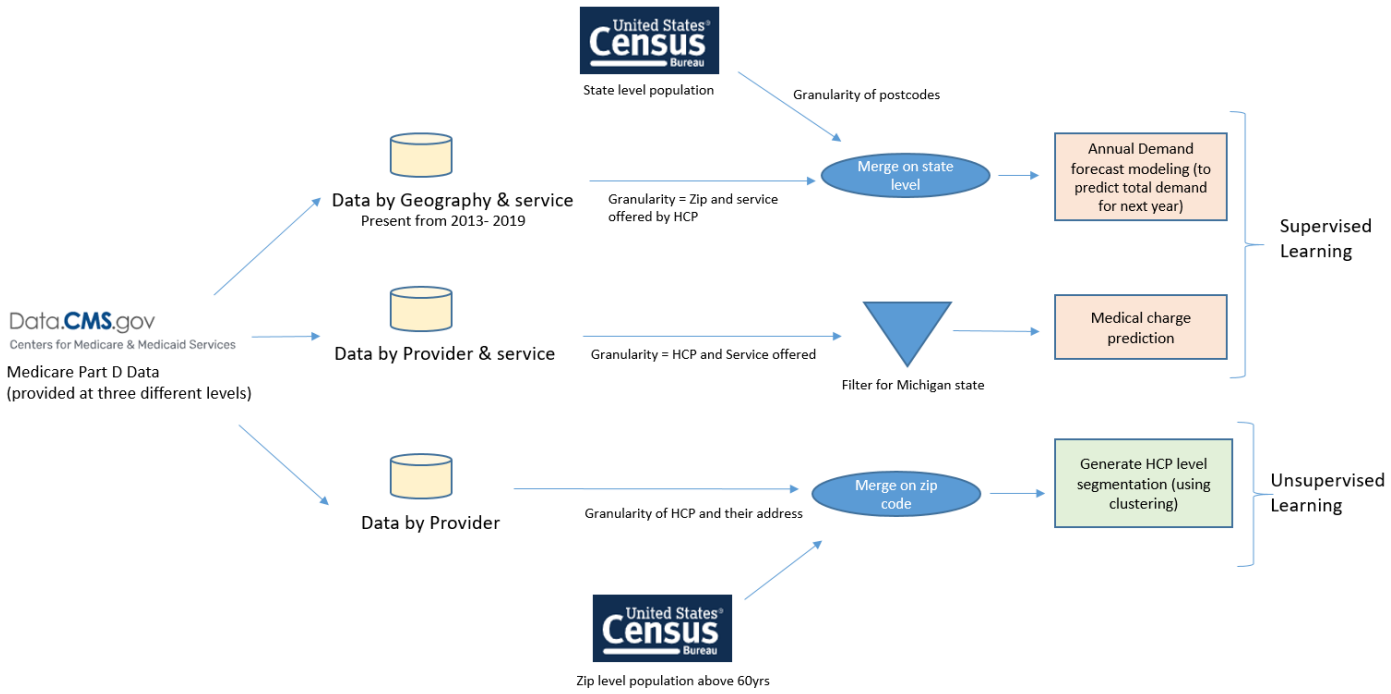
Medical Service Demand/ Cost Prediction and Provider Segmentation

Team members: Himank Kansal, Louis Pienaar, Cindy Xie

Overview

The Centers for Medicare and Medicaid Services (CMS) is a U.S. government agency that provides health coverage to more than 100 million people through Medicare, Medicaid, and other Health Insurance Programs. The Medical Physicians & Other Practitioners dataset is based on information gathered from CMS administrative claims data for Original Medicare Part B beneficiaries available from the CMS Chronic Conditions Data warehouse dataset. It contains 3 different levels of aggregated data for all claims submitted to Medicare by all medical service providers for the years 2013 to 2019.

This project utilizes the datasets to analyze medical physicians' medical practice (charges, service, etc) and gather some insights regarding medical cost, demand, and segmentation. The result can be beneficial to different participants in the healthcare field such as patients, physicians, pharmacy companies, health insurance companies, government agencies like CMS, etc. An overview of the datasets and different parts of the project illustrates below:



Part A-1. Supervised Learning - Cost Prediction Model

Motivation

Medical expense is rising, our motivation is based on the fact that different providers charge different amounts for the same service, so the medical expense is not transparent and hard to plan. With this CMS dataset, we assume that one provider's charge for the same service is not based on the type of insurance patients have, so we can use the charges a provider submitted to CMS on all their potential future patients. Through this model, we'll be able to help people to have a general idea of medical service and cost, especially helpful for people with long-term treatment, to plan their expenses for the coming years.

