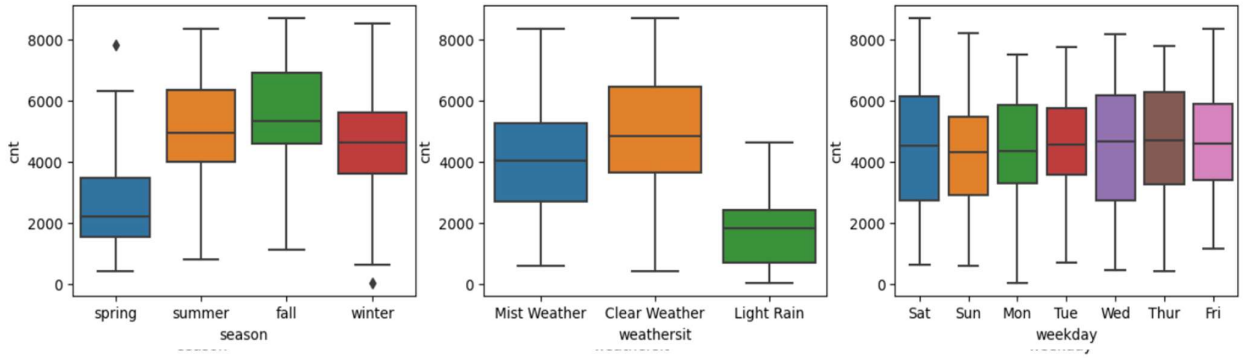
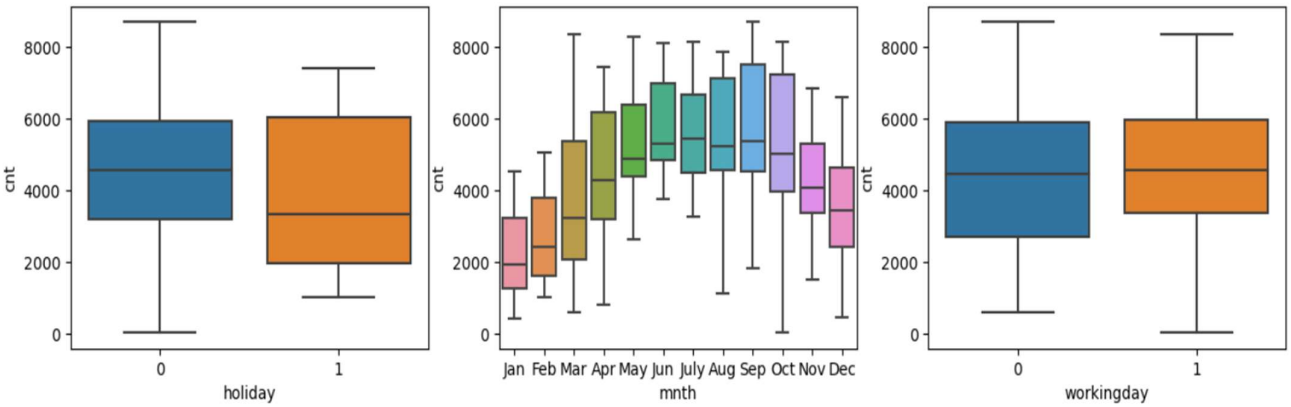
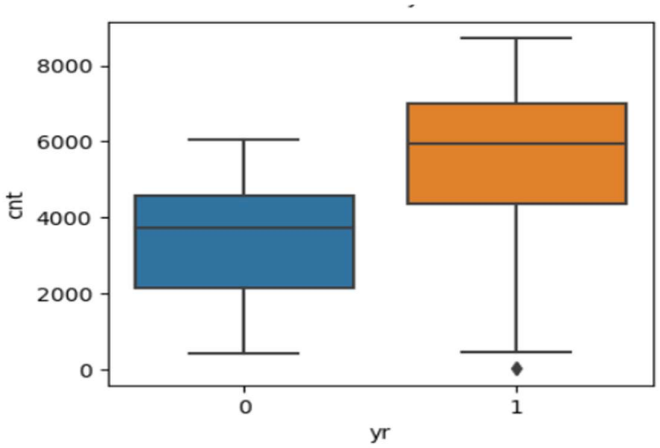
Assignment-based Subjective Questions

