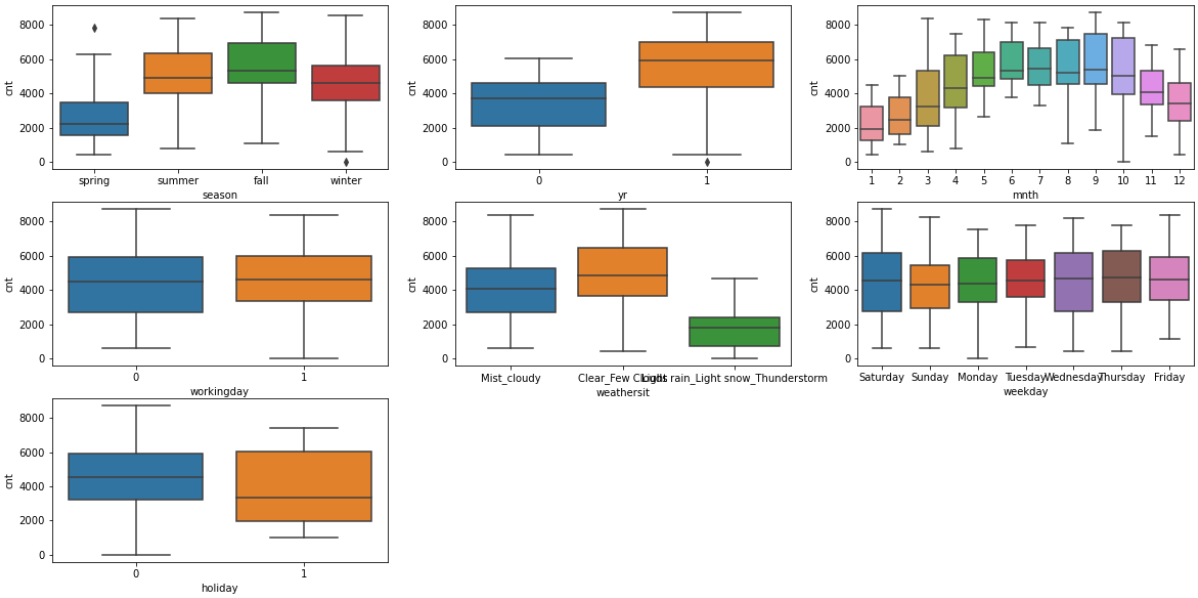
Assignment-based Subjective Questions

1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)

Answer:

There about six categorical variable in the dataset namely season,mnth,yr,weekday, working day and weathersit.



These categorical variable has a major effect on the variable cnt. Demand for bikes is the highest in the months of August , September & October. Demand for bike is visibly low during Spring Season There was a considerable increment in user base from 2018 to 2019

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

Answer:

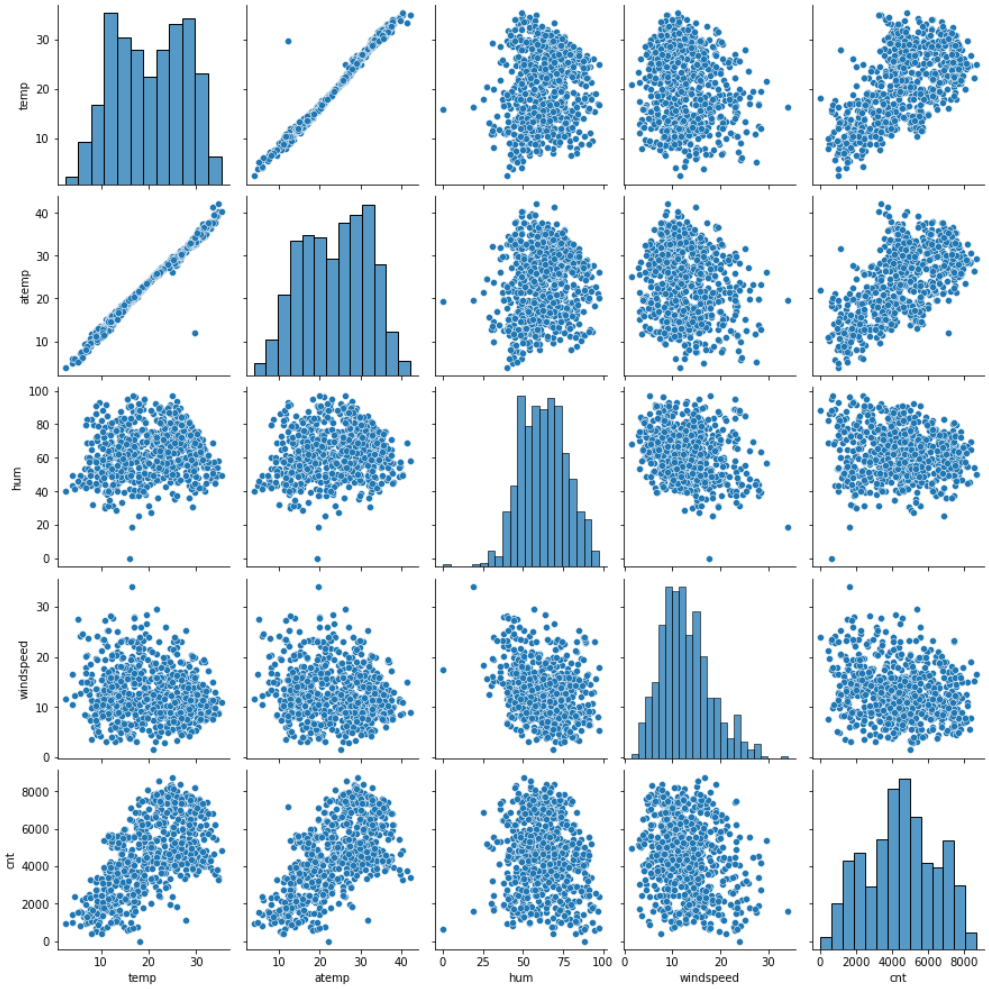
Dummy Variables is used to enable numerical representation of categorical variables for use in statistical models that require numerical data like regression. On dummy variable creation category becomes a new binary variable (0 or 1) indicating its presence or absence. 'drop_first=True:' is used for preventing Perfect Collinearity (Dummy Variable Trap)If you create a dummy variable for every category, they'll be perfectly correlated. This is problematic for statistical models like linear regression.

Example: With categories "Red", "Green", and "Blue", knowing a data point is not "Red" and not "Green" automatically means it's "Blue".
drop_first=True prevents this by removing one dummy variable, making the remaining ones independent.

Drop first function reduces redundancy so that it can infer the missing category's value from the others.
Example: If "Red" and "Green" are 0, the model knows it's "Blue".

Drop first also reduces the number of variables, improving model efficiency and interpretability.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)



The ‘temp’ and ‘atemp’ variables have highest correlation when compared to the rest with target variable as ‘cnt’ as per pairplot.

4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

1) Linearity: This means that there is a straight-line relationship between the independent and dependent variables. You can validate this assumption by plotting the data and looking for a linear pattern. You can also use statistical tests, such as the F-test, to determine whether the relationship is statistically significant.
2) Homoscedasticity: This means that the variance of the errors is constant across all levels of the independent variables. You can validate this assumption by plotting the residuals (the difference between the predicted and actual values) against the independent variables. If there is a pattern in the residuals, it suggests that homoscedasticity is not met.
3) Normality: This means that the errors are normally distributed. You can validate this assumption by looking at a histogram of the residuals. If the histogram is roughly bell-shaped, it suggests that normality is met.
4) Independence: This means that the errors are independent of each other. You can validate this assumption by looking at a plot of the residuals over time. If there is a pattern in the residuals, it suggests that independence is not met.

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

Top 3 features that has significant impact towards explaining the demand of the shared bikes are temperature, year and season.

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)

It's a statistical method for modeling the relationship between two variables: a dependent variable (what you want to predict) and an independent variable (what you use to make the prediction). Goal of the model is to find the best-fitting line (or hyperplane in higher dimensions) that represents the relationship between these variables.

