Honor code

By submitting this assignment, I affirm the following:

1. All work presented in this assignment is my own. I have not collaborated with others or copied work from any unauthorized source.
2. If I used AI tools like ChatGPT, Co-Pilot, etc., I only sought guidance or clarification. Any generated content has been fully understood and appropriately modified to align with the assignment.
3. I understand the submitted code and can explain my work if asked.

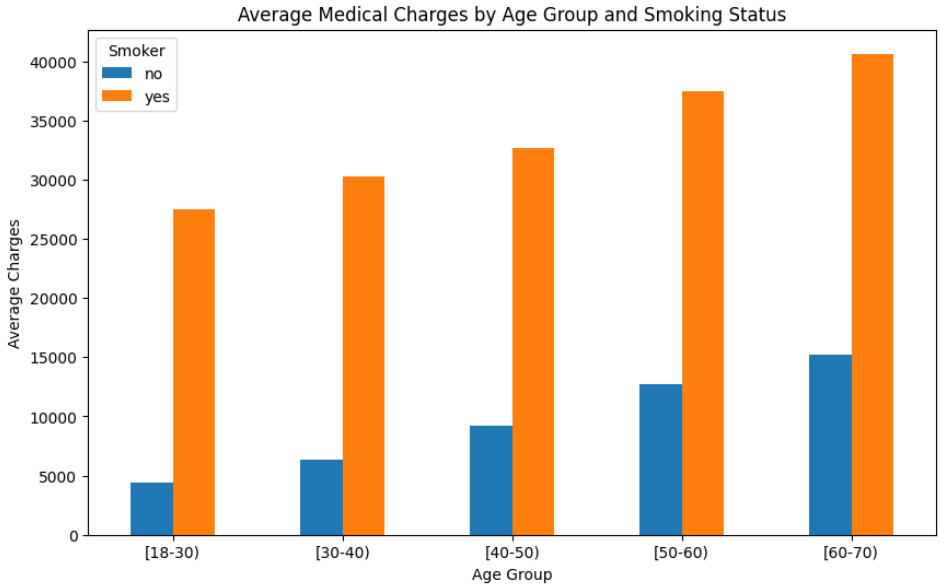
I declare that I have read, understood, and agree to abide this honor code.

Name: Kevin Vu

Student number: 713837tv

Date: 15/11/2024

1. Age group with the largest difference in charges between smokers and non-smokers: [60-70)



Value of the maximum difference: 25397.99

2. Correlation table:

	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

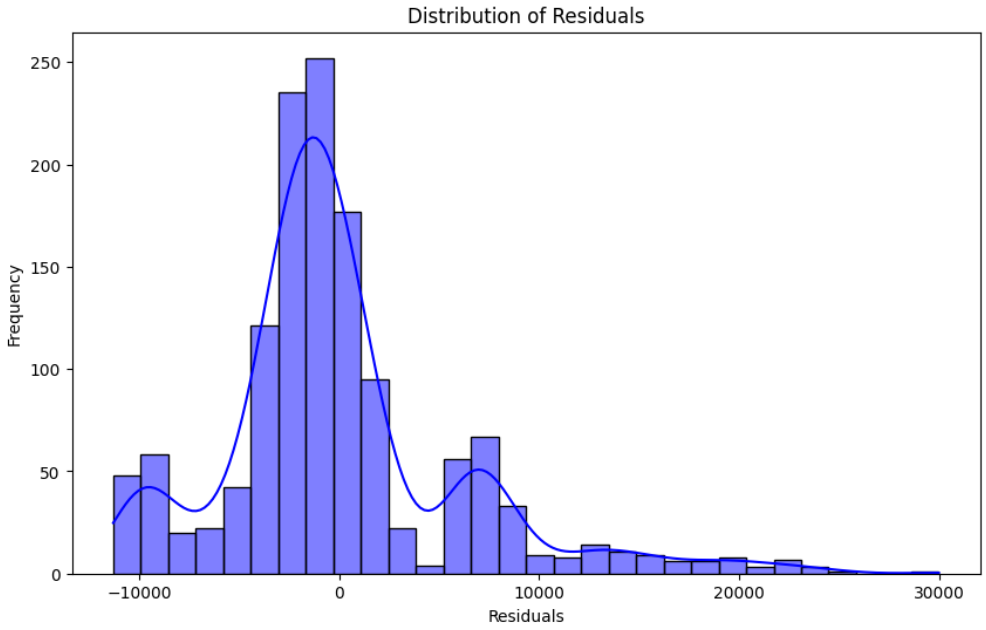
Variable that has strongest correlation with charges: 'age'

