Prediction of Insurance Charges

People's lives are centered around their health and happiness. However, since it's impossible to avoid all risks, the financial industry has developed various products to protect individuals and organizations from these risks using financial resources. One such product is insurance, which aims to reduce or eliminate the expenses associated with different types of risks.

Through this project, we are aiming to understand the combined interplay of smoking behavior, body mass index (BMI), and age on insurance charges in the Prediction of Insurance Charges dataset. By investigating these interaction effects, we can gain insights into how lifestyle choices and health-related attributes interact to determine insurance charges.

Research Question
How do age and various health-related factors interact to determine insurance charges?

Exploratory Data Analysis
Before analyzing the data, we cleaned any duplicate and null values that were present.

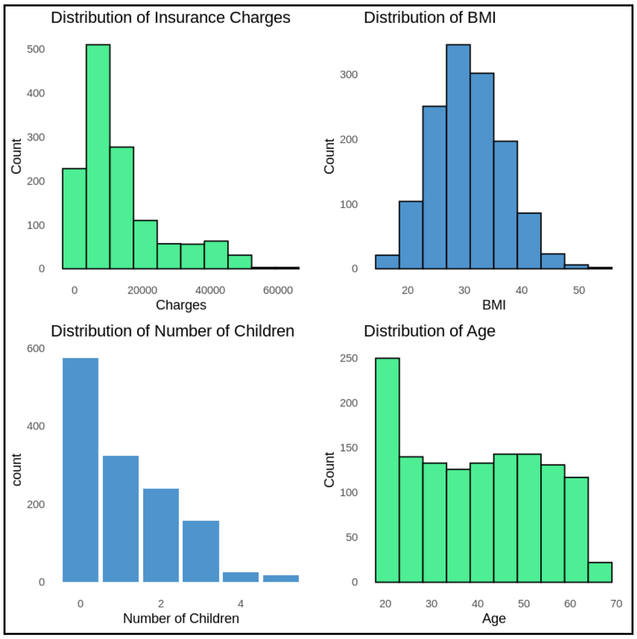

Figure 1: Distributions of Numeric Variables

- ➢ The distribution of insurance charge amount is right-skewed, with most of the charges ranging between $0 and $15,000.
- ➢ The distribution of patient BMI is relatively normally distributed with most patients' BMI ranging from 25 to 35. Most people in the dataset do not have children.

- There are a few hundred people that have 1, 2, or 3 children, and approximately 50 or less that have 4 or 5 children.
- There are more than 250 people from the age of 18 to 25 which forms the largest age group in the dataset.

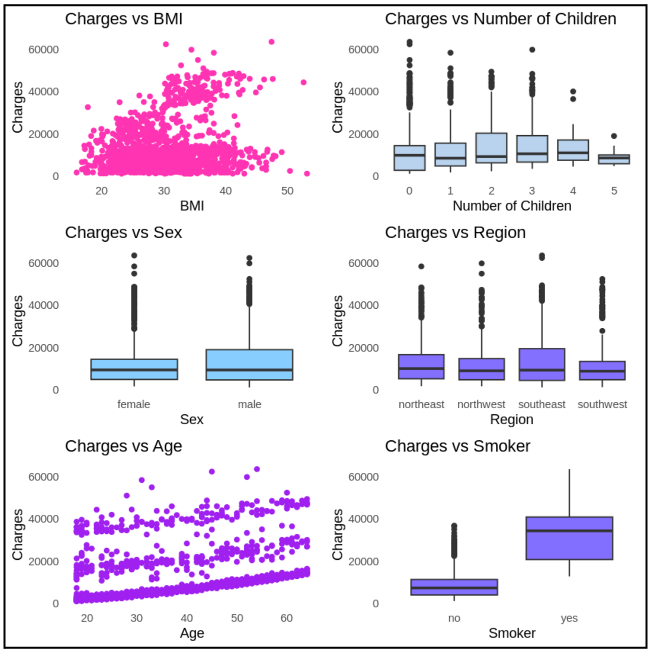

Figure 2: Response Variable vs Explanatory Variables

- We observe that there exists some correlation between Charges vs BMI and Charges vs Age. But, in both plots, the points are divided into groups. This could be due to the influence of another variable. We can understand this better with a multivariate visualization.
- There is a significant difference between the mean of charges for smokers and non-smokers.
- The regions evidently have less effect on charges as there are no notable differences between them.
- The average insurance charge for males is $13975 which is higher than that of women at $12569.58. This could be due to the influence of smoking behavior as the proportion of male smokers (23.5%) is more than that of females (17.4%).
- The number of children also seems to have a weak correlation with charges and there are a few outliers for the people with no children.

