Health Insurance Marketplace Analysis Using Hadoop

Authors: Ayushi Porwal, Zalak Patel, Clifford T Lin, Kajal Bhandare

Department of Information Systems, California State University Los Angeles
CIS5200 System Analysis and Design

aporwal@calstatela.edu, zpatel6@calstatela.edu, clin22@calstatela.edu, kbhanda3@calstatela.edu

Abstract

The health insurance marketplace in the United States represents a pivotal nexus within the nation's healthcare landscape, embodying the principles of accessibility, affordability, and coverage inclusivity. Envisioned under the Affordable Care Act (ACA) of 2010, this marketplace serves as a central platform where individuals, families, and small businesses can navigate and procure diverse health insurance plans.

This Term paper aims to analyze Health Insurance Marketplace data of the USA and provide insights on enrollment details, plan information, user demographics, and healthcare utilization statistics. We begin by describing the dataset and its features, including the plan types, participating network providers, geographical distribution, benefits offered, and insurance rates. We then conduct exploratory data analysis to identify trends and patterns in the data, such as the most sold benefits, and leading network providers distributed by states. Overall, our analysis provides valuable insights into the trends of the Health Insurance market in the USA and helps stakeholders analyze it for improvements.

1. Introduction

This project centers on scrutinizing the assortment of insurance plans provided by the US government, employing Hadoop and Hive for a comprehensive analysis. The focal point revolves around discerning trends within diverse health insurance plans accessible to individuals, families, and small businesses, emphasizing enrollment services. The project's selection stems from its rich and expansive attributes, which are ideal for in-depth analysis. Across three years, we meticulously assessed the prime benefits various network providers offer across states. Our analysis encompassed the evolution of plan rates over these years concerning variations in the number of dependents. We identified the prevalent types of plans and highlighted the top network providers within different states, conducting a thorough investigation into the popularity of various plan types. Furthermore, our analysis encompassed a detailed breakdown of the types of plans available across states and the distribution among insurance network providers. Extensive research into the Health Insurance marketplace concept preceded our utilization of a 3.15GB dataset sourced from Kaggle, comprising four .csv files. Implementation-wise, a Hadoop cluster version 3.1.2, featuring 2 master nodes and 3 data nodes, facilitated the execution of this project, enabling a profound exploration of the USA government's insurance plan landscape. We visualized the data using Tableau Software and Excel 2D and 3D Map. Overall, our analysis provides a unique perspective on the healthcare insurance marketplace dataset and offers valuable insights into benefits, rates, cost per person, and plan type and network provider trends in the healthcare domain.

2. Related Work

In recent years, there has been a growing interest in analyzing the Health Insurance Marketplace dataset to better understand the US Health Insurance industry. One such related work is a study by Sachin Selar[1] that focuses on analyzing the trends and finding interesting patterns in Health Insurance Marketplace data from year 2014 – 2016. He utilized the database to identify how plan rates and benefits vary across states using Python. His research helped us understand the terminologies used in Health insurance like ‘deductible’, ‘coinsurance’, and ‘out of pocket limit’.

Another example is a study by Caitlin N McKillop 1 2, Teresa M Waters 3 4, Cameron M Kaplan 3, Erin K Kaplan 5, Michael P Thompson 3 6, Ilana Graetz [2] that analyzes shifting market dynamics, such as increasing premiums and alterations in plan availability, raise concerns about its reliability for Americans. Analyzing instability's impact on consumer choices can reveal reasons for plan switching among those sensitive to prices. This insight can guide future policy decisions regarding the marketplace. Our study focused on examining spatial and temporal changes in benefits, diverse plan types, and rate fluctuations across the United States between 2014 and 2016. Employing Big Data through Hive, we scrutinized U.S. data, investigating trends in benefit plans and varying plan rates among individuals, couples, and dependents. Visualization was done using Tableau Software and Excel Power Map. Overall, our analysis presents a distinctive viewpoint on the Health Insurance marketplace dataset, yielding valuable insights into evolving trends within the USA's insurance plans and benefits landscape.

