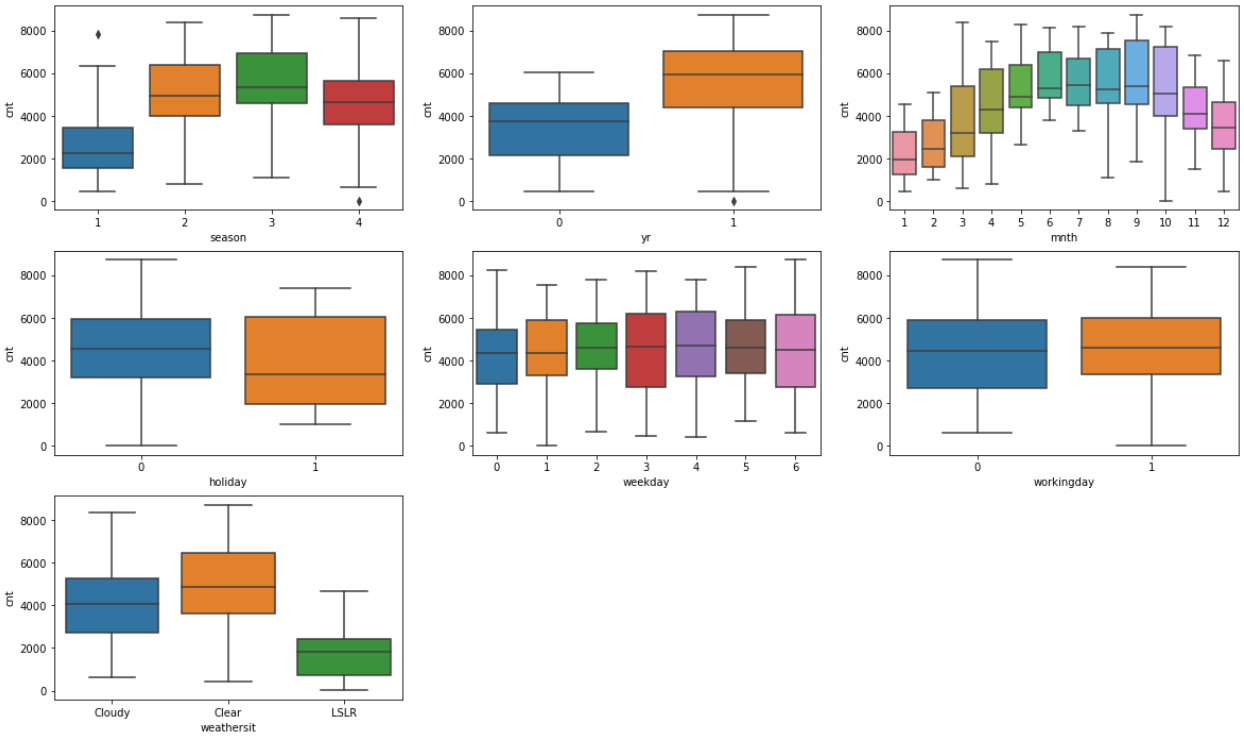
Assignment-based Subjective Questions
Sudeep K S ML38 UPGRAD

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

Ans:

We can use the box plots of categorical variables to infer the effects of dependent variables



