

**Health Insurance Marketplace Data Analysis Report**  
**Mike Wang**  
**CS 5266**

## Introduction

This report summarizes data analysis run on health insurance marketplace data come from the Centers for Medicare & Medicaid Service (CMS). Starting in March 2010, Affordable Care Act is designed to provide benefits for all Americans in bringing reform to the health insurance market. This project aims to analysis the data including everything related to all Federal Health Insurance in the United States.

Since 2019, the Pandemic has brought huge impacts on various markets and society, how did it affect the health insurance marketplace in the past four years? The purpose of this report is to provide findings and insights regarding different health insurance data on the Federal level.

## Data Exploration

Health Insurance Exchange Public Use Files (PUFs) are health insurance files released by the Centers for Medicare & Medicaid Services for easy access to data regarding each year's information on health insurance. The data captures the characteristics of individuals and families who have enrolled in health plans through the state and federal health insurance exchanges and provides a comprehensive view of the current health insurance landscape.

PUFs are all in CSV file format, there are more than 13 million entries in total. Data are available for plan years 2014 to 2023, for the purpose of this report, I am using the federal-level data range from 2019 to 2022. Data used in this project was organized and aggregated from the following original datasets:

- **Rate PUF:** Dataset on rates based on an eligible subscriber's age, tobacco use, and geographic location; and family-tier rates in different family sizes.
- **Benefits and Cost-Sharing PUF:** Dataset on essential health benefits, coverage limits, and cost-sharing.
- **Business Rules PUF:** Dataset on rating business rules, such as maximum age for a dependent, and allowed dependent relationships.
- **Plan ID Crosswalk PUF:** mapping plans offered in the previous plan year to plans offered in the current plan year.

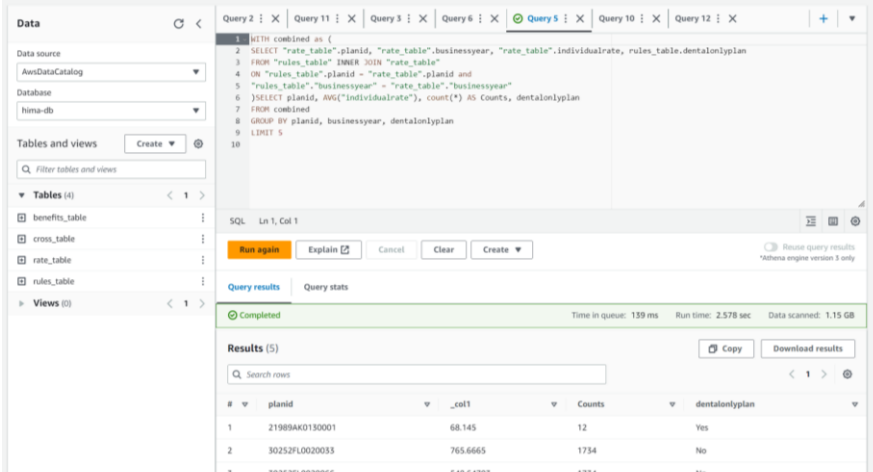
## Solution Structure

I first took a subsample of each of the datasets to do a quick EDA by checking whether there are duplicated or missing values with the interested data features.

During the process, the Rate datasets were found to obtain some outliers that are way higher than the rest of the data. After careful investigation, the outliers were identified as unknown values which were given 9,999 in the plan rate data field. Removing all the outliers provides a roughly normal distribution of the plan rate which is expected.

The Crosswalk dataset is another dataset that requires preprocessing before we can use the data. Each dataset contains the previous year's health insurance and its linked health insurance for the current year. Some of the plans do not carry the same plan id for the next year. To be able to connect and compare all the plans from 2019 to 2022, there is a data field named "ReasonforCrosswalk" to help. According to CMS, a value of "1" for "ReasonforCrosswalk" indicates that the two plans are the same plan with different plan ids.

