Assignment-based Subjective Questions

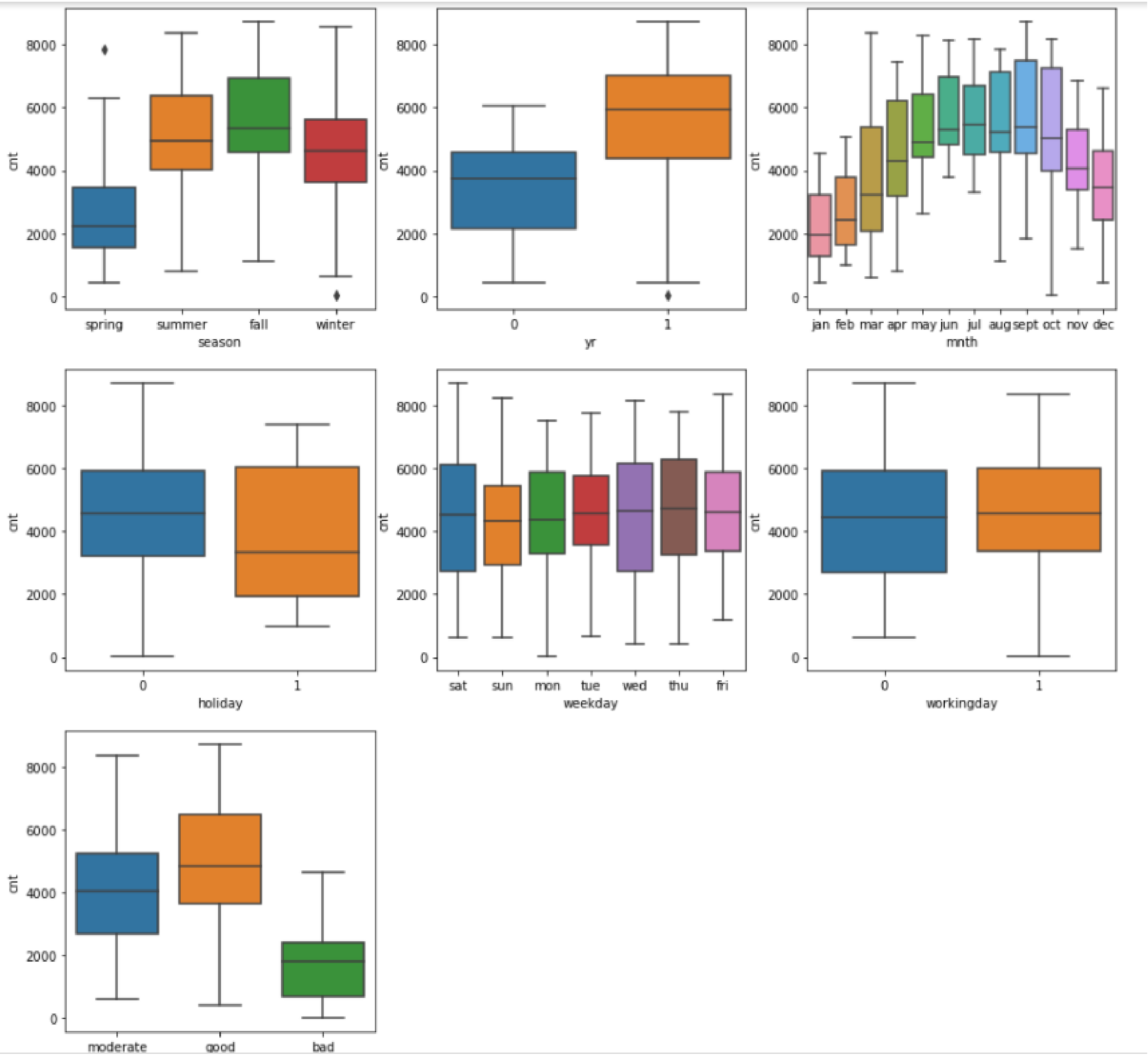
1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

Ans:

1. From my Analysis, of the categorical variable 'yr' we can say that the demand of bikes has substantially increased in the year 2019 from the previous year, which means that the demand was growing before the pandemic and should continue to grow after the pandemic is over.
2. The categorical variable 'season' infers that the demand of bikes are high during 2 seasons 1. Summer and 2. Fall from this observation the company can manage resources and maintain the bikes before the season starts and repeat the cycle after the season end, this would really help the company to plan accordingly and manage the demand in the coming months.
3. The variables 'months' says the same thing that the demand is high during certain period, due to some reason and the company can use this opportunity well manage the resources.
4. The categorical variable 'Holiday' states that the demand is marginally low on Holidays.
5. 'weekdays' however has not much to offer except for the fact the demand is almost similar throughout the week, which is a good sign for the company.
6. 'Workingday' also has not much to offer for this analysis.
7. The variable 'weathersit' is highly corelated to the count of bikes, and the study shows that demand for bikes are high when the weather is good.

