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)
   - “season”, ”yr”, ”mnth”, ”weathersit”, ”holiday” are independent variables that are possible to have a high influence on the dependent variable, whereas “weekday”, ”workingday” may influent less on the dependent variable.

2. Why is it important to use drop_first=True during dummy variable creation? (2 mark)
   - Using drop_first=True during dummy variable creation helps to prevent multicollinearity, improve model interpretability, and simplify the model by excluding one category as the baseline.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)
   - “atemp” 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)
   - (1) Residual analysis: the difference between actual and predicted value should be normally distributed with a mean of 0. (2) Homoscedasticity: plot the residuals against the independent variables and it should be consistent. (3) Linearity: plot the actual values against the independent variables, the data points should follow a straight line. (4) Multicollinearity: the VIF for each independent variable should be less than 10.

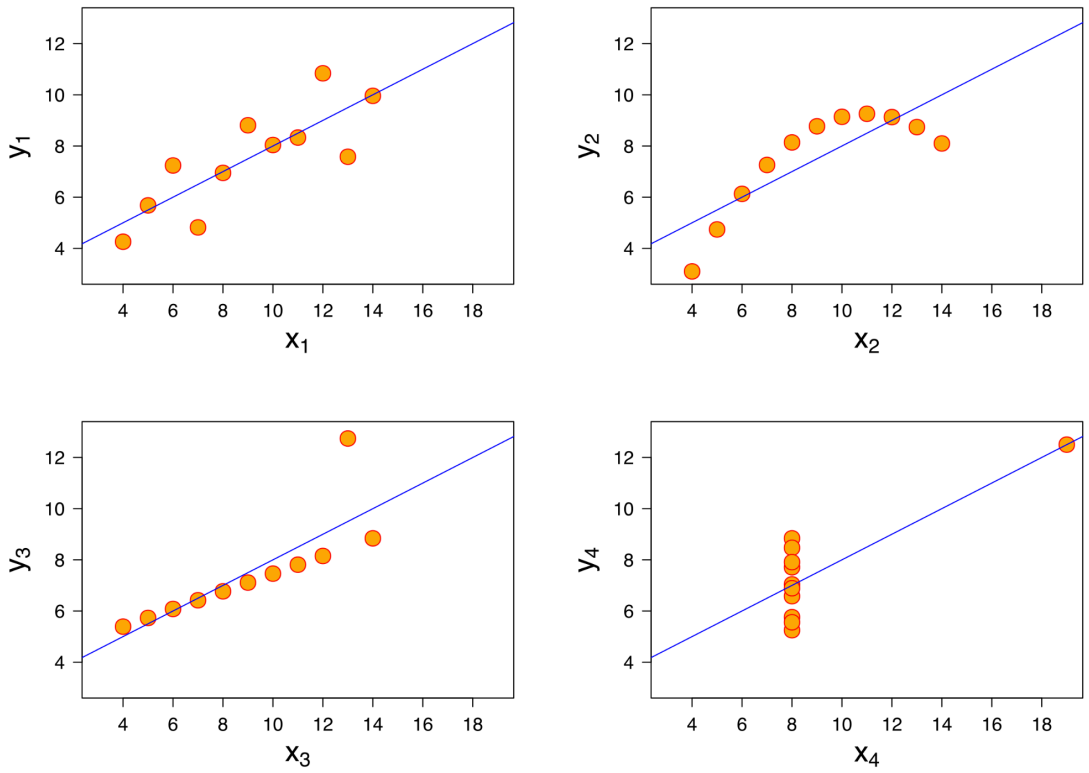
5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)
   - “temp”(temperature in Celsius), ”yr”(year), ”LightSnow”(LightSnow weathersit) are the 3 top features.

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)
   - Linear regression is a supervised learning algorithm used for predicting a continuous dependent variable based on one or more independent variables. It assumes a linear relationship between the independent variables and the dependent variable. The algorithm aims to find the best-fitting line or hyperplane that minimizes the sum of the squared differences between the predicted and actual values. It contains the following steps:
     - Data preparation and EDA
     - Model building and Feature election
     - Model Evaluation and optimization

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

- Anscombe's quartet refers to a set of four datasets that have nearly identical statistical properties but exhibit different patterns when visualized. The quartet was introduced by the statistician Francis Anscombe in 1973 to emphasize the importance of visualizing data and not relying solely on summary statistics. The 4 datasets have the features below:



- Top left appears to be a simple linear relationship, corresponding to two variables correlated where y could be modelled as gaussian with mean linearly dependent on x.
- Top right is not linear, and the Pearson correlation coefficient is not relevant. A more general regression and the corresponding coefficient of determination would be more appropriate.
- Bottom left relationship is linear, but should have a different regression line. The calculated regression is offset by the one outlier which exerts enough influence to lower the correlation coefficient from 1 to 0.816.
- Bottom right shows an example when one high-leverage point is enough to produce a high correlation coefficient, even though the other data points do not indicate any relationship between the variables

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

- Pearson's R is a statistical measure that quantifies the strength and direction of the linear relationship between two continuous variables. It ranges from -1 to +1. +1 indicates a perfect positive linear relationship; -1 indicates a perfect negative linear relationship; 0 it indicates no linear relationship between the variables.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)
   - Scaling is a preprocessing step in data preparation that transforms the values of variables to a specific range or distribution. It is performed to bring all variables to a similar scale, ensuring that no variable dominates or biases the analysis due to its larger magnitude.
   - The reasons for scaling are (1) Equalizing Variable Importance; (2) Enhancing Model Performance; (3) Interpretation of coefficients or features more meaningful.
   - Difference between normalized scaling and standardized scaling:
     - (1) Normalized Scaling transforms variables to a 0-1 range. It uses the minimum and maximum values of each variable to perform the transformation. The formula for normalized scaling is:X' = (X - X_min) / (X_max - X_min). Normalized scaling preserves the relative relationships between data points but can be sensitive to outliers.
     - (2) Standardized Scaling (Z-score Scaling): Standardized scaling transforms variables to have a mean of 0 and a standard deviation of 1. The formula for standardized scaling is:X' = (X - mean) / standard deviation. Standardized scaling centers the data around zero and ensures that it has a unit standard deviation. It is less sensitive to outliers compared to normalized scaling.
     - The choice between the two scaling techniques depends on the specific requirements of the analysis or the algorithm being used.

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)
   - VIF = 1 / (1 - R^2). When VIF is infinite, R^2 = 1. It means that one predictor variable can be expressed as an exact linear combination of other predictor variables. The infinite VIF values indicate that one or more predictor variables in the model are redundant due to perfect multicollinearity. It implies that the information provided by these variables is already captured by other variables in the model. We need to identify and remove the redundant variables from the model.

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 plot (Quantile-Quantile plot) is a graphical tool used to assess the distributional similarity between a sample of data and a theoretical distribution. The Q-Q plot compares the quantiles of the observed data against the quantiles of the theoretical distribution. The process involves ordering the data from smallest to largest and calculating the corresponding quantiles for both the observed data and the theoretical distribution. These quantiles are then plotted against each other, typically on a scatter plot.
   - Q-Q plots provide a visual and intuitive way to assess the distributional assumptions of linear regression models. They help identify departures from normality, detect outliers, evaluate model fit, and compare data against alternative distributions, thereby assisting in making informed decisions and improving the accuracy of regression analysis.