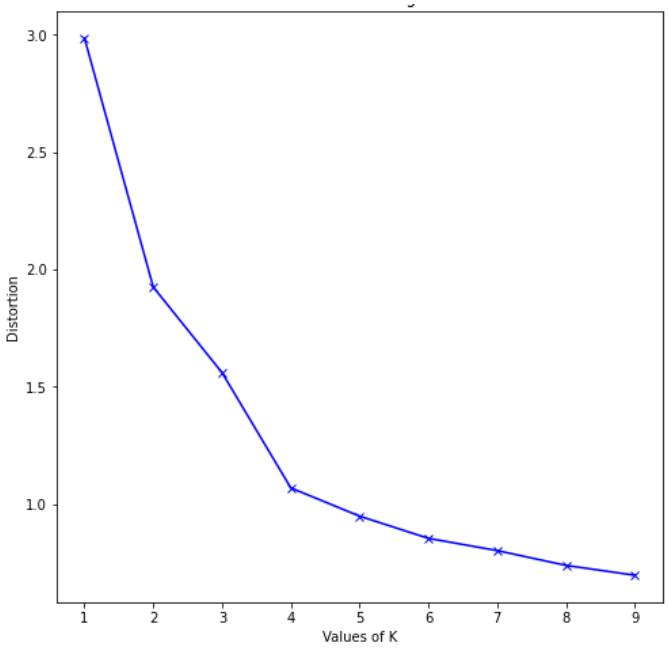


Figure 5: The Elbow Method using Distortion

The plots of cluster labels for low-dimensional and high-dimensional KMeans are given below:

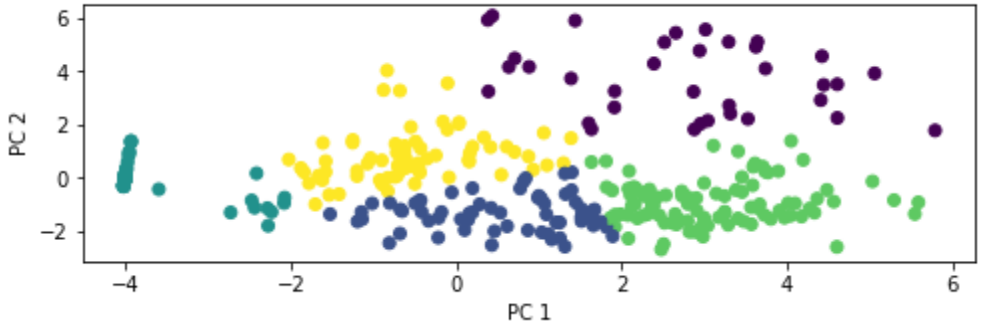


Figure 6: Cluster labels for low-dimension K-Means

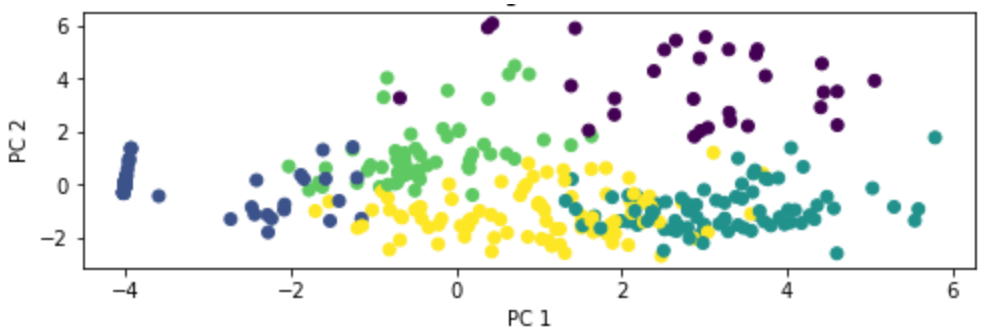


Figure 7: Cluster labels for high-dimensional K-Means

Our clusters remain consistent for the high-dimensional and low-dimensional KMeans as can be seen by the plots above. This shows that the PCA-reduced data maintained the correct distance between the points.

4.3. Regression Comparison

4.3.1. KNN with 5-fold cross validation

a) Date-Based Split

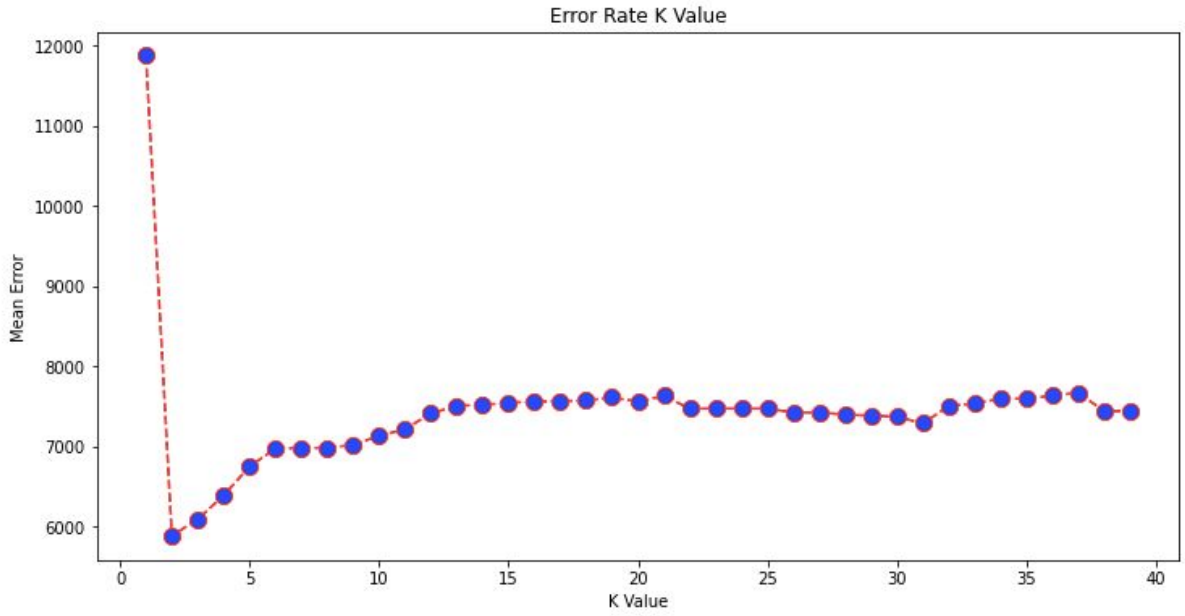


Figure 8: Graph representing mean error of KNN as a function of K Value with a date-based split

