Chapter 2

Types of Analytics

2.1 What Is Analytics?

Analytics, derived from its root word "analysis" or "analytical," is the systematic exploration and interpretation of data to uncover meaningful insights and patterns. At its core, raw data serves as the foundation for analytics, with various analyses and analytical operations applied to derive valuable information. The complexity of analytics varies depending on the depth of analysis performed, with more intricate operations often requiring additional time but yielding higher levels of insight.

In essence, analytics encompasses a spectrum of methodologies and techniques aimed at extracting actionable insights from data. From basic descriptive analytics that summarize historical data to predictive analytics that forecast future trends, the field of analytics offers a range of approaches to address diverse business challenges. As organizations continue to leverage analytics to inform decision-making and drive strategic initiatives, understanding the different types of analytics becomes increasingly crucial. Throughout this chapter, we will explore various types of analytics and their applications across different industries.

2.2 A Simple Example of Analytics—Test Scores

Take this example. There are 10 students in a class, and below are their recorded test scores. Given that this is a dataset, we have one variable, which is called score.

Score 1	Score 2	Score 3	Score 4	Score 5	Score 6	Score 7	Score 8	Score 9	Score 10
10	15	12	14	17	17	20	8	5	13

This is a simple table with 10 observations. Some analytical observations from this:

1. Number of observations

$$n = 10$$

2. Minimum value in the dataset

$$\textit{Minimum of dataset} = 5$$

3. Maximum value in the dataset

$$\textit{Maximum of dataset} = 20$$

4. Average of the dataset

$$Average = \frac{Sum\,of\,Observations}{Number\,of\,observataions}$$

$$= \frac{(Score\ 1 + Score\ 2 + Score\ 3 + Score\ 4 + Score\ 5 + Score\ 6 + Score\ 7 + Score\ 8 + Score\ 9 + Score\ 10)}{n}$$

$$= \frac{(10 + 15 + 12 + 14 + 17 + 17 + 20 + 8 + 5 + 13)}{10}$$

$$\frac{131}{10}$$

$$= 13.1$$

5. Standard deviation of the dataset

X value	Variance squared	Variance squared
10	$(10 - 13.1)^2$	9.61
15	$(15 - 13.1)^2$	3.61
12	$(12 - 13.1)^2$	1.21
14	$(14 - 13.1)^2$	0.81
17	$(17 - 13.1)^2$	15.21
17	$(17 - 13.1)^2$	15.21
20	$(20 - 13.1)^2$	47.61
8	$(8 - 13.1)^2$	26.01
5	$(5 - 13.1)^2$	65.61
13	$(13 - 13.1)^2$	0.01
$\sum \text{Variance squared} = 194.46$ $Average\ of\ Variance\ Squared = \frac{194.46}{10} = 19.446$		

$$Standard\,Deviation = \sqrt{Average\,of\,\text{Variance squared}}$$

$$Standard\,Deviation = \sqrt{19.446}$$

$$Standard\,Deviation = 4.408$$

In a snapshot, the insight we can draw from this set of numbers is the following:

- This is a sample of 10 students with their test scores.
- The lowest score is 5, and the highest score is 20.
- The average score for this test is 13.1.
- The standard deviation is 4.408.

The above bullets describe the dataset and therefore called descriptive analytics.

2.3 Descriptive Analytics

As the name suggests descriptive analytics describes the data. Descriptive analytics serves as the foundation of data analysis, focusing on answering the question "what" in a dataset. It provides straightforward insights into the basic characteristics of data, helping to illuminate its features and trends.

For instance, in the above dataset of student scores, descriptive analytics helps us understand the key metrics such as the number of scores, the range from lowest to highest, the average score, and how much scores vary around this average.

Some questions that guide the thinking in descriptive analytics are:

- How many datapoints are in the dataset?
- What is the minimum of the dataset?
- What is the maximum of the dataset?
- What is the average of the dataset?
- What is the standard deviation of the dataset?
- What is the proportion of students who are above the average?
- What is the proportion of students who are below the average?

All these questions describe the dataset, and these insights act as guideposts, enabling us to navigate through the data landscape and identify significant patterns and outliers. By grasping the size, distribution, and variability of scores, we gain valuable insights into areas of strength and areas that may require attention.

Descriptive analytics holds particular importance for businesses and organizations, offering a practical way to make sense of large datasets. Whether it's optimizing marketing strategies or improving operational efficiency, descriptive analytics provides essential clarity.

