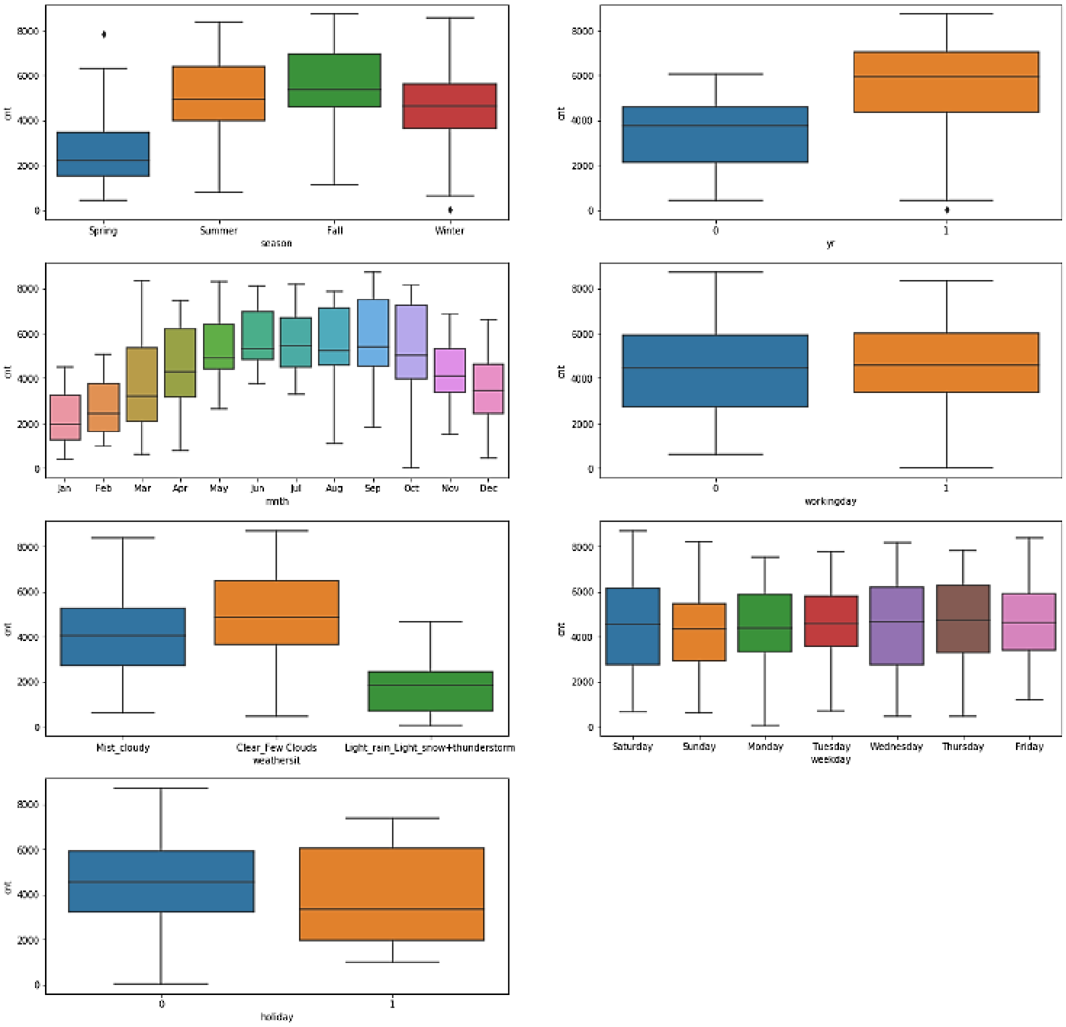
Assignment-based Subjective Questions

1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)

Ans. The categorical variable in the dataset are Season, Year, Months, Working day, Weather, Weekday and Holiday. These were visualized using a boxplot as below.



Below are the findings:

- **Season:** In the season plot we can clearly see, the category 3 : Fall, has the highest median, which shows that the demand was high during this season, followed by Summer and Winter. It is least for Spring.
- **Year:** From the year plot we can say, average rented bikes has increased in 2019 almost double that of 2018.
- **Months:** We can see a similar average count of rented bikes in June, July, August & September, followed by May & October. Company should make sure they prepare with high availability during these month. December, January, February have the least demand probably due to winter season.
- **Working day:** There are similar demands whether it's a working day or not.
- **Weather:** We clearly see that there is more demand when Weather is Clear. The count of total users is in between 4000 to 6000 during clear weather. Company should leverage and look up for forecast of weather to fullfill demands.
- **Weekday:** The bike demand is almost constant throughout the week.
- **Holiday:** There is a decrease of demand if it is a holiday

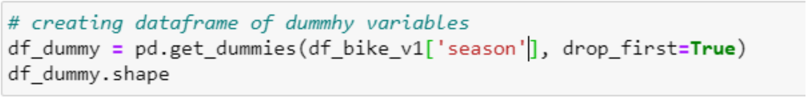
2. Why is it important to use drop_first=True during dummy variable creation? (2 mark)

Ans. If we don't drop the first column then the dummy variables will be correlated (redundant). This may affect some models adversely and the effect is stronger when the cardinality is smaller.

