# Analyzing COVID-19 Search Trends and Hospitalization

Julian Armour (260804046), Timothy Keding (260955532), Saksham Mungroo (260768072)

## Abstract

In this project we tried to visualize the evolution in popularity of various symptoms across different regions over time. We found out that there might exist a correlation between the number of hospitalization cases and the search trends of those symptoms. Using PCA, we took a high dimensional dataset of all symptoms and reduced it to 2D and 3D graphs that could be visualized. We then determined the optimal amount of clusters given the resulting data, and used k-means to visualize clusters within the displayed data. We also explored possible groups in search trends data using knn and concluded that those groups remained similar in lower dimensions. Finally, we explored the predictive performance of K-NN and Decision Trees. Validation of these models was performed using two different schemes. The performance of both models remained similar for both schemes, although the errors in prediction varied greatly depending on the scheme that was used.

## Introduction

The purpose of this project was to analyze google search trends for medical symptoms and their relation to new hospitalization cases in the US. A dataset from google on search trends was merged with a dataset on hospitalization cases, and trimmed to exclude regions not in the US, as well as searched symptoms with inefficient data to be relevant to our analysis. The first task was to visualize the trend of the most popular symptoms over time. The 15 most popular symptoms aggregated across all regions were sampled and their popularity was plotted over time. Then, PCA was used to visualize the data in 2 and 3 dimensions and K-means was used to find clusters. One important observation was the consistency of the clusters which remained 73% similar in 3D compared to the original dimensions. The next task was training regression K-nearest neighbors (K-NN) and a decision tree on the data to predict new hospitalization cases. Both model performances were measured by Mean Squared Error (MSE) and evaluated using two different validation schemes.

## Datasets

There are two datasets used for this project. The first contained google search trends in the US for various medical symptoms over time. The second contained covid-19 information across multiple regions, including hospitalization data over time. The google trend data is given in a weekly time series and the hospitalization as a daily time series. Both datasets divide their samples by region. Hospitalization data was resampled to weekly starting on monday (in order to match the google data) and was merged into the search trend set so that regions and timestamps match. The google search trends dataset only contained data for regions within the US, so any hospitalization data not from the US was removed. The google search trend data is only normalized within regions, so comparing two regions with this data is non-sensible. Thus, we introduced the following normalization scheme: for each region we find the mean of all symptoms popularity data and divide each datapoint by it. This gives us a relative popularity to the mean for that region, which allows comparison between regions. Depending on the tasks at hand, there are several filtering functions that were used. Among those filtering options, the ones which provided the most valuable results included: removing symptoms with little data, removing regions with little data, and keeping the 10-15 most popular symptoms.

# Results

Following cleaning of the data which involved removing all symptoms that had no or very little entries and aggregation of all the regions with respect to date, the 15 most popular symptoms were sampled. The search trends for each of these symptoms was then visualized over time, and a selection can be seen in figure 1.

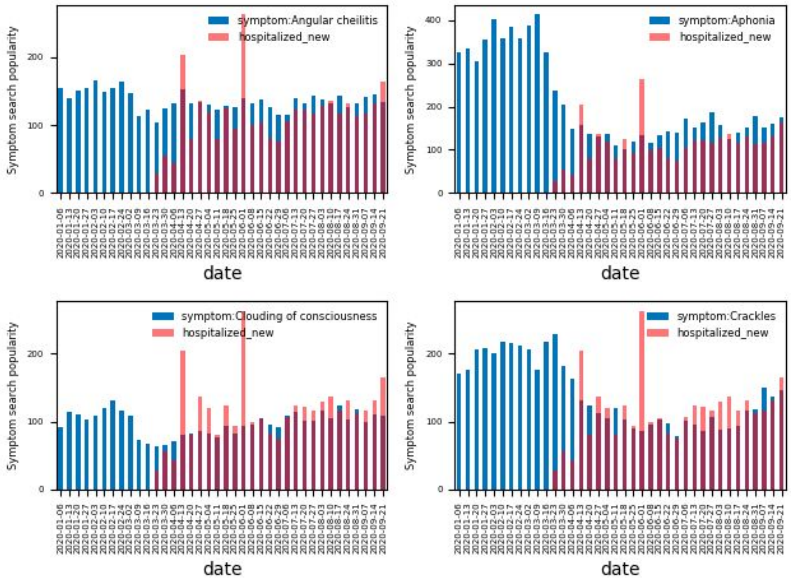
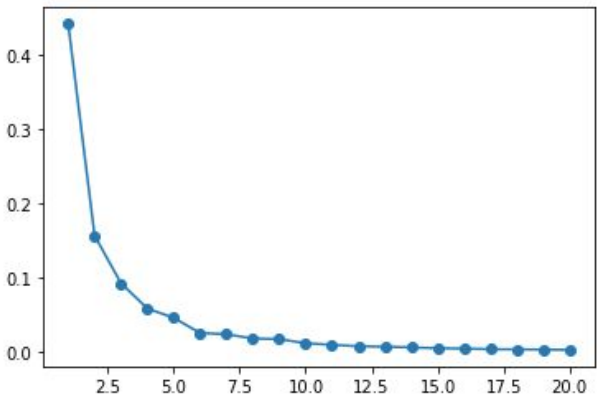



