Final Project

Ibrahim Shkoukani

gv7097@wayne.edu

Wayne State University

INF 6490

Professor Smith

Dec 12, 2025

Introduction

The research examines the Medical Cost Personal dataset (insurance.csv) which includes 1,338 records that link personal characteristics to health information and medical insurance expenses amounting to USD annually. The study includes age, sex, BMI, number of children, smoker status, region, and charges as its variables. The dataset contains appropriate data which enables researchers to study healthcare cost determinants while developing charge prediction models.

The main goal of this research involves using demographic information and health data to both understand and forecast medical insurance expenses while focusing on how smoking and body mass index affect treatment expenses. It will also aim to answer the following research questions:

- Do smokers have significantly higher insurance charges than non-smokers?
- Is BMI associated with charges (and is that association statistically significant)?
- Do average charges differ by geographic region?
- How well can charges be predicted from the available predictors using a regression model with train–test validation?

Recommendations

The research team performed an assessment of missing data points and data inconsistencies within the collected dataset. The analysis did not contain any missing data, so researchers did not need to perform imputation or remove any rows from the dataset. The statistical analysis required the conversion of sex and smoker and region variables into factor types because these variables needed appropriate handling during statistical tests and regression modeling.

The exploratory analysis showed that medical insurance costs followed a right-skewed distribution because few people needed to pay very high amounts for their coverage. The model requires a log transformation of charges to solve the skewness problem which will enhance its predictive power. The model transformation enables the regression model to detect proportional cost variations more effectively while minimizing the impact of outliers that maintain all available data points.

Descriptive Statistics

Continuous Variables

The analysis used descriptive statistics to calculate central tendency values and dispersion measures for all continuous data points. The dataset contained 39.21 years as the average age of participants (SD = 14.05) while the median age reached 39 years. The participants had an average BMI of 30.66 (SD = 6.10) which indicates that most people in this group have weights that fall into the overweight or obese categories.

Medical insurance charges exhibited substantial variability. The average yearly cost amounted to $13,270.42 but the middle value of $9,382.03 showed that the data followed a right-skewed pattern. The standard deviation of $12,110.01 shows that costs between people vary extensively. The substantial gap between the mean and median values confirms that a log transformation should be applied for modeling purposes.

The data showed that children's numbers followed a mean of 1.09 and a median of 1 while many people had no dependents since 0 was the most frequent value.

Categorical Variables

The research team created frequency tables together with proportions to analyze all categorical data points. The study showed that 79.5% of participants did not smoke while 20.5% of participants were smokers. The research participants consisted of 50.5% male and 49.5% female subjects who were distributed equally between sexes.

The study maintained a fair distribution of participants throughout different geographic areas, but the Southeast region contained the most participants at 27.2%. The distributions establish essential background information for subsequent hypothesis testing and regression analysis.

Data Visualization

The analysis used histograms and boxplots as univariate visualizations to study data distributions while searching for unusual values that might exist in the data. The medical charge histogram revealed a strong rightward bias in the data distribution which boxplots confirmed that smokers paid more than non-smokers for their medical expenses.

The research team employed multivariate visualizations through scatterplots and correlation heatmaps to study the connections between different variables. The scatterplots showed that BMI and medical costs had a positive relationship which became more evident when patients smoked. The visual patterns helped determine which statistical tests to use while confirming that smoking status and BMI should be used as primary predictor variables.

Statistical Analysis and Hypothesis Testing

Smoking Status and Medical Charges (t-test)

A Welch two-sample t-test was conducted to compare mean medical charges between smokers and non-smokers. The results indicated a statistically significant difference ($t = 32.75$, $p < 0.001$). Non-smokers had a mean charge of $8,434.27, while smokers had a mean charge of $32,050.23.

The estimated mean difference was $23,615.96, with a 95% confidence interval of [$22,197.21, $25,034.71]. This result provides strong evidence that smoking is associated with dramatically higher medical insurance costs.

Regional Differences in Charges (ANOVA)

The research used one-way analysis of variance (ANOVA) to determine if charges between different geographic areas showed any significant differences. The research data showed that medical costs vary significantly between different regions because region acts as a statistically significant factor ($F(3, 1334) = 2.97$, $p = 0.0309$).

The study shows that different regions have distinct patterns, but the F-value shows that these variations remain small when compared to smoking status and other factors. The research shows that healthcare expenses depend on personal health choices rather than the location where patients receive treatment.

BMI and Medical Charges (Correlation Analysis)

The research used Pearson and Spearman correlation tests to study how BMI affects medical treatment expenses. The statistical analysis using Pearson's correlation method showed a significant positive relationship between the variables (r = 0.198, 95% CI [0.146, 0.249], $p < 0.001$). The Spearman rank correlation analysis demonstrated a substantial relationship between the variables ($\rho = 0.119$, $p < 0.001$) because the charge data followed a non-normal distribution pattern.

