STAT 306 Group Project Report

Group B6

Introduction

Source of the Data:

Prediction of Insurance Charges
https://www.kaggle.com/datasets/thedevastator/prediction-of-insurance-charges-using-age-gender

About the dataset:

This data has been gathered from a variety of sources and contains information such as age, sex, region, smoking status, number of children, and BMI values for each customer.

Response Variable:
The insurance charges for the customer. (Double)

Explanatory Variables:

Variable	Description
Age	The age of the customer. (Integer)
Children	The number of children the customer has. (Integer)
Sex	The gender of the customer. (Character)
Smoker	Whether or not the customer is a smoker. (Character)
Region	The region the customer lives in. (Character)
BMI	The BMI of the customer. (Double)

Age is expectedly one of the most important variables as younger customers are far less likely than an older person to suffer serious health issues, thus bearing lesser insurance charges. Similarly, sex is also potentially influential as traditionally gender roles have dictated premiums with men paying more than women for the same coverage on many policies. Lastly, BMI and smoker status should also be taken into account when making any predictions regarding insurance costs due to health risk factors associated with obesity and smoking being considered by premium pricing decisions made by insurers.

Research Motivation:

People's lives are centered around their health and happiness. However, sinceit'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. The cost of insurance charges varies from person to person since various factors influence the cost of an insurance plan.
Through this project, we are aiming to understand the combined influence of smoking behavior with body mass index (BMI) and age on insurance charges. 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?

Data Visualization and Analysis

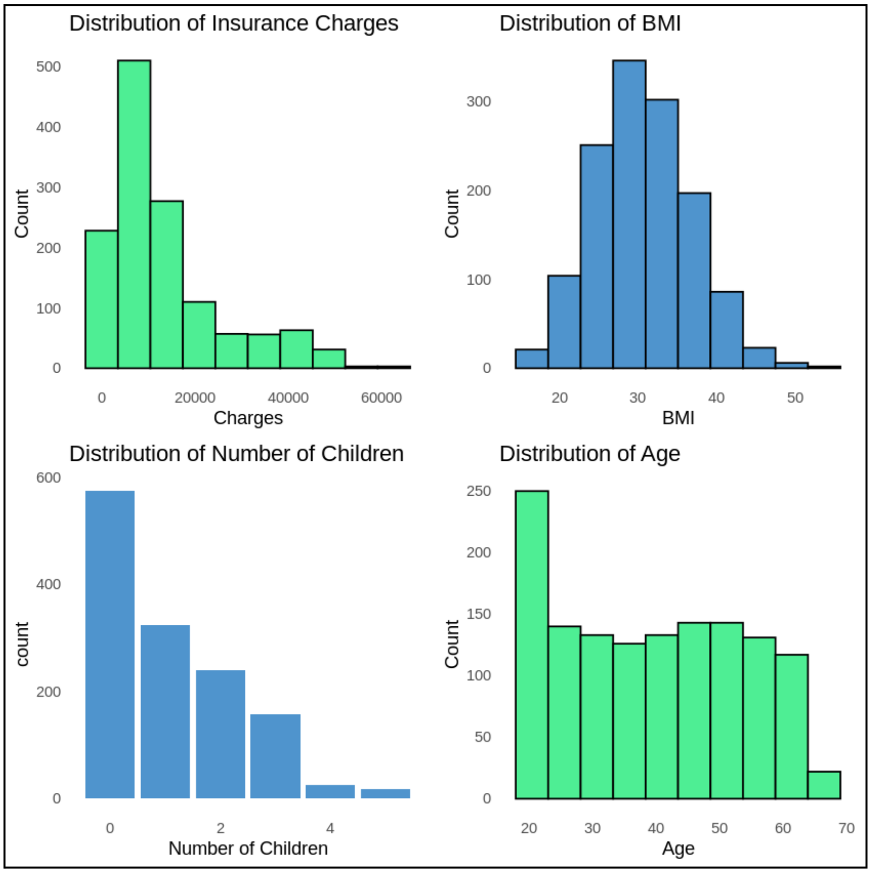


Figure 1: Distributions of Numeric Variables

- ➢ 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.

