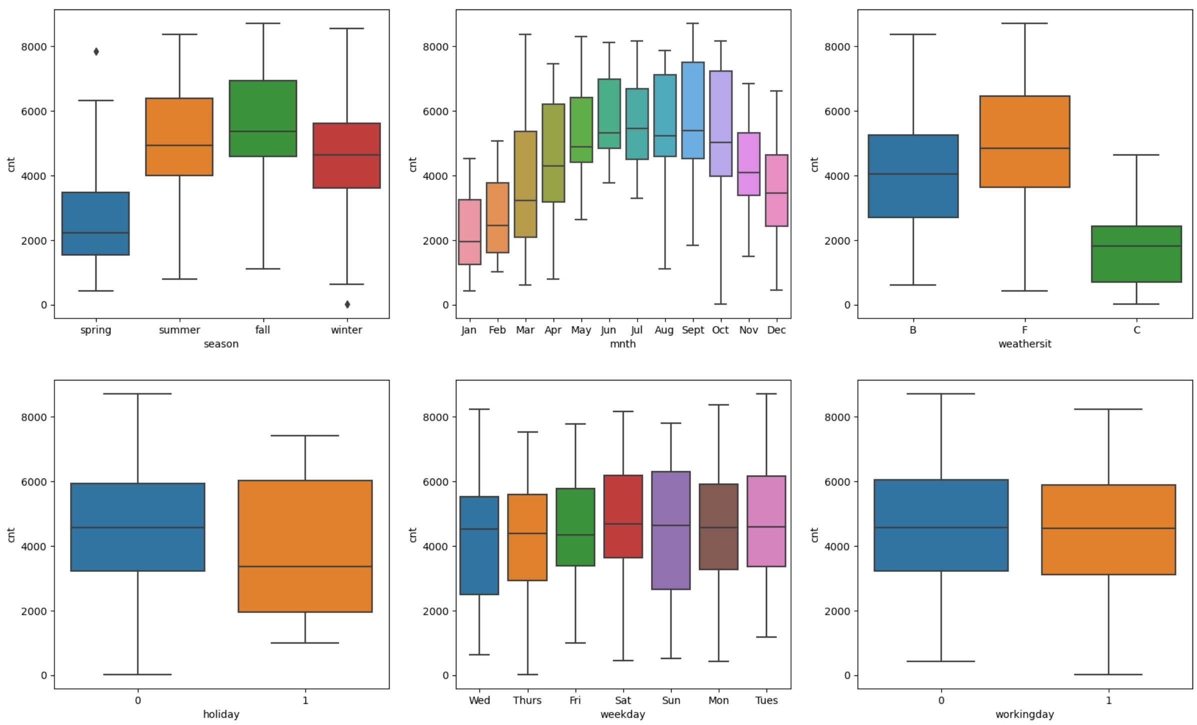
**Assignment-based Subjective Questions and Answers:**

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:

Season, Weather Situation, Holiday, Month, Working day, and Weekday were the categorical variables in the dataset, A boxplot was used to visualize these.




These Variables influenced our dependent variable in the following ways:

1. Season: Around 32% of bike bookings occurred in season 3 (fall), followed by season 2 (summer) and season 4 (winter) at 27% and 25%, respectively. Season appears to be a potential predictor for the dependent variable.
2. Month: Approximately 10% of bike bookings took place in months 5, 6, 7, 8, and 9, with a median exceeding 4000 bookings per month. Month displays a discernible trend and could be a valuable predictor.
3. Weathersit: Nearly 67% of bike bookings happened during 'weathersit1,' with a median close to 5000 bookings. Weathersit2 followed with 30% of total bookings. Weathersit demonstrates a trend and may be a useful predictor.
4. Holiday: A striking 97.6% of bike bookings occurred on non-holidays, indicating a biased dataset. Holiday is unlikely to be a reliable predictor.
5. Weekday: The weekday variable exhibits a consistent trend, contributing 13.5%-14.8% of total bookings across all days with medians ranging from 4000 to 5000. Its influence on the predictor is uncertain and left to the model's discretion.
6. Workingday: About 69% of bike bookings took place on 'workingday,' suggesting it could be a meaningful predictor for the dependent variable

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

Ans:
When creating dummy variables for categorical features, the parameter **drop_first=True** is used to avoid the dummy variable trap. The dummy variable trap occurs when two or more dummy variables perfectly predict each other. This can lead to multicollinearity issues in regression analysis.

