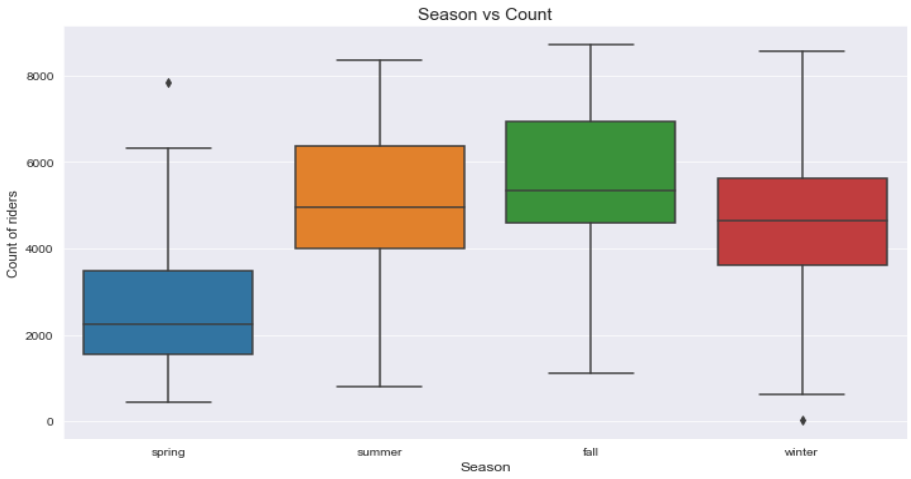
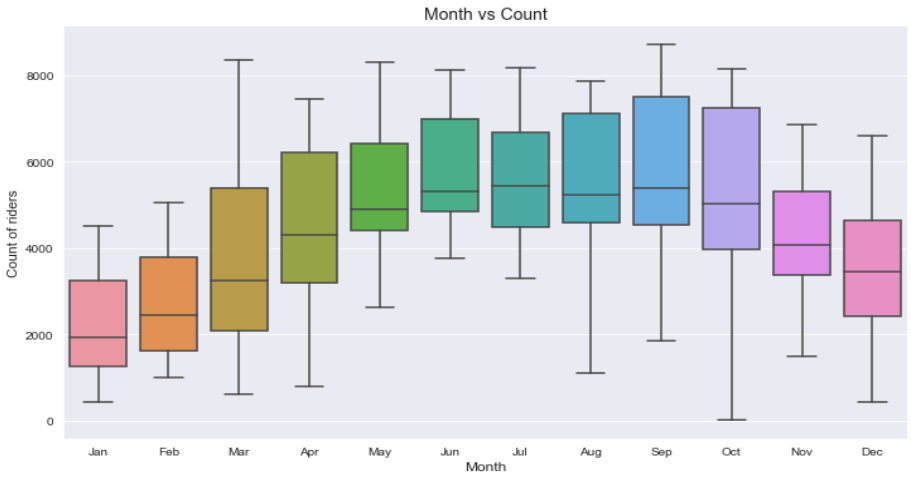
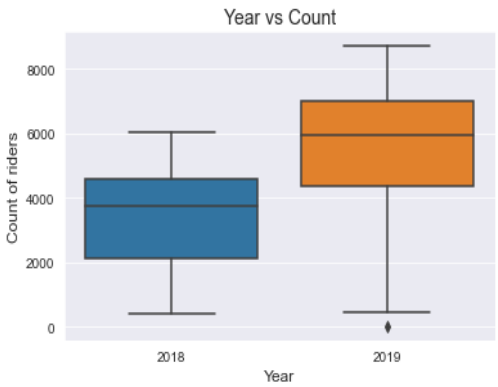
Assignment-based Subjective Questions

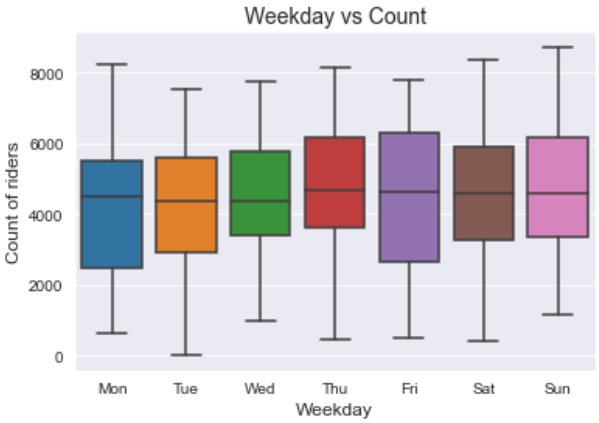
1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

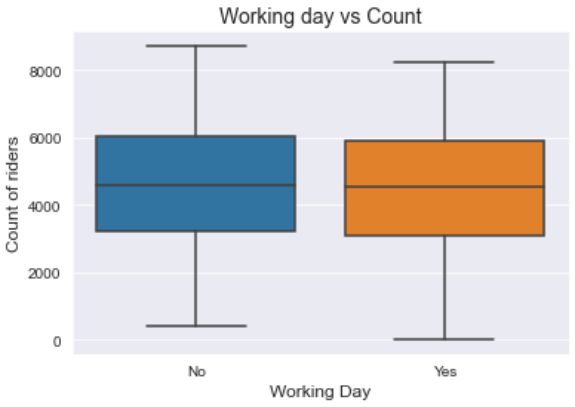
Visualizations for analysing effect of categorical variables from the dataset on the dependent variable:

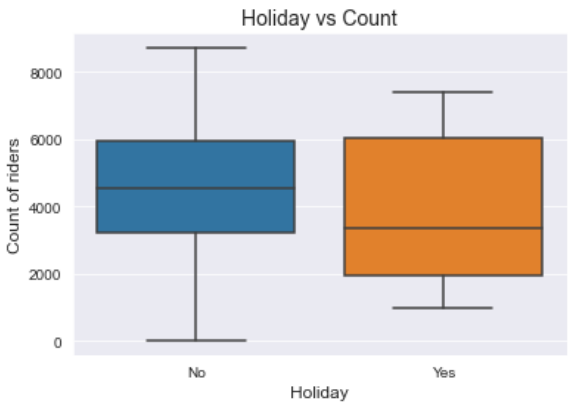


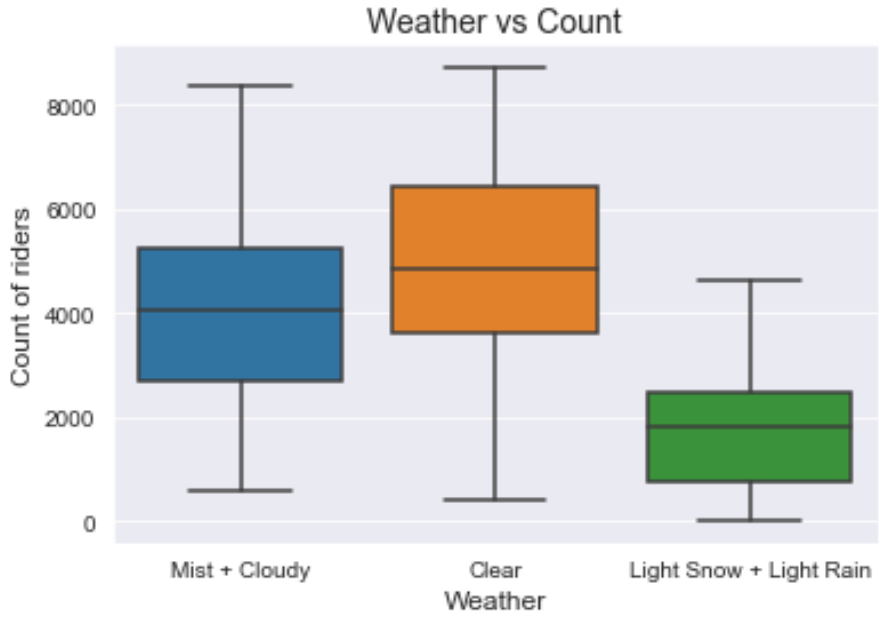












Inferences:

- The median number of riders is the minimum for spring season. The median number of riders for summer fall and winter are almost same and is considerably higher compared to spring. However, the median and the range of count of riders for fall is slightly higher than summer and winter.

  We can infer that fall season seems to be the most lucrative time for bike riders followed by summer and winter. Spring seems to be the least lucrative season.

- When we look at the distribution of riders across months, a similar trend appears where the median number of riders increases from January to June and then it stabilizes till around September before decreasing from October to December. However, one observation that we see among months is that March has a very high range of count of riders compared to other spring months. We observe a similar pattern for October as well.

- We can infer that year plays a very crucial role in the number of riders. From 2018 to 2019 the range of count of riders has increased drastically. In fact the median number of riders in 2019 is almost equal to the maximum riders in 2018.

- Almost all days of the week have similar range of count of riders with a similar median number of riders for all days. Saturday and Sunday seem to have a slightly higher range though it's not very significant.

