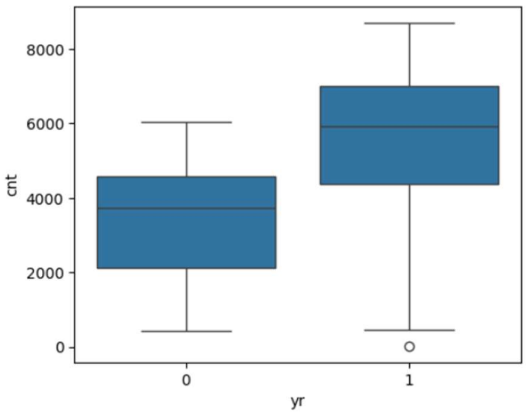
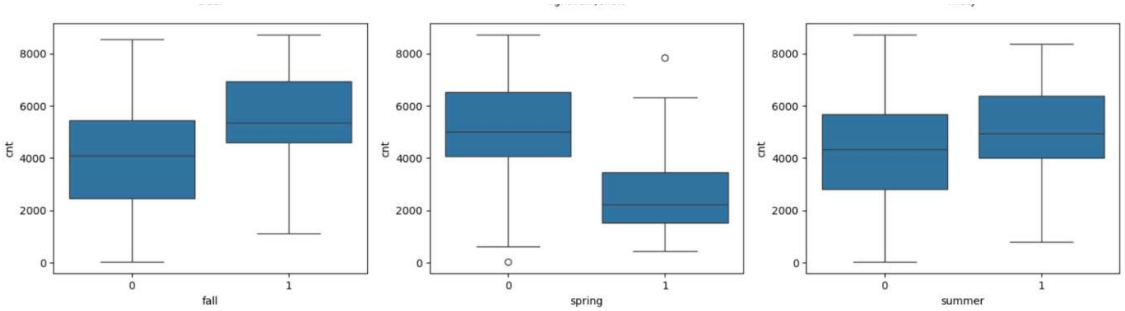
Assignment-based Subjective Questions

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

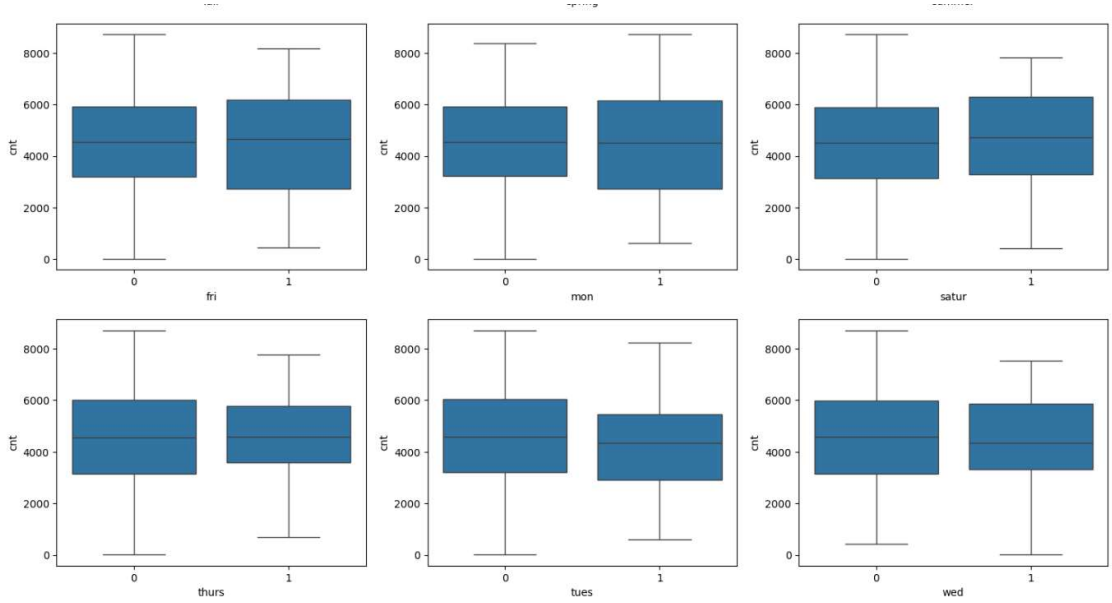
    - Bike rented in the year 2019(represented as 1 in data) is higher as compared to 2018(represented as 0)



