Health Insurance Data Analysis

Team Members

* Malak Lebdjiri
* Aderonke Obatuyi
* Lucy Gachogu

Project Documentation

20th December 2023

Health Insurance Data Analysis

**CONTENT**

* Objective
* Introduction
* Methodology
* Insights
* Recommendations
* Conclusion
* **Objective**: Analyze the attached dataset and create a compelling Power BI dashboard that showcases your analytical prowess. Discover the hidden relationships between the features and use the data to tell a meaningful story.
* **Introduction**:

This project focuses on the analysis of a health insurance dataset, aiming to uncover key insights and trends that could potentially influence insurance charges. The dataset, provided to us, comprises various demographic, health, and financial variables of 1338 customers, including age, gender, number of children, BMI, smoking status, region, and charges.

The methodology adopted for this project involved several stages. While no data gathering or cleaning was required, data transformation was performed to enhance the dataset for analysis. The transformed data was then visualized using a dashboard, enabling us to derive meaningful insights and make recommendations.

The insights derived from this analysis provide a comprehensive understanding of the factors influencing health insurance charges. These insights can be valuable for health insurance companies to understand their customer base and make informed decisions.

This document will further detail the methodology, insights, recommendations, and conclusion of the project. It will also discuss the specific roles of the various analysis tools used.

* **Methodology:**

**Dataset Description:**

The dataset provided for this project is a comprehensive collection of health insurance data for 1338 customers. It includes the following variables:

1. **Age**: This represents the age of the primary beneficiary, ranging from 18 to 64 years.
2. **Sex**: This indicates the gender of the insurance contractor. The dataset includes 662 females and 676 males.
3. **Children**: This shows the number of children covered by health insurance or the number of dependents, ranging from 0 to 5.
4. **BMI**: Body Mass Index provides an understanding of the body, indicating weights that are relatively high or low relative to height. It’s an objective index of body weight (kg / m ^ 2) using the ratio of height to weight, ideally 18.5 to 24.9. In this dataset, it ranges from 15 to 53.
5. **Smoker Status**: This indicates whether the beneficiary is a smoker or not, represented as ‘yes’ or ‘no’. The dataset includes 1064 non-smokers and 274 smokers.
6. **Region**: This represents the beneficiary’s residential area in the US, categorized as Northwest, Northeast, Southwest, or Southeast.
7. **Charges**: This shows individual medical costs billed by health insurance, ranging from $1121 to $637770.

The dataset also includes an overview of the distribution of these variables among the customers. For instance, it provides information on gender distribution, smoking status, BMI categories, regional distribution, number of children, and age distribution. This data is crucial for understanding the demographic and health characteristics of the customers, which can influence their health insurance charges.

**Data Preparation:**

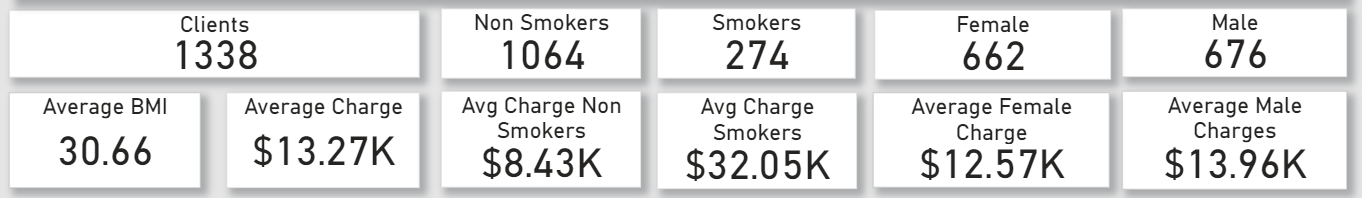
**Data Transformation:**

* 1. **Data Profiling:** This step involves checking the quality of the data to ensure its accuracy and completeness.
  2. **Adding an Index Column:** We added a new column, converted its data type to text, and used it to uniquely identify each individual in the record. We named this column “Clients”.
  3. **Creating Age Groups:** We utilized the bins feature to categorize individuals into age groups (10s, 20s, 30s, 40s, 50s and 60s).
  4. **Creating BMI Categories:** We added a new column to categorize individuals based on their Body Mass Index (BMI) into four categories: Underweight, Healthy Weight, Overweight, and Obese.

