2. Data Preparation for Machine Learning

This part 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 part helps a beginner take the first step towards effective modelling and solving a machine learning problem.

2.1 INTRODUCTION

We have 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 implies learning from past data, also called training data, which has 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.

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.

We also 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.

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 from 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.

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:

(1) Understand the type of data in the given input data set.

(2) Explore the data to understand the nature and quality.

(3) Explore the relationships amongst the data elements, e.g. inter-feature relationship.

(4) Find potential issues in data.

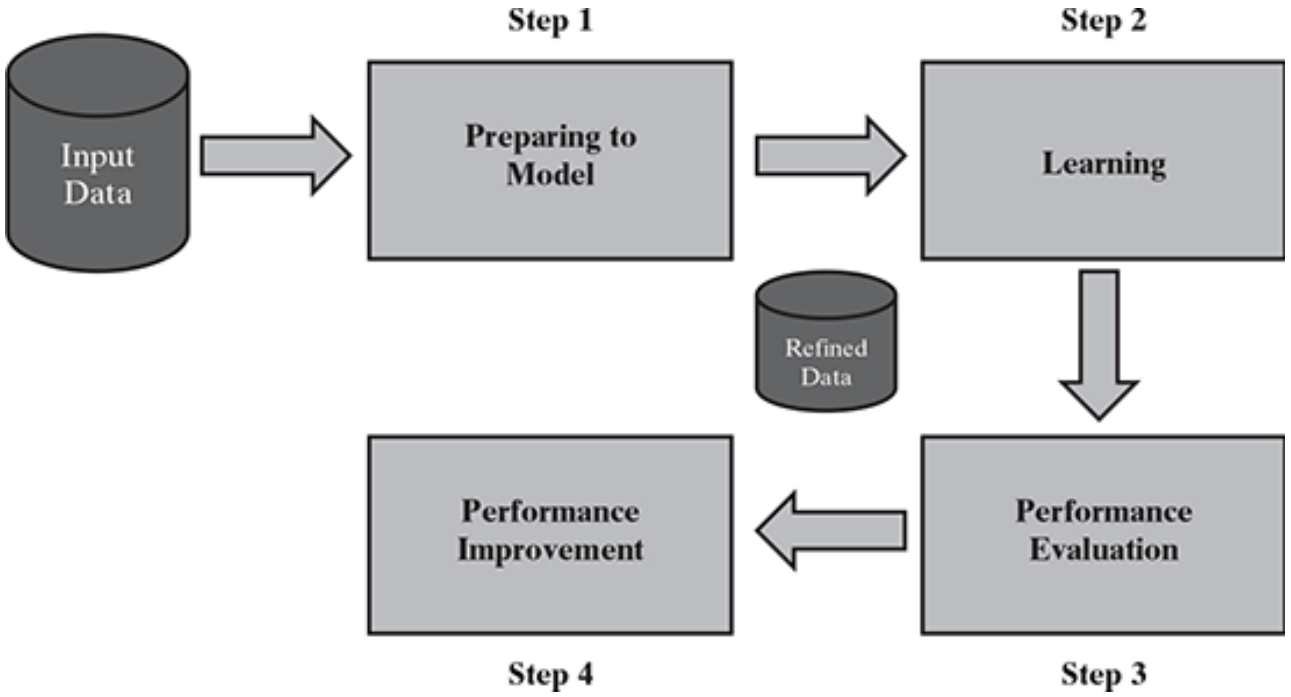
(5) Do the necessary remediation, e.g. impute missing data values, etc., if needed.

(6) Apply pre-processing steps, as necessary.

(7) 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 unsupervised 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


Step 1
Step 2
Input Data
Preparing to Model
Learning
Refined Data
Performance Improvement
Performance Evaluation
Step 4
Step 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. As shown in the following example, 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.

