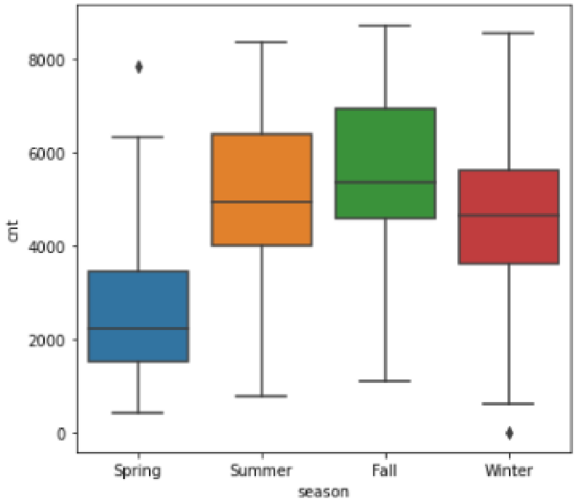
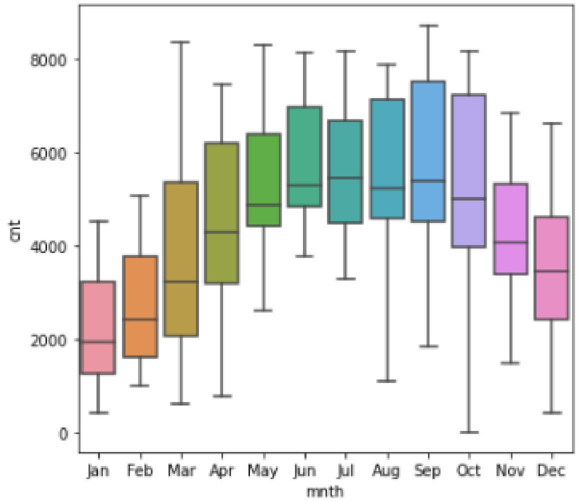
Bike Sharing Assignment – Linear Regression

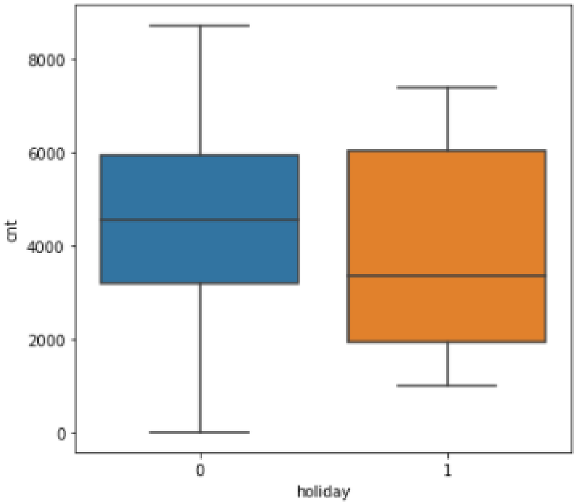
Assignment-based Subjective Questions

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: Categorical variables given in the dataset are season, yr, mnth, holiday, weekday, workingday and weathersit. We can use box plot to see their effect on target variable cnt.



