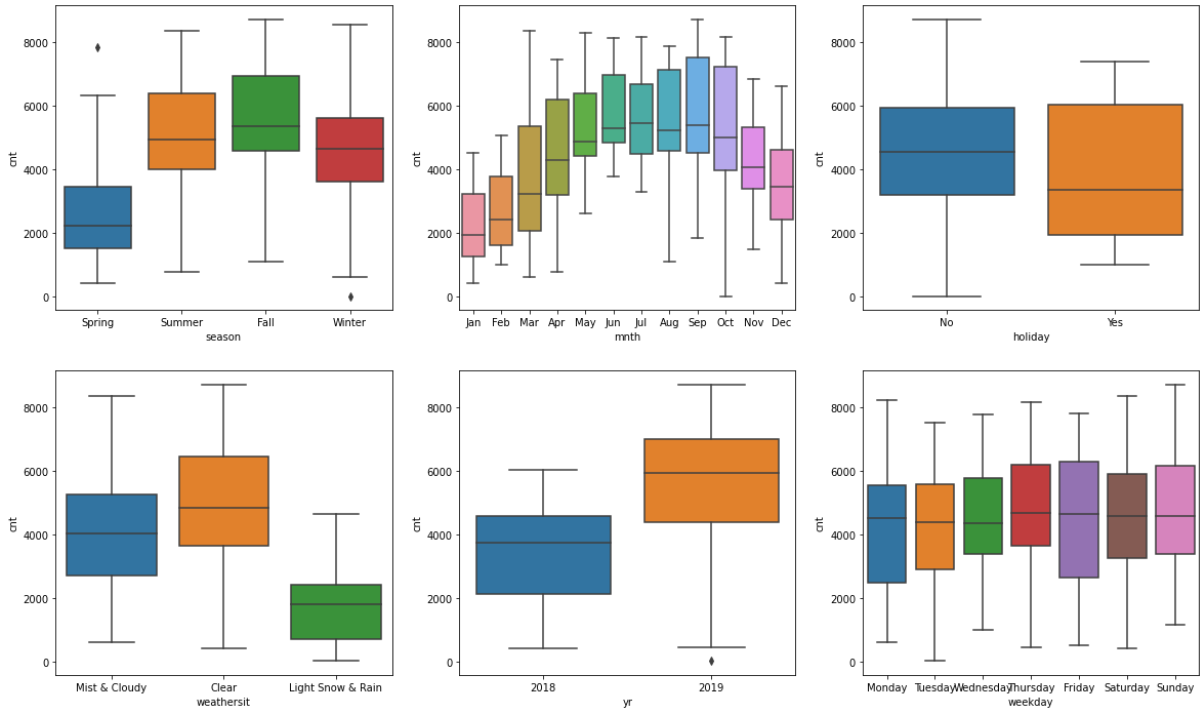
Assignment based Subjective Questions

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

Answer: From my analysis, Categorical variables are: **season**, **mnth**, **holiday**, **weekday**, **yr**, **weathersit**. I used boxplot for visualization of these variables to check how they affect **cnt** variable (our target variable):



Season: Spring season has least value of count as its median is very low as compared to others and Fall has maximum number of count. Rest of them - Summer and Winter had an intermediate number of counts.

Weathersit: There are no users when there is Heavy Rain/Snow. That indicates that this weather is extremely unfavourable. Highest count for rentals is for **Clear weather,** followed by **Mist & Cloudy** and least numbers for **Light Snow & Rain.**

Yr: The number of rentals in 2019 was more than 2018.

Mnth: July has the highest number of rentals while January has least. The observation is the same as weathersit, January usually has heavy snowfall.

Holiday: The median of count decreases during holidays but overall IQR range is more during holidays. Mostly bike booking were happening when it is not a holiday

Weekday: weekday variable shows very close trend having their independent medians between 4000 to 5000 bookings. This variable can have some influence towards the predictor.

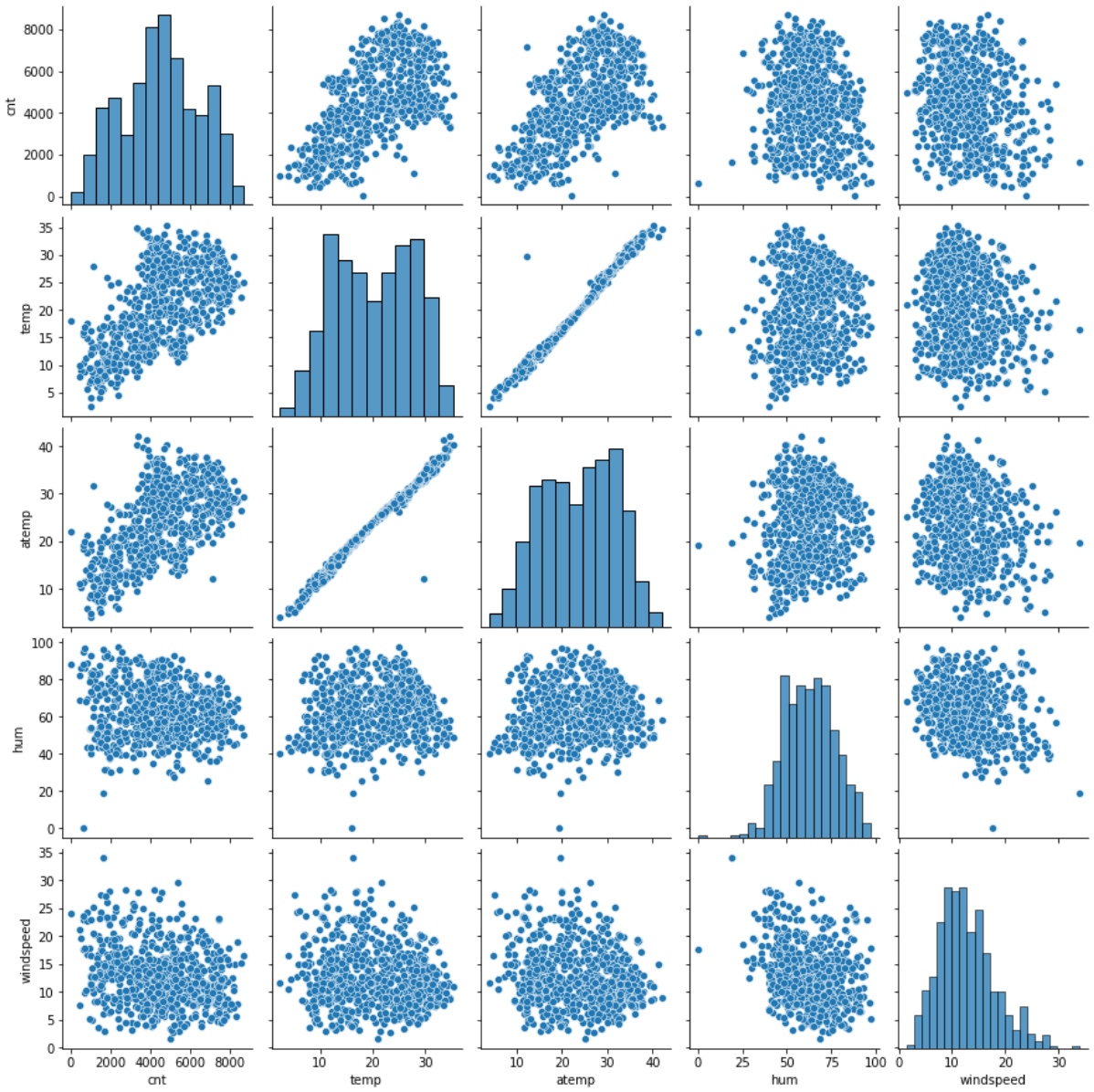
Question 2: Why is it important to use drop_first=True during dummy variable creation?

Answer: If we don't drop the first column then our dummy variables will be correlated or redundant. This may affect the models strongly. Another reason: If we have all dummy variables it leads to Multicollinearity between the dummy variables. To keep this under control, we should drop_first column.

e.g. Let's say we have 3 types of values in a Categorical column and we want to create a dummy variable for that column. If one variable is not furnished and semi_furnished, then It is obviously unfurnished. So we do not need a 3rd variable to identify the unfurnished.

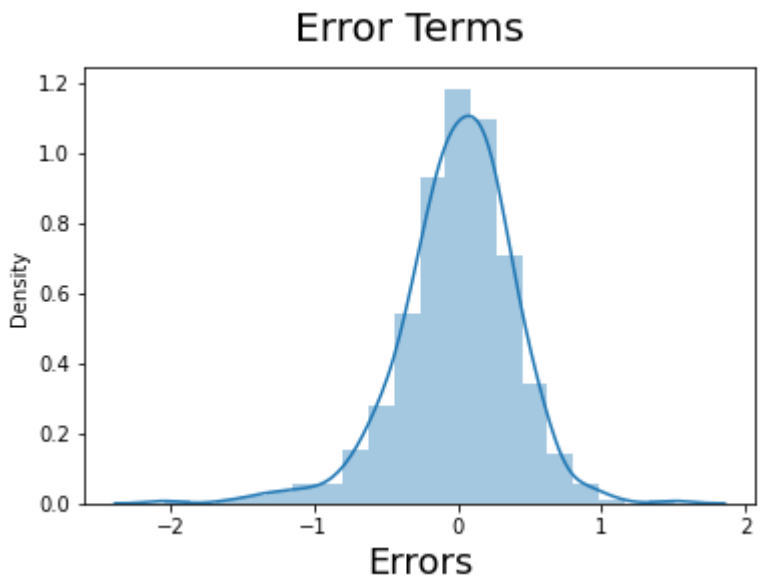
Question 3: Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

Answer: "temp" and "atemp" are the two numerical variables which are highly correlated with the target variable (cnt).



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

Answer: Residuals distribution should follow normal distribution and centred around 0 (i.e. mean = 0). We validate this assumption about residuals by plotting a distplot of residuals and see if residuals are following Normal Distribution or not.The below image is showing same trend:



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

Answer: The top features are:

1. ***yr_2019***. coefficient - 1.045673
2. ***temp***. coefficient - 0.439012
3. ***weathersit_Light Snow & Rain***. coefficient - -1.316081

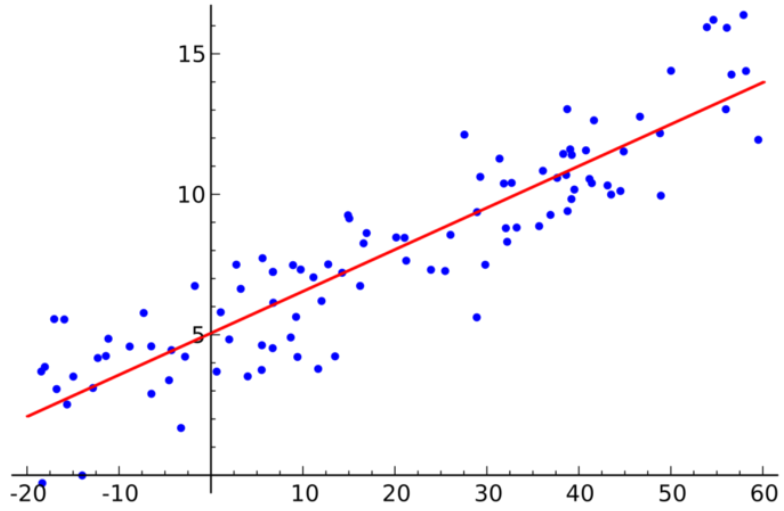
General Subjective Questions

Question 1: Explain the linear regression algorithm in detail.

Answer: Before explaining what linear regression is, let me explain regression. Regression is a method of modelling a target value based on independent predictors. This method is mostly used for forecasting and finding out the cause and effect relationship between variables. Regression techniques mostly differ based on the number of independent variables and the type of relationship between the independent and dependent variables.

Linear Regression is the basic form of regression analysis. It assumes that there is a linear relationship between the dependent variable and the predictor(s). In

regression, we try to calculate the best fit line which describes the relationship between the predictors and the predictive/dependent variable.



The red line in the above graph is referred to as the best fit straight line. Based on the given data points, we try to plot a line that models the points the best. The line can be modelled based on the linear equation: y = b_0 + b_1 * x
Through the best fit line, we can describe the impact of change in independent variables on the dependent variable.
The cost function helps us to figure out the best possible values for b_0 and b_1 which would provide the best fit line for the data points.

