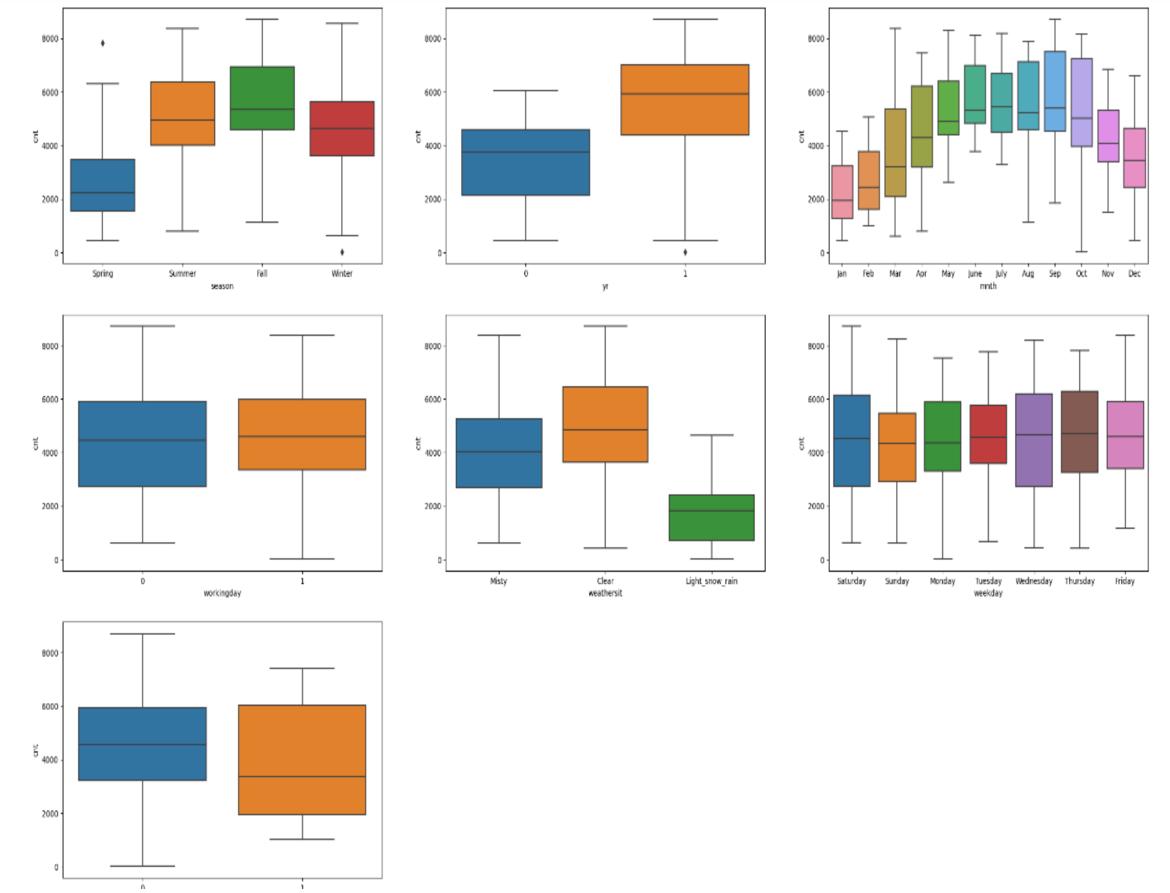
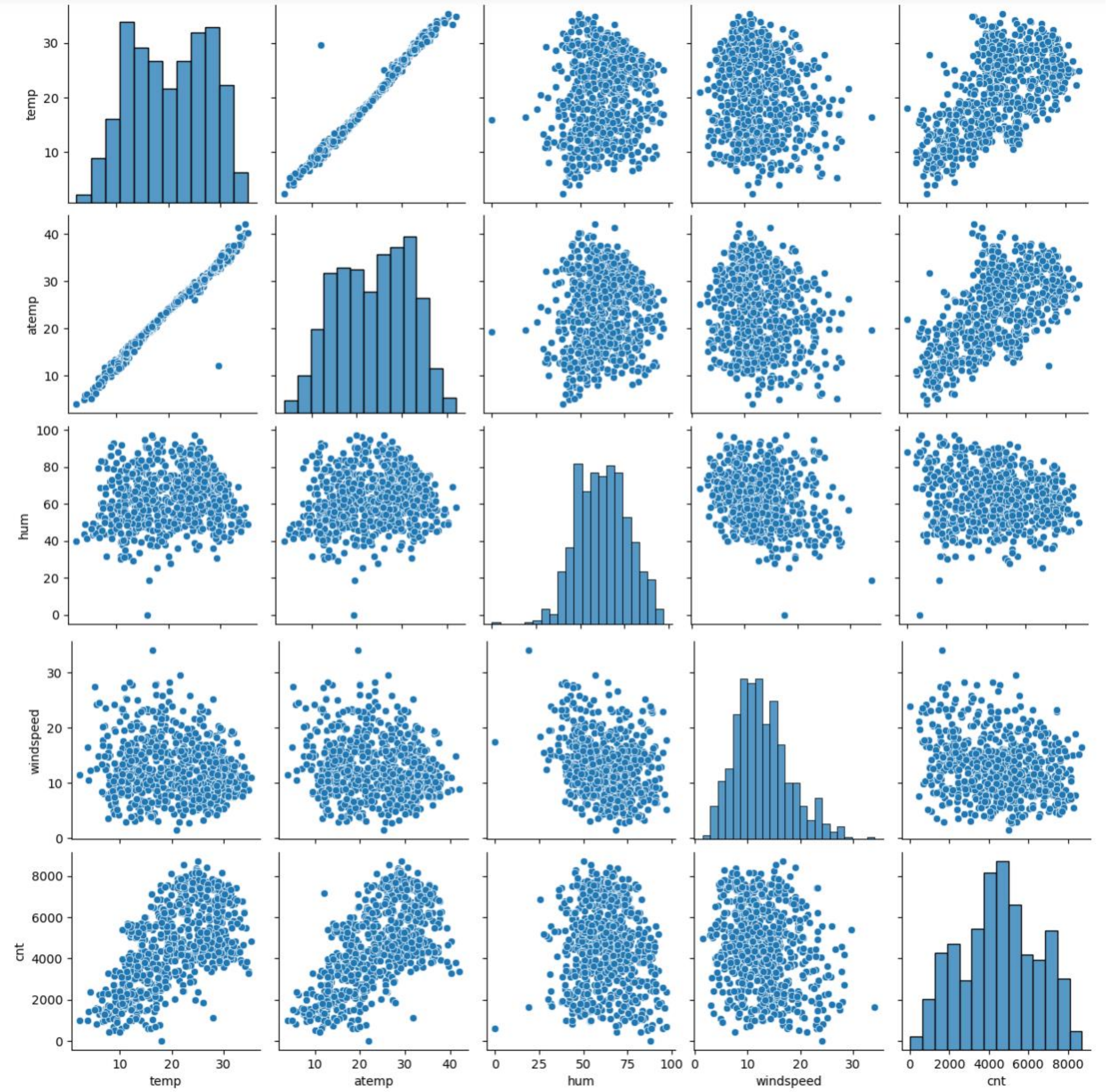
Assignment-based Subjective Questions

