Data Visualization

Final Project Report

Project Title: Health Insurance Marketplace Data Analysis

Team:

Radhika Kulkarni and Ravinder Lambadi worked on this project starting from Raw data collection from Kaggle site, Data cleaning, Analysis, Visualization and Reporting for Health Insurance Marketplace.

Abstract :

Health Insurance Marketplace dataset consists of health insurance and dental plans offered through healthcare.gov between 2014 and 2016. It encompasses rates for smokers and non-smokers, separately listed for each age group, benefits included in the plans, states in which the plans were offered, and other information. Steps Followed are data cleaning from given raw data, Analysis of the plan rates with respect to smokers and non-smokers, age groups and U.S. states.

Introduction:

The project analyzes health insurance and dental plans offered through healthcare.gov. It encompasses rates for smokers and non-smokers, separately listed for each age group, benefits included in the plans, states in which the plans were offered, and other information. The analysis steps includes the processing of raw data from Kaggle datasets and further Cleaning, Analysis and Visualization.

Data Analysis of the Health Insurance Marketplace Public Use Files:

https://www.cms.gov/cciio/resources/data-resources/marketplace-puf.html

The files contain information on health insurance plans for 2014, 2015 and 2016 that were offered through www.healthcare.gov. The data is made available through the US Department of Health and Human Services on Kaggle:

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

This project contains a Jupyter notebook with a Python/Pandas analysis of the files Rate.csv and PlanAttributes.csv, which can be downloaded from the Kaggle website. The files contain the raw data combined for 2014, 2015 and 2016.

The script performs a thorough cleaning of the dataset, extraction of monthly premiums for individuals, and comparison of average rates across the US.

Research Questions And Working Hypotheses:

Project problem is very important because it analyses significant increase in premiums for non-smokers in the years 2014-2016. In the same period, the median rates for smokers. Also it analyses median health insurance rates for non-smokers (and smokers as well) between the states that participated in the health insurance marketplace. I have taken a special case of Montana state for detailed analysis.

The project will try to find the answers for few of the below questions based on Analysis with some data explanation.

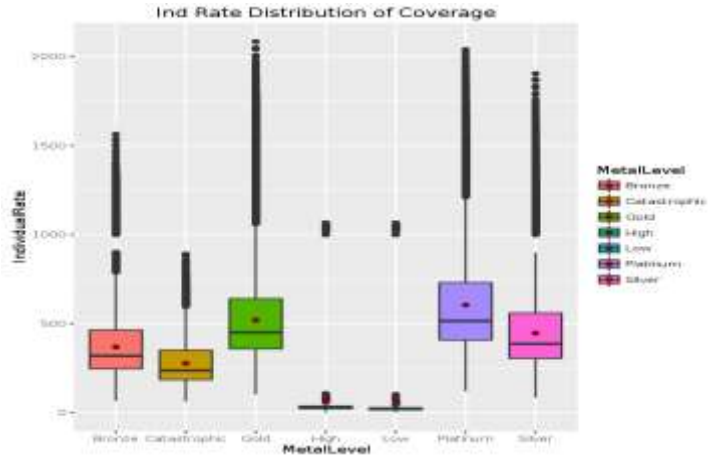
- Project analyses significant increase in premiums for non-smokers in the years 2014-2016. In the same period, the median rates for smokers.
- It helps to analyze median health insurance rates for non-smokers (and smokers as well) between the states that participated in the health insurance marketplace.
- Median plan rates increase for about 25 states that participated in the marketplace between 2014 and 2016.Also a decrease observed.
- It also helps to identify where the Health insurance for non-smoking individuals is cheapest and most expensive. Also, the median rate across the states.
- The large spread in premiums in states are real or not . What are the reasons for it and in all other states, this type of surgery is covered by all plans or not and how the cost for these surgeries is distributed among all payees
- What are the low-cost plans
- Which are the states where Health insurance for non-smoking individuals is cheapest or Highest.

Background And Related Work:

The existing analysis use filed Rate and BenefitsAttributes files and focused on the plan year of 2015 Individual plans only.
he median monthly premium distribution gives a brief overview of the monthly premium being offered by state. It shows a quite wide range of the median premium range. That inspires a series of research questions.

I. Plan Coverage Type

Plans in the Health Insurance Marketplace are presented in 4 "metal" categories: Bronze, Silver, Gold, and Platinum. The boxplot of premium distribution by the metal coverage categories shows the difference in premium levels by plans.