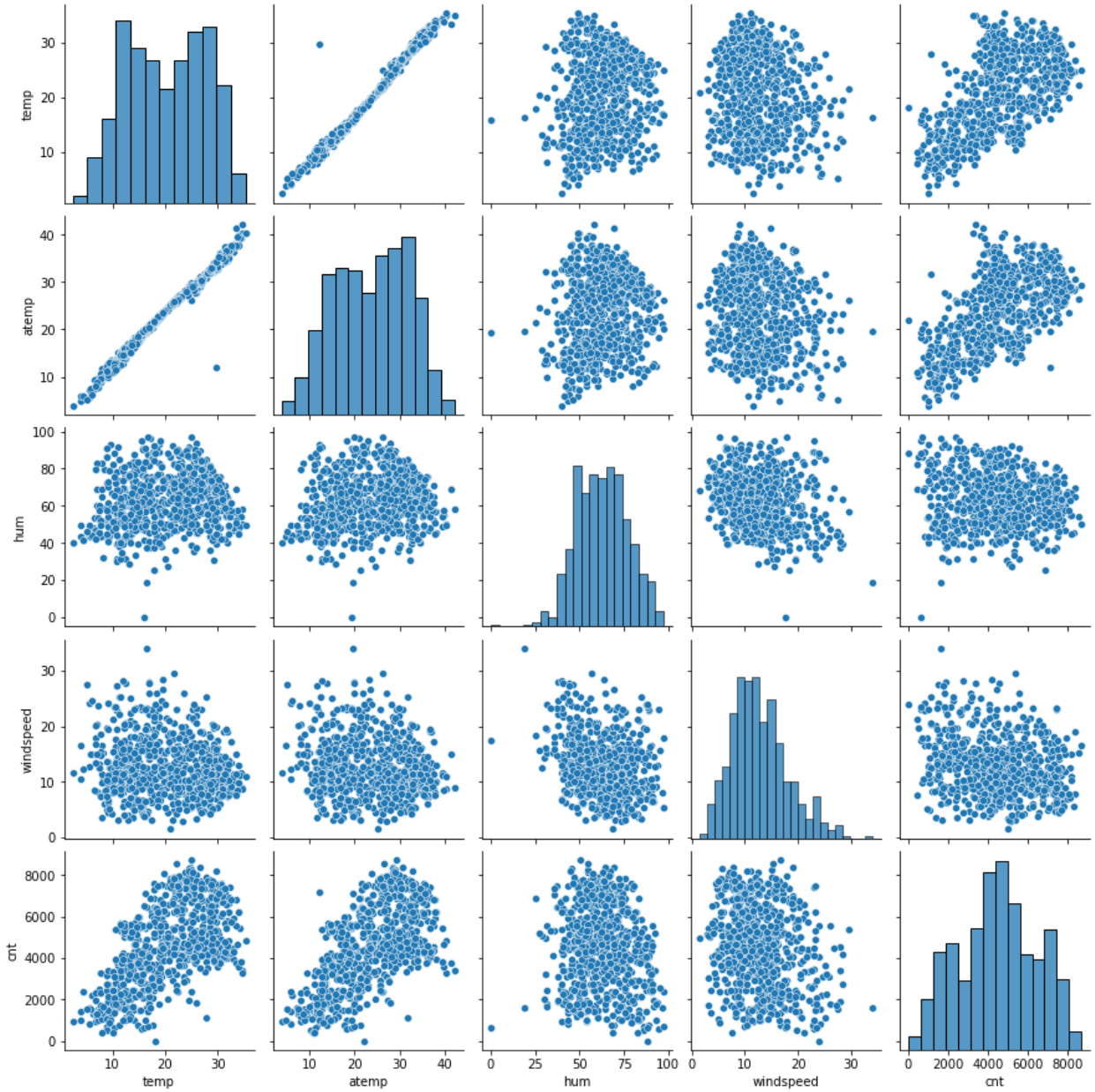
- From the above plots we can clearly see that except weekday and working day all other variables influence cnt variable
- Fall season has the highest number of bookings followed by summer
- The demand peaks in September and is lowest in January.
- The median of the year in 2019 is greater than 2018 which says the cnt variable is more in year 2019.
- The bookings are low during holidays when compared to weekdays
- On Clear days the bookings are high when compared to other days.

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

Ans: When we use one host encoding technique to convert categorical variable into a form that can be used to work. This reduces the extra column created during dummy variable creation, reducing correlation created between dummy variables.