The data above shows the average mean squared error (MSE) we get using cross validation for each k = 1 to 39 where k is the number of neighbours we set for the KNN classifier. We first sort the data based on date and split it into 5 equal sized continuous chunks

for cross validation. We see that k = 2 gives the least MSE but this makes our model more prone to noise since 2 is really small. Thus, we can choose k = 8 instead.

b) Region-Based Split

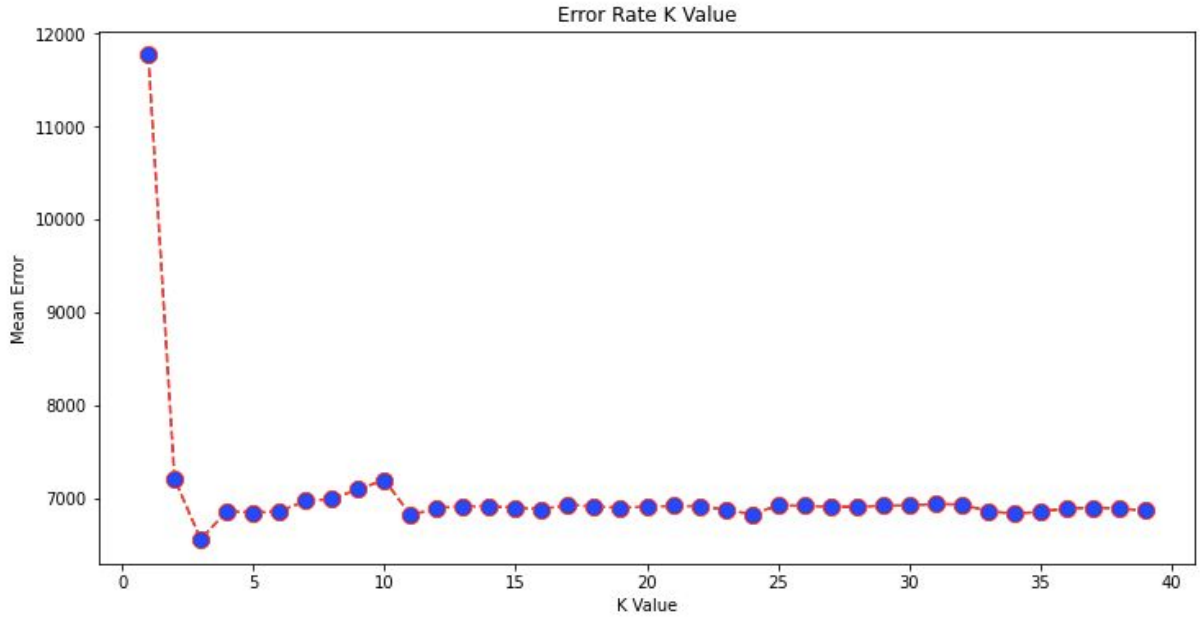


Figure 9: Graph representing mean error of KNN as a function of K Value with a region-based split

The data above is really similar to the previous data except that now we sort the data based on its region. We see that k = 3 gives the least MSE but this makes our model more prone to noise since 3 is really small. Thus, we can choose k = 11 instead which also has a similar MSE but will be less prone to noise.

We see that splitting based on region and based on time both create really similar results so it isn't really important how we split our data for KNN.

4.3.2. Decision Trees

a) Date Based Split:

The data below shows the average mean squared error (MSE) we get using cross validation for each k = 1 to 39 where k is the max depth we set for the Decision Tree classifier. We first sort the data based on date and split it into 5 equal sized continuous chunks for cross validation. We see that k = 13 gives the least MSE. When we don't provide a max depth to the classifier, it optimizes itself (eg. without overfitting) so we also tried that and the depth was 28.

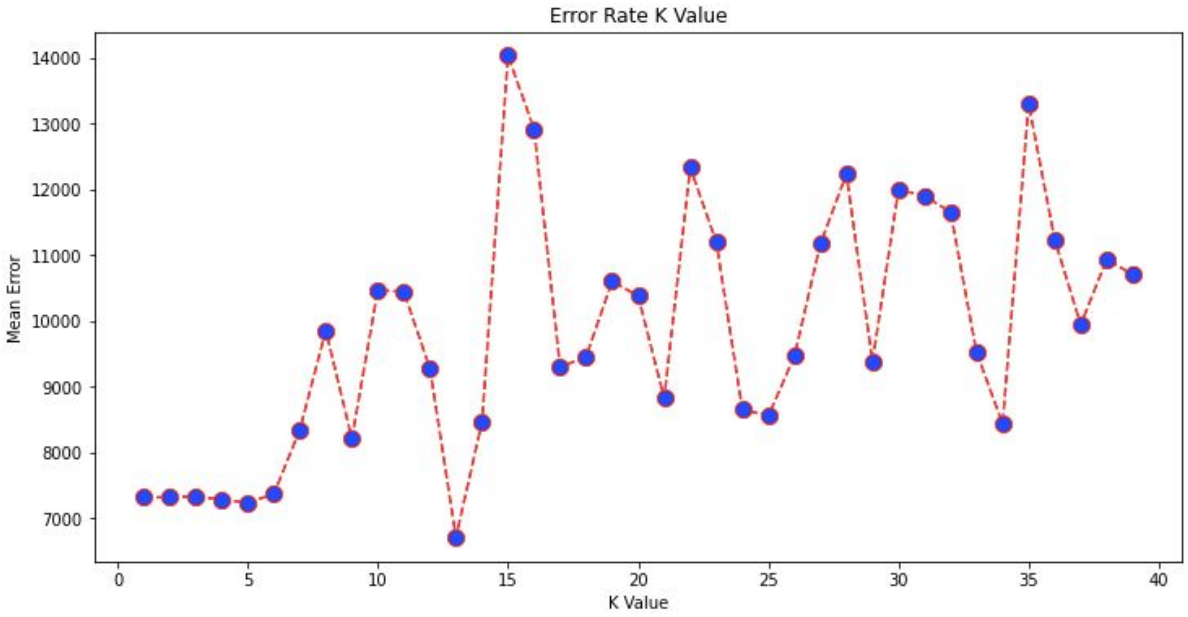


Figure 10: Graph representing mean error of Decision Trees as a function of K Value with a date-based split

b) Region Based Split:

The data below is really similar to the previous data except that now we sort the data based on its region. We see that k = 13 gives the least MSE. We also tried not providing a max depth to the classifier and the optimized depth was 27.