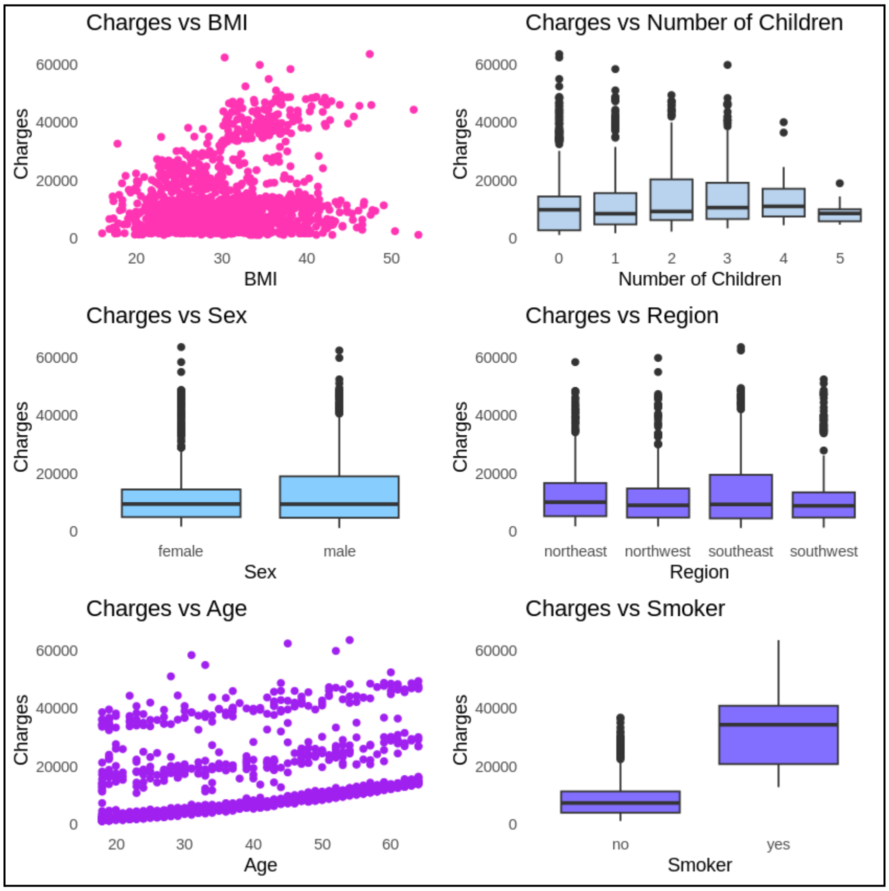


Figure 2: Response Variable vs Explanatory Variables

- ➢ 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.

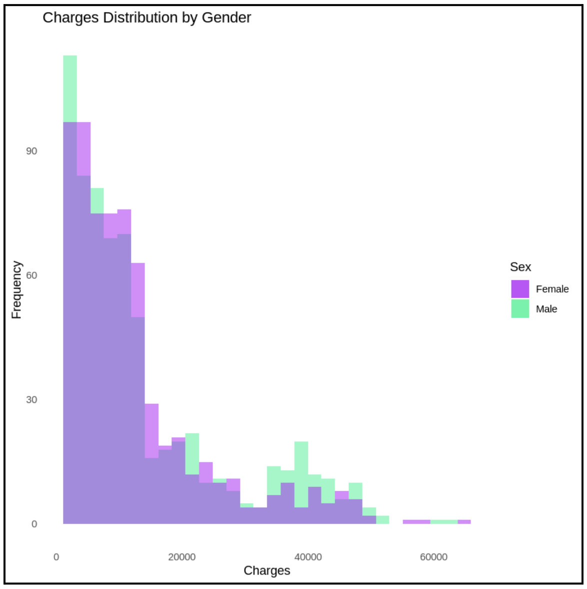
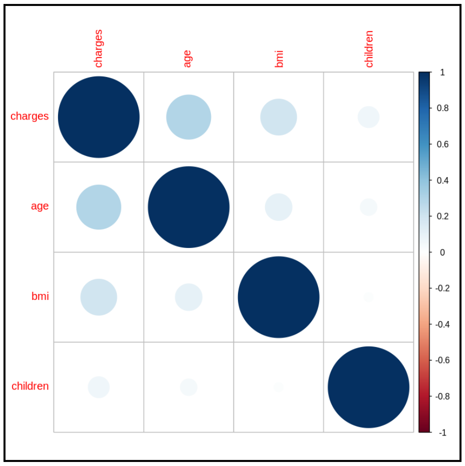
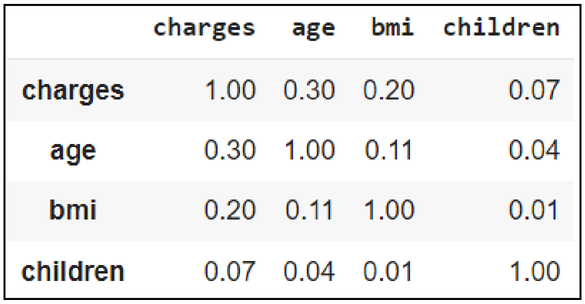