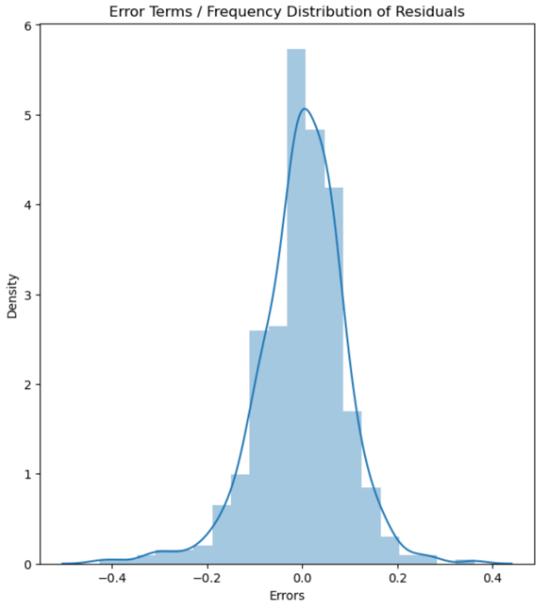
1.	From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)
Answer:	We have 7 categorical variables ['season', 'yr', 'mnth', 'weathersit', 'holiday', 'weekday', 'workingday'], to infer their effect on the dependent variable we have boxplots:  The following conclusions can be drawn from the data: a) The dependent variable 'cnt' is highly correlated/dependant on categorical variable such as season, weather, workingday etc. b) season: Season 3/Fall had the highest proportion of bike bookings, with 32% of the total and a median of over 5000 bookings for the two-year period. Season 2/Summer and season 4/Winter followed with 27% and 25% respectively. This suggests that season is a

