Assignment-based Subjective Question:

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

Answer1:Below are the insights that can be drawn from the above plots

1. Fall season shows the highest rental bike demand.
2. Year 2019 demonstrates greater demand for rental bikes compared to the previous year.
3. Demand for bikes increases consistently from January to September, followed by a decrease in the subsequent months.
4. Holidays see higher demand compared to working days.
5. Weekends experience higher demand than weekdays.
6. Good weather conditions correspond to higher demand than moderate conditions, while bad weather conditions exhibit the lowest demand.

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

Answer2: Using drop_first=True during dummy variable creation is crucial to prevent the "dummy variable trap." This trap arises when we include all dummy variables without dropping one, leading to multicollinearity issues. By dropping the first dummy variable, we ensure independence among variables and accurate model interpretation.

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

Answer3: Looking at the pair-plot among the numerical variables, temp and atemp features have the highest correlation (0.63) with the target variable (cnt).

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

Answer2:

1. Residual Analysis of Training Data:

The errors exhibited a normal distribution with a mean of 0, indicating that the model first assumption have met.

Then i found out that the model's predictions for the training data closely align with the actual values of the training dataset.

Then i observed that my error terms are randomly distributed for train data which suggested that my model is capturing underlying patterns and relationships in data fairly well.

R2_score value for test predictions: R2 value for predictions on test data (0.829) is very close to R2 value of train data(0.811). This is a good R-squared value, hence we can see our model is performing good even on unseen data (test data)

Homoscedacity: I observed that the dispersion of the residuals (error terms) remains fairly consistent across predictions, indicating that the variability of the error term doesn't significantly change as the predictor variable's value varies.

Plotting Test vs Predicted value test: As i observed that prediction for test data is very close to the actual test data which means our model is performing well.

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 towards explaining the demand of the shared bikes are:

1. Year(Yr)
2. Feels Like Temperature(atemp)
3. Season(Season_spring)

General Subjective Question

1. Explain the linear regression algorithm in detail

Answer:

Linear regression is a powerful algorithm used to understand the relationship between a dependent variable and one or more independent variables. It helps us find the best-fitting straight line that represents this relationship. The algorithm follows a step-by-step process:

1. Data Preparation: Gather and clean the data, making sure there are no missing values or errors. This step ensures the data is ready for analysis.
2. Model Initialization: Set up the linear regression model by defining the dependent variable and independent variables. The model assumes a linear relationship between them.
3. Fitting the Model: Estimate the coefficients of the linear equation that best fits the data. This is done using a technique called Ordinary Least Squares (OLS), which minimizes the sum of squared errors.
4. Assessing Model Fit: Evaluate how well the model fits the data. The coefficient of determination (R-squared) tells us the proportion of variance in the dependent variable explained by the independent variables. Other statistical measures like the F-statistic and p-values help assess the significance of the model.
5. Assumptions and Residual Analysis: Validate the assumptions of linear regression, such as linearity, independence of errors, constant variance, and normality of residuals. Residuals are the differences between the observed and predicted values.
6. Prediction and Inference: Use the fitted model to make predictions on new data. Additionally, interpret the coefficients to understand the impact of independent variables on the dependent variable.
7. Model Evaluation and Improvement: Evaluate the performance of the model on unseen data to ensure its reliability. Techniques like cross-validation can help assess the model's predictive accuracy. If necessary, refine the model by adding or removing variables or considering more advanced regression techniques.
8. In summary, linear regression is a valuable algorithm for understanding the relationship between variables. It involves data preparation, model fitting, assessing model fit and assumptions, prediction and inference, and model evaluation and improvement.

2. Explain the Anscombe’s quartet in detail?

Answer:

Anscombe's quartet is a set of four datasets created by the statistician Francis Anscombe in 1973. These datasets are designed to have nearly identical statistical properties, but they exhibit distinct patterns and relationships when plotted. Anscombe's quartet serves as a powerful example to highlight the importance of visualizing data and the limitations of relying solely on summary statistics. Here's a detailed explanation of the Anscombe's quartet:

1. The quartet consists of four sets of data, each containing 11 (x, y) points. Despite having similar statistical measures such as means, variances, and correlation coefficients, the datasets display different patterns that challenge the assumption that summary statistics alone provide a complete understanding of the data.
2. Dataset I: This dataset exhibits a simple linear relationship between x and y, where y increases linearly with x. The relationship is strong and can be accurately described by a linear regression model
3. . Dataset II: In contrast to Dataset I, Dataset II showcases a non-linear relationship. The points follow a curved pattern, indicating that a linear model would not be appropriate for this dataset. This demonstrates the importance of considering alternative models based on the shape of the data.
4. Dataset III: Dataset III comprises two distinct groups of data points. When the entire dataset is considered, it appears to have a linear relationship, similar to Dataset I. However, when the groups are analyzed separately, they reveal different patterns. This emphasizes the significance of exploring subsets of data to gain a comprehensive understanding.
5. Dataset IV: Dataset IV illustrates the impact of outliers on statistical measures. Most of the data points exhibit a clear linear relationship, but a single outlier significantly affects the overall correlation and regression line. This highlights the need to identify and handle outliers appropriately during data analysis.
6. By presenting these four datasets together, Anscombe's quartet demonstrates that relying solely on summary statistics can be misleading. Visualizing the data helps uncover patterns, relationships, and anomalies that are not apparent from summary measures alone. It serves as a reminder of the importance of data exploration and visualization in gaining a deeper understanding of the underlying patterns and relationships within the data.