After that, data from different years were concatenated into a single file that was then uploaded to AWS S3 Bucket for later processing. I used Athena to process querying and transformation for this project. Using SQL queries, it is easy to join related datasets and run queries including complex and nested queries. Following is a query sample that returns back aggregated dataset of all plans grouped by business year and whether it is a dental plan or not:



The screenshot shows the AWS Athena console interface. On the left, the 'Data' pane shows the 'AwsDataCatalog' database and 'hima-db' database. The 'Tables and views' section lists 'benefits\_table', 'cross\_table', 'rate\_table', and 'rules\_table'. The 'Query 5' tab is active, displaying the following SQL query:

```

1 WITH combined AS (
2   SELECT "rate_table".planid, "rate_table".businessyear, "rate_table".individualrate, rules_table.dentalonlyplan
3   FROM "rules_table" INNER JOIN "rate_table"
4   ON "rules_table".planid = "rate_table".planid and
5   "rules_table"."businessyear" = "rate_table"."businessyear"
6 ) SELECT planid, AVG(individualrate), count(*) AS Counts, dentalonlyplan
7 FROM combined
8 GROUP BY planid, businessyear, dentalonlyplan
9 LIMIT 5
10

```

The query results are displayed in a table with 5 columns: #, planid, \_col1, Counts, and dentalonlyplan. The results show 3 rows of data:

#	planid	_col1	Counts	dentalonlyplan
1	21989AA0130001	68.145	12	Yes
2	30252P10020033	765.6665	1734	No
3	30252P10020066	548.64703	1734	No

By Connecting to Athena Client using the boto3 library on Python, queries were passed as parameters into the Athena execution function, and later downloaded back into a CSV file for analyzing. Using the results, I would be able to solve different questions using Pandas data frames to answer questions like the following:

- **How did Health Plan Rate / Copay / Coinsurance change across years?**
- **What are Rate / Copay / Coinsurance differences across US States?**

- **Will and how Rate differ based on applicant's condition (Tobacco / Family size / Dependent / age)?**

### Visualization and Insights

To illustrate the trends and patterns, I used Plotly library on Python to visualize findings related to the different questions. Plotly is good at showing multiple groups of variables on the same plot which helped me to illustrate changes and patterns across different groups.

First, I took a general look at the average plan individual rates in general from 2019-2022.

year	2019.0	2020.0	2021.0	2022.0
individual_rate	325.764151	342.201209	378.964193	412.188214

Table 1

As expected, there is clear growth in the individual plan rates from 2019 to 2022. By grouping each plan into a range of rates, we can see the counts in each group and get a rough distribution of the plan rates in US. Fig. 1 showed that majority of health insurance plan rates fall between the range of \$401-\$800. This is a result that excluded dental-only plans.

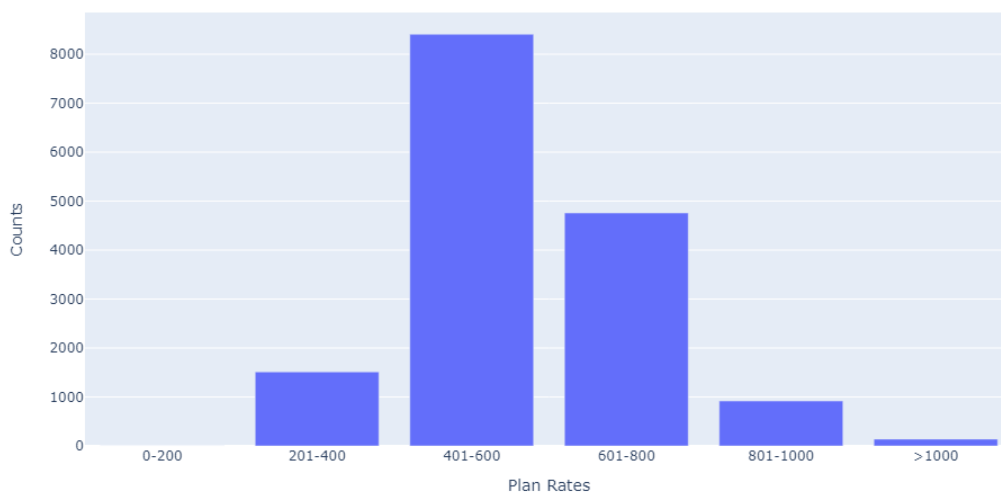


Figure 1

Further investigating plan rates by the groups by age, I found big differences among age groups when it comes to individual plan rates as shown in Fig. 1. Children (Age < 18) in general had lower plan rates and the older adults, the higher the rate of plan tends to be. Interesting fact is that, although children and younger adults in general have lower rates compared to older adults, their plan rates have the highest growth rate since 2019 (Fig. 2). After the age of 20, the growth rate of adults' plan

rates started to slowly climb up and getting close to children's growth rate.

