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)**

Answer:

Below are few inferences–

- Fall season seems to have the highest demand. Winter and Summar stand close to Fall however spring have far lower demand than any other season.
- Most of the bookings has been done during the month of may, june, july, aug, sep and oct. Trend increased starting of the year till mid of the year and then it started decreasing as we approached the end of year.
- Season and months are highly correlated.
- Clear weather situation gets more demand.
- Thu, Fir, Sat and Sun have a greater number of bookings as compared to the start of the week.
- Holidays see less demand.
- Demand does not seem to be dependent on working day or non-working day.
- 2019 shows more demand compared to the previous year across all the categorical variables.

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

Answer:
Dummy variables have high multicollinearity since they are mutually exclusive. Setting drop_first = True reduces the number of variables by one hence reducing the correlations created among dummy variables. (not eliminating though)

Example: If we have 3 distinct values (say A, B, C) in a Categorical variable and we want to create dummy variable for that column. If we just have two dummy variables A and B. Below table illustrate it.

OriginalVariable	DummyA	DummyB
A	1	0
B	0	1
C	0	0

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

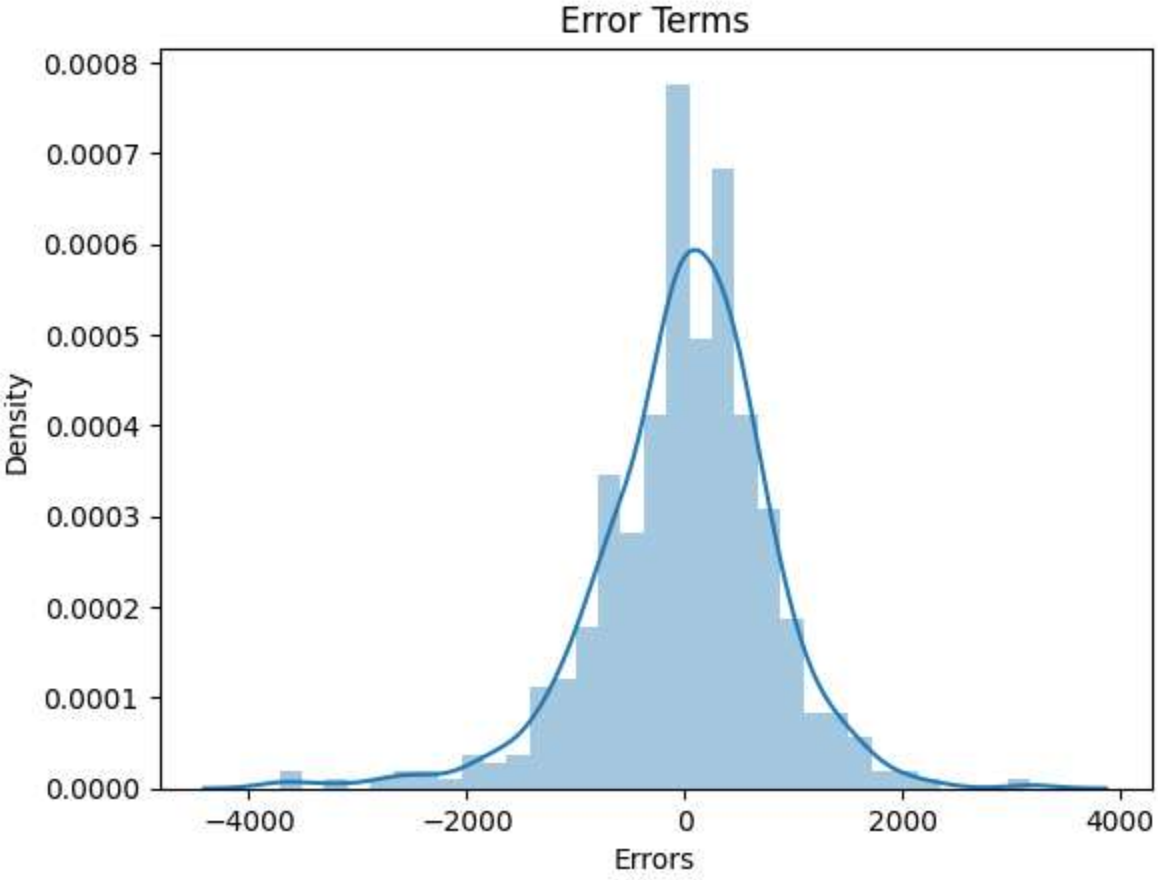
Answer:
‘temp’ variable has the highest correlation with the target variable.

