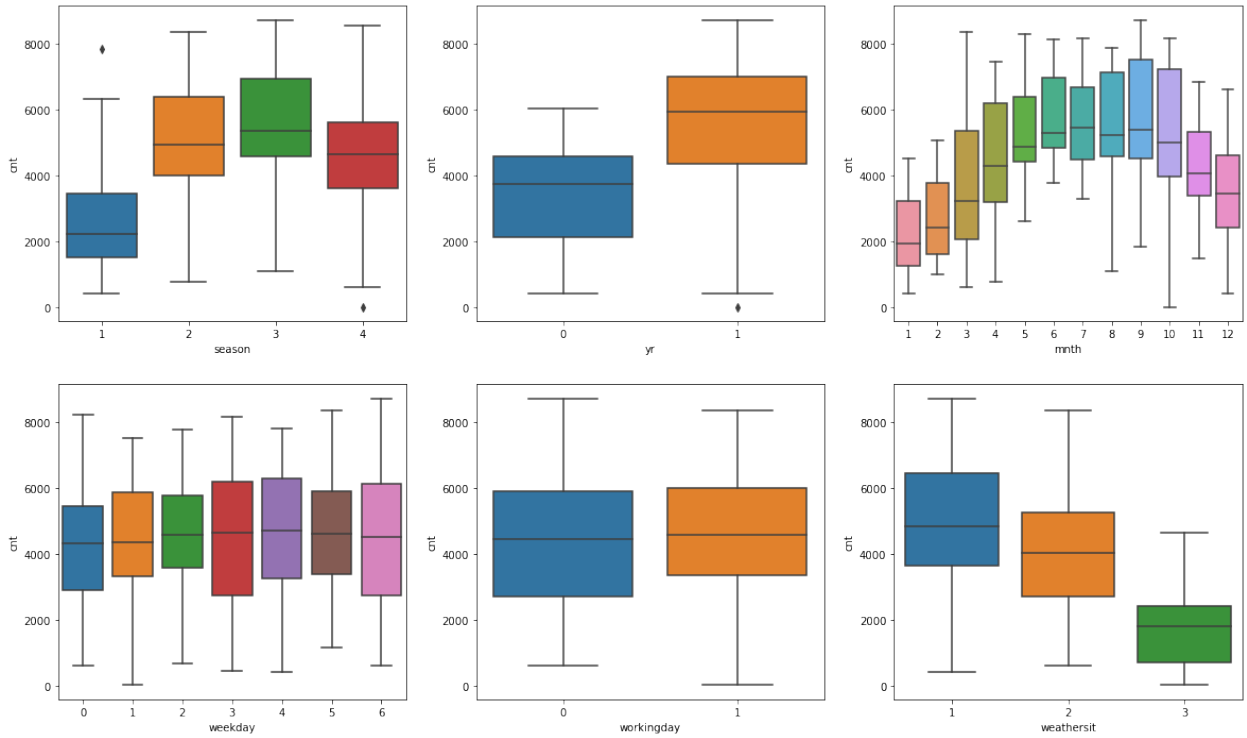
LINEAR REGRESSION QUESTIONS

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?**

Answer:



From above boxplot of categorical variables against count, below observations can be inferred:

- Demand is high in the fall season.
- Demand has grown from year 2018 to 2019.
- More number of customers showed up from June to October.
- Nothing much can be inferred from the box plots of weekday, workingday and count as their median is similar.
- There is a high demand when the weather is Clear, Few clouds, partly cloudy.

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

Answer:

'drop_first' is necessary to generate n-1 variables for a column which has n categorical values. This will reduce the redundant columns.
If drop_first is not used, there would be n variables representing n values.