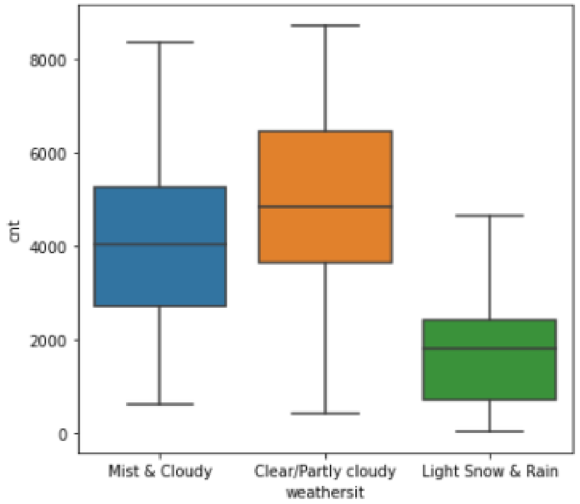


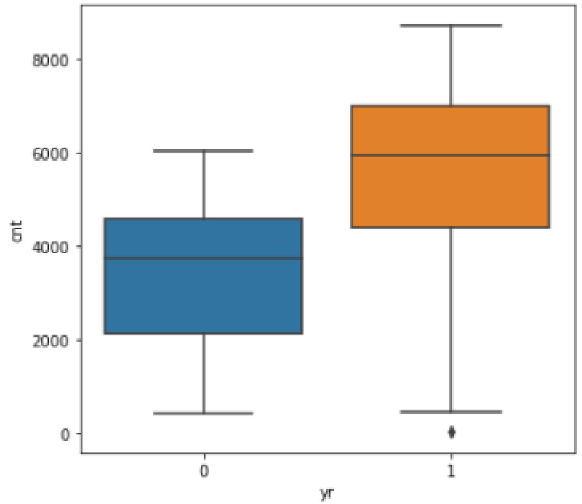


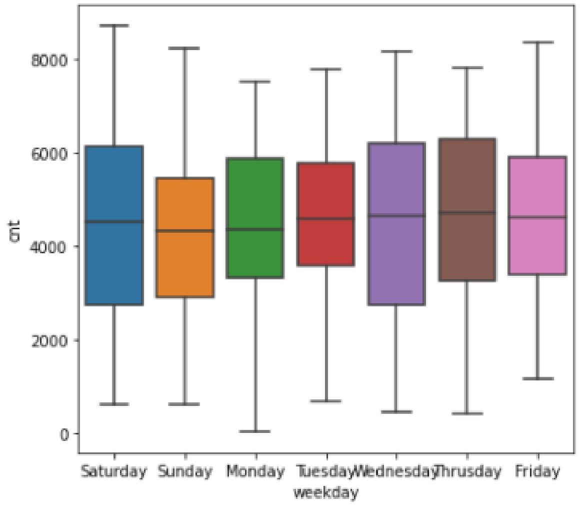
Season : The boxplot showed that spring season had least value of cnt whereas fall had maximum value of cnt, Summer and winter had intermediate value of cnt.

Mnth : September saw highest no of rentals while December saw least.

Holiday : rentals reduced during holiday







Weathersit: There are no users when there is heavy rain/ snow, Highest count was seen when the weathersit was ‘Clear/Partly cloudy’

Yr: The number of rentals in 2019 was more than 2018

Weekday: Bike rentals are high on Saturdays and Low on Mondays.

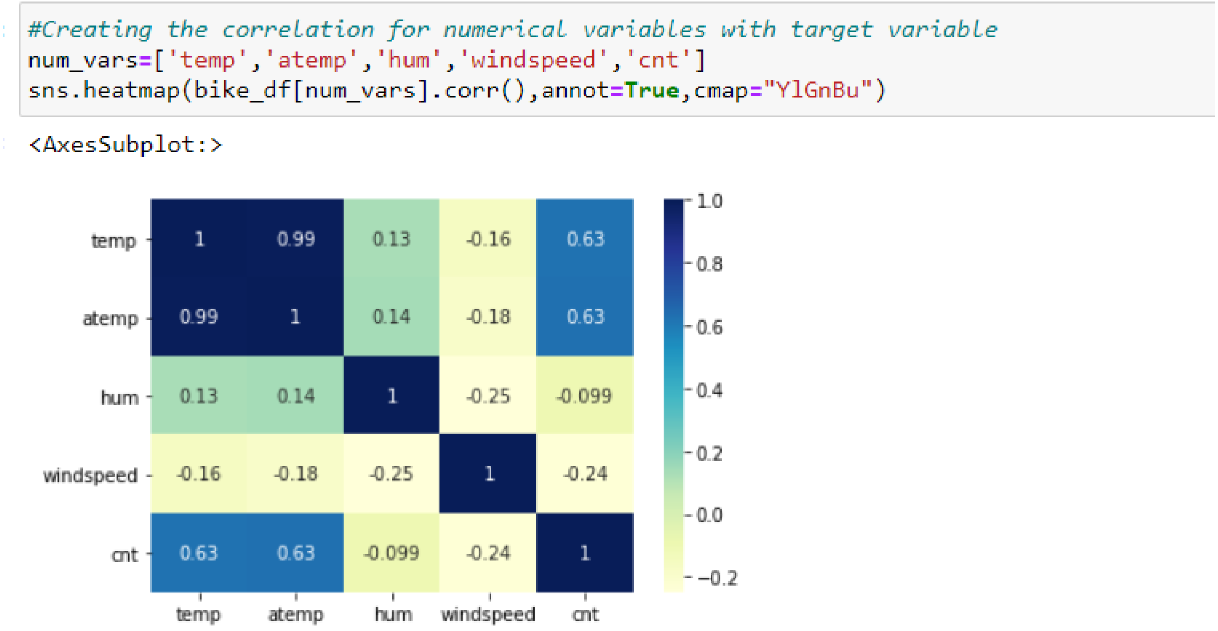
2. Why is it important to use drop_first=True during dummy variable creation? (2 mark)