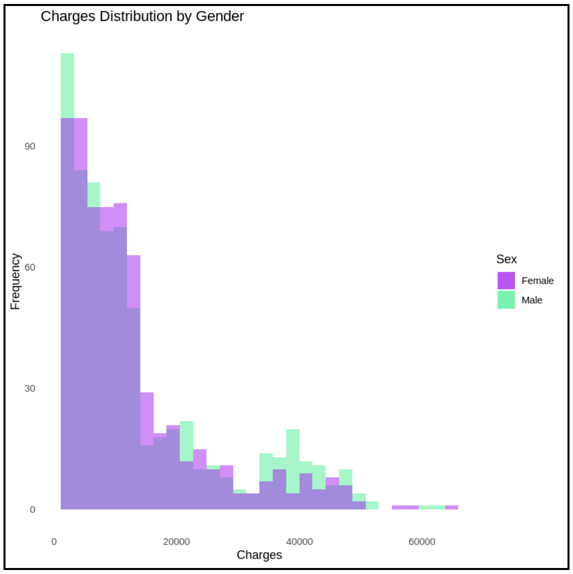

Figure 3: Charges Distribution by Gender

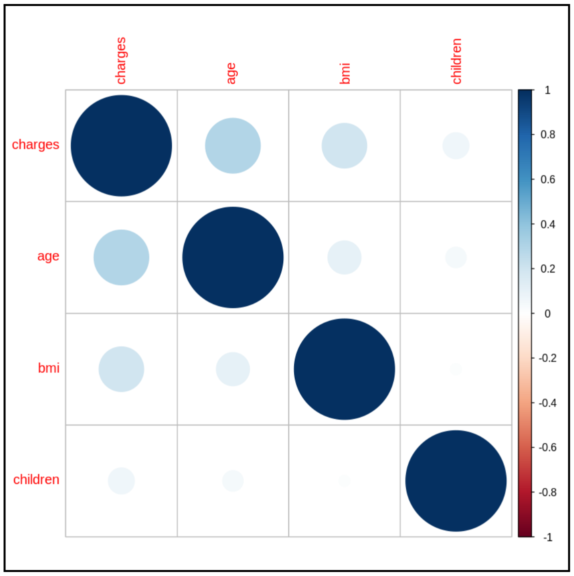

Figure 4: Correlation Matrix Visualization

Figure 3: Both male and female have a right skewed distribution for charges. There are a higher number of males towards the right side which means that they incur higher charges which could be caused due to aforementioned reasons like smoking. Age and BMI seem to have a low positive correlation with charges.

Figure 4: The correlation between the number of children and charges is almost negligible. So, we have 3 variables with positive correlation to charges, namely: smoker, age, BMI.