Consider a categorical variable with three categories (A, B, C). If we create dummy variables without dropping the first one, we end up with two dummy variables, say, **dummy_A**, **dummy_B**, and **dummy_C**. Here's an example:

Original categorical variable: A, B, C

Dummy variables without drop_first=True:

- dummy_A: 1 if category is A, 0 otherwise
- dummy_B: 1 if category is B, 0 otherwise
- dummy_C: 1 if category is C, 0 otherwise

Now, if **dummy_A**, **dummy_B**, and **dummy_C** are known, the value of **dummy_C** can be perfectly predicted because if both **dummy_A** and **dummy_B** are 0, then the category must be C. This perfect predictability introduces multicollinearity, which can lead to issues in regression analysis.

By using **drop_first=True**, we eliminate one of the dummy variables, addressing the multicollinearity problem. Let's see the modified dummy variables:

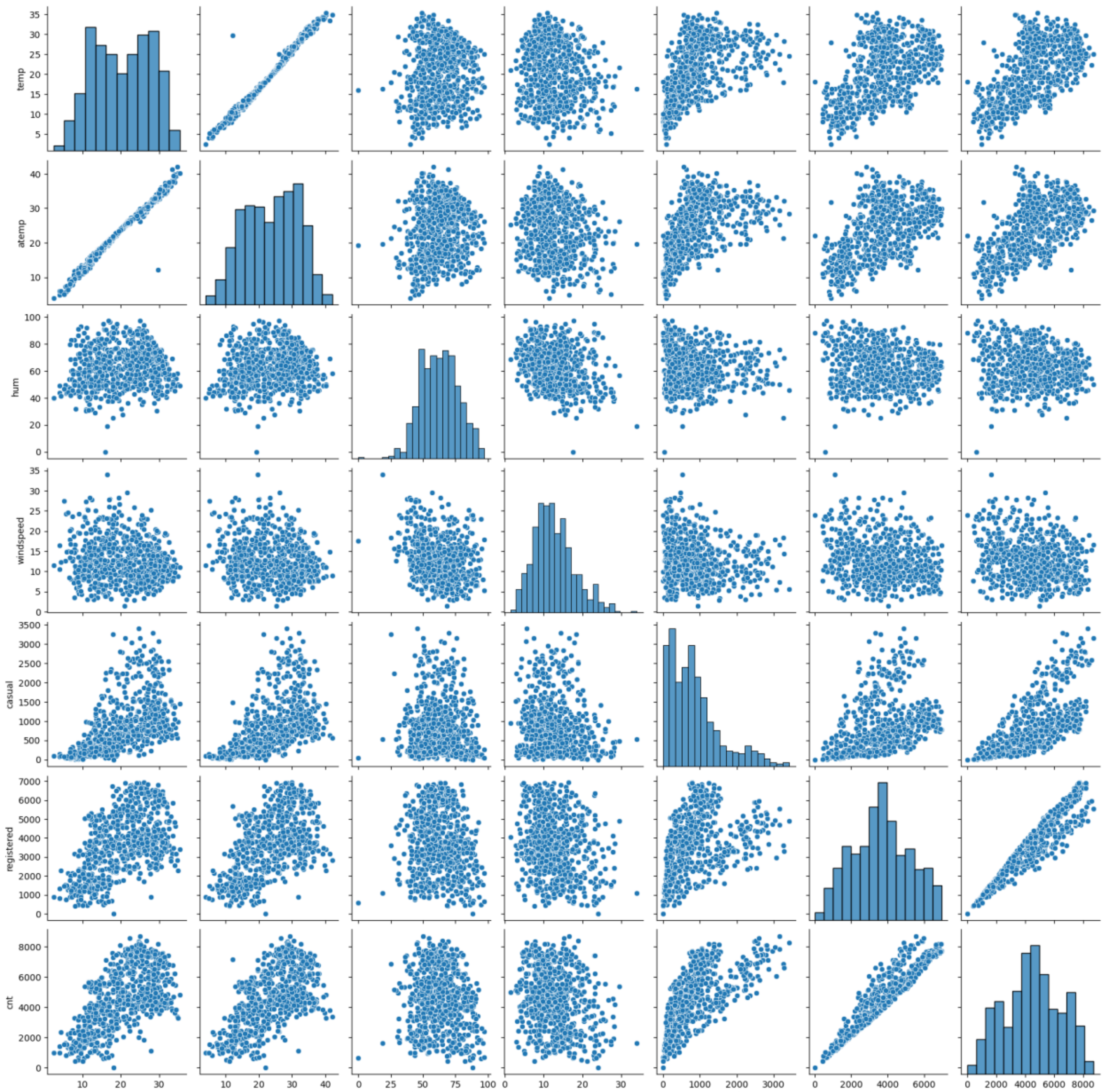
Dummy variables with drop_first=True:

- dummy_B: 1 if category is B, 0 otherwise
- dummy_C: 1 if category is C, 0 otherwise

Now, if both **dummy_B** and **dummy_C** are 0, we know that the category must be A. This approach avoids the perfect predictability issue and helps in a more stable and accurate regression model

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

Ans:




Registered has highest correlation with target variable ‘cnt’.
Temp and atemp has highest correlation with each other.

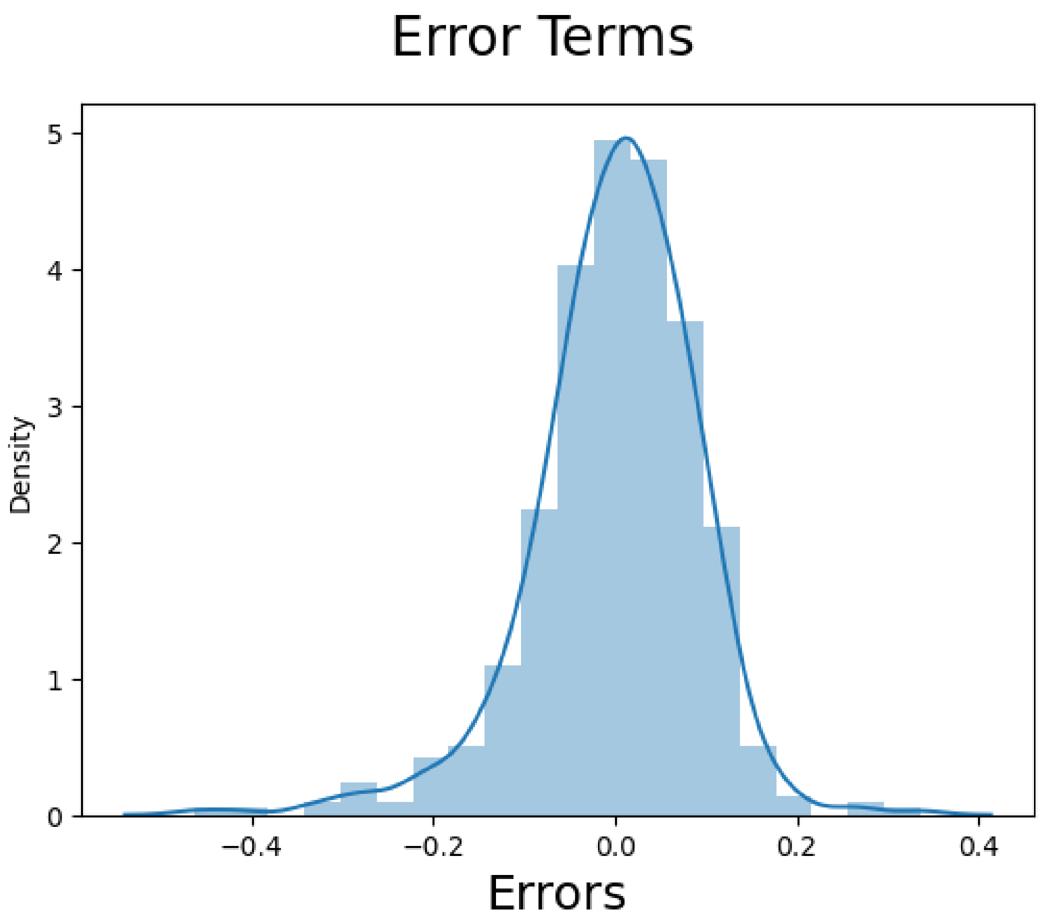
4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

Ans: After building the model on the training set, I carried out the following analysis:
Assumptions of Linear Regression:

1. There is a linear relationship between X and Y
2. Error terms are normally distributed with mean zero (not X, Y)
3. Residual Analysis of Training Data proves that the Residuals are normally distributed.

