Knowledge Discovery in Databases
Summer 2016
Project Assignment: Analysis of Health Insurance Marketplace

I. Identification of Team and Members

Name	ID	Email Id
Sonal Kaulkar	800935313	skaulkar@uncc.edu
Devarsh Jhaveri	800927705	djhaveri@uncc.edu

II. Dataset

We have data set of 10.57 GB. 1627370 records were present in BenefitsCostSharing.csv, 12694445 records were present in Rate.csv and 77353 records were present in PlanAttributes.csv. Thus we worked with huge amount of data. The data set is also divided on the basis of year and the data of each year is given.

a. Description and Purpose

The data which is used is from the Centers for Medicare & Medicaid Services (CMS).

In the BenefitsCostSharing csv file each record relates to the coverage of a single benefit by one issuer's insurance plan. This data contains plan-level data on essential health benefits, coverage limits, and cost sharing for each QHP and SADP.

In the Rate csv file - This csv describes the variables contained in the Rate-PUF. Each record relates to one issuer's rates based on plan, geographic rating area, and subscriber eligibility requirements. The RatePUF is available for plan year 2014, plan year 2015, and plan year 2016.

In the BenefitsCostSharing csv file - This csv describes the variables contained in the BenCS-PUF. Each record relates to the coverage of a single benefit by one issuer's insurance plan. The BenCS-PUF is available for plan year 2014 and plan year 2015

b. Source

https://www.kaggle.com/hhsgov/health-insurance-marketplace

c. Other work that has been completed (document anything you use from Kaggle, etc.)

III. Business Research/Understanding:

a. What is the domain?

We have worked on The Health Insurance domain.

b. What do we need to know about the domain?

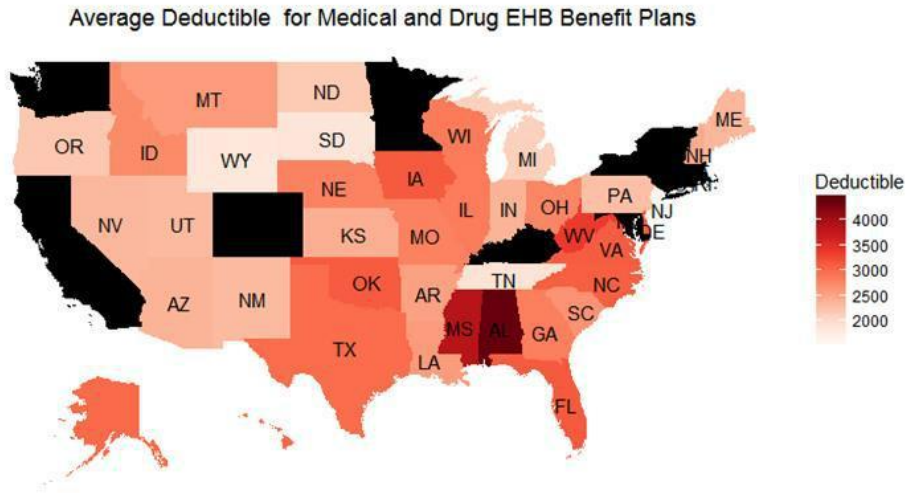
i. Three scholarly references (journal article, text)

https://www.healthcare.gov/glossary/ - The glossary has a list of health insurance present.

https://www.cms.gov/CCIIO/Resources/Data-Resources/Downloads/2-General_Information_Factsheet-05032016_draft.pdf - This document outlines important information about the Health Insurance Marketplace Public Use Files (Marketplace PUF), including source data, file size, variables, key assumptions, analytic utility, and support information. A data dictionary is also available for each of the separate files within the Marketplace PUF.

http://www.ixshealth.com/ - In this website we got to know what kind of insurance are present in the market and who can use which health insurance plans.

ii. Focus on information that is relevant to your:
    1. EDA
        1. Exploration of Deductible Amount for Medical and Drug EHB Benefit Plan R: Using choropleth maps
        Output:



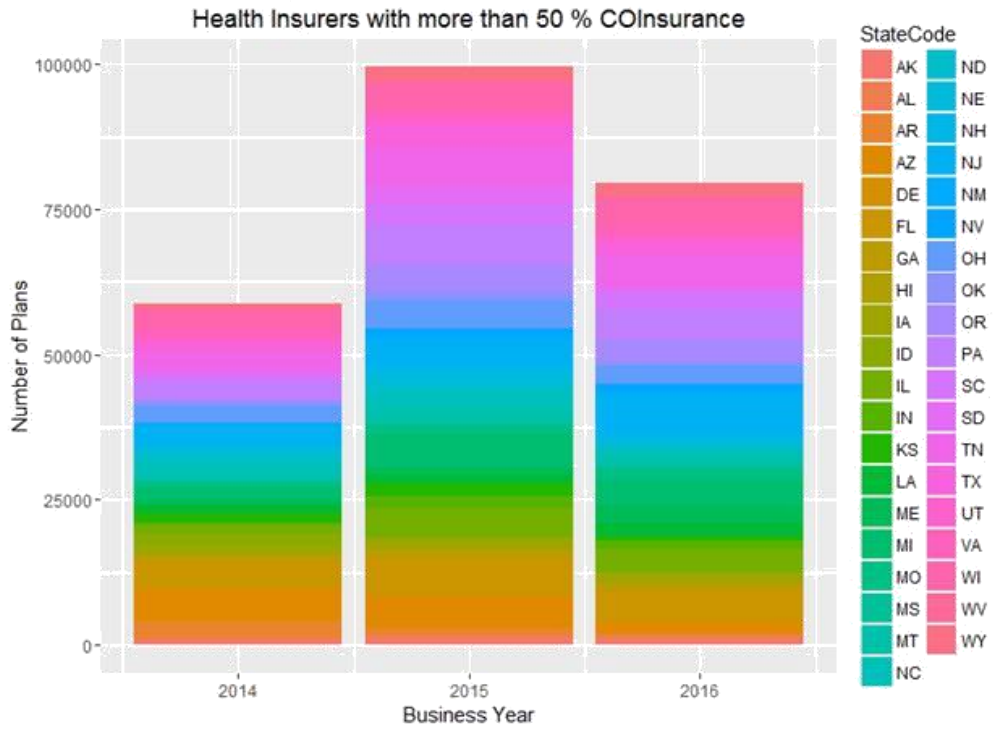
Analysis: The above choropleth map shows the average deductible of Medical plans state wise. Darker the shade more is the mean deductible of the state and vice versa. Thus on exploration we can see that States like Alabama, Mississippi have a greater mean of Deductible than Tennessee, Wyoming, South Dakota.

