**Capstone Project in Business Analytics -BA-64099-010**

**Project Title: Predictive Modelling of Health Insurance Charges**

**Table of Contents**

1. **Introduction**
   1. **Problem Statement**
   2. **Goal & Objective**
2. **Business Context**
3. **Methodology** 
   1. **Data Preparation**
   2. **EDA**
   3. **Feature Selection**
   4. **Model Building**
   5. **Model Evaluation**
4. **Data Overview**
5. **Results and Discussion**
   1. **Data Preparation**
   2. **EDA**
   3. **Feature Selection**
   4. **Model Building**
   5. **Model Evaluation**
6. **Conclusion**

**6.1 Recommendations**

1. **Ethical Considerations**
2. **References**
3. **Introduction**
   1. **Problem Statement:**

The challenges that the insurance industry is facing lie in the correctness of assessment regarding health insurance charges, which results in underpricing, overpricing, and incorrect allocation of resources. In most cases, traditional methods lack preciseness and fail to capture the full complexity of interactions between demographic and health-related factors that influence insurance costs. There is a great need for novel predictive modeling methods that enhance risk assessment, improve premium calculation, and inform decision-making in the insurance sector.

* 1. **Goal & Objective:**

Designing predictive models of health insurance charges using regression analysis techniques. These models will help the insurance company in risk assessment, premium pricing, and potentially expensive policyholders, making the insurance industry better decisions and leading to better financial stability.

Predictive Modeling: Develop regression models that can accurately predict health insurance charges by using individual demographic and health-related characteristics.

1. **Business Context**

In the competitive landscape of the insurance industry, precise risk assessment and premium pricing are crucial for maintaining financial stability and customer satisfaction. By leveraging advanced predictive modeling techniques, insurance companies can better understand the factors influencing health insurance charges, leading to more accurate pricing strategies and efficient resource allocation. This project provides insights that can help insurance firms enhance their actuarial practices, optimize pricing models, and improve overall profitability.

**3. Methodology**

**3.1 Data Preparation:**

Handling Missing Values: Identify and address missing values in the dataset using appropriate techniques, including imputation or deletion.

Detection of outliers: In the dataset, outlier detection will be done using statistical methods or visualization techniques; the decision to remove, transform, or retain them rests on domain knowledge.

Standardization of Data: Standardize numerical features to have a mean of 0 and a standard deviation of 1 to ensure consistent scaling across variables.

Encoding Categorical Variables: Converting categorical variables into numerical levels

**3.2 Exploratory Data Analysis:**

Univariate Analysis: The distribution of variables individually is considered with histograms, box plots, or summary statistics to understand their range and variability.

Bivariate Analysis: Investigating relationships between two variables by using scatter plots, correlation matrices, or grouped summaries to identify potential correlations or patterns.

Multivariate Analysis: This includes the interactions between several variables using different techniques such as heatmaps or clustering to find complex relationships between the data.

Visualization: Formulating plots of bar charts, pie charts, or heatmaps to display the major results and insights from the EDA process.

**3.3 Feature Selection**: Techniques like ANOVA, correlation and PCA will be employed to reduce and select the appropriate features

**3.4 Model Building**:

Data splitting into training and testing sets. I will use different regression models including Linear regression model and decision tree regression model, to identify which algorithm works best in the prediction of health insurance charges.

**3.5 Model Evaluation:**

Assessing the performance of trained models using appropriate evaluation metrics such as MSE, RMSE, and R-squared on a separate validation dataset to estimate their predictive capability and generalizability and interpreting the results.