**Visualizations:**

The visualization used in this project is a Power BI dashboard that displays various data visualizations including bar graphs, pie charts, cards, clustered column chart, stacked bar chart and a slicer.

1. **Cards:**



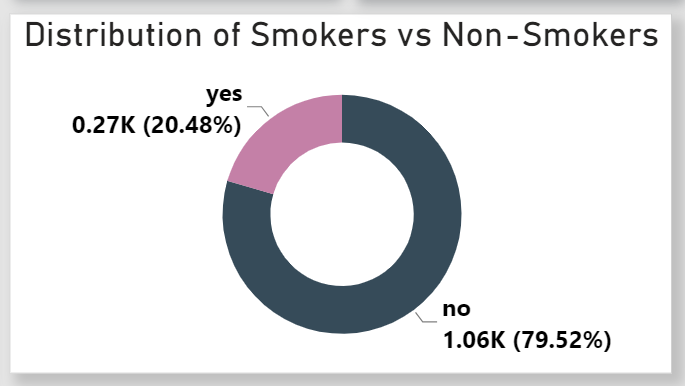
* + 1. **Number of Clients**: This card displays the total number of customers in the dataset, which is 1338.
    2. **Average BMI**: This card shows the average Body Mass Index of the customers, which is 30.66.
    3. **Average Charge**: This card presents the average insurance charge for all customers, which is $13.27K.
    4. **Average Charge for Non-Smokers**: This card displays the average insurance charge for non-smokers, which is $8.43K.
    5. **Average Charge for Smokers**: This card shows the average insurance charge for smokers, which is $32.05K.
    6. **Average Female Charge**: This card presents the average insurance charge for female customers, which is $12.57K.
    7. **Average Male Charge**: This card displays the average insurance charge for male customers, which is $13.96K.

1. **Slicer**:



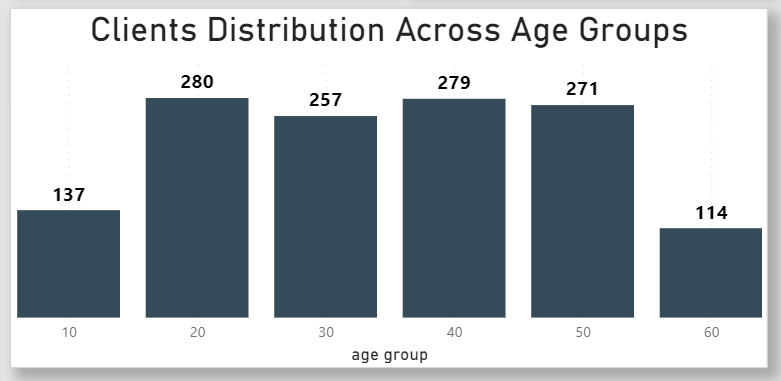
The “Region” slicer in the Power BI dashboard allows users to filter data based on the beneficiary’s residential area in the US. This interactive tool enhances the dashboard’s interactivity, enabling region-specific analysis of trends and patterns in the data.

1. **Pie Chart**:



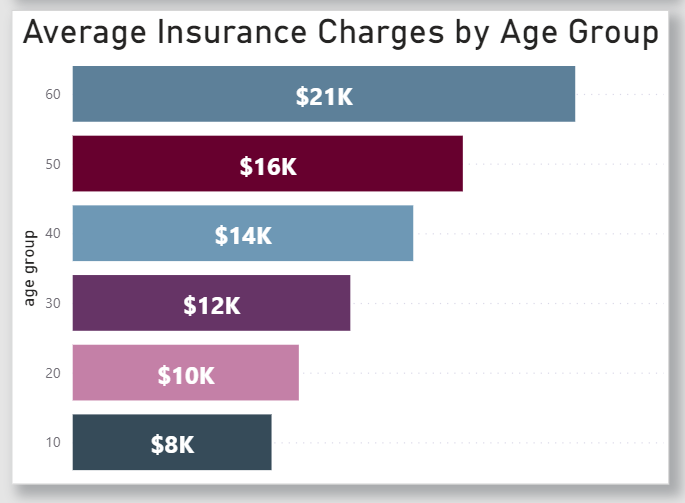
This chart shows the proportion of smokers and non-smokers in the dataset. It provides a clear view of the smoking status of the customers.

1. **Clustered Column Chart**:



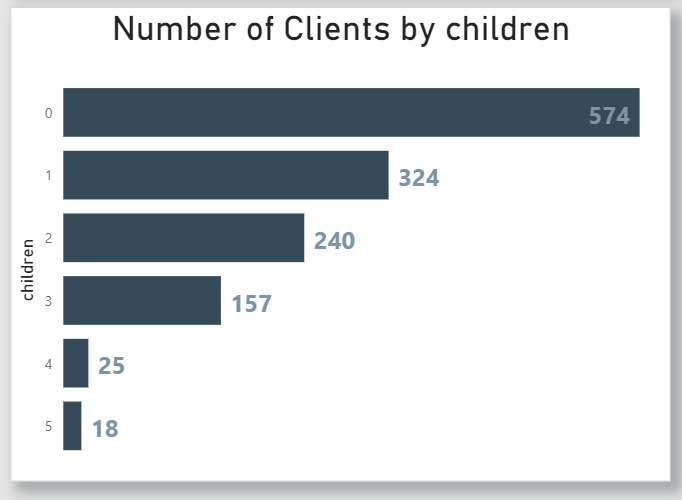
This chart displays the distribution of customers across different age groups. It helps to understand the age demographics of the customers.

1. **Stacked Bar Chart**:



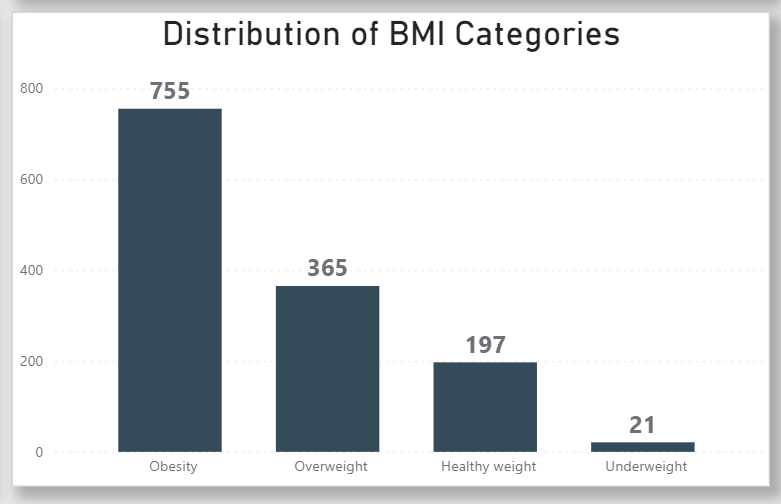
This chart shows the average insurance charges for each age group. The bars are divided into segments that represent the average charge for each age group.

1. **Stacked Bar Chart**:



This chart displays the number of customers segmented by the number of children they have. Each bar represents a different number of children.

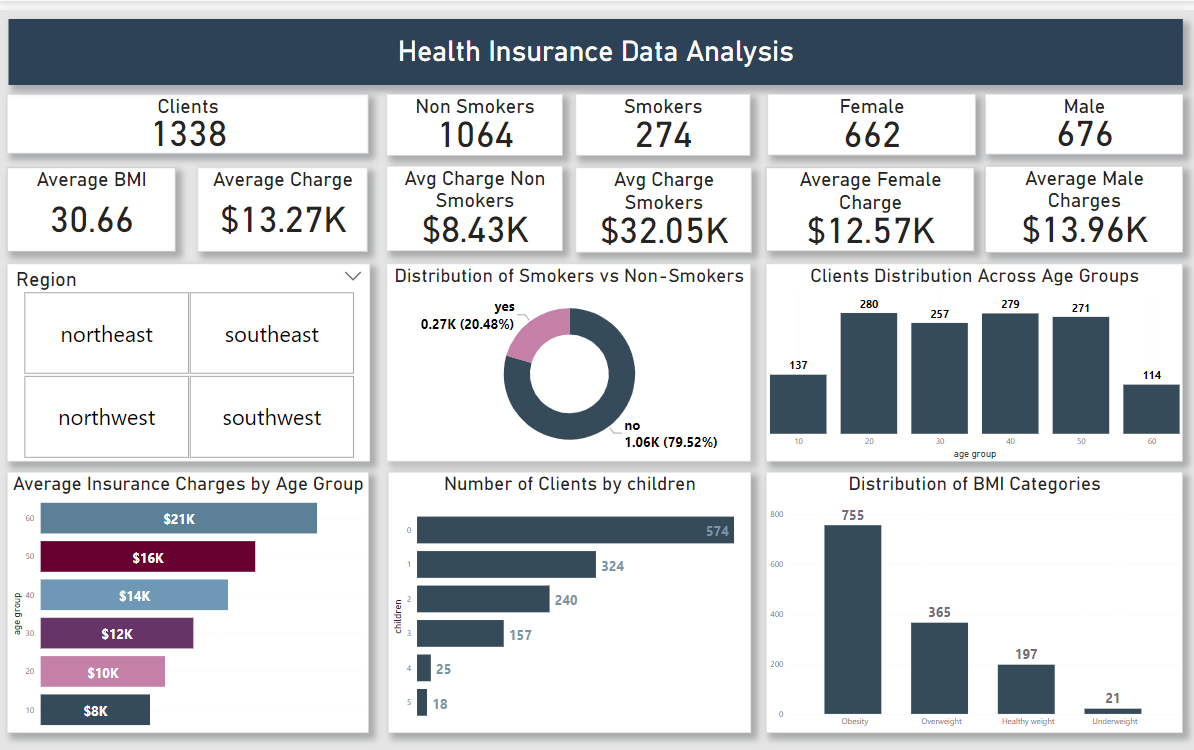
1. **Clustered Column Chart:**



This chart shows the distribution of customers across different BMI categories. It provides a clear view of the BMI status of the customers.

* + Each of these visualizations provides a unique perspective on the data, helping to uncover trends, patterns, and relationships that can inform decision-making and strategy.

**Dashboard:**



* **Insights:**

**Age and Charges:**

* **Trend**: The data suggests a positive correlation between age and insurance charges. As the age of the primary beneficiary increases from 18 to 64, the average insurance charge also increases from $8k to $21k.
* **Insight**: This suggests that as individuals age, they may be more prone to health issues, leading to higher healthcare costs due to the increased need for medical care.

**BMI and Charges:**

* **Trend**: The data shows a trend where beneficiaries with a higher BMI have higher average insurance charges. Beneficiaries with a BMI of 30.0 or higher have an average insurance charge of $15k, those with a BMI ranging from 25.0 to 29.9 have an average charge of $11k, those with a BMI of 18.5 to 24.9 have an average charge of $10k, and those with a BMI below 18.5 have an average charge of $8k.
* **Insight**: This could be because a higher BMI is often associated with a greater risk of chronic diseases, leading to higher healthcare costs.

**Smoking Impact:**

* **Trend**: The significant difference in average insurance charges between smokers and non-smokers highlights the financial impact of smoking on health insurance costs. On average, smokers have insurance charges of $32k, while non-smokers have charges of $8k.
* **Insight**: This could be due to the higher health risks associated with smoking, leading to more frequent and costly medical treatments.

