Academy Overview
From STC
Data Analytics By IBM

Program Details:

- **Duration:** 4 weeks
- **Language:** English
- **Start Date:** 28 October 2024
- **Enrolment Window:** Until 24 October 2024

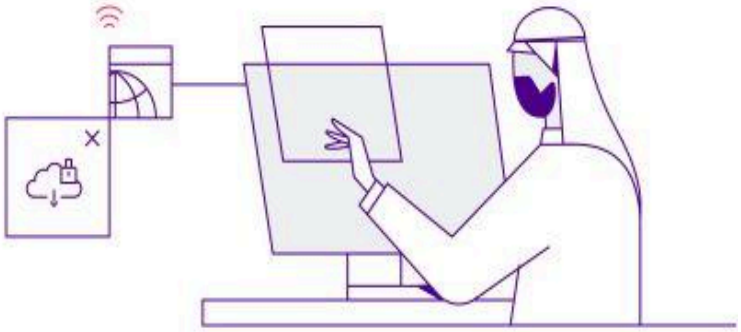
Upon Successful Completion:

- Learners will earn a globally recognized digital credential verified by Credly.
- An IBM Certificate will also be awarded.

Register Today
Join this opportunity to gain invaluable insights, learn from industry experts, and earn a prestigious digital credential and certificates!

Please find the registration link in the caption below.

academy
from stc



IBM data analytics

duration	language	enrolment window	start date
4 weeks	english	until 24.October.2024	28.October.2024

upon successful completion of the program learners will earn

globally recognized digital credential verified
by Credly and an IBM Certificate.

---

register

today and join this opportunity to gain invaluable insights, learn from industry experts, and earn a prestigious digital credential and certificates!
* Please find registration link in the caption below.

Introduction to Data Concepts

Do you know that data is everywhere, and data analytics is an essential part of your daily life? It’s what you do with the data that matters. That’s where data analytics comes in.

Data analytics helps people:

- Keep airplanes on schedule so people can get to their destinations on time.
- Ensure the most popular sizes and colors of clothing are in stock so customers can find a perfect fit.
- Shorten disaster response times so emergency responders can get to people faster.
- Optimize hospitals so operating rooms and beds are available when patients need them.
- Identify threats to stop crime in communities to keep people safe.

Learning Objectives

After completing this course, you should be able to:

- Explain the importance of data in a digital world.
- Differentiate between unstructured and structured data.
- Identify the purpose of a database.
- Differentiate between quantitative and qualitative data.
- Describe the five V's of big data.
- Describe the four types of data analytics.
- Explain the work involved in each step of the data analysis process.
- Identify the purpose of data visualization.

- Recognize different charts to display data visualizations in the best way.

Estimated Duration:

1 hour, 15 minutes

Completion Criteria

To get credit for completing this course, you need to:

- Visit all the pages in the course.
- Complete the interactive activities and quizzes.
- Pass the final assessment.

You must pass the quiz at the end of each module to continue to the next module. There is a final, graded assessment at the end of the course. You must achieve 80 percent to pass. Don’t worry! If you aren’t successful at first, you can retake the quiz as many times as needed to pass.

Copyright Information

© Copyright IBM Corporation 2024

Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

Version Information

Version: 1.1.0

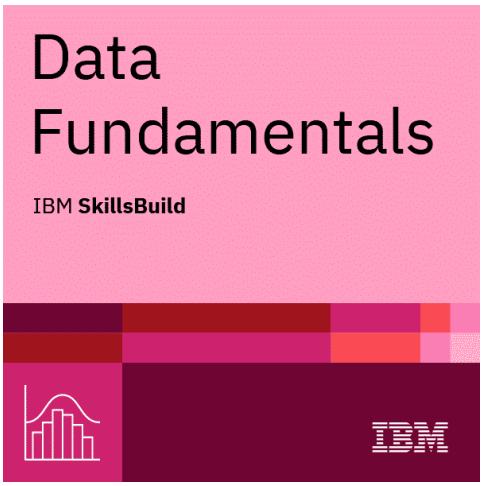
Date: 20-September-2024

Release Notes: Improved final assessment questions, including adding correct and incorrect feedback; added instructional text to quizzes.

Ready to Get Started?

Data Fundamentals Series

Earn the Data Fundamentals Digital Credential from IBM!



This course is one of several in the **Data Fundamentals** series. Complete all the courses in the Data Fundamentals learning plan and earn the **Data Fundamentals** digital credential from IBM!

Credential:

Data Fundamentals

For more information about this credential, visit [Data Fundamentals] on Credly.

What's a Digital Credential?

A credential is an industry-recognized and validated indicator of knowledge and skills acquisition. **IBM SkillsBuild** credentials are aligned to job-market standards and frameworks to help achievers gain visibility with potential employers. IBM SkillsBuild issues its digital credentials through Credly.

Share Your Credentials

Share your credentials on LinkedIn and other social networks with prospective employers, or include them on your resume to amplify your skills!

Course Design

Learn at Your Own Pace with Structured Modules

This course is designed so you can learn at your own pace and easily navigate using the organized module structure. You'll find:

- **Concepts, graphics, and examples** to help you learn.
- **Interactions** to reinforce what you learn.
- **Quizzes** to check your knowledge.
- **Helpful resources** so you can explore more.
- **A final assessment** to test your understanding.

Tips to Be a Successful Learner

- The course is designed for you to complete the modules in sequential order.
- You can pace your learning over a period that works for you and your schedule.
- Your progress is tracked, so you can complete a module or lesson and then come back to pick up where you left off.

Types of Data

Introduction

In this module, you will learn about the importance of data in your everyday life and the forms and differences between the forms of data. You'll also learn about the differences between quantitative and qualitative data and review some examples.

Learning Objectives

After completing this module, you should be able to:

- Explain the importance of data in a digital world.
- Differentiate between unstructured and structured data.
- Identify the purpose of a database.
- Differentiate between quantitative and qualitative data.

Why Is Data Important?

Our world is rich with data. Any business that has a website or a social media presence and accepts electronic payments is collecting data. Businesses collect data about customers, user habits, web traffic, demographics, and more. All of that data is filled with potential if you can learn how to find it.

Data is Critical for Business

Data is one of the most valuable assets organizations can have, whether in business, finance, healthcare, retail, technology, marketing, or other industries. The number of companies using data insights continues to grow. Data insights have the potential to help many companies:

- Improve operations
- Better understand end users or customers
- Drive efficiency
- Reduce costs
- Increase profits
- Find new innovations

Data Can Be a Life Saver

Nonprofit organizations, governments, and academic facilities also strive to find insights in data that can help solve global challenges like poverty, climate change, disease, inequality, and terrorism. They use innovative technologies and research for social, health, and environmental benefits.

Data is a Problem Solver

Data analysts make sense of data. They follow a process to gather, cleanse, and transform data with the objective of obtaining insights to answer the question, *What is the data telling us?* The goal is to use the insights to help make better and more informed decisions, and solve problems.

Explore More!

Check out Our World in Data, whose mission is to publish research and data to make progress against the world's largest problems.

How Much Data is There?

From traffic patterns and music downloads to web histories and medical records, data is recorded, stored, and analyzed to enable the technology and services that the world relies on every day.

How Much Data Does the World Create Every Day?

It's estimated that the world creates **2.5 quintillion bytes of data** every day. A quintillion is the number 1 followed by 18 zeroes! That's a big "wow factor!"

Data Creation Statistics:

- People make **3 billion minutes** of calls on Skype per day.
- People share **5 billion Snapchat videos and photos** per day.
- People send **333.2 billion emails** per day.
- People spend **1 million USD per minute** online.

The Data Explosion Continues

By 2025, global data creation is projected to grow to more than **180 zettabytes**. A zettabyte is the number 1 followed by 21 zeroes!

In 2020, the amount of data created and replicated reached a new high. The growth was higher than previously expected due to the increased demand during the

COVID-19 pandemic as more people worked and learned from home, and used home entertainment options more frequently.

Data Growth Visualization

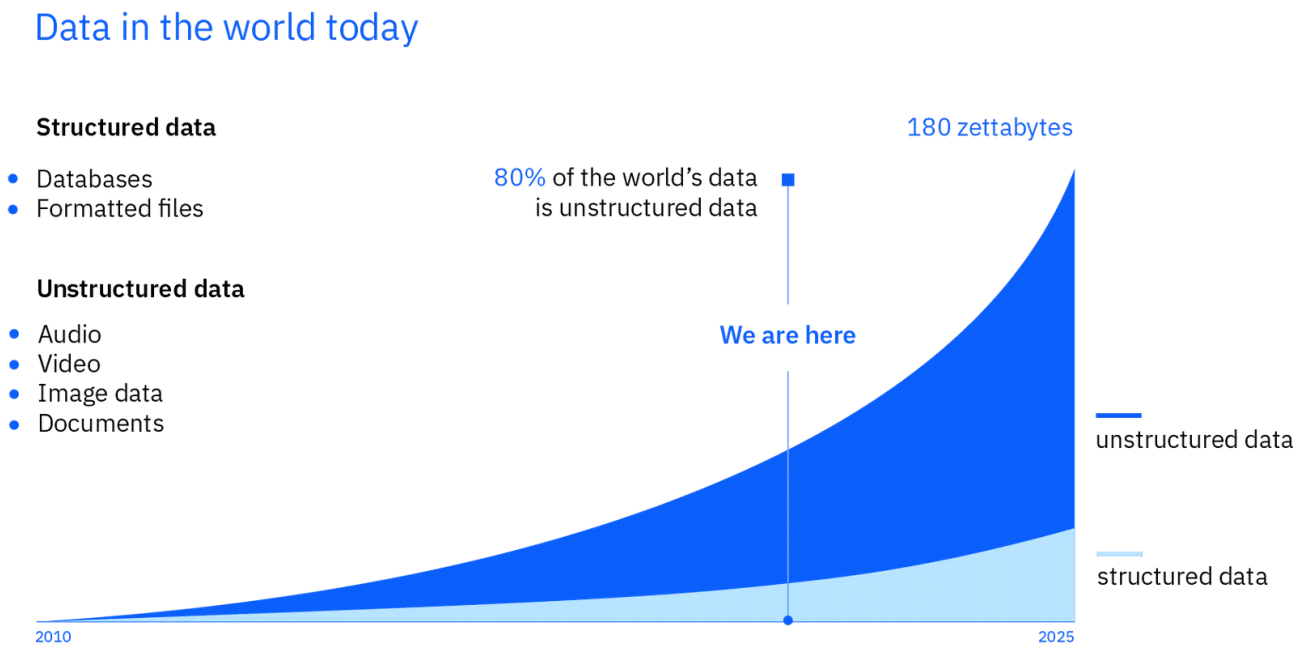
This diagram shows the exponential growth of data:

- **Structured data** includes databases and formatted files.
- **Unstructured data** includes audio, video, image data, and documents.
- **80% of the world's data** is unstructured.

The chart shows the growth from 2010 to 2025, with unstructured data increasing exponentially.

Explore More!

Check out [What Happens in an Internet Minute](), where you can see all the incredible things people are doing online every minute—from checking emails to buying products and watching videos.



Understanding Structured and Unstructured Data

What is Data?

Data is raw information that can take many forms, including facts, statistics, opinions, voices, photos, names, and even dance moves!

Data analysts organize this raw information into two main forms: **structured data** and **unstructured data**. Understanding the difference between these types of data is essential, as it impacts how data analysts work with them.

Structured Data

Structured data is information that can be organized in rows and columns. This type of data is highly organized, making it easier to manipulate and analyze. Examples include:

- Names, dates, addresses
- Credit card numbers
- Stock market information
- Geolocation data

Example of Structured Data: Hardware Store Customer Records

Customer Number	Last Name	Phone Number	Last Order Date
00001	Ajay	(555) 678-9012	12-Mar-2020
00002	Thompson	(555) 345-6789	14-Aug-2021
00003	Smith	(555) 432-1098	01-Aug-2021
00004	Kim	(555) 665-5443	16-Nov-2021
00005	Gonzalez	(555) 912-9945	29-Dec-2019
00006	Wangzi	(555) 212-3767	30-Jun-2022
00007	Colbert	(555) 866-0922	09-Jan-2022
00008	Jarrah	(555) 778-1845	12-Sep-2020
00009	Magnusson	(555) 395-7677	14-Apr-2022
00010	Taki	(555) 550-5515	30-Oct-2021

Why is Structured Data Important?

Structured data is widely used because it is easy to analyze and process using tools like spreadsheets (Google Sheets, Microsoft Excel) and databases (SQL). It allows businesses to:

- Organize and retrieve data efficiently
- Improve customer insights
- Optimize operations
- Enhance decision-making

Unstructured Data

Unstructured data refers to information that does not follow a predefined format. This type of data is often more challenging to work with but holds significant value for uncovering insights.

Examples of Unstructured Data:

- Images and videos
- Social media posts (tweets, comments, blogs)
- Customer reviews and feedback
- Medical records
- Song lyrics

Structured vs. Unstructured Data: Key Differences

- **Department of Motor Vehicles (DMV)**
  - A collection of driver's license photos = **Unstructured data**
  - A table listing names, addresses, and license numbers = **Structured data**
- **Digital Library**
  - The text of thousands of books = **Unstructured data**
  - A catalog listing book names, authors, and dates = **Structured data**

The Challenge of Unstructured Data

While structured data is easier to process, unstructured data holds untapped potential. It can help organizations:

- Identify customer sentiment from reviews
- Analyze trends in social media posts
- Detect patterns in medical records

Interesting Fact!

Approximately **80% of the world's data is unstructured**, while only **20% is structured**. This highlights the growing need for advanced data analytics tools to interpret unstructured data effectively.

Test Your Knowledge!

Match the data types to their correct category:

- **Structured Data:** Dates, Credit card numbers, Addresses
- **Unstructured Data:** Videos, Social media posts, Song lyrics

---

Final Thoughts

Understanding structured and unstructured data is crucial in today's data-driven world. Data analysts use various tools and techniques to organize, analyze, and extract valuable insights from both types of data.

Are you ready to dive deeper into data analytics? Let's explore more!

Databases

What is a Database?

- An organized collection of structured data in a computer system.
- Acts as a container or repository with many columns and rows.
- Ensures data is easily accessible by queries and computation software.

Importance of Data Organization

- Transforming data into a standard format (tidy data) simplifies storage and analysis.

Relational Databases

- Consist of multiple datasets or tables that are linked together.
- Example:
    - One table lists names and addresses.
    - Another table lists properties and their owners.
    - If some owners appear in both tables, they can be linked to form a relational database.

Think of Data as Qualitative or Quantitative

Two Main Types of Data:

1. **Quantitative Data** (Numerical Data)

    - Represents things that can be measured and assigned values.
    - Can be counted and measured.
    - Examples: height, weight, length, blood pressure, temperature.
2. **Qualitative Data** (Categorical Data)

    - Represents characteristics, attributes, properties, and qualities.
    - Describes data using language rather than numbers.
    - Examples: smell, location, color, texture, marital status.

Scenario: Identifying Quantitative vs. Qualitative Data

- A woman is walking a small dog.
- **Quantitative data about the dog:**
    - The dog has four legs.
    - The dog weighs 12 pounds.
- **Qualitative data about the woman:**
    - The woman has gray hair.
    - The woman is wearing sunglasses.

Summary

- **Quantitative Data:** Numerical, countable, measurable (Think **quantity**).
- **Qualitative Data:** Descriptive, categorical, non-numerical (Think **quality**).

Further Distinctions in Data

1. **Quantitative Data** can be:
    - **Discrete Data**: Whole numbers that can't be divided (e.g., number of rooms in a house, number of people in a theater).
    - **Continuous Data**: Can be divided into finer levels (e.g., weight of a car, height, temperature, speed of an airplane).
2. **Qualitative Data** can be:
    - **Nominal Data**: Names or labels without a meaningful order (e.g., hair color, religion, marital status, ethnicity).
    - **Ordinal Data**: Ordered data with meaningful ranking (e.g., school letter grades, clothing sizes, customer satisfaction levels).

Visual Representation

- **Quantitative Data:**
  - **Discrete:** Whole numbers (e.g., 3 children, 76 DVDs).
  - **Continuous:** Measurable data with decimal points (e.g., 13.15 minutes, 56.327 km/h).
- **Qualitative Data:**
  - **Nominal:** Labels (e.g., Greek, blonde hair, married).
  - **Ordinal:** Ordered data (e.g., Likely, Neutral, Unlikely).

This breakdown helps to better understand and categorize data effectively!

Big Data

Introduction

In this module, you will learn about big data and its characteristics, known as the **5 V's**:

1. **Volume** - The massive amount of data generated every second.
2. **Variety** - The different types of data, including structured, unstructured, and semi-structured.
3. **Velocity** - The speed at which data is created, collected, and processed.
4. **Veracity** - The reliability and accuracy of data.
5. **Value** - The usefulness of data in decision-making.

Learning Objectives

After completing this module, you should be able to:

- Describe the **5 V's of Big Data**.
- Understand how these characteristics impact data analysis and decision-making.
- Recognize the importance of managing and processing large datasets effectively.

Big data plays a crucial role in modern analytics, helping organizations gain insights, make predictions, and drive innovation. This module will explore real-world applications and techniques used to handle big data efficiently.

Big Data

Introduction

In this module, you will learn about big data and its characteristics, known as the **5 V's**:

1. **Volume** - The massive amount of data generated every second.
2. **Variety** - The different types of data, including structured, unstructured, and semi-structured.
3. **Velocity** - The speed at which data is created, collected, and processed.
4. **Veracity** - The reliability and accuracy of data.
5. **Value** - The usefulness of data in decision-making.

Understanding Big Data

Many people interact daily with an increasing number of internet-connected devices that record data, from text messages and school records to travel habits and movie preferences. In a digital world, every user leaves a trail of data, commonly referred to as **big data**.

While the term **big data** has existed for years, there is no universal definition. Below are perspectives from experts in the field:

Everything People Do Leaves a Trace That Can Be Analyzed

Bernard Marr, a thought leader in business and technology, describes big data as:

> "The basic idea behind the phrase 'Big Data' is that everything we do is increasingly leaving a digital trace (or data) which we (and others) can use and analyze to become smarter. The driving forces in this brave new world are access to ever-increasing volumes of data and our

> ever-increasing technological capability to mine that data for commercial insights."
>
> — *Big Data: Using SMART Big Data, Analytics, and Metrics To Make Better Decisions and Improve Performance (2015)*

Big Data is More Complex and Can Tackle New Problems

According to **Oracle**:

> "Put simply, big data is larger, more complex data sets, especially from new data sources. These data sets are so voluminous that traditional data processing software just can't manage them. But these massive volumes of data can be used to address business problems you wouldn't have been able to tackle before."

Big Data Demands Innovative Processing to Get Insights

The research firm **Gartner** defines big data as:

> "Big Data is high-volume, high-velocity, and/or high-variety information assets that demand cost-effective, innovative forms of information processing that enable enhanced insight, decision making, and process automation."

Learning Objectives

After completing this module, you should be able to:

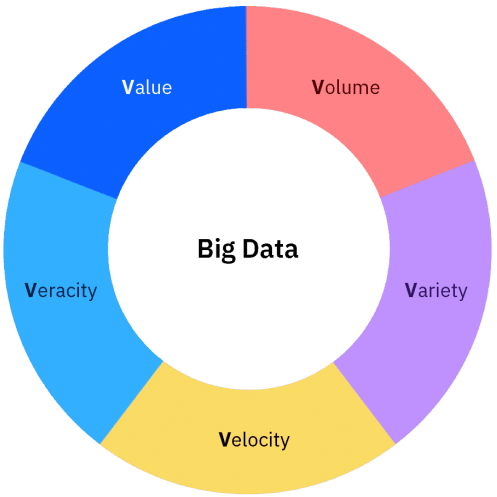
- Describe the **5 V's of Big Data**.
- Understand how these characteristics impact data analysis and decision-making.

- Recognize the importance of managing and processing large datasets effectively.
- Identify real-world applications of big data across various industries.

Big data plays a crucial role in modern analytics, helping organizations gain insights, make predictions, and drive innovation. This module will explore real-world applications and techniques used to handle big data efficiently.

Big Data

The 5 V's of Big Data



The elements of big data are often explained using five broad characteristics called the **5 V's**: **Volume**, **Variety**, **Velocity**, **Veracity**, and **Value**. These help data scientists to understand and manage what they are working with. Below is an overview of each of the 5 V's.

---

1. Volume

What amount of data do you need to process?

Volume refers to the vast amount of data being generated. The scale of data is growing exponentially, with units such as:

- **One gigabyte (GB)** = approximately 1 billion bytes
- **One exabyte (EB)** = one billion gigabytes
- **One zettabyte (ZB)** = 1,000 exabytes
- **One yottabyte (YB)** = 1,000 zettabytes

The increasing number of data sources, driven by the widespread use of digital devices (smartphones, computers, game consoles, etc.), contributes to this growing volume of data.

Companies must adopt robust tools and storage solutions capable of managing and processing these massive volumes of data.

---

2. Variety

Is the data structured or unstructured?

Variety refers to the different types of data being processed.

- **Structured data**: This includes well-organized data like spreadsheets, relational databases, and other formats that are easy to process.
- **Unstructured data**: This includes things like social media posts, emails, audio files, images, blog content, and other forms of data that don't follow a specific structure.

With the increasing variety of data, businesses need diverse tools and techniques to process both structured and unstructured data effectively.

3. Velocity

How quickly is the data being generated from its diverse sources?

Velocity refers to the speed at which data is generated, processed, and moved.

Data is constantly being generated at an incredibly fast rate. For example:

- Every minute on Facebook, there are 510,000 comments, 293,000 status updates, and 240,000 photos uploaded by active users.

As data moves across systems in real time or near-time, businesses need tools capable of processing and utilizing this data quickly.

---

4. Veracity

How accurate is the data?

Veracity refers to the quality and trustworthiness of data. Not all data is of high quality, and it's essential to consider the source to ensure that the data is:

- Complete
- Reliable
- Traceable

For example, social media posts like tweets may include slang, abbreviations, or typographical errors that reduce their accuracy. Therefore, businesses need methods and resources to ensure the data they work with is clean and trustworthy.

---

5. Value

Can you extract value and insights from the data?

Value refers to the ability to extract meaningful insights and actionable information from the data. The main goal of working with big data is to derive value from it.

This value might not always be monetary. It could be in the form of:

- Social or medical benefits
- Improving customer or employee satisfaction
- Gaining insights into business performance

Companies must understand the value they want to obtain from their big data and focus on extracting the most relevant, valuable information while filtering out the noise.

Resources for Further Learning

- **Our Digital World and Big Data**: A free online course offered by Future of Tech from CompTIA.
- **How big is a petabyte, exabyte, or yottabyte?**: A blog post by ZME Science that explains the data units from bytes to yottabytes.
- **Big Data in Practice**: A blog by Bernard Marr that covers 10 major areas where big data is used.

About This Module

Module 3: Types of Data Analytics

Introduction

In this module, you will learn about the four types of data analytics: **Descriptive Analytics**, **Diagnostic Analytics**, **Predictive Analytics**, and **Prescriptive Analytics**. Each type of analytics plays a crucial role in helping businesses and organizations gain insights from data. You will explore the questions that each type of data analytics can answer and how they are applied in various real-world scenarios.

Learning Objectives

After completing this module, you should be able to:

- Describe the four types of data analytics:
    - **Descriptive Analytics**
    - **Diagnostic Analytics**
    - **Predictive Analytics**
    - **Prescriptive Analytics**

Data Analytics Answers Questions

In a constantly changing business environment, predicting the next move can be challenging. That's where data analytics comes in. A successful data analytics initiative can provide a clear picture of where you are, where you've been, and where you should go.

The information technology (IT) industry recognizes four types of data analytics:

- **Descriptive Analytics**
- **Diagnostic Analytics**
- **Predictive Analytics**
- **Prescriptive Analytics**

Each type of data analytics serves a different purpose in the data analysis process and answers a specific question. The following diagram illustrates the degree of complexity and added value of each type of data analytics.

Data Analytics Complexity and Value

A diagram showing the four types of data analytics that grow in value as they grow in complexity, going in order from **Descriptive Analytics** to **Diagnostic Analytics** to **Predictive Analytics** to **Prescriptive Analytics**.

As depicted in the diagram, the difficulty and resources required increase with each type of data analytics, but so does the level of insight and value.

Continue the lesson to explore each of the types of data analytics and understand how they build upon each other.

Descriptive Analytics: What Is Happening?

Descriptive analytics is the simplest and most common type of data analytics.

It answers the question: **"What is happening?"**

Descriptive analytics provides a snapshot of business trends and patterns by analyzing historical and current data. It manipulates raw data from various sources to offer insights into past performance and key business metrics. While it identifies what has occurred, it doesn't explain why it happened. However, it can point to areas that need further investigation.

Examples of Descriptive Analytics Data:

- Number of subscribers for a service
- Customer satisfaction survey results

- Monthly revenue reports
- Daily stock reports
- Demographic data (e.g., 30% of customers are self-employed)

Diagnostic Analytics: Why Is It Happening?

After asking, “What is happening?” the next step is to dive deeper and ask, “Why are trends and patterns happening?”

Diagnostic analytics builds upon descriptive analytics by finding the causes of observed issues. It helps identify patterns of behavior and connects data points to uncover the reasons behind certain trends or outcomes.

Examples of Diagnostic Analytics:

- A freight company investigates the cause of slow shipments in a region.
- A healthcare company examines diagnoses and prescribed medications to determine their influence.
- An IT company analyzes server ticket data to identify servers causing service outages.

Predictive Analytics: What Is Likely to Happen in the Future?

Predictive analytics uses historical data to forecast future events. It involves predicting future outcomes, trends, or behaviors based on past data. Predictive models allow businesses to make informed decisions by understanding the likelihood of future events.

Predictive analytics often relies on machine learning and deep learning techniques, making it more advanced than descriptive and diagnostic analytics.

Examples of Predictive Analytics:

- A software company uses customer segmentation to determine potential sales leads.
- An automotive manufacturer forecasts the failure rate of a vehicle part to recommend service actions.
- A weather forecaster analyzes current conditions to predict weather patterns in different regions.

Prescriptive Analytics: What Should Happen?

Prescriptive analytics takes insights from descriptive, diagnostic, and predictive analytics to recommend specific actions. It not only suggests what to do but also provides guidance on the best course of action to take to address a problem or capitalize on an opportunity.

Prescriptive analytics requires significant technological resources, including machine learning, business rules, and algorithms. While it's complex to implement, it offers valuable insights for decision-making.

Examples of Prescriptive Analytics:

- A traffic application helps people choose the best route home, considering distance, travel speed, and current traffic conditions.
- An exam timetable checks for student scheduling conflicts.
- Artificial intelligence (AI) systems used by companies like Facebook, TikTok, and Netflix to recommend content.

Conclusion

Each type of data analytics serves a distinct purpose and provides different insights. Moving from descriptive analytics to predictive and prescriptive analytics requires advanced technical skills but provides much more value to organizations. The progression from one type to the next unlocks more sophisticated insights, enabling businesses to make better-informed decisions.

Sources:

- 4 Types of Data Analytics to Improve Decision-Making by Alex Bekker, May 2019
- The 4 Types of Data Analytics by Thomas Maydon, July 2017
- Types of Data Analysis by Patrick Gibson, 2021

Steps to Analyze Data

Introduction

In this module, you will learn about the four essential steps to follow in the data analysis process: **Collect**, **Clean**, **Analyze**, and **Visualize**. These steps are crucial for ensuring that data is properly prepared, interpreted, and presented to uncover valuable insights.

Learning Objectives

After completing this module, you should be able to:

- Explain the work involved in each step of the data analysis process:
    - **Collect**
    - **Clean**
    - **Analyze**
    - **Visualize**

Steps You Can Easily Remember

Data analysis is the process of collecting, cleaning, and transforming data to obtain insights that help make better and more informed decisions. In today's data-driven world, this is essential for companies of all sizes to solve everyday business problems. Each company has its own team, processes, and tools for data analysis projects.

Before beginning a data analysis project, a good starting point is to have a business objective in mind and a clearly defined problem. **What questions do you want to answer?** After all, 90% of analytics is asking the right questions! You must also determine the metrics to measure performance. **What are you measuring, and how are you measuring it?**

While there is no one-size-fits-all data analysis process, the following steps are typically followed:

Walkthrough with a Data Analyst

Photograph of a man with glasses, representing Charles Ito.

Meet **Charles Ito**, a marketing data analyst with over two years of experience at his current company. Charles will guide you through the data analysis process, offering insight into each step: **Collect**, **Clean**, **Analyze**, and **Visualize**.

Introduction

Charles will take you through his data analysis journey. As a data analyst, he collaborates with various team members like data scientists and data journalists. Understanding the workflow Charles follows on a project will help you appreciate the roles and processes of data analysis.

Step 1: Collect

“This is ‘square one’ in the process. This step is all about collecting the right data and just enough data for the project’s questions or problems that we want to research.

I first determine the data I can collect from any existing sources and databases that we already have that relate to the problem my company wants to solve. I always collect this data first!

Then, I figure out if my project needs new sources of data because this could mean more time for the project and potentially more of an investment from my business group.

My team and I use our company's data collection tools and follow the data collection guidelines. We're careful to securely store the data on our cloud servers, too.

One crucial point I'd like to make is you have to collect enough in your data set, so you don't skew the results of your analysis."

Step 2: Clean

"Next, not all the data I collect will be useful, so it's time to clean it up!

Data cleaning is the process of detecting and correcting missing or inaccurate records from a data set.

A big part of this step is making sure that the data is in a usable format. This involves searching for what we call 'outliers,' dealing with null values, and looking for data that may have been incorrectly input. Simply put, raw data will have missing and inaccurate values that I need to address.

You might have heard the term, 'data wrangling.' I 'wrangle' the data so it's in a usable format for my project in our database system. For example, I will search for duplicate records and remove them.

No two data sets are the same, so how I clean the data can vary. I clean the data based on the context. In one case, seeing a blank entry might equal a zero entry, so it's good and valuable data. But, in another case, seeing a blank entry could mean it's incomplete data that I need to exclude. This is the art of data science!

Always save your data since this is an iterative process!

Oh, and here's an interesting fact. This is where I spend most of my time, cleaning the data! I'd estimate that data analysts typically spend about 70-80% of their time cleaning data. It's a lot of hard work. But, it's a must so I can move on to analysis."

Step 3: Analyze

"Once I have the relevant data and it's cleansed, it's time to analyze. This is the step where data analysts spend about 20-30% of their time. It's the fun and rewarding part!

I get to be curious and investigate. And, my problem-solving skills come into play. Here, I use different statistical and analytical methods and software tools. It's important that I align my methods of analytics, so they match the intent of the problem.

Basically, I identify issues and use analytics to determine the root causes of issues. I analyze trends, correlations, variations, and outliers to help me focus on answering the questions (and any questions or objections others might have).

As I manipulate data, I might find I have the exact data I need, but, more likely, I might need to revise my original questions or collect more data. This can drive additional analysis and is one reason to always save your data."

Step 4: Visualize

"After analyzing the data, I interpret the results. I think about, 'What did I learn from the results of my analysis? Does the data answer my original questions? How?'

To interpret the results, I create data visualizations. They help me compare data sets and observe relationships. We have a couple of tools that I just love using for this.

And, once I can make my conclusions, the visualizations help me graphically show my results in a way my team and leadership can easily and concisely understand them. I find I often choose bar charts, line graphs, scatter charts, and maps. I like figuring out the best way to depict the data.

We typically have a team presentation and meeting to review the results and data visualizations. Our goal is to make sure the insights that I've concluded can be applied to the business problem and we can take actions, or make changes, or refocus efforts however is needed.

This step is a good ending to the process. It's all about gaining insights from the data and being able to apply the insights to the business."

Summary

This walkthrough presented the four essential steps that Charles takes in his data analysis process: **Collect**, **Clean**, **Analyze**, and **Visualize**. Although the specifics can vary based on the project, these steps offer a solid framework for understanding the core responsibilities of a data analyst.

Good to Know! Extract, Transform, and Load (ETL)

You may hear the term "ETL" used in relation to data analytics and warehousing. **ETL** stands for **Extract**, **Transform**, and **Load**. It's a data integration process used to combine data from various sources into a single, consistent data store.

ETL allows organizations to:

- Extract data from legacy systems
- Cleanse data for consistency and quality
- Load the data into a target database for further analysis

ETL forms the foundation for analytics and machine learning projects, making it a crucial concept for data analysts to understand.

Visualize the Data

Introduction

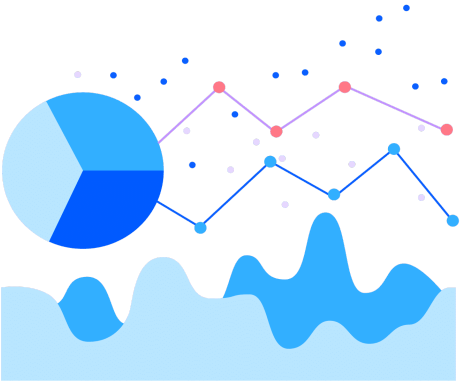
In this module, you will learn about the value of data visualizations and how to tell a story with data. You'll explore different types of visualizations and see examples of how to use them effectively.