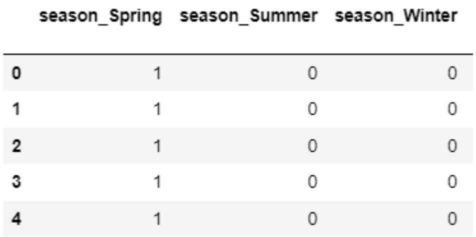
This leads to dummy variable trap. The dummy variable trap occurs when one or more dummy variables are redundant, meaning they can be predicted from the other variables. To avoid the dummy variable trap, we need to drop one of the dummy variables from each category.

With **drop_first=True,** it drops the first column during dummy variable creation. Suppose, you have a column for gender that contains 4 variables- "Male", "Female", "Other", "Unknown". So a person is either "Male", or "Female", or "Other". If they are not either of these 3, their gender is "Unknown".

Eg: From the case study, we have 4 values for Season: Fall, Spring, Summer and Winter. By dropping first in dummy value creation, “Fall” is dropped.



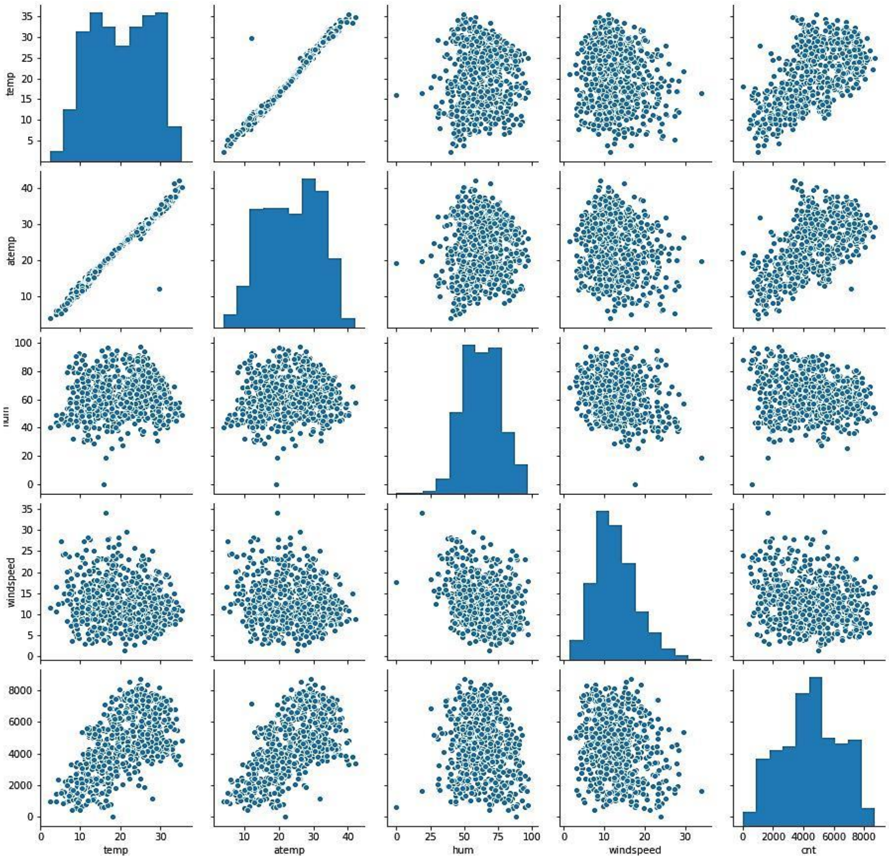
```
# creating dataframe of dummhy variables
df_dummy = pd.get_dummies(df_bike_v1['season'], drop_first=True)
df_dummy.shape
```



	season_Spring	season_Summer	season_Winter
0	1	0	0
1	1	0	0
2	1	0	0
3	1	0	0
4	1	0	0

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

Ans: Using the below pairplot we can say, “temp” and “atemp” are the two numerical variables which are highly correlated with the target variable (cnt).

