University of Khartoum

Faculty of Mathematical Sciences and Informatics

A Review Of Quartiles , Percentiles & Deciles

Zeinab Al Fatih Dafallah	**16-116**
Hiba Eldaw babo abdulraheem	**15-124**
Khadiga Gamal Mohi Eldin Mohamed	**16-111**

Supervisor: Dr. Ibrahim Elabid

ABSTRACT:

This study helped in exploring the essential, foundational concepts, the related concepts ,and the manual, software practical applications of quartiles, deciles, and percentiles,which considered the key measures of position in the world of statistical analysis. These measures are invaluable for understanding data distribution, identifying outliers, and making informed decisions across various fields such as healthcare, finance, and education. The research consolidates definitions, methodologies, and applications of these measures, offering a comprehensive guide for researchers, analysts, and decision-makers, helping them find an integrated, useful reference to their research. Using the R programming language, the study illustrates how to compute these measures efficiently, demonstrating their importance in data visualization, benchmarking, and comparative analysis. This work aims to bridge the gap between theoretical knowledge and real-world implementation of positional statistics.

TABLE OF CONTENTS:

LIST OF FIGURES:

CHAPTER ONE

INTRODUCTION

1.1 Overview and background:

In this chapter, we will introduce the reader this research under study, including a summary of the research, research motivation and objectives that must be achieved, the study aim, assumptions and limits.

Measures of position:

There are some values that divides the distribution into **n** equal parts, these values we call them measures of position. A Measure of position (or measure of location or fractile) is a value in which a specified fraction or percentage of the observations in a given set must fall.
They are used to describe or determine the position or the relative standing of a data point within a data set (a single value in relation to other values in a sample, a population data set).
A measure can tell us whether a data point or a value is about the average , or whether it's usually high or low.
They are used for Quantitative data that falls on some numerical scale, sometimes they can be applied to ordinal variables that have an order , like first , second .. etc.
measures of position can also show how values from different distributions or measurement scale are compared (for example: a person's height (measured in feet) , and weight (measured in pounds) can be compared by converting the measurements to z- scores).

Common measures of positions:

- Box and whiskers Plot
- Deciles
- Five number summary
- Interquartile range (IQR)
- Outliers
- Percentiles
- Quartiles
- Standard score (z-scores).
- Tukey's upper hinge and lower hinge.

In this research we're going to talk about three of these measures, which give us away to see where a certain data point or value falls in a sample or distribution.

Quartiles are three values that split sorted data into four parts , each with an equal number of observations , the first quartile (Q1) is the 25th percentile ,the second quartile (Q2) is the median or 50th percentile , and the third quartile (Q3) is the 75th percentile.
Where ***Percentiles*** divide a dataset into 100 equal parts .each represents 1% of the data .for Example: the 25th percentile (P25) is the same as Q1 , and the 50th percentiles (P50) is the median.
And ***deciles*** divide a dataset into ten equal parts .each decile represents 10% of the data . For For example , the first decile (D1) is the 10th percentile, the second decile (D2) is the 20th percentile , and so on.

1.2 Research motivation:

Understanding quartiles, percentiles, and deciles is of the essence in today's data-driven world. These statistical measures play an essential role in summarizing and interpreting data, enabling analysts , researchers, decision makers to attain insights into distributions and trends.
These measures are fundamental in identifying outliers, and patterns in datasets, which can have a great impact on a lot of fields such as healthcare, finance, education.

1.3 Research aim:

Our research aim is to collect the information, definitions and the methods of calculating quartiles, deciles and percentiles in one place , to provide the information that's needed to answer important questions.

1.4 Research question:

- ❖ How do quartiles, deciles, and percentiles differ in their definitions and application in statistical analysis ?
- ❖ What are the advantages and limitations of using quartiles, deciles and percentiles ?
- ❖ In what ways do quartiles , deciles and percentiles influence decision making processes in various industries ?

1.5 Research objectives:

The objectives of our research:

- ➢ Describing the distribution of a dataset
- ➢ Identifying the position of a value within the distribution
- ➢ Facilitating decision-making and analysis

- Enabling data visualization and interpretation
- Facilitating comparisons and benchmarking
- Identifying outliers and anomalies.

1.6 Research contribution:

The main contribution of our research to the statistical field is in that: quartiles help researchers in identifying outliers anomalies in a dataset, which is important for understanding the distribution of data, where deciles provide a more detailed view of the distribution of a dataset than quartiles, allowing anyone who is interested in identifying the patterns and trends that exist in the studied datasets. And percentiles are helping researchers extract the values that correspond to specific percentages of the data.

Each of these three measures provide different perspectives for data distribution, while using them together allows us to gain a more comprehensive understanding of the data.

CHAPTER TWO

THEORETICAL FRAMEWORK

In this chapter we are going to introduce the basic concepts and concepts related to our research that give us a deep understanding about it.

2.1 Overview:

Data is a simple record or collection of different numbers, characters, images, and others that are processed to form information. In statistics, we have different types of data that are used to represent various information. We analyze the data to obtain any meaningful information and thus categorizing data into different types is very important. Data types in statistics help us to make informed decisions about what type of process is used to analyze the data.

2.1.1 Types of data:

Different data require different methods of summarising, describing and analysing. There are four main types of data: nominal, ordinal, interval and ratio. It is important to be able to identify which type of data you have in order to choose appropriate statistical methods.

Nominal: nominal data are named variables. Nominal data is unordered, categorical and mutually exclusive – which means that each category is separate and can not occur at the same time.

Examples of nominal data are:

- different countries: Afghanistan, Brazil, China, etc.
- yes or no answers. When a variable only takes two values, we call this dichotomous.

Ordinal: ordinal data are named variables that have a meaningful order. Ordinal data is ordered, categorical and mutually exclusive (cannot happen at the same time).

Examples of ordinal data are:

- Body mass index (BMI): underweight, normal weight, overweight, obese.
- Response to questionnaires: strongly disagree, disagree, neutral, agree and strongly agree. This is known as a likert scale, and is often coded using numbers.

Interval: interval data is quantitative (numbered data) that has a meaningful interval between data points but does not have a meaningful zero, where zero means nothing.

Example of interval data:

- Temperature (in degrees celsius, 0 degrees celsius is cold, not no temperature)
- size of shoes (EU size 0? this shoe size does not exist).

Ratio: Ratio data is quantitative (numbered data) that has a meaningful interval between data points. Unlike interval data, ratio data has a meaningful zero – known as a true zero.

Examples of ratio data:

- Length (in cm, where 0 cm has no length).
- Weight (in kg, where 0 kg has no weight).

Additional statistical terminology that is used to summarise the types of data detailed above:

1. Quantitative data: is data that can be quantified. This can be as simple as reporting a percentage of yes to no responses or more complex, reporting if results are statistically significantly different. These pages focus on quantitative data.

Where Qualitative data is further categorized into two categories that includes:

- Discrete data
- Continuous data

where,

- ***Discrete data*** only takes certain values. For example, the number of patients in a trial, the role of a dice 1, 2, 3, 4, 5 or 6.
- ***Continuous:*** continuous data is numbered that can take any value within a range. Examples include, height (cm) weight (kg) and race times (seconds).

2. Qualitative data: refers to data that represent opinions. This data cannot be captured as a quantity, but can be used to provide deeper insight into a topic. Examples of qualitative data are open-ended questions, responses from focus groups, interviews and observation.
And Further, the qualitative data is categorized into:

- nominal data
- ordinal data

We also have,
Categorical: categorical data that can either be ordered or unordered. Nominal and ordinal data are often referred to as categorical data.
Scale: scale data is used to describe numerical data that has a meaningful scale. Interval and ratio data are often referred to as scale data.

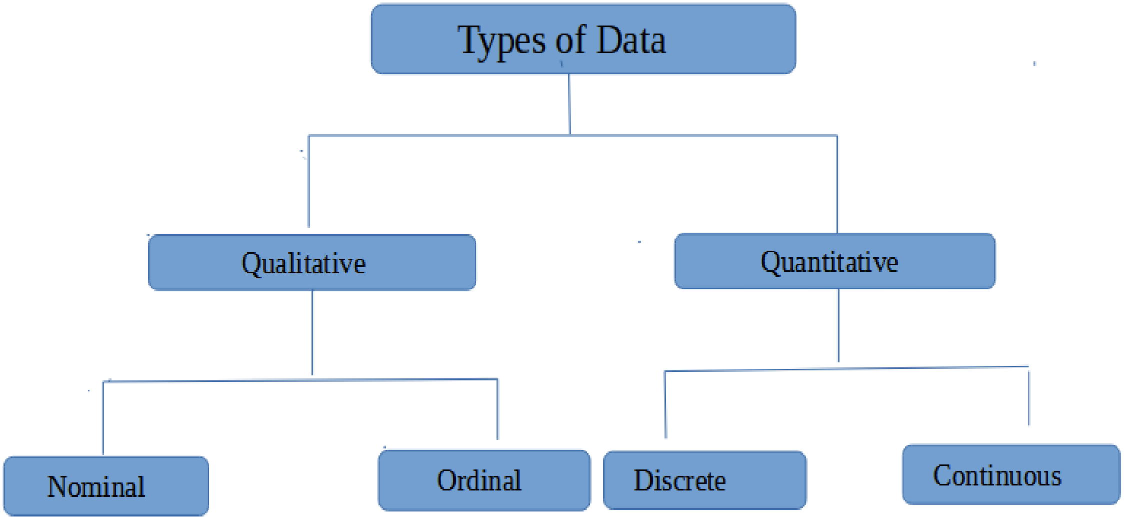


Figure No.1 - Types Of Data

2.2 Consistency and homogeneity of data:

2.2.1 Consistency:

In statistics, consistency of procedures, such as computing confidence intervals or conducting hypothesis tests, is a desired property of their behaviour as the number of items in the data set to which they are applied increases indefinitely. In particular, consistency requires that as the dataset size increases, the outcome of the procedure approaches the correct outcome.
Use of the terms consistency and consistency in statistics is restricted to cases where essentially the same procedure can be applied to any number of data items. In complicated applications of statistics, there may be several ways in which the number of data items may grow. For example records for rainfall within an area might increase in three ways: records for additional time periods; records for additional sites with a fixed area; records for extra sites obtained by extending the size of the area. In such cases, the property of consistency may be limited to one or more of the possible ways a sample size can grow.

Tests: a consistent test is one for which the power of the test for a fixed untrue hypothesis increases to one as the number of data items increases.
Classification: in statistical classification, a consistent classifier is one for which the probability of correct classification, given a training set, approaches, as the size of the training set increases, the best probability theoretically possible if the population distributions were fully known.

2.2.2 Homogeneity and heterogeneity:

in statistics, homogeneity and its opposite, heterogeneity, arise in describing the properties of a dataset, or several datasets. They relate to the validity of the often convenient assumption that the statistical properties of any part of an overall dataset are the same as any other part. In **meta-analysis**, which combines the data from several studies, homogeneity measures the differences or similarities between the several studies.
Homogeneity can be studied to several degrees of complexity. For example, considerations of homoscedasticity examine how much the variability of the data-values changes throughout a dataset. However, questions of homogeneity apply to all aspects of the statistical distributions, including the location parameter. Thus , a more detailed study would examine changes to the whole of the marginal distribution. An intermediate-level study might move from looking at the variability to studying changes in the skewness.
The concept of homogeneity can be applied in many different ways and, for certain types of statistical analysis, it is used to look for further properties that might need to be treated as varying within a dataset once some initial types of non-homogeneity have been dealt with.
Tests: a test for homogeneity, in the sense of exact equivalence of statistical distributions, can be based on an E-statistic. A location test tests the simpler hypothesis that distributions have the same location parameter.

In the general sense, homogeneity refers to items that are alike, identical, and equal. On the other hand, heterogeneity refers to things that are different, distinct, unlike, and non-equivalent. Homogeneous items can be thought of as consistent, regular, and uniform, while heterogeneous items are assorted, diverse, and mixed.
Homogeneity and heterogeneity have slightly more nuanced definition [1]:

A homogeneous population has a uniform character, where all the set elements are believed to be the same or similar in nature or can be altered to make those characteristics. Homogeneous populations can be combined or mixed together in different ways.

A heterogeneous population is diverse in nature, composed of different and dissimilar elements that are, and should be, separate.

In homogeneous sampling, a purposive sampling technique , is where all items in the sample share similar or identical traits or that are relevant to the research, such as age, location, or size. To illustrate, suppose you want to examine the impact of a new drug on a group of adults. In this case, you could use homogeneous sampling to create a sample of adults who share the same, age, gender, and race. This ensures that the sample is representative of the entire population, thus minimizing bias in the study's findings.

On the other hand, maximum variation sampling, where a researcher deliberately creates samples made up of items with regard to a specific characteristic such as age, income or size. For example, drug safety testing studies are often performed on adults who represent a wide range of ages, weights, and races. The FDA requires that drug safety testing studies include a diverse set of participants.

Homogeneous samples are typically smaller and consist of similar cases, such as 20 people who are overweight. On the other hand, a heterogeneous sample is made up of diverse characteristics. Such as a mixture of 18 to 80-year olds of different heights, weights and races.

From a data analysis standpoint, a data set is homogeneous when it is made up of variables of the same type, such as all binary variables or all categorical variables. A mixed set, such as one composed of binary and categorical variables is heterogeneous.

To determine whether or not a data set is homogeneous, data sets can be compared using boxplots or descriptive statistics such as variance, standard deviation, and interquartile range. Some statistical tests are specifically designed for homogeneity assessment. These tests are crucial for various data analyses, as many hypothesis tests assume some level of data homogeneity; for example, an ANOVA test assumes equal variances between populations.

One popular test of homogeneity is the chi-square test of homogeneity, which looks at whether two populations come from the same unknown distribution, determining whether they are homogeneous or not. The test follows the standard chi-square test procedure, where the X2 statistic is calculated and the null hypothesis- that the data come from the same distribution- is either accepted or rejected.

Homogeneity of variance:

homogeneity of variance (also called *homoscedasticity*) is used to describe data with the same variance. On scatterplot, homoscedastic data will have the same scatter. If data does not have the same variance, it will show a dissimilar scatter pattern.

2.3 Measures of Central Tendency & Dispersion

The abundance of data types (descriptive and quantitative in all their forms) has led to the development of advanced statistical methods for analyzing this data in all its forms, whether numerical or descriptive.Statistics has become very important measures in all types of data and is used by the largest educational, governmental, financial, electronic and private institutions. These measures are:

1.Measures of Central Tendency.
2.Measures of Dispersion.
3.Frequency Measures.
4.Measures of Relative Position.

In this research, we will discuss the most important statistical measures in the case of a large number of data, which are:

2.3.1 Measures of Central Tendency:

“measures of central tendency, the several types of average yield numbers or words that attempt to describe, most generally, the middle or typical value for a distribution. There are three different measures of central tendency—the mode, median, and mean. Each of these has its own uses, but the mean is the most "important average in both descriptive and inferential statistics” [John Wiley & Sons, 2017, 2010, 2007-p.47]
As shown in the following figure (Figure No. 1)

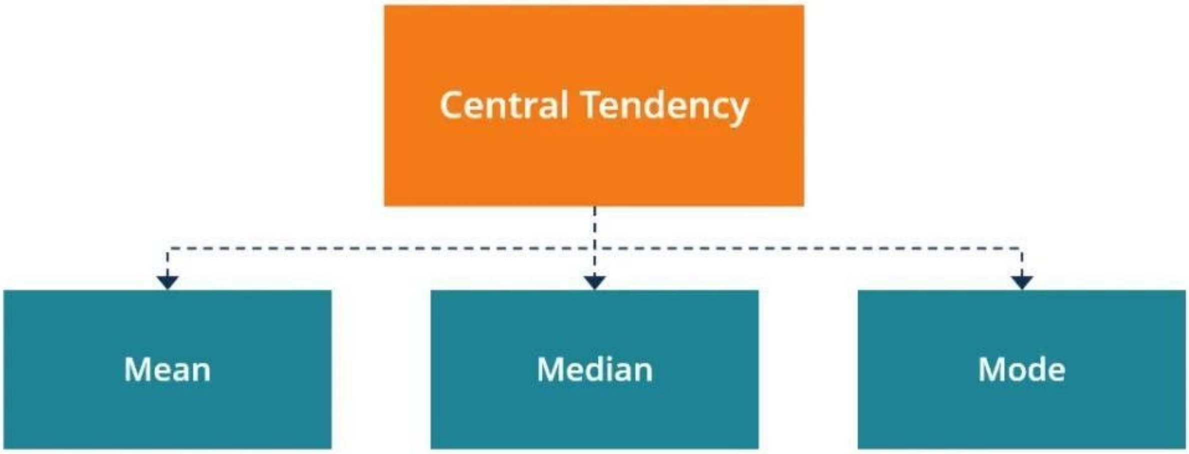


Figure No.2
Measures of Central Tendency

MODE:
It is the most frequently occurring element.

More Than One Mode:
Distributions can have more than one mode (or no mode at all). Distributions with two obvious peaks, even though they are not exactly the same height, are referred to as bimodal.

Example:

Table 1-1 shows the stock prices of the top 20 American companies in one year:

Table 1-1

$14.25	$19.00	$11.00	$28.00
$24.00	$23.00	$43.25	$19.00
$27.00	$25.00	$15.00	$7.00
$34.22	$15.50	$15.00	$22.00
$19.00	$19.00	$27.00	$21.00

Solution:

We arrange the elements from Table 1-1 in ascending order:

7.00 11.00 14.25 15.00 15.00 15.50 19.00 19.00 19.00 19.00
21.00 22.00 23.00 24.00 25.00 27.00 27.00 28.00 34.22 43.25

Now we choose the most frequently repeated element among them:

7.00 11.00 14.25 15.00 15.00 15.50 **19.00 19.00 19.00 19.00**
21.00 22.00 23.00 24.00 25.00 27.00 27.00 28.00 34.22 43.25

This grouping makes it easier to see that 19.00 is the **most frequently** occurring number

MEDIAN:

The median is the middle when observations are ordered from least to most . For an array with an odd number of terms, the median is the middle number. For an array with an even number of terms, the median is the average of the two middle numbers.

Example:

Suppose a business researcher wants to determine the median for the following numbers.

15 11 14 3 21 17 22 16 19 16 5 7 19 8 9 20 4

Solution:

ordered the numbers:

3 4 5 7 8 9 11 14 **15** 16 16 17 19 19 20 21 22

Because the array contains 17 terms (an odd number of terms), the median is the middle number, **or 15.**

If the number 22 is eliminated from the list, the array would contain only 16 terms.

