# **Health Insurance Exchange Analysis**

## **Introduction:**

Health insurance matters to all of us. Most of the United States population have healthcare insurance but not till 2010. There were more than 16% people in United States who were not insured. On an average, US Citizens spends more than $1000 on monthly premiums for family coverage. Such a high premium leads people to not to opt for the Health Insurance which makes them vulnerable to financial breakdown in case of medical emergencies. To mitigate such problems, the **Patient Protection and Affordable Care Act** (**PPACA**), often shortened to the **Affordable Care Act** (**ACA**) or nicknamed **Obamacare**, was introduced. **ACA** 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 database is available online, I can take full advantage of this resource to analyse and try to unravel the hidden invaluable information. The insurance companies always keep pricing decision in a black box, so by this study I will try to discover those patterns using numerous methodologies of Exploratory Data Analysis, Data Visualization and Predictive Analytics. I try to predict the rates of the plans for the next year and suggests the best plan to enroll in as per the individual needs.

## **Data Acquisition:**

The Centers for Medicare & Medicaid Services (CMS) Center for Consumer Information and Insurance Oversight (CCIIO) is committed to increasing transparency in the Health Insurance Exchange. While health plan information including benefits, copayments, premiums, and geographic coverage is publically available on Healthcare.gov, CMS also publishes downloadable public use files so that researchers and other stakeholders can more easily access Exchange data.

The Health Insurance Exchange Public Use Files (Exchange PUFs) are available for plan years 2014 to 2019 to support timely benefit and rate analysis. The link to the dataset is below:

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

<http://www.nber.org/data/cms-marketplace.html>

The Exchange PUFs consist of ten separate files but for our study I will be concentrating on below files:

1. **Plan Attributes file** : Plan-level data on essential health benefits, coverage limits, and cost sharing
2. **Rate File** : Plan-level data on individual rates based on an eligible subscriber's age, tobacco use, and geographic location
3. **Benefits and Cost Sharing file**: Plan-level data on maximum out of pocket payments, deductibles, cost sharing, HSA eligibility, formulary ID, and other plan attributes

All other files like 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 contains plan-level data on individual rates based on an eligible subscriber’s age and has over 12 million data entries. The BCS table contains plan-level data on essential health benefits, coverage limits, cost sharing (copay and coinsurance), and so on. 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, so I must manually select key variables based on our understanding of the healthcare industry.

## **Problem Statement:**

There are number of plans available in the marketplace which are having different benefits and pricing for different states. It’s a nightmare for the member to choose and buy the right kind of plan which can serve more benefits with lesser premium. So, my study will help in making the decision to choose the right plan for the individual and predict the plan rate in future.

During exploratory analysis, I will try to answer how do plan rates and benefits varies across states, how do plan benefits relate to plan rates, how do plan rates relate to age, how do plan vary across insurance network providers?

## **Data Wrangling:**

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.

I have merged the data for different years and different types of data in one data frame, so that I can analyze all the data with the single source. It will require data merging and reshaping techniques (Split-Apply-Combine).

There are few data columns which were not captured from the year 2017 onwards which make me to read the manual so that I can identify all the common Columns in all the years which helps me to merge them together.

I have read all the 3 files from the website for all the year 2014-2019 and merge them separately. The major problem with he files like Plan rates and Benefits and Cost sharing, is they are very huge and takes time to process. So, I have kept them separately and wrangle them individually.

In all files I must deal with all kind of Data cleaning process and Imputing methodologies to make it clean enough to do Visualization and Analysis. The common to deal with are:

* + - * **Missing data (NaN and Spaces)** :

To handle them, it was a logical imputation rather than just deleting them.

There are numerous ways that you can identify whether the dataset has Missing data or not.

I have used the below way to do it:

1. DataFrame.info() will give you an idea how many NaNs are present in the dataframe.
2. After that, I have used .isna() fucntion to get the data of all the missing values in the dataframe

To fill the data with the logical data I have used the below methods:

1. *fillna()* method: There are lot of options available with the fillna() method to deal with the missing values.

**For example:** *fillna(0)* will replace all the NaN values with 0(zero) 2.

1. *dropna()* method: to drop the entire data point if there are missing values in all the columns.
   * + - **Outliers :**

For such kind of data, I majorly deleted them as they can interfere with the Algorithm and can mislead the prediction algorithm. Outliers in input data can skew and mislead the training process of machine learning algorithms resulting in longer training times, less accurate models and ultimately poorer results.

* + - * **Duplicate Rows:**

Removing duplicates is necessary as it can lead the model to overfit and will work badly on the unseen data.

There are columns in the data frames which has floating numbers representing the amount with “$” sign as prefix. For my analysis I had to change the column as numeric so that I can perform mathematical calculation on the columns. To convert these kinds of values to numerical values I must get rid of the “$” sign.

To do this I have used replace function on data frame column name and then pandas *to\_numeric* function to convert the object datatype into numeric datatype.

Similar kind of conversion I did on Age column of the rate data frame, which has the values as object datatype and with the text values as “0-20” where I have to get rid of “-“ and covert it to numerical values.

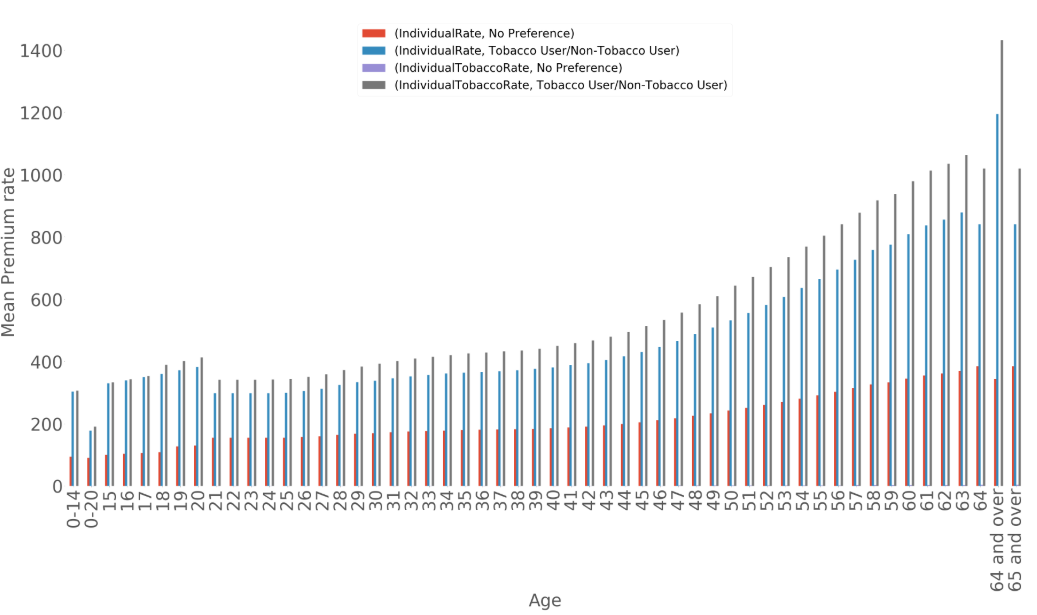
Also, *RatingAreaId* column has the similar values, like “Rating Area 10”, in these types of values I have to get rid of space and then convert these areas into numbers. As these data will require in doing prediction of rates so I need to convert them in numeric quantity.

Below are few analyses that was done to understand the behavior of data points.

1. **Rate vs Age vs Tobacco:** Here I study he behavior of Plan rates with respect to the age for Tobacco/Non-Tobacco users.

The grey line is almost the double of the red line in the below graph. It suggests that **as age increases the rate difference will be higher between Tobacco user Individual rate and Tobacco user Individual tobacco rate.**The premium rate for Tobacco user is more than double of Non Tobacco user. This is a good incentive for living a healthy life.

This clearly suggests that these 2 populations are different and could be an outlier for each other. While applying Machine learning algorithm we have to keep this in consideration.



1. **How are plans being offered in the subsequent years ?**

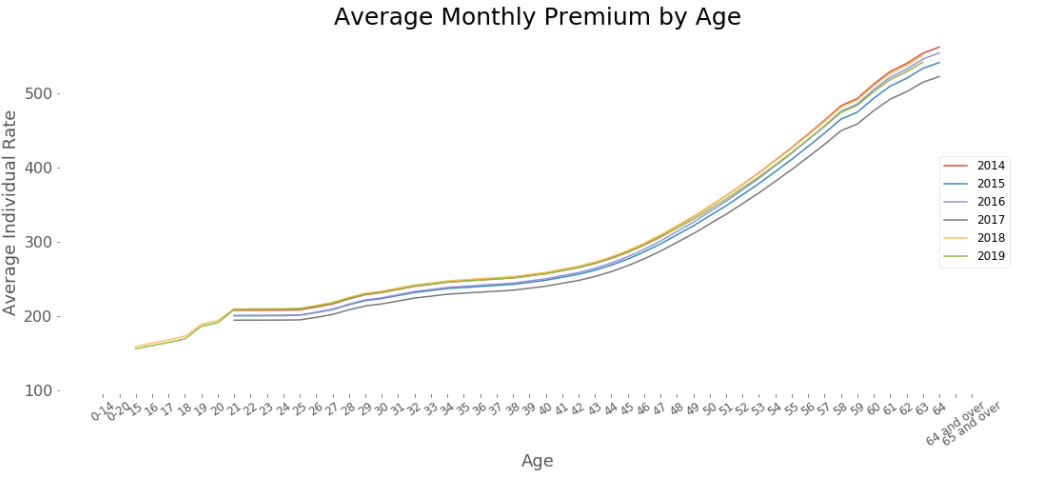
There was a huge change in the number of plans offered in the marketplace from the time of its inception. The below graph suggests the same story.



