Analyzing COVID-19 Search Trends and Hospitalization

COMP 551 Project 1

Zhangyuan Nie - 260924723   Linghui Huang - 260831346   Chen Hu - 260859053

Abstract

In the project, we explored different machine learning techniques to analyze probable correlations in the data between symptom search trends and the hospitalization cases for COVID-19 in different regions in the United States. We discovered that several regions shared similar search patterns by applying Principal Component Analysis (PCA), and visualizing the data using k-means clustering. Furthermore, we investigated the performance of two supervised learning regression models, namely, K-Nearest Neighbours (KNN) and Decision Trees, on predicting COVID-19 hospitalization cases from related symptoms search trends. We found that Decision Trees achieved the best accuracy overall when we split by dates, although the error rates of all methods were quite high to draw any meaningful conclusions.

Introduction

The COVID-19 outbreak is one of the worst global pandemics that have occurred in the 21st century. With the ever-increasing internet penetration rate around the world, researchers are able to analyze possible correlations between the internet usage data and the pandemic. The Search Trends dataset[3] and the COVID hospitalization cases dataset[1] were used in this project to perform analysis in order to recognize possible patterns in the data. Similar work has already been done on these datasets. For example, Ayyoubzadeh et al. used it to predict COVID-19 incidences in Iran with deep learning techniques[2], and Heerfordt et al. also used Google trends data to investigate interest in smoking cessation during the pandemic[4]. The Search Trends dataset demonstrated relative regional popularity of certain symptoms from google search, while the hospitalization dataset includes time-series data for COVID-19 hospitalizations. We discovered that several regions shared similar search patterns for some symptoms by applying Principal Component Analysis (PCA) to visualize the search trend data in lower dimensions and performing K-Means clustering. We then used K-Nearest Neighbours (KNN) and Decision Trees to predict the number of hospitalizations. The dataset was split into training and validation sets using two different strategies: splitting by region and splitting by date. Splitting by date strategy produced a lower error rate in both regression models. We found that, when splitting by dates, Decision Trees achieved the lowest validation error overall, although all methods had quite significant error rates because there were a fair amount of missing values in the datasets.

Datasets

The Search Trends dataset provided by Google reflects the relative volume of weekly Google searches performed on about 400 health symptoms related to COVID-19 organized by geographic region in the United States[3]. The data is already normalized such that comparisons between any two symptoms over any time interval can be made. In order for us to compare the popularities across regions, we normalized them again by multiplying the given value by the population of a region, fetched from the U.S. Census Bureau[5]. We choose this normalization scheme for its simplicity which should make comparisons across regions more meaningful. However, this method does not take into account the actual Internet population that performs Google searches. For example, some states may have a lower internet penetration rate while others may have more accessible medical professionals. For this dataset,

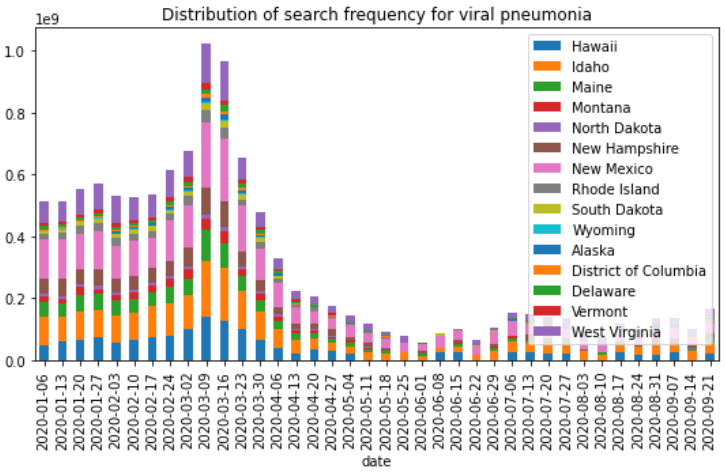

Figure 1: Distribution of search frequency of viral pneumonia in different regions

a symptom is dropped if more than half of the values are missing. We also chose to replace missing data with the median of each column for PCA and regressions.

