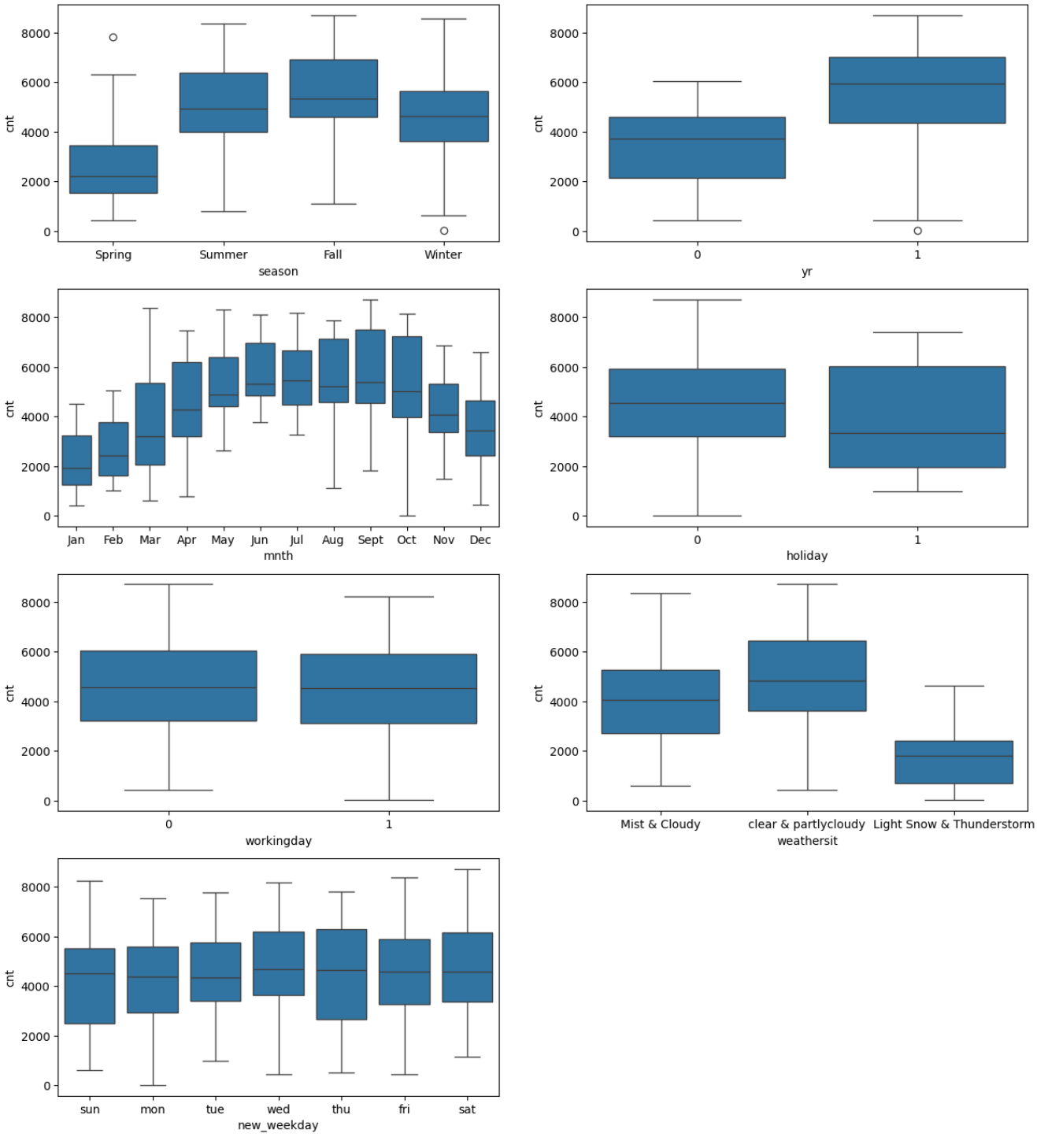
Assignment-based Subjective Questions

By Krishna Sai Sangaraju

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

Ans: Below figure shows the boxplot of categorical variables (season, yr, mnth, holiday, workingday, weathersit, new_weekday) with target variable (cnt).



