Analyzing the Impact of Smoking on Insurance Costs

Author:Yahia Taraf

Abstract

This project explores the relationship between personal attributes and health insurance costs by leveraging a publicly available dataset of 1,339 records. The dataset includes key variables such as age, body mass index (BMI), smoking habits, number of dependents, and geographic region, all of which are analyzed to understand their influence on individual medical charges. Our primary goal is to identify significant trends and actionable insights that can inform both insurers and policyholders.

Through a combination of **Exploratory Data Analysis (EDA)**, regression modeling, and data visualization techniques, we examine how each factor contributes to variations in health insurance premiums. For example, attributes such as smoking status and BMI are expected to show strong correlations with higher costs, while other factors like age and family size may reveal more nuanced relationships. By calculating correlations, building predictive models, and visualizing results, we aim to quantify the individual and combined effects of these variables on insurance charges.

The results discovered highlighting the key drivers of costs and promoting awareness about behaviors or conditions that may lead to higher expenses. Ultimately, this project demonstrates the value of data-driven decision-making in creating more equitable and informed healthcare solutions.

Introduction

Health insurance companies rely on accurate risk assessment to determine premium rates. To develop the best medical insurance products, the insurer needs access to historical data to approximate the medical costs of each user. With this data, a medical insurer can develop more accurate pricing models, plan a particular insurance outcome, or manage a big portfolio. Smoking, age, and BMI are the most notable risk factors affecting these rates. This project seeks to quantify the financial impact of various attributes on insurance premiums, highlighting trends and outliers. The project provides insights to optimize pricing strategies by leveraging statistical models and data visualization.

Data Collection and Description

Medical Cost Personal Dataset Overview

The **Medical Cost Personal Dataset** from Kaggle contains 1,339 records of medical insurance data. The dataset includes various variables of different data types, which are used to calculate and predict individual insurance costs. Below is a detailed breakdown of the dataset variables:

Numerical Variables

These variables contain either discrete or continuous values:

- **age**:
  The age of the primary beneficiary.
  Type: Continuous (integer, represents a range of values).
- **bmi**:
  Body Mass Index (BMI), provides an understanding of body weight relative to height. BMI is calculated as weight (in kg) divided by height (in meters).
- *Type*: Continuous.
  Range: Ideally between 18.5 and 24.9.
- **children**:
  The number of children or dependents covered by health insurance.
  Type: Discrete (integer).
- **charges**:
  The individual medical costs are billed by health insurance.
  Type: Continuous.
  Note: This is the target variable used to predict insurance costs.

Categorical Variables

These variables represent distinct categories:

- **sex**:
  The gender of the insurance contractor.
  Categories: Male, Female.
- **smoker**:
  Indicates whether the beneficiary is a smoker.
  Categories: Yes, No.
- **region**:
  The residential area of the beneficiary in the United States.
  Categories: Northeast, Southeast, Southwest, Northwest.

This dataset provides a mix of numerical and categorical variables, allowing for comprehensive analysis and prediction of medical insurance costs.

Methodology 1

Before conducting any analysis, it is crucial to first understand the data and its structure. This is where **Exploratory Data Analysis (EDA)** plays a vital role. To gain insights into the data and identify potential issues, we utilized the **Diagnose** function from the dlookr package. This function provided key information for all variables in the dataset, including the count and percentage of missing values, as well as the count and rate of unique values. With this information, we successfully transitioned to the next step of preparing and manipulating the data.

During this process, we examined the data types of each variable and identified an issue: the sex and smoker columns contained character-based values, which could lead to errors in our analysis. To resolve this, we used the mutate function to create new columns with binary values, such as is_male, is_female, and is_smoker. These replaced the original columns, allowing for more effective numerical analysis.

Additionally, the Diagnose function revealed that the region column had four unique values, representing regions of the United States. To further analyze patterns, we split the dataset into four subsets based on the region column, ensuring the original dataset remained intact and no data was lost. These subsets would allow for better analysis of the variables and allow for region-based visualizations, like heatmaps, to be used in our analysis.

