Health Insurance Coverage Study

1. Introduction

1.1. Insurance coverage v.s. Medical cost

Recent years, US medical expenses are becoming unaffordable for more and more families. It is the third largest cause of bankruptcy for a family, and the first cause is the loss of job due to medical problems[i]. And, at the main time, healthcare cost increased tremendously. According to Brett O'Hara's research, the total healthcare cost growth rate in 1996 was only 2%, while in 2001, it was already 10%[ii]. Then from 2008 to 2010, the cost increased at a rate about three times inflation, reported by Charu Chandra, ect[iii]. For individuals, Zeldin and Rukavina pointed out that in 2007, average indebted amounts of credit cards for non-medical expenses is $8333, and for medical expenses is $12515[iv]. Medical cost obviously is a huge burden, even for middle-class Americans[v]. In order to avoid going bankruptcy, besides improving health quality, people can only turn to health insurance for help. And data showed that, health insurance can significantly minimize wealth depletion for older adults[vi]. People without health insurance had 49% more debt than the average, however, those with Medicare or Medicaid had 29% and 13% less than the average, respectively. It seems that people should all buy health insurance to lower their risk of bankruptcy. However, in reality, many regions have relatively low health insurance coverage. Then what kind of people are willing to buy more health insurances? What is the commonalities for the areas with high insured rate or low insured rate?

In order to study in what kind of places, insured rate is usually high or low. We choose certain factors which are relative to insurance coverage.

1.2. Insurance coverage v.s. Population

Due to different cultural environments between different geographical locations of the United States, you will often see different educational background, racial components, etc. plays a role in people's spending behavior. According to Tang[vii], the attitude towards money of people in Taiwan varies differently by different background factors like social, political, religious, etc. Similar theories about financial attitudes varies across different geographical regions in the States is also pointed out by another study (Jorgensen, Foster, Jensen and Vieira)[viii]. Thus, it is reasonable to suspect that since insurance buying behavior is strongly associated with financial attitudes, which could affected by different regional factor listed above. In fact, many health providers requires a health insurance coverage or charge a prohibitively high fee as an alternative (Himmelstein et al. 2005[ix]; Institute of Medicine 2003[x]; Kasper et al. 2000[xi]). Thus, Insurance expenditure plays a crucial part of people's financial managements and is strongly affected by different financial attitudes across the States, so we can suspect that people's insurance-buying behavior varies in in different geographic regions. According to the past survey by US census bureau, the insured rate is the lowest in the Midnorth, the Midsouth and the Southeast part of the country.[xii] Another study also discovered the same regional trend for people under 65 years old, that the uninsured rate is higher in most of the Southern states.[xiii] One of the main parts of our study is to verify this trend in health insurance coverage in different regions in United States.

1.3. Insurance coverage v.s. Population

In economics, population density in a certain area is always considered as a measurement of the economic growth there. Chenavaz, R., & Escobar, O stated that population could be used as a GDP indicator and it has a positive correlation within the scope of United States. In their

study, they introduced the concept of "effective area". It refers to the area where is directly used for production and living at the time under consideration.[xiv] And population is positively correlated with the effectiveness of that area.[xv] Another finding of their study is that the effectiveness of the area is positively associated with economic growth.[xvi] Hence, the population is a useful indicator to study economic growth when using the concept of "effective area" as a bridge in the middle.[xvii]

In the meanwhile, it seems natural to believe that in a prosperous areas, individuals will have more spare cash and larger purchasing power. Then, will high insured rate shows up in that areas? In order to check this theory, we choose to examine whether more economically robust area which is the area with larger population density always have higher insurance coverage.

1.4. Our study

From above background information, we know that insurance rate can varied in different kind of areas. Then in what kind of regions, will more people buy insurance than other places? Are the population density positively correlated to the insurance-purchasing behavior? Or the medical cost acts as the same way? These informations can be really helpful for an insurance company to cut down their administrative cost. Studies shows that nowadays the administrative costs are expensive for the insurance companies, and roughly 12% of their total revenue goes into these costs.[xviii] If insurance companies can relatively estimates their revenue by knowing how much people are willing to purchase health insurance in one region, they might relocated their resources accordingly to minimize the administrative cost. In other word, if the insurance company could gain information about certain featured population and insured rate, they can prevent many over-investing in advertising resources in many locations. The companies can use this money on lowering their products price so that more people will be willing to purchase cheaper insurance plans and safeguard their family's financial future. From the perspective of the development of the insurance company, if we discovered the trend that the three factors we want to examine (geographical region, population and hospital expenditure) are not independent with the insurance coverage, the insurance company could benefit heavily from these informations in terms of campaign strategies and advertisement distribution. They can find those niche groups that fit in the demographic features of a highly insured group but have a low insurance coverage in reality to be their target audience for insurance product. The result from our study, if statistically significant, could benefit the insurer and the insuree in the US as a whole.

In our study, we studied the relationship between the health insurance coverage, the medical expenses, the population and the geographical position.

2. Data and Statistical Method

2.1. Data Resources

In order to conduct our study, we searched and collected two relevant datasets. The first one is about the insurance coverage in the United States, and the second one is about the medical expenses by hospital in US. These two datasets share the attributes of their geographic location, which means they are able to be combined into a single dataset by using the combination of county and state as the primary key, so that we can proceed following analysis procedures.

Here is the list of all variables after removed duplicate attributes representing the same thing, along with their description:

Variables	
County_State	County name and State name.
Number_Insured	Number of people with health insurance.
NInsured_Cl_LowerBound	The lower bound of 90% confidence interval of the insured number.
NInsured_Cl_UpperBound	The upper bound of 90% confidence interval of the insured number.
Number_Uninsured	Number of people without health insurance.
NUninsured_Cl_LowerBound	The lower bound of 90% confidence interval of the uninsured number.
NUninsured_Cl-UpperBound	The upper bound of 90% confidence interval of the uninsured number.
Time	Survey time.

Table 1. Useful attributes in first dataset.

Why are these variables useful:

County_State: We need to merge two data sets together using County and States information and also it is also helpful to keep a track of the hospital expenses in terms of geographical location.

Number_Insured & Number_Uninsured: We are here to investigate the relationship between people with health insurance and hospital expenses so the number of people with health insurance is critical. Our key attributes—insured rate by county, which is obtained using the formula: Number of people with insurance/ (Number of people with insurance + number of people without insurance). Besides, we can calculate the population size of each county by adding the number of insured and number of uninsured.

NInsured_CI_Lowerbound & Ninsurance_CI_Upperbound: The 90% confidence interval is obtained here so we can get a whole picture of how the mean of the insured population of this county is distributed.

NUninsured_Lowerbound & NUinsured_Upperbound: The 90% confidence interval is obtained here so we can get a whole picture of how the mean of the uninsured population of this county is distributed.

Time: As the Medicare data was collected within a late-2013 to early-2015 period, it is critical to get the values in these years and the average value will be more representative. In this way, the result of our investigation is time-consistent.

Variables	
Provider_id	The id that hospital registered in Medicare.
County	County name.
State	State name.
Lower_Payment_Est	The estimated lower bound for certain treatments in this hospital.
Ave_Payment	Average payment for certain treatments in this hospital.
Higher_Payment_Est	The estimated upper bound for certain treatments in this hospital.
Measure_id	Type of treatment that indicates which major healthcare treatment the payment corresponds to, including heart, knees, etc.

Table 2. Useful attributes in second dataset.

Why are these variables useful:

Provider_id: This is the unique ID-like code for each hospital investigated in this dataset. It serves as an index for each of these hospitals and it makes us easier for us to locate each row of hospital information.

County: Used as one of the geographical components to associate the hospital information with the insurance information in the previous date set. Together with the county code, it is used to merge two data set together.

State: Used as one of the geographical components to associate the hospital information with the insurance information in the previous date set. Together with the state code, it is used to merge two data set together.

Lower_Payment_Est & Higher_Payment_Est: The range is obtained here so we can get a whole picture of how much money people spend for certain treatments in this hospital.

Ave_Payment: This is the most representative measurement for how much people spend for certain treatments in this hospital.

