Answer Sheet:
Submitted By: Bibhu Sundar Sahoo

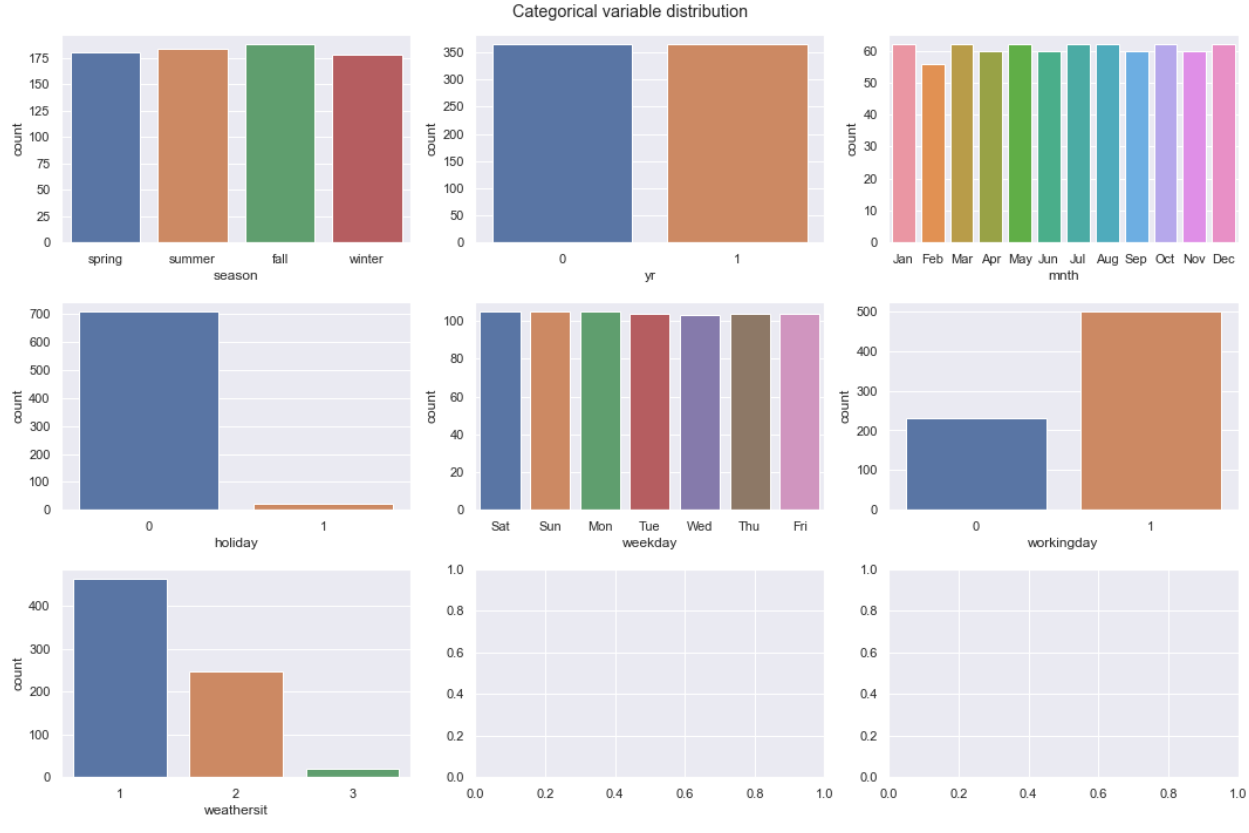
Assignment-based Subjective Questions

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

The categorical model used in the model includes- 'season', 'yr', 'mnth', 'holiday', 'weekday', 'workingday', 'weathersit'

Few of the numerical columns for instance have ordinality but are not continuous for this analysis so we have treated it as a categorical variable.

In the EDA exercise we put a rule to treat any numerical columns as categorical if the maximum number of unique values in the column is less than 15. As we have a dataset of 730 rows, it is fair to apply such rule.



Season- Summer and Fall are the most desired season for biking.

Year(yr)- We see increase in bike hire from 2018 to 2019

Month(mnth)- Bike hire is prominently seen around the year except in February.

