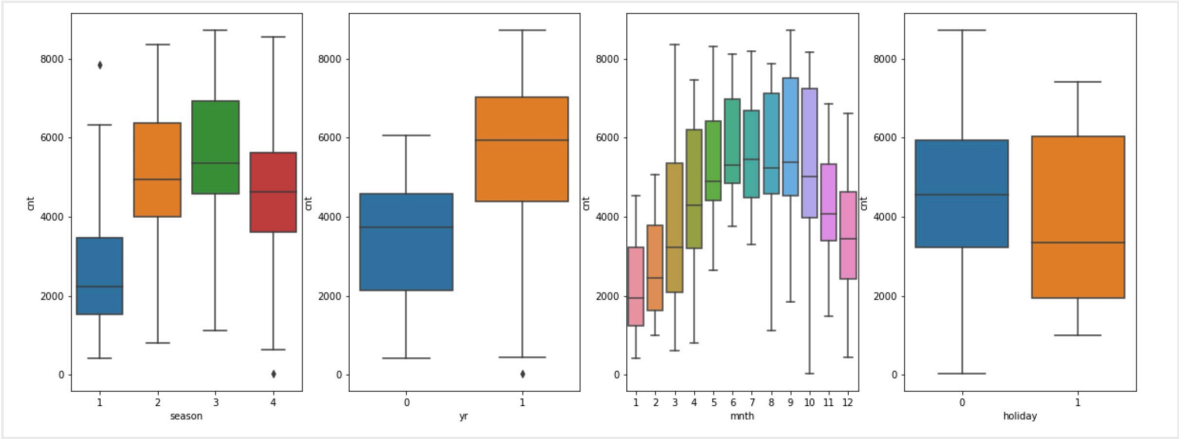
Assignment-based Subjective Questions

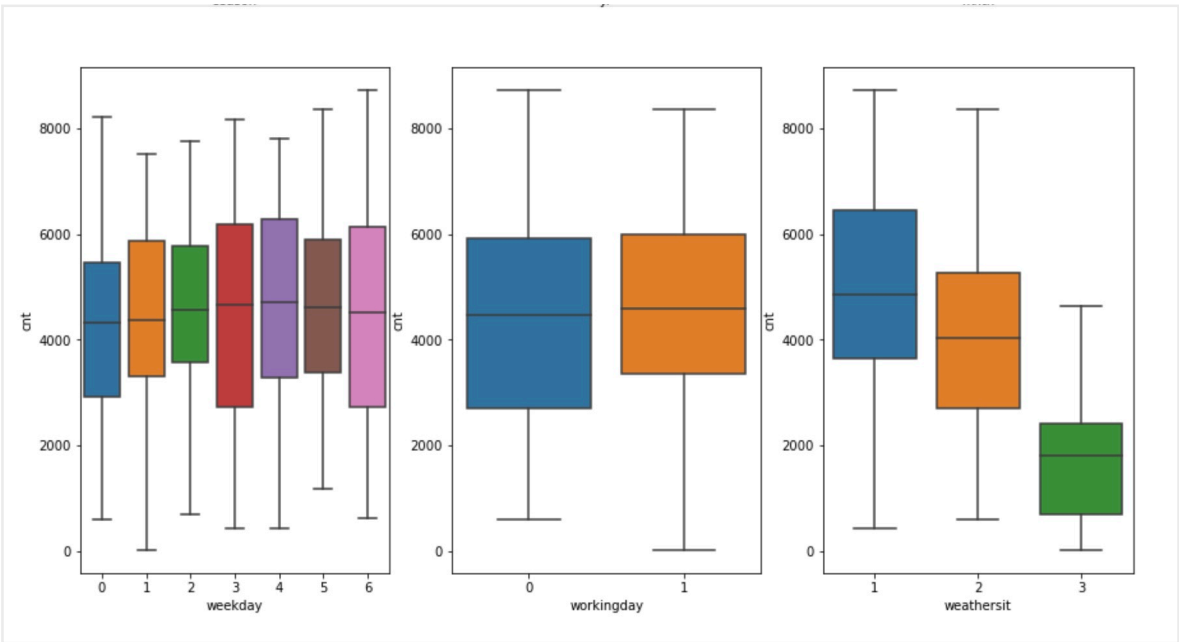
1.From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

- We have 7 main categorical variables in the dataset, namely:
  - season
  - yr
  - mnth
  - holiday
  - weekday
  - workingday
  - weathersit

Boxplot was used to find their

relationship with the dependent variable , ‘cnt’.





