**PROJECT REPORT**

**Factors Influencing Medical Costs**

**Group 10**

INFO B -518 Professor - Dr. Zeyana Hamid

Neha Kowtharapu, Nilaya Ramireddy, Rupasri Pamulapati, Sapna Vibhaker, Srija Dammannagari

**INTRODUCTION**

Healthcare expenses continue to rise globally, prompting the need for a deeper understanding of the factors that drive these costs. Our project, "Factors Influencing Medical Costs," aims to shed light on the key elements that contribute to individual medical expenses using the Medical Cost Personal Dataset. This dataset provides a rich compilation of health-related information including demographics, lifestyles, and medical histories, which are instrumental in analyzing healthcare spending patterns.

Our research will focus on examining the impact of age, Body Mass Index (BMI), and the number of children on medical costs. Through meticulous data analysis using R for data cleaning, visualization, and statistical testing, we aim to uncover significant predictors of healthcare expenditures. This study not only contributes to academic knowledge but also provides practical insights that can guide policymakers, insurers, and individuals in making informed decisions about healthcare management and cost containment. The goal is to foster a more sustainable healthcare system where strategic interventions are based on robust empirical evidence.

By the end of this project, we anticipate presenting findings that will highlight critical drivers of medical costs, ultimately aiding in the development of targeted strategies to manage healthcare expenses efficiently.

**DATASET**

**Link:** <https://www.kaggle.com/datasets/mirichoi0218/insurance>

**Dataset overview**

The dataset includes variables like age, gender, BMI, smoking status, and region, along with corresponding medical costs. We have chosen this data set as the medical charges are one of the most important aspects in one’s life. In this dataset we have chosen the variables such as age, smoking status and the other variables which are used for interpreting the medical costs (insurance charges)

**Variables**

The dataset includes both categorical (Sex, Smoker, Region) and numerical (Age, BMI, Children, Charges) variables. Categorical variables provide demographic information such as gender, smoking habits, and geographical region, while numerical variables offer insights into age, body composition (BMI), family size (Children), and insurance charges**.**

**Data cleaning and pre-processing**

We began by importing the "insurance.csv" dataset into a data frame and conducted initial data cleaning checks. We confirmed that there are no missing values (NA) across all columns, which simplifies the preprocessing stage.

As part of our data exploration, we delved into the unique values present in categorical variables such as sex, smoker, and region. This exploration revealed a dataset encompassing both genders, smoking statuses, and diverse geographical regions from across the United States, all of which are ready for further analysis.

**Summary**

The dataset consists of 1338 observations across 7 variables, detailing factors influencing insurance charges. It includes a diverse age range from 18 to 64 years, with an average age of 39.21. The "sex" variable categorizes individuals as "female" or "male." BMI values stretch from 15.96 to 53.13, averaging 30.66, indicating varied body compositions. The "children" variable shows a range from 0 to 5, with an average of 1.095 children per individual. Smoking status is noted as "yes" or "no," and geographical diversity is represented by four regions: "southwest," "southeast," "northwest," and "northeast." Charges vary significantly, ranging from 1122 to 63770, with an average of approximately 13270. This overview sets the stage for deeper analysis into how these factors impact insurance charges.

**EXPLORATORY DATA ANALYSIS**

Our analysis commenced with univariate exploration, utilizing scatter plots and box plots to visualize the distribution of insurance charges across different age groups, revealing a clear proportional relationship between age and charges. Similarly, scatter plots were employed to investigate the impact of BMI on charges, showcasing stagnant charges for BMI 0-30 and a positive trend for BMI >30. Notably, smokers were found to have significantly higher insurance costs compared to non-smokers, as evidenced by box plots illustrating the disparity in charges between the two groups. Following this, multivariate analysis techniques, including scatter plots and correlation analysis, were leveraged to delve deeper into the relationships between critical variables such as age, BMI, gender, smoking status, region, and number of children. These analyses elucidated their combined effects on insurance charges, providing comprehensive insights into the dataset's characteristics and the multifaceted factors influencing insurance costs.

**Data Visualization**

We utilised scatter plots, heat maps, and box plots to depict and analyse the connections among variables.