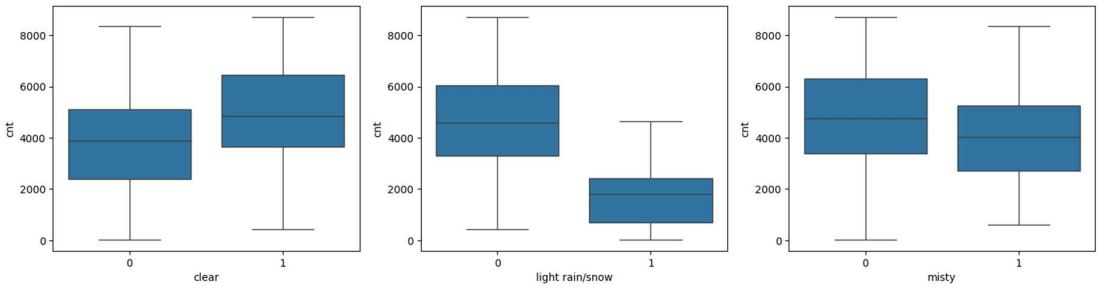
    - Among the season, bike rented is high in fall and low in spring



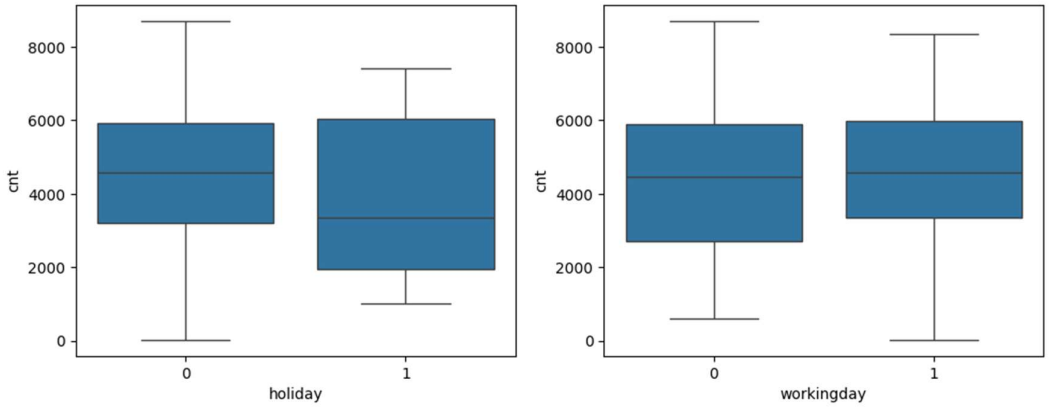
    - Bike rented (the median value) is almost similar throughout all the weekdays



- Bike rented is high when weather is clear and low when it is misty/cloudy or light rain/snow



- Bike rented doesn't change depending on the working day but is relatively lower if it is a holiday

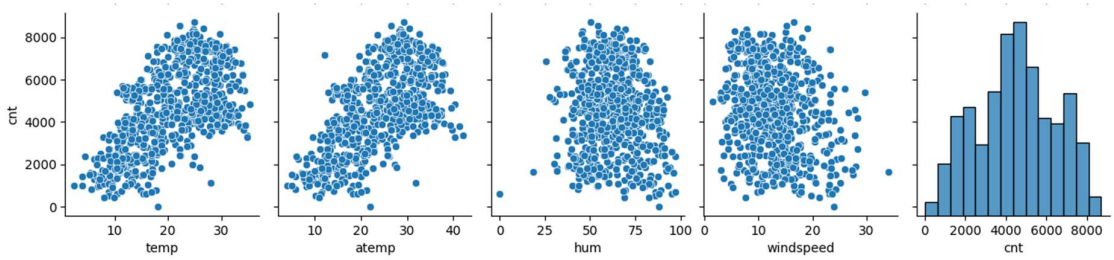


2. Why is it important to use *drop_first=True* during dummy variable creation?

For a categorical variable having 'n' no of categories, the total no of dummy variables needed is 'n-1'.
e.g. A 'gender' column having 2 values 'Male' & 'Female' can be derived as a single dummy variable named 'Male' for which 0 value would denote Female category and 1 as Male.
For this reason, while getting dummy variables using 'get_dummies' method of pandas an extra parameter 'drop_first' can be passed as 'True' to drop the extra column from the created dummy variables.
However this can be achieved manually as well by choosing and removing the column which is thought as necessary.

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

   'temp' and 'atemp' is having the highest correlation with the target variable 'cnt' among all other numerical variables which is found as 63% from correlation



4. **How did you validate the assumptions of Linear Regression after building the model on the training set?**
   - The linear relationship among the X and y variable
   - Normal distribution of error terms with mean value as 0
   - Constant variance of the error terms or Homoscedasticity.

5. **Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?**
   - yr
   - spring
   - light rain/snow

General Subjective Questions

1. **Explain the linear regression algorithm in detail.**
   Linear regression is a supervised machine learning algorithm.
   It predicts the value of a target variable based on different dependent variables which have a linear relationship with the target and can be given by the equation:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_n x_n$$

   where y is the target variable and $x_1, x_2 \ldots x_n$ are the dependent variables and $\beta_0, \beta_1, \beta_2, \ldots \beta_n$ are the coefficients.
   Linear regression model helps us find the values of $\beta_0, \beta_1, \beta_2, \ldots \beta_n$ which gives the relationship of the dependent variables with the target.
   β_0 gives the value of target variable when all other dependent variables are 0.
   $\beta_1, \beta_2, \beta_3, \ldots \beta_n$ respectively gives the amount of change in the target variable when the corresponding dependent variable changes by 1 keeping others constant.

   The linear regression model determines the values of these unknown variables by finding a best fit line by using ordinary least square (OLS) method. In this method, the model minimises the error terms [given by $(y_{actual} - y_{predicted})$].

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

Anscombe's Quartet is a group of four data sets which was first constructed by Francis Anscombe, a British statistician. These data are nearly identical in terms of statistics (means, variance, R-squared), however they have very different distributions and appear differently when plotted on scatter plots.
This tells us about the importance of visualising the data before building a model which would effectively become wrong if solely decided on statistical analysis.

3. **What is Pearson's R ?**

Pearsons'R is the correlation coefficient which gives the relationship between two linearly dependent variables. The value of the coefficient lies between -1 to +1. A very high value of this tells that the correlation between the variables is very high (either negatively or positively, depending on the value ($-ve$ or $+ve$)).

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

Scaling is a process of standardizing/normalizing the numerical data by transforming.

Scaling is performed mainly due to two reasons -

- Scaling reduces the range of the numerical data and the distributions become similar.
- The speed and accuracy increase for gradient descent approach.

Normalization Scaling – Also known as Min-Max Scaling. It maps the minimum value to 0 and maximum to 1 and thus brings all of the data in the range of 0 and 1.

Min-Max Scaling is done using the formula $x = \frac{x - x_{min}}{x_{max} - x_{min}}$

Standardization Scaling – It replaces the values by their corresponding Z values. After the transformation, the mean of the data becomes 0 and standard deviation 1.

Standardization is done using the formula $x = \frac{x - \mu}{\sigma}$

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

If there is very high correlation between two variables, then the VIF value is infinity. It indicates very high multicollinearity among the variables.
VIF is given by the formula,

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

Now, if the value of R^2 is very high (≈ 1) (in case of high correlation), the denominator would become 0 resulting in inf VIF.
An infinite VIF value also indicates that the corresponding variable can be expressed by a linear combination of other variables in the dataset.

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 scatterplot created by plotting two sets of quantiles against one another. Q-Q plots allows to graphically compare two probability distributions to determine normal distribution which is used in linear regression model to check the normality of the error terms by using the actual and predicted data.