Analyzing COVID-19 Search Trends and Hospitalizations

Jay Abi-Saad - 260801368
Julien Phillips - 260804197
Rayan Osseiran - 260803963

October 21, 2020
McGill University
COMP 551

Abstract. The purpose of this project is to better understand the relation between symptom search trends and COVID-19 hospitalizations across several US states. Google's symptom search trend dataset and their hospitalizations dataset were used as the principal datasets for this project. We used heatmaps to visualize high-dimensional search trend data over different regions, which revealed that search trends vary by region and that there are three main symptom search trends. We reduced the dimensionality of our data using PCA. We found that our lower-dimensional data accurately represented our original data. Moreover, we used scatter plots to visualize the low-dimensional data, which revealed that data points corresponding to the same region tend to end up in the same cluster. Our supervised learning analysis used k-nearest neighbours (KNN) and decision tree models to predict COVID-19 hospitalization cases from the searched symptoms. Whilst both models performed similarly, KNN achieved either equal or noticeably better performance depending on the prediction strategy. It was also observed that naive split strategies for the data resulted in less accurate predictions with higher error.

I. INTRODUCTION

This project required the use of two datasets. One dataset contains the trends in search patterns across various US states over time while the other contains the new and cumulative hospitalizations over time from various regions around the world. Our first task was to combine these datasets, which had different time resolutions, and additionally clean the data in both datasets in order to remove regions and features with invalid or too many missing data entries.

Our second task was to utilize dimensionality reduction techniques and visualize the data in different ways for both high and low dimensions. For high dimensional data, we picked the 4 symptom searches that had the most non-zero data across regions and time and visualized it using heatmaps. We found that symptom search trends vary between regions, and three main trends appear; a symptom search trend with a single peak of popularity during the first wave of COVID-19, a symptom search trend of constant popularity over time, and a symptom search trend with a first peak of popularity during the first wave of COVID-19, and a second peak of popularity during the second wave of the virus.

We reduced the dimensionality of our data by using principal component analysis. To decide how many principal components were needed to accurately describe our data, we used the knee and cumulative variance methods, both of which seemed to indicate that 3 principal components were needed to accurately model our data in lower-dimensional space. We visualized the low dimensional data by plotting a scatter plot of 2 of the top 3 PCs that we got from PCA. We found that the data points corresponding to the same region tend to cluster together in the plot.

We evaluated possible groups in the search trends dataset by using the k-means clustering method on both high dimensional and low dimensional data. To find the optimal number of clusters (hyperparameter K), we plotted the sum of squared errors (SSE) over 10 different cluster values ranging from 1 to 10. We then used the elbow method to find that both the low and high dimensional data would have an optimal clustering with $K = 3$ clusters. Lastly, we plotted both high-dimensional and low-dimensional k-means with 3 clusters each. We found that clusters seem to remain approximately consistent between the low and high dimensional data, which indicates that our top 2 principal components give us an accurate representation of our high-dimensional data.

The next step was to perform supervised learning in order to predict hospitalizations based on search trends data. This was done through KNN as well as decision trees and there were four separate strategies in terms of splitting data into train and validation sets. This included:

- Using data after 2020/08/10 as validation data and training on the rest.
- Splitting 80% of regions into a training set and 20% into a validation set.
- Treating each region as a separate model.
- Treating each season as a separate model.

Throughout all of the above experiments, KNN consistently produced either an equal or superior mean squared error to decision trees. The minimum error was selected by testing for all possible values of K in KNN as well as varying the max-depth of the tree in decision trees, however, this does not necessarily reflect the correct parameter selection for testing. Results were also visualized and are further detailed in the following section.