Inferences:

- User activity is highest during the summer and fall months, as evidenced by the higher median values.
- There was a significant increase in the number of users in 2019 compared to previous years.
- User activity gradually increases from January to July, followed by a decrease in the later months.
- More bikes are used when there is no holiday
- For working day users used either on working day or non-working day are almost equal.
- User engagement is higher when the weather is clear or partly cloudy (around 5000 users) compared to misty or cloudy conditions (around 4000 users).
- There is not much significant change in median between weekday, so it does not have much impact.

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

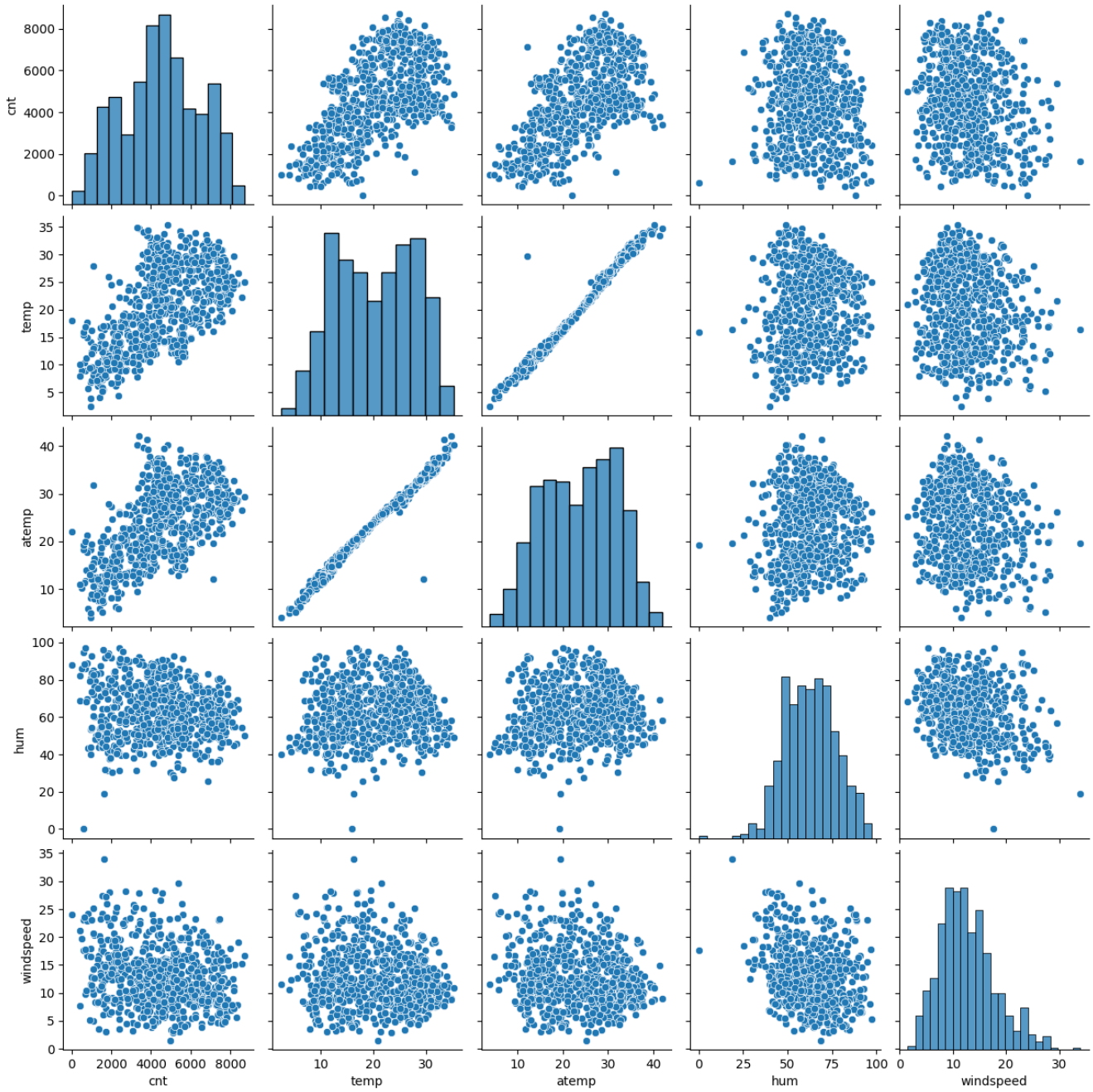
Ans**:** In order to prevent falling into the dummy variable trap and lessen multicollinearity in the dataset, it is crucial to utilise drop_first=True while creating dummy variables. Making dummy variables will result in n columns if your category variable has n distinct levels. But to represent all the data from the original categorical variable, just n-1 columns are required. This is because it indicates that the observation falls into the category that is represented by the dropped column if all other n-1 dummy columns are 0. Perfect multicollinearity, in which one column can be precisely predicted by the others, would result from having an additional column.

