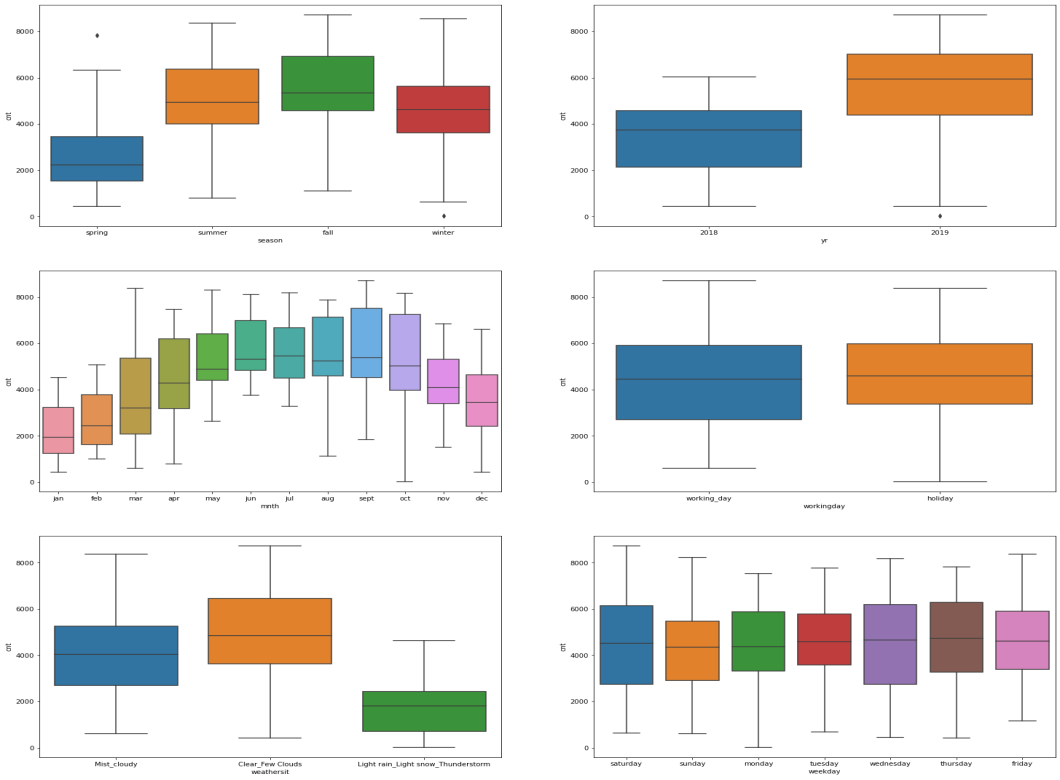
Assignment-based Subjective Questions:

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: The categorical variables in the dataset are season, yr, mnth, weathersit, weekday, workingday. Boxplot is used to visualize the data. Below are the inferences:

1. **Season:** Spring season has the lowest rentals and demand is high in fall.
2. **Yr:** 2019 has the highest demand for Bike rentals.
3. **Mnth:** The demand gradually increases from jan till sept and drops for the next 3 months.
4. **Workingday:** The number of rentals is high in workingday.
5. **Weathersit:** the rentals are highest in 'Clear_Few Clouds' weather condition and lowest in 'Light rain_Light snow_Thunderstorm'.
6. **Weekday:** there is not much difference in rentals in the weekdays; however we can see Sunday has low rentals.



