# 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)**

   Based upon the analysis of the categorical variables from the dataset, we can infer the below points about their effect on the dependent variable –

   - **season:** Fall season seems to have attracted more booking (with a median of over 5000 booking) followed by summer and winter. This indicates, season can be a good predictor for the dependent variable.
   - **year:** 2019 attracted fairly a greater number of bookings as compared to the previous year 2018 (median roughly shot from 4000 to 6000), which shows good progress in terms of business.
   - **month:** Most of the bookings were during the month of May, June, July, August, September, and October (across 2018 and 2019) with a median of 5000 or above booking per month. This indicates, month has some trend for bookings and can be a good predictor for the dependent variable.
   - **weekday:** Weekday does not give a clear picture about demand. Weekday variable shows very close trend having their independent medians between 4000 to 5000 bookings. This variable can have some or no influence towards the predictor. We would decide its influence on the booking demand based upon our model evaluation. Otherwise, the spread of booking is more during Monday, Friday, and Sunday.
   - **workingday:** The demand for bike rental seems to be almost invariable whether it is a working day or not. The median booking is almost 4500 in both working and non-working days. This indicates, workingday cannot be a good predictor for the dependent variable.
   - **weathersit:** Clear weather attracted more booking which seems obvious followed by misty weather. This indicates, weathersit does show some trend towards the bike bookings can be a good predictor for the dependent variable.

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

   Setting drop_first = True causes get_dummies to exclude the dummy variable for the first category of the variable we are operating on. When we have a categorical variable with N mutually exclusive categories, we actually only need N – 1 new dummy variables to encode the same information. This is because if all the existing dummy variables equal 0, then implicitly the value should be 1 for the remaining dummy variable. For example, if season_Summer == 0, and season_Winter == 0, and season_Autumn == 0, then season_Spring must equal 1. This is implied by the existing 3 dummy variables, so we don't need the 4th variable.

   The extra dummy variable literally contains redundant information. 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.

   So, it's a common convention to drop the dummy variable for the first level of the categorical variable that we are encoding.

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

   The numeric variable ‘temp’ has the highest correlation with the target variable ‘cnt’ with a value of 0.63. This is evaluated using heatmap in python notebook.

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

   I have validated the assumption of Linear Regression Model based on below 5 assumptions -

   1. **Linear relationship** is visible between X and Y
   2. Error terms are **normally distributed** (not X, Y) – We validated the assumption of Linear regression by plotting a distplot of the residuals and analysing it to see if it is a normal distribution or not if it has mean equals 0. The diagram below shows that it is normally distributed with mean equals 0.
   3. **Multicollinearity check** - There should be insignificant multicollinearity among variables.
   4. Error terms are independent of each other. **No autocorrelation**.
   5. Error terms have constant variance (**homoscedasticity**). There is no visible pattern.

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

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

   - **Year -** A coefficient of 0.247992 indicates that a unit increase in 'year' variable, increases the bike hire numbers by 0.247992 units. Based on previous data it is expected to have a boom in number of users once situation comes back to normal, compared to 2019.
   - **Light_snowrain -** A coefficient of -0.303393 indicates that w.r.t weathersit_1 (clear weather), a unit increase in Light_snowrain variable decreases the bike hire numbers by 0.303393 units. There would be less bookings during Light Snow or Rain. So the company could probably use this time to service the bikes without having business impact.
   - **Spring -** A coefficient of -0.258069 indicates that, w.r.t season_1 (fall), a unit increase in spring variable, decreases the bike hire numbers by 0.258069 units. So the company should not focus on expanding business during Spring.

## General Subjective Questions

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

   Linear regression algorithm is a statistical technique which is used to do predictive analysis. The algorithm is used to create a model that explains the linear relationship between a dependent variable a set of independent variables. By predictive analysis and linear relationship, we mean the following-

   (1) if a set of input (independent) variables can fairly predict the value of an outcome (dependent) variable.