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

Below are details of validations used for linear regression:

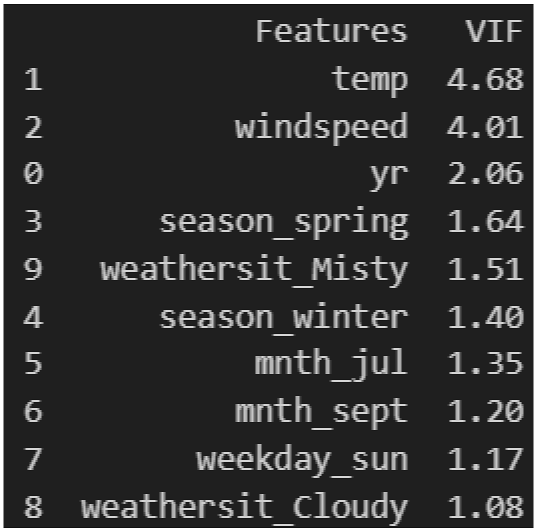
- Normality of error terms

  Use seaborn to plot distribution plot of error in prediction at every term.
  Verified that it is normally distributed.



- Multicollinearity check

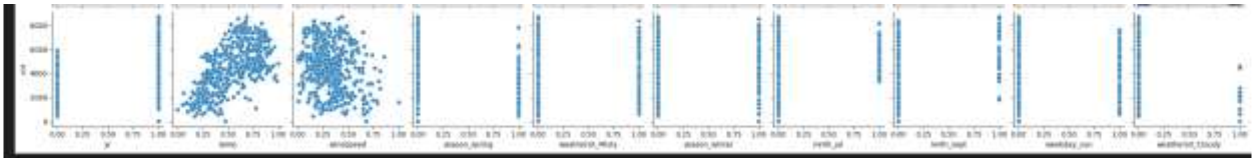
  Used variance_inflation_factor method from statsmodels to calculate VIF for all features present in model and verified that they are all less than 5.



```
             Features   VIF
1                temp  4.68
2           windspeed  4.01
0                  yr  2.06
3       season_spring  1.64
9    weathersit_Misty  1.51
4       season_winter  1.40
5            mnth_jul  1.35
6           mnth_sept  1.20
7         weekday_sun  1.17
8   weathersit_Cloudy  1.08
```