Following insights were obtained:

1. Overall , all the categorical variables did show a good relationship with the

dependent variable. Though for some, this relationship was stronger.

2. In terms of the medians of the boxplots, maximum bookings occurred in season 3(32%), followed by season 2(28%), season 4(26%) and least in season 1(14%).
3. Months 5, 6, 7, 8, 9 and 10 showed higher booking medians over other months.
4. With 69% bookings, Weathersit 1 is most favourable for bookings followed by Weathersit 2 (30%) and lastly Weathersit 3.
5. 98% bookings took place when it was not a holiday. Hence holiday has a very weak relationship with the dependent variable.
6. Weekday bookings don't show much of difference in terms of percentages. However we cannot say that this relationship is weak. Let's discover this further in the model.
7. Workingday : 70% bookings are

taking place on a working day. This shows a strong relationship with the dependent variable ‘cnt’.

8. 62% bookings took place in 2019. This shows a strong relationship with the dependent variable ‘cnt’.

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

This is important to avoid the dummy variable trap, that is the variables become highly correlated.
Simply, we need ‘k-1’ levels of dummy variable to explain a variable with ‘k-levels’. Hence we use drop_first =True during dummy variable creation.

Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?
We can see that temp and atemp have

the highest correlation with the target variable.

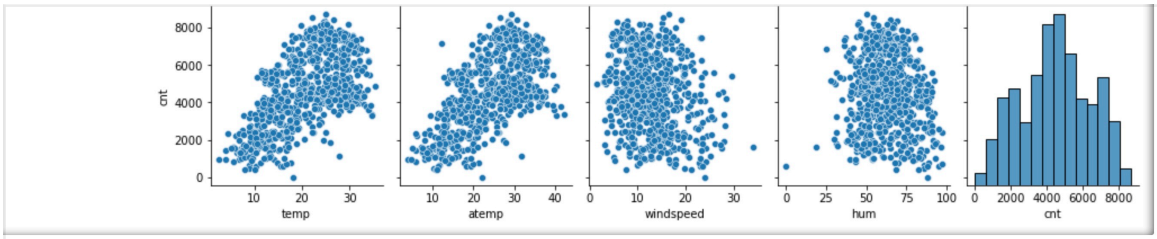
3.Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

Both temp and atemp showed very similar correlation with the target variable. And this correlation was the highest.

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

1. **Linearity**: Made a pair-plot with the variables in the model. The plot

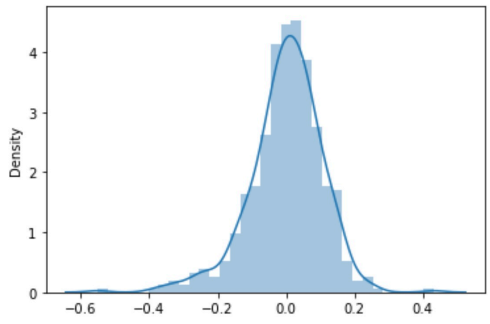
satisfied the condition of linearity between the independent and the dependent variables.



There seems to be some linear relationship between the dependent variable 'cnt' and the independent variables. So the assumption of linearity also stands validated.

