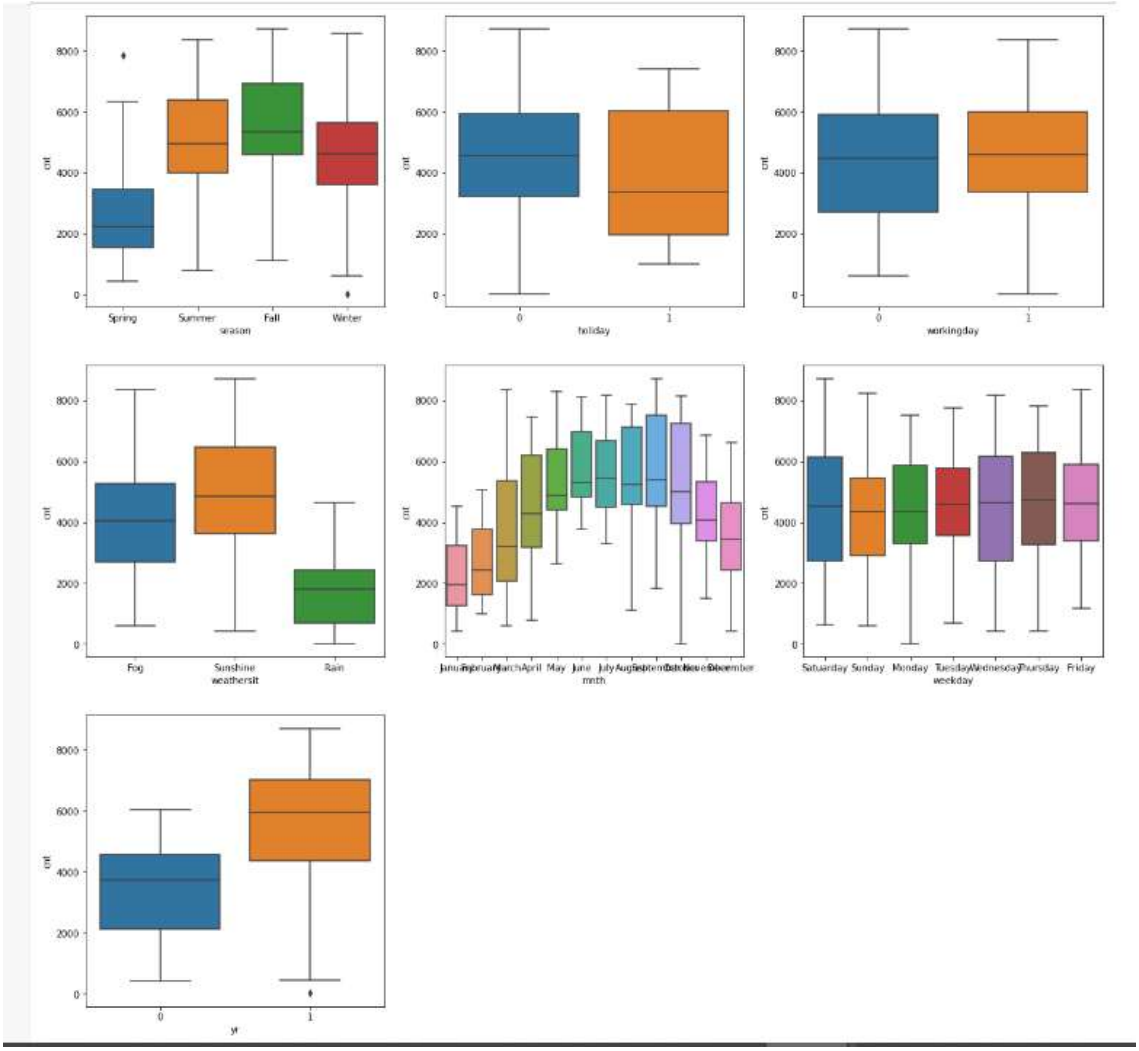
Assignment-based Subjective Questions

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:

1. During the Fall session the rental cnt is increased and during the spring it is decreased
2. From month May to Oct the rental cnt is increased
3. Year 2018 has higher bookings than 2022
4. In rain time the rental count is decreased significantly



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

Ans:

To create n-1 dummies out of n categorical levels by removing the first level.

It helps in reducing the extra column created during dummy variable creation and it 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? (1 mark)**

Ans:

Temp /Atemp has highest correlation

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

Ans:

Using residual analysis

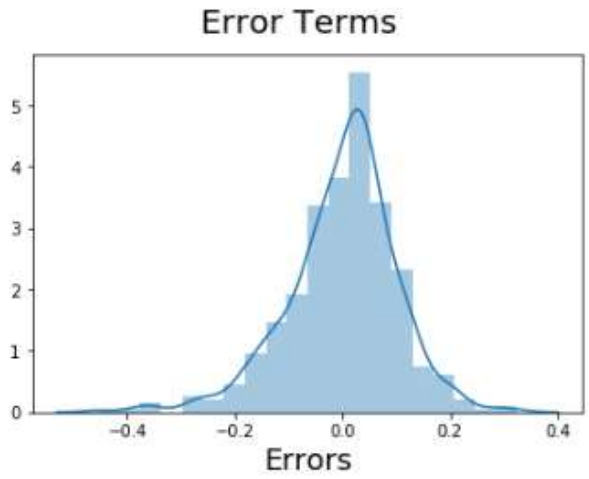
```
### Residual Analysis

Before we make predictions on the test set, let's first analyse the residuals.
```

```
75]: y_train_cnt = lm3.predict(X_train_rfe2)
```

```
76]: # Plot the histogram of the error terms
     fig = plt.figure()
     sns.distplot((y_train - y_train_cnt), bins = 20)
     # Plot heading
     fig.suptitle('Error Terms', fontsize = 20)
     # Give the X-label
     plt.xlabel('Errors', fontsize = 18)
```

```
76]: Text(0.5, 0, 'Errors')
```



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:

1. Temperature
2. Year
3. Holiday

General Subjective Questions

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

Ans:

Linear Regression is a machine learning algorithm based on supervised learning.

