Glimpse of the Data

The first five rows of the dataset are shown below.

Out[ ]:

	age	sex	bmi	children	smoker	region	charges
0	19	female	27.900	0	yes	southwest	16884.92400
1	18	male	33.770	1	no	southeast	1725.55230
2	28	male	33.000	3	no	southeast	4449.46200
3	33	male	22.705	0	no	northwest	21984.47061
4	32	male	28.880	0	no	northwest	3866.85520

Resource: The dataset is derived from the book "Machine Learning with R" by Brett Lantz.

Dimension: The dataset has 1,338 rows (observations) and 7 columns (variables).

Variable Description:

1. **age:**
   - *Description:* Age of the primary beneficiary.
   - *Details:* It has a range from 18 to 64 years with an average age of approximately 39.2 years.
2. **sex:**
   - *Description:* Insurance contractor gender.
   - *Details:* It has two categories: male and female. There are 676 males in the dataset.
3. **bmi:**
   - *Description:* Body Mass Index (BMI).
   - *Details:* It's an objective index of body weight (kg / m^2) using the ratio of height to weight. The average BMI in the dataset is approximately 30.66 with a range from 15.96 to 53.13.
4. **children:**
   - *Description:* Number of children covered by health insurance / Number of dependents.
   - *Details:* The range is from 0 to 5 children, with an average of approximately 1.1 children.
5. **smoker:**
   - *Description:* Indicates whether the individual is a smoker or not.
   - *Details:* It has two categories: yes and no. There are 1,064 non-smokers in the dataset.
6. **region:**
   - *Description:* The beneficiary's residential area in the US.

- *Details:* It has four categories: northeast, southeast, southwest, and northwest. The southeast region has the highest number of observations with 364.

7. **charges:**

- *Description:* Individual medical costs billed by health insurance.
- *Details:* The average charge in the dataset is approximately 13,270.42 dollars with a range from $1,121.87 to 63,770.43$.

Statement of Research:

Objective: The primary aim is to craft a predictive model capable of accurately estimating individual medical costs (charges). This estimation will leverage predictors such as age, sex, BMI, number of children, smoking status, and region of residence.

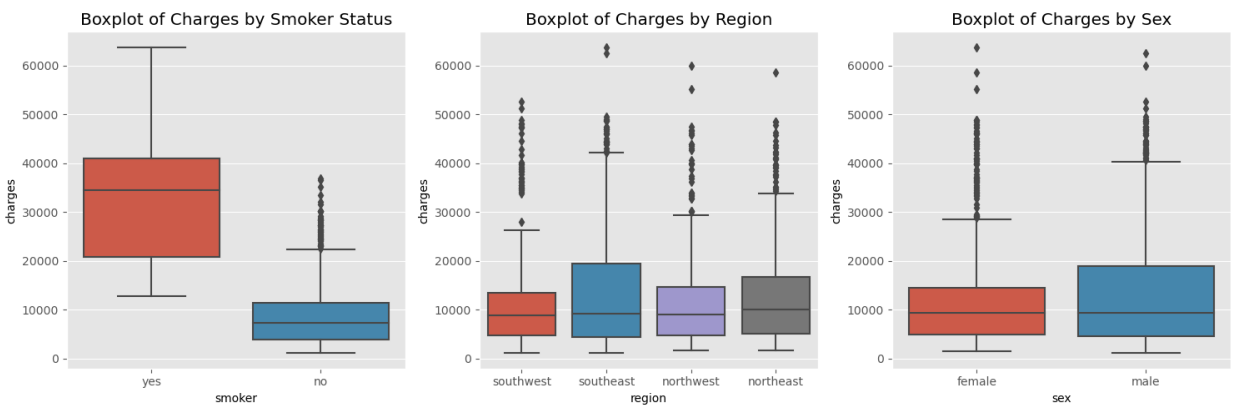
Benefits: This model stands to serve as a pivotal tool for health insurance companies, allowing them to ascertain medical costs for individuals grounded on their health and demographic profile.

Model Utility:

- **Insurance Companies:**
  1. Identification of key determinants influencing medical costs.
  2. Facilitation in the design of apt insurance plans for clientele.
- **Individuals:**
  1. An accessible tool to forecast medical costs based on personal health and demographic information.

Exploratory Data Analysis

Before delving into the intricacies of our predictive model, let's visually explore the underlying relationships and distributions among our predictors and medical costs through boxplots, a correlation heatmap, and pairplots.



The boxplots above reveal that smokers consistently face higher medical charges than non-smokers, with a substantial difference in median costs. Regional differences in charges are less pronounced, though the southeast shows a slight elevation in median expenses. Meanwhile, charges based on gender are closely aligned, with both males and females exhibiting similar median values and variability. In essence, smoking status is the most impactful factor on medical charges among the variables displayed.

Observations from Pair Plots:

1. **Age Distribution:**
   - The distribution appears fairly uniform, with no specific age group dominating the dataset.
2. **Age vs Medical Charges:**
   - Charges tend to increase with age, especially evident for smokers. The trend is clearer for smokers than non-smokers.
3. **Smoker Distribution across BMI:**

- Both smokers and non-smokers are uniformly spread across the BMI range for all age groups.

4. **BMI Distribution:**
    - The distribution appears approximately normal, centering around a BMI value of 30. This hints that many beneficiaries are on the borderline or are classified as overweight.
5. **BMI vs Medical Charges for Smokers and Non-smokers:**
    - For non-smokers, there's no pronounced increase in charges with BMI. In contrast, smokers with a BMI exceeding 30 witness a significant surge in charges, implying higher medical expenses for overweight or obese smokers.
6. **Distribution of Beneficiaries with Children:**
    - A majority have fewer children, mainly ranging from 0 to 2. The number of beneficiaries dwindles as the number of dependents rises.
7. **Distribution of Medical Charges:**
    - The distribution is right-skewed, suggesting that while most beneficiaries have modest medical expenses, a minor fraction incurs considerably high charges.
8. **Impact of Smoking on Medical Charges:**
    - Smokers consistently face higher medical charges than non-smokers, with two distinct clusters observable: one for non-smokers with modest charges and another for smokers with considerably elevated charges.

