Assignment

LINEAR REGRESSION

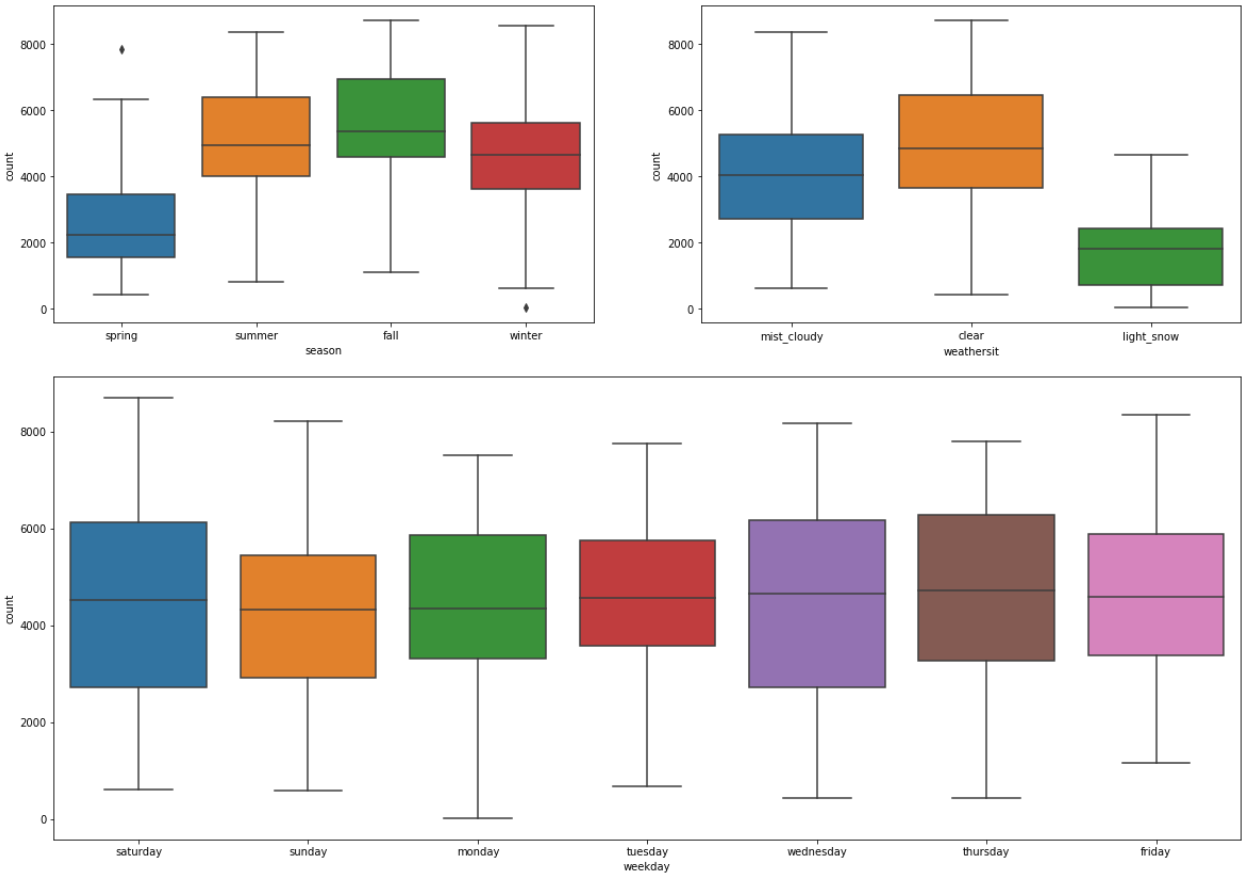
LAKSHAY KAKKAR

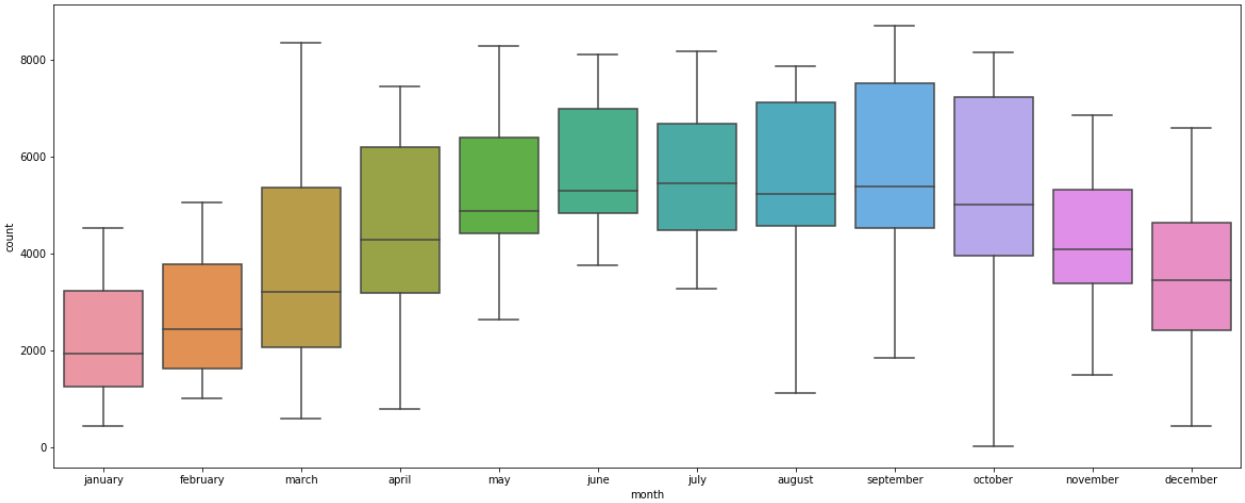
Assignment-based Subjective Questions

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

Ans. Effect of Categorical Variables ‘season’, ‘weathersit’, ‘weekday’ and ‘month’ on Target ‘count’ is visualized and following inferences are drawn:

1. season:
    a. ‘spring’ shows minimum number of median counts among all seasons, but outliers can be observed.
    b. ‘fall’ shows the maximum number of median counts among all seasons.
2. weather:
    a. ‘light_snow’ shows significantly smaller number of counts among all weathers.
    b. ‘heavy_snow’ has 0 counts indicating that services are unavailable during this weather condition.
3. weekday:
    a. The number of median counts is generally same for every day.
4. month:
    a. The beginning and ending months of year shows smaller number of counts.
    b. The number of counts in the mid of the year is (June, July and August) significantly high.





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

Ans. A dummy variable is a numeric variable that represents categorical data, such as season, days of week, gender, race, marital status, educational qualification, etc.

