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?
   - season : majority of the bike bookings were happening in season_2 and season_3, with median between 4000-6000 bookings.
   - weathersit : majority of the bike bookings were happening in weathersit_1 followed by weathersit_2, with median at 5000 and 4000 bookings respectively.
   - mnth : majority of the bike bookings were happening in the months 5,6,7,8 & 9 with a median close to 5000 bookings per month
   - workingday : workingday variable shows almost equal booking on working and non-working days with a median close to 5000 bookings for the 2 years period.
   - weekday: weekday variable shows more or less the same trend with a median between 4000 to 5000 bookings.
   - holiday: Many bikes were booked when it was not a holiday which indicates bias. Therefore, holiday may not be an accurate predictor of bike booking.

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

   The concept of dummy variables is to create '(n-1)' levels of dummy variables for categorical columns (or variables) having 'n' levels. Therefore, the original categorical column becomes redundant, while the data can still be comprehended in its absence. Hence, the 'drop_first = True' term is used, to remove this original categorical column.

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

   'temp' and 'atemp' variables have the highest correlation with the target variable (cnt).

4. How did you validate the assumptions of Linear Regression after building the model on the training set?
   - Residual Analysis on the train dataset to get normal distribution of error terms.
   - Plotting the numerical predictor variables against the target varible (cnt) to observe a linear relationship between them.
   - Test of Multicollinearity using VIF values method, to observe multicollinearity among predictor variables.

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?
   - Temperature (temp) with a correlation coefficient of 0.5174, meaning a unit increase in temp variable would increase the bike demand by 0.5174 units.
   - weathersit_3 (i.e., Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds) with a correlation coefficient of -0.2828, meaning a unit increase in weathersit_3 variable would decrease the bike demand by 0.2828 units.
   - Year (yr) with a correlation coefficient of 0.2325, meaning a unit increase in yr variable would increase the bike demand by 0.2325 units.

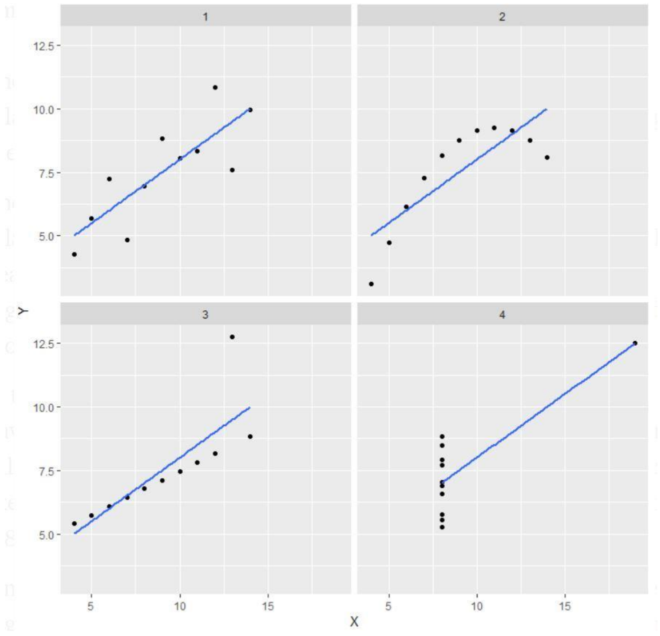
General Subjective Questions

1. Explain the linear regression algorithm in detail.
   - Linear Regression is a form of a Supervised Machine learning algorithm. It predicts/forecasts the best-fit linear relationship between the independent variables (or predictor variables) and a single continuous dependent variable (or target variable). This is achieved using the Residual Sum of Squares.
   - Linear Regression can be further categorized into two types:
     - ➢ Simple linear regression – when 1 independent variable is involved in the model.
     - ➢ Multiple linear regression – model contains more than one independent variables.
   - The linear regression model is represented by the formula:

     $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n$

     This model explains marginal changes in the target variable 'y' for marginal changes in the independent variable 'x_i'. For example, β_1 is the change in 'y' for a unit increase in the magnitude of x_1, when all other variables are kept constant. Similarly, the intercept β_0, is the response when all the other predictor variables are treated as a constant or considered zero.
   - The assumptions of Linear Regression are:
     - ➢ *Linearity Assumption* – A linear relationship is assumed between the independent and dependent variables.
     - ➢ *No Multicollinearity among the independent variables* – The independent variables are assumed to be linearly independent of each other.
     - ➢ *Normal Distribution of residuals with a mean value of zero* – If the error terms are not normally distributed, their randomness does not hold true. Then, the model fails to explain the relation to the data.
     - ➢ *Constant variance of residual terms (homoscedasticity)* – If a funnel pattern exists between the residual and fitted plot (heteroscedasticity), then the model performance decreases and the confidence interval gets relatively high or low values.
     - ➢ *No correlation among the error terms* – error terms are independent of each other.

2. Explain the Anscombe's quartet in detail.
   - Anscombe's Quartet is a collection of four data sets which have the same statistical observations, but vary in distribution when plotted on scatter plots.
   - This highlights the importance of data visualisation before applying the different algorithms for model building. Plotting the data is necessary to understand the distribution and detect anomalies in the data like outliers, non-linearity, etc.
   - The data sets provide the same descriptive statistical information, which involves the variance, mean and correlation of all x,y points in the four data sets.
   - Explanation of the four data sets:
     - ➢ Dataset 1: the linear regression model fits the data well.
     - ➢ Dataset 2: the linear regression can only model linear relationships and cannot handle any other data.
     - ➢ Dataset 3: the linear regression model is sensitive to outliers and hence is not a good fit to the data.
     - ➢ Dataset 4: the linear regression model is sensitive to outliers and hence is not a good fit to the data.



3. What is Pearson’s R?
    - Pearson’s correlation coefficient is a statistic measure of the **linear relationship** or association, between two continuous variables. It is based on the method of covariance. It provides information on the magnitude of association and direction of the relationship.
    - The Pearson’s correlation coefficient ranges between -1 and +1:
        - r = 1 signifies linear data with a positive slope (both variables change in the same direction i.e., increase-increase or decrease-decrease)
        - r = -1 signifies linear data with a negative slope (both variables change in different directions i.e., increase-decrease or decrease-increase)
        - r = 0 signifies non-linear association
        - 0 < r < 5 signifies a weak association
        - 5 < r < 8 signifies a moderate association
        - r > 8 signifies a strong association
    - The following points to remember when interpreting correlations:
        - Correlations can or cannot indicate **causal relations**. However, the inverse does not always hold true i.e., causal relations may or may not indicate correlation between the variables.
        - Correlations are sensitive to **outliers**. These outliers can be detected by inspecting the scatterplot.
    - Pearson correlation between variables X and Y is calculated by:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

where, n is the sample size

x_i and y_i are the individual sample data points

$\bar{x}$ and $\bar{y}$ are the sample means for x and y respectively

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?
    - Scaling is a method of pre-processing the data before model building. The independent variables are normalized to convert the data within a specific range. It also speeds up the optimization of cost functions in an algorithm.
    - Generally, the data set that is collected contains features of varying magnitudes, range and units. Not scaling data leads to incorrect modelling, because only magnitudes will be considered for the algorithm, and not the units. Scaling converts all the variables in the dataset to the same level of magnitude. It does not affect the statistic parameters like p-values, t-statistic, R-squared, F-statistic, etc.
    - **Normalization** or **Min-Max scaling** is a scaling technique which rescales the data values in the range between 0 and 1.

- ➢ The formula for normalization is given as:

$$X^{'} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

  Here,
  Xmax and Xmin are the maximum and the minimum values of the feature variables respectively

- **Standardization** is a scaling technique which centres the data values around the mean with a unit standard deviation. In other words, the data is rescaled with a mean of zero and a standard deviation equal to one.
  - ➢ The formula for normalization is given as:

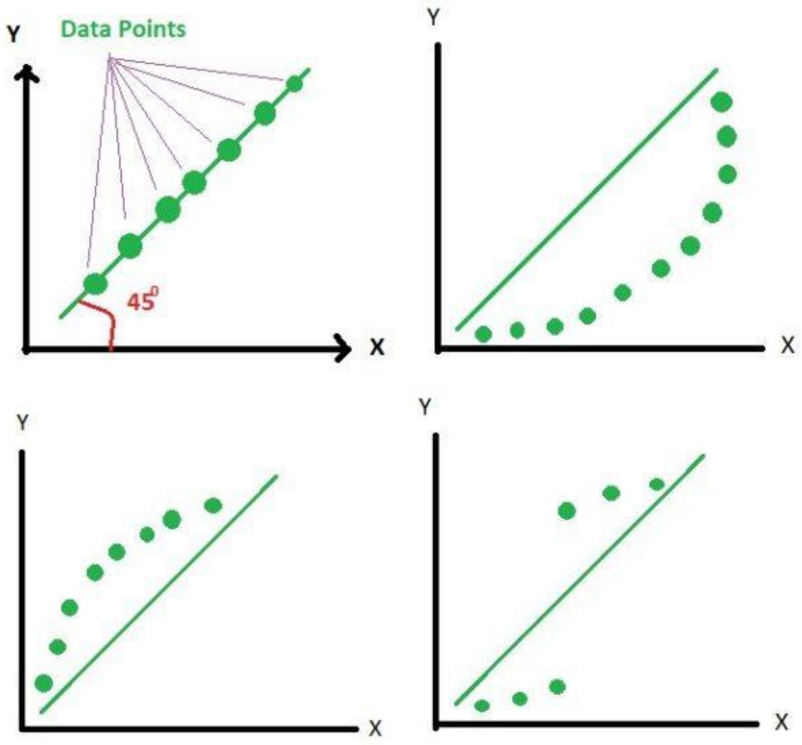
$$X^{'} = \frac{X - \mu}{\sigma}$$

  Here,
  μ is the mean and σ is the standard deviation of the feature variables respectively.
- Unlike standardization, normalization has a bounding range (0 to 1). Therefore, normalization loses data about the outliers.

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?
   - An infinite VIF value indicates a perfect correlation between two predictor/independent variables. For perfect correlation, R-squared is 1, which gives 1/(1-R^2) = ∞. To prevent this, we need to drop one of the variables out of two with perfect multicollinearity from the dataset.
   - Multicollinearity is a phenomenon where the independent variables are correlated in a regression model. High correlation between the variables gives an incorrect model and wrong interpretation of results, as the variables are required to be independent of each other.
   - VIF > 10 indicates high correlation and the variable should be eliminated from the model.
     VIF > 5 may be okay, but needs inspection whether to include or eliminate the variable.
     VIF < 5 is an acceptable value of VIF, for including the variable in the model.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.
   - Quantile-Quantile (Q-Q) plot, helps in the assessment of whether a dataset belongs to some theoretical Normal, Uniform or exponential distribution. It further provides information whether two datasets belong to populations with a common distribution.
   - In linear regression where training and test datasets are available separately, we can confirm using Q-Q plot that the datasets belong to populations with the same distribution.
   - Advantages of Q-Q plots:
     - ➢ It can be applied to sample datasets too
     - ➢ Other inferences from this plot are presence of outliers, changes in scale and symmetry, etc.
   - Uses of Q-Q plots:
     - ➢ When two data sets,

     i. belong to populations having a common distribution
     ii. have similar scale and location
     iii. have similar shape of data distribution
     iv. have similar tail behaviour
   - A Q-Q plot plots the quantiles of the first dataset to the quantiles of the second dataset.
     - ➢ **Similar distribution**: The quantiles values lie on or close to straight line at a 45-degree angle from x -axis
     - ➢ **Y-values < X-values**: The y-quantiles are lower than the x-quantiles.
     - ➢ **X-values < Y-values**: The x-quantiles are lower than the y-quantiles.
     - ➢ **Different distribution**: The quantiles values lie away from the straight line at a 45-degree angle from x -axis.


Y
Data Points
45⁰
X
Y
X
Y
X
Y
X