Mesure_id: It will help generate an average overall healthcare payment of each hospital.

2.2. Data Cleaning

The cleaning strategies we adopted contains 4 major parts, Cleanliness check, data cleaning, restructuring and binning. Cleanliness check is a function determining whether the specific pattern of formats each attribute is accordance with, and what percentage of all types of missing values (‘’, nan, and ‘Not Available’) each attribute has, and whether the mean attributes is inside its upper-lower bound range, and whether the location attributes are consistent matching real location.

Missing value exists in both of the datasets. In the missing percentages are extremely small (under 0.1%) for most of the incomplete attributes. In the second datasets, even the percentages are nearly 30 percent in the average payment and its range attributes, by digging

into the dataset, we found out it connects to certain treatment, as we finally only need the average value for all kinds of treatment and hospitals in each county, it won't hurt the final result. Therefore, we removed all rows with missing values. Besides, the second dataset has some locations that do not exist, we removed them as well. Inconsistent value is a big issue as two datasets word the county name differently, after checking the data, we found the reason is that the US Census Bureau uses the full name of each county, which is not normal in our daily life, so we revise those name by cutting the 'county', 'borough' and others off.

In order to merge two datasets into a single one, where county and state is the unique key (treated as index), we group all rows by county and state and calculated the mean value of each attributes (no categorical presents here). We then add a 'Region' attribute by using the geographic region identified by National Geographic (The official magazine of the National Geographic Society). Then, we generate some important derivative attributes including population size and county insured rate. Last, we removed the outliers in our datasets, and binned the three factors and target factor we willing to investigate in an equal width way.

Eventually, our dataset has 1906 rows and 17 columns, representing 1906 counties in the US with 17 features.

2.3. Statistical Method

2.3.1. Correlation analysis.

2.3.2. Clustering Analysis:

- K-mean clustering, DBSCAN Clustering, Hierarchical Clustering.
- Accuracy measuring mechanism: silhouette score, PCA decomposition

2.3.3. Association rule Analysis: apriori model.

2.3.4. Network Analysis.

2.3.5. Machine learning Analysis:

- KNN, Decision tree, SVM, Naïve Bayes, Random forest.

3. Data Analysis

3.1. Exploratory Analysis

Exploratory data analysis (EDA) is an approach to analyzing data sets to summarize their main characteristics, often with visual methods. A statistical model can be used or not, but primarily EDA is for seeing what the data can tell us beyond the formal modeling or hypothesis testing task.

Firstly, we calculated descriptive statistics, including mean, standard deviation and median of numeric attributes, and mode of categorical attributes so we can get a general idea about the range and the distribution of the data. The result is shown in the following table.

	Mean	Median	Std
Number_Insured	33421.18625	21400.25	33016.68378
NInsured_CI_LowerBound	32973.66815	21059	32680.6705
NInsured_CI_UpperBound	33868.70435	21721.5	33353.55294

Number_Uninsured	5040.937303	3415.75	4678.520558
NUninsured_CI_LowerBound	4593.419203	3091.5	4307.062047
NUninsured_CI_UpperBound	5488.455404	3721.5	5050.666131
Population	38462.12356	25008	37109.56917
Insured_rate	0.858655535	0.862762381	0.046196201
Lower_Payment_Est	15893.22557	15963.9375	2067.280226
Ave_Payment	17748.72822	17874.25	1944.270856
Higher_Payment_Est	19718.67617	19820.66667	1949.03712

Table 3. Summary of Numeric Attributes

	Mode
Regions	midwest
Insured_rate_EqWidth	Fair
Ave_Payment_EqWidth	Very_Cheap
Pop_EqWidth	Very_Small

Table 4. Summary of Categorical Attributes

According to the results above, all of the mean values of each attributes are greater than corresponding median value, which means the distribution of all these attributes are right-skewed. Besides, the standard deviation of the 'Number_Insured', 'NInsured_CI_LowerBound', 'NInsured_CI_UpperBound', 'Number_Uninsured', 'NUninsured_CI_LowerBound' and 'NUninsured_CI_UpperBound' attributes is very large compared to their mean value, which indicates the data points are spread out over a wider range of values, while the standard deviation of the 'Lower_Payment_Est', 'Ave_Payment' and 'Higher_Payment_Est' attributes is relatively not that large compared to their mean value, which means the data points are spread out over a lesser range of values compared to previous attributes. For categorical attributes, most of the counties are in midwest region, the most common insured rate is fair, most of counties have very cheap healthcare costs, and most of counties have small population. The histograms of the attributes are shown in figures below.

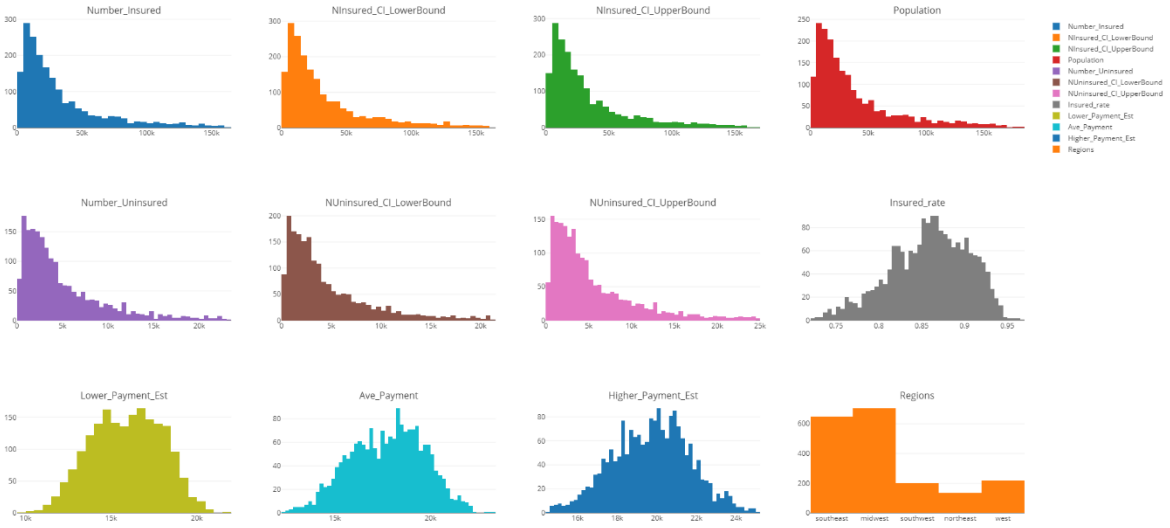


Figure 1. Distributions for each variables.

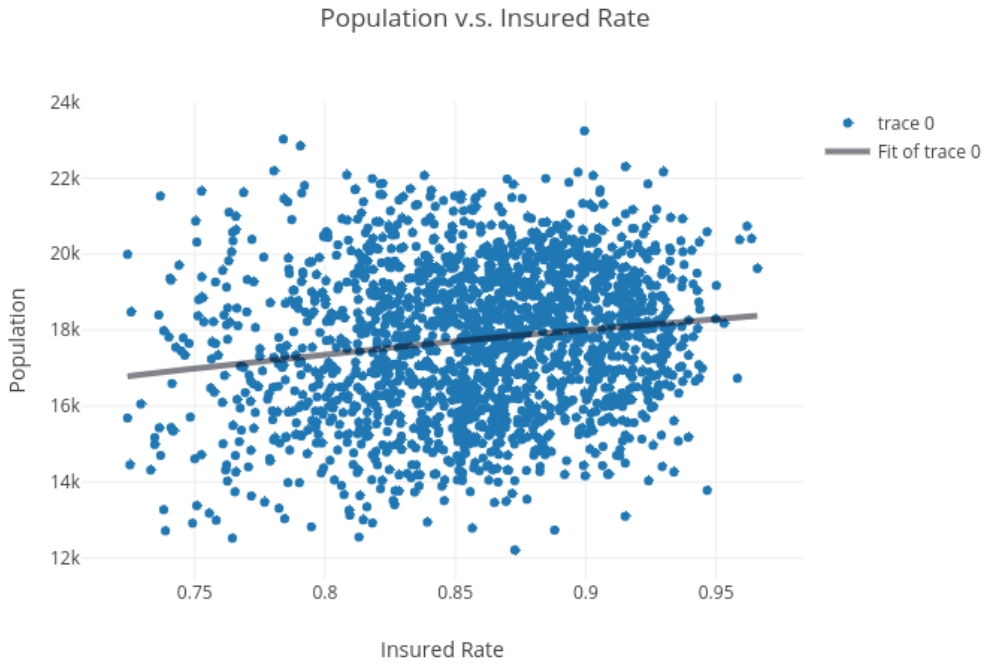


Figure 2. Scatter plot of Insured rate v.s. Population.

