# 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?**

A. There were 7 categorical variables provided in the dataset out of which 3 categorical variables were binary in nature i.e., 3 variables were of the form 'Yes' or No' they are as follows:

- Year (2019) - Yes or No
- Holiday – Yes or No
- Working day – Yes or no

And the remaining variables had various levels of categories. Following are the categorical variables with various levels of categories in each of them:

- Season: 1 – Spring, 2 – Summer, 3 – Fall, 4 - Winter
- Month: 1 – Jan, 2 – Feb, 3 – Mar, 4 – Apr, 5 – May, 6 – Jun, 7 – Jul, 8 – Aug, 9 – Sep, 10 – Oct, 11 – Nov, 12 – Dec
- Weekday: 0 – Tuesday, 1 – Wednesday, 2 – Thursday, 3 – Friday, 4 – Saturday, 5 – Sunday, 6 – Monday,
- Weather Situation: 1 – clear, 2 – misty, 3 – rainy, 4 - thunderstorm - But there are no entries in the provided dataset with thunderstorm value. Hence thunderstorm can be ignored

Following Inference was derived when demand was visualized against each of the categorical variable listed above:

a. The demand is least in Spring season and peaks in Fall, Summer and Winter respectively

b. The demand is low in the months of Jan, Feb, Nov, Dec and there is substantial increase in demand from Mar to October

c. There is no significant pattern noticed for demand against weekdays apart from the fact that demand is least on Tuesdays

d. The demand is least when the weather situation is rainy whereas the demand is high in misty and clear sky weather situation

e. The demand is more on days when there is no public holiday.

f. The demand is more or less the same whether it is a working day or not. No significant pattern noticed.

g. The demand has increased drastically 2019 when compared to 2018. There is a fair amount of chance that with popularity, the demand should increase subsequently

(P.T.O)

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

A. Linear Regression model only takes in numeric data to build a model. But in real world there will be categorical variables as well which might be useful in predicting a model. One way to deal with categorical variable is creating dummy variables. The key idea behind creating dummy variables is that for a categorical variable with ‘n’ levels, we create ‘n-1’ new columns each indicating whether that level exists or not using a zero or one. The reason why we choose n-1 variables is can be explained with the help of following example:
Season categorical variable has 4 levels: spring, summer, fall and winter. When converted into dummy variable, each of the level is represented as follows:

Fall: 1 0 0 0

Spring: 0 1 0 0

Summer: 0 0 1 0

Winter: 0 0 0 1

The above information can be effective represented as follows:

Fall: 0 0 0

Spring: 1 0 0

Summer: 0 1 0

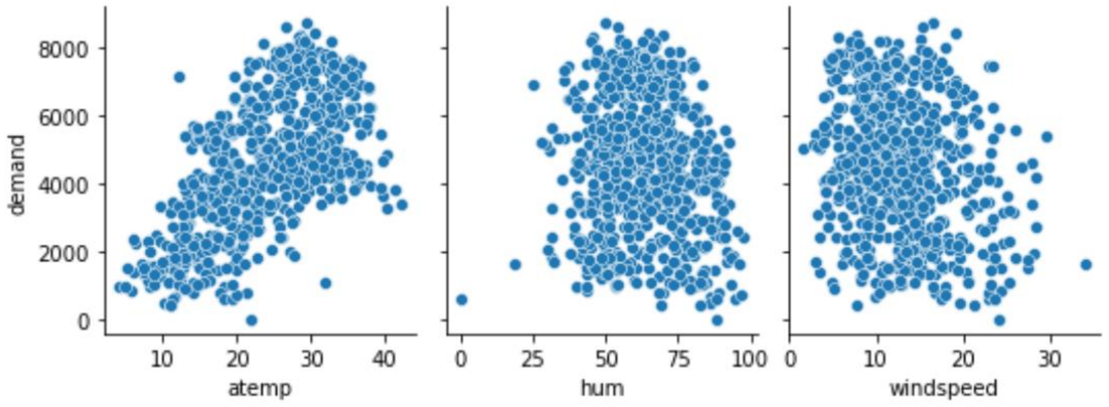
Winter: 0 0 1

Hence essentially only 3 dummy variables can be used describe/identify which season it is. This is done to simplify the analysis (more technically, this will avoid making the data matrix rank-deficient)

(P.T.O)