Ans: 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

If we have a categorical variables with n-levels, then we need to use n-1 columns to represent the dummy variables.

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

Ans: ‘temp’ and ‘atemp’ are two numerical variables which are highly correlated with target variable cnt.



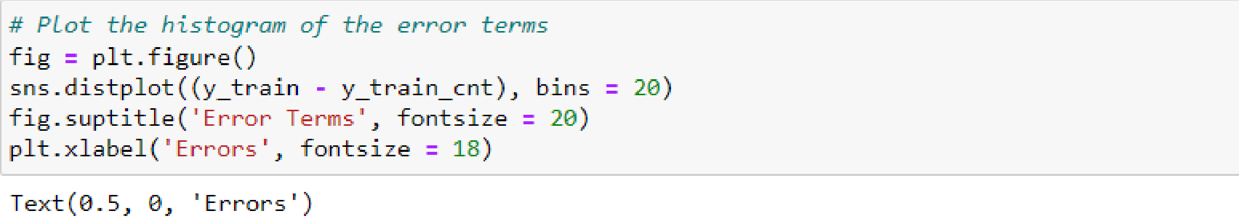
```
#Creating the correlation for numerical variables with target variable
num_vars=['temp','atemp','hum','windspeed','cnt']
sns.heatmap(bike_df[num_vars].corr(),annot=True,cmap="YlGnBu")
```

```
<AxesSubplot:>
```

temp 1 0.99 0.13 -0.16 0.63
atemp 0.99 1 0.14 -0.18 0.63
hum 0.13 0.14 1 -0.25 -0.099
windspeed -0.16 -0.18 -0.25 1 -0.24
cnt 0.63 0.63 -0.099 -0.24 1
temp
atemp
hum
windspeed
cnt
1.0
0.8
0.6
0.4
0.2
0.0
−0.2

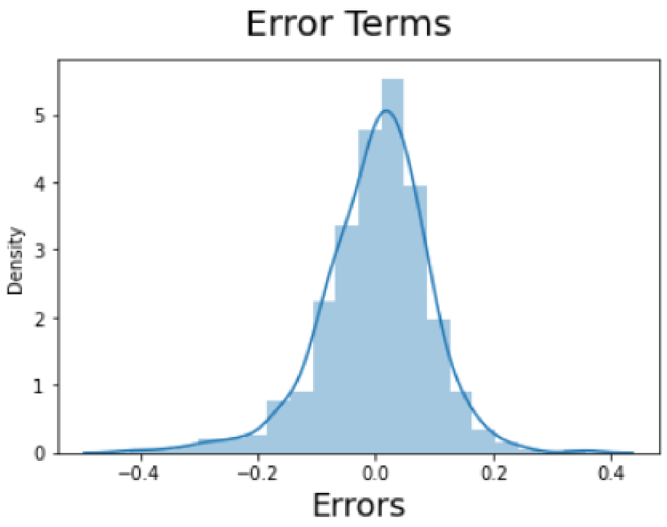
4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