Below is a snapshot of the boxplots that shows the above analysis:

–



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

Ans: Suppose we have a categorical variable, ‘Holiday’ when we create dummy variables on the holiday column it will create 2 columns naming ‘0’ which corresponds to ‘Not a holiday’ and ‘1’ for ‘Holiday’ in such case we don’t need 2 columns because, we can explain the same information with a single column, for eg: if the day is holiday the row will say ‘1’ else ‘0’, hence we use **drop_first = True.**

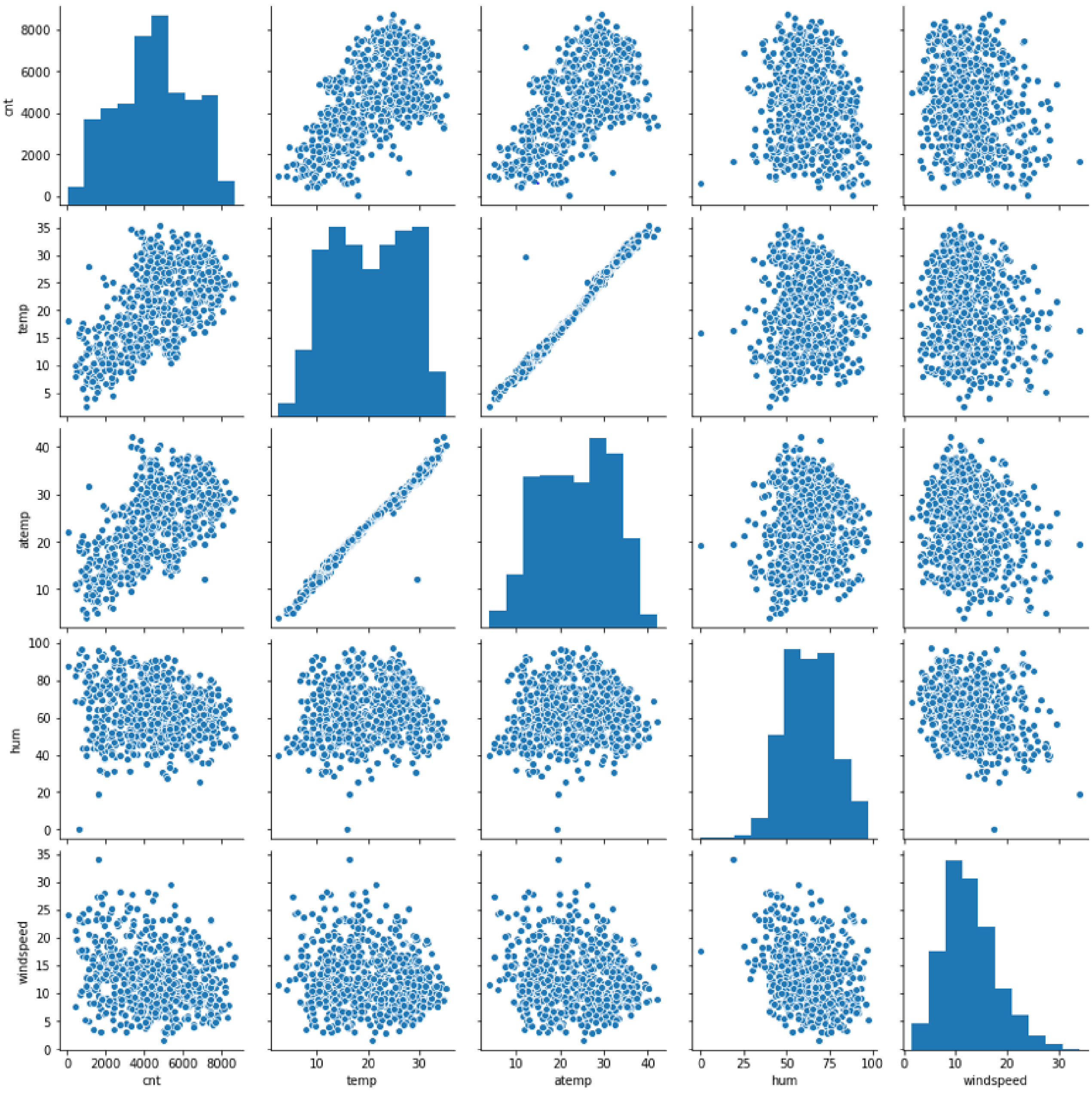
another reason for dropping the 1^{st} column can be the gradient descent algorithm will have less data to process.

