Health Insurance Marketplace Analysis using Hadoop & Hive

Authors: Rishabh Sureshkumar Shah,
Department of Information Systems, California State University Los Angeles
CIS5200 System Analysis and Design

Abstract: The strategy and procedure employed for manipulating the health insurance market and subsequent analyses are explained in the article. Giving a clear flow for processing massive data files and data cleaning procedures utilizing Hadoop and Hive is the project's focus. Furthermore, this data is analyzed using Excel and Tableau, which displays visuals like maps, timelines, and charts on insurance plans and relevant important data. To provide timely benefits, rate analysis, copayments, premiums, geographic coverage, etc. for our data analysis, Health Insurance Exchange Public Use Files (Exchange PUFs) are accessible for plan years 2017 through 2021 [1]. We may, in essence, design a clinical strategy that meets the needs of families, people, and employees and enables them to browse different healthcare insurance plans and select the most important one based on their income and expenses.

1. Introduction

Our project is based on Big Data technologies such as Hadoop and Hive to store, process, and transform the Health Insurance Market dataset. The dataset mainly consists of information about insurance plans, types, cover and ge, individual/family/ group bene, fits, etc., We probably hear people talking about their health insurance all the time. You might, however, feel as though you don't understand it. The United States is the only significant industrialized nation without universal health insurance; over the past six years, the standard of care has dropped. Your ability to pay for your medical bills and receive the care you need can both be facilitated by health insurance. We have extracted data from CMS (The Centers for Medicare & Medicaid Services) resides in the U.S. Department of Health and Human Services as a Federal agency [1].

2. Related Work

There are numerous works publicly available based on Health insurance marketplace data. One of the works based on similar lines is available in the Hamilton project -smart policies on health insurance use the Healthcare.gov dataset [2]. The outcome of this study is to create a platform to suggest smart policies to users based on metrics such as cost reductions and insurance coverage offered by employers whereas in our analysis, although the dataset is different, we concluded the best pick for users in different age groups, cost of plans, highly suited plan types and essential benefits coverage along with reduction strategy. Furthermore, our workflow architecture and approach are different as we are using big data file management. Another similar work was presented by the NAIC (National Association of Insurance Commissioner) [3]. This agency's work focuses more on the business side of Health insurance marketplace data wherein they drew the following conclusion over 10 years from 2012 to 2021, there's tremendous growth in issuing plans but a downfall is observed in their net earnings% and profit margin%. In contrast to this work, we showed the trend and behavior of analysis towards the consumer and help them pick the best-suited ones according to the insurance companies' area coverage, plan types, reduction cost%, and EHB% coverage.

The work presented by AMA (American medical association) emphasizes more on health insurance companies with wider market coverage and focuses on insurers' enrollment% and small groups% with different stats based on their defined metrics [4]. Our synopsis is different from the above-mentioned work with regards to deliverables, since we are focusing more on individual/family plan needs with the help of interactive dashboards thus, giving more insights into various insurance plans.

3. Specification

The dataset comprises of insurance plans related to planning rates, types, coverage, rules, family members, etc. The dataset is of size 3.3GB having a duration from 2017 to 2021 [1]. Table 1 shows numerous files, file types, and, their size of it.

Table 1 Data Specification

Data Set	**Size (Total 3.3 GB)**
rate.csv	1,578,237,718 KB
benefits_cost_sharing.csv	1,492,760,687 KB
plan_attributes.csv	54,091,603 KB
business_rules.csv	8,916,090 KB
service_area.csv	5,862,295 KB
transparency_in_coverage.csv	1,436,757 KB
network.csv	505,880 KB
machine_readble_url.csv	82,728 KB

The below table shows the hardware specifications for the Hadoop cluster.