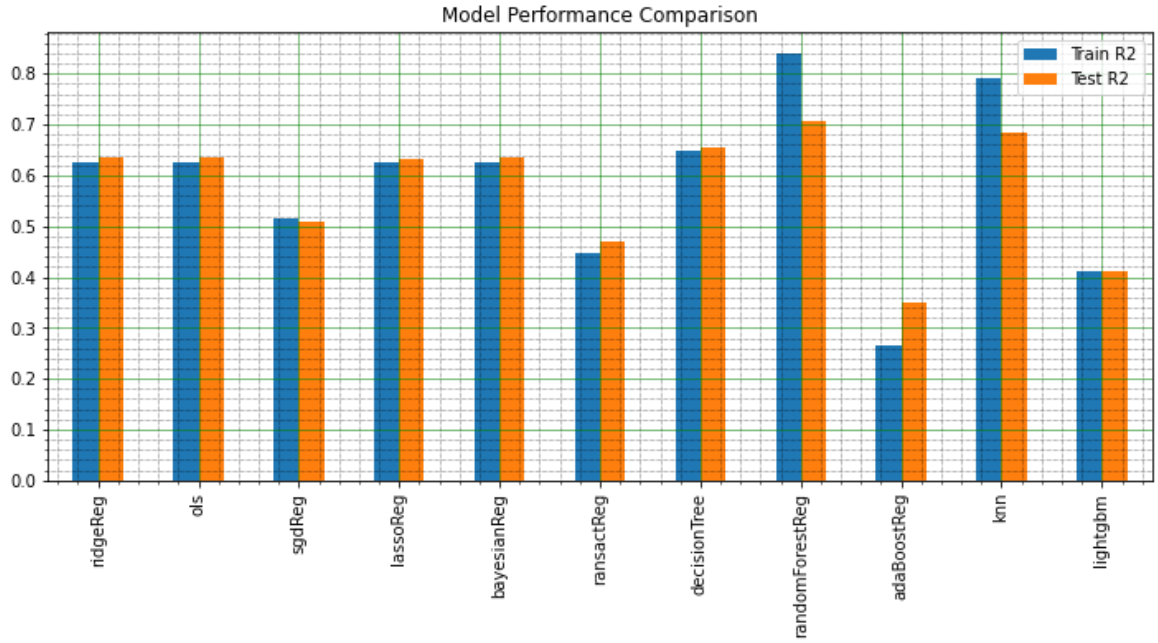
Data Source

File Name / Time Period	Medicare Physician & Other Practitioners - by Provider and Service, the year of 2019: MUP_PHY_R21_P04_V10_D19_Prov_Svc.csv
File Format / Size/Records	CSV file **3GB** with **29 columns** and total records **10,140,228**. Due to the size, filtered to Michigan provider only and reduce the records to **337,351**
Purpose / Variables	Most detail level data by provider and service. Columns include provider info (NPI, name, gender, address, credentials, type, if accept medicare, total patients, total services, etc); treatment info(HCPCS code and description) and cost info including provider submitted charge amount, medicare allowed amount, payment amount and standardized amount, etc.

