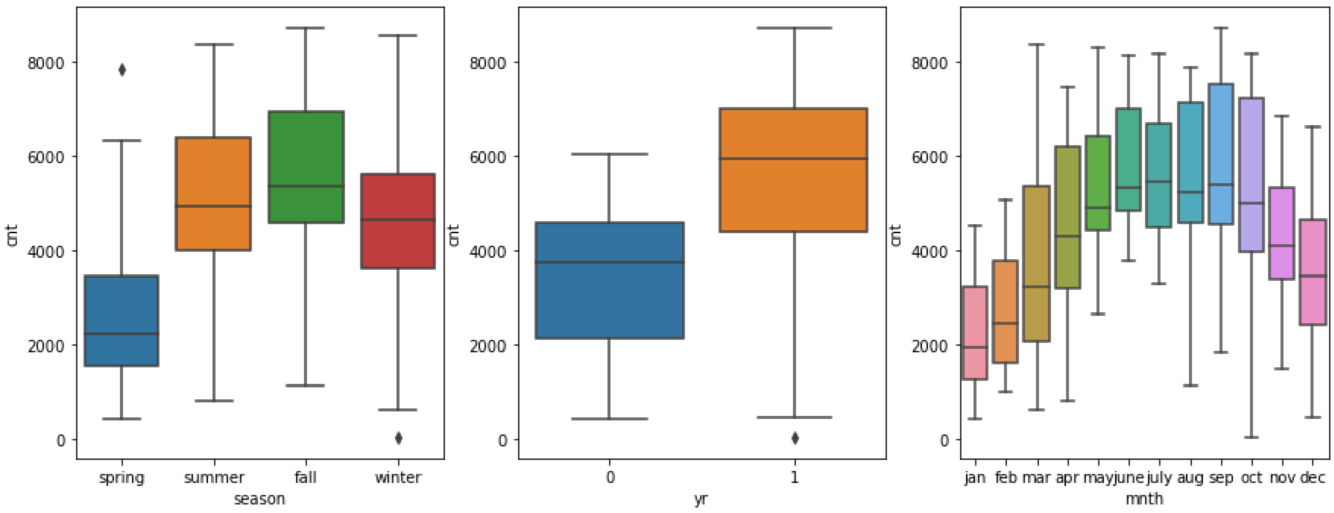
Linear Regression – Subjective Questions