Figure 3: Charges Distribution by Gender

- ➢ 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.





	charges	age	bmi	children
charges	1.00	0.30	0.20	0.07
age	0.30	1.00	0.11	0.04
bmi	0.20	0.11	1.00	0.01
children	0.07	0.04	0.01	1.00

Figure 4: Correlation Matrix Visualization

- ➢ Age and BMI seem to have a low positive correlation with charges. The correlation between number of children and charges is almost negligible. So, we have 3 variables with positive correlation to charges, namely: smoker, age, bmi.

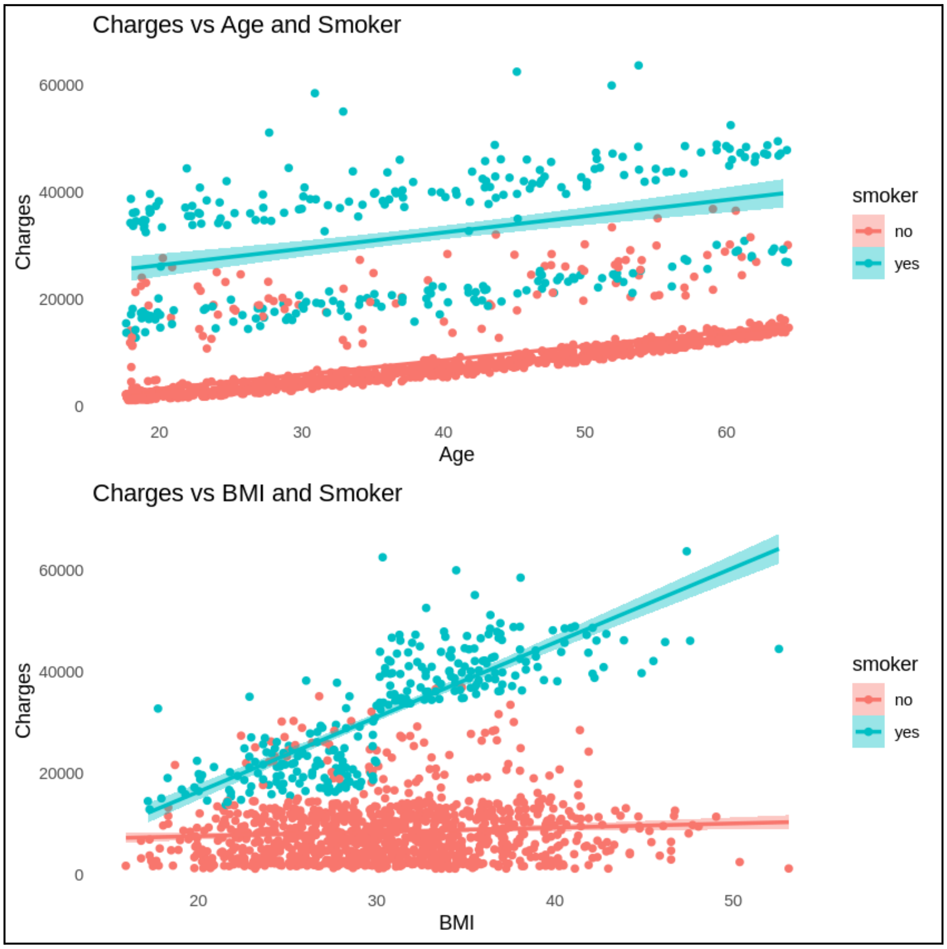


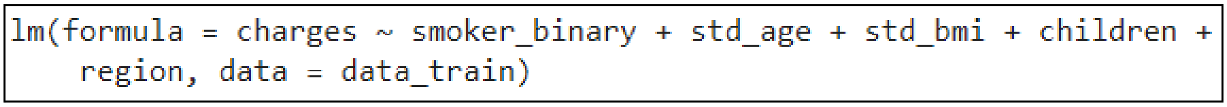
Figure 5: Multivariable Plots

- ➢ 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 3. 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

