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?**

   The categorical variables in the dataset are season, yr, mnth, weathersit, weekday and workingday.

   Boxplot is used to visualize the data. Below are the inferences:

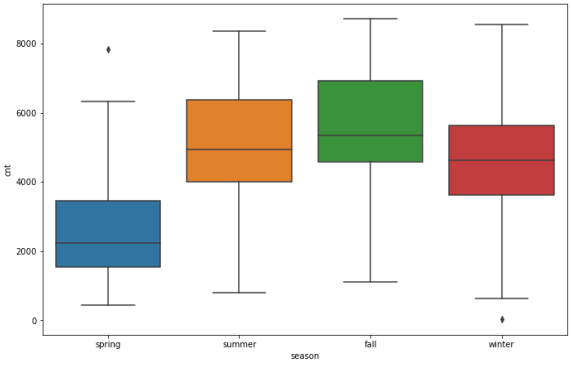
   1. **Season:** Spring season has the lowest rentals and highest rentals is in fall.

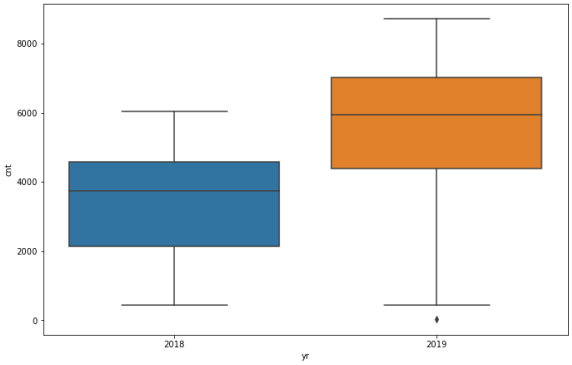
   2. **Mnth:** The demand gradually increases from jan to sept and drops for the next 3 months.

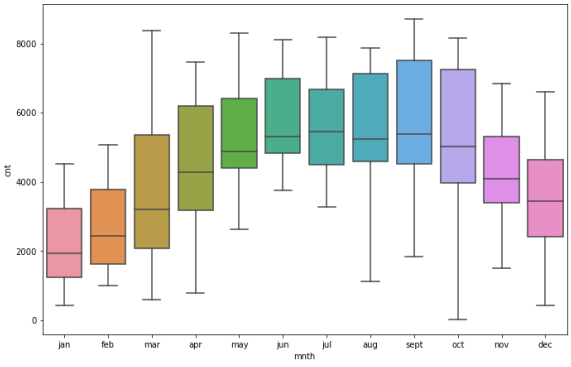
   3. **Yr:** 2019 has the highest demand for Bike rentals.

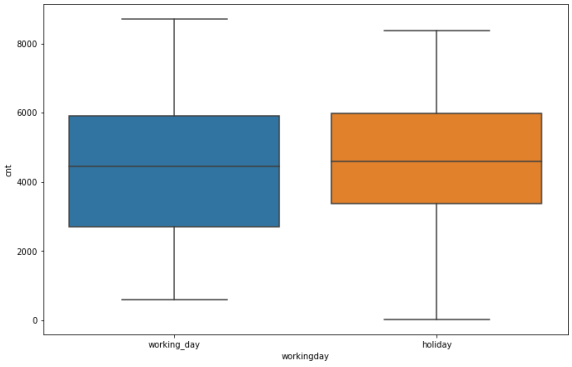
   4. **Weathersit:** The rentals are highest in 'Clear_Few clouds'

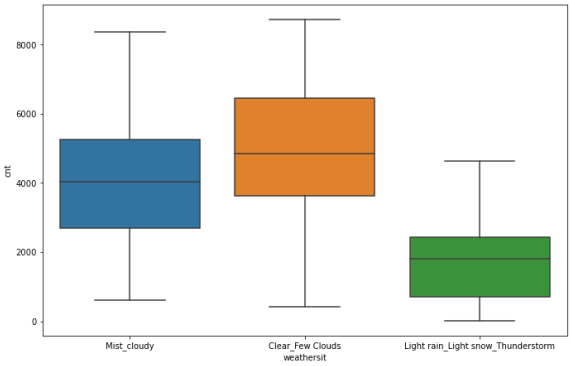
   5. **Workingday:** The number of rentals is high in workingday.

