International
Institute of Information
Technology Bangalore

upGrad

Bike sharing Assignment

Prepared and submitted by

FAHAD BAIG

Assignment-based Subjective Questions

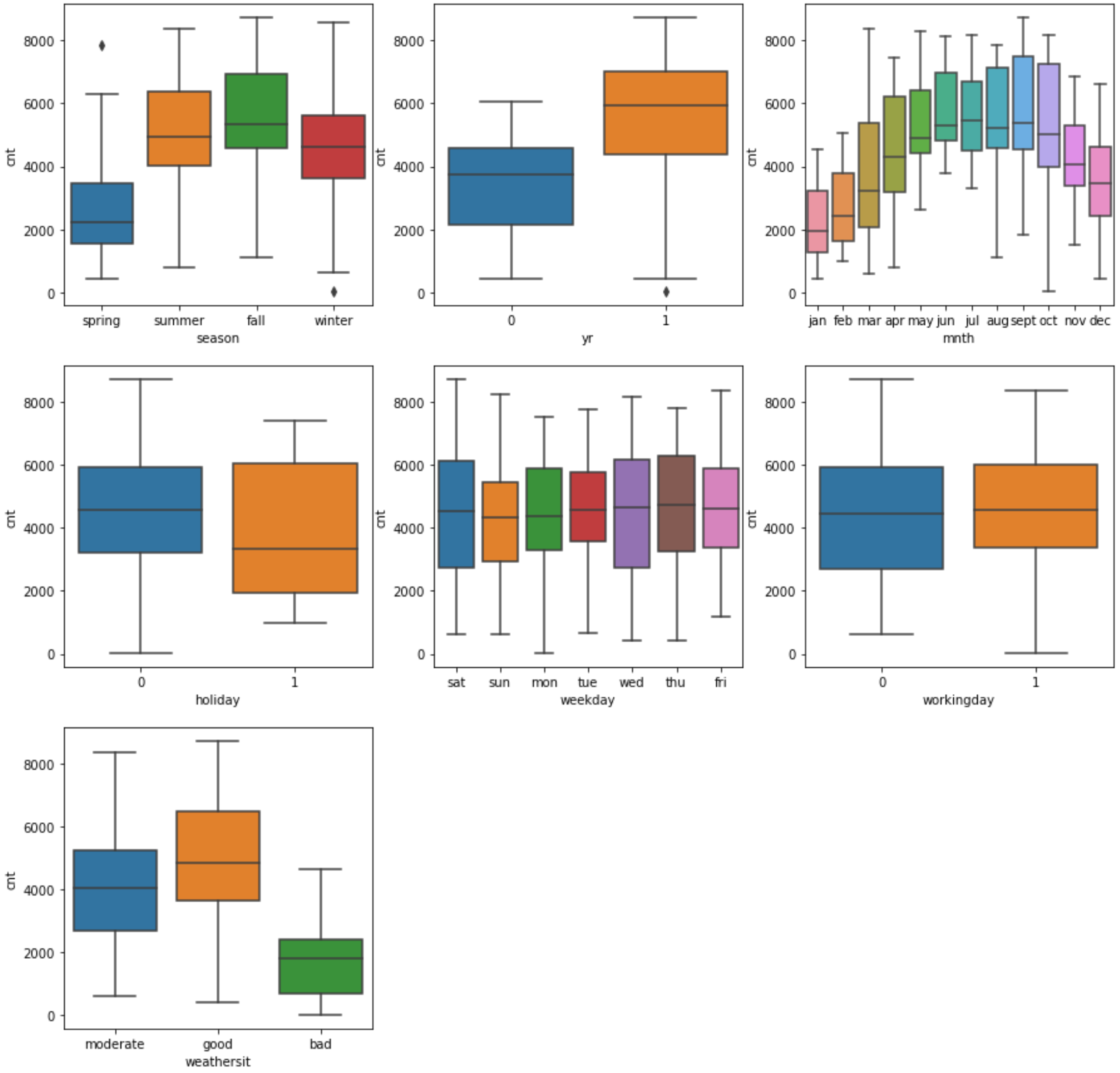
1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)

Observations:-

- Season: 3:fall has highest demand for rental bikes.
- I see that demand for next year has grown.
- Demand is continuously growing each month till June, then September month has highest demand and demand is increasing after September.
- When there is a holiday, demand has decreased.
- Weekday is not giving clear picture about demand.
- It is clear that weather situation has highest demand
- Hence, We can conclude that, During September, bike sharing is more. During the year end and beginning, it is less, could be due to extreme weather conditions.

