A PROJECT REPORT

ON

PREDICTING MILEAGE ON MOTAR TREND CARS ROAD TEST

Submitted to
Osmania University
In partial fulfillment of the requirements for the award of

MASTER OF SCIENCE
IN
STATISTICS

DEPARTMENT OF STATISTICS
UNIVERSITY COLLEGE OF SCIENCE
OSMANIA UNIVERSITY, HYDERABAD – INDIA

By

N.ABHISARIKA	Roll No: 1007-17-507-040
M.SANDHYA RANI	Roll No: 1007-17-507-043
V.SHIRISHA	Roll No: 1007-17-507-035
V.SAI SUMANA	Roll No: 1007-17-507-038
K.VINEETH KUMAR	Roll No: 1007-17-507-037
K.NAVEEN KUMAR	Roll No: 1007-17-507-039
FARAH ABUDULLA MOHAMMED	Roll No: 1007-17-507-045

Under the Supervision of

Dr. M. VENUGOPALA RAO

2018

A PROJECT REPORT

ON

PREDICTING MILEAGE ON MOTOR TREND CARS ROAD TEST

Submitted to
Osmania University
In partial fulfillment of the requirements for the award of

MASTER OF SCIENCE
IN
STATISTICS

DEPARTMENT OF STATISTICS
UNIVERSITY COLLEGE OF SCIENCE
OSMANIA UNIVERSITY, HYDERABAD – INDIA
By

N.ABHISARIKA	Roll No: 1007-17-507-040
M.SANDHYA RANI	Roll No: 1007-17-507-043
V.SHIRISHA	Roll No: 1007-17-507-035
V.SAI SUMANA	Roll No: 1007-17-507-038
K.VINEETH KUMAR	Roll No: 1007-17-507-037
K.NAVEEN KUMAR	Roll No: 1007-17-507-039
FARAH ABDULLAH MOHAMMED	Roll No: 1007-17-507-045

Under the Supervision of
Dr. M. VENUGOPALA RAO

2018

<u>CERTIFICATE</u>

This is to certify that

N.ABHISARIKA	1007-17-507-040
M.SANDHYARANI	1007-17-507-043
V.SHIRISHA	1007-17-507-035
V.SAI SUMANA	1007-17-507-037
K.VINEETKUMAR	1007-17-507-038
K.NAVEEN KUMAR	1007-17-507-039
FARAHABDULLA	
MOHAMMED	100-717-507-045

have submitted the project titled **“Motor Trend Cars Road Test”** in partial fulfilment for the degree of Master of Science in Statistics.

Head

Department of Statistics

Internal Examiner External Examiner

DECLARATION

The research presented in this project has been carried out in the **Department of Statistics, Osmania University, Hyderabad.** The work is original has not been submitted so far, in part or full, for any other degree of diploma of any university.

N.ABHISARIKA

M.SANDHYA RANI

V.SHIRISHA

V.SAI SUMANA

K.VINEETH KUMAR

K.NAVEEN KUMAR

FARAH ABDULLAH MOHAMMED

Department of Statistics
Osmania University
Hyderabad – 500007, T.S.
INDIA.

ACKNOWLEDGEMENT

I deem it a great pleasure to express my deep sense of gratitude and indebtedness to my research supervisor **Dr. M. VENUGOPALA RAO**, Statistics department, University College of Science, Osmania University for his valuable guidance, and enlightening discussions throughout the progress of my project work.

I also express my sincere and heartfelt thanks to **Pro. C. JAYALAKSHMI,** Head of Department, Department of Statistics, Osmania University for providing the necessary support and facilities in the department for completion of this work successfully.

It is indeed with great pleasure I record my thanks to **Dr. G. JAYASREE**, Chairperson, Board of Studies, Department of Statistics, Osmania University for having provided with all the facilities to carry out our work.

I thank **Dr. N. Ch. BHATRACHARYULU, Dr. K.VANI, Dr. S. A. JYOTHI RANI**, **Dr. G.SIRISHA, Mrs. J. L. PADMA SHREE** for their encouragement and constant help during the research.

I would like to express my deepest gratitude to **T.SANDHYA, BALA KARTHIK** for their advice, guidance and involvement at various stages of this work, I would also like to thank them for their understanding and constant encouragement throughout this project.

I thank all Non-Teaching members of the Department of Statistics, who helped me during my Thesis work.

I am thankful to the Osmania University for permitting me to carry out this work.

CONTENTS

CHAPTER-1

INTRODUCTION

1. Introduction

1.1 Scope Of The Problem:

- Objective: The objective is to predict the mpg (**mileage**) for the given attributes.

1.2 Description:

The data was extracted from the 1974 *Motor Trend* US magazine, and comprises fuel consumption and 10 aspects of automobile design and performance for 32 automobiles (1973–74 models).

1.2.1 Source:

Henderson and **Velleman** (1981), Building multiple regression models interactively. *Biometrics*, 37, 391–411.

1.2.2 Attribute Information:

- A data frame with 32 observations on 11 (numeric) variables.

[1]	mpg	Miles/(US) gallon
[2]	Cyl	Number of cylinders
[3]	disp	Displacement (cu.in.)
[4]	Hp	Gross horsepower
[5]	drat	Rear axle ratio
[6]	Wt	Weight (1000 lbs)
[7]	qsec	1/4 mile time
[8]	Vs	Engine (0 = V-shaped, 1 = straight)
[9]	Am	Transmission (0 = automatic, 1 = manual)
[10]	gear	Number of forward gears
[11]	carb	Number of carburettors

Data set characteristics	:	Multivariate
Number of instances	:	32
Area	:	Computer
Attributes characteristics	:	Integer
Number of attributes	:	11

1.3 Review of the Chapters:

Chapter-2 gives the brief introduction about machine learning techniques like need of ML today, types of ML Algorithms and various models in each algorithm and what technique to use when and how to validate, Tune the ML algorithms and how to measure the performance of the ML model.

Chapter 3 describes the various results obtained for the problem. This section contains all the outputs generated through the ML algorithms applied on the data as well as validation and performance matrices.

Chapter-4 describes the summary and conclusions followed by bibliography.

Chapter-2

Review Of Machine Learning Process

2. Review of machine learning process:

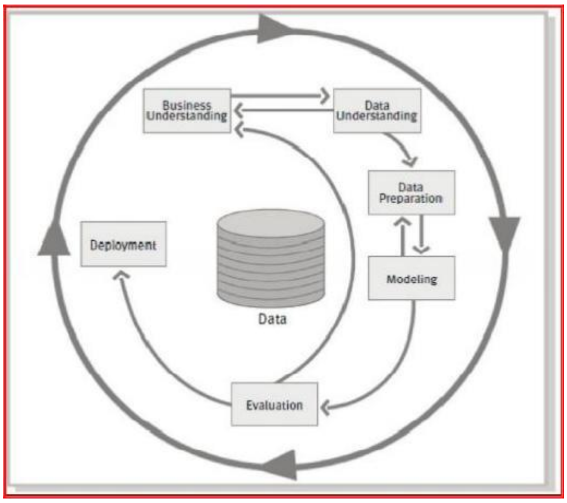
2.0 Need of Machine Learning:

In this age of modern technology, there is one resource that we have in abundance: a large amount of structured and unstructured data. In the second half of the twentieth century, machine learning evolved as a subfield of artificial intelligence that involved the development of self-learning algorithms to gain knowledge from that data in order to make predictions. Instead of requiring humans to manually derive rules and build models from analysing large amounts of data, machine learning offers a more efficient alternative for capturing the knowledge in data to gradually improve the performance of predictive models, and make data-driven decisions. Not only is machine learning becoming increasingly important in computer science research but it also plays an ever greater role in our everyday life.

2.1 Machine Learning Process:

The CRISP-DM (Cross-Industry Standard Process for Data Mining) Process was designed specifically for the data mining. However, it is flexible and thorough enough that it can be applied to any analytical project whether it is predictive analytics, data science, or Machine learning. The Process has the following six phase

- Business Understanding
- Data Understanding
- Data preparation
- Modelling
- Evaluation
- Deployment



(fig 2.1 CRISP-DM Process)

And, each phase has different steps covering important tasks which are mentioned below:

2.1.1) Business Understanding:

It is very important step of the process in achieving the success. The purpose of this step is to identify the requirements of the business so that you can translate them into analytical objectives. It has the following tasks:

1) Identify the Business objective
2) Assess the situation
3) Determine the Analytical goals
4) Produce a project plan

2.1.2) Data Understanding:

After enduring the all-important pain of the first step, you can now get your hands on the data. The task in this process consist the following.

1) Collect the data
2) Describe the data
3) Explore the data
4) Verify the data Quality

2.1.3) Data Preparation:

This step is relatively self-explanatory and in this step the goal is to get the data ready to input in the algorithms. This includes merging, feature engineering, and transformations. If imputation for missing values / outliers is needed then, it happens in this step. The key five tasks under this step are as follows:

1) Select the data
2) Clean the data
3) Construct the data
4) Integrate the data
5) Format the data

2.1.4) Modelling:

Oddly, this process step includes the consideration that you already thought of and prepared for. In this, one will need at least a modicum of an idea about how they will be modelling. Remember, that this is flexible, iterative process and some strict linear flow chart such as an aircrew checklist. Below are the tasks in this step:

1) Select a modelling technique
2) Generate a test design
3) Build a model
4) Assess a Model

Both cross validation of the model (using tran/test or K fold validation) and model assessment which involves comparing the models with the chosen criterion (RMSE, Accuracy, ROC) will be performed under this phase.

2.1.5) Evaluation:

In the evaluation process, the main goal is to confirm that the work that has been done and the model selected at this point meets the business objective. Ask yourself and others, have we achieved the definition of success? And, here are the tasks in this step:

1) Evaluate the results
2) Review the process
3) Determine the next steps.

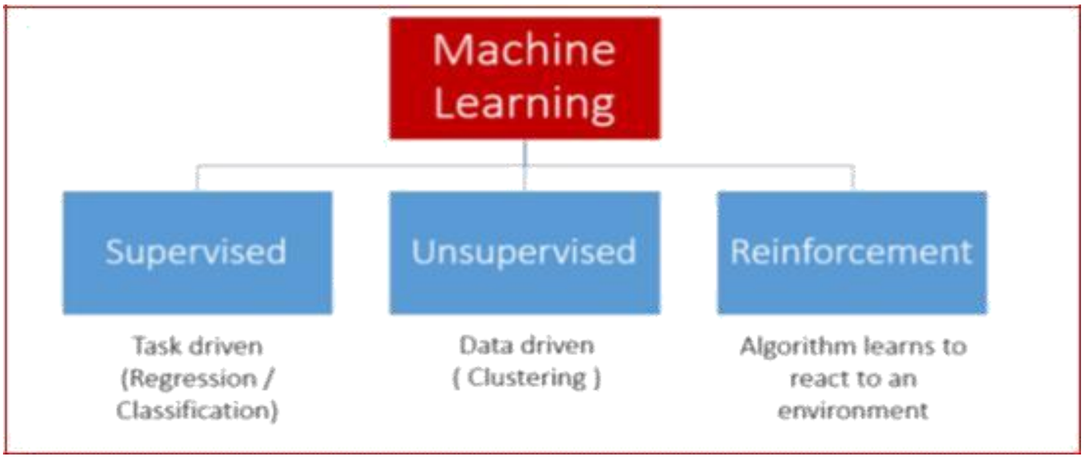
2.1.6) Deployment:

If everything is done according to the plan up to this point, it might come down to flipping a switch and your model goes live. Here are the tasks in this step:

1) Deploying the plan
2) Monitoring and maintenance of the plan
3) Producing the final report

2.2 Types of Machine Learning:

Broadly, the Machine Learning Algorithms are classified into 3 types:


Machine Learning
Supervised
Unsupervised
Reinforcement
Task driven (Regression / Classification)
Data driven (Clustering)
Algorithm learns to react to an environment

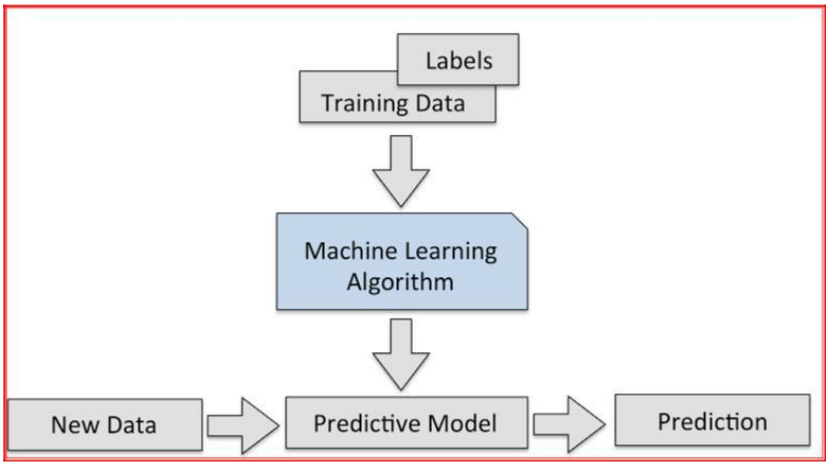
(fig. 2.2 Types of machine learning)

2.2.1) Supervised Learning:

This algorithm consists of a target / outcome / dependent variable which is to be predicted from a given set of predictors / independent variables. Using these set of variables,

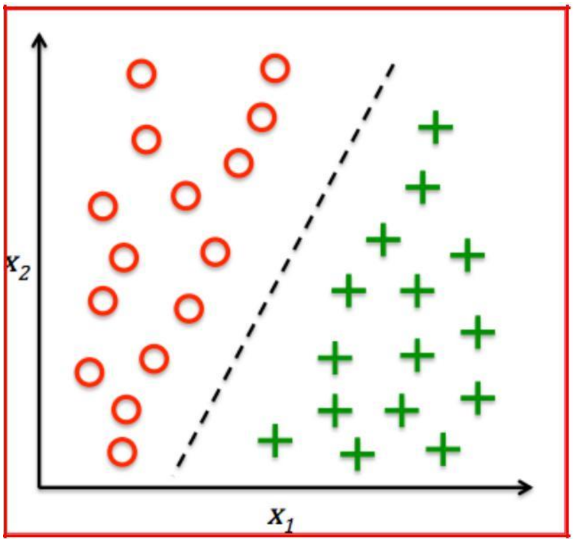
We generate a function that maps inputs to desired output. The training process continues until the model achieves a desired level of accuracy on the training data.