The hospitalization dataset provided by The Atlantic contains daily COVID-19 hospitalization cases in the United States separated by regions[1]. We dropped the regions with invalid data and converted the daily data to weekly. This way, we were able to merge the two datasets by region and date.

To gain a better understanding of the data, we first visualized the distribution of search frequency for a given symptom over time in different regions (Fig. 1). This graph shows when and where the highest search frequency for the symptom occurred. Then we plotted a heatmap of the search frequency for each region and each symptom (Fig. 2), which helps us to understand the different search trends between each region.

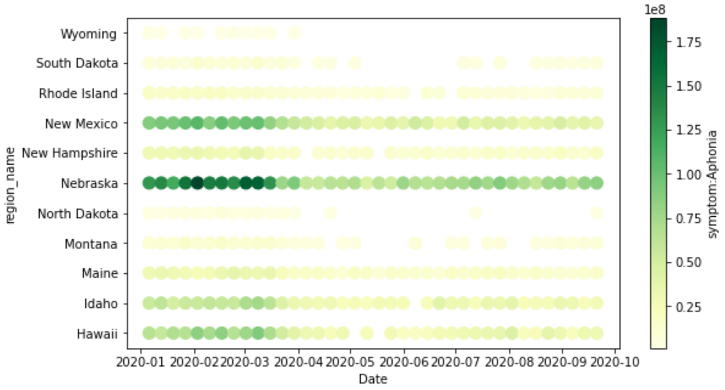

Figure 2: Heatmap of search frequency of aphonia by region

Lastly, we plotted the ratio of search frequency of each symptom by comparing the frequency of each region to the mean of that symptom in that region. Pink regions had more searches during the given week for the symptom. Green regions had fewer related searches. Some regions may have insufficient data for a given week (Fig. 3.1 to 3.4).

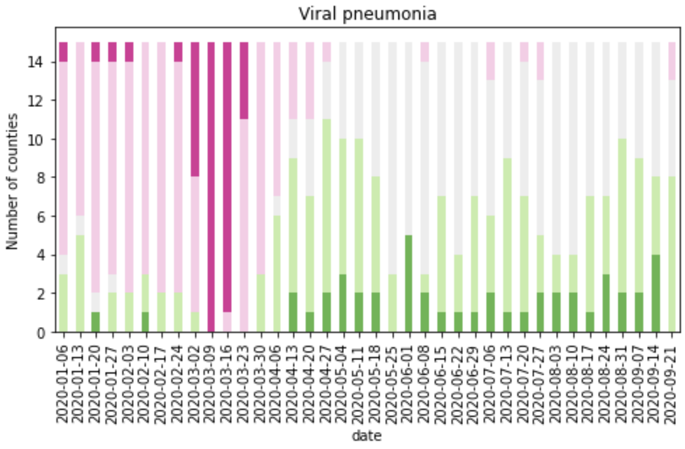

Figure 3.1: Ratio of search frequency for viral pneumonia

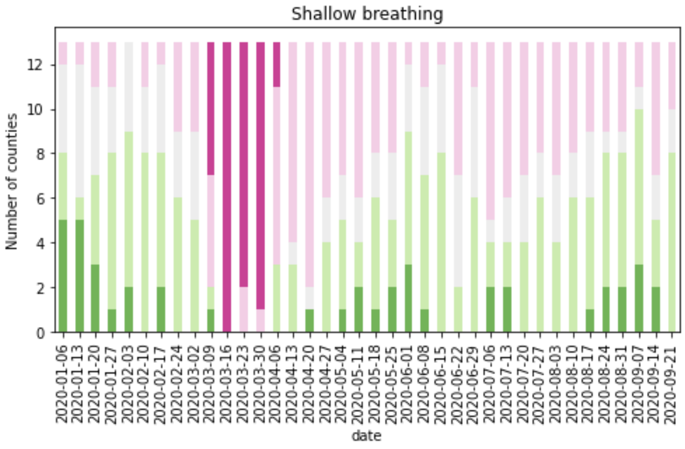

Figure 3.2: Ratio of search frequency for shallow breathing

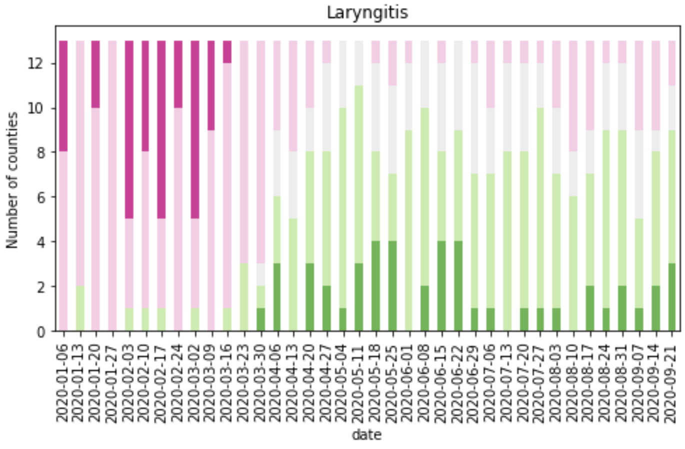

Figure 3.3: Ratio of search frequency for laryngitis

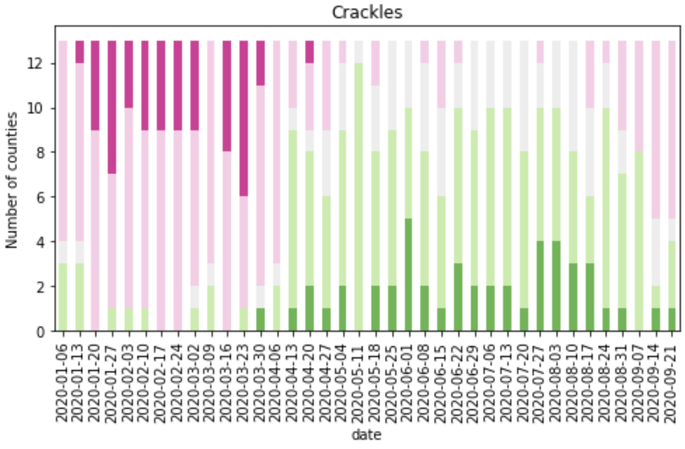

Figure 3.4: Ratio of search frequency for crackles

We can easily see some patterns from our visualizations: All the symptoms that are related to COVID-19, such as laryngitis, seem to be most searched for during March, which is arguably the start of the outbreak in the US.

Results

To further understand the data, principal component analysis (PCA) was performed to visualize the search trends data in 2D (Fig. 4, left). The missing values in the dataset were replaced by the median of the column, and the PCA function from sklearn package was used to generate the principal components. A scree plot was made to visualize the quality of the first few principal components (Fig. 4, right). The height of each bar represents the variance ratio of each principal component.

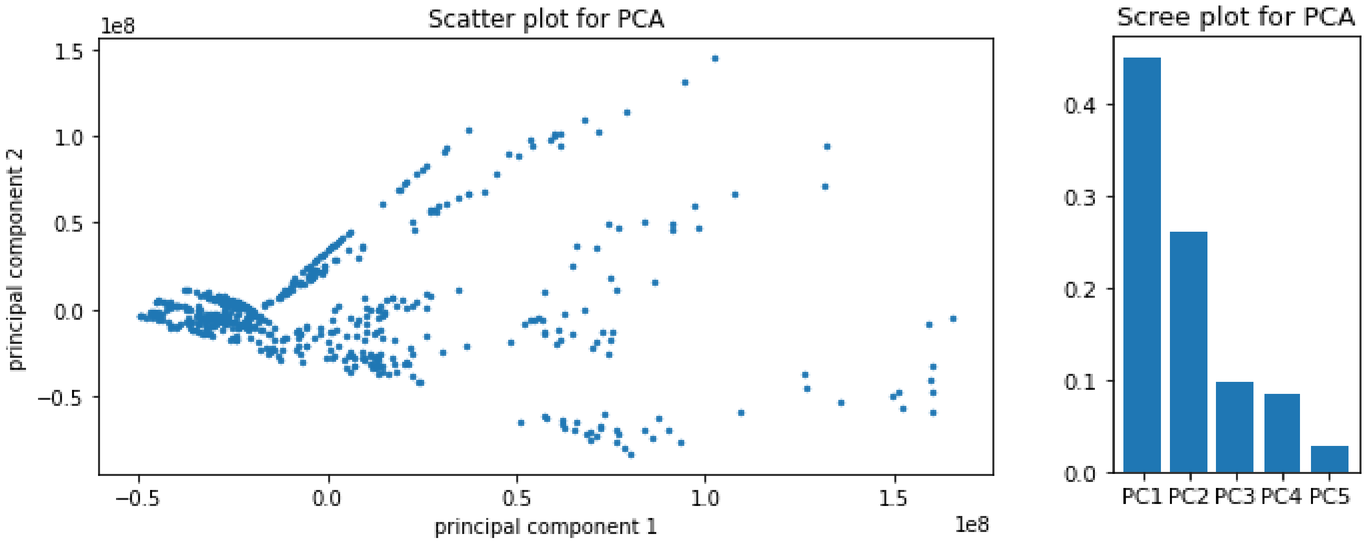

Figure 4: Principal component analysis plot of search trend data, and the corresponding scree plot

To investigate the possible similarities between search trends data in different regions at different times, K-Means clustering was performed on original and PCA-reduced search trends data. Firstly, the k value representing the number of clusters was determined by the Elbow method. Within-Cluster-Sum of Squared Errors (SSE), also known as inertia, were calculated for different values of k (Fig. 6), and the k was chosen for which SSE first starts to diminish, at the "elbow point". The elbow point was determined using KneeLocator from kneed package, and the resulting k value for K-Means clustering was 3 for both original data and PCA-reduced data.

Then K-Means clustering was performed on original and PCA-reduced search trend data (Fig. 7) using KMeans function from sklearn package. Both datasets were plotted by principal components 1 and 2 to visualize them. As shown from the graphs, the groupings were mostly the same except for about 2 points. This indicates that the PCA did a good job reducing the dimensionality without losing too much information.

In addition, to compare the performance of two supervised learning models, K-Nearest Neighbors (KNN) and Decision Trees, for predicting the hospitalization cases given the search trends data, the dataset was split based on regions and based on time, and the corresponding validation errors were reported.

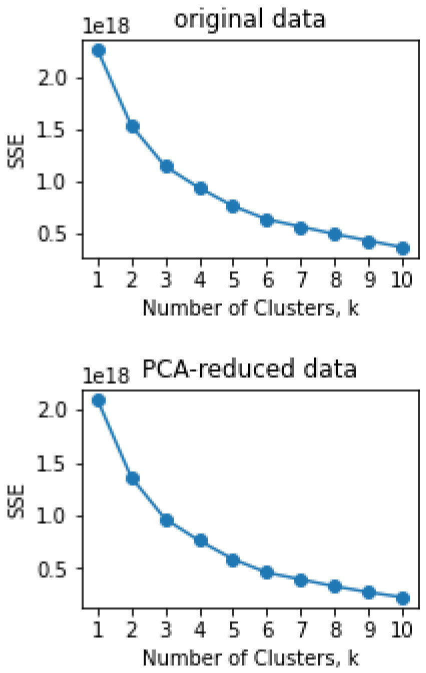

Figure 6: Within-Cluster-Sum of Squared Errors (SSE) plots for original and PCA-reduced search trends data. k value for K-Means clustering was taken from the elbow point at k=3 for both data

