WOMEN TECHSTERS FELLOWSHIP

CLASS OF 2024

MONTH 2 MINI PROJECT

HEALTH INSURANCE ANALYSIS

2.0 TABLE OF CONTENTS

HEALTH INSURANCE DATASET PROJECT

3.0 INTRODUCTION

Technology for Social Change and Development Initiative (Tech4Dev) is a non-profit social enterprise established in 2016 that creates access to decent work and opportunities for Africans through digital skills empowerment and advocacy. Her vision is to equip Africans with digital and life skills that foster economic prosperity, financial freedom, and sustainable development, contributing immensely to the tech ecosystem and the economy of the world at large, starting with Africa.

3.1 AIMS OF THE PROJECT

This project was given by our facilitator for Data Science and Artificial Intelligence (Group C) Yasmin Osama. The goal is to give practical experience obtainable in a business environment as a data scientist. The main idea is to encourage individual analytical thinking of the project. As a fellow in the Data science and Artificial intelligence track, this project is part of the prerequisite to completing the course.

3.2 PROJECT OBJECTIVES

The main objective of this project is to analyze the given dataset to understand patterns and trends in the data, and to create a compelling Power BI dashboard that showcases our analytical prowess and discover hidden relationships between the features and to draw conclusions relating to how factors like region, age, gender and pre-existing conditions affect a customer's insurance charges telling a meaningful story with the data.

3.3 ABOUT THE DATASET

The dataset contains 1338 rows of insured data, where the insurance charges are given against the following attributes of the insured: Age, Sex, BMI, Number of Children, Smoker, and Region. The attributes are a mix of numeric and categorical variables.

Columns discretion:

- age: age of primary beneficiary
- sex: insurance contractor gender, female, male
- bmi: Body mass index, providing an understanding of body, weights that are relatively high or low relative to height, objective index of body weight (kg / m ^ 2) using the ratio of height to weight, ideally 18.5 to 24.9.
- children: Number of children covered by health insurance / Number of dependents.
- smoker: Smoking
- region: the beneficiary's residential area in the US, northeast, southeast, southwest, northwest.
- charges: Individual medical costs billed by health insurance.

3.4 MATERIALS AND TOOLS

To generate a good insight, the below materials and tools were used for analysis and visualization.

1. Insurance Dataset.
2. Microsoft PowerBI.
3. Microsoft Word.

4.0 DATA DESIGN AND ANALYSIS

4.1 Data Pre-Processing

To import the insurance dataset into POWERBI desktop, the following steps were carried out.

1. On the PowerBI Desktop, click on Get data.
2. On the Get Data option, click on Text/Csv file.
3. Transform the data into power query.

4.2 Data Cleaning

1. Check for Null values using column profiling. Column profiling helps to check if there are null values in any of the columns.

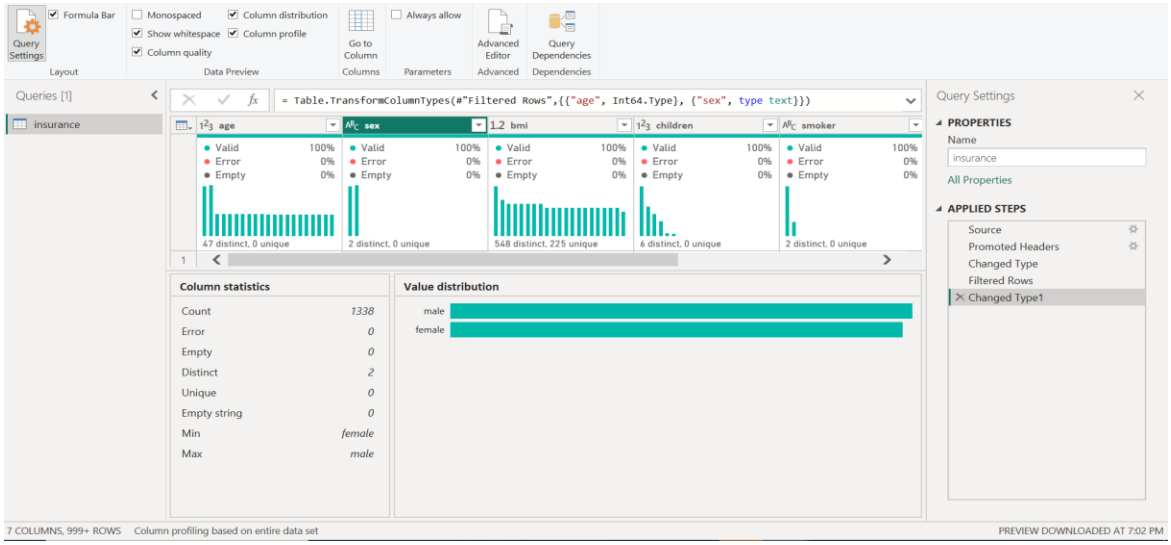

Diagram I

2. Check the data type for each of the columns.
3. Create a new unique ID for each individual. Add an index and begin from 1(Data type should be Text)
4. Creating a new column for Age group (ranging from 60-Above, 50-59, 40-49, 30-39, 20-29 and below 20), Age Category (Youth, Young Adult, Old Adult and Elder) and BMI Category (Underweight, Normal, Overweight and Obese) This aims at categorizing the Age and BMI columns.
5. Change the data types to text.
6. Apply and close.

