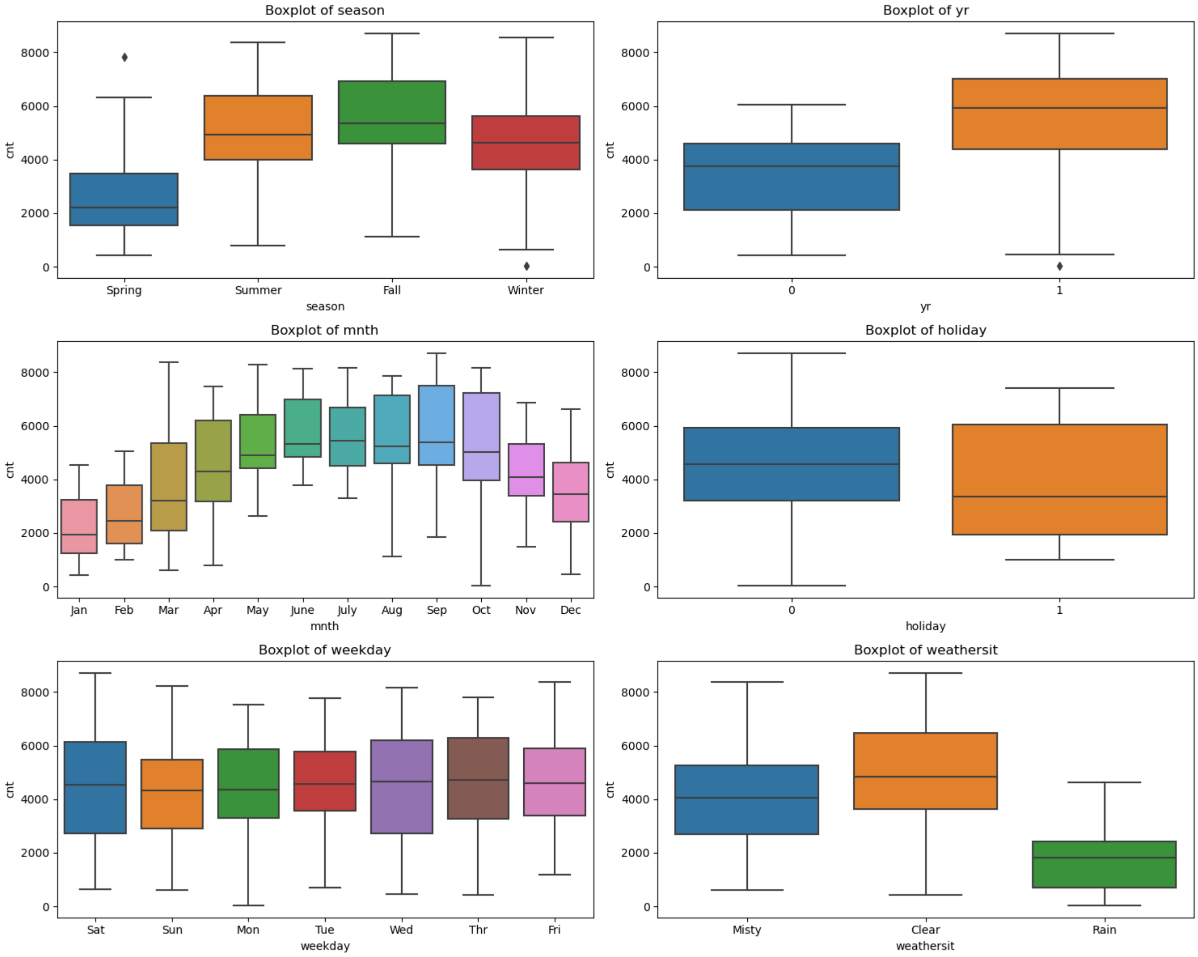
Assignment-based Subjective Questions

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



The categorical variable used in the dataset: season , yr(year) , holiday, weekday , weathersit(weather situation) and mnth(month) . These were visualized using a boxplot.

These variables had the following effect on our dependant variable:

- By season, fall has the highest rentals followed by summer.
- By year, 2019 had 2000 median increase in rentals compared to 2018.
- By weathersit, clear weather has the highest rentals followed by Misty.
- Holidays have lesser rentals than working days, also variation in demand on holidays is huge.
- September has the highest rentals followed by October and August
- There is no significant difference in rentals by weekdays.