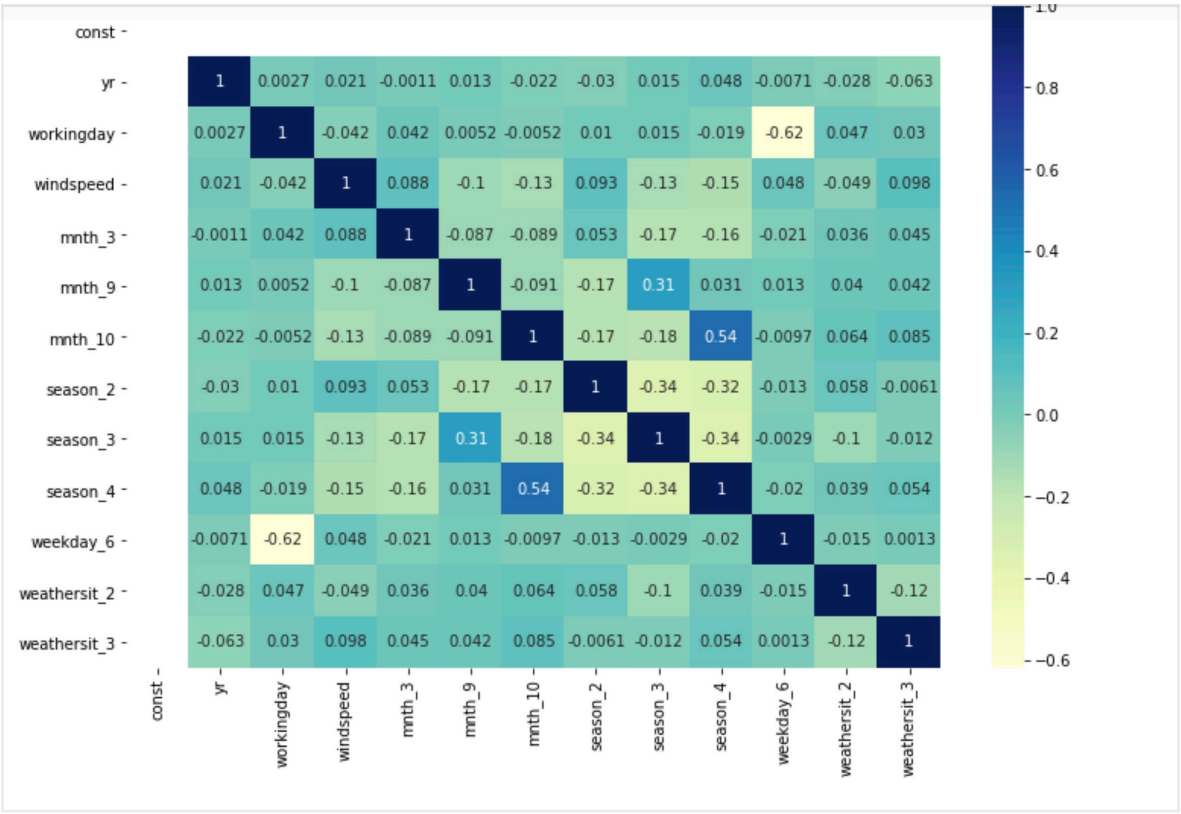
2. **Normality of Error Terms:** We did a residual analysis and found out that the error terms were normally distributed.

```
In [330]: ## plotting residuals
          sns.distplot(res_train)
          plt.show()
```



The residuals seem to be showing more or less a normal distribution. Hence, the assumption of normality of error terms stand validated.

3. **Little or No Multicollinearity among Predictors:** Plotted heat map and also saw VIF scores. Heatmap showed no or little co-relations among the predictors ,while the VIF score for all the predictors I our model were below 5.