- For both working and non-working day, the median number of riders is almost same. However, the range for count of riders for non-working day is slightly higher

- The median number of riders and the range of count of riders for a non-holiday are higher compared to a holiday. However, we can see that on holidays the range of count of riders in the 25th to 75th percentile is much more comparatively

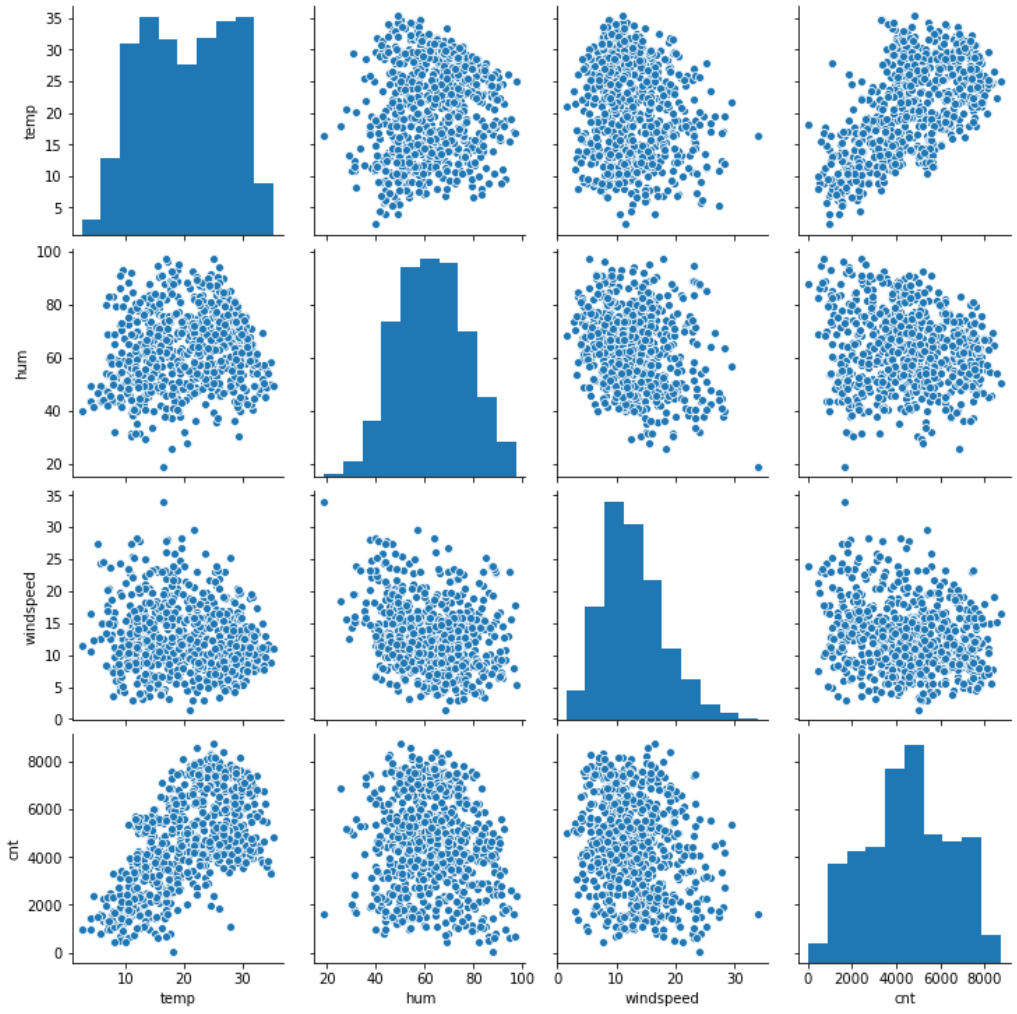
- Clear weather has the highest median count of riders as we can expect. This is followed by mist and cloudy weather. The range of count of riders for both clear and cloudy weather is similar.
  However, the median number of riders as well as the range of count of riders is comparatively much less when the weather is light snow or light rain.

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

When we create dummy variables, it creates dummy variable columns for each category. **drop_first=True** is important to use, as it helps in reducing the extra column created during dummy variable creation. This in turn reduces the correlations created among dummy variables.

For Example, if we have a variable gender, we 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. Hence if we have categorical variable with n-levels, then we need to use n-1 columns to represent the dummy variables.