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

Ans: the columns **temp** has the highest correlation with the target variable.

The numerical variable ‘registered’ has the highest correlation with the target variable ‘cnt’, if we consider all the features. But after data preparation, when I have dropped registered and atemp due to multicollinearity and high VIF value. The numerical variable ‘temp’ has the highest correlation with the target variable ‘cnt’.

shown below:

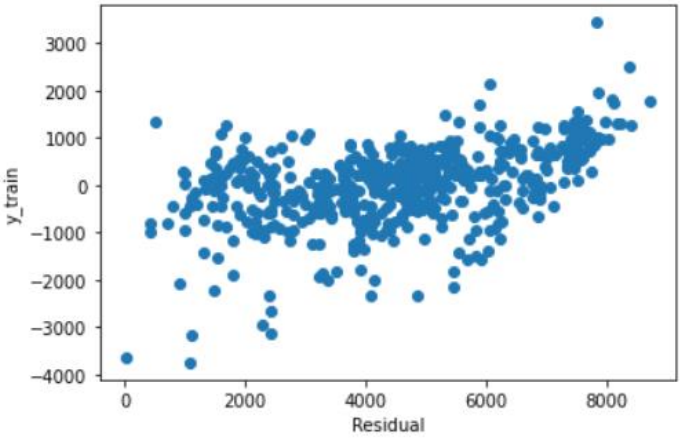


4. How did you validate the assumptions of Linear Regression after building the model on the training set?

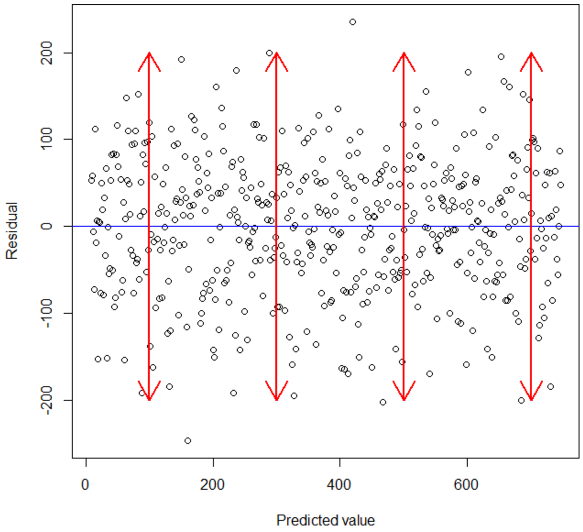
Ans: Assumption of Linear Regression validated with below measures.

- Linear Relationship.
- Homoscedasticity.
- Absence of Multicollinearity
- Independence of residuals
- Normality of Errors

Linear Relationship: A scatterplots would be helpful in validating the linearity assumption as it is easy to visualize a linear relationship on a plot.