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

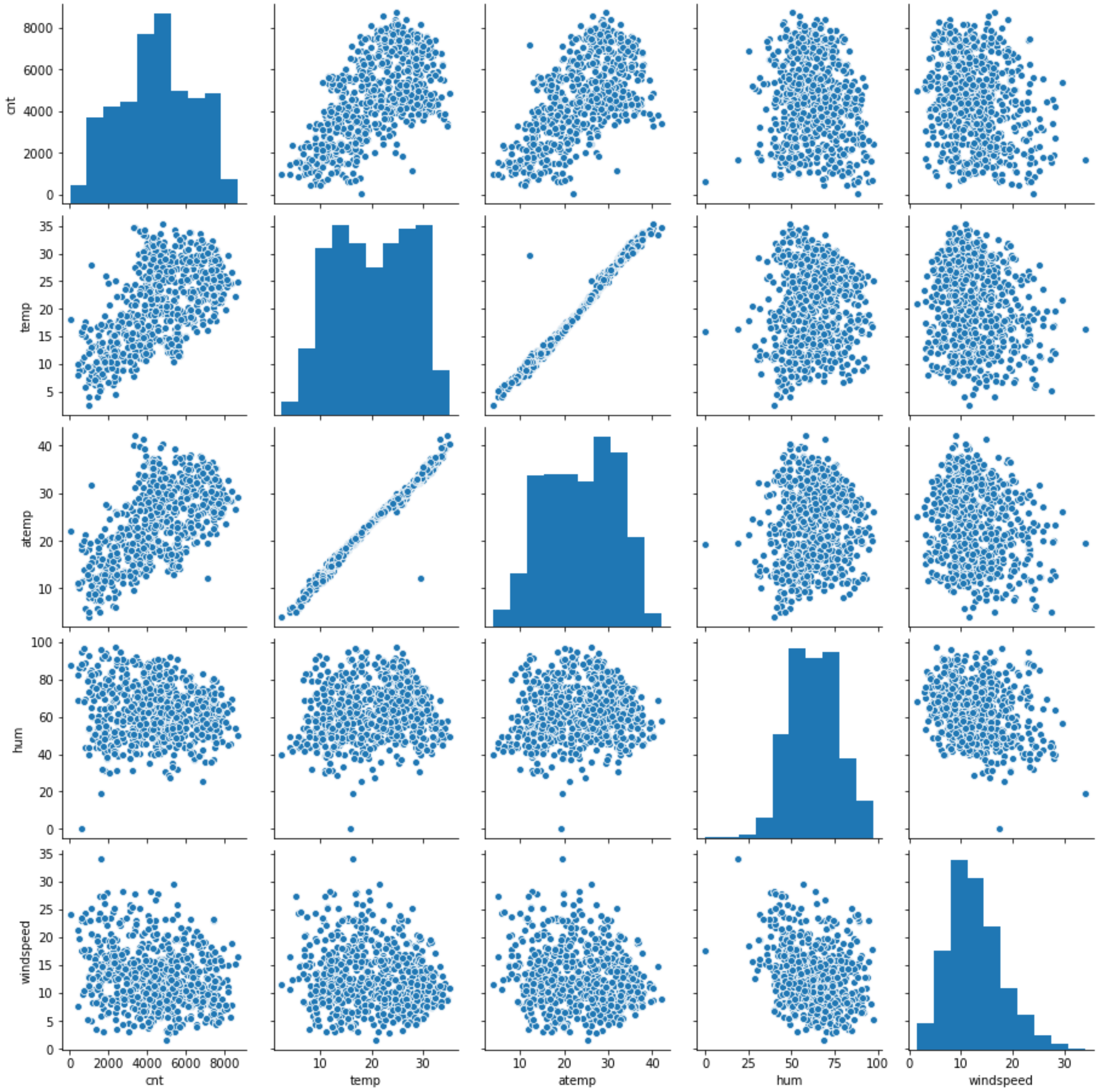
drop_first = True is important to use, as it helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables.