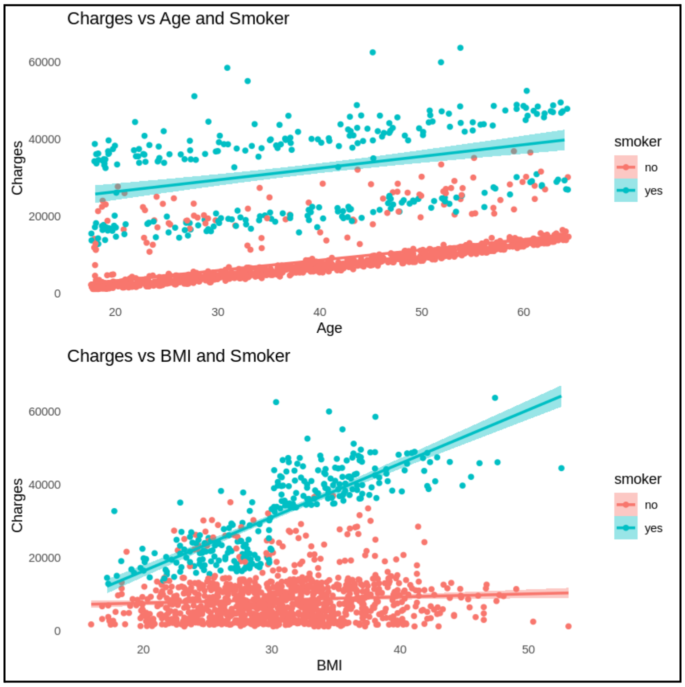

Figure 5: Multivariable Plots

- ➢ Visualizing the relationship between Charges vs Age/BMI and Smoker, we can better understand why there exist separate groups of points in the plots in Figure 2. The multivariate plots suggest a positive relationship of smoking behavior with age and BMI in context of charges.
- ➢ Evidently, using a model with interactions would provide us with a better fit as it would be able to capture the underlying patterns in the data.

Models Explored

We used linear regression for our predictions as it also provides a lens to interpret relationships between variables and quantify their impact.

★ **Additive Model with Health-Related Variables: Focusing on Age and Health**

This model reveals a more nuanced picture, emphasizing the substantial impact of these factors on insurance charges as it presented results similar to the model chosen with exhaustive selection.

★ **Model with Interactions: Capturing Complexity**

Recognizing that variables might not operate in isolation, the model unfolds the dynamic relationships between smoking habits, age, and BMI, shedding light on how these factors interact to influence insurance charges.

★ **Model with Interactions and Transformation: Unmasking Skewed Realities**

As we confront the challenge of skewed data in insurance charges, a transformation comes into play. The log transformation addresses the right-skewness shown in figure 3, producing a more normal distribution of charges.

★ **Polynomial Model: Navigating Non-Linearity**

Linear relationships might not always capture the intricacies of insurance charges. The model detects quadratic effects, thus taking care of potential nonlinearities that the prior models might overlook.

Diagnostics Table

Model	adj.r.squared	RMSE	AIC
Model 1	0.751431	6451.997	20234.03
Model 2	0.84335	5391.474	19774.31
Model 3	0.8153159	18618.09	993.2015

Model 4	0.8444047	5321.946	19769.54

Caveats

- ➢ The presence of individuals with exceptionally high charges may distort results, especially in models without robust transformations.
- ➢ Insurance pricing is a multifaceted domain influenced not only by health-related variables but also by a myriad of social factors not present in our data.
- ➢ Normality of residuals does not hold across the models although some of the residual histograms show that normality is followed to some degree. Using the shapiro-wilk test we found that none of the models have normal residuals.

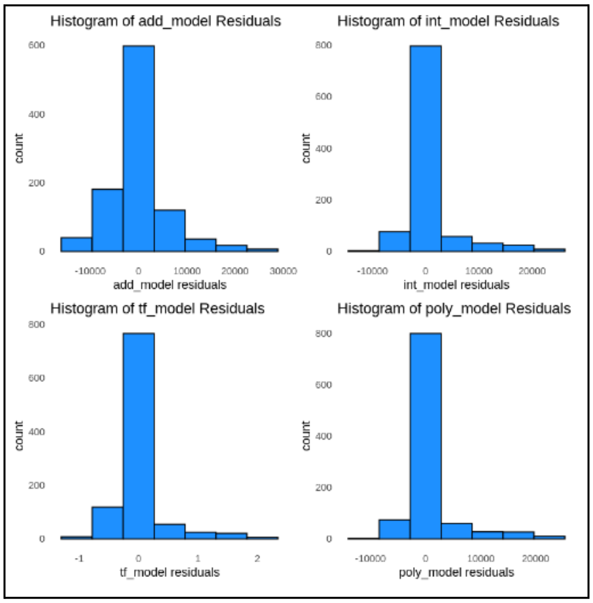

Figure 6: Histogram of Residuals

Interpretation and Conclusion

The model incorporating interactions between health variables demonstrated a better performance compared to the additive model. This enhancement allowed us to capture nuanced relationships and dependencies among health variables, resulting in more accurate predictions and a better fit to the data.

Analyzing the diagnostic table, we observe that while a log transformation might improve the fit of the model by better capturing the linear relationship between variables, it does not guarantee better predictive accuracy in all cases. The RMSE came out to be the highest compared to the rest of the models. This could be because of the effects of outliers or influential points that are extremely far from the rest of the data.

Through rigorous analysis and model comparison, we have not only uncovered a model with strong explanatory power but also emphasized the importance of holistically considering the interaction of health-related attributes in shaping insurance premiums.