CYBR 520

Chapter 5: Descriptive Statistics

Hands-On Exploratory Data Analysis with Python: Perform EDA techniques to understand, summarize, and investigate your data.

West Virginia University
JOHN CHAMBERS COLLEGE OF BUSINESS AND ECONOMICS

PLEASE REFER TO MODULE 2/CH5 ON GITHUB TO OBTAIN THE CODES OR REFER TO THE BOOK’S GITHUB REPO

TOPICS

- Understanding statistics
- Measures of central tendency
- Measures of dispersion

Understanding statistics

- In data science, both qualitative and quantitative analyses are important aspects.
- Quantitative analysis of any dataset requires an understanding of statistical concepts.
  - Statistics is a branch of mathematics that deals with collecting, organizing, and interpreting data.
  - Hence, by using statistical concepts, we can understand the nature of the data, a summary of the dataset, and the type of distribution that the data has.

Let’s talk Distributions

- Understanding different data distributions is essential when working with basic descriptive statistics for several reasons:
  1. **Selection of Appropriate Descriptive Statistics**
  2. **Interpretation of Summary Statistics**
  3. **Identifying Outliers**
  4. **Choosing Visualization Techniques**
  5. **Inference and Hypothesis Testing**
  6. **Communicating Results**
  7. **Data Transformation**

Selection of Appropriate Descriptive Statistics

- different types of data distributions require different descriptive statistics to effectively summarize and interpret the data. For example:
  - For normally distributed data, mean and standard deviation are often used.
  - For skewed data, median and quartiles may be more appropriate.
  - For categorical data, mode and frequency counts are relevant.

Interpretation of Summary Statistics

- Understanding the data distribution helps in interpreting the summary statistics. For instance:
  - In a symmetric distribution, the mean and median are similar.
  - In a positively skewed distribution, the mean is typically greater than the median.
  - In a bimodal distribution, it may be important to identify and describe each mode separately.

Identifying Outliers

- Knowing the typical data distribution allows you to spot outliers more effectively. Outliers can significantly affect measures like the mean, making them less representative of the central tendency. Being familiar with the data distribution helps identify whether an extreme value is genuinely an outlier or just a characteristic of the data.

Choosing Visualization Techniques

- Data distributions influence the choice of visualization methods. For example:
- Histograms are suitable for visualizing the shape of continuous data distributions.
- Bar charts are useful for categorical data.
- Box plots are great for showing the central tendency and spread of data, including outliers.

Inference and Hypothesis Testing

- Understanding the underlying data distribution is crucial when performing hypothesis tests and making statistical inferences. Many statistical tests assume specific distributions, and violations of these assumptions can lead to incorrect conclusions.

Communicating Results

- When communicating your findings to others, it's essential to describe the data distribution. This helps the audience grasp the nature of the data and the context of your analysis.

Data Transformation

- In some cases, you might need to transform the data to achieve a more normal distribution, which can be a prerequisite for certain statistical analyses.
- Understanding the original distribution guides your decision on whether and how to transform the data.

Distribution function

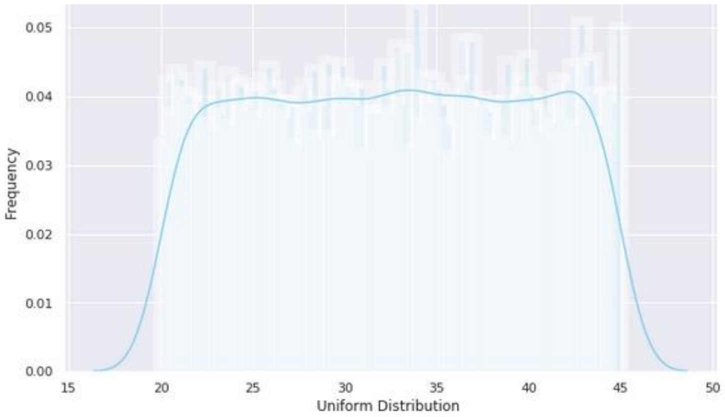
- The distribution function, often referred to as the cumulative distribution function (CDF), is a fundamental concept in probability and statistics.
- It's a bit like a summary of how likely different outcomes are in a random experiment or process.

How likely something is to be less than or equal to a certain value.