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

When we have a categorical variable of 'n' levels, the idea of dummy variable creation is to build 'n-1' variables, indicating the levels. For a variable say, 'Relationship' with three levels namely, 'Single', 'In a relationship', and 'Married', we would create a dummy table like the following:

Relationship Status	Single	In a relationship	Married
Single	1	0	0
In a relationship	0	1	0
Married	0	0	1

But we can clearly see that there is no need of defining three different levels. If we drop a level, say 'Single', we would still be able to explain the three levels.

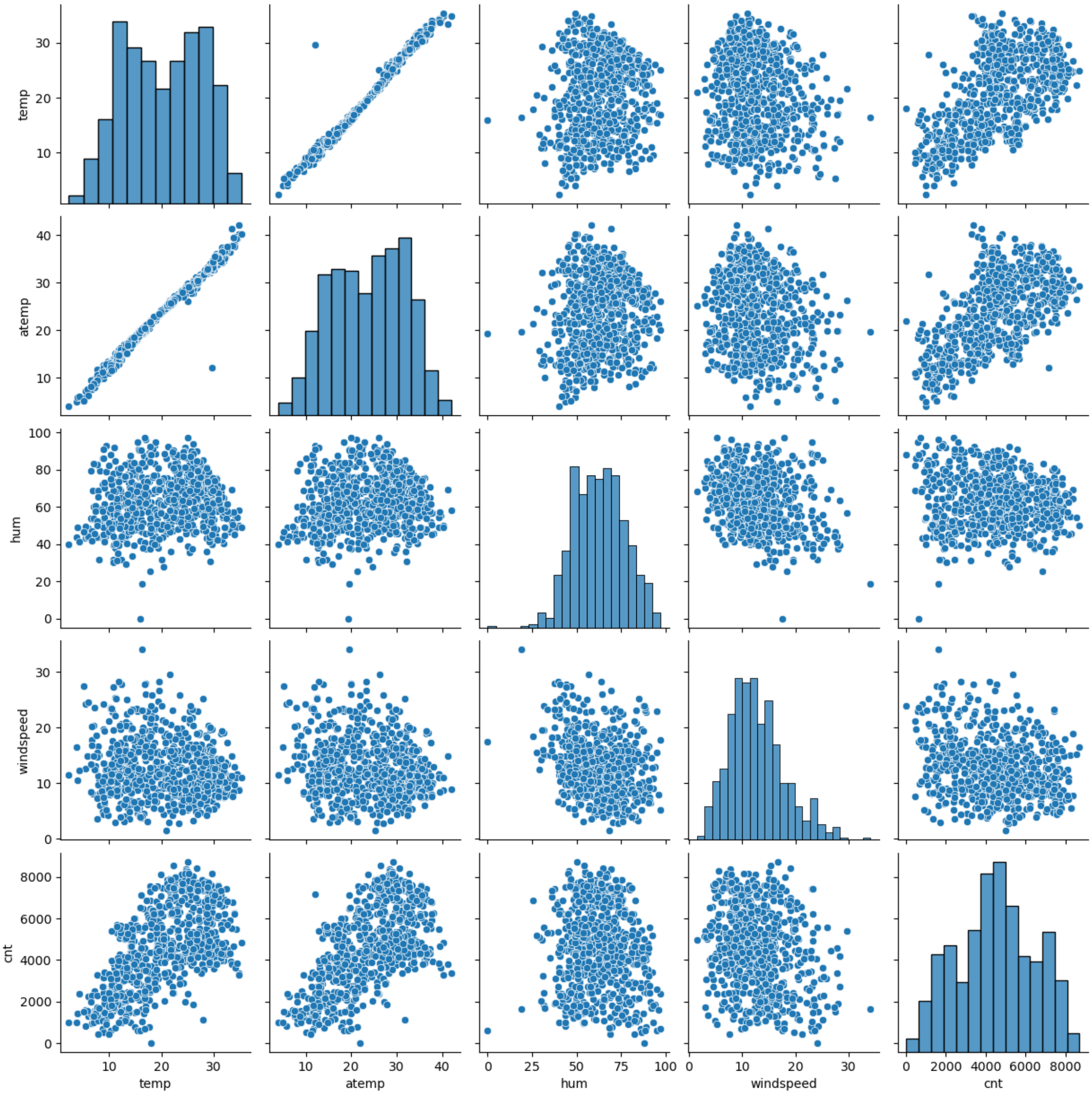
Let's drop the dummy variable 'Single' from the columns:

Relationship Status	In a relationship	Married
Single	0	0
In a relationship	1	0
Married	0	1

If both the dummy variables namely 'In a relationship' and 'Married' are equal to zero, that means that the person is single. If 'In a relationship' is one and 'Married' is zero, that means that the person is in a relationship and finally, if 'In a relationship' is zero and 'Married' is 1, that means that the person is married.

In Python, we use drop_first=True to implement the above logic. This will create ‘n-1’ dummy variables for a categorical variable of 'n' levels.

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

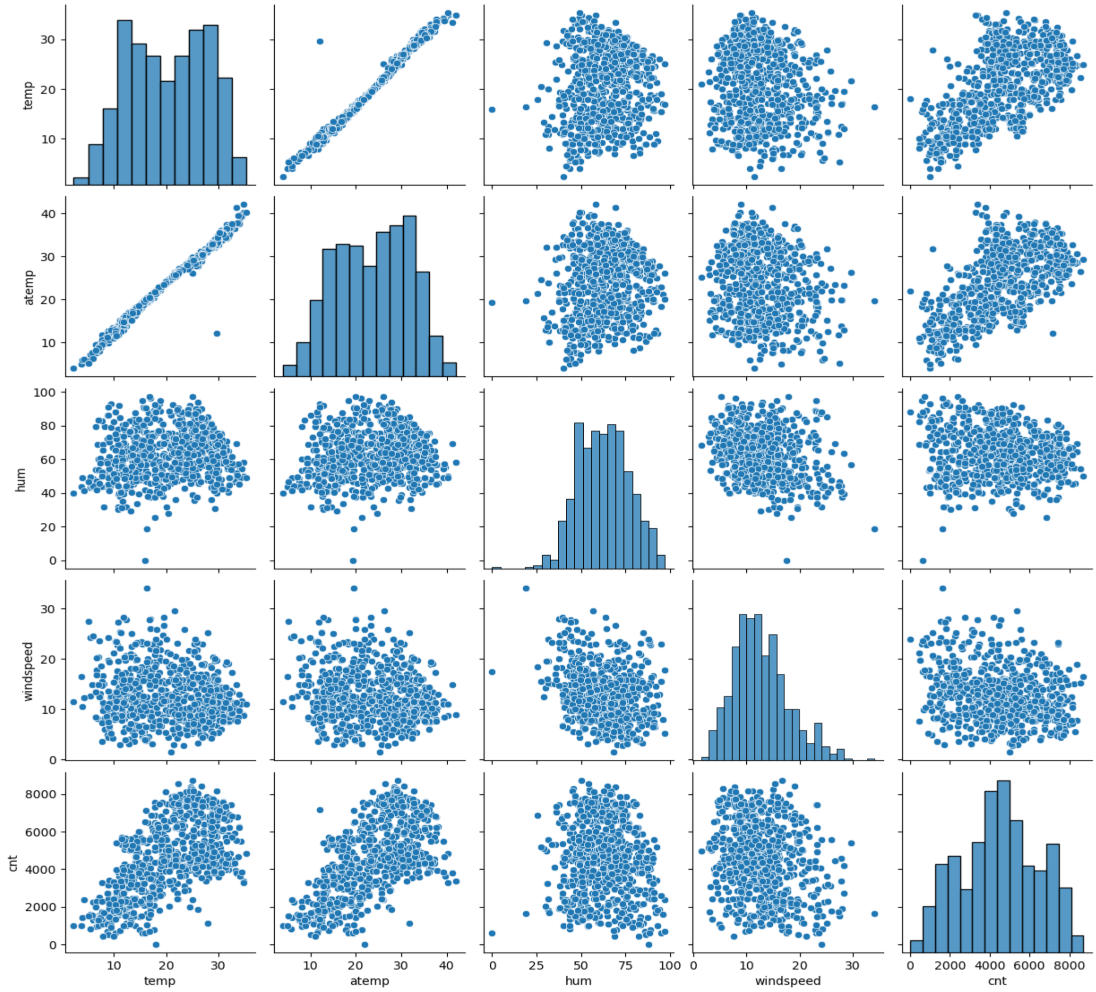


As we can see from the above pairplot, “temp” and “atemp” are highly correlated with the target variable (cnt). Pearson’s R for “atemp” and “temp”, with target variable, are 0.630 and 0.627 respectively.