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

Ans: Here is the pair-plot among numerical variables:

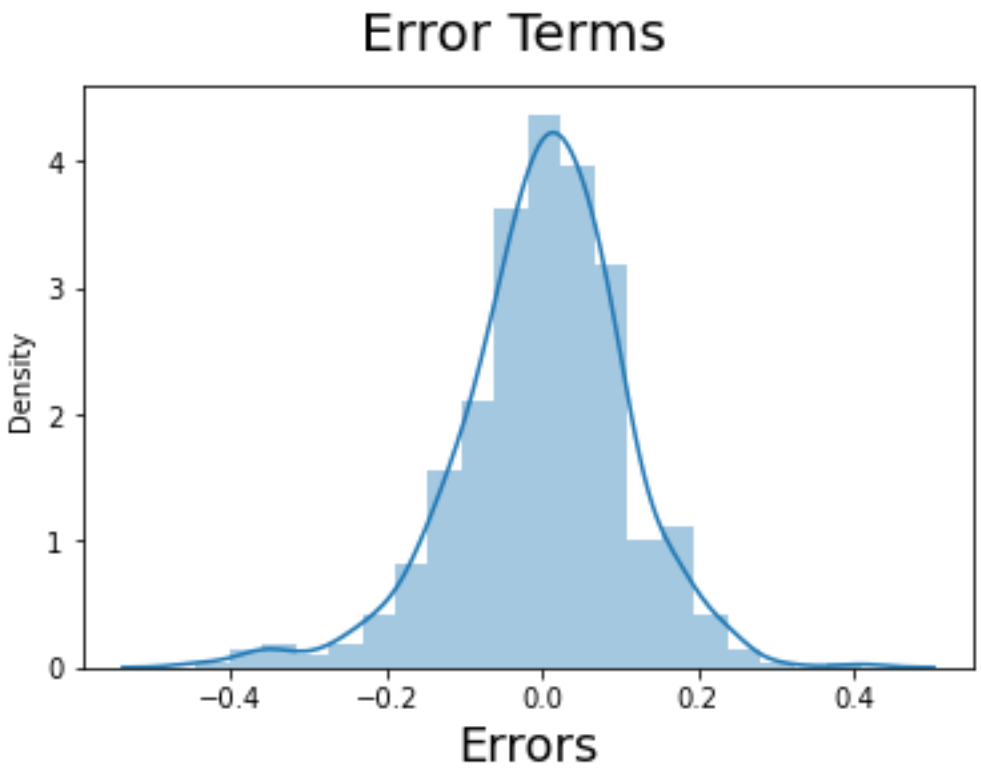


From the above plots we can say that **temp** and **atemp** are highly related **cnt** variable.

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

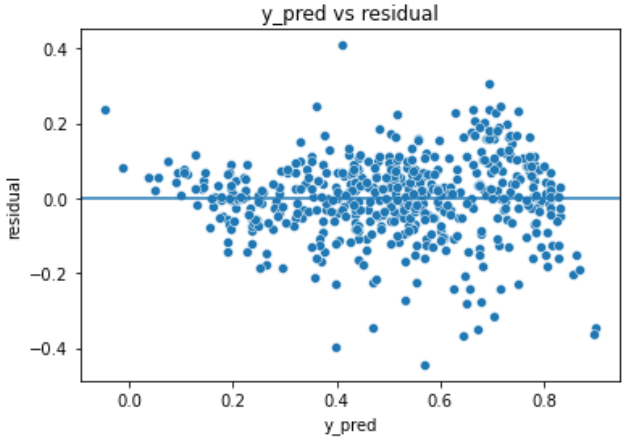
Ans: There are Five assumptions of Linear Regression that was validated:

- **Linear Relationship:** This was verified by scatter plots between dependent and independent variables.
- **Normal distribution of error terms:**