Steps in Linear Regression:

- Collect Data: Gather a dataset containing both the dependent and independent variables.
- Visualize Data: Plot the data to check for a linear relationship.
- Choose a Model: Select either simple linear regression (one independent variable) or multiple linear regression (multiple independent variables).
- Calculate Parameters: Use a method like Ordinary Least Squares (OLS) to estimate the coefficients (slope) and intercept of the line of best fit. OLS minimizes the sum of squared residuals.
- Make Predictions: Use the fitted model to predict values of the dependent variable for new values of the independent variable(s).
- Evaluate Performance: Assess the model's accuracy using metrics like R-squared, Mean Squared Error (MSE), and Root Mean Squared Error (RMSE).

Key Assumptions:

Linearity: The relationship between the variables is linear.
Independence: The observations are independent of each other.
Homoscedasticity: The variance of the residuals is constant.
Normality: The residuals are normally distributed.

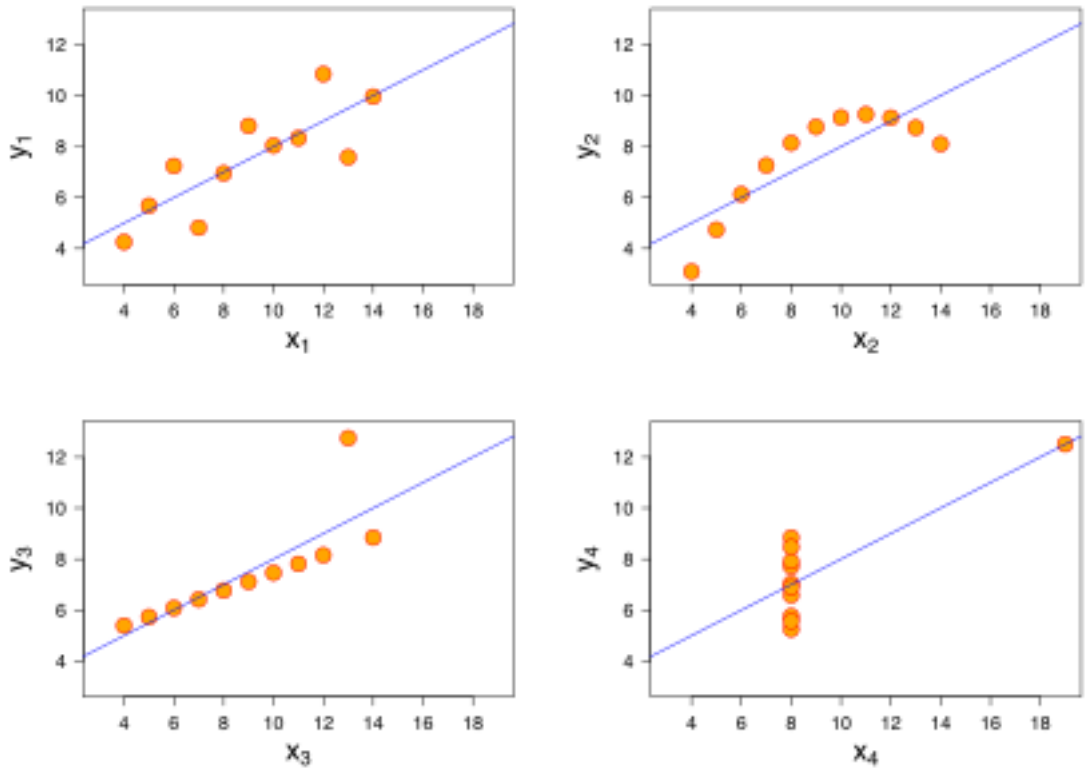
Key Concepts:
Line of Best Fit: The model's output, representing the linear relationship between the variables.
Slope (Coefficient): Indicates how much the dependent variable changes for each unit change in the independent variable.
Intercept: The value of the dependent variable when the independent variable is zero.
Residuals: The differences between the actual data points and the predicted values on the line of best fit.

Common Applications:

Predicting sales based on advertising spending
Estimating house prices based on square footage and location

2. Explain the Anscombe’s quartet in detail. (3 marks)

Anscombe's quartet is a set of four data sets that have nearly identical descriptive statistics (mean, variance, correlation coefficient, etc.) but look completely different when plotted. Statistician Francis Anscombe created it in 1973 to highlight the importance of visualizing data before analysing it.



Despite their identical summary statistics, the four data sets look very different when plotted. This shows that descriptive statistics can be misleading and that it is important to always visualize data before drawing conclusions from it.