➢ **Additive model chosen by exhaustive method (model_chosen)**

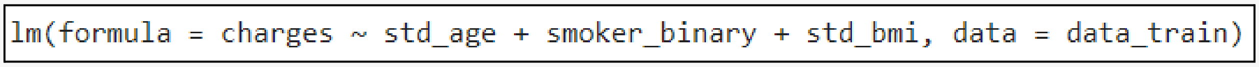
The “regsubsets” selection algorithm was used to compare all possible combinations of variables and the following additive model was chosen. The model with the highest adjusted-r^2 of 0.7541034 and lowest Mallows’ C_p of approximately 6.443 was selected.



```
lm(formula = charges ~ smoker_binary + std_age + std_bmi + children +
    region, data = data_train)
```

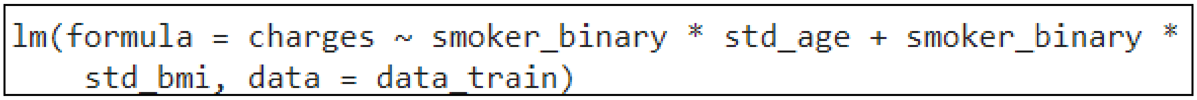
Given that the primary scope of our project centers around investigating the relationship between age and health-related variables that impact insurance charges, our intention is to select the most appropriate model from the ones built using these specific variables only. Below are the models with health-related variables that were analyzed:

➢ **Additive Model with Health Related Variables (add_model)**



```
lm(formula = charges ~ std_age + smoker_binary + std_bmi, data = data_train)
```

➢ **Model with Interactions (int_model)**

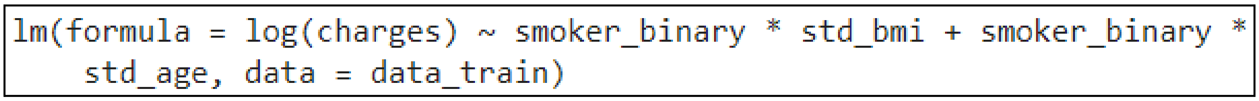


```
lm(formula = charges ~ smoker_binary * std_age + smoker_binary *
    std_bmi, data = data_train)
```

➢ **Model with Interactions and Transformation (tf_model)**

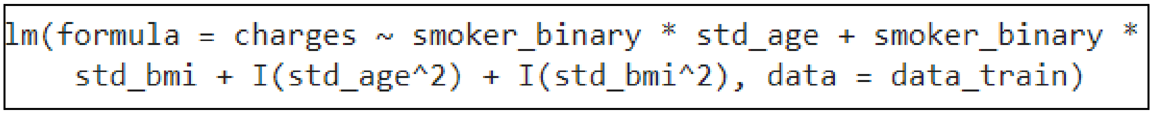
Since the ‘charges' variable in the data is right-skewed, as a few individuals may have exceptionally high charges, a log transformation can help make the data more symmetric and reduce the influence of extreme values.

By applying a log transformation to the charges variable, we can compress the higher values and spread out the lower values. This can make the data conform more closely to the assumptions of linear regression, such as linearly distributed residuals.



```
lm(formula = log(charges) ~ smoker_binary * std_bmi + smoker_binary *
    std_age, data = data_train)
```

➢ **Polynomial Model (poly_model)**



```
lm(formula = charges ~ smoker_binary * std_age + smoker_binary *
    std_bmi + I(std_age^2) + I(std_bmi^2), data = data_train)
```

Diagnostics Table

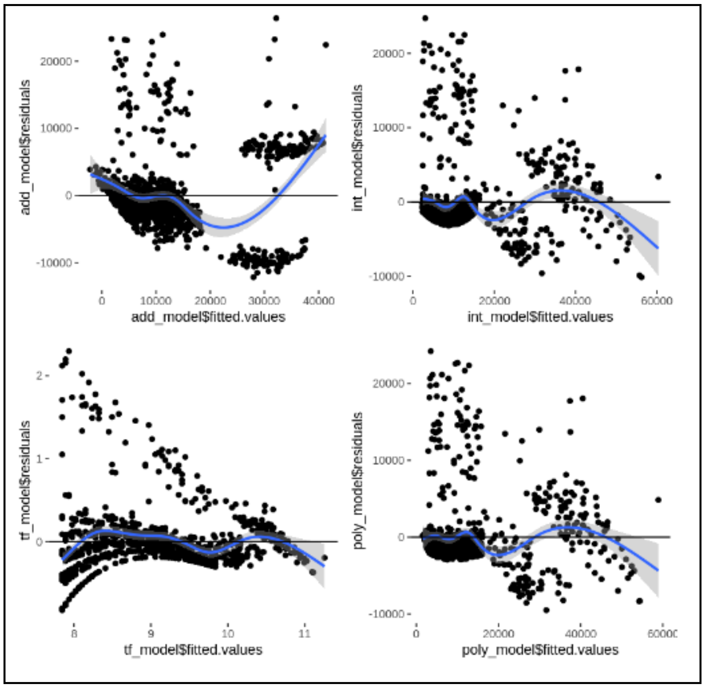
Model	adj.r.squared	RMSE	AIC
Model 1	0.751431	6451.997	20234.03
Model 2	0.84335	5391.474	19774.31
Model 3	0.8153159	18618.09	993.2015
Model 4	0.8444047	5321.946	19769.54