1. Exploration of Number of plans having a Coinsurance of more than 50% or so

Number of plans with co-insurance greater than 50% have been analyzed state wise as well as year-wise

R: Using bar chart with qplot function of ggplot package

Output:

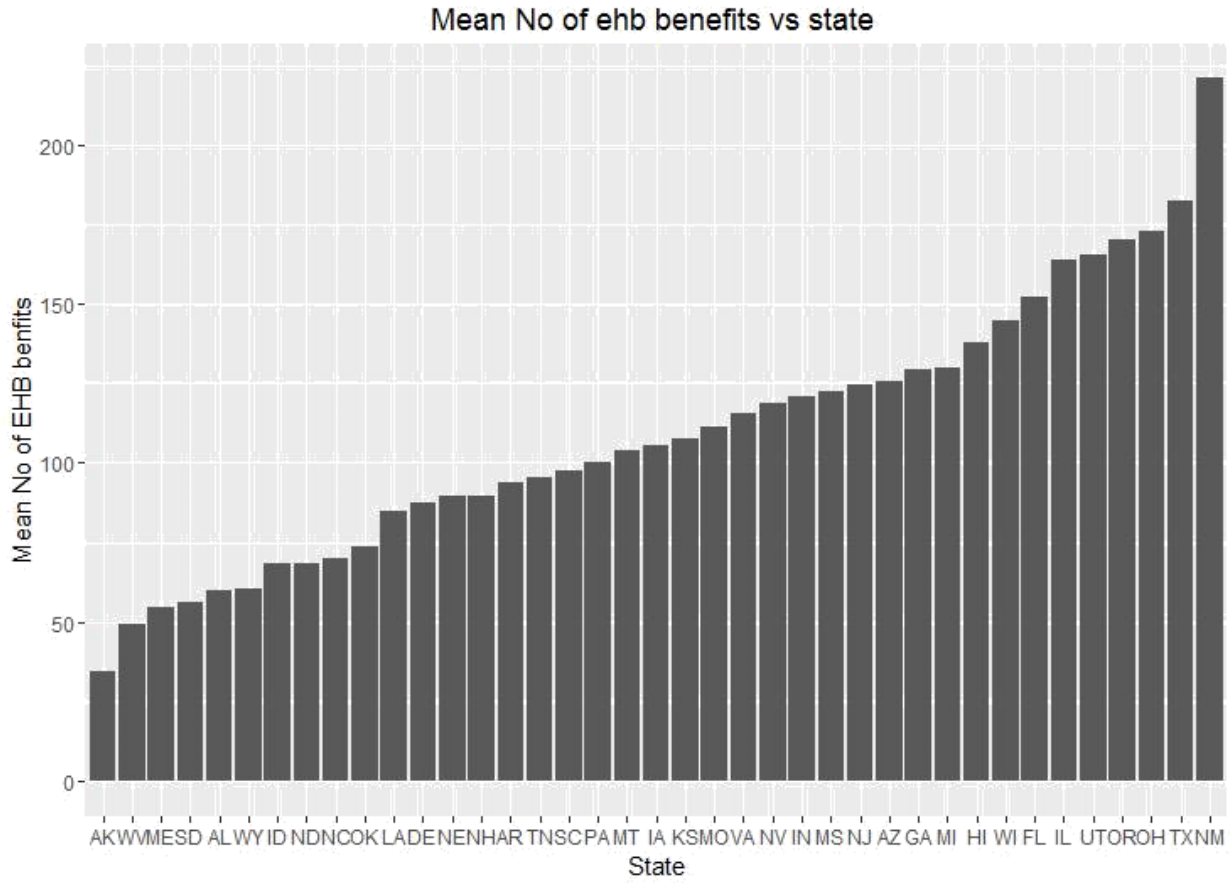


Analysis: The X-axis is the Business Year and the Y- axis denotes the Number of plans with coinsurance greater than 50%.There is an increase in the number of plans from 2014 to 2015 with number of plans increasing on an average in most of the states. But there is an overall l decrease in the number of plans from 2015 to 2016.However in states like Nevada, Ohio, Oklahoma the number of plans over the years.

1. Exploration of Number of plans which more than 10 EHB benefits.

With Obama care act it was mandatory to include for insurance plans to include at least 10 EHB benefits .Following is the pareTO to chart for C mean of the number of benefits provided by plans.

R: ggplot from ggplot package
Output:



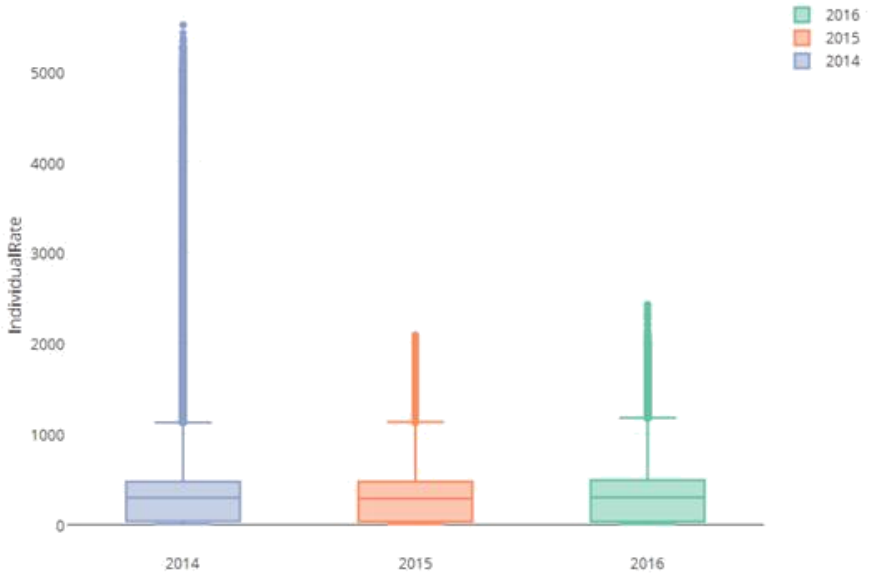
Analysis: As shown in the pareto chart, New Mexico has the highest number of EHB benefit plans with Alaska having the least.

