Linear Regression Subjective Questions

1. **From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?**

   From the Exploratory data analysis the following could be inferred;

   a. The total rental seems to be higher for season 2 and 3 i.e. for summer and fall and starts to increase from March and remains high till October (very much in sync with the season).
   b. Also, rentals have increased from 2018 to 2019.
   c. It is also found to be highest for weather 1 (Clear Weather) and lowest for weather 3 (light snow or rain). No rentals are there for weather 4 (heavy rain or snow).
   d. Not much variation is visible w.r.t. holiday, weekday or working day.

   I also visualized the variation w.r.t. casual and registered rentals. No. of Registered users is also showing the same trend as total rentals except for the fact that registered rentals are higher in working day than in holiday and also no. of registered rentals rises mildly during the mid of the week. Distribution of no. of casual rentals with respect to ther categorical variables shows many outliers, clearly showing the unpredictable nature of number of casual rentals. Also, no. of casual rentals increases on weekends and decreased on weekdays (similar to increases on holidays and reduces on workdays)
   It can be said that distribution of casual and registered rentals is significantly different due to the purpose of the rental. While casual users mainly rent bikes for leisure and recreation, registered users use them for commuting to work.

   After the final model creation the following categorical variable were found to be significant;
   a. Holiday
   b. Season
   c. Year
   d. Month
   e. Weather

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

   drop_first = True is important to use, as it helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables.
   Syntax - drop_first: bool, default False, which implies whether to get k-1 dummies out of k categorical levels by removing the first level. Let's say we have 3 types of values in Categorical column and we want to create dummy variable for that column. If one variable is not A and B, then It is obvious C. So we do not need 3rd variable to identify the C.

3. **Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?**
   'temp' 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?**
   I have validated the assumption of Linear Regression Model based on below 5 assumptions -
   - Normality of error terms
     - Error terms should be normally distributed
   - Multicollinearity check
     - There should be insignificant multicollinearity among variables.
   - Linear relationship validation
     - Linearity should be visible among variables
   - Homoscedasticity
     - There should be no visible pattern in residual values.
   - Independence of residuals
     - No auto-correlation

5. **Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?**
   - Temperature
   - Humidity
   - Year

General Subjective Questions

1. **Explain the linear regression algorithm in detail.**

   Linear Regression is a statistical method for modeling the relationship between a dependent variable and one or more independent variables. It assumes a linear relationship, represented by the equation:

$$y = \beta_0 + \beta_1 x + \epsilon$$

   where:

   -y is the dependent variable.

   -x is the independent variable.

   -$\beta_0$ is the intercept.

-$\beta_1$ is the slope.

-$\epsilon$ is the error term.

The goal is to find the best-fitting line by minimizing the sum of squared residuals (differences between observed and predicted values). This is achieved using the Least Squares Method. The coefficients $\beta_0$ and $\beta_1$ are estimated to minimize:

$$\sum y_i - (\beta_0 + \beta_1 x_i)^2$$

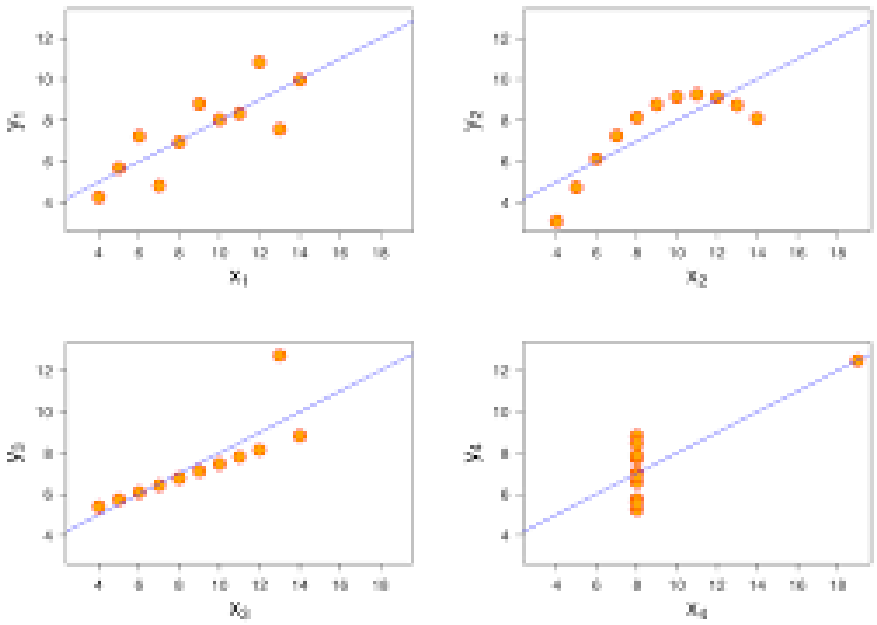
Once the model is fitted, it can be used to predict the dependent variable for new values of the independent variable(s). Linear Regression is simple and interpretable, making it widely used for predictive analysis.

2. Explain the Anscombe's quartet in detail.

Anscombe's quartet is a set of four datasets created by the statistician Francis Anscombe in 1973. These datasets have nearly identical simple descriptive statistics, such as mean, variance, correlation, and linear regression line. Despite these similarities, they have very different distributions and appearances when graphed. Anscombe's quartet demonstrates the importance of graphing data before analyzing it and cautions against relying solely on statistical properties.

The four datasets are as follows:

1. **Dataset I:** A simple linear relationship with some random noise.
2. **Dataset II:** A nonlinear relationship, showing a clear curvature.
3. **Dataset III:** A linear relationship but with an outlier that significantly influences the regression line.
4. **Dataset IV:** A horizontal line with one outlier, distorting the statistical properties.