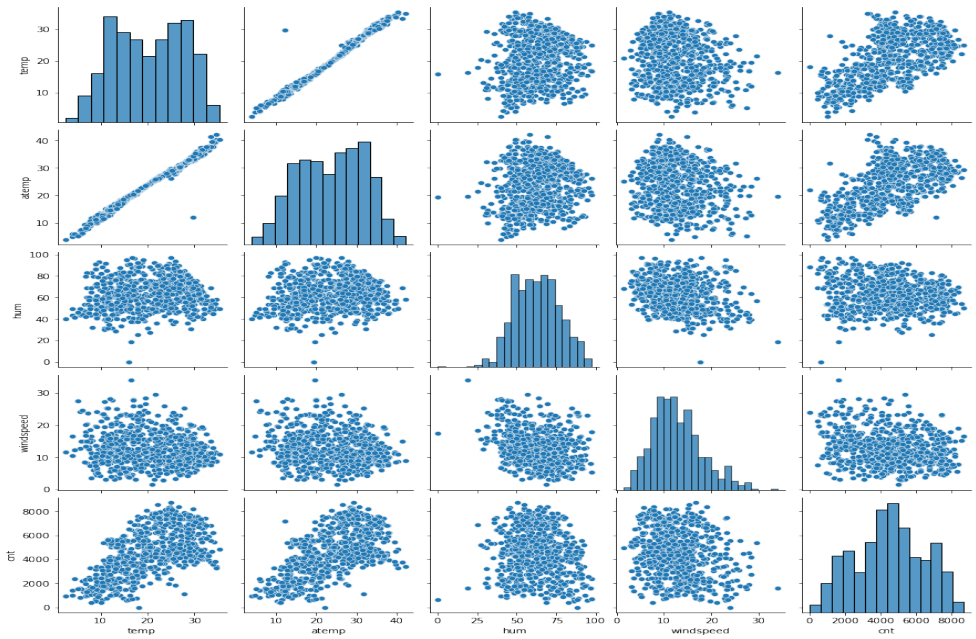
Example: X=[1,2]
Dummy variables without drop_first:

X_1	X_2
1	0
0	1

Here a single column is self explanatory and the other one is not necessary. Having it would make the dataset redundant.
X_2 =0 means X=1 and X_2=1 means X=2
If drop_first=True is used, X_1 column will be dropped and only X_2 will be returned.

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

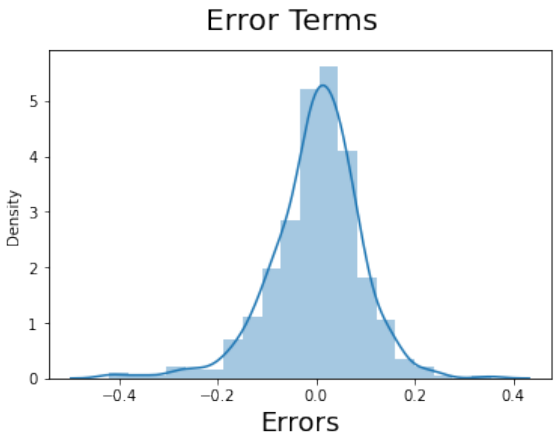
Answer:



From the above pair plot, ‘atemp’ and ‘temp’ seem to have the highest correlation with target variable ‘cnt’.

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

   Answer:



   In the above graph, we can validate one of the major assumptions that Error terms follow a normal distribution and the mean lies at 0.
   Also, from the results I was able to validate that the model has a linear relation.

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

   Answer:
   Temp, season, weathersit are the top significant features which help in determining the demand of shared bikes.

General Subjective Questions:

1. **Explain the linear regression algorithm in detail.**

   Answer:

   Linear regression is one of the supervised learning techniques. It is a predictive algorithm. It involves considering some variables out of which one must be a dependent variable and the other are independent ones. As the name says, it gives out a linear relataionship between all the independent or predictor variables and the dependent or target variable. It helps in finding the correlation among the variables. In a case where there is a single independent variable, the type of regression performed on it would be simple linear regression. When there are multiple independent variables, it is called multiple linear regression.

   After building a suitable model using the input variables, a line is achieved which is a best-fit. It explains the relationship between the variables. Mathematically, the line is represented by y = mx + c, where m is the slope and c is the y-intercept. In case of n input variables, the line is represented as y = c + mx1 + mx2 + mx3 + …… + mxn.

   Example:

   Given, a dataset of symptoms and the diseases that the patient could possibly have, the symptoms would be the predictor variables and disease would be the target variable.