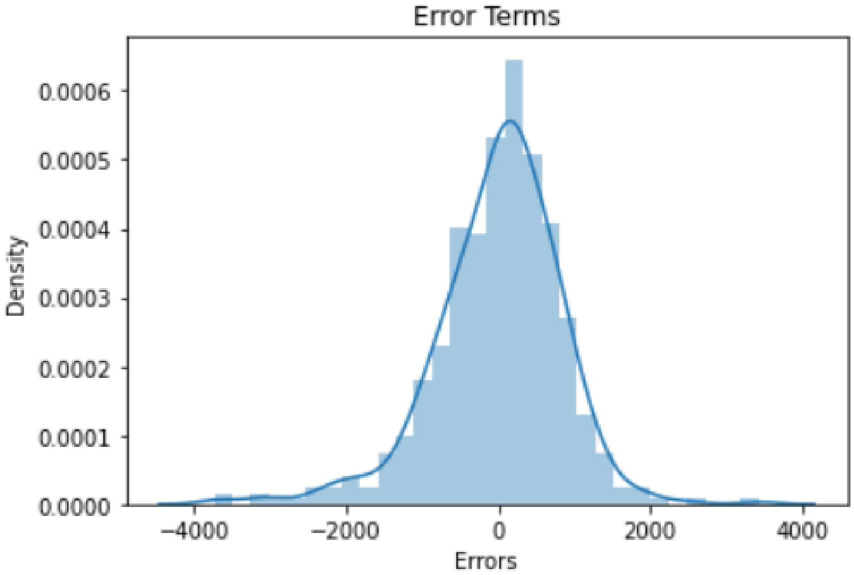
Homoscedasticity: It assumed that when we plot the individual error (residual terms) against the predicted value, it should be constant. As shown in the figure below, the length of the red lines (a proxy of itsvariance) is the same. It is also known as the assumption of homogeneity.



Absence of Multicollinearity: Multicollinearity is a state of very high inter-correlations among the independent variables. It is therefore a type of disturbance in the data if present weakens the statistical power of the regression model. Pair plots and heatmaps can be used for identifying highly correlated features.

Independence of residuals: It is assumed that the error terms orresiduals in linear regression are independent of each other.

Normality of Errors: If the residuals are not normally distributed, Ordinary Least Squares (OLS), and thus the regression, may become biased. We can draw a simple histogram of residuals to see the normal distribution.



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

Ans:

1. Temperature (temp), A coefficient value of ‘0.549936’ indicated that a temperature has significant impact on bike rentals

3. Light Rain & Snow (weathersit), A coefficient value of ‘-0.288021’ indicated that the light snow and rain deters people from renting out bikes

4. Year (yr), A coefficient value of ‘0.233056’ indicated that a year wise the rental numbers are increasing It is recommended to give utmost importance to these three variables while planning to achieve maximum bike rental booking. As high temperature and good weather positively impacts bike rentals, it is recommended that bike availability and promotions to be increased during summer months to further increase bike rentals.

General Subjective Questions

1. Linear Regression Algorithm in detail

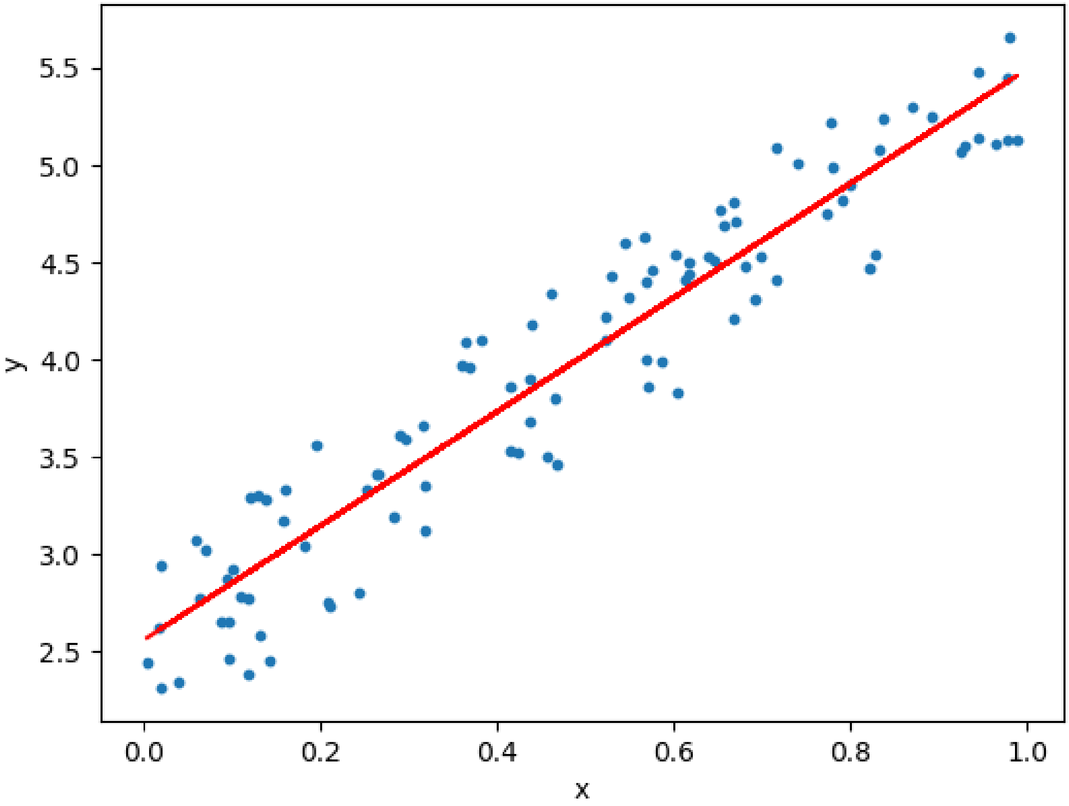
Linear regression is used for finding linear relationship between dependent and one or more independent variables. There are two types of linear regression- **Simple** and **Multiple**.