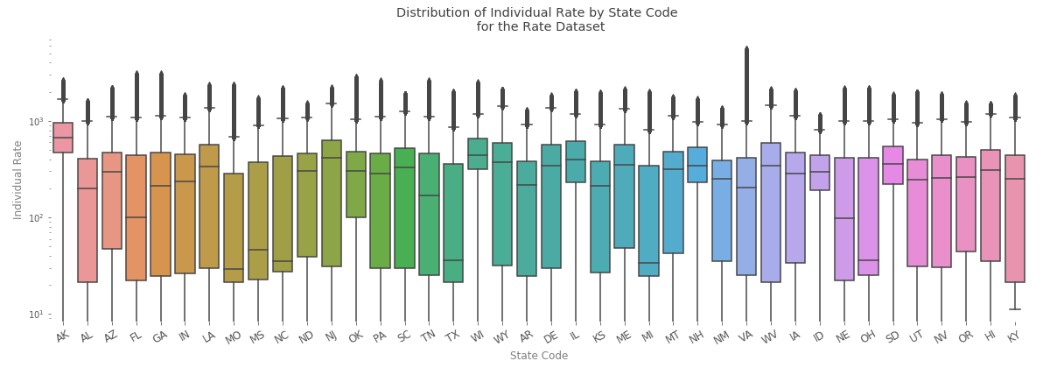
1. **Individual rates variations with Ages in all years**

In the year 2014 to 2017, there was one group of 0-20 age. But from 2018 the age group is reduced to 0-14 and Kids older than 14 are covered differently. Also we can see that upper limit of age is changed from '64 and Above' to '65 and above' after 2018.

Average premium mostly remains same throughout the years. Premium increases with increase in Age, which is normal in health Insurance industry.

