Introduction to Business Analytics

Final Project

Insurance Charges Analysis and Model Evaluation

Submitted by:

Hiral Shah - 101539997

Rajesh Adep - 101540051

Shakshi Maheshwari - 101539638

Manushka Mhatre - 101543219

Table of Contents

1. Introduction

This report presents an analysis of a dataset containing information on health insurance charges. The goal of this study is to explore the relationships between various demographic factors (age, sex, BMI, children, smoker status, region) and insurance charges. Additionally, a linear regression model is built to predict insurance charges using selected features. This report is divided into four major sections:

- Exploratory Data Analysis (EDA) Results
- Data Visualizations and Insights
- Linear Regression Analysis and Performance Evaluation
- Conclusions and Recommendations for Improving the Model

2. Exploratory Data Analysis (EDA) Results

2.1 Dataset Overview

The dataset contains 1338 rows and 7 columns:

- age: Age of the individual.
- sex: Gender of the individual.
- bmi: Body Mass Index, a measure of one's body fat that is based on height and weight.
- children: Number of children/dependents covered by the insurance.
- smoker: Smoking status (yes/no).
- region: Geographic region where the individual lives.
- charges: The medical insurance charges billed to the individual.

2.2 Duplicate and Missing Data

There was 1 duplicate record, which was removed, reducing the dataset to 1337 rows.

No missing data was found in any of the columns, so no imputation was necessary.

2.3 Summary Statistics

Key summary statistics for the numerical features are as follows:

- Age: The individuals range from 18 to 64 years old, with a mean age of 39.22.
- BMI: The BMI values range from 15.96 to 53.13, with a mean of 30.66, indicating the presence of overweight and obese individuals.
- Charges: The insurance charges vary significantly, from $1121.87 to $63,770.43, with an average of $13,279.12.
- Children: Most individuals have between 0 to 5 children, with the mean number of children being 1.09.

2.4 Outlier Detection

- BMI: 9 outliers were detected in the BMI values, indicating individuals who are significantly underweight or overweight compared to the majority.
- Charges: 139 outliers were found in the charges, indicating that some individuals have significantly higher medical costs.

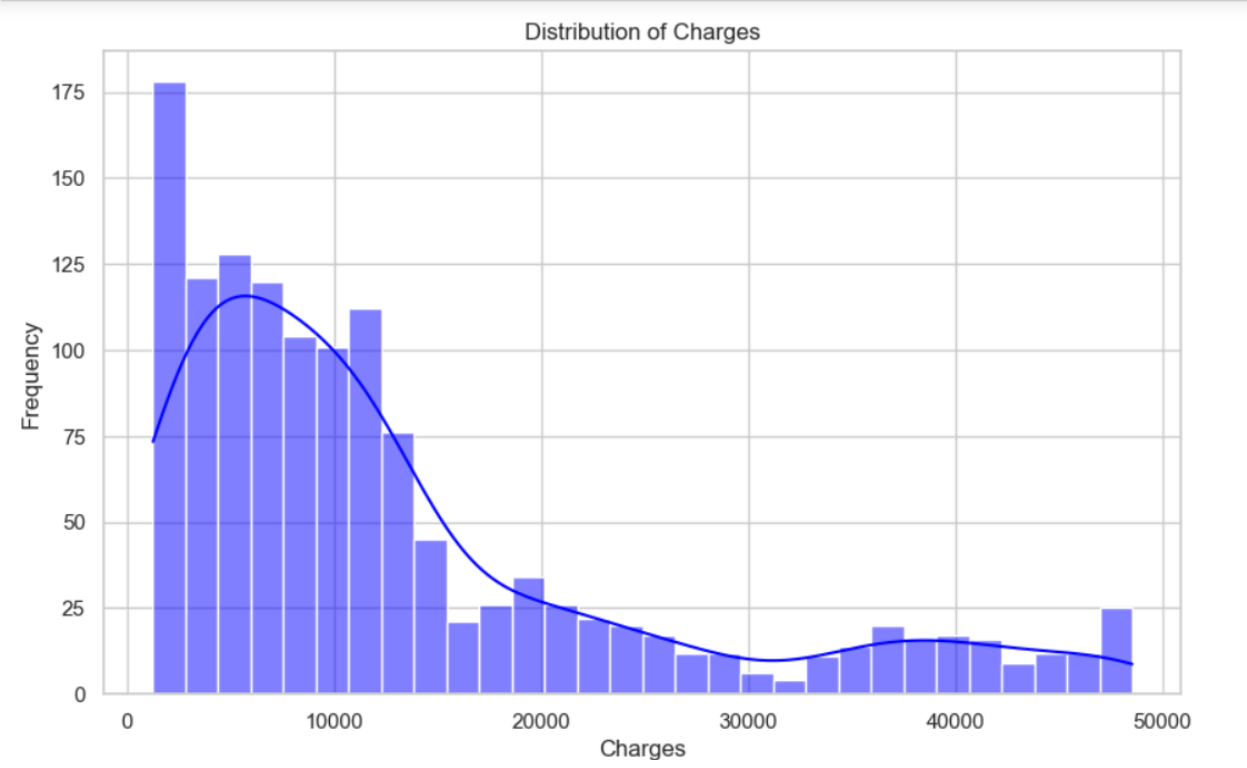
These outliers were detected using the IQR method. These outliers were capped at the 1st and 99th percentiles to avoid their undue influence on the regression model.