*Figure 1: 4 of the 15 most popular symptoms, displaying the search results for a particular symptom, as well as weekly hospitalization cases aggregated across all regions.*

Most of the symptoms presented an increase in popularity during the period of late february to mid march followed by a rapid decrease and another slight growth. To investigate further, the number of new hospitalizations over the same time period was visualized on the subplots. The number of new hospitalizations was multiplied by a constant c (c ≃ 1/60) such that its trend could easily be compared to the evolution in popularity of our 15 chosen symptoms. This procedure does not affect the relative trend evolution between symptom popularity and number of new cases since we are not interested in the actual values but really in their relative evolution (does “hospitalized_new” increases when the popularity of symptom x increases for example?). For the majority of the 15 symptoms chosen, a positive correlation between their popularity and the number of new hospitalizations was reported.

Once symptoms had been visualized over time, we proceeded to take a high dimensional graph across all symptoms, and reduce the dimensionality using PCA. Utilizing a calculation for variance, we concluded that reducing the PCA to a dimensionality of 6 produced the optimal results, compensating for 80 percent of the variance, as can be seen in figure 2.1.



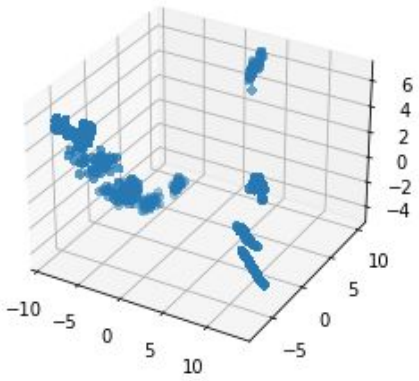




*Figure 2.1: An elbow graph demonstrating the variance captured by each additional dimension when reducing using PCA. The first dimension captures roughly 43 percent of the variance, while the second dimension only accounts for an additional 12 percent.*
*Figure 2.2: A three dimensional graph displaying points represented by weeks, where the dimensions are the eigenvectors of a high dimensional graph where each dimension is a searched symptom.*

Once we're able to reduce the dimensionality, an attempt to determine the ideal number of clusters in 2D, 3D, 6D (optimal dimension found previously from the optimal number of PCs) and original dimensions was made using the elbow method. Figure 3 contains a graph showing the result for 6 dimensions.

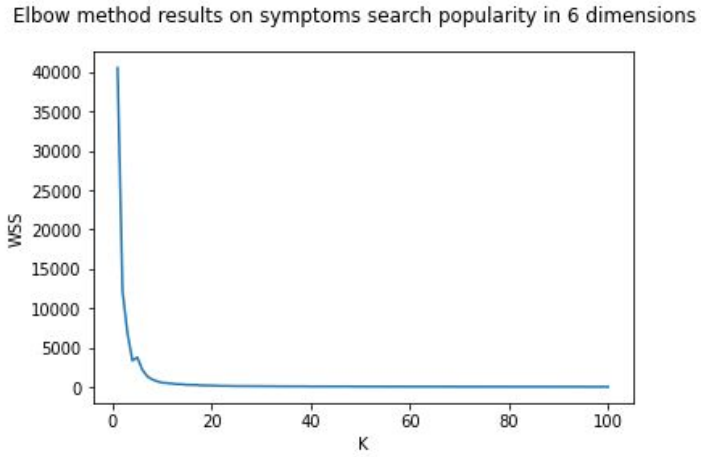



*Figure 3: Elbow method results on symptoms search trends reduced to 6 dimensions.*