1. Analysis of Rate versus the Tobacco Preference of plans.

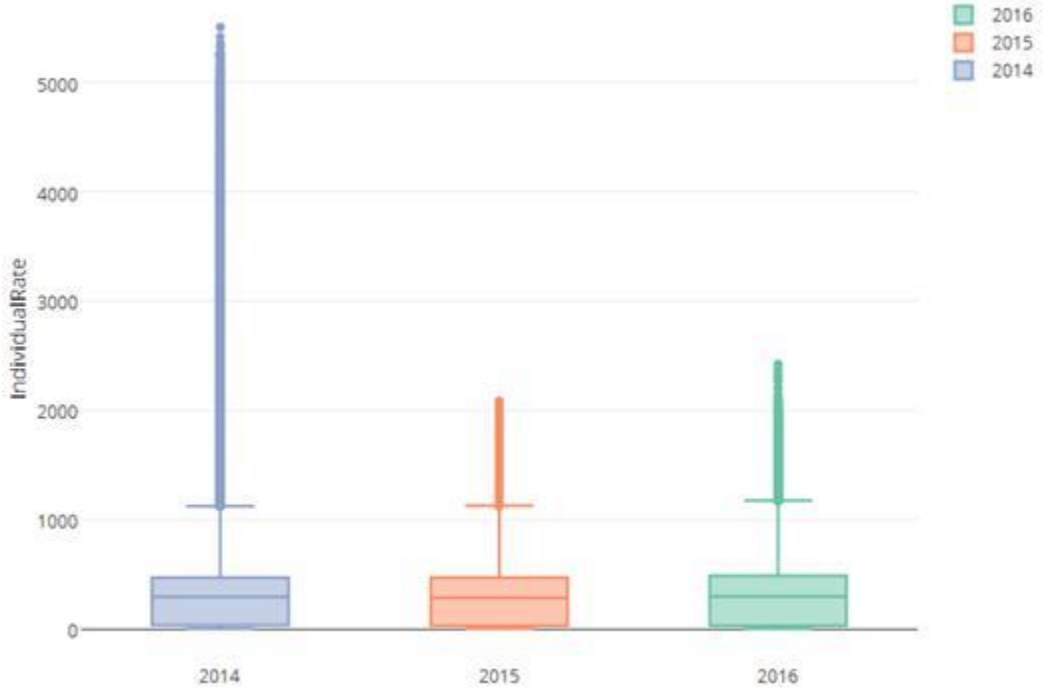
   R: Using Boxplot of plotly package

   Output:

Boxplot of Individual rate for non-tobacco preferred plans



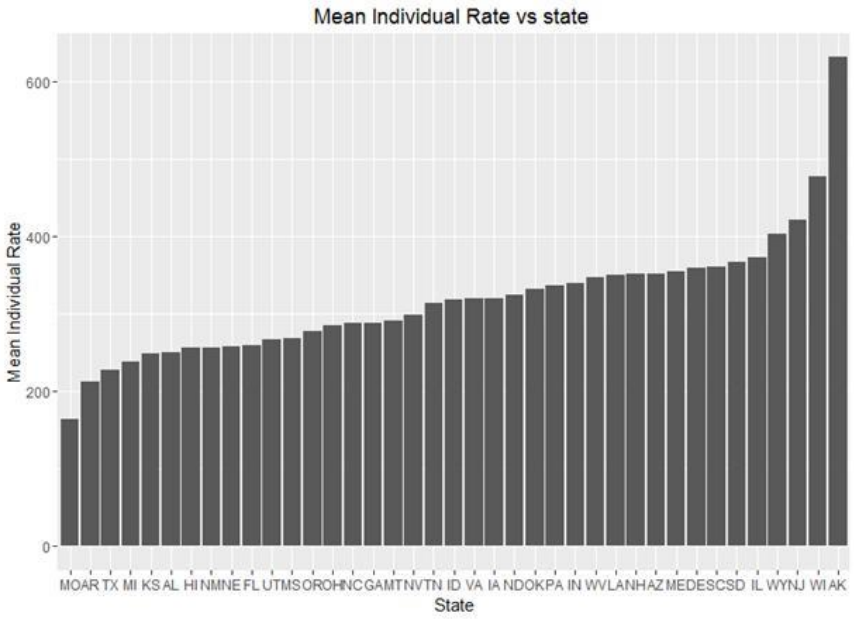
Boxplot of Individual rate for tobacco preferred plans



Analysis: Comparing the above two boxplots we can conclude that there the plan rates are not affected by the factor of tobacco preference in them.

1. Exploration of Individual Plan rates across states.

R: ggplot of plotly package
Output:



Analysis: Shown above is a pareto chart of mean individual rate state wise resulting into

•Alaska has the highest mean of individual premium rates for insurance rates.

•Missouri has the least mean of individual premium rates for insurance rates.

2. Hypothesis Testing

In the hypothesis testing we have imported the Rate.csv file and made subset of the dataset into different types based on the year. As there are many individual rate values which has value of 9999 so we considered they are a value given to missing data. Thus we have filtered the data and considered those values which are less than 6000.

We have made a data frame with attributes BusinessYear, StandardComponentId and DentalOnlyPlan. The we have filtered the data which has no dental plan. Then we have merged the data on the basis of year (2014, 2015 and 2016). We calculated the means of the rate for these years. For the year 2014 the mean of rate is 427.2714, 2015 the mean is 448.8524 and for 2016 the mean is 479.8148.

```
> mean.test2014$statistic
        t
-402.3906
> mean.test2014$p.value
[1] 0
> mean.test2014$conf.int
[1] 427.1460 427.3969
attr(,"conf.level")
[1] 0.95
```

```
> mean.test2015 <- t.test(x= ratemedicalplans2015 $Rate,
+                         mu=ratesallyearmean,
+                         conf.level= 0.95)
> mean.test2015$statistic
        t
-78.62506
> mean.test2015$p.value
[1] 0
> mean.test2015$conf.int
[1] 448.7484 448.9565
attr(,"conf.level")
[1] 0.95
> mean.test2016 <- t.test(x= ratemedicalplans2016 $Rate,
+                         mu=ratesallyearmean,
+                         conf.level= 0.95)
> mean.test2016$statistic
       t
444.1553
>
> mean.test2016$p.value
[1] 0
> mean.test2016$conf.int
[1] 479.6966 479.9330
attr(,"conf.level")
[1] 0.95
```

Hypothesis t-test was performed to compare the means of the Individual Rate of samples with the population of all the three years 2014-16 Results: Small p values Thus the sampling based on year is not representative of the population.

