Module-3

Applications and Careers in Data Science

In the first lesson, you learn about the power of data science applications and how organizations leverage this power to drive business goals, improve efficiency, make predictions, and even save lives. You also reviewed the process you will follow as a data scientist to help your organisation accomplish these ends. In the second lesson, you investigate what companies seek in a competent, experienced data scientist. You will learn how to position yourself to get hired as a data scientist. Amidst the diverse backgrounds of data scientists, you identify the qualities they share and skills that consistently set them apart from other data-related roles. You will complete a peer-reviewed final project by looking at a job posting for a data scientist and identifying commonalities between the job and what you learned in this course. You will also walk through a case study, where you learn about Sarah and her data science journey.

Learning Objectives

- Describe the contents of a data science job posting.
- Describe the application of data science in healthcare.
- Explain how companies can start on their data science journey.
- Describe how consumers generate data.
- Describe how businesses such as Netflix, Amazon, UPS, Google, and Apple use data generated by their consumers and employees.
- Compare some of the qualities that differentiate data scientists from qualities of other data professionals.
- Articulate the purpose of the final deliverable of a data science project and the role of storytelling in the final deliverable.
- Describe what the final report of a Data Science project should cover and how it should be structured for best results.
- Demonstrate your understanding of data science by articulating what data scientists do and what a data science report contains.

Data Science Application Domains

Lesson Overview: Data Science Application Domains

In this lesson, "Data Science Application Domains," you'll embark on a journey to explore the vast and impactful realms where data science plays a pivotal role. This engaging module delves into various activities that shed light on the diverse applications of data science today. You'll uncover how data science drives innovation and transformation across different sectors, from revolutionizing industries to saving lives.

Dive into this lesson to discover the real-world applications that define the dynamic landscape of data science.

Asset name and type	Description
“How Should Companies Get Started in Data Science?” video	Gain insights into how organizations can embark on their data science journey effectively.
“Old Problems with New Data Science Solutions” video	Discover how data science offers innovative solutions to age-old and real-world problems.
“Applications of Data Science” video	Explore the wide-ranging applications of data science across various industries and sectors.
“How Data Science is saving lives” video	In this video, you will learn about the life-saving potential of data science in healthcare and beyond.
“The Final Deliverable” reading	Dive into the details of what constitutes the final deliverable in data science projects.
Practice quiz	Test your understanding of the previous reading.
“Lesson Summary” video	Recap the essential takeaways from this module with a lesson summary.
Practice quiz	Take a practice quiz to evaluate how well you’ve understood the material presented in this lesson.
Glossary	Use this glossary of terms to review the terminology presented in this lesson.

How Should Companies Get Started in Data Science?

📊 Importance of Measurement in Business

- **Core Principle**:
  - Businesses cannot improve what they don’t measure.
  - Measurement is fundamental to both cost reduction and profit maximization.
- **Cost Measurement**:
  - Companies must record detailed cost data:
    - Labor costs
    - Material costs
    - Cost per product
    - Total operational cost
- **Revenue Analysis**:
  - Evaluate revenue sources:
    - Identify major contributors (e.g., does 80% of revenue come from 20% of customers?)
    - Use Pareto analysis to understand customer value distribution.

🧮 Data Collection & Archiving

- **Initial Steps**:
  - Start capturing data immediately.
  - If data collection has already begun:
    - Ensure consistent archiving
    - Never overwrite past data.
- **Value of Old Data**:
  - Historical data retains value indefinitely.
    - Data from 100 or 200 years ago can still be relevant.
  - Helps in long-term business analysis and trend prediction

- **Archiving Best Practices**:
    - Maintain consistency and structure in data formats.
    - Document everything clearly for future usability.
        - Important for someone reviewing data 20 years later

---

📈 Transition to Analytics

- **Preconditions for Data Science**:
    - Quality data is the foundation.
    - Companies often jump to advanced analytics without proper data collection.
- **Rule of Thumb**:
    - "Garbage in, garbage out"
        - Poor data leads to poor analysis and insight.s
- **Data Cleaning**:
    - If data exists, begin cleaning and organising it.
    - If data does not exist, begin collection and proper documentation.

🧠 Building the Right Data Science Team

- **Team Composition**:
    - Not just one data scientist, but a team with diverse strengths
    - Each member should specialise in different areas of data science.
- **Employee Engagement**:
    - Encourage interest in data science within the organisation.
    - Employee enthusiasm leads to better engagement and results.
- **Culture of Data**:
    - Foster a company-wide culture that values data.
    - Make data science part of the company’s strategic foundation.

Summary

This video emphasizes that the cornerstone of business improvement lies in effective measurement and data collection. Companies must start by rigorously capturing and categorizing their cost and revenue data. Old data should never be discarded, as it holds enduring value. Only after establishing solid data practices should businesses explore data science and analytics. Success in this field depends not only on the quality of data but also on the quality of the team interpreting it. A committed, skilled team and a culture that values data are essential for leveraging insights and driving business growth.

Table: What We Learnt in the Video

Topic/Section	What We Learnt
Importance of Measurement	Businesses must measure costs and revenues to improve and grow
Cost Data Breakdown	Record labour, material, per-product, and total operational costs
Revenue Sources	Analyze customer contributions (e.g., 80/20 rule)
Data Collection	Start collecting data now; don't delay
Data Archiving	Archive all data consistently; never overwrite old data
Value of Historical Data	Even 100–200-year-old data remains relevant
Documentation	Maintain detailed records and formats for future understanding
Transition to Analytics	Data science only works if clean, accurate data is available
Rule: Garbage In, Garbage Out	Analysis quality depends on input data quality
Building a Data Science Team	Form teams with diverse strengths, not just one data scientist
Employee Engagement	Cultivate interest in data science across the company
Culture of Data	Create a culture that respects, values, and utilizes data

Old Problems, New Data Science Solutions

📊 The Role of Data Science in Organizations

- **Purpose of Data Science in Organizations**
    - Used to discover optimal solutions to existing problems.
    - Leverages the vast amount of available data.
    - Applied across industries for innovation and efficiency.

🚗 Transport Industry Applications

Uber: Real-Time Demand and Supply Management

- **How Uber Uses Data Science:**
  - Collects real-time user data.
  - Determines the number of available drivers.
  - Identifies demand to activate surge pricing.
  - Ensures:
    - A right number of drivers.
    - Right place.
    - Right time.
    - Cost is acceptable to riders.

Toronto Transportation Commission (TTC): Improving Urban Traffic Flow

- **Problem:** Inefficient traffic flow is causing commuter delays.
- **Data Science Applications by TTC:**
  - Collected data on streetcar operations.
  - Identified intervention areas.
  - Analyzed customer complaint data.
  - Probe data was used to evaluate traffic on main routes.
  - Formed a specialized team to use big data for:
    - Planning.
    - Operations.
    - Evaluation.
- **Impact:**
  - Focused on peak-hour traffic clearance.
  - Identified high congestion routes.
  - Reduced commuter hours lost to congestion from **4.75 hours (2010)** to **3 hours (2014)**.

🌎 Environmental Science Applications

Tackling Harmful Cyanobacterial Blooms in Lakes

- **Problem:** Cyanobacteria threaten lake ecosystems and public health.
- **Project Overview:**
  - Conducted across the U.S. East Coast (Maine to South Carolina).
  - Involves interdisciplinary teams of scientists.
- **Data Collection Tools:**
  - Robotic boats.
  - Buoys.
  - Drones with cameras.
- **Collected Data:**
  - Physical.
  - Chemical.
  - Biological.
- **Outcomes:**
  - Creation of algorithmic models.
  - Better prediction of harmful algal bloom events.
  - Enables **proactive approaches** to:
    - Protect public health.
    - Maintain ecological balance.
- **Significance:**
  - Supports lakes used for drinking water and recreation.
  - Prepares next-generation scientists through interdisciplinary training.

