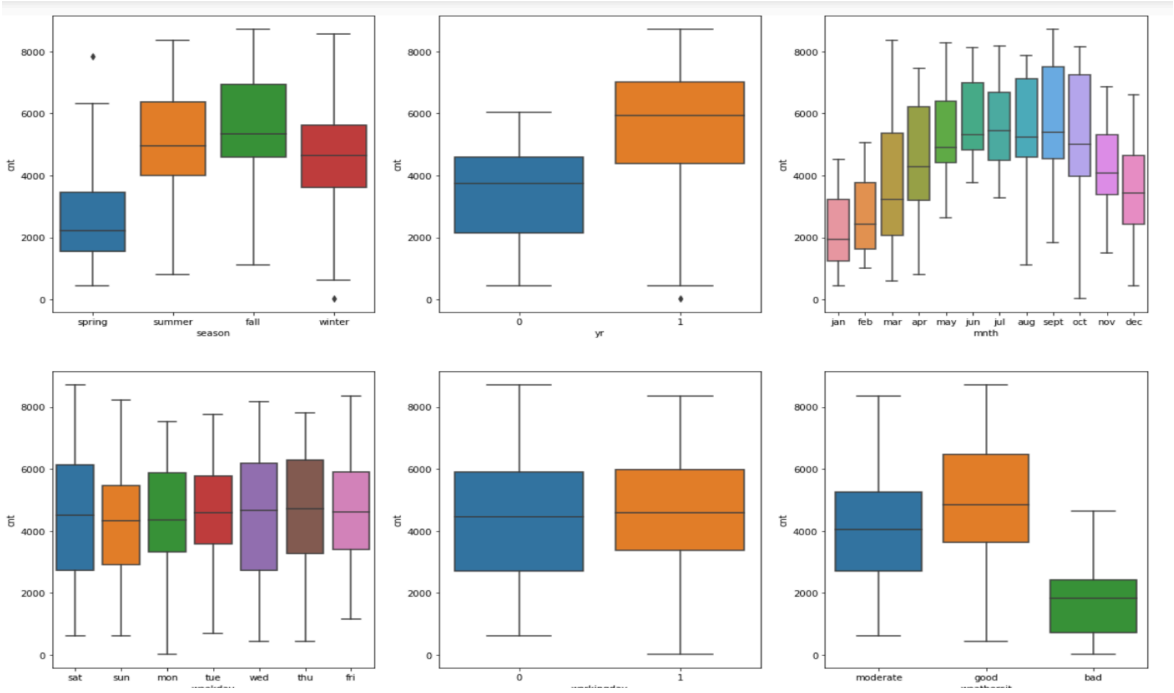
Assignment-Based Subjective Questions

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

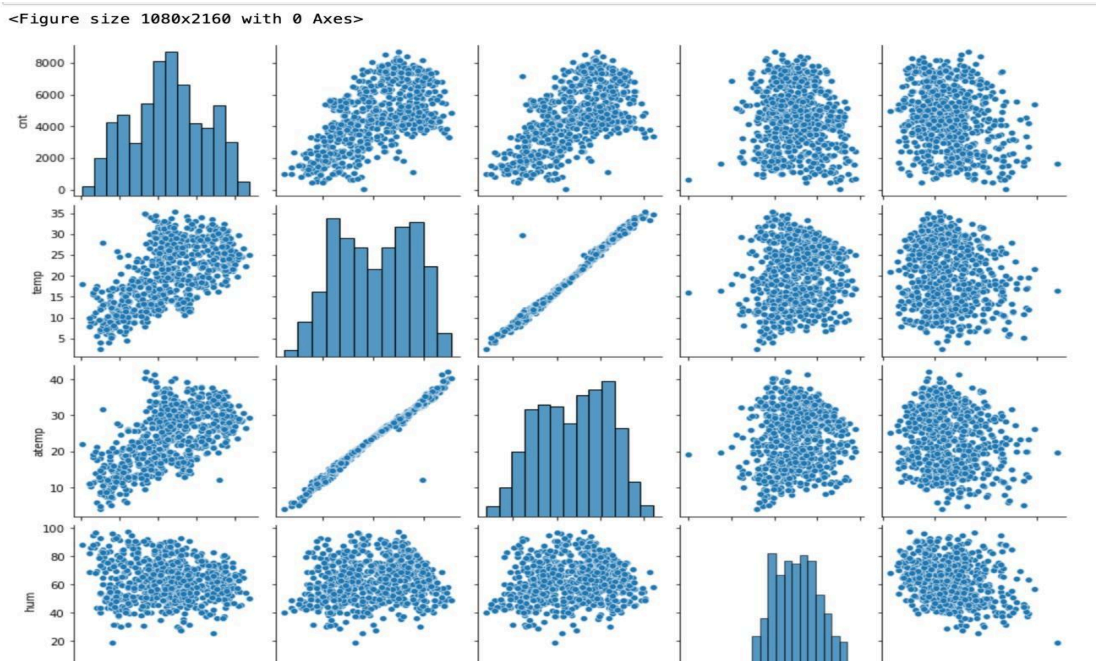
ANSWER: These categorical variables—season, month, year, weekday, working day, and weather situation—significantly impact the dependent variable 'cnt'. The correlation among these variables is visualized using both bar plots and box plots.



