Assignment-based Subjective Questions

From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)

Answer: Effect of categorical variables on bike sharing count :

1. Bike rentals are more in Fall season.
2. Bike rentals are slightly more on weekdays.
3. Bike rental median is more in 2019 than in 2018.
4. Bike rentals are more on Friday and Thursday
5. Bike rentals are more on Non Holiday
6. Bike rentals are more on May, July, August,October
7. Bike rentals are ore on clear weather condition (Clear, Few clouds, Partly cloudy, Partly cloudy)
8. Days with high humidity and strong windspeeds have a negative impact on bike usage

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

Answer :Using drop_first=True during the creation of dummy variables is crucial in regression analysis for several reasons. It helps prevent multicollinearity by eliminating the perfect correlation between dummy variables, ensuring stable and reliable coefficient estimates. Additionally, it enhances model interpretability by making coefficients represent the change in the dependent variable concerning specific categories rather than relative to a reference category. This simplifies the model, reducing the number of variables and making estimation more efficient, particularly in cases involving numerous categories. Most importantly, it helps avoid the "dummy variable trap," a situation where the inclusion of all dummy variables for a categorical feature leads to model estimation issues. In essence, drop_first=True is a best practice that ensures more accurate and meaningful results in regression analysis.

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

Answer :Column 'atemp'

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

Answer: Validating the assumptions of Linear Regression is a crucial step to ensure the model's reliability and interpretability. After building the model on the training set, you can use various diagnostic techniques to assess these assumptions.

Linearity Assumption: We Created a scatterplot of the residuals (predicted values) against the fitted values. A random, evenly distributed pattern indicates linearity. Deviations may suggest non-linearity.

Normality of Residuals: Here we Created a histogram and Q-Q plot of residuals. These should roughly follow a normal distribution. No or Little Multicollinearity: we Calculated the Variance Inflation Factor (VIF) for each predictor. VIF values greater than 10 suggest high multicollinearity

There are other methods : Use the Durbin-Watson test to check for autocorrelation. Values around 2 indicate no autocorrelation, deviations suggest otherwise. Homoscedasticity (Constant Variance): Inspect the residuals vs. fitted values plot. Ensure the spread of residuals is roughly constant across fitted values. Use the Goldfeld-Quandt test to formally test for heteroscedasticity.

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

Answer : yr (Year): The coefficient for the "yr" variable is 0.2305 with a very low p-value ($p < 0.001$). This indicates that the year has a highly significant positive impact on bike demand. As the year increases, the demand for shared bikes is expected to increase substantially.

temp (Temperature): The coefficient for the "temp" variable is 0.5096 with a very low p-value ($p < 0.001$). This suggests that temperature has a highly significant positive influence on bike demand. As the temperature rises, the demand for shared bikes is expected to increase significantly.

Best (Weather Category - Best): The coefficient for the "Best" weather category is 0.2475 with a very low p-value ($p < 0.001$). This implies that weather conditions categorized as "Best" have a highly significant positive effect on bike demand. When the weather is clear, with few clouds or partly cloudy, bike demand is expected to be significantly higher.

These three features, "yr," "temp," and "Best" weather conditions, are the top contributors to explaining the demand for shared bikes in the model. Their positive coefficients and low p-values indicate their importance and significance in influencing bike rental demand.

General Subjective Questions

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

Answer : Linear regression is a fundamental statistical and machine learning algorithm used for modeling the relationship between a dependent variable (target) and one or more independent variables (predictors). Its main goal is to establish a linear relationship that predicts the dependent variable based on the independent variables. Here's a detailed explanation of the linear regression algorithm:

Types of Linear Regression:

Simple Linear Regression: Involves one independent variable. Multiple Linear Regression: Involves two or more independent variables. Polynomial Regression: Models nonlinear relationships by introducing polynomial terms. Logistic Regression: Used for binary classification tasks, predicting probabilities. Linear Equation: Linear regression aims to find the linear equation of the form: $Y = \beta0 + \beta1X1 + \beta2X2 + ... + \betanXn$ Where:

Y is the dependent variable (target). X1, X2, ..., Xn are independent variables (predictors). β0 is the intercept (constant term). β1, β2, ..., βn are coefficients that represent the strength and direction of the relationship between each predictor and the target.

The algorithm calculates the coefficients (β0, β1, β2, ...) that minimize the difference between predicted and actual values. This is typically done using the method of least squares, which minimizes the sum of squared differences (residuals) between predicted and actual values.

Linear regression relies on certain assumptions, including linearity, independence of errors, homoscedasticity (constant variance of errors), normality of errors, and no multicollinearity.

Common evaluation metrics include R-squared (coefficient of determination), which measures the proportion of the variance in the dependent variable explained by the model, and Mean Squared Error (MSE) or Mean Absolute Error (MAE) to assess prediction accuracy.

Linear regression is sensitive to outliers and may not capture nonlinear relationships well. In such cases, more advanced regression techniques or feature engineering may be needed. Overall, linear regression is a versatile and widely used algorithm for understanding and predicting the relationship between variables in various domains. It serves as a foundation for more complex regression and machine learning methods.