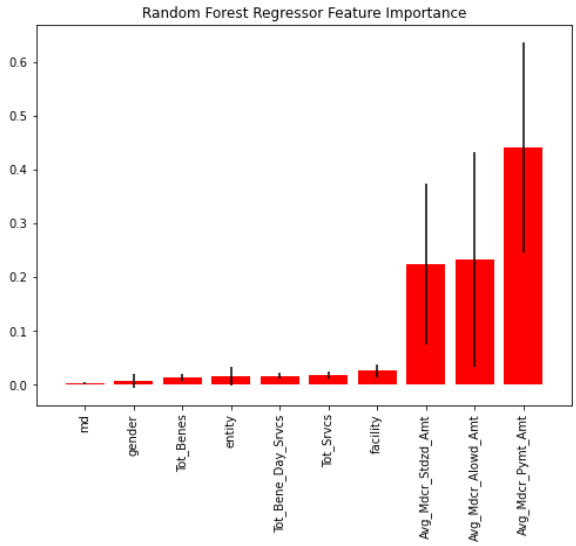
Methods & Evaluation

This model intends to provide transparency and fairness of medical charges for people looking for medical service. There are a rich set of features provided in the dataset including physician information(NPI, credential, gender, specialty, etc), geography information(state, zip code, etc), and service information(HCPCS code, drug indicator, place of service, etc). After initial data analysis, we selected some features to build a medical charge prediction model. We explored and try several different models including decision trees, different linear regression models, and lightggbm. Then we selected the top 3 best performance models and did further tuning of the model through parameter tuning and grid search. We also explored vectorized HCPCS code and then use dimension reduction through PCA to reduce the feature to a reasonable amount due to there are over 2300 unique HCPCS codes.

For evaluation, we use the R^2 ("r-squared") Regression Score along with well-known accuracy measures like RMSE(root mean squared error) and MAE(mean absolute error). Following is the comparison chart for different models



The final winning model is a Rain Forest Regressor with most important features including medicare allowed amount and if the service is performed in a facility or not as shown below:



Failure analysis

We analyze one of the biggest failed predictions is for HCPCS code 22514. The charge amount submitted is $52,000, but medicare only allows $6480. A further investigation found for this same HCPCS, there are over 20 claims, and the submitted charge range from $950 to $52,000. Such a big range makes medical expenses unpredictable. This is one of the reasons we try to build this model. It also tells us we may consider removing outliers for future improvements.

Part A-2. Supervised Learning - Demand Forecasting Model

Motivation

The Medicare Physician & Other Pthe ractitioners - by Geography and Service Data, contains the number of services provided for each specific medical service furnished by the provider, aggregated by geographic coding. The ability to accurately forecast the number of services to be provided will aid supply-side decision-makers. The number of medical services must be partly determined by the population size and population characteristics. The goal of this model is to utilize the historic values recorded as the population estimates from census data to build a demand forecast model to predict the number of services to be provided.

Data Source

File Name / Time Period	Medicare Physician & Other Practitioners - by Geography and Service, year 2013 to 2019: Medicare_Physician_Other_Practitioners_by_Geography_and_Service_****.csv
File Format / Size/Records	2019 **CSV file** is **41MB** with **15 columns** and total **records 273,211**. Other years are similar.
Purpose / Variables	demographic data on the providers and numerical data on the number and amount of claims per medical service. geography level data, columns: Rndrng_Prvdr_Geo_Lvl Rndrng_Prvdr_Geo_Cd Rndrng_Prvdr_Geo_Desc HCPCS_Cd HCPCS_Desc HCPCS_Drug_Ind Place_Of_Srvc Tot_Rndrng_Prvdrs Tot_Benes Tot_Srvcs Tot_Bene_Day_Srvcs Avg_Sbmtd_Chrg Avg_Mdcr_Alowd_Amt Avg_Mdcr_Pymt_Amt Avg_Mdcr_Stdzd_Amt

The CMS dataset is freely available for download from the CMS website. It contains records for the period 2013 to 2019, detailing a number of measures aggregated by geographic region and by a service identifier (HCPCS code). Each year is presented in a comma-separated value file (CSV) and contains a combined number of 1 760 035 records. In this dataset, there are only a few features of interest; The state (Rndrng_Prvdr_Geo_Desc), the service code (HCPCS_Cd) and the number of services offered (Tot_Srvcs).

The second dataset contains the population estimates for each US state, obtained from the census website. It is a set of CSV files, containing the population estimates for each state for each age group and race group. It contains 197 370 lines of data.

Methods & Evaluation

An exploratory data analysis highlighted the possible features that could be used to predict the number of services. The two datasets were joined using the state as the key field. Thereafter the dataset was melted to a wide format. Each row represents the features for the state and HCPCS grouping. Therefore each row contains the historic number of services for the years prior to the year being predicted, as well as the population estimates for those years and the population estimate for the year is predicted. The target value being predicted is therefore the number of services provided per state per HCPCS code.

An important transformation was to convert the features and the targets by using the log of the values. The training set used the years 2013 up to and including 2016 as the feature set and the target for prediction 2017. The test set used the data from 2014 up to and including 2017, and the target value for prediction 2018.

