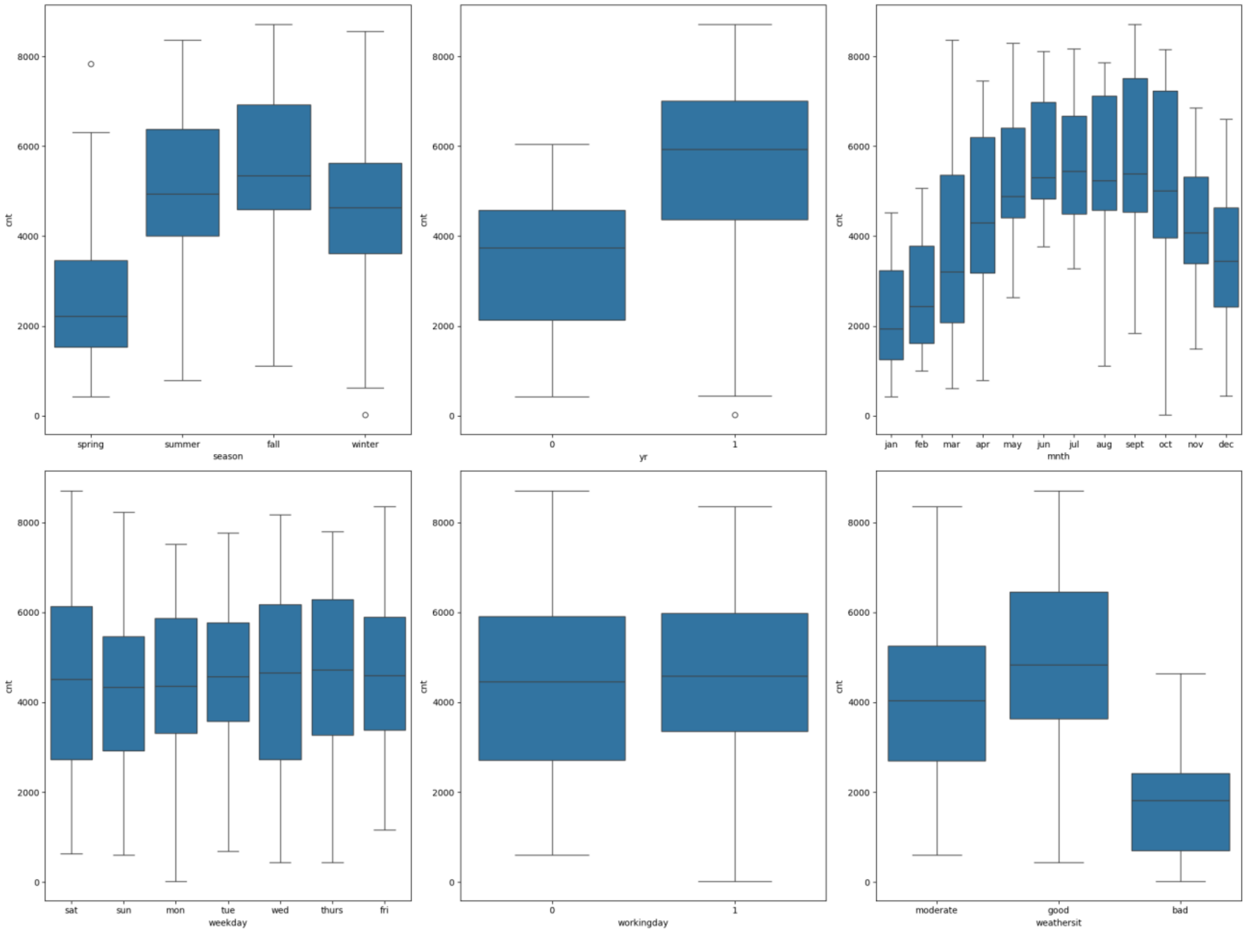
1) From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