Nachiket Shembekar

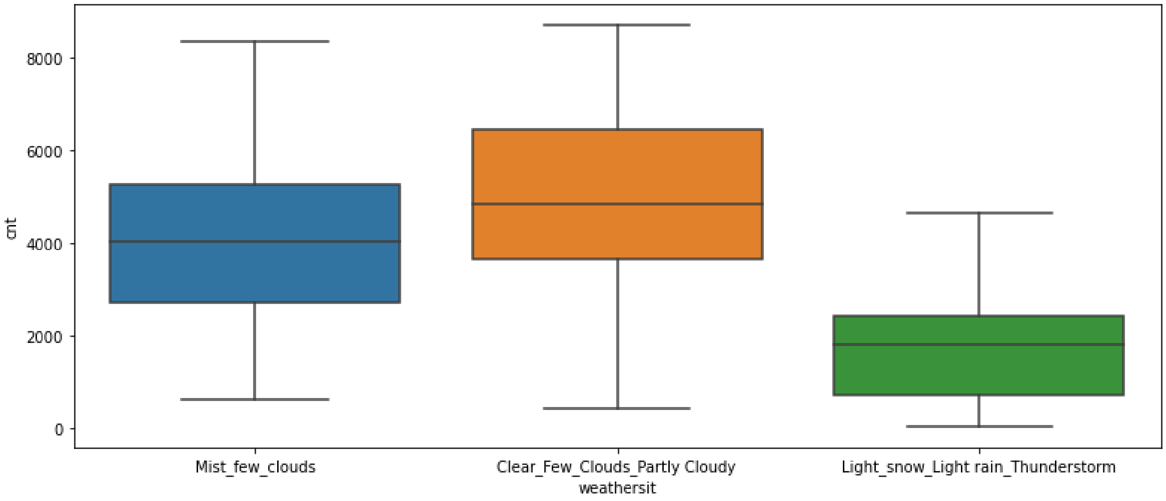
Assignment-based Subjective Questions

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

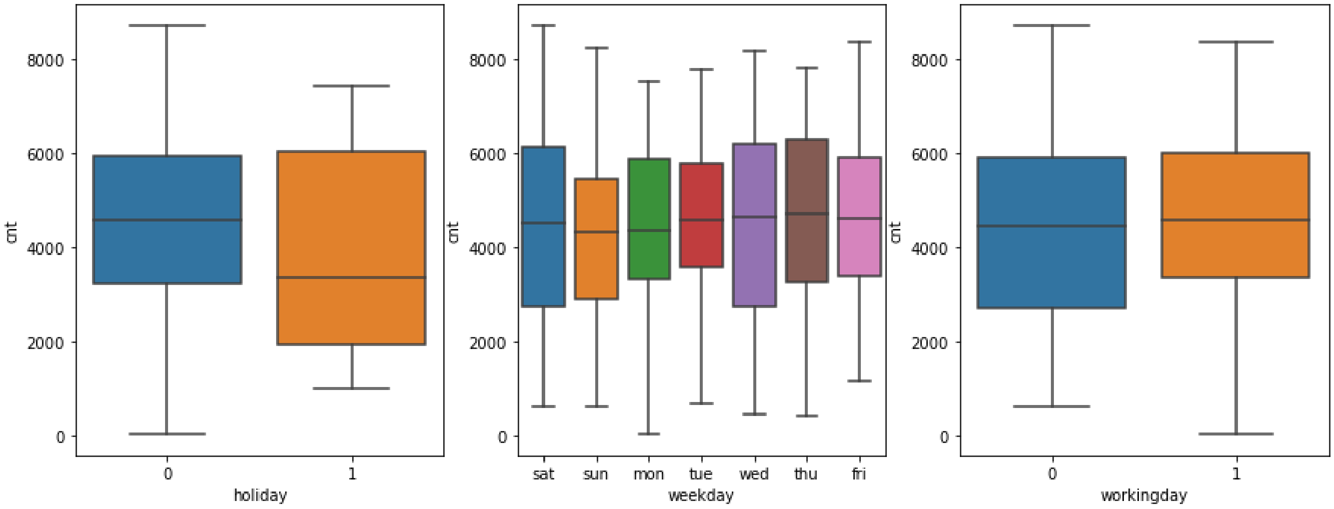
Below is the analysis of categorical variables and their effect on the dependent variable.



1. season : Highest bike rentals are observed in Fall season followed by Summer. Spring season is least preferred. Boombikes can expand during summer and fall with strategic advertising and may give some discounted rates during Spring.
2. yr : There's significant rise in the bike rentals in 2019 as compared to 2018, which may be due to increased awareness for the benefits and/or good marketing.
3. mnth :The bike rentals are higher between June to October, particularly in September & October, which may be because of good weather conditions.



4. weathersit : Weather plays an important role, since most customers prefer to rent bikes during the days with clear sky or partly cloudy. They avoid extreme weather conditions such as snow, thunderstorm which is logical. Adverse weather conditions such as cloudy, windy or extreme ones such as snow/rain/thunderstorm have negative impact as customers do not prefer to rent bikes during such conditions.



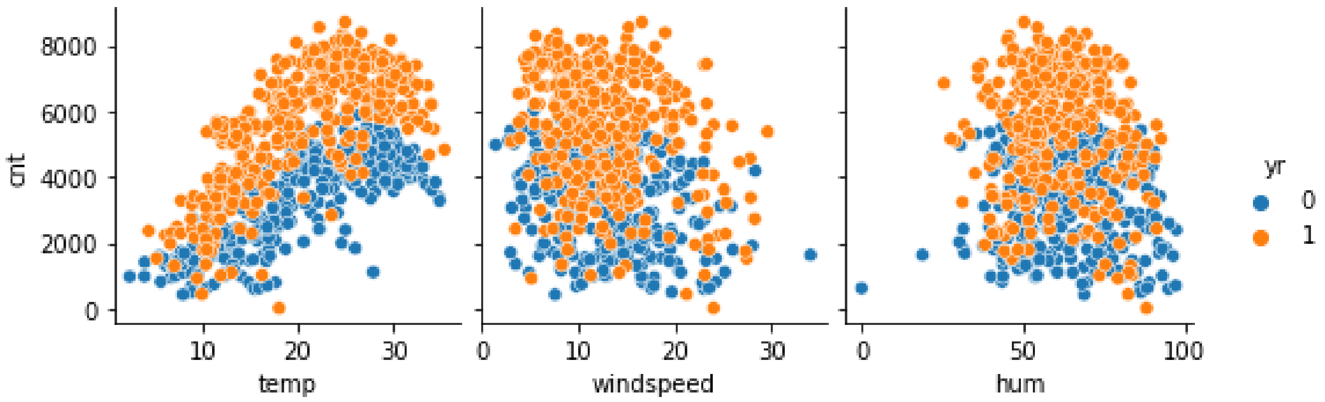
5. weekday/workingday/holiday : Not much impact on the rental counts whether it's workingday, which means this variable won't play significant role our decision making. The rentals are lesser on holidays. Probable reasons could be them out of town or even due to higher rates on those days. Boombikes can think about having some exciting offers or discounts especially around holidays.