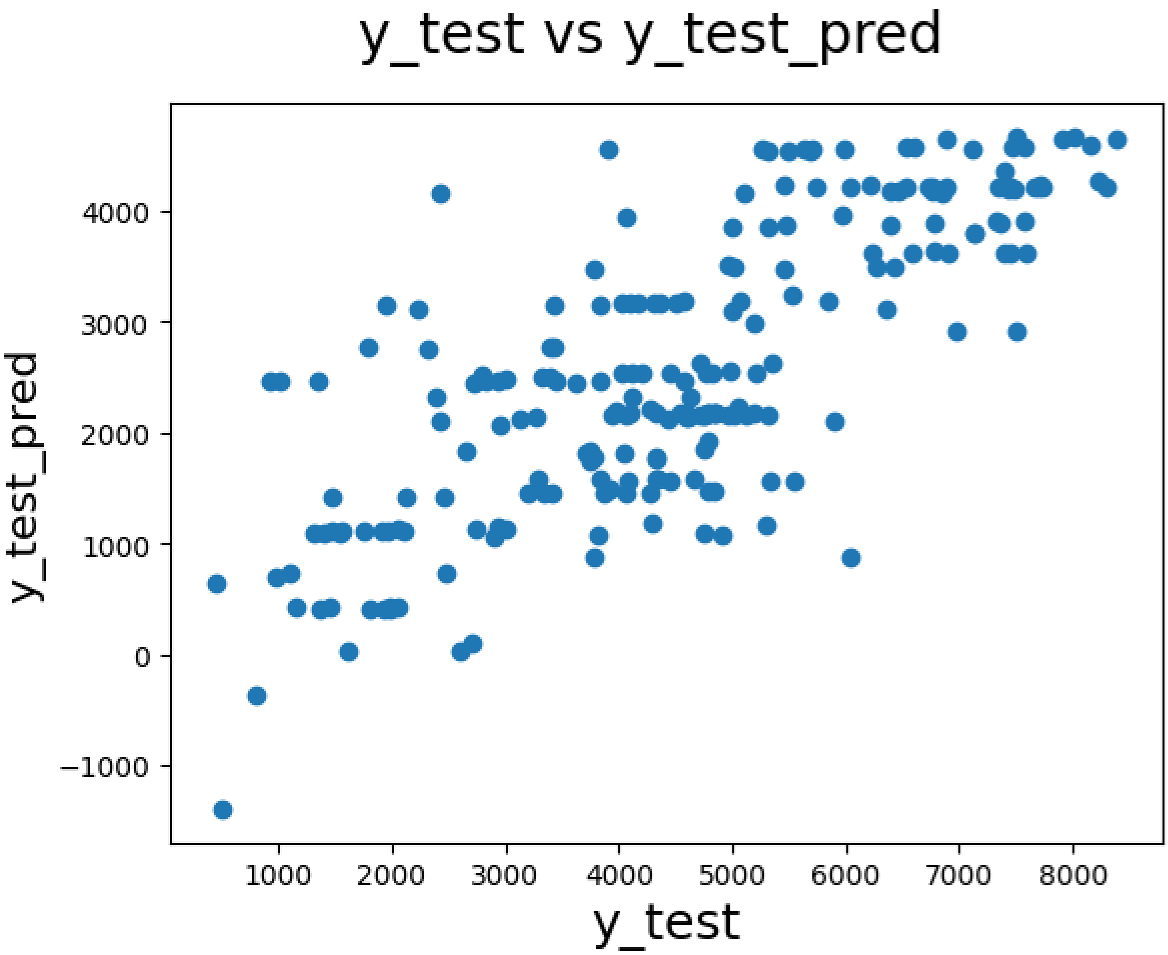
- Linear relationship validation

Validated linear relationship between some of independent variables with dependent variable.



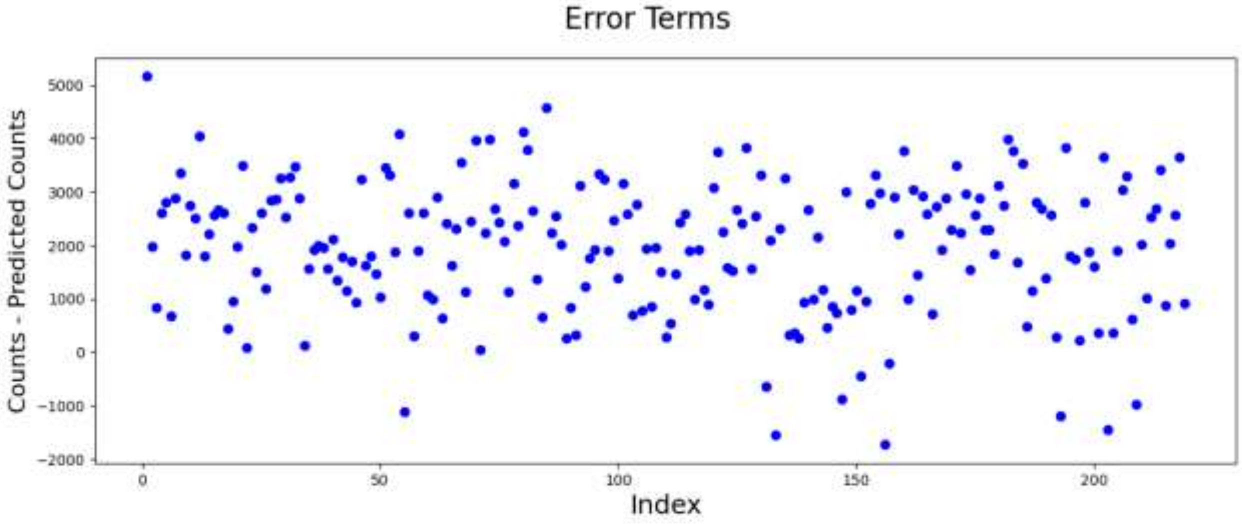
- Homoscedasticity

Plotted y_test and y_test_pred to understand the spread. Validated that it is linearly correlated.