II. DATASETS

1. Data Processing

In order to be able to conduct exploratory analysis of the data and supervised learning, both datasets had to be cleaned and merged together. We began by removing all regions from the hospitalization dataset that were not present in the symptom search trends dataset. All data not pertaining to the number of hospitalizations was dropped from the hospitalizations dataset. In addition, we grouped the hospitalizations dataset into weeks from days, since keeping a daily time resolution would make it incompatible with the symptom search trends dataset, which has a weekly resolution. The search trend data was then cleaned so that any symptom with entirely missing data was removed from the dataset. Both datasets could then be merged together successfully. Other regions and symptoms were dropped in subsequent tasks according to the needs of the tasks. Following our initial testing of the data, we opted to transform new hospitalizations into new hospitalizations per hundred thousand individuals. In order to do this, a small dataset obtained from World Population Review (1) of US states and their respective populations was used. Adjusting new hospitalizations based on the state population allowed us to take the population of different US states into account in the supervised learning process.

2. Understanding the Data

To visualize how the distribution of search frequency of symptoms aggregated across different regions changes over time, we plotted a heat map for 4 of the 121 most popular symptoms. We defined the most popular symptoms to be the symptoms which have the least 0 popularity values in their respective column. Each heat map shows the weekly evolution of the popularity of the given symptom for each of the 16 regions. A darker patch indicates higher popularity of the given symptom at a given week, whereas a lighter patch indicates the opposite.

We used PCA to visualize search trends over lower-dimensional space. To determine the number of principal components to use for our analysis, we used the knee method as well as the cumulative variance method. To visualize the search trends over lower-dimensional space, we plotted a 2D scatter plot of the first two PCs, that is, the two PCs which best describe the data.

We used the k-means clustering method to evaluate possible groups in search trends between high dimension and low dimension data. To find the optimal number of clusters for each, we plotted the sum of squared errors for various cluster values K between 1 and 10 (hyperparameter) and used the elbow method to find the optimal value of K.

3. Additional Processing for Learning

In order to refine the data for learning, additional data processing was done. First, it was observed that all hospitalizations before March 16 were 0. The first step was to drop all data below that date threshold. Once this was done, there were still certain regions that remained after the March 16 point with virtually no hospitalizations. This was also dropped. For the purpose of the experiment, there is not much value in

performing learning on this portion of the dataset. We also explored dropping highly correlated features, but observed that this did not make a notable impact.

III. RESULTS

1. Data Visualization and Clustering

A) High-Dimensional Data Visualization

The 4 top symptoms are: aphonia, crackles, dysautonomia, and ventricular fibrillation. As seen in figures A.1 to A.4, we plotted a heatmap for each. A first observation that can be made is that weekly search trends seem to vary from one state to another for a given symptom. Moreover, for any given symptom or region, we notice 3 types of search trends over time. The first is when the popularity of a given symptom in a given region remains approximately constant from one week to another. An example of such a trend can be found for the ventricular fibrillation symptom in the US-WV region. The second is when the popularity of a given symptom has a single peak during the first wave of COVID-19 in January-February-March, but then continually loses popularity. An example of such a trend can be found for the aphonia symptom in the US-NE region. The third is when the popularity of a given symptom in a given region peaks during the first wave of COVID-19, then decreases over April, May and June, and then peaks again in July, August and September when the second wave of COVID-19 starts. An example of this trend is found for the dysautonomia symptom in the US-NH region.

B) Principal Component Analysis for Dimensionality Reduction

As seen in figure A.5, both the knee method and the cumulative variance method seem to indicate that the optimal number of principal components to use for our analysis is around 3. Therefore, we performed PCA with 3 PCs. The scatter plot seen in figure A.6 of the top two PCs seems to indicate that the data points representing each of the 16 regions are approximately clustering together.

C) Low-Dimensional Data Visualization

The plot of the sum of squared errors over the number of cluster K for high-dimensional k-means seen in figure A.7 seems to indicate that the elbow value is around K = 3. For the low-dimensional plot in figure A.8, it seems to indicate that the elbow value is also around K = 3.

