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

We used box plots to explore the relationship between cnt and the categorical variables. Based on the resultant graphs, we note down the following observations.

1) Season: - Cnt does vary with seasons. Fall having the highest demand and spring resulting in the lowest demand.
2) Yr: - We see that the year 2019 saw a big increase in the demand for bikes.
3) Weathersit: - We observed that when the weather was *Clear, Few clouds, Partly cloudy, Partly cloudy* the demand was the maximum. Weather *Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds* with the minimum demand. *Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist* showed moderate demand.
4) Mnth, workingday, holiday, weekday did not show any observable patterns or variations with respect to demand.

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

Dummy variables are created for categorical variables. For each category of the categorical variable, only one of them can be 1 at a time and the rest would always be 0. Hence, for a variable with n categories if we know the values of n-1 levels the value of the nth level is easily derivable.

Consider a variable gender with 3 levels, Male, Female and Other.

If Male and Female are values 0 and 0 respectively, it would mean Other is 1.

If any one of Male or Female is 0 or 1 it would automatically mean Other is 0.

Hence, one of these levels can be dropped.

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

Temp variable which represents temperature of that particular day has the highest correlation with cnt.

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

The assumptions of Linear Regression that have been validated is explained below

a. Error terms should be normally distributed with mean 0. Once the predicted values of count were available. We plotted a seaborn distplot for the residuals that is the (actual y value – predicted y value) and we could see that the errors were normally distributed and centered at 0.

b. The error terms were independent of each other and there were no visible patterns that indicated any relationship between error terms.
c. Homoscedasticity. We also see that the error terms have constant variance all through in the graph. We used the same graph to validate all the above points.
d. Multi-Collinearity: - For interpretation we have validated the fact that the VIF's .
e. Linearity: - We used the stats model plot_ccpr to observe the linearity of each parameter selected by the model and we could successfully validate a reasonable linear relationship between most.

5) **Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?**
   a. Temperature
   b. Weather 3 - Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds
   c. Year

General Subjective Questions: -

1) **Explain the linear regression algorithm in detail.**

Linear regression in machine learning is a supervised learning algorithm used for predicting a continuous target variable based on one or more input features. It's a simple yet powerful technique that finds the best-fitting linear relationship between the input features and the target variable. Here's a detailed explanation of the linear regression algorithm in the context of machine learning:

Input: We have a dataset with a set of independent variables (features) represented as X1, X2, ..., Xn.

Output: The corresponding target variable we want to predict, denoted as Y.

Linear Model Hypothesis: Linear regression assumes that the relationship between the input features and the target variable can be represented by a linear equation:

$$Y = \beta0 + \beta1 * X1 + \beta2 * X2 + ... + \beta n * Xn + \varepsilon$$

Y is the predicted target variable.

β0 is the intercept (bias), representing the value of Y when all X's are 0.

β1, β2, ..., βn are the coefficients of the input features, indicating how much the predicted Y changes for a one-unit change in the corresponding X while keeping other features constant.

X1, X2, ..., Xn are the input features.

ε is the error term, accounting for the difference between the actual target values and the values predicted by the linear model.

Cost Function: The goal is to find the optimal values of β0, β1, β2, ..., βn that minimize the difference between the actual target values and the values predicted by the linear model. This is achieved by defining a cost function, typically the Mean Squared Error (MSE), which measures the average squared difference between the actual and predicted values.

Parameter Estimation: The process of finding the optimal values of β0, β1, β2, ..., βn that minimize the cost function is usually done using optimization techniques. The most common method is the Ordinary Least Squares (OLS) method, which involves minimizing the sum of the squared differences between the actual and predicted values.

Training the Model: The linear regression model is trained on a labeled dataset, meaning it learns the optimal values of the coefficients (β's) that best fit the training data. The training process involves finding the values of β0, β1, β2, ..., βn that minimize the chosen cost function.

