STAT 515 – 001

Applied Statistics & Visualization for Analytics

Final Project Report

George Mason University

Prof. Dr. Tokunbo Fadahunsi, PhD.

Group-10 Project Members

Utkarsh Desai
Pallavi Vegesana
Vyoma H. Podapati

Project Title: Health Insurance Data Visualization and Analysis

1. INTRODUCTION

Understanding the nuances of health insurance is pivotal in contemporary healthcare research and policymaking. As the United States grapples with ongoing debates and reforms in its healthcare system, delving into the factors influencing insurance charges becomes imperative. This dataset serves as a microcosm, reflecting the diverse demographic and lifestyle factors that contribute to the complex landscape of health insurance costs. Exploring the interplay between variables like age, BMI, smoking habits, and regional disparities can unravel patterns that inform insurers, policymakers, and healthcare providers about the dynamics of risk assessment, pricing strategies, and the overall sustainability of the health insurance market.

The dataset at hand comprises essential information derived from the realm of US health insurance, providing a comprehensive snapshot of individuals and their associated attributes in the context of healthcare coverage. This dataset encapsulates key variables such as age, sex, body mass index (BMI), number of children, smoking habits, region, and corresponding insurance charges. Analysing this dataset offers valuable insights into the intricate dynamics of health insurance in the United States, shedding light on patterns, correlations, and factors that influence the pricing and distribution of insurance charges.

2. DATASET DESCRIPTION

This health insurance dataset provides a comprehensive overview of individuals' demographic and health-related attributes, coupled with corresponding insurance charges. The dataset encompasses information on a diverse set of individuals, including details such as age, sex, body mass index (BMI), number of children, smoking habits, region of residence, and the associated insurance charges. It provides a real-world glimpse into how factors such as age, lifestyle choices, and geographical location impact the cost of health insurance coverage. The inclusion of diverse variables allows for a nuanced analysis, offering insights that can be applicable to both individual policyholders and the broader health insurance industry.

	A	B	C	D	E	F	G
1	age	sex	bmi	children	smoker	region	charges
2	19	female	27.9	0	yes	southwest	16884.92
3	18	male	33.77	1	no	southeast	1725.552
4	28	male	33	3	no	southeast	4449.462
5	33	male	22.705	0	no	northwest	21984.47
6	32	male	28.88	0	no	northwest	3866.855
7	31	female	25.74	0	no	southeast	3756.622
8	46	female	33.44	1	no	southeast	8240.59
9	37	female	27.74	3	no	northwest	7281.506
10	37	male	29.83	2	no	northeast	6406.411
11	60	female	25.84	0	no	northwest	28923.14
12	25	male	26.22	0	no	northeast	2721.321
13	62	female	26.29	0	yes	southeast	27808.73
14	23	male	34.4	0	no	southwest	1826.843
15	56	female	39.82	0	no	southeast	11090.72
16	27	male	42.13	0	yes	southeast	39611.76
17	19	male	24.6	1	no	southwest	1837.237
18	52	female	30.78	1	no	northeast	10797.34
19	23	male	23.845	0	no	northeast	2395.172
20	56	male	40.3	0	no	southwest	10602.39

Figure 1: Dataset

3. RESEARCH QUESTION

Our interest in exploring this dataset is fuelled by the practical relevance it holds for us as students at GMU. Notably, we have observed a substantial increase of $700 in our insurance expenses compared to the previous semester. Motivated by this real-world impact on our finances, we aim to conduct a

detailed analysis of various factors influencing insurance policies. Through this examination, we hope to uncover insights that not only address our immediate concerns but also contribute valuable knowledge to the broader discourse on the determinants of insurance costs.

- How does insurance charges vary with respect to BMI, and number of children?
- Is there a significant difference in insurance charges between smokers and non-smokers?
- How does insurance charges vary across different regions?
- Which is the best fitting model for our dataset? Which variables are the most impactful on the insurance charges?

