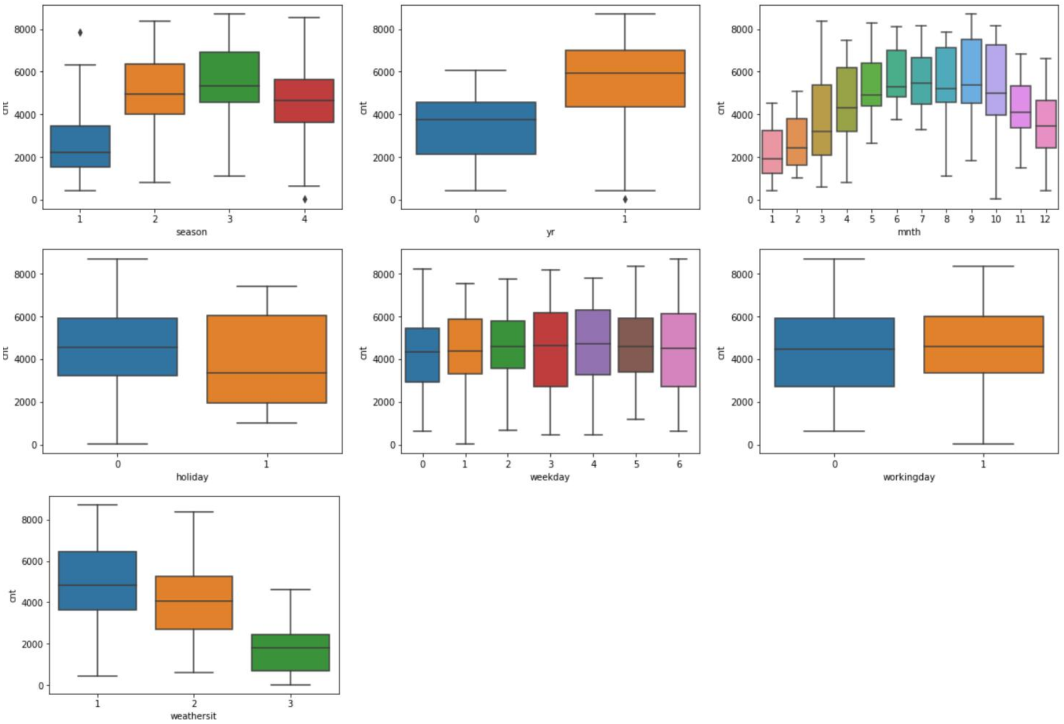
Assignment-based Subjective Questions

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

Answer: 1. ″cnt″ is high in ’2’ and ’3’ seasons.
2. ″cnt″ is high mid of the year and is less at the end, which can be visualised in plot between ″cnt″ and ″mnth″ .
3. ″cnt″ also depends on weather situation.



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

Answer: Dummy variables are created to get encoding for categorical variables. If we have three categories for a variable, we will get three dummy variables if we don’ t use **drop_first=True.** But two dummy variables are enough to uniquely encode. So we use drop_first=True.

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

Answer: Among all numerical variables “registered” has highest correlation as “cont”= “casual”+” registered”.Other than that we have “temp” , “atemp “.

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

Answer: By doing residual analysis, which is res=y_train - y_train_pred and residual should be normally distributed, with mean=0.

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

Answer: yr,temp,month

General Subjective Questions

1. **Explain the linear regression algorithm in detail.**

Answer: Liner regression Algorithm is to find out linear relation ship between dependent variable(y) and one or more independent variables (X) .
Simple Linear Regression: y = b0 + b1 * X
Multiple Linear Regression: y = b0 + b1 * X1 + b2 * X2 …… + bn * Xn

* First the correlation between independent variables with dependent variable is checked.
* Among them, the most linearly correlated independent variables are

selected.
* If in case of multiple independent variables, multicollinearity is checked and the redundant variables are dropped.
* Model is built using the training data.
* R2, Adj. R2 metrics a calculated { 0<R2 <1 }.
* Residual Analysis is performed: res=y_train - y_train_pred.
* Residuals should be normally distributed (mean=0).
* After that , using test data prediction is done and R2 is calculated.
* A difference of 2-3% between R2 score of train and test is acceptable.

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

Answer: Anscombe's quartet comprises four datasets that have nearly identical simple statistical properties, yet appear very different when graphed. Each dataset consists of eleven (x,y) points.
They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data before analyzing it and the effect of outliers on statistical properties.

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

Answer: Pearson's R is referred as Bivariate correaltion or Pearson correlation coefficient , is a correlation between two data sets.
PCC or r = cov(variable1, variable2)/ (std(variable1)* std(variable2))

Result always lies between -1 and +1.
* If r=1, positively correlated
* If r=-1, Negatively correlated
* If r=0, there is no linear association.

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

Answer:
Scaling of variables is an important step because, as you may have noticed, the variable 'temp' is on a different scale with respect to all other numerical variables, which take very small values. Also, the categorical variables that you encoded earlier take either 0 or 1 as their values. Hence, it is important to have everything on the same scale for the model to be easily interpretable.

How to do it:

1. Min-Max scaling (normalisation): Between 0 and 1

normalisation: (x-xmin)/(xmax-xmin)

2. Standardisation (mean=0, sigma=1)

standardisation: (x- mu)/ sigma

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

Answer: If there is perfect correlation, then VIF = infinity. This shows a perfect correlation between two independent variables.
To solve this problem we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

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

Answer: Q-Q plot is plot between two quantiles against each other.
The purpose of Q-Q plots is to find out if two sets of data come from the same distribution.
If the distributions are linearly related, the points in the Q - Q plot will approximately lie on a line.