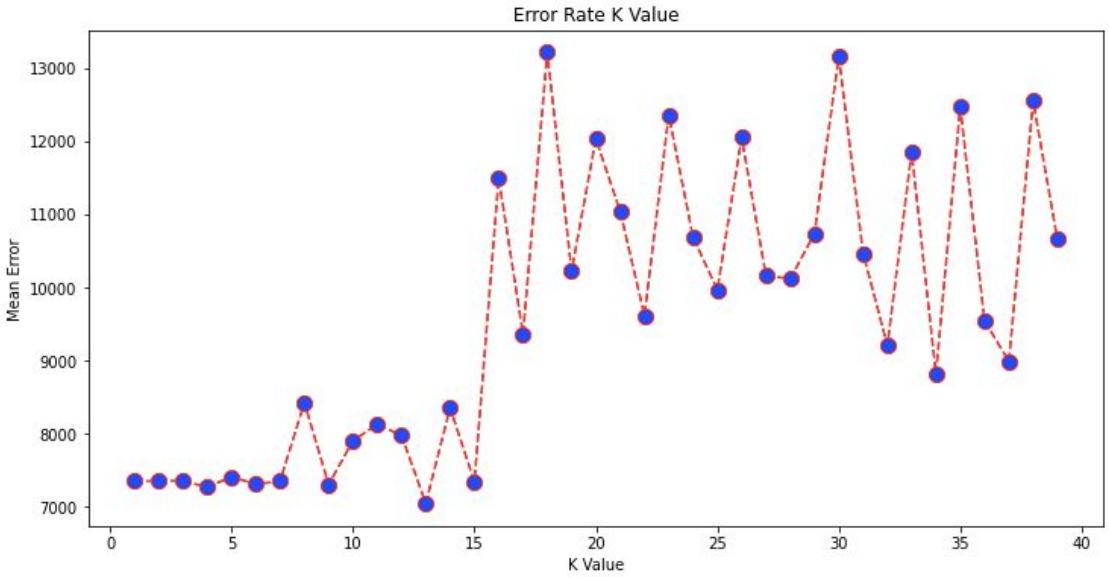


Figure 11: Graph representing mean error of Decision Trees as a function of K Value with a region-based split

We see that splitting based on region and based on time both create really similar results with decision trees as well so it isn't really important how we split our data for decision trees as well.

4.3.3. Comparison

We observed that KNN had a slightly better performance than the Decision Tree approach based on their average mean squared errors using cross validation (500-1000 difference in MSE). We tried to remove the noise from the data by cleaning and normalizing but we still believe that there is a lot of noise present in the current data. KNN performs instance-based learning, thus a well chosen k (#neighbours) can model complex decision spaces having arbitrarily complicated decision boundaries, which is not easily modeled by Decision Trees which explains the performance difference.

4.4. A Third Approach: Time Series Analysis

Intuition: The change in hospitalization cases is time dependent, and therefore potentially displays aspects of trend, seasonality, and cyclicity. Models such as KNN and Decision Trees discard this time dependency and the potential features we can extract from it, thus we can use Time Series Analysis to exploit them and build a more robust forecasting model.

Validation: This method of forecasting COVID-19 cases is already investigated in several academic papers [3], [4], [5].

Results: We split our data into train and test sets (80-20 split). Then, we choose a model that fits the characteristics of our time series' training set (in terms of autocorrelation, stationarity etc.), i.e. ARIMA, and optimized its parameters p,q, and r according to the Akaike Information Criterion.

Then, by fitting the model into our data and forecasting for all points from the start of the test set, we observe a MAPE of ~ 0.4619, meaning we reached an accuracy of about 53.81%. We have attained an even lower MAPE of ~ 0.4167 when forecasting 1-day ahead (or 58.33% accuracy), as shown in the figure below.

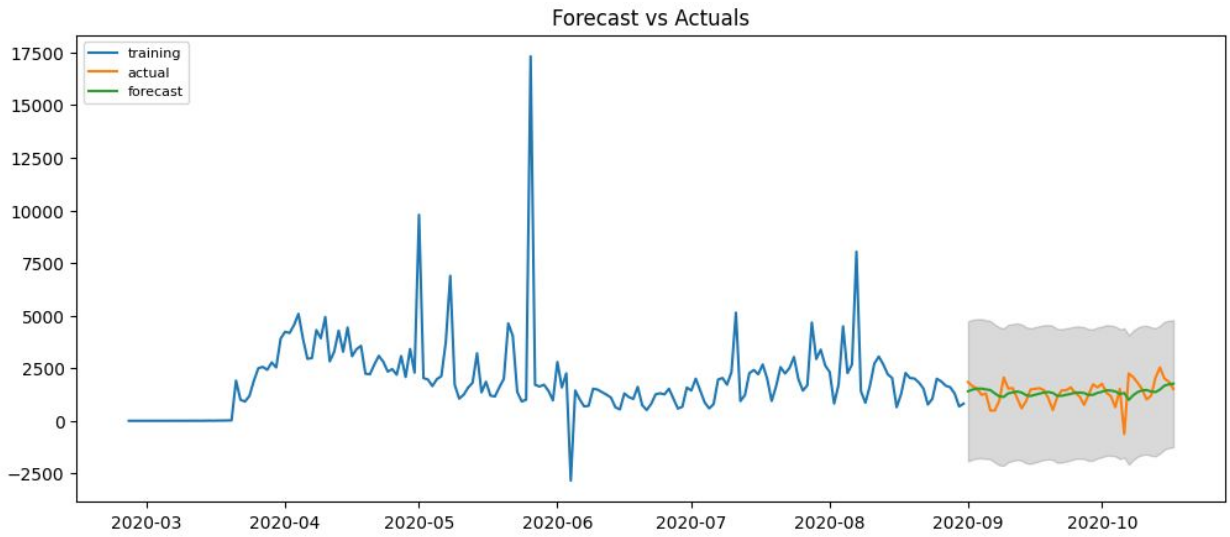


Figure 12: Time series representing the actual vs. the forecasted change in hospitalization cases due to COVID-19