4. EXPLORATORY DATA ANALYSIS

The primary goal of this project is to unravel the intricate relationships and patterns within the health insurance dataset, employing visualizations and predictive models to discern the impact of demographic and lifestyle factors on insurance charges. By leveraging data-driven insights, the project aims to contribute valuable information for informed decision-making in the context of health insurance, with a focus on understanding and potentially mitigating the factors influencing insurance costs.

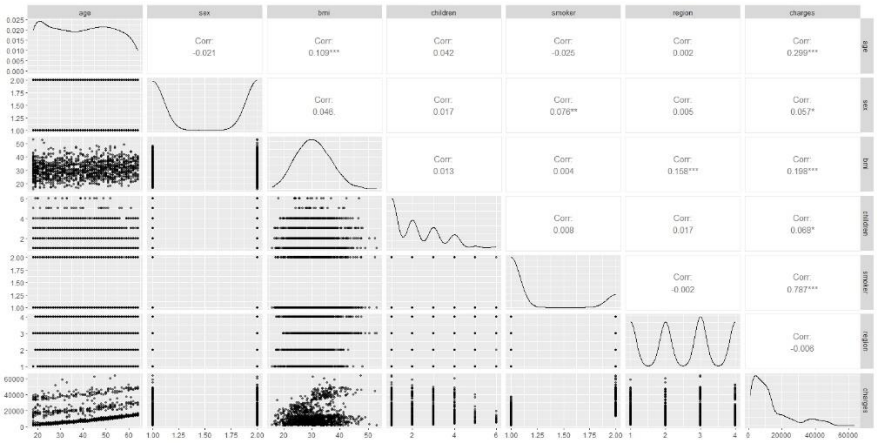

Figure 2: Parametric Analysis

Through an exploratory data analysis (EDA), we have gained a comprehensive understanding of the dataset's structure, integrity, diversity, and the relationships between its variables. This analytical exploration has allowed us to pinpoint crucial aspects of the data, identifying key features that play a significant role in shaping its characteristics.

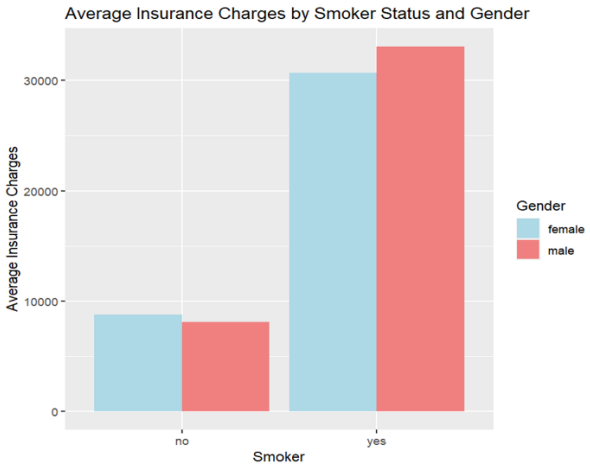

Figure 3: Bar plot for Insurance Charges and Smoker Status

This bar chart illustrates the average insurance charges (Y-axis) categorized by smoker status (X-axis) for both men and women in the dataset. Each bar represents the mean insurance charges for individuals classified as smokers and non-smokers. The chart aims to provide a visual comparison of the average costs associated with smoking and non-smoking individuals, offering insights into the potential impact of smoking habits on health insurance charges for both genders.

Our observations reveal that, within the dataset, insurance charges are lower for non-smoking men compared to non-smoking women. Conversely, for individuals who smoke, the insurance charges are higher for men than for women.