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

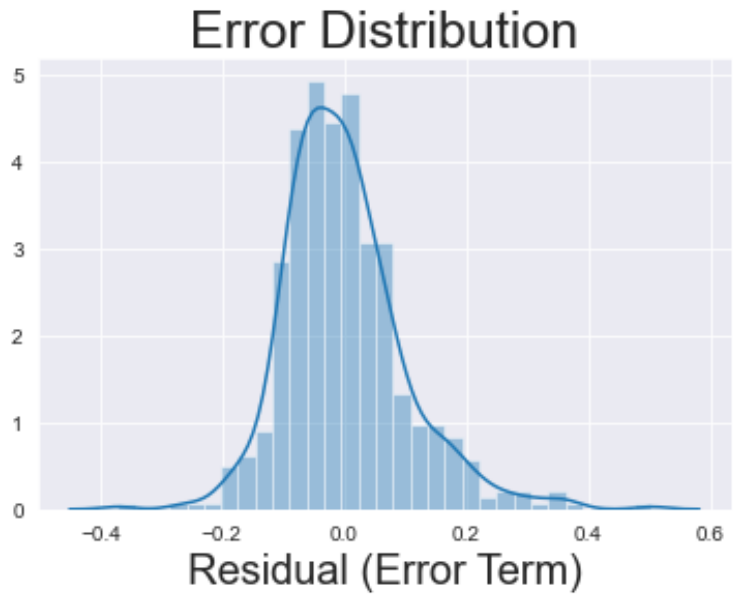


Looking at the pair plot above, among the numerical features, temp has the highest correlation with the target variable

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

After building the model on the training set, I validated the assumptions of Linear Regression by performing residual analysis on the training data. First I calculated the residuals or error terms by taking the difference of the predicted value of the training data and the actual values. After that, we can perform the analysis as below:

Assumption - Normally Distributed Error Terms

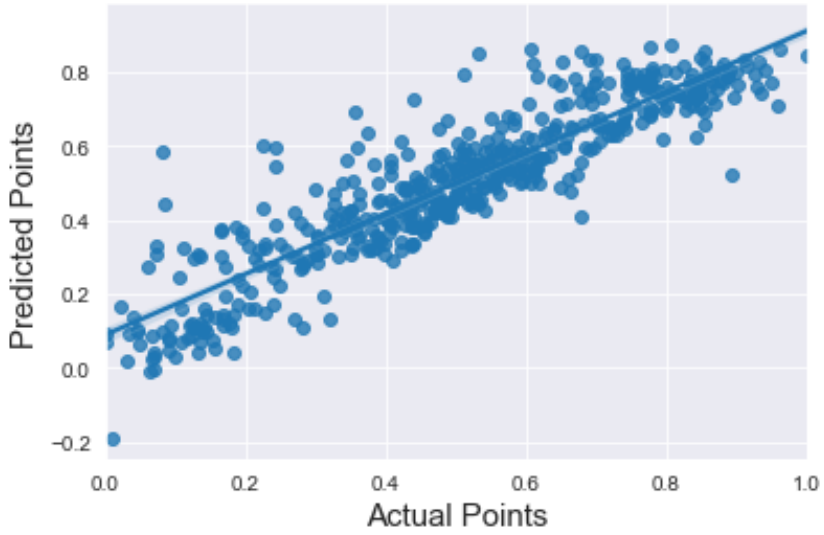


When we plot the histogram of the error terms, it is evident that Error Distribution Is normally distributed around 0, which indicates that our model has handled the assumption of Error Normal Distribution properly

Assumption - Error Terms are Independent

When we plot the residual vs predicted values, we see that there is no specific pattern between Residual & Predicted Value. This shows that the error terms are independent and no pattern exists inherently in the error terms.

Assumption – Homoscedasticity



When we plot the predicted values vs actual values, we can say that residuals are equally distributed across predicted value.

This means there is equal variance and we do not observe high concentration of data points in any particular region.

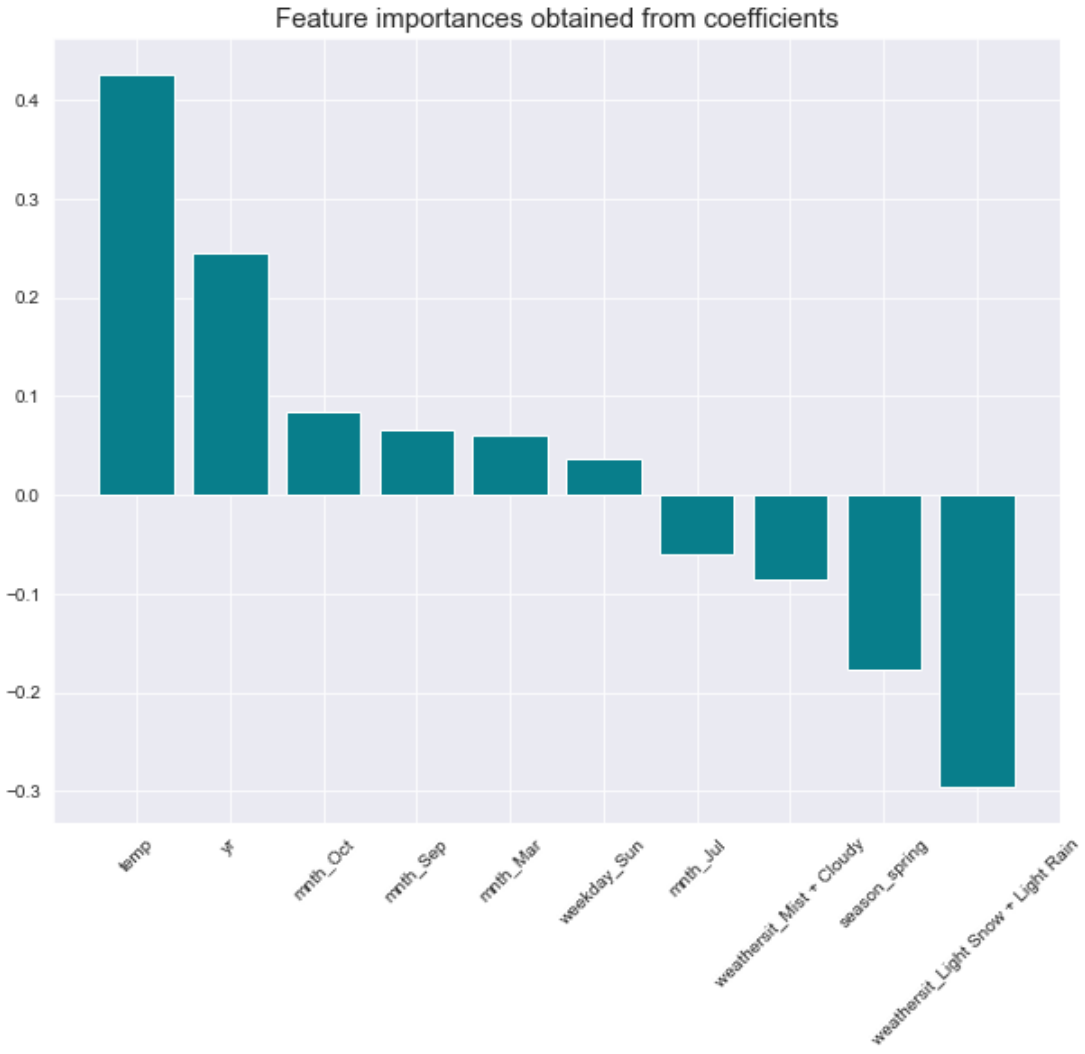
This proves Homoscedasticity of Error Terms

Assumption – Multi-Collinearity

We can infer that there is no multicollinearity among the independent variables based on their VIF values.

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

I have checked for feature importance by examining the final model's coefficients.
If an assigned coefficient is a large (negative or positive) number, it has some influence on the prediction. On the contrary, if the coefficient is zero, it doesn't have any impact on the prediction.



Based on the above criteria, the top three features contributing significantly towards explaining the demand of the shared bikes are:

- **Temperature (temp)** - A coefficient value of 0.4265 indicates that a unit increase in temp variable keeping other variables constant increases the bike hire numbers by 0.4265 units.
- **Weather - Light Snow + Light Rain** - A coefficient value of -0.2961 indicates that a unit increase in weathersit_Light Snow + Light Rain variable keeping other variables constant decreases the bike hire numbers by 0.2961 units.
- **Year (yr)** - A coefficient value of 0.2462 indicates that a unit increase in yr variable keeping other variables constant increases the bike hire numbers by 0.2462 units.

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Linear regression is a linear approach to modelling the relationship between a scalar response (y) and one or more explanatory variables (X1, X2,…Xn). The case of one explanatory variable is called **simple linear regression**; for more than one, the process is called **multiple linear regression**. Different techniques can be used to prepare or train the linear regression equation from data, the most common of which is called **Ordinary Least Squares**.

The linear equation assigns one scale factor to each input value or column, called a coefficient and represented by the capital Greek letter Beta (B). One additional coefficient is also added, giving the line an additional degree of freedom (e.g. moving up and down on a two-dimensional plot) and is often called the intercept or the bias coefficient.

For example, in a simple regression problem (a single x and a single y), the form of the model would be: y = B0 + B1*x

In higher dimensions when we have more than one input (x), the line is called a plane or a hyper-plane. The representation therefore is the form of the equation and the specific values used for the coefficients (e.g. B0 and B1 in the above example).