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

A. ‘atemp’ is the continuous numeric variable that has the highest correlation with the target variable demand

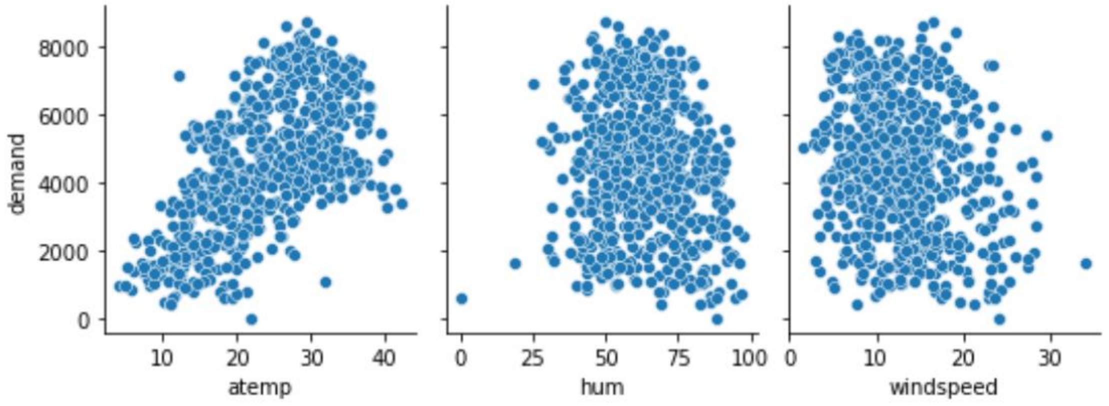


4. How did you validate the assumptions of Linear Regression after building the model on the training set?

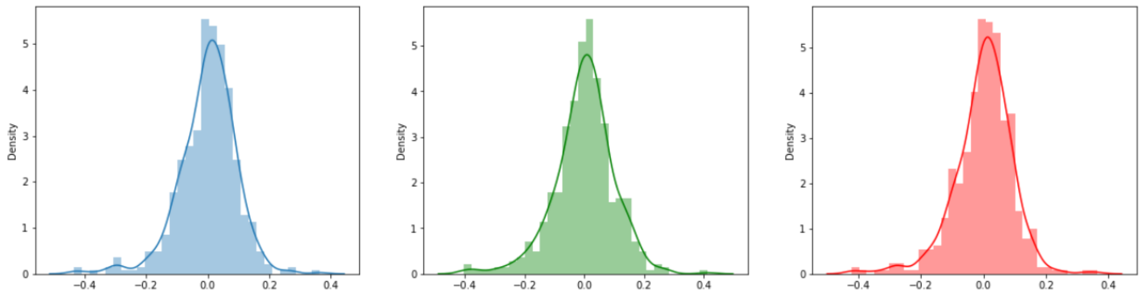
A. The assumptions of Linear Regression are:

   a. There is linear relationship between X-axis and Y-axis
   b. Residuals/Error terms are normally distributed (not X, Y)
   c. Residuals/Error terms are independent of each other
   d. Residuals/Error terms have constant variance (homoscedasticity)

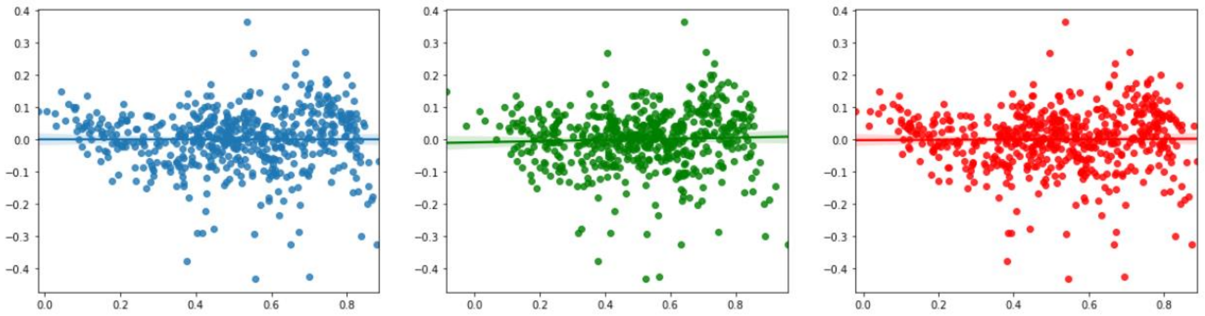
   Using pair plot we observed that there is at least 1 variable that has linear relationship with our target variable.



   We then plotted distribution plot for residuals (y_train-y_pred) to visualize if the error terms are normally distributed about 0.