**Rationale**

Our project addresses the critical need to understand healthcare charge determinants by analyzing variables such as age, sex, BMI, smoking status, and region. Through multiple linear regression, we aim to quantify the impact of these factors on healthcare charges, providing actionable insights for cost management and resource allocation in the healthcare sector. This approach allows us to uncover nuanced relationships and inform decision-making processes to enhance cost-effectiveness and efficiency in healthcare delivery.

**Appropriateness**

Our project's appropriateness lies in its comprehensive approach to analyzing healthcare charge determinants. Through exploratory data analysis (EDA), we gain initial insights into variables like age, sex, BMI, smoking status, and region, setting the stage for further analysis. Statistical tests such as the Shapiro-Wilk test ensure data validity, while methods like ANOVA and Tukey's HSD test allow us to explore group differences and regional impacts on healthcare charges. Utilizing multiple linear regression enables a holistic assessment of these factors, informing cost management and resource allocation strategies in healthcare.

**STATISTICAL METHODS**

**Statistical Tests**

Shapiro-Wilk Normality Test

We first assessed the distribution of charges using the Shapiro-Wilk test for four groups: male smokers, female smokers, male non-smokers, and female non-smokers. The results indicated significant deviations from a normal distribution in all groups, with p-values well below the 0.05 threshold. Most of our data is normal except for smoking groups.

ANOVA

Following the normality checks, we conducted an ANOVA to investigate the differences in charges among smokers across different geographic regions. The test revealed a significant difference (p-value = 0.02), indicating that location significantly impacts charges for smokers.

Tukey’s HSD

Using Tukey’s Honest Significant Difference test, we pinpointed that the significant difference in charges among smokers was primarily between the Southeast and Northeast regions, with a p-value of 0.02. No significant differences were found between the other groups.

T-Tests

Another t-test comparing charges between males and females also indicated a significant difference (p-value = 0.02934), with males incurring higher average charges.

**Multiple Linear Regression**

The statistical analysis utilized multiple linear regression to assess how age, smoking status, BMI, and the number of children affect healthcare charges. All variables demonstrated strong statistical significance with p-values below 0.05. Specifically, the results show that each additional year of age increases healthcare charges by approximately $263.55, smoking status raises charges by about $23,467.36, each unit increase in BMI leads to a $343.18 rise in charges, and each additional child adds $425.57 to the costs.