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

Using one-hot encoding the dummy variables are created to cover the range of values of categorical variable. Each dummy variable has 1 and 0 values. 1 is used to depict the presence and 0 for absence of the respective category. This means if the category variable has 3 categories, there will be 3 dummy variables.

The **drop_first = True** is used while creating dummy variables to drop the base/reference category. It helps in reducing the extra column created during dummy variable creation and avoiding the multi-collinearity getting added into the model if all dummy variables are included. The reference category can be easily deduced where 0 is present in a single row for all the other dummy variables of a particular category.

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



Temperature (temp) seems to be the most significant and dominant feature and has highest correlation (0.63) with the target variable. Warmer temperature seems to attract more customers to rent the bikes and increases business.

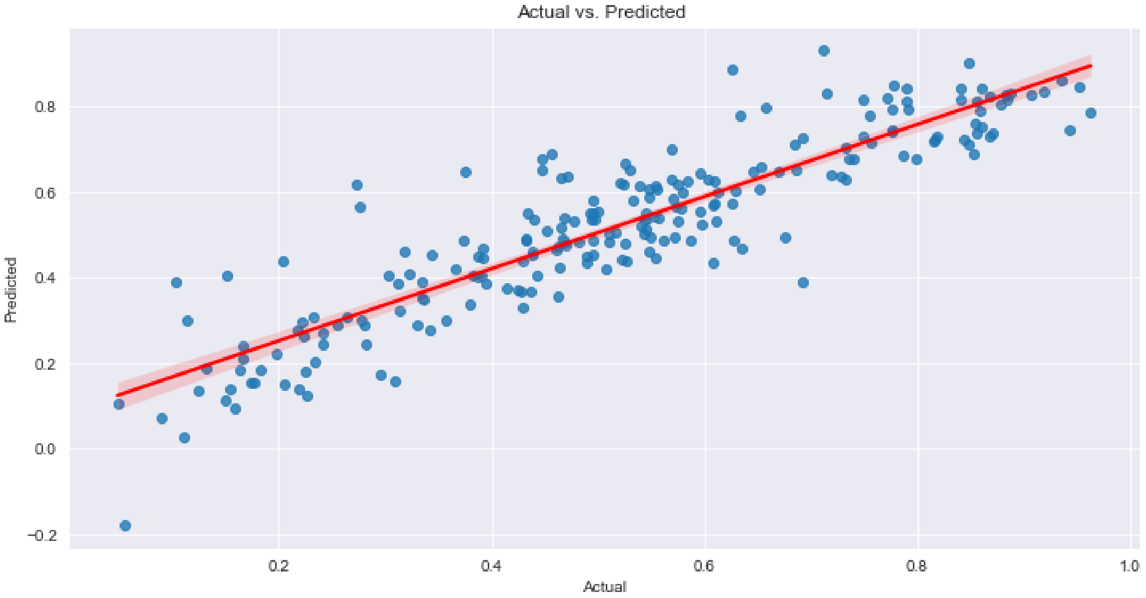
“atemp” is the derived parameter from temp, humidity and windspeed, hence it was dropped in the model preparation.

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

Below is the explanation of how each of the Linear regression assumptions was validated.