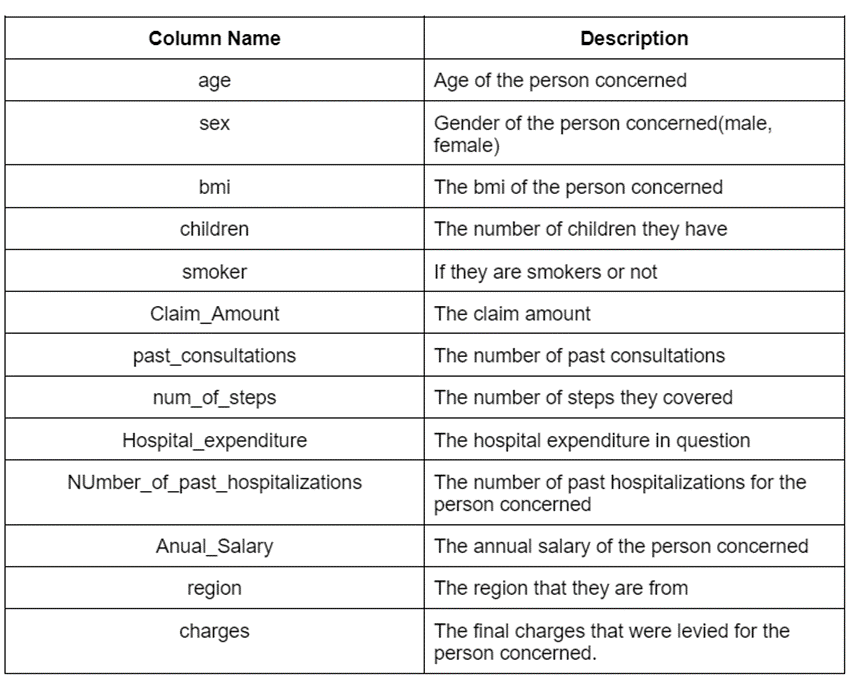
**4. Data Overview:**

There are over 10,000 records in this dataset; each record represents individual health insurance charges and various demographic and health-related attributes. Important features include gender, BMI, age, number of children, smoking status, region of residence, past consultations, hospital expenditures, and claim amounts. The target variable is the health insurance charges, which are influenced by these attributes.

Target Variable – Charges

Number of Variables – 13 Variables, 3- categorical variables, 10 – numerical

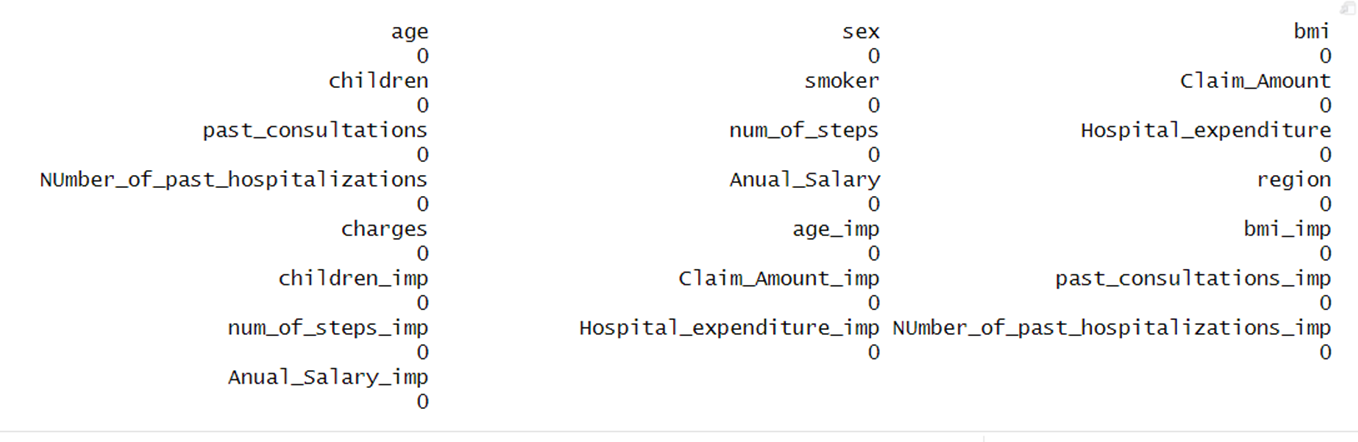
Missing Values – Yes



*Fig.1. Variable overview from dataset*

1. **RESULTS & DISCUSSION**

**5.1 Data Preparation:**

* + Handled missing values using imputation techniques – K-NN imputation has been performed with k = 5
  + Detected and managed outliers based on domain knowledge – Used box plots to visualize outliers and dealt with them.
  + Standardized numerical features and encoded categorical variables – Converted categorical data to numerical and normalized the data
  + After Data preparation and pre-processing, sample size has been reduced to 9006 from 10,008.

**5.2 Exploratory Data Analysis (EDA):**

**A graph of insurance coverage

Description automatically generated5.2.1 Univariate Analysis**: Conducted **univariate analysis** to understand variable distributions. From graphs, the distribution is not normal and is skewed to right.

A graph of numbers and a number of numbers

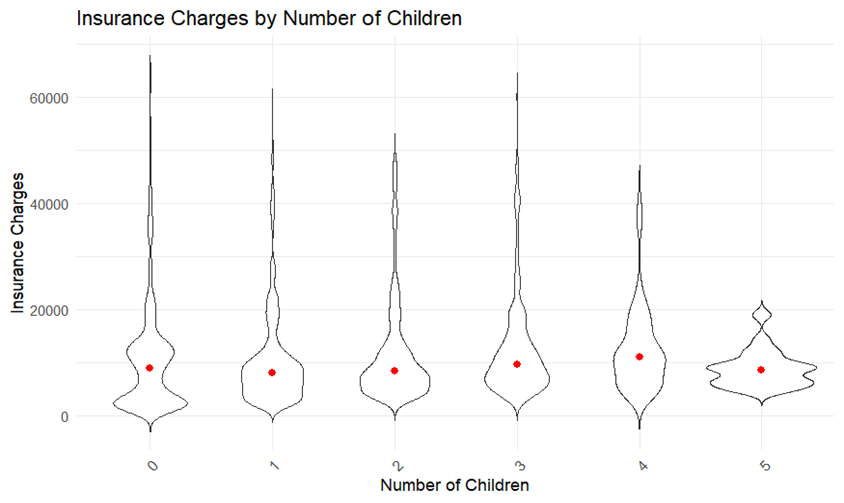
Description automatically generated with medium confidence

