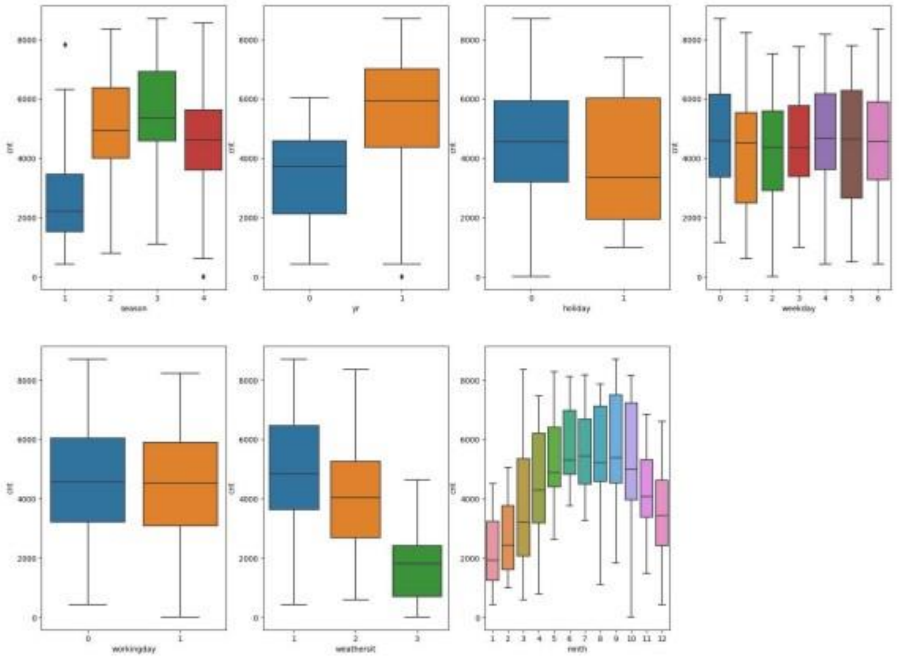
Assignment-based Subjective Questions

By Asif iqbal

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

Answer- Categorical variables used in the dataset: season, year. Holiday, Day of the Week, Working Day and Weather Conditions and Month. These were visualized using a boxplot.



<u>These variables had the following effect on our dependent variable:-</u>

* The fall season seems to have attracted more bookings. At the same time, there has been a huge increase in the number of bookings in every season from 2018 to 2019.

* Weekday - The bike demand is almost constant throughout the week.

* There were higher number of bookings in 2019 compared to the previous year, which shows there will be good progress in business matters.

* It seems that the number of bookings goes down when there are no holidays Suitable because during holidays people will want to spend time at home and enjoy with family

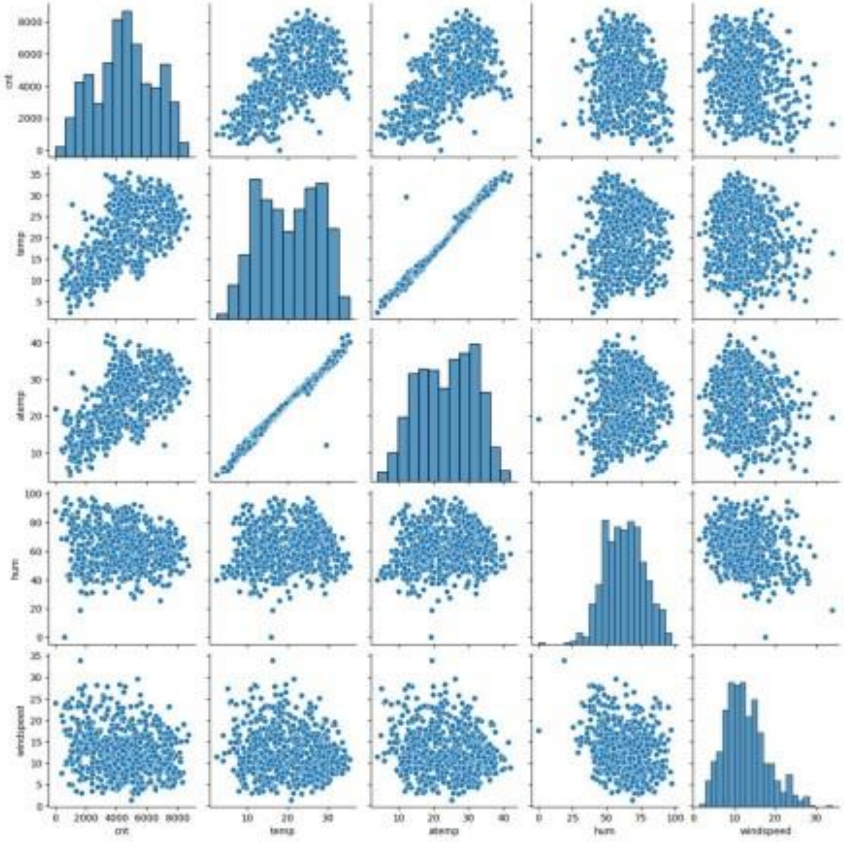
2 .Why is it important to use drop first=True during dummy variable creation?