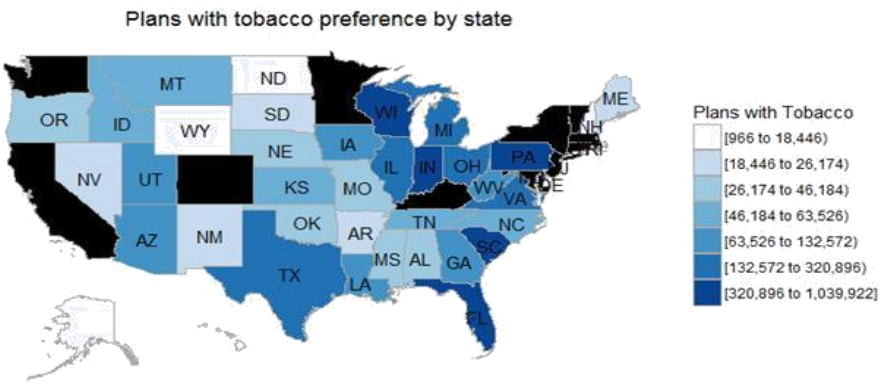
iii. Describe what you want to discover 1. Hypothesis

2. EDA

    1. Exploration of states with the highest number of plans that consider tobacco preference in their individual rates.

R: choropleth maps

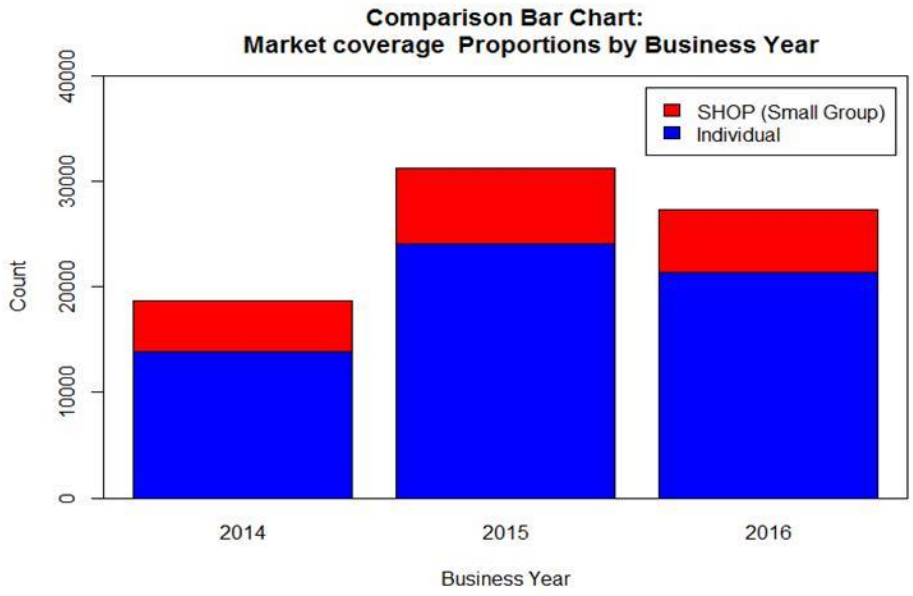
Output:



Analysis: States which are black on the map represent the ones with missing values for tobacco preference.
Lighter the shade of the state less is the number of plans with tobacco preference and vice versa. Thus Wyoming has the least number of plans with Wisconsin and Pennsylvania and Florida and South Carolina with more number of plans with tobacco preference.

1. Exploration of Market coverage year wise: To find out whether with time does the number of plans belonging to Individual category is more than that belonging to a Small group of business?

R: Bar graph
Output:



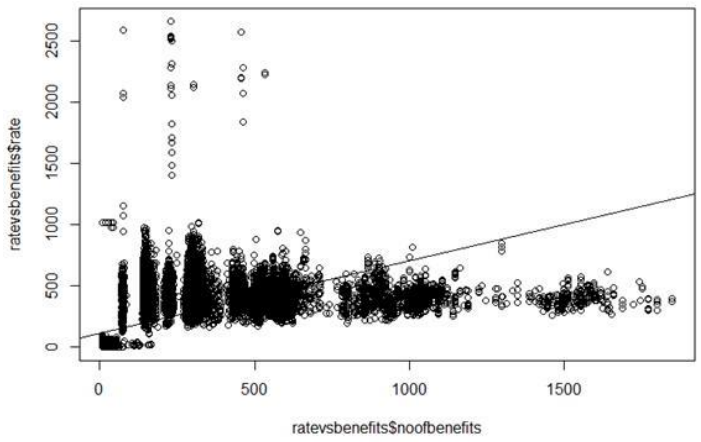
Analysis: The Number of individual plans increases considerably from 2014 to 2015 but the decreases slightly from 2015 to 2016. Similar behavior is demonstrated in the number of plans belonging to the Small group of individuals.

IV. Data Understanding

a. Description

b. Estimation (regression, confidence intervals, correlation)

1. Regression model of Number of benefits vs Individual premium:
   R: Using lm in r
   Output:

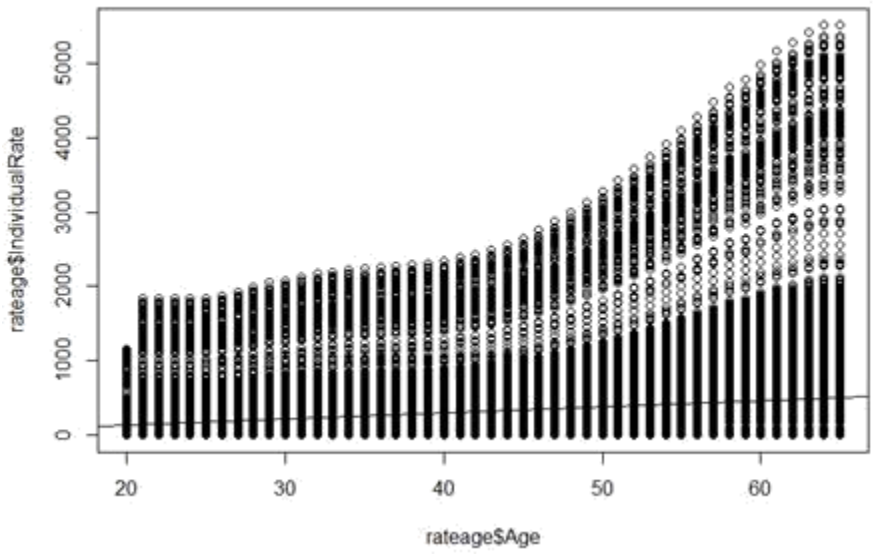


Analysis: For the above plot, rate.csv and BenefitsCostSharing.csv per plan to analyze the change in rate with number of benefits per plan.