5. Discussion and Conclusion

Based on our investigation and analysis, it can be concluded that k-nearest neighbour regression approach achieved better accuracy than decision trees approach. Even though KNN outperformed Decision Trees in terms of accuracy, both models yielded a significant MSE. This was predicated by the Principal Component Analysis, which showed the search trends across regions in the US are perhaps only indicative, but not predictive of new hospitalizations relating to COVID-19. The multiple cases of invalid and corrupted data in the dataset also negatively affected the performance.

To improve the forecasts regarding new hospitalization, we can explore collecting more varied data, such as weather conditions across states over time, news-based sentiment analysis

etc. Moreover, getting the raw data (that does not have the scaling factor) to make the preprocessing more flexible in removing the need to reverse-engineer, for instance a good normalization formula for our data.

6. Statement of Contributions

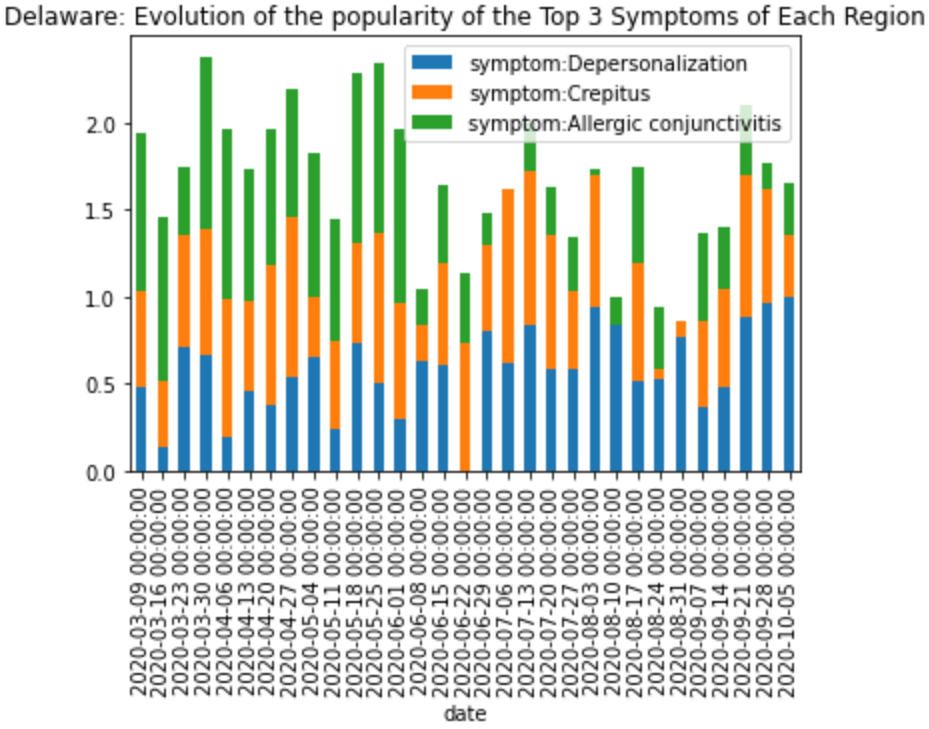
The group members and the tasks they worked on are given below:

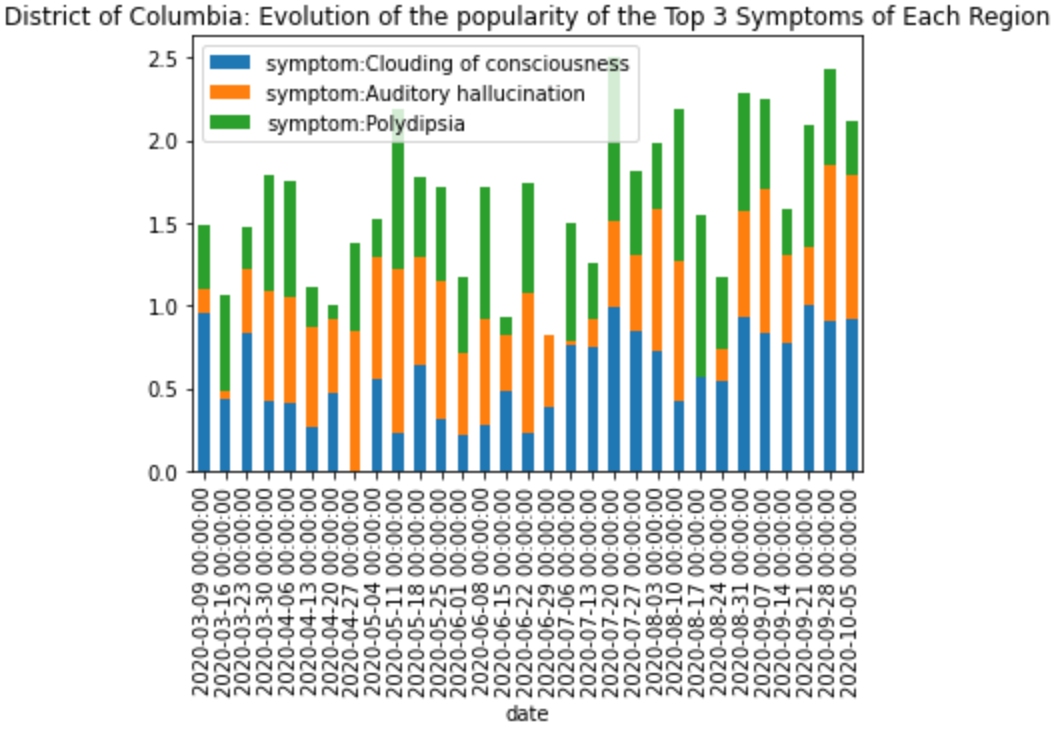
Alain Daccache: acquiring, preprocessing, and analyzing data for the two datasets; visualizing the evolution of popularity of various symptoms; supervised learning for KNNs; time series analysis; and report writing.