Learning Objectives

After completing this module, you should be able to:

- **Identify the purpose of a data visualization**: Understand why visualizations are important in data analysis and how they help convey insights.
- **Recognize different charts to display data visualizations in the best way**: Learn about various types of charts and how to choose the right one for your data and message.

It's Time to Visualize and Tell a Story



What good is data if you can't communicate it in a way that people understand? This is where visualizations come into play.

A **data visualization** is a graphical display of abstract or complex information. Data analysts use visualizations like charts, graphs, and maps for two key purposes:

1. **To explore and interpret data** during analysis to identify patterns or trends.
2. **To communicate results** and help people understand the insights to make informed decisions.

An abstract rendering of a **pie chart**, **line graph**, and **scatter plot** combined to look like a shooting star over a mountain range.

It has been said that **"a picture is worth a thousand words."** This is particularly true when creating a data visualization. A single chart can answer multiple questions. Graphical representations bring clarity and simplicity to complicated issues.

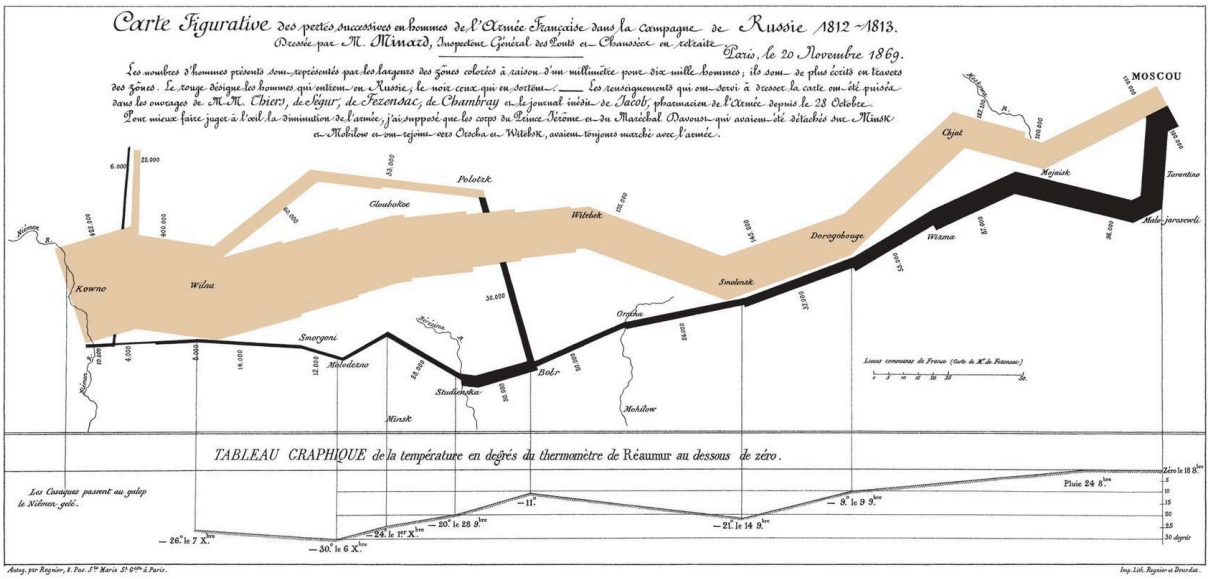
An Early Visualization

Here is **Charles Minard's 1869 chart** showing **Napoleon Bonaparte's 1812 Russian campaign** of his army and their movements.

Details of the chart are described in the course text.

Edward **Tufte**, one of the pioneers in the field of data visualization, considers Minard's chart as **"probably the best statistical graphic ever drawn."** But what makes it so magnificent?

Take a moment to zoom in and study it, then expand to reveal the answer.



Why is Tufte's Chart a Good Visualization?

This visualization vividly depicts **Napoleon's army** departing from the **Polish-Russian border** during the bitter and cold winter of 1812. The chart shows the trajectory of the march, the geography of the landscape, and specific events.

A **thick band** illustrates the size of Napoleon's army at various geographic points during their advance and retreat. The chart displays six types of data in two dimensions:

1. **The number of Napoleon's troops**
2. **The distance traveled**
3. **The temperature**
4. **Latitude and longitude**
5. **Direction of travel**
6. **Location relative to specific dates**, without directly mentioning Napoleon

Source: Wikipedia for Charles Joseph Minard

Tell a Compelling Data Story

You can tell a story about data! The goal is to have a visual that's effective, attractive, and impactful. What's amazing is that, ultimately, **data visualizations turn numbers into narratives.**

Data storytelling is the process of converting data analyses into a simple, understandable story that influences business decisions. With the rise of digital business and data-driven decision-making, data storytelling has become an important skill. The idea is to "connect the dots" between the results and decision-makers, who must be able to interpret the data.

Data storytelling involves a combination of three essential elements:

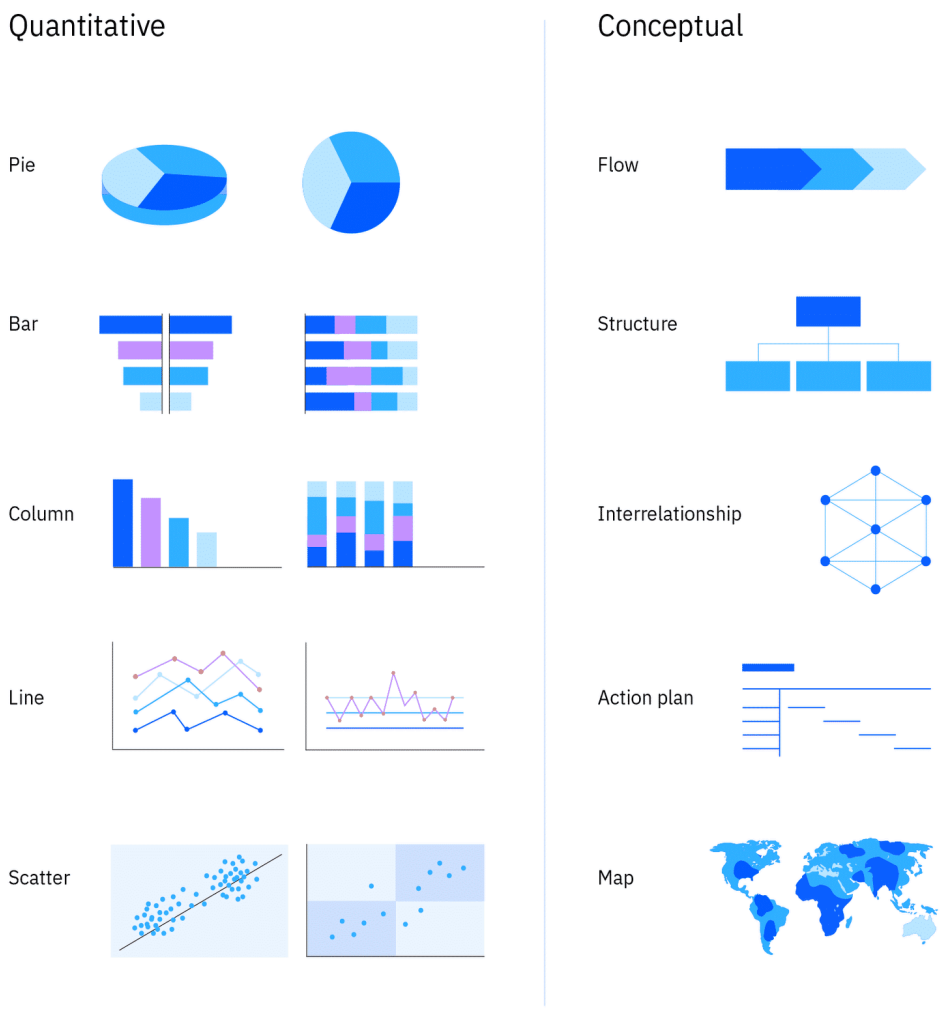
1. **Data**: The foundational facts, figures, and statistics.
2. **Visualizations**: Charts, graphs, and other visuals that help interpret the data.
3. **Narrative**: A clear explanation of what is happening in the data and why it matters.

Why is Data Storytelling Important?

- **When narrative is coupled with data**, it explains to the audience what is happening in the data and why an insight is important.
- **When visualizations are applied to data**, they reveal patterns and trends that might not be easily spotted from raw data alone. These insights might otherwise remain hidden in the rows and columns of a database.
- **When narrative and visualizations come together**, they create a powerful data story that can **influence**, **drive change**, and **engage an audience**.

Types of Visualizations

Data analysts use specific charts to visualize quantitative and qualitative data. The following image contains common charts for visualizing these two types of data. Conceptual charts can show either quantitative or qualitative data. Take a moment to study them.



Charts convey different meanings. Selecting the best chart for your visualization is important. The following image shows the types of charts that work well for

displaying different types of quantitative data comparisons, such as relative proportion, ranking, time, frequency, and correlation. Take a moment to study them.

Relative proportion
is the amount or quantity of a subset present in the population of all data points. For example, if there are 25 students in a class of which 15 are girls and 10 are boys, then the proportion of girls is 15 out of 25 (3 of 5) and the proportion of boys is 10 out of 25 (2 of 5).

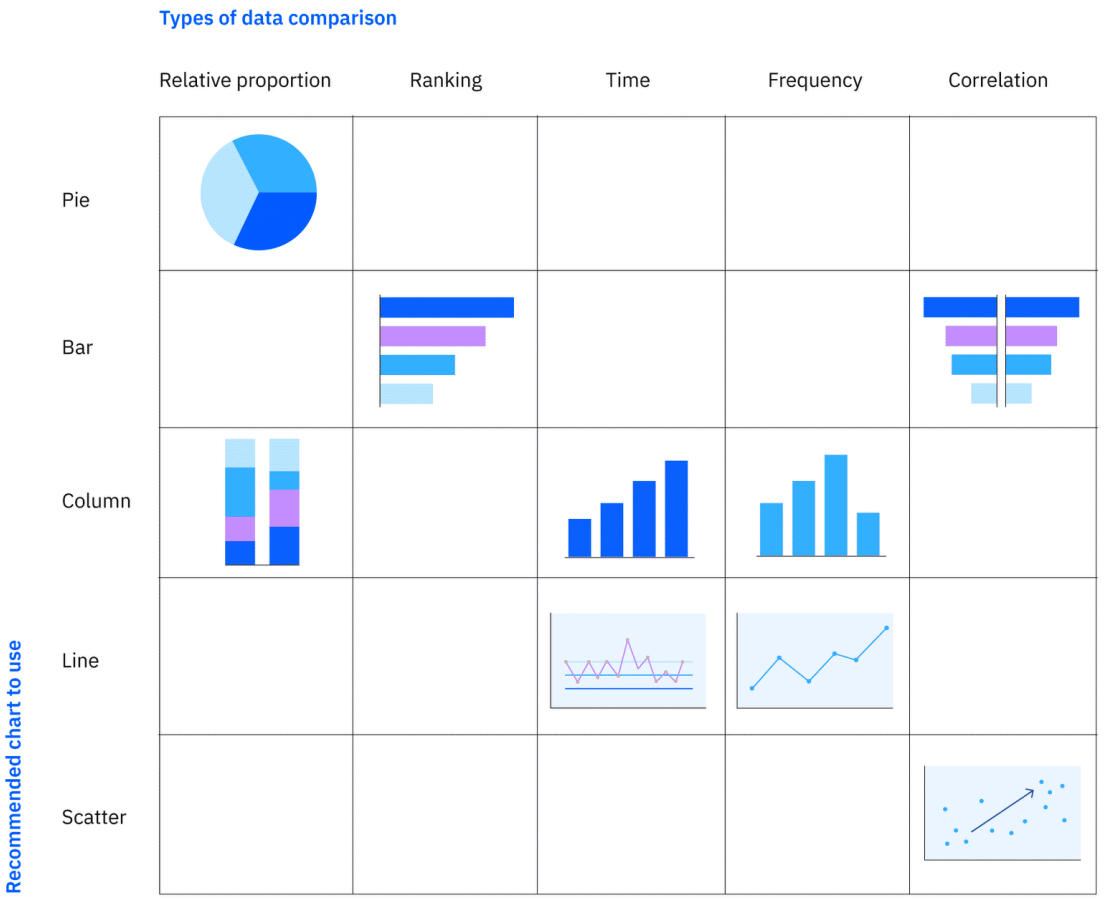
Ranking
is the relationship between a set of items, in which one item is ranked higher, lower, or the same, compared to a second item. For example, video game players can be ranked in order by their highest score in a tournament.

Time
is a series of data points that are listed or sequenced in time order, such as, for example, the daily time of high tide and low tide at a beach.

Frequency
is the number of times a certain event occurs. For example, if it snows two times today, then the frequency of snow on this particular day is 2.

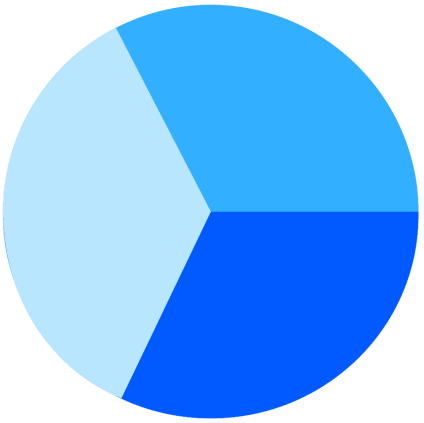
Correlation

is the relationship between two random variables, which are typically related in linear way. For example, there is a correlation between the height of parents and their children.



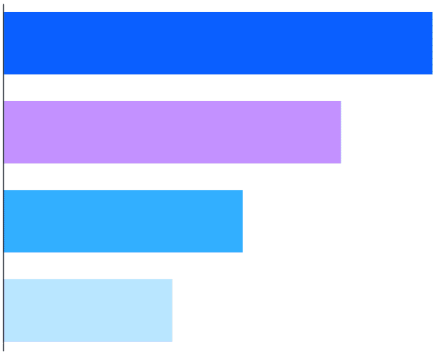
Now, explore samples of common types of quantitative charts.

Pie charts are circular, statistical graphics that are divided into slices to illustrate numerical proportions. The total sum of all “slices” must equal 100%. Pie charts are useful for showing the relative proportion for a small number of items. They can very easily show which category is the largest or has the most impact.



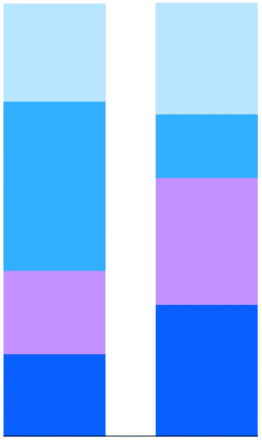
Example: A grocery store chain uses a pie chart to show the percentage of total profit from different grocery store locations and highlight the store earning the most profit.

Bar charts are useful for ranking a large number of categories, showing correlation, and using for before-after analysis. Use bar charts for comparison and ranking. Bar charts help illustrate change over time.



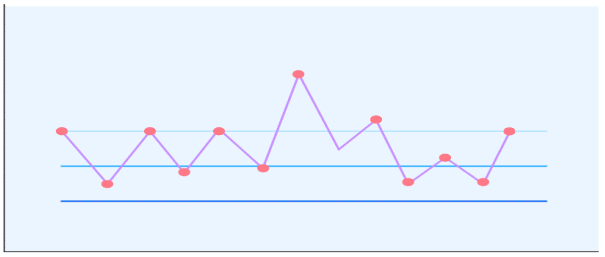
Example: An online pharmacy observes a big jump in web traffic during the COVID-19 pandemic, so they use a bar chart to capture month-to-month traffic for one year to make it easier to identify the increase in web traffic.

Column charts can show relative proportion between items and show data over time and frequency. Column charts display information vertically. They are also useful for showing negative data.



Example: A big-box company tracks the number of accidents that occur in warehouses. When the number of accidents falls below the monthly average, the company uses a column chart to distinguish this in a monthly report.

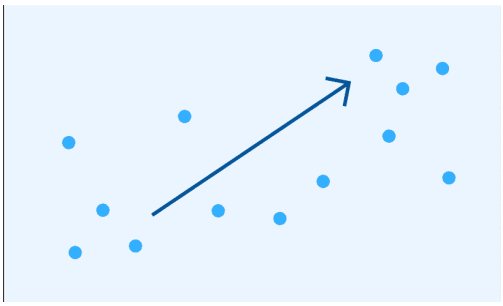
Line charts can display data over time and frequency; however they typically display this data over a continuum. Line charts can track changes over short and long periods of time. Because of this, line charts are useful for indicating small changes. They are effective in showing trends.



Example: An IT help desk must measure their service performance for closing customer issues, so they use a line graph to track how many chats and emails the team responds to per month.

Scatter plot charts display plotted points to show a relationship between two sets of data. They display a large amount of data and make outliers stand out. Scatter plot charts can also help you identify patterns.

Example: A bank uses a scatter plot chart to explain to executives that a data set of thousands of customer calls reveals that customer happiness levels are highest when call response time is lowest.



These different charts are part of a data analyst’s tool kit! Depending upon the type of analysis being performed, there are numerous programs and tools that can be used to generate these different visualizations.

Select the Best Chart

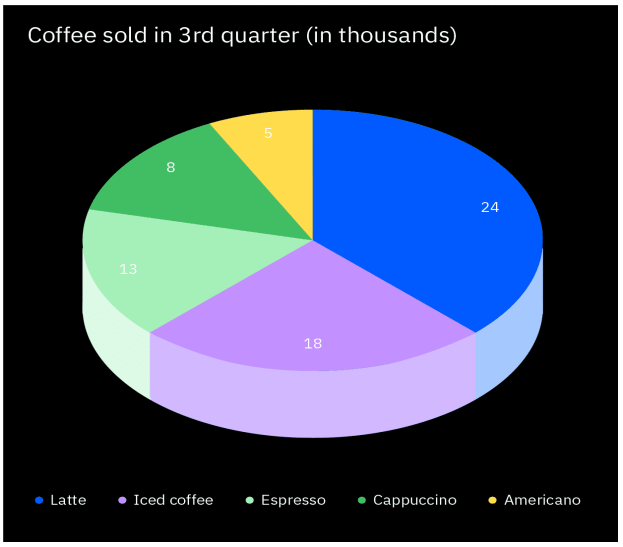
Introduction

Project objective: Design a data visualization for a national coffee company to discover and show the highest-selling type of coffee from the third quarter.

Step 1

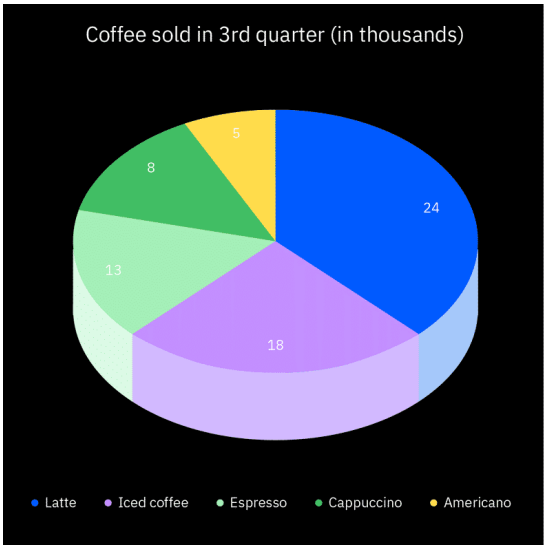
Start with a Pie Chart

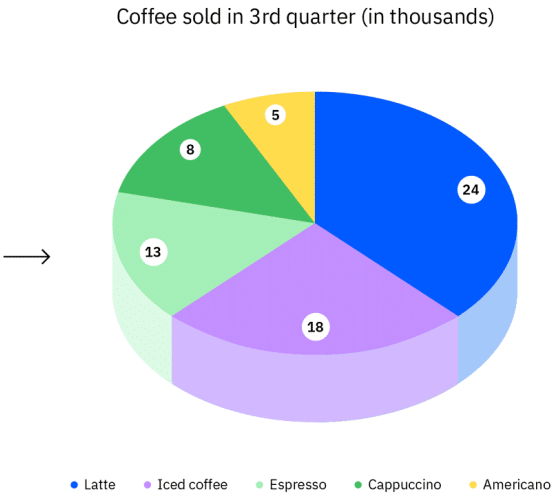
This is the first version of the data visualization. Data from a database has been exported into a three-dimensional pie chart. There are five types of coffee. Latte coffee, shown in blue, is the top-selling coffee at 24,000 units, since it's the biggest "slice" of this pie chart.



Step 2

Remove the Noise

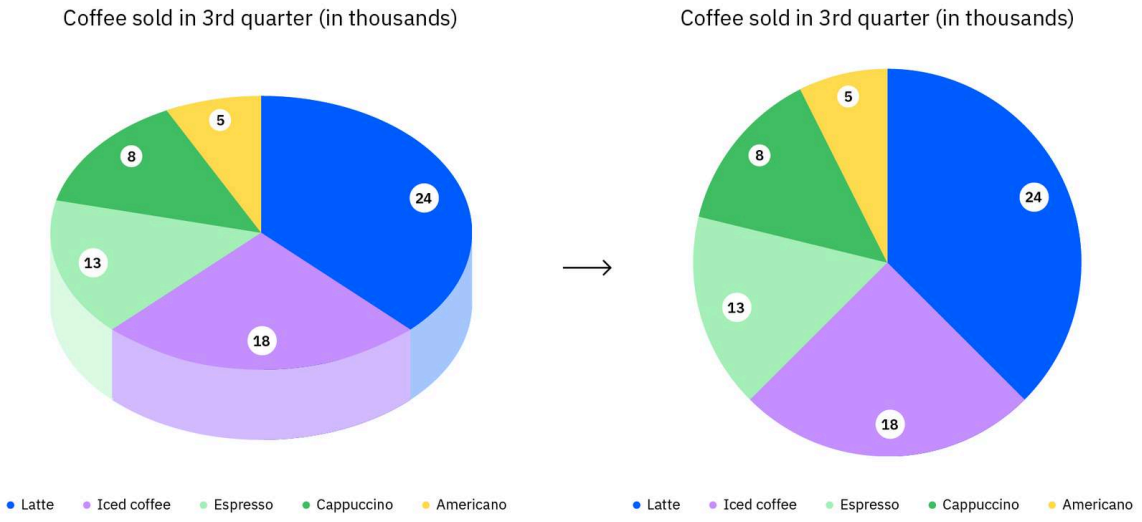




The visualization must support the objective and message it's meant to get across. In doing so, it's important that the design and features of the visualization make it attractive or pleasing. The visualization shouldn't have features that are distracting. There is some distraction here, so you can remove the black background. Also, changing the numbers from white to bold black provides better readability. Now, this is better and less distracting.

Step 3

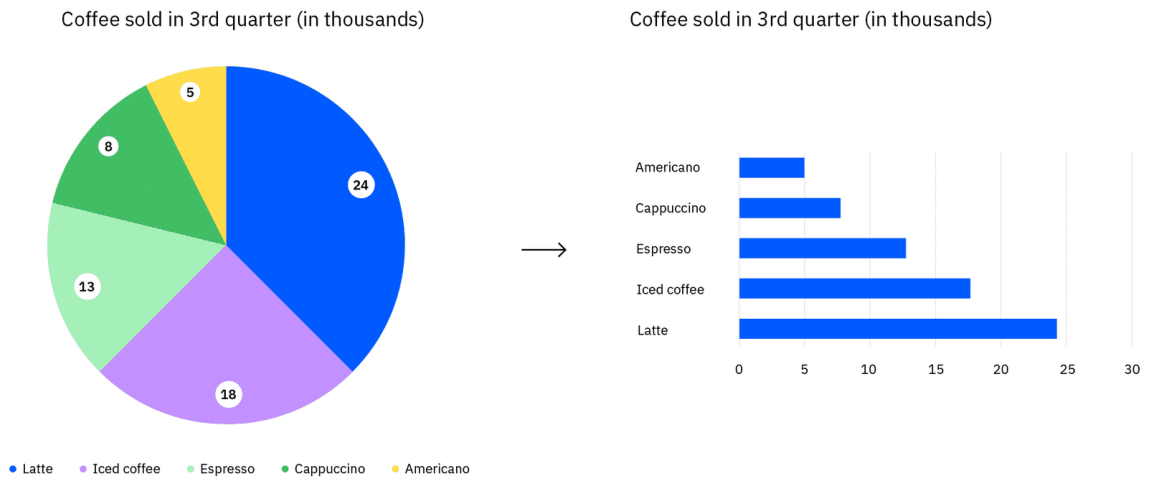
Refine It Further



There is still a lot going on in this pie chart. For instance, does it need the three-dimensional orientation? It seems unnecessary and distracting. It isn't adding any extra information. So, you can remove it so that the pie chart is "flat." We're getting there!

Step 4

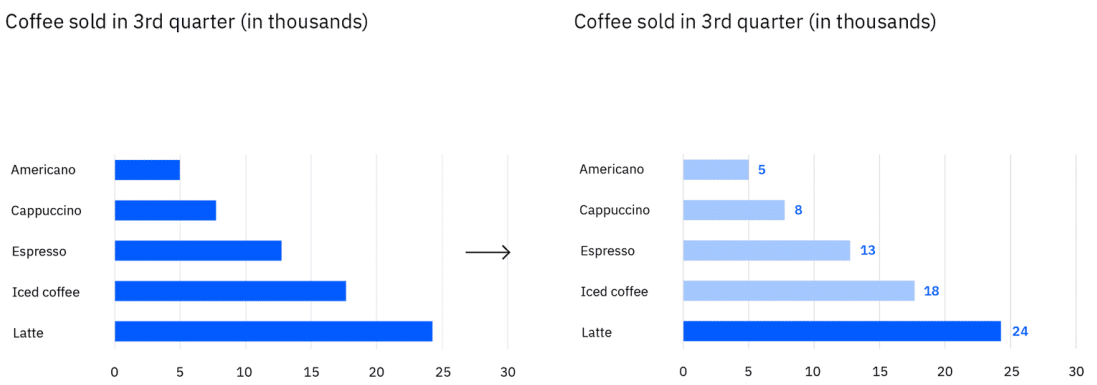
Try Out a More Suitable Chart



Bar charts are useful for comparing different types of things, such as types of coffee. So, a bar chart is better suited to visualize this data. Keep in mind the objective: to discover and show the highest-selling type of coffee from the third quarter. What a difference!

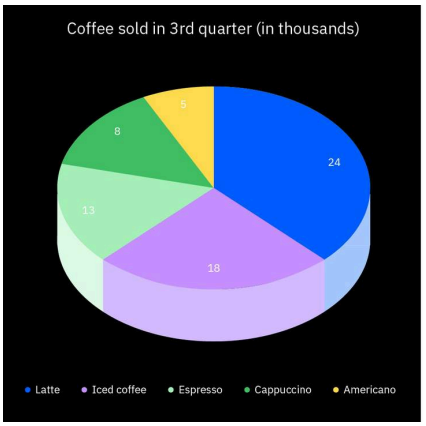
Step 5

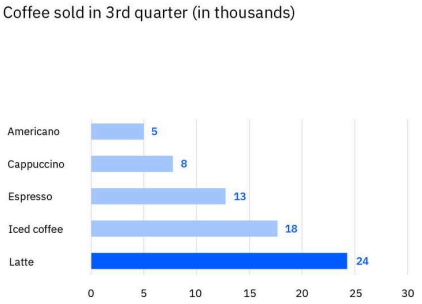
You Found the Best One



This is a bar chart depicting third-quarter coffee sales (in the thousands). The bar chart depicts that latte sales are 24,000, followed by iced coffee (18,000), espresso (13,000), cappuccino (8,000), and americano (5,000).

This is a much better view of the data that helps people conclude—without any confusion—that latte coffee is the highest-selling type of coffee. To make latte coffee stand out, the other bars have a lighter shade of blue. Latte coffee is the only bar with a darker shade of blue. This helps latte coffee stand out in the bar chart.





Data Can Be Misleading

Watch Out!

A real concern amongst the data analytics community is that visualization methods can represent the data in misleading ways. It's not that the data is wrong, but variables can be manipulated to reveal the same data set in a favorable or unfavorable way.

Here is an example of what could be misleading data in a line chart. Take a moment to compare the two line charts.

Evaluate Your Thinking

Does Your Answer Account for the Following?

Expand for Some More Thoughts While each line chart presents the trending price over time for "Stock J," the vertical scale (or y-axis) for Price is different. The scales show the data in two different increments. Notice the second chart is misleading because it doesn't depict $0 to $25 for Price like the first chart does. And, it shows Price in $5 increments. It makes it look like "Stock J" increased in price faster! The first chart is a more accurate depiction because it does not skip the price from $0 to $25 and shows Price consistently in $10 increments.

The key point here is to be precise in how you choose to depict data.

Explore More!

Check out these resources:

- **From Data to Viz** for an online tool to learn more about charts and help determine the appropriate graph for data.
- **How to Choose the Right Data Visualization** blog offered by Chartio - Data Tutorial and authored by Mike Yi and Mel Restori.
- **Viz of the Day** featured by Tableau Public where people around the world share their visualizations.

You've Come So Far!

And Learned So Much!

You learned fundamental data concepts such as types of data, big data, the typical steps in the data analytics process, and data visualization.

Now that you've completed this course, you should be able to:

- Explain the importance of data in a digital world
- Differentiate between unstructured and structured data
- Identify the purpose of a database
- Differentiate between quantitative and qualitative data
- Describe the five V's of big data
- Describe the four types of data analytics
- Explain the work involved in each step of the data analysis process
- Identify the purpose of a data visualization
- Recognize different charts to display data visualizations in the best way.

Key Points to Remember

- **Importance of Data Analysis**

  It’s important for businesses across industries to find insights to potentially improve operations, get a better understanding of customers or end users, increase profits, and much more.

- **Data Analysis Process**

  Data analysis is the process of collecting, cleaning, and transforming data to obtain insights to help make better and informed decisions.

- **Types of Data**

  1. **Structured Data**: Organized in databases with many columns and rows.
  2. **Unstructured Data**: Lacks organization and can include data such as tweets, customer comments, medical records, etc.

- **Quantitative vs Qualitative Data**

  1. **Quantitative Data**: Can be counted and is numerical.
  2. **Qualitative Data**: Cannot be counted and is non-numerical. Think "quality" — something you can categorize about data instead.

- **Big Data Characteristics**

  Big data is characterized by the 5 V’s:

  1. **Volume**
  2. **Variety**
  3. **Velocity**
  4. **Veracity**
  5. **Value**

- **Types of Data Analytics**

  There are four types of data analytics that answer key questions, build on each other, and increase in complexity:

  1. **Descriptive**
  2. **Diagnostic**
  3. **Predictive**
  4. **Prescriptive**

- **Typical Data Analysis Process**

  While there is no single, defined data analysis process, these are the typical steps to follow:

  1. **Collect**
  2. **Clean**
  3. **Analyze**
  4. **Visualize**

- **Data Visualization**

  A data visualization, such as a chart, graph, or map, helps you explore and interpret data during analysis to identify patterns or trends, communicate results, and help people understand the insights to make decisions.

- **Effective Visualization**

  One type of visualization can display data in a better way, compared to another type of visualization. The goal is to have a visualization that's effective, attractive, and impactful.

Data Science in Our World

About This Course

The goal is to turn data into information, and information into insight.

Carly Fiorina, former executive, president, and chair of Hewlett-Packard Company

Data is a precious thing and will last longer than the systems themselves.

Tim Berners-Lee, inventor of the World Wide Web

When we have all data online it will be great for humanity. It is a prerequisite to solving many problems that humankind faces.

Robert Cailliau, Belgian informatics engineer and computer scientist

Introduction

All roads lead to data in our ever-growing, digital world! An increasing number of businesses and organizations are data-driven, meaning they conduct data science projects to uncover and explain the business insights buried in data. The field of data science is and will continue to be critical to business success. Data science is helping to solve some of the biggest global challenges in health, medicine, climate change, agriculture, and technology.

Welcome to the Data Science in Our World course! In this course, you'll learn about the field of data science, including widely adopted methodologies and the steps of a data science methodology process. You'll also learn about the application of data science in our world and the role of data scientists and their fellow colleagues.

Learning Objectives

After completing this course, you should be able to:

- Define data science
- Recognize the importance of being curious to solve problems with data
- Differentiate between the fields of data analytics and data science
- Identify three widely adopted data science methodologies
- Explore a data project scenario and identify key tasks as it moves through a methodology

- Recognize industries and applications of data science that help solve global problems and discover innovations
- Compare the roles and characteristics of a data analyst, data scientist, and data engineer

Estimated Duration

- 2 hours

Completion Criteria

To get credit for completing this course, you need to:

- Visit all the pages in the course
- Complete the interactive activities and quizzes
- Pass the final assessment

You must pass the quiz at the end of each module to continue to the next module. There is a final, graded assessment at the end of the course. You must achieve 80 percent to pass. Don’t worry! If you aren’t successful at first, you can retake the quiz as many times as needed to pass.

As you progress through the course, you’ll see completion check marks beside the module and lesson names in the side navigation.