We then plotted a scatter plot of the error terms/residuals against y_pred to observe if the error terms are independent of each other and have constant variance



From the above scatter plot, we inferred that the error terms are independent as the points are scattered independently without any noticeable pattern and also inferred that the error terms have constant variance as these points are distributed evenly on either side of 0.0 with a min variance of -0.4 and max variance of +0.4.

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

A. Following are the top 3 features contributing towards the demand

   a. atemp (Actual feeling temperature in Celsius) – atemp contributes positively towards the demand of the shared bike service. Unit increase in atemp accounts to 0.43 units of increase in demand of bikes.

   b. weathersit_rainy (Weather Situation: Light rain to Thunderstorm) – weathersit_rainy contributes negatively towards the demand of shared bikes. Unit decrease in weathersit_rainy accounts to 0.29 units of decrease in demand of bikes.

   c. Yr (Year) – yr contributes positively towards the demand of shared bike service. With each year, there is increase in popularity of the service accounting for a lot of demand in bikes each year. With each year, there is 0.23 units of increase in demand of bikes.

Note: It is also noteworthy to point out that there is 0.06 units of increase in the month of September being represented by variable – mnth_sep

# General Subjective Question:

1. **Explain the linear regression algorithm in detail.**

A. Regression algorithm is a type of Machine Learning model that falls under Supervised Learning Method in which the model is trained and built on historical data with labels and thereby predict the output target variable which is always a continuous variable.
Linear Regression is a type of Regression which is built on the assumption that the relationship between the dependent variable and the independent variable is linear (straight line). There are 2 types of linear regressions:

   a. **Simple Linear Regression** – Explains the relationship between a dependent variable and one independent variable using a straight line

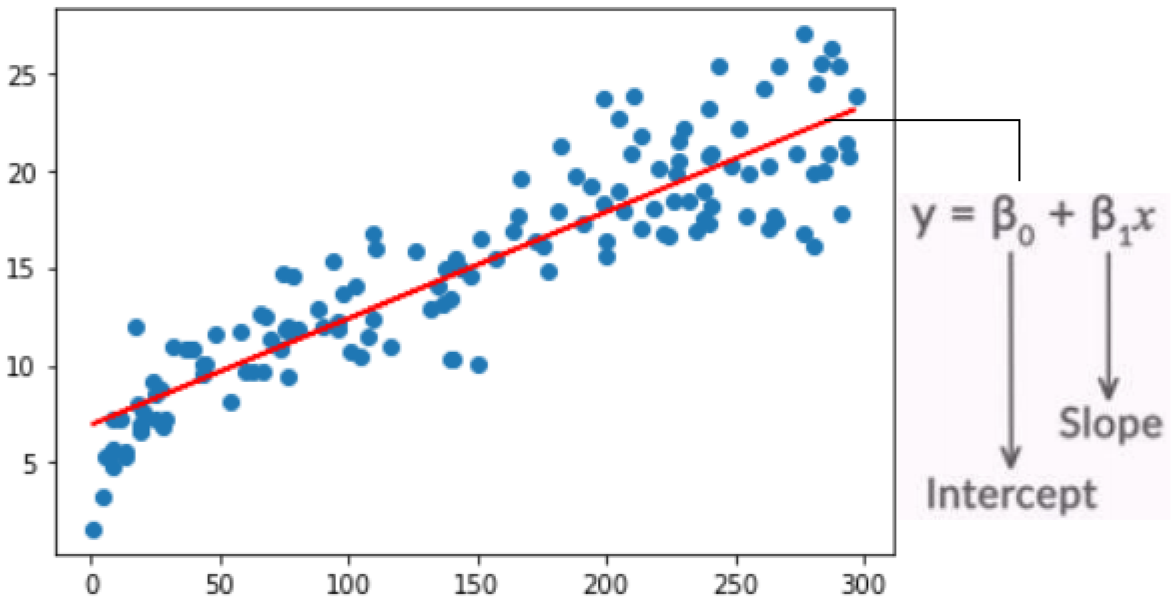
   b. **Multiple Linear Regression** – Explains relationship between one dependent variable and several independent variables and is represented by hyperplane

## Simple Linear Regression:

As pointed out earlier a simple linear regression model is represented using a straight line. The equation of this line is given by:

$$Y = \beta_0 + \beta_1 X$$

The same can be visually represented as follows:



Graph of data in X-axis linearly related to the data in Y-axis

- The unexplained error in the actual value of Y and the value predicted by our model is called Residual.
- To build a model with the best fit, we need to make sure that cost function, the Sum of Squares of Residuals is minimum. This absolute cost function is called Residual Sum of Squares which is given by the following formula:

$$\mathrm{RSS} = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2$$

- But cost function being an absolute value, not relative/universal. This relative approach of identifying the error is given by Total Sum of Squares which is given by the formula:

$$\mathrm{TSS} = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

- We still need better ways to find the effectiveness of the model to describe the variance in Y for given X. This can be done with the help of R-squared(R2). R2 is given by the following formula:

R2 = Explained deviation/Total Deviation => 1 - (unexplained deviation/total deviation)

Therefore, R2 = 1 - (RSS/TSS)

Higher R2 signifies better goodness of fit of the model and thereby better predictions

There are certain assumptions that are made for linear regression to hold good:

i. There is linear relationship between Y axis and X axis
ii. The residuals/error terms are normally distributed
iii. The error terms are independent of each other
iv. The error terms have constant variance (homoscedasticity)

Multiple Linear Regression:

Multiple linear regression (MLR) is an extension of Simple linear regression. Multiple linear regression defines relationship between 1 dependent and multiple independent variables.

Turning to multiple linear regression is beneficial because of the following reasons:

- Adding more variables helps add information about the variance in Y (R2)
- Explanatory power will increase with increase in variables

The equation of the hyperplane formed by Multiple linear regression is given by:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon$$

- Model now fits a ‘hyperplane’ instead of a line as there are multiple dimensions in contrast to what we have seen in Simple Linear Regression
- The cost function is still given by residual sum of squares(RSS) and the best fit model is obtained by minimizing RSS
- The assumptions from Simple Linear Regression still hold good:
    - Independent, normally distributed error terms that have constant variance and zero mean
- The inference part in multiple linear regression also, largely, remains the same as Simple linear regression

Although the above-mentioned concepts are similar to what we have noticed in Simple Linear regression there are some new considerations to be taken, which are as follows:

- Overfitting: Adding more isn’t always useful. Model may overfit by becoming too complex and end up fitting too well only for the train dataset instead of generalizing the dataset
- Multicollinearity: Presence of predictor variables that are highly correlated to each other

While building an MLR model, we need to be mindful to select features that do not follow multicollinearity or that do not overfit the train data.

The selection of such features can be done in the following approaches:

i. Manual Feature Elimination:
   a. Start with all variables and build a model
   b. Drop features that are least helpful for prediction (high p-value)
   c. Drop features that are redundant (high VIF where VIF = 1/(1-R2))
   d. Rebuild model and repeat

ii. Automated Feature Elimination:
   a. Recursive Feature Elimination (RFE)
   b. Forward/Backward/Stepwise selection based on AIC
   c. Regularization – Lasso

iii. Balanced Approach: Automated feature elimination married with Manual feature elimination gives balanced approach

- The effectiveness of MLR to describe the variance in Y for given set of X is given by Adjusted R-squared(R2) which is given by:

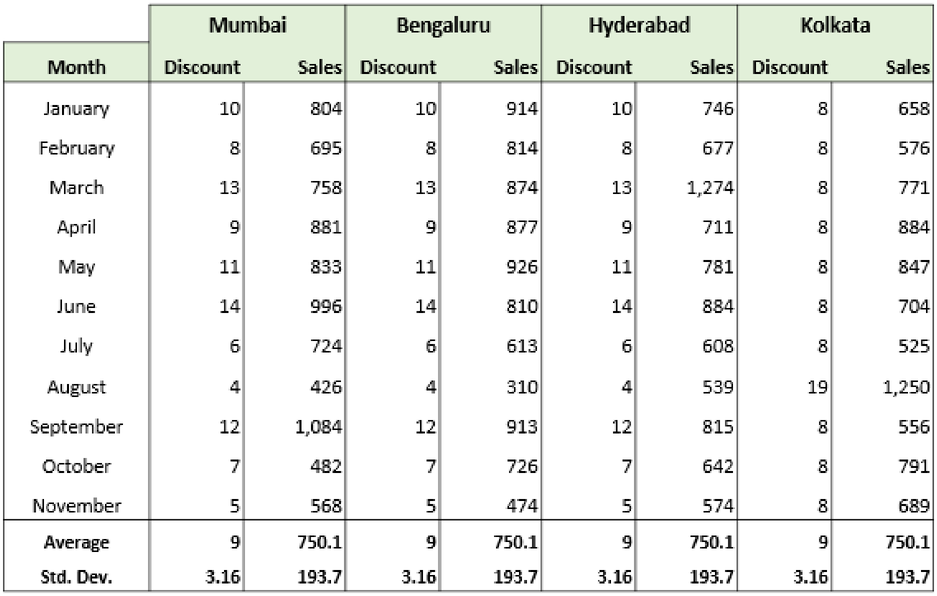
$$Adjusted\ R^2 = 1 - \frac{(1 - R^2)(N - 1)}{N - p - 1}$$

Higher the adjusted R2, better is the capacity of the MLR model to define the variance in Y for change in X.