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%

Each row of a data set is called a sample. 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 4 features or dimensions; hence they are said to have 4-dimensional data space. A row or sample represents a point in the 4-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 sample to sample. For example, if we refer to the first two samples in the Student data set, the value of attributes Name, Gender, and Age are different

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

(2) Nationality: Indian, American, British, etc.

(3) Gender: Male, Female

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.

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.

Since ordering is possible in case of ordinal data, median, and quartiles can be identified. 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

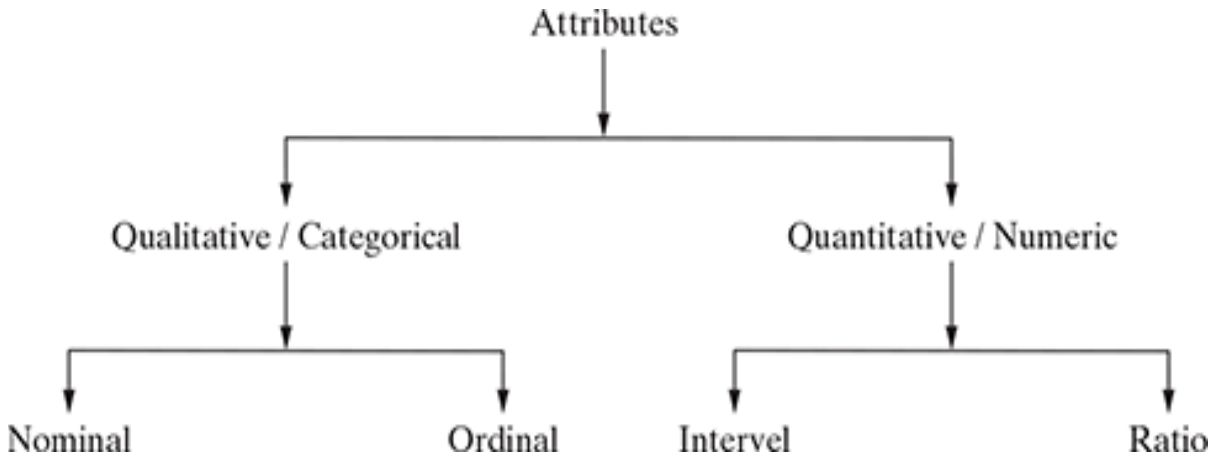
(2) Ratio data

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.

The following gives a summarized view of different types of data that we may find in a typical machine learning problem



Type of data in a machine learning problem

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.

In general, nominal and ordinal attributes are discrete. On the other hand, interval and ratio attributes are continuous

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 from 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 (https://archive.ics.uci.edu/), a collection of 400+ data sets which serve as benchmarks for researchers and practitioners in the machine learning community.

The data set, illustrated below, is the Auto MPG data set available in the UCI repository:

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

As shown in the dataset, 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 there are only finite number of values in these attributes. The remaining of the numeric attributes, i.e. ‘mpg’, ‘displacement’, ‘horsepower’, ‘weight’, and ‘acceleration’ can assume any real value.

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 both 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 mark 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. The 3rd element in the ordered list is considered as the median. Hence, the median value of this set of data is 67.

Why two measures of central tendency are reviewed? The reason is that 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. The comparison between mean and median for all the attributes is shown below:

	mpg	cylin-ders	dis-place-ment	horse-power	weight	accel-eration	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

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.

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

mpg	cylinders	displace-ment	horse-power	weight	accel-eration	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

Mode is the number that occurs most often. For example, we have 9 values

13, 13, 13, 13, 14, 14, 16, 18, 21

13 occurs most often, therefore the mode is 13.

2.4.1.2 Understanding data spread

We have explored the central tendency of the different numeric attributes, and 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 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:

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

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:

$$\sigma(x) = \sqrt{var(x)}$$

In the above case, $var(attribute\ 1) = 2$, while $var(attribute\ 1) = 79.6$.

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.

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'. The following table captures a summary of the range of statistics for the attributes.

