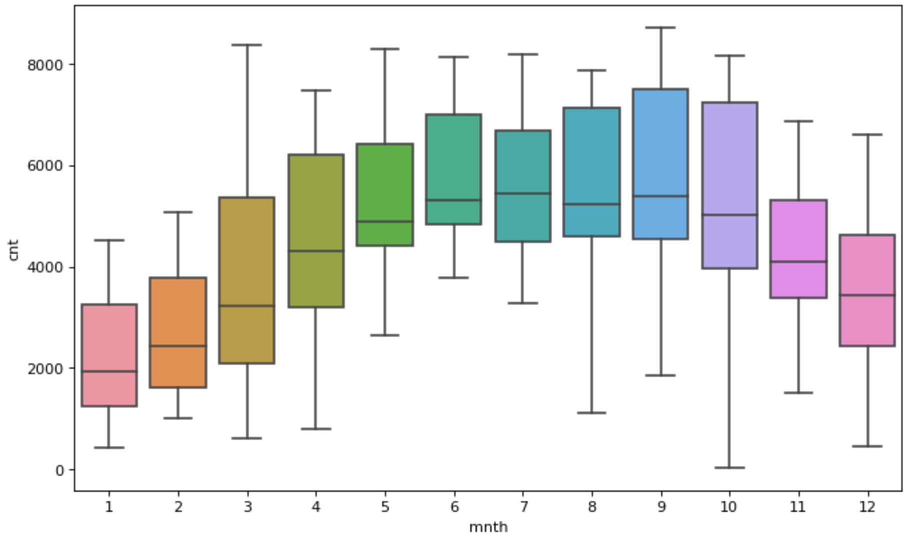
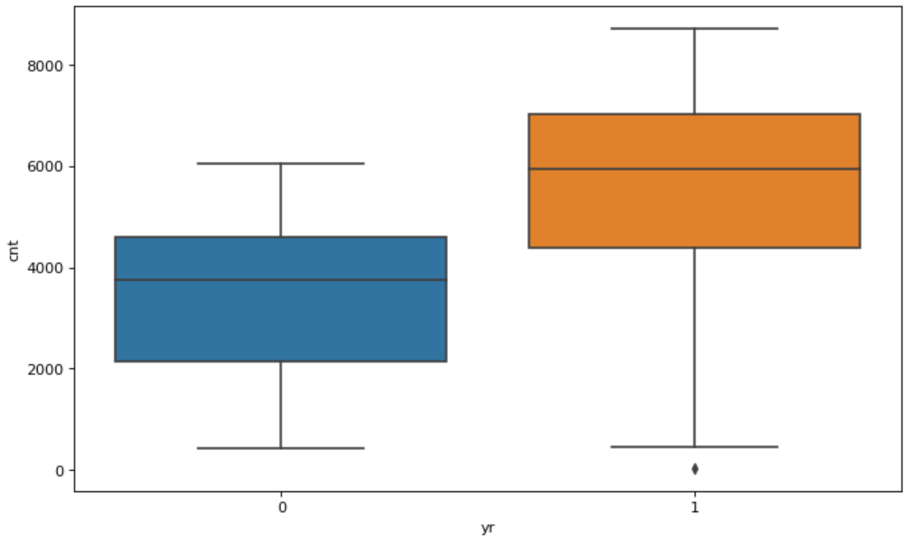
Assignment-based Subjective Questions

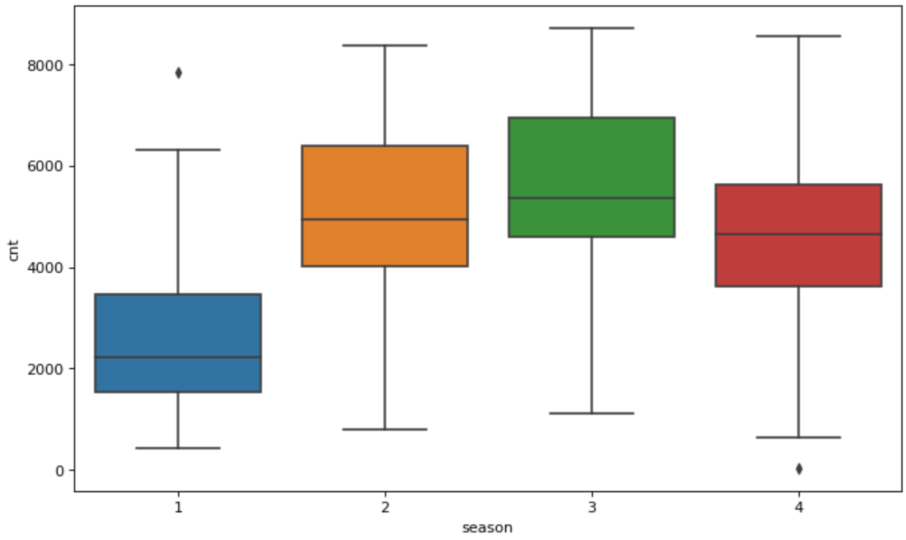
1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)

Season, weathersit, year and month have higher influence. This can be seen in the following diagrams:

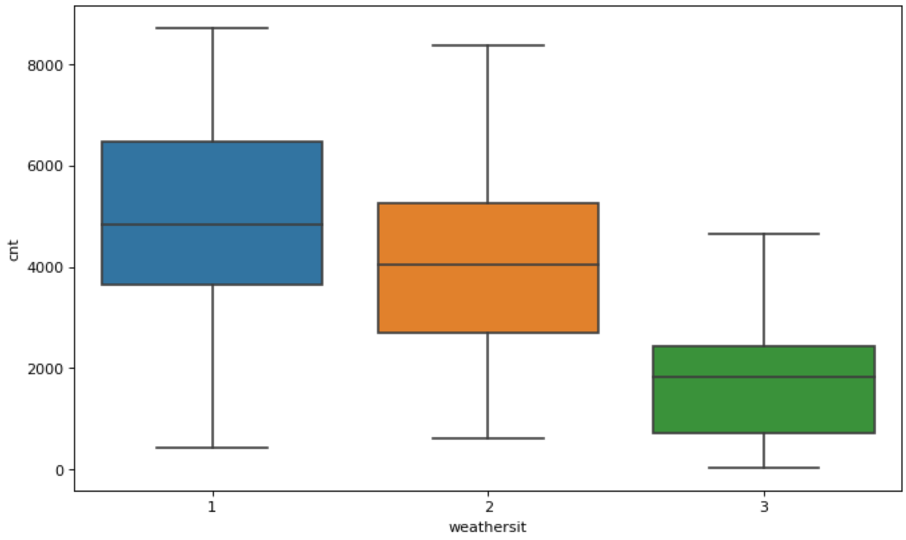




8000
6000
4000
2000
0
cnt
0
1
yr



8000
6000
4000
2000
0
cnt
1
2
3
4
season



8000
6000
4000
2000
0
cnt
1
2
3
weathersit

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

Because for n levels, n-1 columns are created which are enough to represent all the levels as they have information in form of 0 and 1. So drop_first drops the first column.

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

Atemp

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

By checking the following:

1. there exists linearity between x and y
2. error terms are independent of each other
3. error has normal distribution
4. homoscedasticity

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

Season, year and month

General Subjective Questions

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

The link between the dependent (goal variable) and independent factors is explained by a form of predictive modelling technique called linear regression (predictors). Given that linear regression shows a linear relationship, it is possible to utilise it to ascertain how the value of the dependent variable evolves as a function of the value of the independent variable. If there is just one input variable, this type of linear regression is referred to as simple linear regression (x). If there are numerous input variables, this sort of linear regression is also referred to as multiple linear regression. In the linear regression model, a sloping straight line represents the relationship between the variables.

2. Explain the Anscombe’s quartet in detail. (3 marks)

Anscombe's Quartet is a collection of four data sets that, despite having almost equal appearances in simple descriptive statistics, contain a number of anomalies that would trick a regression model if one were to be created. They have noticeably distinct distributions and appear differently on scatter plots. It was created to show the value of graphing data before analysis and model building, as well as the influence of extra observations on statistical aspects. These four data set plots all present the same statistical information, including variance and mean values for both x and y points for all four datasets, and essentially similar statistical findings.

3. What is Pearson’s R? (3 marks)

The most popular method for determining a linear connection is the Pearson correlation coefficient (r). The intensity and direction of the link between two variables is expressed as a number between -1 and 1. Both variables change in the same way when one of them is altered.

$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[\,n\Sigma x^2 - (\Sigma x)^2\,][\,n\Sigma y^2 - (\Sigma y)^2\,]}}$$

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

Scaling is a method for reducing the independent features in the data to a predetermined range. It is done as part of the pre-processing of the data to deal with extremely variable magnitudes, values, or units.

In standardised scaling, the values are rescaled to be centred around the mean with a unit standard deviation, i.e., mean = 0 and standard deviation = 1. In normalised scaling, the feature values are rescaled to range between 0 and 1, while in standardised scaling, the values are rescaled to range between 0 and 1.

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

In essence, VIF (VarianceInflationFactor) aids in the explanation of how one independent variable interacts with every other independent variable. The VIF formulation is shown below:
While a VIF value over 10 is unquestionably high, one over 5 should also not be disregarded and require thorough inspection.
A perfect correlation between two independent variables is indicated by a very high VIF value. If the correlation is perfect, we have R2 = 1, which results in 1/(1-R2) infinite. The variable that is producing this perfect multicollinearity must be removed from the dataset in order to remedy the issue.

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

The quantiles of two probability distributions can be compared using a probability plot, often known as a Q-Q plot.
If you want to discover if a set of data may have come from a theoretical distribution like a normal, exponential, or uniform distribution, you can use a graphic approach called a quantile-quantile (Q-Q) plot.
Use the QQ plot to compare two distributions and see if they are similar.
If they are somewhat comparable, you can assume that the QQ plot will be more linear. The linearity assumption can best be tested using scatter plots. To do a linear regression analysis, all variables must be multivariate normal. A histogram or Q-Q-Plot is the most effective tool for testing this premise.

Importance of QQ Plot in Linear Regression: In linear regression, a Q-Q plot may be produced to ascertain whether the train dataset and test dataset are both taken from a population with the same distribution or not. It can also be used to sample sizes. The presence of outliers as well as variations in position, scale, symmetry, and other distributional properties can all be detected using this graphic.