3 4 5 7 8 9 11 **14 15** 16 16 17 19 19 20 21

Now, for an even number of terms, the statistician determines the median by averaging the two middle values**, 14 and 15**. The resulting median value is 14.5.

MEAN:

The arithmetic mean is the average of a group of numbers and is computed by adding all scores and then dividing by the number of scores. Because the arithmetic mean is so widely used, most statisticians refer to it simply as the mean.

The population mean is represented by the Greek letter mu (μ). The sample mean is represented by x^{-}. The formulas for computing the population mean and the sample mean are given in the boxes that follow. As shown in **figure No. 2**

The capital Greek letter sigma (Σ) is commonly used in mathematics to represent a summation of

POPULATION MEAN

$$\mu = \frac{\Sigma x}{N} = \frac{x_1 + x_2 + x_3 + \cdots + x_N}{N}$$

SAMPLE MEAN

$$\bar{x} = \frac{\Sigma x}{n} = \frac{x_1 + x_2 + x_3 + \cdots + x_n}{n}$$

all the numbers in a grouping. Also, **N** is the number of terms in the population, and **n** is the number of terms in the sample.

Example:

Suppose a company has five departments with 24, 13, 19, 26, and 11 workers each. What is the population mean number of workers in each department.

Solution:

$$\sum_{\square}^{\square} x^{-} = 24+13+19+26+11=93$$

$$\mu = \frac{\sum_{\square}^{\square} x^{-} = \frac{93}{5} = 18.6}{\square}$$

Relationship Between *MEAN*, *MEDIAN & MODE* Under Different Skewness
As shown in the following figure (Figure No. 2)

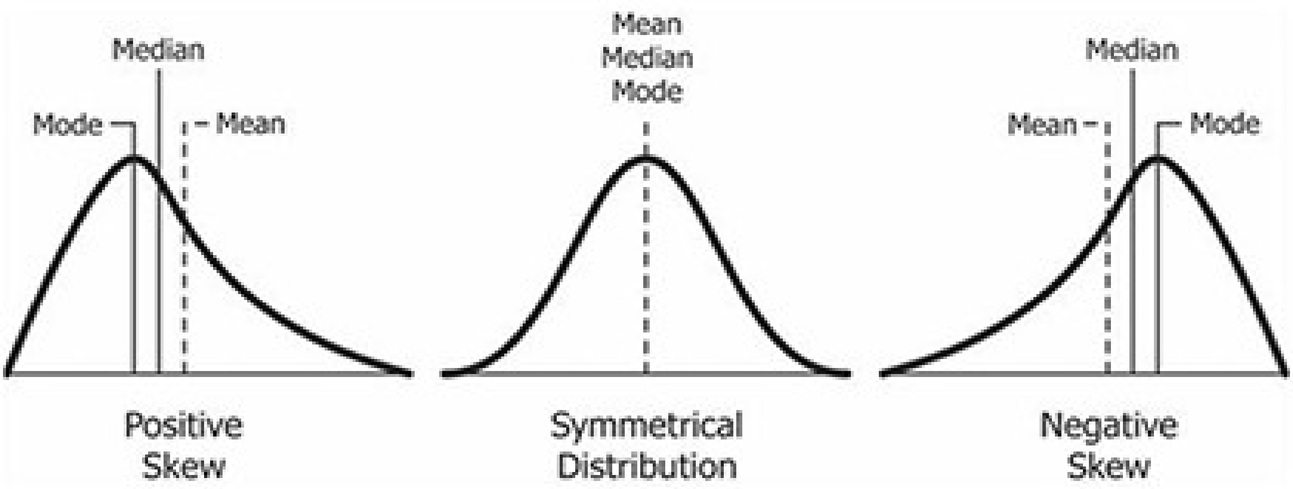


Figure No. 3

Relationship Between *MEAN, MEDIAN & MODE*

2.3.2 Measures of Dispersion:

“measures of variability, that is, measures of the amount by which scores are dispersed or scattered in a distribution. This chapter describes several measures of variability, including the range, the, the variance, and most important, the standard deviation.”[John Wiley & Sons, 2017, 2010, 2007-p.47]

Numerical measures used to measure the variance or dispersion of data. The variance or dispersion of a set of data is the amount of separation, divergence, or spread of the data among them. The variance of the data is small if the data are close to each other and vice versa. As for the data that is equal, there is neither variance or dispersion in it. shown in the following figure (Figure No. 3)

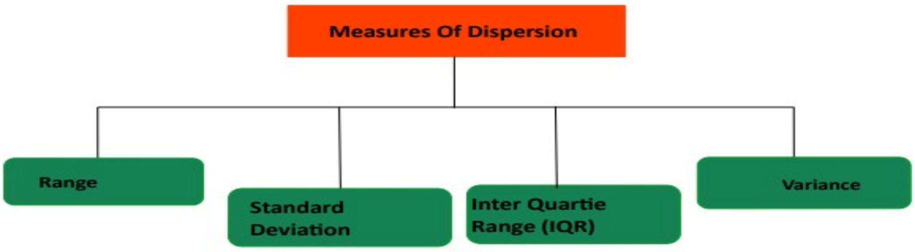


Figure No.4

Measures Of Dispersion

RANGE:

The difference between the highest and lowest values in a dataset.

Range=Maximum value−Minimum value

Example:

The data in **Table 1-1** represent the offer prices for the 20 largest U.S. initial public offerings in a recent year. The lowest offer price was $7.00 and the highest price was $43.25.

Solution:

The range of the offer prices can be computed as the difference of the highest and lowest values:

Range = Highest - Lowest = $43.25 - $7.00 = $36.25

VARIANCE:

The average of the squared differences between each data point and the mean.

For a population:

$$\sigma^{2}=\frac{\sum\left(x_i-\mu\right)2}{N}$$

For a sample:

$$s^{2}=\frac{\sum\left(x_i-x^{-}\right)2}{n-1}$$

Example:

Suppose a small company started a production line to build computers. During the first five weeks of production, the output is 5, 9, 16, 17, and 18 computers, respectively. Which descriptive statistics could the owner use to measure the early progress of production? In an attempt to summarize these figures, the owner could compute a mean.

Solution:

Table 2.2

Number (x)	**Deviations from the Mean (x -µ)**	$(x-\mu)^2$
5	5-13=-8	64
9	9-13=-4	16
16	16-13=+3	9
17	17-13=+4	16
18	18-13=+5	25
Σx=65	**Σ(x-µ)=0**	$\sum_{\square}^{\square} (x-\mu)^2 = 130$

Deviations from the Mean for Computer Production

$$variance = \sigma^2 = \frac{SS_x}{N} = \frac{\Sigma (x-\mu)^2}{N} = \frac{130}{5} = 26.5$$

STANDARD DEVIATION:

The square root of the variance, providing a measure of dispersion in the same units as the data.

For a population:

$\sigma = \sqrt{\square}$

For a sample:

$s = \sqrt{\square}$

Example:

From the previous example:

Solution:

$$variance = \sigma = \sqrt{\square}$$

INTERQUARTILE RANGE:

The range between the first quartile (Q1) and the third quartile (Q3), representing the middle 50% of the data.

IQR=Q3−Q1

Example:

The following data indicate the top 15 trading partners of the United States in exports in a recent year according to the U.S. Census Bureau.

Country	Exports ($ billions)	Country	Exports ($ billions)
Canada	213.1	**Netherlands**	**30.5**
Mexico	119.4	France	25.8
China	61.0	Taiwan	24.8
Japan	58.1	Singapore	23.6
United Kingdom	45.4	Belgium	23.0
Germany	44.3	Brazil	21.7
South Korea	33.0	Australia	17.9
Netherlands	**30.5**	India	16.3

Solution:

For $Q_1 = P_{25}$when N=15:

$$i = \frac{25}{100}(15) = 3.75$$

Because i is not a whole number,P_{25} is found as the fourth term from the bottom.

$$Q_1 = P_{25} = 23.0$$

for $Q_3 = P_{75}$:

$$i = \frac{75}{100}(15) = 11.25$$

Because i is not a whole number, P_{75} is found as the 12th term from the bottom.

$$Q_3 = P_{75} = 58.1$$

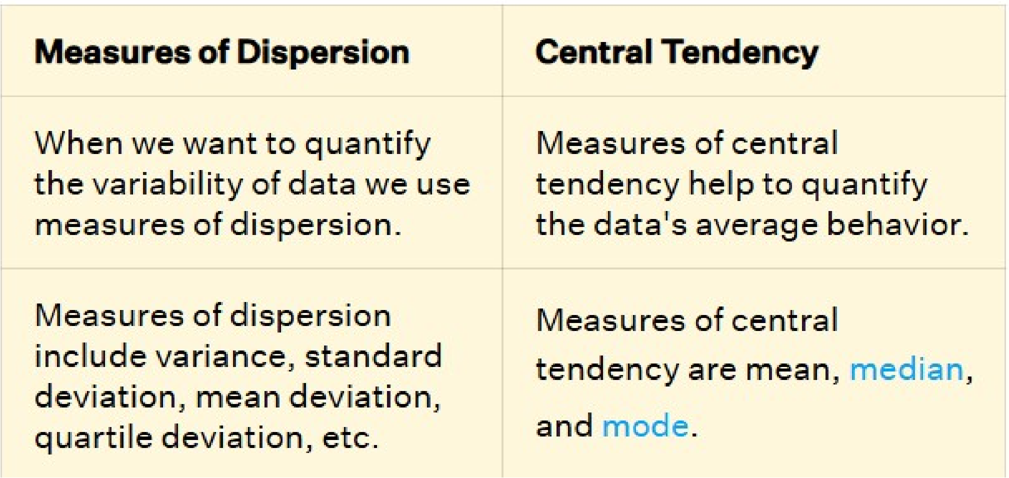
The interquartile range is:

$$Q_3 - Q_1 = 58.1 - 23.0 = 35.1$$

The middle 50% of the exports for the top 15 U.S. trading partners spans a range of 35.1 ($ billions).

Measures of Dispersion and Central Tendency:

Both measures of dispersion and measures of central tendency are used to describe data. **Figure No.4** given below outlines the difference between the measures of dispersion and central tendency.



Measures of Dispersion	Central Tendency
When we want to quantify the variability of data we use measures of dispersion.	Measures of central tendency help to quantify the data's average behavior.
Measures of dispersion include variance, standard deviation, mean deviation, quartile deviation, etc.	Measures of central tendency are mean, median, and mode.

Figure No.5

2.3.3 Merits and demerits of both measures of central tendency and dispersion:

Merits:

- Ease of understanding and interpretation.
- Summarize data.
- Measure data diversity.
- Determine the stability or volatility of data

Demerits:

- Affected by extreme values.
- Complexity.
- Limited range (in case of unique values).
- Inappropriate for some applications.

2.4 Why use measures of position or location instead of measures of central tendency and measures of dispersion?

Measures of position, tendency and dispersion serve different goals in the analysis of data in general, and to choose between them depends on the specific goals wanted to be achieved in the analysis process.
Here are some reasons why measures of position might be used instead of the other measures:

- Percentiles, quartiles, and deciles indicate where a particular value lies within a data distribution. These three measures are very useful when someone wants to understand how individual data points compare, for example: determining the 90th percentile score on an exam shows the score below which 90% of the exam takers fall.
- Sometimes, when measures of central tendency like mean, median and mode do not provide enough details, because they summarise the whole data set with only a single value, and also they don't explain the distribution’s shape and the ranks. While for measures of dispersion such as range, standard deviation and variance which describe variability but never tell about relative position of data points. For example: if the mean of a population is known is does not tell where specific individual stands relative to other individuals.
- When data is not distributed symmetrically, measures like mean, maybe misleading, should be using a measure of position like percentiles, because they are more robust to skewed distributions.

- Extreme values (minimum and maximum values in a dataset) can be outliers, which deviate from the central part of the dataset significantly, so we use measures of position to identify and evaluate these outliers.

2.5 Introduction to Quartiles, Deciles and Percentiles:

The origins of quartiles, deciles and percentiles can be traced back to the development of descriptive statistics over many centuries. In 1890s the use of quartiles and percentiles began to formalize, the term “percentile” was first introduced by Francis Galton, who explored the distribution of human characteristics in his work on eugenics and statistics, where in 1900 the concept of percentiles was introduced in Pearson’s statistical writings further advanced the analysis of data distributions. Although Galton played a foundational role, the terms and concepts were further developed by statisticians building on his work.

2.5.1 Quartiles:

There are three quartiles, i.e. Q1, Q2 and Q3 which divide the total data into four equal parts when it has been orderly arranged. Q1, Q2 and Q3 are termed as first quartile, second quartile and third quartile or lower quartile, middle quartile and upper quartile, respectively.
The first quartile, Q1, separates the first one-fourth of the data from the upper three-fourths and is equal to the 25th percentile. The second quartile, Q2, divides the data into two equal parts (like median) and is equal to the 50th percentile. The third quartile, Q3, separates the first three-quarters of the data from the last quarter and is equal to 75th percentile. Where,

- The first quartile, denoted Q1 , divides the bottom 25% of the data from the top 75%. Therefore, the first quartile is equivalent to the 25th percentile.
- The second quartile, Q2 , divides the bottom 50% of the data from the top 50%; it is equivalent to the 50th percentile or the median.
- The third quartile, Q3 , divides the bottom 75% of the data from the top 25%; it is equivalent to the 75th percentile.

In Other Words
The first quartile, Q1
is equivalent to the 25th
percentile, P25.
The 2nd quartile, Q2
is equivalent to
the 50th percentile, P50
which is equivalent to the median,M.
Finally, the third quartile,
Q3 , is equivalent to the 75th
percentile, P75.

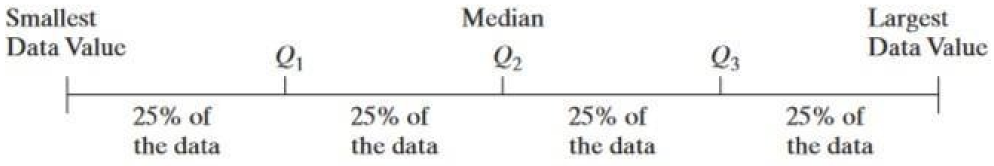


Figure No.6
Illustrate The Concept Of Quartiles.

Quartiles Formula:

The quartile formula helps in calculating the value that divides a list of numbers into quarters. The data is firstly arranged into ascending order and then divided into quartiles.

- 1st quartile is also known as the lower quartile
- 2nd quartile is the same as the median dividing data into 2 equal parts
- 3rd quartile is also called the upper quartile; the interquartile range is calculated as
- upper quartile – lower quartile.

using the following formulas we can calculate each quartile, for the ungrouped data:

- First quartile Q1 = ((n+1) x 1/4)th term
- Second quartile Q2 , or the median = ((n+1) x 2/4) th term
- Third quartile Q3 = ((n+1)x ¾) th term

where for the grouped data we use the following formula:

Example:

find the quartiles Q1, Q2, and Q3 of the following data 20, 30, 25, 23, 22, 32, 36

Solution:

Arrange data in ascending form, and n = 7 odd number

20
22
23
25
30
32
36

Using:

$Q_n = Kn/4$

$q_1 = (1/4) \text{ x } n$

$= (1/4) \text{ x } 7 = 1.75$

$q_1 = 2$

$Q_1 = 22$

$q_2 = (2/4) \text{ x } n$

$= (2/4) \text{ x } 7 = 3.5$

$q_2 = 4$

$Q_2 = 25$

$q_3 = (3/4) \text{ x } n$

$= (3/4) \text{ x } 7 = 5.25$

$q_3 = 6$

$Q_3 = 32$

Example:

find the quartiles Q1, Q2, and Q3 of the following data.

Columns Load	Frequency (fi)
50 - 69	3
70 - 89	7
90 - 109	4
110 – 129	4
130 – 149	9

Solution:

1- find the cumulative frequency and the summation of frequencies and real interval limit.

Column Load	Frequency (fi)	Cumulative frequency	Real Interval
50 - 69	3	3	49.5 – 69.5
70 - 89	7	10	69.5 – 89.5 → Q_1
90 - 109	4	14	89.5 – 109.5
110-129	4	18	109.5 – 129.5
130-149	9	27	129.5 – 149.5

2- Find the arrangement number of quartiles to find quartile interval 1.

q1 = (1/4)x n = (1/4)x 27 = 6.75.

The interval of quartile number 1 has the cumulative frequency = 10 using this formula:

$$Q1 = a + \left[\frac{q1 - n_1}{f_q}\right].C$$

where,

a = the real lower limit of quartiles interval = 69.5

n_1 = the cumulative frequency of the previous interval of the quartiles interval = 3

f_q = the frequency of quartiles interval = 7

c = the length of quartiles interval = 20.

therefore,

Q_1 = 69.5 + [(6.75-3)/7].20 = 80.2

3. find the arrangement number of quartiles to find quartile interval 2.

q_2 = (2/4)x n = (2/4)x 27 = 13.5.

The interval of quartile number is have the cumulative frequency = 14

Q_1 = 89.5+[(13.5-10)/4].20 =107

4. find the arrangement number of quartiles to find quartile interval 3.
q_3 = (3/4)x n = (3/4)x 27 = 20.25
The interval of quartile number 3 is have the cumulative frequency = 27
Q_1 = 129.5 +[(20.25-18)/9].20 = 134.5

Column load	Frequency (fi)	Cumulative frequency	Real interval
50 - 69	3	3	49.5 – 69.5
70-89	7	10	69.5 – 89.5
90-109	4	14	89.5 – 109.5
110-129	4	18	109.5 – 129.5
130-149	**9**	**27**	**129.5 – 149.5**

Finding Quartiles:

Step 1 Arrange the data in ascending order.
Step 2 Determine the median, M, or second quartile, Q2 .
Step 3 Divide the data set into halves: the observations below (to the left of) M and the observations above M. The first quartile, Q1 , is the median of the bottom half of the data and the third quartile, Q3 , is the median of the top half of the data.

In Other Words
To find Q2, determine the median of the data set. To find Q1, determine the median of the "lower half" of the data set. To find Q3, determine the median of the "upper half" of the data set.

Examples:

Problem The Highway Loss Data Institute routinely collects data on collision coverage claims. Collision coverage insures against physical damage to an insured individual's vehicle. The data in Table 16 represent a random sample of 18 collision coverage claims based on data obtained from the Highway Loss Data Institute. Find and interpret the first, second, and third quartiles for collision coverage claims.

