# Health Insurance Analysis and Prediction

Aiswarya Sriram,[1] Reuben Cherian,[1] Sindhu Bhoopalam Dinesh[1]

New York University, NY, USA {as14988, rc4610, sb8019}@nyu.edu

**Abstract.** We used the Health Insurance Marketplace Public Use Files which contain information related to health and dental plans offered in the US Health MarketPlace to individuals and small businesses. The analysis of medical insurance data of different states, age groups, and insurance providers can provide us with indicators that can help people make informed decisions to choose an optimal insurance plan.

**Keywords:** Big Data, pyspark, pysparkml, insurance, Individual Rate

## 1 Introduction

Our objective is to analyse medical insurance plan parameters to observe -
1. How plan rates affect people from different states
2. Analysis of plan benefits across the states
3. Effect of health habits and age of a person on the plan rates
4. Distribution of plan rates across the insurance issuers
5. Predict plan rates using applicable features.
Why is this a Big Data problem? Our dataset size is 3.4 GB and this just spans 3 years (2014-2016). As the number of years increase, the scale of the data increases. Moreover, to perform any kind of analysis with machine learning with larger number of features, we would need Big Data infrastructure.

This paper is structured as follows: Section 2 presents our Methodology-dataset overview, procedure and results. Section 3 shows our conclusion.

## 2 Methodology

### 2.1 Dataset Overview

We referred to the Health Insurance Exchange Public Use Files[1] for our dataset.

We used the 3 csv files from the dataset that describe the plan attributes, rates of the plan and benefits of the plan respectively. The size of the overall data is 3.394 GB.

Rates - describes the features that affect the plan rates such as age, smoking preference, dependents and state of users.

Plan Attributes - describes the plan features like child only offering, disease management programs, first tier utilization and so on.

Benefits Cost Sharing - contains details about benefit names, which year they belong to, benefits which are availed according to state code, copay, coinsurance etc. From these columns, we mainly focus on the state code, Benefit Name and Benefit year for our analysis.

### 2.2 Architecture

To perform scalable analytics and machine learning we used PySpark and the libraries it comes with as well as other visualization libraries like plotly, matplotlib and seaborn. The aforementioned dataset can be ingested into PySpark whenever the visualization and prediction needs to be performed. All of these tools are packaged into a docker container for easy distributed deployment as shown in the diagram below.

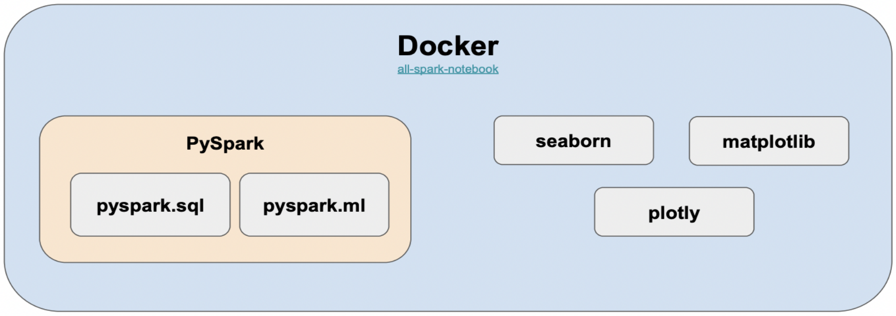


Fig. 1: Architecture Overview

### 2.3 Procedure and Results

#### Part I - Rate vs State Analysis

In this section we find the dependency of the rate of an insurance plan to the state of residence of the individual. To understand this dependency we used correlation ratio [2] compared the intra-category dispersion to the overall dispersion.

Correlation ratio is defined as $\eta^2 = \frac{\sum_x n_x(\bar{y_x} - \bar{y})^2}{\sum_{x,i}(y_{xi} - \bar{y})^2}$ or the ratio between the weighted variance of the category means and the overall variance.

When correlation ratio is 1 then we know that the dispersion is because of the difference between the states suggesting high dependence between Individual Rate and State. When correlation ratio is 0 then we know that the intra-category dispersion is all the same suggesting no dependence between the two columns Individual Rate and State.