Copyright

© Copyright IBM Corporation 2024

Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

Version: 1.1.0

Date: 20-September-2024

Release Notes:

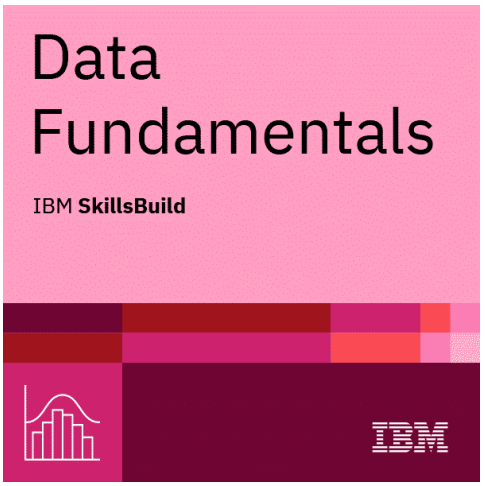
- Improved final assessment questions, including adding correct and incorrect feedback.
- Added instructional text to quizzes.

Version Information

Ready to get started?

Data Fundamentals Series

Earn the Data Fundamentals Digital Credential from IBM!



This course is one of several in the **Data Fundamentals** series. Complete all the courses in the Data Fundamentals learning plan and earn the **Data Fundamentals** digital credential from IBM!

Credential:

Data Fundamentals

For more information about this credential, visit [Data Fundamentals] on Credly.

What's a Digital Credential?

A credential is an industry-recognized and validated indicator of knowledge and skills acquisition. **IBM SkillsBuild** credentials are aligned to job-market standards and frameworks to help achievers gain visibility with potential employers. IBM SkillsBuild issues its digital credentials through Credly.

Share Your Credentials

Share your credentials on LinkedIn and other social networks with prospective employers, or include them on your resume to amplify your skills!

Course Design

This course is designed so you can learn at your own pace and easily navigate using the organized module structure. You'll find:

- Concepts, graphics, and examples to help you learn
- Interactions to reinforce what you learn
- Quizzes to check your knowledge
- Helpful resources so you can explore more
- A final assessment
- Tips to be a successful learner

The course is designed for you to complete the modules in sequential order.

You can pace your learning over a period that works for you and your schedule.

Your progress is tracked, so you can complete a module or lesson and then come back to pick up where you left off.

Be a curious learner and dive into learning!

What is Data Science?

Introduction

In this module, you will learn how you can explain what data science is to people. You will also learn about the 5 Whys technique and how to use it to identify the root cause of a problem. Finally, you'll learn about similarities and differences between data analytics and data science.

Learning Objectives

After completing this module, you should be able to:

- Define data science
- Recognize the importance of being curious to solve problems with data
- Differentiate between the fields of data analytics and data science

What is Science?

- Science is a system or method reconciling practical ends with scientific laws.

What is Data Science?

- Data science is the understanding of the world through the scientific analysis of digital data.
- Combines the scientific method, math, statistics, programming, advanced analytics, AI, and storytelling.
- A multidisciplinary approach to extracting actionable insights from large volumes of data.
- The goal is to extract value from data in all its forms.

The Role of Data Scientists

- Prepare data for analysis and processing.
- Perform advanced data analysis and present results to reveal patterns.
- Use tools and technologies to build models to predict outcomes and discover underlying patterns.
- Provide insights that improve future outcomes.

Data Science as a Commodity

- Data is a commodity; businesses need ways to process and find its value.
- Data science is a repeatable, scientific process.
- Involves unique steps: data collection, data cleansing, machine learning, and visualization.

What Do You Think?

- Imagine explaining data science to a friend or family member.
- Write your response, reflect on your thoughts, and evaluate if your answer covers the following:
    - Data science helps people find value and gain business insights from data.
    - Data science follows a process, uses tools and technologies, and builds models to discover patterns to improve future outcomes.

Be Curious

- Data scientists begin by asking questions that start with "Why?"
    - Example: Why did life move from the primordial oceans to land? Did the moon and the tides have something to do with it?
- **Curiosity** is one of the most important characteristics of a data scientist.

The 5 Whys

- **The 5 Whys** is a valuable problem-solving technique that helps identify the root cause of a problem.
- The technique involves asking "Why?" five times, with each answer informing the next question.
    - The answer to the fifth "Why?" should reveal the root cause of the problem.
- Some problems may have more than one root cause. If this happens, the method should be repeated with a different sequence of questions.
- **Note:** The number of "Why?" questions can vary; five is a rule of thumb.

Example: The 5 Whys in Action

- **Problem:** Your vehicle won't start.
- Start with the **First Why**:
    - **First Why:** Why won't my car start?
        - **Answer:** The battery is dead.
- **Second Why:** Why is the battery dead?

    - **Answer:** The alternator is not functioning.
- **Third Why:** Why isn't the alternator functioning?
    - **Answer:** The alternator belt is broken.

- **Fourth Why:** Why is the alternator belt broken?
    - **Answer:** The alternator belt is beyond its useful service life and has not been replaced.

- **Fifth Why:** Why has the alternator belt not been replaced if it is beyond its useful service life?
    - **Answer:** The car has not been maintained according to the recommended service schedule.

Solution:

- **Solution:** Replace the alternator belt and perform the recommended services on the car. This will enable the car to start and prevent similar or other preventable failures in the future.

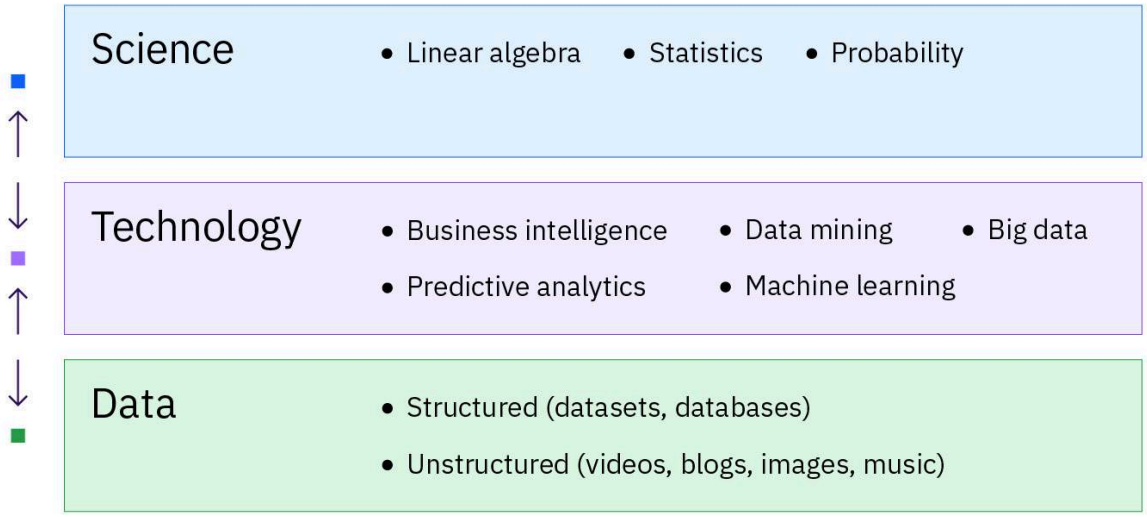
Note:

- When using the 5 Whys technique, avoid making assumptions or skipping to the obvious answer.

Science, Technology, and Data Are Connected

Science, technology, and data are areas that are linked and have given rise to data science. Keep this key concept in mind as you learn about the field of data science.

- A scientific background allows a data scientist to formulate a hypothesis and follow the scientific method.
- Working with technology and tools allows a data scientist to begin classifying data and making predictions.
- A data scientist can gain insights from structured data in relational databases and unstructured data, such as video footage, blogs, and tweets.
- Science, technology, and data have been linked since the early days of science in the 1300s and 1400s. Each area continues to evolve.



At this time in the **science plus technology plus data world**, data science is one of the most important fields.

Data Analytics versus Data Science

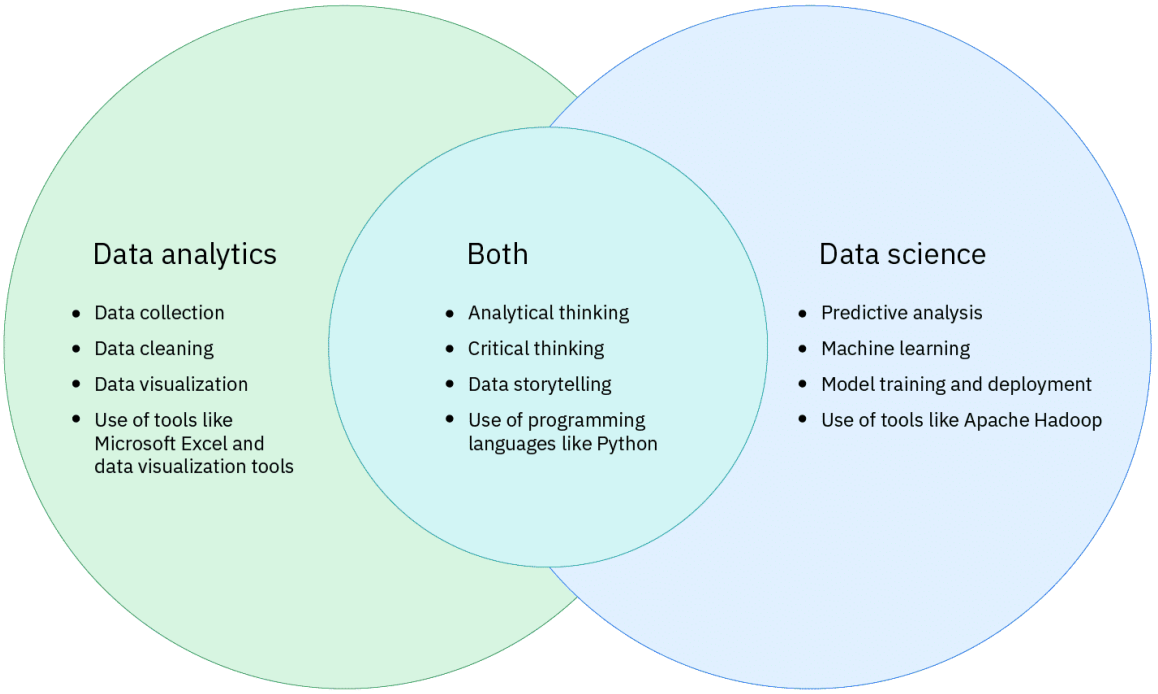
Data analytics and data science are two terms that are often used in the same context. But, it's important to know they have different definitions.

Both fields work with data, and they share the same goal: to translate data analysis into business intelligence.

The main difference between data analytics and data science is in what data analysts and data scientists do with data, meaning the tactics used. Here is how they are differentiated:

- **Data Analysts**:
    - Collect and examine large data sets to identify trends, forecasts, and data visualizations.
    - Tell a compelling story through actionable insights.
    - Help businesses make informed decisions about business needs.
- **Data Scientists**:
    - Design and create new processes for data modeling.
    - Use algorithms, predictive analytics, and statistical analysis.
    - Have technical skills to arrange unstructured data and build their own methodologies to make predictions based on data trends.

This diagram illustrates differences between data analytics, data science, and what they have in common.



Source: Data Analytics vs. Data Science by CompTIA

You'll learn about the role of the data analyst, data scientist, and the teams they work with on data science projects in this course.

Explore More!

Check out these resources:

- **What is Data Science?** – An IBM article that provides a quick tutorial about data science and tools.
- **Data and Analytics** – Provides online resources and articles from CompTIA about the field, skills needed, and the industries that use data.

Data Science Methodologies

Introduction

In this module, you will learn about three traditional data science methodologies so that you have a foundation to learn more about data science projects.

Learning Objectives

After completing this module, you should be able to:

- Identify three widely adopted data science methodologies.

What is a Methodology?

A methodology is a general strategy that guides activities within a process. A methodology doesn't depend on technologies or tools, and it's not a set of techniques or recipes. Rather, a methodology provides data scientists with a framework for how to proceed with whatever methods and processes they will use to obtain answers or results.

Scientists have the scientific method. And like scientists, data scientists need a foundational methodology to guide them for solving problems.

It's Good to Know Your History!

To begin, here are three classic and widely adopted data science methodologies:

- **Cross-Industry Standard Process for Data Mining (CRISP-DM)**
- **Knowledge Discovery in Database (KDD)**
- **Sample, Explore, Modify, Model, Assess (SEMMA)**

CRISP-DM, KDD, and SEMMA:

- Use data mining methods
- Are best suited for structured data
- Are useful for using descriptive and predictive analytics
- Share some common activities, such as data gathering, data transformation, data modeling, and model evaluation

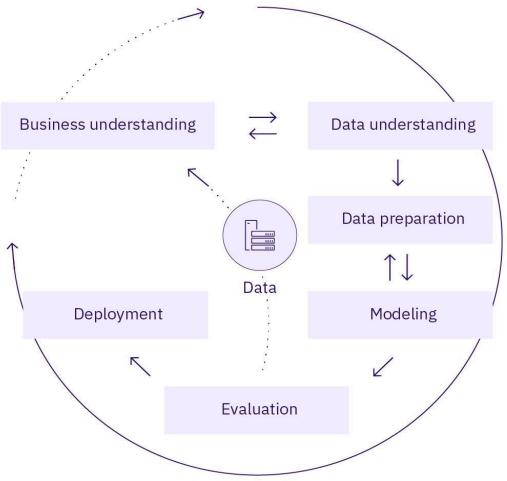
Note: These methodologies are not useful on projects that work with unstructured data, such as images and text.

You will learn about these widely adopted data science methodologies in upcoming lessons.

Example 1: CRISP-DM

CRISP-DM stands for **Cross-Industry Standard Process for Data Mining**.

Founded by the European Strategic Program on Research in Information Technology initiative, CRISP-DM is a proven way to guide data mining efforts. Any industry can use this methodology to help structure a data science project. CRISP-DM is a flexible and comprehensive data science approach.



CRISP-DM Methodology

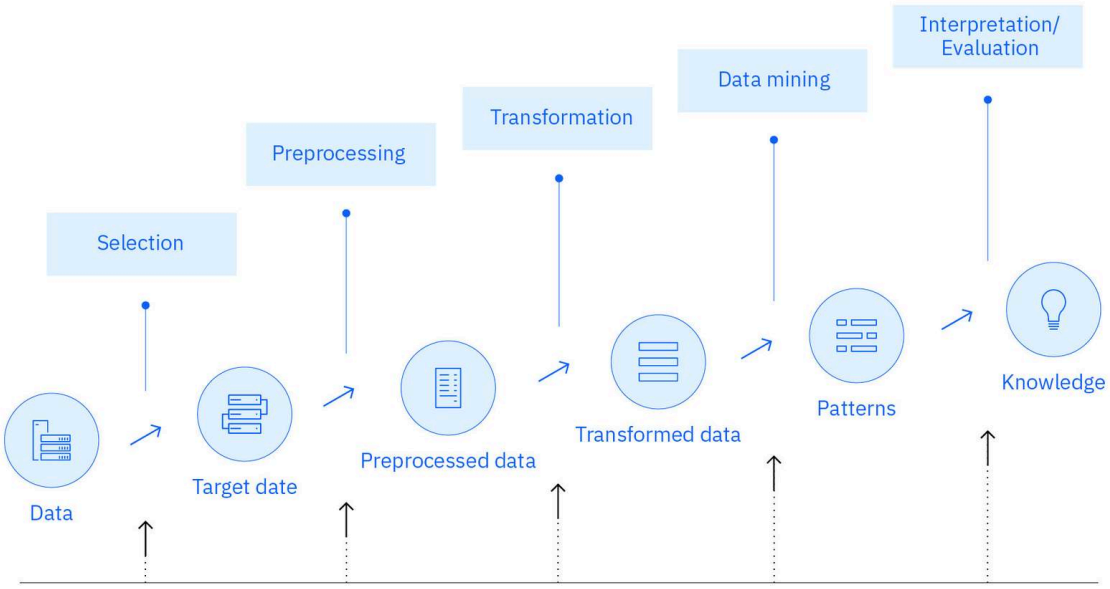
CRISP-DM consists of six phases with arrows indicating the most important and frequent dependencies between phases:

- **Business understanding**
- **Data understanding**
- **Data preparation**
- **Modeling**
- **Evaluation**
- **Deployment**

The sequence of the phases is not strict. CRISP-DM is iterative, meaning that the phases can be repeated to incrementally improve the result. The results of some stages might require the project cycle to go back to earlier stages.

What's unique to the CRISP-DM methodology is that it begins with **Business understanding**. This phase focuses on understanding the project objectives and requirements from a business perspective and defining the data problem to solve.

CRISP-DM is among the most popular approaches for data science projects.



Example 2: KDD

KDD stands for **Knowledge Discovery in Database**.

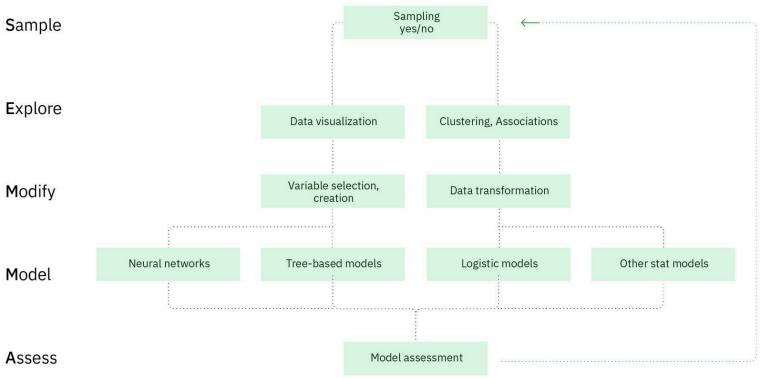
KDD represents the overall process of collecting data and methodically refining it. KDD typically has five steps:

- **Selection**
- **Preprocessing**
- **Transformation**
- **Data Mining**
- **Interpretation/Evaluation**

The KDD methodology can help businesses stay current with customer needs and behaviors and predict future purchasing trends to stay competitive. However, the process doesn't address many of the modern realities of data science projects, such as the setup of big data architecture, considerations of ethics, or the various roles in a data science team.

KDD is iterative, meaning new data can be integrated and transformed to get different and more appropriate results. The knowledge acquired can be cycled back into the process, enhancing its effectiveness.

Example 3: SEMMA



SEMMA stands for its five steps:

- **Sample**
- **Explore**
- **Modify**
- **Model**
- **Assess**

SEMMA methodology

SEMMA is a data science methodology that helps convert data into knowledge. SEMMA can help solve a range of business problems, such as fraud identification, customer retention and turnover, database marketing, customer loyalty, market segmentation, and risk analysis.

The **SAS Institute** developed SEMMA as a process of data mining. SEMMA primarily focuses on the modeling tasks of data mining projects.

SEMMA is also an iterative process, in which answering one set of questions often leads to more interesting and more specific questions.

Source: What is SEMMA? (opens in a new tab), Data Science Process Alliance, by Nick Hotz, May 2021

What do you think?

Now that you've learned about the CRISP-DM, KDD, and SEMMA methodologies, here's one observation on what they have in common:

- **Common Observation**: All three methodologies emphasize iterative processes and focus on **data preparation**, **data transformation**, and **model evaluation**. They are designed for **structured data** and rely on **data mining techniques** to extract valuable insights.

This reflection allows you to consolidate your learning by comparing the core components and similarities across these methodologies.

Does your answer account for the following?

Expand for some more thoughts:

- **Iterative Process**: The three methodologies—CRISP-DM, KDD, and SEMMA—are all **iterative**. This means the phases or steps can be repeated, and the knowledge gained in one phase can be cycled back to improve previous steps. This approach allows data scientists to refine and adjust their models for better insights.

Conclusion:

It's essential to familiarize yourself with the **CRISP-DM**, **KDD**, and **SEMMA** methodologies as a reference and background. Different organizations and businesses typically decide on their own methodology for data science projects. You'll learn more about how these methodologies are applied in upcoming lessons.

Follow a Data Science Methodology

Introduction

In this module, you will learn about the typical tasks in a six-step data science methodology and follow a project scenario to learn from examples as you go through each step.

Learning Objectives

After completing this module, you should be able to:

- Explore a data project scenario and identify key tasks as it moves through a methodology.

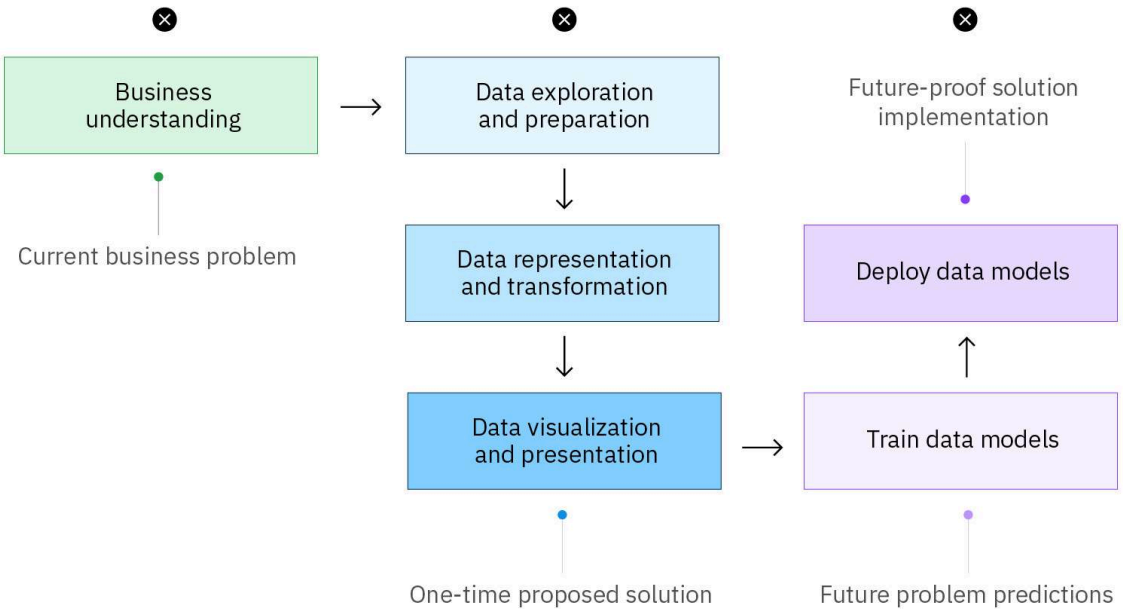
Methodology Overview

In previous lessons, three well-known data science methodologies were presented. To review, they are Cross-Industry Standard Process for Data Mining (CRISP-DM), Knowledge Discovery in Database (KDD), and Sample, Explore, Modify, Model, Assess (SEMMA). In this lesson, you will explore a data science methodology in depth.

Here are the steps in the data science methodology you'll learn about:

- Business understanding
- Data exploration and preparation
- Data representation and transformation
- Data visualization and presentation
- Train data models
- Deploy data models

Take a moment to review the methodology steps in the following diagram. Then, select each “x” button to review an overview.



The Methodology Steps

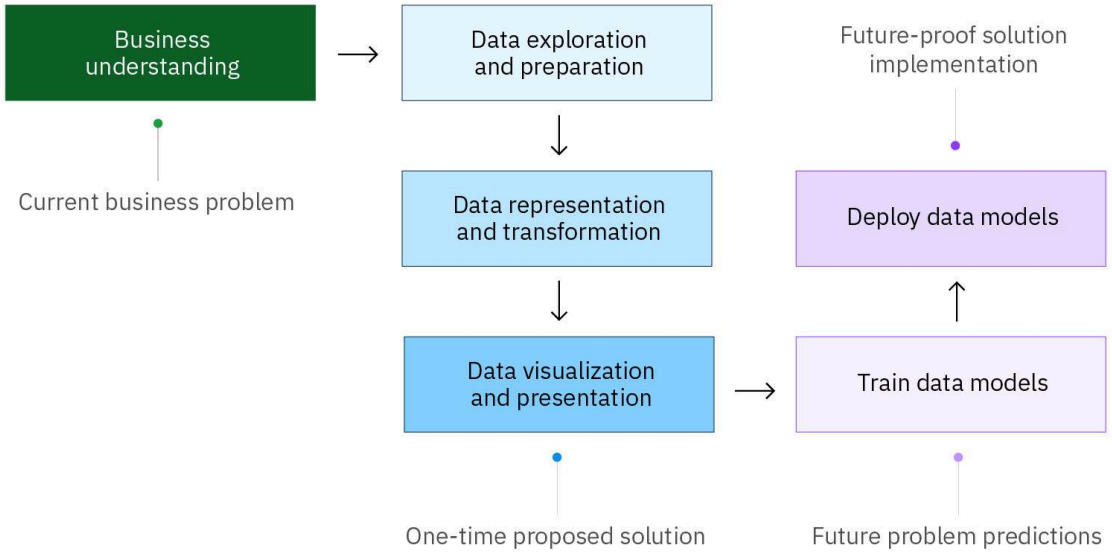
The methodology steps you use might vary, but you can follow these steps to learn about typical tasks that a company might implement for their data science projects. This methodology is independent of technologies or tools and provides a framework for processes that data scientists can use to obtain answers and results. The methodology is also scientific, so the key is that it is repeatable.

Step 1: Business Understanding

Every project, regardless of how big it is, starts with business understanding. Before any data exploration can be done, the team must understand the problem that needs to be solved.

The business sponsor plays a critical role. Business sponsors are in leadership positions. They initiate the project because they have a “pain point,” so they bring the business problem to the data science project team and then support the project.

Flowchart diagram showing the data science methodology for a project, with the first step, Business Understanding, highlighted.



The data science project team examines the business problem. They can do this using design thinking. Design thinking is a problem-solving methodology that focuses on the user, having empathy for the user, and determining the best user experience.

As part of this method, the data science project team might host a design thinking workshop and apply techniques to:

- Define the problem
- Determine the project objectives
- Develop personas or fictional characters that represent typical end users
- Document solution requirements from a business perspective

Defining the problem helps lay the foundation for a potential resolution to the business problem. The business sponsor should also be involved throughout the project to provide expertise, review data findings, and ensure the project stays on track.

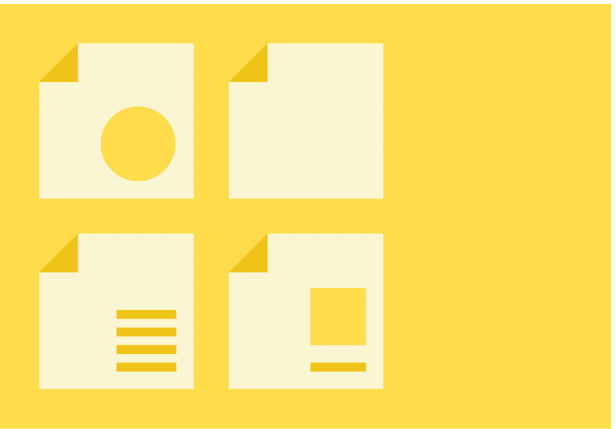
Once the business problem is clearly stated, a data scientist on the team defines the analytic approach to solving the problem. This involves expressing the problem in the context of statistical and machine learning techniques. For example, if the goal is to predict a customer's response in terms of a "yes" or a "no," then the analytic approach could be defined as building, testing, and implementing something called a logistic regression model. Data scientists are experts who have many technologies and methods in their toolbox!

Project Scenario: Business Understanding

A fictitious business, named GAXR, develops software for companies that design video games. GAXR has over 1000 employees. Suddenly, GAXR is facing a human resources (HR) issue. The HR director doesn't know why so many people are leaving the company. There haven't been any company layoffs, salaries are competitive, and the amenities at work are great, including a free onsite gym!

Image of sticky notes

The business problem is that there has been higher-than-expected employee attrition over the past six months. Employee attrition is when employees depart a company for any reason, such as resignation, termination, or retirement. In the case of GAXR, the employee attrition is due to voluntary resignations or people quitting.



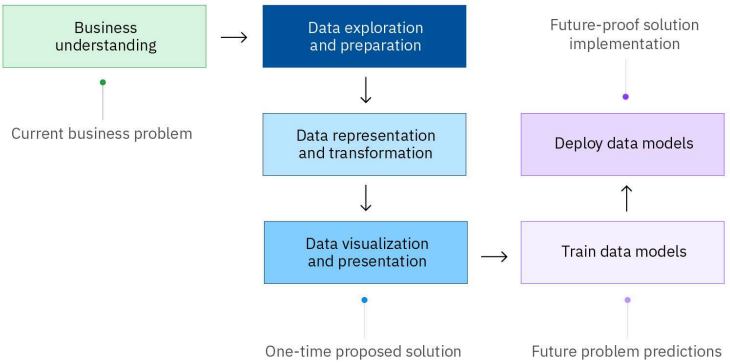
The HR director, Marilyn Shah, is the business sponsor. She gathers a team and hosts a one-day design thinking workshop so the team can understand the business problem and consider how to address the problem.

The data science project begins! The team begins to brainstorm and hypothesize why people are leaving GAXR. The team data scientist, Scott Hill, participates. He partners with his colleagues to define the best analytical approach and determine which company tools the team will use.

Step 2: Data Exploration and Preparation

Data scientists identify and collect data from existing and often new sources in a business. This could be structured and unstructured data that is relevant to the problem. They might retrieve data from sources such as:

- Static files, such as spreadsheets
- Databases
- The internet



Flowchart diagram showing the data science methodology for a project, with the second step, Data Exploration and Preparation, highlighted.

Step 2: Data Exploration and Preparation

If a data scientist encounters any issues or gaps in data collection, they might need to revise the data requirements and collect more data.

Once data is in a format a data scientist can work with, data exploration can begin. The initial exploration of a data set is important because it helps data scientists find patterns and relationships and discover initial insights from the data. Here are a few questions a data scientist might consider during the initial exploration of data:

- Which data characteristics seem promising for further analysis?
- Has exploring revealed new characteristics about the data?
- Has exploring changed the initial hypothesis?

Next, the data scientist prepares the data. Data preparation is very important and often the most time-consuming step in a data science project. It involves constructing the data set that will be used in the modeling step. Data preparation also includes cleaning the data, combining data from multiple sources, and ensuring the data doesn't have any gaps. Additionally, data preparation includes cleaning or "wrangling" the data so it's ready to transform.

Data scientists can't assume that data is ready to use, even if it's structured data. Real-world data usually needs some work because it might be:

- Incomplete or have incorrect values
- Corrupted with broken lines or have fields in the wrong place
- Too random
- Irrelevant
- An outlier, which is a value that lies far away from other values and will skew the data
- A missing value in some fields

Manually verifying large volumes of stored data can be challenging for data scientists. So, they use automated processes and tools to prepare data quickly and accurately.

Project Scenario: Data Exploration and Preparation

Scott Hill, the team data scientist, has a meeting with Lena, the GAXR database administrator, to get her guidance on the types of data that the human resources group collects for its employees. Scott

reviews and determines the employee data set he wants, such as:

- First name
- Last name
- Age
- Department
- Distance from home (to work in kilometers)
- Hourly rate
- Daily rate
- Monthly income
- Marital status

Image of a magnifying glass on a spreadsheet

Scott uses Structured Query Language (SQL) to move the data he needs for the project into a CSV file.

SQL is the most common language for extracting, organizing, and managing data in a relational database to perform various operations on the data. A comma-separated values (CSV) file allows data to be saved in a tabular format.

Here is the CSV file that Scott is cleaning up:

	A	B	C	D	E	F	G	H	I	J
1	Age	BusinessTravel	Department	DistanceFromHome	Gender	JobRole	JobSatisfaction	MaritalStatus	DailyRate	MonthlyIncome
2	41	Travel_Rarely	Sales	1	Female	Sales Executive	4	Single	1102	5993
3	49	Travel_Frequently	Research & Development	8	Male	Research Scientist	2	Married	279	5130
4	37	Travel_Rarely	Research & Development	2	Male	Laboratory Technician	3	Single	1373	2090
5	33	Travel_Frequently	Research & Development	3	Female	Research Scientist	3	Married	1392	2909
6	27	Travel_Rarely	Research & Development	2	Male	Laboratory Technician	2	Married	591	3468
7	32	Travel_Frequently	Research & Development	2	Male	Laboratory Technician	4	Single	1005	3068
8	59	Travel_Rarely	Research & Development	3	Female	Laboratory Technician	1	Married	1324	2670
9	30	Travel_Rarely	Research & Development	24	Male	Laboratory Technician	3	Divorced	1358	2693
10	38	Travel_Frequently	Research & Development	23	Male	Manufacturing Director	3	Single	216	9526
11	36	Travel_Rarely	Research & Development	27	Male	Healthcare Representativ	3	Married	1299	5237
12	35	Travel_Rarely	Research & Development	16	Male	Laboratory Technician	2	Married	809	2426
13	29	Travel_Rarely	Research & Development	15	Female	Laboratory Technician	3	Single	153	4193
14	31	Travel_Rarely	Research & Development	26	Male	Research Scientist	3	Divorced	670	2911
15	34	Travel_Rarely	Research & Development	19	Male	Laboratory Technician	4	Divorced	1346	2661
16	28	Travel_Rarely	Research & Development	24	Male	Laboratory Technician	3	Single	103	2028
17	29	Travel_Rarely	Research & Development	21	Female	Manufacturing Director	1	Divorced	1389	9980
18	32	Travel_Rarely	Research & Development	5	Male	Research Scientist	2	Divorced	334	3298
19	22	Non-Travel	Research & Development	16	Male	Laboratory Technician	4	Divorced	1123	2935
20	53	Travel_Rarely	Sales	2	Female	Manager	4	Married	1219	15427
21	38	Travel_Rarely	Research & Development	2	Male	Research Scientist	4	Single	371	3944
22	24	Non-Travel	Research & Development	11	Female	Manufacturing Director	3	Divorced	673	4011
23	36	Travel_Rarely	Sales	9	Male	Sales Representative	1	Single	1218	3407
24	34	Travel_Rarely	Research & Development	7	Female	Research Director	2	Single	419	11994
25	21	Travel_Rarely	Research & Development	15	Male	Research Scientist	4	Single	391	1232
26	34	Travel_Rarely	Research & Development	6	Male	Research Scientist	1	Single	699	2960

CSV file depicting GAXR employee data

Each line of the file is a data record for one GAXR employee.