🛠️ Steps to Achieve Efficient Data-Driven Solutions

1. **Identify the Problem**
   - Understand the issue clearly.
2. **Gather Data for Analysis**
   - Use relevant sources and data types.
3. **Identify the Right Tools**

    - Select appropriate analytical tools or platforms.

4. **Develop a Data Strategy**
    - Align with business goals and define the process.
5. **Leverage Case Studies**
    - Customize and refine potential solutions.
6. **Build a Machine Learning Model**
    - Apply after meeting all foundational steps.
    - Continuous refinement is needed to reach best practices.

💡 Summary

This video emphasizes how data science is fundamentally about solving real-world problems by leveraging massive amounts of data. It presents three distinct case studies—Uber's dynamic ride supply system, the Toronto Transportation Commission's congestion management, and a U.S. research effort combating harmful lake algae. These examples show how data collection, analysis, and modelling lead to tangible improvements across transportation and environmental systems. Organisations must develop a strong understanding of problems, use the right tools, and build strategic, data-driven models to unlock lasting, scalable solutions.

📘 Table: What We Learnt in the Video

Topic/Section	What We Learnt
Purpose of Data Science	Used to find optimal solutions to persistent problems across industries.
Uber Case Study	Data is used in real-time to effectively match driver supply with rider demand.
Toronto Traffic Management	Data science reduced commuter congestion time through strategic traffic analysis.
Environmental Case Study	Monitoring cyanobacteria with drones/robots enables proactive ecological protection.
Data-Driven Solution Strategy	Steps: Identify problem → Gather data → Use the right tools → Build a strategy/model.
Importance of Interdisciplinary Work	Combines environmental science, engineering, and data analytics to solve complex issues.

Applications of Data Science

📊 The Impact of Data Science and Big Data on Business

- **General Impact**:
  - Transforming day-to-day operations.
  - Improving financial analytics.
  - Enhancing customer interactions.
- **Business Value**:
  - Provides deep insights into consumer behaviour.
  - Drives operational efficiency and competitiveness.

🌐 Data Generation by Consumers

- **Everyday Data Creation**:
  - Consumers generate large volumes of data daily, often unknowingly.
  - This is referred to as a *digital trace*.
  - Online behaviours, transactions, and searches leave data footprints.

🎯 Application Example: Recommendation Engines

- **Platforms Using Recommendation Engines**:
  - Amazon, Netflix, Spotify.
- **How They Work**:
  - Algorithms analyze:
    - Customer preferences.
    - Historical behaviour.
  - Output:

        - Personalized product, show, or music recommendations.

🧠 Application Example: Personal Assistants

- **Siri (Apple Devices)**:
    - Uses data science to interpret and respond to various user questions.
- **Functionality**:
    - Voice recognition.
    - Natural language processing.
    - Query understanding and response generation.

📍 Location-Based Services

- **Google**:
    - Monitors:
        - Online shopping habits.
        - Social media activity.
        - Physical movements.
    - Uses this data to:
        - Recommend nearby restaurants, bars, shops, and attractions.
        - Tailor results based on user preferences and location.

⌚ Wearable Technology & Health Data

- **Devices**:
    - Fitbit, Apple Watch, Android Watches.
- **Data Collected**:
    - Physical activity.
    - Sleep patterns.
    - Heart rate.

- **Purpose**:
  - Enhance personal health tracking.
  - Feed into broader consumer data ecosystems.

🏢 Data Science in Business Operations

- **Historical Prediction (McKinsey, 2011)**:
  - Data science is becoming central to competition, productivity, and innovation.
- **UPS Case Study (2013)**:
  - Data from drivers, vehicles, and customers is used for:
    - Route optimization.
    - Saving time, money, and fuel.

🏆 Gaining Competitive Advantage: Netflix Case Study

- **Data Collection**:
  - Tracks:
    - Viewing times.
    - Pause/rewind/fast-forward behaviour.
    - Searches for actors and directors.
- **Predictive Capabilities**:
  - Combines user preferences to predict successful content.
  - Data-driven investment decisions before production.
- **Example**: *House of Cards*
  - Insights:
    - Users enjoyed David Fincher’s work.
    - Robin Wright’s films had strong viewership.
    - The British version of *House of Cards* was popular.
    - An overlap was found between Fincher fans and Wright fans.

        - Conclusion:
            - Strong potential for U.S. version.
            - Netflix invested → The show became a huge success.

📝 Summary

The video highlights the transformative impact of data science and big data on businesses and consumers. From personalized recommendation engines and intelligent virtual assistants to optimized logistics and strategic content production, data science is driving more thoughtful decisions and competitive advantages. Examples like Amazon, Google, UPS, and especially Netflix illustrate how data-driven insights can forecast consumer behaviour and guide investments. The key takeaway is that leveraging massive data streams allows businesses to understand their customers better and innovate and compete more effectively.

📚 Table: What We Learnt in the Video

Topic/Section	**What We Learnt**
Data Science in Business	Enables better operations, finance, and customer interaction.
Digital Trace	Consumers generate large volumes of data unknowingly.
Recommendation Engines	Use algorithms to offer personalized suggestions (e.g., Amazon, Netflix).
Personal Assistants	Siri uses data science to answer voice-based questions.
Google's Data Use	Analyses behaviour and location to recommend places and services.
Wearables	Track health metrics and add to personal data profiles.
McKinsey Prediction (2011)	Data science is a key competitive factor.
UPS Data System (2013)	Used data to optimize delivery routes and save resources.
Netflix Case Study	Uses user behaviour data to predict content success (e.g., *House of Cards*).

How Data Science is Saving Lives

🧠 Introduction to Data Science and Its Impact

- Data science plays a significant role in improving human lives.
- Applications include:
    - Healthcare: helping professionals offer the best treatment.
    - Disaster Preparedness: predicting natural disasters to save lives.

🏥 Data Science in Healthcare

🔍 Predictive Analytics

- Uses tools such as:
  - Data mining
  - Data modeling
  - Statistics
  - Machine learning
- Purpose:
  - To identify best options for patients.
  - To examine known disease factors:
    - Gene markers
    - Associated conditions
    - Environmental factors
  - Recommends:
    - Appropriate medical tests
    - Clinical trials
    - Tailored treatments

👨‍⚕️ Equalizing Physician Knowledge

- Physicians have varying personal knowledge bases.
- Predictive analytics ensures:
  - Equal access to up-to-date disease information.
  - Consistent treatment plans.
  - Improved patient outcomes.

📊 Case Study: Cancer Diagnosis and Gene Markers

- Conducted by:
  - Boston Consulting Group
  - AdvaMedDx
- Findings:
  - Main factor in test offering: the oncologist’s knowledge.

    - Not all doctors are aware of specific gene-related tests.
- Implication:
    - Data science tools help bridge knowledge gaps.

💾 EMRs and Medical Research

- Source: Electronic Medical Records (EMRs)
- Example:
    - NorthShore University HealthSystem (Chicago)
        - First in the U.S. to achieve top-level EMR deployment.
        - Recognized for both inpatient and outpatient care.
        - Offers guidance on data mining for research.
        - Generated anonymized data for analytics.