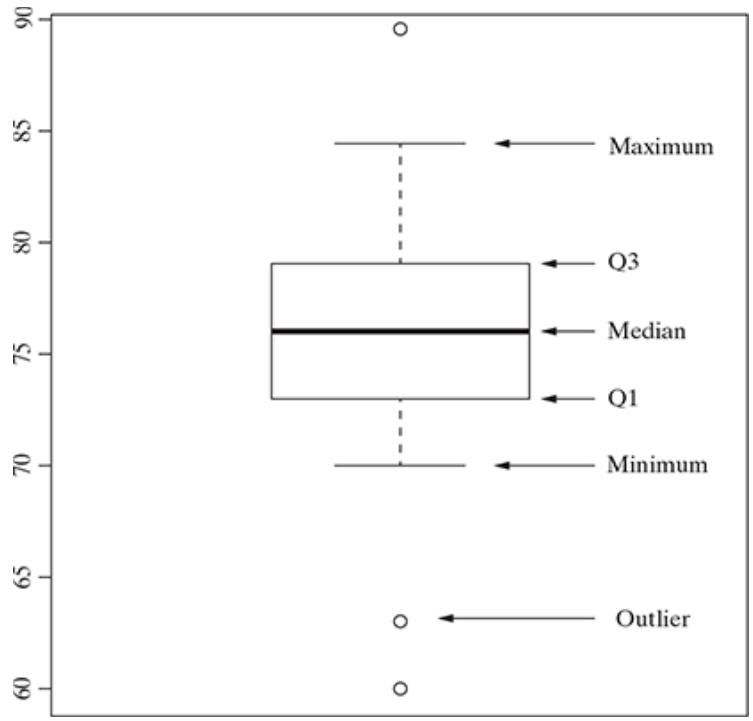
	cylinders	displacement	origin
Minimum	3	68	1
Q1	4	104.2	1
Median	4	148.5	1
Q3	8	262	2
Maximum	8	455	3

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.

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 (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 shows a box plot.



Box plot

(1) The central rectangle or the box spans from first to third quartile (i.e. Q1 to Q3), thus giving the inter-quartile range (IQR).

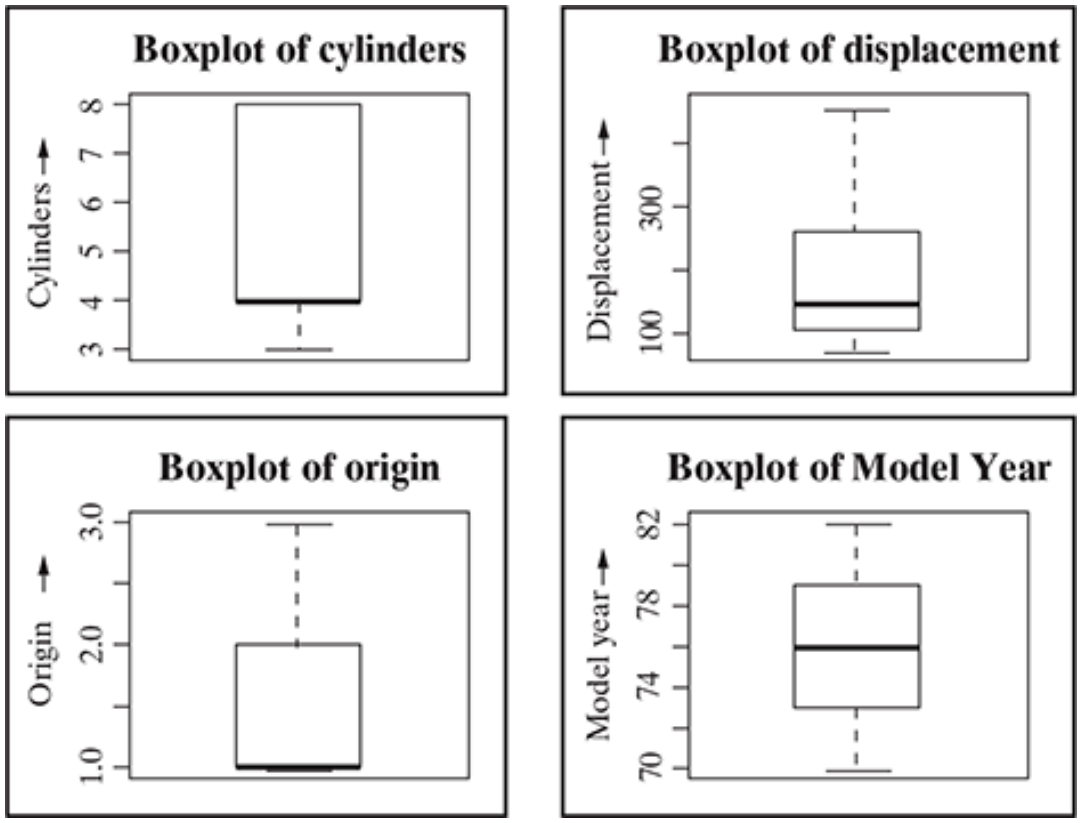
(2) Median is given by the line or band within the box.