**5.2.2 Bivariate Analysis:**

A graph showing different colored rectangles

Description automatically generated

**Interpretation:** Analyzing insurance charges across four regions in the US reveals the South East region as the most expensive, with an average cost of $11,750. Following closely behind is the North East at $11,500. The North West region offers a slight decrease at $11,250, while the South West boast the most affordable insurance with an average of $11,000.



**Interpretation:**

Based on the chart, it appears that people with more children tend to have higher insurance charges. This could be because:

* There are more people to cover under the insurance plan, increasing the overall cost.
* Children may require more medical care, leading to higher healthcare costs for the family.

A graph of age versus insurance charges

Description automatically generated

**Interpretation:**

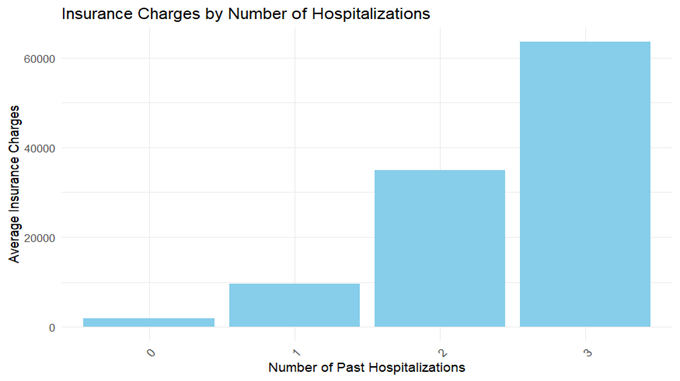
Examining insurance premiums by age indicates a distinct and predictable pattern: insurance costs generally rise with age. This finding corresponds with the increased health risks and medical requirements as individuals age.

A graph showing a number of black dots

Description automatically generated

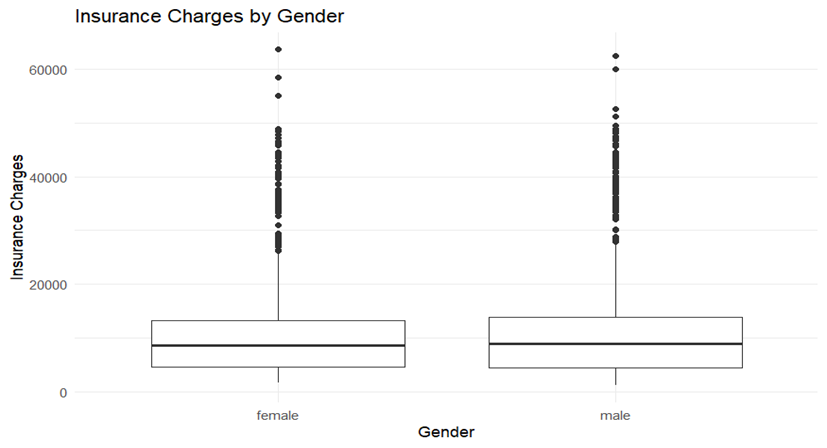
**Interpretation:**

Analyzing the BMI vs Charges graph reveals a positive correlation, meaning as BMI increases, so do insurance charges. This trend likely reflects the association between higher BMI, greater risks for expensive health conditions, and the need for more preventive care.



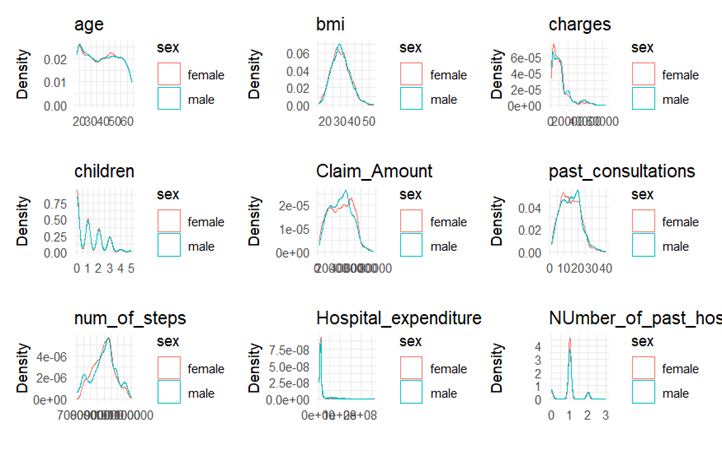
**Interpretation:**

The analysis reveals that insurance charges tend to rise with an increasing number of hospitalizations. As individuals experience more hospitalizations, their insurance costs also increase accordingly.