Answer: 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.

Let’s 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 A and B, then It is obvious C. So we do not need 3rd variable to identify the C.

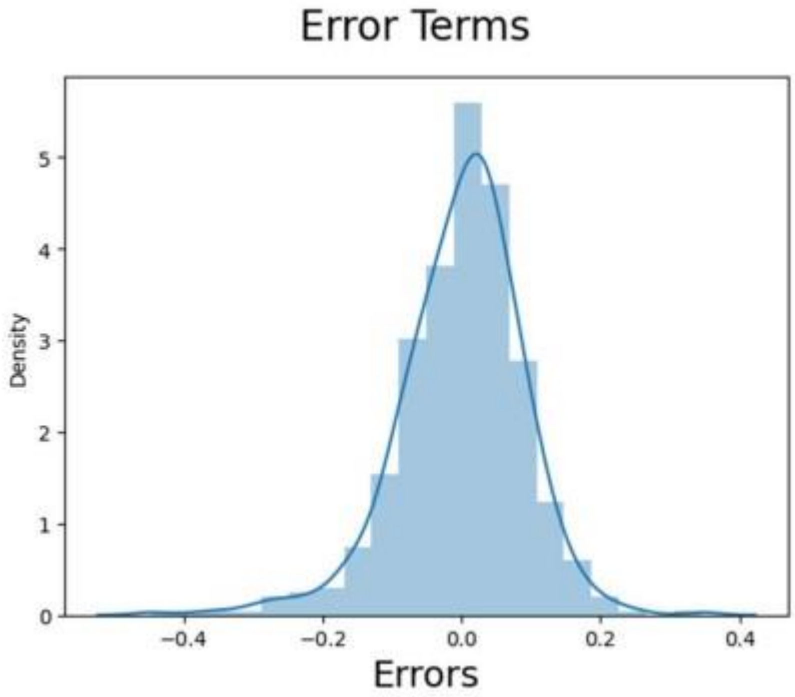
3.Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

Answer: "Temp" and "Attemp" are two numerical variables that are highly correlated target Variable.



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

Answer:- I have validated the assumption of Linear Regression Model based on below 5 assumptions



Normality of error terms

* Error terms should be normally distributed

Multicollinearity check

* There should be insignificant multicollinearity among variables.

Linear relationship validation

* Linearity should be visible among variables

Homoscedasticity

* There should be no visible pattern in residual values.

Independence of residuals

* No auto-correlation

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

Answer: - Below are the top 3 features that are significantly contributing to explaining its demand Shared Bike

1. Temperature 2. Year 3. Weather

General Subjective Questions

Q.1 Explain the linear regression algorithm in detail

Answer: Linear Regression is a fundamental machine learning algorithm used for predictive analysis. It is important to understand its mechanics to understand how it operates and apply it effectively.

Introduction:

The purpose of linear regression is to model the relationship between a dependent variable and one or more independent variables. The goal is to find a linear equation that best represents this relationship.