After plotting the following scatter plot for both low-dimensional in figure A.9 and high-dimensional in figure A.10 with 3 clusters for each, we notice that the clusters remain approximately consistent for raw as well as PCA reduced data. This is expected since it indicates that our low-dimensional data gives an accurate approximation of our raw data.

2. Supervised Learning

A) Time Based Split Strategy

Splitting by time resulted in similar best case performance for both KNN and decision trees. The overall results are summarized in table 1 and the distribution of mean squared errors for differing values of k, and max depth are illustrated in figures A.11 and A.12 respectively.

Overall, it seems that a naive time split may result in inconsistencies and biases with test data. Whilst an RMSE of ~5 may seem low, taking a look at the normalized hospitalizations count would indicate that these are large error margins. There are a couple of possible reasons for this, namely that even after normalization, it may be difficult to predict hospitalizations across regions as the search trends may also vary greatly per region. Moreover, there may be dominant patterns across a specific time period that are not necessarily captured in our training data.

Model	KNN (k = 3)	Decision Tree (max depth = 1)
Mean Squared Error	28.6	28.7
Root Mean Squared Error	5.3	5.4

Table 1: Minimum error achieved in KNN and decision tree regression for time based split.

B) Region Based Split Strategy

Splitting data by regions resulted in similar performance to the time split strategy. The results are documented in table 2. 5-fold cross validation was performed and the distribution of errors for all values of k in KNN is illustrated in figure A.13. In our case, the minimal error was achieved with k = 3 neighbours, however, in practice, this may lead to overfitting, and can lead to poor predictions if there is noise in the dataset. In practice, we would try to select a larger k within one standard deviation of our error, however in this case, the error is relatively consistent and as a result, all values of K remain in 1 standard deviation of our minimum error. The same is true for the max depth where the error is lowest for depth = 1 which can result in large errors with a larger dataset to test. The distribution of errors for differing values of max depth is illustrated in figure A.14. In this case, it would likely be better to select a larger max depth such as 6 whose error remains within 1 standard deviation of our minimum.

Overall, this can once again point to how it may be difficult to compare across regions. Even with normalization, these errors are not negligible and other unknown factors may be impacting search trends in a region.

Model	KNN (k = 3)	Decision Tree (max depth = 1)
Mean Squared Error	32.5	37.4
Root Mean Squared Error	5.3	5.5

Table 2: Minimum error achieved in KNN and decision tree regression for region based split.

C) Individual Region Strategy

Treating each region as an individual model allowed us to better interpret the variations between each region. Moreover, variations in certain errors point to how regions cannot necessarily be compared even after normalization. Table A.1 is a summary of the mean squared errors for each region in both KNN and decision trees. Notice that for certain regions, such as US-MT, we have an effectively negligible error.

This data was processed similar to the time split where the number of neighbours and decision tree depth that produced the lowest error were selected and reported. Once again, the reported number of neighbors here may result in overfitting (or underfitting for larger Ks) to the training data, and may not perform well with test data. With that being said, this experiment confirmed our hunch that comparisons across regions are not necessarily straightforward and that a different strategy may be required.

D) Individual Season Strategy

Treating seasons individually produced a very low mean squared error across the board. This was particularly the case for KNN which once again proved to produce a lower error than decision trees. The results are illustrated in table 3. The primary reasoning here is that we can try to observe seasonal trends or in other words see how accurate of a prediction we can make in each phase of the pandemic. This proved to be the best method that allowed comparisons across regions.

Model	Minimal K	KNN (MSE)	Minimal Depth	Decision Tree (MSE)
Spring	3	0.6	11	4.8

Summer	5	2.2	5	5.3
Fall	5	2.3	8	6.6

Table 3: Minimum error achieved in KNN and decision tree regression for individual season based split.

