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) I have a dataset having 56 variables, in which 4 dependent and 52 independent variables. In those independent variables 45 variables are categorical and 3 dependent variables are categorical remains are continuous. Each variables having 1500 observations. Independent variables are nominal, and dependent Categorical variables are ordinal. I want to check, there is any effect of independent variables on each dependent variables .

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

Ans) I think it depends on the model. If you don't drop the first column then your dummy variables will be correlated (redundant as Dimitre shows in the post below). This may affect some models adversely and the effect is stronger when the cardinality is smaller. For example iterative models may have trouble converging and lists of variable importances may be distorted.

If you have a small number of dummies, i suggest removing the first dummy. For example, if you have a variable gender, you don't need both a male and female dummy. Just one will be fine. If male=1 then the person is a male and if male=0 then the person is female. However if you have a category with hundreds of values, I suggest not dropping the first column. That will make it easier for the model to "see" all the categories quickly during learning (and the adverse effects are negligible).

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

Ans) windspeed is giving the highest correlation with the target variable.

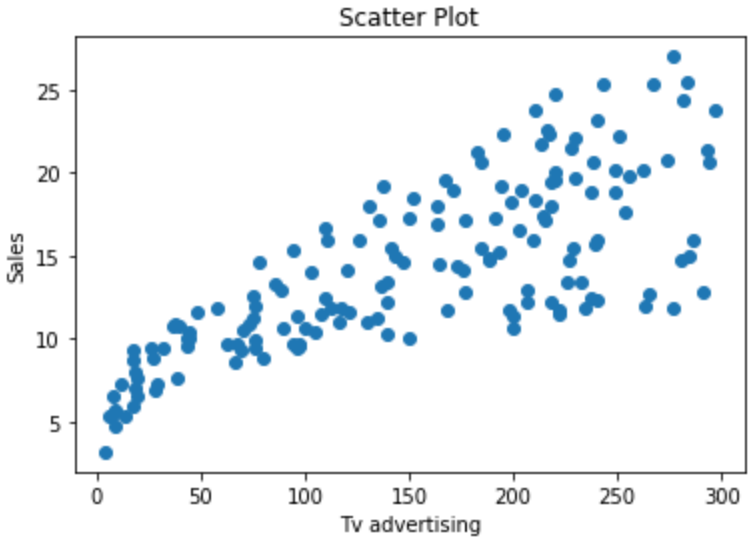
4) How did you validate the assumptions of Linear Regression after building the model on the training set?

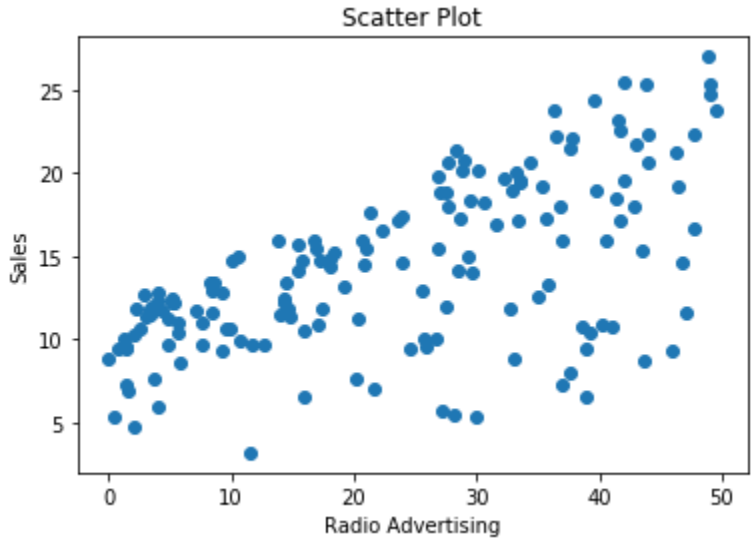
Ans) Assumptions of Linear Regression

There are 5 basic assumptions of Linear Regression Algorithm:

1. **Linear Relationship between the features and target:**

According to this assumption there is linear relationship between the features and target.Linear regression captures only linear relationship.This can be validated by plotting a scatter plot between the features and the target.