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







- The median correspond to the count of total rental bikes is the highest during fall season and lowest during spring season
- The median correspond to the count of total rental bikes is the highest during clear weather and lowest during light rain
- The median correspond to the count of total rental bikes are higher (and almost same) on Wednesday, Thursday and Saturday. It is slightly lower on other days.
- The median correspond to the count of total rental bikes is the highest during holiday
- The median correspond to the count of total rental bikes is the highest in 2019

- The median corresponding to the count of total rental bikes is showing highest in July and September. It is the lowest in Jan.
- There is no significant difference in the median between working day and non-working day

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

Categorical variables in Machine Learning are encoded by 0/1 before being used in model. If there is a categorical variable having m different levels (possible values), then m dummy variables are required at first glance to signify each case uniquely. Example – Lets consider a variable named "Colour" which has three values: "Red", "Yellow", "Green". If we want to encode this variable using 0/1, then we can create these combinations to identify each case uniquely:

	Red	**Yellow**	**Green**
Red:	1	0	0
Yellow:	0	1	0
Green:	0	0	1

Now. "Red" can be identified by either "Red=1 & Yellow=0 & Green=0" or "Yellow=0 & Green=0".

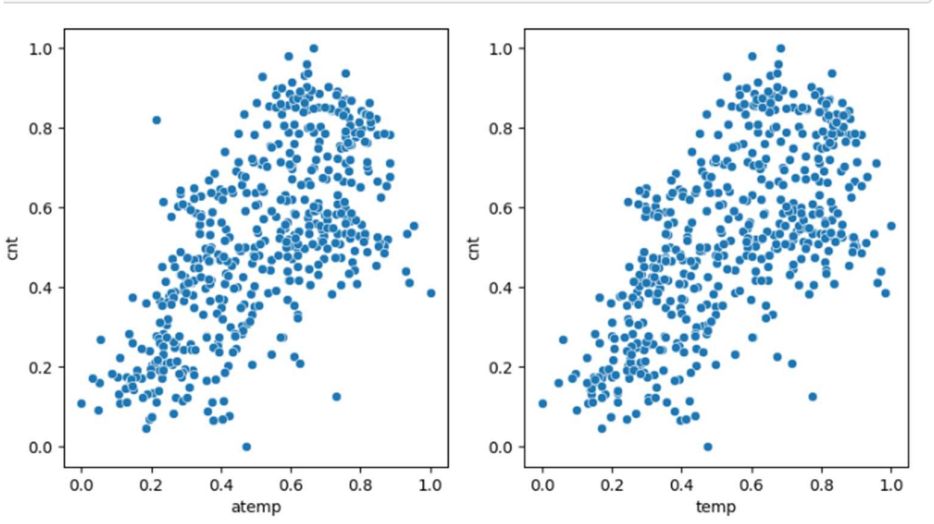
If we use combination of 3 binary variables ("Red=1 & Yellow=0 & Green=0") to identify "Red" there will be chance of multicollinearity among these binary variables which may impact model.

If we use combination of 2 binary variables ("Yellow=0 & Green=0".) then there will not be any multicollinearity issue among new binary variables and the no. of required variables will also be less at the same time.

So, "drop_first=Ture" is used as option of pd.get_dummies() option to drop first column of dummy variables before being used in model.

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

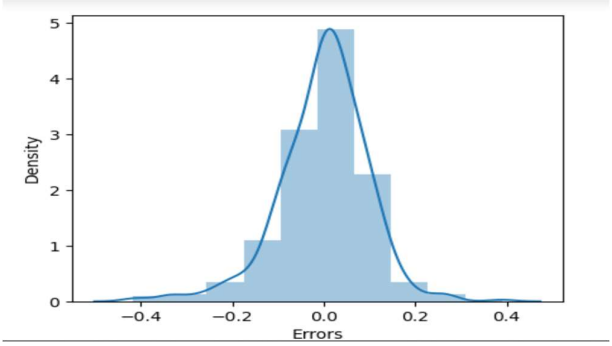
"temp", "atemp"



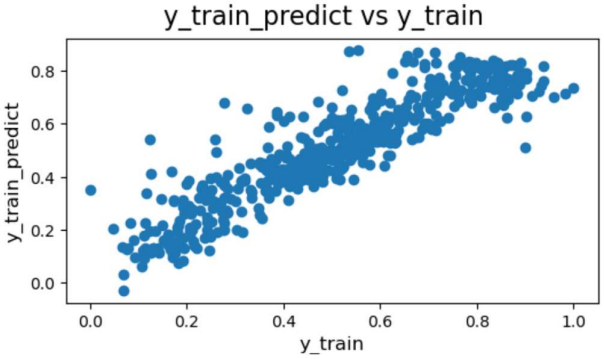
4. How did you validate the assumptions of Linear Regression after building the model on the training set?

   I have verified below points which are assumptions of a linear regression:

   a. Residuals (actual - predicted) are normally distributed with mean at 0 (same plot is present in Jupyter notebook)



   b. Error terms have constant variation (homoscedasticity is present). There is no funnel kind of structure visible in the plot. That means that there is no heteroscedasticity. (same plot is present in Jupyter notebook)



   c. There is no auto corelation
   Durbin-Watson (the measure of correlation of a variable's values over time) value of final dataset is almost 2.0 which means that there is no auto corelation (data mentioned below is present in Jupyter note book)

Omnibus:	58.242	**Durbin-Watson:**	1.999
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	141.824
Skew:	-0.596	**Prob(JB):**	1.60e-31
Kurtosis:	5.292	**Cond. No.**	10.4

   d. There is no multicollinearity. VIF of all predictors is less than 5.
   (data mentioned below is present in Jupyter note book)

	Features	VIF
0	const	19.77
2	temp	1.64
4	spring	1.61
3	windspeed	1.05
5	Light Rain	1.04
6	Mist Weather	1.03
1	yr	1.02

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

   Top 3:

   I. Temp: bike rent increased when there is increase in "temp"
   II. Light rain: bike rent decreased when there is decrease in "LightRain"
   III. Year: bike rent is increased year by year

General Subjective Questions and Answers

1. Explain the linear regression algorithm in detail.

   The linear regression is a statistical method which is used to study the relationship between two variables. One of the variables is called predictor or independent variable (X) and other one is called dependent or output or outcome variable (y).

   It is used as the most commonly used predictive technique in machine learning which predicts output as a continuous variable. It falls under "supervised learning" method (where machine learning model learns from previous data through labels) of machine learning model.

   It is used in different kinds of industries. Some real scenarios where linear regression can be used for prediction:

   a) Advertising spending (X) and revenue (y)
   b) Rain/Water (X) and Crop yield (y) etc.

   The linear regression technique can be classified into two sub categories depending on the no. of independent/predictor variables used to build the model:

   a) Simple linear regression
   b) Multiple linear regression

   a) Simple linear regression:

   In simple linear regression, there is only one predictor variable (X) which is used to predict the output variable (y) and it is explained by plotting a straight line named as regression line or best fit line in a scatter plot of X and y.

$$y = b_0 + b_1X + E$$

Here,

'b_0' : intercept

'b_1' : coefficient or slope

'E' : Random error of measurement

b) Multiple linear regression:

Multiple linear regression is an extension of single linear regression. In multiple linear regression, there are multiple predictor variables (X_1, X_2, .. ,X_N) which are used to predict the output variable (y) and it fits a hyperplane. The equation of multiple linear regression:

$$y = b_0 + b_1X_1 + b_2X_2 + .. + b_NX_N + E$$

Here,

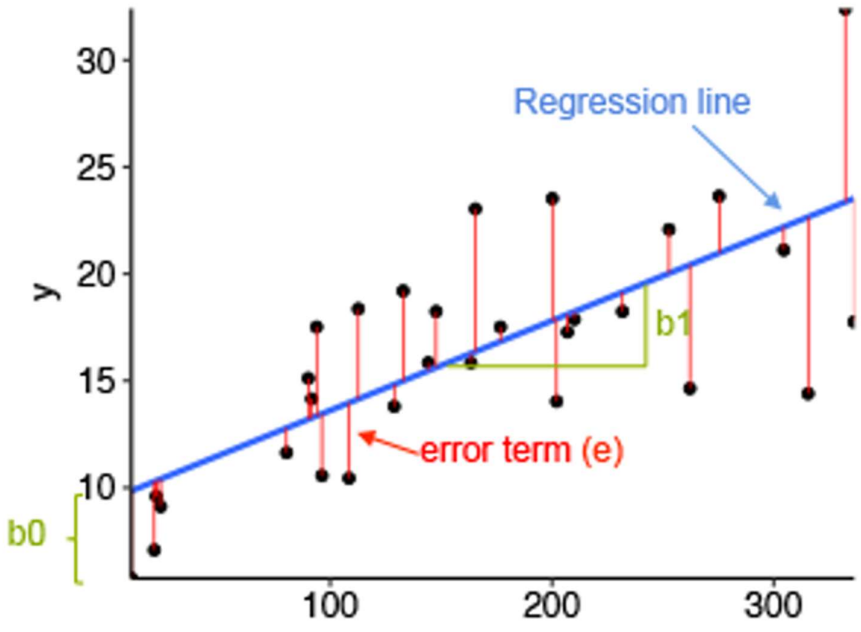
'b_0' : intercept

'$b_1, b_2, .. , b_N$' : coefficient or slope correspond to predictor '$X_1, X_2, .., X_N$'

'E' : Random error of measurement

The main aim of linear regression model is to find the best fit linear line to get the optimal values of intercept and coefficients such that the error is minimized. Here error is the difference between the actual value and predicted value.

The linear aggression can be explained by a simple graph mentioned below:



X

Image Source: Statistical tools for high-throughput data analysis

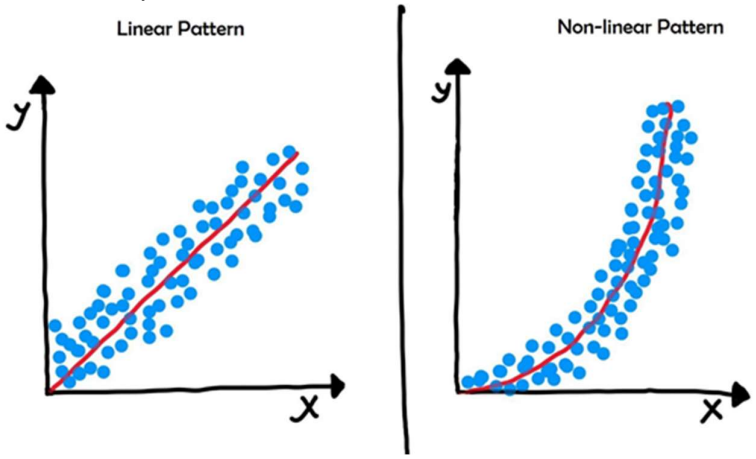
In the above diagram,

- X is the independent variable / predictor whereas y is the dependent / output.
- The blue line is the best fit line containing values predicted by the model
- Black dots are the actual values.
- b_o is the intercept which is 10 and b_1 is the slope of the x variable.
- The vertical distance between the black dots and the regression line is known as error or residual. Each data point has one residual and the sum of all the differences is known as the Sum of Residuals/Errors.

