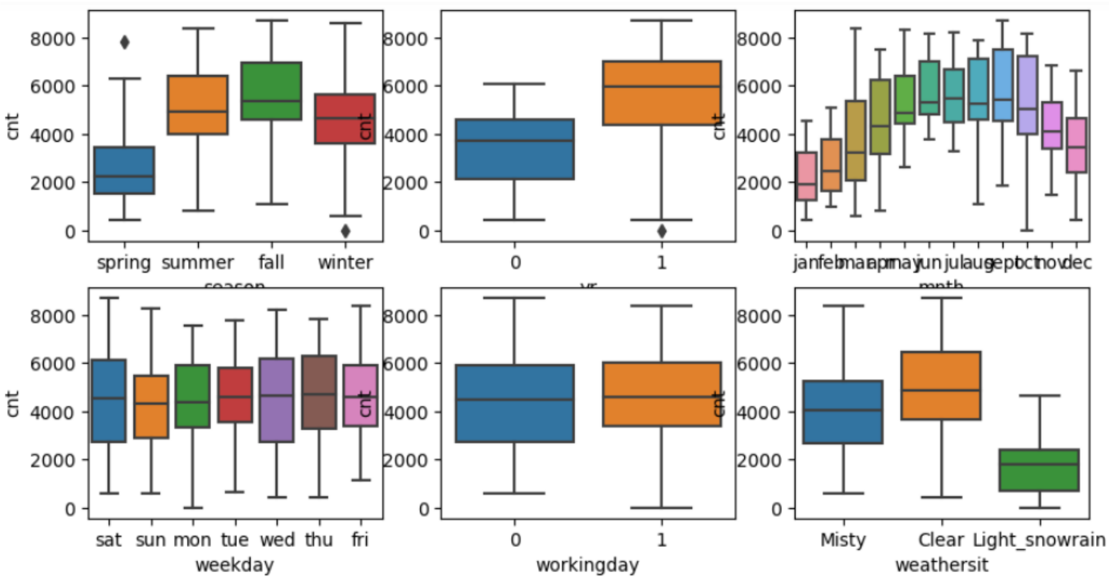
Assignment-based Subjective Questions

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

These categorical variables—season, month, year, weekday, working day, and weather situation—significantly impact the dependent variable 'cnt'.

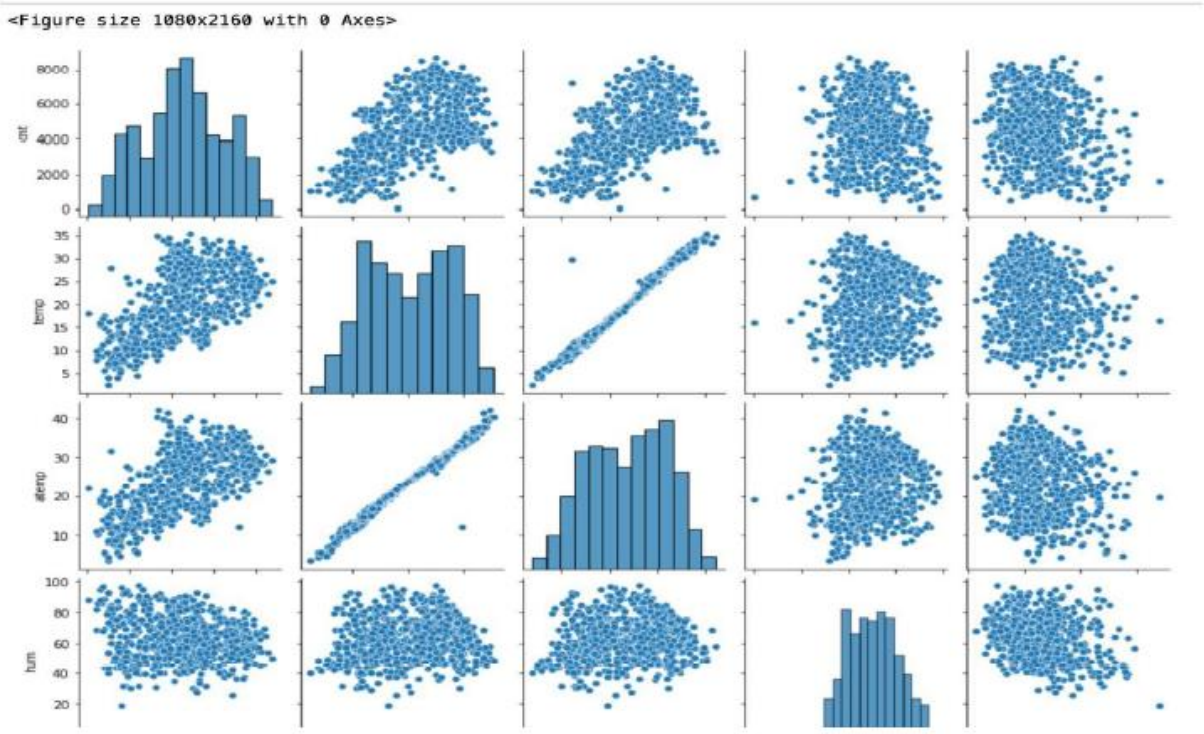


2) **Why is it important to use drop_first=True during dummy variable creation?**

Using drop_first=True during dummy variable creation is important to avoid multi-collinearity, which occurs when one predictor variable in a model can be linearly predicted from others with a high degree of accuracy. In the context of categorical variables, if all dummy variables are included, they are perfectly collinear (they add up to one). Dropping the first dummy variable removes this redundancy and ensures that the model is not over parameterized, leading to more stable and interpretable coefficient estimates.