We would like to minimize the error between the predicted value and the actual value.

$$minimize \frac{1}{n} \sum_{i=1}^{n} (pred_i - y_i)^2$$

$$J = \frac{1}{n} \sum_{i=1}^{n} (pred_i - y_i)^2$$

Performance of Regression

The performance of the regression model can be evaluated by using various metrics like RMSE, R-squared, Adjusted R-squared etc.

Root Mean Square Error (RMSE)

RMSE calculates the square root average of the sum of the squared difference between the actual and the predicted values.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{N}}$$

where y_i is the actual value, $\hat{y}_i$ is the predicted value and n is the number of observations in the dataset.

R-squared values

R-square value depicts the percentage of the variation in the dependent variable explained by the independent variable in the model.

R-square = 1 - RSS/TSS

RSS/Residual sum of squares: It is the measure of the difference between the expected and the actual output. A small RSS indicates a tight fit of the model to the data. It is also defined as follows:

$$RSS = \sum_{i=1}^{n}(y_i - (\beta_0 + \beta_1 X_i))^2$$

TSS/Total sum of squares: It is the sum of errors of the data points from the mean of the response variable.

$$TSS = \sum_{i=1}^{n} (y_i - \bar{y})^2$$, where $\bar{y}$ = average of the dependent variable.

R^2 value ranges from 0 to 1. Higher the R-square value, the better the model. The value of R^2 increases if we add more variables to the model irrespective of the variable contributing to the model or not. This is the disadvantage of using R^2.

Adjusted R-squared values

The disadvantage of R^2 is fixed by the Adjusted R^2 value. Adjusted R^2 value will improve only if the added variable is making a significant contribution to the model. Adjusted R^2 value adds penalty in the model.

$$R^2_{adjusted} = 1 - \frac{(1 - R^2)(N - 1)}{N - p - 1}$$

where

R^2 = sample R-square

p = Number of predictors

N = Total sample size.

where R^2 is R-square value, N = total number of observations, and p = total number of variables used in the model. If we increase the number of variables, the denominator becomes smaller and the overall ratio will be high. Subtracting from 1 will reduce the overall Adjusted R^2. So to increase the Adjusted R^2, the contribution of additive features to the model should be significantly high.

Types of Linear Regression :

Simple Linear Regression and **Multiple Linear Regression**

Simple Linear Regression (SLR) : SLR is used when the dependent variable is predicted using only one independent variable

Equation for the Simple Linear Regression is:

$$Y_i = \beta_i X_i + \beta_0 ,$$

If there is only 1 predictor available then it is known as Simple Linear Regression.

The equation for SLR will be $Y_i = \beta_1 X_{1i} + \beta_0$, β_1 = coefficient for X_1 variable and β_0 is the intercept.

While executing the prediction, there is an error term which is associated with the equation.

$Y_i = \beta_1 X_{1i} + \beta_0 + \varepsilon_i$, ε is the error term associated with each predicted value.

The goal of the SLR model is to find the estimated values of β_1 & β_0 by keeping the error term (ε) minimum.

Multiple Linear Regression (MLR) : MLR is used when the dependent variable is predicted using multiple independent variables.

The equation for MLR will be:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \dots ,$$