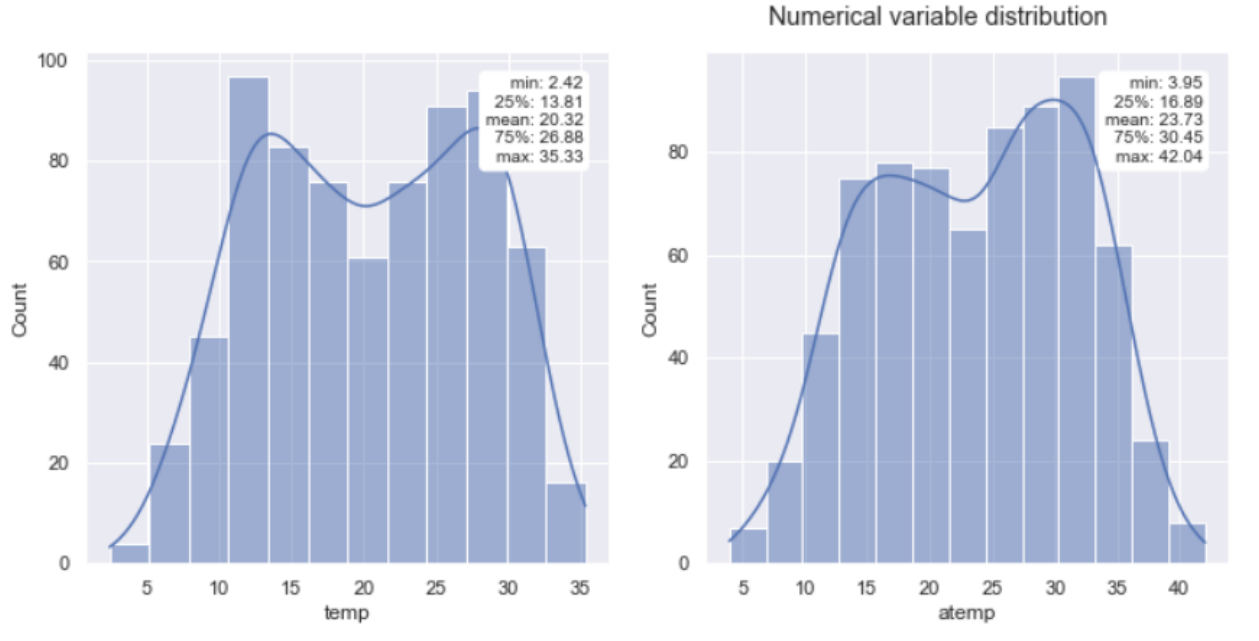
Weekdays(Weekday) - Friday is the peek day for bike hire while Sunday is the worst.

Weather Situation(weathersit)- Clearly on Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog weather the bike hire slows down. As the weather situation goes bad and extreme we see less bike hire.

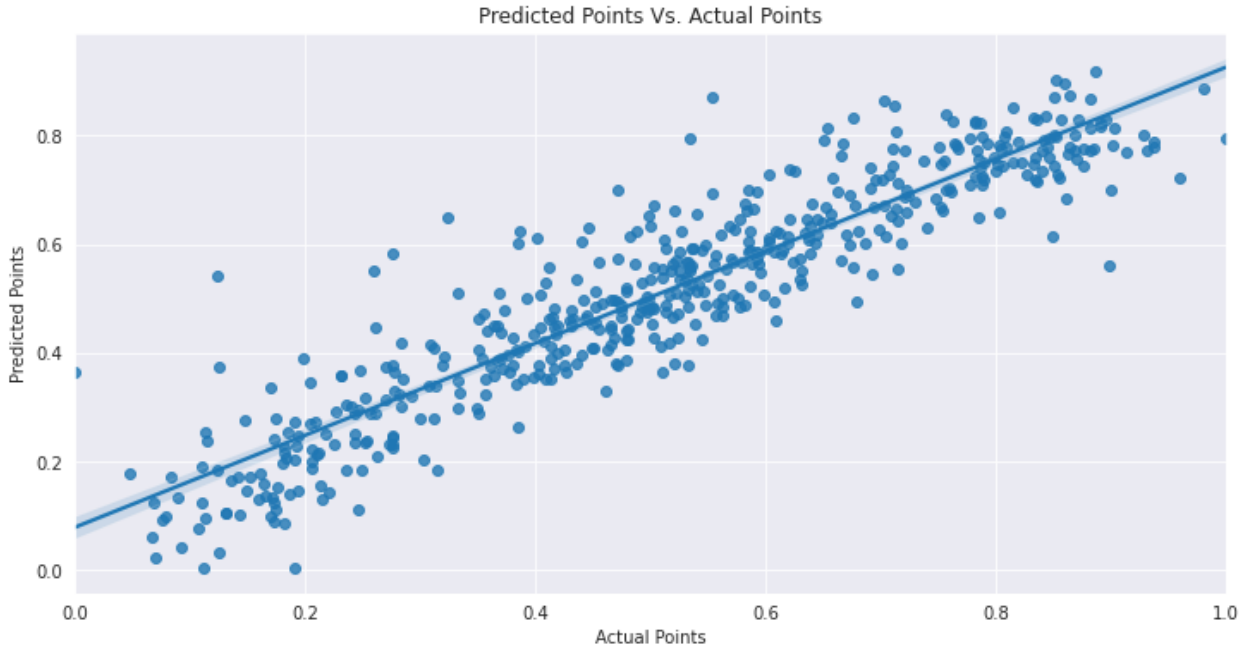
2. Why is it important to use **drop_first=True** during dummy variable creation? (2 mark)
   While creating dummy variables we use drop_first=True as we can have a based or reference category instead of creating a new category. The reason for this also is to avoid the multi-collinearity getting added into the model if all dummy variables are included. The reference category can be deduced from a row where all the dummy variables have zero values.

