Assignment-based Subjective Questions

Question 2.

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

Answer. When we convert the categorical variables to dummies, indirectly we are giving importance to each value in a categorical column by making each value as a column and to avoid redundancy we are dropping a column.

drop_first=True is important to use, as it helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables.

For example, if there are Red, Blue and green are categorical variables. When we convert these into dummies Red, blue and green will be the extra columns created. So, if the particular sample has green value, the Red, Blue will be indicated as Zero. So automatically it indicates the sample belongs to green. If the value of green column can be explained by Red and blue column, then there is no need of green column.

Question 3.

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

Answer – atemp has the highest correlation with the target variable. It has the correlation value of 0.63

Question 4.

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

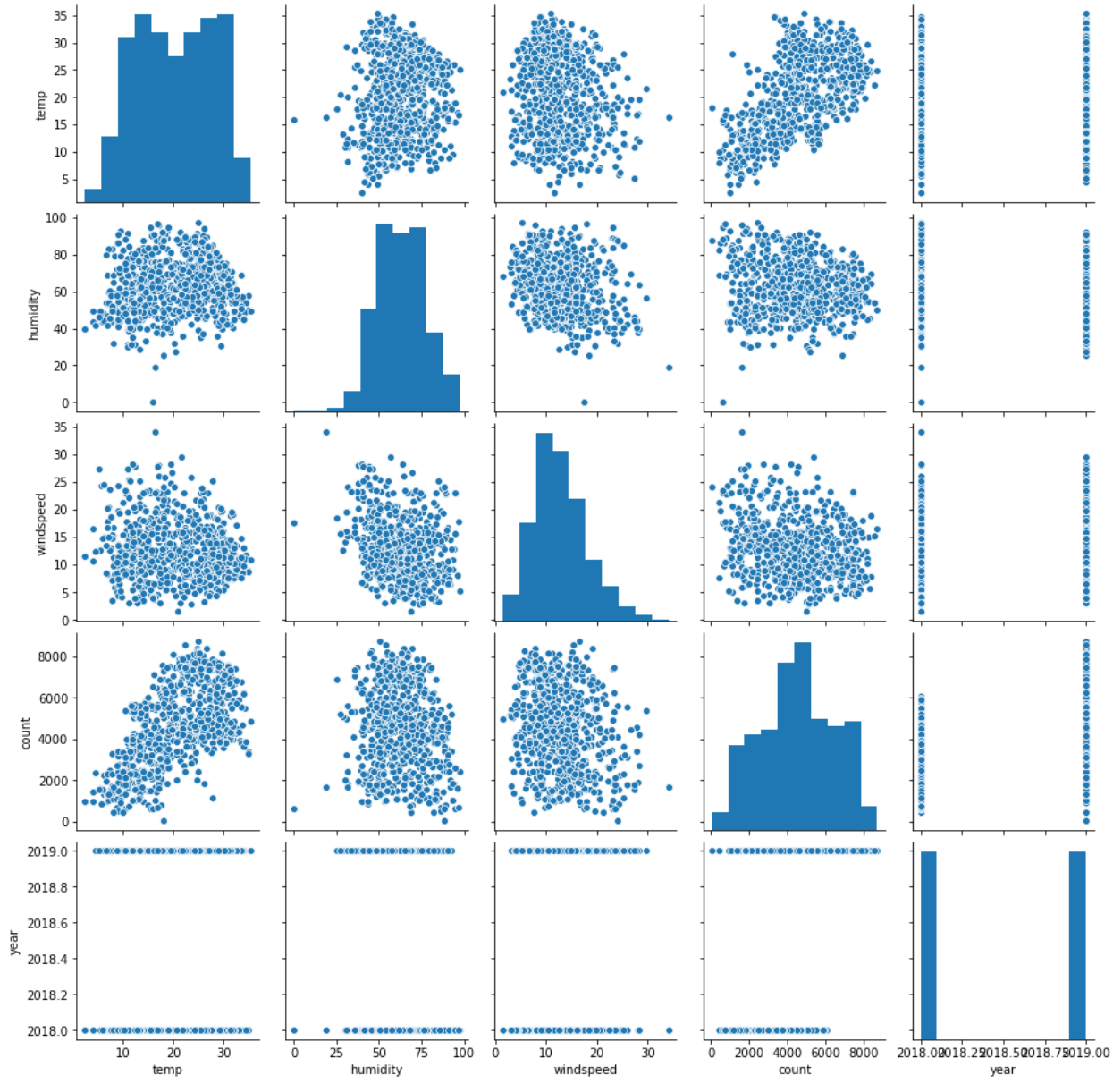
Answer – We can validate the assumptions of Linear Regression with these ways-

