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? (3 marks)

Ans - Analyzing the effect of categorical variables on the dependent variable, cnt (total bike rentals), involves examining how each category within these variables influences the number of bike rentals. In this context, we can infer the potential impacts of the categorical variables season and weathersit on the dependent variable.

<u>Season:</u>

Spring: Moderate rentals.

Summer: Highest rentals due to good weather.

Fall: Slightly fewer rentals than summer, still popular.

Winter: Lowest rentals due to cold weather.

<u>Weather Situation (Weathersit):</u>

Clear: Highest rentals, ideal conditions.

Mist: Slight reduction in rentals.

Light Rain: Noticeable drop in rentals.

Heavy Rain: Lowest rentals, poor conditions.

These effects are inferred from the patterns typically associated with these seasons and weather conditions.

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

Ans - Using drop_first=True helps prevent multicollinearity and simplifies the interpretation of regression coefficients by providing a baseline category for comparison.

Example:

Assume you have a categorical variable Color with three categories: Red, Green, Blue.

Without drop_first=True:

Dummy variables: Color_Red, Color_Green, Color_Blue.

With drop_first=True:

Dummy variables: Color_Green, Color_Blue.

Here, Red becomes the reference category:

Interpretation: The coefficients for Color_Green and Color_Blue show how much the dependent variable changes compared to the Red category.

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

with the target variable? (1 mark)

Ans - Based on typical bike-sharing data, temp (temperature) often has the highest correlation with the total bike rentals (cnt). This makes intuitive sense, as favorable weather conditions typically lead to higher bike usage. However, it's best to confirm this with the actual dataset using the code provided.

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

training set? (3 marks)

Ans - After building the linear regression model on the training set, the assumptions were validated as follows:

Linearity: Checked by plotting actual vs. predicted values and residuals vs. predicted values to ensure no clear pattern in residuals.

Independence: Verified using the Durbin-Watson statistic, aiming for a value close to 2.

Homoscedasticity: Assessed by plotting residuals vs. predicted values, looking for constant variance.

Normality of Errors: Evaluated by plotting a histogram and a Q-Q plot of the residuals, checking for a normal distribution.

Lack of Multicollinearity: Confirmed by calculating the Variance Inflation Factor (VIF) for each predictor, ensuring VIF values are below 10.

5. Based on the final model, which are the top 3 features contributing significantly towards

explaining the demand of the shared bikes? (2 marks)

Ans - Based on the final model, the top 3 features contributing significantly to the demand for shared bikes are:

Temperature (temp)

Weather situation - mist (weathersit_mist)

Humidity (hum)

These features were identified based on the magnitude of their coefficients in the linear regression model.

General Subjective Questions

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

Ans - Linear regression is a simple algorithm used in machine learning and statistics to predict a numeric outcome based on one or more input features.

Here's a step-by-step explanation:

1. **Objective**: The goal is to find a linear relationship between the input features (independent variables) and the output (dependent variable). This relationship is represented by a straight line.

2. **Equation**: For a single feature, the relationship is modeled by the equation:

$y=mx+by=mx+b$

where:

- yy is the predicted output.
- xx is the input feature.
- mm is the slope of the line (how much yy changes for a unit change in xx).
- bb is the y-intercept (the value of yy when xx is 0).

For multiple features, the equation is extended to:

$y=b+m1x1+m2x2+\ldots+mnxny=b+m_1x_1+m_2x_2+\ldots+m_nx_n$

where each xix_i is an input feature, and each mim_i is its corresponding coefficient.

3. **Training**: During training, the algorithm finds the best values for the coefficients (mms) and the intercept (bb) that minimize the difference between the predicted values and the actual values. This difference is measured by a metric called the "cost function," often using the Mean Squared Error (MSE):

$$\mathrm{MSE}=1N\sum i=1N(yi-y\hat{}i)2\mathrm{MSE}=N1_{i=1}\sum_{N}(y_i-y\hat{}_i)_2$$

where:

- NN is the number of data points.
- yiy_i is the actual value.
- $y\hat{}iy\hat{}_i$ is the predicted value.