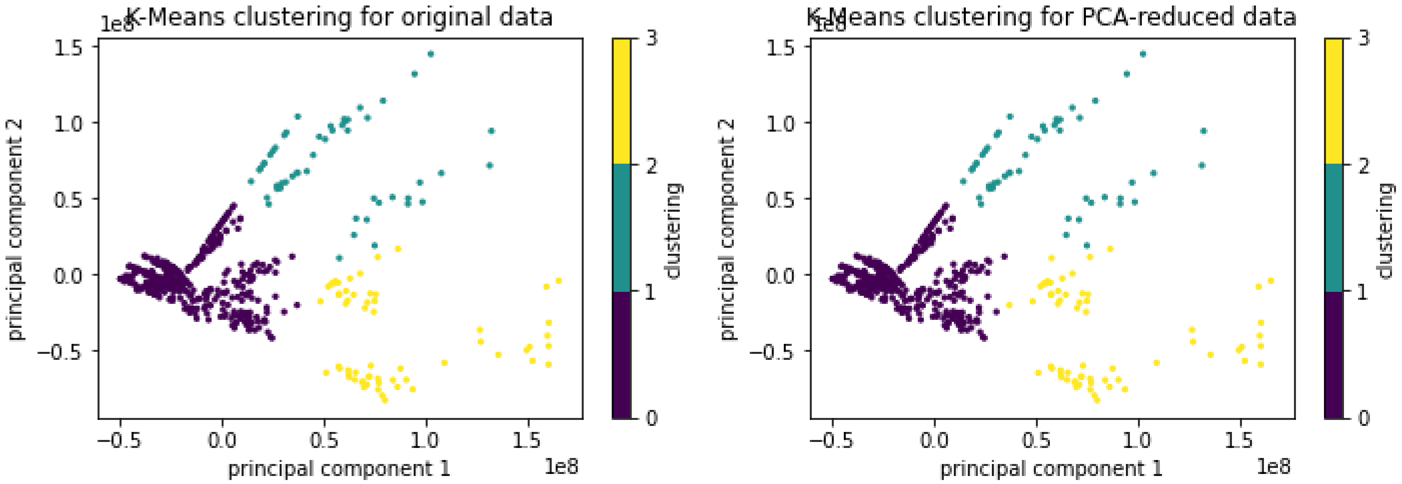

Figure 7: K-Means clustering on original and PCA-reduced search trends data, k = 3

For the splitting by regions strategy, the dataset was first divided into 80% regions in the training set and 20% in the validation set, and doing KNN regression using a different number of neighbors (k values) for 5 times to report the validation errors (mean absolute error, MAE), and the average becomes the 5-fold cross-validation error (Fig. 8, left).

For KNN, we can observe from figure 8 that there are two local minimums on the graph. This indicates that the MAE for KNN regression is small for k = 1 and k = 6. However, if we look at their respective mean squared error (MSE) instead, it becomes apparent that MSE decreases as k increases. This is due to the fact that MAE does not punish large errors in prediction. Very small k for KNN regression will give a finitely textured decision boundary which is more sensitive to noise and outliers while larger k gives more rigid decision boundaries. To predict COVID-19 hospitalization cases from related symptoms search, we have data of a large size with lots of "NaN" values which create noise for the prediction and therefore it's more reliable to compare with more neighbours. The 5-fold cross-validation error was minimal at k = 6.

The same 5-fold cross-validation was also reported using Decision Trees regression when splitting by regions (Fig. 8, right). The mean cross-validation error for Decision Tree is greater than the minimal cross-validation error of KNN, indicating KNN performs better for splitting by regions.

