Zohaib Sheraz Mughal co57jiku

MADE Final Analysis Report

How do demographic factors like income, age, and, sex influence health insurance charges/coverage across different regions?

Introduction:

Discrimination of healthcare services is an important issue in a lifestyle of a person with an average/survival source of income across different demographic groups and regions in America. This problem is crucial because there can be affordability issues of healthcare accessibility among diverse regions across different demographics. This problem is addressed by identifying the relation between disparities/factors like income, age, gender, region/location and Health care charges/coverage that may lead to this unhealthy behavior. The insights aim to uncover the demographic groups that face significant barriers to healthcare accessibility, affordability and treatment.

Used Data:

a. Overview:

There are 2 datasets produced as a result of our ETL pipeline as the SQLite files. The first one is about the US demographics and the second one is about the US Health Insurance.

b. Data Structure:

The **US demographics** dataset contains 33 columns and the **US Health Insurance** contains 8 columns. For the dataset 1(US demographics), only the columns which are directly or closely related to the problem are selected, and for dataset 2(US Health Insurance) all the columns are selected. Both datasets have all the values present in that helps us to focus on the efficient analysis rather than data cleansing and hopefully build a strong foundation to answer the question, fully or partially. In our ETL pipeline only the columns are processed which contains sufficient amount of information.

c. Licenses:

Both data sources are licensed under the **CC0: Public Domain**, allowing unrestricted use, which includes modification, publishing etc. CC0 empowers the choice to opt out of copyright, and the exclusive rights automatically granted to creators – the “no rights reserved”. Although there are no restrictions as per the definition of CC0, but still, we respect its public availability and ensure the ethical use of the data, and transparency in analysis for our project.

Analysis:

a. Preparation:

The datasets are merged on a **region** column before carrying out the analysis. The dataset 1 contains the **State** feature with 52 unique values and dataset 2 contains the **region** feature with 4 unique values. The State feature is transformed/mapped from 52 to 4 unique values of **region** column from dataset 2 and State values are verified from US Department of Labor official website.

For e.g. **California** from **State** column in dataset 1 will be mapped to **southwest** value in the newly created **region** column in **dataset 1** replacing State column. The unique values of **region** feature are **'southwest'**, **'southeast'**, **'northwest'**, and **'northeast'**.

b. Method:

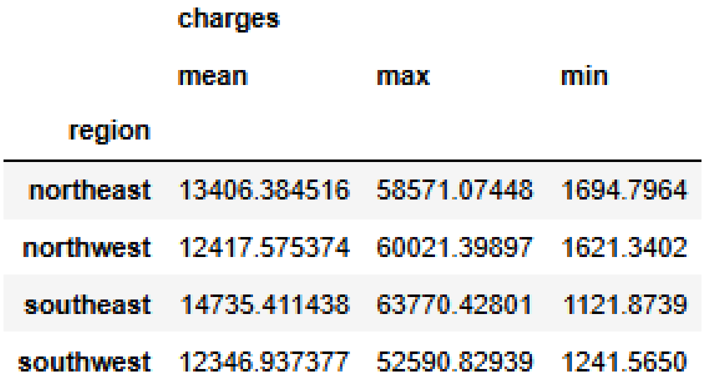
To understand if there is any relationship between the US Demographics and the Health Insurance coverage, we perform **Exploratory Data Analysis (EDA)** to group the features to generate the insights, and visualize to understand more intuitively the accessibility and affordability across different demographics.

In this methodology, we group the different regions with the average income, average insurance charges, and average charges per region with sex/smoker. This grouping provides a holistic view of the statistics of the different regions and the actionable insights can be generated. It enables a better interpretation of the datasets produced that can help for better decisions for different regions.

c. Results:

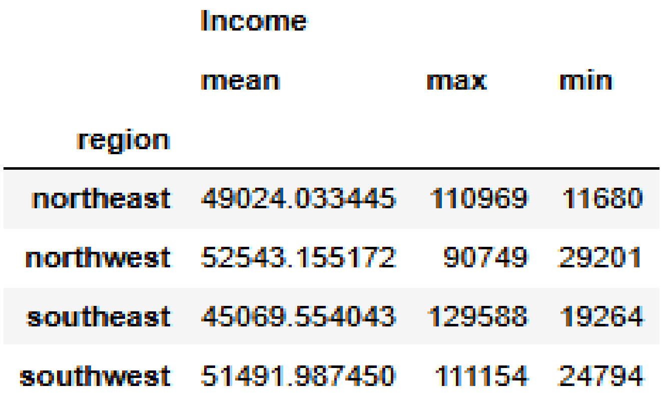
Below are some interesting results per region with income, sex, and smoker:

a. Average, maximum and minimum Charges region wise:



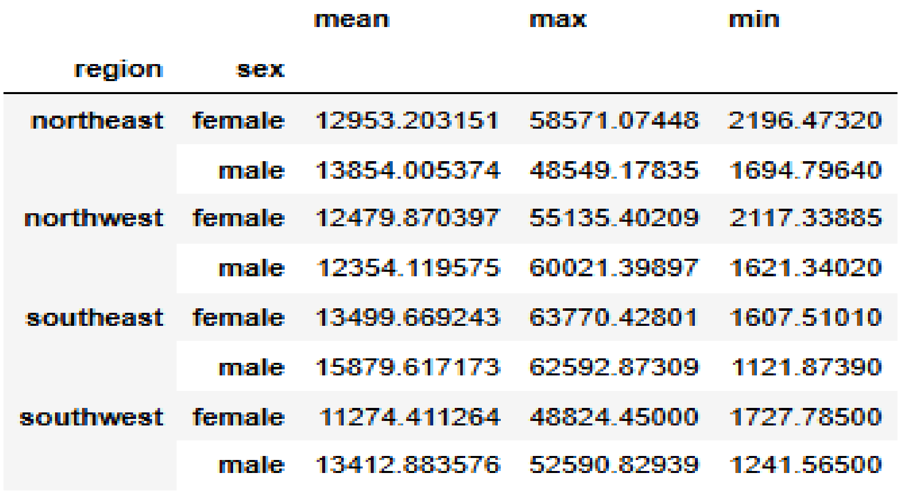
	charges		
	mean	max	min
region			
northeast	13406.384516	58571.07448	1694.7964
northwest	12417.575374	60021.39897	1621.3402
southeast	14735.411438	63770.42801	1121.8739
southwest	12346.937377	52590.82939	1241.5650

b. Average, maximum and minimum Income region wise:



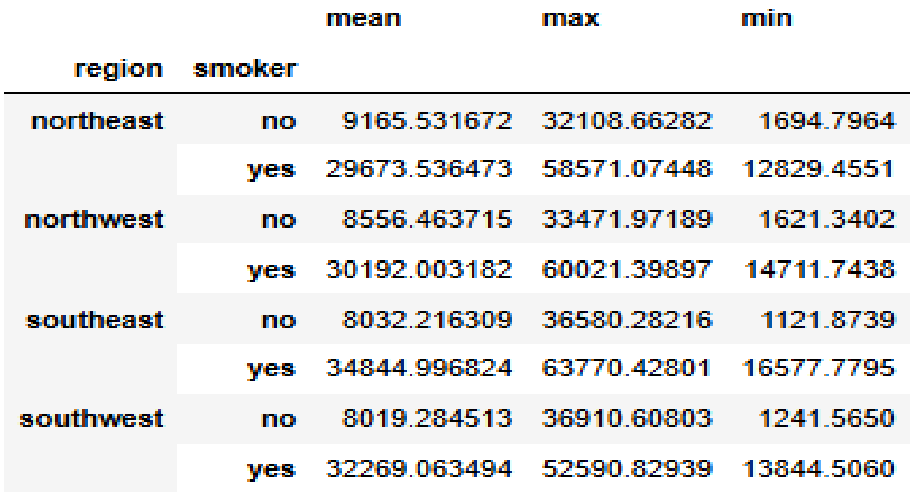
	Income		
	mean	max	min
region			
northeast	49024.033445	110969	11680
northwest	52543.155172	90749	29201
southeast	45069.554043	129588	19264
southwest	51491.987450	111154	24794

c. Average charges region and gender wise:



		mean	max	min
region	sex			
northeast	female	12953.203151	58571.07448	2196.47320
	male	13854.005374	48549.17835	1694.79640
northwest	female	12479.870397	55135.40209	2117.33885
	male	12354.119575	60021.39897	1621.34020
southeast	female	13499.669243	63770.42801	1607.51010
	male	15879.617173	62592.87309	1121.87390
southwest	female	11274.411264	48824.45000	1727.78500
	male	13412.883576	52590.82939	1241.56500

d. Average charges region and smoker wise:



		mean	max	min
region	smoker			
northeast	no	9165.531672	32108.66282	1694.7964
	yes	29673.536473	58571.07448	12829.4551
northwest	no	8556.463715	33471.97189	1621.3402
	yes	30192.003182	60021.39897	14711.7438
southeast	no	8032.216309	36580.28216	1121.8739
	yes	34844.996824	63770.42801	16577.7795
southwest	no	8019.284513	36910.60803	1241.5650
	yes	32269.063494	52590.82939	13844.5060

d. Interpretation:

The **result (a**) provides the basic statistics about the different regions resulting the estimated/average charges a person has to pay in order to be treated in a well manner. Next the **result (b)** provides us with the average incomes per region to showcase the affordability of the healthcare services. Following that, the **result (c) and (d)** shows the breakdown of the health insurance charges in different regions with person being male or female and smoker or not.

Furthermore, we can infer, by comparing **result (a)** and **(c)** that Male's average is slightly more in approximately 3 regions than the normal average. In addition, it can be clearly seen, in **result (a)** and **(d),** the smoker's average is significantly higher than the averages reported in **result (a)**. The minimum and maximum statistics are also reported just for the comparison with the average in case anyone is interested.

The incomes reported also reveal when compared with the healthcare charges that it is quite a good amount which is spent on the healthcare services. The males earning lower than healthcare average might face the significant barriers to the healthcare facilities, same issue females can face if consider earning less.

Conclusions:

In a nutshell, there are some demographic groups which face issues of the healthcare accessibility and affordability that lead to the existence of (partial) relationship between different demographics across different regions. Specifically, the analysis resulted the male and the smoker group might encounter the problems when healthcare services and charges are not in an affordable range. In minor injury/issue, it may not be a big issue but for a severe incident, it may lead to the significant loss.

Limitations:

Although the analysis could produce some insights but there are still limitations / gaps that can be fulfilled and more meaningful insights can be generated. Below are some of them:

1. Both datasets don't contain any common column which can be more beneficial for our purpose
2. Datasets aren't fully suited to the problem defined, but are the best could be found at the moment
3. Lack of temporal data, no trend information/insights
4. Age attribute isn't useful in term of comparison with charges, but might be improved with further preprocessing