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:

- High demand of bikes are their in season 3 fall.
- Their will be an increase in the demand in 2019(next year)
- Demand of bikes are growing till June and the highest demand is in the month of September. After September demand is decreasing.
- Weekday is not giving clear picture about demand.
- On holidays, demand has decreased.
- Clear weathershit has highest demand.

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

Answer:

drop_first = True is important to use, as it helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables.

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

Answer:

'temp' column has highest correlation with target variable. There is a positive relationship between 'cnt' and 'temp' variable.

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

Answer:

1) Normality of Residuals: Residuals/Error terms should be normally distributed.

2) Linearity: Relationship between each predictor variable and the response variable is linear by examining residual plots or using methods like partial regression plots.

3) Multicollinearity: Check whether there is correlation between predictor variables to ensure they are not highly correlated, which can inflate standard errors and affect coefficient interpretations.

4) Homoscedasticity: Ensure that the variance of residuals is constant across all levels of predictor variables using scatter plots of residuals against predicted values or predictor variables.

5) Independence of Errors: Check for autocorrelation or serial correlation in residuals using methods like Durbin-Watson statistic for time-series data or residuals versus time plots.

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

Answer:

1) Temp
2) weathersit_light
3) year

General Subjective Questions

Question 1 :Explain the linear regression algorithm in detail.

Answer:

Linear regression is a fundamental algorithm in machine learning and statistics used to model the relationship between a dependent variable and one or more independent variables. The objective is to find the best-fitting straight line (or hyperplane in the case of multiple variables) through the data points. Here's a detailed explanation of the linear regression algorithm:

Key Concepts

1) Dependent Variable (Y): The outcome or target variable you are trying to predict.

2) Independent Variable (X): The feature(s) or predictor(s) used to predict the dependent variable.

3) Simple Linear Regression: Involves one independent variable.

4) Multiple Linear Regression: Involves multiple independent variables.

The Linear Regression Model

The relationship between the dependent variable yyy and the independent variables XXX is modeled as:

$$y=\beta0+\beta1x1+\beta2x2+\ldots+\betanxn+\epsilon$$

- Where:
- yyy is the dependent variable.
- β0\beta_0β0 is the intercept (the value of yyy when all xxxs are zero).
- β1,β2,…,βp\beta_1, \beta_2, \ldots, \beta_pβ1,β2,…,βp are the coefficients of the independent variables.
- x1,x2,…,xpx_1, x_2, \ldots, x_px1,x2,…,xp are the independent variables.
- ϵ\epsilonϵ is the error term, representing the difference between the actual and predicted values.

Assumptions:

Linear regression relies on several key assumptions:

1) Linearity: The relationship between the dependent and independent variables is linear.

2) Independence: Observations are independent of each other.

3) Homoscedasticity: The variance of residual errors is constant across all levels of the independent variables.

4) Normality: Residuals (errors) are normally distributed.

5) No Multicollinearity: Independent variables are not highly correlated with each other.

Steps in Linear Regression:

1. Data Collection and Preparation

- Collect and clean the data.
- Split the data into training and testing sets.
- Standardize or normalize the data if necessary.

2. Model Specification

- Define the model equation as

$$y=\beta0+\beta1x1+\beta2x2+\ldots+\beta pxpy$$

3. Estimation of Coefficients

- Ordinary Least Squares (OLS): The most common method for estimating the coefficients.
- Minimize the sum of squared residuals (differences between observed and predicted values).

4. Model Fitting

Fit the linear regression model to the training data to estimate the coefficients.

5. Model Evaluation

- Use metrics like R-squared, Adjusted R-squared, Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE) to evaluate the model's performance.
- Validate assumptions by analyzing residuals.

6. Prediction

Use the fitted model to make predictions on new or test data.

7. Interpretation

- Interpret the coefficients to understand the impact of each independent variable on the dependent variable.
- Consider statistical significance (p-values) of the coefficients.

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

Answer:

Anscombe's quartet is a collection of four datasets that are designed to have nearly identical simple descriptive statistics, yet they exhibit markedly different distributions and relationships when visualized graphically. Created by statistician Francis Anscombe in 1973, these datasets highlight the importance of graphing data to uncover underlying patterns that summary statistics alone may not reveal.

- **Key Characteristics**
- Each dataset in Anscombe's quartet has the same:
- Mean of the x-values (9)
- Mean of the y-values (7.5)
- Variance of the x-values (11)
- Variance of the y-values (4.125)
- Correlation between x and y (0.816)
- Linear regression line ($y=3+0.5xy = 3 + 0.5xy=3+0.5x$)
- Despite these similarities, the datasets are dramatically different when plotted.

The Four Datasets

1) Dataset I:

- Exhibits a simple linear relationship with some random noise.
- Points are scattered around the regression line, showing a typical linear pattern.