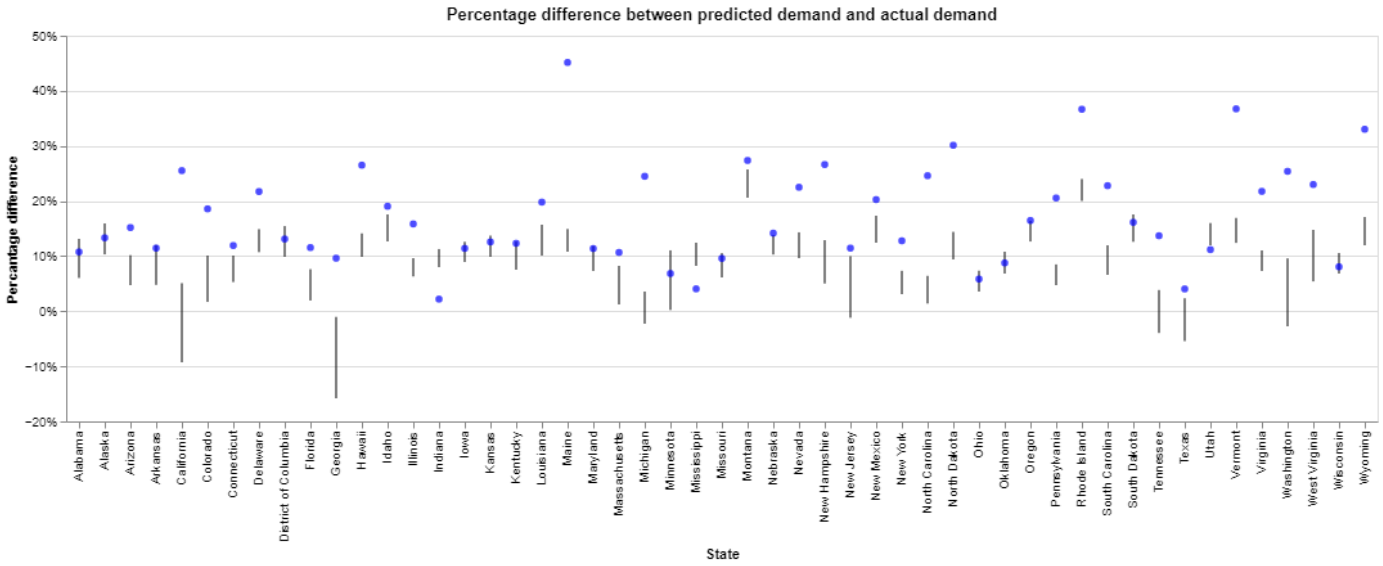
To reduce the complexity of the model, the population estimates were grouped for all races and into only two age groups (above and below 40).

The chosen modeling method was to use regression rather than time series analysis. The modeling accuracy were measured using mean squared error as the main performance indicator. The final two models that were considere was a XGBoost regressor and a lightGBM regressor. The XGboost regressor by far out performed the other models and the lightGBM model.

We used the hyperopt package to aid with finding the optimal set of parameters.

The accuracy of the final model is respectable given that the only features being used is the historic demand and the population estimates for each year.

The final predictions on the test set performed with accuracies on average between 8% and 20%. The graph below shows the accuracies by illustrating the mean accuracies per state as the blue dots and the variance per state as an error bar for each state.



Failure analysis

The strategy of using population estimates (only) and historic values of services provided worked well for most scenarios, but for others, failed. Upon inspection, it was found that population estimates were not available for some of the states in the datasets. These states were dropped for both training and test sets. (They include states like: 'Armed Forces Pacific', 'Foreign Country').

There were entries that would appear to be subject to other external changes. For example, entries that had services offered in excess of 50 000 per year, consistently, and then in the prediction year, that would revert back to zero. These entries were identified and removed from the prediction sets.

From deeper investigations, it is found to be true that other factors that influence the demand can easily be identified and included in the feature set. In their absence, the model would continue to fail in those cases.

Part B. Unsupervised Learning

Motivation

CMS Provider data includes all the individual physicians as well as office-based accounts. These physicians/ accounts have different potential to prescribe any drug. To get the best ROI from the medical rep cost and marketing, Pharma companies usually target physicians with high potential. Just like any customer segmentation done in the industry for ex FMCG etc, segmentation for Pharma companies will be done on HCPs/ Hospitals. Physician-level details like patient potential, specialty, place of practice can be used to identify and build segments based on potential. This helps in utilizing funds effectively and increasing ROI on reps, sample drops, face-to-face calls with physicians, etc.