Another example is Dr. Sadrach Pierre's research, titled "Analyzing Health Insurance Market Data"[3]. Dr. Pierre's quantitative examination of health insurance plans and general healthcare services could significantly contribute to improving transparency in the healthcare system, especially regarding pricing and the implementation of value-based care. His focus lies in identifying the top network providers based on maximum out-of-pocket expenses and deductibles across various plan types, this study helps us to find out the pattern regarding network providers. In contrast, our case study concentrates on a distinct approach, analyzing leading network providers based on enrollment rates across states over the period spanning from 2014 to 2016. Additionally, our study considers different plan types across various states.

3. Specifications

The Health Insurance Marketplace dataset is the collection of all the information related to the insurance plans bought by individuals and small shops for businesses in the United States of America. The dataset contains four CSV files each for network, plan attributes, benefits plan, and insurance rates respectively. It contains data for three years (2014-2016) from all states which sum up to the size of 3.15 GB.

Table 1 shows the files and size of the files from the dataset.

Data Set Size	3.15 GB
Number for files	4
Content Format	CSV

Table 1 Data Specification

Table 2 below shows the specification for the Oracle cluster we are using and the Hadoop specification for our project.

Number of nodes	5 (2 master nodes, 3 worker nodes)
CPU speed	1995.312 MHz
Storage	390 GB

Table 2 Hadoop Specification

4. Implementation Flowchart

Initially, we downloaded the raw dataset from the Kaggle website to our personal computers. Later we used "scp" to copy the dataset from the local machine to our virtual Linux server. Once we got the dataset on the Linux server, we created directories in HDFS and pushed the dataset on HDFS for analysis. The whole process of data manipulation and process flow is shown in *Figure 1 Project Architecture Chart* and *Figure 2 Process Workflow.* The dataset isavailable in CSV. We downloaded the dataset and uploaded it to theHadoop File System. After that, HiveQL is used as a querying language to create the tables' schema, clean data, create a summary table, and export the results. Once the output file had been downloaded in CSV format, we used Excel's 3D map and Tableau to obtain the visualizations.