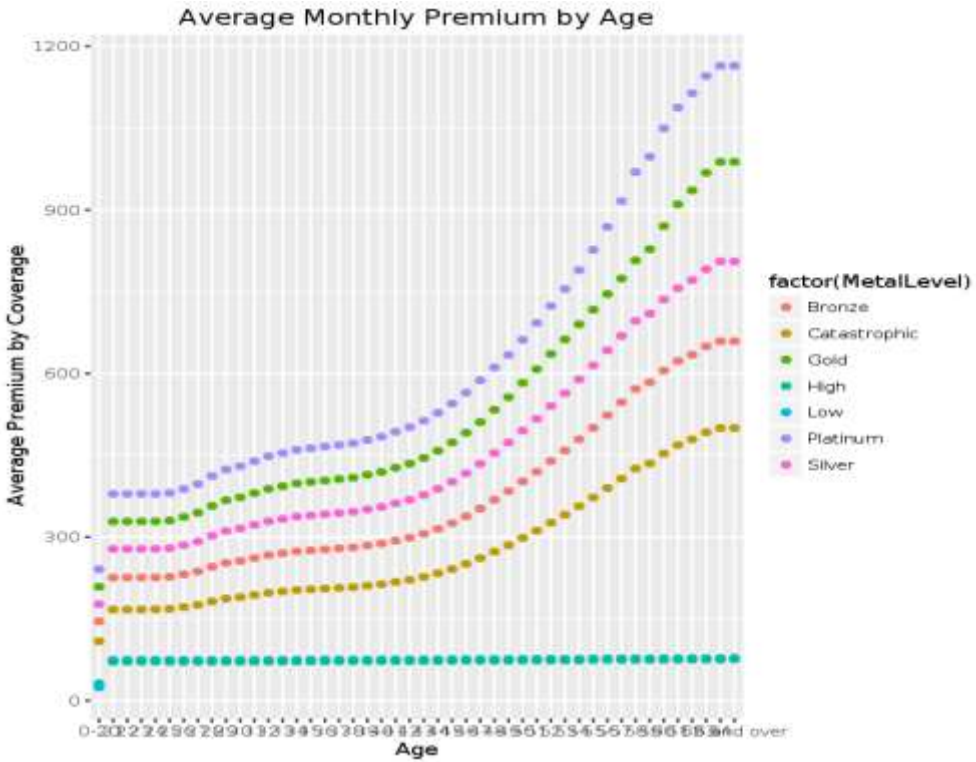


II. Plan Premium By Age

The next questions to assess is how the premium varies with the increase of age. Intuitively the older you are, the more risk you potentially carry for any health related issues. Therefore, an upward trend is expected in their case.

III. State of Residency

The next thing to inspect is whether the state residency will make a significant different in the premium level too.



This analysis confirms that there are at least three variables that affects premium levels: benefit type (metal level), age, and state of residency.
Reference: *https://nycdatascience.com/blog/student-works/2015-health-insurance-marketplace-data-exploration/*

This analysis lacks idea on plan level detailing. Also it lacks the proper visualization for special cases and over the states.
Willing to work on similar lines ,I would analyze the data further for smokers and non-smokers category across all states to identify where the Health insurance for non-smoking individuals is cheapest and most expensive. Also the median rate across the states.
We are analyzing Median plan rates increase for about 25 states that participated in the marketplace between 2014 and 2016 and the count of decrease observed.

DataSet used: U.S. Department of Health and Human Services Data (https://www.kaggle.com/hhs/health-insurance-marketplace)

Process:

It includes Data Cleaning ,Analysis and ,Visualization using Python

Data Cleaning Process:

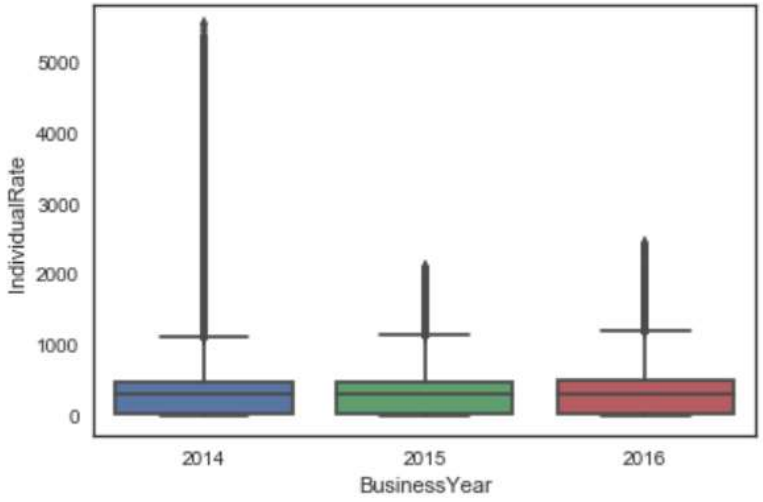
Clean-up step includes typos, non-sensical unique values, NaN' values. Identifying unique values in the columns that contain categories to see if there are any strange values

From the complete data ,I am more interested in columns below after looking at the data.
BusinessYear, StateCode, Age, PlanId, IndividualRate, IndividualTobaccoRate and all the family/couple rates. Data is further broke down into individual U.S. states:

I. Number of states for which data is present: 39 states
II. Number of health insurance plans: 16808
III. From the states data,is clear that not all states make use of the federal network healthcare.gov. as some states have their own health insurance marketplace eg: NY
IV. Some states offer significantly more plans than others. This may be due to the different sizes of the states and not all states have offered plans through healthcare.gov in for all three years.

