Assignment-based Subjective Questions and Answers

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

After performing univariant analysis on season, year, month , holiday, weekday, working day, weather sit categorical columns spring season have less effect on count on which we should work where fall season gives good demand, from 2019 year demand increase which indicate company should enhance their business, from July to December demands increases which means company should work more on years starting month and can launch more offers to create more demand, where holiday, weekday and working day does not give major effect on dependent variable but yes good weather means Clear, Few clouds, Partly cloudy, Partly cloudy gives biggest effect on bike sharing count which increases demand.

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

Using drop first=True when creating dummy variables in pandas is important to avoid multicollinearity, particularly when the dummy variables are used in regression models. By setting drop first=True, pandas drop the first dummy variable, leaving $n-1$ dummy variables. This reduces redundancy and avoids multicollinearity because the dropped category serves as a reference group. The remaining $n-1$ dummy variables represent the difference between each category and the reference group.

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

Looks like the temp and atemp has the highest correlation with the target variable cnt
temp and atemp are highly co-related with each other, atemp is derived from temp so we can drop atemp to resist multicollinearity.

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

We should follow below steps to ensure model is robust and that its predictions are reliable:

- Check Linearity with scatter plots and residuals vs. predicted values plots.
- Test for Independence using the Durbin-Watson test.
- Check Homoscedasticity using residual plots and tests like Breusch-Pagan.
- Assess Normality of residuals with Q-Q plots, histograms, and formal tests.
- Evaluate Multicollinearity with VIF and correlation matrices.
- Consider Endogeneity with domain knowledge and advanced techniques if needed.

By thoroughly validating these assumptions, you can ensure that your linear regression model is robust and that its predictions are reliable.

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

I found that temp, windspeed, season and one more feature is weather sit are the significant.

General Subjective Questions and Answers

1. Explain the linear regression algorithm in detail.

Linear regression is a supervised learning algorithm used for predicting a continuous target variable based on one or more input features (independent variables). The algorithm assumes a linear relationship between the input features and the target variable.

Key Points:

Model Representation: In linear regression, the relationship between the input features XXX and the target variable YYY is modeled as a linear equation:

$Y=\beta 0+\beta 1X1+\beta 2X2+\cdots+\beta nXn+\epsilon$

Here, β0 is the intercept, β1,…,βn are the coefficients (slopes) for each feature, and is the error term (residual).

Objective: The goal is to find the coefficients β0,β1,…,βn that minimize the difference between the predicted values Y^ and the actual values Y. This difference is measured using the Mean Squared Error

Optimization: The coefficients are typically estimated using the Ordinary Least Squares (OLS) method, which minimizes the sum of squared residuals (errors).

Prediction: Once the model is trained, it can predict the target variable YYY for new input data by plugging the input features into the linear equation.

In Summary:

Linear regression fits a line (or hyperplane in higher dimensions) to the data that best describes the relationship between the input features and the target variable. It does so by minimizing the sum of squared differences between the actual and predicted values.

2. Explain the Anscombe's quartet in detail.

Anscombe's quartet consists of four datasets with nearly identical summary statistics—such as mean, variance, and correlation—but very different distributions when plotted. Created by Francis Anscombe, the quartet demonstrates the importance of visualizing data, as relying solely on statistical metrics can be misleading. Despite their similar statistics, the datasets reveal different patterns: one follows a linear trend, one has a vertical line, one is influenced by an outlier, and one shows a nonlinear relationship. This highlights the necessity of graphing data to properly understand its characteristics and ensure appropriate model application.

3. What is Pearson's R?

Pearson's R, also known as the Pearson correlation coefficient, is a measure of the linear relationship between two variables. It ranges from -1 to 1, where:

- r = 1 indicates a perfect positive linear correlation (as one variable increases, the other also increases).

- r = -1 indicates a perfect negative linear correlation (as one variable increases, the other decreases).
- r = 0 indicates no linear correlation between the variables.

Pearson's R only measures linear relationships and does not capture nonlinear associations.

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

What is Scaling?

Scaling is a process in data preprocessing where the features (variables) of a dataset are adjusted to a common scale, usually to improve the performance of machine learning algorithms. The goal is to ensure that no single feature dominates others due to its scale, especially when features have different units or magnitudes.

Why is Scaling Performed?

Scaling is important for several reasons:

1. **Improves Algorithm Performance**: Many machine learning algorithms, such as gradient descent-based methods (e.g., linear regression, logistic regression, neural networks) and distance-based algorithms (e.g., k-nearest neighbors, support vector machines), perform better when the features are on a similar scale.

2. **Faster Convergence**: For algorithms like gradient descent, scaling the features can lead to faster convergence because the algorithm does not have to adjust disproportionately large steps for features with larger scales.

3. **Equal Feature Importance**: Scaling ensures that each feature contributes equally to the result, preventing features with larger scales from disproportionately influencing the model.

Difference Between Normalized Scaling and Standardized Scaling

1. Normalized Scaling (Min-Max Scaling):

Normalization, or min-max scaling, rescales the features to a fixed range, typically [0, 1]. Useful when the data needs to be bounded within a specific range, such as in image processing or neural networks with activation functions like sigmoid or tanh.

2.Standardized Scaling (Z-score Scaling):

Standardization rescales the features to have a mean of 0 and a standard deviation of 1, making the data follow a standard normal distribution. Commonly used in algorithms that assume normally distributed data, such as linear regression, logistic regression, and PCA (Principal Component Analysis).

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

The value of the Variance Inflation Factor (VIF) becomes infinite when there is perfect multicollinearity among the predictor variables in a regression model. Multicollinearity occurs when one predictor variable is a perfect linear combination of one or more other predictor variables.

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

A Q-Q (Quantile-Quantile) plot is a graphical tool used to compare the distribution of a dataset to a theoretical distribution, typically the normal distribution. It plots the quantiles of the sample data against the quantiles of the theoretical distribution.

Use and Importance in Linear Regression:

Assessing Normality: In linear regression, one key assumption is that the residuals (errors) are normally distributed. A Q-Q plot helps visually assess this assumption by plotting the residuals against a normal distribution. If the residuals are normally distributed, the points will approximately lie along a straight diagonal line.

Identifying Deviations: The Q-Q plot can reveal deviations from normality, such as skewness (points deviate at the ends) or heavy tails (points deviate in the middle). These deviations can indicate problems with the model, such as the need for transformations or the presence of outliers.

Improving Model Validity: By using a Q-Q plot to check the normality of residuals, you can ensure that the linear regression model meets its assumptions, leading to more reliable estimates and predictions.