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)**
    - Categorical variables in the dataset include year, month, holiday, weekday, workingday, season_fall, season_spring, season_summer, season_winter, weather_clear, weather_light snow, and weather_misty.
    - From the analysis, we can infer that these categorical variables may have an effect on the dependent variable (count) based on their one-hot encoded values. For instance, season and weather variables indicate weather conditions and seasons, which can influence bike demand. Similarly, weekday and workingday can also impact demand patterns.

2. **Why is it important to use drop_first=True during dummy variable creation? (2 mark)**
   When creating dummy variables for categorical features, setting drop_first=True avoids multicollinearity issues. It drops one level of each categorical variable to prevent perfect multicollinearity, which can lead to unstable and less interpretable models. The omitted level becomes the reference category, and the remaining levels represent the presence or absence of a specific category. This approach helps improve model interpretability and generalization.

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

4. **How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)**
   After building the Linear Regression model on the training set, residual analysis was performed to validate the assumptions.
   Checked residuals vs. predicted values: The residuals should be randomly scattered around the horizontal axis. Any clear patterns might indicate non-linearity or heteroscedasticity.
   Checked the histogram of residuals: The residuals should be approximately normally distributed. Deviations from normality might suggest model misspecification.
   These checks help ensure that the assumptions of linearity, homoscedasticity, and normally distributed residuals are met.

5. **Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)**
   workingday &weekday , season and weather

General Subjective Questions

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

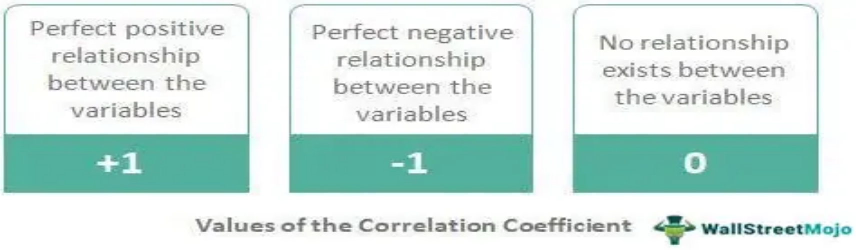
   Linear Regression is a fundamental and widely used statistical algorithm for modelling the relationship between a dependent variable and one or more independent variables. It's a supervised learning algorithm used for both regression (predicting continuous values) and classification (predicting categorical values) tasks.
   Let's dive into the details of the Linear Regression algorithm:
   Objective: The primary goal of Linear Regression is to find the best-fitting linear relationship (a straight line) between the independent variable(s) and the dependent variable. This relationship is represented by the equation:
   $y=mx+b$
   y is the dependent variable (target)
   x is the independent variable (feature)
   m is the slope of the line
   b is the y-intercept
   The goal is to determine the optimal values of m and b that minimize the difference between the predicted values and the actual values.
   Assumptions:
   - Linearity: The relationship between the independent and dependent variables is linear.
   - Independence: The residuals (differences between predicted and actual values) are independent and identically distributed.
   - Homoscedasticity: The residuals have constant variance across all levels of the independent variables.
   - Normality: The residuals are normally distributed.

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

   Anscombe's quartet comprises a set of four dataset, having identical descriptive statistical properties in terms of means, variance, R-Squared, correlations, and linear regression lines but having different representations when we scatter plot on graph

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

   Pearson correlation coefficient, also known as Pearson R statistical test, measures the strength between the different variables and their relationships. Therefore, whenever any statistical test is conducted between the two variables, it is always a good idea for the person analyzing to calculate the value of the correlation coefficient to know how strong the relationship between the two variables is.



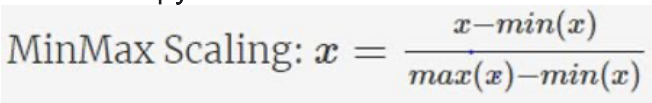
Perfect positive relationship between the variables	Perfect negative relationship between the variables	No relationship exists between the variables
+1	-1	0

Values of the Correlation Coefficient WallStreetMojo

4. **What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)**
   It is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm.
   Normalization/Min-Max Scaling:

   - It brings all of the data in the range of 0 and 1

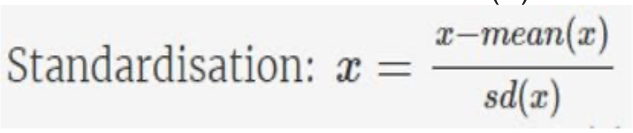
     sklearn.preprocessing.MinMaxScaler helps to implement normalization in python.



$$\text{MinMax Scaling: } x = \frac{x - min(x)}{max(x) - min(x)}$$

Standardization Scaling:

- Standardization replaces the values by their Z scores. It brings all of the data into a standard normal distribution which has mean (μ) zero and standard deviation one (σ).



$$\text{Standardisation: } x = \frac{x - mean(x)}{sd(x)}$$

5. **You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)**
   If there is perfect correlation, then VIF = infinity. A large value of VIF indicates that there is a correlation between the variables. If the VIF is 4, this means that the variance of the model coefficient is inflated by a factor of 4 due to the presence of multicollinearity.

6. **What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)**
   The purpose of Q Q plots is to find out if two sets of data come from the same distribution. A 45 degree angle is plotted on the Q Q plot; if the two data sets come from a common distribution, the points will fall on that reference line. A Q Q plot showing the 45 degree reference line.