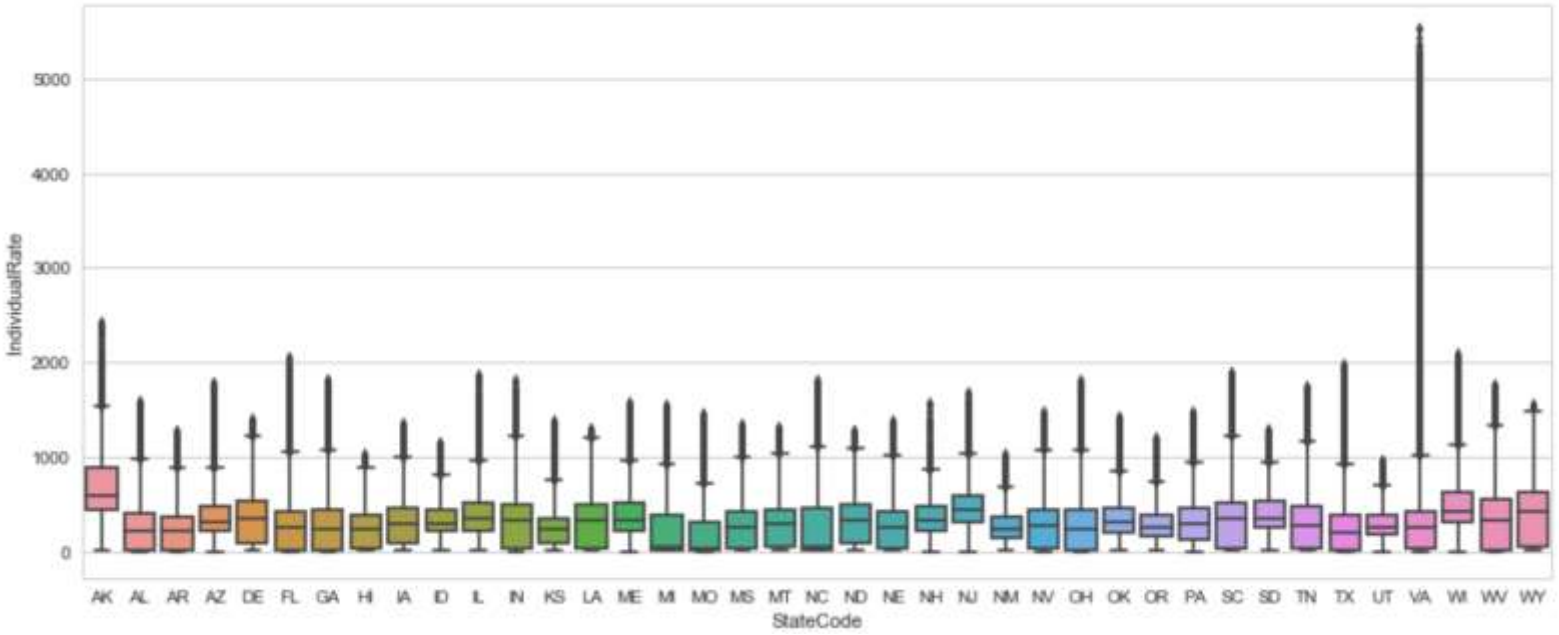
The raw dataset was quite clean already. Only about 250,000 rows in the dataframe contained (for our purposes) redundant data. For the further analysis,we will focus on 15 of the 24 columns.

Let's have a look at trends over the years.



