Spring 2017

Medicare Analytics

Capstone Project

California State University, Fullerton

Mihaylo College of Business and Economics

Department of Information System and Decision Sciences

ISDS 577

Amogh Newgi
Pranali Jayakar
Santosh Konchada
Yashwanth Vattikuti

Table of Contents

Executive Summary

Governed by the US Federal Government, Medicare is a single payer, national social insurance program for people aged 65 and above who have given parts of their monthly salaries as payroll taxes. Medicare provides insurance to millions of people every year and has some potential insights. CMS collects data for services and procedures provided to Medicare beneficiaries by physicians and healthcare professionals. A separate dataset also contains information about which prescription drugs were given to patients and were paid for by Medicare Part D Prescription Drug Program. Data is available for years 2012 through 2014. We used Excel and Python for cleaning of the data. Data extraction and desired output data frames for visualization was obtained using Python. Later, the outputs from python were used for visualization in Tableau. Finally, we used Rapid Miner to detect Medicare fraud by the providers and achieved a model. The overall aim of the project is to assess healthcare utilization.

Medicare Data

The data is freely available on the website - https://www.cms.gov which is a site for Centers for Medicare & Medicaid Services (CMS). The center is an agency that works in partnership with the state government, health insurance portability standards. CMS collects data for services and procedures provided to Medicare beneficiaries by physicians and healthcare professionals. In addition, data is also provided for inpatient and outpatient customers and

what they were charged for the treatments. A separate dataset also contains information about which prescription drugs were given to patients and were paid for by Medicare Part D Prescription Drug Program. Data is available for years 2012 through 2014. However, we have worked on data for 2014.

Data Cleaning

Bad data is worse than no data at all. We initially had a raw data and it had to be cleaned before conducting any kind of analysis. The cleaner the data more accurate the analysis, Inpatient and outpatient data set had no outliers or missing values at all and for prescription data, we had lots of missing and redundant values. We have eliminated rows and columns using excel option missing data handling, which are irrelevant to our analysis. We also removed duplicates using python to remove redundancy in the data. Doing so we obtained a much accurate data which can be used to conduct analysis and run models. In addition, some variables that do not aid the research have also been deleted.

Variables Utilized in the Medicare

Our data set contains many variables, some of the major variables are hospital name, city, address, zip code, Average Medicare Payments, Prescriptions, Diagnostic Conditions etc.

1. **Provider Id**: The CMS Certification Number (CCN) assigned to the Medicare certified hospital facility.
2. **Provider Name**: The name of the provider.
3. **Provider Street Address**: The provider's street address.
4. **Provider City**: The city where the provider is located.

5. **Provider State:** The state where the provider is located.
6. **Total Discharges**: The number of discharges billed by the provider for inpatient hospital services.
7. **Average Covered Charges**: The provider's average charge for services covered by Medicare for all discharges in the MS-DRG. These will vary from hospital to hospital because of differences in hospital charge structures.
8. **drug_name** – The name of the drug filled. This includes both brand names (for drugs that have patent protection) and generic names (for drugs that no longer have patent protection). In the calendar year 2013 data, there are a proportion of cases where the drug name and generic name could not be determined from the NDC on the PDE record. In these instances, the drug name is blank.
9. **total_claim_count** – The number of Medicare Part D claims. This includes original prescriptions and refills. Aggregated records based on total_claim_count fewer than 11 are not included in the data file.
10. **total_day_supply** – The aggregate number of day's supply for which this drug was dispensed.
11. **total_drug_cost** – The aggregate total drug cost paid for all associated claims. This amount includes ingredient cost, dispensing fee, sales tax, and any applicable vaccine administration fees. The total drug cost is based on the amounts paid by the Part D plan, Medicare beneficiary, government subsidies, and any other third-party payers.
12. **Average Total Payments**: The average total payments to all providers for the MS-DRG including the MSDRG amount, teaching, disproportionate share, capital, and outlier

payments for all cases. Also, included in average total payments are co-payment and deductible amounts that the patient is responsible for and any additional payments by third parties for coordination of benefits.

13. **Average Medicare Payments**: The average amount that Medicare pays to the provider for Medicare's share of the MS-DRG. Average Medicare payment amounts include the MS-DRG amount, teaching, disproportionate share, capital, and outlier payments for all cases. Medicare payments DO NOT include beneficiary copayments and deductible amounts nor any additional payments from third parties for coordination of benefits.

Data merging

The comprehensive data that is needed for the project is in three separate excel files. To answer our research question, the data must be consolidated. As an initial solution, the group has decided to create a data warehouse in MS SQL server to be able to get insight from the data. However, the data was neither transactional nor was the database live. Therefore, we felt that data warehouse would not provide any added value to our analysis as the slicing and dicing of the data could be done in python.

Research Questions

An analysis consisting of inpatient costs, diagnostic conditions, total discharges and DRG procedures have been used to examine these 7 questions.

1. Average Inpatient Cost in each city and state

Finding insights into this question can help managers devise solutions to reduce costs by implementing practices from cities and states where the cost is the lowest.

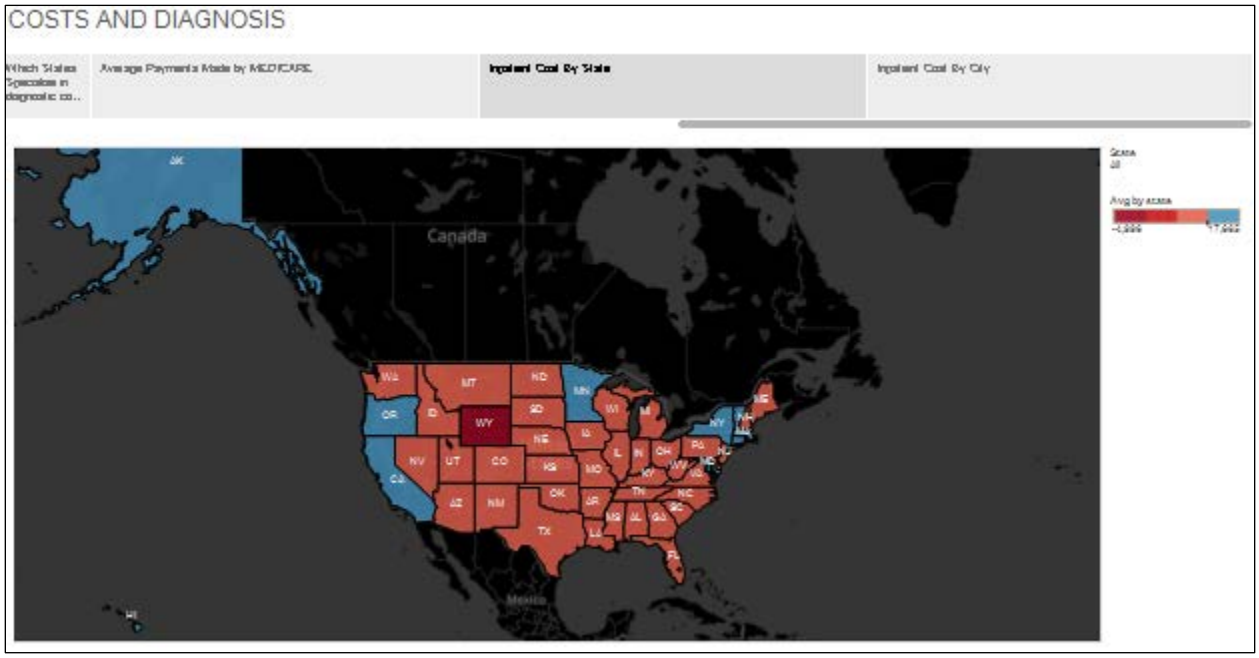

Figure 1: Average Inpatient Cost for each State in United States

The Figure 1 above shows the average inpatient cost across the United State by state. The data was aggregated for all the DRG conditions that were listed in the original data from CMS. The raw data provided details such as total discharges for each DRG, average and total payments made by Medicare at the Provider Id level of granularity. The calculations required to find the average inpatient cost included aggregating cost regardless of the DRG code and then grouping the data at state level. The final output is shown in the graphic above. We found the average cost for inpatient across the United States to be around $12000. This figure was used to color code the graphic. The blue states are the most expensive states in the United States with average costs significantly exceeding the average for the country. On the other hand, the red

states are around the average for the country or on the lower spectrum of the amount. We also see that Wyoming has is the brightest red. This is because data for that state was not available due to state regulations.

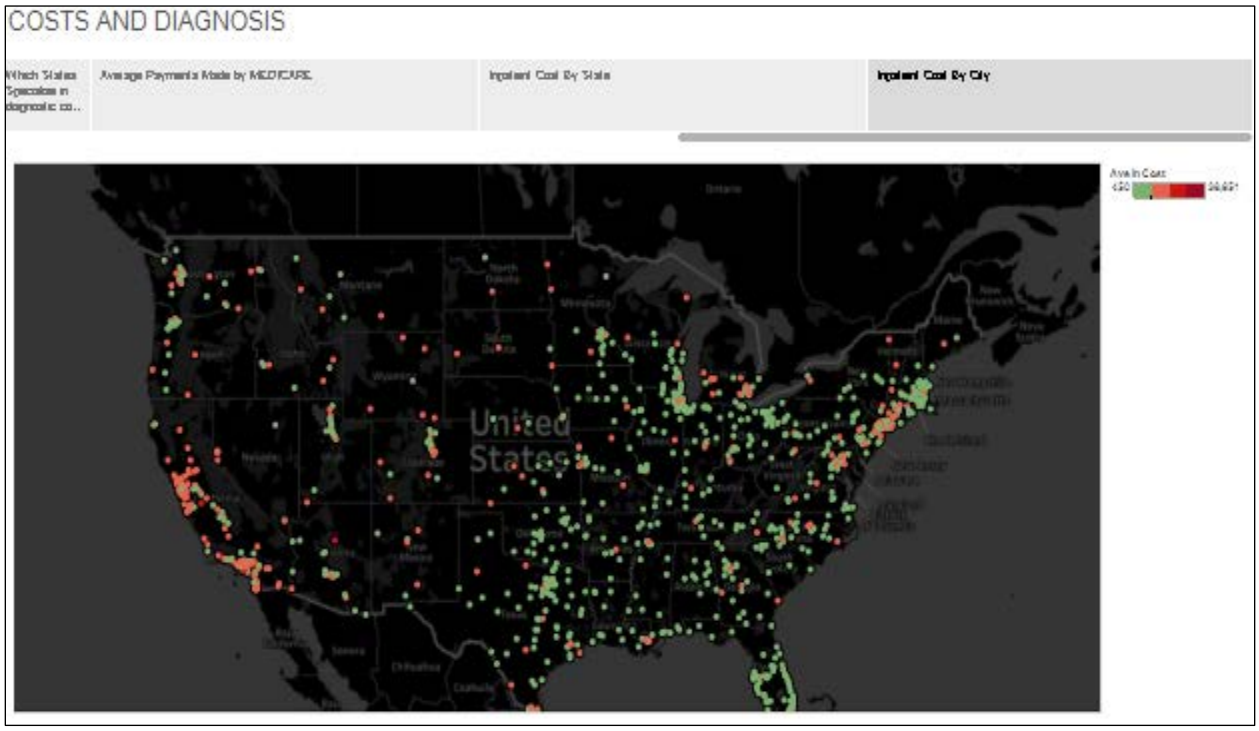

Figure 2: Average inpatient cost by city in United States

The Figure 2 above serves a similar purpose as Figure 1. After analysing inpatient cost for at the state level, we drill down into finding what the inpatient costs are across cities in the United States. The red colored dots show the cities that have costs significantly higher than the average for the country. As we can see most of the more expensive cities are clustered on the east and west coast, particularly in larger cities where standard of living is higher. Another reason we feel this might be because higher population levels drive costs in these areas. A distinction that we can make from our output is that Florida which is one of the most expensive

states in the country does not have cities that are as expensive. The costs in Florida as a state are driven higher by a smaller portion of the cities in the state.

