NAME- NISHU KUMARI
FROM- IIIT B (UpGrad)

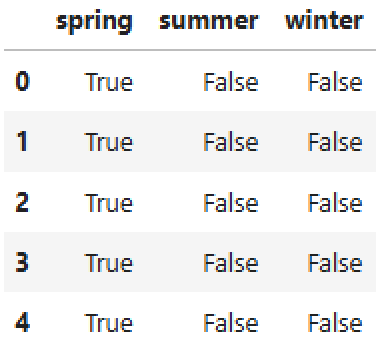
Assignment-based Subjective Questions

Question 1*. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?   (Do not edit)*

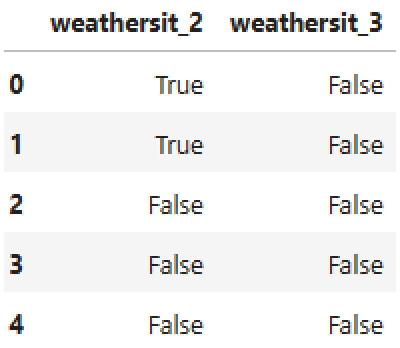
Total Marks*: 3 marks (Do not edit)*

Answer: *<Your answer for Question 1 goes below this line> (Do not edit)*

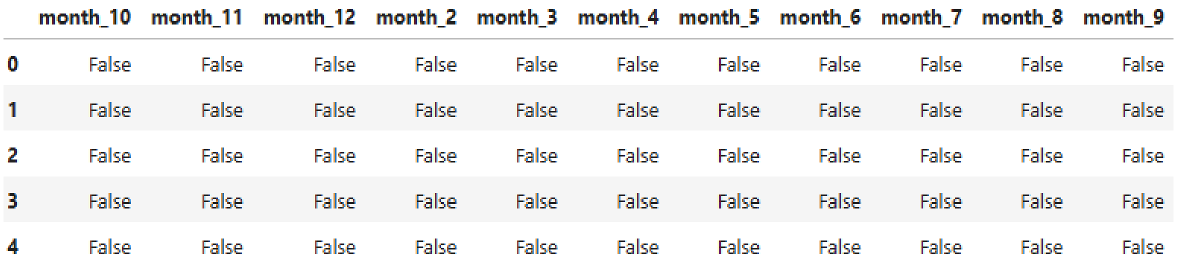
The final Multiple Linear Regression model contains many predictor variables that are categorical in nature and some of them have been encoded to dummy variables.



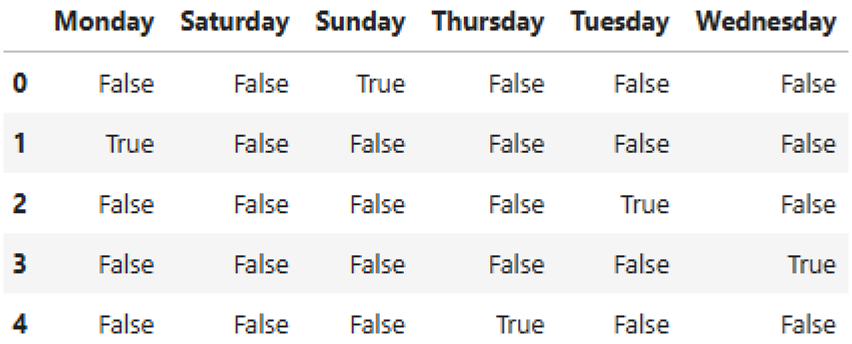
	spring	summer	winter
0	True	False	False
1	True	False	False
2	True	False	False
3	True	False	False
4	True	False	False



	weathersit_2	weathersit_3
0	True	False
1	True	False
2	False	False
3	False	False
4	False	False



	month_10	month_11	month_12	month_2	month_3	month_4	month_5	month_6	month_7	month_8	month_9
0	False	False	False	False	False	False	False	False	False	False	False
1	False	False	False	False	False	False	False	False	False	False	False
2	False	False	False	False	False	False	False	False	False	False	False
3	False	False	False	False	False	False	False	False	False	False	False
4	False	False	False	False	False	False	False	False	False	False	False



	Monday	Saturday	Sunday	Thursday	Tuesday	Wednesday
0	False	False	True	False	False	False
1	True	False	False	False	False	False
2	False	False	False	False	True	False
3	False	False	False	False	False	True
4	False	False	False	True	False	False

Spring, winter falls under season category *and have been dummy encoded****. weathersit_2 and weathersit_3 fall*** *under weathersit category and have been dummy encoded.*

Similarly, ***month variables fall under month category and have been dummy encoded****.*

We can infer from above image that these variables are statistically significant and explain the variance in model very well likewise each separate.

