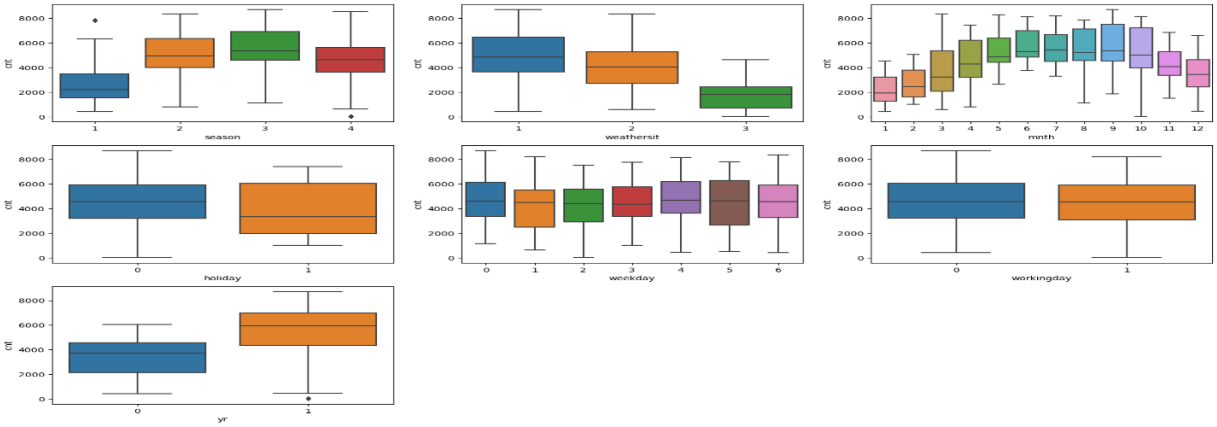
Assignment-based Subjective Questions

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

There were 6 categorical variables in the dataset. We used Box plot to study their effect on the dependent variable ('cnt').



The inference that We could derive were:

season: Almost 32% of the bike booking were happening in season3 with a median of over 5000 booking (for the period of 2 years). This was followed by season2 & season4 with 27% & 25% of total booking. This indicates, season can be a good predictor for the dependent variable.

mnth: Almost 10% of the bike booking were happening in the months 5,6,7,8 & 9 with a median of over 4000 booking per month. This indicates, mnth has some trend for bookings and can be a good predictor for the dependent variable.

weathersit: Almost 67% of the bike booking were happening during 'weathersit1 with a median of close to 5000 booking (for the period of 2 years). This was followed by weathersit2 with 30% of total booking. This indicates, weathersit does show some trend towards the bike bookings can be a good predictor for the dependent variable.

holiday: Almost 97.6% of the bike booking were happening when it is not a holiday which means this data is clearly biased. This indicates, holiday CANNOT be a good predictor for the dependent variable.

weekday: weekday variable shows very close trend (between 13.5%-14.8% of total booking on all days of the week) having their independent medians between 4000 to 5000 bookings. This variable can have some or no influence towards the predictor. I will let the model decide if this needs to be added or not.

workingday: Almost 69% of the bike booking were happening in 'workingday' with a median of close to 5000 booking (for the period of 2 years). This indicates, workingday can be a good predictor for the dependent variable

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

Drop_first = True, it drops the first column created while creating dummy variables because it creates one extra column.

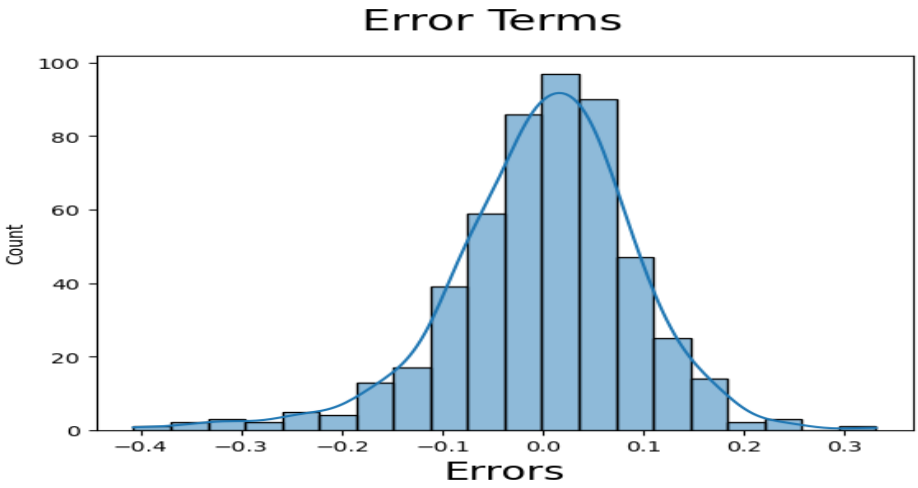
3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

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?

