Assignment-based Subjective Questions

Question 1 – From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

Answer – From the analysis of the categorical variables –

1. We can infer that the presence of certain categories in the data could have a significant impact on the demand of shared bikes.
2. For example, the presence of holidays or festivals could increase the demand for shared bikes as people might prefer using them for short trips or leisure activities.
3. On the other hand, the presence of bad weather conditions could decrease the demand for shared bikes as people might prefer using other modes of transportation.
4. The variable 'yr' also shows a clear increase in demand from 2018 to 2019.
5. Overall, the categorical variables provide valuable insights into the demand for shared bikes, and they should be included in the regression model.

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

Answer – It is important to use drop_first=True during dummy variable creation to avoid the dummy variable trap.

1. When creating dummy variables for categorical variables with k categories, we need to create k-1 dummy variables to avoid perfect multicollinearity.
2. If we don't use drop_first=True, then one of the categories will be used as the reference category and will not have a corresponding dummy variable. This can lead to incorrect estimates of the regression coefficients.

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

Answer – Based on the pair-plot, the variable 'temp' has the highest correlation with the target variable 'cnt'. This is intuitive, as temperature is likely to have an impact on the demand for shared bikes.

- Higher temperatures are generally associated with higher demand for shared bikes. This is an important variable to include in the regression model since temperature has a significant impact on demand.

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

Answer – To validate the assumptions of Linear Regression –

1. We used diagnostic plots to check for linearity, homoscedasticity, and normality of residuals.
2. We also calculated the Variance Inflation Factor (VIF) to check for multicollinearity, and ensured that the VIF was below 5 for all variables.
3. We checked for autocorrelation by plotting the residuals against time and ensuring that there was no clear pattern.
4. For example, we can use a scatterplot of the residuals versus the predicted values to check for homoscedasticity. We can also use a Q-Q plot of the residuals to check for normality.
5. Additionally, we can use a correlation matrix to check for multicollinearity and a histogram of the residuals to check for outliers.

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

Answer – Based on the final model, the top 3 features contributing significantly –

1. Towards explaining the demand of the shared bikes are 'temp', 'weathersit_3', and 'yr_1'.
2. These variables have the highest coefficients and therefore have the most significant impact on the demand for shared bikes.
3. This implies that temperature, weather conditions, and season are the most important factors affecting the demand for shared bikes.

General Subjective Questions

Question 1 – Explain the linear regression algorithm in detail.

Answer –

1. Linear regression is a supervised machine learning algorithm used for regression tasks, where the target variable is continuous.
2. The algorithm fits a linear equation to the data, which can be used to predict the target variable based on the independent variables.
3. Linear regression is a statistical method used to model the linear relationship between a dependent variable and one or more independent variables.
4. The algorithm involves finding the best-fitting line or hyperplane that minimizes the sum of the squared distances between the observed values and the predicted values. The equation for a simple linear regression model is:
   $y = \beta0 + \beta1x + \varepsilon$
   where y is the dependent variable, x is the independent variable, β0 is the y-intercept, β1 is the slope, and ε is the error term. The algorithm uses a method called least squares to find the values of β0 and β1 that minimize the sum of the squared errors.

Question 2 – Explain the Anscombe's quartet in detail.

Answer –

1. Anscombe's quartet is a set of four datasets that have the same summary statistics but different distributions (mean, variance, correlation, and regression line).
2. The dataset was created by Francis Anscombe to demonstrate the importance of visualizing data before analysing it.
3. The four datasets have the same mean and variance for both the dependent and independent variables, and they have the same correlation coefficient.
4. However, when plotted on a scatterplot, it is clear that they have different distributions and relationships between the variables. This highlights the importance of visualizing data before drawing conclusions based on summary statistics alone.

Question 3 – What is Pearson's R?

Answer –

1. Pearson's R is a measure of the linear correlation between two variables. It ranges from -1 to 1, where -1 indicates a perfect negative correlation, 0 indicates no correlation, and 1 indicates a perfect positive correlation.
2. Pearson's R is calculated as the covariance of the two variables divided by the product of their standard deviations.
3. It is a widely used measure of correlation and is often used in linear regression to determine the strength of the relationship between the independent and dependent variables.

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

Answer –

- Scaling is the process of adjusting the values of the independent variables so that they have the same range and distribution.
    1. There are two main types of scaling: normalized scaling and standardized scaling.
    2. Normalized scaling scales the values to a range of 0 to 1, while standardized scaling scales the values to have a mean of 0 and a standard deviation of 1.
- Scaling is performed to ensure that all variables are on the same scale, which can improve the performance of machine learning algorithms that are sensitive to the scale of the input variables. This is done to avoid biasing the regression model towards variables with larger values.
- The difference between the two is that normalized scaling preserves the original distribution of the data, while standardized scaling transforms the data to have a normal distribution.

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

Answer –

1. The value of VIF (Variance Inflation Factor) can be infinite if there is perfect multicollinearity between the independent variables. This occurs when one independent variable is a linear combination of the other independent variables.

2. In such cases, it is impossible to estimate the regression coefficients accurately, and the model is said to be ill-conditioned. To avoid this, it is recommended to remove one of the correlated variables or use techniques such as ridge regression or principal component regression.
3. Multicollinearity occurs when two or more independent variables are highly correlated, making it difficult to determine the individual effect of each variable on the dependent variable. In the case of perfect multicollinearity, one independent variable can be expressed as a linear combination of the other independent variables, causing the design matrix to become singular and the VIF to become infinite. To address this issue, one or more of the correlated variables can be removed or combined, or regularization techniques such as ridge regression can be used.

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

Answer – A Q-Q plot (Quantile-Quantile plot) is a graphical tool used to check the normality of the residuals in a linear regression model. It compares the quantiles of the residuals to the quantiles of a normal distribution.

1. If the residuals are normally distributed, the points on the Q-Q plot will form a straight line. Deviations from a straight line indicate that the residuals are not normally distributed, and the model assumptions are violated.
2. The Q-Q plot is an important diagnostic tool in linear regression and is used to check the validity of the model assumptions.
3. The Q-Q plot is useful in identifying the presence of outliers, skewness, and other non-normality issues. Addressing these issues can improve the reliability and predictive power of the linear regression model.