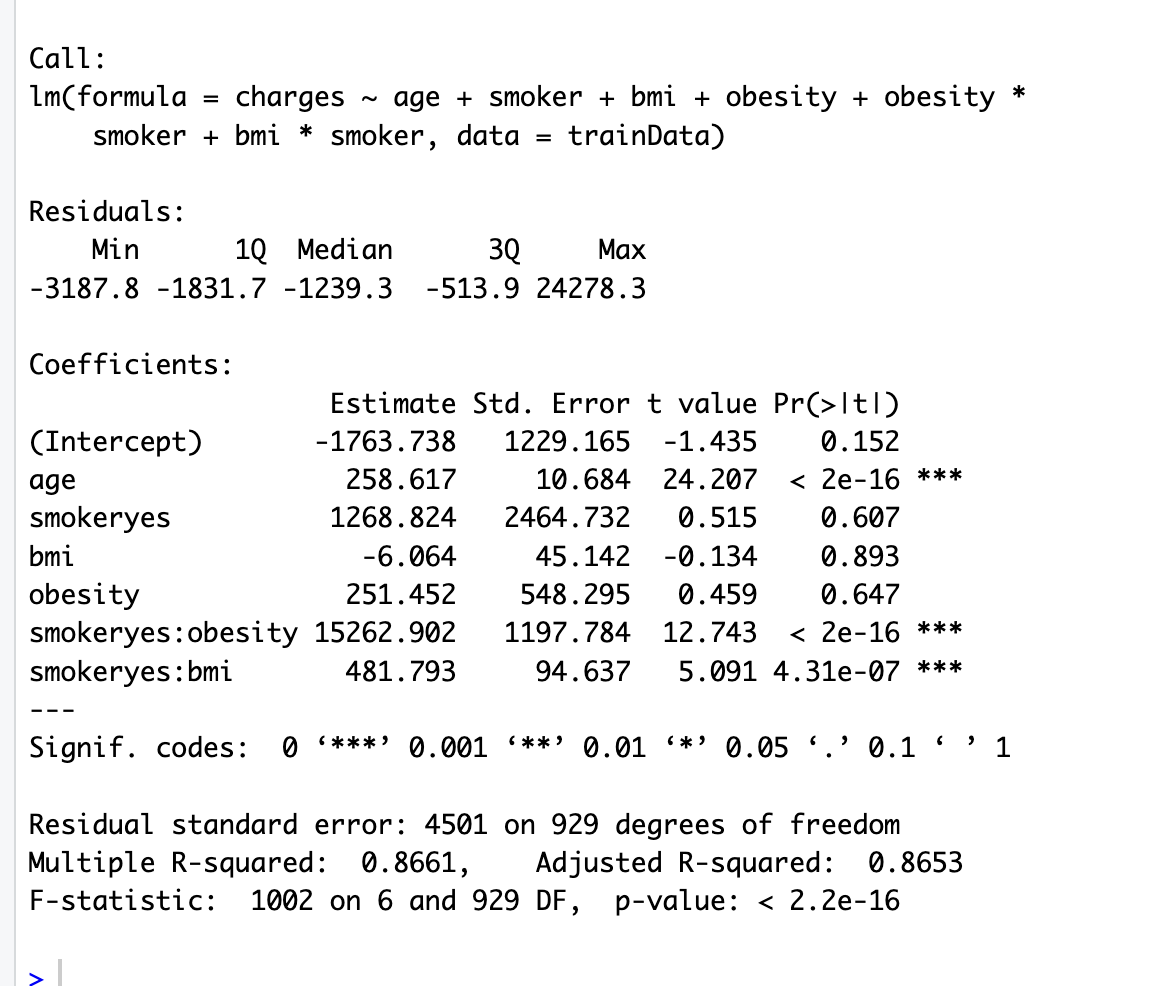
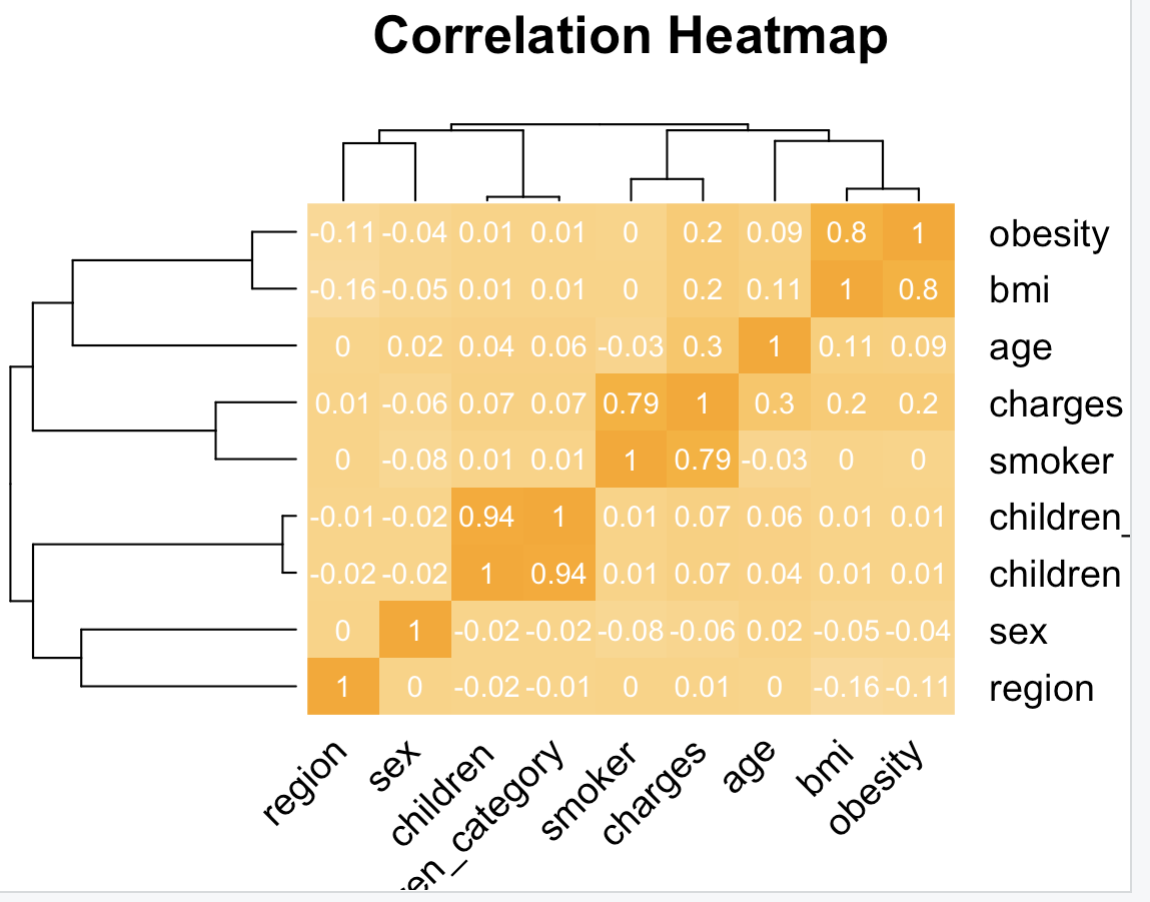
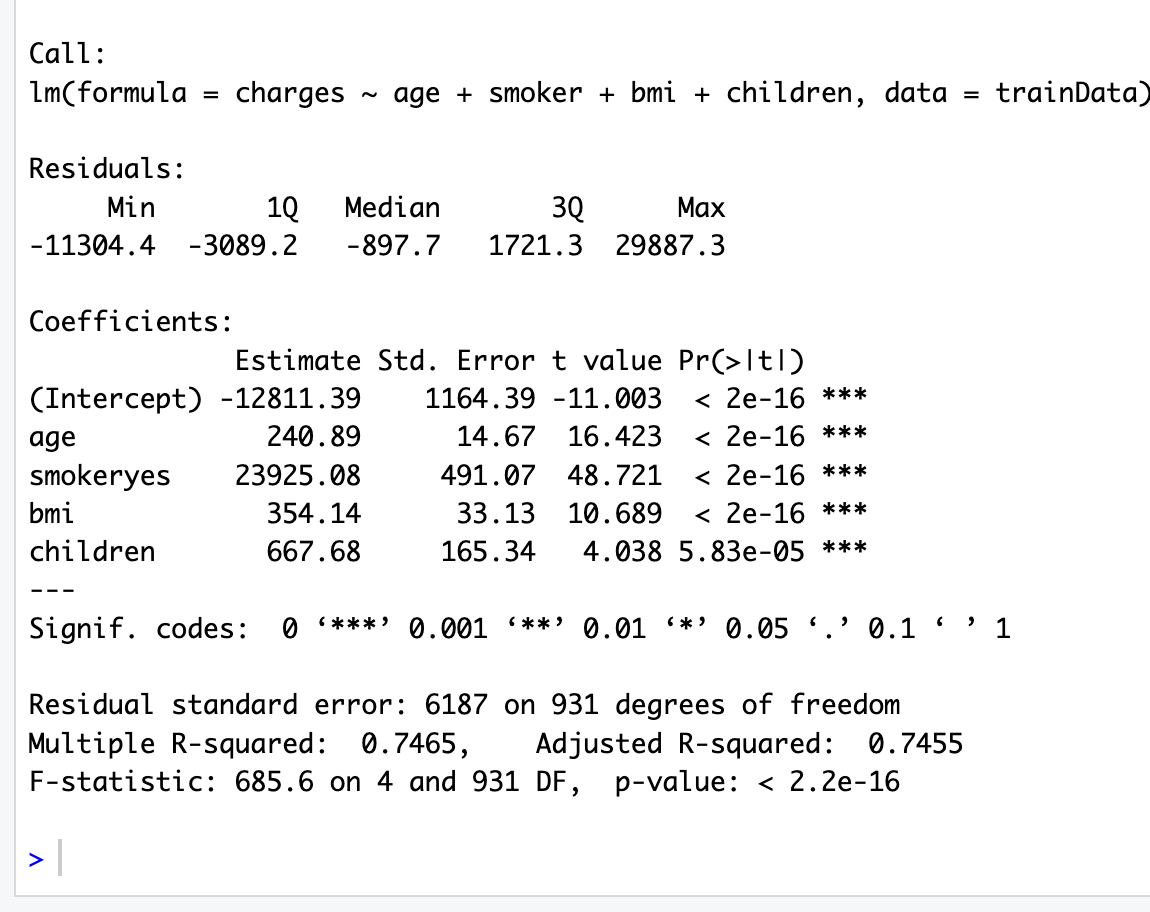
This multiple linear regression analysis provides valuable insights into factors influencing healthcare charges. The results underscore the high cost associated with smoking and the progressive increase in charges with age and BMI. Such information could be instrumental for healthcare providers and policymakers to devise strategies targeted at these significant factors to manage healthcare costs effectively. However, the wide range of residuals suggests that there could be other variables not included in the model that influence healthcare charges.