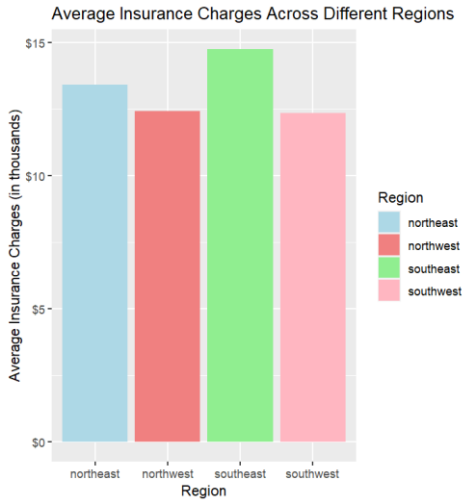

Figure 4: Insurance Charges across Region

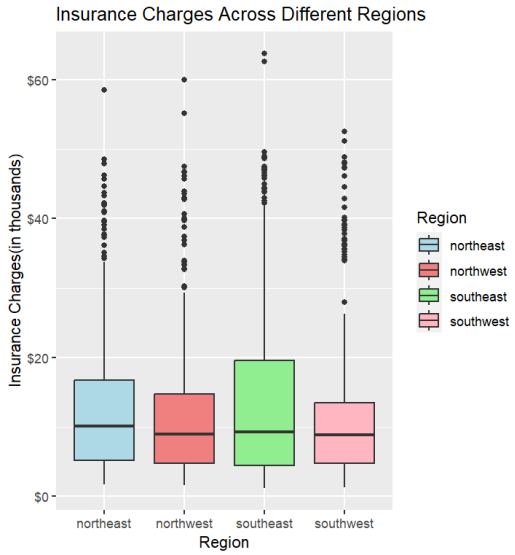

Figure 5: Box Plot for Insurance Charges by Regions

This bar and box plot depicts the average insurance charges across distinct regions, including Southwest, Southeast, Northwest, and Northeast. The Y-axis represents the average insurance charges, and each bar corresponds to a specific region on the X-axis. The chart offers a visual comparison of the mean insurance costs for individuals residing in different geographical areas, providing insights into potential regional variations in health insurance charges.

The visual analysis indicates that individuals in the Southeast region consistently incur the highest insurance charges, with the Northeast region closely trailing behind. This observation suggests the possibility of elevated healthcare expenses in the Southeast, contributing to the higher insurance premiums. Furthermore, it implies that residents in the Southeast and Northeast may experience more diverse health conditions, leading to a wider range of insurance charges. In contrast, the Northwest and Southwest regions exhibit a more limited range in insurance charges, implying a potential homogeneity in healthcare costs within these areas.

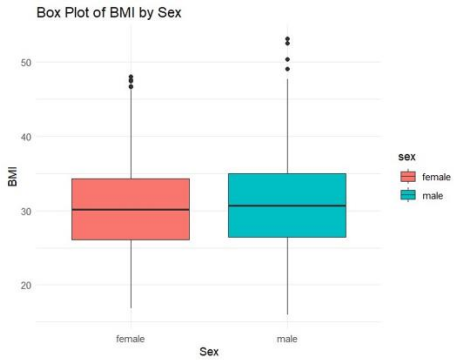

Figure 6: Box Plot for Body Mass Index by Sex

This box plot visually represents the distribution of Body Mass Index (BMI) categorized by gender (male and female). The vertical axis displays the BMI values, while the horizontal axis distinguishes between the two genders. The box plot provides a clear depiction of the central tendency, spread, and potential outliers in BMI for both males and females. Through this visualization, insights into the comparative distribution of BMI between the two genders can be gained, offering a nuanced understanding of body mass patterns within the dataset.

The box plot reveals that the majority of individuals, regardless of gender, have a Body Mass Index (BMI) outside the recommended ideal range of 18 to 25. This observation underscores that a significant portion of the population in the dataset exhibits BMI values that deviate from the commonly considered healthy range.

