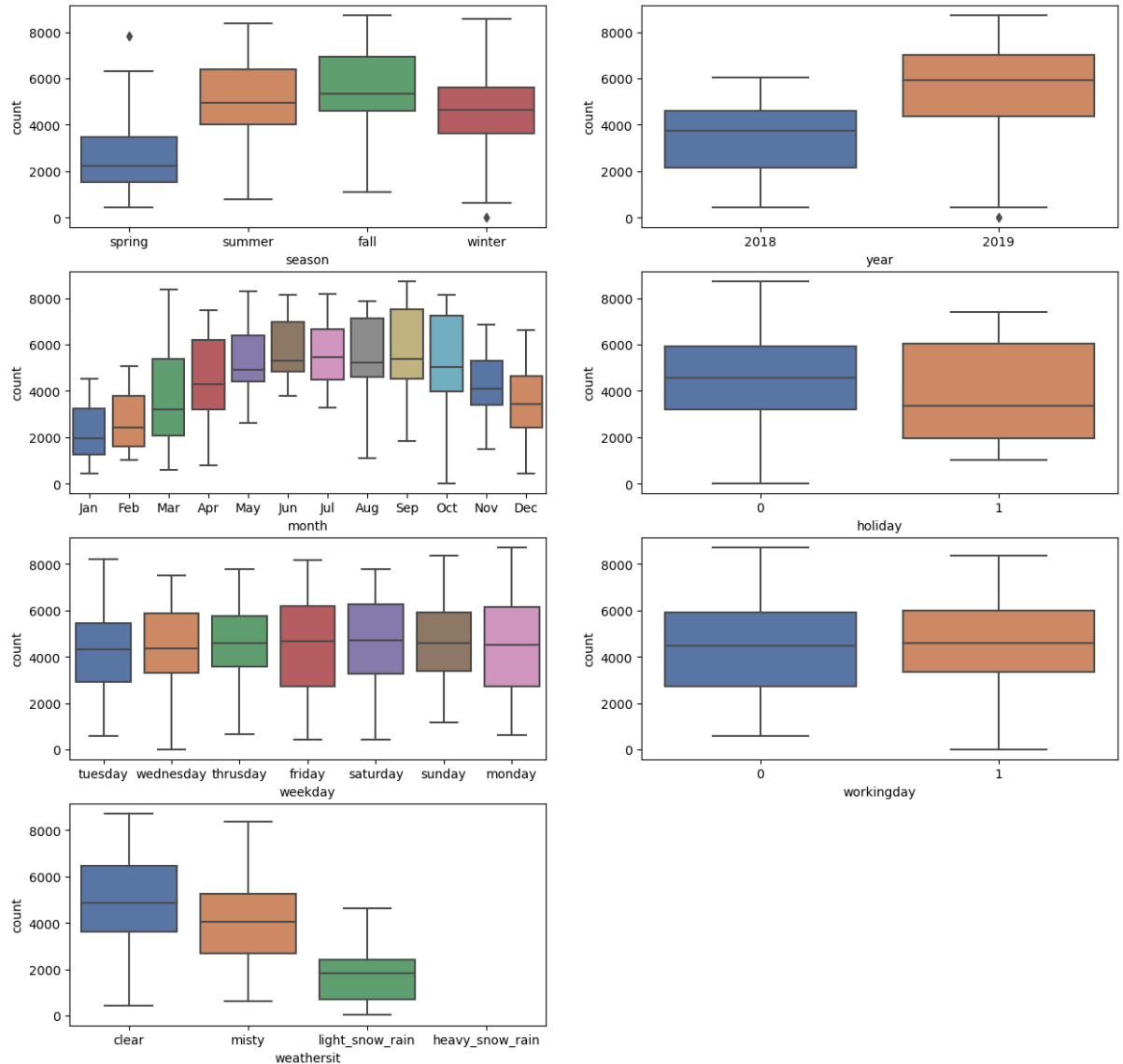# Assignment-based Subjective Questions

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

The dataset's categorical variables—"season," "year," "month," "holiday," "weekday," "workingday," and "weathersit"—were visualized using a boxplot, showing their impact on bike demand.

- Season: Bike demand peaks in fall, is moderate in summer and winter, and is lowest in spring, likely due to weather conditions.
- Year: Demand increased from 2018 to 2019, indicating growing popularity.
- Month: June to September have the highest bike usage, with September leading, while December sees the least rentals due to colder weather.
- Holiday: Bike demand is higher on holidays, likely due to increased leisure time.
- Weekday: Demand is fairly consistent across weekdays, with a slight rise on Fridays and Saturdays.

