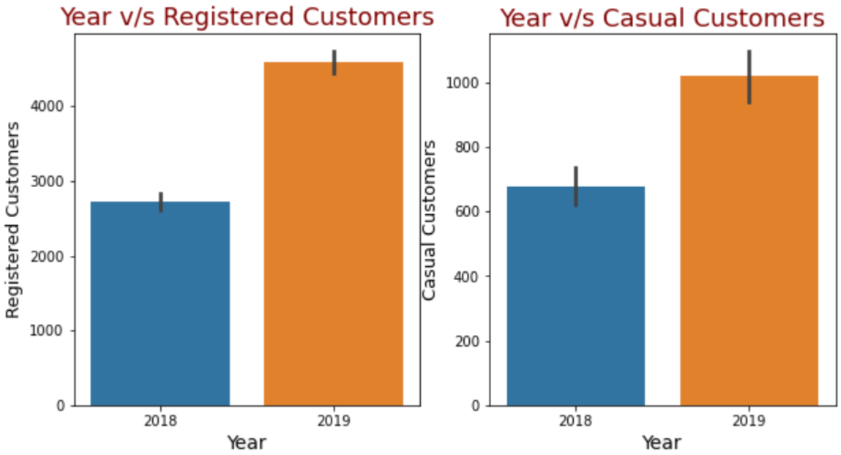
ASSIGNMENT-BASED SUBJECTIVE QUESTIONS

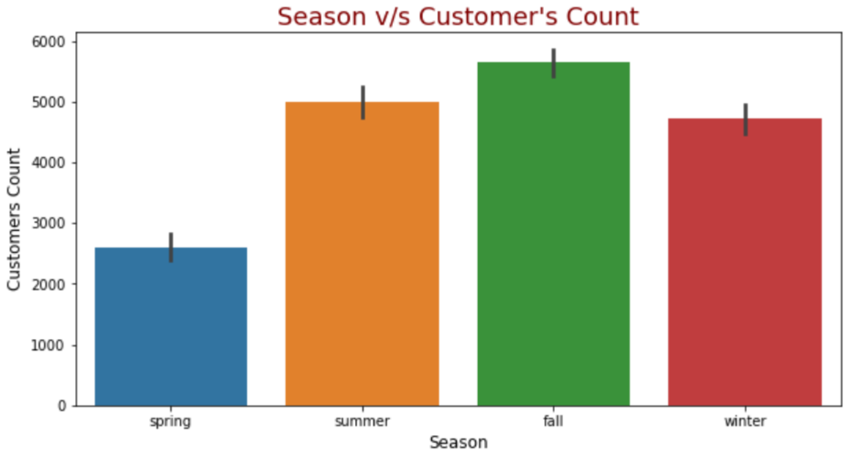
Question 1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

Ans.

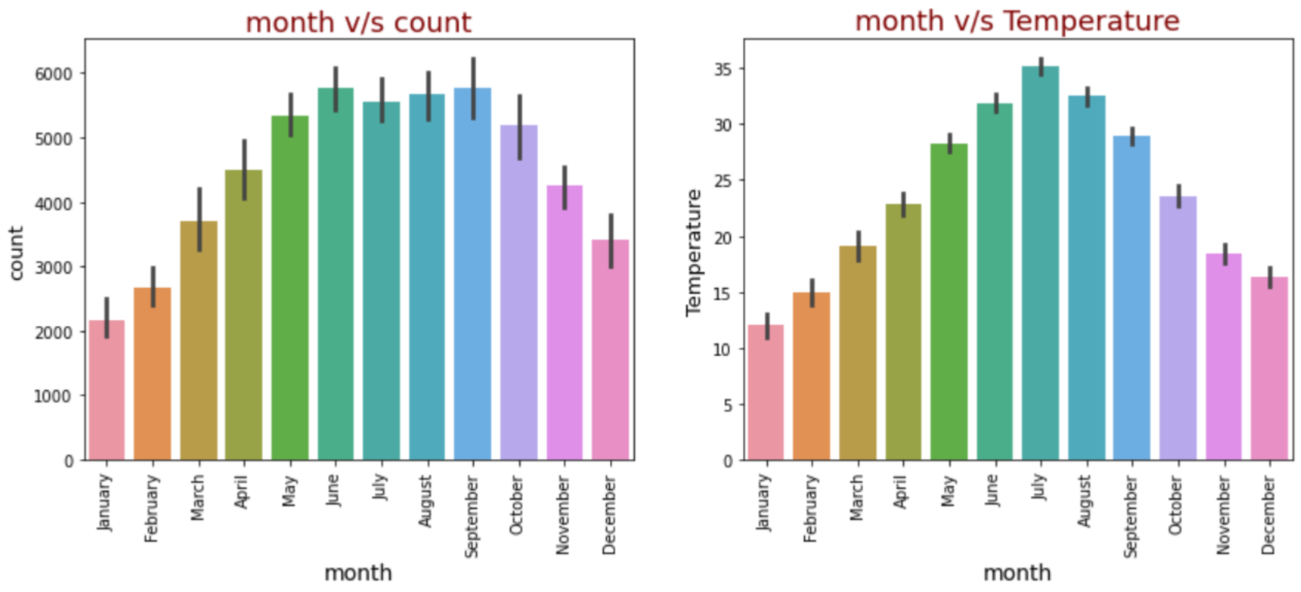
1. 'yr' (year) has positive impact on the bike rentals i.e. every year bike rental sales is increasing as the number of users are increasing.



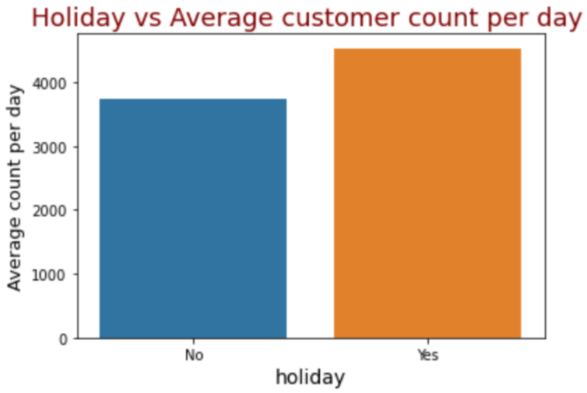
2. In spring season bike rentals reduces may be because in spring season people prefers to walk.



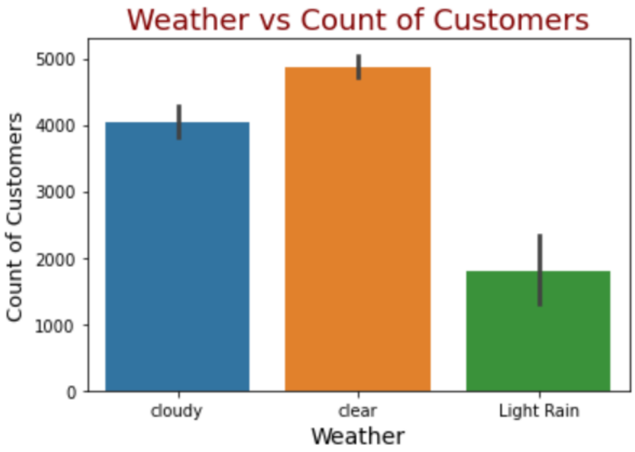
3. In the middle of the year bike rentals increases. This effect is because of the temperature. (I guess the data belongs to country where increase in temperature is good. Definitely not India)



4. On holidays average customer count increases. On holidays people use to rent bike more.



5. In clear weather people use to rent bikes more.



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

Ans: While creating dummy variables for any categorical variable, we need to represent N number of states (values of categorical variable). So, to represent N number of states we only require N-1 variables.

For e.g.: -

Let's say we have a categorical variable "Weather" in our current data set with values

1. Clear
2. Cloudy
3. Light Rain
4. Heavy Rain and Thunder

To represent these states for Weather category we can use:

Clear	Cloudy	Light Rain	Heavy rain
1	0	0	0
0	1	0	0
0	0	1	0
0	0	0	1

We can observe here that none of the combination having (0,0,0,0) as its value so we can remove one of the dummy variable and assign (0,0,0) as its value.

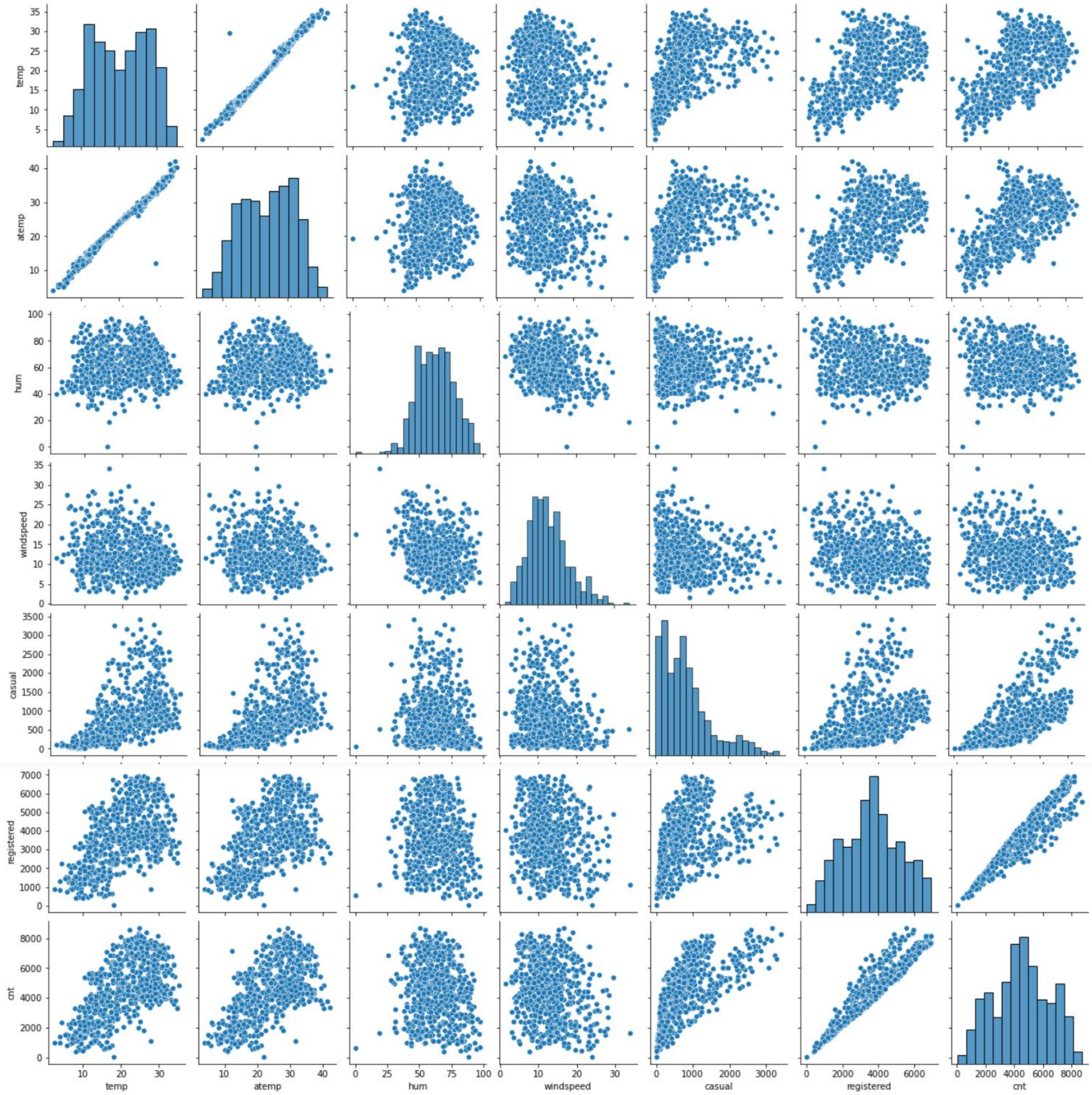
So now after removing one variable we can represent all four variables as:-

	Clear	Cloudy	Light Rain
Clear	1	0	0
Cloudy	0	1	0
Light Rain	0	0	1
Heavy Rain →	0	0	0

And to do that we can use drop_first=True in the Python code.

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

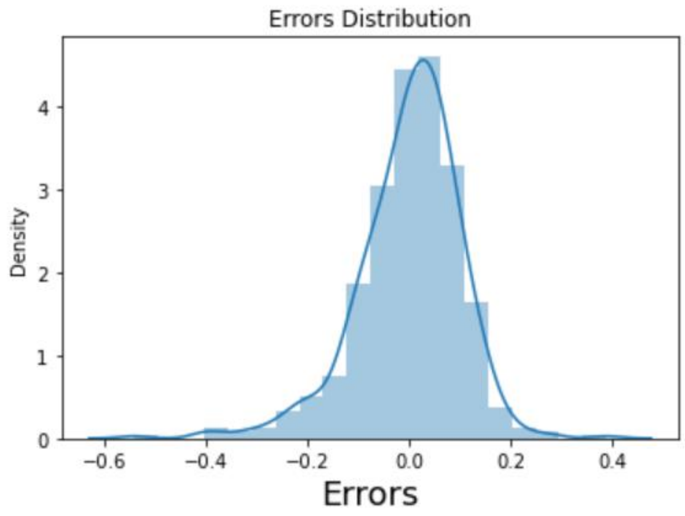
Ans: registered variable has highest correlation with target variable **cnt**.



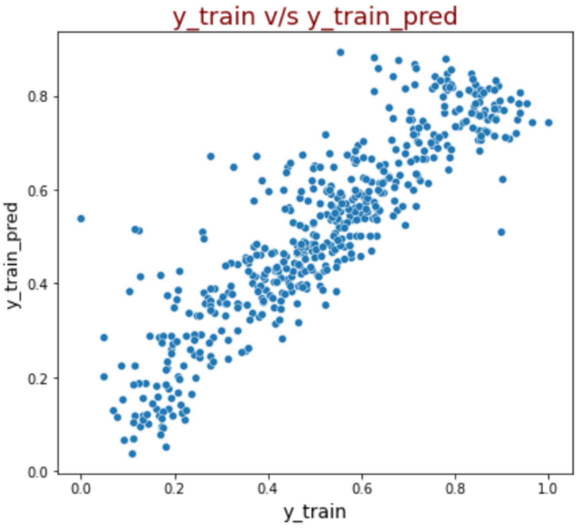
Question 4. How did you validate the assumptions of Linear Regression after building the model on the training set?

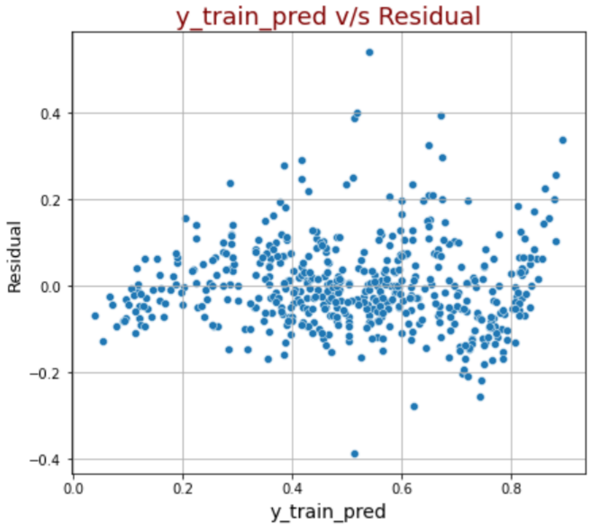
Ans:

1. We assumed while creating the model that the error terms are normally distributed. To verify this, we plot the Residuals (error between original and predicted values) of the Training set. If that distribution is normal then this verifies that our assumption was correct.



2. We assumed that the variance of Residuals (errors) is constant. So, we can plot the residuals and if the plot doesn’t shows any pattern and it seems that the residuals are dispersed evenly around origin then it confirms that our assumption was correct.

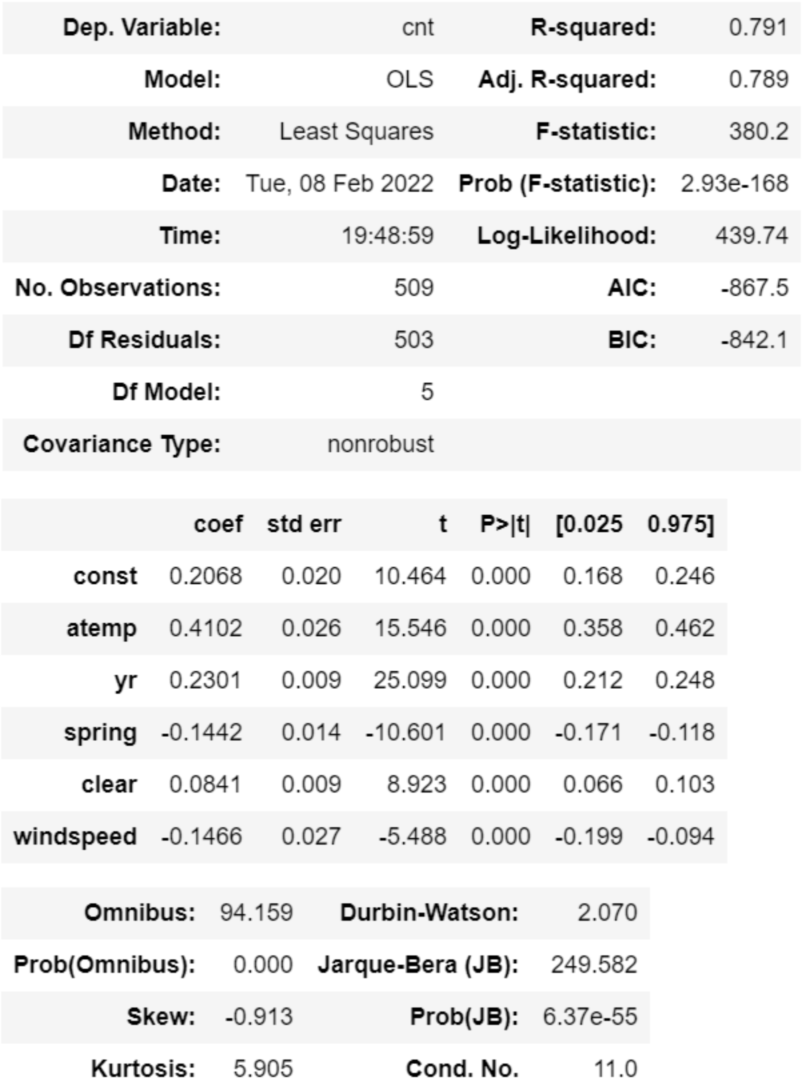




Question 5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

Ans: Significance of any variable can be described by the P(t) value of that variable as well as the Coefficient of that variable (steepness of the slope).

“**atemp**”, “**yr**”, “**windspeed**” are the three variables explaining the demand of the shared bikes.



Dep. Variable:	cnt	**R-squared:**	0.791
Model:	OLS	**Adj. R-squared:**	0.789
Method:	Least Squares	**F-statistic:**	380.2
Date:	Tue, 08 Feb 2022	**Prob (F-statistic):**	2.93e-168
Time:	19:48:59	**Log-Likelihood:**	439.74
No. Observations:	509	**AIC:**	-867.5
Df Residuals:	503	**BIC:**	-842.1
Df Model:	5		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
const	0.2068	0.020	10.464	0.000	0.168	0.246
atemp	0.4102	0.026	15.546	0.000	0.358	0.462
yr	0.2301	0.009	25.099	0.000	0.212	0.248
spring	-0.1442	0.014	-10.601	0.000	-0.171	-0.118
clear	0.0841	0.009	8.923	0.000	0.066	0.103
windspeed	-0.1466	0.027	-5.488	0.000	-0.199	-0.094

Omnibus:	94.159	**Durbin-Watson:**	2.070
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	249.582
Skew:	-0.913	**Prob(JB):**	6.37e-55
Kurtosis:	5.905	**Cond. No.**	11.0

Business Significance of variables:

Price and marketing are such variables which directly relates to bike rental but those variables are not present in the dataset. So, we can look for variables which can be controlled to have indirect impact on rent price and marketing which can be “weather” and “temperature”.

a. When temperature increases then customers (demand) increase, so increase the rental price at that time.

b. When the weather is clear then increase the rental price and when whether is bad decrease it.

c. Specially in spring season decrease the rental price because in that season bike sale decreases.

GENERAL SUBJECTIVE QUESTIONS

Quesntion 1: Explain the linear regression algorithm in detail.

Ans:

- Regression algorithm is a way by which we try to predict the outcome of any unseen data.

- **Linear** word here depicts that we are trying to fit a straight line through a set of data points.

- We have one **output/prediction/Dependent** variable for which we want to predict the value, And the other variables are **Independent/Feature** variables using which we use to generate value for the dependent variable.

- The relation between output and feature variable can be described by the equation of line as shown below-

Y = b0 + b1 * x1 + b2 * x2 + . . . + bN * xN

- Here b0 (a constant) is the intercept of the line (hyperplane in case of multivariate analysis) with y axis. Which shows that from where the line starts.

- b1, b2, …, bN, all decides the slope of the hyperplane. Each Independent variable has its own slope with the dependent variable which depicts the strongness of relationship between dependent variable and independent variable.

- We need to consider few assumptions before starting to compute the equation of hyperplane. They are: -

1. The relation between independent and dependent variable is linear for entire population.

2. There is only single value of Y for any given value of x (keeping other x constant).

3. There is always a distribution of error terms around any predicted y value.

4. Error terms are always normally distributed (Value of X and y are not assumed to be normally distribited)

5. Error terms have constant variance.

- **R^2 score** depicts, how much variance in data is described by the model.

R^2 = 1- (RSS / TSS) where RSS =residual sum of squares, TSS = total sum of squares

- **F-statistics** shows the significance of the model and show whether or not the fit is statistically correct or gained by chance.

- **P(t)** shows whether of not the included variable has any significance or not.

- **Coefficients** shows how strongly a dependent variable is related with the independent variable.

There are two types of Regression Models:

a. Simple Linear Regression
b. Multiple Linear Regression

- In Multiple linear regression model, multiple variables are part of the model which improves the Model R^2 value but may overfit because of the Multicollinearity issue.

- Multicollinearity is a situation when one dependent variables is linearly dependent or other one or more variables.

Question 2: Explain the Anscombe's quartet in detail.

Ans:

- It was constructed by Francis Anscombe in 1973.

- It has four set of data points which has nearly identical descriptive statistics properties but when graphed appears very different.

- This concludes that before building any model over the data we need to first graph the data and then start the analysis.

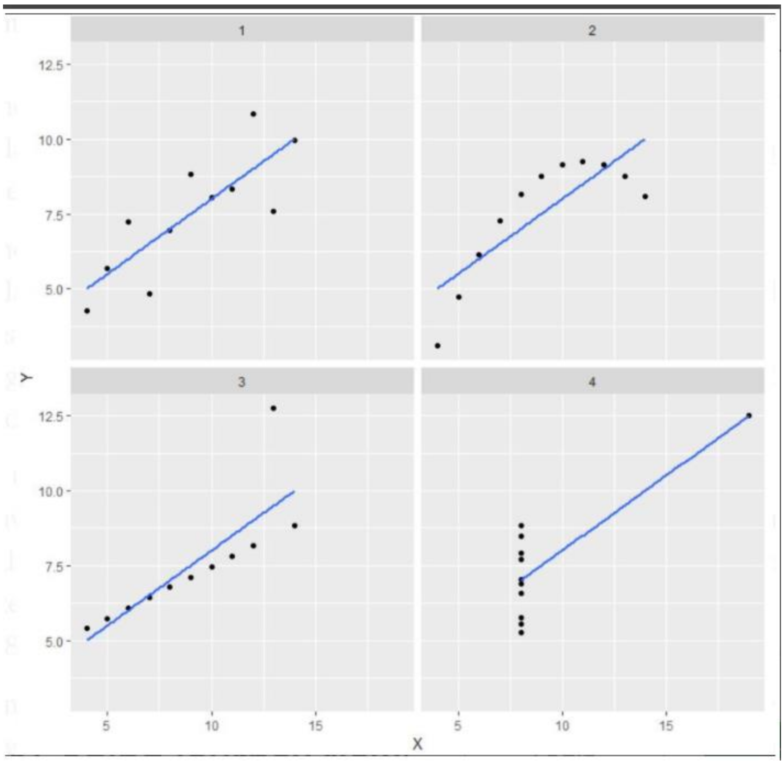


Image credit: https://www.geeksforgeeks.org/anscombes-quartet/

- Here we can see four datasets all four have same number of data points and same statistical properties like mean, std deviation, variance but when graphed we can clearly see the difference.

- In **Top left** dataset it seems that the data is dispersed linearly so we can fit a linear regression model on that data.

- In **Top right** dataset, it clearly shows that's the data is non-linear in nature and linear model can't be fit over that data.

- In **Bottom left** dataset, One outlier is there which hugely affects the slope of regression line which affects the performance of the model. So can't fit linear model with that data point persist in the dataset.

- In **Bottom Right** dataset, Again a outlier which is far away in the model hugely disturbs the linear regression model.

Question 3: What is Pearson's R?

Ans:

- This is a correlation coefficient which measures the strongness of relationship between two datasets.

- Use to measure the linear correlation between two sets of data.

- It is the ration between covariance of two variables and the product of their standard deviation.

$$r_{xy} = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right)$$

Here always **(-1 <= r <= 1)**

- Here first term after sigma symbol is the Z score for x variable and second term is Z score for y variable.

- If the value of **r** is positive that means that with increase or decrease in value of x variable, y variable also increases or decreases respectively.

- If the value of **r** is negative that means that with increase or decrease in value of x variable, y variable decreases or increases respectively.

- If the value of r = 0 that means there is no relation between two datasets ideally.

- Getting r = 1, r = -1 or r = 0 is practically very difficult in real world scenario.

- Value of r =1 and r = -1 means that the points are exactly on the line.

- Value of **r** is positive if (Xi − X) * (Yi − Y) is positive which means Xi and Yi lie on the same side of their respective mean.

- The correlation coefficient is negative if Xi and Yi tends to lie opposite side of their respective means.

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

Ans:

- When we build any model then sometime, we need to compare the significance of each independent variable i.e., how much any variable is affecting the output variable compare to other independent variables.

- So, the significance of any variable in model is provided by the coefficient of that variable (the beta values).

- If all variables are not on the same scale, then it is very difficult sometimes to find out the significance of any variable with respect to other variables.

- To remedy this situation, we can “normalize” or “standardize” all the variables.

- The second use of the scaling techniques is that sometime use of these techniques improves the performance of ML algorithms.

Normalization: means rescaling entire data to fit in range 0 to 1.

- it is best to use Normalization when we assume that the data does not follows any distribution.

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}}$$

- Here X_max and X_min are the min and max value in the dataset and X is the datapoint.

Standardization: is a way for the data to adapt mean = 0 and std deviation = 1.

- it is best to use Standardization when we assume that the data follows normal distribution.

$$X' = \frac{X - \mu}{\sigma}$$

- Here U us the mean of the dataset, sigma is the std deviation of the dataset and X is the data point.

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

Ans:

- Formula for VIF says:-

$$\mathrm{VIF}_i = \frac{1}{1 - R_i^2}$$

- Here in the formula, we can see that VIF depends on the R square value. And R square actually is the square of the correlation coefficient (R) between variables.

- So, if R =1 that means there is a perfect correlation between the variables, in that case R square also have the value = 1.

R = R^2 = 1 makes the value of VIF = Infinity.

- This happens if one variable is linear combination of one or more than one variable. That means one variable is perfectly describes by the linear combination of other variables which makes that variable redundant for the analysis.

- If there is no correlation between any two variable that means R = 0 (ideally) which intends R^2 = 0 and which in turn makes VIF = 1 which makes both of the variables perfectly suitable for the analysis.

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

Ans:

- Q-Q plot means quantile-quantile plot.

- Q-Q plot is used to find out whether a dataset is normally distributed or not. It is more accurate then to plot the distribution of data and check visually.

- Q-Q plot provides the information of whether a plot is skewed, tailed or perfectly normally distributed.

- We plot theoretical quantiles or basically known as the std normal variate (normal distribution with mean = 0 and std deviation = 1) with any sample datapoints from the population which we need to find out whether those are normally distributed or not.

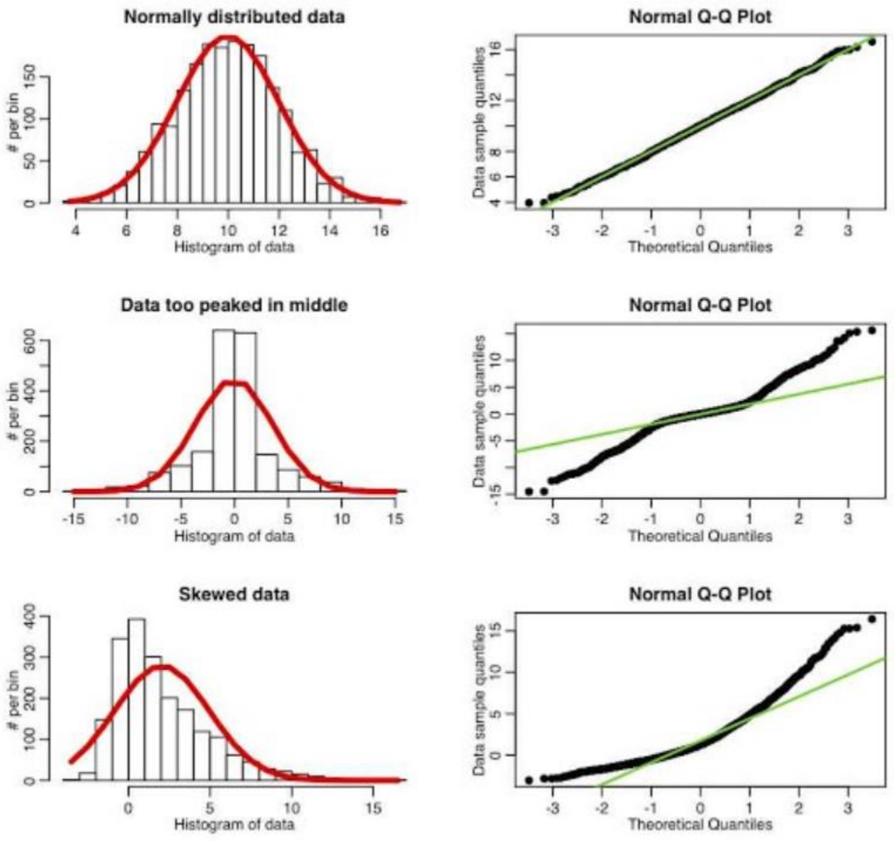


Image Credit: https://towardsdatascience.com/q-q-plots-explained-5aa8495426c0

- These plots are plotted with a 45-degree line in the graph. If the points lie on that line that means that the dataset is normally distributed

- If the lower tail of the graph does not lie on the 45-degree line means dataset is left tailed.

- If the upper tail of the graph does not lie on the 45-degree line means dataset is right tailed.

- If both upper and lower tail of the graph does not lie on the 45-degree line that means graph is long tailed on both sides.

- And if all the points lie on the 45-degree line that means distribution is thin tailed.