Question 2. *Why is it important to use* ***drop_first=True*** *during dummy variable creation? (Do not edit)*

Total Marks: *2 marks (Do not edit)*

Answer: *<Your answer for Question 2 goes below this line> (Do not edit)*

Reason 1-

To avoid dummy variable trap. When we create dummy variables for categorical data, we must avoid multicollinearity, which happens when dummy variables are highly correlated. This can violate the assumptions of linear regression. To prevent this, if we have k categories (where k is 3 or more), we use only k-1 dummy variables. The missing category is represented by the intercept as a base case.

Reason 2-

Handling Nominal Data Properly. Some categorical variables, like colors (Red, Green, Blue), have no natural order. If we simply assign numbers like 1, 2, and 3, the model might mistakenly assume an order-based relationship (Red < Green < Blue), which could lead to bias. Dummy encoding helps by representing each category as a separate binary variable, ensuring the model treats them correctly. Additionally, since machine learning models work with numerical data, converting text categories into numbers is necessary for proper processing.

Question 3. *Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (Do not edit)*

Total Marks: *1 mark (Do not edit)*

Answer: *<Your answer for Question 3 goes below this line> (Do not edit)*

Before building and training the model, the pair plot shows that the ***registered*** *variable has the highest correlation (0.945). However, we are not including* ***casual*** *and* ***registered*** *in our pre-processed training data because their sum equals* ***cnt*** *(casual + registered = cnt). Including them could leak important information into the model, leading to overfitting.*

Since these two variables are excluded, ***atemp*** *has the highest correlation with the target variable* ***cnt****, followed by* ***temp****. According to the correlation heatmap, the correlation coefficient between* ***atemp*** *and* ***cnt*** *is* ***0.631****, while the correlation between* ***temp*** *and* ***cnt*** *is* ***0.627****.*

Question 4. *How did you validate the assumptions of Linear Regression after building the model on the training set? (Do not edit)*

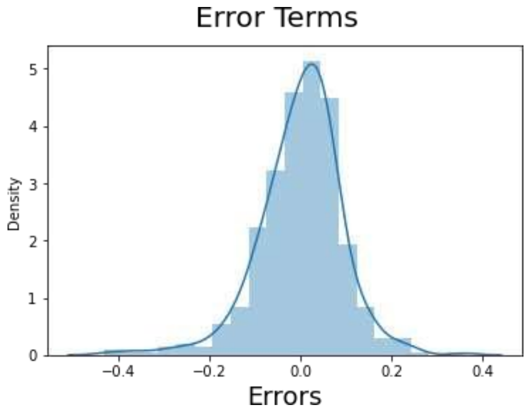
Total Marks: *3 marks (Do not edit)*

Answer: *<Your answer for Question 4 goes below this line> (Do not edit)*

To validate assumptions of the model, and hence the reliability for inference, we go with the following procedures:

- *Residual Analysis:*

We need to check if the error terms are also normally distributed (which is in fact, one of the major assumptions of linear regression). I have plotted the histogram of the error terms and this is what it looks like:



The residuals are normally distribution with a mean 0.

Linear relationship between predictor variables and target variable: *This is happening because all the predictor variables are statistically significant (p-values are less than **0.05**). Also, R-Squared value on training set is **0.832** and adjusted R-Squared value on training set is **0.828**. This means that variance in data is being explained by all these predictor variables.*

Error terms are independent of each other: *Handled properly in the model. The predictor variables are independent of each other. Multicollinearity issue is not there because VIF (Variance Inflation Factor) for all predictor variables are below 5.*

Question 5. *Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (Do not edit)*
Total Marks: *2 marks (Do not edit)*
Answer: *<Your answer for Question 5 goes below this line> (Do not edit)*

The top three features that significantly impact the demand for shared bikes are:

1. *Temperature (temp) – with a coefficient of 0.4495, indicating that temperature has the strongest influence on bike demand.*
2. *Year (yr) – with a coefficient of 0.2341, suggesting that demand has increased over time.*
3. *Month_9 (September) – with a coefficient of 0.0800, meaning that bike usage tends to be higher in September compared to other months.*

General Subjective Questions

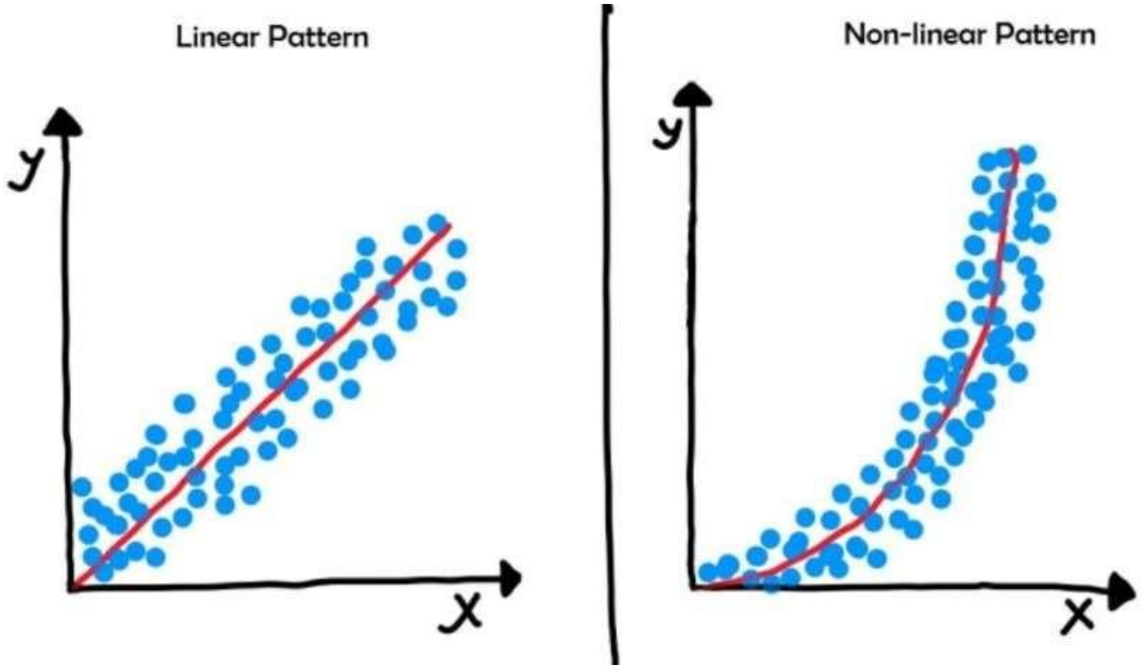
Question 6. *Explain the linear regression algorithm in detail. (Do not edit)*
Total Marks: *4 marks (Do not edit)*
Answer: *Please write your answer below this line. (Do not edit)*

<Your answer for Question 6 goes here>

Linear Regression finds the best linear relationship between the independent and dependent variables. It is a method of finding the best straight-line fitting to the given data. In technical terms, linear regression is a machine learning algorithm that finds the best linear-fit relationship on any given data, between independent and dependent variables. It is mostly done by the Sum of SquaredResiduals Method.

The assumptions of linear regression are:

a. The assumption about the form of the model: It is assumed that there is a linear relationship between the dependent and independent variables.



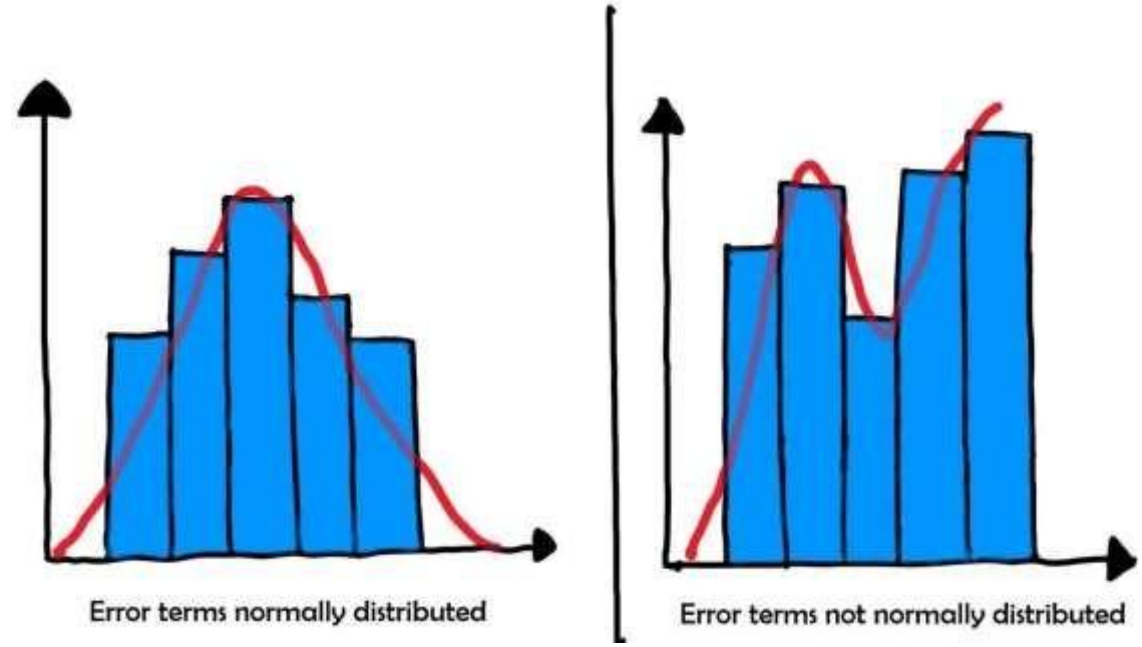
b. Assumptions about the residuals:

1) Normality assumption: It is assumed that the error terms, ε(i), are normally distributed.

2) Zero mean assumption: It is assumed that the residuals have a mean value of zero, i.e., the error terms are normally distributed around zero.

3) Constant variance assumption: It is assumed that the residual terms have the same (but unknown) variance, sigma square. This assumption is also known as the assumption of homogeneity or homoscedasticity.

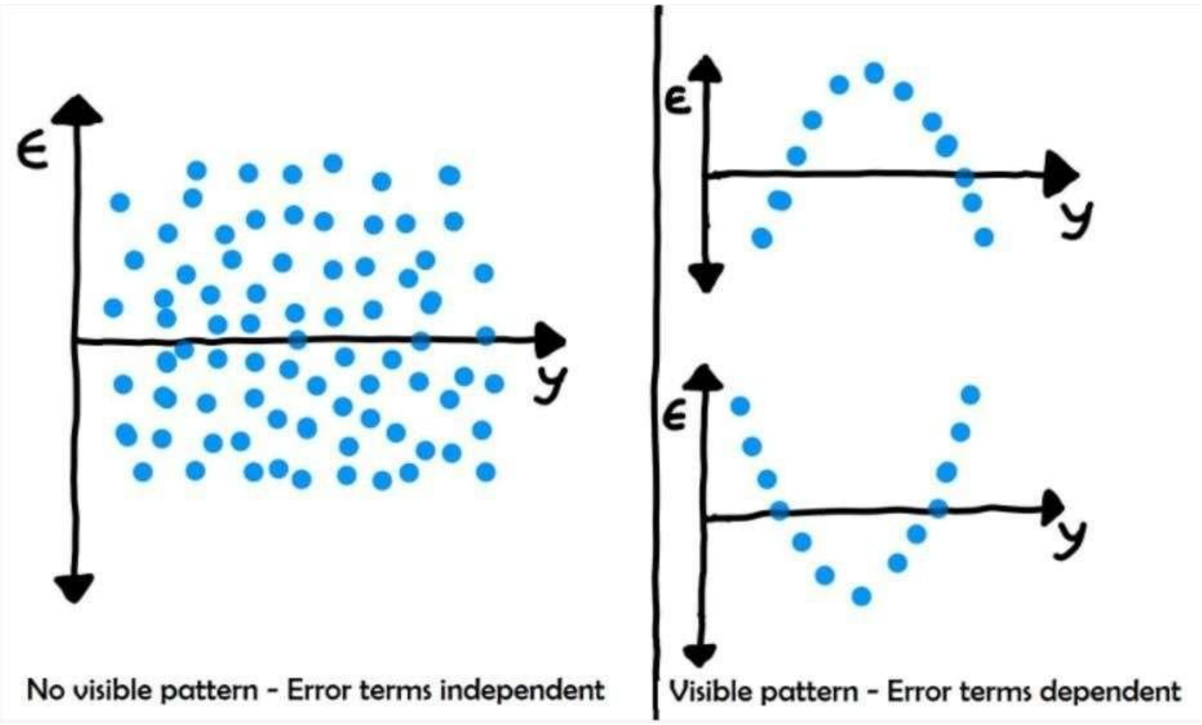
4) Independent error assumption: It is assumed that the residual terms are independent of each other, i.e., their pair-wise covariance is zero.

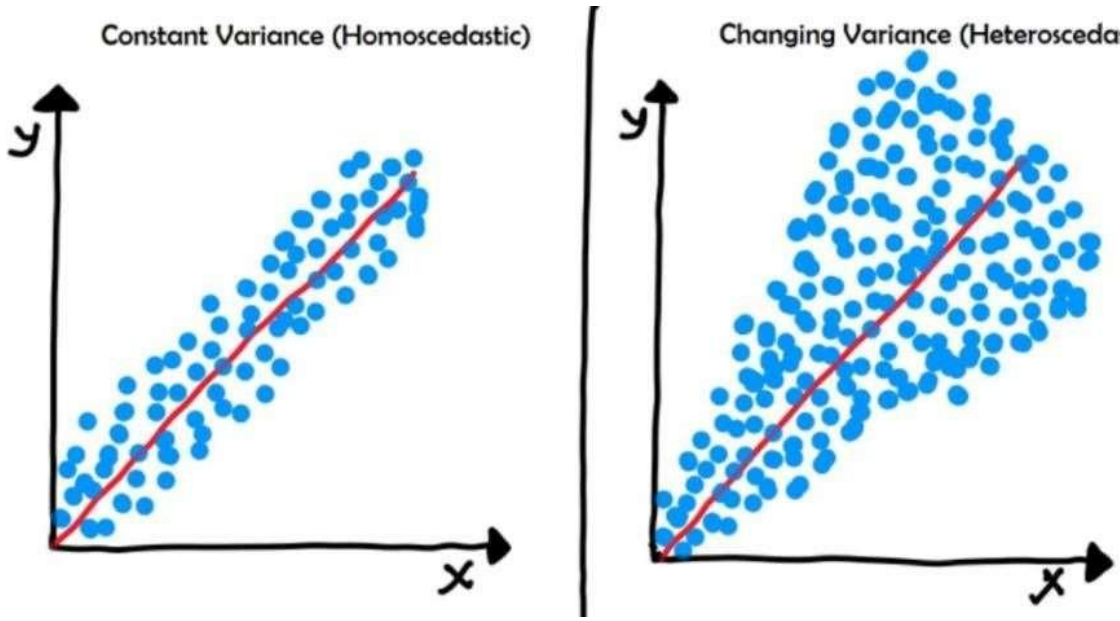


c. Assumptions about the estimators:

1) The independent variables are measured without error.

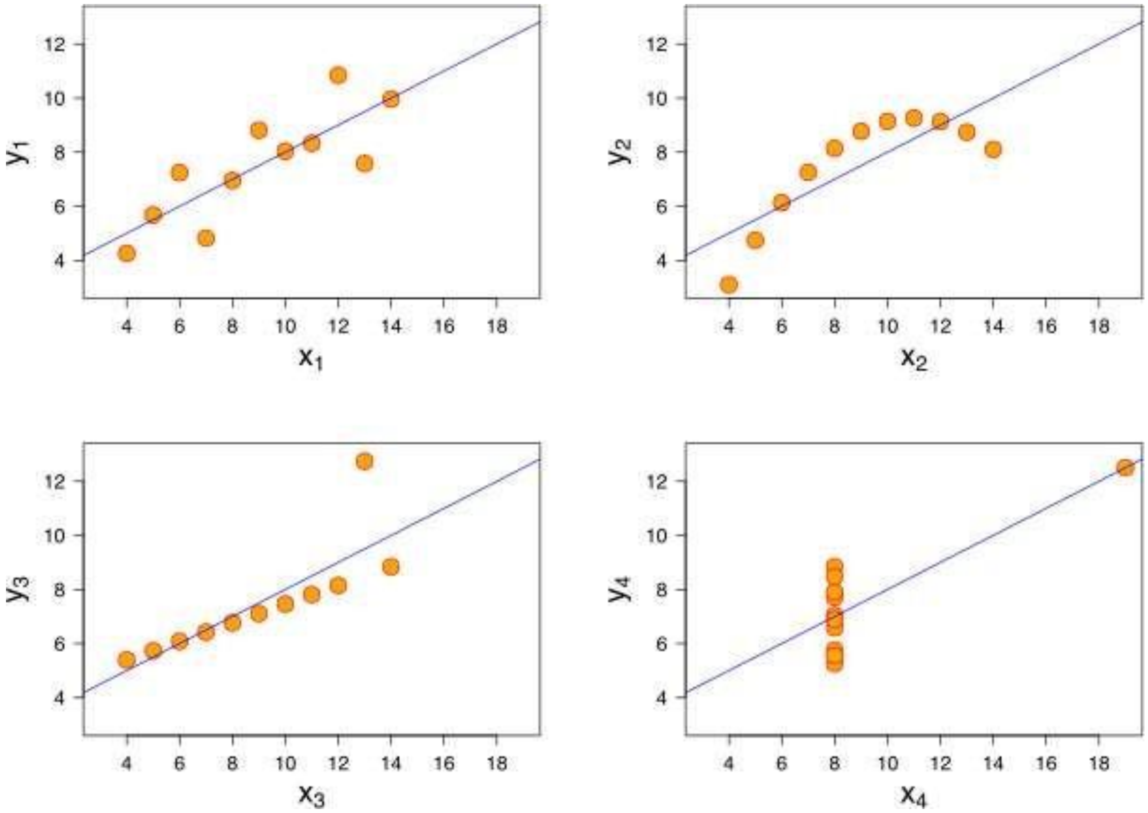
2) The independent variables are linearly independent of each other, i.e., there is no multicollinearity in the data.





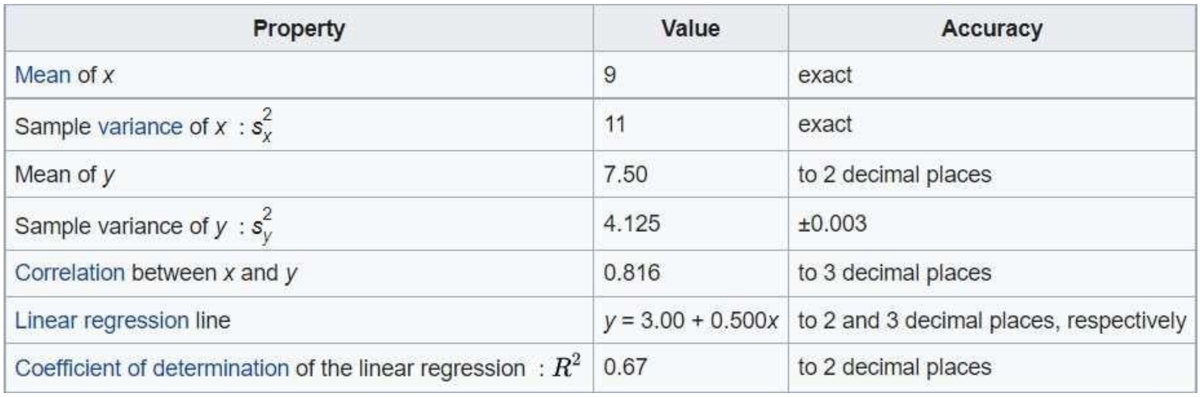
Question 7. *Explain the Anscombe's quartet in detail. (Do not edit)*
Total Marks: *3 marks (Do not edit)*
Answer: *Please write your answer below this line. (Do not edit)*

Anscombe's quartet comprises four data sets that have nearly identical simple descriptive statistics, yet have very different distributions and appear very different when graphed.



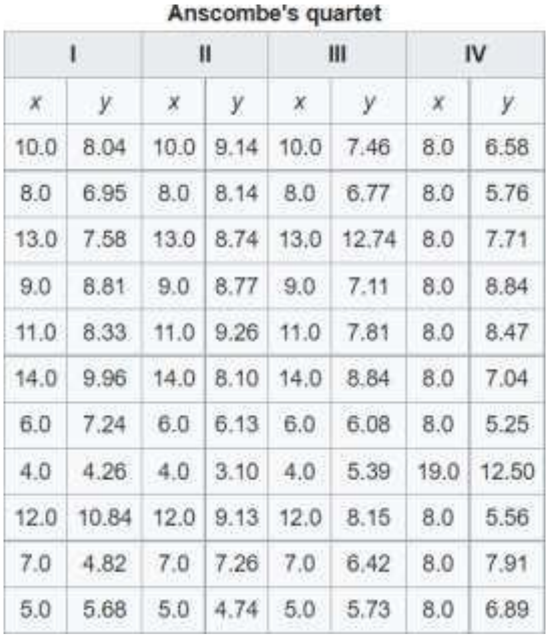
All four sets are identical when examined using simple summary statistics, but vary considerably when graphed.

1) 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.
2) 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.
3) 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.
4) 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.



Property	Value	Accuracy
Mean of x	9	exact
Sample variance of x : s_x^2	11	exact
Mean of y	7.50	to 2 decimal places
Sample variance of y : s_y^2	4.125	±0.003
Correlation between x and y	0.816	to 3 decimal places
Linear regression line	$y = 3.00 + 0.500x$	to 2 and 3 decimal places, respectively
Coefficient of determination of the linear regression : R^2	0.67	to 2 decimal places

Anscombe's quartet

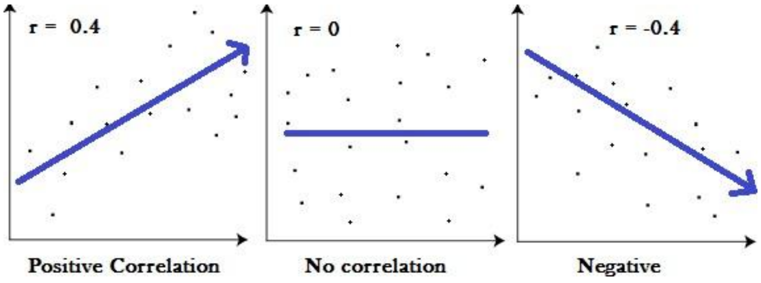


I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	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

Question 8. *What is Pearson's R? (Do not edit)*
Total Marks: *3 marks (Do not edit)*
Answer: *Please write your answer below this line. (Do not edit)*

<Your answer for Question 8 goes here>

Pearson's R or correlation coefficient is a measure of linear correlation between two sets of data. It is the ratio between the covariance of two variables and the product of their standard deviations; thus, it is essentially a normalised measurement of the covariance, such that the result always has a *value between −1 and 1. As with covariance itself, the measure can only reflect a linear correlation of variables, and ignores many other types of relationship or correlation.*



1) A correlation coefficient of 1 means that for every positive increase in one variable, there is a positive increase of a fixed proportion in the other. For example, shoe sizes go up in (almost) perfect correlation with foot length.

2) A correlationcoefficient of -1 means that for every positive increase in one variable, there is a negative decrease of a fixed proportion in the other. For example, the amount of gas in a tank decrease in (almost) perfect correlation with speed.

3) Zero means that for every increase, there isn't a positive or negative increase. The two just aren't related.

The absolute value of the correlation coefficient gives us the relationship strength. The larger the number, the stronger the relationship. For example, |-.95| = .95, which has a stronger relationship than .55.

***Question 9.** What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (Do not edit)*

***Total Marks:** 3 marks (Do not edit)*

***Answer:** Please write your answer below this line. (Do not edit)*

Scaling is a method used to normalize the range of independent variablesor features of data. Since the range of values of raw data varies widely, in some machine learning algorithms, objective functions will not work properly without normalization. For example, many classifiers calculate the distance between two points by the Euclidean distance. If one of the features has a broad range of values, the distance will be governed by this particular feature. Therefore, the range of all features should be normalized so that each feature contributes approximately proportionately to the final distance.

Another reason why feature scaling is applied is that gradient descent converges much faster with feature scaling than without it.

Normalization:

Also known as min-max scaling or min-max normalization, it is the simplest method and consists of rescaling the range of features to scale the range in [0, 1]. The general formula for normalization is given as:

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

Here, max(x) and min(x) are the maximum and the minimum values of the feature respectively.

Standardization:

Feature standardization makes the values of each feature in the data have zero mean and unit variance. The general method of calculation is to determine the distribution mean and standard deviation for each feature and calculate the new data point by the following formula:

$$x' = \frac{x - \bar{x}}{\sigma}$$

Here, σ is the standard deviation of the feature vector, and $\bar{x}$ is the average of the feature vector.

Question 10. *You might have observed that sometimes the value of VIF is infinite. Why does this happen? (Do not edit)*

Total Marks: *3 marks (Do not edit)*

Answer: *Please write your answer below this line. (Do not edit)*

If there is perfect correlation, then VIF = infinity. This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity. To solve this problem, we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

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).

Question 11. *What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (Do not edit)*

Total Marks: *3 marks (Do not edit)*

Answer: *Please write your answer below this line. (Do not edit)*

<Your answer for Question 11 goes here>

Q-Q Plots (Quantile-Quantile plots) are plots of two quantiles against each other. A quantile is a fraction where certain values fall below that quantile. For example, the median is a quantile where 50% of the data fall below that point and 50% lie above it. The purpose of Q-Q plots is to find out if two sets of data come from the same distribution. A 45-degree angle is plotted on the Q-Q plot; if the two data sets come from a common distribution, the points will fall on that reference line.

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.

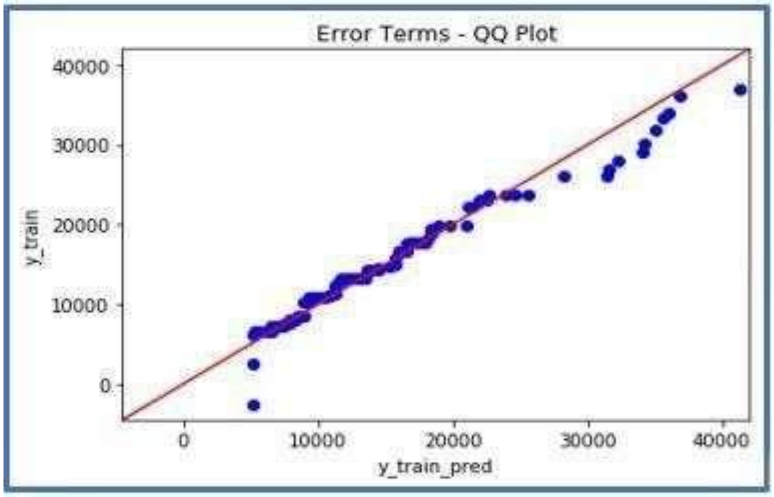
Interpretation:

A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set.

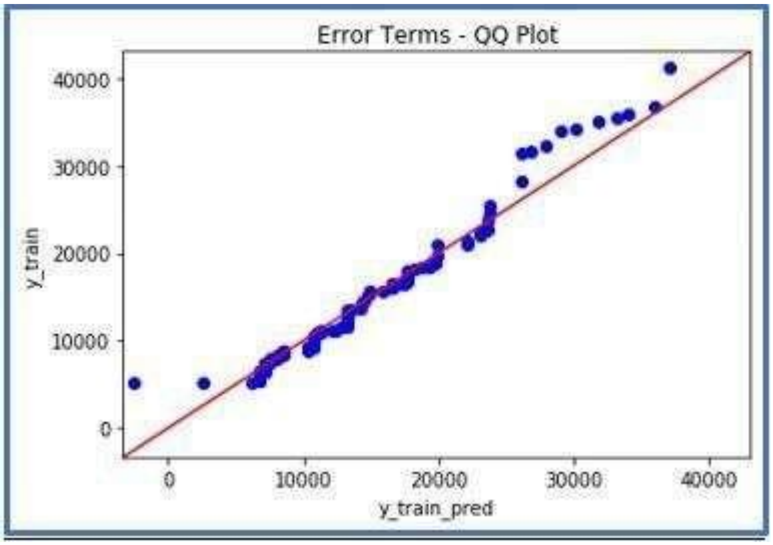
Below are the possible interpretations for two data sets.

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

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



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



4) *Different distribution: If all point of quantiles lies away from the straight line at an angle of 45 degree from x -axis*

statsmodels.api *provide* ***qqplot*** *and* ***qqplot_2samples*** *to plot* ***Q-Q*** *graph for single and two different data sets respectively.*