3. Data Visualizations and Insights

3.1 Distribution of Charges

Description: The distribution of charges is heavily right-skewed, with most individuals incurring lower medical charges. A few individuals have very high charges, creating a long tail.

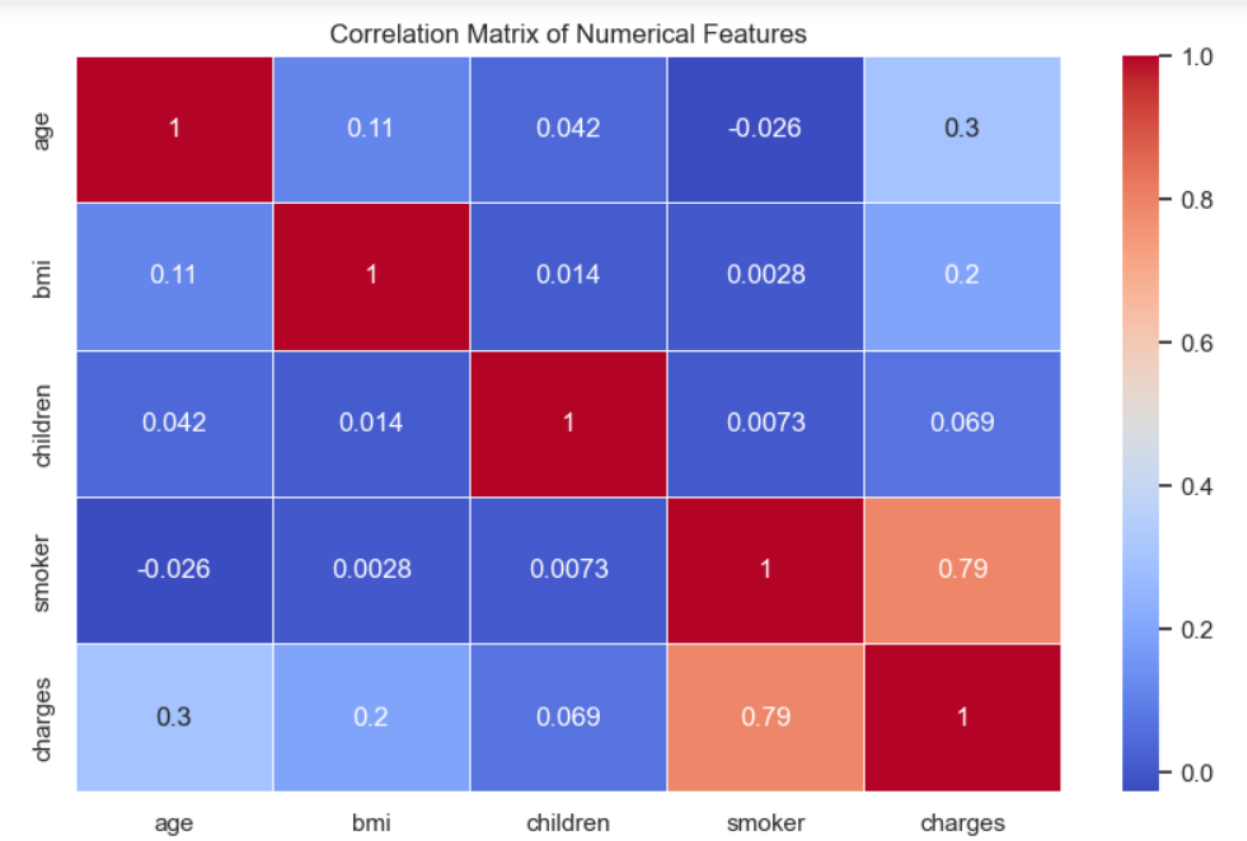
Insight: This skewness is indicative of certain individuals having very high medical costs, potentially due to severe health issues or smoking habits.



3.2 Correlation Matrix

Description: A correlation matrix heatmap shows the relationships between numerical variables (age, BMI, children, smoker, and charges).