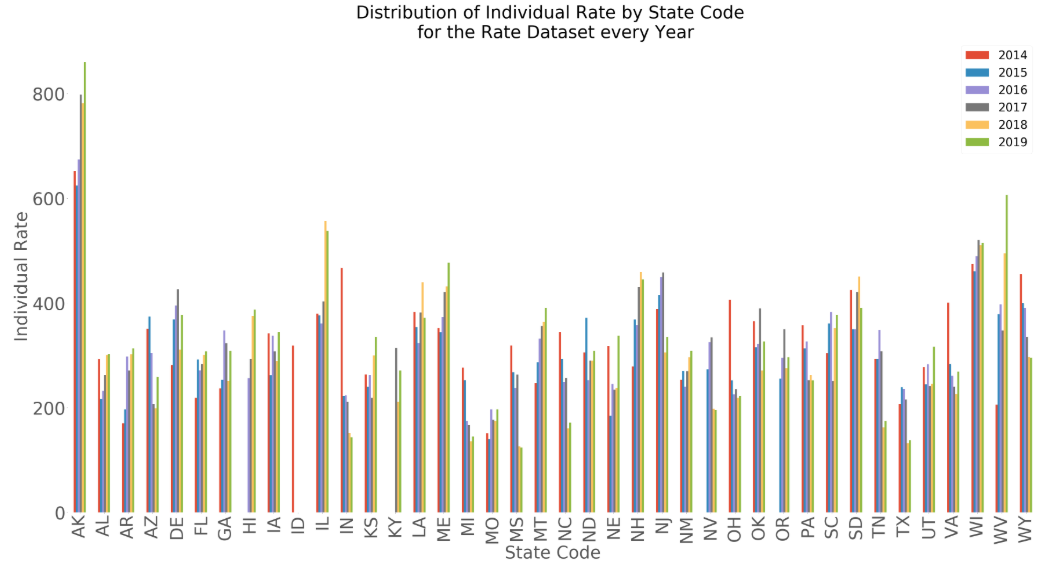


1. Distribution of individual rate across all the states.

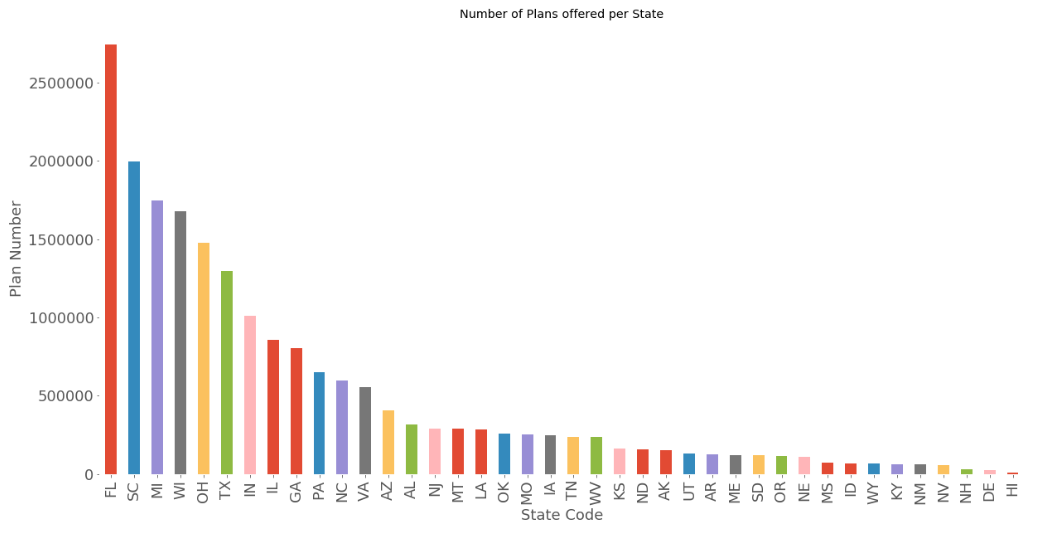


**Observations from the above graph:**

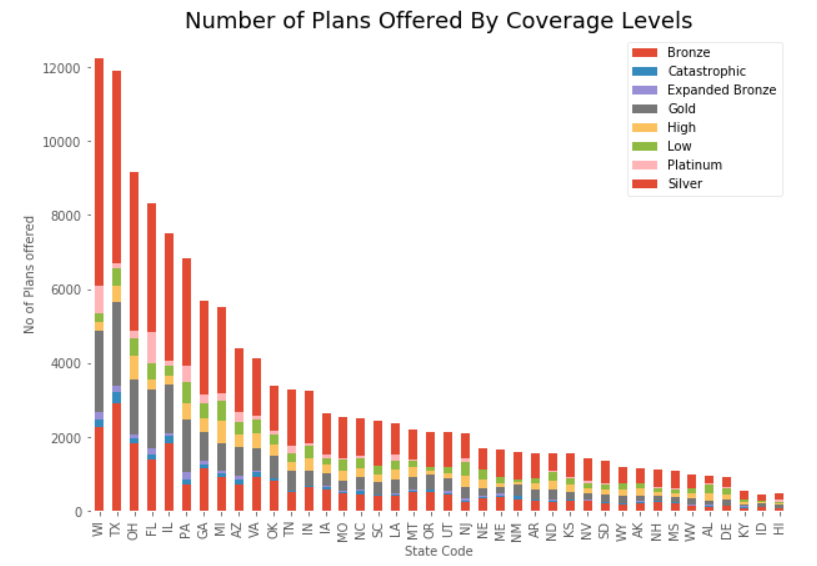
1. Out of 50 states only 40 states have participated in this exchange program
2. There are big differences in insurance premiums among the states. It's clear that the individual rates in Alaska and Illinois are very high.
3. The Median of Alaska is also very high, which means that there are no Plans which comes cheaper in this state
4. Tennessee and Texas are the more reasonable states with regards to Individual Plans.
5. Mean Individual rate per state code per year also tells the same story:
6. Alaska is always a very expensive state.
7. Illinois is again expensive as compare to other states
8. Wyoming is becoming expensive in 2019.



