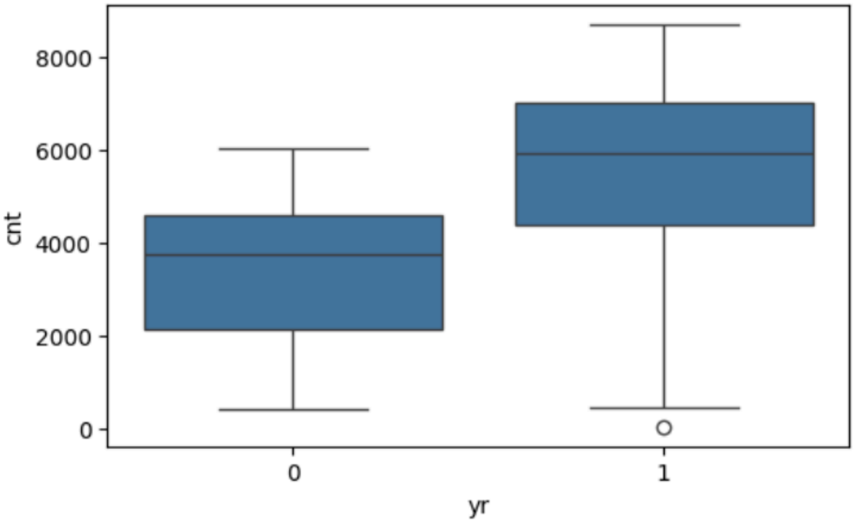
Assignment-based Subjective Questions

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

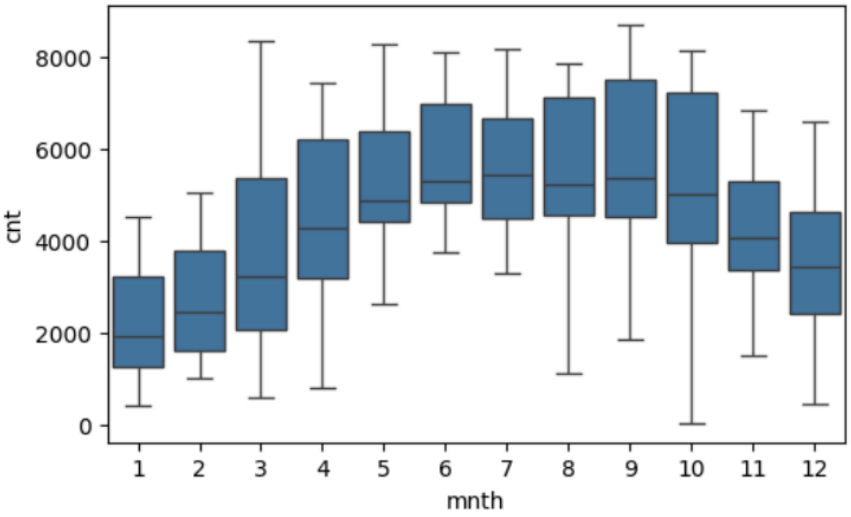
Answer :

For each of the provided categorical variables, these were the effect on dependent variable :

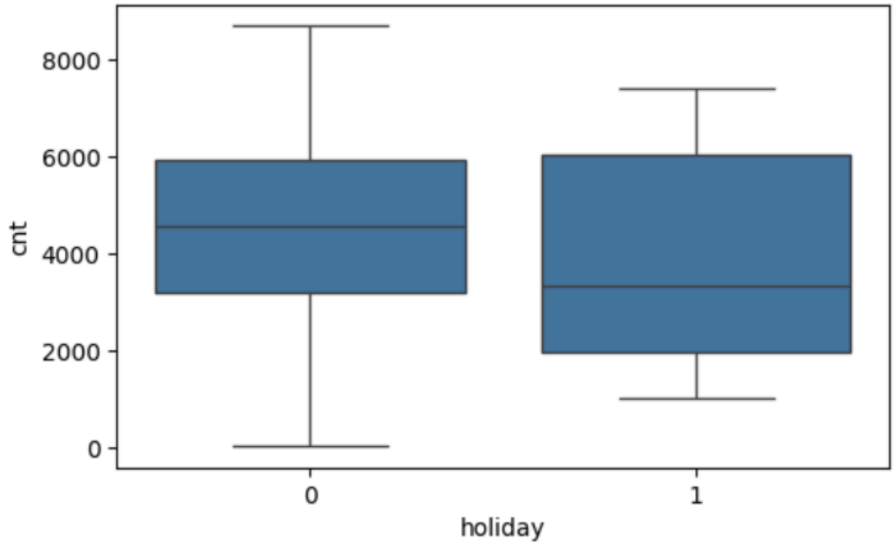
1. \`yr\` : Year on year, count has increased a lot , which is in line with the fact that the company is growing



