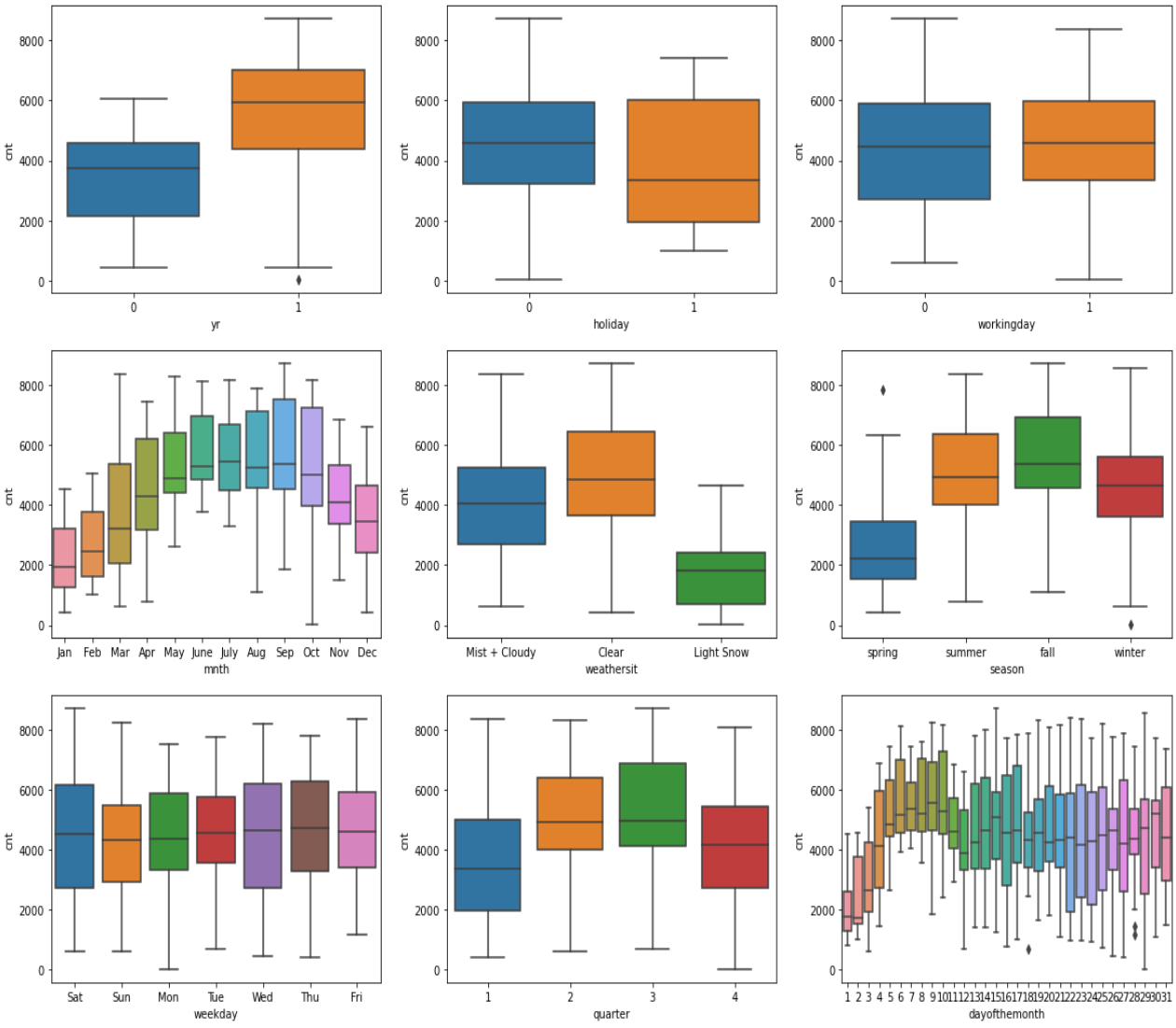
Assignment-Based Subjective Questions

1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)

Observations made after analysing the relationship of the categorical variables with the target variables are as follows:

1. Bike rentals were more in 2019 than 2018 (yr)
2. Day being a holiday leads to a decrease in bike rentals (holiday)
3. Bike rentals are increasing from Jan-June, Sep has the max bike rentals and the demand falls after that (mnth)
4. Bike rentals are more in clear weather and are least in the snowy conditions (weatherSit)
5. Bike rentals are max during Fall, followed closely by summer season (season)
6. Weekday is not leading to a solid conclusion for bike rental behavior (weekday)
7. Bike rentals are high in number in Q2 and Q3 (derived variable Quarter from dteday)
8. Bike rentals are highest from 5th to 12th of a month (derived variable dayofthemonth from dteday)



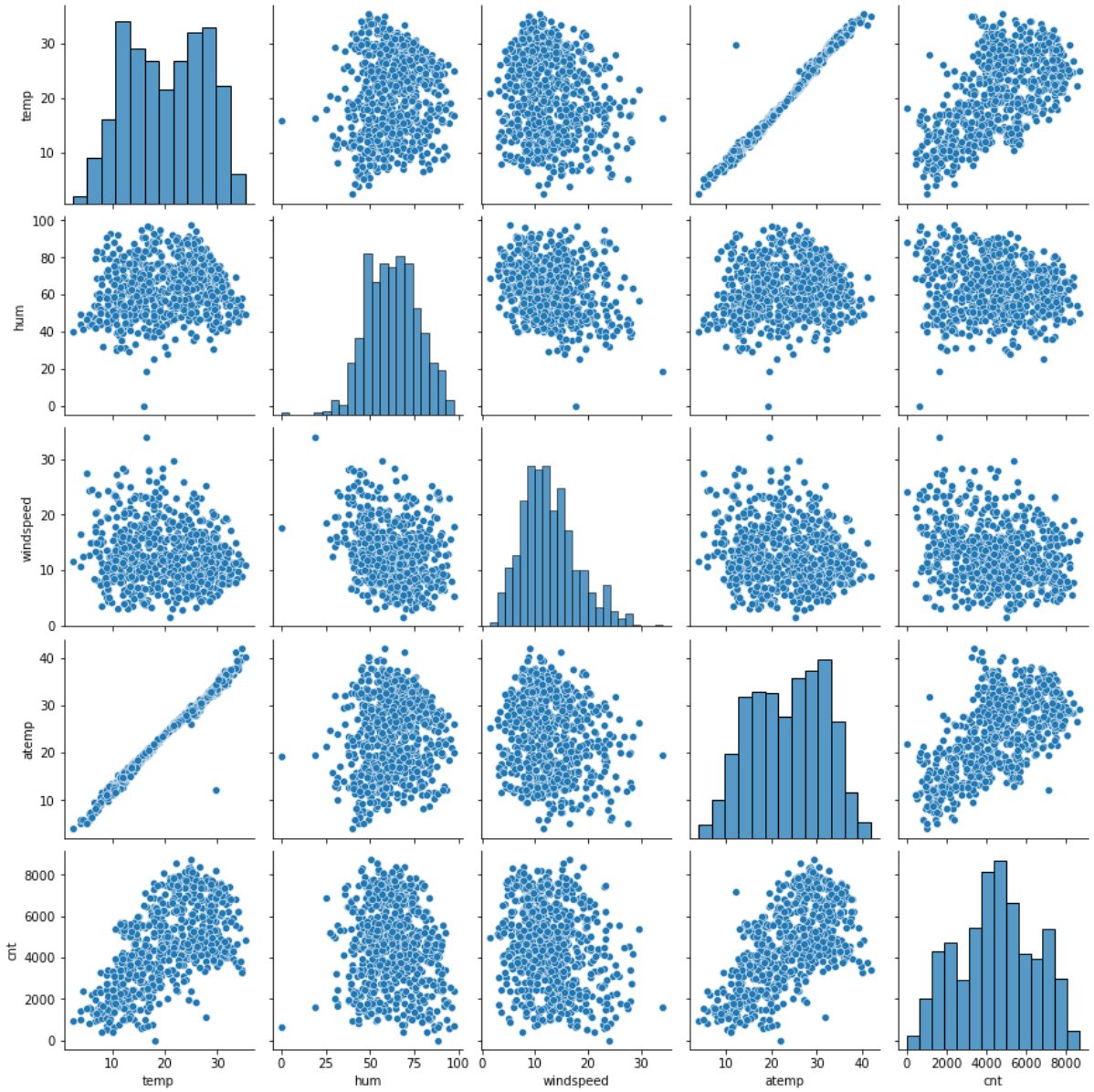
2. Why is it important to use drop_first=True during dummy variable creation? (2 mark)