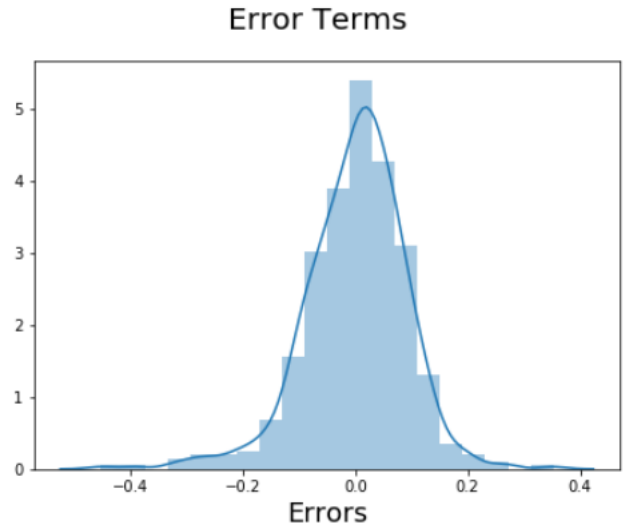


4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

Ans. The following tests were done to validate the assumptions of linear regression:

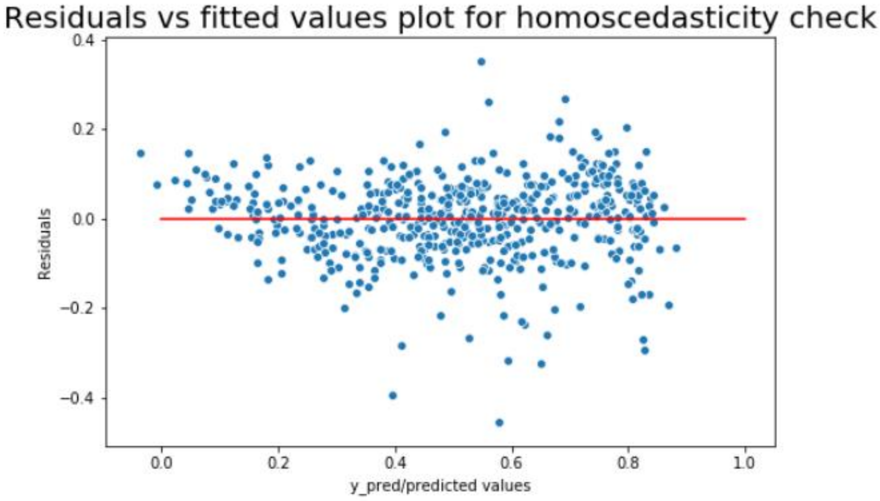
1. **Linearity:** Linear regression needs the relationship between the independent and dependent variables to be linear. We visualised the numeric variables using a pairplot to see if the variables are linearly related or not. Refer to the notebook for more details.

2. **Normality:** Residuals distribution should follow normal distribution and centred around 0 (mean = 0). We validated this assumption about residuals by plotting a distplot of residuals and saw if residuals are following normal distribution or not. The diagram below shows that the residuals are distributed about mean = 0.



3. **No multicollinearity:** Linear regression assumes that there is little or no multicollinearity in the data. Multicollinearity occurs when the independent variables are too highly correlated with each other. We calculated the VIF (Variance Inflation Factor) to get the quantitative idea about how much the feature variables are correlated with each other in the new model. Refer to the notebook for more details.

4. **Homoscedasticity:** The variance should not increase (or decrease) as the error values change. Also, the variance should not follow any pattern as the error terms change. From the below plot, we can see that residuals have equal or almost equal variance across the regression line.



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

Ans: The top 3 features are:
1. **atemp** - coefficient : **0.4117**
2. **yr** - coefficient : **0.2357**
3. **weathersit_Light rain_Light snow+Thunderstorm** - coefficient : **-0.2912**

General Subjective Questions

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

Linear regression is a type of supervised machine learning algorithm that computes the linear relationship between a dependent variable and one or more independent features. It is used for the prediction of numeric values. Linear Regression is the most basic form of regression analysis. Regression is the most commonly used predictive analysis model.

Linear regression is based on the popular equation "y = mx + c".

It assumes that there is a linear relationship between the dependent variable(y) and the predictor(s) / independent variable(x).

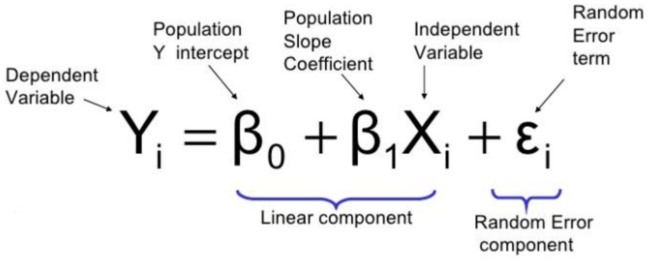
In regression, we calculate the best fit line which describes the relationship between the independent and dependent variable. Regression is performed when the dependent variable is of continuous data type and Predictors or independent variables could be of any data type like continuous, nominal/categorical etc.

