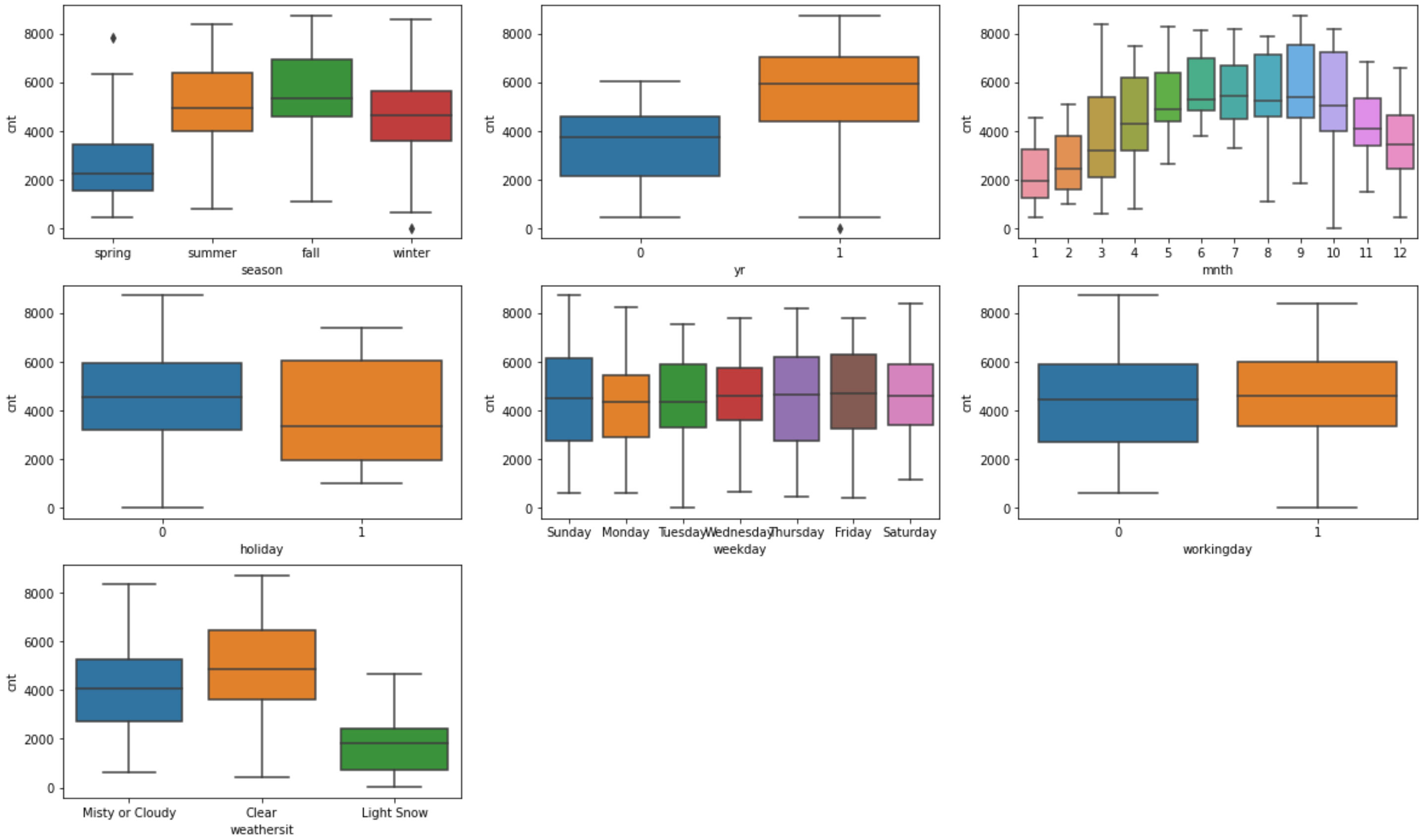
Assignment-based Subjective Questions

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



From the above image we can conclude that,

A) **Cnt vs season** –
There is a significant increase in the total number bike rental in 'summer' and 'fall'. So, these can be good predictors for the total count.

B) **Cnt vs yr** –
Here 0 is 2018 and 1 is 2019. There is a significant increase in the total number bike rental in 2019 i.e., 1. So, this can be good predictor for the total count.

C) **Cnt vs month** –
There is a significant increase in the total number bike rental from 3rd month [March] to till 10th month [October]. So, these can be good predictors for the total count.

D) **Cnt vs holiday** –
Here 0 is holiday and 1 is not a holiday. There is slight increase in the total number bike rental on 'holiday' compared to on 'not a holiday'. So, this can be good predictor for the total count.

E) **Cnt vs weekday** –

Almost all the values have the same count. So, these cannot be good predictors for the total count.

F) **Cnt vs workingday** –
Total count for working day and non-working day is almost same. So, this cannot be good predictor for the total count.