- Check Little or no Multicollinearity between the features:

  Multicollinearity occurs when independent variables in a regression model are correlated.
  This correlation is a problem because independent variables should be *independent*. If the degree of correlation between variables is high enough, it can cause problems when we fit the model and interpret the results.

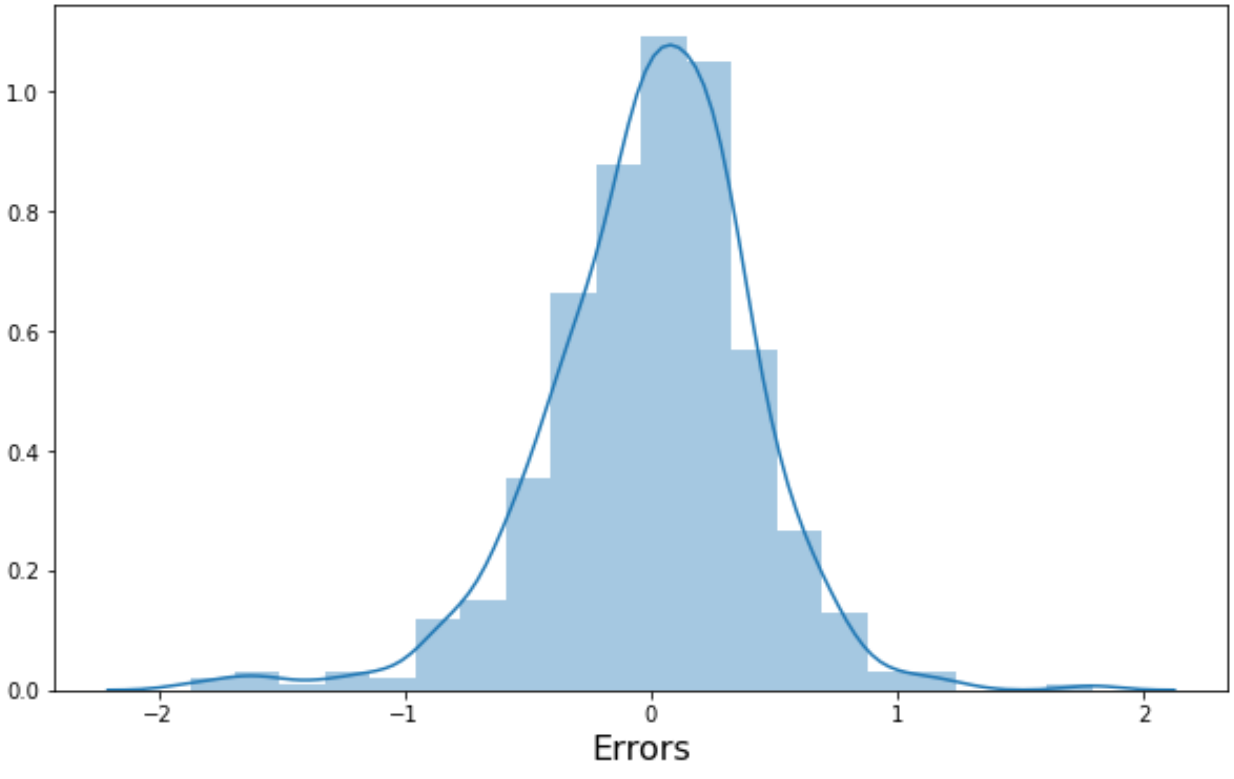
  The idea is that we can change the value of one independent variable and not the others. However, when independent variables are correlated, it indicates that changes in one variable are associated with shifts in another variable. The stronger the correlation, the more difficult it is to change one variable without changing another. It becomes difficult for the model to estimate the relationship between each independent variable and the dependent variable *independently* because the independent variables tend to change in unison.

  In below pairplot we see that there is no or less correlation between the independent variables.



