**STAT 112 – Introduction to Data Processing and Visialization Project**

**Data-driven Insights into Health Insurance Dynamics**

**by**

Baran Erkek

Yiğit Sütlü

Efe Anık

Selin Aytan

İdil Bükre Tuzkaya

January 14, 2024

**ABSRACT**

Ever since the ancient times health has been the number one priority for the mankind. It is an undeniable fact that having good health is most probably the biggest desire of a human being rather than being wealthy and powerful. Without health all riches render useless. Health is a field of utmost importance as its hard to get it back once it has deterioted. Nowadays, purchasing insurance is one of the most common and effective method to secure oneself, since getting insurance prevents excessive expenditures. In today’s world, there are numerous types of insurance for every possible scenerio. In this analysis, the main focus is to explain health insurance within the framework of given data. This work is done by using descriptive statistics and exploratory data analysis techniques with proper data cleaning steps which utilizes features of Python libraries

**INTRODUCTION**

This analysis mainly aims to investigate health insurance preferences of a group along with their age, gender, level of income and health condition throughout years. The dataset contains data from 121 people with respect to 11 unique variables starting from the year 2012 to 2021. Overall, six research questions were asked in order to interpret given data precisely. Research questions as follows:

1. How does income level influence doctor visits? Do individuals with higher incomes visit doctors more frequently?
2. How does the level of income correlate with insurance coverage?
3. Is there a correlation betwen age and coverage level?
4. Do age and health condition have a directly proportional relationship?
5. How has the preference for HMO and PPO insurance types changed over the years?

**DATA CLEANING AND TIDYING STEPS**

To accuretly and effectively analize the data, it is essential to clean and prepare the data before using it. Otherwise the results will be misleading. Data cleaning steps as follows:

* First and foremeost, we examined and explained each variable individually.

| ***variable name*** | ***variable description*** | ***scale of measurament*** |
| --- | --- | --- |
| year | year of the observation | ratio, discrete |
| month | month of the observation | ratio, discrete |
| age | age of the individual | ratio, discrete |
| gender | gender of the individual (F & M) | nominal |
| income | annual income of the individual | ratio, discrete |
| insurance type | type of health insurance (HMO or PPO)\* | nominal |
| coverage level | coverage level of insurance (bronze, silver or gold)\*\* | ordinal |
| health condition | overall health condition of the individual | ordinal |
| doctor visit | numbers of doctor visits in the past year | ratio, discrete |
| premium amount  claims | monthly premium amount paid by the individual  number of insurance claims made | ratio, discrete  ratio, discrete |

\* HMO (Health Maintenance Organization): This type of insurance typically requires members to choose a primary care physician (PCP) and get referrals from the PCP to see specialists.

PPO (Preferred Provider Organization): This type of insurance offers more flexibility in choosing healthcare providers. Members can see specialists without a referral, both inside and outside the provider network.

\*\* Bronze: Basic coverage with lower premiums and higher out-of-pocket costs.

Silver: Moderate coverage with a balance of premiums and out-of-pocket costs.

Gold: Comprehensive coverage with higher premiums and lower out-of-pocket costs.

* To make sure that we have imported data correctly we checked first and last 5 rows and didn’t find any inconsistencies besides header names.
* At first there wasn’t any column name so we have written suitable names for each column. Then, we have removed (“”), ($$), (##), (-) signs from the variable names.
* Every column contains relevant information for our EDA, so we didn’t drop any.
* No duplicate variables were detected after detailed investigation.
* Data in the claim column that were written in letter rather than numbers were converted to float. And instead using full words, we have changed “Females” to F and “Males" to M.
* We had a mixture of upper and lowercase letters in our data so to make them all uniform we have used *“.str.capitalize()”* function.
* We have sorted years and months chronologically and combine them with pd.to\_datetime function by creating a new column.

*df['Date'] = df['YEAR'].astype(str) + ' ' + df['MONTH']*

*df['Date'] = pd.to\_datetime(df['Date'], format='%Y %B')*

* Ranges and standart deviations of numerical variables in our dataset such as age, claim, income, doctor visit and premium amount were calculated in order to examine existence of any unusal behavior. As a result, no unusual behavior were detected.
* We have calculated the IQR for every numerical column. There weren’t any value above or below 1.5 IQR. As a result of our calculations, no outliers were detected.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Income** |  | **Age** |  | **Premium Amount** |  |
| QTL1 | 55750 | QTL1 | 34,75 | QTL1 | 350 |
| QTL3 | 87000 | QTL3 | 50,25 | QTL3 | 500 |
| IQR | 31250 | IQR | 15,5 | IQR | 150 |
| Lower limit | 8875 | Lower limit | 11,5 | Lower limit | 125 |
| Upper limit | 133875 | Upper limit | 73,5 | Upper limit | 725 |

* When we checked our dataset once again in terms of units, none of the variables required unitial transformation except income column. For income column, since no information was given we accepted all values as dollars.
* NA values percentage was less than %60 so we replaced them with median and modes accordingly. For NA values in the age, income, premium amount and claim columns we have replaced them with median since the values were numerical. NA values in gender, insurance type, coverage level and heath condition columns were replaced with mode since the variables are categorical which makes median useless. Since percentage was low we didn’t drop any columns.

**EXPLORATORY DATA ANALYSIS**

Exploratory Data Analysis is an approach used for analyzing and interprating datasets in order to summarize their key features by utilizing statistical tools. EDA targets to understand what data tell us properly and identify correlations between variables. This process is essential for further analysis. In general five research questions were investigated. The process and descriptive statistics as follows:

metin, yazı tipi, siyah beyaz, sayı, numara içeren bir resim

Açıklama otomatik olarak oluşturuldu

**Ouestion 1:** **How does income level influence doctor visits? Do individuals with higher incomes visit doctors more frequently?**

metin, ekran görüntüsü, diyagram, öykü gelişim çizgisi; kumpas; grafiğini çıkarma içeren bir resim

Açıklama otomatik olarak oluşturulduIn today’s world, there is a common belief that individuals with higher income level spend more time on their health. This question targets to reveal whether wealthy people visit doctors more frequently not. In order to show the relation, a bar chart was created. As stated in the graph, it is evident that as the income level increases, the frequecy of doctor visits increases as well.

metin, ekran görüntüsü, diyagram, öykü gelişim çizgisi; kumpas; grafiğini çıkarma içeren bir resim

Açıklama otomatik olarak oluşturuldu**Question 2: How does the level of income correlate with insurance coverage?**

In this question we aimed to find the effects of income levels on coverage levels. Overall we can observe that people with incomes 40-100k has gotten the most insurance. Bronze insurance was least chosen among all coverage types. Only in the income of 40k-50k it was dominantly chosen. Its rates have dropped signifacantely after that.

On the other hand, highest income group only picked gold but their population is too small. The middle groups distribution was mostly balanced. With each income getting higher gold popularity increased and other two type’s decreased. We can interpret that the trend is as income increases there is a shift to the gold from bronze.

**Question 3:** **Is there a correlation betwen age and coverage level?**

diyagram, ekran görüntüsü, dikdörtgen, kare içeren bir resim

Açıklama otomatik olarak oluşturuldu In this particular graph we aimed to find the relationship between age and coverage level preferred. Age is discrete numerical variable and coverage level is categorical. We have used a box plot to demonstrate its spread and distribution. Bronze is mainly prefered among the younger age groups. Bronze coverage level is commonly chosen by middle age groups with a median of 35.5. It has the second highest interquartile range with 13.75 which indicates high variability. As a result of its high variability, it has a high spread . Distribution is skewed to right. Silver level is prefered by slitghly older individuals with a median of 41.0. It has the smallest range 10.25 among all of the coverage levels. Distribution is bell shaped which means it has a even spread. Gold is prefered by older individuals with a median of 48.5. It has the highest spread 18.75 and from this information we can conclude that it is prefered by many age groups. Overall we can see that as people get older they are more likely to switch from bronze coverage level to gold coverage level. In addition, no outliers were detected.

**Question 4: Do age and health condition have a directly proportional relationship?**

diyagram, origami içeren bir resim

Açıklama otomatik olarak oluşturulduIt is possible to count plenty of factors that effects health condition. For this question, our aim is to show the relationship between age which is a numeric variable and health condition which is a categorical variable via violin plot. The reason violin plot was used is to give more detailed information about distribution density. Also; descriptive statistics can easily be examined.

Observing the big picture at first sight, it is possible to notice fluctuations on the first graph whereas the rest display more consistent distribution. The majority of people in fair condition are aged between 30-40 with a median of 36.5. it has the smallest spread out of all the health conditions. Most of population is centered around approximately 34. It has a interquartile range of 9.0 which suggest low variability. For those in good health condition the range goes up to 45-55 median is 48.0 which is the most of every health condition. the age demographic of those in hood health is the highest which also means high spread. Individuals over the age of 70 and under the age 20 are only included in this group. we can see that people who have good health is increase as they age. Mean while we can see the opposite effect in excellent health condition. Individuals in excellent condition predominantly has age range 35-45 with a median of 41.0 which is the middle value. There are no fluctuations but a steady decrease in number as they got older. Its interquartile range is 13.5 which suggest moderate variability. In conclusion These insights underscore the intricate interplay between age and health, crucial for tailored healthcare strategies

**Ouestion 5: How has the preference for HMO and PPO insurance types changed over the years?**

metin, ekran görüntüsü, paralel, dikdörtgen içeren bir resim

Açıklama otomatik olarak oluşturulduThe stacked bar chart illustrates changes in the distribution of insurance types (HMO and PPO) over the years 2012 to 2021. Each bar represents a specific year, with the total height of the bar representing the total number of insurance policies for that year. The individual segments within each bar represent the proportion of HMO and PPO policies. HMO policies have consistently been the majority over the years wheras PPO policies have consistently been the minority compared to HMO.

The number of HMO policies has varied between 6 and 8 and the highest count of HMO policies was observed in 2016 and 2021 with 8 policies. On the other hand, the number of PPO policies has varied between 4 and 6 and the highest count of PPO policies was observed in 2013 and 2014 with 6 policies.

In summary, HMO policies consistently dominate the distribution, with the number of HMO policies being higher than PPO policies each year. While there are variations in the exact numbers, the overall trend indicates a preference or prevalence of HMO policies over the specified period.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |  |  |

**CONCLUSION**

In conclusion, our analysis displays a positive correlation between income and doctor visits. a preference shift from bronze to gold coverage with increasing income and age, and a consistent dominance of HMO policies. These insights provide practical implications for stakeholders. While our analysis spans 2012 to 2021, future research can explore specific demographics for a more nuanced understanding. Overall, our collaborative effort aims to inform decision-making in the dynamic healthcare landscape.