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’.



```
<Figure size 1080x2160 with 0 Axes>
```

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 Multi-collinearity.

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 fundamental statistical and machine learning technique used to understand the relationship between a dependent variable (often denoted as yyy) and one or more independent variables (often denoted as xxx). It assumes that there is a linear relationship between the independent variables and the dependent variable.

The goal of linear regression is to find the values of β0\beta_0β0 and β1\beta_1β1 that minimize the sum of squared residuals (the difference between actual y values and predicted y values). Two types – Simple and Multiple linear regression.

It is easy to implement and interpret. Works well with linearly separable data. Provides insights into relationships between variables. Linear regression serves as a foundational technique in data analysis, predictive modelling, and understanding the influence of variables on outcomes.

2) Explain the Anscombe’s quartet in detail.

Anscombe's quartet is a collection of four datasets that have nearly identical simple descriptive statistics (mean, variance, correlation, etc.) but exhibit very different distributions and graphical properties. The quartet was created by the British statistician Francis Anscombe in 1973 to illustrate the importance of graphical analysis of data and the limitations of relying solely on summary statistics.

Dataset 1 appears to form a linear relationship with a slight amount of noise and fits well to a linear regression line.

Dataset 2 appears to form a curve, not a straight line and linear regression line does not fit well.

Dataset 3 appears to contain an outlier that affects the relationship and without the outlier, the data might fit a linear relationship better.

In Dataset 4, most of the data points are identical, except for one extreme outlier. The outlier heavily influences the regression line, which does not represent the bulk of the data.
Anscombe’s quartet demonstrates that statistical analysis should always be accompanied by graphical representations to understand the true nature of the data.

3) What is Pearson’s R?

Pearson's r, often referred to as Pearson correlation coefficient or Pearson's correlation, is a measure of the linear correlation between two variables X and Y. It quantifies the strength and direction of the linear relationship between the variables, ranging from -1 to +1:

r=+1r = +1r=+1: Perfect positive linear relationship.

r=0r = 0r=0: No linear relationship (variables are not linearly correlated).

r=−1r = -1r=−1: Perfect negative linear relationship (one variable increases as the other decreases).

Pearson's correlation coefficient is widely used in statistics, data analysis, and machine learning for its simplicity and effectiveness in quantifying linear relationships between variables.

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

Scaling in the context of data pre-processing refers to the process of adjusting the range or distribution of data. It is done to ensure that variables are comparable and have similar scales.

Scaling is performed for,

- Comparable Units: Different variables often have different units (e.g., kilograms vs. meters), and scaling ensures that these variables are on a similar scale, making them comparable.
- Algorithm Performance: Many machine learning algorithms perform better or converge faster when the input features are on a similar scale. Algorithms like k-nearest neighbours (KNN), support vector machines (SVM), and gradient descent-based optimization methods (used in linear regression, neural networks, etc.) are particularly sensitive to the scale of input variables.
- Interpretability: Scaling can aid in the interpretability of coefficients or feature importance in models. It prevents variables with larger ranges from dominating or skewing results.

Difference between normalized and standardized scaling are,

- Range: Normalized scaling transforms data to a specific range (typically [0, 1]), while standardized scaling transforms data to have zero mean and unit variance.
- Effect on Distribution: Normalized scaling preserves the original distribution shape, whereas standardized scaling may change the distribution to be more Gaussian.
- Application: Normalized scaling is suitable when you know the distribution does not follow a normal distribution or when you want to preserve the scale of outliers. Standardized scaling is useful when the algorithm assumes normally distributed data or when interpreting feature importance.

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

The phenomenon where the value of VIF (Variance Inflation Factor) becomes infinite typically occurs due to perfect multicollinearity among the predictor variables in a regression model. Variance Inflation Factor (VIF) measures how much the variance of a regression coefficient is inflated due to multicollinearity among the predictors. It quantifies how much the variance of the estimated regression coefficients are increased compared to what would be expected if the predictors were uncorrelated.

infinite VIF values occur when there is perfect multicollinearity among predictor variables in a regression model. This situation arises due to linear dependencies among predictors, making it impossible to calculate VIF values as per the standard formula. Identifying and addressing multicollinearity is crucial for ensuring the reliability and interpretability of regression models.

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

A Q-Q plot (Quantile-Quantile plot) is a graphical tool used to assess whether a given dataset follows a particular theoretical distribution, such as the normal distribution. It compares the quantiles of the dataset against the quantiles of the theoretical distribution being tested.

Use and Importance in Linear Regression:

- Normality Assumption:

  - Linear regression models often assume that the residuals (the differences between observed and predicted values) are normally distributed.
  - A Q-Q plot of the residuals can help verify this assumption by comparing the distribution of residuals against a normal distribution.
  - If the residuals closely follow a straight line on the Q-Q plot, it suggests that the normality assumption holds.

- Identifying Outliers and Skewness:
  - Q-Q plots can also reveal outliers and skewness in the data.
  - Outliers appear as points far from the straight line, indicating potential errors or unusual observations in the data.
  - Skewness can cause the data points to deviate from the straight line in a particular direction, indicating a departure from normality.
- Model Fit and Validity:
  - A well-fitting linear regression model should have residuals that are normally distributed and centered around zero.
  - Q-Q plots provide a visual confirmation of whether the residuals meet these expectations.

Misalignment or significant departures from the straight line in the Q-Q plot may indicate issues with the model's assumptions or data quality.

Q-Q plots are valuable tools in linear regression for assessing the normality of residuals and verifying model assumptions. They provide a clear visual indication of how well the data conform to theoretical distributions, helping analysts make informed decisions about model validity and data quality.