For each dataset:

- The mean of x is 9.
- The mean of y is approximately 7.50.
- The variance of x is 11.
- The variance of y is around 4.12.
- The correlation between x and y is about 0.816.
- The regression line is y=3.00+0.5xy = 3.00 + 0.5xy=3.00+0.5x.

Visualizing these datasets shows their distinct characteristics, highlighting the necessity of graphical analysis to fully understand data distributions and relationships.

### 3. What is Pearson's R?

Pearson's r is a numerical summary of the strength of the linear association between the variables. If the variables tend to go up and down together, the correlation coefficient will be positive. If the variables tend to go up and down in opposition with low values of one variable associated with high values of the other, the correlation coefficient will be negative. The Pearson correlation coefficient, r, can take a range of values from +1 to -1. A value of 0 indicates that there is no association between the two variables. A value greater than 0 indicates a positive association; that is, as the value of one variable increases, so does the value of the other variable. A value less than 0 indicates a negative association; that is, as the value of one variable increases, the value of the other variable decreases. This is shown in the diagram below:

4. **What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?**

Feature Scaling is a technique to standardize the independent features present in the data in a fixed range. It is performed during the data pre-processing to handle highly varying magnitudes or values or units. If feature scaling is not done, then a machine learning algorithm tends to weigh greater values, higher and consider smaller values as the lower values, regardless of the unit of the values. Example: If an algorithm is not using feature scaling method then it can consider the value 3000 meter to be greater than 5 km but that's actually not true and in this case, the algorithm will give wrong predictions. So, we use Feature Scaling to bring all values to same magnitudes and thus, tackle this issue.

| Aspect | Normalized Feature Scaling | Standardized Feature Scaling |
|---|---|---|
| Definition | Rescales the features to a fixed range, usually [0, 1] or [-1, 1]. | Rescales the features so they have a mean of 0 and a standard deviation of 1. |
| Output Range | [0, 1] or [-1, 1] depending on the chosen range | Typically unbounded (depends on the data distribution) |
| Effect on Data | Compresses all data points into a specific range. | Centers the data around 0 and scales it based on standard deviation. |
| Sensitivity to Outliers | High. Outliers can significantly affect the min and max values. | Moderate. Outliers affect mean and standard deviation but less so than min-max. |
| Use Cases | When data needs to be within a specific range, such as in image processing or neural networks. | When the assumption of normal distribution is reasonable, such as in linear regression or PCA. |

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

The Variance Inflation Factor (VIF) measures the extent of multicollinearity in a regression model, specifically indicating how much the variance of a regression coefficient is inflated due to collinearity with other predictors. A VIF value that is infinite occurs when there is perfect multicollinearity, meaning one predictor variable can be perfectly predicted by a linear combination of the other predictor variables.

**Reasons for Infinite VIF**

1. **Perfect Multicollinearity**:
    - If a predictor variable is an exact linear combination of other predictor variables, the regression model cannot uniquely estimate the coefficients, leading to an infinite VIF for that predictor.

2. **Dummy Variable Trap**:
    - This occurs when dummy variables (indicator variables) are used incorrectly. For example, if there are k categories, k dummy variables are used instead of k-1, causing perfect multicollinearity.
3. **Redundant Features**:
    - Including features that are perfectly correlated (e.g., adding both a feature and its exact duplicate or a sum of other features) can cause infinite VIF values.

**Handling Infinite VIF**

To address infinite VIF values, you can:

1. **Remove Perfectly Collinear Variables**:
    - Identify and remove one of the perfectly collinear variables from the model.
2. **Regularization Techniques**:
    - Use techniques like Ridge Regression, which can handle multicollinearity by adding a penalty to the regression coefficients.
3. **Principal Component Analysis (PCA)**:
    - Transform the predictors into a set of uncorrelated components.
4. **Check Dummy Variables**:
    - Ensure proper use of dummy variables by including k−1dummy variables for k categories to avoid the dummy variable trap.

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

A Q-Q (Quantile-Quantile) plot is a graphical tool used to assess if a dataset follows a particular theoretical distribution, most commonly the normal distribution. It plots the quantiles of the data against the quantiles of the specified theoretical distribution. If the data points lie approximately along a straight line, it suggests that the data distribution matches the theoretical distribution.

Interpretation of a Q-Q Plot

Straight Line: Data follows the theoretical distribution.
S-Shaped Curve: Indicates deviations from normality, such as heavy tails.
Upward/Downward Curve: Indicates skewness in the data.

Use and Importance of a Q-Q Plot in Linear Regression

In the context of linear regression, a Q-Q plot is primarily used to check the normality assumption of the residuals. Normality of residuals is important because:

1. Validates Statistical Tests: Many inferential statistics, such as t-tests for coefficients and F-tests for the overall model, assume that the residuals are normally distributed. Non-normal residuals can lead to invalid test results.
2. Improves Model Interpretability: When residuals are normally distributed, it implies that the model errors are randomly distributed, enhancing the reliability of the model predictions.
3. Detects Model Mis-specifications: Deviations from normality can indicate issues such as omitted variables, incorrect functional form, or the presence of outliers.

Steps to Create and Interpret a Q-Q Plot in Linear Regression

1. Fit the Linear Regression Model: Obtain the residuals from the fitted model.
2. Generate Theoretical Quantiles: Based on the assumed distribution (typically normal).
3. Plot the Q-Q Plot: Plot the residual quantiles against the theoretical quantiles.
4.Analyze the Plot:
   - Linearity: If the points lie on the straight line, residuals are normally distributed.
   - Deviations: Systematic deviations from the line suggest non-normality.