To take a simple look at the relationship between the population size and insured rate, we drew a scatterplot between the attributes. In figure 1, even though the points are very dense in the center, the generated regression line will has a tiny slope, which trigged us to further our study.

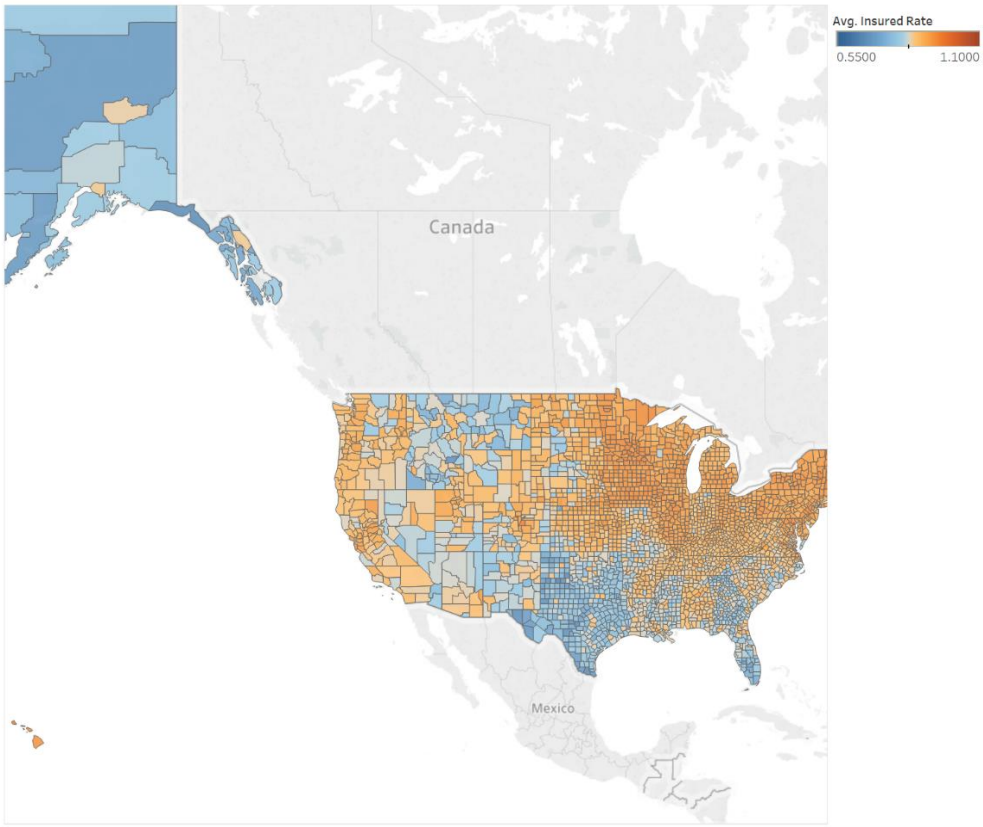


Figure 3. Map for insured rate per county

Then, to overview the insured rate by county, we drew a colorize map, where orange color indicates a high insured rate while blue color indicates a low insured rate. It is obvious that some regions' insured rate higher than other regions. So, we drew a boxplot to dig into the 5 regions we categorized as follows.

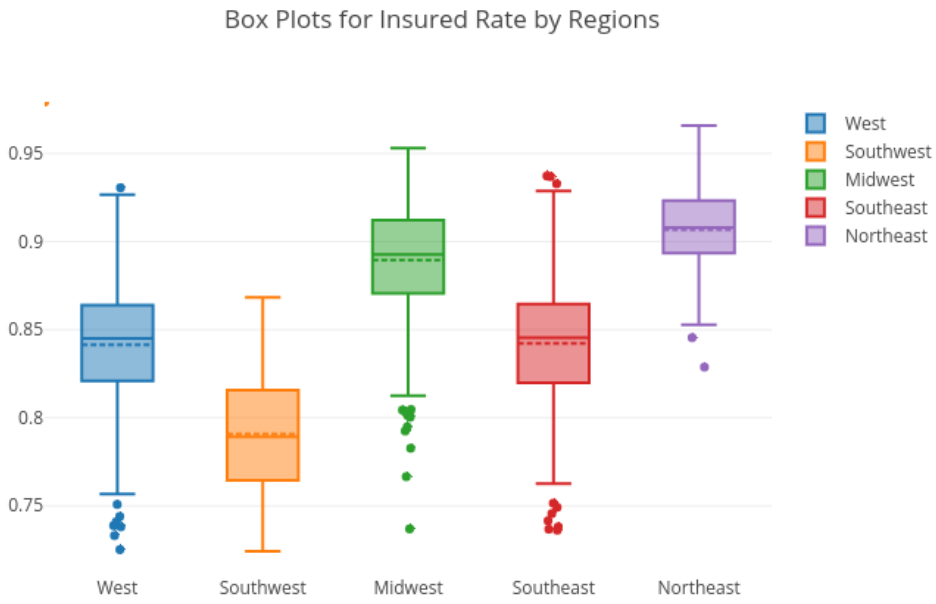


Figure 4. Boxplot for Insured rate by Regions.

From the plot, it is true that Northeast region has the highest insured rate, while the southwest has the lowest insured rate.

3.2. Correlations Analysis

As long as we already learn some basic information about our dataset. Further analysis should be done to inform us general ideas about the relationship between the insured rate and other three factors.

names	umber_Insur	ed_CI_Lower	ed_CI_Upper	mber_Uninsu	red_CI_Lowe	red_CI_Upp	Population	Insured_rate	er_Payment	Ave_Paymen	her_Payment	Regions
umber_Insur	1	1	1	0.86	0.85	0.86	1	0.28	0.48	0.37	0.24	-0.11
ed_CI_Lower	1	1	1	0.86	0.85	0.86	1	0.28	0.48	0.37	0.24	-0.11
ed_CI_Upper	1	1	1	0.86	0.86	0.86	1	0.27	0.48	0.37	0.24	-0.1
nber_Uninsu	0.86	0.86	0.86	1	1	1	0.89	-0.11	0.46	0.35	0.22	0.19
red_CI_Lowe	0.85	0.85	0.86	1	1	1	0.89	-0.12	0.45	0.35	0.22	0.2
red_CI_Upp	0.86	0.86	0.86	1	1	1	0.89	-0.1	0.46	0.36	0.22	0.19
Population	1	1	1	0.89	0.89	0.89	1	0.23	0.48	0.38	0.24	-0.07
Insured_rate	0.28	0.28	0.27	-0.11	-0.12	-0.1	0.23	1	0.18	0.16	0.12	-0.69
er_Payment	0.48	0.48	0.48	0.46	0.45	0.46	0.48	0.18	1	0.97	0.88	-0.11
Ave_Paymen	0.37	0.37	0.37	0.35	0.35	0.36	0.38	0.16	0.97	1	0.97	-0.12
her_Payment	0.24	0.24	0.24	0.22	0.22	0.22	0.24	0.12	0.88	0.97	1	-0.13
Regions	-0.11	-0.11	-0.1	0.19	0.2	0.19	-0.07	-0.69	-0.11	-0.12	-0.13	1

Table 5 . Correlation values of each variable

Here we draw the correlation table between all of the variables and find out that the variables within Group1: the 'Number_Insured', 'NInsured_CI_LowerBound', 'NInsured_CI_UpperBound' and Group 2: 'Number_Uninsured', 'NUninsured_CI_LowerBound', 'NUninsured_CI_UpperBound' and Group 3: 'Lower_Payment_Est', 'Ave_Payment' and 'Higher_Payment_Est' are strongly correlated (correlation coefficient are all very closed to 1.) So we can eliminate the upper bound and the lower bound in these three groups and only use the average value to conduct further analysis.