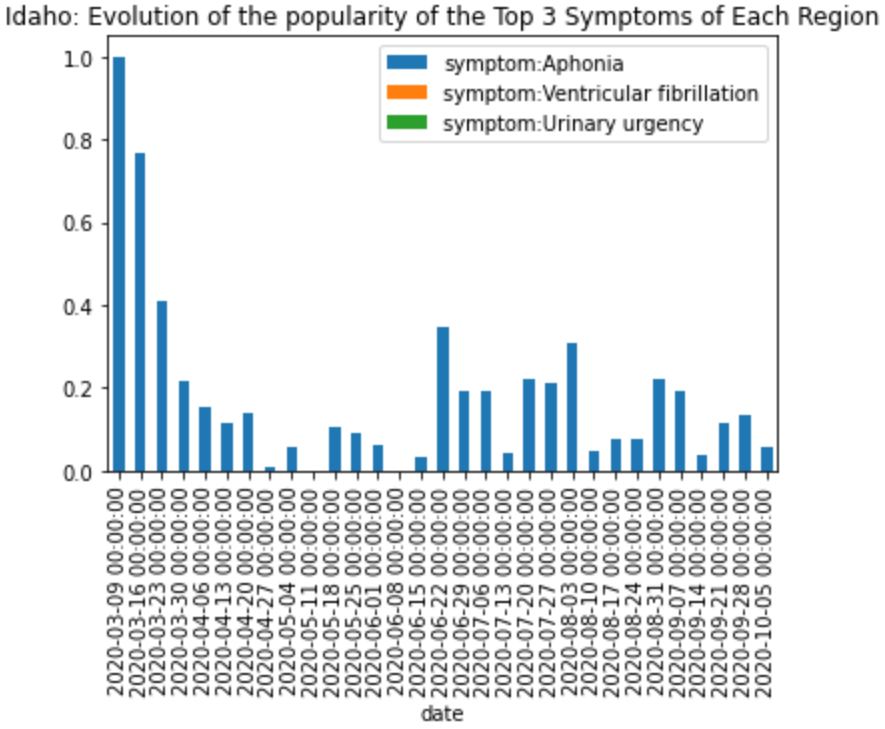
Bera Sogut: Preprocessing and analyzing data for the two datasets; supervised learning for KNNs and decision trees; and report writing.

Shayan Sheikh: Preprocessing and analyzing data for the two datasets; visualizing search trends in lower dimensional space; k-means clustering; and report writing.

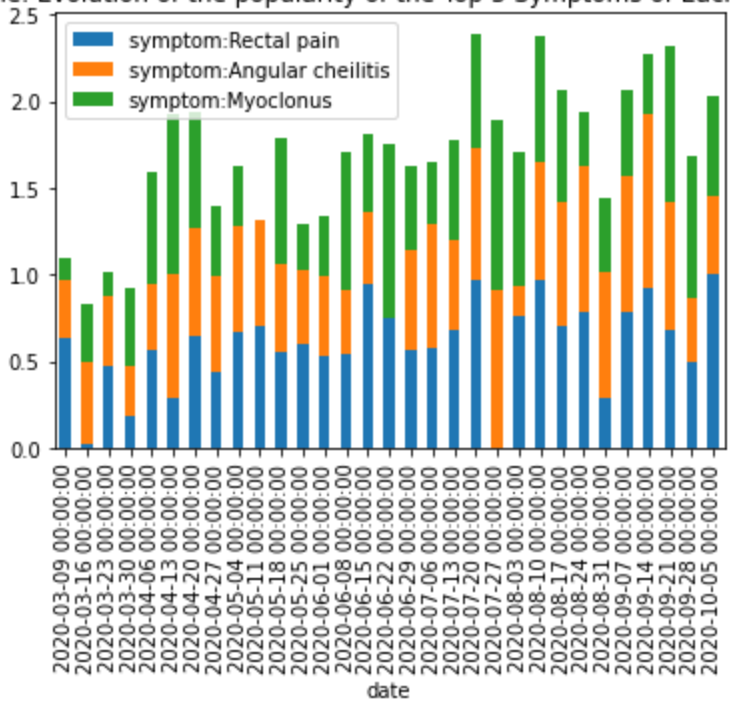
7. Appendix


Delaware: Evolution of the popularity of the Top 3 Symptoms of Each Region
symptom:Depersonalization
symptom:Crepitus
symptom:Allergic conjunctivitis
0.0
0.5
1.0
1.5
2.0
2020-03-09 00:00:00
2020-03-16 00:00:00
2020-03-23 00:00:00
2020-03-30 00:00:00
2020-04-06 00:00:00
2020-04-13 00:00:00
2020-04-20 00:00:00
2020-04-27 00:00:00
2020-05-04 00:00:00
2020-05-11 00:00:00
2020-05-18 00:00:00
2020-05-25 00:00:00
2020-06-01 00:00:00
2020-06-08 00:00:00
2020-06-15 00:00:00
2020-06-22 00:00:00
2020-06-29 00:00:00
2020-07-06 00:00:00
2020-07-13 00:00:00
2020-07-20 00:00:00
2020-07-27 00:00:00
2020-08-03 00:00:00
2020-08-10 00:00:00
2020-08-17 00:00:00
2020-08-24 00:00:00
2020-08-31 00:00:00
2020-09-07 00:00:00
2020-09-14 00:00:00
2020-09-21 00:00:00
2020-09-28 00:00:00
2020-10-05 00:00:00
date