Ans: Residual Distribution shows follow normal distributions and mean was centered around 0. In python sheet, we checked this assumption by plotting residuals with help of distplot and found residual follows normal distribution.



```
# Plot the histogram of the error terms
fig = plt.figure()
sns.distplot((y_train - y_train_cnt), bins = 20)
fig.suptitle('Error Terms', fontsize = 20)
plt.xlabel('Errors', fontsize = 18)
```

```
Text(0.5, 0, 'Errors')
```



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: Top 3 positive features are

Features	Co-efficient
Temp	0.491508
Yr	0.233482
Season_winter	0.083084

Top 3 negative features are

Features	Co-efficient
weathersit_Light Snow & Rain	-0.285155
windspeed	-0.147977
holiday	-0.098013

General Subjective Questions

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

Ans: Linear regression is one of the easiest and most popular Machine Learning algorithms. It is a statistical method that is used for predictive analysis. Linear regression makes predictions for continuous/real or numeric variables such as **sales, salary, age, product price,** etc.

Linear regression algorithm shows a linear relationship between a dependent (y) and one or more independent (y) variables, hence called as linear regression. Since linear regression shows the linear relationship, which means it finds how the value of the dependent variable is changing according to the value of the independent variable.

The linear regression model provides a sloped straight line representing the relationship between the variables. Consider the below image:

There are two types of linear regression – Simple and Multiple.

Simple Linear Regression : for example, if we have dataset to find relationship between 'Count of Newspaper advertisement' and 'sales'.

Goal is to design a model that can predict sales if given the 'Count of newspaper advertisement'. Using the training data, a regression line is obtained which will give minimum error. This Linear equation is then used for any new data. That is, if we give 'count of newspaper advertisement' by as an input, out model should predict their sales with minimum error.

Y(pred) = b0 + b1*x

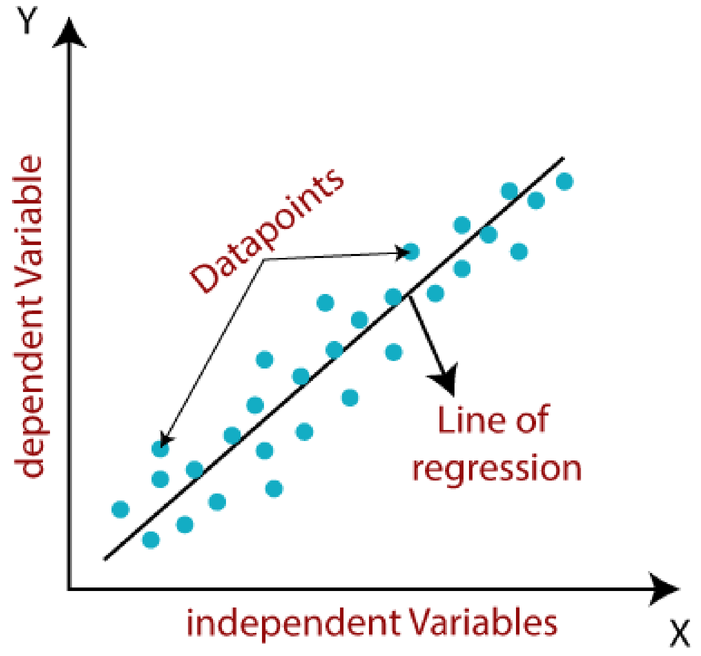
Value of b0 and b1 will be such so that error will be minimized.

Observation on b1:

If b1>0 then x(predictor) and y(target) have positive relationship. That increase in x will increase y.

If b<0 then x(predictor) and y(target) have negative relationship. That increase in x will decrease y.

b0 – intercept/Constant.



There are two types of linear regression – Simple and Multiple.

Multiple Linear Regression : it is used when dependent variable is predicted using multiple independent variables.

Equation of MLR will be Y(pred) = b0 + b1*x + b2*x2+b3*x3..

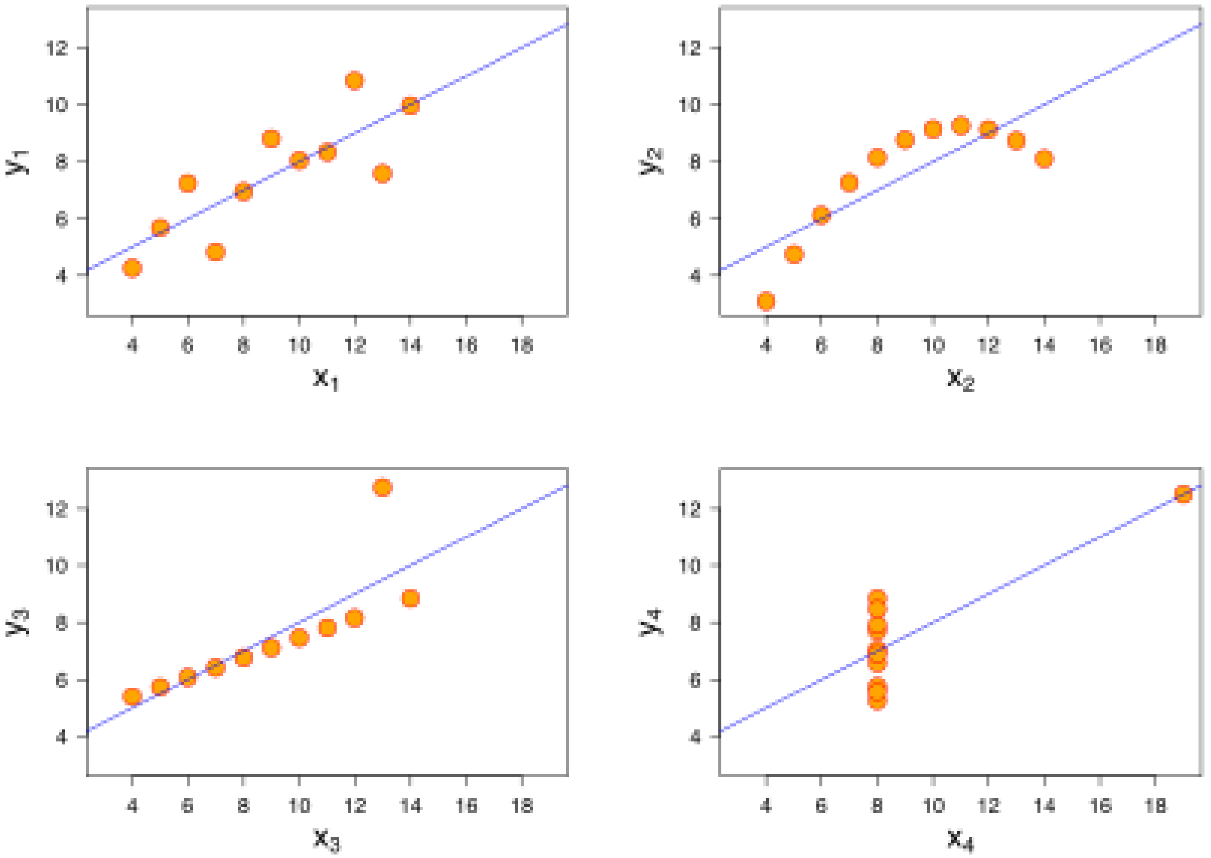
b1- co efficient of x1, b2-co efficient of x2, b3- co efficient of x3.

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

Ans: Anscombe's Quartet was developed by statistician Francis Anscombe. It includes four data sets that have almost identical statistical features, but they have a very different distribution and look totally different when plotted on a graph. It was developed to emphasize both the importance of graphing data before analyzing it and effect of outliers and other influential observations on statistical properties.

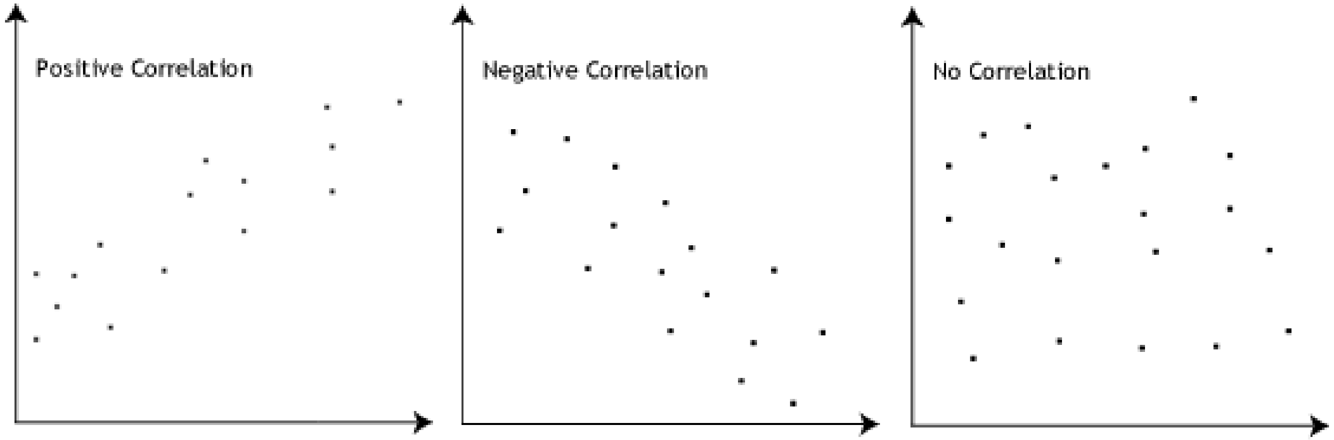
- The first scatter plot x1 vs y1 appears to be a simple linear relationship.
- The second graph x2 vs y2 is not distributed normally, while there is a relation between them, it's not linear.
- In the third graph x3 vs y3, the distribution is linear, but should have a different regression line The calculated 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 x4 vs y4 shows an example when one high-leverage point is enough to product a high correlation coefficient, even through the other data points do not indicate any relationship between the variables.

In brief, four datasets that were intentionally created to describe the importance of data visualization and how any regression algorithm can be fooled by the same. Hence, all the important features in the dataset must be visualized before implementing any machine learning algorithm on them which will help to make a good fit model.



3. What is Pearson’s R? (3 marks)

Ans: 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 donated 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 fir (i.e. how well the data points fit this new model/line of best fit).