2. **Explain the Anscombe's quartet in detail.**

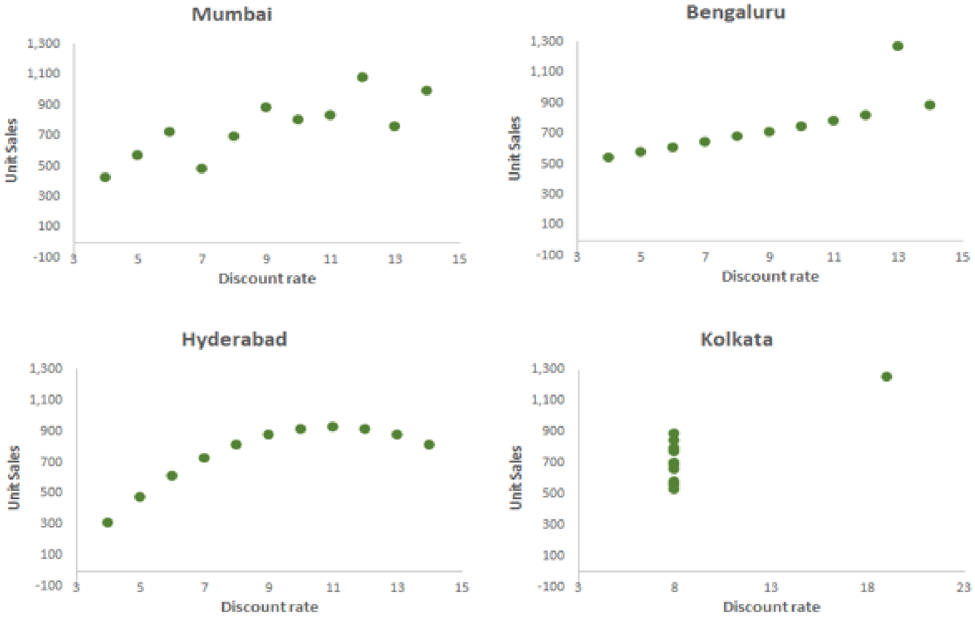
A. Anscombe's quartet comprises four data sets that have nearly identical simple descriptive statistics yet have very different distributions and appear very different when graphed. It is very easy to be deceived by the numbers and summary statistics. This can be better explained with the help of an example:

The table below explains discount and sales observed by various branches of a franchise in 4 different cities in India over 11 months. Each of the branches had similar average sales and discount rates, and the corresponding standard deviations were similar as well, as shown below.



| | Mumbai | | Bengaluru | | Hyderabad | | Kolkata | |
|---|---|---|---|---|---|---|---|---|
| Month | Discount | Sales | Discount | Sales | Discount | Sales | Discount | Sales |
| January | 10 | 804 | 10 | 914 | 10 | 746 | 8 | 658 |
| February | 8 | 695 | 8 | 814 | 8 | 677 | 8 | 576 |
| March | 13 | 758 | 13 | 874 | 13 | 1,274 | 8 | 771 |
| April | 9 | 881 | 9 | 877 | 9 | 711 | 8 | 884 |
| May | 11 | 833 | 11 | 926 | 11 | 781 | 8 | 847 |
| June | 14 | 996 | 14 | 810 | 14 | 884 | 8 | 704 |
| July | 6 | 724 | 6 | 613 | 6 | 608 | 8 | 525 |
| August | 4 | 426 | 4 | 310 | 4 | 539 | 19 | 1,250 |
| September | 12 | 1,084 | 12 | 913 | 12 | 815 | 8 | 556 |
| October | 7 | 482 | 7 | 726 | 7 | 642 | 8 | 791 |
| November | 5 | 568 | 5 | 474 | 5 | 574 | 8 | 689 |
| **Average** | **9** | **750.1** | **9** | **750.1** | **9** | **750.1** | **9** | **750.1** |
| **Std. Dev.** | **3.16** | **193.7** | **3.16** | **193.7** | **3.16** | **193.7** | **3.16** | **193.7** |