The research indicates that people with higher BMI levels tend to spend more on medical care but the connection between BMI and healthcare expenses remains moderate.

Regression Model and Predictive Performance

The model used log-transformed medical charges as its outcome variable while age and BMI and number of children and sex and smoking status and region served as predictor variables.

The model explained most of the charge variability in the training data because it achieved an R^2 value of 0.7785 and an adjusted R^2 value of 0.7768. The research results showed that smoking status served as the most significant factor which predicted medical costs ($\beta = 1.565$, 95% CI [1.50, 1.63]) thus medical expenses for smokers were 4.5 to 5.1 times higher than those of non-smokers.

The statistical analysis revealed that Age together with BMI and number of children served as significant positive factors which increased medical expenses. The analysis showed that male defendants received approximately 10% lower penalties than female defendants after researchers applied all relevant variables. The regional impact on prices remained limited because the Southeast and Southwest maintained lower price levels compared to the Northeast.

Model Validation (Train-Test Split)

The model performance evaluation used an 80/20 train–test split with a fixed random seed for evaluation. The model achieved an R^2 value of 0.7249 when testing the data which meant it explained 72.5% of the log-transformed charge variance. The RMSE on the log scale was 0.4847 which translates to an approximate RMSE of $9,304.21 when the predictions are converted from log scale to dollar units.

The research results show excellent predictive accuracy because healthcare expenses naturally fluctuate in medical settings.

Conclusions

The research shows that smoking status creates the largest impact on medical insurance costs which exceeds all other factors including age and location and demographic characteristics. The combination of BMI and age factors leads to higher healthcare expenses which confirms previous research about how health risk factors result in elevated medical costs.

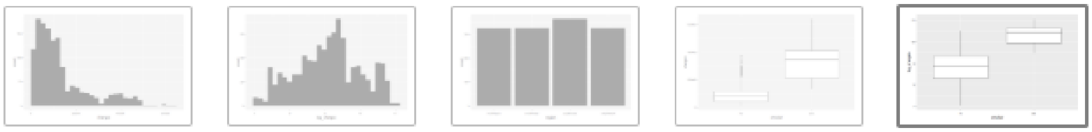
The research shows that regional payment variations exist, but these variations remain limited when compared to the impact of patient behavior and their medical conditions. The regression model delivers robust explanatory and predictive capabilities because it explains most of the medical insurance charge variability through its included variables.