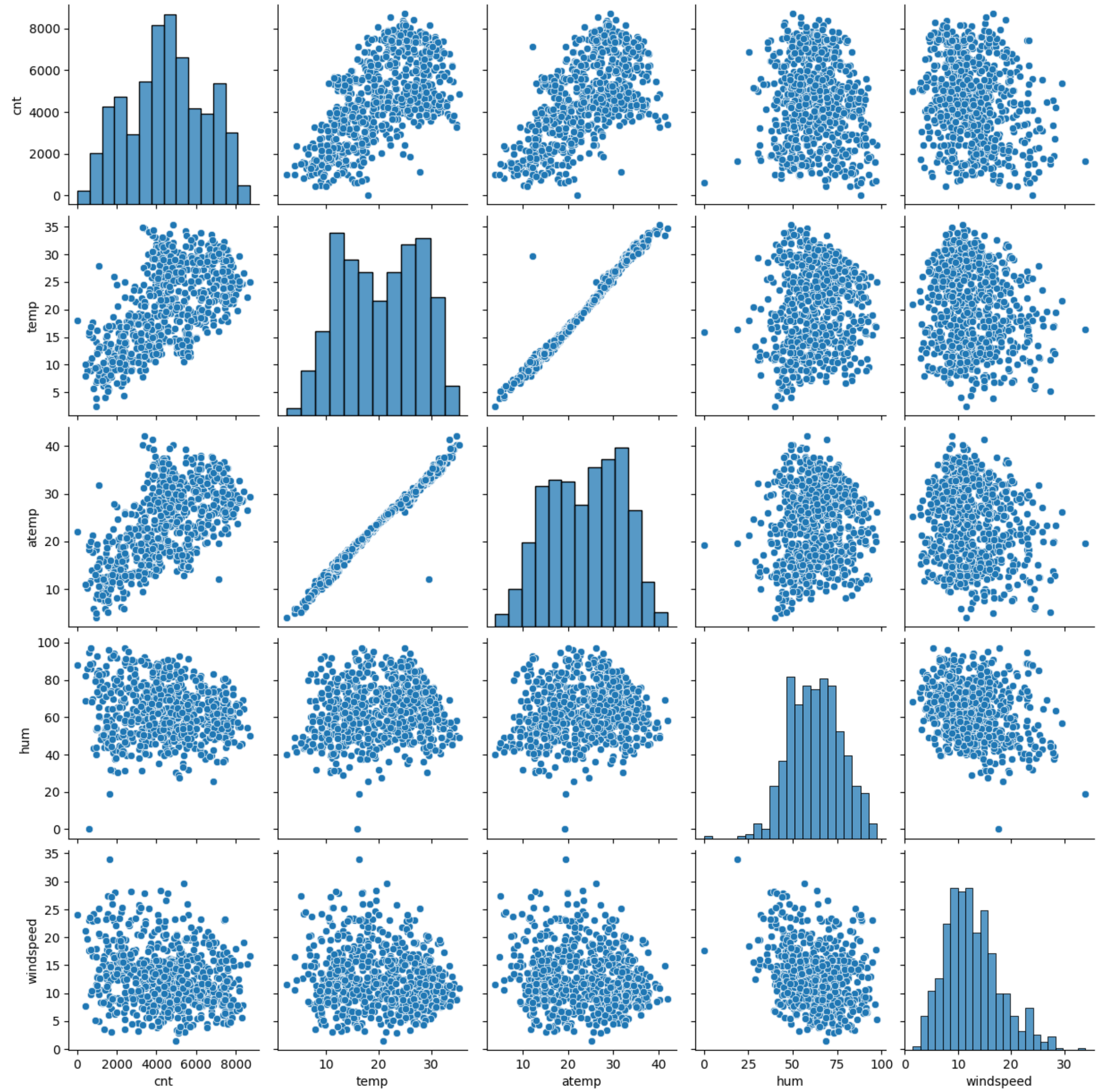
Ans: A few categorical variables are season, month, year, weekday, working day, and weather condition. The dependent variable "cnt" is significantly impacted by these category variables. The figure below displays the association between the same.



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

Ans: drop_first=True should be used since it minimizes the extra column that is produced while creating dummy variables. By doing this, multicollinearity—the formation of correlations between dummy variables—will be less likely to occur.

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



When compared to the other variables, the variables "temp" and "atemp" exhibit the highest correlation, with the target variable being "cnt."