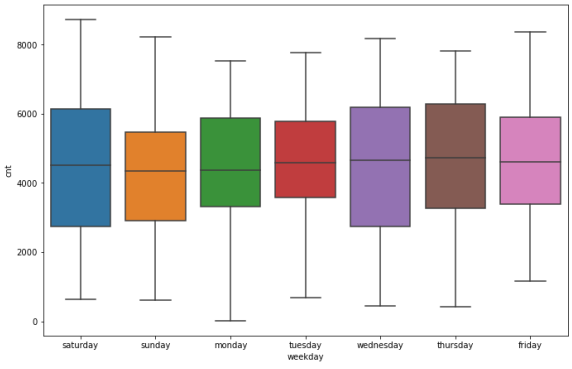












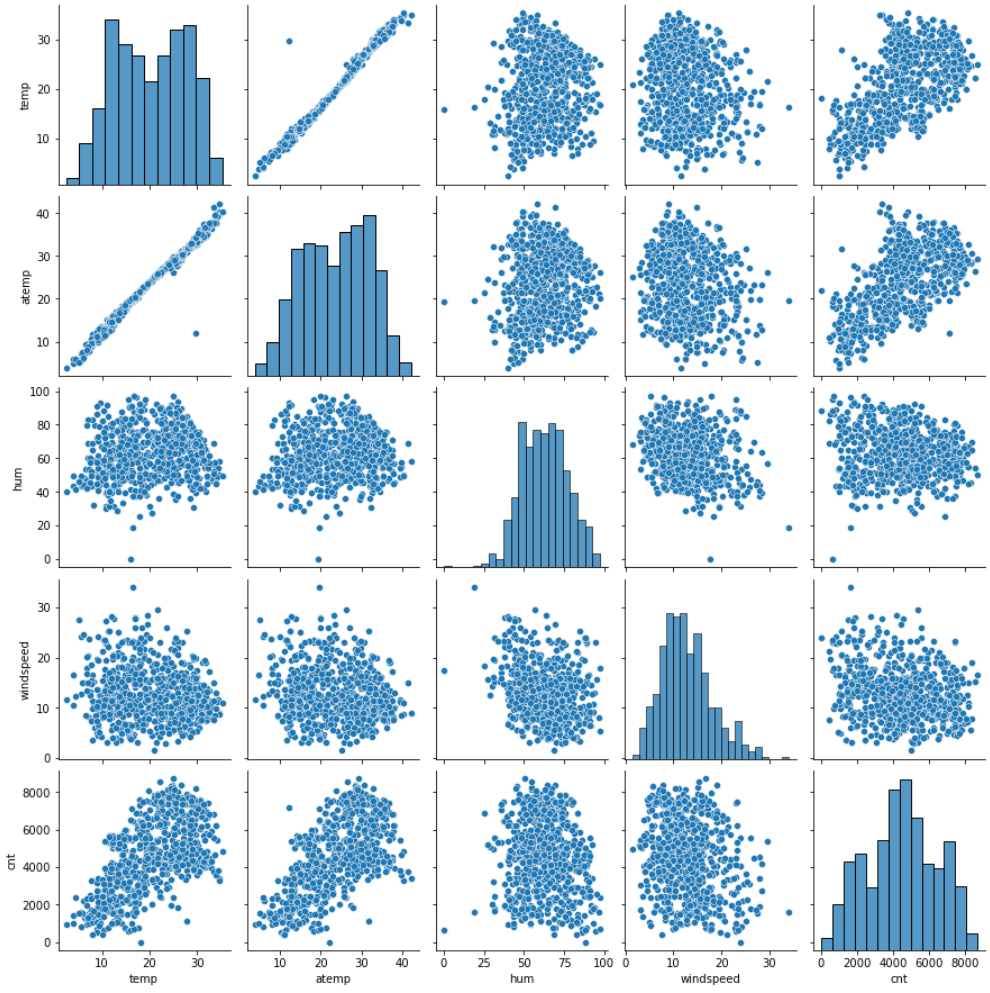
2. **Why is it important to use drop_first=True during dummy variable creation?**

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

Example: Say we have three values (A, B, C) for a particular variable. It can be interpreted as shown below
A-10, B-01, C-00, so only two columns are enough to explain three values.