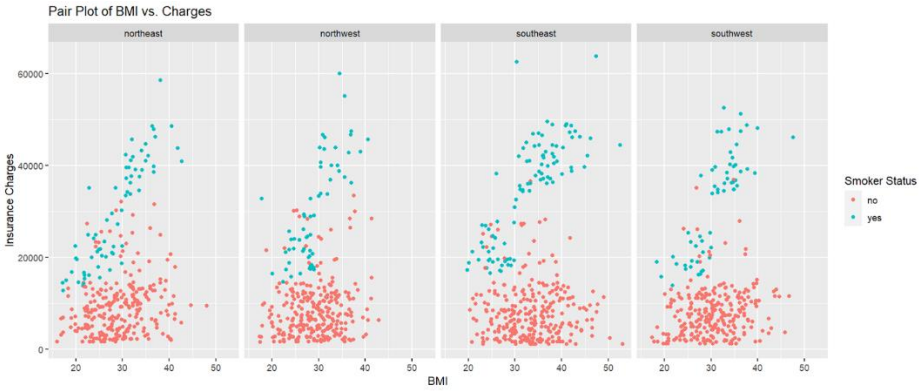

Figure 7: Pair Plot for Body Mass Index by Insurance Charges

This pair plot illustrates the relationship between Body Mass Index (BMI) and Insurance Charges, with data points differentiated based on smoker status. Each point on the plot represents an individual, with BMI values on one axis and corresponding insurance charges on the other. The plot points are color-coded to distinguish between smokers and non-smokers. Through this visualization, we gain insights into how BMI and smoking habits collectively influence insurance charges, allowing for a comprehensive understanding of the interplay between these key variables in the dataset.

In the pair plot analysis, it becomes evident that individuals maintaining a BMI within the recommended range tend to have lower insurance charges. Notably, smokers within the BMI range also exhibit comparatively lower insurance costs. However, the most significant increase in insurance charges is observed among smokers who fall outside the BMI range. This highlights a notable correlation between smoking habits, BMI, and the resulting impact on insurance premiums, emphasizing the potential compounding effect of these factors on healthcare costs.

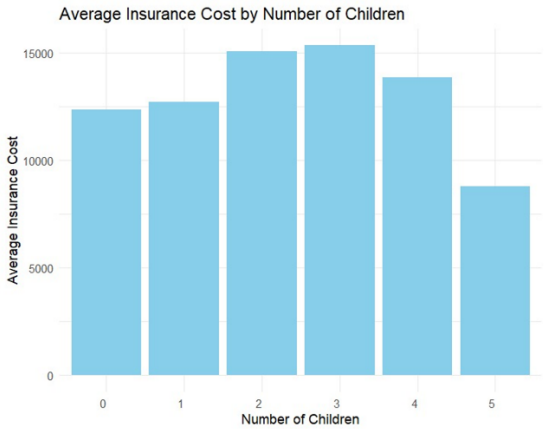

Figure 8: Insurance Charges Across number of Children

This bar plot illustrates the average insurance charges corresponding to different numbers of children. The X-axis represents the count of children, ranging from individuals with no children to those with multiple children. The Y-axis displays the average insurance charges associated with each category. The chart provides a visual representation of how the number of children influences the average cost of insurance. Through this visualization, one can discern any patterns or trends in insurance charges relative to the number of children, offering insights into the potential impact of family size on healthcare expenses.

This plot offers intriguing insights, challenging the intuitive expectation that insurance rates would uniformly rise with an increasing number of children. Surprisingly, the visualization reveals a peak in insurance prices for families with 2 or 3 children, while the rates decrease for families with 5 children. One plausible explanation could be that insurance providers strategically set higher prices for families with 2 or 3 children, considering it a common family size. Conversely, families with no children or only one child show minimal differences in insurance rates. This observation prompts a reconsideration of assumptions about the linear relationship between family size and insurance charges, suggesting a more nuanced dynamic at play.

