Documentation

I downloaded the datasets from cms.gov (healthcare.gov) together rates, plan attributes, crosswalk, benefit descriptions in Excel and reviewed data. In reviewing the data dictionaries, it became apparent that there were too many columns describing various copayments at the medical insurance level for essential health benefits for individuals, couples, 1-3 person families, etc. Explaining copayments and coinsurance for first tier dental benefits was complicated enough. Too many columns and rows were in the data set (hundreds of millions if joined), but it was obvious that the rates table could be aggregated by age to shorten the table. I determined that I would need the state, plan ID, Issuer ID, rating area, and metal level it was available for as many tables as possible to create unique keys to be able to join tables together later. The Service Area, Network, and Business plan tables I quickly figured out I would not need, they had redundant information.

First, I imported Benefits table into Python and made a copy. 2.2 million rows, 24 columns, 0 duplicates, largest number of blanks is 1.8 million for the CopayInnTier1, but it’s because most plans do not have that copayment tier, they have a CoinsInnTier1. Because I was aware that the table had medical benefits as well as dental benefits, I reviewed the values in ‘BenefitName’ and kept only those rows that contained information regarding dental benefits or dental benefit procedures. This shortened the table to 361k rows, with most of the columns having unique values for ‘BenefitName.’ Next, I separated the standard dental benefit names from the medical based ones by making a copy of Benefits column, duplicating it, stripping the regular benefit names, renaming it, and added it back into the dataset using a merge, then reordered the columns. I did attempt some further data cleaning in Python, but it took a long time to find the correct formulas and it was faster to clean the columns in Excel. I also removed most of the columns that would not be needed in the final table and ran data checks. There were many blanks in each column because some of the benefits did not have values for those columns. Because this was mostly the case and there were rarely columns missing data because the information itself was missing, I did keep all blanks or null values. I exported the table to csv.

For the plan attributes table, I imported, copied, and did data checks and found 32k rows and 147 columns, no duplicates, and man blanks again because some plans did not have values for attributes that other plans did. I kept the columns ‘StateCode,’ ‘IssuerMarketPlaceMarketingName’, ‘MarketCoverage,’ ‘DentalOnlyPlan’, ‘StandardComponentId’, ‘PlanMarkingName’, ‘PlanType,’ ‘MetalLevel’ and most of the MEHB columns initially which to 32 columns. I exported the table to a csv.

For the rates table,I had 2.8 million rows, 20 columns, no duplicates, and many blanks, though again, not every row had a value for the attribute. I removed most of the columns and performed aggregation on the rates by age using Group By into ‘Under 14’, ‘15-17’, ‘18-30’, ‘31-40’, ‘41-50’, ‘51-60’, ‘61-64’, ‘Over 64’, and ‘Family’ using a formula. While this might seem to skew the data because not all the groups have the same number of values, for many of the ages under 30, the difference in rate only changed by a few dollars when the age increased and all the rates for ’15-17’ were the same for each age. I also verified exactly which states had rates in the table, 33 states, and re-ordered the columns, only keeping the ‘Individual’ and ‘Couple’ rates initially (I removed ‘Couple’ from the final table when I opted to focus on the ‘Under 14’ age group). The resulting table was 375k rows and 6 columns. I exported the table to csv.

For the Crosswalk ID, I joined the table with the rates, plans, and benefits tables in various configurations to reduce the null values and maximize the number of plans that would return complete rows. Although the tables could be joined, the resulting rows had many null values resulting in the joins not “seeing” all the information in the index and Python kept thinking that there were unique values in each column even though most of the information in the columns was the same most of the time.

Because there was missing information about how states, rating areas, and FIPS codes were related and it was not possible to use the existing datasets to answer the question, I decided to compile a separate table of FIPS codes, counties, and rating areas. The geographic data was based on states, rating areas, and FIPS codes in the tables. I downloaded information from various websites to compile the table as completely and accurately as possible. Additionally, it was suggested that I may want another table with regions of the US so that I could group the states for future aggregations. I saved the table in Excel as an Excel file and CSV.

I determined that the Benefits table would be the primary table for the dataset (’index’) to which all other tables were joined because it had the most information on each individual plan and the most attributes (‘PlanID’, ‘StandardComponentID,’ etc.,) for using as the index for the joins.

Initially I uploaded the full data sets to SQL and attempted to join the tables, however, since two of the tables had over 2 million rows and tens of columns, SQL took up to 30 minutes to create the resulting table. Instead, I imported the tables that had been modified in Python and joined those to produce a dataset with all the attributes I wanted to analyze. As in Python, I tried to join the Plan ID Crosswalk with the other tables, but since it only contributed rating level, ID, and metal level and I already had them in the table, along with the FIPS County Rating Area table, I decided to forgo utilizing the crosswalk.

