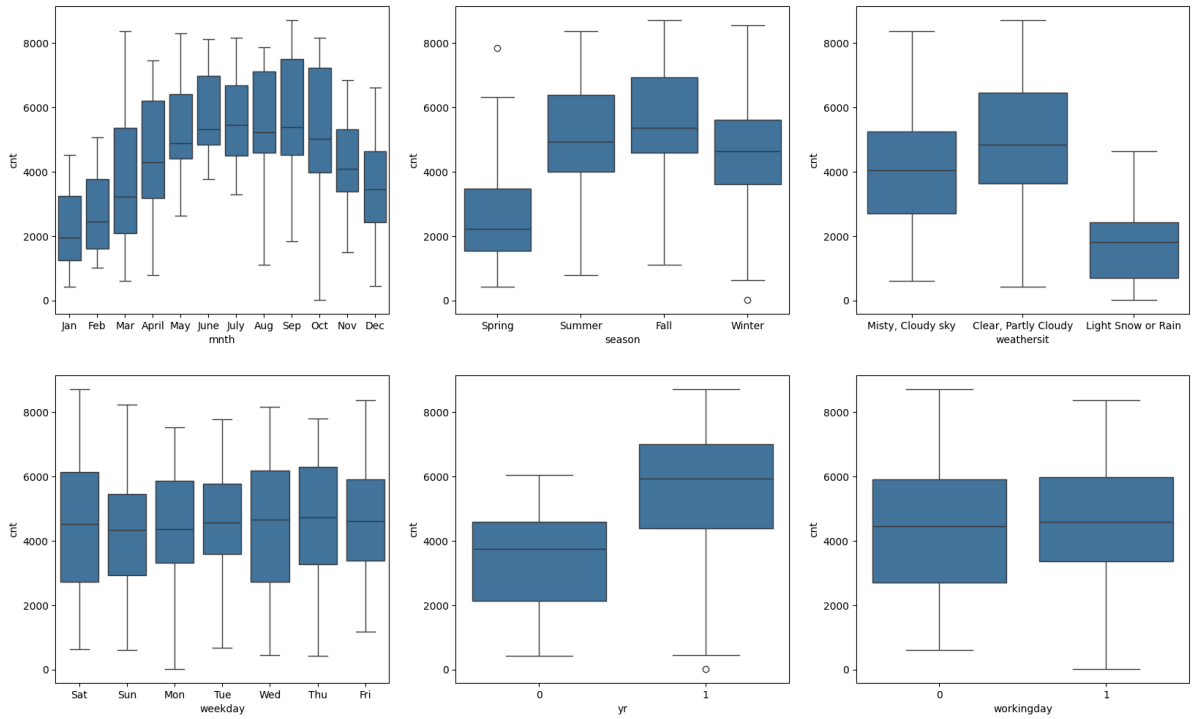
Assignment-based Subjective Questions

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

- There were more users when the weather was clear, average under cloudy conditions and least when it rains.
- The users signed up / used has increased from 2018 - 2019
- The demand was least in spring, this can be seen from the month wise data as well.