Table 1					
$6751	$9908	$3461	$2336	$21,147	$2332
$189	$1185	$370	$1414	$4668	$1953
$10,034	**$735**	**$802**	**$618**	**$180**	**$1657**

Approach: Follow the steps given above.

Solution:
Step 1 The data written in ascending order are given as follows:

$180	**$189**	**$370**	**$618**	**$735**	**$802**	**$1185**	**$1414**	**$1657**
$1953	**$2332**	**$2336**	**$3461**	**$4668**	**$6751**	**$9908**	**$10,034**	**$21,147**

Step 2 There are n = 18 observations, so the median, or second quartile, Q2 , is the mean of the 9th and 10th observations. Therefore, M = Q2 = ($1657 + $1953)/ 2 = $1805.

Step 3 The median of the bottom half of the data is the first quartile, Q1 . As shown next, the median of these data is the 5th observation, so Q1 = $735.

$180 $189 $370 $618 $735 $802 $1185 $1414 $1657
↑
Q1

note: If the number of observations is odd, do not include the median when determining Q1 and Q3 by hand.

The median of the top half of the data is the third quartile, Q3 . As shown next, the the median of these data is the 5th observation, so Q3 = $4668.

$1953 $2332 $2336 $3461 $4668 $6751 $9908 $10,034 $21,147
↑
Q3

Interpretation Interpret the quartiles as percentiles. For example, 25% of the collision claims are less than or equal to the first quartile, $735, and 75% of the collision claims are greater than $735. Also, 50% of the collision claims are less than or equal to $1805, the second quartile, and 50% of the collision claims are greater than $1805. Finally, 75% of the collision claims are less than or equal to $4668, the third quartile, and 25% of the collision claims are greater than $4668.#

Check a Set of Data for Outliers:

When performing any type of data analysis, we should always check for extreme observations in the data set. Extreme observations are referred to as outliers. Outliers can occur by chance, because of error in the measurement of a variable, during data entry, or from errors in sampling. For example, in the 2000 presidential election, a a precinct in New Mexico accidentally recorded 610 absentee ballots for Al Gore as 110. Workers in the Gore camp discovered the data-entry error through an analysis of vote totals.

Outliers do not always occur because of error. Sometimes extreme observations are common within a population. For example, suppose we wanted to estimate the mean price of a European car. We might take a random sample of size 5 from the population of all European automobiles. If our sample included a Ferrari F430 Spider approximately $175,000), it probably would be an outlier, because this car costs much more than the typical European automobile. The value of this car would be considered unusual because it is not a typical value from the data set.

Use the following steps to check for outliers using quartiles.

Checking for Outliers by Using Quartiles

Step 1 Determine the first and third quartiles of the data.
Step 2 Compute the interquartile range.
Step 3 Determine the fences. Fences serve as cutoff points for determining outliers.

Lower fence = Q1 - 1.5(IQR2)
Upper fence = Q3 + 1.5(IQR2)

Step 4 If a data value is less than the lower fence or greater than the upper fence, it is considered an outlier.

Example:

Problem Check the collision coverage claims data in Table 16 for outliers.
Approach Follow the preceding steps. Any data value that is less than the lower fence or greater than the upper fence will be considered an outlier.
Solution
Step 1 The quartiles found in Example 3 are Q1 = $735 and Q3 = $4668.
Step 2 The interquartile range, IQR, is
IQR = Q3 - Q1
= $4668 - $735
= $3933

Step 3 The lower fence, **LF**, is

LF = Q1 - 1.5(IQR2)
= $735 - 1.5($39332)
= - $5164.5

The upper fence, **UF**, is

UF = Q3 + 1.51IQR 2
= $4668 + 1.51$39332
= $10,567.5

Step 4 There are no observations below the lower fence. However, there is an observation above the upper fence. The claim of $21,147 is an outlier.

Merits of Quartiles:

- Ease of Calculation: Quartiles are straightforward to compute, especially in individual and discrete series.
- Open End Intervals: They can be directly determined in cases of open-end class intervals without needing the exact limits.
- Graphical Representation: Quartiles can be located both graphically and tabularly, aiding in data visualization.
- Robustness to Outliers: They are less affected by extreme values compared to other measures like the mean.

Demerits of Quartiles:

- Limited Data Utilization: Quartiles do not consider all data points, potentially overlooking detailed variations.
- Order Requirement: Data must be arranged in ascending or descending order, which can be time-consuming for large datasets.
- Sampling Variability: They may exhibit less stability across different samples, affecting reliability.

2.5.2 Deciles:

A decile is a quantitative method of splitting up a set of ranked or arranged data into 10 equally large subsections or parts.
There are nine deciles i.e. D_1, D_2, D_3….., D_9 and 5th decile is same as median or Q_2, because it divides the data in two equal parts.
In descriptive, a decile is used to categorize large data sets from the highest to lowest values, or vice versa. Like the quartile and the percentile, a decile is a form of a quantile that divides a set of observations into samples that are easier to analyze and measure.
A decile is usually used to assign decile ranks to a data set. A decile rank arranges the data in order from lowest to highest and is done on a scale of one to 10 where each successive number corresponds to an increase of 10 percentage points. In other words there are nine deciles points that have 10% of the observations below it, D2 has 20% of the observations below it, D3 has 30% of the observations falling below it, and so on.

Decile Class Rank:

To split the given data and order it according to some specified metric, statisticians use the decile rank also known as decile class rank.
Once the given data is divided into deciles then each subsequent data set is assigned a decile rank. Each rank is based on an increase by 10 percentage points and is used to order the deciles in an increasing order. The 5th decile of a distribution will give the value of the median..

Deciles Formula:

Ungrouped data : D(X) = (n +1) * x/10
grouped data : D(x)= I +w/f (Nx/10 - C)

The decile formulas can be used to calculate the deciles for grouped and ungrouped data. When data is in its raw form it is known as ungrouped data. When this data is sorted and organized then it forms grouped. These are given as follows:

Deciles formula for ungrouped data: **D(X)** = Value of the decile that needs to be calculated and ranges from 1 to 9. n is the total number of observations in the data set.

Deciles formula for grouped data: **D(x) = I + w/f (Nx/10 – C)**.

where:
I is the lower boundary of the class containing the decile given by (x x cf)/10 , cf is the cumulative frequency of the entire data set
w is the size of the class
N is the total frequency
C is the cumulative frequency of the preceding class.

Example:
suppose a data set consists of the following numbers: 24, 32, 27, 32, 23, 6, 45, 80, 59, 63, 36, 54, 36, 72, 55, 51, 32, 56, 33, 42, 55, 30. the value of the first two deciles has to be calculated. The steps required are as follows:

- **Step 1**: arrange the data in increasing order. This gives 23,24 , 27, 30, 32, 32, 32, 33, 36,36, 42, 45, 51, 54, 55, 55, 56, 57, 59, 62 63, 72, 80.
- **Step 2**: identify the total number of points, here, n = 23
- **Step 3**: apply the decile formula to calculate the position of the required data point. D(1) = (n+1)/ 10 = 2.4. This implies the value of the 2.4th data point has to be determined. This will lie between the scores in the 2nd and 3rd positions. In other words, the 2.4th data is 0.4 of the way between the scores 24 and 27
- **Step 4**:the value of the decile can be determined as [lower score + (distance) (higher score – lower score)]. This is given as 24 + 0.4 * (27 – 24) = 25.2
- **Step 5**: apply steps 3 and 4 to determine the rest of the exiles. D(2) = 2(n+1)/10 = 4.8th data between digit number 4 and 5. thus, 30 + 0.8 * (32-30) = 31.6

Example:

the table below shows the ungrouped scores (out of 100) for 30 exams takers:

48	52	55	57	58	60	61	64	65	66
69	72	73	75	76	78	81	82	84	87
88	90	91	92	93	94	95	96	97	99

Solution: using the information presented in the table, the 1st decile can be calculated as:

D1 = value of [(30+1)/10] th data
= value of 3.1st data which is 0.1 of the way between scores 55 and 57
= 55+2(0.1) = 55.2 = D1
D1 means that 10% of the data set falls below 55.2.

now calculating 3rd decile:

- D3 = Value of 3 (30+1)/10
  = Value of 9.3rd position which is 0.3 between the scores of 65 and 66
- thus, D3 = 65 +1 (0.3) = 65.3
- 30% of the 30 scores in the observation fall below 65.3.

Notes:

- A decile is a quantile that is used to divide a data set into 10 equal subsections.
- The 5th decile will be the median for the data set.
- The decile formula for ungrouped data is given as x(n+1)/10 th term in the data set.
- The decile formula for grouped data is given by I + w/f (Nx/10 – C).

<u>Example</u>: Find the 6th and 9th decile for the data in the above-mentioned example.

Solution: the arranged data is 23, 24, 27, 30, 32, 32, 32, 33, 36, 36, 42, 45, 45, 51, 54, 55, 55, 56, 57, 59, 62, 63, 72, 80

n = 23
D(6) = 6(n+1)/10 = 14.4th data. This lies between 54 and 55.
= 54 + 0.4 * (55-54) = 54.4

D(9) = 9(n+1)/10 = 21.6th data. This lies between 63 and 72

D(9) = 63 + 0.6 *(72-63) = 68.4

Answer: D(5) = 54.4 and D(9) = 68.4

Merits of Deciles:

- Detailed Data Segmentation: Deciles provide a more granular view of data distribution compared to quartiles, dividing data into ten equal parts.
- Enhanced Analysis: They offer a more detailed view of data distribution compared to quartiles or percentiles.

Demerits of Deciles:

- Calculation Complexity: Determining deciles can be more complex, especially with large datasets.
- Data Ordering: Similar to quartiles, data must be ordered, which can be cumbersome for extensive datasets.
- Limited Algebraic Treatment: Deciles are not easily amenable to further algebraic manipulation, limiting their use in advanced statistical analyses.

2.5.3 Percentiles:

Percentiles are the values which divide the arranged data into hundred equal parts.
There are 99 percentiles i.e. $P_{1,}$ P_2, P_3, ……..,P_{99}. The 50th percentile divides the series into two equal parts and $P_{50} = D_5$ = Median.
Similarly the value of $Q_1 = P_{25}$ and value of Q3 = P_{75}
percentiles can be used to detect outliers. When a value is less than the 5th percentile (p_5) or greater than the 95th percentile (p_{95}), it can be categorized as an outlier.
For example, in a group of 20 children, Ben is the 4th tallest and 80% of the children are shorter than you. Hence, it means that Ben is at 80th percentile. It is most commonly use in competitive exams such as SAT, LSAT, etc.

Percentile Formula:

The percentile formula determines the performance of a person over others. A percentile is a number that tells the percentage of scores that fall below the given number.
The percentile formula is used when we need to compare the exact values or numbers over the other numbers from the given data i.e. the accuracy of the number. Often percentile and percentage are taken as one but both are different concepts. A percentage is where the fraction is considered as one term while a percentile of a value is the percentage of the values that are below the given values out of the whole set of values. The next example gives the difference between percentage and percentile. If an exam is conducted out 100 marks, then:

- We say that a student scored 100 “percent” if and only if he had scored 100/100.
- We say that a student scored 100 “percentile” if all the students (100% students) scored less than him.

In our day-to-day life, percentile formulas are usually helpful in grading test scores or biometric measurements. Hence, the percentile formula is;

$$\mathbf{p = (n/N) \times 100}$$

where,

- n = ordinal rank of the given value or value below the numbered
- N = number of values in the data set
- p = percentiles

The percentile of x is the ratio of the number of values below x to the total number of values multiplied by 100. i.e., the percentile formula is

percentile = (number of values below “x” / total number of values) x 100

Calculation of Percentiles:

to calculate the percentile, here are a few steps to use the percentile formula. If q is any number between zero and hundred, the qth percentile is the value that divides the data into two parts i.e the lowest part contains the q percent of the data and the rest of the data is the upper part.

- **Step 1**: collect the data set
- **step 2**: arrange the data set in ascending order
- **step 3**: determine the total number of observations
- **step 4**: identify the data value for which you are interested to find the percentile
- **step 5**: count the number of data values that are less than the above value
- **step 6**: divide the number from step 5 by the number from step 3 to find the percentile of the given data value.

<u>Example:</u> the scores obtained by 10 students are 38, 47, 58, 65, 70, 79, 80, 92. using the percentile formula, calculate the percentile for score 70?

<u>Solution:</u>

given:
scores obtained by students are 38, 47, 58, 49, 58, 60, 65, 70, 80, 92

number of scores below 70 = 6
using the percentile formula

percentile = (number of values below “x” / total number of values) x 100
percentile of 70 = (6/10) x 100 = 0.6 x 100 = 60

therefore the percentile for score 70 = 60%

<u>Example</u>: the weights of 10 people were recorded in kg as 35, 41, 42, 56, 58, 62, 70, 71, 77, 90. How do you find percentile for the weight 58 kg?

<u>Solution</u>:

given:
weight of the people are 35, 41, 42, 56, 58, 62, 70, 71, 77, 90

number of people with weight below 58 kg = 4
using the formula for percentile,
percentile = (number of values below “x” / total number of values) x 100
percentile for weight 58 kg = (4/10) x 100 = 0.4 x 100 = 40%
therefore, the percentile for weight 58 kg = 40%

<u>Example:</u>

In a college, a list of scores of 10 students is announced. The scores are 56, 45, 69, 78, 72, 94, 82, 80, 63, 59. using the percentile formula find the 70th percentile

<u>Solution:</u>

- arrange the data in ascending order :

  45, 56, 59, 63, 69, 72, 78, 80, 82, 94

  find the rank,

  rank = percentile / 100
  rank = 70 /100 = 0.7
  so, the rank is 0.7

  using the formula to calculate the percentile,

  percentile = rank x total number of the data set
  percentile = 0.7 x 10 = 7
  now, counting 7 values from left to right we reach 800, and we can say that all the values below 80 will come under the 70th percentile.
  In other words, 70% of the values are below 80.
  therefore the 70th percentile is 80

Interpreting Percentiles:

Recall that the median divides the lower 50% of a set of data from the upper 50%. The median is a special case of a general concept called the percentile.

Definition: The kth percentile, denoted P_k, of a set of data is a value such that k percent of the observations are less than or equal to the value.

So percentiles divide a set of data that is written in ascending order into 100 parts; Thus 99 percentiles can be determined. For example, P1 divides the bottom 1% of the observations from the top 99%, P_2 divides the bottom 2% of the observations from the top 98%, and so on. Figure displays the 99 possible percentiles.

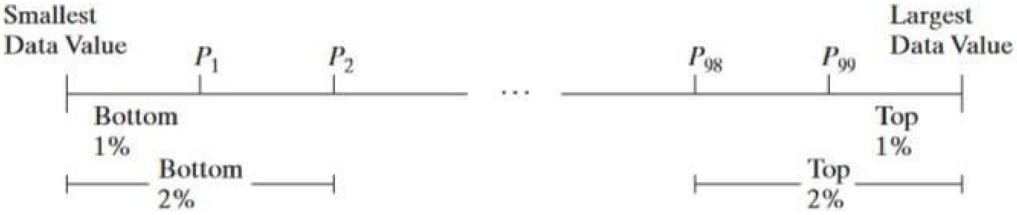


Figure No. 7

Percentiles are used to give the relative standing of an observation. Many standardized exams, such as the SAT college entrance exam, use percentiles to let students know how they scored on the exam in relation to all other students who took the exam.

Example:
Interpret a percentile

Problem: Jennifer just received the results of her SAT exam. Her SAT Mathematics a score of 600 is in the 74th percentile. What does this mean?
Approach:The kth percentile of an observation means that k percent of the observations are less than or equal to the observation.
Interpretation: A percentile rank of 74% means that 74% of SAT Mathematics scores are less than or equal to 600 and 26% of the scores are greater. So 26% of the students who took the exam scored better than Jennifer.

- The first quartile, denoted Q1 , divides the bottom 25% of the data from the top 75%. Therefore, the first quartile is equivalent to the 25th percentile.
- The second quartile, Q2 , divides the bottom 50% of the data from the top 50%; it is equivalent to the 50th percentile or the median.
- The third quartile, Q3 , divides the bottom 75% of the data from the top 25%; it is equivalent to the 75th percentile.

Merits of percentiles:

- High Precision: Percentiles divide data into 100 equal parts, allowing for precise identification of data positions.
- Comprehensive Analysis: They provide a detailed understanding of data distribution, useful in various statistical applications.

Demerits of percentiles:

- Calculation Intensity: Computing percentiles can be labor-intensive, particularly for large datasets.
- Data Arrangement Necessity: Data must be sorted in ascending or descending order, which can be time-consuming.
- Sampling Variability: Percentiles may be influenced by sample fluctuations, affecting their stability.

Merits of Quartiles, Deciles and Percentiles:

- These positional values can be directly determined in case of open end class intervals.
- These positional values can be calculated easily in absence of some data.
- These are helpful in the calculation of measures of skewness.
- These are not affected very much by the extreme items.
- These can be located graphically.

Demerits of Quartiles, Deciles and Percentiles:

- These values are not easily understood by a common man.
- These values are not based on all the observations of a series.
- These values cannot be computed if items are not given in ascending or descending order.
- These values have less sampling stability.

Summary

Partition values divide the data, when arranged in either ascending order or descending order, into different numbers of equal parts. Median is the middle value in a set of arranged data. The place of the median in a series is such that an equal number of items lie on either side of it, i.e. it splits the observations into two halves. We can also say that 50% of the observations lie above median value, while the rest 50% of the observations lie below median value. Quartiles divide the total data into four equal parts when it has been orderly arranged. The first quartile, Q1, separates the first one-fourth of the data from the upper three-fourths and is equal to the 25th percentile. The second quartile, Q2,divides the data into two equal parts (like median) and is equal to the 50th percentile. The third quartile, Q3, separates the first three-quarters of the data from the last quarter and is equal to 75th percentile. Deciles are the partition values which divide the arranged data into ten equal parts whereas percentiles divide the data into hundred equal parts.