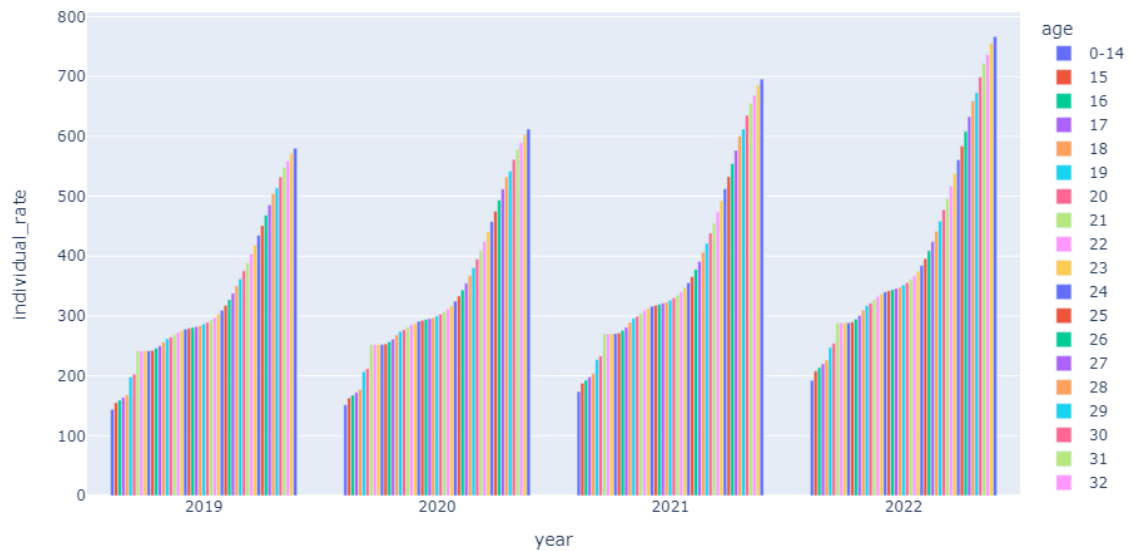


Figure 2

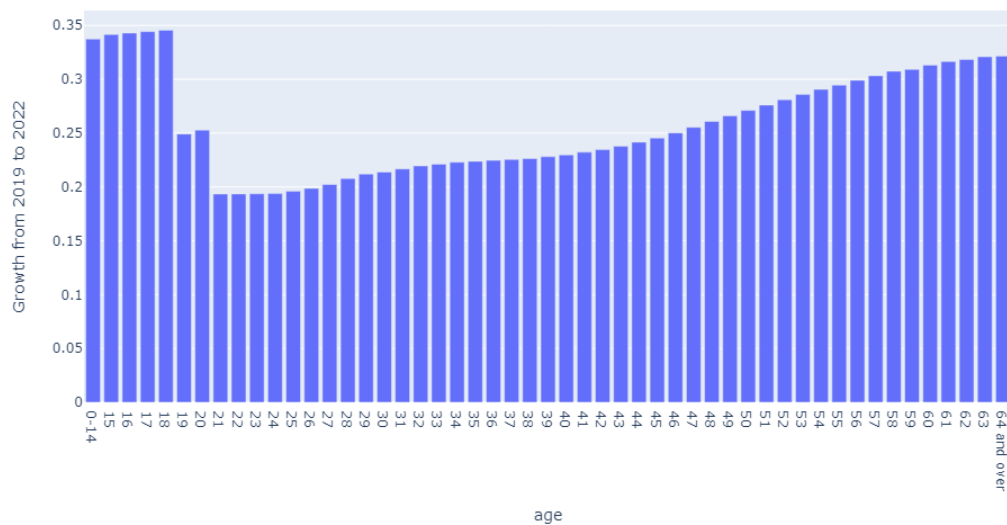


Figure 3

Similar steps were repeated to analyze the trend of family plans from 2019 to 2022. Rather than a charge based on the applicant's age, family plans rate varies differently depending on the family size and number. The data fields are ordered from small to large are: individual\_rate (one person), couple\_rate (couples), primaryondependent\_rate (one person and one dependent), primary\_twodependent\_rate (one person and two dependents), coupleondependent\_rate (couples and one dependents), coupleandtwodependents (couples with two dependents)



Figure 4

As shown in Figure 3, monthly plan rates between different family sizes are significantly different from each other. An individual has the lowest average plan rate ( $M = 30.75$ ), and couples with two dependents have the highest average plan rate ( $M = 105.31$ ). Since we have found that children tend to receive lower plan rates, it makes sense that one primary subscriber with two dependents has lower plan rates than couples with one child. However, the rate for one primary subscriber with one dependent is slightly higher than a couple, which seems to contradict the previous finding. Family plan rates were fairly stable during 2019-2022, except for a slight peak in the year 2021.