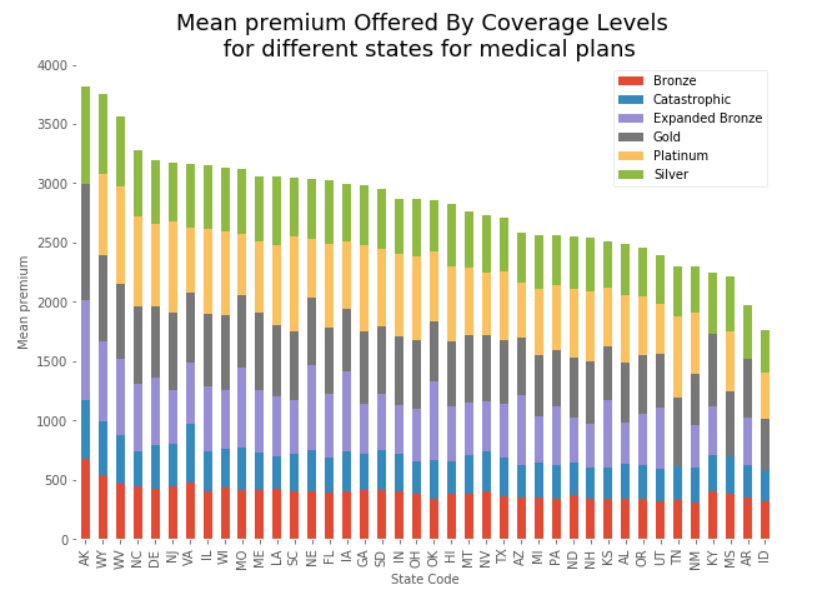
1. The above analysis allows us to think why few states are so expensive than others. Lets check the below graph to answer it. I have checked the number of plans offered in each year, and it gives us a very interesting story. There is huge difference among the states in plan offerings.

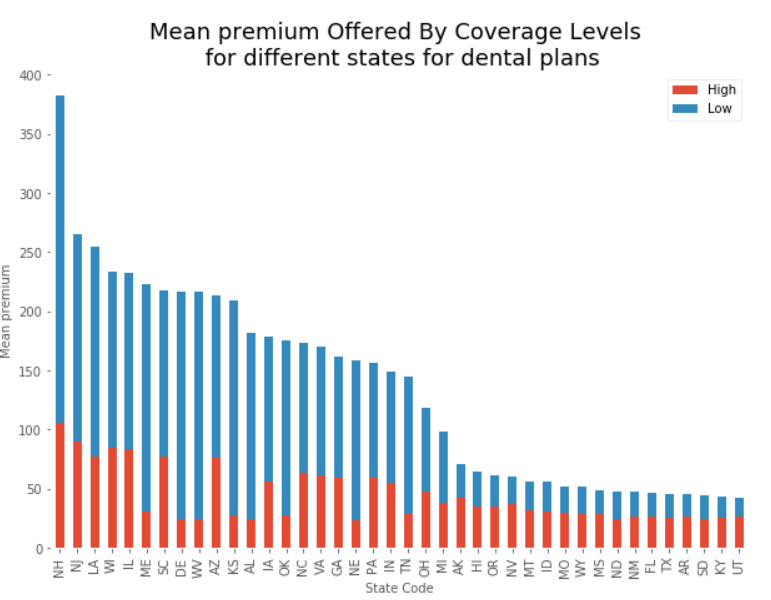


1. Florida has maximum numbers of plans offerings for the population
2. Alaska, Illinois and Wyoming have very less offerings for the plan.
3. We can see some negative co-relation with Number of plans offered with monthly premium in a State
4. There is a difference in offerings of the plans in different states but the ratio of plans in different categories remains same. As per the below graph, all states offer Bronze category plans the most. Bronze covers all the mandatory benefits as suggested by the ACA act but doesn’t provide whole lot of benefits. So, it is a cheap plan with all but few benefits.



1. Now I have combined the plan and rate files together based on Plan ID, state code and Business year. We have the indicator in the plan file which suggests whether the plan is Dental or Medical. So, I have analyzed them differently to see how their rate distribution is among all the states.