CHAPTER THREE

APPLICATIONS OF QUARTILES, DECILES & PERCENTILES IN R

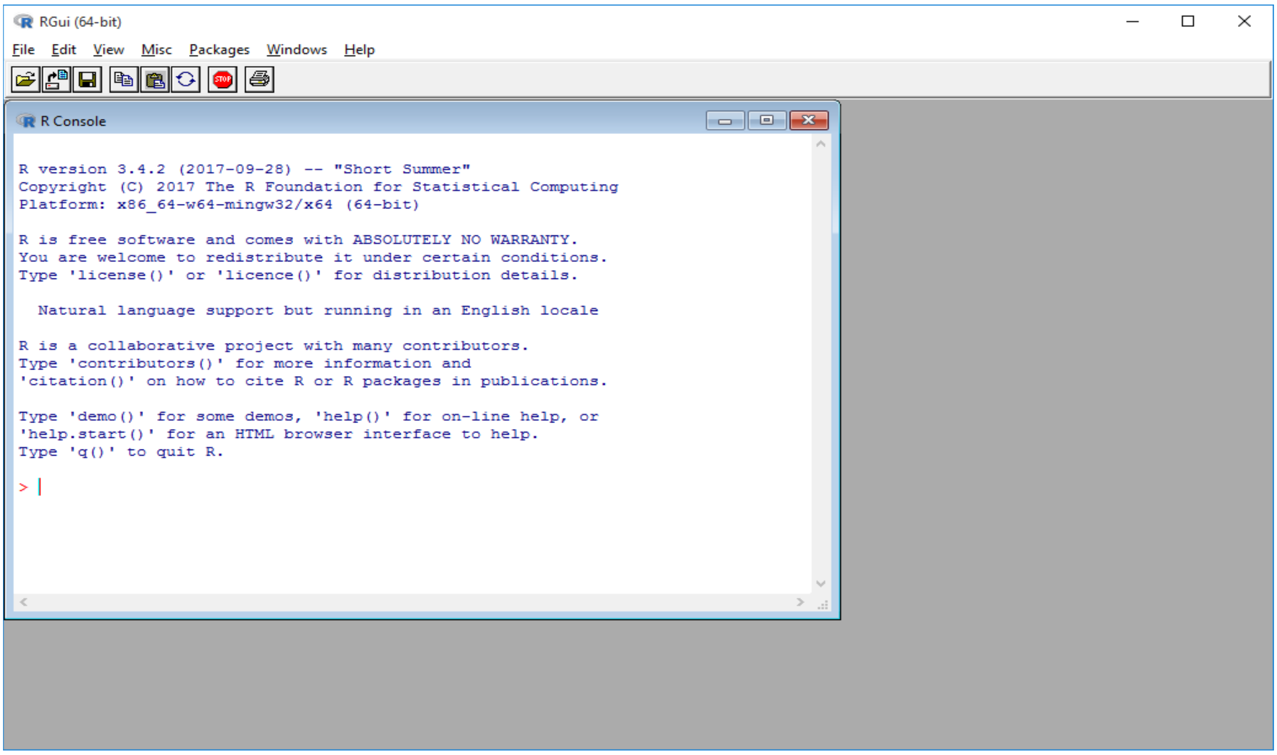
In this chapter, we introduce a background of the R programming language, and how to do calculations and data analysis of quartiles, deciles and percentiles in software using R.

3.1 Background:

R is a wonderful tool for statistical analysis, visualization and reporting. Its usefulness is best seen in the wide variety of fields where it is used.
R is an open-source software package, licensed under the GNU General Public License (GPL). This means that you can install R for free on most desktop and server machines. So we can define R as an integrated suite of software facilities for data manipulation, calculation and graphical display.

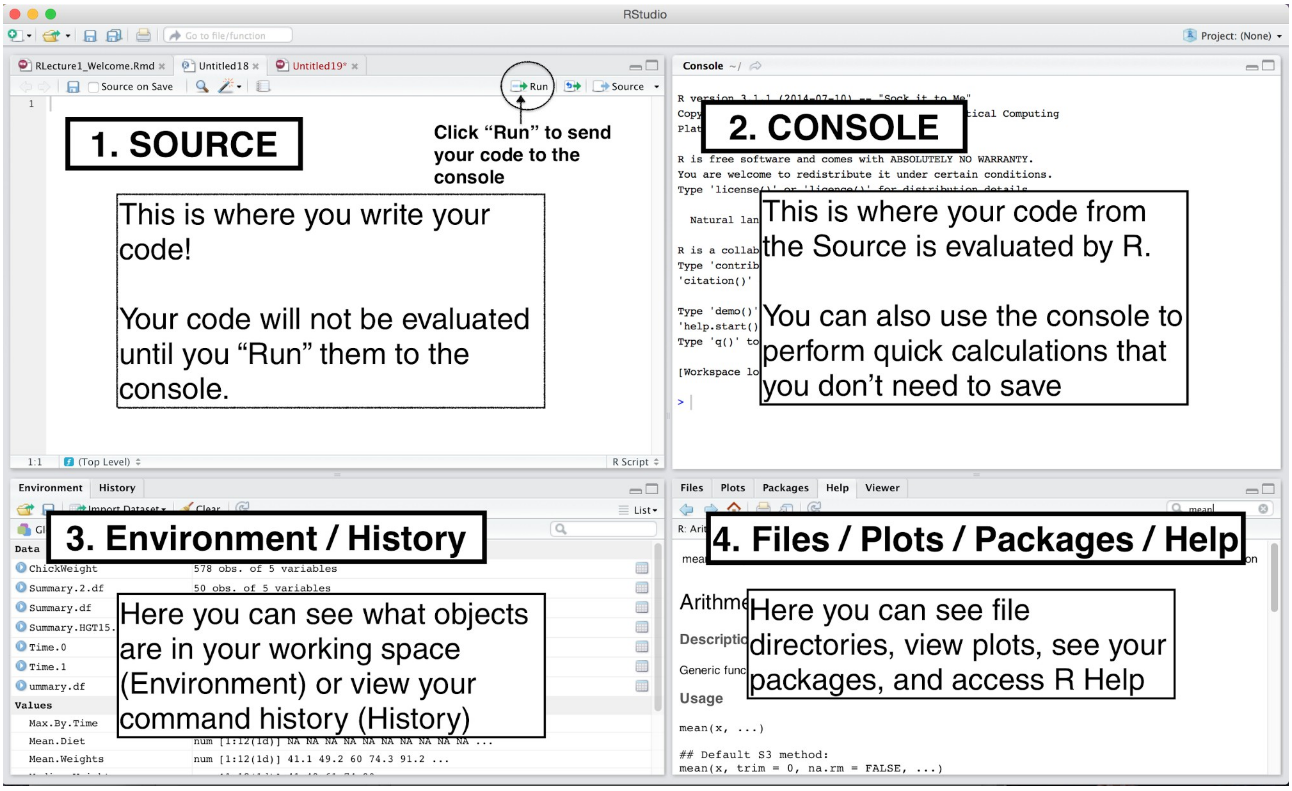
3.2 Uses of R language:

R is used for projects with banks, political campaigns, tech startups, food startups, international development, aid organizations, hospitals, and real estate developers



```
R version 3.4.2 (2017-09-28) -- "Short Summer"
Copyright (C) 2017 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32/x64 (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

  Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> |
```

Figure No.8
The Standard R Interface In Windows.



Click “Run” to send your code to the console

This is where you write your code!

Your code will not be evaluated until you “Run” them to the console.

This is where your code from the Source is evaluated by R.

You can also use the console to perform quick calculations that you don’t need to save

Here you can see what objects are in your working space (Environment) or view your command history (History)

Here you can see file directories, view plots, see your packages, and access R Help

Figure No.9

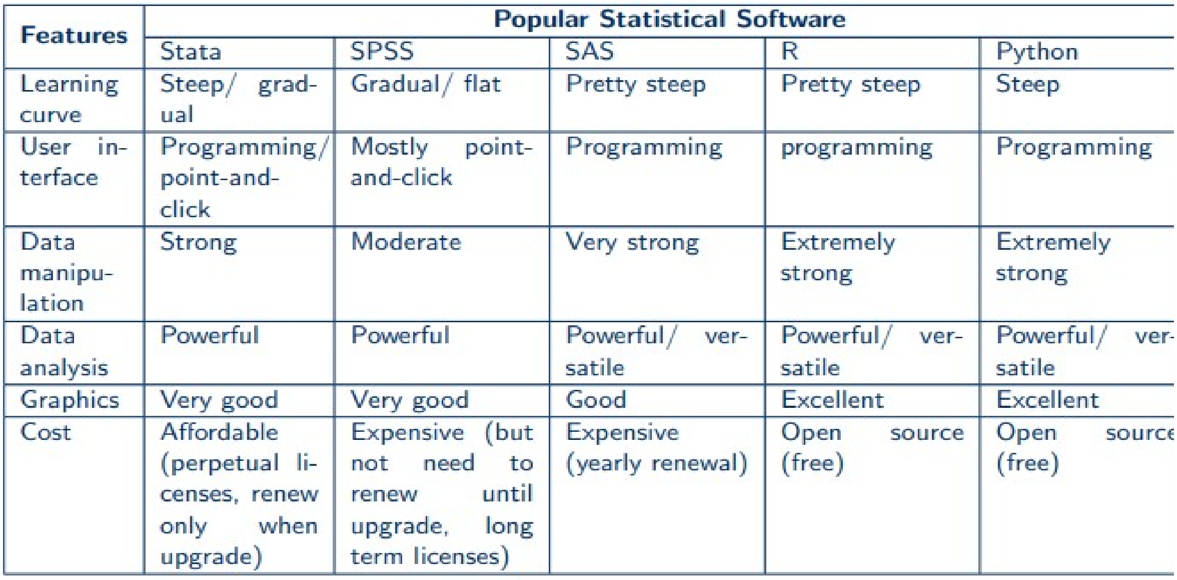
The General Layout Of RStudio.

3.3 R Features:

❑ Effective data handling and storage facility.
❑ A suite of operators for calculations on arrays, in particular matrices.
❑ Large, coherent, integrated collection of intermediate tools for data analysis.
❑ Graphical facilities for data analysis and display either directly at the computer or on hardcopy.
❑ Well developed, simple and effective programming language which includes conditionals, loops, user defined recursive functions and input and output facilities.

3.4 Why did we choose R :

Comparing Some Statistical Softwares:



Features	Popular Statistical Software				
	Stata	SPSS	SAS	R	Python
Learning curve	Steep/ gradual	Gradual/ flat	Pretty steep	Pretty steep	Steep
User interface	Programming/ point-and-click	Mostly point-and-click	Programming	programming	Programming
Data manipulation	Strong	Moderate	Very strong	Extremely strong	Extremely strong
Data analysis	Powerful	Powerful	Powerful/ versatile	Powerful/ versatile	Powerful/ versatile
Graphics	Very good	Very good	Good	Excellent	Excellent
Cost	Affordable (perpetual licenses, renew only when upgrade)	Expensive (but not need to renew until upgrade, long term licenses)	Expensive (yearly renewal)	Open source (free)	Open source (free)

Figure No.10

3.5 Selecting data package in R:

Perhaps the biggest reason for R's phenomenally ascendant popularity is its collection of user contributed packages. As of early February 2017 there were over 10,000 packages available on CRAN1 written by more than 2,000 different people. Odds are good that if a statistical technique exists, it has been written in R and contributed to CRAN. Not only is there an incredibly large number of packages; many are written by the authorities in the field, such as Andrew Gelman, Trevor Hastie, Dirk Eddelbuettel and Hadley Wickham.
A package is essentially a library of prewritten code designed to accomplish some task or a collection of tasks. The survival package is used for survival analysis, ggplot2 is used for plotting and sp is for dealing with spatial data.

When we can calculate Quartiles, Deciles, and Percentiles for any data set that contains numbers (numeric values).
There are many combined data sets that represent the economy and income or can be used in economic analysis that contain numeric data in CRAN such as :

3.6 How calculate the *Quartiles, Deciles & Percentiles* by using *R:*

Firstly:

3.6.1 Preparing R:

Steps:
1-Open R :

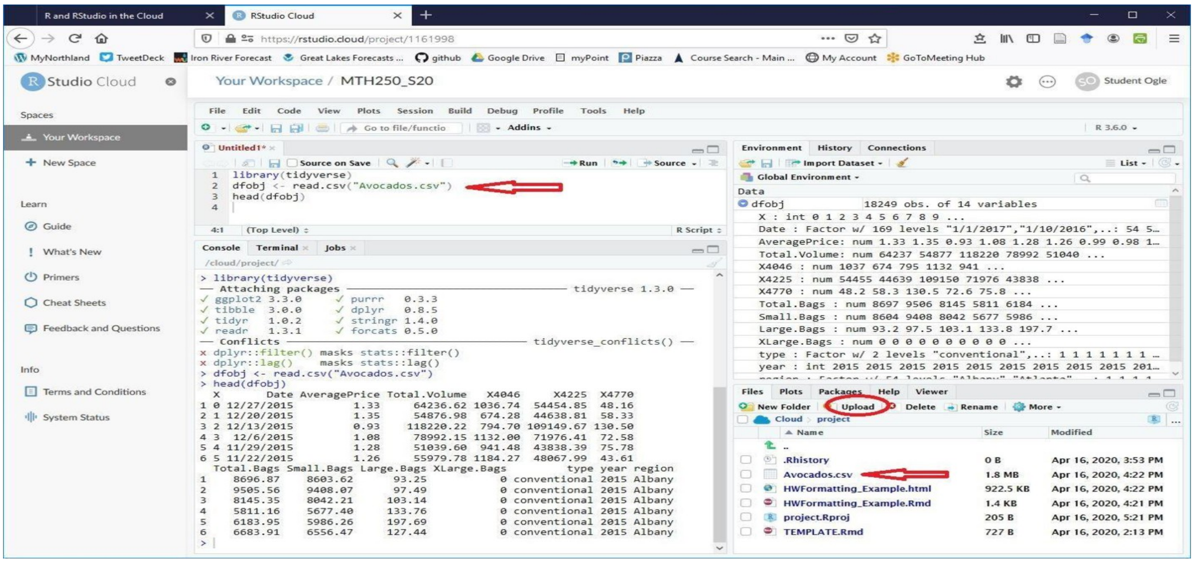
- Install R: Make sure you have R installed on your machine. If not, you can download it from https://cran.r-project.org/ .
- Install RStudio (optional): If you want a user-friendly interface, install RStudio (Here we will use RStudio).

(Return To **Figure 2.2**)

2-Reading Data:

- If you are working with an external dataset, you can import it into R using one of the following commands:

A) Reading CSVs :
Many people also like to use read.csv, which is a wrapper around read.table with the sep argument preset to a comma (,). The result of using read.table is a data.frame.
The first argument to read.table is the full path of the file to be loaded. The file can be sitting on disk or even the Web. **Look at the following figure**



```
library(tidyverse)
dfobj <- read.csv("Avocados.csv")
head(dfobj)
```

```
> library(tidyverse)
── Attaching packages ─────────────────────────── tidyverse 1.3.0 ──
✓ ggplot2 3.3.0     ✓ purrr   0.3.3
✓ tibble  3.0.0     ✓ dplyr   0.8.5
✓ tidyr   1.0.2     ✓ stringr 1.4.0
✓ readr   1.3.1     ✓ forcats 0.5.0
── Conflicts ──────────────────────────────── tidyverse_conflicts() ──
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
> dfobj <- read.csv("Avocados.csv")
> head(dfobj)
  X       Date AveragePrice Total.Volume   X4046     X4225  X4770
1 0 12/27/2015         1.33     64236.62 1036.74  54454.85  48.16
2 1 12/20/2015         1.35     54876.98  674.28  44638.81  58.33
3 2 12/13/2015         0.93    118220.22  794.70 109149.67 130.50
4 3  12/6/2015         1.08     78992.15 1132.00  71976.41  72.58
5 4 11/29/2015         1.28     51039.60  941.48  43838.39  75.78
6 5 11/22/2015         1.26     55979.78 1184.27  48067.99  43.61
  Total.Bags Small.Bags Large.Bags XLarge.Bags         type year region
1    8696.87    8603.62      93.25           0 conventional 2015 Albany
2    9505.56    9408.07      97.49           0 conventional 2015 Albany
3    8145.35    8042.21     103.14           0 conventional 2015 Albany
4    5811.16    5677.40     133.76           0 conventional 2015 Albany
5    6183.95    5986.26     197.69           0 conventional 2015 Albany
6    6683.91    6556.47     127.44           0 conventional 2015 Albany
> |
```

```
Data
dfobj           18249 obs. of 14 variables
   X : int 0 1 2 3 4 5 6 7 8 9 ...
   Date : Factor w/ 169 levels "1/1/2017","1/10/2016",..: 54 5…
   AveragePrice: num 1.33 1.35 0.93 1.08 1.28 1.26 0.99 0.98 1…
   Total.Volume: num 64237 54877 118220 78992 51040 ...
   X4046 : num 1037 674 795 1132 941 ...
   X4225 : num 54455 44639 109150 71976 43838 ...
   X4770 : num 48.2 58.3 130.5 72.6 75.8 ...
   Total.Bags : num 8697 9506 8145 5811 6184 ...
   Small.Bags : num 8604 9408 8042 5677 5986 ...
   Large.Bags : num 93.2 97.5 103.1 133.8 197.7 ...
   XLarge.Bags : num 0 0 0 0 0 0 0 0 0 0 ...
   type : Factor w/ 2 levels "conventional",..: 1 1 1 1 1 1 1 …
   year : int 2015 2015 2015 2015 2015 2015 2015 2015 2015 201…
```

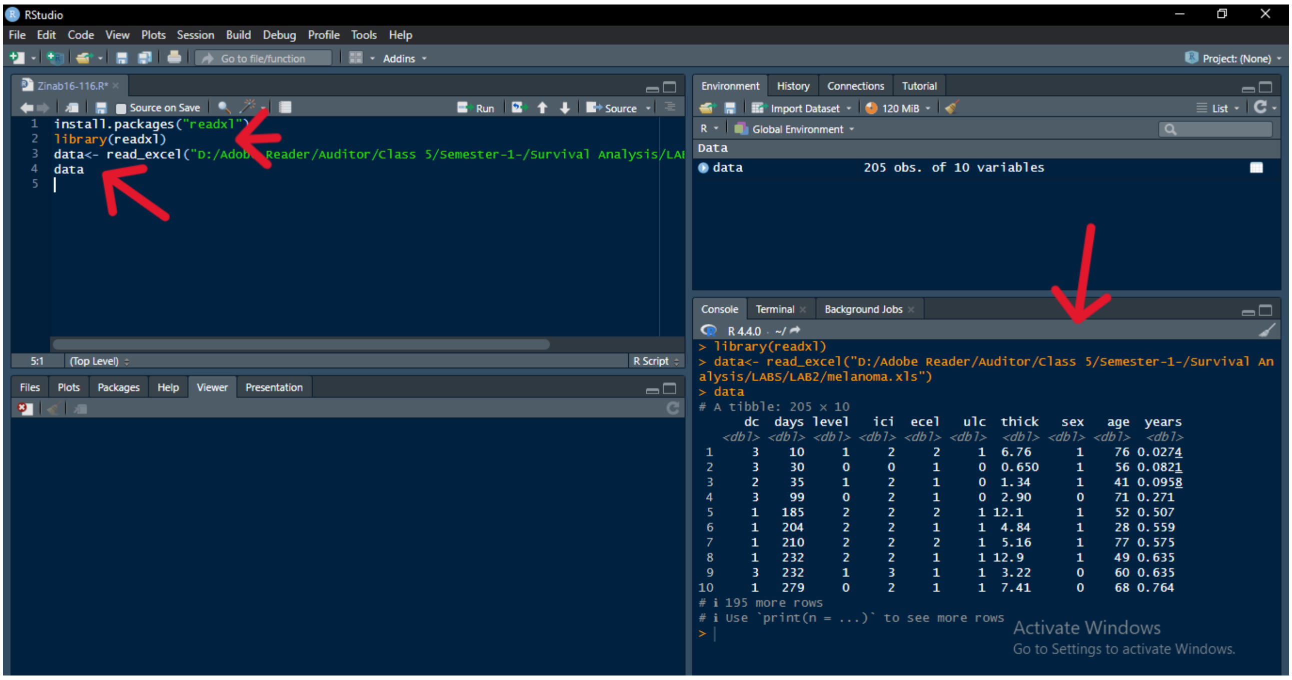
Name	Size	Modified
..		
.Rhistory	0 B	Apr 16, 2020, 3:53 PM
Avocados.csv	1.8 MB	Apr 16, 2020, 4:22 PM
HWFormatting_Example.html	922.5 KB	Apr 16, 2020, 4:22 PM
HWFormatting_Example.Rmd	1.4 KB	Apr 16, 2020, 4:21 PM
project.Rproj	205 B	Apr 16, 2020, 5:21 PM
TEMPLATE.Rmd	727 B	Apr 16, 2020, 2:13 PM

Figure No.11

B) Excel Data:

Excel may be the world’s most popular data analysis tool, and while that has benefits and disadvantages, it means that R users will sometimes be required to read Excel files. Fortunately, for anyone tasked with using Excel data, the package readxl, by Hadley Wickham, makes reading Excel files, both .xls and .xlsx, easy. The main function is read_excel, which reads the data from a single Excel sheet.