About 25% of health insurance plans were calculated depending on applicants' tobacco usage, so I also compared the difference between non-tobacco users and tobacco users and observed that tobacco users in general do receive about 16% higher plan rates than non-tobacco users (Fig. 5). Interestingly, the plan rates for non-tobacco and tobacco users were both reducing overtime, although very slightly.

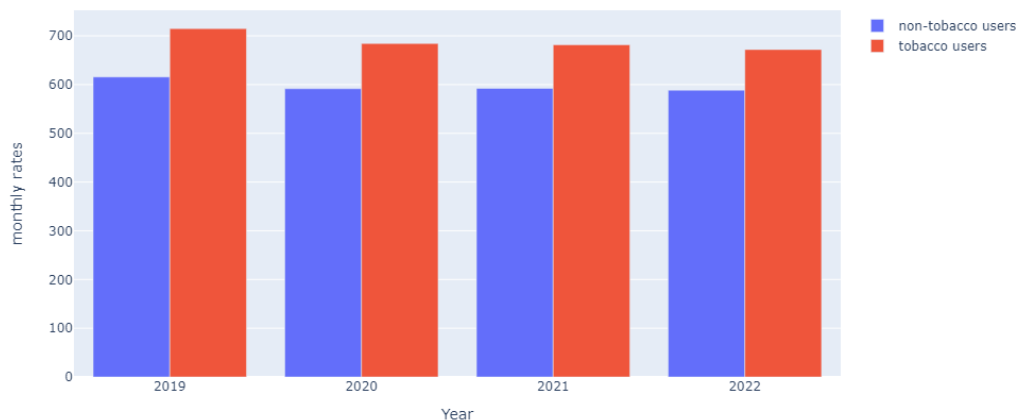


Figure 5

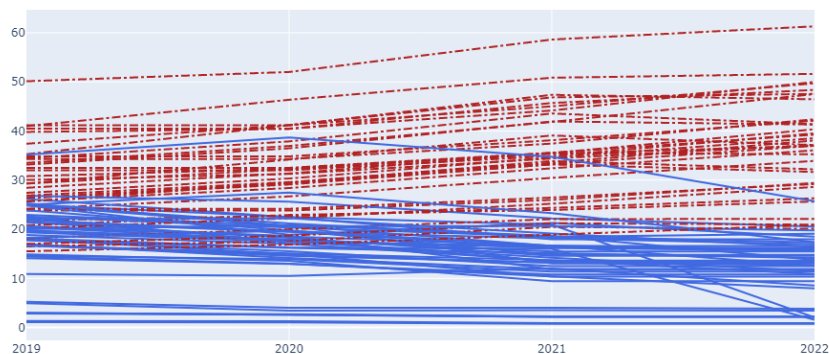


Figure 6

By using the Crosswalk dataset, I was able to connect all the health Insurance plans that were carried over from 2019 to 2022. There are 1643 non-dental plans and 552 dental-only plans. 15.5% (256 out of 1643) of Non-dental plans dropped or increased by at least 20% in the past four years, whereas 14.6% of Dental plans (81 out of 552) of plans dropped or increased by at least 20% in the past four years. As shown in Figure 6, red dotted lines are dental plans that increased by over 20% in the past four years, and blue straight lines are those that dropped by 20% or more. It seems that a higher proportion of expensive dental plans tend to increase their rate, and cheaper dental plans more likely to reduce their plan rates over time.

By using US states as group filters, I calculated average health insurance rates for each available US state and compared the difference. The lack of data in some states due to state law, states such as Washington, California, New York, and Idaho only permit state-based health insurance rather than federal-based health insurance. Future studies could focus on state vs federal analysis.

Among 39 available states, West Virginia has the highest average plan rate ( $M = 721.26$ ) and Michigan has the lowest average plan rate ( $M = 135.35$ ). In general, Mid-West and Northeast states have higher individual plan rates than other states in the US.