- Independence of residuals

Plotted residual in sequence of data points to validate that there is no 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)**

Answer:

Looking at corresponding coefficients, below are the top 3 features contributing significantly towards explaining the demand of the shared bikes –

- temp
- weathersit == 2 (Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist)
- yr

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Linear regression assumes linear algebraic relationships between dependent variable and one or more independent variables. At high level the algorithm guesses the coefficients and intercepts of the linear equations iteratively to reduce the error in prediction as a whole.

Mathematically the relationship can be represented with the help of following equation −
There are two main types of linear regression:

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:
$y=\beta 0+\beta 1X$y=β0+β1X
where: Y is the dependent variable

X is the independent variable

β0 is the intercept

β1 is the slope

Multiple Linear Regression

This involves more than one independent variable and one dependent variable. The equation for multiple linear regression is:
$y=\beta 0+\beta 1X+\beta 2X+\ldots\ldots\ldots\beta nX$y=β0+β1X+β2X+………βnX
where:

Y is the dependent variable, X1, X2, …, Xp are the independent variables

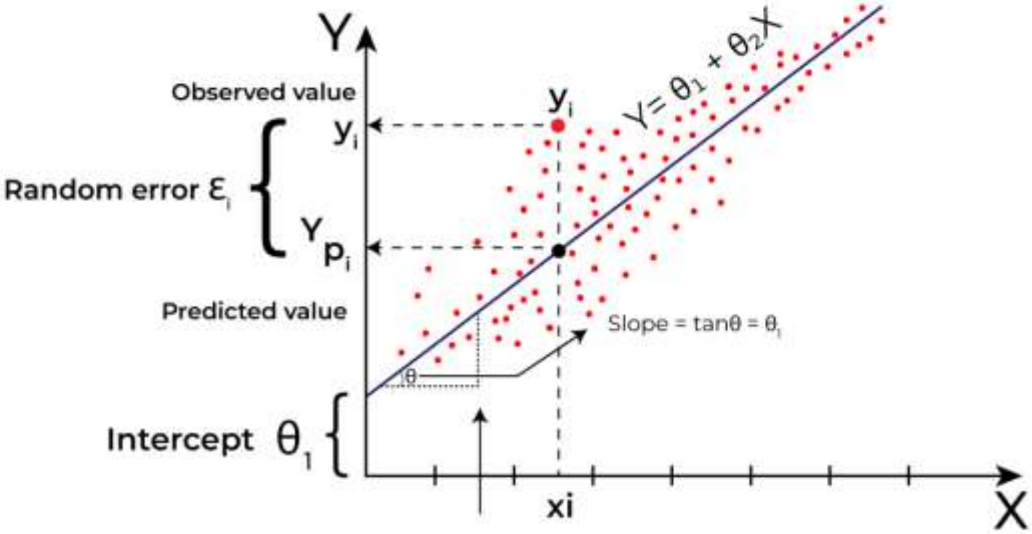
β0 is the intercept

β1, β2, …, βn are the slopes

The goal of the algorithm is to find the best Fit Line equation that can predict the values based on the independent variables.

In regression analysis, we have a dataset containing independent variable X and dependent variable Y. The goal is to learn a function that can predict the value of Y for a new, unknown X. Regression involves determining a function that can forecast a continuous Y value based on X as the independent variable.

The equation for the best fit line provides a linear representation of how the independent variables influence the dependent variable. The line's gradient indicates the magnitude of change in the dependent variable for each unit change in the independent variable(s).


Y
Observed value
y_i
y_i
Random error ε_i
Y_{p_i}
Predicted value
$Y = \theta_1 + \theta_2 X$
Slope = tanθ = θ_i
θ
Intercept θ_1
xi
X

Linear Regression

In this context, Y is referred to as the dependent or target variable, while X is known as the independent variable, also termed the predictor of Y. There exists a variety of functions or models that can be utilized for regression purposes. Among these, a linear function is considered the most basic form. Here, X could represent a single feature or multiple features that characterize the problem.

Linear regression performs the task to predict a dependent variable value (y) based on a given independent variable (x)). Hence, the name is Linear Regression. In the figure above, X (input) is the work experience and Y (output) is the salary of a person. The regression line is the best-fit line for our model.