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

Answer : Anscombe's quartet is a set of four small datasets that have nearly identical simple descriptive statistics but appear very different when graphed. It was created by the statistician Francis Anscombe in 1973 to illustrate the importance of data visualization in understanding and interpreting data. This quartet challenges the notion that summary statistics alone can fully describe a dataset, highlighting the need for graphical analysis.

The four datasets in Anscombe's quartet share the same statistical properties:

Dataset I: It is a simple linear relationship with some random noise. When you calculate the regression line and correlation coefficient, they suggest a strong linear relationship.

Dataset II: This dataset also has a linear relationship but with an outlier. The outlier significantly impacts the regression line and correlation coefficient.

Dataset III: This dataset contains two separate linear relationships. The presence of these two different relationships makes it challenging to describe the data using a single linear regression.

Dataset IV: This dataset appears to have no linear relationship between the variables. Although the summary statistics may not suggest any relationship, visualizing the data shows a clear nonlinear pattern. The key takeaways from Anscombe's quartet are:

Descriptive statistics alone may not provide a complete understanding of the data. Visualizing the data is essential for revealing patterns, outliers, and relationships that may not be apparent from summary statistics. It emphasizes the importance of exploratory data analysis (EDA) and the use of data visualization techniques to gain insights from data.

What is Pearson’s R? (3 marks)

Answer : Pearson's correlation coefficient, often denoted as "r" or "Pearson's R," is a statistical measure used to quantify the strength and direction of the linear relationship between two continuous variables. It was developed by Karl Pearson and is widely used in statistics and data analysis to assess the degree of association between two variables.

The key characteristics of Pearson's R are as follows:

Range and Significance: Pearson's R ranges from -1 to 1. A positive value (closer to 1) indicates a positive linear relationship, meaning as one variable increases, the other tends to increase. A negative value (closer to -1) indicates a negative linear relationship, where one variable tends to decrease as the other increases. A value close to 0 suggests a weak or no linear relationship.

Calculation: The formula for Pearson's R involves calculating the covariance between the two variables and dividing it by the product of their standard deviations. This normalized measure ensures that R is not affected by the scale or units of the variables.

Assumptions: Pearson's R assumes that the relationship between the variables is linear, and it is sensitive to outliers. Non-linear relationships may not be accurately captured by this correlation measure.

Limitations: Pearson's R is restricted to linear relationships and may not capture complex, nonlinear associations. It can be influenced by outliers and may not be suitable when data has a skewed distribution.

What is scaling? Why is scaling performed? What is the difference between normalized scalingand standardized scaling? (3 marks)

Answer : Scaling is a data preprocessing technique used to transform the range of numerical variables in a dataset to make them more suitable for analysis and modeling.

The primary reasons for scaling are to ensure that all variables have a similar influence on the model and to improve the convergence and performance of machine learning algorithms.

The two common scaling methods are normalized scaling and standardized scaling:

Normalized Scaling (Min-Max Scaling): This method scales the data to a specific range, typically between 0 and 1, but it can be adjusted to any desired range. It preserves the relative differences between data points and is useful when you want to maintain the original units of measurement.Min-Max scaling is sensitive to outliers because it constrains the range based on the minimum and maximum values in the dataset.

Standardized Scaling (Z-Score Scaling): Standardization transforms data to have a mean of 0 and a standard deviation of 1. It is suitable for algorithms that assume data is normally distributed, as it ensures that data is centered around zero and has consistent variances. Standardized scaling is robust to outliers because it does not depend on extreme values in the dataset.

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

Answer : Here are the key reasons for infinite VIF values:

Perfect Linear Relationship: When two or more independent variables in the model are perfectly correlated, one can be expressed as a linear combination of the others. This results in a division by zero when calculating VIF, leading to an infinite value.

Redundant Variables: In cases where one variable can be precisely predicted from a combination of other variables, multicollinearity arises. This redundancy makes it impossible to calculate VIF for the affected variables.

Insufficient Data: When the dataset is too small relative to the number of independent variables, VIF calculations can become unstable. In such cases, the estimates are unreliable, and VIF values may be infinite.

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

Answer : A Q-Q plot, short for "Quantile-Quantile plot," is a graphical tool used to assess whether a dataset follows a particular theoretical distribution, such as a normal distribution. It plots the quantiles (percentiles) of the observed data against the quantiles of the chosen theoretical distribution. In a Q-Q plot, if the data points fall along or approximately along a straight line, it suggests that the dataset is consistent with the chosen theoretical distribution.

The use and importance of a Q-Q plot in linear regression include:

Normality Assumption: Linear regression often assumes that the residuals (the differences between observed and predicted values) are normally distributed. Q-Q plots help verify this assumption by assessing whether the residuals follow a normal distribution. Deviations from a straight line in the plot can indicate departures from normality.

Identification of Outliers: Q-Q plots can reveal outliers and extreme values. Outliers can distort regression results and may need special consideration or data cleaning.

Model Evaluation: Q-Q plots are valuable for model evaluation. If the residuals closely match a straight line, it suggests that the model assumptions are met, and the regression results are more reliable.

Data Transformation: When the Q-Q plot indicates non-normality, data transformation techniques may be applied to improve the model's performance and the validity of statistical inferences.