(2) which input variables are significant predictors of the outcome variable and how. By significance we mean the magnitude and by how we mean whether the independent variable positively or negatively impact the value of the outcome variable.

Mathematically the linear relationship can be represented with the help of following equation –

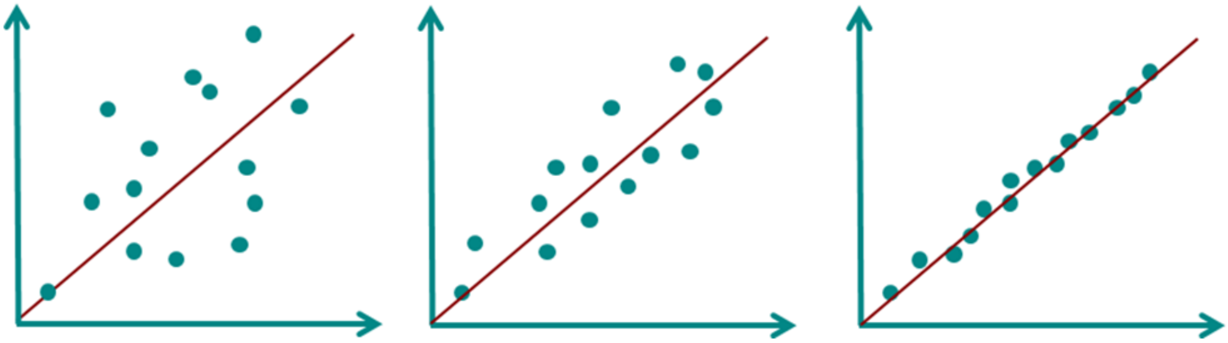
**Y = mX + c**

where, Y is the dependent variable which needs to be predicted.

X is the independent variable which will be used to make predictions.

m is the slope of the regression line (also called regression coefficient) which indicates the significant effect X has on Y

c is a constant, known as the Y-intercept. If X = 0, Y will be equal to c.

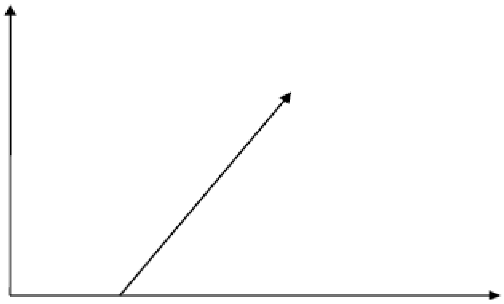


Depending on whether there is one or more independent variables, there are **simple** and **multiple** linear regression algorithms. A **simple linear regression algorithm** analyses the influence of an independent variable on a dependent variable. A **multiple linear regression algorithm** analyses the influence of several independent variables on a dependent variable.

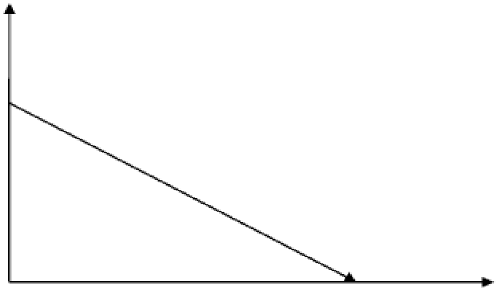
The greater the linear relationship between the output and input variables, the more accurate the prediction is. The stronger the linear relationship is, greater proportion of the variance in the dependent variable can be explained by the independent variable. Visually, the relationship between the variables can be shown in a scatter diagram as above. The greater the linear relationship between the dependent and independent variables, the more the data points lie on a straight line.