Simple Linear Regression

A simple linear regression model attempts to explain the relationship between a dependent and an independent variable using a straight line. The core idea is to obtain a line that best fits the data. The best fit line is the one for which total prediction error (all data points) are as small as possible. Error is the distance between the point to the regression line.

A linear regression line has an equation of the form **Y = a + bX**,

Where, **X** is the explanatory variable and **Y** is the dependent variable. The slope of the line is **b**, and **a** is the intercept (the value of y when x = 0).



Multiple Linear Regression

It represents the relationship between two or more independent variables and a dependent or target variable. Multiple linear regression is needed when one variable might not be sufficient to create a good model and make accurate predictions.

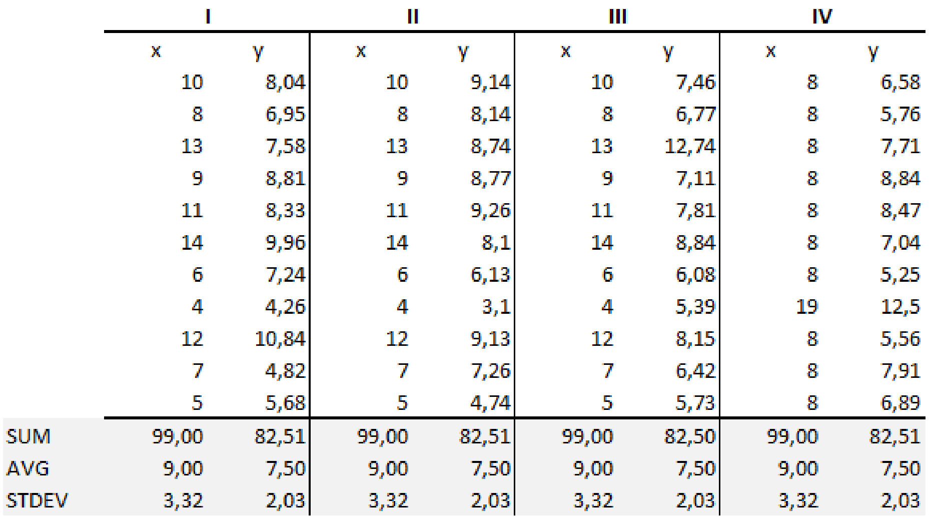
The formulation for predicting the dependent variable in multiple linear regression is given by the following equation below:

$$Y = \beta_0 + \ \beta_1 X1 + \beta_2 X2 + \ldots\ldots\ldots + \beta_n Xn + \varepsilon$$

Where, β_0 is the intercept of the line on the y-axis; $\beta 1$, $\beta 2$, and β_n are the slopes of the relations between y and X1, X2, and Xn, respectively; and **ε** is the random error term. The model now fits a hyperplane instead of a line. It is plotted through n + 1-dimensional space (n independent variables plus one dependent variable).

2. Explain the Anscombe's quartet in detail.

Anscombe's Quartet is the name given to the four numerical datasets which was developed by statistician **Francis Anscombe**. Each dataset consists of eleven (x,y) pairs. The most important thing is that all the four datasets have same descriptive. But when plotted, each graph visualized differently, irrespective of their similarity in summary statistics.



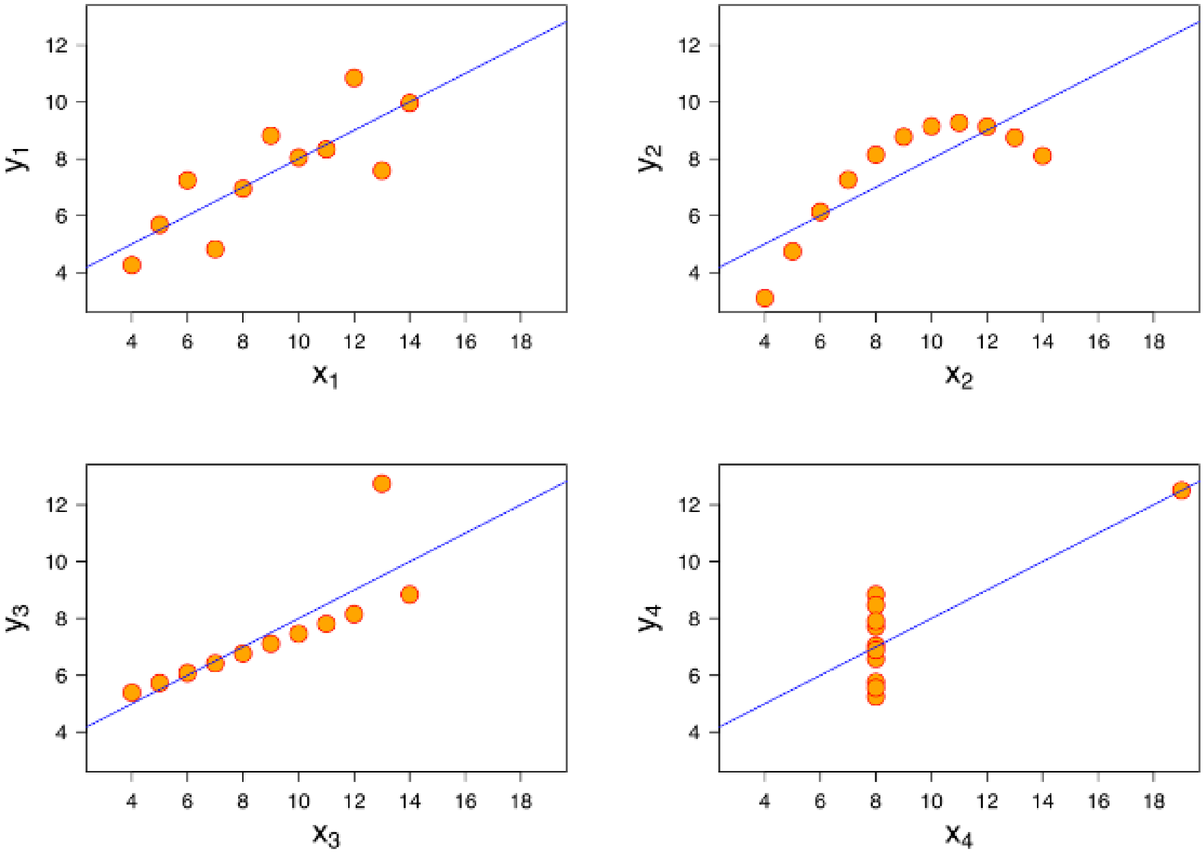
	I		II		III		IV	
	x	y	x	y	x	y	x	y
	10	8,04	10	9,14	10	7,46	8	6,58
	8	6,95	8	8,14	8	6,77	8	5,76
	13	7,58	13	8,74	13	12,74	8	7,71
	9	8,81	9	8,77	9	7,11	8	8,84
	11	8,33	11	9,26	11	7,81	8	8,47
	14	9,96	14	8,1	14	8,84	8	7,04
	6	7,24	6	6,13	6	6,08	8	5,25
	4	4,26	4	3,1	4	5,39	19	12,5
	12	10,84	12	9,13	12	8,15	8	5,56
	7	4,82	7	7,26	7	6,42	8	7,91
	5	5,68	5	4,74	5	5,73	8	6,89
SUM	99,00	82,51	99,00	82,51	99,00	82,50	99,00	82,51
AVG	9,00	7,50	9,00	7,50	9,00	7,50	9,00	7,50
STDEV	3,32	2,03	3,32	2,03	3,32	2,03	3,32	2,03

Quartet's Summary Stats

As we can see in above figure, the summary statistics shows that the means and the variances were identical for x and y across different groups:

- Mean of x is 9 and mean of y is 7.50 for each dataset.
- Similarly, the variance of x is 11 and variance of y is 4.13 for each dataset
- The correlation coefficient (how strong a relationship is between two variables) between x and y is 0.816 for each dataset