**Interpretation:**

The "Insurance Charges by Gender" boxplot reveals a trend of potentially higher charges for females. The median charge for females sits higher than males, and the wider spread in the female box suggests greater variability in their costs. This difference could be due to factors like pregnancy-related care or historical cost structures, but it's important to remember that individual charges depend on various health factors, not just gender.



**Density Plots:**

**Interpretation:**

The density plots indicate that there are no principal differences in the distributions of the variables assessed between males and females. As a matter of fact, for all variables, both genders do not differ too much from one another, ranging from age, BMI, and charges to children, claim amount, past consultation, number of steps, hospital expenditure, and the number of past hospitalizations; thus, showing that gender is basically not a chief determinant of these health and demographic factors in this dataset.

A screenshot of a computer screen

Description automatically generated

**Interpretation:**

The density plots reveal significant differences between smokers and non-smokers in several health and demographic variables. Smokers tend to have higher health insurance charges, claim amounts, hospital expenditures, and more past consultations and hospitalizations. These differences highlight the increased healthcare burden associated with smoking, underscoring the importance of targeted health interventions and policies to reduce smoking rates and associated costs.

A close-up of a chart

Description automatically generated

**Interpretation:**

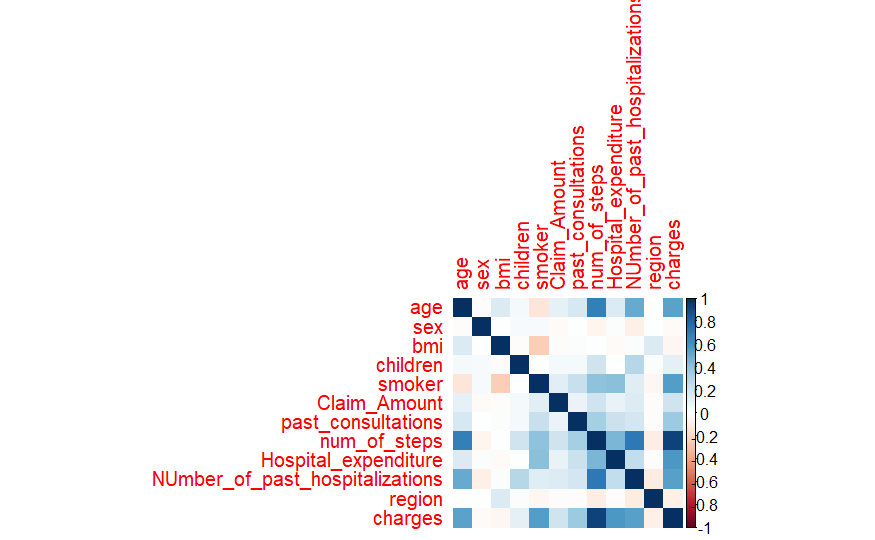
The density plots reveal regional differences in health insurance charges, claim amounts, past consultations, hospital expenditures, and the number of past hospitalizations. The southeast region consistently shows higher healthcare costs and more frequent healthcare utilization. These insights highlight the importance of considering regional variations in healthcare planning and resource allocation. By understanding these regional differences, policymakers and healthcare providers can develop targeted interventions to address the specific needs of each region and improve overall healthcare efficiency and effectiveness.

