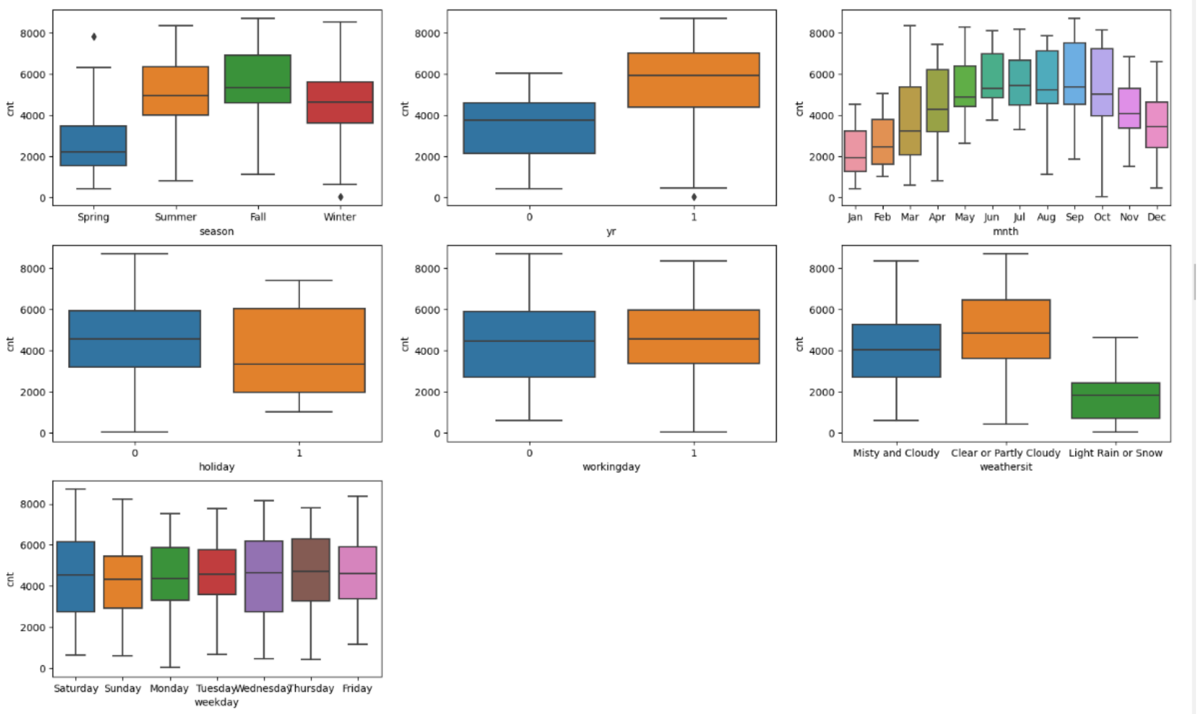
Assignment Based Questions

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

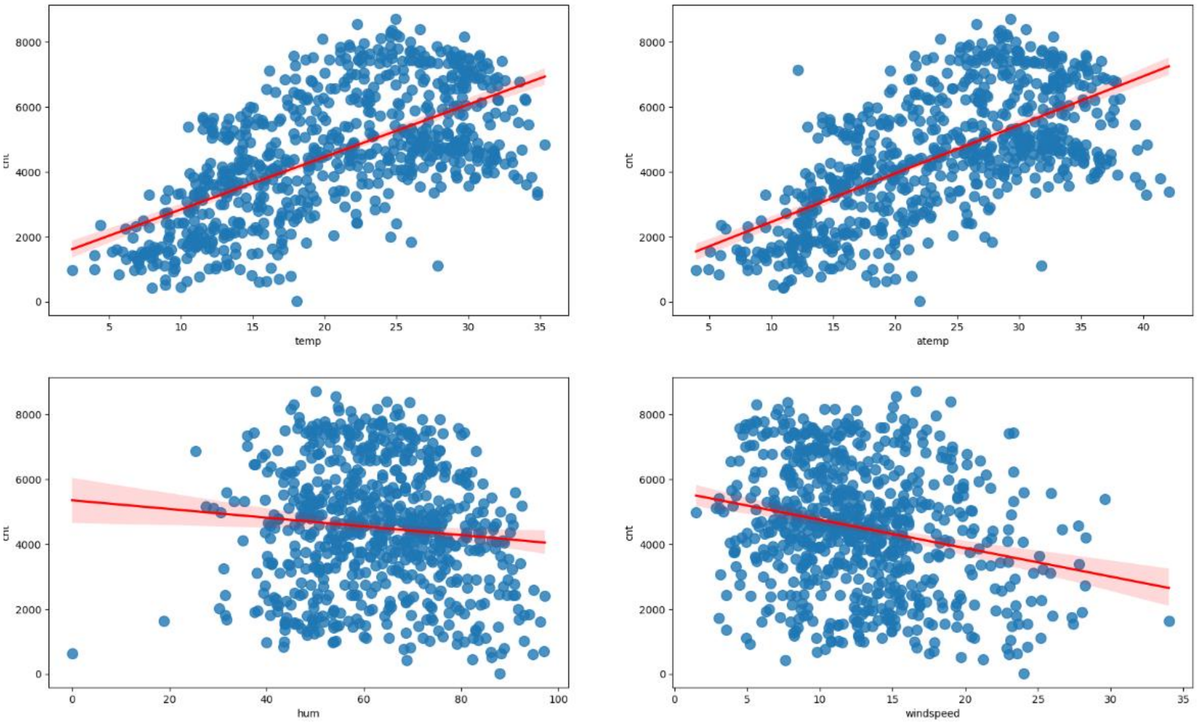


- The maximum demand is in fall, followed by summer and winter
- 2019 has had a increase in business (median 6000) compared to that of 2018 (median 4000).
- The pattern of month on month demand almost forms a bell curve/normal distribution which again signifies that highest demand occurs in fall and summer. (May to October Highest demand)
- Demand is highest on clear days and lowest on rainy/snowy days and almost no demand during heavy snow.