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

When creating dummy variables for a categorical variable with 'n' levels, the approach is to generate 'n-1' new columns. Each column signifies the presence (1) or absence (0) of a specific category level. The parameter `drop_first=True` is utilized to align the resulting dummy variables with 'n-1' levels, effectively reducing correlation among them. For instance, if there are 3 levels, `drop_first=True` removes the first column, ensuring independence among the remaining columns representing each category level.

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



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

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

Linear regression models are evaluated for Linearity, Absence of autocorrelation, Normality of errors, Homoscedasticity, and Multicollinearity.

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

   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 ?**

Linear regression is a predictive modeling technique that establishes the relationship between a dependent variable (target) and independent variables (predictors). It determines how the value of the dependent variable changes with variations in the independent variables. When there is only one input variable (x), it is termed as simple linear regression, while multiple input variables constitute multiple linear regression.

The essence of linear regression lies in finding a straight line that best describes the relationship among the variables. This line can depict either a positive or negative linear relationship. The objective of the linear regression algorithm is to optimize the coefficients (a0 and a1) to derive the line that minimizes error.

In linear regression, methods like Recursive Feature Elimination (RFE), Mean Squared Error (MSE), or cost functions aid in determining the optimal values for a0 and a1. These values ensure that the line fitted to the data points is the best possible fit, minimizing the discrepancy between predicted and actual values.

2. Explain the Anscombe's quartet in detail?

Anscombe's Quartet consists of four datasets that share nearly identical simple descriptive statistics, such as mean and variance for both the x and y variables. However, each dataset exhibits unique characteristics that can mislead a regression model if not properly understood and visualized.

1. **First Dataset**: This dataset shows a clear linear relationship between the x and y variables, making it suitable for linear regression modeling. It adheres to the assumptions of linear regression, where the relationship can be adequately captured by a straight line.
2. **Second Dataset**: Unlike the first dataset, the second dataset does not display a linear relationship between x and y. It demonstrates the importance of not assuming linearity without visual inspection, as a regression model may not accurately represent the data.
3. **Third Dataset**: In this dataset, outliers are present, which can significantly impact the performance of a linear regression model. Outliers can skew the regression line and affect the model's predictive accuracy, highlighting the need for robustness checks in model building.
4. **Fourth Dataset**: The fourth dataset includes a high leverage point, which exerts considerable influence on the regression model's results. This point can strongly affect the correlation coefficient and regression line, underscoring the importance of understanding influential observations in the dataset.

Conclusion: Anscombe's Quartet serves as a compelling demonstration of why data visualization is essential before constructing and analyzing regression models. By visually inspecting the data, analysts can identify non-linear relationships, outliers, influential points, and other nuances that may impact model assumptions and performance. This practice ensures that regression algorithms are applied appropriately and yield reliable insights from the data. Therefore, always visualize and explore data thoroughly before proceeding with machine learning model development.