Hence our assumptions for Linear Regression is valid.

Eliminations and inclusion of independent variables into each model based on VIF and p-values to avoid multi collinearity.




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: As per our final Model, The top 3 predictor variables that influences the bike booking are:

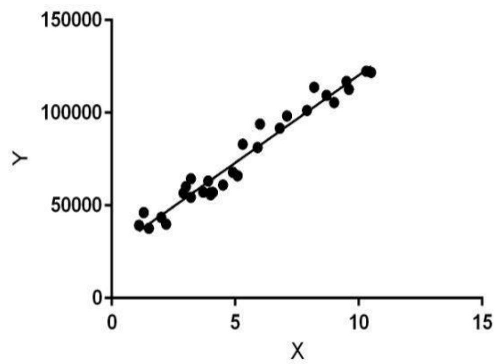
1. Temperature (temp): The coefficient of '0.4207' suggests that a one-unit increase in the temperature variable leads to a 0.4207 -unit increase in bike hires.
2. Weather Situation C (weathersit_C) (Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered): The coefficient of '-0.1903' indicates that, concerning Weathersit_3, a one-unit increase in the Weathersit_3 variable results in a decrease of 0.1903 units in bike hires.
3. Year (yr): With a coefficient value of '0.2333,' a one-unit increase in the year variable corresponds to a 0.2333 -unit increase in bike hires.

**General Subjective Questions and Answers:**

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

Ans:

Linear Regression is a machine learning algorithm based on supervised learning. It performs a regression task. Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting. Different regression models differ based on – the kind of relationship between dependent and independent variables, they are considering and the number of independent variables being used.




Linear regression performs the task to predict a dependent variable value (y) based on a given independent variable (x). So, this regression technique finds out a linear relationship between x (input) and y(output). Hence, the name is Linear Regression.

The linear regression model can be represented by the following equation:

$Y = \theta_0 + \theta_1 x_1 + \theta_1 x_1 + \ldots\ldots + \theta_n x_n$ where,

$Y$ is the predicted value $\theta_0$ is the constant term.

$\theta_1, \ldots, \theta_n$ are the model parameters $x_1, x_2, \ldots, x_n$ are the feature values.

The goal of regression analysis is to create a trend line based on the data you have gathered. This then allows you to determine whether other factors apart from the amount of calories consumed affect your weight, such as the number of hours you sleep, work pressure, level of stress, type of exercises you do etc. Before taking into account, we need to look at these factors and attributes and determine whether there is a correlation between them. Linear Regression can then be used to draw a trend line which can then be used to confirm or deny the relationship between attributes. If the test is done over a long time duration, extensive data can be collected and the result can be evaluated more accurately.

**How to update θ1 and θ2 values to get the best fit line?**

**Cost Function (J):**

By achieving the best-fit regression line, the model aims to predict y value such that the error difference between predicted value and true value is minimum. So, it is very important to update the θ1 and
θ2 values, to reach the best value that minimize the error between predicted y value (pred) and true y value (y).

$$minimize \frac{1}{n} \sum_{i=1}^{n} (pred_i - y_i)^2 \qquad\qquad J = \frac{1}{n} \sum_{i=1}^{n} (pred_i - y_i)^2$$

Cost function(J) of Linear Regression is the **Root Mean Squared Error (RMSE)** between predicted y value (pred) and true y value (y).

**Gradient Descent:**

To update θ1 and θ2 values in order to reduce Cost function (minimizing RMSE value) and achieving the best fit line the model uses Gradient Descent. The idea is to start with random θ1 and θ2 values and then iteratively updating the values, reaching minimum cost.