- Workingday: The "Workingday" boxplot shows that most bookings range between 4000 and 6000, with a consistent median user count throughout the week. There is minimal variation in bookings regardless of whether it is a working day or not.
- Weathersit: Clear weather sees the highest demand, while adverse conditions like snow and rain reduce usage drastically.

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

Using drop_first=True when creating dummy variables is crucial for several reasons:

- **Preventing Multicollinearity**: Including all dummy variables for a categorical feature can cause multicollinearity, where predictor variables are highly correlated. Dropping the first dummy variable avoids this and ensures each variable's impact is distinguishable.
- **Reducing Redundancy**: Dropping the first category decreases the total number of dummies, simplifying the model and enhancing efficiency.
- **Model Interpretability**: This practice keeps the model interpretable by eliminating redundant variables, making it easier to understand the effects of other predictors.

For example, if a categorical column has three values (A, B, C), we can drop A since its absence can be inferred from the presence of B or C. A will be represented as 00, B as 10, and C as 01.

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

The pair-plot analysis reveals that "temp" and "atemp" have the highest positive correlation with the target variable "count," indicating that higher temperatures lead to an increase in bike bookings.

Pairplot Analysis of Target Variables "count" with "temp","atemp"
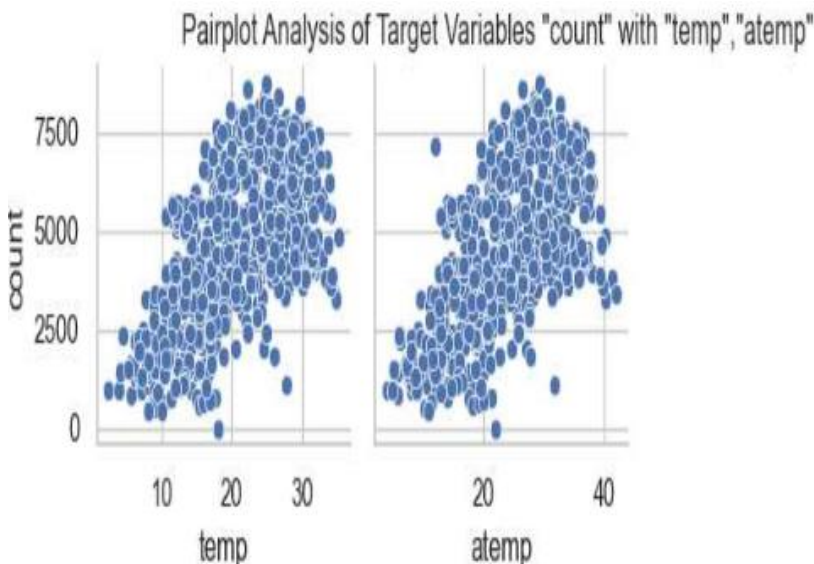
4. **How did you validate the assumptions of Linear Regression after building the model on the training set?**

I validated the assumptions of the Linear Regression Model based on five criteria:

- Normality of Error Terms: I checked this by plotting a histogram and a Q-Q plot of the residuals, both of which should resemble a bell curve if normally distributed.
- Multicollinearity Check: I calculated the Variance Inflation Factor (VIF) for each predictor, confirming that VIF values below 10 indicate no multicollinearity issues.
- Linear Relationship Validation: A residual plot against predicted values showed random scattering around zero, suggesting a linear relationship between predictors and the response variable.
- Homoscedasticity: I assessed residuals against predicted values, and the lack of a clear pattern indicated constant variance.
- Independence of Residuals: The Durbin-Watson statistic was calculated to check for autocorrelation in the residuals.

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

- Temperature ("temp"): Warmer temperatures are linked to greater bike usage.
- Year 2019: Data from 2019 strongly predicts an upward trend in bike usage over time.
- Light Snow or Rain ("weathersit_light_snow_rain"): Poor weather conditions, such as light snow or rain, reduce bike demand by discouraging usage.

General Subjective Questions

1. **Explain the linear regression algorithm in detail.**
   Linear regression is a machine learning algorithm based on supervised learning. It performs a regression task, which means it predicts a continuous output variable (y) based on one or more input variables (x). It is mostly used for finding out the linear relationship between variables and forecasting.
   The basic idea of linear regression is to find a line that best fits the data points, such that the distance between the line and the data points is minimized. The line can be represented by an equation of the form:
   $y = \theta_0 + \theta_1 x$
   where θ_0 is the intercept (the value of y when x is zero) and θ_1 is the slope (the change in y for a unit change in x). These are called the parameters or coefficients of the linear model.