drop_first=True is very important to use while encoding the categorical variable to dummy variables, as it helps us in reducing the extra columns created during the process. Eventually, it reduces the correlations created among dummy variables which is important for the stability of the dataset. Hence, if we have a categorical variable with n-levels, then we need to use n-1 columns to represent the dummy variables required to represent the full extent of information of that categorical variable.

For example, if you have a variable gender, you don't need both a male and female dummy. Just one will be fine. If male=1 then the person is a male and if male=0 then the person is female.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

temp variable has the highest correlation with the cnt variable, closely followed by atemp.

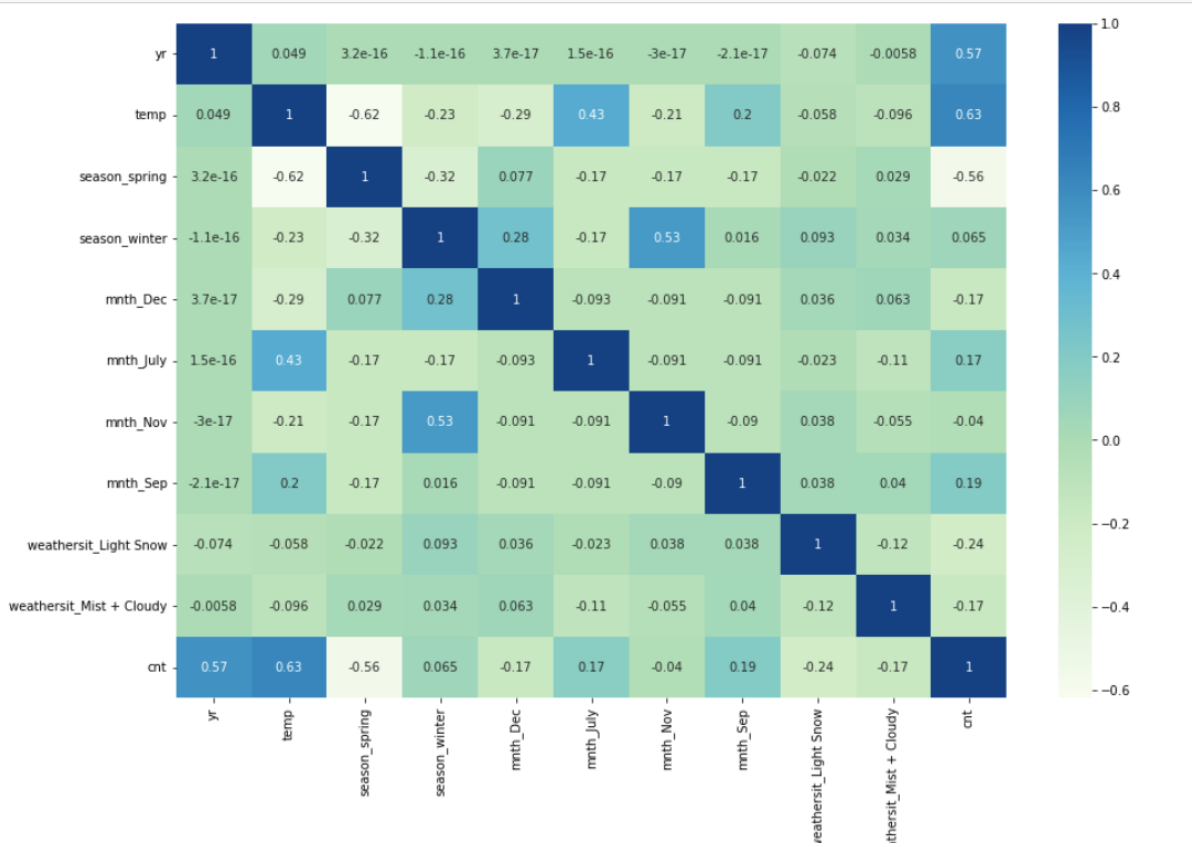


4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

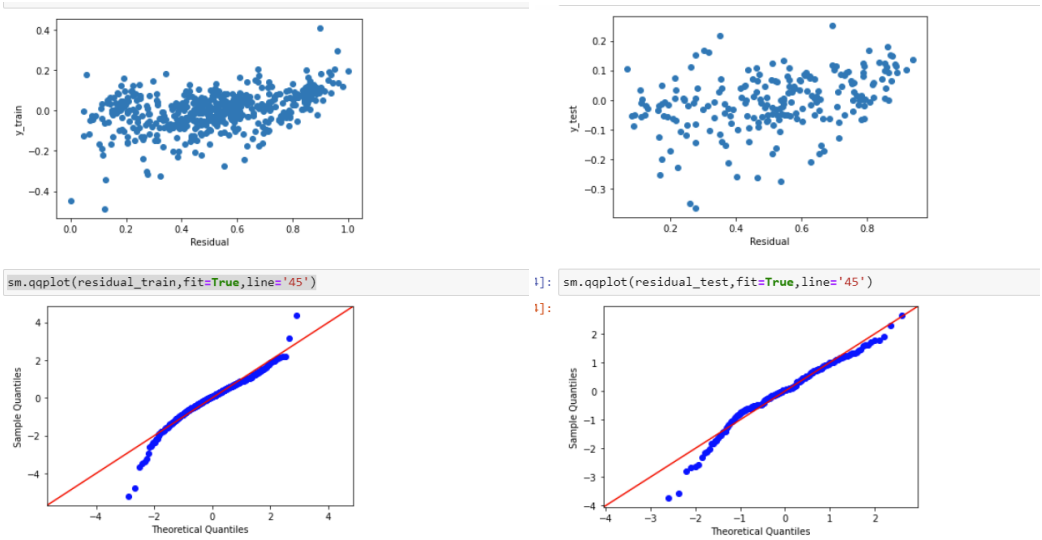
1. To check the multicollinearity assumption, VIF values were calculated and variables with VIF>3 were dropped

	Features	VIF
1	temp	2.90
3	season_winter	2.33
0	yr	2.02
6	mnth_Nov	1.61
9	weathersit_Mist + Cloudy	1.55
4	mnth_Dec	1.38
2	season_spring	1.37
5	mnth_July	1.35
7	mnth_Sep	1.21
8	weathersit_Light Snow	1.07