The regression coefficient m can have different signs, which can be interpreted as follows:

- m > 0: there is a positive relationship between x and y (larger value of x implies larger value of y)



- m < 0: there is a negative relationship between x and y (larger value of x implies smaller value of y)



- m = 0: there is no relationship between x and y

The following are some assumptions about dataset that is made by Linear Regression model –

- **Multi-collinearity –** Linear regression model assumes that there is very little or no multi-collinearity in the data. Basically, multi-collinearity occurs when the independent variables or features have relationship in them.
- **Auto-correlation –** There is very little or no auto-correlation in the data. Basically, Autocorrelation occurs when the residuals are not independent from each other. For instance, this typically occurs in stock prices, where the price is not independent from the previous price.
- **Relationship between variables –** The relationship between predicted and feature variables must be linear.
- **Normality of error terms –** Error terms should be normally distributed.
- **Homoscedasticity –** The variance of the residual, or error term, is constant. There should be no visible pattern in residual values. This also implies that variance is roughly the same across all data points.

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

**Anscombe's Quartet** is the modal example to illustrate the importance of exploratory data analysis using data plotting and visualization and the drawbacks of depending solely on summary statistics. This was developed by the statistician Francis Anscombe in 1973. His findings intended to counter the general understanding among statisticians that "numerical calculations are exact, but graphs are rough".

The example emphasizes the importance of using data visualization to spot trends, outliers, and other crucial details out of a dataset that might not be obvious from summary statistics alone.

It comprises of four datasets and each dataset consists of eleven (x, y) points. The basic thing to analyse about these datasets is that they all have identical descriptive statistics, viz., mean, variance, correlations, standard deviation etc. but different graphical representation. Each graph plot shows the different behaviour irrespective of statistical analysis. When plotted, each dataset reflects unique relation between x and y, in terms of variability patterns and correlation strengths. Despite these variations, each dataset has the same summary statistics, such as the same x and y mean and variance, x and y correlation coefficient, and linear regression line.

The four datasets of Anscombe's quartet are listed in the below table.