And also could do this by visiting a browser or we can stay within R and use download.file.


RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
Project: (None)
Zinab16-116.R*
Source on Save
Run
Source
Environment History Connections Tutorial
Import Dataset
120 MiB
List
Global Environment
Console Terminal Background Jobs
R 4.4.0 · ~/
Files Plots Packages Help Viewer Presentation
(Top Level)
R Script
Activate Windows
Go to Settings to activate Windows.

```
install.packages("readxl")
library(readxl)
data<- read_excel("D:/Adobe Reader/Auditor/Class 5/Semester-1-/Survival Analysis/LA
data
```

```
Data
data                205 obs. of 10 variables
```

```
> library(readxl)
> data<- read_excel("D:/Adobe Reader/Auditor/Class 5/Semester-1-/Survival An
alysis/LABS/LAB2/melanoma.xls")
> data
# A tibble: 205 × 10
      dc  days level   ici  ecel   ulc thick   sex   age  years
   <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  <dbl>
 1     3    10     1     2     2     1  6.76     1    76 0.0274
 2     3    30     0     0     1     0  0.650    1    56 0.0821
 3     2    35     1     2     1     0  1.34     1    41 0.0958
 4     3    99     0     2     1     0  2.90     0    71 0.271
 5     1   185     2     2     2     1 12.1      1    52 0.507
 6     1   204     2     2     1     1  4.84     1    28 0.559
 7     1   210     2     2     2     1  5.16     1    77 0.575
 8     1   232     2     2     1     1 12.9      1    49 0.635
 9     3   232     1     3     1     1  3.22     0    60 0.635
10     1   279     0     2     1     1  7.41     0    68 0.764
# i 195 more rows
# i Use `print(n = ...)` to see more rows
> |
```

Figure No.12

C) Reading from Databases & Embedded Data :

1. Reading data from databases, which are data stored externally and accessed by connecting to databases such as PostgreSQL, MySQL, etc., using packages in R such as RPostgreSQL and RMySQL or via the ODBC interface.

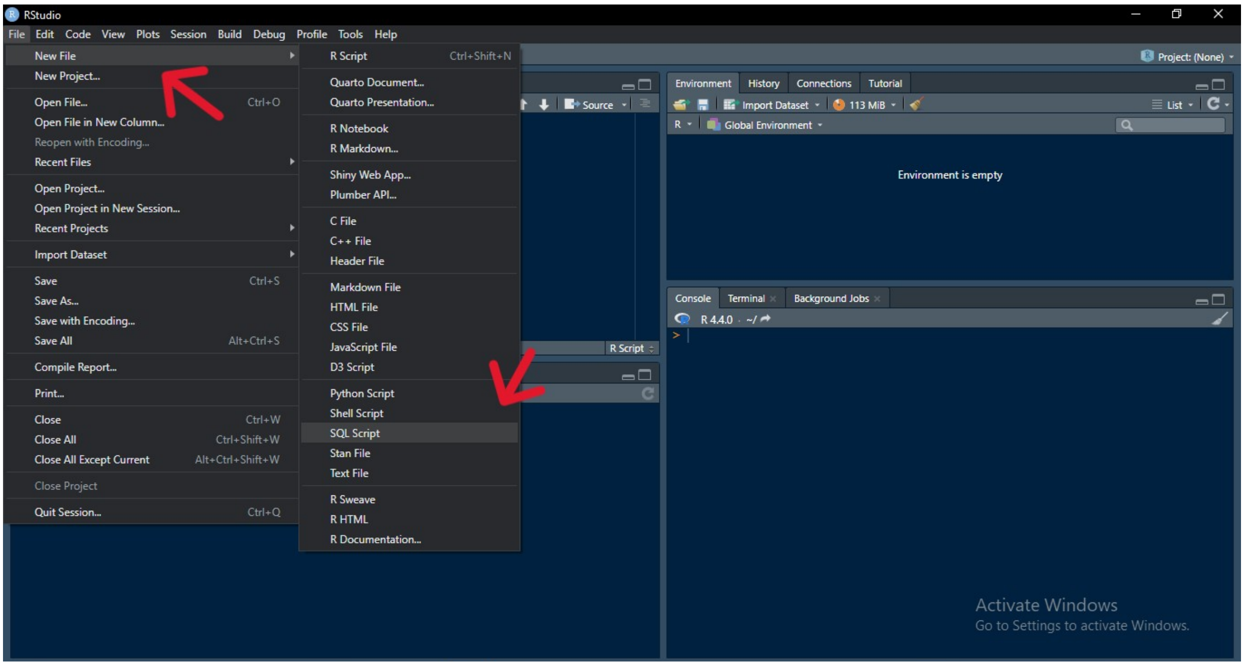
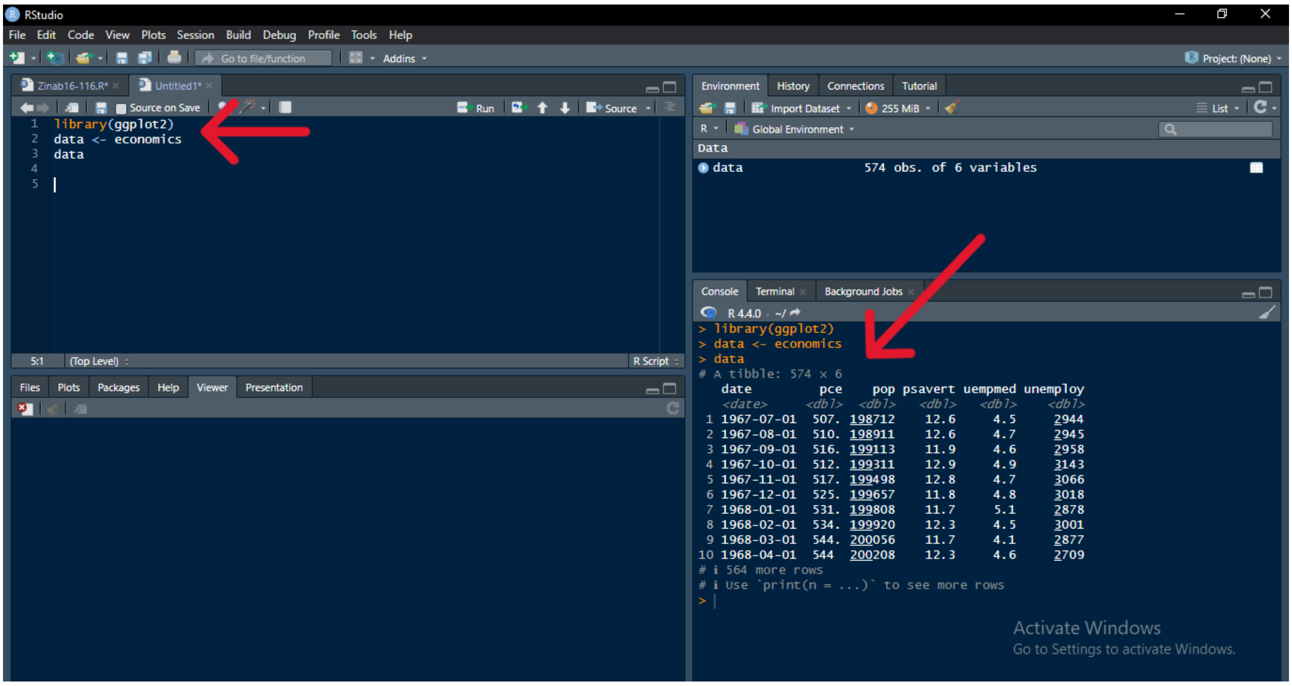


Figure No.13

2. Embedded Data. Embedded data usually refers to data stored internally within the package as ‘ggplot, ggplot2’ or program so that it can be accessed directly without having to load it from external sources or databases . This is what we are interested in in this research.



```
library(ggplot2)
data <- economics
data
```

```
data                574 obs. of 6 variables
```

```
> library(ggplot2)
> data <- economics
> data
# A tibble: 574 × 6
   date          pce    pop psavert uempmed unemploy
   <date>      <dbl>  <dbl>   <dbl>   <dbl>    <dbl>
 1 1967-07-01   507. 198712    12.6     4.5     2944
 2 1967-08-01   510. 198911    12.6     4.7     2945
 3 1967-09-01   516. 199113    11.9     4.6     2958
 4 1967-10-01   512. 199311    12.9     4.9     3143
 5 1967-11-01   517. 199498    12.8     4.7     3066
 6 1967-12-01   525. 199657    11.8     4.8     3018
 7 1968-01-01   531. 199808    11.7     5.1     2878
 8 1968-02-01   534. 199920    12.3     4.5     3001
 9 1968-03-01   544. 200056    11.7     4.1     2877
10 1968-04-01   544  200208    12.3     4.6     2709
# i 564 more rows
# i Use `print(n = ...)` to see more rows
> 
```

RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
Project: (None)
Zinab16-116.R*
Untitled1*
Source on Save
Run
Source
5:1
(Top Level)
R Script
Files Plots Packages Help Viewer Presentation
Environment History Connections Tutorial
Import Dataset
255 MiB
List
R
Global Environment
Data
Console Terminal Background Jobs
R 4.4.0 · ~/
Activate Windows
Go to Settings to activate Windows.

Figure No.14

Secondly:

3.6.2 Calculate The Quartiles, Deciles & Percentiles:

Quartiles in R:

Understanding and calculating quartiles is a fundamental aspect of statistical analysis, providing insights into the distribution of data. In the R programming language, there are specific functions and methodologies to calculate these quartiles, making it easier for professionals and beginners alike to conduct thorough data analysis.
Quartiles dissect a dataset into four equal segments, offering a quick peek into the data’s distribution. This foundational knowledge is not just academic; it’s a practical tool for anyone looking to make informed decisions based on data. They help in identifying the spread and skewness of the dataset, making them indispensable in outlier detection and data normalization.

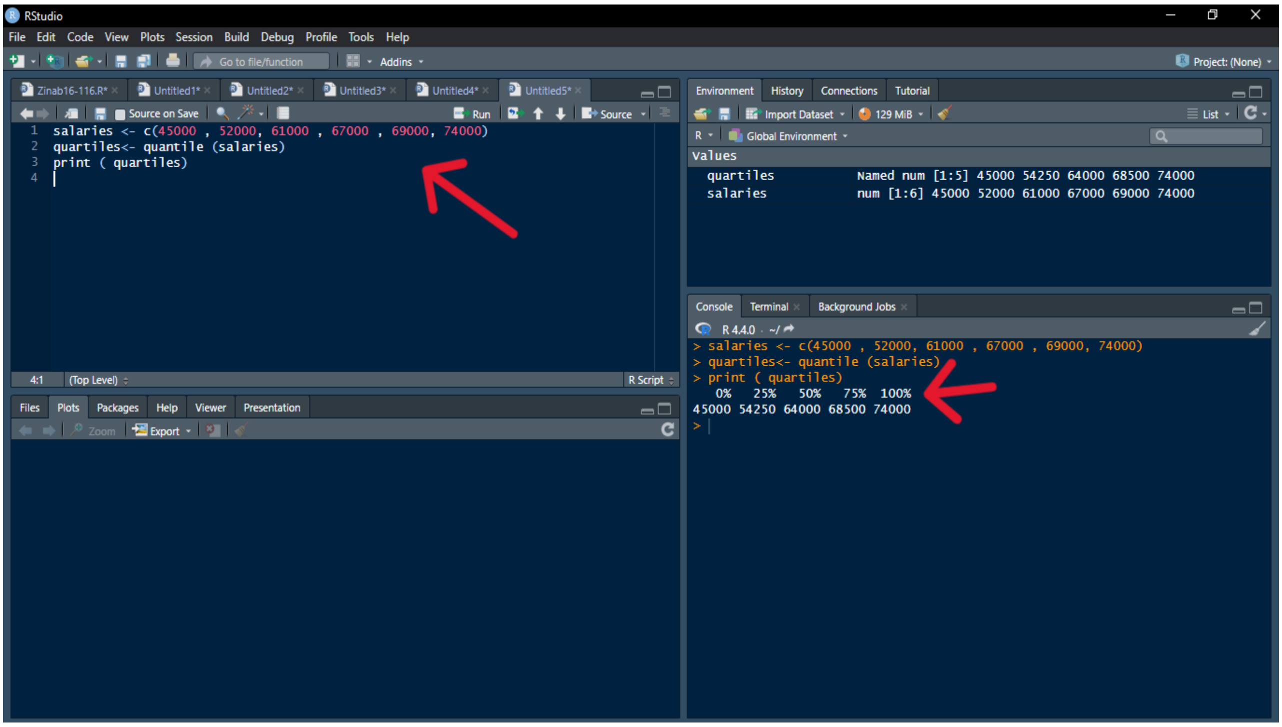
Consider a dataset containing annual salaries of employees within a company. By calculating quartiles, we can determine the distribution of salaries, identify the median salary (Q2), and understand how dispersed the salaries are around the median. This is crucial for HR departments in making compensation decisions, ensuring equity and competitiveness. Quartiles also play a pivotal role in financial data analysis, helping to assess risk and return distributions for investment portfolios.

Practical application:

- data cleaning: identifying outliers that may skew the data..
- policy making: setting thresholds for decision making based on quartile analysis.

In R, the calculation can be as simple as using the **‘ quantile( )’** function on a numeric vector; this code snippet will give you the quartile distribution of the salaries, offering a clear view of data spread.

```
salaries <-  c(45000 , 52000, 61000 , 67000 , 69000, 74000)
quartiles <- quantile (salaries)
print ( quartiles)
```



```
salaries <- c(45000 , 52000, 61000 , 67000 , 69000, 74000)
quartiles<- quantile (salaries)
print ( quartiles)
```

```
Values
  quartiles         Named num [1:5] 45000 54250 64000 68500 74000
  salaries          num [1:6] 45000 52000 61000 67000 69000 74000
```

```
> salaries <- c(45000 , 52000, 61000 , 67000 , 69000, 74000)
> quartiles<- quantile (salaries)
> print ( quartiles)
   0%   25%   50%   75%  100% 
45000 54250 64000 68500 74000 
>
```

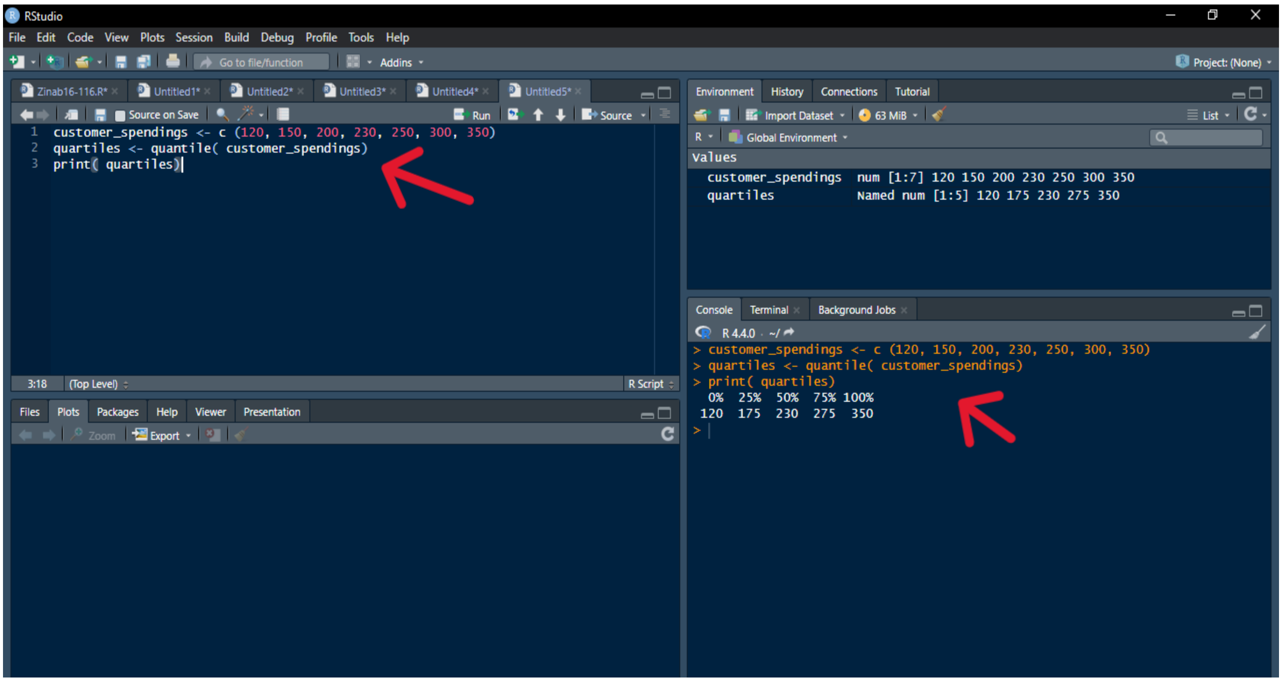
RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Zinab16-116.R* Untitled1* Untitled2* Untitled3* Untitled4* Untitled5*
Source on Save
Run
Source
Environment History Connections Tutorial
Import Dataset
129 MiB
Global Environment
Console Terminal Background Jobs
R 4.4.0 · ~/
4:1 (Top Level) R Script
Files Plots Packages Help Viewer Presentation
Zoom Export
Project: (None)

Figure No.15

Example: analyzing customer spending in a retail setting can reveal which quartile most customers fall into, guiding inventory and marketing strategies.

