**Health Insurance Marketplace Analysis**

Chenhui Huang (BME), Tingran Yang (BME-EE), Jiayi Zhang (EE)

Columbia University

[ch3216@columbia.edu](mailto:ch3216@columbia.edu), [ty2362@columbia.edu](mailto:ty2362@columbia.edu), [jz2856@columbia.edu](mailto:jz2856@columbia.edu)

*Abstract*—**Choosing the most suitable health insurance plan is not easy. There is an abundance of options in the health insurance marketplace. The insurance company always keep pricing decision in a black box. As the data in the healthcare domain is made open for public use, one can discover invaluable information from those data by utilizing the power of predictive modeling and machine learning. Therefore, our group is motivated is to find out the most influential factor in determining the monthly premium rate. In the meanwhile, we have developed an open-source website/software package for consumers to choose their most suitable plan. Also, in this study, a well-rounded exploratory data analysis was performed to study the current landscape of the health insurance marketplace.**

***Keywords-component; health insurance landscape; predictive modeling; monthly premium rate; linear regression; exploratory data analysis; data visualization***

1. Introduction

Health insurance matters to all of us. 90.4% of the United States population have healthcare insurance. On average, the monthly premium for family coverage is $1,462 per month or $17,545 per year.As the Patient Protection and Affordable Care Act (ACA), known as the ObamaCare, became fully operational in 2014, consumers now have the option to purchase different government regulated health care plan that complies with the ACA.In the United States, health insurance marketplaces, also called health exchanges, are organizations in each state through which people can purchase health insurance. It aims to promote individuals to compare and choose from a range of available health plans, and select the one that is most suitable. In the meanwhile, the healthcare reform act proposed the “Meaningful Use” incentive. As a result, much of the data was made public online, and became very easy to access.

Given the massive datasets were made public, consumers now can fully take the advantage of this resource. With the power of machine learning or data mining, one can get invaluable insights and patterns via analyzing the enormous datasets available. However, data in the healthcare domain are notorious for its dirtiness and untidiness. As Hadley Wickham has pointed out in his paper, “tidy data”, it is essential to perform data wrangling to get it ready for meaningful analysis.Therefore, one of the goals for this study is to clean this enormous dirty and untidy dataset. There are many missing values, conflicting data entries, and inconsistent data types.

Another motive of this studies is that we wish to understand what the health insurance landscape looks like. Particularly, we would like to examine the relationship between different variables. For instance, one of the exploration questions we have is how do monthly premium rates vary by different state, plan, or age? There are many more interesting and hidden patterns to be discovered in this dataset.

With a more advanced understanding of the data structure and relationship between variables, we are inspired to create a predictive model to predict individual monthly premium rate based on all the available features in this dataset. Although the price is available across different plans online, but one thing insurance company will not disclose to consumers is how the price are made up or the pricing decisions. In other words, we are curious at what factors are the most important and influential in determining the price of each plan. Based on the model, we are also motivated in turning this model into a web application, that can provide a expected monthly premium rate according to users preference.

1. Related Works

This particular datasets has piqued interests in the Kaggle community. However, most of the efforts have been focused on simple exploratory data analysis. For instance, there were users who explored rate is distributed and varied by different metal level and state. There haven’t been work done on the predictive modeling part. Therefore, we are interested to build a predictive model to predict monthly premium rates based on the findings in those exploratory data analysis.

1. System Overview

The dataset is published by the US department of Health and Human Services. It contains 6 main data tables, with each describing: benefits and cost sharing (BCS), rate, plan attributes (PA), network coverage area, service area, and business rules. Among those, network coverage area, service area, and business rules tables mainly describe issuer-level data. Issuer here means insurance company available in the health insurance marketplace. Since the main goal of this project is to explore the landscape of each individual plan and predict rates for each plan, all issuer-level data was omitted from this study.

The rate.csv (1.97 G) contains plan-level data on individual rates based on an eligible subscriber’s age, and has over 12 million data entries. The BCS table (1.38G) contains plan-level data on essential health benefits, coverage limits, cost sharing (copay and coinsurance), and so on. It has over 5 million data entries. The plan attributes table, as the name suggests, contains plan-level data that describes attributes of each individual plans, such as maximum out of pocket payments, deductibles, number of wellness program offered, and 174 other variables. Not all variables were used in this study, as we have manually select 63 key variables based on our understanding of the healthcare industry.

Data wrangling was primarily done in python at local machines, and the packages utilized include numpy and pandas. Exploratory data analysis for each individual table was done in R at local machine. However, when joining aforementioned tables together, it was impossible to handle data in local machine memory. Therefore, we utilized the power of spark (pyspark and sparklyr). Exploratory data analysis and data visualization was done using package such as dplyr and ggplot2. We built our predictive model using pyspark, and its machine learning library.

A website application was built in html/css. We applied a linear regression model we trained from the training sample of the data and validated from the testing sample. Users can use this web app to predict how much their individual rates will be by entering several factors like ages, expected coinsurance, etc. In this way, consumers can have an initial estimation individual rate and understand the factors that influence the individual rate.

1. Algorithm

The data wrangling process was primarily done programmatically using python’s pandas library. Some of the common functions used to assess the data quality includes pandas’ “info”, “value\_counts”, “describe”, and “isnull”. After the data was fully assessed, cleaning was done using python’s built-function. In many cases, missing values were filled according to dataset’s dictionaries, available at cms.gov website. Outliers were replaced and approximated as the 95th percentile value. Data types were also corrected using python’s astype function. The exploratory data analysis and data visualization was mainly done in R using sparklyr, dplyr, and ggplot2 packages as mentioned in the previous section.

In this study, one of the main goals is to predict the individual rate for each single plan. Before implementing any predictive model, the data was splitted into two chunks, training data which occupies 80% of the original data, and test data (the remaining 20%) to validate the model to ascertain the model is general and not overfitting.

The algorithm/model we used to predict the rate from our data was lasso linear regression. We assumed that there is linear relationship between the dependent variable , which is the individual rate, and the p-vector of regressors which are features picked out by us. This relationship was also modeled through a disturbance term of an unobserved random error variable , which adds noise to the linear relationship between the dependent variable and regressors.

Since we assume that the dependent variable and regressors have linear relationship, the model should take the form as:



For our regressors, we picked 10 columns which we find that might be most meaningful from the earlier exploratory data analysis. In this form is a parameter vector, where is the constant offset term. In this equation, the vector gives out how those variable can affect the dependent variable, thus we can also call them coefficient. Our target is to acquire the parameter vector and the error term by regression algorithm.

