Assignment-based Subjective Questions

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

Answer 1:

Here are the inference points based on Bivariate and Multivariate:

Bivariate:

1. High bookings are seen in fall season
2. High bookings are seen in month of May, June, July, Aug, Sep,Oct
3. High bookings are seen in Clear weather situation
4. High bookings are seen on weekdays Thu, Fir, Sat, and Sun compared to other days.
5. High bookings are seen on Holidays.
6. High bookings are nearly equal on working day vs non-working day.
7. High bookings are seen in year 2019 vs 2018.

Multivariate:

1. High bookings are seen in year 2019 in season of summer and fall season.
2. High bookings are seen in fall on a working day and summer on a non-working day.
3. High bookings are seen on a holiday in summer and fall season.
4. High bookings are seen on a Clear weather situation in summer and fall season.

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

Answer 2:

- When creating dummy variables for categorical data, it is important to use drop_first=True because it helps in reducing the extra column created during dummy variable creation.

- Hence it reduces the correlations created among dummy variables. For example, if we have 3 types of values in a categorical column and we want to create a dummy variable for that column, if one variable is not furnished and semi-furnished, then it is obvious unfurnished.

- So we do not need the 3rd variable to identify the unfurnished. Hence if we have a categorical variable with n-levels, then we need to use n-1 columns to represent the dummy variable.

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

"temp" variable has the highest correlation with the target variable.

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

Validated the assumption of Linear Regression Model based on below 5 assumptions -

1. **Normality of error terms**: - Error terms should be normally distributed.
2. **Multicollinearity check**: - There should be insignificant multicollinearity among variables.
3. **Linear relationship validation**: - Linearity should be visible among variables.
4. **Homoscedasticity**: - There should be no visible pattern in residual values.
5. **Independence of residuals**: - No auto-correlation.

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

The below features contribute significantly:-

- Season: company should focus on Summer & Fall
- Weather: Users prefer the bikes when the weather is pleasant
- Temp: Users prefer bikes when the temperature is moderate
- Year: company should wait for pandemic to get over and things become normal

General Subjective Questions

Question 1. Explain the linear regression algorithm in detail.

Answer:

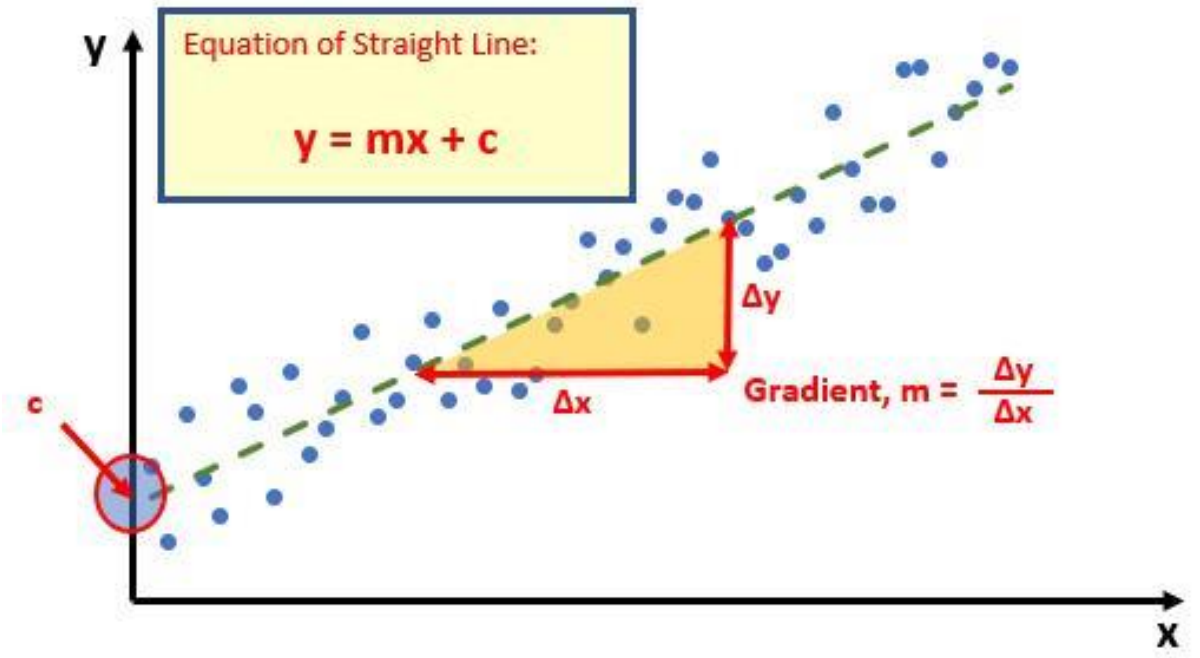
Linear regression is a machine learning algorithm used for predicting a continuous numeric output (also called the dependent variable) based on one or more input features (independent variables). It models the relationship between the input features and the output by fitting a linear equation to the observed data points. The primary goal is to find the best-fitting line that minimizes the difference between the predicted and actual values.

