Assignment-based Subjective Questions

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

Ans:

1. Clear or misty weather had a better chance of bike being rented.
2. Among seasons, summer(march) and fall are preferred season for biking. (Sept, October)
3. Over the years, the % of people preferring rental bikes have increased.
4. It's more likely that a customer will rent a bike on a holiday compared to when its not a holiday.

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

Ans:

It is to prevent the dummy variable trap. Every time we use the pandas.get_dummies () we introduce multicollinearity because 1 out of n columns can be explained by n-1 columns. Hence, we drop the 1st column. Also, n-1 columns are good enough to explain all the categories.

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

Ans: Temp has the highest correlation.

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

Ans:

1) There is linearity between the feature and the target variables. Eg: Temp vs Cnt is linear.
2) Error terms are normally distributed (visualising error terms on a distribution plot)
3) There is Homoscedasticity (variance of the residual terms in constant)
4) Mean of residuals is 0.

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

Ans:

Temp
Windspeed
Holiday
Year

General Subjective Questions

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

Ans:

The linear regression (LR) is used to model the relationship between one or more independent feature variables and a target variable(dependent) by fitting a straight line or a linear equation to the given data.
The LR Model assumes that the relationship between the features and the target variables can be explained by the equation of a straight line.

Equation for linear regression: y = mx + c , where

y = dependent variable(target)

X= independent variables (features)

m= slope (coefficients)

c = y-intercept

In training phase, the algorithm adjusts the values of m and c such that to minimize the difference between predicted and actual values of y in the training data set.

The commonly used loss function in LR is mean squared error (MSE)

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Where n is total number of data points.

The goal is to find the values of m and c that minimize the MSE. This Is done by iteratively adjusting the m and c values by using optimization techniques like **Gradient descent.**

Finally, the Linear regression model is evaluated by using metrics like R-squared which measures the proportion of the variance in the dependent variable that is predictable from the independent variables.

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

Ans:

Anscombe's quartet is a set of four datasets that have nearly identical simple descriptive statistics (mean, variance, correlation, etc.) but differ significantly when graphed, emphasizing the importance of visualizing data.
Created by Francis Anscombe in 1973 to demonstrate the limitations of summary statistics in describing datasets.

It consists of 4 pairs of x and y variables, each with 11 data points:

- Linear: Linear relationship with minimal variance and no outliers.
- Non-linear: Non-linear relationship with one outlier.
- Another linear: Linear relationship with an outlier.
- Vertical line: Perfectly linear relationship but with a single outlier significantly altering the regression line.

Despite having identical summary statistics, the datasets exhibit vastly different characteristics when visualized, highlighting the importance of exploratory data analysis and data visualization in understanding data patterns and relationships.

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

Ans:

Pearson's R is a measure of linear correlation between two variables.
R ranges from -1 to 1
R = 1 perfect positive correlation
R = -1 perfect negative correlation
R = 0 No correlation
Pearson's correlation assumes that there is a linear relationship between variables in the data set. It is also sensitive to outliers.

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

Ans:

Scaling refers to the process of transforming numerical features to a common scale, typically to a range or distribution that allows for easier comparison and processing by machine learning algorithms.

Many ML algorithms are sensitive to the scale of input features. Scaling ensures that no feature dominates the learning process due to its larger magnitude. Scaling can make the interpretation of coefficients or feature importance more meaningful.

Standardization - Transforms features to have a mean of 0 and a standard deviation of 1.

Normalization: Scales features to have unit norm (i.e., length of 1).

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

Ans:

Some of the cases when VIF can be infinite are:

- **Perfect multicollinearity** - When one feature can be perfectly predicted from all other features in the model, it leads to infinite VIF.
- If two or more predictor variables are linearly dependent, meaning one can be expressed as a linear combination of others, it leads to perfect multicollinearity and thus infinite VIF.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

Ans:

Q-Q plot, or Quantile-Quantile plot, is a graphical tool used to assess whether a dataset follows a particular probability distribution (e.g., normal distribution).

Importance in LR:

- Assumption checking: Q-Q plots are used to assess the assumption of normality in linear regression residuals.
- Residual analysis: The residuals should ideally follow a normal distribution. Q-Q plots help visualize whether this assumption holds.