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)

From the final model following categorical variable are relevant in predicting demand:

1. Positively correlated categorical variables are:
   a. yr (Coeff: + 0.2515) ## Year in which demand is being predicted
   b. winter (Coeff: + 0.1359) ## Season is summer or not
   c. Sept (Coeff: + 0.0403) ## Month is September or not
2. Negatively correlated categorical variables are:
   a. holiday (Coeff: - 0.0810) ## weather day is a holiday or not
   b. July (Coeff: - 0.0928) ## Month is July or not
   c. Cloudy (Coeff: - 0.0458) ## weathersit is 'Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist' or not
   d. LRain (Coeff: - 0.2586) ## weathersit is 'Light Snow, Light Rain + Thunderstorm + S cattered clouds, Light Rain + Scattered clouds' or not

Among the categorical variables, ‘yr’ and ‘weathersit’ are variables which has the most impact on demand

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

Using drop_first=True during dummy variable creation drops the first column of dummy variable. This helps in reducing correlation between dummy variables and protects the model from dummy variable trap

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

‘temp’ and ‘atemp’ has the highest correlation of 0.63 with target variable

4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

1. No Multicollinearity – All relevant features have VIF < 5

2. Homoscedasticity – Scatter plot of residual shows no prominent pattern, which means that the residuals have constant variance

3. Normal distribution of error – Error terms are normally distributed with mean = 0

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

Top three features are:

1. temp (Coeff: + 0.7093) ## Temp on the day
2. LRain (Coeff: - 0.2586) ## weathersit is 'Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds' or not
3. yr (Coeff: + 0.2515) ## Year in which demand is being predicted

General Subjective Questions

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

Linear regression is a statistical method for modelling the relationship between a dependent variable (Y) and one or more independent variables (X). The goal is to find the best-fitting line (or hyperplane) that minimizes the sum of squared differences between predicted and actual values. The model is represented by the equation:

$Y=\beta 0+\beta 1X+\epsilon$

Where, $\beta 0$ is the intercept, $\beta 1$ is the slope, and ϵ is the error term.

The method of least squares is used to estimate the coefficients by minimizing the sum of squared differences. Common metrics for evaluation include R-Square and Mean Squared Error. Linear regression assumes linearity, independence, homoscedasticity, and normality of residuals. Multiple linear regression extends the model to include multiple independent variables.

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

Anscombe's quartet is a set of four datasets that have nearly identical simple descriptive statistics, including means, standard deviations, and correlation coefficients, but exhibit vastly different patterns when graphically visualized. It was created by the statistician Francis Anscombe to emphasize the importance of graphically exploring data before drawing conclusions based solely on summary statistics. This quartet is often used to highlight the limitations of relying solely on numerical summaries and the importance of data visualization in understanding the underlying patterns in the data.

Key takeaways from Anscombe's quartet are:

1. Visual Inspection is Crucial: Descriptive statistics alone may not reveal the true nature of the data. Graphical exploration helps identify patterns, outliers, and relationships.
2. Robustness of Summary Statistics: The quartet illustrates that the same summary statistics can apply to datasets with different underlying structures.
3. Importance of Data Visualization: Anscombe's quartet emphasizes the value of data visualization in gaining insights and avoiding misinterpretations.

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

Pearson's correlation coefficient, r, is a statistical measure that quantifies the strength and direction of a linear relationship between two continuous variables. It ranges from -1 to 1, where:

- $r=1$ indicates a perfect positive linear correlation,
- $r=-1$ indicates a perfect negative linear correlation,
- $r=0$ indicates no linear correlation.

The closer $|r|$ is to 1, the stronger the linear correlation.

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

Scaling is a preprocessing technique in data analysis and machine learning that involves transforming the numerical features of a dataset to a standard range or distribution. The goal of scaling is to ensure that all variables contribute equally to the analysis, as many machine learning algorithms are sensitive to the scale of the input features.

Reasons for scaling:

1. **Algorithm Sensitivity:** Some machine learning algorithms are sensitive to the scale of input features. For example, distance-based algorithms like k-nearest neighbors and clustering algorithms may be influenced by the scale of variables.
2. **Convergence Speed:** Optimization algorithms, like gradient descent used in many machine learning models, often converge faster when features are on a similar scale.
3. **Model Performance:** Scaling can improve the performance and interpretability of models. It prevents features with larger scales from dominating those with smaller scales.

Difference between Normalized and Standardized Scaling:

1. Normalized scaling transforms values to a specific range (e.g., 0 to 1), while standardized scaling transforms values to have a mean of 0 and a standard deviation of 1.
2. Normalized scaling is sensitive to outliers, while standardized scaling is more robust.
3. Standardized scaling is often preferred in cases where the distribution of features is unknown or may be skewed.

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

The Variance Inflation Factor (VIF) is a measure used in regression analysis to quantify how much the variance of an estimated regression coefficient increases when your predictors are correlated. VIF values help assess multicollinearity, where predictor variables in a regression model are highly correlated.

The formula for the VIF of a predictor variable Xi is given by:

$$\mathrm{VIF}_i = \frac{1}{1-R_i^2}$$

Here, Ri-Square is the R-Square value obtained by regressing the predictor Xi against all other predictor variables.

If the VIF is infinite, it typically indicates perfect multicollinearity. Perfect multicollinearity occurs when one or more predictor variables in a regression model can be exactly predicted by a linear combination of other variables.

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

A Q-Q (quantile-quantile) plot is a graphical tool used in statistics to assess whether a dataset follows a particular theoretical distribution. It compares the quantiles of the observed data against the quantiles of a theoretical distribution, typically a normal distribution. Q-Q plots are particularly useful in linear regression for checking the assumption of normality of residuals

Interpretation of a Q-Q Plot:

1. **Straight Line:** If the points on the Q-Q plot fall approximately along a straight line, it suggests that the residuals follow a normal distribution.
2. **Deviation from Line:** Deviations from the straight line may indicate departures from normality. Skewness or outliers in the residuals can be visually detected.
3. **Tails of the Plot:** Examination of the tails of the Q-Q plot can reveal information about the tails of the distribution and whether they match the theoretical distribution.