- ***Analysis***

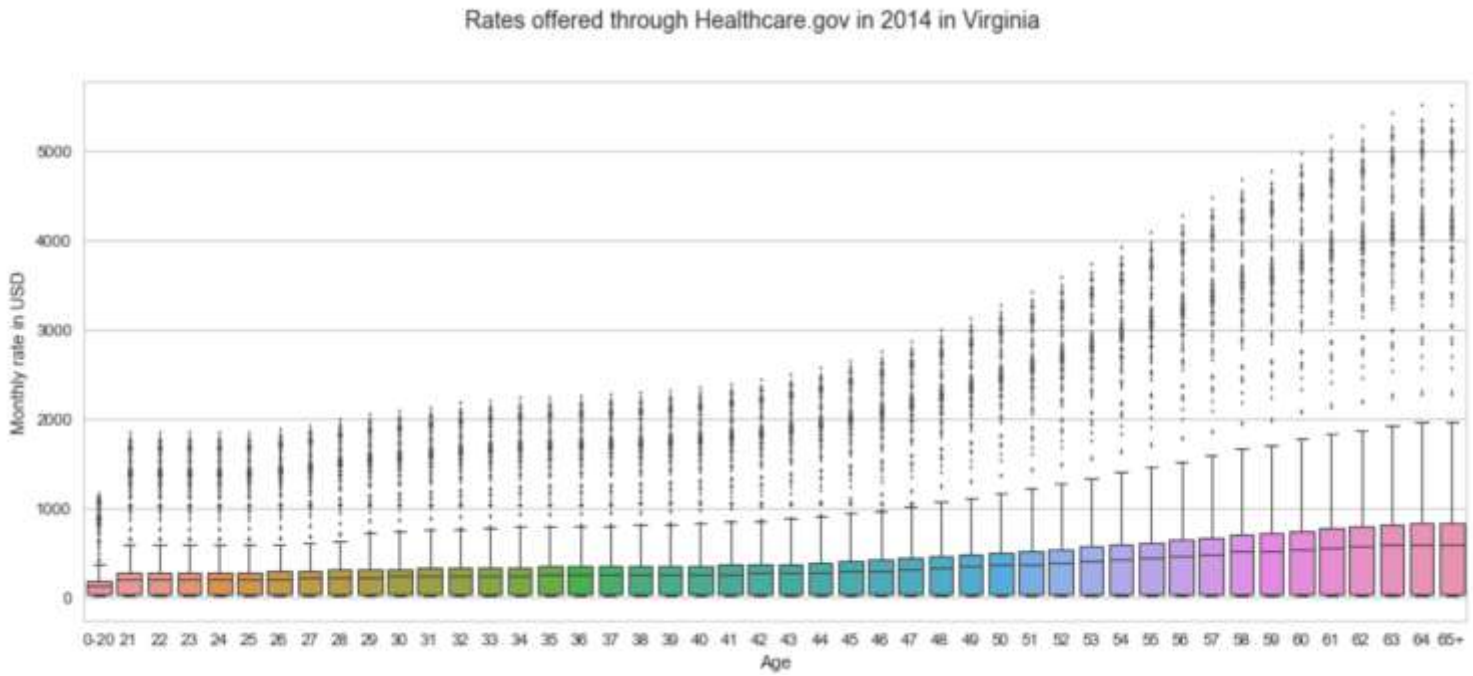
  Looking at the end data in 2014,these outliers make the boxchart unreadable. Need to break it down by states to narrow down the outliers.



- ***Analysis:***

Virginia is the one state that's causing the outliers in the 2014 statistics. Taking look at VA in 2014 in more detail.

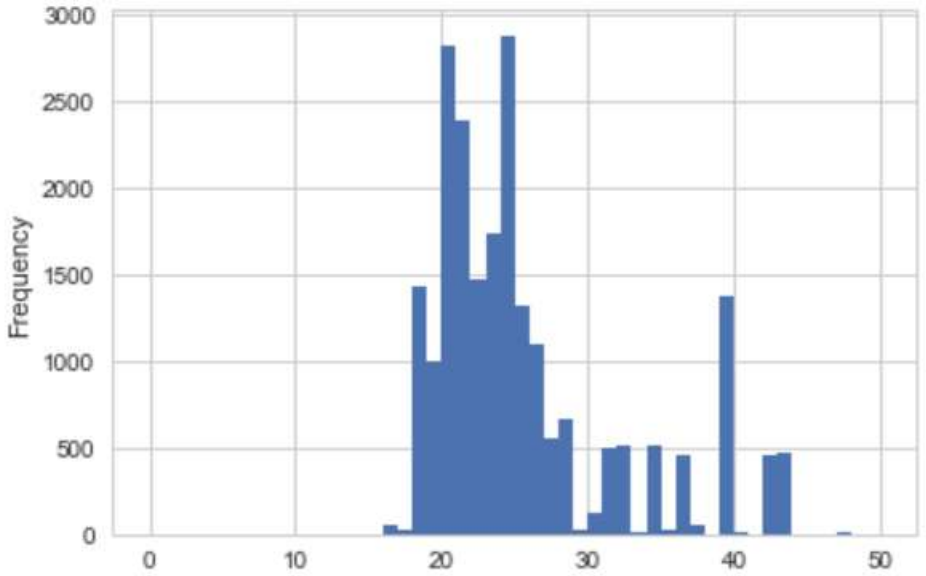
Analyzing the case of Virginia Further



- ***Analysis***

  The large spread in premiums in Virginia is real. After some internet research it turns out that the wide range in monthly premiums in Virginia is due to plans covering gastric bypasses! In all other states, this type of surgery is covered by all plans, so the cost for these surgeries is distributed among all payees (http://www.webmd.com/health-insurance/20131011/why-some-virginia-health-plans-cost-so-much).

I will take another case of Montana where there are really low premiums.