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

Ans:

**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

It is a group of four datasets that appear to be similar when using typical summary statistics, yet tell four different stories when graphed. Each dataset contains of eleven (x, y) pairs 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 |

All the summary statistics for each dataset are identical

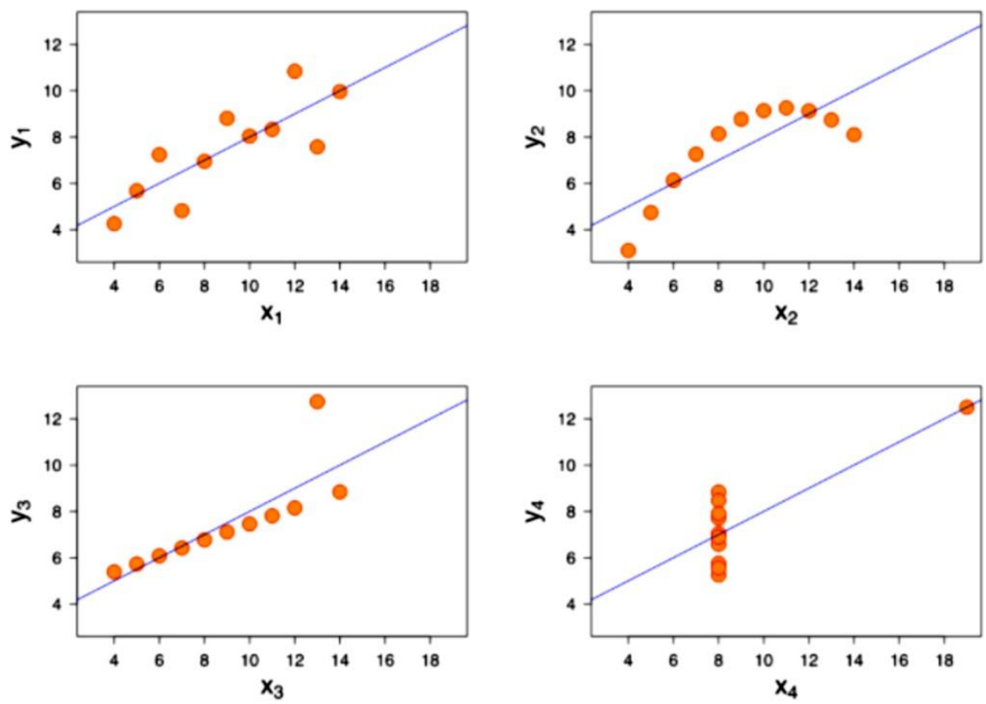
1. The average value of x is 9.
2. The average value of y is 7.5.
3. The variance for x is 11 and y is 4.12

4. The correlation between x and y is 0.816

5. The line of best for is y = 0.5x + 3.

But the plots tell a different and unique story for each dataset.

- The first scatter plot (top left) appears to be a simple linear relationship, corresponding to two variables correlated where y could be modelled as gaussian with mean linearly dependent on *x*.
- The second graph (top right) is not distributed normally; while a relationship between the two variables is obvious, it is not linear, and the Pearson correlation coefficient is not relevant. A more general regression and the corresponding coefficient of determination would be more appropriate.




- In the third graph (bottom left), the distribution is linear, but should have a different regression line (a robust regression would have been called for). 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? (3 marks)

Ans:

Pearson’s R is a numerical summary of the strength of the linear association between the variables. It varies between -1 and +1. If the variables tend to go up and down together, the correlation coefficient will be positive. If the variables tend to go up and down in opposition with low values of one variable associated with high values of the other, the correlation coefficient will be negative. r = 1 means the data is perfectly linear with a positive slope (i.e., both variables tend to change in the same direction)

r = -1 means the data is perfectly linear with a negative slope (i.e., both variables tend to change in different directions)

r = 0 means there is no linear association

r > 0 < 5 means there is a weak association

r > 5 < 8 means there is a moderate association r > 8 means there is a strong association

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

Where:

$N$ = number of pairs of scores

$\Sigma xy$ = sum of the products of paired scores

$\Sigma x$ = sum of x scores

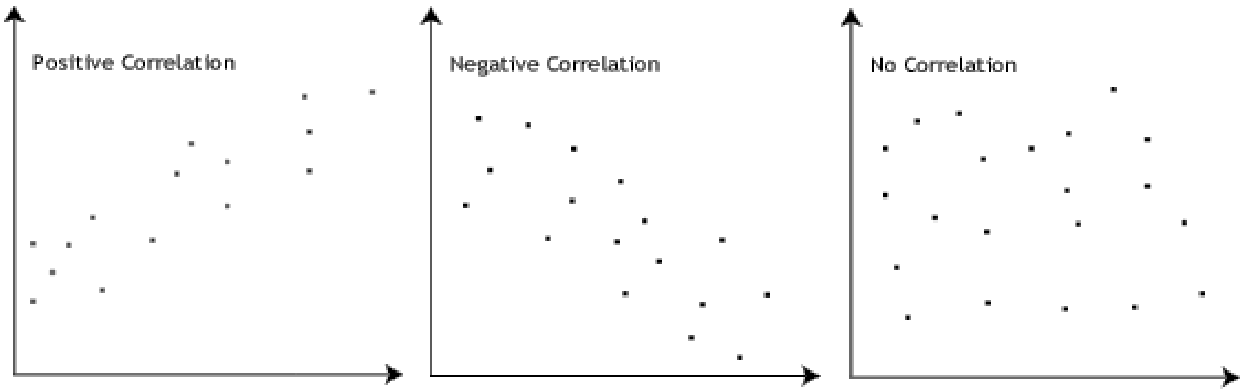
$\Sigma y$ = sum of y scores

$\Sigma x^2$ = sum of squared x scores

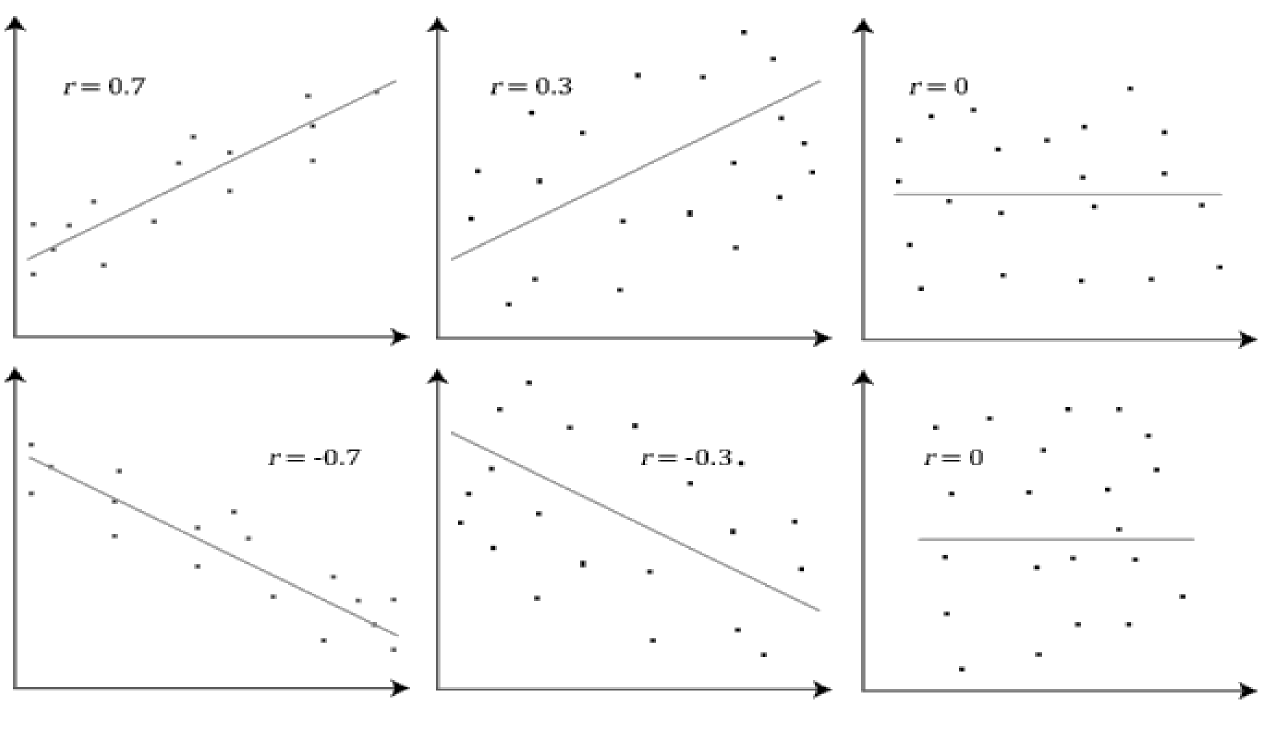
$\Sigma y^2$ = sum of squared y scores

The Pearson product-moment correlation coefficient (or Pearson correlation coefficient, for short) is a measure of the strength of a linear association between two variables and is denoted by *r*. Basically, a Pearson product-moment correlation attempts to draw a line of best fit through the data of two variables, and the Pearson correlation coefficient, *r*, indicates how far away all these data points are to this line of best fit (i.e., how well the data points fit this new model/line of best fit).

The Pearson correlation coefficient, *r*, can take a range of values from +1 to -1. A value of 0 indicates that there is no association between the two variables. A value greater than 0 indicates a positive association; that is, as the value of one variable increases, so does the value of the other variable. A value less than 0 indicates a negative association; that is, as the value of one variable increases, the value of the other variable decreases. This is shown in the diagram below:




The stronger the association of the two variables, the closer the Pearson correlation coefficient, *r*, will be to either +1 or -1 depending on whether the relationship is positive or negative, respectively. Achieving a value of +1 or -1 means that all your data points are included on the line of best fit – there are no data points that show any variation away from this line. Values for *r* between +1 and -1 (for example, *r* = 0.8 or -0.4) 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:



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

Ans:

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 is extremely important to rescale the variables so that they have a comparable scale. If we don't have comparable scales, then some of the coefficients as obtained by fitting the regression model might be very large or very small as compared to the other coefficients.