However, the patterns in the underlying data and the difference became apparent when visualized through appropriate plots as shown below:



1. The scatter plot for Mumbai appears to be simple linear relationship, corresponding to two variables correlated where y could be modelled as gaussian with mean linearly dependent on x
2. The distribution of data in scatter plot for Bengaluru is linear, but should have a different regression line due to the presence of an outlier.
3. The scatter plot for Hyderabad is not distributed normally.
4. The scatter plot for Kolkota shows an example that one high-leverage point is enough to produce a high correlation coefficient, even though the other data points do not indicate any relationship between the variables

It is clear from the above visualization that each of the branches had actually employed a different strategy to calculate its discount rate, and the sales numbers were also quite different across all of them. It is difficult to draw this type of insight and understand the difference between each of the branches using raw numbers alone.

The Anscombe's quartet was constructed by statistician Frances Anscombe **to counter the notion that** "**numerical calculations are exact, but graphs are rough**."

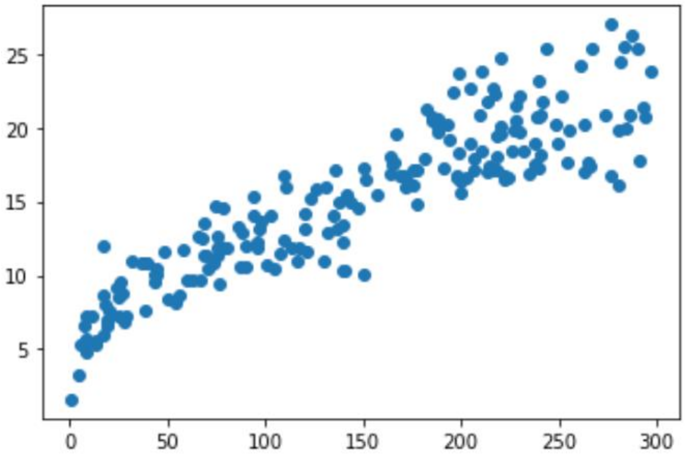
**3. What is Pearson's R?**

A. Pearson's R also known as correlation coefficient is the measure of the strength of relationship between 2 numeric variables. It is mathematically denoted by the letter 'r' and its value varies from –1 to +1.

- –1 denotes linear relationship with negative slope
- +1 denotes linear relationship with positive slope
- 0 denotes no linear relationship

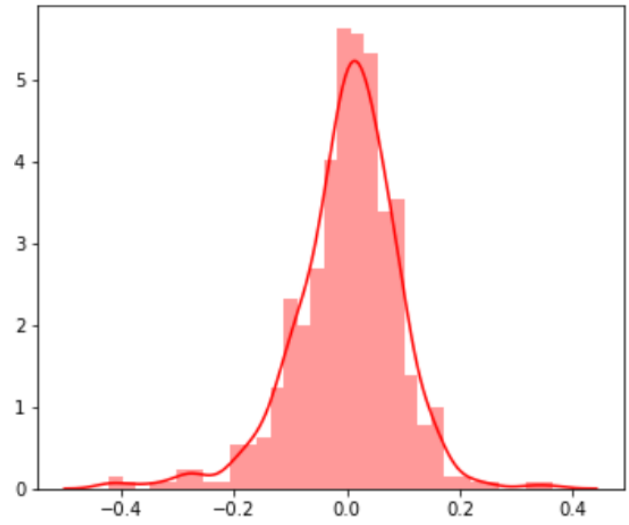
There are certain assumptions for Pearson's R to hold good:

i. There is linear relationship between the involved variables

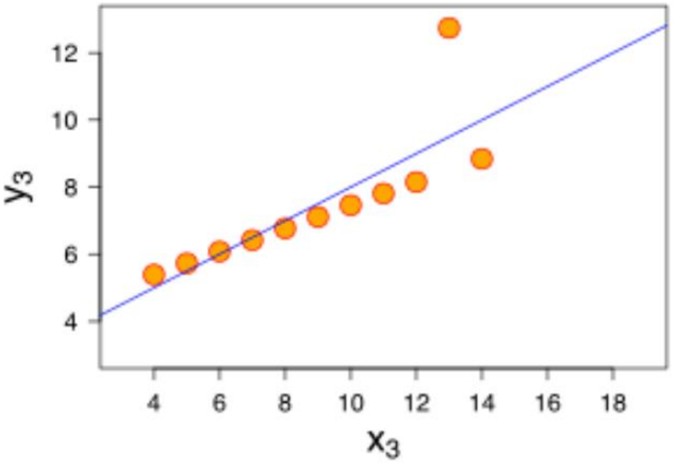


ii. Both variables are normally distributed I.e., have a gaussian curve about the mean when visualized