(3) 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. 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.

(4) 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.

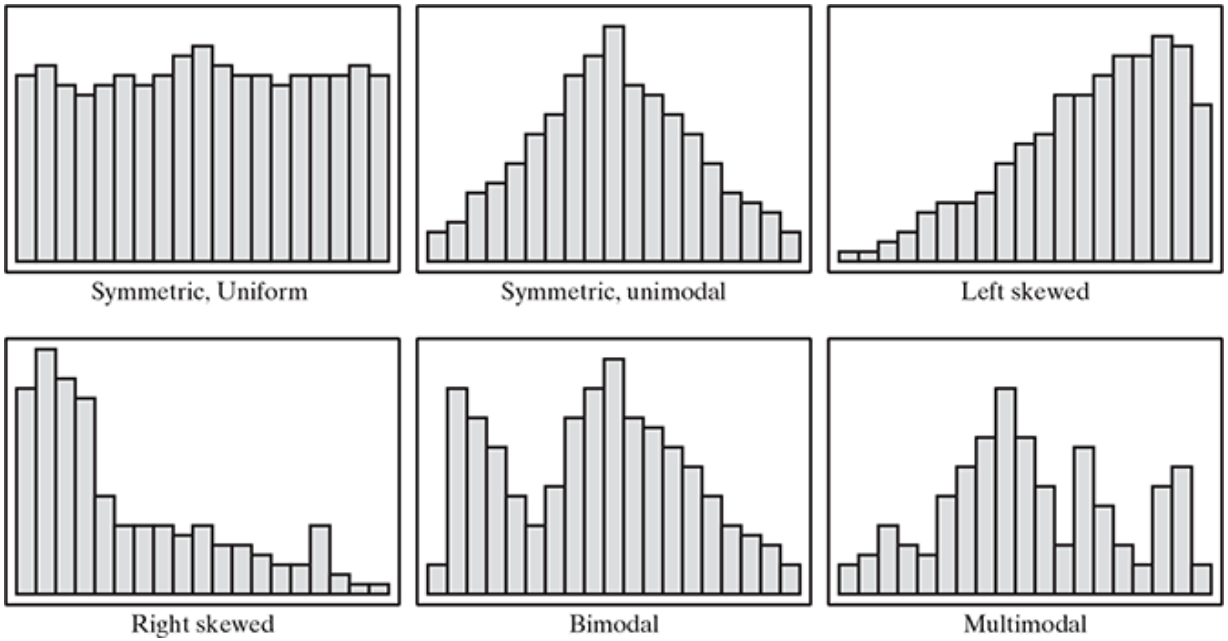
(5) The data values coming beyond the lower or upper whiskers are the ones which are of unusually low or high values respectively. These are the outliers, which may deserve special consideration.

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.