A summary of the Montana 2014 dataframe

- ***Analysis***

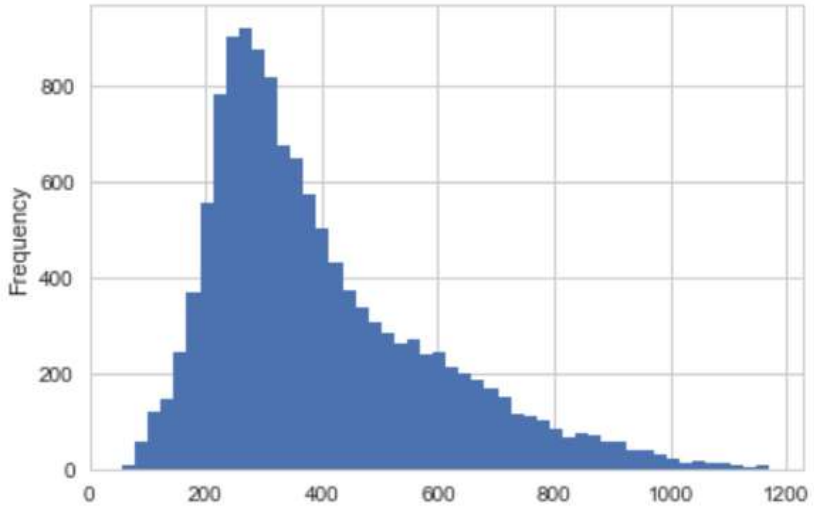
  It's interesting that all couple/family plans have a price range comparable to the low-price bump in the individual rates.The low-cost plans are actually dental-only plans. Excluding them if we want to say anything about health insurance premiums.We can read in another CSV file that can give us some insights on which is which.

Distinguishing between full plans and dental-only plans

- ***Analysis***

One third of all listed plans are dental-only plans.
The goal was to remove these dental-only plans so that we can get some statistics on the prizes of health insurance plans for individuals. So let's get rid of the dental-only plans!

The cleaned datasets for full health insurance plan rates (the case of Montana continued)¶



Calculating the median of this asymmetric distribution

Analysis

The median monthly premium in 2014 in Montana was about $340. That's a reasonable number!

Results and insights:

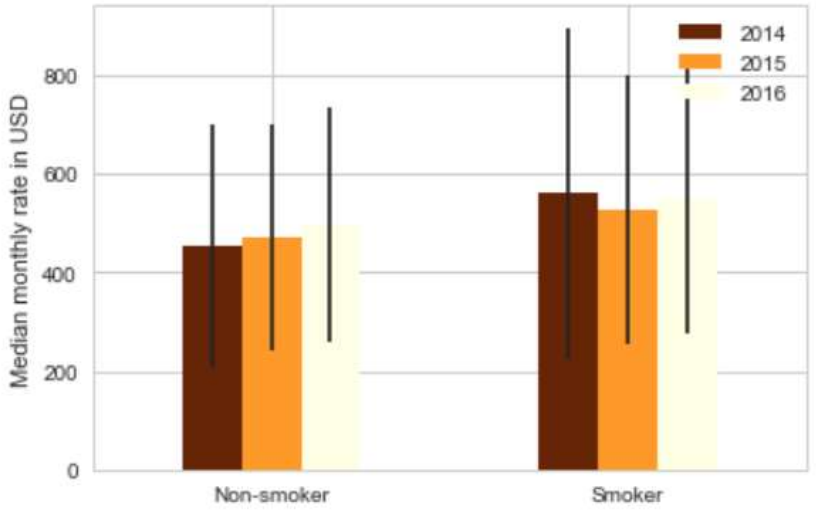
Breaking the whole dataset down by year to see some trends on the individual types of rates

	Unnamed: 0	BusinessYear	IndividualRate	IndividualTobaccoRate	Couple	PrimarySubscriberAndOneDependent	PrimarySubscriberAndTwoDependents	P
count	2.370659e+06	2370659.0	2.370659e+06	1.601995e+06	0.0	0.0	0.0	
mean	1.870614e+06	2014.0	4.524658e+02	5.601592e+02	NaN	NaN	NaN	
std	1.066630e+06	0.0	2.468284e+02	3.345296e+02	NaN	NaN	NaN	
min	1.398000e+03	2014.0	4.907000e+01	5.566000e+01	NaN	NaN	NaN	
25%	1.055422e+06	2014.0	2.902200e+02	3.412000e+02	NaN	NaN	NaN	
50%	1.815117e+06	2014.0	3.872100e+02	4.759600e+02	NaN	NaN	NaN	
75%	2.754092e+06	2014.0	5.810200e+02	6.953100e+02	NaN	NaN	NaN	
max	3.656258e+06	2014.0	5.503850e+03	6.604610e+03	NaN	NaN	NaN	

Analysis

All the couple and family rates were dental-only plans. Maybe that was clear to anyone else.