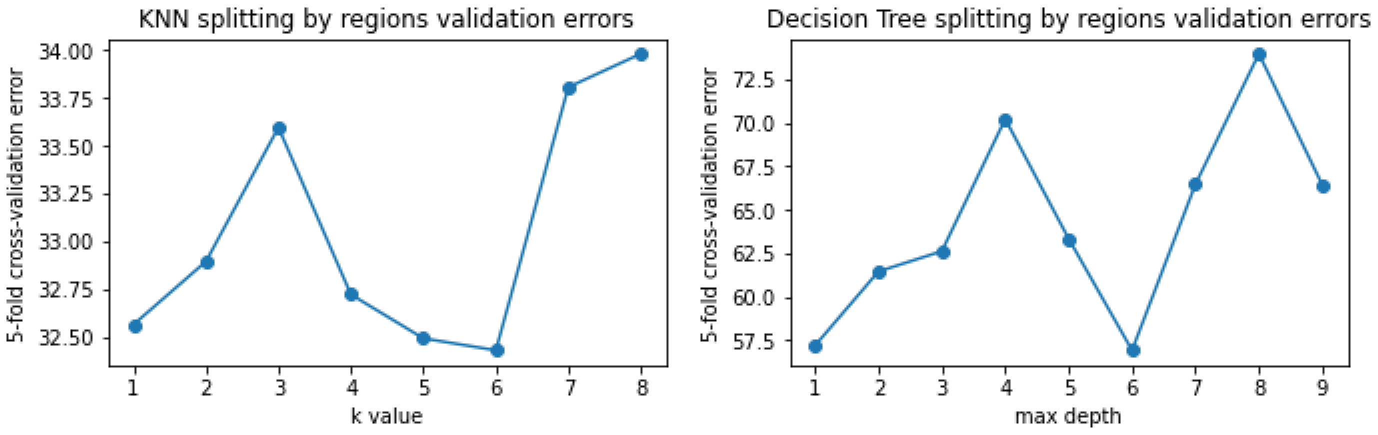

Figure 8: 5-fold cross-validation errors (MAE) when splitting by regions, by K-Nearest Neighbors regression using different k values (distance metric: Euclidean distance), and by Decision Trees regression using different max depths (training criterion: MAE)

For the splitting by dates strategy, search trend data after 2020-08-10 was kept as the validation set, and the rest of the data were used as the training set. The validation errors of KNN using different k values were reported (Fig. 9, left), with the lowest error at k = 4. The validation errors of Decision Tree when splitting by dates were reported as well (Fig. 9, right), with the lowest error at max depth = 7. Decision Trees regression achieved better accuracy than KNN.