When joining the remaining tables, even with aggregation, the SQL query to formulate the dataset took a very long time and produced millions of rows, no matter how the tables were joined. In the end, I used a CTE table in the query to produce the dataset, but I still had to limit the data to the ‘Under 14’ age group to produce a table that was only 330k rows. I did not attempt to analyze the data or create pivot tables in SQL.

SQL query dataset was imported in to Excel where I copied each data set and filtered each column to determine whether the table needed additional cleaning. I realized that I needed to add back in Child Only and Dental Only to analyze the percentage of plans that fell into those categories, so I modified the SQL query to reproduce the new data set and imported it back to Excel. In Excel, I removed any references to ‘Adult’ procedures or ‘Adult Only’ plans from the columns, set the coinsurance column format to %, copay column to $, and ‘AvgIndividualRate’ to $. I also changed the description of benefits to be ‘Routine’, ‘’Check-Up, ‘Basic’, ‘Major’, and ‘Orthodontia’ since all the procedures referred to children. I removed the ‘OtherBenefits’ table because it only had values for the adult procedures, not the child procedures. I removed the ‘Age’ column since the ages were all under 14, then imported the final data set and the FIPS County Rating Area Region set into Tableau for Analysis. While it would have been ideal to combine the geographical data with the SQL dataset, it became apparent that the resultant table would have millions of rows again and be difficult to produce in SQL or to make sure that all the rows had all the information because Excel might not save all the data. Resulting tables were saved as Excel and CSV files to import to Tableau for analysis and producing data visualizations.

After uploading all the tables into Tableau, I created some basic tables pivoting data on geographic location, age, dental plan benefits, etc. First, I focused on the analytics that would be used in the dashboard, so rates in each state, states, rating areas, and counties, numbers of plans in each geographical area, the copayments for each of the benefits and the frequency limitations on the benefits. Then I analyzed numbers of plans and average individual rates of each plan by state and rating Area, then grouped by region. Similar pivot charts were made for numbers of plans and average individual rates by metal (‘Platinum’, ‘Gold’, ‘Silver’, ‘Expanded Bronze’, ‘Bronze’, ‘High’, ‘Low’, and ‘Catastrophic’) and by plan type (Indemnity, PPO, EPO, HMO, POS). Calculated fields were produced so that I could combine copayment and coinsurance in and out-of-network under the same heading (plans either had one method of patient payment or the other). I also created a ’Plan Detail’ where the state, county, and rating area were all in one label so that the dental plans could be filtered simultaneously on all details and return from an additional column, only the plans that were offered in that state, county, rating area. An additional calculated field with plan id, plan type, and metal level was also created for the similar purpose of filtering the copayment/coinsurance information on each benefit type as well as the ‘Frequency Limitations’ (additional calculated field combining unit of limits and the limit description for the benefits so that the dashboard would read ‘2 per year’ for Check-Up benefit (ex.)). Additional pivots included maps with rating areas in each state with county names and average individual rates, pie charts detailing the breakdown for the plan count for child-only coverage, medical/dental, metal level, and plan type, as well as the number of plans that cover each dental benefit. Resulting data visualizations were utilized for a dashboard and PowerPoint presentation.

Findings include:

* Over 32k distinct plans.
* Highest number of dental plans in the South, especially Texas with 1023.
* Wisconsin and Florida also have a high number of dental plans at 624 and 556, respectively.
* West Virginia has the highest average cost for an individual customer $419, followed by Nebraska and South Dakota, $313
* Most common metal level offered in a dental plan is Silver at 31% followed by Expanded Bronze at 25%.
* Most common type of plan utilized is an HMO at 51%, then EPO at 23% and PPO at 22%.
* Blue Shield has the highest number of plans being used throughout the country by +50%, 1938 (after grouping). Ambetter has the next highest at 15%.
* Highest utilization of plans at 97% are adult and child as opposed to child-only, most customers looking for family plans.
* Over 86% of plans are medical plans with dental benefits which explains high average individual rates.
* Orthodontics only covered in about 37% of child dental plans.
* On average, the most common type of plan is an HMO Silver Plan that covers adults and children in the South at around $400 for a medical/dental plan.

What was missing in the data?

* Some of the counties were labelled incorrectly on the websites used for reference and some were spelled differently in Tableau and had to be rectified there. For the plan benefits, there didn’t seem to be any deductible information. For some reason, not all 50 states were in the dataset, even though all the rating areas were on the cms.gov website.
* Most of the data was faithful in filling out a lot of information, however, it would be helpful if the inputs for some of the plans were more limited so that deciphering null values was easier. For instance, many of the tables were missing information on limit quantity and limit units referring to frequency limitations which all dental plans have outlined because they only cover each benefit a certain number of times in a certain time frame, but the information was only present in about 1/3 of the plans.
* It would have been helpful if there was a definite distinction between having a “0” or having a blank. The blank could mean “0” as well, but it wasn’t always clear from the data.
* Another issue was the size of the tables, for analysis of the entire datasets, using AWS or a program made for cleaning and analysis of millions of rows would have made it possible to analyze the entire dataset as opposed to just parts of it.