3. What is Pearson's R?

In statistics, Pearson's correlation coefficient is also known as Pearson's r, Pearson product-moment correlation coefficient (PPMCC), or bivariate correlation. It is a measure of the linear relationship or association between two variables.

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

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

Scaling is a crucial data preprocessing step in machine learning to ensure that all features contribute equally to the model fitting process, regardless of their original magnitudes, units, or ranges. Here's an explanation of the differences between Normalizing Scaling (MinMax Scaling) and Standardizing Scaling (Z-score Scaling):

Normalizing Scaling (MinMax Scaling):

- **Method**: Scales the values to a specific range, typically between 0 and 1 (or -1 and 1).
- **Scaling Formula**: Xscaled=X−XminXmax−XminX_{scaled} = \frac{X - X_{min}}{X_{max} - X_{min}}Xscaled=Xmax−XminX−Xmin
- **Use Case**: Useful when features have varying scales and you want to constrain them within a specific range.
- **Effect on Distribution**: Retains the original distribution shape but squashes or stretches it to fit the defined range.
- **Outlier Sensitivity**: Highly sensitive to outliers because it directly uses minimum and maximum values.
- **Application**: Often used in algorithms that require inputs to be on a bounded interval, like neural networks and algorithms that use distance measures (e.g., K-means clustering).

Standardizing Scaling (Z-score Scaling):

- **Method**: Transforms the data to have a mean of 0 and a standard deviation of 1.
- **Scaling Formula**: Xscaled=X−μσX_{scaled} = \frac{X - \mu}{\sigma}Xscaled=σX−μ, where μ\muμ is the mean and σ\sigmaσ is the standard deviation of the feature.
- **Use Case**: Ensures features have the same scale and are comparable. Useful when features follow a normal distribution or when the algorithm assumes zero-centered data.
- **Effect on Distribution**: Centers the distribution around 0 and adjusts the spread of the data.
- **Outlier Sensitivity**: Less sensitive to outliers compared to MinMax scaling because it uses the mean and standard deviation, which are less affected by outliers.
- **Application**: Commonly used in algorithms that rely on distance metrics (e.g., SVM,

KNN) or require normally distributed data (e.g., linear regression, logistic regression).

Summary:

- **Purpose**: Normalized scaling adjusts values to a specified range (0-1 or -1 to 1), while standardized scaling adjusts values to have zero mean and unit variance.
- **Distribution**: Normalized scaling preserves the distribution shape, while standardized scaling centers the distribution around zero.
- **Outliers**: Normalized scaling is sensitive to outliers due to its reliance on minimum and maximum values, whereas standardized scaling is more robust to outliers.
- **Use Cases**: Choose normalized scaling when the range of values is crucial, and standardized scaling when consistency in scale and distribution is needed.

In practice, the choice between these scaling methods depends on the characteristics of your data and the requirements of your machine learning algorithm.

1.

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

VIF(VarianceInflationFactor) basically helps explain the relationship of one independent variable with all the other independent variables. The formulation of VIF is given below:
A VIF value of greater than 10 is definitely high, a VIF of greater than 5 should also not be ignored and inspected appropriately.
A very high VIF value shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity. To solve this problem we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

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

Q–Q plot is a probability plot, which is a graphical method for comparing two probability distributions by plotting their quantiles against each other.
Quantile-Quantile (Q-Q) plot, is a graphical tool to help us assess if a set of data possibly came from some theoretical distribution such as a Normal, exponential or Uniform distribution.
QQ plot can also be used to determine whether or not two distributions are similar or not. If they are quite similar you can expect the QQ plot to be more linear. The linearity assumption can best be tested with scatter plots. Secondly, the linear regression analysis requires all variables to be multivariate normal. This assumption can best be checked with a histogram or a Q-Q-Plot.

Importance of QQ Plot in Linear Regression :
In Linear Regression when we have a train and test dataset then we can create Q-Q plot by which we can confirm that both the data train and test data set are from the population with the same distribution or not.
Advantages:

- It can be used with sample size also
- Many distributional aspects like shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot

Q-Q plot use on two datasets to check

- If both datasets came from population with common distribution
- If both datasets have common location and common scale
- If both datasets have similar type of distribution shape
- If both datasets have tail behavior