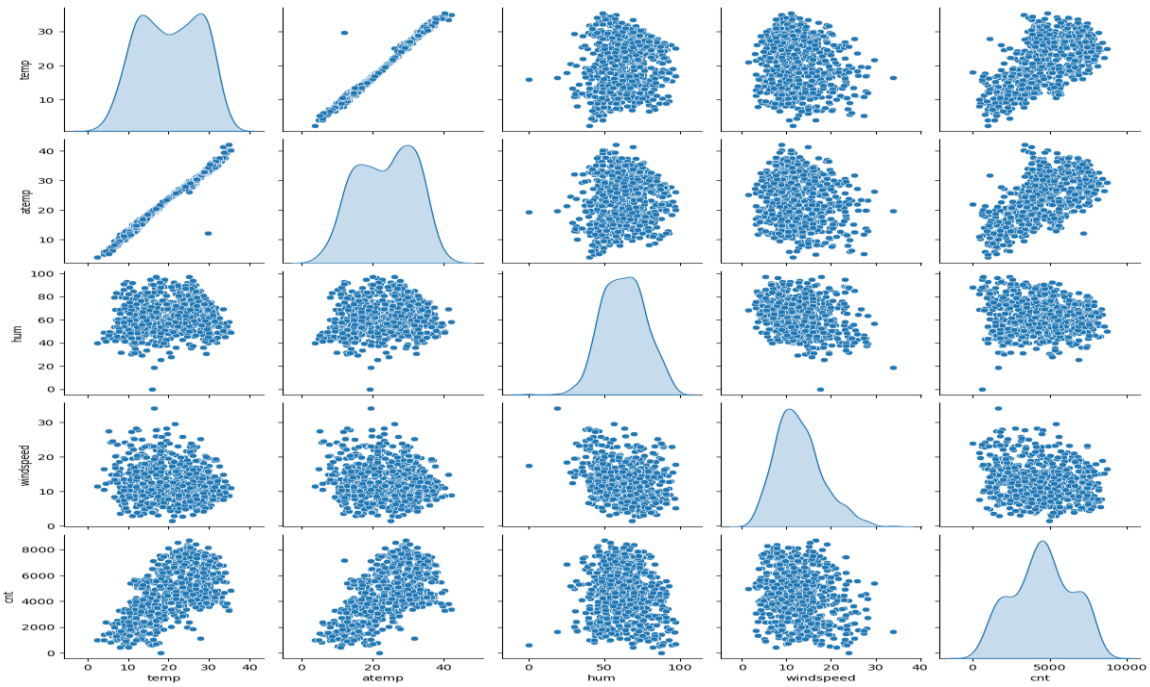
ASSUMPTIONS

1- Error terms are normally distributed with mean zero (not X, Y) Residual Analysis Of Training Data



From the above histogram, we could see that the Residuals are normally distributed. Hence our assumption for Linear Regression is valid.

2- There is a linear relationship between X and Y



Using the pair plot, we could see there is a linear relation between temp and atemp variable with the predictor ‘cnt’.

3- There is No Multicollinearity between the predictor variables

	Features	VIF
2	temp	3.68
3	windspeed	3.05
0	yr	2.00
4	season_2	1.56
7	weathersit_2	1.48
5	season_4	1.38
6	mnth_9	1.20
8	weathersit_3	1.08
1	holiday	1.03

From the VIF calculation we could find that there is no multicollinearity existing between the predictor variables, as all the values are within permissible range of below 5

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

As per our final Model, the top 3 predictor variables that influences the bike booking are:

- Temperature (temp) - A coefficient value of '0.5480' indicated that a unit increase in temp variable increases the bike hire numbers by 0.5480 units.
- Weather Situation 3 (weathersit_3) - A coefficient value of '-0.2838' indicated that, w.r.t Weathersit1, a unit increase in Weathersit3 variable decreases the bike hire numbers by 0.2838 units.
- Year (yr) - A coefficient value of '0.2328' indicated that a unit increase in yr variable increases the bike hire numbers by 0.2328 units.

So, it's suggested to consider these variables utmost importance while planning, to achive maximum Booking The next best features that can also be considered are

- season_4: - A coefficient value of '0.1306' indicated that w.r.t season_1, a unit increase in season_4 variable increases the bike hire numbers by 0.1306 units.
- windspeed: - A coefficient value of '-0.1533' indicated that, a unit increase in windspeed variable decreases the bike hire numbers by 0.1533 units.

NOTE:

The details of weathersit_1 & weathersit_3

- weathersit_1: Clear, Few clouds, Partly cloudy, Partly cloudy
- weathersit_3: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds

The details of season1 & season4

- season1: spring
- season4: winter

General Subjective Questions

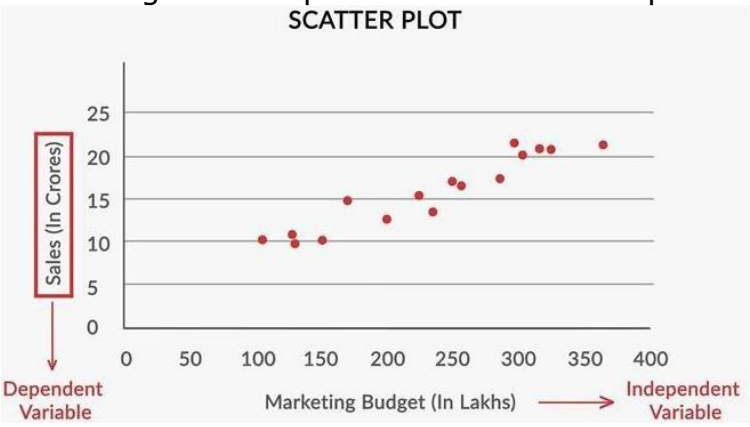
1. Explain the linear regression algorithm in detail.

Regression analysis is graphing a line on a set of data points that most closely fits the overall shape of the data. In other words, Regression shows the changes in a dependent variable on the y-axis to the changes in the explanatory variable on x-axis.
Linear Regression is the relationship between the dependent (target variable) and independent variables (predictors). There are two types of linear regression:

1- Simple linear regression
2- Multiple linear regression

Simple Linear Regression:

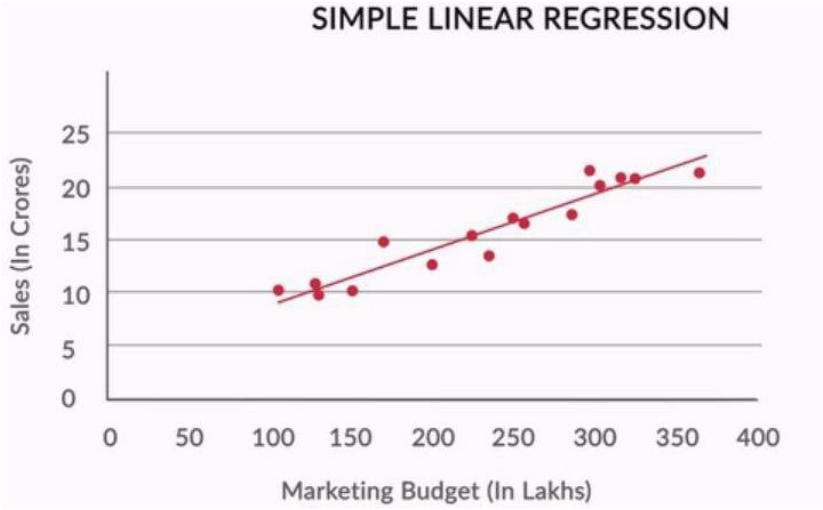
The most elementary type of regression model is the simple linear regression which explains the relationship between a dependent variable and one independent variable using a straight line. The straight line is plotted on the scatter plot of these two points.



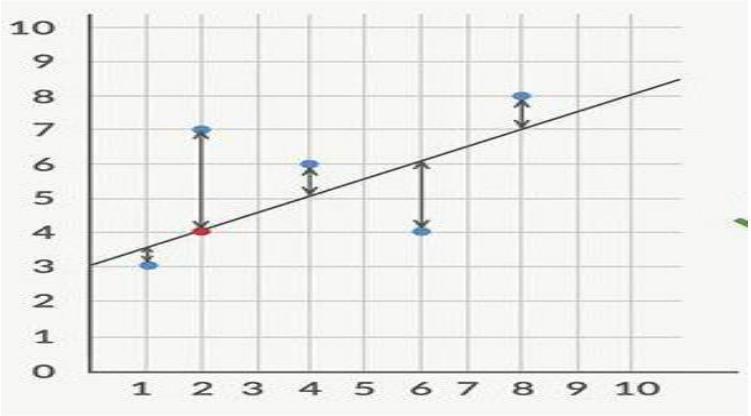
The standard equation of the regression line is given by the following expression:

$$Y = \beta_0 + \beta_1 X$$

Where, β_0 is intercept and β_1 is slope.



The best-fit line is found by minimising the expression of RSS (Residual Sum of Squares) which is equal to the sum of squares of the residual for each data point in the plot. Residuals for any data point is found by subtracting predicted value of dependent variable from actual value of dependent variable:



Residuals:

Intercept Slope

$e_i = y_i - y_{pred}$

Ordinary Least Squares Method:

$e_1^2 + e_2^2 + \_ \_ \_ + e_n^2 = RSS$ (Residual Sum Of Squares)

$RSS = (Y_1-\beta_0-\beta_1 X_1)^2 + (Y_0-\beta_0-\beta_1 X_2)^2 + \_ \_ \_ \_ + (Y_n-\beta_0-\beta_1 X_n)^2$

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

The strength of the linear regression model can be assessed using 2 metrics:

1. R^2 or Coefficient of Determination
2. Residual Standard Error (RSE)

<u>R^2 or Coefficient of Determination</u>

An alternative way of checking the accuracy of our model, which is R2 statistics. R2 is a number which explains what portion of the given data variation is explained by the developed model. It always takes a value between 0 & 1. In general term, it provides a measure of how well actual outcomes are replicated by the model, based on the proportion of total variation of outcomes explained by the model, i.e. expected outcomes. Overall, the higher the R-squared, the better the model fits your data. Mathematically, it is represented as:

$$R^2 = 1 - (RSS / TSS)$$

Where, RSS is Residual sum of square
TSS is Sum of errors of the data from mean

<u>RSS (Residual Sum of Squares):</u> In statistics, it is defined as the total sum of error across the whole sample. 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:

$$RSS = \sum_{i=1}^{n} (y_i - (\alpha + \beta x_i))^2$$

<u>TSS (Total sum of squares):</u> It is the sum of errors of the data points from mean of response variable. Mathematically, TSS is:

$$\text{TSS} = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

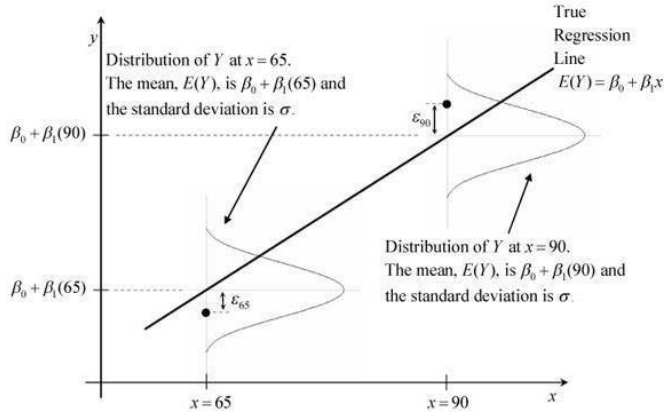
Importance of RSS/TSS: Think about it for a second. If you know nothing about linear regression and still have to draw a line to represent those points, the least you can do is have a line pass through the mean of all the points.

Assumptions of Simple Linear Regression

Taking a more statistical view:

- Linear regression, at each X, finds the best estimate for Y
- At each X, there is a distribution on the values of Y

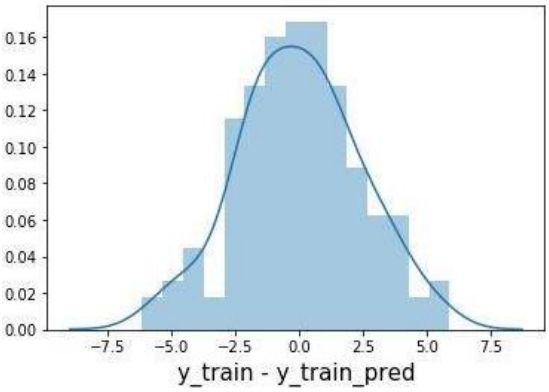
Model predicts a single value, therefore there is a distribution of error terms at each of these values as can be seen from the figure below.



the assumptions of simple linear regression were:

1. Linear relationship between X and Y
2. Error terms are normally distributed (not X, Y)
3. Error terms are independent of each other
4. Error terms have constant variance (homoscedasticity)
5. Multicollinearity

there is NO assumption on the distribution of X and Y, just that the error terms must have a normal distribution. The normal distribution of the residual terms is a very crucial assumption when it comes to making inferences from a linear regression model. Hence, it is very important that we analyse these residual terms before we can move forward. The simplest method to check for the normality is to plot a histogram of the error terms and check whether the error terms are normal.



After we have determined that the coefficient is significant, using p-values, we need some other metrics to determine whether the overall model fit is significant. To do that, you need to look at a parameter called the F-statistic.

So, the parameters to assess a model are:

1. **t statistic:** Used to determine the p-value and hence, helps in determining whether the coefficient is significant or not
2. **F statistic:** Used to assess whether the overall model fit is significant or not. Generally, the higher the value of F statistic, the more significant a model turns out to be
3. **R-squared:** After it has been concluded that the model fit is significant, the R-squared value tells the extent of the fit, i.e. how well the straight line describes the variance in the data. Its value ranges from 0 to 1, with the value 1 being the best fit and the value 0 showcasing the worst.

Multiple Linear Regression:

Multiple linear regression is a statistical technique to understand the relationship between one dependent variable and several independent variables. The objective of multiple regression is to find a linear equation that can best determine the value of dependent variable Y for different values independent variables in X.

adding more variables increases the R-squared and it might be a good idea to use multiple variables to explain a feature variable. Basically:

1. Adding variables helped add information about the variance in Y!
2. In general, we expect explanatory power to increase with increase in variables

Hence, this brings us to multiple linear regression which is just an extension to simple linear regression. The formulation for multiple linear regression is also similar to simple linear linear regression with the small change that instead of having beta for just one variable, you will now have betas for all the variables used. The formula now can be simply given as:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon$$

Apart from the formula, a lot of other ideas in multiple linear regression are also similar to simple linear regression, such as:

1. Model now fits a 'hyperplane' instead of a line
2. Coefficients still obtained by minimizing sum of squared error (Least squares criterion)
3. For inference, the assumptions from from Simple Linear Regression still hold ○ Zero mean, independent, Normally distributed error terms that have constant variance

○ The inference part in multiple linear regression also, largely, remains the same.

most of the ideas in simple and multiple linear regression are the same, there are a few new considerations that you need to make when moving to multiple linear regression, such as

Overfitting: When you add more and more variables, for example, let's say you keep on increasing the degree of the polynomial function fitting the data, your model might end up memorizing all the data points in the training set. This will cause major problems with generalisation, i.e. now when the model runs on the test data, the accuracy will drop tremendously since, it doesn't generalise well. This is a classical symptom of overfitting.

Multicollinearity: Multicollinearity is the effect of having related predictors in the multiple linear regression model. In simple terms, in a model which has been built using several independent variables, some of these variables might be interrelated, i.e. some of these variables might completely explain some other independent variable in the model due to which the presence of that variable in the model is redundant. So in order to know, where the effect on the feature variable is coming from, we need to drop some of these related independent variables. Basically, multicollinearity affects:

1. Interpretation: Does "change in Y, when all others are held constant" apply?
2. Inference: a. Coefficients swing wildly, signs can invert

b. p-values are, therefore, not reliable

But there are a few aspects that multicollinearity does not affect, such as:

a. The predictions and the precision of the predictions

b. Goodness-of-fit statistics such as R-squared

Hence, dealing with multicollinearity is extremely important. There are two ways to detect multicollinearity in a model:

1- Correlations

2- VIF (Variance Inflation Factor)

after any multicollinearity has been detected in the model, you need to deal with it appropriately in order to avoid building an unnecessarily complex model with a lot of redundant variables. The few methods to deal with multicollinearity are:

1. Dropping variables
2. Create new variable using the interactions of the older variables

Feature Scaling**:** Another important aspect to consider is feature scaling. When you have a lot of independent variables in a model, a lot of them might be on very different scales which will lead a

model with very weird coefficients that might be difficult to interpret. So we need to scale features because of two reasons:
1. Ease of interpretation

2. Faster convergence for gradient descent methods You can scale the features using two very popular method:
1- Standardizing 2- MinMax Scaling

Feature Selection: dropping features from our model. But choosing to drop the correct features (that are redundant and not adding any value to the model) is quite essential. There are various methods for optimal feature selection:
1. Try all possible combinations (**2p** models for p features) ○ Time consuming and practically unfeasible
2. Manual Feature Elimination

○ Build model

○ Drop features that are least helpful in prediction (high p-value)

○ Drop features that are redundant (using correlations, VIF)

○ Rebuild model and repeat
3. Automated Approach ○ Recursive Feature Elimination (RFE)

○ Forward/Backward/Stepwise Selection based on AIC (not covered)
It is generally recommended that we follow a balanced approach, i.e., use a combination of automated (coarse tuning) + manual (fine tuning) selection in order to get an optimal model.

2. Explain the Anscombe's quartet in detail.

Anscombe's Quartet is a set of four datasets that have nearly identical simple descriptive statistics but differ significantly when graphed. It was created by the Francis Anscombe in 1973.

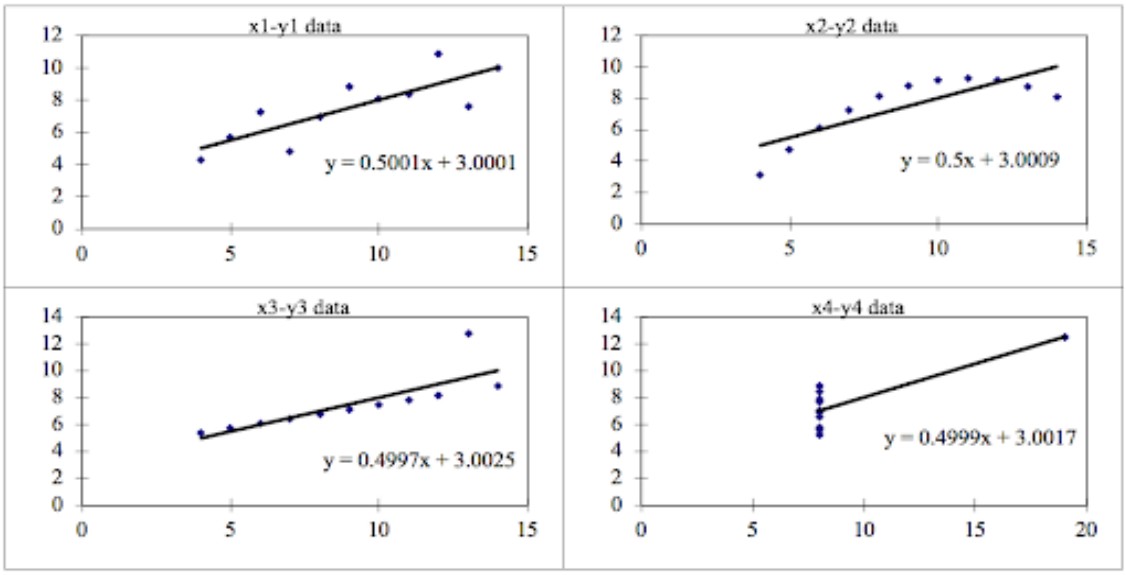
Anscombe's quartet highlights the importance of visualizing data before applying various algorithms to build models. This suggests the data features must be plotted to see the distribution of the samples that can help you identify the various anomalies present in the data (outliers, diversity of the data, linear separability of the data, etc.). Moreover, the linear regression can only be considered a fit for the data with linear relationships and is incapable of handling any other kind of data set.

