Chapter 3

Data Exploration

Data Exploration

- Data science helps decipher the hidden useful relationships within data.
- Before venturing into any advanced analysis of data using statistical, machine learning, and algorithmic techniques, it is essential to perform basic data exploration to study the basic characteristics of a dataset.
- Data exploration helps with understanding data better, to prepare the data in a way that makes advanced analysis possible, and sometimes to get the necessary insights from the data faster than using advanced analytical techniques.

Data Exploration

- Data exploration can be broadly classified into two types—

  I. descriptive statistics.

  II. data visualization

DESCRIPTIVE STATISTICS

Descriptive statistics refers to the study of the aggregate quantities of a dataset. Descriptive statistics can be broadly classified into

- Univariate Exploration
- Multivariate Exploration

Univariate Exploration

Univariate data exploration denotes analysis of one attribute at a time. The example Iris dataset for one species, I. setosa, has 50 observations and 4 attributes. Here some of the descriptive statistics for sepal length attribute are explored

Measure of Central Tendency:

The objective of finding the central location of an attribute is to quantify the dataset with one central or most common number.

Mean: The mean is the arithmetic average of all observations in the dataset. It is calculated by summing all the data points and dividing by the number of data points

Median: The median is the value of the central point in the distribution. The median is calculated by sorting all the observations from small to large and selecting the mid-point observation in the sorted list. If the number of data points is even, then the average of the middle two data points is used as the median. The median for sepal length is in centimeters is 5.0000.

Mode: The mode is the most frequently occurring observation. In the dataset, data points may be repetitive, and the most repetitive data point is the mode of the dataset. In this example, the mode in

centimeters is 5.1000.

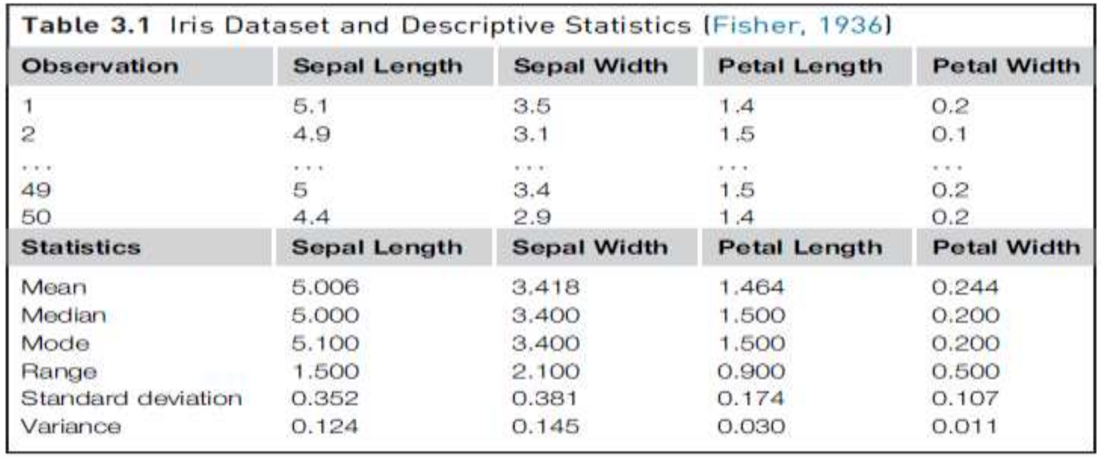
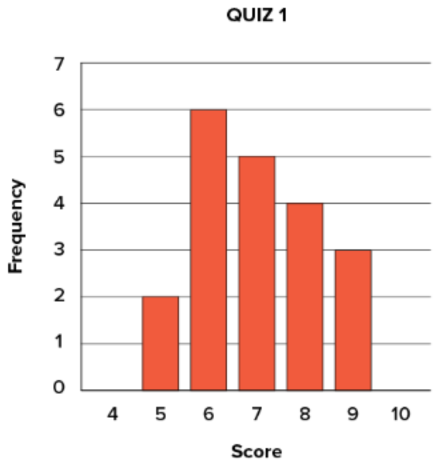