2) Dataset II:

- Displays a linear relationship but includes one significant outlier.
- The outlier influences the correlation and regression line, demonstrating how a single data point can impact statistical measures.

3) Dataset III:

- Shows a clear non-linear (curvilinear) relationship.
- While the summary statistics suggest linearity, the scatter plot reveals a distinct curve.

4) Dataset IV:

- Consists of a vertical line with an extreme outlier.
- This dataset has most points at the same x-value, except for one outlier with a much higher y-value, drastically affecting the statistical properties.

Importance of Anscombe's Quartet

1) Graphical Analysis:

- Emphasizes the necessity of visualizing data. Summary statistics alone can be misleading, and graphs can uncover hidden patterns, outliers, and relationships.

2) Robustness of Statistical Methods:

Demonstrates that a few anomalous points can significantly affect statistical summaries and regression analyses. Visual inspection helps identify these anomalies.

3) Contextual Understanding:

Highlights the need to interpret statistical metrics in the context of the data. Relying solely on numerical summaries without graphical representation can lead to incorrect conclusions.

Question 3: What is Pearson's R?

Answer:

Pearson's R, also known as the **Pearson correlation coefficient**, is a statistical measure that quantifies the strength and direction of the linear relationship between two continuous variables. It is denoted by the symbol rrr.

Key Characteristics

1) Range:

- The value of rrr ranges from -1 to +1.
- r=+1r = +1r=+1: Perfect positive linear relationship.
- r=−1r = -1r=−1: Perfect negative linear relationship.
- r=0r = 0r=0: No linear relationship.

2)Direction:

- Positive rrr indicates that as one variable increases, the other variable tends to also increase.
- Negative rrr indicates that as one variable increases, the other variable tends to decrease.

3) **Strength**:

- The closer the value of rrr is to ±1, the stronger the linear relationship between the variables.
- Values closer to 0 indicate a weaker linear relationship.

Interpretation

1) Perfect Positive Correlation (r=+1r = +1r=+1):

- Indicates that the two variables move in perfect synchronization in the same direction.

2)Perfect Negative Correlation (r=−1r = -1r=−1):

- Indicates that the two variables move in perfect synchronization in opposite directions.

3) No Correlation (r=0r = 0r=0):

- Suggests that there is no linear relationship between the variables.

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

Answer:

Scaling is a data preprocessing technique used to adjust the range of features in a dataset so that they can be compared on a common scale. This process is essential in many machine learning algorithms where the distance between data points influences the model's performance and outcomes.

Why is Scaling Performed?

- **Improving Model Performance**: Algorithms such as k-nearest neighbors (KNN), support vector machines (SVM), and gradient descent optimization methods (used in linear regression, logistic regression, neural networks, etc.) are sensitive to the scales of the input features. Scaling ensures that features contribute equally to the result, improving the performance and convergence speed of these algorithms.
- **Avoiding Dominance**: Features with larger scales can dominate the learning process, leading to biased models. Scaling ensures that no single feature disproportionately affects the model.
- **Facilitating Gradient Descent**: In algorithms that rely on gradient descent, like neural networks, scaling helps achieve faster convergence by ensuring that the gradient steps are well-proportioned.

Types of Scaling

1)Normalization (Min-Max Scaling):

- Rescales the data to a fixed range, usually [0, 1].
- Sensitive to outliers, which can skew the scaled values

2)Standardization (Z-score Normalization):

- Transforms the data to have a mean of 0 and a standard deviation of 1.
- Less sensitive to outliers, as it standardizes based on the mean and standard deviation.

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

Answer:

Infinite VIF values occur due to perfect multicollinearity, which means one predictor variable can be perfectly predicted by a linear combination of other predictors. This often happens because of:

- **Perfect Multicollinearity**: Predictors are perfectly linearly related.
- **Dummy Variable Trap**: Including all categories of a categorical variable without dropping one.
- **Redundant Features**: Features that are exact linear combinations of others.

Infinite VIF indicates severe multicollinearity, making regression coefficients unreliable.

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 that compares the quantiles of a dataset against the quantiles of a theoretical distribution, typically the normal distribution, to assess how well the data fits that distribution.

Use and Importance in Linear Regression

- **Normality Check**: In linear regression, a Q-Q plot is used to check if the residuals (errors) are normally distributed.
- **Model Diagnostics**: Helps assess the goodness-of-fit of the regression model and validate the assumption of normality.
- **Detection of Deviations**: Identifies skewness, kurtosis, and outliers that may indicate issues with the model.
- **Assumption Validation**: Ensures that statistical inferences drawn from the model, such as hypothesis tests and confidence intervals, are valid.

A good fit in a Q-Q plot indicates that the residuals are normally distributed, supporting the reliability of the regression model.