Making Predictions: Once the model is trained, it can be used to make predictions on new, unseen data. To make a prediction, you plug the values of the input features into the linear equation, and the model calculates the predicted target variable (Y).

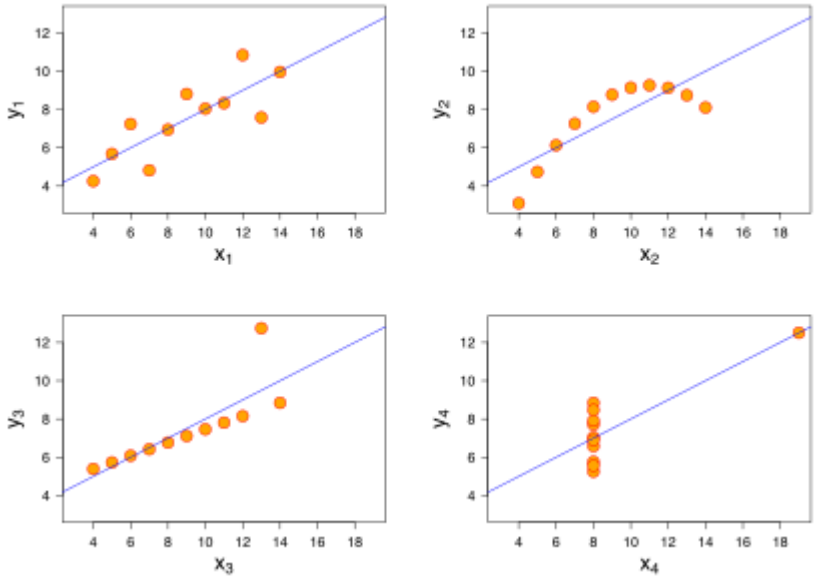
Evaluation: The performance of the linear regression model is evaluated using various metrics, such as R-squared (coefficient of determination), Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), etc. These metrics help assess how well the model fits the training data and how well it is expected to generalize to new data.

Assumptions and Considerations: Linear regression makes certain assumptions about the data, such as linearity, independence of errors, homoscedasticity (constant variance of errors), and normality of errors. It's important to check whether these assumptions are met before interpreting the results and using the model for predictions.

2) Explain the Anscombe's quartet in detail.

Anscombe's quartet is a fascinating statistical demonstration that illustrates the importance of visualization in data analysis. The quartet consists of four datasets that have nearly identical simple statistical properties, yet they have significantly different

characteristics when examined visually and in terms of more complex statistical properties. This demonstration underscores the limitations of relying solely on summary statistics without visualizing the data.



You will see that for all the 4 data sets plotted above, a lot of the properties such as mean for X and Y, variance for X and Y, linear regression coefficients fitted for Y, and the R squared value for best fit models on the data set. However, we still observe that these datasets are significantly different from each other and have their own unique distributions as shown in the figure above. This shows the power of visualization of data over raw descriptive statistics.

3) What is Pearson's R?

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

- +1 indicates a perfect positive linear correlation, meaning that as one variable increases, the other variable also increases proportionally.

- -1 indicates a perfect negative linear correlation, meaning that as one variable increases, the other variable decreases proportionally.

- 0 indicates no linear correlation between the variables; they are independent of each other.

Pearson's correlation coefficient is a measure of linear association and assumes that the relationship between the variables follows a linear pattern. It does not capture non-linear relationships.

The formula for Pearson's correlation coefficient (r) between two variables X and Y is:

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

Where:

- X and Y are the respective values of the two variables.

- X bar and Y bar are means of X and Y respectively.

- The summations are taken over all the data points in the dataset.

Interpreting Pearson's correlation coefficient:

- If r > 0, there is a positive correlation, meaning that as X increases, Y tends to increase.

- If r < 0, there is a negative correlation, meaning that as X increases, Y tends to decrease.

- The closer r is to +1 or -1, the stronger the correlation.