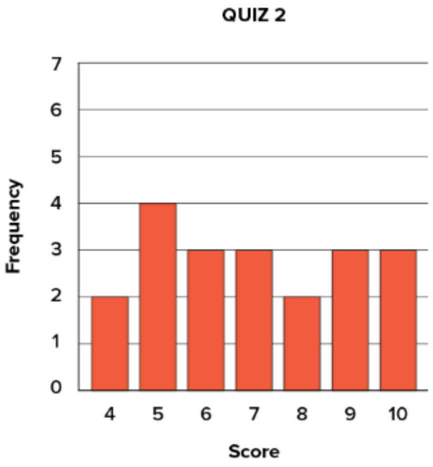
Table 3.1 Iris Dataset and Descriptive Statistics (Fisher, 1936)

Observation	Sepal Length	Sepal Width	Petal Length	Petal Width
1	5.1	3.5	1.4	0.2
2	4.9	3.1	1.5	0.1
...	...	...	...	...
49	5	3.4	1.5	0.2
50	4.4	2.9	1.4	0.2
Statistics	**Sepal Length**	**Sepal Width**	**Petal Length**	**Petal Width**
Mean	5.006	3.418	1.464	0.244
Median	5.000	3.400	1.500	0.200
Mode	5.100	3.400	1.500	0.200
Range	1.500	2.100	0.900	0.500
Standard deviation	0.352	0.381	0.174	0.107
Variance	0.124	0.145	0.030	0.011

Measure of Spread

- **Measure of Spread** : The terms spread, variability and dispersion are synonyms and refer to how spread out a distribution is. To see what we mean by spread out, consider the followings graphs.



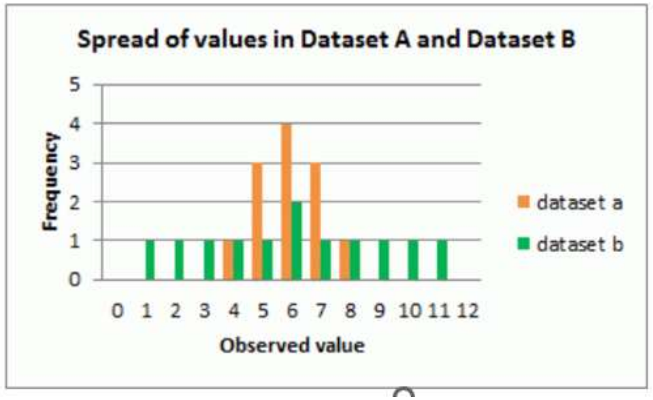


- These graphs represent the scores on two quizzes.
- The mean score for each quiz is 7.0
- Despite the equality of means, you can see that the distributions are quite different.
- Specifically, the scores on Quiz 1 are more densely packed and those on Quiz 2 are more spread out.
- The differences among students were much greater on Quiz 2 than on Quiz 1.

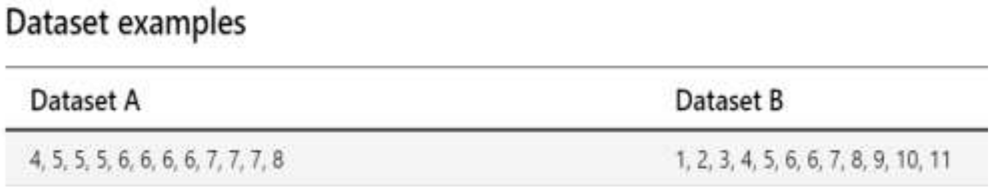
Reasons to measure spread

- The measure of spread (or dispersion) describes how data points are distributed around the center of a dataset. It indicates the variability or consistency within the data, helping to understand whether the data points are closely clustered or widely scattered. Measures of spread summarize the data in a way that shows how scattered the values are and how much they differ from the mean value.