The government benefits from the findings by being able to produce more efficient offerings for dental insurance for individuals. For state and federal dollars, the primary concerns should be the ROI for the patients and state/country and being able to justify the cost for the benefit to the American taxpayer and government representatives to keep the plans funded.

The final analysis was conducted in a much more streamlined fashion than the EDA, though the steps followed to clean and analyze the data were pretty much the same. Adjustments had to be made for which tool to use for which step due to the size of the data tables and the final size of the dataset used. For instance, even though the raw data could be imported into DBeaver and joined, the resulting tables were so large that trying to shape the data with that program became too time consuming and crashed my computer several times. Likewise, shaping the tables in Python was relatively straightforward, but joining the tables and trying to analyze and pivot the data was very time consuming in terms of researching the appropriate Python commands to use and then when the tables were joined, we still ended up with millions of rows in the final set. Ultimately, using Python to shape the data, SQL to join and produce a dataset, and Excel for EDA and cleaning and reformatting to produce the final dataset ended up being the most efficient process. Deciding not to join the geographic data set directly with the dental plan benefits data set until using Tableau produced the most efficient outcome of all.

The only additional data added to the sets was the inclusion of US geographical regions to aggregate the data regarding rates and benefit costs.

Mostly averages, percentages, and count statistics were used to create pivot charts and calculated fields to analyze the data because there was so much variation in the plans, plan types, metal levels, even within the same set or the same plan type or metal level that comparing the statistics on them would have yielded data where you were comparing Gold plans that were really structured more like Bronze plans and vice versa depending on the geographic location, plan type, etc.

There were some inconsistencies in the data. Initially, I had put together the FIPS County Rating Area dataset by researching the current names of all the counties. However, Tableau did not have the most current information or names for some of the counties, so I was getting far too many null values until I reconciled the discrepancies county name by county name in Tableau to match my geographical area with the county name that Tableau described. There is still one null datapoint in New Hampshire, but Tableau seems to think everything is reconciled without recognizing that the null has not been, so I removed it from the data. Also, there may have been some inconsistencies in the individual rate data for the insurance plans, some of the costs seem a little high for an individual rate for each member of the family, even with it being a medical/dental plan. However, since I removed the ‘Family’ rate from the data initially to focus on the rate for ‘Under 14,’ that may be a matter of how you are defining what type of data you are looking for. It was not obvious from the start that most of the plans were medical/dental and covered adults and children, but if it had been, trying to find the rates for families instead might have been the objective. It would not have been possible in that time frame to produce a data set with ‘Family’ and ‘Under 14’ values, the data set was already too large as it was.

Recommendations include:

* Most of the plans seem to be formulated to maximize preventative coverage and cover most of the more expensive procedures, though not all (Silver Plans covered at 70% with Routine procedures covered 100% of the time).
* If the government were to invest more money on these plans, offering Silver Level HMO plans for the family would probably have the best ROI.
* Educate the patients- choosing dental care plan is highly dependent on what your family’s needs are. Educating the parents on the best coverage for their child’s treatment plan may increase dental insurance participation for ACA plans.
* Possible to structure the plans at a family rate so that the investment can be prioritized for the family member who needs it most at the time for more efficiency and greater ROI.
* Some of the plans may have “rollover” benefits, may be beneficial to have rollover on the yearly maximums in addition to the ortho max for the plans that have an Orthodontia benefit.

Next steps include:

* Considering this our “sample set” and investigating other tools and methods for cleaning, shaping, and analyzing data with larger data sets.
* Adding data from the remaining 18 states to see if the trends in the data hold up regarding the highest usage being a Silver HMO Family plan and how much the average individual rate would be for that plan.
* Incorporating patient utilization data into the dataset to see if the assumptions about being able to optimize the ROIs for the plans are supported.
* Discussing and implementing patient education protocols with insurer participants to help educate the patients and increase interest and use of the ACA plans.
* Investigating the pros and cons of the ‘rollover’ option.

References:

Downloads from:

* https://www.cms.gov/CCIIO/Programs-and-Initiatives/Health-Insurance-Market-Reforms/state-gra
* https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=&cad=rja&uact=8&ved=2ahUKEwib28bCmZP-AhUDmWoFHagbAKsQFnoECAkQAQ&url=https%3A%2F%2Fwww.mdreducation.com%2Fpdfs%2FUS\_FIPS\_Codes.xls&usg=AOvVaw36EFL9luETN\_q-sTDErMI3
* https://transition.fcc.gov/oet/info/maps/census/fips/fips.txt
* **https://simple.wikipedia.org/wiki/List\_of\_ZIP\_Code\_prefixes**
* https://namecensus.com/zip-codes/alaska/
* https://www.cms.gov/cciio/resources/data-resources/marketplace-puf

Documentation steps also outlined in Jupyter Notebook.