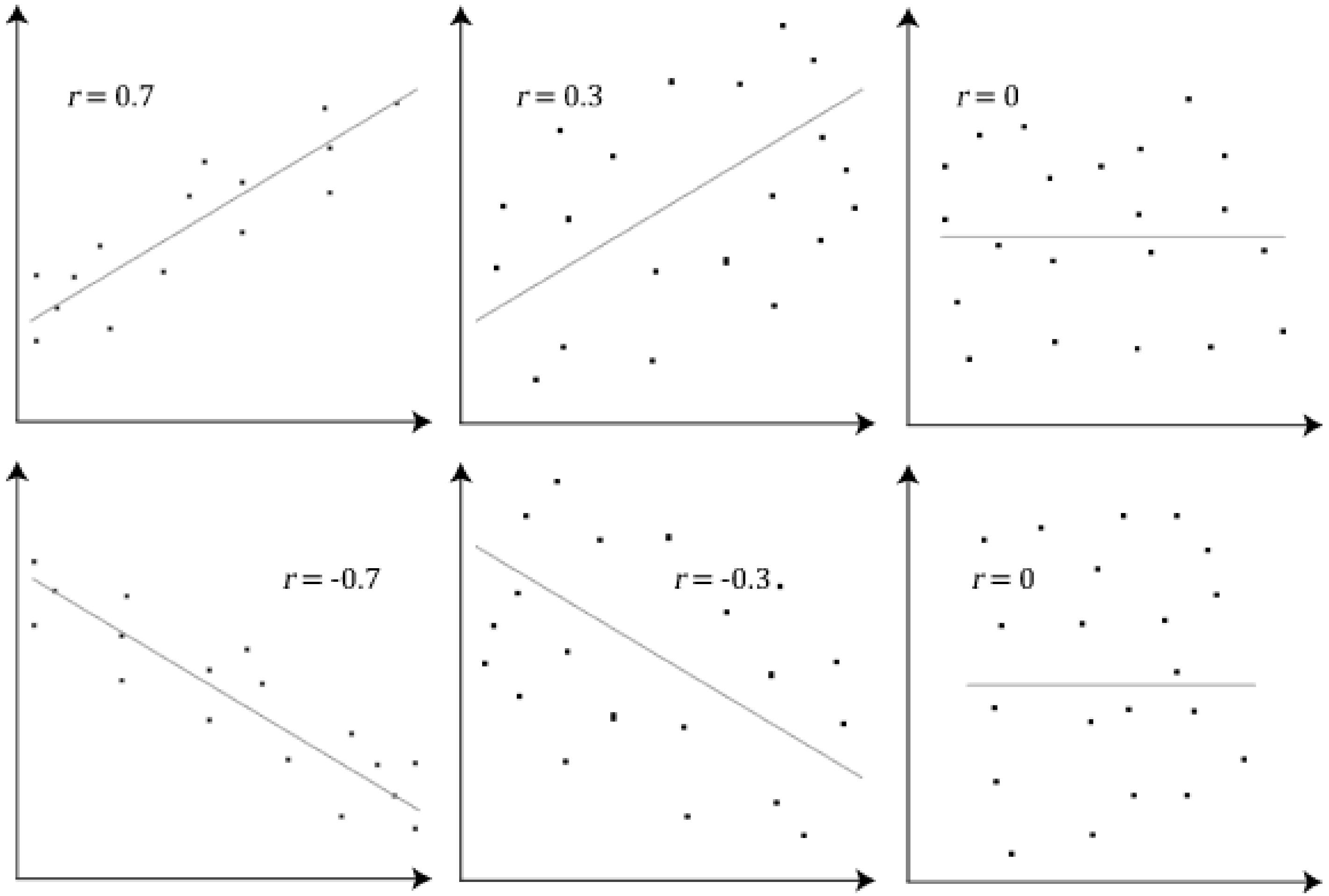
r=1 means data is perfectly linear with positive slope.

r=-1 means data is perfectly linear with negative slope.

r=0 means there is no linear relationship.

determine the strength of association based on the Pearson correlation coefficient

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:


r = 0.7
r = 0.3
r = 0
r = -0.7
r = -0.3
r = 0

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

Ans: 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. 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.

Example: If an algorithm is not using the feature scaling method then it can consider the value 3000 meters to be greater than 5 km but that’s actually not true and in this case, the algorithm will give wrong predictions. So, we use Feature Scaling to bring all values to the same magnitudes and thus, tackle this issue.

Techniques to perform Feature scalling:

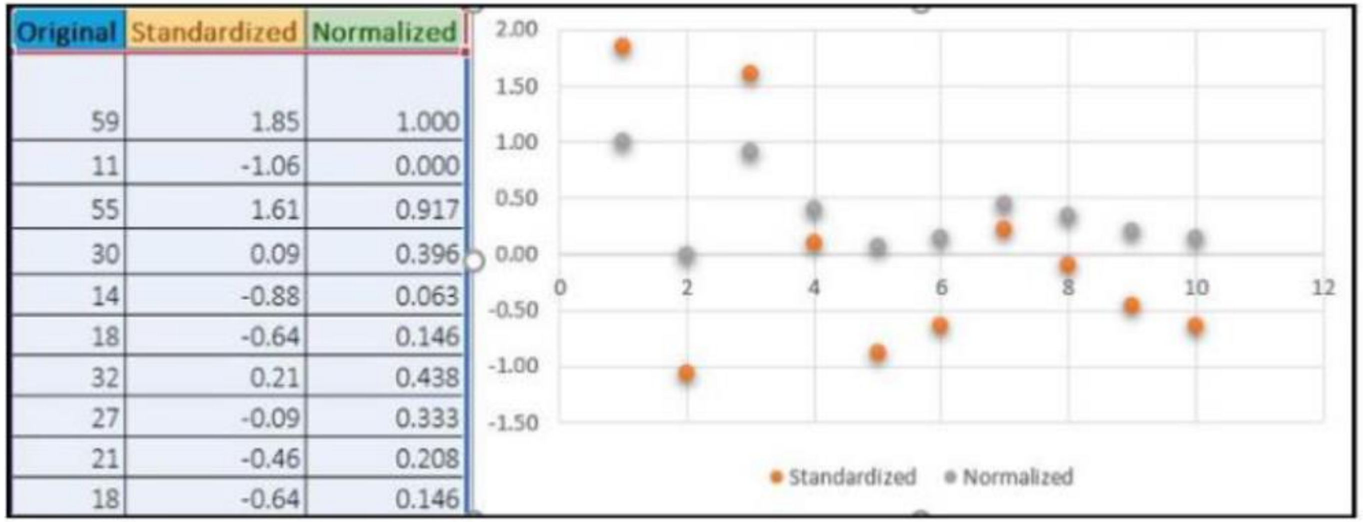
Min-Max Normalization: This technique re-scales a feature or observation value with distribution value between 0 and 1.

$$X_{\text{new}} = \frac{X_i - \min(X)}{\max(x) - \min(X)}$$

Standardization: It is a very effective technique which re-scales a feature value so that it has distribution with 0 mean value and variance equals to 1.

$$X_{\text{new}} = \frac{X_i - X_{\text{mean}}}{\text{Standard Deviation}}$$

One disadvantage of normalization over standardization is that is losses some information in data, especially outliers.
Below Example shows 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

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

Ans: VIF is an index that provides a measure of how much the variance of an estimated regression coefficient increases due to collinearity. In order to determine VIF, we fit a regression model between the independent variables. For example, we would fit the following models to estimate the coefficient of determination R1 and use this value to estimate theVIF:

X_1=C+α_2X_2+α_3X_3+⋯

〖VIF〗_1=1/(1-R_1^2)

Next, we fit the model between X2 and the other independent variables to estimate the coefficient of determination R2:

X_2=C+ α_1 X_1+α_3 X_3+⋯

〖VIF〗_2=1/(1-R_2^2)

VIF is infinite means there is a perfect relationship between two independent variables. In case of perfect co-relation we get R value as 1. which leads to VIF infinite, to overcome this we need to drop one of variables from dataset which is causing perfect multi-co linearity.

In Short, an infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables.

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

Ans: The quantile-quantile (q-q) plot is a graphical technique for determining if two data sets come from populations with a common distribution.

A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set. By a quantile, we mean the fraction (or percent) 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 advantages of the q-q plot are:

1. The Sample sizes do not need to be equal.
2. Many distributional aspects can be simultaneously tested. For example, shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot. For example, if the two data sets come from populations whose distributions differ only by a shift in location, the points should lie along a straight line that is displaced either up or down from the 45-degree reference line.

The q-q plot is formed by:

Vertical axis: estimated quantiles from data set 1

Horizontal axis: estimated quantiles from data set 2

Both axes are in units of their respective data sets. That is, the actual quantile level is not plotted. For a given point on the q-q plot, we know that the quantile level is the same for both points. But now what that quantile level actually is.

If the data sets have the same size, the q-q plot is essential a plot of sorted data set 1 against sorted data set 2. if the data sets are not of equal size, the quantiles are usually picked to correspond to the sorted values from the smaller data set and then the quantiles for the larger data set are interpolated.

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 common location and scale?
- Do two data sets have similar distributional shapes?
- Do two data sets have similar tail behaviour?