1) From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?
Answer: Yr, Season, mnth, holiday and weathersit are correlated with dependent variable in the give order. But weekday has very less correlation with dependent variable.

2) Why is it important to use drop_first=True during dummy variable creation?
Answer: We drop first to avoide Multicollinearity. When we convert a categorical variable with n categories to dummy variables. It create n columns. But if we drop first column. The remaing n-1 columns can still give is inffo of deleted column if all other n-1 columns are 0 and deleted column is 1.

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

4) How did you validate the assumptions of Linear Regression after building the model on the training set?
Answer: By residual analysis of the train data. As errors in residual plot is normally distributed. That validates the assumptions of Linear Regression

5) Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?
Answer: temp, yr, light_snow

6) Explain the linear regression algorithm in detail.
Answer: Linear Regression is a supervised learning algorithm in machine learning, which is used for solving regression problems. Regression is a type of machine learning problem where the goal is to predict a continuous output variable based on one or more input variables.
In Linear Regression, the goal is to find the best-fitting linear equation to describe the relationship between the input variables (also known as predictors or features) and the output variable (also known as the response or target or dependent variable).

The equation for a simple linear regression model can be written as follows:
$y = b1 * x + b0$

Here, y is the dependent variable, x is the independent predictor or feature variable, b0 is a constant, and b1 is the slope coefficient (the change in y for a unit change in x).

The goal of Linear Regression is to find the best values for b0 and b1 such that the line best fits the data points, minimizing the errors or the difference between the predicted values and the actual values.

Types of linear regressions are as follows:
Simple Linear Regression: Equation with one predictor variable

$$y = b1 * x + b0$$

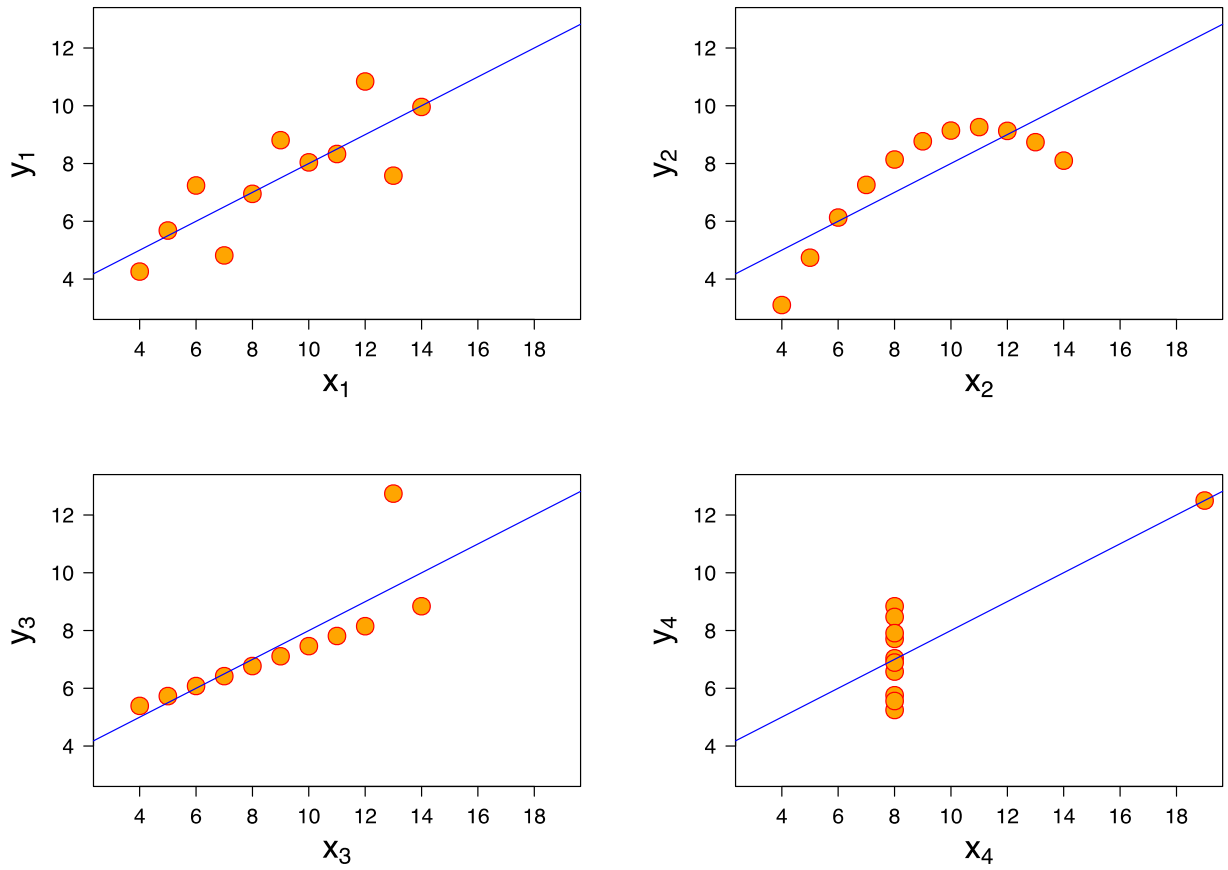
Multiple Linear Regression: Equation with more then one predictor variable

