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?

Ans: Based on analysis of the categorical variables, the following equation is obtained for the best fit straight line.

cnt = 0.2494 + 0.0765*Season_2 + 0.1511*Season_4 – 0.0432*month_12 – 0.0483*month_1 – 0.0401*month_7 – 0.0450*month_11 + 0.0836*month_9 -0.0440*weekday_1– 0.2469*weather_3 – 0.0543*weather_2 + 0.2265*yr + 0.5595*temp – 0.1636*humidity – 0.1933*windspeed

- Windspeed - a unit increase in variable, will *increase* bike hiring by 0.1933.
- Humidity - a unit increase in variable, will *decrease* bike hiring by 0.1636.
- Temperature - a unit increase in variable, will *increase* bike hiring by 0.5595.
- YR – a unit increase in variable, will *increase* bike hiring by 0.2265.
- Weather_2 - a unit increase in variable, will *decrease* bike hiring by 0.0543.
- Weather_3 - a unit increase in variable, will *decrease* bike hiring by 0.2469.
- Weekday_1 - a unit increase in variable, will *decrease* bike hiring by 0.0440.
- Month_9 - a unit increase in variable, will *increase* bike hiring by 0.0836.
- Month_11 - a unit increase in variable, will *decrease* bike hiring by 0.0450.
- Month_7 - a unit increase in variable, will *decrease* bike hiring by 0.0401.
- Month_1 - a unit increase in variable, will *decrease* bike hiring by 0.0483.
- Month_12 - a unit increase in variable, will *decrease* bike hiring by 0.0432.
- Season_4 - a unit increase in variable, will *increase* bike hiring by 0.1511.
- Season_2 - a unit increase in variable, will *increase* bike hiring by 0.0765.

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

Ans: While creating dummy variables, drop_first = True would reduce one column which would lead to redundant data. For example: for a column representing day of week, one would expect seven columns to be created while creating dummy. But if this flag is set to True, the number of dummy columns created would be six. One can easily deduce the seventh column data based on the other six columns. This features therefore reduces the correlation among dummy variables.

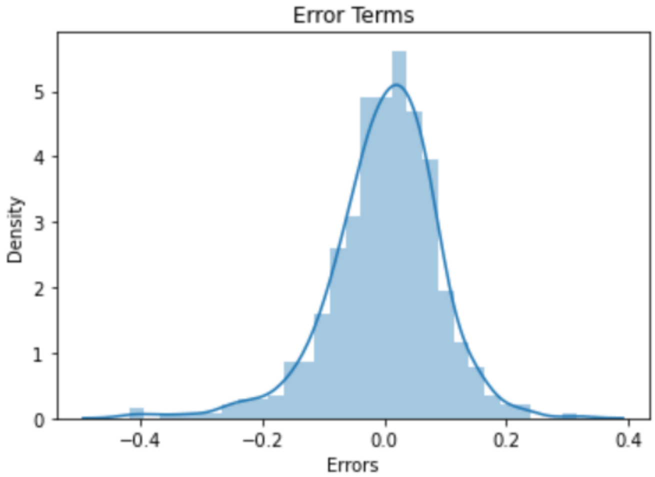
3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

Ans: with respect to the pair-plot, 'temp' has the highest correlation with the target 'cnt'.