This is the way we potentially answer the above questions:

- How many datapoints are in the dataset?
  - This is data of 10 students.
- What is the minimum of the dataset?
  - The lowest score a student earned was 5.
- What is the maximum of the dataset?
  - The highest score a student earned was 20.
- What is the average of the dataset?
  - On average, the students earned a score of 13.1.
- What is the standard deviation of the dataset?
  - The scores of the students are approximately 4.408 points apart.
- What is the proportion of students who are above the average?
  - Scores 2, 4, 5, 6, and 7 are above the average (=13.1).
  - 5 out of 10 students performed above the average.
  - Approximately 50% of the students scored higher than the average.
- What is the proportion of students who are below the average?
  - Scores 1, 3, 8, 9, and 10 are below the average (=13.1).
  - 5 out of 10 students performed below the average.
  - Approximately 50% of the students scored lower than the average.

Descriptive analytics takes the least amount of time and provides very basic value and insight to the business problem in hand. For any data science problem, the first step is descriptive analytics, then it moves onto diagnostic analytics.

2.4 Diagnostic Analytics

Diagnostic analytics builds upon the foundation laid by descriptive analytics, aiming to delve deeper into the underlying reasons or causes behind observed patterns or trends in data. It is the phase where we seek to answer the question “why” certain phenomena are occurring.

In the context of test scores, diagnostic analytics raises questions that probe into the reasons behind the observed data points. These questions include

- Why is this specific number of data points considered optimal for this problem?
- What factors contribute to score 9 being the lowest in the dataset?
- What circumstances lead to score 7 emerging as the highest in the dataset?
- What factors might explain fluctuations in the average score compared to previous years?
- Why is there a variance in the standard deviation compared to previous years?
- What underlying factors contribute to the differences in the proportion of students scoring above or below the average?

Answering these questions requires deeper domain knowledge and an understanding of the industry context. It involves analyzing factors such as teaching methodologies, curriculum changes, student demographics, or external influences that may impact test scores.

Moreover, diagnostic analytics plays a crucial role in designing effective key performance indicators (KPIs) to evaluate the success of new initiatives. By understanding the underlying factors influencing outcomes, organizations can tailor KPIs to assess the effectiveness of programs or interventions accurately.

For instance, let us consider a scenario where a new educational program was introduced in 2023 with the goal of improving student scores. With this additional context, diagnostic analytics can help identify whether the observed changes in test scores are attributable to the implementation of the new program or other external factors. It provides valuable insights into the effectiveness of the initiative and informs future decision-making strategies.

With the additional context provided, the insights gleaned from diagnostic analytics become more valuable as they offer potential explanations and avenues for further exploration within the dataset. Let’s consider the possible answers to each question:

- Why are these number of data points ideal for this problem?
  - The new program was only implemented with a small sample of students.
- Why is score 9 the lowest in the dataset?
  - Possibility 1: The student was potentially onboarded to the program late.
  - Possibility 2: The new program did not cater to the learning style of Student 9.
  - Possibility 3: This could be an outlier.
  - Possibility 4: There could have been a data entry error.
  - Note: This data point contradicts the initial hypothesis that the program helps increase test scores.
- Why is score 7 the highest in the dataset?
  - Possibility 1: The student was potentially a talented and gifted student.
  - Possibility 2: The new program catered to the learning style of Student 7.
  - Possibility 3: This could be an outlier.
  - Possibility 4: There could have been a data entry error.
  - Note: This data point supports the initial hypothesis that the program helps increase test scores.
- Why is the average the same?
  - The new program did not have an impact on student scores.

- Why is the average lower?
  - The new program had a negative impact on student scores.
- Why is the average higher?
  - The new program had a positive impact on student scores.
- Why is the standard deviation the same?
  - The new program did not impact the differences between student scores.
- Why is the standard deviation lower?
  - Most students had similar scores with little to no difference, indicating a small range of scores.
- Why is the standard deviation higher?
  - Most students had very different scores with large differences, indicating a wide range of scores.
- Why is there a difference in the proportion of students who are above the average?
  - Currently, 50% of students scored above the average.
  - An increase would indicate that the new program is effective.
  - Additional information from previous years can determine if there was an increase or decrease.
- Why is there a difference in the proportion of students who are below the average?
  - Currently, 50% of students scored below the average.
  - A decrease would indicate that the new program is effective.
  - Additional information from previous years can determine if there was an increase or decrease.

In conclusion, diagnostic analytics offers valuable insights into correlations between variables, but it does not establish causation. Despite this limitation, it empowers businesses to make informed decisions and optimize strategies based on observed patterns. However, to uncover causal relationships, predictive analytics becomes essential.