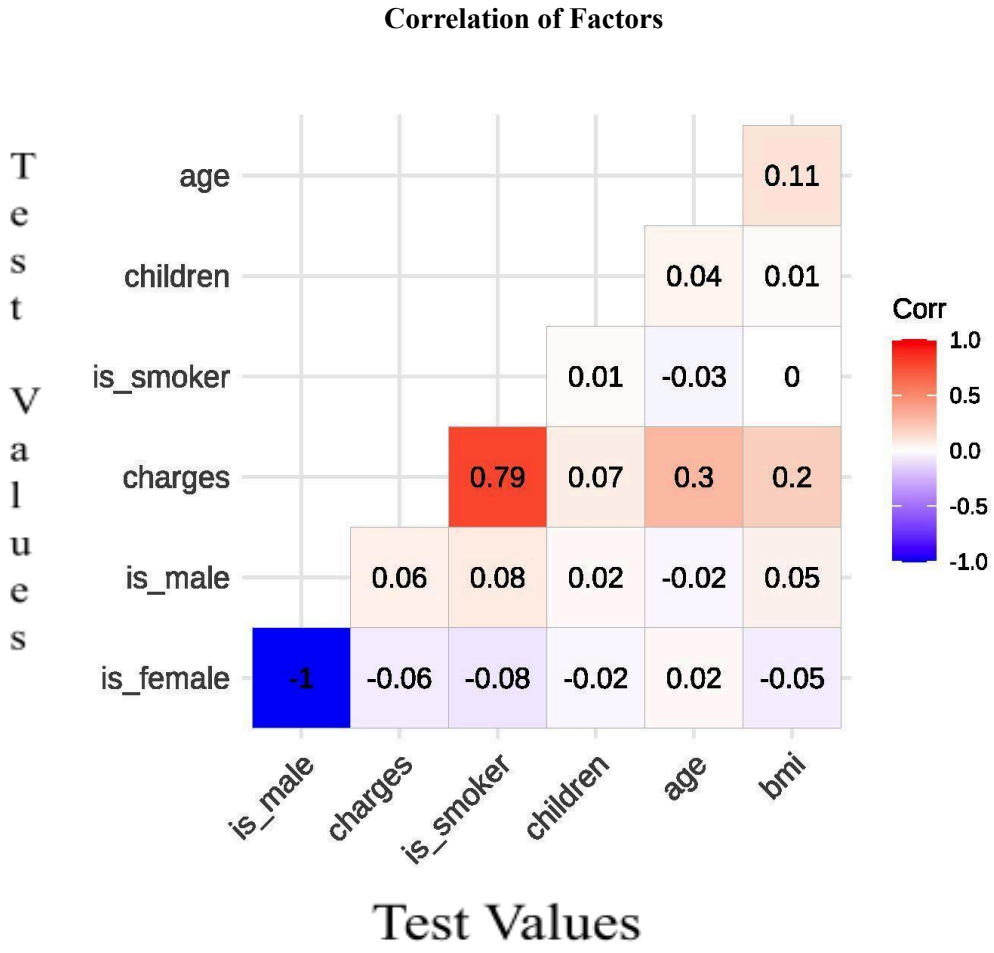
In the final stage of EDA, we aimed to identify which factors had the greatest impact on our target variable, **charges**. To achieve this, we calculated correlations between all numeric variables. Using the ggcorrplot library, we visualized these relationships in a correlation matrix. This matrix highlighted the key variables influencing charges and set the foundation for our subsequent analysis.

Methodology 2

In order to attempt to predict the charges, we've decided to use a multilinear regression model as a way to assess the influences of different variables. Our initial set of independent variables were age, BMI, and smoking status. We started with these to see how statistically significant our model would be in predicting the outcome with a few variables. Later we decided to add Gender as well as the amount of children in order to get a more accurate model.

Before we were able to start the linear regression model we had to check if our assumptions were correct. The three assumptions that we checked were linearity, independence, and normality. When analyzing normality we used a Q-Q residuals plot to see that the data was roughly normally distributed. In order to check independence we created a correlation matrix to test the different correlations not only with the outcome variable but also between the predictors to see if there were any multicollinearity. Finally checked linearity where we also saw through visualization of different scatter plots that there was a linear relationship between the variables.

