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?

Ans1: Following data cleaning and feature engineering, five categorical variables were identified: year, month, season, weather, and working/non-working. A correlation analysis revealed a strong positive association between the 'season' variable and the target variable 'cnt.' Specifically, the 'winter' season exhibited the highest correlation. Conversely, the 'misty' and 'light snow' weather conditions were found to have a negative impact on bike demand. The 'working/non-working' variable, however, appeared to have a negligible influence on the target variable.".

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

Ans 2: The get_dummies() method in pandas is used to convert categorical variables into numerical representations. By default, it creates a separate column for each unique category. However, to avoid redundant information and potential multicollinearity, the drop_first=True parameter can be used. This drops the first category column, as its presence can be inferred from the absence of values in the remaining columns. For example, in a housing dataset with a 'furnishing status' column containing three categories ('furnished,' 'semi-furnished,' and 'unfurnished'), using drop_first=True would create only two dummy columns: 'furnished' and 'semi-furnished.' A value of '1, 0' in these columns would indicate a 'furnished' status, '0, 1' would represent 'semi-furnished,' and '0, 0' would imply 'unfurnished.' This approach effectively encodes categorical information while minimizing the dimensionality of the dataset.

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

Ans 3 A pair plot analysis revealed a strong positive correlation between the 'aTemp' variable and the target variable 'cnt.' This visual observation was further confirmed by a correlation matrix, which quantified the correlation coefficient between 'aTemp' and 'cnt' as 0.63.

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

Ans 4: **Linear Regression Assumptions:**

1. **Linearity:**
   - **Visual Inspection:** Scatter plots were generated to assess the relationship between each independent variable and the dependent variable. The plots exhibited a linear pattern, suggesting a linear relationship.
2. **Multicollinearity:**
   - **Variance Inflation Factor (VIF):** The VIFs of all independent variables were calculated. Low VIF values (typically less than 5) indicate a minimal degree of multicollinearity among the predictors.
3. **Normality of Residuals:**
   - **Distribution Plot:** A distribution plot of the model's residuals was created. The residuals appeared to follow a normal distribution with a mean of 0 and a standard deviation of 1. This suggests that the errors are normally distributed, which is a key assumption of linear regression.

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

Ans 5: The final model identified the following variables as significant predictors of bike demand: year, aTemp (apparent temperature), windspeed, mistyCloudy, summer, and winter. The coefficients associated with these variables are as follows:

Year: 2075.3028

aTemp: 5295.4750

Windspeed: -1490.9861

MistyCloudy: -540.6742

Summer: 609.2490

Winter: 909.5742

Based on the absolute values of the coefficients, temperature (aTemp), windspeed, and year emerged as the top three features with the most substantial impact on bike demand.

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Ans 1Linear 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 and uses a mathematical equation to predict the value of the dependent variable based on the values of the independent variables.

Think of linear regression like a detective trying to solve a mystery. The detective (the algorithm) wants to figure out how different clues (independent variables) are connected to the main mystery (the dependent variable). The detective assumes that there's a direct, straightforward relationship between the clues and the mystery, just like a linear equation.

Here are some things the detective needs to keep in mind:

1. The clues need to fit together: The detective can't solve the mystery if the clues don't make sense together. This means the independent variables need to have a linear relationship with the dependent variable.
2. The detective needs to be unbiased: The detective can't let their own biases cloud their judgment. This means the errors (or "residuals") between the detective's guesses and the actual solution should be evenly distributed.
3. The clues can't be too similar: If all the clues are basically the same, it's hard to figure out which one is most important. This means the independent variables shouldn't be too highly correlated with each other.
4. The clues should be consistent: The detective can't rely on clues that are sometimes strong and sometimes weak. This means the errors should be consistent, no matter what the clues are.

2. Explain the Anscombe's quartet in detail.

Ans 2: **Anscombe's Quartet** is a collection of four datasets that, despite having identical summary statistics (mean, variance, correlation), exhibit radically different visual patterns when plotted. This demonstrates the limitations of relying solely on numerical summaries for data analysis and underscores the importance of **exploratory data analysis** (EDA). The quartet serves as a cautionary tale, highlighting how seemingly similar datasets can conceal distinct underlying relationships that are only discernible through visualization.

3. What is Pearson’s R?

Ans 3: Pearson Correlation Coefficient is a parametric statistical measure that quantifies the linear relationship between two variables. It ranges from -1 to +1, where -1 indicates a perfect negative correlation, +1 signifies a perfect positive correlation, and 0 suggests no correlation. The coefficient is calculated by dividing the covariance of the variables by the product of their standard deviations. This method is most suitable for normally distributed data as it relies on the mean and standard deviation of the variables.

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

Ans 4: Scaling is a data preprocessing technique that transforms data points to a common range, facilitating model interpretability and convergence. For instance, variables like salary (in lakhs) and age (in years) have disparate scales, which can hinder the interpretation of coefficients in a linear model.

Normalization (Min-Max Scaling): This method rescales data to a range of 0 to 1 by linearly transforming the values. It preserves the relative distances between data points while ensuring comparability across variables.

Standardization (Z-Score Scaling): Standardization transforms data to have a mean of 0 and a standard deviation of 1. This is equivalent to converting data points to their Z-scores.

Standardization is particularly useful when the data distribution is approximately normal or when outliers are present, as it helps mitigate their influence.

By applying appropriate scaling techniques, it becomes easier to interpret model coefficients and improve the overall performance of machine learning algorithms.

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

Ans 5: Imagine you're trying to build a house. You have a bunch of different materials, like wood, bricks, and glass. If all of these materials are exactly the same (100% correlated), it's like trying to build a house with only one kind of material. The house won't be very strong or stable.

VIF is like a measure of how similar these materials are. If VIF is high, it means the materials are too similar, and the house (or your model) won't be very reliable. It's like trying to build a house with a bunch of identical bricks. The house might look nice on the outside, but it won't be very sturdy.

So, a high VIF means your model is overfitting. It's too focused on the data it was trained on and might not work well with new data.

VIF represents the correlation between the independent variables. It is calculated as below:

$$VIF = 1/1-R^2$$

This means when R^2 is 1 i.e. 100%, VIF will be infinite. That means the model is working with no error. This implies overfitting.

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

Ans 6: Imagine you have two sets of data, like a bunch of test scores from two different classes. You want to see if the scores from both classes are similar. One way to do this is to compare the scores one by one, from lowest to highest. If the scores are similar, you'd expect the highest score in one class to be close to the highest score in the other class, and so on.

A Q-Q plot is like a visual way to do this comparison. It plots the scores from one class against the scores from the other class, with the scores arranged from lowest to highest. If the scores are similar, you'll see a straight line on the plot. If they're not, the line will be curved or scattered.

This is useful for checking assumptions in statistics. For example, if you're running a statistical test that assumes your data is normally distributed, you can use a Q-Q plot to compare your data to a normal distribution. If the plot is a straight line, your data is likely normally distributed.