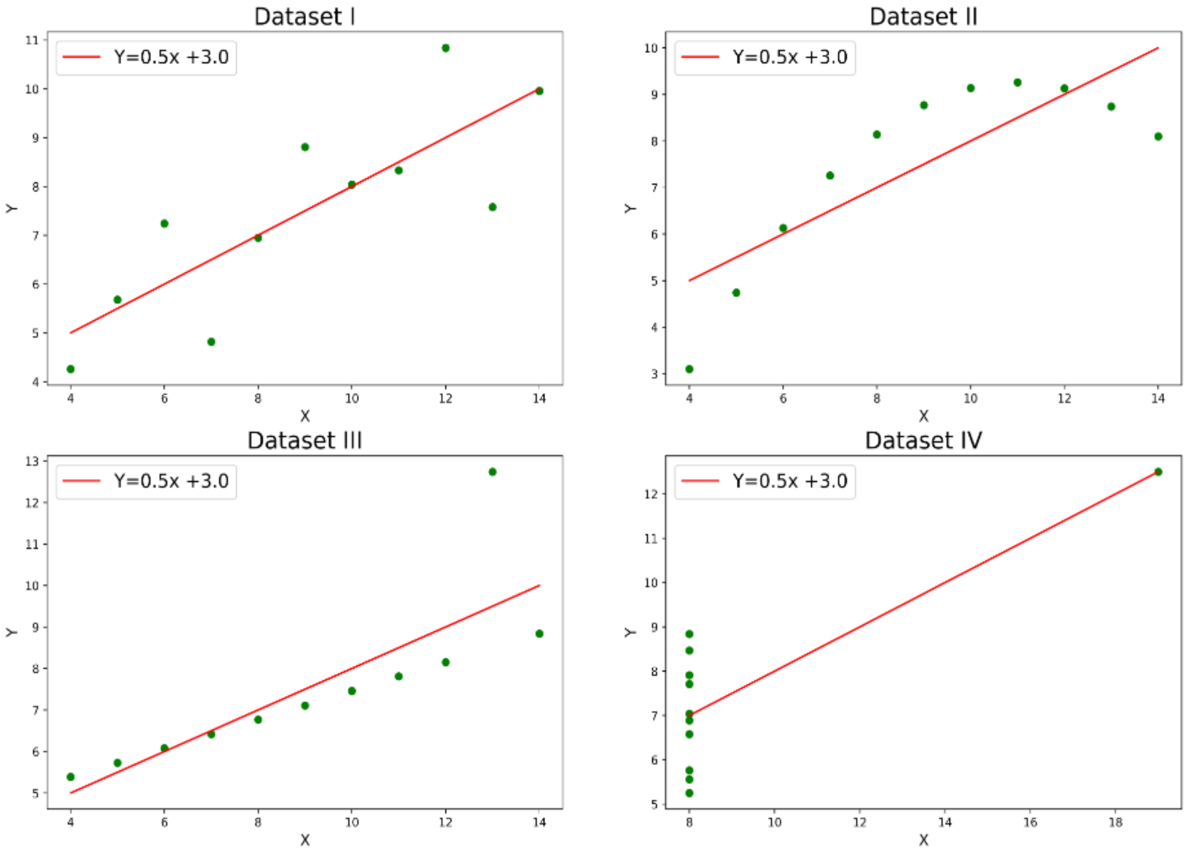
| x1 | x2 | x3 | x4 | y1 | y2 | y3 | y4 |
|---|---|---|---|---|---|---|---|
| 10 | 10 | 10 | 8 | 8.04 | 9.14 | 7.46 | 6.58 |
| 8 | 8 | 8 | 8 | 6.95 | 8.14 | 6.77 | 5.76 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 13 | 13 | 13 | 8 | 7.58 | 8.74 | 12.74 | 7.71 |
| 9 | 9 | 9 | 8 | 8.81 | 8.77 | 7.11 | 8.84 |
| 11 | 11 | 11 | 8 | 8.33 | 9.26 | 7.81 | 8.47 |
| 14 | 14 | 14 | 8 | 9.96 | 8.1 | 8.84 | 7.04 |
| 6 | 6 | 6 | 8 | 7.24 | 6.13 | 6.08 | 5.25 |
| 4 | 4 | 4 | 19 | 4.26 | 3.1 | 5.39 | 12.5 |
| 12 | 12 | 12 | 8 | 10.84 | 9.13 | 8.15 | 5.56 |
| 7 | 7 | 7 | 8 | 4.82 | 7.26 | 6.42 | 7.91 |
| 5 | 5 | 5 | 8 | 5.68 | 4.74 | 5.73 | 6.89 |

The following are the descriptive statistical properties for all four datasets.

| **Descriptive Statistics** | **I** | **II** | **III** | **IV** |
|---|---|---|---|---|
| Mean_x | 9 | 9 | 9 | 9 |
| Variance_x | 11.000 | 11.000 | 11.000 | 11.000 |
| Mean_y | 7.500909 | 7.500909 | 7.5 | 7.500909 |
| Variance_y | 4.127269 | 4.127629 | 4.12262 | 4.123249 |
| Correlation | 0.816421 | 0.816237 | 0.816287 | 0.816521 |
| Linear Regression slope | 0.500091 | 0.5 | 0.499727 | 0.499909 |
| Linear Regression intercept | 3.000091 | 3.000909 | 3.002455 | 3.001727 |

Below is the scatter plot and linear regression line for each dataset.



Below is an explanation of this graphical plot of each dataset:

- In the first one (top left), the scatter plot illustrates a simple linear relationship between x and y.
- In the second one (top right), the scatter plot illustrates a non-linear relationship between x and y.
- In the third one (bottom left), the scatter plot illustrates a perfect linear relationship between x and y for all the data points except one which seems to be an outlier as indicated far away from that line.
- Finally, the fourth one (bottom right) illustrates one outlier which is enough to produce a high correlation coefficient, even though the other data points do not indicate any relationship between the variables.