Regression method tries to find the best fit line which shows the relationship between the dependent variable and predictors with least error.

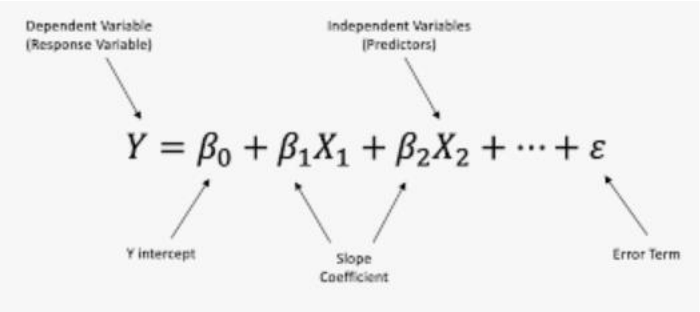
In regression, the output/dependent variable is the function of an independent variable and the coefficient and the error term.

Regression is broadly divided into simple linear regression and multiple linear regression.

1. **Simple Linear Regression:** This is the simplest form of linear regression, and it involves only one independent variable and one dependent variable.
The equation for simple linear regression is:



2. **Multiple Linear Regression:** This involves more than one independent variable and one dependent variable. The equation for multiple linear regression is:



$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip} + \epsilon$$

$$\textbf{where, for } i = n \textbf{ observations:}$$

$$y_i = \text{dependent variable}$$

$$x_i = \text{expanatory variables}$$

$$\beta_0 = \text{y-intercept (constant term)}$$

$$\beta_p = \text{slope coefficients for each explanatory variable}$$

$$\epsilon = \text{the model's error term (also known as the residuals)}$$

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

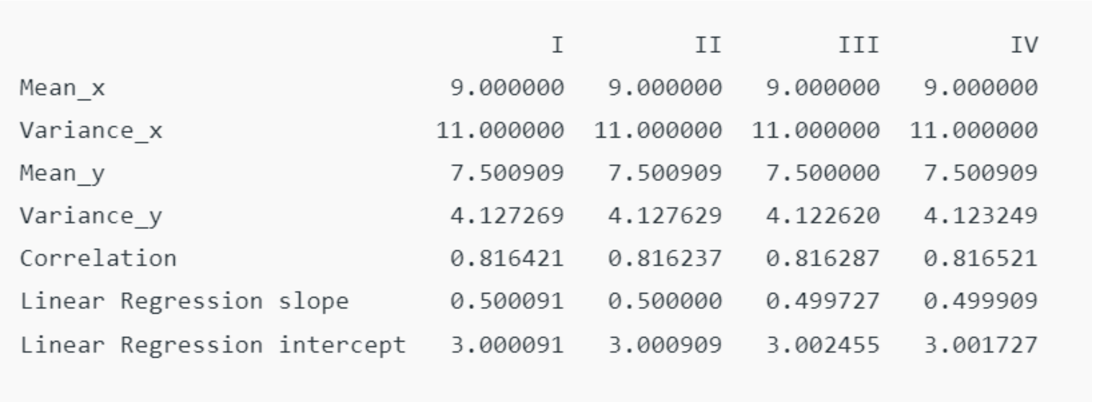
Anscombe's quartet comprises a set of four datasets, having identical descriptive statistical properties in terms of means, variance, R-squared, correlations, and linear regression lines but having different representations when we scatter plots on a graph.

The datasets were created by the statistician Francis Anscombe in 1973 to demonstrate the importance of visualizing data and to show that summary statistics alone can be misleading.

The four datasets that make up Anscombe's quartet each include 11 x-y pairs of data. When plotted, each dataset seems to have a unique connection between x and y, with unique variability patterns and distinctive 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.

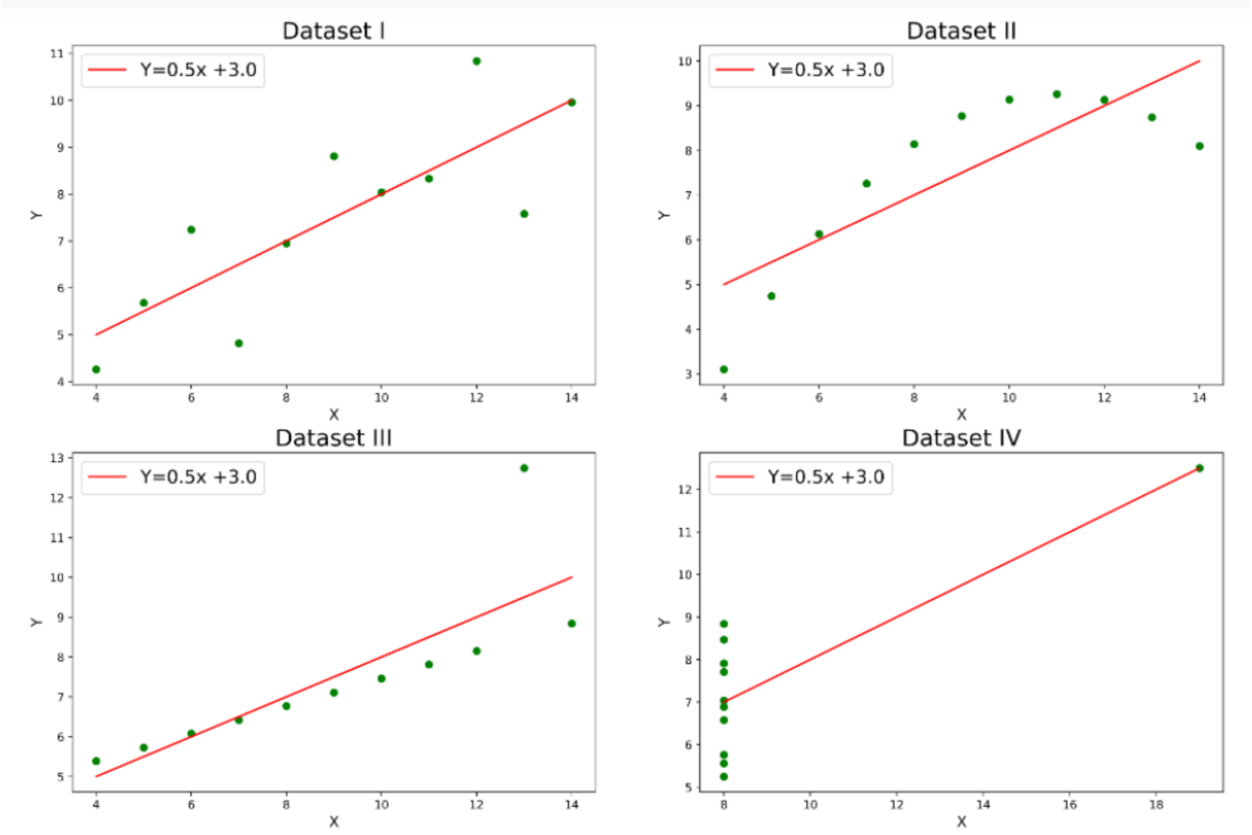
Purpose of Anscombe's Quartet
Anscombe's quartet is used to illustrate the importance of exploratory data analysis and the drawbacks of depending only on summary statistics. It also emphasizes the importance of using data visualization to spot trends, outliers, and other crucial details that might not be obvious from summary statistics alone.

Descriptive Statistical Properties for the all four Dataset



```
                                    I          II         III          IV
Mean_x                       9.000000    9.000000    9.000000    9.000000
Variance_x                  11.000000   11.000000   11.000000   11.000000
Mean_y                       7.500909    7.500909    7.500000    7.500909
Variance_y                   4.127269    4.127629    4.122620    4.123249
Correlation                  0.816421    0.816237    0.816287    0.816521
Linear Regression slope      0.500091    0.500000    0.499727    0.499909
Linear Regression intercept  3.000091    3.000909    3.002455    3.001727
```

The scatter plot and linear regression line for each datasets



Anscombe's quartet Plot

Explanation of this output:

- 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 which is indicated be 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.

Conclusion

While the descriptive statistics of Anscombe's Quartet may appear uniform, the accompanying visualizations reveal distinct patterns, showcasing the necessity of combining statistical analysis with graphical exploration for robust data interpretation.

3. What is Pearson’s R? (3 marks)

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.

A value of +1 is the result of a perfect positive relationship between two or more variables. Positive correlations indicate that both variables move in the same direction. Conversely, a value of -1 represents a perfect negative relationship. Negative correlations indicate that as one variable increases, the other decreases; they are inversely related. A zero indicates no correlation.

The Pearson correlation coefficient is also an inferential statistic, meaning that it can be used to test statistical hypotheses. Specifically, we can test whether there is a significant relationship between two variables.