3. Report variables that are statistically significant and the significance level:
   - Smoker (Yes): Significant at the 0.01 level ($p < 0.01$)
   - Age: Significant at the 0.01 level ($p < 0.01$)
   - BMI: Significant at the 0.01 level ($p < 0.01$)
   - Children: Significant at the 0.01 level ($p < 0.01$)
   - Region:
     - Southeast: Significant at the 0.05 level ($p < 0.05$)
     - Southwest: Significant at the 0.05 level ($p < 0.05$)

4. Interpretation of coefficients (1-2 sentences each):
   - Age: Each additional year of age increases charges by approximately 256.9 units, indicating that older age is associated with higher costs.
   - Gender: Gender does not significantly impact charges in this model, suggesting no strong Difference between males and females.
   - BMI: Each additional unit of BMI raises charges by around 339.2 units, meaning higher BMI is linked to higher insurance costs.
   - Region (Northeast is ref):
     - Northwest: Compared to the Northeast, charges are lower in the Northwest (-353 units).
     - Southeast: Charges are also lower in the Southeast relative to the Northeast (-1035 units).

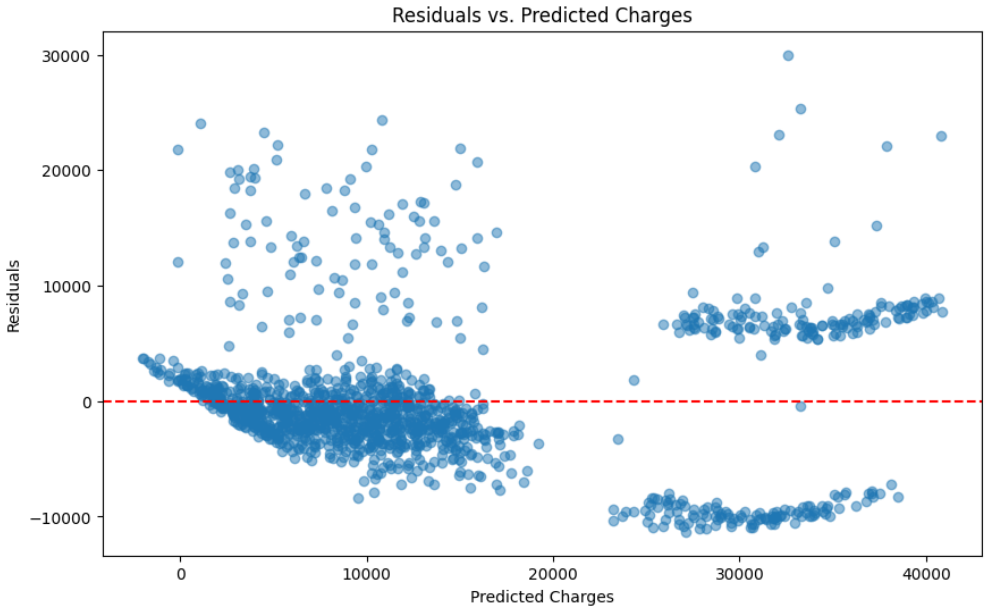
- Southwest: Similarly, the Southwest has lower charges than the Northeast (-960 units).

5. a) Histogram of residuals:



Comment on the histogram:
The residuals appears roughly bell-shaped and centered around 0. They suggest an roughly normal distribution with outliers. There are also noticeable **deviations** from perfect symmetry, and there is a presence of **right skew** indicates that the model might be underpredicting charges more frequently than overpredicting.

