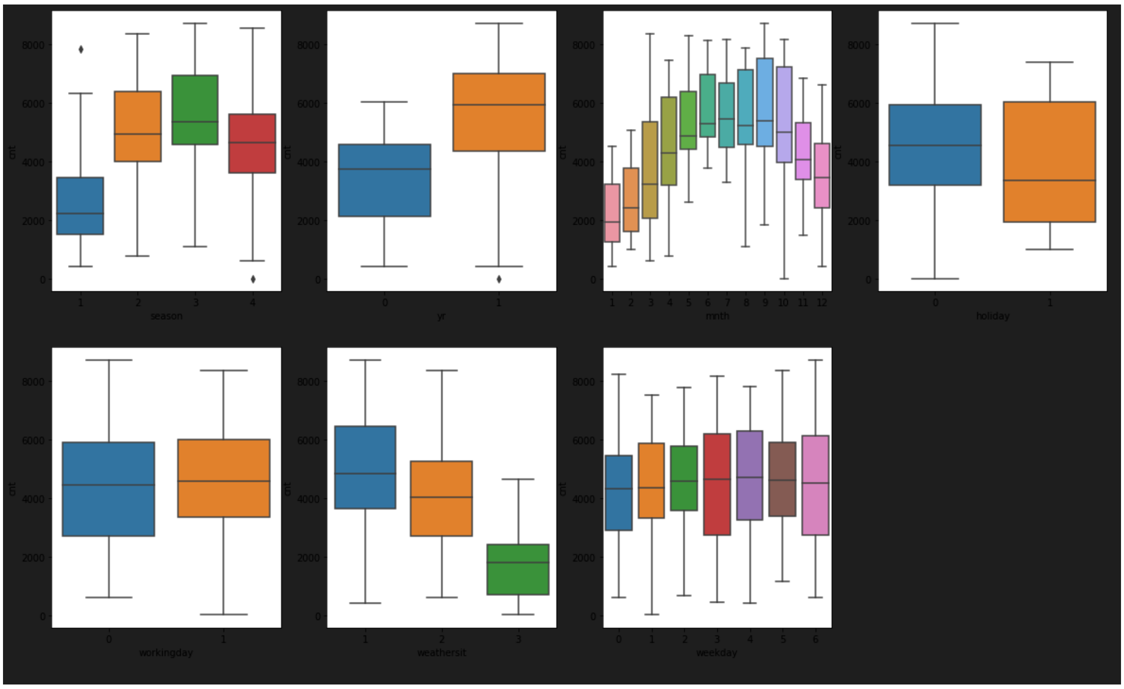
Assignment related questions:

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

   Below plots exhibit, the relationship between the different categorical variables (Season, year, month, holiday, working day, weather situation and weekday) and the dependent variable count.



   As evident from the above inferences, median of count on season 1 is comparatively lower when compared to other three seasons. So, there is some relationship between season and number of users using bicycle. This is justifying the assumption that the variable summer and winter are two key variables in the model.

   Similarly, the count of user is high on working day compared to holiday, and holiday is another variable in the final model. For weather situation 2 and 3, the median of the number of users were less. And in final model, situation 2 and 3 are present with negative coefficient.

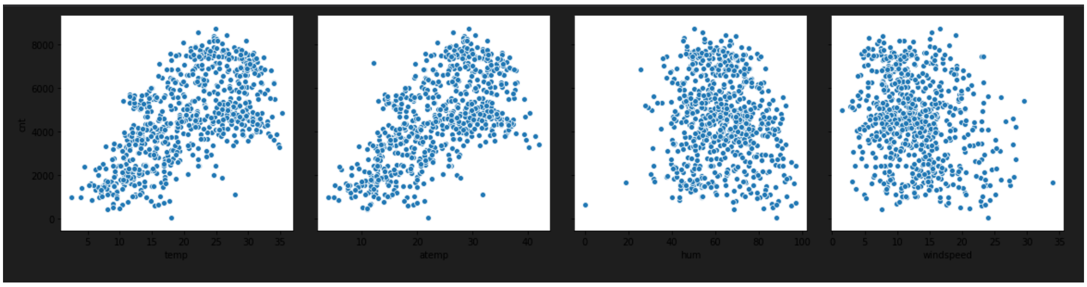
   Hence, in summary, the initial inference from the analysis of box plot for categorical variable were proven correct while the final model was created.