$$y = b0 + b1 * x1 + b2 * x2 + \ldots + bn * xn$$

7) Explain the Anscombe's quartet in detail.
Answer: Anscombe's quartet is a group of datasets (x, y) that have the same mean, standard deviation, and regression line, but which are qualitatively different.

The four datasets that make up Anscombe's quartet each include 11 x-y pairs of data. When plotted, each dataset seems to have a unique connection between x and y, with unique variability patterns and distinctive correlation strengths. Despite these variations, each dataset has the same summary statistics, such as the same x and y mean and variance, x and y correlation coefficient, and linear regression line.

Purpose of Anscombe's Quartet: Anscombe's quartet 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.



The four datasets composing Anscombe's quartet. All four sets have identical statistical parameters, but the graphs show them to be considerably different

8) What is Pearson's R?
Answer: Pearson's R is a measure of the linear relationship between two variables that have been measured on interval or ratio scales. It can only be used to measure the relationship between two variables which are both normally distributed. It is usually denoted by **r** and it can only take values between -1 and 1.

r = 1 : Perfect positive linear correlation
r = -1 : Perfect negative linear correlation
r = 0: No correlation
0 < r < 1 : Positive linear correlation increases as we move towards 1
-1 < r < 0 : Negative linear correlation increases as we move towards -1

Steps to calculting Pearson's R:

- We can check correlation by making scattered plot after removing the outliers. And by being able to see the distribution of your data you will get a good idea of the strength of correlation of your data before you calculate the correlation coefficient.
- Check if data is normally distributed and linearly correated.
- Finally you can calculate the correlation coefficient using the following formula:
    - $$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}},$$

- xi and yi are your data points,
- x and y cap are mean of x- values and y-values

9) What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?
Answer: Machine learning scaling is part of data preparation as this technique brings data points that are far from each other closer in order to increase the algorithm effectiveness and speed up the Machine Learning processing. Scaling data enables the model to learn and actually understand the problem.

normalized scaling: The data scales between min and max mostly 0 and 1.

standardized scaling: The data centers around mean of zero.

Examples: if we are preparing a model for property selling in an area as number of propety got sale per day will be in a few hundreds but the value of the transactions will be in millions. If we scale this data set both property count and price will come in same range. Like mean round 0(normalized scaling) or minmax (standardized scaling)

10) You might have observed that sometimes the value of VIF is infinite. Why does this happen?
Answer: When there is a perfect correlation in varibales

11) 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) plot is a graphical method for determining if a dataset follows a certain probability distribution or whether two samples of data came from the same population or not. Q-Q plots are particularly useful for assessing whether a dataset is normally distributed or if it follows some other known distribution. They are commonly used in statistics, data analysis, and quality control to check assumptions and identify departures from expected distributions.

In Linear regression, we can validate assumption of linear regression using Q-Q plot. By ploting graph on influnce.