4.3 FINDINGS FROM THE DATA

FINDING I

We want to know the region that is mostly represented among the beneficiaries and has the highest insurance charges?

INSIGHT

The result shows that Southeast has the highest beneficiaries with an average sum of $14,735.41.

FINDING II

We want to know the percentage of the beneficiaries who are smokers and non-smokers.

INSIGHT

The result shows that 79.52% of the beneficiaries are non-smokers while only 20.48% are smokers.

FINDING III

We want to know the correlation between insurance charges, BMI category and smokers.

INSIGHT

The result shows the Average insurance charges for smokers whose BMI is above Normal is x2 higher than the sample average (18.5 -24.9)

FINDING IV

We want to know the average charges for each age group.

INSIGHT

The result shows that insurance charges for beneficiaries above 60 years are higher with an average charge of 21,248.02 with beneficiaries under 20 years having the least charges with 8,407.35.

FINDING V

We want to know the number of beneficiaries by age category.

INSIGHT

The result shows that Old Adults are more beneficiaries with 573 while Elders have the least beneficiaries with 91. 417 customers belong to the youth category and 257 customers are young adults.

FINDING VI

How does the age bracket affect the individual medical cost billed by the insurance company (distribution)

INSIGHT

The result shows that there is a correlation between the age groups and the charges. It is observed that the charges increased with the age group with the customers with age group below 20years paying the least average amount of $8,407 and the customers within the age group above 60years paying the highest average amount of $21,248.

FINDING VII

Which gender paid the highest insurance charge?

INSIGHT

The result in our findings shows that the male gender pays an average insurance charge of $19,956 which is higher than the average amount of $12,569 paid by the female gender.

FINDING VIII

We want to find out what the total cost of insurance is by region.

INSIGHT

The total cost of insurance for the North-east region is $4,343,668.

The total cost of insurance for the North-west region is $4,035,712.

The total cost of insurance for the South-east region is $5,363,689.

The total cost of insurance for the South-west region is $4,012,754.

The result shows that the South-east region pays the highest amount of insurance while the lowest amount of insurance charges is from the South-east.

FINDING IX

We want to know if the number of children of beneficiaries has an influence on their insurance charges.

INSIGHT

The result in our findings shows that beneficiaries who have 3 children have the highest insurance charges with average charges of $15,355.32 while customers with 5 children have the least average insurance charges of $8,706.04.

FINDING X

We want to know if the BMI Category of beneficiaries has any correlation with the average insurance charges of customers.

INSIGHT

The result in our findings shows that beneficiaries whose BMI is above the Normal standard range (18.5 – 24.9) are being charged more. Customers who are Obese has highest average insurance charges of $15,491.54.

FINDING X1

We want to find out whether being a smoker or not affects the average insurance charge of a beneficiary?

INSIGHT

The result from our finding shows that beneficiaries who are smokers are being charged more with an average insurance charge of $32,050.23 (79.17%) which is approximately 4x the insurance charges for non- smokers.

5.0 DATA VISUALIZATION

After a thorough analysis of several questions to generate an insight, we created various charts and graphs to visualize the relationships between different variables, such as ribbon to show the correlation between Smoker, BMI and gender, and table to compare the average insurance charges for different regions using Microsoft Power BI tool. An interactive dashboard is then built for better presentation to the stakeholders.

6.0 CONCLUSION

The analysis of the insurance dataset has revealed several key findings that shed light on various factors influencing insurance charges among beneficiaries. These insights can inform strategic decisions and interventions aimed at managing and optimizing insurance costs. Here are the main conclusions drawn from the findings:

- ***Regional Impact on Insurance Charges:*** Southeast dominates with the highest beneficiaries and charges.
- ***Smoking Trends -*** This insight emphasizes the prevalence of non-smokers in the dataset.
- ***BMI and Smoking Impact:*** Smokers with a high BMI incur double the average charges.
- ***Age Variation:*** Charges rise with age, with those above 60 facing the highest costs.
- ***Demographic Distribution:*** Old Adults lead beneficiaries, while Elders have the fewest.
- ***Gender Disparity:*** Males pay significantly higher charges than females.

7.0 RECOMMENDATION

These recommendations aim to optimize insurance costs, improve health outcomes, and enhance overall customer satisfaction. Tailoring strategies to specific demographics and addressing lifestyle-related factors can contribute to a more effective and efficient healthcare insurance system.

- Target interventions in the Southeast, the dominant region with higher charges.
- Launch smoking cessation programs to reduce the smoking rate.
- Implement health programs targeting smokers with high BMI to mitigate elevated charges.
- Develop age-specific insurance plans to cater to varied healthcare needs.
- Tailor marketing strategies based on the dominant Old Adults demographic.
- Design gender-specific health initiatives to address charge disparities.
- Investigate and manage costs in regions with high insurance charges.
- Conduct in-depth studies on the impact of children on charges, BMI correlations, and smoking habits.
- Implement campaigns to educate customers on lifestyle impacts on charges.
- Establish a system for ongoing monitoring of demographics, behaviors, and charges.