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’. Histograms might be of different shapes depending on the nature of the data as shown below



Histogram

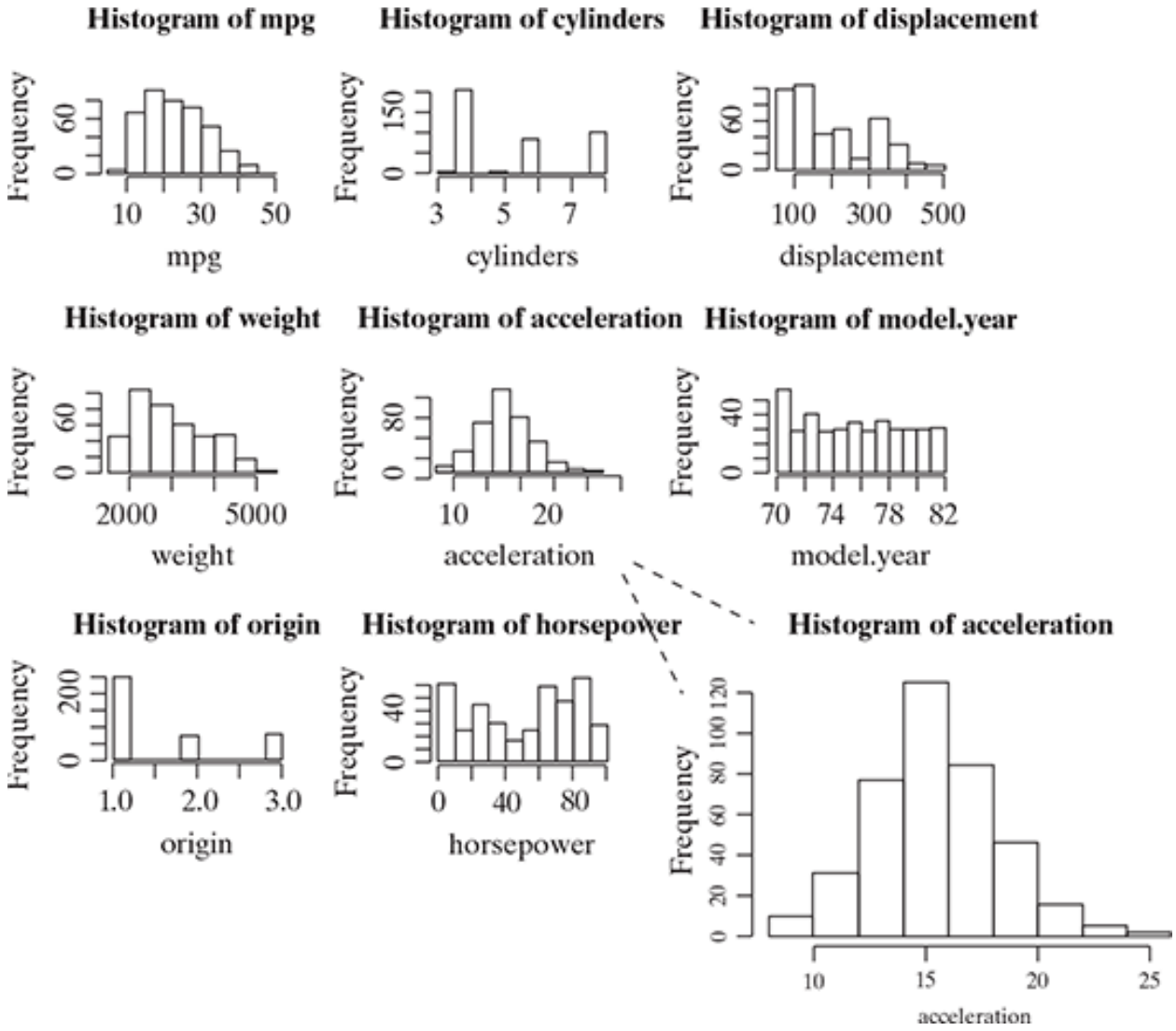
The important difference between histogram and box plot is

(1) 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.