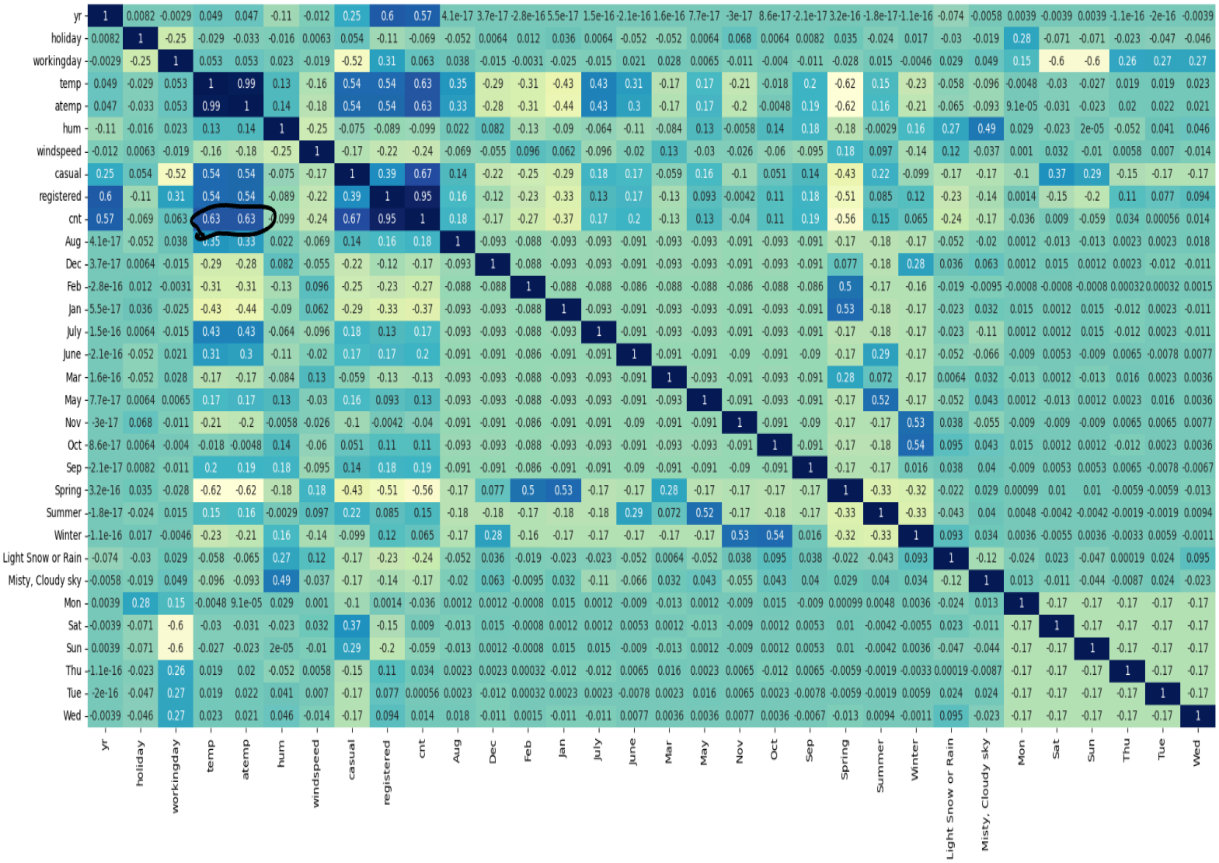


2. Why is it important to use drop_first=True during dummy variable creation?

- This is to not add the columns that do not offer additional information. For example, if there are 4 seasons, Spring, Summer, Fall and Winter, if it is not Summer / Fall or Spring, it must be Winter so we do not need to waste more space conveying the same message.

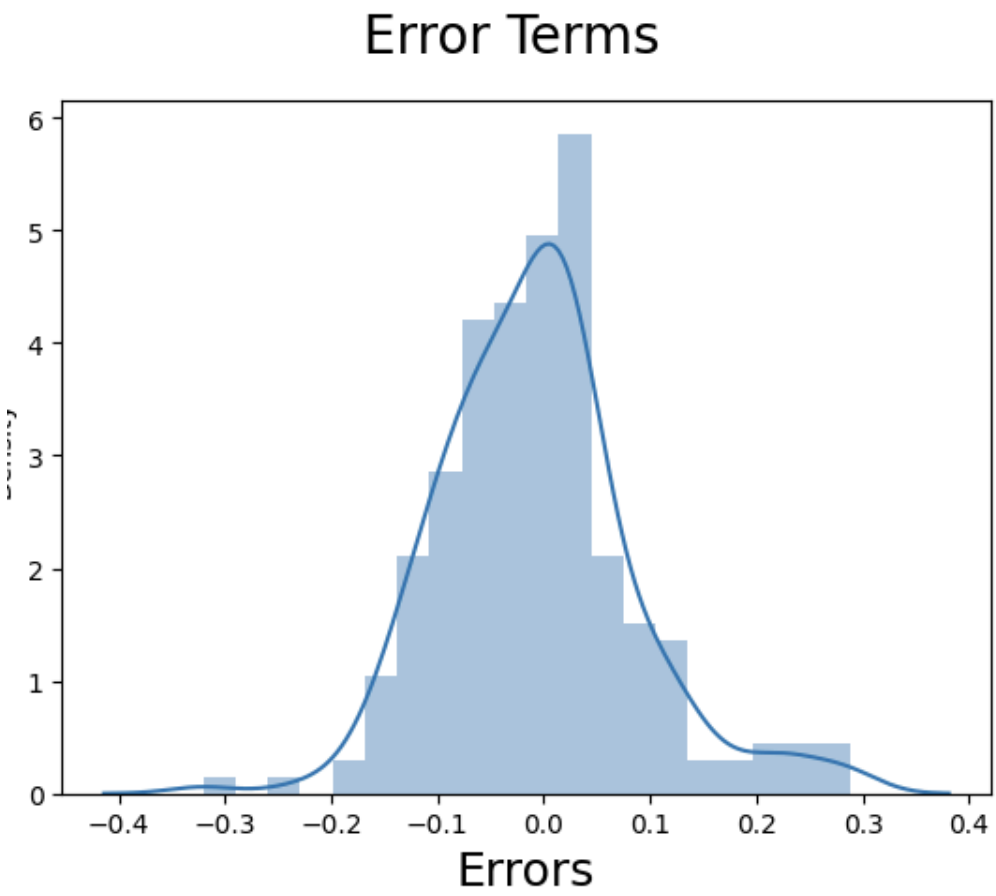
3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

Tmp and Atmp has the highest corelation, if we remove the obvious casual and registered since cnt is sum of them.



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

- The error was normally distributed with mean around zero.



- R-sqaured of the predicted values was 0.82 which is decent.

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

- Temperature had positive correlation
- Weather - especially if it was Raining / Snowing had a negative correlation
- Windspeed also had a positive correlation if we do not consider ‘year’.

General Subjective Questions

1. Explain the linear regression algorithm in detail.

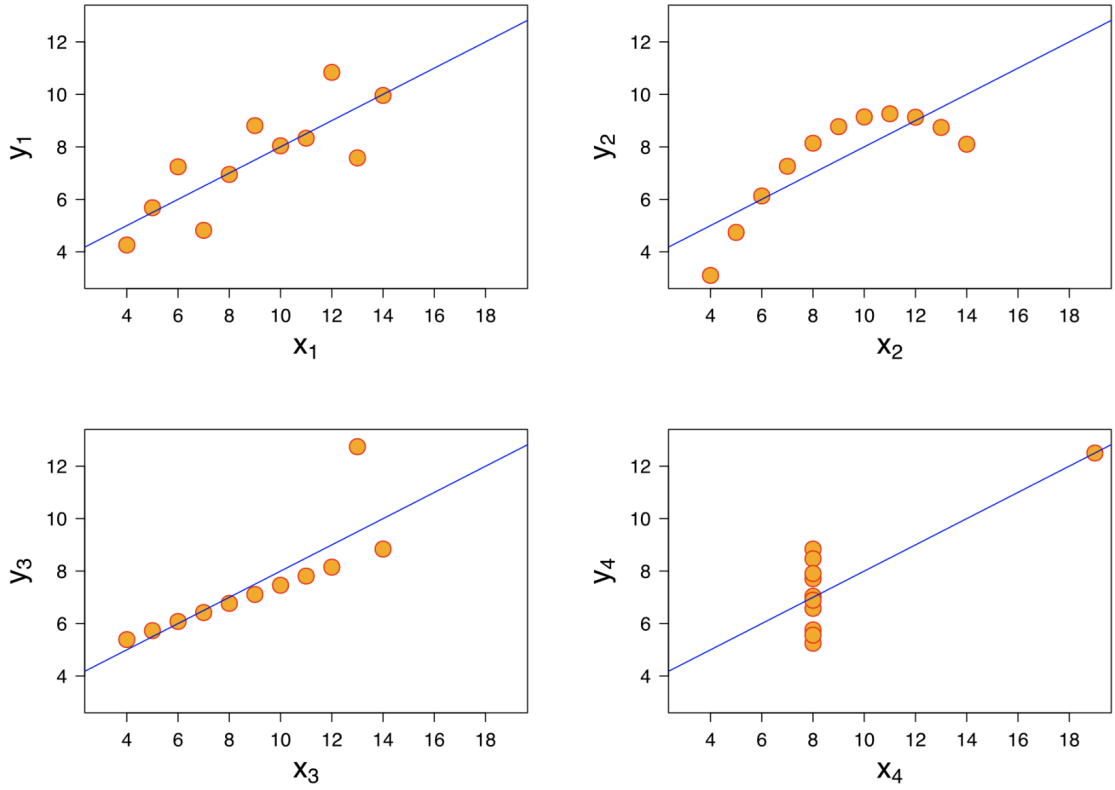
Linear regression is a statistical model which estimates the linear relationship between a scalar response and one or more explanatory variables.The algorithm tries to fit the

values represented by a straight line. There are two types Simple and Multiple Linear Regression. It involves following steps.

- Reading and understanding the data
- Visualizing the data to identify relationships
- Data cleansing and preparation
- Split to training / test set
- Train the model.
- Evaluate the model.

2. Explain the Anscombe's quartet in detail.

Anscombe's quartet comprises four data sets that have nearly identical simple descriptive statistics, yet have very different distributions and appear very different when graphed. Each dataset consists of eleven (x, y) points.



3. What is Pearson's R?

Pearson's R, also called the Pearson correlation coefficient, is a numerical value that is used to assess the **linear relationship** between two continuous variables. It's a number between -1 and 1. Positive number mean a positive correlation. The closeness to 1 indicates the strength of correlation.

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

Scaling:

Scaling is a process of bring the values of all the variables in to a common range.

Why is Scaling performed:

This is to improve convergence of the data and provide equal footing to avoid larger valued data from influencing the mode.

Standardized vs Normalized:

Normalization scales the features to a range between 0 and 1 (or sometimes -1 and 1), using the min and max values.

Standardization scales the features by subtracting the mean and then dividing by the standard deviation.

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

This can happen when there is a perfect multicolinearity. Since VIF is calculated based on R-squared, if R-squared is 1, VIF can be infinity.

It is most likely because of unreliable coefficients or misleading interpretations. For example we might have assumed two variables as independent when they were not.

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

Q-Q plot(Quantile-Quantile plot) is a tool used to assess the distribution in linear regression. It helps us visually compare the distribution of the residuals (errors) in the model. A point (x, y) on the plot corresponds to one of the quantiles of the second distribution (y-coordinate) plotted against the same quantile of the first distribution (x-coordinate). This defines a parametric curve where the parameter is the index of the quantile interval.