📈 Progression of Analytics in Healthcare

- From:
    - Descriptive analytics (what happened)
- To:
    - Predictive analytics (what might happen)

🌍 Data Science in Disaster Preparedness

🚨 Predictive Tools and Applications

- Improves response and preparation for:
    - Earthquakes
    - Hurricanes & Tornadoes
    - Floods
    - Volcanic eruptions
- Large datasets used for:
    - Faster warnings
    - Saving lives

📱 Use of Social Media in Prediction

- Research from:

    - University of Warwick (UK)
- Method:
    - Analyze social media (photos, keywords)
    - Combine with data from scientists and weather stations
- Benefit:
    - Enhances localized weather event predictions

🎓 Educational Response

- Universities integrating data science into curricula
- Example:
    - University of Chicago Graham School
        - Master of Science in Threat and Response Management

---

🛠️ Capabilities of Data Science Tools

- Analyze massive, diverse datasets
- Provide clear, actionable insights.
- Potentially save hundreds of lives by improving decision-making.

🧾 Summary

The video emphasises how data science transforms critical sectors like healthcare and disaster preparedness. In healthcare, predictive analytics helps doctors choose the best treatment options by analyzing a wide range of patient-specific factors, ensuring all physicians can access the latest knowledge. A case study illustrates how a lack of awareness among oncologists affects patient testing. Institutions like NorthShore University are leading the way in using electronic medical records for research. In disaster management, predictive tools and social media analytics enhance warning systems. Education is adapting to these changes, with universities offering specialised data science programs. Ultimately, data science is essential in saving lives and improving outcomes across disciplines.

📘 Table: What We Learnt in the Video

Topic/Section	What We Learnt
Impact of Data Science	Enhances healthcare and disaster preparedness, with broad human benefits.
Predictive Analytics in Healthcare	Analyzes disease factors to recommend personalized tests and treatments.
Physician Knowledge Equality	Predictive systems ensure all doctors access the same up-to-date info.
Cancer Gene Marker Case Study	Patient care is impacted by the physician’s awareness of specific diagnostic tools.

EMRs and NorthShore University	EMR data is used for research; NorthShore leads the implementation.
Analytics Progression	Healthcare is evolving from descriptive to predictive analytics.
Disaster Prediction	Data science predicts natural disasters, enabling earlier alerts.
Social Media Use in Weather Tracking	Social data helps track and predict weather events in real time.
Education in Data Science	Schools now teach data science applications in threat and response management.
Data Science Capabilities	Handles large data sets for life-saving insights and decisions.

The Final Deliverable

Sure! Below is a structured, comprehensive set of study notes based on the excerpt from **Chapter 3, pages 52–53** of *Getting Started with Data Science* by **Murtaza Haider**. These notes focus on the importance of the **final deliverable** in analytics, using the **Deloitte U.S. Economic Forecast Report** as a prime example.

🔍 Purpose of the Final Deliverable

- **Goal of Analytics**: Communicate insights to policymakers, strategists, or decision-makers.
- **Means of Communication**:
  - Tables and plots summarize findings.
  - Narratives **interpret and explain** the findings to give them meaning.
- **Academic Deliverables**:
  - Essays and reports.
  - Typically 1,000–7,000 words.
- **Business/Consulting Deliverables**:
  - Short reports (< 1,500 words) with visuals.
  - Or large, comprehensive reports (hundreds of pages).
  - Purpose: Showcase expertise and provide actionable insights.

📝 Example: Deloitte's "United States Economic Forecast"

- **Nature**:
  - A 24-page report (Dec 2014).
  - Focused on the U.S. economy.

    - Example of **storytelling with data**.
- **Opening Strategy**:
    - Uses a **grabber**: Claims U.S. economy and job growth stronger than perceived.
- **Thematic Approach**:
    - Cites **Voltaire** — adds depth and authority.
    - Presents a **positive economic outlook**:
        - Growth in **manufacturing investments**.
        - **Higher consumer spending** due to **lower oil prices**.

📈 Use of Graphics and Analytics

- **GDP Growth Time Series**:
    - Shows recession and recovery.
    - Projects **four future scenarios**.
- **Consumer Spending Plot**:
    - Visualizes spending trends.
    - Narrative links to **income inequality** (cites **Thomas Piketty**).
    - Highlights that **real income** did not rise, but **spending levels stayed up**.
- **Other Areas Covered**:
    - **Housing**
    - **Business and Government sectors**
    - **International Trade**
    - **Labor and Financial Markets**
    - **Prices**
- **Appendix**:
    - Four tables support the four projected scenarios.

📢 Importance of the Narrative

- **Critical Role**:

    - Narrative links data to broader themes and context.
    - Makes the report **persuasive and memorable**.
- **Hypothetical Comparison**:
    - If presented only as PowerPoint slides (with visuals but no story):
        - Would **fail to convey depth and meaning**.
        - Loss of literary references, context, and message.

🛠️ Planning the Final Deliverable

- **Work Backwards Strategy**:
    - Start with a **clear idea** of the report's **purpose and key message**.
    - Identify the **required data** and **analytical methods** to support it.
- **Warning**:
    - Jumping into analytics without planning the deliverable leads to:
        - A disjointed, ineffective report.
        - Poor integration between narrative and data.

📌 Summary

This section from *Getting Started with Data Science* emphasises crafting a meaningful final deliverable in analytics. Using Deloitte's *United States Economic Forecast* as an example, it shows how a combination of data visualization and strong narrative can effectively convey insights. The narrative, supported by visuals and references, helps build a compelling argument. Analysts must plan the final message before analysing to ensure the narrative and analytics are aligned, persuasive, and impactful.

📊 Table: What We Learnt

Topic/Section	**What We Learnt**
Purpose of Final Deliverable	To communicate data-driven insights through a blend of visuals and narrative.
Academic vs. Business Reports	Academic: Long essays; Business: Short or extensive documents with graphics.
Deloitte Report Overview	A 24-page thematic report focused on economic optimism and analytical rigour.

Use of Storytelling	References (Voltaire, Piketty) and themes help strengthen the message.
Graphics and Tables	Used to illustrate trends in GDP, spending, housing, trade, etc.
Four Economic Scenarios	Visualised and tabulated in the appendix to show possible future outcomes.
Narrative Importance	Adds depth, context, and persuasion beyond what visuals alone can achieve.
Planning Backwards	Start with a message, find supporting data and analysis to create a cohesive report.

Careers and Recruiting in Data Science

Lesson Overview: Careers and Recruiting in Data Science

In the lesson "Data Science Application Domains," you'll embark on a journey to explore the vast and impactful realms where data science plays a pivotal role. This engaging module delves into various activities that shed light on the diverse applications of data science today.

Dive into this lesson to discover the real-world applications that define the dynamic landscape of data science.

Asset name and type	Description
“How Can Someone Become a Data Scientist?” video	Discover the pathways to becoming a proficient data scientist, exploring the skills and knowledge required.
“Recruiting for Data Science” video	Gain insights into organisations' strategies and considerations for recruiting data science talent.
“Careers in Data Science” video	Explore the diverse career opportunities and roles available in the dynamic field of data science.
“Importance of Mathematics and Statistics for Data Science” video	Understand the fundamental role of mathematics and statistics in data science, emphasizing their significance.
“The Report Structure” reading	Delve into the intricacies of structuring reports within data science projects, enhancing your understanding of this essential aspect.
Practice quiz	Test your understanding of the previous reading.
“Infograph on Roadmap” reading	Explore an informative infographic detailing the roadmap for success in data science careers.

“Lesson Summary” video	Recap the essential takeaways from this module with a lesson summary.
Practice quiz	Take a practice quiz to evaluate how well you’ve understood the material presented in this lesson.
Glossary	Use this glossary of terms to review the terminology presented in this lesson.
Grade Quiz	Evaluate your knowledge of data science in a business setting with this graded quiz.
“Data Science in Business” Reading	Summarize your learning journey in this module, reviewing the key takeaways and insights gained.

How Can Someone Become a Data Scientist?