Mathematical Approach:

Residual/Error = Actual values – Predicted Values
Sum of Residuals/Errors = Sum(Actual- Predicted Values)
Square of Sum of Residuals/Errors = (Sum(Actual- Predicted Values))^2
i.e

Rsq, AdjRsq, MSE,RMSE,MAE – 5 evaluation metrics

Let see R Squared (R2) approach:

$$\sum {e_i}^2 = \sum (Y_i - \hat{Y}_i)^2$$

As our regression line moves towards perfection, R2 score move towards one. And the model performance improves.

The normal case is when the R2 score is between zero and one like 0.8 which means your model is capable to explain 80 per cent of the variance of data.

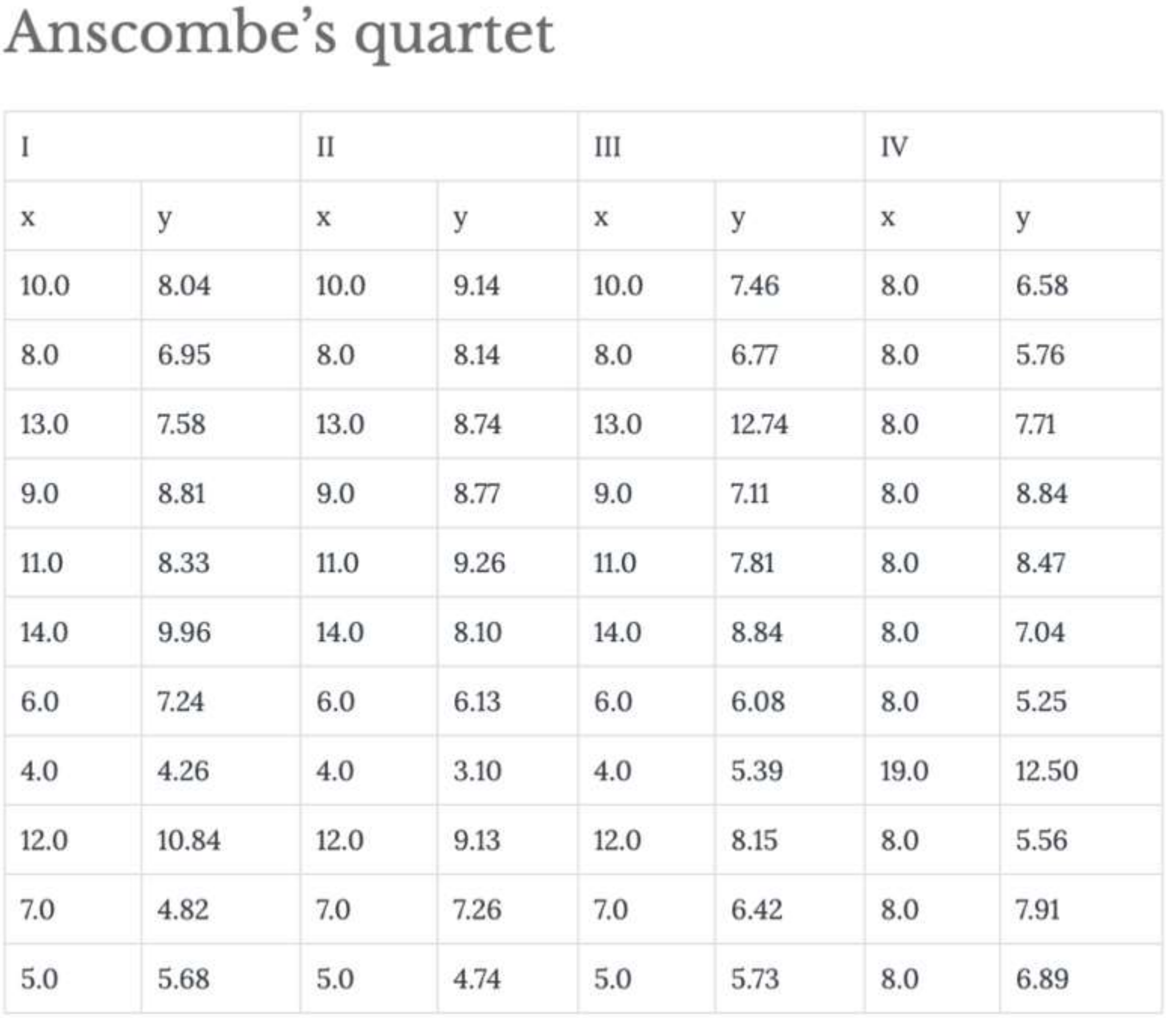
```
from sklearn.metrics import r2_score

r2 = r2_score(y_test,y_pred)

print(r2)
```