Dummy variables are dichotomous and quantitative variables, technically i.e., their range of values is small and they can take on only two quantitative values. Regression results are easiest to interpret when dummy variables are limited to two specific values, 1 or 0. Where, 1 represents the presence of a qualitative attribute, and 0 represents the absence.

To represent a categorical variable with k different values, one needs to define k - 1 dummy variables, only. As, k^{th} dummy variable is redundant; it carries no new information and **it creates a severe multicollinearity problem for the analysis**. Using k dummy variables when only k - 1 dummy variables are required is known as the dummy variable trap.

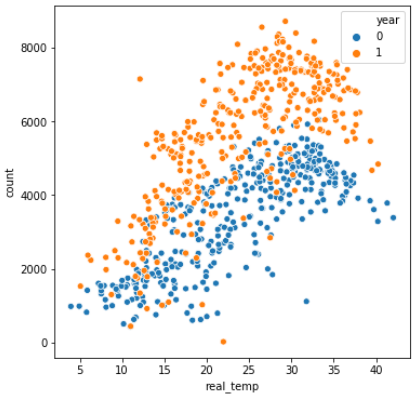
For example, suppose an example where one is interested in marital status, a categorical variable that can have three values - Single, Married, or Divorced. This information can be represented with only 2 variables, i.e., k – 1 variable:

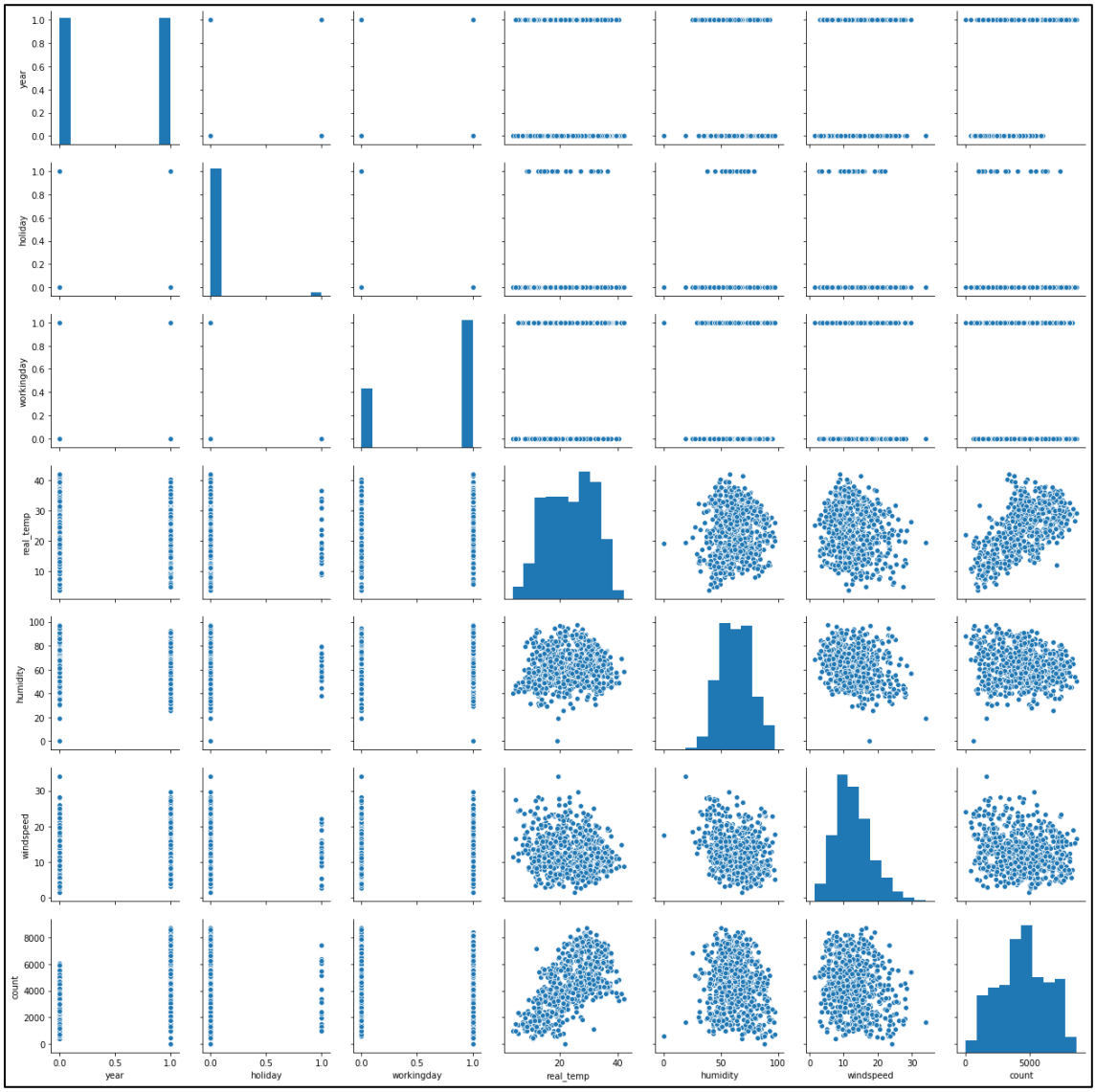
- M = 1, if Married; M = 0, otherwise.
- D = 1, if Divorced; D = 0, otherwise.