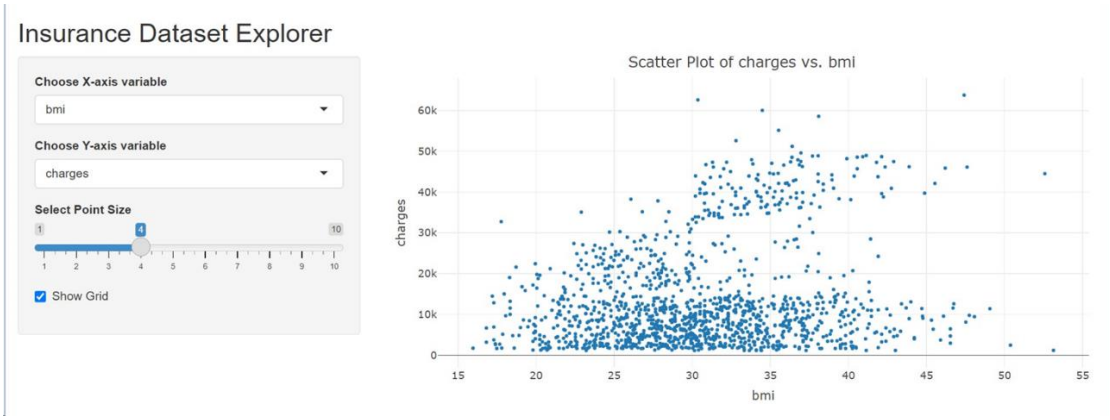

Figure 9: Interactive Plot to display relationships between various parameters

The interactive scatter plot, crafted using Plotly and Shiny functions, encompasses a holistic exploration of the entire health insurance dataset. Each data point on the plot represents an individual, with the scatter plot incorporating all parameters from the dataset, including age, sex, BMI, number of children, smoker status, region, and insurance charges. Users can interactively engage with the plot, hovering over data points to reveal specific details, and dynamically adjusting parameters to observe their impact on insurance charges. This visualization provides a comprehensive and customizable view of the dataset, enabling a nuanced understanding of the relationships between various factors and their collective influence on health insurance costs.

5. STATISTICAL MODELLING FOR THE DATASET

Linear Regression Model

Summary Statistics:

Using these statistics, we can say that the model is a good fit and it also showcases some of the important parameters and their coefficients, this could be used to determine predicted changes in Insurance Charges based on unit change for various parameters while keeping the others constant.

```
> model_summ

Call:
lm(formula = log(charges) ~ ., data = df)

Residuals:
     Min       1Q   Median       3Q      Max
-1.09068 -0.19752 -0.04952  0.06012  2.15598

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.5147310  0.0877320  62.859  < 2e-16 ***
age          0.0346556  0.0008726  39.717  < 2e-16 ***
sex         -0.0753414  0.0244326  -3.084  0.00209 **
bmi          0.0122826  0.0020352   6.035 2.06e-09 ***
children     0.1023950  0.0101049  10.133  < 2e-16 ***
smoker       1.5498384  0.0302346  51.260  < 2e-16 ***
region      -0.0476067  0.0111534  -4.268 2.11e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4449 on 1331 degrees of freedom
Multiple R-squared:  0.767,     Adjusted R-squared:  0.7659
F-statistic: 730.2 on 6 and 1331 DF,  p-value: < 2.2e-16
```

