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:

There are 7 categorical columns.

Season, holiday, weekday, workingday, weathersit, mnth, yr

High demand of the bike

- In season 2 and 4.
- On holidays
- When weathersit is Clear, Few clouds, Partly cloudy, Partly cloudy

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

Ans:

- it can be used to delete extra column while creating dummy variables.
- It is also used to reduce the collinearity between dummy variables

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

Ans: atemp and temp

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

Ans:

There are 28 variables including dummy variables. We found r2 value 84% with all variables.

We have created another model using RFE with 9 variables and found r2 value is 77% and it seems very good model.

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

Ans:

Temp, casual, registered, cnt

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Ans:

Linear regression is a statistical method used to model the relationship between a dependent variable (often denoted as Y) and one or more independent variables (often denoted as X). It assumes that there is a linear relationship between the independent variable(s) and the dependent variable.

The linear regression model is represented by the equation:

$Y= B_0 + B_1X_1 + B_2X_2 + .. + B_nX_n + E$

Where Y = dependent variable

B_0 to B_n are the coefficients

X_1 to X_n are independent variables.

E represents the error term.

2. Explain the Anscombe's quartet in detail?

Ans:

Anscombe's quartet is a set of four small datasets that have nearly identical simple descriptive statistics, yet have very different distributions and appear very different when graphed.

Anscombe's quartet demonstrates the importance of visualizing data and not relying solely on summary statistics. It shows that different datasets can have similar summary statistics but very different underlying structures. This highlights the value of exploratory data analysis and visualization in understanding the nature of the data.

3. What is Pearson's R?

Ans:

Pearson's r (also known as the Pearson correlation coefficient) is a statistical measure that quantifies the strength and direction of a linear relationship between two continuous variables. It is denoted by the symbol r and ranges from -1 to +1.

Here's what these values represent:

r=+1: A perfect positive linear relationship.

r=0: No linear relationship (the variables are not correlated).

r=−1: A perfect negative linear relationship.

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

Ans:

Scaling, in the context of data preprocessing in machine learning, refers to the process of transforming or adjusting the range or distribution of features or variables. This is done to ensure that all variables have a similar scale or range, which can be important for certain machine learning algorithms to perform effectively.

Scaling Performed

Preventing Dominance of Features: In many machine learning algorithms, features with larger scales might dominate over features with smaller scales. This can lead to biased results. Scaling ensures that no particular feature has a disproportionate impact on the algorithm.

Improving Convergence in Optimization Algorithms: Some optimization algorithms, like gradient descent, converge faster when features are on similar scales. This can speed up the training process.

Improving Interpretability: For models like linear regression, the coefficients represent the importance of each feature. If features are on different scales, comparing the importance becomes difficult.

Enhancing Model Performance: Some algorithms, like Support Vector Machines (SVMs) and K-Nearest Neighbors (KNN), are sensitive to the scale of features. Scaling can lead to more accurate predictions.

Types of Scaling:

1. Normalized Scaling (Min-Max Scaling):

Normalized scaling transforms the data so that it falls within a specific range, usually [0, 1].

The formula for Min-Max scaling is:

$X_{scaled} = X – X_{min} / X_{max} - X_{min}$

Where:

X is an individual data point.

X min is the minimum value of the feature.

X max is the maximum value of the feature.

2. Standardized Scaling (Z-Score Scaling):

Standardized scaling transforms the data so that it has a mean of 0 and a standard deviation of 1.

The formula for Standardized scaling is:

$X_{scaled} = X – \mu / \sigma$

Where:
X is an individual data point.
μ is the mean of the feature.
σ is the standard deviation of the feature.
Standardized scaling is especially useful when the features have different units or are measured in different ways.

<u>Differences between Normalized and Standardized Scaling:</u>

Range of Values:

- Normalized scaling transforms data to fall within the range [0, 1].
- Standardized scaling transforms data to have a mean of 0 and a standard deviation of 1.

Impact on Outliers:

- Normalized scaling can be influenced by outliers since it depends on the range of the data.
- Standardized scaling is less affected by outliers because it is based on the mean and standard deviation.

Interpretability:

- Normalized scaling maintains the interpretability of the original data in terms of the range.
- Standardized scaling may make interpretation less intuitive, as the values are in terms of standard deviations from the mean.

The choice between normalized and standardized scaling depends on the specific requirements of the problem and the characteristics of the data.

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

Ans:

The Variance Inflation Factor (VIF) is a measure used in regression analysis to assess the severity of multicollinearity in a set of independent variables. Multicollinearity occurs when independent variables in a regression model are highly correlated with each other, which can lead to issues in interpreting the model's coefficients.

A VIF value of infinity (or a very large value) typically occurs when there is perfect multicollinearity in the dataset. Perfect multicollinearity means that one or more independent variables can be perfectly predicted from the other variables in the 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 (Quantile-Quantile) plot is a graphical tool used to assess whether a given dataset follows a particular probability distribution. It compares the distribution of a dataset to a theoretical distribution (usually the normal distribution) by plotting the quantiles of the dataset against the quantiles of the theoretical distribution.

Use and Importance in Linear Regression:

- Checking Normality Assumption
- Identifying Skewness and Outliers
- Assessing Model Assumptions
- Guiding Model Refinement
- Avoiding Misinterpretation of Results
- Ensuring Reliable Inferences