- Normal distribution of error terms:

  In the below distplot we see that the error terms is normally distributed and the peek is at point 0.



Question 5.

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

Answer

top 3 features contributing significantly towards explaining the demand of the shared bikes are -

- Spring season : -0.6842
- Temperature : 0.3999
- Mist : -0.3647

General Subjective Questions

Question 1.

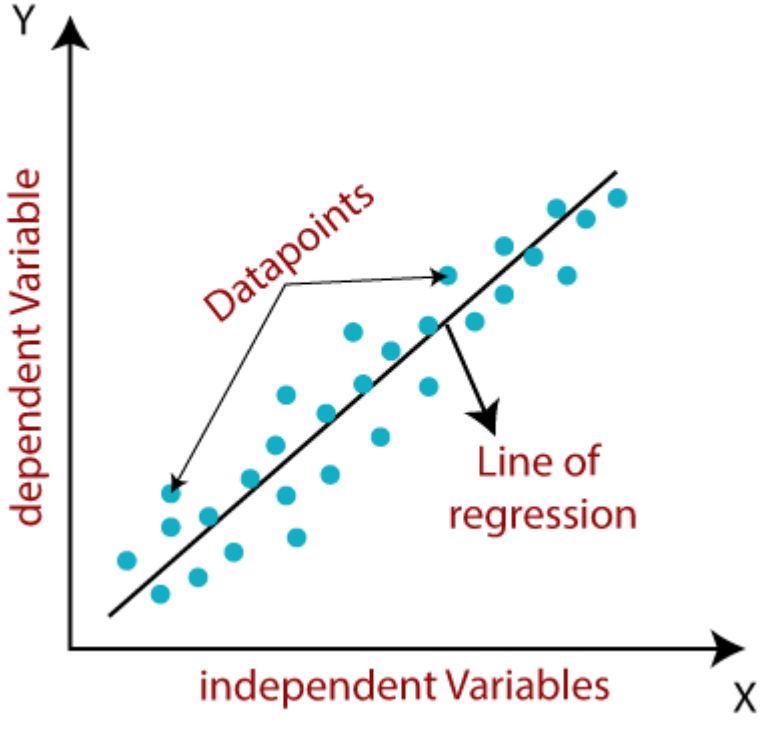
Explain the linear regression algorithm in detail.

Answer –

Linear regression is one of the easiest and most popular Machine Learning algorithms. It is a statistical method that is used for predictive analysis. Linear regression makes predictions for continuous/real or numeric variables such as **sales, salary, age, product price,** etc.

Linear regression algorithm shows a linear relationship between a dependent (y) and one or more independent (y) variables, hence called as linear regression. Since linear regression shows the linear relationship, which means it finds how the value of the dependent variable is changing according to the value of the independent variable.

The linear regression model provides a sloped straight line representing the relationship between the variables.



Mathematically, we can represent a linear regression as:

$$y= a_0+a_1x+ \varepsilon$$

Here,

Y= Dependent Variable (Target Variable)
X= Independent Variable (predictor Variable)

a0= intercept of the line (Gives an additional degree of freedom)
a1 = Linear regression coefficient (scale factor to each input value).
ε = random error

The values for x and y variables are training datasets for Linear Regression model representation.

Types of Linear Regression

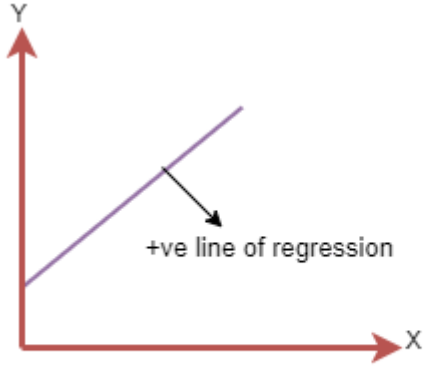
Linear regression can be further divided into two types of the algorithm:

- **Simple Linear Regression:**
  If a single independent variable is used to predict the value of a numerical dependent variable, then such a Linear Regression algorithm is called Simple Linear Regression.
- **Multiple Linear regression:**
  If more than one independent variable is used to predict the value of a numerical dependent variable, then such a Linear Regression algorithm is called Multiple Linear Regression.