Also, from the histogram for all of the variables in our data, we can see that the within-group trends for three groups we mentioned above are homogenous so this is another evidence that we can only use one of them for further study.

names	Insured_rate	Ave_Payment	Population	Regions
Ave_Payment	0.16	1	0.38	-0.12
Population	0.23	0.38	1	-0.07
Regions	-0.69	-0.12	-0.07	1
Insured_rate	1	0.16	0.23	-0.69

Table 6. Correlation values of each variable

We extract 4 critical variables from our data further more. They are: 'Ave_Payment', 'Population', 'Regions', 'Insured_rate'. Here we use 'insured_rate' because different counties have different population, so here solely use the number of insured people in the counties could be really biased by the total population of the counties. Instead, we use 'insured_rate' which is just the rate of the number of insured people over the total population of the counties.

We can see from the above table:

The correlation coefficient between 'Instured_rate' and 'Ave_Payment' is: 0.16.

The correlation coefficient between 'Instured_rate' and 'Population' is: 0.23.

The correlation coefficient between ‘Instured_rate’ and ‘Regions’ is: -0.69.

The ‘Insured_rate’ is slightly correlated with 'Ave_Payment', 'Population', 'Regions'.

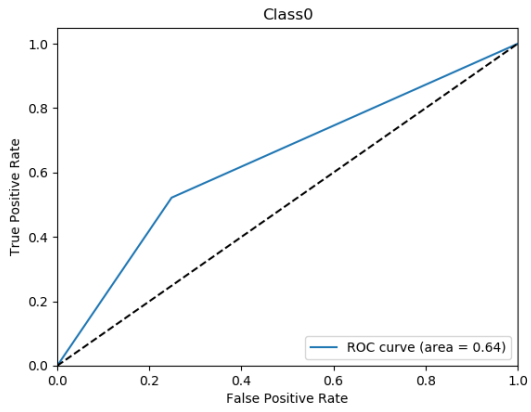
We can have two results here. First, the insured rate is strongly correlated with regions. Second, the insured rate is the slightly correlated with Population (corr = 0.23) and Ave_Payment (corr = 0.16).

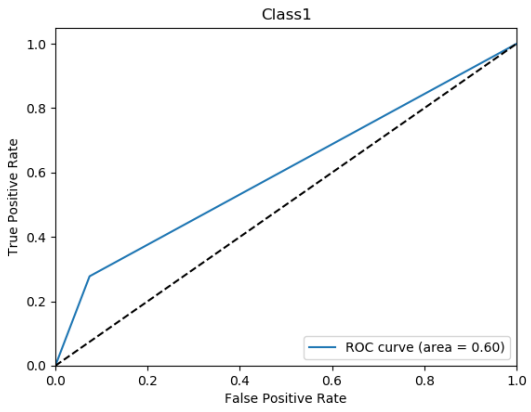
To be noted is that, here we encoded 5 geographic regions in the States as numbers from 0 to 4 and chose the ordered permutation of these 5 numbers that makes the correlation coefficient between ‘Regions’ and ‘Insured_Rate’ the largest as the order of region labels. That is, we labeled the 5 regions in the States in a manner that their labels correlate with ‘Insured_Rate’ is the strongest. Although the meaning of the number itself is meaningless, this method does guarantee us to locate a trend when insured_rate increases, the regions switch in certain directions in an obvious way (correlation coefficient = -0.69).

3.3. Machine learning Analysis

From above results, we known that the insured rate is correlated to the regions, population, and medical expenditure. However the correlation is not big enough for us to be sure. And in order to further confirm it, we performed the machine learning analysis to our data.

First, we check the relationship between the insured rate and geographical position. Using decision tree method, we obtain the figures below. It is obviously that all these five classes, the Midwest, Northeast, Southeast, Southwest and West are all classified better than random guess, the straight line. Then checked with the prediction accuracy score, we can further confirm the result. Because there are 5 regions, if we randomly guess, the probability of guessing correctly is 0.2. However the actual accuracy score using Decision tree algorithm is 0.399, which is twice the probability of randomly guessing. It means that in our data, there exist a certain pattern for insured rate and regions so that using machine learning can accurately predict the result. And the insured rate actually correlated to regions.





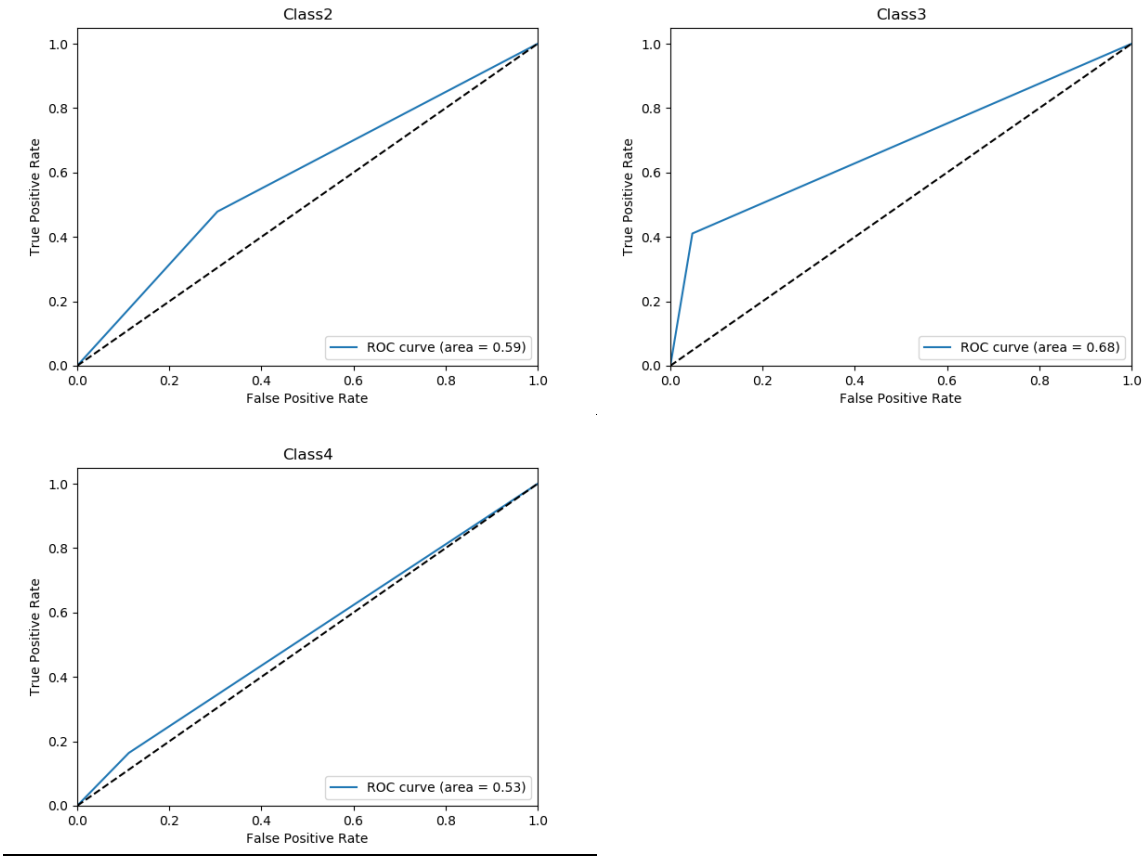


Figure 5. The ROC curve of classifier performance for classifying insured rate based on five regions.

Then we study whether the insured rate has a relationship with medical expenditure. From the KNN classifying result, we can find out that the accuracy score is about 0.26, and it is 1.3 times of random guess, which is 0.2. And it is also can be verified by a Decision Tree model with accuracy score 0.264. Also, the below figures can better explicit this result. All these five ROC curves are mildly above the random guess line. This indicates that the correlation between insured rate and population although are weak but still exist.

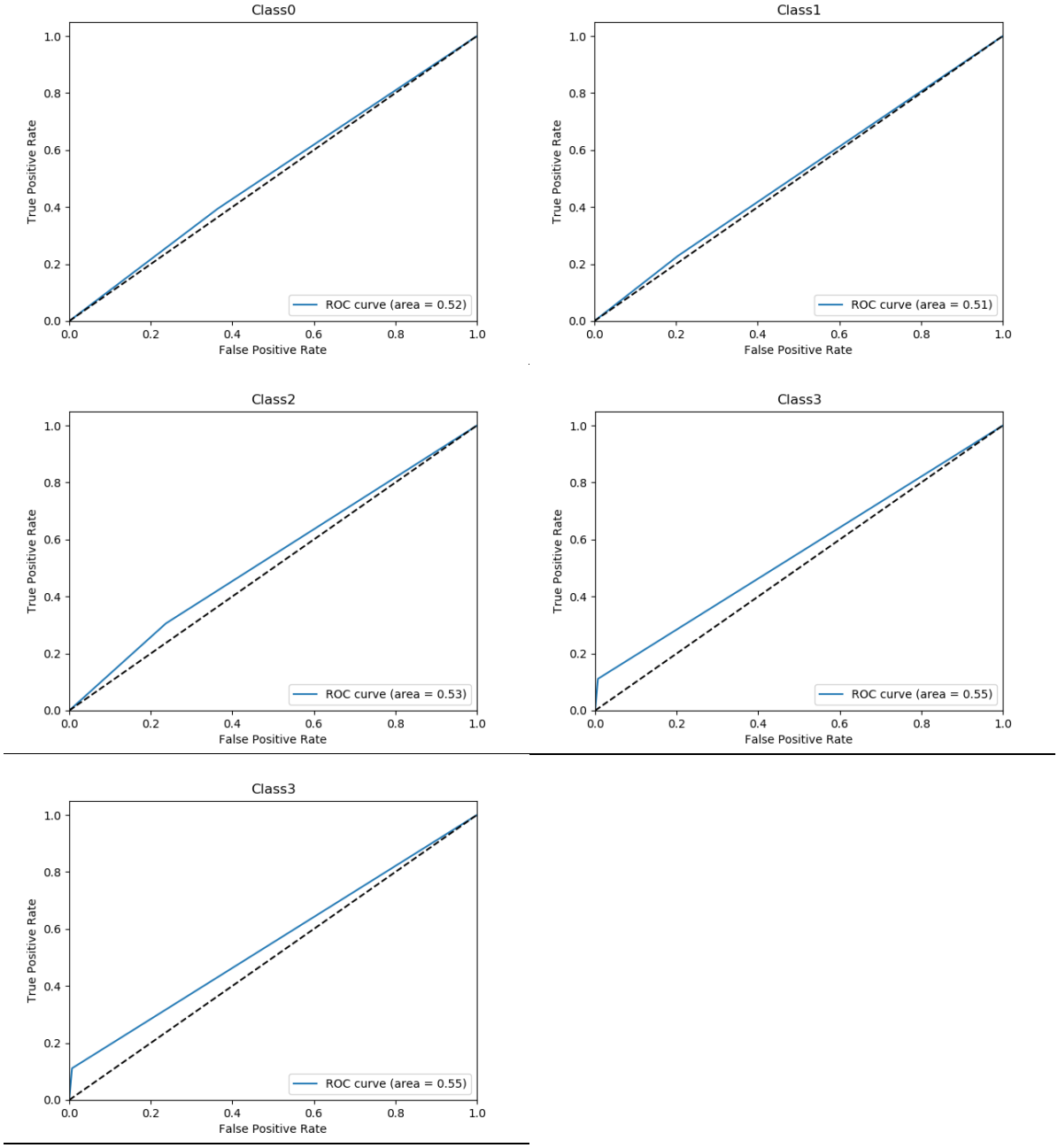


Figure 6. The ROC curve of classifier performance for classifying insured rate based on population.

Finally, we want to find out whether insured rate is related to the average medical expenditure in a certain area. And in order to do that we use a machine learning method to predict the insured rate level based on medical cost. The accuracy score is about 0.25 with SVM model, and 0.3 with model Naïve Bayes. This score seems to be fine, compared with 0.2. However, from the ROC curve below, we can see that the independent classifier performance are not well. Then considering the above two results, we can only say that there may be a slightly correlation of insured rate and average medical cost. And in order to verify this finding, we need to perform more analysis.

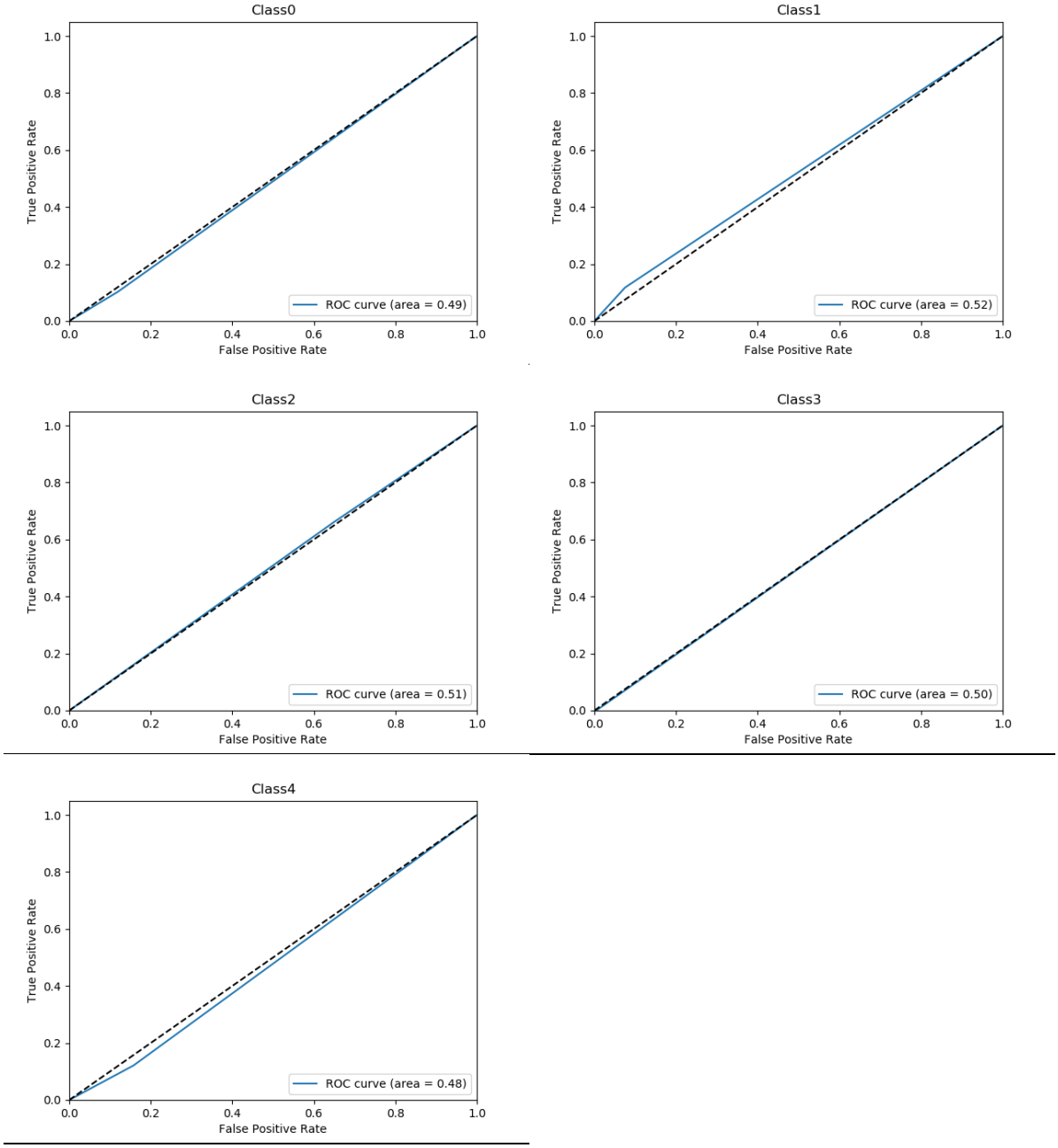


Figure 7. The ROC curve of classifier performance for classifying insured rate based on medical expenditure.

3.4. Association Rule Analysis

After the above exploratory analyses, we conjecture that insured rate should have relationship with other three factors to certain level. In order to confirm our assumption, we conducted the association rule analysis. Association Rule Analysis (ARA) aims to discover strong association rules in the datasets for identifying high frequent patterns within a set of attributes. Considering the three factors the study is focusing on, the subset of 4 categorical attributes 'Insured_rate_EqWidth', 'Regions', 'Insured_rate_EqWidth', 'Ave_Payment_EqWidth' become our most representative attributes to further investigate the inner connection between the insured rate, region, population and major health treatment costs.

By setting a minimum support of 0.2, we found out several top frequent itemset as follows:

X_Y	Support
{'Cheap'}	0.310073
{'Fair'}	0.270724
{'Low'}	0.265477
{'Moderate'}	0.278594
{'Very_Cheap'}	0.33893
{'Very_Low'}	0.262329
{'Very_Small'}	0.66999
{'midwest'}	0.368835
{'southeast'}	0.339979
{'Very_Small', 'Cheap'}	0.20724
{'Very_Small', 'Very_Cheap'}	0.29958
{'midwest', 'Very_Small'}	0.273872

Table 7. Frequent itemset with minimum support of 0.2.

Figure 8. Frequent itemset with minimum support of 0.2.

In table 7 above, both ‘Very_Low’ and ‘Low’ stood out among those frequent itemset in the whole dataset with support of 0.26 and 0.27, which means the majority of counties have a below average insured rate. In order to dig into the inner patterns regarding insured rate, we lowered the support threshold to 0.001, then set the minimum threshold of confidence as 0.8, and respectively filtered the addition item, which is the ‘Y’ part in the association rule, as ‘High’ or ‘Very_High’. We got some interesting rules:

X_Y	X	Y	Support	Confidence
{'northeast', 'High', 'Medium', 'Cheap'}	{'northeast', 'Medium', 'Cheap'}	{'High'}	0.002623	0.833333
{'northeast', 'High', 'Very_Large', 'Cheap'}	{'northeast', 'Very_Large', 'Cheap'}	{'High'}	0.001049	1
{'northeast', 'High', 'Expensive', 'Very_Large'}	{'northeast', 'Expensive', 'Very_Large'}	{'High'}	0.001049	1
{'midwest', 'High', 'Medium', 'Very_Cheap'}	{'midwest', 'Medium', 'Very_Cheap'}	{'High'}	0.001049	1
{'Moderate', 'High', 'Very_Large', 'midwest'}	{'Moderate', 'Very_Large', 'midwest'}	{'High'}	0.004197	0.8

Table 8. Frequent itemset with fixed ‘High’ item.

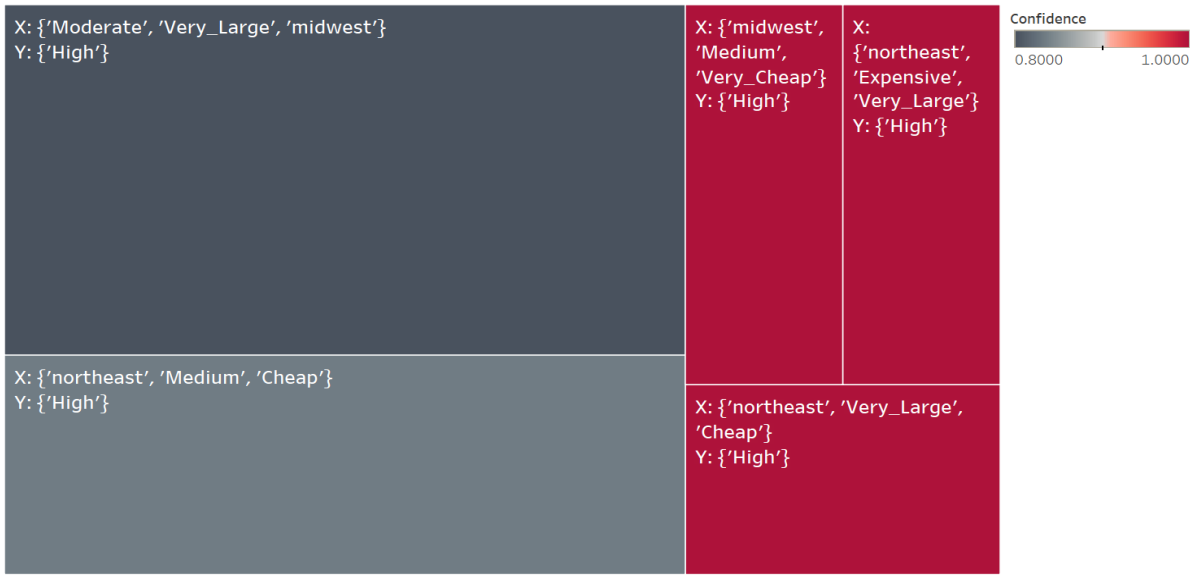


Figure 9. Frequent itemset with fixed ‘High’ item, colored by confidence.

Although all of them have very small support level (under 0.01), but we acknowledged that a county in Northeast region with very large population is definitely having a high insured rate, regardless of their level of healthcare expenditure. Besides, a county in Midwest region, with median population and very cheap health treatment expenses are surely having a high insured rate.

On the contrary, we filter the addition item as ‘Very_Low’ and ‘Low’ with the same support and confidence minimums.

X_Y	X	Y	Support	Confidence
{'southwest', 'Very_Low'}	{'southwest'}	{'Very_Low'}	0.094439	0.891089
{'Very_Low', 'southwest', 'Cheap'}	{'southwest', 'Cheap'}	{'Very_Low'}	0.026233	0.909091
{'Expensive', 'southwest', 'Very_Low'}	{'Expensive', 'southwest'}	{'Very_Low'}	0.008395	1
{'Moderate', 'southwest', 'Very_Low'}	{'Moderate', 'southwest'}	{'Very_Low'}	0.017838	0.871795
{'Small', 'southwest', 'Very_Low'}	{'Small', 'southwest'}	{'Very_Low'}	0.016789	0.864865
{'Very_Low', 'southwest', 'Very_Cheap'}	{'southwest', 'Very_Cheap'}	{'Very_Low'}	0.040923	0.866667
{'Very_Expensive', 'southwest', 'Very_Low'}	{'Very_Expensive', 'southwest'}	{'Very_Low'}	0.001049	1
{'Very_Small', 'southwest', 'Very_Low'}	{'Very_Small', 'southwest'}	{'Very_Low'}	0.070829	0.918367
{'Large', 'southeast', 'Low', 'Cheap'}	{'Large', 'southeast', 'Cheap'}	{'Low'}	0.003148	0.857143
{'Small', 'Very_Low', 'southwest', 'Cheap'}	{'Small', 'southwest', 'Cheap'}	{'Very_Low'}	0.004722	0.818182
{'Very_Small', 'Very_Low', 'southwest', 'Cheap'}	{'Very_Small', 'southwest', 'Cheap'}	{'Very_Low'}	0.020462	0.975
{'Expensive', 'southwest', 'Medium', 'Very_Low'}	{'Expensive', 'southwest', 'Medium'}	{'Very_Low'}	0.001049	1
{'Small', 'Expensive', 'southwest', 'Very_Low'}	{'Small', 'Expensive', 'southwest'}	{'Very_Low'}	0.002623	1
{'Very_Small', 'Expensive', 'southwest', 'Very_Low'}	{'Very_Small', 'Expensive', 'southwest'}	{'Very_Low'}	0.004197	1
{'Moderate', 'Small', 'southwest', 'Very_Low'}	{'Moderate', 'Small', 'southwest'}	{'Very_Low'}	0.005247	1

{'Moderate', 'Very_Small', 'southwest', 'Very_Low'}	{'Moderate', 'Very_Small', 'southwest'}	{'Very_Low'}	0.010493	0.869565
{'Very_Small', 'Very_Low', 'southwest', 'Very_Cheap'}	{'Very_Small', 'southwest', 'Very_Cheap'}	{'Very_Low'}	0.035677	0.894737

Table 9. Frequent itemset with fixed 'Low' and 'Very_Low' items.

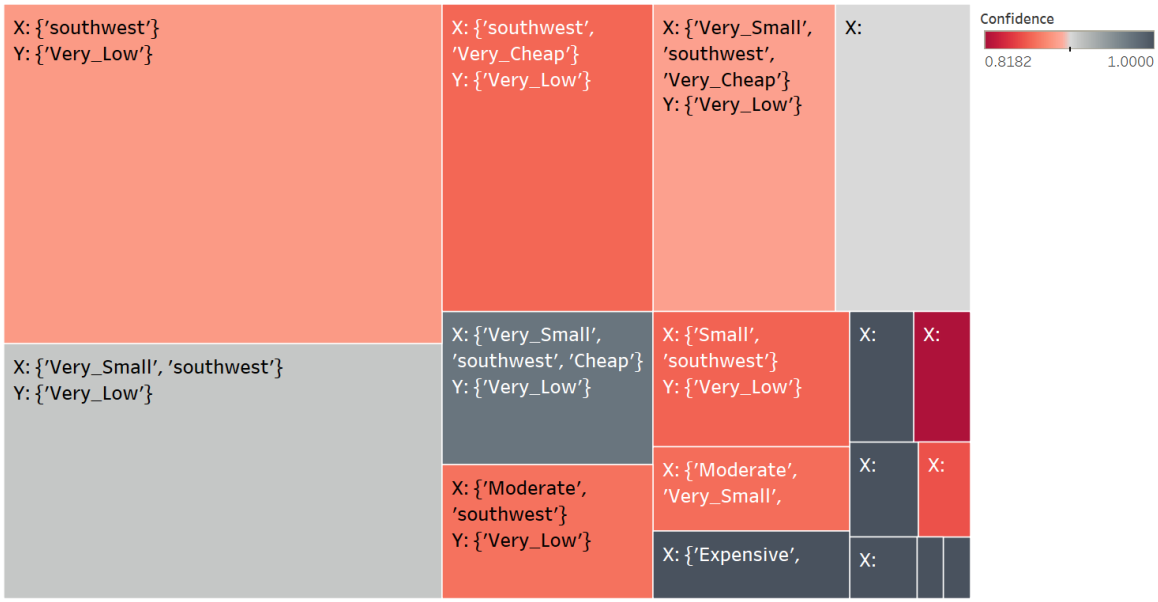


Figure 10. Frequent itemset with fixed 'Low' and 'Very_Low' items, colored by confidence.

By observing table above, the association rule {'Very_Small', 'southwest'}->{'Very_Low'}' is the most frequent pattern compared to others. And a county in Southwest region with expensive and very expensive healthcare costs are guaranteed to have a very low insured rate. And {'Large', 'southeast', 'Cheap'}->{'Low'} is the only one here in the southeast region and the only one with a large population size that have a low insured rate. Based on that, we informed an idea that Southwest Region more likely to have a lower insured rate.

3.5. Network Analysis

Association rule analysis shows that item 'High', 'Low' and 'Very-Low' which represent high, low, and very low insured rate, seems to be relative to several certain itemset. Then we assume that our data should contains several patterns. Therefore, we perform a network analysis to search for these potential patterns.

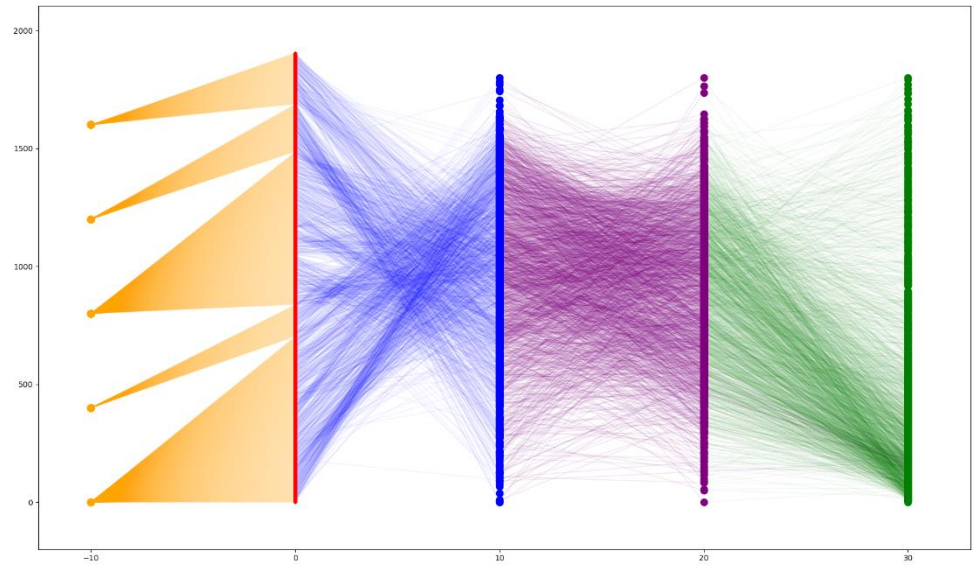


Figure 11. Network of five regions, counties, insured rate, medical expenditure, and population.

In the above figure, the orange nodes represent five regions, from the bottom up, the Midwest, Northeast, Southeast, Southwest and West. While the red, blue, purple, and green nodes represent 1757 counties, insured rate, average medical expenditure and population in every county, respectively.

We perform a network analysis to study the relationship between the county and the insured rate, medical cost, population and region, finding certain patterns of these variables. Therefore, we draw a network with fixed nodes position and edges as the figure below. Because the edges are artificially designed as one county to one region and one county to one insured rate to one medical expenditure to one population, the degree of nodes in the middle three rows and green nodes are 2 and 1, respectively. Form the figure, we can find out that it is a very sparse network, the betweeness and clustering coefficient of each nodes are all nearly 0. For the same reason, the global network metrics, density, average centrality and triangles of this network are all very small, almost 0. Also, as we designed the network as a very sparse one, it cannot be clustered into several clearly separated clusters with high modularity. And the clustering result confirms this.

As shown above, there is an obvious flow from Midwest to relatively high insured rate to moderate medical expenditure to low population. It seems that people in Midwest and Northeast region are more willing to buy health insurance than other regions, and regions with high insured rate shows a tendency of having moderate medical expenditure and low population. These result coincides with the one from the association rule.

It seems that the insurance company should consider relocating more investment in counties in Midwest and Northeast regions with relatively low population and moderate or low medical costs. Because in generally, these areas have larger probability of high insured rate. If one county which fits these characters has relatively low insured rate, then it is a potential business objective for the insurance company to increase the insured number there.

3.6. Clustering Analysis

From the above two parts, we find out two patterns in our data. Then in order to further check this result, we also perform a clustering analysis to see if there are two clusters in our data. After we properly pre-process the data (categorical variable encoding and normalization), we are able to conduct Kmeans Clustering, Hierarchical Clustering and DBSCAN clustering analysis. Due to the strong correlation with some groups of variables,we only have to choose Number_Insured, Numeber_Uninsured and Ave_Payment to represent their lower and upper bound. The features we are using to perform clustering analysis are: 'Number_Insured','Number_Uninsured','Population','Insured_rate','Ave_Payment','Regions'.

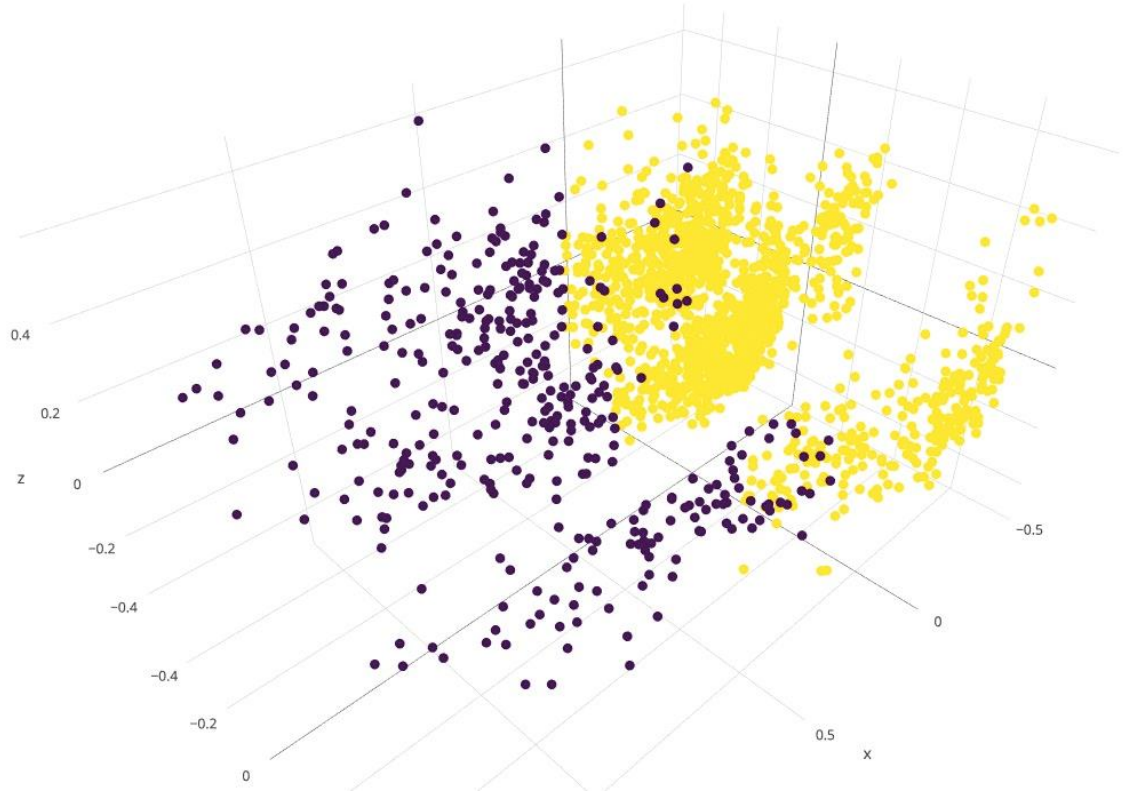


Figure 12. PCA for KNN clustering.

By using Hierarchical Clustering, we get 2 clusters because from the PCA plot we can see that vaguely there are two clusters for this dataset. We use silhouette score to measure the accuracy of the clustering result. The silhouette score is 0.342. So we can see that there are relatively two clusters for our data. This implies there vaguely two trends to be discovered from our further analysis.

4. Discussion

After conducting all kinds of analyses, we confirm that the insured rate actually related to regions, population and average medical cost to some degree. And it seems to exist some patterns for the combinations of these four variables.

First, from the correlation analysis, we find out that the correlation between the insured rate and the region is 0.56, which is strong. While the correlation between the insured rate and population is slightly weak, only 0.22. And for the medical cost, the correlation score is only 0.14, which is very weak. And these results can also be seen in the machine learning analysis. Therefore, it is safe to say that insured rate will be influenced greatly by regions, and slightly affected by population. Then we want to find out the certain patterns between these variables.

The network analysis shows very clear two trends. First, Midwest area plus moderate or cheap medical expenditure and small population usually associated with high insured rate. Then, Southwest and West regions plus high population and low medical payment always associated with low insured rate. In the meanwhile, the association rule analysis represents some frequent item sets consistent with above findings. First, with fixed high insured rate, cheap expenditure, medium population and Midwest or Northeast shows up with high frequency. Second, cheap payment, small population, and Southwest shows frequently with low insured rate. Therefore, we have a strong conjecture that there should be two clearly clusters for this dataset. With further analysis, our clustering analysis represents two clearly clusters, further proving our conclusions.

5. Conclusions

- There are two clear trends for the insurance rate buying behavior in the States:
  - High insurance rate is usually associated with moderate or cheap medical expenditure, small population and Midwest region of the United States;
  - Low insurance rate is usually associated with low medical expenditure and high population and Southwest region of the States.
- Data limitation: There is no distinctly separated clusters in the data. Sample points in some categories are evenly distributed so when doing machine-learning training, the result only converge to one of the category that contains most sample points.

6. References:

i O'Hara, Brett. "Do medical out-of-pocket expenses thrust families into poverty?" Journal of Health Care for the Poor and underserved 15.1 (2004): 63-75.

ii O'Hara, Brett. "Do medical out-of-pocket expenses thrust families into poverty?" Journal of Health Care for the Poor and underserved 15.1 (2004): 63-75

iii Zeldin, C., and M. Rukavina. "Borrowing to stay healthy: how credit card debt is related to medical expenses. The Access Project and Demos, 2007." (2008).

iv Zeldin, C., and M. Rukavina. "Borrowing to stay healthy: how credit card debt is related to medical expenses. The Access Project and Demos, 2007." (2008).

v Himmelstein, David U., et al. "Medical bankruptcy in the United States, 2007: results of a national study." The American journal of medicine 122.8 (2009): 741-746.

vi Kim, Hyungsoo, Wonah Yoon, and Karen A. Zurlo. "Health shocks, out-of-pocket medical expenses and consumer debt among middle-aged and older Americans." Journal of Consumer Affairs46.3 (2012): 357-380.

vii Tang, T. L. (1993). The meaning of money: Extension and exploration of the money ethic scale in a sample of university students in Taiwan. Journal of Organizational Behavior, 14(1), 93–99. doi:10.1002/job.4030140109

viii Jorgensen, B. L., Foster, D., Jensen, J. F., & Vieira, E. (2016). Financial Attitudes and Responsible Spending Behavior of Emerging Adults: Does Geographic Location Matter? Journal of Family and Economic Issues, 38(1), 70-83. doi:10.1007/s10834-016-9512-5

ix Himmelstein, D. U., Warren, E., Thorne, D., & Woolhandler, S. (2005). Illness and injury as contributors to bankruptcy. Health Affairs (Millwood). doi:10.1377/hlthaff.w5.63.

x Institute of Medicine. (2002). Unequal treatment: Confronting racial and ethnic disparities in health care. Washington, DC: The National Academies of Sciences.

xi Kasper, J. D., Giovannini, T. A., & Hoffman, C. (2000). Gaining and losing health insurance: strengthening the evidence for effects of access to care and health outcomes. Medical Care Research and Review, 57, 298–318.

xii U.S. Census Bureau, 2015 1-Year American Community Survey. https://www.census.gov/content/dam/Census/library/publications/2016/demo/p60-257.pdf

xiii U.S. Census Bureau, 2014 and 2015 Small Area Health Insurance Estimates (SAHIE). https://www.census.gov/content/dam/Census/library/publications/2017/demo/p30-01.pdf

xiv Chenavaz, R. and Escobar, O. (2012) Effective area as a measure of land factor, Economics Bulletin, 32, 1962–69.

xv Chenavaz, R. and Escobar, O. (2012) Effective area as a measure of land factor, Economics Bulletin, 32, 1962–69.

xvi Chenavaz, R., & Escobar, O. (2015). Population distribution, effective area and economic growth. Applied Economics, 47(53), 5776-5790. doi:10.1080/00036846.2015.1058907

xvii Chenavaz, R., & Escobar, O. (2015). Population distribution, effective area and economic growth. Applied Economics, 47(53), 5776-5790. doi:10.1080/00036846.2015.1058907

xviii Diamond, P. (1992). Organizing the Health Insurance Market. Econometrica, 60(6), 1233. doi:10.2307/2951520