Figure 10: Summary Statistics of Linear Regression Model

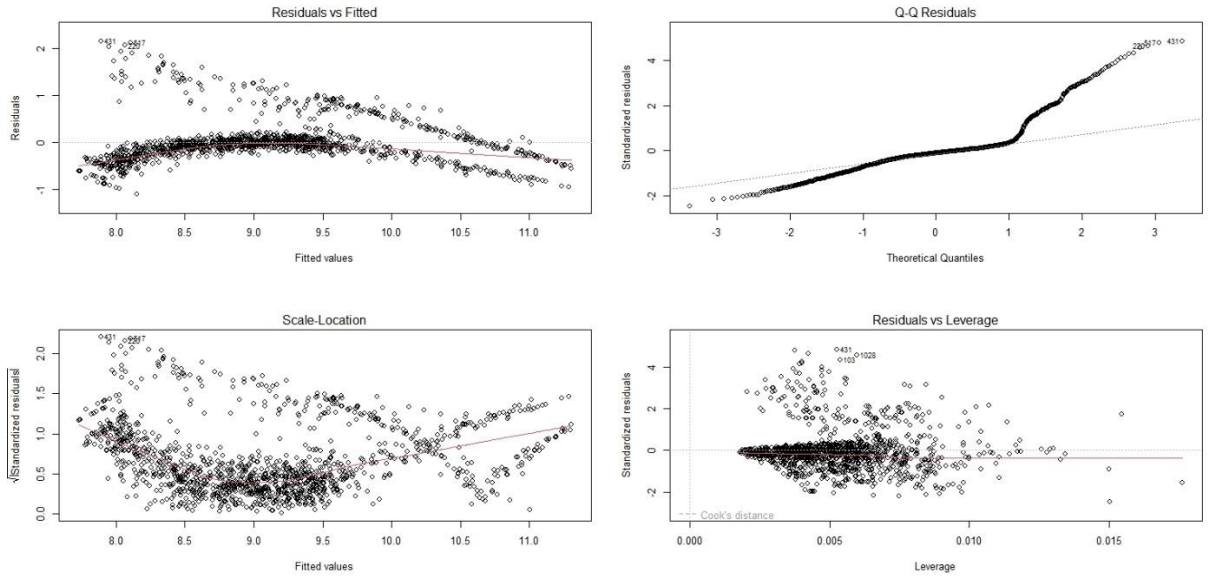

Figure 11: Plots for Linear Model

Our Linear Regression model is meeting the linearity assumption, but is not able to satisfy the normality assumption and equal variance assumption(heteroscedastic).

Best Subset Selection

```
> regfit_full <- regsubsets(log(charges)~., df)
> summary(regfit_full)
Subset selection object
Call: regsubsets.formula(log(charges) ~ ., df)
6 Variables  (and intercept)
         Forced in Forced out
age          FALSE      FALSE
sex          FALSE      FALSE
bmi          FALSE      FALSE
children     FALSE      FALSE
smoker       FALSE      FALSE
region       FALSE      FALSE
1 subsets of each size up to 6
Selection Algorithm: exhaustive
         age sex bmi children smoker region
1  ( 1 ) " " " " " " " "      "*"    " "
2  ( 1 ) "*" " " " " " "      "*"    " "
3  ( 1 ) "*" " " " " "*"      "*"    " "
4  ( 1 ) "*" " " "*" "*"      "*"    " "
5  ( 1 ) "*" " " "*" "*"      "*"    "*"
6  ( 1 ) "*" "*" "*" "*"      "*"    "*"
>
```

Figure 11: Best Subset Selection

By employing this approach, we can ascertain the selection of parameters suitable for constructing single linear regression, two-variable multiple linear regression, three-variable, four-variable, and so forth. This method not only aids in the identification of pertinent variables but also facilitates a comprehension of the hierarchical influence of these parameters on the target variable.

Insights from the Linear Regression Model:

Applying a linear regression model to the health insurance dataset can provide insights into the quantitative relationships between various factors and insurance charges. For example, it can reveal the average increase in charges associated with each additional year of age, the impact of BMI on charges, and whether smoking status significantly contributes to higher insurance costs. This model can be a valuable tool for understanding the linear associations within the dataset, providing a basis for predictions and aiding in the identification of influential factors that contribute to the variability in health insurance charges.

Decision Tree (Regression Tree) Model

The decision tree output serves as an insightful tool for understanding the decision rules employed by the model in predicting insurance charges. It facilitates interpretability and transparency in model decision-making, offering valuable insights into the factors influencing the target variable. The pruning process contributes to a more generalized and robust model.