Normalized scaling means to scale a variable to have values between 0 and 1, while standardized scaling refers to transform data to have a mean of zero and a standard deviation of 1

***What?***

It is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm.

***Why?***

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.

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.

**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**).

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

- **sklearn.preprocessing.scale** helps to implement standardization in python.
- One disadvantage of normalization over standardization is that it **loses** some information in the data, especially about **outliers**.

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

Ans:

The Variance Inflation Factor (VIF) is a measure of colinearity among predictor variables within a multiple regression. It is calculated by taking the the ratio of the variance of all a given model's betas divide by the variane of a single beta if it were fit alone.

If there is perfect correlation, then VIF = infinity. A large value of VIF indicates that there is a correlation between the variables. An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables (which show an infinite VIF as well).

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

Ans:

Quantile-Quantile (Q-Q) plot, is a graphical tool to help us assess if a set of data plausibly 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.

This 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.

Few advantages:

a) It can be used with sample sizes also

b) Many distributional aspects like shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot.

It is used to check following scenarios: If two data sets.

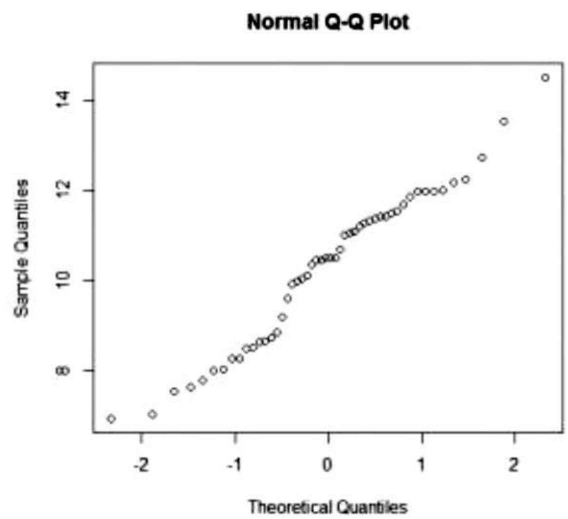
i. come from populations with a common distribution

ii. have common location and scale

iii. have similar distributional shapes

iv. have similar tail behavior

A Q-Q plot is a scatterplot 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 i.e.


Normal Q-Q Plot
Sample Quantiles
Theoretical Quantiles
14
12
10
8
-2
-1
0
1
2