Distribution function

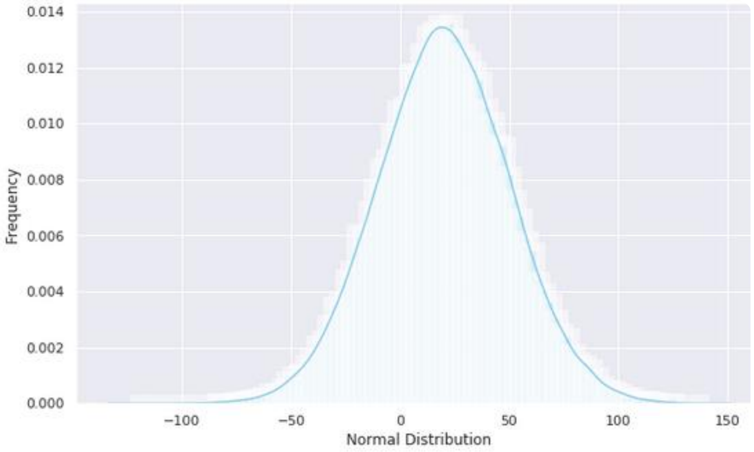
- Imagine you have a magical machine that can predict how likely different things are to happen. This magical machine uses something called a "distribution function."
- Now, to understand this "distribution function," think of it like a line or curve on a graph.
- If this line or curve is super smooth without any sudden jumps or gaps:
  - we call it a "continuous function" It's like a perfectly straight road with no potholes or speed bumps.
  - Otherwise, it is “discrete function”

Uniform distribution



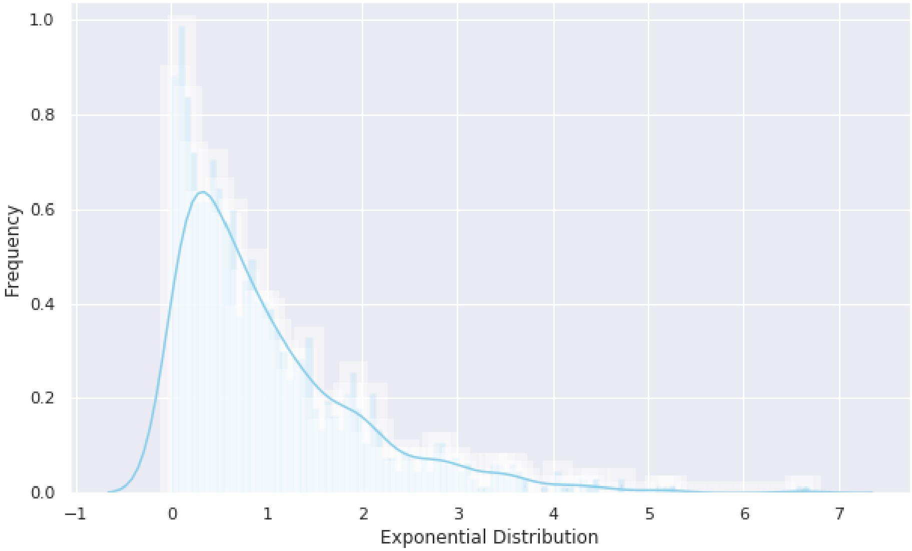
- In the context of probability and statistics, a uniform probability distribution, often denoted as "U(a, b)," describes a situation where a continuous random variable has an equal probability of taking any value within a specified interval, [a, b].
- The probability density function (PDF) for a continuous uniform distribution is characterized by a constant value within this interval and is zero outside of it.

Normal distribution



- The normal distribution, also known as the Gaussian distribution, is a fundamental concept in statistics and data analysis.
- It is characterized by a symmetric bell-shaped curve when graphed. The key features of a normal distribution are its mean (μ) and standard deviation (σ).