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 target variable i.e. 0.63.
   - atemp, casual and registered have been dropped as they are represented by other variables in the dataset in the EDA Steps.

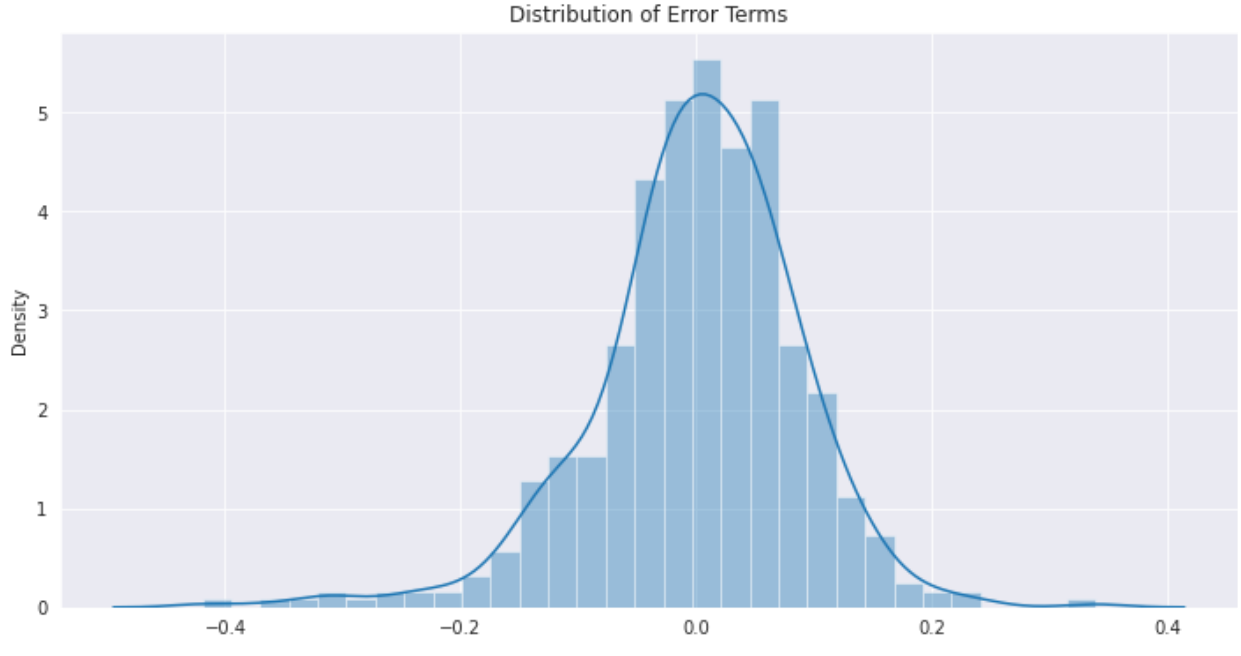


4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)
   1. **Linearity:** Linear relationship exists between X and y variables