β_1 = coefficient for X_1 variable

β_2 = coefficient for X_2 variable

β_3 = coefficient for X_3 variable and so on…

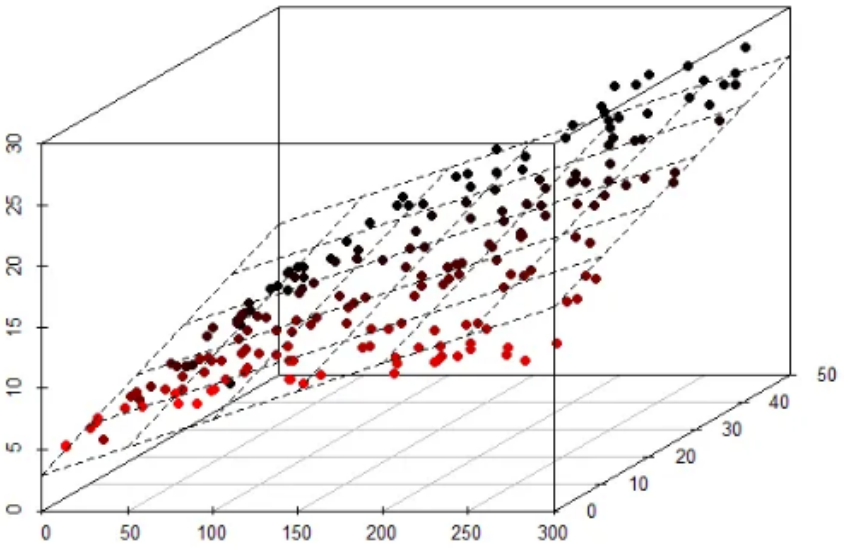
β_0 is the intercept (constant term). While doing the prediction, there is an error term which is associated with the equation.

$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \ldots + \varepsilon_i$, ε_i is the error term associated with each predicted value.

The goal of the MLR model is to find the estimated values of $\beta_0, \beta_1, \beta_2, \beta_3 \ldots$ by keeping the error term (ε i) minimum.

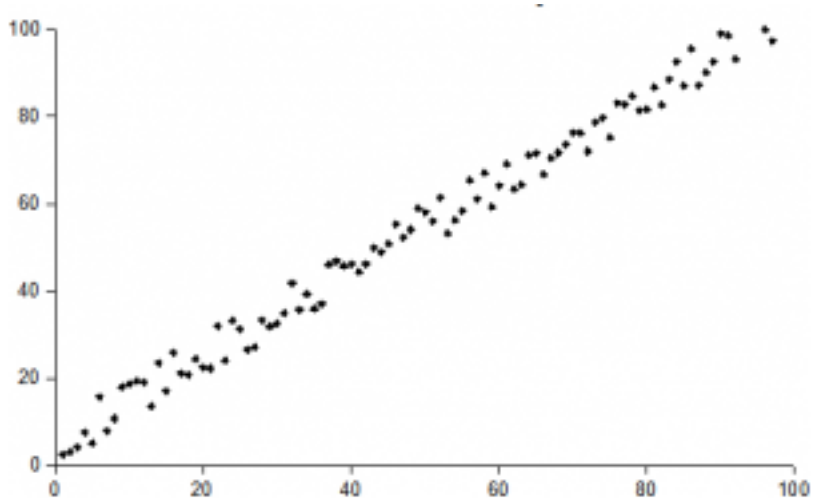
Assumptions for Multiple Linear Regression

1. Linearity: There should be a linear relationship between dependent and independent variables like shown in the below example graph.



2. Multicollinearity: There should not be high correlation between two or more independent variables. Multicollinearity can be checked using correlation matrix, Tolerance and Variance Influencing Factor (VIF).

3. Homoscedasticity: If Variance of errors are constant across independent variables, then it is called Homoscedasticity. The residuals should be homoscedastic. Standardized residuals versus predicted values is used to check homoscedasticity as shown in the below figure. Breusch-Pagan and White tests are the famous tests used to check Homoscedasticity. Q-Q plots are also used to check homoscedasticity.



4. Multivariate Normality: Residuals should be normally distributed.

5. Categorical Data: Any categorical data present should be converted into dummy variables.

Question 2: Explain the Anscombe's quartet in detail.

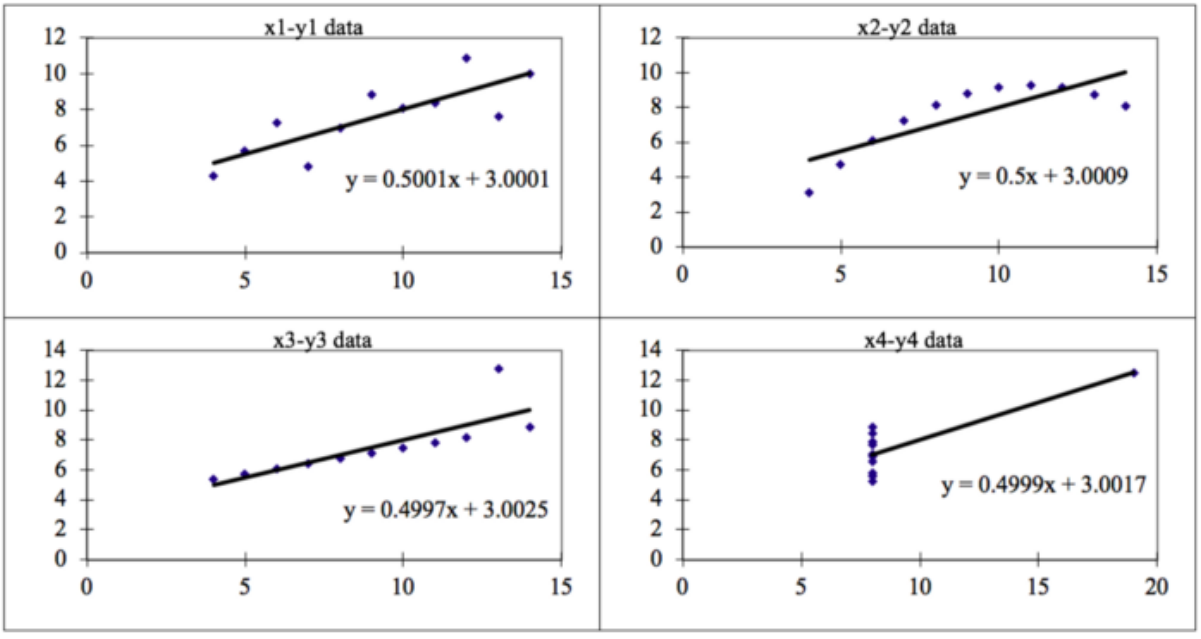
Answer. Anscombe's quartet comprises four datasets that have nearly identical simple descriptive statistics, yet have very different distributions and appear very different when graphed.

Anscombe's Data										
Observation	x1	y1		x2	y2		x3	y3		x4	y4
1	10	8.04		10	9.14		10	7.46		8	6.58
2	8	6.95		8	8.14		8	6.77		8	5.76
3	13	7.58		13	8.74		13	12.74		8	7.71
4	9	8.81		9	8.77		9	7.11		8	8.84
5	11	8.33		11	9.26		11	7.81		8	8.47
6	14	9.96		14	8.1		14	8.84		8	7.04
7	6	7.24		6	6.13		6	6.08		8	5.25
8	4	4.26		4	3.1		4	5.39		19	12.5
9	12	10.84		12	9.13		12	8.15		8	5.56
10	7	4.82		7	7.26		7	6.42		8	7.91
11	5	5.68		5	4.74		5	5.73		8	6.89

The statistical information for all these four datasets are approximately similar and can be computed as follows:

Anscombe's Data										
Observation	x1	y1		x2	y2		x3	y3		x4	y4
1	10	8.04		10	9.14		10	7.46		8	6.58
2	8	6.95		8	8.14		8	6.77		8	5.76
3	13	7.58		13	8.74		13	12.74		8	7.71
4	9	8.81		9	8.77		9	7.11		8	8.84
5	11	8.33		11	9.26		11	7.81		8	8.47
6	14	9.96		14	8.1		14	8.84		8	7.04
7	6	7.24		6	6.13		6	6.08		8	5.25
8	4	4.26		4	3.1		4	5.39		19	12.5
9	12	10.84		12	9.13		12	8.15		8	5.56
10	7	4.82		7	7.26		7	6.42		8	7.91
11	5	5.68		5	4.74		5	5.73		8	6.89
				Summary Statistics							
N	11	11		11	11		11	11		11	11
mean	9.00	7.50		9.00	7.500909		9.00	7.50		9.00	7.50
SD	3.16	1.94		3.16	1.94		3.16	1.94		3.16	1.94
r	0.82			0.82			0.82			0.82	

When these models are plotted on a scatter plot, all datasets generates a different kind of plot that is not interpretable by any regression algorithm which is fooled by these peculiarities and can be seen as follows:



The four datasets can be described as:

1. Dataset 1: This fits the linear regression model pretty well.
2. Dataset 2: This could not fit linear regression model on the data quite well as the data is non-linear.

3. Dataset 3: shows the outliers involved in the dataset which cannot be handled by linear regression model and it can be overfitted.

4. Dataset 4: shows the outliers involved in the dataset which cannot be handled by linear regression model

The above data sets clearly show the importance of data visualisation and how any regression algorithm can be misguided by the same. Hence, all the important features in the dataset must be visualised before implementing any machine learning algorithm on them which will help to make a good fit model.

Question 3: What is Pearson's R?

Answer: Correlation is a bi-variate analysis that measures the strength of association between two variables and the direction of the relationship. In terms of the strength of relationship, the value of the correlation coefficient varies between +1 and -1. Pearson's R Correlation is a type of Correlation.

Pearson R Correlation

Pearson correlation coefficient is a measure of the strength of a linear association between two variables — denoted by r.

r = 1 means that the data is perfectly linear with a positive slope

r = -1 means that data is perfectly linear with negative slope.

r = 0 means that there is no linear association.

Assumptions

1. For the Pearson r correlation, both variables should be **normally distributed**. i.e. the normal distribution describes how the values of a variable are distributed. This is sometimes called the 'Bell Curve' or the 'Gaussian Curve'.
2. There should be **no significant outliers**. Pearson's correlation coefficient, r, is very sensitive to outliers.

3. Each variable should be **continuous** i.e. interval or ratios for example weight, time, height, age etc.
4. The two variables have a **linear relationship**. Scatter plots will help us tell whether the variables have a linear relationship.
5. The observations are **paired observations**. For example if we are calculating the correlation between age and weight. If there are 12 observations of weight, we should have 12 observations of age. i.e. no blanks.
6. **Homoscedascity**: A scatter-plot makes it easy to check for this. If the points lie equally on both sides of the line of best fit, then the data is homoscedastic.

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

Answer: **Scaling** : 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.

Why it is performed:

Most of the time, the collected data set contains features highly varying in magnitudes, units and range. If scaling is not done then the algorithm only takes magnitude in account and not units, hence incorrect modelling. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude. Scaling can be done by two methods:

1. Normalization/ Min-Max Scaling
2. Standardization

Difference between scaling methods:

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.

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

- sklearn.preprocessing.scale helps to implement standardization in python.
- **One disadvantage of normalization over standardization is that it loses some information in the data, especially about outliers.**

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

Answer: If there is perfect correlation, then VIF = infinity. This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1, which leads to 1/(1-R2) infinity. To solve this problem we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables (which show an infinite VIF as well).

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

Answer: Quantile-Quantile (Q-Q) plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal, exponential or Uniform distribution. Also, it helps to determine if two data sets come from populations with a common distribution.

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 the same distributions.
Python Libraries for Q-Q plot are: statsmodels.api
provide qqplot and qqplot_2samples to plot Q-Q graph for single and two different data sets respectively.

Advantages of Q-Q plot:

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.

It is used to check following scenarios:

If two data sets —

i. come from populations with a common distribution
ii. have common location and scale
iii. have similar distributional shapes
iv. have similar tail behavior

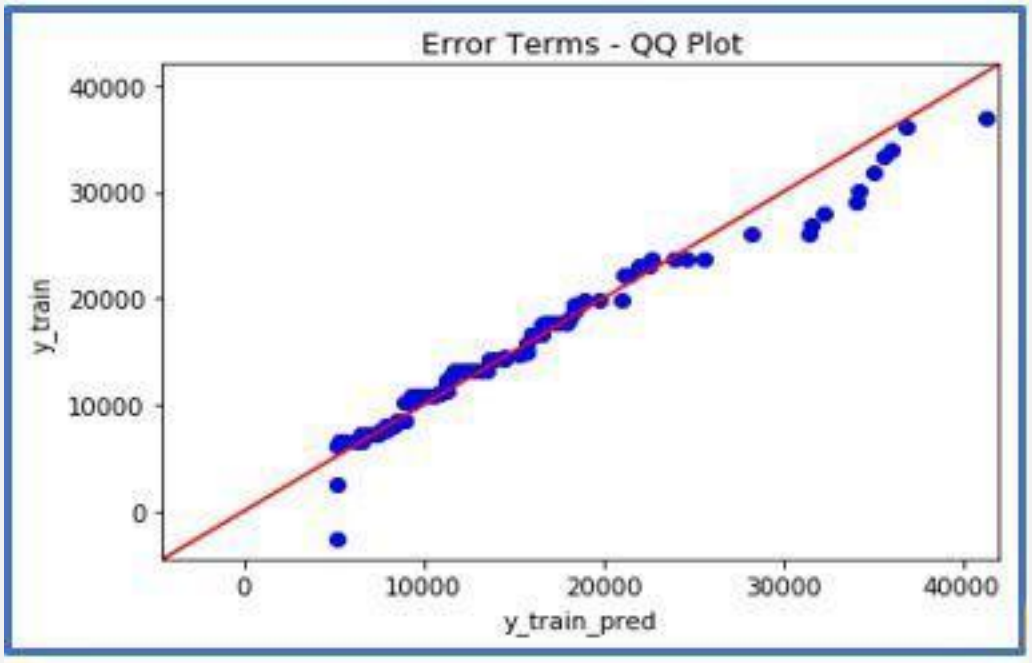
Interpretation:

A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set.

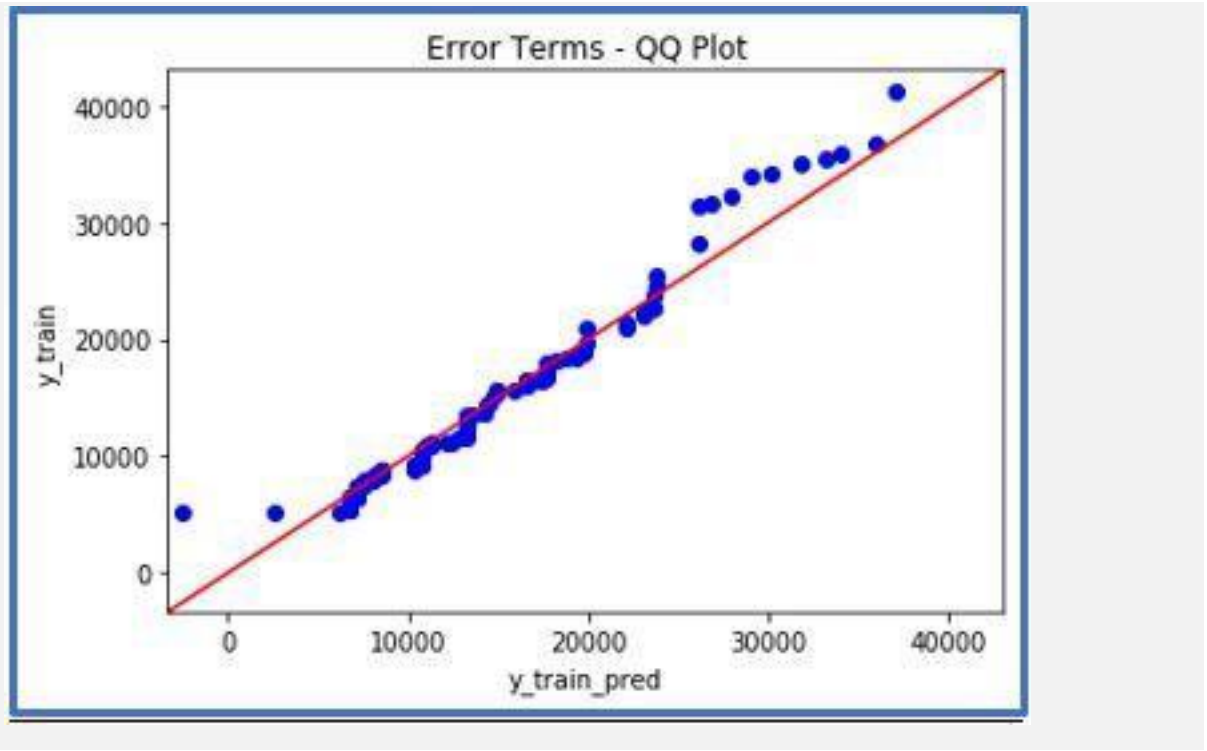
Below are the possible interpretations for two data sets.

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