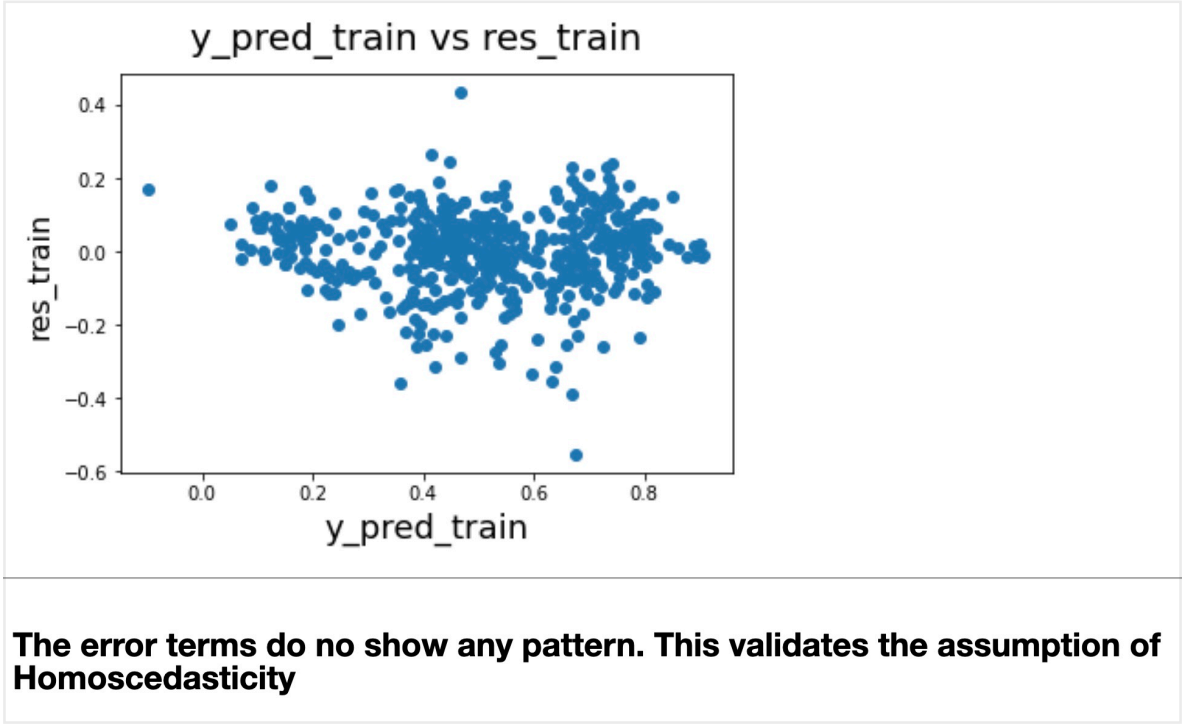
	Features	VIF
0	const	17.91
9	season_4	2.16
8	season_3	1.86
2	workingday	1.63
10	weekday_6	1.63
7	season_2	1.54
6	mnth_10	1.46
5	mnth_9	1.17
4	mnth_3	1.13
3	windspeed	1.09
12	weathersit_3	1.05
11	weathersit_2	1.04
1	yr	1.01

4. **No Autocorrelation of the Error Terms :** Checked the Durbin-Watson value from the model statistics. The Durbin-Watson value of 1.978 of the model is almost equal to 2, which indicates no autocorelation of the error terms. Even if we don't round-off this value, still 1.978 lies within the acceptable range of 1.5 to 2.5 for Durbin-Watson statistic.

Omnibus:	69.993	**Durbin-Watson:**	1.978
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	164.601
Skew:	-0.723	**Prob(JB):**	1.81e-36
Kurtosis:	5.378	**Cond. No.**	11.3

5. **Homoscedasticity : Same variance within the error terms:** Made a scatter plot between predicted values from the trainset and the residuals from the difference between the

actual values and predicted values. The residuals did not show any patterns of distribution. This validated the assumption of Homoscedasticity.



The error terms do no show any pattern. This validates the assumption of Homoscedasticity

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

1. Weathersit_3
2. Season_3

3. Season_2

General Subjective Questions

1.Explain the linear regression algorithm in detail.

Lets first look into the word linear-regression.
Regression stands for a method by which we find out the value of a dependent variable(y) based on the values of independent variable/s (X).
‘Linear’ , represents the fact that this relationship between the dependent variable(y) and the independent variable/s (X) follows a linear trend.
Now this ‘linear-relationship ‘ can be simple, characterised by the equation

$$y = m*x + c$$

Or, it can be multiple, characterised by the equation

y = c+ m1*x 1 + m2*x2 + m3*x 3+ m4*x4 +.....(and so on)

The algorithm for linear regression is based on finding the ‘best-fitted’ line, running through the dataset.
This line can predict the values of the dependent variable based on the value/s of the independent variable/s.

Through this model, we first train our machine to learn to predict the dependent variables based on the dataset that we feed to it. This dataset is known as the training set.
Based on its learning of the Training set, the model predicts values for the dependent variable.