**Improved Multiple Linear Regression**

We improved our healthcare charges regression model by adding interaction terms between key variables like BMI and smoking status, and obesity and smoking status. This helped capture synergistic effects that individual variables couldn't fully explain. Our model's R-squared value increased significantly to 0.8678, indicating that around 86.78% of healthcare charge variability is now explained. Notably, the interaction term for smoker and obesity showed a substantial impact on costs. This nuanced approach reveals deeper insights into cost drivers, aiding targeted interventions and policy adjustments for cost management.

**FINDINGS**

The comprehensive analysis conducted on the dataset revealed significant relationships between various factors and healthcare charges. Initial tests, such as the Shapiro-Wilk Normality Test, highlighted normal distributions of charges across different groups. Subsequent ANOVA and Tukey’s HSD tests unveiled regional disparities in charges among smokers, particularly emphasizing differences between the Southeast and Northeast regions. T-tests emphasized significant differences in charges between smokers and non-smokers, as well as between males and females. Multiple linear regression provided deeper insights, showing that age, smoking status, BMI, and number of children significantly influence healthcare charges, with smokers and individuals with higher BMI facing notably higher costs. Further enhancing the regression model with interaction terms revealed the complex interplay between smoking and obesity on healthcare costs. Overall, the findings reject the null hypothesis, affirming a significant relationship between demographic and lifestyle factors and healthcare charges, underscoring the need for targeted interventions to manage healthcare costs effectively.



**LIMITATIONS:**

One limitation could be the potential presence of unobserved confounding variables not accounted for in the model, such as lifestyle factors or pre-existing health conditions, which could affect medical charges.

The appropriateness of the model relies on the assumption of linear relationships between the independent variables and the dependent variable, which might not hold true in all cases.

**CONCLUSION:**

After conducting exploratory data analysis and scrutinizing the data patterns, we have arrived at a pivotal conclusion: rejecting the null hypothesis. This indicates a significant relationship between at least one of the predictors—age, sex, BMI, number of children, smoking status, and region—and medical costs.

Our analysis reveals that smoking status exerts the most substantial influence on insurance charges, followed closely by obesity (as indicated by BMI) and age. Notably, individuals who both smoke and are obese (BMI > 30) tend to face notably higher insurance charges. This observation underscores the compounding effect of these two risk factors on healthcare expenses.

Moreover, we have employed interaction terms, such as smoker \* BMI, to enrich our model and capture the combined effects of smoking and BMI. This strategic approach allows us to more comprehensively understand and account for the intertwined influences of these variables on insurance charges, enhancing the accuracy and robustness of our analysis.

**References**

Dieleman, J. L., Squires, E., Bui, A. L., Campbell, M., Chapin, A., Hamavid, H., Horst, C., Li, Z., Matyasz, T., Reynolds, A., Sadat, N., Schneider, M., & Murray, C. J. L. (2017). Factors associated with increases in US health care spending, 1996-2013. JAMA, 318(17), 1668. https://doi.org/10.1001/jama.2017.15927

https://www.kaggle.com/datasets/mirichoi0218/insurance

**Appendix**