Given the representation is a linear equation, making predictions is as simple as solving the equation for a specific set of inputs.

Linear regression makes the following assumptions:

1. There should be a **linear and additive relationship** between dependent (response) variable and independent (predictor) variable(s). A linear relationship suggests that a change in response Y due to one unit change in X^1 is constant, regardless of the value of X^1. An additive relationship suggests that the effect of X^1 on Y is independent of other variables.
2. There should be **no correlation between the residual** (error) terms.
3. The independent variables should not be correlated. Absence of this phenomenon is known as multicollinearity.
4. The error terms must have constant variance. This phenomenon is known as **Homoscedasticity**. The presence of non-constant variance is referred to heteroscedasticity.
5. The error terms must be normally distributed.

2. Explain the Anscombe's quartet in detail.

Anscombe's quartet comprises four datasets that have nearly identical simple descriptive statistics but there are some peculiarities in the dataset that **fools the regression model** if built. They have very different distributions and **appear differently** when plotted on scatter plots.

They were constructed by statistician Francis Anscombe to demonstrate both the **importance of graphing data** before analysing it and **the effect of outliers** on statistical properties.

The four datasets are:

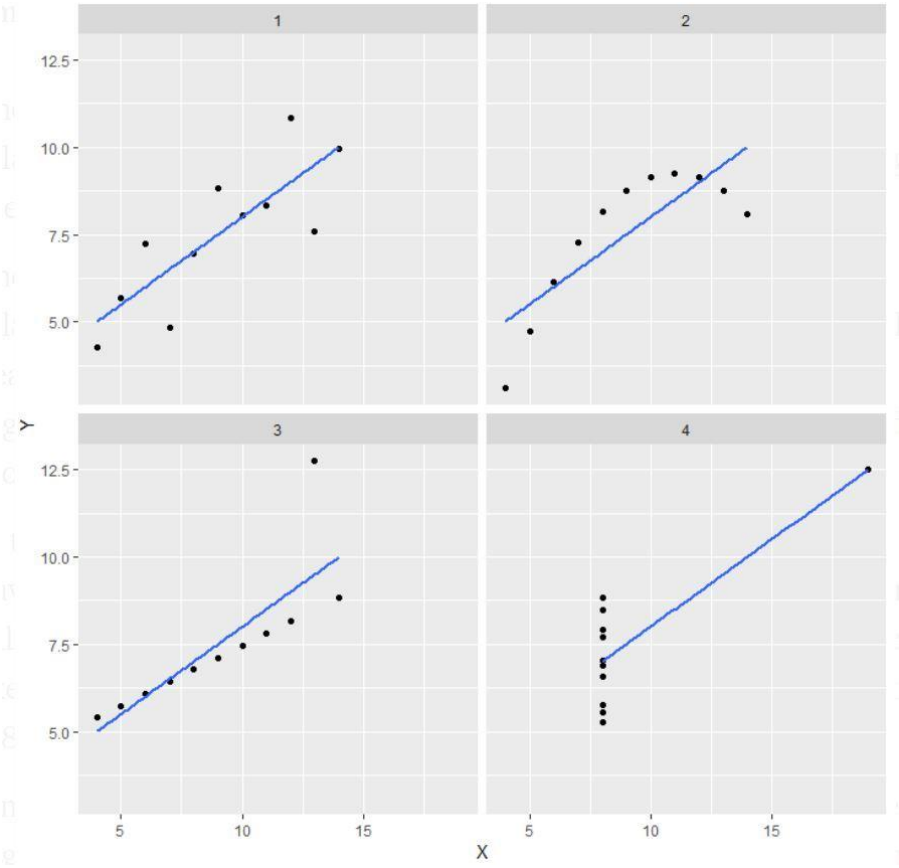
Anscombe's Data										
Observation	x1	y1		x2	y2		x3	y3		x4	y4
1	10	8.04		10	9.14		10	7.46		8	6.58
2	8	6.95		8	8.14		8	6.77		8	5.76
3	13	7.58		13	8.74		13	12.74		8	7.71
4	9	8.81		9	8.77		9	7.11		8	8.84
5	11	8.33		11	9.26		11	7.81		8	8.47
6	14	9.96		14	8.1		14	8.84		8	7.04
7	6	7.24		6	6.13		6	6.08		8	5.25
8	4	4.26		4	3.1		4	5.39		19	12.5
9	12	10.84		12	9.13		12	8.15		8	5.56
10	7	4.82		7	7.26		7	6.42		8	7.91
11	5	5.68		5	4.74		5	5.73		8	6.89