Exponential distribution



- The exponential distribution is a fundamental probability distribution used to model the time between events in a Poisson point process. In this process, events occur continuously and independently at a constant average rate.
- The probability density function (PDF) of the exponential distribution is described as:

f(x; λ) = λ * e^(-λx) for x >= 0, 0 otherwise

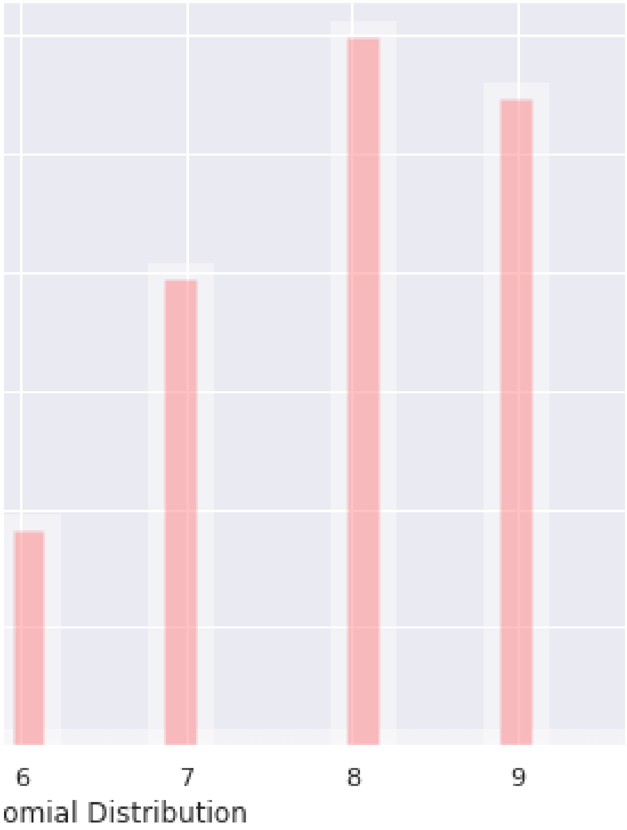
- Here's what it means:

"λ" (lambda) represents the average rate at which events occur per unit of time.

- The function describes the probability of waiting a specific amount of time "x" for the next event to happen.

Binomial distribution

- The binomial distribution is a discrete probability distribution that deals with situations characterized by a fixed number of independent trials, each having only two possible outcomes: success or failure. These outcomes are not required to have equal probabilities, and the outcome of one trial does not influence the others.
- Key features of the binomial distribution:
  - It's named "binomial" because it involves "bi" or two possible outcomes.
  - Parameters include the number of trials (n), the probability of success in a single trial (p), and the number of successful outcomes you're interested in (k).



Descriptive statistics

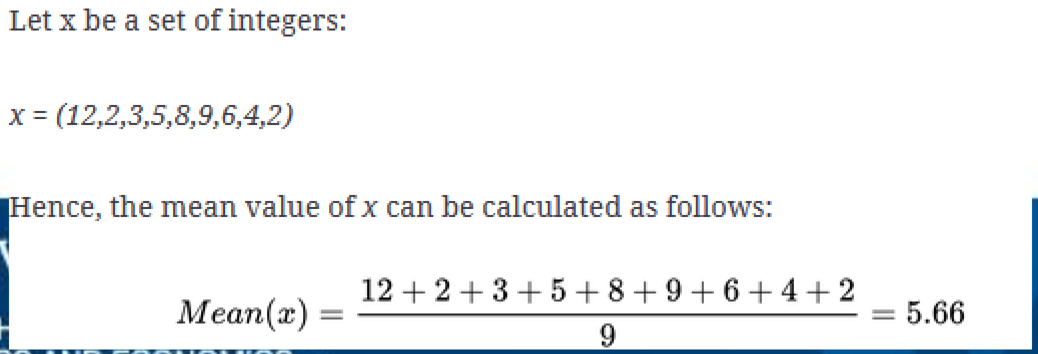
- Descriptive statistics deals with the formulation of simple summaries of data so that they can be clearly understood. The summaries of data may be either numerical representations or visualizations with simple graphs for further understanding.
- Typically, such summaries help in the initial phase of statistical analysis. There are two types of descriptive statistics:
  1. Measures of central tendency
  2. Measures of variability (spread)

Measures of central tendency

- Measures of central tendency are statistical tools used to identify the central or representative value in a dataset.
- There are three primary measures:
  - Mean
  - Median
  - Mode

Mean

- **Mean:** The mean, often referred to as the average, is calculated by summing all the values in the dataset and dividing by the total number of values. It's sensitive to extreme values and is suitable for data that follows a roughly symmetrical distribution.