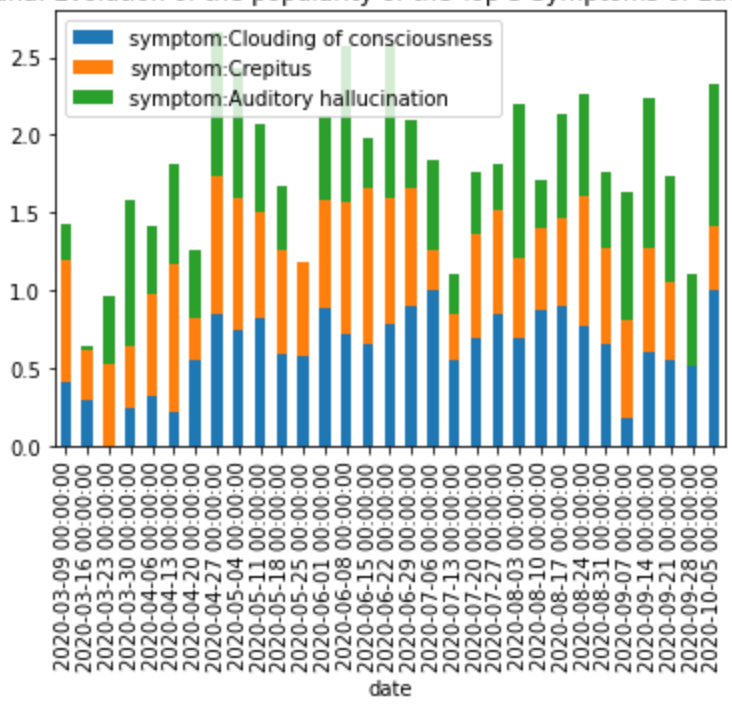

District of Columbia: Evolution of the popularity of the Top 3 Symptoms of Each Region
symptom:Clouding of consciousness
symptom:Auditory hallucination
symptom:Polydipsia
0.0
0.5
1.0
1.5
2.0
2.5
2020-03-09 00:00:00
2020-03-16 00:00:00
2020-03-23 00:00:00
2020-03-30 00:00:00
2020-04-06 00:00:00
2020-04-13 00:00:00
2020-04-20 00:00:00
2020-04-27 00:00:00
2020-05-04 00:00:00
2020-05-11 00:00:00
2020-05-18 00:00:00
2020-05-25 00:00:00
2020-06-01 00:00:00
2020-06-08 00:00:00
2020-06-15 00:00:00
2020-06-22 00:00:00
2020-06-29 00:00:00
2020-07-06 00:00:00
2020-07-13 00:00:00
2020-07-20 00:00:00
2020-07-27 00:00:00
2020-08-03 00:00:00
2020-08-10 00:00:00
2020-08-17 00:00:00
2020-08-24 00:00:00
2020-08-31 00:00:00
2020-09-07 00:00:00
2020-09-14 00:00:00
2020-09-21 00:00:00
2020-09-28 00:00:00
2020-10-05 00:00:00
date


Idaho: Evolution of the popularity of the Top 3 Symptoms of Each Region
symptom:Aphonia
symptom:Ventricular fibrillation
symptom:Urinary urgency
1.0
0.8
0.6
0.4
0.2
0.0
2020-03-09 00:00:00
2020-03-16 00:00:00
2020-03-23 00:00:00
2020-03-30 00:00:00
2020-04-06 00:00:00
2020-04-13 00:00:00
2020-04-20 00:00:00
2020-04-27 00:00:00
2020-05-04 00:00:00
2020-05-11 00:00:00
2020-05-18 00:00:00
2020-05-25 00:00:00
2020-06-01 00:00:00
2020-06-08 00:00:00
2020-06-15 00:00:00
2020-06-22 00:00:00
2020-06-29 00:00:00
2020-07-06 00:00:00
2020-07-13 00:00:00
2020-07-20 00:00:00
2020-07-27 00:00:00
2020-08-03 00:00:00
2020-08-10 00:00:00
2020-08-17 00:00:00
2020-08-24 00:00:00
2020-08-31 00:00:00
2020-09-07 00:00:00
2020-09-14 00:00:00
2020-09-21 00:00:00
2020-09-28 00:00:00
2020-10-05 00:00:00
date

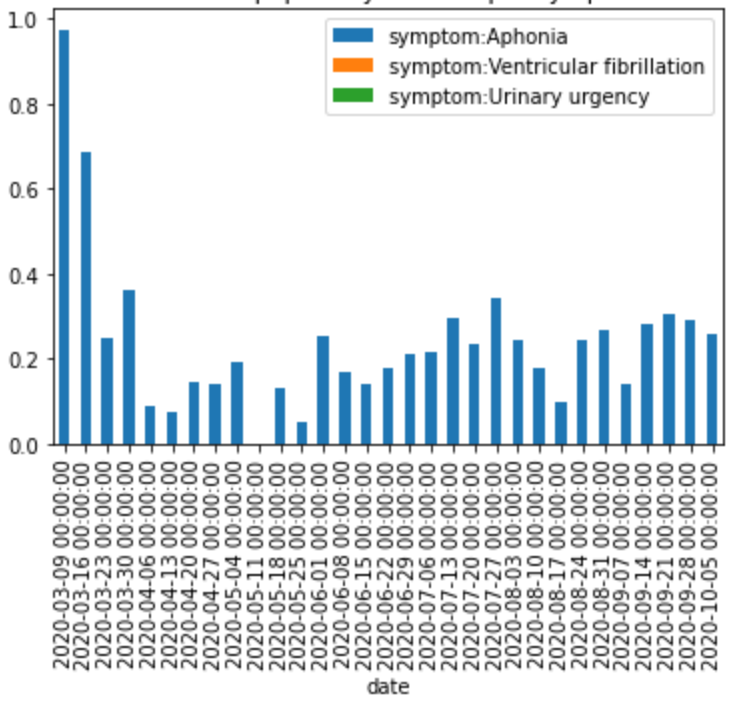
Maine: Evolution of the popularity of the Top 3 Symptoms of Each Region



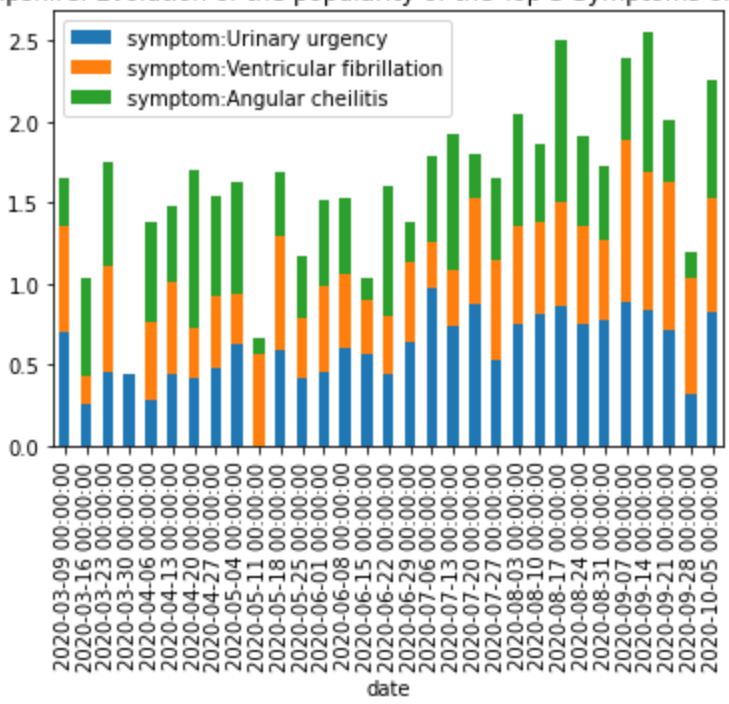
Montana: Evolution of the popularity of the Top 3 Symptoms of Each Region



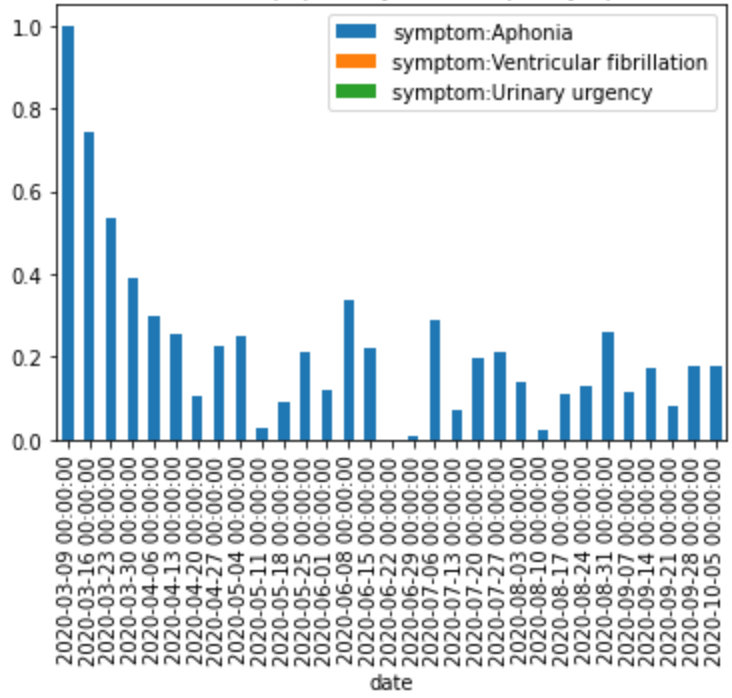
Nebraska: Evolution of the popularity of the Top 3 Symptoms of Each Region



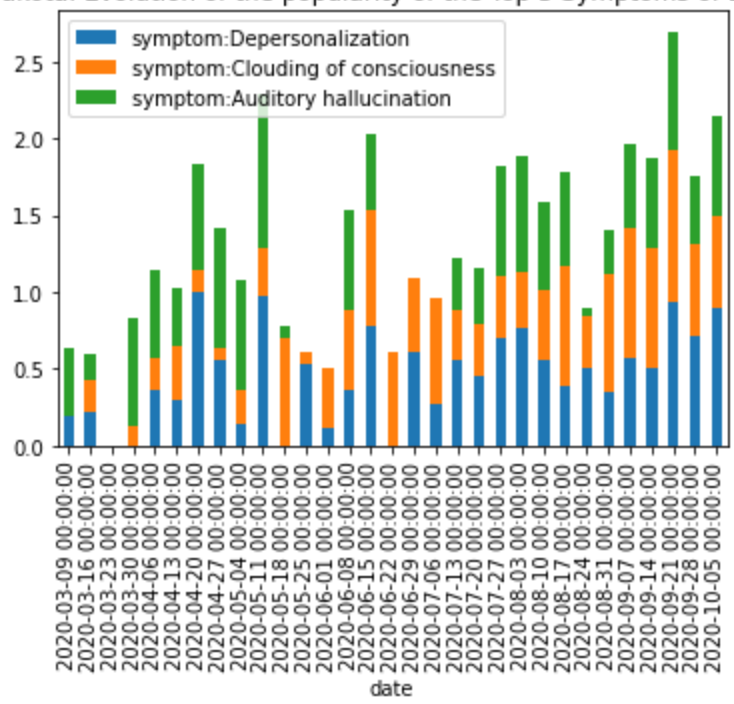
New Hampshire: Evolution of the popularity of the Top 3 Symptoms of Each Region



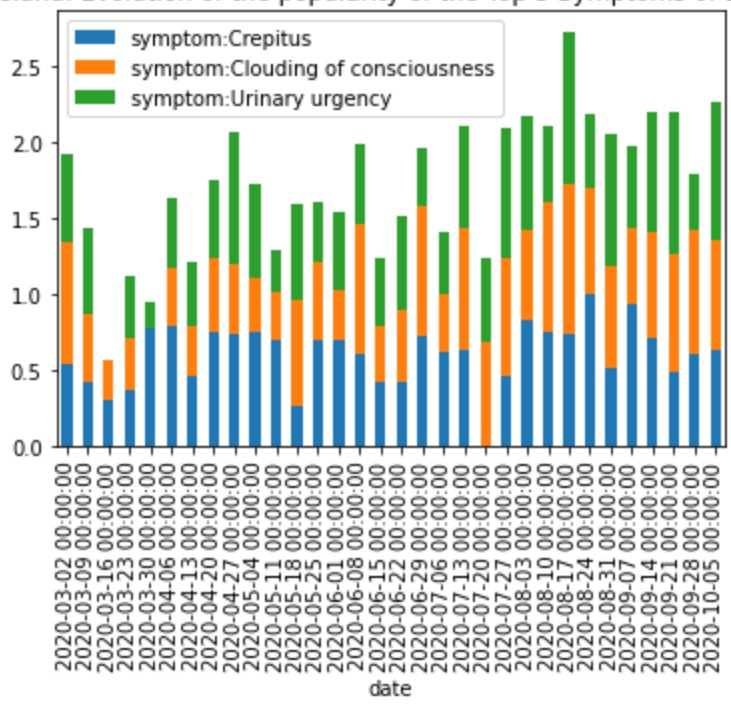
New Mexico: Evolution of the popularity of the Top 3 Symptoms of Each Region



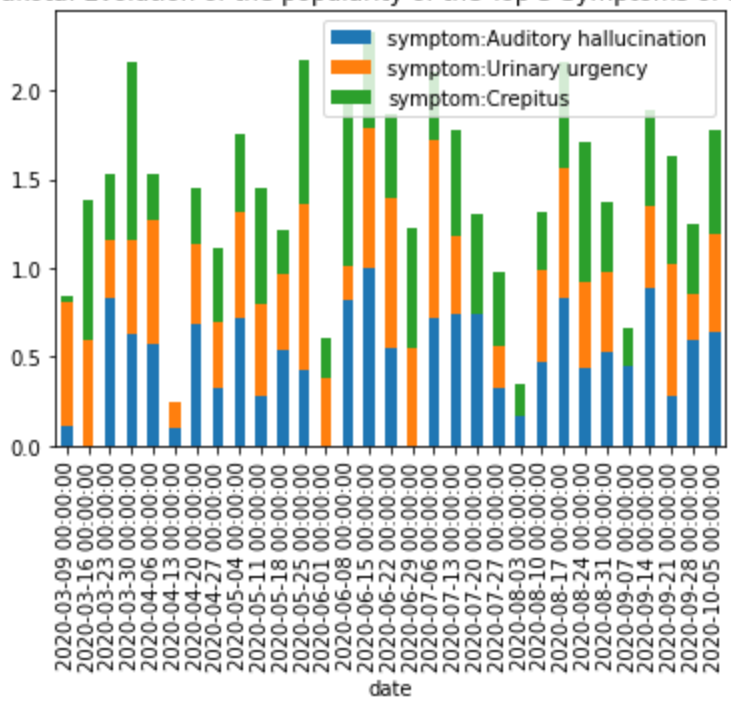
North Dakota: Evolution of the popularity of the Top 3 Symptoms of Each Region

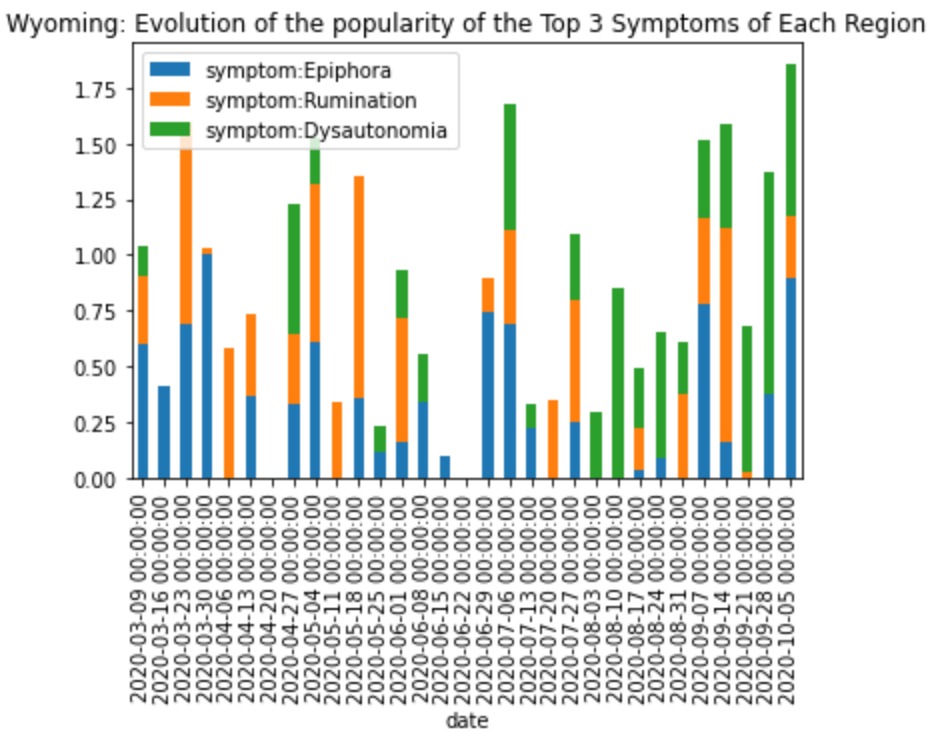


Rhode Island: Evolution of the popularity of the Top 3 Symptoms of Each Region



South Dakota: Evolution of the popularity of the Top 3 Symptoms of Each Region




Wyoming: Evolution of the popularity of the Top 3 Symptoms of Each Region
symptom:Epiphora
symptom:Rumination
symptom:Dysautonomia
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
2020-03-09 00:00:00
2020-03-16 00:00:00
2020-03-23 00:00:00
2020-03-30 00:00:00
2020-04-06 00:00:00
2020-04-13 00:00:00
2020-04-20 00:00:00
2020-04-27 00:00:00
2020-05-04 00:00:00
2020-05-11 00:00:00
2020-05-18 00:00:00
2020-05-25 00:00:00
2020-06-01 00:00:00
2020-06-08 00:00:00
2020-06-15 00:00:00
2020-06-22 00:00:00
2020-06-29 00:00:00
2020-07-06 00:00:00
2020-07-13 00:00:00
2020-07-20 00:00:00
2020-07-27 00:00:00
2020-08-03 00:00:00
2020-08-10 00:00:00
2020-08-17 00:00:00
2020-08-24 00:00:00
2020-08-31 00:00:00
2020-09-07 00:00:00
2020-09-14 00:00:00
2020-09-21 00:00:00
2020-09-28 00:00:00
2020-10-05 00:00:00
date