2. Most Common Inpatient Diagnostic Conditions in the United States

Finding out the main reasons why a patient was admitted can help in implementing solutions to reduce costs. Insight into which were the most diagnosed condition can be vital into deciding which diseases need more research funding.

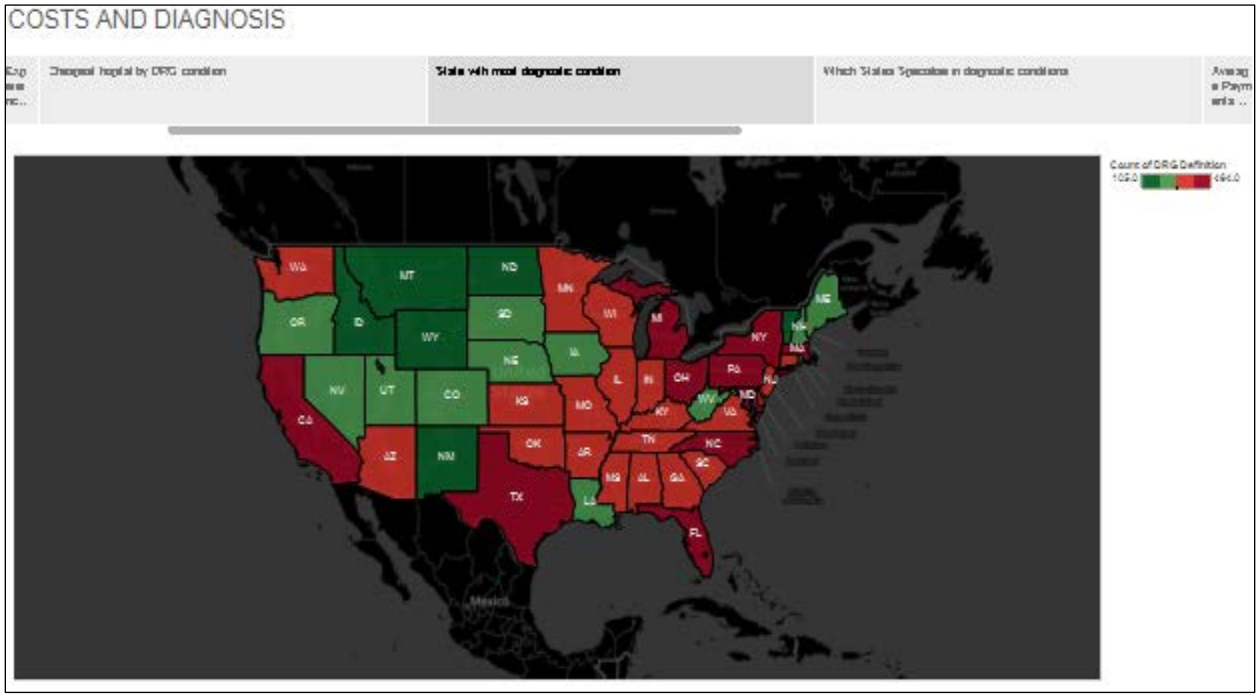

Figure 3: Most Common Diagnostic conditions across states

Figure 3, as shown above shows the states that have claims for the all the diagnostic conditions. The color coding shows the count of diagnostic conditions in each state with states that are red have the most diagnostic conditions claims and the color fades or is green as the number of conditions across the states decreases. As we can see the states of Florida, California, Texas and New York among some other states have the most number of

diagnosed conditions in the country. Each of these states have had claims for at least 400 out of the 564 possible diagnostic conditions in the data. The state with the least DRG diagnosis is Alaska. The differences in these states could possibly be explained by older people moving to the higher states after retirement.

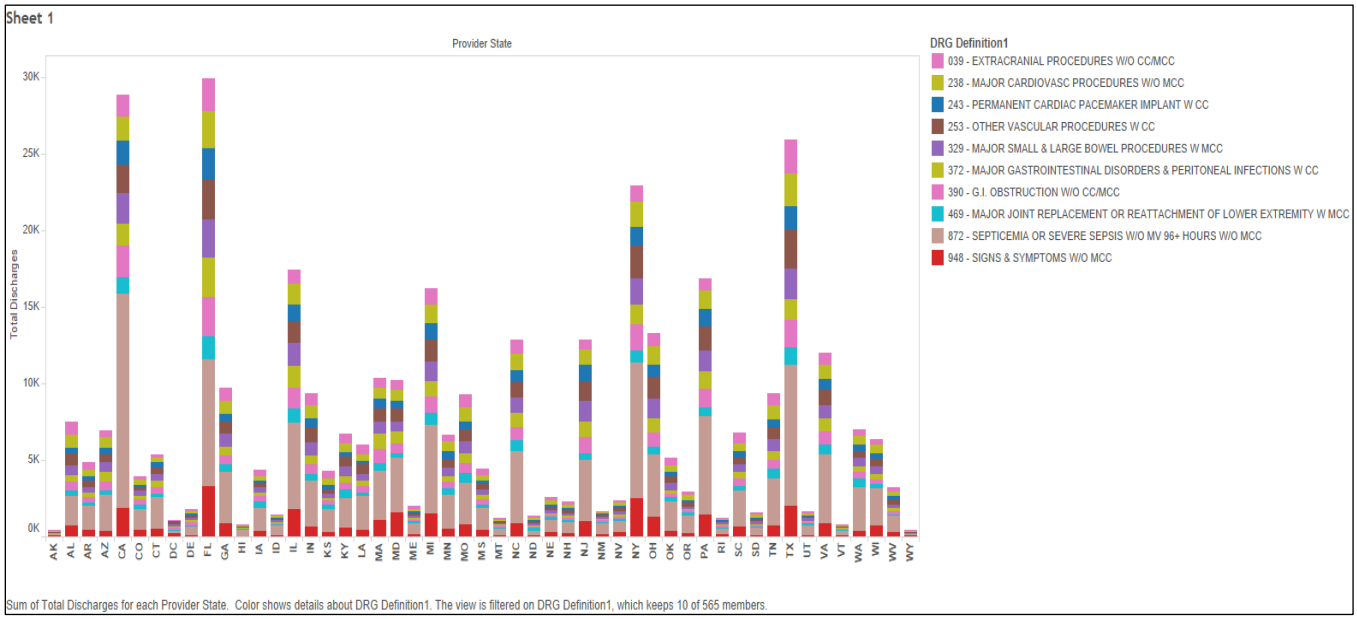

Figure 4: Most Diagnostic Condition among all the States

After analyzing which states had the most claims for all the DRG diagnosis codes, we drill down into which are the top diagnostic conditions in the United States and what states they occur in most. Look at eh graphic in Figure 4 above, we see that Septicemia is the most common diagnostic condition in the United States. We then see how these diagnostic conditions spread across each state. Again, we see the same states we analyzed above with the most DRG codes. Major joint replacement is also another condition that seems to be very prevalent with the population above 65 years of age.

3. Average state wise payments for all the diagnostic conditions

In the following figure, we see the average payments made by Medicare for the diagnostic conditions state wise across United States. The red shade shows the maximum average amount which gradually changes to green as the average amount comes down. Undoubtedly the bigger states like California, Texas, Florida and New York are the top four states where the average payments are the highest. Population could be one of the reasons for the vast difference amongst the states.

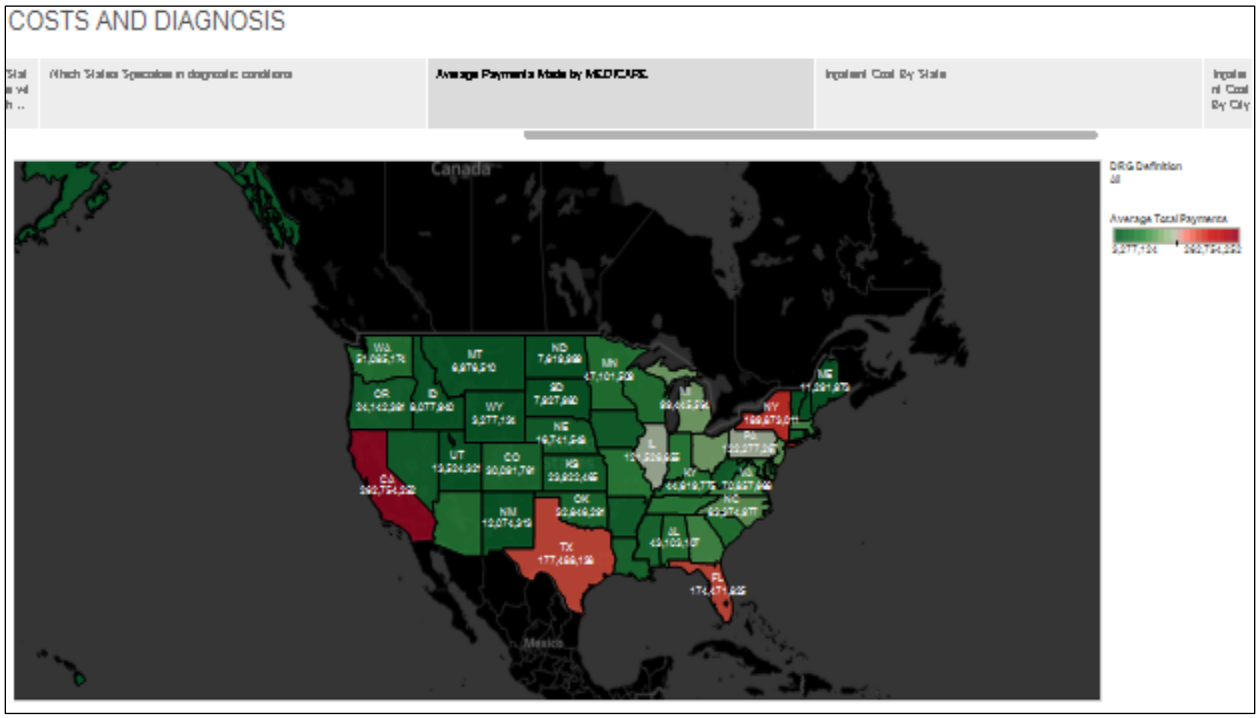

Figure 5: States with most diagnostic conditions

4. Highest Number of Cases for each diagnostic condition in a State

Figure.6 shows which states in the United States have the highest number of cases for each diagnostic condition. We have added a filter in order to fit in all the DRG definitions and show state wise specializations accordingly. In the figure, we have filtered the DRG definition to 'Heart Transplant' to show which states performing that procedure have the most cases. Here we see that greener the color of the state, in this case, Texas, Illinois and California, maximum the number of the discharges are for the procedure. Whereas, the red shade shows the discharges that were comparatively lesser than the green ones. The darker the shade of red, lesser is the number of discharges for that particular disease.