2.5 Predictive Analytics

Predictive analytics represents the frontier of data analysis, where we peer into the future using mathematical models and algorithms. It's the realm of forecasting and anticipation, leveraging insights gleaned from descriptive and diagnostic analytics to project potential outcomes and trends. While descriptive analytics helps us understand the past and diagnostic analytics uncovers hidden correlations, predictive analytics enables us to proactively shape the future.

Imagine we have a dataset containing student scores, but we want to predict future scores based on additional characteristics, known as features or variables. These features could include factors like study habits, socioeconomic background, or attendance records. By incorporating these variables into our dataset, predictive analytics can generate forecasts and insights beyond what's immediately evident from the data.

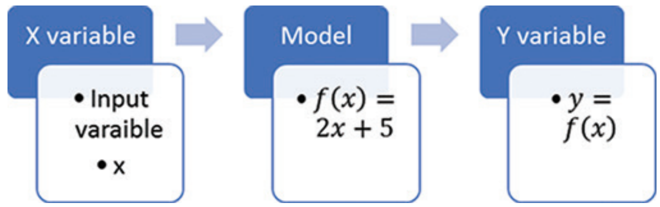
For instance, predictive analytics use case might involve predicting future exam scores based on factors such as student demographics, previous academic performance, and extracurricular activities. By analyzing historical data and identifying patterns, predictive analytics can provide valuable insights into potential academic outcomes, allowing educators to tailor interventions and support systems to maximize student success.

In essence, predictive analytics empowers organizations to anticipate future trends, mitigate risks, and capitalize on opportunities before they arise. By harnessing the power of mathematical modeling and machine learning algorithms, predictive analytics enables proactive decision-making and strategic planning, driving innovation and growth in an increasingly dynamic and competitive landscape.

Predictive analytics models are mathematical functions. A simple example of a function is

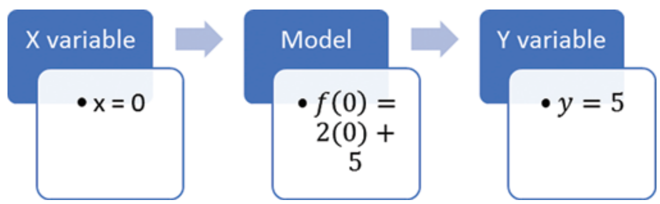
$$f(x) = 2x + 5$$

This function has a variable *x* with a coefficient 2. The constant is 5.

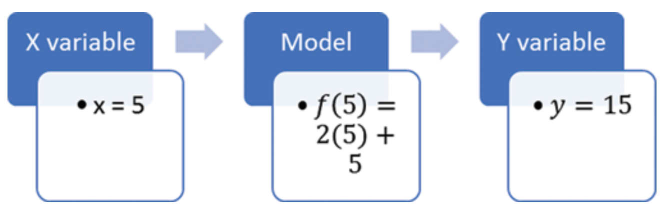


© Anusha Vissapragada

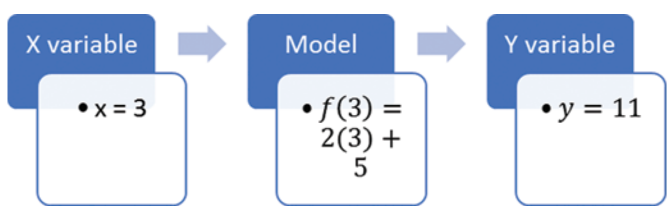
The results of the function vary depending on the value of the *x* variable, making it the dependent variable. Conversely, the *x* values, which remain unchanged, are referred to as independent variables. The equation representing this relationship constitutes a basic model.



© Anusha Vissapragada



© Anusha Vissapragada



© Anusha Vissapragada

Now, let's go back to our case. The unknown variables are future scores. The *x* values are going to be other characteristics and features that we need to collect information about to predict future scores.

Let us say that we have additional context that more number of teachers the students have, the better their scores are.

Quantifying this further, research findings reveal that students with no teachers achieved a baseline score of 15. Moreover, for each additional teacher, the score increased by 3 points. This empirical evidence provides valuable insights into the relationship between the number of teachers and student scores, enhancing the predictive capabilities of our analytics model.