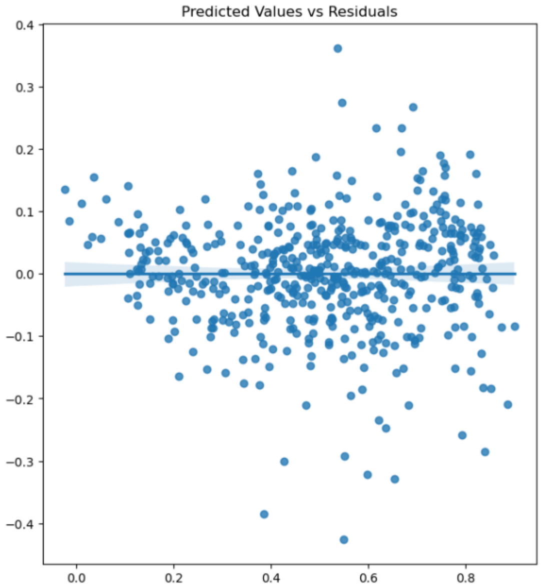
	good predictor for the dependent variable. c) mnth: The months May, June, July, Aug, and Sept had the most bike bookings, with 10% each and a median of over4000bookings per month. This suggests that mnth has some trend for bookings and can be a good predictor for the dependent variable. d) weathersit: Weathersit 1 had the most bike bookings, with 67% of the total and a median of close to 5000 bookings for the two-year period. Weathersit 2 followed with 30% of the total. This suggests that weathersit has some influence on the bike bookings and can be a good predictor for the dependent variable. e) holiday: Only 2.4% of the bike bookings occurred on holidays, which means this data is clearly biased. This suggests that holiday cannot be a good predictor for the dependent variable. f) weekday: Weekday variable showed a similar trend (between13.5%-14.8%of the total bookings on all days of the week) with their independent medians between 4000 to 5000 bookings. This variable may have some or no effect on the predictor. I will let the model decide if this needs to be added or not. g) workingday: Workingday had 69% of the bike bookings, with a median of close to 5000 bookings for the two-year period. This suggests that workingday can be a good predictor for the dependent variable. h) yr: Year also had a strong correlation with cnt, which can be seen from the boxplot diagram.
2.	Why is it important to use **drop_first=True** during dummy variable creation? (2 mark)
Answer:	Using drop_first=True during dummy variable creation is important because it helps to avoid the dummy variable trap, which is a situation where one or more of the dummy variables are redundant and can be predicted by the others. This can cause multicollinearity, which is a problem for some regression models that assume the predictor variables are independent of each other. By dropping the first column, we reduce the number of dummy variables by one and ensure that they are not perfectly correlated. This way, we avoid the dummy variable trap and multicollinearity. Ex: Let us say we have 3 types of values in Categorical column and we want to create dummy variable for that column. If one variable is not furnished and semi furnished, then It is obvious unfurnished. So, we do not need 3rd variable to identify the unfurnished.