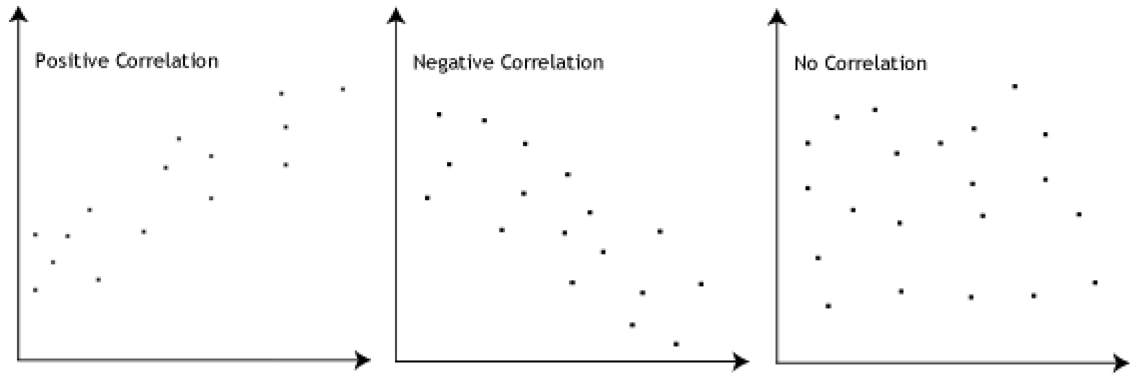
### 3. What is Pearson's R? (3 marks)

The Pearson correlation coefficient is a descriptive statistic. It summarizes the characteristics of a dataset by specifically describing the strength and direction of the linear relationship between two quantitative variables. So, it is also an inferential statistic that can be used to test statistical hypotheses.

**Values of r:** Denoted by r, Pearson correlation coefficient can take a range of values from +1 to -1.

| **Pearson correlation coefficient (r)** | **Correlation type** | **Interpretation** |
|---|---|---|
| Between 0 and 1 | Positive correlation | As the value of one variable increases, the value of the other variable also increases. |
| 0 | No correlation | There is no relationship between the variables. |
| Between 0 and –1 | Negative correlation | As the value of one variable increases, the value of the other variable decreases. |

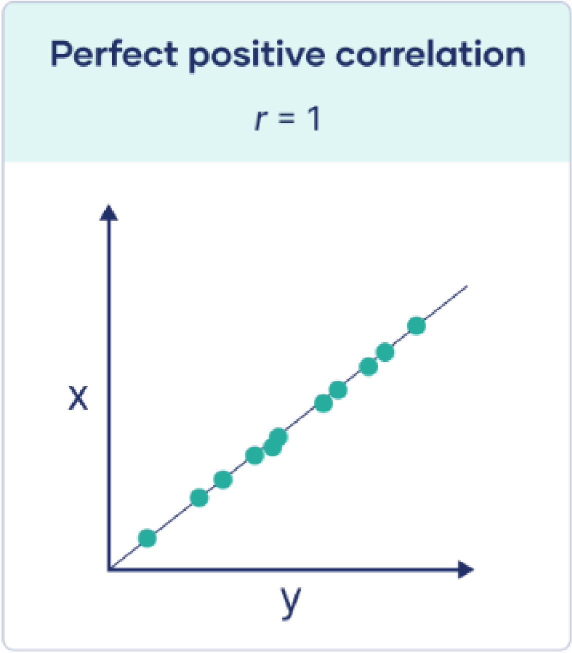
Below graphs further illustrate the different types of correlation.

