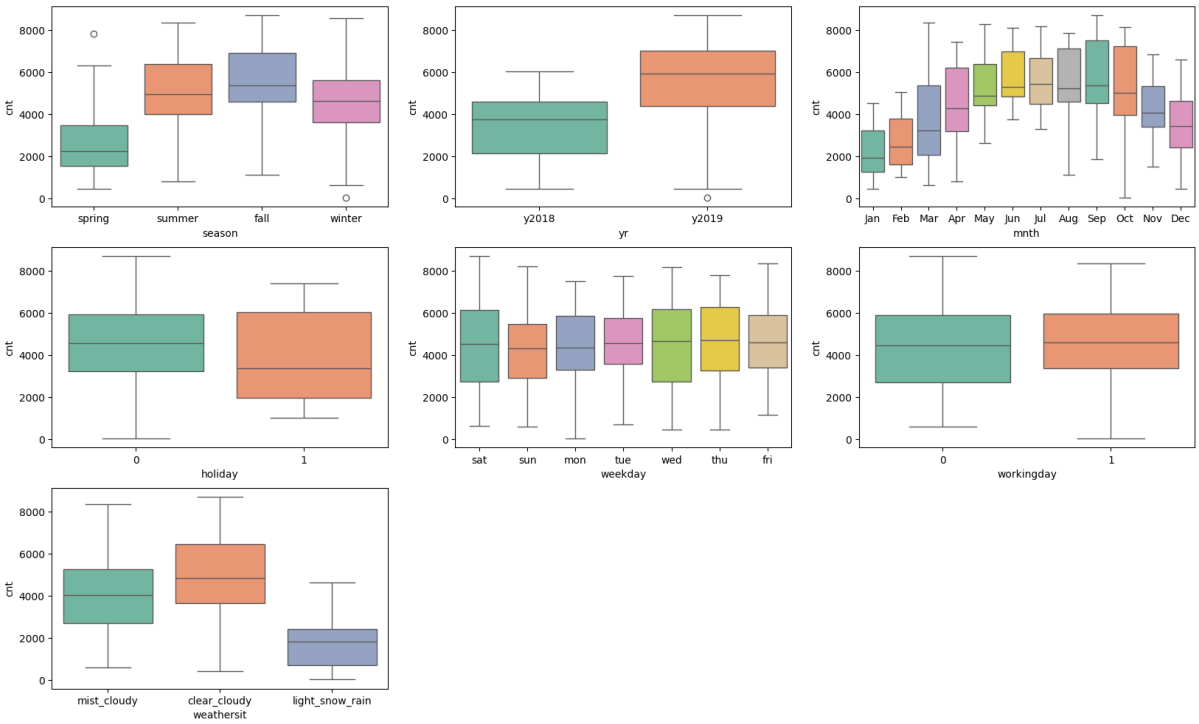
Assignment-based Subjective Questions

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

Answer: The dependent variable for this assignment is **cnt.** Following are the important inferences that were derived using Exploratory Data Analysis (EDA)

- Business is less in `spring` season. It meanes **`spring` is negatively correlated**.
- Business **improved almost *50%*** in the year 2019. **Median in 2018 is ~4000 and 2019 is ~6000.**
- Demand is high in the middle of the year. The months **`Jun`,`Jul`,`Aug` and `Sep`** are positively correlated.
- Demand is almost ***30%* high on non-holidays** compared to the demand on holidays.
- Business is stable withrespect to **`weekday`** and **`workingday`** variables. Meaning, they have minimal or **no correlation**.
- Looks like people do not prefer bikes during rain. The demand is **significantly less during Light Snow/Rain** and **no business during Heavy Rains**.

The boxplots below were created and analysed to make above inferences.