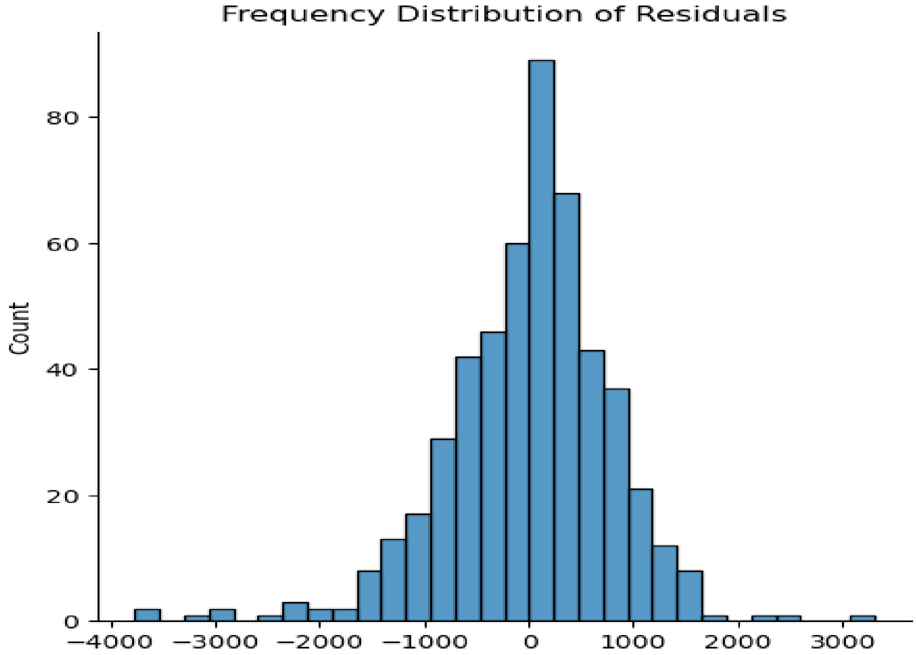
4. **How did you validate the assumptions of Linear Regression after building the model on the training set?**

Following are the ways in which the assumptions of Linear Regression can be validated for the given dataset:

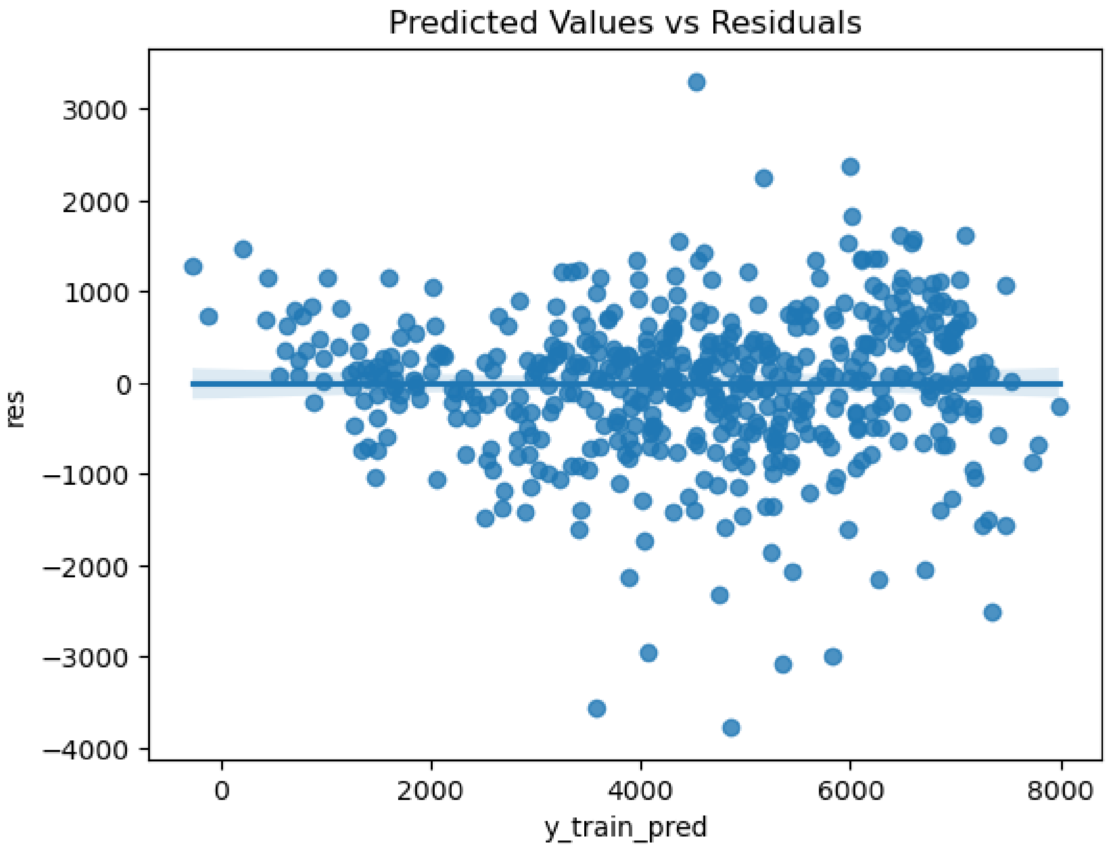
- There is a linear relationship between dependent and independent variables:



- Residual errors follow normal distribution, i.e. the mean is 0:



- Error terms are independent of each other:



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

   The top three features contributing significantly towards explaining the demand of the shared bikes are:

   - **Temp** with **positive** coefficient of **4368.639370**
   - **Yr** with **positive** coefficient of **2020.888840**
   - Weathersit – **Rain** with **negative** coefficient of **2644.415685**

Feature	MLR Co-eff
temp	4368.639370
yr	2020.888840
Rain	-2644.415685

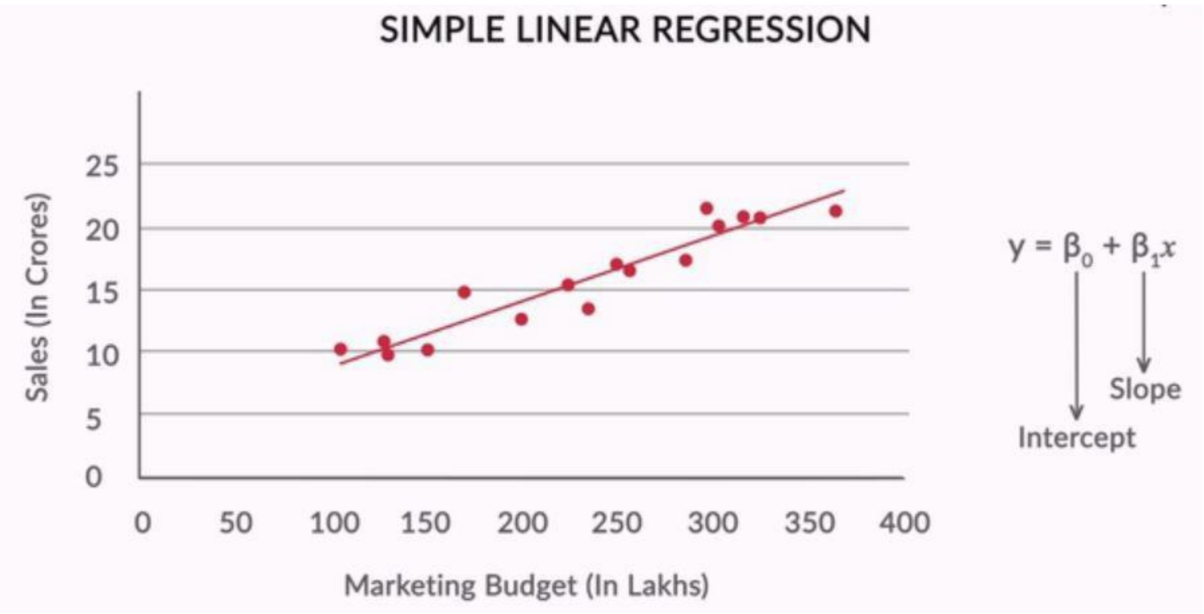
General Subjective Questions

1. **Explain the linear regression algorithm in detail.**

The most used and simple regression model is the simple linear regression which explains the relationship between a dependent variable and one independent variable using a straight line.