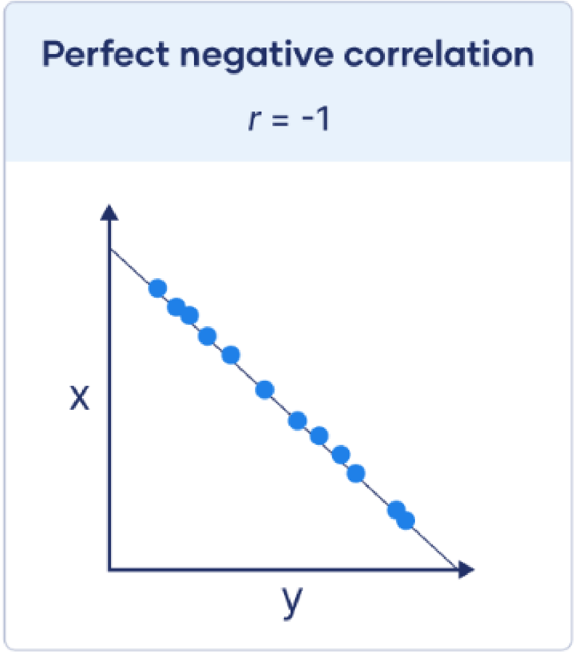


**Visualizing the Pearson correlation coefficient:** Pearson correlation coefficient (r) attempts to draw a line of best fit through the data of two variables. It is a measure of how close the observations are to a line of best fit. It tells whether the slope of the line of best fit is negative or positive.

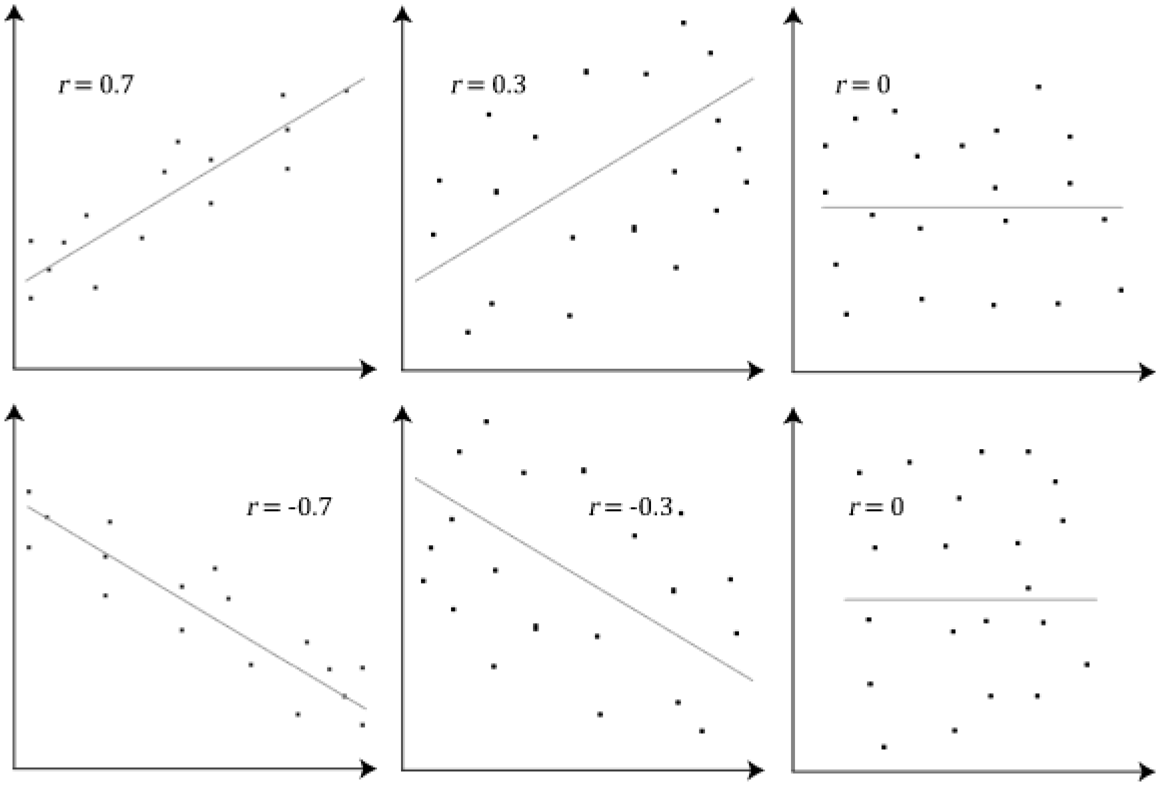
- When the slope is negative, r is negative.
- When the slope is positive, r is positive.