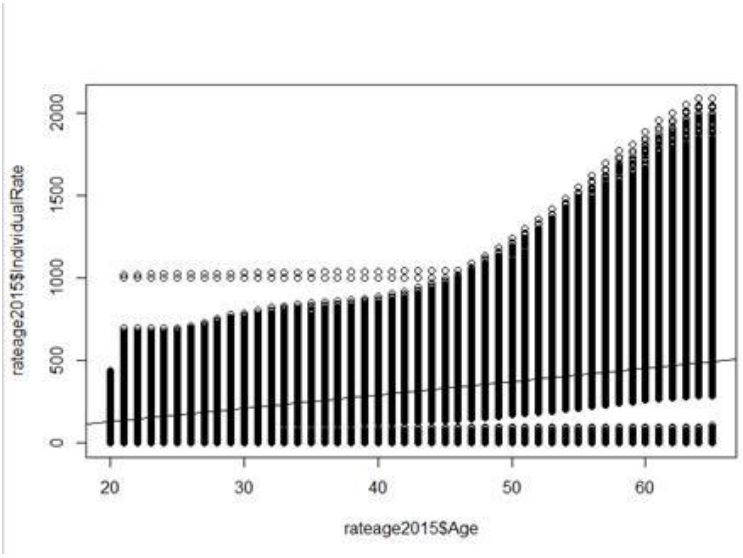
Equation: Individual Premium Rate ~ Number of benefits provided

An abline with positive slope indicates that as number of benefits increases so does the individual premium of the plans.

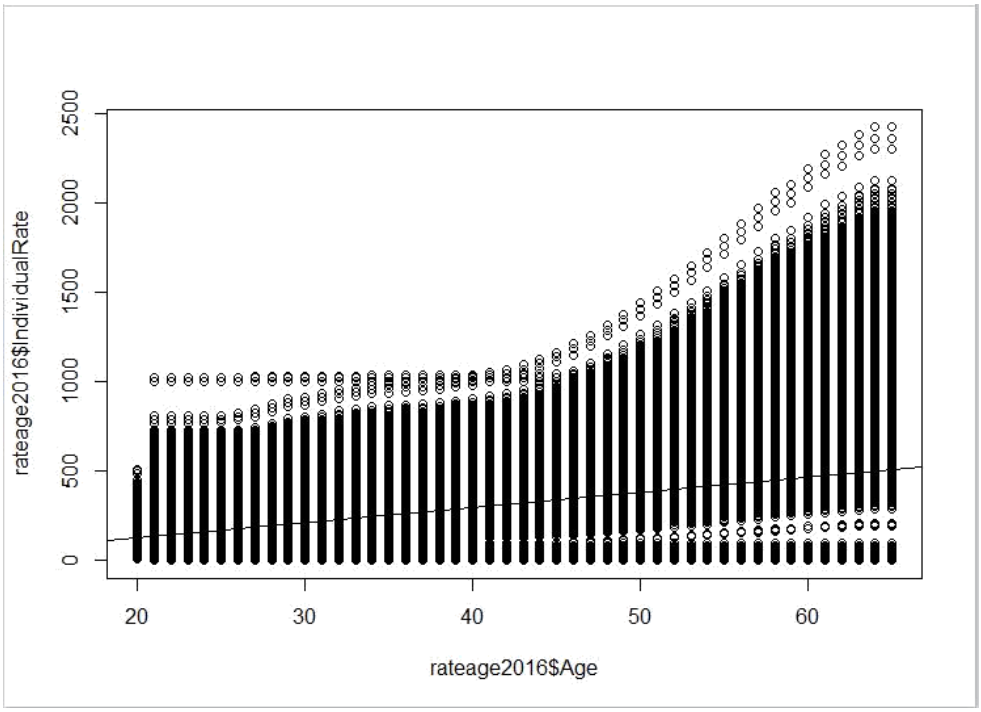
1. Regression model of Individual rate vs age groups done year wise: For business Year 2014:



As seen for business year 2014 as age goes on increasing the increase in the individual rate has a small positive slope. Thus there is no such considerable increase in rate with age
For Business Year 2015:



In the business year of 2015, there slope is higher than the previous year and the rate of change individual premium with respect to age is thus more.
For Business year 2016:



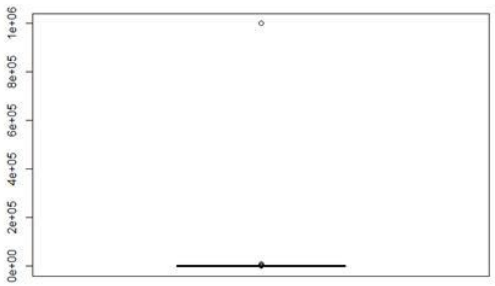
For the business year 2016 there is considerable increase in the rate as age increases. Thus with every year from 2014 to 2016 the Individual rate is dependent on the age of the groups.

V. Data Preparation

a. Cleansing

1. After analysis of data, the rate.csv file contains attribute individual rate with has values of 999999 for missing values. Also the attribute is filled with these values in a huge subset of records. This can be visualized by the boxplot.

Output:



Thus it is required to cleanse the individual rate for values less than 6000.

1. The attribute value TEHBDedInnTier1Individual in PlanAttributes.csv is cleansed as follows:

a. Records with value “Not Applicable” are eliminated as they are not applicable for the analysis.

b. Dollar sign is the values is removed

c. Comma in higher values of deductible are replaced by nulls thus converting it into a pure numeric variable.

1. Cleansing was also performed at the time of clustering to reduce the dataset from 12 million to 6 million by eliminating duplicate records having the same planed, metal level and dentalonlyplan factor variable.

b. Combining with Other Data (find one other dataset to combine)

We have combined the following datasets for analysis:

- Rate.csv and BenefitsCostSharing.csv to compare Number of benefits versus the individual rate in rate.csv
- Rate.csv and PlanAttributes.csv for clustering of dataset for attributes of rate and age
- Rate.csv and PlanAttributes.csv for generation of association rules with predictors being MarketCoverage(PlanAttributes.csv) and MetalLevel (PlanAttributes.csv). The target variable used is OutOfCountryCoverage (PlanAttributes). Association is however performed on the one of the clusters formed using rate and age belonging to Rate .csv

c. Creation of New Variables (at least one)

- Conversion of age from a factor to a numeric variable: Age attribute belonging to Rate.csv is filtered as follows: -
  - Records with "Family Option" are excluded
  - Records with age value "65 and Above" is converted to numeric value 65
  - Records with age value "0 to 20 "are converted into numeric value 20

d. Binning, Discretization (at least one)

Some algorithm prefers categorical rather than continuous predictors, in which case we would need to do partition any numerical predictors into bins or bands. For example, here we have tried to partition the numerical predictor SBCHavingDiabetesCoinsurance into low, medium and high. We have used equal width binning here.