(P.T.O)

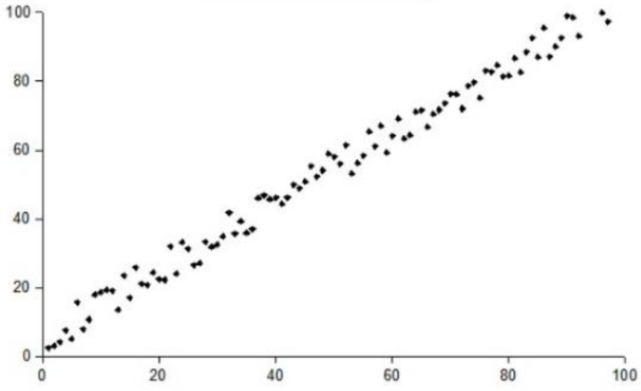


iii. There are no significant outliers to skew the value of Pearson’s coefficient



iv. Both variables are continuous numeric variables

v. There is constant variance among the error terms to display homoscedasticity



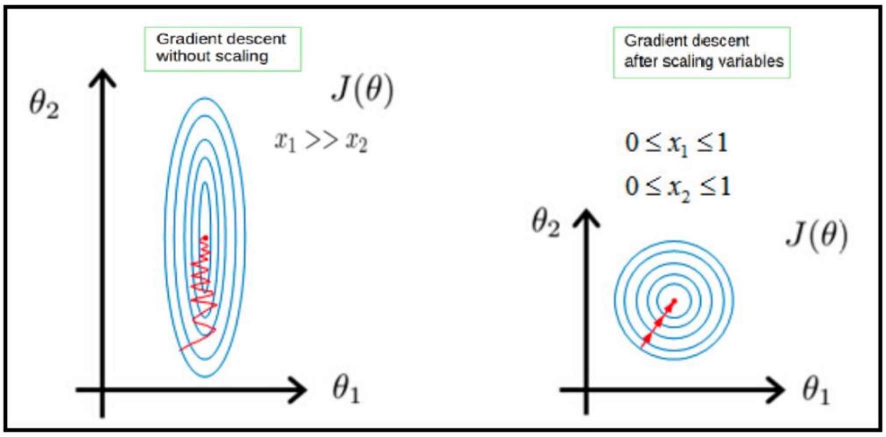
4. **What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?**

A. Feature Scaling is a technique to standardize/normalize the independent features present in the data in a fixed range. It is performed during the data pre-processing to handle highly varying magnitudes or values or units.

When you have a lot of independent variables in a model, a lot of them might be on very different scales which will lead a model with very weird coefficients that might be difficult to interpret. Hence scaling is required for 2 primary reasons:

i. Ease of interpretation

ii. Faster convergence for gradient descent methods as shown below:



It is noteworthy to mention that scaling just affects the coefficients and none of the other parameters like t-statistic, F statistic, P-Values or R-square.

In general scaling only affects the interpretation of variables but not the prediction of the model.

There are 2 types of scaling: 1. MinMax Scaling/ Normalized Scaling and 2. Standardised Scaling

MinMax Scaling: The variables are scaled in such a way that all the values lie between zero and one using the maximum and the minimum values in the data.

