Assignment-based Subjective Questions

Q1. Based on the analysis you did for the categorical variables of the dataset, what can you infer about their effect on the dependent variable?
Ans –

- The certain inferences that could be obtained from the analysis of the categorical variables are as follows:

    1. Season:

        - The summer season (season_2), fall season (season_3), winter season (season_4) has been considerably positive in their effect on bike demand compared to the spring season, as inferred from their positive coefficients from the OLS summary.
        - Hence, this means one can say that demand for bikes increases in these seasons, with fall having the most positive impact among them.

    2. Weather Situation:

        - As the weather conditions get worse, shared bikes are in extremely low demand.
        - Mist or cloudy weather reduces the counts of bike rentals.
        - Light snow or rain, weathersit_3, is even more adversely affecting.

    3. Month:

        - September would thus suggest a positive impact of demand; this also captures favourable fall weather in the early part of the season.
        - April (mnth_4), November (mnth_11), and December (mnth_12) negatively affect demand for perhaps less good weather conditions or holiday seasons.
- This analysis depicts that seasonality, weather condition, and month of the year can make a huge difference in demand.

Q2. Why is it so important to apply drop_first=True while creating a dummy variable? (2 marks)

Ans –

- drop_first=True while creating the dummy variables is important because it helps

    1. Dummy Variable Trap: In the case of high multicollinearity between the dummy variables, inclusion of all dummy variables corresponding to a categorical variable results in perfect multicollinearity.
    2. You would have ensured that the model used one category as a reference, and the coefficients of the remaining categories be considered with respect to this reference by removing the first category.

Q3. Which is the numerical variable most strongly related to the target variable from the pair-plot?
Ans –

- Based on the pair-plot, the atemp- feeling temperature variable with the value of 0.630685 generally has the highest correlation with the target variable cnt for the simple fact that people go out more when the weather is pleasant - not too hot or cold, hence more use of bicycles.

Q4. How did you check assumptions of Linear Regression after modelling on the training set? (3 marks)
Ans –
• Linearity: Scatter plot was done; actual vs predicted values showed linearity between the two.
• Normality of Residuals: Distribution of the residuals has been checked with the help of a QQ- Plot. The residuals are showing some skewness, hence it suggests further data transformation of the variables.
• Multicollinearity: VIF was computed to make sure that there is no high multicollinearity among predictors. VIF values over 5-10 would indicate multicollinearity for which variables should be removed or other corrective actions taken.

Q5. Based on the final model, which three features are most significant in explaining the demand of shared bikes?
Ans –

- According to the coefficients and p-values of the final model, the 3 most significant features in explaining the demand of the bike are:
    1. Year-yr [25.267]: Bike demand has increased incredibly from 2018 to 2019, reflecting in growing popularity.
    2. Fall Season (season_3) [23.131]: This has a big contribution to increased demand for bikes.
    3. Weather Situation (weathersit_3) [-10.411]: Forcing variable, bad weather conditions, either in the form of snow or rain, severely depresses demand.

General Subjective Questions

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

Ans –

- Linear Regression is a supervised learning algorithm used to predict a continuous dependent variable (y) based on one or more independent variables (X).
- The model is represented by the equation:

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots \beta_n X_n + \epsilon$$

Where:

- β_0 is the intercept.
- $\beta_1, \beta_2, \cdots, \beta_n$ are the coefficients of the independent variables.
- $\in$ is the error term.
- The algorithm minimizes the sum of squared differences between actual and predicted values (Ordinary Least Squares).
- Linear regression assumes linearity, no multicollinearity, homoscedasticity, and normal distribution of residuals.

2. Explain the Anscombe's quartet in detail. (3 marks).

Ans –

- Anscombe's quartet consists of four datasets that have nearly identical statistical properties (e.g., mean, variance, correlation) but different distributions when graphed.
- The key takeaway is that relying on summary statistics alone can be misleading. Visualization is crucial:
    - Dataset 1 shows a typical linear relationship.
    - Dataset 2 has a clear non-linear relationship.
    - Dataset 3 has an outlier affecting the regression.
    - Dataset 4 forms a vertical or horizontal line.

This demonstrates the importance of data visualization before drawing conclusions from statistics.

3. What is Pearson's R? (3 marks).

Ans –

- Pearson's R (correlation coefficient) measures the strength and direction of a linear relationship between two variables. It ranges from -1 to +1:
    - R = +1 indicates a perfect positive correlation.
    - R = −1 indicates a perfect negative correlation.
    - R = 0 means no linear correlation.
- The formula for Pearson's R is:

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}$$

  Where $\bar{X}$ and $\bar{Y}$ are the means of variables X and Y.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks).

Ans –

- Scaling transforms features to a common range to prevent variables with large magnitudes from dominating the model.
- It is crucial for algorithms that are sensitive to feature magnitudes (e.g., gradient descent, k-NN).
- Types of Scaling:
    1. Normalized Scaling: Rescales values between 0 and 1 using the min-max scaling formula:

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

    2. Standardized Scaling: Centers the data around the mean with unit variance:

$$z = \frac{x - \mu}{\sigma}$$

Where μ is the mean and σ is the standard deviation.

4. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks).

Ans –

- An infinite VIF (Variance Inflation Factor) occurs when there is perfect multicollinearity, meaning one independent variable is an exact linear combination of others.
- This causes the regression model to fail to compute the inverse matrix needed for fitting.
- When VIF is infinite, the variable is redundant and should be removed or combined with others.

5. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks).

Ans –

- A Q-Q plot (Quantile-Quantile plot) compares the quantiles of the residuals from a model to the quantiles of a normal distribution.
- It helps assess whether the residuals follow a normal distribution. A straight line in the Q-Q plot suggests normally distributed residuals.
- In linear regression, normality of residuals is an assumption for hypothesis testing (e.g., t-tests for regression coefficients). A deviation from normality could indicate non-normality or the presence of outliers.