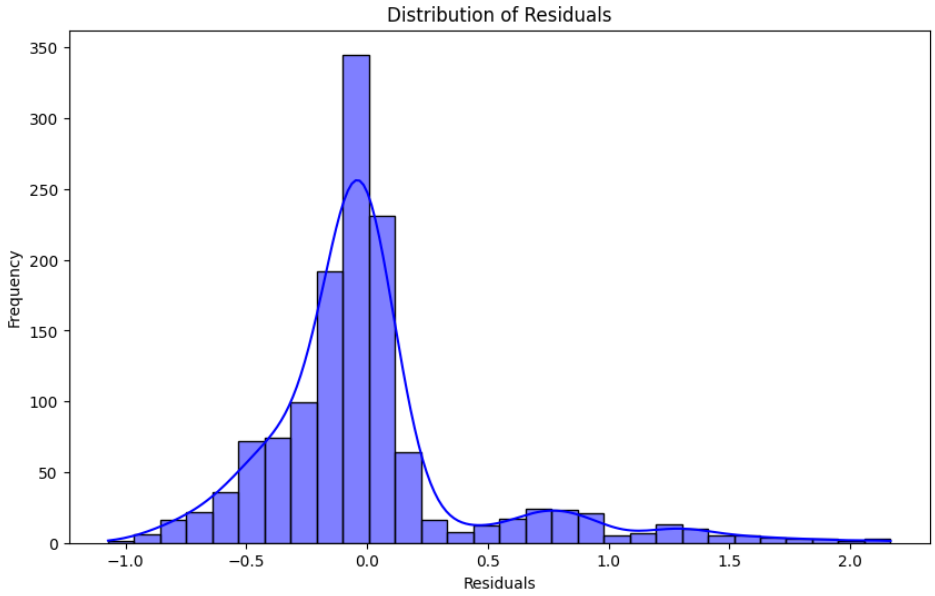
b) Residuals vs. predicted charges



Comment on homoscedasticity assumption:
There is a **funnel shape** or **fan-like pattern**, where the spread of residuals increases with higher predicted values. This pattern indicates **heteroscedasticity** because the variance of residuals is not constant across the range of predicted charges.

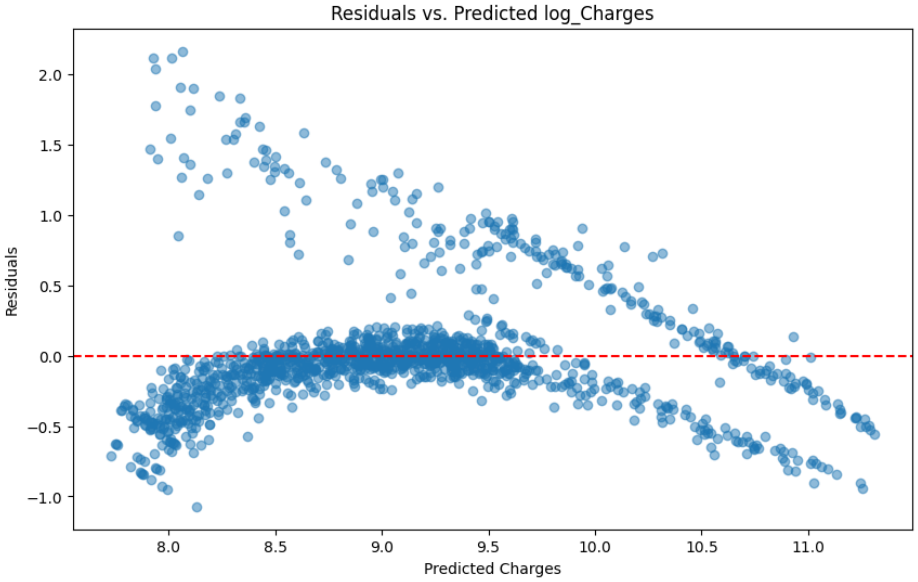
6. a) Interpretation of the coefficient of age: coefficient of 'age' = 0.0346 when the target variable is "log_charges". This mean that for each additional year of age, the predicted Insurance charges Increase by approximately 3.46%, holding all other variables constant.

   b) Histogram of residuals:



   Comment on the histogram:
   The histogram appears to be bell-shaped and centered around 0, which means that the residuals are approximately normally distributed. The normality assumption is likely met.

   c) Residuals vs. predicted charges:



   Comment on homoscedasticity assumption:
   In this plot, the **funnel-like pattern** indicates **heteroscedasticity**, meaning the variance of the residuals **increases** as the predicted values change. The residuals do not maintain a consistent spread but instead fan out, suggesting a violation of the homoscedasticity assumption.