Hypothesis testing

---

1. Difference in Charges Between Smokers and Non-Smokers

- **Hypothesis**:
    - **H0**: There's no difference in charges between smokers and non-smokers.
    - **H1**: There's a significant difference in charges between smokers and non-smokers.
- **Test Used**: Two-sample t-test
- **Result**: t = 46.665, p = 0.000
- **Interpretation**: We reject the null hypothesis (H0) as the p-value is less than 0.05. This indicates that there is a statistically significant difference in charges between smokers and non-smokers.

---

2. Difference in Charges Based on Gender

- **Hypothesis**:
    - **H0**: There's no difference in charges between males and females.
    - **H1**: There's a significant difference in charges between males and females.
- **Test Used**: Two-sample t-test
- **Result**: t = -2.098, p = 0.036
- **Interpretation**: We reject the null hypothesis (H0) as the p-value is less than 0.05. This suggests that there is a statistically significant difference in charges based on gender.

3. Difference in Charges Across Different Regions

- **Hypothesis**:
  - **H0**: There's no difference in charges across different regions.
  - **H1**: At least one region has a different average charge than the others.
- **Test Used**: One-way ANOVA
- **Result**: F = 2.970, p = 0.031
- **Interpretation**: We reject the null hypothesis (H0) since the p-value is less than 0.05, indicating that there's a significant difference in charges across the regions.

---

4. Relationship Between Age and Charges

- **Hypothesis**:
  - **H0**: Age does not have any effect on charges.
  - **H1**: Age does have an effect on charges.
- **Test Used**: Pearson correlation test
- **Result**: r = 0.299, p = 0.000
- **Interpretation**: We reject the null hypothesis (H0) as the p-value is less than 0.05. The positive correlation indicates that as age increases, the charges also tend to increase.

---

5. Difference in BMI Between Smokers and Non-Smokers

- **Hypothesis**:
  - **H0**: There's no difference in BMI between smokers and non-smokers.
  - **H1**: There's a significant difference in BMI between smokers and non-smokers.
- **Test Used**: Two-sample t-test
- **Result**: t = 0.137, p = 0.891
- **Interpretation**: We fail to reject the null hypothesis (H0) as the p-value is greater than 0.05, suggesting that there's no significant difference in BMI between smokers and non-smokers.

---

6. Smoking Habits Across Genders

- **Hypothesis**:
  - **H0**: The proportion of smokers is the same between males and females.
  - **H1**: The proportion of smokers is different between males and females.
- **Test Used**: Chi-squared test for independence
- **Result**: Chi2 = 7.393, p = 0.007
- **Interpretation**: We reject the null hypothesis (H0) since the p-value is less than 0.05, suggesting that the proportion of smokers differs between males and females.

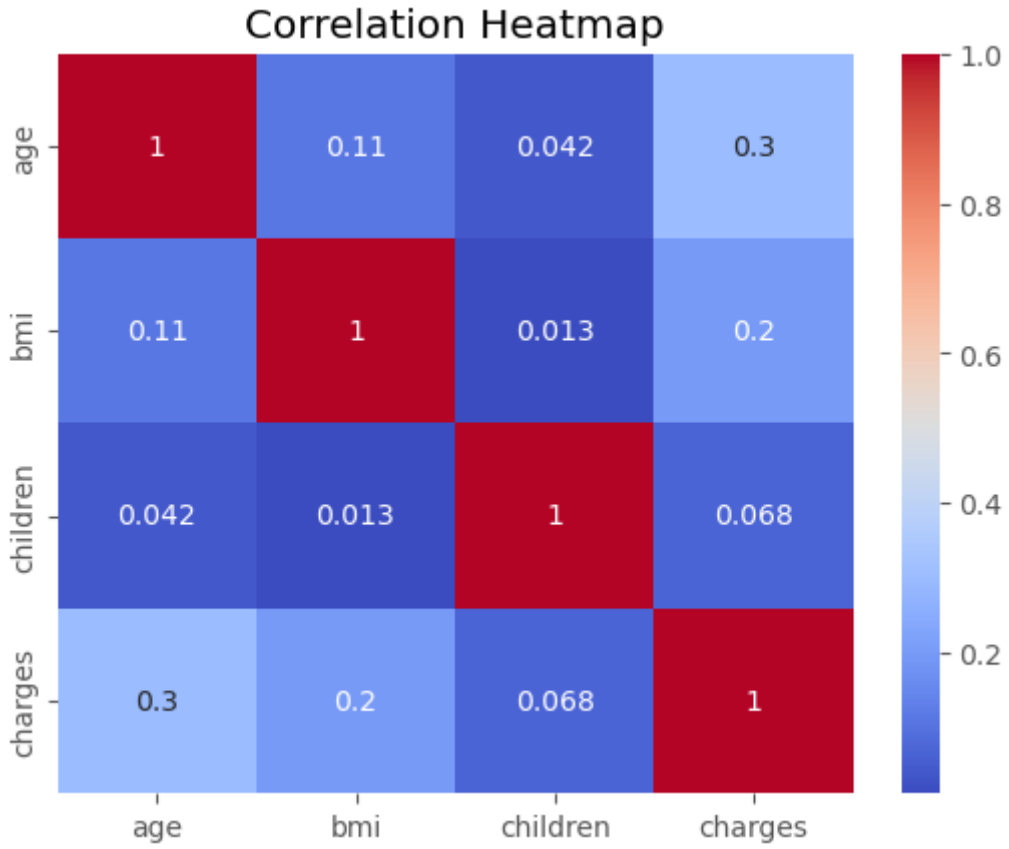
---

7. Association Between Number of Children and Smoking Habits

- **Hypothesis**:

  - **H0**: The number of children a person has is independent of their smoking habits.
  - **H1**: The number of children a person has is associated with their smoking habits.
- **Test Used**: Chi-squared test for independence
- **Result**: Chi2 = 6.888, p = 0.229
- **Interpretation**: We fail to reject the null hypothesis (H0) as the p-value is greater than 0.05, suggesting that the number of children a person has is not significantly associated with their smoking habits.

check for multicollinearity



Correlation Analysis Highlights

1. **Age & Charges:** A moderate positive correlation of **0.3** suggests that medical charges tend to escalate with increasing age.