**5.2.3 Multivariate Analysis:**

A graph of a person and person

Description automatically generated

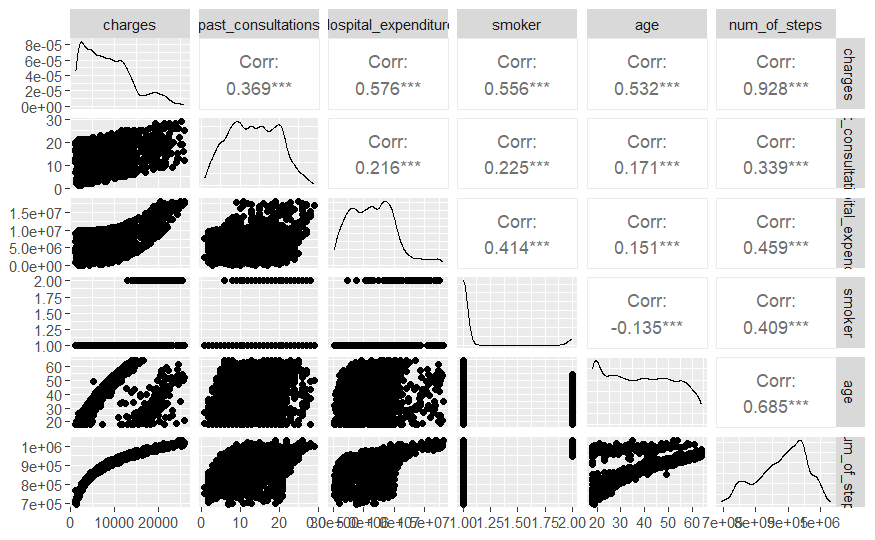
**Interpretation:**

While non-smokers of both genders benefit from lower insurance costs, a curious trend emerges for smokers. Interestingly, women who smoke appear to have lower insurance premiums compared to men who smoke. This suggests that although smoking typically increases insurance costs, there might be gender-specific factors influencing how these costs are calculated.

**Correlation Matrix:**

**Interpretation:**

* Charges are highly positively correlated with Hospital\_expenditure and Claim\_Amount.
* Smoker status shows a strong positive correlation with charges and Claim\_Amount.
* BMI has a moderate positive correlation with charges and hospital expenditure.
* Number of past hospitalizations is positively correlated with hospital expenditure and past consultations

**Pair Metrics**:

**Interpretation:**

The pair plot and correlation matrix reveal significant relationships between the variables. Higher health insurance charges are associated with more past consultations, higher hospital expenditures, smoking status, and older age. Past consultations are also positively correlated with hospital expenditure, smoker status, age, and the number of steps. Smokers and older individuals tend to incur higher hospital expenditures. Older individuals also tend to take more steps.

These insights highlight the complex interactions between demographic factors, health behaviors, and healthcare costs. Understanding these relationships can help healthcare providers and policymakers develop targeted interventions to manage healthcare costs and improve patient outcomes. For instance, initiatives to reduce smoking rates or promote healthier lifestyles among older adults could potentially reduce healthcare expenditures and improve overall health.

**Feature Selection:**

**ANOVA**

**A screenshot of a computer

Description automatically generated**

From the Analysis of the Variance Table, all the variables are significant. This could be due to correlation among the features which can be further addressed in the PCA

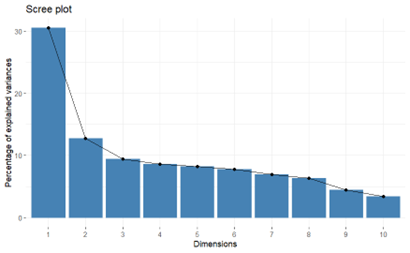
**PCA:**

From the Scree plot, it is evident that the first two principal components capture most of the variation. (STHDA, n.d.)

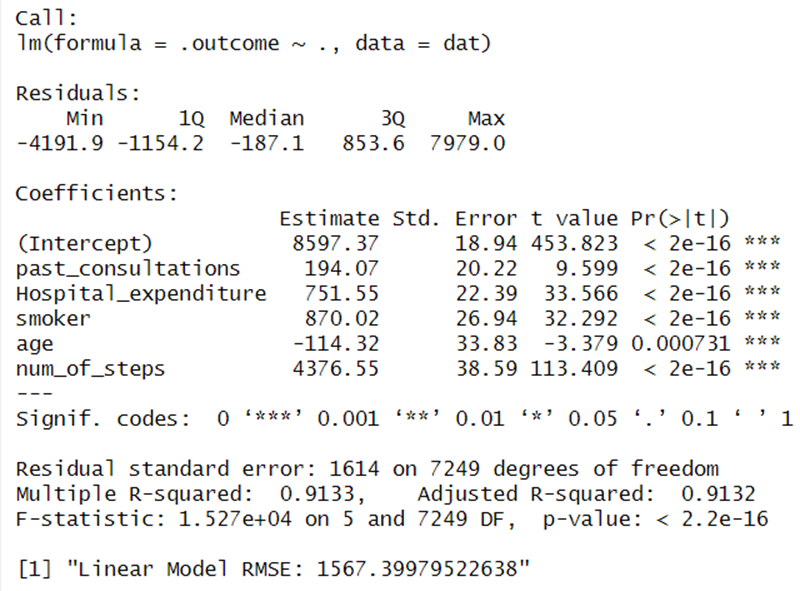
From the PCA Biplot, the selected features are past consultations, Hospital expenditure, smoker, age, and num\_of\_steps from underlying assumptions

**A graph with text on it

Description automatically generated**

****

**5.3 Predictive Analytics: Model Building**

 **5.3.1 Linear Regression Model**(Scribbr, 2022)

**Key Findings:**

The model accounts for a very high proportion of variance in the outcome variable, with an adjusted R-squared of 0.9132. This dictates how strong the relationship between independent variables and the outcomes is.