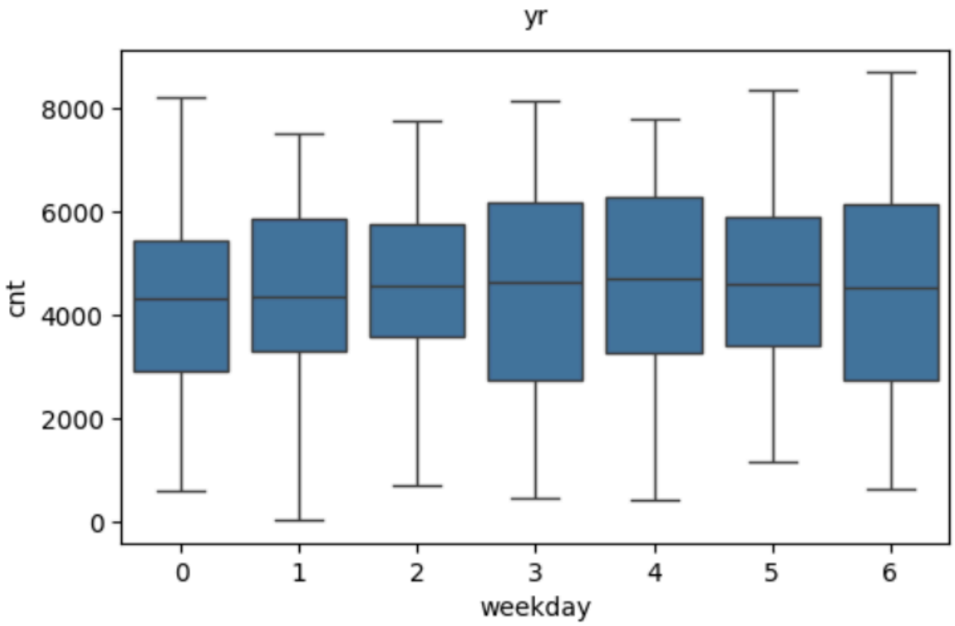
2. \`mnth\`: Months 8, 9, 10 seem to have really high count, whereas month 1, 2, 10, 12 seem to have very low count. We can clearly see people prefer some months over the other



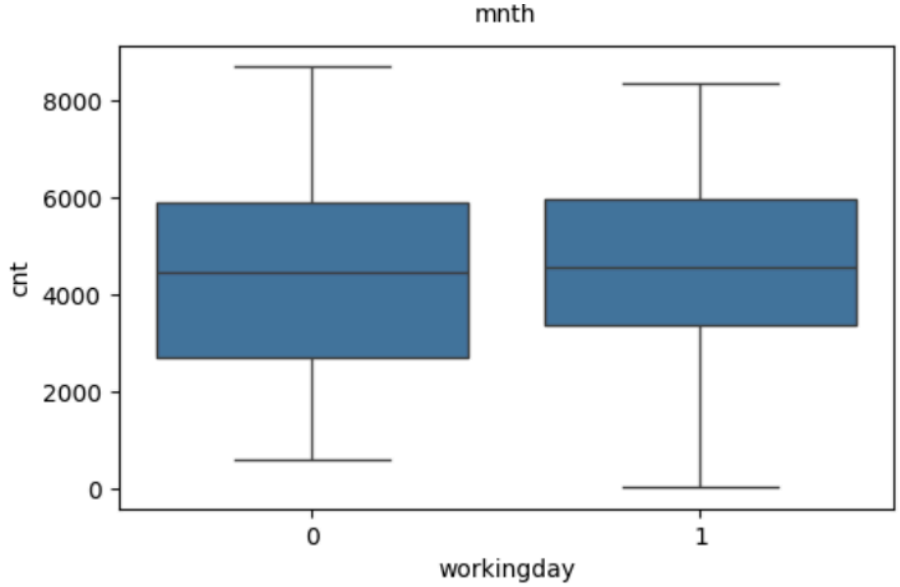
3. `holiday`: We can see that median is higher for non holiday period, but 75th percentile is pretty much the same so we cant say too much from this categorical variable



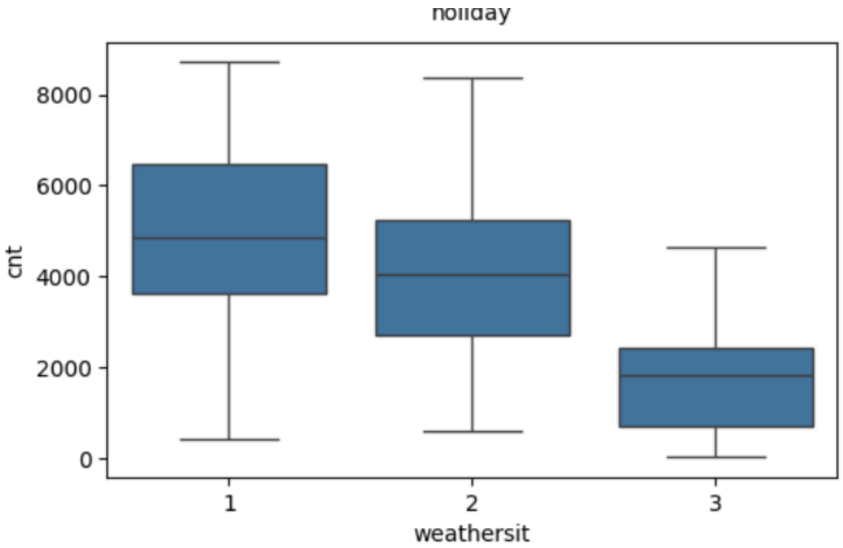
4. `weekday`: Count is pretty much evenly distributed across all weekdays



5. `workingday`: median and 75th percentile is basically same for the two values, effect is not obvious



6. `weathersit`: Relatively clear weather has more count of users, the worse the weather gets, the less the count. Clear effect is there

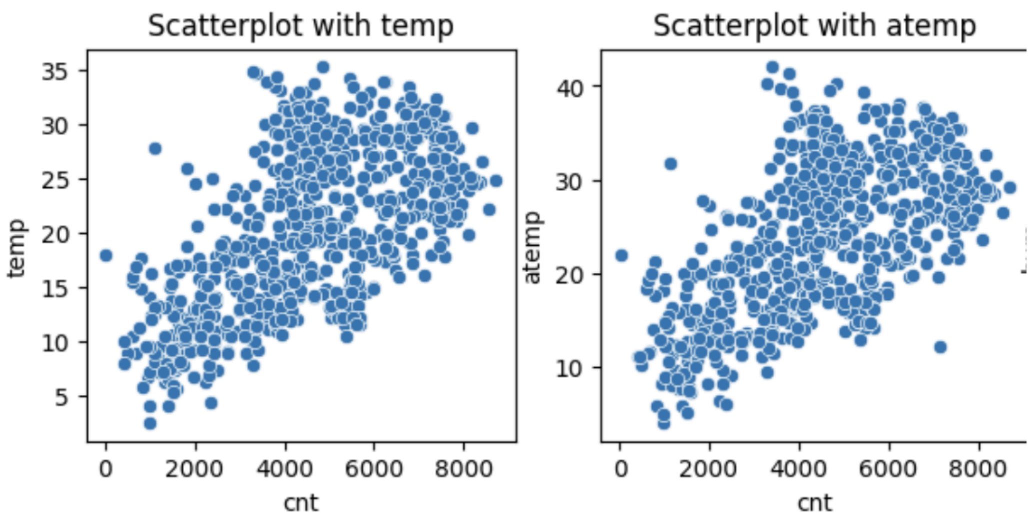


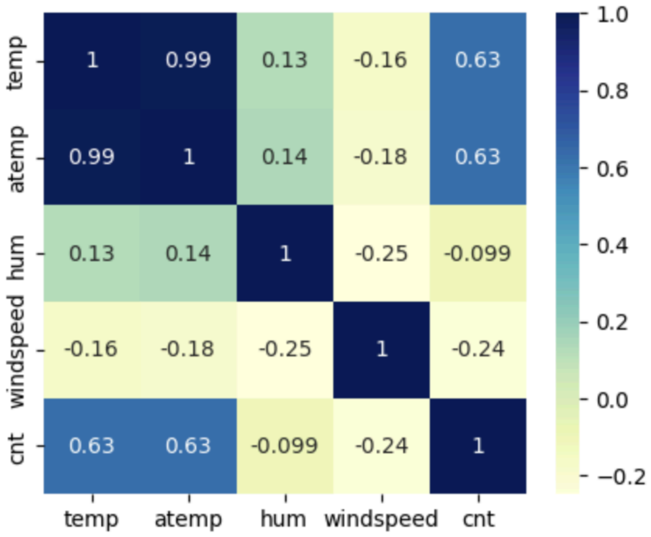
Q.2 Why is it important to use drop_first=True during dummy variable creation?

Answer:
If categorical variable has n possible values, we only need n-1 dummy variables. This is because the nth value can be represented by having all dummy variables as 0. drop_first=True ensures that only n-1 dummy variables are created.

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

Answer:
From the pair plot, `temp` and `atemp` both have high correlation with the target variable. This is evident from the scatterplot, which shows both of them having a sort of linear relationship with the target variable . We can also plot correlation matrix, which shows both `atemp` and `temp` having 63% correlation with target.



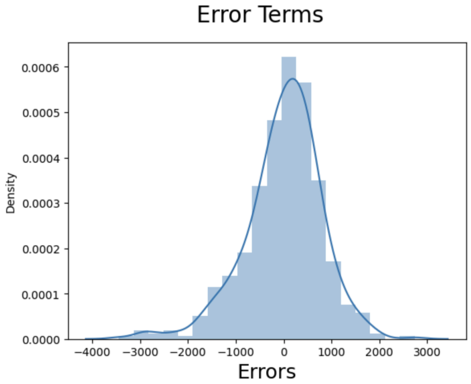


	temp	atemp	hum	windspeed	cnt
temp	1	0.99	0.13	-0.16	0.63
atemp	0.99	1	0.14	-0.18	0.63
hum	0.13	0.14	1	-0.25	-0.099
windspeed	-0.16	-0.18	-0.25	1	-0.24
cnt	0.63	0.63	-0.099	-0.24	1

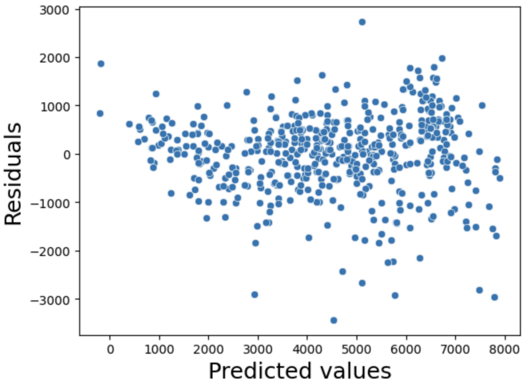
Q.4 How did you validate the assumptions of Linear Regression after building the model on the training set?

It was observed that R-squared value was 83% for the model, which shows that predictor variables and target follow a somewhat linear relationship.

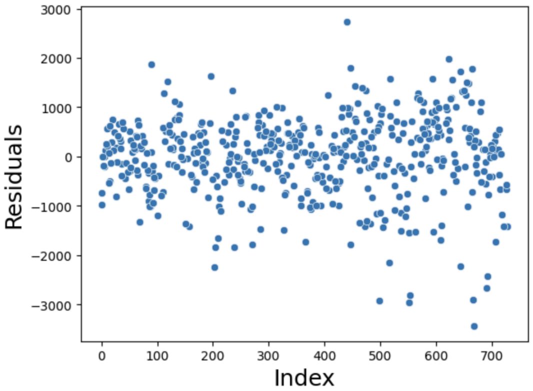
Residual analysis was performed on training data, to verify the other assumptions. First of all, error terms were plotted and it was observed that they followed a normal distribution with mean around 0.



After that, to verify that error terms have constant variance, error terms were plotted against target variable, and it was observed that increase in target had no significant impact on the error term.



To verify that error terms are independent, error terms were plotted against index, to see if their is any visible pattern, none were found, proving that they were independent.