For instance, generating three columns using dummy variables will result in a gender variable with three categories (Male, Female, and Other). Nonetheless, an observation must be other if you are certain that it is neither Male (column 1 is 0) nor Female(column 2 is 0). As a result, the Other column is unnecessary and may be removed without affecting the data.

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

Ans: The numerical variables which have the highest level correlation with the target variable (cnt) are temp and atemp.

As you can see from pair plot below

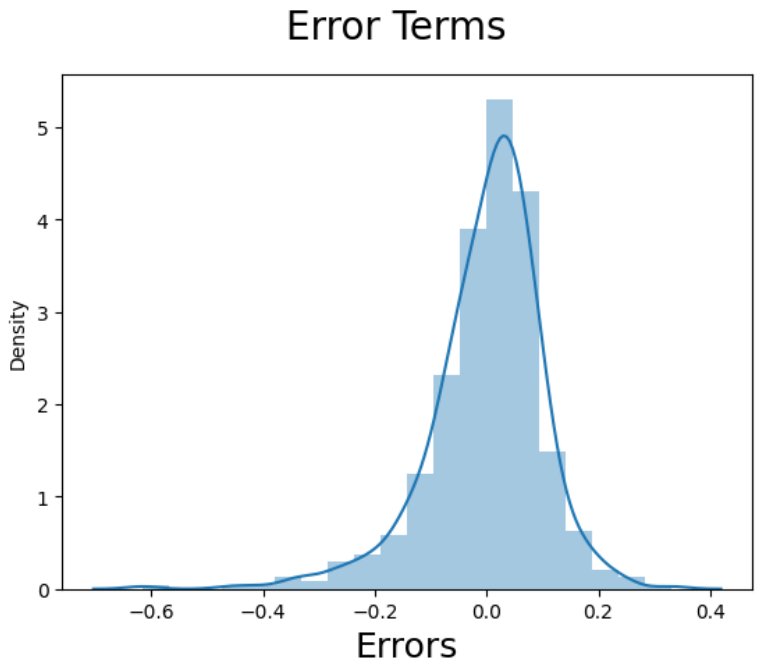


4. How did you validate the assumptions of Linear Regression after building the model on the training set?

Ans: To validate the linear regression assumptions on training data,

- There is a linear relationship between X and Y
- Error terms are normally distributed with mean zero (not X, Y)
- Residual Analysis of Training Data proves that the Residuals are normally distributed.

Hence our assumption for Linear Regression is valid. Eliminations and inclusion of independent variables into each model based on VIF and pvalues to avoid multi collinearity. The residuals are dispersed around mean = 0, as can be seen in the diagram below.



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

Ans. Here is the final equation of model

cnt = 0.239669 x yr + (-0.083884) x holiday + 0.487582 x temp + (-0.183457) x windspeed + (-0.054892) x season_Spring + 0.045609 x season_Summer + 0.068005 x season_Winter + (-0.033143) x mnth_Jan + (-0.053569) x mnth_Jul + 0.066688 x mnth_Sept + (-0.067414) x weathersit_Mist & Cloudy + 0.200957

The top 3 features contributing to demand of shared bikes based on final model are

1. **Temperature (temp) :** It has coefficient value as (0.487582)
2. **Year (yr):** It has coefficient value as (0.239669)
3. **Winter Season (season_Winter):** It has coefficient values as (0.068005)

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Ans: Linear regression is a type of supervised machine learning algorithm that computes the linear relationship between the dependent variable and one or more independent features by fitting a linear equation to observed data.
When there is only one independent feature, it is known as Simple Linear Regression, and when there are more than one feature, it is known as Multiple Linear Regression.
Linear regression is not merely a predictive tool; it forms the basis for various advanced models. Techniques like regularization and support vector machines draw inspiration from linear regression, expanding its utility. Additionally, linear regression is a cornerstone in assumption testing, enabling researchers to validate key assumptions about the data.

Types of Linear Regression

There are two main types of linear regression:

1. Simple Linear Regression

This is the simplest form of linear regression, and it involves only one independent variable and one dependent variable. The equation for simple linear regression is:

$y=\beta 0+\beta 1Xy=\beta 0+\beta 1X$

where:

- Y is the dependent variable
- X is the independent variable
- β0 is the intercept
- β1 is the slope

2. Multiple Linear Regression

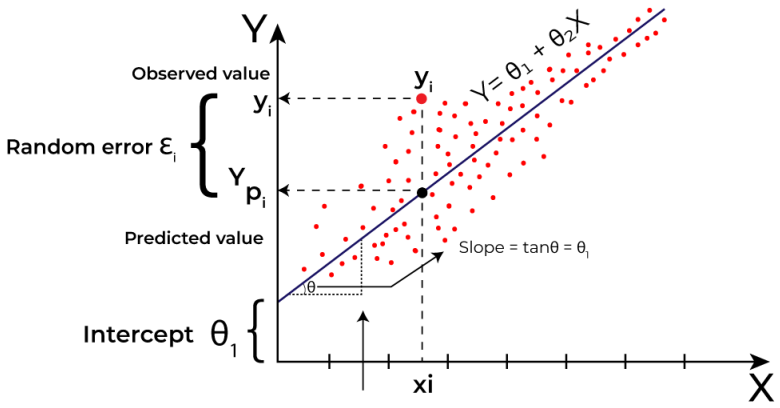
This involves more than one independent variable and one dependent variable. The equation for multiple linear regression is:

$y=\beta 0+\beta 1X+\beta 2X+\ldots\ldots\ldots\beta nXy=\beta 0+\beta 1X+\beta 2X+\ldots\ldots\ldots\beta nX$

where:

- Y is the dependent variable
- X1, X2, …, Xp are the independent variables
- β0 is the intercept
- β1, β2, …, βn are the slopes

The goal of the algorithm is to find the best Fit Line equation that can predict the values based on the independent variables.
Our primary objective while using linear regression is to locate the best-fit line, which implies that the error between the predicted and actual values should be kept to a minimum. There will be the least error in the best-fit line.
The best Fit Line equation provides a straight line that represents the relationship between the dependent and independent variables. The slope of the line indicates how much the dependent variable changes for a unit change in the independent variable(s).



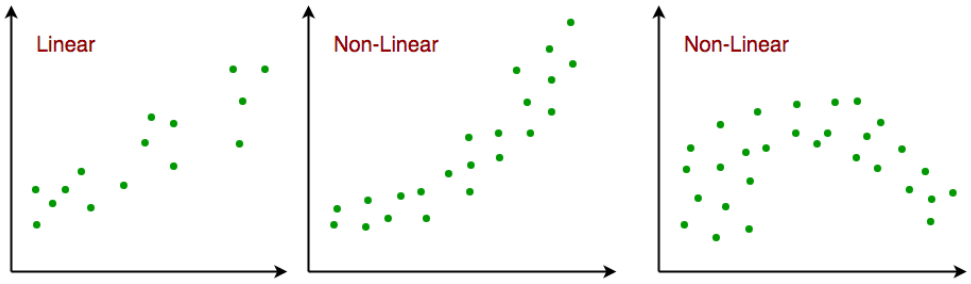
Linear Regression

Here Y is called a dependent or target variable and X is called an independent variable also known as the predictor of Y. There are many types of functions or modules that can be used for regression. A linear function is the simplest type of function. Here, X may be a single feature or multiple features representing the problem.

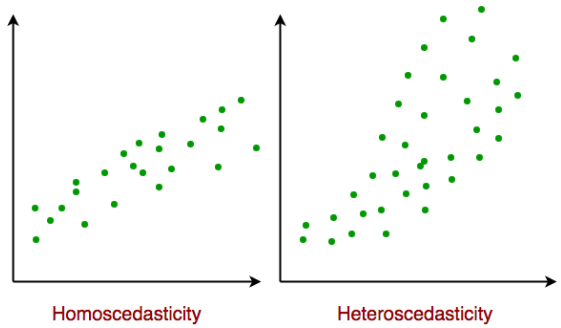
Assumptions of Simple Linear Regression

Linear regression is a powerful tool for understanding and predicting the behavior of a variable, however, it needs to meet a few conditions in order to be accurate and dependable solutions.

1. **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. This means that there should be a straight line that can be drawn through the data points. If the relationship is not linear, then linear regression will not be an accurate model.



2. **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. If the observations are not independent, then linear regression will not be an accurate model.
3. **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. If the variance of the residuals is not constant, then linear regression will not be an accurate model.



4. **Normality**: The residuals should be normally distributed. This means that the residuals should follow a bell-shaped curve. If the residuals are not normally distributed, then linear regression will not be an accurate model.

Assumptions of Multiple Linear Regression

For Multiple Linear Regression, all four of the assumptions from Simple Linear Regression apply. In addition to this, below are few more:

1. **No multicollinearity**: There is no high correlation between the independent variables. This indicates that there is little or no correlation between the independent variables. Multicollinearity occurs when two or more independent variables are highly correlated with each other, which can make it difficult to determine the individual effect of each variable on the dependent variable. If there is multicollinearity, then multiple linear regression will not be an accurate model.
2. **Additivity:** The model assumes that the effect of changes in a predictor variable on the response variable is consistent regardless of the values of the other variables. This assumption implies that there is no interaction between variables in their effects on the dependent variable.
3. **Feature Selection:** In multiple linear regression, it is essential to carefully select the independent variables that will be included in the model. Including irrelevant or redundant variables may lead to overfitting and complicate the interpretation of the model.
4. **Overfitting:** Overfitting occurs when the model fits the training data too closely, capturing noise or random fluctuations that do not represent the true underlying relationship between variables. This can lead to poor generalization performance on new, unseen data.

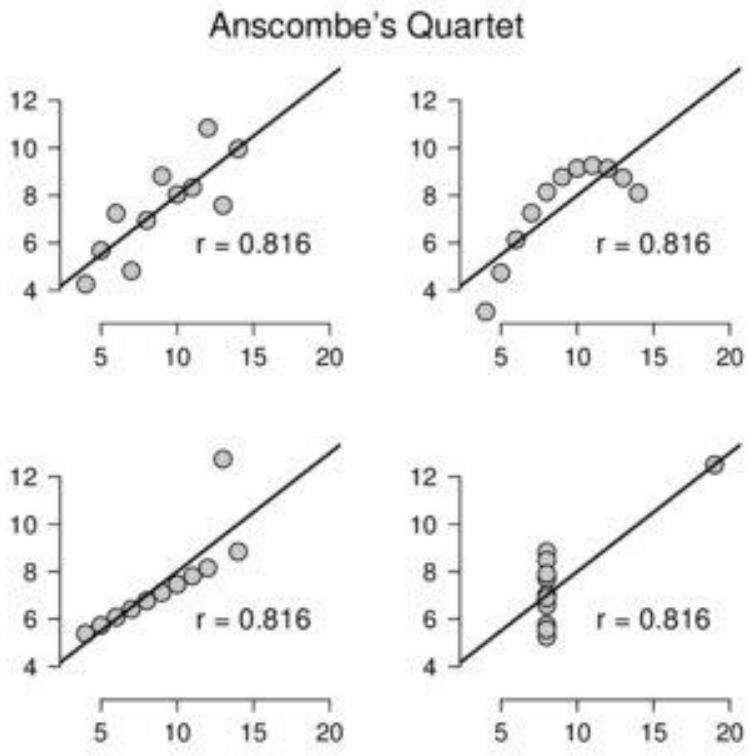
2. Explain the Anscombe's quartet in detail.

Ans: Anscombe's quartet was constructed in 1973 by statistician Francis Anscombe to illustrate the importance of plotting data before you analyze it and build your model. These four data sets have nearly the same statistical observations, which provide the same information (involving variance and mean) for each x and y point in all four data sets. However, when you plot these data sets, they look very different from one another. Anscombe's quartet intended to counter the impression among statisticians that "numerical calculations are exact, but graphs are rough.

These datasets for four plots are shown along with their statistical information:

	Chart I		Chart II		Chart III		Chart IV	
Point #	x	y	x	y	x	y	x	y
1	10.00	8.04	10.00	9.14	10.00	7.46	8.00	6.58
2	8.00	6.95	8.00	8.14	8.00	6.77	8.00	5.76
3	13.00	7.58	13.00	8.74	13.00	12.74	8.00	7.71
4	9.00	8.81	9.00	8.77	9.00	7.11	8.00	8.84
5	11.00	8.33	11.00	9.26	11.00	7.81	8.00	8.47
6	14.00	9.96	14.00	8.10	14.00	8.84	8.00	7.04
7	6.00	7.24	6.00	6.13	6.00	6.08	8.00	5.25
8	4.00	4.26	4.00	3.10	4.00	5.39	19.00	12.50
9	12.00	10.84	12.00	9.13	12.00	8.15	8.00	5.56
10	7.00	4.82	7.00	7.26	7.00	6.42	8.00	7.91
11	5.00	5.68	5.00	4.74	5.00	5.73	8.00	6.89
Sum	**99.00**	**82.51**	**99.00**	**82.51**	**99.00**	**82.50**	**99.00**	**82.51**
Average	**9.00**	**7.50**	**9.00**	**7.50**	**9.00**	**7.50**	**9.00**	**7.50**
St.dev	**3.16**	**1.94**	**3.16**	**1.94**	**3.16**	**1.94**	**3.16**	**1.94**

When these models are plotted on a scatter plot, all datasets generates a different kind of plot that is not interpretable by any regression algorithm which is fooled by these peculiarities and can be seen as follows



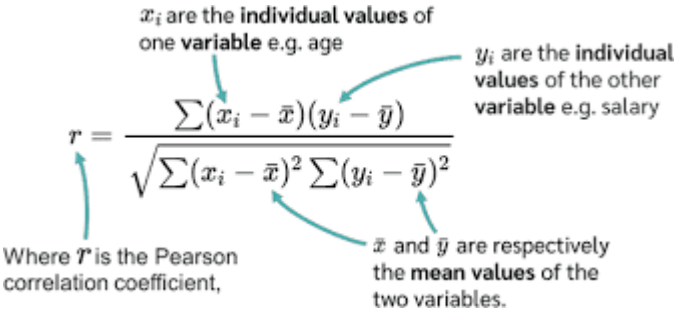
The four datasets can be described as:

- The first scatter plot (top left) appears to be a simple linear relationship,
- The second graph (top right); cannot fit the linear regression model because the data is non-linear
- In the third graph (bottom left), the modelled relationship is linear, but should have a different regression line. It shows the outliers involved in the dataset which cannot be handled by linear regression model
- Finally, the fourth graph (bottom right) shows the outliers involved in the dataset which cannot be handled by linear regression model. It 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. It shows the outliers involved in the dataset which cannot be handled by linear regression model

3. What is Pearson's R?

Ans**:** The Pearson correlation coefficient (r) is the most common way of measuring a linear correlation. It is a number between –1 and 1 that measures the strength and direction of the relationship between two variables.

Which can be calculated by



Pearson correlation coefficient (*r*) value	Strength	Direction
Greater than .5	Strong	Positive
Between .3 and .5	Moderate	Positive
Between 0 and .3	Weak	Positive
0	None	None
Between 0 and –.3	Weak	Negative
Between –.3 and –.5	Moderate	Negative
Less than –.5	Strong	Negative

The Pearson correlation coefficient is a good choice when all of the following are true:

- Both variables are quantitative: You will need to use a different method if either of the variables is qualitative.
- The variables are normally distributed: You can create a histogram of each variable to verify whether the distributions are approximately normal. It's not a problem if the variables are a little non-normal.
- The data have no outliers: Outliers are observations that don't follow the same patterns as the rest of the data. A scatterplot is one way to check for outliers—look for points that are far away from the others.
- The relationship is linear: "Linear" means that the relationship between the two variables can be described reasonably well by a straight line. You can use a scatterplot to check whether the relationship between two variables is linear

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

Ans:
Scaling is a data preprocessing technique used to standardize the range of independent variables or features of data. It is an important step in many machine learning algorithms, as it can significantly impact the performance and convergence of the model.