IV. DISCUSSION & CONCLUSION

One issue pertaining to the symptom search trend dataset is that the scaling of normalized popularity values for each region makes it meaningless to compare across regions. Ideally, providing the scalar values used to scale each region's data would allow us to revert to normalized popularity values. Since these scalars are not made available, some other technique to make search trend data comparable across regions would likely be necessary. Although it is not immediately clear how this can be achieved, it is worth noting that it may be possible to improve our learning error by normalizing or standardizing the data by region or symptom. Performing these operations on the symptom search trend data may be worth investigating in the future.

Heatmaps are a great way to visualize the evolution of the popularity of symptoms for different regions over time. They helped us to visualize the three symptom search trend types; a symptom search trend with a single peak of popularity during the first wave of COVID-19, a symptom search trend of constant popularity over time, and a symptom search trend with a first peak of popularity during the first wave of COVID-19, and a second peak of popularity during the second wave of the virus. With the polarized opinions about COVID-19 in the US, it would be interesting to compare heatmaps of the search trends between Democratic and Republican states.

Moreover, using scatter plots to visualize low dimensional data over different regions showed us that data points coming from the same region seem to be in the same clusters. It would be interesting to visualize the low dimensional data with different techniques that could give us more insight into all three principal components.

Lastly, the k-means cluster technique showed us that clusters remain approximately consistent between the low and high dimensional data. This indicates that our top 2 principal components give us an accurate representation of our high-dimensional data. Future experiments could try to visualize the lower-dimensional data with different more robust techniques.

As for learning, KNN performed similarly or better than decision trees in all cases. Furthermore, the results pointed to the importance of the dataset, as the model is only as good as its data. Based on our split strategies, it was observed that naive splits that simply compared across regions were often not successful in making predictions with a low error rate. The lowest errors were observed only when looking at each region individually, or when trying to compare data that was collected during a similar phase of the pandemic. As a result, it would be interesting to explore other strategies that also attempt to take advantage of patterns such as the seasonal one. This, along with the additional data standardization mentioned earlier, would enable us to make better predictions when comparing across regions.

V. STATEMENT OF CONTRIBUTIONS

Julien: Responsible for task 1 and contributed to task 2 and 3. Contributed to the project write up.

Jay: Responsible for task 2 and contributed to task 1 and 3. Contributed to the project write up.

Rayan: Responsible for task 3 and contributed to task 1 and 2. Contributed to the project write up.