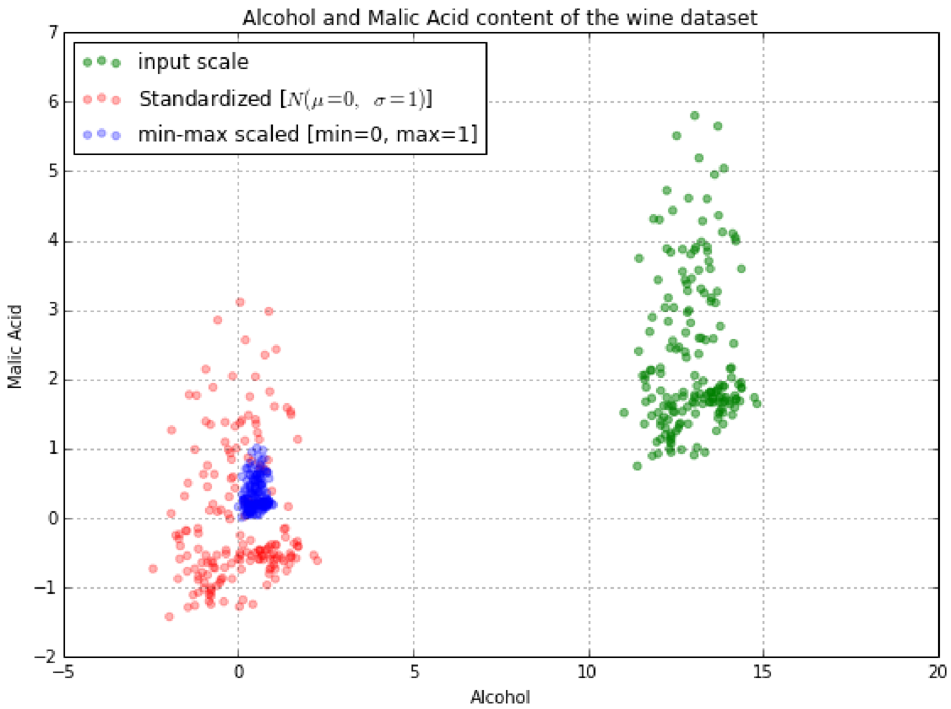
$$x = \frac{x - mean(x)}{sd(x)}$$

Standardized Scaling: The variables are scaled in such a way that their mean is zero and standard deviation is one.

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

The difference between normalizing and standardizing can be visualized in the graph of a clustering example shown below:

(P.T.O)



It is noteworthy to point out that as part of normalization because all datapoints are represented between 0 and 1, it takes care of the outliers present in the original dataset whereas standardization will not handle outliers.

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

A. VIF basically helps explaining the relationship of one independent variable with all the other independent variables. In order words, to determine VIF, we fit a regression model between the independent variables

VIF is given by the following formula:

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

Mathematically speaking, an expression 1/x will near infinity as the value of x tends to near 0. In our present scenario, VIF will be equal to infinity only if the value of R-square tends to near 1. In other words, we see VIF=infinity when there is perfect correlation among the predictor variables. This will lead to multi collinearity and there by affect our inference and interpretation of the model and hence we will not be able to present our analysis to the business and help them out. There are various ways to deal with this. 1 common method is to drop these highly correlated features.

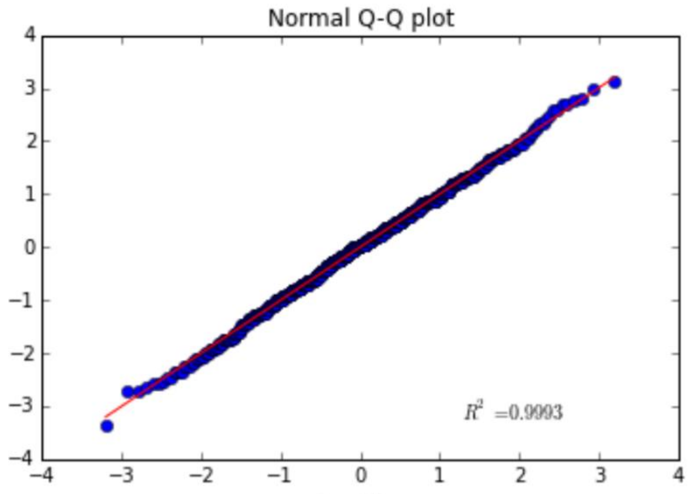
VIF can range from 1 to infinity. The common heuristic we follow for the VIF values is:

- > 10: Definitely high VIF value and the variable should be eliminated.
- > 5: Can be okay, but it is worth inspecting.
- < 5: Good VIF value. No need to eliminate this variable.

(P.T.O)

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

A. A Q-Q plot is a 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 that's roughly straight



The Q in Q-Q plot stands of ‘Quantile’. 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.

Importance of Q-Q plot in linear regression is as follows:

a. The quantile-quantile (q-q) plot is a graphical technique for determining if two data sets come from populations with a common distribution.
b. As shown above, while plotting the scatter plot, a 45-degree reference line is also plotted.
c. If the two sets come from a population with the same distribution, the points should fall approximately along this reference line.
d. 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.
e. In real world scenarios for models where linear regression is implemented, we would have received 2 different sets of data, 1 for training and 1 for test. In these scenarios we can use Q-Q in our linear regression to assess if the two datasets we received are from population with common distribution or not. This would help us reliably evaluate our model.

The advantages of q-q plot are as follows:

a. The sample sizes do not need to be equal.
b. Many distributional aspects can be simultaneously tested. For example, shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot.