Precision in Healthcare Costs: Modeling Individual Characteristics to Refine Insurance Charge Estimation

Arun Kumar Soora[1], Kalyan Raj Chinigi[1], Pallavi vandanapu[1], Srihari Myla Venkata[1], Yugala Ramula[1]

[1]Indiana University-Purdue University Indianapolis, Indianapolis, IN 46202, United States of America

Abstract:

This study investigates the impact of demographic and lifestyle factors, including age, BMI (Body Mass Index), smoking status, and geographical location, on health insurance charges. By analyzing a diverse range of variables, this research aims to elucidate the complex interplay between these factors and insurance premiums. Utilizing advanced statistical methods and machine learning algorithms, we develop a predictive model to discern patterns and relationships within the data. The findings shed light on the differential effects of demographic and lifestyle factors on health insurance charges, providing valuable insights for insurers, policymakers, and consumers alike. Understanding these influences can lead to more informed decision-making, tailored risk assessment, and potentially fairer pricing structures in the health insurance market. This research contributes to the ongoing discourse on healthcare affordability and accessibility, offering avenues for improving insurance pricing strategies and promoting public health initiatives (Carvalho et al., 2018) .

Keywords: Insurance, premiums, BMI, age, smoking

1. Project Scope

1.1 Introduction

The insurance industry operates on the principle of risk assessment, pivotal for setting policy premiums and allocating resources effectively. In recent years, the intricacies of insurance charge prediction have become more pronounced, driven by diverse and complex variables influencing individual policy costs. Traditional models, often linear in their assumptions, struggle to accommodate the nuanced interactions between these variables, such as age, health status, lifestyle choices, and the specific details of the coverage required. As a result, there is a growing demand within the industry for more sophisticated analytical techniques that can handle this complexity and deliver more accurate predictions.

This paper addresses the problem by deploying advanced statistical methods and machine learning algorithms to analyze and predict insurance charges. Our objective is to develop a predictive model that not only captures the subtle interdependencies between the multitude of variables affecting insurance premiums but also enhances the accuracy and granularity of these predictions. Through this approach, we aim to contribute a robust analytical tool to the industry, which can lead to more informed decision-making, optimized risk assessment, and ultimately, fairer pricing for consumers. This work is crucial as it not only supports economic efficiency but also promotes a more equitable allocation of insurance costs based on a detailed understanding of individual risk factors.

1.2. Objective (Research Problem)

This study aims to investigate the relationship between demographic factors (age and geographical location) and lifestyle factors (BMI and smoking status) on health insurance charges. By framing hypotheses, we seek to understand how these variables impact risk assessment and pricing strategies in the industry, ultimately providing insights for enhancing the accuracy and fairness of insurance premiums.

Null Hypothesis (H0):

There is no significant relationship between demographic and lifestyle factors (such as age, BMI, smoking status) and health insurance charges.

Alternate Hypothesis (H1):

There is a significant relationship between demographic and lifestyle factors (such as age, BMI, smoking status) and health insurance charges.

1.3. Purpose:

The dataset is often used for predictive modeling tasks, such as estimating insurance costs based on demographic and lifestyle factors, as well as for exploring relationships between variables, identifying trends, and understanding the impact of different factors on insurance charges.

2. Methodology

2.1. Steps of the Project

Our project is dedicated to analyzing the multitude of factors affecting insurance. We utilized R for statistical analyses and data visualization to deepen our understanding of these complex interactions. Our methodology comprises five key stages:

1. Data Collection: We sourced and collected the relevant dataset required for our research objectives.

2. Data Extraction and Cleaning: The acquired data was carefully extracted from its original source and prepared for future analysis by ensuring cleanliness and integrity.

3. Data Analysis: Rigorous statistical techniques were employed to unveil meaningful insights, identifying patterns and correlations between various factors influencing insurance.

4. Data Visualization: Utilizing R, we visualized the results of our statistical tests and regression models to effectively communicate our findings.

