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)

Answer=> According to Analysis, 'weathersit' & 'season' are categorical variables from the dataset.
As per the change in these variables, it would impact change in target variable 'cnt',
e.g Summer/winter could be reason the 'cnt' of people would rent a bike based on climate conditions Or Based on meteorological conditions, people would decide the mindset to avail bike on rent or not.

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

Answer=> It is important to use drop_first=True during dummy variable creation as It would help in reducing one extra column from the model analysis.

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

Answer=> 'atemp' has the highest correlation with the target variable

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

Answer=> On Training Set, while validating the assumptions of Linear Regression, below variables were checked during building the model

A) the const coefficient is not approaching towards 1

B) Whatever the variable that has been added to the train set, its coefficient also not crossing one.

C) Along with R-squared values should be reaching towards +1

D) And the Probability should not reach to 1.

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

Answer=>The top 3 features contributing significantly towards explaining the demand of the shared bikes

1. atemp - feeling temperature in Celsius
2. yr - Year in which the request was demanded
3. mnth - month in which demand was raised

General Subjective Questions

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

Answer=>

Linear regression is commonly used for predictive analysis and modeling based on supervised learning.

Linear Regression performs a regression task where models a target prediction value based on independent variables.

It is mostly used for finding out the relationship between variables and forecasting.

step 1. Read & understand the data
step 2. Start Visualizing the data
step 3. check with paiplot about the multicollinearity within the variables
step 4. Visualize the categorical variables
step 5. Convert Categorical variables into dummy variables
step 6. Prepare Data with associated numerical values
step 7. Split the Dataset into Train& Test Set
step 8. Apply the scaling on variables apart from Dummy & unwanted variables
step 9. Check variables are highly co-related with each other & correlation coefficients
step 10. Divide into X and Y sets for the model building
step 11. Build a linear model
step 11-1. Add one variable into trainset & check the R-squared value, i.e check with another highly correlated variable
step 11-2. Repeat the step 11-1
step 12. Check OLS Regression Results, If R2-Square value is approaching to nearby 1. then training is done significantly
step 13. Do Residual analysis to check error terms also whether they are normally distributed or not.
step 14. Based on step no. 13, make predictions using final model, i.e scaler.transform
step 15. Divide X_test & y_test
step 16. Make prediction using latest model
step 17. Once step no. 16 is completed, Evaluate the model, by calculating R_Square value
step 18. Based on step no. 17, make the predictions based on target variables.

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

Answer=> Anscombe's Quartet is a group of four data sets which are nearly identical in simple descriptive statistics,

but there are some peculiarities in the dataset that fools the regression model if built.

They have very different distributions and appear differently when plotted on scatter plots.

Datasets can be described as:

Dataset 1: Fits the linear regression model pretty well.

Dataset 2: Could not fit linear regression model on the data quite well as the data is non-linear.

Dataset 3: Shows the outliers involved in the dataset which cannot be handled by linear regression model

Dataset 4: Shows the outliers involved in the dataset which also cannot be handled by linear regression model

Anscombe's quartet helps to understand the importance of data visualization.

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

Answer=> Pearson's R was developed by Karl Pearson and it is a correlation coefficient which is a measure of the strength of a linear association between two variables and it is denoted by 'r'. it has a value between +1 and −1, where 1 is total positive linear correlation, 0 is no linear correlation, and −1 is total negative linear correlation.

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

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.

Scaling is performed because often collected data set contains features highly varying in magnitudes and range. If scaling is not done, then algorithm only takes magnitude in account and not units hence incorrect modelling. To overcome this, scaling to bring all the variables to the same level of magnitude.

Scaling just affects the coefficients and none of the other parameters like t-statistic, F-statistic, p-values, R-squared, etc.

Normalization Scaling brings all of the data in the range of 0 and 1.

Whereas Standardization Scaling replicates the values by their Z scores.

It brings all of the data into a standard normal distribution which has mean zero and standard deviation one.

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

Answer=> The value of VIF is calculated by the below formula:

VIFi = 1/1-(Ri_Square)

Where, ‘i’ refers to the ith variable.

If Ri_squared value is equal to 1 then the denominator of the above formula become 0 and the overall value become infinite.

It happens because of perfect correlation in variables.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.(3 marks)

Answer=> The Q-Q plot or quantile-quantile plot is a graphical technique for determining if two data sets come from populations with a common distribution.

A Q-Q plot is a scatterplot created by plotting two sets of quantiles against one another.

If both sets of quantiles came from the same distribution, the points forming a line that’s roughly straight.

Use of Q-Q plot in Linear Regression: The Q-Q plot is used to see if the points lie approximately on the line. If they don’t, it means, our residuals aren’t Gaussian (Normal) and thus, our errors are also not Gaussian.

Importance of Q-Q plot: Below are the points:

1. The sample sizes do not need to be equal.
2. Many distributional aspects can be simultaneously tested. For example, shifts in location, shifts in scale, changes in symmetry, and the presence of outliers.
3. The q-q plot can provide more insight into the nature of the difference than analytical methods.