We created these medical provider segments (Super targets, Tier 1, Tier 2, Tier 3 as well as Non-targets) utilizing physician specialties, services offered by the provider, cost of service and Medicare claims data (a proxy for potential) from this CMS dataset, and few additional other demographics data. These segments play a key role in defining rep visits to these physicians.

Data Source

1. **Medicare Physician & Other Practitioners - by Provider:**

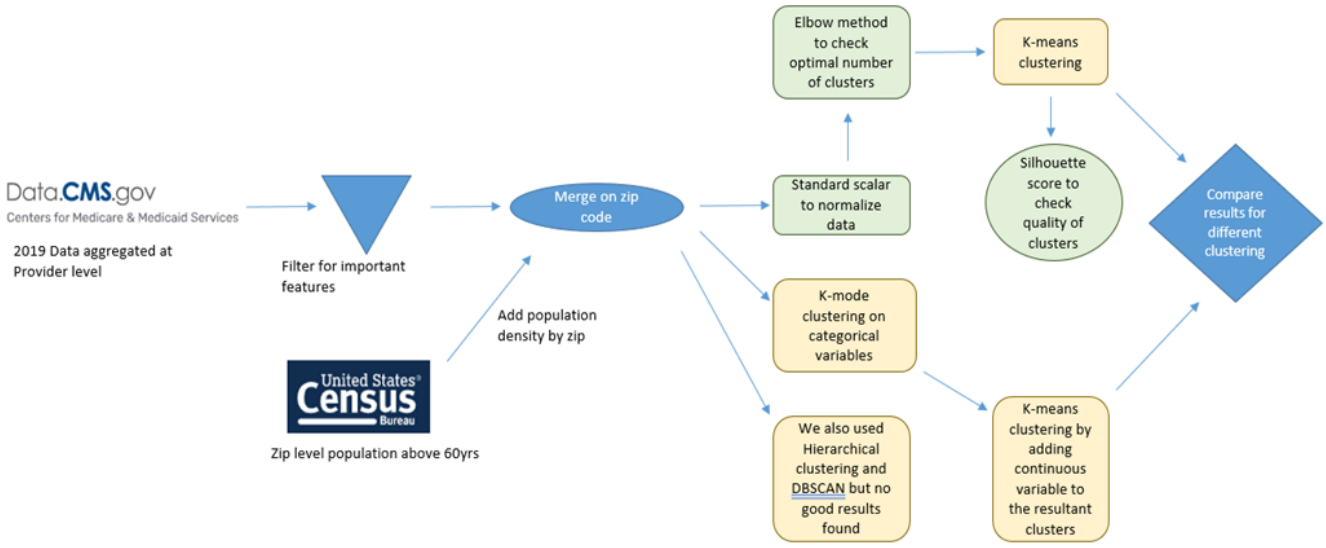
File Name / Time Period	Medicare Physician & Other Practitioners - by Provider: UP_PHY_R21_P04_V10_D19_Prov.csv Link to download File: Centers for Medicare & Medicaid Services Data (cms.gov)
File Format / Size/Records	Downloaded 2019 **CSV file 458 MB** with **73 columns** and total **records 1,155,870.** **Note on Time period:** Just like other aggregation level used in supervised learning, this data is also available from 2013 to 2019. We have used only the latest year i.e. 2019 for segmentation.
Purpose / Variables	This is a provider level data with features like their demographics information (name, gender, address, zip code) credentials, type if accept Medicare, total patients, total services, etc.; treatment info(HCPCS code and description) and cost info including provider submitted charge amount, Medicare allowed amount, payment amount and standardized amount, etc. ~15 Important variables representing above details are selected for clustering

2. US population by age and by zip code:

File Name / Time Period	US population by age and by zip code. This is US census data but available in required format from Kaggle. Hence, we have used directly from Kaggle **Location to download:** https://www.kaggle.com/census/us-population-by-zip-code/version/1?select=population_by_zip_2010.csv
File Format / Size/Records	Downloaded csv file for population of 2010 by age and by zip code. **Size** – 58 MB **Columns** = 6, **Records** = 1,048,576
Purpose / Variables	This file is used to estimate 60+ population for 2019 by using growth rates