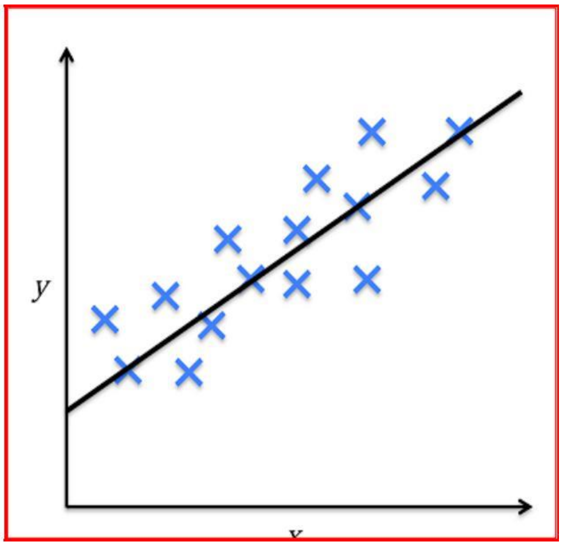
The process of Supervised Learning model is illustrated in the below picture :



(fig 2.2.1 supervised learning)

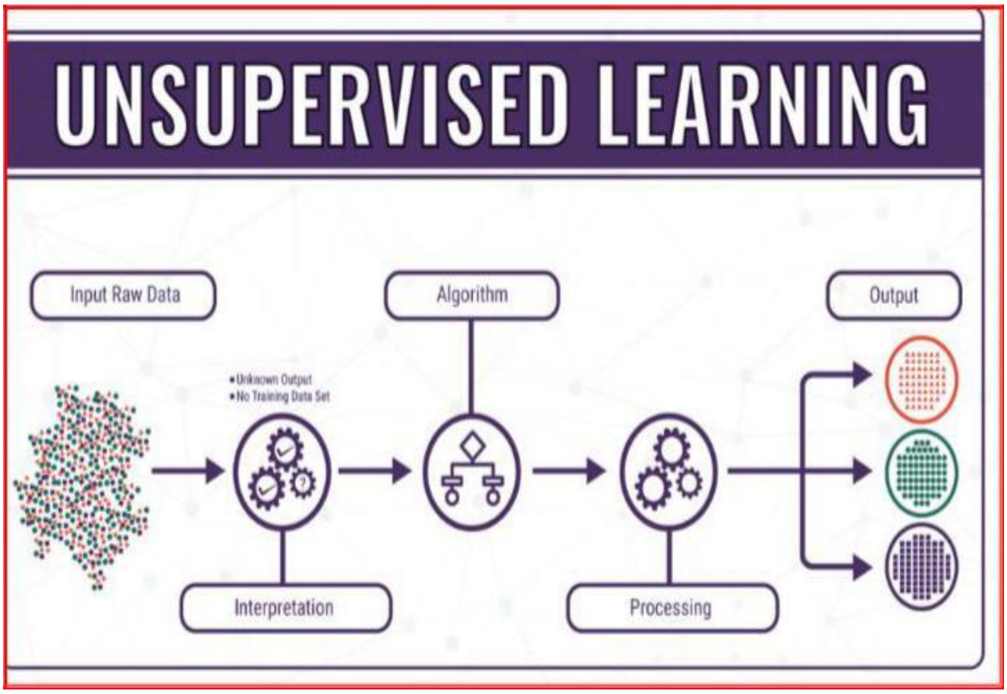
Examples of Supervised Learning: Regression, Decision Tree, Random Forest, KNN, Logistic Regression, etc





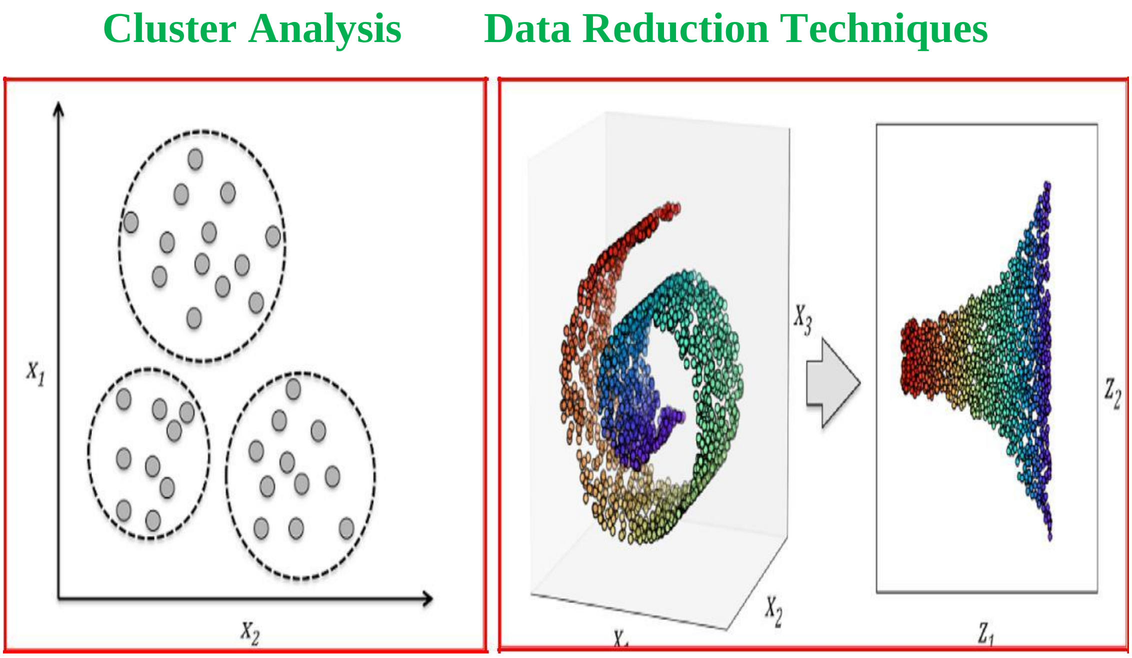
2.2.2) Unsupervised Learning:

In this algorithm, we will not have any target or outcome variable to predict / estimate. It is used for clustering population into different groups, which is widely used for segmenting customers in different groups for specific intervention. (More of Exploratory Analysis)



(fig 2.2.2 unsupervised learning)

Examples of Unsupervised Learning: Data reduction techniques, Cluster Analysis, Market Basket Analysis, etc...



Cluster Analysis

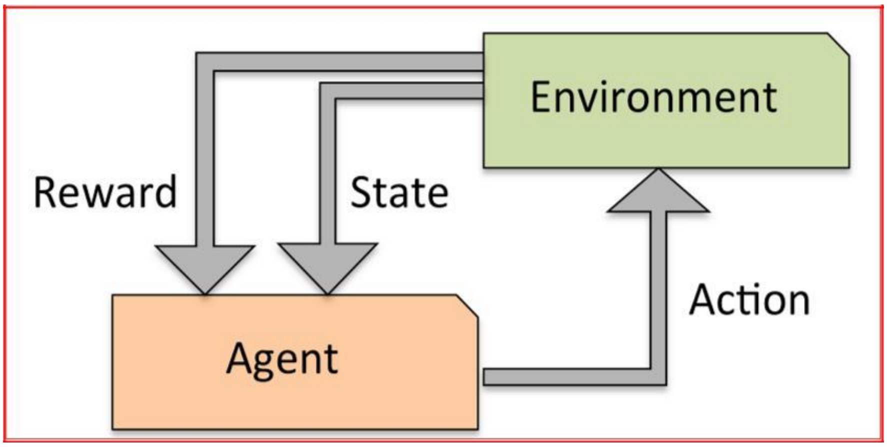
Data Reduction Techniques

(fig 2.2.2 Unsupervised learning techniques)

2.2.3) Reinforcement Learning:

Using this algorithm, the machine is trained to make specific decisions. It works this way: the machine is exposed to an environment where it trains itself continually using trial and error. This machine learns from past experience and tries to capture the best possible knowledge to make accurate business decisions.

The process of reinforcement learning is illustrated in the below picture:



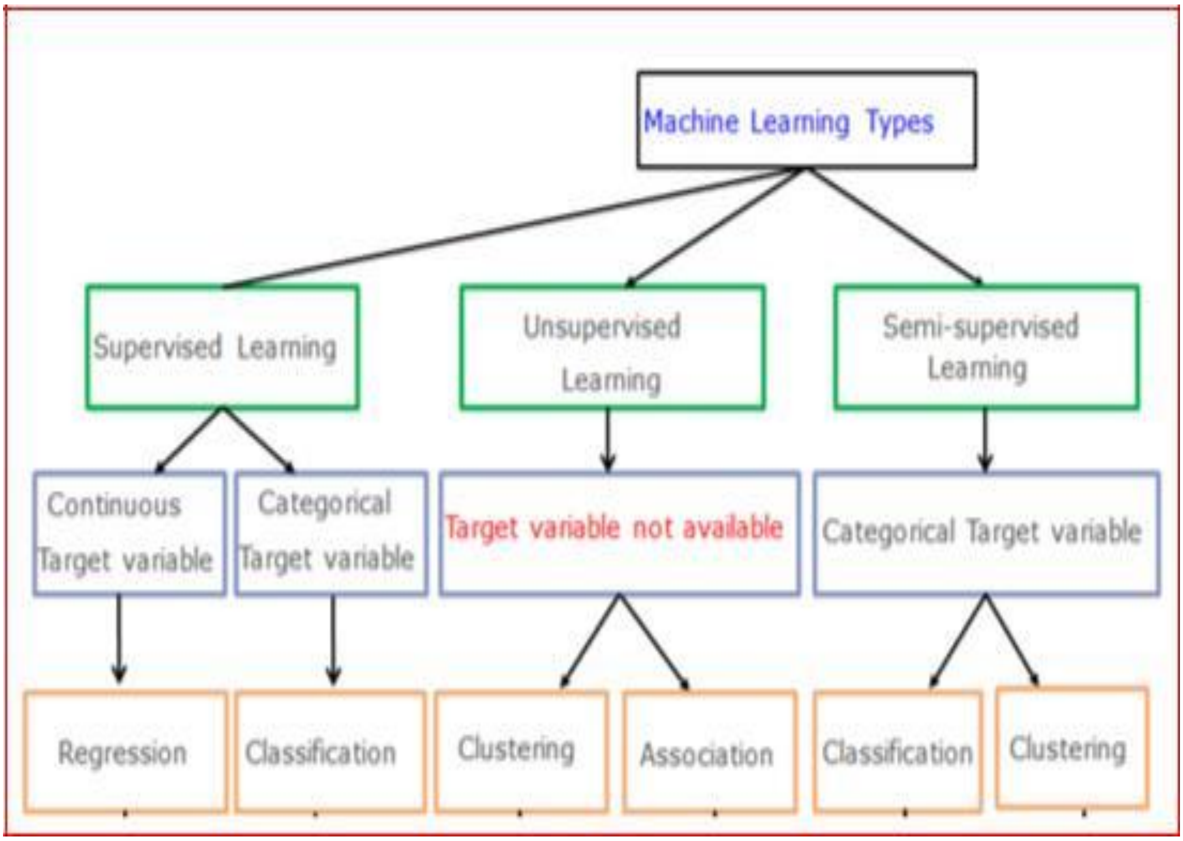
(fig 2.2.3 Reinforcement learning)

Examples of Reinforcement Learning: Markov Decision Process, Self-driving cars, etc

2.3 Choosing the algorithm:

Choosing the right algorithm will depend on the type of the problem we are solving also depends on the scale of the dependent variable. In case of continuous target variable, we will use regression algorithms and in case of categorical target, we will use classification algorithms and for the model which doesn’t have target variable, we will use either cluster analysis / data reduction techniques.

Below picture describes the process of choosing the right algorithm



(fig 2.3 process of choosing the algorithm)

2.3.1) Types of Regression Algorithms:

There are many Regression algorithms in machine learning, which will be used in different regression applications. Some of the main regression algorithms are as follows:

a) **Simple Linear Regression:-** In simple linear regression, we predict scores on one variable from the data of second variable. The variable we are forecasting is called the criterion variable and referred to as Y. The variable we are basing our predictions on is called the predictor variable and denoted as X.

b) **Multiple Linear Regression:** Multiple linear regression is one of the algorithms of regression techniques, and is the most common form of linear regression analysis. As a predictive analysis, the multiple linear regressions are used to explain the relationship between one dependent variable with two or more independent variables. The independent variables can be either continuous or categorical.

c) **Polynomial Regression:-** Polynomial regression is another form of regression in which the maximum power of the independent variable is more than 1. In this regression technique, the best fit line is not a straight line instead it is in the form of a curve.

d) **Support Vector Machines:-** Support Vector Machines can be applied to regression problems as well as Classification. It contains all the features that characterises maximum margin algorithm. Linear learning machine maps a non-linear function into high dimensional kernel-induced feature space. The system capacity will be controlled by parameters that do not depend on the dimensionality of feature space.

e) **Decision Tree Regression:-** Decision tree builds regression models in the form of a tree structure. It breaks down the data into smaller subsets and while at the same time an associated decision tree is incrementally developed. The final result is a tree with decision nodes and leaf nodes.

f) **Random Forest Regression:-** Random Forest is also one of the algorithms used in regression technique. It is very a flexible, easy to use machine learning algorithm that produces, even without hyper - parameter tuning, a great result most of the time. It is also one of the most widely used algorithms because of its simplicity and the fact that can used for both regression and classification tasks. The forest it builds is an ensemble of Decision Trees, most of the time trained with the "bagging" method.

Other than these we have regularized regression models like **Ridge, LASSO** and **Elastic Net regression** which are used to select the key parameters and these is also **Bayesian regression** which works with the Bayes theorem.

2.3.2) Types of Classification Algorithms:

There are many Classification algorithms in machine Learning, which can be used for different classification applications. Some of the main classification algorithms are as follows:

a) **Logistic Regression/Classification:-** Logistic regression falls under the category of supervised learning; it measures the relationship between the dependent variable which is categorical with one or more than one independent variables by estimating probabilities using a logistic/sigmoid function. Logistic regression can generally be used when the dependent variable is Binary or Dichotomous. It means that the dependent variable can take only two possible values like "Yes or No", "Living or dead".

b) **K -Nearest Neighbours: -** k-NN algorithm is one of the most straight forward algorithms in classification, and it is one of the most used ML algorithms. An object is classified by a majority vote of its neighbours, with the object being assigned to the class most common among its k nearest neighbours. It can also use for regression — output is the value of the object (predicts continuous values). This value is the average (or median) of the values of its k nearest neighbours.

c) **Naive Bayes :-** Naive Bayes is a type of Classification technique based on Bayes' theorem, with an assumption of independence among predictors. In simple terms, a Naive Bayes classifier assumes that the presence of a Particular feature in a class is unrelated to the presence of any other function. Naive Bayes model is accessible to build and particularly useful for extensive datasets.

d) **Decision Tree Classification:-** Decision tree builds classification models in the form of a tree structure. It breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with decision nodes and leaf nodes. A decision node has two or more branches. Leaf node represents a classification or decision. The first decision node in a tree which corresponds to the best predictor called root node. Decision trees can handle both categorical and numerical data.

e) **Support Vector Machines:-** A Support Vector Machine is a type of Classifier, in which a discriminative classifier is formally defined by a separating hyperplane. The algorithm outputs an optimal hyperplane which categorises new examples. In two dimensional space, this hyperplane is a line dividing a plane in two parts where in each class lay in either side.

f) **Random Forest Classification:-** Random Forest is a supervised learning algorithm .It creates a forest and makes it somehow random. The forest it builds is an ensemble of Decision Trees, most of the times the decision tree algorithm trained with the "bagging" method. The general idea of the bagging method is that a combination of learning models increases the overall result. And Random Forest is also very powerful to find the variable importance in classification/ Regressionproblems

2.3.3) Types of Unsupervised Learning:

Clustering is the type of unsupervised learning in which an unlabelled data is used to draw inferences. It is the process of grouping similar entities together. The goal of this unsupervised machine learning technique is to find similarities in the data points and group similar data points together and also to figure out which cluster should a new data point belong to.