Figure 6: Experience through practice

5. Most Prescribed Medication and Cost in each state across United States

This question can help pharmaceutical companies forecast production of specific prescription drugs. The following figure shows the top 5 medications prescribed in the United States. The number shows the total cost of that particular medication in millions. We clearly see that Levothyroxine Sodium is the most prescribed medicine which treats underactive Thyroid (Hypothyroidism).

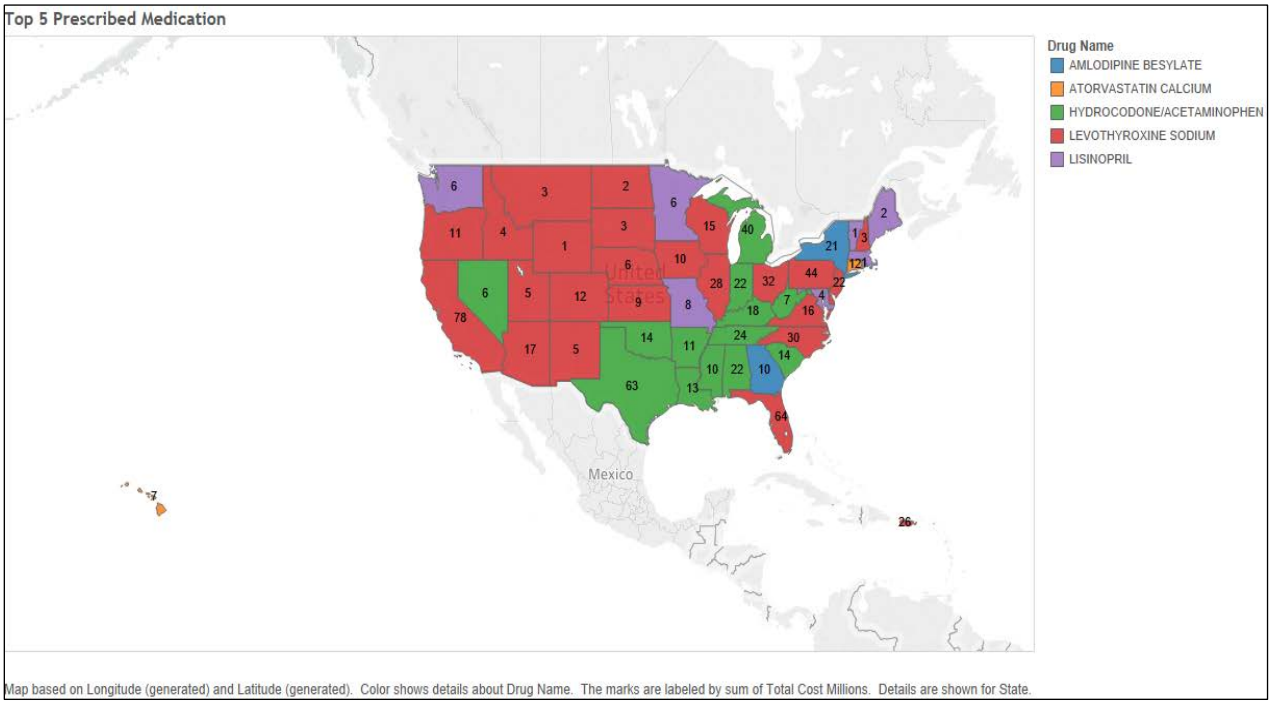

Figure 7: Most prescribed medications across United States

6. Most Experienced Hospital or State for a Particular Procedure

The below figure represents the most experienced hospital across all the states in United States for a particular procedure. We can say that a hospital is experienced for a particular procedure based on the number of operations it has performed in our case it is total discharges. The above visualization initially shows the most experienced hospital as Florida

hospital by accounting for all the 564 procedures we have in our data. When we drill down to a particular procedure say 'Heart transplant' it will show Barnes Jewish hospital as the most experienced hospital in USA with total discharges of 77 and it is in the state of Missouri. Similarly, we can look for other procedures through this visualization, we can further add another filter as state and find out which is the most experienced hospital for a particular procedure in a particular state.