🎓 Skills Required for High-End Data Scientists

- **Educational Background:**
  - Most high-end data scientists are PhDs.
  - Common fields: Physics, Statistics, Mathematics, and Computer Science.
- **Essential Knowledge Areas:**
  - Computer Science
  - Mathematics
  - Databases
  - Probability and Statistics
- **Statistical Knowledge:**
  - Understanding different statistical distributions
  - Basic and advanced probability and statistics concepts

🧠 Skills for Entry-Level Data Scientists

- **Foundational Skills:**
  - Programming, or at least computational thinking
  - Introductory algebra and analytical geometry
  - Calculus (basic level)
  - Basic statistics and probability

    - Understanding of relational databases
- **Starting Point:**
    - Learn relational databases:
        - Core to understanding how data is stored and queried
        - Can be applied to significant data clusters
    - No need to fully understand MapReduce at the early stages

---

🔁 Self-Learning & Practical Experience

- **Learning by Doing:**
    - Importance of hands-on practice: Building and experimenting with technology
    - Example: Built an HPC (High Performance Computing) cluster after researching the Beowulf cluster concept
- **Tools for Learning:**
    - Online learning platforms
    - IPython & Jupyter Notebooks
    - Apache Zeppelin
- **Effective Learning Techniques:**
    - Interactive experimentation with code
    - Visualization and exploration to grasp complex ideas
- **Key Driver: Motivation**
    - Challenge: Maintaining learner motivation
    - Solution: Badge systems (e.g., Big Data University) help encourage progress
    - An individual must set goals and stay committed.

🏢 Organizational Placement of Data Science

- **Where Data Science Should NOT Sit:**
    - Under the Chief Information Officer (CIO)
        - Many CIOs come from accounting or finance.

        - Often lack technical expertise in data science.

- **Ideal Placement:**
    - Research departments
    - Companies with a research agenda:
        - Pharmaceuticals
        - Finance
        - Technology companies
- **Industry Demand:**
    - High demand for Phd-level data scientists
    - Top companies hiring:
        - Facebook
        - LinkedIn
        - Uber
        - Lyft
    - Attractive salaries and cutting-edge problems (e.g., optimizing Uber car schedules)

Summary

This video highlights the structured skill set required to become a data scientist, especially at the PhD level. It stresses the interdisciplinary foundation in math, statistics, computer science, and databases. It is recommended that beginners start with programming and understanding relational databases. Self-learning through practice, building projects, and using interactive platforms like Jupyter is emphasized. Motivation is a key factor in learning, and gamification, like badges, can help. Organizationally, data science functions best in research-driven environments rather than under traditional IT leadership like CIOs. The demand for skilled data scientists, particularly those with PhDs, is surging across tech and research sectors.

Table: What We Learnt in the Video

Topic/Section	**What We Learnt**
High-End Data Scientists	Often PhDs with backgrounds in physics, stats, math, and CS; require deep expertise
Entry-Level Skills	Must know programming, algebra, basic calculus, probability, stats, and databases
Relational Databases	Easy starting point; foundational for understanding big data
Self-Learning Approach	Learn by doing; build projects; experiment and explore

Learning Tools	Jupyter, IPython, and Zeppelin enable interactive learning
Motivation Strategies	Key to progress: badge systems help; self-driven goals are necessary
Role of CIOs	Not ideal leaders for data science; often lack technical depth
Ideal Team Placement	Data science should align with research teams
Industry Demand	Phd-level data scientists are in high demand at top tech companies
Real-World Problems	Example: Uber's scheduling using massive data and optimisation

Recruiting for Data Science

🔍 Understanding the Hiring Challenge in Data Science

- **The Myth of the "Unicorn" Data Scientist**
    - Companies tend to look for individuals with:
        - Deep domain knowledge
        - Proficiency in structured & unstructured data analysis
        - Excellent storytelling and presentation skills
    - Such candidates (having all skills) are rare, often referred to as *unicorns*.
    - Necessary to recognize the improbability of finding a perfect candidate
- **Hiring Strategy**
    - Focus on alignment with the company's **DNA** and **mission.**
    - **Trainability** is key — analytics skills can be taught.
    - Core attributes to prioritize:
        - Passion for the business domain
        - Curiosity
        - Sense of humour
        - Communication & storytelling skills

🧠 Key Traits to Look for in Candidates

- **1. Curiosity**

    - General curiosity about surroundings, not just data
    - Interest in understanding the "why" behind things
- **2. Sense of Humor**
    - Keeps work lighthearted
    - Encourages creativity and resilience
- **3. Social & Communication Skills**
    - Ability to collaborate across departments
    - Being relatable and approachable
- **4. Storytelling Ability**
    - Capacity to identify stories in data
    - Convey insights in a compelling and engaging way

🛠️ Technical Skills and Role Clarity

- **Importance of Technical Skills**
    - Should be considered **after** evaluating soft skills
    - Necessary but **trainable**
    - Roles should define technical requirements:
        - Engineers
        - Architects
        - Visualization experts
        - Statisticians
- **Steps for Technical Skill Evaluation**
    - Determine domain focus:
        - Structured data (e.g., Market Research)
        - Unstructured data (e.g., Text, Weblogs)
        - Big Data (e.g., Logs, IoT, Health Data)
- **Examples of Technical Tools by Domain**

Domain Type	Tools & Platforms

Structured Data	R, Stata, Python
Unstructured Data	Python is preferred over R
Big Data	Hadoop, Spark

🧩 Soft Skills & Communication

- **Communication and Presentation**
    - Equally important as technical expertise
    - Need to:
        - Present data in simple, engaging ways
        - Translate numbers into business value.
        - Help audiences feel *"Aha!"* moments.
    - Examples:
        - Large tables → simplified insights
        - Presentations that “sing” findings
        - Sharp-turn storytelling → surprise and delight

🧭 How to Build a Data Science Team

- **Understand the Business Needs First**
    - Define the role clearly before hiring.
    - Ask:
        - What are we trying to achieve with data?
        - Do we need more technical engineers or insight storytellers?
- **Grow Strategically**
    - Start with core skill sets.
    - Gradually add specialists depending on need.

Summary

The video offers a thoughtful approach to hiring for data science teams, emphasizing that companies often search for an idealised "unicorn" candidate — someone with all technical, domain, and communication skills. However, the speaker stresses that this is unrealistic. Instead, hiring should focus on intrinsic qualities like curiosity, passion for the domain, sense of humour, and storytelling ability, which are challenging to teach. Technical skills, while important, are secondary and trainable. Understanding the role and business need is vital before building a data science team. The goal is to find candidates who resonate with the company's culture and can grow into the role.

📘 Table: What We Learnt in the Video

Topic/Section	**What We Learnt**
The "Unicorn" Myth	Perfect candidates with all skills are rare; focus on teachable and intrinsic traits
Qualities to Prioritize	Curiosity, humour, communication, storytelling > technical skills
Role-Specific Hiring	Define business needs before hiring; different domains need different skillsets
Technical Tools	Tool choice depends on data type and use case (e.g., R, Python, Hadoop, Spark)
Communication Skills	The ability to simplify, present, and engage is critical for data scientists
Team Building	Start with clear goals, grow with a mix of engineers, analysts, and storytellers.

Careers in Data Science

🌐 The Rise of Data Science

- **Technological Context**
  - Growth of the **Internet of Things (IoT)** and **distributed computing**
  - Massive volumes of data are now available
  - Technological advances allow robust data analysis.
  - Data must be shaped into actionable insight.s
  - This gave rise to **Data Science** as a field.

📈 Data Science as a Career

- **Employment Trends**
  - Platforms like **LinkedIn, Glassdoor, Indeed, and Dice** show a sharp rise in data science job postings.
  - Since **2016**, Data Scientist has been among the **top career choices.**
  - In **2020**, still ranked in the **top 3 most promising careers**
  - Dice reports data science jobs are spread **across many industries**, not just tech.
- **Market Growth**

    - **Global Industry Analysts Inc.** predicts:
        - The data science platform market will grow by **USD 314.8 billion by 2025**
        - Driven by a **compound annual growth rate (CAGR)** of **38.2%**
- **Talent Shortage**
    - **McKinsey Global Institute** warned of **massive talent shortages by 2018**
    - **Forrester Research (Brandon Purcell)** in 2019:
        - Demand for data scientists will continue to grow as organisations become more **data-driven**
    - Result: **High demand** but **low supply** of skilled data scientists

---

👩‍💻 What Motivates a Data Scientist?

- **Field Versatility**
    - Applicable across **almost every industry and discipline**
    - Suitable for people who:
        - Enjoy working with data.
        - Like coding
        - Are open to learning **math and statistics**
        - Possess **storytelling** skills (to convey data insights effectively)
- **Lifelong Learning**
    - Success requires:
        - Constantly updating tools and techniques.
        - Willingness to learn new methods and platforms

👩‍🔬 Women in Data Science (WiDS)

- **WiDS Initiative**
    - Launched by the **Stanford Institute for Computational and Mathematical Engineering**
    - Aims to:
        - **Inspire and educate** data scientists globally.
        - **Support women** in the data science field.

        - Promote **diversity and inclusion.**

🎯 Getting Into Data Science

- **Career Preparation Tips**
    - Align your **skill set** with the **specific role** you are targeting.
    - Use **online education platforms** to:
        - Gain missing skills
        - Learn programming, statistics, data handling, machine learning, etc.
    - Prepare for a career that is:
        - **Fascinating**
        - **Rewarding**
        - **Continuously evolving**

📝 Summary

The video highlights the evolution of data science as a crucial and rapidly growing field, driven by the explosion of data and advancements in technology like IoT and distributed computing. Since 2016, data science has ranked among the top career paths due to its versatility and importance across industries. However, there's a significant talent shortage. Success in this field requires curiosity, coding ability, and a knack for storytelling, alongside continuous learning. Initiatives like Women in Data Science (WiDS) support diversity and skill development, and various online resources make entering the field more accessible than ever.

📊 Table: What We Learnt in the Video

Topic/Section	What We Learnt
Technological Background	IoT and distributed computing created the data boom that data science responds to
Career Trends	Data science ranks among the top 3 careers; demand spans multiple industries
Market Growth	$314.8B market value expected by 2025, with a 38.2% CAGR
Talent Gap	Severe global shortage of skilled data scientists
Motivation & Fit	Field suits coders, math lovers, storytellers; requires cross-disciplinary skills

Learning & Upskilling	Ongoing learning is essential; online resources are widely available
Women in Data Science (WiDS)	Promotes global participation and support for women in data science

Importance of Mathematics and Statistics for Data Science

🎓 Getting Started with Data Science

- **Recommended Skills to Begin:**
    - Learn how to program
    - Study some math
    - Take a course in probability
    - Learn basic statistics
- **Approach to Learning:**
    - Emphasize **hands-on experience**:
        - Build things (not necessarily physical, can be software or systems)
        - Write programs
        - Construct statistical systems

🔍 Learning Through Doing

- **Iterative Discovery:**
    - Learning what tools or concepts you need becomes clearer once you start building.
        - E.g., encountering a new concept like "inner product" leads to learning it naturally.
- **Advantage Over Time:**
    - Early exposure gives students a **head-start** in college.
    - Leads to better preparation for job markets post-college.
    - Results in financial success and **increased happiness** due to enjoyable work.

💡 Encouragement for High School Students

- **What High Schoolers Can Do Now:**
    - Get familiar with **databases** and **SQL.**
    - Take **computer science** courses if available.
    - Start thinking about **systems and software development.**

🧠 Fostering Curiosity and Creativity

- **Traits of a Good Data Scientist:**
    - Curiosity is critical — likened to:
        - Detective games 🕵️
        - Treasure hunts 🪙
    - These hobbies foster:
        - Problem-solving ability
        - Pattern recognition
        - Experimental thinking

🧪 Data Science is Modern Scientific Inquiry

- **Science Fair Analogy:**
    - Similar to science fairs, but instead of baking soda volcanoes:
        - Work with real **datasets**
        - Ask and answer meaningful questions with data
- **Election Season Example:**
    - Great opportunity to discuss:
        - How polling works
        - How predictions are made from surveys
        - Validity and interpretation of data

💼 Career Outlook for Data Scientists

- **High-Demand Profession:**

    - Data science is a **knowledge profession** in high demand globally.
    - Data scientists help businesses:
        - Make more thoughtful, more efficient decisions.
        - Solve real-world problems
    - Career is:
        - Fulfilling
        - Financially rewarding
        - Long-term sustainable

🧮 Overcoming Fear of Math

- **Math Struggles Are Common:**
    - Many successful data scientists were not strong at math initially.
    - School math may seem irrelevant or abstrac.t
- **Making Math Meaningful:**
    - When math is used to solve **real-world problems**, it becomes easier and more interesting.
    - Knowing the people who benefit from your work adds purpose and motivation.

📌 Summary

The video encourages aspiring data scientists—especially students—to begin with foundational skills in programming, math, probability, and statistics. It highlights the importance of learning by building and being curious. Real-world problems like election polling provide opportunities to understand and apply data science. High school students are advised to explore SQL and computer science and view data science as a modern form of the scientific method. Despite everyday struggles with math, relevance and context make learning easier. Data science is portrayed as a fun, impactful, and high-demand career path.

📊 Table: What We Learnt in the Video

Topic/Section	**What We Learnt**
Starting Skills	Learn programming, math, probability, and statistics before diving into projects.
Learning Approach	Build systems and solve problems to discover what tools and knowledge are needed.
Early Exposure Advantage	Early in high school can give a big head start in college and the job market.

Tools to Learn Early	In school, SQL, databases, and computer science are functional areas to focus on.
Curiosity and Creativity	Detective games and treasure hunts develop a curious and analytical mindset essential for data science.
Science Fair Analogy	Like science fairs, data science is about asking questions and learning through experimentation, using data instead of volcanoes.
Real-World Relevance	Discussing election polls helps relate data science to current events and understand predictive analytics.
Career Outlook	Data science is a highly valued profession across industries, offering rewarding and meaningful work.
Math Anxiety	Even those not strong in math can succeed; real-world relevance helps make math more understandable and valuable.

The Report Structure

📘 Report Length and Purpose

- **Brief Reports (≤5 pages)**:
    - Concise, focused summary of key findings.
    - Often commentary on current trends or issues attracting attention.
    - Suitable for fast consumption.
- **Detailed Reports (>100 pages)**:
    - Comprehensive analysis with:
        - Review of relevant works
        - Detailed research methodology
        - Data sources
        - Intermediate and final results
    - Often includes original research (data collection, expert interviews).

📑 Essential Structure of a Report

Even short reports should follow a **standard format**:

- **Cover Page**
- **Table of Contents (ToC)**
- **Executive Summary/Abstract**
- **Detailed Contents**

- **Acknowledgments**
- **References**
- **Appendices** (if needed)

🧾 Cover Page

- Frequently omitted even by advanced writers and firms.
- **Should include**:
  - Title of the report
  - Author(s) name(s)
  - Affiliation and contact info
  - Institutional publisher (if applicable)
  - Date of publication
- Importance:
  - Helps in citation.
  - Facilitates reader contact.
  - Adds professional appearance.

🗺️ Table of Contents (ToC)

- Acts like a **map** for the reader.
- Especially essential for reports ≥5 pages.
- Includes:
  - Major headings
  - List of tables and figures
- **Purpose**:
  - Helps navigate the document.
  - Prepares reader for the content journey.

📋 Executive Summary / Abstract

- Recommended **even for short documents**.

- Aim:
  - Convey core ideas in ≤3 paragraphs.
  - Serve as a standalone summary.
- In long reports:
  - Can be extended in length proportionally.

📖 Introductory Section

- Introduces the problem/topic to readers unfamiliar with it.
- Should:
  - Establish context.
  - Provide a smooth entry into complex material.

📚 Literature Review

- Reviews prior research.
- **Length depends on topic complexity**:
  - **Simple consensus** → brief section with core citations.
  - **Nuanced debates** → longer with more references.
- Helps to:
  - Identify knowledge gaps.
  - Introduce research questions and hypotheses.

⚙️ Methodology

- Describes tools, data, and methods used.
- If using original data:
  - Explain collection process in detail.
- Justify:
  - Variable choices
  - Analytical approach

📊 Results Section

- Presents findings through:

  - Descriptive statistics
  - Graphics (See Chapters 4, 5, 10)
  - Hypothesis testing (See Chapter 6)
- May use:
  - Regression (Chapter 7)
  - Categorical analysis (Chapters 8, 2)
  - Time-series (Chapter 11)
  - Data mining (Chapter 12)
- Business reports often:
  - Prioritize visuals
  - Minimize complex stats

🧩 Discussion Section

- Interprets results.
- Refers back to:
  - Research questions
  - Literature review
- Goal:
  - Build a narrative from data.
  - Fill identified knowledge gaps.
- Accept limitations:
  - Partial answers
  - Caveats

🧠 Conclusion

- Generalizes findings.
- Adopts a marketing tone:
  - Reinforces key messages.

    - Pushes significance of work.

- Suggests future research and applications.

🗂️ Housekeeping Sections

- **References**: Cite all works used.
- **Acknowledgements**: Credit contributors and supporters.
- **Appendices**: Include extra info like data, code, figures, etc.

✅ Writing Checklist (from *Transport Policy Journal*)

Authors must ensure:

1. Have you explained the benefit to the reader?
2. Is your aim clear?
3. Is your contribution's significance stated?
4. Have you provided adequate background and references?
5. Have you addressed practicality and usefulness?
6. Have you discussed possible future developments?
7. Is the structure clear and logical?

🧾 Summary

This chapter emphasises the importance of planning a report's structure before analysing. Reports should be tailored by length and purpose, from concise briefings to extensive studies. Regardless of length, every report should include formal sections such as a cover page, executive summary, introduction, methodology, results, and conclusion. The writing must be strategic, not just analytical, to effectively communicate insights. A checklist from *Transport Policy* offers valuable guidance to ensure clarity, relevance, and completeness in report writing.

📋 Table: What We Learnt in the Chapter

Topic/Section	**What We Learnt**
Report Length	Length depends on purpose; brief = summary, long = detailed research & analysis.
Standard Structure	All reports need a cover, ToC, a summary, a body, references, and acknowledgements.
Cover Page	Include title, authors, contacts, publisher, and date — often neglected.
Table of Contents	Offers roadmap for the reader; vital if report >5 pages.
Executive Summary / Abstract	Critical to summarize key findings; short yet powerful.

Introduction	Gently brings new readers into the topic.
Literature Review	Reviews existing work, identifies gaps, and supports research questions.
Methodology	Details research methods and data sources; justifies analytical choices.
Results	Presents findings using stats and visuals; varies by technique and audience.
Discussion	Builds narrative from results; revisits hypotheses and explains significance.
Conclusion	Generalizes results, emphasizes impact, and proposes future research directions.
Housekeeping	References, acknowledgements, and appendices complete the report.
Writing Checklist	Ensure aim, clarity, structure, context, significance, and usefulness are addressed.

Lesson Summary: Careers and Recruiting in Data Science

1. Overview of the Data Science Job Market

- Companies are actively recruiting data scientists.
- They often seek individuals with a wide range of skills.
- However, expecting one person to have *all* the desired skills is unrealistic.
- Instead, companies should:
  - Build diverse teams with complementary skills.
  - Hire people with potential and help them grow.

2. Qualities Companies Look for in Data Scientists

- **Passion for the Industry**
  - Enthusiasm for the business domain is vital.
  - Retail data scientists may not be suitable for healthcare or IT environments if they lack interest.
  - Passion leads to higher productivity.
- **Curiosity**
  - Essential for asking meaningful, impactful questions.
  - Curiosity drives motivation and innovation.
  - It helps data scientists remain engaged and seek deeper insights.

- **Self-Learning & Tinkering**
    - Being proactive in learning and experimenting is key.
    - Playing with data and visualizations helps develop understanding and skill.

3. Core Skills of a Data Scientist

- **Analytical & Computational Thinking**
    - Critical for deriving insights from complex data.
- **Mathematics, Statistics & Probability**
    - Fundamental for ensuring conclusions are valid and reliable.
- **Computer Programming**
    - Languages commonly used:
        - **Python**
        - **R** (especially for statistical analysis)
- **Data Management**
    - Proficiency in handling both structured and unstructured data.
    - Understanding of storage and retrieval systems is crucial.
- **Machine Learning**
    - Familiarity with standard algorithms is necessary to extract insights.
    - Knowledge of AI techniques enhances analytical capabilities.

4. Communication and Storytelling

- **Importance of Communication Skills**
    - Data scientists must explain findings clearly and persuasively.
    - Instructional and presentational skills enhance team impact.
- **Storytelling Techniques**
    - Final reports must:
        - Engage the reader with a narrative.
        - Provide value, clarity, and surprise.
        - Include goals, background, contributions, and practical implications.

- Use of analogies (e.g., driving around a mountain to discover a valley) emphasizes the emotional impact of well-communicated insights.

5. Team-Based Approach

- Companies should:
  - Avoid seeking a "super data scientist" with all skills.
  - Build balanced teams where members bring different strengths.
  - Ensure teams include:
    - Domain experts
    - Analytical minds
    - Programming specialists
    - Visualization and storytelling communicators

🧾 Summary

This lesson emphasizes that while companies often look for data scientists with broad skills, it's more practical to form diverse, well-rounded teams. Key traits such as passion for the domain, curiosity, self-learning, and strong communication are vital. In addition to technical expertise in math, programming, and data analysis, the ability to convey findings through compelling storytelling is essential. Companies should focus on finding individuals excited about the field who can contribute to a collaborative data science team.

📊 Table: What We Learnt in the Video

Topic/Section	**What We Learnt**
Job Expectations	Companies often want data scientists to have wide-ranging skills, but this is rare.
Team-Based Hiring	Teams should be built with members having complementary skills.
Industry Passion	Passion for the specific industry boosts productivity and engagement.
Curiosity	Drives exploration, better questions, and sustained motivation.

Self-Learning	Important for growth and innovation in handling data.
Core Technical Skills	Required: Math, stats, probability, Python, R, and machine learning.
Data Handling	Must know how to manage structured and unstructured data.
Communication & Storytelling	Critical for presenting insights clearly and with impact.
Effective Reports	Should have clear goals, context, significance, and engaging narrative.
Ideal Data Science Team Makeup	Should include domain experts, programmers, analysts, and communicators.

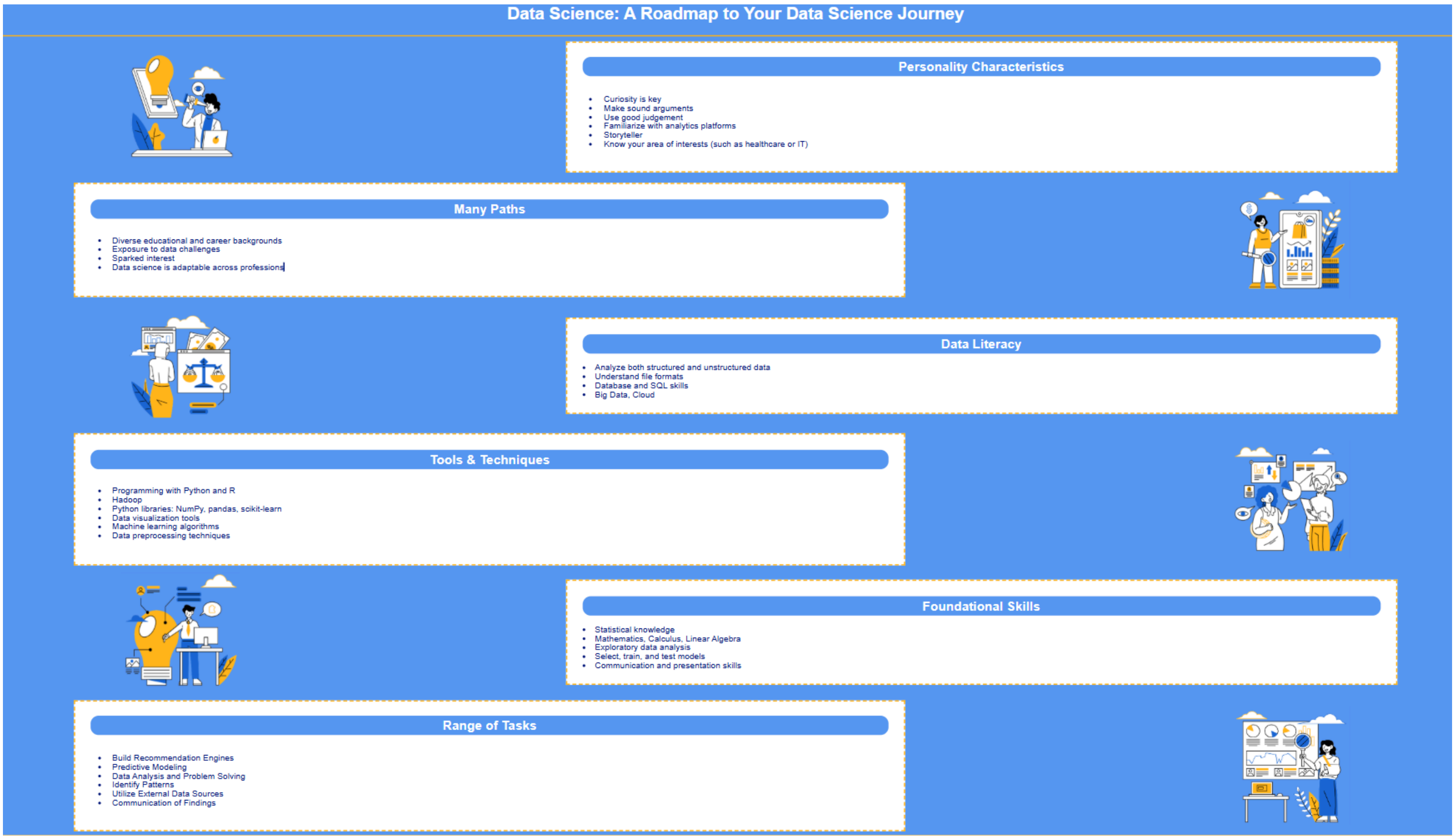
Summary: Careers and Recruiting in Data Science

Congratulations! You have completed this module. At this point, you know that:

- Data Science helps physicians provide the best treatment for their patients, helps meteorologists predict the extent of local weather events, and can even help predict natural disasters like earthquakes and tornadoes.
- Companies can start on their data science journey by capturing data. Once they have data, they can begin analysing it.
- Everyone who uses the Internet generates massive amounts of data daily.
- Amazon and Netflix use recommendation engines, and UPS uses data from customers, drivers, and vehicles to efficiently use the drivers' time and fuel.
- The purpose of the final deliverable of a Data Science project is to communicate new information and insights from the data analysis to key decision-makers.
- The report should present a thorough data analysis and communicate the project findings.
- Companies should seek someone excited about working with the data in their particular industry. They should seek out someone curious who can ask interesting, meaningful questions about the types of data they intend to collect. They should hire people who love working with data, are fluent in statistics, and are competent in applying machine learning algorithms.
- A organised and logical report should communicate the following to the reader:
    - What they gain by reading the report
    - Clearly defined goals
    - The significance of your contribution
    - Appropriate context by giving sufficient background
    - Why is this work practical and useful
    - Conjecture plausible future developments that might result from your work

Final Assignment

A Roadmap to Your Data Science Journey



Data Science: A Roadmap to Your Data Science Journey

Personality Characteristics

- Curiosity is key
- Make sound arguments
- Use good judgement
- Familiarize with analytics platforms
- Storyteller
- Know your area of interests (such as healthcare or IT)

Many Paths

- Diverse educational and career backgrounds
- Exposure to data challenges
- Sparked interest
- Data science is adaptable across professions

Data Literacy

- Analyze both structured and unstructured data
- Understand file formats
- Database and SQL skills
- Big Data, Cloud

Tools & Techniques

- Programming with Python and R
- Hadoop
- Python libraries: NumPy, pandas, scikit-learn
- Data visualization tools
- Machine learning algorithms
- Data preprocessing techniques

Foundational Skills

- Statistical knowledge
- Mathematics, Calculus, Linear Algebra
- Exploratory data analysis
- Select, train, and test models
- Communication and presentation skills

Range of Tasks

- Build Recommendation Engines
- Predictive Modeling
- Data Analysis and Problem Solving
- Identify Patterns
- Utilize External Data Sources
- Communication of Findings

Case Study: Final Assignment

Case Study: Lila's Journey to Becoming a Data Scientist: Her Working Approach on the First Task

This case study explores the data scientist's career path and key attributes, highlighting the skills, education, and experiences required to excel in this dynamic field. We'll follow the story of Lila, a fictional individual who aspires to become a successful data scientist.

There will be a quiz after this reading based on the contents of this case study.

Education and Skill Acquisition

With an undergraduate degree in economics and a substantial background in data analysis, Lila finds data science and its potential to drive meaningful change captivating. Inspired by her experiences, she decides to transition her career and step into the data scientist role.

Lila realises that she needs to enhance her skills and knowledge to embark on her data science journey. She enrolled in the IBM Data Science Professional Certificate online program that covers key topics like statistics, machine learning, data analysis, and programming languages like Python and SQL. She diligently completes coursework and practices her coding skills on real datasets.

Building a Strong Foundation

As she progresses in her studies, Lila deeply understands data science fundamentals such as data manipulation and visualisation with Python libraries like NumPy, Pandas, and Matplotlib. This strong foundation equips her with essential skills for data analysis.

Visualization for Storytelling

Lila knows she must communicate her findings effectively, so she learns which types of data visualizations will be most informative. She learns to create charts and graphs visually representing sales trends, customer segmentation, and product popularity, allowing stakeholders to grasp the data's significance. These visualisations help in storytelling and decision-making.

Hands-On Experience

Lila understands that practical experience is invaluable in data science. She started participating in Kaggle competitions and working on personal data projects. These experiences expose her to real-world data problems and help her develop problem-solving skills. Furthermore, she created her GitHub account and uploaded her projects to build her profile.

Data Wrangling and Preprocessing

Lila learns that data scientists spend a significant portion of their time on data cleaning and preprocessing. She worked on various datasets, learned data preprocessing using sed, NumPy and pandas Python libraries. She became skilled in handling missing data, outlier detection, and feature engineering to improve model performance.

Communication and Storytelling

Recognizing that data scientists must communicate their findings effectively, Lila honed her data storytelling skills. She learned various tools like matplotlib and plotly while she pursued her IBM Data Science Professional Certificate. She knew how to create compelling visualizations and present her insights clearly and understandably.

Networking and Collaboration

Lila actively participates in data science communities and attends meetups and conferences. She collaborates on open-source projects, connects with fellow data scientists, and gains exposure to various industries when she attends the IBM TechXchange Conference.

Domain Expertise

Understanding that domain knowledge is crucial, Lila chooses a niche that aligns with her interests. She looks deeply into several domains, including e-commerce, healthcare, finance, and other fields, to which she could effectively apply her data science skills. Since her master's in economics, she chose e-commerce as her core domain to land a data science career.

Landing the First Job

After months of preparation, Lila started applying for data scientist positions. She tailors her resume to highlight her relevant skills and projects. Her online portfolio showcases her capabilities and demonstrates her commitment to the field.

Lila's Approach to Working on Her First Task as a Data Scientist

As a newly hired junior data scientist at a retail company, Lila uses data insights to improve customer service. Her first assignment involves diving into customer data to identify patterns and anomalies that could impact customer service. She uses data analysis to enhance the overall customer experience.

Dataset Selection and Sourcing

In the initial phase of her data science journey, Lila faced the challenge of selecting a suitable dataset and procuring it from different sources. Apart from the historical data available for the organisations for the past four years, she scoured various repositories, websites, and databases to find the right datasets for her project. Upon collecting data from diverse sources, Lila encountered another crucial decision point. She had to decide how to harmonise and integrate these disparate datasets into a cohesive whole. She contacted product professionals, data engineers, and domain specialists, seeking their input and expertise in merging datasets.

Data Understanding and Cleaning

Lila begins by importing the dataset into her data analysis environment using Python and SQL. She loads the data and examines the first few rows to understand its structure and contents. Upon acquiring the dataset, Lila encounters her first challenge: data cleaning. Lila checks for missing values, duplicates, and outliers in the dataset. She addresses missing data by imputing or removing rows or columns with missing values. Outliers are identified and treated appropriately based on their impact on the analysis.

Exploratory Data Analysis (EDA)

As she delves into exploratory data analysis, Lila faces numerous choices. She must determine which summary statistics, visualisations, and distribution analyses will best reveal insights into customer behaviour and sales trends. Each choice she makes during EDA influences the story the data will tell. Lila conducts EDA to gain insights into the dataset. She generates summary statistics and visualisations (histograms, scatter plots) and explores the distribution of variables. EDA helps her understand customer behaviour, popular products, and sales trends.

Feature Engineering

Lila recognises the potential for feature engineering to enhance her analysis. She assesses whether creating new features, such as calculating total purchase amounts, will improve the dataset's utility for her project.

Statistical Analysis, Machine Learning

Lila evaluates whether statistical tests or machine learning algorithms are necessary. She employs regression analysis to understand relationships between variables and explore machine learning models for demand forecasting or customer segmentation tasks. Lila also performs statistical tests to uncover patterns in the data. She uses regression analysis to understand the impact of unit price on sales.

Presentation and Reporting

At the culmination of her analysis, Lila faces the challenge of presenting her findings. Lila compiles her analysis and conclusions using a Jupyter Notebook into a comprehensive report and presentation. She highlights actionable insights and recommendations for the e-commerce platform's stakeholders.

Continuous Learning

After completing her first project, Lila continues to refine her skills, explores more complex datasets, and tackles increasingly challenging data science tasks.

Machine Learning Skills

Although Lila took an introductory course on Machine Learning as part of the IBM Data Science Professional Certificate, the field intrigues her, and she wants to develop her skills further by taking the IBM Machine Learning Professional Certificate. She identified Machine Learning Repository datasets in the course and experimented with various algorithms. Lila dives into machine learning to excel as a data scientist, wherein she studies various algorithms, such as linear regression, decision trees, and deep learning models. She continues to gain expertise in selecting and fine-tuning algorithms based on specific data problems.

Explore Data Science Job Listings

Review and evaluate a data science job post.

Assignment overview

For this project, find a data science job posting on a job board of your choice, such as LinkedIn, Indeed, Zip Recruiter, Glassdoor, Monster, Naukri, or USAjobs.gov, that interests you.

Analyze the posting by responding to the following questions and statements. You do not need to submit your responses. This is an exercise to familiarise yourself with actual data science-related jobs.

Identify the following aspects of a data science job post:

1. What is the name of the company advertising the job?
2. What is the job title?
3. Where is the role located?
4. What is the expected salary or salary range?
5. What is the total number of results from the search for the job post?
6. What is one technical responsibility from the job post related to something you learned about in this course?
7. What are two technical skills required for the job post?
8. What are at least two ideas or concepts you learned about in this course relevant to these jobs?

Course Summary

Congratulations! You have completed this course. At this point, you know that:

- Data science is extracting valuable insights from vast datasets to guide strategic decision-making.
- Data science careers offer diverse paths, often involving mathematics, programming, and a curiosity for data exploration.
- Successful data scientists exhibit curiosity, critical judgment, and an aptitude for constructive argumentation.
- The data science field is characterised by high demand, resulting in attractive remuneration for skilled professionals.
- A Data Scientist's daily routine can vary significantly depending on the project's nature.
- A wide array of algorithms is available for extracting insights from data.
- Big Data plays a pivotal role in driving digital transformation across industries.
- Cloud computing is a fundamental technology in modern data science.
- Data mining techniques are essential for uncovering patterns and knowledge from data.
- Tools like Hadoop, HDFS, Hive, and Spark are employed to process big data.
- Deep learning, machine learning, and regression are critical data science topics.
- Data science applications span diverse domains, solving complex problems.
- Companies can harness data science to address age-old challenges with innovative solutions.
- Data science contributes significantly to saving lives and improving various aspects of society.
- Careers in data science offer exciting opportunities, with mathematics and statistics being essential

foundations.

- Reports in data science adhere to specific structures, and career roadmaps provide guidance.
- Case studies and projects offered practical application of the knowledge acquired during the course.

Congrats & Next Steps

Congratulations on completing this course! We hope you enjoyed it.

As you have learned, data science is an emerging field in high demand. To meet these, data scientists must combine soft skills like curiosity and knowledge of data science tools.

If you aspire to become a data scientist, we encourage you to complete the optional Honours project at the end of the course and explore jobs in this field.

If you want to know more about data science, we encourage you to take the next step by pursuing either the Tools for Data Science course or the IBM Data Science Professional Certificate. The introductory course for both programs is as follows: What is Data Science? The course, which you have now completed.

We also encourage you to leave your feedback and rate this course so that we can continue to improve our content.

Good luck!

Course Team and Acknowledgements

The entire course team thanks you for taking this course. We hope you enjoyed it and wish you the best in applying your new knowledge and skills.

Please rate the course and provide a review. Your feedback is much appreciated.

Best regards,

Course Team

- - -

This course has been brought to you through the involvement of the following team of contributors:

Primary Instructors

- Rav Ahuja

Alex Aklson

- Ph.D., Data Scientist

Other Contributors and Staff

Project Lead: Rav Ahuja

Instructional Designer: Bethany Hudnutt, Pratik Choudhary

Lab Author: Dr. Pooja

Technical Advisor: Dr. Pooja

Production Team

Publishing: Rachael Jones, Jada Harrison

QA: Mercedes Schneider, Deyanira Mares

Project Manager: Vishali (Karpagam Sangameswaran)

Video Production: Vaishali Rani, Alex Jones and Sohini Biswas.

Teaching Assistants and Forum Moderators

Teaching Assistants:

- Anamika Agarwal, Lavanya T S, Lavanya Rajalingam, K Sundararajan, Sapthashree K S, Vandana, Danie Kulsum, Rajshree Patil, Rithika Joshi and Manvi Gupta.