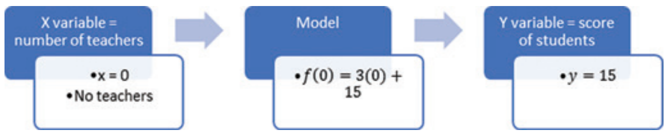
Breaking this down further:

- X variable = number of teachers.
- Y variable = test score.
- Equation:

$$f(x) = 3x + 15$$

The function tells us that for every increase in teachers, the score of the student increases by 3. For a student with no teachers, the score is 15.

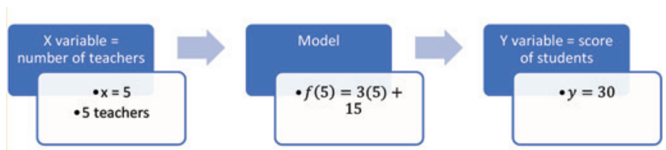
Let us test this out:



© Anusha Vissapragada

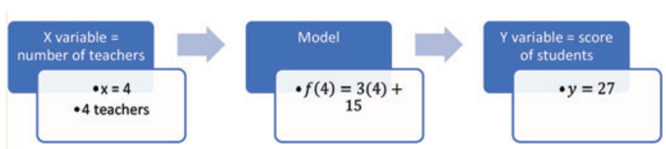
Now, let's look at some more examples:

Example 1—5 teachers



© Anusha Vissapragada

Example 2—4 teachers

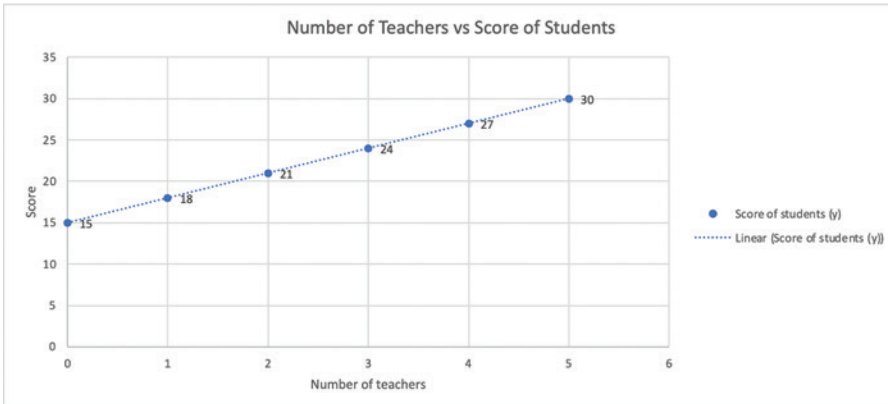


© Anusha Vissapragada

In conclusion, we can tabulate our results:

Number of teachers (x)	**Score of students (y)**
0	15
1	18
2	21
3	24
4	27
5	30

When we plot our graph, the number of teachers is our independent x variable, which would be on the x axis. The score is our dependent variable on the x, which would be on the y-axis.



© Anusha Vissapragada

This equation follows a linear form with a slope and intercept, indicative of a linear relationship. Further exploration of linear regression will be conducted in upcoming chapters. In this context, predictive analytics would aim to determine the optimal number of teachers, allocate more hiring resources effectively, and forecast student scores based on the number of teachers present.

Predictive analytics is the third step that is more complex, requires more time, and results in a higher value for the analysis. A key difference for predictive analytics is that it looks at causal relationships. The equation tells us that the number of teachers has an impact on student scores, showing us causality.

As trends persist and patterns emerge consistently, predictive analytics transitions into a policymaking tool known as prescriptive analytics. This approach goes beyond forecasting to prescribe specific actions or strategies based on predictive insights, enabling organizations to make proactive decisions and optimize outcomes.

2.6 Prescriptive Analytics

Prescriptive analytics represents a theoretical concept often applied in policymaking. When tested hypotheses consistently yield common results, predictive models become more accurate, naturally leading to prescriptive analytics as the next step. An example of prescriptive analytics includes tried-and-tested methods that reliably produce desired outcomes and have become common knowledge. Recommendations and policymaking processes often leverage prescriptive analytics to inform decision-making.

Given the complexity and high value of the insights provided, prescriptive analytics typically requires the most time to generate. This is due to the rigorous analysis and validation processes involved in ensuring the reliability and effectiveness of the recommended actions or policies. As such, while prescriptive analytics offers invaluable guidance for decision-makers, it also demands careful consideration and thorough examination of the data and underlying assumptions.

2.7 What Type of Analytics to Use?

There is never an answer to this question, as it depends on the needs of the business at that point in time. However, some guidelines to keep in mind are that time, money, value, and complexity are directly proportional to each other in analytics.