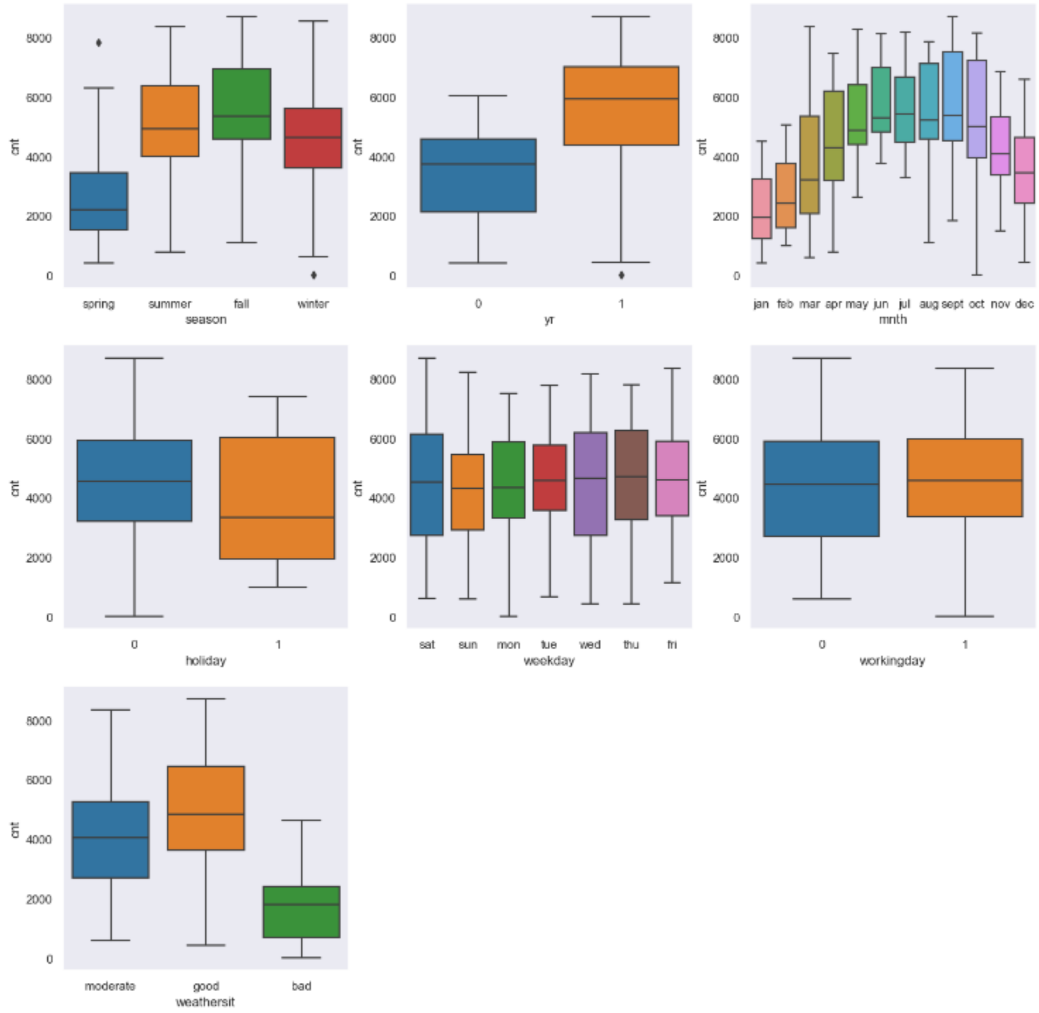
# Assignment-based Subjective Questions

**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:

There are a couple of categorical variables namely season, mnth, yr, holiday, weekday, working day and weathersit. These categorical variables have a major effect on the dependent variable ‘cnt’. The below fig shows the correlation among the same. These were visualized using a boxplot (Fig. attached).