But when plotted four datasets on an x and y coordinate plane, we can observe that they show the same regression lines but each dataset has different visualization:



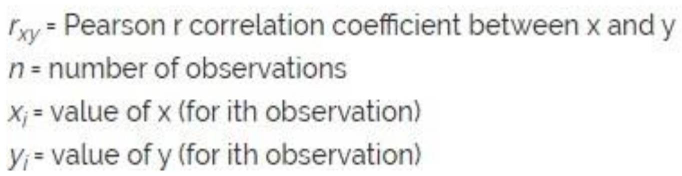
- Figure I appears to have clean and well-fitting linear models.
- Figure II is not distributed normally.
- In Figure III the distribution is linear, but the calculated regression is thrown off by an outlier.
- Figure IV shows that one outlier is enough to produce a high correlation coefficient.

The Anscombe’s quartet emphasizes that visualization is very important part of Data Analysis. Simply looking at the data reveals a pattern and a clear picture of the dataset.

3. What is Pearson’s R?

Pearson r correlation is the correlation statistic to measure the degree of the relationship between pair of linearly related variables. It is used to measure the degree of relationship between the two variables. The following formula is used to calculate the Pearson r correlation:

$$r_{xy} = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n\sum x_i^2 - (\sum x_i)^2}\,\sqrt{n\sum y_i^2 - (\sum y_i)^2}}$$



r_{xy} = Pearson r correlation coefficient between x and y

n = number of observations

x_i = value of x (for ith observation)

y_i = value of y (for ith observation)

Assumptions of Pearson’s R Correlation:

- For finding Pearson r correlation, both the variables should be normally distributed.
- Both the variables should have a straight line relationship.

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

Generally, dataset contain features that highly varying in magnitudes, units and range. For example, in a class performance dataset there is a feature age in years while marks out of 100. But machine learning algorithms use **Eucledian distance** between two data points in their computations of prediction, which is a problem.

Means machine learning algorithms only take the magnitude of features while ignoring the units. Due to this, the results would vary drastically. The features having higher magnitudes will weigh more in the distance calculations than features with low magnitudes. To nullify this effect, we need to bring all features to the same level of magnitudes let say between 0 and 1. This method of scaling down magnitudes to same level for all features is called **Scaling**.

Normalized Scaling

Normalization rescales the values into a range of [0, 1]. This might be useful in some cases where all features need to have the same positive scale. However, the outliers from the data set are lost.

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

Standardized Scaling

Standardized scaling makes the values of each feature in the data to have zero-mean (when subtracting the mean in the numerator) and unit-variance.

$$X_{changed} = \frac{X - \mu}{\sigma}$$

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

Variance inflation factors show the degree to which a regression coefficient will be affected because of the variable's redundancy with other independent variables. The variance inflation factor is computed as

$$VIF = \frac{1}{1 - R^2}$$

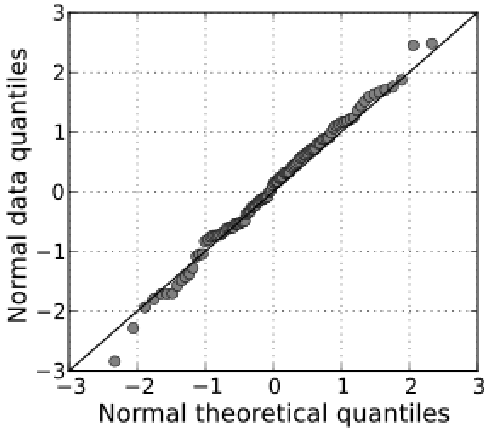
Where, R^2 is the R – squared statistic of the regression.

If the R^2 of any predictor variable with the other predictors approaches unity, the corresponding VIF becomes **infinite**. When VIF of any predictor become infinite it means there is a perfect collinearity i.e., presence of completely redundant variables.

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

The quantile-quantile (q-q) plot is a probability plot, which is a graphical technique for determining if two data sets come from populations with a common distribution.

A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set. By a quantile, we mean the fraction (or percent) of points below the given value. That is, the 0.3 (or 30%) quantile is the point at which 30% percent of the data fall below and 70% fall above that value.



A 45-degree reference line is also plotted. If the two sets come from a population with the same distribution, the points should fall approximately along this reference line. The greater the departure from this reference line, the greater the evidence for the conclusion that the two data sets have come from populations with different distributions.