The assunmption of linear regression:

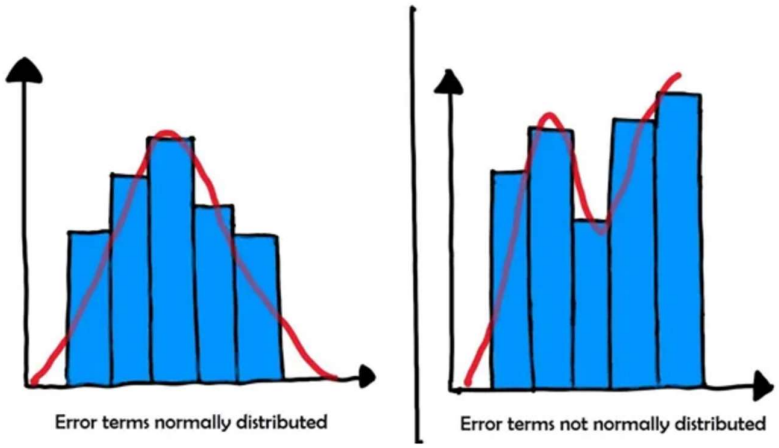
I. Linear relationship between X and y:

Linear regression model can be built only if there is linear relationship between X and y.



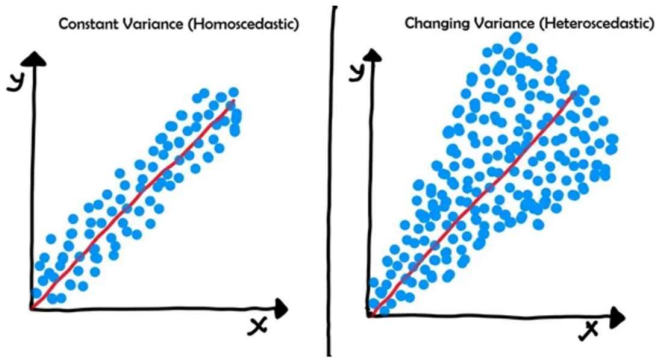
II. Error terms are normally distributed

The error terms (predicted - actual) are normally distributed with mean at 0



III. Constant variation of error terms (Homoscedasticity)

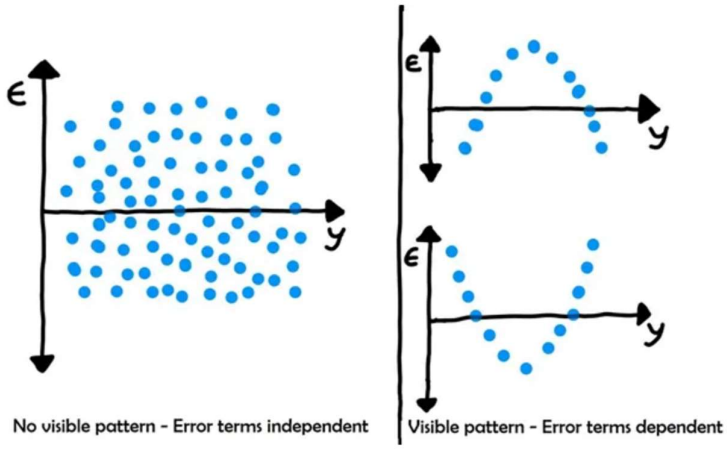
The variance of the error terms should be constant for all values of X. This assumption can be checked by plotting a residual plot. If the assumption is violated then the points will form a funnel shape and that scenario is called heteroscedasticity.



IV. Error terms are independent of each other

The error terms should not be dependent on one another – there should not be any correlation among values of the same variable at different interval.

It can be tested using “Durbin Watson” technique. The value of the test lies between 0 and 4. The value close to 2 means no autocorrelation, positive autocorrelation when value is less than 2 and negative autocorrelation when value is greater than 2



V. No multicollinearity

Multicollinearity occurs when independent variables are correlated. It can happen when an independent variable is computed from other variables in the data set or if two independent variables provide similar results. It does not impact regression estimate but it can be difficult to determine the influence of independent variables on the dependent variable individually. **"Variance Inflation Factor" (VIF)** , a statistical method, can detect and measure the amount of collinearity in a multiple regression model. In general, a VIF value less than 5 is accepted whereas VIF > 5 is not accepted.

