Assignment-based Subjective Questions

1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)

types of categorical data: Ordinal, Nominal and Binary

Different types of statistical methods are used based on what type of categorical data it is.

2. Why is it important to use drop_first=True during dummy variable creation? (2 mark)

Drop-first is used to avoid multicollinearity issue between the variables

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

temp

4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

via Residual Analysis

the errors must be normally distributed

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)

   y = mx + c, a general statement of maths is used to represent liner regression

   Linear regression is a statistical algorithm used to model the relationship between a dependent variable and one or more independent (predictor) variables. It assumes a linear relationship between the variables,

   Linear regression can be represented as $y=\beta0+\beta1x+\varepsilon$

   y is the dependent variable.

   x is the independent variable (predictor).

   β0 is the intercept (constant term).

   β1 is the coefficient for the independent variable x.

   e is the error term

The steps to compute linear regression:

- Create X and y
- Create train and test sets (70-30, 80-20)
- Train your model on training set (i.e., learn the coefficient)
- Evacuate the model (training set, test set)

2. Explain the Anscombe's quartet in detail. (3 marks)

comprises a set of four dataset, having identical descriptive statistical properties in terms of means, variance, R-Squared, correlations, and linear regression lines but having different representations when we scatter plot on graph.

The four datasets that make up Anscombe's quartet each include 11 x-y pairs of data.
When plotted, each dataset seems to have a unique connection between x and y, with unique variability patterns and distinctive correlation strengths.
each dataset has the same summary statistics, such as the same x and y mean and variance, x and y correlation coefficient, and linear regression line.

used to illustrate the importance EDA and the drawbacks of depending only on summary statistics

3. What is Pearson's R? (3 marks)

Pearson correlation coefficient, also known as Pearson R statistical test, measures the strength between the different variables and their relationships.

whenever any statistical test is conducted between the two variables, it is always a good idea for the person analysing to calculate the value of the correlation coefficient to know how strong the relationship between the two variables is.

Range from the value +1 to the value -1, where +1 indicates the perfect positive relationship between the variables considered, -1 indicates the perfect negative relationship between the variables considered, and 0 value indicates that no relationship exists between the variables considered.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

Scaling is a method by which values are transformed in a dataset of similar scale.

Scaling is performed to ensure all features contribute equally to the model and to avoid the domination of features with larger values.

Normalized scaling and standardized scaling:

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?

(3 marks)

The greater the VIF, the higher the degree of multicollinearity. In the limit. when multicollinearity is perfect (i.e., the regressor is equal to a linear combination of other regressors), the VIF tends to infinity.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

(3 marks

A Q–Q plot (quantile-quantile plot) is a probability plot, a graphical method for comparing two probability distributions by plotting their quantiles against each other.

Used to graphically analyse and compare two probability distributions by plotting their quantiles against each other.

The most fundamental question answered by Q-Q plot is:

Is this curve Normally Distributed?