Linear Regression Line

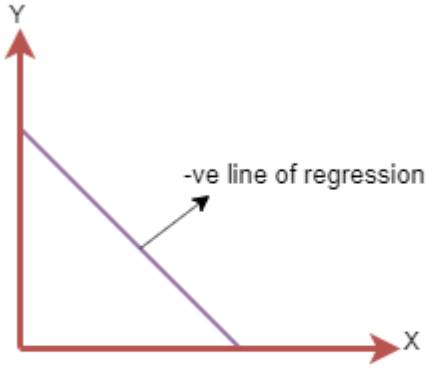
A linear line showing the relationship between the dependent and independent variables is called a **regression line**. A regression line can show two types of relationship:

- **Positive Linear Relationship:**
  If the dependent variable increases on the Y-axis and independent variable increases on X-axis, then such a relationship is termed as a Positive linear relationship.



The line equation will be: $Y= a_0+a_1x$

- **Negative Linear Relationship:**
  If the dependent variable decreases on the Y-axis and independent variable increases on the X-axis, then such a relationship is called a negative linear relationship.



The line of equation will be: $Y= -a_0+a_1x$

Finding the best fit line:

When working with linear regression, our main goal is to find the best fit line that means the error between predicted values and actual values should be minimized. The best fit line will have the least error.

The different values for weights or the coefficient of lines (a_0, a_1) gives a different line of regression, so we need to calculate the best values for a_0 and a_1 to find the best fit line, so to calculate this we use cost function.

Cost function-

- The different values for weights or coefficient of lines (a_0, a_1) gives the different line of regression, and the cost function is used to estimate the values of the coefficient for the best fit line.
- Cost function optimizes the regression coefficients or weights. It measures how a linear regression model is performing.
- We can use the cost function to find the accuracy of the **mapping function**, which maps the input variable to the output variable. This mapping function is also known as **Hypothesis function**.

For Linear Regression, we use the **Mean Squared Error (MSE)** cost function, which is the average of squared error occurred between the predicted values and actual values. It can be written as:

For the above linear equation, MSE can be calculated as:

$$MSE= 1\frac{1}{N}\sum_{i=1}^{n}(y_i - (a_1x_i + a_0))^2$$

Where,

N=Total number of observation
Y_i = Actual value
$(a1x_i+a_0)$= Predicted value.

Residuals: The distance between the actual value and predicted values is called residual. If the observed points are far from the regression line, then the residual will be high, and so cost function will high. If the scatter points are close to the regression line, then the residual will be small and hence the cost function.

Gradient Descent:

- Gradient descent is used to minimize the MSE by calculating the gradient of the cost function.
- A regression model uses gradient descent to update the coefficients of the line by reducing the cost function.
- It is done by a random selection of values of coefficient and then iteratively update the values to reach the minimum cost function.

Model Performance:

The Goodness of fit determines how the line of regression fits the set of observations. The process of finding the best model out of various models is called **optimization**. It can be achieved by below method:

R-squared method:

- R-squared is a statistical method that determines the goodness of fit.
- It measures the strength of the relationship between the dependent and independent variables on a scale of 0-100%.
- The high value of R-square determines the less difference between the predicted values and actual values and hence represents a good model.
- It is also called a **coefficient of determination,** or **coefficient of multiple determination** for multiple regression.
- It can be calculated from the below formula:

$$\text{R-squared} = \frac{\text{Explained variation}}{\text{Total Variation}}$$

Assumptions of Linear Regression

Below are some important assumptions of Linear Regression. These are some formal checks while building a Linear Regression model, which ensures to get the best possible result from the given dataset.

- **Linear relationship between the features and target:**
  Linear regression assumes the linear relationship between the dependent and independent variables.
- **Small or no multicollinearity between the features:**
  Multicollinearity means high-correlation between the independent variables. Due to multicollinearity, it may difficult to find the true relationship between the predictors and target variables. Or we can say, it is difficult to determine which predictor variable is affecting the target variable and which is not. So, the model assumes either little or no multicollinearity between the features or independent variables.