3.	Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)
	Here is the pair-lot among numerical variables:  The pairplot shows a linear relationship between temp, atemp, and the target variable 'cnt'. This means that as the temperature (either actual or felt) increases, so does the number of bike bookings.
4.	How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

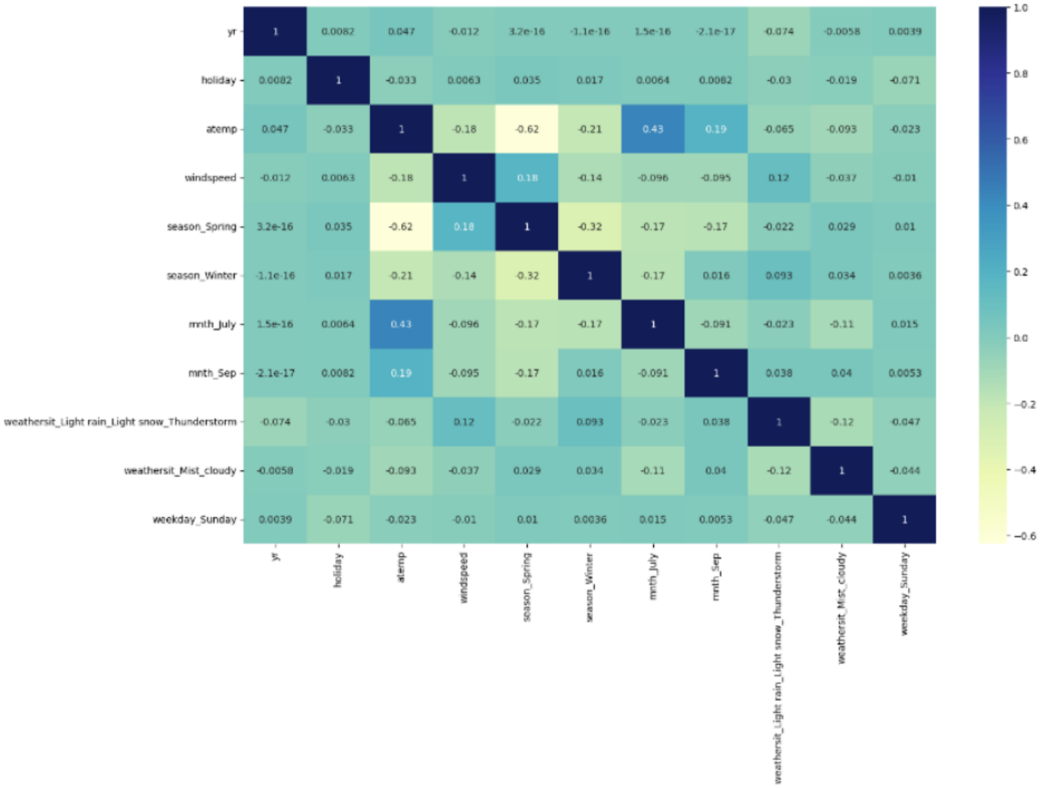
Answer:

a) The errors have a normal distribution:





b) The predictor variables are not correlated with each other. (No Multicollinearity).



	yr	holiday	atemp	windspeed	season_Spring	season_Winter	mnth_July	mnth_Sep	weathersit_Light rain_Light snow_Thunderstorm	weathersit_Mist_cloudy	weekday_Sunday
yr	1	0.0082	0.047	-0.012	3.2e-16	-1.1e-16	1.5e-16	-2.1e-17	-0.074	-0.0058	0.0039
holiday	0.0082	1	-0.033	0.0063	0.035	0.017	0.0064	0.0082	-0.03	-0.019	-0.071
atemp	0.047	-0.033	1	-0.18	-0.62	-0.21	0.43	0.19	-0.065	-0.093	-0.023
windspeed	-0.012	0.0063	-0.18	1	0.18	-0.14	-0.096	-0.095	0.12	-0.037	-0.01
season_Spring	3.2e-16	0.035	-0.62	0.18	1	-0.32	-0.17	-0.17	-0.022	0.029	0.01
season_Winter	-1.1e-16	0.017	-0.21	-0.14	-0.32	1	-0.17	0.016	0.093	0.034	0.0036
mnth_July	1.5e-16	0.0064	0.43	-0.096	-0.17	-0.17	1	-0.091	-0.023	-0.11	0.015
mnth_Sep	-2.1e-17	0.0082	0.19	-0.095	-0.17	0.016	-0.091	1	0.038	0.04	0.0053
weathersit_Light rain_Light snow_Thunderstorm	-0.074	-0.03	-0.065	0.12	-0.022	0.093	-0.023	0.038	1	-0.12	-0.047
weathersit_Mist_cloudy	-0.0058	-0.019	-0.093	-0.037	0.029	0.034	-0.11	0.04	-0.12	1	-0.044
weekday_Sunday	0.0039	-0.071	-0.023	-0.01	0.01	0.0036	0.015	0.0053	-0.047	-0.044	1