Linear regression relies on certain assumptions:

Linearity: The relationship between the features and the output is linear.
Independence: The residuals (differences between predicted and actual values) are independent and do not show patterns.
Homoscedasticity: The variance of the residuals is constant across all levels of the predictor variables.
Normality: The residuals are normally distributed.

Simple Linear Regression:

In simple linear regression, there's only one input feature (independent variable) and one output (dependent variable). The goal is to fit a line (equation) that best describes the linear relationship between the input and output.
The linear regression equation: y = mx + c



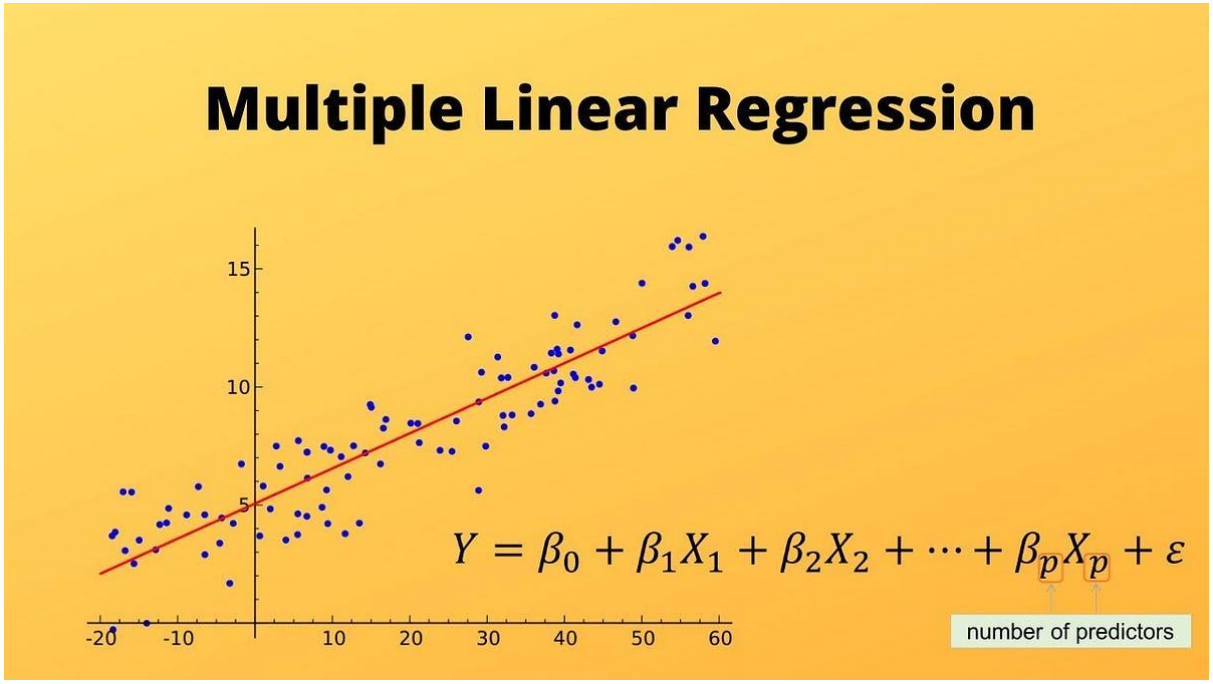
where y is the predicted output, m is the slope of the line, x is the input feature, and c is the y-intercept.

Multiple Linear Regression:

When there are multiple input features, it's called multiple linear regression. The goal remains the same: find the best-fitting hyperplane that represents the linear relationship between multiple input features and the output.

The linear regression equation becomes:

$y = \beta0 + \beta1X1 + \beta2X2 + ... + \betanXn$



where y is the predicted output, β0 is the intercept, β1, β2, ..., βn are the coefficients for each input feature X1, X2, ..., Xn.

Model Training:

The goal of training is to find the coefficients (slopes and intercept) that minimize the difference between the predicted values and the actual values in the training dataset. This is usually done by using a cost function like Mean Squared Error (MSE), which measures the average squared difference between predicted and actual values.