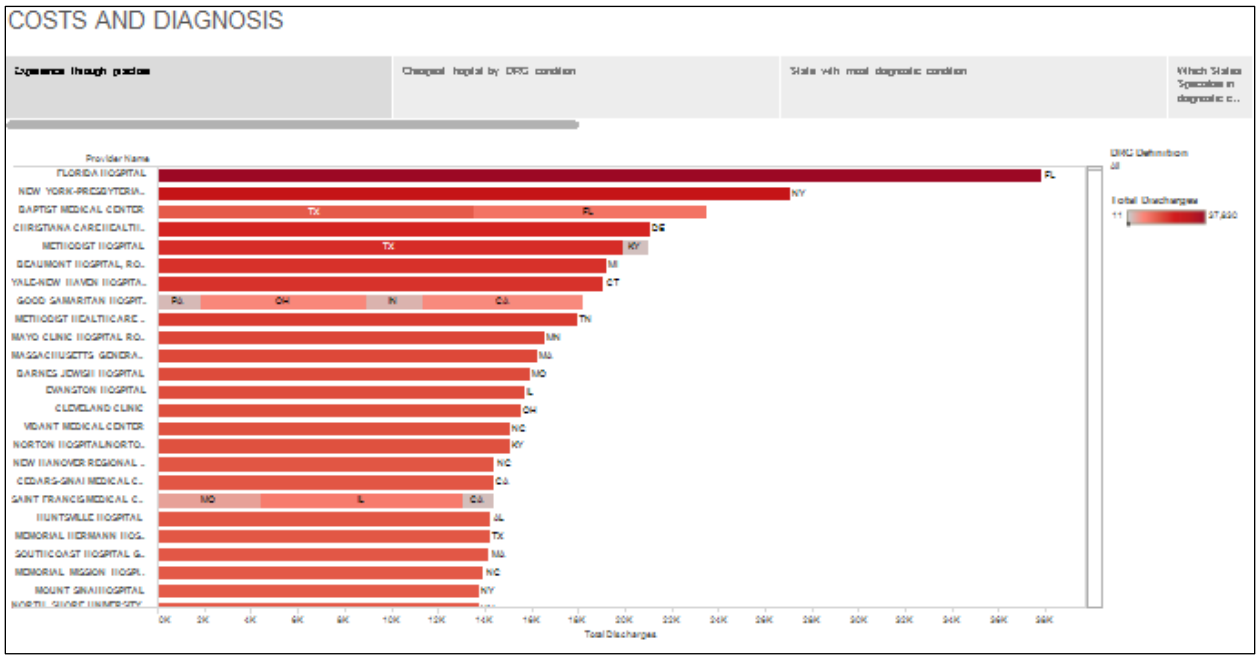

Figure 8: Most experienced Hospital across United States for a particular procedure

7. Cheapest Hospital or State for a Particular Procedure

The below figure represents the cheapest hospital across all the states in United States for a particular procedure. A hospital is said to be cheapest for a particular procedure based on the average total payments. The above visualization initially shows the cheapest hospital as pioneer community hospital by accounting for all the 564 procedures with the average total payment of $3,138. When we drill down to a particular procedure say heart transplant it will

show Abbott northwestern hospital as the most experienced hospital in USA with total average payment of $1,63,954 and it is in the state of Minnesota. Similarly, we can look for other procedures through this visualization, we can further add another filter as state and find out which is the most cheapest hospital for a particular procedure in a particular state.

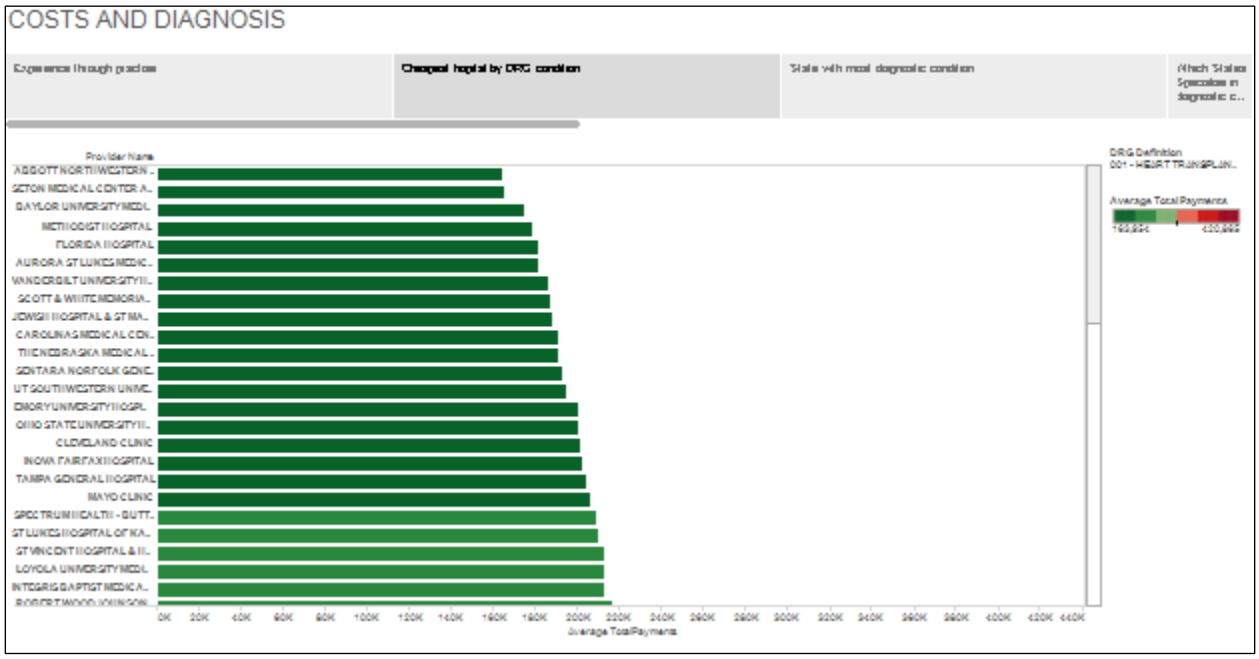

Figure 9: cheapest Hospital across United States for a particular procedure

Supplemental Analysis: Fraud Detection

Researching the internet for malpractices we found various ways by which providers can commit fraud. One of the notable cases we found involved a clinic charging a group of Alzheimer patients for group therapy when they were watching a movie in the waiting lounge of the clinic. This peaked our interest and one of the main objectives of our analysis was to detect fraud in terms of the charged amount for each claim with the diagnostic condition and the prescribed medicine.

Fraud detection

In this section, we will be talking about models we performed to detect fraud in Medicare. For detecting fraud, we have analyzed Part D: Prescriber and Medication data set. This data set is that it contains Provider id, the total amount of prescriptions, total cost and number of days prescribed. Interesting thing about this data is we can create a few features to qualify the chance of each physician to be a fraudster.

The big question is why would a physician defraud Medicare. Maybe to make more money by either selling drugs (fake prescriptions) or by charging too much money from Medicare for some drugs. So, we focused on detecting the drugs which have higher cost because there is an unlikely chance for committing fraud for smaller amounts of money. To detect this high cost prescriptions, we performed anomaly detection using Rapid Miner.