However one may encounter independence of data in this model. If different covariates are not independent from each other, that is to say, some of those data are collinear, then the value of may not be uniquely determined. If goes too large, it can make prediction error worse. Therefore we need to improve the prediction error by limiting large regression coefficients to be subject a constant in order to reduce overfitting in our model and give out better performance.

Lasso was originally introduced in the context of least squares, and it should be considered as a sample consisting of N cases, each of which consist of p covariates and a single outcome. Let be the outcome and be the covariate vector for the *i*th case. Then the objective of LASSO is to solve:



Here is a prespecified free parameter that determines the amount of regularization. Thus we can rewrite our object of LASSO as:



In the so-called Lagrangian form we can rewrite the objective as:



In this form the exact relationship between and is data dependent.

1. Software Package Description

We developed our website application in html/css as a static website. We used html to establish the frame of the website by adding header, columns, and components like button and text boxes. Then we added style to the frame. We set the body background in light blue, and changed the font size and alignment. We put some of the most important factors which has relative large coefficients in our linear regression model on the website for users to input for an estimation. We wrote a JavaScript code to collect the input of users, and we could calculate and output an estimated individual rate.

Figure 1 shows the entire website UI interface. All input and output are operated and displayed on this website.

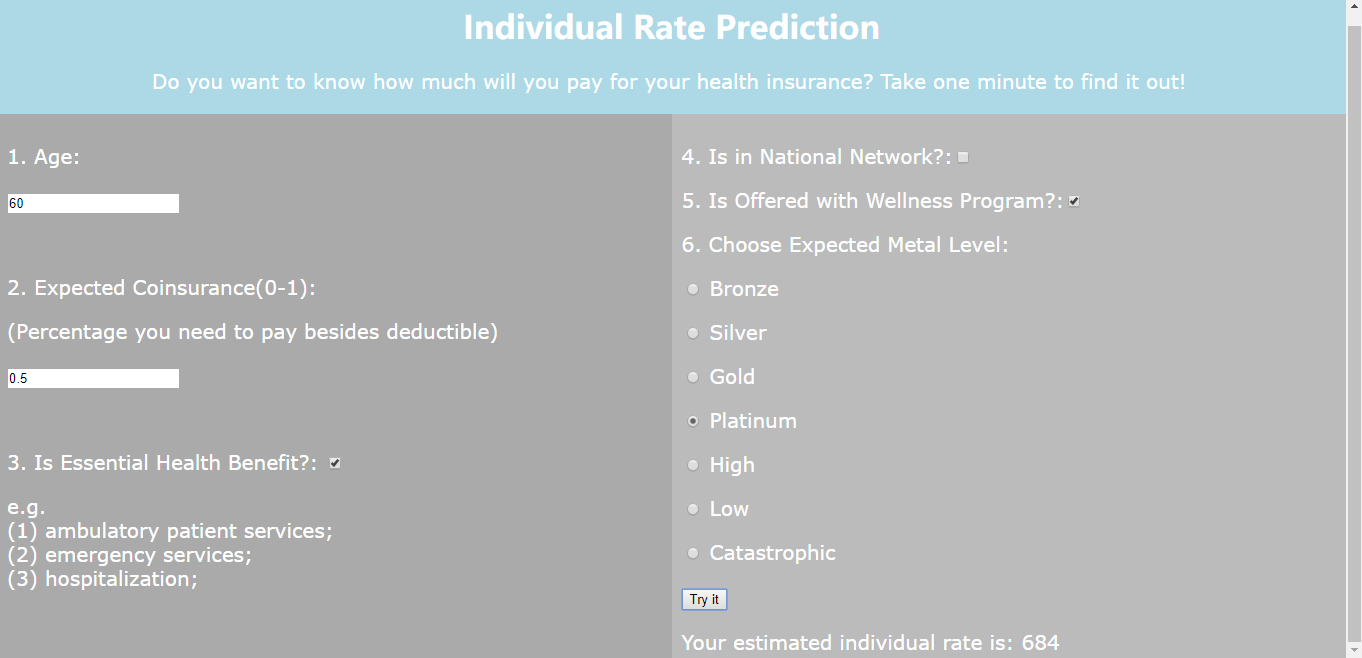


Figure 1. Individual rate prediction website applictaion.

In the part 1 of the questions, users need to fill in ages, which should be a positive integer within a normal applicable range. The coinsurance means the the percentage of costs of a covered health care service you pay (20%, for example) after you've paid your deductible. User also needs to fill in whether or not the plan Essential health benefit is included in the plan.

In the part 2 of the questions, whether or not the plan is nationwide covered and whether or not it is offered with wellness program. Users also need to choose an expected metal level. Finally, after all inputs above, user could click the button “Try it” to see the result.

1. Experiment Results

The exploratory data analysis started with asking exploratory question such as, “How does the monthly premium vary by age?” or “How does the monthly premium vary in different state?” Figure 1 and figure 2 illustrates those problems respectively. In figure 1, we can see there is a very obvious trend where the median monthly premium increases as the age increases, and the slope also increases when age is 45 years and older. Figure 2 shows the landscape of the median monthly premium rate in different state. The scale runs from yellow to blue, where yellow indicates a lower premium rate and blue indicates a more expensive premium rate. States in black indicate that the data is not available in those states.