The statistical information for all these four datasets are approximately similar and can be computed as follows:

Summary

Set	mean(X)	sd(X)	mean(Y)	sd(Y)	cor(X,Y)
1	9	3.32	7.5	2.03	0.816
2	9	3.32	7.5	2.03	0.816
3	9	3.32	7.5	2.03	0.816
4	9	3.32	7.5	2.03	0.817

When these models are plotted on a scatter plot, all datasets generates a different kind of plot as shown below:



The above four datasets can be described as:

- In the first one (top left) if you look at the scatter plot you will see that there seems to be a linear relationship between x and y.
- In the second one (top right) if you look at this figure you can conclude that there is a non-linear relationship between x and y.
- In the third one (bottom left) you can say when there is a perfect linear relationship for all the data points except one which seems to be an outlier and is indicated by it being far away from that line.
- Finally, the fourth one (bottom right) shows an example when one high-leverage point is enough to produce a high correlation coefficient.

3. What is Pearson's R?

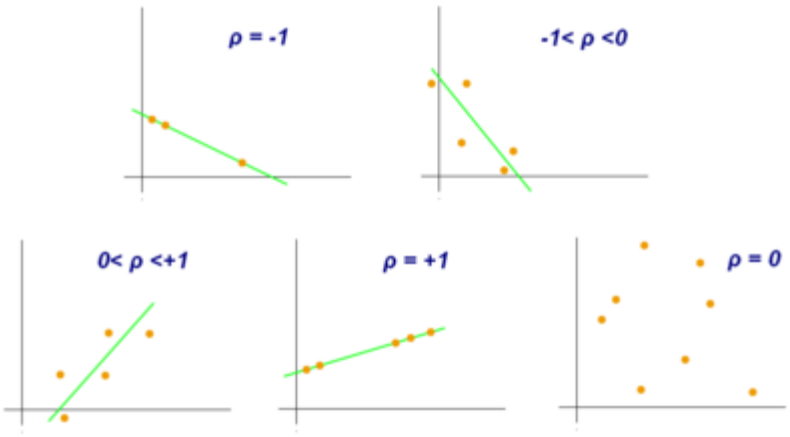
Pearson's R is a measure of linear correlation between two sets of data. It is the covariance of two variables, divided by the product of their standard deviations; thus it is essentially a normalised measurement of the covariance, such that the result always has a value between −1 and 1. As with covariance itself, the measure can only reflect a linear correlation of variables, and ignores many other types of relationship or correlation.

$$\rho_{X,Y} = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y}$$

The Pearson's correlation coefficient varies between -1 and +1 where:

- r = 1 means the data is perfectly linear with a positive slope (i.e., both variables tend to change in the same direction)
- r = -1 means the data is perfectly linear with a negative slope (i.e., both variables tend to change in different directions)
- r = 0 means there is no linear association

Examples of scatter plots with different values of Pearson's R



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

Feature scaling is a step of data Pre-Processing. It is applied to independent variables to normalize the data within a particular range.

Most of the times, collected data set contains features highly varying in magnitudes, units and range. If scaling is not done then algorithm only takes magnitude into account and not units. This leads to incorrect modelling. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude.

For example, many classifiers calculate the distance between two points by the Euclidean 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 scaled so that each feature contributes approximately proportionately to the final distance.

It also helps in speeding up the calculations in an algorithm as gradient descent converges much faster with feature scaling than without it.

Normalization: It involves in rescaling the range of features to scale the range in [0, 1]

$$\text{MinMax Scaling: } x = \frac{x - min(x)}{max(x) - min(x)}$$

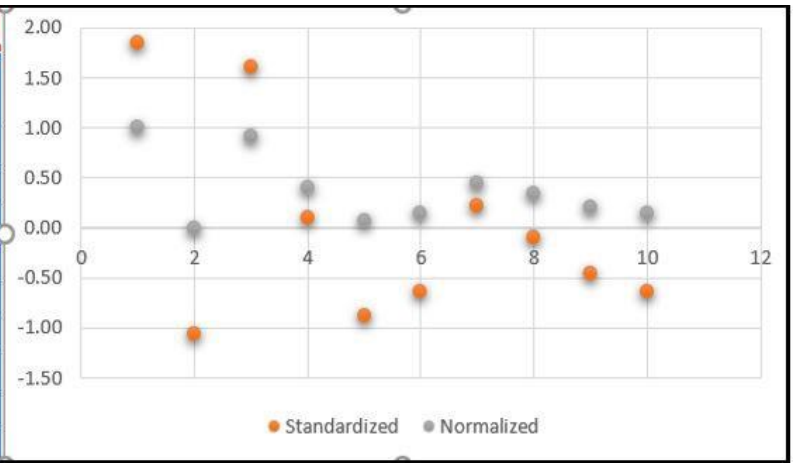
Standardization: Standardization replaces the values by their Z scores. It brings all of the data into a standard normal distribution which has mean (μ) zero and standard deviation one (σ)

$$\text{Standardisation: } x = \frac{x - mean(x)}{sd(x)}$$

One disadvantage of normalization over standardization is that it **loses** some information in the data, especially about **outliers**.

Example of standardization and normalization on original values:

Original	Standardized	Normalized
59	1.85	1.000
11	-1.06	0.000
55	1.61	0.917
30	0.09	0.396
14	-0.88	0.063
18	-0.64	0.146
32	0.21	0.438
27	-0.09	0.333
21	-0.46	0.208
18	-0.64	0.146



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

If there is perfect correlation, then the value of VIF is infinity. This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) as infinity. To solve this problem we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.
An infinite VIF value for a variable might also indicate that the corresponding variable may be expressed exactly by a linear combination of other variables (which show an infinite VIF as well).

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

Quantile-Quantile (Q-Q) plot is a graphical tool to help us assess if a set of data came from some theoretical distribution such as a Normal, Exponential or Uniform distribution.
Also, it helps us determine if two data sets come from populations with a common distribution.

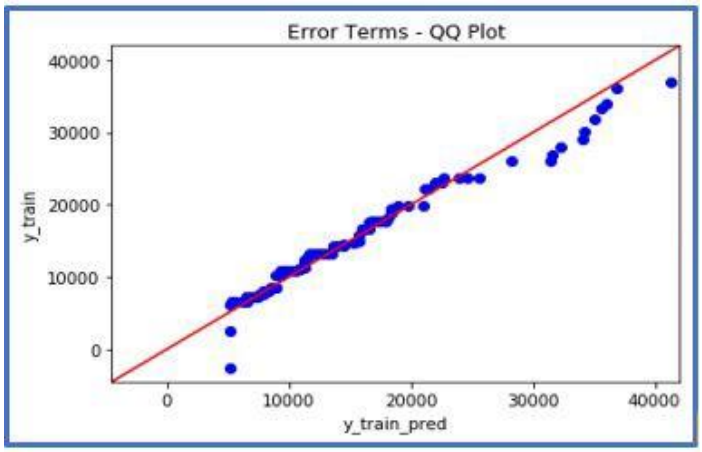
This helps in a scenario of linear regression when we have training and test data set that we have received separately. We can use Q-Q plot to confirm whether both the data sets are from populations with same distributions.

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

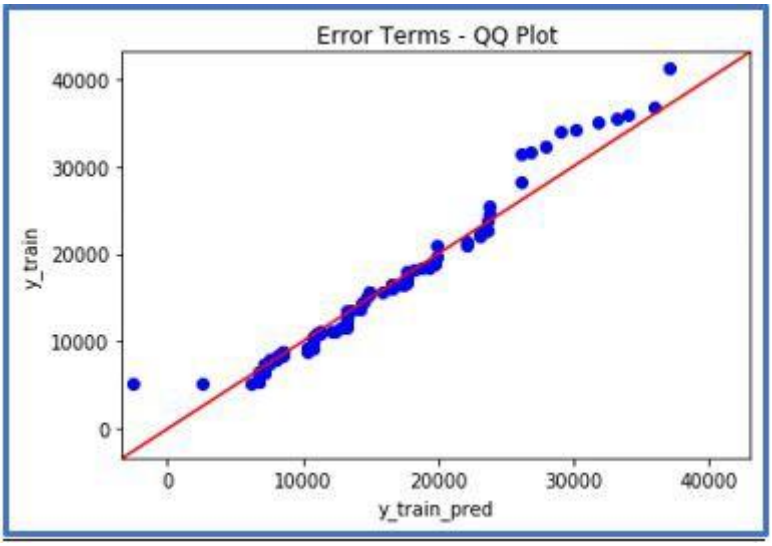
Below are the possible interpretations for two data sets.

- **Similar distribution:** If all point of quantiles lies on or close to straight line at an angle of 45 degree from x –axis

- **Y-values < X-values:** If y-quantiles are lower than the x-quantiles.



- **X-values < Y-values:** If x-quantiles are lower than the y-quantiles.



- **Different distribution:** If all point of quantiles lies away from the straight line at an angle of 45 degree from x -axis