Analyzing COVID-19 Search Trends and Hospitalization

COMP 551, Fall 2020, Mcgill University

October 21, 2020

Team members:

Narry Zendehrooh Kermani 260700556

Elias Tamraz 260871813

Abdelhadi Ghalmi 260587573

Abstract

Coronavirus disease 2019 (COVID-19) is a contagious respiratory and vascular disease. The disease was first identified in December 2019 in Wuhan, China. The outbreak was declared a Public Health Emergency of International Concern in January 2020, and a pandemic in March 2020. In order to help improve on understanding how COVID-19 impacts communities and detecting outbreak earlier, we will be using the COVID-19 Search Trends symptoms dataset, which is a time series of Google searches for a set of health symptoms, signs and conditions and the COVID-19 Hospitalization Cases dataset, which is a public time series for COVID-19 data sources. We will investigate the performance of standard time series prediction models namely, k-nearest neighbours and decision trees regression, on predicting COVID-19 hospitalization cases from related symptoms search. We found that the k-nearest neighbour regression approach achieves better accuracy than the decision tree but was slower to train.

Introduction

The tasks consist of processing the data, visualization and clustering, and finally, deploy supervised learning frameworks such as K-nearest neighbours (KNN) and decision trees. In preprocessing the hundreds of features were trimmed down to 16. The graphs for symptoms' popularity (5) across time for different regions show a common pattern in which search frequency appears to remain stable or decay to a baseline level. The PCA analysis helped visualize the 16-features dataset search trends in 2D, closely followed by the clusters obtained on the original data and the clusters on the reduced dimensionality. Finally, KNN was deployed on 5-fold cross-validation on region splits (80-20 split) and on date splits (before and after 2020-08-10) to find the model's performance (MSE and running time). The results can be compared with a decision tree model showing KNN is consistently a better option.

