Assignment 2

Maximum points: 20 Due date: 21st Nov. 2025 at 23:00

Guidelines

- Individual Assignment: This is an individual assignment. Please do not seek help from others or collaborate with classmates.
- Problem Understanding: Carefully read the problem description and each question before starting your analysis.
- Choice of Software: You can use either R or Python to perform your analysis.
- Code Requirements: Your code should be well-structured and free of errors. Please use clear and descriptive variable names.
- Deliverables: Submit the following two files on Canvas:
    1. Code file (.R/.Rmd/.py/.ipynb): This file should contain the complete executable code for the analysis.
    2. Report file (.pdf): This document should contain your answers to all the questions asked below. Please use the solution template provided on Canvas (Assignment_2_Template.docx) to answer the questions. On the cover page, please write your name, student number, and date. After writing your answers, submit the document as a PDF file.
- Plagiarism Warning: Ensure all the code and analysis are your original work. Plagiarism will lead to disqualification and academic consequences.
- Late Submissions: Late submissions will be progressively penalized.
- Points Distribution: The distribution of points and the grading rubric for the questions is provided at the end of this document.

Context

The dataset insurance.csv contains anonymized information about a sample of individuals and their medical charges. Each entry includes demographic information, lifestyle factors, and the resulting medical costs for the individual. Key factors such as age, BMI, and smoking status are included, which will allow you to investigate patterns and correlations that impact healthcare expenses.

Data dictionary

Please access the file insurance.csv from Canvas.

Column	Data type	Description
age	integer	Age of the primary beneficiary
gender	categorical	Gender of the beneficiary
bmi	numerical	Body Mass Index of the beneficiary
children	integer	# of children covered by the insurance policy
smoker	binary	Indicates whether the beneficiary smoked or not
region	categorical	Residential area of the beneficiary in the US
charges	numerical	Yearly medical costs billed by health insurance provider

Import the insurance data as a DataFrame and name it as insurance.

Exploratory Data Analysis

1. How much more do smokers pay in medical charges compared to non-smokers across different age groups? To explore this, start by dividing the dataset into the following age groups: [18, 30), [30, 40), [40, 50), [50, 60), and [60, 70). For each age group, calculate the average medical charges separately for smokers and non-smokers. Create a bar chart that shows the average charges for smokers and non-smokers within each age group. Finally, identify the age group with the largest difference in charges between smokers and non-smokers, and report the value of this maximum difference. Do all the computations in R/Python.

2. Generate a correlation matrix for the numerical variables in the insurance dataset, including age, bmi, children, and charges, and visualize it using a heatmap. Identify which variable has the strongest correlation with charges.

Linear Regression

3. Build a linear regression model using all available variables to predict charges (the target variable). Use age, gender, bmi, children, smoker, and region as predictor variables. After building the model, report which variables are statistically significant predictors of charges. Identify the significance level (e.g., 0.01, 0.05, or 0.1) for each significant variable. Handle the categorical variables carefully.

4. Provide interpretations for the coefficients of age, gender, bmi, and region.

   Note: The region variable has multiple (4) categories, resulting in three binary (dummy) variables in the model, each comparing a specific region to a reference region. Identify the reference region and interpret the coefficients of each dummy variable for region.

5. Let us now assess whether the model meets some of the core assumptions.
   a. Create a histogram of residuals to visually inspect its distribution and comment on whether the residuals appear approximately normal.
   b. Plot the residuals vs. the predicted charges to assess homoscedasticity and comment on whether the homoscedasticity assumption is met.

6. Create a new target variable named log_charges by applying a logarithmic transformation to the original charges variable. Refit the linear regression model using log_charges as the target variable instead of charges. For this model,
   a. Interpret the coefficient of age.
   b. Create a histogram of residuals to visually inspect its distribution and comment on whether the residuals appear approximately normal.
   c. Plot the residuals vs. the predicted charges to assess homoscedasticity and comment on whether the homoscedasticity assumption is met.

7. Evaluate the overall fit of the model (developed in question 6) by reporting the R-squared and adjusted R-squared value. Explain the meaning of these statistics in the context of this model.

8. Using your model (developed in question 6), predict charges for five new observations based on the characteristics listed below:

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

Logistic Regression

9. Let μ denote the mean of charges. Create a new variable binary_charges as follows: For each observation, assign binary_charges a value of 0 if charges is less than or equal to μ and a value of 1 if charges is greater than μ. Report the % of observations assigned to each class (0 or 1) of binary_charges.

10. Build a logistic regression model using age, gender, bmi, children, smoker, and region as predictor variables and binary_charges as the target variable. After building the model, report which variables are statistically significant predictors of charges. Identify the significance level (e.g., 0.01, 0.05, or 0.1) for each significant variable.

11. Provide interpretations for the coefficients of age, gender, bmi, and region.

    Note: The region variable has multiple (4) categories, resulting in three binary (dummy) variables in the model, each comparing a specific region to a reference region. Identify the reference region and interpret the coefficients of each dummy variable for region.

12. Using the logistic regression model predict binary_charges for five new observations based on the characteristics listed below:

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

Decision Tree

13. Let q_1 and q_3 denote the first and third quartile of the charges variable. Create a new variable multiclass_charges as follows: For each observation, assign multiclass_charges a value of 'low' if charges is less than or equal to q_1, a value of 'medium' if charges is greater than q_1 and less than or equal to q_3, and a value of 'high' if charges is greater than q_3. Report the % of observations assigned to each class ('low', 'medium', 'high') of multiclass_charges.

14. Build a decision tree model to predict multiclass_charges using age, gender, bmi, children, smoker, and region as predictor variables. Set a maximum depth of 4 for the decision tree to avoid overfitting. Report the number of leaf nodes in the tree.

15. Plot the decision tree. Pick one of the paths from the root node to a leaf node and explain how the decision tree makes the classification decision on that path.

16. Using the decision-tree model predict multiclass_charges for five new observations based on the characteristics listed below:

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

Points distribution and grading rubric

Question	Point(s)	Grading criteria		
		Correct answer	Incorrect answer, but coding logic partially correct	Incorrect answer and coding logic
1	2	2	1	0
2	1	1	0.5	0
3	1	1	0.5	0
4	1	1	0.5	0
5a	1	1	0.5	0
5b	1	1	0.5	0
6a	1	1	0.5	0
6b	1	1	0.5	0
6c	1	1	0.5	0
7	1	1	0.5	0
8	1	1	0.5	0
9	1	1	0.5	0
10	1	1	0.5	0
11	1	1	0.5	0
12	1	1	0.5	0
13	1	1	0.5	0
14	1	1	0.5	0
15	1	1	0.5	0
16	1	1	0.5	0