Next, the consistency of the clusters from the original dimensions to lower dimensions was calculated (see methodology for calculating the consistency on "optimal_num_clusters.ipynb" file). The following table summarizes the optimal K (number of clusters) as well as the similarity of those clusters with respect to their original dimensions:

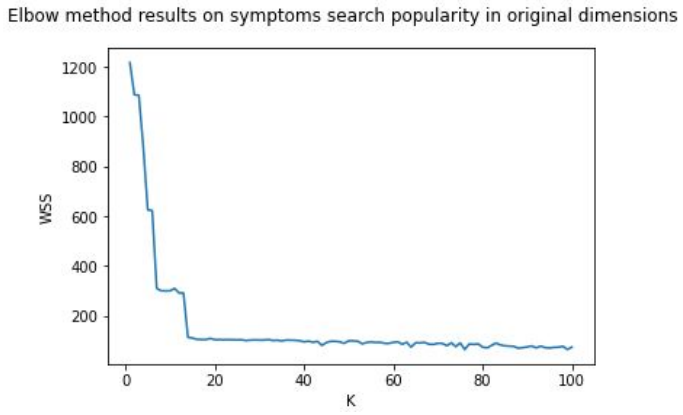
| Dimensions reduced to | Optimal K | Consistency of cluster (%) for K = 15 |
|---|---|---|
| Original (no reduction) | ≈ 15 | 100 |

| | | |
|---|---|---|
| 2D | ≈ 5 | ≈ 75 |
| 3D | ≈ 5 | ≈ 73 |
| 6D | ≈ 5 | ≈ 69 |

*Figure 4: Table presenting the optimal number of clusters and consistency across clusters from high dimensions to lower dimensions.*

It should be noted that even though the optimal K for 2,3 and 4 dimensions was 5, the consistency of the clusters was calculated using K = 15, being the optimal K for our raw data. This choice was made so that we get a good idea of how consistent the clustering would be if we reduced our data to lower dimensions to be able to visualize them. The idea being to determine whether it is possible to faithfully visualize the clusters present in our raw data in 3D or 2D.

In order to properly visualize the reduced data, we reduced the high dimensional data to three dimensions, compensating for 69% of variance in the data. Once reduced we're given the graph seen in figure 2.2. Once we have obtained the reduced data we use kmeans to visualize these clusters.



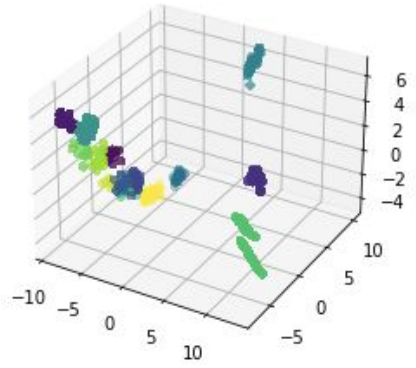




*Figure 5.1: An elbow graph comparing accuracy to the number of clusters k. By utilizing this graph we can determine that the optimal clusters k for the raw data is approximately 15.*
*Figure 5.2: The same three dimensional graph seen in figure 2.2, however using 15 clusters to portray the greatest accuracy to the original dimensions, as determined by figure 4. Kmeans has been utilized to determine optimal cluster centers.*

K-NN and Decision Trees were used as models to predict new hospitalization cases. These models were validated using two schemes. The first performs cross validation by splitting on regions so that 20% of the regions are used in the validation step in each fold. This lets the models predict hospitalization for new regions.The second splits the dataset into two parts, separated by a date, and the validation set is the second half. This allows the models to predict recent hospitalization.

K-NN is run 100 times where K is increased incrementally. The Decision Tree is also run 100 times and its hyperparameter is the minimum number of samples per leaf node. The best hyperparameter and its Mean Squared Error were recorded.

The following table summarizes the results of K-NN and Decision Tree performance.

| | K-NN | Decision Tree |
|---|---|---|

| | | |
|---|---|---|
| Region split | Best K=100, MSE=2306.6 | Best min_samples_leaf=28, MSE=2258.6 |
| Time split | best K = 1, MSE = 972.1 | Best min_samples_leaf = 3, MSE = 1222.3 |

*Table 6. Validation Results*

PCA was used on the dataset to see how effective the models would be if they used the data with reduced dimensions. Using 5-fold cross validation, the best min_samples_leaf was 153 with MSE 2177.7 and the best K for K-NN was 190 with a MSE of 2242.0.