Anscombe's quartet is a reminder that data analysis is not just about numbers. It is also about understanding the context of the data and using common sense to interpret the results.

3. What is Pearson's R? (3 marks)

Pearson's R, also known as the Pearson correlation coefficient, is a statistical measure that quantifies the linear relationship between two continuous variables. It ranges from -1 to 1, with:

Pearson's R value 1 indicates a perfect positive linear relationship: As one variable increases, the other increases proportionally. Imagine a straight line with a positive slope. Value -1 indicates a perfect negative linear relationship: As one variable increases, the other decreases proportionally. Value 0 indicates no linear relationship: There is no predictable relationship between the two variables. Values closer to 1 or -1 suggest a stronger linear relationship, either positive or negative. Values closer to 0 suggest a weaker or no linear relationship.

Pearson's R only measures linear relationships. It cannot detect non-linear relationships, such as curves or U-shapes. Pearson's R does not imply causation. Just because two variables are correlated does not mean that one causes the other. Pearson's R is sensitive to outliers. A single outlier can significantly affect the value of R.

The formula for Pearson's R is:

$r = (\sum(xi - \bar{x})(yi - \bar{y})) / \sqrt{(\sum(xi - \bar{x})^2)}\sqrt{(\sum(yi - \bar{y})^2)}$
where:

xi and yi are the individual data points for variables X and Y, respectively.
$\bar{x}$ and $\bar{y}$ are the mean values of X and Y, respectively.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

Scaling is a data preprocessing technique that involves transforming the values of features (independent variables) in a dataset to a common scale, typically within a specified range.

Scaling performed for the following:
Reduces Bias: Prevents features with larger ranges from dominating those with smaller ranges, leading to biased models.
Enhances Interpretability: Makes model coefficients more comparable, aiding interpretation of feature importance.
Enhances Accuracy: Many algorithms, especially distance-based ones (e.g., k-nearest neighbors, SVMs) are sensitive to feature magnitudes. Scaling ensures features contribute equally to the model's decision-making.

Difference between normalized scaling and standardized scaling:

Normalization (Min-Max Scaling)	Standardization (Z-Score Normalization)
Transforms features to a range between 0 and 1	Subtracts mean and divides by standard deviation, resulting in a distribution with a mean of 0 and a standard deviation of 1.
Formula: X_scaled = (X - X_min) / (X_max - X_min)	Formula: X_scaled = (X - X_mean) / X_std
Ideal for algorithms sensitive to outliers, like neural networks.	More robust to outliers than normalization. Often preferred for algorithms assuming a Gaussian distribution of features, like linear regression.
Range 0 to 1	No fixed range, but typically around 0
Sensitive to outliers	More robust to outliers
Preserves original distribution shape	Transforms to standard normal distribution

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

A VIF value of infinity can occur in two main scenarios:

1. Perfect multicollinearity:

This happens when two or more independent variables in your regression model are perfectly correlated with each other. When this occurs, the regression model tries to estimate the coefficient for each of the collinear variables, but the information is redundant and leads to an unstable and undefined estimate. This instability translates to an infinite VIF value.
2. Insufficient data points:
If you have fewer data points than independent variables in your model, VIF values can become unreliable, including reaching infinity. This is because the regression model doesn't have enough information to accurately estimate the relationships between the variables, leading to inflated variances and consequently, infinite VIF values.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

A Q-Q plot, short for Quantile-Quantile plot, is a graphical tool used to compare the distribution of a sample against a theoretical reference distribution. It does this by plotting the quantiles of both distributions against each other.

Importance in Linear Regression:

In linear regression, a Q-Q plot is primarily used to visually assess whether the residuals (the difference between predicted and actual values) follow the assumed normal distribution. This assumption is crucial for many aspects of linear regression analysis, including:

1.Validating model assumptions: Normality is one of the key assumptions of linear regression. If the residuals are not normally distributed, it can potentially lead to biased estimates and unreliable statistical tests.

2. Diagnosing model problems: Deviations from a straight line in the Q-Q plot can indicate specific issues, such as non-linearity, heteroscedasticity (unequal variance of errors), or outliers.

3.Interpreting results: Knowing if the residuals are normally distributed helps us interpret the p-values of our hypothesis tests and confidence intervals more accurately.