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? (3 marks)

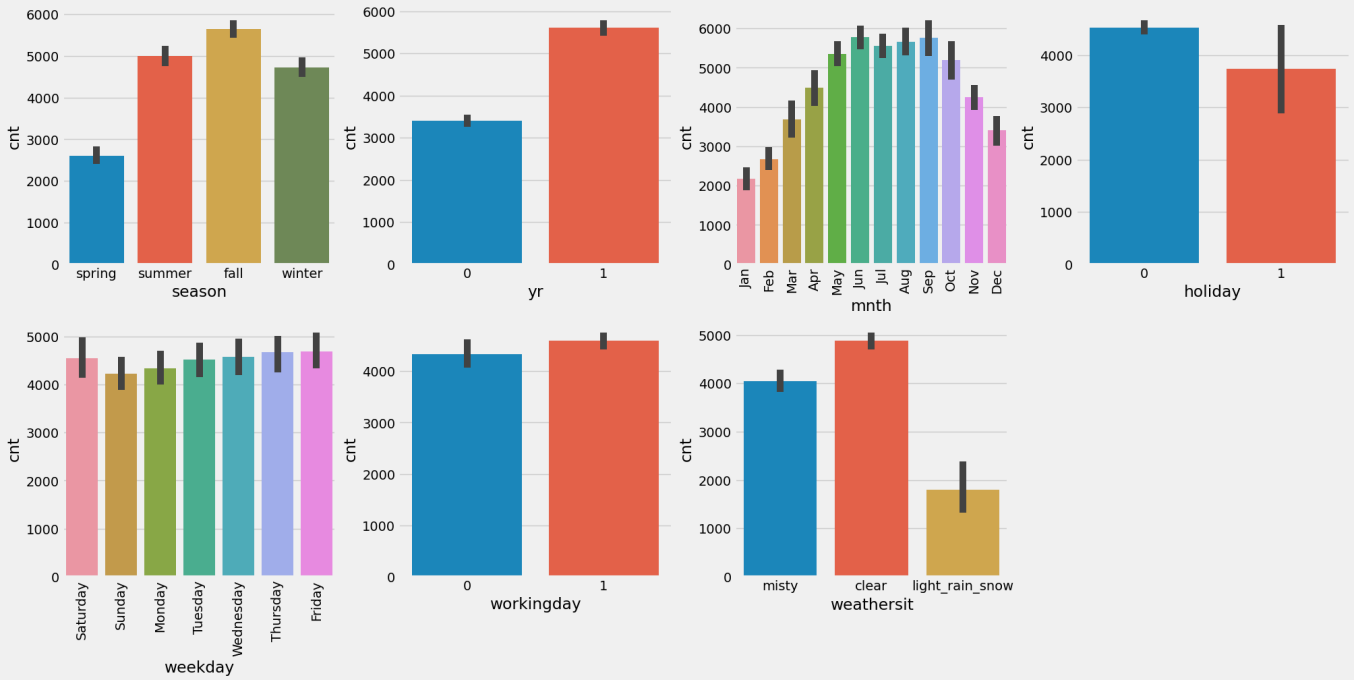


Figure: Mean plot of categorical variables w.r.t **cnt**

- **Season:** A higer trend is seen in the fall season followed by summer
- **Year:** Year 2019 saw a higher trend on the dependent variable compared to 2018
- **Month:** Higher trend seen in the middle months
- **Holiday, Weekday & WorkingDay:** Constant trends seen for all categories among these columns
- **WeatherSit:** High trend seen for clear weather whereas low trend seen in light rain snow

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

Setting `drop_first=True` when creating dummy variables avoids multicollinearity, improves model performance, enhances coefficient interpretability, and leads to a more efficient model representation.

For example in our case, we have a categorical variable **"season"** (spring, summer, fall, winter), dropping the first dummy variable (e.g., "season_spring") ensures meaningful interpretation and avoids perfect multicollinearity.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

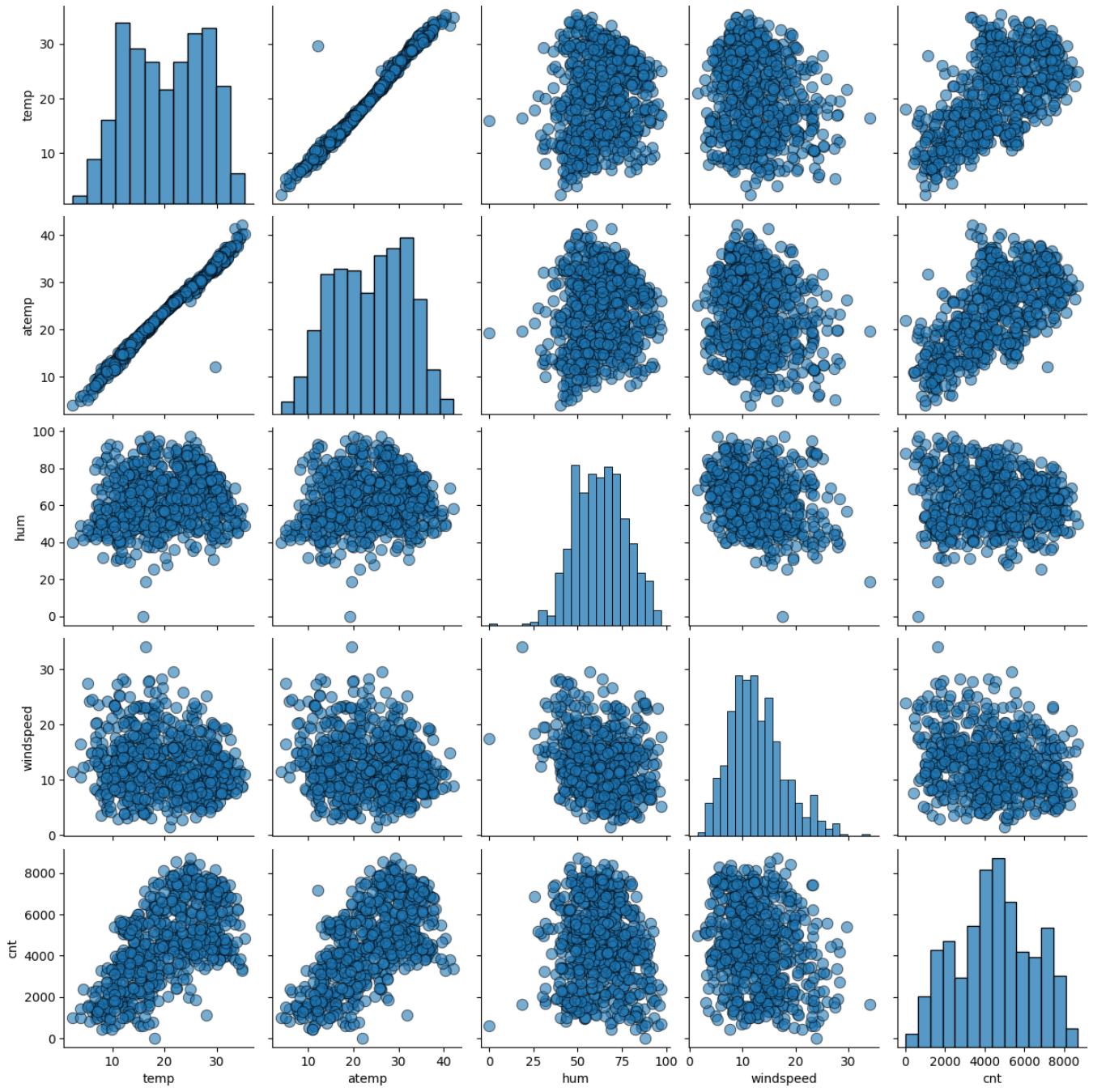


Figure: Pair plot among the numerical variables

From the pair plot, we can infer that the columns **temp** & **atemp** has the highest correlation with the target variable 'cnt'.