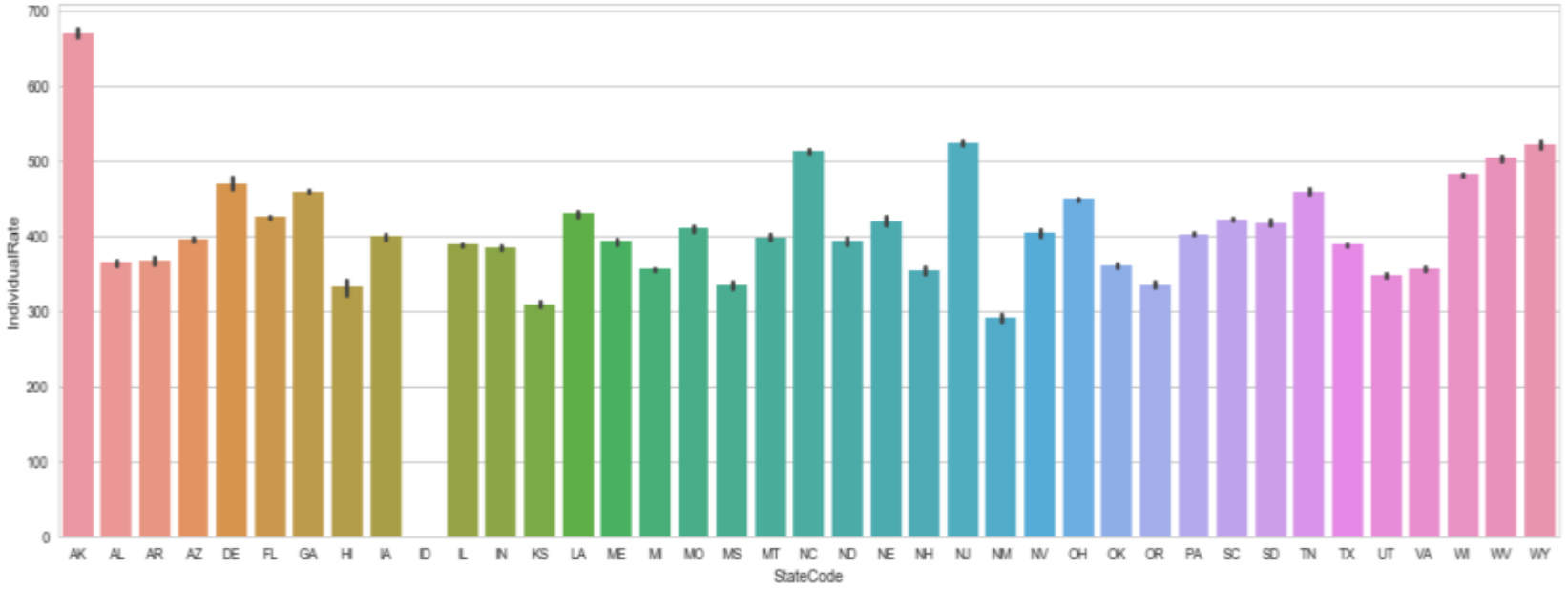
Comparing smoker and non-smoker rates then! We can start with simple statistics such as median, mean and standard deviation



- ***Analysis***

There has been a significant increase in premiums for non-smokers in the years 2014-2016. In the same period, the median rates for smokers remained roughly constant

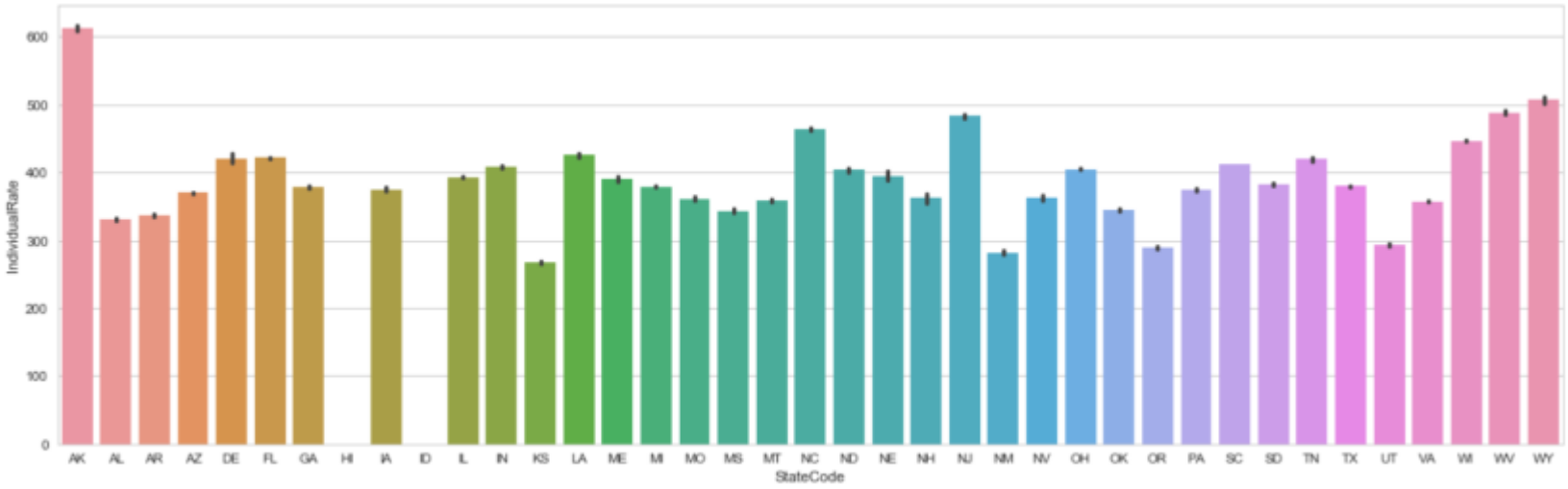
Let us look at the individual rates for each of the three years broken down by states. Did the prizes go up in all of the states? We can use Seaborn's barplot to show the data for the 38 states. It offers us the possibility to select an estimator for a quantity that we want to compare across categories (which we can set to numpy's median) and even includes a bootstrapping routine like the one we used above.