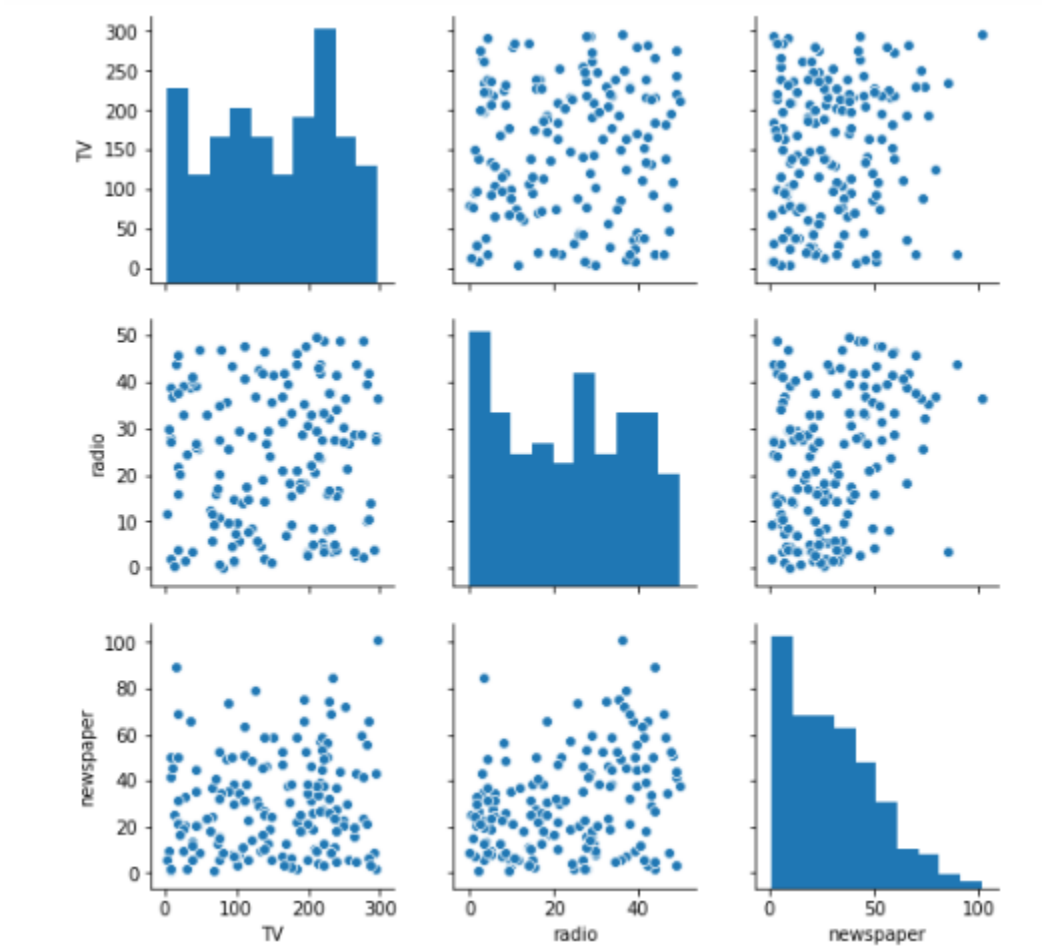




The first scatter plot of the feature TV vs Sales tells us that as the money invested on Tv advertisement increases the sales also increases linearly and the second scatter plot which is the feature Radio vs Sales also shows a partial linear relationship between them,although not completely linear.

2.Little or no Multicollinearity between the features:

Multicollinearity is a state of very high inter-correlations or inter-associations among the independent variables.It is therefore a type of disturbance in the data if present weakens the statistical power of the regression model.Pair plots and heatmaps(correlation matrix) can be used for identifying highly correlated features.

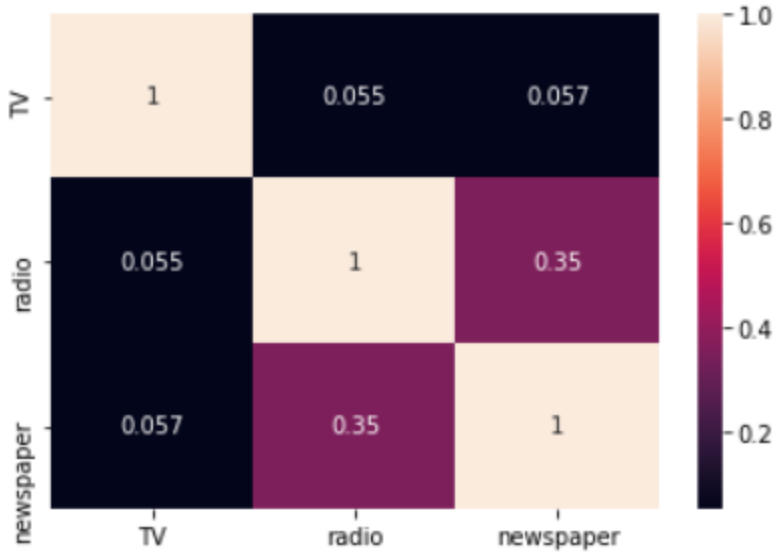


Pair plots of the features

The above pair plot shows no significant relationship between the features.

```
sns.heatmap(df_new.corr(),annot=True)
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x1676c
```