VI. REFERENCES

1. US state population dataset from https://worldpopulationreview.com/states

VII. APPENDIX

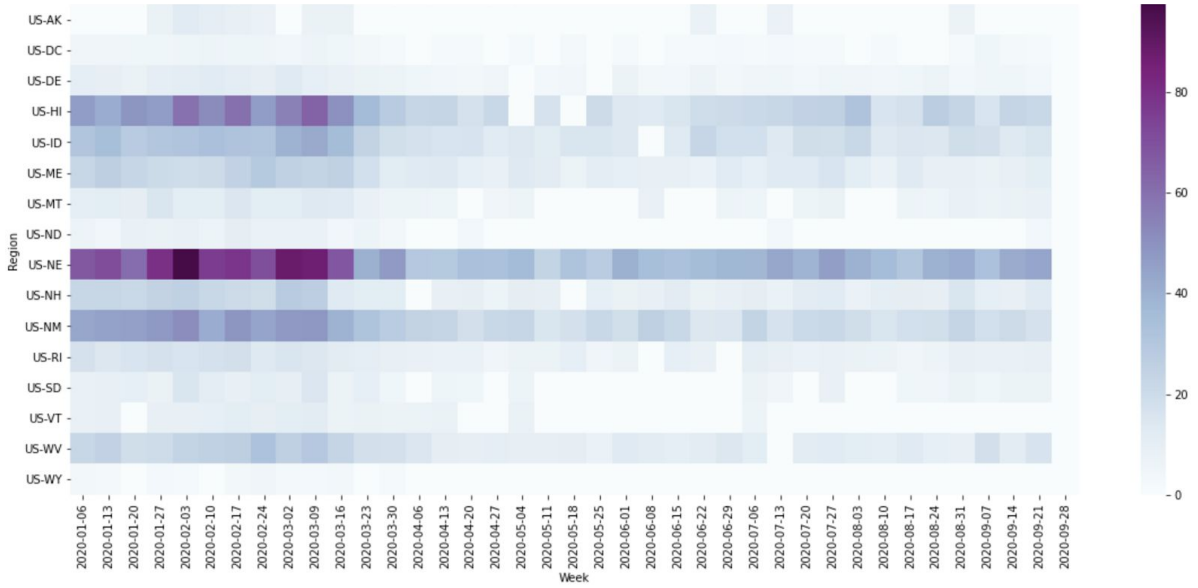

Figure A.1: Weekly evolution of popularity of the Aphonia search for US regions from 2020-01-06 to 2020-09-28.

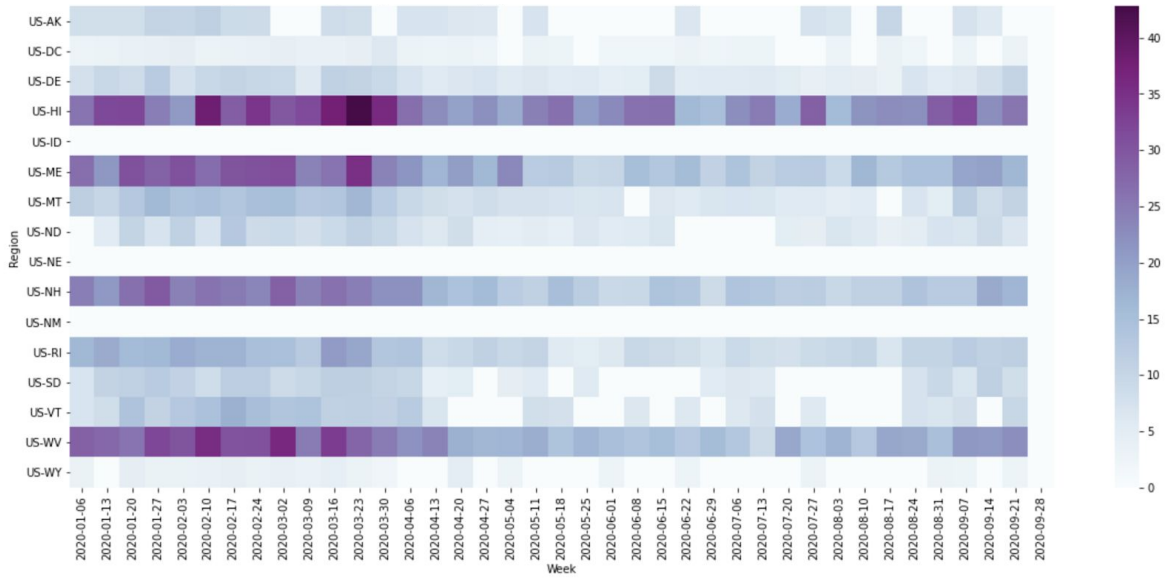

Figure A.2: Weekly evolution of popularity of the Crackles search for US regions from 2020-01-06 to 2020-09-28.