- holidays have slightly lower demand than that of weekday.

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

The drop_first=True should be in place to avoid dummy variable multi collinearity among the derived dummy variables. Since the independent variables have to be truly independent for the model to be perfect, and the dummy variables are derived from the first categorical variable which will skew the model and increase the correlation among these variables.

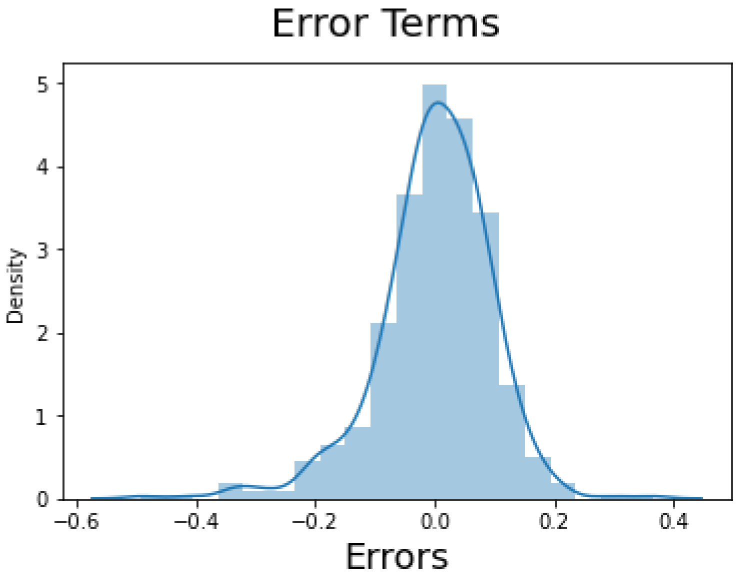
3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?



Temperature has the highest correlation to the target variable.

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

The assumption is that the dependent variable has linear correlation with the independent variables. This can be confirmed by plotting the residuals. If it is normally distributed we can conclude that linear regression is valid.



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

The top 3 features are:

1. temp (Temperature): Coefficient = 0.4502. For every one-unit increase in temperature, cnt is expected to increase by approximately 0.4502 units.

2. weathersit_Light Rain or Snow: Coefficient = -0.2886. In light rain or snow weather conditions, cnt is expected to decrease by approximately 0.2886 units.

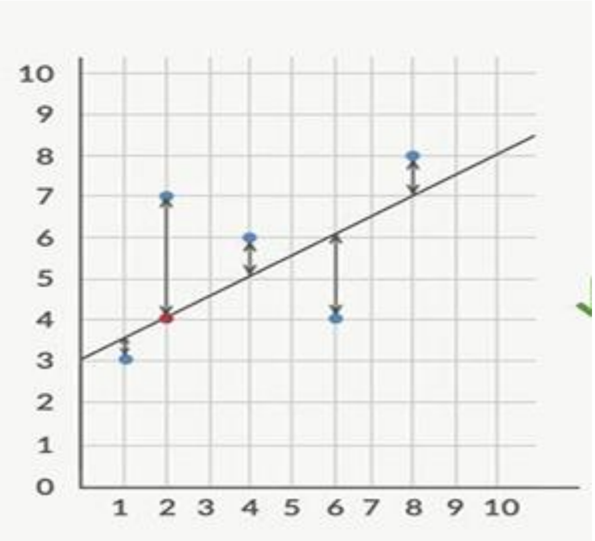
3. yr (Year): Coefficient = 0.2340 : For each additional year, cnt is expected to increase by approximately 0.2340 units. holiday: 0.0423

General Subjective Questions

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

A linear regression algorithm is when trying to predict out a dependent variable using independent variable(s) based on the existing data from a sample dataset.

- If we plot a scatter plot between two variables, say x and y; then for every x there is a y value at a given instance. After plotting the graph with the existing dataset, if we see a pattern and it's not completely random. Some patterns are as follows:
  - When the value of x increases, y value increases (or vice versa)
  - When the value of x increases, y value decreases
- The goal is to pass a line through this scatter plot, or find the line called as the best fit line.
- The best fit line is the least erroneous when compared other passing lines through the scatter plot.



In a single linear Regression, the equation of a line would be y=mx+c and we try to find the m and c.

Whereas in a multiple linear regression model, the algorithm tries to get the coefficients that result in the least erroneous line.

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

Anscomb's quartet is a set of four identical dataset with identical descriptive statistics but varies so much when graphed.

Anscombe's quartet is a powerful illustration of the importance of visual exploration in data analysis and the limitations of relying solely on summary statistics to understand the nature of data.

3. **What is Pearson's R?**

Pearson's R is the numerical summary of the strength of the linear association between variables. It varies between -1 and +1.

R = 0 means no linear association between variables.

R= 1 implies strong positive slope.

R = -1 implies perfect negative slope.

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

Scaling is the process of bringing all the numerical values in a dataset to a standard scale.

Scaling is performed to bring every numerical value to the same platform so that each variable doesn't influence each other or the target variable differently.

- It avoids sensitivity.

Normalized Scaling brings the entire dataset within a particular range, usually [0,1]

Formula is Xnorm = (X - Xmin)/(Xmax – Xmin)

Standardized Scaling is to transform the dataset to have a mean of 0 and standard deviation of 1.

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

If the R^2 is 1 the VIF value would be infinite.

Which means that the dataset might have been duplicated, the dummy variable trap or the perfect fit.

Or if the algorithm memorized the dataset.

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

A Q-Q (Quantile-Quantile) plot is a graphical tool used to assess whether a set of observed data follows a particular theoretical distribution, such as the normal distribution.