Assignment-based Subjective Questions

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

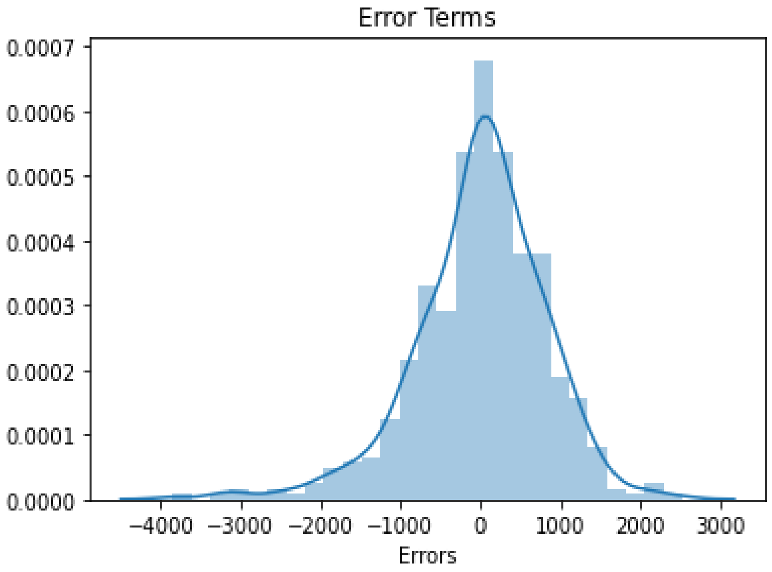
- Looks like the temp and atemp has the highest corelation with the target variable cnt.
- temp and atemp are highly co-related with each other

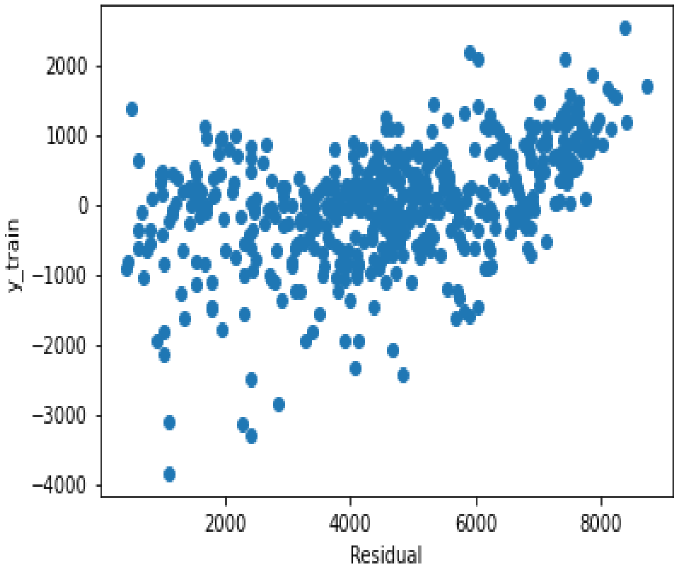


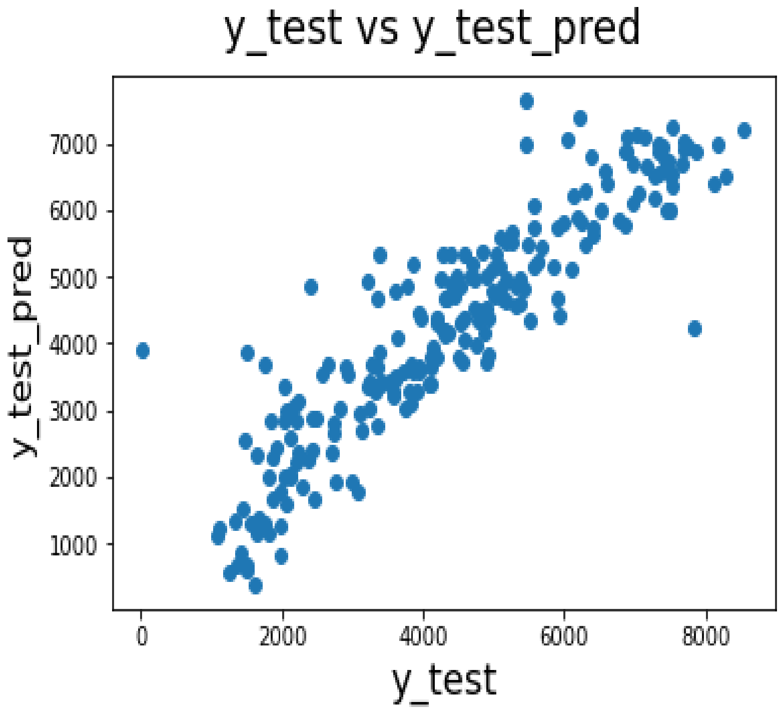
Assignment-based Subjective Questions

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

Following predictions were made :-







Normality of error terms

- Error terms should be normally distributed

Linearity Check

- There should be insignificant linearity among variables

Independence of residuals

o No auto-correlation

Homoscedasticity

o There should be no visible pattern in residual values.