The standard equation of the regression line is given by the following expression:

$Y = \beta_0 + \beta_1 X$



To find the best values of β_0 and β_1, we need to define a cost function that measures how well the line fits the data. A common choice is the mean squared error (MSE), which is the average of the squared differences between the actual y values and the predicted y values:

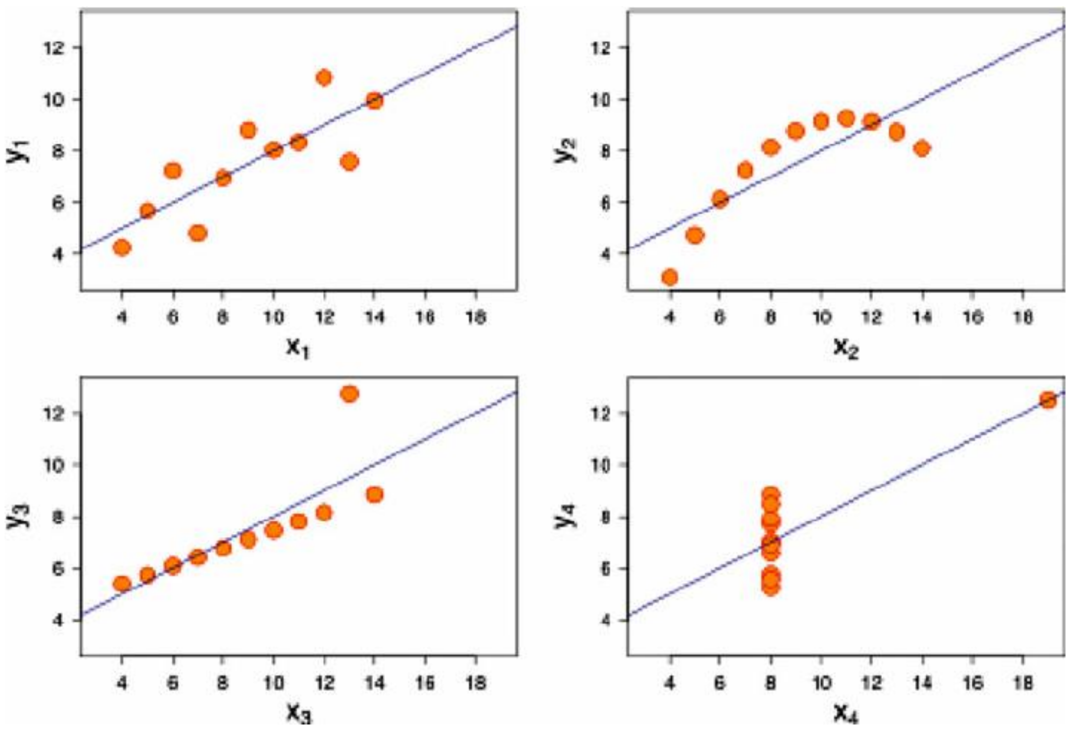
MSE = (1/n) * ∑(y - y')^2

where n is the number of data points, y is the actual value, and y' is the predicted value.

The goal is to minimise MSE by optimizing finding optimal values of β_0 and β_1.

2. **Explain the Anscombe's quartet in detail.**

Anscombe's quartet comprises four datasets that have nearly identical simple descriptive statistics yet have very different distributions and appear very different when graphed. Statistician Francis Anscombe constructed them to demonstrate both the importance of graphing data when analysing it, and the effect of outliers and other influential observations on statistical properties.