G) **Cnt vs weathersit** –
There is a significant increase in the total number bike rental in 'Misty or Cloudy' and 'Clear'. So, these can be good predictors for the total count.

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

To describe n-levels of categorical variables n-1 dummy variables are required.
If we call the generate dummy variable function for n levels without mentioning 'drop_first=True', it will create
'n' dummy variables. So, to avoid this 'drop_first=True' can be used to create 'n-1' dummy variables.

Ex :

	Holiday
1/1/2022	No
2/1/2022	No
3/1/2022	Yes

A) Without drop_first=True

	Holiday - Yes	Holiday- No
1/1/2022	0	1
2/1/2022	0	1
3/1/2022	1	0

B) With drop_first=True

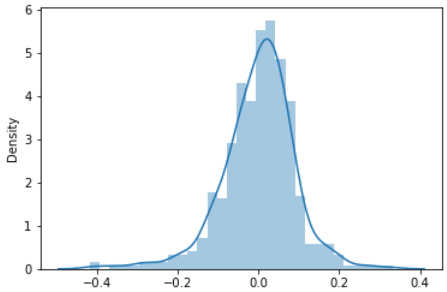
	Holiday - No
1/1/2022	1
2/1/2022	1
3/1/2022	0

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

Both temp and atemp have the highest correlation with the target variable. Since temp has been removed , 'atemp' will have the highest correlation with target variable.

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

   A) Error term normally distributed and centered at 0.



   B) Predicted values for both train and test data are nearly equal.
   Train data set - 83.67% and Test data set – 82.11%

   C) There is no multicollinearity among the variables.

   D) There is no visible patterns in distribution of error term.

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

   Linear regression equation is ,
   cnt = 0.2989 + yr * 0.2325 + atemp * 0.4575 + hum * (-0.1612)+ windspeed * (-0.1627) + spring * (-0.0789)+ winter * 0.0992 + Light Snow * (-0.2309) + Misty or Cloudy * (-0.0537) + 3 * 0.0618 + 4 * 0.0585 + 5 * 0.0938 + 6 * 0.0649 + 8 * 0.0628 + 9 * 0.1156

   top 3 features contributing significantly towards explaining the demand of the shared bikes are ,

   1) atemp
   2) yr
   3) Month (9)

General Subjective Questions

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

   Linear regression is a machine learning algorithm based on the surprised learning that supports finding the correlation among the variables. It is a statistical regression used for predictive analysis and it will show the relationship between the continuous variables. It shows that there is a linear relationship between independent and dependent variable(s).

Majorly linear regression can be divided as,

1) Simple linear regression – There will one target (dependent) variable which we need to predict from one independent variable.
   $y = \beta_0 + \beta_1 x$
   Here, x = Independent variable
   y = Dependent variable
   β_1 = slope
   β_0 = intercept

   Ex: Predict the salary of an employee based on year of experience. The recent data from the company indicates the relationship between experience and salary. Year of experience is an independent variable, and the salary of an employee is a dependent variable, as the salary of an employee is dependent on the experience of an employee. Using this insight, we can predict the future salary of the employee based on current & past information by using the simple linear regression.

2) Multiple linear regression – There will one target (dependent) variable which we need to predict from multiple independent variables.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon$$

   Here, x1, x2………. xp = Independent variables
   y = Dependent variable
   β_1, β2………… βp = coefficient of each independent variable
   β_0 = intercept

   Ex: prediction of used-car prices based on make, model, year, shift, mpg and color.

The best-fit line is found by minimizing the expression of RSS (Residual Sum of Squares) which is equal to the sum of squares of the residual for each data point in the plot.
Error(e) = y – y-predicted
RSS = $\sum_0^i e2$

Accuracy of your model can be checked by R2 statistics. R2 is a number which explains what portion of the given data variation is explained by the developed model. It always takes a value between 0 & 1. In general term, it provides a measure of how well actual outcomes are replicated by the model, based on the proportion of total variation of outcomes explained by the model, i.e., expected outcomes. Overall, the higher the R-squared, the better the model fits your data.

Assumption of Regression Model:

- Linearity: The relationship between dependent and independent variables should be linear.
- Homoscedasticity: Constant variance of the errors should be maintained.
- Multivariate normality: Multiple Regression assumes that the residuals are normally distributed.
- Lack of Multicollinearity: It is assumed that there is little or no multicollinearity in the data.

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

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 when analyzing 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.