- If r is close to 0, there is little to no linear correlation between the variables.

It's important to note that correlation does not imply causation. Even if two variables have a strong correlation, it doesn't necessarily mean that changes in one variable cause changes in the other. Correlation only measures the strength and direction of the linear relationship between the variables.

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

Scaling is a technique used in machine learning to eliminate the effect of magnitudes or units of variables to cause bias in a machine learning algorithm. If feature scaling is not done then machine learning algorithms tend to provide higher weightage to independent variables whose values are higher in magnitude and lower weightage to variables with lower magnitude.

For example: -

Weight in kilograms for a person always has a higher range of values than height in feet for the person. This does not mean to say that weight in any way is more important to the model than height. But since the magnitudes of these variables vary significantly, the algorithm provides higher weightage to weight than height just because of their ranges. To eliminate this effect, we can either decide to either perform min max scaling (normalized scaling) or Standardized Scaling.

Normalized scaling: -

a) Values are computed based on the highest and lowest variable in the dataset. For example, min max scaling uses (x-min(x))/(max(x)-min(x)) to restrict the values of variables within 0 and 1.
b) Outlier values are also brought within the same range of 0 and 1.
c) No restriction that normalized values have to centered at 0.

Standardized scaling: -

a) Uses mean and standard deviation to compute standardized values. The formula is (X - mean)/(S.D). The values are not restricted to any specific range. The standardized value can vary with the mean and Standard deviation.
b) Outlier values are not restricted to 0 and 1.
c) The standardized values show a normal distribution with mean 0 and S.D 1.

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

The formula for VIF is 1/(1 – R squared). The value of VIF can be infinite only when the denominator becomes 0 and this happens when R squared is 1. R squared happens to be 1 when the model predicted can exactly predict all the values of the dependent variable. In the context of VIF, this means that excluding the target variable, for an independent variable, the other predictors in the dataset are able to perfectly predict it's value. This essentially means that there is multi collinearity in the system and the predictor with infinite VIF isn't going to add any value to the model and can be eliminated.

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

A Q-Q (Quantile-Quantile) plot is a graphical tool used to assess whether a given dataset follows a specific theoretical distribution, such as the normal (Gaussian) distribution. The Q-Q plot compares the quantiles of the dataset to the quantiles of the theoretical distribution. If the dataset follows the theoretical distribution, the points on the Q-Q plot will roughly lie along a straight line.

1. Sort the Data: First, you sort the data in your dataset in ascending order.

2. Compute Theoretical Quantiles: For each data point, you compute the theoretical quantile that it corresponds to in the theoretical distribution. These theoretical quantiles are calculated based on the chosen distribution (e.g., normal distribution) using a statistical function or a lookup table.

3. Plot the Data: You plot the sorted data points (observed quantiles) against the corresponding theoretical quantiles of the chosen distribution.

4. Interpret the Q-Q Plot:

- If the points on the Q-Q plot approximately form a straight line, it suggests that the dataset roughly follows the chosen theoretical distribution.

- If the points deviate from a straight line, it indicates that the dataset does not follow the chosen theoretical distribution.

The use and importance of a Q-Q plot in linear regression are as follows:

1. Assumption Checking: In linear regression, one of the key assumptions is that the errors (residuals) are normally distributed. By creating a Q-Q plot of the residuals you can visually assess whether the residuals follow a normal distribution. If the residuals deviate significantly from a straight line on the Q-Q plot, it suggests that the assumption of normality might be violated.

2. Detecting Outliers and Skewness: Q-Q plots can help you detect outliers and non-normality in your data. If the points on the Q-Q plot deviate dramatically from a straight line in the tails, it may indicate the presence of outliers or skewness in the data.

3. Model Evaluation Q-Q plots are a valuable tool for model evaluation and diagnostics. They can reveal patterns in the residuals that might not be apparent from summary statistics alone. Deviations from normality in the residuals can indicate areas where the model might be making systematic errors.