For a sample, the sample correlation coefficient or the sample Pearson correlation coefficient is given by

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

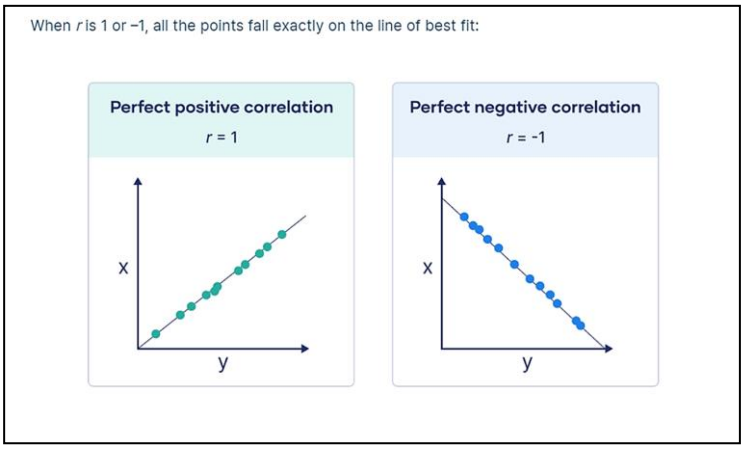
where

- n is sample size
- x_i, y_i are the individual sample points indexed with *i*
- $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ (the sample mean); and analogously for $\bar{y}$.

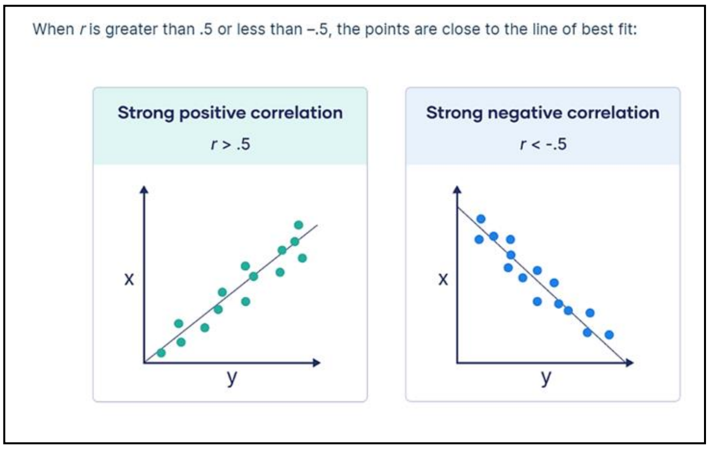
Visualizing the Pearson correlation coefficient

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.

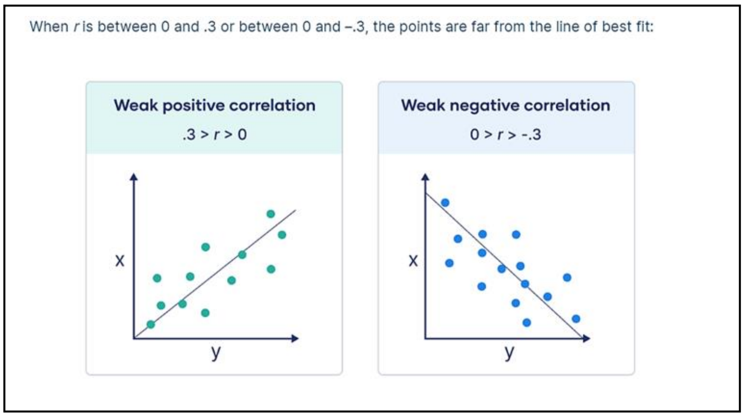
The Pearson correlation coefficient also tells you whether the slope of the line of best fit is negative or positive. When the slope is negative, r is negative. When the slope is positive, r is positive.



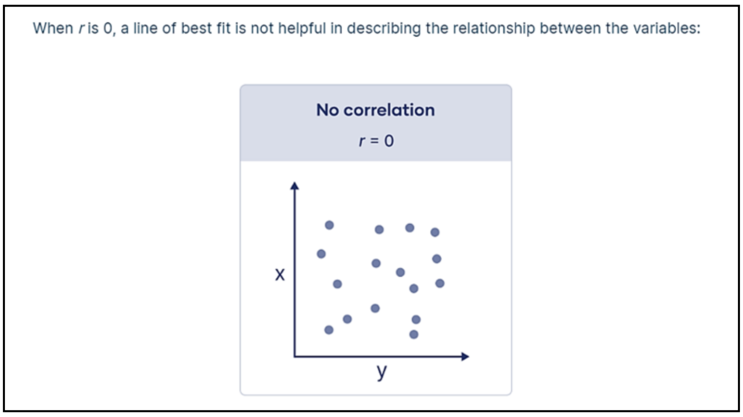
When *r* is 1 or −1, all the points fall exactly on the line of best fit:



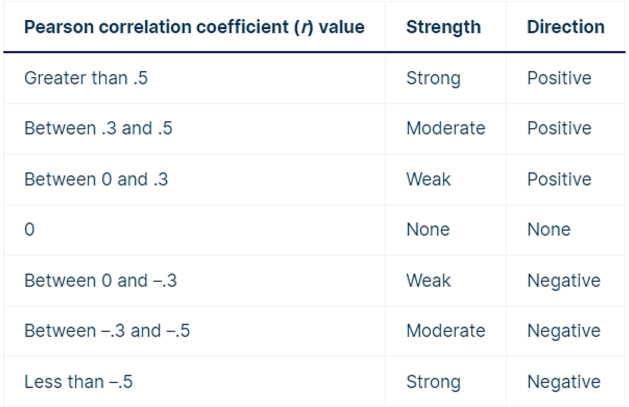
When *r* is greater than .5 or less than −.5, the points are close to the line of best fit:



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:



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

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

Feature Scaling is a technique to standardize the independent features present in the data in a fixed range. It is performed during the data pre-processing to handle highly varying magnitudes or values or units. It also helps in speeding up the calculations in an algorithm.

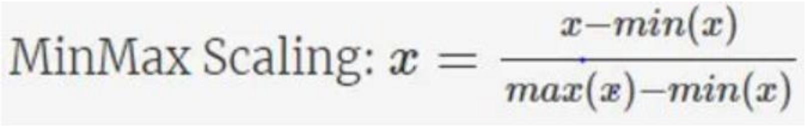
Scaling is performed because:

It is a data pre-processing procedure used to normalize data within a specific range by applying it to independent variables. Additionally, it aids in accelerating algorithmic calculations. The majority of the time, the obtained data set includes characteristics that vary greatly in magnitudes, units, and 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 in account and not units hence incorrect modelling. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude.

Normalization/Min-Max Scaling:

It brings all of the data in the range of 0 and 1. sklearn.preprocessing.MinMaxScaler helps to implement normalization in python.

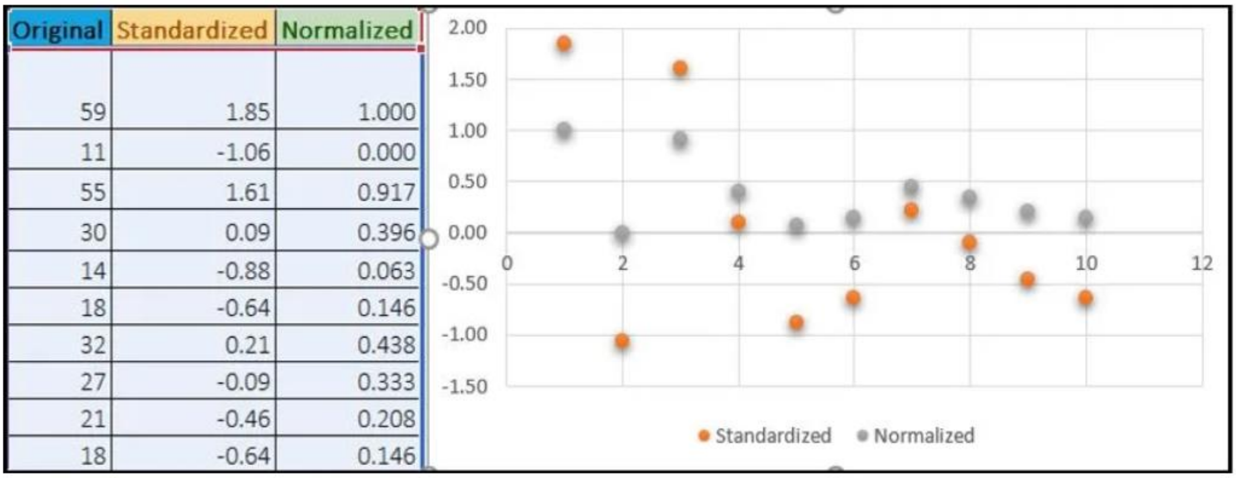


$$\text{MinMax Scaling: } x = \frac{x - min(x)}{max(x) - min(x)}$$

Standardization Scaling:
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).

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

Example:
Below shows example of Standardized and Normalized scaling 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? (3 marks)

A variance inflation factor (VIF) is a measure of the amount of multicollinearity in regression analysis. Multicollinearity exists when there is a correlation between multiple independent variables in a multiple regression model. This can adversely affect the regression results. Thus, the variance inflation factor can estimate how much the variance of a regression coefficient is inflated due to multicollinearity.

Formula of VIF
The formula for VIF is:

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

where:

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

What VIF Value indicates:
When Ri^2 is equal to 0, and therefore, when VIF or tolerance is equal to 1, the ith independent variable is not correlated to the remaining ones, meaning that multicollinearity does not exist.

- 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

The higher the VIF, the higher the possibility that multicollinearity exists, and further research is required. When VIF is higher than 10, there is significant multicollinearity that needs to be corrected.

When does VIF tends to infinity:

The greater the VIF, the higher the degree of multicollinearity.