1. **BMI & Charges:** With a correlation of **0.2**, there's a slight indication that higher BMI values might be associated with increased charges.

1. **General Observations:** Correlations between other variable combinations are notably weak (below **0.2**), indicating limited linear relationships amongst them.

VIF

Along the same vein as the correlation plot analysis, the VIF values for all independent variables were below 10, indicating that multicollinearity is not a significant concern in our dataset. Lower VIF values suggest that the variance in the estimated coefficients is not substantially inflated due to multicollinearity.

	VIF Factor	features
0	32.863734	Intercept
1	1.008878	sex[T.male]
2	1.006457	smoker[T.yes]
3	1.015129	age
4	1.014578	bmi
5	1.002242	children

Initial Full Model Fitting

Let's now fit a full model with all the predictors and evaluate its performance.

Out[ ]:

OLS Regression Results

Dep. Variable:	charges	**R-squared:**	0.751
Model:	OLS	**Adj. R-squared:**	0.749
Method:	Least Squares	**F-statistic:**	500.8
Date:	Sat, 07 Oct 2023	**Prob (F-statistic):**	0.00
Time:	21:47:52	**Log-Likelihood:**	-13548.
No. Observations:	1338	**AIC:**	2.711e+04
Df Residuals:	1329	**BIC:**	2.716e+04
Df Model:	8		
Covariance Type:	nonrobust		

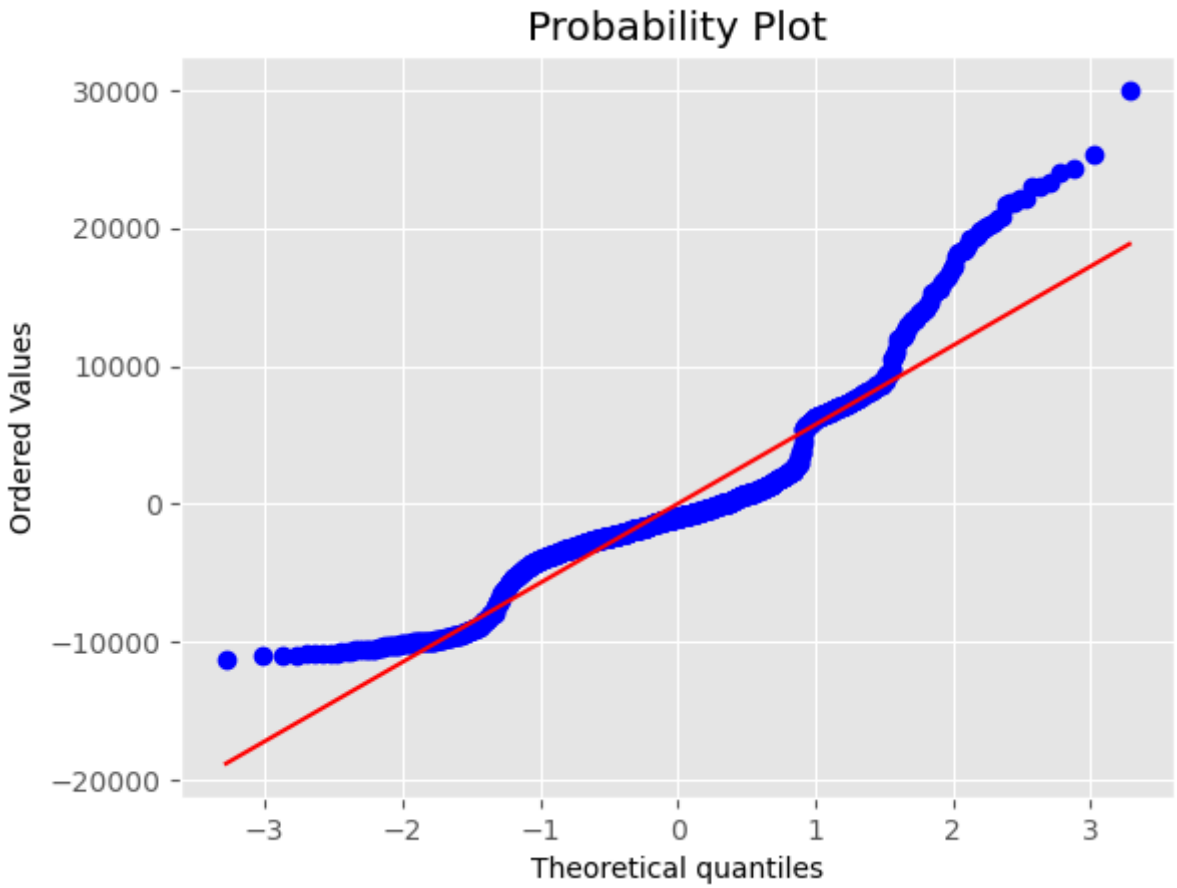
	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-1.194e+04	987.819	-12.086	0.000	-1.39e+04	-1e+04
C(sex)[T.male]	-131.3144	332.945	-0.394	0.693	-784.470	521.842
C(smoker)[T.yes]	2.385e+04	413.153	57.723	0.000	2.3e+04	2.47e+04
C(region)[T.northwest]	-352.9639	476.276	-0.741	0.459	-1287.298	581.370
C(region)[T.southeast]	-1035.0220	478.692	-2.162	0.031	-1974.097	-95.947
C(region)[T.southwest]	-960.0510	477.933	-2.009	0.045	-1897.636	-22.466
age	256.8564	11.899	21.587	0.000	233.514	280.199
bmi	339.1935	28.599	11.860	0.000	283.088	395.298
children	475.5005	137.804	3.451	0.001	205.163	745.838

Omnibus:	300.366	**Durbin-Watson:**	2.088
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	718.887
Skew:	1.211	**Prob(JB):**	7.86e-157
Kurtosis:	5.651	**Cond. No.**	311.