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





**Strength of association based on r:** The stronger the association of the two variables, the closer the value of r, will be to either +1 or -1 depending on whether the relationship is positive or negative, respectively. A value of +1 or -1 means that all data points are included on the line of best fit, i.e., there are no data points that lie away from this line. Values of r between +1 and -1 indicate that there is variation around the line of best fit. The closer the value of r to 0, the greater the variation around the line of best fit. Different relationships and their correlation coefficients are shown in the diagram below:



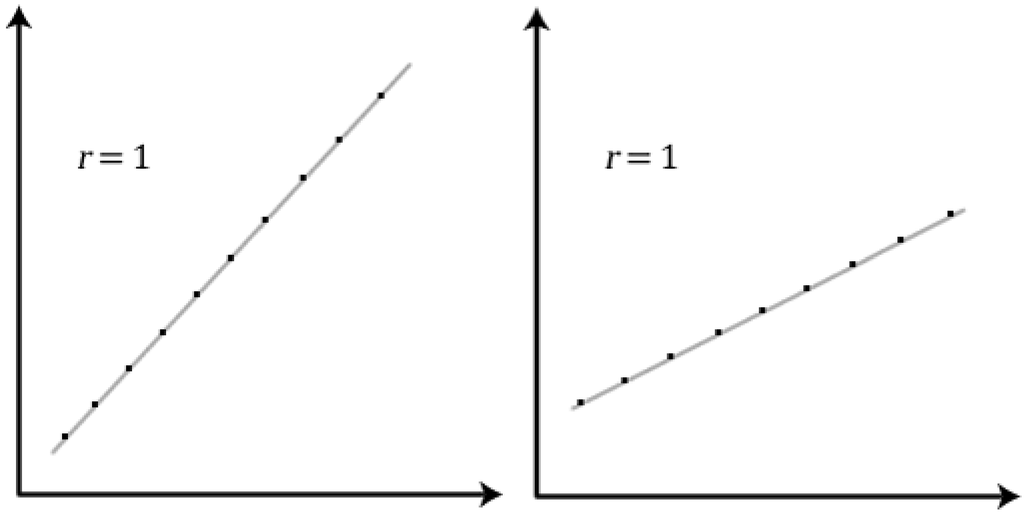
Though may vary based upon context, the following is a general rule of thumb to interpret the relationship strength between two variables.

| **Pearson correlation coefficient (r) value** | **Strength** | **Direction** |
|---|---|---|
| Greater than .5 | Strong | Positive |
| Between .3 and .5 | Moderate | Positive |
| Between 0 and .3 | Weak | Positive |
| 0 | None | None |
| Between 0 and –.3 | Weak | Negative |
| Between –.3 and –.5 | Moderate | Negative |
| Less than –.5 | Strong | Negative |

**Assumptions of Pearson's correlation:**

- Two variables should be measured on a continuous scale.
- Two continuous variables should be paired, i.e., each case has two values, one for each variable x and y.
- There should be independence of cases, which means that the two observations (x, y) of one case should be independent of the two observations (x, y) of other case.

**Pearson correlation coefficient does not indicate the slope of the line:** Pearson correlation coefficient, r, does not represent the slope of the line of best fit. So, r value of +1 this does not mean that for every unit increase in one variable there is a unit increase in another. It means that there is no variation between the data points and the line of best fit. This is illustrated below:



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

Feature scaling is a data preprocessing technique used to transform the values of features or variables in a dataset to a similar scale. It is a critical step in building accurate and effective machine learning models. The purpose is to ensure that all features contribute equally to the model building and to avoid the domination of features with larger values.

**Purpose of scaling:** Feature scaling is required 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.

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.

