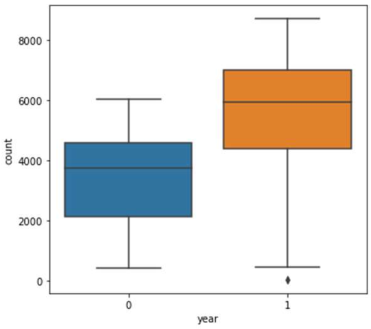
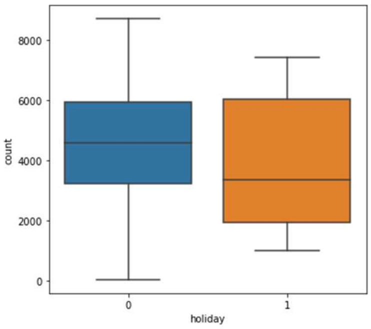
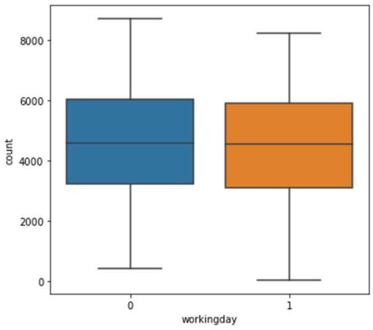
Assignment-based Subjective Questions

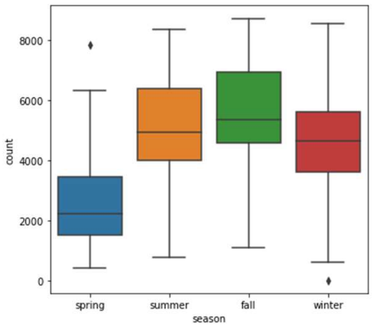
1. **From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)**

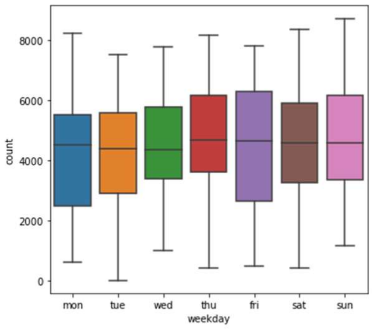
1. Fall season has much greater demand for rental bike compared to other seasons. Summer is second closet.
2. Weekday does not seem to have any significant impact on demand for rental bikes.
3. May-September has significantly higher demand for rental bikes.
4. Clear and misty weather conditions results in higher demand for rental bikes. Whereas rains have negative effect on the same.

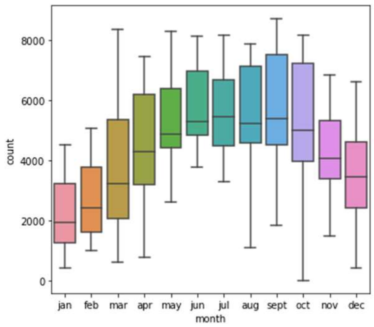


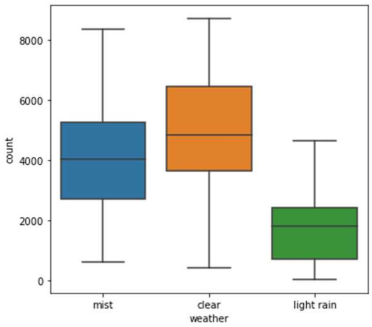












2. **Why is it important to use drop_first=True during dummy variable creation? (2 mark)**

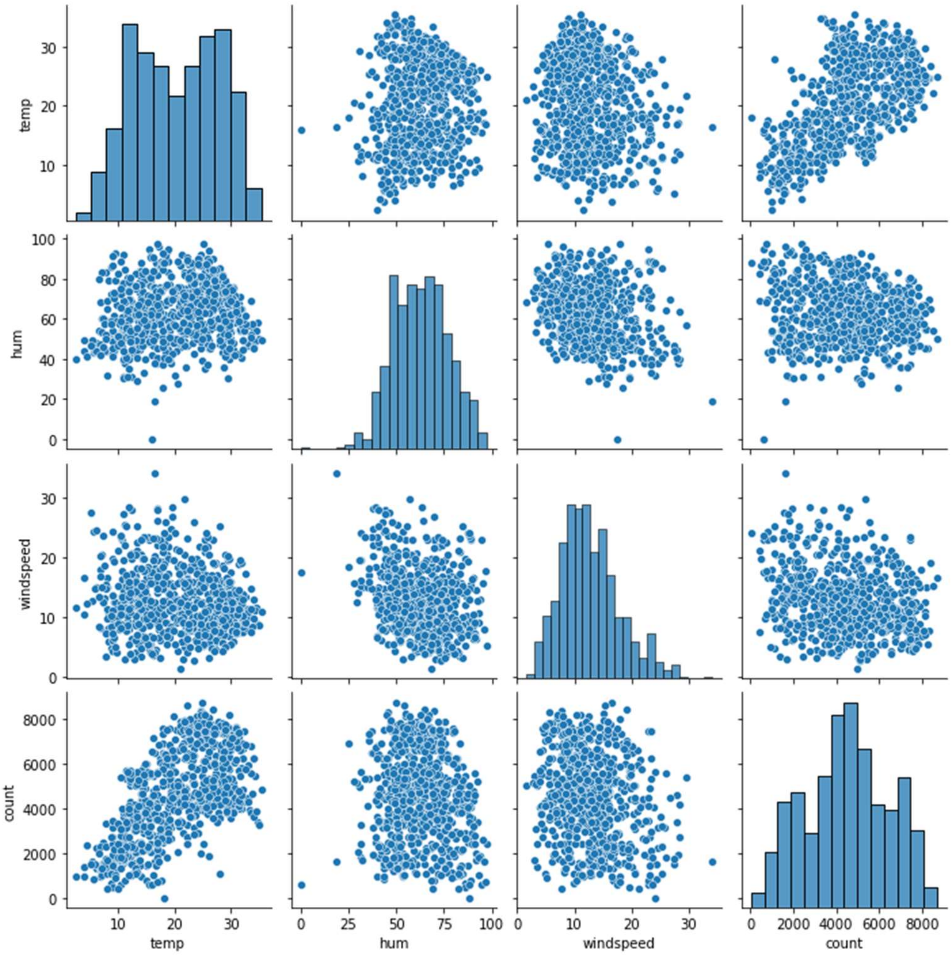
It helps in reducing the extra column created during dummy variable creation.