4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

- **There is a linear relationship between X and Y:** We found from the pair plot that the numberical variables are in linear relationship with the target variable (cnt)
- **Error terms are normally distributed with mean zero(not X, Y):**

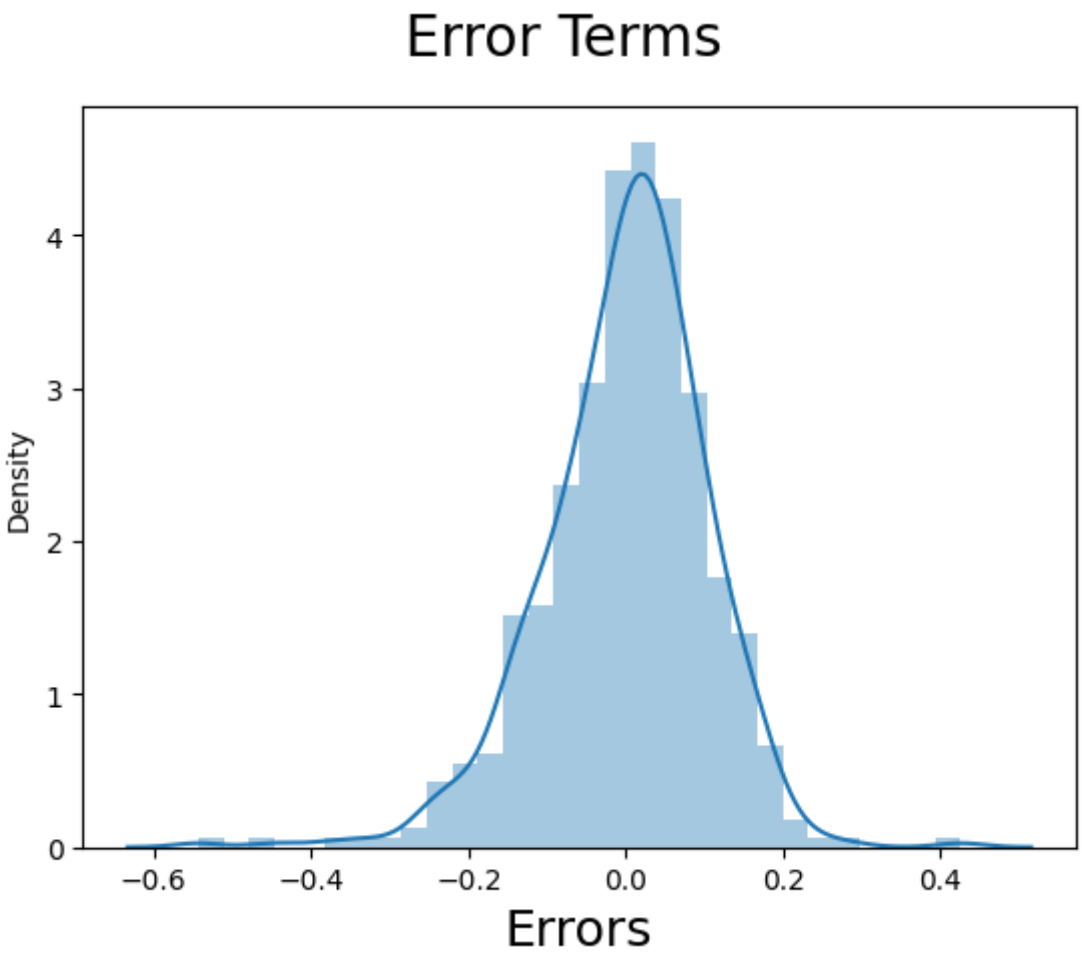


Figure: Error Terms plot

We can observe that the error terms are normally distributed with mean close to 0

- **Error terms are independent of each other and Error terms have constant variance (homoscedasticity):**

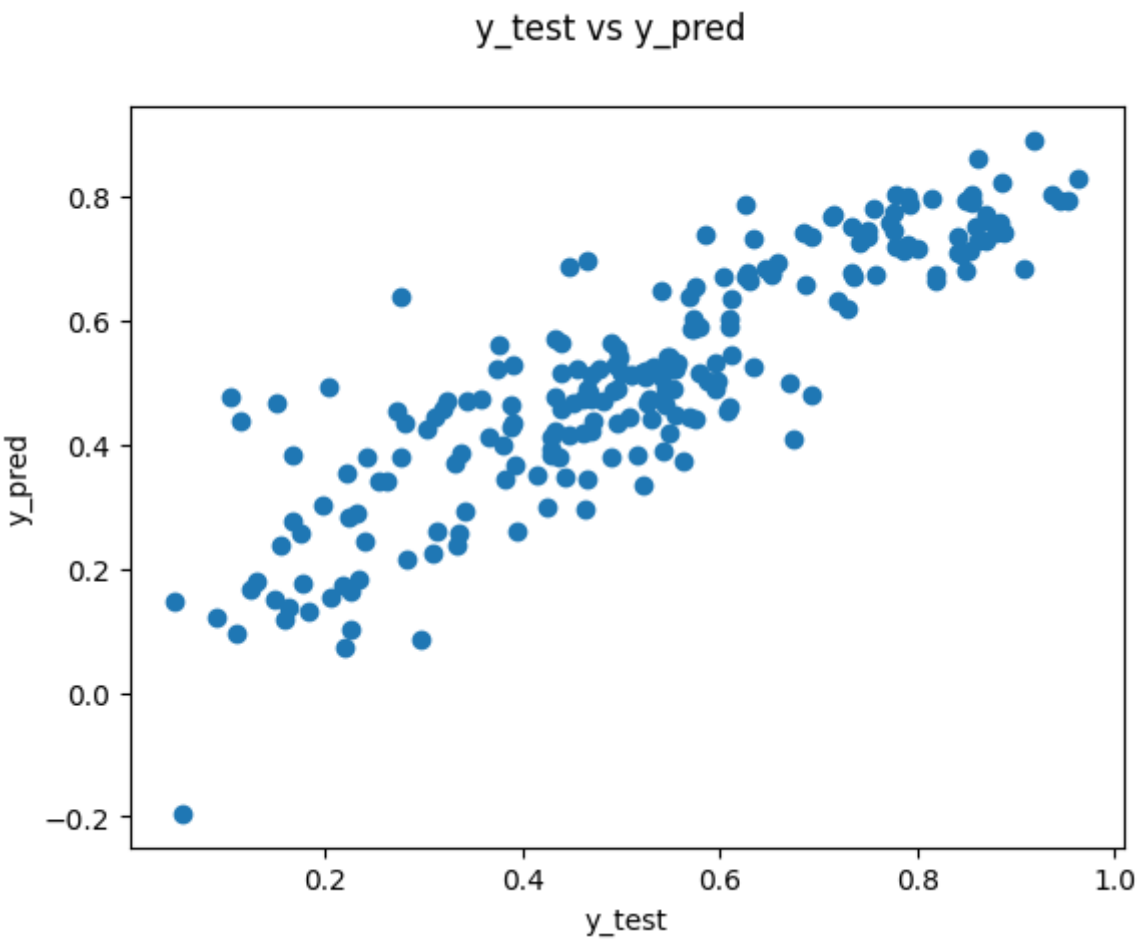


Figure: Plot of y_test vs y_pred

We can see that the variance remains similar as the values change.