Types of Linear Regression:-

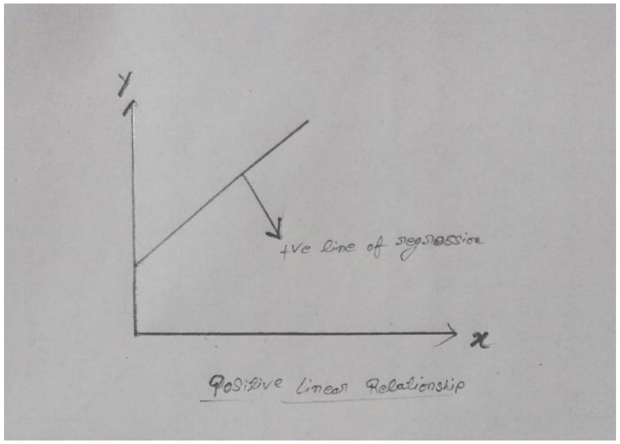
Simple Linear Regression: A single independent variable is used to predict the value of a numerical dependent variable, such linear regression algorithm is called simple linear regression.

Multiple Linear regression: If more than one independent variable is used to predict the value of a numerical dependent variable, then such a Linear Regression algorithm is called Multiple Linear Regression.

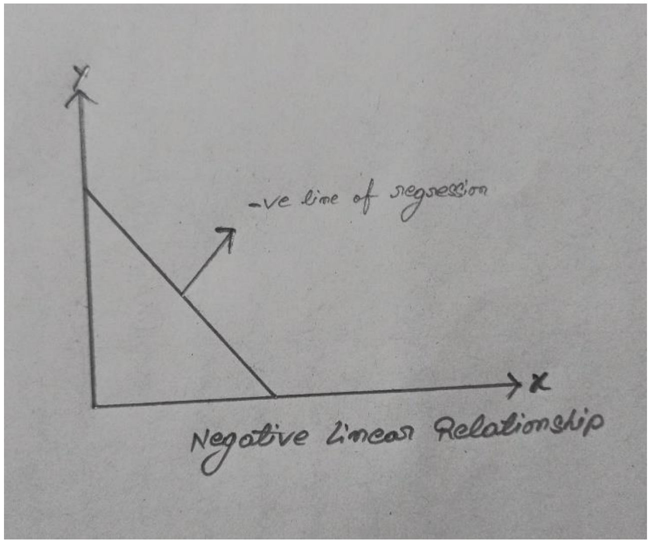
Linear Regression Line: - A linear line showing the relationship between the dependent and independent variables is called a regression line.

A regression line can show two types of relationship:

Positive Linear Relationship: If the dependent variable increases on the Y-axis and independent variable increases on X-axis, then such a relationship is termed as a Positive linear relationship.

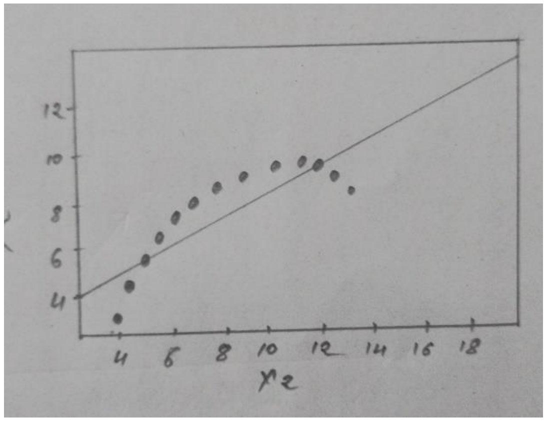


Negative Linear Relationship: If the dependent variable decreases on the Y-axis and independent variable increases on the X-axis, then such a relationship is called a negative linear relationship.

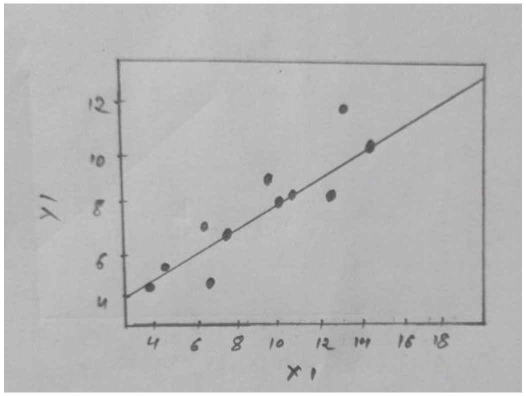


2. Explain the Anscombe’s quartet in detail.

Answer It is used to illustrate the importance of exploratory data analysis and the drawbacks of depending only on summary statistics. It also emphasizes the importance of using data visualization to spot trends, outliers, and other crucial details that might not be obvious from summary statistics alone.