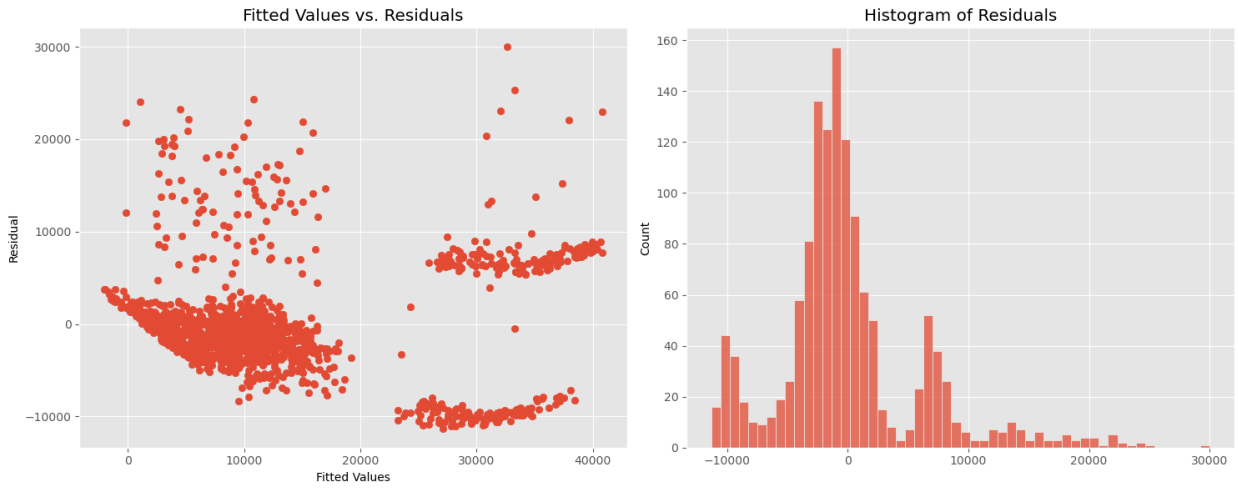
Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

The adjusted R^2 is 0.749. In other words, approximately 74.9% of the variance in the predictors can be accounted for by the independent variables included in the model. But before we proceed with model inference, let's check the assumptions of linear regression.

Model Diagnostics



The given QQ plot illustrates the distribution of the data against a theoretical normal distribution. The significant deviations of the data points (blue) from the red line, especially at both tails, suggest that the data is not normally distributed. The curvature in the middle and the heavier tails indicate potential skewness and the presence of outliers.



The histogram of residuals is clearly not normal. Interesting to see is the modality of the distribution, which suggests that there are three maybe 4 distinct groups of residuals. This is evident in the residual vs fitted plot. Let's explore these clusters in more detail:

Cluster 1:

- **Smoking Status:** All are non-smokers.
- **Age Profile:** The age range is from young adults to seniors (from 18 to 64).
- **Charges Insight:** Charges are relatively high despite being non-smokers. This may indicate that other factors, like age or region, are influencing the charges for this cluster.

Cluster 2:

- **Smoking Status:** All are non-smokers.
- **Age Profile:** The age range is relatively young to middle-aged (from 18 to 46).
- **Charges Insight:** Charges are relatively low, consistent with their non-smoking status and younger age.

Cluster 3:

- **Smoking Status:** All are smokers.
- **Age Profile:** Age range is from young adults to middle-aged (from 22 to 34).
- **Charges Insight:** Charges are very high, likely driven by their smoking status combined with other health factors.

Cluster 4:

- **Smoking Status:** All are smokers.
- **Age Profile:** The age range is young (from 19 to 53).
- **Charges Insight:** Despite being smokers, their charges are relatively lower compared to Cluster 3. This might be due to their younger age or possibly other unobserved health factors.

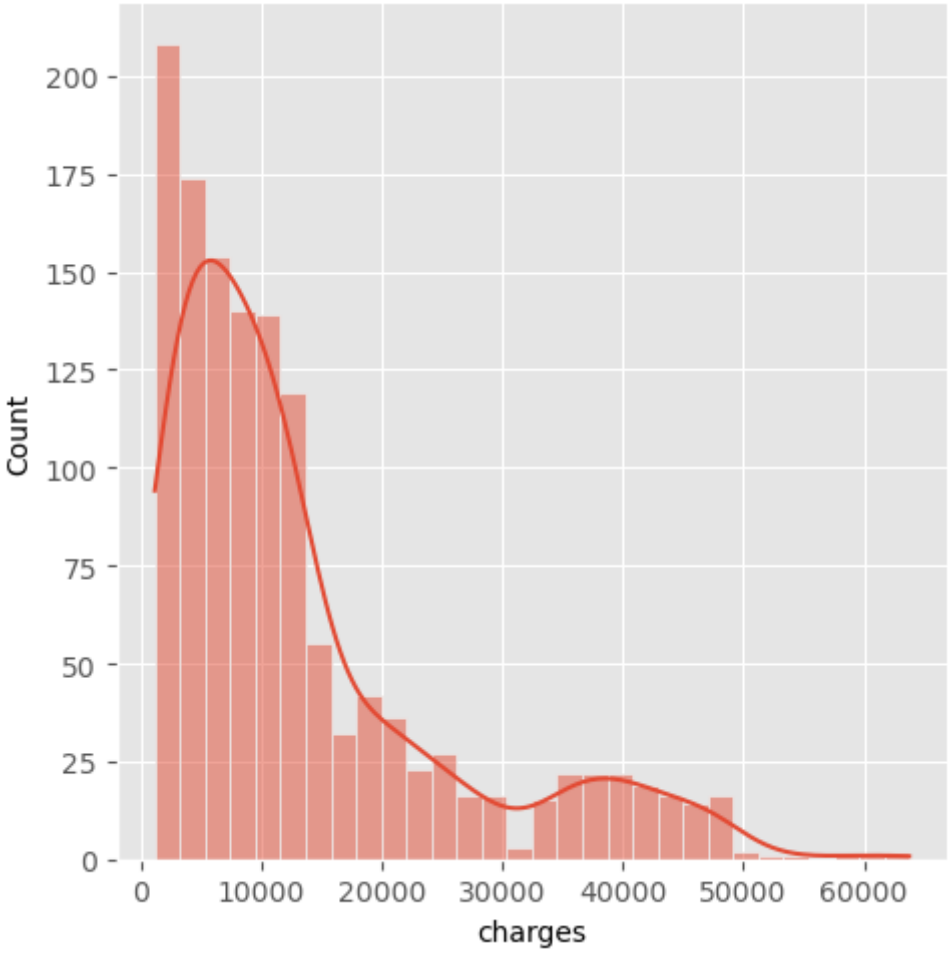
Summary & Insights:

- **Smoking Differentiator:** Smoking appears to be a significant differentiator among the clusters. Clusters with smokers (3 & 4) have higher charges than non-smokers (1 & 2).
- **Age Influence:** Age plays a role in the charges. Cluster 1 with older non-smokers has higher charges than Cluster 2 with younger non-smokers. Similarly, Cluster 3 with middle-aged smokers has higher charges than Cluster 4 with younger smokers.
- **Regional Influence:** The region varies across clusters, which might influence the charges. However, from the provided data, it's challenging to ascertain a clear pattern.

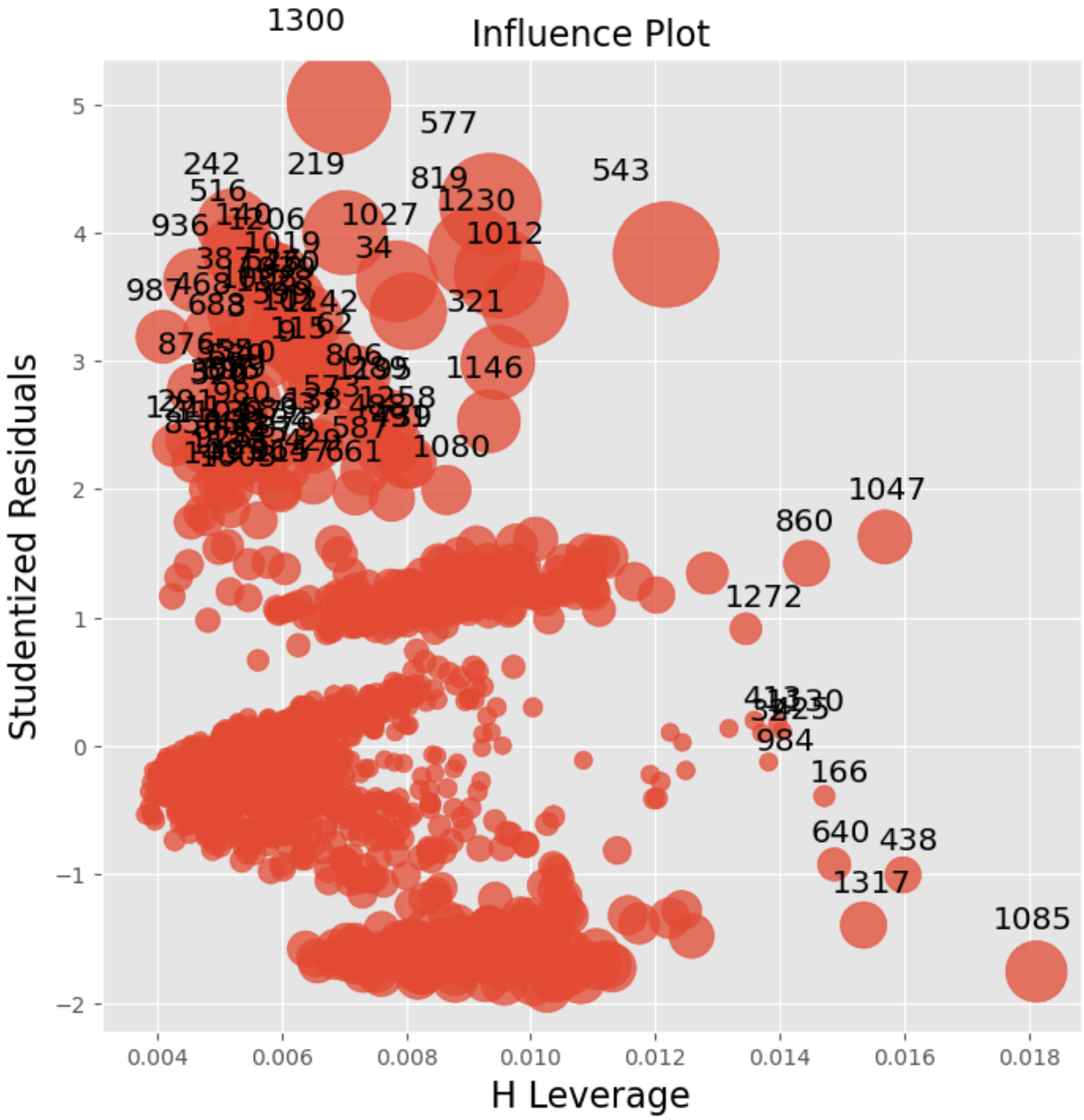
From all this, we can infer the importance of including interaction terms in our model that involve smoking.

We also notice the distribution of the response variable charges is highly skewed to the right. Therefore another adjustment we can make is to take the log transform as our new response variable.

Out[ ]: `<seaborn.axisgrid.FacetGrid at 0x7fa1e8bceb90>`



The Breusch-Pagan test's high LM Statistic of 121.74 and an extremely low p-value (extremely close to zero) for the model "charges ~ age + C(sex) + bmi + children + C(smoker) + C(region)" suggest the presence of heteroscedasticity, indicating that the model's residuals have non-constant variance.



The influence plot visualizes the studentized residuals against the leverage of each observation, with the size of the circles representing the influence of the points. Several data points, such as 1300 and 577 have high leverage and residuals, indicating potential outliers with considerable influence on the regression model.

In the analysis, we applied multiple criteria to detect influential observations in the regression model. First, we utilized studentized residuals, identifying points that had residuals exceeding the 97.5th percentile of the t-distribution. Next, Cook's distance was computed to determine the influence of each observation on the fitted values, with observations surpassing a threshold of 4/n being flagged. Finally, high-leverage points were found based on their hat-values, with observations having values greater than 2p/n deemed high-leverage. Combining these criteria, we pinpointed the following four outlier points: 577, 321, 1012, and 543.