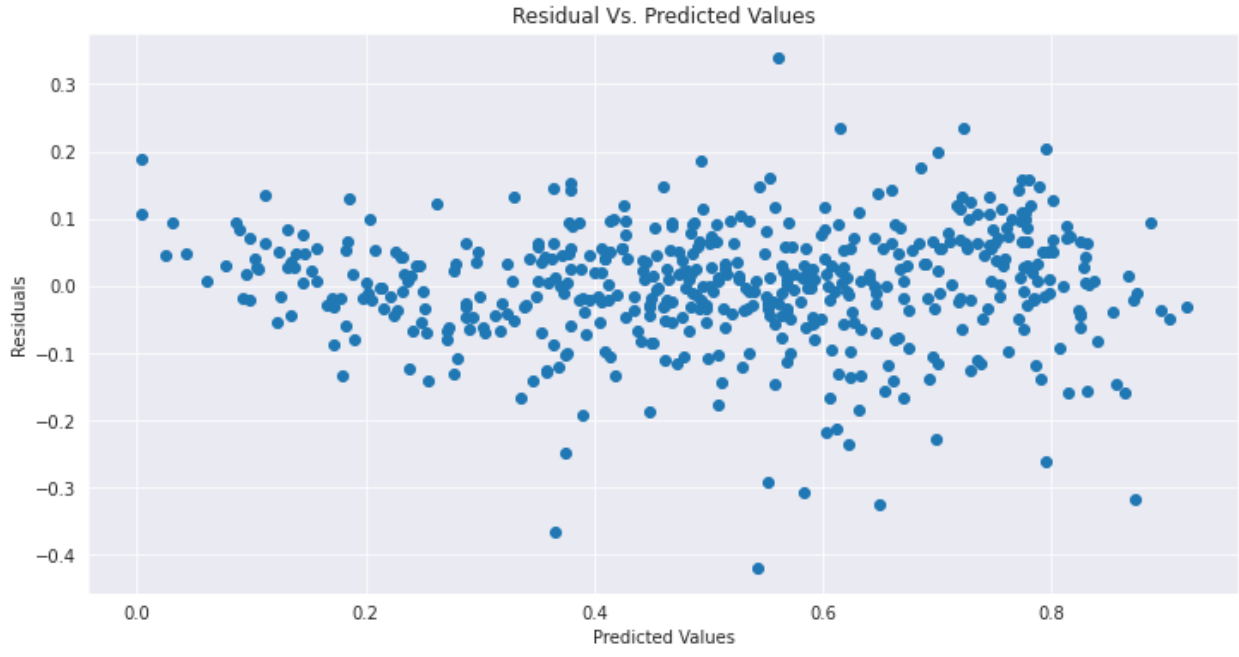
The symmetric distribution of points around the diagonal line confirms the linearity of the model.

2. Error terms are **normally distributed** (not the X and y variables)



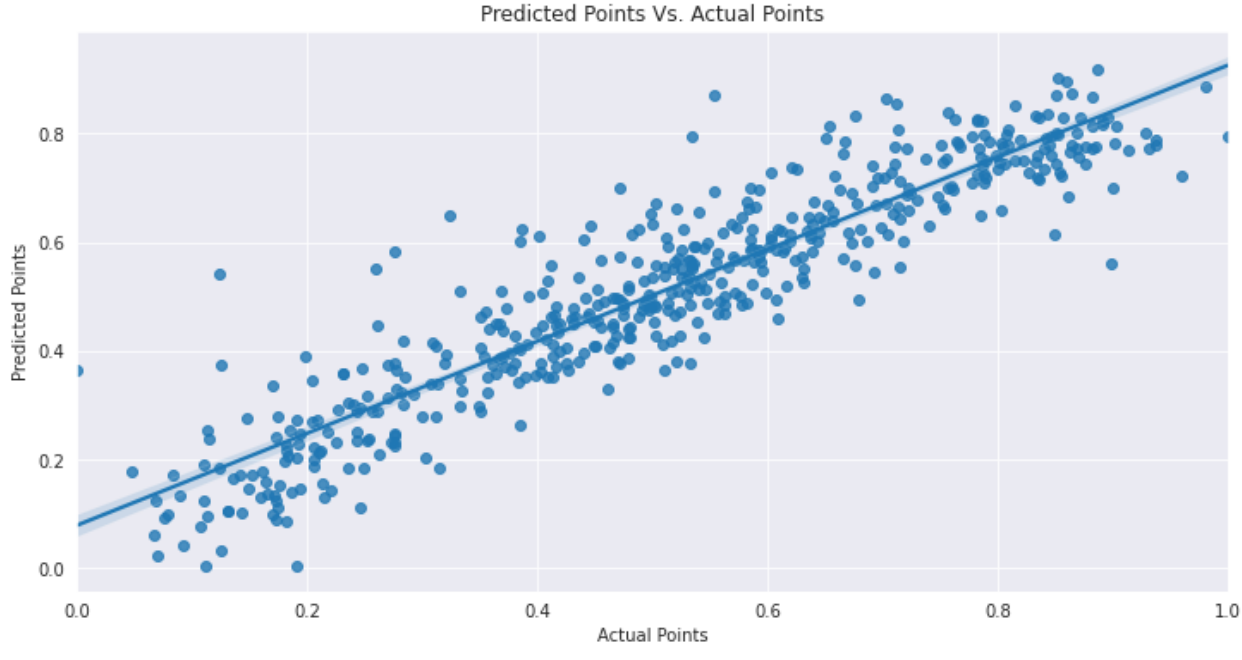
We did a distribution plot of Error terms to check if they follow normal distribution with mean zero.