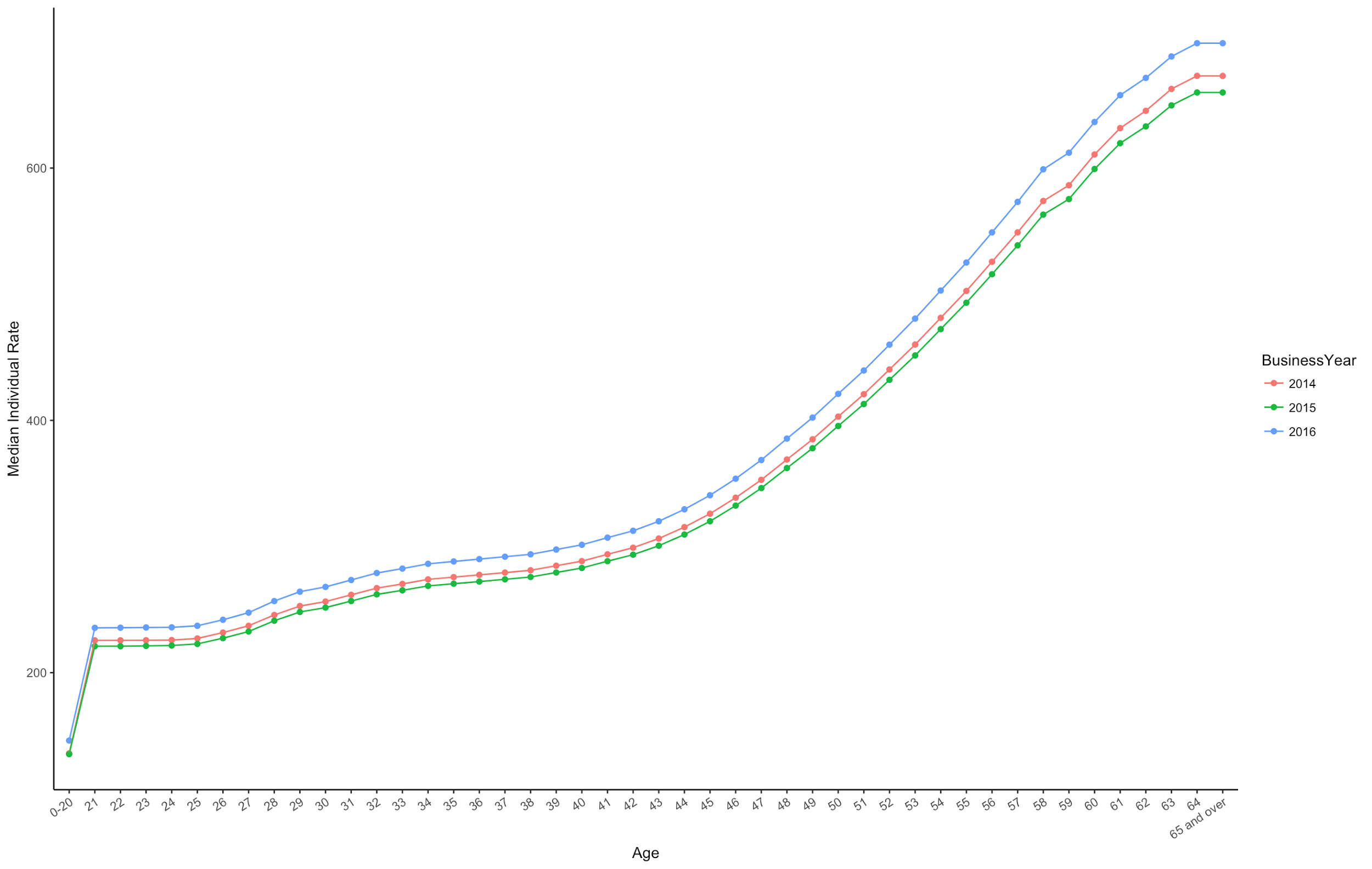
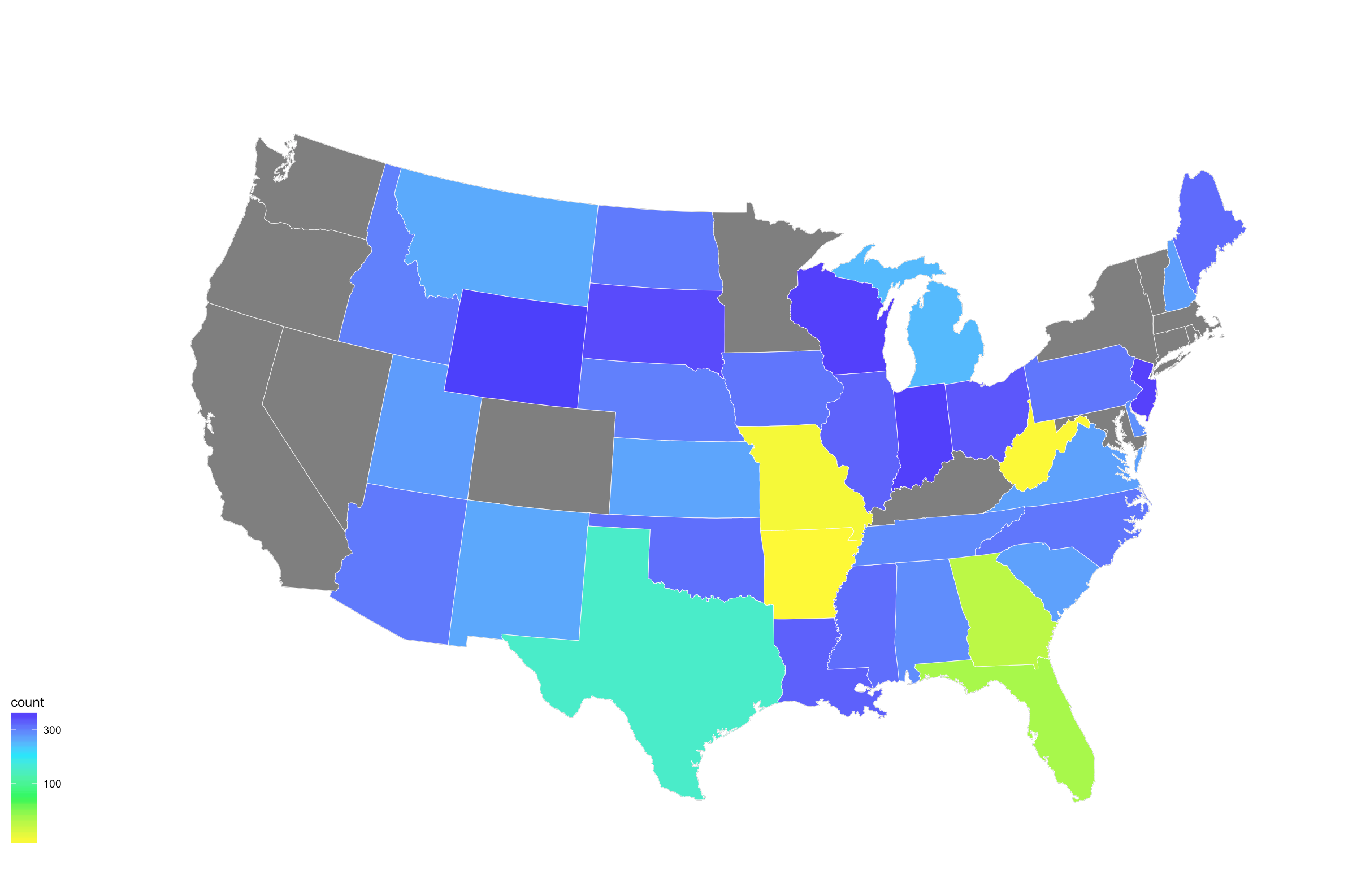


Figure 2. Age vs Median Monthly Premium Rates



Further exploration was done when tables (BCS, rate, PA) were joined together using sparklyr. Since it is impossible to plot every single entries in a single machine, we have decided to visualize the aggregated data. We were most curious and interested in finding out how each individual plan is different from each other. The first aggregated data for analysis was grouped by planID and issuerID. From the aggregated data, one can calculate the mean monthly premium rate, mean copay, mean coinsurance, mean maximum out of pocket pay, etc. In most cases, the distribution made common sense and fitted our expectation. For example, figure 4 is a boxplot between the median coinsurance and plan metal level. Metal Level has 4 categories, with Platinum being the most premium plan, followed by Gold, Silver, and Bronze. We did see a steady drop in the rate as the plan gets more premium. However, there are also instance where the relationship is unintuitive to our understanding. In figure 5, the amount of copay go up as the plan gets more premium.