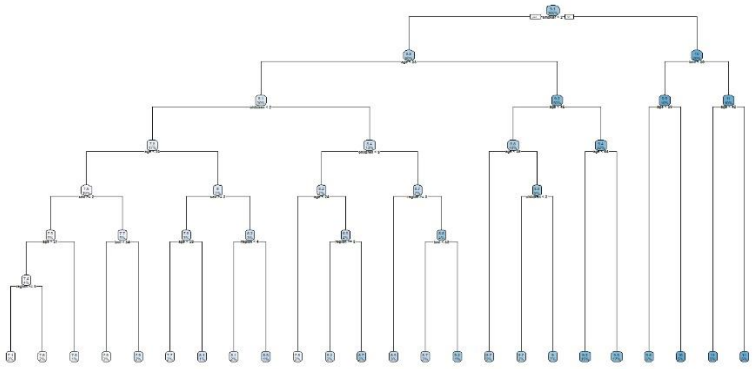

Figure 12: Decision Tree

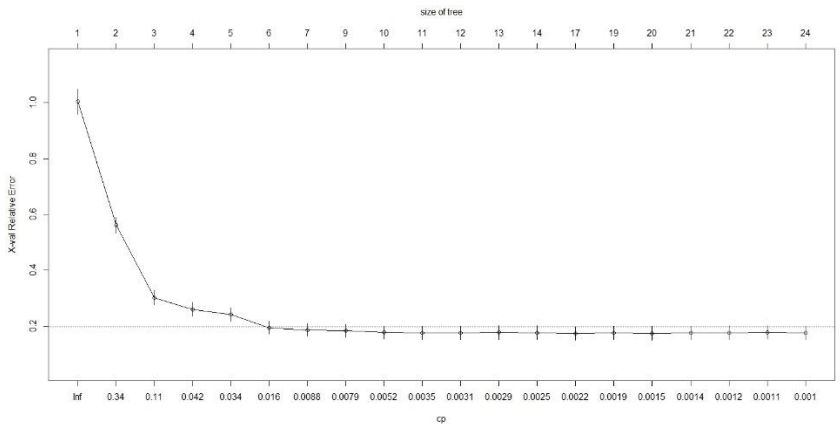

Figure 13: Complexity Parameter Analysis

Here we've created complexity parameter analysis to decide the best CP value which helps us identify and create a pruned decision tree with least error rate.

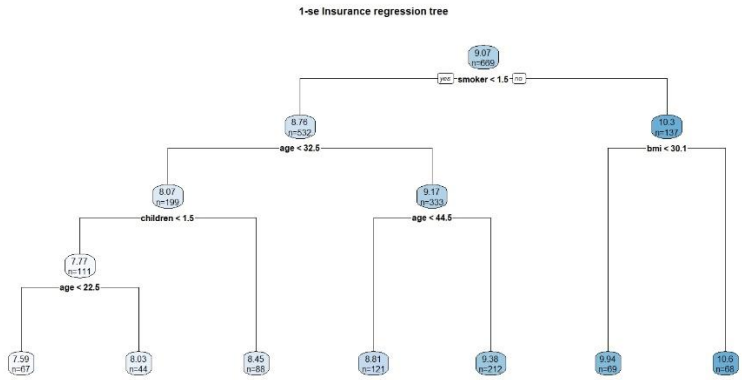

Figure 14: Pruned Decision Tree

Random Forest Regression Model

In the Random Forest regression analysis of this model, the outcome provides a comprehensive understanding of the significance and impact of each parameter on the target variable. The model leverages the collective decision-making of multiple decision trees, offering insights into complex relationships, interactions, and non-linearities within the dataset. The output includes variable importance metrics, allowing for the identification of key features that contribute significantly to the variability in the target variable.

```
> rf.ins

Call:
 randomForest(formula = log(charges) ~ ., data = df, ntree = 100,
E)
               Type of random forest: regression
                     Number of trees: 100
No. of variables tried at each split: 2

          Mean of squared residuals: 0.1522497
                    % Var explained: 81.98
>
```