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

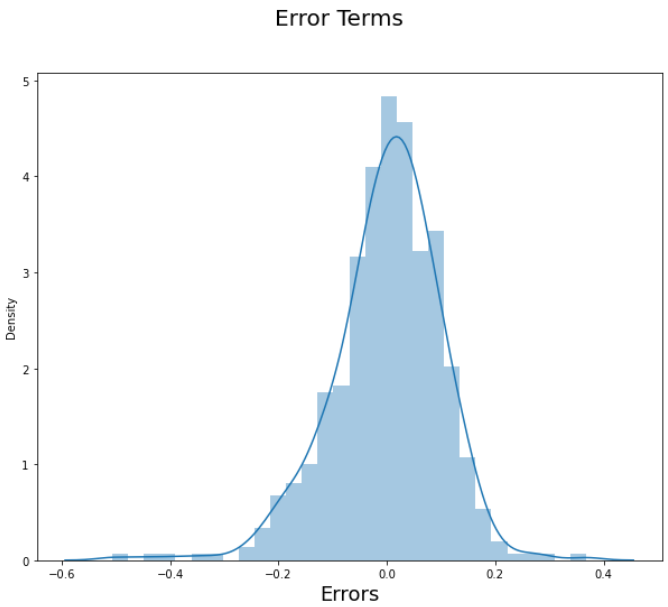
"Temp" and "Atemp" have the highest correlation with the target variable cnt.



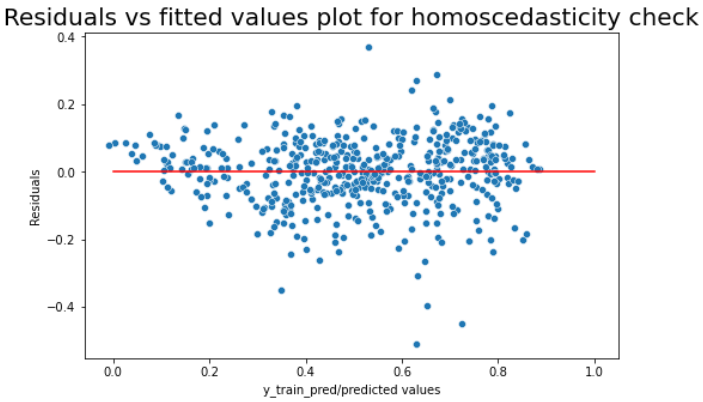
4. **How did you validate the assumptions of Linear Regression after building the model on the training set?**

   It was validated based on the below mentioned points:

a. Residual Analysis: Errors are normally distributed with mean centred at 0.



b. There should be linearity between target and input variables [straight line].
c. Homoscedasticity: Cone shape should not be present and there is no significant pattern observed in the plot.



d. Errors should be independent; this can be checked using the DW value. [in our final model it was: **2.001**].
e. 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?**

   The top three features contributing significantly towards the demand of shared bikes are:
   - **Weathersit**: Light rain_Light snow_Thunderstorm with coefficient 0.3045.
   - **Yr:** Year 2019 with coefficient 0.2468.
   - **Season:** Spring season with coefficient 0.1973.

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Linear Regression is a supervised ML model in which the model finds the best fit line which uses independent variables to predict the dependent variable, provided the output variables to be predicted are continuous.
The model finds the linear relation between predictor and target variables.
We will use straight line plot to predict the possible outcomes for a particular value of independent variable.
We use ordinary least square method to identify the best fit line.
Equation of straight line is y = mx+c, where m is the slope and c is the intercept.
Linear regression can be classified into the following:

- Simple LR: Dependent variable is predicted using only One independent variable
- Multiple LR: Dependent variable is predicted using multiple independent variable

