2. Data Preparation for Machine Learning

2.1 MACHINE LEARNING ACTIVITIES

- Machine learning activity starts with data.

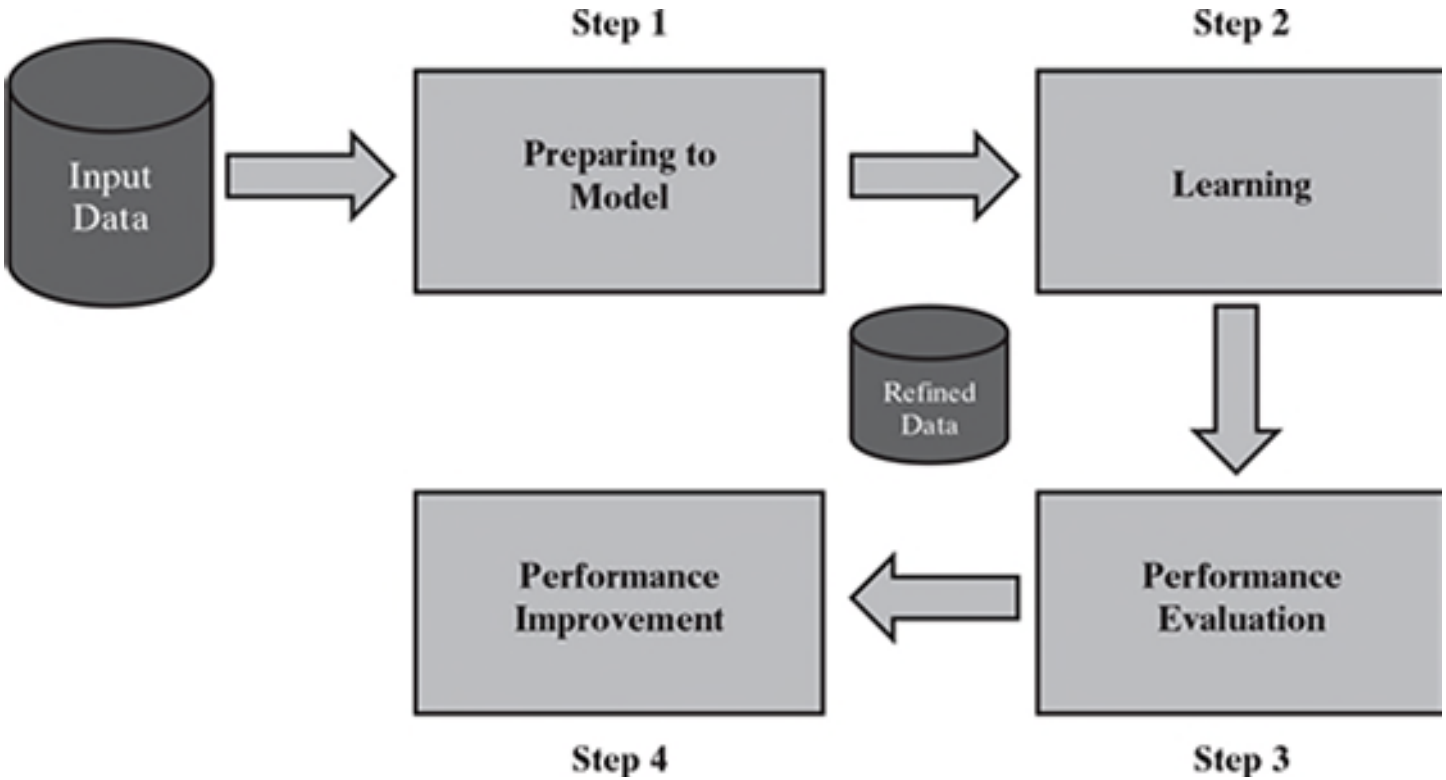
- ❑ In case of supervised learning, it is the labelled training data set followed by test data which is not labelled.
- ❑ In case of unsupervised learning, there is no question of labelled data, but the task is to find patterns in the input data.

Following are the typical preparation activities:

- 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, then the learning tasks start

- The input data is first divided into two parts – the training data and the test data. 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 it to unknown test data. For unsupervised learning, directly apply the learning algorithm to the input data.
- Evaluate the model performance


Step 1
Step 2
Input Data
Preparing to Model
Learning
Refined Data
Performance Improvement
Performance Evaluation
Step 4
Step 3

2.2 BASIC TYPES OF DATA IN MACHINE LEARNING

A data set is a collection of related information or records.

The following example is a data set on students, where each record consists of information about a specific student.

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

The following is a data set on student performance:

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, such as Roll Number, Maths, Science and Percentage in above example
- Each attribute, also called variable or feature, gives information on a specific characteristic.

Data can broadly be divided into following two types:

- Qualitative data
- Quantitative data

Qualitative data provides information about the quality or information which cannot be measured:

- The quality of performance, in terms of 'Good', 'Average', and 'Poor', is qualitative data.
- Name or roll number of students are information that cannot be measured using some scale of measurement, are qualitative data.

Qualitative data is also called categorical data.

Qualitative data can be further subdivided into two types:

- Nominal data
- Ordinal data

Nominal data is one which has no numeric value, but a named value. For example:

- Blood group: A, B, O, AB
- Nationality: Indian, American, British, etc.
- Gender: Male, Female

For nominal data:

- Mathematical operations such as addition, subtraction, multiplication, etc. cannot be performed.
- Therefore, mean and variance etc. also cannot be applied.

Ordinal data can be naturally ordered.

Examples of ordinal data are:

- Customer satisfaction: ‘Happy’, ‘Unhappy’, etc.
- Grades: A, B, C, etc.
- Hardness of Metal: ‘Very Hard’, ‘Hard’, ‘Soft’, etc.

For ordinal data:

- Since ordering is possible, median and quartiles can be identified.
- Mean still cannot be calculated.

Quantitative data relates to information about the quantity of an object

- Can be measured.
- Quantitative data is also termed as numeric data.

There are two types of quantitative data:

- Interval data
- Ratio data

For Interval data, not only the order is known, but the exact difference between values is also known. Examples:

- Celsius temperature
- Date, time, etc.