For the Family Plan Rates, only 21 states have applicable data. Among all the states, Alaska has the highest average family plan rate ( $M = 162.76$ ), and Georgia State has the lowest average family plan rate ( $M = 40.72$ ).

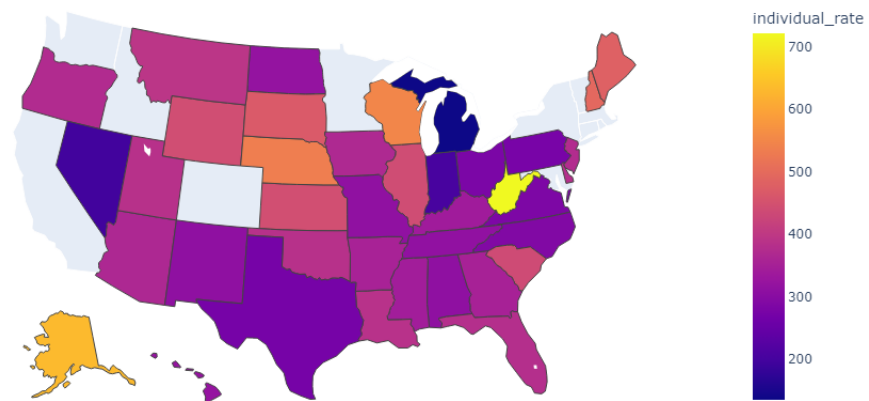


Figure 7

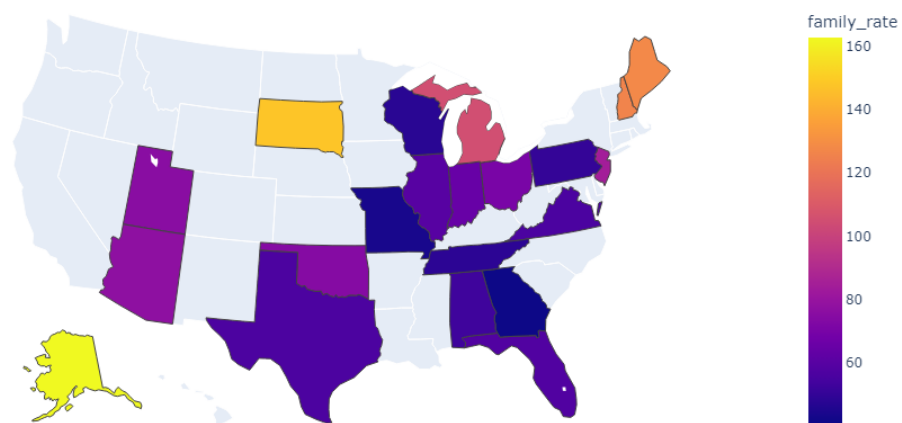


Figure 8

The benefits dataset requires text analysis to filter different copay and coinsurance conditions based on the string provided. For Copay amount (fixed amount patients pay for each visit), there are three main types of copayment which are: Copay after deductible, Copay with deductible, and Copay without deductible. The copay amount vary from \$0 all the way to more than \$1000 with a median at \$141. Following are the count plots of copay amount from 2019 to 2021:



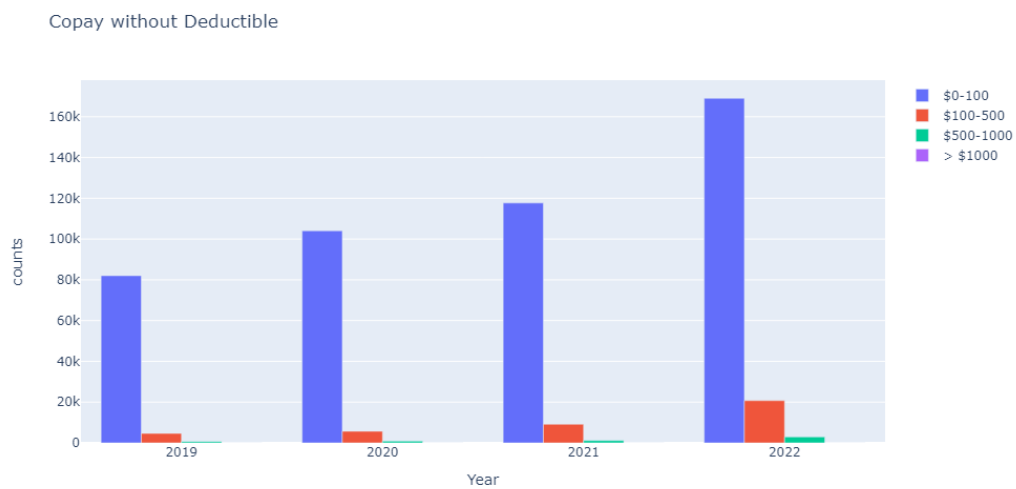


Figure 9

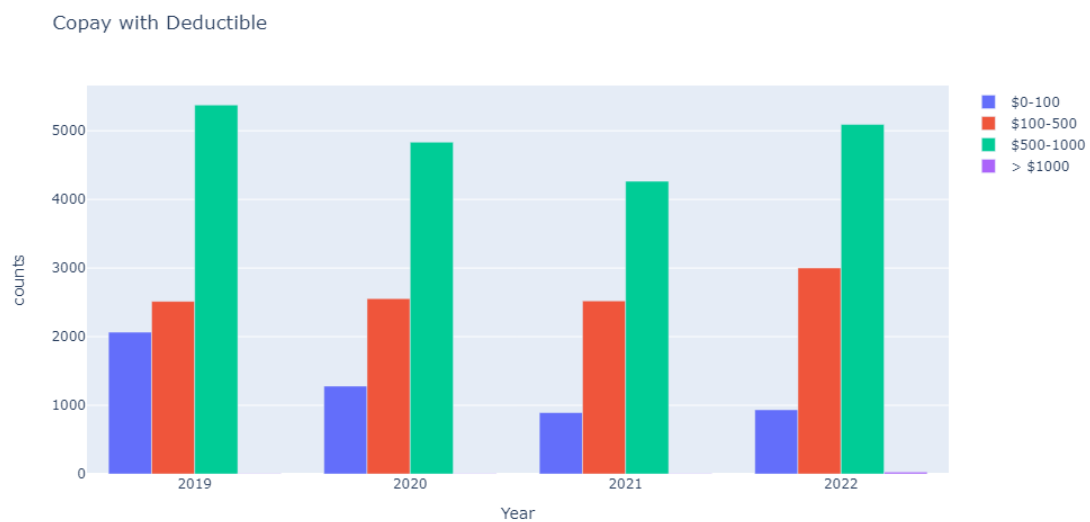


Figure 10

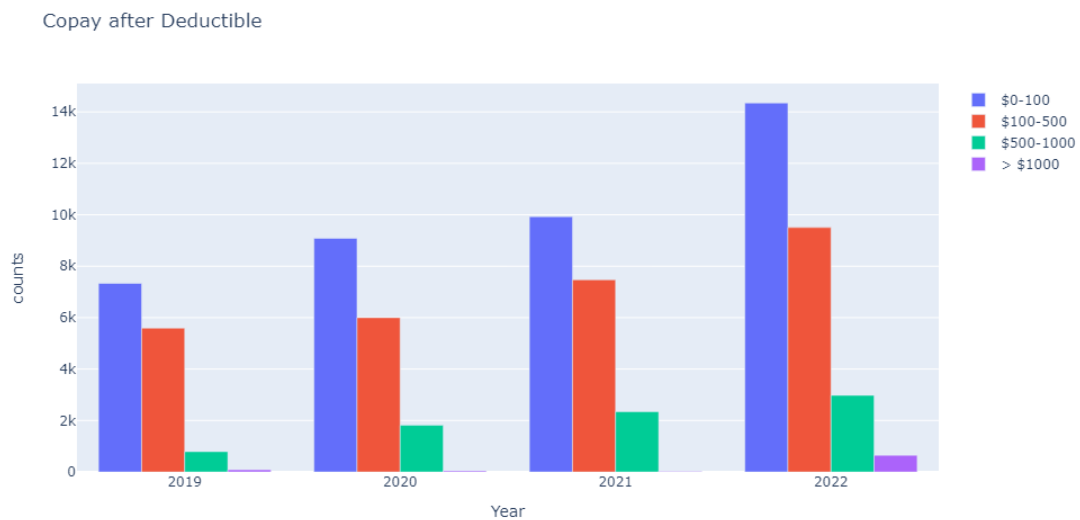


Figure 11

Comparing across three plots, we can see that almost every copay amount without deductible was under \$100, whereas copay with Deductible has the highest amount of plans between \$500-\$1000. Although most of copays without deductible were under \$100, there is a slight growth in plans that require copayment higher than \$500.

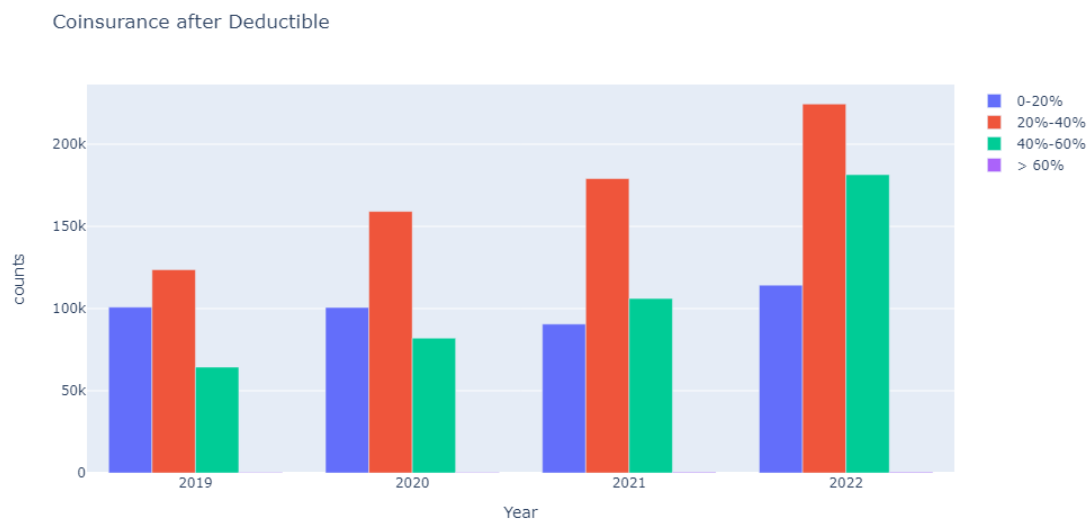


Figure 12

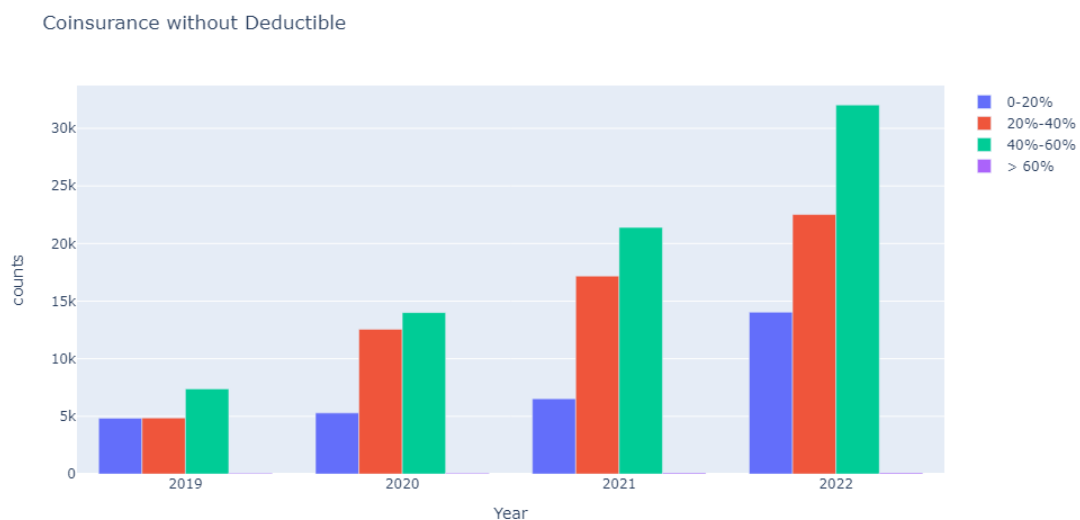


Figure 13

Coinsurance (Percentage of costs patients pay) was grouped into two types: Coinsurance after Deductible and Coinsurance without Deductible. Same visualization was done to coinsurance, the only difference is that coinsurance use percentage as a unit so the data range is between 0 to 1. Coinsurance after deductible overall provides a better percentage than coinsurance without a deductible.

### Resources

#### Health Insurance Exchange Public Use Files (Exchange PUFs):

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

<https://www.cms.gov/ccio>

#### Copay vs coinsurance:

<https://www.verywellhealth.com/whats-the-difference-between-copay-and-coinsurance-1738506>