Heat Map(Correlation Matrix)

This heatmap gives us the correlation coefficients of each feature with respect to one another which are in turn less than 0.4.Thus the features aren't highly correlated with each other.

Why removing highly correlated features is important?

The interpretation of a regression coefficient is that it represents the mean change in the target for each unit change in an feature when you hold all of the other features constant. However, when features are correlated, changes in one feature in turn shifts another feature/features. The stronger the correlation, the more difficult it is to change one feature without changing another. It becomes difficult for the model to estimate the relationship between each feature and the target independently because the features tend to change in unison.

How multicollinearity can be treated?

If we have 2 features which are highly correlated we can drop one feature or combine the 2 features to form a new feature,which can further be used for prediction.

3.Homoscedasticity Assumption:

Homoscedasticity describes a situation in which the error term (that is, the "noise" or random disturbance in the relationship between the features and the target) is the same across all values of the independent variables.A scatter plot of residual values vs predicted values is a goodway to

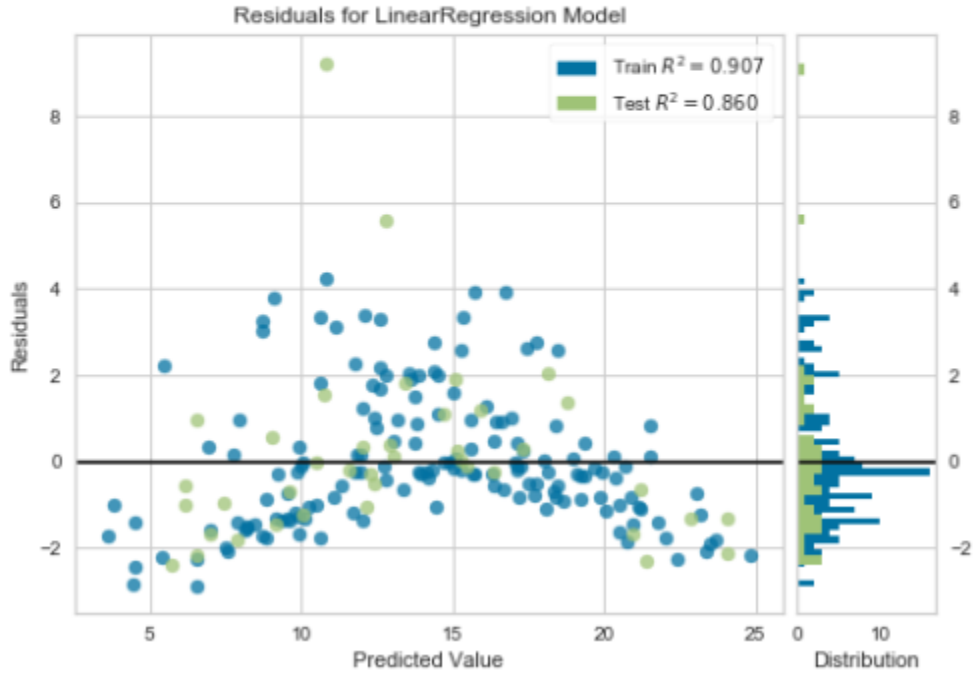
check for homoscedasticity.There should be no clear pattern in the distribution and if there is a specific pattern,the data is heteroscedastic.