Loaded the data from the Rate.csv using PySpark and selected the two columns which are required that is Individual Rate and State Code. Next we calculated the overall variance, overall count and overall mean using the default PySpark functions. Using these values we calculated the overall denominator in the formula for correlation ratio as $overallvariance * overallcount$.

Now, to find the numerator used the window function to split on every state and found the count and mean per state and used it along with overall mean in the formula $\sum_x n_x(\bar{y_x} - \bar{y})^2$.

Dividing the overall numerator with overall denominator we got a value of 0.003207, this tells us that the correlation ratio is low and that there is no dependence between insurance rate charged and the state that an individual is from.

Further, it may be useful to visualise the average/mean rate paid by the individuals of a state, this was done with the help of chloropleth map shown below.

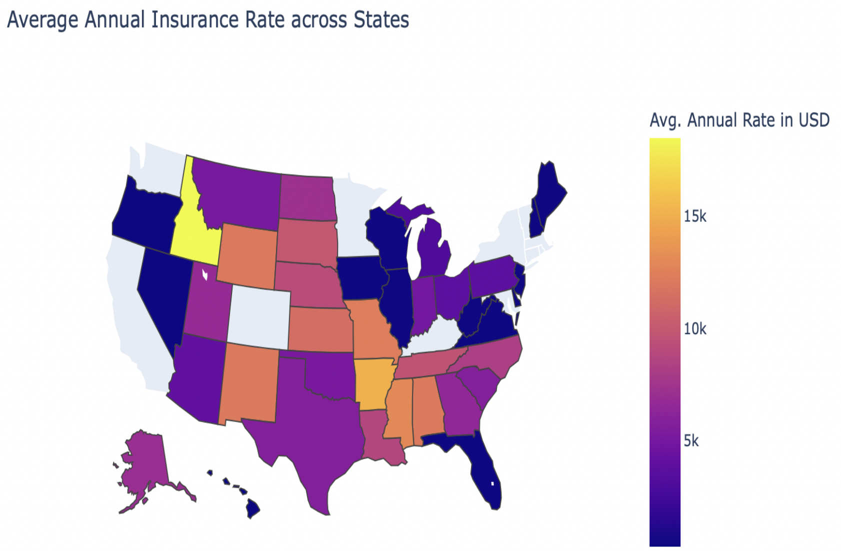


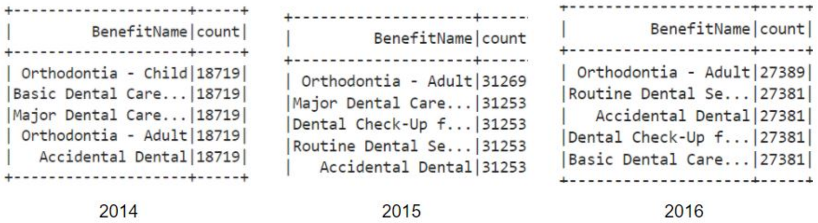
Fig. 2: Chloropleth map of Average Insurance Rate across States

## Part II - Analysis of Plan Benefits

What is a Plan Benefit?

The health care items or services covered under a health insurance plan are called Plan Benefits. These could be related to dental care, ophthalmology, neurology,

pediatry, gynecology etc. **Plan Benefits Across the years:**
We used the BenefitsCostSharing csv and focused on the columns of Benefit Name and Benefit year. We grouped the benefits by year and found the count of each of the health benefits in a year. Then we sorted this data by the count to find the top benefits in each year.



```
+--------------------+-----+
|         BenefitName|count|
+--------------------+-----+
| Orthodontia - Child|18719|
|Basic Dental Care...|18719|
|Major Dental Care...|18719|
| Orthodontia - Adult|18719|
|    Accidental Dental|18719|
+--------------------+-----+
```

2014

```
+--------------------+-----
|         BenefitName|count
+--------------------+-----
| Orthodontia - Adult|31269
|Major Dental Care...|31253
|Dental Check-Up f...|31253
|Routine Dental Se...|31253
|   Accidental Dental|31253
```

2015

```
+--------------------+-----+
|         BenefitName|count|
+--------------------+-----+
| Orthodontia - Adult|27389|
|Routine Dental Se...|27381|
|   Accidental Dental|27381|
|Dental Check-Up f...|27381|
|Basic Dental Care...|27381|
+--------------------+-----+
```

2016

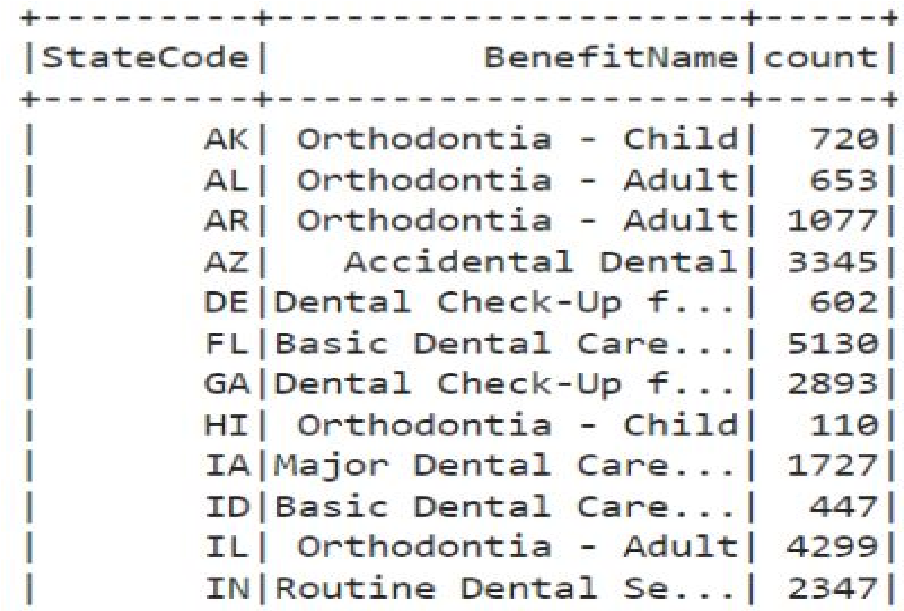
Fig. 3: Top benefits availed in each year

Fig 3. shows a snapshot of the top 5 benefits that we got for each of the years 2014, 2015 and 2016. We can see that the benefit that has been availed the most is related to some form of Dental care/Orthodontia. Thus dental care is the most sought after benefit. Insurance providers may use this information to design their insurance plans.

**Most Availed Plan Benefit per state:**
To find the most availed plan benefit per state, we first found the distinct states. This required some pre-processing of the data to get rid of states that were labeled wrongly with numerical values. After the junk values were cleaned out, we grouped the data according to the State Code and benefit name to find the benefit counts. We then used the window function to extract the benefit with the highest count from each state. This gave us the most availed benefit in each state.

Fig 4. shows a snapshot of the results obtained. For example, we can see that in state AK(Alaska), the most used benefit is orthodontia for children with 720 people using this benefit.

**State vs benefit count graph:**
In order to find the variation of the benefit count(No of benefits used) across the state, we first grouped by the state code and then found the total number of benefits associated with each state. We ordered the count of benefits in descending order so that we can see the state that uses the maximum number of benefits first Then we used seaborn to plot our graph. It has the states on the y axis and the benefit counts on the x axis.



```
+---------+--------------------+-----+
|StateCode|         BenefitName|count|
+---------+--------------------+-----+
|       AK| Orthodontia - Child|  720|
|       AL| Orthodontia - Adult|  653|
|       AR| Orthodontia - Adult| 1077|
|       AZ|   Accidental Dental| 3345|
|       DE|Dental Check-Up f...|  602|
|       FL|Basic Dental Care...| 5130|
|       GA|Dental Check-Up f...| 2893|
|       HI| Orthodontia - Child|  110|
|       IA|Major Dental Care...| 1727|
|       ID|Basic Dental Care...|  447|
|       IL| Orthodontia - Adult| 4299|
|       IN|Routine Dental Se...| 2347|
```

Fig. 4: Top benefits availed per state

Fig 5. shows the graph obtained. We can see that state WI(Wisconsin) is using maximum number of benefits, followed by TX(Texas) and the FL(Florida) The state using least number of benefits is Hawaii. This means that insurance providers are currently offering a wider range of benefits in states like Wisconsin and Texas.

### Part III - Effect of Age and Health Habits on the Insurance Plan Rates

**Effect of age of a person on the insurance plan rate:**
To find the effect of age on the rate of an insurance plan, we first group by age and plot the average annual insurance rate per age group. This showed that the average annual insurance rate was almost the same for a range of age groups as seen in Fig. 6. Thus, we separate the dental and the non-dental plans to analyse the plan rates differently in each case.

For non-dental plans, we group by age and plot the average annual insurance rate per age group. As we can see from Fig. 7, the plan rates increase as the age increases. We also find the correlation value between age and the annual insurance rate. To do this, we first pre-process the age groups to include only numeric values, which removes the groups like 'Family Option' and then cast the age column to integer. Next, we find the Pearson Correlation between age and annual insurance rate which has the value of 0.7599.

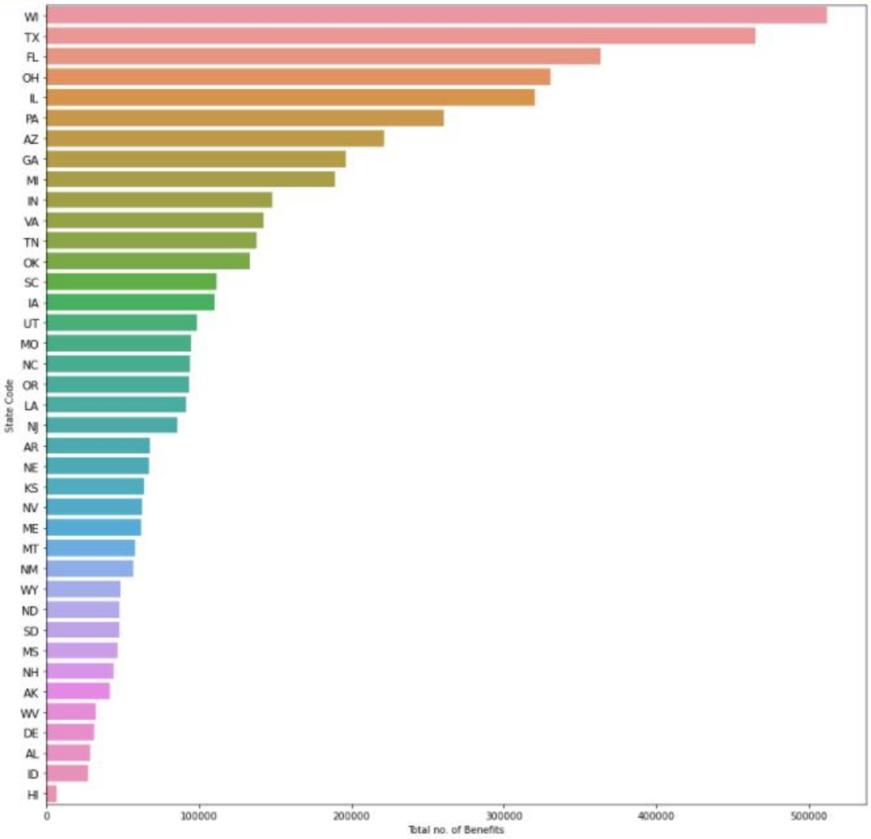


Fig. 5: State vs Benefit Count

For dental plans, we group by age and plot the average annual insurance rate per age group. As we can see from Fig. 8, the plan rates are almost a flat line for a wide range of age groups. We also find the correlation value between age and the annual insurance rate. To do this, we first pre-process the age groups to include only numeric values, which removes the groups like 'Family Option' and then cast the age column to integer. Next, we find the Pearson Correlation between age and annual insurance rate which has the value of 0.0037.

**Effect of health habits of a person on the insurance plan rate:** For analysing the effects of health habits of a person on the insurance plan rates, we consider the tobacco preferences of the person. Some plan users have particular tobacco preferences and some have no preference.For this analysis, we consider the users having tobacco preference and use the individual tobacco rate of that user to compare with the insurance plan rates. In order to do this, we

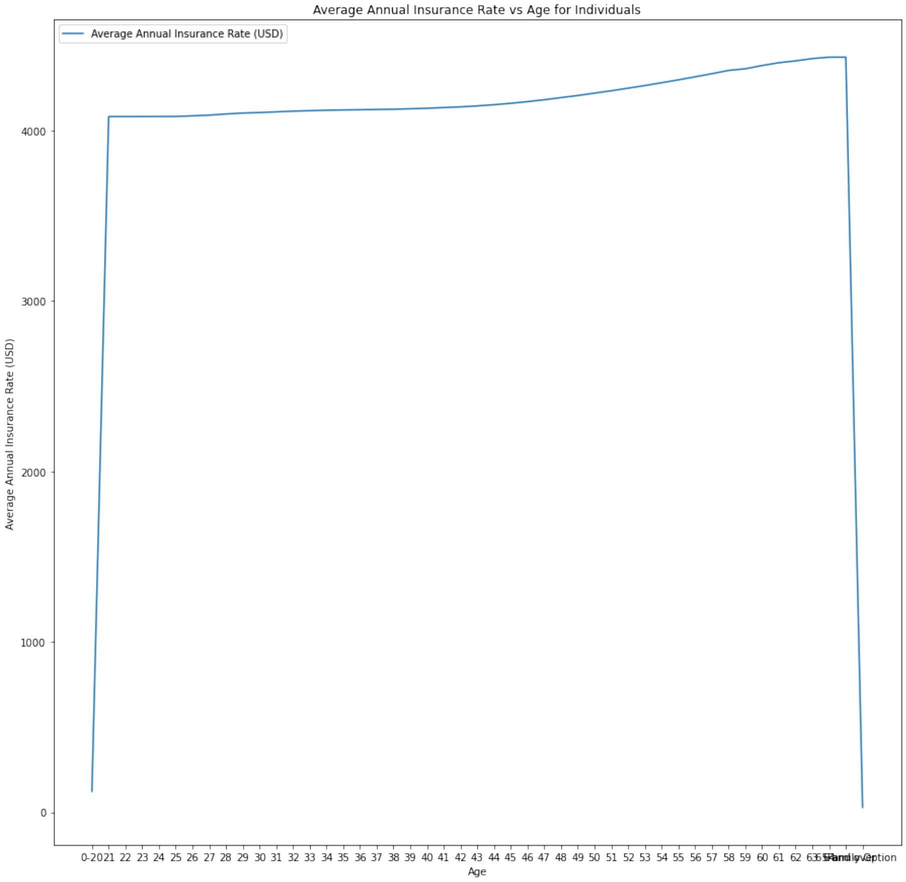


Fig. 6: Age vs Average Annual Insurance Rate

first filter the users having tobacco preference. Considering the Individual Tobacco Rate for users with tobacco preference, the correlation between Individual Tobacco Rate and Plan Rates is 0.9737 with the plan rates increasing as tobacco rates increase.

## Part IV - Distribution of Rates across Insurance Issuers

Used the window function to group by Issuer Id and calculated the mean, standard deviation and count using the default PySpark functions. These metrics gives us an idea about how spread out the data is in each insurance issuer id group. We display the issuer id groups with the most standard deviation since these groups are the ones with the most diverse portfolios as shown in Fig. 9.

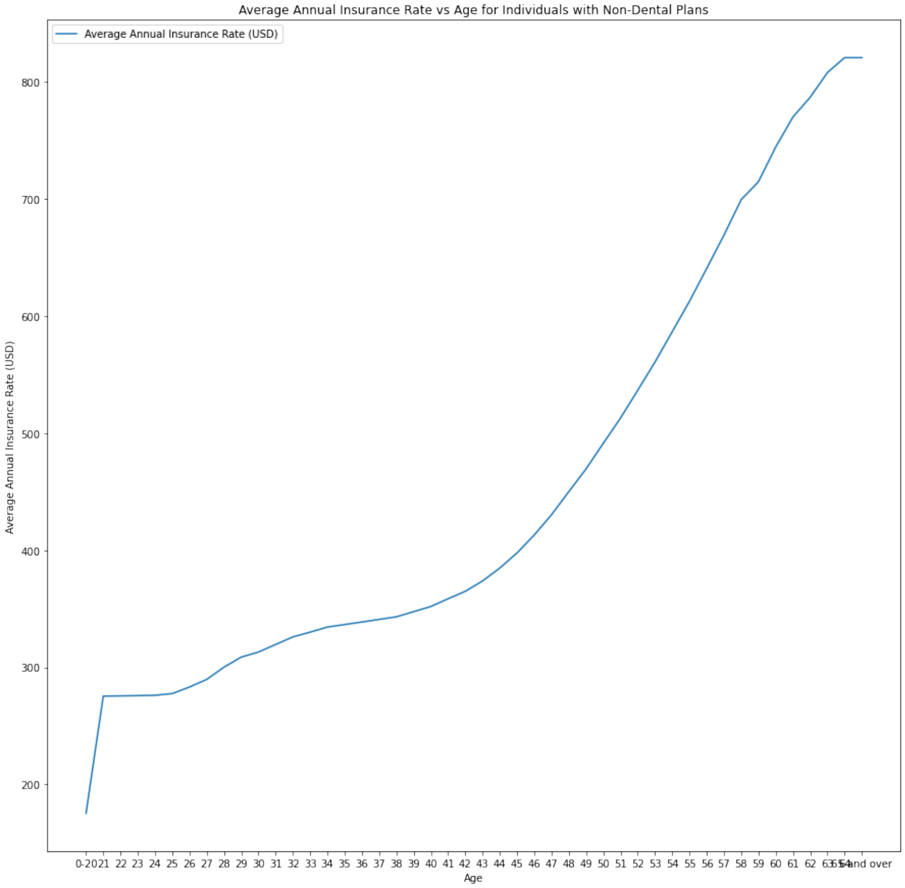


Fig. 7: Age vs Average Annual Insurance Rate for Non-Dental Plans

## Part V - Predicting Individual Insurance Rate

We created a linear regression model to predict individual insurance rate. The features we used for this model were the age and Tobacco Rate. We chose these two features since they showed a good correlation with individual insurance rate in our above analysis. We wanted to predict the Individual Rate given these two features.We used a vector Assembler on our features and then passed the vector into an ML pipeline. We did a train test split of 70-30 and then fitted our data by using the Linear regression model. We obtained an $R^2$ value of 0.94 and an accuracy of 94.95%.

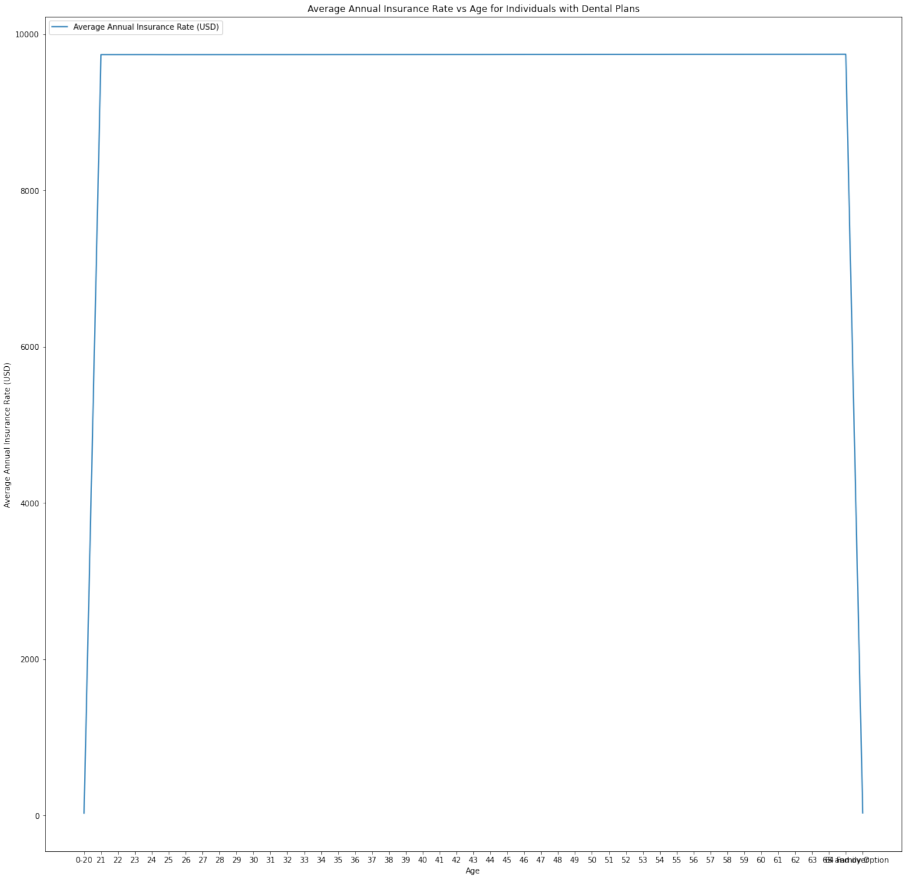
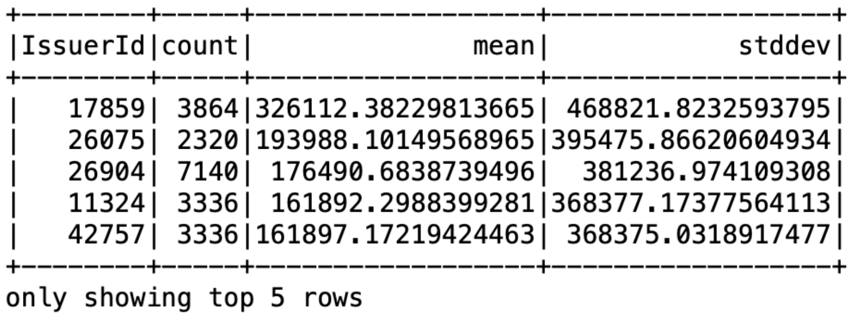


Fig. 8: Age vs Average Annual Insurance Rate for Non-Dental Plans

## 3 Overall Conclusions

Part I - Due to the low correlation ratio we can conclude that there is no dependence between Individual Rate and the State that an individual is from. Also gave a useful visualization of how average rate differs from state to state.
Part II - The most common benefit used in every state is related to dental care. This is useful for insurance providers and lets them know they should invest in dental care options.
Part III - Age and insurance plan rates are strongly dependent for non-dental plans. Insurance users can keep this in mind while setting aside money for an insurance plan. However, age and insurance plan rates don't seem to be dependent for dental plans.Also, the dental insurance plan rates are relatively higher. Thus dental insurance is sought after, irrespective of the age.



```
+--------+-----+------------------+------------------+
|IssuerId|count|              mean|            stddev|
+--------+-----+------------------+------------------+
|   17859| 3864|326112.38229813665| 468821.8232593795|
|   26075| 2320|193988.10149568965|395475.86620604934|
|   26904| 7140| 176490.6838739496|  381236.974109308|
|   11324| 3336| 161892.2988399281|368377.17377564113|
|   42757| 3336|161897.17219424463| 368375.0318917477|
+--------+-----+------------------+------------------+
only showing top 5 rows
```

Fig. 9: Distribution of Rates across Insurance Issuers

The insurance plan rates increase with the increase in individual tobacco rate. The insurance providers could look to make more affordable plans for Tobacco users.
Part IV - Displayed the issuer groups with the most diverse portfolio by seeing how the rate data is distributed across insurance issuers.
Part V - Using the result from the analysis conducted we built a linear regression recommendation system with features Age and Tobacco Rate. This model has an $R^2$ value of 0.94 and an accuracy of 94%.

## 4 Future Work

To gain more insights, we could analyze correlation between other variables to determine the best features to use to predict insurance rates more reliably. We could also use the new features to build a recommendation system for the user.

## References

1. Dataset - https://www.cms.gov/cciio/resources/data-resources/marketplace-puf
2. Correlation Ratio Wiki - https://en.wikipedia.org/wiki/Correlation_ratio