Using PCA to keep two dimensions yields the following scatter plot, where the z-axis (up) are the hospitalization labels

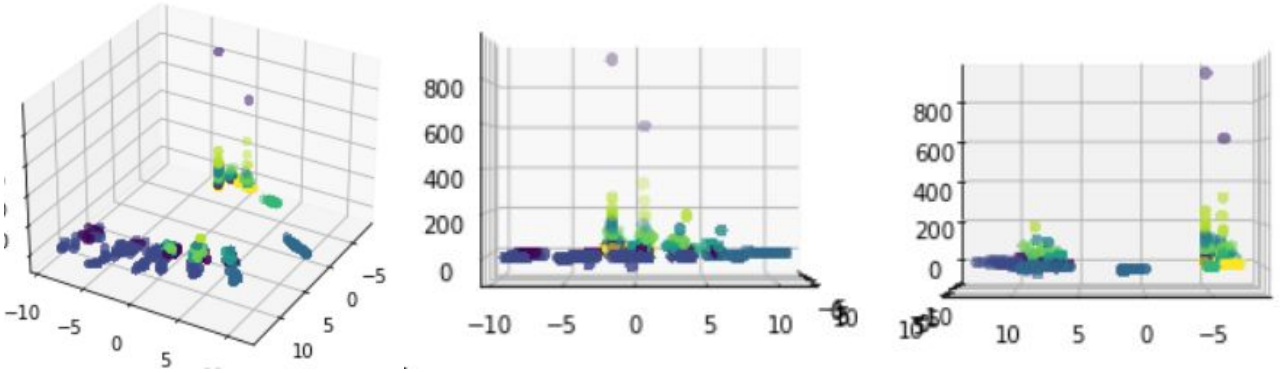



*Figure 7. PCA reduced (D=2) dataset with labels added on the z-axis*

# Discussion and Conclusion

The increase in popularity of various symptoms during the period of late February to March suggests a possible link to the start of the covid-19 epidemic in the US. This is reinforced by the correlation between symptoms search trends increase and the number of new hospitalizations due to covid.

The good consistency of the clusters from their original dimensions to 3D and 2D allowed a faithful visualization of our data clustering in lower dimensions (see clustering graph in 3 dimensions, *Figure 5.2*). Therefore any assumptions involving clustering of our data in lower dimensions would fairly fit for the raw data as well.

By analyzing the graph in figure 5.2, you can see clear clusters, determined by utilizing kmeans. Without more data it is difficult to determine what these clusters represent, but we can make a few hypotheses. We could make a hypothesis that the clusters are points within a similar time range, where search trends for the same symptoms have similar results across regions. Another possibility is that certain search results are much more concentrated in particular regions, and the clusters represent a certain collection of symptoms given a handful of regions.

The performance of the K-NN and Decision Tree was twice as good when predicting recent hospitalization trends when trained on older data as opposed to making predictions on new regions with other region data. But the performance of both models remained about the same for each validation scheme. Training both models on PCA-reduced data yielded similar MSEs. Both the models' MSEs are quite high, so these models are not great. Figure 07 gives a picture of why this is the case. There is a lot of clustering on features, but for each cluster the variance in the labels is very high.

# Statement of Contribution

Julian: Wrote code for data import, merge, normalization, and some other utility functions to filter the data. Wrote the jupyter notebook `models.ipynb` containing code for K-NN and the Regression Tree, their error performance, the error of training on PCA-reduced data, and a visualization of the data with labels in low dimensions. Wrote sections in the lab report pertaining to the above statements.

Tim: Wrote the jupyter notebooks `pca_dimensions.ipynb`, and `kmeans_graphics.ipynb`. Also wrote their respective `pca.py` and `kmeans.py` classes. Wrote sections in the report regarding the related work.

Saksham: Wrote optimal_num_clusters.ipynb containing the code to find the optimal number of clusters and consistency of clusters from high dims to low dims. Wrote symptoms_pop.ipynb containing the code to represent the evolution in popularity of various symptoms. Wrote some methods used in the data pre-processing.