3. Error terms are **independent of each other**



No specific patterns can be seen in error terms wrt to predicted values, hence we can say the error terms are independent of each other.

4. Error terms have constant variance ie. **Homoscedasticity**



Error Terms have approximated constant variable, hence we meet the criteria of homoscedasticity.

5. No multi-collinearity

We looked into p-value and variance inflation factor(VIF) to reduce the multicollinearity between X variables and tried reducing it as possible. We used recursive feature elimination to achieve this.

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

Top predictor variables:

temperature(temp): coefficient of 0.442, indicates a unit increase in temperature variable will increase the bike hire number by 0.442.

year: coefficient of 0.234, indicates a unit increase in year will increase the bike hire number by 0.234.

weather Situation(weathersit): coefficient of -0.1932, indicates if the weather situation gets bad by one level up then bike hire number goes down by -0.193.

General Subjective Questions

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

   Linear Regression is a machine learning supervised learning algorithm.

   Linear regression algorithms are used to find relationship between target and one or more predictors

   Linear regression shows correlation(relationship only) not causation. Using this we can interpolate data.

   The algorithm uses the best fitting line to map the association between target and predictor variables.

   There are 2 types of linear regression algorithms

   - Simple Linear Regression – Single independent variable is used.
   - 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.
     - $\beta_0 = value\ of\ the\ Y\ when\ X = 0\ (Y\ intercept)$
     - $\beta_1, \beta_2, \ldots, \beta_p = Slope\ or\ the\ gradient.$

   Use Cases of Linear Regression:

   a. Prediction of trends and Sales targets.
   b. Price Prediction- predict the change in price of stock or product.
   c. Risk Management- analysis of Risk Management in the financial and insurance sector.

2. Explain the Anscombe's quartet in detail. (3 marks)

Anscombe's quartet : These are four datasets that have nearly identical simple statistical description, yet have very different distributions and appear very different when graphed.

Anscombe's quartet tells us about the importance of visualizing the data before applying various algorithms out there to build models out of them which suggests that 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 like outliers, diversity of the data, linear separability of the data, etc. Also, the Linear Regression can only be considered a fit for the data with linear relationships and is incapable of handling any other kind of datasets.

The four datasets can be defined as:

1. One which fits the linear regression model well.
2. One could not fit linear regression model on the data quite well as the data is non-linear.
3. One which shows the outliers involved in the dataset which can be handled by linear regression model.
4. One which shows the outliers involved in the dataset which cannot be handled by linear regression model.

Key points from **Anscombe's quartet**:

1. Plotting the data is very important and a good practice before analyzing the data.
2. Outliers should be removed while analyzing the data.
3. Descriptive statistics do not fully depict the data set in its entirety.

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

The Pearson's R (aka Pearson's correlation coefficients) measures the strength between the different variables and the relation with each other.

The Pearon's R returns values between -1 and 1. The interpretation of the coefficients are:

- -1 coefficient indicates strong inversely proportional relationship.
- 0 coefficient indicates no relationship.
- 1 coefficient indicates strong proportional relationship.

$r = n(\Sigma x * y) - (\Sigma x) * (\Sigma y) \sqrt{[n\Sigma x^2 - (\Sigma x)^2] * [n\Sigma y^2 - (\Sigma y)^2]}$

Where:

N = the number of pairs of scores
Σxy = the sum of the products of paired scores
Σx = the sum of x scores
Σy = the sum of y scores
Σx2 = the sum of squared x scores
Σy2 = the sum of squared y scores

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

The scaling is done data preparation step for ML model. The scaling normalizes the varied scaled datatypes to a particular data range.