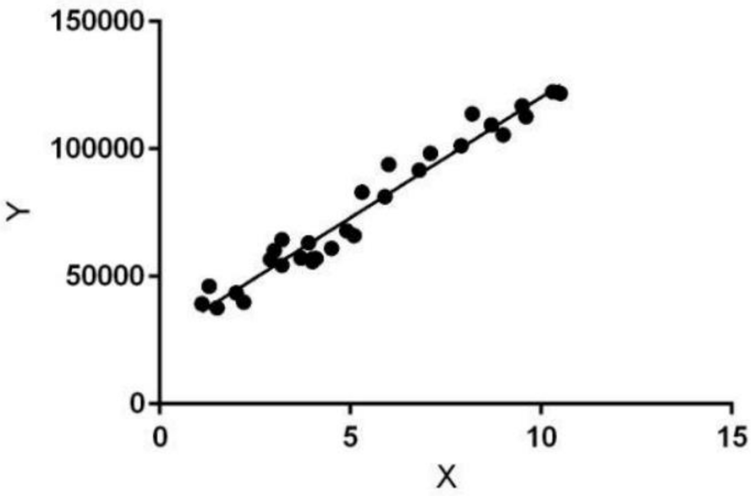
c) There is a linear relationship between temp, atemp and cnt or in other words the number of bike bookings increases with the temperature (both actual and felt). (Pairplot is shown in Question no. 3)

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:	The features that have the highest positive correlation with the target variable are atemp/temp, yr, season_winter and mnth_Sep, with correlation values of 0.4632, 0.2350, 0.0412 and 0.0587 respectively. The following factors have a significant impact on the bike bookings: (considering positive coef only): • temp: The users prefer to ride bikes when the temperature is moderate and comfortable. • yr: Demand increases(as coefficient is positive) in case of yr • season: The company should target seasons, when the demand is higher based on positive coef's. • weather: The users prefer to ride bikes when the weather is pleasant and clear. **We can see that the equation for best fitted line is:** **cnt = 0.2620 + 0.2350 X yr - 0.1028 X holiday + 0.4632 X atemp - 0.1254 X windspeed - 0.1167 X season_Spring + 0.0412 X season_Winter - 0.0657 X mnth_July + 0.0587 X mnth_Sep - 0.2872 X weathersit_Light rain_Light snow_Thunderstorm - 0.0837 X weathersit_Mist_cloudy -0.0484 X weekday_Sunday**

General Subjective Questions

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

Answer: Linear regression is a method of finding the best straight line fitting to the given data, finding the best linear relationship between independent and dependent variables. 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 Sum of Squared Residuals Method.



The line in the above graph is referred as the best fit straight line. Based on the given data points

we try to plot a line that models the points the best.

Mathematically the relationship can be represented with the help of following equation –

Y = mX + b

Here, Y is the dependent variable we are trying to predict

X is the dependent variable we are using to make predictions.

m is the slop of the regression line which represents the effect X has on Y

b is a constant, known as the Y-intercept. If X = 0, Y would be equal to b

Assumption for Linear Regression Model:

- Linear regression is a powerful tool for understanding and predicting the behaviour of a variable, however, it needs to meet a few conditions to be accurate and dependable solutions.
- Linearity: The independent and dependent variables have a linear relationship with one another. This implies that changes in the dependent variable follow those in the

independent variable(s) in a linear fashion.