All the independent variables—past consultations, hospital expenditure, smoker status, age, and number of steps—have a statistically significant, or p-value less than 0.05, relationship with the outcome variable.

The intercept (8597.37) is what the model predicts when all of its independents are zero.

The RMSE for the model is 1567.40, and this is the average difference between the observed and fitted outcome values.

**Model Coefficients:**

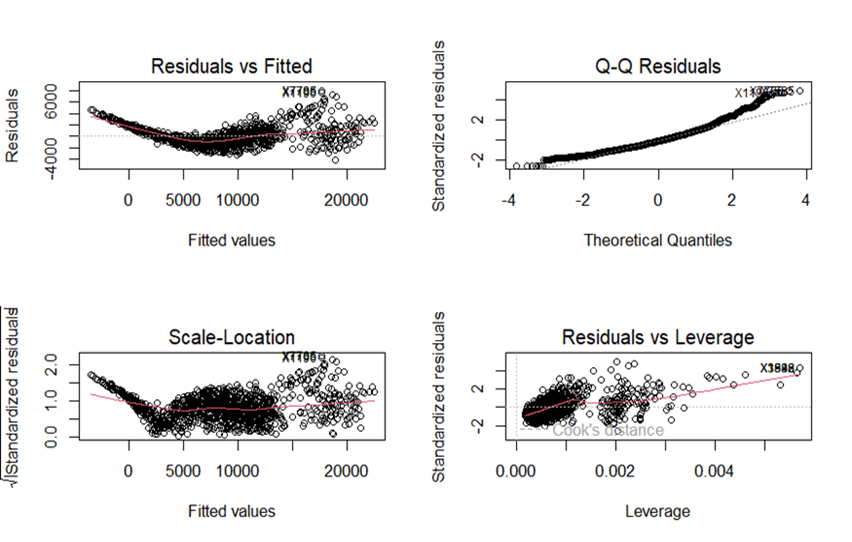
It is seen that, for each of the independents, the coefficients reflect their signs and strengths of relationship with the dependent or outcome variable. For example, "past\_consultations" increases the outcome variable by 194.07 units when it increases one unit while holding other variables in the model constant. The coefficient for age is negative, indicative of a decrease in the outcome with increasing age.

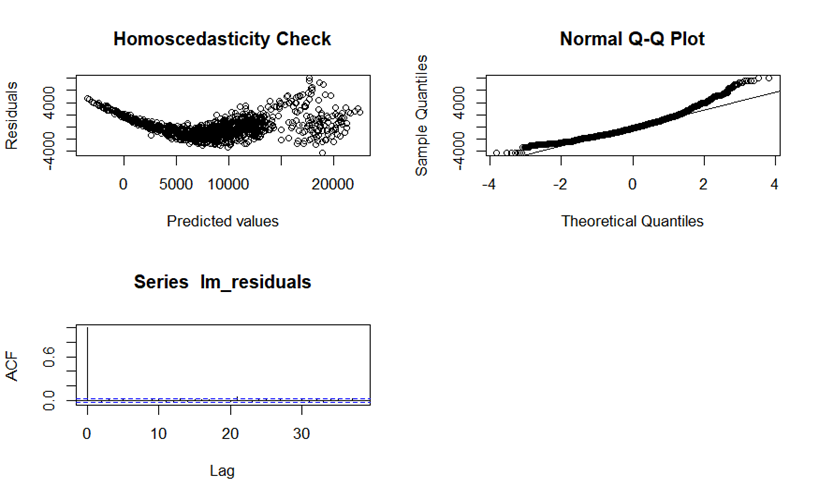
**Residuals:**

The residuals are the differences between the real values of outcomes and the values predicted by the model. It gives a summary of the minimum, quartiles, and maximum values of the residuals.

From this model, one can deduce that there is a strong relationship—a statistically significant relationship—between the independent variables and the outcome variable.

Further analysis is justified, such as residual plot inspections, checking model assumptions, and looking for potential outliers.

**Model Assumptions:**

****

**1. Residuals vs Fitted**

Ideally, the residuals should be randomly scattered around the horizontal line (residuals = 0) with no clear pattern.

In the plot, the residuals display a curved pattern, which suggests that the linearity assumption may be violated. This indicates that the relationship between the predictors and the response variable may not be purely linear.

**2. Normal Q-Q**

Ideally, the residuals should fall along the 45-degree reference line.

In the plot, most residuals fall along the line, but there are deviations at the tails (particularly at the upper end), indicating some departure from normality. This suggests that the residuals may not be perfectly normally distributed, but moderate deviations are usually acceptable.