Let x be a set of integers:

x = (12,2,3,5,8,9,6,4,2)

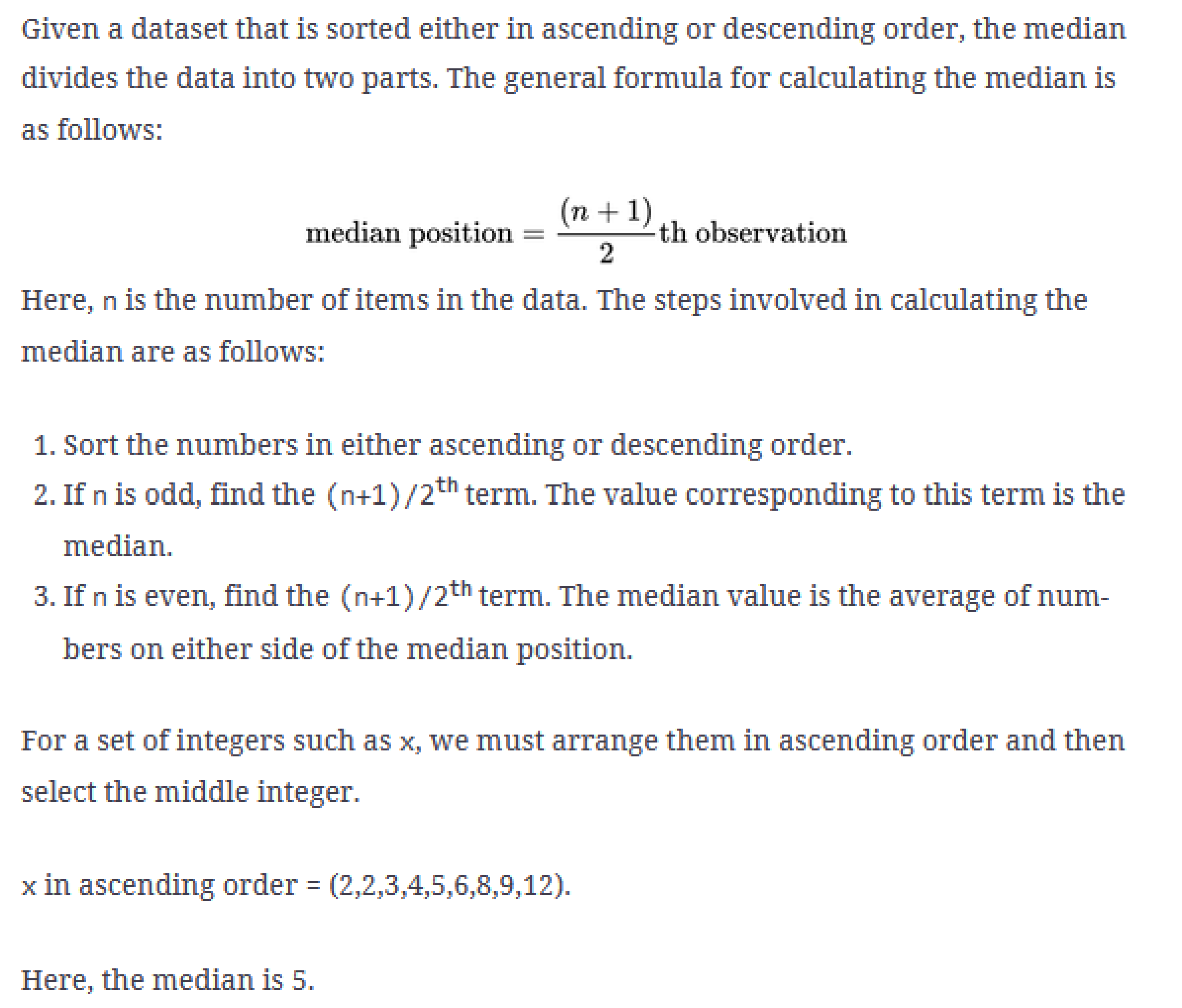
Hence, the mean value of *x* can be calculated as follows:

$$Mean(x) = \frac{12+2+3+5+8+9+6+4+2}{9} = 5.66$$

Median

- **Median:** The median is the middle value in a dataset when it's arranged in ascending or descending order. If there's an even number of data points, it's the average of the two middle values. The median is robust to extreme values and works well with skewed data.

Median



Given a dataset that is sorted either in ascending or descending order, the median divides the data into two parts. The general formula for calculating the median is as follows:

$$\text{median position} = \frac{(n+1)}{2}\text{th observation}$$

Here, n is the number of items in the data. The steps involved in calculating the median are as follows:

1. Sort the numbers in either ascending or descending order.
2. If n is odd, find the $(n+1)/2^{th}$ term. The value corresponding to this term is the median.
3. If n is even, find the $(n+1)/2^{th}$ term. The median value is the average of numbers on either side of the median position.

For a set of integers such as x, we must arrange them in ascending order and then select the middle integer.

x in ascending order = (2,2,3,4,5,6,8,9,12).

Here, the median is 5.

Mode

- **Mode:** The mode is the value that occurs with the highest frequency in the dataset. There can be one mode (unimodal), more than one mode (multimodal), or no mode at all. It's useful for identifying the most common value in a dataset.
- Mode is 2 in the previous example

Measures of variability

- Also known as measures of spread, provide insights into the dispersion or spread of data points within a dataset. Here are three commonly used measures:
  1. **Range**
  2. **Variance**
  3. **Standard Deviation**

Range

- The range is the simplest measure of spread.
- It's calculated as the difference between the maximum and minimum values in the dataset.
- While easy to compute, it can be sensitive to outliers.

Variance

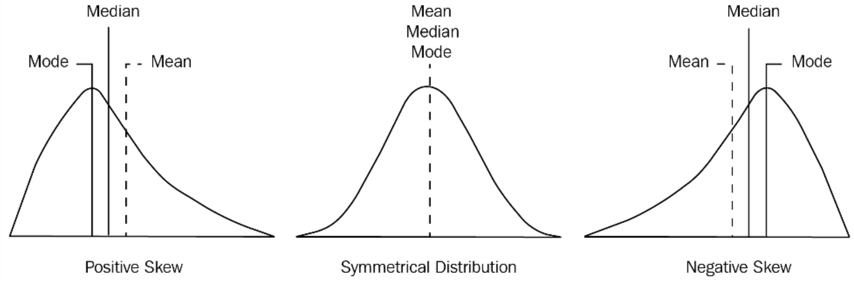
- Variance quantifies how individual data points deviate from the mean.
- It involves squaring the difference between each data point and the mean, summing these squared differences, and dividing by the number of data points.
- High variance indicates greater spread.

Standard Deviation

- The standard deviation is the square root of the variance.
- It provides a measure of the average distance between each data point and the mean.
- It's widely used because it shares the same units as the original data and is easier to interpret than variance.

Skewness

- Skewness is a statistical measure used in probability theory and statistics to quantify how asymmetrical a dataset is in relation to its mean.
- Skewness can take positive, negative, or undefined values, indicating the direction and degree of skewness in the data distribution.



- In the provided illustration, you can observe the following:
- The rightmost graph has a longer tail on the left side, indicating left-skewness. This means that if you pick a point in the longer left tail, the mean is less than the mode, a condition known as negative skewness.
- The leftmost graph has a longer tail on the right side, indicating right-skewness. If you select a point on the longer right tail, the mean value is greater than the mode, a condition known as positive skewness.
- The middle graph has tails on both sides that are equal in length, representing a symmetrical distribution with zero skewness.