Spread of values in Dataset A and Dataset B



- The mode , median and mean of both datasets is 6.
- If we just looked at the measures of central tendency, we may assume that the datasets are the same.
- However, if we look at the spread of the values in the following graph, we can see that Dataset B is more dispersed than Dataset A.
- Used together, the measures of central tendency and measures of spread help us to better understand the data



Dataset examples

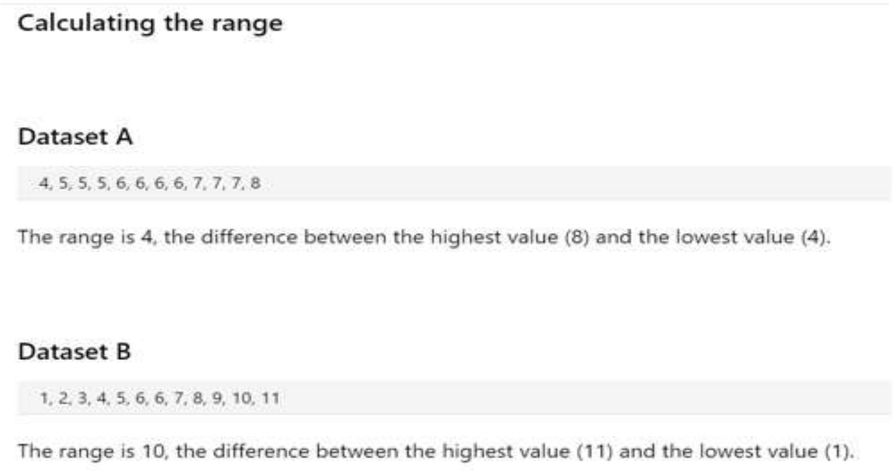
Dataset A	Dataset B
4, 5, 5, 5, 6, 6, 6, 6, 7, 7, 7, 8	1, 2, 3, 4, 5, 6, 6, 7, 8, 9, 10, 11

Measure of Spread

We will discuss measures of the spread of a distribution. There are four frequently used measures of spread: **Range, Interquartile Range(IQR),variance, and standard deviation.**

Range: The range is the difference between the maximum value and the minimum value of the attribute.

we can see below the range of values for Dataset B is larger than Dataset A.



Calculating the range

Dataset A

4, 5, 5, 5, 6, 6, 6, 6, 7, 7, 7, 8

The range is 4, the difference between the highest value (8) and the lowest value (4).

Dataset B

1, 2, 3, 4, 5, 6, 6, 7, 8, 9, 10, 11

The range is 10, the difference between the highest value (11) and the lowest value (1).

Interquartile Range(IQR)

IQR: The Interquartile Range (IQR) is a measure of statistical dispersion and represents the spread of the middle 50% of a dataset. The IQR is the difference between the third quartile (Q3) and the first quartile (Q1).

$$IQR=Q3-Q1$$

Q1:The value below which 25% of the data falls.

Q3: The value below which 75% of the data falls.

Interquartile Range(IQR)

Dataset: {4,4,10,11,15,7,14,12,6,100} find the IQR value of this dataset.

We need to sort the data from least to greatest.