Table 2 H/W Specification

Number of nodes	5
CPU speed	1995.309 MHz
Memory	251.3GB
Storage	390.7GB
Hadoop Cluster Version	3.1.2

4. Implementation Flowchart

Health Insurance Marketplace dataset is downloaded from the official website of CMS (Centers for Medicare & Medicaid Services). The entire implementation process is shown in the

algorithm/flow chart below (Figure 1). The raw data set was downloaded from the year 2017 to 2021 to our local system and then 8 data logs in CSV format were uploaded to the Linux system. To transfer the files from Linux to the Hadoop server, we logged in to our remote HDFS server and later created subfolders under the main folder to upload files. From HDFS, we connected to Hive using Beeline-Client to create 8 subsequent tables schemas using HiveQL. Data transformation, cleaning, and summary tables are performed using the same query language. The aggregated results from the summary tables have been exported and imported into Tableau to highlight business insights through interpretation and visualization.

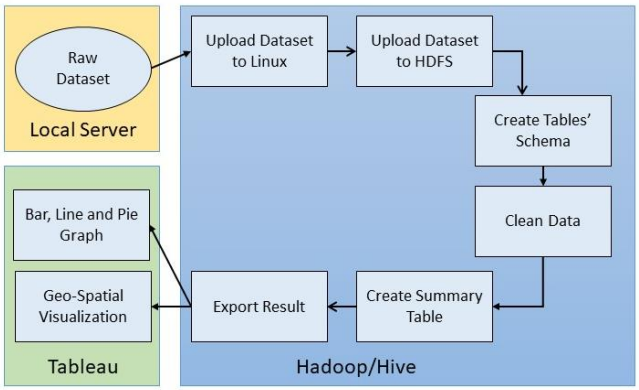

Figure 1 - Implementation Flowchart

5. Data Cleaning

Data cleaning was performed on all the files having a different set of information using various built-in functions such as regular expressions, nested-if else statements, type casting & Date difference, case when, joining tables, group by & aggregate functions.

Since the file size is huge such as rate.csv and benefits_cost_sharing.csv are having records above 10 million which are difficult to process, we created subsets of these tables using views or tables using the above cleaning methods which brought down the data from 3.3 GB to 55 MB.

Below are the highlighted methods of data cleaning.

1) **Regular expression** is used to extract domain names or insurance company names offering various plans.
   regexp_extract (m1.Tech_POC_Email, '@(.*)\\.', 1) as insurance_company_name
2) **Aggregate functions** in the group by clause

max (IsEHB) as IsEHB, max (IsCovered) as discovered, max (Exclusions) as Exclusions

3) Nested if-else condition

if (cast (age as int)>=15 and cast (age as int) <=44 or age='0-20','15-44', if (cast (age as int)>=45 and cast (age as int) <=64,'45-64', if (age ='64 and over','65 and over, age))) as age_new,

avg (if (IndividualRate=-1, 0, IndividualRate)) as IndividualRate,

4) Casting

cast ((cast ((datediff (cast (RateExpirationDate as timestamp), cast (RateEffectiveDate as timestamp))) as int)/364) as decimal (5,2)) as total_plan_year,

5) Case –When the condition

The case when cast (RateExpirationDate as timestamp) >= current_timestamp then cast ((cast ((datediff (current_timestamp, cast (RateExpirationDate as timestamp))) as int)/364) as decimal (5,2))

else 0 ends as active_plan_years_left,

Thus, a summary table is a small view/table of aggregated data that helps in analyzing and picking the best health insurance plans as per individual/shop needs across the USA.

6. Analysis and Visualization

After creating summary tables, we extracted the files and performed data analysis and visualization using Tableau. For our analysis, we highlighted deep insights by using several charts, graphs, and geospatial forms to showcase the best insurance plans as per individual/group needs.

6.1 TreeMap and Geo-spatial Graph

The below chart shows (Figure 2), the top 15 Health Insurance Plans are oss USA based on total subscriber count. From this analysis we came to know, various dental check-ups for children plan is the most issued one as they require frequent dental check-up during their growing stage. Thus, the following geospatial map shows (Figure 3), the coverage area for these top insurance plans.