Calculating these quartiles in R involves the ‘ quantile’ function, applied to any numeric dataset; this example shows the spending distribution across different customer segments, enabling targeted marketing efforts.

```
customer_spendings <- c (120, 150, 200, 230, 250, 300, 350)
quartiles <- quantile( customer_spendings)
print( quartiles)
```



```
customer_spendings <- c (120, 150, 200, 230, 250, 300, 350)
quartiles <- quantile( customer_spendings)
print( quartiles)
```

```
Values
  customer_spendings  num [1:7] 120 150 200 230 250 300 350
  quartiles           Named num [1:5] 120 175 230 275 350
```

```
> customer_spendings <- c (120, 150, 200, 230, 250, 300, 350)
> quartiles <- quantile( customer_spendings)
> print( quartiles)
  0%  25%  50%  75% 100% 
 120  175  230  275  350 
>
```

RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Zinab16-116.R* Untitled1* Untitled2* Untitled3* Untitled4* Untitled5*
Source on Save
Run
Source
Environment History Connections Tutorial
Import Dataset
63 MiB
Global Environment
Console Terminal Background Jobs
R 4.4.0 · ~/
3:18 (Top Level) R Script
Files Plots Packages Help Viewer Presentation
Zoom Export
Project: (None)

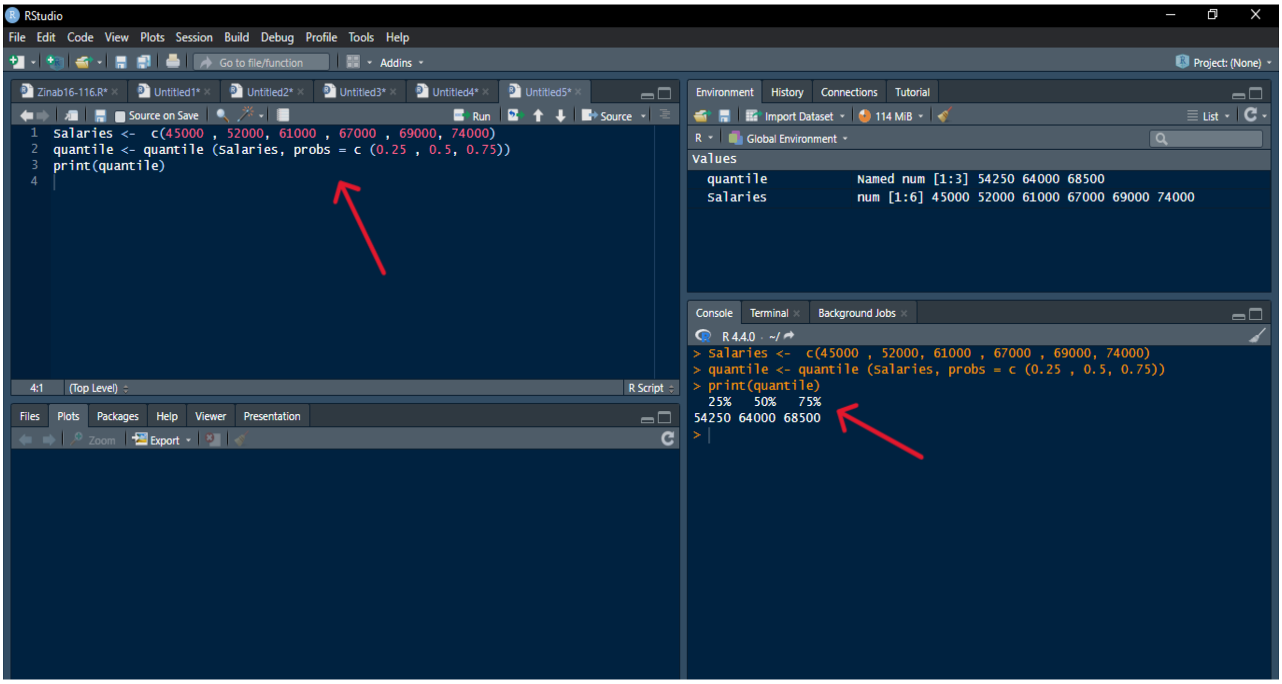
Figure No.16

Harnessing the power of the quantile function in R:

Introduction to the quantile function:

The ‘quantile’ function in R the Go-to tool for quartile calculation. It’s not just about findin the middle value; it’s about understanding the entire data distribution through quartiles. Here’s a basic syntax to get you started:

```
salaries <-  c(45000 , 52000, 61000 , 67000 , 69000, 74000)
quantile <- quantile (salaries, probs = c (0.25 , 0.5, 0.75))
print( quartiles)
```



RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
Project: (None)
Zinab16-116.R* Untitled1* Untitled2* Untitled3* Untitled4* Untitled5*
Source on Save
Run
Source
Environment History Connections Tutorial
Import Dataset
114 MiB
List
Global Environment
Console Terminal Background Jobs
R 4.4.0 · ~/
4:1 (Top Level)
R Script
Files Plots Packages Help Viewer Presentation
Zoom
Export

```
salaries <-  c(45000 , 52000, 61000 , 67000 , 69000, 74000)
quantile <- quantile (salaries, probs = c (0.25 , 0.5, 0.75))
print(quantile)
```

```
Values
  quantile            Named num [1:3] 54250 64000 68500
  salaries            num [1:6] 45000 52000 61000 67000 69000 74000
```

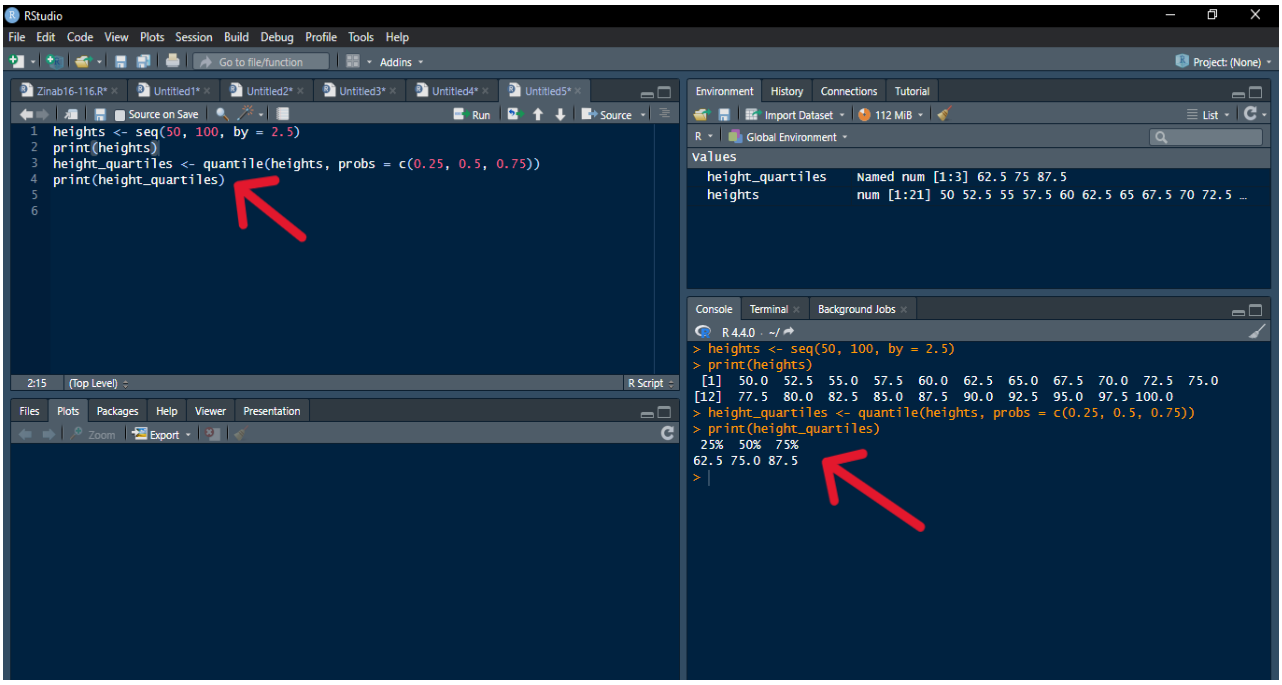
```
> salaries <-  c(45000 , 52000, 61000 , 67000 , 69000, 74000)
> quantile <- quantile (salaries, probs = c (0.25 , 0.5, 0.75))
> print(quantile)
  25%   50%   75% 
54250 64000 68500 
> 
```

Figure No.17

in this snippet, ‘salaries’ represents your datasets, and ‘probs’ specifies the quartile. The command returns the first, second (median), and the third quartiles of ‘salaries’.

practical application: imagine you have a dataset ‘ height’ containing the heights of students in a class. Calculating the quartiles would give you insights into the distribution:

```
height_quartiles <- quantlie(heights, probs = c(0.25 , 0.5, 0.75))
print( height_quartiles)
```



```
heights <- seq(50, 100, by = 2.5)
print(heights)
height_quartiles <- quantile(heights, probs = c(0.25, 0.5, 0.75))
print(height_quartiles)
```

```
Values
  height_quartiles    Named num [1:3] 62.5 75 87.5
  heights             num [1:21] 50 52.5 55 57.5 60 62.5 65 67.5 70 72.5 …
```

```
> heights <- seq(50, 100, by = 2.5)
> print(heights)
 [1]  50.0  52.5  55.0  57.5  60.0  62.5  65.0  67.5  70.0  72.5  75.0
[12]  77.5  80.0  82.5  85.0  87.5  90.0  92.5  95.0  97.5 100.0
> height_quartiles <- quantile(heights, probs = c(0.25, 0.5, 0.75))
> print(height_quartiles)
 25%  50%  75%
62.5 75.0 87.5
> 
```

Figure No.18

This code snippet provides a clear, tangible understanding of how the heights are distributed across quartiles, highlighting the simplicity and power of the ‘ quantile’ function in R.

Quartiles in large datasets:

Working with large datasets requires a more nuanced approach to calculating quartiles, as the volume of data can significantly affect performance and interpretation.

Example:
Let's consider a more complex dataset, such as a large sales record over several years. For the sake of this example, assume we have loaded our dataset into R as ‘ ‘.

Calculating quartiles for such a dataset can be performed similarly, but attention must be paid to data preparation and handling missing values or outliers.

```
# Assuming sales_data is already loaded
quartile <- quantile (sales_data$Sales amount, na.rm = TRUE)
print(quartiles)
```

In this example, ‘na.rm = TRUE’ ensures that missing values are ignored in the quartile calculation, which is crucial for maintaining the integrity of the analysis.

Interpretation:

For large datasets, the quartile calculation not only provides insights into the distribution of data but also highlights potential areas for further analysis, such as seasonal trends or outlier transactions. This step is vital for making informed decisions based on data.

Interpreting quartile results in R:

Once the quartiles are calculated in R, the next crucial step is interpreting these results to glean insights into the used dataset. This understanding can significantly influence decision making in data analysis.

Analyzing quartile output:

Interpreting the output of quartile calculations in R provides a comprehensive view of your data distribution. Let’s delve into practical applications with examples.

- ***Understanding the spread***: the distance between the first quartile (Q1) and the third quartile(Q3) is known as the interquartile range (IQR). It offers a measure of the data spread. A larger IQR indicates a wider spread of data.

```
# calculate IQR
IQR(data$column)
```

- ***Identifying the median***: the second quartile(Q2) is the median, providing a central value of the dataset. Comparing the median to Q1 and Q3 can help identify the skewness in the data.
- ***Skewness detection***: if Q2 is closer to Q1 than to Q3, the data might be skewed left. Conversely, if Q2 is closer to Q3, the data might be skewed right.

Understanding these elements enables interpreting quartile results effectively, providing a clear picture of the dataset’s distribution.

Conclusion:

quartile calculation in R is a powerful tool for data analysis, offering insights into data distribution, outliers, and overall dataset characteristics. By understanding and applying the methods detailed in this guide, one can enhance data analysis skills and make more informed decisions based on quartile analysis.

Deciles in R

Understanding and calculating deciles is a key aspect of statistical analysis, helping to divide data into ten equal parts, each representing 10% of the dataset. In R, the process is simplified with built-in functions, allowing professionals and beginners to quickly analyze data distributions and derive meaningful insights.

Deciles provide a detailed view of data segmentation, enabling the identification of patterns or trends within the dataset. For example, they are widely used in income distribution analysis to group individuals into ten percentiles for comparative studies or policy-making.

Practical Application:

Data Segmentation: Deciles are particularly useful in understanding customer spending behavior, where each decile may represent a specific group of customers. For instance, the top decile (10th) might represent high-spending customers, enabling businesses to tailor marketing strategies.

Steps:

1-Data Preparation:

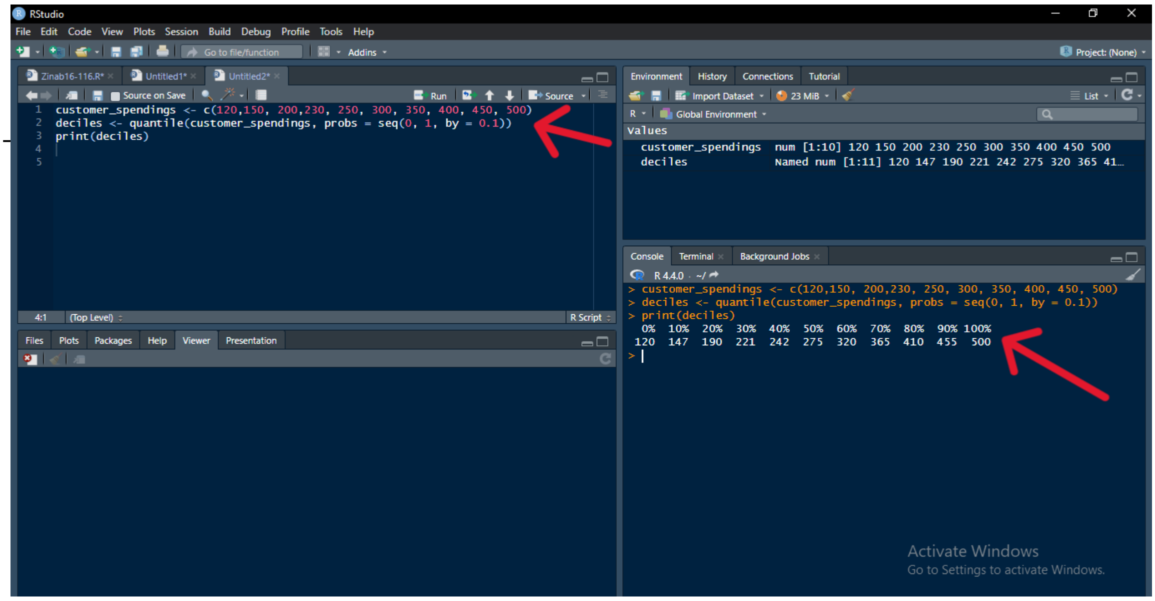
Ensure your dataset is numeric and free of inconsistencies. If the data contains missing values, use appropriate measures (e.g., na.rm = TRUE) to handle them.