❖ Types of Clustering Algorithms:-

There are many Clustering algorithms in machine learning, which can be used for different clustering applications. Some of the main clustering algorithms are as follows:

a) **Hierarchical Clustering:-** Hierarchical clustering is one of the algorithms of clustering technique, in which similar data is grouped in a cluster. It is an algorithm that builds the hierarchy of clusters. This algorithm starts with all the data points assigned to a bunch of their own. Then, two nearest groups are merged into the same cluster. In the end, this algorithm terminates when there is only a single cluster left. It starts by assigning each data point to its bunch. Finds the closest pair using Euclidean distance and merges them into one cluster. This process is continued until all data points are clustered into a single cluster.

b) **K -Means Clustering:-** K-Means clustering is one of the algorithms of clustering technique, in which similar data is grouped into a cluster. K-means is an iterative algorithm that aims to find local maxima in each iteration. It starts with K as the input which is the desired number of clusters. Input k centroid in random locations in your space. Now, with the use of the Euclidean distance method, calculates the distance between data points and centroids, and assign data point to the cluster which is close to its centroid. Re calculate the cluster centroids as a mean of data points attached to it. Repeat until no further changes occur.

❖ **Types of Dimensionality Reduction Algorithms:-** There are many dimensionality reduction algorithms in machine learning, which are applied for different dimensionality reduction applications. One of the main dimensionality reduction techniques is Principal Component Analysis (PCA) / Factor Analysis.

Principal Component Analysis (Factor Analysis):-

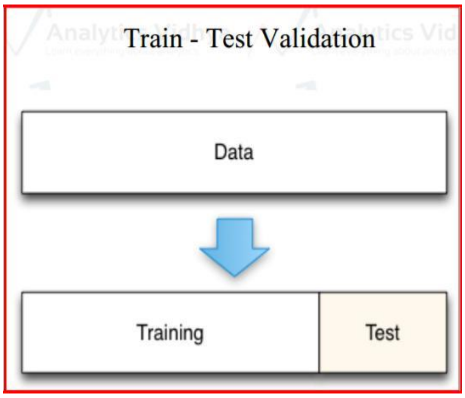
Principal Component Analysis one of the algorithms of Dimensionality reduction. In this technique, it transforms data into a new set of variables from input variables, which are the linear combination of real variables. These Specific new set of variables are known as principal components. As a result of the transformation, the first primary component will have the most significant possible variance, and each following component in has the highest possible variance under the constraint that it is orthogonal to the above components. Keeping only the best m < n components, reduces the data dimensionality while retaining most of the data information.

2.4 Choosing and comparing models through Pipelines:

When you work on machine learning project, you often end up with multiple good models to choose from. Each model will have different performance characteristics. Using resampling methods like k-fold cross validation; you can get an estimate of how accurate each model may be on unseen data. You need to be able to use these estimates to choose one or two best models from the suite of models that you have created.

2.4.1) Model Validation:

When you are building a predictive model, you need to evaluate the capability or generalization power of the model on unseen data. This is typically done by estimating accuracy using data that was not used to train the model, often referred as cross validation.



(fig 2.4.1 Model validation)

A few common methods used for Cross Validation

1) The Validation set Approach (Holdout Cross validation):

In this approach, we reserve large portion of dataset for training and rest remaining portion of the data for model validation. Ideally people will use 70-30 or 80-20 percentages for training and validation purpose respectively.

A major disadvantage of this approach is that, since we are training a model on a randomly chosen portion of the dataset, there is a huge possibility that we might miss-out on some interesting information about the data which, will lead to a higher bias.

2) K-fold cross validation:

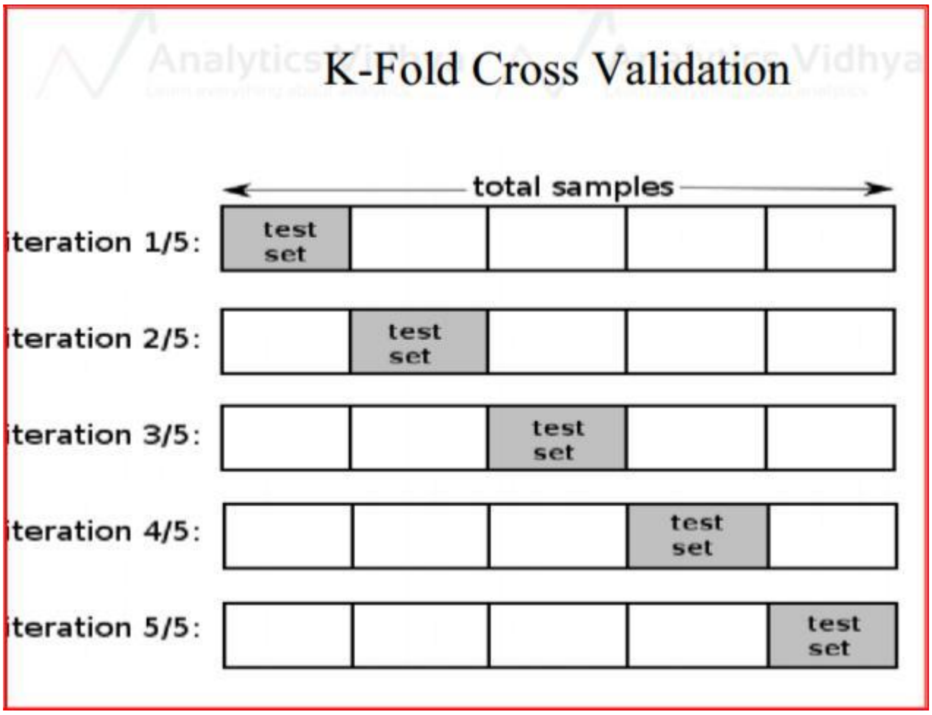
As there is never enough data to train your model, removing a part of it for validation may lead to a problem of under fitting. By reducing the training data, we risk losing important patterns/ trends in data set, which in turn increases error induced by bias. So, what we require is a method that provides ample data for training the model and also leaves ample data for validation. K Fold cross validation does exactly that.

In K Fold cross validation, the data is divided into k subsets. Now the holdout method is repeated k times, such that each time, one of the k subsets is used as the test set/ validation set and the other k-1 subsets are put together to form a training set. The error estimation is averaged over all k trials to get total effectiveness of our model. As can be seen, every data point gets to be in a validation set exactly once, and gets to be in a training set(k-1) times. This significantly reduces the bias as we are using most of the data for fitting, and also significantly reduces variance as most of the data is also being used in validation.

Interchanging the training and test sets also adds to the effectiveness of this method. As a general rule and empirical evidence, K = 5 or 10 is preferred, but nothing's fixed and it can take any value.

Below are the steps for it:

- Randomly split your entire dataset into k ”folds”
- For each k-fold in your dataset, build your model on k – 1 folds of the dataset. Then, test the model to check the effectiveness for kth fold.
- Record the error you see on each of the predictions.
- Repeat this until each of the k-folds has served as the test set.
- The average of your k recorded errors is called the cross-validation error and will serve as your performance metric for the model.
- Below is the visualization of a k-fold validation when k=5.



(fig 2.4.1 k-fold cross validation)

How to choose K:

- Smaller dataset: 10-fold cross validation is better
- Moderate dataset: 5 or 6 fold cross validation works mostly
- Big dataset: Train – Val split for validation.

Other than this, we have Leave one out cross validation (LOOCV), in which each record will be left over from the training and then, the same will be used for testing purpose. This process will be repeated across all the respondents.

2.5 Model Diagnosis with over fitting and under fitting:

2.5.1) Bias and Variance:

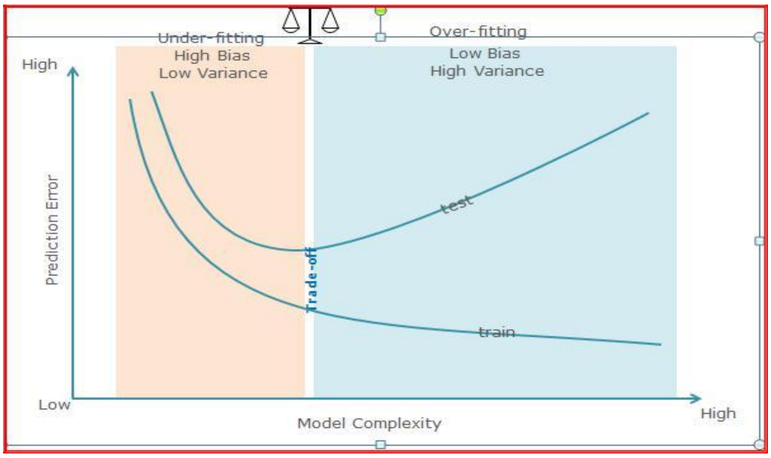
1. A fundamental problem with supervised learning is the bias variance trade-off. Ideally, a model should have two key characteristics.
2. Sensitive enough to accurately capture the key patterns in the training dataset.
3. Generalized enough to work well on any unseen dataset.

Unfortunately, while trying to achieve the above-mentioned first point, there is an ample chance of over-fitting to noisy or unrepresentative training data points leading to a failure of generalizing the model. On the other hand, trying to generalize a model may result in failing to capture important regularities.

I f model accuracy is low on a training dataset as well as test dataset, the model is said to be under-fitting or that the model has high bias. The **Bias** refers to the simplifying assumptions made by the algorithm to make the problem easier to solve. To solve an under-fitting issue or to reduce bias, try including more meaningful features and try to increase the model complexity by trying higher-order interaction.

The **Variance** refers to sensitivity of a model changes to the training data. A model is giving high accuracy on a training dataset, however on a test dataset the accuracy drops drastically then, the model is said to be over-fitting or a model that has high variance.

To solve the over-fitting issue Try to reduce the number of features, that is, keep only the meaningful features or try regularization methods that will keep all the features. Ideal model will be the trade-off between Under-fitting and over-fitting like mentioned in the below picture.



(fig 2.5.1. Bias and variance)

And, the Hyper parameters will be tuned in the below mentioned ways to reach the optimal solution:

1)Grid Search

2)Random Search

3)Manual Tuning.

2.5.2) Model Performance Matrix:

Model evaluation is an integral part of the model development. Based on model evaluation and subsequent comparisons, we can take a call whether to continue our efforts in model enhancement or cease them and select the final model that should be used / deployed.

Evaluating Classification models:

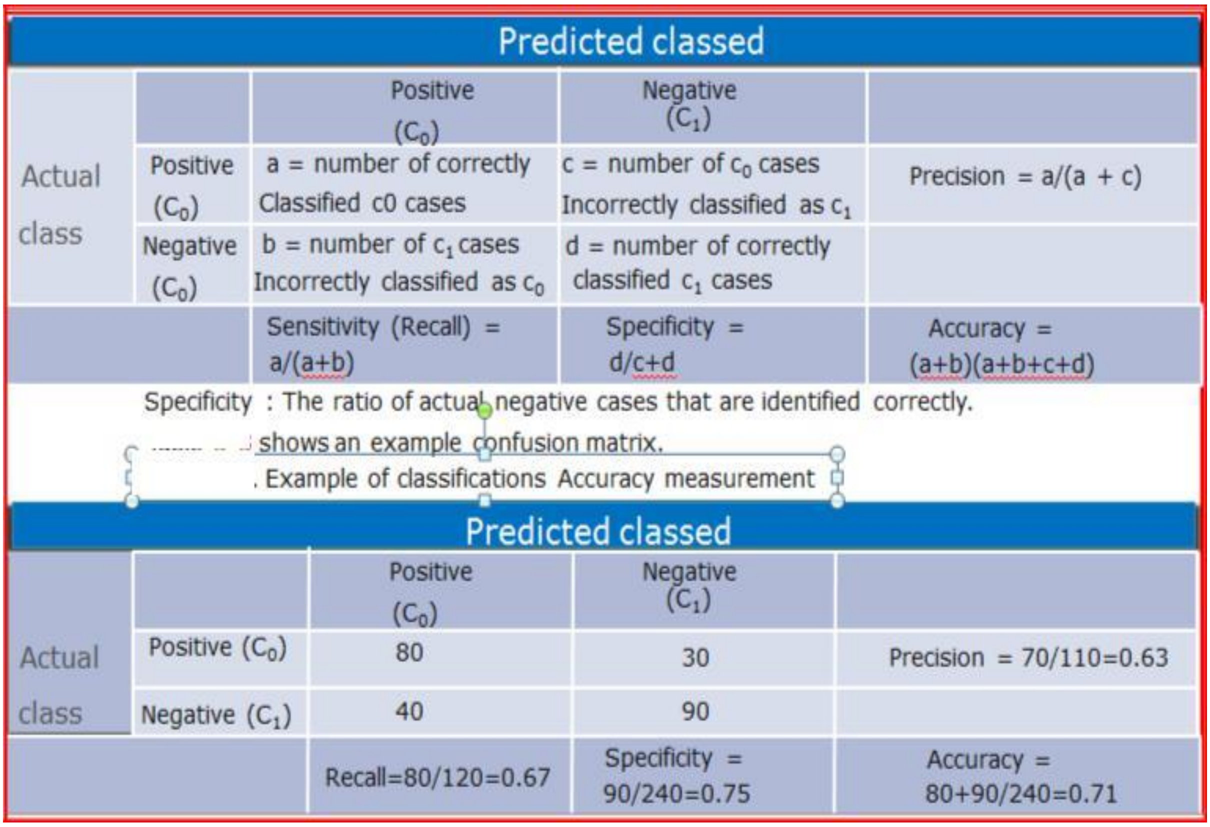
1. Confusion matrics:

Confusion matrix is one of the most popular ways to evaluate a classification model. A confusion matrix can be created for a binary classification as well as a multi-class classification model.

A confusion matrix is created by comparing the predicted class label of a data point with its actual class label. This comparison is repeated for the whole

dataset and the results of this comparison are compiled in a matrix or tabular format.

(Table 2.5.2 Confusion matrics)



		Predicted classed		
		Positive (C_0)	Negative (C_1)	
Actual class	Positive (C_0)	a = number of correctly Classified c0 cases	c = number of c_0 cases Incorrectly classified as c_1	Precision = a/(a + c)
	Negative (C_0)	b = number of c_1 cases Incorrectly classified as c_0	d = number of correctly classified c_1 cases	
		Sensitivity (Recall) = a/(a+b)	Specificity = d/c+d	Accuracy = (a+b)(a+b+c+d)

Specificity : The ratio of actual negative cases that are identified correctly.

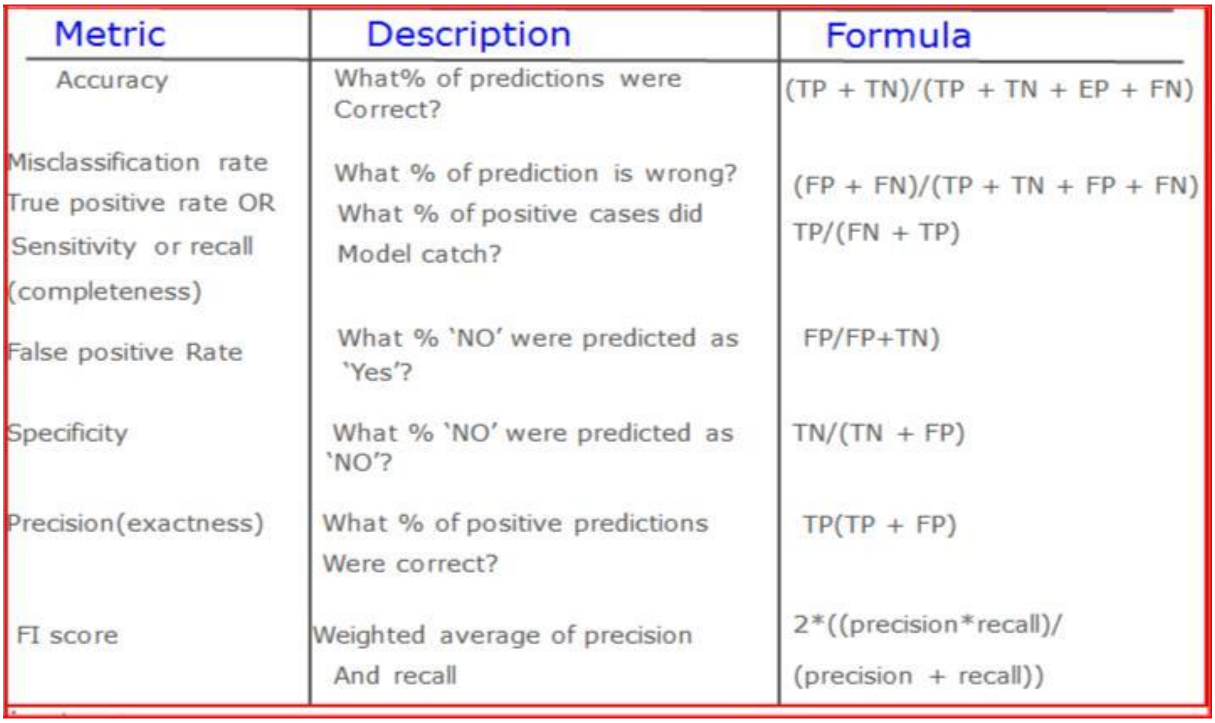
------- -- shows an example confusion matrix.

. Example of classifications Accuracy measurement

		Predicted classed		
		Positive (C_0)	Negative (C_1)	
Actual class	Positive (C_0)	80	30	Precision = 70/110=0.63
	Negative (C_1)	40	90	
		Recall=80/120=0.67	Specificity = 90/240=0.75	Accuracy = 80+90/240=0.71

And, below are the various measures that will be used to assess the performance of the model based on the requirement of the problem and as well as data.

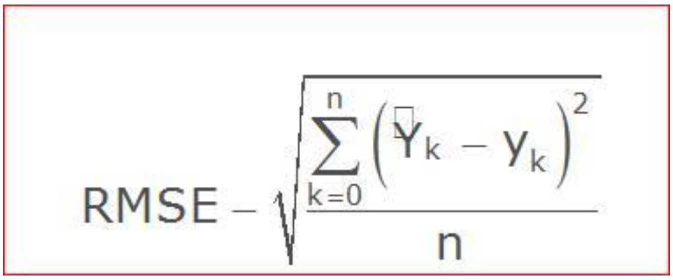
(Table 2.5.3 various measures)



Metric	Description	Formula
Accuracy	What% of predictions were Correct?	(TP + TN)/(TP + TN + EP + FN)
Misclassification rate	What % of prediction is wrong?	(FP + FN)/(TP + TN + FP + FN)
True positive rate OR Sensitivity or recall (completeness)	What % of positive cases did Model catch?	TP/(FN + TP)
False positive Rate	What % ‘NO’ were predicted as ‘Yes’?	FP/FP+TN)
Specificity	What % ‘NO’ were predicted as ‘NO’?	TN/(TN + FP)
Precision(exactness)	What % of positive predictions Were correct?	TP(TP + FP)
FI score	Weighted average of precision And recall	2*((precision*recall)/ (precision + recall))

2. Regression Model Evaluation:

A regression line predicts the y values for a given x value. Note that the values are around the average. The prediction error (called as root-mean-square error or RSME) is given by the following formula:



$$\text{RMSE} - \sqrt{\frac{\sum_{k=0}^{n}\left(\hat{Y}_k - y_k\right)^2}{n}}$$

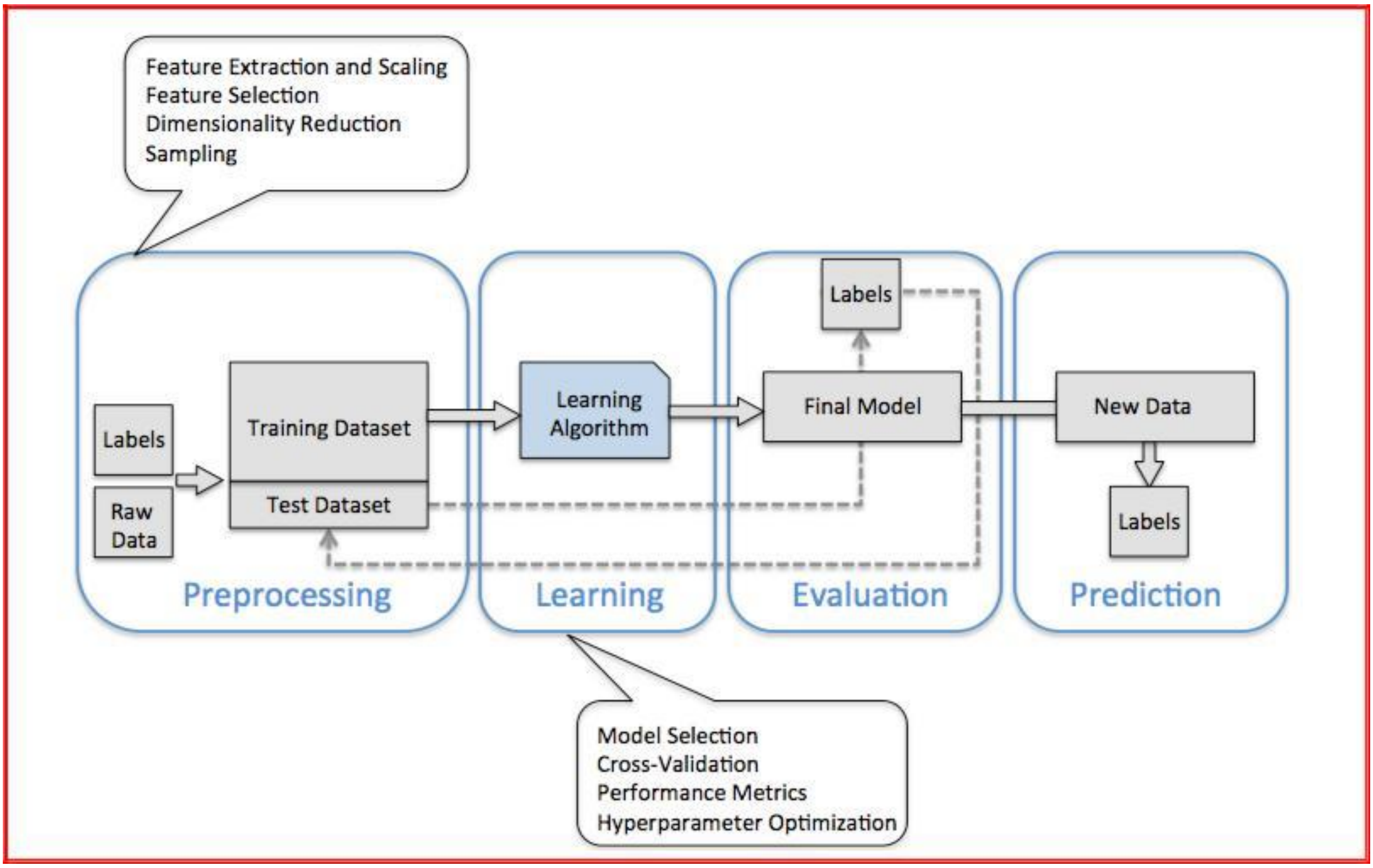
And, the regression will also assessed by R square (Co efficient of determination).

3. Evaluating Unsupervised Models:

The Unsupervised algorithms will be assessed by the profile of the factors/ clusters which were derived through the model.

2.6 Overall Process of Machine Learning:

To put overall process together, below is the picture that describes the road map for building ML Systems



(fig 2.6 Overall process of machine learning)

Chapter 3

Machine learning-At work

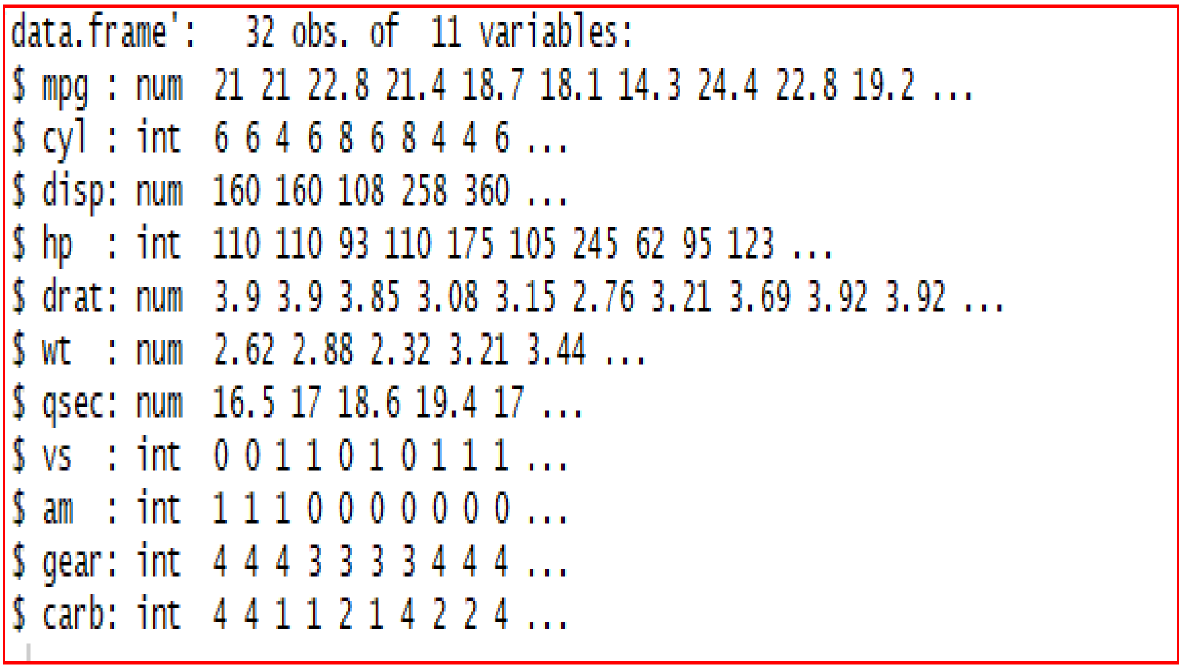
3. Machine Learning at work

3.1 An Approach to the problem:

In order to carry out the analysis, we have extracted MTCARS records from the 1974 Motor trend US magazine and the information of the same is mentioned in chapter 1.

In this chapter, we are going to discuss about the results of different Machine Learning methods used in order to obtain the solution for the problem mentioned in chapter 1.

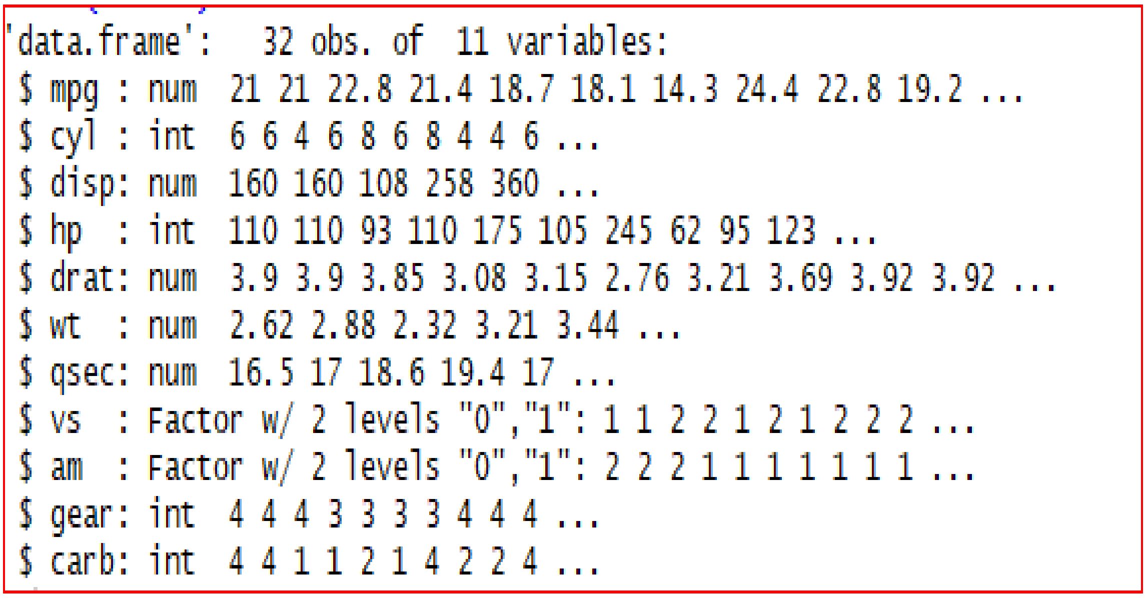
As mentioned in chapter 2, the first step of a ML Algorithm is data cleaning and preparing data for the modelling. As a first step, we have to check whether the data was read properly and all the scale types are as per the data.



```
data.frame':   32 obs. of  11 variables:
$ mpg : num  21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...
$ cyl : int  6 6 4 6 8 6 8 4 4 6 ...
$ disp: num  160 160 108 258 360 ...
$ hp  : int  110 110 93 110 175 105 245 62 95 123 ...
$ drat: num  3.9 3.9 3.85 3.08 3.15 2.76 3.21 3.69 3.92 3.92 ...
$ wt  : num  2.62 2.88 2.32 3.21 3.44 ...
$ qsec: num  16.5 17 18.6 19.4 17 ...
$ vs  : int  0 0 1 1 0 1 0 1 1 1 ...
$ am  : int  1 1 1 0 0 0 0 0 0 0 ...
$ gear: int  4 4 4 3 3 3 3 4 4 4 ...
$ carb: int  4 4 1 1 2 1 4 2 2 4 ...
```

Output:- 3.1.1Description of data.

As we have observed some of the scales have not alligned properly. we need to convert the scale types that are needed to chance the data type as factors for both VS and AM variables.

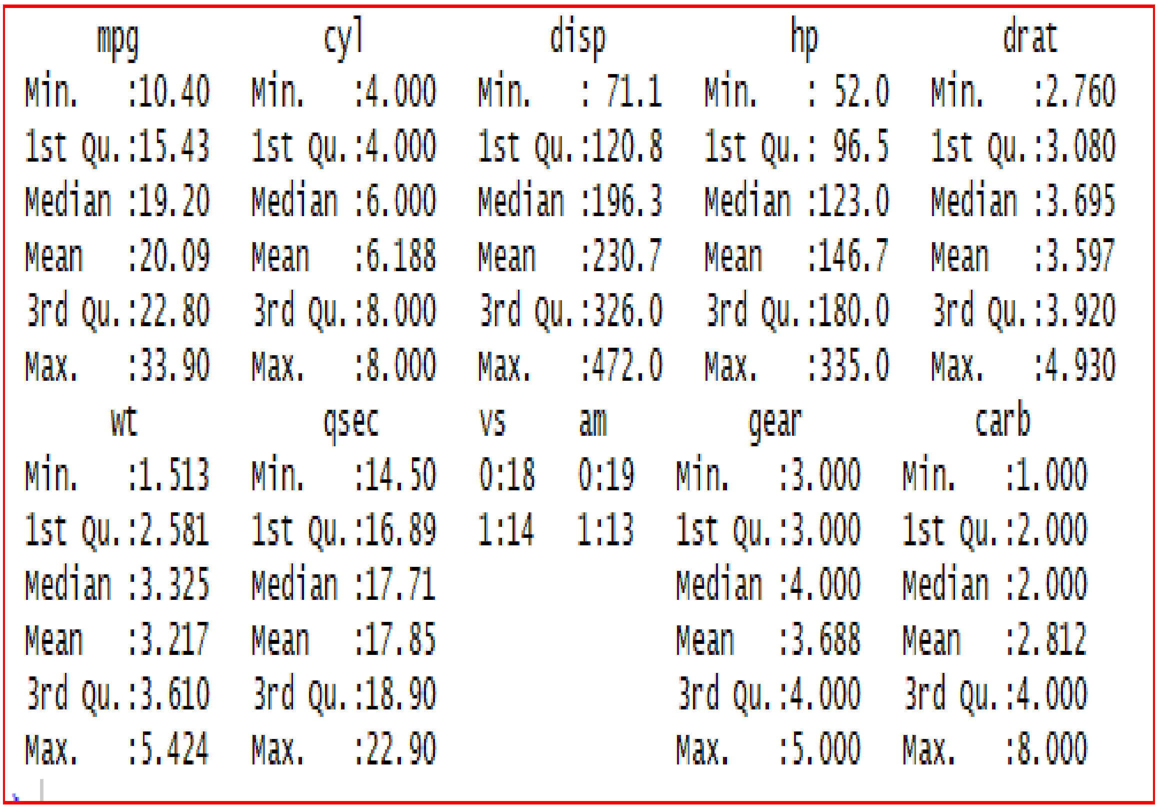


```
'data.frame':   32 obs. of  11 variables:
 $ mpg : num  21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...
 $ cyl : int  6 6 4 6 8 6 8 4 4 6 ...
 $ disp: num  160 160 108 258 360 ...
 $ hp  : int  110 110 93 110 175 105 245 62 95 123 ...
 $ drat: num  3.9 3.9 3.85 3.08 3.15 2.76 3.21 3.69 3.92 3.92 ...
 $ wt  : num  2.62 2.88 2.32 3.21 3.44 ...
 $ qsec: num  16.5 17 18.6 19.4 17 ...
 $ vs  : Factor w/ 2 levels "0","1": 1 1 2 2 1 2 1 2 2 2 ...
 $ am  : Factor w/ 2 levels "0","1": 2 2 2 1 1 1 1 1 1 1 ...
 $ gear: int  4 4 4 3 3 3 3 4 4 4 ...
 $ carb: int  4 4 1 1 2 1 4 2 2 4 ...
```

Output:-3.1.2 converting integers into factors

❖ Understanding data using Descriptive Statistics:

- We will look at the summary of the data.



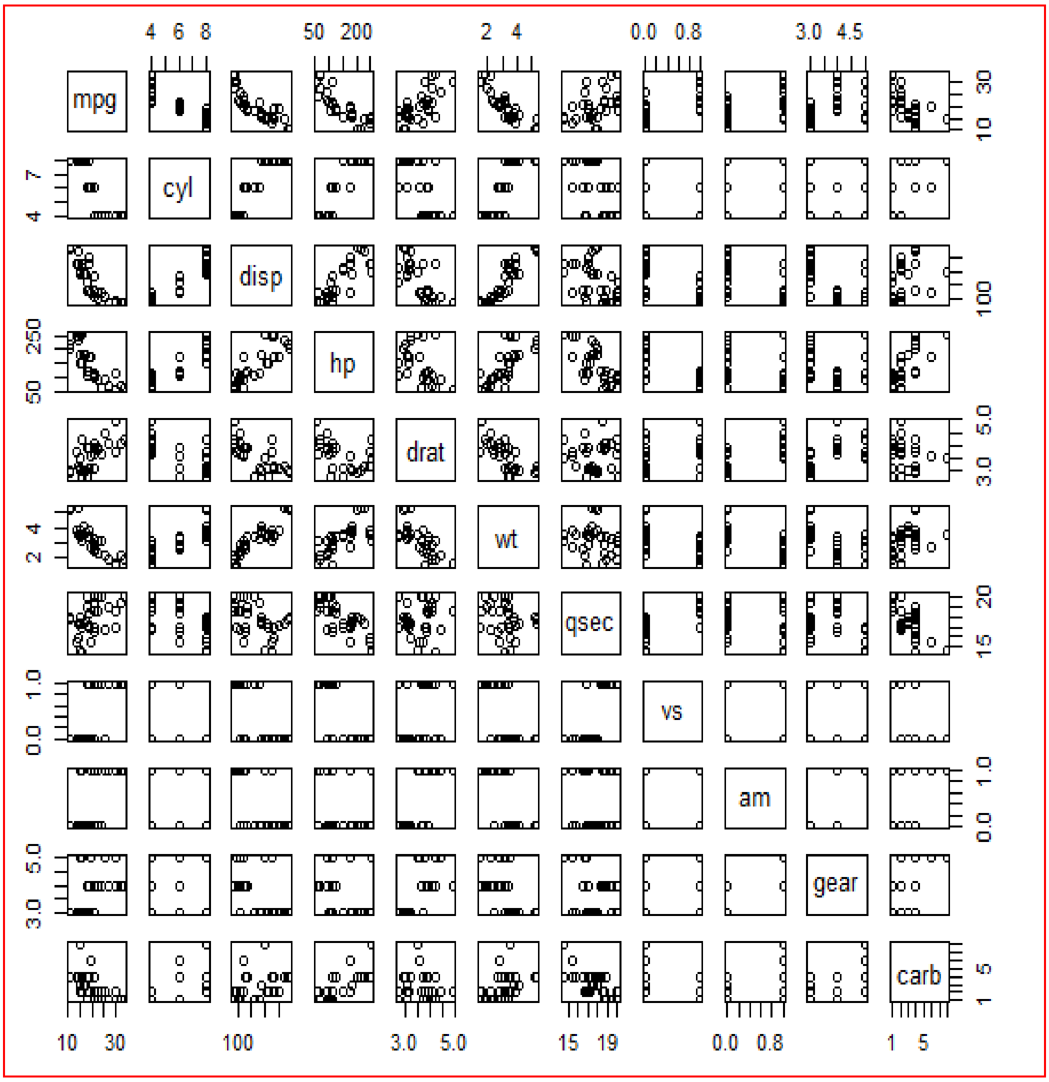
```
      mpg             cyl             disp             hp             drat      
 Min.   :10.40   Min.   :4.000   Min.   : 71.1   Min.   : 52.0   Min.   :2.760  
 1st Qu.:15.43   1st Qu.:4.000   1st Qu.:120.8   1st Qu.: 96.5   1st Qu.:3.080  
 Median :19.20   Median :6.000   Median :196.3   Median :123.0   Median :3.695  
 Mean   :20.09   Mean   :6.188   Mean   :230.7   Mean   :146.7   Mean   :3.597  
 3rd Qu.:22.80   3rd Qu.:8.000   3rd Qu.:326.0   3rd Qu.:180.0   3rd Qu.:3.920  
 Max.   :33.90   Max.   :8.000   Max.   :472.0   Max.   :335.0   Max.   :4.930  
       wt             qsec       vs     am          gear            carb      
 Min.   :1.513   Min.   :14.50   0:18   0:19   Min.   :3.000   Min.   :1.000  
 1st Qu.:2.581   1st Qu.:16.89   1:14   1:13   1st Qu.:3.000   1st Qu.:2.000  
 Median :3.325   Median :17.71                 Median :4.000   Median :2.000  
 Mean   :3.217   Mean   :17.85                 Mean   :3.688   Mean   :2.812  
 3rd Qu.:3.610   3rd Qu.:18.90                 3rd Qu.:4.000   3rd Qu.:4.000  
 Max.   :5.424   Max.   :22.90                 Max.   :5.000   Max.   :8.000  
```

Output:- 3.1.3 summary of the data..

Understanding data visually

Also, look at the data visually to understand the relationships between and

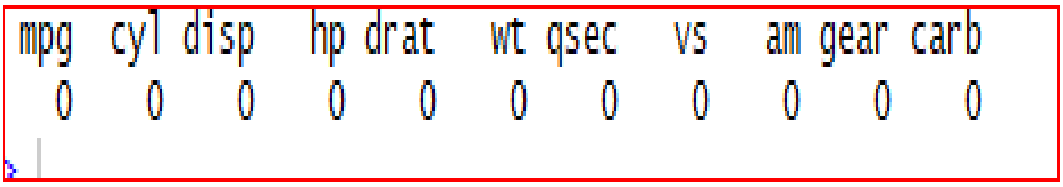
Within the variable.



Output:- 3.1.4 Data visualisation.

From the above tables we can check the location, spread proportion of all the Variables in the data.

❖ Checking for missing Values:



```
 mpg  cyl disp   hp drat   wt qsec   vs   am gear carb
   0    0    0    0    0    0    0    0    0    0    0
>
```

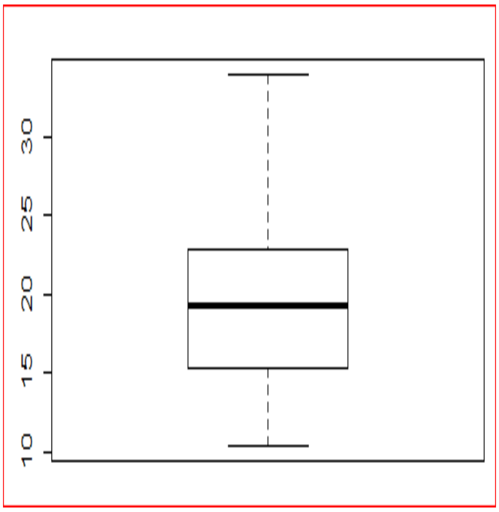
Output:-3.1.5 Missing values.

The missing values for the continuous variables will be imputed using Mean or Median value of the valid records and the categorical variables will be imputed Using Mode value.

❖ Checking for Outliers:

We used Box-plots to check for Outliers in each of the continuous variables.

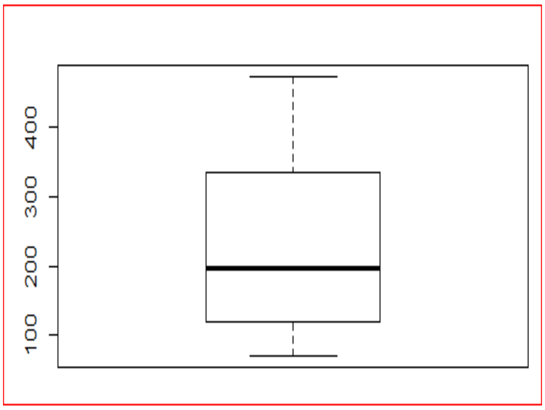
(Box plot for mpg):



❖ There are no outliers.

Output:- 3.1.6 Box plot for mpg

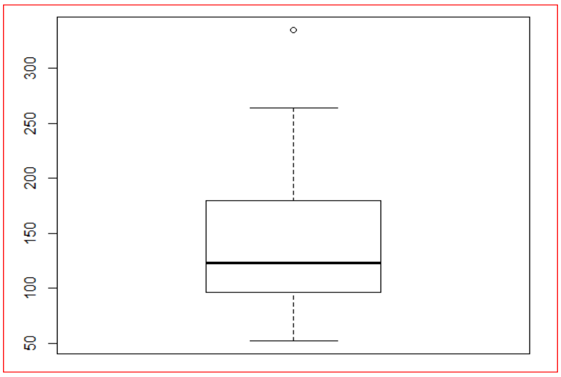
(Box plot for disc):



- ❖ There are no outliers.

Output:- 3.1.7 Box plot for disc

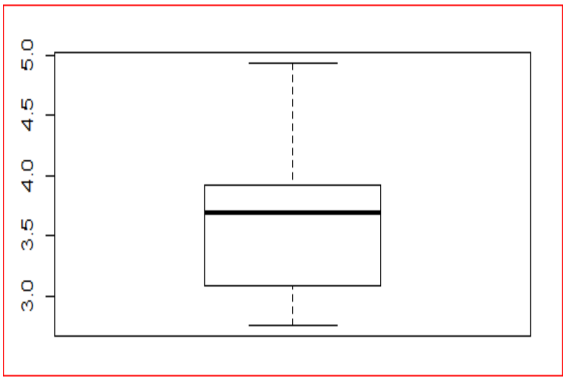
(Box plot for hp):



- ❖ There is an outlier.

Output:- 3.1.8 Box plot for hp

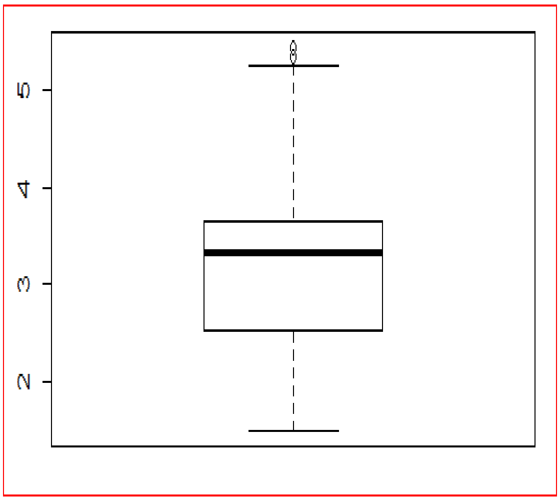
(Box plot for drat):



- ❖ There are no outliers.

Output:- 3.1.9 Box plot for drat

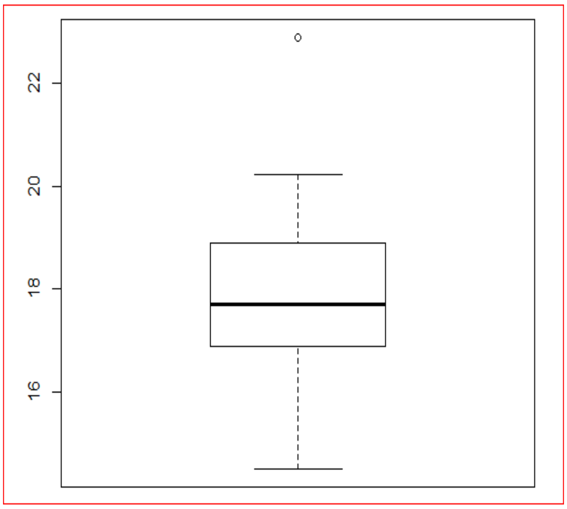
(Box plot for wt)



- ❖ There are two outliers.

Output:- 3.1.10 Box plot for wt

(Box plot for qsec):



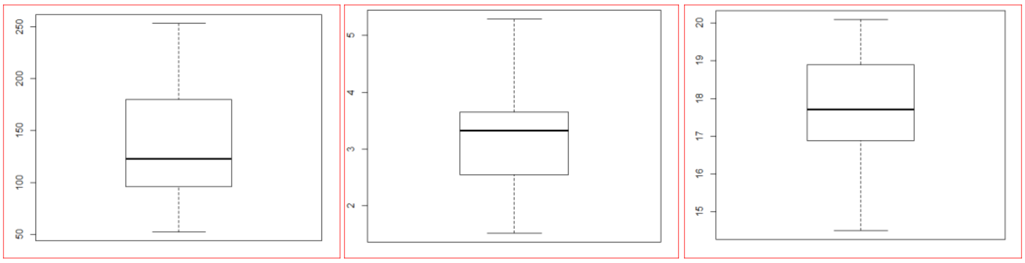
- ❖ There is an outlier.

Output:- 3.1.11Box plot for qsec

Values more than 95th percentile will be imputed using the 95th percentile value and the values less than 5th percentile will be imputed using 5th percentile value.

Box plots after removing Outliers:

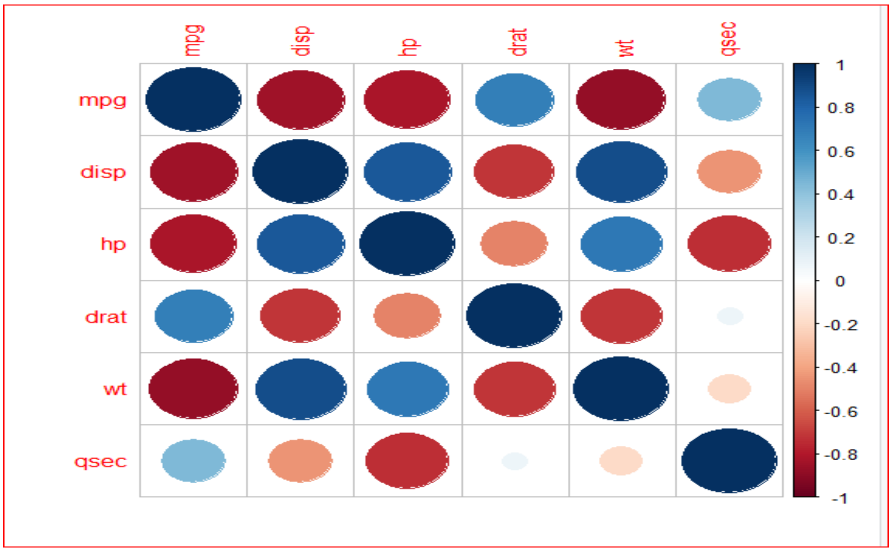
Output :- 3.1.12 Removing outliers for hp, wt, qsec



Here, we can see that there are no more Outliers in all these variables.

Understanding relationships between variables:

For the continuous variables, we will look at the Correlation plots to understand the relationships between variables.



Output:- 3.1.13 Correlation plot

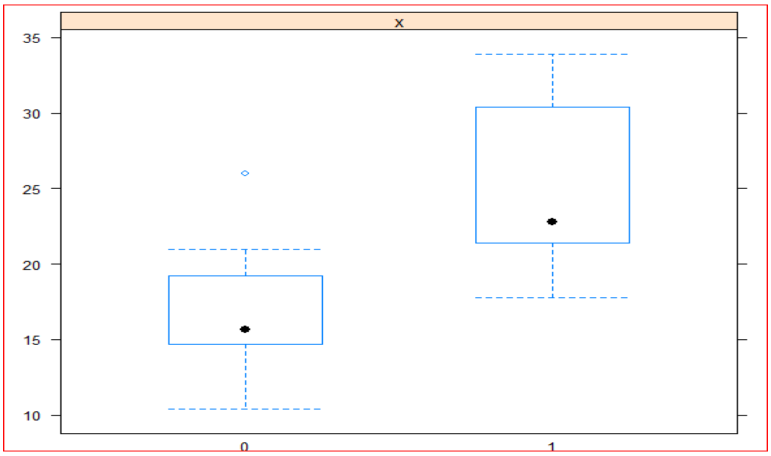
Here, the circle size refers to the strength of the relation and colour refers to the direction of the relationship.

From the plot, we can see that mpg, wt are negatively correlated and wt, disp are positively correlated.

❖ Feather Plots :

For the continuous v/s categorical variable, we will look at Feature plots to understand the relationships.

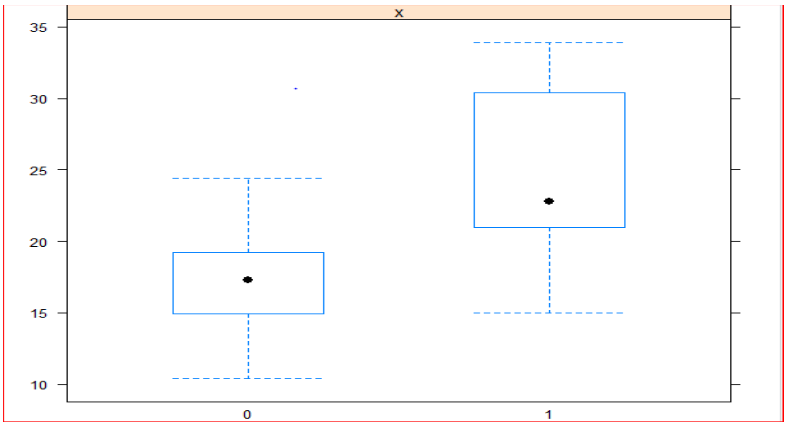
MPG v/s VS physical test



Output:-3.1.14 feature plot

From the above plot, we observe that straight engine has more mpg than v-shaped engines.

.MPG v/s AM physical test:



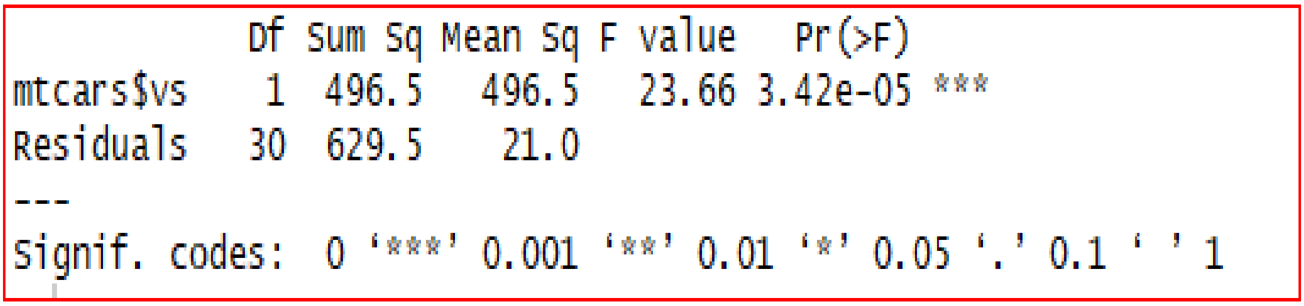
Output:- 3.1.15 feature plot

From the above plot we observe that manual transmission has more mpg than automatic transmission.

❖ Checking for the significance difference between variables:

To test the significance difference between continues and categorical variables, we will look at the t-test values.

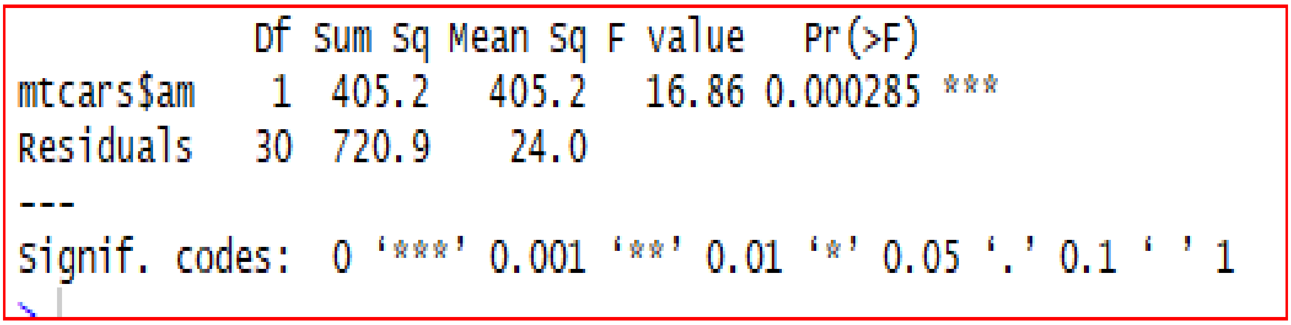
❖ **Output**:-3.1.16 Mpg v/s vs



```
              Df Sum Sq Mean Sq F value   Pr(>F)    
mtcars$vs      1  496.5   496.5   23.66 3.42e-05 ***
Residuals     30  629.5    21.0                     
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

From the above table as the p-value <0.05, we conclude that mpg and vs are more significant.

❖ **Output**:-3.1.17 Mpg v/s Am

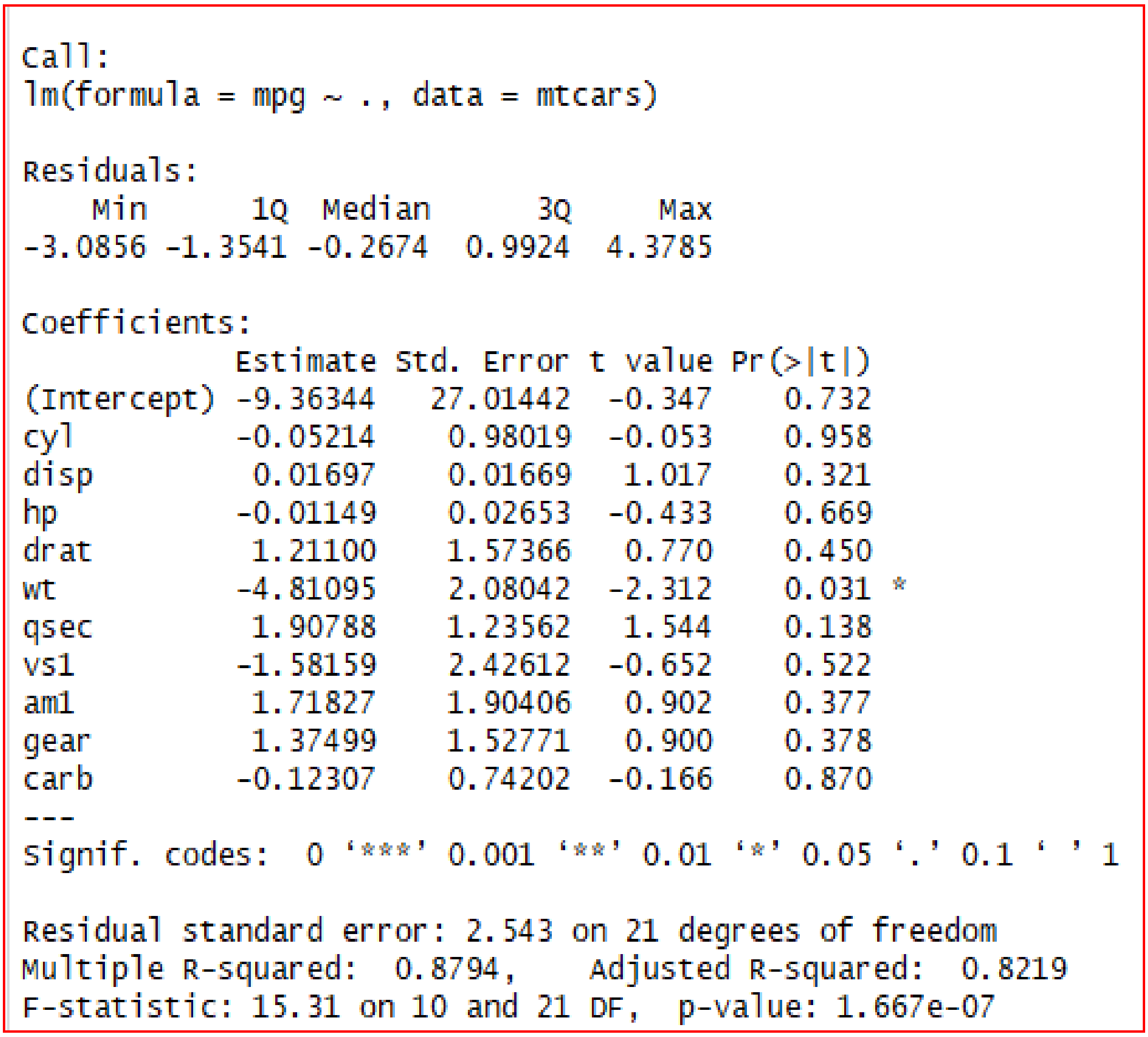


```
              Df Sum Sq Mean Sq F value   Pr(>F)    
mtcars$am      1  405.2   405.2   16.86 0.000285 ***
Residuals     30  720.9    24.0                     
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

From the above table as the p-value <0.05 .we conclude that mpg and am are more significant.

❖ Fitting a model:

To predict the mileage, we fitted a linear model and below is an output obtained for the same:



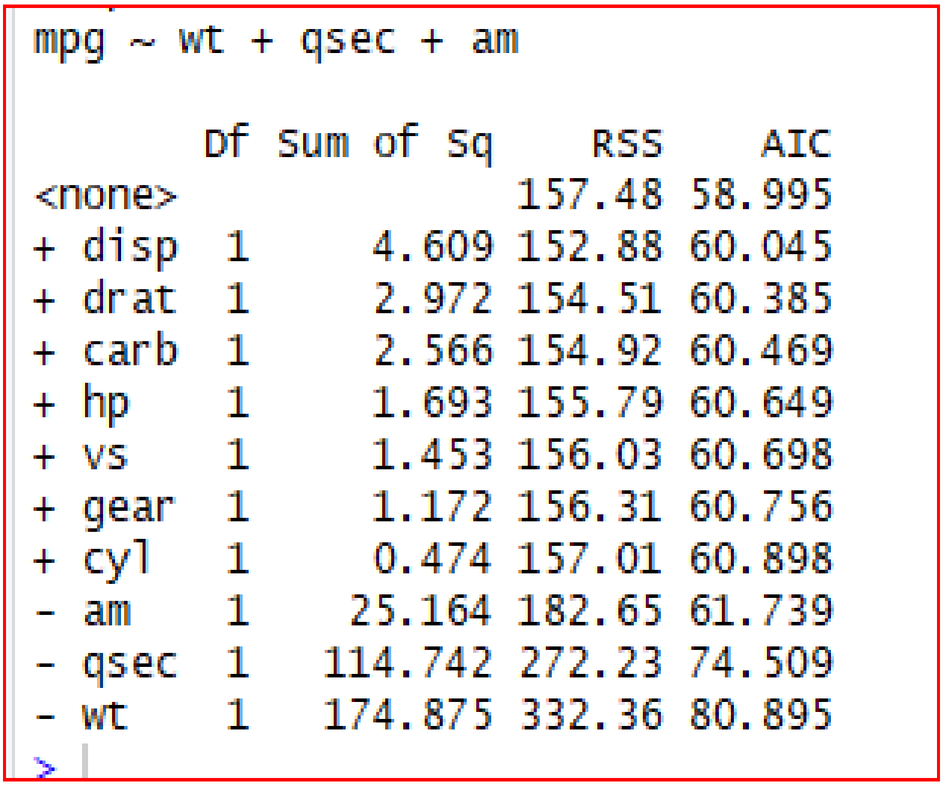
```
Call:
lm(formula = mpg ~ ., data = mtcars)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.0856 -1.3541 -0.2674  0.9924  4.3785 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)  
(Intercept) -9.36344   27.01442  -0.347    0.732  
cyl         -0.05214    0.98019  -0.053    0.958  
disp         0.01697    0.01669   1.017    0.321  
hp          -0.01149    0.02653  -0.433    0.669  
drat         1.21100    1.57366   0.770    0.450  
wt          -4.81095    2.08042  -2.312    0.031 *
qsec         1.90788    1.23562   1.544    0.138  
vs1         -1.58159    2.42612  -0.652    0.522  
am1          1.71827    1.90406   0.902    0.377  
gear         1.37499    1.52771   0.900    0.378  
carb        -0.12307    0.74202  -0.166    0.870  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.543 on 21 degrees of freedom
Multiple R-squared:  0.8794,	Adjusted R-squared:  0.8219 
F-statistic: 15.31 on 10 and 21 DF,  p-value: 1.667e-07
```

Output:-3.1.18 Model fitting.

From the above coefficients , we can conclude that our model is explaining about 82% of the total variance in mpg(Mileage), using all the variables as input.

❖ Finding the key variable using step-wise model:

In order to find key variables from the list of all variables, we run the step-wise model and obtained key parameters.



```
mpg ~ wt + qsec + am

       Df Sum of Sq    RSS    AIC
<none>              157.48 58.995
+ disp  1     4.609 152.88 60.045
+ drat  1     2.972 154.51 60.385
+ carb  1     2.566 154.92 60.469
+ hp    1     1.693 155.79 60.649
+ vs    1     1.453 156.03 60.698
+ gear  1     1.172 156.31 60.756
+ cyl   1     0.474 157.01 60.898
- am    1    25.164 182.65 61.739
- qsec  1   114.742 272.23 74.509
- wt    1   174.875 332.36 80.895
>
```

Output:- 3.1.19 key variables.

From the above table we can conclude that the key variables are wt, qsec and am.

CHAPTER-4

SUMMARY

4. Summary

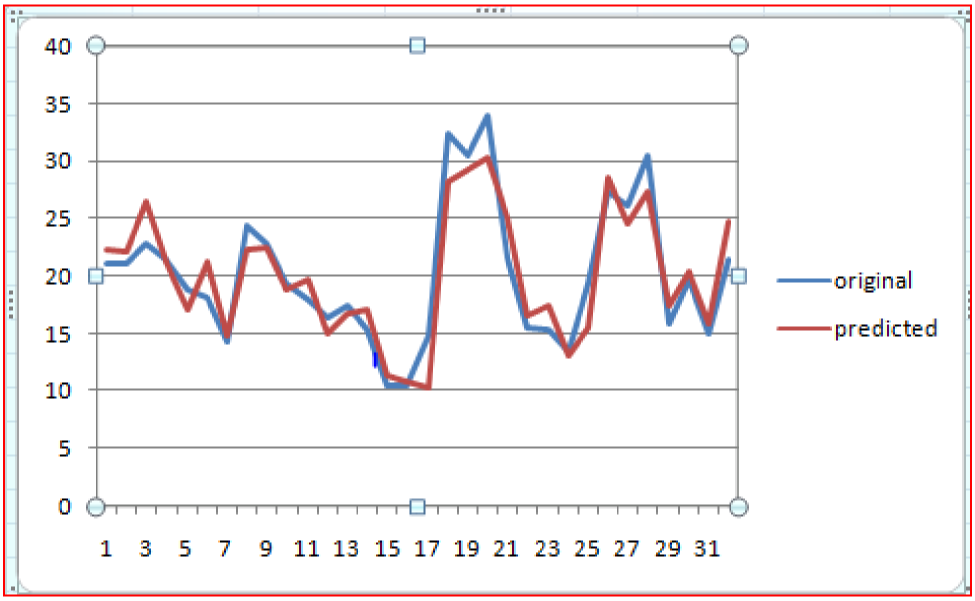
In order to solve the above problem, we have applied multiple linear regression model and shortlisted the linear- regression based on R-square value. And also we applied the step-wise method to find the key variables for predicting the mileages and the key variables are wt, qsec and am.

Hence we have applied the liner regression model for this key variables and obtained the test data and computed the R-square value for the test as below.

R-SQUARE OF TEST DATA IS :

0.8233, i.e. 82% accuracy

Graphical Representation



Output:-Since the test data is more accurate. Hence this model is generalized. So we can use this model to predict the future data.

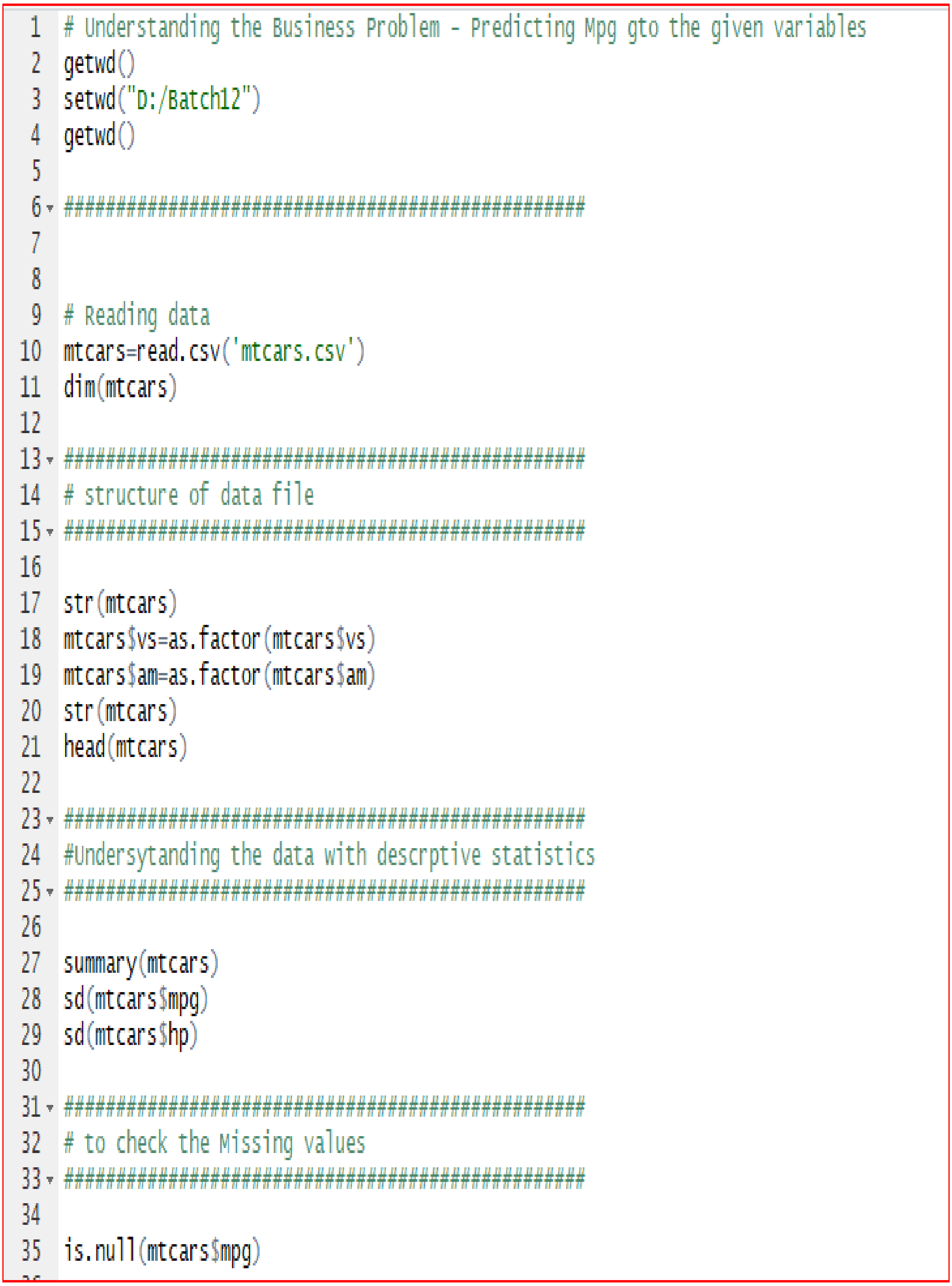
5.APPENDIX

R-CODE

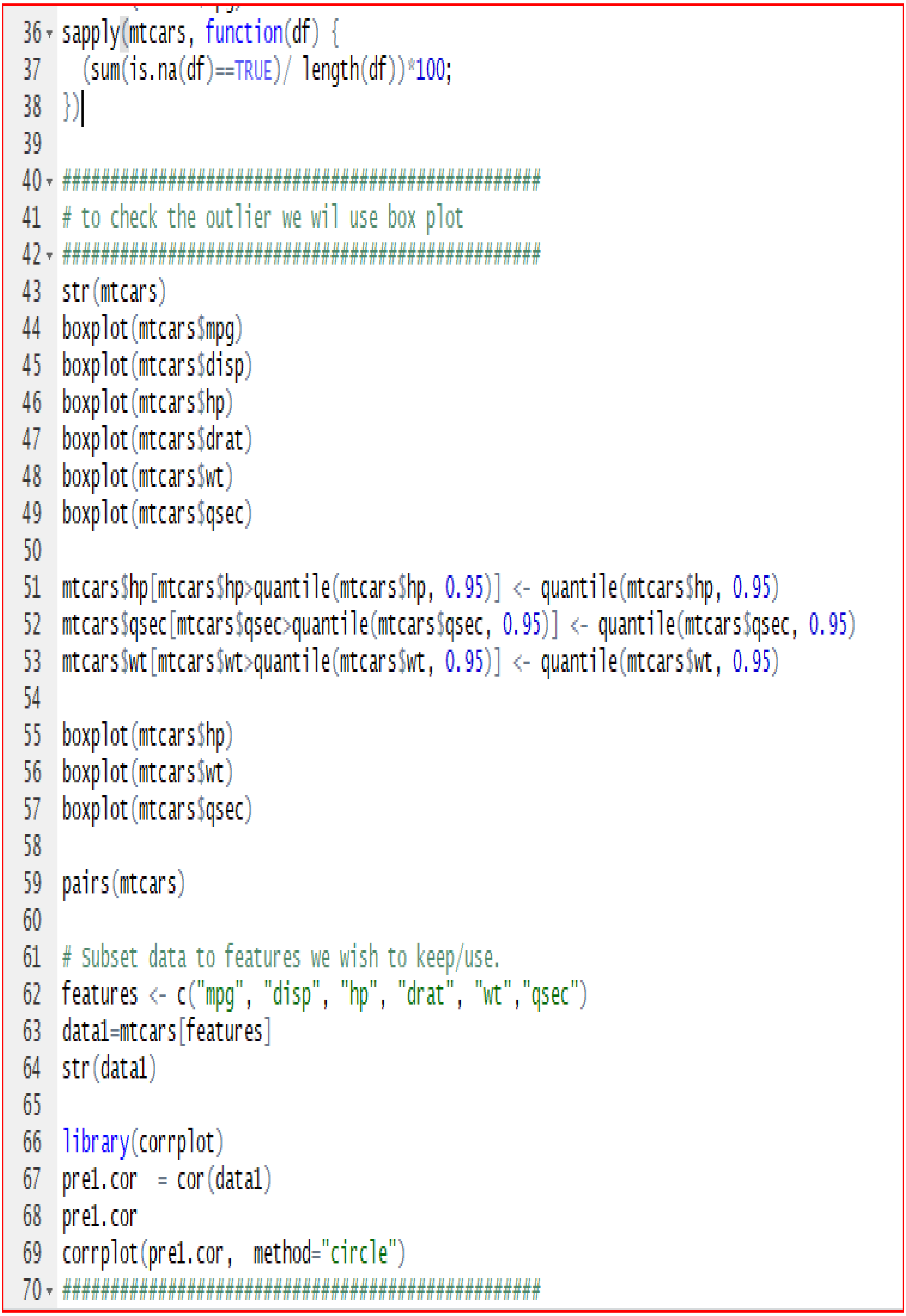
DATASET

BIBILOGRAPHY

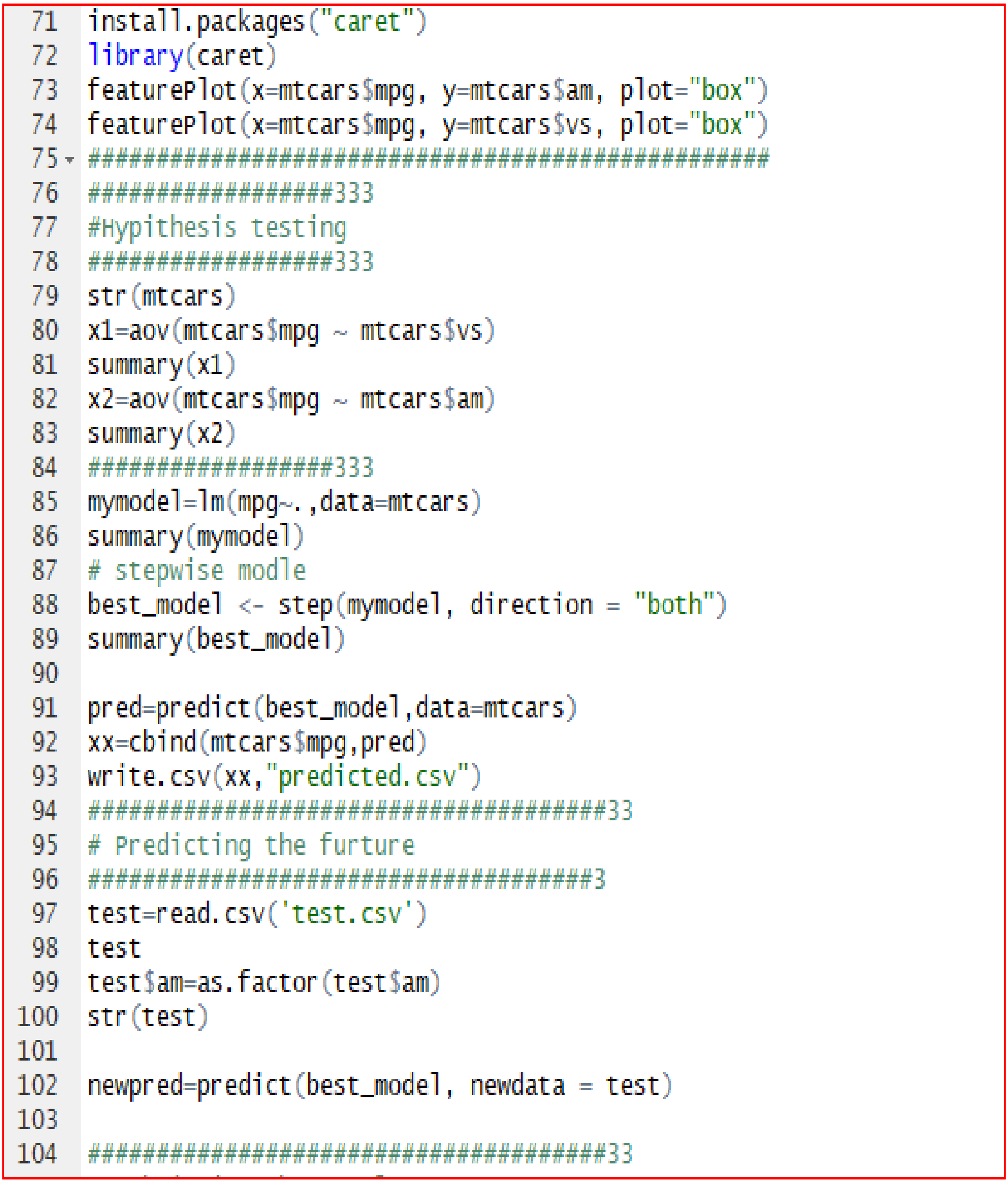
R-CODE



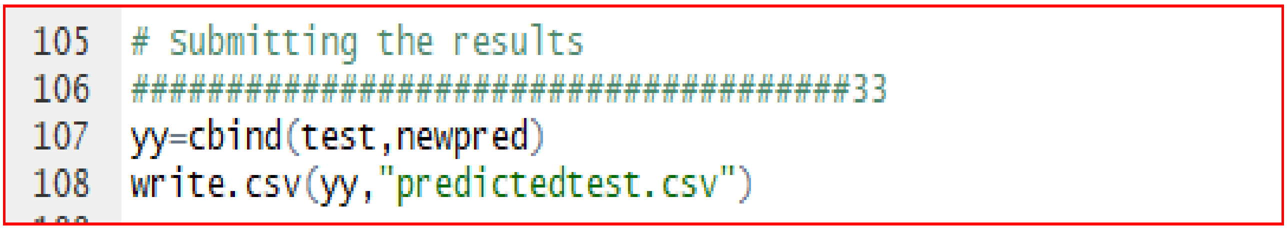
```
# Understanding the Business Problem - Predicting Mpg gto the given variables
getwd()
setwd("D:/Batch12")
getwd()

##################################################


# Reading data
mtcars=read.csv('mtcars.csv')
dim(mtcars)

##################################################
# structure of data file
##################################################

str(mtcars)
mtcars$vs=as.factor(mtcars$vs)
mtcars$am=as.factor(mtcars$am)
str(mtcars)
head(mtcars)

##################################################
#Undersytanding the data with descrptive statistics
##################################################

summary(mtcars)
sd(mtcars$mpg)
sd(mtcars$hp)

##################################################
# to check the Missing values
##################################################

is.null(mtcars$mpg)
```



```
sapply(mtcars, function(df) {
  (sum(is.na(df)==TRUE)/ length(df))*100;
})

##################################################
# to check the outlier we wil use box plot
##################################################
str(mtcars)
boxplot(mtcars$mpg)
boxplot(mtcars$disp)
boxplot(mtcars$hp)
boxplot(mtcars$drat)
boxplot(mtcars$wt)
boxplot(mtcars$qsec)

mtcars$hp[mtcars$hp>quantile(mtcars$hp, 0.95)] <- quantile(mtcars$hp, 0.95)
mtcars$qsec[mtcars$qsec>quantile(mtcars$qsec, 0.95)] <- quantile(mtcars$qsec, 0.95)
mtcars$wt[mtcars$wt>quantile(mtcars$wt, 0.95)] <- quantile(mtcars$wt, 0.95)

boxplot(mtcars$hp)
boxplot(mtcars$wt)
boxplot(mtcars$qsec)

pairs(mtcars)

# Subset data to features we wish to keep/use.
features <- c("mpg", "disp", "hp", "drat", "wt","qsec")
data1=mtcars[features]
str(data1)

library(corrplot)
pre1.cor  = cor(data1)
pre1.cor
corrplot(pre1.cor,  method="circle")
##################################################
```

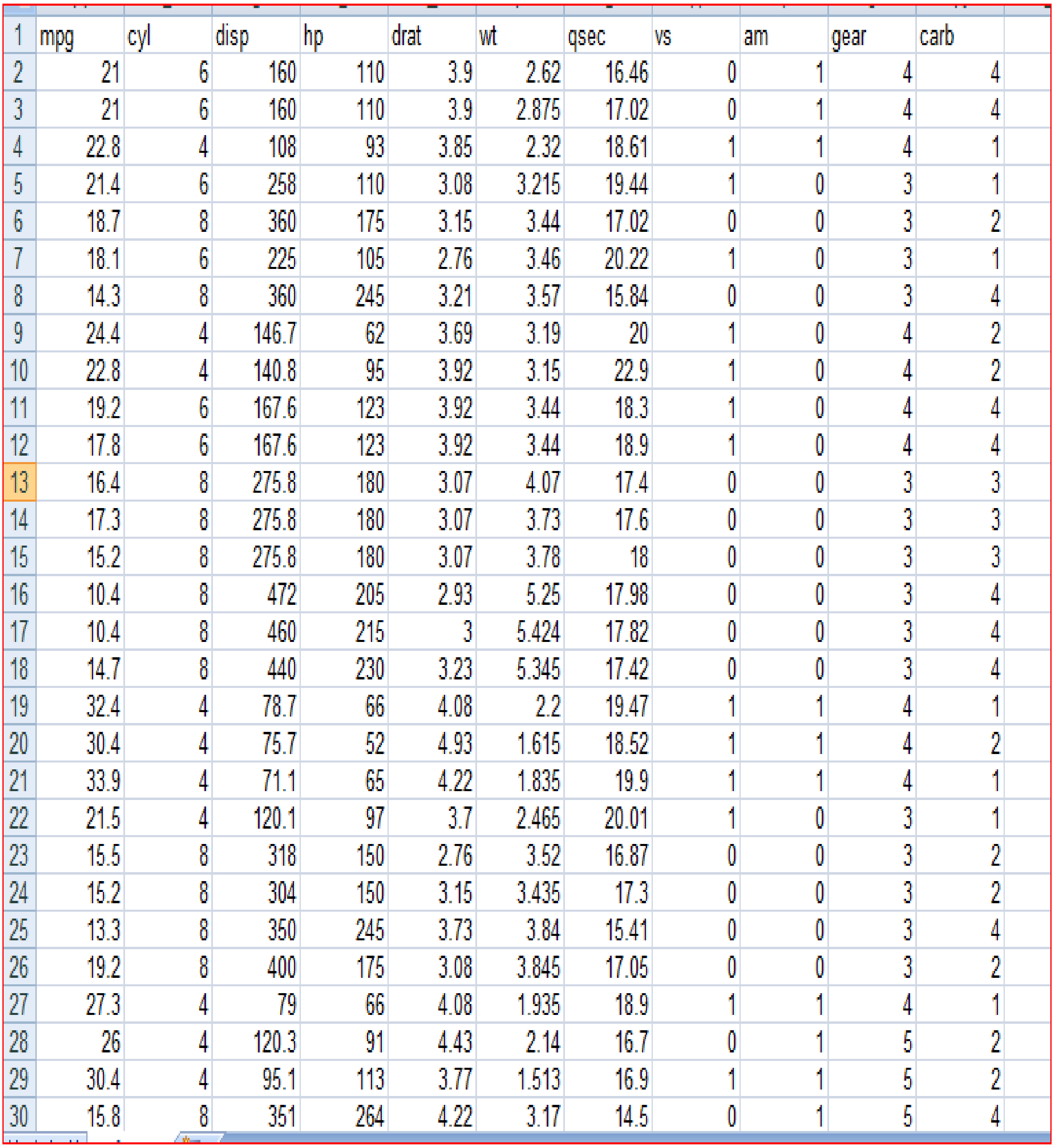


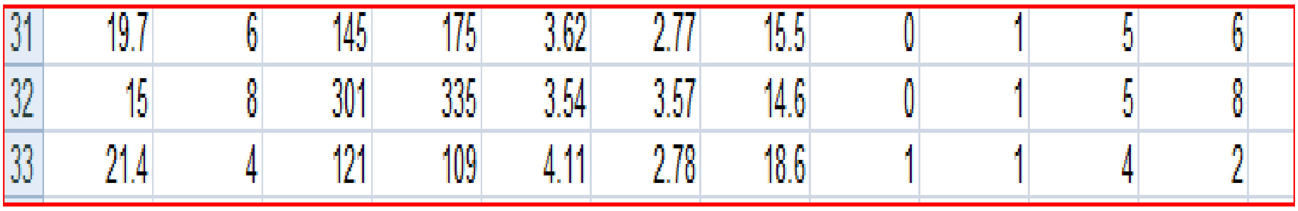
```
install.packages("caret")
library(caret)
featurePlot(x=mtcars$mpg, y=mtcars$am, plot="box")
featurePlot(x=mtcars$mpg, y=mtcars$vs, plot="box")
###################################################
##################333
#Hypithesis testing
##################333
str(mtcars)
x1=aov(mtcars$mpg ~ mtcars$vs)
summary(x1)
x2=aov(mtcars$mpg ~ mtcars$am)
summary(x2)
##################333
mymodel=lm(mpg~.,data=mtcars)
summary(mymodel)
# stepwise modle
best_model <- step(mymodel, direction = "both")
summary(best_model)

pred=predict(best_model,data=mtcars)
xx=cbind(mtcars$mpg,pred)
write.csv(xx,"predicted.csv")
#######################################33
# Predicting the furture
######################################3
test=read.csv('test.csv')
test
test$am=as.factor(test$am)
str(test)

newpred=predict(best_model, newdata = test)

#######################################33
```



```
# Submitting the results
######################################33
yy=cbind(test,newpred)
write.csv(yy,"predictedtest.csv")
```

DATA SET





1	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
2	21	6	160	110	3.9	2.62	16.46	0	1	4	4
3	21	6	160	110	3.9	2.875	17.02	0	1	4	4
4	22.8	4	108	93	3.85	2.32	18.61	1	1	4	1
5	21.4	6	258	110	3.08	3.215	19.44	1	0	3	1
6	18.7	8	360	175	3.15	3.44	17.02	0	0	3	2
7	18.1	6	225	105	2.76	3.46	20.22	1	0	3	1
8	14.3	8	360	245	3.21	3.57	15.84	0	0	3	4
9	24.4	4	146.7	62	3.69	3.19	20	1	0	4	2
10	22.8	4	140.8	95	3.92	3.15	22.9	1	0	4	2
11	19.2	6	167.6	123	3.92	3.44	18.3	1	0	4	4
12	17.8	6	167.6	123	3.92	3.44	18.9	1	0	4	4
13	16.4	8	275.8	180	3.07	4.07	17.4	0	0	3	3
14	17.3	8	275.8	180	3.07	3.73	17.6	0	0	3	3
15	15.2	8	275.8	180	3.07	3.78	18	0	0	3	3
16	10.4	8	472	205	2.93	5.25	17.98	0	0	3	4
17	10.4	8	460	215	3	5.424	17.82	0	0	3	4
18	14.7	8	440	230	3.23	5.345	17.42	0	0	3	4
19	32.4	4	78.7	66	4.08	2.2	19.47	1	1	4	1
20	30.4	4	75.7	52	4.93	1.615	18.52	1	1	4	2
21	33.9	4	71.1	65	4.22	1.835	19.9	1	1	4	1
22	21.5	4	120.1	97	3.7	2.465	20.01	1	0	3	1
23	15.5	8	318	150	2.76	3.52	16.87	0	0	3	2
24	15.2	8	304	150	3.15	3.435	17.3	0	0	3	2
25	13.3	8	350	245	3.73	3.84	15.41	0	0	3	4
26	19.2	8	400	175	3.08	3.845	17.05	0	0	3	2
27	27.3	4	79	66	4.08	1.935	18.9	1	1	4	1
28	26	4	120.3	91	4.43	2.14	16.7	0	1	5	2
29	30.4	4	95.1	113	3.77	1.513	16.9	1	1	5	2
30	15.8	8	351	264	4.22	3.17	14.5	0	1	5	4
31	19.7	6	145	175	3.62	2.77	15.5	0	1	5	6
32	15	8	301	335	3.54	3.57	14.6	0	1	5	8
33	21.4	4	121	109	4.11	2.78	18.6	1	1	4	2

6. BIBLIOGRAPHY

1. Multivariate data analysis (Fifth Edition) --- Joseph F. Hair, RolphE.Anderson, Ronald l Tatham and William C. Black

2. Data Mining- Theories, Algorithms, and Examples – NoNG YE

3. A Practical Guide to Data Mining for Business and Industry -- Andrea Ahlemeyer-Stubbe, Shirley Coleman

4. Data Mining and Predictive Analytics – Daniel T. Larose, Chantal D.Lorse

5. machine_learning_mastery_with_r. – Jason Brownlee

6. master_machine_learning_algorithms -- Jason Brownlee

7. statistical_methods_for_machine_learning - Jason Brownlee

8. Machine Learning Using R -- KarthikRamasubramanian ,Abhishek Singh

9. Data Science for Business - Forster Provost & Tom Fawcett

10. Deep learning with Deep learning R by François Chollet