Evaluation of linear regression model

a. R Squared:

R2 is explains what portion of the given data variation is explained by the developed model. It always takes a value between 0 & 1. Overall, the higher the R-squared, the better the model fits your data. Mathematically, it is represented as:

R2 = 1 - (RSS / TSS)

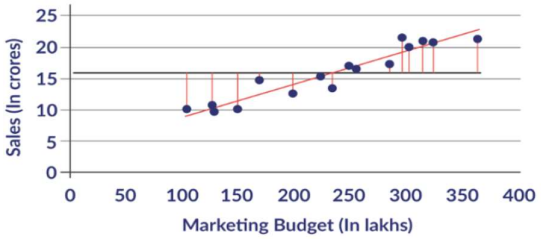
$$R^2 = 1 - \frac{SS_{RES}}{SS_{TOT}} = \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}$$

Where RSS = Residual Sum of Square (difference between actual and predicted)

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

TSS = Total Sum of Square (sum of errors of the data points from mean of response variable)

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



b. Adjusted R Square:

Normal R Square has a drawback that it never decreases. It either remains same or increases with addition of new predictor into the model irrespective of the fact that new predictor is redundant or not. Adjusted R Square handles this. It is an improvement over normal R Square. Adjusted R Square considers number of predictor variables during calculation. In that way, it can determine whether addition of any new variable will be a good fit for the model or not. Adjusted R Square can decrease if new predictor is not a good fit for the model. This is not the case for R Square which never decreases.

$$Adjusted\ R^2 = \{1 - [\frac{(1 - R^2)(n-1)}{(n-k-1)}]\}$$

R2 = Normal R2
K = No. of predictor
N = Sample size

c. Mean Squared Error (MSE):

It is the mean of the squared difference of actual vs predicted values.

d. Root Mean Squared Error (RMSE):

Root of the mean difference of Actual and Predicted values. RMSE penalizes the large errors whereas MSE doesn't.

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

d. MAE (Mean Absolute Error)

Mean absolute error (MAE) is a common measure of how far the predicted values are from the actual values in a dataset. It is the average absolute difference between the predicted and actual values. In simple terms, MAE measures the average distance between the predicted values and the actual values in a dataset, and it is often used in regression analysis and machine learning to evaluate the performance of a model. The lower the value of MAE, the better the model's performance.

Steps / Flow chart of building a linear regression model in Python:

1. Reading and cleaning the Data using EDA (Exploratory Data Analysis) technique
2. Visualising the data through different plots to understand linear relationship between outcome and predictor variables
3. Data preparation for Modelling –
    a. This step includes creating dummy variables which contain 0/1 value and represent each categorical data uniquely
    b. Scaling numeric variable so that all variables are in same range – between 0 and 1
    c. Splitting data set into train and test

