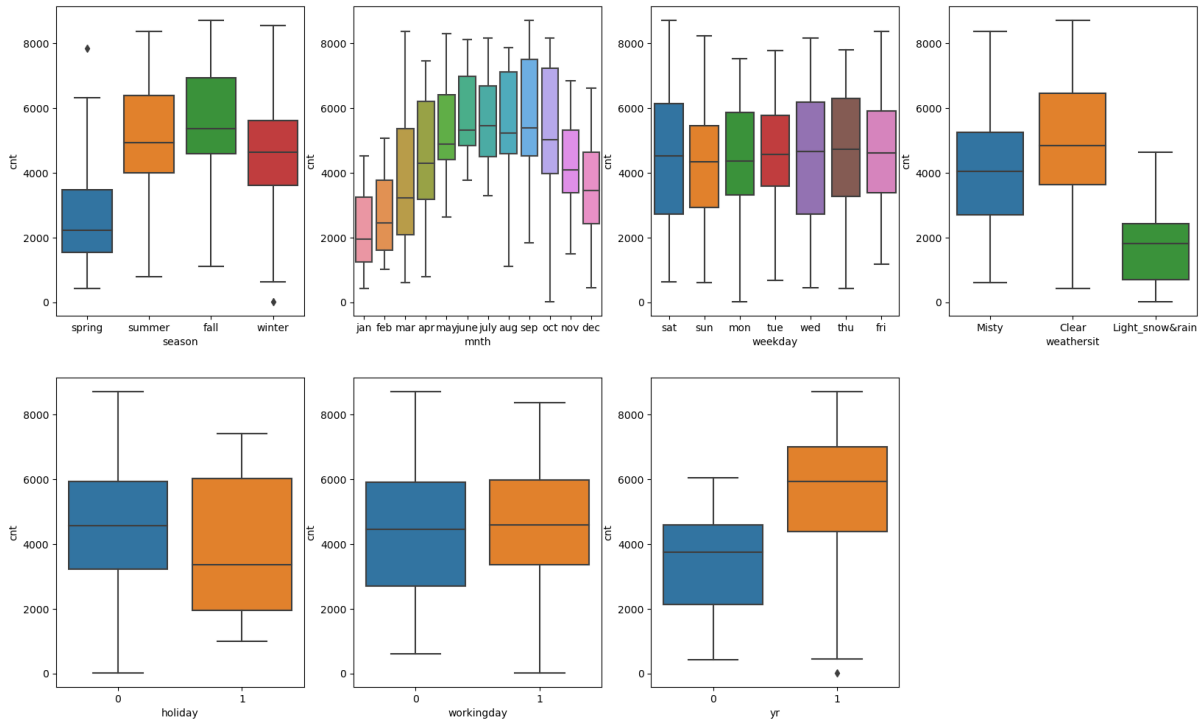
1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

Ans. From the fig we can say that



More Demands in

a. Year 2019 – year box plot
b. Fall season -season box plot
c. Holiday – holiday box plot
d. Not working day - workingday box plot
e. September month – month box plot
f. Saturday – weekday box plot
g. Clear weather - weather

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

Ans.

a. This will drop first category of each categorical variable and will create n-1 dummy variable, where n is number of dummy variables of categorical variable and do the encoding.
b. It reduces extra column while creating dummy variable. Because more dummy features make it harder for the algorithm to fit or even worse make it easier to overfit.
c. Thereafter reduces the corelation between dummy variable.

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

Ans. ‘temp’ has the highest correlation with the target variable i.e. 0.65.

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

Ans.

a. By checking the linear relationship between input and output variables.
b. Histogram curve of error term was observed in the model
c. R square value
d. Less multi collinearity, dependence of input variable on error term

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

Ans.

'temp', 'yr' positively and 'weathersit_light_snow&rain' negatively contributing significantly.

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Ans.

Linear regression is a regression model that uses a straight line to describe the relationship between variables. It finds the line of best fit through your data by searching for the value of the regression coefficient(s) that minimizes the total error of the model.

a. Understanding and cleaning the data
b. Visualize the data and its corelation with the target variable – exploratory data analysis
c. Checking collinearity
d. Spitting data into train and test data
e. Training the model
f. Building linear model checking the factors contributing like R sqr, p value, vif
g. Residual analysis of train data
h. Model prediction and evaluation.

2. Explain the Anscombe's quartet in detail.

Ans.

a. Anscombe's quartet has four datasets, variance, R-squared, correlations, and linear regression lines similar to statistical description but different when scatter plot on graph
b. Each data set has 11 x-y pairs of data. When plotted each dataset have unique characteristics but same statistical summary such as mean, variance
c. **It** is used to illustrate the importance of EDA and the disadvantage of depending only on summary statistics. It focuses on the importance of using data visualization to spot trends, outliers, and other crucial details that might not be obvious from summary statistics alone.

3. What is Pearson's R?

Ans.

a. It is the most common way of measuring a linear correlation.
b. It is the ratio between the covariance of two variables and the product of their standard deviations
c. It is a number between –1 and 1 that measures the strength and direction of the relationship between two variables.

d. When one variable changes, the other variable changes in the same direction.
e. It is used to see the relation between 2 quantitative variable.

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

Ans.

a. The data sets contain various types of units and different rang of magnitude. To bring them to same level to perform operation and to get the proper result scaling is done.
b. Standardization centres data around a mean of zero and a standard deviation of one, while normalization scales data to a set range, often [0, 1], by using the minimum and maximum values.

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

Ans.

a. Perfect corelation means vif is infinity between input and output variable
b. When the R sqr value is 1 , vif becomes infinity.
c. Vif = 1/ 1- R sqr

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

Ans.

Q-Q plot is a graphical tool to assess if sets of data come from the same statistical distribution. It is particularly helpful in linear regression when we are given testing and training datasets differently. In this scenario, it becomes important to check whether both the data comes from the same background, in order to maintain the sanity of the model