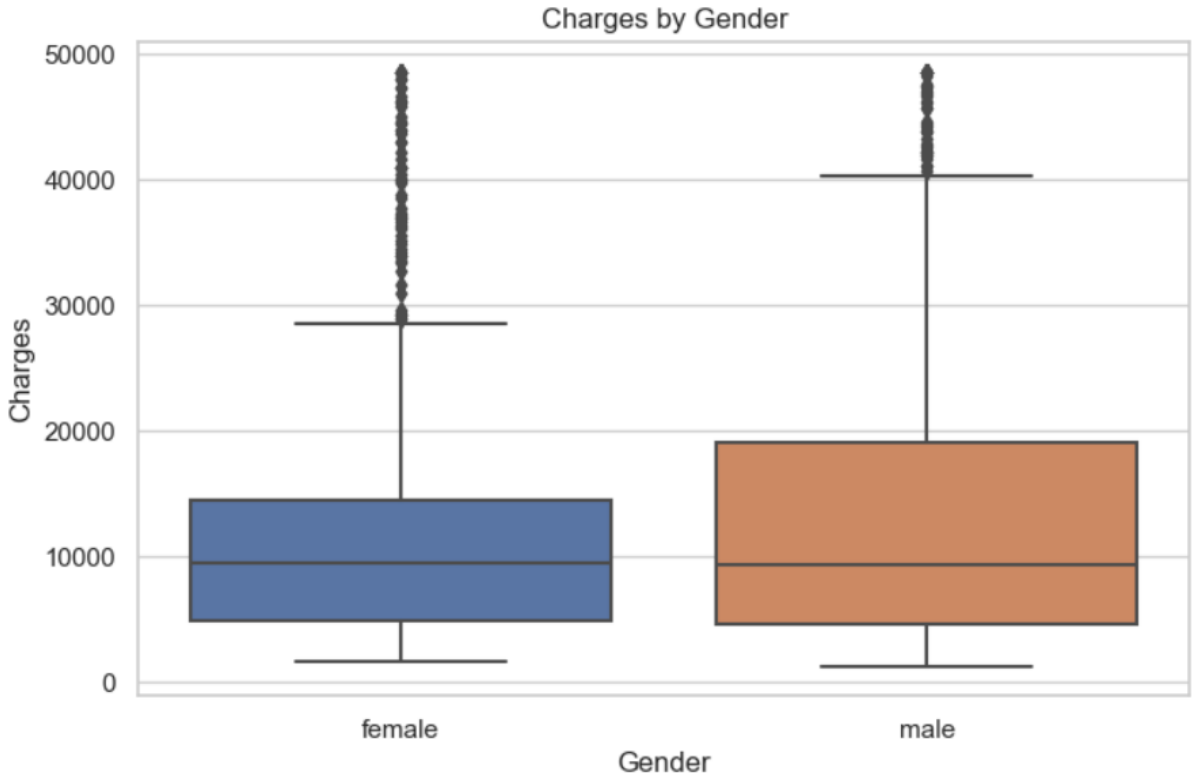
Insight: Smoking status has the strongest positive correlation with charges (0.79), followed by age (0.30) and BMI (0.20). This suggests that smoking is a major determinant of medical costs.



3.3 Charges by Gender

Description: A box plot comparing insurance charges across gender.

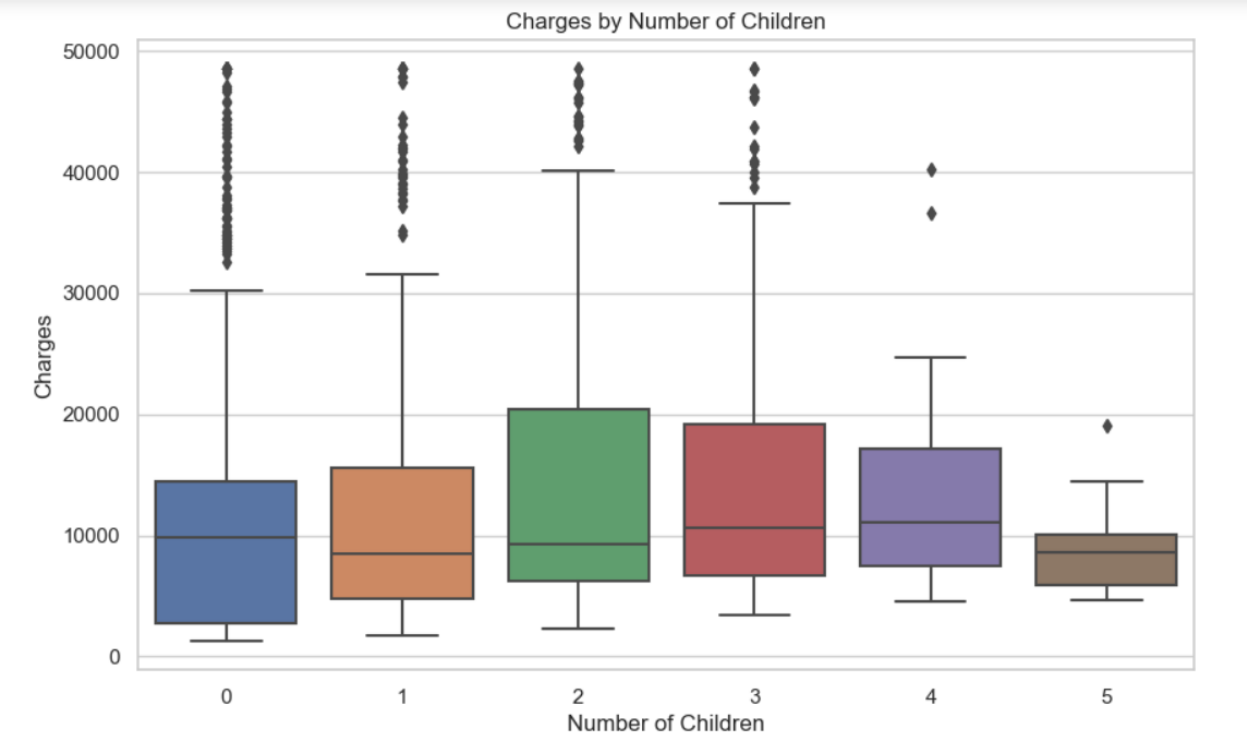
Insight: There is no significant difference in charges between males and females, with both genders having similar median. Though Males have a higher IQR whereas Females having more outliers.



3.4 Charges by Number of Children

Description: A box plot comparing charges across individuals with different numbers of children.