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

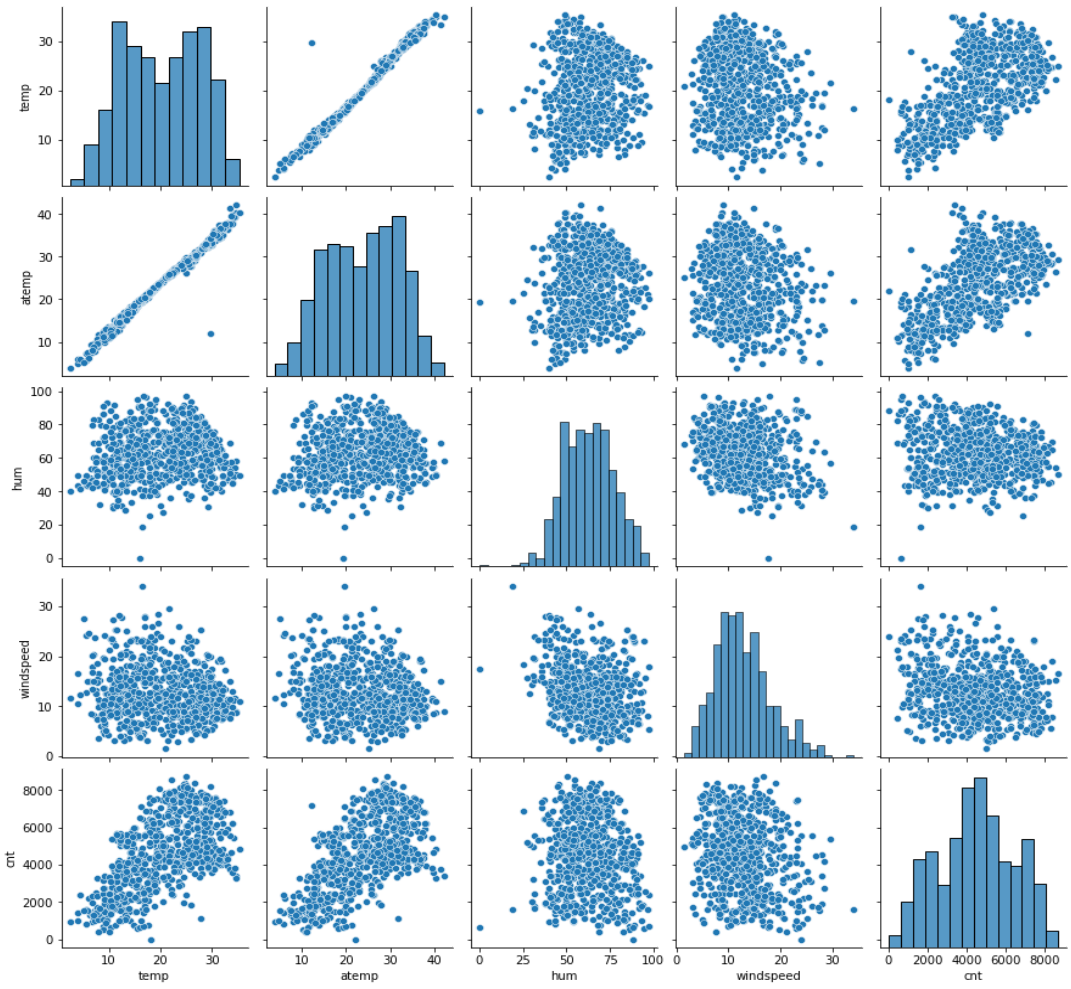
ANS: drop_first =True is used to reduce the number of columns used. When the data can be explained with two columns instead of three, it would help in the readability and reduces the correlation created among the dummy variables.

Taking an Ex:

Say we have three values (A, B, C) for a particular variable. We can interpret using 10 for A, 01 for B, 00 for C, so only two columns are enough to explain all the three values.

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

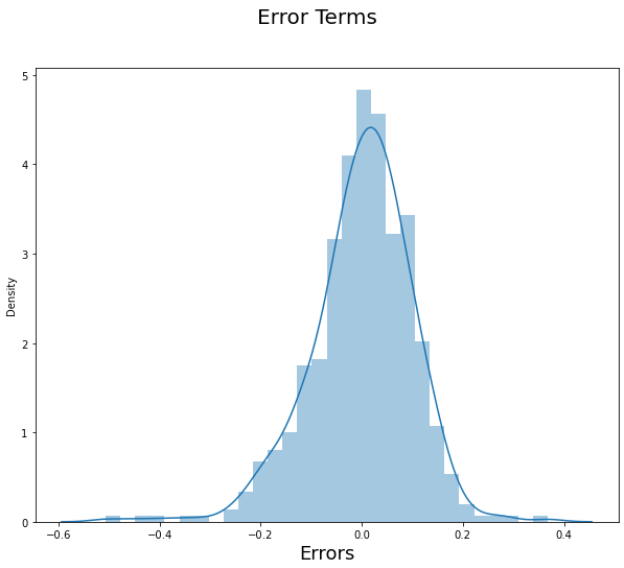
ANS: 'temp' and 'atemp' have the highest correlation with the target variable 'cnt'.



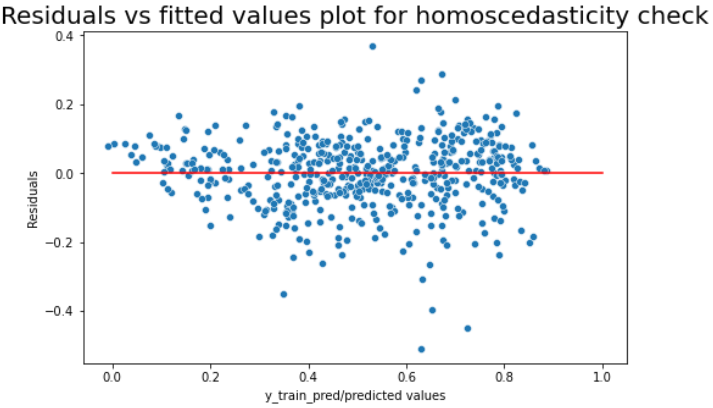
4. **How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)**

ANS:

- Residual Analysis: Errors are normally distributed with mean 0.



- Linearity between target and input variables [straight line].
- Homoscedasticity: Cone shape should is not present and there is no pattern observed in the plot.



- Errors should be independant; this can be checked using the DW value. [In our case it was: 2.001].
- VIF and p-values are optimal.

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: The top three features contributing significantly towards the demand of shared bikes are:

- **Weathersit:** Light rain_Light snow_Thunderstorm with coefficient 0.3045.
- **Yr:** 2019 year with coefficient 0.2468.
- **Season:** spring season with coefficient 0.1973.