These variables had the following effect on our dependent variable:

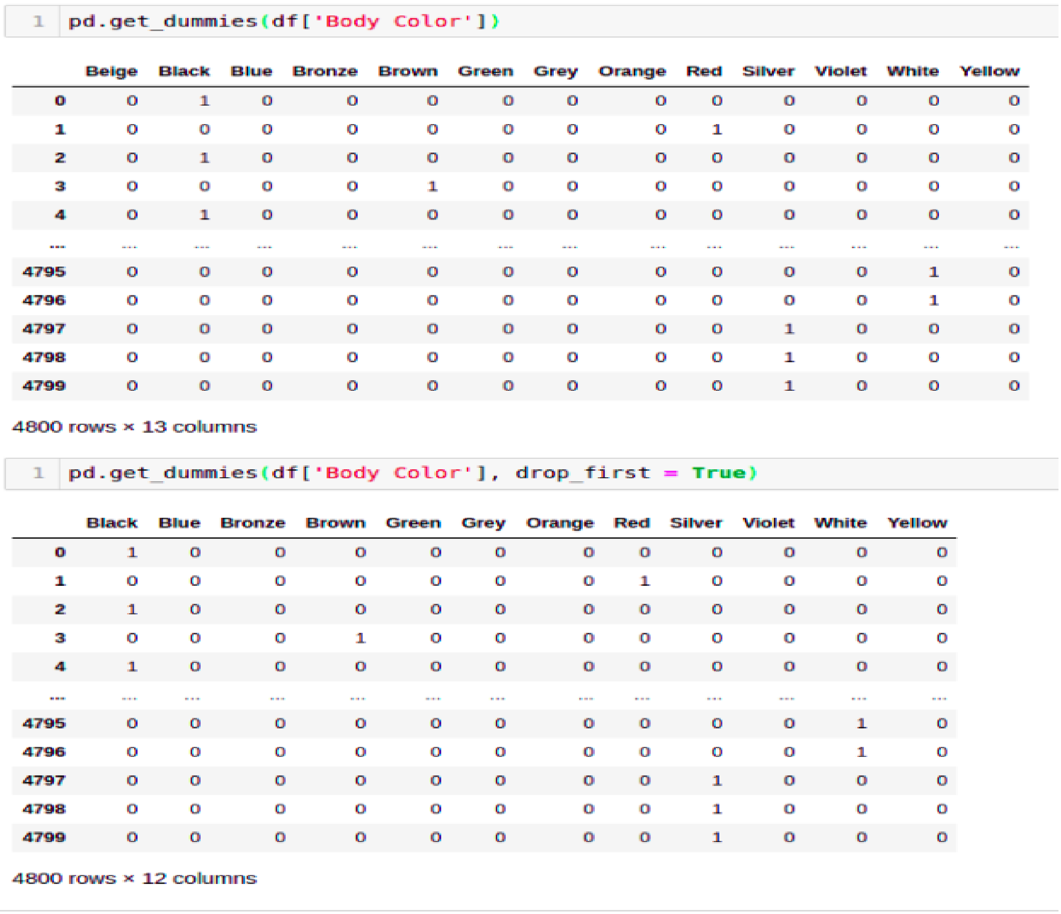
- Season - The boxplot showed that the spring season had the least value of cnt whereas fall had the maximum value of cnt. Summer and winter had intermediate values of cnt.
- Weathersit - There are no users when there is heavy rain/ snow indicating that this weather is extremely unfavorable. The highest count was seen when the weathersit was’ Clear, Partly Cloudy’.
- Yr - The number of rentals in 2019 was more than 2018.
- Holiday - rentals reduced during holiday.

- Mnth - September saw the highest no of rentals while December saw the least. This observation is in accordance with the observation made in weathersit. The weather situation in December is usually heavy snow due to which the rentals might have dropped.
- Weekday - The count of rentals is almost even throughout the week.
- Workingday – The median count of users is constant almost throughout the week.