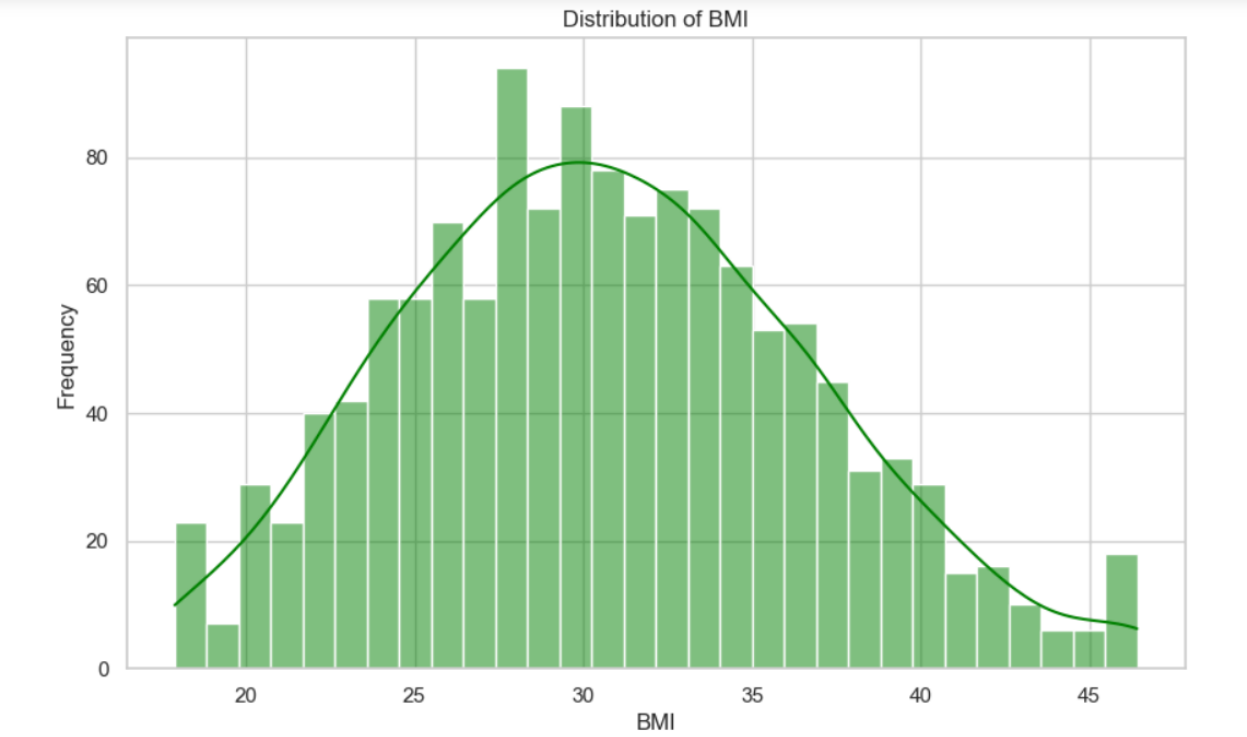
Insight: The number of children does not have a significant impact on charges, as median values are similar and close across different groups.



3.5 Distribution of BMI

Description: A histogram of BMI values shows that most individuals fall in the 20-40 range.

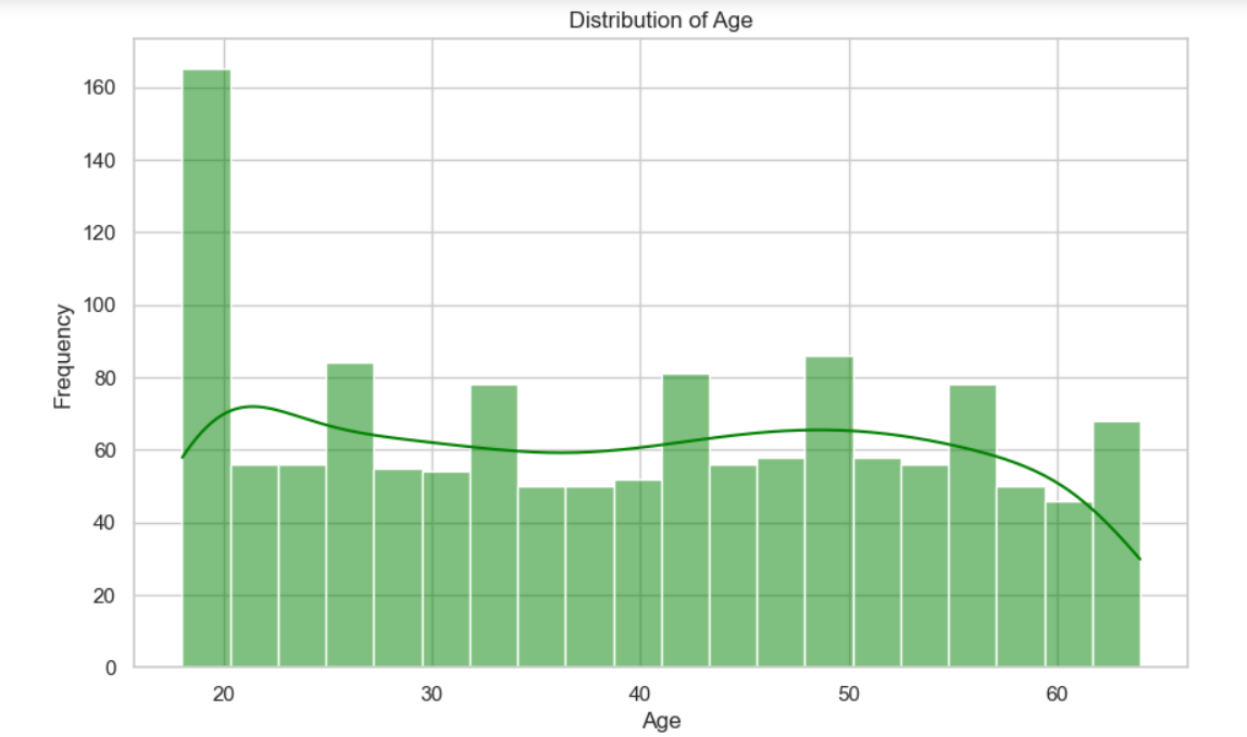
Insight: This distribution highlights that most of the population in the dataset is either overweight or obese, given that BMI values above 25 are considered overweight.