3. What is Pearson’s R?

Answer:

Pearson's R-squared is a statistical measure that ranges between 0 and 1. It represents the proportion of the total variation in the dependent variable that can be accounted for by the independent variable(s) in a linear regression model. An R-squared value of 0 indicates that the independent variable(s) have no explanatory power and cannot explain any of the variation in the dependent variable. On the other hand, an R-squared value of 1 indicates that the independent variable(s) perfectly explain all the variation in the dependent variable, resulting in a perfect fit. In practical terms, R-squared measures how well the observed data points fit the regression line. A higher R-squared value suggests that the model provides a better fit to the data, indicating that a larger proportion of the variability in the dependent variable is captured by the independent variable(s). However, it is important to note that a high R-squared does not guarantee that the model is accurate or that the relationship is causal. R-squared is often interpreted as the percentage of variation in the dependent variable explained by the independent variable(s). For example, an R-squared value of 0.75 means that 75% of the variation in the dependent variable can be explained by the independent variable(s), while the remaining 25% is attributed to other factors or random variation.

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

Answer:

Scaling refers to the process of rescaling the values of variables/features in a dataset. It is performed for the following key reasons:

1. To avoid attribute dominance: Some attributes may have larger numeric ranges than others. This can cause those attributes with larger scales to dominate comparisons made during the modeling process. Scaling helps balance the influence of all attributes.
2. To improve algorithm efficiency: Some machine learning algorithms work better when features/variables are constrained within a specific range, for example -1 to 1 or 0 to 1. Scaling helps prepare the data to work with these algorithms.
3. To avoid numerical difficulties during calculation: Calculations related to distant features can cause numerical problems. Scaling helps reformulate the values into smaller ranges to avoid underflow or overflow errors.
4. There are two main types of scaling performed: Normalized Scaling (Min-Max Scaling): This scales and translates each attribute/feature to fall between a specific range, typically 0 to 1. It is done by rescaling the original values to lie between these bounds. This preserves all relationships in the original data.
5. Standardized Scaling (Z-Scaling): This transforms the values of each attribute/feature to have zero-mean and unit variance. It is done by subtracting the mean and then dividing the result by the standard deviation of each attribute. This results in a standard distribution with mean 0 and standard deviation 1. This transformation results in all attributes being measured on comparable scales.
6. In summary, scaling helps handle attribute dominance issues, improves algorithm performance, and avoids numerical problems during calculation. Normalized scaling preserves relationships while standardized scaling results in a standardized distribution for improved comparability across attributes.

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

Answer:

VIF standing for Variance Inflation Factor is a useful statistical measure for detecting multicollinearity (multiple linear correlation) between independent variables in a multiple regression model. VIF values higher than 5 are usually taken to indicate multicollinearity. The value of VIF becomes infinite when there is perfect multicollinearity between two or more independent variables.

This happens due to the following reasons: When one independent variable is perfectly predicted by the other independent variables, its VIF value will become infinite. In other words, if one independent variable can be written as a linear combination of others, it indicates perfect multicollinearity. The VIF is calculated as 1/(1-R2) where R2 is the coefficient of determination from the regression of that independent variable against other independent variables. When R2 is equal to 1, it means the independent variable can be completely explained or predicted by the other variables, making the denominator equal to 0 and VIF infinite. Mathematically, in a standard multiple linear regression model, the covariance matrix becomes singular in the case of perfect multicollinearity making it impossible to calculate the regression coefficients. This causes statistical programmes to return error messages or produce infinite VIF values. So in summary, an infinite VIF value occurs when there is a perfect linear relationship between two or more independent variables due to which their inclusion together in the model makes the covariance matrix singular leading to computational issues. It is an indication of perfect multicollinearity.

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, or quantile-quantile plot, is a graphical tool that helps us assess if a set of data plausibly came from some theoretical distribution such as a normal, exponential, or uniform distribution. It is also used to compare two data sets to see if they come from populations with the same distribution.

In linear regression, a Q-Q plot can be used to assess the normality of the residuals. The residuals are the differences between the observed values and the predicted values from the regression model. If the residuals are normally distributed, then the Q-Q plot will be a straight line. However, if the residuals are not normally distributed, then the Q-Q plot will deviate from the straight line.

The importance of a Q-Q plot in linear regression is that it can help us identify potential problems with the model. For example, if the residuals are not normally distributed, then this may indicate that the model is not a good fit for the data. It may also indicate that there are some outliers in the data that are distorting the results of the model. Here are some of the things that we can learn from a Q-Q plot: Whether the data is normally distributed. Whether two data sets come from populations with the same distribution. The presence of outliers in the data. The skewness and kurtosis of the data. To interpret a Q-Q plot, we can look at how closely the points follow the 45-degree reference line. If the points are close to the line, then the data is likely normally distributed. However, if the points deviate from the line, then the data is not likely normally distributed.

The following are some of the common patterns that we may see in a Q-Q plot: If the points are all below the line, then the data is negatively skewed. If the points are all above the line, then the data is positively skewed. If the points are more dispersed towards the tails of the distribution, then the data has high kurtosis. If the points are more dispersed towards the center of the distribution,then the data has low kurtosis.