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

**ANSWER:**

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. For example, iterative models may have trouble converging and lists of variable importance may be distorted. In the figure below, check the number of columns, instead of 13 we got 12 columns. It removes the first column of the get_dummies dataframe. The first column for the "Body Color" column is Beige. If there is a beige car, all columns are 0. When all columns are 0, the model knows it's a beige car. More columns mean less performance and more training time. Imagine we have 20 columns that are not numerical. If we use 'drop_first', we get 20 columns less. So it is useful to use the drop_first = True parameter for model performance.



```
pd.get_dummies(df['Body Color'])
```

| | Beige | Black | Blue | Bronze | Brown | Green | Grey | Orange | Red | Silver | Violet | White | Yellow |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **1** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| **2** | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **3** | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **4** | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| **4795** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| **4796** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| **4797** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| **4798** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| **4799** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |

4800 rows × 13 columns

```
pd.get_dummies(df['Body Color'], drop_first = True)
```

| | Black | Blue | Bronze | Brown | Green | Grey | Orange | Red | Silver | Violet | White | Yellow |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **1** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| **2** | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **3** | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **4** | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| **4795** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| **4796** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| **4797** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| **4798** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| **4799** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |

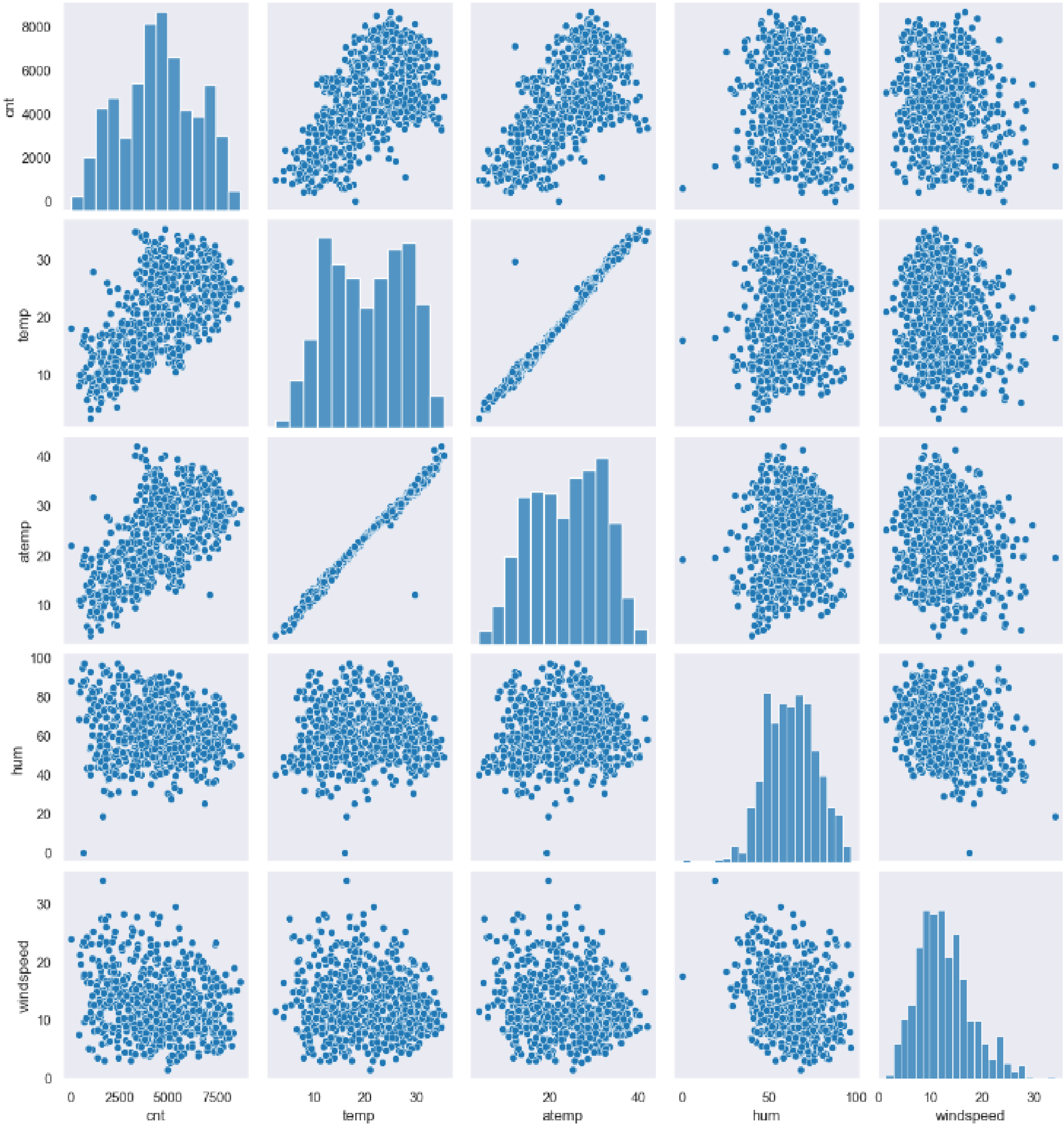
4800 rows × 12 columns

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

**ANSWER:**

Using the below PAIRPLOT and HEATMAP the "temp" and "atemp" are the two numerical variables that are highly correlated when compared to the rest with target variable as 'cnt'.

**PAIRPLOT analysis below:**



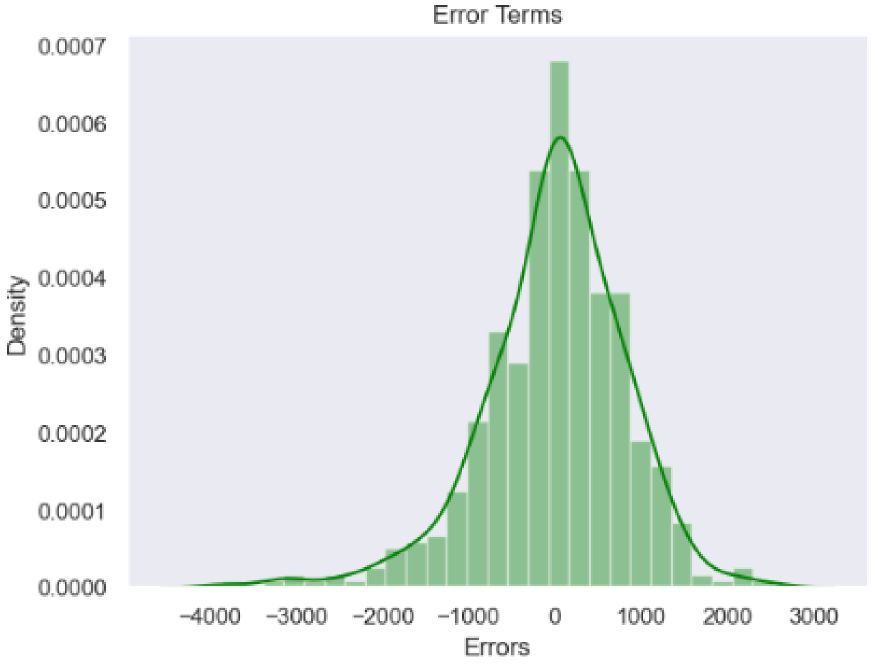
### 4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

**ANSWER:**

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

1. First, linear regression needs the relationship between the independent and dependent variables to be linear. We visualized the numeric variables using a pairplot to see if the variables were linearly related or not. Refer to the notebook for more details.

2. Secondly, residual distribution should follow a normal distribution and be centered around 0 (mean = 0). We validated this assumption about residuals by plotting a distplot of residuals and saw if residuals were following normal distribution or not. The diagram below shows that the residuals are distributed about mean = 0



3. Thirdly, 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.

### 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:**

Below are the top 3 features contributing significantly towards explaining the demand for the shared bikes:

1) **temp**: for a unit increase in the ambient temperature, the target variable increases.
2) **yr**: for a unit increase in the feature "yr", the target variable increases.
3) **clear_partlycloudy**: for a unit increase in this variable, the target variable increases.

# General Subjective Questions

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

**ANSWER:**

Linear Regression is a type of supervised Machine Learning algorithm that 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 variables. 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. The regression method tries to find the best-fit line which shows the relationship between the dependent variable and predictors with the 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: SLR is used when the dependent variable is predicted using only one independent variable.

The equation for SLR will be:

$$\overset{\text{Dependent Variable}}{Y_i} = \underbrace{\overset{\text{Population Y intercept}}{\beta_0} + \overset{\text{Population Slope Coefficient}}{\beta_1} \overset{\text{Independent Variable}}{X_i}}_{\text{Linear component}} + \underbrace{\overset{\text{Random Error term}}{\varepsilon_i}}_{\text{Random Error component}}$$

2. Multiple Linear Regression: MLR is used when the dependent variable is predicted using multiple independent variables.

The equation for MLR will be:

$$\textit{observed data} \rightarrow \quad y = b_0 + b_1x_1 + b_2x_2 + \cdots + b_px_p + \varepsilon$$

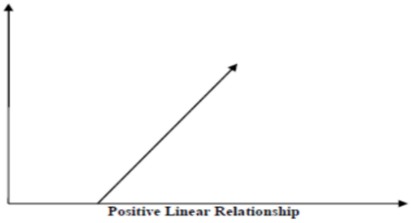
$$\textit{predicted data} \rightarrow \quad y' = b_0 + b_1x_1 + b_2x_2 + \cdots + b_px_p$$

$$\textit{error} \rightarrow \quad \varepsilon = y - y'$$

B1 = coefficient for X1 variable
B2 = coefficient for X2 variable
B3 = coefficient for X3 variable and so on…
B0 is the intercept (constant term)

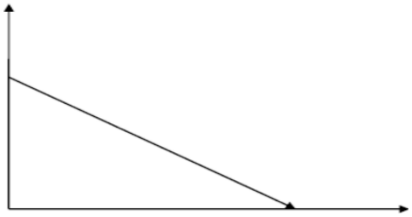
Furthermore, the linear relationship can be positive or negative in nature as explained below−

- Positive Linear Relationship:
  - A linear relationship will be called positive if both independent and dependent variable increases. It can be understood with the help of the following graph –



Positive Linear Relationship

- Negative Linear relationship:
  - A linear relationship will be called positive if the independence increases and the dependent variable decreases. It can be understood with the help of the following graph –



Negative Linear Relationship

Linear regression is of the following two types –

- Simple Linear Regression
- Multiple Linear Regression

Assumptions –

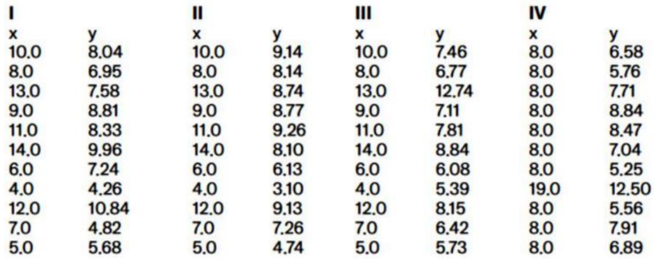
The following are some assumptions about the dataset that are made by the Linear Regression model :

- Multi-collinearity –
  - The 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 a dependency in them.
- Auto-correlation –
  - Another assumption the Linear regression model assumes is that there is very little or no auto-correlation in the data. Basically, auto-correlation occurs when there is a dependency between residual errors.
- Relationship between variables –
  - The linear regression model assumes that the relationship between response and feature variables must be linear.
- Normality of error terms –
  - Error terms should be normally distributed ⎫ Homoscedasticity – o There should be no visible pattern in residual values.