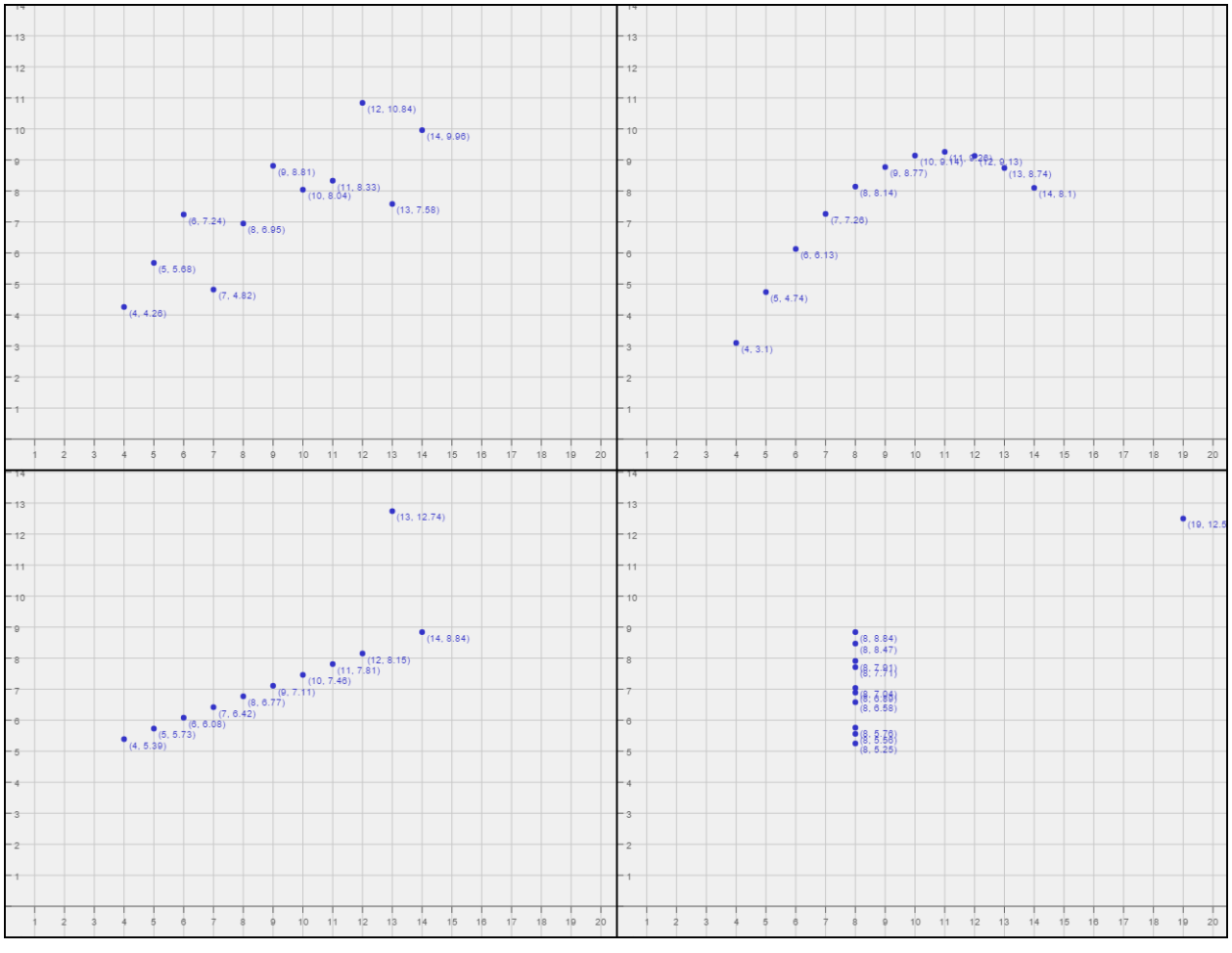
- Homoscedasticity Assumption:
  Homoscedasticity is a situation when the error term is the same for all the values of independent variables. With homoscedasticity, there should be no clear pattern distribution of data in the scatter plot.
- Normal distribution of error terms:
  Linear regression assumes that the error term should follow the normal distribution pattern. If error terms are not normally distributed, then confidence intervals will become either too wide or too narrow, which may cause difficulties in finding coefficients.
  It can be checked using the q-q plot. If the plot shows a straight line without any deviation, which means the error is normally distributed.
- No autocorrelations:
  The linear regression model assumes no autocorrelation in error terms. If there will be any correlation in the error term, then it will drastically reduce the accuracy of the model. Autocorrelation usually occurs if there is a dependency between residual errors.