7. R-squared and Adjusted R-squared value and explanation:
    - R-squared: 0.7679 -> 76.79% of the variability in the log-transformed insurance charges can be explained by the predictor variables, which include age, gender, BMI, children, smoking status, and region. The predictor variables explain a large part of the variation in healthcare costs
    - Adjusted R-squared: 0.7666 -> 76.66% is very close to the R-squared value, which suggests that most of the predictors in the model are relevant and contribute meaningfully to explaining the variation in log_charges. This means that the predictors included are appropriate and relevant.

8. Model prediction:

age	gender	bmi	children	smoker	region	charges
25	male	28.0	1	no	northeast	4007.947498
45	female	35.2	3	yes	southeast	47109.137290
32	male	30.5	0	no	northwest	4473.376054
54	female	24.7	2	yes	southwest	51917.388975
29	female	22.8	1	yes	southeast	18720.126574

9. % of observations assigned to class 0: 68.61
   % of observations assigned to class 1: 31.39

10. Report variables that are statistically significant and the significance level:
    - **Smoker (yes)**: **Coefficient**: 8.3972| **p-value < 0.01** (significant at the 0.01 level)
      This indicates that being a smoker significantly increases the likelihood of having charges above the mean, with a very large positive impact.
    - **Region (southwest)**: **Coefficient**: -0.6648 | **p-value < 0.05** (significant at the 0.05 level)
      This suggests that individuals from the **southwest** are less likely to have charges above the mean compared to those in the reference region (northeast).
    - **Age**: **Coefficient**: 0.0713 | **p-value < 0.01** (significant at the 0.01 level)
      Each additional year of age is associated with a higher likelihood of having charges above the mean.

11. Interpretation of coefficients (1-2 sentences each):
    - Age: coef = 0.0713 indicates that for each additional year of age, the odds of having charges above the mean increase by about **7.13%**, holding all other factors constant. This suggests that older individuals are more likely to have higher medical costs.
    - Gender: coef = -0.2747 suggests that a male Individual will have a slightly decreased odd of having charges above the mean compared to females. Although this is not significant (p-value > 0.05).
    - BMI: coef = 0.0182 means that for each additional unit increase In BMI, the odds of having charges above the mean increase by approximately 1.82%. Despite that, this effect is not statistically significant (p-value > 0.05), suggesting that BMI does not strongly impact on whether the charges are above the mean.
    - Region (Northeast is ref):
        - Northwest: The coefficient (-0.1578) for individuals in the northwest indicates a slight decrease in the odds of having charges above the mean compared to the reference region (northeast), but this effect is not statistically significant (p-value > 0.05).
        - Southeast: The coefficient (-0.2034) for individuals in the southeast suggests a small decrease in odds compared to the northeast, but this is not statistically significant (p-

value > 0.05).

- Southwest: The coefficient (-0.6648) for individuals in the southwest is significant (p-value < 0.05), indicating that individuals from the southwest have significantly lower odds of having charges above the mean compared to those in the northeast.