   To find the best values of θ_0 and θ_1, we need to define a cost function that measures how well the line fits the data. A common choice is the mean squared error (MSE), which is the average of the squared differences between the actual y values and the predicted y values:

   $$MSE = (1/n) * \sum(y - y')^2$$

   where n is the number of data points, y is the actual value, and y' is the predicted value.

   The goal is to minimize the MSE by adjusting θ_0 and θ_1. There are different methods to do this, such as gradient descent, normal equation, or using libraries like scikit-learn.

   Linear regression can also be extended to multiple input variables (x1, x2, …, xn), in which case the equation becomes:

   $$y = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + … + \theta_n x_n$$

   Limitations are: it assumes a linear relationship between the input variables and the output variable, which may not always be the case. Another limitation is that it may be sensitive to outliers or multicollinearity.

2. **Explain the Anscombe's quartet in detail.**

   Anscombe's Quartet was developed by statistician Francis Anscombe. It includes four data sets that have almost identical statistical features, but they have a very different distribution and look totally different when plotted on a graph. It was developed to emphasize both the importance of graphing data before analyzing it and the effect of outliers and other influential observations on statistical properties.

   ● The first scatter plot (top left) appears to be a simple linear relationship.
   ● The second graph (top right) is not distributed normally; while there is a relation between them is not linear.
   ● In the third graph (bottom left), the distribution is linear, but should have a different regression line the calculated regression is offset by the one outlier which exerts enough influence to lower the correlation coefficient from 1 to 0.816.

● Finally, the fourth graph (bottom right) shows an example when one high-leverage point is enough to produce a high correlation coefficient, even though the other data points do not indicate any relationship between the variables.
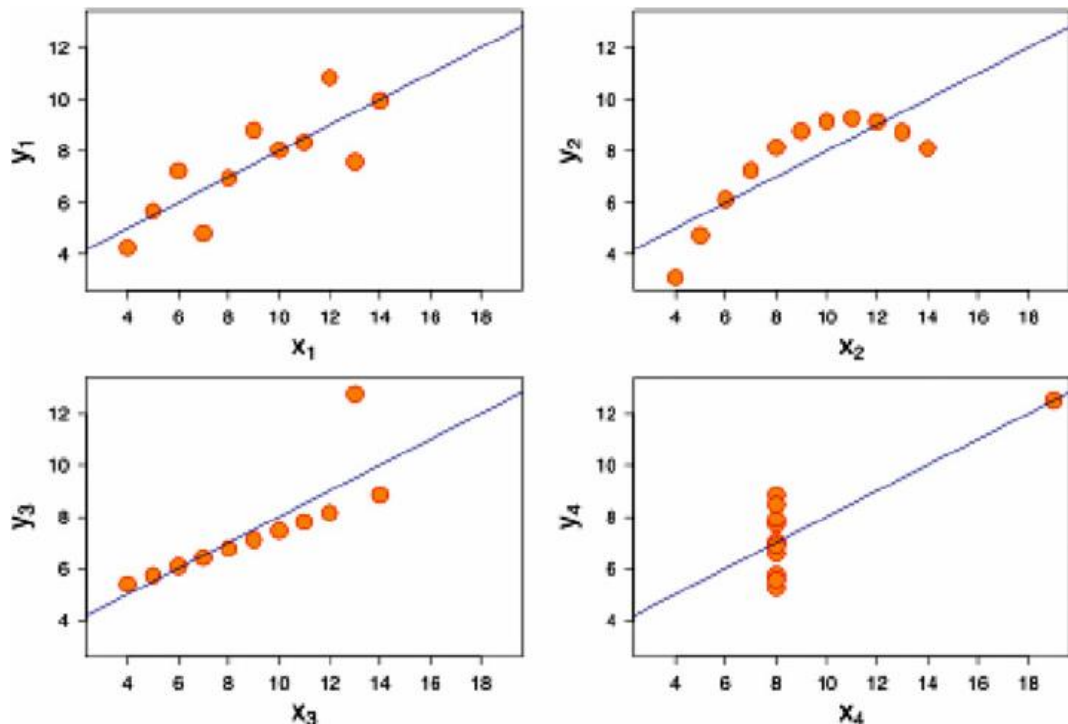
3. **What is Pearson's R?**

   Pearson's r is a numerical summary of the strength of the linear association between the variables. It value ranges between -1 to +1. It shows the linear relationship between two sets of data. In simple terms, it tells us "can we draw a line graph to represent the data?"

   Formula

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

   r = correlation coefficient

   x_i = values of the x-variable in a sample

   $\bar{x}$ = mean of the values of the x-variable

   y_i = values of the y-variable in a sample

   $\bar{y}$ = mean of the values of the y-variable

As can be seen from the graph below, r = 1 means the data is perfectly linear with a positive slope r = -1 means the data is perfectly linear with a negative slope r = 0 means there is no linear association
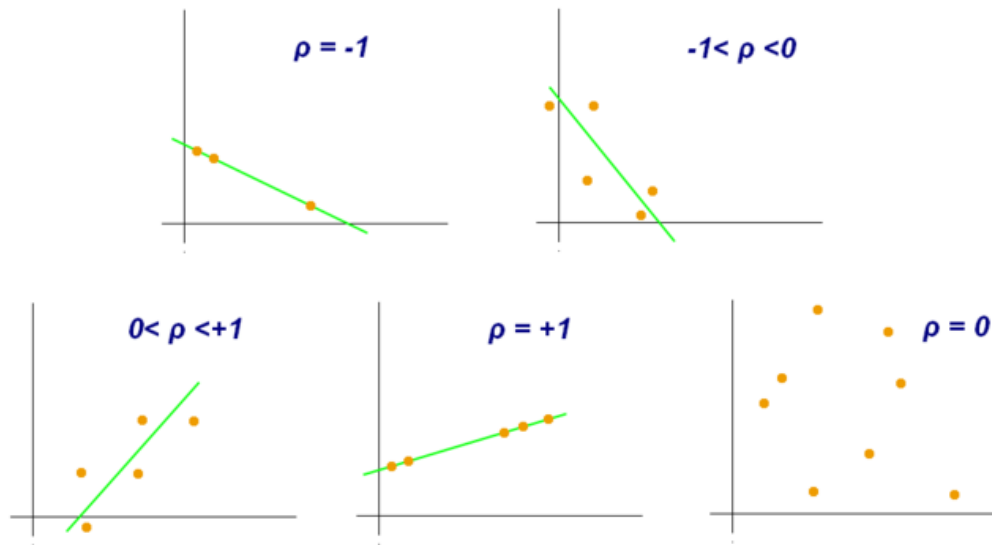
4. **What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?**

Feature scaling is a method used to normalize or standardize the range of independent variables or features of data. It is performed during the data pre-processing stage to deal with varying values in the dataset. If feature scaling is not done, then a machine learning algorithm tends to weigh greater values, higher and consider smaller values as the lower values, irrespective of the units of the values.

● Normalization is generally used when you know that the distribution of your data does not follow a Gaussian distribution. This can be useful in algorithms that do not assume any distribution of the data like K-Nearest Neighbours and Neural Networks.

● Standardization, on the other hand, can be helpful in cases where the data follows a Gaussian distribution. However, this does not have to be necessarily true. Also, unlike normalization, standardization does not have a bounding range. So, even if you have outliers in your data, they will not be affected by standardization.

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

The VIF (Variance Inflation Factor) gives how much the variance of the coefficient estimate is being inflated by collinearity. If there is perfect correlation, then VIF = infinity. It gives a basic quantitative idea about how much the feature variables are correlated with each other. It is an extremely important parameter to test our linear model.

$$VIF = \frac{1}{1 - R^2}$$

Where R-1 is the R-square value of that independent variable which we want to check how well this independent variable is explained well by other independent variables. If that independent variable can be explained perfectly by other independent variables, then it will have perfect correlation and it's R-squared value will be equal to 1. So, VIF = 1/(1-1) which gives VIF = 1/0 which results in "infinity" The numerical value for VIF tells you (in decimal form) what percentage the variance (i.e. the standard error squared) is inflated for each coefficient. For example, a VIF of 1.9 tells you that the variance of a particular coefficient is 90% bigger than what you would expect if there was no multicollinearity — if there was no correlation with other predictors.

A rule of thumb for interpreting the variance inflation factor:

- 1 = not correlated.
- Between 1 and 5 = moderately correlated.
- Greater than 5 = highly correlated.

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

A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set. It is used to compare the shapes of distributions .A Q-Q plot is a scatterplot created by plotting two sets of quantiles against one another. If both sets of quantiles came from the same distribution, we should see the points forming a line that's roughly straight. The q-q plot is used to answer the following questions:

- Do two data sets come from populations with a common distribution?
- Do two data sets have common location and scale?
- Do two data sets have similar distributional shapes?
- Do two data sets have similar tail behaviour?