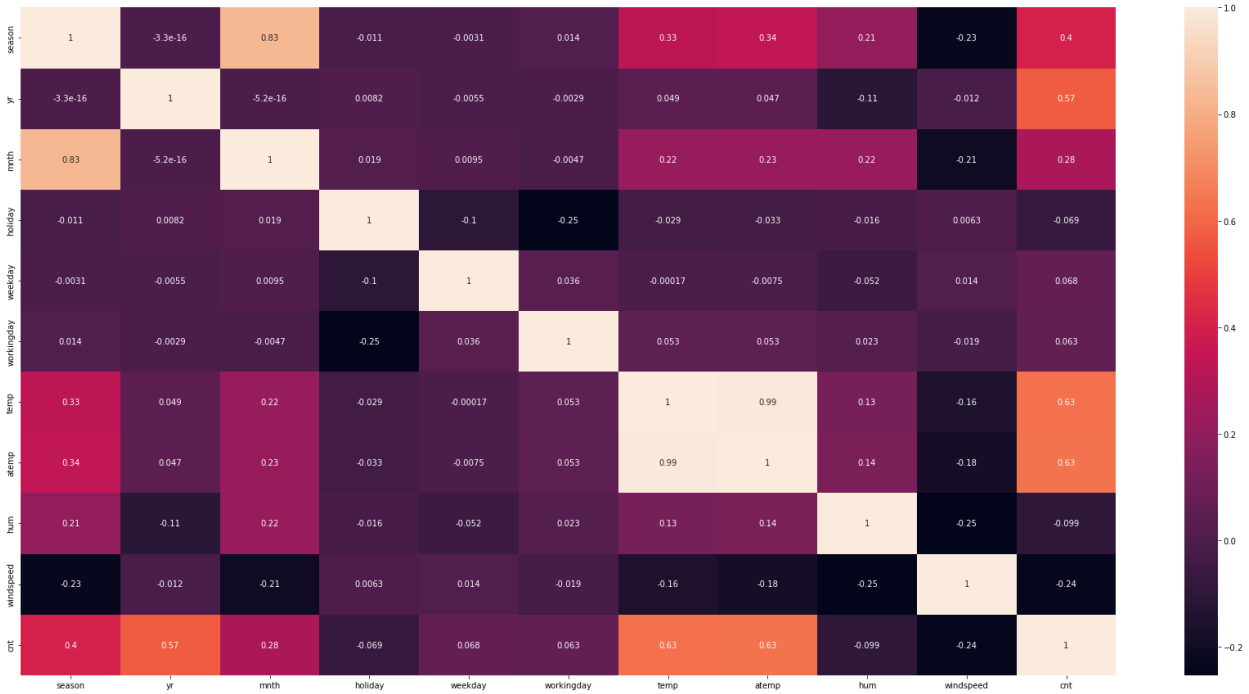


By doing residual analysis on train data, With the above plot we can confirm that there is normal distribution of error term

- **Homoscedasticity:** Homoscedasticity in a model means that the error is constant along the values of the dependent variable. After the plotting of residual vs y_pred plot



- **Multicollinearity**: Using a heat map we can see collinearity between multiple variables



- **Independence of Residuals:** No auto correlation or independence.
  Durbin-Watson Statistic will always assume a value between 0 and 4.
  Below are the interpretations:
  DW = 2: no autocorrelation
  DW< 2: positive correlation
  DW > 2: negative correlation

In our OLS Regression Results of the final model, the DW value is 1.977 which is approximately equal to 2. Hence, we say that there is little or no auto-correlation

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

Ans: By looking at coefficients, top 3 features contributing significantly towards explaining the demand of the shared bikes are:

$$cnt = 0.0892 + 0.2341 \times yr + 0.6381 \times atemp + 0.095 \times seasonwinter - 0.12 \times windspeed - 0.02409 \times weathersitLightSnow$$

- atemp
- year
- weathersit_Light Snow

General Subjective Questions

1. Explain Linear Regression Algorithm in detail.

Ans: Linear Regression is a machine learning algorithm which is based on supervised learning category. It finds a best linear-fit relationship

on any given data, between independent (Target) and dependent (Predictor) variables. In other words, it creates the best straight-line fitting

to the provided data to find the best linear relationship between the independent and dependent variables. Mostly it uses Sum of Squared Residuals Method.