Next, Scott cleans the data! For instance, he removes the Employee Count field since all GAXR employees have the value of 1 (because each line is for one employee). This is unnecessary data. Scott also updates the Hourly Rate field and all other monetary fields to remove decimal places and round up to the nearest US dollar.

Scott also removes personal information (PI) since it's not pertinent to his data science project. This includes the First Name and Last Name fields.

In order to run data comparisons in the tool, Scott also removes columns that are not helpful in making predictions, such as the Telephone Number.

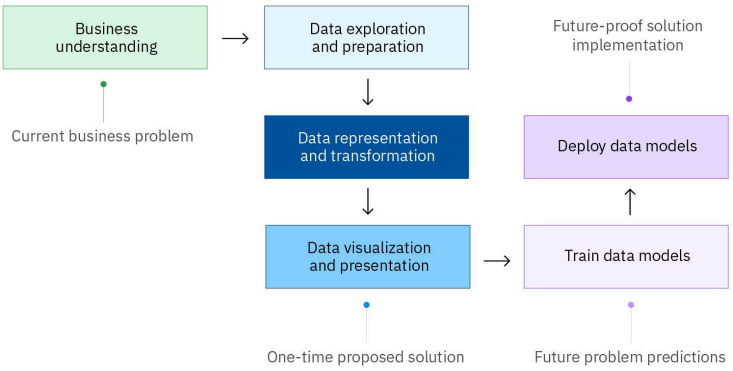
Additionally, Scott ensures there aren't any gaps and that he has enough data. In all, he has 1470 employee records to use for his analysis. He revisits and confirms his

hypothesis: Employees are resigning from GAXR because they need a higher salary in today’s economy.

Step 3: Data Representation and Transformation

The data representation and transformation step in the data science methodology is about:

- Understanding the data
- Assessing data quality
- Discovering initial insights about the data



Flowchart diagram showing the data science methodology for a project, with the third step, Data Representation and Transformation, highlighted.

Step 3: Data Representation and Transformation

Think of data scientists as detectives investigating a case. Data scientists use many techniques, tools, and types of analysis to better represent and transform data to detect insights. This lesson presents some of these techniques, but remember that there are many more.

Descriptive Statistics

To understand data, a data scientist can use a mathematical approach, such as descriptive statistics. Descriptive statistics quantitatively summarizes a data set. It can answer the question, “What is happening?” Data scientists can build a table to describe a large, complex data set and make quick observations about:

- **Number (N)**: What is the total number of observations?
- **Mean**: What is the average of a set of two or more numbers?
- **Median**: What is the middle number or “center” in a sorted list of numbers?

- **Mode**: What is the most observed value in a data set?
- **Minimum**: What is the minimum extreme of a data set?
- **Maximum**: What is the maximum extreme of a data set?
- **Standard deviation**: How spread out is the data in relation to the mean?

For instance, the following table indicates the mean household income in US dollars for a company's customers, based on their education level. The data is fictitious and for illustrative purposes.

Education Level	**Did not complete high school**	**High school degree**	**Some college**	**College degree**
Mean	51.48	52	56.90	70.94
Standard Deviation	51.855	56.370	53.836	67.940
N (Count)	246	527	333	310
Median	36.00	35.00	39.00	49.00
Minimum	15	12	13	15
Maximum	497	533	403	512

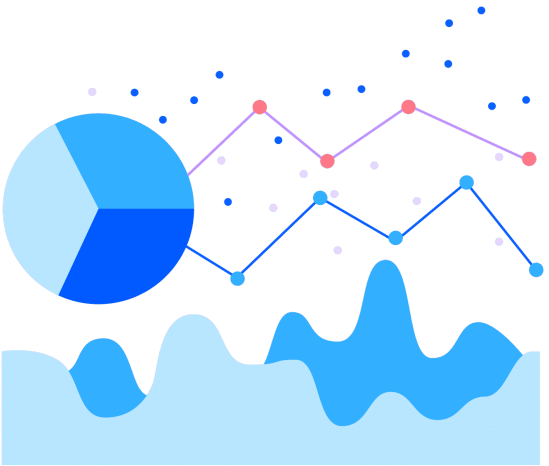
Did you complete this table?

What conclusions can you make about this descriptive statistics table? Consider:

- The mean income increases with education level.
- The standard deviation varies significantly across education levels, suggesting different ranges of income within each group.
- The median values are consistently lower than the mean, indicating some higher-income outliers.

Exploratory Visualizations

Sometimes, statistics can be misleading on their own, so it's important to use additional techniques. Data scientists also rely on exploratory visualizations, which help make complex data more accessible and revealing. Initial visualizations, such as charts, graphs, and maps, help uncover distributions, find patterns, and understand trends.



Abstract representation of different visualizations, including a pie chart and line chart.

Descriptive statistics, visualization techniques, and many other techniques help data scientists understand data and assess its quality. Data science teams must validate the quality of the data they use as input for predictive modeling because poor quality data will lead to poor model performance later in the process.

Transforming Data with Machine Learning

Data representation is followed by data transformation. Regardless of the format of the source data, a data scientist structures and organizes the data in a way that supports the most efficient machine-learning model. Machine learning is a branch of artificial intelligence (AI) and computer science that focuses on using data and algorithms to imitate the way humans learn, gradually improving its accuracy. Machine learning models understand only numbers, not text or images, so data scientists might need to transform unstructured data into a “0” or “1”.

For instance, a data scientist can break up text into words, phrases, symbols, or other meaningful elements called tokens. Then, a list of tokens becomes input for further processing into numbers. This technique is called **tokenization**, which is one of many data transformation techniques.

Data scientists also use data transformation tools and languages, such as **Python**, to conduct more in-depth analysis and create more complex visualizations. They perform programmatical activities on data formatting, including determining which models may be the best fit for the data.

To make analysis easier, data scientists often need to standardize data into model-friendly formats so that it’s "tidy data" rather than "messy data."

Data scientists might need to refine their analytics approach, which was selected in the **Business Understanding** step.

Project Scenario: Data Representation and Transformation

Scott Hill, the team data scientist, continues with the data science project to find out why employees are leaving GAXR.

Scott realizes he needs to access another data source to better understand all of the data. He needs data on why a particular employee left the company. Scott uses SQL to merge more data into his CSV file. He adds a new column for **Attrition**. This is an important step!

Take a moment to study column B—the new, highlighted **Attrition** column—in the following image of the GAXR CSV file.

	A	B	C	D	E	F	G	H	I	J	K
1	Age	Attrition	BusinessTravel	Department	DistanceFromHome	Gender	JobRole	JobSatisfaction	MaritalStatus	DailyRate	MonthlyIncome
2	41	Yes	Travel_Rarely	Sales	1	Female	Sales Executive	4	Single	1102	5993
3	49	No	Travel_Frequently	Research & Development	8	Male	Research Scientist	2	Married	279	5130
4	37	Yes	Travel_Rarely	Research & Development	2	Male	Laboratory Technician	3	Single	1373	2090
5	33	No	Travel_Frequently	Research & Development	3	Female	Research Scientist	3	Married	1392	2909
6	27	No	Travel_Rarely	Research & Development	2	Male	Laboratory Technician	2	Married	591	3468
7	32	No	Travel_Frequently	Research & Development	2	Male	Laboratory Technician	4	Single	1005	3068
8	59	No	Travel_Rarely	Research & Development	3	Female	Laboratory Technician	1	Married	1324	2670
9	30	No	Travel_Rarely	Research & Development	24	Male	Laboratory Technician	3	Divorced	1358	2693
10	38	No	Travel_Frequently	Research & Development	23	Male	Manufacturing Director	3	Single	216	9526
11	36	No	Travel_Rarely	Research & Development	27	Male	Healthcare Representativ	3	Married	1299	5237
12	35	No	Travel_Rarely	Research & Development	16	Male	Laboratory Technician	2	Married	809	2426
13	29	No	Travel_Rarely	Research & Development	15	Female	Laboratory Technician	3	Single	153	4193
14	31	No	Travel_Rarely	Research & Development	26	Male	Research Scientist	3	Divorced	670	2911
15	34	No	Travel_Rarely	Research & Development	19	Male	Laboratory Technician	4	Divorced	1346	2661
16	28	Yes	Travel_Rarely	Research & Development	24	Male	Laboratory Technician	3	Single	103	2028
17	29	No	Travel_Rarely	Research & Development	21	Female	Manufacturing Director	1	Divorced	1389	9980
18	32	No	Travel_Rarely	Research & Development	5	Male	Research Scientist	2	Divorced	334	3298
19	22	No	Non-Travel	Research & Development	16	Male	Laboratory Technician	4	Divorced	1123	2935
20	53	No	Travel_Rarely	Sales	2	Female	Manager	4	Married	1219	15427
21	38	No	Travel_Rarely	Research & Development	2	Male	Research Scientist	4	Single	371	3944
22	24	No	Non-Travel	Research & Development	11	Female	Manufacturing Director	3	Divorced	673	4011
23	36	Yes	Travel_Rarely	Sales	9	Male	Sales Representative	1	Single	1218	3407
24	34	No	Travel_Rarely	Research & Development	7	Female	Research Director	2	Single	419	11994
25	21	No	Travel_Rarely	Research & Development	15	Male	Research Scientist	4	Single	391	1232
26	34	Yes	Travel_Rarely	Research & Development	6	Male	Research Scientist	1	Single	699	2960

An abstract spreadsheet followed by a refresh icon

The **Attrition** column contains what's called **dichotomous data**, meaning each cell has only two possible values: **Yes** or **No**. These values can help Scott quickly answer the question, "Did the employee leave GAXR or not?"

Scott ensures the data is in the right format for the model the data science project team plans to use.

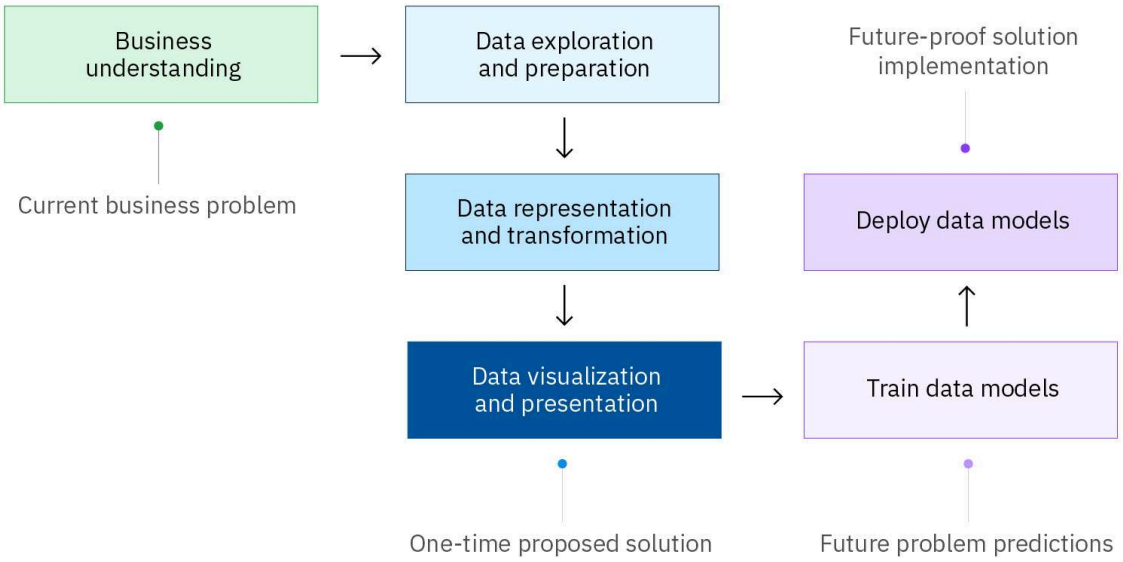
Scott also uses a data visualization tool to begin creating different charts and graphs. The charts and graphs can provide Scott with initial insights that help him draw conclusions about the data.

Important note: Column B, for **Attrition**, will be used as training data to help build a **supervised learning model**. The model will help determine the cause of employee attrition, which the company can also use to consider future strategies.

Step 4: Data Visualization and Presentation

Overview:

The data visualization and presentation step is about displaying the insights from data transformation through charts and graphs. These visualizations help to test hypotheses, check assumptions, and answer the business question being addressed. The key goal is to communicate findings effectively to stakeholders.



Types of Data Visualizations:

1. **Pie Charts**
    - **Purpose**: To show proportions of a whole.
    - **Use Case**: Best used when there are a small number of categories and you want to see relative sizes.
    - **Example**: A grocery store chain uses a pie chart to show the percentage of total profit from various locations.

2. **Bar Charts**
    - **Purpose**: To compare different categories or track changes.
    - **Use Case**: Best for ranking and comparing categories or showing changes over time.
    - **Example**: An online pharmacy uses a bar chart to show traffic increases during the COVID-19 pandemic.
3. **Column Charts**
    - **Purpose**: Similar to bar charts but used to show data over time.
    - **Use Case**: Effective for comparing items over time or tracking trends.
    - **Example**: A company uses column charts to track the number of accidents in warehouses over several months.
4. **Line Charts**
    - **Purpose**: To show trends over time, particularly with smaller changes.
    - **Use Case**: Ideal for showing changes over time or continuous data.
    - **Example**: An IT helpdesk uses a line chart to measure performance trends in closing customer issues.
5. **Scatter Plots**
    - **Purpose**: To examine the relationship between two variables.
    - **Use Case**: Great for identifying correlations or outliers in large data sets.
    - **Example**: A bank uses a scatter plot to analyze how customer happiness is affected by call response time.

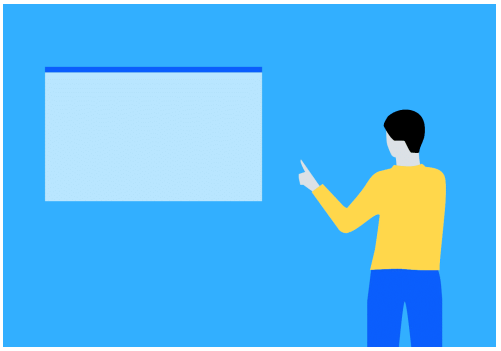
Considerations for Data Presentation:

To ensure the visualization communicates the data clearly and effectively, consider these factors:

- **Purpose**: What is the problem you are trying to address with your visualization? Why will the visualization help solve it?
- **Audience**: Who will be viewing the visualization? Tailor the message and design to your audience's needs.
- **Data**: Is the data represented in the best format for the intended insight? Will it need updates?
- **Context**: Where will the visualization be viewed (e.g., software, website, report)?

Project Scenario: Data Visualization and Presentation:

In this project, Scott Hill, the data scientist at GAXR, uses data visualization tools to present the findings of employee attrition at GAXR.

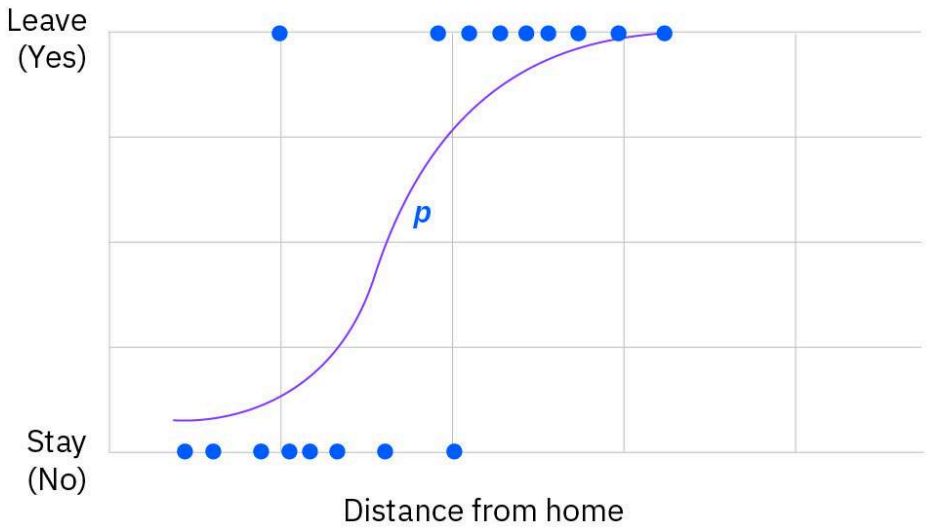


- **Business Problem**: GAXR wants to understand why employees are leaving the company.
- **Data Visualization**: Scott uses a **scatter plot** with an **S-curve** to show the relationship between the **distance from home** and **employee attrition**.
    - The visualization reveals that employees who live farther from the office are more likely to leave the company.

- **Insight**: The scatter plot shows an increasing probability of attrition as the distance from home increases.
- **Proposed Solution**: Offer employees the option to work from home two days per week to reduce turnover.

Final Presentation:

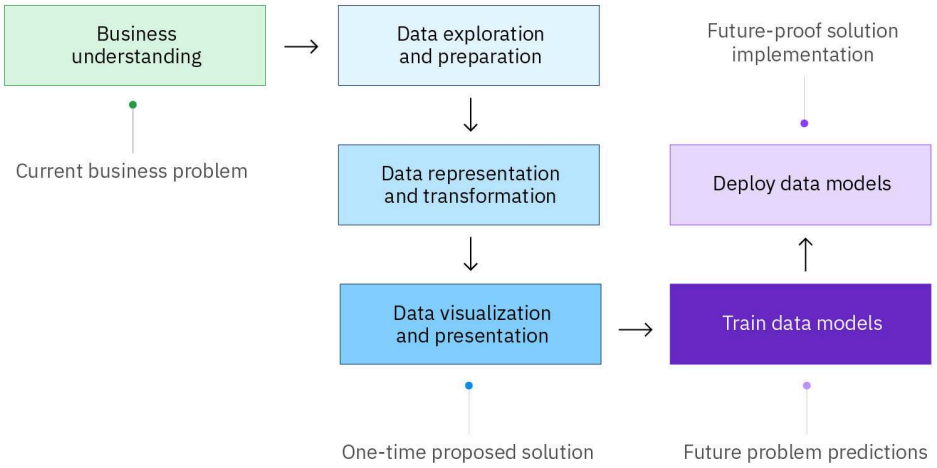
Scott prepares a presentation for Marilyn, the HR director:



- **What Scott Covers**:
  - **The Business Problem**: Employee attrition at GAXR.
  - **Data Sources**: Information from employee records and attrition data.
  - **Data Preparation Steps**: Merging datasets and ensuring the data is in a suitable format for analysis.
  - **Data Visualization**: The scatter plot showing the relationship between distance from home and attrition.
  - **Proposed Solution**: Offer remote work options to employees living far from the office.

- **Conclusion**: Marilyn will present the findings and solution to company leadership.

Step 5: Train Data Models



What is a Model?

A data model identifies the data, its attributes, and the relationships or associations with other data. It provides a generalized view of the data that represents a real business scenario and the underlying system. The goal of building a model is to predict future outcomes more accurately and gain a deeper understanding of the system being modeled.

Why Build a Model?

Building a model allows a data scientist to systematically address a business problem and make better predictions. It is a critical part of the data science process,

allowing organizations to predict and understand the factors that impact their business operations.

Machine Learning and Model Training

Training a model involves teaching a machine to make predictions or decisions based on data. This is done through machine learning, where a computer "learns" from the data it is provided, improving its performance over time. A machine learning algorithm ingests data and adapts to make more accurate predictions.

Types of Machine Learning:

1. **Supervised Learning**
    - **Description**: In supervised learning, a machine is trained on labeled data (input and output pairs). The machine learns the relationship between the inputs and the outputs, allowing it to make predictions on unseen data.
    - **Example**: A model is trained with labeled images of dogs and cats. Once trained, the model can predict whether a new image is of a dog or a cat.
    - **Use Case**: Fraud detection, image classification, market forecasting.
2. **Unsupervised Learning**
    - **Description**: In unsupervised learning, a machine is trained on unlabeled data. The machine must identify patterns and relationships within the data without explicit guidance on what the outputs should be.

- **Example**: A model analyzes customer behavior and groups them into segments (e.g., frequent buyers, occasional buyers).
- **Use Case**: Customer segmentation, anomaly detection, recommendation systems.

3. **Reinforcement Learning**
   - **Description**: In reinforcement learning, a machine learns by interacting with its environment and receiving feedback in the form of rewards (for good actions) or penalties (for bad actions). This process helps the machine learn through trial and error.
   - **Example**: A self-driving car learns how to navigate by receiving feedback from the environment about its actions.
   - **Use Case**: Robotics, game-playing AI, autonomous vehicles.

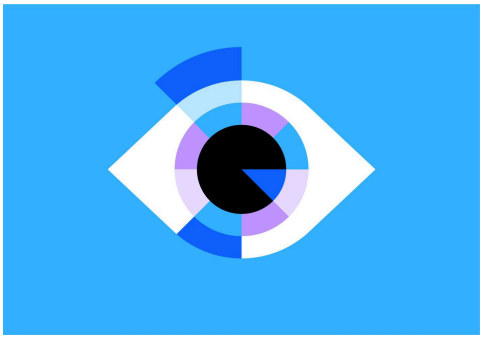
Comparison: Supervised vs. Unsupervised Learning

Supervised Learning	**Unsupervised Learning**
Process: Input and output variables are given.	**Process**: Only input data is given.
Data Type: Algorithms are trained using labeled data.	**Data Type**: Algorithms are used against unlabeled data.

Complexity: Simpler method.	**Complexity**: Computationally complex.
Accuracy: Highly accurate and trustworthy.	**Accuracy**: Less accurate and trustworthy.
Use Cases: Fraud detection, image classification, weather forecasting.	**Use Cases**: Customer segmentation, targeted marketing.

Project Scenario: Train Data Models

In this scenario, Marilyn Shah, the HR director at GAXR, wants to predict future employee attrition. Scott Hill, the data scientist, prepares to train and validate a model using the attrition data from GAXR.



- **Business Problem**: GAXR is experiencing voluntary employee attrition, and the goal is to predict which employees are most likely to leave the company.
- **Training Data**: Scott uses the "Attrition" column (with "yes" or "no" values) to train the model. Since the data is dichotomous, Scott decides to use **supervised learning** and builds a **logistic regression model**.

Logistic Regression Model:

Logistic regression is a popular method used for binary classification problems (e.g., "yes" or "no" outcomes). It estimates the probability of a categorical variable occurring based on independent variables.

- **Example**: Predicting whether an employee will leave the company (yes or no) based on factors like distance from home, marital status, and business group.
- **Visualization**: Logistic regression is often represented using an S-curve, where the probability of the outcome (e.g., attrition) is plotted against an independent variable (e.g., distance from home).

Training and Validating the Model:

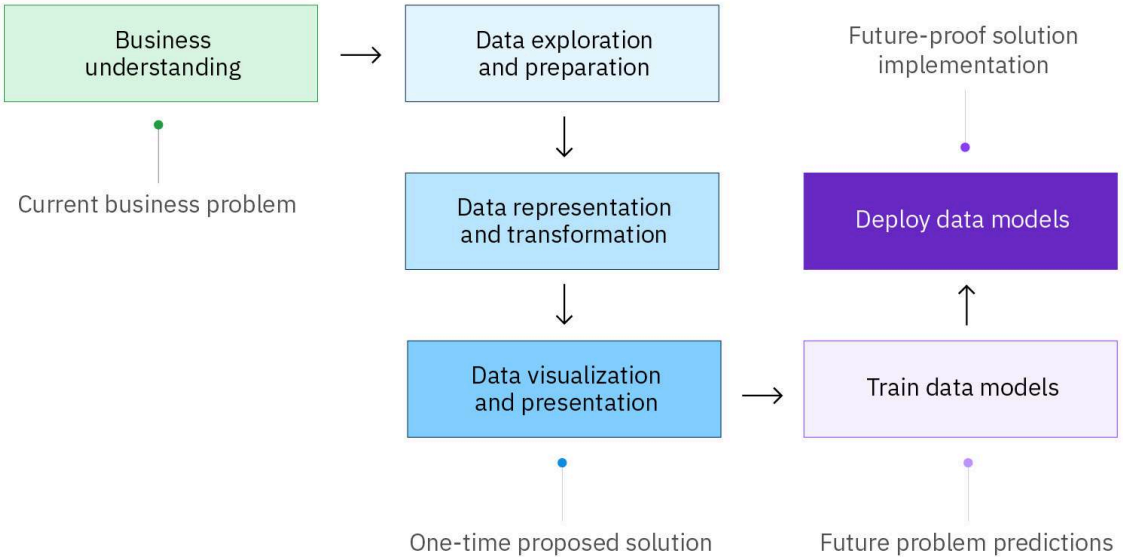
Scott uses **AutoAI** in **IBM Watson Studio** to automate the training and validation of the logistic regression model. After training, Scott reviews the model's performance by comparing the correct and incorrect predictions.

- **Key Findings**: The model does not predict that marital status or business group affects employee attrition. Instead, it finds that the distance from home is a significant predictor of attrition. Employees living farther from the office are more likely to leave the company.

Scott concludes that the model is effective in predicting future employee attrition based on distance from home. This insight will help GAXR address potential attrition by offering employees the option to work from home.

This step is crucial in the data science methodology, as it trains a model to predict future outcomes, helping the business make informed decisions and plan for the future.

Step 6: Deploy Data Models



What is Model Deployment?

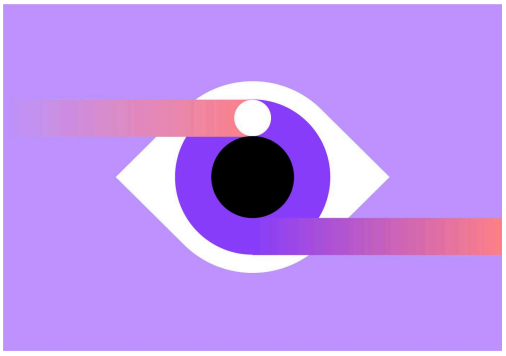
Deploying a model involves integrating the machine learning model into a business' production environment. After the model is trained and saved, it is deployed so it can be used to make predictions or generate insights continuously. This step ensures that the model operates according to a schedule and is maintained over time by a data scientist.

Why Deploy a Model?

Deployment is the final step in the data science pipeline. Once a model is deployed, it can be used to make predictions for real-world applications, helping businesses solve problems and make better decisions based on the insights provided by the model. Model deployment ensures that the model's predictions are actionable and impactful.

Project Scenario: Deploy Data Models

In this scenario, Scott Hill, the data scientist at GAXR, deploys the model to help mitigate employee attrition. By integrating the trained model into GAXR's production system using **IBM AutoAI technology**, the company can use future employee data to predict and prevent attrition due to long commute times.



Once deployed, the model is now a machine learning file type, which can be applied to real-time data to make predictions on employee attrition based on factors like commute distance.

Key Considerations for Model Deployment:

Scott needs to consider the following questions for successful deployment and ongoing maintenance of the model:

1. **How will I evaluate the model's performance in production?**
    - It's essential to track how well the model performs in the real-world environment, ensuring it continues to provide valuable and accurate predictions.
2. **How frequently do I plan on re-training or running the model?**
    - Models may need to be retrained periodically with new data to maintain accuracy and relevance as business conditions change.

3. **Will the format of the production input data differ drastically from the model training data?**
    - If the production data differs significantly from the data used during training, adjustments or retraining may be necessary to ensure the model remains effective.
4. **Will the data come in batches or as a stream?**
    - Understanding whether the model will process data in real-time (stream) or at scheduled intervals (batch) is crucial for deciding how to manage its deployment.
5. **Do I need to run the model offline?**
    - In some cases, models may need to be run offline or in a non-live environment due to operational constraints or data privacy concerns.

Maintaining the Deployed Model:

Once deployed, the model must be maintained and regularly evaluated to ensure its continued effectiveness. This may involve re-training with updated data, monitoring performance, and adjusting the model as business requirements evolve.

Deployment is a crucial phase, as it takes the model from a development environment into one where it can provide continuous value to the business.

This step is essential for ensuring that the insights and predictions generated by the model have practical applications in the business environment. By addressing the maintenance and performance evaluation questions, businesses can ensure the model continues to function effectively over time.

Applying Data Science to Our World

Introduction

In this module, you will learn how data science projects in our world impact industries such as healthcare, transportation, sports, e-commerce, and social platforms.

Learning Objectives

After completing this module, you should be able to:

- Recognize industries and applications of data science in our world to solve problems and help discover innovations.

Data Scientists Have Changed Almost Every Industry

Data science has revolutionized the way data is perceived. There are many data science applications in healthcare, banking, e-commerce, manufacturing, and more. Big data companies like Amazon, Google, and Facebook use data science concepts for business insights and decisions for their organizations.

Data science projects are happening around the world, from predicting the side effects of medications to optimizing routes during traffic rush hours. Data science is used in a variety of industries and all kinds of organizations—some might even surprise you.

Data Science Can:

- Identify and predict disease and personalize recommendations in healthcare
- Optimize shipping routes in real time for transportation
- Accurately evaluate athletes' performance in sports
- Prevent tax evasion and predict incarceration rates for governments
- Automate digital ad placement in e-commerce
- Improve online experiences for gaming
- Create algorithms to pinpoint compatible partners for social media

Source: 22 Data Science Applications and Examples, Built In, by Mae Rice and updated by Jessica Powers, June 2022.

Explore examples of how companies are applying data science to improve and solve global problems.

Expand each section to study applications of data science.

Healthcare

One of the biggest opportunities for data science is in healthcare. Healthcare is an enormous industry, prime for disruption and new insights. Think about the huge amounts of electronic medical records, patient populations, and medical imaging.

Each person will generate 1 million gigabytes (GB) of health-related data in their lifetime, which is equivalent to about 300 million books. In that data is the secret to health and well-being.

Examples of Data Science in Healthcare:

- Google developed a tool, LYNA (short for Lymph Node Assistant), to identify breast cancer tumors that metastasize to nearby lymph nodes. This can be difficult for the human eye to see, especially when the new cancer growth is small. In one trial, LYNA accurately identified metastatic cancer 99 percent of the time using its machine-learning algorithm. More testing is required before doctors can use it in hospitals.

- Oncora's software uses machine learning to create personalized recommendations for current cancer patients based on data from past ones. Healthcare facilities using the company’s platform include UT Health San Antonio and Scripps Health. Their radiology team collaborated with Oncora data scientists to mine 15 years’ worth of data on diagnoses, treatment plans, outcomes, and side effects from more than 50,000 cancer records. Based on this data, Oncora’s algorithm learned to suggest personalized chemotherapy and radiation regimens.

Source: 22 Data Science Applications and Examples, Built In, by Mae Rice and updated by Jessica Powers, June 2022.

Transportation and Logistics

More cities, transit organizations, and departments of transportation are using data science to solve problems and prioritize investments.

Examples of Data Science in Transportation:

- StreetLight uses data science to model traffic patterns for cars, bikes, and pedestrians on North American streets. Based on a monthly influx of trillions of data points from smartphones, in-vehicle navigation devices, and more, Streetlight's traffic maps stay up-to-date. The company's maps inform various city planning enterprises, including commuter transit design.

- The United Parcel Service (UPS) uses data science to optimize package transport from drop-off to delivery. The company's integrated navigation system, ORION, helps drivers choose fuel-efficient routes. ORION has saved UPS approximately 100 million miles and 10 million gallons of fuel per year with the use of advanced algorithms, AI, and machine learning.

Source: 22 Data Science Applications and Examples, Built In, by Mae Rice and updated by Jessica Powers, June 2022.

Sports

The sports industry also benefits from data science, which can be used to calculate statistics across different sports.

Examples of Data Science in Sports:

- WHOOP makes wearable devices that track athletes' physical data, like resting heart rate, sleep cycle, and respiratory rate. The goal is to help athletes understand when to push their training and when to rest, ensuring they're maximizing their physical performance.

- Trace provides soccer coaches with recording gear and an AI system that analyzes game film. Players wear a tracking device, called a Tracer, while its specially designed camera records the game. The AI bot then analyzes that footage and compiles the most important moments in a game.

Source: 22 Data Science Applications and Examples, Built In, by Mae Rice and updated by Jessica Powers, June 2022.

---

e-Commerce

e-Commerce has exploded in popularity, with companies using data science to track everything from marketing to purchases to consumer trends.

For example, if you use social media and search for a product online, you might see ads for that product everywhere immediately afterward. This is how e-commerce companies use data science for advertising.

Example of Data Science in e-Commerce:

- Instagram uses data science to target its sponsored posts, selling everything from trendy sneakers to influencers posting sponsored ads. The company's data scientists gather data on users, including age and education, to craft algorithms predicting what products users are likely to buy.