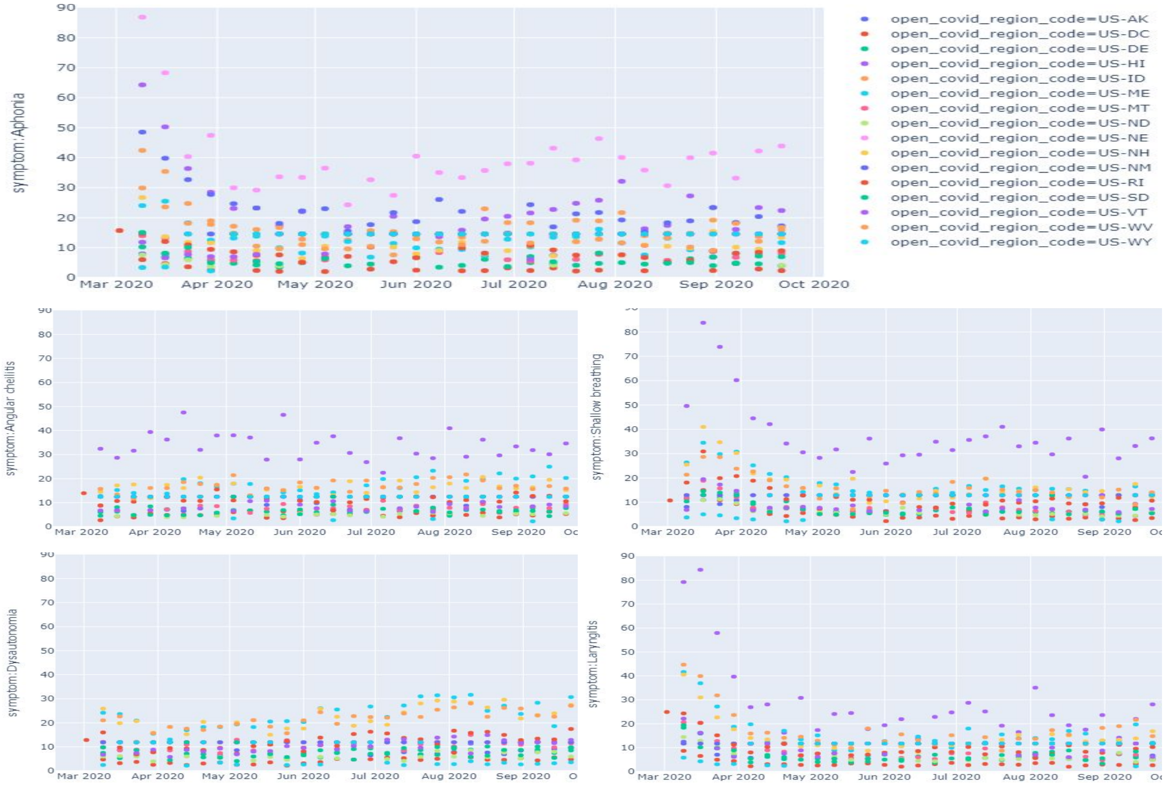
Datasets

We are going to use the COVID-19 Search Trends symptoms and COVID-19 Hospitalization Cases Dataset for our tasks. COVID-19 Search Trends symptoms reflect the volume of Google searches for a broad set of health symptoms, signs, and conditions. For each day, it counts the searches mapped to each of these symptoms and organizes the data by geographic regions of the US. The resulting dataset is a weekly time series for each region showing the relative frequency of searches for each symptom. COVID-19 Hospitalization Cases is an open-source pipeline that aggregates public COVID-19 data sources into a single dataset. The data includes daily time series data for COVID-19 cases, deaths, tests, hospitalizations, discharges, etc. for all countries around the world. From the first dataset, we only keep the symptoms that have data available for at least 35% of the rows, so only 9 symptoms remain. From the second dataset, we only pick the new hospitalized cases for US regions. Afterward, we convert daily data of the second dataset into weekly data to be able to merge it with the first dataset. The merged dataset consists of 9 symptoms and hospitalized cases from March 9, 2020, to September 21, 2020. In the end, we replace the empty values with the mean of their column.

Results

In order to be able to visualize the distribution of search frequency of each symptom aggregated across different regions changes over time for the most popular k symptoms. An algorithm was run to systematically sum the frequencies of all features for each region and the results were sorted to obtain the 5 most popular symptoms.

Figure 1: Popular symptoms and their search frequencies across regions and time.



Since that data set contains a big number of features then we need to lower the dimensional space in order to be able to visualize the data set. This has been done using Principal Component Analysis (PCA) to lower the dimensional space into (2D).

Figure 2: PCA reduction along 2 principal components

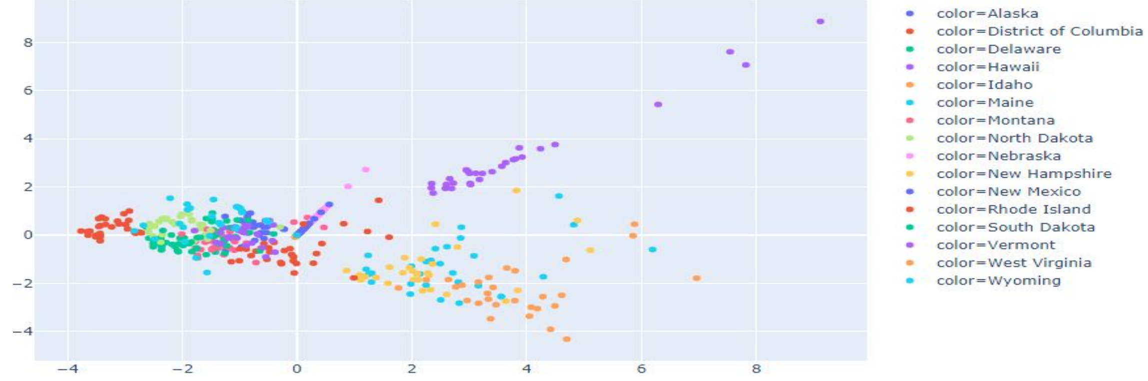
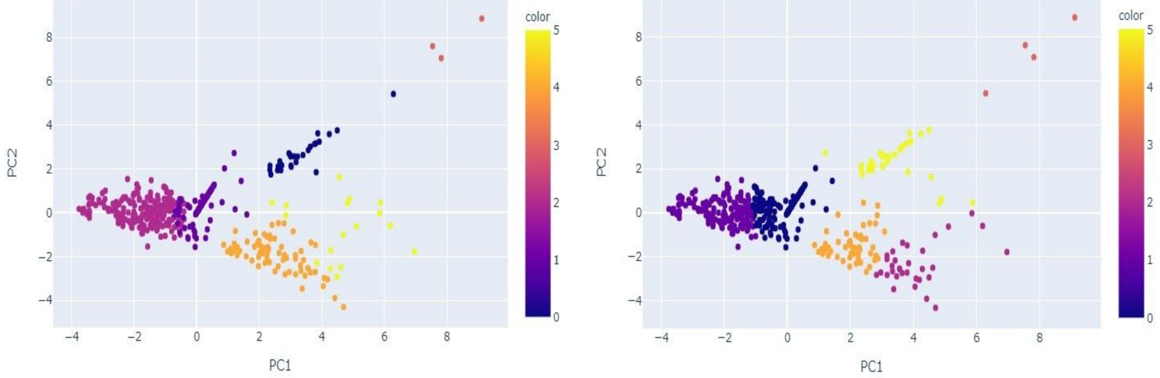


Figure 4: Clustering (k-means) on PCA original data [left], lower dimensional data [right]



We found that the PCA-reduced data and the original data clusters were somehow consistent, the majority of the pointes were colored the same way (even if the color is different between the two graphs).