For interval data, mathematical operations such as addition and subtraction are possible.

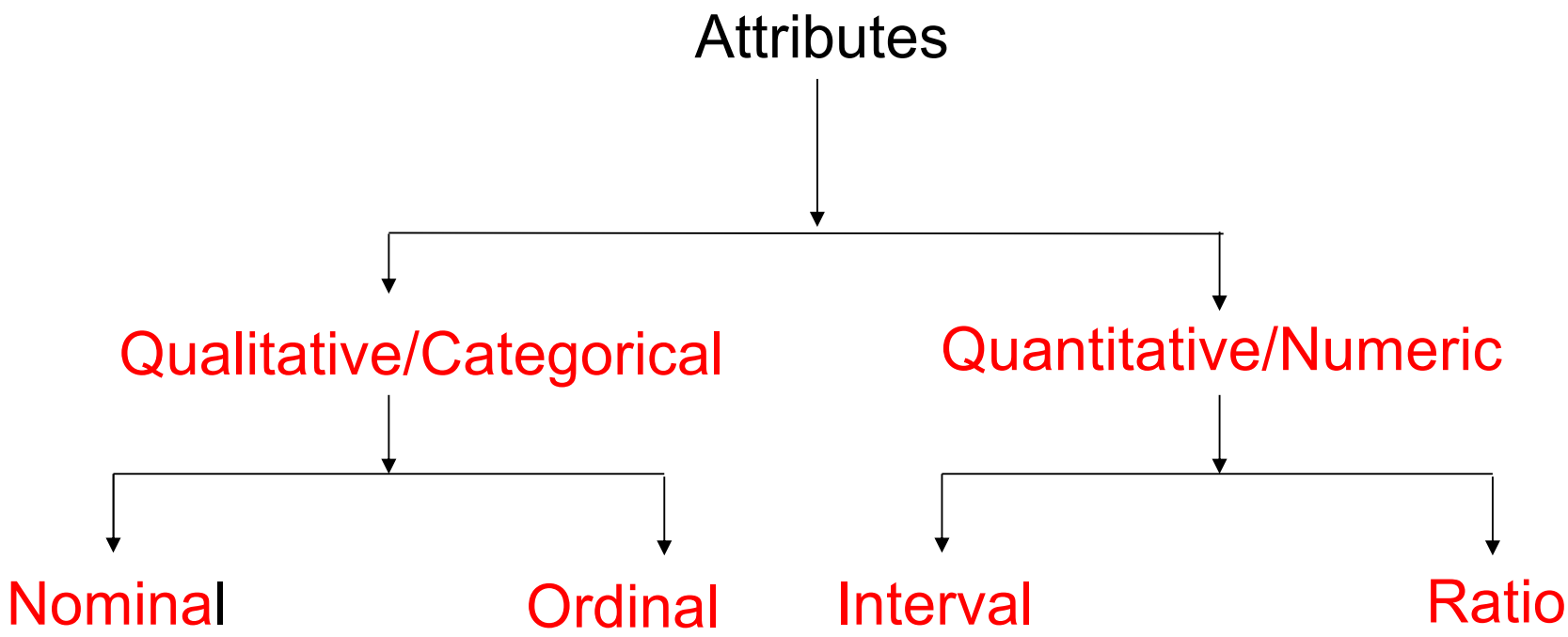
For this reason,

- The central tendency can be measured by mean, median, or mode.
- Standard deviation can be calculated.
- Addition and subtraction applies for interval data
- The ratio cannot be applied.

Ratio data represents numeric data for which exact value can be measured.

- Absolute zero is available for ratio data.
- Can be added, subtracted, multiplied, or divided.
- The central tendency can be measured by mean, median, or mode.
- The dispersion such as standard deviation can be measured.

The following gives a summarized view of different types of data:



Attributes can also be categorized into types based on the number of values that can be assigned:

- Discrete or
- Continuous

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
- Numeric attributes such as count, rank of students, etc. can have countably infinite values.
- Binary attribute is a special type of discrete attribute with two possible values, such as, positive/negative, yes/no, etc.

Countably infinite (*from ChatGPT):*

A set is **countably infinite** if it has the same size as the set of natural numbers $\mathbb{N} = \{1, 2, 3, 4, \dots\}$, meaning its elements can be put in a one-to-one correspondence with $\mathbb{N}$. In simpler terms, you can "count" the elements, even if there are infinitely many.

Examples of Countably Infinite Sets:

1. **Natural numbers** $\mathbb{N}$
2. **Integers** $\mathbb{Z} = \{\dots, -2, -1, 0, 1, 2, \dots\}$
3. **Rational numbers** $\mathbb{Q}$ (fractions like 1/2, -3/4, etc.)
4. **Finite-length binary strings** (e.g., "0", "10", "1101", 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.
- Interval and ratio attributes are continuous

2.3 EXPLORING STRUCTURE OF DATA

We can now delve deeper into understanding a data set.

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.

Example

Auto MPG

Donated on 7/6/1993

Revised from CMU StatLib library, data concerns city-cycle fuel consumption

Dataset Characteristics	Subject Area	Associated Tasks
Multivariate	Other	Regression
Feature Type	**# Instances**	**# Features**
Real, Categorical, Integer	398	7

Dataset Information

Additional Information

This dataset is a slightly modified version of the dataset provided in the StatLib library. In line with the use by Ross Quinlan (1993) in predicting the attribute "mpg", 8 of the original instances were removed because they had unknown values for the "mpg" attribute. The original dataset is available in the file "auto-mpg.data-original".

"The data concerns city-cycle fuel consumption in miles per gallon, to be predicted in terms of 3 multivalued discrete and 5 continuous attributes." (Quinlan, 1993)

SHOW LESS ^

Has Missing Values?

