Data Science

2024/06/30

List of Figures

List of Tables

Data Science and Visualisation

This is a placeholder for course materials I am creating for a Data Science and Visualisation module. This should be treated as *very* alpha quality.

The course is delivered in ten blocks. Ideally students will study these over ten weeks rather than bite the whole thing off in a single week.

Week 1

In week one, we will introduce the subject and give you a flavour of what you can expect.

Week 2

In week two, we will look at how to find and evaluate existing datasets.

Week 3

In week three, we will look at how to extract data from various sources.

Week 4

In week three, we will look at how to manipulate data.

Week 5

In week five, we will look at hypothesis testing.

Week 6

In week six, we will look at exploratory data analysis.

Week 7

In week seven, we will look at how to visualise data.

Week 8

In week eight, we will look at how to tell stories with data.

Week 9

In week nine, we will look at other tools that can be used in data science and visualisation.

Week 10

Week 1: Introduction

First Things First

> Look at the module guide and the assessment briefs. Make sure you understand what is expected of you and what you need to do to pass the module. Make sure you understand the deadlines and the assessment criteria. Whilst the terminology may be unfamiliar at this stage, it should give you an understanding of what is required, when. If you have any questions, let your tutor know. It is a good idea to get the deadlines in your calendar now, so you can plan your time effectively.

Data Science is the study of data to find insights and trends. It is a multidisciplinary field that uses techniques from statistics, machine learning, and computer science to analyse and interpret complex data. Data visualisation is the process of presenting data in a visual format, such as charts, graphs, and maps, to help people understand the data and make informed decisions.

In this course we will explore the key concepts and techniques of data science and data visualisation, and learn how to apply them to real-world problems. We will cover topics such as data cleaning, data wrangling, data analysis, and data visualisation. The course purposefully doesn't focus on specific tools or programming languages, but rather on the underlying concepts and techniques that are common to all data science and data visualisation projects. As a data scientist, you may not have free-choice of tools, so it is important to understand the underlying principles and techniques that are common to all data science projects.

What Is Data Science?

Data Science is an essential part tool in the modern economy. It underpins the AI revolution and is used in a wide range of industries, from finance to healthcare to marketing. Data Science underpins our ability to:

1. Understand our customers better
2. Make informed decisions
3. develop new products and services

==Expand with examples==

==quote from Data Now/ doughnut economics/ Big data now==

Though you may associate Data Science with big tech companies or finance, it is increasingly essential to a broad range of industries. Universities will use data science to track and understand student performance to improve teaching and support. Healthcare providers use data science to improve patient outcomes. Governments use data science to improve public services. Data science is *everywhere*.

Types of Data Science

Data Science can be divided into four main areas. These areas are not mutually exclusive, and many data scientists will work across multiple areas.

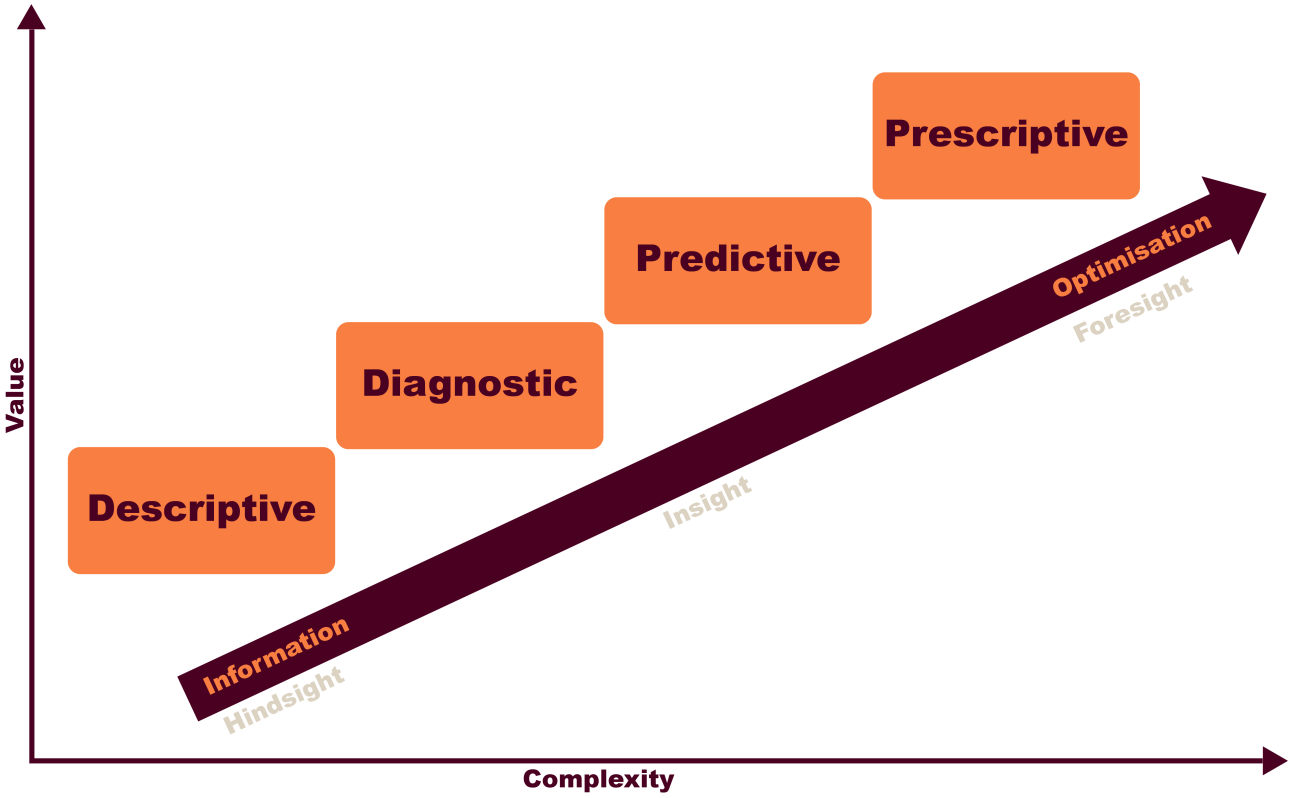


Figure 1 – Types of Data Science

Descriptive Analytics Descriptive analytics uses data to describe *what* has happened in the past. For example, a company might use descriptive analytics to analyze sales data to understand trends and patterns. You'll try your hand at descriptive analytics later in this week.

Diagnostic Analytics Diagnostic analytics uses data to understand *why* something happened. For example, a company might use diagnostic analytics to analyze customer feedback to understand why sales have dropped. You'll try your hand at diagnostic analytics later in this course.

Predictive Analytics Predictive analytics uses historical data to predict *how* likely something is to happen in the future. For example, a company might use predictive analytics to forecast sales or customer churn.

Prescriptive Analytics Prescriptive analytics uses data to recommend actions to achieve a desired outcome. For example, a company might use prescriptive analytics to optimise its supply chain or marketing

strategy.

What is a Data Scientist?

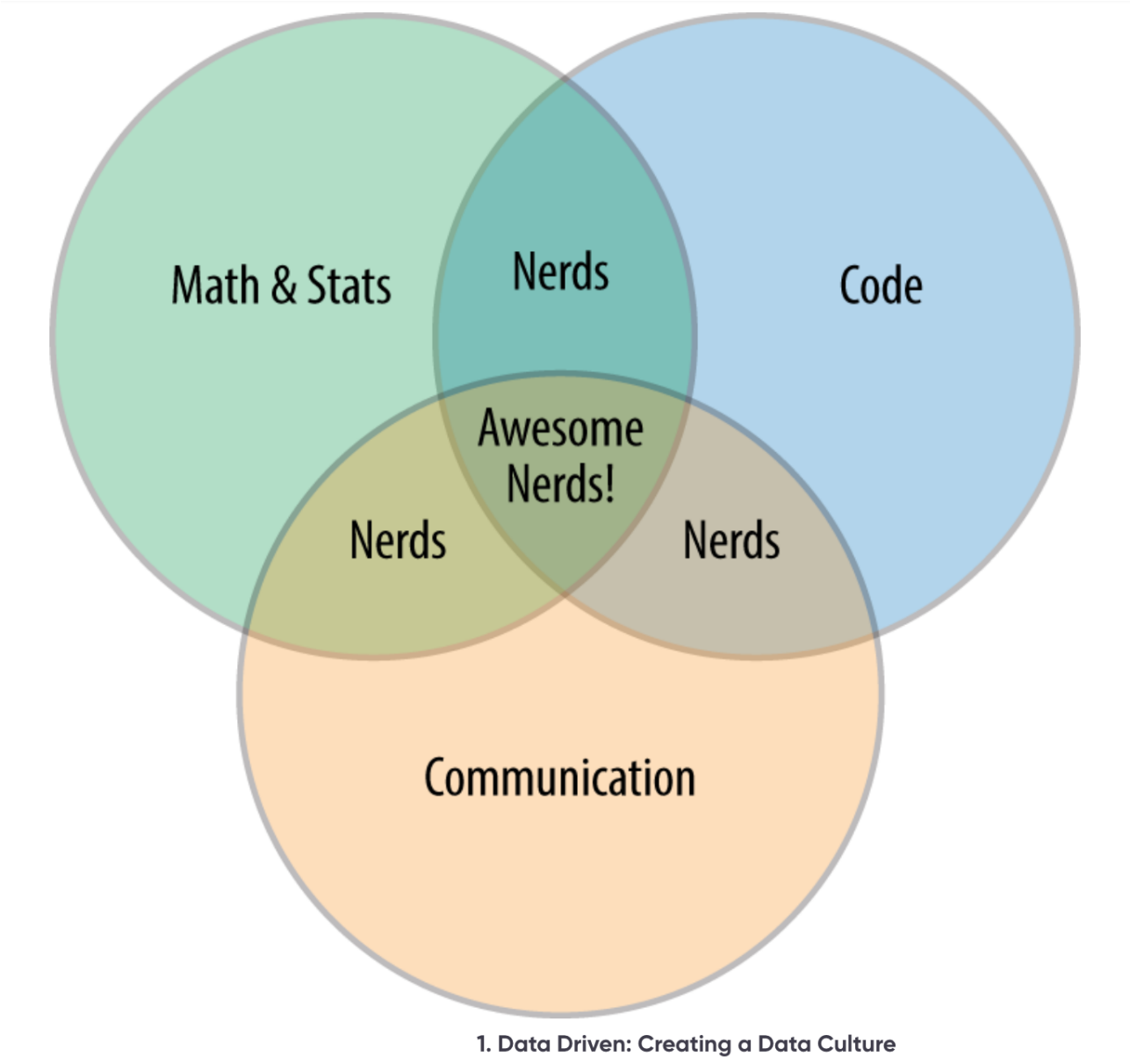


Figure 2 – The attributes of a Data Scientist

Image courtesy of Patil and Mason (2015, p. 3)

Case Study: The Housing Market Crash of 2008

In 2008, the housing market crash triggered a global financial crisis that led to the worst recession since the Great Depression. It has been depicted in films such as The Big Short and Inside Job. This crash

This housing market crash wasn't caused by a single factor, but rather a combination of interconnected issues:

Housing Bubble Easy access to credit and low interest rates inflated housing prices beyond their true value, creating a bubble. The American hosing market had traditionally been stable. Nate Silver reports that

> "After adjusting for inflation a $10,000 investment in a home in 1896 would be worth just $10,600 in 1996. The rate of return had been less in a century than the stock market typically produces in a year."

Silver (2013, p. 30)

You can see the relative US house prices in the graph below. Note the sharp increase in house prices starting in 2003, and the subsequent crash in 2008.

Activity: recreate this graph There is a simple dataset here that you can use to recreate this graph. I've created an `csv` (Comma Separated Values) file of the data in activities You can use Excel, Google Sheets, or any other tool you are comfortable with.

Discussion ==This'll be a half-hour minute "how I did it" session==

Subprime Mortgages Lenders offered risky loans to borrowers with poor credit history (subprime) at adjustable interest rates. These borrowers struggled to afford payments when rates went up. In the space of just three years, the share of the market for subprime mortgages had grown from 8.3% of the market to 23.5% of the market. This near-tripling of the market share meant that almost a quarter of mortgages were at

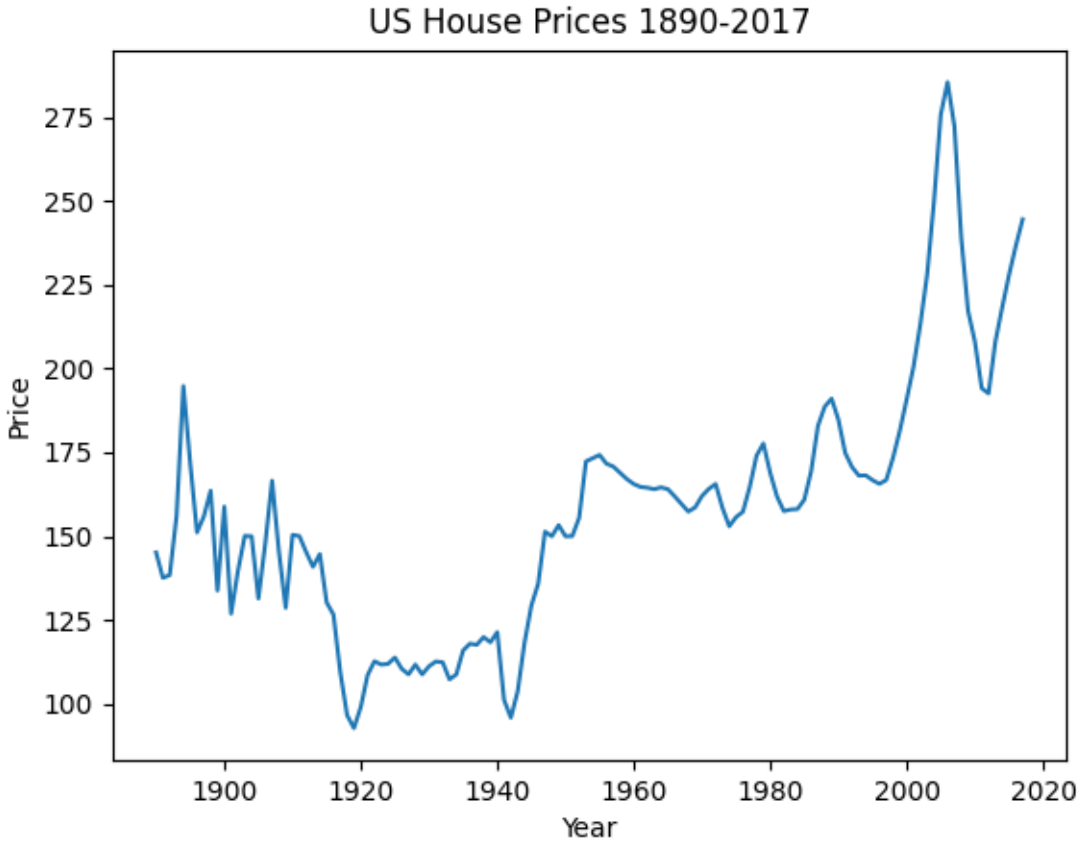
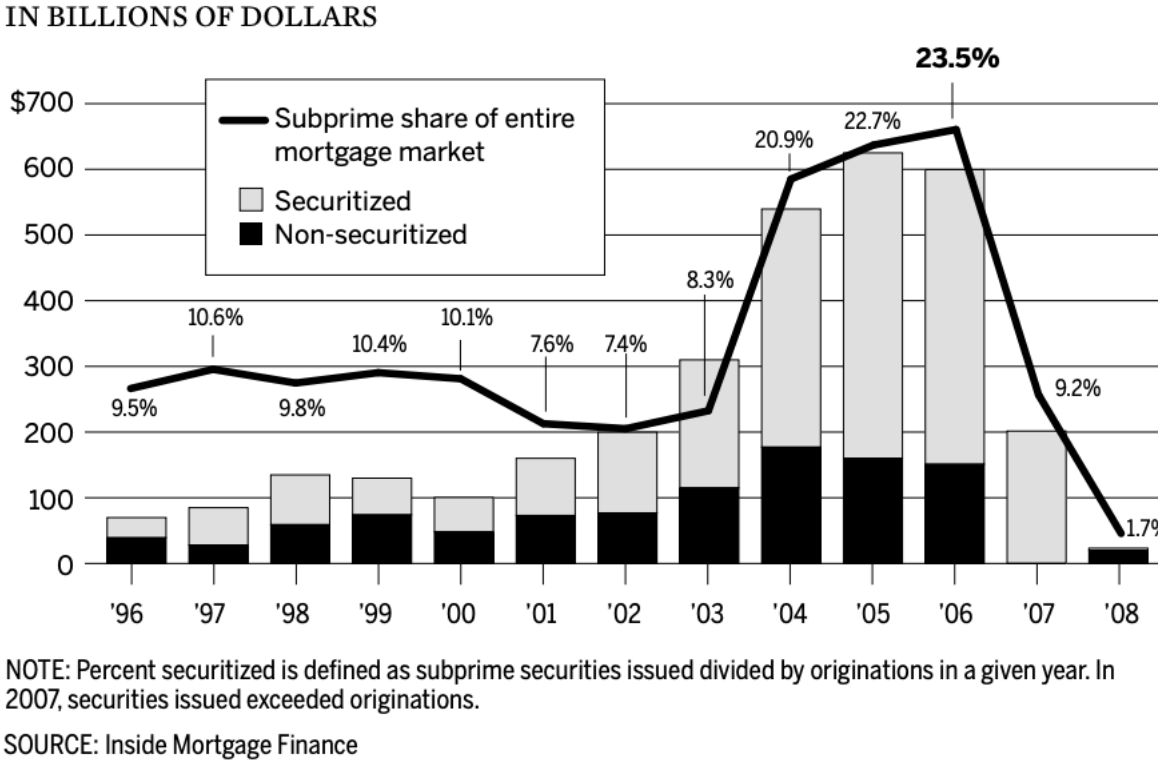


Figure 3 – US House Prices 1890 - 2017



NOTE: Percent securitized is defined as subprime securities issued divided by originations in a given year. In 2007, securities issued exceeded originations.

SOURCE: Inside Mortgage Finance

active risk of default.

Image courtesy of COMMISSION (2011, p. 70)

Extend your learning There is an excellent, albeit very technical explanation of the subprime mortgage crisis here Whilst the technicalities are beyond the scope of this course, it will give a great insight into the rigour of the analysis that data scientists and economists undertake.

Mortgage-Backed Securities One of the underlying reasons these risky mortgages were able to be offered was the development of new financial products. Financial institutions such as Lehman Brothers and Bear Stearns had developed new financial products that allowed them to bundle these risky mortgages together and sell them as investments. These Mortgage-Backed Securities (MBS) were sold to investors, spreading the risk but also making it hard to assess the underlying quality of the loans. Each Mortgage Backed Security is essentially a bet that the underlying mortgages will be repaid. These bets were then bundled together and sold on to other investors. This meant that the risk was spread across the financial system, but also meant that it was hard to assess the underlying quality of the loans. Lehman Brothers effectively bet thirty one times *on each mortgage* being repaid (Lioudis, 2024). When the housing market crashed, the value of these Mortgage Backed Securities plummeted, and financial institutions buckled under the weight of bad debt. This triggered a domino effect that led to the global financial crisis.

Predatory Lending These MBS were a primary contributing factor to the increase in predatory lending practices. Loan providers knew that they could sell on the risk, so they were less concerned about the underlying quality of the loans. Alongside this, mortgage brokers were incentivised to sell as many loans as possible, as they received a commission for each loan they sold. Silver reports that ===what=== Unethical lending practices included misleading borrowers about loan terms and inflating their income to qualify for loans. This meant that they could offer loans to people who were unlikely to be able to repay them. This also had the effect of accelerating the housing bubble; borrowers who were previously unable to afford properties were now able to buy them, which in turn accelerated the rise in house prices.

Lax Regulation Underlying all of these issues was a lack of regulation. The US government had deregulated the financial sector in the 1980s and 1990s, which allowed financial institutions to take on more risk. The government also failed to regulate the mortgage industry, which allowed predatory lending practices to flourish. The government also failed to regulate the financial products that were being developed, which allowed the creation of complex financial products that were difficult to understand and assess. Compounding this was the fact that salaries at credit agencies were a third of those at investment banks, meaning that the best talent ended up working for the banks, rather than the credit agencies. The gamekeeper was effectively working for the poacher.

These factors all came together to create a ticking time bomb. When the housing bubble burst, homeowners defaulted on their mortgages, the value of MBS plummeted, and financial institutions buckled under the weight of bad debt. This triggered a domino effect that led to the global financial crisis.

Could Data Science Have Prevented the 2008 Financial Crisis? The underlying causes of the 2008 financial crisis were complex, and it was only after the fact that the pieces of the puzzle came together. ===BULK THIS OUT with examples===

Despite the complexity of the contributing factors and the hidden risks, data science could have helped identify the warning signs of the impending crisis. Here's how:

Now do this Use the library search to find out more about the housing crisis. You could start with the following search terms:

- Housing crisis 2008
- Subprime mortgages
- Mortgage-backed securities
- Predatory lending
- Lax regulation

Asking the right questions

Not all problems are as complex as the 2008 financial crisis. ===BULK THIS OUT=== A good place to start with data science is understanding what problem you are trying to solve or interrogate.

Gutman and Goldmeier (2021, p4) suggest that the first step in any data science project is to define the problem you are trying to solve. They introduce a series of questions that can help you define the problem:

1. Why is this problem important?
2. Who does this problem affect?
3. What if we dont have the right data?
4. When is the project over?
5. What if we dont like the results?

Now read this:

Becoming a Data Head - How to Think, Speak, and Understand Data Science, Statistics, and Machine Learning Chapter 1

Activity 1.2

> Are there any other questions you should consider when defining a data science problem?

Discussion Whilst the questions above are a good starting point, there are many other questions you could consider when defining a data science problem. For example, you could consider the following questions:

Do I have enough data to draw robust conclusions? Whilst there are vast datasets available, it is important to consider whether you have enough data to draw robust conclusions. If you are working with a small dataset, you may need to consider whether you need to collect more data or whether you can draw robust conclusions from the data you have.

What are the limitations of the data? It is important to consider the limitations of the data you are working with. For example, you may need to consider whether the data is representative of the population you are interested in, whether the data is accurate, and whether the data is up-to-date.

Do I need to aggregate the data in some way? Online datasets are often purposefully aggregated to protect privacy. You should consider whether the the problem you are trying to solve needs de-aggregated data. If it does, perhaps you need to collect your own data, or aggregate multiple datasets.

What is the quality of the data? Often, raw data is messy and needs to be cleaned before it can be analysed. You should consider whether the data you are working with is clean, or whether you need to clean it before you can analyze it.

Week 2: Finding and working with data

Identifying datasets

Data Scientists often have to choose between using existing datasets or collecting their own data. This week we will look at how to find and evaluate existing datasets.

Evaluating Datasets

You may have come across the "PROMPT" criteria for evaluating texts. The PROMPT criteria are a set of questions that can be used to evaluate the quality of a text. The PROMPT criteria are:

- **Presentation**
  - Is the information presented clearly?
  - Is the language appropriate?
  - Is it succinct?
  - Can I understand it?
- **Relevance**
  - Does this information match my needs right now?
  - What is it mostly about?
- **Objectivity**
  - Is there bias in what you are reading?
  - Might the author/s have any hidden agendas? Have they been selective with their evidence?
  - Is the language used emotive?
  - Are opinions expressed?
  - Are there sponsors?
  - What are they selling? A particular product, a corporate view?
  - Is there contribution from different viewpoints by diverse authors to provide a balanced overview?
  - Are you selecting sources which confirm your own biases or seeking a broad range of perspectives on an issue?
- **Method**
  - Is it clear how any research was carried out?
  - How was data gathered?
  - If statistical data is presented, what is this based on?
  - Do researchers address any differences in outcomes between groups (e.g., ethnic/racial groups)?
  - Were the methods appropriate, rigorous, etc.?

- **Provenance**
  - Is it clear who produced this information?
  - Where does it come from? Whose opinions are these?
  - Do you trust this source of information?
  - Are there references/citations that lead to further reading, and are they trustworthy sources?
- **Timeliness**
  - When was it produced or published? Do any of the sources reinforce stereotypes or represent other outdated views?
  - Is it current?
  - Has the climate/situation changed since this information was made available?
  - Is it still up to date?

Extraction

Cleaning

> Cleaning the data is often the most taxing part of data science, and is frequently 80% of the work.

Data types

Missing data

Outliers

Duplicates

Normalisation

Wrangling

Making steps reproducible

Data extraction, cleaning, ‘wrangling’, etc

Week 3: Creating your own datasets

Last week we looked at how to find and evaluate existing datasets. This week we will look at how to create your own datasets.

https://www.linkedin.com/learning/sql-essential-training-20685933/

https://mystery.knightlab.com/

Week 4: Manipulating Data

When analyzing univariate data, we focus on its overall shape and characteristics. Key questions include:

- Location and spread of data points: What are the typical, minimum, and maximum values?
- Distribution pattern: Are data points evenly spread or clustered?
- Data set size: Is it large or small?
- Symmetry: Is the distribution symmetric or skewed?
- Tail weight: Are there many points far from the center, or are outliers rare?
- Clusters: Number, location, and size of any clusters.
- Outliers: Presence of significantly different data points.
- Other features: Any unusual aspects like gaps, sharp cutoffs, or peculiar values. Even simple data sets can reveal complex features.

Context

> The Numbers have no way of speaking for themselves. We speak for them. We imbue them with meaning.

Silver (2013)

Context is the information that surrounds the data. It is the information that tells us what the data is about. Fundamentally, data without context is meaningless. Data with context is *information* Context is essential both in understanding the raw data, and presenting our findings. Without context, we cannot hope to understand whether our analysis is accurate or not.

Often, the context is not explicit in the data. It is up to the data scientist to find the context, and to use it to interrogate the data. This is why data science is as much an art as a science. It is the art of finding the context, and using it to interpret and present the data.

Activity 4.1 How many trees are there on our planet?

This question was posed by David Spiegelhalter in his book "The Art of Statistics" Spiegelhalter (2019). How could we possibly answer this? We could jump into looking at forests and working out the density of trees, and extrapolating a value from that, but there is a more fundamental first step, we need to know *what is a tree*? Fortunately for us, there are definitions we can use.

> Trees are woody plants having a more or less erect perennial stem(s) capable of achieving at least 3 inches (7.6 cm) in diameter at breast height, or 5 inches (12.7 cm) diameter at root collar and a height of 16.4 feet (5.0 meters) at maturity in situ. Nelson *et al.* (2020). Though disputed (e.g., Brokaw and Thompson (2000)) this provides a starting point for our investigation.

Reading Activity

Context is explored in the the first chapter of "Storytelling with Data: A Data Visualization Guide For Professionals" by Cole Nussbaumer Knaflic. This book is available in the library.

Hypothesis Testing

Ethics in Data Science

You'll cover fundamental ethics in other modules, but there are a couple of ethical considerations that are particularly relevant to data science. These are:

- asymmetries of information
- biases in data

- privacy and security

> Whats the most evil thing that can be done with this? […]By asking the team to imagine what their impact could be if you abandon all constraints, you allow for a conversation that will help you identify opportunities that you would otherwise miss, and refine good ideas into great ones. We dont want to build evil products, but subversive thinking is a good way to get outside the proverbial box. (Patil Data Driven)

Asymmetries of Information

An asymmetry of information gives the party with more information an advantage in a transaction. Asymmetries of information are when one party in a transaction has more information than the other. This gives the party with more information an advantage in the transaction. For instance, if you are buying a used car, the seller knows more about the car than you do. This means that the seller has an advantage in the transaction. This is an example of an asymmetry of information.

In Week One we considered the Financial Crisis of 2008. No single factor caused to the financial crisis, and no single person knew the risks. Instead it was a combination of asymmetries of information that led to the crisis. The crisis was caused by this combination of partial truths:

Borrowers knew that they were taking on risky loans, but they believed that they would be able to refinance or sell their homes before the rates went up. **Lenders** knew that the loans they were offering were risky, but they believed that the risk was spread across the financial system, so they would not be affected if the loans went bad. **Investment Banks** knew that the MBS they were buying were potentially risky, but they believed that the underlying mortgages were sound. **Regulators** knew that the financial system was taking on more risk, but they believed that the market would self-correct.

The lenders had more information than the borrowers, the investment banks had more information than the lenders, and the regulators had more information than the investment banks. This meant that the risks were not fully understood, and the consequences were not fully appreciated.

Activity 5.1: look up the seminal paper on this subject by George Akerlof, "The Market for Lemons: Quality Uncertainty and the Market Mechanism" (1970). This paper is a classic in the field of economics and is a clear example of how asymmetries of information can affect markets.

https://doi-org.libezproxy.open.ac.uk/10.1016/B978-0-12-214850-7.50022-X

Privacy and Security

Data science relies on data, and data is often personal. This means that data science can raise privacy and security concerns. It is important to be aware of these concerns and to take steps to protect privacy and security.

There are laws and regulations that govern the use of data. For instance the GDPR (General Data Protection Regulation) seeks to limit the use of personal data and to protect the privacy of individuals. It has

Whilst these uses are important, it is also important to remember that data science can be used for good or for bad. It is important to be aware of the ethical implications of data science and to use data science responsibly. Zuboff (2019) argues that data science is being used to manipulate people and to control society. She argues that data science is being used to create a new form of power that she calls “surveillance capitalism”. It is important to be aware of these issues and to use data science responsibly.

now watch this:

https://www.youtube.com/watch?v=8HzW5rzPUy8

Target’s Predictive Analytics

In 2012, Forbes reported that Target (a major supermarket in the US) had sent a teenager a coupon book for expectant mothers Hill (2012). The teenager’s father was furious, but when he went to the store to complain, he found out that his daughter was indeed pregnant. Target had used predictive analytics to predict that the teenager was pregnant based on her shopping habits. Whilst likely apocryphal Piatetsky (2014), this story illustrates the power of predictive analytics and the ethical implications of using data science.

Activity:

The New York Times has a more nuanced take on this story. Read the article and consider the ethical implications of using predictive analytics in this way.

https://www.nytimes.com/2012/02/19/magazine/shopping-habits.html

Week 6: Exploratory Data Analysis

Week 7: Visualisation

Week 8: Storytelling

Week 9: Other Tools

- R
- Python
- SQL
- NoSQL
- Excel
- Tableau
- PowerBI
- D3.js

Week 10: Presentations

References

Brokaw, N. and Thompson, J. (2000) 'The H for DBH', *Forest Ecology and Management*, 129(1-3), pp. 89–91. Available at: https://doi.org/10.1016/S0378-1127(99)00141-3.

COMMISSION, T.F.C.I. (2011) *THE FINANCIAL CRISIS INQUIRY REPORT*.

Hill, K. (2012) 'How Target Figured Out A Teen Girl Was Pregnant Before Her Father Did', *Forbes*. https://www.forbes.com/sites/kashmirhill/2012/02/16/how-target-figured-out-a-teen-girl-was-pregnant-before-her-father-did/.

Lioudis, N. (2024) 'The Collapse of Lehman Brothers: A Case Study', *Investopedia*. https://www.investopedia.com/articles/brothers-collapse.asp.

Nelson, M.D. *et al.* (2020) *Defining the United States land base: A technical document supporting the USDA Forest Service 2020 RPA assessment*. NRS-GTR-191. Madison, WI: U.S. Department of Agriculture, Forest Service, Northern Research Station, pp. NRS-GTR-191. Available at: https://doi.org/10.2737/NRS-GTR-191.

Patil, D.J. and Mason, H. (2015) *Data Driven*. 1st edition. O'Reilly Media.

Piatetsky, G. (2014) 'Did Target Really Predict a Teen's Pregnancy? The Inside Story', *KDNuggets*. https://www.kdnuggets.com/did-target-really-predict-a-teens-pregnancy-the-inside-story.

Silver, N. (2013) *The Signal and the Noise: The Art and Science of Prediction*. 1st edition. London: Penguin.

Spiegelhalter, D. (2019) *The Art of Statistics: Learning from Data*. Pelican.

Zuboff, P.S. (2019) *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*. London: Profile Books.