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

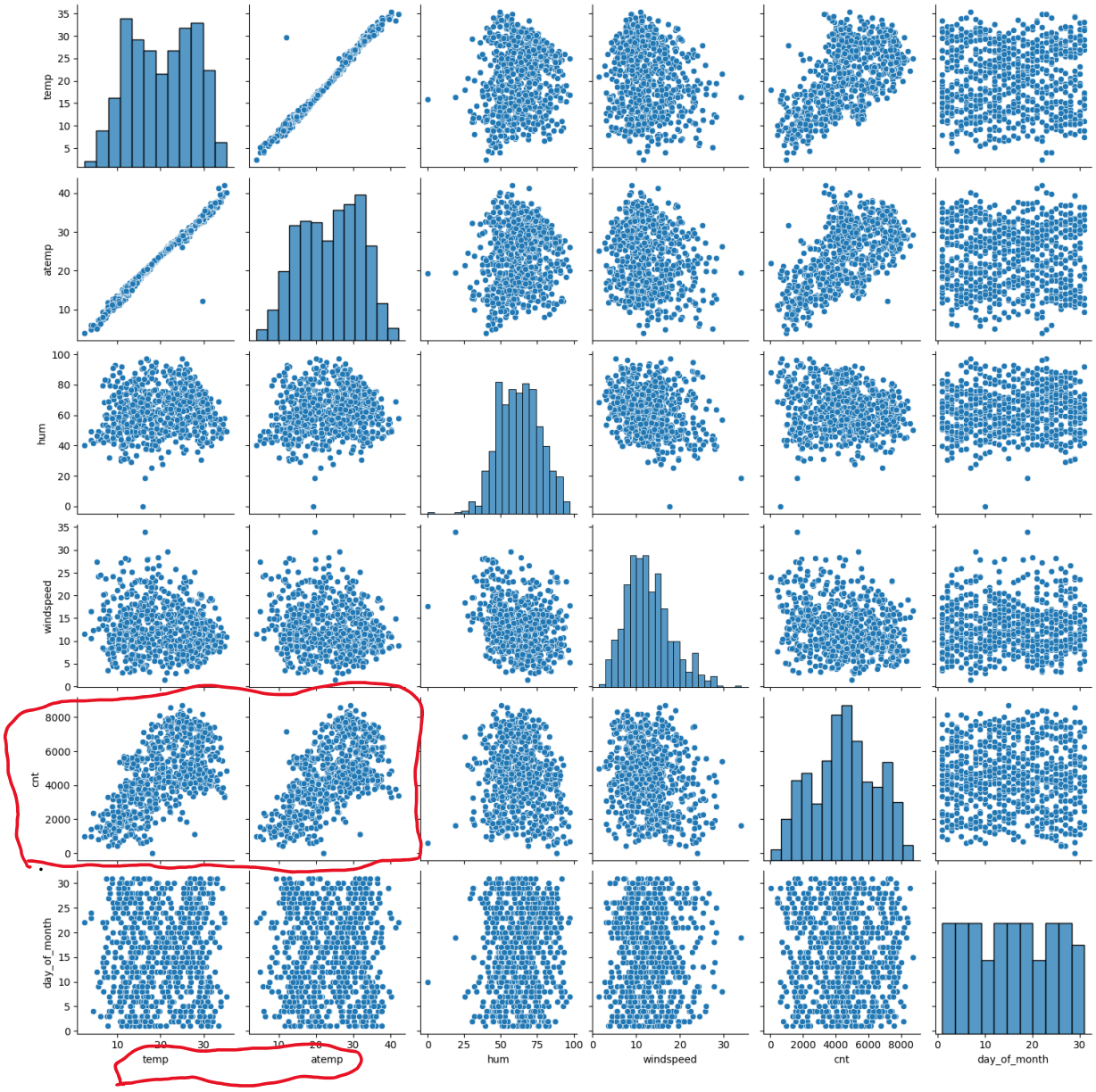
Answer: When creating dummy variables (also known as one-hot encoding) in statistical modelling, the parameter **drop_first=True** plays a crucial role in preventing a common statistical issue known as **multicollinearity**.

When we convert a categorical variable with N categories into dummy variables without dropping a category, we end up with N dummy variables. However, keeping all N dummy variables **introduces multicollinearity** because the sum of these N dummy variables will always equal 1 (each observation falls into one category). This perfect multicollinearity means one variable can be predicted perfectly from the others.

By using **drop_first=True**, one category is dropped (usually the first based on alphabetical order in case of character data), and it will **eliminate the multicollinearity** problem.

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

Answer: The variables **temp** and **atemp** have the highest correlation with the target variable **cnt.**



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

Answer: Below are the assumptions of Linear Regression

1. **Linearity**: The scatter plots show the linearity between the independent variables and dependent variable
2. **Normal Distribution of Errors**: Histogram was plotted for the residuals / errors, and it represents the normal distribution
3. **Independence of errors:** The errors are observed using the scatter plot with the target variable **cnt** and noticed no patterns.
4. **Homoscedasticity:** A scatter plot is created between predicted values and residuals. No variance is observed.