2. Explain the Anscombe’s quartet in detail.

Anscombe's Quartet consists of four distinct datasets, each displaying the same statistical descriptors—mean, variance, correlation, and linear regression lines—yet they reveal diverse distributions upon graphical representation. Conceived by statistician Francis Anscombe in 1973, these datasets serve as a pivotal illustration of the criticality of graphical data analysis, cautioning against the sole reliance on summary statistics. Each of the quartet's datasets contains 11 pairs of x-y data points. The scatter plots of these datasets exhibit unique relationships and variability patterns between x and y, along with varying strengths of correlation. Nevertheless, all four datasets share identical summary statistics, maintaining the same mean, variance, and correlation coefficients for both x and y, as well as congruent linear regression lines.



Anscombe’s quartet

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

Purpose of Anscombe’s Quartet

Anscombe's Quartet serves as a compelling example of why exploratory data analysis is crucial and the pitfalls of relying solely on summary statistics. It further accentuates the significance of employing data visualization techniques to identify patterns, anomalies, and other essential aspects that may not be immediately apparent through summary statistics.

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.

Pearson correlation coefficient (*r*)	**Correlation type**	**Interpretation**	**Example**
Between 0 and 1	Positive correlation	When one variable changes, the other variable changes in the **same direction**.	Spending on Marketing and Revenue. More we spend on marketing more we get more revenue.
0	No correlation	There is **no relationship** between the variables.	House price & Exterior color of house.
Between 0 and –1	Negative correlation	When one variable changes, the other variable changes in the **opposite direction**.	Elevation & air pressure: The higher the elevation, the lower the air pressure.

Limitation of Pearson correlation coefficient:

The Pearson correlation coefficient (*r*) is one of several correlation coefficients that you need to choose between when you want to measure a correlation. The Pearson correlation coefficient is a good choice when **all** of the following are true:

- Both variables are quantitative
- The variables are normally distributed
- The data have no outliers
- The relationship is linear

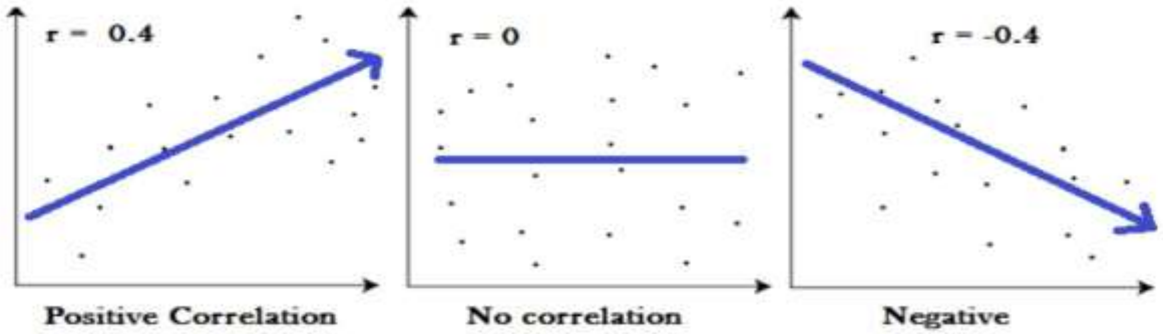
Calculating the Pearson correlation coefficient

Below is a formula for calculating the Pearson correlation coefficient (*r*):

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

Correlation coefficient formulas are used to find how strong a relationship is between data. The formulas return a value between -1 and 1, where:

- • 1 indicates a strong positive relationship.
- • -1 indicates a strong negative relationship.
- • A result of zero indicates no relationship at all.



Graphs showing a correlation of -1, 0 and +1

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

Scaling, also referred to as data normalization, is a technique utilized to normalize the range of data features. Given the potential for a broad spectrum of data values, scaling becomes an essential step in data pre-processing for machine learning algorithms. Through the process of scaling, or min-max scaling, data is transformed to fit within a designated range, such as [0, 1], where x' represents the normalized value. This procedure is primarily conducted because datasets often contain features with significant variations in magnitudes, units, and ranges. However, as most machine learning algorithms calculate the Euclidean distance between two data points, this presents a challenge. Without scaling, these algorithms would consider only the magnitude of features, disregarding the units, leading to skewed results across different units, such as 5kg versus 5000gms. Features with larger magnitudes would disproportionately influence the distance calculations compared to those with smaller magnitudes. To mitigate this issue and equalize the magnitude levels across all features, scaling is implemented.

The difference between normalized scaling and standardized scaling is as follows:

- Normalization maps feature values to the [0, 1] range by setting the minimum value to 0 and the maximum value to 1.
- Standardization transforms data to have a mean of 0 and a standard deviation of 1, without enforcing a specific range.
- Normalization is affected by outliers, while standardization is less affected.
- Normalization is based on min max value of the variable while Standardization is based on mean and standard deviation.

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

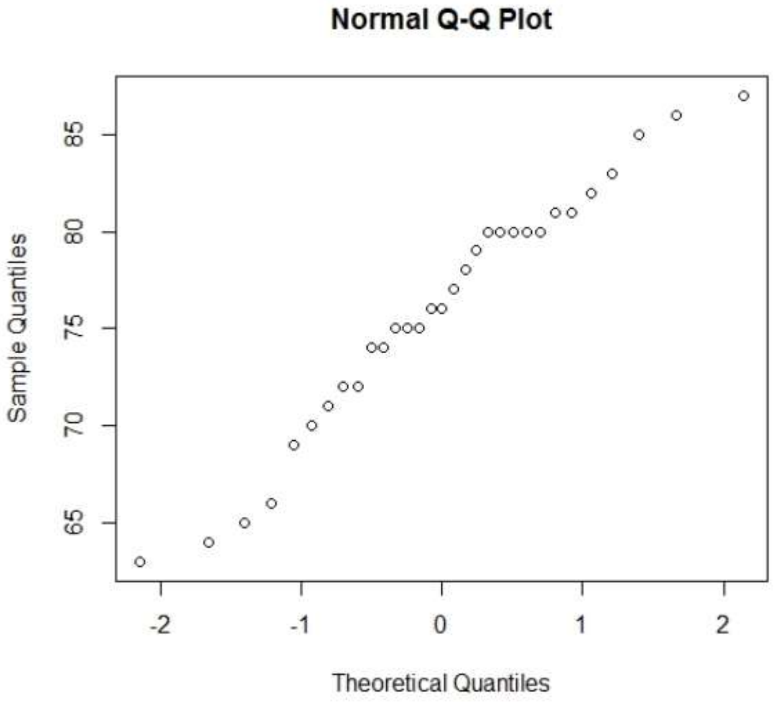
The degree of correlation between a predictor and the rest of the predictor variables in a linear regression model can be measured by the R-squared statistic, which is derived from a regression where the predictor in question is estimated using all other predictor variables. The variance inflation for a variable is then computed as:

$$VIF = \frac{1}{1 - R^2}$$

When R-squared reaches 1, VIF reaches infinity. When R-squared reaches 1 then it means multicollinearity exists. Different variables are highly correlated with each other.

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

The Q-Q plot, also known as the quantile-quantile plot, is a visual instrument utilized to determine whether a dataset likely originates from a certain theoretical distribution, such as Normal or exponential. For instance, in statistical analyses predicated on the normal distribution of the dependent variable, a Normal Q-Q plot can be employed to verify this presumption. While it is a subjective method and not a definitive proof, it provides a quick visual indication of the validity of our assumption. Should the assumption prove incorrect, the Q-Q plot can reveal how it is breached and identify the data points responsible for this breach. Essentially, a Q-Q plot is a scatterplot constructed by plotting two sets of quantiles against each other. If both quantile sets derive from the same distribution, the plotted points should align along a nearly straight line. An example is a Normal Q-Q plot where both quantile sets are from Normal distributions, which can indicate whether residuals are normally distributed. The alignment of residuals along a straight dashed line is ideal, whereas significant deviations would suggest otherwise.



Q-Q plots organize your sample data in ascending order and plot it against quantiles from a theoretical distribution. The quantile count is tailored to your sample size. While Normal Q-Q Plots are commonly used due to the assumption of normality in many statistical methods, Q-Q Plots can be generated for any distribution. These plots are a staple in statistics, typically used to validate a linear regression model by checking the alignment of data points with the line. If the points don't align, it suggests non-Gaussian residuals and errors, indicating that for small samples, the Gaussian assumption for estimators may not hold, rendering standard confidence intervals and significance tests potentially invalid.