- Binning of SBCHavingDiabetesCoinsurance attribute belonging to PlanAttributes.csv is given by :-
  - Records with $0 \leq X < 1000$ denoted as low.
  - Records with $1000 \leq X < 2000$ denoted as medium
  - Records with $2000 \leq X < 3000$ denoted as high.

e. Normalization of Numeric Variables

Following are the normalized attributes:

- For prediction analysis of rate vs age, we have normalized:
  - IndividualRate variable from the Rate.csv after cleaning it using Z-Score standardization
  - Age variable is also normalized using Z-Score normalization after converting it into a numeric variable as stated above.

VI. Modeling

a. Prediction

1. Estimation of Individual Rate by Age

Year wise:

Business Year 2014:

```
> pred.confidence1 <- predict(lr1,
+                             data.frame(age2014=10) ,
+                             interval = "confidence")
> pred.confidence1
      fit      lwr      upr
1 49.5369 48.79237 50.28142
> pred.prediction1 <- predict(lr1,
+                             data.frame(age2014 = 10),
+                             interval = "prediction")
> pred.prediction1
      fit       lwr      upr
1 49.5369 -494.7605 593.8343
>
```

Business Year 2015:

```
> pred.confidence2 <- predict(lr2,
+                             data.frame(age2015=10) ,
+                             interval = "confidence")
> pred.confidence2
      fit      lwr      upr
1 48.6779 48.01178 49.34401
> pred.prediction2 <- predict(lr2,
+                             data.frame(age2015 = 10),
+                             interval = "prediction")
> pred.prediction2
      fit       lwr      upr
1 48.6779 -494.6095 591.9653
```

Business Year 2016:

```
> pred.confidence3 <- predict(lr3,
+                             data.frame(age2016=10) ,
+                             interval = "confidence")
> pred.confidence3
       fit      lwr      upr
1 44.09338 43.37964 44.80713
> pred.prediction3 <- predict(lr3,
+                             data.frame(age2016 = 10),
+                             interval = "prediction")
> pred.prediction3
       fit       lwr     upr
1 44.09338 -508.8902 597.077
>
```

1. Estimation of Individual rate by Number of benefits per plan

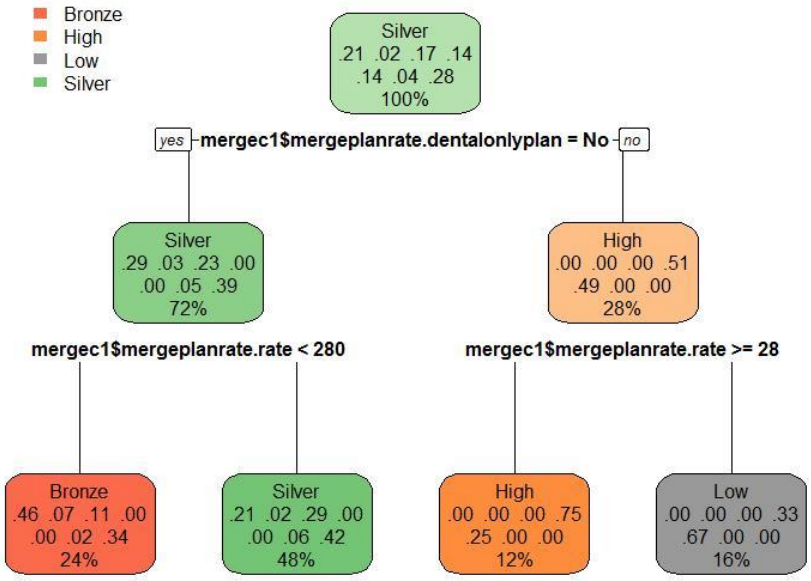
```
> pred.confidence <- predict(lr1,
+                            data.frame(noofben = 100),
+                            interval = "confidence")
> pred.confidence
      fit      lwr      upr
1 248.8449 245.1281 252.5617
> pred.prediction <- predict(lr1,
+                            data.frame(noofben = 100),
+                            interval = "prediction")
> pred.prediction
      fit       lwr      upr
1 248.8449 -156.7176 654.4073
```

b. Classification (supervised – you have to identify a target variable, show your knowledge of CART and C5.0 in R)

Predictors: IndividualRate from rate.csv and DentalPlanOnly from PlanAttributes.csv Target Variable: Metal Level
Steps involved in Classification:

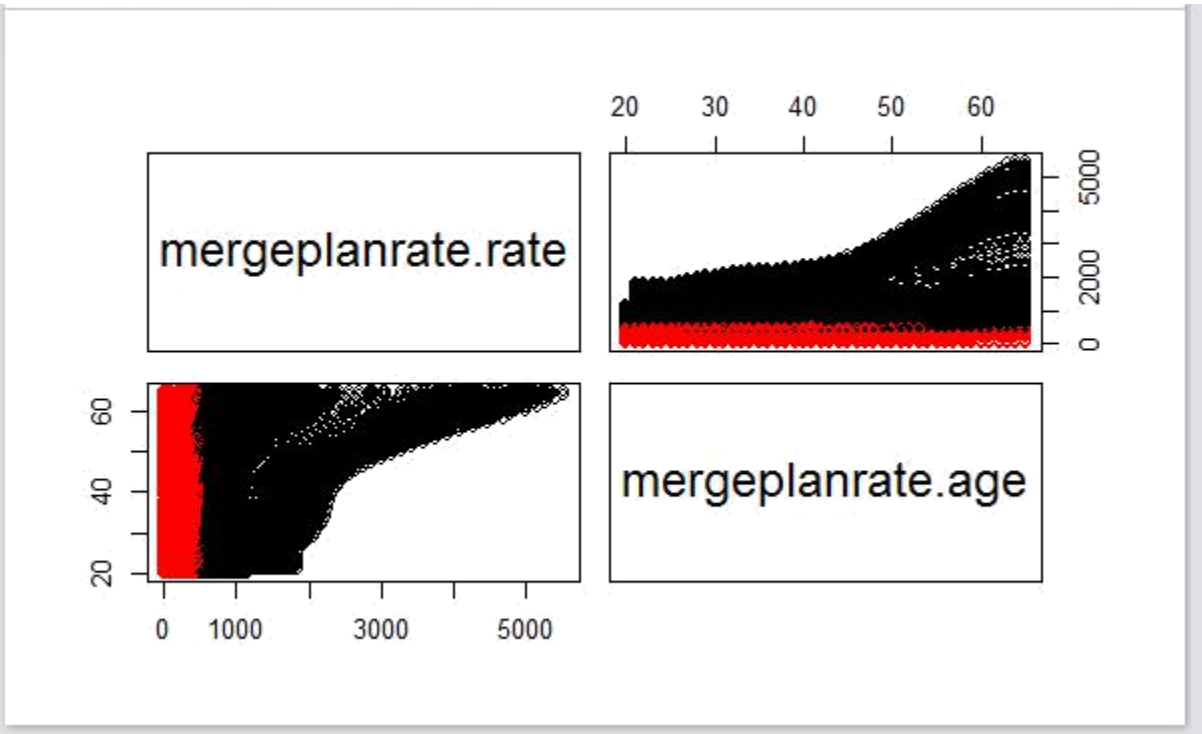
- Extraction of PlanId, Metallevel and DentalPlan Only from PlanAttributes.csv into a new dataframe calling it planfile
- Removal of duplicated entries of the planfile
- Extraction of Rate and PlanId from Rate.csv
- Clustering based on IndividualRate and Age Variables.
- Use of cluster 1 for classifying purpose
- Conversion of metallevel ,dentalonlyplan attributes into factors
- Normalization of IndividualRate by z-score normalization
- Classification using C5.0 algorithm with the following formula
    - metallevel ~ planrate + dentalonlyplan

  Output :



c. Clustering

While exploring the data and after cleaning the data we still had many records about 6 million records. Because we wanted to use the sample of that data we decided to use clustering. Thus we used the kmeans algorithm for the clustering where we found 2 separate clusters after using kmeans algorithm. We used age and rate for the clustering both located in the Rate.csv file. After we got the clusters we used one of the clusters for association as well as classification. Below is the plot of the age and the rate with respect to the different clusters. As we can see the dots is red represent one cluster and the dots in the black represent other cluster. A




```
> km $centers
  mergeplanrate.rate mergeplanrate.age
1           707.0254          54.83345
2           241.2183          37.25501
> km $size
[1] 2011508 4739462
>
```

In the image above we can see the different centers of age and rate for different clusters and also the size of the clusters

d. Association

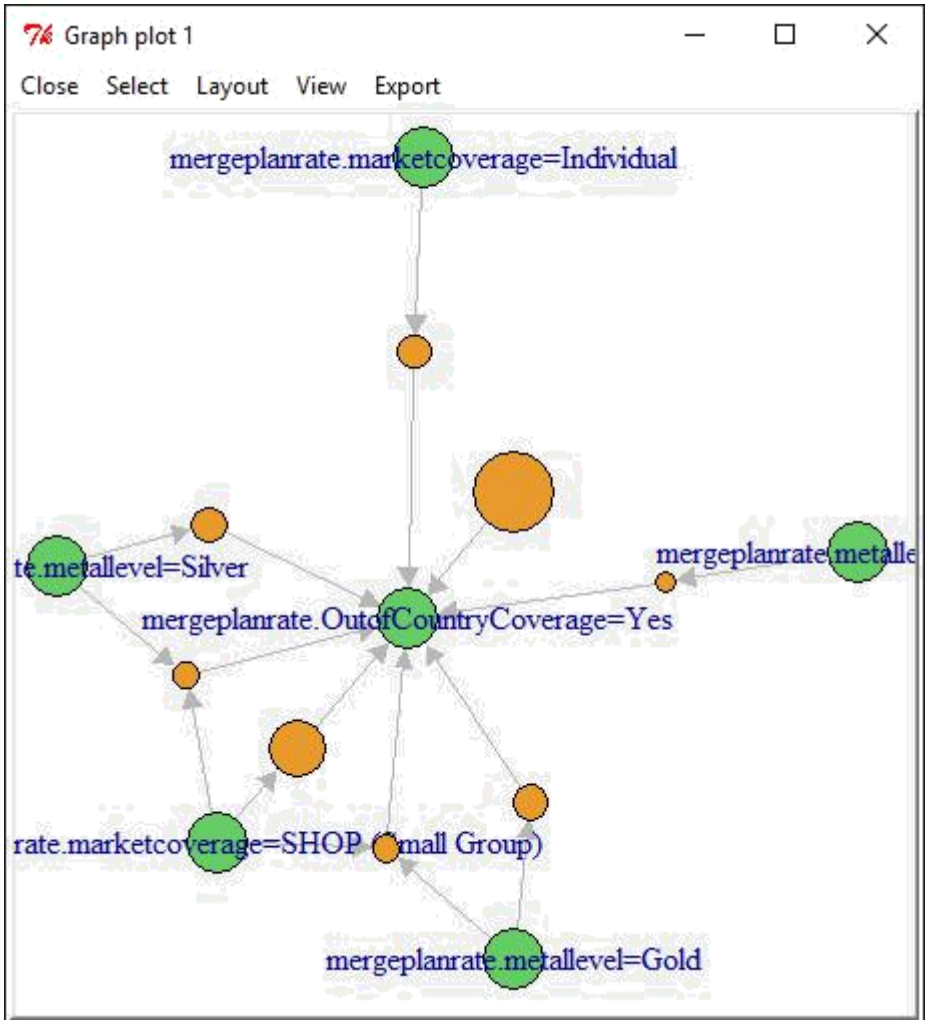
In association we are associating the columns OutofCountryCoverage, marketcoverage and metallevel all of these columns are located in the PlanAttributes file. We are taking one of the cluster that we got while doing clustering and we combine all of the records with the PlanAttributes file. The reason for choosing these attributes was that people will choose plan according to Metallevel which are gold, silver, bronze, platinum for the medical plans also low and high for the dental plans. Also people also has a preference for choosing the plans from either an individual or a small group. Thus we wanted to find the combination of those attributes for which maximum Out of country coverage is provided

```
> inspect(sort(rules))
  lhs                                                    rhs
       support confidence      lift
1 {}                                                  => {mergeplanrate.OutofCountryCoverage
=Yes} 0.8376780  0.8376780 1.0000000
2 {mergeplanrate.marketcoverage=SHOP (Small Group)} => {mergeplanrate.OutofCountryCoverage
=Yes} 0.5535703  0.8551878 1.0209028
3 {mergeplanrate.metallevel=Silver}                   => {mergeplanrate.OutofCountryCoverage
=Yes} 0.3045481  0.8348667 0.9966440
4 {mergeplanrate.metallevel=Gold}                     => {mergeplanrate.OutofCountryCoverage
=Yes} 0.3008410  0.8418434 1.0049726
5 {mergeplanrate.marketcoverage=Individual}           => {mergeplanrate.OutofCountryCoverage
=Yes} 0.2841077  0.8055415 0.9616363
6 {mergeplanrate.metallevel=Gold,

   mergeplanrate.marketcoverage=SHOP (Small Group)} => {mergeplanrate.OutofCountryCoverage
=Yes} 0.2165147  0.8584621 1.0248116
7 {mergeplanrate.metallevel=Silver,

   mergeplanrate.marketcoverage=SHOP (Small Group)} => {mergeplanrate.OutofCountryCoverage
=Yes} 0.1972132  0.8610181 1.0278628
8 {mergeplanrate.metallevel=Bronze}                   => {mergeplanrate.OutofCountryCoverage
=Yes} 0.1265821  0.8215224 0.9807138
>
```

