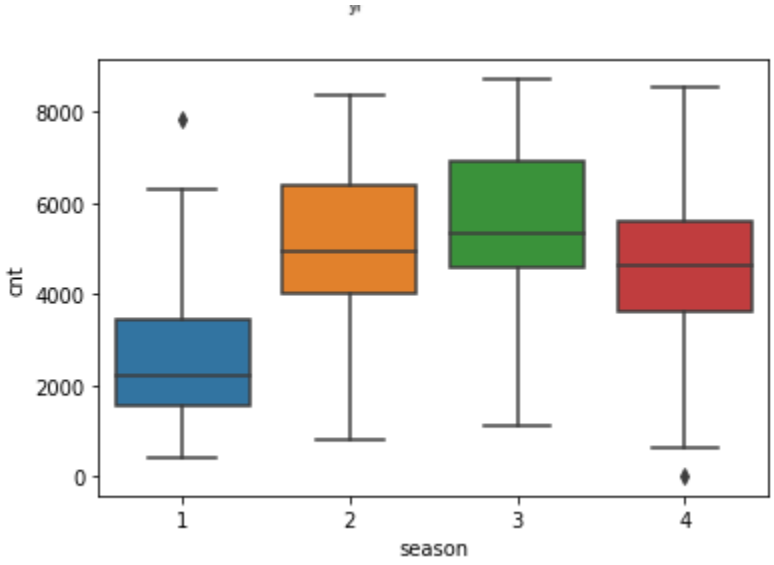
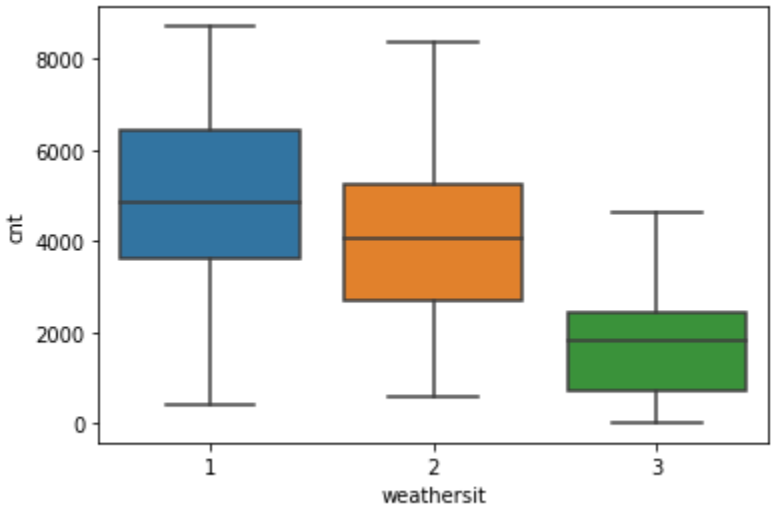
Assignment-based Subjective Questions

1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)

Ans: season and weathersit are two categorical columns. Which are again processed with dummy variables.

Both seasons and weathesit has effect on the dependent variales.





In plot graphs it shows the clear effect on target variables in summer and fall in season plot graphs.In weathersit the mean is more for clear and mist+cloud weather situation.

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

Ans: It removes the first column created for first unique value

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

Ans: temp, atemp has highest correlation with cnt target variable

4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

By building linear regression with OLS and calculating VIF
Predection for using final model after building model OLS and VIF

```
from sklearn.metrics import r2_score
r2_score(y_test, y_pred)
```

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

Ans: **workingday hum, and temp are the top 3 features contributing to demand of shared bikes**

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)

Ans: It is used to perform regression analysis.Linear regression performs the task to predict a dependent variable y based on independent variables x.
It provides the relationship between variables.

Below is the formula for

y=B0+B1*x+B2*x+...+Bn*x

2. Explain the Anscombe’s quartet in detail. (3 marks)
Ans: Anscombe's quartet comprises four data sets that have nearly identical simple descriptive statistics, yet have very different distributions and appear very different when graphed

3. What is Pearson’s R? (3 marks)

Ans: Pearson's R is a measure of linear correlation of two data sets.Its the ratio between the two covariance of two variables and the product of there standard deviations.

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

Ans: it is extremely important to rescale the variables so that they have a comparable scale. If we don't have comparable scales, then some of the coefficients as obtained by fitting the regression model might be very large or very small as compared to the other coefficients. This might become very annoying at the time of model evaluation. So it is advised to use standardization or normalization so that the units of the coefficients obtained are all on the same scale. As you know, there are two common ways of rescaling:

Min-Max scaling
Standardisation (mean-0, sigma-1)

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

Ans: If the correlation is perfect then VIF is infinite.It happens when the corresponding variables may be expressed exactly by a linear combination of other variables.

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

Ans: A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set

QQ plot is used to compare the shapes of distributions providing graphical views.

It is a tool to show if the two data sets come from same distribution.