Assignment-based Subjective Questions

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

As per the Final Model is designed, the features which are contributing significantly towards explaining the demand of the shared bikes are :-

- ✓ Temp
- ✓ Weathersit(Light Snow, Mist + Cloudy)
- ✓ Months(January, July, September, November, December) i.e., nothing but Years (as we can see September has the more demand for bikes)

But there are still more features which can be explaining us the demand of the shared bikes which are as follows:-

- ➢ Holiday
- ➢ Hum
- ➢ Windspeed
- ➢ Season
- ➢ Sunday

General Subjective Questions

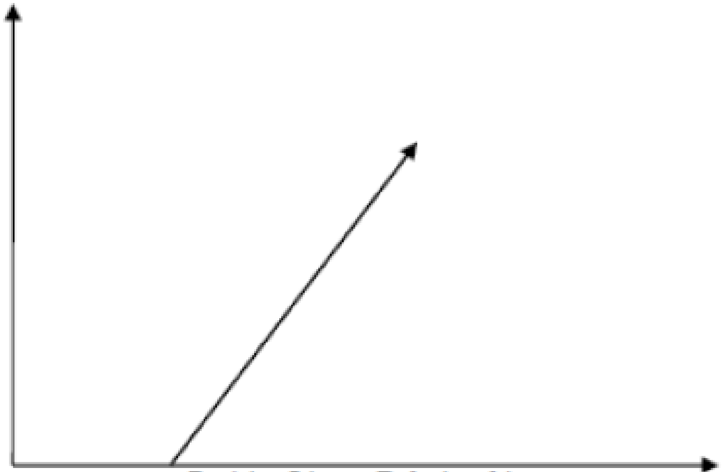
1. Explain the linear regression algorithm in detail. (4 marks)

Linear regression may be defined as the statistical model that a linear relationship between a dependent variable with given set of indepenc variables. Linear relationship between variables means that when the value more independent variables will change (increase or decrease), the value of dependent variable will also change accordingly (increase or decrease). Mathematically the relationship can be represented with the help of following equation − $Y = mX + c$

Here, **Y** is the dependent variable we are trying to predict.
X is the independent variable we are using to make predictions.
m is the slope of the regression line which represents the effect X has on Y
c is a constant, known as the Y-intercept. If X = 0, Y would be equal to c.

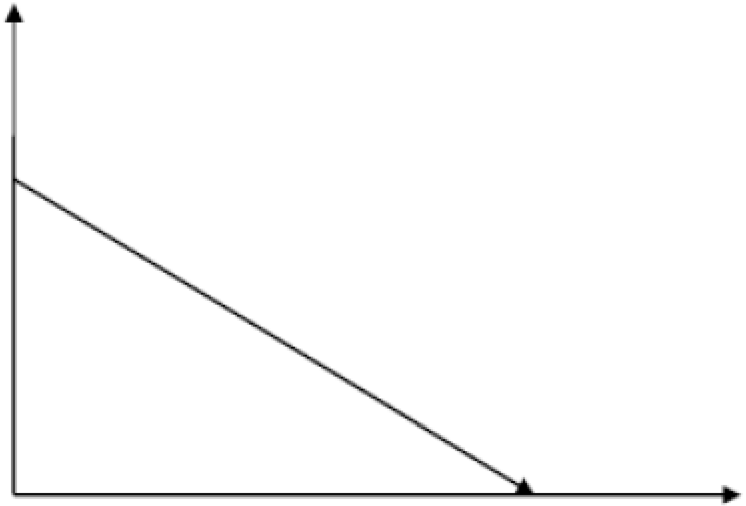
Furthermore, the linear relationship can be positive or negative in nature as ex below−

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



Positive Linear Relationship

- Negative Linear relationship:
  A linear relationship will be called positive if independent increase dependent variable decreases. It can be understood with the help of following



Negative Linear Relationship

General Subjective Questions

Linear regression is of the following two types –

- Simple Linear Regression
- Multiple Linear Regression

Assumptions -
The following are some assumptions about dataset that is made by Linear Regression model –

- Normality of error terms –

Error terms should be normally distributed

- Linearity Check –

There should be insignificant linearity among variables

- Homoscedasticity –

There should be no visible pattern in residual values.

- Independence of residuals –

No auto-correlation

General Subjective Questions

	I		II		III		IV	
	x	y	x	y	x	y	x	y
	10	8,04	10	9,14	10	7,46	8	6,58
	8	6,95	8	8,14	8	6,77	8	5,76
	13	7,58	13	8,74	13	12,74	8	7,71
	9	8,81	9	8,77	9	7,11	8	8,84
	11	8,33	11	9,26	11	7,81	8	8,47
	14	9,96	14	8,1	14	8,84	8	7,04
	6	7,24	6	6,13	6	6,08	8	5,25
	4	4,26	4	3,1	4	5,39	19	12,5
	12	10,84	12	9,13	12	8,15	8	5,56
	7	4,82	7	7,26	7	6,42	8	7,91
	5	5,68	5	4,74	5	5,73	8	6,89
SUM	99,00	82,51	99,00	82,51	99,00	82,50	99,00	82,51
AVG	9,00	7,50	9,00	7,50	9,00	7,50	9,00	7,50
STDEV	3,32	2,03	3,32	2,03	3,32	2,03	3,32	2,03

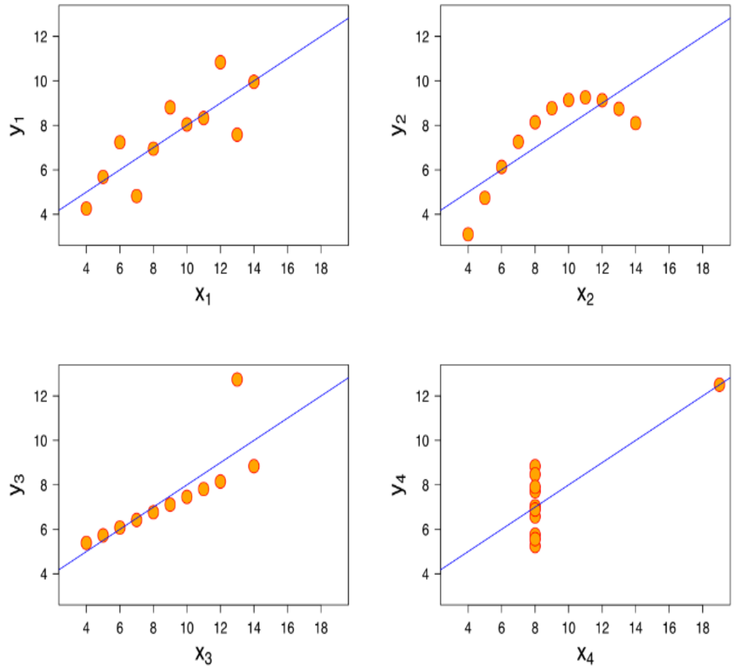
2. Explain the Anscombe's quartet in detail. (3 marks)

Anscombe's Quartet was developed by statistician Francis Anscombe. It comprises four datasets, each containing eleven (x, y) pairs. The essential thing to note about these datasets is that they share the same descriptive statistics. But things change completely, and I must emphasize COMPLETELY, when they are graphed. Each graph tells a different story irrespective of their similar summary statistics.

The summary statistics show that the means and the variances were identical for x and y across the groups:

- Mean of x is 9 and mean of y is 7.50 for each dataset.
- Similarly, the variance of x is 11 and variance of y is 4.13 for each dataset
- The correlation coefficient (how strong a relationship is between two variables) between x and y is 0.816 for each dataset

When we plot these four datasets on an x/y coordinate plane, we can observe that they show the same regression lines as well but each dataset is telling a different story:



Dataset I appears to have clean and well-fitting linear models.
- Dataset II is not distributed normally.
- In Dataset III the distribution is linear, but the calculated regression is thrown off by an outlier.
- Dataset IV shows that one outlier is enough to produce a high correlation coefficient.