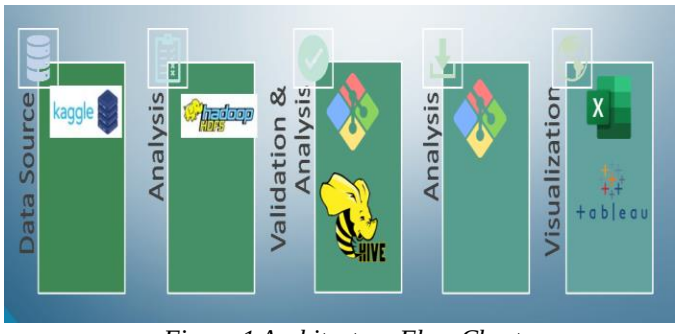

Figure 1 Architecture Flow Chart

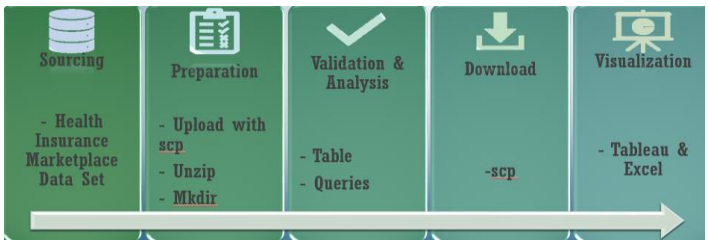

Figure 2 Process Flow Chart

5. Data Cleaning

We followed certain steps to perform data cleaning in our project. Downloading Raw Files from Source Website: This involves obtaining the original, unprocessed data files from their source on the internet. Uploading Files to a Linux Server and raw data files retrieved earlier are transferred and stored within a Linux server environment using HDFS, a distributed file system designed to handle large volumes of data. Beeline is a command-line interface that connects to Hive, a data warehouse infrastructure built on Hadoop. This step involves utilizing Beeline to load the data from the stored files into tables within the Hadoop ecosystem. The dataset comprises multiple CSV files and tables with a considerable number of columns. To streamline the analysis, a decision was made to create a "clean" table containing only the specific columns necessary for the analysis. A new table was constructed, likely through SQL queries, containing a subset of columns relevant to the analysis. This process reduces complexity and focuses on pertinent data for the intended analysis. Additionally, any rows containing NULL values were likely removed, ensuring data quality and usability, and addressing Sensitive Data and Dataset Readiness. This ensures that the data is sanitized and devoid of sensitive personal details, making it safe for analysis and usage. During the creation of external tables, serialization/deserialization (SERDE) functions were employed to specify how the data should be read and parsed. Using SERDE functions ensures that the tables are created with all data types defaulted to string datatype, providing a flexible approach to handling different data formats. Overall, these steps outline our systematic process of acquiring, storing, refining, and preparing data for analysis within a Hadoop-based environment, particularly focusing on the health insurance dataset in this scenario.

6. Analysis and Visualization

After data cleaning and preparation for further analysis, files were extracted into Tableau and Excel. We used different interactive maps to show statistics based on different states and three consecutive years i.e., average plan costs per person, leading network providers with highest enrollment rates, most popular plan type, and popular plan types among different network providers across all states of the USA

Insurance providers offered different types of plans across the state of the USA. Figure 3 is a visual representation of plan type across the state on an Excel 3D map. This visualization uses a bubble chart to represent each state with the size of the bubble indicating the plan count. The layer with the different colors of the bubble indicates the plan type counts for different plan types (EPO, HMO, Indemnity, POS & PPO) across the state. The figure shows that South Carolina state offers the most extensive variety of EPO plan types, Wisconsin state provides HMO and POS plans in abundance, and Ohio state boasts the highest number of Indemnity and PPO plans available.

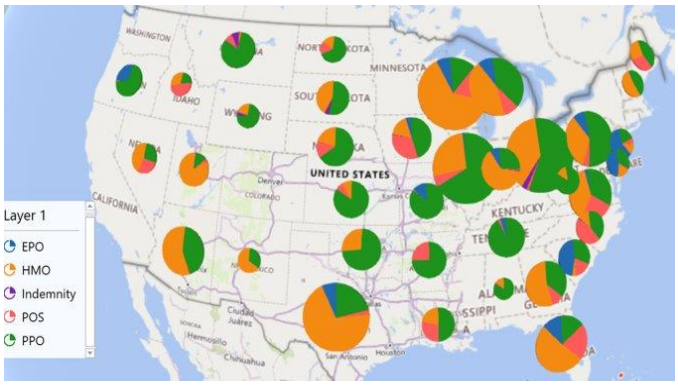

Figure 3 Plan Types Across States

Insurances provide many different types of plans and rates depending on the number of dependents being insured. Figure 4 is a visual representation of the average plan cost per person on an Excel bar graph. The costs per person are depicted on the y-axis and the different categories of plan holders are on the x-axis. The categories of plan holders are displayed as per health insurance terminologies purchasing insurance planeside., individuals, couples, dependent, dependent 2, dependent 3, couple dependent, couple dependent 2 and couple dependent 3.

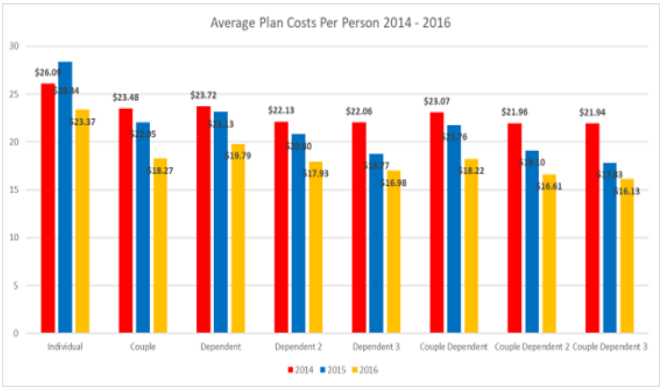

Figure 4 Average Plan Costs per Person

In Figure 5, the Excel bar graph provided depicts a visual representation of the top 5 benefits that were sold over three consecutive years, commencing from 2014. The horizontal axis (x-axis) is labeled with the names of the different benefit plans, while the vertical axis (y-axis) quantifies the total number of benefit plans sold. This graphical representation offers a clear and concise overview of the popularity and sales trends of specific benefit plans across the specified time frame. The figure indicates that the highest number of benefits was sold in 2015 among the three years. Additionally, Orthodontia - Adult stands out as the most frequently sold benefit across all years.

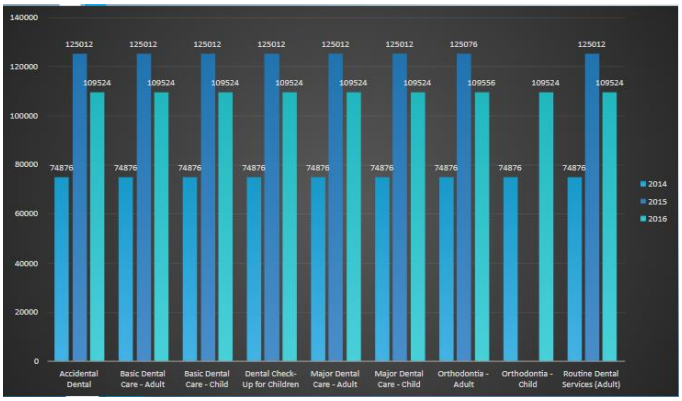

Figure 5 Top 5 Benefits Sold for Different Years

The dashboard in the next Figure 6 talks about the number of benefits offered across different states. The darkest shade of blue represents the highest count of benefits offered. Another graph shown shows the count of benefit plans on the y-axis and different states on the x-axis. The highest number of Benefits are provided in the state of Michigan followed by Arizona, and Florida. Idaho is the state with the lowest benefits plan offered - only 71!

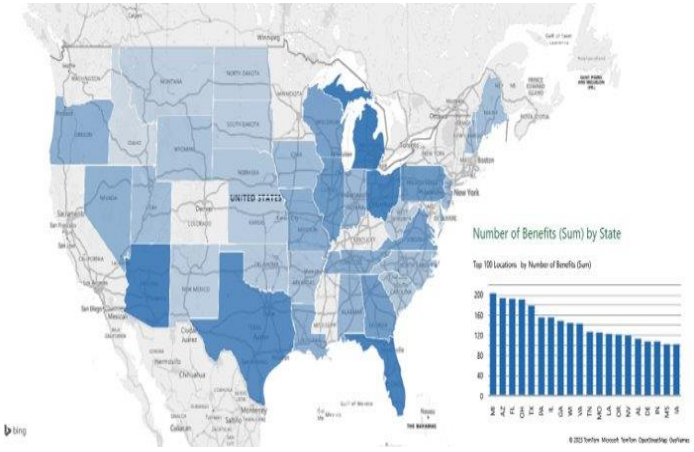

Figure 6 Benefits Changes Across States

Across the years 2014 to 2016, the trends in leading network providers across all states in the USA were visually represented through stacked bars in Figure 7, offering a clear insight into the shifting enrollment patterns. In 2014, HealthPlan Select held the top position in Ohio, boasting the highest enrollment rates, which peaked at 41.9k individuals. This highlighted its dominance within that specific state's insurance landscape for that year. The subsequent year, there was a notable shift in leadership as Blue Advantage HMO surged ahead, taking the lead in Texas with a substantial enrollment of 89.3k individuals. Its ascendancy marked a significant change in the competitive landscape, and notably, it managed to maintain its top position across the year. The enrollment figures surged to an impressive 47.6k individuals, indicating both its sustained popularity and its ability to attract even more enrollees, solidifying its dominant position within Texas' insurance market during that period.

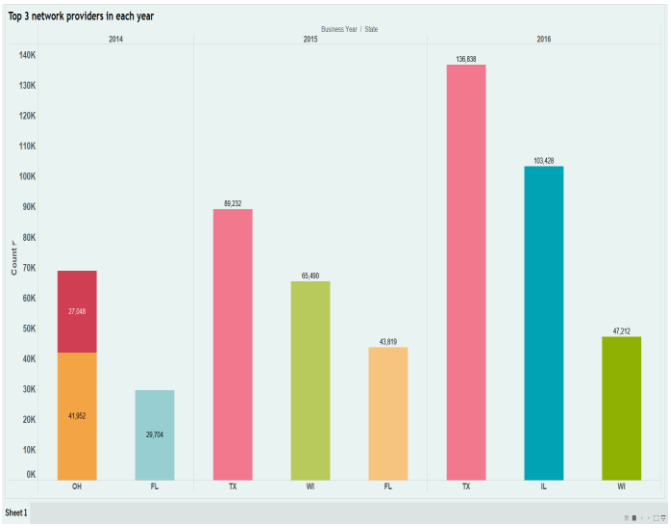

Figure 7 Leading Network Providers

Figure 8 is the visual representation of the total unique network provider per plan type. The categories of plan type (EPO, HMO, Indemnity, POS & PPO) are depicted in the x-axis, and the distinct network name counts are depicted in the y-axis. To visualize this analysis, an Excel area graph is used. The figure shows that the HMO plan comprises a total of 443 unique providers, the PPO plan includes 400 network providers, while the POS plan has 177 network providers. The EPO plan type encompasses 107 network providers, and the Indemnity plan type has the lowest network, with only 20 providers.

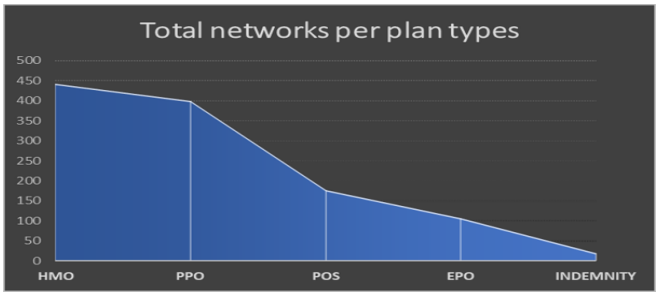

Figure 8 Total Unique Network Providers Per Plan Types

7. Conclusion

As a result of our analysis, we were able to identify that when individuals or couples add dependents to their insurance plans, there's a significant reduction in the cost per person. This is due to the pooling effect, where the overall cost of the plan is divided among more individuals, thereby decreasing the cost burden per person. Among the top-selling insurance plans, every benefit included falls under the category of dental benefits. This suggests a significant emphasis on dental coverage within these plans, possibly indicating the high demand or importance consumers place on dental care. The Blue Advantage Health Maintenance Organization (HMO) is a health insurance plan offered by Blue Cross Blue Shield Insurance Company and it stands out as a popular choice among various network providers. Its prominence suggests it's a preferred option among consumers in populated states like Texas and Florida, possibly due to its offered benefits, network coverage, or affordability compared to other providers. Among the array of plans available, the Health Maintenance Organization (HMO) emerges as the most popular choice. This indicates a consumer preference for this plan, which might be due to its cost-effectiveness, comprehensive coverage, or flexibility in accessing healthcare services within a designated network.

For more information, dashboards, and code visit the project's GitHublink.[2]

References

[1] shelars1985. (2017, November 13). Exploring Health Insurance Marketplace. Kaggle. https://www.kaggle.com/code/shelars1985/exploring-health-insurance-marketplace

[2] McKillop CN;Waters TM;Kaplan CM;Kaplan EK;Thompson MP;Graetz I; (n.d.). *Three years in - changing plan features in the U.S. Health Insurance Marketplace*. BMC health services research. https://pubmed.ncbi.nlm.nih.gov/29902996/

[3] Sadrach Pierre, Ph. D. (2020, September 16). Analyzing health insurance market data. Medium. https://towardsdatascience.com/analyzing-health-insurance-market-data-71b1cf00e97d

[1] The outlying islands (Alaska) are not considered in the list of states.

[2] GitHub Link: https://github.com/ayushiporwal13/HealthInsuranceAnalysis