Figure 6: KNN validation MSE for region split scheme

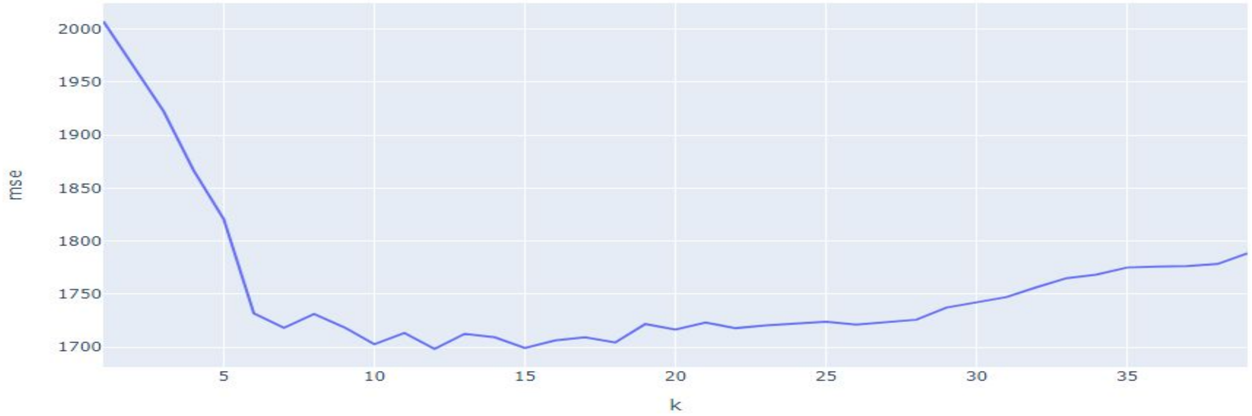
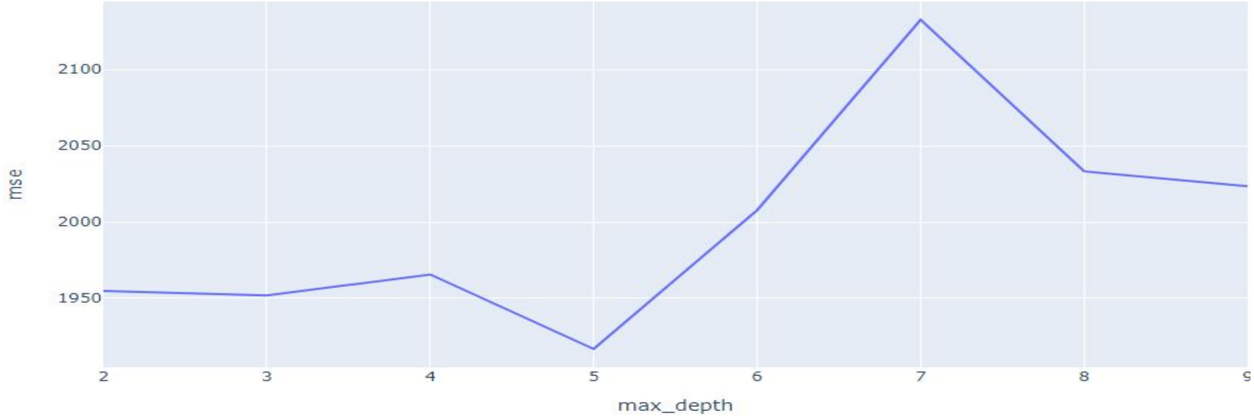


Figure 7: Decision Tree validation MSE for region split scheme



Note that both methods are non-parametric. The hyper-parameters (number of neighbours and maximum tree depth) are varied and plotted against MSE. The argmin is chosen for each model (k = 12 and max_depth = 5).

Table 1: regression performance (MSE) for KNN and decision trees

	KNN	DT
Region split (80-20)	639	736
Date split (before/after 2020-08-10)	140	286

Discussion and Conclusion

The point that stood out the most when plotting the symptoms' popularity across regions and time is that a subset of symptoms was extremely popular in the search frequencies at the beginning of the pandemic while others were more or less frequent around a baseline level (Figure 1). As months were passed the search frequencies converged around the baseline level. This is true for at least most specific signs and symptoms [aphonia, shallow breathing, angular cheilitis and laryngitis] with the outlier being [dysautonomia], a more general condition associated with multiple symptoms. While no strong conclusion can be inferred from this result it is an interesting find reflecting the panic-driven nature of this pandemic during its early stages.

Statement of Contributions

Narry Zendehrooh Kermani: Task 1, Code Cleanup, Task 3.3
Elias Tamraz: Task 2
Abdelhadi Ghalmi: Task 3, Bonus