Yes

Variables Table

Variable Name	Role	Type	Description	Units	Missing Values
displacement	Feature	Continuous			no
mpg	Target	Continuous			no
cylinders	Feature	Integer			no
horsepower	Feature	Continuous			yes
weight	Feature	Continuous			no
acceleration	Feature	Continuous			no
model_year	Feature	Integer			no
origin	Feature	Integer			no
car_name	ID	Categorical			no

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

‘mpg’, ‘cylinders’, ‘displacement’, ‘horsepower’, ‘weight’, ‘acceleration’, ‘model year’, and ‘origin’ are all numeric:

- ‘cylinders’, ‘model year’, and ‘origin’ are discrete in nature, as they have only finite number of values.
- ‘mpg’, ‘displacement’, ‘horsepower’, ‘weight’, and ‘acceleration’ can assume any real value.

‘cylinders’ or ‘origin’ have a small number of possible values, one may prefer to treat it as a categorical or qualitative attribute.

‘car name’ is of type categorical, or more specifically nominal.

2.4 EXPLORE NUMERICAL DATA

2.4.1 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.

For example, mean of a set of observations: 21, 89, 34, 67, and 96 is calculated as below:

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

Median, is the value of the element appearing in the middle of an ordered list of data elements. For example:

$$21, 34, 67, 89, 96$$

The median value of this set of data is 67.

Mean and median are impacted differently by data values appearing at the beginning or at the end of the range.

- Mean is sensitive to outliers.
- If the difference between mean and median is quite high, we should find out the root cause along with the need for remediation.

Mode is the number that occurs most often.

For example, we have 9 values

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

Value 13 occurs the most often, therefore the mode is 13.

Mode is usually used for measuring central tendency of categorical attributes.

For the Auto MPG data set, we can find out if the deviation between median and mean 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

- For ‘mpg’, ‘weight’, ‘acceleration’, and ‘model.year’, the deviation between mean and median is not significant which means the chance of having too many outlier values is less.
- 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.
- There is some problem in the values of the attribute ‘horsepower’ because the mean and median calculation is not possible.

The problem is occurring because 6 data elements do not have value for ‘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

This is the so called missing value problem.

2.4.2 Understanding data spread

We next take a granular view of the data spread in the form of

- Dispersion of data
- Position of the different data values

2.4.2.1 Measuring data dispersion

To measure the extent of dispersion of a data, the variance of the data is measured using the formula given below:

$$var(X) = \frac{1}{N}\sum_{k=1}^{N} x^2(k) - \left(\frac{1}{N}\sum_{k=1}^{N} x(k)\right)^2$$

Standard deviation of a data is measured as follows:

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

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

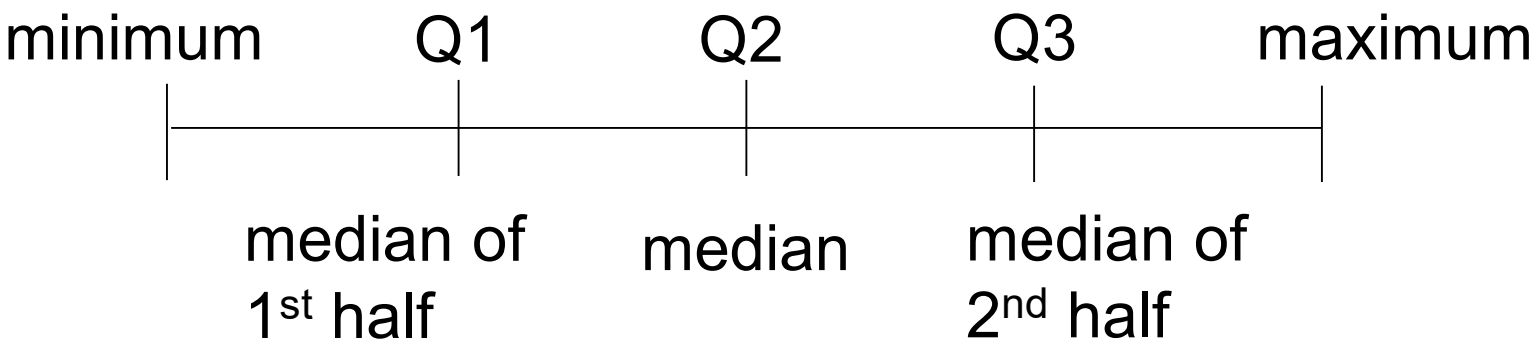
In the above case,

- $var(attribute\ 1) = 2$, which means attribute 1 values are quite concentrated around the mean
- $var(attribute\ 2) = 79.6$, which means attribute 2 values are extremely spread out.

2.4.2.2 Measuring data value position

Median gives the central data value, which divides the entire data set into two halves:

- The median of the first half is called first quartile or Q1.
- The median of the second half is called Q3.
- The overall median is second quartile or Q2.
- Any data set has five values - minimum, Q1, Q2, Q3, and maximum.



The summary of the range of attributes ‘cylinders’, ‘displacement’, and ‘origin’:

	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

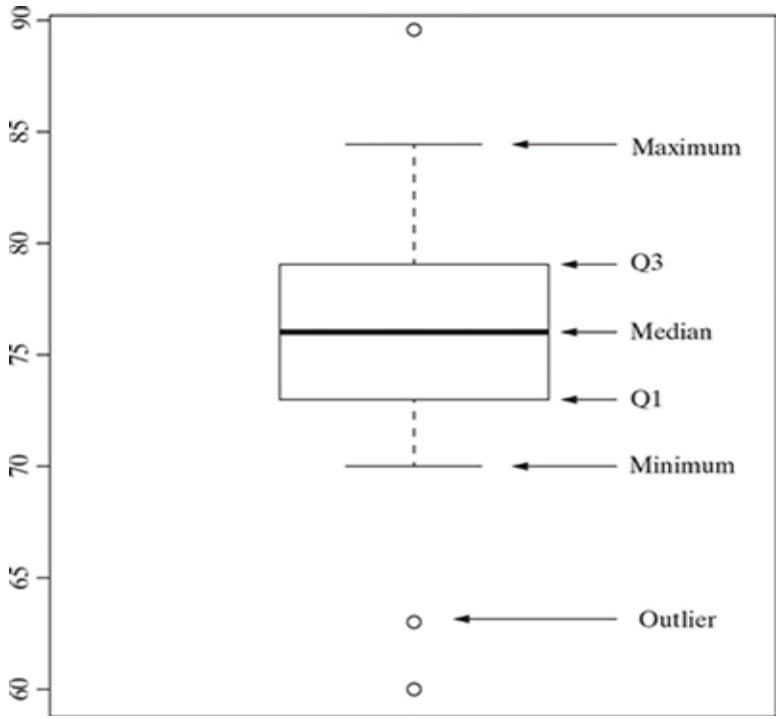
Take ‘displacement’ as an example:

- The difference between minimum value and Q1 is 36.2 and difference between Q1 and median is 44.3.
- The difference between median and Q3 is 113.5 and difference between Q3 and the maximum value is 193.
- Therefore, the larger values are more spread out than the smaller ones.

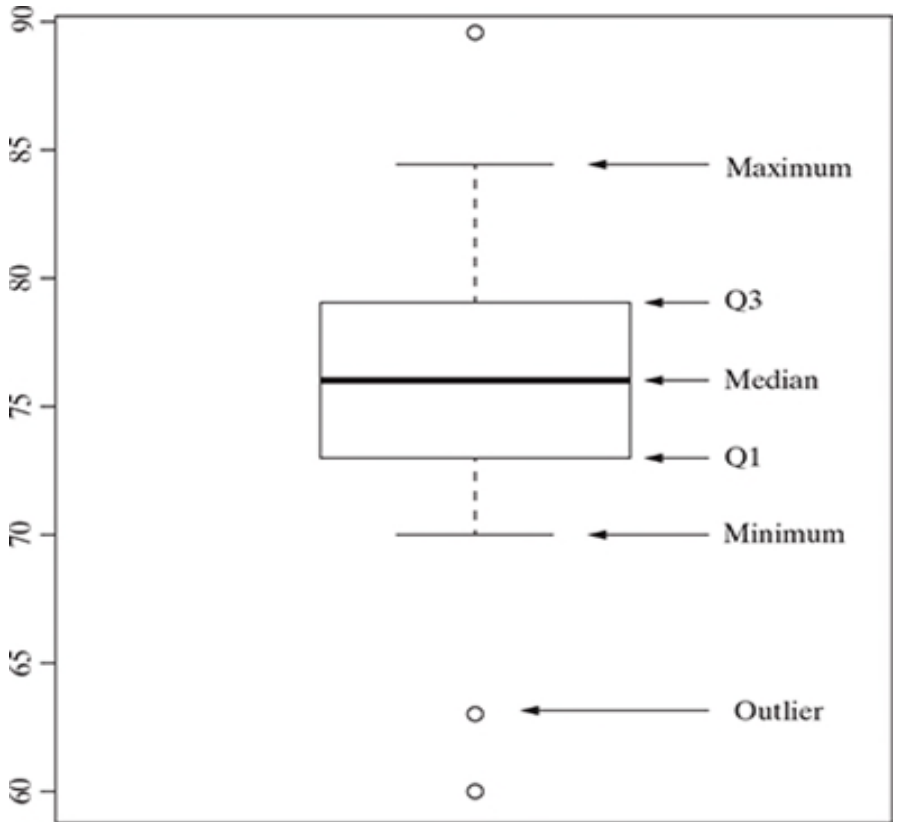
2.4.3 Plotting and exploring numerical data

2.4.3.1 Box plots

A box plot gives a standard visualization of the five-number summary statistics:



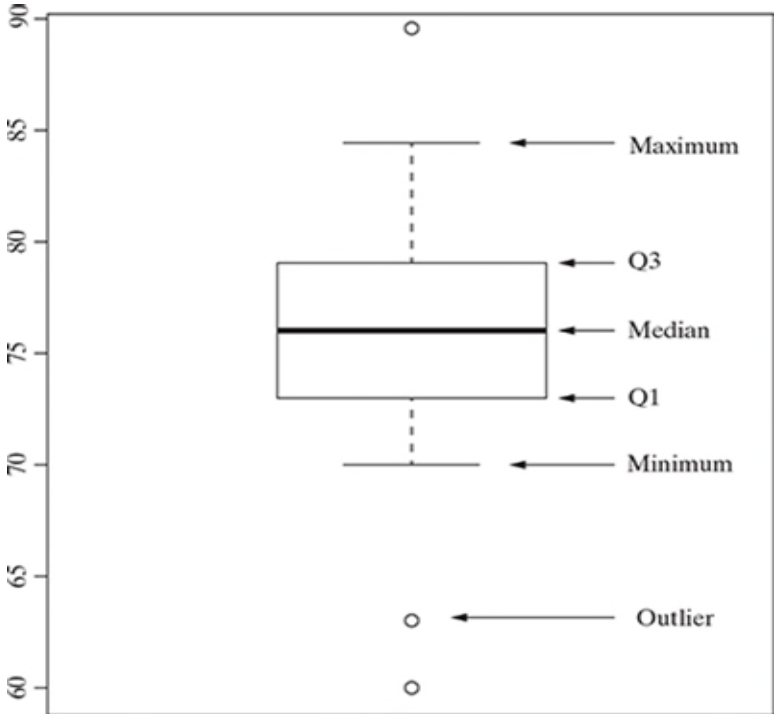
- The central rectangle or the box spans from Q1 to Q3. The inter-quartile range IQR=Q3-Q1.
- Median is given by the line or band within the box.



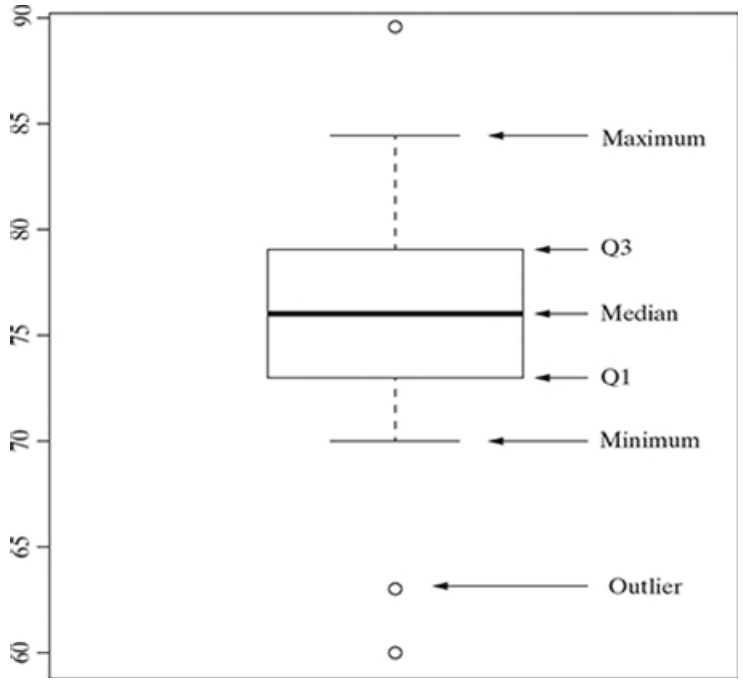
- The lower whisker extends up to 1.5 times of the IQR from Q1.

Say Q1 = 73, median = 76, Q3 = 79, IQR=Q3-Q1=6. Then Q1 – 1.5 × IQR= 73 – 1.5 × 6 = 64.

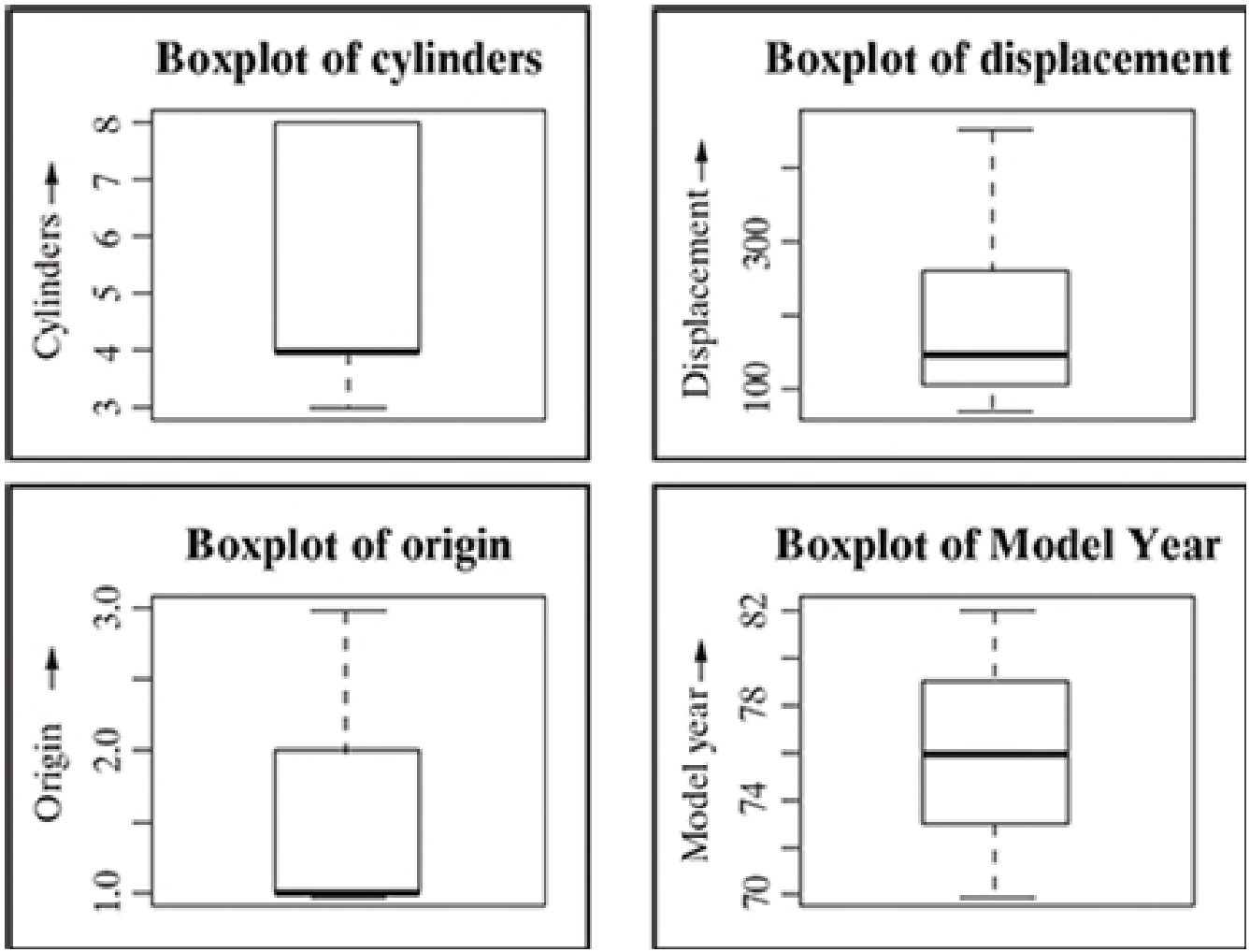
- If there are lower range data values such as 70, 63, and 60. So, the lower whisker will come at 70 as this is the lowest data value larger than 64.



- The upper whisker extends up to 1.5 as times of IQR from Q3. For example, Q3 + 1.5 × IQR = 79 + 1.5 × 6 = 88.
- If there is higher range of data values like 82, 84, and 89, 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 outliers.

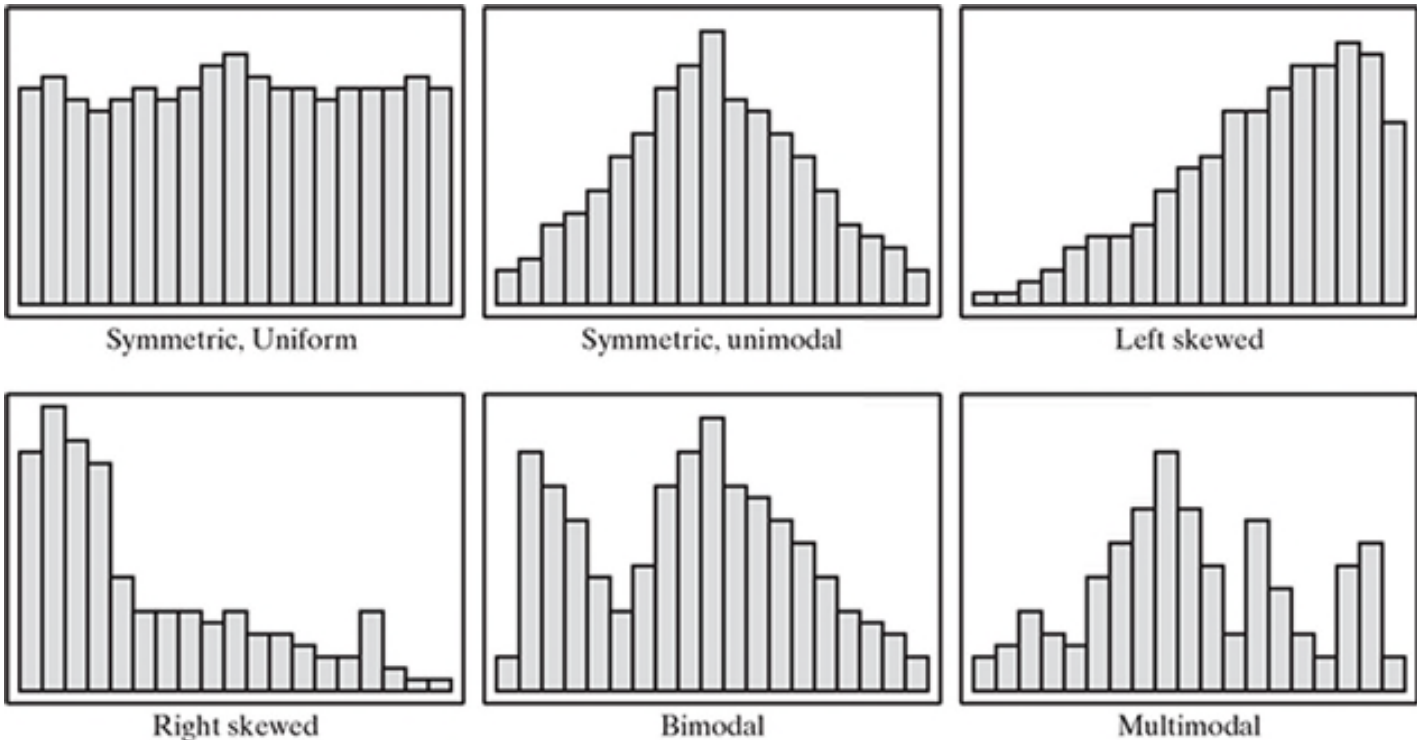


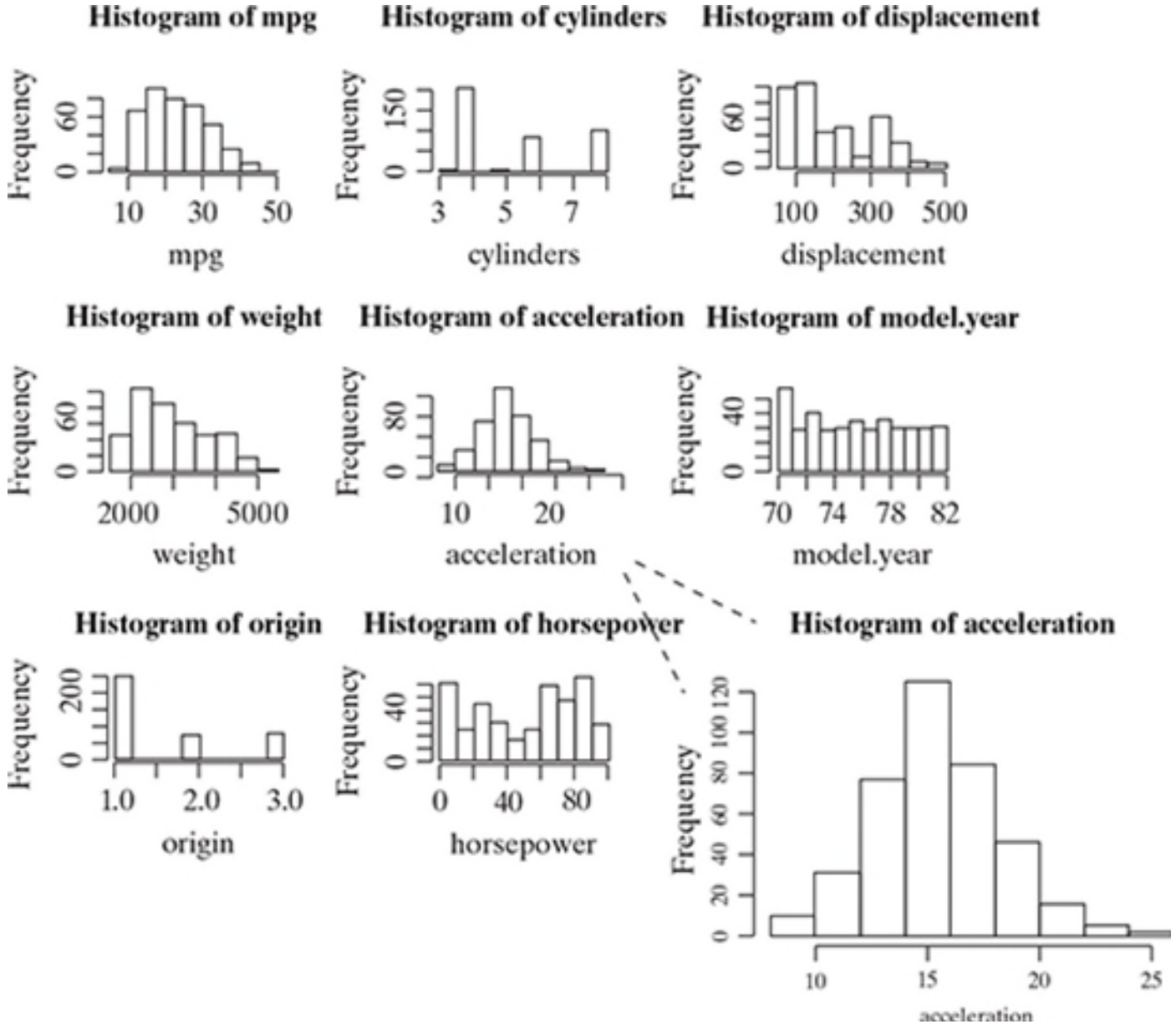
The following are the box plot for four attributes:



2.4.3.2 Histogram

Histogram helps in understanding the distribution of a numeric data into series of intervals, also termed as 'bins'. It uses bars to show how often different ranges (bins) of values appear in a dataset




Histogram of mpg
Frequency
0
60
10
30
50
mpg
Histogram of cylinders
Frequency
0
150
3
5
7
cylinders
Histogram of displacement
Frequency
0
60
100
300
500
displacement
Histogram of weight
Frequency
0
60
2000
5000
weight
Histogram of acceleration
Frequency
0
80
10
20
acceleration
Histogram of model.year
Frequency
0
40
70
74
78
82
model.year
Histogram of origin
Frequency
0
200
1.0
2.0
3.0
origin
Histogram of horsepower
Frequency
0
40
0
40
80
horsepower
Histogram of acceleration
Frequency
0
20
40
60
80
100
120
10
15
20
25
acceleration

2.5 EXPLORING CATEGORICAL DATA

There are not many options for exploring categorical data. For exam, for the Auto MPG data set, the first summary is how many unique values are there for ‘car.name’ and ‘cylinders’:

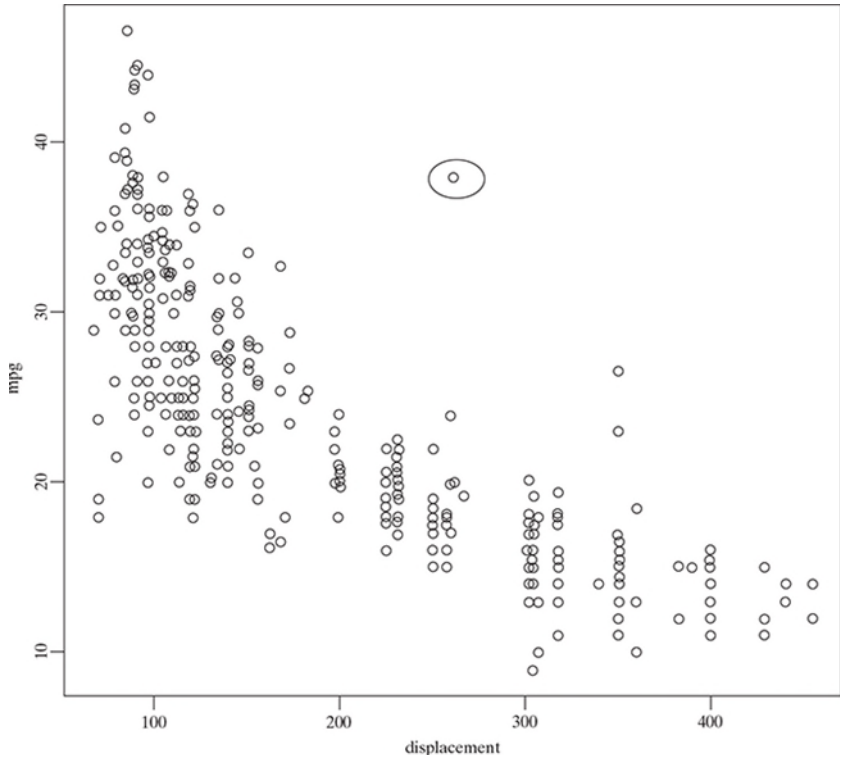
Attribute Value	amc ambas-sador brougham	amc ambas-sador dpl	amc ambassa-dor sst	amc concord	amc concord d/l	amc con-cord dl 6	amc gremlin	…
Count	1	1	1	1	2	2	4	…

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

2.6 EXPLORING RELATIONSHOP BETWEEN VARIABLES

2.6.1 Scatter plot

A scatter plot is a two-dimensional plot in which points or dots are drawn on coordinates provided by values of the attributes.



2.6.2 Two-way cross-tabulations

Two-way cross-tabulations, also called cross-tab or contingency table, has a matrix format that presents a summarized view of the bivariate frequency distribution.

It helps to understand how much the data values of one attribute changes with the change in data values of another attribute.

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

2.7 DATA QUALITY AND REMEDIATION

2.7.1 Data quality

Two types of commonly encountered problems:

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

Multiple factors could lead to these data quality issues:

- Incorrect sample set selection
- Errors in data collection

2.7.2 Data remediation

2.7.2.1 Handling outliers

Outliers are samples with an abnormally high or low value. We may consider one of the following approaches to handle outliers.

- Remove outliers: If the number of outliers is not many, a simple approach may be to remove them.
- Imputation: Impute the value with mean or median or mode. The mean/median/mode of the most similar samples may also be used for imputation.

- Capping: For values that lie outside the 1.5 times of IQR limits, we can cap them by replacing those observations with the value of 5th percentile or the value of 95th percentile.

  - ✓ If there is a significant number of outliers, they should be treated separately in the statistical model: the data should be treated as two different datasets, the model should be built for both.

  - ✓ However, if the outliers are natural, i.e. because of a valid reason, then we should not amend it.

2.7.2.2 Handling missing values

In a data set, one or more data elements (i.e. attributes) may have missing values. There are multiple strategies to handle missing values.

❑ Eliminate samples having missing values

In case the proportion of data having missing values is within a tolerable limit, a simple but effective approach is to remove the samples having miss values.

This is possible if the quantum of data left after removing the samples having missing values is sizeable.

In the case of Auto MPG data set, 6 out of 398 samples, the value of attribute ‘horsepower’ is missing. If we get rid of those 6 samples, we will still have 392 samples. So, we can very well eliminate the 6 samples and keep working with the remaining data set.