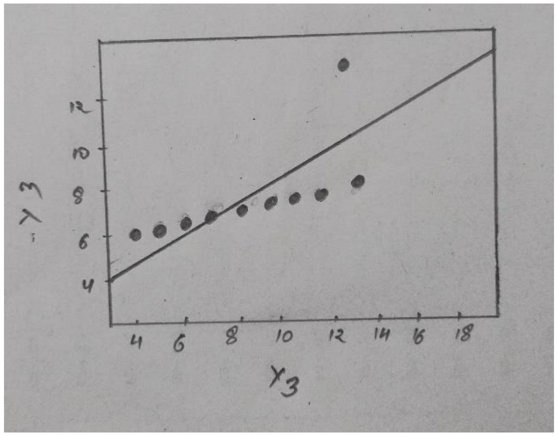


1. The first scatter plot appears to be a simple linear relationship

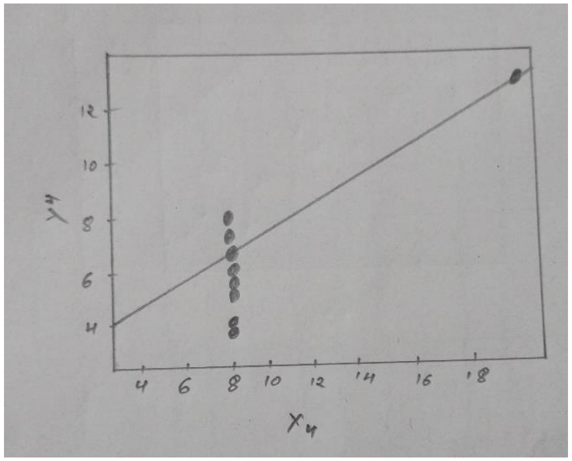


2 .The second graph is not distributed normally; while there is a relation between them is not linear

3. In the third graph the distribution is linear, but should have a different regression line the calculated regression is offset by the one outlier which exerts enough influence to lower the correlation coefficient from 1 to 0.816.



4. Finally, the fourth graph 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



3. What is Pearson’s R?

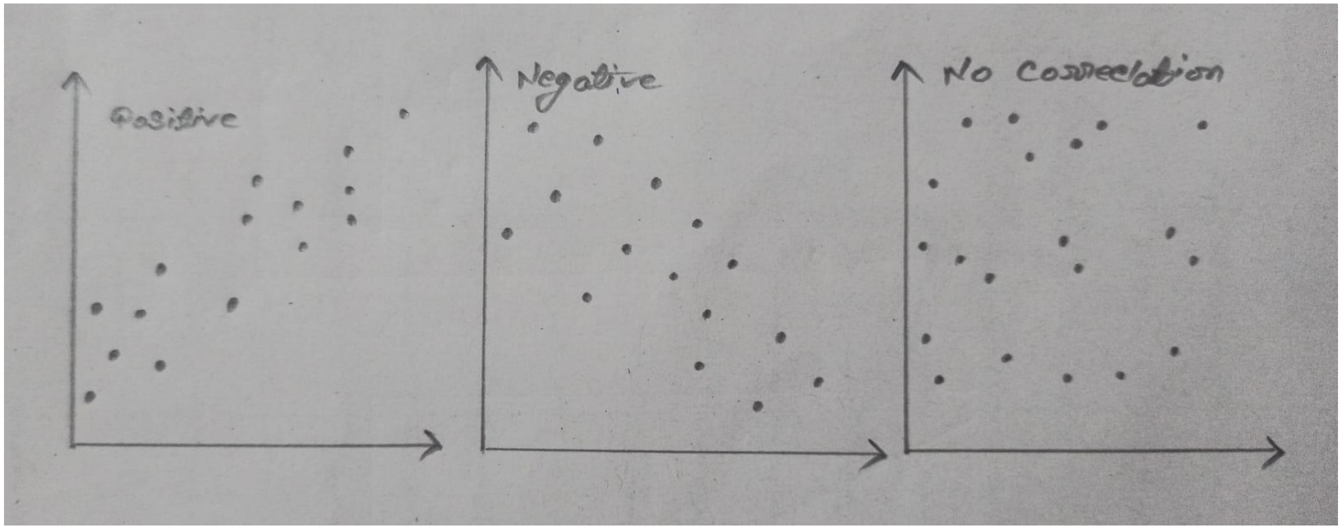
Answer: - Pearson's correlation coefficient, often denoted as "Pearson's R," is a statistical measure that quantifies the strength and direction of a linear relationship between two variables. It ranges from -1 to +1.