Sorted Dataset:{4,4,6,7,10,11,12,14,15,100}

Q1(First Quartile)= 6

Q3(Third Quartile)=14

IQR=Q3-Q1=14-6=8

Measure of Spread

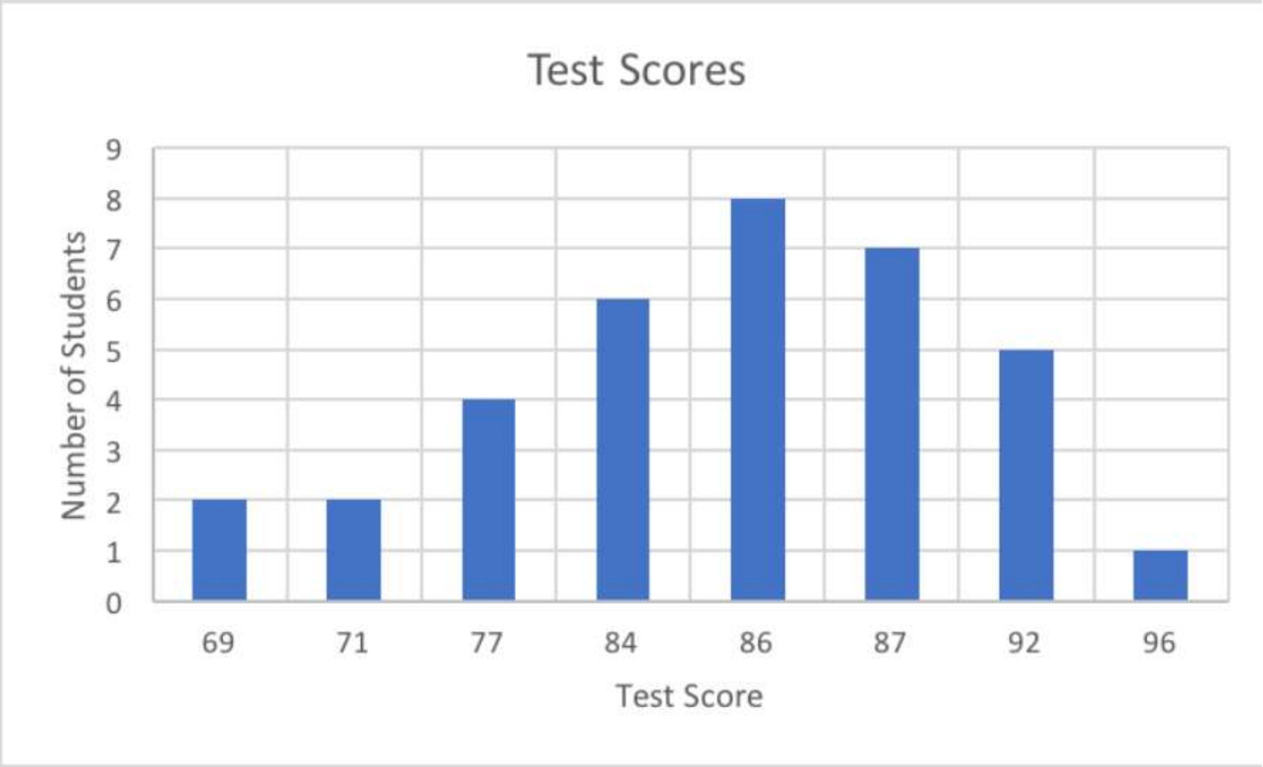
Variance: The variance and standard deviation measures the spread, by considering all the values of the attribute. Deviation is simply measured as the difference between any given value (xi) and the mean of the sample (μ). The variance is the sum of the squared deviations of all data points divided by the number of data points. For a dataset with N observations, the variance is given by the following equation.

:

$$\text{Variance} = s^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$$

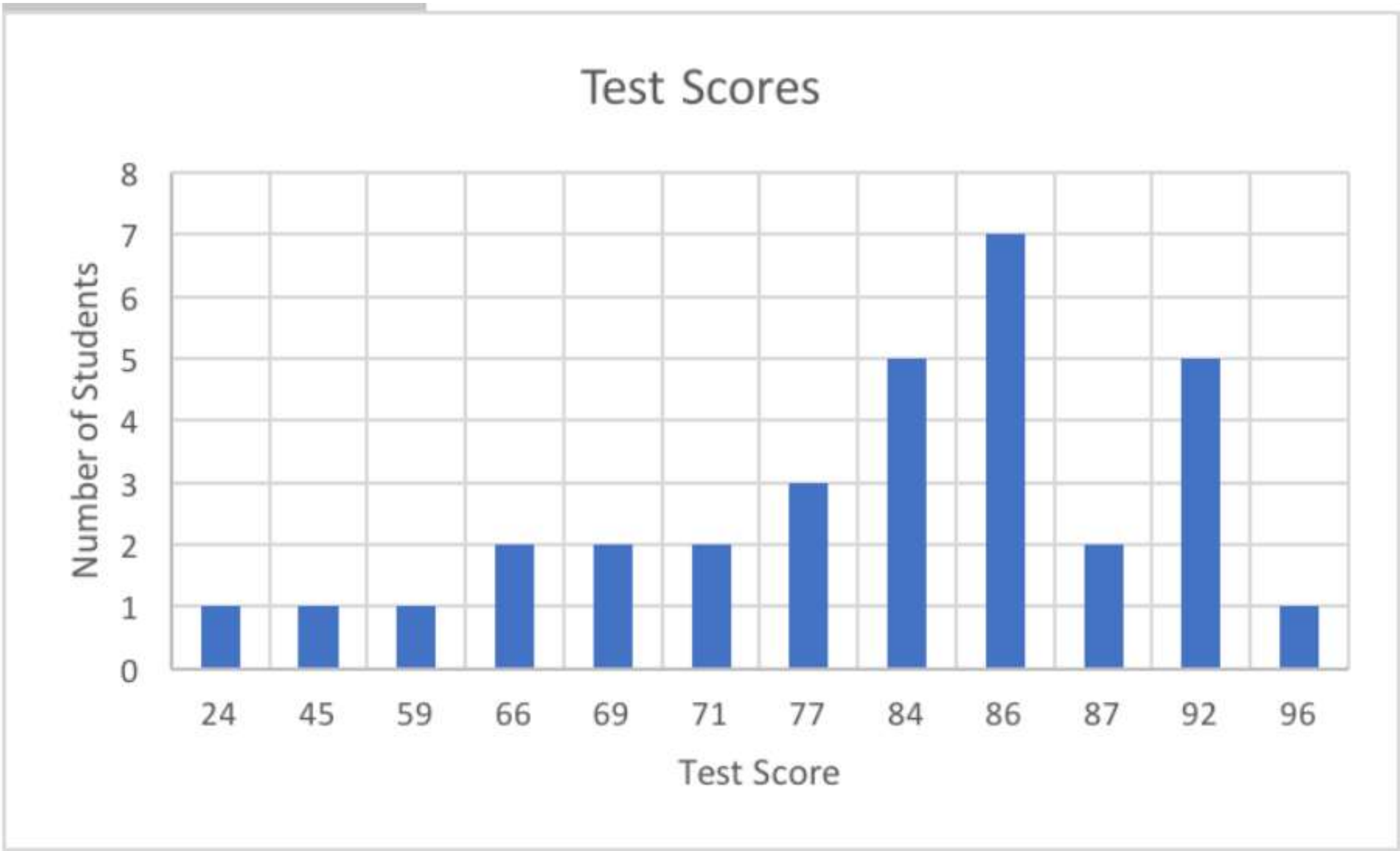
Measure of Spread

Standard Deviation: Standard deviation is the square root of the variance. Standard deviation measures how much your entire data set differs from the mean. High standard deviation means the data points are spread widely around the central point. Low standard deviation means data points are closer to the central point. In the following graph, the mean is 84.47, the standard deviation is 6.92 and the distribution look