Rapid Miner Model Selection

Anomaly detection is similar to finding outliers in the data which do not belong to rest of the data. For detecting these outliers, we used K nearest neighbor(KNN). This model is run based on a formulation of distance-based outliers on the basis of distance of a point from its k-th nearest neighbor. All the points are given a ranking based on its k-th nearest neighbor. The top n points in this ranking are declared as outliers. Different distance functions are supported in this model the function we selected is Euclidian distance. we proceeded to run the model by the following steps. Firstly, we must take a sample data because rapid miner does not support the complete data set that we have. We took sample data set of 2000 instances and imported into rapid miner and we chose a subset of attributes which are key

in our analysis like provider id, drug cost, number of prescriptions etc. The model detects outlier based on the variable 'total drug cost' which we have selected beforehand. later the model is executed to find outliers.

After executing the operator adds a new Boolean attribute named as outlier. If the value of this attribute is true, then it is considered as outlier and vice versa. The below figure is the sample output file we obtained after model execution.

DRUG_NAME	TOTAL_DAY_SUPPLY	TOTAL_DRUG_COST	NPI	TOTAL_CLAIM_COUNT	outlier
CITALOPRAM HBR	1301.0	268.9	1710030812.0	47.0	false
BUMETANIDE	540.0	211.3	1437130549.0	18.0	false
CITALOPRAM HBR	990.0	111.7	1932491032.0	11.0	false
OMEPRAZOLE	9444.0	2850.4	1457429359.0	374.0	false
GAMMAPLEX	336.0	55290.4	1043248727.0	12.0	true
ENALAPRIL MALEATE	1264.0	608.6	1346248358.0	42.0	false
METOPROLOL TARTRATE	1020.0	86.1	1194060848.0	34.0	false
RANITIDINE HCL	748.0	263.9	1669482907.0	26.0	false
FUROSEMIDE	4396.0	396.3	1528086857.0	108.0	false
NITROFURANTOIN MONO-MACRO	156.0	650.1	1154384782.0	21.0	false
DIGOXIN	1110.0	985.1	1750572087.0	33.0	false
ANDROGEL	330.0	7258.7	1821153305.0	11.0	false

Figure 10: Outlier Detection using K Nearest Neighbor

From the above figure, we can observe that the first five columns are attributes and the highlighted column is where outliers are detected. The red marked true Boolean value indicates that the particular drug cost is an outlier and false indicates they are not outliers.

The below graph is visual cluster representation of Drug name and total drug cost.

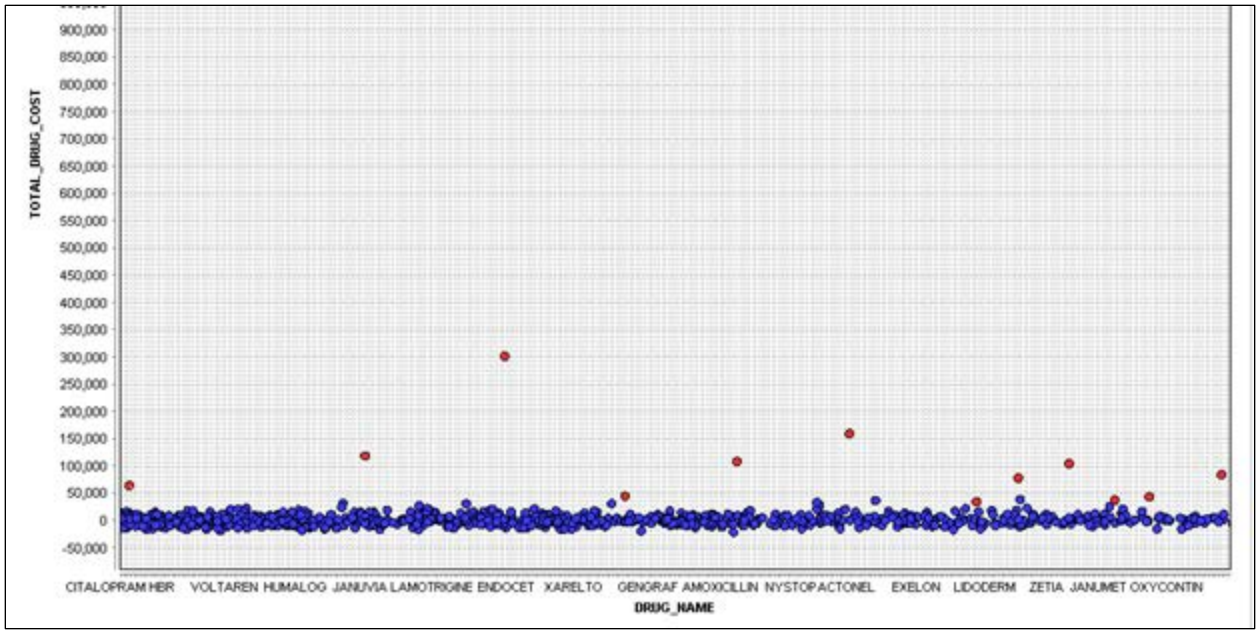

Figure 11: Comparison of Drug costs associated with Drug name

We can see from the graph that most of the drug costs fall under an average price range. The red dots in the above graph represents as outliers as the prices are far from the other prices.in other words, for a particular drug different providers from different states are selling them at variable prices the red marked dots indicate there is a big price difference from the rest of the providers.

From the analysis of this model we can assume that there is a chance that these providers might be committing fraud. Although we got the desired output we are not sure about the accuracy of these results. To verify the model, we ran another model known as cross validation in Rapid Miner.