- ***Analysis***

  Median health insurance rates for non-smokers (and smokers as well) vary strongly between the states that participated in the health insurance marketplace.

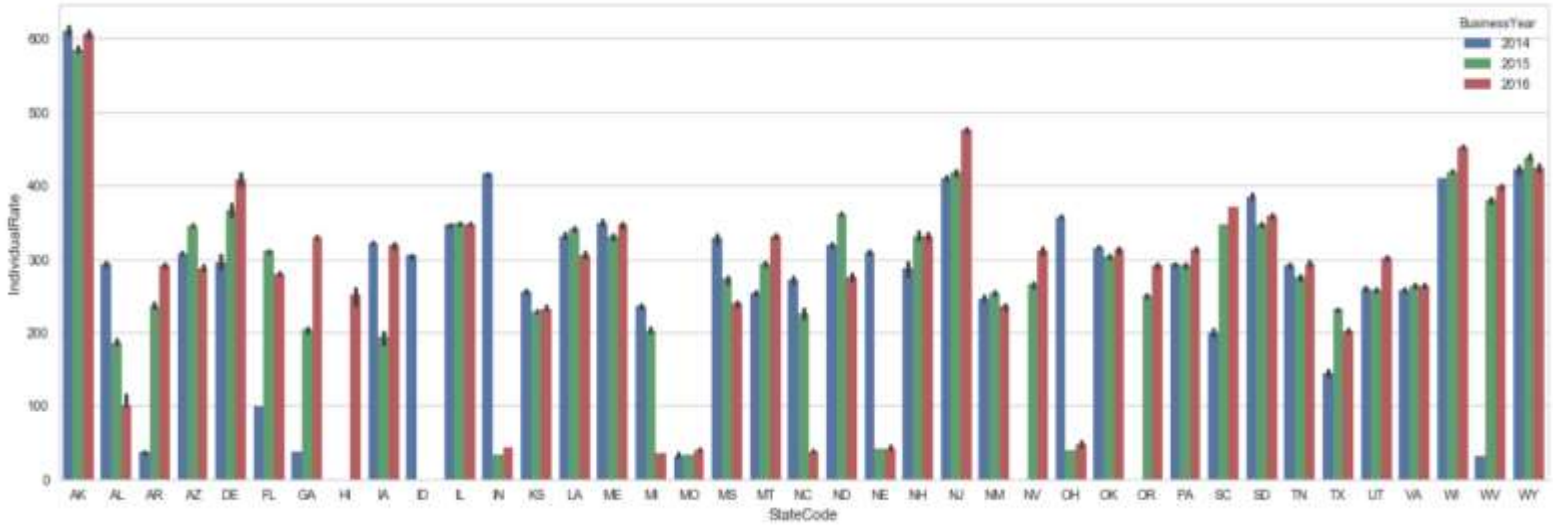
Let's also look at the other years. 2015:



- ***Analysis***

  Not all states participated in the federal health insurance marketplace in all three years.