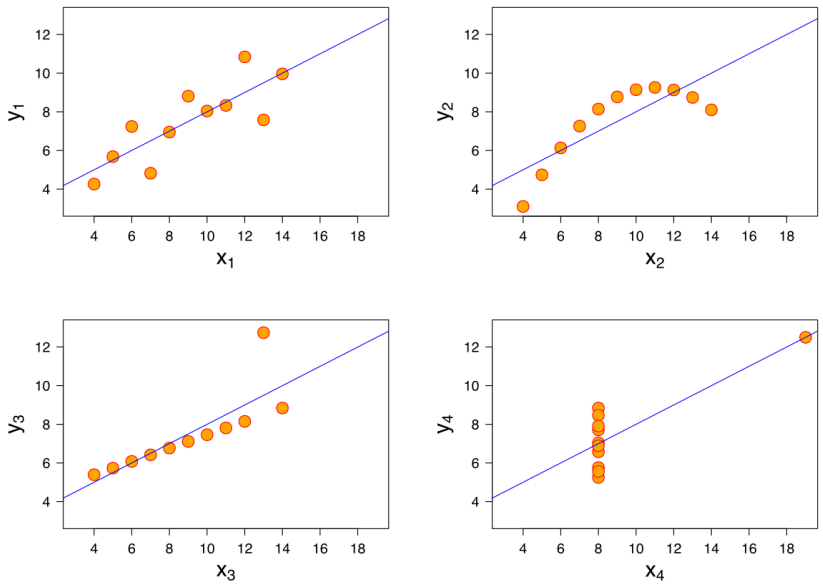
Assumptions of Linear Regression:

a. Error terms are normally distributed and mean is centred at zero.
b. Error terms are independent of each other
c. Error terms have same variance(homoscedasticity)
d. Linear relationship between target and input variable.

2. Explain the Anscombe's quartet in detail.

Anscombe's quartet has four data sets that are identical in simple descriptive statistics, however when plotted they have very different distributions.
It was built by Francis Anscombe to demonstrate the importance of visualizing the data using graphs before model building.

The below figure is plotted using 11 data points.



The statistical information such as mean and std dev for these four datasets are approximately similar.

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

3. What is Pearson's R?

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?

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:
Normalization/Min-Max Scaling: It brings all of the data in the range of 0 and 1. sklearn.preprocessing.MinMaxScaler helps to implement normalization in python.

x= (x – min(x)) / (max(x)- min(x))

Standardization Scaling: Standardization replaces the values by their Z scores. It brings all of the data into a standard normal distribution which has mean (μ) zero and standard deviation one (σ).
sklearn.preprocessing.scale helps to implement standardization in python.

x = (x – mean(x)) / sd(x)

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

   If there is perfect correlation, then VIF = infinity. This shows a perfect correlation between two independent variables.
   In this 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.

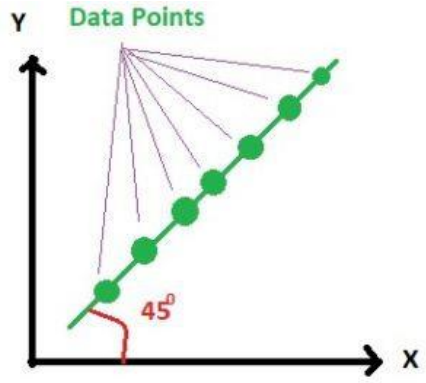
   An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables.

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

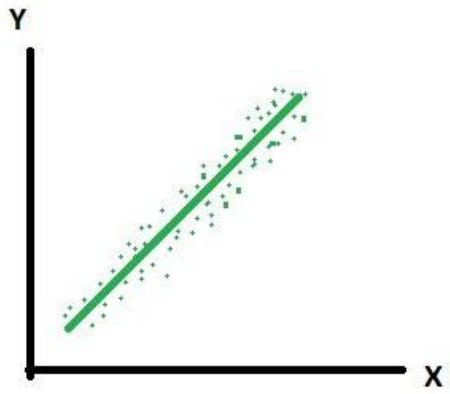
   When the quantiles of two variables are plotted against each other, then the plot obtained is known as quantile – quantile plot or qqplot. This plot provides a summary of whether the distributions of two variables are similar or not with respect to the locations.

   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.



   - And 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


Y
X