5. Feature Selection and Model Refinement: Feature selection techniques were applied in post-initial analyses to refine our models, focusing on the most significant predictors, and facilitating a deeper analysis.

3. Data Collection

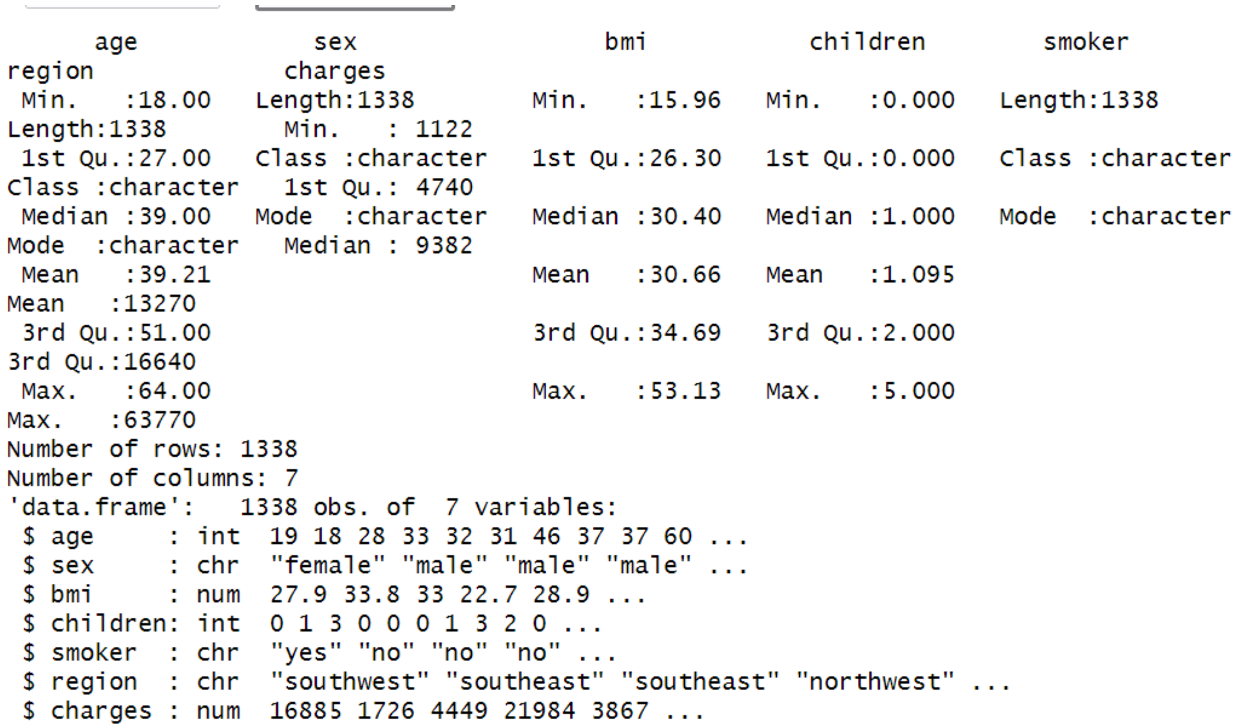
The dataset, comprising information for 1,338 individuals, was obtained from reputable health insurance databases. It includes variables such as age, geographical location, BMI, smoking status, and insurance charges. Each record represents an individual case, facilitating predictive modeling tasks and exploration of relationships between variables to understand their impact on insurance charges.

Link for dataset: https://www.kaggle.com/datasets/mirichoi0218/insurance

4. Data Extraction

4.1 Sampling and Summaries:

To ensure the reliability and generalizability of our analysis, the dataset employs a stratified sampling technique, focusing on achieving a balanced representation across different geographical regions. This stratification extends to maintaining a proportional distribution between smokers and non-smokers, thereby allowing for more accurate assessments of smoking's impact on insurance charges.



```
      age            sex                  bmi           children         smoker
region            charges
 Min.   :18.00   Length:1338        Min.   :15.96   Min.   :0.000   Length:1338
Length:1338        Min.   : 1122
 1st Qu.:27.00   Class :character   1st Qu.:26.30   1st Qu.:0.000   Class :character
Class :character   1st Qu.: 4740
 Median :39.00   Mode  :character   Median :30.40   Median :1.000   Mode  :character
Mode  :character   Median : 9382
 Mean   :39.21                      Mean   :30.66   Mean   :1.095
Mean   :13270
 3rd Qu.:51.00                      3rd Qu.:34.69   3rd Qu.:2.000
3rd Qu.:16640
 Max.   :64.00                      Max.   :53.13   Max.   :5.000
Max.   :63770
Number of rows: 1338
Number of columns: 7
'data.frame':   1338 obs. of  7 variables:
 $ age     : int  19 18 28 33 32 31 46 37 37 60 ...
 $ sex     : chr  "female" "male" "male" "male" ...
 $ bmi     : num  27.9 33.8 33 22.7 28.9 ...
 $ children: int  0 1 3 0 0 0 1 3 2 0 ...
 $ smoker  : chr  "yes" "no" "no" "no" ...
 $ region  : chr  "southwest" "southeast" "southeast" "northwest" ...
 $ charges : num  16885 1726 4449 21984 3867 ...
```

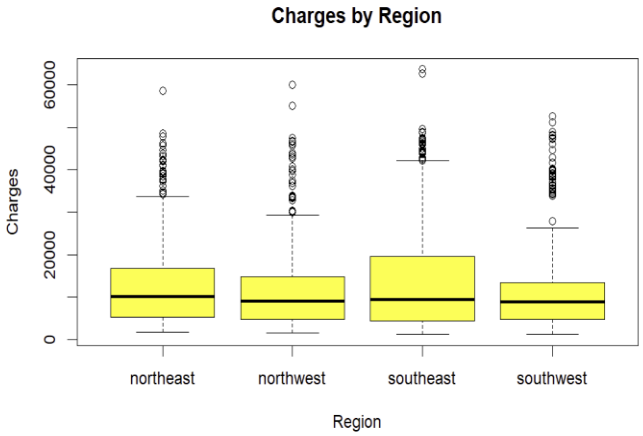
the average age of the individuals is 39.21 years, suggesting a middle-aged cohort predominantly. The mean BMI is recorded at 30.40, positioning the average participant at the threshold of obesity, as defined by the Centers for Disease Control and Prevention (CDC). These metrics are crucial as they significantly influence insurance premiums and are indicative of potential health risks.

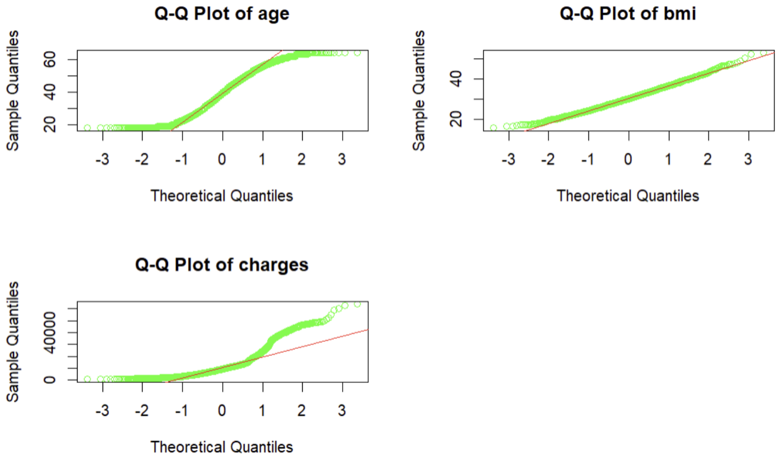
4.2 Data Preprocessing:

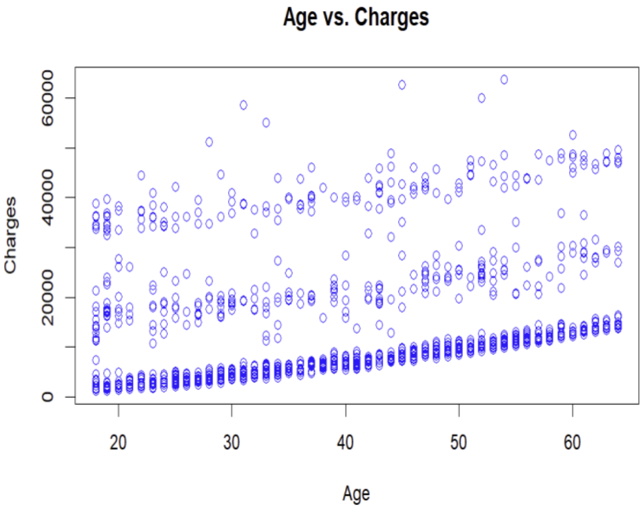
The dataset was rigorously prepared for analysis by transforming categorical variables like sex, smoker status, and region into numerical factors, ensuring compatibility with machine learning algorithms (James et al 2021).

5. Exploratory Data Analysis:

In our exploratory data analysis (EDA), we used visualizations like histograms, scatter plots, and boxplots to understand the distribution of insurance charges and explore relationships with variables like age, BMI, and smoker status. These visualizations helped identify outliers, compare charge distributions across categories like gender and region, and assess the normality of key variables. This EDA phase provided crucial insights into our dataset, guiding further analysis and model development for our project.





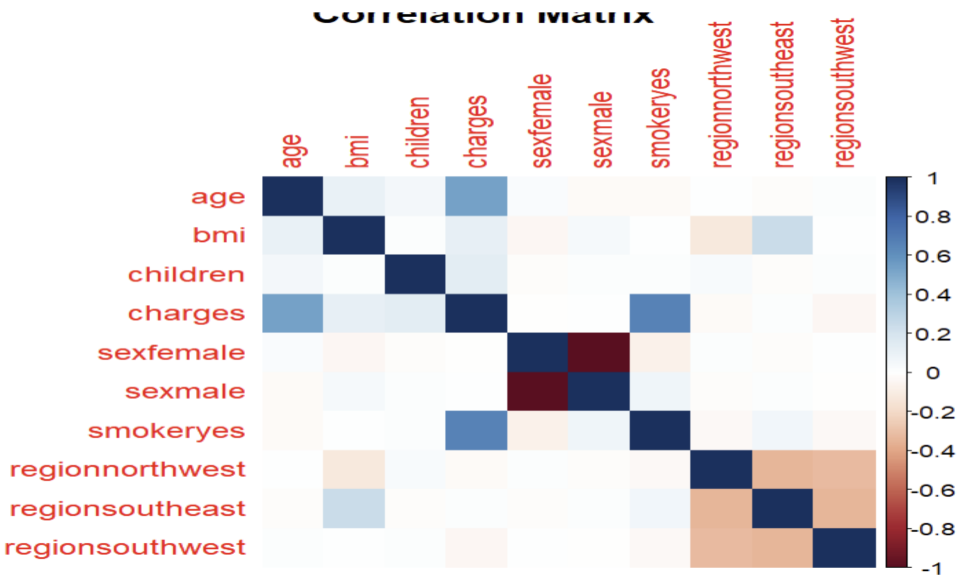


6. Statistical Methods:

6.1. The Spearman's Correlation Analysis of Demographic and Lifestyle Factors with Insurance Charges:

The analysis of the dataset revealed significant correlations among its variables. Age displayed a moderate positive correlation (0.53) with insurance charges, indicating that as individuals age, their insurance costs tend to rise. Similarly, BMI exhibited a weak positive correlation (0.12) with charges, suggesting a slight increase in charges with higher BMI values. The number of children also showed a weak positive correlation (0.13) with charges, indicating slightly higher costs for individuals with more children. However, the most notable correlation was observed with smoking status, displaying a strong positive correlation coefficient (0.66) with charges. This signifies a substantial impact of smoking on insurance costs, with smokers consistently facing significantly higher charges than non-smokers. In contrast, correlations between regional factors and charges varied, with geographical location demonstrating weaker associations compared to smoking status. These findings underscore the multifaceted nature of factors influencing insurance charges, with smoking status emerging as a particularly influential determinant.