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

### ANSWER:

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 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 |

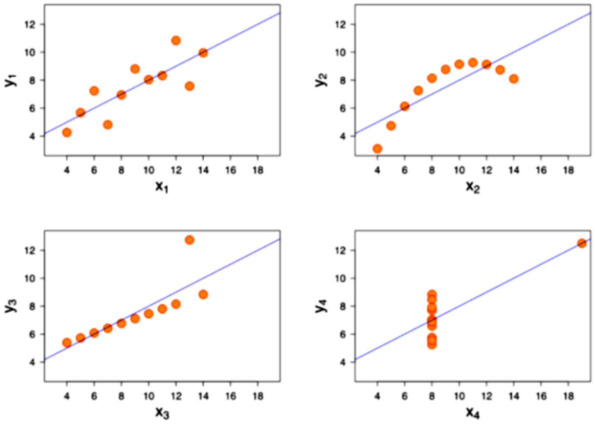
All the summary statistics for each dataset are identical.

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.

However, 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 modeled 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)

### ANSWER:

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

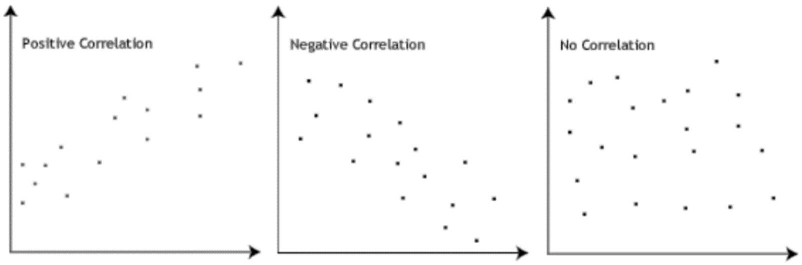
$\Sigma x^2$ = sum of squared x scores

$\Sigma y^2$ = 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).

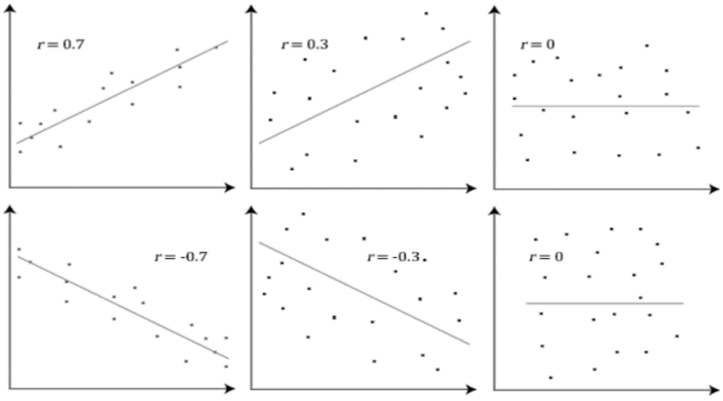
The Pearson correlation coefficient, r, can take a range of values from +1 to -1. A value of 0 indicates that there is no association between the two variables. A value greater than 0 indicates a positive association; that is, as the value of one variable increases, so does the value of the other variable. 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)

### ANSWER:

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 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 is scaling?**

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 is scaling performed?**

Most of the time, the collected data set contains features highly varying in magnitudes, units, and range. If scaling is not done, then the algorithm only takes magnitude into 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.

**Difference between normalized scaling and standardized scaling:**

| S.NO. | Normalized scaling | Standardized scaling |
|---|---|---|
| 1 | Minimum and maximum values of features are used for scaling. | The mean and standard deviation is used for scaling. |
| 2 | It is used when features are of different Scales. | It is used when we want to ensure zero mean and unit standard deviation. |
| 3 | Scales values between [0, 1] or [-1, 1]. | It is not bound to a certain range. |
| 4 | It is really affected by outliers. | It is much less affected by outliers. |
| 5 | Scikit-Learn provides a transformer called MinMaxScaler for Normalization. | Scikit-Learn provides a transformer called StandardScaler for standardization. |

**Formula:**

- In **Normalized scaling** the sklearn.preprocessing.MinMaxScaler helps to implement normalization in python.

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

- **Standardization Scaling** 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).