2-Using the quantile() ***Function***:

The quantile() function can calculate deciles by dividing the dataset into ten parts. Here's an example:

```
# Sample dataset

customer_spendings <- c(120,150, 200,230, 250, 300, 350, 400, 450, 500)

# Calculate deciles

deciles <- quantile(customer_spendings, probs = seq(0, 1, by = 0.1))

# Print deciles

print(deciles)
```

This snippet computes the deciles for customer_spendings, offering a clear view of how values are distributed across ten equal intervals.



```
customer_spendings <- c(120,150, 200,230, 250, 300, 350, 400, 450, 500)
deciles <- quantile(customer_spendings, probs = seq(0, 1, by = 0.1))
print(deciles)
```

```
Values
  customer_spendings  num [1:10] 120 150 200 230 250 300 350 400 450 500
  deciles             Named num [1:11] 120 147 190 221 242 275 320 365 41…
```

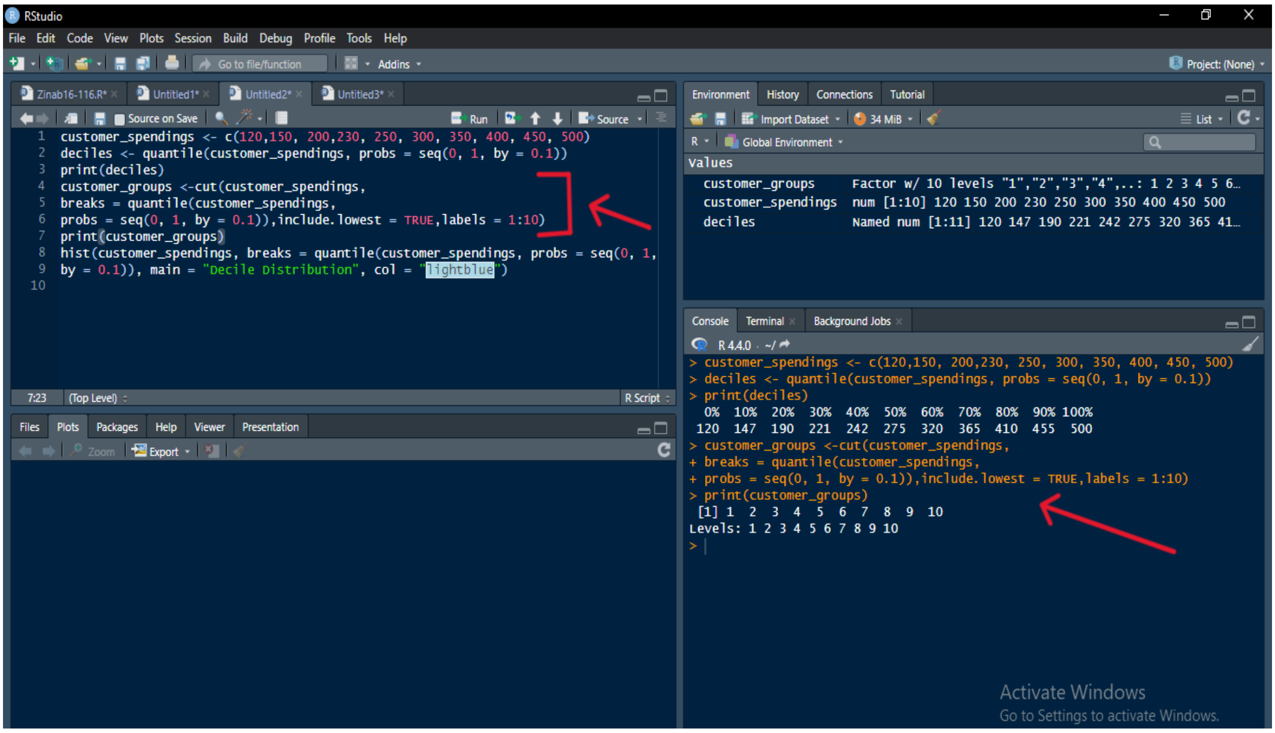
```
> customer_spendings <- c(120,150, 200,230, 250, 300, 350, 400, 450, 500)
> deciles <- quantile(customer_spendings, probs = seq(0, 1, by = 0.1))
> print(deciles)
  0%  10%  20%  30%  40%  50%  60%  70%  80%  90% 100% 
 120  147  190  221  242  275  320  365  410  455  500 
> |
```

Figure No.19

3-Grouping Data into Deciles:

To assign each data point to a specific decile group:

```
# Assign to decile groups
customer_groups <-
cut(customer_spendings, breaks =
quantile(customer_spendings,
probs = seq(0, 1, by = 0.1)),include.lowest = TRUE,labels = 1:10)

# Print grouped data
print(customer_groups)
```



```
customer_spendings <- c(120,150, 200,230, 250, 300, 350, 400, 450, 500)
deciles <- quantile(customer_spendings, probs = seq(0, 1, by = 0.1))
print(deciles)
customer_groups <-cut(customer_spendings,
breaks = quantile(customer_spendings,
probs = seq(0, 1, by = 0.1)),include.lowest = TRUE,labels = 1:10)
print(customer_groups)
hist(customer_spendings, breaks = quantile(customer_spendings, probs = seq(0, 1,
by = 0.1)), main = "Decile Distribution", col = "lightblue")
```

```
Values
  customer_groups     Factor w/ 10 levels "1","2","3","4",..: 1 2 3 4 5 6…
  customer_spendings  num [1:10] 120 150 200 230 250 300 350 400 450 500
  deciles             Named num [1:11] 120 147 190 221 242 275 320 365 41…
```

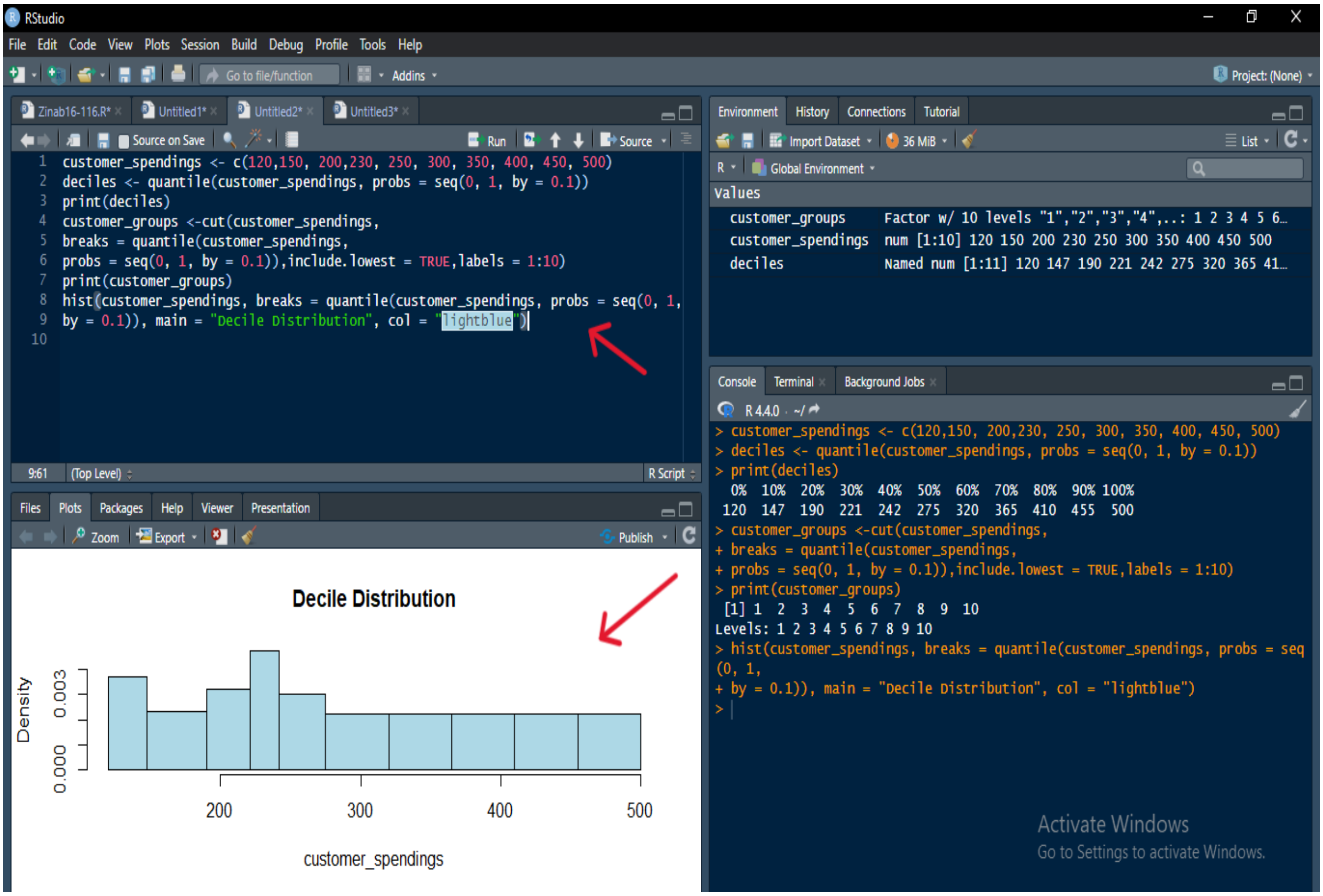
```
> customer_spendings <- c(120,150, 200,230, 250, 300, 350, 400, 450, 500)
> deciles <- quantile(customer_spendings, probs = seq(0, 1, by = 0.1))
> print(deciles)
  0%  10%  20%  30%  40%  50%  60%  70%  80%  90% 100% 
 120  147  190  221  242  275  320  365  410  455  500 
> customer_groups <-cut(customer_spendings,
+ breaks = quantile(customer_spendings,
+ probs = seq(0, 1, by = 0.1)),include.lowest = TRUE,labels = 1:10)
> print(customer_groups)
 [1] 1  2  3  4  5  6  7  8  9  10
Levels: 1 2 3 4 5 6 7 8 9 10
> 
```

Figure No.20

4-***Visualization***:

A histogram or bar chart can illustrate the distribution of deciles:

hist(customer_spendings, breaks = quantile(customer_spendings, probs = seq(0, 1, by = 0.1)), main = "Decile Distribution", col = "lightblue") ; **Figure No.21**



```
customer_spendings <- c(120,150, 200,230, 250, 300, 350, 400, 450, 500)
deciles <- quantile(customer_spendings, probs = seq(0, 1, by = 0.1))
print(deciles)
customer_groups <-cut(customer_spendings,
breaks = quantile(customer_spendings,
probs = seq(0, 1, by = 0.1)),include.lowest = TRUE,labels = 1:10)
print(customer_groups)
hist(customer_spendings, breaks = quantile(customer_spendings, probs = seq(0, 1,
by = 0.1)), main = "Decile Distribution", col = "lightblue")
```

```
Values
  customer_groups     Factor w/ 10 levels "1","2","3","4",..: 1 2 3 4 5 6…
  customer_spendings  num [1:10] 120 150 200 230 250 300 350 400 450 500
  deciles             Named num [1:11] 120 147 190 221 242 275 320 365 41…
```

```
> customer_spendings <- c(120,150, 200,230, 250, 300, 350, 400, 450, 500)
> deciles <- quantile(customer_spendings, probs = seq(0, 1, by = 0.1))
> print(deciles)
  0%  10%  20%  30%  40%  50%  60%  70%  80%  90% 100% 
 120  147  190  221  242  275  320  365  410  455  500 
> customer_groups <-cut(customer_spendings,
+ breaks = quantile(customer_spendings,
+ probs = seq(0, 1, by = 0.1)),include.lowest = TRUE,labels = 1:10)
> print(customer_groups)
 [1] 1  2  3  4  5  6  7  8  9  10
Levels: 1 2 3 4 5 6 7 8 9 10
> hist(customer_spendings, breaks = quantile(customer_spendings, probs = seq
(0, 1,
+ by = 0.1)), main = "Decile Distribution", col = "lightblue")
> 
```

Interpretation:

•Decile Values: The output of quantile() provides the thresholds for each decile. For instance, the 5th decile (median) divides the dataset into two halves.

•Data Spread: Identifying clusters or gaps between deciles can reveal data characteristics, such as concentration or dispersion.

By using deciles, decision-makers can analyze data in finer detail, ensuring precise and ***actionable insights for a wide range of applications***

Percentiles in R

Calculating percentiles is a fundamental statistical task widely used in data analysis to understand the distribution of data points. In the R programming language, various methods exist to perform this calculation, each suited for different scenarios. This guide provides beginners with a comprehensive understanding of these methods, accompanied by practical R code samples. Whether you're analyzing exam scores, market data, or any other dataset, mastering percentiles in R will significantly enhance your data analysis skills.

What Exactly are Percentiles?

Percentiles divide a dataset into 100 equal parts, helping to understand data distribution. For example, the 25th percentile indicates the value below which 25% of data falls. Percentiles have practical uses, such as assessing student performance relative to peers, guiding curriculum adjustments, and creating personalized learning plans.

Steps:

1-Calculating percentages using quantile():

Use the quantile() function to determine percentage values (25%, 50%, 75%) directly.

# Sample dataset

set.seed (123)

- Specifies a fixed random value so that the same random numbers are generated each time the code is run. This helps in obtaining repeatable results.

data <- rnorm (100)

- Creates a dataset of 100 normally distributed data points with mean 0 and standard deviation One.

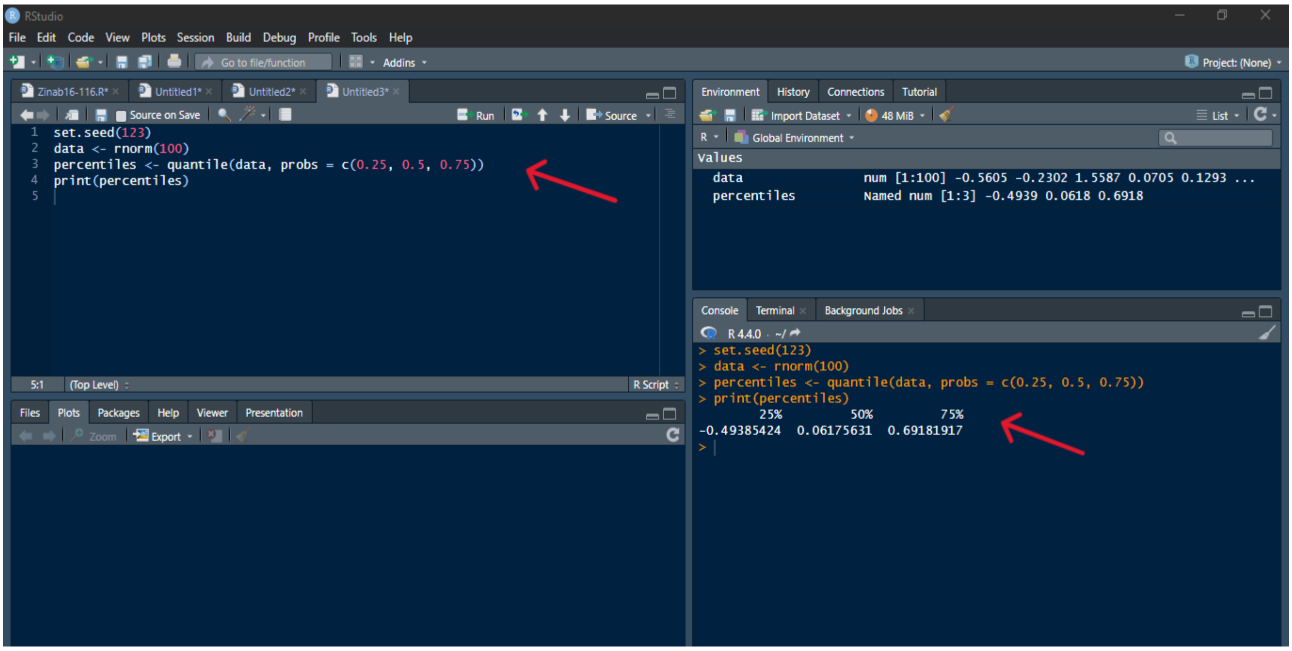
# Calculating the 25th, 50th, and 75th percentiles

percentiles <- quantile(data, probs = c(0.25, 0.5, 0.75))

- The quantile function calculates the percentile values specified by probs.
  0.25: represents 25% (first quartile).
  0.5: represents 50% (median).
  0.75: represents 75% (third quartile).

print (percentiles)

- Displays the resulting values (25th, 50th, and 75th percentile) in the output window.



```
set.seed(123)
data <- rnorm(100)
percentiles <- quantile(data, probs = c(0.25, 0.5, 0.75))
print(percentiles)
```

```
Values
  data           num [1:100] -0.5605 -0.2302 1.5587 0.0705 0.1293 ...
  percentiles    Named num [1:3] -0.4939 0.0618 0.6918
```

```
> set.seed(123)
> data <- rnorm(100)
> percentiles <- quantile(data, probs = c(0.25, 0.5, 0.75))
> print(percentiles)
        25%         50%         75% 
-0.49385424  0.06175631  0.69181917 
> 
```

Figure No.22

2-Using the ecdf() function to calculate cumulative percentages:

```
#Creating an ECDF
set.seed(123)
data <- rnorm(100)
ecdf_function <- ecdf(data)
```

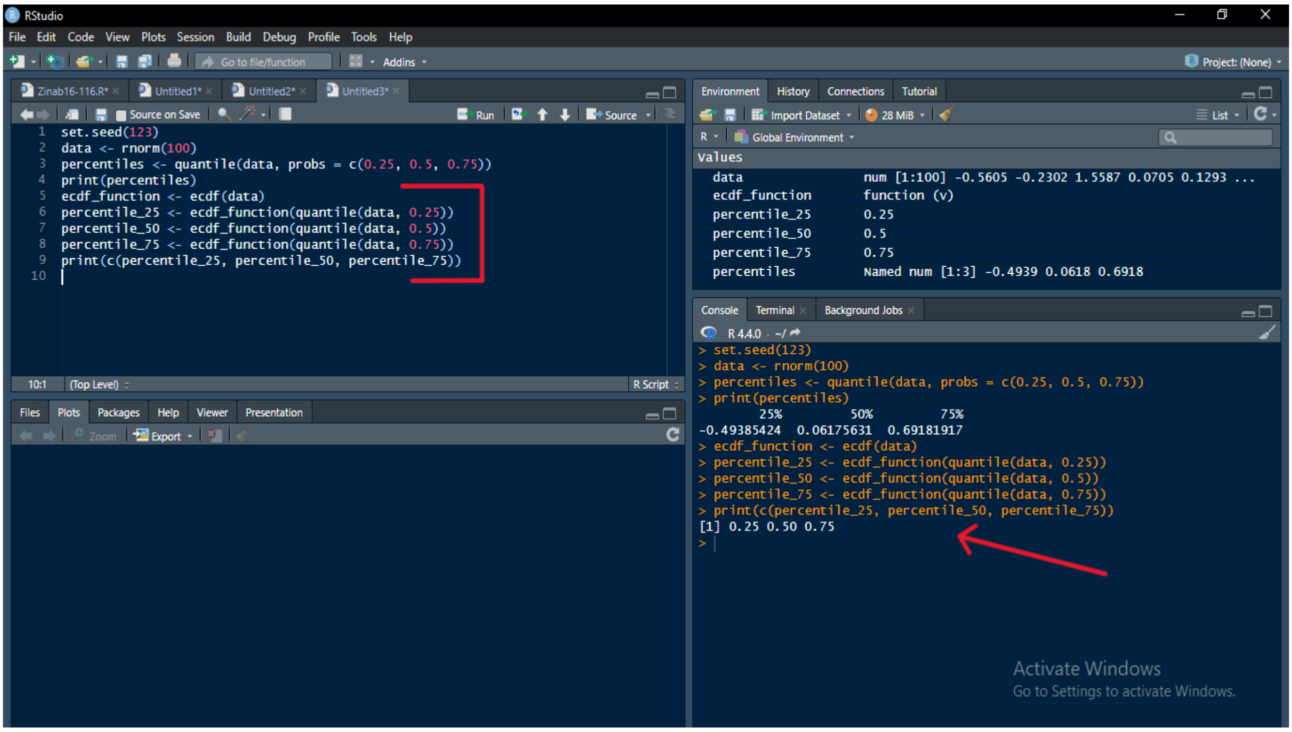
- Creates an Empirical Cumulative Distribution Function (ECDF) from the dataset.
- The ecdf() function creates a cumulative step function for each data point.

```
# Using the ECDF to find percentiles
percentile_25 <- ecdf_function(quantile(data, 0.25))
percentile_50 <- ecdf_function(quantile(data, 0.5))
percentile_75 <- ecdf_function(quantile(data, 0.75))
```

- quantile(data, 0.25):
  Returns the value representing 25% (the first quartile) of the data.
- ecdf_function(quantile(data, 0.25)):
  Applies the cumulative distribution function to the point representing the first quartile. The result gives the cumulative percentage (here 0.25).
- This is true for both (ecdf_function(quantile(data, 0.5)) and ecdf_function(quantile(data, 0.75))).

```
print (c(percentile_25, percentile_50, percentile_75)
```

- Displays the cumulative values of the 25%, 50%, and 75% percentages in the output window.



```
set.seed(123)
data <- rnorm(100)
percentiles <- quantile(data, probs = c(0.25, 0.5, 0.75))
print(percentiles)
ecdf_function <- ecdf(data)
percentile_25 <- ecdf_function(quantile(data, 0.25))
percentile_50 <- ecdf_function(quantile(data, 0.5))
percentile_75 <- ecdf_function(quantile(data, 0.75))
print(c(percentile_25, percentile_50, percentile_75))
```

```
Values
  data              num [1:100] -0.5605 -0.2302 1.5587 0.0705 0.1293 ...
  ecdf_function     function (v)
  percentile_25     0.25
  percentile_50     0.5
  percentile_75     0.75
  percentiles       Named num [1:3] -0.4939 0.0618 0.6918
```

```
> set.seed(123)
> data <- rnorm(100)
> percentiles <- quantile(data, probs = c(0.25, 0.5, 0.75))
> print(percentiles)
        25%         50%         75% 
-0.49385424  0.06175631  0.69181917 
> ecdf_function <- ecdf(data)
> percentile_25 <- ecdf_function(quantile(data, 0.25))
> percentile_50 <- ecdf_function(quantile(data, 0.5))
> percentile_75 <- ecdf_function(quantile(data, 0.75))
> print(c(percentile_25, percentile_50, percentile_75))
[1] 0.25 0.50 0.75
> 
```

Figure No.23

When to use these methods?

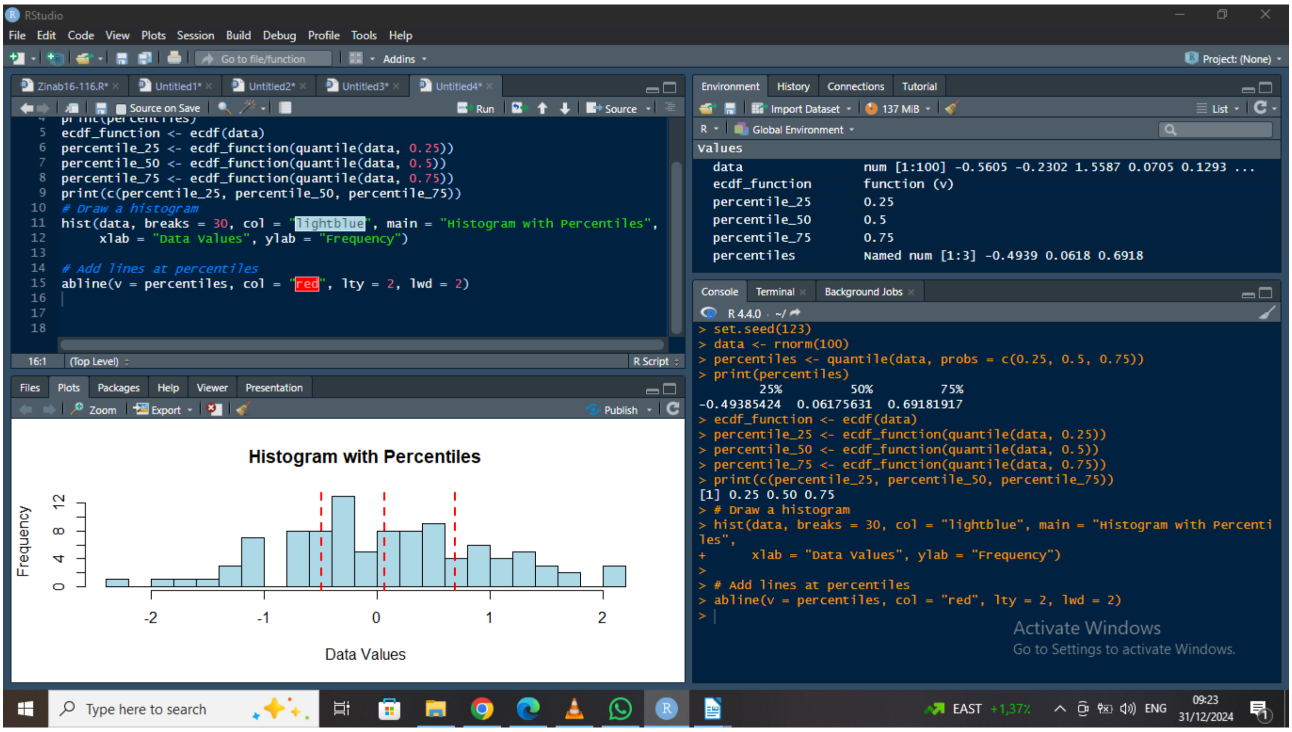
- quantile(): When you need the percentile values directly without analyzing the accumulation of points.
- ecdf(): If you want to display or analyze the accumulation of data via percentages, especially in graphs.

3-Visualizing:

A) Drawing with basic R functions:

Draw a histogram
hist(data, breaks = 30, col = "lightblue", main = "Histogram with Percentiles", xlab = "Data Values", ylab = "Frequency")
hist(): Draws a histogram.
abline(): Adds vertical lines at percentile values using v = percentiles

Add lines at percentiles
abline(v = percentiles, col = "red", lty = 2, lwd = 2)
col: Specifies the color (red here).
lty: Specifies the line type (dotted).
lwd: Specifies the line thickness.



```
print(percentiles)
ecdf_function <- ecdf(data)
percentile_25 <- ecdf_function(quantile(data, 0.25))
percentile_50 <- ecdf_function(quantile(data, 0.5))
percentile_75 <- ecdf_function(quantile(data, 0.75))
print(c(percentile_25, percentile_50, percentile_75))
# Draw a histogram
hist(data, breaks = 30, col = "lightblue", main = "Histogram with Percentiles",
     xlab = "Data Values", ylab = "Frequency")

# Add lines at percentiles
abline(v = percentiles, col = "red", lty = 2, lwd = 2)
```

```
Values
  data            num [1:100] -0.5605 -0.2302 1.5587 0.0705 0.1293 ...
  ecdf_function   function (v)
  percentile_25   0.25
  percentile_50   0.5
  percentile_75   0.75
  percentiles     Named num [1:3] -0.4939 0.0618 0.6918
```

```
> set.seed(123)
> data <- rnorm(100)
> percentiles <- quantile(data, probs = c(0.25, 0.5, 0.75))
> print(percentiles)
        25%         50%         75% 
-0.49385424  0.06175631  0.69181917 
> ecdf_function <- ecdf(data)
> percentile_25 <- ecdf_function(quantile(data, 0.25))
> percentile_50 <- ecdf_function(quantile(data, 0.5))
> percentile_75 <- ecdf_function(quantile(data, 0.75))
> print(c(percentile_25, percentile_50, percentile_75))
[1] 0.25 0.50 0.75
> # Draw a histogram
> hist(data, breaks = 30, col = "lightblue", main = "Histogram with Percenti
les",
+      xlab = "Data Values", ylab = "Frequency")
> 
> # Add lines at percentiles
> abline(v = percentiles, col = "red", lty = 2, lwd = 2)
> 
```

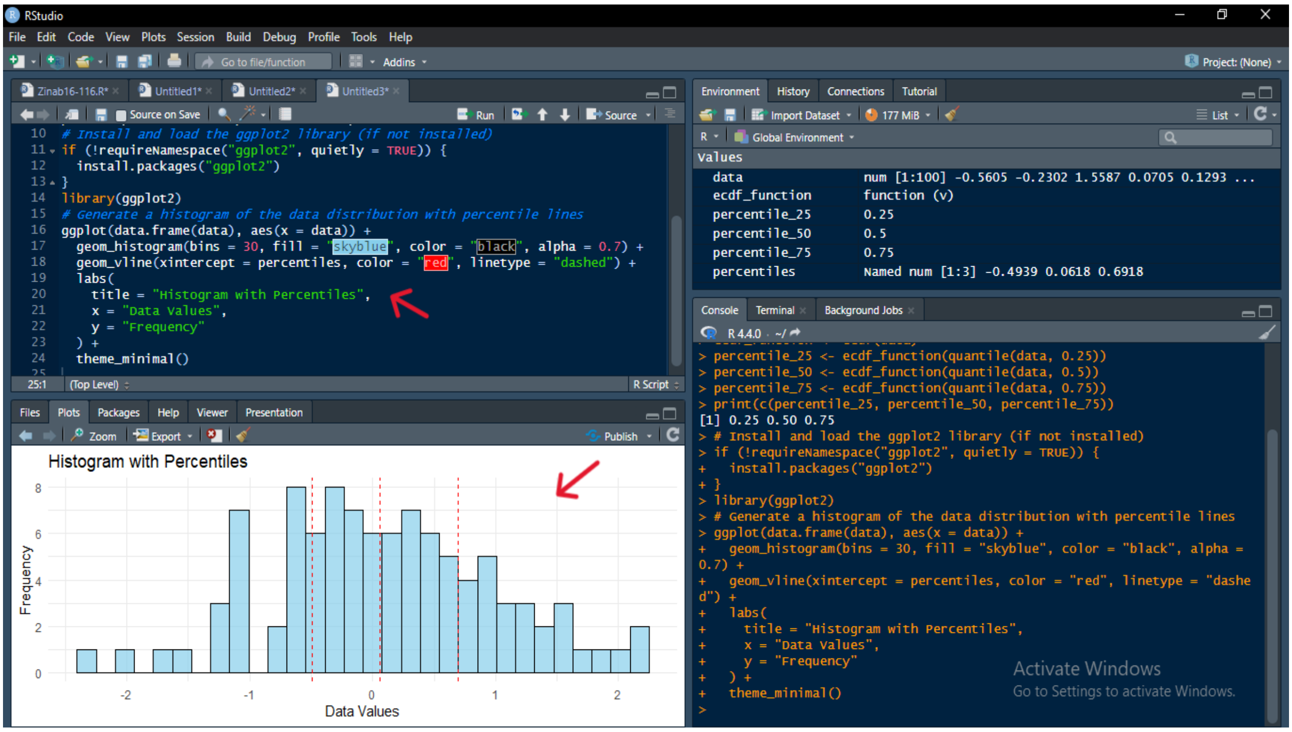
Figure No.24

- Advantages:
  Fast and simple.
- Disadvantages:
  Less flexible and aesthetic.

B) Plotting using ggplot2 library:

```
# Install and load the ggplot2 library (if not installed)
if (!requireNamespace("ggplot2", quietly = TRUE)) {
install.packages("ggplot2")
}
library(ggplot2)
# Generate a histogram of the data distribution with percentile lines
ggplot(data.frame(data), aes(x = data)) +
geom_histogram(bins = 30, fill = "skyblue", color = "black", alpha = 0.7) +
geom_vline(xintercept = percentiles, color = "red", linetype = "dashed") +
labs(
title = "Histogram with Percentiles",
 x = "Data Values",
 y = "Frequency"
 ) +
 theme_minimal()
```

- geom_histogram(): Creates a histogram showing the distribution of data.
- geom_vline(): Draws vertical lines at quartiles.
- labs(): Used to add titles and axis labels.
- theme_minimal(): Improves the appearance of the histogram.



```
# Install and load the ggplot2 library (if not installed)
if (!requireNamespace("ggplot2", quietly = TRUE)) {
  install.packages("ggplot2")
}
library(ggplot2)
# Generate a histogram of the data distribution with percentile lines
ggplot(data.frame(data), aes(x = data)) +
  geom_histogram(bins = 30, fill = "skyblue", color = "black", alpha = 0.7) +
  geom_vline(xintercept = percentiles, color = "red", linetype = "dashed") +
  labs(
    title = "Histogram with Percentiles",
    x = "Data Values",
    y = "Frequency"
  ) +
  theme_minimal()
```

```
Values
  data             num [1:100] -0.5605 -0.2302 1.5587 0.0705 0.1293 ...
  ecdf_function    function (v)
  percentile_25    0.25
  percentile_50    0.5
  percentile_75    0.75
  percentiles      Named num [1:3] -0.4939 0.0618 0.6918
```

```
> percentile_25 <- ecdf_function(quantile(data, 0.25))
> percentile_50 <- ecdf_function(quantile(data, 0.5))
> percentile_75 <- ecdf_function(quantile(data, 0.75))
> print(c(percentile_25, percentile_50, percentile_75))
[1] 0.25 0.50 0.75
> # Install and load the ggplot2 library (if not installed)
> if (!requireNamespace("ggplot2", quietly = TRUE)) {
+   install.packages("ggplot2")
+ }
> library(ggplot2)
> # Generate a histogram of the data distribution with percentile lines
> ggplot(data.frame(data), aes(x = data)) +
+   geom_histogram(bins = 30, fill = "skyblue", color = "black", alpha = 
0.7) +
+   geom_vline(xintercept = percentiles, color = "red", linetype = "dashe
d") +
+   labs(
+     title = "Histogram with Percentiles",
+     x = "Data Values",
+     y = "Frequency"
+   ) +
+   theme_minimal()
>
```

Figure No.25

- Advantages:
  Flexible, elegant, and allows for great customization.
- Disadvantages:
  Requires additional library installation.

CONCLUSION:

measures of position such as quartiles, deciles, and percentiles serve as fundamental tools in summarizing and analyzing data distributions. They not only help identify patterns and outliers but also provide great and meaningful insights into different datasets, enabling informed decision-making in various fields and industries. This research draws attention to their significance, portraying that quartiles divide data into four equal parts, offering a foundational understanding of its distribution. Deciles provide a more gritty segmentation, splitting data into ten parts, while percentiles give greater and more precise understanding by dividing data into one hundred equal segments. Each measure contributes uniquely to analyzing data and understanding its distribution.

The practical application of these measures is indicated using the R programming language, calling attention to its capabilities for efficient data handling, computation, and visualization. By employing R's quantile function, analysts can seamlessly calculate quartiles, deciles, and percentiles, playing upon these insights for tasks like detecting anomalies, segmenting customers, or analyzing income distributions. This study also underscores the importance of integrating these measures into decision-making frameworks, particularly in data-driven industries like finance, healthcare, and education.

While the simplicity of quartiles makes them best possible for quick overviews, deciles and percentiles offer more detailed insights, supplying the needs of advanced analyses.as with any statistical tool, careful consideration must be given to data preparation, handling of outliers, and the context in which these measures are applied. Misinterpretation or reliance on a single measure without considering others can lead to skewed or unreliable results, especially in datasets with extreme values.
So, In conclusion, quartiles, deciles, and percentiles remain invaluable for statistical analysis. Their integration into modern tools like R ensures that data analysts and decision-makers can support their perspectives fully, transforming raw data into actionable and usable knowledge. This study contributes to the understanding and application of these measures, connecting and feeling the gap between statistical theory and practical implementation, and enabling a deeper conception of data distributions across different domains.

REFERENCES:

- Galton, F. (1885): “Regression towards mediocrity in Hereditary stature”.
- Pearson, K. (1895): “Contributions to the mathematical theory of statistics”.
- Nield, T. (2022). Essential math for data science: take control of your data with fundamental linear algebra, probability, and statistics (1st ed.) O’Reilly Media.
- Statistics- mathematical statistics functions – Python 3.10.5 documentation.Docs.python.org. (2022). retrieved 20 july 2022, from http://docs.python.org/3/library/statistics.html.
- Cuemath. (n.d). percentile formula. http://www.cuemath.com/percentile-formula/
- Cuemath. (n.d). Decile formula.
- http://www.cuemath.com/data/decile/
- Bluman, A. G.(2017). Elemeantry statistic: A step-by-step approach (10th ed.).
- McGraw-Hill Education.
- Starnes , D.S , yates, D., and Moore, D.S. (2018). The practice of statistics (6th ed).
- Khan academy.(n.d) . statistics and probability. From http://www.khanacademy.org/math/statistics-probabilitiy
- stat trek.(n.d). statistics tutorial , from http://statterk.com/
- university of sydeny.(n.d.). measures of spread and position, from http://www.syndey.edu.au
- Witte, R. S., & Witte, J. S. (2017). Statistics (11th ed.). John Wiley & Sons
- Black, K. (2010). Business statistics for contemporary decision making. John Wiley & Sons, Inc.
- Triola, M. F. (2018). Elementary statistics (18th ed.). Pearson.
- Corporate Finance Institute. (n.d.). *Central Tendency*. Retrieved December 21, 2024, from https://corporatefinanceinstitute.com/resources/data-science/central-tendency/
- Study.com. (n.d.). Mean, median, mode, and range: Measures of central tendency. Study.com. Retrieved December 22, 2024, from https://study.com/learn/lesson/mean-median-mode-range-measures-central-tendency.html
- Plantlet.org. (n.d.). Measures of dispersion. Plantlet.org. Retrieved December 22, 2024, from https://plantlet.org/measures-of-dispersion/#google_vignette
- Shiha, M. (n.d.). Chapter 4: Measures of dispersion [Document]. Retrieved December 22, 2024, from https://fac.ksu.edu.sa/sites/default/files/ch_04_shiha_2nd_ver.doc
- GeeksforGeeks. (n.d.). *Data types in statistics*. Retrieved January 21, 2025, from https://www.geeksforgeeks.org/data-types-in-statistics/
- Wikipedia. (n.d.). *Homogeneity and heterogeneity (statistics)*. Retrieved January 21, 2025, from https://en.wikipedia.org/wiki/Homogeneity_and_heterogeneity_(statistics)
- Wikipedia. (n.d.). *Consistency (statistics)*. Retrieved January 21, 2025, from https://en.m.wikipedia.org/wiki/Consistency_(statistics)/
- Scribbr. (n.d.). *Quartiles and quantiles*. Retrieved January 21, 2025, from https://www.scribbr.com/statistics/quartiles-quantiles/
- Cuemath. (n.d.). *Quartile formula*. Retrieved January 21, 2025, from https://www.cuemath.com/quartile-formula/
- Cuemath. (n.d.). *Decile*. Retrieved January 21, 2025, from https://www.cuemath.com/data/decile/
- Investopedia. (n.d.). *Quartile*. Retrieved January 21, 2025, from https://www.investopedia.com/terms/q/quartile.asp/
- Cuemath. (n.d.). *Percentile formula*. Retrieved January 21, 2025, from https://www.cuemath.com/percentile-formula/