In this example, it can be noticed that there is no need to create a dummy variable to represent the "Single" category of marital status. If M equals zero and D equals zero, it can be obvious that person is neither Married nor Divorced. Therefore, person must be Single.

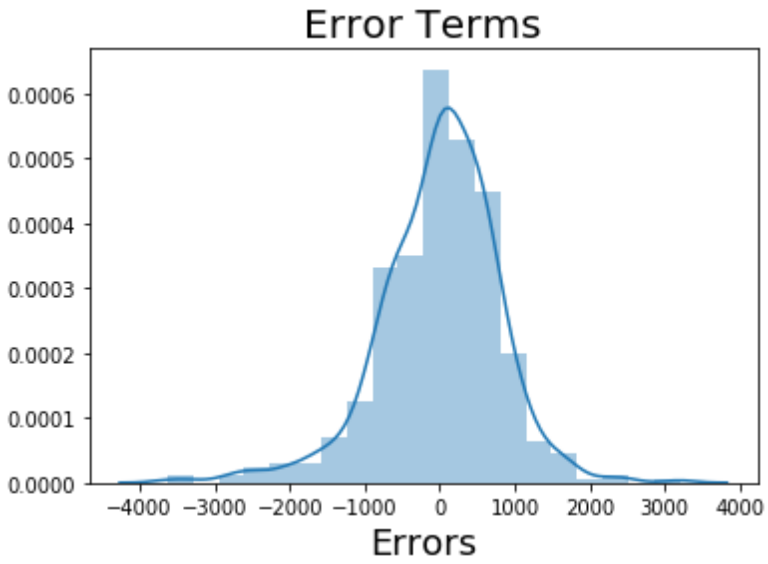
3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

Ans. ‘real_temp’ has the highest correlation with target ‘count’. The graph suggests that ‘count’ increases with increase in ‘real_temp’ in a linear fashion.
Also, the pairplot below descripts the same.





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



Ans. By observing the distribution of residuals: The mean of residuals should follow a normal distribution with mean equal to zero or close to zero. This is done in order to check whether the selected line is actually the line of best fit or not.

If the error terms are non-normally distributed, the confidence intervals may become too wide or narrow. Once the confidence interval becomes unstable, it leads to difficulty in estimating coefficients based on minimisation of least squares. This also suggests that there are a few unusual data points which must be studied closely to make a better model.

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

Ans. Top 3 are

1. 'real_temp': positive relationship
2. 'light_snow' weather: negative relationship
3. year: positive relationship

```
year            2043.571645
holiday         -826.484887
real_temp       4034.912310
windspeed      -1092.018884
spring         -1020.561779
winter           338.691695
july            -590.994742
september        521.978849
light_snow     -2450.461573
mist_cloudy     -710.158488
```

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)

Ans: Linear Regression is a machine learning technique based on supervised learning, that allows us to associate one or more predictor variables with a dependent variable. All machine learning models try to approximate f(x) [the function that accurately describes the relationship between the independent (predictor) and dependent variables], in Linear Regression, it is assumed that f(x) is linear.

A linear algebraic function y is given by y = mx + b, where y is the dependent variable, m is the slope, or derivative, and b, is the intercept, or the value of y when x, the predictor variable is equal to 0.

The goal of linear regression is to obtain a line that best fits the data. In other words, the best suited values for m and b, for which total prediction error (all data points) is as small as possible.

