Assignment-based Subject Qs

Q1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

A1. There are a few categorical variables in the dataset. Some variables like Weather Situation having values Misty/Cloudy or Light Rain/Snowy negatively effect the bike demand. While seasons Summer & Winter, month September, and day of week Saturday positively effect the demand.

Q2. Why is it important to use drop_first=True during dummy variable creation?

A2. Its easy to represent a categorical variable having n values with n-1 dummy variables. Example:

Consider furnished column having values unfurnished, furnished, and semi-furnished. If we drop unfurnished and just keep the other 2 dummy columns, then value combination 00 will represent unfurnished.

Q3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

A3. temp variable has highest correlation with target cnt variable

Q4. How did you validate the assumptions of Linear Regression after building the model on the training set?

A4. I have done residual analysis and checked if

1. The residuals are centered around mean
2. No correlation between residuals and y_train_pred
3. Also, homoscedasticity was also checked between y_train vs y_train_pred

Q5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

A5. Temp, Year, season Light Rain or Snowy contribute significantly towards explaining the demand

General Subjective Questions

Q1. Explain the linear regression algorithm in detail.

A1. Linear regression is a statistical method used to model the relationship between a dependent variable and one or more independent variables. The goal is to find the best-fitting line (or hyperplane in higher dimensions) that describes the relationship between these variables.

Key Ideas:

1. Dependent Variable (y): The variable we are trying to predict or explain.
2. Independent Variables (X): The variables we use to make predictions.
3. Regression Coefficients (β): The parameters of the model that are estimated from the data.

Simple Linear Regression

Simple linear regression deals with one independent variable. The model can be represented as: $y=\beta0+\beta1X+\epsilon$ where:

- y is the dependent variable.
- X is the independent variable.
- β0 is the y-intercept of the regression line.
- β1 is the slope of the regression line.
- ϵ is the error term (the difference between the actual and predicted values).

Multiple Linear Regression

Multiple linear regression deals with multiple independent variables. The model can be represented as: $y=\beta0+\beta1X1+\beta2X2+\cdots+\beta nXn+\epsilon$ where:

- y is the dependent variable.
- X1,X2,…,Xn are the independent variables.
- β0 is the y-intercept.
- β1,β2,…,βn are the coefficients for the independent variables.
- ϵ is the error term.

Assumptions of Linear Regression

1. Linearity: The relationship between the dependent and independent variables is linear.
2. Independence: The residuals (errors) are independent.
3. Homoscedasticity: The residuals have constant variance at every level of XXX.
4. Normality: The residuals of the model are normally distributed.

Q2. Explain the Anscombe's quartet in detail.

A2. Anscombe's quartet is a collection of four datasets that have nearly identical simple descriptive statistics, yet appear very different when graphed. It was created by the statistician Francis Anscombe in 1973 to demonstrate the importance of graphing data before analyzing it and to illustrate the effect of outliers and the influence of other statistical properties on data analysis.

Q3. What is Pearson's R?

A3. Pearson's R, also known as the Pearson correlation coefficient, is a measure of the linear correlation between two variables. It is named after Karl Pearson, who developed the formula. The coefficient quantifies the strength and direction of the linear relationship between two continuous variables.

Q4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling

A4. Scaling is a data preprocessing technique used to adjust the range of features in a dataset. The goal is to transform the data such that it fits within a specific scale. Its performed for ML algorithm efficiency, model interpretability, and numerical stability.

Q5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?

A5. When R2 becomes 1 for a variable (when explained by other independent variables), then VIF becomes infinite. Causes:

1. When one predictor variable can be perfectly predicted using a linear combination of other predictor variables.
2. Including redundant variables that do not add new information but are perfectly correlated with existing predictors.

Q6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

A6. A Q-Q (Quantile-Quantile) plot is a graphical tool to assess if a dataset follows a particular theoretical distribution, such as the normal distribution. It plots the quantiles of the dataset against the quantiles of the specified theoretical distribution. If the points on the Q-Q plot form a roughly straight line, the dataset follows the theoretical distribution.

In linear regression, the assumptions about the residuals (errors) play a crucial role in validating the model. The Q-Q plot helps in assessing these assumptions, particularly the assumption of normality, detection of outliers, and model diagnostics.