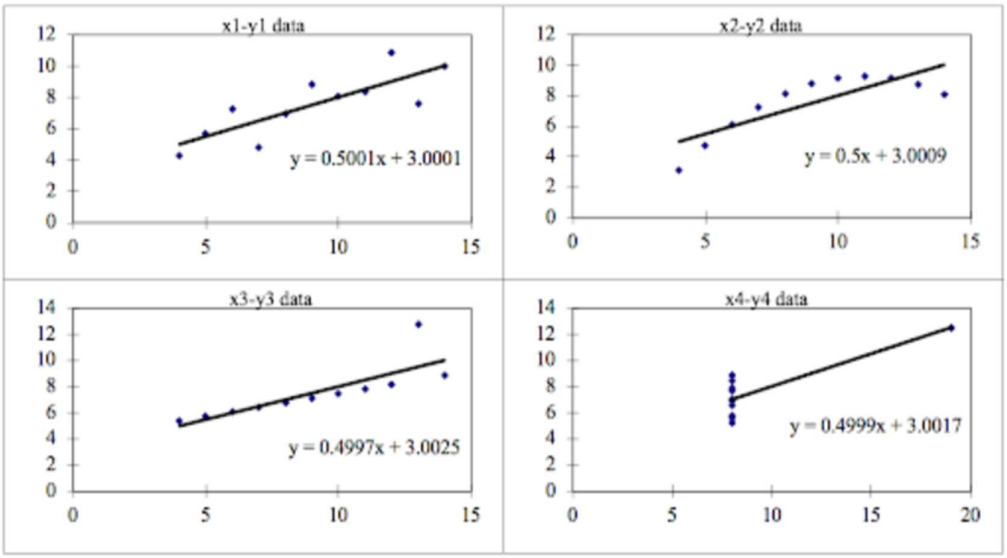
4. Building a linear model using statsmodel and scikit-learn – VIF (should be <5), p value (<0.05), R2, Adjusted R2, F statistics are checked to know which variables are best fit for the linear model
5. Residual analysis of the training data to check normal distribution of error terms
6. Making prediction using final model – the plot of y_predict vs y on training set is done to check homoscedasticity and then applied on test set
7. Model evaluation is done by comparing R2 between train and test data based result
8. Summary – Final linear regression equation is created and dominant predictors are identified

2. Explain the Anscombe's quartet in detail

Anscombe's quartet tells us about the importance of visualizing data before applying various algorithms to build models. It actually suggests to visualize the distribution of the samples to see whether there is any anomaly present in the data (outliers, non-linear relationship between X & y, etc.) or not.

This was constructed and explained by statistician Francis Anscombe in 1973 through four data sets having same descriptive statistics (variance, mean etc.) but having different plots.

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	



Data set x1-y1: the data set is kind of linear

Data set x2-y2: data set is not linear. Linear regression model can't be fitted here.

Data set x3-y3: data set contains some outliers. Linear model is sensitive to outlier

Data set x4-y4: data set contains outlier. Linear model is sensitive to outlier

3. What is Pearson's R?

Pearson correlation coefficient or Pearson's R is defined as the measurement of the strength of the linear relationship between two variables and the direction of relationship between each other. It basically calculates the effect of change in one variable when the other variable changes. It is the most common measure of corelation in statistics to find the linear relationship.

Pearson R returns value between +1 and -1:

a) +1: It indicates strong positive (every positive increase in the value of one variable causes the positive increase in the value of another variable) linear corelation

b) -1: It indicates strong negative (every positive increase in the value of one variable causes decrease in the value of another variable) linear corelation.

c) 0: It indicates no linear corelation

The formula used to calculate Pearson's R:

$$r = \frac{N\Sigma xy-(\Sigma x)(\Sigma y)}{\sqrt{[N\Sigma x^2-(\Sigma x)^2][N\Sigma y^2-(\Sigma y)^2]}}$$

N = the number of pairs of scores

Σxy = the sum of the products of paired scores

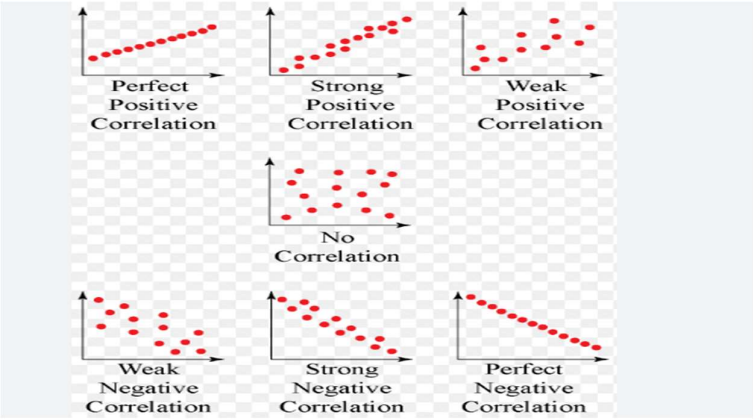
Σx = the sum of x scores

Σy = the sum of y scores

Σx2 = the sum of squared x scores

Σy2 = the sum of squared y scores

Graphical representation of positive, negative and zero corelation:



The limitation of Pearson's R:

a. It will not be useful if two variables are not linear.
b. It is sensitive to outliers.

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

a. What is scaling?

It is the process of normalising the value of independent variables within a specific range. It is performed during data pre-processing stage to manage highly varying magnitude, unit.

b. Why is scaling performed?

I. Scaling guarantees that all features are on a comparable scale and have comparable ranges so that the models give equal importance to all features instead of being biased towards the features of higher magnitude otherwise the resultant coefficient will be impacted

II. Algorithm performance is improved and it converges more quickly

c. What is the difference between normalized scaling and standardized scaling?

I. Normalized Scaling (Min-Max scaling):

It is a scaling technique that scales the features in the range of 0 to 1. As its formula depends on the minimum and maximum value of features, hence we need to ensure those values are correct and are not affected by outliers.

$$x_{\text{norm}} = \frac{x - \min(x)}{\max(x) - \min(x)}$$

It does not assume that the data is normally distributed.

II. Standardization (Z-Score Scaling):

Standardization is a technique that scales the feature in such a way that its mean becomes zero and standard deviation 1. There is no predefine range between feature will get scaled (like in normalization range is 0 to 1).

$$x_{\text{stand}} = \frac{x - \text{mean}(x)}{\text{standard deviation } (x)}$$

It assumes that the data is normally distributed.

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

The VIF formula:

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

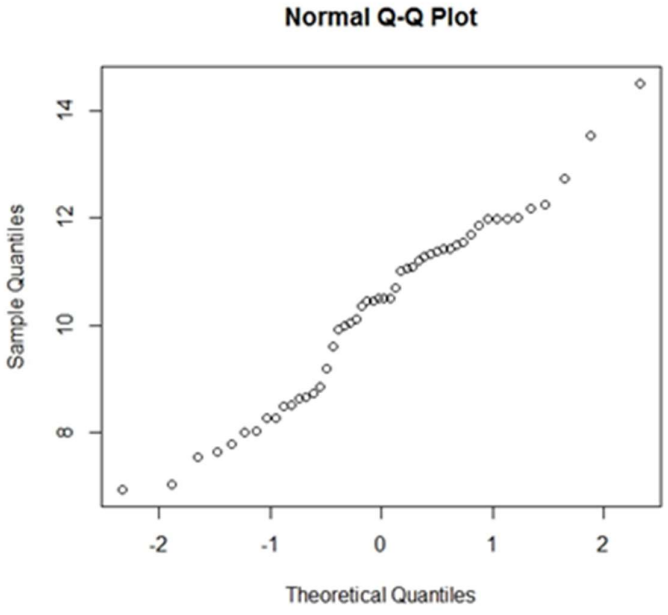
'i' refers to the 'I' th variable.

If R-squared value is equal to 1 then the denominator of the above formula become 0 and the overall value becomes infinite. It denotes perfect correlation in variables.

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

The Q-Q plot or quantile-quantile plot is a graphical technique for determining if two data sets come from populations with a common distribution.

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. Here's an example of a Normal Q-Q plot when both sets of quantiles truly come from Normal distributions.



Use of Q-Q plot in Linear Regression:

The Q-Q plot is used to see if the points lie approximately on the line. If they don't, it means, our residuals aren't Gaussian (Normal) and thus, our errors are also not Gaussian.

Importance of Q-Q plot: Below are the points:

I. The sample sizes do not need to be equal.

II. Many distributional aspects can be simultaneously tested. For example, shifts in location, shifts in scale, changes in symmetry, and the presence of outliers.

III. The q-q plot can provide more insight into the nature of the difference than analytical methods.