Homoscedasticity vs Heteroscedasticity

The leftmost graph shows no definite pattern i.e constant variance among the residuals,the middle graph shows a specific pattern where the error increases and then decreases with the predicted values violating the constant variance rule and the rightmost graph also exhibits a specific pattern where the error decreases with the predicted values depicting heteroscedasticity

Python code for residual plot for the given data set:

```
from yellowbrick.regressor import ResidualsPlot
from sklearn.linear_model import LinearRegression
Lr=LinearRegression()
visualizer = ResidualsPlot(Lr)
visualizer.fit(x_train,y_train)     # Fit the training data to the model
visualizer.score(x_test,y_test)     # Evaluate the model on the test data
visualizer.poof() visualizer.poof() # Draw/show/poof the data
```

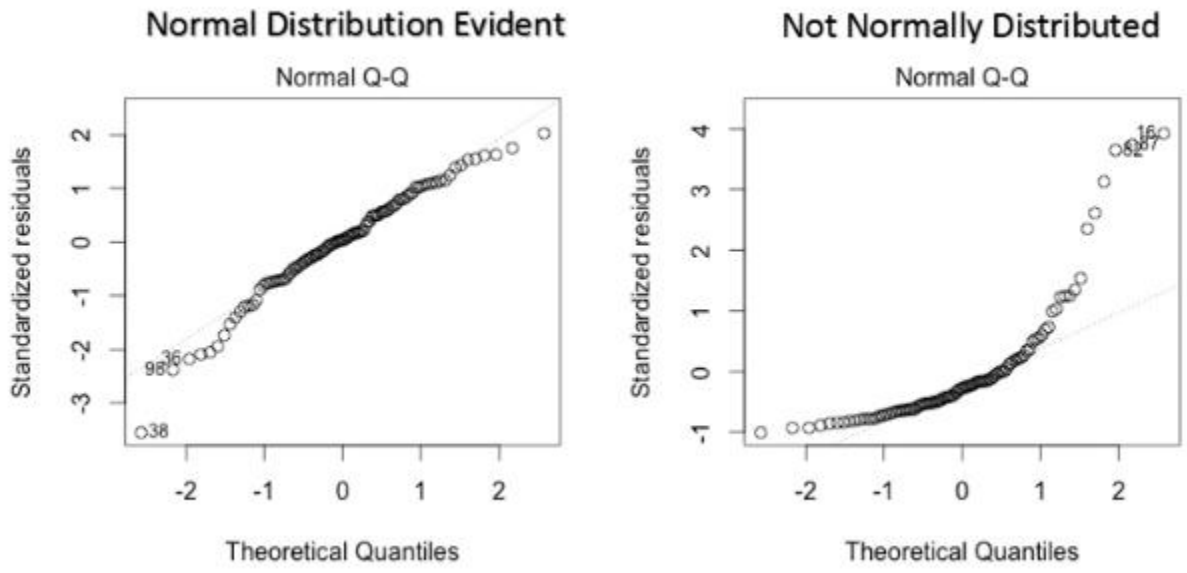


Error(residuals) vs Predicted values

4.Normal distribution of error terms:

The fourth assumption is that the error(residuals) follow a normal distribution.However, a less widely known fact is that, as sample sizes increase, the normality assumption for the residuals is not needed. More precisely, if we consider repeated sampling from our population, for large sample sizes, the distribution (across repeated samples) of the ordinary least squares estimates of the regression coefficients follow a normal distribution. As a consequence, for moderate to large sample sizes, non-normality of residuals should not adversely affect the usual inferential procedures. This result is a consequence of an extremely important result in statistics, known as the central limit theorem.

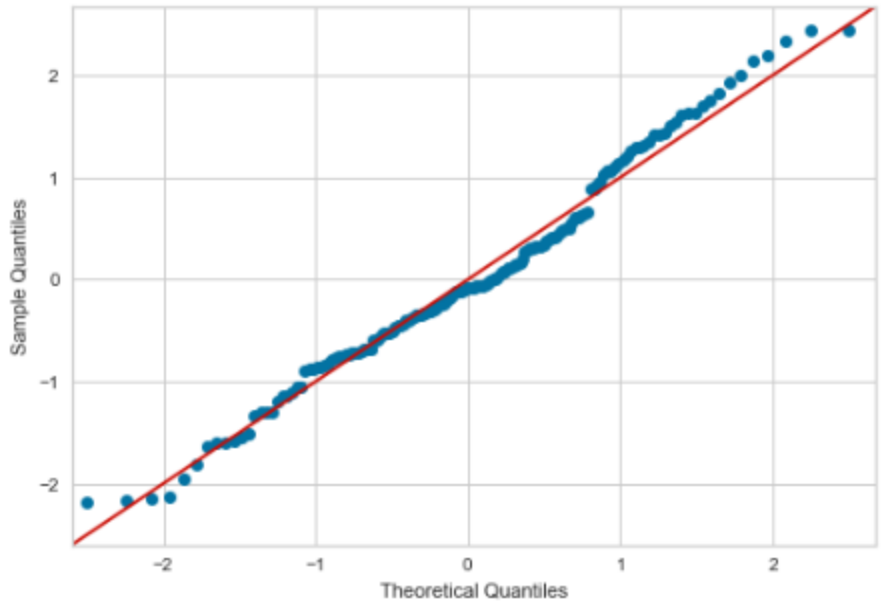
Normal distribution of the residuals can be validated by plotting a q-q plot.



Q-Q Plots

Using the q-q plot we can infer if the data comes from a normal distribution. If yes, the plot would show fairly straight line. Absence of normality in the errors can be seen with deviation in the straight line.

```
import statsmodels.api as sm
mod_fit = sm.OLS(y_train,x_train).fit()
res = mod_fit.resid # residuals
fig = sm.qqplot(res,fit=True,line='45')
plt.show()
```



Q-Q Plot for the advertising data set

The q-q plot of the advertising data set shows that the errors(residuals) are fairly normally distributed.The histogram plot in the “Error(residuals) vs Predicted values” in assumption no.3 also shows that the errors are normally distributed with mean close to 0.

5.Little or No autocorrelation in the residuals:

Autocorrelation occurs when the residual errors are dependent on each other.The presence of correlation in error terms drastically reduces model’s accuracy.This usually occurs in time series models where the next instant is dependent on previous instant.

Autocorrelation can be tested with the help of Durbin-Watson test.The null hypothesis of the test is that there is no serial correlation. The Durbin-Watson test statistics is defined as:

$$\sum_{t=2}^{T}((e_t - e_{t-1})^2)/\sum_{t=1}^{T} e_t^2$$

The test statistic is approximately equal to 2*(1-r) where r is the sample autocorrelation of the residuals. Thus, for r == 0, indicating no serial correlation, the test statistic equals 2. This statistic will always be between 0 and 4. The closer to 0 the statistic, the more evidence for positive serial correlation. The closer to 4, the more evidence for negative serial correlation.

```
model = sm.OLS(y_train,x_train)
results = model.fit()
print(results.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  sales   R-squared:                       0.984
Model:                            OLS   Adj. R-squared:                  0.984
Method:                 Least Squares   F-statistic:                     3191.
Date:                Fri, 24 May 2019   Prob (F-statistic):          2.03e-140
Time:                        22:03:45   Log-Likelihood:                -331.22
No. Observations:                 160   AIC:                             668.4
Df Residuals:                     157   BIC:                             677.7
Df Model:                           3
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
TV             0.0528      0.001     36.622      0.000       0.050       0.056
radio          0.2327      0.010     23.630      0.000       0.213       0.252
newspaper      0.0136      0.007      1.868      0.064      -0.001       0.028
==============================================================================
Omnibus:                        2.123   Durbin-Watson:                   1.885
Prob(Omnibus):                  0.346   Jarque-Bera (JB):                2.140
Skew:                           0.273   Prob(JB):                        0.343
Kurtosis:                       2.847   Cond. No.                         12.5
==============================================================================
```

Summary of the fitted Linear Model

From the above summary note that the value of Durbin-Watson test is 1.885 quite close to 2 as said before when the value of Durbin-Watson is equal to 2, r takes the value 0 from the equation 2*(1-r),which in turn tells us that the residuals are not correlated.

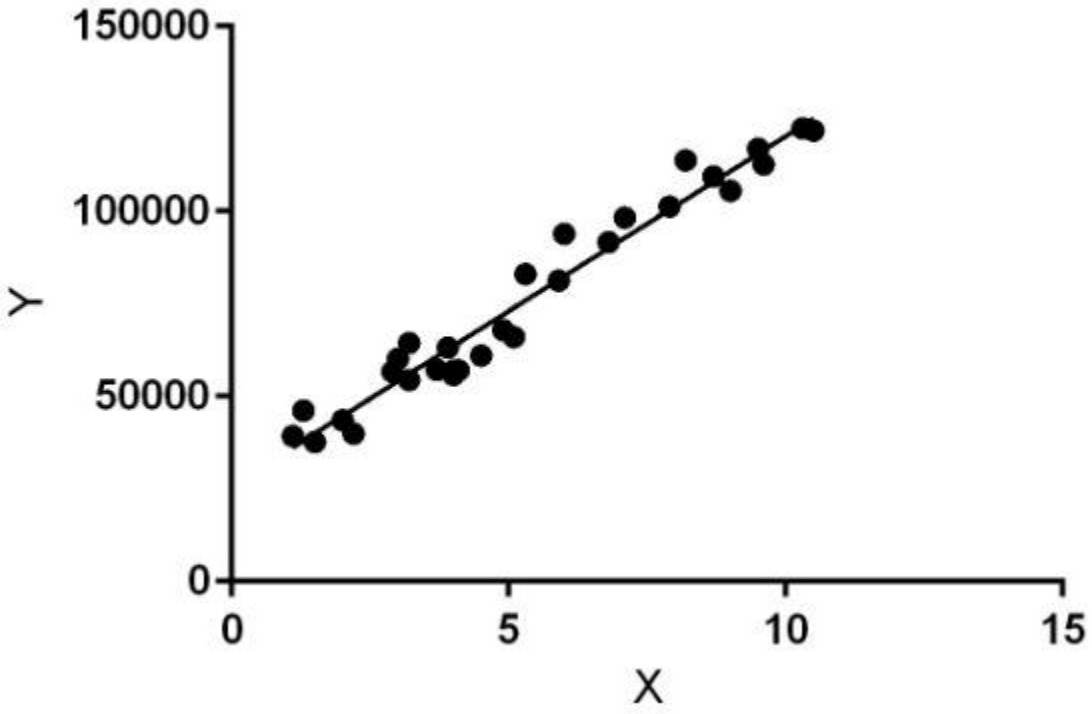
Conclusion:

We have gone through the most important assumptions which must be kept in mind before fitting a Linear Regression Model to a given set of data.These assumptions are just a formal check to ensure that the linear model we build gives us the best possible results for a given data set and these assumptions if not satisfied does not stop us from building a Linear regression mode

General Subjective Questions

1) Explain the linear regression algorithm in detail

Ans) Linear Regression is a machine learning algorithm based on supervised learning. It performs a regression task. Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting. Different regression models differ based on – the kind of relationship between dependent and independent variables, they are considering and the number of independent variables being used.



Linear regression performs the task to predict a dependent variable value (y) based on a given independent variable (x). So, this regression technique finds out a linear relationship between x (input) and y(output). Hence, the name is Linear Regression.
In the figure above, X (input) is the work experience and Y (output) is the salary of a person. The regression line is the best fit line for our model.

Hypothesis function for Linear Regression :

$$y = \theta_1 + \theta_2.x$$

While training the model we are given :
x: input training data (univariate – one input variable(parameter))
y: labels to data (supervised learning)

When training the model – it fits the best line to predict the value of y for a given value of x. The model gets the best regression fit line by finding the best θ_1 and θ_2 values.
θ_1: intercept
θ_2: coefficient of x

Once we find the best θ_1 and θ_2 values, we get the best fit line. So when we are finally using our model for prediction, it will predict the value of y for the input value of x.

Cost Function (J):
By achieving the best-fit regression line, the model aims to predict y value such that the error difference between predicted value and true value is minimum. So, it is very important to update the θ_1 and θ_2 values, to reach the best value that minimize the error between predicted y value (pred) and true y value (y).

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

Cost function(J) of Linear Regression is the **Root Mean Squared Error (RMSE)** between predicted y value (pred) and true y value (y).

Gradient Descent:
To update θ_1 and θ_2 values in order to reduce Cost function (minimizing RMSE value) and achieving the best fit line the model uses Gradient Descent. The idea is to start with random θ_1 and θ_2 values and then iteratively updating the values, reaching minimum cost.

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. Each dataset consists of eleven (*x*,*y*) points. They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data before analyzing it and the effect of outliers and other influential observations on statistical properties. He described the article as being intended to counter the impression among statisticians that "numerical calculations are exact, but graphs are rough."

For all four datasets:

Property	Value	Accuracy
Mean of x	9	exact
Sample variance of x :	11	exact
Mean of y	7.50	to 2 decimal places
Sample variance of y :	4.125	±0.003
Correlation between x and y	0.816	to 3 decimal places
Linear regression line	$y = 3.00 + 0.500x$	to 2 and 3 decimal places, respectively
Coefficient of determination of the linear regression :	0.67	to 2 decimal places

The first scatter plot (top left) appears to be a simple linear relationship, corresponding to two variables correlated where y could be modelled as gaussian with mean linearly dependent on x.

- The second graph (top right) is not distributed normally; while a relationship between the two variables is obvious, it is not linear, and the Pearson correlation coefficient is not relevant. A more general regression and the corresponding coefficient of determination would be more appropriate.
- In the third graph (bottom left), the distribution is linear, but should have a different regression line (a robust regression would have been called for). The calculated regression is offset by the one outlier which exerts enough influence to lower the correlation coefficient from 1 to 0.816.
- Finally, the fourth graph (bottom right) shows an example when one high-leverage point is enough to produce a high correlation coefficient, even though the other data points do not indicate any relationship between the variables.

The 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

The datasets are as follows. The x values are the same for the first three datasets.

Anscombe's quartet

I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71

9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

It is not known how Anscombe created his datasets. Since its publication, several methods to generate similar data sets with identical statistics and dissimilar graphics have been developed

3)What is Pearson's R?

Ans) In statistics, the Pearson correlation coefficient, also referred to as Pearson's *r*, the Pearson product-moment correlation coefficient (PPMCC), or the bivariate correlation,is a statistic that measures linear correlation between two variables *X* and *Y*. It has a value between +1 and −1. A value of +1 is total positive linear correlation, 0 is no linear correlation, and −1 is total negative linear correlation

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

Ans) Scaling is the procedure of measuring and assigning the objects to the numbers according to the specified rules. In other words, the process of locating the measured objects on a continuous sequence of numbers to which the objects are assigned is called as scaling.

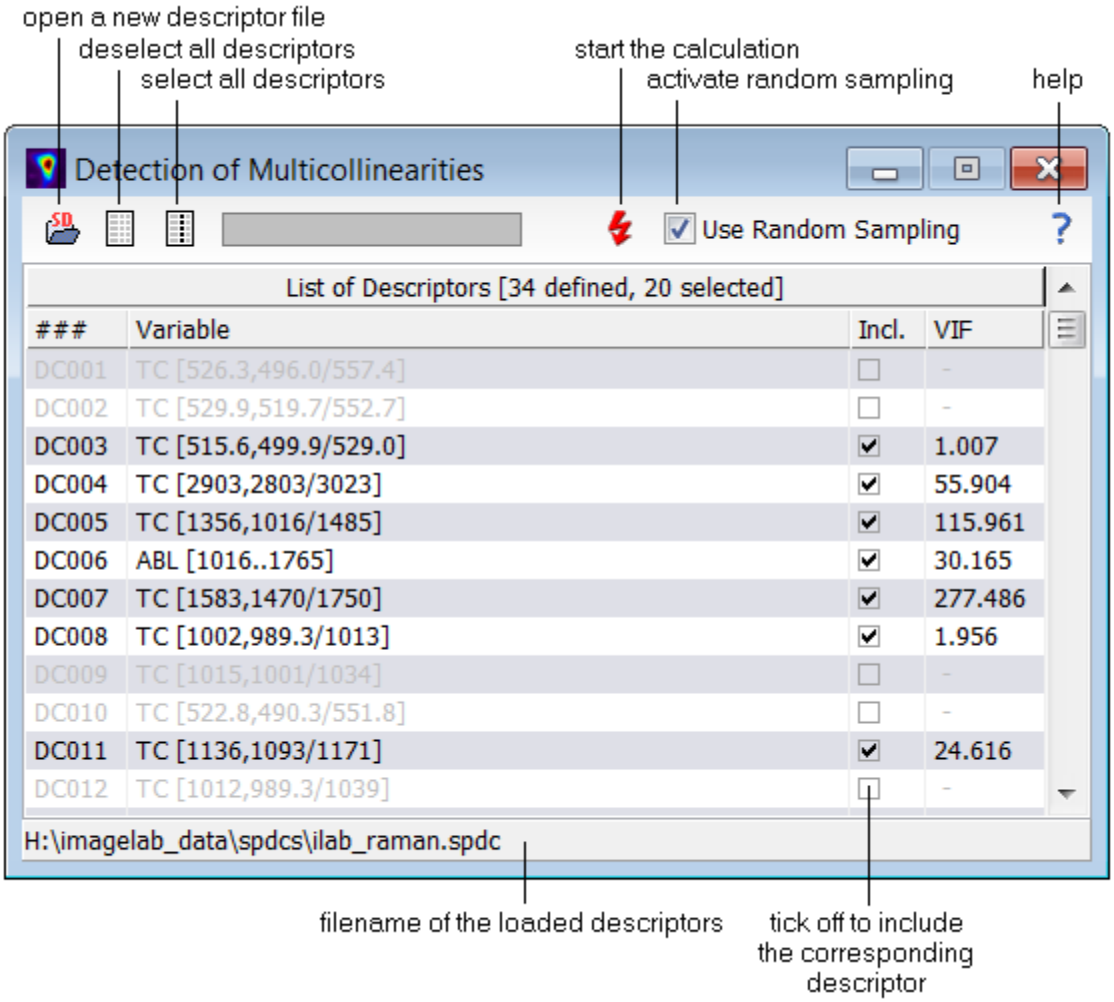
Feature Scaling is a technique to standardize the independent features present in the data in a fixed range. It is performed during the data pre-processing to handle highly varying magnitudes or values or units. If feature scaling is not done, then a machine learning algorithm tends to weigh greater values, higher and consider smaller values as the lower values, regardless of the unit of the values.