Model Assumptions

1. Linearity

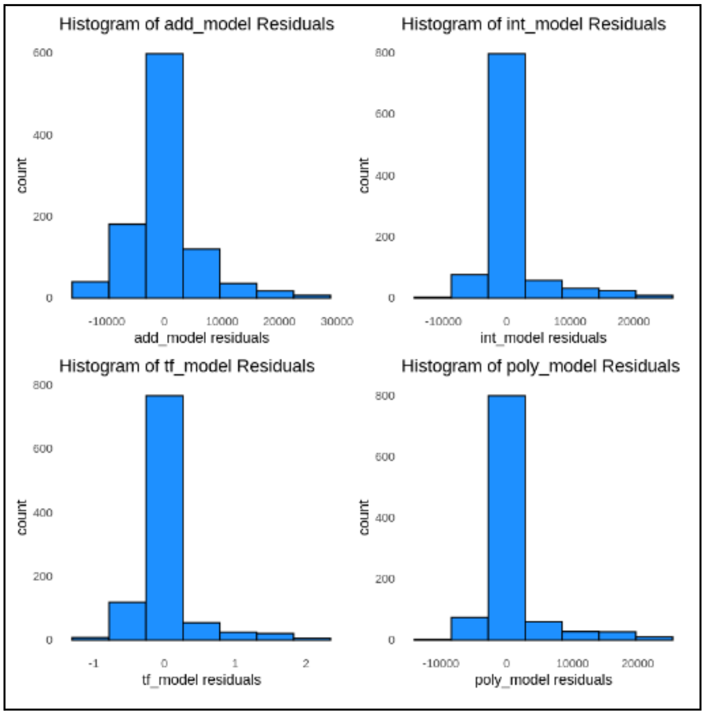
To check for non-linearity in the relationship between explanatory variables and the response, we use residual plots. A plot with a pattern indicates that the model does not meet the linearity assumption.

The log transformation has clearly aided in satisfying the linearity assumption in tf_model (model 3). Model 1 with only addition of variables is non-linear and shows a pattern of more negative residuals as the fitted values increase. Model 2 and 4 are somewhat linear as they show no visible pattern.



2. Normality

To check if the residuals follow a normal distribution we plot histograms and use Shapiro-Wilk Test to validate our findings. Although some of the histograms show that the residuals follow a normal distribution to some degree, using the shapiro-wilk test we found that none of the models have normal residuals.

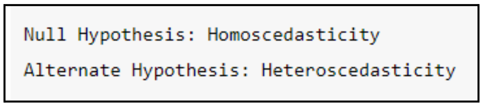


3. Multicollinearity

Using VIF values to check for multicollinearity, we find that all the models have VIF < 10 implying that there is no multicollinearity in any of the models.

4. Homoscedasticity

Using the Breusch-Pagan Test to check for homoscedasticity.



```
Null Hypothesis: Homoscedasticity
Alternate Hypothesis: Heteroscedasticity
```

We reject the assumption of homoscedasticity for the additive model (model 1), and the model with interactions + transformation (model 3) as the p-value < 0.05.

Conclusion:

The model incorporating interactions between health variables demonstrated a better performance compared to the additive model. Clearly, the health variables did not have simple linear relationships with the target variable as shown in the data visualization. As observed from the results of the analysis, it is evident that the additive model presents a simplistic understanding of older people being more susceptible to falling sick, higher BMI contributing to chronic conditions, and smoking increasing the likelihood of a person necessitating medical care. So, this enhancement from an additive model to an interaction model, allowed the model 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. Although the AIC of the transformed model was the lowest and the residual standard error was almost 0, 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. These extreme values can lead to larger prediction errors (higher RMSE) for certain observations. The presence of such large outliers in an insurance dataset could be due to some people requiring expensive health treatments. Overall, the best model was model 4 with RMSE 5321.946 and adjusted-r^2 0.844 meaning that it explains 84.4% of variance in our target variable ‘charges’. Model 2, albeit simpler, also had a similar outcome with adjusted-r^2 0.84335.

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.