How the model finds the ‘best-fitted’ line?
It calculates the sum of squares of the distance between the predicted values of the dependent variable and the actual

values of the dependent variable present in the dataset.
There can be n number such lines.
So how does the model choose one line?
It does do through minimising what is known as the ‘cost-function’.
Mathematically it is represented as ;

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

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

We also need to control the learning rate, that is the rate at which the machine is learning. Too fast a learning rate can lead to over-simplification and

missing out on data, while too slow a learning rate can lead to inclusion. Of unnecessary data and over-complicated model. So we need an optimal rate of learning which is indicated by what is known as the gradient descent .

Now the model can take various approaches like:
The model performance and the effect of the variables is judged on the parameters like p-value, R-sqaured, Adjusted R-sqaured, VIF, F-Statistic etc.

1. Forward selection: Here we start with one independent variable to predict the dependent variable. We then keep on adding more independent variables (generally 1 at a time), to see how the inclusion of the variable improves or deteriorates our model.
2. Backward-elimination : Here we start with all the independent variable/s and then we start eliminating one variable

at a time to get to an optimal model.
3. Automated(Recursive Function Elimination): We use the in-built function and libraries to build a model.
4. Hybrid: We first begin with Automated to make a coarse model and then fine-tune it manually using Forward selection and Backward-elimination to get to an optimal model.

2.Explain the Anscombe's quartet in detail.

Anscombe's quartet is a very simple yet effective representation to illustrate the pitfalls of blindly relying on summary-statistics while analysing data. It also highlights the importance of data-visualisation while performing data-analysis.

Anscombe's quartet comprises four datasets that have nearly identical

simple statistical properties, yet 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 on statistical properties.(source: Wikipedia)

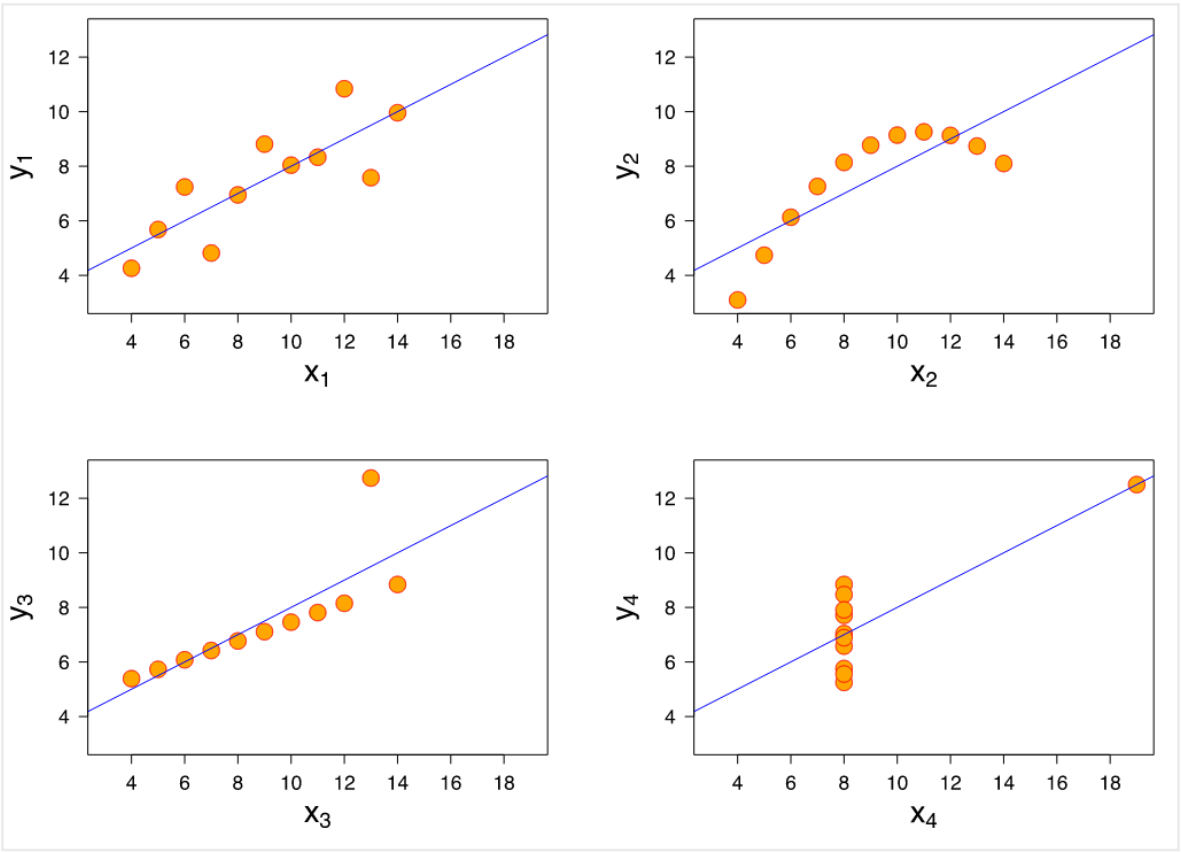
The dataset looks something like this:

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

For this dataset, all the summary statistics are close to identical:

- 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 $y = 0.5x + 3$

But when we plot these four data sets on an x/y coordinate plane, we get the following results:



This visualisation represents the real relationships among the datasets

Dataset I consists of a set of points that appear to follow a rough linear relationship with some variance.
Dataset II fits a neat curve but doesn't follow a linear relationship .
Dataset III looks like a tight linear relationship between *x* and *y*, except for one large outlier.
Dataset IV looks like *x* remains constant, except for one outlier as well.

Computing summary statistics or staring at the data wouldn't have told us any of these stories. Instead, it's important to visualize the data to get a clear picture of what's going on.

3.What is Pearson's R?

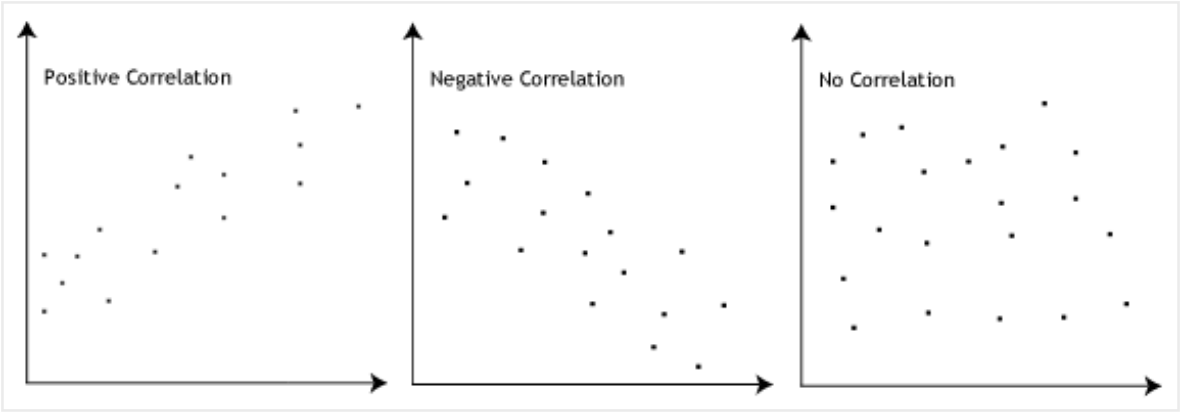
Pearson's R is one of the most popular type of correlation coefficient.
Also called Pearson's coefficient, it is commonly used to analyse linear regression models.
First let us understand what is a correlation coefficient.
Correlation coefficients are a measure of strength of relationship is between two