4. **Optimization**: The coefficients and intercept are adjusted using an optimization algorithm like Gradient Descent, which iteratively updates the values to reduce the cost function.

5. **Prediction**: Once the model is trained, it can predict the output for new input data by applying the learned linear relationship.

In summary, linear regression finds the best-fitting straight line through the data points by minimizing the prediction errors, allowing it to predict numeric outcomes based on input features.

2. Explain the Anscombe’s quartet in detail.

Ans - Anscombe's Quartet is a set of four datasets created by statistician Francis Anscombe to demonstrate the importance of graphing data before analyzing it. Each dataset has nearly identical statistical properties (mean, variance, correlation, and linear regression line), but they look very different when graphed.

Here's a breakdown of Anscombe's Quartet:

1. **Statistical Similarity**: All four datasets have:

- The same mean of xx and yy.
- The same variance of xx and yy.

- The same correlation coefficient between xx and yy.
- The same linear regression line ($y=3+0.5xy=3+0.5x$).

2. **Graphical Differences**:

    - **Dataset 1**: Shows a typical linear relationship between xx and yy.
    - **Dataset 2**: Appears to have a clear curve, indicating a non-linear relationship.
    - **Dataset 3**: Contains an outlier that greatly affects the regression line.
    - **Dataset 4**: Most data points are the same, with one outlier that affects the regression line significantly.

Importance:

- **Visualization**: Anscombe's Quartet emphasizes that relying solely on statistical measures can be misleading. Visualizing data is crucial to understand its true nature.
- **Data Integrity**: It highlights how different patterns and anomalies in data can be hidden behind similar statistical summaries.

In short, Anscombe's Quartet teaches us that data visualization is essential for accurate data analysis, ensuring we don't overlook underlying patterns or outliers.

3. What is Pearson’s R?

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

1 indicates a perfect positive linear relationship.

-1 indicates a perfect negative linear relationship.

0 indicates no linear relationship.

The formula for Pearson's R is: r = Cov(X,Y)/σXσY

Cov(X,Y) is the covariance between variables X and Y

σX and σY are the standard deviations of X and Y, respectively.

In simple terms, Pearson's R tells you how well two variables are linearly related.

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

Ans - Scaling is the process of transforming data so that it fits within a specific range or distribution. This is crucial for many machine learning algorithms that perform better or converge faster when the input features are on a similar scale.

Why Scaling is Performed:

Improves Algorithm Performance: Some algorithms (e.g., gradient descent-based methods) are sensitive to the scale of the data.

Speeds Up Convergence: Scaling can make optimization faster and more stable.

Enhances Model Accuracy: Ensures that each feature contributes equally to the result, avoiding bias towards features with larger values.

In summary:

Normalization rescales data to a fixed range, typically [0, 1].

Standardization centres data 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 value of the Variance Inflation Factor (VIF) becomes infinite when there is perfect multicollinearity between the predictors in a regression model. This means that one predictor can be exactly predicted by a linear combination of the other predictors.

Why This Happens:

Perfect Multicollinearity: Occurs when one variable is a perfect linear function of one or more other variables. In mathematical terms, the correlation coefficient between the variables is exactly ±1.

Division by Zero, this results in division by zero, causing VIF to be infinite.

In summary, VIF is infinite when predictors are perfectly collinear, indicating a redundancy in the predictors used in the regression model.

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

Ans - A Q-Q plot (Quantile-Quantile plot) is a graphical tool used to compare the distribution of a dataset to a theoretical distribution, typically the normal distribution.

Use in Linear Regression:

Assess Normality: Helps check if the residuals (errors) from a linear regression model follow a normal distribution, which is an assumption of many linear regression methods.

Interpretation:

If the points on the Q-Q plot lie approximately along a straight line, the residuals are normally distributed.

Deviations from the line suggest departures from normality, such as skewness or heavy tails.

Importance:

Model Validation: Ensuring that residuals are normally distributed validates the model assumptions, leading to more reliable inference and predictions.

Identifying Problems: Helps detect issues like outliers, skewness, or other deviations from normality that might affect model performance.

In summary, a Q-Q plot is essential in linear regression for validating the assumption of normally distributed residuals, ensuring the model's reliability and accuracy.