General Subjective Questions:

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

ANS: Linear regression is a method of finding the best straight line fitting to the given data, which is finding the best linear relationship on any given data between the independent and dependent variables.

It is mostly done by the Sum of Squared Residuals Method.

Equation of straight line is y = mx+c, where m is the slope/gradient and c is the interpret (y value when X is zero).

Linear regression can be classified into Simple LR and Multiple LR.

Simple LR: Model with only 1 independent variable.

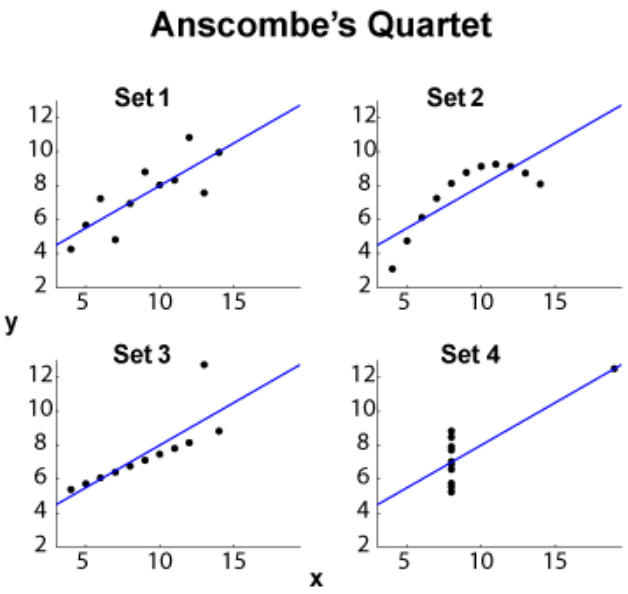
Multiple LR: Model with more than 1 independent variables.

Assumptions of Simple Linear regression are:

a. Linear relationship between target and input variable.
b. Error terms are normally distributed.
c. Error terms are independent of each other.
d. Error terms have constant variance (Homoscedascity).

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

ANS: Anscombe's quartet has four data sets that are identical in simple descriptive statistics, however when plotted they have very different distributions. Each dataset consists of eleven (x,y) points.



It was built by Francis Anscombe to demonstrate the importance of visualizing the data using graphs before model building.

The statistical information such as mean and standard deviation for these four datasets are approximately similar.

The four datasets can be explained as:

- The first dataset X1, fits the linear regression model well.
- X2, could not fit linear regression model on the data as the data is non-linear.
- X3 shows the distribution is linear, however the outliers involved in the dataset are not handled by linear regression model .
- X4 shows that one outlier is enough to produce a high correlation coefficient.

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

ANS: In Statistics, the Pearson's Correlation Coefficient is also referred to as Pearson's r. It is a statistic that measures the linear correlation between two variables. Like all correlations, it also has a numerical value that lies between - 1.0 and +1.0. It cannot capture nonlinear relationships between two variables and cannot differentiate between dependent and independent variables. Pearson's correlation coefficient is the covariance of the two variables divided by the product of their standard deviations.

Pearson's R is given by

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

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

Σx2 = the sum of squared x scores

Σy2 = the sum of squared y scores

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

ANS: It is a data processing step which helps to normalize/standardize the independent variables within a particular range. This step is performed to maintain all variables in the same range for better results. Scaling only affects the coefficients and not the T-statistics, F-statistics, P-statistics, or the R-squared values.

Scaling methods:

a. Normalisation/MinMax scaling: MinMax scaling brings all of the data in the range of 0 and 1.

$$\text{Minmax scaling} = X - \min(x)/\max(X) - \min(X)$$

b. Standardization: Standardization basically brings all the data into a standard normal distribution with mean zero and standard deviation one.

$$\text{Standard scaling} = X - \text{mean}(X)/\text{std}(X)$$

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

ANS: 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 R^2 =1, which lead to 1/(1-R^2) infinity.

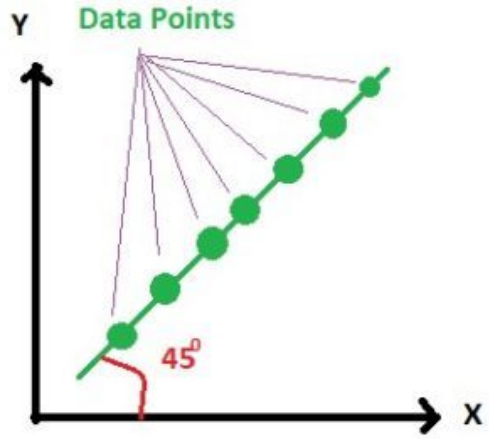
To solve this problem, we need to drop one of the variables from the dataset which is causing this perfect multicollinearity. An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables.

6. **What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)**

ANS: When the quantiles of two variables are plotted against each other, then the plot obtained is known as quantile – quantile plot or qqplot. This plot provides a summary of whether the distributions of two variables are similar or not with respect to the locations.

Interpretations:

All point of quantiles lie on or close to straight line at an angle of 45 degree from x – axis. It indicates that two samples have similar distributions.



In practice it is always not possible to get such a 100 percent clear straight line but the plot looks like below. Here the points are lying nearly on the straight line.