Two basic types of scaling:

1. Standardization:

   In this the features will be rescaled so that they'll have the properties of a standard normal distribution with

   μ=0 and σ=1

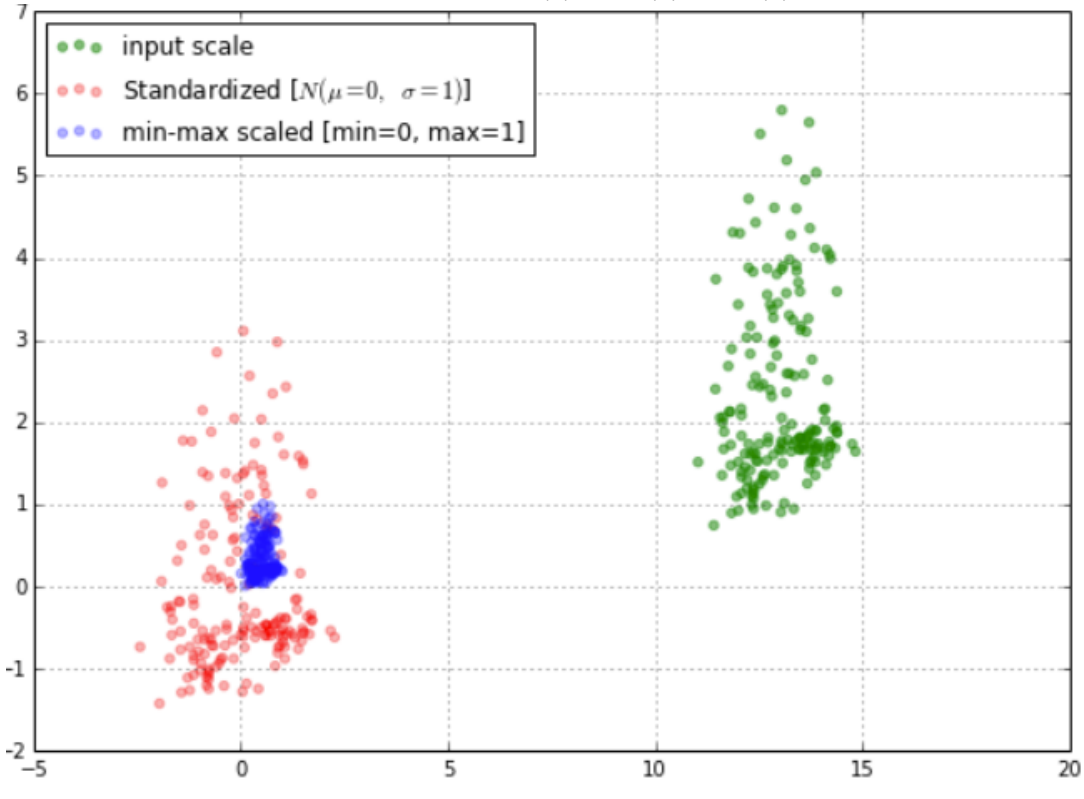
   where μ is the mean (average) and σ is the standard deviation from the mean.

   $$Standardization: x = x - mean(x)/sd(x)$$

2. Normalization:

   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.

   $$MinMaxScaling: x = x - \min(x)/\max(x) - \min(x)$$



In the above plot input data set in green is scaled using Standardization technique in red and Normalization technique in Blue.

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

From formula we can say, if the R2 is 1 then the VIF is infinite. The reason for R2 to be 1 is that there is a perfect correlation between 2 independent variables, i.e the independent variables are orthogonal to each other.

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

(3 marks)

Q-Q plots are used to find the type of distribution for a random variable whether it be a Gaussian Distribution, Uniform Distribution, Exponential Distribution or even Pareto Distribution, etc.

Q-Q Plots (Quantile-Quantile plots) are plots of two quantiles against each other. A quantile is a fraction where certain values fall below that quantile. For example, the median is a quantile where 50% of the data fall below that point and 50% lie above it. The purpose of Q-Q plots is to find out if two sets of data come from the same distribution. A 45 degree angle is plotted on the Q-Q plot; if the two data sets come from a common distribution, the points will fall on that reference line.