Chapter 2

Preparing to Model

OBJECTIVE OF THE CHAPTER:

This chapter gives a detailed view of how to understand the incoming data and create basic understanding about the nature and quality of the data. This information, in turn, helps to select and then how to apply the model. So, the knowledge imparted in this chapter helps a beginner take the first step towards effective modelling and solving a machine learning problem.

2.1 INTRODUCTION

In the last chapter, we got introduced to machine learning. In the beginning, we got a glimpse of the journey of machine learning as an evolving technology. It all started as a proposition from the renowned computer scientist Alan Turing – machines can 'learn' and become artificially intelligent. Gradually, through the next few decades path-breaking innovations came in from Arthur Samuel, Frank Rosenblatt, John Hopfield, Christopher Watkins, Geoffrey Hinton and many other computer scientists. They shaped up concepts of Neural Networks, Recurrent Neural Network, Reinforcement Learning, Deep Learning, etc. which took machine learning to new heights. In parallel, interesting applications of machine learning kept on happening, with organizations like IBM and Google taking a lead. What started with IBM's Deep Blue beating the world chess champion Gary Kasparov, continued with IBM's Watson beating two human champions in a Jeopardy competition.Google also started with a series of innovations applying machine learning. The Google Brain, Sibyl, Waymo, AlphaGo programs – are all extremely advanced applications of machine learning which have taken the technology a few notches

up. Now we can see an all-pervasive presence of machine learning technology in all walks of life.

We have also seen the types of human learning and how that, in some ways, can be related to the types of machine learning – supervised, unsupervised, and reinforcement. Supervised learning, as we saw, implies learning from past data, also called training data, which has got known values or classes. Machines can 'learn' or get 'trained' from the past data and assign classes or values to unknown data, termed as test data. This helps in solving problems related to prediction. This is much like human learning through expert guidance as happens for infants from parents or students through teachers. So, supervised learning in case of machines can be perceived as guided learning from human inputs. Unsupervised machine learning doesn't have labelled data to learn from. It tries to find patterns in unlabelled data. This is much like human beings trying to group together objects of similar shape. This learning is not guided by labelled inputs but uses the knowledge gained from the labels themselves. Last but not the least is reinforcement learning in which machine tries to learn by itself through penalty/ reward mechanism – again pretty much in the same way as human self-learning happens.

Lastly, we saw some of the applications of machine learning in different domains such as banking and finance, insurance, and healthcare. Fraud detection is a critical business case which is implemented in almost all banks across the world and uses machine learning predominantly. Risk prediction for new customers is a similar critical case in the insurance industry which finds the application of machine learning. In the healthcare sector, disease prediction makes wide use of machine learning, especially in the developed countries.

While development in machine learning technology has been extensive and its implementation has become widespread, to start as a practitioner, we need to gain some basic understanding. We need to understand how to apply the array of tools and technologies available in the machine learning to solve a problem. In fact, that is going to be very specific to the

kind of problem that we are trying to solve. If it is a prediction problem, the kind of activities that will be involved is going to be completely different vis-à-vis if it is a problem where we are trying to unfold a pattern in a data without any past knowledge about the data. So how a machine learning project looks like or what are the salient activities that form the core of a machine learning project will depend on whether it is in the area of supervised or unsupervised or reinforcement learning area. However, irrespective of the variation, some foundational knowledge needs to be built before we start with the core machine learning concepts and key algorithms. In this section, we will have a quick look at a few typical machine learning activities and focus on some of the foundational concepts that all practitioners need to gain as pre-requisites before starting their journey in the area of machine learning.

Points to Ponder

No man is perfect. The same is applicable for machines. To increase the level of accuracy of a machine, human participation should be added to the machine learning process. In short, incorporating human intervention is the recipe for the success of machine learning.

2.2 MACHINE LEARNING ACTIVITIES

The first step in machine learning activity starts with data. In case of supervised learning, it is the labelled training data set followed by test data which is not labelled. In case of unsupervised learning, there is no question of labelled data but the task is to find patterns in the input data. A thorough review and exploration of the data is needed to understand the type of the data, the quality of the data and relationship between the different data elements. Based on that, multiple pre-processing activities may need to be done on the input data before we can go ahead with core machine learning activities. Following are the typical **preparation** activities done once the input data comes into the machine learning system:

- Understand the type of data in the given input data set.

- Explore the data to understand the nature and quality.
- Explore the relationships amongst the data elements, e.g. inter-feature relationship.
- Find potential issues in data.
- Do the necessary remediation, e.g. impute missing data values, etc., if needed.
- Apply pre-processing steps, as necessary.
- Once the data is prepared for modelling, then the learning tasks start off. As a part of it, do the following activities:
  - The input data is first divided into parts – the training data and the test data (called holdout). This step is applicable for supervised learning only.
  - Consider different models or learning algorithms for selection.
  - Train the model based on the training data for supervised learning problem and apply to unknown data. Directly apply the chosen un-supervised model on the input data for unsupervised learning problem.

After the model is selected, trained (for supervised learning), and applied on input data, the performance of the model is evaluated. Based on options available, specific actions can be taken to improve the performance of the model, if possible.

Figure 2.1 depicts the four-step process of machine learning.

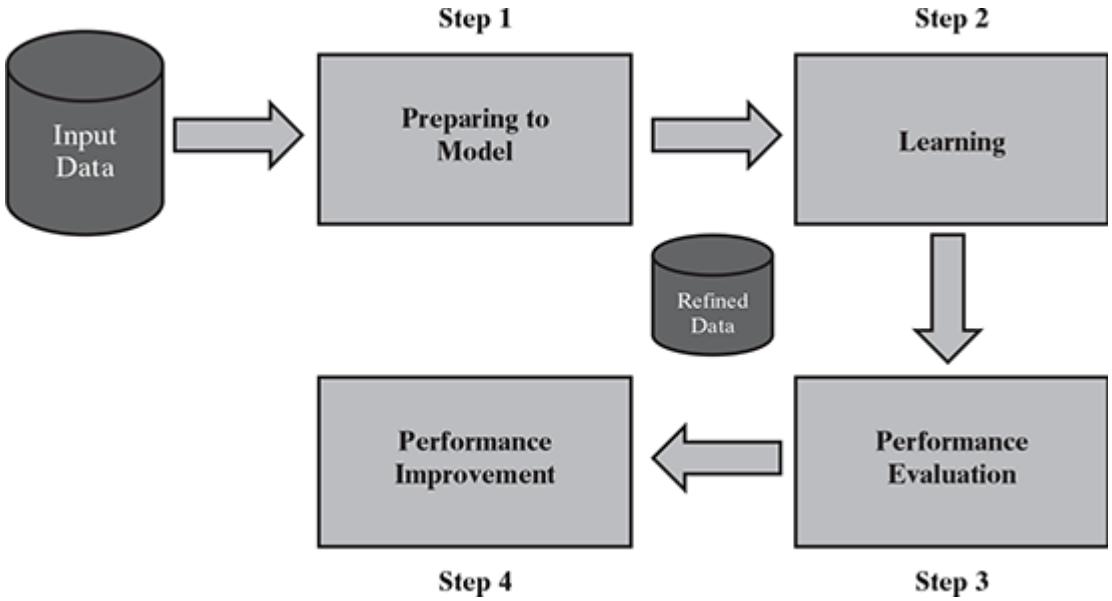

FIG. 2.1 Detailed process of machine learning

Table 2.1 contains a summary of steps and activities involved:

Table 2.1 *Activities in Machine Learning*

Step #	Step Name	Activities Involved
Step 1	Preparing to Model	• Understand the type of data in the given input data set • Explore the data to understand data quality • Explore the relationships amongst the data elements, e.g. inter-feature relationship • Find potential issues in data • Remediate data, if needed • Apply following pre-processing steps, as necessary: ✓ Dimensionality reduction ✓ Feature subset selection
Step 2	Learning	• Data partitioning/holdout • Model selection • Cross-validation
Step 3	Performance evaluation	• Examine the model performance, e.g. confusion matrix in case of classification • Visualize performance trade-offs using ROC curves
Step 4	Performance improvement	• Tuning the model • Ensembling • Bagging • Boosting

In this chapter, we will cover the first part, i.e. preparing to model. The remaining parts, i.e. learning, performance evaluation, and performance

improvement will be covered in Chapter 3.

2.3 BASIC TYPES OF DATA IN MACHINE LEARNING

Before starting with types of data, let's first understand what a data set is and what are the elements of a data set. A data set is a collection of related information or records. The information may be on some entity or some subject area. For example (Fig. 2.2), we may have a data set on students in which each record consists of information about a specific student. Again, we can have a data set on student performance which has records providing performance, i.e. marks on the individual subjects.

Each row of a data set is called a record. Each data set also has multiple attributes, each of which gives information on a specific characteristic. For example, in the data set on students, there are four attributes namely Roll Number, Name, Gender, and Age, each of which understandably is a specific characteristic about the student entity. Attributes can also be termed as feature, variable, dimension or field. Both the data sets, Student and Student Performance, are having four features or dimensions; hence they are told to have four-dimensional data space. A row or record represents a point in the four-dimensional data space as each row has specific values for each of the four attributes or features. Value of an attribute, quite understandably, may vary from record to record. For example, if we refer to the first two records in the Student data set, the value of attributes Name, Gender, and Age are different (Fig. 2.3).

Student data set:

Roll Number	Name	Gender	Age
129/011	Mihir Karmarkar	M	14
129/012	Geeta Iyer	F	15
129/013	Chanda Bose	F	14
129/014	Sreenu Subramanian	M	14
129/015	Pallav Gupta	M	16
129/016	Gajanan Sharma	M	15

Student performance data set:

Roll Number	Maths	Science	Percentage
129/011	89	45	89.33%
129/012	89	47	90.67%
129/013	68	29	64.67%
129/014	83	38	80.67%
129/015	57	23	53.33%
129/016	78	35	75.33%

FIG. 2.2 Examples of data set

Roll Number	Name	Gender	Age
129/011	Mihir Karmarkar	M	14
129/012	Geeta Iyer	F	15

FIG. 2.3 Data set records and attributes

Now that a context of data sets is given, let's try to understand the different types of data that we generally come across in machine learning problems. Data can broadly be divided into following two types:

1. Qualitative data
2. Quantitative data

Qualitative data provides information about the quality of an object or information which cannot be measured. For example, if we consider the quality of performance of students in terms of 'Good', 'Average', and 'Poor', it falls under the category of qualitative data. Also, name or roll number of students are information that cannot be measured using some scale of measurement. So they would fall under qualitative data.

Qualitative data is also called **categorical data**. Qualitative data can be further subdivided into two types as follows:

1. Nominal data
2. Ordinal data

Nominal data is one which has no numeric value, but a named value. It is used for assigning named values to attributes. Nominal values cannot be quantified. Examples of nominal data are

1. Blood group: A, B, O, AB, etc.
2. Nationality: Indian, American, British, etc.
3. Gender: Male, Female, Other

Note:

A special case of nominal data is when only two labels are possible, e.g. pass/fail as a result of an examination. This sub-type of nominal data is called 'dichotomous'.

It is obvious, mathematical operations such as addition, subtraction, multiplication, etc. cannot be performed on nominal data. For that reason, statistical functions such as mean, variance, etc. can also not be applied on nominal data. However, a basic count is possible. So mode, i.e. most frequently occurring value, can be identified for nominal data.

Ordinal data, in addition to possessing the properties of nominal data, can also be naturally ordered. This means ordinal data also assigns named values to attributes but unlike nominal data, they can be arranged in a sequence of increasing or decreasing value so that we can say whether a value is better than or greater than another value. Examples of ordinal data are

1. Customer satisfaction: 'Very Happy', 'Happy', 'Unhappy', etc.
2. Grades: A, B, C, etc.
3. Hardness of Metal: 'Very Hard', 'Hard', 'Soft', etc.

Like nominal data, basic counting is possible for ordinal data. Hence, the mode can be identified. Since ordering is possible in case of ordinal data, median, and quartiles can be identified in addition. Mean can still not be calculated.

Quantitative data relates to information about the quantity of an object – hence it can be measured. For example, if we consider the attribute 'marks', it can be measured using a scale of measurement. Quantitative data is also termed as numeric data. There are two types of quantitative data:

1. Interval data
2. Ratio data

Interval data is numeric data for which not only the order is known, but the exact difference between values is also known. An ideal example of interval data is Celsius temperature. The difference between each value remains the same in Celsius temperature. For example, the difference between 12°C and 18°C degrees is measurable and is 6°C as in the case of difference between 15.5°C and 21.5°C. Other examples include date, time, etc.

For interval data, mathematical operations such as addition and subtraction are possible. For that reason, for interval data, the central tendency can be measured by mean, median, or mode. Standard deviation can also be calculated.

However, interval data do not have something called a 'true zero' value. For example, there is nothing called '0 temperature' or 'no temperature'. Hence, only addition and subtraction applies for interval data. The ratio cannot be applied. This means, we can say a temperature of 40°C is equal to the temperature of 20°C + temperature of 20°C. However, we cannot say the temperature of 40°C means it is twice as hot as in temperature of 20°C.

Ratio data represents numeric data for which exact value can be measured. Absolute zero is available for ratio data. Also, these variables can be added, subtracted, multiplied, or divided. The central tendency can be measured by mean, median, or mode and methods of dispersion such as standard deviation. Examples of ratio data include height, weight, age, salary, etc.

Figure 2.4 gives a summarized view of different types of data that we may find in a typical machine learning problem.

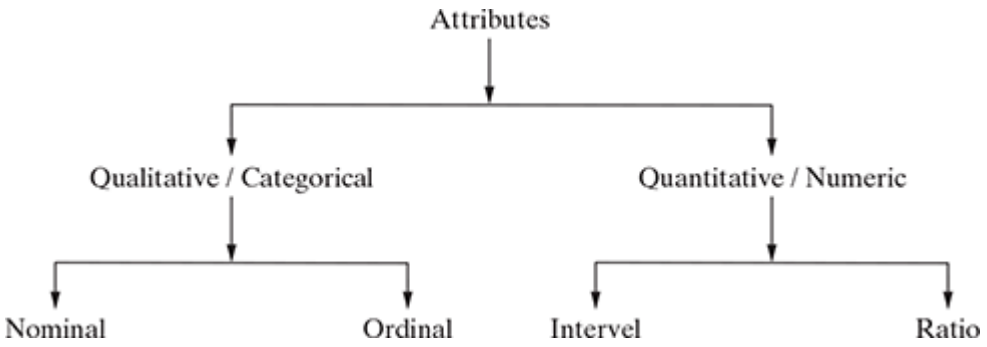

FIG. 2.4 Types of data

Apart from the approach detailed above, attributes can also be categorized into types based on a number of values that can be assigned. The attributes can be either discrete or continuous based on this factor.

Discrete attributes can assume a finite or countably infinite number of values. Nominal attributes such as roll number, street number, pin code, etc. can have a finite number of values whereas numeric attributes such as count, rank of students, etc. can have countably infinite values. A special type of discrete attribute which can assume two values only is called binary attribute. Examples of binary attribute include male/ female, positive/negative, yes/no, etc.

Continuous attributes can assume any possible value which is a real number. Examples of continuous attribute include length, height, weight, price, etc.

Note:

In general, nominal and ordinal attributes are discrete. On the other hand, interval and ratio attributes are continuous, barring a few exceptions, e.g. 'count' attribute.

2.4 EXPLORING STRUCTURE OF DATA

By now, we understand that in machine learning, we come across two basic data types – numeric and categorical. With this context in mind, we can delve deeper into understanding a data set. We need to understand that in a data set, which of the attributes are numeric and which are categorical in nature. This is because, the approach of exploring numeric data is different than the approach of exploring categorical data. In case of a standard data set, we may have the data dictionary available for reference. Data dictionary is a metadata repository, i.e. the repository of all information related to the structure of each data element contained in the data set. The data dictionary gives detailed information on each of the attributes – the description as well as the data type and other relevant details. In case the data dictionary is not available, we need to use standard library function of the machine learning tool that we are using and get the details. For the time being, let us move ahead with a standard data set from UCI machine learning repository.

Did you know?

University of California, Irvine (UCI) Machine Learning Repository (http://archive.ics.uci.edu/ml/index.php) is a collection of 400+ data sets which serve as benchmarks for researchers and practitioners in the machine learning community.

The data set that we take as a reference is the Auto MPG data set available in the UCI repository. Figure 2.5 is a snapshot of the first few rows of the data set.

mpg	cylinder	displace-ment	horse-power	weight	accel-eration	model year	origin	car name
18	8	307	130	3504	12	70	1	Chevrolet chev-elle malibu
15	8	350	165	3693	11.5	70	1	Buick skylark 320
18	8	318	150	3436	11	70	1	Plymouth satellite
16	8	304	150	3433	12	70	1	Amc rebel sst
17	8	302	140	3449	10.5	70	1	Ford torino
15	8	429	198	4341	10	70	1	Ford galaxie 500
14	8	454	220	4354	9	70	1	Chevrolet impala
14	8	440	215	4312	8.5	70	1	Plymouth fury iii
14	8	455	225	4425	10	70	1	Pontiac catalina
15	8	390	190	3850	8.5	70	1	Amc acbassador dpl
15	8	383	170	3563	10	70	1	Dodge challenger se
14	8	340	160	3609	8	70	1	Plymouth ' cuda 340
15	8	400	150	3761	9.5	70	1	Chevrolet monte carlo
14	8	455	225	3086	10	70	1	Buick estate wagon (sw)
24	4	113	95	2372	15	70	3	Toyota corona mark ii
22	6	198	95	2933	15.5	70	1	Plymouth duster
18	6	199	97	2774	15.5	70	1	Amc hornet

FIG. 2.5 Auto MPG data set

As is quite evident from the data, the attributes such as ‘mpg’, ‘cylinders’, ‘displacement’, ‘horsepower’, ‘weight’, ‘acceleration’, ‘model year’, and ‘origin’ are all numeric. Out of these attributes, ‘cylinders’, ‘model year’, and ‘origin’ are discrete in nature as the only finite number of values can be assumed by these attributes. The remaining of the numeric attributes, i.e. ‘mpg’, ‘displacement’, ‘horsepower’, ‘weight’, and ‘acceleration’ can assume any real value.

Note:

Since the attributes ‘cylinders’ or ‘origin’ have a small number of possible values, one may prefer to treat it as a categorical or qualitative attribute and explore in that way. Anyways, we will treat these attributes as numeric or quantitative as we are trying to show data exploration and related nuances in this section.

Hence, these attributes are continuous in nature. The only remaining attribute ‘car name’ is of type categorical, or more specifically nominal. This data set is regarding prediction of fuel consumption in miles per gallon, i.e. the numeric attribute ‘mpg’ is the target attribute.

With this understanding of the data set attributes, we can start exploring the numeric and categorical attributes separately.

2.4.1 Exploring numerical data

There are two most effective mathematical plots to explore numerical data – box plot and histogram. We will explore all these plots one by one, starting with the most critical one, which is the box plot.

2.4.1.1 Understanding central tendency

To understand the nature of numeric variables, we can apply the measures of central tendency of data, i.e. mean and median. In statistics, measures of central tendency help us understand the central point of a set of data. Mean, by definition, is a sum of all data values divided by the count of data elements. For example, mean of a set of observations – 21, 89, 34, 67, and 96 is calculated as below.

$$\text{Mean} = \frac{21 + 89 + 34 + 67 + 96}{5} = 61.4.$$

If the above set of numbers represents marks of 5 students in a class, the mean marks, or the falling in the middle of the range is 61.4.

Median, on contrary, is the value of the element appearing in the middle of an ordered list of data elements. If we consider the above 5 data elements, the ordered list would be – 21, 34, 67, 89, and 96. Since there are 5 data elements, the 3rd element in the ordered list is considered as the median. Hence, the median value of this set of data is 67.

There might be a natural curiosity to understand why two measures of central tendency are reviewed. The reason is mean and median are impacted differently by data values appearing at the beginning or at the end of the range. Mean being calculated from the cumulative sum of data values, is impacted if too many data elements are having values closer to the far end of the range, i.e. close to the maximum or minimum values. It is especially sensitive to outliers, i.e. the values which are unusually high or low, compared to the other values. Mean is likely to get shifted drastically even due to the presence of a small number of outliers. If we observe that for certain attributes the deviation between values of mean and median are quite high, we should investigate those attributes further and try to find out the root cause along with the need for remediation.

So, in the context of the Auto MPG data set, let's try to find out for each of the numeric attributes the values of mean and median. We can also find out if the deviation between these values is large. In Figure 2.6, the comparison between mean and median for all the attributes has been shown. We can see that for the attributes such as 'mpg', 'weight', 'acceleration', and 'model.year' the deviation between mean and median is not significant which means the chance of these attributes having too many outlier values is less. However, the deviation is significant for the attributes 'cylinders', 'displacement' and 'origin'. So, we need to further drill down and look at some more statistics for these attributes. Also, there is some problem in the values of the attribute 'horsepower' because of which the mean and median calculation is not possible.

	mpg	cylinders	displacement	horsepower	weight	acceleration	model year	origin
Median	23	4	148.5	?	2804	15.5	76	1
Mean	23.51	5.455	193.4	?	2970	15.57	76.01	1.573
Deviation	2.17	26.67%	23.22%		5.59%	0.45%	0.01%	36.43%
	Low	High	High		Low	Low	Low	High

FIG. 2.6 Mean vs. Median for Auto MPG

With a bit of investigation, we can find out that the problem is occurring because of the 6 data elements, as shown in Figure 2.7, do not have value for the attribute 'horsepower'.

mpg	cylinders	displacement	horsepower	weight	acceleration	model year	origin	car name
25	4	98	?	2046	19	71	1	Ford pinto
21	6	200	?	2875	17	74	1	Ford maverick
40.9	4	85	?	1835	17.3	80	2	Renault lecar deluxe
23.6	4	140	?	2905	14.3	80	1	Ford mustang cobra
34.5	4	100	?	2320	15.8	81	2	Renault 18i
23	4	151	?	3035	20.5	82	1	Amc concord di

FIG. 2.7 Missing values of attribute 'horsepower' in Auto MPG

For that reason, the attribute 'horsepower' is not treated as a numeric. That's why the operations applicable on numeric variables, like mean or median, are failing. So we have to first remediate the missing values of the attribute 'horsepower' before being able to do any kind of exploration. However, we will cover the approach of remediation of missing values a little later.

2.4.1.2 Understanding data spread

Now that we have explored the central tendency of the different numeric attributes, we have a clear idea of which attributes have a large deviation between mean and median. Let's look closely at those attributes. To drill down more, we need to look at the entire range of values of the attributes, though not at the level of data elements as that may be too vast to review manually. So we will take a granular view of the data spread in the form of

1. Dispersion of data
2. Position of the different data values

2.4.1.2.1 Measuring data dispersion

Consider the data values of two attributes

1. Attribute 1 values : 44, 46, 48, 45, and 47
2. Attribute 2 values : 34, 46, 59, 39, and 52

Both the set of values have a mean and median of 46. However, the first set of values that is of attribute 1 is more concentrated or clustered around the mean/median value whereas the second set of values of attribute 2 is quite spread out or dispersed. To measure the extent of dispersion of a data, or to find out how much the different values of a data are spread out, the variance of the data is measured. The variance of a data is measured using the formula given below:

Variance $(x) = \frac{\sum_{i=1}^{n} x_i^2}{n} - \left(\frac{\sum_{i=1}^{n} x_i}{n}\right)^2$, where x is the variable or attribute whose variance is to be measured and n is the number of observations or values of variable x.

Standard deviation of a data is measured as follows:

$$\text{Standard deviation } (x) = \sqrt{\text{Variance } (x)}$$

Larger value of variance or standard deviation indicates more dispersion in the data and vice versa. In the above example, let's calculate the variance of attribute 1 and that of attribute 2. For attribute 1,

$$\text{Variance} = \frac{\sum_{i=1}^{n} x_i^2}{n} - \left(\frac{\sum_{i=1}^{n} x_i}{n}\right)^2$$

$$= \frac{44^2 + 46^2 + 48^2 + 45^2 + 47^2}{5} - \left(\frac{44 + 46 + 48 + 45 + 47}{5}\right)^2$$

$$= \frac{1936 + 2116 + 2304 + 2025 + 2209}{5} - \left(\frac{230}{5}\right)^2 = \frac{10590}{5} - (46)^2 = 2$$

For attribute 2,

$$\text{Variance} = \frac{\sum_{i=1}^{n} x_i^2}{n} - \left(\frac{\sum_{i=1}^{n} x_i}{n}\right)^2$$

$$= \frac{34^2 + 46^2 + 59^2 + 39^2 + 52^2}{5} - \left(\frac{34 + 46 + 59 + 39 + 52}{5}\right)^2$$

$$= \frac{1156 + 2116 + 3481 + 1521 + 2704}{5} - \left(\frac{230}{5}\right)^2 = \frac{10978}{5} - (46)^2 = 79.6$$

So it is quite clear from the measure that attribute 1 values are quite concentrated around the mean while attribute 2 values are extremely spread out. Since this data was small, a visual inspection and understanding were possible and that matches with the measured value.

2.4.1.2.2 Measuring data value position

When the data values of an attribute are arranged in an increasing order, we have seen earlier that median gives the central data value, which divides the entire data set into two halves. Similarly, if the first half of the data is divided into two halves so that each half consists of one-quarter of the data set, then that median of the first half is known as first quartile or Q_1. In the same way, if the second half of the data is divided into two halves, then that median of the second half is known as third quartile or Q_3. The overall median is also known as second quartile or Q_2. So, any

data set has five values - minimum, first quartile (Q1), median (Q2), third quartile (Q3), and maximum.

Let's review these values for the attributes 'cylinders', 'displacement', and 'origin'. Figure 2.8 captures a summary of the range of statistics for the attributes. If we take the example of the attribute 'displacement', we can see that the difference between minimum value and Q1 is 36.2 and the difference between Q1 and median is 44.3. On the contrary, the difference between median and Q3 is 113.5 and Q3 and the maximum value is 193. In other words, the larger values are more spread out than the smaller ones. This helps in understanding why the value of mean is much higher than that of the median for the attribute 'displacement'. Similarly, in case of attribute 'cylinders', we can observe that the difference between minimum value and median is 1 whereas the difference between median and the maximum value is 4. For the attribute 'origin', the difference between minimum value and median is 0 whereas the difference between median and the maximum value is 2.

	cylinders	displacement	origin
Minimum	3	68	1
Q1	4	104.2	1
Median	4	148.5	1
Q3	8	262	2
Maximum	8	455	3

FIG. 2.8 Attribute value drill-down for Auto MPG

Note:

Quantiles refer to specific points in a data set which divide the data set into equal parts or equally sized quantities. There are specific variants of quantile, the one dividing data set into four parts being termed as quartile. Another such popular variant is percentile, which divides the data set into 100 parts.

However, we still cannot ascertain whether there is any outlier present in the data. For that, we can better adopt some means to visualize the data. Box plot is an excellent visualization medium for numeric data.

2.4.2 Plotting and exploring numerical data

2.4.2.1 Box plots

Now that we have a fairly clear understanding of the data set attributes in terms of spread and central tendency, let's try to make an attempt to visualize the whole thing as a box-plot. A box plot is an extremely effective mechanism to get a one-shot view and understand the nature of the data. But before we get to review the box plot for different attributes of Auto MPG data set, let's first try to understand a box plot in general and the interpretation of different aspects in a box plot. As we can see in Figure 2.9, the box plot (also called box and whisker plot) gives a standard visualization of the five-number summary statistics of a data, namely minimum, first quartile (Q1), median (Q2), third quartile (Q3), and maximum. Below is a detailed interpretation of a box plot.

- The central rectangle or the box spans from first to third quartile (i.e. Q1 to Q3), thus giving the inter-quartile range (IQR).
- Median is given by the line or band within the box.
- The lower whisker extends up to 1.5 times of the inter-quartile range (or IQR) from the bottom of the box, i.e. the first quartile or Q1. However, the actual length of the lower whisker depends on the lowest data value that falls within (Q1 − 1.5 times of IQR). Let's try to understand this with an example. Say for a specific set of data, Q1 = 73, median = 76 and Q3 = 79. Hence, IQR will be 6 (i.e. Q3 – Q1).So, lower whisker can extend maximum till (Q1 – 1.5 × IQR) = 73 – 1.5 × 6 = 64. However, say there are lower range data values such as 70, 63, and 60. So, the lower whisker will come at 70 as this is the lowest data value larger than 64.
- The upper whisker extends up to 1.5 as times of the inter-quartile range (or IQR) from the top of the box, i.e. the third quartile or Q3.

Similar to lower whisker, the actual length of the upper whisker will also depend on the highest data value that falls within (Q3 + 1.5 times of IQR). Let's try to understand this with an example. For the same set of data mentioned in the above point, upper whisker can extend maximum till (Q3 + 1.5 × IQR) = 79 + 1.5 × 6 = 88. If there is higher range of data values like 82, 84, and 89. So, the upper whisker will come at 84 as this is the highest data value lower than 88.

- The data values coming beyond the lower or upper whiskers are the ones which are of unusually low or high values respectively. These are the outliers, which may deserve special consideration.

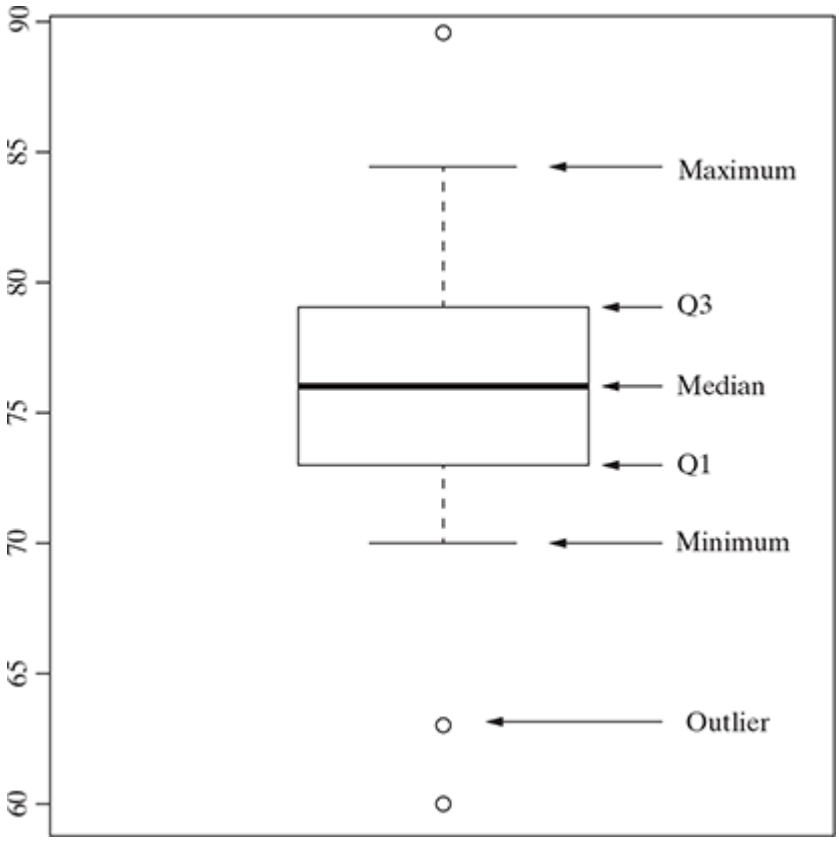

FIG. 2.9 Box plot

Note:

There are different variants of box plots. The one covered above is the Tukey box plot. Famous mathematician John W. Tukey introduced this type of box plot in 1969.

Let's visualize the box plot for the three attributes - 'cylinders', 'displacement', and 'origin'. We will also review the box plot of another attribute in which the deviation between mean and median is very little and see what the basic difference in the respective box plots is. Figure 2.10 presents the respective box plots.

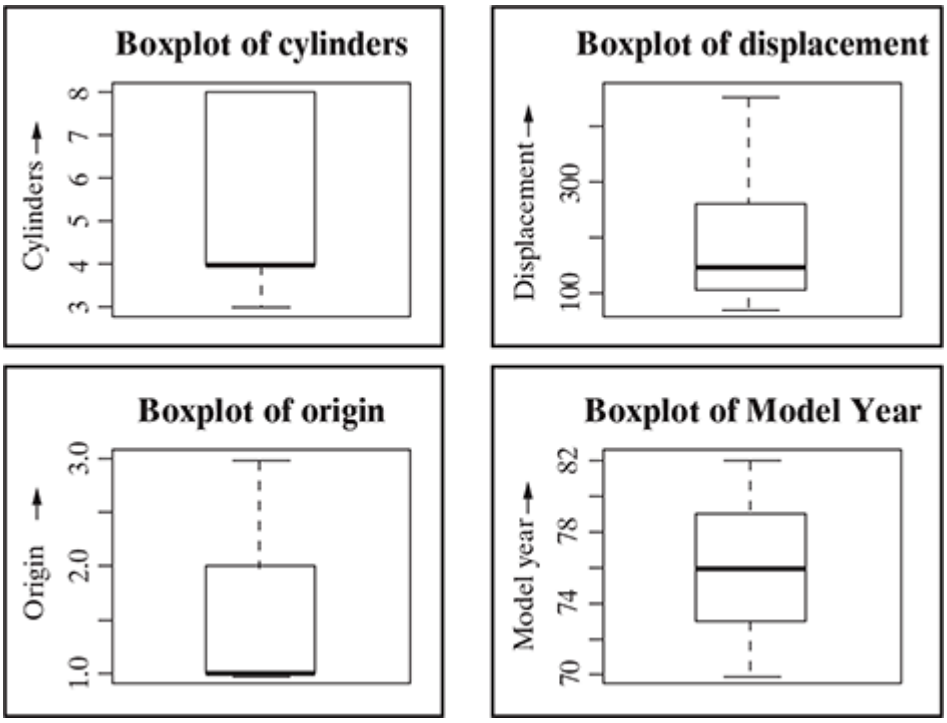

FIG. 2.10 Box plot of Auto MPG attributes

2.4.2.1.1 Analysing box plot for 'cylinders'

The box plot for attribute 'cylinders' looks pretty weird in shape. The upper whisker is missing, the band for median falls at the bottom of the box, even the lower whisker is pretty small compared to the length of the box! Is everything right?

The answer is a big YES, and you can figure it out if you delve a little deeper into the actual data values of the attribute. The attribute 'cylinders' is discrete in nature having values from 3 to 8. Table 2.2 captures the frequency and cumulative frequency of it.

Table 2.2 *Frequency of "Cylinders" Attribute*

Cylinders	Frequency	Cumulative Frequency
3	4	4
4	204	208 (= 4 + 204)
5	3	211 (= 208 + 3)
6	84	295 (= 211 + 84)
7	0	295 (= 295 + 0)
8	103	398 (= 295 + 103)

As can be observed in the table, the frequency is extremely high for data value 4. Two other data values where the frequency is quite high are 6 and 8. So now if we try to find the quartiles, since the total frequency is 398, the first quartile (Q1), median (Q2), and third quartile (Q3) will be at a cumulative frequency 99.5 (i.e. average of 99th and 100th observation), 199 and 298.5 (i.e. average of 298th and 299th observation), respectively. This way Q1 = 4, median = 4 and Q3 = 8. Since there is no data value beyond 8, there is no upper whisker. Also, since both Q1 and median are 4, the band for median falls on the bottom of the box. Same way, though the lower whisker could have extended till -2 (Q1 – 1.5 × IQR = 4 – 1.5 × 4 = –2), in reality, there is no data value lower than 3. Hence, the lower whisker is also short. In any case, a value of cylinders less than 1 is not possible.

2.4.2.1.2 Analysing box plot for 'origin'

Like the box plot for attribute 'cylinders', the box plot for attribute 'cylinders' also looks pretty weird in shape. Here the lower whisker is missing and the band for median falls at the bottom of the box! Let's verify if everything right?

Just like the attribute 'cylinders', attribute 'origin' is discrete in nature having values from 1 to 3. Table 2.3 captures the frequency and cumulative frequency (i.e. a summation of frequencies of all previous intervals) of it.

Table 2.3 *Frequency of "Origin" Attribute*

origin	Frequency	Cumulative Frequency
1	249	249
2	70	319 (= 249 + 70)
3	79	398 (= 319 + 79)

As can be observed in the table, the frequency is extremely high for data value 1. Since the total frequency is 398, the first quartile (Q1), median (Q2), and third quartile (Q3) will be at a cumulative frequency 99.5 (i.e. average of 99th and 100th observation), 199 and 298.5 (i.e. average of 298th and 299th observation), respectively. This way Q1 = 1, median = 1, and Q3 = 2. Since Q1 and median are same in value, the band for median falls on the bottom of the box. There is no data value lower than Q1. Hence, the lower whisker is missing.

2.4.2.1.3 Analysing box plot for 'displacement'

The box plot for the attribute 'displacement' looks better than the previous box plots. However, still, there are few small abnormalities, the cause of which needs to be reviewed. Firstly, the lower whisker is much smaller than an upper whisker. Also, the band for median is closer to the bottom of the box.

Let's take a closer look at the summary data of the attribute 'displacement'. The value of first quartile, Q1 = 104.2, median = 148.5, and third quartile, Q3 = 262. Since (median – Q1) = 44.3 is greater than (Q3 – median) = 113.5, the band for the median is closer to the bottom of the box (which represents Q1). The value of IQR, in this case, is 157.8. So the lower whisker can be 1.5 times 157.8 less than Q1. But minimum data value for the attribute 'displacement' is 68. So, the lower whisker at 15% [(Q1 – minimum)/1.5 × IQR = (104.2 – 68) / (1.5 × 157.8) = 15%] of the permissible length. On the other hand, the maximum data value is 455. So the upper whisker is 81% [(maximum – Q3)/1.5 × IQR = (455 – 262) / (1.5 × 157.8) =

81%] of the permissible length. This is why the upper whisker is much longer than the lower whisker.

2.4.2.1.4 Analysing box plot for 'model Year'

The box plot for the attribute 'model. year' looks perfect. Let's validate is it really what expected to be.

For the attribute 'model.year':

First quartile, Q1 = 73

Median, Q2 = 76

Third quartile, Q3 = 79

So, the difference between median and Q1 is exactly equal to Q3 and median (both are 3). That is why the band for the median is exactly equidistant from the bottom and top of the box.

I QR = Q3 9 Q1 = 7 9 - 7 3 = 6

Difference between Q1 and minimum data value (i.e. 70) is also same as maximum data value (i.e. 82) and Q3 (both are 3). So both lower and upper whiskers are expected to be of the same size which is 33% [3 / (1.5 × 6)] of the permissible length.

2.4.2.2 Histogram

Histogram is another plot which helps in effective visualization of numeric attributes. It helps in understanding the distribution of a numeric data into series of intervals, also termed as 'bins'. The important difference between histogram and box plot is

- The focus of histogram is to plot ranges of data values (acting as 'bins'), the number of data elements in each range will depend on the data

distribution. Based on that, the size of each bar corresponding to the different ranges will vary.

- The focus of box plot is to divide the data elements in a data set into four equal portions, such that each portion contains an equal number of data elements.

Histograms might be of different shapes depending on the nature of the data, e.g. skewness. Figure 2.11 provides a depiction of different shapes of the histogram that are generally created. These patterns give us a quick understanding of the data and thus act as a great data exploration tool.

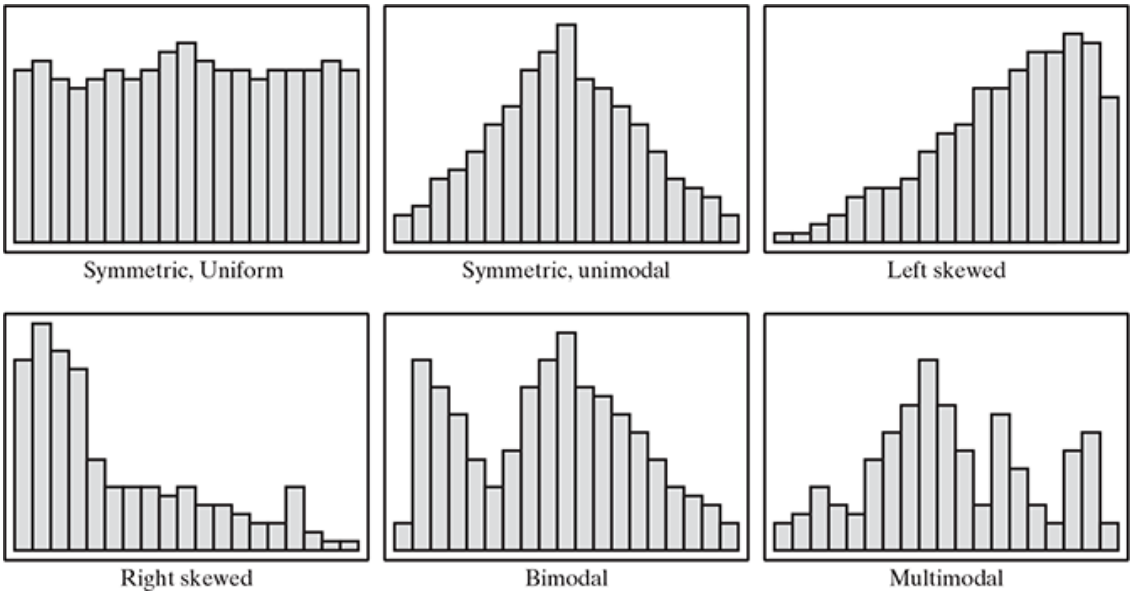

FIG. 2.11 General Histogram shapes

Let's now examine the histograms for the different attributes of Auto MPG data set presented in Figure 2.12. The histograms for 'mpg' and 'weight' are right-skewed. The histogram for 'acceleration' is symmetric and unimodal, whereas the one for 'model.year' is symmetric and uniform. For the remaining attributes, histograms are multimodal in nature.

Now let's dig deep into one of the histograms, say the one for the attribute 'acceleration'. The histogram is composed of a number of bars, one bar appearing for each of the 'bins'. The height of the bar reflects the total count of data elements whose value falls within the specific bin value, or

the frequency. Talking in context of the histogram for acceleration, each ‘bin’ represents an acceleration value interval of 2 units. So the second bin, e.g., reflects acceleration value of 10 to 12 units. The corresponding bar chart height reflects the count of all data elements whose value lies between 10 and 12 units. Also, it is evident from the histogram that it spans over the acceleration value of 8 to 26 units. The frequency of data elements corresponding to the bins first keep on increasing, till it reaches the bin of range 14 to 16 units. At this range, the bar is tallest in size. So we can conclude that a maximum number of data elements fall within this range. After this range, the bar size starts decreasing till the end of the whole range at the acceleration value of 26 units.

Please note that when the histogram is uniform, as in the case of attribute ‘model. year’, it gives a hint that all values are equally likely to occur.

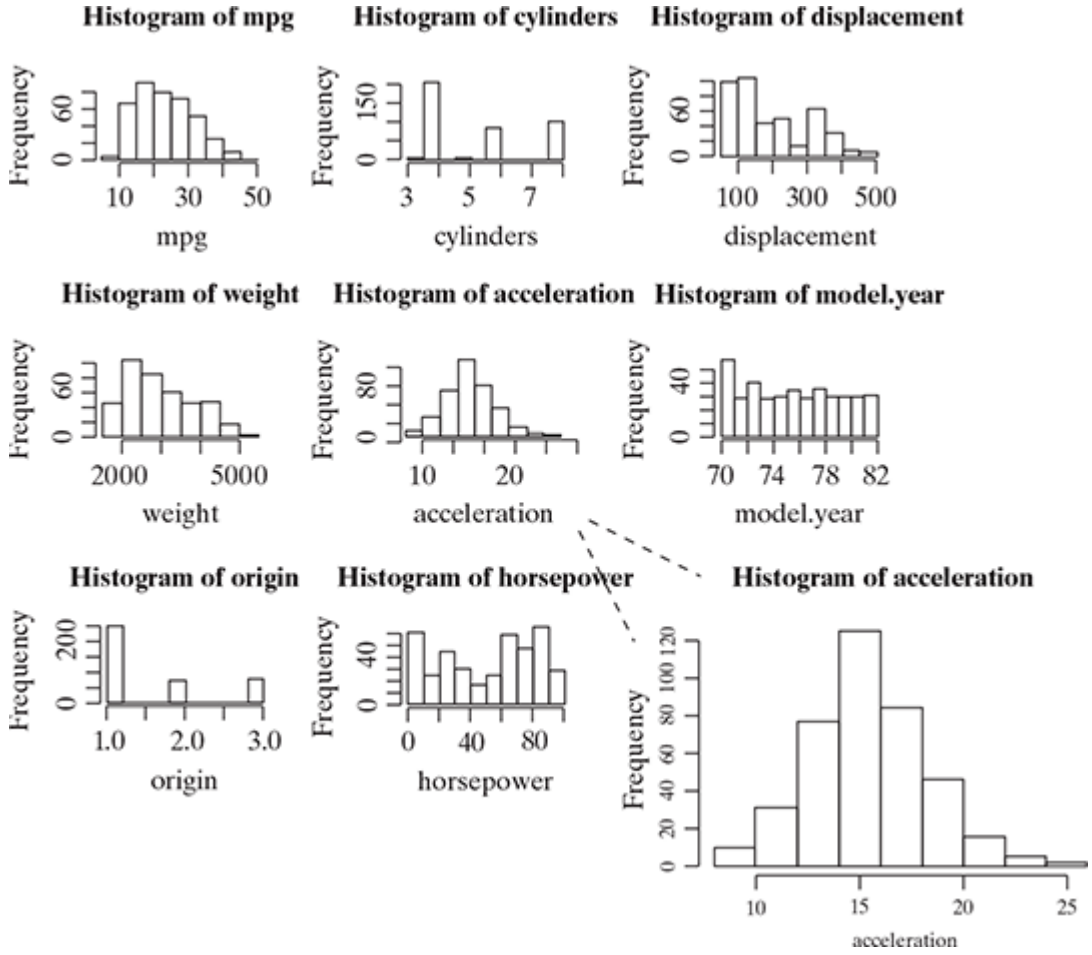

FIG. 2.12 Histogram Auto MPG attributes

2.4.3 Exploring categorical data

We have seen there are multiple ways to explore numeric data. However, there are not many options for exploring categorical data. In the Auto MPG data set, attribute 'car.name' is categorical in nature. Also, as we discussed earlier, we may consider 'cylinders' as a categorical variable instead of a numeric variable.

The first summary which we may be interested in noting is how many unique names are there for the attribute 'car name' or how many unique values are there for 'cylinders' attribute. We can get this as follows:

For attribute 'car name'

1. Chevrolet chevelle malibu

2. Buick skylark 320
3. Plymouth satellite
4. Amc rebel sst
5. Ford torino
6. Ford galaxie 500
7. Chevrolet impala
8. Plymouth fury iii
9. Pontiac catalina
10. Amc ambassador dpl

For attribute ‘cylinders’

8 4 6 3 5

We may also look for a little more details and want to get a table consisting the categories of the attribute and count of the data elements falling into that category. Tables 2.4 and 2.5 contain these details.

For attribute ‘car name’

Table 2.4 *Count of Categories for ‘car name’ Attribute*

Attribute Value	amc ambassador brougham	amc ambassador dpl	amc ambassador sst	amc concord	amc concord d/l	amc concord dl 6	amc gremlin	…
Count	1	1	1	1	2	2	4	…

For attribute “cylinders”

Table 2.5 *Count of Categories for 'Cylinders' Attribute*

Attribute Value	3	4	5	6	8
Count	4	204	3	84	103

In the same way, we may also be interested to know the proportion (or percentage) of count of data elements belonging to a category. Say, e.g., for the attributes 'cylinders', the proportion of data elements belonging to the category 4 is 204 ÷ 398 = 0.513, i.e. 51.3%. Tables 2.6 and 2.7 contain the summarization of the categorical attributes by proportion of data elements.

For attribute 'car name'

Table 2.6 *Proportion of Categories for '"Cylinders' Attribute*

Attribute Value	Amc ambas-sador brougham	Amc ambassa-dor dpl	Amc ambassa-dor sst	Amc concord	Amc concord d/l	Amc concord dl 6	Amc gremlin	…
Count	0.003	0.003	0.003	0.003	0.005	0.005	0.01	…

For attribute 'cylinders'

Table 2.7 *Proportion of Categories for "Cylinders" Attribute*

Attribute Value	3	4	5	6	8
Count	0.01	0.513	0.008	0.211	0.259

Last but not the least, as we have read in the earlier section on types of data, statistical measure "mode" is applicable on categorical attributes. As we know, like mean and median, mode is also a statistical measure for

central tendency of a data. Mode of a data is the data value which appears most often. In context of categorical attribute, it is the category which has highest number of data values. Since mean and median cannot be applied for categorical variables, mode is the sole measure of central tendency.

Let's try to find out the mode for the attributes 'car name' and 'cylinders'. For cylinders, since the number of categories is less and we have the entire table listed above, we can see that the mode is 4, as that is the data value for which frequency is highest. More than 50% of data elements belong to the category 4. However, it is not so evident for the attribute 'car name' from the information given above. When we probe and try to find the mode, it is found to be category 'ford pinto' for which frequency is of highest value 6.

An attribute may have one or more modes. Frequency distribution of an attribute having single mode is called 'unimodal', two modes are called 'bimodal' and multiple modes are called 'multimodal'.

2.4.4 Exploring relationship between variables

Till now we have been exploring single attributes in isolation. One more important angle of data exploration is to explore relationship between attributes. There are multiple plots to enable us explore the relationship between variables. The basic and most commonly used plot is scatter plot.

2.4.4.1 Scatter plot

A scatter plot helps in visualizing bivariate relationships, i.e. relationship between two variables. It is a two-dimensional plot in which points or dots are drawn on coordinates provided by values of the attributes. For example, in a data set there are two attributes – attr_1 and attr_2. We want to understand the relationship between two attributes, i.e. with a change in value of one attribute, say attr_1, how does the value of the other attribute, say attr_2, changes. We can draw a scatter plot, with attr_1 mapped to *x*-axis and attr_2 mapped in *y*-axis. So, every point in the

plot will have value of attr_1 in the *x*-coordinate and value of attr_2 in the *y*-coordinate. As in a two-dimensional plot, attr_1 is said to be the independent variable and attr_2 as the dependent variable.

Let's take a real example in this context. In the data set Auto MPG, there is expected to be some relation between the attributes 'displacement' and 'mpg'. Let's try to verify our intuition using the scatter plot of 'displacement' and 'mpg'. Let's map 'displacement' as the *x*-coordinate and 'mpg' as the *y*-coordinate. The scatter plot comes as in Figure 2.13.

As is evident in the scatter plot, there is a definite relation between the two variables. The value of 'mpg' seems to steadily decrease with the increase in the value of 'displacement'. It may come in our mind that what is the extent of relationship? Well, it can be reviewed by calculating the correlation between the variables. Refer to chapter 5 if you want to find more about correlation and how to calculate it. One more interesting fact to notice is that there are certain data values which stand-out of the others. For example, there is one data element which has a mpg of 37 for a displacement of 250. This record is completely different from other data elements having similar displacement value but mpg value in the range of 15 to 25. This gives an indication that of presence of outlier data values.

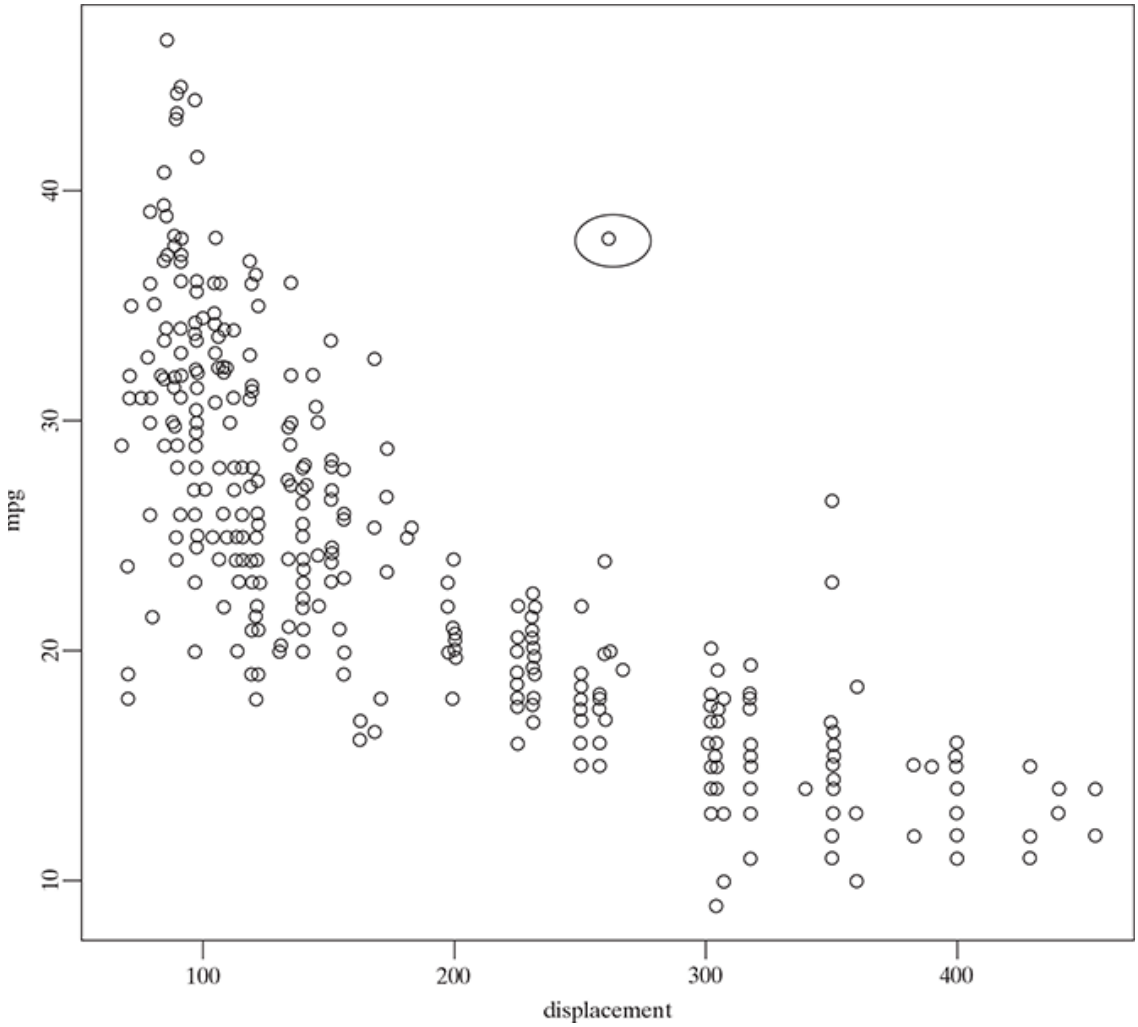

FIG. 2.13 Scatter plot of 'displacement' and 'mpg'

In Figure 2.14, the pair wise relationship among the features – 'mpg', 'displacement', 'horsepower', 'weight', and 'acceleration' have been captured. As you can see, in most of the cases, there is a significant relationship between the attribute pairs. However, in some cases, e.g. between attributes 'weight' and 'acceleration', the relationship doesn't seem to be very strong.

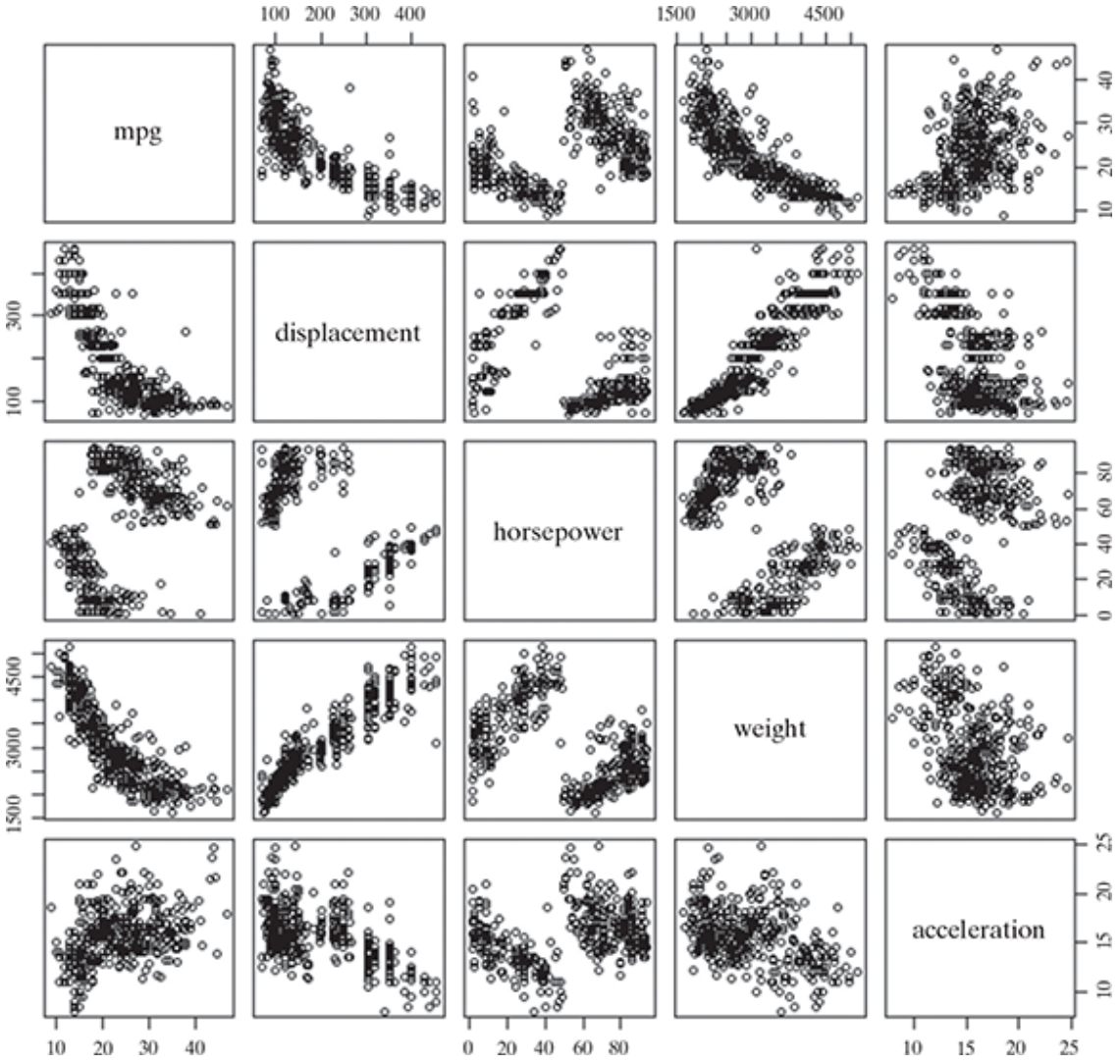

FIG. 2.14 Pair wise scatter plot between different attributes of Auto MPG

2.4.4.2 Two-way cross-tabulations

Two-way cross-tabulations (also called cross-tab or contingency table) are used to understand the relationship of two categorical attributes in a concise way. It has a matrix format that presents a summarized view of the bivariate frequency distribution. A cross-tab, very much like a scatter plot, helps to understand how much the data values of one attribute changes with the change in data values of another attribute. Let's try to see with examples, in context of the Auto MPG data set.

Let's assume the attributes 'cylinders', 'model.year', and 'origin' as categorical and try to examine the variation of one with respect to the other.

As we understand, attribute 'cylinders' reflects the number of cylinders in a car and assumes values 3, 4, 5, 6, and 8. Attribute 'model.year' captures the model year of each of the car and 'origin' gives the region of the car, the values for origin 1, 2, and 3 corresponding to North America, Europe, and Asia. Below are the cross-tabs. Let's try to understand what information they actually provide.

The first cross-tab, i.e. the one showing relationship between attributes 'model. year' and 'origin' help us understand the number of vehicles per year in each of the regions North America, Europe, and Asia. Looking at it in another way, we can get the count of vehicles per region over the different years. All these are in the context of the sample data given in the Auto MPG data set.

Moving to the second cross-tab, it gives the number of 3, 4, 5, 6, or 8 cylinder cars in every region present in the sample data set. The last cross-tab presents the number of 3, 4, 5, 6, or 8 cylinder cars every year.

We may also want to create cross-tabs with a more summarized view like have a cross-tab giving a number of cars having 4 or less cylinders and more than 4 cylinders in each region or by the years. This can be done by rolling up data values by the attribute 'cylinder'. Tables 2.8–2.10 present cross-tabs for different attribute combinations.

'Model year' vs. 'origin'

Table 2.8 *Cross-tab for* 'Model year' *vs.* 'Origin'

Origin \ Model Year	70	71	72	73	74	75	76	77	78	79	80	81	82
1	22	20	18	29	15	20	22	18	22	23	7	13	20
2	5	4	5	7	6	6	8	4	6	4	9	4	2
3	2	4	5	4	6	4	4	6	8	2	13	12	9

'Cylinders' vs. 'Origin'

Table 2.9 *Cross-tab for 'Cylinders' vs. 'Origin'*

Cylinders \ Origin	1	2	3
3	0	0	4
4	72	63	69
5	0	3	0
6	74	4	6
8	103	0	0

'Cylinders' vs. 'Model year'

Table 2.10 *Cross-tab for 'Cylinders' vs. 'Model year'*

Cylinders \ Model Year	70	71	72	73	74	75	76	77	78	79	80	81	82
3	0	0	1	1	0	0	0	1	0	0	1	0	0
4	7	13	14	11	15	12	15	14	17	12	25	21	28
5	0	0	0	0	0	0	0	0	1	1	1	0	0
6	4	8	0	8	7	12	10	5	12	6	2	7	3
8	18	7	13	20	5	6	9	8	6	10	0	1	0

2.5 DATA QUALITY AND REMEDIATION

2.5.1 Data quality

Success of machine learning depends largely on the quality of data. A data which has the right quality helps to achieve better prediction accuracy, in case of supervised learning. However, it is not realistic to expect that the data will be flawless. We have already come across at least two types of problems:

1. Certain data elements without a value or data with a missing value.
2. Data elements having value surprisingly different from the other elements, which we term as outliers.

There are multiple factors which lead to these data quality issues. Following are some of them:

- **Incorrect sample set selection:** The data may not reflect normal or regular quality due to incorrect selection of sample set. For example, if we are selecting a sample set of sales transactions from a festive period and trying to use that data to predict sales in future. In this case, the prediction will be far apart from the actual scenario, just because the sample set has been selected in a wrong time. Similarly, if we are trying to predict poll results using a training data which doesn't comprise of a right mix of voters from different segments such as age, sex, ethnic diversities, etc., the prediction is bound to be a failure. It may also happen due to incorrect sample size. For example, a sample of small size may not be able to capture all aspects or information needed for right learning of the model.
- **Errors in data collection:** resulting in outliers and missing values
  - In many cases, a person or group of persons are responsible for the collection of data to be used in a learning activity. In this manual process, there is the possibility of wrongly recording data either in terms of value (say 20.67 is wrongly recorded as 206.7 or 2.067) or in terms of a unit of measurement (say cm. is wrongly recorded as m. or mm.). This may result in data elements which have abnormally high or low value from other elements. Such records are termed as *outliers*.
  - It may also happen that the data is not recorded at all. In case of a survey conducted to collect data, it is all the more possible as survey responders may choose not to respond to a certain question. So the data value for that data element in that responder's record is *missing*.

2.5.2 Data remediation

The issues in data quality, as mentioned above, need to be remediated, if the right amount of efficiency has to be achieved in the learning activity. Out of the two major areas mentioned above, the first one can be reme-

died by proper sampling technique. This is a completely different area – covered as a specialized subject area in statistics. We will not cover that in this book. However, human errors are bound to happen, no matter whatever checks and balances we put in. Hence, proper remedial steps need to be taken for the second area mentioned above. We will discuss how to handle outliers and missing values.

2.5.2.1 Handling outliers

Outliers are data elements with an abnormally high value which may impact prediction accuracy, especially in regression models. Once the outliers are identified and the decision has been taken to amend those values, you may consider one of the following approaches. However, if the outliers are natural, i.e. the value of the data element is surprisingly high or low because of a valid reason, then we should not amend it.

- **Remove outliers:** If the number of records which are outliers is not many, a simple approach may be to remove them.
- **Imputation:** One other way is to impute the value with mean or median or mode. The value of the most similar data element may also be used for imputation.
- **Capping:** For values that lie outside the $1.5 \times$ IQR limits, we can cap them by replacing those observations below the lower limit with the value of 5th percentile and those that lie above the upper limit, with the value of 95th percentile.

If there is a significant number of outliers, they should be treated separately in the statistical model. In that case, the groups should be treated as two different groups, the model should be built for both groups and then the output can be combined.

2.5.2.2 Handling missing values

In a data set, one or more data elements may have missing values in multiple records. As discussed above, it can be caused by omission on part of the surveyor or a person who is collecting sample data or by the respon-

der, primarily due to his/her unwillingness to respond or lack of understanding needed to provide a response. It may happen that a specific question (based on which the value of a data element originates) is not applicable to a person or object with respect to which data is collected. There are multiple strategies to handle missing value of data elements. Some of those strategies have been discussed below.

2.5.2.2.1 Eliminate records having a missing value of data elements

In case the proportion of data elements having missing values is within a tolerable limit, a simple but effective approach is to remove the records having such data elements. This is possible if the quantum of data left after removing the data elements having missing values is sizeable.

In the case of Auto MPG data set, only in 6 out of 398 records, the value of attribute 'horsepower' is missing. If we get rid of those 6 records, we will still have 392 records, which is definitely a substantial number. So, we can very well eliminate the records and keep working with the remaining data set.

However, this will not be possible if the proportion of records having data elements with missing value is really high as that will reduce the power of model because of reduction in the training data size.

2.5.2.2.2 Imputing missing values

Imputation is a method to assign a value to the data elements having missing values. Mean/mode/median is most frequently assigned value. For quantitative attributes, all missing values are imputed with the mean, median, or mode of the remaining values under the same attribute. For qualitative attributes, all missing values are imputed by the mode of all remaining values of the same attribute. However, another strategy may be identify the similar types of observations whose values are known and use the mean/median/mode of those known values.

For example, in context of the attribute 'horsepower' of the Auto MPG data set, since the attribute is quantitative, we take a mean or median of the remaining data element values and assign that to all data elements having a missing value. So, we may assign the mean, which is 104.47 and assign it to all the six data elements. The other approach is that we can take a similarity based mean or median. If we refer to the six observations with missing values for attribute 'horsepower' as depicted in Table 2.11, 'cylinders' is the attribute which is logically most connected to 'horsepower' because with the increase in number of cylinders of a car, the horsepower of the car is expected to increase. So, for five observations, we can use the mean of data elements of the 'horsepower' attribute having cylinders = 4; i.e. 78.28 and for one observation which has cylinders = 6, we can use a similar mean of data elements with cylinders = 6, i.e. 101.5, to impute value to the missing data elements.

Table 2.11 *Missing Values for 'Horsepower' Attribute*

mpg	cylinders	displacement	horsepower	weight	acceleration	model year	origin	car name
25	4	98	?	2046	19	71	1	Ford pinto
21	6	200	?	2875	17	74	1	Ford maverick
40.9	4	85	?	1835	17.3	80	2	Renault lecar deluxe
23.6	4	140	?	2905	14.3	80	1	Ford mustang cobra
34.5	4	100	?	2320	15.8	81	2	Renault 18i
23	4	151	?	3035	20.5	82	1	Amc concord dl

2.5.2.2.3 Estimate missing values

If there are data points similar to the ones with missing attribute values, then the attribute values from those similar data points can be planted in place of the missing value. For finding similar data points or observations, distance function can be used.

For example, let's assume that the weight of a Russian student having age 12 years and height 5 ft. is missing. Then the weight of any other Russian student having age close to 12 years and height close to 5 ft. can be assigned.

2.6 DATA PRE-PROCESSING

2.6.1 Dimensionality reduction

Till the end of the 1990s, very few domains were explored which included data sets with a high number of attributes or features. In general, the data sets used in machine learning used to be in few 10s. However, in the last two decades, there has been a rapid advent of computational biology like genome projects. These projects have produced extremely high-dimensional data sets with 20,000 or more features being very common. Also, there has been a wide-spread adoption of social networking leading to a need for text classification for customer behaviour analysis.

High-dimensional data sets need a high amount of computational space and time. At the same time, not all features are useful – they degrade the performance of machine learning algorithms. Most of the machine learning algorithms perform better if the dimensionality of data set, i.e. the number of features in the data set, is reduced. Dimensionality reduction helps in reducing irrelevance and redundancy in features. Also, it is easier to understand a model if the number of features involved in the learning activity is less.

Dimensionality reduction refers to the techniques of reducing the dimensionality of a data set by creating new attributes by combining the original attributes. The most common approach for dimensionality reduction is known as Principal Component Analysis (PCA). PCA is a statistical technique to convert a set of correlated variables into a set of transformed, uncorrelated variables called principal components. The principal components are a linear combination of the original variables. They are orthogonal to each other. Since principal components are uncorrelated, they capture the maximum amount of variability in the data.

However, the only challenge is that the original attributes are lost due to the transformation.

Another commonly used technique which is used for dimensionality reduction is Singular Value Decomposition (SVD).

More about these concepts have been discussed in Chapter 4.

2.6.2 Feature subset selection

Feature subset selection or simply called feature selection, both for supervised as well as unsupervised learning, try to find out the optimal subset of the entire feature set which significantly reduces computational cost without any major impact on the learning accuracy. It may seem that a feature subset may lead to loss of useful information as certain features are going to be excluded from the final set of features used for learning. However, for elimination only features which are not relevant or redundant are selected.

A feature is considered as irrelevant if it plays an insignificant role (or contributes almost no information) in classifying or grouping together a set of data instances. All irrelevant features are eliminated while selecting the final feature subset. A feature is potentially redundant when the information contributed by the feature is more or less same as one or more other features. Among a group of potentially redundant features, a small number of features can be selected as a part of the final feature subset without causing any negative impact to learn model accuracy.

There are different ways to select a feature subset. In Chapter 4, we will be discussing feature selection in details.

2.7 SUMMARY

- A data set is a collection of related information or records.
- Data can be broadly divided into following two types
  - Qualitative data

    - Quantitative data
- Qualitative data provides information about the quality of an object or information which cannot be measured. Qualitative data can be further subdivided into two types as follows:
    - Nominal data: has named value
    - Ordinal data: has named value which can be naturally ordered
- Quantitative data relates to information about the quantity of an object – hence it can be measured. There are two types of quantitative data:
    - Interval data: numeric data for which the exact difference between values is known. However, such data do not have something called a 'true zero' value.
    - Ratio data: numeric data for which exact value can be measured and absolute zero is available.
- Measures of central tendency help to understand the central point of a set of data. Standard measures of central tendency of data are mean, median, and mode.
- Detailed view of the data spread is available in the form of
    - Dispersion of data:extent of dispersion of a data is measured by variance
    - Related to the position of the different data values there are five values:minimum, first quartile (Q1), median (Q2), third quartile (Q3), and maximum
- Exploration of numerical data can be best done using box plots and histograms.
- Options for exploration of categorical data are very limited.
- For exploring relations between variables, scatter-plots and two-way cross-tabulations can be effectively used.
- Success of machine learning depends largely on the quality of data. Two common types of data issue are:
    - Data with a missing value
    - Data values which are surprisingly different termed as outliers
- High-dimensional data sets need a high amount of computational space and time. Most of the machine learning algorithms perform better if the dimensionality of data set is reduced.

- Some popular dimensionality reduction techniques are PCA, SVD, and feature selection.

SAMPLE QUESTIONS

MULTIPLE-CHOICE QUESTIONS (1 MARK QUESTIONS) :

1. Temperature is a
    1. Interval data
    2. Ratio data
    3. Discrete data
    4. None of the above
2. Principal component is a technique for
    1. Feature selection
    2. Dimensionality reduction
    3. Exploration
    4. None of the above
3. For bi-variate data exploration, _____ is an effective tool.
    1. Box plot
    2. Two-way cross-tab
    3. Histogram
    4. None of the above
4. For box plot, the upper and lower whisker length depends on
    1. Median
    2. Mean
    3. IQR
    4. All of the above
5. Feature selection tries to eliminate features which are
    1. Rich
    2. Redundant
    3. Irrelevant
    4. Relevant
6. When the number of features increase
    1. Computation time increases
    2. Model becomes complex

3. Learning accuracy decreases
4. All of the above

7. For categorical data, ____ cannot be used as a measure of central tendency.
    1. Median
    2. Mean
    3. Quartile
    4. None of the above
8. For understanding relationship between two variables, ____ can be used.
    1. Box plot
    2. Scatter plot
    3. Histogram
    4. None of the above
9. Two common types of data issue are
    1. Outlier
    2. Missing value
    3. Boundary value
    4. None of the above
10. Exploration of numerical data can be best done using
    1. Boxplots
    2. Histograms
    3. Scatter plot
    4. None of the above
11. Data can broadly divided into following two types
    1. Qualitative
    2. Speculative
    3. Quantitative
    4. None of the above
12. Ordinal data can be naturally ____.
    1. Measured
    2. Ordered
    3. Divided
    4. None of the above

SHORT-ANSWER TYPE QUESTIONS (5 MARKS QUESTIONS):

1. What are the main activities involved when you are preparing to start with modelling in machine learning?
2. What are the basic data types in machine learning? Give an example of each one of them.
3. Differentiate:
    1. Categorical vs. Numeric attribute
    2. Dimensionality reduction vs. Feature selection
4. Write short notes on any two:
    1. Histogram
    2. Scatter plot
    3. PCA
5. Why do we need to explore data? Is there a difference in the way of exploring qualitative data vis-a-vis quantitative data?
6. What are different shapes of histogram? What are ‘bins’?
7. How can we take care of outliers in data?
8. What are the different measures of central tendency? Why do mean, in certain data sets, differ widely from median?
9. Explain how bivariate relationships can be explored using scatter plot. Can outliers be detected using scatter plot?
10. Explain how cross-tabs can be used to understand relationship between two variables.

LONG-ANSWER TYPE QUESTIONS (10 MARKS QUESTIONS) :

1. What are the main activities involved in machine learning? What is meant by data pre-processing?
2. Explain qualitative and quantitative data in details. Differentiate between the two.
3. Prepare a simple data set along with some sample records in it. Have at least one attribute of the different data types used in machine learning.
4. What are the different causes of data issues in machine learning? What are the fallouts?

5. Explain, with proper example, different ways of exploring categorical data.
6. When there are variables with certain values missing, will that impact the learning activity? If so, how can that be addressed?
7. Explain, in details, the different strategies of addressing missing data values.
8. What are the different techniques for data pre-processing? Explain, in brief, dimensionality reduction and feature selection.
9. 1. What is IQR? How is it measured?
   2. Explain, in details, the different components of a box plot? When will the lower whisker be longer than the upper whisker? How can outliers be detected using box plot?
10. 1. Write short notes on any two:
        1. Interval data
        2. Inter-quartile range
        3. Cross-tab
    2. Write the difference between (any two):
        1. Nominal and ordinal data
        2. Box plot and histogram
        3. Mean and median