Assignment Questions

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)**

   a. `season`, `yr`, `mnth`, `weathersit` is significant for `cnt`

   b. `holiday`, `weekday`, `workingday` is insignificant for `cnt`

   c. on `holiday`, `cnt` demands increase more

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

   a. For `n` level categorical variable, we can infer the rest variable by previous `n-1`. It does not affect the result.

   b. The training process will be more efficient with less variable.

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

   `atemp` seems has the highest correlation with `cnt`.

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

   a. Residual Analysis the train data.

   b. Check the `VIF` of the final train data, if all the variables' VIF are less than `5`.

   c. Check R squared of the final model and the prediction, if the values is greater than `80%`.

   d. Check the difference of R squared between the final model and the prediction, check if the difference is greater than `5%`.

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

   `temp`, `yr` and `LightSnow`.

General Subjective Questions

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

   Following are steps of the linear regression algorithm:

   1. **Data Preparation:** Collect dataset and divide understand it, which one is features and witch one is target variable.
   2. **Cost Function:** Define a cost function that quantifies the difference between the predicted values and the actual values.
   3. **Gradient Descent:** Adjust the weights to minimize the cost function by taking partial derivatives of the cost function with respect to each weight. These derivatives indicate how the cost function changes as we change the weights.
   4. **Training the Model:** Start with initial random values for the weights and iteratively update the weights using gradient descent until we reach a point where the cost function is minimized or falls below a certain threshold.
   5. **Prediction:** Use the trained model to make predictions on new, unseen data.
   6. **Evaluation:** Using the mean squared error (MSE), root mean squared error (RMSE), mean absolute error (MAE), and R-squared (coefficient of determination) to evaluate the model.

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

   Anscombe’s quartet is a set of four datasets that have nearly identical statistical properties, despite their vastly different visual representations. The quartet demonstrate the importance of visualizing data before analyzing it.

   Each dataset in the quartet consists of 11 (x, y) pairs, and all have the same mean, variance, correlation coefficient, and linear regression line. However, the datasets

have very different distributions and patterns when plotted.

The first dataset appears to have a relatively strong linear relationship, while the second dataset has a non-linear relationship that would be overlooked if we relied solely on measures of central tendency and correlation. The third dataset has an outlier that has a significant effect on the regression line, and the fourth dataset has a perfect fit with a curve that is not linear.

The Anscombe's quartet demonstrates that descriptive statistics, such as mean and correlation, can be misleading when used alone, and that visualizing data is crucial to obtaining a complete understanding of it.

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

   Pearson's R is a statistical measure that quantifies the linear relationship **between two continuous variables**. It assesses the **strength and direction** of the linear association between **two variables**.

   Pearson's R is calculated as the covariance of the two variables divided by the product of their standard deviations. It is defined within the range of `-1` to `1`, where:

   - A value of `1` indicates a perfect positive linear relationship, meaning that as one variable increases, the other variable increases proportionally.
   - A value of `-1` indicates a perfect negative linear relationship, meaning that as one variable increases, the other variable decreases proportionally.
   - A value of `0` indicates no linear relationship between the variables.

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

   - Scaling is the process of transforming the data so that it fits within a specific scale. This is done **to ensure that all the features contribute equally to the model** and that no feature dominates the others.
   - Scaling is performed to ensure that all features are on the same scale. **If the features are not on the same scale, the algorithm may give too much importance to the features with larger values**.

- Normalized scaling and Standardized scaling are two common scaling techniques.
    - Normalization scales the features between `0` and `1`.
    - Standardization scales the features so that they have a mean of `0` and a standard deviation of `1`.

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

   Infinite VIF values indicate a severe issue of multicollinearity in the regression model. It implies that the linear relationship between **the predictor variable and the other variables** can be expressed precisely by a combination of those other variables.

   In this case, if we use `casual` or `registerd` variables and use `cnt` as target variable, will get this issue. Because `cnt` was calculated from `casual` and `registered`.

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 is a graphical tool used to **compare two probability distributions.** It is commonly used in linear regression to assess whether the residuals follow a normal distribution.

   The Q-Q plot plots the quantiles of the residuals against the theoretical quantiles of a normal distribution. If the residuals are normally distributed, the points in the plot should follow a straight line. If the residuals are not normally distributed, the points in the plot will deviate from the straight line.