Model Evaluation and Selection

In our analysis, we undertook a comprehensive model selection approach, incorporating not only individual predictors but also interaction terms to capture potential synergistic effects between variables. Initially, we applied a best subset selection technique, examining all possible combinations of predictors, including interactions, for predicting log_charges. We evaluated each subset based on its adjusted R-squared value, aiming to maximize explanatory power while considering model complexity. Subsequently, we adopted a forward selection approach based on the Akaike Information Criterion (AIC). Starting with an empty model, we iteratively added predictors or interactions that resulted in the most substantial reduction in AIC, continuing until no further improvements were detected. This blend of exhaustive and stepwise methodologies provided a thorough evaluation of the predictor space. The final outcomes, including top-performing models and their respective metrics, are concisely presented in a structured table format.

Below are the top-performing models based on the adjusted R-squared metric:

# Predictors	Predictors	Adjusted R-squared
6	C(sex)[T.male], smoker[T.yes], region[T.northwest], region[T.southeast], region[T.southwest], age, smoker[T.yes]:age, children, bmi, smoker[T.yes]:bmi, region[T.northwest]:bmi, region[T.southeast]:bmi, region[T.southwest]:bmi	0.82433
7	C(sex)[T.male], C(smoker)[T.yes], smoker[T.yes], region[T.northwest], region[T.southeast], region[T.southwest], bmi, smoker[T.yes]:bmi, region[T.northwest]:bmi, region[T.southeast]:bmi, region[T.southwest]:bmi, children, age, smoker[T.yes]:age	0.82433
8	C(sex)[T.male], C(smoker)[T.yes], smoker[T.yes], region[T.northwest], region[T.southeast], region[T.southwest], age, smoker[T.yes]:age, bmi, smoker[T.yes]:bmi, region[T.northwest]:bmi, region[T.southeast]:bmi, region[T.southwest]:bmi, children	0.82433
9	C(sex)[T.male], C(smoker)[T.yes], C(region)[T.northwest], C(region)[T.southeast], C(region)[T.southwest], smoker[T.yes], region[T.northwest], region[T.southeast], region[T.southwest], age, smoker[T.yes]:age, bmi, smoker[T.yes]:bmi, region[T.northwest]:bmi, region[T.southeast]:bmi, region[T.southwest]:bmi, children	0.82433
5	C(sex)[T.male], smoker[T.yes], region[T.northwest], region[T.southeast], region[T.southwest], children, bmi, smoker[T.yes]:bmi, region[T.northwest]:bmi, region[T.southeast]:bmi, region[T.southwest]:bmi, age, smoker[T.yes]:age	0.82433
4	smoker[T.yes], region[T.northwest], region[T.southeast], region[T.southwest], children, bmi, smoker[T.yes]:bmi, region[T.northwest]:bmi, region[T.southeast]:bmi, region[T.southwest]:bmi, age, smoker[T.yes]:age	0.82241
3	smoker[T.yes], children, bmi, smoker[T.yes]:bmi, age, smoker[T.yes]:age	0.81760
2	smoker[T.yes], bmi, smoker[T.yes]:bmi, age, smoker[T.yes]:age	0.79867
1	smoker[T.yes], age, smoker[T.yes]:age	0.77653

Below is the performance of all models based on the AIC metric:

Formula	AIC
log_charges ~ age	3138.62
log_charges ~ C(sex)	3575.53
log_charges ~ bmi	3551.81
log_charges ~ children	3540.29
log_charges ~ C(smoker)	2792.84
log_charges ~ C(region)	3575.36
log_charges ~ smoker*bmi	2728.04
log_charges ~ smoker*age	1571.59
log_charges ~ region*bmi	3551.19
log_charges ~ smoker*age + age	1571.59
log_charges ~ smoker*age + C(sex)	1566.81
log_charges ~ smoker*age + bmi	1546.83
log_charges ~ smoker*age + children	1456.61
log_charges ~ smoker*age + C(smoker)	1571.59
log_charges ~ smoker*age + C(region)	1564.58
log_charges ~ smoker*age + smoker*bmi	1434.03
log_charges ~ smoker*age + region*bmi	1530.31
log_charges ~ smoker*age + smoker*bmi + age	1434.03
log_charges ~ smoker*age + smoker*bmi + C(sex)	1423.31
log_charges ~ smoker*age + smoker*bmi + bmi	1434.03
log_charges ~ smoker*age + smoker*bmi + children	1302.89
log_charges ~ smoker*age + smoker*bmi + C(smoker)	1434.03
log_charges ~ smoker*age + smoker*bmi + C(region)	1411.90
log_charges ~ smoker*age + smoker*bmi + region*bmi	1407.67
log_charges ~ smoker*age + smoker*bmi + children + age	1302.89
log_charges ~ smoker*age + smoker*bmi + children + C(sex)	1289.20
log_charges ~ smoker*age + smoker*bmi + children + bmi	1302.89
log_charges ~ smoker*age + smoker*bmi + children + C(smoker)	1302.89
log_charges ~ smoker*age + smoker*bmi + children + C(region)	1276.87
log_charges ~ smoker*age + smoker*bmi + children + region*bmi	1273.05
log_charges ~ smoker*age + smoker*bmi + children + region*bmi + age	1273.05
log_charges ~ smoker*age + smoker*bmi + children + region*bmi + C(sex)	1259.54
log_charges ~ smoker*age + smoker*bmi + children + region*bmi + bmi	1273.05
log_charges ~ smoker*age + smoker*bmi + children + region*bmi + C(smoker)	1273.05
log_charges ~ smoker*age + smoker*bmi + children + region*bmi + C(region)	1273.05

Formula	AIC
log_charges ~ smoker*age + smoker*bmi + children + region*bmi + C(sex) + age	1259.54
log_charges ~ smoker*age + smoker*bmi + children + region*bmi + C(sex) + bmi	1259.54
log_charges ~ smoker*age + smoker*bmi + children + region*bmi + C(sex) + C(smoker)	1259.54
log_charges ~ smoker*age + smoker*bmi + children + region*bmi + C(sex) + C(region)	1259.54

Best Model (Forward Selection)	AIC	Adjusted R-squared
log_charges ~ smoker*age + smoker*bmi + children + region*bmi + C(sex)	1259.54	0.82433

The summary of the best performing model based on our selection procedure is shown below:

Out[ ]:

OLS Regression Results

Dep. Variable:	log_charges	**R-squared:**	0.826
Model:	OLS	**Adj. R-squared:**	0.824
Method:	Least Squares	**F-statistic:**	483.6
Date:	Sat, 07 Oct 2023	**Prob (F-statistic):**	0.00
Time:	21:48:03	**Log-Likelihood:**	-615.77
No. Observations:	1338	**AIC:**	1260.
Df Residuals:	1324	**BIC:**	1332.
Df Model:	13		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	6.9582	0.116	60.002	0.000	6.731	7.186
smoker[T.yes]	1.2521	0.147	8.541	0.000	0.965	1.540
region[T.northwest]	0.0673	0.164	0.409	0.682	-0.255	0.390
region[T.southeast]	0.2833	0.152	1.869	0.062	-0.014	0.581
region[T.southwest]	-0.0369	0.160	-0.231	0.817	-0.350	0.276
C(sex)[T.male]	-0.0834	0.021	-3.929	0.000	-0.125	-0.042
age	0.0415	0.001	48.808	0.000	0.040	0.043
smoker[T.yes]:age	-0.0335	0.002	-17.803	0.000	-0.037	-0.030
bmi	0.0073	0.004	1.937	0.053	-9.48e-05	0.015
smoker[T.yes]:bmi	0.0521	0.004	12.304	0.000	0.044	0.060
region[T.northwest]:bmi	-0.0045	0.006	-0.820	0.412	-0.015	0.006
region[T.southeast]:bmi	-0.0136	0.005	-2.845	0.005	-0.023	-0.004
region[T.southwest]:bmi	-0.0040	0.005	-0.769	0.442	-0.014	0.006
children	0.1059	0.009	12.068	0.000	0.089	0.123

Omnibus:	858.619	**Durbin-Watson:**	1.992
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	8191.616
Skew:	2.958	**Prob(JB):**	0.00
Kurtosis:	13.580	**Cond. No.**	1.24e+03

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.24e+03. This might indicate that there are strong multicollinearity or other numerical problems.

1. Model Metrics:

- **R-squared:** This value (0.826) indicates that the model explains 82.6% of the variability in the logarithm of medical charges. This is a relatively high R-squared, indicating the model fits the data well.
- **Adj. R-squared:** Adjusted R-squared (0.824) is slightly less than the R-squared. It adjusts for the number of predictors in the model, penalizing for unnecessary complexity. Since it's very close to R-squared, it indicates that most predictors are relevant.
- **F-statistic & Prob (F-statistic):** The F-statistic is 483.6 with an associated p-value of 0.00. This tests the hypothesis that all of the regression coefficients are equal to zero. A p-value of 0.00 indicates that the model's predictors, as a set, are statistically significant.

2. Model Coefficients:

Intercept (6.9582, p-value: 0.000)

The base log medical charge when other predictors are at their reference levels or zero (though some zeroes might be non-practical). This coefficient is highly significant.

Region

Compared to the omitted baseline region:

- **Northwest (`region[T.northwest]` , coef: 0.0673, p-value: 0.682):** Living in the northwest seems to have an effect on log medical charges, but this effect is not statistically significant.
- **Southeast (`region[T.southeast]` , coef: 0.2833, p-value: 0.062):** Residing in the southeast results in a borderline significant increase in log charges by 0.2833 units.
- **Southwest (`region[T.southwest]` , coef: -0.0369, p-value: 0.817):** Living in the southwest seems to have an effect on log medical charges, but this effect is not statistically significant.

Sex (`C(sex)[T.male]` , coef: -0.0834, p-value: 0.000)

Being male, as compared to female, is associated with a statistically significant decrease in log medical charges by about 0.0834 units.

Age and Smoker-Age Interaction

Considering the interaction with smoking status, the effect of age on log medical charges for smokers is `0.0415 - 0.0335 = 0.008` units (both coefficients are highly significant). This means that, for smokers, each additional year increases log medical charges by 0.008 units more than it does for non-smokers.

BMI, Smoker-BMI, and Region-BMI Interactions:

For non-smokers in the reference region, a unit increase in BMI increases the log charges by 0.0073 units, which is borderline significant. However, for smokers in the reference region, a unit increase in

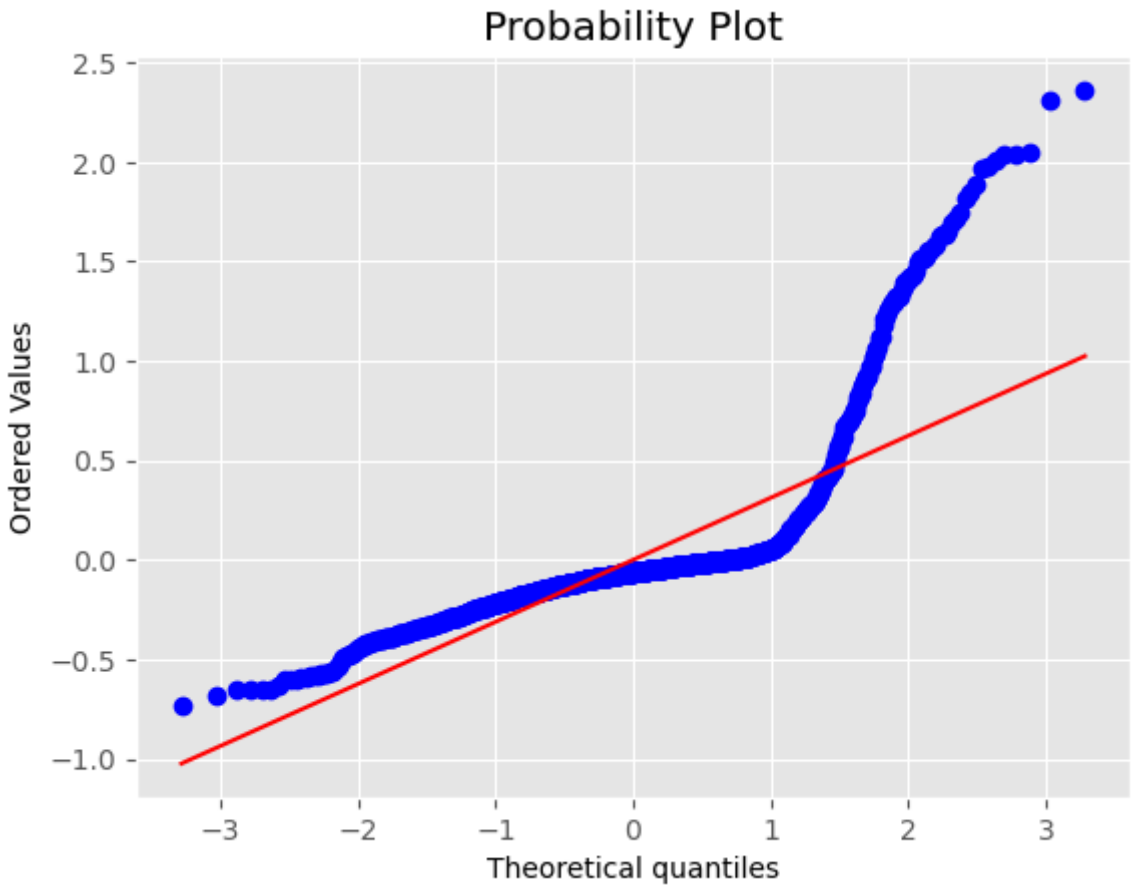
BMI affects the log charges by `0.0073 + 0.0521 = 0.0594` units, which is highly significant.

- For individuals in the southeast, a unit increase in BMI changes the log medical charges by `0.0073 - 0.0136 = -0.0063` units, a significant decrease compared to the baseline region.
- The interactions involving northwest and southwest with BMI are statistically insignificant.

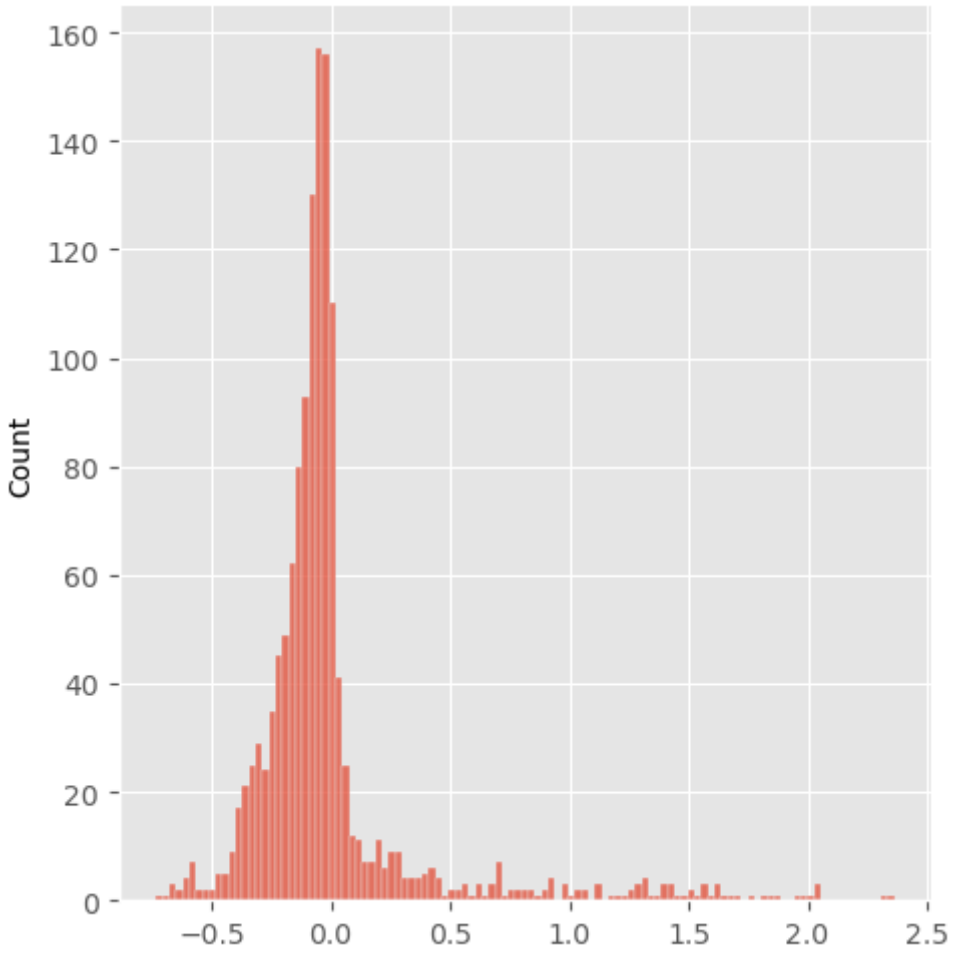
Children (`children` , coef: 0.1059, p-value: 0.000)

Having an additional child leads to a significant increase in the log medical charges by 0.1059 units.

If we look at the model diagnostic plots again, we see an improvement. Though, there are still clear departures from assumptions that one needs to be wary of when drawing conclusions from the model and our research study.



Out[ ]: `<seaborn.axisgrid.FacetGrid at 0x7fa1e9bf1390>`



The breusch pagan test still shows heteroscedasticity, but the LM statistic has reduced from 121.74 to 72.74, showing a marked improvement.

Summary, Conclusion, and Recommendations for Future Analysis

Drawing on data from "Machine Learning with R" by Brett Lantz, our in-depth analysis set out to construct a predictive model to accurately estimate individual medical costs, encompassing variables such as age, gender, BMI, number of dependents, smoking habits, and regional differences. Confirming our research objectives, the selected model explains 82.6% of the variance in medical charges, highlighting particularly pronounced effects of smoking, age, and their interaction. Moreover, BMI and its interactions with smoking status and region also showcased important, albeit more nuanced, implications for predicting medical expenses. Insurance companies can leverage these insights to tailor insurance plans more adeptly, aligning premiums with individual risk profiles. Meanwhile, individuals gain a tool to forecast prospective medical costs based on their health and demographic markers.

While our model provides a robust foundation for predicting medical charges, there are several avenues for future research. First, the dataset could be expanded to include additional variables, such as the type of insurance plan, the number of hospital visits, and the number of prescriptions. Second, the model could be extended to incorporate non-linear relationships between the predictors and response

variable. Finally, the model could be applied to a larger dataset to assess its generalizability and performance on unseen data.