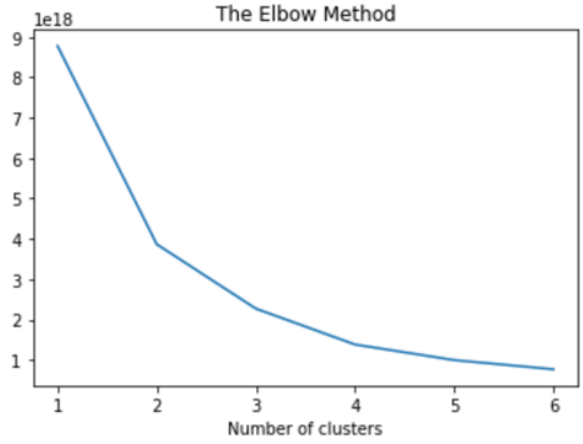
Unsupervised Learning Methods

Below is an outline of steps we used to arrive at final segmentations:



Tuning of hyper parameters:

1. **Arriving at an optimal number of clusters using Elbow method**: Calculate the Within-Cluster-Sum of Squared Errors (WSS) for different values of k. In the plot of WSS-versus-k, this is visible as an elbow. We finally created 3 clusters. One can argue to create 2 clusters from the below graph but taking a business objective, having just 2 clusters (segments) would not be useful for any pharma organization



1. **Arriving at initial cluster centroids**: We tried a few runs with random points of cluster centroids and then also used K-means++ as the init parameter. This algorithm ensures a smarter initialization of the centroids and we observed improved quality of the clustering

Key challenges we observed:

Getting clusters where very few points are associated with one of the clusters and others being heavy. This was overcome by using the init parameter (initializing with k means++ algorithm)

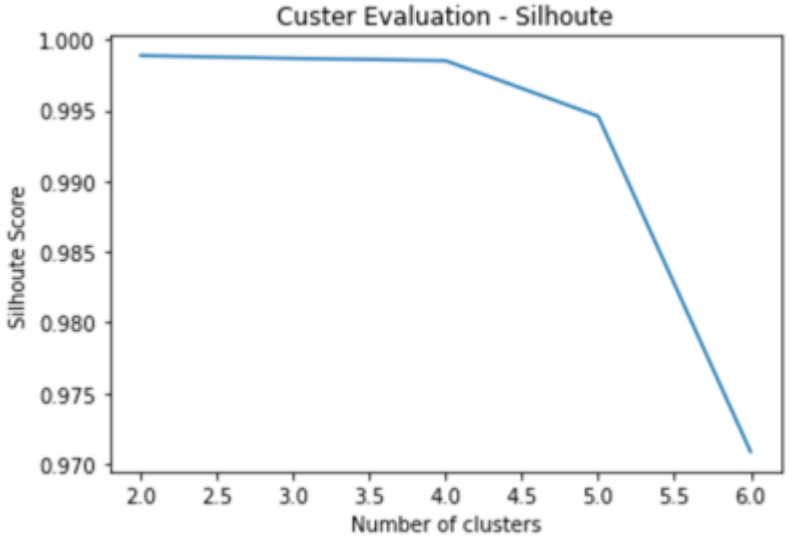
1. **Standardizing the data**: As we are using data columns with different scales, for example, a number of patients seen by a doctor, a number of services being offered by HCPs/ Hospitals, the total amount claimed by Medicare insurance, avg age of patients in each service, etc, we have to normalize the data to bring it at the same scale for which we used standard scaler before applying clustering

2. **Using categorical and continuous variables**: Dataset also contains categorical variables like service codes (HCPCS code), specialty. As K-means works best on a continuous variable, we first used K-mode to create clusters on categorical variables and then used these clusters as input to K-means to arrive on final segmentation
3. Apart from above methodology challenges, one big **logistic challenge** we faced initially was computing power for handling these algorithms as well as a platform to work on this huge dataset. To overcome this challenge, we executed our final algorithms on University provided great lakes systems and to test multiple algorithms and create the code we took a sample of 10-15% of data to work on our local or google colab and deepnote.