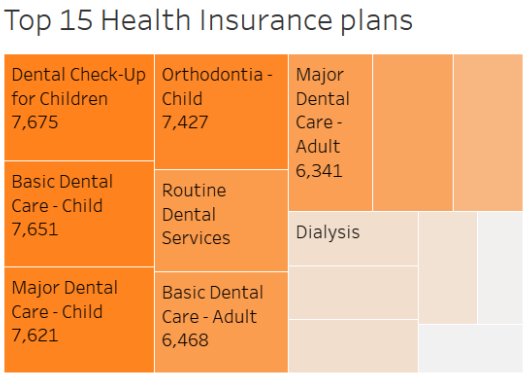

Figure 2 – Top 15 Health Insurance Plans

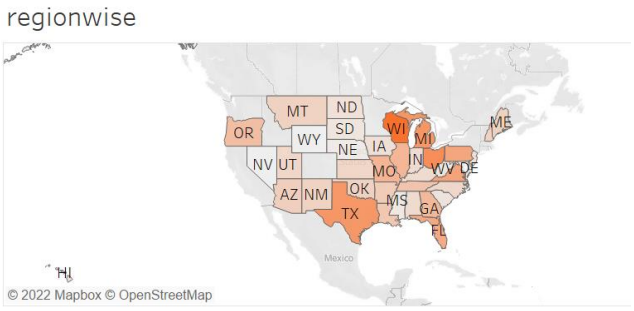

Figure 3 – No of Insurance Plan issued State-wise distribution

The next chart shows (Figure 4) the lowest trend of the Denplan plan issued in Arizona state.

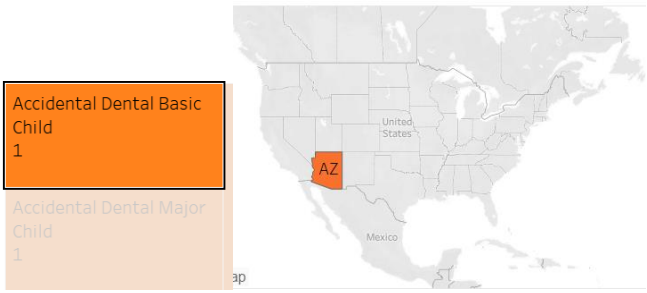

Figure 4 – Lowest insurance plan count with its covered state

6.2 Bar graph and Geo-spatial graph

In the previous analysis, we showcased popular health insurance plans, thus supporting the following statement. We are interpreting well-recognized health insurance companies through the below graph and their coverage area across the USA.

From Figure 6, we observe that Health the insurance provider has the most customer base issuing the above health insurance plans. In contrast to the above analysis, we interpret that Argusdental insurance company has the least customer base in Florida (Figure 7).

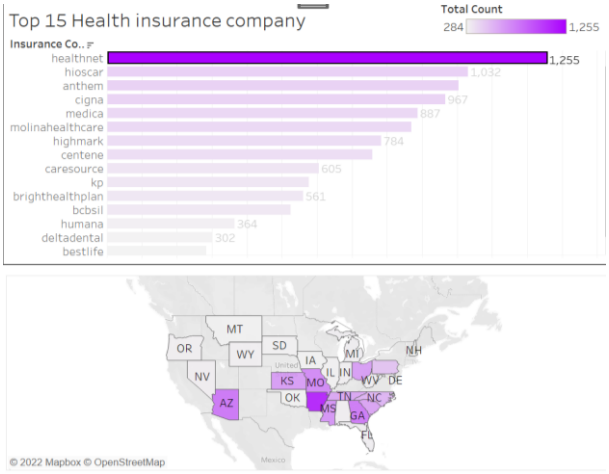

Figure 6 – Topmost health insurance company and its coverage area

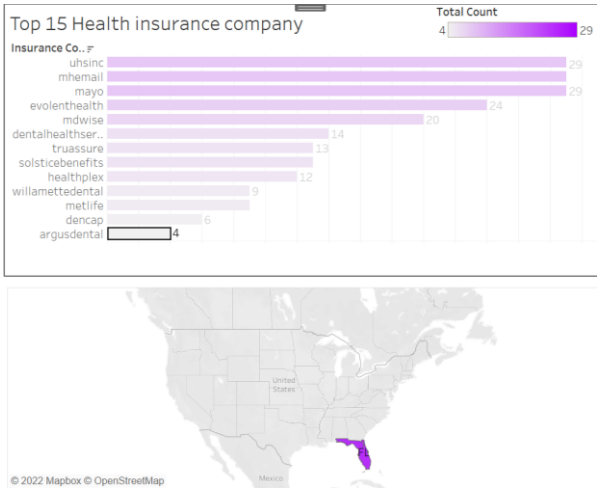

Figure 7 – Lowest health insurance company covered across one state

6.3 Bar graph and Stacked bar graph

From the below bar graph (Figure 8), HMOs are the highest subscribed plan with 43,,266, and PPOs, EPOs, POSs, and Indemnity respectively less subscribed plans from the year 2017 to 2021.

There are several types of plans available in the marketplace. (1) EPO: Exclusive Provider Organization (2) HMO: Health Maintenance Organization (3) POS: Point of Service (4) PPO: Preferred Provider Organization

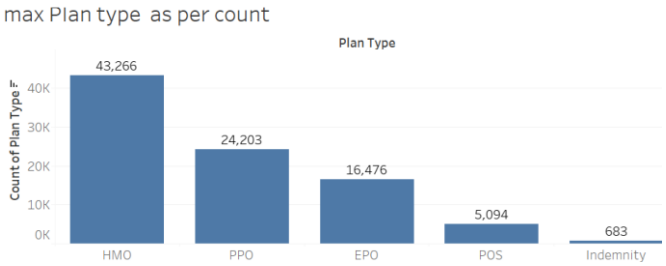

Figure 8 – Max plan type as per count for overall year distribution

Over the years (Figure 9), HMOs are the highest subscribed plan, and POSs and Indemnity are the least subscribed plans. But EPOs subscription increases and PPOs subscription decreases over time.

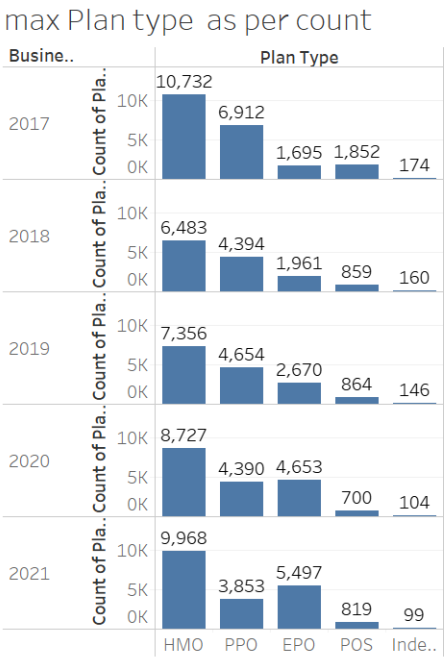

Figure 9 – Max plan type as per count for year-wise distribution

In a nutshell, we gathered deep insights into the HMO plan, and thus below-stacked bar graph shows the cost reduction value for the HMO plan over the duration between 2017 and 2021.

The various metal level is cost reduction parameters for a defined user set. For example, if a person has opted for the Silver plan, their cost gets reduced by 73%. Hence, this is the most opted plan combined with either HMO, PPO, POS, EPO, or Indemnity plans.

The stacked bar chart has two metrics: (1) EHB% (2) CSRV% (Figure 10).

EHB% defines the coverage of essential health benefits under an insurance plan and it's shown by the width size of the bar chart.

CSRV% defines the cost reduction percentage from the actual insurance cost, and thus represented by saturation in colors.

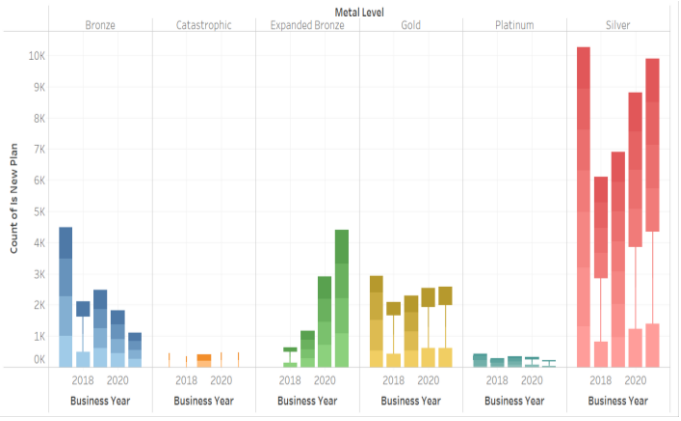

Figure 10 – CSRV% & EHB% for HMO plan over the duration between 2017 and 2021

In contradiction to the above analysis, we observe that HMO has the highest subscriber base, yet EPO has the most coverage across the USA.

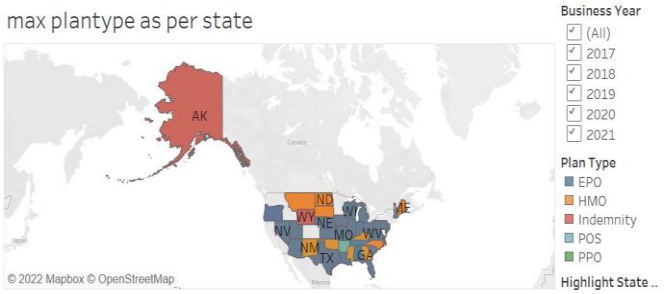

Figure 11 – Max plan type as per state

6.4 Pie chart and Tabular graph

The below pie distribution shows (Figure 12) the average count of a different age group issuing various health insurance plans.
The tabular graph (Figure 13)depicts the average cost for individuals, individuals and dependents, couples, couples, and dependents.
Therefore, average individual rates falling under various age groups gradually increase with 65 and over being the highest amount compared to the 0-14 age group.
There is a drastic difference in cost if an individual opts for the Family Option plan because the cost substantially decreases with co-dependents.

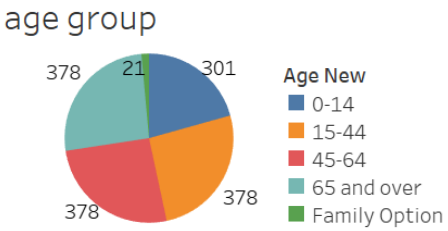

Figure 12 - Average count of different age groups

rate as per age group

Age New	Avg Ind..	Avg. In..	Avg. Co..	Avg. Co..
0-14	185.5	0.0	0.0	0.0
15-44	267.3	0.0	0.0	0.0
45-64	428.6	0.0	0.0	0.0
65 and over	481.6	0.0	0.0	0.0
Family Option	31.0	59.9	51.5	85.0

Figure 13 – Average rates as per age group

7. Conclusion

To summarize our presented analysis, we finally conclude that:
1) Dental plans are the most issued health insurance plan amongst children of age group 0-14 across the USA.
2) HMO & EPO plan types are the most subscribed plans having great network coverage, and service area and partnered with Tier-1 org across the globe.
3) EHB% and CSRV% should be higher in terms of getting maximum essential health benefits with a great reduction in value combined either with HMO or EPO plan types.
4) Individual rates are higher compared to codependents, thus ideally one should get a family option insurance plan to save more on pockets.
From the available dataset provided by the CMS website, we were able to achieve and provide great insights through our summarized dataset and analysis. In addition to this, further insights could be found using dashboards and code present in the GitHub repository.

References

[1] Dataset – Centers for Medicare & Medicaid Services (CMS) https://www.cms.gov/cciio/resources/data-resources/marketplace-puf

States included in the dataset:

AK	IA	MO	NM	TN
AL	IL	MS	NV	TX
AR	IN	MT	OH	UT
AZ	KS	NC	OK	VA
DE	KY	ND	OR	WI
FL	LA	NE	PA	WV
GA	ME	NH	SC	WY
HI	MI	NJ	SD	

[2] Ben Handel; Jonathan Kolstad-Smart policies on Health Insurance choice, University of California, Berkely, 2015 https://www.hamiltonproject.org/assets/files/smart_policies_on_health_insurance_choice_final_proposal.pdf

[3] National Association of Insurance Commisioners(NAIC)-2021 Annual Results, USA, 2021 https://content.naic.org/sites/default/files/2021-Annual-Health-Insurance-Industry-Analysis-Report.pdf

[4] American Medical Association(AMA)-Competition Health Insurance US markets, Chicago, 2021 https://www.ama-assn.org/system/files/competition-health-insurance-us-markets.pdf