# Report on Insurance Dataset Analysis

Renzo Nicolas Daziano

## Author Information

Renzo Nicolas Daziano
Software Design at IT University of Copenhagen, Denmark

## Introduction

This report presents the analysis of an insurance dataset to identify key insights about the factors influencing medical insurance charges. The analysis focuses on the average age of patients, regional distributions, cost differences between smokers and non-smokers, and the impact of having children on the average age of patients.

## Data Source

The source of the data is Kaggle Insurance Dataset.

## Course Information

This project is part of the career paths "Data Engineering" and "Business Intelligence Analyst" at Codecademy.

## Data Description

The dataset consists of several columns: age, sex, BMI, number of children, smoking status, region, and insurance charges. It contains records for numerous patients, providing a rich source of data for analysis.

## Exploratory Data Analysis (EDA)

- **Age Distribution**: Histogram showing the age distribution of the patients.
- **BMI Distribution**: Boxplot depicting the distribution of BMI values.

- **Charges Distribution**: Histogram of the insurance charges.
- **Region Distribution**: Bar chart showing the count of patients from each region.

## Correlation Analysis

A heatmap showing the correlation between features such as age, BMI, number of children, and insurance charges can help identify strong relationships between variables.

## Analysis Results

### 1. Average Age of Patients

The average age of patients in the dataset is **39.21** years. This indicates that the dataset primarily consists of middle-aged individuals.

### 2. Regional Distribution

The majority of the patients are from the **southeast** region. This could imply a regional bias in the dataset or a higher concentration of insured individuals in this area.

### 3. Cost Analysis

**Average Cost for Non-Smokers**

The average medical insurance cost for non-smokers is **$8,434.27**. This serves as a baseline for understanding the cost impact of smoking on insurance charges.

**Average Cost for Smokers**

The average medical insurance cost for smokers is significantly higher at **$32,050.23**. This stark difference highlights the substantial financial burden of smoking on healthcare costs.

**Cost Difference between Smokers and Non-Smokers**

The cost difference between smokers and non-smokers is **$23,615.96**. This substantial difference underscores the impact of smoking on medical expenses, reinforcing the importance of smoking cessation programs and policies.

### 4. Average Age with Children

The average age of patients with at least one child is **39.78** years. This is slightly higher than the overall average age, suggesting that individuals with children tend to be marginally older.

## Predicted Influential Features

Based on the analysis, the following features are predicted to be the most influential for an individual's medical insurance charges:

1. **Smoking Status**: Significantly impacts insurance costs, with smokers incurring much higher charges.
2. **Age**: Older individuals generally have higher medical costs.
3. **BMI**: Higher BMI values are often associated with higher medical costs.
4. **Number of Children**: More children might indicate a higher cost due to additional family coverage.
5. **Region**: Certain regions might have different healthcare costs, as indicated by the majority representation from the southeast.

## Potential Biases in the Data

While conducting the analysis, several potential biases were identified that could impact the results and their generalizability:

- **Sample Bias**: The dataset may not represent the entire population accurately. Overrepresentation of certain age groups, regions, or socioeconomic statuses can skew the findings.
- **Measurement Bias**: Inconsistent or inaccurate data collection methods can introduce biases. For example, self-reported BMI might be less accurate, affecting the analysis of BMI-related costs.
- **Socioeconomic Factors**: Insurance costs can be influenced by factors not included in the dataset, such as income or education level, potentially skewing the results.
- **Access to Healthcare**: Differences in access to healthcare services across regions can affect medical costs and insurance charges, which may not be fully captured in the dataset.

## Discussion

The findings highlight significant differences in insurance costs based on smoking status and other factors. The high-cost difference between smokers and non-smokers suggests that smoking cessation programs could lead to substantial savings in healthcare costs.

## Recommendations

- Implement targeted smoking cessation programs to reduce healthcare costs.
- Consider regional healthcare policies to address disparities in costs.
- Include socioeconomic factors in future analyses to improve the accuracy of insurance cost predictions.

## Limitations

The analysis is limited by the potential biases in the dataset, such as sample and measurement biases. Additionally, the dataset does not include all possible factors influencing insurance costs, such as income and education level.

## Conclusion

The analysis of the insurance dataset provides valuable insights into the factors influencing medical insurance charges. Smoking status, age, BMI, number of children, and region are identified as the most influential features. However, potential biases in the data must be considered to ensure the accuracy and fairness of the findings. Addressing these biases in future analyses can lead to more reliable and equitable insights.

By understanding these key factors, policymakers and insurers can better design targeted interventions and policies to manage healthcare costs and improve public health outcomes.