Cross Validation

This model is based on patient information by training and applying a gradient boosted tree model. This is used to check the accuracy fit and performance of our K nearest neighbor model. There are 3 steps involved in running fraud detection model. As you can observe from the figure they are in the order of retrieving data, remove correlation and cross validation.

Figure 12: Operators used in Medical Fraud Detection model

In the first step data is imported into the model and is converted into numbers to run the algorithm. This operator performs full meta data processing so that all the data transformations are possible.

The second step is removing correlations, a correlation is a number between 1 and -1 that measures the degree of association between any two attributes. A positive value means a positive association and vice versa for negative. For example, if large values of one attribute are associated with large values of other or small values of one in associated with small values of other those attributes are said to be positively correlated. This operator can be used for removing correlated or uncorrelated attributes depending on the setting of parameters. We used a 90% setting for our analysis meaning that we automatically remove attributes which are correlated more than 90 percent. The used correlation function in this

operator is the Pearson correlation. Correlated attributes are usually removed because they are similar in behavior and will have similar impact in prediction calculations, so keeping attributes with similar impacts is redundant. Removing correlated attributes saves space and time of calculation of complex algorithms.

The third and final step is cross validation. This operator performs a cross-validation in order to estimate the statistical performance of a learning operator. This is used mainly to check the accuracy of a model. It uses an GBT algorithm to infer fraudulent behavior. The cross validation is a nested operator which has two sub processes namely training and testing. The trained model is applied to testing sub process. The input data is partitioned into k subsets of same size. A single subset from these subsets is retained and used as the testing data set and remaining subsets are used as an input for training sub process. The cross-validation process is repeated for all the subsets with each one of the subset selected as the testing data. Later, the results obtained from all the iterations are used to produce a single estimation. The learning processes usually optimize the model to make it fit the training data as well as possible.

On performing all the steps the output we achieved is a confusion matrix which is shown in the below figure.

accuracy: 98.90% +/- 1.37% (mikro: 98.90%)

	true false	true true
pred. false	980	6
pred. true	5	9
class recall	99.49%	60.00%

Figure 13: Confusion Matrix between KNN and Cross validation

From the confusion matrix, we can see that we achieved 60 percent in true false column. This means that the outliers from KNN model and outliers from fraud detection model are matching 60 percent of the time. To increase the accuracy of the model we performed the same process with different number of outliers. However, the results obtained are similar to the mentioned above.

Conclusion and Recommendations

Our dataset was downloaded from the CMS website. The data was contained with two separate files, one for inpatient claims and other for the claims that included the medication prescribed by the provider. We used python to clean the data for handling missing value and eliminating variables that we did not use for our analysis. After slicing and dicing the data in python, we created the output files to be visualized in a story format in tableau. We have found out that Florida and California among some other states along the coast are states with the most claims, both in terms of the number of discharges as well the different diagnoses and conditions treated. The higher number of claims from these states could be explained by the fact that a lot of people above the age of 65 move there after retirement, which is particularly true for Florida. We also analyzed which states had the most experience for each diagnostic condition available, in terms of the number of discharges that had been done. Again, the same states mentioned above were among the results. Analyzing the cost, we also found the same states with the most diagnosed conditions had the highest cost across the United States. However, when analyzing the cost across cities, we found that it was a small portion of hospitals or cities in a few states that were driving the costs higher rather than the whole state being expensive over all. We further drilled down into the data to find out which states and hospitals were experienced and what were the costs for specific diagnostic conditions at these hospitals, similar to the analysis conducted for state.

In addition, we conducted anomaly detection on the Part D: Prescriber and Medication data available from CMS. The aim was to find any claims that showed behavior different to the

other claims that were similar to it. The major part of our analysis was based on finding if any claims were charging more money for medications than all other claims that had similar characteristics in terms of the cost of the drug for number of people being treated across all the states, the amount of medication required to treat that diagnostic condition.

Analyzing our outputs, people that need treatment can find out which specific hospital, in which state and city would be the best for them when considering several parameters. The decision could be based on parameters such as cost or the amount of experience each hospital has with specific diagnostic conditions. In addition, the data for CMS was made freely available in 2014 under instructions from Obama. One of the main reason for this was to reduce and stop fraud that was being conducted in the healthcare industry. Our method can provide a good gateway into some of the methods that can be used to detect unusual patterns in the data.

References

1. Hongxing He, Warwick Graco, Xin Yao (1999). “Application of Genetic Algorithm and k-Nearest Neighbour Method in Medical Fraud Detection.” Retrieved From: https://link.springer.com/chapter/10.1007/3-540-48873-1_11

2. Pierre Gutiereez (2016), “Detecting Medicare Fraud.” Dataiku. Retrieved From: https://blog.dataiku.com/2015/08/12/medicare-fraud

3. Dallas Thornton, Roland M. Mueller, Paulus Schoutsen, Jos van Hillegersberg (2013). “Predicting Healthcare Fraud in Medicaid: A Multidimensional Data Model and Analysis Techniques for Fraud Detection.” Retrieved From: http://www.sciencedirect.com/science/article/pii/S2212017313002946

4. Rasim Muzaffer Musal (2010). “Two models to investigate Medicare fraud within unsupervised databases. “Retrieved From: http://www.sciencedirect.com/science/article/pii/S0957417410005993

5. Joseph Goedert(2015). “CMS Data Chief Urges Use of Medicare Data Analytics, Tools.” Retrieved From: https://www.healthdatamanagement.com/news/cms-data-chief-urges-use-of-medicare-data-analytics-tools

6. Terry Wing (2016), “Advanced data analysis helping to nab Medicare cheaters.” Retrieved From: https://federalnewsradio.com/big-data/2016/08/advanced-data-analysis-helping-nab-medicare-cheaters/