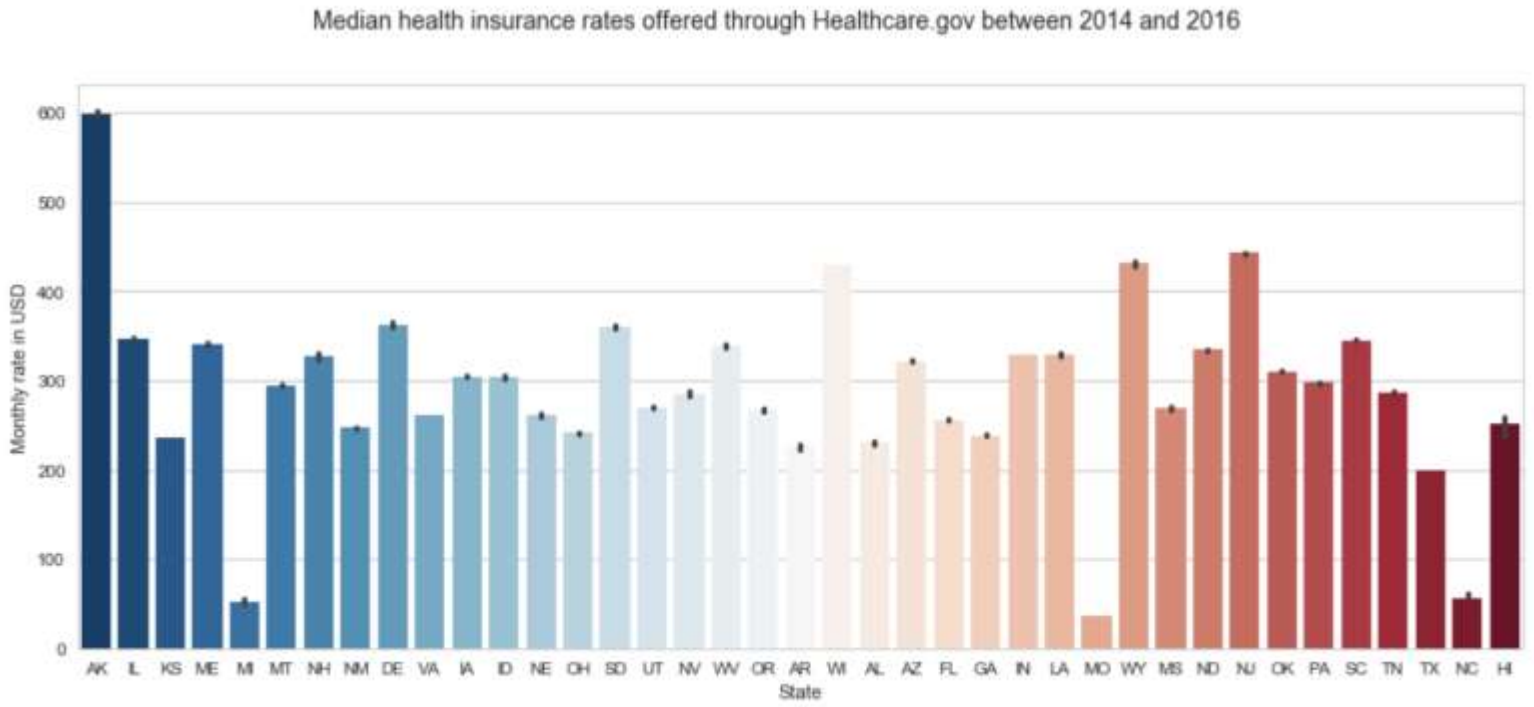
  Other than that, these three plots don't tell us much. We can combine all of them into one figure to see how the median rates changed with the years.



- ***Analysis***

Median plan rates significantly increased for about 25 states that participated in the marketplace between 2014 and 2016. A decrease can be observed in about 8 states.

Making a sorted comparison plot of median, non-smoker plan rates across all states that participated in the program.



Analysis

Health insurance for non-smoking individuals is cheapest in Kansas and most expensive in Alaska. The median rate is about twice as high in Alaska as it is in Kansas, New Mexico, Utah, Oregon or Hawaii. *

Conclusion And Future work:

From the analysis of this research ,we have observed the major results as :

- The low-cost plans are actually dental-only plans. Excluding them if we want to say anything about health insurance premiums.

- One third of all listed plans are dental-only plans.
    - The goal was to remove these dental-only plans so that we can get some statistics on the prizes of health insurance plans for individuals. So let's get rid of the dental-only plans!

- The median monthly premium in 2014 in Montana was about $340. That's a reasonable number!

- There has been a significant increase in premiums for non-smokers in the years 2014-2016. In the same period, the median rates for smokers remained roughly constant

- Median health insurance rates for non-smokers (and smokers as well) vary strongly between the states that participated in the health insurance marketplace.

- Not all states participated in the federal health insurance marketplace in all three years.
    - Other than that, these three plots don't tell us much

- Median plan rates significantly increased for about 25 states that participated in the marketplace between 2014 and 2016. A decrease can be observed in about 8 states

- Health insurance for non-smoking individuals is cheapest in Kansas and most expensive in Alaska. The median rate is about twice as high in Alaska as it is in Kansas, New Mexico, Utah, Oregon or Hawaii.

- **Future work** may mainly focus on why in some states Health insurance rates for non-smokers are very high like Alaska .Can there be any steps taken to reduce it so its reasonable for general population ?

References:

The Center for Consumer Information & Insurance Oversight
https://www.cms.gov/cciio/resources/data-resources/marketplace-puf.html

Healthcare site reference:
https://www.healthcare.gov/

Kaggle Dataset :
https://www.kaggle.com/hhs/health-insurance-marketplace

Python:
https://docs.Python.com/manual/tutorial/install-Python-enterprise-on-red-hat/

Python libraries for graph plotting:
https://matplotlib.org/users/pyplot_tutorial.html