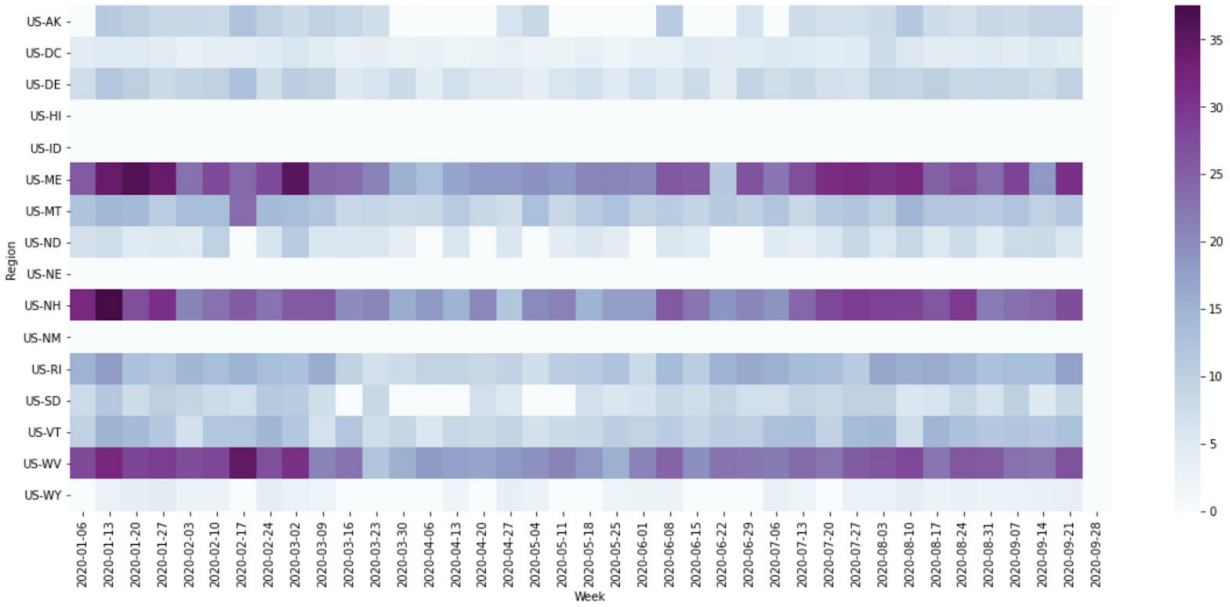

Figure A.3: Weekly evolution of popularity of the Dysautonomia search for different US from 2020-01-06 to 2020-09-28.

Figure A.4: Weekly evolution of popularity of the Ventricular fibrillation search for US regions from 2020-01-06 to 2020-09-28.

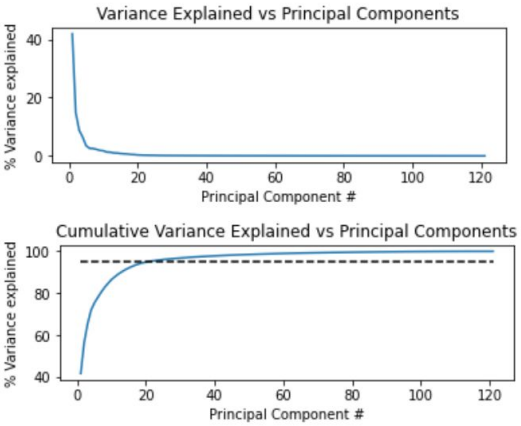

Figure A.5: Cumulative variance and knee method plots.

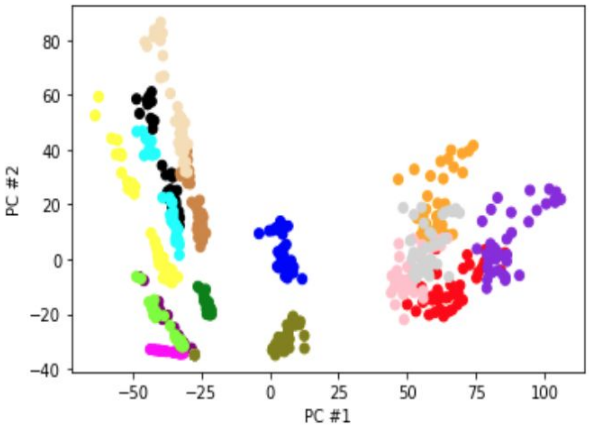

Figure A.6. Scatter plot of first two PCs colored by region.