Gradient Descent:

Gradient Descent is an optimization technique used to find the optimal coefficients that minimize the cost function. It iteratively adjusts the coefficients based on the calculated gradients until convergence.

Model Evaluation:

After training, the model's performance is evaluated using evaluation metrics such as R-squared, Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), etc., on the validation or test dataset. These metrics help assess the model's accuracy and generalization to new data.

Predictions:
Once the model is trained and evaluated, it can be used to make predictions on new data by plugging in the input features into the regression equation.

Linear regression is a foundational algorithm and serves as a basis for more complex regression techniques. It's important to note that while linear regression is powerful for capturing linear relationships, it might not perform well when relationships are highly nonlinear or have interactions. In such cases, more advanced regression techniques might be required.

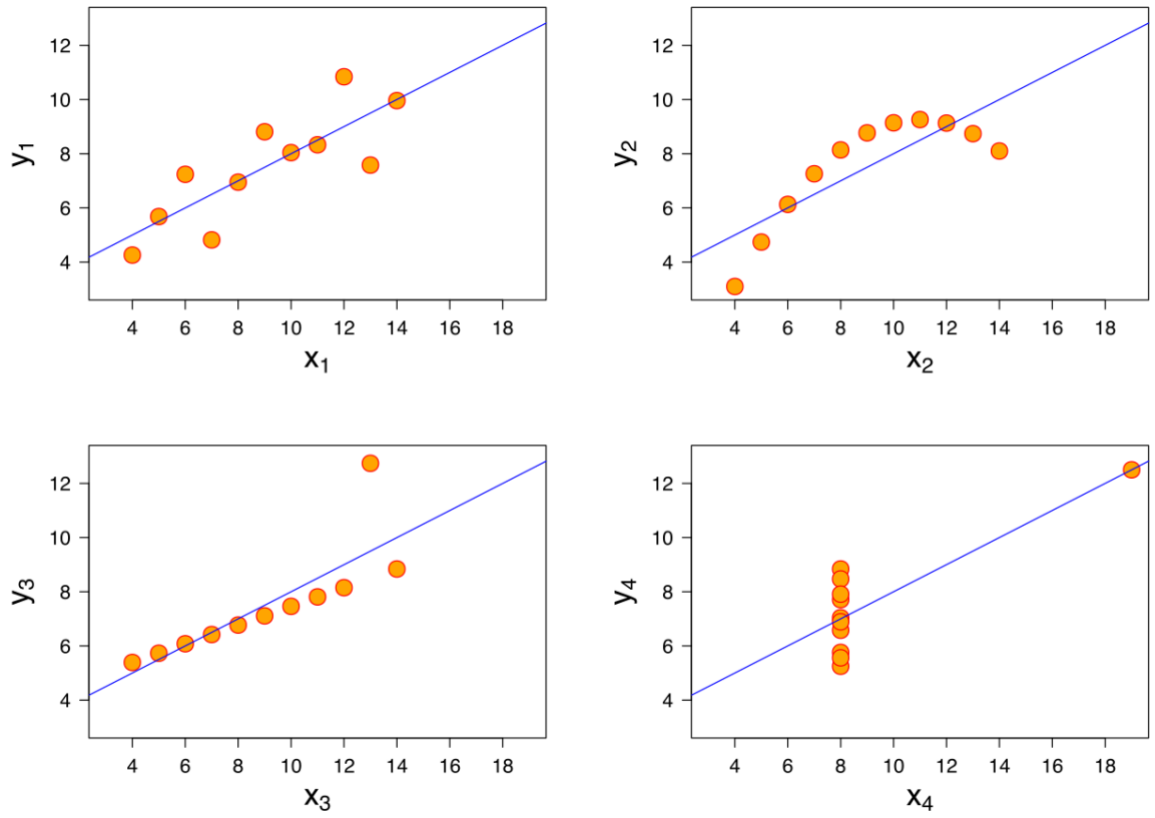
Question 2. Explain the Anscombe's quartet in detail.
Answer:

Anscombe's quartet is a set of four datasets that have identical or nearly identical statistical properties, but when graphed, they reveal very different patterns. This concept was introduced by the statistician Francis Anscombe in 1973 to highlight the importance of visualizing data and not solely relying on summary statistics. Anscombe's quartet serves as a cautionary example against blindly trusting summary statistics without visual inspection.

Here's a detailed explanation of Anscombe's quartet:

Dataset Characteristics:
The quartet consists of four small datasets, each containing 11 data points. These datasets have similar properties in terms of means, variances, correlations, and linear regression coefficients.