(2) 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 as shown in the figure above. These patterns give us a quick understanding of the data and thus act as a great data exploration tool.

Let’s now examine the histograms for the different attributes of Auto MPG data set presented next. 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.



Histograms of Auto MPG data set

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.

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 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 these as follows:

Attribute Value	amc ambas-sador brougham	amc ambas-sador dpl	amc ambassa-dor sst	amc concord	amc concord d/l	amc con-cord dl 6	amc gremlin	…
Count	1	1	1	1	2	2	4	…

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

2.4.4 Exploring relationship between variables

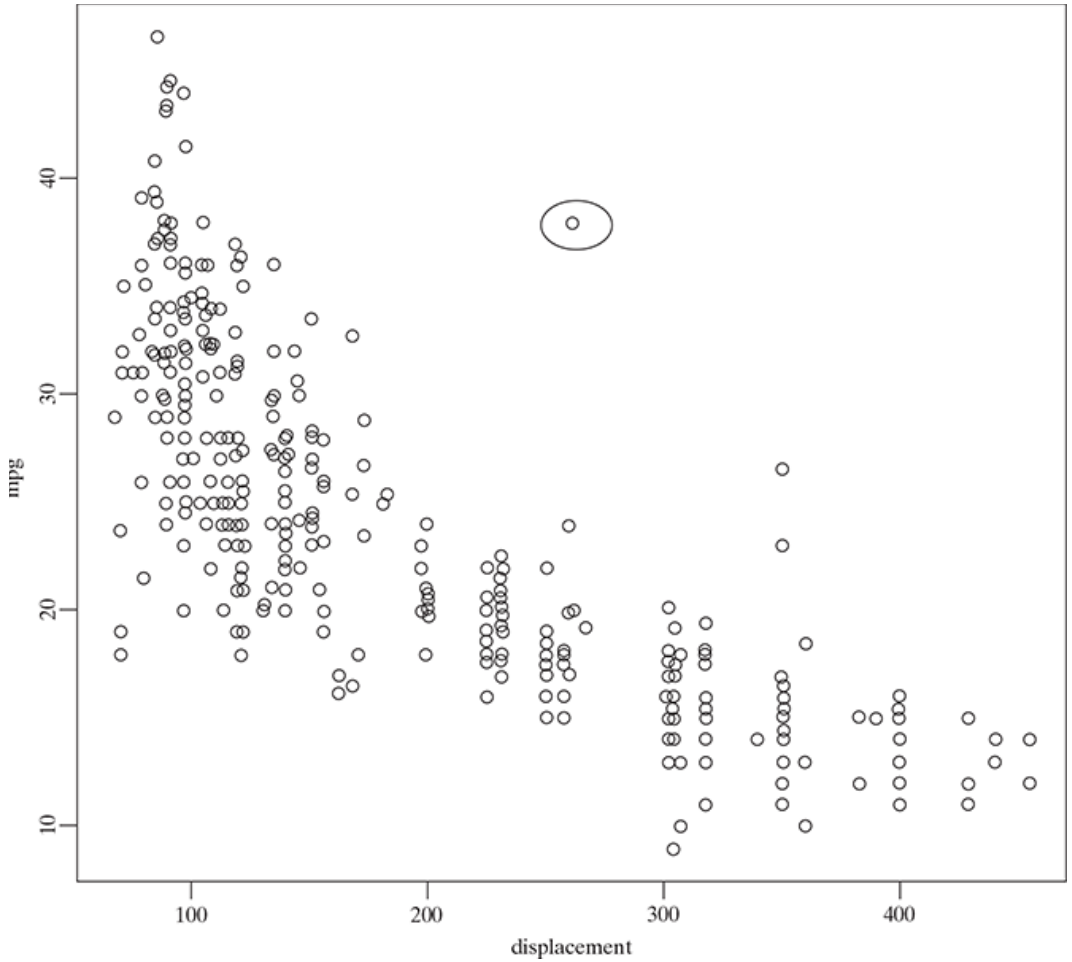
Till now we have been exploring single attributes in isolation. One more important angle of data exploration is to explore relationship between attributes. There are multiple plots to enable us explore the relationship between variables. The basic and most commonly used plot is scatter plot.

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.

In the data set Auto MPG, there is expected to be some relation between the attributes ‘displacement’ and ‘mpg’.

Let’s map ‘displacement’ as the *x*-coordinate and ‘mpg’ as the *y*-coordinate. The scatter plot comes as below


mpg
40
30
20
10
100
200
300
400
displacement

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 next cross-tab shows the relationship between attributes 'model. year' and 'origin', which helps 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.

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

The following cross-tab gives the number of 3, 4, 5, 6, or 8 cylinder cars in every region present in the sample data set.

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

We may create cross-tabs for different attribute combinations to gain a more thorough view of the relationship between attributes.

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:

(a) 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.

(b) 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 different 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, 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. Out of the two major errors mentioned above, the first one can be remedied by proper sampling technique. This is a completely different area – covered as a specialized subject area in statistics.

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.

(1) **Remove outliers:** If the number of records which are outliers is not many, a simple approach may be to remove them.

(2) **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.

(3) **Capping:** For values that lie outside the 1.5 times of 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 data should be treated as two different datasets, the model should be built for both and then the output can be combined.

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.

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 of a person who is collecting sample data or by the responder, primarily due to his/her unwillingness to respond or lack of understanding needed to provide a response.

There are multiple strategies to handle missing value of data elements. Some of those strategies are 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 is to 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’, and ‘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.

As shown in the table below, 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.

mpg	cylin-ders	dis-place-ment	horse-power	weight	accel-eration	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 Scaling, normalization, and standardization

Attribute or feature scaling, normalization and standardization involves transforming the data to make it more suitable for machine learning. These techniques can help to improve model performance, reduce the impact of outliers, and ensure that the data is on the same scale. In this article, we will explore the concepts of scaling, normalization, and standardization, including why they are important and how to apply them to different types of data.

2.6.1.1 What is Feature Scaling?

Feature scaling is a data preprocessing technique used to transform the values of features or attributes in a dataset to a similar scale. The purpose is to ensure that all features contribute equally to the model and to avoid the domination of features with larger values.

Feature scaling becomes necessary when dealing with datasets containing features that have different ranges, units of measurement, or orders of magnitude. In such cases, the variation in feature values can lead to biased model performance or difficulties during the learning process.

There are several common techniques for feature scaling, including standardization, normalization, and min-max scaling. These methods adjust the feature values while preserving their relative relationships and distributions.

By applying feature scaling, the dataset's features can be transformed to a more consistent scale, making it easier to build accurate and effective machine learning models. Scaling facilitates meaningful comparisons between features, improves model convergence, and prevents certain features from overshadowing others based solely on their magnitude

2.6.1.2 Why Should We Use Feature Scaling?

Some machine learning algorithms are sensitive to feature scaling, while others are virtually invariant. Let's explore these in more depth.

(1) Distance-based algorithms

Distance-based algorithms like K-nearest neighbour (KNN), K-means clustering and support vector machines (SVM) are most affected by the range of features. This is because, behind the scenes, they are using distances between data points to determine their similarity.