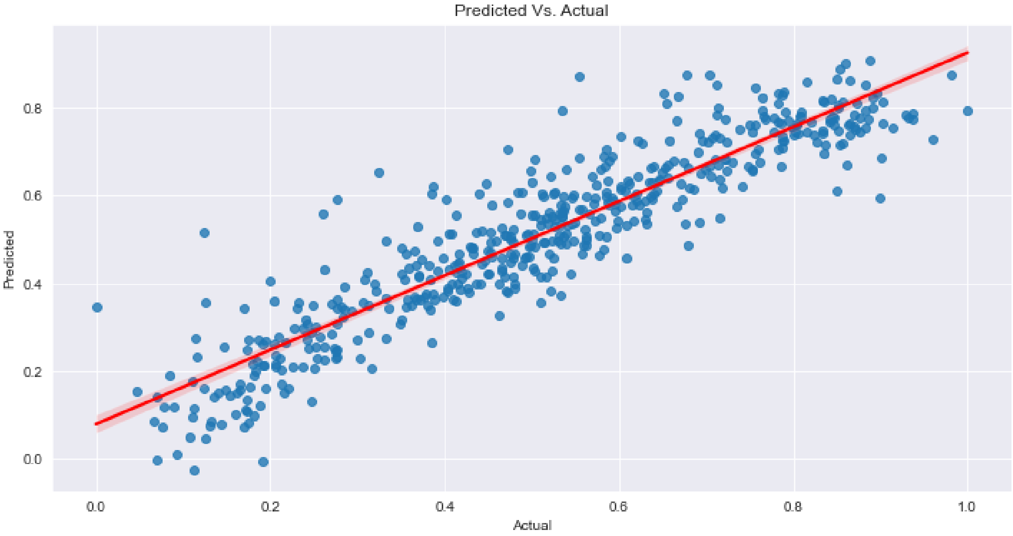
1. Linearity

   The linearity is validated by looking at the points distributed symmetrically around the diagonal line of the actual vs predicted plot as shown in the below figure.

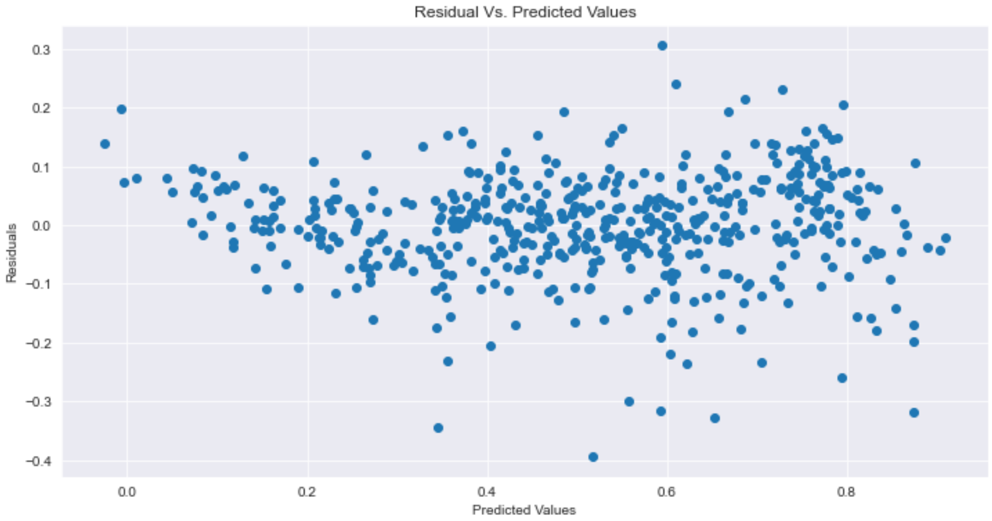


2. Homoscedasticity

   Error Terms nearly have a Constant Variance; hence they can be considered to follow the Assumption of Homoscedasticity.

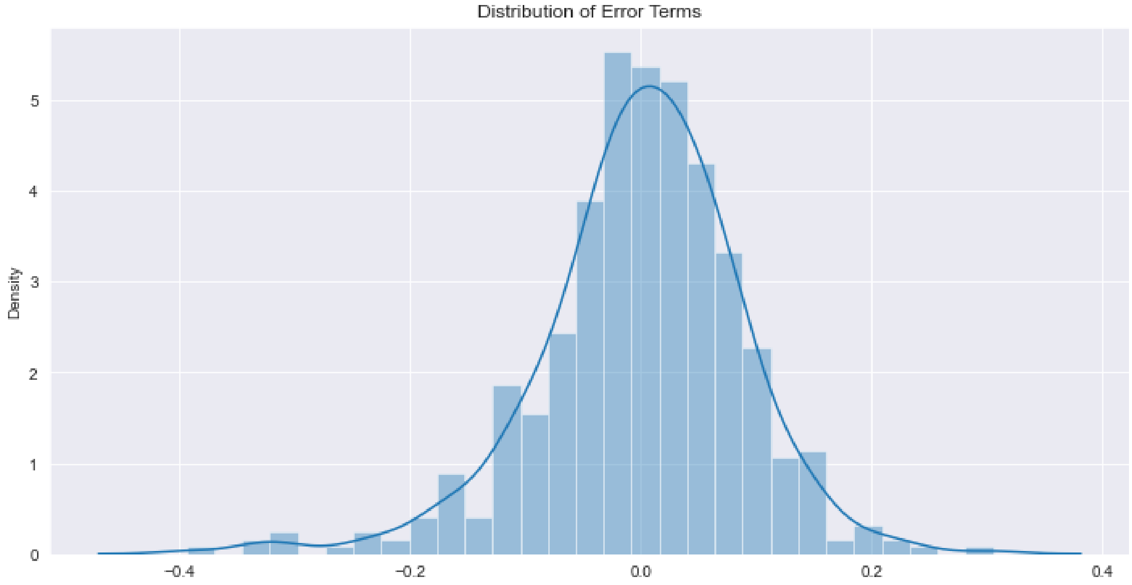


3. Independence Error Terms :
   No specific Pattern is observed in the Error Terms with respect to Prediction. Thus Error terms can be considered independent of each other.



4. The error terms should be normally distributed

   Error Terms are normally Distributed with mean Zero. Hence our model is following Normality of error terms.



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

Based on the final model, below are the features having most significant contribution;

1. Temp : Temperature is the most significant feature with coefficient of 0.57. Warmer temperature attracts more customers to rent the bikes and increases business.
2. Yr : Year is the 2nd most significant feature with coefficient of 0.23. There's an increase in the demand in 2019 as compared to 2018. This could be due to increased awareness or an effect of good marketing that Boombikes may be doing already. This shows a positive trend to invest more in future.
3. Season : Winter season has been preferred by the customers over the rainy seasons, where conditions are unfavourable.

General Subjective Questions

1. Explain the linear regression algorithm in detail.

What is Linear Regression ?

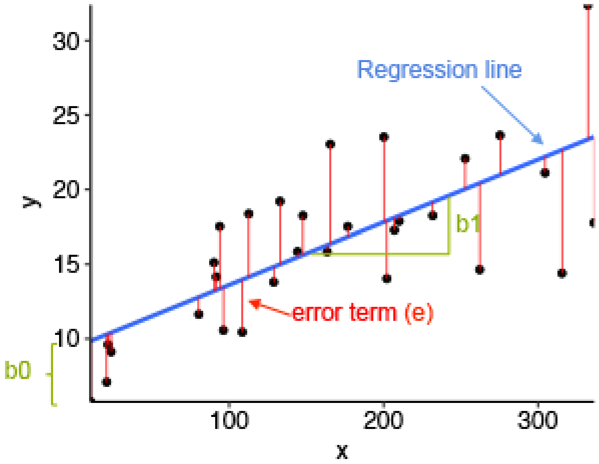
Linear Regression is the Supervised Machine Learning (ML technique where models are trained on labeled data i.e., output variable is provided in these types of problems) model in which the model finds the best fit linear line between the independent and dependent variable i.e., it finds the linear relationship between the dependent and independent variable.

The algorithm uses the best fitting line to map the association between independent variables with dependent variable.

Objective of Linear Regression :

A Linear Regression model's main aim is to find the best fit linear line and the optimal values of intercept and coefficients such that the error is minimized.

Error is the difference between the actual value and Predicted value and the goal is to reduce this difference.



In the above diagram,

- x is dependent variable which is plotted on the x-axis and y is the dependent variable which is plotted on the y-axis.
- Black dots are the data points i.e the actual values.
- b_o or $\beta_0$is the intercept which is 10 and b_1 (or β_1) is the slope of the x variable.
- The blue line is the best fit line predicted by the model i.e the predicted values lie on the blue line.

The vertical distance between the data point and the regression line is known as error or residual. Each data point has one residual and the sum of all the differences is known as the Sum of Residuals/Errors.

- Residual/Error = Actual values – Predicted Values
- Sum of Residuals/Errors = Sum(Actual- Predicted Values)
- Square of Sum of Residuals/Errors = (Sum(Actual- Predicted Values))2

Types of Linear Regression :

- There are 2 types of linear regression algorithms
    - Simple Linear Regression – Single independent variable is used and the model has to find the linear relationship of it with the dependent variable
        - $Y = \beta_0 + \beta_1 X$ is the line equation used for SLR.
    - Multiple Linear Regression – Multiple independent variables are used.
        - $Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \in$ is the line equation for MLR.
    - $\beta_0 = value\ of\ the\ Y\ when\ X = 0\ (Y\ intercept)$
    - $\beta_1, \beta_2, \ldots, \beta_p = Slope\ or\ the\ gradient.$

Assumptions of Linear Regression

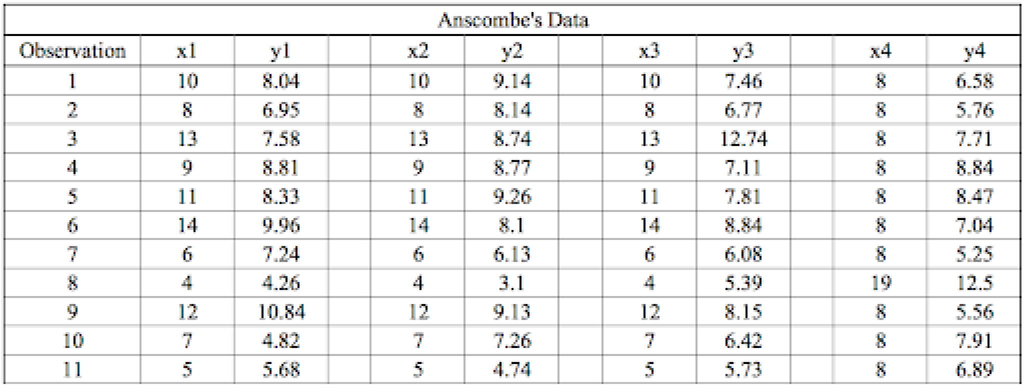
1. Linearity: It states that the dependent variable Y should be linearly related to independent variables.
2. Normality: The X and Y variables should be normally distributed. Histograms, KDE plots, Q-Q plots can be used to check the Normality assumption.
3. Homoscedasticity: The variance of the error terms should be constant i.e., the spread of residuals should be constant for all values of X. This assumption can be checked by plotting a residual plot. If the assumption is violated, then the points will form a funnel shape otherwise they will be constant.
4. Independence/No Multicollinearity: The variables should be independent of each other i.e., no correlation should be there between the independent variables. To check the assumption, we can use a correlation matrix or VIF score. If the VIF score is greater than 5 then the variables are highly correlated.
5. The error terms should be normally distributed. Q-Q plots and Histograms can be used to check the distribution of error terms.
6. No Autocorrelation: The error terms should be independent of each other. Autocorrelation can be tested using the Durbin Watson test. The null hypothesis assumes that there is no autocorrelation. The value of the test lies between 0 to 4. If the value of the test is 2 then there is no autocorrelation.

2. Explain the Anscombe’s quartet in detail.

Statistics like variance and standard deviation are usually considered good enough parameters to understand the variation of some data without looking at every data point

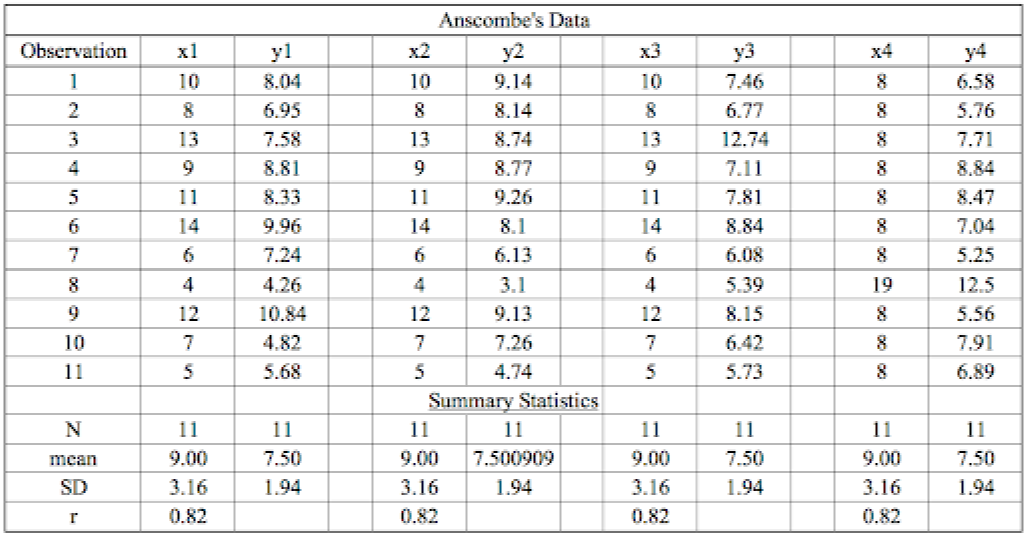
Anscombe’s quartet was constructed in 1973 by statistician Francis Anscombe to illustrate the importance of plotting data before analysing and building the model. These four data sets have nearly the same statistical observations, which provide the same information (involving variance and mean) for each x and y point in all four data sets. However, while plot these data sets, they look very different from one another.

We can define these four plots as follows:



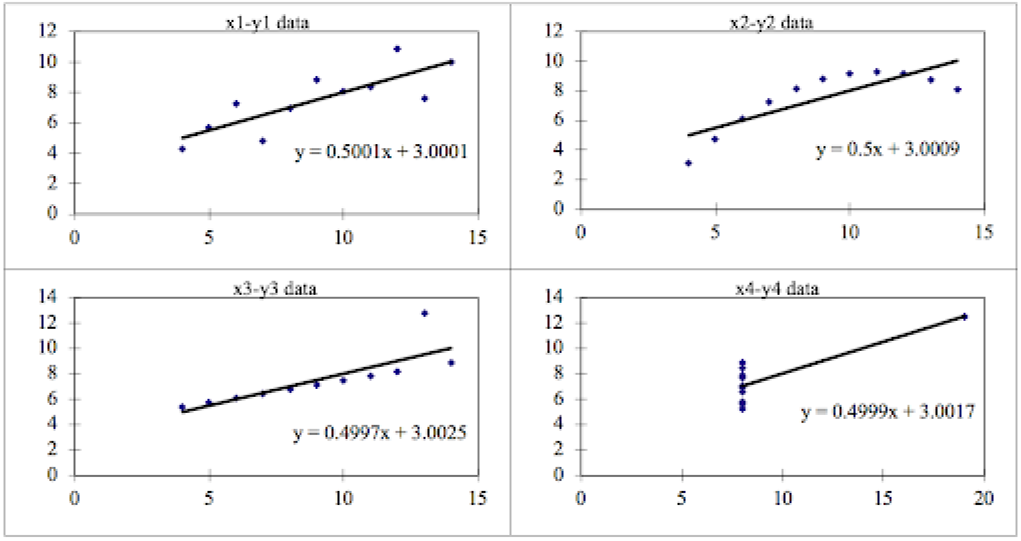
Anscombe's Data											
Observation	x1	y1		x2	y2		x3	y3		x4	y4
1	10	8.04		10	9.14		10	7.46		8	6.58
2	8	6.95		8	8.14		8	6.77		8	5.76
3	13	7.58		13	8.74		13	12.74		8	7.71
4	9	8.81		9	8.77		9	7.11		8	8.84
5	11	8.33		11	9.26		11	7.81		8	8.47
6	14	9.96		14	8.1		14	8.84		8	7.04
7	6	7.24		6	6.13		6	6.08		8	5.25
8	4	4.26		4	3.1		4	5.39		19	12.5
9	12	10.84		12	9.13		12	8.15		8	5.56
10	7	4.82		7	7.26		7	6.42		8	7.91
11	5	5.68		5	4.74		5	5.73		8	6.89

The statistical information for these four data sets are approximately similar. We can compute them as follows:



Anscombe's Data											
Observation	x1	y1		x2	y2		x3	y3		x4	y4
1	10	8.04		10	9.14		10	7.46		8	6.58
2	8	6.95		8	8.14		8	6.77		8	5.76
3	13	7.58		13	8.74		13	12.74		8	7.71
4	9	8.81		9	8.77		9	7.11		8	8.84
5	11	8.33		11	9.26		11	7.81		8	8.47
6	14	9.96		14	8.1		14	8.84		8	7.04
7	6	7.24		6	6.13		6	6.08		8	5.25
8	4	4.26		4	3.1		4	5.39		19	12.5
9	12	10.84		12	9.13		12	8.15		8	5.56
10	7	4.82		7	7.26		7	6.42		8	7.91
11	5	5.68		5	4.74		5	5.73		8	6.89
					Summary Statistics						
N	11	11		11	11		11	11		11	11
mean	9.00	7.50		9.00	7.500909		9.00	7.50		9.00	7.50
SD	3.16	1.94		3.16	1.94		3.16	1.94		3.16	1.94
r	0.82			0.82			0.82			0.82	

However, when these models are plotted on a scatter plot, each data set generates a different kind of plot that isn't interpretable by any regression algorithm, as shown below :



Anscombe’s Quartet Four Datasets :

1. Data Set 1: fits the linear regression model pretty well.
2. Data Set 2: cannot fit the linear regression model because the data is non-linear.
3. Data Set 3: shows the outliers involved in the data set, which cannot be handled by the linear regression model.
4. Data Set 4: shows the outliers involved in the data set, which also cannot be handled by the linear regression model.

It tells us about the importance of visualizing data before applying various algorithms to build models. This suggests the data features must be plotted to see the distribution of the samples that can help you identify the various anomalies present in the data (outliers, diversity of the data, linear separability of the data, etc.). Moreover, the linear regression can only be considered a fit for the data with linear relationships and is incapable of handling any other kind of data set.

3. What is Pearson’s R?

Pearson's Correlation Coefficient is also referred to as Pearson's r, the Pearson product-moment correlation coefficient (PPMCC), or bivariate correlation.

It measures the strength between the different variables and the relation with each other. The Pearson’s R returns values between -1 and 1. The interpretation of the coefficients are:

• r = 1 means the data is perfectly linear with a positive slope (i.e., both variables tend to change in the same direction)

• r = -1 means the data is perfectly linear with a negative slope (i.e., both variables tend to change in different directions)

• r = 0 means there is no linear association

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

Where:

r=correlation coefficient

xi=values of the x-variable in a sample

x¯=mean of the values of the x-variable

yi=values of the y-variable in a sample

y¯=mean of the values of the y-variable

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

- **What is Scaling:**

  Scaling is step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. It helps in speeding up calculations in the algorithm. Scaling just affects the coefficients and none of the other parameters like t-statistic, F-statistic, p-values, R-squared, etc.

- **Why is Scaling performed :**
  Most of the times the feature data is collected at public domains where the interpretation of variables and units of those variables are kept open collect as much as possible. This results into the high variance in units and ranges of data. If scaling is not done on these data sets, then the chances of processing the data without the appropriate unit conversion are high. Also, the higher the range then higher the possibility that the coefficients are impaired to compare the dependent variable variance. The scaling only affects the coefficients. The prediction and precision of prediction stays unaffected after scaling.

- **Difference between normalized & Standardized scaling :**

  1. Normalization/Min-Max scaling – The Min max scaling normalizes the data within the range of 0 and 1. The Min max scaling helps to normalize the outliers as well.

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

  2. Standardization converges all the data points into a standard normal distribution where mean is 0 and standard deviation is 1.

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

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

VIF formula is as below

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

Whenever there’s a perfect correlation between 2 independent variables, R^2 becomes 1. Thus, VIF in this case becomes infinite.

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). To solve this problem, one of the variables from the dataset causing this perfect multicollinearity should be dropped.

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

Quantile-Quantile (Q-Q) plot is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal, exponential, or Uniform distribution.

In linear regression, when we have the training and test data sets received separately and then we can confirm using a Q-Q plot that both the data sets are from populations with the same distributions.

The theoretical distributions could be of type normal, exponential, or uniform. This is another method to check the normal distribution of the data sets in a straight line with patterns explained below

- **Use & Interpretation :**
  - Similar distribution: If all the data points of the quantile are lying around the straight line at an angle of 45 degrees from the x-axis.
  - Y values < X values: If y-values quantiles are lower than x-values quantiles.
  - X values < Y values: If x-values quantiles are lower than y-values quantiles.
  - Different distributions – If all the data points are lying away from a straight line.

- **Importance & Benefits :**
  - Distribution aspects like shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot
  - The plot can be used with sample size as well.