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)

   To infer the effect of the categorical variables (season, yr, mnth, weathersit) on the dependent variable CNT, we rely on the p-values and co-efficients.

   A positive co-efficient will indicate an increase in the sales of bike.

   The p-values determines if the fit is statistically significant.

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

   The drop_first=True is used while creating the dummy variable. By giving this statement, we will have one variable reduced during analysis. This will ensure the model does not impacted by multi collinearity.

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

   Temperature variable 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? (3 marks)

   After building the model, we validated the assumptions by using
   a. R-squared and Adjusted R square
   b. P-value and co-efficient
   c. Higher F-statistic value
   d. Checking the VIF value
   e. Residual Analysis on the Training data
   f. Model evaluation using Test data R quare

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

   The top 3 features are –
   1. TEMP
   2. YR
   3. weather

General Subjective Questions

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

   Linear regression is a way to find a straight line that best represents the relationship between two things. Imagine plotting points on a graph, where one thing (like studying time) is on the x-axis, and another thing (like exam score) is on the y-axis. Linear regression helps us find the line that comes closest to touching all those points. This line helps us make predictions, like estimating your exam score based on how much you study. It's a simple but powerful tool for understanding and predicting relationships in data.

2. E_x_p_l_a_i_n_ _t_h_e_ _A_n_s_c_o_m_b_e_’s_ _q_u_a_r_t_e_t_ _i_n_ _d_e_t_a_i_l_._ _(3 marks)

   Anscombe's quartet is a famous statistical demonstration that highlights the importance of visualizing data and not relying solely on summary statistics. It consists of four distinct datasets, each with 11 data points, which have nearly identical summary statistics (mean, variance, correlation) but exhibit vastly different patterns when plotted.

   Here are key points about Anscombe's quartet:

   - **The Datasets**: Anscombe's quartet contains four datasets, labeled I, II, III, and IV.
   - **Similar Summary Statistics**: Each dataset has nearly identical summary statistics, such as the mean, variance, and correlation coefficient. This makes them appear very similar when looking at just the numbers.
   - **Different Patterns**: When you visualize these datasets by plotting them, you'll see that they have very different shapes and relationships between variables. One might be linear, another quadratic, another might have outliers, and so on.
   - **Statistical Inference**: It underscores the importance of not making strong statistical inferences based solely on summary statistics and highlights the value of exploratory data analysis and visualization.

   In summary, Anscombe's quartet is a powerful reminder that looking at data visually can reveal insights that summary statistics might miss. It encourages data analysts and statisticians to explore and visualize their data before drawing conclusions or making decisions based solely on mathematical summaries.

3. W_h_a_t_ _i_s_ _P_e_a_r_s_o_n_’s_ _R_?_ _(3 marks)

Pearson's correlation coefficient, often denoted as "r" or "Pearson's r," is a statistical measure of the strength and direction of the linear relationship between two continuous variables. It quantifies how well the variation in one variable can be predicted by the variation in another variable when both are measured on an interval or ratio scale.

Pearson's correlation coefficient is a widely used statistic to measure the strength and direction of linear relationships between two continuous variables. It provides valuable insights into how variables relate to each other, but it's important to interpret the results with caution and consider the context of the data.

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

Scaling in data preprocessing is the process of adjusting the range or distribution of your data. It's done for several reasons:

Two common types of scaling are:

- **Normalized Scaling (Min-Max Scaling)**: This scales data to a specific range (usually 0 to 1), preserving relative relationships between values.
- **Standardized Scaling (Z-Score Scaling)**: This scales data to have a mean of 0 and a standard deviation of 1. It's useful when the distribution of the data matters.

In simple terms, scaling helps make your data more compatible with different machine learning methods, improves model performance, and ensures that no single variable overwhelms the others.

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

(3 marks)

VIF formula is 1 / (1 – R2), whenever R square value is 1. Then 1/0 will result in infinte

In certain situations, variables are created in such a way that they are perfectly related. For example, creating dummy variables for a categorical variable without omitting one category (dummy variable trap) can lead to perfect multicollinearity.

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

(3 marks)

Q-Q plot (Quantile-Quantile plot) is a graphical tool used in statistics to assess whether a dataset follows a particular theoretical probability distribution, such as the normal distribution. It helps compare the quantiles (percentiles) of the observed data against the quantiles of the expected theoretical distribution. The Q-Q plot is particularly useful for checking the assumption of normality in linear regression and other statistical analyses.