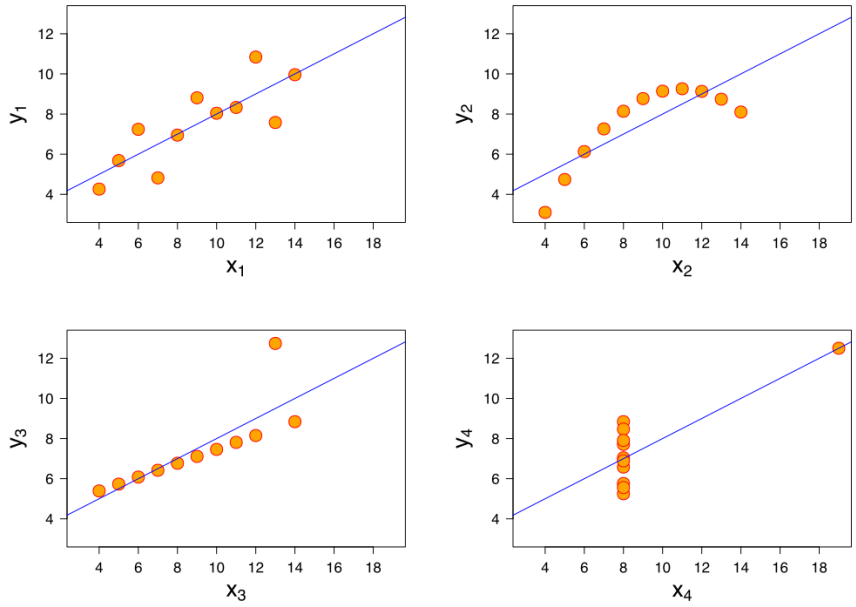
   Using this information, we could build a model which can predict the disease the patient might have, using the set of symptoms.

2. **Explain the Anscombe's quartet in detail.**

   Answer:

   It is a group of four datasets which have almost same statistical values by looking at which one can presume that the graphs and information produced by them would be similar. But, when the four are graphed, they all produced very different distributions and unexpected observations. It seems that the scientist who produced this wanted to prove this point wrong**.**

"Numerical calculations are exact, but graphs are rough."
In the below picture, following inferences can be made:
- 1st graph has a linear relationship
- 2nd graph does not have a linear relationship
- 3rd graph cannot have a linear relation with low residual errors due to the presence of outliers
- 4th graph has many points concentrated at a single point which also has high residual errors.



3. What is Pearson's R?

Answer:
Pearson's correlation coefficient(R) is the most used descriptive statistic to measure the relation between two variables.
- When R is between -1 and 0 then it is said to have a negative correlation
- When R is between 0 and 1 then it is said to have a positive correlation
- When R is 0 then it is said to have a no correlation

Formula:

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

4. **What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?**

   Answer:

   Scaling:

   Scaling is done to put all variables on a common scale. All the data points are squeezed into a defined interval so that going forward there will be no discrepancies in model building and evaluation.

   In normalized scaling, minimum is mapped to 0 and maximum value is mapped to 1.

   Formula:

   Z = x-mean(x)/max(x)-min(x)

   In standardized scaling, the values are put into an interval, but they are transformed so that the mean is 0 and standard deviation is 1

   Formula:

   Z = X-mean/S. D

   Differences:

   - Normalization has an interval but standardization does not.

   - Normalization is used when the data distribution is now known. Standardization is used when data distribution is known.

   - Normalization is considered when the algorithms do not make assumptions about the data distribution. Standardization is used when algorithms make assumptions about the data distribution.

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

   Answer:

   Formula of VIF: 1/(1-R^2)

   Therefore when R-squared value is 1, VIF will become infinity.

   Variance Inflation Factor is a measure of the relationship between two independent variables.

   VIF = 1 indicates a perfect correlation between the variables. It says that the variable can be expressed exactly by a linear combination of other variables. It causes a perfect multicollinearity.

   VIF < 5 is acceptable.

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

   Answer:
   Q-Q plot also known as Quantile-Quantile plot which is plotted with on theoretical quantiles X-axis and sample quantiles on y-axis.

   Use:
   - Looking at the Q-Q plot one can say the distribution of the dataset.
   - It can be used to validate the assumption that the data follows a normal distribution. The points in the scatter plot fall on a straight line if the distribution is normal.
   - It is also used to see whether two distributions are similar or not.

   Importance:
   - In linear regression, it can be used to check whether the training and testing data follow the same distribution or not
   - It can be used to check many common aspects in 2 models like shape, distribution, outliers, change in statistical values etc.