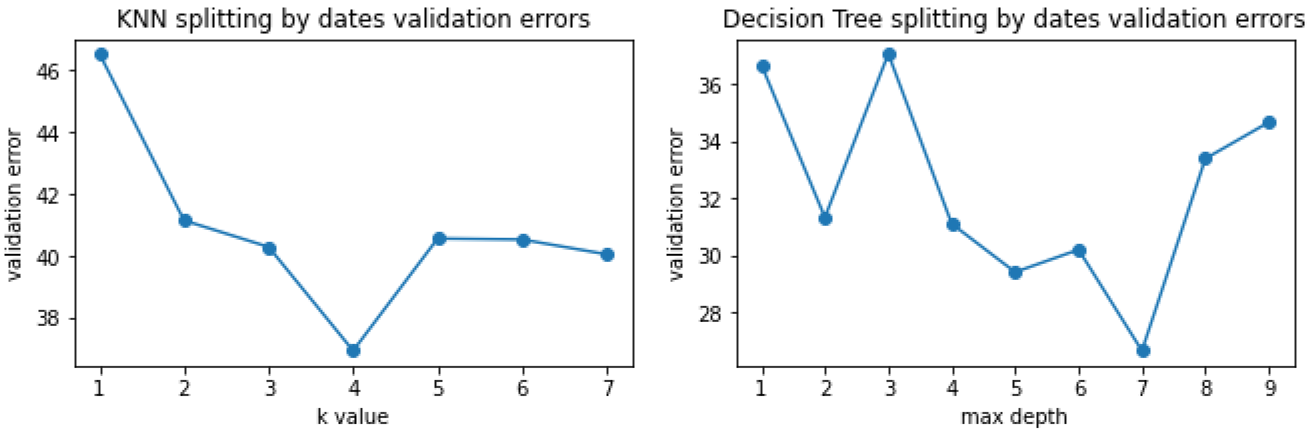

Figure 9: validation errors (MAE) when splitting by dates, by K-Nearest Neighbors regression using different k values (distance metric: Euclidean distance), and by Decision Trees regression using different max depths (training criterion: MAE)

Overall, we had better accuracy for both KNN and Decision Trees models when splitting by date than by regions. This indicates that it is more reasonable to predict the future from the past instead of predicting other regions based on some regions. The regression method that achieved the lowest validation error was Decision Tree when splitting by dates at max depth = 7.

We also attempted to train data in each region separately. The mean validation error for KNN was 39.6 with k = 2, and that for Decision Trees was 33.4 with max depth = 5. The validation errors were unexpectedly higher than previous methods for both models, which may be due to the large portion of missing values in a lot of the regions.

Discussion and Conclusion

K-Nearest Neighbors regression uses feature similarity to predict the values for new data points. Since the output value depends on the neighbors of the points, the k value should be wisely selected. Some downsides of KNN included the large computational cost when the sample size is large, and the need for proper scaling for fair treatment of features. Compared to KNN, Decision Tree doesn’t rely much on data normalization. Since the information gain (entropy) is used as the criteria for selecting the termination condition for the leaves, a large input dataset helps this recursive algorithm give a more precise prediction. It also has some disadvantages, including that it’s more likely to overfit the training set and the model’s sensitivity to outliers. Thus the selection of these regression methods shall depend on the type of data you’re dealing with.

When predicting hospitalization cases given search trends data, Decision Tree with splitting by dates had the lowest cross-validation error, and hence was the best method in this case. However, since we do not have the search trends data from previous years, we could not know which data were related to COVID-19 and which were due to other diseases like influenza and pneumonia and could not normalize the data accordingly. Also, the datasets contained a lot of missing values. Those might be some reasons behind the high error rates of the regression models.

Some possible improvements for future directions includes doing analysis using more data, normalizing the data based on internet usage instead of just population, trying other better split methods such as splitting based on PCA groupings, and optimizing other parameters of the regression models.

Statement of Contributions

Zhangyuan Nie worked on task 1, task 2.1, and task 2.2. Linhui Huang worked on task 1, task 2.2, 2.3, and a small part of task 3. Chen Hu worked on task 3 exclusively. All of us wrote the corresponding sections of the present report.

References

[1] The Atlantic. "COVID-19 Tracking Project." https://github.com/COVID19Tracking/covid-tracking-data/tree/master/data. Accessed 6 October 2020.

[2] Ayyoubzadeh, Seyed Mohammad, et al. "Predicting COVID-19 Incidence Through Analysis of Google Trends Data in Iran: Data Mining and Deep Learning Pilot Study." *JMIR Public Health and Surveillance*, 2020. doi:10.2196/18828.

[3] Google LLC. "Google COVID-19 Search Trends symptoms dataset." http://goo.gle/covid19symptomdataset. Accessed 7 October 2020.

[4] Heerfordt, C., and I.M. Heerfordt. "Has there been an increased interest in smoking cessation during the first months of the COVID-19 pandemic? A Google Trends study." *Public Health*, vol. 183 (2020): 6-7. doi:10.1016/j.puhe.2020.04.012

[5] U.S. Census Bureau. "State Population Totals and Components of Change: 2010-2019." https://www.census.gov/data/tables/time-series/demo/popest/2010s-state-total.html. Accessed 12 October 2020.