And it used to for target prediction based on independent variables and find the relationship between variables and forecasting

There are two types of regression:

1. Simple linear regression
2. Multiple linear regression

Simple linear regression:

Simple linear regression is used to estimate the relationship between two quantitative variables.

- How strong the relationship is between two variables
- The value of the dependent variable at a certain value of the independent variable

Regression models describe the relationship between variables by fitting a line to the observed data. Linear regression models use a straight line, while logistic and nonlinear regression models use a curved line. Regression allows you to estimate how a dependent variable changes as the independent variable(s) change.

Multiple linear regression:

It is used to estimate the relationship between two or more independent variables and one dependent variable. You can use multiple linear regression when you want to know:

1. How strong the relationship is between two or more independent variables and one dependent variable
2. The value of the dependent variable at a certain value of the independent variables

Multiple linear regression formula:
The formula for a multiple linear regression is:

$$y = \beta_0 + \beta_1 X_1 + \ldots + \beta_n X_n + \epsilon$$

y = the predicted value of the dependent variable

B_0 = the y-intercept (value of y when all other parameters are set to 0)
$B_1 X_1$ = the regression coefficient (B_1) of the first independent variable (X_1)
… = do the same for however many independent variables you are testing
$B_n X_n$ = the regression coefficient of the last independent variable
ϵ = model error

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

Ans:

It uses descriptive statistics for grouping together four similar datasets. It gives weightage on the visualization of the dataset first and then using algorithms.

And it says that linear regression is the best algorithm for determining the linear relationship between the variables. But when the model is plotted using scatter plot, it has difference even when the datasets are similar.

Anscombe's quartet

I		II		III		IV	
x	*y*	*x*	*y*	*x*	*y*	*x*	*y*
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

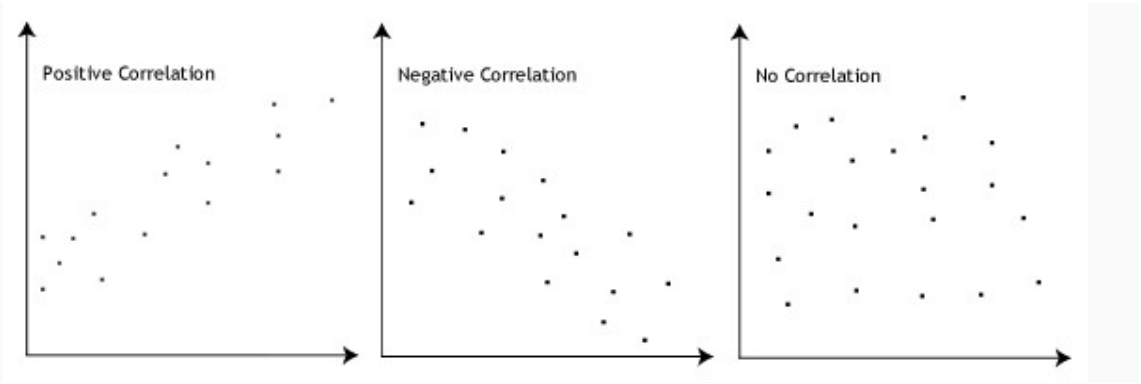
3**. What is Pearson's R? (3 marks)**

Ans:

Pearson's correlation coefficient is the covariance of the two variables divided by the product of their standard deviations.

It is a number between -1 to 1 that measures the strength direction of the relation between two variables.

r = 1 means the data is perfectly linear with a positive slope (i.e., both variables tend to change in the same direction)
r = -1 means the data is perfectly linear with a negative slope (i.e., both variables tend to change in different directions)
r = 0 means there is no linear association
r > 0 < 5 means there is a weak association
r > 5 < 8 means there is a moderate association
r > 8 means there is a strong association



Pearson r Formula

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

Here,

r = correlation coefficient
x_i = values of the x-variable in a sample
$\bar{x}$ = mean of the values of the x-variable
y_i = values of the y-variable in a sample
$\bar{y}$ = mean of the values of the y-variable

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

Ans:
Scaling provides an automatic solution to outliers. If there is a huge deviation from one value to another value.
Scaling helps to order between 0 and 1 for a better learning for the model.
Min max scaling is most importantly done for numerical features. If scaling is prior, then there will be huge coefficient of weights and the learning will be improper.

i. Normalized Scaling:

- Scale values are either between (-1,1) or (0,1)
- Outlier detection is not that good in normalized scaling compared to Min max.

ii. Standardized scaling:

- It doesn't have ranges. Mean is 0 and standard deviation is always 1. Best for outliers.

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

Ans:

VIF is infinity, If there is perfect correlation.

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.

An infinite VIF value indicates that the corresponding variable 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:

Q-Q Plots are scattered plots of two quantiles against each other.

The use of Q-Q plots is to find out if two sets of data come from the same distribution.