Properties of the Datasets:
Despite having similar statistical properties, the datasets are quite distinct in terms of their graphical representations. The datasets are numbered I, II, III, and IV.



Dataset I:

Linear relationship.
High correlation (approximately 0.816).
Simple linear regression model fits well.
Dataset II:

Nonlinear relationship.
High correlation (approximately 0.816).
Simple linear regression model doesn't fit well.
Dataset III:

Linear relationship.
Low correlation (approximately 0.12).
Outliers significantly impact the regression line.
Dataset IV:

Outlier significantly influences the regression model.
The regression model without the outlier is different from the one with the outlier.
Implications:

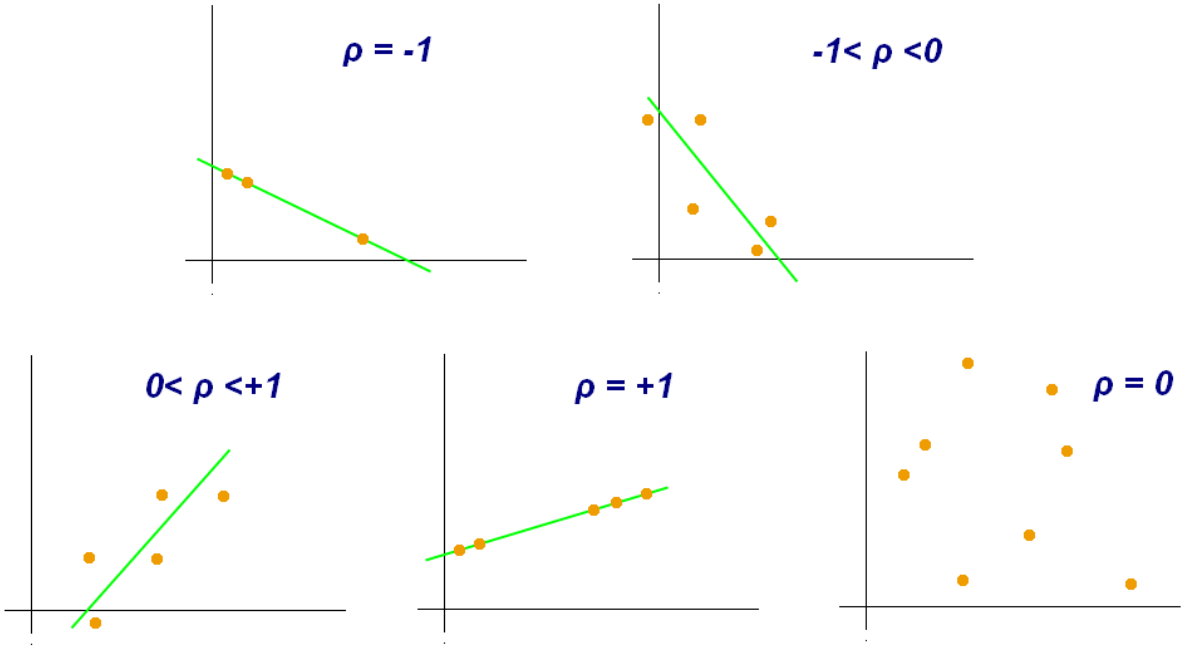
Anscombe's quartet emphasizes the importance of visualizing data. Relying solely on summary statistics like means, variances, and correlations can lead to incorrect conclusions. Graphical representations allow us to detect patterns, relationships, and outliers that might not be apparent from just looking at the numbers.

Lesson Learned:
Anscombe's quartet illustrates that summary statistics can't capture the full complexity of a dataset. It emphasizes the need for exploratory data analysis (EDA) through visualization to uncover insights, patterns, and potential issues.

By highlighting the differences between visually similar datasets, Anscombe's quartet demonstrates that visualization is a critical step in understanding data and making informed decisions based on it. It serves as a reminder for analysts, researchers, and data scientists to always visualize data before drawing conclusions or making predictions.

Question 3. What is Pearson's R?
Answer:
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 is used to determine how closely the values of the two variables move together.