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

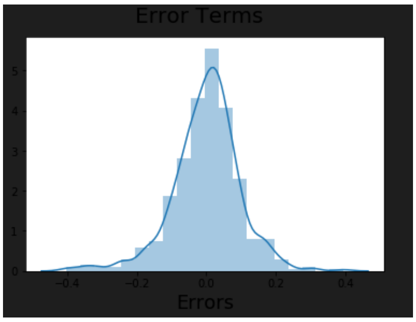
   “drop_first=True” is used, as it helps in reducing the extra column created during dummy variable creation. When we have a categorical variable with, say, 'n' levels, the idea of dummy variable creation is to build 'n-1' variables, indicating the levels. Hence this process reduces the correlations created among dummy variables.

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

Looking at the pair-plot below, both ‘temp’ and ‘atemp’ have similar behaviour and they are the one with highest correlation.



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

Sl#	Perspective	Details
1	Absence of Multicollinearity	• Computed the variance Influence Factor of each predictor variable. • For the predictor variables used in the model, the VIF is less than 5, hence confirmed the absence of multicollinearity.
2	Normal residual distribution	• When the residuals are not normally distributed, then the hypothesis that they are a random dataset, takes the value NO. • This means that in that case regression model does not explain all trends in the dataset. • Histogram of the error term when plotted found to be normal. 

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

Top three contributors are –

- Feeling temperature in Celsius
- Light snow rain – having negative correlation. Decrease in demand in case of light snow and rain
- Year – strategies used in 2019 shows a positive response over 2018 count. Hence this is also a key feature.

General subjective questions

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

   Linear Regression is a machine learning algorithm based on supervised learning. This regression technique finds out a linear relationship between x (input) and y (output). Hence, the name is Linear Regression.

   Linear regression model is represented as Y=C+C1*X1, where Y is the output, C is the interceptor and C1 is the slope and X1 is the predictor variable.

2. Explain the Anscombe's quartet in detail.

   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. 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 analysing it and the effect of outliers and other influential observations on statistical properties. He described the article as being intended to counter the impression among statisticians that "numerical calculations are exact, but graphs are rough.

3. What is Pearson's R?

   Pearson's correlation coefficient (r) is a measure of the strength of the association between the two variables. Pearson's correlation coefficient (r) for continuous (interval level) data ranges from -1 to +1. Positive correlation indicates that both variables increase or decrease together, whereas negative correlation indicates that as one variable increases, so the other decreases, and vice versa.

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

   Machine learning algorithms like linear regression, logistic regression, neural network, etc. that use gradient descent as an optimization technique require data to be scaled.

   - Normalization is a scaling technique in which values are shifted and rescaled so that they end
   - up ranging between 0 and 1. It is also known as Min-Max scaling.

   - Standardization is another scaling technique where the values are centred around the mean with a unit standard deviation. This means that the mean of the attribute becomes zero and the resultant distribution has a unit standard deviation.

   Normalization is good to use when you know that the distribution of your data does not follow a Gaussian distribution. This can be useful in algorithms that do not assume any distribution of the data like K-Nearest Neighbours and Neural Networks.

   Standardization, on the other hand, can be helpful in cases where the data follows a Gaussian distribution. However, this does not have to be necessarily true. Also, unlike normalization, standardization does not have a bounding range. So, even if you have outliers in your data, they will not be affected by standardization.

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

An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables (which show an infinite VIF as well). Sometimes a variable is expressed by a constant, then VIF may become infinity.

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

Quantile-Quantile (Q-Q) plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal, exponential or Uniform distribution. Also, it helps to determine if two data sets come from populations with a common distribution. This helps in a scenario of linear regression when we have training and test data set received separately and then we can confirm using Q-Q plot that both the data sets are from populations with same distributions.
It is used in scenarios where two data sets —

- come from populations with a common distribution
- have common location and scale
- have similar distributional shapes
- have similar tail behaviour