Normalization is useful when your data has varying scales and the algorithm you are using does not make assumptions about the distribution of your data, such as k-nearest neighbors and artificial neural networks. Standardization assumes that your data has a Gaussian (bell curve) distribution

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

Ans) The command Tools > VIF of Descriptors... supports the detection of multicollinearities by means of the variance inflation factor (VIF). The user has to select the variables to be included by ticking off the corresponding check boxes. In general one starts with the selection of all variables, and proceeds by repeatedly deselecting variables showing a high VIF. Ideally, the VIF values should be below 10.

As the calculation of the VIF can be quite time consuming, you may choose to use only a random sample of 1000 pixels to calculate the VIF. This increases the speed of calculation considerably, even though the accuracy of the VIF values is degraded. However this should be sufficient to get a rough overview.



For descriptor sets with less than 25 spectral descriptors the calculation of the VIF values is performed automatically upon following any change of the selected descriptors. If a set contains a higher number of descriptors the user can decide when to calculate the VIF values by clicking the "start the calculation" button ( ).

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

Ans) 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 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

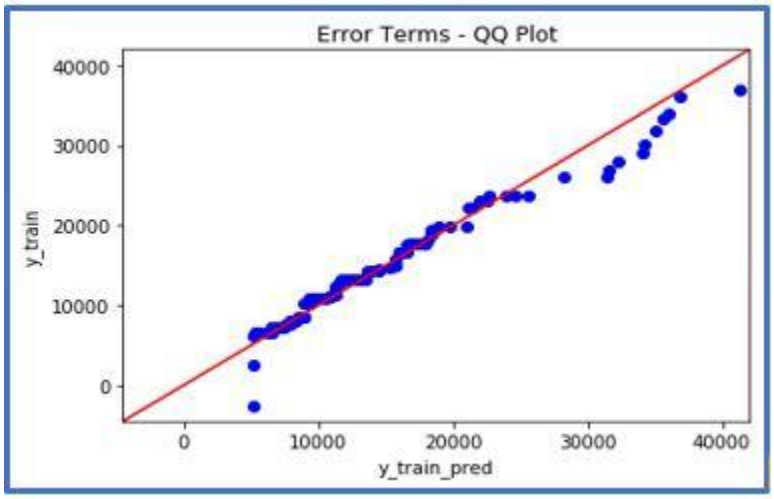
Interpretation:

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

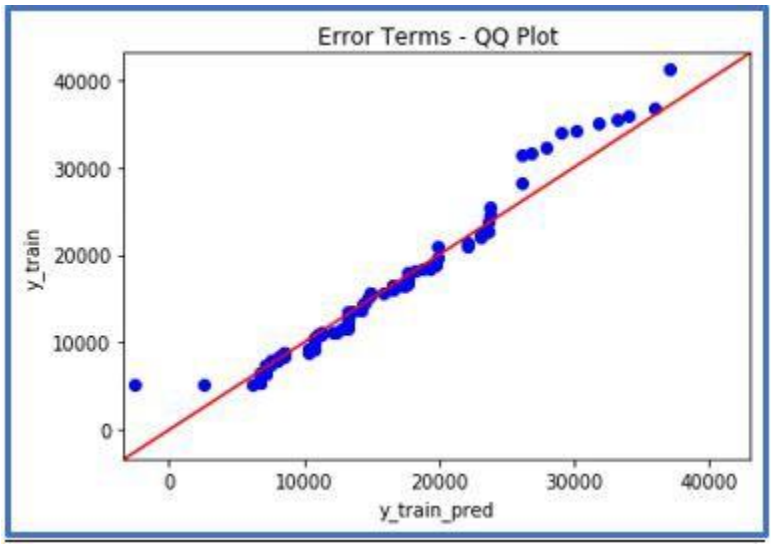
Below are the possible interpretations for two data sets.

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