Figure 3. Median Monthly Premium Rates in states

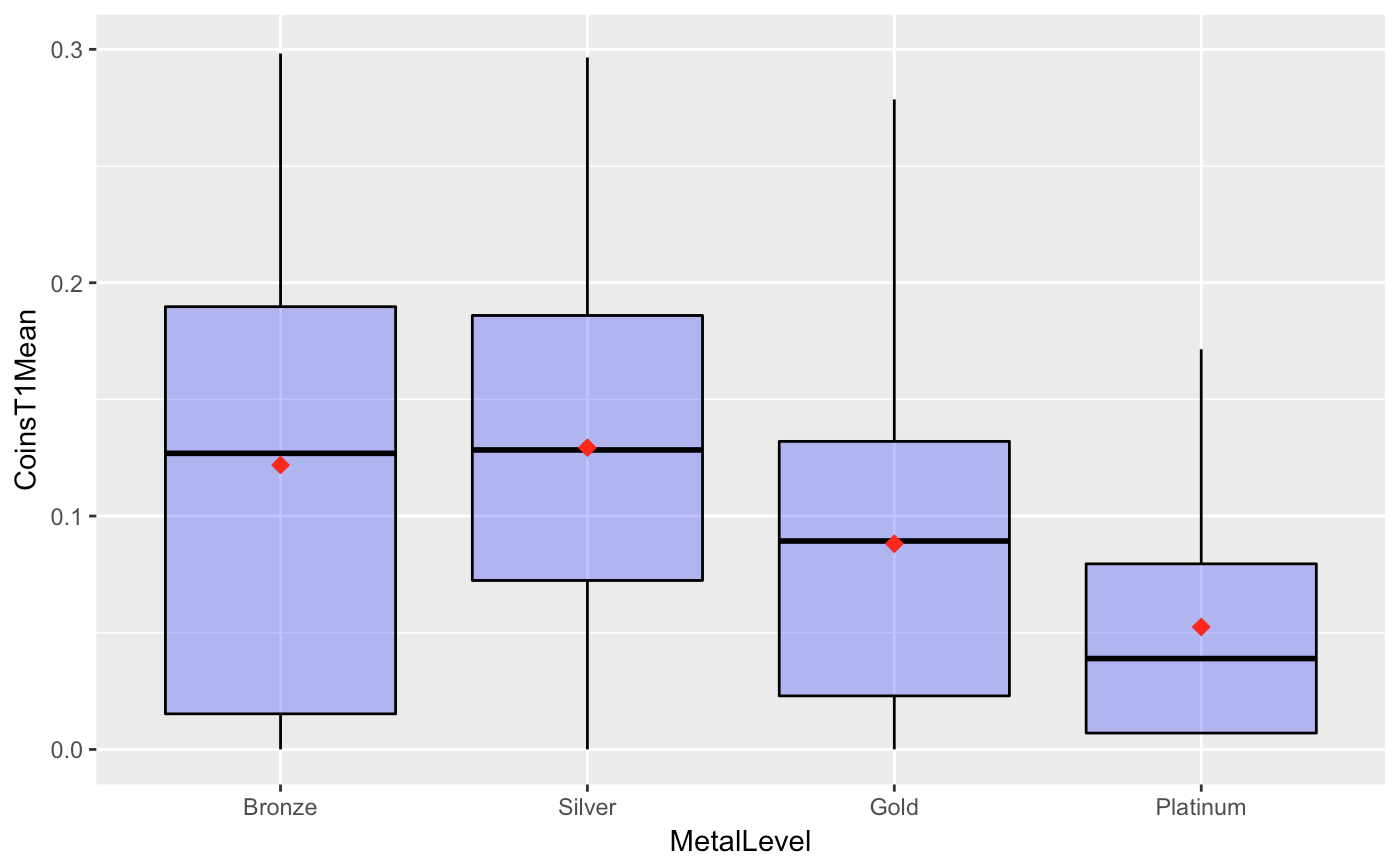


Figure 4. Median Coinsurance by Plan Metal Level

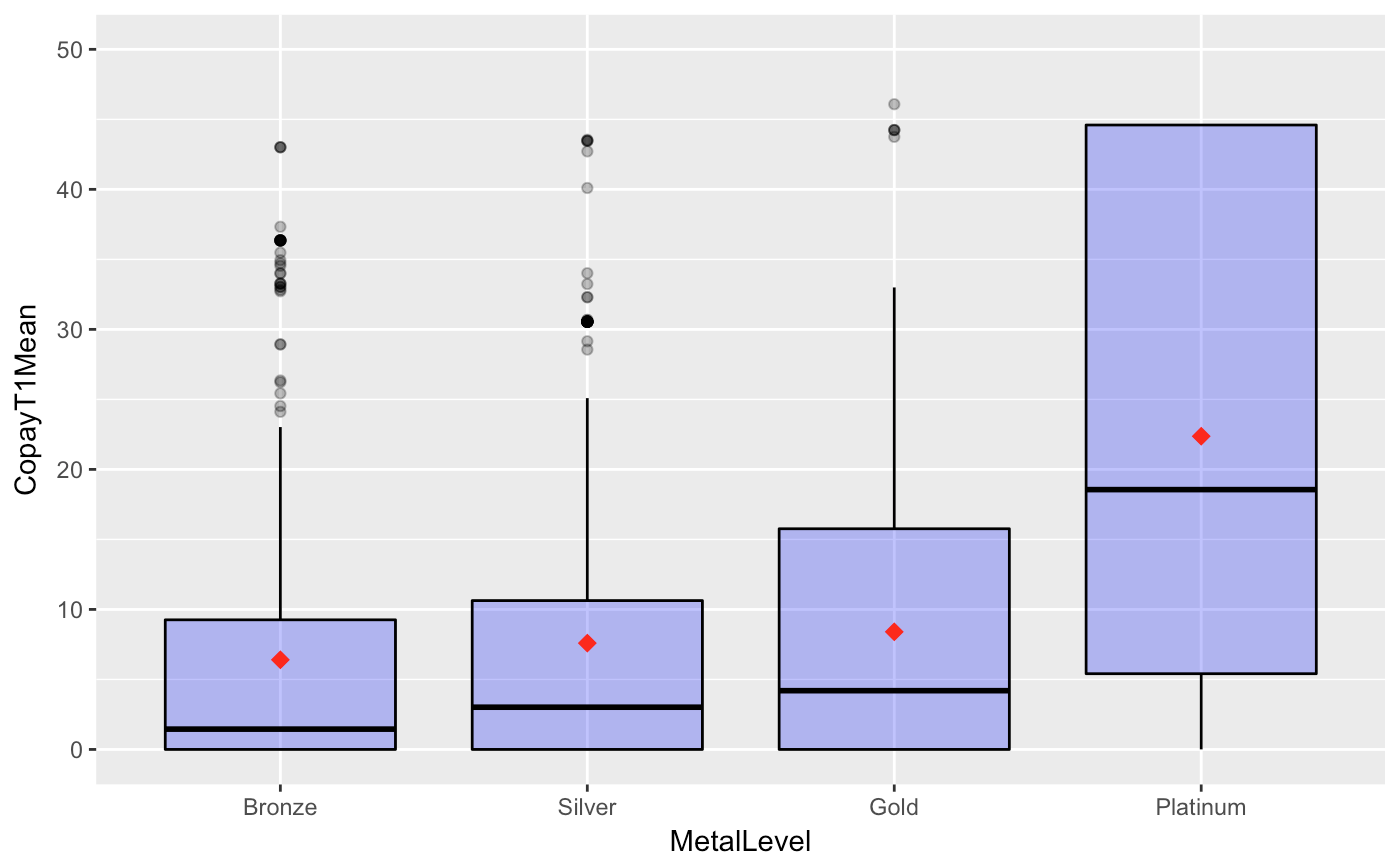


Figure 5. Median Copay by Plan Metal Level

We have also drawn scatter plots between the rates and other variables, such as the amount of deductibles (figure 6) and issuer’s actuarial value (figure 7). The issuer’s actuarial value is the percentage of total average costs for covered benefits that a plan will cover calculated by the issuer. As we can see, the distribution fitted our expectations: the rate increases as the amount of deductible decreases (figure 6) and the rate increases as the issuer actuarial level goes up (figure 7). A coefficient coefficient matrix was also made to better visualize the relationship between each individual variables, as figure 8 illustrate.

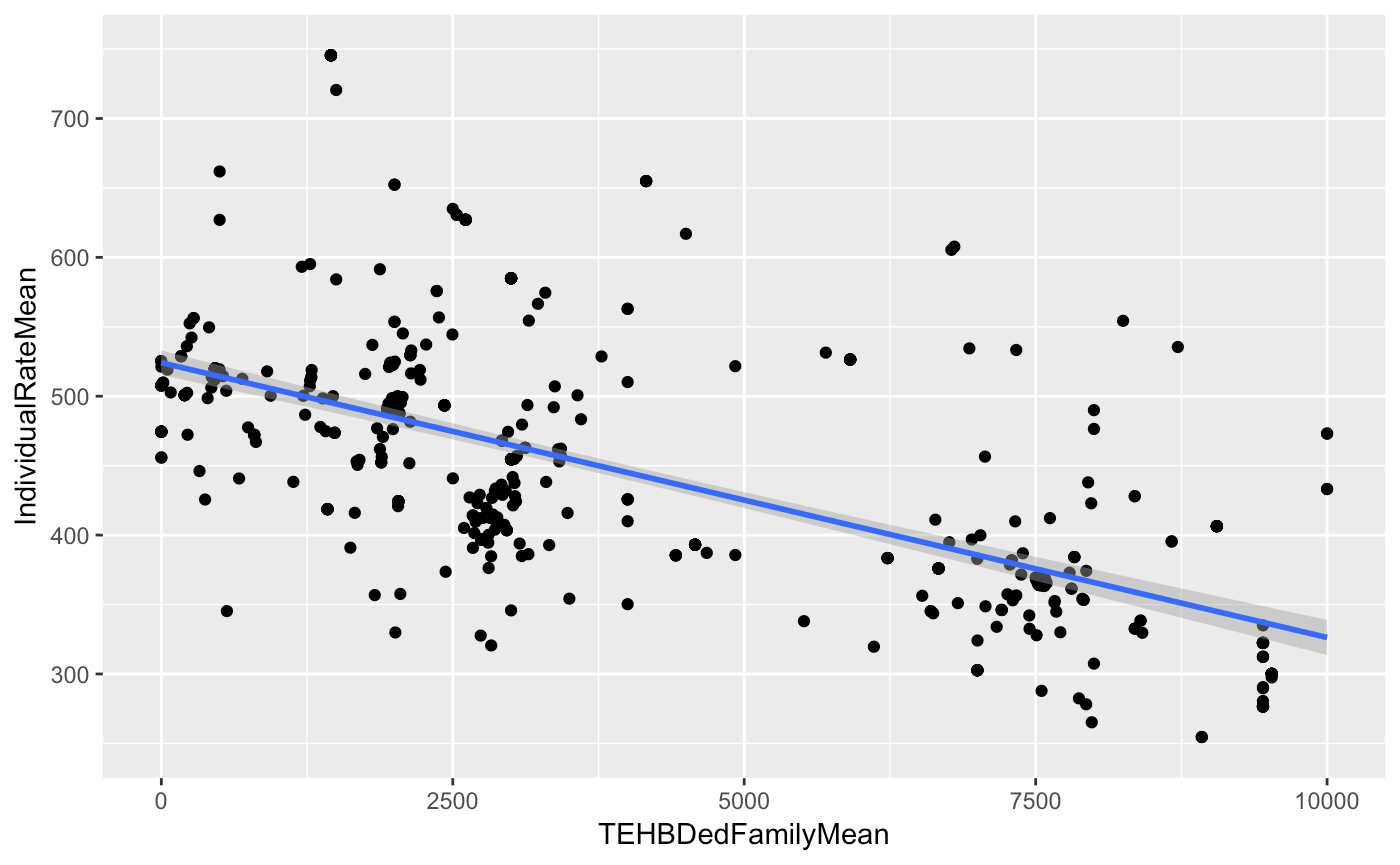


Figure 6. Rate vs Amount of Deductibles

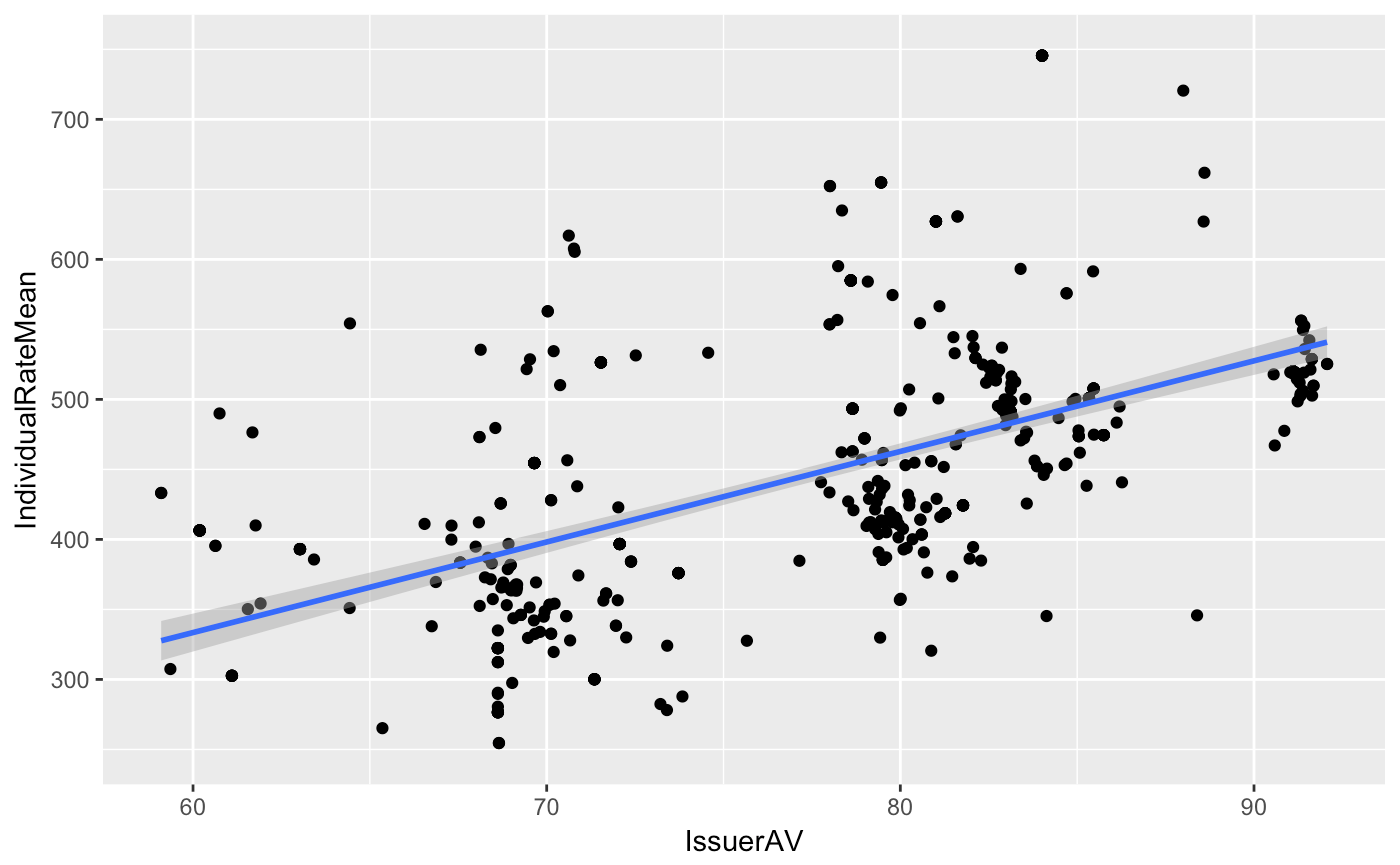


Figure 7. Rate vs Issuer’s Actuarial Value

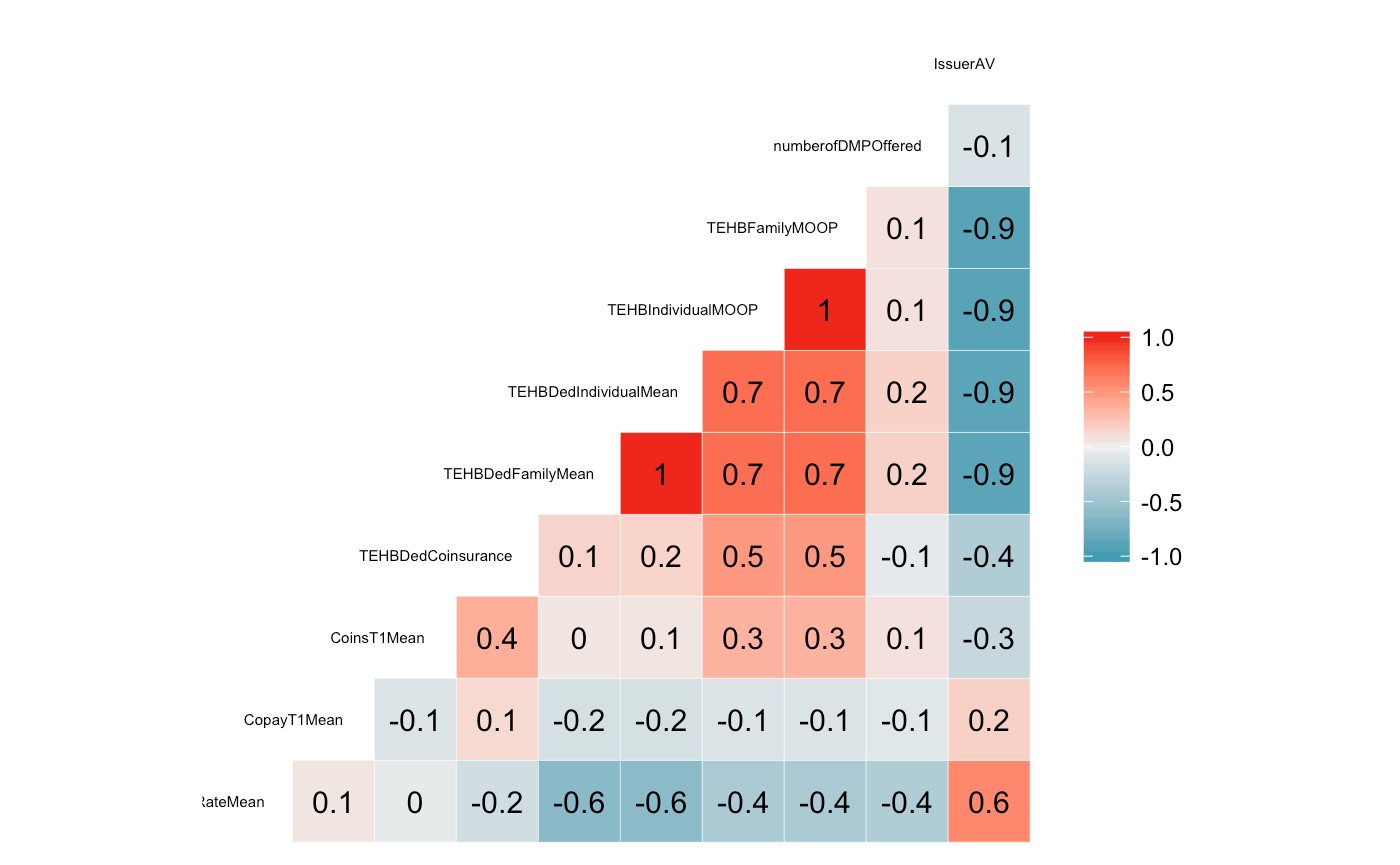


Figure 8. Correlation Coefficient Matrix

Another aggregated data we looked at is the data grouped by different benefits. Based on the aggregated results, we can find out some of the more expensive covered benefits (figure 8), such as gym membership reimbursement, massage therapy, which makes a lot of sense.

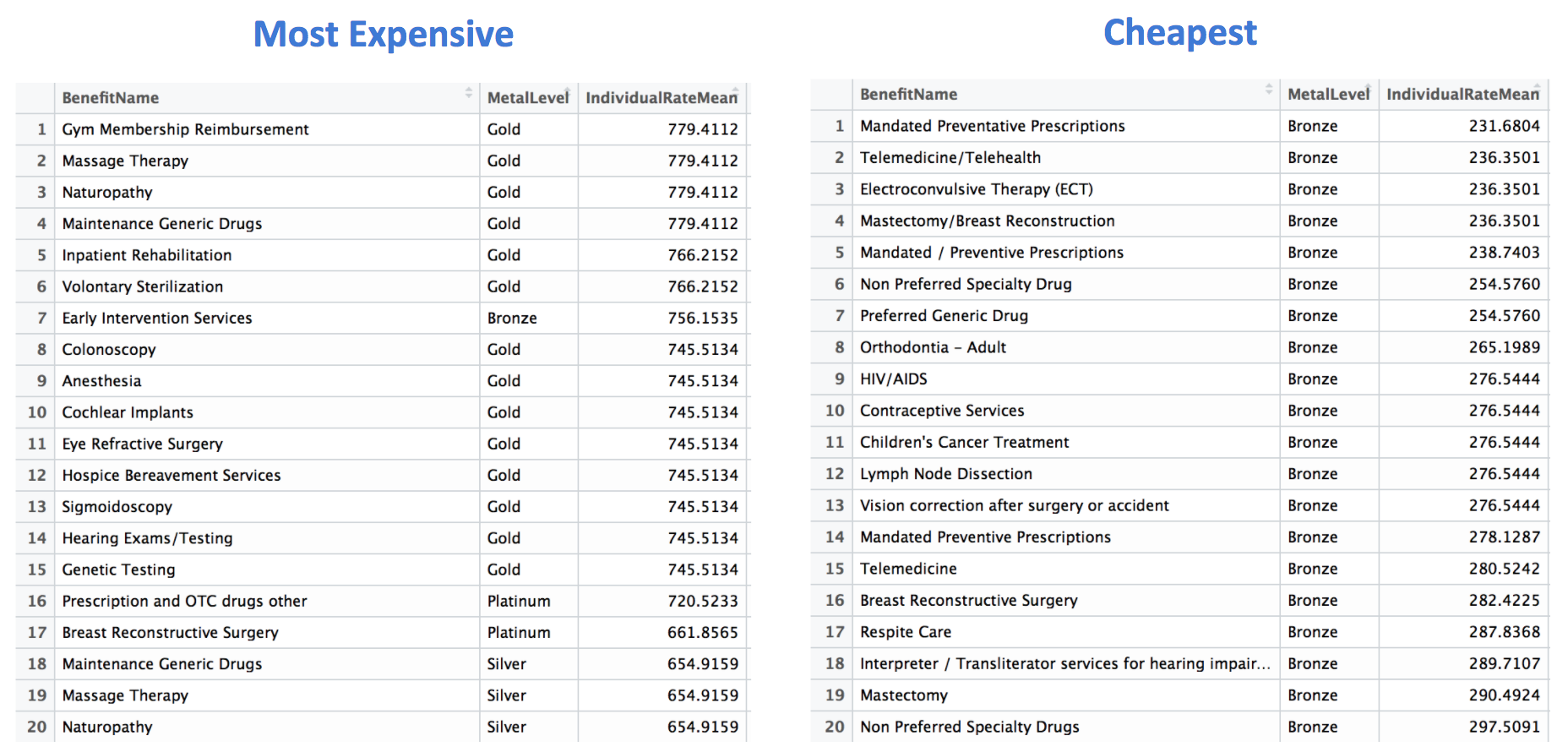


Figure 9. Most expensive (left) and Cheapest (right) Benefits

As mentioned earlier, the data was splitted to training and testing chunks before any model was built. Before fitting into regression model, we have to do some data wrangling to make our data numerical thus can be fit into regression model. For example, we replaced the value “0-20” in age column to 19, and we replaced the value “65 and over” to 66. For binary value “yes” and “no”, we simply convert them into 1 and 0. As for ordered categorical value such as “bronze” or “gold” in metal level column, we replaced them with number 0 to 4 to make them numerical. Then when can construct our training data with one label column, which is the individual rate, and ten attributes columns which will form the feature vector later into model.

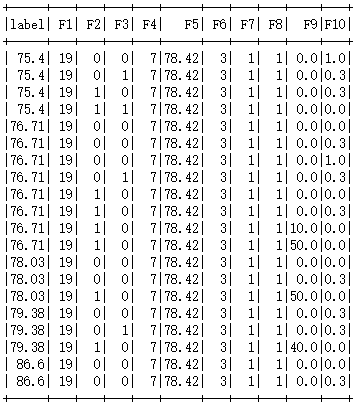


Figure 10. Feature Vectors

After applying LASSO linear regression to our data we can see the different coefficient of every single variable which are used to predict the individual rate. We can find that the age of users affect the rate most with its coefficient 10.67. Since the age attribute varies from 19 to 66, it gives most change to the predicted individual rate, which is exactly the same as what we have found in the exploration part. Obviously, the metal level also affects the rate a lot with its coefficient 32.13. People expect higher individual rate with higher metal level. The “National Network” variable also has a remarkable coefficient of 30.85, which means that the rate will goes higher if the plan is nationally covered. Besides the coinsurance variable has a coefficient of -38.19. Plans with high coinsurance rate receive low rate and plans with low coinsurance receive relatively high individual rate. And one thing interesting and quite unintuitive, if the plan is offered with wellness program, it is very strange to see that the individual rate will go down a lot, with the coefficient -71.40. Using the model we fitted, we could predict the individual rate now.

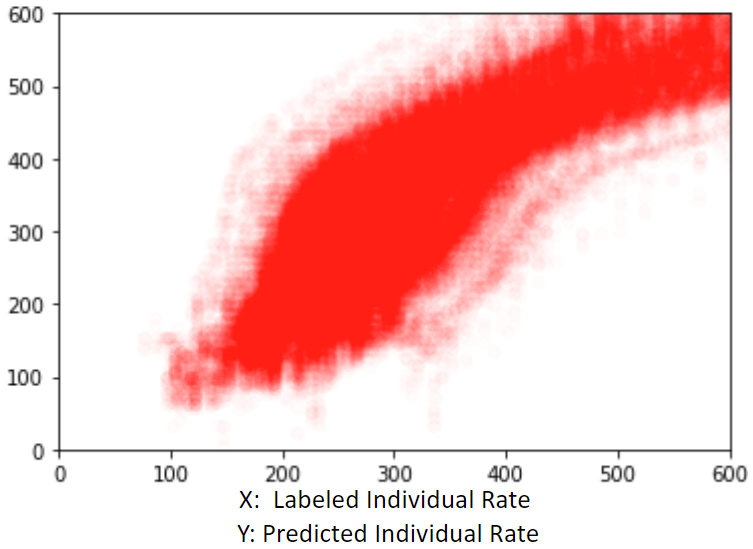
 The R squared of labeled individual rate and predicted individual rate is 0.71. In conclusion, we found that variables including the age of users, the metal level of plan, the coinsurance of plan, whether the plan is offered with wellness program and whether is plan is covered in national network matters the most.

Figure 11. Plot of Individual Rate Prediction Result

We tested our website by entering several different settings of the factors to see the different results. Firstly we set the factors as 40, 0.2, Yes, Yes, No, Gold, respectively by the order of Age, Expected Coinsurance, Is EHB, Is in National Network, Is offered with Wellness Program, Expected Metal Level. The estimated individual rate is 552$ / month. Then we alter the age from 40 to 30, the rate goes down to 446$ / month, which indicates that as age goes up, the individual rate goes up as well. We take another try to alter the metal level from gold to platinum. As a result, the individual rate goes up to 585$ / month. The results are intuitively straightforward and make sense to us.

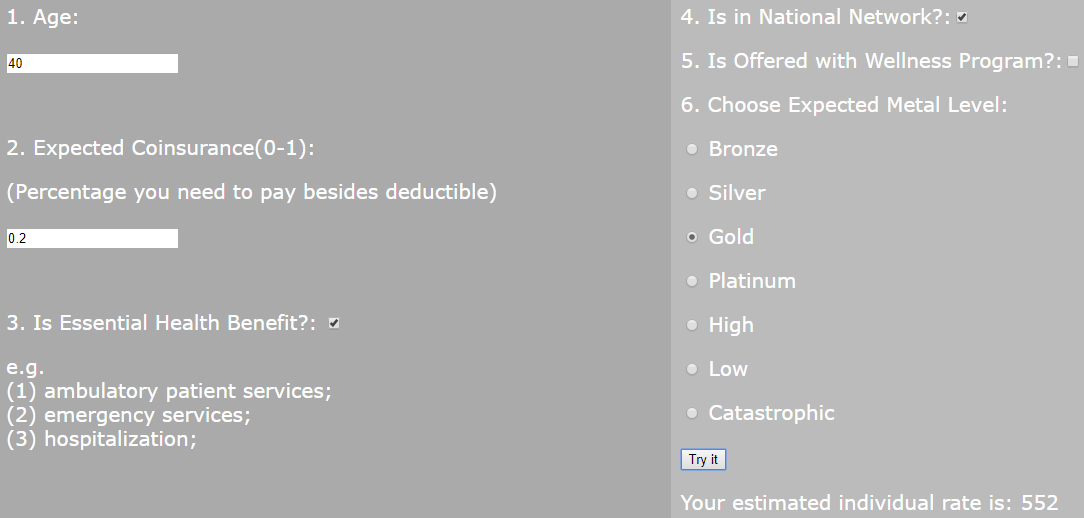


Figure 12. Individual Rate Estimation of a set of user inputs.

1. Conclusion

Our project conducted a detailed data exploratory analysis of the health insurance marketplace dataset to see how monthly premium rates vary by different state, plan, or age and how insurance coverage distributions vary by geographical region, etc. We achieved a linear regression model to predict individual rate and discovered the factors that are most important and influential in determining the price of each plan. In this project, we have distributed work evenly from the beginning, and all of us contributed as much as we could.

For future work, we would like to improve our model using larger data size and try different model besides linear regression. Furthermore we would like to develop a dynamic website to restore user profile and display some data exploratory analysis results according to users’ demand. In addition we would like to try to predict more potential interesting facts.

Acknowledgment

The authors would like to thank Dr. Lin for the instruction and resources during the project. We are grateful for any help given to us from our TAs. We would like to express our gratitude to our families, friends, and whoever helped us.

Appendix

Individual Contribution:

Chenhui Huang is responsible for data wrangling of table benefits, Plan Attribute and most Exploratory Data Analysis work.

JiaYi Zhang is responsible for data wrangling of table rate and developed the model of rate prediction.

Tingran Yang is responsible for data wrangling of table Network, Service Area and Business Rule and developed the website application.

References

1. <https://www.zanebenefits.com/blog/faq-how-much-does-it-cost-to-provide-health-insurance-to-employees>
2. <https://en.wikipedia.org/wiki/Health_insurance_marketplace>
3. H. Wickham, “Tidy Data” in Journal of Statistical Software, vol. 59, Issue 10, 2014.
4. https://www.kaggle.com/ruonanding/rate-by-coverage-level-metallevel
5. https://en.wikipedia.org/wiki/Lasso\_(statistics)