Source: 22 Data Science Applications and Examples, Built In, by Mae Rice and updated by Jessica Powers, June 2022.

Social Platforms

The rise of social networks has completely altered how people socialize. Now that many relationships begin online, data about people impacts who they get to know next.

Example of Data Science Shaping Human Connection:

- When singles match on Tinder, they can thank the company's data scientists. A carefully crafted algorithm boosts the probability of matches, prioritizing users based on their activity, location, and similar interests, enhancing the chances of a successful connection.

Source: 22 Data Science Applications and Examples, Built In, by Mae Rice and updated by Jessica Powers, June 2022.

In Conclusion:

Data science is being applied across industries in many ways:

- Data science has changed the way pharmaceutical drugs are discovered and developed for people.
- Retailers can develop new products that respond to the needs of customers.
- Companies can detect and prevent cybersecurity threats.
- Organizations can work on solving global challenges like climate change, poverty, inequity, and terrorism.

Data tells businesses a lot and provides a potentially limitless horizon for research and technology.

Explore More!

Check out these resources:

- Applications of Data Science and Business Analytics – A blog by KDnuggets examining how businesses realize the potential of data science.
- Data Science Use Cases Guide – A blog by DataCamp, authored by Elena Kosourova, on implementing data science in various industries.
- Is Data at the Center of Healthcare for the Future? – A blog describing the rise of big data in healthcare.

Data Science is a Team Sport

Introduction

In this module, you will learn about the role of a data analyst, data scientist, and data engineer, and how they collaborate on a project.

Learning Objectives

After completing this module, you should be able to:

- Compare the roles and characteristics of a data analyst, data scientist, and data engineer.

The Data Science Project Team

Data science isn't just about data sources and methodologies—it's also about people. In practice, several people work on a team. It's not just one data scientist! Analysis results are only as good as the team that is responsible for collecting, analyzing, and interpreting the data.

A data science project team works on projects with the goal of extracting valuable insights from data to answer business problems.

In this lesson, you'll learn about and compare the unique roles of the:

- Data Analyst
- Data Scientist
- Data Engineer

Roles in the Data Science Team

What does Joseph do?

Joseph is in the office day-to-day, working with structured data in databases and using statistics.

Joseph's primary responsibility is to collect, organize, clean, and analyze mountains of data. He uses standardized methods and tools from his company to identify trends, find patterns, and make forecasts. Joseph knows the industry and uses his knowledge of the business. He communicates findings and depicts insights to the business sponsors using data visualizations and presentations to help them make decisions and drive action.

Helpful Characteristics and Skills to Have:

- Methodical
- Critical thinker
- Skilled in data transformation tools to clean data and visualization tools to show insights
- Great communicator with presentation skills

What does Zhanna do?

Zhanna is in the office and involved from the beginning to the end of a project.

Given the business problem, Zhanna develops a hypothesis to research and find hidden patterns. She works with data from many sources. Depending on the project, she might travel or go into the field to collect data and measurements. The data can be unstructured or structured if it has already been “wrangled.” Zhanna conducts experiments to build custom models using the company’s data science methodology and tools. She uses techniques such as machine learning to build and train models to predict future outcomes. Overall, Zhanna transforms data into knowledge to produce actionable insights that can be used to improve future outcomes.

Helpful Characteristics and Skills to Have:

- Always curious and wondering “why?”
- Research and scientific mindset
- Problem solver
- Math, statistics, and machine learning
- Skilled in data analytics tools

What does Ronnie do?

Ronnie manages the data infrastructure for the company. His primary responsibility is to set up systems and processes for the data analysts and data scientists to use and rely on when working with data.

Ronnie understands the flow of data and transforms large volumes of raw data into usable “pipelines” for projects. He is focused on tools and uses advanced programming techniques. Ronnie works on the steps to test and deploy machine-learning models into production for the company.

Helpful Characteristics and Skills to Have:

- Tech savvy
- Math, statistics, machine learning
- Skilled in programming
- Familiar with infrastructure architecture (IA)

Steps of the Data Science Project

Step 1:

Joseph, the data analyst, sources a spreadsheet that’s from a company source managed by Ronnie, the data engineer. The spreadsheet contains thousands of rows and hundreds of columns. Joseph uses data transformation tools to perform the task of cleaning or wrangling the data records. Are there missing values? Are there any redundant columns? He removes the “noise” from the data set.

Step 2:

Joseph sends the clean data set to Zhanna, the data scientist, and shares his business knowledge and initial data findings.

Step 3:

Zhanna constructs a hypothesis and asks herself, “What predictions can we make?” She uses statistical techniques and analytics tools to classify the data. She then builds and trains a model with the data.

Step 4:

Zhanna works with Ronnie, the data engineer, to deploy the trained model at the company.

Step 5:

Zhanna works with Joseph and Ronnie to validate the success of the trained model. Is her hypothesis correct? Should she use different techniques? She continues experimenting.

Step 6:

Zhanna extracts the insights and determines actions the company can take to improve future outcomes.

Step 7:

Joseph creates a meaningful and impactful data visualization to “tell the story”. He shares it with the team so they can schedule a presentation with the business sponsor.

Collaboration and Impact

Skilled data analysts, scientists, and engineers are transformative figures in modern business. They collaborate on data science projects to solve problems and speed up innovation. Here is a brief, high-level overview of how Joseph, Zhanna, and Ronnie interact and collaborate on a project.

Interesting Notes:

- Everyone on the team is responsible for understanding the business problem so they can work towards proposing a solution.
- Organizations use different job titles for the roles that work with data. Depending on the size of an organization, a person might have a combined role and responsibilities.
- There are other roles which aren't covered here. For instance, you may find someone who is a **data journalist**. A data journalist maintains a client-focused mindset, and their primary responsibility is to communicate results. Although this is not a technical role, a data journalist is a highly skilled communicator who turns facts into a compelling story with insights and data visualizations.

Data Science Course Overview

1. **Definition of Data Science**:

   Data science involves using scientific methods, algorithms, and systems to extract insights from data, both structured and unstructured.

2. **Importance of Curiosity**:

   Curiosity is crucial in data science as it drives the exploration needed to solve problems and uncover valuable insights.

3. **Difference between Data Analytics and Data Science**:
   - **Data Analytics**: Focuses on analyzing past data to identify trends.
   - **Data Science**: Uses predictive modeling, machine learning, and advanced analytics for deeper insights.

4. **Widely Adopted Data Science Methodologies**:
   - CRISP-DM
   - KDD
   - SEMMA

5. **Data Project Scenario**:

   A data project follows stages such as problem definition, data collection, cleaning, exploration, modeling, and evaluation, with continuous optimization.

6. **Industries and Applications**:
   - **Healthcare**: Medical research, patient care optimization, predictive analytics.
   - **Finance**: Fraud detection, risk management, forecasting.
   - **Retail**: Customer segmentation, inventory management, sales forecasting.
   - **Technology**: AI model development, user experience improvement, big data analytics.

7. **Roles Comparison**:
    - **Data Analyst**: Analyzes data for business decisions (descriptive analytics).
    - **Data Scientist**: Builds predictive models using machine learning and advanced techniques.
    - **Data Engineer**: Designs and builds data infrastructure, ensuring data is accessible for analysis.

Data Science Course Overview

1. **Definition of Data Science**:

   Data science involves using scientific methods, algorithms, and systems to extract insights from data, both structured and unstructured.

2. **Importance of Curiosity**:

   Curiosity is crucial in data science as it drives the exploration needed to solve problems and uncover valuable insights.

3. **Difference between Data Analytics and Data Science**:
   - **Data Analytics**: Focuses on analyzing past data to identify trends.
   - **Data Science**: Uses predictive modeling, machine learning, and advanced analytics for deeper insights.

4. **Widely Adopted Data Science Methodologies**:
   - CRISP-DM
   - KDD
   - SEMMA

5. **Data Project Scenario**:

   A data project follows stages such as problem definition, data collection, cleaning, exploration, modeling, and evaluation, with continuous optimization.

6. **Industries and Applications**:
   - **Healthcare**: Medical research, patient care optimization, predictive analytics.
   - **Finance**: Fraud detection, risk management, forecasting.
   - **Retail**: Customer segmentation, inventory management, sales forecasting.
   - **Technology**: AI model development, user experience improvement, big data analytics.

7. **Roles Comparison**:
    - **Data Analyst**: Analyzes data for business decisions (descriptive analytics).
    - **Data Scientist**: Builds predictive models using machine learning and advanced techniques.
    - **Data Engineer**: Designs and builds data infrastructure, ensuring data is accessible for analysis.

Key Points to Remember in Data Science

1. **Data Science Overview**:

   Data science combines the scientific method, mathematics, statistics, specialized programming, advanced analytics, AI, and storytelling to uncover and explain business insights from data.

2. **Methodologies**:

   Data scientists use repeatable methodologies, which include processes or activities to obtain results. These methodologies are based on scientific principles.

3. **Widely Used Methodologies**:
   - **CRISP-DM** (Cross-Industry Standard Process for Data Mining)
   - **KDD** (Knowledge Discovery in Database)
   - **SEMMA** (Sample, Explore, Modify, Model, and Assess)

4. **Key Activities in a Data Science Project**:
   - Business understanding
   - Data exploration and preparation
   - Data representation and transformation
   - Data visualization and presentation
   - Train data models
   - Deploy data models

5. **Purpose of Models**:

   Data scientists use technologies and tools to build models that predict outcomes or uncover patterns. The goal is to gain insights that lead to actions, improving future results.

6. **Machine Learning in Data Science**:

   Data scientists use machine learning to train models. Machine learning allows

a computer to learn from data, solving problems without needing explicit programming rules.

7. **Data Visualization**:

   Data visualizations should be effective, attractive, and impactful, clearly communicating the insights from data.

8. **Curiosity**:

   Curiosity is one of the most important characteristics of a data scientist, driving the search for insights and innovative solutions.

9. **Applications of Data Science**:

   Data science is applied in various industries, including healthcare, transportation, sports, e-commerce, and more, helping solve business problems and drive innovation.

10. **Collaboration in Data Science**:

    Data science is a team effort, with data analysts, data scientists, and data engineers collaborating to solve business problems and generate insights.

Overview of Data Tools and Languages

About this Course

There are no perfect tools, just good tools for people with certain goals.
Lisa Charlotte Rost, Data Visualization Enthusiast

Those who work in the field of data science solve problems and answer questions every day. They build models to predict outcomes and discover underlying patterns. The outcomes and discoveries enable data science practitioners, such as data scientists, to gain insights that can lead to actions that improve future outcomes. Data transformation tools and technologies that data scientists use are vital to reaching those outcomes.

Welcome to the *Overview of Data Tools and Languages* course! In this course, you'll get an overview of several data analysis and visualization tools. You'll also learn about data tools, languages, libraries, frameworks, software, and more used in the data science field. The purpose of this course is to introduce you to common data science tools so that you're familiar with and can recognize how they're used in data science projects.

Learning Objectives

After completing this course, you should be able to:

- **Recognize the value of collaborating and using open source in the field of data science.**
- **Identify the purpose of GitHub.**

- **Identify the purpose of a selection of common tools to analyze and visualize data, including:**
    - **Microsoft Excel**
    - **Google Sheets**
    - **Structured Query Language (SQL)**
    - **Python**
    - **IBM Watson Studio**
    - **Tableau**
    - **Matplotlib**
- **Identify factors businesses can consider when selecting a data transformation tool.**
- **Identify factors individuals can consider when selecting a data transformation tool.**

Estimated Duration

1 hour, 30 minutes

Completion Criteria

To get credit for completing this course, you need to:

- **Visit all the pages in the course.**
- **Complete the interactive activities and quizzes.**
- **Pass the final assessment.**

You must pass the quiz at the end of each module to continue to the next module.

There is a final, graded assessment at the end of the course. You must achieve 80 percent to pass. Don't worry! If you aren't successful at first, you can retake the quiz as many times as needed to pass.

As you progress through the course, you'll see completion check marks beside the module and lesson names in the side navigation.

Copyright

© Copyright IBM Corporation 2024

Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

z

About this Course

There are no perfect tools, just good tools for people with certain goals.

Lisa Charlotte Rost, Data Visualization Enthusiast

Those who work in the field of data science solve problems and answer questions every day. They build models to predict outcomes and discover underlying patterns. The outcomes and discoveries enable data science practitioners, such as data scientists, to gain insights that can lead to actions that improve future outcomes. Data transformation tools and technologies that data scientists use are vital to reaching those outcomes.

Welcome to the *Overview of Data Tools and Languages* course! In this course, you'll get an overview of several data analysis and visualization tools. You'll also learn about data tools, languages, libraries, frameworks, software, and more used in the data science field. The purpose of this course is to introduce you to common data science tools so that you're familiar with and can recognize how they're used in data science projects.

Learning Objectives

After completing this course, you should be able to:

- Recognize the value of collaborating and using open source in the field of data science.
- Identify the purpose of GitHub.
- Identify the purpose of a selection of common tools to analyze and visualize data, including:
    - Microsoft Excel
    - Google Sheets
    - Structured Query Language (SQL)
    - Python
    - IBM Watson Studio
    - Tableau
    - Matplotlib
- Identify factors businesses can consider when selecting a data transformation tool.
- Identify factors individuals can consider when selecting a data transformation tool.

Estimated Duration

1 hour, 30 minutes

Completion Criteria

To get credit for completing this course, you need to:

- Visit all the pages in the course.
- Complete the interactive activities and quizzes.
- Pass the final assessment.

You must pass the quiz at the end of each module to continue to the next module.

There is a final, graded assessment at the end of the course. You must achieve **80 percent** to pass. Don't worry! If you aren't successful at first, you can retake the quiz as many times as needed to pass.

As you progress through the course, you'll see completion check marks beside the module and lesson names in the side navigation.

Copyright

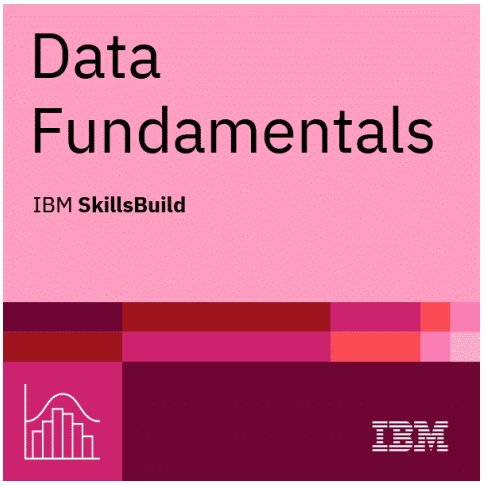
© Copyright IBM Corporation 2024

Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

Ready to get started?

Data Fundamentals Series

Earn the Data Fundamentals Digital Credential from IBM!



This course is one of several in the **Data Fundamentals** series. Complete all the courses in the Data Fundamentals learning plan and earn the **Data Fundamentals** digital credential from IBM!

Credential:

Data Fundamentals

For more information about this credential, visit [Data Fundamentals] on Credly.

What's a Digital Credential?

A credential is an industry-recognized and validated indicator of knowledge and skills acquisition. **IBM SkillsBuild** credentials are aligned to job-market standards and frameworks to help achievers gain visibility with potential employers. IBM SkillsBuild issues its digital credentials through Credly.

Share Your Credentials

Share your credentials on LinkedIn and other social networks with prospective employers, or include them on your resume to amplify your skills!

Course Design

Learn at Your Own Pace with Structured Modules

This course is designed so you can learn at your own pace and easily navigate using the organized module structure. You'll find:

- **Concepts, graphics, and examples** to help you learn.
- **Interactions** to reinforce what you learn.
- **Quizzes** to check your knowledge.
- **Helpful resources** so you can explore more.
- **A final assessment** to test your understanding.

Tips to Be a Successful Learner

- The course is designed for you to complete the modules in sequential order.
- You can pace your learning over a period that works for you and your schedule.
- Your progress is tracked, so you can complete a module or lesson and then come back to pick up where you left off.

Course Design

This course is designed so you can learn at your own pace and easily navigate using the organized module structure. You'll find:

- Concepts, graphics, and examples to help you learn
- Interactions to reinforce what you learn
- Quizzes to check your knowledge
- Helpful resources so you can explore more
- A final assessment
- Tips to be a successful learner

The course is designed for you to complete the modules in sequential order.

- You can pace your learning over a period that works for you and your schedule.
- Your progress is tracked, so you can complete a module or lesson and then come back to pick up where you left off.

Be a curious learner and dive into learning!

The Value of Open Source

Introduction

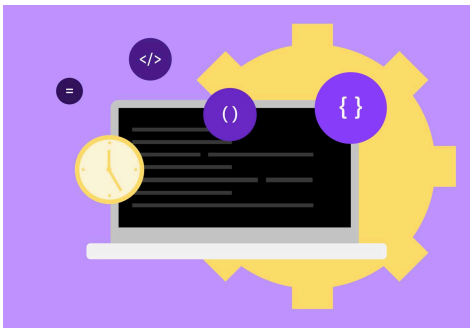
In this module, you'll learn about the collaborative way of working with open source software and have an opportunity to set up a GitHub account to practice adding data.

Learning Objectives

After completing this module, you should be able to:

- Recognize the value of collaborating and using open source in the field of data science
- Identify the purpose of GitHub

Open source software is software with code that is published publicly and can be used by anyone.



Open source software is collaborative, meaning it relies on a virtual community of people to review, change, and share source code with each other. Developers share insights, ideas, and code to create more innovative software solutions.

Here are four key aspects of open source software to keep in mind:

- **Use**: Anyone can use and execute (or run) the software for any purpose, under the license.
- **View**: Anyone can view the source code to understand how the software works.
- **Modify**: Enhancements, bug fixes, and solutions can come from anyone.
- **Share**: Contributions are based on a common, shared purpose.

Now, review some basic terminology and concepts used in the open source world.

- **Community**: Anyone related to an open source project. This is a group of people who receive some benefit from the project.
- **Contributor**: Someone who contributes back to a project.
- **Committer**: Someone who reviews and approves changes to project source code. A committer has write access (or write permission) to a source code repository. This means they are authorized to update the data.
- **Code of Conduct**: Protects the community and provides guidelines about acceptable behavior. All projects should have a code of conduct.

- **Contribution Guidelines**: Describe how to contribute and collaborate. They provide rules about how the community can participate in the open source project.

Open source is a way of working on a project that relies on access and collaboration.

But, there are some common misconceptions about open source.

Why is Open Source Software Valuable?

From the perspective of a data scientist or data engineer who works with code on a project, open access to new and key technologies is critical.

Open source software provides opportunities to learn new software and develop skills. Open source communities can:

- Be a resource for mentorship
- Provide opportunities for developers to interact and showcase their subject matter expertise
- Potentially help people who are beginning a coding journey advance their career

From the perspective of a business, open source software provides many benefits.

- Businesses can increase innovation using the community model. For example, the code a project might need might already have been written and is available in an open source community.
- There is a potential for cost savings when developing software and documenting, testing, and fixing bugs.
- Starting a project or contributing to an existing project can help businesses influence the direction of technological development.
- Open source software can help businesses increase software security and reliability and keep pace with their competitors.

Open source software makes source code available to users and invites a community of technical and non-technical contributors to build on and improve it.

Open Source Flashcards

It's free!

Open source is free, but there are some considerations.

- First, open source gives freedom to those who use it.
- Second, it’s free because there’s no cost upfront to use, modify, or distribute it.
- But keep in mind that open source software rarely comes with the technical support and warranty services that proprietary software provides. Businesses might have expenses later for software maintenance and training.

You have the source code, so it means the software is open source.

Having access to the source code of an application doesn’t mean it is open source.

- The creators still need to grant users access to view and modify the source code.

You can use open source however you like.

Not quite.

- Open source software comes with a license that users must follow.
- Each open source project might have different usage terms and conditions.

You downloaded it for free, so that means it’s open source.

Just because you can download software without paying for it doesn’t mean it is open source.

Git and GitHub

Now, you will learn about open source industry tools. Git and GitHub are two related, but separate, platforms that are extremely popular and widely used by open source contributors.

Git

Git is a version control system for tracking changes in source code during software development. It helps coordinate work among programmers. Git is open source software that's licensed and installed locally on a computer.

You can use Git without using GitHub, but you cannot use GitHub without using Git.

GitHub

GitHub is an online service that provides a place to host source code as well as contribute and collaborate. It lets people work together on projects from anywhere. GitHub is a service, not a software. It offers free, professional, and enterprise accounts.

What can you do in GitHub?

People work with repositories in GitHub. A repository is like a folder for your project. You can have multiple public and private repositories in GitHub. Repositories can contain files, images, videos, spreadsheets, and datasets. GitHub provides the revision history for all files in a repository.

In general, in GitHub you can:

- Host your own open source project. To do this, you create an online repository and add files.
- Contribute to an existing open source project that's public. To do this, you access a copy of the project's repository, make updates, and request a review of the changes you want to contribute.

Activity: Open source in action!

Step 1: Create a GitHub account

Creating a GitHub account is free. It's like creating a new social media account or signing up for a newsletter.

1. Go to the GitHub Join page: GitHub Join.
2. Complete the online form.
3. Type a username. If a username is not available, keep trying.
4. Type your email address and create a password.
5. Verify your account.
6. Select **Create account**.

Done! Now, you're ready to explore GitHub, check out topics, or find out what's trending.

Step 2: Create a repository

1. Go to the GitHub home page and log in.
2. Find **Recent Repositories** and select the **New** button.
3. The **Create a new repository** page displays. Notice that your username appears in the **Owner** field.
4. Type a name in the **Repository name** field.

5. Type a description in the **Description** field (optional but recommended).
6. Select **Public** to make your repository visible to anyone.
7. Check **Add a README file**.
8. Skip **Add .gitignore** and **Choose a license** for now.
9. Select **Create repository**.

Congratulations! You've just created a public repository in your personal GitHub account. Take a moment to check it out. Notice your README file and the icon to edit it.

Step 3: Add a data file

1. Download and save this CSV file: german_credit_data_biased_training.csv.
2. Go to your GitHub repository.
3. From the **Add file** drop-down, select **Upload files**.
4. Drag the CSV file into your repository or select **choose your files**.
5. Under **Commit changes**, put a name in the **Add files via upload** field.
6. Optionally, add a description.
7. Select **Commit changes**.

Now, you can view the CSV file in your repository! Your repository is dated and stored with version control.

Summary

You've successfully created your own GitHub account, set up a repository, and added data! You're ready to explore more across GitHub repositories and communities.

Explore more!

Check out these resources:

- GitHub Docs – online tutorials by GitHub for getting started and all of the online how-to documentation you might need.
- The Open Source Way 2.0 Guidebook – an online guide about best practices for creating and maintaining open source communities.

Tools to Analyze and Visualize Data

Introduction

In this module, you will learn about a selection of data analysis and visualization tools that professionals commonly use in projects. You will gain an understanding of the purpose of each tool, review examples, and explore suggested online resources for further learning.

Learning Objectives

After completing this module, you should be able to:

- Identify the purpose of a selection of common tools to analyze and visualize data, including:
    - Microsoft Excel
    - Google Sheets
    - Structured Query Language (SQL)
    - Python
    - IBM Watson Studio
    - Tableau
    - Matplotlib

Tools Overview

1. Microsoft Excel

- **Purpose**: A spreadsheet tool widely used for analyzing small to medium datasets. Excel allows for data manipulation, calculations, and simple visualizations (charts, graphs).

- **Example**: Use Excel to clean data, perform statistical analysis, and create pivot tables and visualizations.

2. Google Sheets

- **Purpose**: A cloud-based spreadsheet tool similar to Excel, allowing real-time collaboration and integration with other Google Workspace tools.
- **Example**: Use Google Sheets for data analysis with the added benefit of collaborative editing and cloud accessibility.

3. Structured Query Language (SQL)

- **Purpose**: SQL is a programming language used for managing and querying relational databases. It helps in retrieving, manipulating, and analyzing large datasets stored in databases.
- **Example**: Writing SQL queries such as `SELECT * FROM sales WHERE region = 'North'` to retrieve specific data from a database.

4. Python

- **Purpose**: A versatile programming language used for data analysis, visualization, and machine learning. Python offers libraries like Pandas, NumPy, and Matplotlib for data manipulation and visualization.
- **Example**: Analyze data using Pandas, perform statistical analysis with SciPy, or create visualizations with Matplotlib.

5. IBM Watson Studio

- **Purpose**: An integrated environment for data science and machine learning, offering tools for data cleaning, analysis, and visualization. It also enables the creation and deployment of machine learning models.
- **Example**: Build a predictive model and visualize its performance using various chart types in Watson Studio.

6. Tableau

- **Purpose**: A powerful data visualization tool that allows the creation of interactive and shareable dashboards. It helps users transform complex data into easily interpretable visual formats.
- **Example**: Create interactive dashboards to visualize trends, such as sales over time, with dynamic charts and filters.

7. Matplotlib

- **Purpose**: A Python library for creating static, interactive, and animated visualizations. Matplotlib is commonly used in data science for producing high-quality charts and plots.
- **Example**: Plot a line graph using Matplotlib to show the relationship between two variables, such as sales and time.

Learning Resources

After completing this module, you can explore tutorials and online resources to deepen your understanding of each tool.

Overview

There are many useful data analysis tools and platforms available. This module introduces you to some of the most common tools used by data science professionals to analyze and visualize data. The tools covered in this module include:

- **Microsoft Excel**
- **Google Sheets**
- **Structured Query Language (SQL)**
- **Python**
- **IBM Watson Studio**
- **Tableau**
- **Matplotlib**

While this is not an exhaustive list, it will provide you with an orientation to the essential tools, giving you a solid foundation to explore further.

Microsoft Excel

Working with Microsoft Excel

Microsoft Excel (MS Excel) is a spreadsheet application created by Microsoft and one of the most widely used tools for data analysis. It allows users to enter, examine, and interpret data in various ways, making it a versatile tool for data analysis tasks.

Key Features of Microsoft Excel:

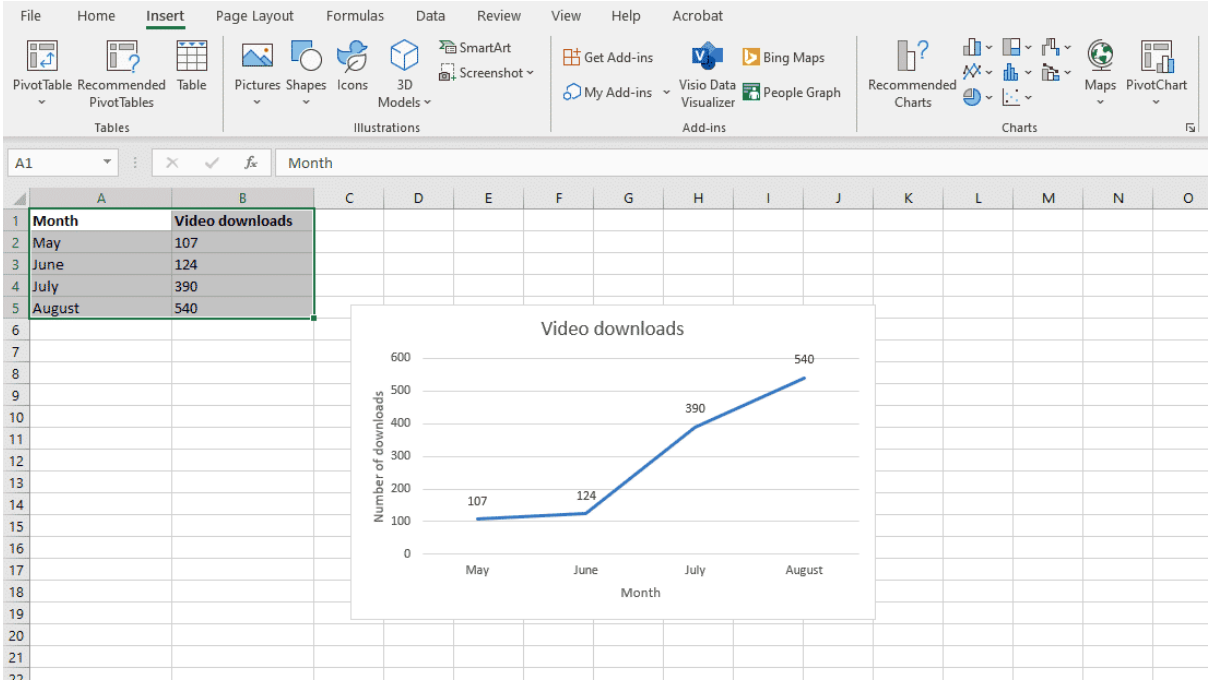
- **Data Manipulation**: You can manipulate and clean data in rows and columns before analysis.
- **Built-in Functions**: Excel offers built-in data analysis functions and features, such as filtering, formulas, and pivot tables.
- **Data Visualization**: MS Excel enables users to visualize data sets and gain insights through charts and graphs.

Chart Types in Microsoft Excel:

MS Excel offers different types of charts to represent data:

- **Pie Charts**
- **Bar Charts**
- **Line Charts**
- **Scatter Plots**

You can also use built-in templates or let Excel recommend a chart type based on your data. Additionally, MS Excel allows combining multiple chart types in a single spreadsheet for more detailed analysis.



Example: Analyzing Video Downloads

Here's an example of how you can use Excel to visualize data. Suppose you have the following data showing the number of video downloads over four months:

Month	Video Downloads
May	107
June	210
July	380
August	540

In this example, you can:

1. Enter the data into columns (e.g., column A for months, column B for downloads).
2. Select the data and insert a chart (in this case, a **Line Chart**) to indicate the increasing trend in video downloads.
3. Customize the chart by adjusting text, colors, labels, and other features to suit your needs.

Note: A potential downside of MS Excel is that it primarily offers two-dimensional charts, with data represented on an x-axis and a y-axis. This can limit the ability to incorporate multiple variables into a single chart.

Explore More!

Here are some resources to help you get started with Microsoft Excel:

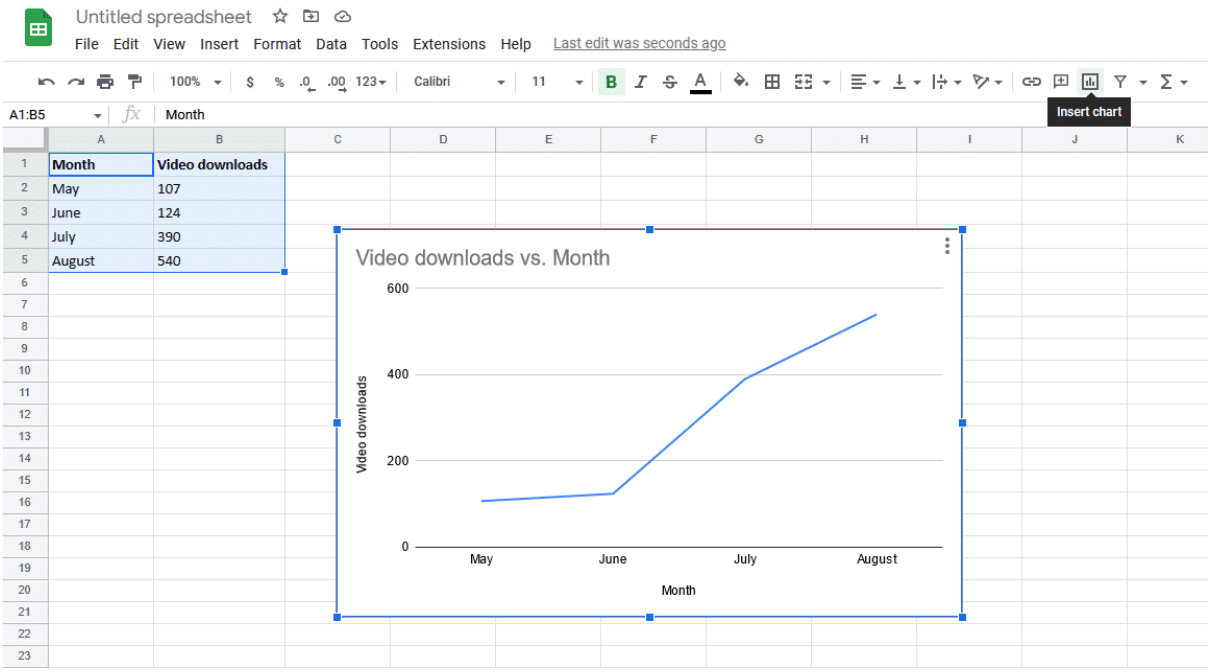
- **Excel Video Training** – A series of introductory online resources from Microsoft to help you get started with Excel.
- **Excel Easy** – Online tutorials from Excel Easy that teach you how to use various MS Excel features, with practical examples.

Google Sheets

Google Sheets is a free tool you can use to perform tasks like entering, analyzing, and visualizing data to make data-driven decisions.

Take a moment to review a Google Sheets example:

This Google Sheets spreadsheet shows the same video download data as the Microsoft Excel spreadsheet example presented in a previous lesson. Note the similarities between the spreadsheet features. In Google Sheets, you can also enter data and insert a line chart to visualize the data.



A screen capture of an example Google Sheets spreadsheet in which the months of May, June, July, and August are in the first column, A. The number of video downloads is in the second column, B. The data in columns A and B is selected to yield a line chart displayed in the foreground of the spreadsheet. The line chart shows the number of video downloads increasing from 107 in May to 540 in August.

Note: Google Sheets is suitable for collaborating online because multiple people can edit the same sheet at once. However, many businesses use Microsoft Excel due to statistical and data analysis features and capabilities that are not available in Google Sheets.

Check it out ...

You can access it right now. Try Google Sheets (opens in a new tab)!

Explore more!

Check out this resource:

Google Sheets training and help – online tutorials from the Google Workspace Learning Center to learn the basics about Sheets.

Here is the structured document you requested:

Structured Query Language (SQL)

What is SQL?

Structured Query Language (SQL) is a standard language used to communicate with databases. As the name suggests, SQL is used when working with structured data in a relational database. SQL has been a standard of the American National Standards Institute (ANSI) since 1986 and the International Organization for Standardization (ISO) since 1987.

SQL is a query language, not a programming language. The purpose of SQL is to ask questions or "query" a relational database or modify its contents. You can use SQL queries to perform operations such as selecting, retrieving, updating, and deleting data.

Standard SQL commands such as “Select,” “Insert,” “Update,” “Delete,” “Create,” and “Drop” can accomplish most database operations.

With SQL, you can:

- Execute queries against a database
- Retrieve data from a database
- Insert records into a database
- Update records in a database
- Delete records from a database
- Create new databases
- Create new tables in a database

- Create stored procedures in a database
- Create views in a database
- Set permissions on tables, procedures, and views

Take a moment to review an SQL example:

Here is a sample database of employee data for a tech company in the United States.

First	Last	ID	Age	City	State
Mike	Jonas	99001	47	Sacramento	California
Santosh	Mohan	99102	26	San Diego	California
Julie	Smith	88323	39	San Diego	California
Lisa	Adams	88678	31	San Jose	California
John	Cho	99544	40	Palo Alto	California
Maya	Rodriguez	934029	24	Palo Alto	California

A common SQL statement is the `SELECT` statement. It queries a database and retrieves selected data that matches the criteria you specify.

Here is the format of a simple `SELECT` statement:

sql

```
SELECT "column1"

     [, "column2", etc]

FROM "tablename"

[WHERE "condition"];
```

To select only employees whose last names end with an “s,” you would use the `LIKE` operator, which matches patterns in the data. You can use the percent sign (`%`) as a "wild card" to match any characters before or after the specified characters.

The `SELECT` statement would look like this:

sql

```
SELECT first, last

FROM empinfo

WHERE last LIKE '%s';
```

This would retrieve employee data for Mike Jonas and Lisa Adams since their last names end with an “s”.

What is NoSQL?

NoSQL stands for "Not Only SQL." While SQL works with structured data in relational databases, NoSQL is used for unstructured data in non-relational databases. NoSQL is suitable for big data and real-time web applications. For example, companies like Twitter, which collect vast amounts of user data daily, may use NoSQL.

NoSQL is commonly used to work with large, unstructured datasets, such as images or text. Types of NoSQL databases include document databases, wide-column databases, and graph databases.

Data scientists also use NoSQL to manage and analyze unstructured data, making it a vital tool in data science.

Explore more!

Check out these resources:

- SQL Tutorial – A series of extensive online tutorials by w3schools to learn SQL and get hands-on practice with exercises.
- Learn SQL – A free, online beginner course by Codecademy to learn SQL (Note: You will need to sign up to access the free course).
- SQL Commands Cheat Sheet – A blog by freeCodeCamp offering a quick start to SQL for beginners.

- SQL vs. NoSQL: What's the difference? – A video by IBM Technology that breaks down the top five differences between SQL and NoSQL databases.

Python

Python is Popular

Python is a free, open-source, and general-purpose programming language that's available for everyone to use. Created by Guido van Rossum and released in 1991, Python was designed to be easy and fun to use. The name "Python" is in acknowledgment of the British comedy group, Monty Python.

Python can be used for a wide range of applications, including creating web applications and supporting data science projects. Its design philosophy emphasizes readability, and it is known for its use of whitespace.

Python's syntax is similar to the English language, with some influence from mathematics. This simplicity allows programmers to write cleaner code and reduce the number of lines needed compared to other programming languages.

Python Capabilities:

- Connect to database systems
- Read and modify files
- Handle big data and perform complex mathematics
- Pair with data manipulation and analysis libraries, like pandas, to obtain insights and create data visualizations

Take a Moment to Review a Python Example

Here's an example of Python code that checks whether a number entered by the user is even or odd.

python

CopyEdit

```python
# Example Python Program to Check if a Number is Odd or Even

number = int(input("Enter a number: "))

if number % 2 == 0:

    print(f"{number} is an even number.")

else:

    print(f"{number} is an odd number.")
```

You can try running this Python program on an online compiler, such as Programiz, where you can enter a number and check if it's even or odd.

Python Community and Learning Resources

Python has a large and growing community, and there is a wealth of open-source packages and libraries available. Whether you are a first-time programmer or experienced with other programming languages, you can learn Python at your own pace.

Most online Python tutorials assume you know how to run a program on your computer, but some websites offer in-browser coding to help you practice.

Explore More!

Check out these resources to help you get started with Python:

- Beginner's Guide to Python – An official online guide to help you get started quickly.

- Python Tutorial for Beginners – A free, online tutorial that covers basic to advanced concepts, with examples to explain the concepts.
- Python Tutorial by w3schools – A series of extensive online tutorials from w3schools to learn Python and get hands-on practice with exercises.

IBM Watson Studio

IBM Watson Studio Can Refine Data

IBM Watson Studio is an integrated development environment (IDE) designed for data science and machine learning. Named after IBM's founder, this platform combines the most useful development and analytic tools, providing a powerful environment that meets large-scale challenges while being simple enough for developers to master quickly.

Key features of IBM Watson Studio:

- Collaborative data science and machine learning environment.
- Works with open-source tools.
- Offers a graphical interface with built-in operations.
- No coding knowledge is required to use the tool.
- Includes a built-in **Data Refinery** tool for transforming raw data into high-quality data.

Data Refinery Tool in IBM Watson Studio

The **Data Refinery** allows you to prepare and transform large amounts of raw data into high-quality data. With this tool, you can:

- Visualize data using built-in charts and graphics to better understand the distribution of your data.
- Schedule jobs for repeatable outcomes.

Data Refinery Example

In IBM Watson Studio, you can analyze data through a project interface with three main tabs: **Data**, **Profile**, and **Visualizations**. For example, auto insurance claims data is uploaded and displayed in the interface, where it can be analyzed and refined using the Data Refinery tool.

Projects / Auto Insurance / AutoInsClaims_csv_shaped / Refine data

Steps | Use a code template to add a step

Data | Profile | Visualizations

	INCIDENT_... Integer	CLAIM_STA... Integer	ODOMETER... Decimal	LOSS_EVEN... Date	CLAIM_INI... Date	POLICE_RE... Integer	CLAIMS_AT... Integer	LOSS_LOCA... Decimal	LOSS_LOCA... Decimal	CLAIM_AM... Decimal	FLAG_FOR Integer
1	3	1	157654.9	2017-04-25	2017-04-28	1	1	41.90210313	-87.75562353	35765	1
2	3	3	226154.5	2018-08-26	2018-08-31	0	1	41.96356191	-87.73139688	1909	0
3	1	1	83968.6	2016-01-07	2016-01-11	0	2	41.73660156	-87.6049681	25730	1
4	1	1	309570.3	2016-12-11	2016-12-18	0	1	41.90992525	-87.73155722	40880	1
5	3	3	136633.9	2017-06-06	2017-06-08	0	1	41.9237502	-87.78988086	2130	0
6	1	3	326514.1	2018-03-06	2018-03-19	0	1	41.90925707	-87.78505659	1970	0
7	2	3	58477.9	2017-08-13	2017-08-18	0	1	41.92839511	-87.7964676	2290	0
8	1	3	176476.9	2017-02-10	2017-02-18	0	1	41.88571554	-87.72837561	2990	0
9	3	3	277812.7	2018-02-16	2018-02-24	0	1	41.89505399	-87.74566176	1170	0
10	5	3	190541.3	2016-12-28	2017-01-11	0	1	41.89197561	-87.61458027	2090	0
11	1	3	290975.1	2017-08-12	2017-08-25	0	1	41.91122338	-87.63865603	2130	0
12	3	1	159873.2	2018-07-30	2018-07-30	1	5	41.73583545	-87.66786637	33040	1
13	1	3	391866.9	2018-02-22	2018-02-26	0	1	41.73416431	-87.55121062	2400	0
14	3	1	384000.6	2017-02-06	2017-02-06	0	3	41.79185207	-87.8013778	21330	1
15	2	3	251644.8	2018-07-10	2018-07-18	1	1	42.00593081	-87.68066049	3427	0
16	2	1	300279.3	2017-07-29	2017-08-10	0	1	41.69340052	-87.61232172	35385.5	1
17	2	3	73419.5	2017-12-30	2018-01-03	0	1	41.84014456	-87.66133946	1570	0
18	2	3	118923.9	2018-01-16	2018-01-28	0	1	41.97517487	-87.7683442	2370	0
19	1	1	182681.3	2017-04-25	2017-05-07	1	1	41.89017255	-87.64999164	26250	1
20	5	1	160429.9	2018-02-10	2018-02-19	1	3	41.84065787	-87.72589914	25830	1
21	2	1	165363.7	2018-01-23	2018-01-24	1	4	41.88470505	-87.66704583	23483	1
22	5	3	240639.2	2016-12-01	2016-12-10	1	1	41.73618257	-87.64365982	2173.5	0
23	4	3	388591.3	2017-07-12	2017-07-12	1	1	41.76203035	-87.61169575	1750	0
24	1	1	59498.7	2016-01-06	2016-01-08	1	1	42.00642207	-87.67812041	20310	1
25	1	3	94043.6	2018-10-13	2018-10-25	0	1	41.81360176	-87.704338	2810	0

Check It Out

If you'd like to dive deeper, watch this 4-minute demo video titled **Preparing Data with Data Refinery**. In the video, a data science team predicts whether a customer is likely to default on their home loan. The team collects three data sets, cleans the data, and selects an initial visualization. You can watch the video to see how the Data Refinery tool is used to shape and clean raw data.

Watch the demo video

Explore More!

Here are some helpful resources to get started with IBM Watson Studio:

- IBM Watson Studio Trial – Start with a free trial of IBM Watson Studio on Cloud Pak for Data (sign-up required).
- Data Refinery Tool – More information about the Data Refinery tool in IBM Watson Studio.
- Use the Data Refinery to Shape Raw Data – A video that demonstrates how to shape raw data using the Data Refinery tool.

Tableau: Creating Interactive Visualizations

Tableau is a popular data visualization and business intelligence software that helps businesses derive meaningful insights from data. It allows users to create powerful pictorial and graphical representations of data to inform business decisions.

With Tableau, you can:

- **Analyze large volumes of data** quickly and efficiently.
- **Create various dashboards, charts, graphics, maps**, stories, and more to aid in decision-making.
- **Perform tasks without needing programming experience**, as it offers an intuitive interface.
- **Design interactive visualizations** that engage viewers with the data.

Tableau Example: Web Traffic Dashboard

Here's an example of a Tableau data visualization dashboard displaying incoming web traffic for a fictitious website. The dashboard features different types of visualizations a business could use to monitor its web traffic, including:

- A **bar chart** showing traffic sources.
- A **scatter plot** analyzing traffic distribution.
- A **map** displaying the geographic locations of users.

The example visualization is interactive. You can visit the interactive dashboard, roll over elements, and select parts of the visualizations to view more detailed information.

Explore More!

Here are some helpful resources to get started with Tableau:

- **Tableau Free Trial** – Sign up for a free trial and explore Tableau.
- **Tableau Free Training Videos** – A series of videos to help you learn how to prepare data and create visualizations (sign-up required).
- **10 Examples of Interactive Map Data Visualizations** – View and interact with 10 examples of map-based visualizations.
- **Data Visualization Tips for More Effective and Engaging Design** – Tips on making data visualizations more dynamic and relevant.
- **Viz of the Day** – A public forum where you can see a new data visualization every weekday from Tableau users.

Tableau: Creating Interactive Visualizations

Tableau is a popular data visualization and business intelligence software that helps businesses derive meaningful insights from data. It allows users to create powerful pictorial and graphical representations of data to inform business decisions.

With Tableau, you can:

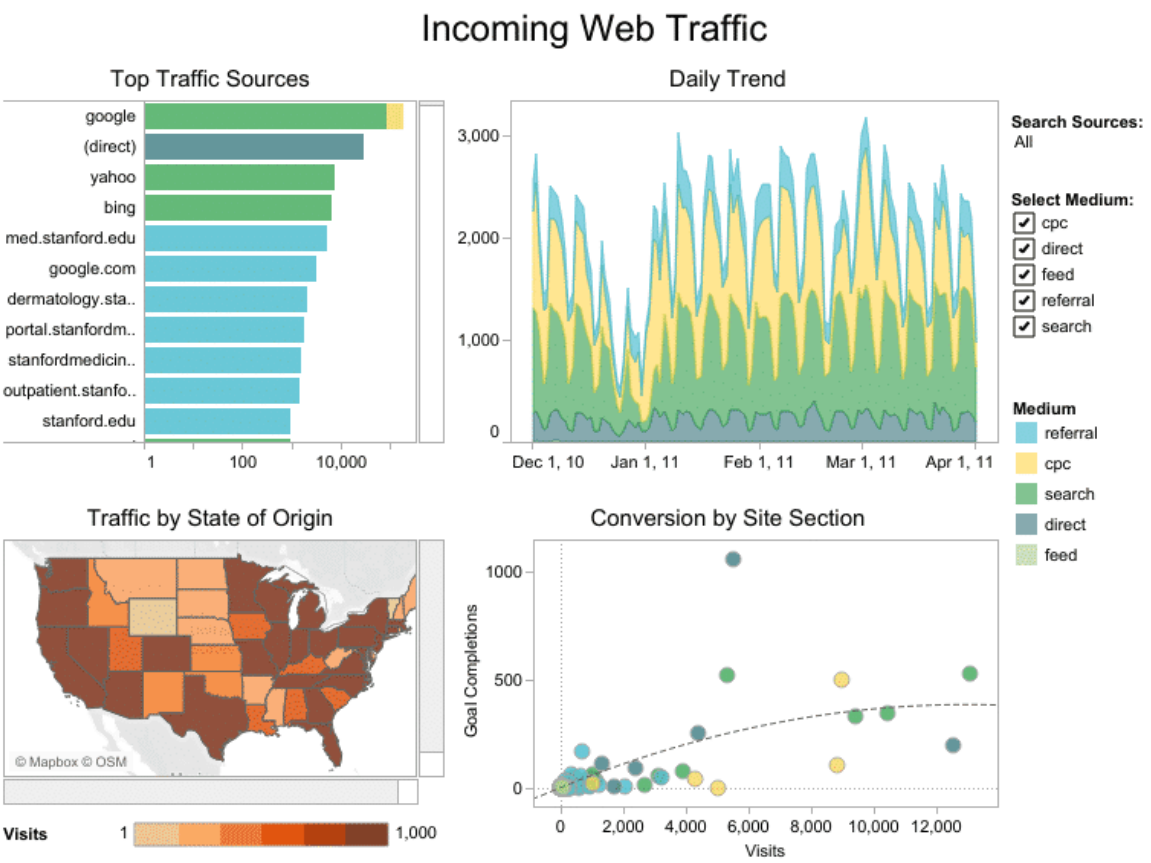
- **Analyze large volumes of data** quickly and efficiently.
- **Create various dashboards, charts, graphics, maps**, stories, and more to aid in decision-making.
- **Perform tasks without needing programming experience**, as it offers an intuitive interface.
- **Design interactive visualizations** that engage viewers with the data.

Tableau Example: Web Traffic Dashboard

Here's an example of a Tableau data visualization dashboard displaying incoming web traffic for a fictitious website. The dashboard features different types of visualizations a business could use to monitor its web traffic, including:

- A **bar chart** showing traffic sources.
- A **scatter plot** analyzing traffic distribution.
- A **map** displaying the geographic locations of users.

The example visualization is interactive. You can visit the interactive dashboard, roll over elements, and select parts of the visualizations to view more detailed information.



Explore More!

Here are some helpful resources to get started with Tableau:

- **Tableau Free Trial** – Sign up for a free trial and explore Tableau.
- **Tableau Free Training Videos** – A series of videos to help you learn how to prepare data and create visualizations (sign-up required).
- **10 Examples of Interactive Map Data Visualizations** – View and interact with 10 examples of map-based visualizations.
- **Data Visualization Tips for More Effective and Engaging Design** – Tips on making data visualizations more dynamic and relevant.
- **Viz of the Day** – A public forum where you can see a new data visualization every weekday from Tableau users.

Matplotlib: A Library for Plots in Python

Matplotlib is one of the most widely used libraries in Python for creating two-dimensional plots from data. It is a cross-platform, open-source library maintained by its community. Matplotlib was introduced by John Hunter in 2002 and allows you to create various types of plots, such as scatter plots, histograms, and bar charts, with just a few lines of code.

You can create static, animated, and interactive visualizations using Matplotlib, and the plots can be exported to various file formats.

Matplotlib Example: Grouped Bar Chart

Here's an example of a Matplotlib script that generates a grouped bar chart, where the data for men and women is plotted side by side.

```
import matplotlib.pyplot as plt

import numpy as np


labels = ['G1', 'G2', 'G3', 'G4', 'G5']

men_means = [20, 34, 30, 35, 27]
women_means = [25, 32, 34, 20, 25]

x = np.arange(len(labels))  # the label locations

width = 0.35  # the width of the bars

fig, ax = plt.subplots()

rects1 = ax.bar(x - width/2, men_means, width, label='Men')

rects2 = ax.bar(x + width/2, women_means, width, label='Women')

# Add labels, title, and custom x-axis tick labels

ax.set_ylabel('Scores')

ax.set_title('Scores by group and gender')

ax.set_xticks(x)

ax.set_xticklabels(labels)

ax.legend()

ax.bar_label(rects1, padding=3)

ax.bar_label(rects2, padding=3)

fig.tight_layout()

plt.show()
```

This script creates a grouped bar chart with labels displaying the scores for each group (G1 to G5) for men and women.

Explore More!

- **Matplotlib Introductory Tutorials** – Learn how to create visualizations with Matplotlib and explore best practices.
- **The Lifecycle of a Plot** – A tutorial that guides you through the creation of a plot from start to finish using Matplotlib.

Additional Tools and Programming Languages

So far, you have learned about various data analysis and visualization tools, along with some suggested resources for further exploration.

Here are some additional data science technologies in the data science ecosystem:

Apache Spark

Apache Spark is an open-source, multi-language analytics engine. It executes data engineering, data science, and machine learning tasks on single-node machines or clusters. Apache Spark is particularly useful for handling very large datasets.

Jupyter Notebook

Jupyter Notebook is an open-source web application that serves as an interactive notebook containing live code, visualizations, equations, and text. Jupyter Notebooks can be collaborative, allowing multiple users to create and share computational documents. They offer powerful, interactive tools for fast and flexible data experimentation and analysis.

R and RStudio

R is a free programming language used for statistical computing and graphics. Many data analysts and data scientists begin with R or one of its popular interfaces. R

offers numerous useful packages for data visualization, such as ggplot2. RStudio is the integrated development environment (IDE) designed for R programming.

Scala

Scala (short for "scalable language") is a robust, statically typed, general-purpose programming language that supports both object-oriented and functional programming. Scala is highly useful for analyzing large datasets while maintaining high performance.

IBM SPSS Statistics

IBM SPSS Statistics is a statistical software suite developed by IBM. It is designed for data management, advanced analytics, multivariate analysis, and business intelligence. SPSS covers the entire analytical process, from data collection and preparation to analysis, reporting, and deployment. Its user-friendly interface enables efficient data organization and extraction of actionable insights.

Summary

Data science tools and programming languages integrate computer science, statistics, predictive analytics, and other fields to enable deeper insights from data. These tools facilitate the collection, manipulation, and analysis of business data to derive valuable insights. By leveraging these technologies, data scientists can

perform complex tasks more efficiently and help businesses develop solutions for success.

Is There a Right Tool?

Introduction

In this module, you'll learn about what businesses consider when selecting data tools and programming languages. Additionally, you'll explore what factors individuals can consider when choosing a data transformation tool to learn.

Learning Objectives

After completing this module, you should be able to:

- Identify factors businesses consider when selecting a data transformation tool.
- Identify factors individuals consider when selecting a data transformation tool.

Key Considerations for Businesses

When selecting a data transformation tool, businesses typically evaluate:

1. **Scalability and Performance** – Can the tool handle large datasets efficiently?
2. **Cost and Licensing** – Is the tool cost-effective for the organization?
3. **Integration with Existing Systems** – Does it work well with current databases and workflows?
4. **Security and Compliance** – Does the tool meet industry security and regulatory requirements?
5. **Support and Community** – Is there adequate documentation, support, and a strong user community?

Key Considerations for Individuals

When selecting a data transformation tool to learn, individuals may consider:

1. **Ease of Learning and Usability** – Is the tool beginner-friendly and well-documented?
2. **Popularity and Industry Demand** – Is the tool widely used and in demand for job opportunities?
3. **Compatibility with Personal Projects** – Does it align with personal or freelance work needs?
4. **Availability of Learning Resources** – Are there tutorials, courses, and community support available?

Conclusion

Selecting the right data transformation tool depends on various factors, whether for business applications or individual learning. By understanding these considerations, you can make informed decisions to enhance productivity and career growth.

Your Toolbox You can analyze data with "this" tool or "that" programming language.

You can visualize data with "that" tool or "this" programming language.

Each data transformation tool is designed with a certain use in mind. Sometimes, tools overlap and serve the same purpose. But every tool has its advantages and disadvantages. The experiences can vary, and people will have their own preferences.

(Illustration of a hammer, screwdriver, and wrench in a toolbox.)

For instance, Python is a free, open-source programming language that's popular among data analysts and data scientists because it contains many libraries they can use. Microsoft Excel might be someone's preferred tool for creating a chart because it offers many types of built-in charts.

So, there isn't just one right tool for everyone! There are all kinds of tools in the data "toolbox" and factors to consider for which to use.

How Does a Business Select Tools?

There are many factors that a business might consider when choosing data analysis and visualization tools. Here are several questions a business might consider:

- What type of data will be analyzed?
- Do specific processes, such as data modeling, need to be performed?
- What is the volume of data? Is it big data?
- What type of visualizations are needed? For instance, do they need to be interactive?
- What level of technicality is needed, and what job roles will be using the tools? For instance, will data scientists or marketing professionals use the tool?
- What is the budget?

It's important for businesses to select the right tools for their needs, at the right time!

Tools to Consider During a Project

From determining the problem to presenting the solution, there's a variety of tools that align with the steps of a data science project.

Learn how each step in a typical data science project has certain questions to ask, techniques to use, and tools that help.

This outline presents a structured approach to a data science project, covering key stages and questions for each phase. Here's a breakdown of the steps and tools:

Step 1: Determine the Problem

- **Questions to ask:** What's the business problem?
- **Techniques to consider:** Ask questions, use the 5 Why's.
- **Tools to consider:** Design thinking.

Step 2: Collect and Clean Data

- **Questions to ask:** Where is the data? Is it the right data? How can the data be collected? What needs to change?
- **Techniques to consider:** Structured and/or unstructured data collection and cleaning.
- **Tools to consider:** SQL, NoSQL, Microsoft Excel, Google Sheets.

Step 3: Analyze Data

- **Questions to ask:** What stands out?
- **Techniques to consider:** Regression, classification, clustering, initial visualizations to look for trends.
- **Tools to consider:** Python, IBM Watson Studio, IBM SPSS, RStudio.

Step 4: Validate Analysis and Model

- **Questions to ask:** Does this answer the business problem? Do the results make sense?
- **Techniques to consider:** Machine learning, business models.
- **Tools to consider:** Python, IBM Watson Studio, IBM SPSS, RStudio.

Step 5: Visualize Data

- **Questions to ask:** What's the best way to depict the data? Does the visualization connect insights to the problem?
- **Techniques to consider:** Storytelling, communicating insights, presenting findings.
- **Tools to consider:** Tableau, Matplotlib, Microsoft PowerPoint.

Summary

This approach demonstrates how different techniques and tools can be applied to a data science project, aligning with business objectives and problem-solving. It emphasizes decision-making and taking action based on insights gathered at each stage.

How to Choose a Tool to Learn

When it comes to data analysis and visualization, there is a wide range of tools available, each catering to different levels of experience and project needs. Some tools are simple to use, while others are more complex. Not all tools are suitable for every learner or every type of project.

The visualization below helps illustrate the relationship between a tool's flexibility and its ease of use, particularly for beginners. You may recognize some tools, while others might be new to you.

Note: This is a general overview and not an exhaustive list of all tools available. It is meant to guide your thinking when selecting a tool.

Take a moment to review the visualization:

- The **Y-axis** shows the range from **Highly Flexible** to **Not Flexible**.
- The **X-axis** shows the range from **Easy to Learn** to **Hard to Learn**.
- Each app or library is represented by a dot in one of the four quadrants, indicating its level of flexibility and ease of learning.

Source and inspiration: *What I Learned Recreating One Chart Using 24 Tools: Finding the best tool means thinking hard about your goals and needs* by Lisa Charlotte Rost, 2016.

Key Takeaways from the Visualization

- **Easier-to-learn apps** tend to be more limited in features.
- **Programming languages or libraries** may require a steeper learning curve, but they offer greater flexibility and more options.
- Remember: "There are no perfect tools, just good tools for specific goals."

Reflection: What do you think?

What data tool or programming language are you interested in learning, and why?

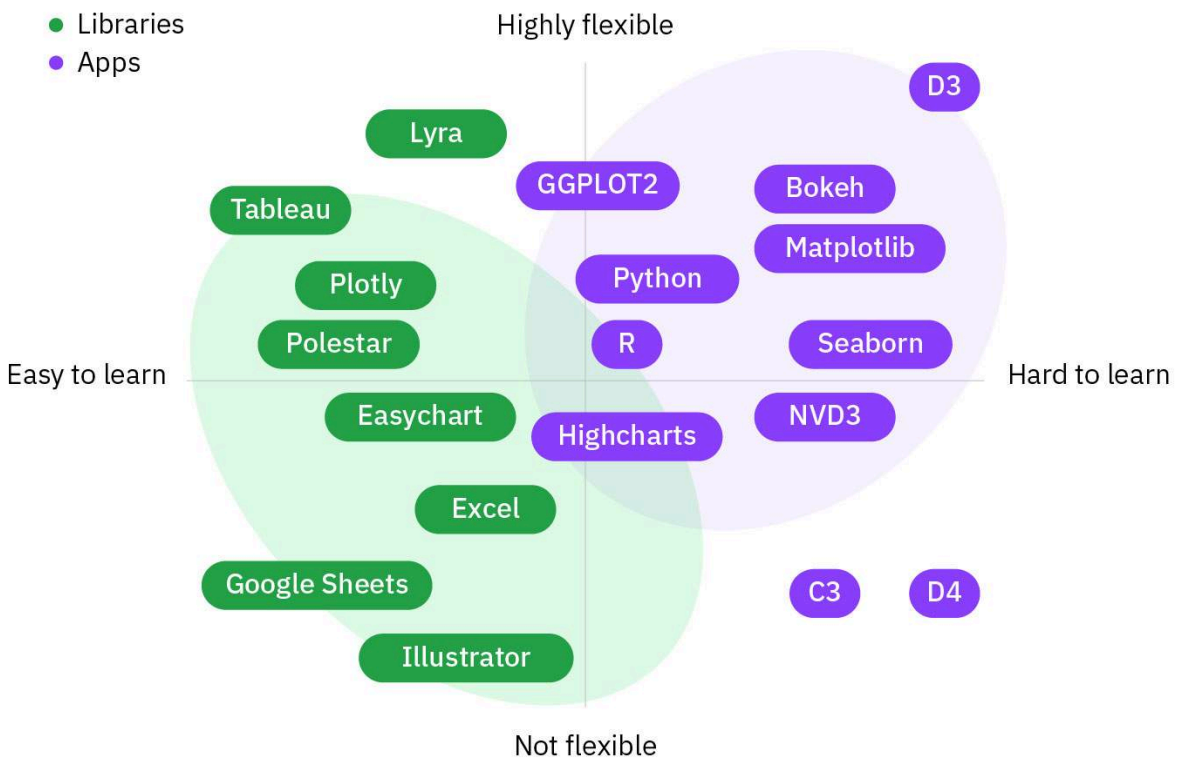
Note: Feel free to enter your response in the provided text box. When finished, click DONE. Writing down your thoughts is a great way to process them. You can also download and save your responses, which will not be retained in the text box when you move on in the course.

It's fantastic to be a lifelong learner! Many data analytics and visualization tools are available, and many of them are beginner-friendly, making them great options for new learners.

How to Choose a Tool to Learn

When it comes to data analysis and visualization, there is a wide range of tools available, each catering to different levels of experience and project needs. Some tools are simple to use, while others are more complex. Not all tools are suitable for every learner or every type of project.

The visualization below helps illustrate the relationship between a tool's flexibility and its ease of use, particularly for beginners. You may recognize some tools, while others might be new to you.



Note: This is a general overview and not an exhaustive list of all tools available. It is meant to guide your thinking when selecting a tool.

Take a moment to review the visualization:

- The **Y-axis** shows the range from **Highly Flexible** to **Not Flexible**.
- The **X-axis** shows the range from **Easy to Learn** to **Hard to Learn**.
- Each app or library is represented by a dot in one of the four quadrants, indicating its level of flexibility and ease of learning.

Source and inspiration: *What I Learned Recreating One Chart Using 24 Tools: Finding the best tool means thinking hard about your goals and needs* by Lisa Charlotte Rost, 2016.

Key Takeaways from the Visualization

- **Easier-to-learn apps** tend to be more limited in features.
- **Programming languages or libraries** may require a steeper learning curve, but they offer greater flexibility and more options.
- Remember: "There are no perfect tools, just good tools for specific goals."

Reflection: What do you think?

What data tool or programming language are you interested in learning, and why?

Note: Feel free to enter your response in the provided text box. When finished, click DONE. Writing down your thoughts is a great way to process them. You can also download and save your responses, which will not be retained in the text box when you move on in the course.

It's fantastic to be a lifelong learner! Many data analytics and visualization tools are available, and many of them are beginner-friendly, making them great options for new learners.

You've come so far!

And learned so much! You've explored open source and received an introduction to some common data analysis and visualization tools, equipping you with the knowledge to recognize how they are applied in data science projects.

Now that you've completed the Overview of Data Tools and Languages course, you should be able to:

- **Recognize the value of collaborating and using open source** in the field of data science.
- **Identify the purpose of GitHub**.
- **Identify the purpose of a selection of common tools** used to analyze and visualize data, including:
    - Microsoft Excel
    - Google Sheets
    - Structured Query Language (SQL)
    - Python
    - IBM Watson Studio
    - Tableau
    - Matplotlib
- **Identify factors businesses can consider when selecting a data transformation tool**.
- **Identify factors individuals can consider when selecting a data transformation tool**.

Key Points to Remember

1. Open Source Software

Open source software is publicly available code that anyone can use and collaborate on within a virtual community. It enables data science professionals to learn new software and develop skills. Businesses benefit from open source through **accelerated innovation, cost savings, and enhanced competitiveness**.

2. GitHub

GitHub is a **free, online platform** used to **host, contribute, and collaborate** on source code using repositories.

3. Microsoft Excel & Google Sheets

- **Microsoft Excel** is a powerful tool for **entering, manipulating, analyzing, and visualizing data**.
- **Google Sheets** is a **free alternative** to Microsoft Excel.

4. Structured Query Language (SQL)

SQL is a **standard language** used to **query and communicate** with databases, allowing users to **modify and retrieve data efficiently**.

5. IBM Watson Studio

IBM Watson Studio provides a **data refinery tool** with **built-in charts and graphics** to help understand data distribution.

6. Tableau

Tableau is a **powerful data visualization software** used to analyze **large datasets** and create **interactive dashboards**.

7. Matplotlib

Matplotlib is a **Python library** used for **creating various types of plots and charts**.

8. Selecting the Right Tools

There are **many data science tools and programming languages** available. **Choosing the right tool at the right time is crucial** for a successful data science project.

9. Accessibility & Learning

- Many tools **do not require prior coding experience**.
- Some tools offer **free trials** or **online tutorials**, making it easier to get started.

Clean, Refine, and Visualize Data with IBM Watson Studio

About this Course

Do you think you can work on a data science project? Can you understand a data set in IBM Watson Studio? Can you conclude insights from that data?

Welcome to the **Clean, Refine, and Visualize Data with IBM Watson Studio** course! In this course, you'll practice cleaning, refining, and visualizing data, in a series of simulations, using IBM Watson Studio with the data refinery tool.

Learning Objectives

After completing this course, you should be able to:

- Explain the purpose and key features of IBM Watson Studio.
- Set up a new project in IBM Watson Studio.
- Import a data set.
- Clean data using the data refinery tool.
- Refine data using the data refinery tool.
- Create a data visualization.
- Conclude insights from a data visualization.

Estimated Duration

1 hour, 35 minutes

Completion Criteria

To get credit for completing this course, you need to:

- Visit all the pages in the course.
- Complete the simulations.
- Pass the quizzes.

You must pass the quiz at the end of each module to continue to the next module. Don’t worry! If you aren’t successful at first, you can retake the quiz as many times as needed to pass.

As you progress through the course, you’ll see completion check marks beside the module and lesson names in the side navigation.

Copyright

© Copyright IBM Corporation 2024

Course materials may not be reproduced in whole or in part without the prior written permission of IBM.

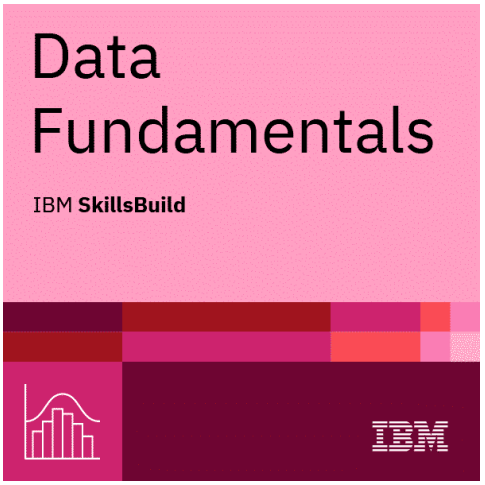
Version Information

- **Version**: 1.0.1
- **Date**: 20-September-2024

- **Release Notes**: Added instructional text to quizzes.

Data Fundamentals Series

Earn the Data Fundamentals Digital Credential from IBM!



This course is one of several in the **Data Fundamentals** series. Complete all the courses in the Data Fundamentals learning plan and earn the **Data Fundamentals** digital credential from IBM!

Credential:

Data Fundamentals

For more information about this credential, visit [Data Fundamentals]() on Credly.

What's a Digital Credential?

A credential is an industry-recognized and validated indicator of knowledge and skills acquisition. **IBM SkillsBuild** credentials are aligned to job-market standards and frameworks to help achievers gain visibility with potential employers. IBM SkillsBuild issues its digital credentials through Credly.

Share Your Credentials

Share your credentials on LinkedIn and other social networks with prospective employers, or include them on your resume to amplify your skills!

Course Design

Learn at Your Own Pace with Structured Modules

This course is designed so you can learn at your own pace and easily navigate using the organized module structure. You'll find:

- **Concepts, graphics, and examples** to help you learn.
- **Interactions** to reinforce what you learn.
- **Quizzes** to check your knowledge.
- **Helpful resources** so you can explore more.
- **A final assessment** to test your understanding.

Tips to Be a Successful Learner

- The course is designed for you to complete the modules in sequential order.
- You can pace your learning over a period that works for you and your schedule.
- Your progress is tracked, so you can complete a module or lesson and then come back to pick up where you left off.

Course Design

This course is designed so you can learn at your own pace and easily navigate using the organized module structure. You'll find:

- **Engaging, hands-on simulations** so you can learn step-by-step.
- **Quizzes** to assess your knowledge.
- **Tips** to be a successful learner.

Sequential Learning

The course is designed for you to complete the modules in sequential order.

- You can pace your learning over a period that works for you and your schedule.
- Your progress is tracked, so you can complete a module or lesson and then come back to pick up where you left off.

Learning Approach

Be a curious learner and dive into learning!

About this Module

Introducing IBM Watson Studio

Introduction

In this module, you will learn about IBM Watson Studio features and one of its tools, the data refinery tool.

Learning Objectives

After completing this module, you should be able to:

- Explain the purpose and key features of IBM Watson Studio.

IBM Watson Studio

How Can You Turn Data into Knowledge?

IBM has a solution called **IBM Watson Studio**. It's an integrated development environment (IDE) named after IBM's founder. This platform combines the most useful development and analytic tools, wrapping them in a powerful development platform that meets large-scale challenges yet remains simple enough for developers to master quickly.

Solving AI Development Problems

With IBM Watson Studio, businesses can simplify data projects through a streamlined process to extract insights from data, enabling them to become smarter faster. The platform offers an easy-to-use, collaborative data science and machine learning environment that allows users to:

- Build and train models.
- Prepare and analyze data.
- Share insights—all within a single environment.

Key Features

IBM Watson Studio provides a unified environment for:

- Sharing work to solve problems within the system rather than starting from scratch with each new issue.
- Quickly building machine learning and deep learning algorithms.

Benefits for Data Scientists

IBM Watson Studio delivers:

- A streamlined process for data projects.
- A collaborative data science and machine learning environment.
- Easy-to-create visualizations.
- Access to open-source tools.
- A comprehensive way to develop, train, manage, and deploy AI-powered applications.

IBM Watson Studio is a comprehensive, all-in-one environment to explore more possibilities with data and artificial intelligence (AI).

Data Refinery Tool

IBM Watson Studio offers many tools, including the **data refinery tool**. It's beneficial because it saves data preparation time by transforming large amounts of raw data into consumable, high-quality information that's ready for analytics.

Capabilities for Data Scientists

With the data refinery tool, data scientists can:

- Create a workflow to clean and shape data.
- Understand the quality and distribution of data using dozens of built-in charts, graphs, and statistics.
- Schedule data jobs for repeatable outcomes.
- Visualize data to discover insights.

Key Features

The data refinery tool is interactive and easy to use. You don't need coding skills to work with it. It includes:

- Over 100 built-in operations to help you transform data.
- Intuitive interfaces for quick data preparation and visualization.

Next, you're going to experience how the data refinery tool in IBM Watson Studio works.

Start Your Project

Introduction

In this module, you will learn about the scenario for your data science project. You will also get hands-on practice in a simulation to set up a new project in IBM Watson Studio and import data into your project environment.

Learning Objectives

After completing this module, you should be able to:

- Set up a new project in IBM Watson Studio.
- Import a data set into your project.

Let me know if you'd like any adjustments or additional details!

Your Data Science Project

Business Problem and Goal

You are a data scientist working at an auto insurance company. The director of the Fraud Claims Department is kicking off a new data science project and serves as your business sponsor.

Key Details:

- The Fraud Claims Department has improved the customer experience by offering online and mobile platforms for submitting automobile insurance claims. This advancement allows the company to process transactions more quickly, but it has resulted in claim processors having less time to identify and counteract fraud.
- Fraud is a growing concern, and the company has already lost millions of dollars due to fraudulent claims.
- The goal of the project is to predict fraud and identify suspicious activity to significantly reduce losses due to fraud.

Your To-Do List

Using IBM Watson Studio and the Data Refinery tool, you will perform the following tasks:

- **Provision the Watson Studio Service**
- **Set up a new project**
- **Import auto insurance claims data**
- **Clean the data set**
- **Refine the data set**

- **Create and study a data visualization to gain insights**

Your Data Set

You will be working with a fictitious data set, modeled after a real auto insurance company's recording of events. The data is in a **comma-separated values (CSV)** file format, with the following specifications:

- **975 rows** of data
- **38 columns** of data
- Each row represents a single person's claim data. The claim was submitted for auto repairs after an accident. All of these claims have already been approved by the auto insurance company.

Important Notes About Auto Insurance:

- Drivers usually sign a six-month policy with an auto insurance company.
- The driver pays a premium to the company either monthly or in a lump sum.
- Several factors determine the cost of an auto insurance policy, including:
    - **The type of car insured**: particularly its safety record and repair costs.
    - **The driver's record**: for example, more speeding tickets make a driver riskier to insure.
    - **The driver's age**: younger drivers, like teenagers, tend to have higher premiums due to inexperience.
- Drivers with fewer accidents and tickets, part-time drivers, those who take driver education courses, and families with multiple cars generally enjoy lower-cost premiums.

Important Note:

You do not need to download the data set, access or purchase IBM Cloud, or work directly in IBM Watson Studio to complete this course. Simulations will be used to practice the steps in a safe, pretend environment.

Simulation: Start a Project and Upload Data

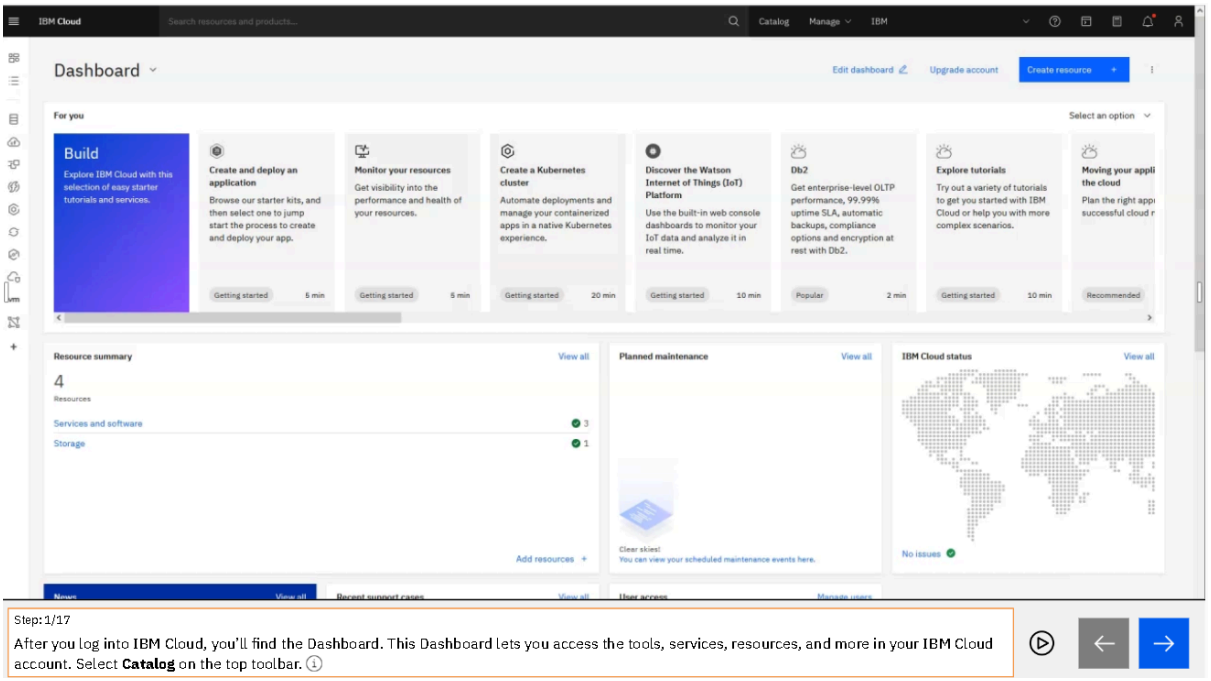
Your first step is to access **IBM Watson Studio** to get your data science project started.

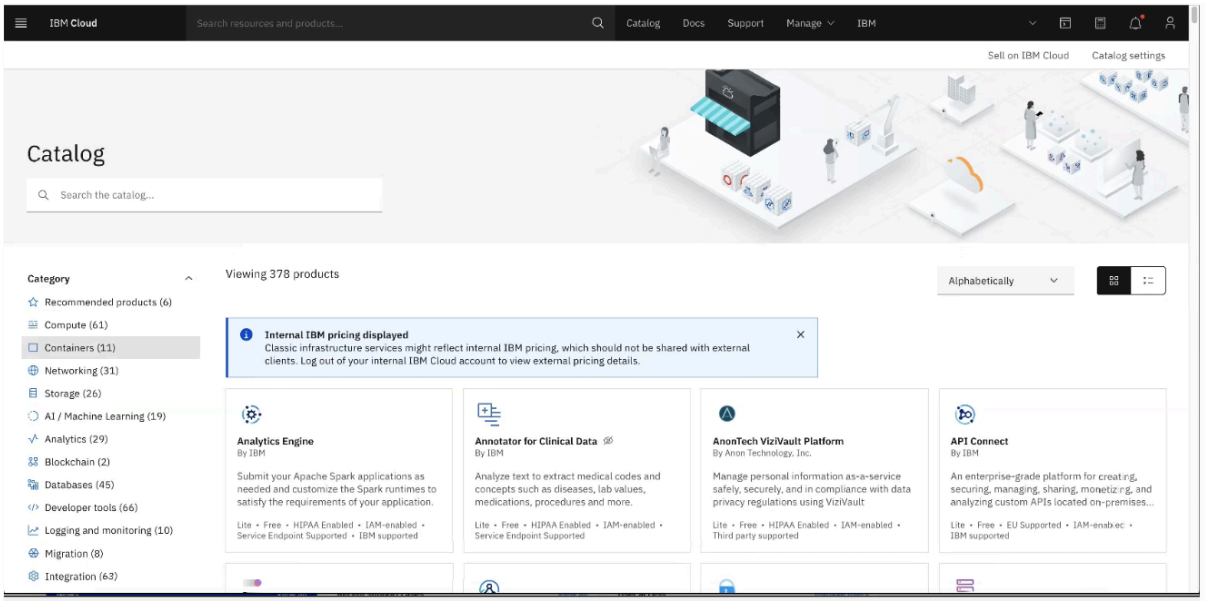
In this simulation, you will use IBM Watson Studio to perform the following steps:

1. **Provision the service**: Set up the necessary services in IBM Watson Studio.
2. **Start a new project**: Create a new project within IBM Watson Studio.
3. **Upload the data set**: Import the auto insurance claims data set into your project.

Steps to Complete

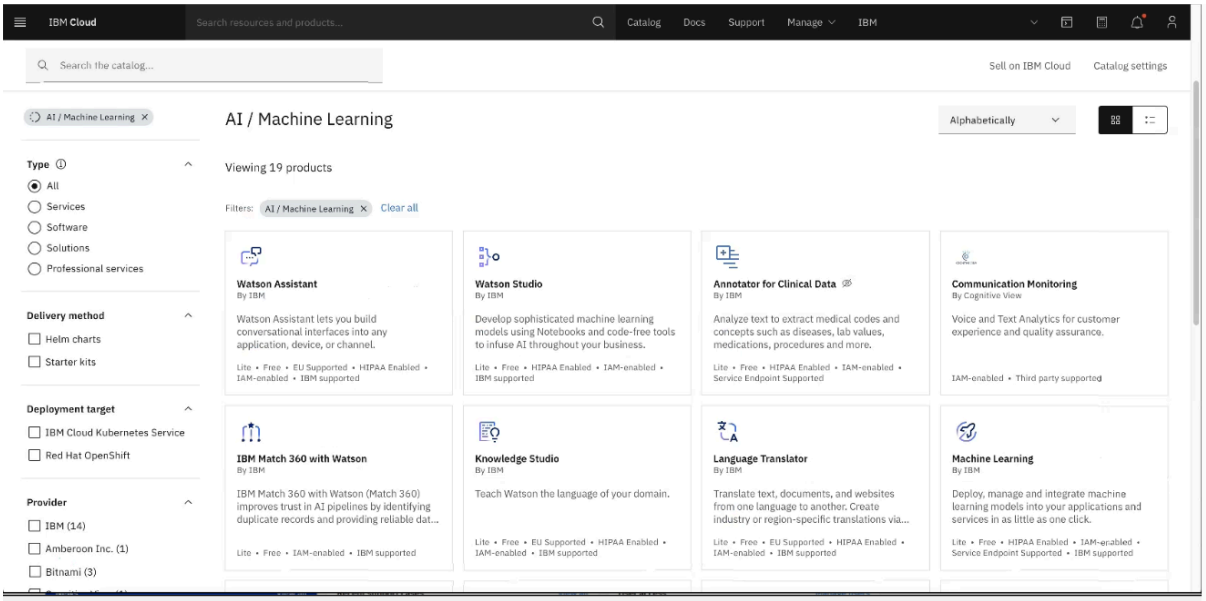
There are **17 steps** in this simulation to guide you through the process of setting up your project and uploading the data.





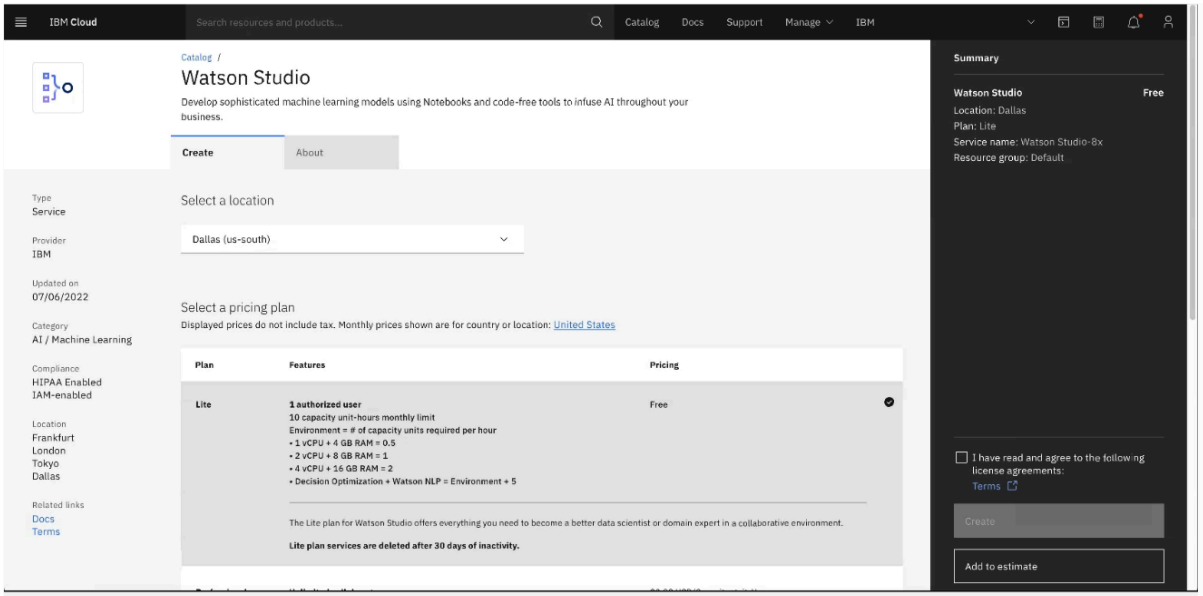
Step: 2/17

This is the catalog of IBM products and services. Notice on the left side under **Category** that there is a large collection of available tools. Select the **AI / Machine Learning** category.



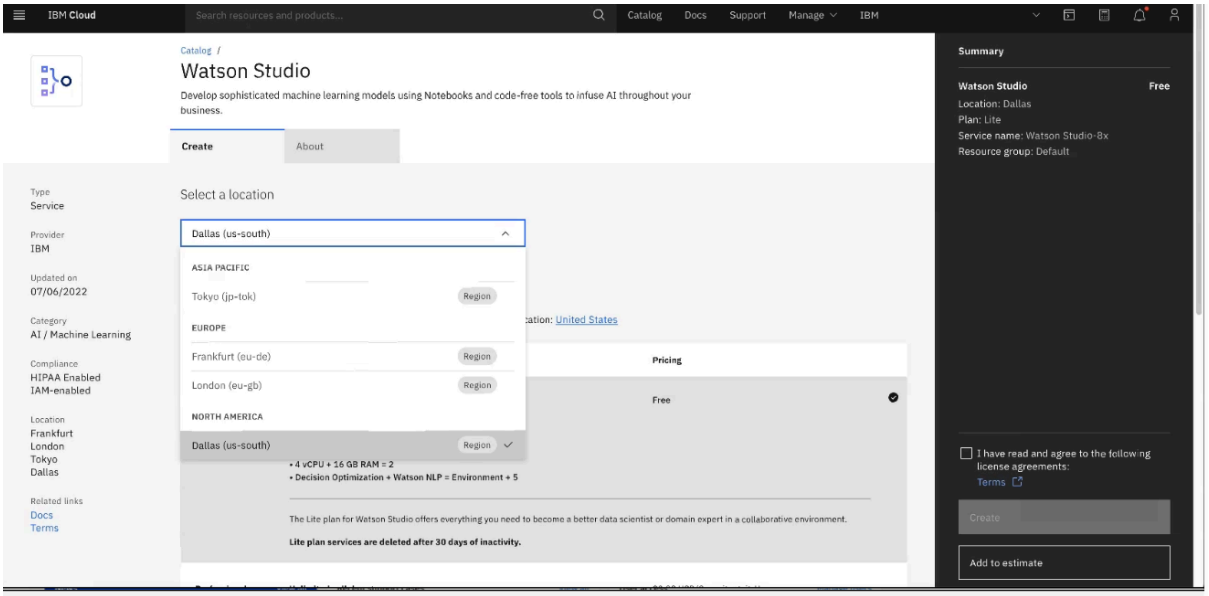
Step: 3/17

Notice there are several AI and machine learning products listed. Find and select the **Watson Studio** block.



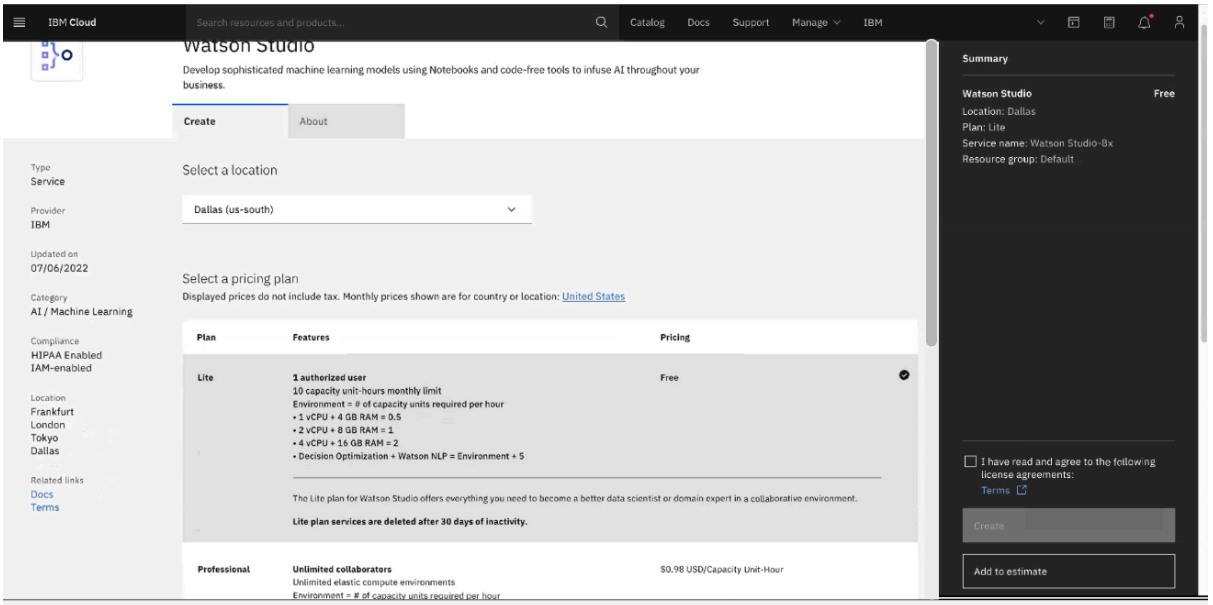
Step: 4/17

Now it's time to provision IBM Watson Studio as your service. It's good to pick one location close to you and keep using the same location across your projects. Under **Select a location**, select **Dallas (us-south)** from the drop-down list.



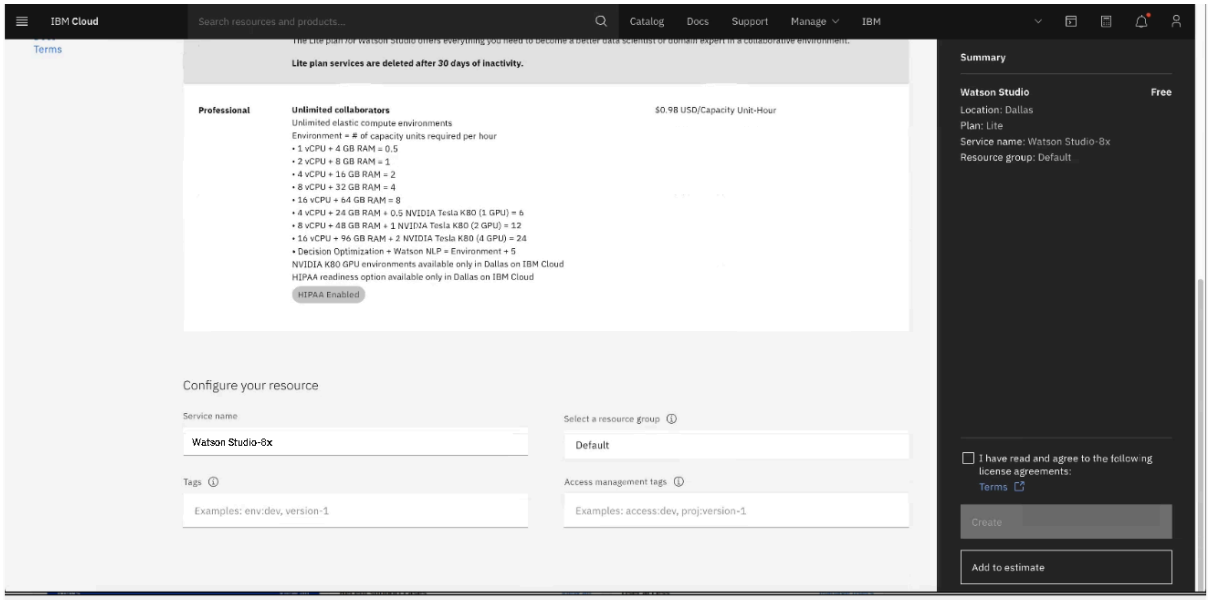
Step: 4/17

Now it's time to provision IBM Watson Studio as your service. It's good to pick one location close to you and keep using the same location across your projects. Under the **Create** tab and select a location, select **Dallas (us-south)** from the drop-down.



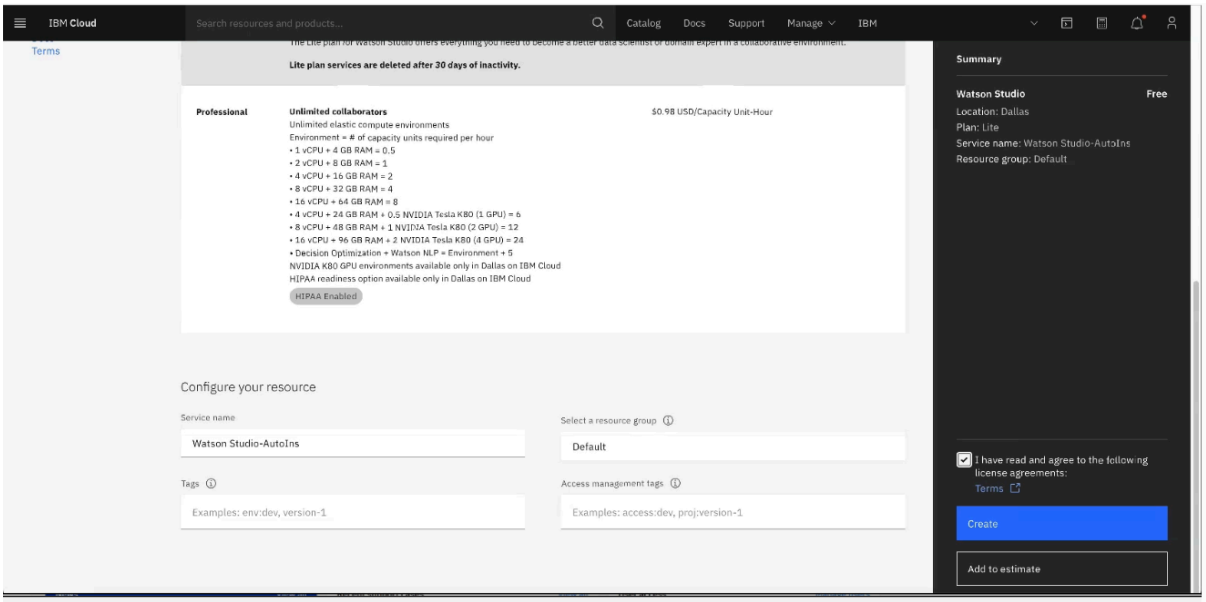
Step: 5/17

For this simulation, you are using a free account, so you don't need to review a pricing plan. Scroll down and select the **Next arrow** to continue.



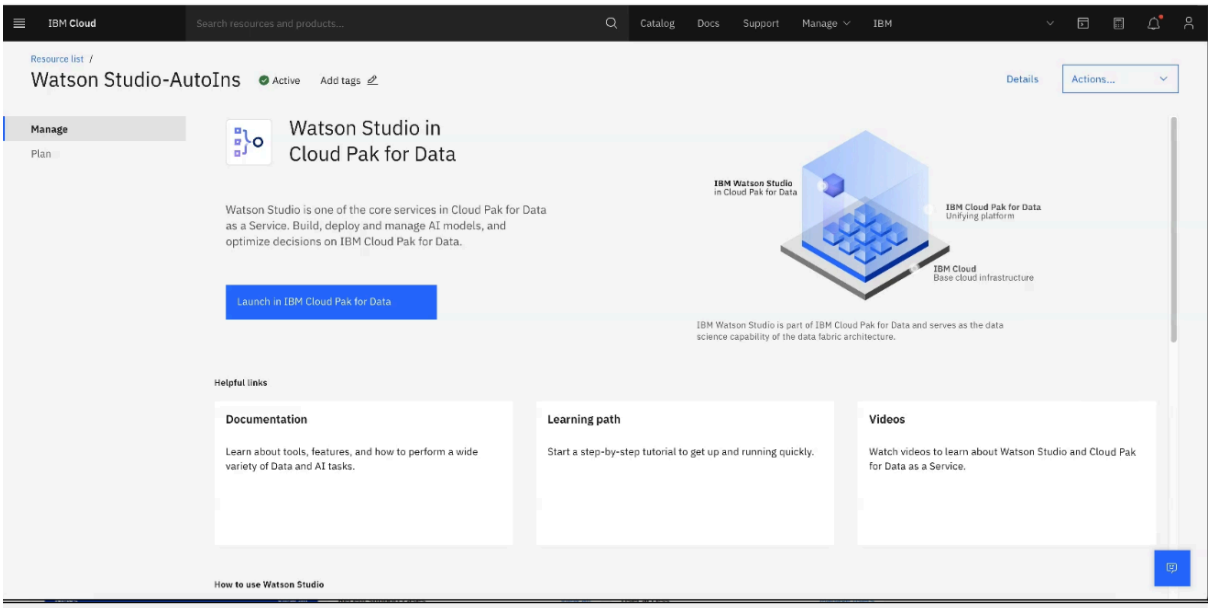
Step: 6/17

Under the **Configure your resource** section, there's an automatically generated **Service name**. It's a good practice to keep "Watson Studio" and then add a name that's relevant to your project. Type "Watson Studio-AutoIns" in the **Service name** field and press **Enter**.



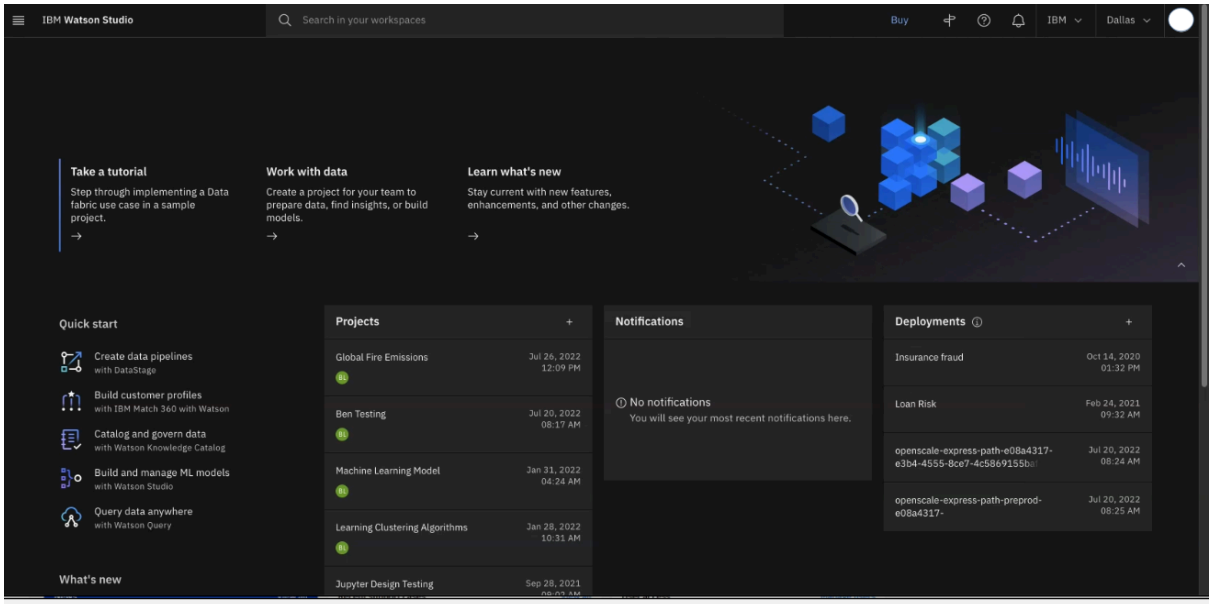
Step: 7/17

You don't need to fill out the other fields. In the lower right, assume that you've checked that you have read and agree to the license agreements, then select **Create.**



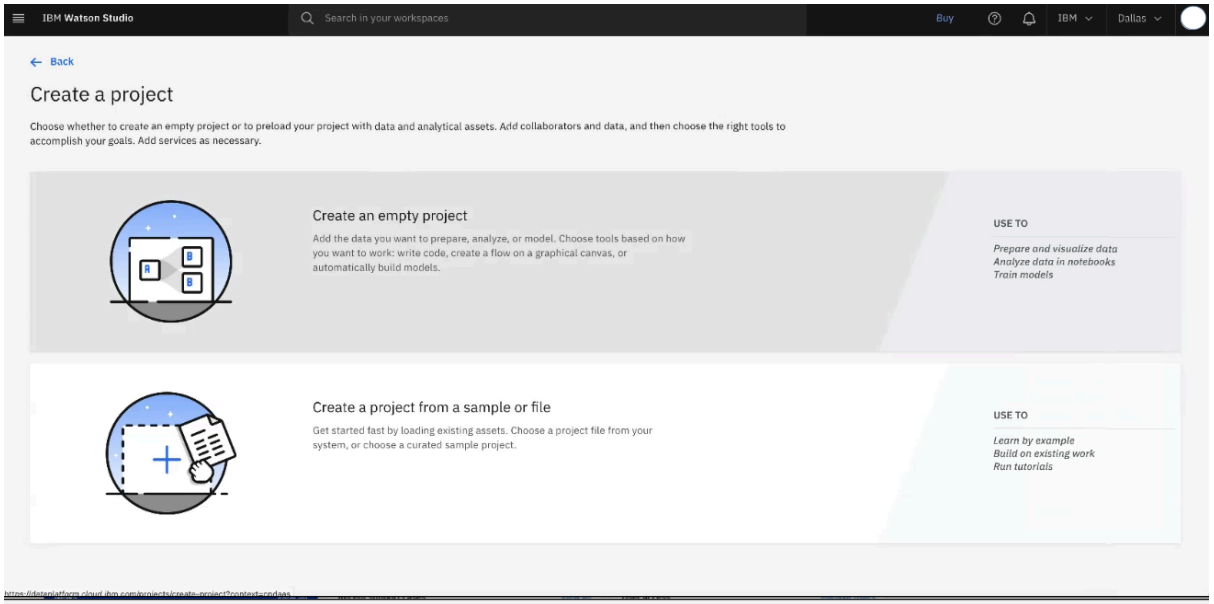
Step: 8/17

You've provisioned the IBM Watson Studio service! Notice the service name you picked displays in the upper left. Now, you need create a new project. Select **Launch in IBM Cloud Pak for Data.**



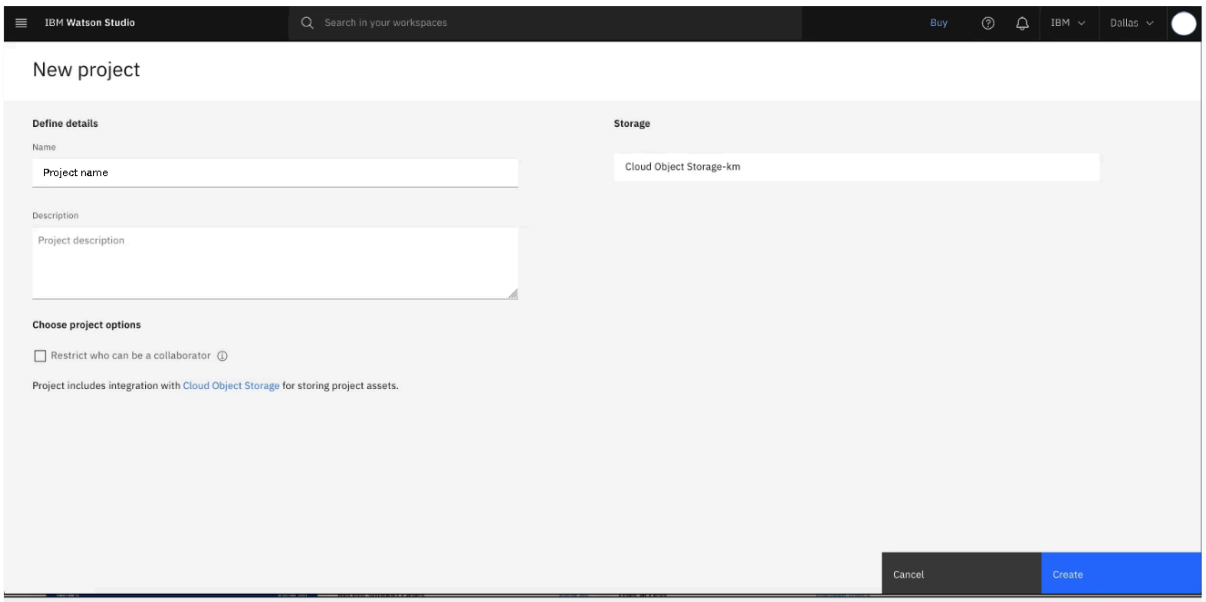
Step: 9/17

You are working on a new project, so select **+** next to **Projects** and then select **Create an empty project.**



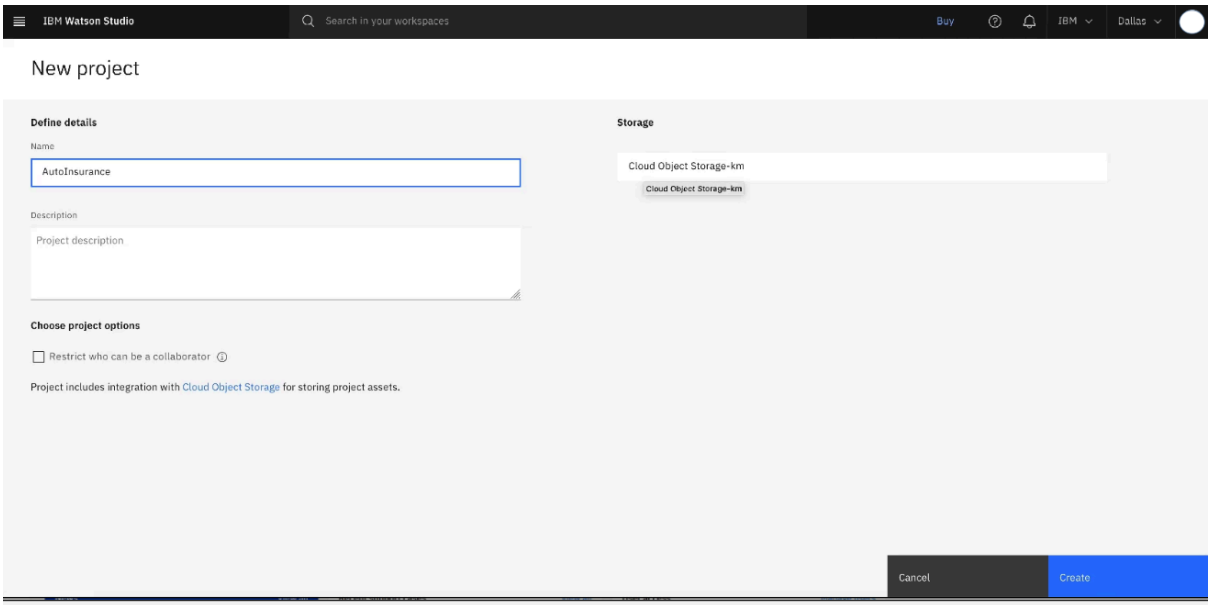
Step: 9/17

You are working on a new project, so select **+** next to **Projects** and then select **Create an empty project.**



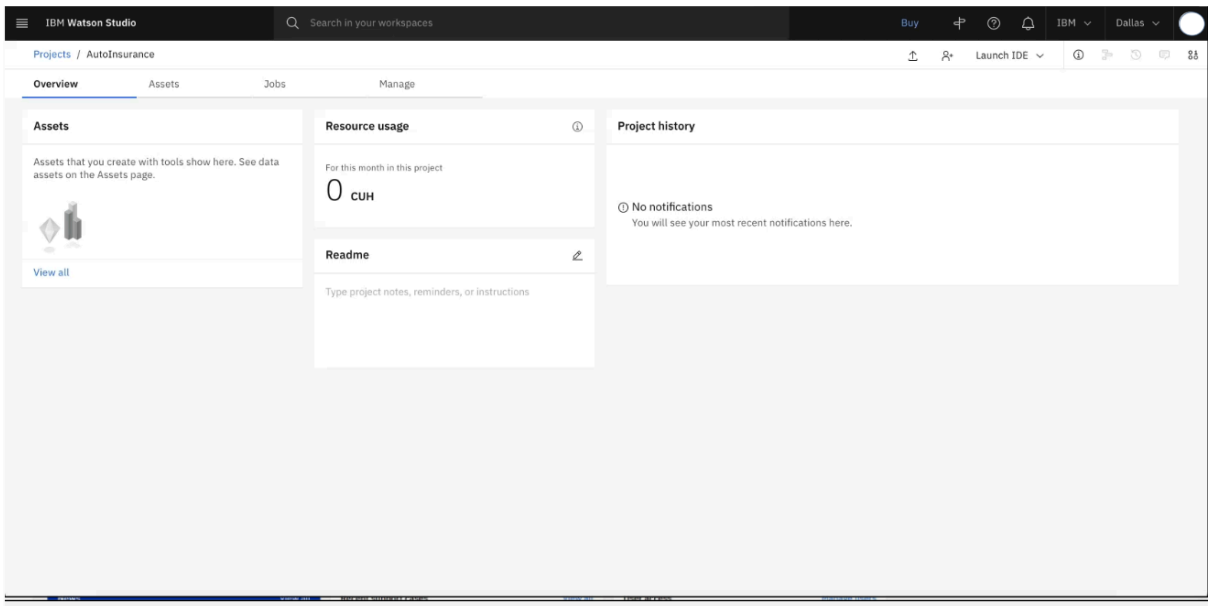
Step: 10/17

Under the **Define details** section, you can name your project. Type "AutoInsurance" in the **Name** field and press **Enter**.



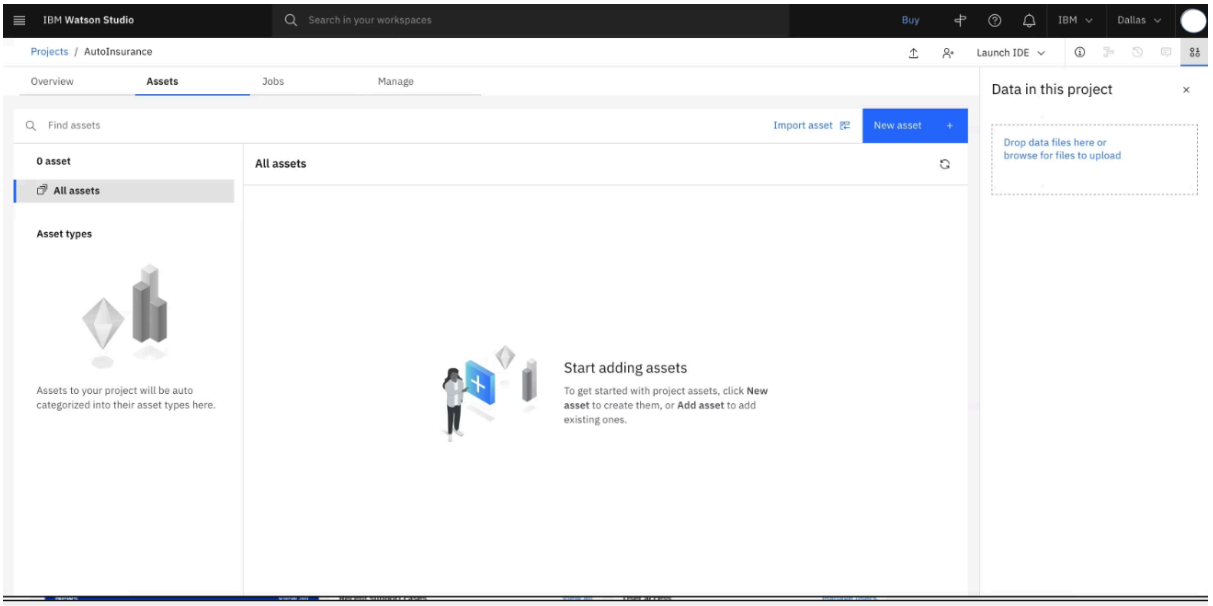
Step: 11/17

Notice under **Storage** that IBM Watson Studio already provisioned Cloud Object Storage (COS) so you have a database. This COS can store unstructured data like images and text. Soon, you'll work with structured data in a comma-separated values (CSV) file. Select **Create.**



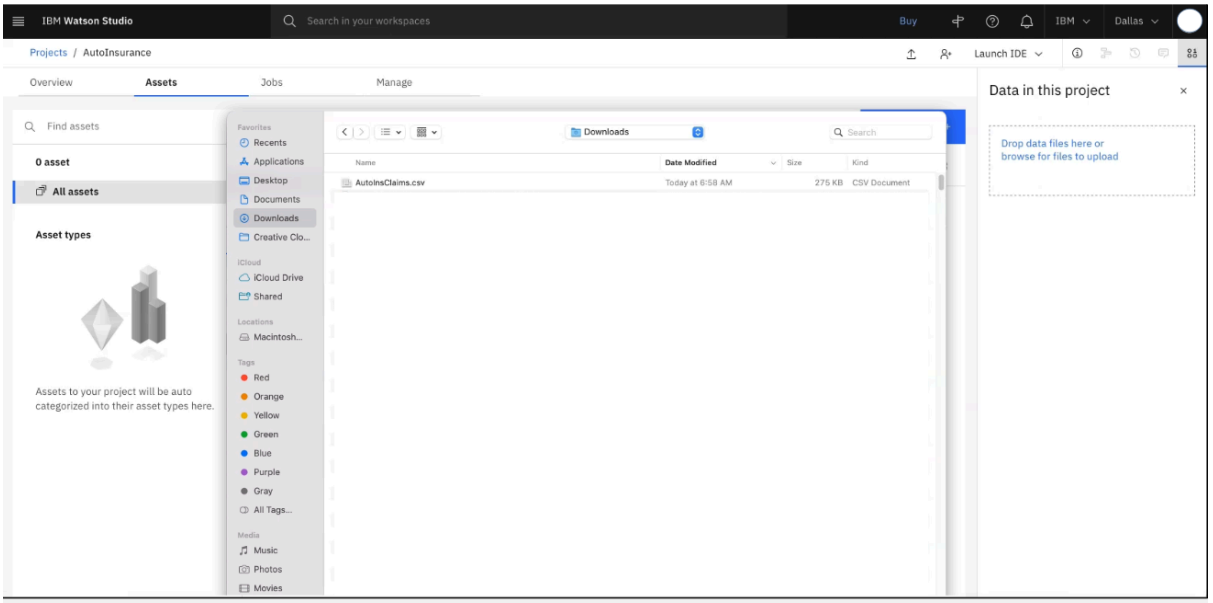
Step: 12/17

You've created the project! This is the IBM Watson Studio Projects dashboard. You can see your project name at the top. This is the **Overview** tab. Select the **Assets** tab.



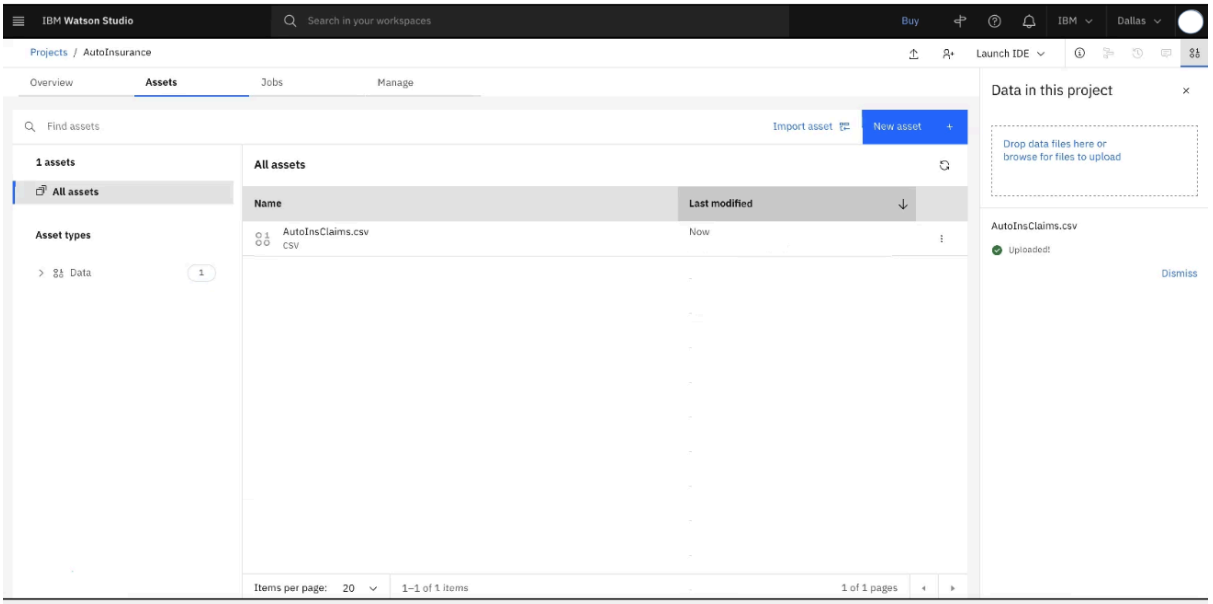
Step: 13/17

It's time to upload your data set. Select **New asset** to open your local drive and find the AutoInsClaims.csv file that you'll be working with for the project.



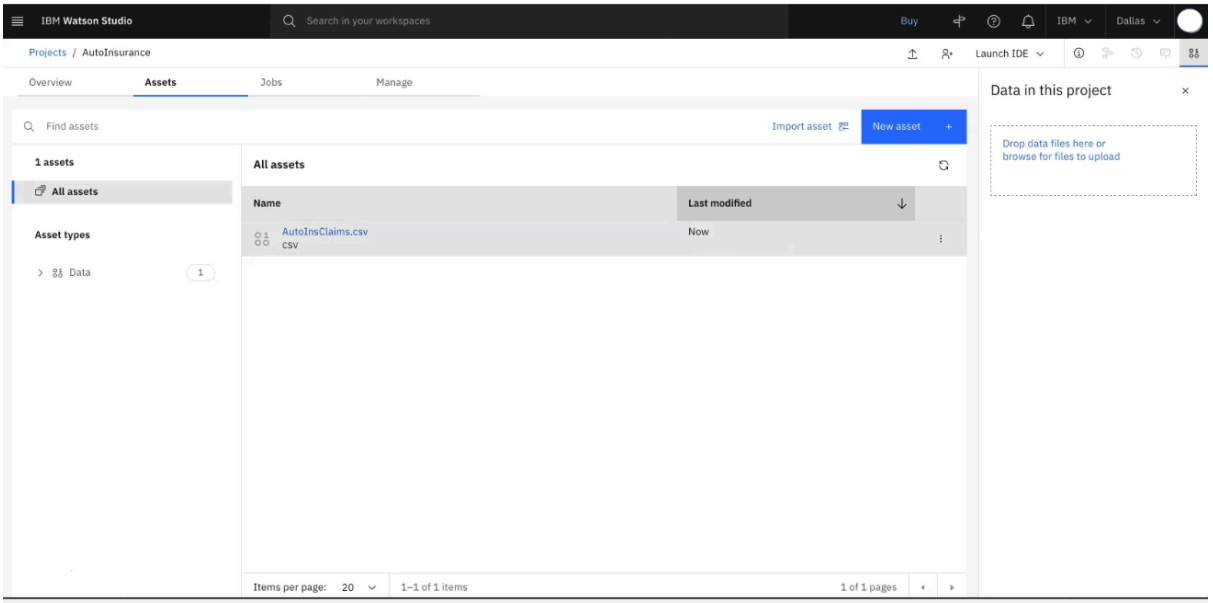
Step: 14/17

Drag and drop the **AutoInsClaims.csv** file to **Drop data files here or browse for files to upload** under **Data in this project.**



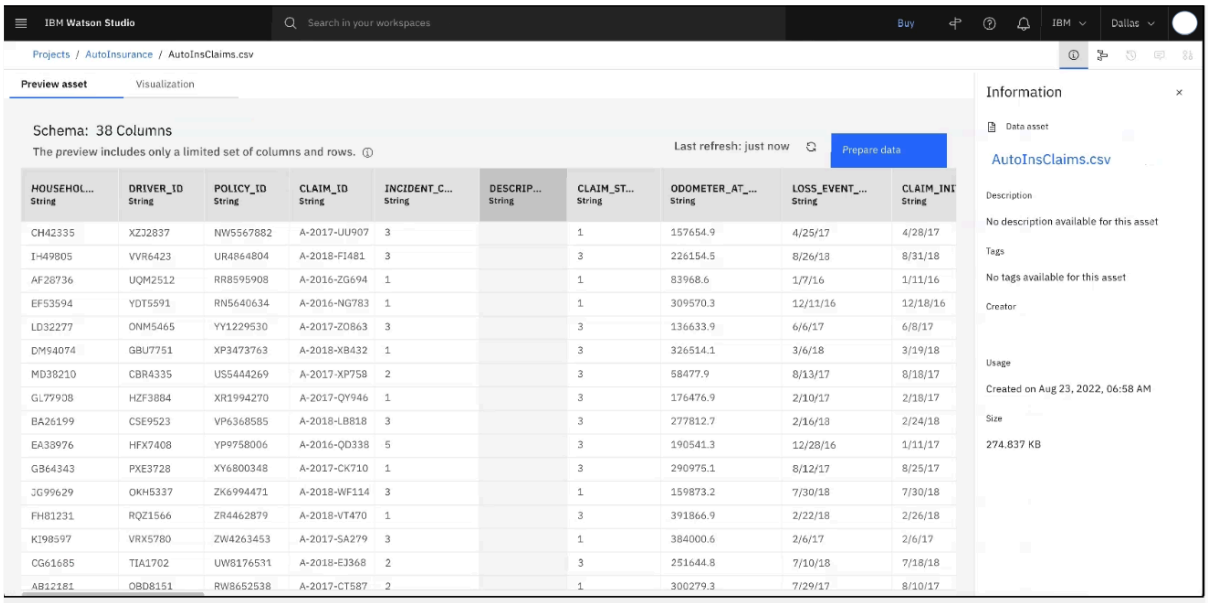
Step: 15/17

After a few moments, the data will upload. On the left, select **All assets.**



Step: 16/17

Notice your CSV file is right there for you. Hover over and select "**AutoInsClaims.csv**".



HOUSEHOL... String	DRIVER_ID String	POLICY_ID String	CLAIM_ID String	INCIDENT_C... String	DESCRIP... String	CLAIM_ST... String	ODOMETER_AT_... String	LOSS_EVENT_... String	CLAIM_INI String
CH42335	XZJ2837	NW5567882	A-2017-UU907	3		1	157654.9	4/25/17	4/28/17
I-49805	VVR6423	UR4864804	A-2018-FI481	3		3	226154.5	8/26/18	8/31/18
AF28736	UQM2512	RR8595908	A-2016-ZG694	1		1	83968.6	1/7/16	1/11/16
EF53594	YDT5591	RN5640634	A-2016-NG783	1		1	309570.3	12/11/16	12/18/16
LD32277	ONM5465	YY1229530	A-2017-ZO863	3		3	136633.9	6/6/17	6/8/17
DM94074	GBU7751	XP3473763	A-2018-XB432	1		3	326514.1	3/6/18	3/19/18
MD38210	CBR4335	US5444269	A-2017-XP758	2		3	58477.9	8/13/17	8/18/17
GL77908	HZF3884	XR1994270	A-2017-QY946	1		3	176476.9	2/10/17	2/18/17
BA26199	CSE9523	VP6368585	A-2018-LB818	3		3	277812.7	2/16/18	2/24/18
EA38976	HFX7408	YP9758006	A-2016-QD338	5		3	190541.3	12/28/16	1/11/17
GB64343	PXE3728	XY6800348	A-2017-CK710	1		3	290975.1	8/12/17	8/25/17
JG99629	OKH5337	ZK6994471	A-2018-WF114	3		1	159873.2	7/30/18	7/30/18
FH81231	RQZ1566	ZR4462879	A-2018-VT470	1		3	391866.9	2/22/18	2/26/18
KJ98597	VRX5780	ZW4263453	A-2017-SA279	3		1	384000.6	2/6/17	2/6/17
CG61685	TIA1702	UW8176531	A-2018-EJ368	2		3	251644.8	7/10/18	7/18/18
AB12181	OBD8151	RW8652538	A-2017-CT587	2		1	300279.3	7/29/17	8/10/17

Step: 17/17

A **Preview asset** page opens so you can now preview your data. Notice that you're working with 38 columns and many rows of data. Select the **Next arrow** to continue.

Simulation Transcript: Start a Project and Upload Data

Follow these steps to start your project in **IBM Watson Studio** by provisioning the service, creating a new project, and uploading a data set.

Step-by-Step Instructions:

1. **Access the IBM Cloud Dashboard**
    - Log into IBM Cloud to access the Dashboard. This dashboard allows you to manage tools, services, and resources in your IBM Cloud account.
    - Select **Catalog** from the top toolbar.
2. **Open the AI / Machine Learning Category**
    - In the catalog, locate the left-side menu under **Category**.
    - Select the **AI / Machine Learning** category.
3. **Choose Watson Studio**
    - Browse the list of AI and machine learning products.
    - Find and select the **Watson Studio** block.
4. **Provision Watson Studio**
    - Choose a location for your service. For this simulation, select **Dallas (us-south)** from the drop-down menu under **Select a location**.
    - Using a free account, there's no need to review a pricing plan. Scroll down and select the **Next** arrow.

5. **Configure Your Service**

    - In the **Configure your resource** section, an automatically generated service name will appear.
    - Update the **Service name** field to “Watson Studio-AutoIns.”
    - Leave other fields blank. Agree to the license agreements, then select **Create**.
6. **Launch Watson Studio**

    - Once the Watson Studio service is provisioned, the service name you provided will display in the upper left.
    - Select **Launch in IBM Cloud Pak for Data**.
7. **Create a New Project**

    - On the Projects page, select the **+** icon next to **Projects** and choose **Create an empty project**.
8. **Name Your Project**

    - In the **Define details** section, enter “AutoInsurance” in the **Name** field. Press Enter.
9. **Check Storage Provisioning**

    - Under **Storage**, IBM Watson Studio automatically provisions Cloud Object Storage (COS) to store unstructured data like images and text.
    - Select **Create** to finalize the project creation.

10. **View the Project Dashboard**

    - The **IBM Watson Studio Projects dashboard** will open, showing the project name at the top under the **Overview** tab.
    - Navigate to the **Assets** tab.

11. **Upload Your Data Set**

    - Select **New asset** to open your local drive and locate the **AutoInsClaims.csv** file.

12. **Drag and Drop Your File**

    - Drag and drop the **AutoInsClaims.csv** file into the **Drop data files or browse for files to upload** section under **Data in this project**.

13. **Verify Upload**

    - After a brief upload process, your data will be available.
    - On the left, select **All assets**.

14. **Preview the Data File**

    - Locate and hover over the **AutoInsClaims.csv** file.
    - Select it to open the **Preview asset** page.

15. **Inspect Your Data**

    - Preview the data set, which contains 38 columns and multiple rows of data.
    - Select the **Next** arrow to proceed.

Completion

You have successfully:

- Provisioned the IBM Watson Studio service.
- Set up a new project.
- Uploaded the data set.

You’re now ready to begin working with your project.

Good Work So Far!

Achievements So Far

Congratulations on your progress! Here's what you've accomplished:

1. **Provisioned IBM Watson Studio**
    - Set up Watson Studio as a service to support your data science project.
2. **Created a New Project**
    - Initiated a project in Watson Studio named "AutoInsurance."
3. **Uploaded the Data Set**
    - Successfully imported the **AutoInsClaims.csv** file containing auto insurance data from your company's Fraud Claims Department.

Next Steps

You're now ready to move forward and:

- **Understand and clean the data** to prepare it for analysis.

But before diving in, take a short quiz to reinforce what you've learned so far.

Understand and Clean the Data

Introduction

In this module, you will gain hands-on practice using a simulation to understand and clean data in **IBM Watson Studio** with the **data refinery tool**.

Learning Objectives

By the end of this module, you will be able to:

- **Clean data** using the **data refinery tool** in IBM Watson Studio.

Simulation: Understand and Clean the Data

Overview

Understanding and cleaning your data is an essential process to ensure accurate and meaningful results in your project. This involves:

- **Removing "noisy" data**: Irrelevant or unnecessary information that doesn't add value to your analysis.
- **Retaining key data**: Valuable information that can assist in predicting fraud for your company.

Objective

In this simulation, you will use **IBM Watson Studio** and the **data refinery tool** to:

- Gain insights into the data set.
- Clean the data to prepare it for analysis.

Details

- **Steps to Complete**: 38
- **Tool Used**: IBM Watson Studio with the data refinery tool.

By the end of this simulation, you will have practical experience in preparing a data set for analysis by eliminating irrelevant data and focusing on valuable information.

Simulation Transcript: Understand and Clean the Data

Overview

Understanding and cleaning your data is essential for accurate and meaningful results. In this simulation, you will:

- Remove irrelevant or “noisy” data.
- Retain useful data to predict fraud for your company.

Steps

1. **Upload the Data**: Preview your uploaded CSV file. Start exploring it or select **Prepare Data** to begin refining.

2. **Data Refinery Tool**: Wait for the data refinery tool to load. Once loaded, your dataset becomes editable. Familiarize yourself with the **Data**, **Profile**, and **Visualizations** tabs.

3. **Identify Columns**: The first four columns (“HOUSEHOLD_ID,” “DRIVER_ID,” “POLICY_ID,” “CLAIM_ID”) are identifiers and not predictive. Remove them by selecting the three dots next to each column and choosing **Remove**

Column.

4. **Profile Tab**: Use the Profile tab to analyze descriptive statistics and assess column relevance.

5. **Delete Columns**: Remove columns that are not useful, such as those with unique occurrences or missing data.

6. **Examine Incident Causes**: Investigate the “INCIDENT_CAUSE” column using the Profile tab. Understand the standard deviation to evaluate data dispersion.

7. **Handle Missing Data**: Remove columns like “DESCRIPTION” that have blank entries.

8. **Keep Relevant Data**: Retain columns such as “CLAIM_STATUS,” “ODOMETER_AT_LOSS,” “LOSS_EVENT_TIME,” and “CLAIM_INIT_TIME” as they hold predictive value.

9. **Curiosity in Data Science**: Recognize that date and time data can reveal patterns. Keep “LOSS_EVENT_TIME” and “CLAIM_INIT_TIME.”

10. **Save Often**: Regularly save your work using the Save icon.

11. **Police Report Column**: Retain the “POLICE_REPORT” column as it indicates whether a report was filed (0 for no, 1 for yes).

12. **Claim Amount**: Keep “CLAIM_AMOUNT” as it provides cost details of the auto accident.

13. **Fraud Flags**: Retain the “FLAG_FOR_FRAUD_INV” column, as it is essential for training a supervised model.

14. **Personal Information**: Remove columns like “FIRST_NAME” and “LAST_NAME” that contain personal information, as they do not predict fraud.

15. **Data Types**: Verify column data types. For example:

    - Ensure “ODOMETER_AT_LOSS” is Decimal.
    - Convert “LOSS_EVENT_TIME” from String to Date for proper calculations.

16. **Save and Create a Job**:

    - Save your work and create a refinery job.
    - Name the job (“Predicting fraud in auto insurance”).
    - Accept default configurations, leave scheduling and notifications off.
    - Create and run the job.

17. **Review Outputs**: The tool creates a new CSV file with “_shaped” appended to the file name. This file will be your refined dataset.

18. **Complete the Process**: You have successfully:

    - Removed irrelevant and missing data.
    - Corrected data types for accurate calculations.
    - Created a clean dataset ready for analysis.

Data Cleaning and Understanding Summary

Good work so far!

The steps you just completed to understand and clean your dataset took time. It's an important and time-consuming step in a data science project. Data analysts and data scientists spend a lot of time on this type of work.

You've gained valuable practice in this process!

Key Actions Completed:

1. **Reviewed the Data Using the Data Tab:**
    - Explored various columns to better understand the dataset in the context of researching your company's business problem about fraudulent claims.
2. **Selected Relevant Columns:**
    - Identified and retained columns with predictive data that could help detect fraud, such as the incident cause and columns with time and date information.
3. **Removed Non-Predictive Columns:**
    - Excluded columns that lacked predictive data, such as those:
        - With missing values.
        - Containing identification numbers.
        - Holding personal information.

4. **Utilized the Profile Tab:**

    - Analyzed descriptive statistics provided by the data refinery tool, including standard deviation.
    - Used these insights to decide which columns were interesting and likely to contain predictive data worth keeping.
5. **Corrected Data Types:**

    - Adjusted the data type of a column to:
        - Enable accurate calculations.
        - Ensure the dataset's quality and suitability for the project.

Next Steps:

Now that you've laid a solid foundation by understanding and cleaning your data, you're ready to refine it further. But first, take a short quiz to consolidate your learning!