**3. Scale-Location (or Spread-Location)**

In the plot, the spread of the standardized residuals increases with the fitted values, suggesting heteroscedasticity (non-constant variance). This means that the variance of the residuals increases as the fitted values increase, which violates the homoscedasticity assumption.

**4. Residuals vs Leverage**

Points that are far from others horizontally (high leverage) and have large residuals vertically (outliers) can be influential.

In the plot, a few points are highlighted (e.g., point labeled X38880) with high leverage or high Cook's distance, indicating they are potentially influential. Such points can have a significant impact on the regression model and should be investigated further.

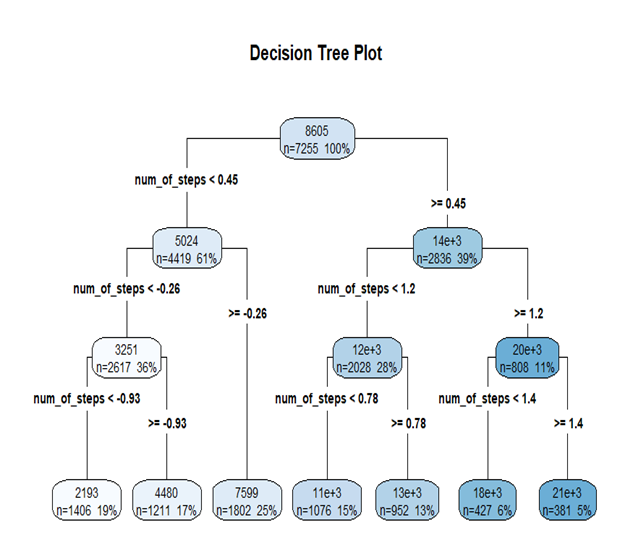
**Homoscedasticity**

In this plot, it appears there is a slight trend upwards as the predicted values increase. This suggests that the variance of the residuals might be increasing with increasing predicted values, violating the assumption of homoscedasticity.

**Normality of residuals**

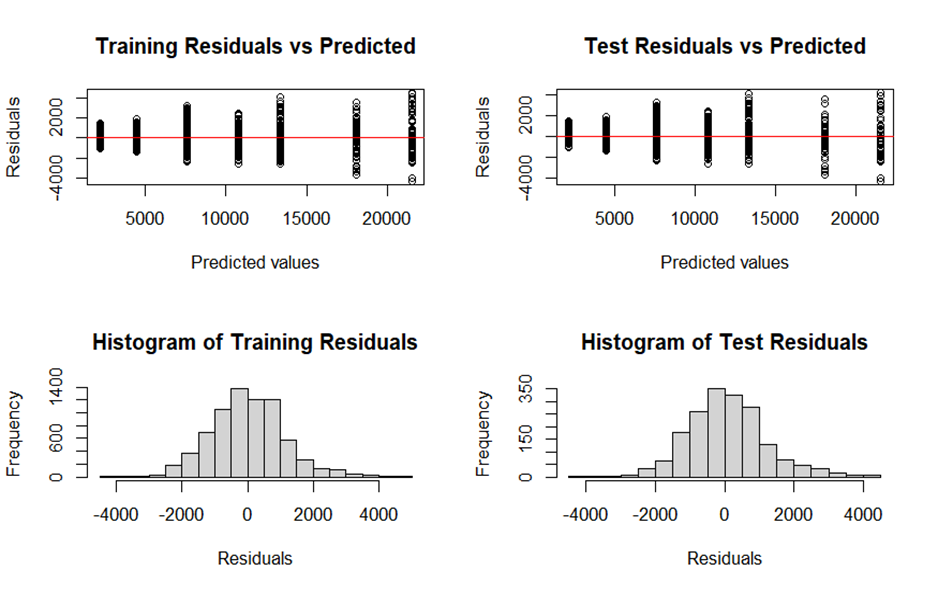
If the residuals are normally distributed, the points should fall approximately along a straight diagonal line.

In this plot, the points deviate from the straight line, particularly for the tails of the distribution. This suggests that the residuals may not be normally distributed.

* + 1. **Decision Tree Model**

**Key Findings:**

It is seen from the decision tree analysis that steps are the most critical variable predicting the outcome, followed by hospital expenditure, smoking status, age, and past consultations. The tree node is from the mean 8605.064 and MSE 2.999775e+07, splitting based on the steps that affect the dependent variable to a great extent. The following splits are very consistent, and the lower the means constantly go along with fewer steps, lower hospital expenditure, and non-smoking status. Age and past consultations play a role to a very minor extent. Tree's nodes show at the end, varying means and MSEs, an indication of heterogeneity in data. Especially, the analysis indicates that increased steps and costs mean better outcomes, while higher steps and costs have higher means indicating poorer outcomes.