Result: As you can see above these are the association rules we got after using the apriori algorithm. From the association rules we can see that when the marketcoverage is small group and the metallevel is gold or silver the lift is 1.0248 and 1.0278 respectively compared to other combination of the attributes. Thus we can say that for combination of these values there is a high chance the plan would have Out of country coverage as compared to other values. Also below you can find the graph of all the association rules



VII. Evaluation

Evaluation of various modeling techniques is done as follows:

Description: The data mining models deployed are as transparent as possible. Also with respect to each model, the results that are generated to appeal to the domain knowledge with insights into future predictions that may result into profitable actions.

Estimation and Prediction:

1. Prediction of Individual Rate by Age a. Business Year :2014

   Coefficient of determination: 277.7
   Mean Square Error: 13.88%

   b. Business Year: 2015

   Coefficient of determination: 12.98% Mean Square Error:277.2

   c. Business Year: 2016

   Coefficient of determination :13.68 % Mean Square Error:282.1

1. Prediction of Individual Rate by Number of benefits of plan

   Coefficient of determination: 18.93%
   Mean Square Error: 206.9

Classification Task:

Constructing a contingency Table of correct and incorrect classifications to evaluate the classification task of dental plans classified as High and low

		Predicted Category		
Actual Category		High	Low	Total
	High	398491 = TP	244231 = FN	642722
	Low	131643 = FP	496842 = TN	628485
	Total	530134	741073	1271207

Overall Error rate: $\frac{FN + FP}{TN + FN + FP + TP}$ = 375874/1271207 = 0.295

Overall Accuracy = 1 − Overall Error Rate = $\frac{TN + TP}{TN + FN + FP + TP}$ = 0.705

Proportion of False Positives = PFP = $\frac{FP}{FP + TP}$ = 0.248

Proportion of False Negatives = PFN = $\frac{FN}{FN + TN}$ = 0.329

Sensitivity = $\frac{\text{Number of true positives}}{\text{Number of actual positives}}$

$= \frac{TP}{TP + FN}$ = 0.612

Specificity = $\frac{\text{Number of true negatives}}{\text{Number of actual negatives}}$

$= \frac{TN}{FP + TN}$ = 0.790

VIII. Report of Results

a. Knowledge Discovered

The knowledge we discovered after exploring and various tests and algorithms on the dataset were as follows

- States like Mississippi and Alabama have the highest mean deductible whereas states like Tennessee, Wyoming and South Dakota have the lowest mean deductible as we found out while plotting the states on the choropleth map

- While plotting the bar graph of mean number of ebb benefits per plan with respective to each state we found out that New Mexico provides highest number of mean ehb benefits per plan while Alaska has the least
- We plotted the box plot of the rate of the insurance with every year for tobacco and non-tobacco preference separately and found no change in the rates
- By plotting pareto chart with respect to mean individual rate we found that Alaska has the highest mean individual rates while Missouri has the lowest mean individual rates.
- By conducting hypothesis test on the sample of the rates for every year from 2014-2016 we found that p value is very small thus sample is not representative of the population.
- When plotting the choropleth graph of tobacco preference for every state we found out that Wisconsin, Pennsylvania, Florida and South Carolina has the most number of plans with tobacco preference while Wyoming and North Dakota has the least.
- Comparing the market coverage plans for both small groups and individuals for every year we found out that number of plans increases significantly from 2014 to 2015 for both while there is a slight decrease in the number of plans from 2015 to 2016 for both individual and small group
- When we did the regression plot for the Individual Premium Rates versus Benefits provided we found out that as the number of benefits increased the rates also increased.
- While performing regression models for age versus rate for different years we found at the regression line for the year 2014 is almost constant while it is linear for the years 2015 and 2016.
- By doing association we found out when the metal level is silver or gold and the market coverage is the small group there is a high chance the plan will have out of country coverage compared to the combination of other variables.

b. Predictive Capabilities

- By using the regression models for individual years from 2014-2016 for rate versus age we can predict for future years that rate versus age regression line would be either constant, or linearly increasing in nature. Thus as the age will increase the rates will also increase
- Using the regression model for the rate versus no of benefits for the years 2014-2016 we can predict for future years that as the number of benefits will increase so will the rate. Thus the regression line is going to be linearly increasing.
- By using the number of records in our dataset we can classify the new records according to their metal level, dental plan and the rate by comparing it with our classification model we used for our existing dataset
- Using the clustering in future we can identify which record will be in which cluster using our existing dataset
- With the help of the association rules from the existing dataset if someone wants an out of country coverage plan we can predict what kind of plan and from which market will benefit them the most i.e. with the help of the metallevel, marketcoverage and the OutofCountryCoverage columns in the planattributes.csv dataset

IX. Limitations

The main limitation we had was that the dataset was not the customer dataset or the dataset of companies giving insurance to client. We had the dataset of the agents who issues the insurance who were mentioned by issuer id in every dataset. Also there were columns in the

dataset which we could not understand what exactly they meant. Also the dataset being very large we had very less time exploring each and every bit of the dataset.

Sources :

https://www.kaggle.com/ruonanding/d/hhsgov/health-insurance-marketplace/plans-and-carriers- by-state/code

https://www.kaggle.com/benhamner/d/hhsgov/health-insurance-marketplace/median-monthly- premiums-by-state

https://www.kaggle.com/benhamner/d/hhsgov/health-insurance-marketplace/rates-strange- values

https://favorableoutcomes.wordpress.com/2012/10/19/create-an-r-function-to-convert-state- codes-to-full-state-name/