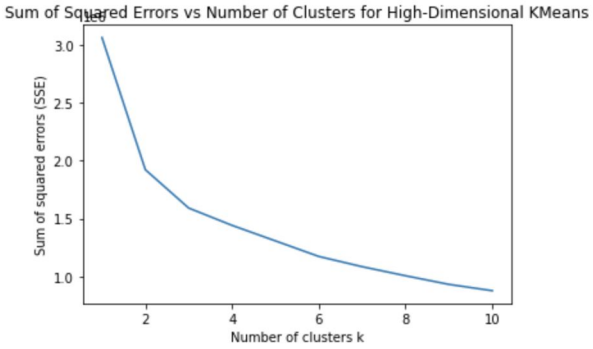

Figure A.7. Elbow method to estimate number of clusters K for high-dimensional data.

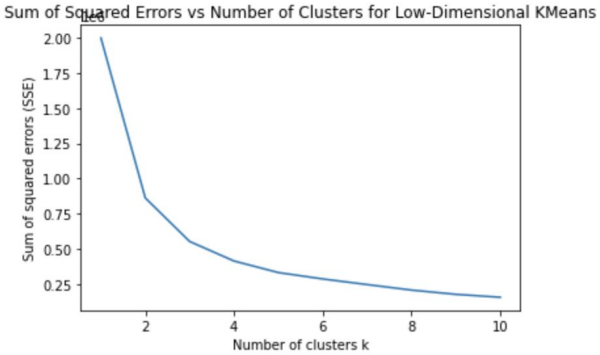

Figure A.8: Elbow method to estimate number of clusters K for low-dimensional data .

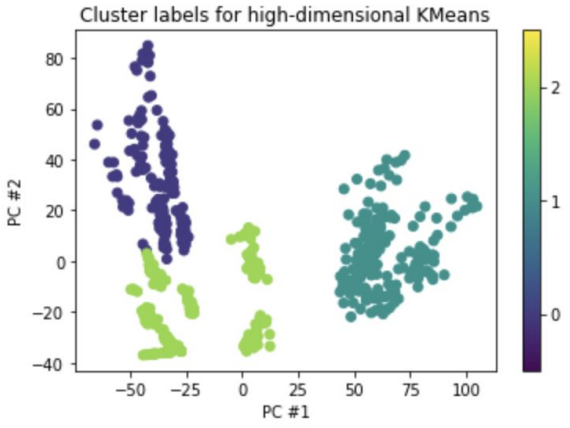

Figure A.9: KMeans cluster for high-dimensional data using K=3 clusters.

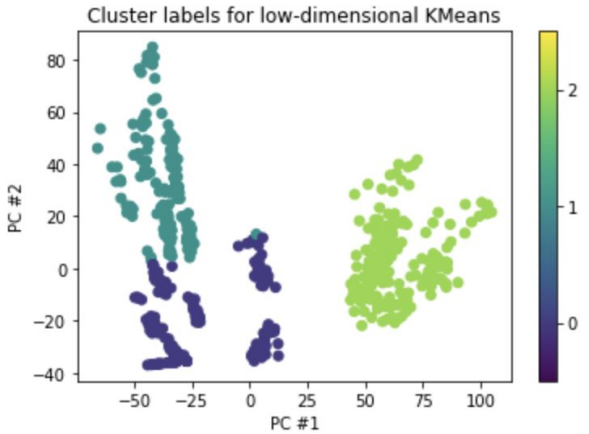

Figure A.10: KMeans cluster for low-dimensional data using K=3 clusters.