Key characteristics of Pearson's correlation coefficient:



Range and Interpretation:
The value of Pearson's r ranges from -1 to +1:

+1: Perfect positive correlation (as one variable increases, the other also increases proportionally).
0: No linear correlation (variables are not linearly related).
-1: Perfect negative correlation (as one variable increases, the other decreases proportionally).
Calculation:
Pearson's correlation coefficient is calculated using the following formula:

$$r = (\Sigma((X - \bar{X})(Y - \bar{Y}))) / (\text{sqrt}(\Sigma(X - \bar{X})^2) * \text{sqrt}(\Sigma(Y - \bar{Y})^2))$$

where:

X and Y are the individual data points of the two variables.
$\bar{X}$ and $\bar{Y}$ are the means of the respective variables.
Strength of Relationship:
The absolute value of r indicates the strength of the relationship:

Close to +1 or -1: Strong linear relationship.
Close to 0: Weak or no linear relationship.
Direction of Relationship:
The sign of r (+ or -) indicates the direction of the linear relationship:

Positive r: As one variable increases, the other tends to increase.
Negative r: As one variable increases, the other tends to decrease.
Assumptions:
Pearson's correlation assumes:

Linearity: The relationship between variables is linear.
Homoscedasticity: The variance of the two variables is roughly constant across all levels.
Normality: The variables follow a normal distribution.
Use Cases:
Pearson's r is commonly used in data analysis to:

Assess the strength and direction of relationships between variables.
Identify potential predictor variables for regression analysis.
Examine associations in scientific research and social sciences.
It's important to note that Pearson's correlation coefficient specifically measures linear relationships. It may not capture nonlinear relationships, outliers, or other complex interactions. In such cases, alternative correlation measures or data transformation techniques might be more appropriate.

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

Answer:
Scaling is a pre-processing technique used in data analysis and machine learning to transform the features (variables) of a dataset so that they have similar scales or ranges. The goal of scaling is to bring the features to a common scale, which can improve the performance and effectiveness of various algorithms, especially those that rely on distance or magnitude-based calculations.

Why is Scaling Performed?
Scaling is performed for several reasons:

Algorithm Sensitivity: Many machine learning algorithms, like k-nearest neighbors (KNN) and gradient descent-based algorithms, are sensitive to the scale of features. Features with larger magnitudes can dominate the learning process.

Convergence Speed: Scaling can help gradient-based optimization algorithms converge more quickly, as features with similar scales lead to smoother loss surfaces.

Distance Metrics: Scaling ensures that distance-based metrics (e.g., Euclidean distance) are meaningful across all features, preventing one feature from dominating the distance calculation.

Model Interpretability: Scaling makes it easier to interpret the coefficients or importance of features in models like linear regression.

- **Normalized Scaling**:
- Normalized scaling (also known as Min-Max scaling) transforms features so that they are within a specific range, usually.
  The formula for normalized scaling is:
  X_scaled = (X - X_min) / (X_max - X_min)
  where X is the original feature value, X_min is the minimum value of the feature, and X_max is the maximum value of the feature.

- **Standardized Scaling:**

Standardized scaling (also called z-score scaling or standardization) transforms features to have a mean of 0 and a standard deviation of 1. The formula for standardized scaling is:

X_scaled = (X - X_mean) / X_std

where X is the original feature value, X_mean is the mean of the feature, and X_std is the standard deviation of the feature.

Difference Between Normalized and Standardized Scaling:

- **Normalized Scaling**: Transforms features to a specific range (usually [0, 1]).
- **Standardized Scaling**: Transforms features to have a mean of 0 and a standard deviation of 1.

Impact of Outliers:

- **Normalized Scaling**: Sensitive to outliers, as they can significantly affect the scale.
- **Standardized Scaling**: Less sensitive to outliers due to the use of mean and standard deviation.

Interpretation:

- **Normalized Scaling**: Preserves the original distribution, but doesn't handle outliers well.
- **Standardized Scaling**: Centers the distribution around 0 and is often preferred for algorithms that assume normal distribution.

Both scaling methods have their advantages and are chosen based on the characteristics of the dataset and the requirements of the algorithm being used.

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

Answer:

An infinite value of VIF for a given independent variable indicates that it can be perfectly predicted by other variables in the model.

This happens when R2 approaches 1. In other words, some variables are able to create perfect multiple regressions on other variables which would explain why all the VIF are infinity.

To identify them, one can try to do some actual regressions $Xj = X\backslash j\beta + \epsilon$ and check the coefficients in order to try to identify the problematic variables.

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

Answer: A Q-Q plot (quantile-quantile 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. It is a scatterplot created by plotting two sets of quantiles against one another. If both sets of quantiles came from the same distribution, we should see the points forming a line that's roughly straight.

In linear regression, Q-Q plots are used to check the normality assumption of residuals. The use and importance of Q-Q plots in linear regression is that it helps us to check whether the residuals are normally distributed or not. If the residuals are normally distributed, then it

means that the model is correctly specified and we can trust the results of our linear regression model. If the residuals are not normally distributed, then it means that there is something wrong with our model and we need to investigate further.