- The first scatter plot (top left) appears to be a simple linear relationship.
- The second graph (top right) is not distributed normally; while there is a relation between them is not linear.
- 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.
- Finally, the fourth graph (bottom right) shows an example when one high-leverage point is enough to produce a high correlation coefficient, even though the other data points do not indicate any relationship between the variables.

3. What is Pearson's R?

The Pearson correlation coefficient (r) is the most common way of measuring a linear correlation. It is a number between –1 and 1 that measures the strength and direction of the relationship between two variables.

Formula

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

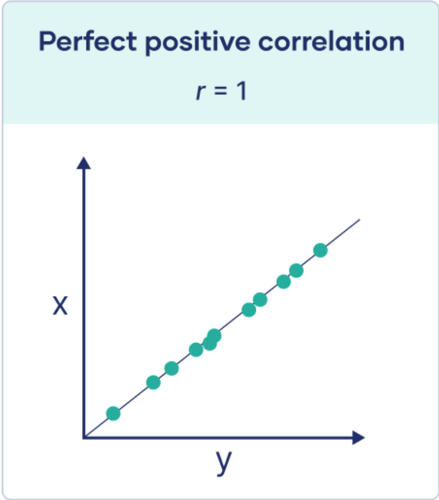
r = correlation coefficient

x_i = values of the x-variable in a sample

$\bar{x}$ = mean of the values of the x-variable

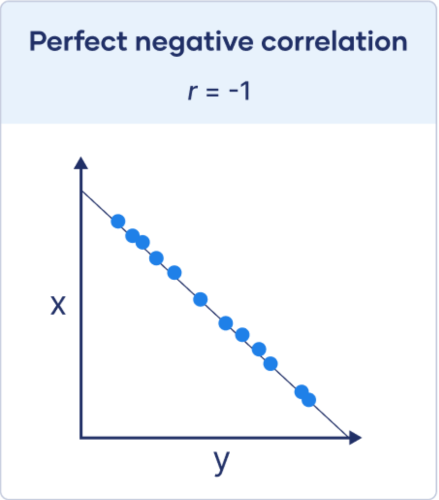
y_i = values of the y-variable in a sample

$\bar{y}$ = mean of the values of the y-variable



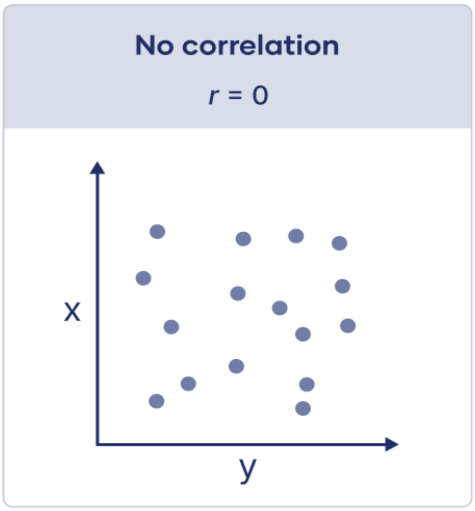
Perfect positive correlation

r = 1



Perfect negative correlation

r = -1



No correlation

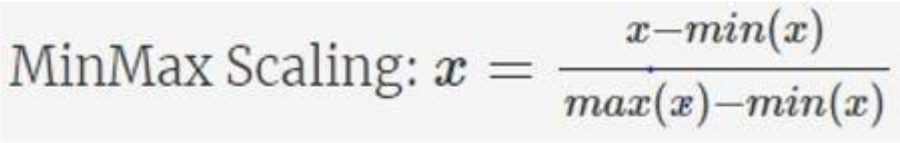
r = 0

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

Feature scaling is a method used to normalize or standardize the range of independent variables or features of data. It is performed during the data pre-processing stage to deal with varying values in the dataset. 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, irrespective of the units of the values.