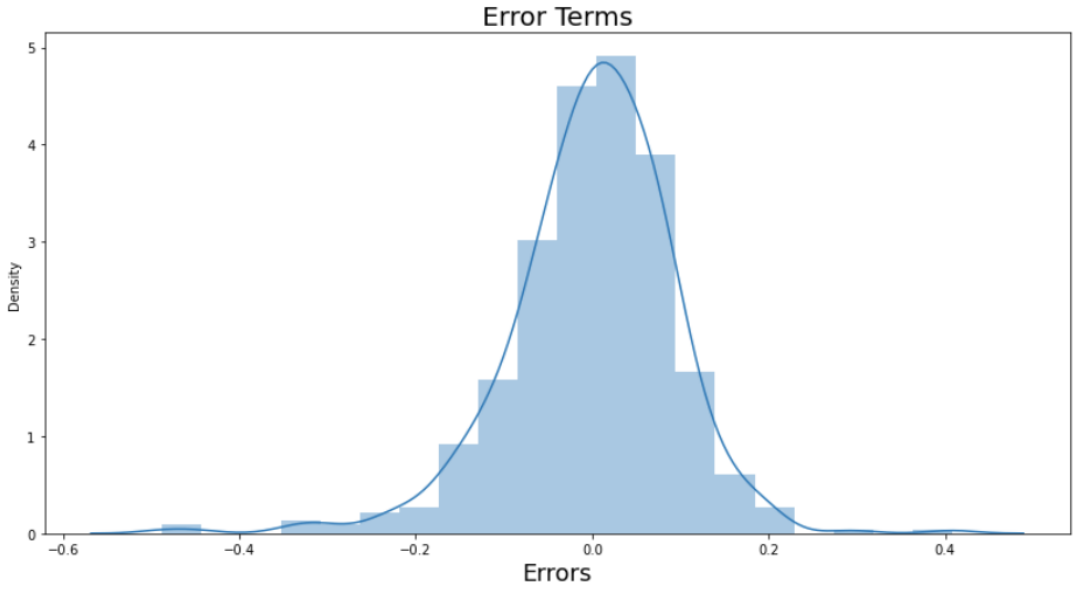
2. For the linearity assumption, a heatmap of correlation matrix was plotted and correlation values were checked. 'Cnt'(target variable) seems to be linearly related to all the independent variables which are coming out to be significant in the final model.



3. For the homoscedasticity assumption, scatter plot was built for y-test vs test_residual as well as y-train vs train_residual and also Q-Q plots have been plotted for both test and train datasets. Equally spread residuals around a horizontal line without distinct patterns are a good indication of having the linear relationships. Error terms are randomly distributed and are independent of one another. Thus, Homoscedasticity assumption is getting satisfied.



4. The residuals should be normally distributed which is validated by plotting a histogram on the residual on the train dataset and mean=0



5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

Top 3 features contributing significantly towards explaining the demand of the shared bikes are:

Final Model equation:
*cnt= 0.2316 + (0.2366 * yr) + (0.4146 * temp) + (-0.1295 * season_spring) + (0.0915 * season_winter) + (-0.0749 * mnth_Dec) + (-0.0517 * mnth_July) + (-0.0634 * mnth_Nov) + (0.0523 * mnth_Sep) + (-0.3074 * weathersit_Light Snow) + (-0.0827 * weathersit_Mist + Cloudy)*

1. **Temperature** - A coefficient value of 0.4146 indicates that temperature has a significant impact on bike rentals
2. **Weathersit (3) Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds** - A coefficient value of -0.3074 indicates that light snow and rainy conditions deter people from taking rental bikes

3. **Year**- A coefficient of 0.2366 indicates that year wise the bike rental numbers are increasing. An increase of 0.2366 units is recorded moving from 2018 to 2019 (with a unit increase in year).

OLS Regression Results

Dep. Variable:	cnt	R-squared:	0.825
Model:	OLS	Adj. R-squared:	0.822
Method:	Least Squares	F-statistic:	235.7
Date:	Wed, 13 Apr 2022	Prob (F-statistic):	7.14e-182
Time:	17:37:54	Log-Likelihood:	484.50
No. Observations:	510	AIC:	-947.0
Df Residuals:	499	BIC:	-900.4
Df Model:	10		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.2316	0.023	9.919	0.000	0.186	0.277
yr	0.2366	0.008	27.935	0.000	0.220	0.253
temp	0.4146	0.033	12.651	0.000	0.350	0.479
season_spring	-0.1295	0.016	-8.069	0.000	-0.161	-0.098
season_winter	0.0915	0.014	6.648	0.000	0.064	0.119
mnth_Dec	-0.0749	0.017	-4.392	0.000	-0.108	-0.041
mnth_July	-0.0517	0.018	-2.910	0.004	-0.087	-0.017
mnth_Nov	-0.0634	0.020	-3.242	0.001	-0.102	-0.025
mnth_Sep	0.0523	0.016	3.293	0.001	0.021	0.083
weathersit_Light Snow	-0.3074	0.024	-12.825	0.000	-0.354	-0.260
weathersit_Mist + Cloudy	-0.0827	0.009	-9.143	0.000	-0.100	-0.065

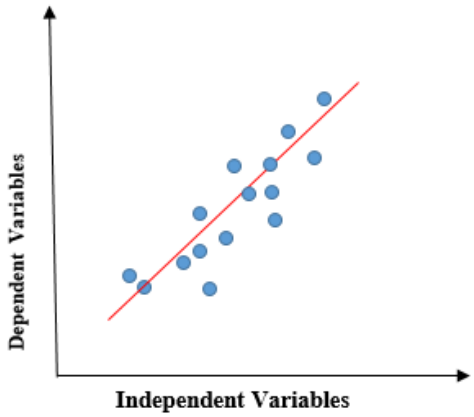
Omnibus:	85.275	Durbin-Watson:	1.933
Prob(Omnibus):	0.000	Jarque-Bera (JB):	288.936
Skew:	-0.747	Prob(JB):	1.81e-63
Kurtosis:	6.371	Cond. No.	14.1

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

General Subjective Questions

1. Explain the Linear Regression algorithm in detail

Linear Regression is a supervised learning technique that helps us performs the task of regression. Regression is applied where the target variable is continuous in nature. LR shows the linear relationship between the independent variable (X-axis) and the dependent/target variable (Y-axis), in case of a single independent variable, LR is termed as Simple Linear Regression and in case of multiple independent predictor variables, it's termed as Multiple/Multivariate Linear Regression. The red line shown below is termed as the best fit regression line. The independent variables must have a linear relationship with the target variable and must be statistically significant in predicting the target variable.



Source: Link

The linear regression algorithm tries to reduce the errors in predictions using the Ordinary Least Squares (OLS) method.

Key assumptions the dataset should satisfy in order to be able to implement the linear regression algorithm are -

1. Linear relationship between independent and dependent variables
2. Multivariate normality for all variables
3. Little to no multicollinearity in the data
4. Little or no autocorrelation in the data
5. Homoscedasticity in the data

Equation of Simple Linear Regression - $y = B0 + B1x1$
(where y - dependent variable, B0 - intercept, B1 - coefficient, x1 - independent variable)

Equation of Multiple Linear Regression - $y = B0 + B1x1 + B2x2 + .. + Bnxn$
(where y - dependent variable, B0 - intercept, B1 - coefficient of x1, x1 - first independent variables)

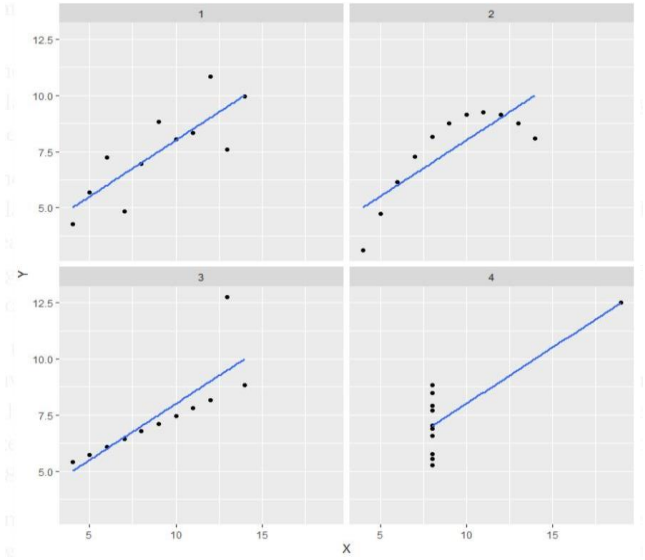
2. Explain the Anscombe's quartet in detail

Anscombe's quartet comprises of four different datasets that have nearly identical simple statistical properties but they appear very different from each other when graphed. Each dataset consists of eleven (x,y) points.
They were constructed in 1973 by the statistician Francis Anscombe to demonstrate:

- the importance of graphing data before analysing it
- the effect of outliers on statistical properties.

I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

Source:https://www.geeksforgeeks.org/anscombes-quartet/



Explanation of this output- by looking at the tables, it seems like the datasets are not very different from each other but the graphs explain otherwise:

- First one (top left): Linear relationship between x and y.
- Second one (top right): Non-linear relationship between x and y.
- Third one (bottom left): Perfect linear relationship for all the data points except one outlier
- Fourth one (bottom right): shows an example when one high-leverage point is enough to produce a high correlation coefficient.

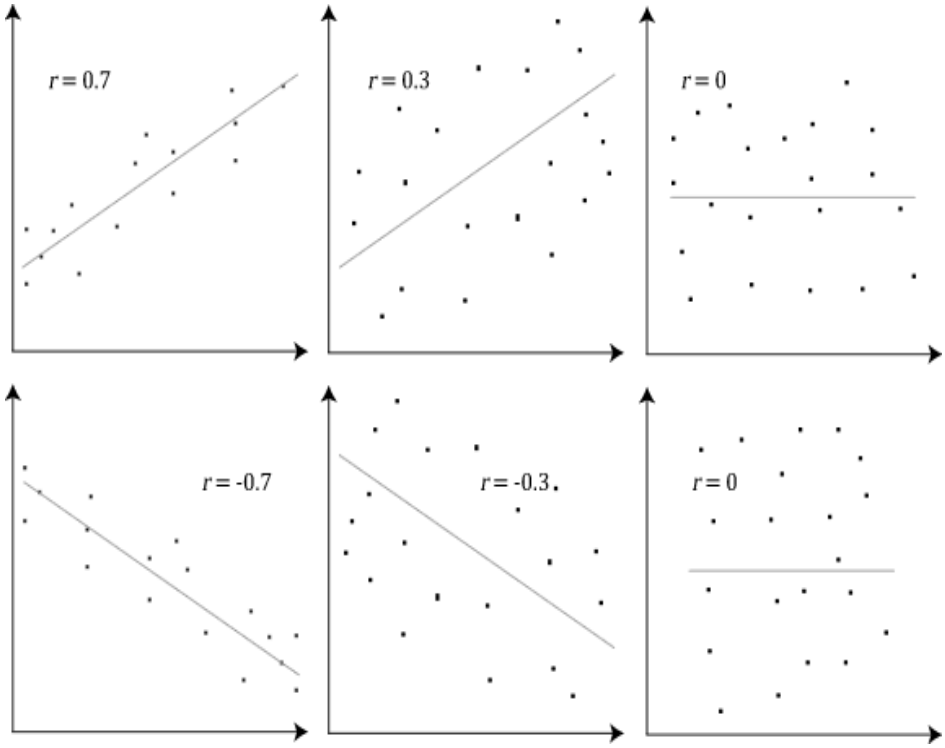
3. **What is Pearson's R** *(Answer taken from my own article on Medium on the same topic)*

Pearson's correlation coefficient is also referred to as Pearson's *r.* This bivariate correlation is a statistic that measures the linear correlation between two variables *X* and *Y*. It has a value between +1 and −1. A value of +1 is a total positive linear correlation, 0 is no linear correlation, and −1 is a total negative linear correlation.

Important Inference to keep in mind: The Pearson correlation can evaluate ONLY a linear relationship between two continuous variables (A relationship is linear only when a change in one variable is associated with a proportional change in the other variable)

Example use case: We can use the Pearson correlation to evaluate whether an increase in age leads to an increase in blood pressure.

Below is an example of how the Pearson correlation coefficient (r) varies with the **strength and the direction of the relationship** between the two variables. Note that when no linear relationship could be established (refer to graphs in the third column), the Pearson coefficient yields a value of zero.



Source: Wikipedia

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

What and why of Scaling:
Most of the times, our datasets have different variables varying highly in magnitudes, units and ranges. If we use those variables as is, then the regression algorithm takes into account the magnitude and not the units corresponding to each variable and leads to incorrect modelling. To solve

this issue, scaling comes into picture. We use scaling to bring all the variables to the same level of magnitude so that the modelling algorithm will not differentiate between the variable's basis on their magnitude.