- The first scatter plot appears to be a simple linear relationship, corresponding to two variables correlated where y could be modelled as gaussian with mean linearly dependent on *x*.
- The second graph while a relationship between the two variables is obvious, it is not linear, and the Pearson correlation coefficient is not relevant. A more general regression and the corresponding coefficient of determination would be more appropriate.
- In the third the modelled relationship is linear, but should have a different regression line (a robust regression would have been called for). The calculated regression is offset by the one outlier which exerts enough influence to lower the correlation coefficient from 1 to 0.816.
- Finally, the fourth graph shows an example when one high-leverage point is enough to produce a high correlation coefficient, even though the other data points do not indicate any relationship between the variables.

The quartet is still often used to illustrate the importance of looking at a set of data graphically before starting to analyze according to a particular type of relationship, and the inadequacy of basic statistic properties for describing realistic datasets.

3. **What is Pearson’s R? (3 marks)**

In statistics, the Pearson correlation coefficient also known as Pearson's *r*, the Pearson product-moment correlation coefficient (PPMCC), the bivariate correlation, or colloquially simply as the correlation coefficient is a measure of linear correlation between two sets of data. It is the ratio between the covariance of two variables and the product of their standard deviations; thus it is essentially a normalized measurement of the covariance, such that the result always has a value between −1 and 1. As with covariance itself, the measure can only reflect a linear correlation of variables, and ignores many other types of relationship or correlation. As a simple example, one would expect the age and height of a sample of teenagers from a high school to have a Pearson correlation coefficient significantly greater than 0, but less than 1 (as 1 would represent an unrealistically perfect correlation).

Pearson's correlation coefficient is the covariance of the two variables divided by the product of their standard deviations. The form of the definition involves a "product moment", that is, the mean (the first moment about the origin) of the product of the mean-adjusted random variables; hence the modifier *product-moment* in the name.

For a population[edit]

Pearson's correlation coefficient, when applied to a population, is commonly represented by the Greek letter ρ (rho) and may be referred to as the *population correlation coefficient* or the *population Pearson correlation coefficient*. Given a pair of random variables (X,Y), the formula for ρ is:

$\rho X,Y = cov (X, Y) / \sigma X \sigma Y$

where: cov is the covariance

σX is the standard deviation of X

σY is the standard deviation of Y

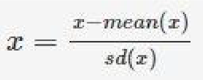
4. **What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)**

Scaling: Another important aspect to consider is feature scaling. When you have a lot of independent variables in a model, a lot of them might be on very different scales which will lead a model with very weird coefficients that might be difficult to interpret. So we need to scale features because of two reasons:

1. Ease of interpretation

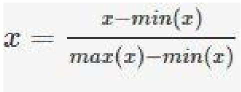
2. Faster convergence for gradient descent methods You can scale the features using two very popular method:

1. Standardizing: The variables are scaled in such a way that their mean is zero and standard deviation is one.



$$x = \frac{x - mean(x)}{sd(x)}$$

2. MinMax Scaling: The variables are scaled in such a way that all the values lie between zero and one using the maximum and the minimum values in the data.



$$x = \frac{x - min(x)}{max(x) - min(x)}$$

It is important to note that scaling just affects the coefficients and none of the other parameters like t-statistic, F statistic, p-values, R-square, etc.

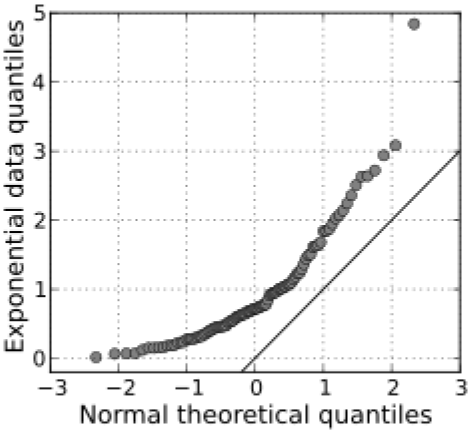
5. **You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)**
If there is perfect correlation, then VIF = infinity. This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity. To solve this problem, we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

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).

A large value of VIF indicates that there is a correlation between the variables. A general rule of thumb is that if VIF > 10 then there is multicollinearity.

6. **What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)**
Q-Q Plots (Quantile-Quantile plots) are plots of two quantiles against each other. A quantile is a fraction where certain values fall below that quantile. For example, the median is a quantile where 50% of the data fall below that point and 50% lie above it. The purpose of Q Q plots is to find out if two sets of data come from the same distribution. A 45-degree angle is plotted on the Q Q plot; if the two data sets come from a common distribution, the points will fall on that reference line.

A Q Q plot showing the 45-degree reference line:



If the two distributions being compared are similar, the points in the Q–Q plot will approximately lie on the line y = x. If the distributions are linearly related, the points in the Q–Q plot will approximately lie on a line, but not necessarily on the line y = x. Q–Q plots can also be used as a graphical means of estimating parameters in a location-scale family of distributions.

A Q–Q plot is used to compare the shapes of distributions, providing a graphical view of how properties such as location, scale, and skewness are similar or different in the two distributions.