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 –

- Bike renting is increasing from March - Oct and from May - Sep demand is high.
- Most bikes are rented on clear or partially cloudy days.
- Bikes were rented more in 2019 as compared to the previous year 2018.
- Bikes were rented more in the fall while least in Spring
- Weekdays and holiday's has not much effect on bike renting but still more bikes were on when there is no holiday.

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

Ans- It is important to use drop_first=True while dummy encoding categorical variables to avoid the dummy variable trap. If we don't use drop_first=True, it would result in perfect multicollinearity as one of the dummy variables becomes redundant.

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

Ans - temperature has the highest correlation with the target variable.

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

Ans - To validate the assumptions of linear regression on the training set, I checked for the following after building the model:
Linearity - Using residual plots
Homoscedasticity - Using residual plots
Normality of errors - Using Q-Q plots
Absence of multicollinearity - Using VIF scores for all independent variables

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

Ans - Based on the final model results, the top 3 features contributing significantly to explaining bike rental demand are:
Temperature
Humidity
Winter season

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Ans –

- Linear regression is a statistical algorithm to model the relationship between a dependent (target) variable and one or more independent (predictor) variables, with the relationship being linear. It fits a straight line over the data by finding best-fit values for the slope and intercept using the least squares method. The line enables us to predict new data values too.
- It assumes the relationship is linear and fits a straight line equation of the form: y = b0 + b1x1 + b2x2 + ... + bn*xn
- b0 is the intercept
- b1 to bn are the regression coefficients for each independent variable
- The goal is to find the best-fit line that minimizes the sum of squared residuals (error between actual and predicted y)
- The least squares method is used to estimate the coefficients
- Makes predictions by using the independent variables, multiplying with coefficients, and adding intercept
- Assumptions of Linear Regression:
    - Linearity - The relationship between dependent and independent variables is linear
    - Homoscedasticity - The error terms have constant variance
    - Independence - No correlation between consecutive error terms
    - Normality - Error terms follow a normal distribution
    - No Multicollinearity - Independent variables are not highly correlated

2. Explain the Anscombe’s quartet in detail.

2. Ans - Anscombe's Quartet

- In 1973, Francis Anscombe demonstrated the limitation of only using summary statistics to describe data
- He created 4 datasets with very different distributions (linear, quadratic, upwards curve, outliers)
- But they had identical descriptive statistics like mean, variance, correlation coefficient (0.816)

- Plotting the datasets makes the differences clearly visible
- Highlights the importance of visualizing data rather than just using numeric statistical summaries
- Descriptive statistics can be insufficient and misleading if not supported by data visualization

3. What is Pearson's R?

Ans –

- Pearson's correlation coefficient (Pearson's R) measures the strength and direction of the linear relationship between two continuous variables.
- Its values range from +1 to -1: +1: Total positive linear correlation 0: No linear correlation -1: Total negative linear correlation
- Positive value: Positive Linear relationship (As X increases, Y increases)
- Negative value: Negative linear relationship (As X increases, Y decreases)
- It only measures linear relationships
- Sensitive to outliers which can wrongly skew results

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

Ans –

- Scaling
    - Features have differing scales, ranges, units
    - Features with a wider range dominate the ML algorithm
    - Makes comparison between features difficult
- Normalization (Min-Max Scaling):
    - Values scaled to a fixed range of 0 to 1
    - Done using: x_scaled = (x - min(x)) / (max(x) - min(x))
    - Makes features more comparable
    - Reduces influence of outliers
- Standardization (Z-score normalization):
    - Values scaled to have mean 0 and SD 1
    - Done by: $z = (x - \mu)/\sigma$
    - Makes distribution normal
    - Outlier treatment is the same as the original data
- Difference:
    - Normalization bounds range
    - Standardization centers data around mean

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

Ans -

- VIF or Variance Inflation Factor quantifies multicollinearity
- High VIF indicates increased standard errors of coefficients
- Makes the estimates unstable and difficult to interpret

- Infinite VIF arises when one independent variable is a perfect linear function of others
- Creates a singular matrix which cannot be inverted during regression
- Requires removal of redundant variables causing perfect collinearity

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 method to check if a set of data follows a specified distribution, typically the normal distribution.

In the context of linear regression, a Q-Q plot is used to validate one of the key assumptions - that the error terms (residuals) are normally distributed.

How Q-Q Plots Work:

- The quantiles of the sample data set are plotted against the theoretical quantiles of a standard normal distribution
- If the sample data follows a normal distribution, the points in the Q-Q plot will lie approximately on the 45 degree reference line
- Deviation from this straight line indicates deviation from normality

Use of Q-Q Plots in Linear Regression:

- The residuals from a fitted linear regression model are plotted on the Q-Q plot
- If the points closely follow the reference line, we can assume a normal distribution of errors
- Any significant deviations e.g. curvature or "S" shaped pattern indicates non-normal errors

Importance of Q-Q Plots:

- Assessing the normality of errors is critical to ensure the validity of linear regression
- Non-normal errors can undermine inferences and predictions made by the model
- Q-Q plots provide fast, visual assessment instead of just relying on numerical tests
- Helps catch multiple departures like skewness, heavy tails, outliers
- Critical check before further analysis and interpretation of the fitted model

In summary, Q-Q plots allow easy visualization of error distribution to validate a key linear regression assumption, instead of just using statistical tests. This graphical assessment is very important.