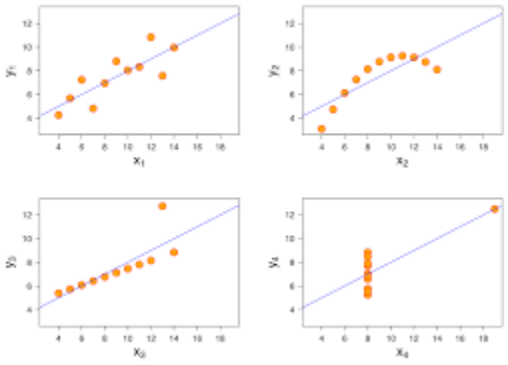
Basic Equation for Linear Regression is

$$Y=\beta 0+\beta 1X1 +\epsilon$$

We use Root Mean Square error as cost Function for Linear Regression. Our aim is to build the Line with minimizing the Cost Function.

2. Explain the Anscombe's quartet in detail.

Ans: Anscombe's quartet comprises four datasets that have nearly identical simple statistical properties yet appear very different when graphed. Each dataset consists of eleven co-ordinates.



3. What is Pearson's R?

Ans: Pearson's R is a statistic that measures the Correlation between two variables. It has a Numerical value which ranges from -1 to +1. Mathematically it is defined as covariance of two variables divide by product of their standard deviations.

If the variables tend to go up and down together, the correlation coefficient will be positive. If the variables tend to go up and down in opposition with low values of one variable associated with high values of the other, the correlation coefficient will be negative.

A value of 0 indicates that there is no association between the two variables. A value greater than 0 indicates a positive association; that is, as the value of one variable increases, so does the value of the other variable. A value less than 0 indicates a negative association; that is, as the value of one variable increases, the value of the other variable decreases.

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

Where,
r = correlation coefficient
xi = values of the x variable in a sample
x bar = mean of x values
yi = values of y variable in a sample
y bar = mean of y values

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

Ans: Feature Scaling is a technique to standardize the independent features present in the data for a fixed range. It is performed during the data pre-processing to handle varying magnitudes or values or units. If feature scaling is not done, then a machine learning algorithm wights for different values will vary drastically.

Most common methods of scaling are Normalization and standardization.

- Normalization scales the values to a range between 0 and 1

  Normalized scaling: x = x-min(x) / max(x) – min(x)

- Standardization scales data to have a mean of 0 and a standard deviation of 1

  Standardized scaling: x = x-min(x) / sd(x)

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

Ans: VIF is calculated by the below formula:

VIFi = 1 / (1 – Ri2) Where***, 'i'*** refers to the ***ith*** variable.

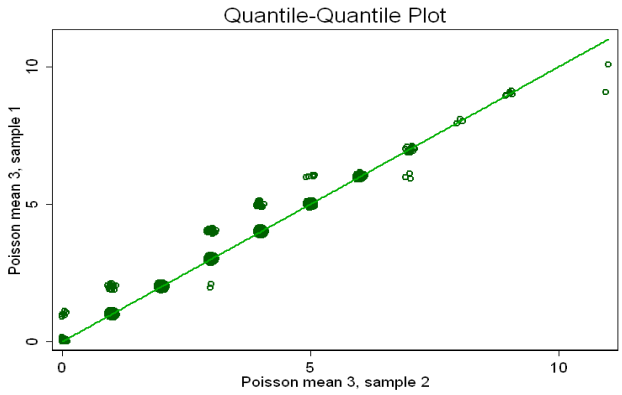
If R-squared value is equal to 1 then the denominator of the above formula become 0 and the overall value become infinite. It denotes perfect correlation in variables

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

Ans:

The Q-Q plot or quantile-quantile plot is a graphical technique for determining if two data sets come from populations with a common distribution. A Q-Q plot is a scatterplot created by plotting two sets of quantiles against one another.

If both sets of quantiles came from the same distribution, we should see the points forming a



An example of a Normal Q-Q plot when both sets of quantiles truly come from Normal distributions.

Use of Q-Q plot in Linear Regression:

The Q-Q plot is used to see if the points lie approximately on the line. If they don't, it means, our residuals aren't Gaussian (Normal) and thus, our errors are also not Gaussian.

Importance of Q-Q plot:

Below are the points:

- The sample sizes do not need to be equal.
- Many distributional aspects can be simultaneously tested. For example, shifts in location, shifts in scale, changes in symmetry, and the presence of outliers.
- The q-q plot can provide more insight into the nature of the difference than analytical methods.