Question 2.

Explain the Anscombe's quartet in detail.

Answer

Statistics have long been used to describe data in general terms, Things like variance and standard deviation allow us to understand how much variation there was in some data without having to look at every data point individually. They give us a rough idea as to how consistent data is. However, knowing variance alone does not give us the full picture as to what the data truly is in its native form.
Statistics are great for describing general trends and aspects of data, but statistics alone can't fully depict any data set. Francis Anscombe realized this in 1973 and created several data sets, all with several identical statistical properties, to illustrate it. These data sets, collectively known as "Anscombe's Quartet," are shown below.
Graphs generated by meta-calculator.com
All four of these data sets have the same variance in x, variance in y, mean of x, mean of y, and linear regression. But, as we can clearly tell, they are all quite different from one another. So, what does this mean to we as a statistician?
Well, to start, Anscombe's Quartet is a great demonstration of the importance of graphing data to analyze it. Given simply variance values, means, and even linear regressions can not accurately portray data in its native form. Anscombe's Quartet shows that multiple data sets with many similar statistical properties can still be vastly different from one another when graphed.
Additionally, Anscombe's Quartet warns of the dangers of outliers in data sets. Think about it: if the bottom two graphs didn't have that one point that strayed so far from all the other points, their statistical properties would no longer be identical to the two top graphs. In fact, their statistical properties would more accurately resemble the lines that the graphs seem to depict.
how to analyze our data. For example, while all four data sets have the same linear regression, it is obvious that the top right graph really shouldn't be analyzed with a linear regression at all because it's a curvature. Conversely, the top left graph probably *should* be analyzed with a linear regression because it's a scatter plot that moves in a roughly linear manner. These observations demonstrate the value in graphing our data before analyzing it.
Anscombe's Quartet reminds us that graphing data prior to analysis is good practice, outliers should be removed when analyzing data, and statistics about a data set do not fully depict the data set in its entirety.



Question 3.

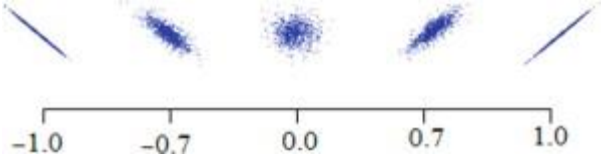
What is Pearson's R?

Answer

Correlation

The Pearson correlation method is the most common method to use for numerical variables; it assigns a value between − 1 and 1, where 0 is no correlation, 1 is total positive correlation, and − 1 is total negative correlation. This is interpreted as follows: a correlation value of 0.7 between two variables would indicate that a significant and positive relationship exists between the two. A positive correlation signifies that if variable A goes up, then B will also go up, whereas if the value of the correlation is negative, then if A increases, B decreases.
For further reading on the Pearson Correlation Method, see:

Considering the two variables "age" and "salary," a strong positive correlation between the two would be expected: as people get older, they tend to earn more money. Therefore, the correlation between age and salary probably gives a value over 0.7. Figure 6.2 illustrates pairs of numerical variables plotted against each other, with the corresponding correlation value between the two variables shown on the x-axis. The right-most plot shows a perfect positive correlation of 1.0, whereas the middle plot shows two variables that have no correlation whatsoever between them. The left-most plot shows a perfect negative correlation of − 1.0.



A correlation can be calculated between two numerical values (e.g., age and salary) or between two category values (e.g., type of product and profession). However, a company may also want to calculate correlations between variables of different types. One method to calculate the correlation of a numerical variable with a categorical one is to convert the numerical variable into categories. For example, age would be categorized into ranges (or buckets) such as: 18 to 30, 31 to 40, and so on.