Figure 15: Summary for Random Forest Model

```
> importance(rf.ins)
           %IncMSE IncNodePurity
age      44.1649057    344.370520
sex       0.3694383      9.701133
bmi      11.9012638     88.088618
children 14.5715225     46.728813
smoker   75.2559265    477.287916
region    5.9371074     19.367817
>
```

Figure 16: Impact of Parameters

In our analysis, we employed Random Forest Regression to assess the significance of each parameter in influencing the target variable. By integrating this information with insights from the Linear Regression Model, we derived a confident and conclusive hierarchy, ranking each parameter based on its impact.

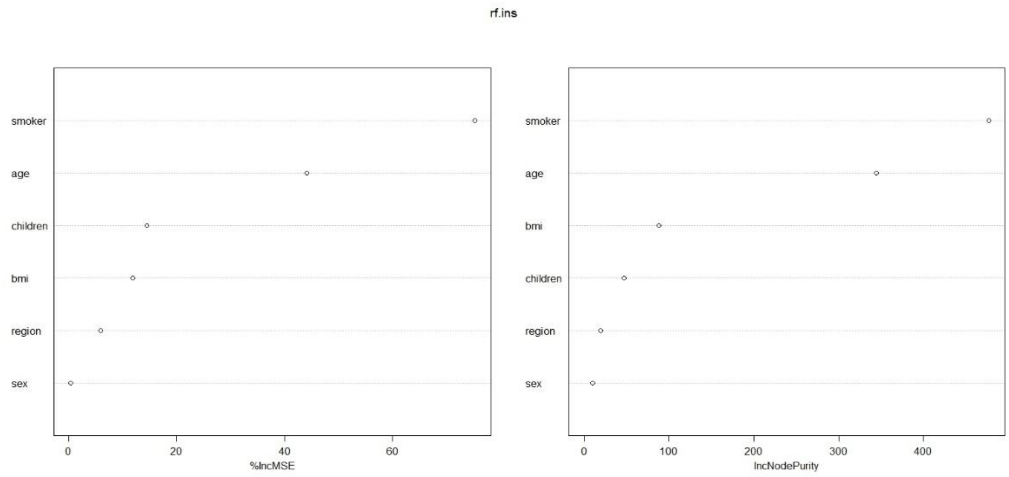

Figure 17: Random Forest Model

Mean Squared Error Statistics

We were able to see a very low error rate for both models but as the LR model is more interpretable and could be easily explained to a layman we recommend the LR model for this dataset.

MSE of Linear Regression		MSE of Random Forest	
MSE	0.196865	MSE	0.1522497
RMSE	0.44369	RMSE	0.3901919

TABLE 1: MEAN SQUARED ERRORS VALUES

6. FUTURE SCOPE

Here in this report, we discussed the best fitting model and the impact of various parameters on the target variable. Further we can use these models and try doing predictions using them. We can further dive into questions like why our linear model shows that the model fails the normality and equal variance assumptions and what modifications or calculations we can make to overcome them.

7. CONCLUSION

In conclusion, the comprehensive analysis of the health insurance dataset has provided valuable insights into the intricate dynamics influencing insurance charges. Through exploratory data analysis, visualizations, and the construction of a Linear Regression, Random Forest and decision tree model, we have unravelled patterns and relationships among demographic, lifestyle, and regional factors. The observed disparities in insurance charges across different regions, the nuanced impact of smoking habits, and the unexpected variations related to family size highlight the complexity of determining healthcare costs.

8. REFERENCES

[1] *US health insurance dataset*. (2020b, February 16). Kaggle. https://www.kaggle.com/datasets/teertha/ushealthinsurancedataset

[2] *An introduction to statistical learning*. (n.d.). An Introduction to Statistical Learning. https://www.statlearning.com/