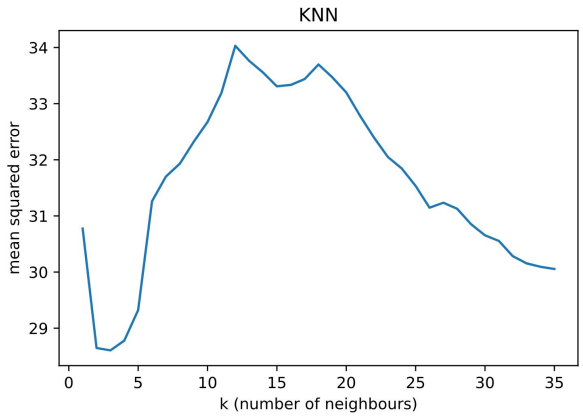

Figure A.11: KNN Regression performance for all possible values of K.

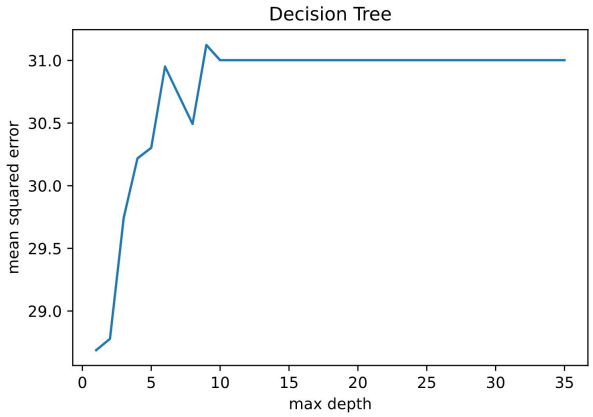

Figure A.12. Decision tree Regression performance for all possible values of max depth.

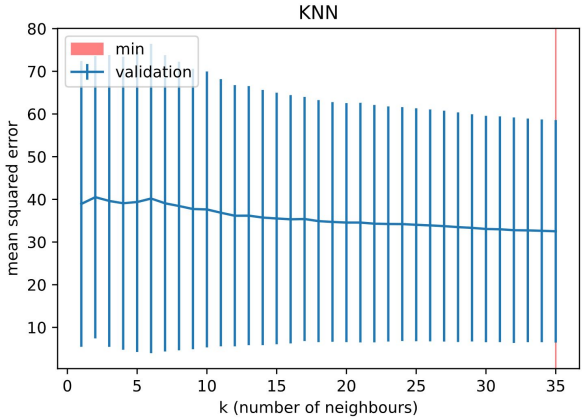

Figure A.13: KNN cross validation performance for all possible values of K.

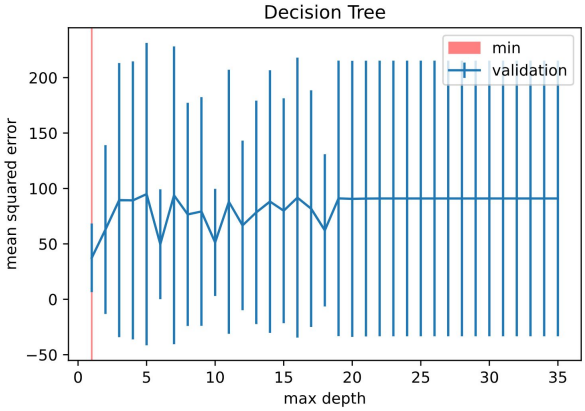

Figure A.14: Decision tree cross validation performance for all possible values of max depth.

Region	Minimal K	KNN (MSE)	Minimal Depth	Decision Tree (MSE)
US-HI	1	4.9	1	4.9
US-ID	8	4.5	2	8.2
US-ME	2	0.2	1	0.2
US-MT	1	1.2	1	1.2
US-ND	2	0.7	4	1.1
US-NE	2	3.9	2	3.9
US-NH	5	0.6	2	1.2
US-NM	22	8.8	2	28.8
US-RI	8	193.5	4	221.6
US-SD	1	0.9	1	0.9
US-WV	22	15.7	22	15.7
US-WY	20	1.0	1	2.0

Table A.1: Minimum error achieved in KNN and decision tree regression for individual region based split.