6.2. T-test for Differences in Charges between Smokers and Non-Smokers:



The results of the Welch Two Sample t-tests reveal notable variations in insurance charges across different demographic categories. Age emerges as a significant determinant, with individuals in higher age brackets bearing substantially higher charges compared to their younger counterparts. Specifically, those in the elder age group face a mean charge of $16,261.71, contrasting sharply with the $10,094.77 mean charge among the younger cohort. Similarly, BMI levels wield considerable influence, as individuals with higher BMI experience significantly heftier charges, with a mean of $15,655.74, compared to the $10,877.96 mean charge observed among those with lower BMI. Furthermore, the number of children also plays a role, with individuals having more children encountering higher charges, averaging $13,949.94, compared to $12,365.98 for those with fewer children. Perhaps most strikingly, smoking status showcases a stark divide, with smokers facing substantially higher charges than non-smokers. Smokers are burdened with a mean charge of $32,050.23, while non-smokers bear a significantly lighter financial load, with a mean charge of $8,434.27. This discrepancy is further emphasized by the confidence interval, which indicates consistently lower charges among non-smokers.

6.3. Chi-square tests to find influence of Demographics on Insurance Charges Relative to the Median:

The Pearson's Chi-squared tests revealed significant insights into the relationship between demographic variables and insurance charges relative to the median.

Regarding sex, the analysis found no substantial association (X-squared = 0.0029899, df = 1, p = 0.9564), indicating that sex does not significantly impact insurance charges compared to the median.

Conversely, smoking status exhibited a strong association (X-squared = 342.05, df = 1, p < 2.2e-16), highlighting a significant influence on insurance charges relative to the median.

In contrast, regional differences showed no significant association (X-squared = 4.466, df = 3, p = 0.2153), suggesting that geographical location does not notably affect insurance charges compared to the median.

These findings underscore the pivotal role of smoking status in shaping insurance charges, while sex and region appear to have minimal influence in comparison.

6.4. Shapiro-Wilk test for normality for charges:

A Shapiro-Wilk normality test was conducted to assess the normality of the distribution of insurance charges. The test revealed a significant departure from normality (W = 0.81469, p-value < 2.2e-16), indicating that the distribution of charges may not follow a normal distribution. Since the "charges" variable in our dataset is not normally distributed, we will be using non-parametric tests instead of parametric tests.

6.5. Kruskal-Wallis Test Analysis for Differences in Insurance Charges:

The Kruskal-Wallis tests were conducted to assess differences in insurance charges across categorical and numerical variables. For categorical variables, significant disparities were observed based on smoker status (p < 0.001), while sex and region showed no significant differences. Among numerical variables, age (p < 0.001) and the number of children (p = 0.0000186) displayed significant variations in charges, indicating influences of age and family size on insurance costs. However, BMI did not show significant differences in charges. These findings underscore the importance of considering various demographic factors when analyzing insurance charges, with smoker status and age emerging as significant determinants.

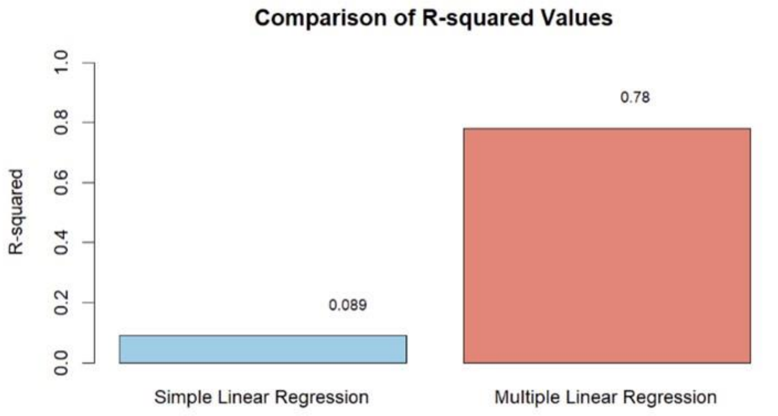
7. Linear Regression Model:

The linear regression model was employed to explore the relationship between various predictors and insurance charges. The analysis revealed several key findings: Age, BMI, and the number of children exhibited significant associations with charges (p < 0.001 for age and BMI, p = 0.001 for children). For each year's increase in age, charges increased by $256.9, while each unit increase in BMI was associated with a $339.2 increase in charges. Additionally, each additional child led to a $475.5 increase in charges. Smoking status emerged as a highly significant predictor (p < 0.001), with smokers facing substantially higher charges, averaging $23,848.5 more than non-smokers. Among regions, the southeast and southwest showed significant differences in charges compared to the baseline region (northeast), with individuals in these regions facing lower charges by $1,035.0 and $960.0, respectively. Gender did not show a significant association with charges (p = 0.693). Overall, the model exhibited a strong fit (Adjusted R-squared = 0.7494), underscoring the importance of age, BMI, smoking status, and region in predicting insurance charges. These findings offer valuable insights into factors influencing insurance costs, aiding in pricing strategies and risk assessment.

8. Multiple Regression Model:

The multiple linear regression model was utilized to predict insurance charges based on several predictor variables including age, sex, BMI, children count, smoking status, and region. Upon evaluating the model, it was found that age, BMI, number of children, smoking status, and a binary charges variable significantly contributed to predicting insurance charges. Specifically, each additional year of age was associated with a $144.02 increase in charges, while an increase of one unit in BMI led to a $314.46 rise in charges. Furthermore, smokers incurred significantly higher charges compared to non-smokers, with an estimated increase of $19880.02. However, the coefficients for region and sex were not statistically significant. The model explained approximately 77.86% of the variance in insurance charges, indicating a strong fit to the data.

9. Model Performance:



The multiple linear regression model, which likely includes more than one independent variable, has a considerably higher R-squared value of 0.78 compared to the simple linear regression model with an R-squared of 0.089. This suggests that the multiple linear regression model explains a larger portion of the variance in the dependent variable and has a better fit to the data.

10. Conclusion:

In conclusion, our exploratory data analysis of the insurance dataset yielded valuable insights into the factors influencing insurance charges. Through rigorous statistical analysis, we discovered a significant relationship between predictor variables such as age, BMI, smoking status, and region, and insurance charges. By rejecting the null hypothesis and accepting the alternate hypothesis, we affirm that these variables play pivotal roles in determining insurance costs.

Moreover, our comparison between the simple linear regression model and the multiple linear regression model revealed that the latter outperformed the former in predicting insurance charges. By considering multiple factors simultaneously, including age, BMI, smoking status, and region, the multiple regression model provided deeper insights into the determinants of insurance pricing. This underscores the importance of adopting a comprehensive approach in modeling insurance charges for more accurate predictions.

These findings hold practical implications for insurance companies, enabling them to better assess risk factors and tailor pricing strategies accordingly. By leveraging these insights, insurers can enhance their decision-making processes, optimize pricing strategies, and ultimately improve their operational efficiency. Additionally, our study contributes to a deeper understanding of the dynamics within the insurance industry, highlighting the value of data-driven approaches in navigating complex market landscapes.

11. References

Medical cost personal datasets. (2018). Kaggle. https://www.kaggle.com/datasets/mirichoi0218/insurance

Vu, J., & Harrington, D. (2020). *Introductory Statistics for the Life and Biomedical Sciences: Preliminary Edition*.

Carvalho, J. N. de, de Camargo Cancela, M., & de Souza, D. L. B. (2018). Lifestyle factors and high body mass index are associated with different multimorbidity clusters in the

Brazilian population. *PLOS ONE*, *13*(11), e0207649. https://doi.org/10.1371/journal.pone.0207649

Appendix:

Exploratory Data Analysis:

Data Visualization:

```
# Explore the relationship between age and charges

plot(data$age, data$charges, main = "Age vs. Charges", xlab = "Age", ylab = "Charges", col = "blue")

# Explore the relationship between BMI and charges

plot(data$bmi, data$charges, main = "BMI vs. Charges", xlab = "BMI", ylab = "Charges", col = "green")

# Compare charges by gender
```

~~~
boxplot(charges ~ sex, data = data, main = "Charges by Gender", xlab = "Gender", ylab = "Charges", col = c("pink", "lightblue"))
# Compare charges by smoker status
boxplot(charges ~ smoker, data = data, main = "Charges by Smoker Status", xlab = "Smoker Status", ylab = "Charges", col = c("lightgreen", "orange"))
# Compare charges by region
boxplot(charges ~ region, data = data, main = "Charges by Region", xlab = "Region", ylab = "Charges", col = "yellow")
# Explore the relationship between the number of children and charges
plot(data$children, data$charges, main = "Number of Children vs. Charges", xlab = "Number of Children", ylab = "Charges", col = "purple")
# Identify outliers or anomalies in charges
boxplot(data$charges, main = "Charges Boxplot", ylab = "Charges", col = "red")
# Create a scatterplot matrix to visualize relationships between variables
pairs(data[, c("age", "bmi", "children", "charges")], main = "Scatterplot Matrix", col = "blue")
# Add Q-Q plots
par(mfrow = c(2, 2)) # Set the layout to 2x2 for Q-Q plots
for (variable in c("age", "bmi", "charges")) {
  if (is.numeric(data[[variable]])) {
    qqnorm(data[[variable]], main = paste("Q-Q Plot of", variable), col = "green")
    qqline(data[[variable]], col = "red")
  }
}
```
~~~

Spearman's correlation:

~~~
```{r}
# Load necessary library
library(Hmisc)

# Check for missing values
sum(is.na(insurance_data))
# Remove rows with missing values (if any)
~~~

~~~
insurance_data <- na.omit(insurance_data)
# Convert non-numeric columns to numeric if possible
insurance_data <- type.convert(insurance_data, as.is = TRUE)
# Convert categorical variables to factors
insurance_data$sex <- as.factor(insurance_data$sex)
insurance_data$smoker <- as.factor(insurance_data$smoker)
insurance_data$region <- as.factor(insurance_data$region)
# Identify categorical columns
categorical_cols <- c("sex", "smoker", "region")
# Convert categorical columns to numerical using one-hot encoding
encoded_data <- model.matrix(~ 0 + ., data = insurance_data[, categorical_cols])
# Combine the encoded columns with the original dataframe
insurance_data_numeric <- cbind(insurance_data[, !names(insurance_data) %in% categorical_cols], encoded_data)
# Compute Spearman's correlation
cor_matrix <- rcorr(as.matrix(insurance_data_numeric), type = "spearman")
# Print the correlation matrix
print(cor_matrix$r)
```
```{r}
# Visualize the correlation matrix using a heatmap
corrplot(cor_matrix$r, method = "color", main = "Correlation Matrix")
```
~~~

Statistical Analysis:

Performing T-Test:

~~~
```{r}
# Filter all numerical columns (excluding the target variable 'charges')
numeric_cols <- data %>%
  select_if(is.numeric)
# Exclude the target variable 'charges'
numeric_cols <- numeric_cols %>%
~~~

~~~
  select(-charges)
# Filter encoded numerical columns (if you've encoded categorical variables)
encoded_numeric_cols <- data %>%
  select(starts_with("encoded_"))  # Adjust this based on the column names after encoding
# Combine all numerical columns
all_numeric_cols <- cbind(numeric_cols, encoded_numeric_cols)
# Define a function to binarize numeric variables
binarize_numeric <- function(x) {
  ifelse(x >= median(x), "high", "low")
}

# Perform t-tests for each numeric column
t_test_results <- lapply(all_numeric_cols, function(col) {
  # Binarize the numeric variable
  binarized_col <- binarize_numeric(col)
  # Perform t-test
  t_test_result <- t.test(data$charges ~ binarized_col, data = data)
  return(t_test_result)
})
# Print t-test results
names(t_test_results) <- names(all_numeric_cols)
t_test_results
```
~~~

Chi-square tests for each categorical column:

~~~
```{r}
# Perform chi-square tests for each categorical column
chi_square_results <- lapply(categorical_cols, function(col) {
  chi_square_result <- chisq.test(table(data[[col]], data$charges > median(data$charges)))
  return(chi_square_result)
})
~~~

~~~
# Print chi-square test results
names(chi_square_results) <- categorical_cols
chi_square_results
```
~~~

Test for Normality:

~~~
```{r}
# Perform Shapiro-Wilk test for normality
shapiro_test_result <- shapiro.test(data$charges)
# Print the test result
shapiro_test_result
```
~~~

Non-Parametric Tests:

Kruskal-Wallis test:

~~~
```{r}
# Filter categorical and numerical variables
categorical_vars <- c("sex", "smoker", "region")
numeric_vars <- c("age", "bmi", "children")
# Perform Kruskal-Wallis test for categorical variables
kruskal_cat <- lapply(categorical_vars, function(var) {
  kruskal.test(charges ~ get(var), data = data)
})
# Perform Kruskal-Wallis test for numerical variables
kruskal_num <- lapply(numeric_vars, function(var) {
  kruskal.test(data$charges ~ data[[var]])
})
# Print results for categorical variables
cat_results <- data.frame(
  Variable = categorical_vars,
~~~

~~~
  Statistic = sapply(kruskal_cat, function(x) x$statistic),
  P_Value = sapply(kruskal_cat, function(x) x$p.value)
)
print("Kruskal-Wallis test results for categorical variables:")
print(cat_results)
# Print results for numerical variables
num_results <- data.frame(
  Variable = numeric_vars,
  Statistic = sapply(kruskal_num, function(x) x$statistic),
  P_Value = sapply(kruskal_num, function(x) x$p.value)
)
print("Kruskal-Wallis test results for numerical variables:")
print(num_results)
```
~~~

Linear Regression Model:

~~~
```{r}
# Fit a linear regression model
linear_model <- lm(charges ~ age + bmi + children + sex + smoker + region, data = data)
# Summary of the linear regression model
summary(linear_model)
```
~~~

Multiple Regression Model:

~~~
```{r}
# Fit a multiple regression model
multiple_model <- lm(charges ~ ., data = data)
# Summary of the multiple regression model
summary(multiple_model)
```
~~~

Model Performance:

~~~
```{r}
~~~

~~~
# Linear Regression
# Simple Linear Regression
simple_lm <- lm(charges ~ age, data = data)
r_squared_simple <- summary(simple_lm)$r.squared
# Multiple Linear Regression
multiple_lm <- lm(charges ~ ., data = data)
r_squared_multiple <- summary(multiple_lm)$r.square
# Combine R-squared values with model names
model_names <- c("Simple Linear Regression", "Multiple Linear Regression")
r_squared_values <- c(r_squared_simple, r_squared_multiple)
# Create a bar plot to visualize R-squared values
barplot(r_squared_values,
        names.arg = model_names,
        ylab = "R-squared", main = "Comparison of R-squared Values",
        col = c("skyblue", "salmon"), ylim = c(0, 1)
# Add text labels for R-squared values
text(x = 1:2, y = r_squared_values + 0.05, round(r_squared_values, 3), pos = 3, cex = 0.8, col = "black")
```
~~~