As well as the correlation, the covariance of two variables is often calculated. In contrast with the correlation value, which must be between − 1 and 1, the covariance may assume any numerical value. The covariance indicates the grade of synchronization of the variance (or volatility) of the two variables.

two variables that have the highest correlations are profession with income (US $), with a correlation of 0.85, and age with income (US $), with a correlation of 0.81. The lowest correlations are cell phone usage with income (0.25) and cell phone usage with profession (0.28). Hence the initial conclusion is that cell phone usage doesn't have a high correlation with any other variable, so it could be considered for exclusion from the input variable set. Table 6.1 also shows that cell phone usage has a significantly lower reliability (0.3) than the other variables and this could have repercussions on its correlation value with the remaining variables. Also, profession only has a high correlation with income; however, it will be seen that this correlation pair (income, profession) is important to the type of business. Given that each variable has a correlation with every other variable, the values are repeated around the diagonal. Therefore, the values on one side of the diagonal can be omitted. Note that all the values are equal to 1 on the diagonal, because these are the correlations of the variables with themselves.

Question 4.

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

Answer-

Scaling

In scaling (also called *min-max scaling*), we transform the data such that the features are within a specific range e.g. [0, 1].

$x' = x - x_{min} x_{max} - x_{min} x' = x - x_{min} x_{max} - x_{min}$

where x' is the normalized value.

Scaling is important in the algorithms such as support vector machines (SVM) and k-nearest neighbors (KNN) where distance between the data points is important.

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.
Feature scaling is a method used to standardize the range of independent variables or features of data. In data processing, it is also known as data normalization and is generally performed during the data preprocessing step.
Since the range of values of raw data varies widely, in some machine learning algorithms, objective functions will not work properly without normalization. For example, the majority of classifiers calculate the distance between two points

by the distance. If one of the features has a broad range of values, the distance will be governed by this particular feature. Therefore, the range of all features should be normalized so that each feature contributes approximately proportionately to the final distance.
Another reason why feature scaling is applied is that gradient descent converges much faster with feature scaling than without it.

Standardization:

Standardizing the features around the center and 0 with a standard deviation of 1 is important when we compare measurements that have different units. Variables that are measured at different scales do not contribute equally to the analysis and might end up creating a bais.
For example, A variable that ranges between 0 and 1000 will outweigh a variable that ranges between 0 and 1. Using these variables without standardization will give the variable with the larger range weight of 1000 in the analysis. Transforming the data to comparable scales can prevent this problem. Typical data standardization procedures equalize the range and/or data variability.
As we discussed earlier, standardization (or Z-score normalization) means centering the variable at zero and standardizing the variance at 1. The procedure involves subtracting the mean of each observation and then dividing by the standard deviation:
The result of standardization is that the features will be rescaled so that they'll have the properties of a standard normal distribution with

$\mu=0$ and $\sigma=1$

where μ is the mean (average) and σ is the standard deviation from the mean.

Normalization:

Similarly, the goal of normalization is to change the values of numeric columns in the dataset to a common scale, without distorting differences in the ranges of values. For machine learning, every dataset does not require normalization. It is required only when features have different ranges.
For example, consider a data set containing two features, age, and income(x2). Where age ranges from 0–100, while income ranges from 0–100,000 and higher. Income is about 1,000 times larger than age. So, these two features are in very different ranges. When we do further analysis, like multivariate linear regression, for example, the attributed income will intrinsically influence the result more due to its larger value. But this doesn't necessarily mean it is more important as a predictor. So we normalize the data to bring all the variables to the same range.
In this approach, the data is scaled to a fixed range — usually 0 to 1.
In contrast to standardization, the cost of having this bounded range is that we will end up with smaller standard deviations, which can suppress the effect of outliers. Thus MinMax Scalar is sensitive to outliers.
A Min-Max scaling is typically done via the following equation:

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

Question 5.

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

Answer - VIF is an index that provides a measure of how much the variance of an estimated regression coefficient increases due to collinearity. In order to determine VIF, we fit a regression model between the independent variables.
If all the independent variables are orthogonal to each other, then VIF = 1.0. If there is perfect correlation, then VIF = infinity. A large value of VIF indicates that there is a correlation between the variables.
This happens when there is perfect correlation between two independent variables. When we get R2 =1, then it lead to 1/(1-R2) infinity.

If there is perfect correlation, then VIF = infinity. A large value of VIF indicates that there is a correlation between the variables. If the VIF is 4, this means that the variance of the model coefficient is inflated by a factor of 4 due to the presence of multicollinearity.
What does infinite VIF mean?
The user has to select the variables to be included by ticking off the corresponding check boxes. An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables (which show an infinite VIF as well).

Question 6.