3.6 Distribution of Age

Description: A histogram of age distribution shows that the dataset is fairly balanced across age groups, but slightly skewed toward middle-aged individuals.

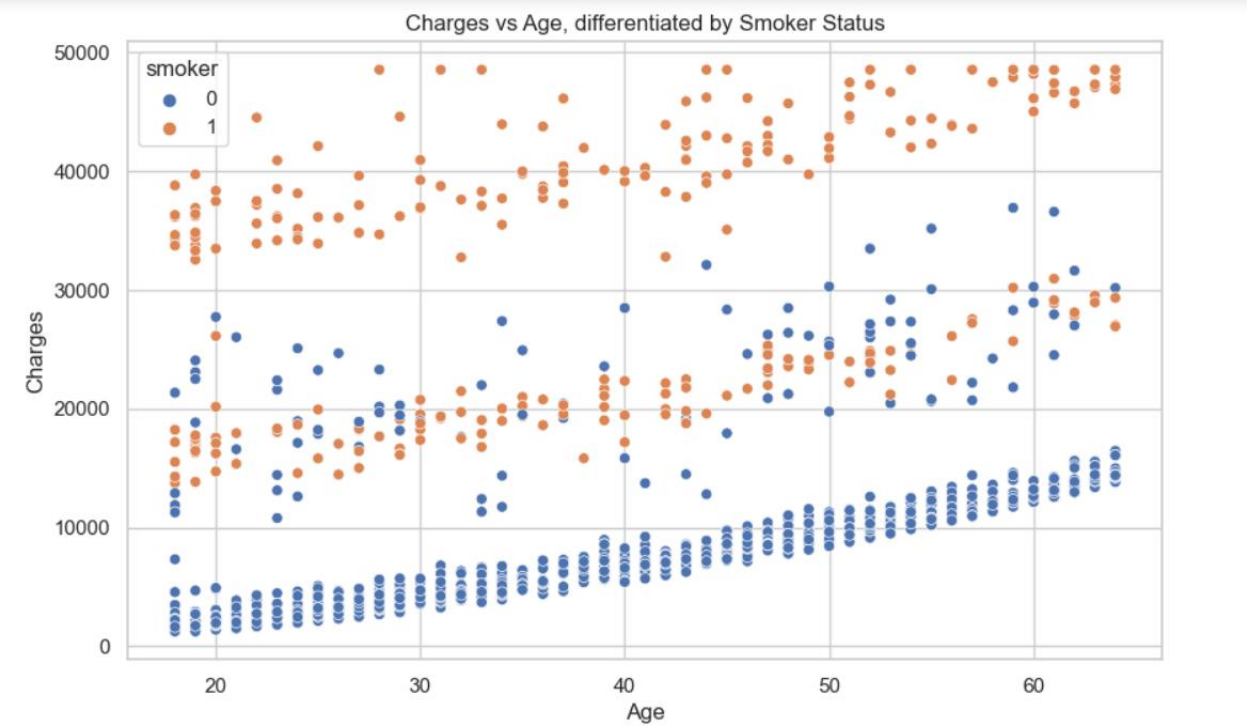
Insight: The distribution is representative, with a healthy spread across ages.



3.7 Charges vs. Age (Smoker vs. Non-Smoker)

Description: A scatter plot of charges vs age, with smoker status highlighted.

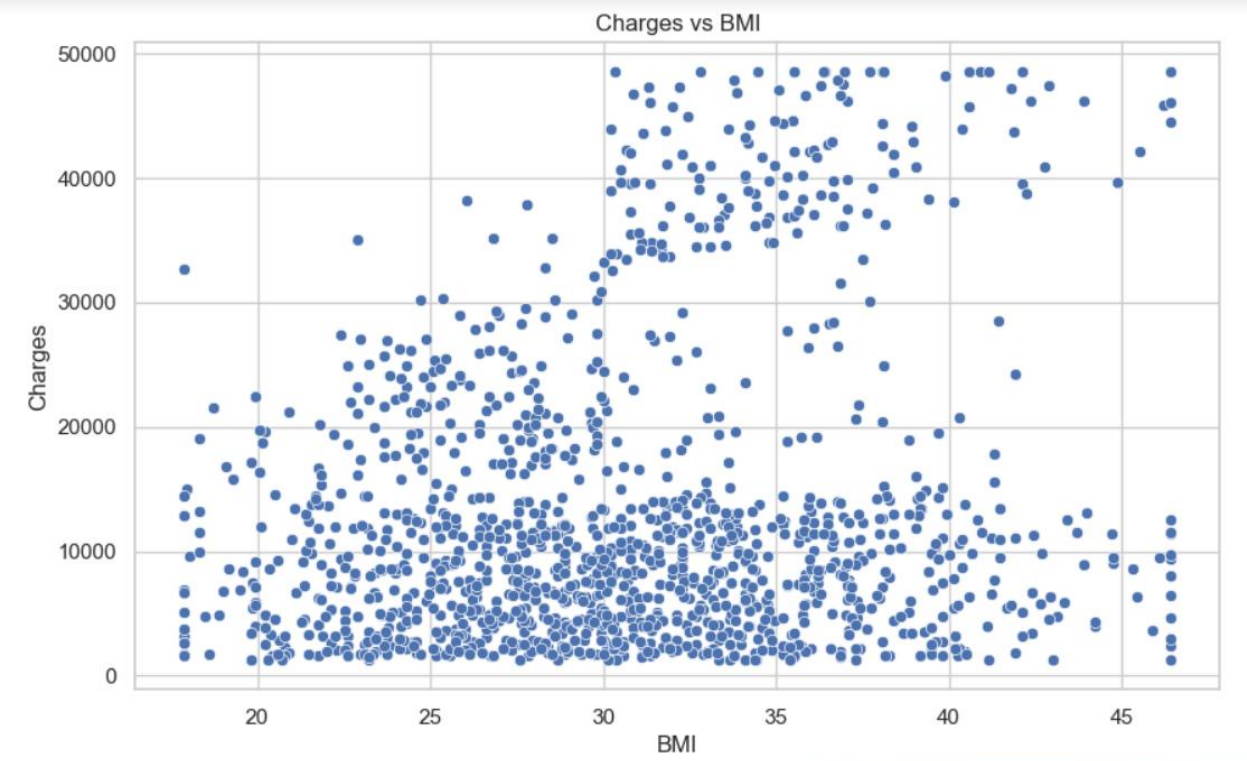
Insight: Smokers consistently have higher charges, especially as they age, while non-smokers generally have lower, more stable charges.