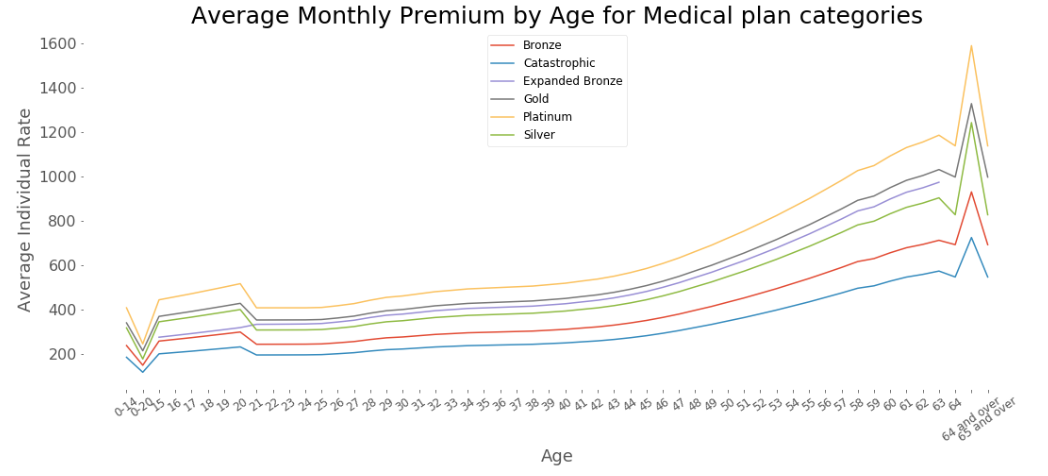


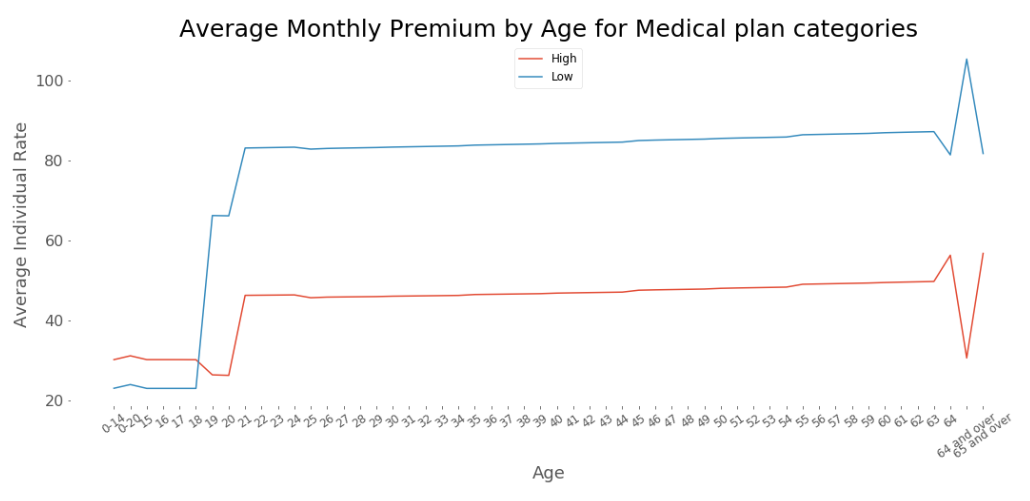


We can see that maximum premium is charged by Alaska in Medical plan and Cheapest plans are offered by Idaho

In Dental plan NH seems to be charging way too high than other states.

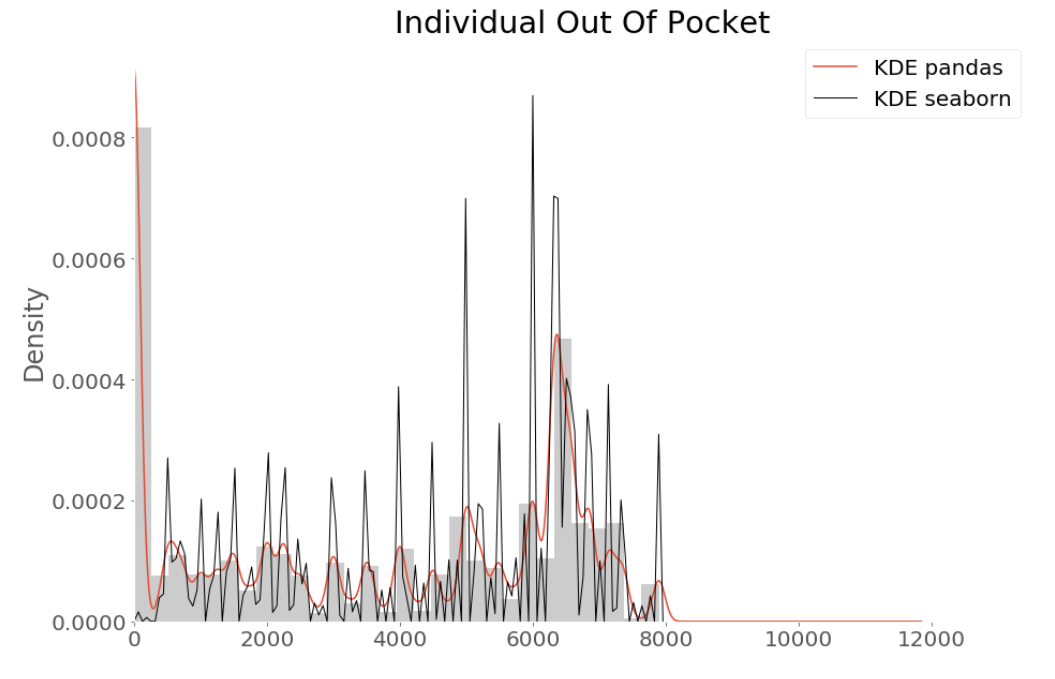
1. We have already checked how plan rates vary by age for different years. Now we will analyze how rates vary by age for different plan categories.
2. As expected, Platinum plan category is expensive
3. Bronze plan category is not the least expensive. Catastrophic plans are for specific needs only so they are cheaper but covers few benefits only
4. The variations at the ends of both graphs are due to the policy changes in the year 2017. Age limit for Old enrollees is changed from 64 to 65. Policies for children is reduced from 20 to 15.
5. Delta Plans has very little variations with Age after 20 years. Category 'Low' is quite double the category 'High'





1. Out of Pocket analysis

For any plan OutOfPocket is the amount which everyonse should care about. Apart from Monthly Premium, population must fetch out extra amount for every Service he/she takes. Let’s see how they are different for In Network and Out Network. I am going to look into EHB benefits for all the plans. I will use KDE to analyze it.



As the KDE chart shows, lots of plan limits are 0, that means these plans are for the poor people who need help. And we can find that most limits are more than 12000, that means the U.S. insurance insurers like focusing on high-end plan products. Maybe that's why Americans usually spend loads of money on their health insurance.

1. Unique plan offered to Americans: 981

Number of states participated in the Exchange program: 40

Over the period of 6 years, American were offered 981 unique benefits. Interestingly not all the 50 states have participated in the marketplace or Exchange. For example California, New york. States who are not participating in the federal run marketplace they have their own state-run exchange. There are only 39 states who are using federal run health insurance marketplace. Idaho moved to state run marketplace after the first year of enrollment. And state-run marketplace of Kentucky was dismantled in year 2016 and joined federal run exchange. The above number 40 is the number of states who were/are associated with federal run exchange since its inception.

1. List of benefit type topping the chart in a given business year is given below. There is no surprise that top of the benefits offered are of dental as these are mandatory for children and its also come with Medical benefits.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Benefits** | | | |
| **BusinessYear** | **count** | **unique** | **top** | **freq** |
| 2014 | 1164869 | 496 | Orthodontia -Adult | 18719 |
| 2015 | 2079286 | 517 | Orthodontia -Adult | 31269 |
| 2016 | 1774255 | 421 | Orthodontia -Adult | 26997 |
| 2017 | 1324275 | 281 | Orthodontia -Child | 21371 |
| 2018 | 829652 | 252 | Major Dental Care- Adult | 13857 |
| 2019 | 967050 | 244 | Basic Dental Care -Child | 15695 |

Lets check how the statistics change after dividing them into dental and Medical data.