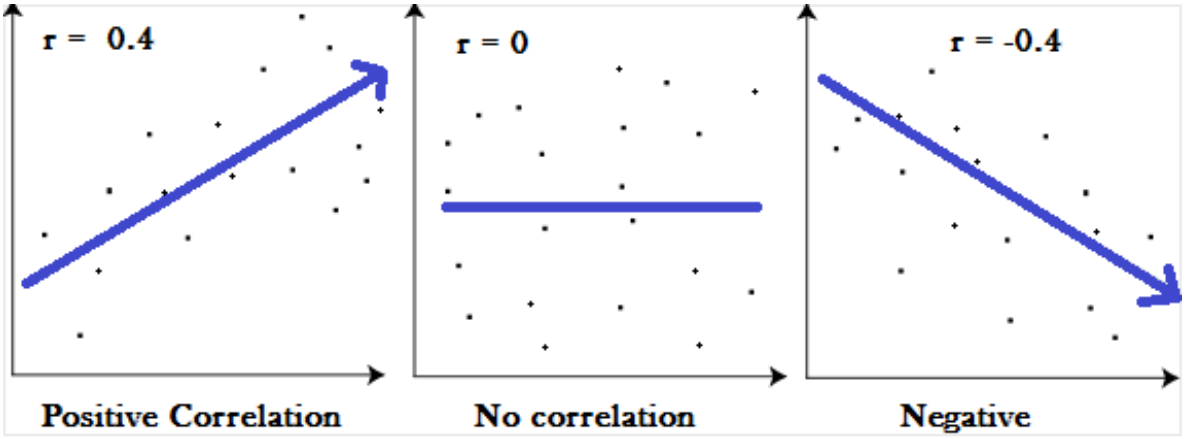
variables. Normally we find the strength for the relationship between the independent variable/s(X) and dependent variable(y)

Basically, a Pearson correlation attempts to draw a line of best fit through the data of two variables, and the Pearson correlation coefficient, *r*, indicates how far away all these data points are to this line of best fit (i.e., how well the data points fit this new model/line of best fit).

The value of Pearson correlation, represents how well the dependent variable is explained by the independent variable/s.
The correlation coefficient ranges from −1 to 1. A value of 1 implies that a linear equation describes the relationship between *X* and *Y* perfectly, with all data points lying on a line for which *Y* increases as *X* increases. A value of −1 implies that all data points lie

on a line for which *Y* decreases as *X* increases. A value of 0 implies that there is no linear correlation between the variables





4.What is scaling? Why is scaling performed? What is the difference between

normalized scaling and standardized scaling?

Scaling is a way to bring all the features in the dataset to the same scale, that is, within the same range.

Many machine learning models that use the gradient descent as an optimisation technique(linear regression, logistic regression etc) and distance algorithm like KNN, K-means and SVM are affected by the range of values of the features(variables). As such without scaling, the results of such models are not reliable. Though running such models won't show any errors but their results can not be trusted.
If feature scaling is not done, then a machine learning algorithm(which is sensitive to scaling) tends to weigh greater values, higher and consider

smaller values as the lower values, regardless of the unit of the values. As such, features with greater values dominate over others.
Hence, to avoid these problems, we perform feature scaling. Scaling is done during the data pre-processing stage. However Tree-based algorithms are fairly insensitive to the scale of the features.

Normalization or Min-Max Scaling is used to transform features to be on a similar scale.
Mathematically, it is represented by the equation:
X_new = (X - X_min)/(X_max - X_min)
This brings the features on a scale between 0 and 1.
Normalization is useful when there are no outliers as it cannot cope up with them because it forces the outliers between the given range of 0 to 1.

Standardization or Z-Score Normalization is the transformation of features by subtracting from mean and dividing by standard deviation. This is often called as Z-score.
Mathematically, it is represented by the equation:
X_new = (X - mean)/Std
Standardization can be helpful in cases where the data follows a Gaussian distribution, but this is not always true.
Standardization does not get affected by outliers because there is no predefined range of transformed features.

Serial Number	Normalisation	Standardisation

1.	Minimum and maximum value of features are used for scaling	Mean and standard deviation is used for scaling.
2.	It is used when features are of different scales.	It is used when we want to ensure zero mean and unit standard deviation.
3.	Scales values between [0, 1] or [-1, 1].	It is not bounded to a certain range.
4.	It is really affected by outliers.	It is much less affected by outliers.

5.	Scikit-Learn provides a transformer called MinMaxScaler for Normalization.	Scikit-Learn provides a transformer called StandardScaler for standardization.
6.	This transformation squishes the n-dimensional data into an n-dimensional unit hypercube.	It translates the data to the mean vector of original data to the origin and squishes or expands.

7.	It is useful when we don’t know about the distribution	It is useful when the feature distribution is Normal or Gaussian.
8.	It is a often called as Scaling Normalizati on	It is a often called as Z-Score Normalizati on.