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

From the final equations of the best fitted line i.e.,

```
 Count = 0.59 * const + 0.25 * yr + -0.19 * windspeed + -0.26 * seasonspring
+ -0.04 * seasonsummer + -0.11 * seasonwinter + -0.1 * mnthJan + -0.01 *
mnthJul + 0.1 * mnthOct + 0.08 * mnthSep + 0.01 * weekdaySaturday + -0.05 *
weekdaySunday + -0.32 * weathersitlightrainsnow + -0.09 * weathersitmisty
```

We can tell that the top 3 features contributing significantly are:

- weathersit_light_rain_snow (-0.32)
- season_spring (-0.26)
- yr (0.25)

General Subjective Questions

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

Linear regression 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.

Types of Linear Regression

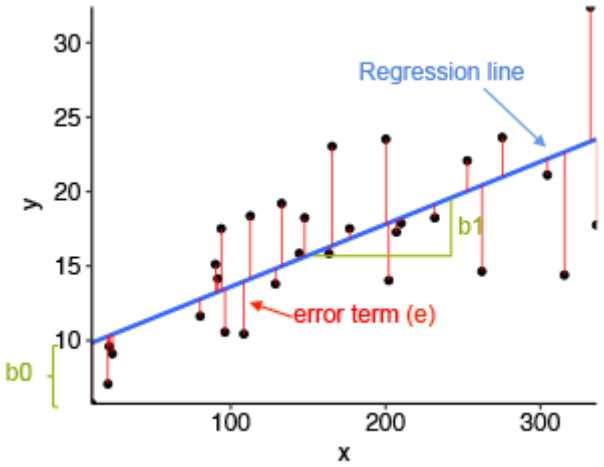
Linear regression can be further divided into two types of the algorithm:

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.

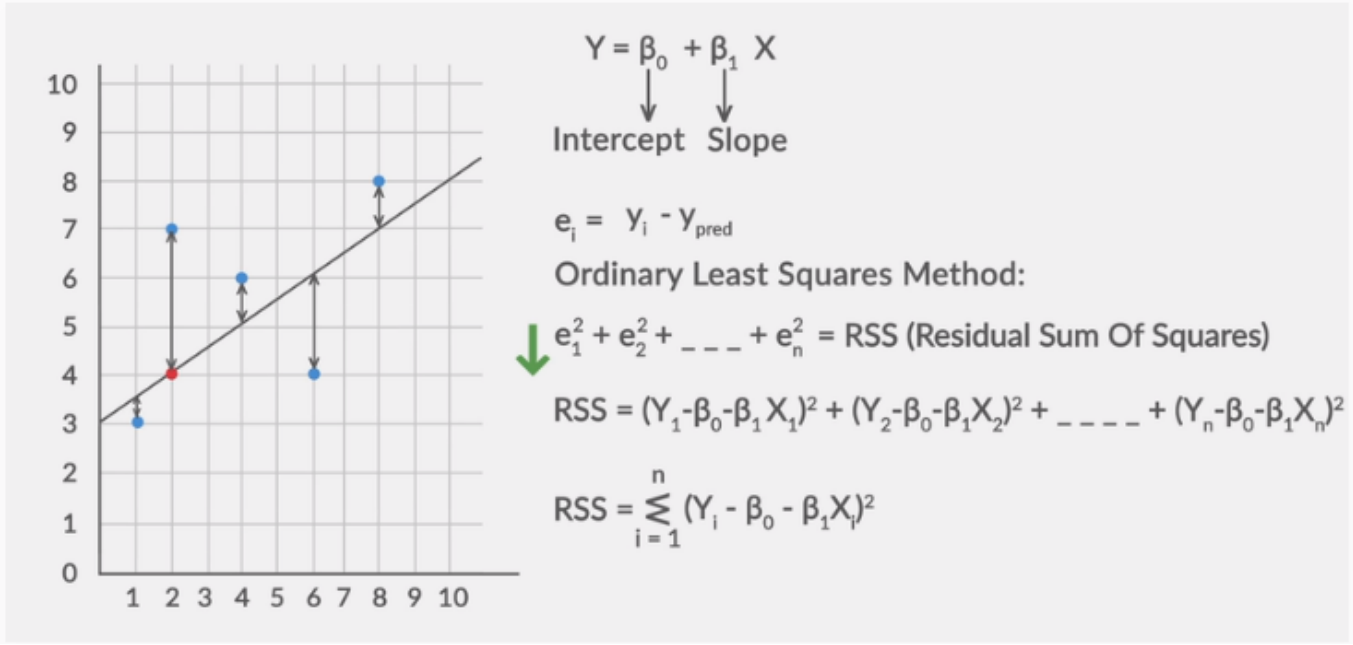
Linear regression finds the straight line, called the **least squares regression line** or LSRL, that best represents observations in a bivariate data set. Suppose Y is a dependent variable, and X is an independent variable. The population regression line is: A linear regression line has an equation of the form:

$$Y = \beta 0 + \beta 1X$$



RSS (Residual Sum of Squares)

The best fit line is derived by minimizing the **RSS (Residual Sum of Squares)** which is the cost function in this case. Residuals for any data point is found by subtracting predicted value of dependent variable from actual value of dependent variable:



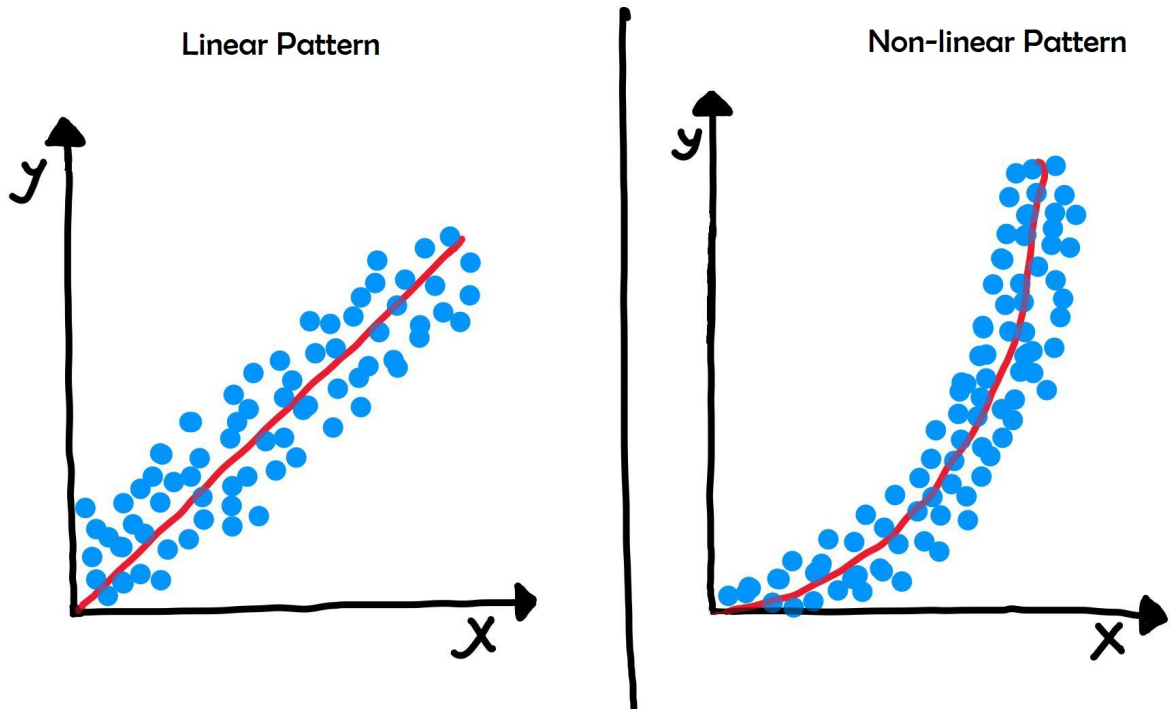
Gradient Descent

Gradient is one optimization method which can be used to optimize the Residual sum of squares cost function. There can be other cost functions. Basically it starts with an initial value of β0 and β1 and then finds the cost function. It then increases or decreases the parameters to find the next cost function value. This is done till a minima is found. Gradient descent expects that there is no local minimal and the graph of the cost function is convex.

Assumptions of Simple Linear Regression

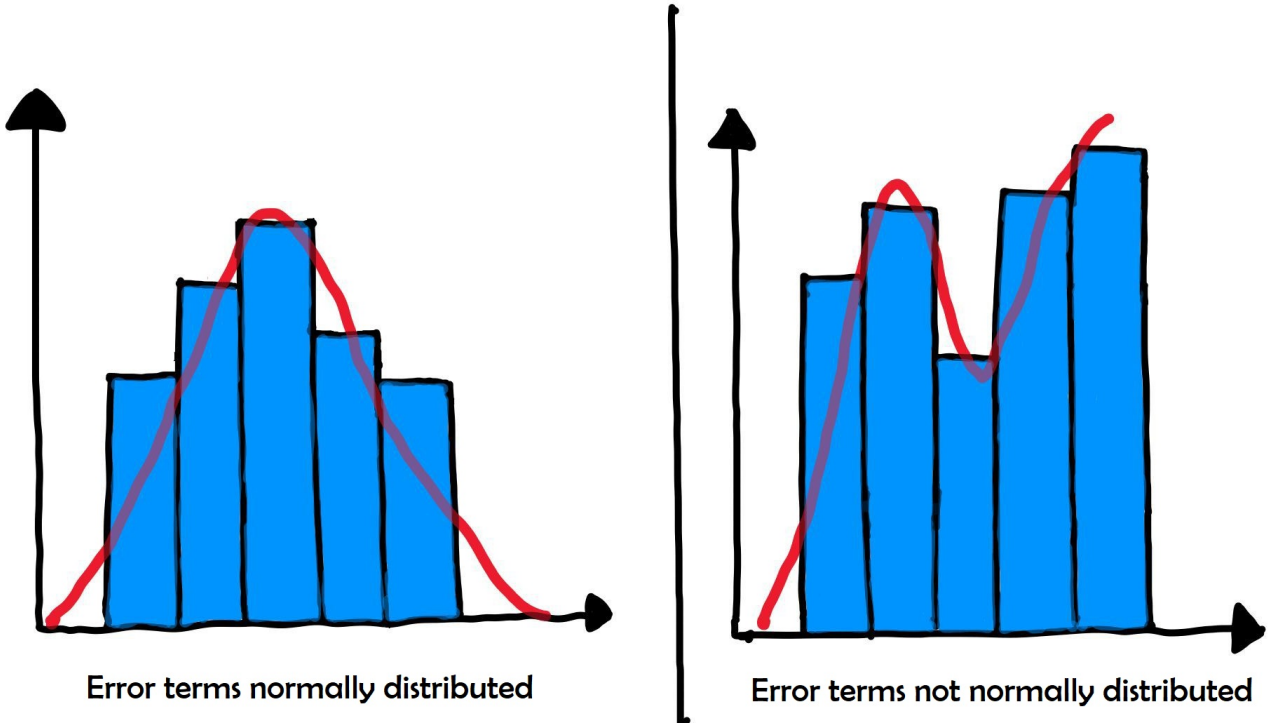
There is a linear relationship between X and Y

X and Y should display some sort of a linear relationship; otherwise, there is no use of fitting a linear model between them.



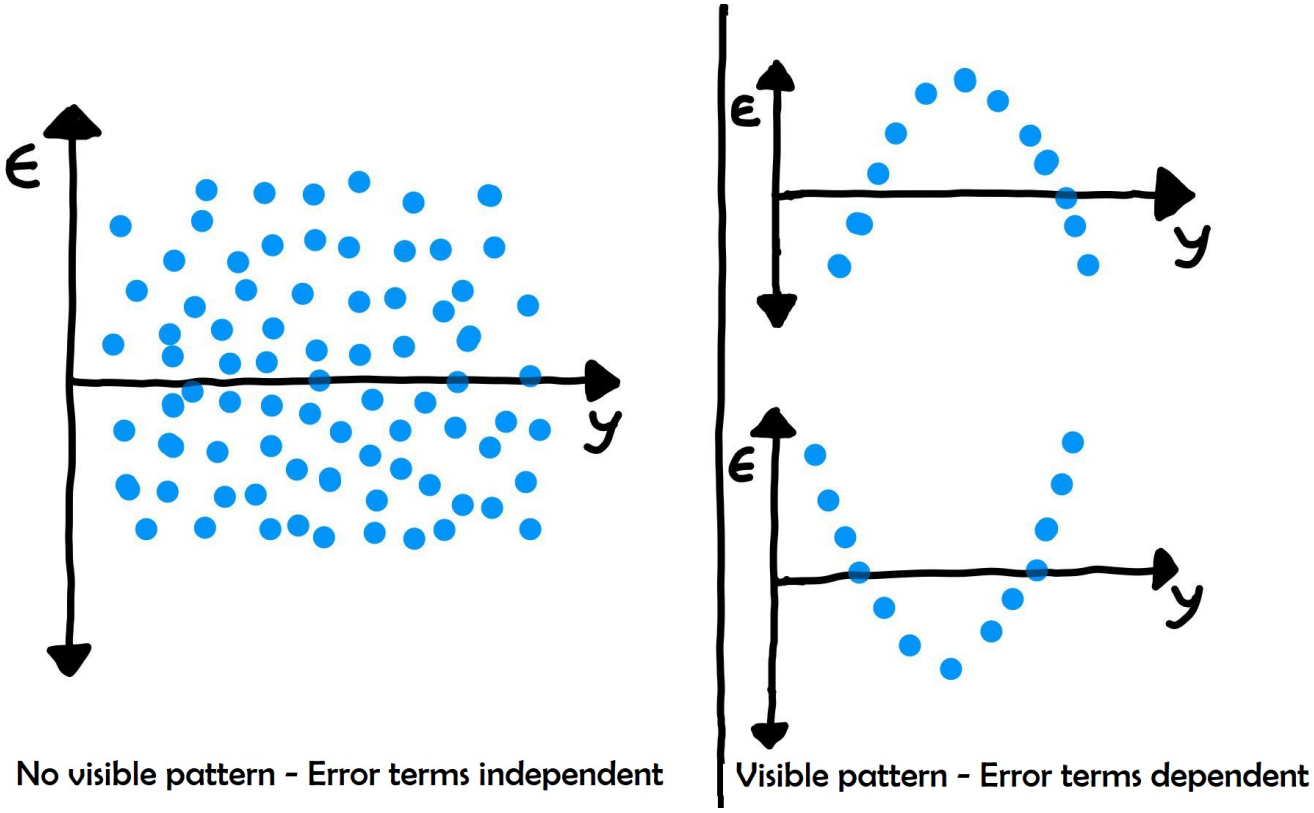
Error terms are normally distributed with mean zero(not X, Y)

The residuals should follow a normal distribution. Once you obtain the residuals from your model, this is relatively easy to test using either a histogram or a QQ Plot.



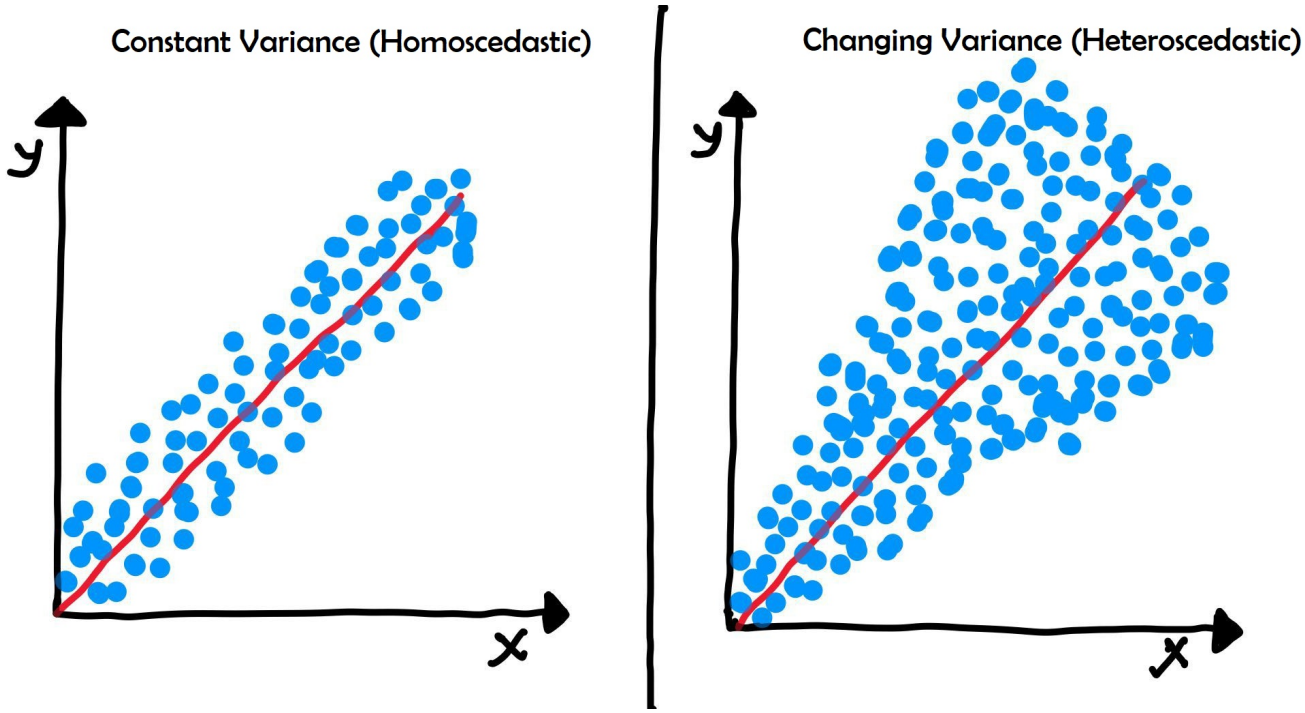
Error terms are independent of each other

The error terms should not be dependent on one another (like in a time-series data wherein the next value is dependent on the previous one).



Error terms have constant variance (homoscedasticity)

The variance should not increase (or decrease) as the error values change.



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.

The eqaution of Multiple Linear Regression (MLR) can be seen below:

Multiple Linear Regression

$$\gamma = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \varepsilon$$

Considerations

Overfitting

When we add more and more variables, the model might end up memorizing all the data points in the training set. This will cause major problems with the generalization on unseen data.

Multicollinearity

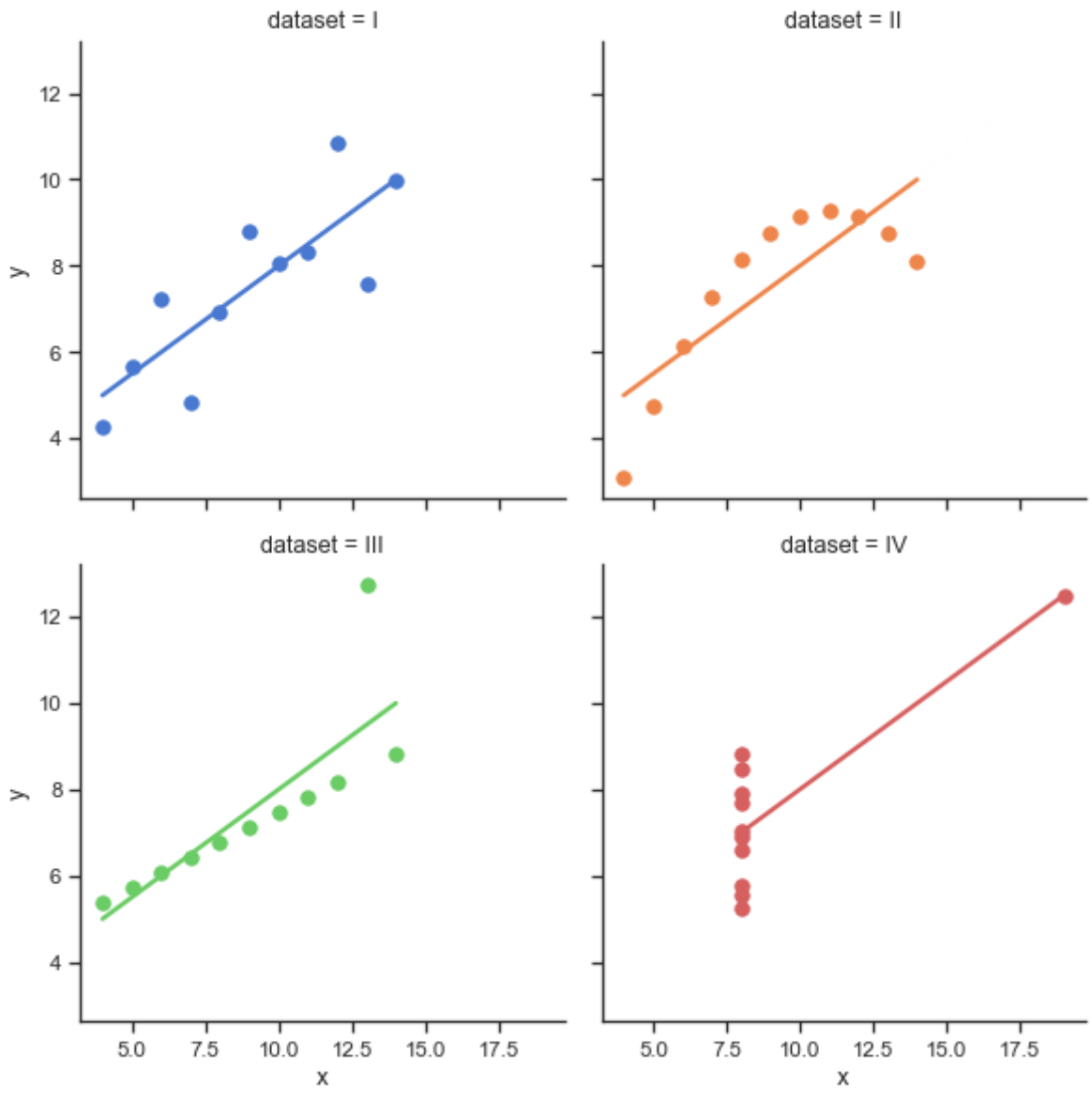
A model that has been built using several independent variables, some of these variables might be interrelated, i.e. some of these variables might completely explain some other independent

variable in the model due to which the presence of that variable in the model is redundant. This affects the model interpretation and coefficients of variables.

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

Anscombe's Quartet can be defined as a group of four data sets which are nearly identical in 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.

It was constructed in 1973 by statistician **Francis Anscombe** to illustrate the importance of plotting the graphs before analyzing and model building, and the effect of other observations on statistical properties.There are these four data set plots which have nearly same statistical observations, which provides same statistical information that involves variance, and mean of all x,y points in all four datasets.



This tells us about the importance of visualising the data before applying various algorithms out there to build models out of them which suggests that the data features must be plotted in

order to see the distribution of the samples that can help you identify the various anomalies present in the data like outliers, diversity of the data, linear separability of the data, etc. Also, the Linear Regression can be only be considered a fit for the data with linear relationships and is incapable of handling any other kind of datasets. These four plots can be defined as follows:

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

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

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
				Summary Statistics							
N	11	11		11	11		11	11		11	11
mean	9.00	7.50		9.00	7.500909		9.00	7.50		9.00	7.50
SD	3.16	1.94		3.16	1.94		3.16	1.94		3.16	1.94
r	0.82			0.82			0.82			0.82	

When these models are plotted on a scatter plot, all datasets generates a different kind of plot that is not interpretable by any regression algorithm which is fooled by these peculiarities and can be seen in the above graph.

The four datasets can be described as:

- **Dataset 1:** this fits the linear regression model pretty well.
- **Dataset 2:** this could not fit linear regression model on the data quite well as the data is non-linear.
- **Dataset 3:** shows the outliers involved in the dataset which cannot be handled by linear regression model
- **Dataset 4:** shows the outliers involved in the dataset which cannot be handled by linear regression model

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

Pearson correlation coefficient, also known as Pearson R statistical test, measures the strength between the different variables and their relationships. Therefore, whenever any statistical test is conducted between the two variables, it is always a good idea for the person analyzing to calculate the value of the correlation coefficient to know how strong the relationship between the two variables is.

The Pearson Correlation Coefficient **formula** is as follows:

$$\mathbf{r} = \frac{\mathbf{n}(\sum \mathbf{xy}) - (\sum \mathbf{x})(\sum \mathbf{y})}{\sqrt{[\mathbf{n}\sum \mathbf{x^2} - (\sum \mathbf{x})^2][\mathbf{n}\sum \mathbf{y^2} - (\sum \mathbf{y})^2]}}$$

The Pearson correlation coefficient represents the relationship between the two variables, measured on the same interval or ratio scale. It measures the strength of the relationship between the two continuous variables.

The coefficient not only states the presence or absence of the correlation between the two variables but also determines the exact extent to which those variables are correlated. It is independent of the unit of measurement of the variables where the values of the correlation coefficient can range from the value +1 to the value -1. However, it is insufficient to tell the difference between the dependent and independent variables.

It is independent of the unit of measurement of the variables. For example, suppose the unit of measurement of one variable is in years while the unit of measurement of the second variable is in kilograms. In that case, even then, the value of this coefficient does not change.

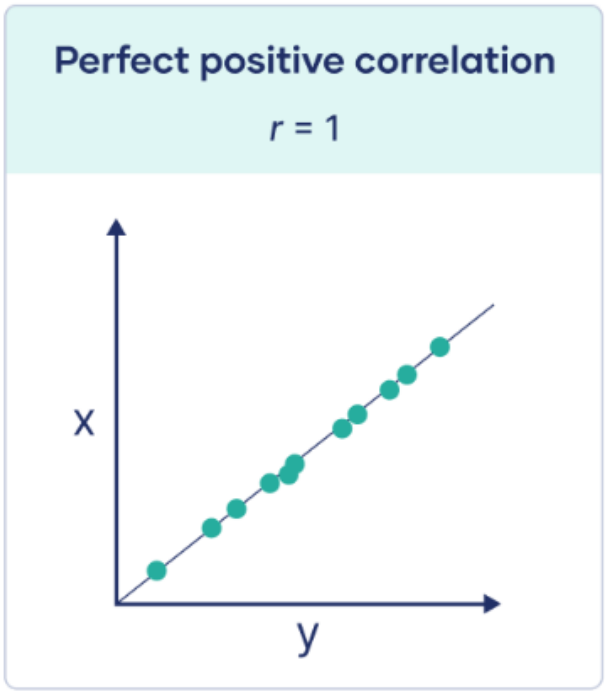
The correlation coefficient between the variables is symmetric, which means that the value of the correlation coefficient between Y and X or X and Y will remain the same.

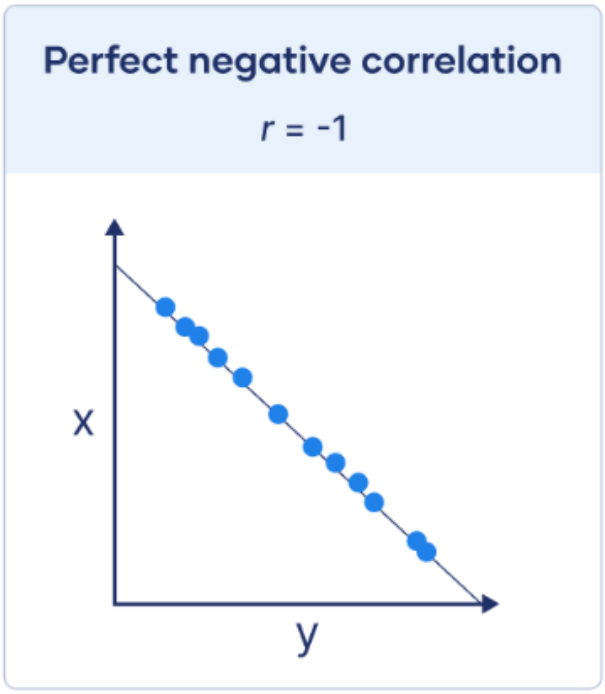
Visualizing the Pearson correlation coefficient

Another way to think of the Pearson correlation coefficient (r) is as a measure of how close the observations are to a line of best fit.

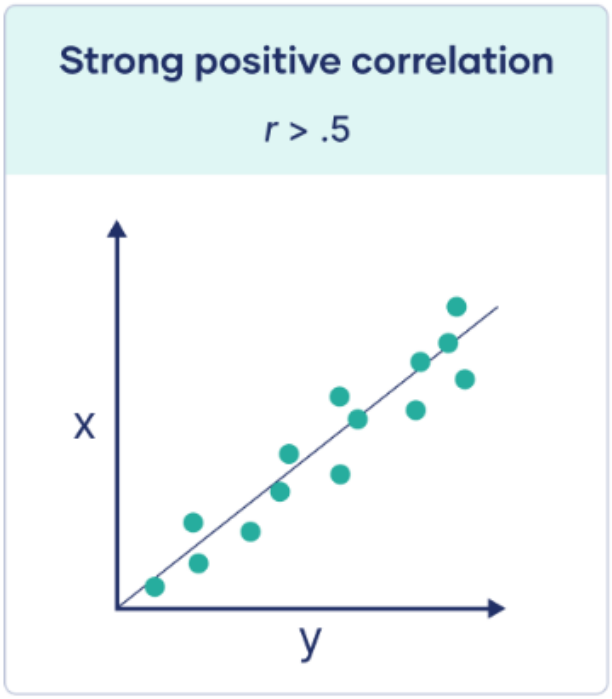
The Pearson correlation coefficient also tells you whether the slope of the line of best fit is negative or positive. When the slope is negative, r is negative. When the slope is positive, r is positive.

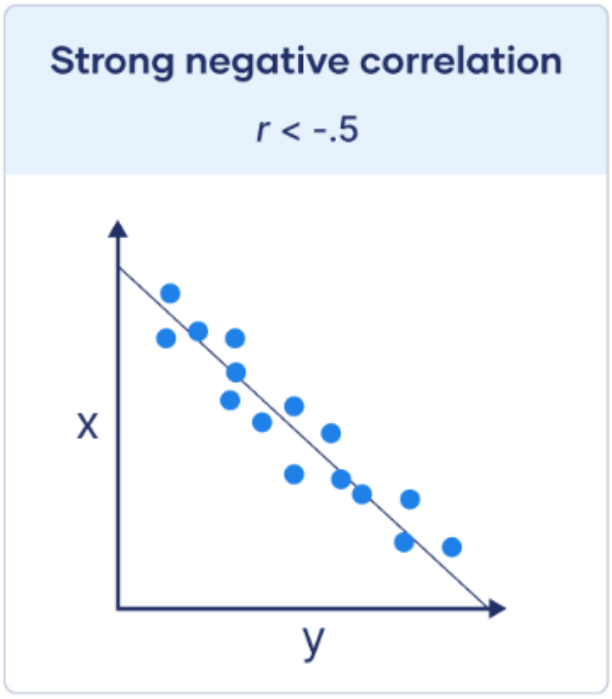
- When r is 1 or –1, all the points fall exactly on the line of best fit:



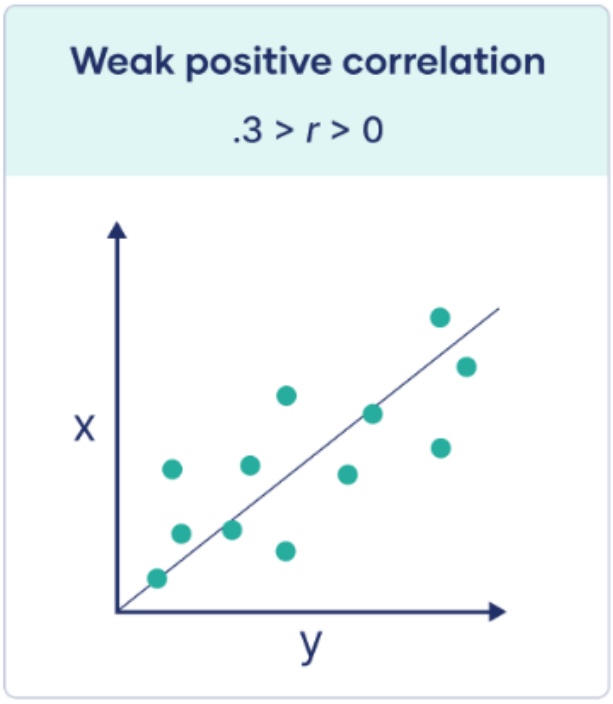


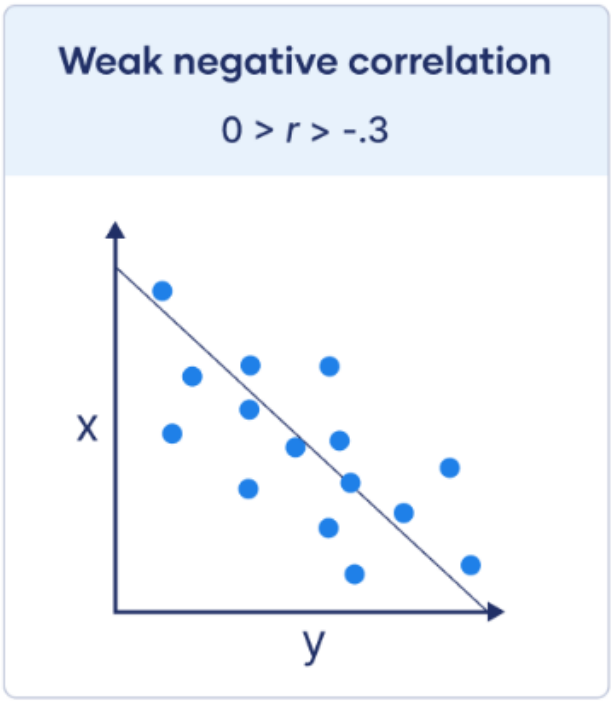
- When r is greater than .5 or less than –.5, the points are close to the line of best fit:



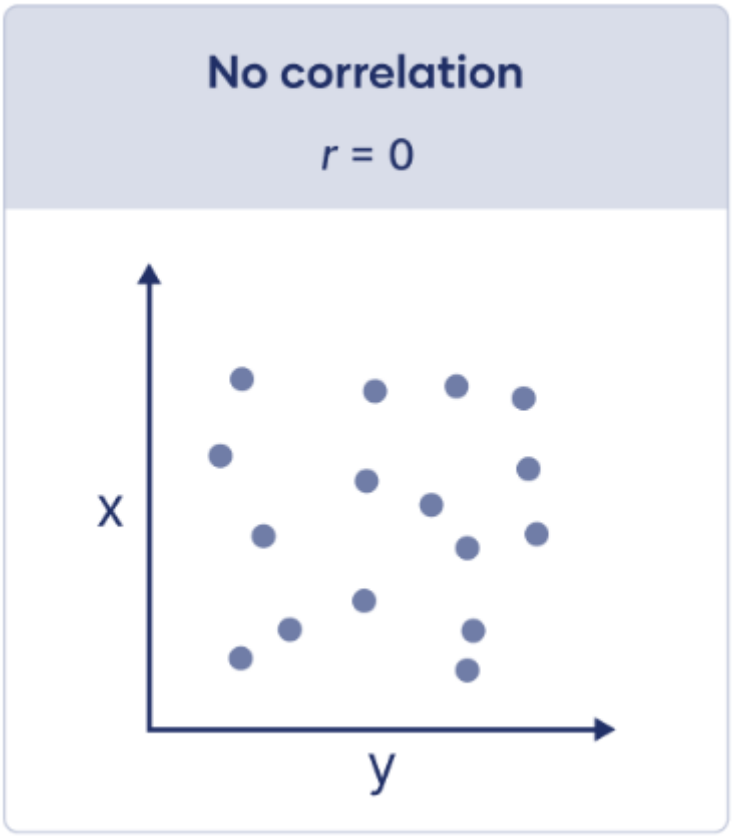


- When r is between 0 and .3 or between 0 and –.3, the points are far from the line of best fit:





- When r is 0, a line of best fit is not helpful in describing the relationship between the variables:



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

What is scaling?

Feature scaling is a data preprocessing technique used to transform the values of features or variables in a dataset to a similar scale. The purpose is to ensure that all features contribute equally to the model and to avoid the domination of features with larger values.

Feature scaling becomes necessary when dealing with datasets containing features that have different ranges, units of measurement, or orders of magnitude. In such cases, the variation in feature values can lead to biased model performance or difficulties during the learning process.

Why is scaling performed?

Some machine learning algorithms are sensitive to feature scaling, while others are virtually invariant. Let's explore these in more depth:

Gradient Descent Algorithm

Machine learning algorithms like linear regression, logistic regression using this algorithm as their basic function. Gradient descent is a first-order iterative optimization algorithm for finding a local minimum of a different function. Basically in any algorithm, the gradient descent function slides through the data set while applied to the data set, step by step. So if the distance between the data points increases the size of the step will change and the movement of the function will not be smooth. Take a look at the formula.

$$\theta_j := \theta_j - \alpha \frac{1}{m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)}) x_j^{(i)}$$

So if the distance range between feature values increases the movement will increase and function will not work properly. In that situation, we will be required to have a data set well rescaled so that the function can better help in the development of the machine learning model.

Distance Based-Algorithm

Distance algorithms like KNN, K-means clustering, and SVM(support vector machines) are most affected by the range of features. This is because, behind the scenes, they are using distances between data points to determine their similarity.

For example, in a corporate office the salary of the employees are totally dependent on the experience and there are people who are newcomers and some are well experienced and some of those have medium experience. We need to make a model which can predict the salary and if the number of employees of any class is more then the model will be prone to that class of employees to prohibit the situation. We need to rescale the data so the data is well spread in the space and algorithms can learn better from it.

What is the difference between normalized scaling and standardized scaling?

Normalization is a suitable choice when your data's distribution does not match a Gaussian distribution. A practical transformation approach that helps your model perform and be more accurate is normalization. Normalization of a machine learning model is helpful when you are unsure about the precise feature distribution. To put it another way, the data's feature distribution does not have a Gaussian distribution. Outliers in your data will be impacted by normalization because it needs a wide range to function correctly.

When you are entirely aware of the feature distribution of your data, or, to put it another way, when your data has a Gaussian distribution, standardization in the machine learning model is useful. This need not necessarily be the case, though. In contrast to Normalization, Standardization does not always have a bounding range; therefore, any outliers in your data won't be impacted by it.

Scales for normalization fall between [0,1] and [-1,1]. Standardization has no range restrictions. When the algorithms don't make any assumptions about the distribution of the data, Normalization is taken into account. When algorithms create predictions about the data distribution, standardization is applied.

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

The VIF value can be infinite if there is perfect correlation between two independent variables. This means that the two variables are perfectly related to each other, and knowing one variable perfectly predicts the other. In this case, the R-squared of the regression of one variable on the other will be 1, and the VIF will be infinite.

For example, consider the following two variables:

- `height` in centimeters
- `height` in inches
  These two variables are perfectly correlated, because one can be perfectly converted to the other. Therefore, the VIF of height in inches will be infinite.

In general, a VIF value of infinity indicates that there is perfect multicollinearity between two variables. This can be a problem for regression models, because it can make the coefficients of the model unstable and unreliable. If you see a VIF value of infinity, you should consider removing one of the variables from the model.

Here are some other reasons why the VIF value might be infinite:

- The model has too few observations.
- The variables are not scaled correctly.
- There is a dummy variable trap.

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

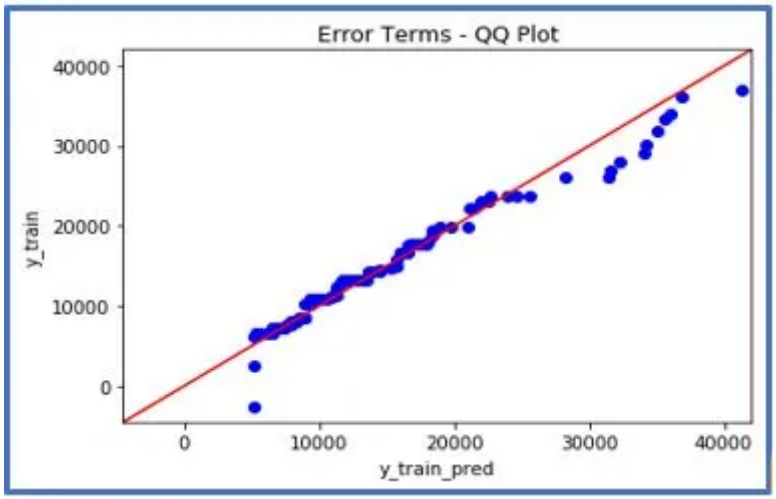
The quantile-quantile (q-q) plot is a graphical technique for determining if two data sets come from populations with a common distribution.

For example, if you are testing if the distribution of age of employees in your team is normally distributed, you are comparing the quantiles of your team members' age vs quantile from a normally distributed curve. If two quantiles are sampled from the same distribution, they should roughly fall in a straight line.

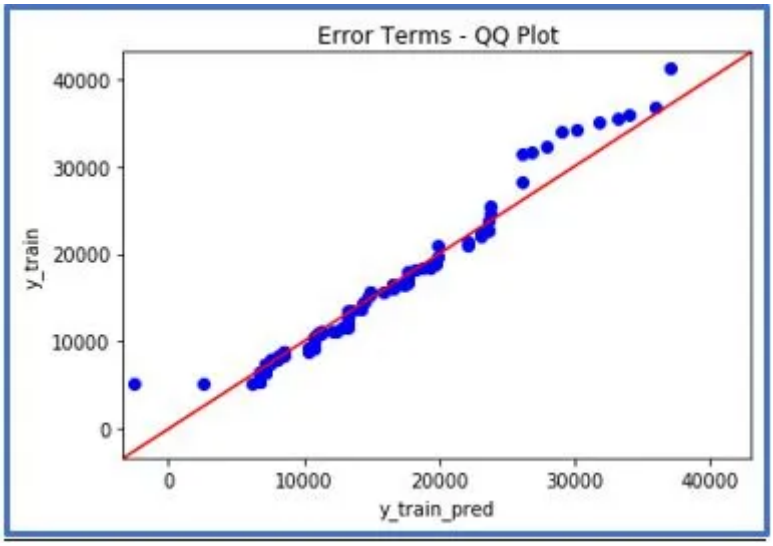
A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set.

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