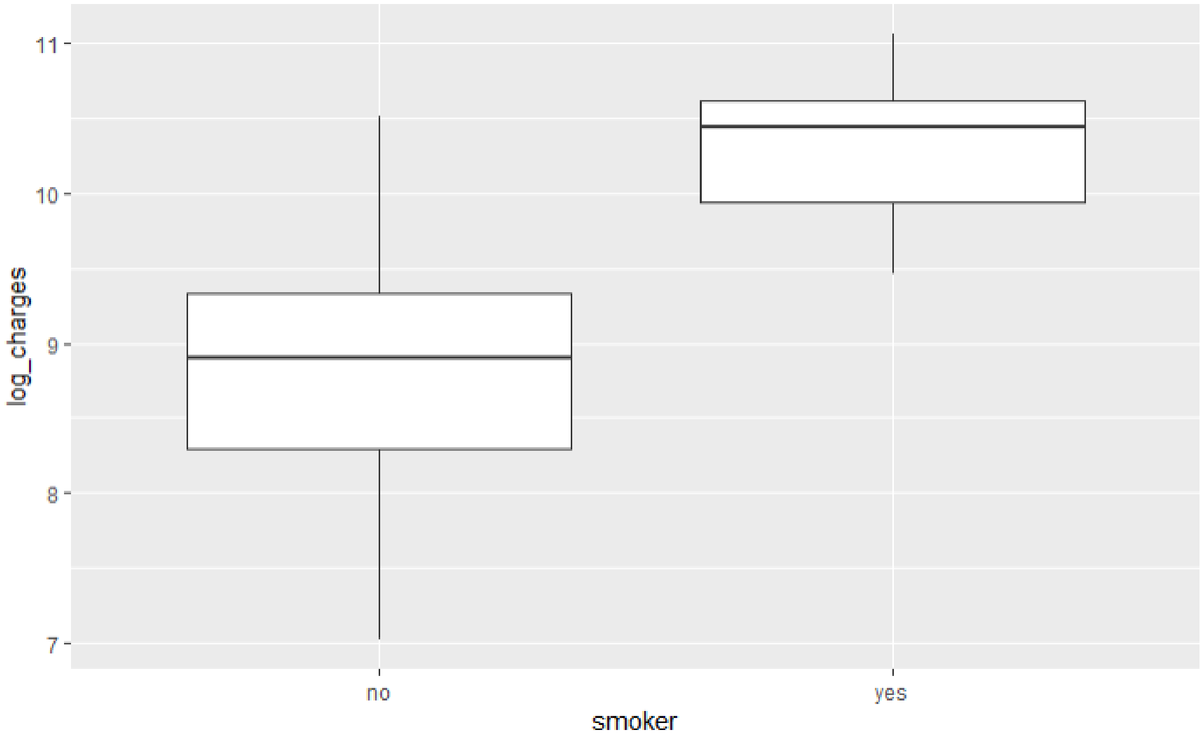
Limitations and Extensions

The research uses observational data which prevents researchers from establishing cause-and-effect relationships. The analysis lacks three essential elements which are income levels and pre-existing medical conditions and insurance plan specifics because these factors could affect thetotal expenses. The log transformation makes it difficult to interpret results in terms of direct dollar values.

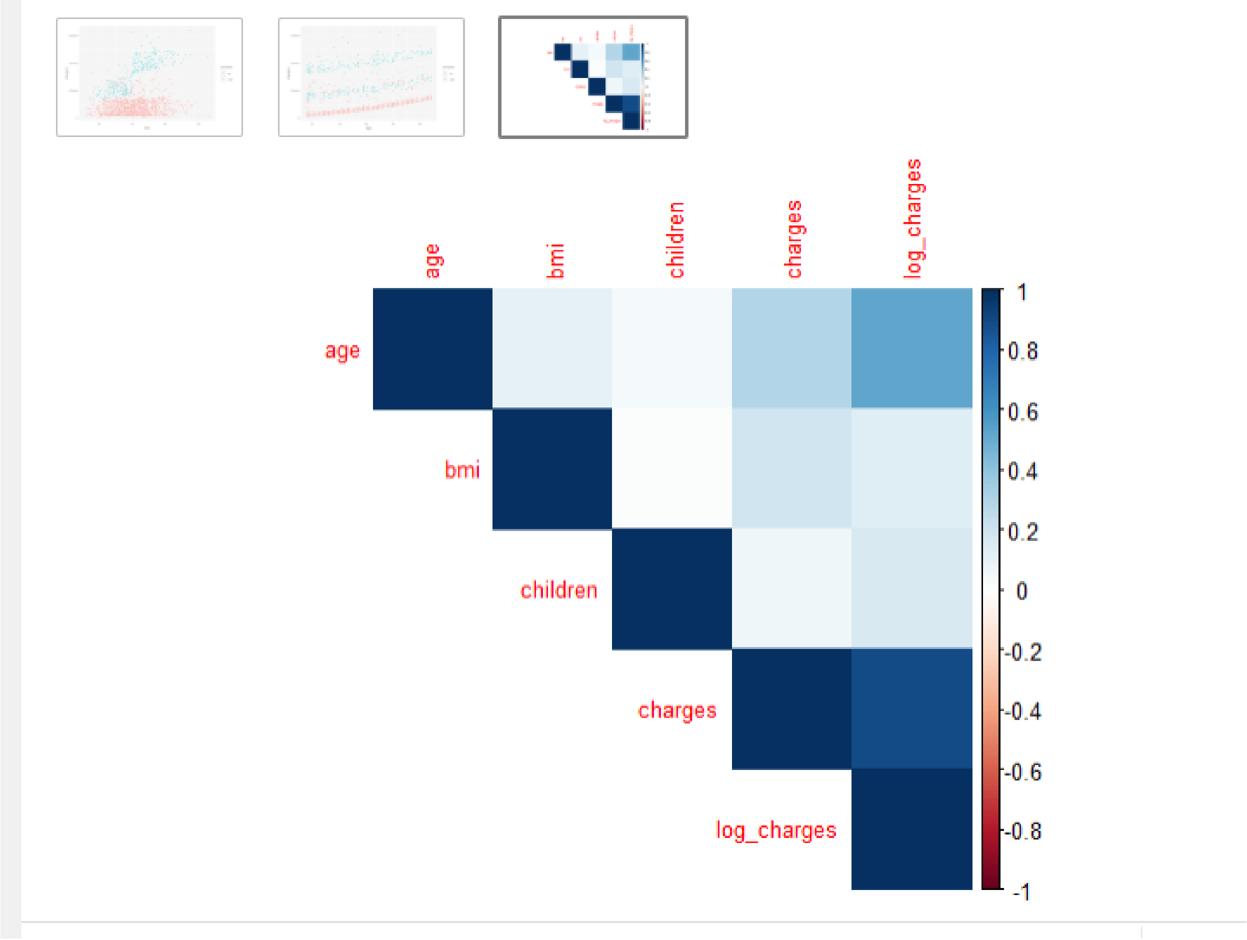
Research should investigate how BMI interacts with smoking status and other factors while adding new variables to the model and testing different predictive methods to enhance forecasting performance.

Appendix A





Appendix B


age
bmi
children
charges
log_charges
age
bmi
children
charges
log_charges
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1