It was constructed to illustrate the importance of plotting the graphs before analyzing and model building, and the effect of other observations on statistical properties.There are these four data set plots which have nearly same statistical observations, which provides same statistical information that involves variance**,** and mean of all x,y points in all four datasets.

Also, the Linear Regression can be only be considered a fit for the data with linear relationships and is incapable of handling any other kind of datasets. These four plots can be defined 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

However, when these models are plotted on a scatter plot, each data set generates a different kind of plot that isn't interpretable by any regression algorithm, as you can see below:



ANSCOMBE'S QUARTET FOUR DATASETS

Data Set 1: fits the linear regression model pretty well.
Data Set 2: cannot fit the linear regression model because the data is non-linear.
Data Set 3: shows the outliers involved in the data set, which cannot be handled by the linear regression model.
Data Set 4: shows the outliers involved in the data set, which also cannot be handled by the linear regression model.

Application:

The Anscombe's quartet is still often used to illustrate the importance of looking at a set of data graphically before starting to analyze according to a particular type of relationship, and the inadequacy of basic statistic properties for describing realistic datasets.

3. What is Pearson's R?

The Pearson's R is a correlation coefficient, that measures linear coorreletion between the two sets of data. It was developed by Karl Pearson. The Pearson coefficient is the ratio between the covariance of two variables and the product of their standard deviations. Thus, it is essentially a normalized measurement of the covariance, such that the result always has a value between −1 and 1. As with covariance itself, the measure can only reflect a linear correlation of variables, and ignores many other types of relationships or correlations. As a simple example, one would expect the age and height of a sample of teenagers from a high school to have a Pearson correlation coefficient significantly greater than 0, but less than 1 (as 1 would represent an unrealistically perfect correlation).
Pearson's correlation coefficient is commonly represented by 'r'.
The formula for Pearson's R is:

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

where

- n is sample size

- x_i , y_i are the individual sample points indexed with *i*
- $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ (the sample mean) and analogously for r_{xy}

Pearson's correlation coefficient is sensitive to the scale, means it is influenced by the units in which the variables are measured. It only measures the strength of a linear relationship and may not accurately represent nonlinear associations. The correlation does not imply causation. Even if two variables are correlated, it does not necessarily mean that one variable causes the other, correlation simply indicates a statistical association.

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

Scaling 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. The purpose of scaling is to bring all variables to a similar scale or level playing field, which can be crucial for certain machine learning algorithms and statistical techniques.Most of the times, collected data set contains features highly varying in magnitudes, units and range. If scaling is not done then 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.It is important to note that **scaling just affects the coefficients** and none of the other parameters like **t-statistic, F-statistic, p-values, R-squared**, etc.

There are two common types of scaling: normalized scaling and standardized scaling.

Normalized Scaling (Min-Max Scaling): This method scales the values of a variable to a specific range, usually between 0 and 1. The formula for normalized scaling is:

X normalized = X−min(X) /max(X)−min(X)

Where, X is an individual data point, min(X) is the minimum value of the variable, and max(X) is the maximum value of the variable.

Standardized Scaling (Z-score Standardization): This method scales the values of a variable to have a mean of 0 and a standard deviation of 1. The formula for standardized scaling is:

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

Standardized scaling is particularly useful when the distribution of the variable is approximately normal, as it helps make the variable's values interpretable in terms of standard deviations from the mean.

The difference between normalized scaling and standardized scaling:

The values of a normalized dataset will always fall between 0 and 1. A standardized dataset will have a mean of 0 and a standard deviation of 1, but the maximum and minimum values are not constrained by any specified upper or lower bounds.

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

Variance Inflation Factor (VIF):

The **variance inflation factor** (**VIF**) is the ratio of the variance of estimating some parameter in a model that includes multiple other terms by the variance of a model constructed using only one term. It quantifies the severity of multicollinearity in an ordinary least squares regression analysis. It provides an index that measures how much the variance of an estimated regression coefficient is increased because of collinearity.

$$\mathrm{VIF_i} = \frac{1}{1 - R_i^2}$$

When R_i^2 is equal to 0, and therefore, when VIF or tolerance is equal to 1, the i^{th} independent variable is not correlated to the remaining ones, meaning that multicollinearity does not exist.