Linear Regression

Simple Linear Regression

It is a model with only one independent variable. It attempts to explain the relationship between a dependent (target) variable y and an independent (predictor) variable x using a straight line given by:

$y = f(x) = \theta_0 + \theta_1 x$

Multiple Linear Regression

It represents the relationship between one dependent variable (target) and several independent variables (predictor variables). The objective of multiple regression is to find a linear equation that can best determine the value of the dependent variable y for different values independent variables in x, and is given as:

$y = f(x_1, x_2, x_{3,\ldots,} x_k) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3 + \theta_4 x_4 + \ldots + \theta_k x_k$

In simple linear regression, the equation for best fitting line is given by:
$\boldsymbol{y_{predicted} = \beta_0 + \beta_1 X}$

Since the aim is to obtain the best values for β0 and β1, the search problem can be converted into a minimization problem where the error between the predicted value and the actual value (or the residual) needs to be minimised. The function to be minimised is called the **Cost Function**.

The most common method for fitting a regression line is the method of least-squares. This method calculates the best-fitting line for the observed data by minimizing the sum of the squares of the residuals (RSS). Because the residuals are first squared, then summed, there are no cancellations between positive and negative values. Small RSS indicates a tight fit of the model to the data.

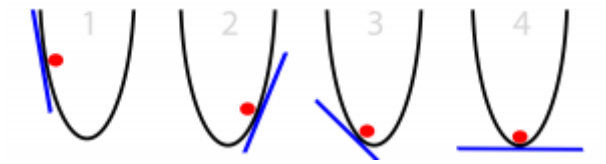
$$RSS = \sum_{i=1}^{n} \left(y_i^{Actual} - y_i^{Predicted} \right)^2 = \sum_{i=1}^{n} \left(y_i - \left(\beta_o + \beta_1 x_i \right) \right)^2$$

This provides the average squared error over all the data points. Therefore, this cost function is also known as the Mean Squared Error (MSE) function.

A regression model uses a technique called **Gradient Descent** to update the coefficients of the line by reducing the cost function. It is done by a random selection of values of coefficients and then iteratively update the values to reach the minimum cost function. Gradient descent helps us on how to update the values.

The derivative of the cost function with respect to each value being updated is calculated and is called the gradient. The sign of the derivative (whether +ve or -ve) decides in which direction next step is to be taken while updating the values in the cost function.

If the sign of gradient/derivative is negative, the value is increased thus, the function will move to the right of the slope (Initially being on the left side of the minima). Similarly, if the sign is positive (i.e. the initial position is on the right of minimal as in the figure) then the value will decrease thus moving towards the minima at the left.



To refer to an analogy, consider a valley which one wants to descend. The best way is to move a step and check for the slope of the valley i.e. whether it is going up or down. Then, proceed to follow the downward slope of the valley, and repeat the same step again and again until the minimum is reached.

Python Code - Simple Linear Regression using Statsmodels

1. Visualize Data

```
>> sns.pairplot(df,x_vars=[columns],y_vars=column,size=4,aspect=1,kind= 'scatter'
)
```

2. Check Collinearity

```
>> sns.heatmap(df.corr(), cmap= "YlGnBu" , annot= True )
```

3. Create Train & Test Data

```
>> from sklearn.model_selection import train_test_split
>> X_train,X_test,y_train,y_test = train_test_split(X, y, train_size=0.7,
test_size=0.3, random_state=100)
```

4. Scaling Features

```
>> from sklearn.preprocessing import StandardScaler, MinMaxScaler
>> scaler = MinMaxScaler()
>> X_train_scaled = scaler.fit_transform(X_train.values.reshape(-1,1))
>> X_test_scaled = scaler.transform(X_test.values.reshape(-1,1))
# In statsmodels intercept variable needs to be added explicitly
>> import statsmodels.api as sm
>> X_train_scaled_sm = sm.add_constant(X_train_scaled)
>> X_test_scaled_sm = sm.add_constant(X_test_scaled)
```

5. Train Model

```
>> model = sm.OLS(y_train, X_train_scaled_sm).fit()
```

6. Analyze Model

```
>> model.params
>> model.summary()
# Statsmodel provides an extensive summary of various metrics
>> plt.scatter(X_train_scaled, y_train)
>> plt.plot(X_train_scaled, coeff_A + coeff_B * X_train_scaled, 'r' )
```

7. Analyze Residuals

```
>> y_train_pred = model.predict(X_train_scaled_sm)
>> residual = (y_train - y_train_pred)
>> sns.distplot(residual, bins = 15) # Checking if residuals are normally
distributed
>> plt.scatter(X_train_scaled, residual) # Checking for independence of residuals
```

8. Predict

```
>> y_pred = model.predict(X_test_scaled_sm)
```

9. Evaluate Model

```
>> from sklearn.metrics import mean_squared_error
>> from sklearn.metrics import r2_score
>> rsme = np.sqrt(mean_squared_error(y_test, y_pred))
>> r_squared = r2_score(y_test, y_pred)
```