Unsupervised Evaluation

Learnings from running different clusters:

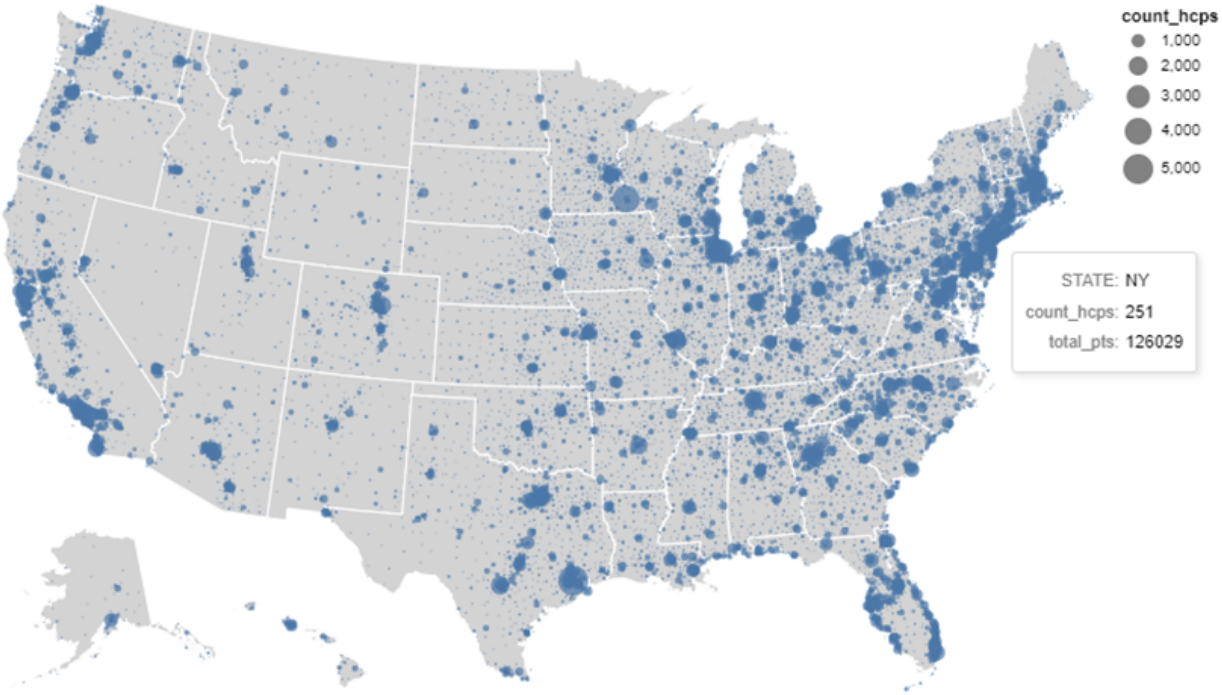
1. Most visits by patients are to general physicians (Like MD, PCP, GP, Internal medicine etc.). The specialty visits are quite low but the treatment costs are too high. Therefore, if don't include specialty in segmentation we might end having incorrect segments
2. As most of our variables are dependent on patient volume at each HCP, it is important to also take into consideration the area where the HCP is practicing. A New York hospital will have more patients compared to sub-urban regions. We took zip level population and used population density to normalize this attribute
3. Due to the number of data points (more than million), hierarchical clustering is not a very good option. This is computationally very heavy and didn't provide good clusters
4. We tested silhouette score on our final usage of different clusters using K Means. It gives good score for 3 clusters that we end up doing:



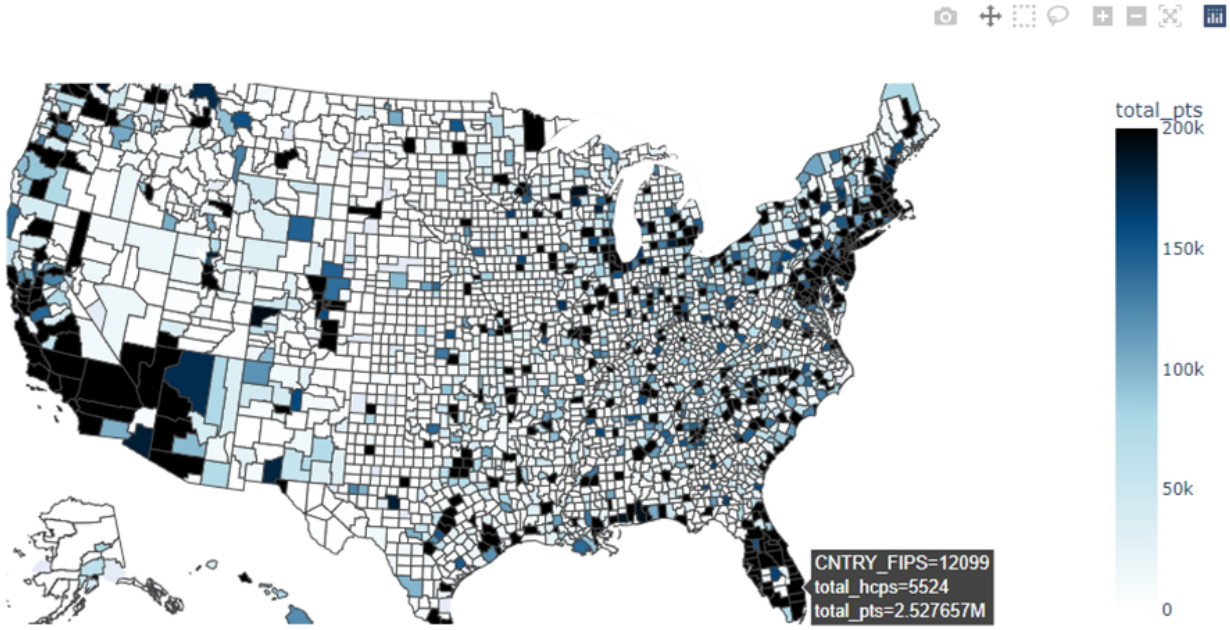

```
For 2 clusters, we have silhoute score: [0.9988673753919458]
For 3 clusters, we have silhoute score: [0.9988673753919458, 0.9986709651805992]
For 4 clusters, we have silhoute score: [0.9988673753919458, 0.9986709651805992, 0.9985104056726504]
For 5 clusters, we have silhoute score: [0.9988673753919458, 0.9986709651805992, 0.9985104056726504, 0.994586182787601]
For 6 clusters, we have silhoute score: [0.9988673753919458, 0.9986709651805992, 0.9985104056726504, 0.994586182787601, 0.97
08631175224681]
```

Final deliverable for business use (i.e. our initial objective of increasing ROI by providing segments, we provide two set of charts for decision making)

1. Location of physicians for top clusters (to help companies target specific geographies) – this chart helps to decide number of reps required to target top segment - by geography



2. Patient population by counties to help focus on selected areas for maximum ROI: Depending on the budget, pharma company can choose which counties to target (only top counties – i.e. darkest shade)



Discussion

In supervised training we discovered that the HCPCS codes are erroneous and vast, however, using good modeling techniques, they still hold predictive power. We were also pleasantly surprised that a regressor can yield good predictions as opposed to using normal time series techniques. Although we believe that this analysis in general is free from causing any potential harm. The nature of the data, in general, has the potential to shift the power to the select few that can analyze this data. When power shifts, there are potential ethical concerns that need to be checked.

In unsupervised training, where we found top segments to target for the best ROI, we are surprised to see how few doctors can give great ROI to pharma companies as they generate very high sales. This ROI keeps on decreasing as they start to include lower segments. If Pharma companies only target the top 1% HCPs, they can get the best ROI but their drug will not reach as many patients as it should be, which should be the goal of developing any drug. By investing more time, this project can be continued to determine the size of field force required to target the physicians and also identify territories for reps that companies can focus on (for ex, rocky mountain regions like Utah, Wyoming, etc. can be deprioritized whereas east and west like NY, Florida, California can be the priority or first experiment)

In part B, Issues that might arise is sharing demographics information of HCPs which can be considered as private. However, to overcome this, we can exclude PDRP physicians from the analysis. PDRP HCPs are those who have restrictions on using their personal information and also do not allow reps to visit them.

Statement of Work

We worked on this project together and have helped each other with any work they are doing. However, at overall, each one of us took end to end ownership of one part each (and others have helped to reach their goals)

1. Part A-1 Cost prediction model: was fully designed and lead by Cindy Xie
2. Part A-2 Demand Forecasting: was fully designed and lead by Louis Pienaar
3. Part B Segmentation: was fully designed and lead by Himank Kansal

Other logistics responsibilities we shared were:

1. Louis Pienaar was responsible for setting up deepnote and collaboration environment
2. Cindy Xie was responsible for setting up and managing Great Lakes environment
3. Himank Kansal was responsible for sourcing and finalizing the datasets to be used in the project

Reference

Code Repository: https://github.com/louisza/MADS-Medicare-Project

Project Report Link:
https://docs.google.com/document/d/16FTtughz1Owduuz1tzx04qO0uneKl9q8t-shNZUVQUE/edit?usp=sharing