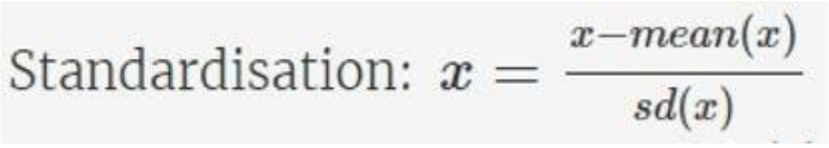
It is important to note that scaling just affects the coefficients and none of the other parameters like **t-statistic, F-statistic, p-values, R-squared, etc.**

- It brings the data in the range of 0 and 1:



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

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



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

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

The value of VIF is calculated by the below formula:

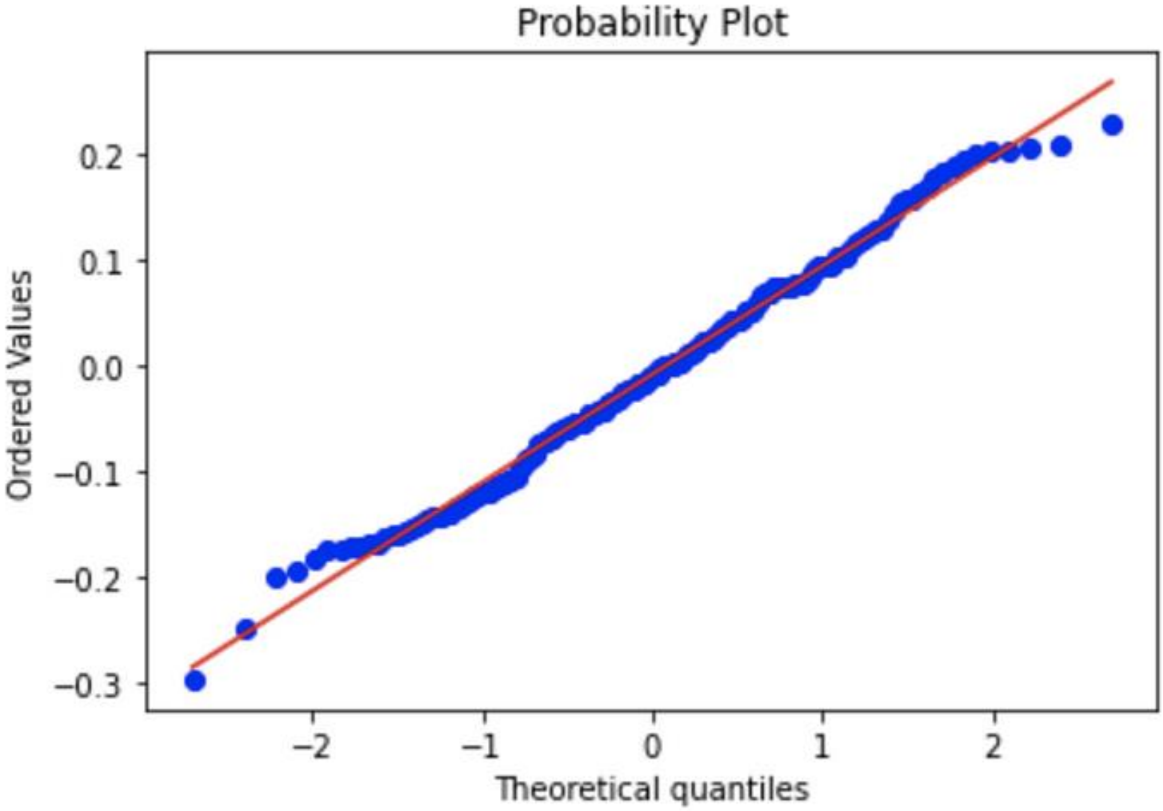
$$\mathrm{VIF_i} = \frac{1}{1 - R_i^2}$$

Where, ‘i’ refers to the ith variable.

If R-squared value is equal to 1 then the denominator of the above formula become 0 and the overall value become infinite. It denotes **perfect correlation** in variables.

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

The Q-Q plot or quantile-quantile plot is a graphical technique for determining if two data sets come from populations with a common distribution. In a Q-Q plot, scatterplot is 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 that’s roughly straight. Here’s an example of a Normal Q-Q plot when both sets of quantiles truly come from Normal distributions.



Use of Q-Q plot in Linear Regression:
The Q-Q plot is used to see if the points lie approximately on the line. If they don't, it means, our residuals aren't Gaussian (Normal) and thus, our errors are also not Gaussian.

Importance of Q-Q plot: Below are the points:

- 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.
- The q-q plot can provide more insight into the nature of the difference than analytical methods.