For example, let's say we have data containing high school CGPA scores of students (ranging from 0 to 5) and their future incomes (in thousands):

Student	CGPA	Salary '000
1	3.0	60
2	3.0	40
3	4.0	40
4	4.5	50
5	4.2	52

Since both the features have different scales, there is a chance that higher weightage is given to features with higher magnitudes. This will impact the performance of the machine learning algorithm; obviously, we do not want our algorithm to be biased towards one feature. The data after scaling is as follow:

Student	CGPA	Salary '000
1	-1.184341	1.520013
2	-1.184341	-1.100699
3	0.416120	-1.100699
4	1.216350	0.209657
5	0.736212	0.471728

The effect of scaling is conspicuous when we compare the Euclidean distance between data points for students 1 and 2, and students 2 and 3 , before and after scaling, as shown below:

- Distance between students 1 & 2 before scaling

$$\sqrt{(40-60)^2+(3-3)^2}=20$$

- Distance between students 2 & 3 before scaling

$$\sqrt{(40-40)^2+(4-3)^2}=1$$

- Distance between students 1 & 2 after scaling

$$\sqrt{(-1.10-1.52)^2+(-1.18+1.18)^2}=2.6$$

- Distance between students 2 & 3 after scaling

$$\sqrt{(-1.1+1.1)^2+(0.42+1.18)^2}=1.59$$

(2) Gradient descent-based algorithms

Many machine learning algorithms use gradient descent as an optimization technique require data to be scaled. Take a look at the formula for gradient descent below:

$$\theta_i(k+1) = \theta_i(k) - \eta \frac{\partial L(k)}{\partial \theta_i(k)} x_i(k)$$

The presence of feature value x in the formula will affect the update of parameter. The big difference in the ranges of features will cause very different update for the parameters. Having features on a similar scale can help the gradient descent converge more quickly towards the minima.

2.6.1.3 Normalization

Normalization is a data preprocessing technique used to adjust the values of features in a dataset to a common scale. This is done to facilitate data analysis and modelling, and to reduce the impact of different scales on the accuracy of machine learning models.

Normalization is a scaling technique in which values are shifted and rescaled so that they end up ranging between 0 and 1. It is also known as Min-Max scaling:

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}}$$

Where x_{min} and x_{max} denotes the minimum and maximum value of x

6.3.1.4 Standardization

Standardization is another scaling method to centre the data around the mean with a unit standard deviation. This means that the mean of the attribute becomes zero, and the resultant distribution has a unit standard deviation. Here's the formula for standardization:

$$x' = \frac{x - \mu}{\sigma}$$

Where μ and σ denotes the mean and standard deviation of x. Note that, in this case, the values are not restricted to a particular range.

2.6.1.5 Normalization or Standardization?

Normalization	Standardization
Rescales values to a range between 0 and 1	Centers data around the mean and scales to a standard deviation of 1
Useful when the distribution of the data is unknown or not Gaussian	Useful when the distribution of the data is Gaussian or unknown
Sensitive to outliers	Less sensitive to outliers
Retains the shape of the original distribution	Changes the shape of the original distribution
May not preserve the relationships between the data points	Preserves the relationships between the data points

However, at the end of the day, the choice of using normalization or standardization will depend on your problem and the machine learning algorithm you are using. There is no hard and fast rule to tell you when to normalize or standardize your data. You can always start by fitting your model to raw, normalized, and standardized data, and compare the performance for the best results.

2.6.2 Dimensionality reduction

Till the end of the 1990s, very few data sets have 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, high-dimensional data is frequently encountered.

High-dimensional data sets need a high amount of computational space and time. At the same time, not all features are useful. 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.

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

2.6.3 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 the 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 on model accuracy.

There are different ways to select a feature subset. We will be discussing feature selection in details later.

2.7 SUMMARY

(1) A data set is a collection of related information of samples/records.

(2) Data can be broadly divided into following two types

- Qualitative data
- Quantitative data

(3) 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

(4) 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.

(5) 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.

(6) 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, including minimum, first quartile (Q1), median (Q2), third quartile (Q3), and maximum

(7) Exploration of numerical data can be best done using box plots and histograms.

(8) Options for exploration of categorical data are very limited.

(9) For exploring relations between variables, scatter-plots and two-way cross-tabulations can be effectively used.

(10) 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

(11) Data needs to be scaled by either normalization or standardization

(12) 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.

(13) Some popular dimensionality reduction techniques are PCA, SVD, and feature selection.