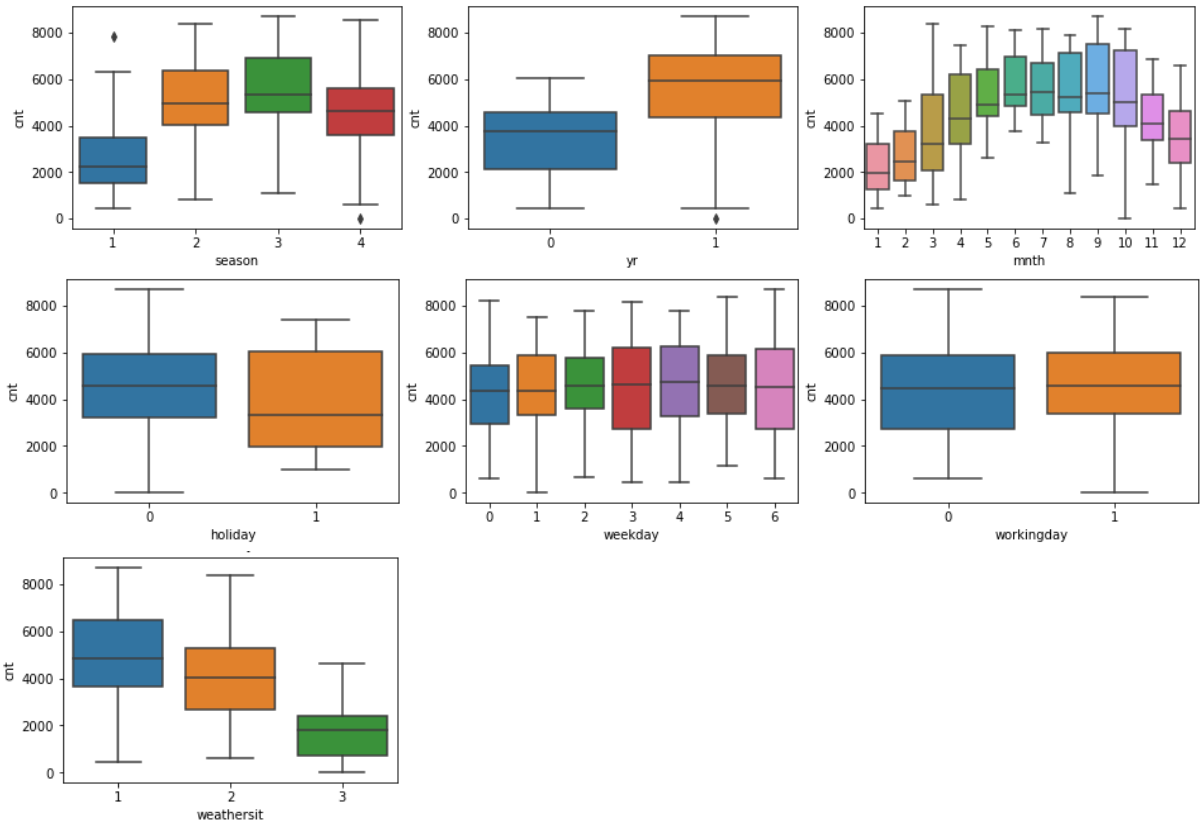
Assignment-based Subjective Questions

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:

- season3(spring) has the highest demand.
- year 2019 got more demand than 2018.
- for the months, first it is increasing and then decreasing.
- holidays have fatty demand for rental bike.
- for weekdays, percentage for 3rd and 6th days is more compared to other days.
- weather1 has more demand than others.

Refer the below image for reference:



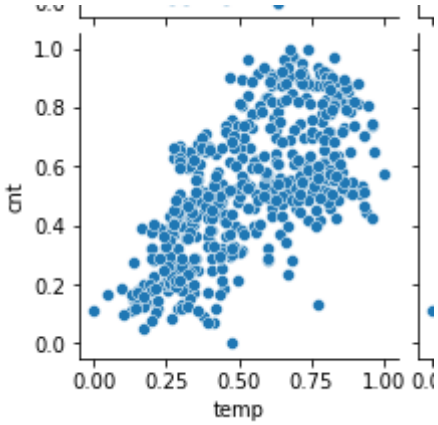
2. Why is it important to use 'drop_first=True' during dummy variable creation? (2 mark)

Ans:

It helps in reducing the extra column or unwanted column created during dummy variable creation. Also, it avoids multicollinearity.