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

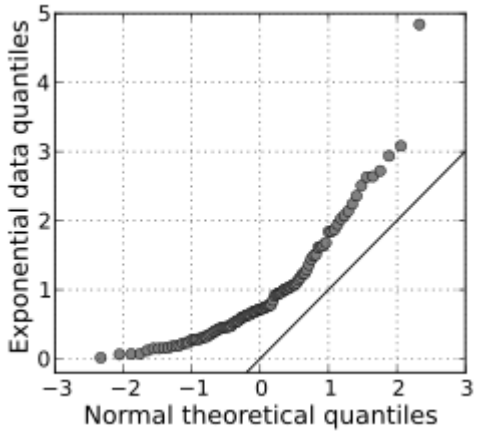
Answer. The quantile-quantile (q-q) plot is a graphical technique for determining if two data sets come from populations with a common distribution.
A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set. By a quantile, we mean the fraction (or percent) of points below the given value. That is, the 0.3 (or 30%) quantile is the point at which 30% percent of the data fall below and 70% fall above that value.
A 45-degree reference line is also plotted. If the two sets come from a population with the same distribution, the points should fall approximately along this reference line. The greater the departure from this reference line, the greater the evidence for the conclusion that the two data sets have come from populations with different distributions.

The quantile-quantile (q-q) plot is a graphical technique for determining if two data sets come from populations with a common distribution.
A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set. By a quantile, we mean the fraction (or percent) of points below the given value. That is, the 0.3 (or 30%) quantile is the point at which 30% percent of the data fall below and 70% fall above that value.
A 45-degree reference line is also plotted. If the two sets come from a population with the same distribution, the points should fall approximately along this reference line. The greater the departure from this reference line, the greater the evidence for the conclusion that the two data sets have come from populations with different distributions.

The advantages of the q-q plot are:

The sample sizes do not need to be equal.
Many distributional aspects can be simultaneously tested. For example, shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot. For example, if the two data sets come

from populations whose distributions differ only by a shift in location, the points should lie along a straight line that is displaced either up or down from the 45-degree reference line.
The q-q plot is similar to a probability plot. For a probability plot, the quantiles for one of the data samples are replaced with the quantiles of a theoretical distribution.



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

The points plotted in a Q–Q plot are always non-decreasing when viewed from left to right. If the two distributions being compared are identical, the Q–Q plot follows the 45° line $y = x$. If the two distributions agree after linearly transforming the values in one of the distributions, then the Q–Q plot follows some line, but not necessarily the line $y = x$. If the general trend of the Q–Q plot is flatter than the line $y = x$, the distribution plotted on the horizontal axis is more dispersed than the distribution plotted on the vertical axis. Conversely, if the general trend of the Q–Q plot is steeper than the line y = x, the distribution plotted on the vertical axis is more dispersed than the distribution plotted on the horizontal axis. Q–Q plots are often arced, or "S" shaped, indicating that one of the distributions is more skewed than the other, or that one of the distributions has heavier tails than the other.

Although a Q–Q plot is based on quantiles, in a standard Q–Q plot it is not possible to determine which point in the Q–Q plot determines a given quantile. For example, it is not possible to determine the median of either of the two distributions being compared by inspecting the Q–Q plot. Some Q–Q plots indicate the deciles to make determinations such as this possible.

The intercept and slope of a linear regression between the quantiles gives a measure of the relative location and relative scale of the samples. If the median of the distribution plotted on the horizontal axis is 0, the intercept of a regression line is a measure of location, and the slope is a measure of scale. The distance between medians is another measure of relative location reflected in a Q–Q plot. The "probability plot correlation coefficient" (PPCC plot) is the correlation coefficient between the paired sample quantiles. The closer the correlation coefficient is to one, the closer the distributions are to being shifted, scaled versions of each other. For distributions with a single shape parameter, the probability plot correlation coefficient plot provides a method for estimating the shape parameter – one simply computes the correlation coefficient for different values of the shape parameter, and uses the one with the best fit, just as if one were comparing distributions of different types.

Another common use of Q–Q plots is to compare the distribution of a sample to a theoretical distribution, such as the standard normal distribution N(0,1), as in a normal probability plot. As in the case when comparing two samples of data, one orders the data (formally, computes the order statistics), then plots them against certain quantiles of the theoretical distribution.[3]