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

  - 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)

## ANSWER:

VIF - Variance Inflation Factor

The VIF gives how much the variance of the coefficient estimate is being inflated by collinearity. If there is a perfect correlation, then VIF = infinity. It gives a basic quantitative idea about how much the feature variables are correlated with each other. It is an extremely important parameter to test our linear model.

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

Where R-1 is the R-square value of that independent variable, we want to check how well this independent variable is explained by other independent variables. If that independent variable can be explained perfectly by other independent variables, then it will have a perfect correlation, and its R-squared value will be equal to 1. So, VIF = 1/ (1-1) which gives VIF = 1/0 which results in "infinity"

The numerical value for VIF tells you (in decimal form) what percentage of the variance (i.e., the standard error squared) is inflated for each coefficient. For example, a VIF of 1.9 tells you that the variance of a particular coefficient is 90% bigger than what you would expect if there was no multicollinearity — if there was no correlation with other predictors.

A **rule of thumb** for interpreting the VIF (variance inflation factor):

- 1 = not correlated.
- Between 1 and 5 = moderately correlated.
- Greater than 5 = highly correlated.

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

## ANSWER:

A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set. It is used to compare the shapes of distributions. 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.

**Some advantages of q-q plot:**

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

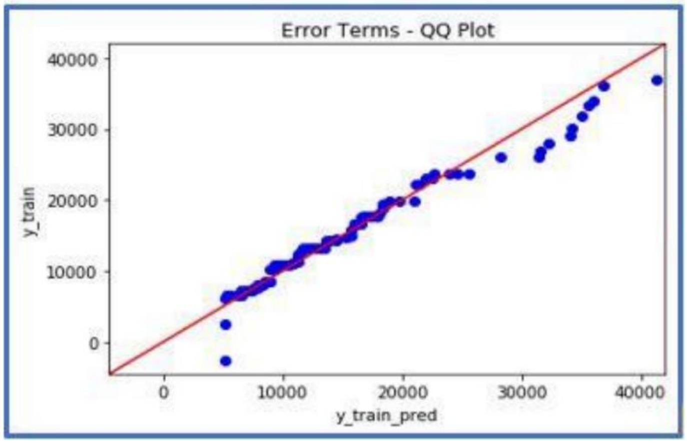
**The q-q plot is used to answer the following questions:**

- Do two data sets come from populations with a common distribution?
- Do two data sets have a common location and scale?
- Do two data sets have similar distributional shapes?
- Do two data sets have similar tail behavior?

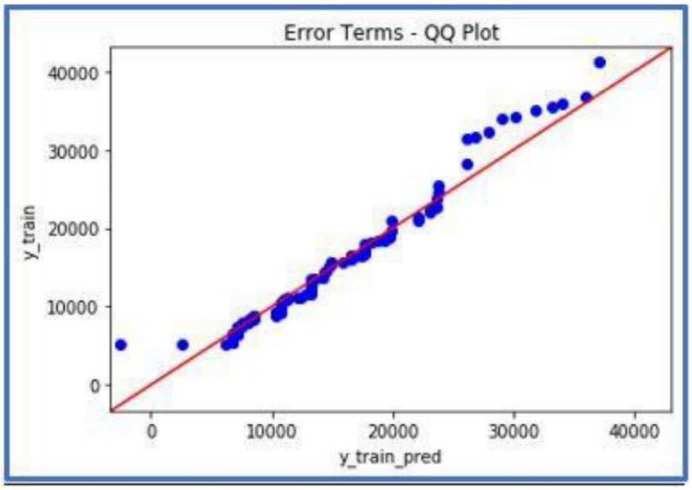
**Below are the possible interpretations for two data sets using a Q-Q plot:**

A) Similar distribution: If all point of quantiles lies on or close to a straight line at an angle of 45 degrees from x -axis

B) Y-values < X-values: If y-quantiles are lower than the x-quantiles.



c) X-values < Y-values: If x-quantiles are lower than the y-quantiles.


Error Terms - QQ Plot
y_train
y_train_pred