10. Visualize Model

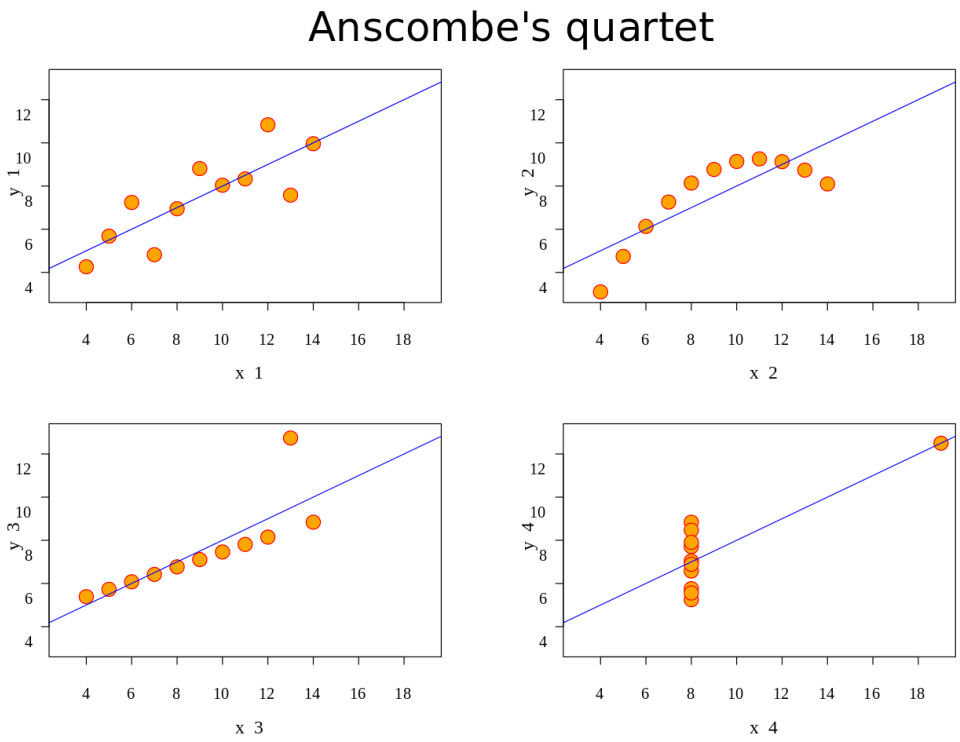
```
>> plt.scatter(X_test_scaled, y_test)
>> plt.plot(X_test_scaled, coeff_A + coeff_B * X_test_scaled, 'r' )
```

2. Explain Anscombe's quartet in detail. (3 marks)

Ans: Anscombe generated a quartet of made-up data in the early 1970's. Each dataset has 11 {x, y} pairs of numbers. The means of the x values are almost identical for all four sets and the means of the y values are also almost identical. The summary statistics of the data is observed to be as follows:

- The average x value is 9 for each dataset
- The average y value is 7.50 for each dataset
- The variance for x is 11 and the variance for y is 4.12
- The correlation between x and y is 0.816 for each dataset
- A linear regression (line of best fit) for each dataset follows the equation y = 0.5x + 3

With just looking at the numbers, it might be concluded that these four datasets are virtually the same. But, by the graphs of the data and the fit to them, the datasets look quite different.



Following explanations can be derived from the above visualizations:

- 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 not linear, and the Pearson correlation coefficient is not relevant.
- In the **third graph** (bottom left), the distribution is linear, but should have a different regression line. The calculated regression is offset by the one outlier which exerts enough influence to lower the correlation coefficient from 1 to 0.816.
- 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.

Anscombe constructed these quartets to demonstrate both the importance of graphing the data before analysing it and the effect of outliers and other influential observations on statistical properties of the data.

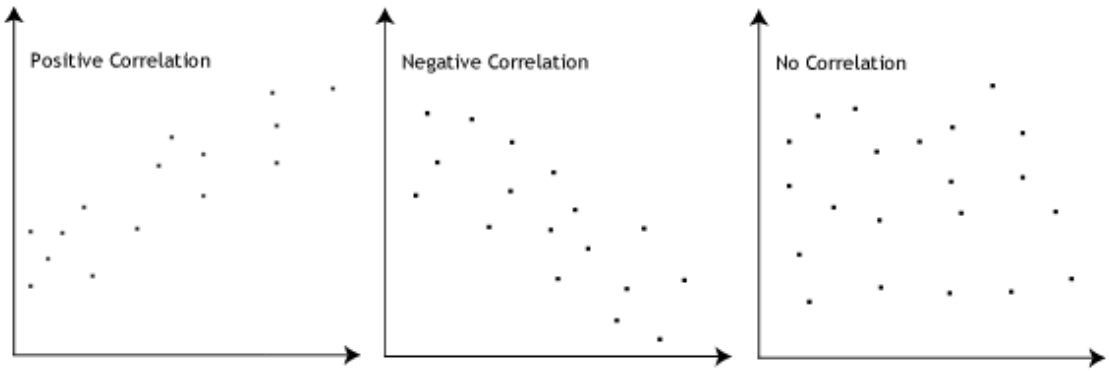
From the graphs the real relationships in the datasets start to emerge. Computing the summary statistics or staring at the data is not just enough and had not given any such insights. Therefore, it is important to visualize the data to get a clear picture of what lies inside the data.

3. What is Pearson's R? (3 marks)

Ans: Correlation is a technique for investigating the relationship between two quantitative, continuous variables. Pearson's correlation coefficient (R) or the Pearson's R is a measure of the strength of the association between the two variables.

It can take a range of values from +1 to -1. A value of 0 indicates that there is no association between the two variables. A value greater than 0 indicates a positive association; that is, as the value of one variable increases, so does the value of the other variable. A value less than 0 indicates a negative association; that is, as the value of one variable increases, the value of the other variable decreases.

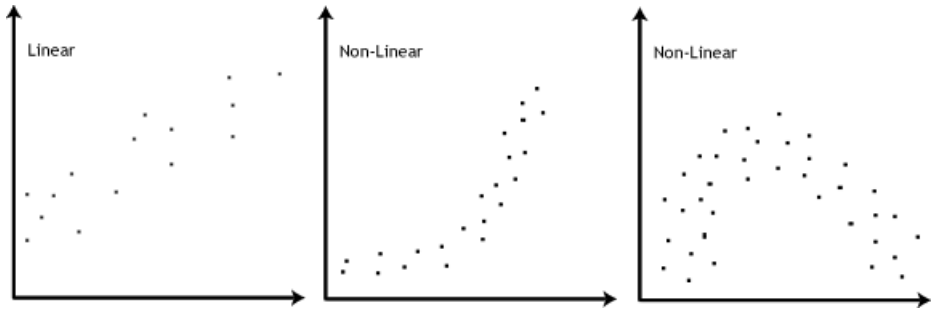
This is shown in the diagram below:



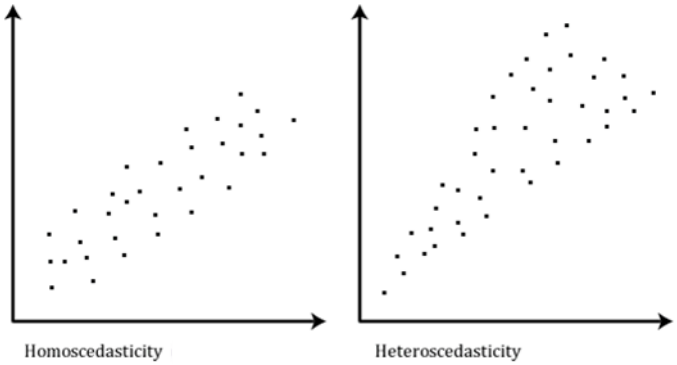
It is important to understand that Pearson's correlation will only give valid results if the data meets the following requirements:

- The two variables should be measured on a continuous scale.
- The two continuous variables should be paired, which means that each case has two values: one for each variable.
- There should be independence of cases, which means that the two observations for one case should be independent of the two observations for any other case.
- There should be a linear relationship between your two continuous variables.

To check whether the two variables form a linear relationship a scatterplot can give a visual inspection of the shape of graph. It would not be appropriate to analyse a non-linear relationship using Pearson's R.



- There should be homoscedasticity, which means that the variances along the line of best fit remain similar as you move along the line. If the variances are not similar, there is heteroscedasticity.



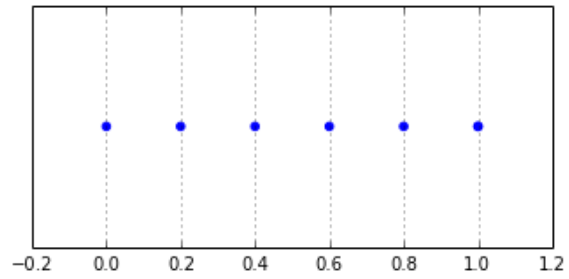
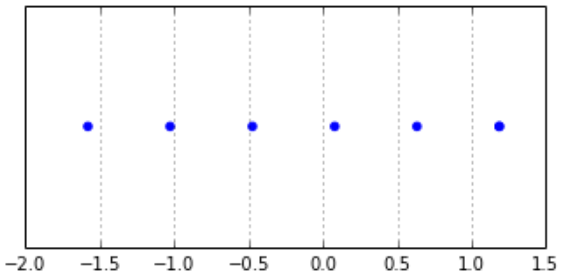
4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

Ans: Scaling is a technique to standardize the predictor variables (independent features) present in the data to a fixed range.

It is performed during the data pre-processing to handle highly varying magnitudes or units. If feature scaling is not done, then a machine learning model tends to weigh greater values, higher and smaller values as the lower values, regardless of the unit of the values.

Normalised Scaling and Standardised Scaling:

Normalized Scaling	Standardized Scaling
This technique is to re-scale features with a distribution value between 0 and 1. For every independent feature, the minimum value of that feature gets transformed into 0, and the maximum value gets transformed into 1.	In this scaling technique, the features are rescaled to ensure the mean and the standard deviation of the data points to be 0 and 1, respectively.
Technique Also known as Min-Max Scaling.	Also known as Z-score normalization.

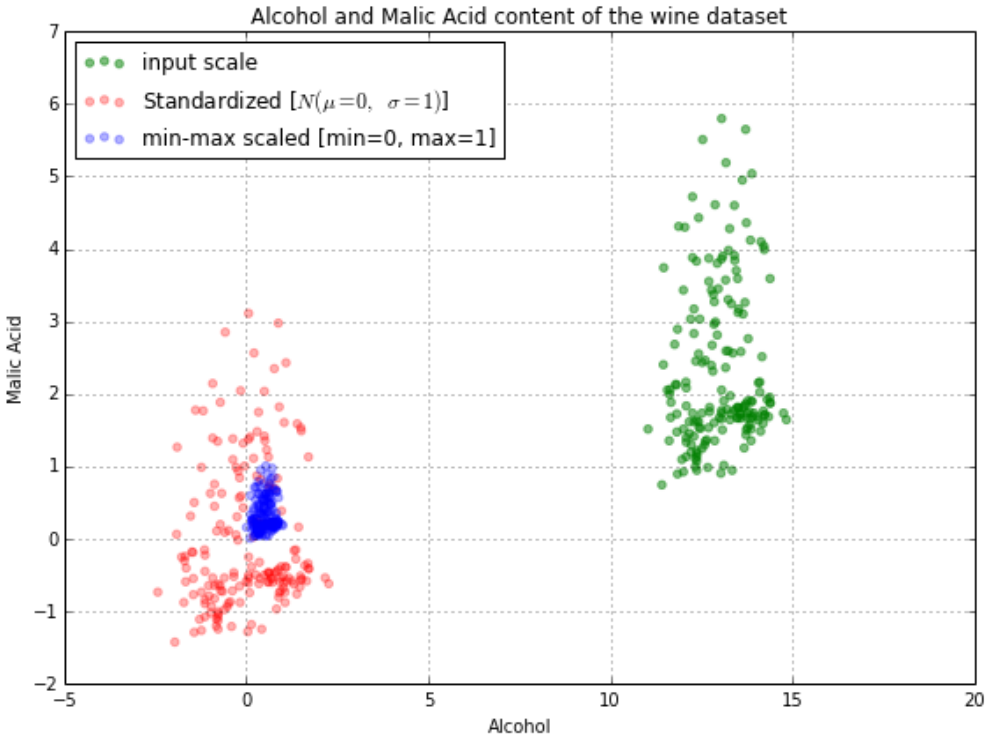
$x_{\text{norm}} = \frac{x - \min(x)}{\max(x) - \min(x)}$	$x_{\text{stand}} = \frac{x - \text{mean}(x)}{\text{standard deviation } (x)}$
	

Example:

	Class label	Alcohol	Malic acid
0	1	14.23	1.71
1	1	13.20	1.78
2	1	13.16	2.36
3	1	14.37	1.95
4	1	13.24	2.59

As it can be observed from the table, the features Alcohol (percent/volume) and Malic acid (g/l) are measured on different scales, so that Feature Scaling is necessary important prior to any comparison or combination of these data.

The plot below includes the wine datapoints on all three different scales: the input scale where the alcohol content was measured in volume-percent (green), the standardized features (red), and the normalized features (blue).



5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

Ans: In regression analysis, variance inflation factor (VIF) is a measure or multicollinearity. Multicollinearity is referred to as the correlation between predictors (independent variables) in a regression model, whose presence can significantly affect the results. The VIF estimates how much the variance of a regression coefficient is inflated due to multicollinearity in the model.

Using the R-squared (R2) value, VIF can be calculated by the formula:

It ranges from 1 and upwards. The numerical value for VIF tells you (in decimal form) what percentage the variance (i.e. the standard error squared) of a regression coefficient is inflated due to multicollinearity in the model.

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

A rule of thumb for interpreting the variance inflation factor:

- 1 = not correlated.
- Between 1 and 5 = moderately correlated.
- Greater than 5 = highly correlated.

The infinite value of VIF shows a perfect correlation between two independent variables. In the case of the perfect correlation we get $R^2 = 1$, which leads $1 / (1 – R^2)$ to infinity. To solve this problem, we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

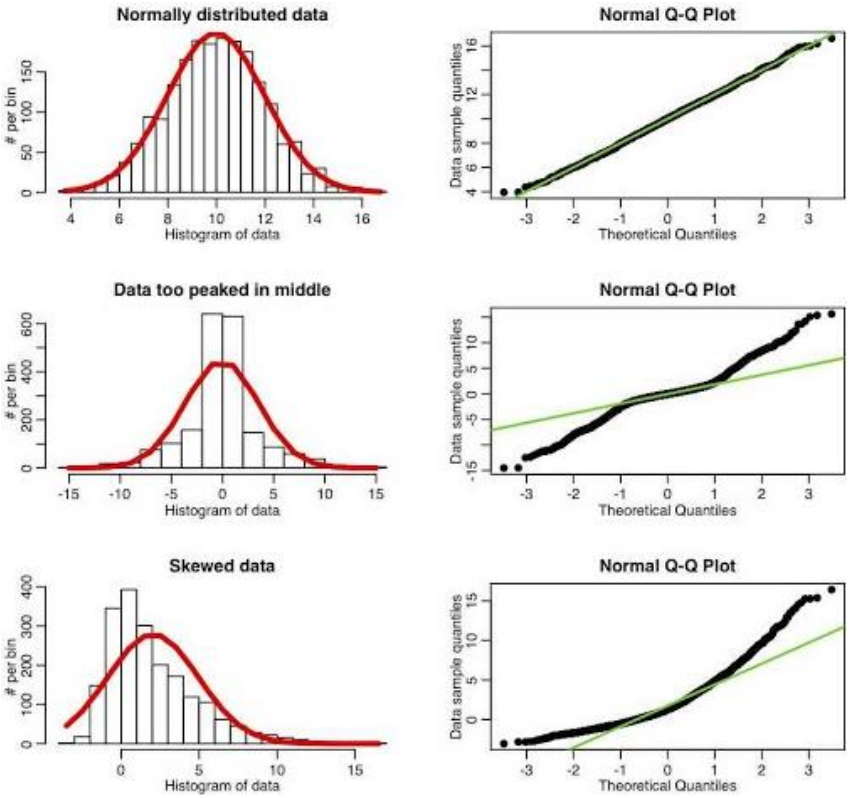
Ans: Q-Q (quantile-quantile) plots are an essential tool to graphically analyse and compare the probability distributions of two datasets by plotting their quantiles against each other. If the two distributions are exactly equal then the points on the Q-Q plot lie perfectly on a straight line (y = x).

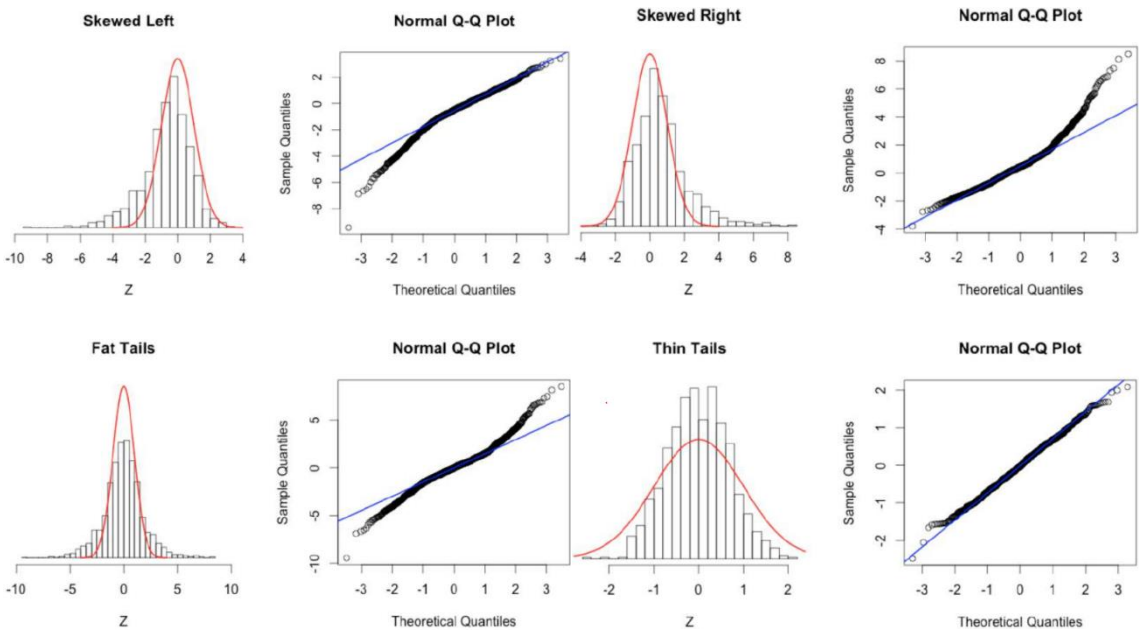
Q-Q plots are helpful in determining the type of distribution for a random variable whether it is a Gaussian Distribution, Uniform Distribution or Exponential Distribution, etc. Also, looking at Q-Q plots can give insights about the skewness and the measure of tailedness of the distribution.

If the bottom end of the Q-Q plot deviates from the straight line and the upper end does not, it can be concluded that the distribution is left-skewed (negatively skewed). Similarly, if the upper end deviates from the straight line but not the lower then it is right-skewed (positively skewed).

The fat-tailed distribution will have both the ends of the Q-Q plot deviating but its central part following the straight line, whereas there will be very less or negligible deviation at the ends for a thin-tailed distribution.

Q-Q plots illustrations:





It should be noted that when the datapoints are fairly less the observations from Q-Q plots fails to conclude any behaviour of the distribution but it can give a significant result when involving a large dataset.

Q-Q plots in Linear Regression: In Linear Regression it serves as a type of diagnostic aid and is used to assess if the residuals are normally distributed.

Simply, a linear regression model is fitted and checked if the points lie approximately on the line, and if they don't, then residuals aren't Gaussian which implies that for small sample sizes, it can't be assumed that β is Gaussian either, so the standard confidence intervals and significance tests are invalid.

If the points do not fall on the line, still some important insights can be learnt from the distribution of the data.

- The slope tells whether the steps in the data are too big or too small (or just right).
- A steeply sloping section of the plot means that in this part of the data, the observations are more spread out than expected to be if they were normally distributed (the cause of this would be an unusually large number of outliers present in the data).
- A flat plot means that the data is more bunched together than expected from a normal distribution.