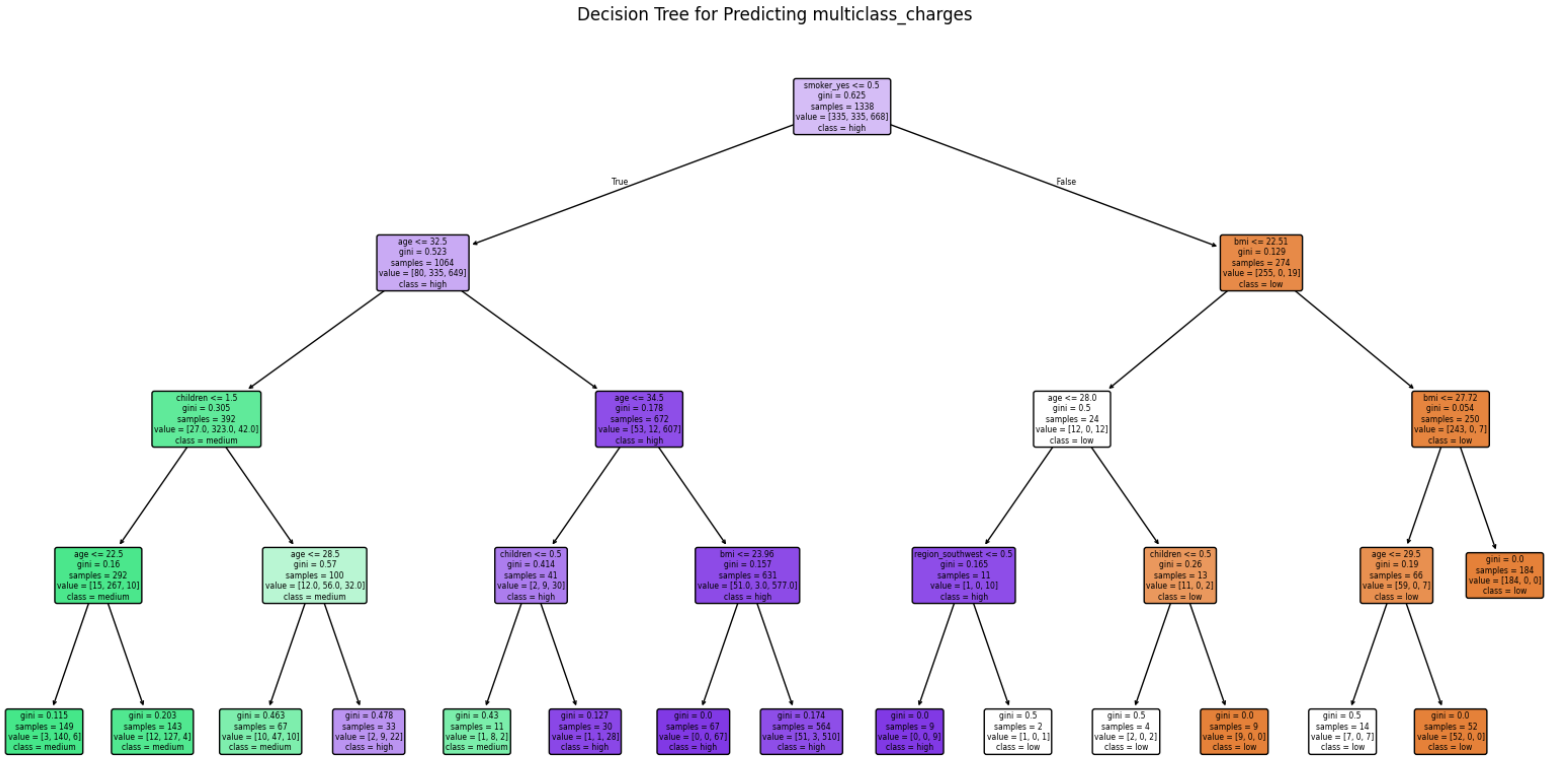
12. Model prediction:

age	gender	bmi	children	smoker	region	binary_charges
25	male	28.0	1	no	northeast	0.040098
45	female	35.2	3	yes	southeast	0.999164
32	male	30.5	0	no	northwest	0.051805
54	female	24.7	2	yes	southwest	0.999048
29	female	22.8	1	yes	southeast	0.995861

13. % of observations assigned to class ‘low’: 25.04
    % of observations assigned to class ‘medium’: 49.92
    % of observations assigned to class ‘high’: 25.04

14. # of leaf nodes in the decision tree: 15

15. Decision tree plot:



Explanation of path decision:

- **Root Node**: smoker_yes <= 0.5
  -> If **not a smoker** (True), move **left**.
- **Next Node**: age <= 32.5
  -> If **age ≤ 32.5**, move **left**.
- **Next Node**: children <= 1.5

-> If **children ≤ 1.5**, move **left**.

- **Next Node**: age <= 22.5
  -> If **age ≤ 22.5**, move **left** to the **leaf node**.
- **Leaf Node**:
  -> Class = medium

16. Model prediction:

age	gender	bmi	children	smoker	region	multiclass_charges
25	male	28.0	1	no	northeast	low
45	female	35.2	3	yes	southeast	high
32	male	30.5	0	no	northwest	low
54	female	24.7	2	yes	southwest	high
29	female	22.8	1	yes	southeast	high