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-

- Bike demand is higher in 2019 than in 2018.
- Bike demand is higher on non-holidays than holidays
- Bike demand is higher for clearer weather
- Median is same for working and non-working days but spread is higher for non-working days

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

Answer-

If we have categorical variable with n-levels, then we need to use n-1 columns to represent the dummy variables.

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

Answer-

'temp' and 'atem' has the highest correlation with the target variable 'cnt' which is 0.63

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

Answer-

Assumptions:

- There is a linear relationship between the independent variable, X and the dependent variable, y – this can be verified if the scatter plot between X and y fall on roughly a straight line.
- The residuals are independent – this can be verified by Durbin-Watson test. 'statsmodels' linear regression summary gives the DW value. Ideally DW=2
- The residuals have constant variance, should not have heteroscedasticity – this can be verified by creating a fitted value vs residual plot. If the plot shows a funnel shape pattern, then heteroscedasticity is present
- The residuals are normally distributed – this can be verified by the distribution plot and see if it's a normal distribution.
- No Multicollinearity – this can be verified by calculating Variance Inflation Factor. VIF<5 means little to moderate multicollinearity which can be allowed.

5. 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 explaining the demand of the shared bikes are:

- Temp
- Light Snow/Rain
- year

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Answer –

Linear regression algorithm is a supervised learning based algorithm. It's used to predict values within a continuous range, a target value is predicted based on independent variables. There are 2 types, Simple Linear Regression predicts using a single variable, "y=mx + c" where c is a constant and m is the intercept/slope of the model. Multiple Linear Regression is when multiple independent variable predict the target, "y=B1X1 + B2X2 + B3X3 + B0" where Bo is the constant and B1,B2, etc are the respective coefficient values

2. Explain the Anscombe's quartet in detail.

Answer –

Anscombe's quartet is a group of 4 datasets with almost identical simple statistic properties, yet appear different when graphed. Each consists of 11 data points. They were constructed to demonstrate both the importance of graphing data before analyzing it and the effect of outliers on statistical properties. Statistician Frank Anscombe constructed them in 1973.

3. What is Pearson's R?

Answer –

Pearson's R is a statistic that measures the linear correlation between 2 variables . Like all correlations, it also has a numerical value that lies between -1.0 and +1.0. Pearson's correlation coefficient is the covariance of the two variables divided by the product of their standard deviations. Pearson's r, the Pearson product-moment correlation coefficient (PPMCC), or bivariate correlation. Whenever we discuss correlation in statistics, it is generally Pearson's correlation coefficient. Pearson's Correlation Coefficient is named after Karl Pearson. He formulated the correlation coefficient from a related idea by Francis Galton in the 1880s.

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

Answer –

Scaling is a step of data pre-processing which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm.

If scaling is not done then algorithm only takes magnitude in account and not units hence incorrect modelling. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude. It is important to note that scaling just affects the coefficients and none of the other parameters.

Normalized scaling brings all of the data in the range of 0 and 1.

Standardized scaling replaces the values by their Z scores. It brings all of the data into a standard normal distribution which has mean (μ) zero and standard deviation one (σ).

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

Answer –

The formula of VIF = 1/(1-R^2). VIF is infinite means there is a perfect correlation between 2 variables. The solution for this is to drop one of the variables from the dataset which is causing perfect multicollinearity.

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

Answer –

The quantile-quantile (q-q) plot is a graphical technique for determining if two data sets come from populations with a common distribution. A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set.

A Q–Q plot is used to compare the shapes of distributions, providing a graphical view of how properties such as location, scale, and skewness are similar or different in the two distributions and the sample sizes do not need to be equal.