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

Answer:
The top 3 features contributing significantly towards explaining the demand of the shared bikes, judged by the absolute values of their coefficients, are:

1. temp (Temperature): Coefficient = 5157.4381

   This indicates that temperature has the strongest positive impact on bike demand among the variables included in the model. A unit increase in temperature is associated with an increase of approximately 5157.44 in the demand for shared bikes, holding other variables constant.
2. hum (Humidity): Coefficient = -2433.7771

   Humidity has a significant negative impact on bike demand. A unit increase in humidity is associated with a decrease of approximately 2433.78 in the demand for shared bikes, holding other factors constant.
3. yr (Year): Coefficient = 1972.5545

   This suggests a year-over-year increase in bike demand, possibly reflecting growing popularity or expansion of the bike-sharing system. A shift from the base year (presumably 0 for 2018 and 1 for 2019) is associated with an increase of approximately 1972.55 in the demand for shared bikes.

These variables are critical in understanding how environmental conditions and time (year) influence the demand for shared bikes. Temperature, being the variable with the highest coefficient, shows a strong positive relationship, meaning that warmer conditions likely boost bike-sharing demand. In contrast, increased humidity tends to reduce demand, possibly due to discomfort. The positive coefficient for the year indicates an overall increase in demand, suggesting that the bike-sharing system is becoming more popular or accessible over time.

General Subjective Questions

Q.1 Explain the linear regression algorithm in detail.

Answer:
Linear regression is a fundamental statistical and machine learning method used to model the relationship between a dependent variable and one or more independent variables. The goal of linear regression is to find the best-fitting line through the data points that minimizes the differences between the observed values and the values predicted by the model. This method is

widely used in predictive analytics to forecast values and identify relationships between variables.

The Linear Regression Model

The basic form of a linear regression model with one independent variable is expressed as:

y = mX + c + e

Where:
y is the dependent variable we aim to predict or explain.
X is the independent variable used to predict y.
c is the intercept of the regression line, representing the predicted value of y when x is 0.
m is the slope of the regression line, representing the change in y for a one-unit change in x.
e is the error term, accounting for the difference between the observed and predicted values.

In multiple linear regression, where there are two or more independent variables, the model is extended as:

y = m1*X1 + m2*X2 + …. + mn*Xn + c + e

Estimating the Model Parameters

The process of fitting a linear regression model involves estimating the parameters (m1, m2, …, mn) that minimize the sum of the squared differences between the observed and predicted values. This method is known as the Least Squares method. The best-fitting line is the one where the sum of the squares of these differences is the smallest.

Assumptions of Linear Regression

Linear regression relies on several key assumptions:

Linearity: The relationship between the dependent and independent variables should be linear.
Independence: The residuals (errors) should be independent of each other.
Homoscedasticity: The variance of the error terms should be constant across all levels of the independent variables.
Normal Distribution of Errors: The error terms should be normally distributed.

Model Evaluation

After fitting the model, it's crucial to evaluate its performance. Common metrics for this purpose include:

R-squared: Represents the proportion of the variance in the dependent variable that is predictable from the independent variables. It ranges from -inf to 1, with higher values indicating a better fit.
Adjusted R-squared: Adjusted for the number of predictors in the model, providing a more accurate measure when comparing models with different numbers of independent variables.

Mean Squared Error (MSE) and Root Mean Squared Error (RMSE): Indicate the average squared difference and the square root of the average squared difference, respectively, between observed and predicted values.

P value of coefficients : High p value means that the coefficient isn’t significant and for better interpretation we might as well get rid of it.

Q.2 Explain the Anscombe’s quartet in detail.

Answer:
Anscombe's quartet consists of four datasets that are designed to have nearly identical simple descriptive statistics, yet they have very different distributions and appear very different when graphed. Each dataset consists of eleven points. This quartet was crafted by the statistician Francis Anscombe in 1973 to demonstrate the importance of visualizing data before analyzing it, as well as the effect outliers and other anomalies can have on statistical properties.

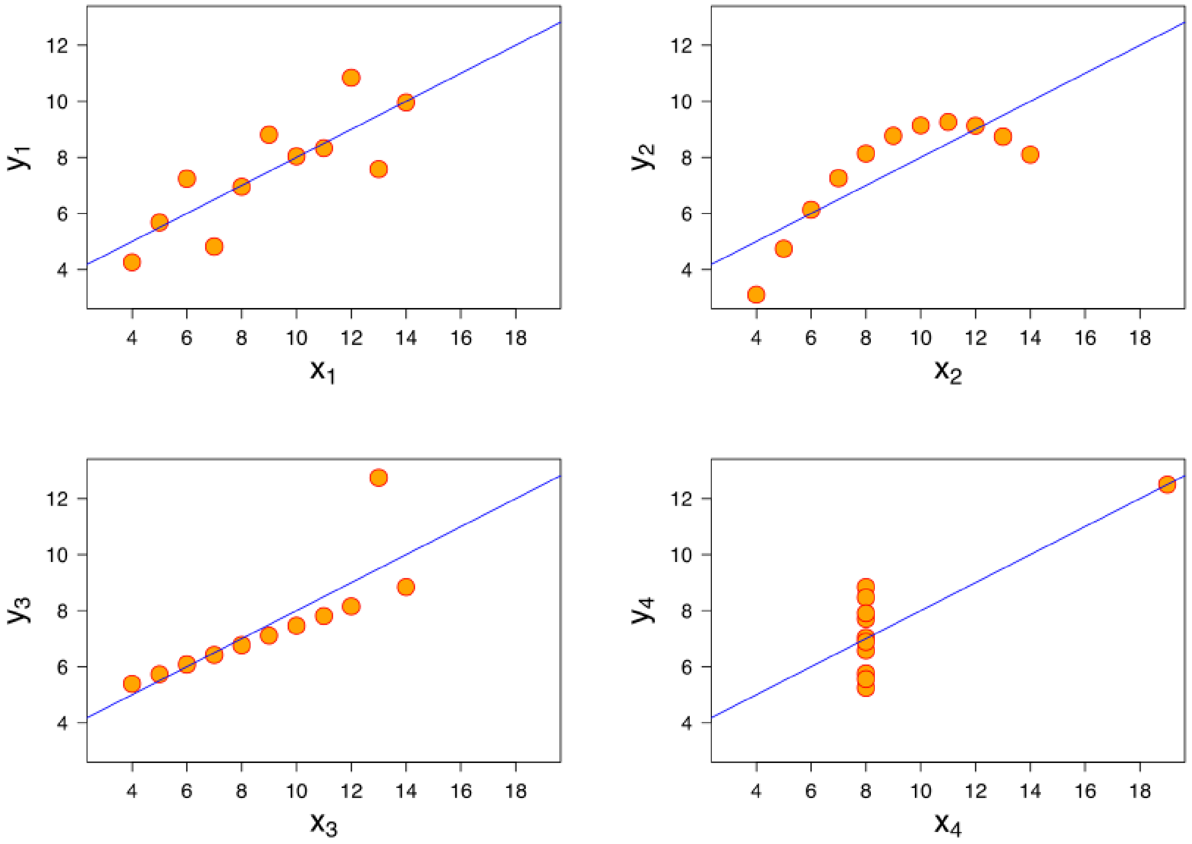
Key Properties:
Despite their different appearances when plotted, all four datasets share the following statistical properties:

1. The mean of the x values is approximately 9 for all datasets, and the mean of the y values is approximately 7.5.
2. The variance of x values is the same for all datasets, as is the variance of y values.
3. The correlation between x and y values is approximately 0.816 for each dataset.
4. The linear regression line (the best fit line) for each dataset is y = 3.00 + 0.500x.

These similarities suggest that the datasets might be similar in terms of their underlying relationship between x and y values. However, when each dataset is graphed, they reveal very different relationships:

1. Dataset I shows a simple linear relationship, matching our expectations for a typical dataset where the assumptions of linear regression hold true.
2. Dataset II demonstrates a clear nonlinear relationship. The points follow a curved pattern, illustrating that a linear model would not be appropriate.
3. Dataset III appears to be a linear relationship similar to Dataset I, but with one clear outlier that drastically influences the slope of the regression line.
4. Dataset IV shows that when one point is a far outlier in terms of the x variable (horizontal outlier), it can have an excessive influence on the regression line, even if the relationship between the other points does not support that trend.

Anscombe's quartet is a powerful demonstration of why graphical analysis is an essential part of data analysis, alongside numerical methods. It illustrates how critical it is to not solely rely on statistical properties when analyzing data, as very different datasets can share the same statistical summaries. The quartet also highlights the impact of outliers and the importance of considering the appropriateness of the model being used for data analysis.



Q.3 What is Pearson’s R ?

Answer:

Pearson's R, also known as Pearson's correlation coefficient, is a measure of the strength and direction of the linear relationship between two continuous variables. It is a widely used statistical metric that quantifies the degree to which two variables linearly relate to each other. Pearson's R values range from -1 to 1, where:

1. 1 indicates a perfect positive linear relationship: as one variable increases, the other variable also increases in a proportional manner.
2. -1 indicates a perfect negative linear relationship: as one variable increases, the other variable decreases in a proportional manner.
3. 0 indicates no linear relationship: there is no consistent pattern of increase or decrease between the two variables.

The formula to calculate Pearson's R is:

Pearson's R = (cov(X, Y)) / (σX * σY)

Where:

1. cov(X, Y) is the covariance of the two variables, representing how changes in one variable are associated with changes in another.
2. σX and σY are the standard deviations of X and Y, respectively, measuring the spread of each variable.

Pearson's R is most appropriate for data that is both normally distributed and linearly related. It should be used with caution when these assumptions are not met, as it may not accurately represent the relationship between the variables. Pearson's correlation coefficient is a powerful tool for exploratory data analysis, helping researchers and analysts to identify potential relationships that warrant further investigation.

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

Answer:
Scaling is a key data preprocessing technique in machine learning that adjusts the range of feature variables. Scaling is crucial because many algorithms, particularly those that compute distances or gradients, assume that all features are on similar scales. Without scaling, features with larger numerical ranges can dominate those with smaller ranges, potentially leading to biased or inefficient learning.

Why is Scaling Performed?

1. Equal Importance: Ensures every feature contributes equally to the model, preventing features with larger scales from overshadowing those with smaller scales.
2. Optimization Efficiency: Algorithms using gradient descent, like linear and logistic regression or neural networks, converge faster when features are on the same scale, as it improves the path to the minimum of the cost function.

Normalized Scaling vs. Standardized Scaling

Normalization rescales data to a specific range, often 0 to 1 or -1 to 1. It's done by subtracting the minimum value of a feature and then dividing by the range of that feature. This method is useful when you need values to be in a specific range, but it can be sensitive to outliers since it directly uses the minimum and maximum values for scaling.

Standardization transforms data to have a mean of 0 and a standard deviation of 1. It's achieved by subtracting the mean of the feature from each value, then dividing by the standard deviation of the feature. Standardization is less affected by outliers and is beneficial when the feature distribution is not uniform or normal.

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

Answer:
An infinite VIF occurs when there is perfect multicollinearity in the data, meaning one independent variable can be perfectly linearly predicted from the others with absolute certainty. In mathematical terms, this happens when the independent variables have an exact linear relationship among themselves, leading to a determinant of zero in the matrix used to calculate VIF values. Since VIF is calculated as:

VIF_i = 1 / (1 - R_i^2)

where R_i^2 is R-squared value when other predictor variables together try to predict i-th predictor variable
If R_i^2 = 1, indicating perfect multicollinearity, the denominator becomes zero, making the VIF infinite.

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

Answer:
A Q-Q (Quantile-Quantile) plot is a graphical tool used to compare two probability distributions by plotting their quantiles against each other. If both distributions are similar, the points in the Q-Q plot will approximately lie on the line y = x. In the context of linear regression, Q-Q plots are primarily used to analyze the distribution of residuals (errors) from the model.

Use of Q-Q Plot in Linear Regression:

In linear regression, one of the key assumptions is that the residuals are normally distributed. A Q-Q plot is an excellent tool for checking this assumption. By comparing the quantiles of the residuals to the expected quantiles from a normal distribution, the Q-Q plot can show whether the residuals follow a normal distribution pattern.

How to Interpret a Q-Q Plot:

1. Points Lying on the 45-degree Line: If the points in the Q-Q plot lie approximately along the 45-degree line that passes through the origin, it suggests that the residuals have a distribution similar to the theoretical distribution (often the normal distribution in this context).

2. Deviations from the Line: If the points deviate systematically from the line, it indicates that the residuals do not follow a normal distribution. For instance, if the points curve upwards away from the line, the residuals may be right-skewed; if they curve downwards, the residuals may be left-skewed.

Importance of a Q-Q Plot in Linear Regression:

1. Validity of Statistical Tests: Many statistical tests used in the context of linear regression, like the t-test for coefficients, assume normality of residuals. A Q-Q plot helps validate these assumptions, ensuring the reliability of the test results.
2. Model Diagnostics: A Q-Q plot can reveal issues with the model, such as skewness or outliers in the residuals, that could affect the model's performance. Identifying and addressing these issues can improve the model's accuracy and reliability.
3. Choosing Appropriate Transformations: If residuals are not normally distributed, transformations of the dependent variable or the use of alternative distributions in generalized linear models might be necessary. The Q-Q plot helps in identifying the need for such transformations.