For example, if an algorithm does not use any feature scaling method, then it can consider the value 50 centimetres to be greater than 2 meters which is not true and, so, the algorithm will end up making wrong predictions. In such scenarios, feature scaling is used to bring all values to same magnitudes and thus, handle this issue. So, feature scaling helps to easily interpret data.

A few techniques for feature scaling include standardization, normalization, and min-max scaling. These methods adjust the feature values while preserving their relative relationships and distributions. Scaling just affects the coefficients and none of the other parameters like t-statistic, F statistic, p-values, R-square, etc.

| Sl. No. | Normalized scaling | Standardized scaling |
|---|---|---|
| 1 | Minimum and maximum value of features are used for scaling | Mean and standard deviation is used for scaling. |
| 2 | Formula for normalization:<br>$X' = \frac{X - X_{min}}{X_{max} - X_{min}}$<br>Where, Xmax and Xmin are the maximum and the minimum values of the feature, respectively. | Formula for standardization:<br>$X' = \frac{X - \mu}{\sigma}$<br>Where, μ is the mean of the feature values and σ is the standard deviation of the feature values. |
| 3 | Used when features are of different scales. | Centers data around the mean and scales to a standard deviation of 1 |
| 4 | Rescales values to a range between 0 and 1 | It is not bounded to a certain range. |
| 5 | It is really affected by outliers. | It is much less affected by outliers. |
| 6 | Scikit-Learn provides a transformer called MinMaxScaler for Normalization. | Scikit-Learn provides a transformer called StandardScaler for standardization. |
| 7 | Useful when the distribution of the data is unknown or not Gaussian | Useful when the distribution of the data is Gaussian or normal. |

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

A variance inflation factor (VIF) provides a measure of multicollinearity among the independent variables in a multiple regression model. In general,

- VIF equal to 1 = variables are not correlated
- VIF between 1 and 5 = variables are moderately correlated
- VIF greater than 5 = variables are highly correlated

A large value of VIF indicates that there is a correlation between the variables. If there is perfect correlation, then VIF = infinity. The formula for VIF is:

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

where:
$R_i^2$ = Unadjusted coefficient of determination for regressing the ith independent variable on the remaining ones.

When the value of VIF is infinite it shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R-squared ($R^2$) =1, which leads to 1/ (1-$R^2$) = infinity. To solve this, we need to drop one of the variables from the dataset by trial which may potentially cause this perfect multicollinearity.

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

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

A Q-Q plot is a scatterplot that compares the quantiles of two distributions. One distribution is usually the observed data, and the other is a theoretical or reference distribution, such as the normal distribution. The idea is to see how well the data fit the expected distribution by checking if the points lie on or near a straight line.

**Use of Q-Q plot:**

A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second dataset. By a quantile, we mean the percentage 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. The first quantile is that of the variable we are testing the hypothesis for and the second one is the actual distribution we are testing it against. A 45-degree reference line is also plotted. If the two sets come from a population with the same distribution, the points should roughly fall along this reference line. The more the points are away from this reference line, the greater the evidence for the conclusion that the two data sets have come from populations with different distributions.

**Importance of Q-Q plot in linear regression:**

A Q-Q plot can be used in regression models to check some of the assumptions that are required for valid inference. For example, a Q-Q plot can be used to check if the residuals of the model are normally distributed, which is an assumption for many parametric tests and confidence intervals. If there is a significant deviation from the mean, we may need to check the distribution of the feature variable and consider transforming them into a normal shape.

A Q-Q plot can also be used to check if the residuals have a constant variance, which is an assumption for the homoscedasticity of the model. To do this, we need to create a Q-Q plot for the residuals of the model and compare them with the normal distribution.

Q-Q plot can also be used to test distribution amongst 2 different datasets. For example, if in dataset 1, the height variable has 200 records and in dataset 2, the height variable has 20 records, it is possible to compare the distributions of these datasets to see if they are indeed the same. This can be particularly helpful in machine learning, where we split data into train-validation-test to see if the distribution is indeed the same.