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?

Ans:

Season: The demand for bikes is a slight increase in fall and summer. That means people tend to rent bikes more in favourable weather conditions.

Month: There is no significant variation in bike demand across different months.

Weekday: Bike demand is consistent across all days of the week, showing that people use bikes equally on weekdays and weekends.

Weather: Clear and few clouds have the highest bike demand, while light snow and light rain conditions have the lowest. This indicates that adverse weather conditions negatively impact bike rentals.

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

Ans:

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.

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

Ans:

The numerical variable with the highest correlation with the target variable - cnt is temp.

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

Ans:

Linearity: Scatter plots between predicted values and residuals.

Normality of residuals: Distribution plot (histogram) of residuals.

Multicollinearity: Checking VIF values to ensure they are below the threshold (less than 5)

Normality of error terms: Error term should be normally distributed.

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

Ans:

temp: Positive correlation indicating higher bike demand with increasing temperature.

2019: Positive correlation showing increased bike demand in 2019 compared to 2018.

winter: Positive correlation indicating higher bike demand in winter compared to other seasons.

General Subjective Questions

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

Ans: Linear Regression is a statistical method used to model the relationship between a dependent variable and one or more independent variables. It assumes a linear relationship between the variables, which can be expressed as: Y = beta_0 + beta_1X_1 + beta_2X_2 + ……+ beta_nX_n

Where:

- Y is the dependent variable.• X_1, X_2, … , X_n are the independent variables.

- beta_0 is the y-intercept.•

- beta_1, beta_2, …, beta_n are the coefficients of the independent variables.

Steps in Linear Regression:

1. Data Collection and Preparation: Gather and clean the data, handling missing values and outliers.

2. Exploratory Data Analysis (EDA): Understand the data distribution and relationships between variables.

3. Model Building:

- Choosing the Model: Simple Linear Regression (one independent variable) or Multiple LinearRegression (multiple independent variables).

-Splitting the Data: Divide the data into training and testing sets.

-Fitting the Model: Use the training data to fit the linear regression model and estimate thecoefficients using methods like Ordinary Least Squares (OLS).

- Model Evaluation: Performance Metrics: Evaluate the model using metrics like R-squared, Adjusted R-squared• Residual Analysis: Check the residuals to ensure they are normally distributed with constant variance.

- Model Interpretation: Interpret the coefficients to understand the impact of each independent variable on the dependent variable.6. Prediction: Use the model to make predictions on new data

2. Explain Anscombe’s quartet in detail.

Ans:

Anscombe’s quartet consists of four datasets that have nearly identical simple statistics(mean, variance, correlation, etc.) but differ significantly when graphed. This demonstrates the importance of graphical analysis of data.

Each dataset in Anscombe’s quartet has:

- Nearly the same mean of (x) and (y)

-Nearly the same variance of (x) and (y)

-Nearly the same correlation between (x) and (y)

- Nearly the same linear regression line

Despite these similarities, plotting the datasets reveals that they have very different distributions and Patterns. Anscombe’s quartet highlights the importance of visualising data before analysing itstatistically. Graphs can reveal patterns, trends, and outliers that summary statistics might not capture.

3. What is Pearson’s R?

Ans:

Pearson’s R, also known as the Pearson correlation coefficient, measures the strength and direction of the linear relationship between two variables. It ranges from -1 to 1, where:

1 indicates a perfect positive linear relationship.

-1 indicates a perfect negative linear relationship.

0 indicates no linear relationship.

Pearson’s R is sensitive to outliers, which can skew the correlation. It assumes a linear relationshipbetween the variables and requires interval or ratio scale data.

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

Ans:

Scaling is the process of transforming the features in a dataset so that they have a comparable scale,often to improve the performance of machine learning algorithms.

Why Scaling is Performed:

- Ensures that features are treated equally by algorithms.

-Helps gradient-based algorithms converge faster.

- Improves the performance of algorithms sensitive to the scale of data.

Types of Scaling:

1. Normalized Scaling (Min-Max Scaling): Transforms features to a fixed range, usually [0, 1] and Retains the relationship between the values but changes their scale.

2. Standardized Scaling (Z-score Standardization): Transforms features to have a mean of 0 and a standard deviation of 1.

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

Ans:

The Variance Inflation Factor (VIF) measures how much the variance of a regression coefficient is inflated due to multicollinearity in the model. A VIF value can become infinite when there is perfect multicollinearity, meaning one predictor variable is an exact linear combination of one or more other predictor variables. This perfect correlation causes the denominator in the VIF calculation to be zero, leading to an infinite value.

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

Ans:

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

Use and Importance in Linear Regression:

- Normality Check: Assess whether the residuals of a regression model are normally distributed, a key assumption in linear regression.

-Detecting Deviations: Identify deviations from normality, such as skewness or kurtosis.

-Model Validation: Ensure that residuals meet the assumptions of the regression model for valid inference.

In a Q-Q plot, if the points lie approximately along the reference line (typically a 45-degree line), the sample distribution matches the theoretical distribution. Deviations from the line indicate departures from the assumed distribution.