For example: See the image below and observe the scales of salary Vs Work experience Vs Band level. Due to the higher scale range of the attribute Salary, it can take precedence over the other two attributes while training the model, despite whether or not it actually holds more weight in predicting the dependent variable.

Employee Name	Salary	Work Ex (Years)	Band Level
John	1,10,000	13	3
Anna	67,000	8	4
Scarlette	23,000	3	7
Shiva	75,000	10	4
Sean	43,000	5	5

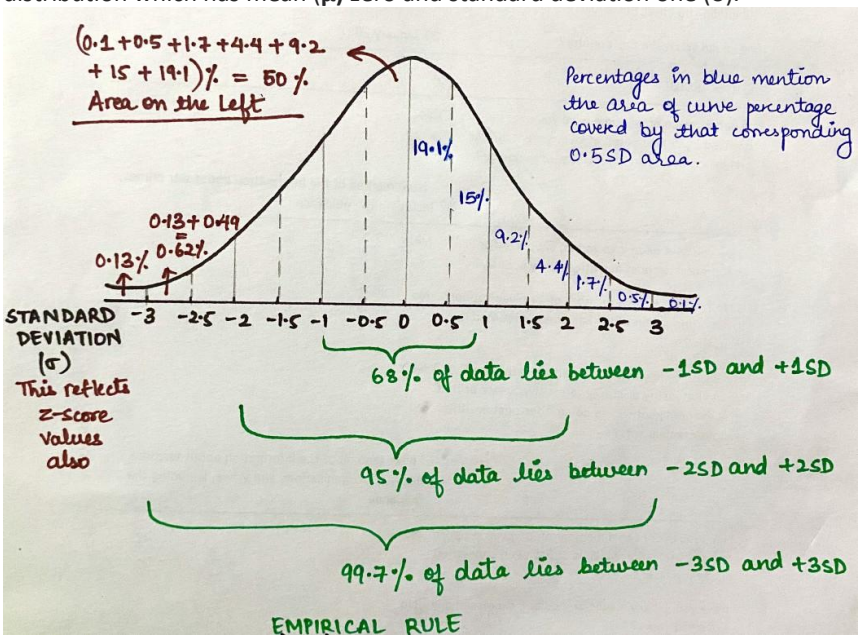
<u>Difference between normalized and standardized scaling</u>

Normalization/Min-Max Scaling:

It brings all of the data in the range of 0 and 1.
sklearn.preprocessing.MinMaxScaler helps to implement normalization in python.

$$\text{MinMax Scaling: } x = \frac{x - min(x)}{max(x) - min(x)}$$

Standardization Scaling:
Standardization replaces the values by their Z scores. It brings all of the data into a standard normal distribution which has mean (**μ**) zero and standard deviation one (**σ**).



$$\text{Standardisation: } x = \frac{x - mean(x)}{sd(x)}$$

sklearn.preprocessing.scale helps to implement standardization in python.
One disadvantage of normalization over standardization is that it **loses** some information in the data, especially about **outliers**.

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

If the value of VIF for an independent variable is infinity, i.e. R squared = 1 (VIF = 1/(1-R2)), it means that the variable is highly collinear with other independent variables and we should definitely drop this variable as it will hamper the interpretability of the model results.
An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables (which show an infinite VIF as well).

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

Q-Q plots are also known as Quantile-Quantile plots. As the name suggests, they plot the quantiles of a sample distribution against quantiles of a theoretical distribution. The power of Q-Q plots lies in their ability to summarize any distribution visually.

Q-Q plots are used to determine:

1. If two populations are of the same distribution
2. If residuals follow a normal distribution. Having a normal error term is an assumption in regression and we can verify if it's met using this.
3. Skewness of distribution

In Q-Q plots, we plot the theoretical Quantile values with the sample Quantile values. Quantiles are obtained by sorting the data. It determines how many values in a distribution are above or below a certain limit.
In case of linear regression, we can plot the Q-Q plots on training data and test data separately to confirm if both are from populations with same distributions.