There are two main reasons why scaling is performed:

1. To avoid features in greater numeric ranges dominating those in smaller numeric ranges: If one feature has a range of values from 1 to 1000 and another from 1 to 10, the algorithm may think the first feature is more important simply because of its larger range, even though the smaller range feature may be just as relevant.
2. To ensure numerical stability: Many machine learning algorithms, such as gradient descent, perform better when all features are on a similar scale. Scaling helps prevent features with larger ranges from dominating the objective function and can lead to faster convergence.

The two most common scaling techniques are normalization and standardization:

Normalization (Min-Max Scaling):

- Scales the features to a common range, typically 0 to 1.
- Uses the formula: x_scaled = (x - min(x)) / (max(x) - min(x))
- Preserves the shape of the original distribution.
- Sensitive to outliers, as they can skew the min and max values.

Standardization (Z-score Normalization):

- Transforms the features to have a mean of 0 and a standard deviation of 1.
- Uses the formula: x_scaled = (x - mean(x)) / std(x)
- Does not bound the values to a specific range.
- More robust to outliers compared to normalization.
- Assumes the data follows a Gaussian distribution.

The choice between normalization and standardization depends on the specific problem and algorithm being used. Normalization is preferred when the data follows a different distribution than the normal distribution, while standardization is more suitable when the data is normally distributed. Standardization is also commonly used when the scale of the features is not important, such as in algorithms that use the dot product between data points (e.g., linear discriminant analysis). In summary, scaling is an essential preprocessing step in machine learning that helps improve the performance and stability of many algorithms. Normalization and standardization are the two most common scaling techniques, with different properties and use cases.

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

Ans: The reason VIF (Variance Inflation Factor) becomes infinite in linear regression is due to perfect multicollinearity among the independent variables.

VIF essentially tells you how much the variance of an estimated regression coefficient is inflated due to collinearity with other independent variables. A VIF value of 1 indicates no inflation, while higher values suggest increasing inflation.

Perfect multicollinearity: This occurs when one independent variable can be perfectly predicted by a linear combination of the other independent variables. In simpler terms, one variable is entirely redundant and contains no unique information compared to the others.

Impact on VIF: When perfect multicollinearity exists, the regression model cannot uniquely estimate the coefficient for the perfectly collinear variable. This leads to a situation where the denominator in the VIF calculation becomes zero as it's R-squared value will be equal to 1. Dividing by zero is mathematically undefined, hence the VIF value becomes infinite.

So, VIF = 1/(1-1) which gives VIF = 1/0 which results in "infinity"

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression?

Ans: Quantile-Quantile (Q-Q) plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal, exponential or Uniform distribution. Also, it helps to determine if two data sets come from populations with a common distribution. 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.

Advantages:
a) It can be used with sample sizes also
b) 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.

Here's how a Q-Q plot works:

a) Rank the residuals: You order the residuals from your model, from smallest to largest. Calculate theoretical quantiles: You calculate the expected quantiles of a normal distribution with the same number of data points as your residuals.
b) Plot the quantiles: You plot the ranked residuals against the theoretical quantiles.
c) Interpreting the Plot: Perfect fit: If the points in your plot fall roughly along a straight diagonal line, it suggests that your residuals are close to being normally distributed. This is a good sign for your linear regression model.
d) Deviations from the line: Deviations from the straight line indicate departures from normality. For example, if the points curve upwards, it might suggest a right-skewed distribution of residuals.

Importance of Normality:

a) Validity of p-values: The p-values generated by your regression model rely on the normality assumption. If residuals aren't normal, p-values may not be reliable indicators of statistical significance.
b) Confidence intervals: Confidence intervals for your model's coefficients also depend on the normality assumption. Departures from normality can lead to unreliable confidence intervals.
c) Robustness: While linear regression can still be useful even with non-normal residuals, the model might be less robust and susceptible to outliers.