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

Answer: The top 3 contributing features are:

1. **temp:** Temperature in Celsius – **0.568**
2. **y2019:** Year (0: 2018, 1:2019) – **0.233**
3. **winter:** Winter season – **0.126**

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Answer: Linear regression is a fundamental statistical and machine learning technique used to predict a dependent variable (target) based on the values of independent variable(s). It is used across a vast range of industries from economics to biology, and is extremely valuable because of its simplicity and interpretability. Here's a detailed explanation of the linear regression algorithm:

Linear regression models the relationship between a scalar dependent variable (y) and one or more independent variables (or predictors) denoted (X). The case of one independent variable is called simple linear regression; for more than one, the process is called multiple linear regression.

The linear regression model can be represented by the following equation:

$$y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \varepsilon$$

Where:

- y_i is the dependent variable.
- X1, X2, …, X_p are the independent variables.

- $\beta_1, \beta_2, ..., \beta_p$ are the coefficients which the algorithm will estimate to predict (y).
- β_0 is the intercept.
- ε is the error term

2. Explain the Anscombe's quartet in detail.

Answer: Anscombe's quartet comprises four different datasets that have nearly identical simple statistical properties, yet appear very different when graphed. Each dataset consists of eleven points (x, y) and was constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance and limitations of statistical summary properties (like the mean, variance, and correlation) and to emphasize the value of graphing data before analysing it.

Key Statistical Properties

The intriguing aspect of Anscombe's quartet is that all four datasets provide nearly the same statistical properties, which often leads to similar analytical summaries from a purely numerical perspective. Specifically:

- The mean of the x values is approximately 9 for all datasets.
- The mean of the y values is approximately 7.50 for all datasets.
- The variance of x is approximately 11 for all datasets.
- The variance of y is approximately 4.12 for all datasets.
- The correlation between x and y is approximately 0.816 for all datasets.
- The regression line for predicting y from x is (y = 3.00 + 0.500x) for all datasets.

3. What is Pearson's R?

Answer: Pearson's R, also known as the Pearson correlation coefficient (PPMCC), is a measure of the linear correlation between two variables X and Y. It provides a value between -1 and +1 inclusive, where:

- +1 indicates a perfect positive linear relationship,
- -1 indicates a perfect negative linear relationship,
- 0 indicates no linear correlation between the variables.

Formula

Pearson's R is calculated as follows:

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

Where:

- r = correlation coefficient
- x_i = values of the x-variable in a sample
- x bar = mean of the values of the x-variable
- y_i = values of the y-variable in a sample
- y bar = mean of the values of the y-variable

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

Answer: Scaling is a method of transforming data to fit within a specific scale, such as 0-1 or with a mean of 0 and a standard deviation of 1. It is commonly used in preprocessing data during the data preparation phase of machine learning and data science projects.

Types of scaling:

1. **Normalized Scaling (Min-Max Scaling):** Normalization rescales the data into a range of [0, 1] or [-1,1]. It subtracts the minimum value of the feature and then divides by the range (the difference between the maximum and the minimum values).
2. **Standardized Scaling (Z-score Normalization):** Standardization rescales data to have a mean (average) of 0 and a standard deviation of 1. It subtracts the mean value of the feature from each data point and divides by the standard deviation.

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

Answer: An infinite VIF arises specifically when there is perfect multicollinearity or exact linear relationship between some of the independent variables. This means that one independent variable can be expressed as an exact linear combination of other independent variables.

The calculation of VIF is based on the R-squared value obtained by regressing one independent variable against all the other independent variables. The formula for VIF is:

VIF = 1 / 1 - R^2

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 used to compare two probability distributions by plotting their quantiles against each other. If the two distributions being compared are similar, the points in the Q-Q plot will approximately lie on the line (y = x). Q-Q plots are commonly used to compare the distribution of a sample to a theoretical distribution, typically the normal distribution, which is a common assumption in many statistical models.

In the context of comparing a sample to a theoretical normal distribution:

- The x-axis represents the theoretical quantiles from the normal distribution.
- The y-axis represents the ordered sample values (quantiles from the empirical distribution).

Use and Importance in Linear Regression:

1. Checking Normality of Residuals:
2. Identifying Outliers:
3. Comparing Distributions: