DATA7001
Introduction to Data Science

Student workbook
Semester 2, 2020

About this workbook

This workbook is for a specific semester offering and you should make sure you have your own copy. However, you are encouraged to discuss and share your understanding with your peers and lecturers.

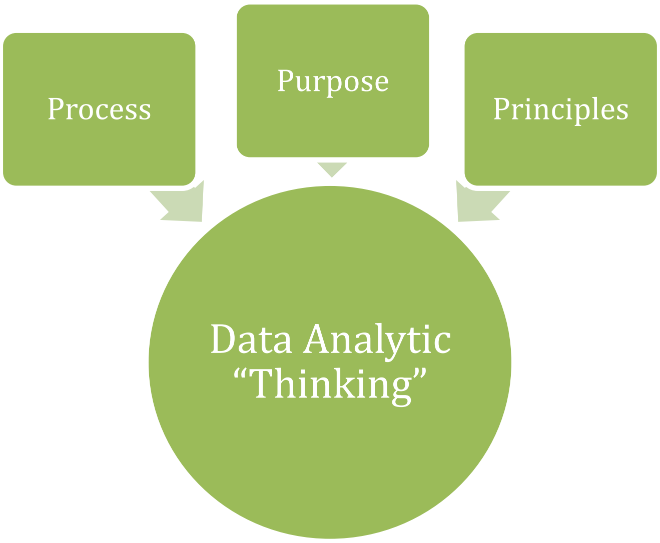
Table of Contents

About DATA7001

The aim of this course is to allow students to transform from a mindset of ‘learning about data science’ to ‘becoming a data scientist’.

- Apply design thinking methodology to data science problems
- Design effective data science processes from problem formulation to persuasive story telling with data
- Reason with the fitness of basic computational analytical models in data science scenarios
- Develop data-centric approaches to complex business and scientific problems

Towards developing the Data Analytic ‘Thinking”, the course has three focus areas of Process, Purpose and Principles.



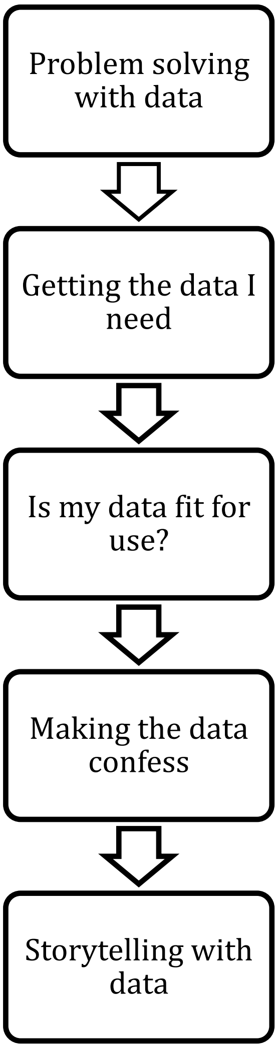
Process

DATA7001 modules are designed to align with a typical data science process.

Each step of the data science process has significant technical and business challenges. Overcoming those challenges will require learning in several techniques, methods and algorithms.

This course will introduce some representative techniques that will enable you to develop a data-driven mindset towards complex problems and equip you with the essential skills to navigate the data science process.

Various courses in the Master of Data Science program will cover additional advanced techniques relating to various aspects of the data science process.



→ DATA7901 Data Science Capstone Project 1

→ DATA7002 Responsible Data Science
→ DATA7201 Data Analytics at Scale

→ DATA7201 Data Analytics at Scale
→ DATA7202 Statistical Methods for Data Science

→ DATA7201 Big Data Analytics at Scale
→ DATA7202 Statistical Methods for Data Science

→ DATA7901 Data Science Capstone Project 1
→ DATA7902/3 Data Science Capstone Project 2

Purpose

Data Science is an emerging field that bring together knowledge from multiple disciplines including computer science, statistics and mathematics. Although the term has existed for more than 50 years, it has gained importance and visibility due to the *Big Data* phenomena.

According to estimates 2.5 quintillion bytes of data (2.5 followed by 18 zeros) is created every day from sensors used to gather environmental data, or posts made to social media sites, including photos and videos, business transaction data, GPS data from mobile devices and vehicles, and many more. It is even claimed that 90% of the data in the world today has been created in the last two years alone. Collectively this is termed Big Data.

Big Data is characterized by so-called Vs:

- Volume - terabytes, petabytes, …
- Velocity - batch, real-time, streams, …
- Variety - structured, unstructured, multimedia, …
- Veracity - reliability, availability, quality, ….
- Value - insights, foresights, actions/decisions, ….

Each V holds a number of technical and non-technical challenges. Value, is especially elusive as without value, there will be no return on investment in big data by any government, business, social or scientific endeavor. Value can be monetary but it can also be for social good or scientific progress.

There are two main ways to work with the data with the aim of generating value:

Data First

Region	**State**	**Population**	**City**	**Max**	**UV**
Southeast	Queensland	2572526	Brisbane	27	16
Southeast	Queensland	2572526	Ipswich	32	16
Southeast	Queensland	2572526	Logan	28	15
Wide Bay Burnett	Queensland	244847	Bundaberg	29	18
Wide Bay Burnett	Queensland	244847	Maryborough	28	17
Fitzroy	Queensland	187916	Gladstone	33	16
…	…	…	…	…	…

→ What insights can I find from this data?

Purpose First
Run a targeted campaign on sun safe awareness in Queensland?

→ What data do I need to serve the purpose?

This course takes a **Purpose First** approach and uses Design Thinking as the guiding methodology to formulate authentic data science problems and develop well-targeted solutions.

Principles

This course has five modules, corresponding to the data science process. For each module, a number of terms, concepts, methods and technologies will be introduced (collectively referred to as data science techniques). Given the vastness of the sub-disciplines of data science, these techniques can neither be complete in coverage nor in depth, in the span of one course.

The use of the 'first-principles' approach is intended to ensure that as you develop a mature understanding of these techniques, you continue to develop data science solutions that are explainable (why) and reproducible (how).

How to use this workbook

In the remaining workbook, you will find sections for each of the 5 modules. Each section includes:

Resources. A number of resources are provided which you may (or may not) need depending on your background. Where required, you are encouraged to utilize these resources to ensure you progress through the module at the expected pace. There is also a wealth of online resources, which you are encouraged to use where relevant and helpful, however you should discuss with course coordinators when in doubt to ensure authenticity of any external resources.

Practical work. A short description of the practical work associated with the module is provided. All practicals will take place in the designated lab (see timetable for the semester)

Reflective Questions. These questions are designed to enable you to set the scope of each module, and reflect/test and cement your understanding of the key techniques, methods or concepts (see index) related to the module. Use your lecture notes, readings, prac work, and peer discussions to document your understanding by (1) answering the questions and (2) creating descriptions of the terms in the index where relevant.

Discussion Questions. Additional questions are provided to help you expose deeper insights within the content of each module. You should prepare your response to these questions ahead of the tutorial session for the module.

1 Problem solving with data

In this module you will be introduced to some fundamentals. You will study the characteristics of (big) data and how they led to the emergence of data science [2] as a discipline. There are a number of online articles and blogs that discuss the distinctive characteristics of big data, generally referred to as the Big Data V's. Although there is some debate on which V's are more important [5] or even relevant, there is general consensus that these V's signify technical challenges [3] as well as business (or social, political and cultural) [4] challenges.

[1] Viktor Mayer-Schonberger and Kenneth Cukier. Big Data – A Revolution that will Transform how We Live, Work and Think. John Murray Publishers, UK 2013.
[2] Daniel Price. Surprising Facts and Stats about the Big Data Industry. Cloudtweaks.com, Mar 2015. http://cloudtweaks.com/2015/03/surprising-facts-and-stats-about-the-big-data-industry/
[3] Jagadish et al., Big data and its technical challenges, Communications of ACM, Vol 57, No 7, July 2014.
[4] https://hbr.org/2012/10/big-data-the-management-revolution.
[5] https://medium.com/@brennash/the-ten-fallacies-of-data-science-9b2af78a1862

The second topic in this module is on problem solving with design thinking. Use the provided further readings from the lecture notes and the learning from the design thinking task to supplement your understanding of this topic.

1.1 Practical Work

In this practical you will:

- Familiarise yourself with Jupyter Notebook, python and R
- Learn the basics of Unix filesystems and how to navigate them using the Unix shell

No preparatory tasks are needed before the practical. This practical is not assessed, but will be needed as a reference for future practicals. Many of the practicals this semester will be using *Jupyter notebook.* Each student has been provisioned their own notebook, which can be found at:

https://data7001-sXXXXXXX.uqcloud.net, replacing sXXXXXXX with your student number. Login to your personal notebook and navigate to the "Prac 1" directory. Then click the "Prac 1…" notebook to view the rest of the prac material.

Following additional resources can be used to get further familiarity with these environments:

[1] Refer to "Data Science BootCamp" lecture
[2] Jupyter: Jupyter notebook is a web tool that allows you to create documents that contain live code and explanatory text. You can read more about it at the

project's webpage, http://jupyter.org; Another useful resource is https://www.datacamp.com/community/tutorials/tutorial-jupyter-notebook. Note that you are provided with a pre-installed jupyter notebook for your prac sessions. However some of the general introductions to the notebook use can be helpful

[3] R: https://www.datacamp.com/courses/free-introduction-to-r

[4] Unix: https://swcarpentry.github.io/shell-novice/

1.2 Reflective Questions

Fundamentals

1. What are the Big Data Vs – explain with examples?
2. Explain the properties and give examples for the three generations of data, namely Transactional, Interactional and Sensory
3. Explain why the following are major challenges for Big Data initiatives: Lack of purpose, Cultural divide, Human intelligence and Data quality?

Design Thinking

1. What is Design Thinking and why is it relevant for Data Science?
2. Outline the importance of empathy and human centred design for data science

1.3 Discussion Questions

Review the Design Thinking Task sheet and prepare for the stakeholder interview

2 Getting the data I need

In this module you will be introduced to different types of data, and how they can be ingested into various data storage systems [1]. You will also be introduced to when the data you need should actually not be used (privacy, ethical or legal concerns) [2] and when you need to, or can only work with, a sample of the data [3].

[1] Andy Oram. Managing the Data Lake. O'Reilly Media, Inc. Sep 2015.
[2] Latanya Sweeney. k-anonymity: a model for protecting privacy. Int. J. Uncertain. Fuzziness Knowl.-Based Syst. 10, 5 (October 2002), 557-570.
[3] Section 4.2 of Nina Zumel and John Mount. Practical Data Science with R. Manning Publications Co., 2014

2.1 Practical Work

In this practical you will:

- Assess the ethical use of data using K-Anonymity
- Learn how to ingest data in different storage systems
- Reason with sampling strategies

Like Prac 1, this week's practical will be available on your *Jupyter notebook.* You can access your Jupyter notebook at https://data7001-sXXXXXX.uqcloud.net, replacing sXXXXXX with your student number. Login to your personal notebook and navigate to the "Prac 2" directory. Then click the "Prac2…" notebook to view the rest of the prac material.

The practical for this topic includes work on data ingestion. We will use two different data storage systems for ingesting data – a relational database management system (MySQL) and a big data file system (HDFS Hadoop Distributed File System). MySQL is used in the context of private data, where as the use of HDFS is general purpose.

The following resources can be used to get familiarity with MySQL and HDFS.

[1] MySQL: MySQL will be available through a web client called phpmyadmin. This will already be setup for you in the lab. You can refer to https://www.siteground.com/tutorials/phpmyadmin/ for a basic overview of the phpmyadmin interface and

[2] https://www.siteground.com/tutorials/phpmyadmin/phpmyadmin_mysql_query.htm to find more details on querying a MySQL database with SQL[1].

[3] HDFS: http://hadoop.apache.org provides an overview of the Hadoop framework. Review http://hadoop.apache.org/docs/current/hadoop-project-dist/hadoop-hdfs/HdfsDesign.html#Introduction to gain a basic understanding of the Hadoop distributed file system.

You will also do a range of activities in R to practice various sampling techniques (refer to introductory references for R in module 1).

2.2 Reflective Questions

Data Types

1. What are the prominent Data Types – use examples to explain?
2. What is Meta-data and why is it important for data science?
3. Give an example of each of the following data types: Structured Data, Text Data, Spatial Data, Time Series Data, Graph Data and Multimedia Data.

Data Ingestion

1. What is the purpose of Data Ingestion?
2. What are the potential pitfalls of creating a Data Lake?
3. Hadoop is a commonly used file system for Big Data, what are the features of Hadoop that make it appropriate for big data ingestion?

Data Privacy

1. Provide 1-3 ethical and legal considerations important for data scientists.
2. What are the two principles of private data release?
3. What is anonymisation failure?
4. How does k-anonymity support privacy preservation, and what makes it vulnerable? Give an example where an adversary's knowledge can violate a 2-anonymous dataset

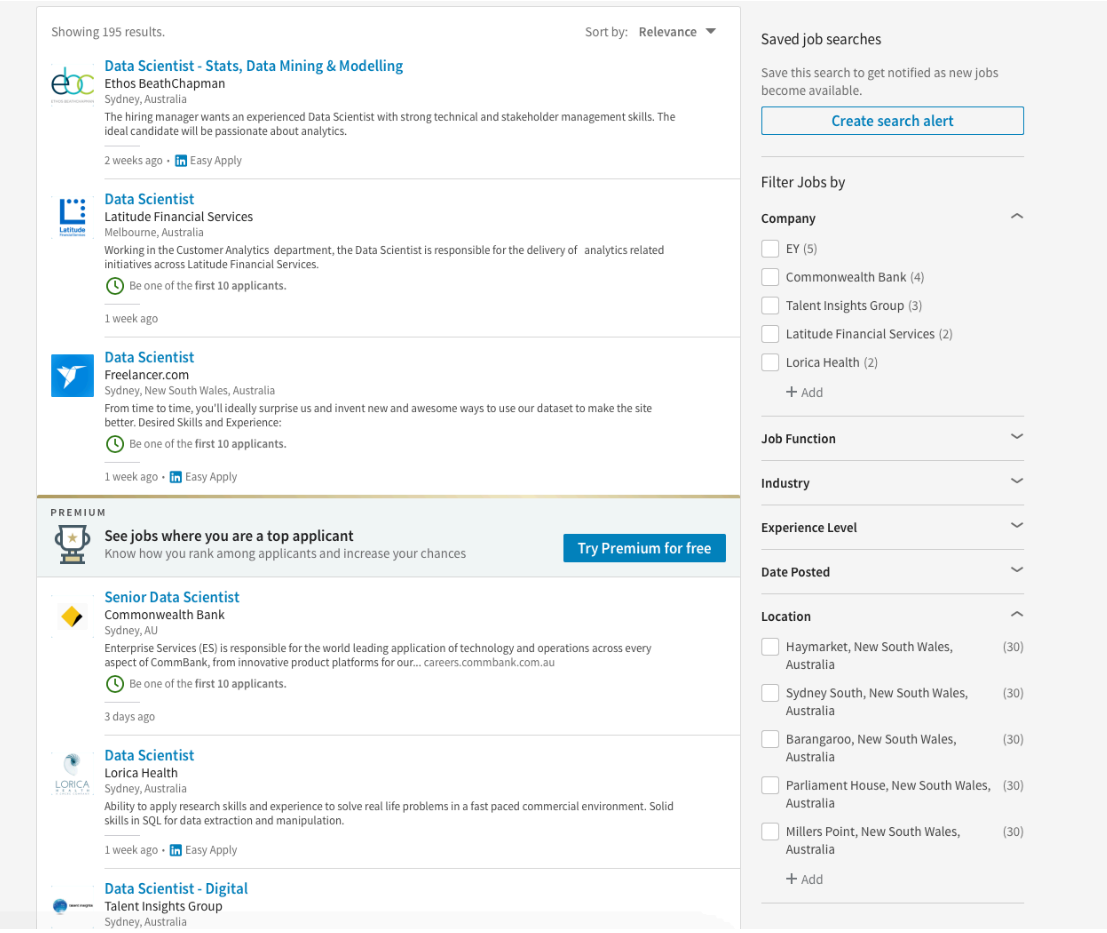
Data Sampling

1. What is sampling?
2. When and why is sampling used for Big Data?
3. Give examples when each of the following sampling approach is suitable: Simple Random Sampling, Weighted Random Sampling, Stratified Sampling?
4. What is the difference between categorical data, ordinal data and measurement data?

2.3 Discussion Questions

[1] If you have not done SQL in your undergraduate then you should consider taking INFS7901 as part of your data science program.

1. Which data types can you identify in the following (Tip: you may choose from Meta-data, Structured data, Text data, Spatial/Location data etc.)



Showing 195 results.

Sort by: Relevance

Data Scientist - Stats, Data Mining & Modelling
Ethos BeathChapman
Sydney, Australia
The hiring manager wants an experienced Data Scientist with strong technical and stakeholder management skills. The ideal candidate will be passionate about analytics.
2 weeks ago · Easy Apply

Data Scientist
Latitude Financial Services
Melbourne, Australia
Working in the Customer Analytics department, the Data Scientist is responsible for the delivery of analytics related initiatives across Latitude Financial Services.
Be one of the first 10 applicants.
1 week ago

Data Scientist
Freelancer.com
Sydney, New South Wales, Australia
From time to time, you'll ideally surprise us and invent new and awesome ways to use our dataset to make the site better. Desired Skills and Experience:
Be one of the first 10 applicants.
1 week ago · Easy Apply

PREMIUM
See jobs where you are a top applicant
Know how you rank among applicants and increase your chances
Try Premium for free

Senior Data Scientist
Commonwealth Bank
Sydney, AU
Enterprise Services (ES) is responsible for the world leading application of technology and operations across every aspect of CommBank, from innovative product platforms for our... careers.commbank.com.au
Be one of the first 10 applicants.
3 days ago

Data Scientist
Lorica Health
Sydney, Australia
Ability to apply research skills and experience to solve real life problems in a fast paced commercial environment. Solid skills in SQL for data extraction and manipulation.
1 week ago · Easy Apply

Data Scientist - Digital
Talent Insights Group
Sydney, Australia

Saved job searches

Save this search to get notified as new jobs become available.

Create search alert

Filter Jobs by

Company
□ EY (5)
□ Commonwealth Bank (4)
□ Talent Insights Group (3)
□ Latitude Financial Services (2)
□ Lorica Health (2)
+ Add

Job Function

Industry

Experience Level

Date Posted

Location

□ Haymarket, New South Wales, Australia	(30)
□ Sydney South, New South Wales, Australia	(30)
□ Barangaroo, New South Wales, Australia	(30)
□ Parliament House, New South Wales, Australia	(30)
□ Millers Point, New South Wales, Australia	(30)

+ Add

2. Give an example of source and adversary data (tables) where *k*-anonymity will result in a privacy breach

3. Explain a situation for which data sampling is appropriate.

4. What is the difference between sampling with replacement (WR) and sampling without replacement (WOR)? Why do we opt for WR sampling in this course?

5. Describe in detail how to take a stratified random sample, with probabilities proportional to size, of total size 10 from a set of 100 observations with two strata (20 observations in one stratum and 80 in the other stratum).

3 Is my data fit for use?

In this module you will learn about data quality and its importance in data science. The knowledge of data quality is used to assess the fitness of a given dataset for a specific analytical task. Since fitness for use is highly contextual, you will apply different techniques to explore, transform and enrich your data and improve its fitness for the task at hand.

The following resources can be used as a further reference to supplement the lecture material and gain familiarity with exploratory data analysis and imputation methods in R.

Data Quality:

[1] Shazia Sadiq (Editor). Handbook of Data Quality Management – Research and Practice. Springer-Verlag Berlin Heidelberg, 2013.

[2] http://dke.uqcloud.net/DataQualityPatterns/

Exploratory Data Analysis:

[3] Y. Cohen and J. Y. Cohen. Statistics and Data with R: An applied approach through examples. John Wiley & Sons, Chichester UK, 2008.
Part I: Chapters 1, 2, and 3. Part III: Chapter 8.
http://onlinelibrary.wiley.com.ezproxy.library.uq.edu.au/book/10.1002/9780470721896

[4] G. J. Myatt and W. P. Johnson. Making Sense of Data I: A Practical Guide to Exploratory Data Analysis and Data Mining, Second Edition. John Wiley & Sons, Hoboken NJ, 2014.
Chapters 1, 2, and 3.
http://onlinelibrary.wiley.com.ezproxy.library.uq.edu.au/book/10.1002/9781118422007

[5] J. W. Tukey. Exploratory data analysis. Addison-Wesley, Reading MA, 1977.
UQ Library: HA29 .T783 1977
Chapters 1, 2, and 3

[6] Tamraparni Dasu and Theodore Johnson. Exploratory Data Mining and Data Cleaning. Wiley, Hoboken New Jersey, 2003
Chapters 1 and 2.

Data Imputation:

[7] Stef van Buuren. Flexible Imputation of Missing Data. CRC Press, Boca Raton FL, 2012.
Part I: Chapters 1, 2, and 3
http://ebookcentral.proquest.com.ezproxy.library.uq.edu.au/lib/uql/detail.action?docID=888580

[8] Therese D. Pigott (2001). A Review of Methods for Missing Data. Educational Research and Evaluation, Vol. 7, No. 4, pp. 353-383

[9] R.J.A. Little and D.B. Rubin. Statistical analysis with missing data. Wiley, New York, 1987, [2ed, 2002].
Chapter 1.

3.1 Practical Work

In this practical you will learn how to:

- Explore and discover the quality characteristics of your data
- Transform, clean and enrich your data as needed while ensuring that the changes you have made are valid
- Perform exploratory data analysis using R

Like previous pracs, this week's practical will be available on your *Jupyter notebook.* You can access your Jupyter notebook at https://data7001-sXXXXXX.uqcloud.net, replacing sXXXXXX with your student number.

Login to your personal notebook and navigate to the "Prac 3" directory. Then click the "Prac 3 ..." notebook to view the rest of the prac material.

3.2 Reflective Questions

Data Quality

1. Name and explain briefly with examples 5 different data quality dimensions that can be detected or measured without user input?
2. Give two different examples of data incompleteness.
3. What is the difference between schema quality and data quality?
4. Is there a difference between the following two dimensions of data quality - accuracy to reference source and accuracy to reality?

Data Transformation

1. What is data curation and why is it important for data analytics?
2. Give an example each of three different ways to transform (clean) data to make it fit for use, namely cleaning from rules, cleaning from filter and cleaning from source.

Data Exploration

1. What is the purpose of data exploration?
2. What are sources of variability in data?
3. What is the difference between quantitative and qualitative variables?
4. What is the difference between discrete and continuous variables?
5. What are three techniques for visualising data?
6. What are common shape characteristics of univariate data?
7. How are outliers defined?

Data Enrichment (Data Imputation)

1. What are Rubin's missing data mechanisms?
2. What is an example of MCAR Missing Completely at Random, MAR Missing at Random, and MNAR Missing Not at Random?
3. What is the purpose of data imputation?
4. When and why is data imputation used?
5. How does the multiple imputation process work?

Data Enrichment (Data Integration)

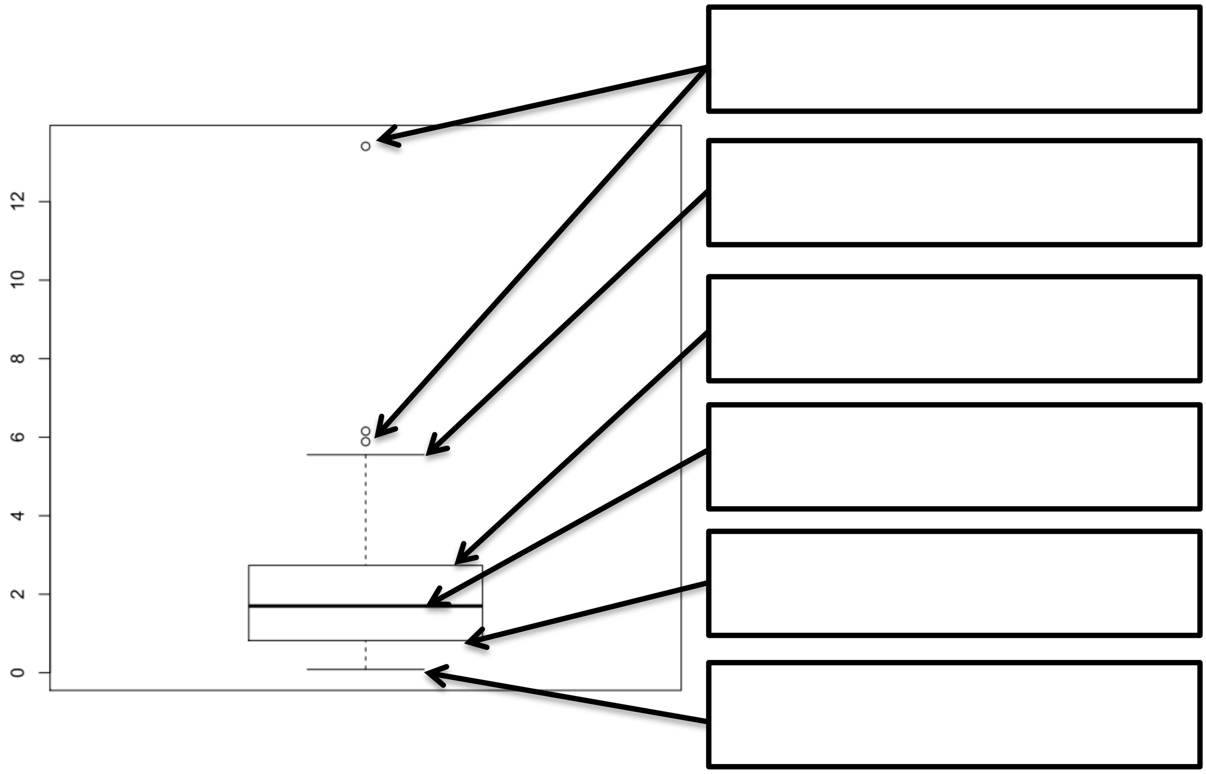
1. What is data integration and why is it needed by data scientists?
2. Give an example of data integration problems at schema level.
3. Give an example of data integration problems at instance level.
4. What is the difference between field matching and record matching?
5. Explain with an example the purpose of ‘edit distance’ in data integration.

3.3 Discussion Questions

1. Consider the following relational table. Name and briefly describe the data quality dimensions that this table violates.

caseNumber	sentence	registrationDate	category	firstName	lastName	age	sex	race	district	post
106A12593	Guilty		Incarcerated	Joshua	Porter	20	Male	Black	NORTHWESTERN	612
106D13687	Guilty	12/27/2011	Incarcerated	Brian	Mcclurkin	23	Male	Black	NORTHWESTERN	612
106D13999	Guilty	02/18/2011	Incarcerated	Joseph	Griffin	19	Male	Black	WESTERN	733
106D88900	guilty	07/26/2010	Incarcerated	Sterling	Melton	23	Male	Black		
âˆž	Guilty		Incarcerated	Corey	Alston	21	Male	Black	NORTHWESTERN	621
âˆž	Guilty	8/9/12	Baltimore	Mark	Taylor	25	Male	Black	NORTHWESTERN	624
âˆž	Guilty		Incarcerated	Abram	Hall	19	Male	Black	NORTHERN	522
âˆž	Guilty	10/23/2012	Baltimore	Donald	Actie	22	Male	Black	NORTHWESTERN	623
âˆž	GIUTLY	04/13/2012	Incarcerated	Terrell	Kennedy	22	Male	Black	NORTHWESTERN	634
âˆž	Guilty		Incarcerated	Kevin	Barber	27	Male	Black	EASTERN	321
106F10599	Guilty		Incarcerated	Jonathan	Oliver	22	Male	Black	NORTHWESTERN	632
106F11867	Guilty	1/11/13	Baltimore	James	Hamlet	21	Male	Black	NORTHWESTERN	622
106F14206	Guilty	1/2/13	Baltimore	Damien	Saunders	30	Male	Black	NORTHWESTERN	624
01K04873	60	10/24/2012	Baltimore	Joseph	Griffin	18	Male	Black	EASTERN	334
023A10838	Guilty		Incarcerated	Markie	Cole	24	Male	Black	NORTHEASTERN	435
026C09059	Guilty	1/2/13	Baltimore	Kali	Moulton	33	Male	Black	SOUTHEASTERN	211
026K19034	Guilty		Incarcerated	Melvin	Cummings	3119	Male	Black	NORTHWESTERN	622
âˆž	Guilty		Incarcerated	Bobby	Williams	19	Male	Black	SOUTHEASTERN	213
âˆž	Guilty	03/28/2012	Incarcerated	Harold	Singfield	23	Male	Black		
036G07553	Guilty		Incarcerated	Jason	Moody	37	Male	Black		
039H04732	Guilty		Incarcerated	Damaim	Thompson	31	Male	Black		
03L03830	Guilty		Incarcerated	Demetrius	Mayes	21	Male	Black	NORTHEASTERN	421
043L16294	Guilty	07/18/2012	Baltimore	Antonio	Bowens	19	Male	Black	SOUTHWESTERN	834
044J12261	Guilty		Incarcerated	Percy	Johnson	20	Male	Black	EASTERN	315

2. Label the following boxplot.



3. Describe a scenario which leads to data that is missing not at random (MNAR).

4. Explain a situation in which data imputation should be used.

5. What are two problems that arise from using deterministic regression imputation?

4 Making the data confess

In this module you will learn how to interrogate your data using a range of techniques for extracting hind-sights, insights and foresights from data.

The following resource can be used to get familiar with statistical learning methods in R:

[1] G. James, D. Witten, T. Hastie, R. Tibshirani. An Introduction to Statistical Learning: with Applications in R. Corrected 6th printing. Springer, NY, 2015. Chapters 1, 2, 3, 4, 5, 9, and 10 https://link-springer-com.ezproxy.library.uq.edu.au/book/10.1007/978-1-4614-7138-7/page/1

Following resources provide further general reading on the topic:

[2] Christopher D. Manning, Prabhakar Raghavan, Hinrich Schütze. An Introduction to Information Retrieval. Cambridge University Press, 2008.
Chapters 13, 14, and 15.

[3] N. Zumel and J. Mount. Practical Data Science with R. Manning Publications Co., 2014
Chapters 5, 6, 7, and 8.

[4] Andrew Gelman and Jennifer Hill. Data Analysis Using Regression and Multilevel/Hierarchical Models. Cambridge University Press, Cambridge, 2006.
Chapters 2, 3, and 4.

[5] http://www.tomdavenport.com/books/

4.1 Practical Work

In this practical you will learn how to:

- Perform a range of analytical queries using SQL to gain hindsight from a given dataset(s).
- Perform a range of statistical methods using R to gain insight and foresight from a given dataset(s).

Like previous pracs, this week's practical will be available on your *Jupyter notebook.* You can access your Jupyter notebook at https://data7001-sXXXXXX.uqcloud.net, replacing sXXXXXX with your student number.

Login to your personal notebook and navigate to the "Prac 4" directory. Then click the "Prac 4 …" notebook to view the rest of the prac material.

The practical for this topic will use SQL to demonstrate aggregation queries for big (structured / relational) data and R for a number of analytical techniques as discussed in lectures.

4.2 Reflective Questions

Hindsight

1. What is an aggregation query?
2. What is the difference between a transactional database and a data warehouse?
3. Give an example of a fact table and its associated star schema with three dimensions i.e a data cube. Provide examples of 2 aggregation (OLAP) queries that can be performed on your example data cube.

Insight

1. What is the difference between supervised learning and unsupervised learning?
2. What is the main focus of statistical inference?
3. What is the difference between regression and classification?
4. What is overfitting?
5. What is a technique one can use for each of: regression, classification, and clustering?
6. How can one measure the quality of a regression model?
7. How can one measure the quality of a binary classifier?

Foresight

1. What is the difference between training data and test data?
2. What is the purpose of prediction?
3. What is the purpose of cross validation?
4. For a regression model, what is the difference between a prediction interval and a confidence interval?

4.3 Discussion Questions

The following reading materials cover some basic concepts that will be helpful for understanding Module 4.

- Basics of probability: Sec. 1.2.1, Sec. 1.2.2, Sec. 1.2.4 from Pattern Recognition and Machine Learning. The book can be downloaded from https://www.microsoft.com/ en-us/research/publication/pattern-recognition-machine-learning/.

- Equations for lines/hyperplanes: https://en.wikipedia.org/wiki/Linear_equation.

- The concepts of estimation and hypothesis testing: https://ocw.mit.edu/courses/ mathematics/18-443-statistics-for-applications-fall-2003/lecture-notes/ lec1.pdf.

Use Q1 and Q2 to check your understanding.

1. (Normal distribution) A normal/Gaussian distribution with mean μ and variance σ^2, denoted by $N(\mu,\sigma^2)$, is defined as the distribution with probability density function

$$f(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2} .$$

The distribution $N(0,1)$ is called the standard normal distribution. Consider a standard normal variable X (i.e. a random variable following the standard normal distribution).

Scipy contains various useful functions for normal distributions in scipy.stats.norm module (https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.norm.html). Import the module as follows.

```
from scipy.stats import norm
```

(a) The cumulative distribution function of X is $\Phi(x) = P(X \leq x)$. This can be evaluated using the function scipy.stats.norm.cdf as follows.

```
x = 1.96
print(norm.cdf(x))
```

Run the above code, and determine the probability that $P(X \leq 1.96)$. What is the probability that $P(X \leq -1.96)$? What is the probability that $P(-1.96 \leq X \leq 1.96)$?

(b) The p percentile of X is denoted by z_p, that is, $P(X \le z_p) = p$. Thus $\Phi(z_p) = p$, or $z_p = \Phi^{-1}(p)$. The percentile can be evaluated using the function scipy.stats.norm.cdf as follows.

```
p = 0.975
print(norm.ppf(p))
```

What is $z_{0.975}$? What is $z_{0.995}$?

(c) Suppose that the input $x \in \mathbf{R}$ is related to the output $y \in \mathbf{R}$ via a linear model $y = 2x + 1 + \epsilon$, where $\epsilon \sim N(0,4)$. When $x = 2$, what is the distribution of y? Find $\delta > 0$ such that $P(y \in [5 - \delta, 5 + \delta]) = 0.95$.

(d) What is the function for sampling from a normal distribution? Use it to draw 1000 numbers from $N(5,4)$. What is the sample standard deviation? What is the standard error of the mean? What is the proportion of the numbers falling in the interval [1.08,8.92]?

2. (Hyperplanes)

(a) Find the equation of the straight line passing through (1,2) and (2,3). What is the intercept? What is the slope? Find a normal of the straight line.

(b) A plane in 3D space has a normal (1,−1,1), and its intercept with the z-axis is 3. Write down an equation for the plane.

3. Given the following Fact table for a data warehouse. What is the fact and what are the dimensions?

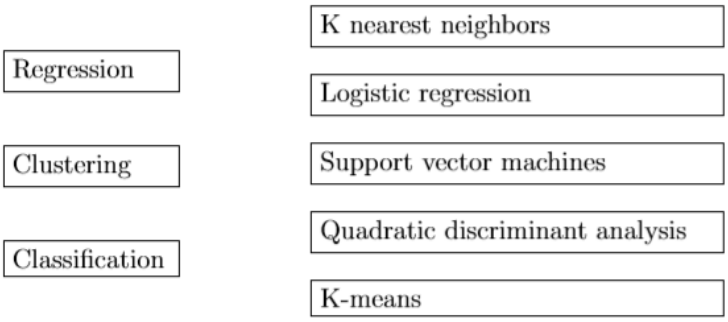
Day	Product	Store	Sales (AUD)
9.2.04	Milk	Toowong	3412
10.2.04	Milk	Toowong	2918
9.2.04	Bread	Toowong	2918
10.2.04	Bread	Toowong	3445
9.2.04	Milk	Sunnybank	5440
10.2.04	Milk	Sunnybank	4992
9.2.04	Bread	Sunnybank	2918
10.2.04	Bread	Sunnybank	3067

4. (Machine learning overview)

(a) Machine learning is only concerned with finding out how data is generated (that is, find the probability distribution generating the data). True or false? Explain your answer.

(b) What are supervised learning and unsupervised learning? Name a supervised learning algorithm and an unsupervised learning algorithm.

(c) Match the labels on the left and right.



Regression

Clustering

Classification

K nearest neighbors

Logistic regression

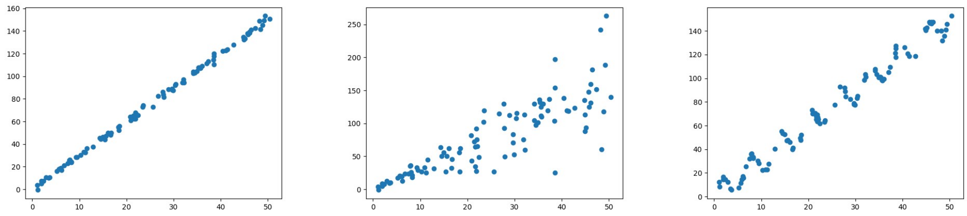
Support vector machines

Quadratic discriminant analysis

K-means

5. (Regression)

   (a) Linear regression often assumes that the output Y is related to the predictors $X_1,...,X_p$ as follows
   $Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p + \epsilon,$
   where ∈ is a random noise following a normal distribution $N(0,\sigma^2)$, and β_i's are model parameters. The following diagrams show the scatter plots for a few datasets. For each of them, explain whether it can be a collection of data points independently drawn from a linear model shown above.



(a) (b) (c)

   (b) Suppose the sales amount Y is related to the total budget for advertisement X by the following simple linear regression model
   $Y = \beta_0 + \beta_1 X + \epsilon,$
   where β_0 = 100, β_1 = 2 and$\epsilon \sim N(0, 4)$.

   i. What is the expected sales amount if X = 10?

   ii. When X = 10, can you construct a range $[l,u]$ such that the probability that $Y \in [l,u]$ is 95%?

   iii. What is the increase in the expected sales amount if the advertisement budget is increased by 1?

6. (Classification)

   (a) Given a training set consisting of three points (−1,0), (1,0) (0,1) labelled as red, green, blue respectively. Draw a diagram to show regions that are classified as red, green, and blue by 1NN.

(b) Given a training set with three positive examples (−2,−1), (−1,0), (−1,1), and three negative examples (1,−1), (1,0), (2,1). Draw the examples on the *xy*-plane. Are they linearly separable? If yes, what is the separating hyperplane (a straight line in this case) chosen by hard-margin SVM?

(c) Consider the following logistic regression model

$$P(Y = 1 \mid x, \boldsymbol{\beta}) = \frac{e^{\beta_0+\beta_1 x}}{1 + e^{\beta_0+\beta_1 x}},$$

where $\boldsymbol{\beta} = (\beta_0, \beta_1)$. If *x* is increased by 1, how much does the odds change?

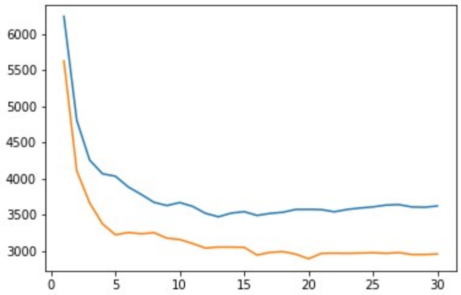
(d) Spams are relatively rare as compared to non-spam emails. In spam detection, we want to be able to filter out most spams but avoid filtering out non-spams. Is accuracy is a good performance measure for spam detection? If not, what performance measures are appropriate? Explain your answer.

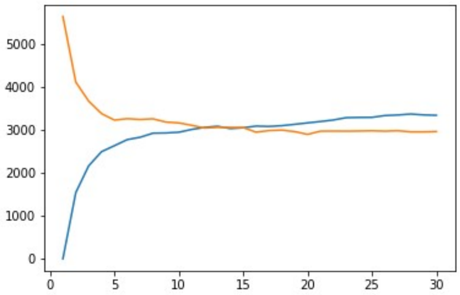
7. (Clustering) Consider the following points on the real line: -6, -5, -4, -1, 0, 1, 4, 5, 6. Use *k*-means algorithm to cluster them into three groups, assuming that initially they are grouped as follows: {-6, 6, 0}, {-5, -4, 1}, {5, 4, -1}. Show your working. What are the centers of the final clusters?

8. (Model selection)

(a) If we train a linear model on a training set and its sum of squared error on the training set is 0, then given any new example (*x*,*y*), the model can correctly predict that *x*'s value is *y*. True or false? Justify your answer.

(b) The following are two typical plots showing for *k*NN classifier (a) how the training and test errors vary as *k* increases, and (b) how the cross validation and test and test errors vary as *k* increases. Which is (a)? Which is (b)?





5 Story telling with data

In this module you will learn about key approaches to effective data visualization and storytelling.

Visualising data is an important step in the data storytelling process. The ability to incorporate infographics and visualisations into stories allows us to see and present information in new and interesting ways. Data can be made into visual narrative. Further, interactive elements of data visualisations allow people to explore the information and make sense of it on their own.

To study further on visualization from R, the following can be used as a reference:

[1] T. Rahlf. Data Visualisation with R: 100 Examples. Springer, Cham, Switzerland, 2017. https://link-springer-com.ezproxy.library.uq.edu.au/book/10.1007/978-3-319-49751-8/page/1

5.1 Practical Work

In this practical you will learn how to:

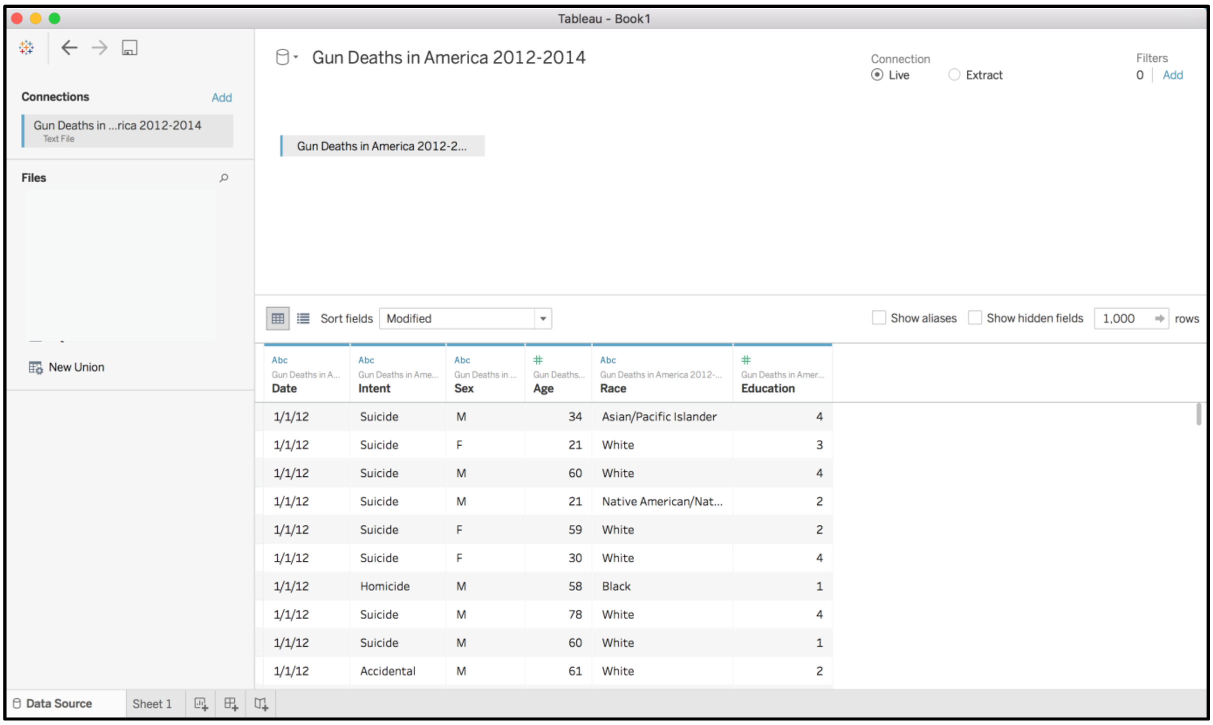
- Use Tableau to visualize different data sets.
- Present an 'interesting' story located within your data

The practical for this topic will use Tableau to demonstrate data visualization. Although the Tableau software is already setup for you in the lab, you can also install it on your own device to gain further knowledge. Tableau provides training material accessible after registration from https://www.tableau.com/learn/tutorials/on-demand/introduction-tableau.

There are three exercises in this practical detailed below.

Exercise 1: Visualising data using Tableau
Tableau is a visualisation tool that allows for the creation of a range of interactive data visualisations. It uses a drag and drop interface – meaning it is relatively easy to use once you get the hang of it. Today's practical will introduce you to the basics of Tableau. Note that there are many useful visualisation tools available, including powerful libraries such as Vega and d3.js that can be used to develop visualisation solutions in your own projects.
-> Spend some time getting used to the functionality of Tableau

Importing data
Begin by importing the 'Gun Deaths' data into Tableau using the options on the left hand side. If your data is in .xls format, then select the 'Excel' document option. If your data is in .csv format should be imported using the 'Text' option. You'll then be brought to the data import screen.



Gun Deaths in America 2012-2014

Connection ● Live ○ Extract

Filters 0 | Add

Gun Deaths in America 2012-2...

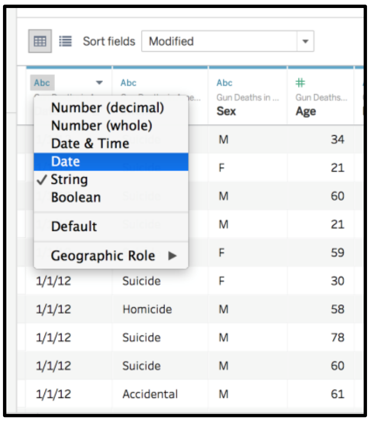
Sort fields: Modified

□ Show aliases □ Show hidden fields 1,000 rows

Abc Gun Deaths in A... Date	Abc Gun Deaths in Ame... Intent	Abc Gun Deaths in ... Sex	# Gun Deaths... Age	Abc Gun Deaths in America 2012-... Race	# Gun Deaths in Amer... Education
1/1/12	Suicide	M	34	Asian/Pacific Islander	4
1/1/12	Suicide	F	21	White	3
1/1/12	Suicide	M	60	White	4
1/1/12	Suicide	M	21	Native American/Nat...	2
1/1/12	Suicide	F	59	White	2
1/1/12	Suicide	F	30	White	4
1/1/12	Homicide	M	58	Black	1
1/1/12	Suicide	M	78	White	4
1/1/12	Suicide	M	60	White	1
1/1/12	Accidental	M	61	White	2

Here you can get a preview of the data and make the final necessary adjustments before beginning your visualisation. The headings of each column are displayed in bold. If Tableau doesn't automatically import your column headings, simply click the grey box containing file name in the middle of the screen, and select 'Field Names are in First Row'.

Tableau will automatically detect the kind of data contained in each column. For example, in our Gun Deaths data, Tableau has labelled data in the 'Age' column as numeric, while it has categorised data in the 'Race' column as text data. However, Tableau doesn't always automatically categorise data correctly. For example, data in the 'Date' column has been labelled as 'text' data, rather than time data. In order to make sure we can create the visualisation we want, we must make sure all the columns are correctly categorised. We can change the data type of a column by clicking the data type icon (#, ABC, etc.) and selecting the appropriate type. For our Gun Deaths data, we need to categorise the 'Date' column as 'date' data.



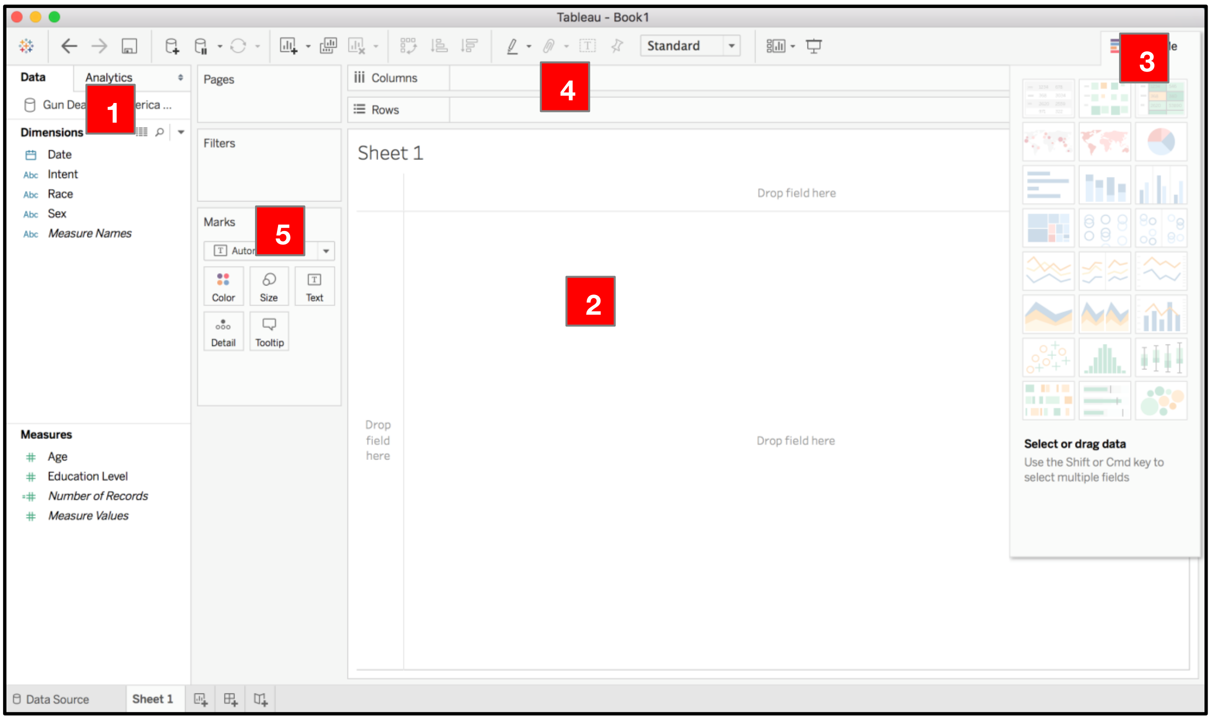
Sort fields: Modified

Abc	Abc	Abc Gun Deaths in ... Sex	# Gun Deaths... Age
		M	34
		F	21
		M	60
		M	21
		F	59
1/1/12	Suicide	F	30
1/1/12	Homicide	M	58
1/1/12	Suicide	M	78
1/1/12	Suicide	M	60
1/1/12	Accidental	M	61

The Tableau workspace

At the base of the screen you will see a tab called 'Sheet 1'. Click this, and you will be taken to the main Tableau workspace. In Tableau, you may have several sheets operating at the same time: each sheet allows you to make a different visualisation using the same imported data.

There are a number of key elements in the Tableau workspace:



Select or drag data

Use the Shift or Cmd key to select multiple fields

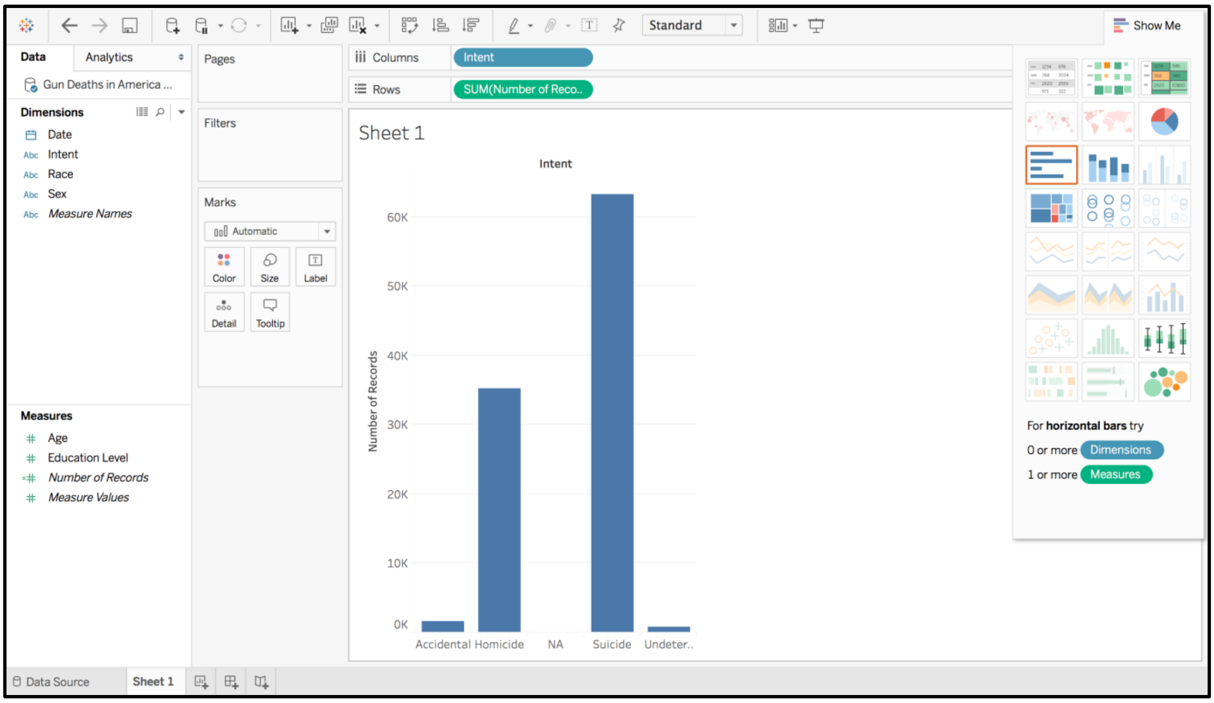
[1] In this left hand pane is where the elements of your dataset are listed. They are separated into dimensions (discrete data such as text, location or dates) and measures (numerical data). You can drag data from this pane into the visualisation.

[2] This space is where the main visualisation will appear.

[3] This pane on the right hand side lists the type of visualisations you can create. Once you have dragged data into the visualisation, use this pane to select and switch between visualisation types.

[4] This is where you drag the data that you would like to include in the visualisation. Data dragged into the columns shelf will make up the column headings of the table, while (you guessed it!) data in the rows shelf will be listed as rows. Whether your data is listed in the rows or columns will alter your visualisation. As with most elements of Tableau – the best way to get the hang of this is to try out different combinations.

[5] This is the 'Marks' tab, where you can change the design of your visualisation, including colours, labels, and tooltips.



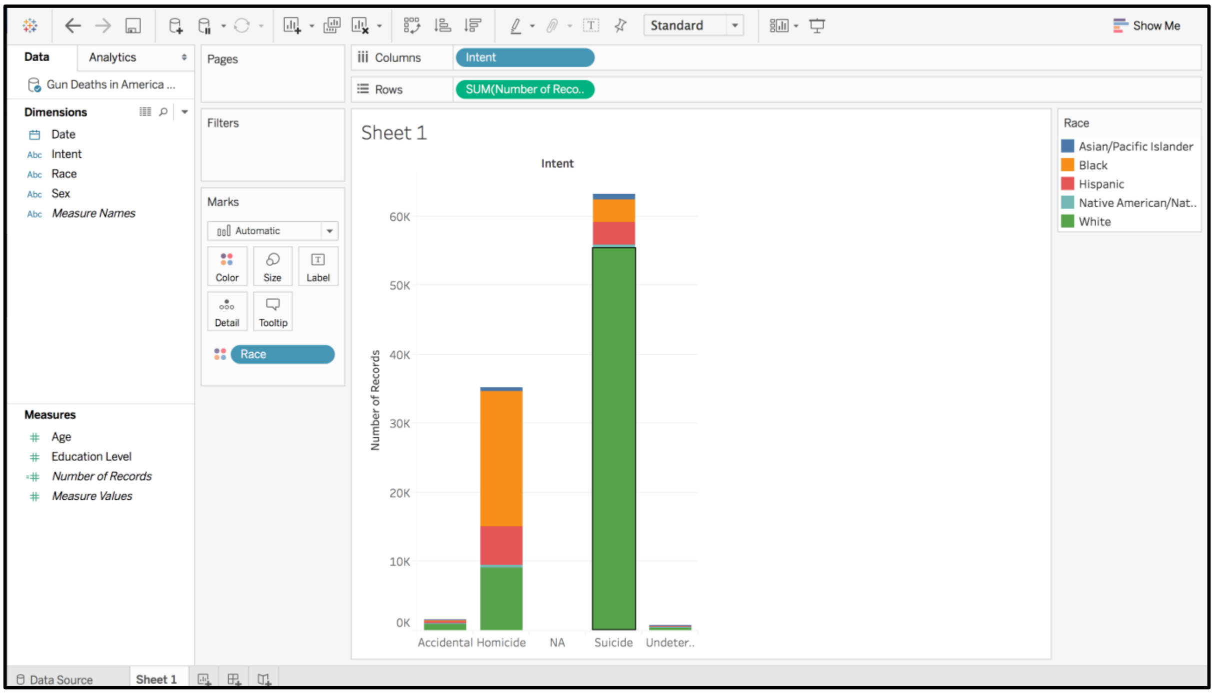
Columns: Intent

Rows: SUM(Number of Reco..

Let’s try out making some visualisations with our Gun Deaths data. From our ‘Gun Deaths’ data, try dragging the ‘Intent’ data into the ‘Columns’ shelf.

You’ll note that in addition to your own column headings, there are some automatically generated data types. Of particular interest is the ‘*number of records’* entry. This essentially works as a tally function, and will prompt Tableau to simply add up how many instances there are for each different entry in a data type. If we pair this with ‘Intent’ data, and drag the ‘Number of records’ data to the ‘Rows’ shelf, Tableau will generate a bar chart that shows us how many times each entry was recorded. We can clearly see that suicide was the most common.

So what if we add some other data to this graph? For example, can we tell if intent is linked with race? If we drag the ‘Race’ data onto the ‘Colour’ tab, Tableau will colour each bar according to race.



Columns: Intent

Rows: SUM(Number of Reco..

Marks: Automatic

Color: Race

Now we can see that most suicides were Whites, while most Homicides were Blacks. What about sex? If we change the colour tab to display 'Sex' instead of 'Race', Tableau will recolour the bars according to Male and Female.

Play around with different data types in different places and see what interesting links you can discover. Switch out 'Intent' for 'Race' or 'Sex' or change the visualisation type around and see if a different view reveals new information. Try adding 'Date' to the columns and add 'Number of records' to the rows. Make it a line graph: what do you see? (Hint – you might need to expand the date data using the plus!). Remember you can make multiple visualisations by simply adding another sheet at the base of Tableau.

Exporting a visualisation

Once you're done, you can export a visualisation using the Worksheet menu -> Export.

Exercise 2: Visualising different datasets using Tableau

Now that you've got the hang of using Tableau, try importing some of the other datasets you have been using in the course. The choice is yours; the focus is on using the tool to try to discover interesting patterns in the data.

Tableau contains support for storyboarding. Navigate to and find the storyboard features in the top menu and use this to build a data story. If you happen to find something interesting include it in the story, remembering what was discussed in lectures regarding the link between story elements and the wider data context. Are there individual datapoints that allow you to contextualise this data?

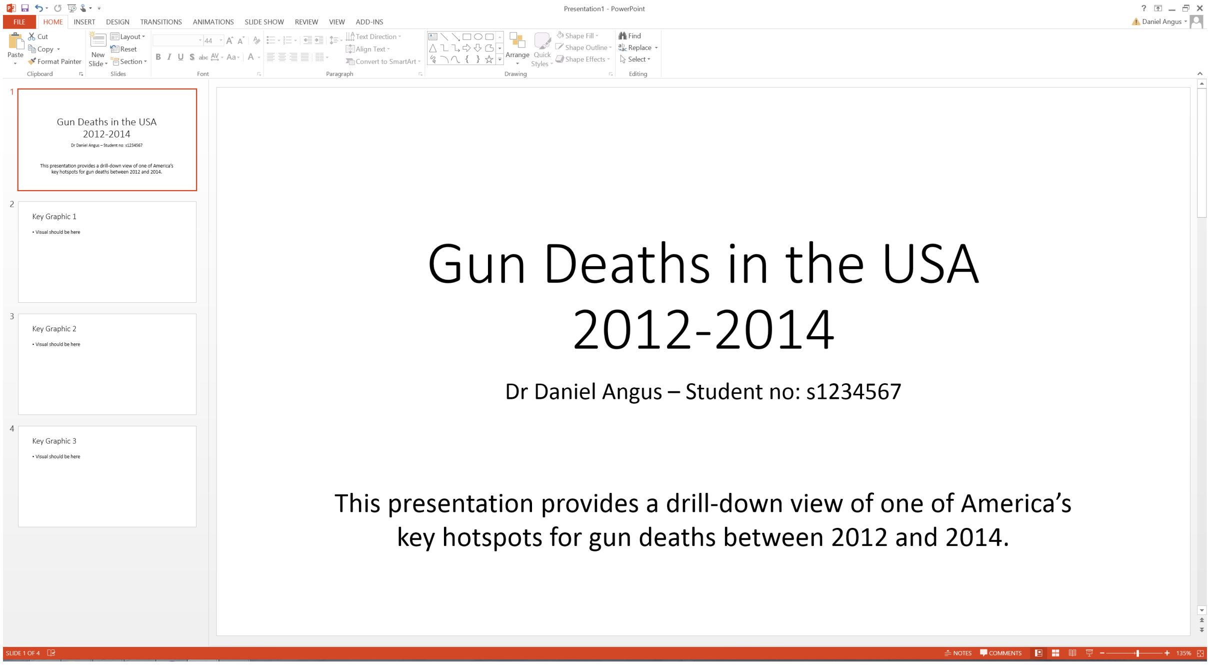
-> Construct a storyboard with at least 3 panels using any dataset you like.

Exercise 3: Storytelling

Once the storyboard is complete the next stage is communicate this story to an external audience. The final task of the practical is to exercise your data storytelling skills. Using your storyboard above, or otherwise construct a 4 slide powerpoint deck with the notes section containing the key points to make for each visual. Use the first slide to provide a heading and your student name and number, and a brief introduction to the dataset you have chosen. The remaining three slides should provide insight into the data you have chosen using one of the Tableau storytelling styles (e.g. drill-down, outlier, contrast).

-> Submit your powerpoint file (if using Mac please export as Powerpoint from Keynote) for assessment.

Example title slide (using gun deaths data set as an example. Note you can use any data you choose):



1

Gun Deaths in the USA 2012-2014

Dr Daniel Angus – Student no: s1234567

This presentation provides a drill-down view of one of America's key hotspots for gun deaths between 2012 and 2014.

2

Key Graphic 1

- Visual should be here

3

Key Graphic 2

- Visual should be here

4

Key Graphic 3

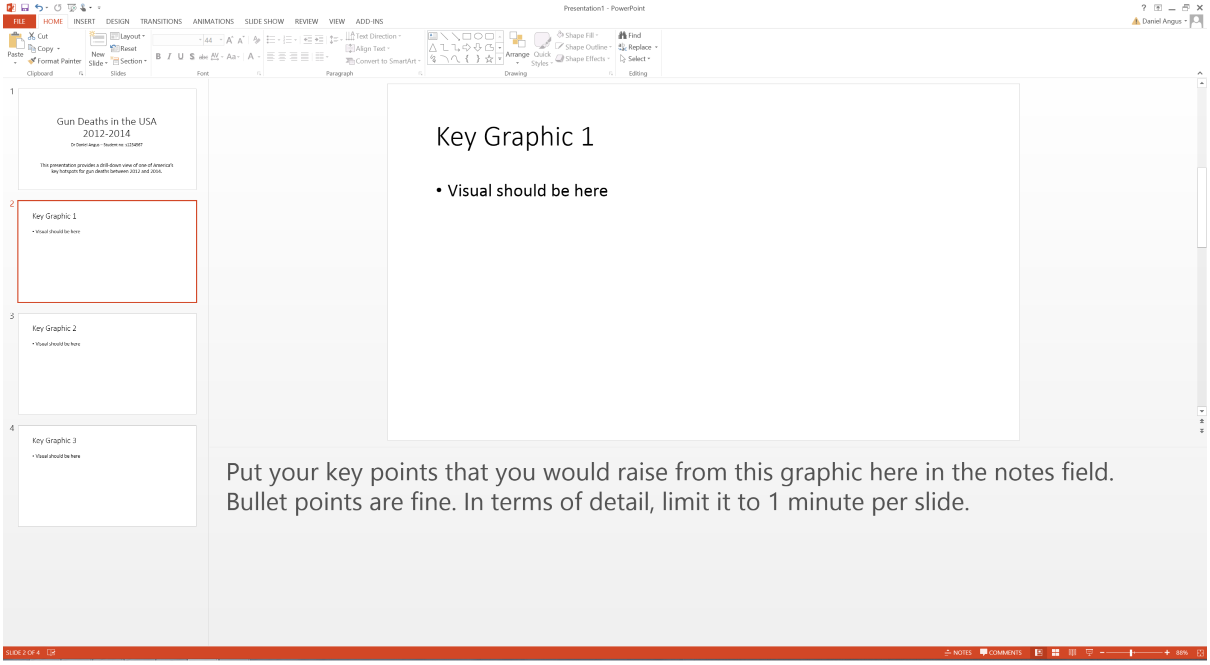
- Visual should be here

Gun Deaths in the USA 2012-2014

Dr Daniel Angus – Student no: s1234567

This presentation provides a drill-down view of one of America's key hotspots for gun deaths between 2012 and 2014.

Example showing how notes section should be used:



1

Gun Deaths in the USA 2012-2014

Dr Daniel Angus – Student no: s1234567

This presentation provides a drill-down view of one of America's key hotspots for gun deaths between 2012 and 2014.

2

Key Graphic 1

- Visual should be here

3

Key Graphic 2

- Visual should be here

4

Key Graphic 3

- Visual should be here

Key Graphic 1

- Visual should be here

Put your key points that you would raise from this graphic here in the notes field. Bullet points are fine. In terms of detail, limit it to 1 minute per slide.

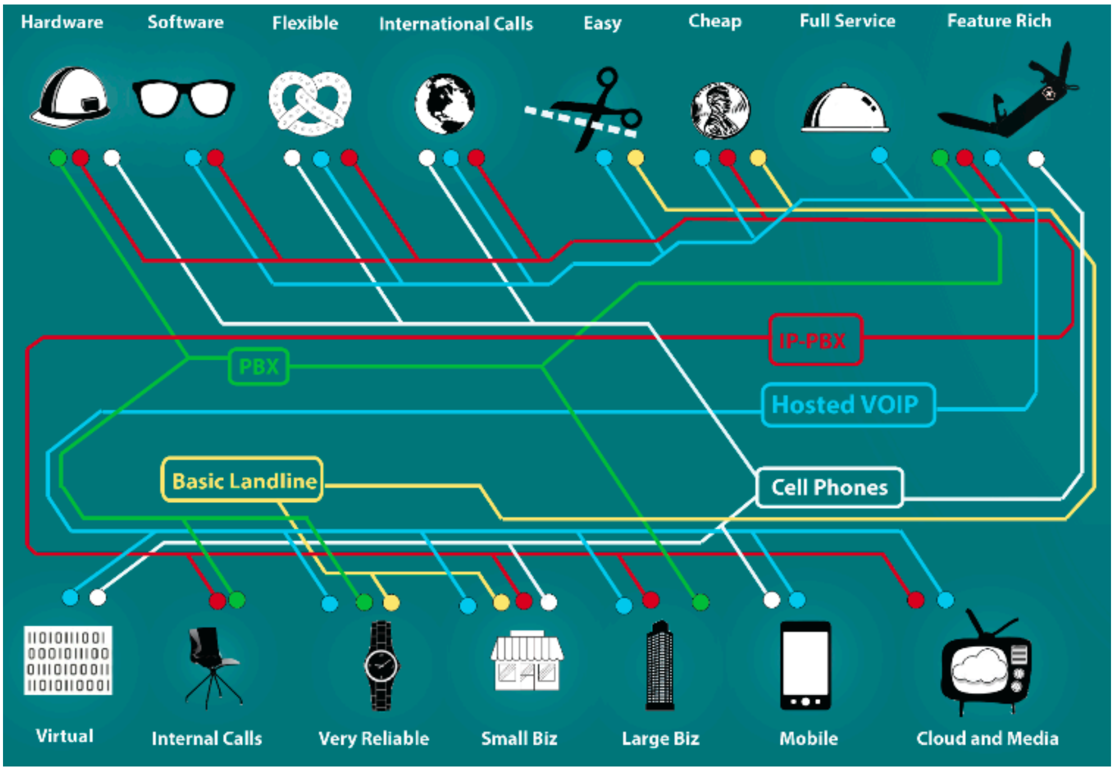
5.2 Reflective Questions

1. What are the seven visual storytelling methods?
2. What are the Gestalt laws, and why would we need to be aware of them when designing a visualisation?
3. What distinguishes visual analytics from visual storytelling?
4. What is a figurative element and why might we use it in visual storytelling?

5.3 Discussion Questions

1. Using these examples of data journalism:
   http://tinyurl.com/m9taokt
   http://tinyurl.com/kkntmdn
   http://tinyurl.com/kslv74w
   Determine the kinds of stories and story styles being presented/used
   Critique these data stories, what do you think works well or not so well in each example?

2. Visualisation Critique: Identify the figurative element(s) in this visualization, do they adhere to the principles of figurative visualization discussed in the lecture, briefly discuss why or why not? In what other way is this an example of good or bad visualization design?



3. Visual storytelling: If you were provided with a dataset of sports statistics for Australian Rules Footballers over the 2016 season, containing data such as the number of kicks, handballs, goals, fouls, and other player specific data, in what way might you use the 'outlier' data storytelling technique?

Index