4. How did you validate the assumptions of Linear Regression after building the model on the training set?

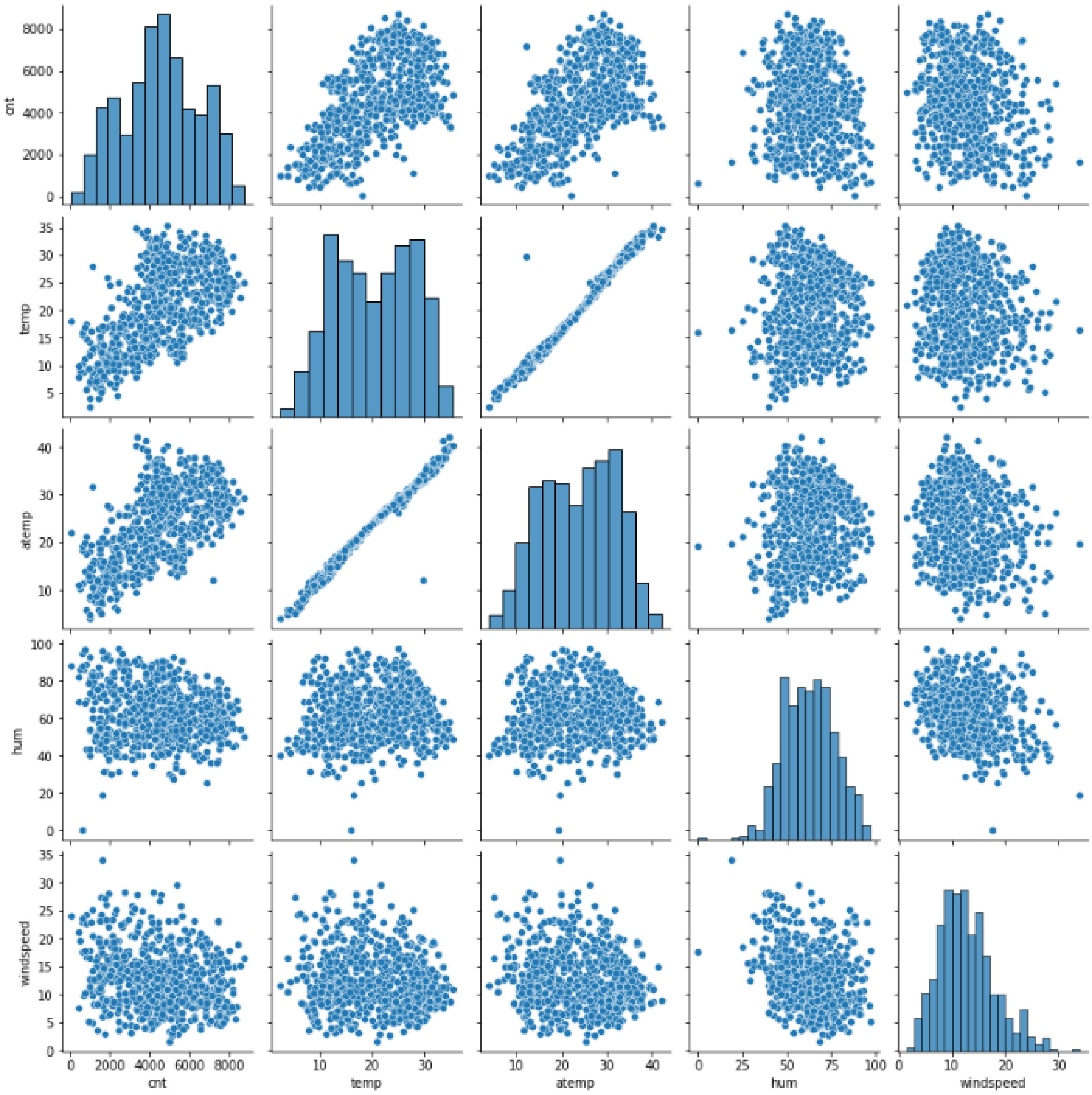
Ans: The following methods are employed to validate linear regression:

- We plot the residuals using histogram, which should be normally distributed. Also, the maximum of the error terms should be around zero.



-

- We employ pair-plot to see if there is a linear relation between temp and cnt.



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

Ans: Based on co-efficients, we can infer that the following variables are significantly contributing:

- Temperature - a unit increase in variable, will *increase* bike hiring by 0.5595.
- YR – a unit increase in variable, will *increase* bike hiring by 0.2265.
- Windspeed - a unit increase in variable, will *increase* bike hiring by 0.1933.

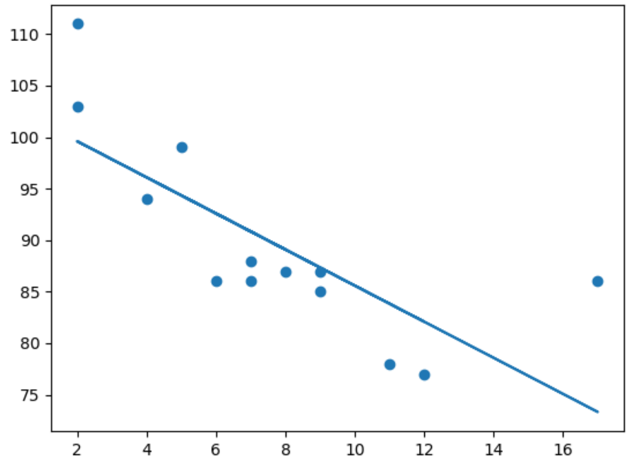
General Subjective Questions

1. Explain the linear regression algorithm in detail.

Ans: Linear Regression is a Supervised Learning method, where the predicted output will be continuous in nature. It is a fundamental statistical and machine learning technique used for modelling the relationship between a dependent variable (also known as the target or response variable) and one or more independent variables (predictors or features).

The goal is to establish a linear equation that best represents the association between these variables, allowing us to make predictions and draw insights from the data.

Idea is to find the "best-fit" line that minimizes the difference between the predicted values and the actual observed values.



This best-fit line is defined by a linear equation of the form:

$$Y = b_0 + b_1X_1 + b_2X_2 + ... + b_nX_n$$

In this equation:

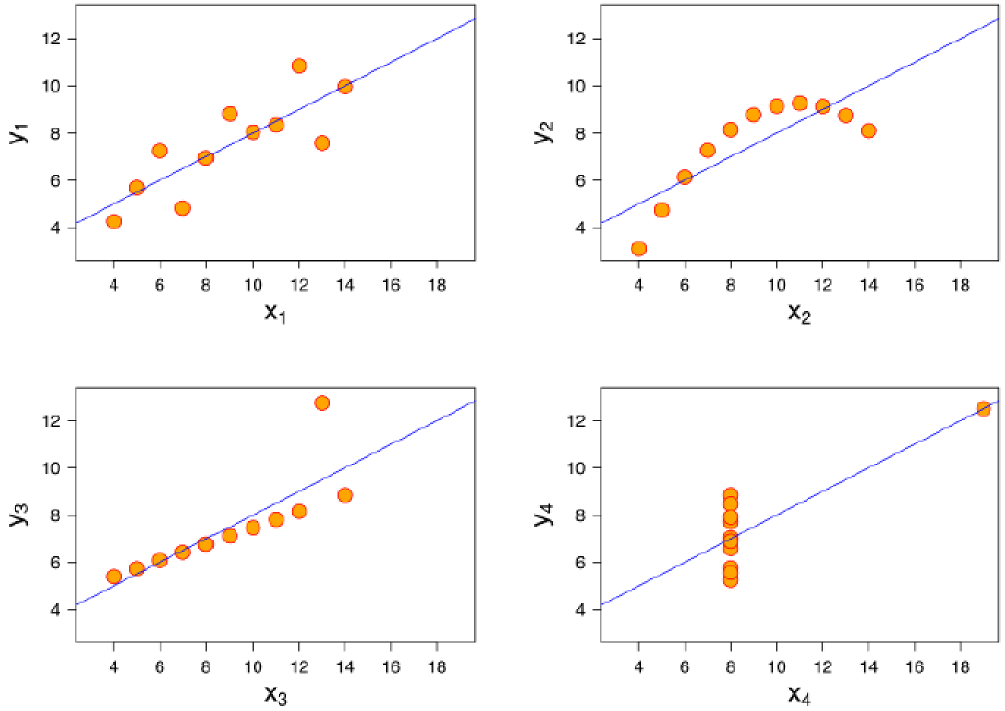
1. Y represents the dependent variable we want to predict.
2. $X_1, X_2, ..., X_n$ are the independent variables or features.
3. b_0 is the intercept (the value of Y when all X values are zero).
4. $b_1, b_2, ..., b_n$ are the coefficients that determine the relationship between each independent variable and the dependent variable.

Linear regression assumes that there is a linear relationship between the predictors and the target variable.

2. Explain the Anscombe's quartet in detail.

Ans: Anscombe's quartet comprises four data sets that have nearly identical simple descriptive statistics, yet have very different distributions and appear very different when graphed.

The essential thing to note about these datasets is that they share the same descriptive statistics. But things change completely, and I must emphasize completely**,** when they are graphed. Each graph tells a different story irrespective of their similar summary statistics.



- Dataset I appear to have clean and well-fitting linear models.
- Dataset II is not distributed normally.
- Dataset III shows that the distribution is linear, but the calculated regression is thrown off by an outlier.
- Dataset IV shows that one outlier is enough to produce a high correlation coefficient.

This quartet emphasizes the importance of visualization in Data Analysis.

3. What is Pearson's R?

Ans: Pearson's correlation coefficient is the test statistics that measures the statistical relationship, or association, between two continuous variables. It is known as the best method of measuring the association between variables of interest because it is based on the method of covariance. It gives information about the magnitude of the association, or correlation, as well as the direction of the relationship.

It is represented by the letter 'R' and is commonly used in defining the linear regression.

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

Ans: 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 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.

Normalized Scaling:

This method of scaling requires below two-step:

1. First, we are supposed to find the minimum and the maximum value of the column.
2. Then we will subtract the minimum value from the entry and divide the result by the difference between the maximum and the minimum value.

$$X_{\text{scaled}} = \frac{X_i - X_{\text{min}}}{X_{\text{max}} - X_{\text{min}}}$$

As we are using the maximum and the minimum value this method is also prone to outliers but the range in which the data will range after performing the above two steps is between 0 to 1. To fix the issue of outliners, we use normalized scaling. In this we subtract each entry by the mean value of the whole data and then divide the results by the difference between the minimum and the maximum value.

$$X_{\text{scaled}} = \frac{X_i - X_{\text{mean}}}{X_{\text{max}} - X_{\text{min}}}$$

Standardized Scaling:

This method of scaling is basically based on the central tendencies and variance of the data.

1. First, we should calculate the mean and standard deviation of the data we would like to normalize.
2. Then we are supposed to subtract the mean value from each entry and then divide the result by the standard deviation.

This helps us achieve a normal distribution (if it is already normal but skewed) of the data with a mean equal to zero and a standard deviation equal to 1.

$$X_{\text{scaled}} = \frac{X_i - X_{\text{mean}}}{\sigma}$$

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

Ans: VIF is an index that provides a measure of how much the variance of an estimated regression coefficient increases due to collinearity. In order to determine VIF, we fit a regression model between the independent variables.

For example, we would fit the following models to estimate the coefficient of determination R1 and use this value to estimate the VIF:

X_1 = C+ α_2 X_2+α_3 X_3+⋯

VIF_1 = 1/(1-R_1^2)

Next, we fit the model between X2 and the other independent variables to estimate the coefficient of determination R2:

X_2 = C+ α_1 X_1+α_3 X_3+⋯

VIF_2 = 1/(1-R_2^2)

If all the independent variables are orthogonal to each other, then VIF = 1.0. 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.

Ans: A Q-Q plot is a scatterplot created by plotting two sets of quantiles against one another. If both sets of quantiles came from the same distribution, we should see the points forming a line that's roughly straight.

This helps in a scenario of linear regression when we have training and test data set received separately and then we can confirm using Q-Q plot that both the data sets are from populations with same distributions.

Few advantages:

a) It can be used with sample sizes also.

b) Many distributional aspects like shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot.

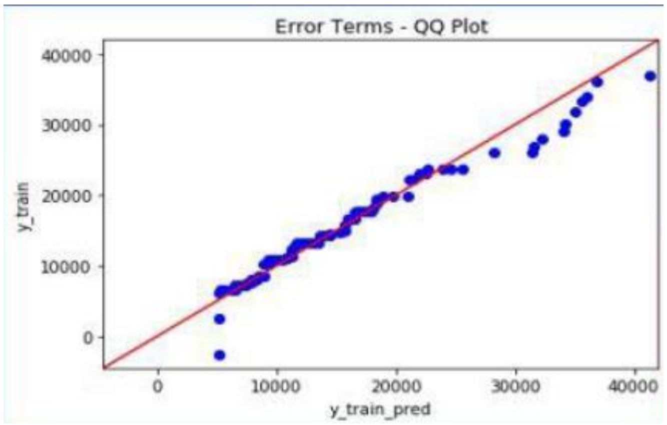
If two data sets —

- Come from populations with a common distribution.
- Have common location and scale.
- Have similar distributional shapes.
- Have similar tail behavior.

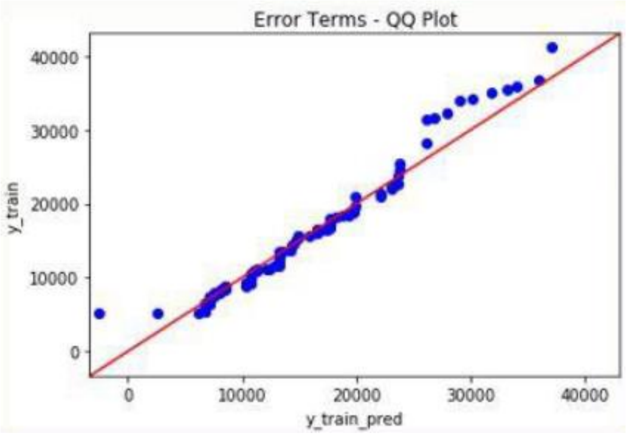
Below are the possible regressions for two data sets:

a) **Similar distribution**: If all point of quantiles lies on or close to straight line at an angle of 45 degree from x -axis

b) **Y-values < X-values:** If y-quantiles are lower than the x-quantiles.



c) **X-values < Y-values:** If x-quantiles are lower than the y-quantiles.



d) **Different distribution:** If all point of quantiles lies away from the straight line at an angle of 45 degree from x –axis.