3.8 Charges vs. BMI

Description: A scatter plot comparing charges and BMI.

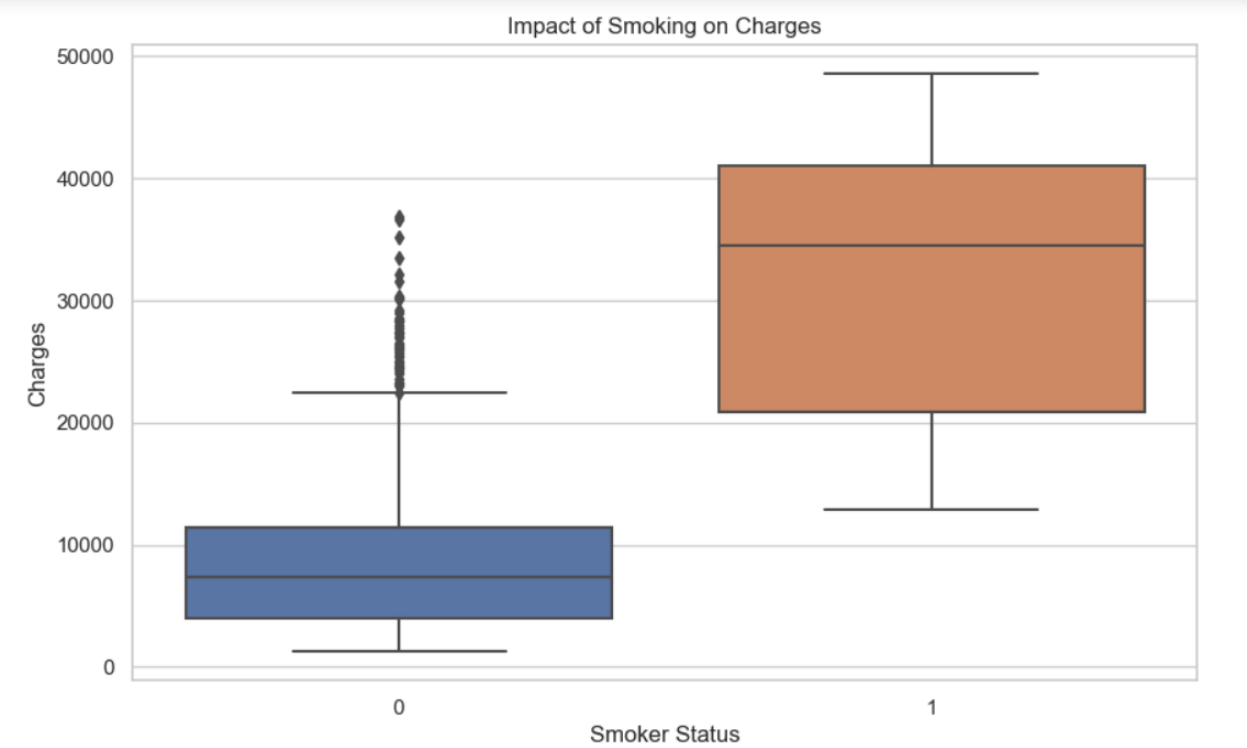
Insight: BMI is not a strong predictor of charges, though some individuals with high BMIs do incur higher medical costs.



3.9 Impact of Smoking on Charges

Description: A box plot comparing charges for smokers and non-smokers.