In the limit, when multicollinearity is perfect (i.e., the regressor is equal to a linear combination of other regressors), the VIF tends to infinity.

i.e When the independent variable can be explained perfectly by other independent variables, then it will have perfect correlation and it's R-squared value will be equal to 1. So, VIF = 1/(1-1) which gives VIF = 1/0 which results in "**infinity**" .

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) plot is a graphical method for determining if a dataset follows a certain probability distribution or whether two samples of data came from the same population or not. Q-Q plots are particularly useful for assessing whether a dataset is normally distributed or if it follows some other known distribution. They are commonly used in statistics, data analysis, and quality control to check assumptions and identify departures from expected distributions.

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

Interpretation of Q-Q plot:

- If the points on the plot fall approximately along a straight line, it suggests that your dataset follows the assumed distribution.
- Deviations from the straight line indicate departures from the assumed distribution, requiring further investigation.

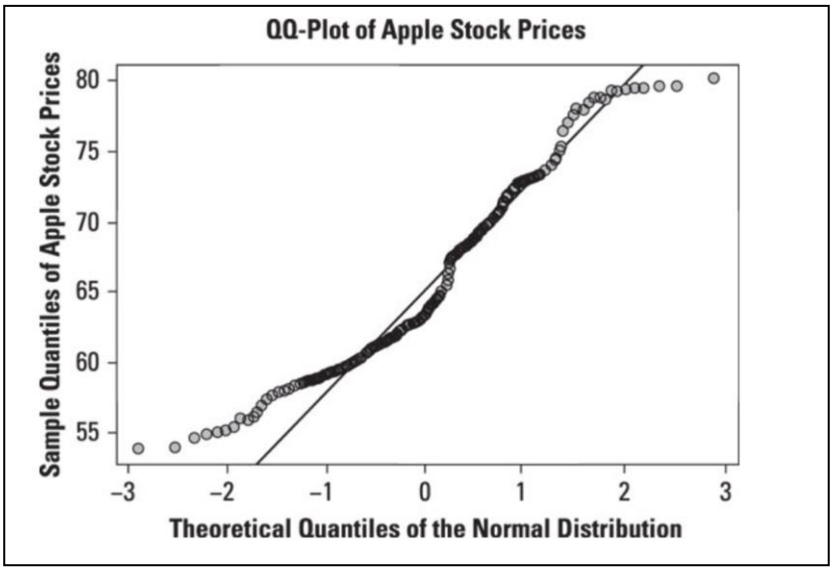
Advantages of Q-Q plot:

- **Flexible Comparison:** Q-Q plots can compare datasets of different sizes without requiring equal sample sizes.
- **Dimensionless Analysis:** They are dimensionless, making them suitable for comparing datasets with different units or scales.
- **Visual Interpretation:** Provides a clear visual representation of data distribution compared to a theoretical distribution.
- **Sensitive to Deviations:** Easily detects departures from assumed distributions, aiding in identifying data discrepancies.
- **Diagnostic Tool:** Helps in assessing distributional assumptions, identifying outliers, and understanding data patterns.

Q-Q Plot:

With a QQ-plot, the quantiles of the sample data are on the vertical axis, and the quantiles of a specified probability distribution are on the horizontal axis. The plot consists of a series of points that show the relationship between the actual data and the specified probability distribution. If the elements of a dataset perfectly match the specified probability distribution, the points on the graph will form a 45 degree line.

For example, this figure shows a normal QQ-plot for the price of Apple stock from January 1, 2013 to December 31, 2013.



The QQ-plot shows that the prices of Apple stock do not conform very well to the normal distribution. In particular, the deviation between Apple stock prices and the normal distribution seems to be greatest in the lower left-hand corner of the graph, which corresponds to the left tail of the normal distribution. The discrepancy is also noticeable in the upper right-hand corner of the graph, which corresponds to the right tail of the normal distribution.

The graph shows that the smallest prices of Apple stock are not small enough to be consistent with the normal distribution; similarly, the largest prices of Apple stock are not large enough to be consistent with the normal distribution. This shows that the tails of the Apple stock price distribution are too "thin" or "skinny" compared with the normal distribution. The conclusion to be drawn from this is that the Apple stock prices are not normally distributed.

Use and importance of a Q-Q plot in linear regression:

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

- A Q-Q plot can be used in regression models to check some of the assumptions that are required for valid inference. For example,

  - You can use a Q-Q plot to check if the residuals of the model are normally distributed, which is an assumption for many parametric tests and confidence intervals.

  - You can also use a Q-Q plot to check if the residuals have a constant variance, which is an assumption for the homoscedasticity of the model.

  - To do this, you need to create a Q-Q plot for the residuals of the model and compare them with the normal distribution.