In general terms,

- VIF equal to 1 = variables are not correlated
- VIF between 1 and 5 = variables are moderately correlated
- VIF greater than 5 = variables are highly correlated

The higher the VIF, the higher the possibility that multicollinearity exists, and further research is required. When VIF is higher than 10, there is significant multicollinearity that needs to be corrected.

VIF measures the strength of the correlation between the independent variables in regression analysis. This correlation is known as multicollinearity, which can cause problems for regression models.

While a moderate amount of multicollinearity is acceptable in a regression model, a higher multicollinearity can be a cause for concern. as the information provided by these variables is redundant.

This situation indicates an extremely high degree of multicollinearity, where one or more predictors in the model can be almost perfectly predicted by a linear combination of the other predictors. In practice, infinite VIF values suggest that the associated predictor is redundant because its variation is almost completely explained by the other predictors.

To address the issue of multicollinearity, we may need to consider removing one or more correlated predictors from the model or using techniques such as feature selection or dimensionality reduction. Additionally, centering or scaling variables can sometimes mitigate multicollinearity, but extreme cases may still result in high VIF values, potentially leading to computational instability.

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

A Q-Q plot, which stands for Quantile-Quantile plot, It is a graphical tool used to assess whether a dataset follows a particular theoretical distribution, such as the normal distribution. In the context of linear regression, Q-Q plots are often used to check the assumption of normality of residuals.
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.
It is used to check following scenarios:
If two data sets —

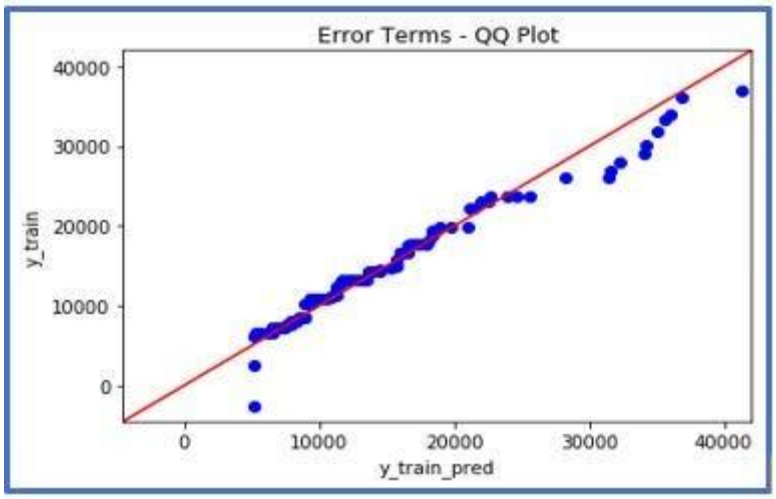
i. come from populations with a common distribution
ii. have common location and scale
iii. have similar distributional shapes
iv. have similar tail behavior

Interpretation:

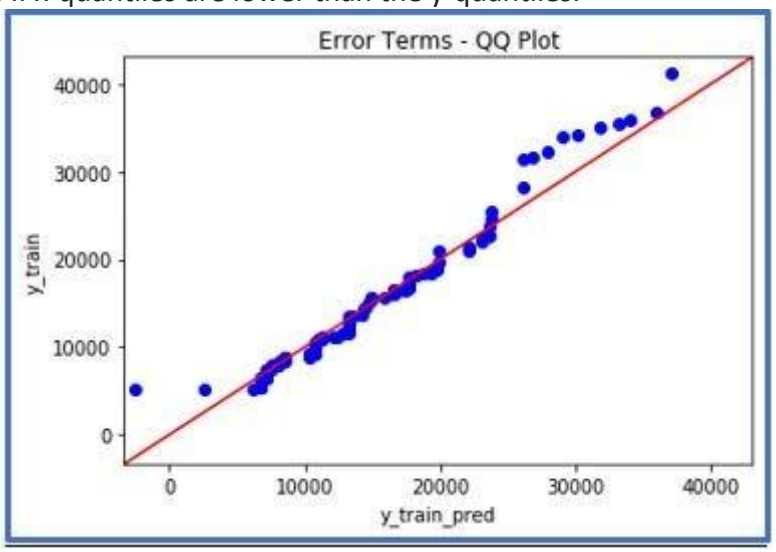
A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set.
Below are the possible interpretations for two data sets.

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