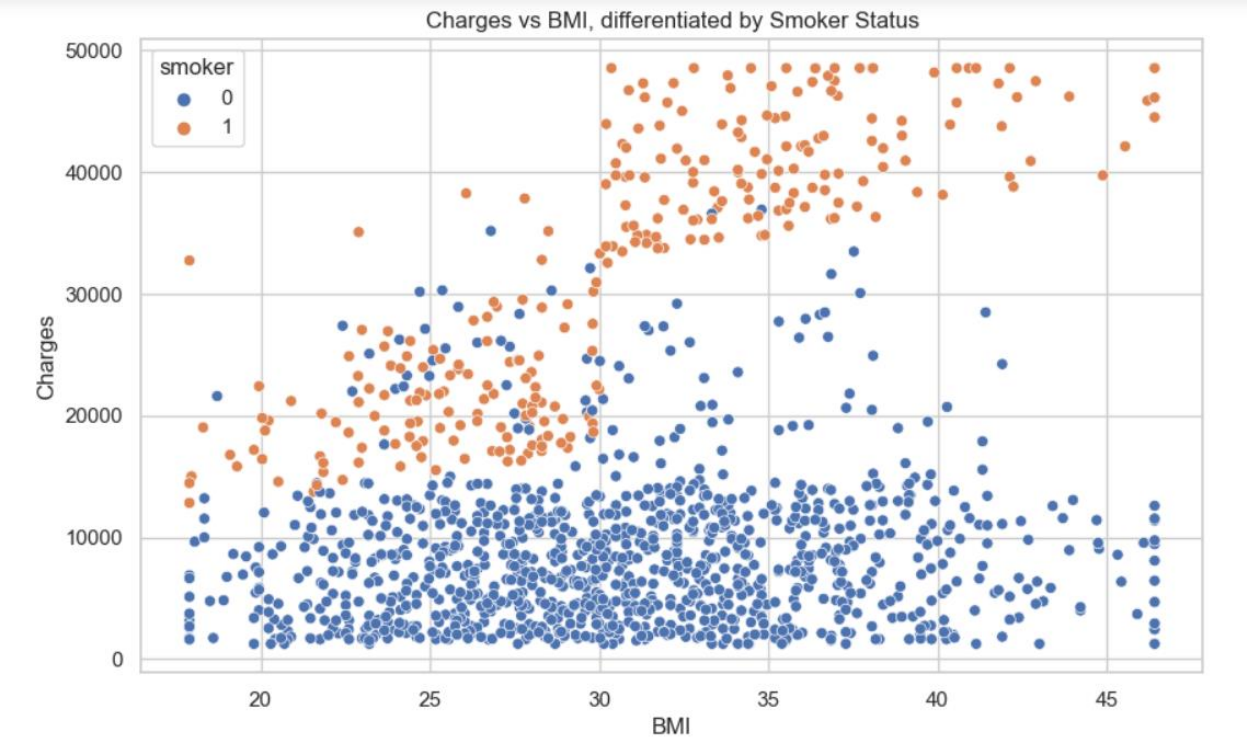
Insight: Smoking dramatically increases medical costs. Smokers have a much wider range of charges, with a significantly higher median compared to non-smokers.



3.10 Charges vs. BMI (Smoker vs. Non-Smoker)

Description: A scatter plot of charges vs BMI, with smoker status highlighted.

Insight: Smokers incur significantly higher medical charges than non-smokers, with charges for smokers increasing as BMI rises, especially for those with a BMI above 30.



4. Linear Regression Analysis and Performance Evaluation

4.1 Model Construction

A linear regression model was built using three key features:

- age
- bmi
- smoker (binary: yes = 1, no = 0)

The target variable is charges.

The dataset is split in the ratio of 80% for training and 20% testing.

4.2 Model Output

Coefficients:

- Age: 251.25
- BMI: 307.55
- Smoker: 22,941.73

Intercept: -10,825.66

R-squared value: 0.8191

4.3 Interpretation of Coefficients

Age: For each additional year of age, charges increase by approximately $251.

BMI: A unit increase in BMI increases charges by approximately $307.

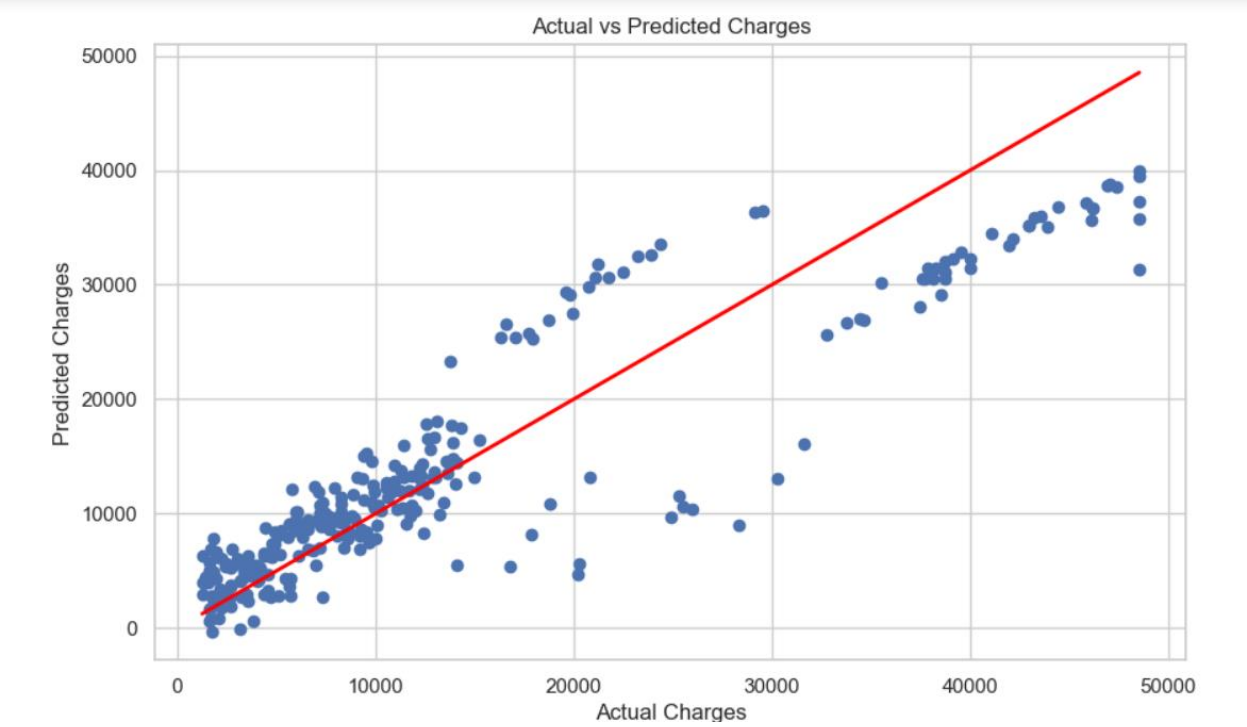
Smoking: Being a smoker increases insurance charges by $22,941, indicating a substantial impact of smoking on healthcare costs.

4.4 Model Performance

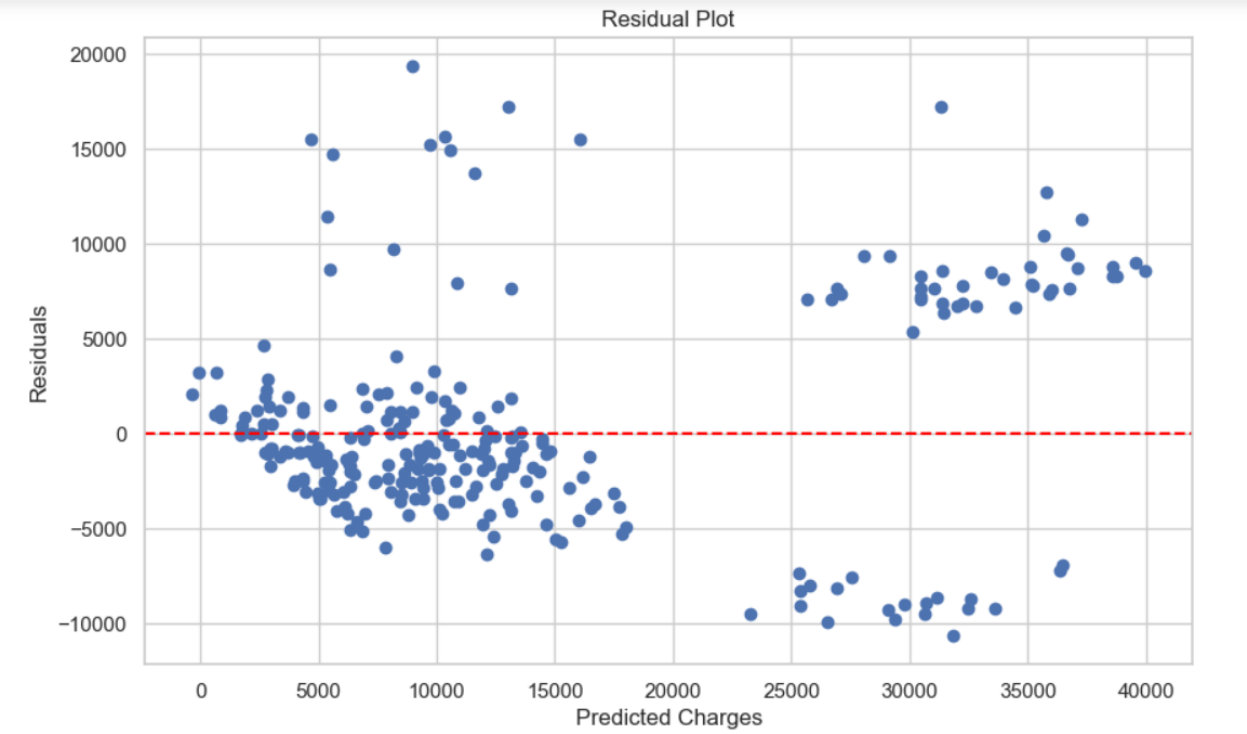
The R-squared value of **0.8191** indicates that the model explains approximately **81.91%** of the variance in charges, which is a strong fit for a simple linear regression model. However, there may still be room for improvement, especially in capturing some of the variation due to other factors not included in this model (e.g., region or gender interactions).

4.5 Prediction and Residual Analysis

Actual vs Predicted Charges: The scatter plot of actual vs predicted values shows a good fit, but some predictions deviate significantly from the actual values, especially at the higher end.



Residual Plot: The most residuals are randomly scattered around zero, indicating that the model performs well in capturing the linear relationships without major systematic errors.



5. Conclusions and Recommendations for Improving the Model

5.1 Key Findings

- Smoking is the most significant factor driving insurance charges, with a huge impact on costs.
- BMI and age also have notable, though smaller, impacts on charges.
- The current linear regression model explains 81.91% of the variation in insurance charges, but further improvement is possible.

5.2 Recommendations for Improving the Model

- Feature Engineering: Additional variables such as region or gender could improve the model.
- Non-Linear Relationships: Given the skewed distribution of charges and outliers, exploring non-linear models (e.g., decision trees) might yield better predictive performance.
- Outlier Treatment: While outliers were capped, further exploration of robust regression techniques that are less sensitive to outliers could improve model predictions.