**Model Assumptions:**

* 1. **Model Evaluation:**

**5.4.1 Linear Regression**

A screenshot of a computer code

Description automatically generated

**Interpretation:**

The RMSE is 1614.438. This is an average prediction error on the cross-validation sets. The lower the RMSE, the better the model's performance; the RMSE of 1614.438 result says that the model predictions deviate from the actual values by an average of 1614.438 units according to the target variable, insurance charges. The R-squared is 0.9135376.

It quantifies the proportion of variance in the dependent variable predictable from the independent variables. The value taken by it follows: An R-squared of 0.9135376 indicates that approximately 91.35 percent of insurance charges variation is explained by this model.

MAE: 1244.023

It is the average of the absolute errors between predicted and actual values. An MAE of 1244.023 means, for instance, that the model on average predicts 1244.023 units of the target variable off.

**Test Set Results:**

RMSE: 1567.4

A little lower than the cross-validation RMSE: good generalization to new data.

R-squared: 0.9172124

The test set R-squared is quite close to the cross-validation R-squared, so it explains a bit more of the variance in the test set.

* + 1. A screenshot of a computer

       Description automatically generated**Decision Tree**

RMSE: 1995.027

This is the average prediction error using this cp value. An RMSE of 1995.027 means that the model's predictions are off from the real values by an average of 1995.027 units of the target variable unit is insurance charges.

R-squared: 0.8637927

An R-squared of 0.8637927 tells that about 86.38 % of the variance in the insurance charges is explained by the model.

MAE: 1652.353

An MAE of 1652.353 indicates that on average, the predictions made by the model are off by 1652.353 units of the target variable.

Conclusions for other cp values:

cp = 0.15723563

RMSE: 2620.251

R-squared: 0.7641581

MAE: 2169.359

cp = 0.66600622

RMSE: 4556.487

R-squared: 0.6604994

MAE: 3705.757

Optimal Tuning Parameter:

The final model used a value for cp of 0.09267839, which gave the lowest RMSE.

Results on the Test Set:

RMSE: 2362.393

It can be seen that the test set RMSE is higher than the cross-validation RMSE, which may indicate that this model does not generalize as well to new data. The R-squared in this case is 0.8115849. One can easily see that the test set R-squared is lower than the cross-validation R-squared, indicating the model explains a smaller proportion of variance in the test set.

1. **Conclusion**

* From the model evaluation and cross-validation results, the linear regression model was found to be more accurate than the decision tree model with low RMSE and high R-squared value.
* The project successfully demonstrates the application of regression analysis techniques to predict health insurance charges, providing valuable insights for the insurance industry. Linear Regression and Decision Tree Regression models showed strong predictive capabilities, with the Linear Regression model achieving an adjusted R-squared of 0.9132. However, further refinement and validation are needed to ensure the models' robustness and generalizability.

**6.1 Recommendations**

1. Model Refinement: An in-depth refinement of the model, including more features and possibly complex algorithms, can make use of ensemble methods to ensure more accurate predictions.

2. Regular Updates: Continuously update models due to data changes so that they remain relevant and accurate; the accuracy increases with time.

3. Ethical Practices: Ensure preservation of data privacy along with fairness and transparency in model development and deployment.

4. Stakeholder Engagement: Engage stakeholders in the process in order to line up models with business goals and ethical standards.

5. Full Validation: The models shall be subject to detailed validation and stress-testing to ensure that they work well in every respect and are not biased.

1. **Ethical Considerations**

Data Protection: Ensure the application of strict measures on data handling with the highest confidentiality concerning a person's information and in full compliance with the data protection regulations like HIPAA.

Fairness and Bias: These models were scrutinized for identification so that biases do not perpetuate already existing imbalances in access and costs of health care, which could have been generated during the modeling process.

Transparency: The project kept lucid the application of predictive models and various factors taken into consideration in building them. How this data base influenced their insurance charges was made open to policyholders.

Informed Consent: Maintained awareness among all that they had consented to the use of their data in building the predictive models, done with the highest ethical standards.

1. **References**

Ali, Z. (n.d.). Insurance data with 10,000 records [Data set]. Kaggle. <https://www.kaggle.com/datasets/zulqarnainalipk/insurace-data-with-10000-records/data>

STHDA. (n.d.). Principal component analysis in R: prcomp vs princomp. Retrieved from <http://www.sthda.com/english/articles/31-principal-component-methods-in-r-practical-guide/118-principal-component-analysis-in-r-prcomp-vs-princomp/#google_vignette>

Scribbr. (2022, June 23). Linear regression in R. <https://www.scribbr.com/statistics/linear-regression-in-r/>