Analysis and Results

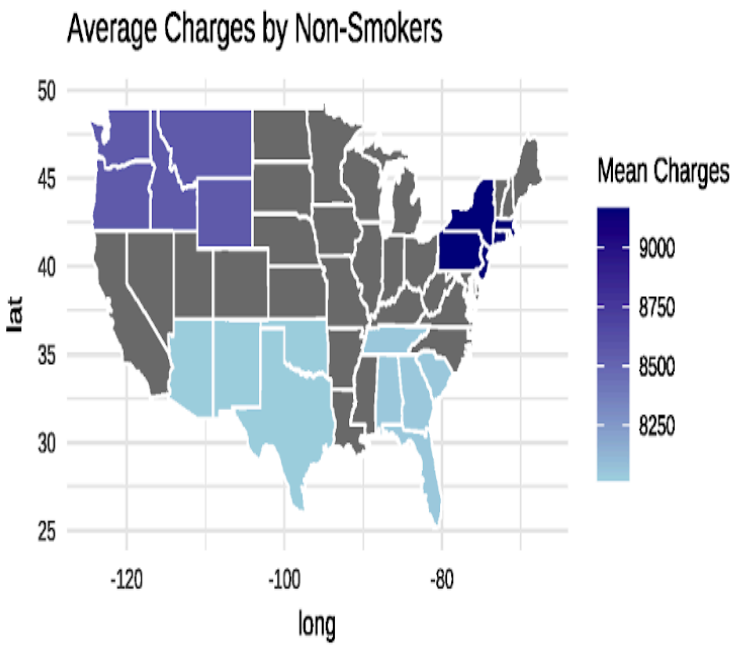


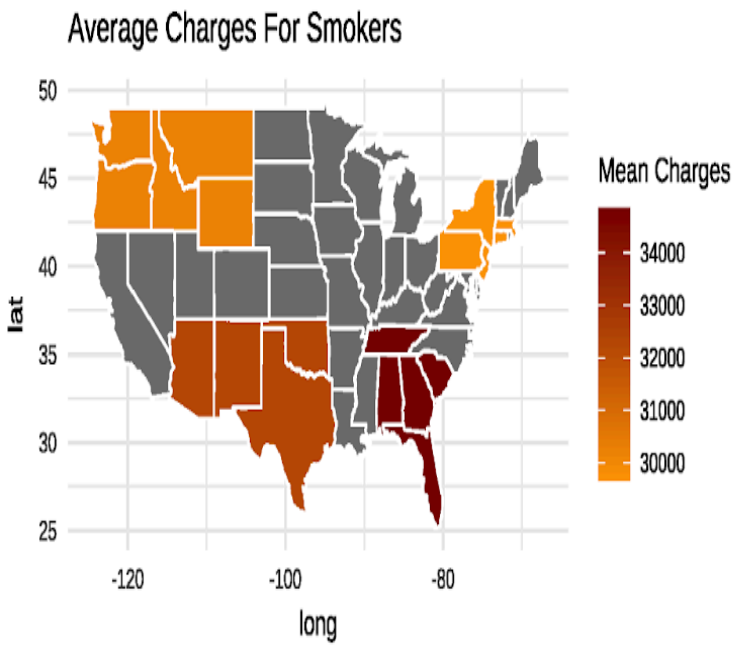
This matrix provides a clear visualization of the factors that have the greatest influence on the charges variable. Notably, the smoker status exhibits the strongest correlation with charges, with a coefficient of **0.79**, indicating a substantial positive relationship. This suggests that being a smoker is strongly associated with higher medical costs.

The second most influential factor is age, which shows a correlation of **0.30** with charges. While the relationship is weaker than that of smoker status, it still demonstrates that older individuals tend to incur higher medical expenses, likely due to increased health risks associated with aging.

These findings highlight that smoking behavior and age are the most significant predictors of medical charges in this dataset, with smoking status having a particularly pronounced impact. This underscores the importance of targeted health interventions and policies to address high-risk behaviors like smoking, which can dramatically increase healthcare costs.

Charges for Non-Smokers and Smokers by Region





The heat maps provide a clear depiction of regional variations in average medical charges for non-smokers and smokers in the United States.

The heat map on the left illustrates the average charges for non-smokers across 4 regions. The gradient, ranging from light blue to dark blue, represents mean charges, with darker shades indicating higher costs. Notably, the charges for non-smokers remain relatively consistent, ranging between $8,250 and $9,000. This suggests that medical expenses for non-smokers are fairly uniform across the country, with no significant regional disparities.