• Independence: The observations in the dataset are independent of each other. This means that the value of the dependent variable for one observation does not depend on the value of the dependent variable for another observation.

• Homoscedasticity: Across all levels of the independent variable(s), the variance of the errors is constant. This indicates that the amount of the independent variable(s) has no impact on the variance of the errors.

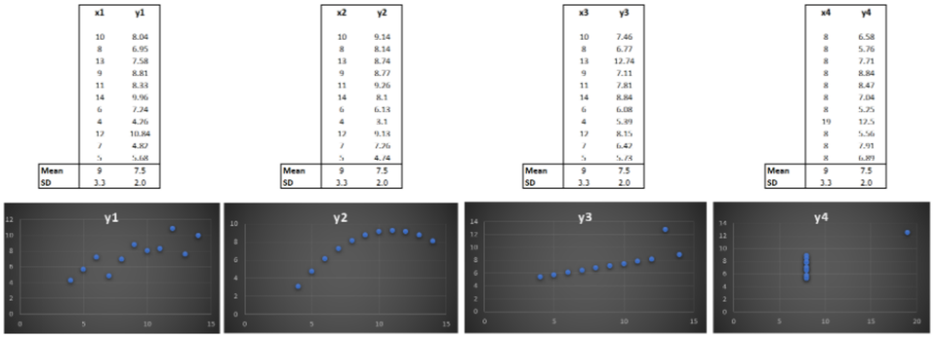
• Normality: The errors in the model are normally distributed.

• No multicollinearity: There is no high correlation between the independent variables. This indicates that there is little or no correlation between the independent variables.

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

Answer: Anscombe's quartet is a famous statistical example that highlights the importance of visualizing data before drawing conclusions. It consists of four datasets, each containing 11 data points, and when analysed using common summary statistics, they appear to be very similar. However, when you plot the data, you will see that they have drastically different characteristics. This quartet was created by the statistician Francis Anscombe in 1973 to illustrate the limitations of relying solely on summary statistics and the value of data visualization.

The datasets and their graphical representation are shown in the following Excel snapshot:



	x1	y1
	10	8.04
	8	6.95
	13	7.58
	9	8.81
	11	8.33
	14	9.96
	6	7.24
	4	4.26
	12	10.84
	7	4.82
	5	5.68
Mean	9	7.5
SD	3.3	2.0

	x2	y2
	10	9.14
	8	8.14
	13	8.74
	9	8.77
	11	9.26
	14	8.1
	6	6.13
	4	3.1
	12	9.13
	7	7.26
	5	4.74
Mean	9	7.5
SD	3.3	2.0

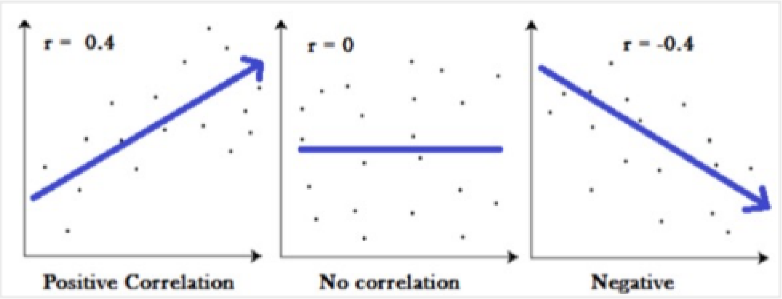
	x3	y3
	10	7.46
	8	6.77
	13	12.74
	9	7.11
	11	7.81
	14	8.84
	6	6.08
	4	5.39
	12	8.15
	7	6.42
	5	5.73
Mean	9	7.5
SD	3.3	2.0

	x4	y4
	8	6.58
	8	5.76
	8	7.71
	8	8.84
	8	8.47
	8	7.04
	8	5.25
	19	12.5
	8	5.56
	8	7.91
	8	6.89
Mean	9	7.5
SD	3.3	2.0

Despite the variations in each dataset, they have the same summary statistics such as same mean, same standard deviations (SD), correlational coefficient, and linear regression line.