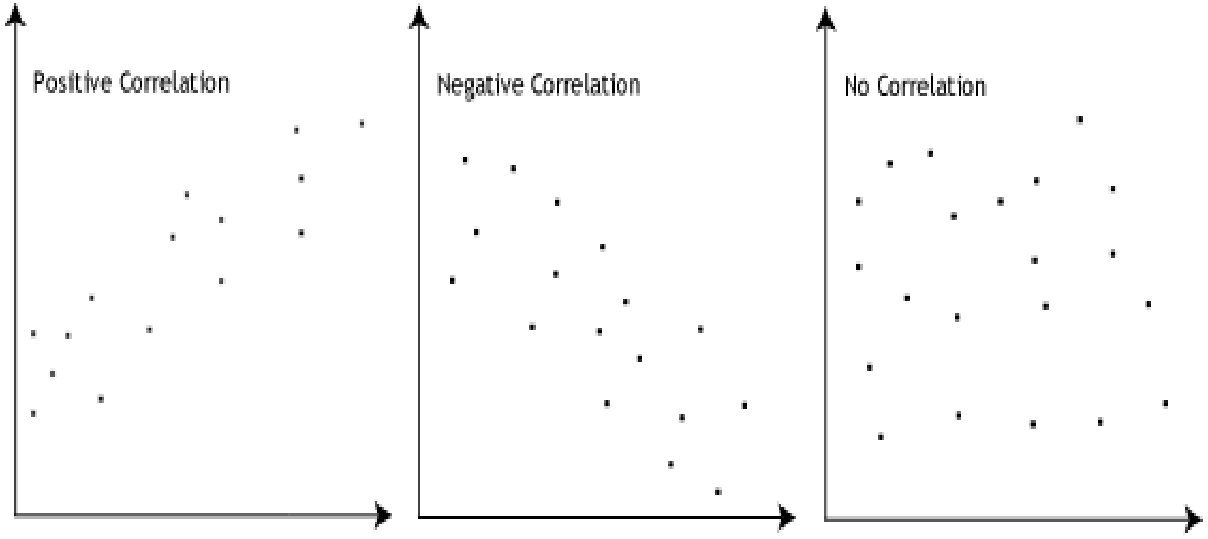
This quartet emphasizes the importance of visualization in Data Analysis. Looking at the data reveals a lot of the structure and a clear picture of the dataset.

General Subjective Questions

3. What is Pearson's R? (3 marks)

Pearson's r is a numerical summary of the strength of the linear association between the variables. 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.

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:



$$r = \frac{\sum x_i y_i - \frac{\sum x_i \sum y_i}{n}}{\sqrt{\left(\sum x_i^2 - \frac{(\sum x_i)^2}{n}\right)}\sqrt{\left(\sum y_i^2 - \frac{(\sum y_i)^2}{n}\right)}}$$

Pearson's R is always between -1 and +1

- The correlation coefficient lies between -1 and +1. *i.e.* $-1 \leq r \leq 1$
- A positive value of '*r*' indicates positive correlation.
- A negative value of '*r*' indicates negative correlation
- If $r = +1$, then the correlation is perfect positive
- If $r = -1$, then the correlation is perfect negative.
- If $r = 0$, then the variables are uncorrelated.

General Subjective Questions

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. 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 feature scaling method then it can consider the value 3000 meter 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 same magnitudes and thus, tackle this issue.

SL No.	Normalized scaling	Standardized scaling
1.	Minimum and maximum value of features are used for scaling	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 bounded 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.

General Subjective Questions

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

If there is perfect correlation, then VIF = infinity. A large value of VIF indicates that there is a correlation between the variables. If the VIF is 4, this means that the variance of the model coefficient is inflated by a factor of 4 due to the presence of multicollinearity.

When the value of VIF is infinite it shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R-squared (R2) =1, which lead to 1/ (1-R2) infinity. To solve this we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

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 is a graphical technique for determining if two data sets come from populations with a common distribution.

Use of Q-Q plot:

A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second dataset. 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. A 45-degree reference line is also plotted. If the two sets come from a population with the same distribution, the points should fall approximately along this reference line. The greater the departure from this reference line, the greater the evidence for the conclusion that the two data sets have come from populations with different distributions.

Importance of Q-Q plot:

When there are two data samples, it is often desirable to know if the assumption of a common distribution is justified. If so, then location and scale estimators can pool both data sets to obtain estimates of the common location and scale. If two samples do differ, it is also useful to gain some understanding of the differences. The q-q plot can provide more insight into the nature of the difference than analytical methods such as the chi-square and Kolmogorov-Smirnov 2-sample tests.