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

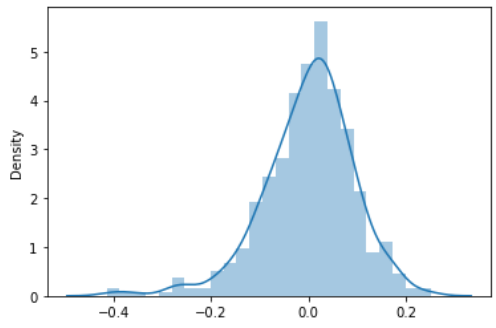
Ans: it is variable 'Temp'. Refer the below image for reference:



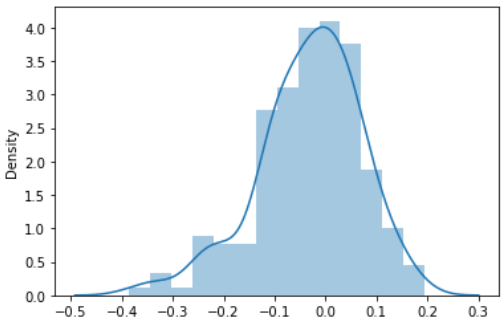
4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

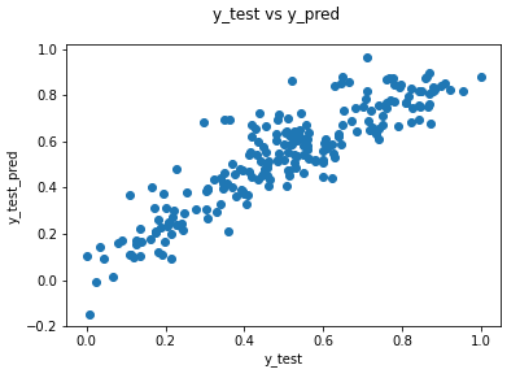
Ans: to evaluate model, made plot for residual train and test set. Also y test and y test pred

```
y_train_pred = lr_model8.predict(X_train_sm8)
res = y_train-y_train_pred
sns.distplot(res)
plt.show()
```




```
y_test_pred = lr_model8.predict(X_test_sm8)
res_test = y_test-y_test_pred
sns.distplot(res_test)
plt.show()
```





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: Temp (co-efficient: 0.5983), year (co-efficient: 0.2350), and winter season (co-efficient: 0.1257),

General Subjective Questions

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

Ans: In Regression, there is a graph we plot between the variables which best fit the given data points. The machine learning model can deliver predictions regarding the data, which is basically used for prediction, forecasting, time series modelling, and determining the causal-effect relationship between variables.

Linear regression shows the linear relationship between the independent variable on X-axis and the dependent variable on Y-axis and that is why it is called as linear regression. It is a regression method used for predictive analysis and shows the relationship between the continuous variables. If there is a single input variable (x), it is called **simple linear regression**. if there is more than one input variable, then it is called **multiple linear regression**.

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

Ans: Anscombe's Quartet is a group of four data sets which are nearly identical in simple descriptive statistics, but there are some peculiarities in the dataset that fools the regression model if built. They have very different distributions and appear differently when plotted on scatter plots.

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

it is generally Pearson's correlation coefficient, whenever we discuss correlation in statistics.
Pearson's Correlation Coefficient is also referred as **Pearson's r, the Pearson product-moment correlation coefficient (PPMCC), or bivariate correlation**.

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

Where,
N = the number of pairs of scores
Σxy = the sum of the products of paired scores
Σx = the sum of x scores
Σy = the sum of y scores
Σx^2 = the sum of squared x scores
Σy^2 = the sum of squared y scores

the value of correlation coefficients lies between -1 and +1 and the magnitude tells us the strength of the relationship while the sign suggests the direction like positively correlated or negatively correlated and a zero correlation implies no linear relationship at all.

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

Ans: it is data pre-processing step which is applied on independent variables for normalized scaling and standardized scaling.

Difference:

Normalization: is also called min-max scaling and range is 0-1 or -1 to 1. Minimum and maximum value are used for scaling

$$X_{changed} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

Standardization: it is also called Z-Score Normalization. It doesn't have any certain range. Mean and standard deviation is used for scaling.

Standardization: $\frac{X-\mu}{\sigma}$

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

Ans: It is because of correlation between independent variables. If VIF is infinity, it may show a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1, which lead to VIF (1/(1-R2)) infinity. To avoid this multicollinearity, we used to drop the independent variables one by one until we get this resolved.

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 is also called quantile-quantile plot. it is a fraction where certain values fall below a quantile. For instance, 50% of the data lie above and below of the median, being a quantile. 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.

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.