The first dataset appears to be a simple linear relationship, where y increases as x increases. The second dataset, shows a linear trend, a single outlier affects the regression line, creating a misleading representation of the data.

	Now, the third dataset takes an unexpected turn. It follows a perfectly quadratic relationship, with a clear curve. This highlights the fact that data can exhibit nonlinear patterns, and relying solely on linear regression can lead to incorrect conclusions. Finally, the fourth dataset adds a new layer of complexity to the situation. There is one data point that stands out from the others and entirely contradicts the pattern, which causes the linear regression line to shift in a significant way. Anscombe's Quartet shows us that we should not blindly trust summary statistics or standard methods of analysis. It tells us to look closely at our data, question our assumptions, and use a variety of analytical tools to get a full picture. This concept emphasizes the importance of visualizing data, as graphs can reveal patterns and outliers that summary statistics alone may overlook.
3.	What is Pearson's R? (3 marks)
Answer:	Pearson's r is a statistic that measures the linear correlation between two variables. Like all correlations, it also has a numerical value that lies between -1.0 and +1.0. Whenever we discuss correlation in statistics, it is generally Pearson's correlation coefficient. However, it cannot capture nonlinear relationships between two variables and cannot differentiate between dependent and independent variables. Pearson correlation coefficient (r): $r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[n\Sigma x^2 - (\Sigma x)^2][n\Sigma y^2 - (\Sigma y)^2]}}$

	Diagrams Depicting correlations:  In summary: ● 1 indicates a strong positive relationship. ● -1 indicates a strong negative relationship. ● A result of zero indicates no relationship at all.
4.	What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)
Answer:	**What is scaling -** Scaling is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. **Why is scaling performed -** Scaling helps in speeding up the calculations in an algorithm. In general, collected data set contains features varying in magnitudes, units and range. If scaling is not done, then algorithm only takes magnitude in account and not unit. This results in incorrect modelling. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude. 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** **Normalization Scaling -** It brings all of the data in the range of 0 and 1. It is also called as MinMax Scaling MinMax Scaling: $x = \frac{x - min(x)}{max(x) - min(x)}$

	sklearn.preprocessing.MinMaxScaler helps to implement normalization in python. Standardised Scaling - Standardization replaces the values by their Z scores. It brings all the data into a standard normal distribution which has mean (μ) zero and standard deviation one (σ) 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)
Answer:	The value of VIF is infinite when there is perfect correlation between a given independent variable and other variables in the model. This means that the given variable can be perfectly predicted by a linear combination of the other variables. In other words, the given variable is redundant and does not add any new information to the model. This situation can cause problems for some regression methods that rely on the inverse of the covariance matrix, which becomes singular when there is perfect correlation. To avoid infinite VIF values, one should check for multicollinearity among the independent variables and remove any variables that are highly correlated with others. Alternatively, one can use regularization methods such as ridge or lasso regression that can handle multicollinearity by shrinking the coefficients of correlated variables.
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, or quantile-quantile plot, is a graphical tool that helps us assess if a set of data plausibly came from some theoretical distribution, such as a normal, exponential, or uniform distribution. It can also help us compare if two data sets come from populations with a common distribution. A Q-Q plot is created by plotting the quantiles of the first data set against the quantiles of the second data set. If the points lie approximately on a straight line, it means that the two data sets have similar distributions. The slope and intercept of the line indicate the relative location and scale of the two data sets. In linear regression, a Q-Q plot is often used to check the normality assumption of the error terms or residuals. By plotting the standardized residuals against the theoretical quantiles of a standard normal distribution, we can see if the residuals are normally distributed. If the residuals are not normally distributed, it implies that the standard confidence intervals and significance tests for the regression coefficients may

	be invalid. A Q-Q plot is important in linear regression because it can help us diagnose potential problems with our model, such as outliers, skewness, heteroscedasticity, or non-linearity. It can also help us decide if we need to transform our data or use a different regression method to improve our model fit.