Book Bikes Assignment

Name: Abhishek Singh
Dated: 30 May 2021
Email: aulakh.abhishek@gmail.com
Cohort – PGDML – March-21
Course: Linear Regression

Assignment-based Subjective Questions

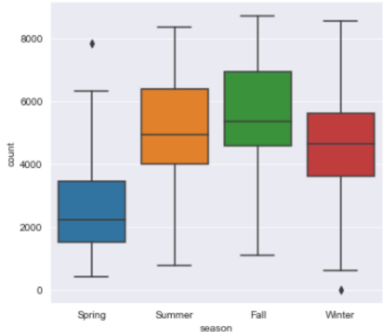
Question 1- From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)

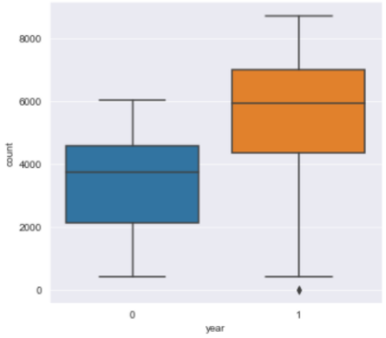
Answers 1- In our dataset 'days.csv', The dependent variable is 'cnt' (Count) with following categorical predictor variables –

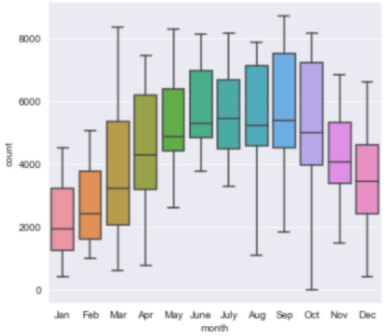
- season: season (1: spring, 2: summer, 3: fall, 4: winter)
- yr: year (0: 2018, 1:2019)
- mnth: month (1 to 12)
- holiday:
- weekday
- workingday: (if day is neither weekend nor holiday is 1, otherwise is 0.)
- weathersit: (1: 'Clear',2:'Mist_n_Cloudy',3:'Light_Snow',4:'Snow_n_Fog')

Based on the trends observed via Box Plots (shown below) for target variables vs categorial predictor variable, we can infer that

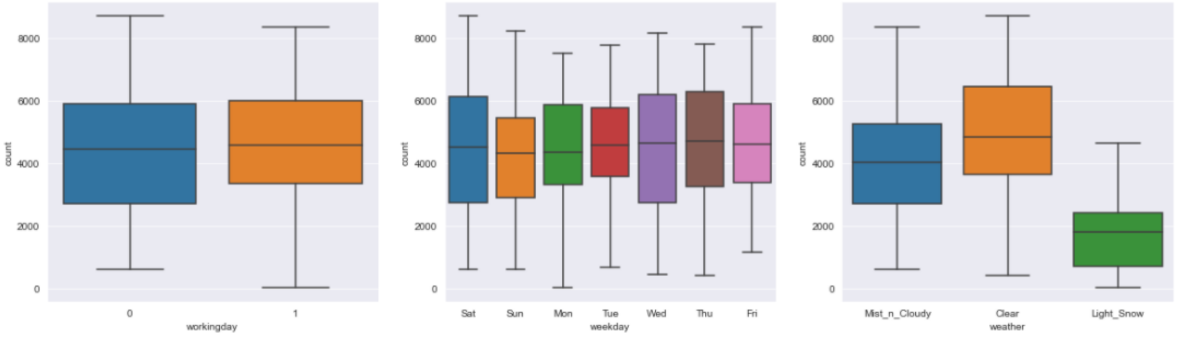
1. Count of Bikes (i.e., Demand) remains relatively low during Spring season while the demand increased significantly in the Summer and Fall season and fall slightly in the Winter season. The negative coefficient of Spring in our final Linear equation is an evidence of this.
2. The Demand of Bikes is higher in 2019 compared to 2018.
3. On the similar lines with Season, The Demand increases from March onwards each month and remain high between the months of May (Summer) and October (Fall) and then remains low for months of Winter (Nov – Feb)





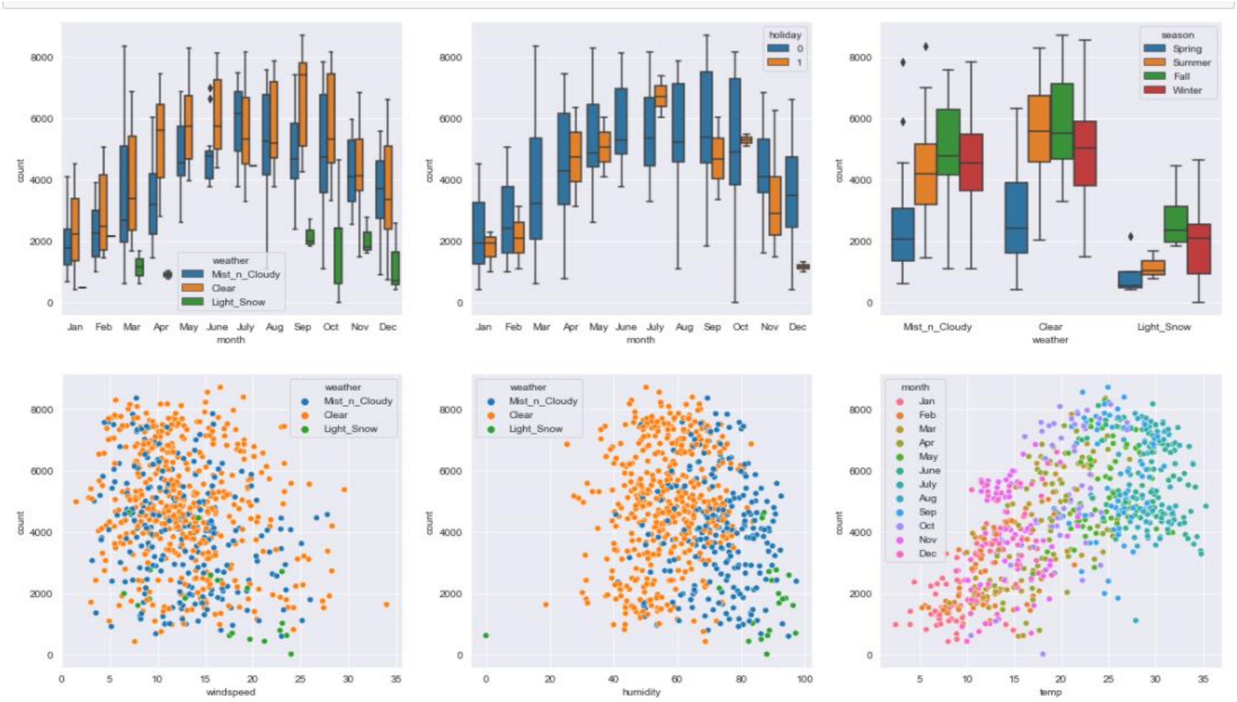


4. The categorical variables – ‘Working Day’ contains all the information provided the variable Holiday, meaning that Working day is 0 for all 0 values of Holiday and non-zero when it is not a holiday. If we infer the first plot, the demand is not affected much if it is a holiday or a working day.
5. The second plot represents the distribution of Demand over the weekdays and as we can see the median value of the count is somewhat similar on all days of the week, so again Demand is not affected if it is one specific day of the week or other including weekends.
6. The demand though is greatly affected by the weather. The demand remains fairly highly when it is clear weather outside, as can be seen in the third plot in the row and is very less on the days when it is raining or snowing.



If we combine few of these variables and plot them together, we observe following.

i. The Demands is high from May till October but relatively higher when the weather is clear. The demand is less when it is Snowing or Raining.
ii. The demand per month is higher when it is not a Holiday.
iii. In the plot-3, clear weather during Summer and Fall leads to High Demand. Same as point 1
iv. The demand is high when the windspeed remains less than 15 and reduces as the windspeed increases. The effect clear weather can be seen as well with Orange dots.
v. Demand remains higher when humidity is between 40 and 60 and reduces if it increases above 60.
vi. Rise in temperature clearly draws more Bike demand as can be seen in last plot. The High demand in Summer is an evidence of this.



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

Answer – 2: It is important to use '*drop_first=True'* in the get_dummies() method from pandas library to avoid the **Dummy Variable trap**, which is a problematic situation having predictor variables in the model which are highly correlated (Multi-collinearity) and one of the variable can predicts the value of others.

The Multiple linear equations assumes that coefficient of a predictor variable is the amount by which the dependent variable will change when all other predictor variables are kept constant, but in case if two predictor variables (p and q) are highly correlated then change in the value of one (let's say p) not only affects the dependent variable(y) but also the other predictor variable (q). In such cases since both predictor variable provides similar effect to dependent variable and are introducing error; it is best to drop one of them to avoid redundancy and Multi-collinearity.

When Dummy variables are created using *one hot encoding* for the categorical data, then one of dummy variable (attribute) can be predicted with the help of other dummy variables resulting into high correlation amongst the dummy variables. This situation fails Non-Multicollinearity and is named as *dummy variable trap*.

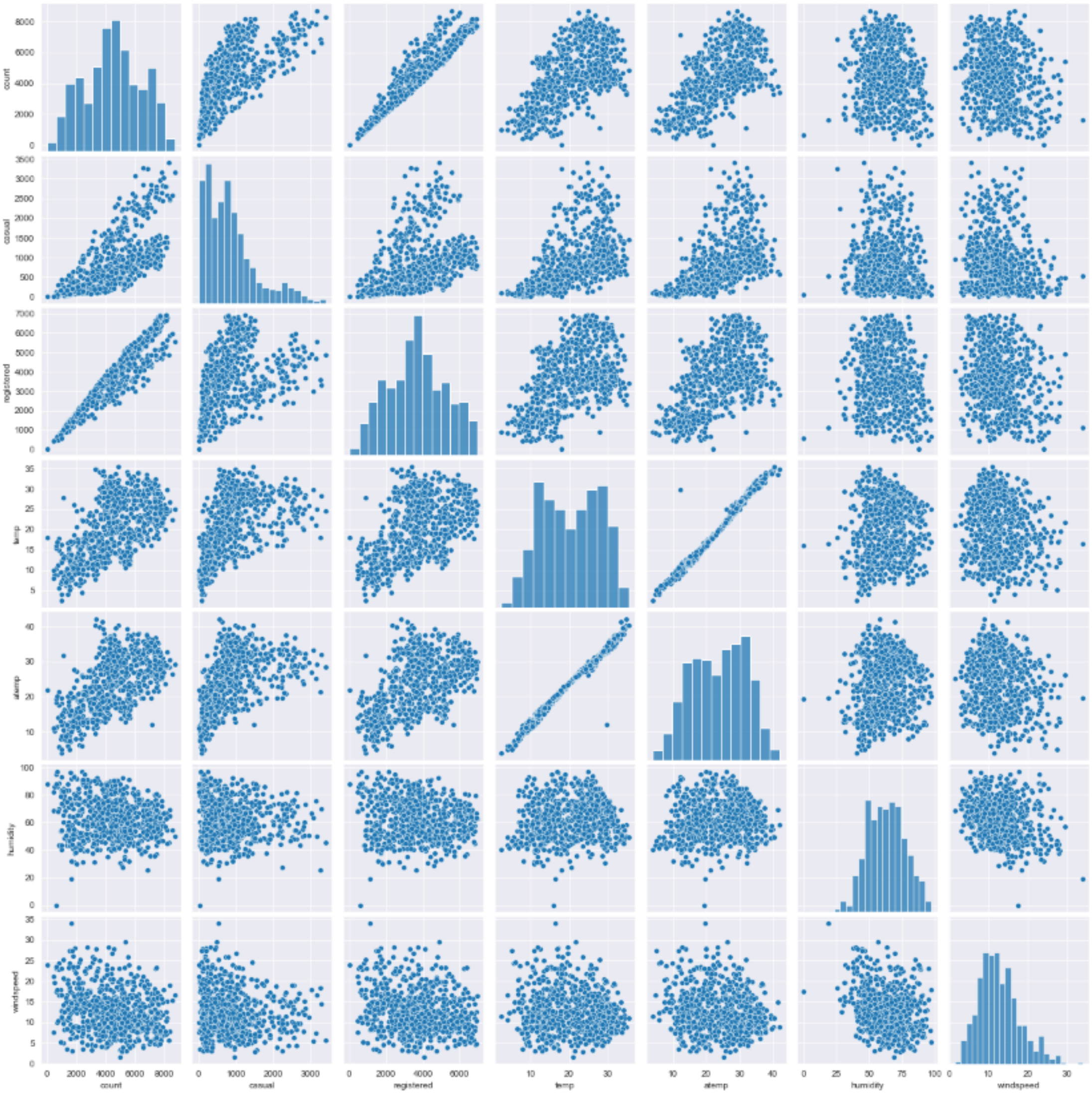
As an example, considering the case of marital status having only two possible categories as *Married (*0 and 1) and *Unmarried (*0 and 1). If the categorical data of marital status is One Hot Encoded it will provide two dummy variable columns, one each for Married and Unmarried. Clearly the two columns will be correlated as one can be predicted from the value of other because marital status has only two possible values and if a person is not Married (0) it is obvious that person is Unmarried (1).
Hence, including both the dummy variables will cause redundancy of information and multi-collinearity and we can safely drop one of the dummies.

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

Answer – 3: Below is the pair-plot of all the numerical variables amongst each other, it can seen clearly that target variable is highly correlated to **‘registered’** but it is because ‘cnt’ (Count) is actually the sum of ‘casual’ and ‘registered’. Both - ‘casual’ and ‘registered’ explains all the variance there is in the target variable and R-Square will always be 1. It is important to remove these two variables from the predictor variable to build the correct model.

The next most correlated numerical feature with target variable is **Temperature**.

Pair-Plot of Numerical variables



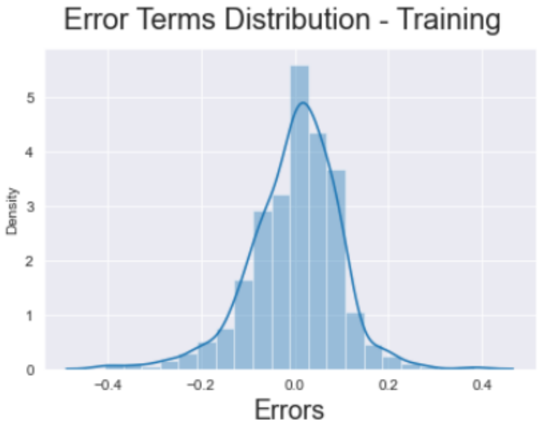
Question 4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

Answer 4: The assumptions of Multi-variate normality, Independence of Error and Homoscedasticity were validated on the Linear Regressions model built on the training dataset.

Multi-variate normality

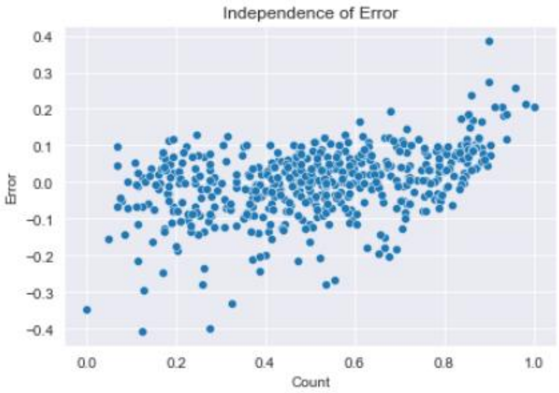
The Residuals or Error terms were calculated as the difference between the actual value and predicted values of the target variable from the Linear Regression.

The Residuals or Error terms when plotted were distributed normally with a mean of 0 and standard deviation of 1, as shown in the below picture. The assumption of Multi-variate normality was validated therefore.



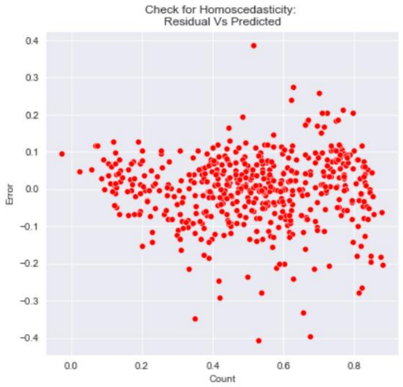
Independence of the Error

Next, the assumption of **Independence of the Error** was also validated for the value of target variable in the training set and the Residual Error. The Error terms were found independent (No pattern) for all possible values of target variable (Count) and are centred around 0 (which is their mean value). Shown in below picture.



Homoscedasticity

Also, the **Homoscedasticity** i.e., the variance of the error remains constant for all values of the target variable is also validated using the training dataset using below plot. The Error is not constant but good enough.



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

Answer 5 – The Final Linear Equation for the Model came out be as below.

*Count = (0.2356 * year) + (0.3937 * temp) + (-0.1525 * windspeed) + (-0.1460 * season_Spring) + (-0.0727 * month_July) + (0.0531 * month_Sep) + (-0.2748 * weather_Light_Snow) + (-0.0804 * weather_Mist_n_Cloudy) + Error*

As can be seen, the top three features with highest absolute value are given as below in descending order, which will have the highest effect on the demand of Bikes –

1. Temperature
2. Year
3. Weather – 3 (Light Rain and Light Snow)

General Subjective Questions

Question 1- Explain the linear regression algorithm in detail. (4 marks)

Answer 1 - Linear regression is a **supervised learning** algorithm to model the linear relationship between a dependent variable and one or more explanatory (Independent) variables and is used when target / dependent variable is a **continues** real number.

Linear Regression tries to establish the relationship between dependent variable vector (y) and one or more independent variable (X) using best fit line. The basic principle of ordinary least square (OLS) / Mean square error (MSE) is used to find the best fit line.

The below equation represents the Linear Model with B0 (Beta-0) as the intercept and B1 (Beta-1) to Bi (Beta-i) as the coefficient of the explanatory variables x and yi is the dependent variable. The Epsilon at the end represents Error (Noise).

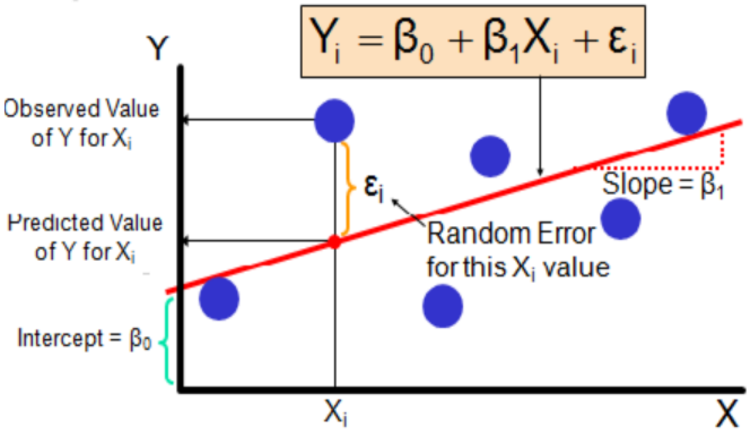
$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i = \mathbf{x}_i^\mathsf{T}\boldsymbol{\beta} + \varepsilon_i, \qquad i = 1, \ldots, n,$$

When the equation contains only one explanatory variable x, and the dependent variable y, the equation is called Simple Linear Regression.

As simpler representation of Simple Linear equation with

y = mx + c { m = Slope (Beta-1) & c is intercept (Beta-0}

In the picture below, the blue dot represents the actual value of independent variable Xi, while the red line represents the Linear Regression Model or the 'best fit line' with values of Beta-0, Beta-1 and Error Term


$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$
Y
Observed Value of Y for X_i
Predicted Value of Y for X_i
Intercept = β_0
ε_i
Random Error for this X_i value
Slope = β_1
X_i
X

Linear Regression Model (Red Line) and distribution of data points.

There are few basic assumptions which needs to be validated before applying the Linear Regression Model on a dataset, as follows.

- Linearity
    - The response variable (target) is a linear combination of the parameters (regression coefficients) and the predictor variables.

- Homoscedasticity.
    - This means that the variance of the errors does not depend on the values of the predictor variables and refers to a condition in which the variance of the residual, or error term, in a regression model is constant. That is, the error term does not vary much as the value of the predictor variable changes.
- Independence of errors.
    - This assumes that the errors of the response variables are uncorrelated with each other.
- Multivariate Normality
    - Multiple regression assumes that the residuals (Error terms) are normally distributed around a mean of 0 and standard deviation of 1.
- Non-Multicollinearity
    - Multicollinearity exists whenever an independent variable is highly correlated with one or more of the other independent variables in a multiple regression equation. Multicollinearity is a problem because it undermines the statistical significance of an independent variable. One independent variable is not only explaining the dependent variable but also one or more independent variable.

Performance of Linear Aggression Model

The Linear Regression model tries to find the best fit line using Ordinary Least Square method in which the goal is to minimize the sum of squared differences (Residual Sum of Square) between observed dependent variable in the given data set and those predicted by linear regression function.

The Line with least possible value of RSS fits the model best as it contains least Error.

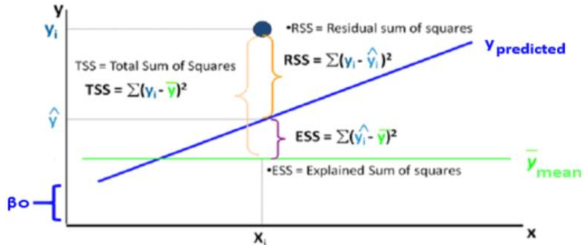
RSS

$$\text{Error} = \sum_{i=1}^{n} (Actual\_output - predicted\_output)\text{^}2$$

The other metric used is TSS (Total Sum of Square) which is a sum of the square of the difference variable actual value and mean value of the explanatory variables. TSS represents the mere average model.

TSS

$$\text{Error} = \sum_{i=1}^{n} (\text{Actual\_output} - \text{average\_of\_actual\_output})\text{^}2$$


y
y_i
•RSS = Residual sum of squares
$y_{predicted}$
TSS = Total Sum of Squares
$TSS = \sum(y_i - \bar{y})^2$
$RSS = \sum(y_i - \hat{y}_i)^2$
$\hat{y}$
$ESS = \sum(\hat{y}_i - \bar{y})^2$
$\bar{y}_{mean}$
•ESS = Explained Sum of squares
βo
X_i
x

The R-square is a metric which have a value between 0 and 1 and provides percentage indication how much of the variance of the target variable is explained by the Linear Regression Model.

$$R^2 = 1 - \frac{RSS}{TSS} = \frac{Explained\ Variation}{Total\ Variation}$$

The higher value for R-Square represents better model.

Question - 2. Explain the Anscombe’s quartet in detail. (3 marks)

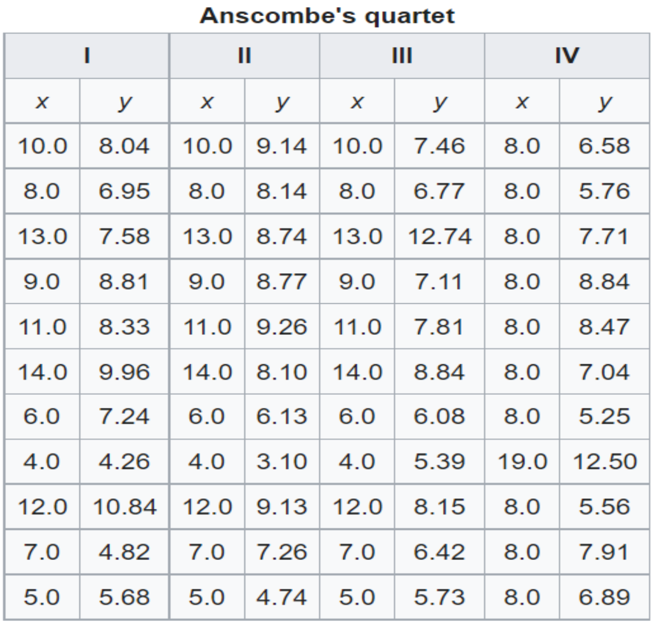
Answer -2: The basic concept of Anscombe’s quartet is that the descriptive statistics of a dataset does not provide all required information including distribution of datapoint and can be misleading in some cases.

In 1973, A statistician Francis Anscombe, created four data sets that had nearly identical simple descriptive statistics like Mean, yet have very different distributions but appeared very different when plotted on graph

Each of the four-dataset contained of eleven data (x, y) points. The dataset were constructed to demonstrate both the importance of plotting the data points before analysing them and the effect of outliers and other influential observations on statistical properties.

Anscombe described the experiment being intended to counter the impression among statisticians that "numerical calculations are exact, but graphs are rough." Anscombe’s quartet has been rendered as an actual musical quartet.

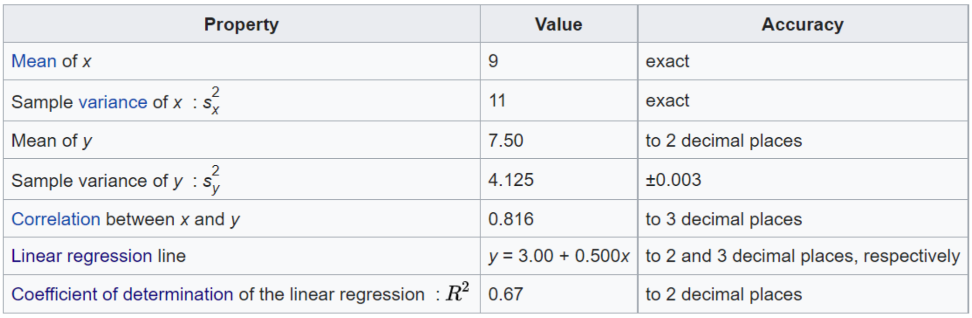
The 4 Dataset with 11 datapoints in Anscombe’s quartet



Anscombe's quartet

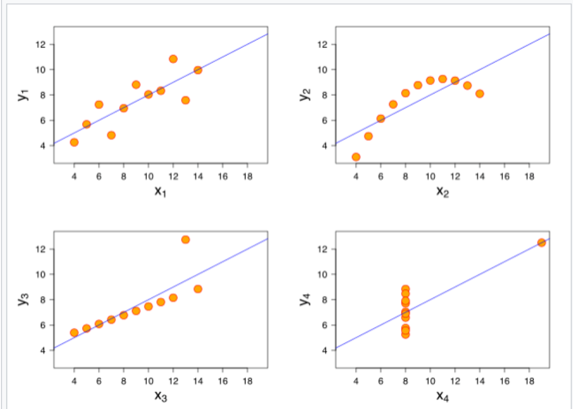
I		II		III		IV	
x	*y*	*x*	*y*	*x*	*y*	*x*	*y*
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	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

Descriptive Summary Statistics for all Four dataset and if we see the mean, variance and other stats for all 4 datasets is exactly same.



Property	Value	Accuracy
Mean of *x*	9	exact
Sample variance of *x* : s_x^2	11	exact
Mean of *y*	7.50	to 2 decimal places
Sample variance of *y* : s_y^2	4.125	±0.003
Correlation between *x* and *y*	0.816	to 3 decimal places
Linear regression line	*y* = 3.00 + 0.500*x*	to 2 and 3 decimal places, respectively
Coefficient of determination of the linear regression : R^2	0.67	to 2 decimal places

But when datasets are plotted on a X-Y plane, They gave a very different picture about the distribution and relationship between the variables.



So, it is always advisable to plot a graph between independent variables and dependent variable to understand if the relationship along with descriptive statistics to find the central tendencies and other metrics.

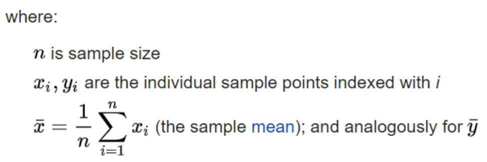
Question 3. What is Pearson's R? (3 marks)

Answer 3: Correlation is a measure of how strong a relationship is between two variables. There are several types of correlation coefficient to quantify the strength of the relationship , but the most popular one is Pearson's R or Pearson's correlation.
The linear relationship between the variables can be positive (0 < r <= 1) or negative or be none at all.
Pearson R describes three type of relationships which can be further described with an example as below.

- Positive linear relationship (0 < r <= 1).
  - Increase in the value of first variable follows up with increase in the value other variable and vice versa for decrease in value.
  - E.g., In most cases, the income of a person increases as his/her age increases.
- Negative linear relationship (-1 <= r < 0)
  - Increase in the value of first variable follows up with decrease in the value other variable and vice versa.
  - E.g., If the vehicle increases its speed, the time taken to travel decreases, and vice versa.
- No Correlation (r = 0).
  - Increase or decrease in the value of first variable have no effect whatsoever in the value other variable and vice versa.
  - E.g., Person's Age and temperature outside have no correlation if both are present in the same dataset.

Formula for Pearson R correlation Coefficient

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



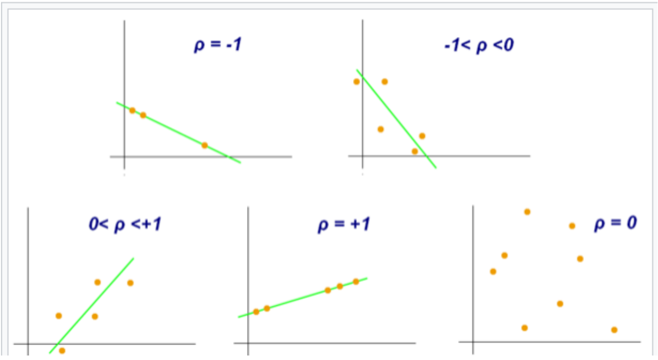
where:

n is sample size

x_i, y_i are the individual sample points indexed with i

$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ (the sample mean); and analogously for $\bar{y}$

The Pearson R value of -1 means Strong Negative Correlation and +1 (Strong Positive Correlation) while 0 represents no collinearity at all.



Question 4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

Answer 4 - Scaling is a method used to normalize the range of independent variables or features of data.

If the spread in the values of independent variables is very large, for example if few of the explanatory variables have fractional decimal value while other are in thousands (or millions), scaling is applied on the model to normalize all values of the independent variables between a definite range for the **linear model to perform better**.

Why Feature Scaling is important -

As the range of values of dependent variable varies widely, the model functions will not work efficiently properly without normalization, many model which calculates the distance between two points as Euclidean distance and If one of the features has a broad range of values, the distance will be governed by this particular feature. Therefore, it is advisable that the range of all features should be brought on same scale so that each feature contributes approximately proportionately to the final distance. Another reason to use feature scaling is that Gradient Descent algorithm converges much faster with feature scaling than without it.

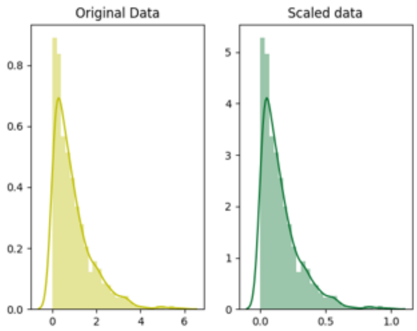
Normalization

Normalization refers to method of scaling all the feature variables using the maximum (1) and minimum value (0) present in the dataset into a range between 0 and 1. The method is also known as min-max scaling or min-max normalization, is the simplest method and consists in rescaling the range of features to scale the range in [0, 1] (or sometimes [−1, 1]).

The general formula for a min-max of [0, 1] is given as:

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

x' refers to the scaled value.
x refers to actual value in the dataset.
min(x) refers to the minimum value for x feature in the dataset.
max(x) refers to the maximum value for x feature in the dataset.



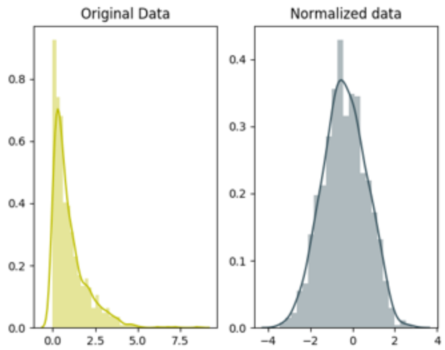
Normalization

Standardization

The Standardization (or **Z-score normalization**) is a method of scaling where the independent variables are rescaled to ensure having a mean of 0 and standard deviation of 1. The value of feature variable generally lies between -3 and 3. This method is preferable to avoid the effect of outliers in the dataset.
The equation used in the Standardization is as below:

$$x' = \frac{x - \bar{x}}{\sigma}$$

x' refers to the scaled value.
x refers to actual value in the dataset.
x-bar refers to the mean value for feature x in the dataset.
Sigma refers to the Standard deviation feature x in the dataset.



Standardization

The Standardization handles the cases better than Normalization when you have outliers in the feature variables of the dataset. As normalizing your data will scale most of the data to a small interval, which means all features will have the same scale but does not handle outliers well.

Note that If a model algorithm is not distance-based, feature scaling is unimportant.

Question - 5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

Answer -5 - Variance inflation factor(VIF) is a metric to detects multicollinearity in regression analysis. Multicollinearity is present when there's correlation between explanatory variables (i.e. independent variables) in a model.

The VIF estimates how much the variance of a regression coefficient is inflated due to multicollinearity in the model.

VIF for a predictor variable Xi is derived from R-square value obtained when all other independent variables (X1- Xi-1) are used to predict the value of Xi.

Formula for VIF of i-th predictor variable is given by -

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

<u>Case for VIF to be Infinity -</u>

If there is a perfect correlation between the i-th predictor variable and any other predictor variable in the dataset then R-square will always be equal to 1, because that correlated variable (along with other variable) will be able to explain all the variance in the i-th variable. This situation will cause VIF to turn out as 'INF' (Infinity) as denominator will be equal to zero if R-sq is 1.

The 1/0 is invalid fraction with value tending to Infinity. The 'INF' value for VIF indicates Multi-collinearity amongst the predictor variables and R-square of value 1.

Question - 6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

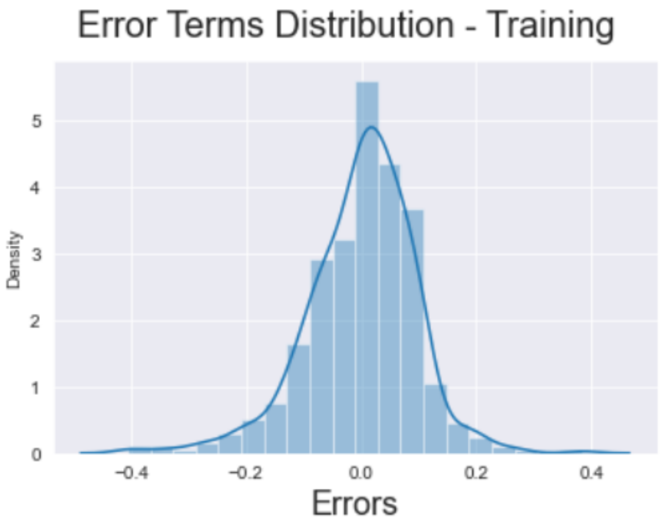
Answer – 6: A **Q–Q (quantile-quantile)** plot is a probability plot, which is a graphical method for comparing two probability distributions by comparing their quantiles against each other

It is basically a graphical tool to help assess if a set of data came from some theoretical distribution such as a Normal or Exponential.

A Q-Q 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 line which is roughly straight (line y = x). If the scatter points on the Q-Q plots follows a straight line then the distributions of the data are normal.

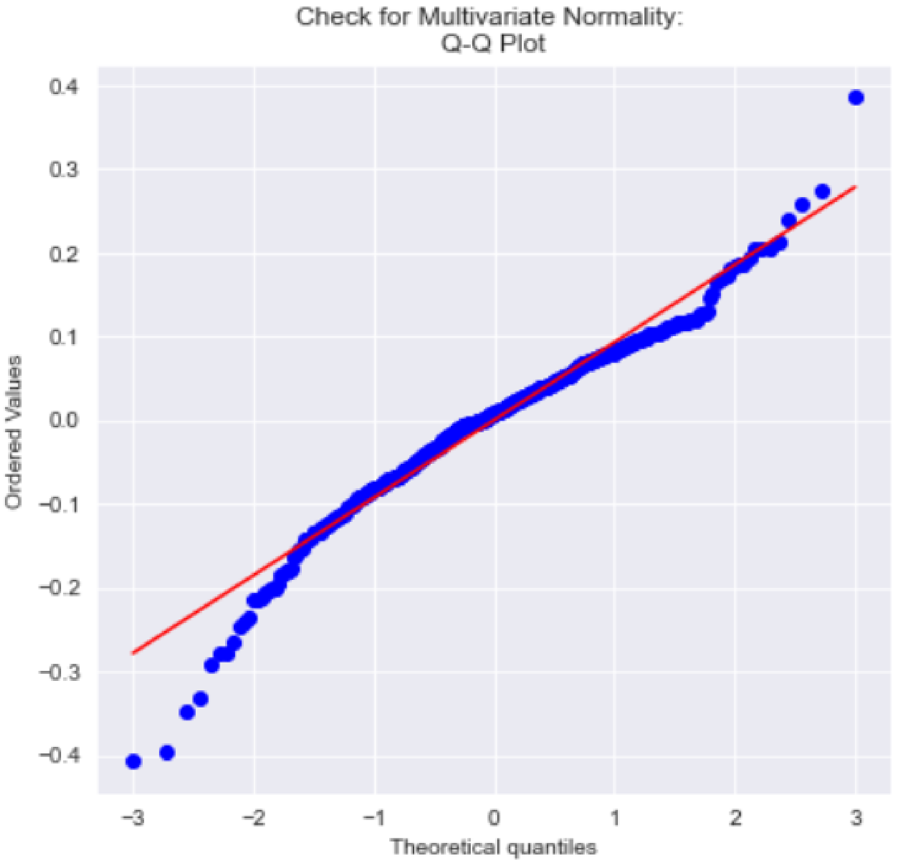
The Q-Q Plots are used to validate the assumption of Multi-variate normality in Linear Regressions Models. Multi-variate normality assumes that the Error or Residuals are normally distributed for all values for target variable.

Once the Linear Model is built for target variables, The Residuals, the difference between the target variable and its predicted value, are plotted then if the distribution in normal then the assumption of Multi-variate normality is validated.



Same validation can be proved using Q-Q plots as shown in the below plot in our Boom Bike Sharing assignment.

Since the data points on the scatter plot closely follows the straight line (y=x) on the X-Y plane, it shows that Distribution of **Error terms is normally distributed.**


Check for Multivariate Normality:
Q-Q Plot
Ordered Values
0.4
0.3
0.2
0.1
0.0
−0.1
−0.2
−0.3
−0.4
−3
−2
−1
0
1
2
3
Theoretical quantiles