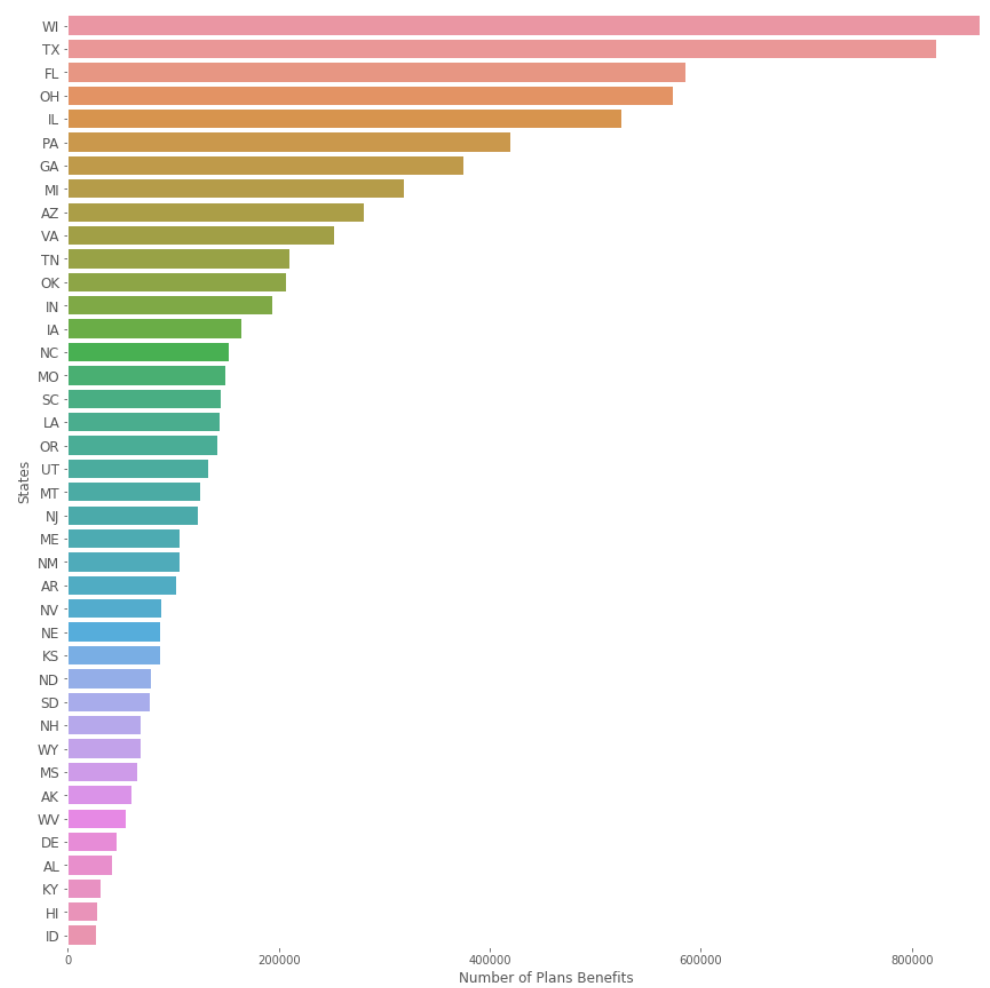
Medical Plans benefits overview

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Benefits** | | | |
| **BusinessYear** | **count** | **unique** | **top** | **freq** |
| 2014 | 1134699 | 339 | Mental/Behavioral Health Outpatient Services | 15247 |
| 2015 | 2038625 | 387 | Home Health Care Services | 26991 |
| 2016 | 1737894 | 317 | Orthodontia - Adult | 23138 |
| 2017 | 1298219 | 228 | Accidental Dental | 18568 |
| 2018 | 808437 | 191 | Mental/Behavioral Health Outpatient Services | 11565 |
| 2019 | 947765 | 193 | Routine Dental Services (Adult) | 13618 |

Dental Plans benefits overview

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Benefits** | | | |
| **BusinessYear** | **count** | **unique** | **top** | **freq** |
| 2014 | 34508 | 172 | Basic Dental Care - Child | 3532 |
| 2015 | 40661 | 141 | Orthodontia - Adult | 4278 |
| 2016 | 36361 | 164 | Orthodontia - Adult | 3859 |
| 2017 | 26056 | 62 | Basic Dental Care - Child | 2803 |
| 2018 | 21215 | 70 | Basic Dental Care - Child | 2292 |
| 2019 | 19285 | 60 | Basic Dental Care - Child | 2077 |

1. Few of the benefits are still in the Medical Benefit plan as they are the plans which give both Dental as well as Medical benefits
2. Number of Medical plans decreases from the year of inception.
3. Maximum number of plans were offered in 2015 but by 2019 it was reduced to 193 from 387, almost 50% reduction.
4. Same reduction is in Dental care as well.
5. This is not a good sign of the marketplace!!
6. Benefits covered in the plan is the major factor which drives the cost of the plan. More the benefits covered, expensive the plan will be.



So we can say that Wisconsin (WI) offers most number of benefit plans to the population and Texas is second. This can also be viewed on the America’s map like below.

# Summary of Analysis

In the whole analysis I have tried to figure out how Plans, benefits and rates are related to each other and how they are spread among all the states.

There are 40 states who have participated in the Exchange program and California is the biggest state which has not participated in any of these years. In the coming year if Calfiornia participated in the Exchange programs then it will be a huge opportunity for the companies, as California has maximum population to target

As the exchange program getting older, number of plans and benefits are decreasing. This is a huge concern for ACA initiative.

Wisconsin gives maximum number of benefits to choose from and has reasonable Monthly premium. It suggests that Wisconsin state is a huge success for ACA initiative.

Alaska offers minimum number of Benefits and premium is super expensive as compared to other States. It might be due to its harsh conditions and small population to deal with.

Huge premium will need to be paid if a user is Smoker. It is a good penalty imposed on people who are living Unhealthy life.