However, this will not be possible if the proportion of samples having missing value is really high

❑ Imputing missing values

Imputation is a method to assign a value to the data elements having missing values. Mean/mode/median is the most frequently assigned value.

- For quantitative attributes, missing values can be imputed with the mean, median, or mode of the remaining values under the same attribute.
- For qualitative attributes, missing values can be imputed by the mode of all remaining values of the same attribute.
- Another strategy is to identify the similar types of observations whose values are known and use the mean/median/mode of those known values.

- 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.
- The other approach is that we can take a similarity-based mean or median.

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

If we refer to the 6 observations with missing values for attribute ‘horsepower’, and ‘cylinders’ is the attribute which is logically most connected to ‘horsepower’,

- We can use the mean of the ‘horsepower’ attribute of samples having cylinders = 4 to impute value to 5 samples with the missing ‘horsepower’ and with ‘cylinders’ 4;
- For the 1 observation which has cylinders = 6, we can use a similar mean of samples with cylinders = 6, to impute the missing value,

❑ 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.

2.8 DATA PRE-PROCESSING

Scaling, normalization, and standardization

2.8.1 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.

Why Feature Scaling?

Some machine learning algorithms are sensitive to feature scaling, while others are virtually invariant.

- Distance-based algorithms Distance-based algorithms are most affected by the range of features. This is because they are using distances between data points to determine their similarity.

For example, 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.

The data after scaling is:

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

- 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

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

- Gradient descent-based algorithms Many machine learning algorithms use gradient descent as an optimization technique:

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

The big difference in the ranges of features will cause very different update for the parameters.

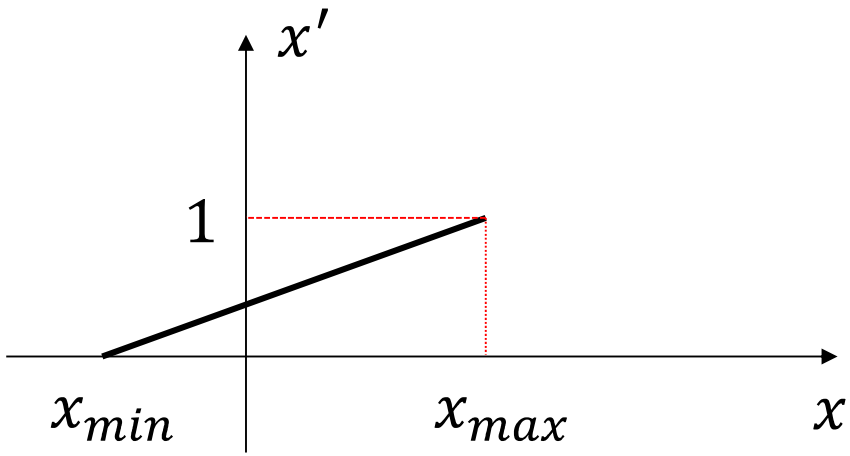
Having features on a similar scale can help the gradient descent converge more quickly towards the minima.

2.8.2 Normalization

Normalization is a scaling technique in which values are shifted and rescaled to the range 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



2.8.3 Standardization

Standardization is another scaling method to centre the data around the mean with a unit standard deviation:

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

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.

After standardization, the mean of the attribute becomes zero, and the standard deviation is one.

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

The choice of using normalization or standardization will depend on your problem and the machine learning algorithm. You can always start by fitting your model to raw, normalized, and standardized data, and compare the performance for the best results.

2.8.4 Dimensionality reduction

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 is reduced.

Dimensionality reduction refers to the techniques of reducing the dimensionality of a data.

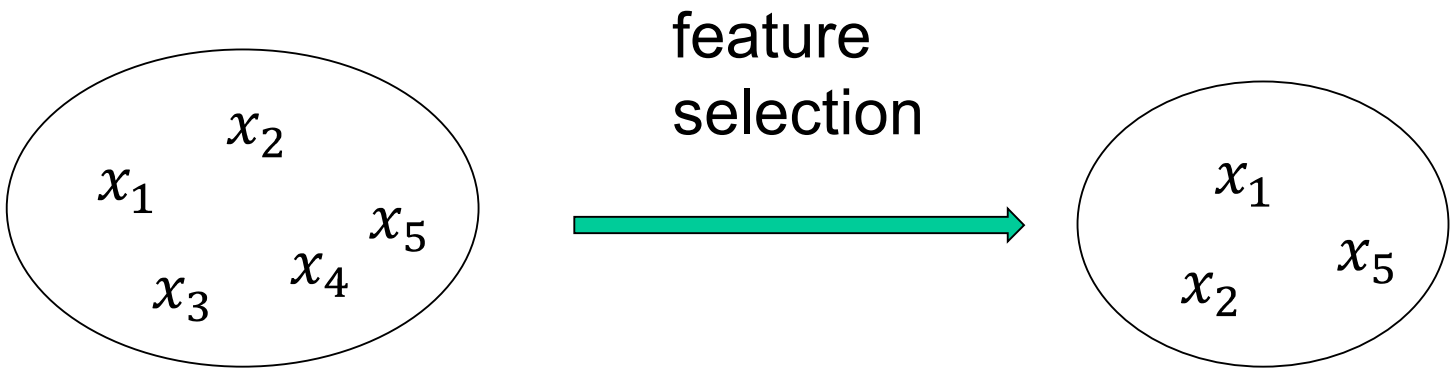
The most common approach for dimensionality reduction is known as Principal Component Analysis (PCA), another is Singular Value Decomposition (SVD).

These methods are also called feature extraction.

Feature subset selection

Feature subset selection or simply called feature selection try to find out the optimal subset of the entire feature set

- Reduces computational cost
- Without any major impact on the learning accuracy.



2.9 SUMMARY

- A data set is a collection of related information of samples 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. It 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:
  - Interval data: numeric data for which the exact difference between values is known. Such data do not have 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.

- Measure of data spread is available in the form of
  - Dispersion of data measured by variance
  - Related to the position of the different data values, there are five values: maximum, minimum, Q1,Q2,Q3

- 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
- Data needs to be scaled by either normalization or standardization

- 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 the feature extraction method such as PCA and SVD, and the feature selection method.