Kurtosis

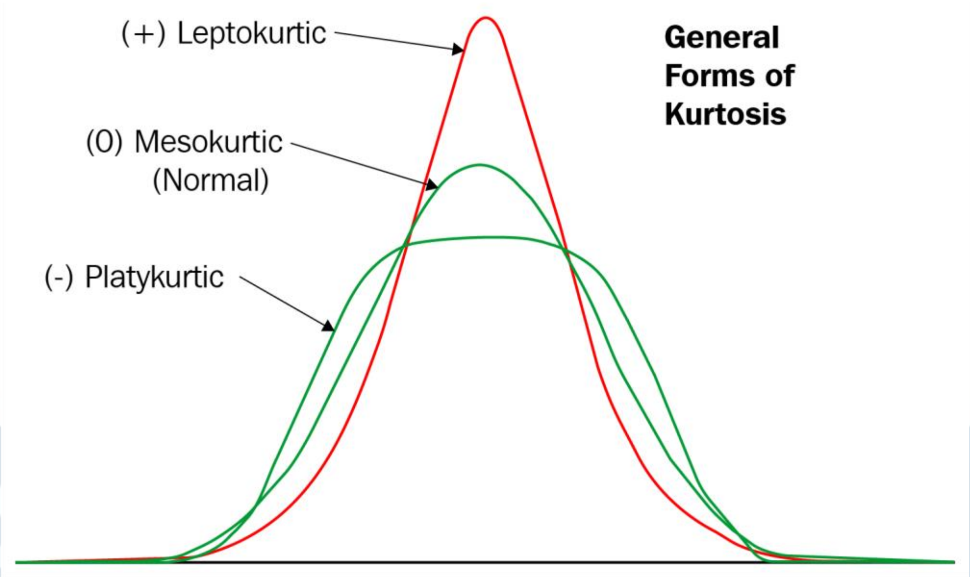
- Kurtosis is a statistical measure that helps us understand how the tails of a data distribution compare to those of a normal distribution (the bell-shaped curve).
- It tells us if the data contains extreme values, or outliers, that deviate significantly from the norm.
- Unlike skewness, which focuses on the symmetry of the distribution, kurtosis assesses the thickness or heaviness of the tails.

Kurtosis

- High kurtosis indicates a distribution with heavy tails, suggesting the presence of outliers, either on the high or low end of the data.
- Low kurtosis, on the other hand, implies thinner tails and fewer extreme values.
- Both high and low kurtosis values signal that further investigation of the data may be necessary.

Kurtosis

- **Mesokurtic**: If any dataset follows a normal distribution, it follows a mesokurtic distribution. It has kurtosis around 0.
- **Leptokurtic**: In this case, the distribution has kurtosis greater than 3 and the fat tails indicate that the distribution produces more outliers.
- **Platykurtic:** In this case, the distribution has negative kurtosis and the tails are very thin compared to the normal distribution.



Calculating percentiles

- Percentiles are statistical values that divide a dataset into 100 equal parts, where each part represents a percentage. To calculate percentiles, you first sort the data. For instance, if you find that the 80th percentile of a dataset is 130, it signifies that 80% of the data points are less than or equal to 130. The formula used for this calculation is:
- Percentile Value = (Position / Total Number of Data Points) * 100

Calculating percentiles – Example1

- For example, in a dataset:
  1, 2, 2, 3, 4, 5, 6, 7, 7, 8, 9, 10

the 4th percentile would be (4/12) * 100 = 33.33%.

- This means that 33.33% of the data values are less than or equal to 4.

Finding the Median (50th Percentile)

- Suppose you have a dataset of exam scores: 60, 70, 75, 80, 85, 90.
- To find the median (50th percentile), you can follow these steps:
  - First, sort the dataset in ascending order: 60, 70, 75, 80, 85, 90.
  - The formula for finding the median (50th percentile) is: (Position / Total Number of Data Points) * 100.
  - In this case, (3 / 6) * 100 = 50. So, the median score is the value at the 50th percentile.
  - The 50th percentile falls between the 2nd and 3rd data points, which are 70 and 75. You can take the average of these two values: (70 + 75) / 2 = 72.5. So, the median score is 72.5.

Finding the 75th Percentile

- The 75th percentile (also known as the third quartile) can be found using the data provided:
  50, 40, 65, 30, 55, 45, 25, 70, 35, 60

1. Arrange the numbers in ascending order: 25, 30, 35, 40, 45, 50, 55, 60, 65, 70
2. Calculate the index for the 75th percentile using the formula: index=(75/100) ×n index=7.5 ×n Where n is the number of data points.

- Using the provided data: 0.75×10=7.5
- The index 7.5 means that the 75th percentile lies between the 7th and 8th data points.

