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?

Answer: The demad of bike is less in the month of spring when compared with other seasons.

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

Answer: drop_first=True is important to use, as it helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables.

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

Answer: has the highest correlation with the target variable is between temp and atemp variable with the predictor 'cnt'.

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

Answer:

- There must be a linear relationship
- No auto-correlation or independence
- No Multicollinearity
- Homoscedasticity
- Normal Distribution of error terms

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

Answer: The Top 3 features contributing significantly towards the demands of share bikes are:

- weathersit_Light_Snow (negative correlation).
- Yr (Positive correlation).
- temp(Positive correlation).

General Subjective

Questions 1. Explain the linear regression algorithm in detail.

Answer: Linear Regression is a machine learning algorithm based on supervised learning. It performs a regression task. Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting.

Linear regression performs the task to predict a dependent variable value (y) based on a given independent variable (x). So, this regression technique finds out a linear relationship between x (input) and y(output). Hence, the name is Linear Regression.

Hypothesis function for Linear Regression :

$$Y = mx + b$$

2. Explain the Anscombe's quartet in detail.

Answer: Anscombe's quartet comprises four data sets that have nearly identical simple descriptive statistics, yet have very different distributions and appear very different when graphed. Each dataset consists of eleven (x,y) points**.**

3. What is Pearson's R?

Answer: The Pearson correlation coefficient also known as Pearson's r, the Pearson product-moment correlation coefficient (PPMCC), the bivariate correlation,or colloquially simply as the correlation coefficient , is a measure of linear correlation between two sets of data. It is the ratio between the covariance of two variables and the product of their standard deviations; thus it is essentially a normalized measurement of the covariance, such that the result always has a value between −1 and 1. As with covariance itself, the measure can only reflect a linear correlation of variables, and ignores many other types of relationship or correlation. For example, one would expect the age and height of a sample of teenagers from a high school to have a Pearson correlation coefficient significantly greater than 0, but less than 1

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Answer: Feature scaling refers to putting the feature values into the same range. Scaling is extremely important for the algorithms considering the distances between observations like k-nearest neighbors. On the other hand, rule-based algorithms like decision trees are not affected by feature scaling.

In normalization, we map the minimum feature value to 0 and the maximum to 1. Hence, the feature values are mapped into the [0, 1] range:

$$z = x - \min(x)/\max(x) - \min(x)$$

In standardization, we don't enforce the data into a definite range. Instead, we transform to have a mean of 0 and a standard deviation of 1:

$$z = x - mu/sigma$$

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

Answer: If there is perfect correlation, then VIF = infinity. This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity.

An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables

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

Answer : Q-Q Plots (Quantile-Quantile plots) are plots of two quantiles against each other. A quantile is a fraction where certain values fall below that quantile. For example, the median is a quantile where 50% of the data fall below that point and 50% lie above it. The purpose of Q Q plots is to find out if two sets of data come from the same distribution. A 45 degree angle is plotted on the Q Q plot. if the two data sets come from a common distribution, the points will fall on that reference line.

If the two distributions being compared are similar, the points in the Q–Q plot will approximately lie on the line y = x. If the distributions are linearly related, the points in the Q–Q plot will approximately lie on a line, but not necessarily on the line y = x. Q–Q plots can also be used as a graphical means of estimating parameters in a location-scale family of distributions.

A Q–Q plot is used to compare the shapes of distributions, providing a graphical view of how properties such as location scale are similar or different in the two distributions.