In contrast, the second heat map highlights the average charges for smokers. The gradient, spanning from light orange to dark red, reveals a much higher range of costs, approximately $30,000 to $34,000. Smokers face significantly higher healthcare expenses compared to non-smokers, with the highest charges concentrated in regions like the Southeast. This indicates a potential link between smoking prevalence, regional healthcare costs, and possibly local healthcare policies or economic factors.

Overall, these heat maps underscore the profound impact of smoking on medical costs. While non-smoker charges remain relatively stable, smokers incur dramatically higher expenses. This highlights the need for targeted health initiatives, particularly in regions with higher costs, to address smoking-related health risks and reduce the financial burden on the healthcare system.

Variable Regression

In our initial model, the x variables were age, BMI, and children. We found that this model had an adjusted R-squared value of 0.118, therefore it was not statistically significant. The x variables from this model do not predict the charges from here. After we faced this challenge, our team came up with a solution, which was to encode gender and smoker variables to binary for our observation.

After encoding the variables and adding them to our model is statistically significant because the p-value is very close to 0, also the F-statistics is 798 on 5 which is very strong. The R-squared is .7497 and the Adjusted R-squared is .7488. Both of these values are a very strong indication of One thing also to notice about the model the significance of the gender column, which is about 70%. Thus, this column does not significantly predict the outcome variable and could be caused by multicollinearity between the variables.

Conclusion

The study provides a comprehensive analysis of the financial impact of personal attributes on health insurance premiums, emphasizing the substantial role of smoking in determining premiums. Key findings underscore the strong correlation between smoking status and higher medical charges, while age also plays a notable, though lesser, role in influencing healthcare costs. The analysis highlights how behavioral factors like smoking and demographic characteristics such as age contribute significantly to healthcare expenses, necessitating targeted interventions for high-risk groups.

However, limitations in the study include potential biases in the data, such as the oversimplification of certain variables and the need for external validation of the findings. To strengthen the conclusions, further research should focus on exploring regional variations in healthcare costs, which may offer deeper insights into the economic impact of health behaviors across different geographical areas. Additionally, incorporating a broader range of demographic variables, such as income, education, and ethnicity, could provide a more comprehensive understanding of the factors influencing health insurance premiums. More data would also likely decrease our risk of overfitting and predicting the model too accurately. This would allow for more nuanced recommendations and better policy-making to address the disparities in healthcare costs.

References

1. **Choi,Mirian.*Insurance*.Kaggle,n.d., https://www.kaggle.com/datasets/mirichoi0218/insurance/data.**
2. **ggplot2**
   - Wickham, H., Chang, W., Henry, L., et al. (2023). *ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics.* Retrieved from https://ggplot2.tidyverse.org.
3. **ggcorrplot**
   - Kassambara, A. (2023). *ggcorrplot: Visualization of a Correlation Matrix Using ggplot2.* R package version 0.1.4. Retrieved from https://CRAN.R-project.org/package=ggcorrplot.
4. **dplyr**
   - Wickham, H., François, R., Henry, L., & Müller, K. (2023). *dplyr: A Grammar of Data Manipulation.* Retrieved from https://dplyr.tidyverse.org.
5. **dlookr**
   - Ryu, D. (2023). *dlookr: Tools for Data Diagnosis, Exploration, and Transformation.* Retrieved from https://CRAN.R-project.org/package=dlookr.
6. **tidyverse**
   - Wickham, H., Averick, M., Bryan, J., et al. (2019). *Welcome to the tidyverse.* Journal of Open Source Software, 4(43), 1686. Retrieved from https://doi.org/10.21105/joss.01686.
7. **maps**
   - Becker, R. A., Wilks, A. R., & Brownrigg, R. (2018). *maps: Draw Geographical Maps.* Retrieved from https://CRAN.R-project.org/package=maps.
8. **patchwork**
   - Pedersen, T. L. (2020). *patchwork: The Composer of Plots.* R package version 1.1.2. Retrieved from https://CRAN.R-project.org/package=patchwork.
9. **car**
   - Fox, J., Weisberg, S. (2019). *An R Companion to Applied Regression.* Sage Publications.
10. **randomForest**
    - Liaw, A., & Wiener, M. (2002). *Classification and Regression by randomForest.* R News, 2(3), 18-22.

Appendices