We used drop_first=True because just like Boolean value. There are only limited number of outcomes is possible.

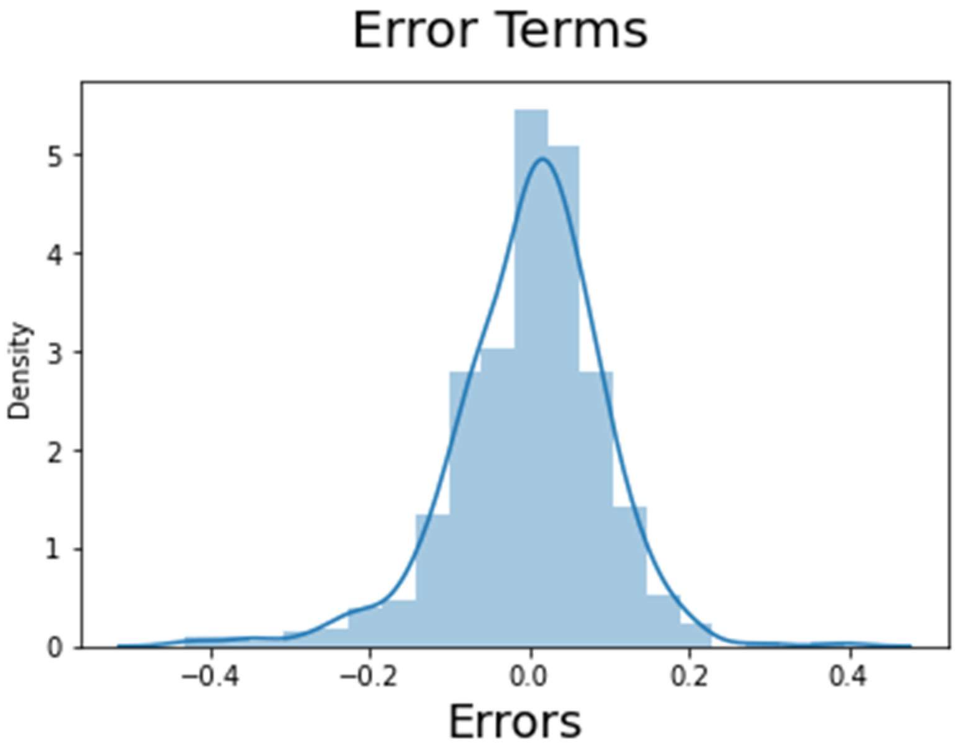
e.g. - Suppose Sky has only three descriptions: clear, cloudy and raining. If sky is not cloudy neither raining. Then it must be clear. With this logic, we don't need to know if it is clear or not. because it either one of those three.

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

Temp variable has the highest correlation with count (target variable.)



4. **How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)**
    1. We found that dependent variable and independent variable has linear relationship. Temp has linear effect on demand for rental bikes.
    2. The residuals are independent and have constant variance no matter the level of the dependent variable.
    3. The residuals of the model are normally distributed.



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

1. **Temperature** has highest impact on demand. Greater the temperature, higher the demand.
2. Weather type **'Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds'** have negative impact on demand for rental bikes.
3. Year **2019** have positive impact on demand.

General Subjective Questions

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

   LR is a machine learning algorithm based on supervised learning. There is always single target/dependent variable and multiple independent variables. Using regression, we can find out relationship between dependant and independent variable. Furthermore, we can make predictions.

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

   As per my understanding, it is importance of data visualization. Datasets with similar mean, variance, correlation-coefficient and line of best fit. Could have drastically different data distribution once plotted because outliers.

   So, it is always a good practice the plot the data even if we have the summary. Visually we can see anomalies and patterns which numbers may not express.

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

   Pearson's R is most used correlation-coefficient in linear regression.

   It shows linear relationship between two variables as R value.

   Pearson's R score is always between -1 and 1.

   Where score -1 signifies there is strong negative correlation between two variables.

   score 0 signifies that there is no correlation between two variables.

   score 1 signifies that there is strong positive correlation between two variables.

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

   Scaling is handling the maximum and minimums of the data. scaling doesn't impact your model all that much. However, it makes it much more useable when doing numerical analysis.

   There are two types of scaling,

   1. Min-Max scaling -

      Min Max scaling clips the outliers on the either side to fixed min and max values.

   2. Standardisation -

      Standardisation maintains the actual ratio of the data but shrinks it.

      e.g., if the dataset has a count variable with min value 10 and max value of 100. We will assume the values after 50 are outliers. Standardisation can shrink the data to min value 1 and max value as 10. Notice the ratio is not changed at all.

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

(3 marks)

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.

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 are used to compare two common distributions. A Q-Q plot is a plot of the quantiles of the first data set against the quantiles of the second data set. For regression, when checking if the data in this sample is normally distributed, we can use a Q-Q plot to test that assumption. Q-Q plots can help you validate the assumption of normally distributed residuals.