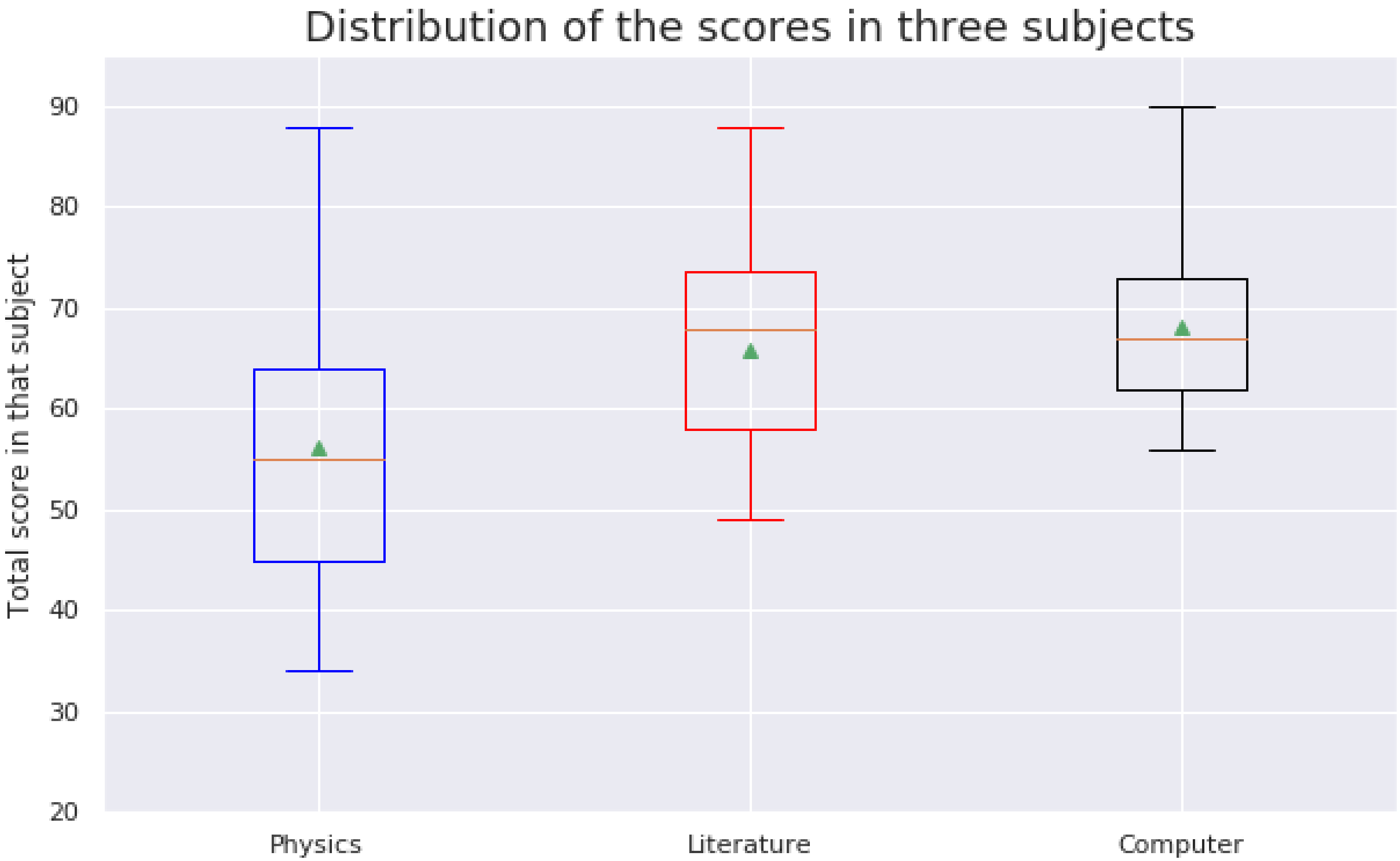
3. Now, take the average of the 7th and 8th numbers. P75=55+602=57.5P75=255+60=57.5

- So, the 75th percentile of the given data set is 57.5.

Quartiles

- Quartiles are statistical measures used to split a dataset into four equal parts. They are often related to percentiles:
  - Q1 (the first quartile) corresponds to the 25th percentile.
  - Q2 (the second quartile) corresponds to the 50th percentile, which is also the median.
  - Q3 (the third quartile) corresponds to the 75th percentile.
- The inter-quartile range (IQR) is a measure of the variability within the middle 50% of the dataset. It's calculated as the difference between the third quartile (Q3) and the first quartile (Q1).
- It tells you how spread out the data is in the central portion of the dataset.

Quartiles



From the graph, it is clear that the minimum score obtained by the students was around 32, while the maximum score obtained was 90, which was in the computer science subject.