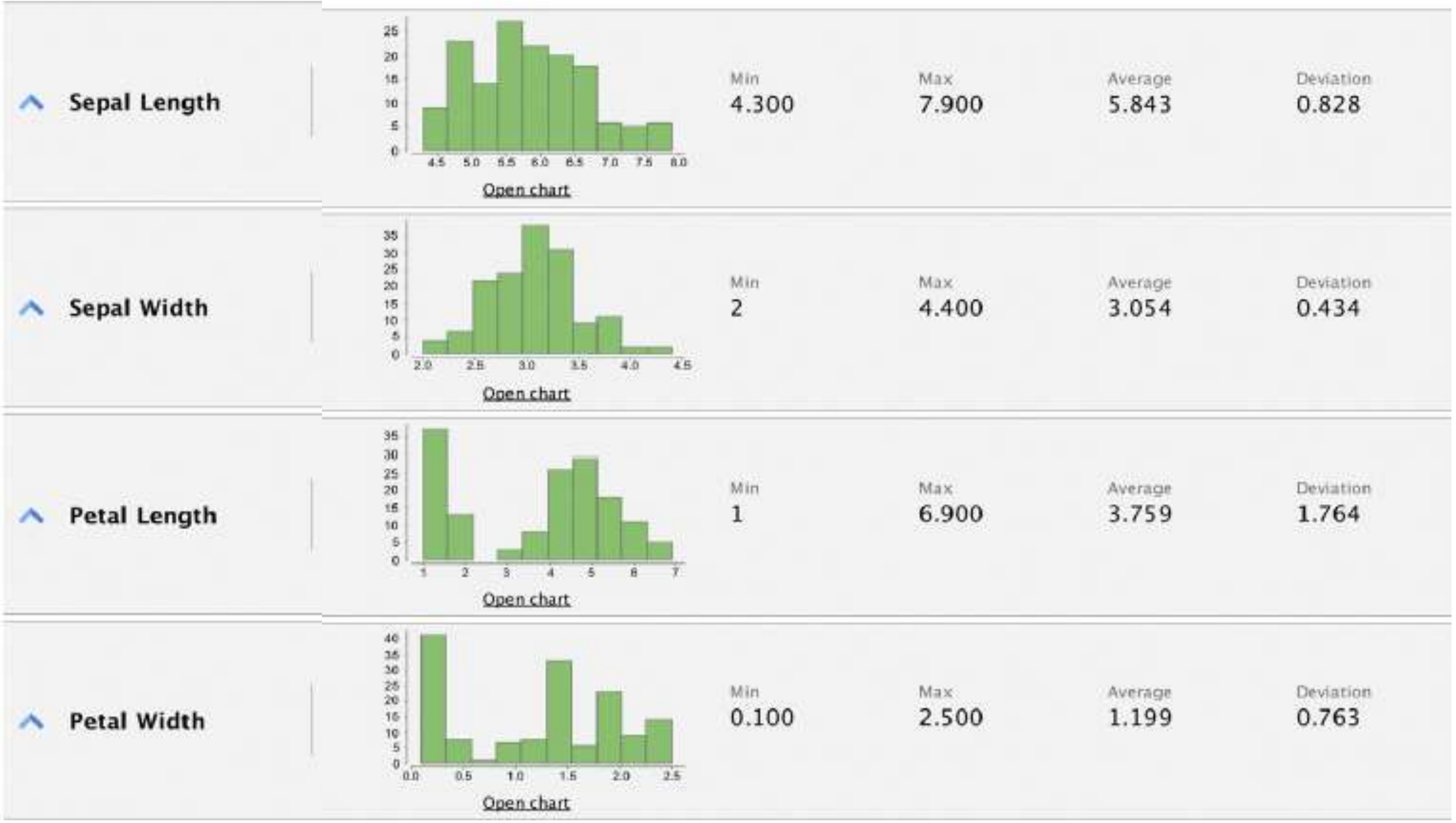
Measure of Spread

In this second graph, the mean is 80, the standard deviation is 14.57 , and the distribution looks like this



Measure of Spread

The following figure provides the univariate summary of the Iris dataset with all 150 observations, for each of the four numeric attributes.



		Min	Max	Average	Deviation
Sepal Length	Open chart	4.300	7.900	5.843	0.828
Sepal Width	Open chart	2	4.400	3.054	0.434
Petal Length	Open chart	1	6.900	3.759	1.764
Petal Width	Open chart	0.100	2.500	1.199	0.763