**Regional Analysis:**

* **Trend**: The average insurance charge varies by region. The Southeast region faces the highest average insurance charge of $14k. This is followed by the Northeast with an average charge of $13k, and both the Northwest and Southwest regions have an average charge of $12k.
* **Insight**: The Southeast region not only has the highest average insurance charge of $14k but also a higher average BMI. This could potentially explain the higher charges in this region, as a higher BMI is often associated with a greater risk of chronic diseases, leading to higher healthcare costs.

**Number of Children and Charges:**

* **Trend:** On average, insurance charges are $11.9k for beneficiaries with 0 children, $10.4k for those with 1 child, $17.4k for those with 2 children, $10.4k for those with 3 children, $14.9k for those with 4 children, and $8.4k for those with 5 children.
* **Insight**: This suggests that the number of children a beneficiary has can influence the cost of insurance, however, the trend does not continue for those with more than 2 children, suggesting other factors may be influencing the charges for these groups.

**Gender and Charges:**

* **Trend**: On average, male beneficiaries have slightly higher insurance charges of $13.4k than female beneficiaries who have charges of $11.2k.
* **Insight**: This could be due to a variety of factors, including differences in health behaviors, use of medical services, or risk factors between genders. Specifically, more males are smokers and have a higher BMI over 30.0, which could contribute to the higher charges

**Outlier:**

* The higher insurance charges for beneficiaries with 2 children could be considered an outlier. This doesn’t follow the expected trend of increasing charges with the number of children. These beneficiaries with 2 children might have other factors contributing to the higher charges such as smoking more or living in regions with higher average charges like the Southeast, which could contribute to the higher charges.

**Pattern:**

* **Observed Pattern:** One pattern that emerges from the data is that certain risk factors, such as higher BMI and smoking, consistently lead to higher insurance charges across different categories of beneficiaries. Another pattern is that certain regions like the Southeast have higher average charges, potentially due to regional health factors or healthcare costs.
* **Recommendations:**

Based on the insights derived from the dataset, here are some recommendations for the health insurance company:

1. **Health Promotion Programs**: Given the high number of smokers and individuals with a high BMI, the company could invest in health promotion and disease prevention programs. These could include smoking cessation programs and initiatives to promote healthy eating and physical activity. This could potentially reduce future medical costs.
2. **Risk-Based Pricing**: The data shows that smokers and individuals with a high BMI tend to have higher medical costs. The company could consider these factors when determining insurance premiums to ensure that prices reflect the risk.
3. **Family Plans**: The data shows that individuals with children tend to have higher medical costs. The company could consider offering family plans or discounts for multiple dependents to attract more customers.
4. **Regional Strategies**: The data shows regional variations in smoking rates and BMI. The company could develop regional strategies to address these differences, such as targeted marketing campaigns or partnerships with local healthcare providers.
5. **Age-Based Services**: With the insight that older individuals tend to have higher medical costs, the company could consider offering additional services tailored to older adults, such as preventative care, wellness programs, or partnerships with providers specializing in geriatric care.

* **Conclusion:**

This project provided valuable insights into the factors influencing health insurance charges. Through the analysis of a dataset of 1338 customers, we were able to identify key trends and patterns related to age, gender, BMI, smoking status, region, and number of children. The use of a Power BI dashboard facilitated a comprehensive and interactive exploration of the data.

The insights derived from this analysis can guide health insurance companies in making informed decisions, such as tailoring insurance premiums based on risk factors, investing in health promotion programs, and developing regional strategies.

Overall, this project demonstrated the power of data analysis and visualization in uncovering hidden relationships and telling meaningful stories with data. It highlighted the potential of using such insights to drive strategic decision-making in the health insurance industry.

**Analysis Tools:**

* 1. **Microsoft Excel:** exploring the data and conducting basic analysis.
  2. **Microsoft Power Query:** Data Transformation.
  3. **Microsoft Power BI:** Data Visualization and dashboard.
  4. **Microsoft Word:** Documentation and Reporting.
  5. **Microsoft Teams:** Collaboration and Communication for Project Progress.