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

VIF or the variance inflation factor is a measure of collinearity between coefficients.
It compares each feature in terms of it’s collinearity with all the other features.
Mathematically , it is represented by the

equation:

- VIF = 1 / (1 – R2)

R2 = R-Squared

Infinite value of VIF represents R-Squared score of 1 or we can say that there is perfect correlation.

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

Q-Q Plots or Quantile-Quantile plots are plots of two quantiles against each other.

A quantile represents a fraction where certain values fall below a percentage threshold.

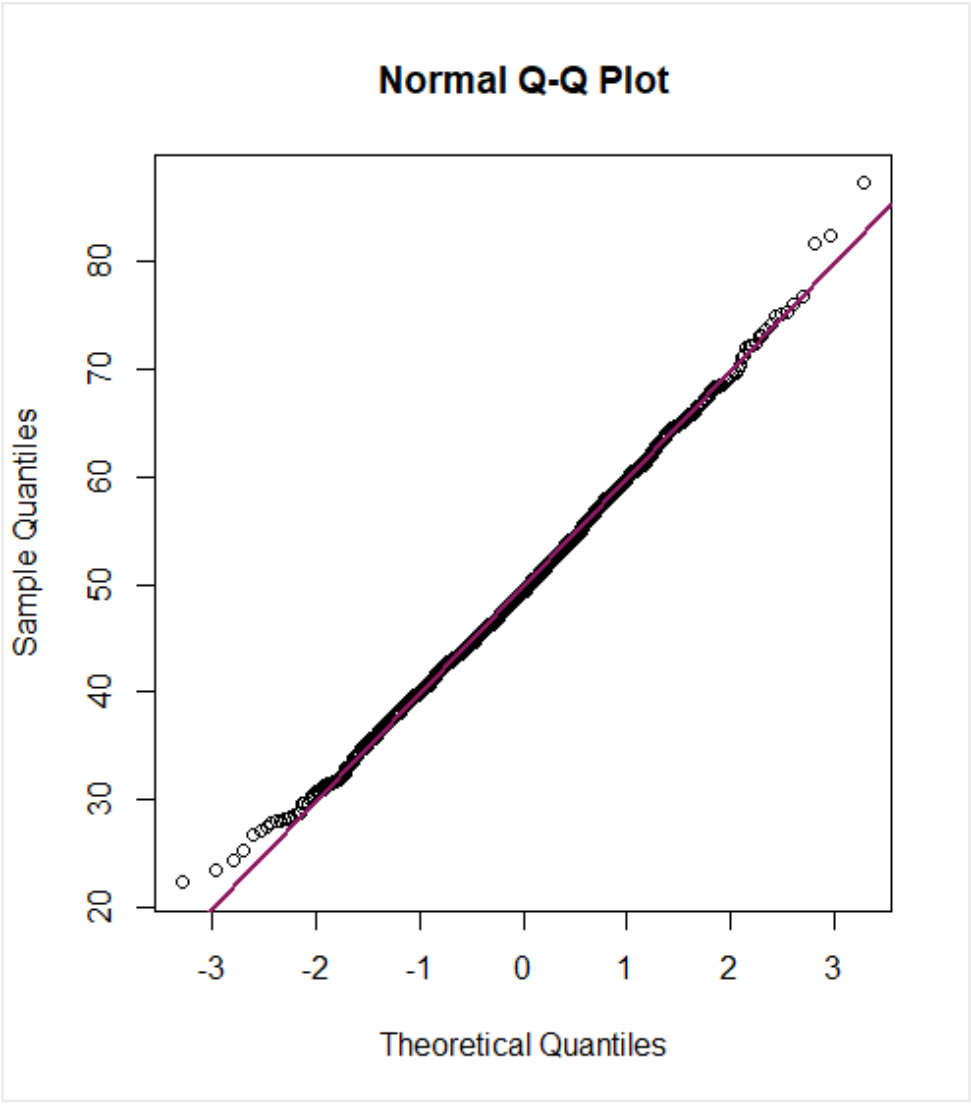
A Q-Q plot tells us the following :

1. Whether the two datasets come from the same sample distribution.
2. Whether the two datasets have common location and scale.
3. Whether the two datasets have similar

distributional shapes.

4. Whether the two datasets

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 (45-degree reference line) that's roughly straight.
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:

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

2. 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.