+1 indicates a perfect positive linear relationship,

-1 indicates a perfect negative linear relationship, and

0 indicates no linear relationship.

In simpler terms, if one variable tends to increase as the other increases, the correlation is positive. If one variable tends to decrease as the other increases, the correlation is negative. If there's no apparent trend, the correlation is close to zero. Pearson's R is widely used in fields like statistics, psychology, and economics to assess relationships between different sets of data**.**

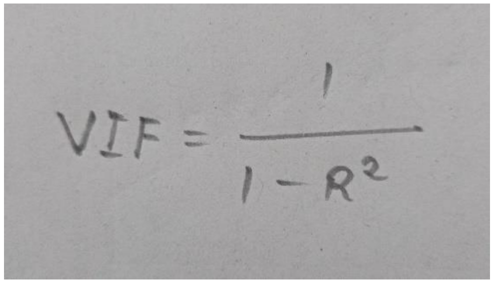


4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Answer: - Scaling is a preprocessing step in data where you transform the values of your variables to be within a specific range or follow a particular distribution.

S.NO	**Normalization Scaling**	**Standardized Scaling**
1.	**Minimum and Maximum value of different scale.**	**Mean and standard deviation is used for scaling.**
2.	**It is use when features are different scales.**	**It is used when we want to ensure zero mean and until standard deviation.**
3.	**Scale value 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. You might have observed that sometimes the value of VIF is infinite. Why does this happen?

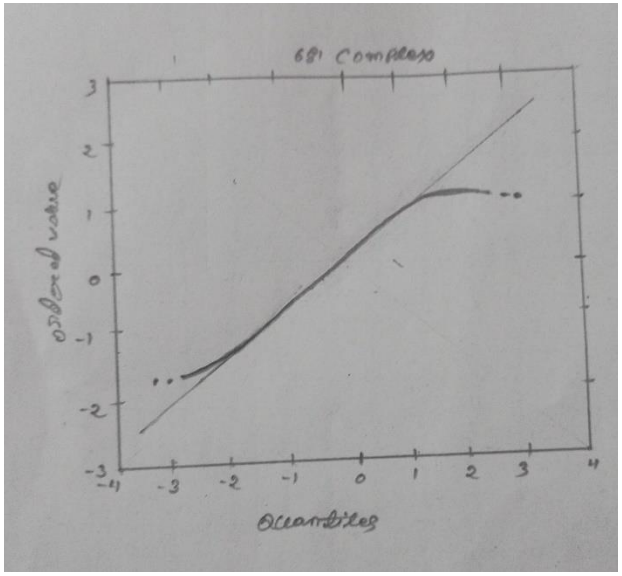


$$VIF = \frac{1}{1 - R^2}$$

Answer: - The VIF (Variance Inflation Factor) gives how much the variance of the coefficient estimate is being inflated by collinearity. If there is perfect correlation, then VIF = infinity. It gives a basic quantitative idea about how much the feature variables are correlated with each other. It is an extremely important parameter to test our linear model.

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

Answer: - The quantile-quantile (q-q) plot is a graphical technique for determining if two data sets come from populations with a common distribution.


681 Complexo
ordered value
Quantiles