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

Ans: Validation criteria for linear regression models

- include linearity,
- No autocorrelation,
- error normality,
- homoscedasticity,
- multicollinearity.

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

Ans: Temperature, year, and season

General Subjective Questions

1) Explain the linear regression algorithm in detail.

Ans: One type of predictive modeling technique that shows us the link between the independent variables (predictors) and the dependent variable (targeted variable) is called linear regression. Given that linear regression illustrates a linear relationship, it determines how the dependent variable's value changes in response to the independent variable's value. Such linear regression is referred to as simple linear regression if there is only one input variable (x). Additionally, this type of linear regression is known as multiple linear regression if there are many input variables. The relationship between the variables is described as a slanted straight line by the linear regression model.
Either a positive or negative linear relationship can be represented by a regression line. The linear regression algorithm's objective is to obtain the best

2) Explain the Anscombe's quartet in detail.

Ans**:** Anscombe's quartet consists of four datasets with virtually similar simple descriptive statistics that, when graphed, show significant differences in distribution and appearance. Every dataset has eleven (x, y) points in it. Francis Anscombe, a statistician, developed it in 1973 to highlight the significance of charting data prior to analysis.

The following are emphasized by the quartet:

- Visualizations are crucial. Statistics that are descriptive may be deceptive. Plotting datasets with comparable statistical characteristics can produce remarkably distinct patterns.
- Outliers can be very influential: Significant distortion of statistical traits and relationships can be caused by outliers or influential observations.
- Not every situation lends itself to linearity: Not all relationships between variables are linear, despite what summary statistics may imply.

Anscombe's quartet highlights the importance of visually exploring data.

3) What is Pearson's R?

Ans: Pearson's R, or Pearson correlation coefficient, is a statistical measure that quantifies the strength and direction of the linear relationship between two continuous variables.

$$r = \Sigma[(xi - \bar{x})(yi - \bar{y})] / \sqrt{[\Sigma(xi - \bar{x})^2 * \Sigma(yi - \bar{y})^2]}$$

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

Ans: it’s a important machine learning technique used in data preprocessing technique to adjust the rang of data

Scaling is performed to ensure the uniformity and improve the algorithm performance, speed up the Convergence.

Normalized vs Standardized scaling .

- Normalized scaling - min-max scaling: its is use full when we need to bound the data within the specific range (typically (0,1) or (-1,1))

- Standardized Scaling (Z-score Normalization): used standardize the data for algorithms for example like linear regression or PCA , the transform data to have a mean of 0 and standard of deviation 1

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

Ans: Variance Inflation Factor (VIF) helps explain the relationship of one independent variable with all the other independent variables. The formulation of VIF is given below:

- A VIF value greater than 10 is definitely high.
- A VIF value greater than 5 should also not be ignored and should be inspected appropriately.

A very high VIF value indicates a perfect correlation between two independent variables. In the case of perfect correlation, we get (R^2 = 1), which leads to (\frac{1}{1 - R^2}) becoming infinite. To solve this problem, we need to drop one of the variables from the dataset that 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.

Ans: A Q–Q plot (Quantile-Quantile plot) is a probability plot, which is a graphical method for comparing two probability distributions by plotting their quantiles against each other. It helps assess if a set of data possibly came from a theoretical distribution such as Normal, Exponential, or Uniform.

Uses of Q-Q Plot:

- Assess Distribution Similarity: Determines whether two distributions are similar. If they are, the Q-Q plot will be more linear.

- Test Linearity Assumption: The linearity assumption can be tested with scatter plots. Linear regression analysis requires all variables to be multivariate normal, which can be checked with a histogram or a Q-Q plot.

Importance in Linear Regression:

- Train and Test Datasets: In linear regression, a Q-Q plot can confirm if both the train and test datasets come from populations with the same distribution.

Advantages:

- Sample Size: Can be used with small sample sizes.
- Detect Distributional Aspects: Shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected.

Q-Q Plot Uses:

- Common Distribution: Check if both datasets came from a population with a common distribution.
- Common Location and Scale: Verify if both datasets have common location and scale.
- Similar Distribution Shape: Determine if both datasets have a similar type of distribution shape.
- Tail Behavior: Assess if both datasets have similar tail behavior.