FIT3152 Data analytics– Lecture 4

Assignment 1: recap, Q&A

The data science industry

Data science workflow models

Dirty, and Tidy data, tidyr

Transforming data:

- Recoding, Extracting subsets,
- Drawing a heatmap.

Consultations on Zoom

Clayton consultations have commenced:

- Any student can attend any consultation.
- Schedule on Moodle, https://lms.monash.edu/
- Current days/times:
- Monday 9:30-10:30AM, 2:00-3:00PM, 6:00-7:00PM,
- Tuesday 9:00-10:00AM, 12:00PM-1:00PM,
- Wednesday 10:00AM-11:00, 11:00-12:00PM,
- Thursday 1:00PM-02:00PM, 6:00PM-7:00PM.
- Please check the schedule for any changes.

Week-by-week

Week Starting	Lecture	Topic	Tutorial	A1	A2
28/2/22	1	Intro to Data Science, review of basic statistics using R	...		
7/3/22	2	Exploring data using graphics in R	T1		
14/3/22	3	Data manipulation in R	T2	Released	
21/3/22	4	Data Science methodologies, dirty/clean/tidy data, data manipulation	T3		
28/3/22	5	Network analysis	T4		
4/4/22	6	Regression modelling	T5		
11/4/22	7	Classification using decision trees	T6		
		Mid-semester Break		Submitted	
25/4/22	8	Naïve Bayes, evaluating classifiers	T7		Released
2/5/22	9	Ensemble methods, artificial neural networks	T8		
9/5/22	10	Clustering	T9		
16/5/22	11	Text analysis	T10		Submitted
23/5/22	12	Review of course, Exam preparation	T11		

Assignment 1

Assignment 1: Summary

FIT3152 Data analytics – 2022: Assignment 1

Your task	• Analyse the activity, language use and social interactions of an on-line community using metadata and linguistic summary from a real on-line forum and submit a report of your findings. • This is an individual assignment.
Value	• This assignment is worth **20%** of your total marks for the unit. • It has 30 marks in total.
Suggested Length	• 6 – 8 A4 pages (for your report) + extra pages as appendix (for your code) • Font size 11 or 12pt, single spacing
Due Date	**11.55pm Friday 22nd April 2022**
Submission	• PDF file only. Naming convention: *FirstnameSecondnameID.pdf* • Via Moodle Assignment Submission. • Turnitin will be used for similarity checking of all submissions.
Late Penalties	• 10% (3 mark) deduction per calendar day for up to one week. • Submissions more than 7 calendar days after the due date will receive a mark of zero (0) and no assessment feedback will be provided.

Assignment 1: Instructions

Instructions

Submit the results of your analysis, answering the research questions and report anything else you discover of relevance. If you choose to analyse only a subset of your data, you should explain why.

You are expected to include at least one multivariate graphic summarising key results. You may also include simpler graphs and tables. Report any assumptions you've made in modelling, and include your R code as an appendix.

There are two options for compiling your report:
(1) You can submit a single pdf with R code pasted in as machine-readable text as an appendix, or
(2) As an R Markup document that contains the R code with the discussion/text interleaved. Render this as an HTML file and print off as a pdf and submit.

Regardless of which method you choose, you will submit a single pdf, and your R code will be machine readable text. We need to conform to this format as the university now requires all student submission to be processed by plagiarism detection software.

Submit your report as a single PDF with the file name ***FirstnameSecondnameID.pdf*** on Moodle.

Assignment 1: Software

Software

It is expected that you will use R for your data analysis and graphics and tables. You are free to use any R packages you need but please document these in your report and include in your R code.

Assignment 1: Questions a & b

Questions

Activity, language use and social interactions in an on-line community. Analyse the metadata and linguistic summary from a real on-line forum and submit a report of your findings. Do the following:

(a) Analyse activity and language on the forum over time:

1. How active are participants over the longer term (that is, over months and/or years)? Are there periods where activity increases or decreases? Is there a trend over time? **(3 Marks)**
2. Looking at the linguistic variables, do the levels of these change over the duration of the forum? Is there a relationship between linguistic variables over the longer term? **(3 Marks)**

(b) Analyse the language used by threads:

We can think of threads as groups of participants posting on the same topic.

1. Using the relevant linguistic variables, is it possible to see whether or not particular threads are happier or more optimistic than other threads, or the forum in general, at different periods in time. **(3 Marks)**

Assignment 1: Question c

(c) Analyse social networks online:

We can think of authors posting to the same thread at similar times (for example during the same month) as having a connection to each other, forming a social network. This is called a two-mode network. When an author posts to more than one network during the same time period their social network extends to include authors from both networks, and so on. We will cover social network analysis in Lecture 5.

1. Create a non-trivial social network of all authors who are posting over a particular time period. For example, over one month. To create this, your social network should include at least 30 authors, some of whom will have posted to multiple (2 or more) threads during this period. Your social network should be connected, although some authors may be disconnected from the main group. Present your result as a network graph. **(3 Marks)**
2. Identify the most important author in the social network you created. Looking at the language they use, can you observe any difference between them and other members of their social network? **(3 Marks)**

Assignment 1: Overall considerations

(d) Overall considerations:

- The quality and clarity of your reasoning and assumptions. **(3 Marks)**
- The strength of support for your findings. **(3 Marks)**
- The quality of your writing in general and communication of results. **(3 Marks)**
- The quality of your graphics throughout, including at least one high-quality multivariate graphic. **(3 Marks)**
- The quality of your R coding. **(3 Marks)**

Assignment 1: Data generation

Data

The data is contained in the file webforum.csv and consists of the metadata and linguistic analysis of posts over the years 2002 to 2011. You will each work with 20,000 posts, randomly selected from the original file. The linguistic analysis was conducted using Linguistic Inquiry and Word Count (LIWC), which assesses the prevalence of certain thoughts, feelings and motivations by calculating the proportion of key words used in communication. See http://liwc.wpengine.com/ for more information, including the language manual http://liwc.wpengine.com/wp-content/uploads/2015/11/LIWC2015_LanguageManual.pdf

Create your individual data as follows:

```
rm(list = ls())
set.seed(XXXXXXXX) # XXXXXXXX = your student ID
webforum <- read.csv("webforum.csv")
webforum <- webforum [sample(nrow(webforum), 20000), ] # 20000 rows
```

Assignment 1: Data fields

Data fields given. (see the language manual for more detail and examples):

Column	Brief Descriptor	Column	Brief Descriptor
ThreadID	Unique ID for each thread	we	"We, us, our" words
AuthorID	Unique ID for each author	you	"You" words
Date	Date	shehe	"She, her "him words
Time	Time	they	"They" words
WC	Word count of the text of the post	posemo	Expressing positive emotions
Analytic	Summary: Analytical thinking	negemo	Expressing negative emotions
Clout	Summary: Power, force, impact	anx	Indicating anxiety
Authentic	Summary: Authentic tone of voice	anger	Indicating anger
Tone	Summary: Emotional tone	sad	Indicating sadness
ppron	"I, we, you" words	focuspast	Expressing a focus on the past
i	"I, me, mine" words	focuspresent	Expressing a focus on the present
focusfuture	Expressing a focus on the future	focusfuture	Expressing a focus on the future

Assignment 1: Data extract

ThreadID	AuthorID	Date	Time	WC	Analytic	Clout	Authentic	Tone	ppron	i	we	you	shehe	they	...
144564	41084	9/8/04	4:46	134	55.23	69.94	63.91	68.05	7.46	2.99	2.24	1.49	0	0.75	...
404119	128515	21/7/07	22:27	12	1	79.76	74.76	25.77	33.33	8.33	0	0	0	25	...
395992	93243	19/6/07	1:02	28	13.85	76.25	1.06	99	7.14	3.57	0	3.57	0	0	...
405421	99958	24/7/07	1:40	16	84.57	89.42	35.37	1	6.25	0	0	6.25	0	0	...
662470	185647	5/12/09	16:05	37	32.06	79.13	21.26	75.85	18.92	8.11	0	0	5.41	5.41	...
420058	53655	13/9/07	22:59	17	26.21	3.89	99	1	11.76	5.88	0	0	0	5.88	...
13933	1740	9/3/02	2:01	61	22.35	37.15	72.51	25.77	11.48	6.56	1.64	0	0	3.28	...
245087	80190	9/11/05	15:06	94	82.45	66.48	44.79	25.77	4.26	2.13	1.06	0	0	1.06	...
442550	47686	6/12/07	5:06	80	61.95	54.96	59.88	96.76	7.5	5	0	1.25	0	1.25	...
352716	26979	5/1/07	21:33	10	8.19	84.14	1	25.77	0	0	0	0	0	0	...
463617	104430	29/2/08	8:02	249	98.57	78.92	15.3	83.06	3.61	0.8	1.61	0	0.8	0.4	...
363541	-1	15/2/07	11:30	26	53.63	87.57	38.39	99	11.54	3.85	0	7.69	0	0	...
258941	44297	1/1/06	13:47	59	94.34	91.23	10.76	6.73	8.47	1.69	1.69	5.08	0	0	...
765163	54960	17/12/10	21:06	139	26.01	58.53	13.52	66.61	7.91	1.44	0.72	2.88	0	2.88	...
263152	79878	18/1/06	7:34	114	48.42	73.03	9.58	1	10.53	4.39	0	2.63	0	3.51	...
228773	166362	6/9/09	4:52	14	13.85	98.33	89.63	25.77	14.29	0	0	14.29	0	0	...
254482	83344	6/1/06	0:17	107	80.6	77.26	24.3	1	2.8	0.93	0	0.93	0	0.93	...
255544	81721	17/12/05	21:46	166	98.84	45.21	34.91	17.07	1.2	0	0.6	0.6	0	0	...
218880	22130	18/7/05	5:07	11	12.85	81.84	99	1	18.18	9.09	0	9.09	0	0	...
244912	41084	8/11/05	2:46	35	99	38.74	13.15	98.56	0	0	0	0	0	0	...
273089	-1	25/2/06	4:22	92	90.46	58.59	68.63	11.64	8.7	2.17	1.09	0	5.43	0	...
265715	38794	2/2/06	0:57	275	81.4	69.47	29.78	20.28	6.55	2.91	0.73	0.73	1.09	1.09	...
198321	21367	17/4/05	22:23	110	54.02	89.83	14.1	94.75	10.91	5.45	0	1.82	0.91	2.73	...
45244	13359	21/12/02	18:01	45	92.84	81.29	10.08	67.75	8.89	4.44	0	0	0	4.44	...
233103	70832	1/10/05	9:19	77	95.05	69.84	65.41	97.38	2.6	0	0	1.3	0	1.3	...
566748	109818	25/3/09	5:25	77	89.94	74.2	9.09	99	2.6	0	1.3	0	0	1.3	...
146671	116703	24/1/07	7:25	38	33.88	1.81	98.54	74.74	7.89	7.89	0	0	0	0	...
745917	105443	1/11/10	6:46	242	27.37	38.61	93.65	6.99	12.81	8.26	1.24	2.48	0	0.83	...
618782	165386	11/7/09	2:46	119	55.71	50	10.42	1	3.36	0.84	0	0	0.84	1.68	...
55689	19796	10/2/03	2:07	12	1	20.24	98.01	25.77	16.67	16.67	0	0	0	0	...
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...

Response to student questions

- May I know what is the purpose of set.seed(studentID)?

  > The set.seed(studentID) and getting a sample after that will generate the same subset of data for you to work with and also for someone else checking your work to be able to regenerate the same data subset.

  > This is specific to yourself since you will be using your own student ID.

Response to student questions

- What do you mean by “how active are participants”? How do you measure activity? Is it referring to the number of posts/threads within a specific timeframe, say 100 posts in Jan 2009?

  > Yes, you can measure activity as posting activity either by individuals or in total over a given time frame. You will need to choose the intervals over which you count this. Years may be too coarse, to see any detail, for example.

Response to student questions

- When looking at user activity "over months and/or years", for monthly activity, are we interested in looking at activities for a specific month of a specific year (eg: Jan 2005, Feb 2006, etc), or just specific months overall (eg: Jan, Feb, etc for all years)?

  > My assumption is that you would look at all months/years but that you might identify periods when activity was high/low or a trend (increasing/decreasing) etc.

Response to student questions

- Question b.1, does it mean particular threads vs the forum at different periods of time, or just the forum alone in general at different periods of time? Or is it up to our interpretation?
  - > It is up to your interpretation but both different points in time and different threads, and the combination of the two are of interest.
- May I know what exactly is a linguistic variable?
  - > Everything except for ThreadID, AuthorID, Date, and Time.

Data manipulation review questions

Please respond using Zoom chat...

Edgar Anderson's Iris data

50 samples from 3 species:

- Iris setosa, – virginica, – versicolor

Four features measured:

- Sepal width and length
- Petal width and length

Is it possible to distinguish species using physical measurements?

- Data is packaged with R: "iris"

http://en.wikipedia.org/wiki/Iris_flower_data_set

Print

> iris *# = print(iris)*

```
    Sepal.Length Sepal.Width Petal.Length Petal.Width    Species
1            5.1         3.5          1.4         0.2     setosa
2            4.9         3.0          1.4         0.2     setosa
3            4.7         3.2          1.3         0.2     setosa
4            4.6         3.1          1.5         0.2     setosa
5            5.0         3.6          1.4         0.2     setosa
6            5.4         3.9          1.7         0.4     setosa
7            4.6         3.4          1.4         0.3     setosa
8            5.0         3.4          1.5         0.2     setosa
9            4.4         2.9          1.4         0.2     setosa
10           4.9         3.1          1.5         0.1     setosa
11           5.4         3.7          1.5         0.2     setosa
12           4.8         3.4          1.6         0.2     setosa
13           4.8         3.0          1.4         0.1     setosa
14           4.3         3.0          1.1         0.1     setosa
15           5.8         4.0          1.2         0.2     setosa
...
```

Question 1

Predict the output from the following command:

```
> aggregate(iris[1:4], iris[5], mean)
```

(a) Matrix of column means

(b) Matrix of row means

(c) Data frame of column means by species

(d) Data frame of row means by species

Question 2

Predict the output from the following command:

```
> by(iris, iris[5], function(df) cor(df$Sepal.Length, df$Sepal.Width))
```

```
(a)    Correlation of sepal length and width
(b)    Correlation by species
(c)    Correlation by species as a table
(d)    Correlation by species as a data frame
```

Question 3

Predict the output format from the following:

> Sepal.cor <- as.data.frame(as.table(by(iris, iris[5], function(df) cor(df[1], df[2]))))

(a) Data frame 3 rows x 1 column
(b) Data frame 3 rows x 2 columns
(c) Data frame 150 rows x 1 column
(d) Data frame 150 rows x 2 columns

Question 4

Predict the output from the following command:

```
> iris[which.max(iris[,3]),]
```

(a) Row number having longest petal

(b) Row data having longest petal

(c) Row number having longest petal by species

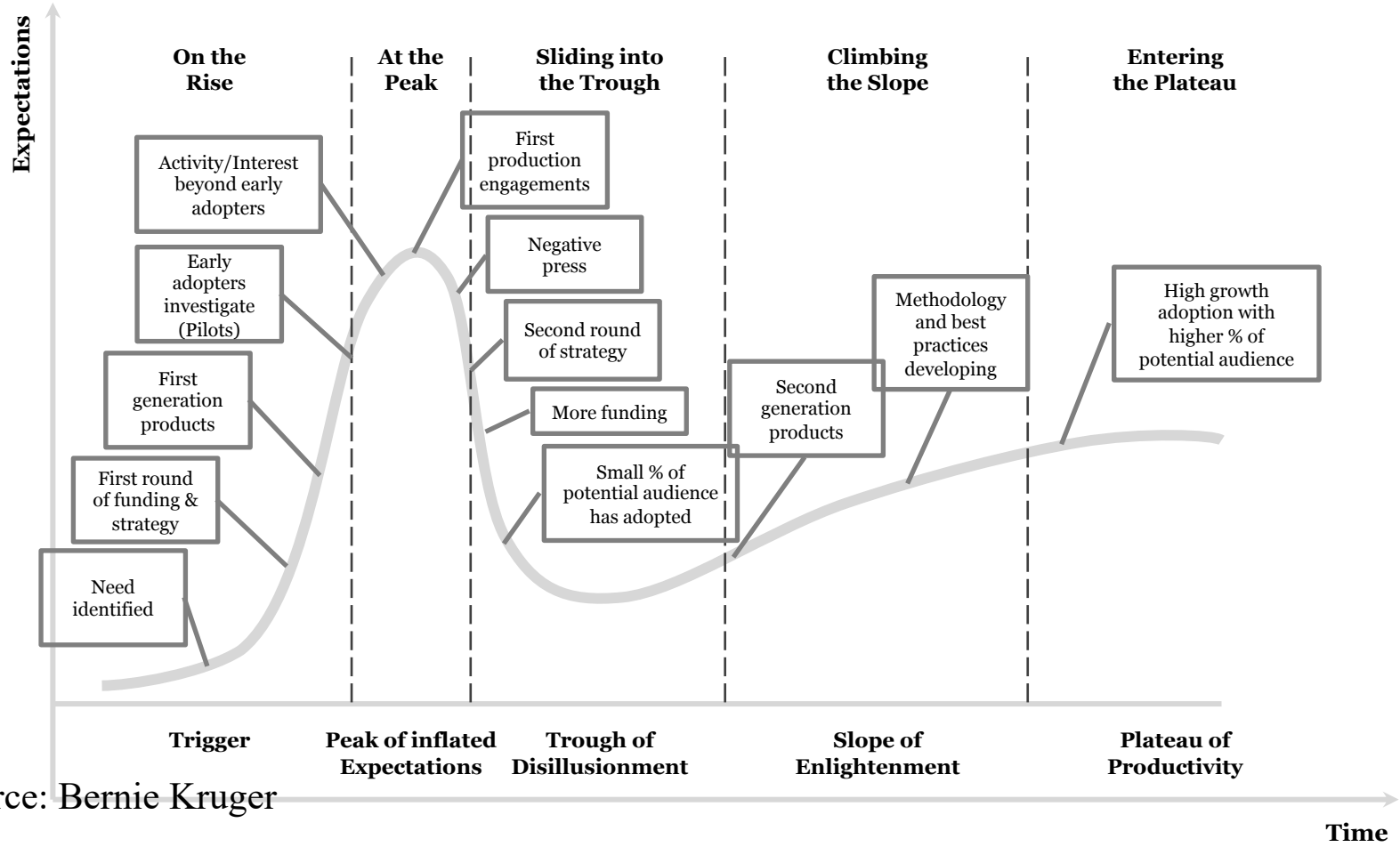
(d) Row data having longest petal by species

The data science industry

Some general thoughts, including those from a former guest speaker:

- Industry trends: The Gartner Hype Cycle
- Key skills
- How data science is being used
- Data science methodologies

Gartner Hype Cycle



Source: Bernie Kruger

Gartner Hype Cycle

How Do Hype Cycles Work?

Each Hype Cycle drills down into the five key phases of a technology's life cycle.

Innovation Trigger: A potential technology breakthrough kicks things off. Early proof-of-concept stories and media interest trigger significant publicity. Often no usable products exist and commercial viability is unproven.

Peak of Inflated Expectations: Early publicity produces a number of success stories — often accompanied by scores of failures. Some companies take action; many do not.

Trough of Disillusionment: Interest wanes as experiments and implementations fail to deliver. Producers of the technology shake out or fail. Investments continue only if the surviving providers improve their products to the satisfaction of early adopters.

Slope of Enlightenment: More instances of how the technology can benefit the enterprise start to crystallize and become more widely understood. Second- and third-generation products appear from technology providers. More enterprises fund pilots; conservative companies remain cautious.

Plateau of Productivity: Mainstream adoption starts to take off. Criteria for assessing provider viability are more clearly defined. The technology's broad market applicability and relevance are clearly paying off.

http://www.gartner.com/technology/research/methodologies/hype-cycle.jsp

Gartner Hype Cycle AI, 2021

https://medium.com/

Gartner Hype Cycle: AI

On the rise:

- General AI, Responsible AI, Human-Centered AI, Transformers (adaptable deep learning models).

At the peak:

- Smart Robots, Digital Ethics, Edge (device-based) AI.

Sliding into the trough:

- Conventional Deep Learning, Machine Learning, NLP, etc.

On the slope:

- Computer vision, autonomous vehicles, semantic search…

Data Science: Key Skills

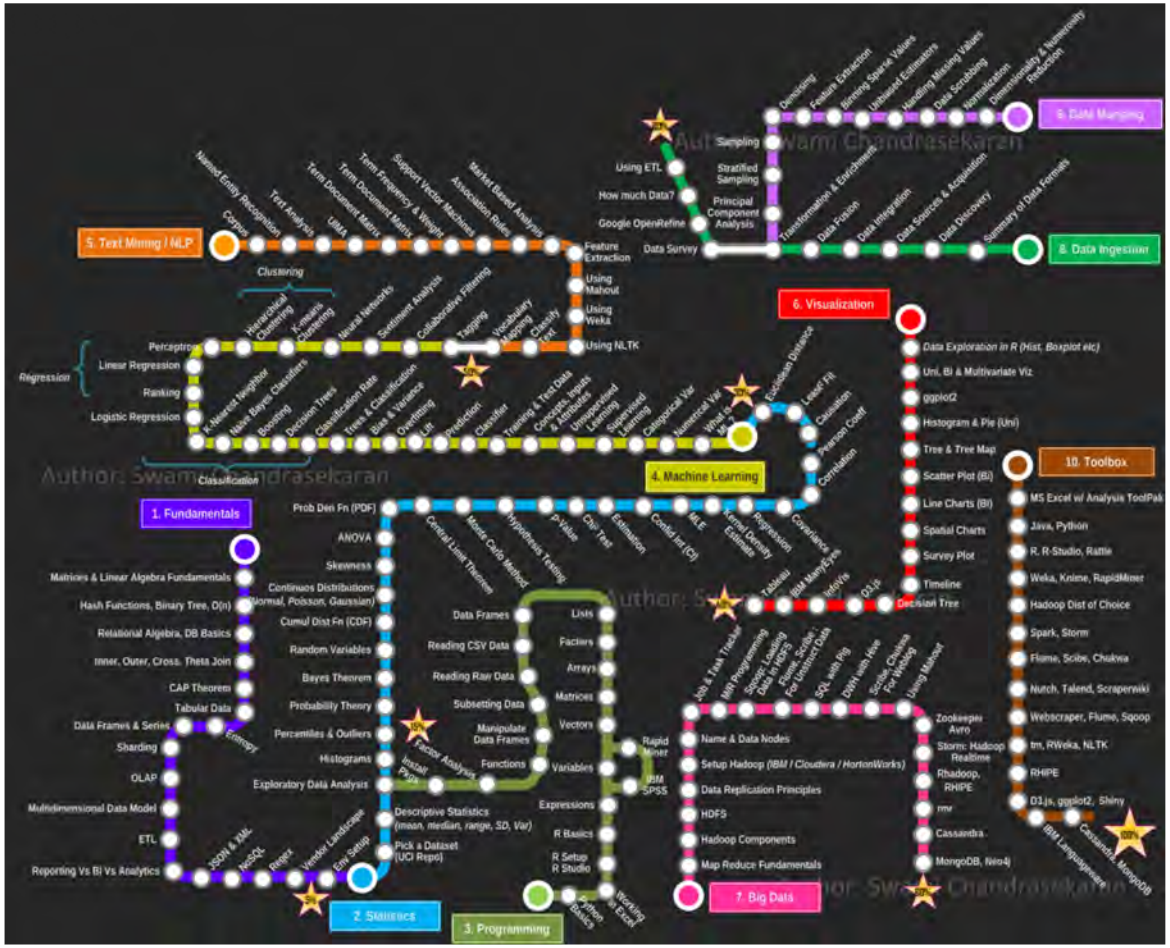
From Swami Chandrasekaran:

- Fundamentals
- Statistics
- Programming
- Machine Learning
- Text Mining/Natural Language Processing
- Data Visualization
- Big Data
- Data Ingestion
- Data Munging
- Toolbox

This is from an old (2013) post, but still quite relevant. You could add ML/AI to this…

http://nirvacana.com/thoughts/becoming-a-data-scientist/

Data Science: Key Skills



http://nirvacana.com/thoughts/becoming-a-data-scientist/

How data science is being used

Five stages of understanding:

- Descriptive: What happened?
- Diagnostic: Why did it happen?
- Explorative: What might be interesting?
- Predictive: What is likely to happen?
- Prescriptive: What can we do about it?

Typical activities for each of these stages are given on the following slides, for reading only.

How data science is being used

Descriptive: What happened?

- Condense data into smaller, more useful pieces of information;
- Standard and ad-hoc reporting;
- Dashboards, mostly static;
- Query & drill down into details; Aim of most MI/BI activities

Diagnostic: Why did it happen?

- Data analysis by employing predefined criteria;
- Rules based data analysis e.g. controls testing, suspicious transaction activity etc.;
- Essential to process rectification and improvement

How data science is being used

Explorative: What might be interesting?

- Manual (interactive dashboard);
- Automated discovery (machine learning, e.g. clustering);
- Non-rule based data discovery;
- Uncover underlying structure, patterns, anomalies

Predictive: What is likely to happen?

- Predictive modelling on historical data to produce future likelihood of events
- Prediction, e.g. Random Forest, NN, Deep Learning, GBM etc.
- Forecasting, e.g. statistical(smoothing/ trend/seasonality), advanced (Arima)

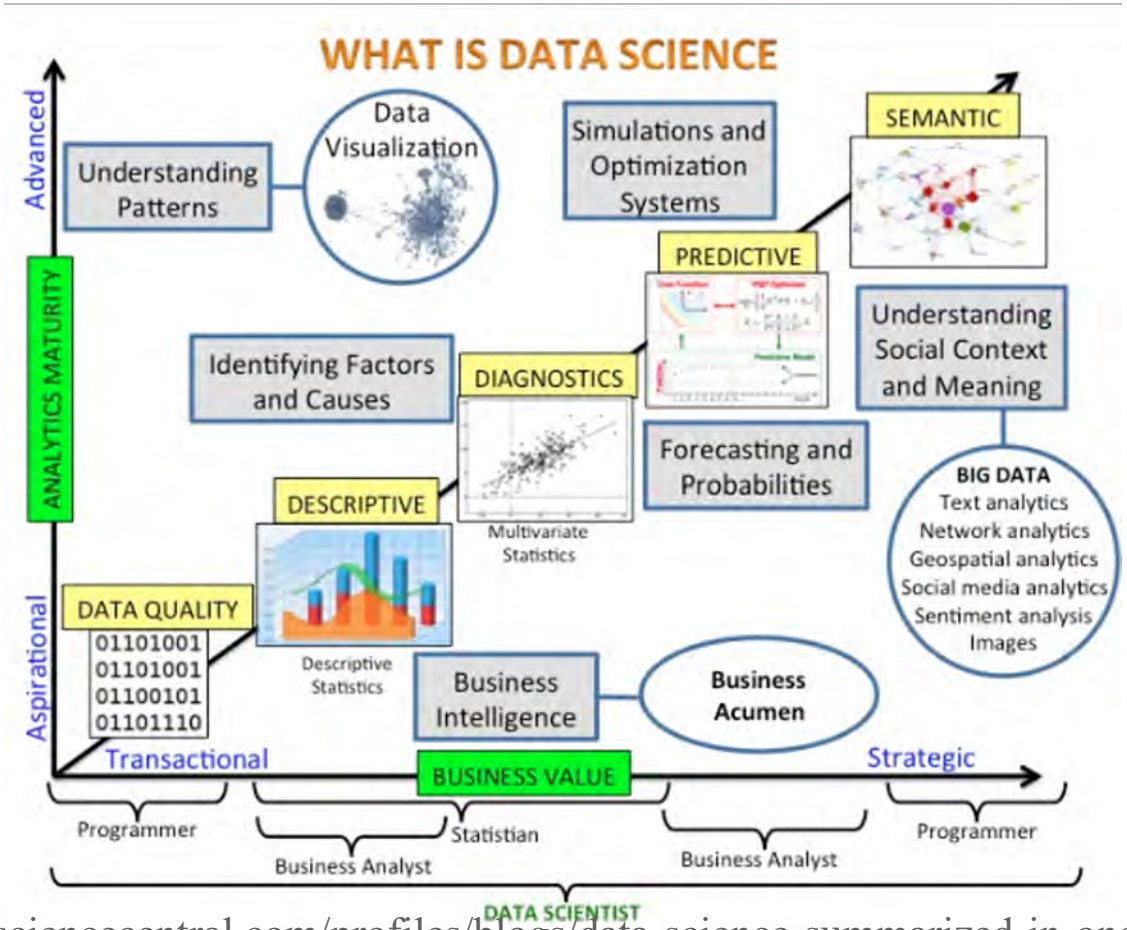
How data science is being used

Prescriptive: What can we do about it?

- Suggests the best option for handling a future scenario;
- Convergence of prior analytic activities
- Optimisation under uncertainty;
- Closed loop between analytics and process;
- Simulation e.g., (Monte Carlo, Markov Chains)

Source: Bernie Kruger

How data science is being used



https://www.datasciencecentral.com/profiles/blogs/data-science-summarized-in-one-picture

ML, AI and Big Data Landscape

Matt Turck

VC at FirstMark

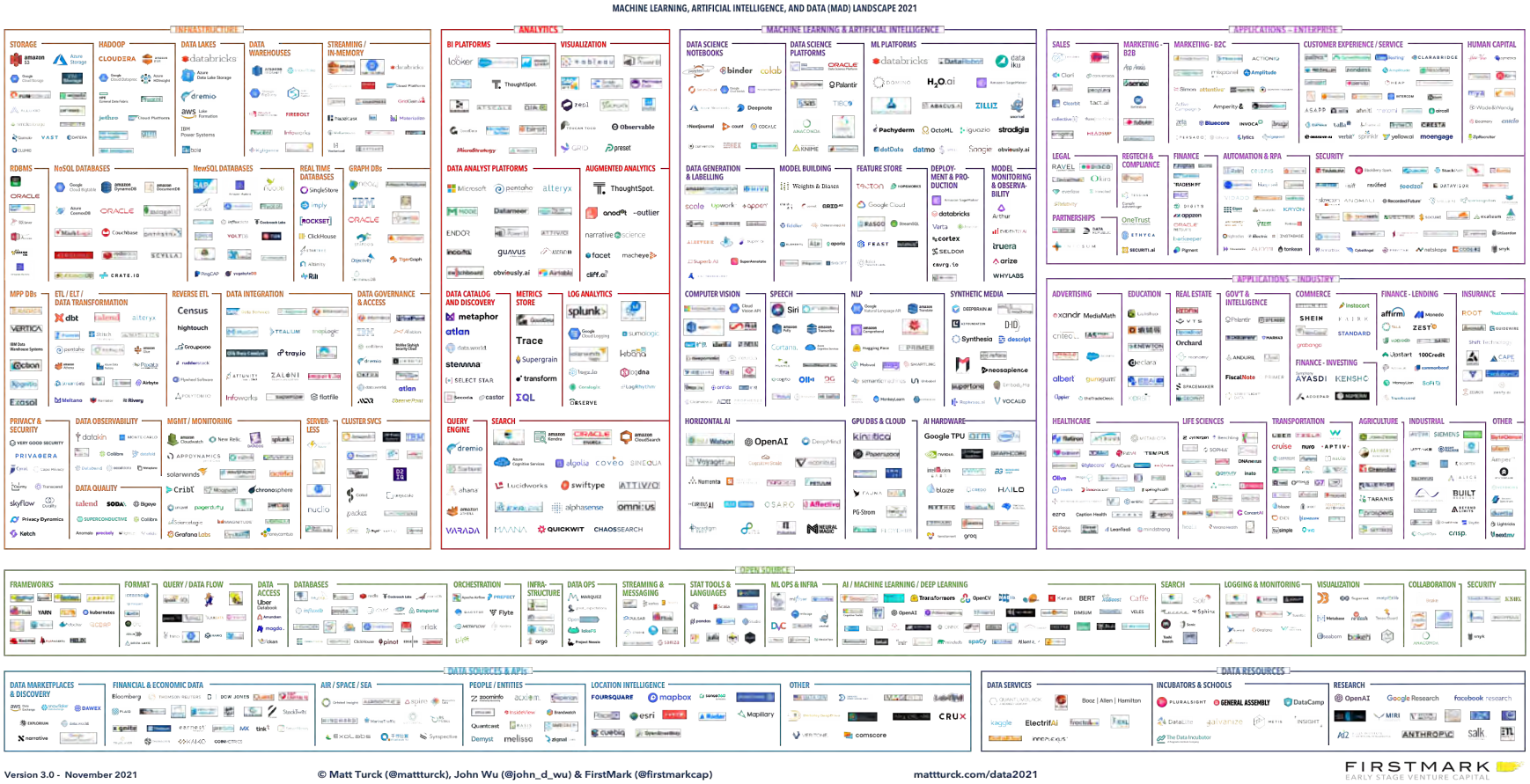
@mattturck

Interesting read if you are interested in how the area is developing.

Red Hot: The 2021 Machine Learning, AI and Data (MAD) Landscape

https://mattturck.com/data2021/

ML, AI and Big Data Landscape



https://mattturck.com/data2021/

ML, AI and Big Data Landscape

It's been a hot, hot year in the world of data, machine learning and AI.

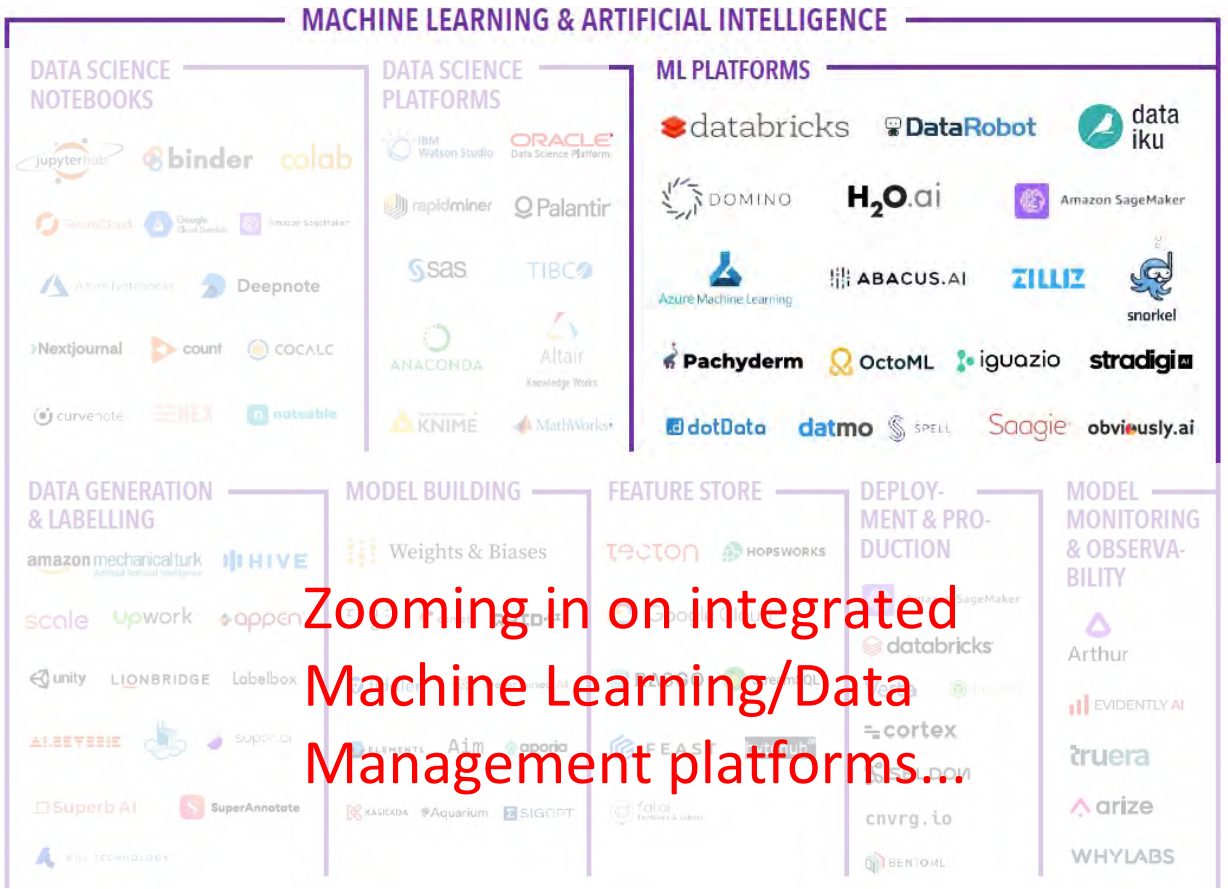
Just when you thought it couldn't grow any more explosively, the data/AI landscape just did: rapid pace of company creation, exciting new product and project launches, a deluge of VC financings, unicorn creation, IPOs, etc.

…

One story has been the maturation of the ecosystem, with market leaders reaching large scale and ramping up their ambitions for global market domination…

https://mattturck.com/data2021/

ML, AI and Big Data Landscape



https://mattturck.com/data2021/

Data science methodologies

The need for a methodology:

- There are so many options, tasks, techniques, tools, formats, and approaches to data analysis that industry specialists find it very difficult to design and implement projects.
- Although methodologies already exist, they are designed for specific software packages. Most of these methodologies use a traditional statistical approach.
- A data mining methodology to meet the specific requirements of industrial procedures is needed.

Bernie Kruger

Data science methodologies

The following slides will present three data science methodologies. Think about the following:

- What is a typical data science methodology?
- What is typical data science workflow?
- What are the key elements of a typical workflow?

Data science methodologies

KDD – (Knowledge Discovery in Databases)

- Broad process of finding knowledge in data and emphasizes the "high-level" application of particular data mining methods.

SEMMA – (Sample, Explore, Modify, Model and Assess)

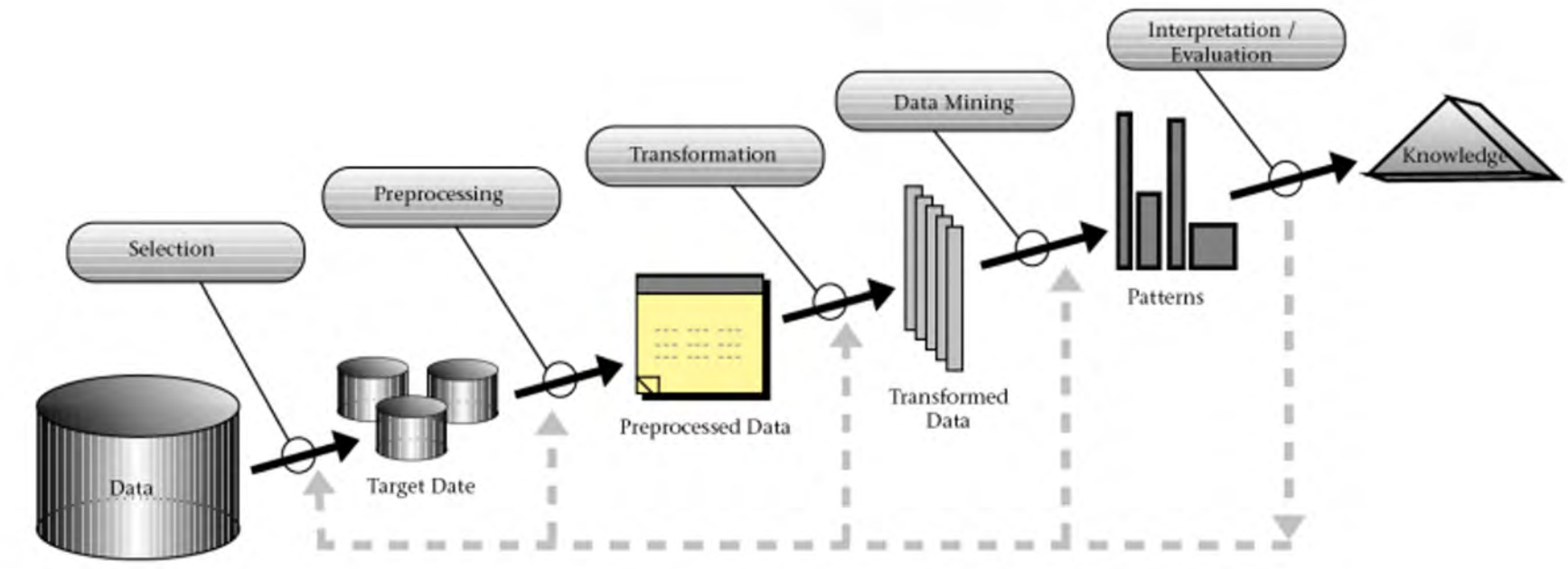
- Methodology for data mining processes proposed by the SAS Institute for the software package Enterprise Miner.

CRISP-DM – (Cross-Industry Standard Process for DM)

- Developed by a consortium of data mining vendors and companies through an effort founded by the European Commission. (CRISP-DM preferred due to inclusion of business aspects)

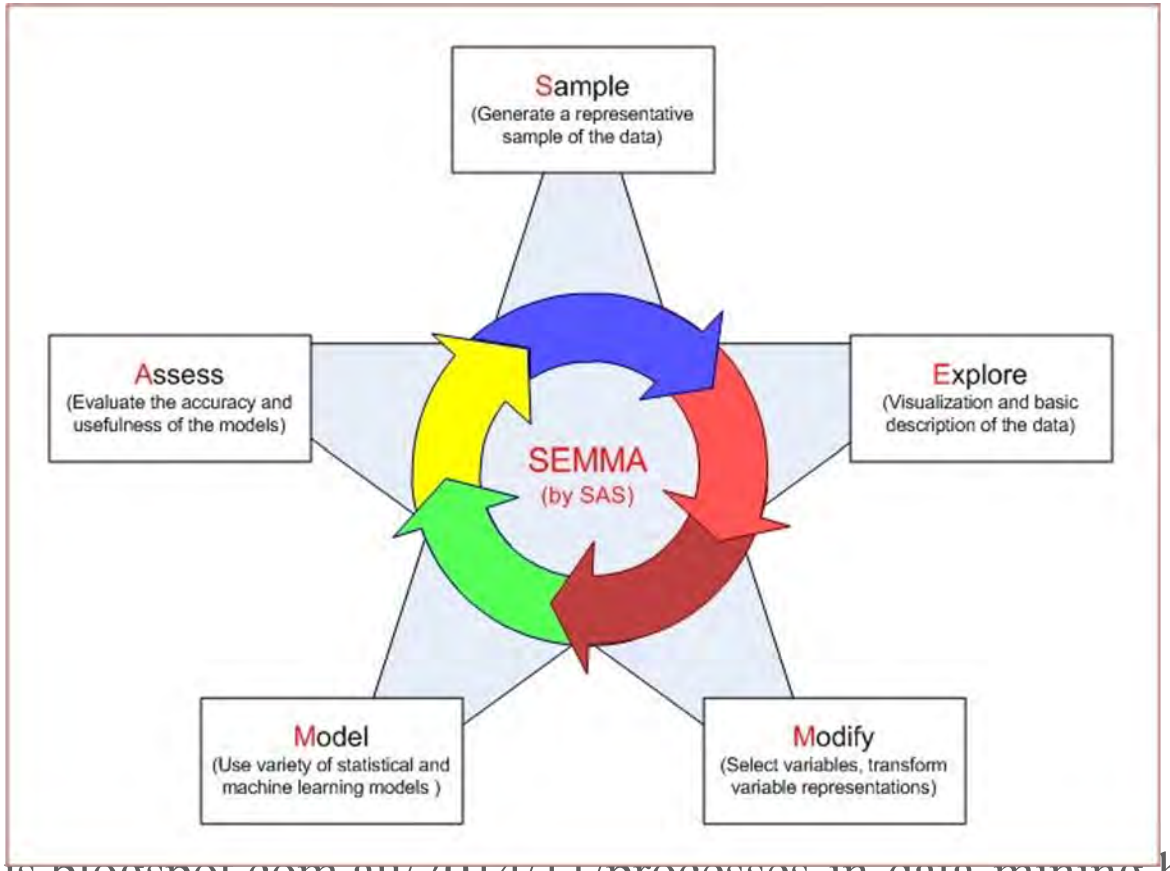
Source: Bernie Kruger

KDD



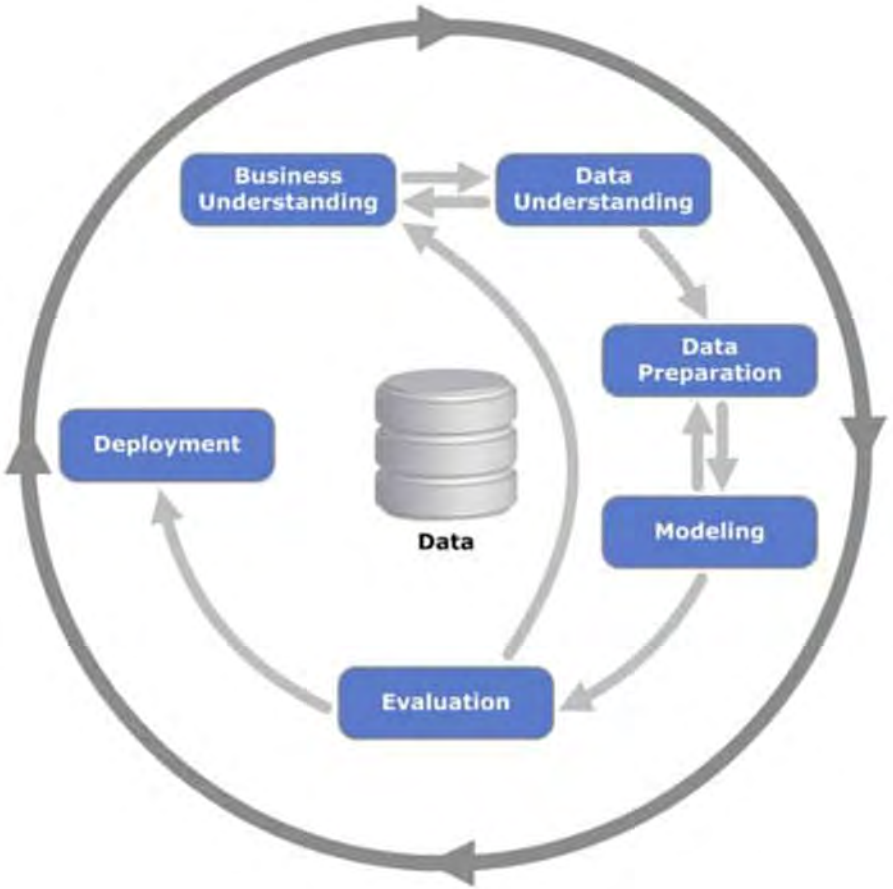
https://infovis-wiki.net/wiki/Knowledge_Discovery_in_Databases_(KDD)

SEMMA



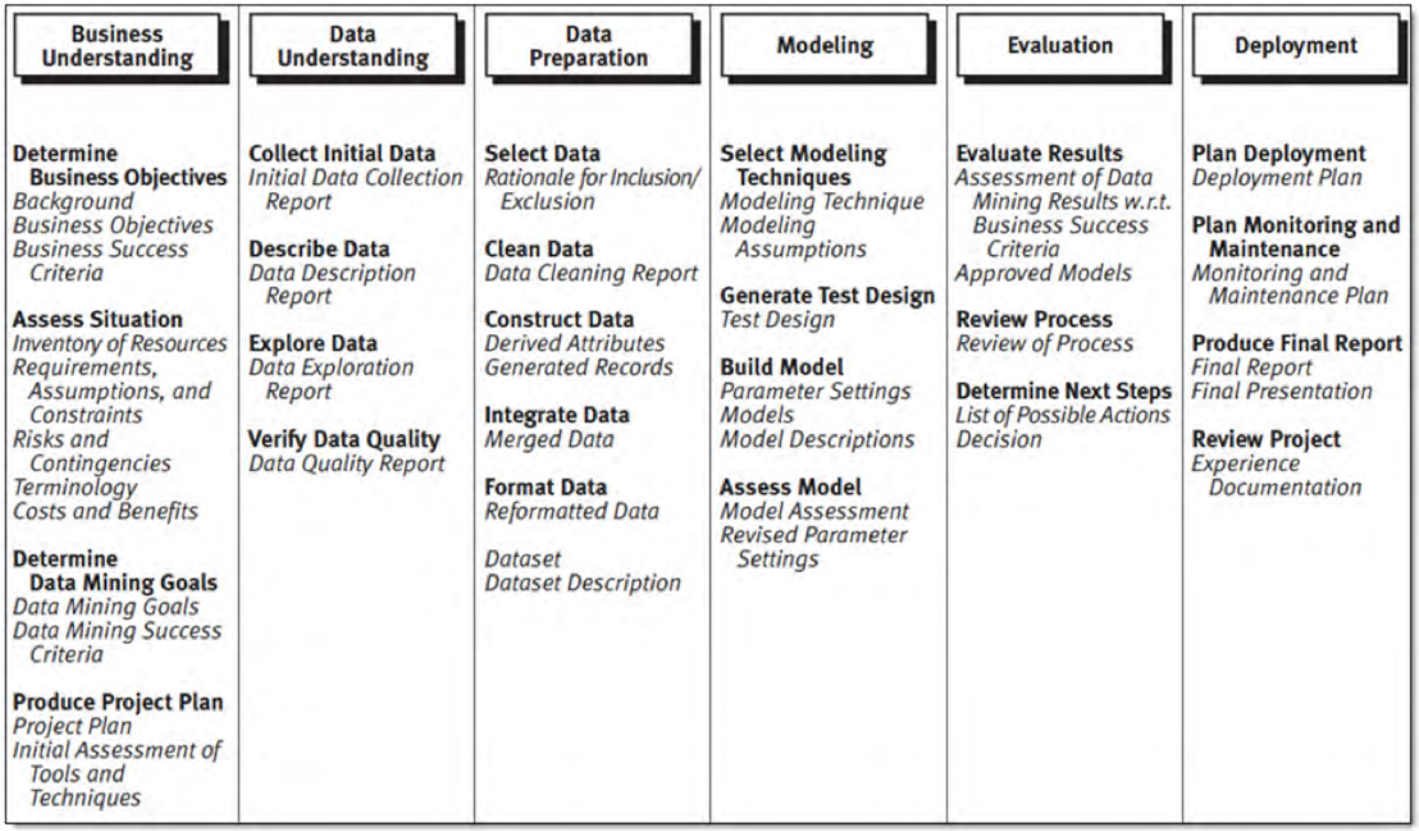
https://sisbinus.blogspot.com.au/2014/11/processes-in-data-mining.html

CRISP–DM



https://www.datascience-pm.com/crisp-dm-2/

CRISP–DM



http://www.havlena.net/en/business-analytics-intelligence/predictive-analytics-project-in-automotive-industry/

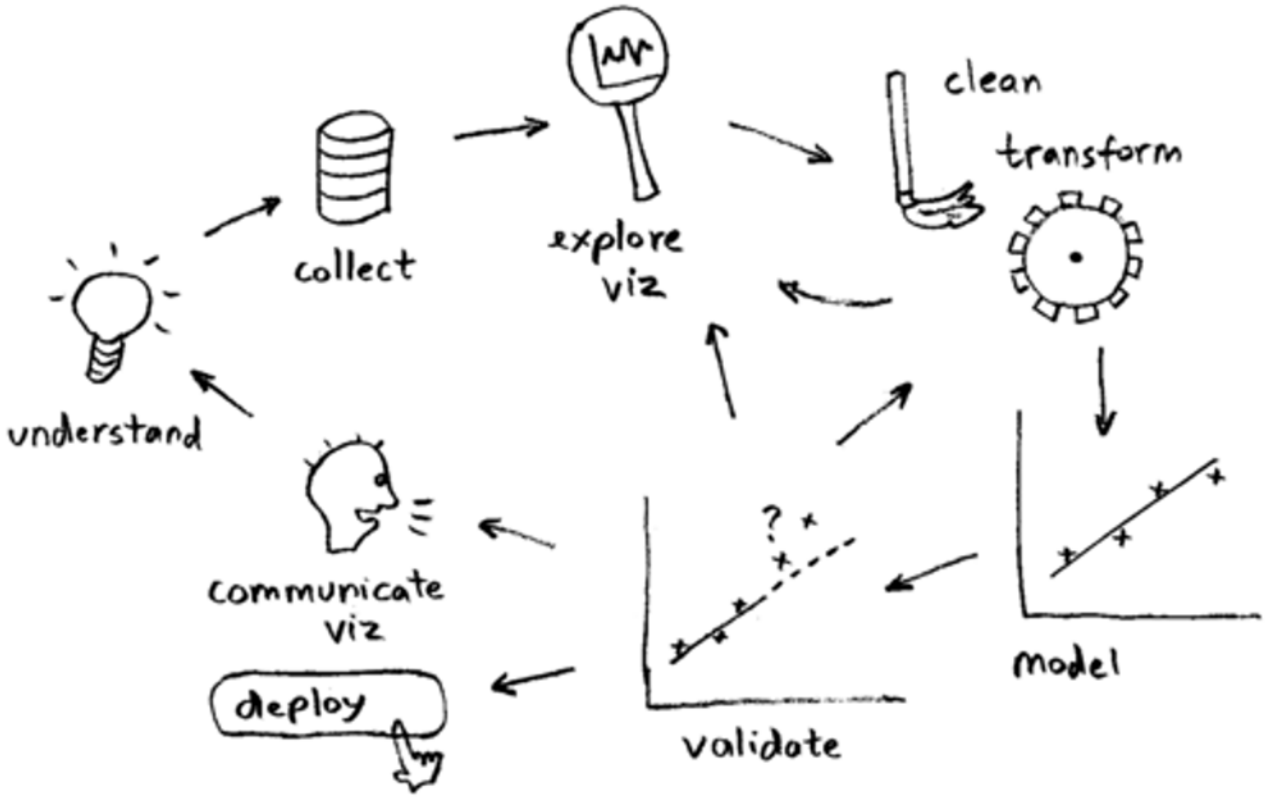
The data science workflow

What then is a typical methodology?

What is typical workflow?

Key elements of a typical workflow?

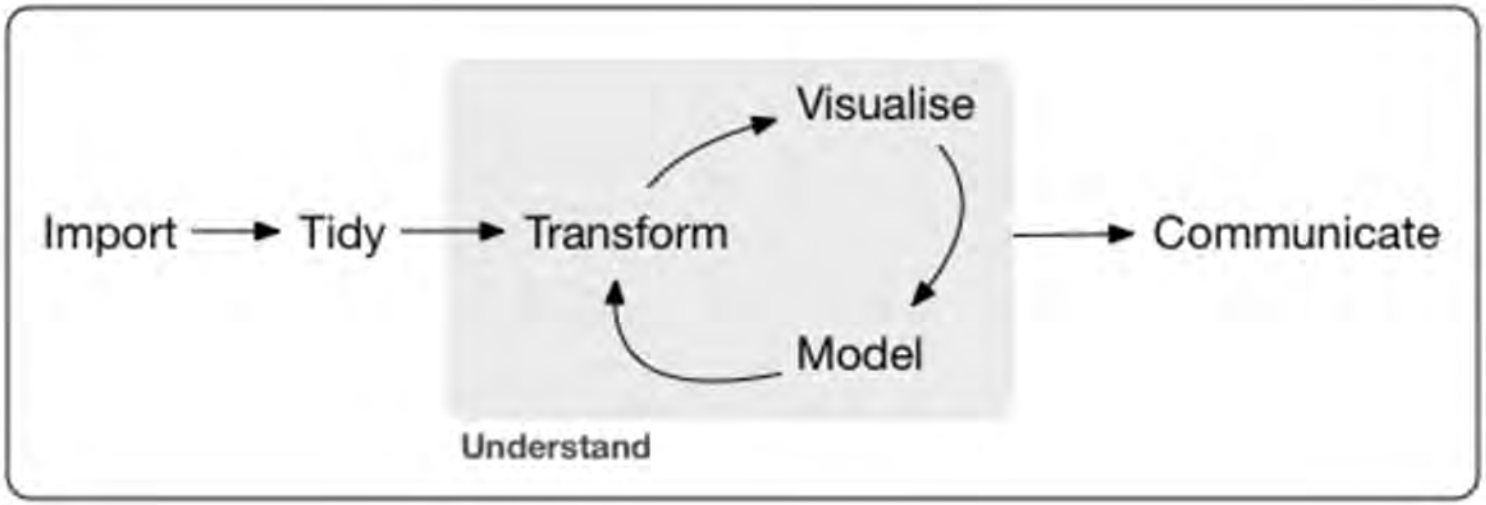
The data science workflow



http://datascience.la/data-science-toolbox-survey-results-surprise-r-and-python-win/

The data science workflow

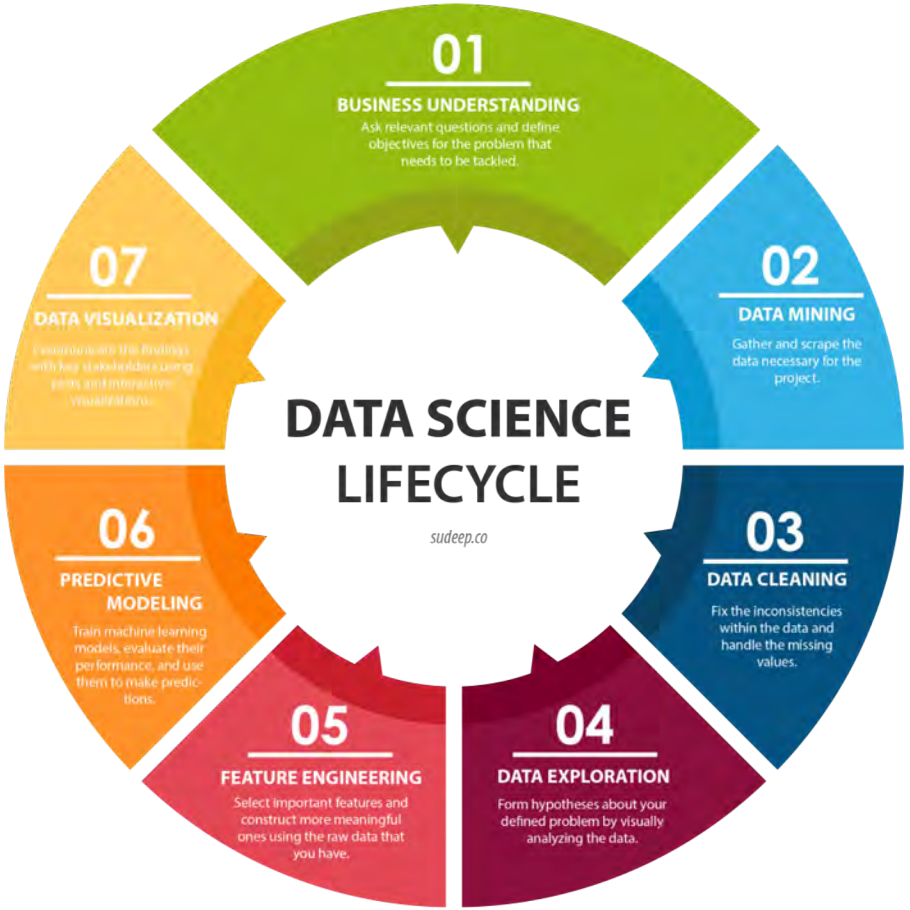
From: R for Data Science



Think about the relevance of this slide and previous slides for how You might tackle Assignment 1.

https://r4ds.had.co.nz/introduction.html

Data Science Lifecycle, another view



http://sudeep.co/data-science/Understanding-the-Data-Science-Lifecycle/

Dirty data

The following slide presents a small section of bibliographic metadata from a collection of books in the British Library.

- Think about the difficulties you would have putting this data into a standard form for analysis:

https://groups.google.com/forum/ (inactive, last accessed 2020)

Dirty data

Identifier	Edition Statement	Place of Publication	Date of Publication	Publisher	Title	Author	Contributors
206		London	1879 [1878]	S. Tinsley & Co.	Walter Forbes. [A novel.] By A. A	A. A.	FORBES, Walter.
216		London; Virtue & Yo	1868	Virtue & Co.	All for Greed. [A novel. The dedication s	A., A. A.	BLAZE DE BURY, Ma
218		London	1869	Bradbury, Evans & C	Love the Avenger. By the author of â€œ	A., A. A.	BLAZE DE BURY, Ma
472		London	1851	James Darling	Welsh Sketches, chiefly ecclesiastical, to	A., E. S.	Appleyard, Ernest Si
480	A new edition, revis	London	1857	Wertheim & Macint	[The World in which I live, and my place	A., E. S.	BROOME, John Hen
481	Fourth edition, revis	London	1875	William Macintosh	[The World in which I live, and my place	A., E. S.	BROOME, John Hen
519		London	1872	The Author	Lagonells. By the author of Darmayne (F	A., F. E.	ASHLEY, Florence Er
667		pp. 40. G. Bryan & Co: Oxford, 1898			The Coming of Spring, and other poems	A., J.|A., J.	ANDREWS, J. - Write
874		London]	1676		A Warning to the inhabitants of England	RemaÊ¿.	ADAMS, Mary.
1143		London	1679		A Satyr against Vertue. (A poem: suppos	A., T.	OLDHAM, John.
1280		Coventry	1802	Printed by J. Turner	An Account of the many and great Loan		CARTE, Samuel.|JAC
1808		Christiania	1859		Erindringer som Bidrag til Norges Histor	AALL, Jacob.	AALL, J. C.|LANGE, C
1905		Firenze	1888		Gli Studi storici in terra d'Otranto ... Fra	AAR, Ermanno - pse	S., L. G. D.|SIMONE,
1929		Amsterdam	1839, 38-54		De Aardbol. Magazijn van hedendaagsch		WITKAMP, Pieter Ha
2836		Savona	1897		Cronache Savonesi dal 1500 al 1570 ... A	ABATE, Giovanni Ag	ASSERETO, Giovann
2854		London	1865	E. Moxon & Co.	See-Saw; a novel ... Edited [or rather, w	ABATI, Francesco.	READE, William Win
2956		Paris	1860-63		Gel odel sie d'une partie de la Haute E	ABBADIE, Antoine T	RADAU, Rodolphe.
2957		Paris	1873		[With eleven maps.]	ABBADIE, Antoine T	RADAU, Rodolphe.
3017	Nueva edicion, anot	Puerto-Rico	1866		[Historia geogral fica, civil y politica de	ABBAD Y LASIERRA,	ACOSTA Y CALBO, Jo
3131		New York	1899	W. Abbatt	The Crisis of the Revolution, being the s	ABBATT, William.	ANDREl , John - Ma
4598		Hull	1814	The Author	Peace: a lyric poem. [With prefatory ad	ABBOTT, Thomas Ea	WRANGHAM, Franc
4884		London	1820	J. Hatchard & Son	Abdallah; or, The Arabian Martyr: a Chr		BARHAM, Thomas F
4976	[Another edition.] A	Oxonii	1800	J. Cooke, etc.	[Abdollatiphi HistoriÃ¦ Ã†gypti compen		WHITE, Joseph - Car
5382		London	1847, 48 [1846-48]	Punch Office	The Comic History of England ... With ...	A'BECKETT, Gilbert A	LEECH, John - Artist
5385	[Another edition.] Il	London	[1897?]	Bradbury, Agnew &	[The comic history of England ... With tv	A'BECKETT, Gilbert A	LEECH, John - Artist
5389	[Another edition.]	London	[1897?]	Bradbury, Agnew &	[The Comic History of Rome ... Illustrate	A'BECKETT, Gilbert A	LEECH, John - Artist
5432		Milano	1893		Signa: opera in tre atti [founded on the	A'BECKETT, Gilbert A	MAZZUCATO, Giova
6036		London	1805	C. & R. Baldwin	The Venetian Outlaw, a drama in three		ELLISTON, Robert W
6821		Aberdeen	1837	J. Davidson & Co.	Description of the Coast between Aberd		DUNCAN, William -

Dirty data: some concerns

Identifier	Edition Statement	Place of Publication	Date of Publication	Publisher	Title	Author	Contributors
206		London	1879 [1878]	S. Tinsley & Co.	Walter Forbes. [A novel.] By A. A	A. A.	FORBES, Walter.
216		London; Virtue & Yo	1868	Virtue & Co.	All for Greed. [A novel. The dedication s	A., A. A.	BLAZE DE BURY, Ma
218		London	1869	Bradbury, Evans & C	Love the Avenger. By the author of â€œ	A., A. A.	BLAZE DE BURY, Ma
472		London	1851	James Darling	Welsh Sketches, chiefly ecclesiastical, to	A., E. S.	Appleyard, Ernest Si
480	A new edition, revis	London	1857	Wertheim & Macint	[The World in which I live, and my place	A., E. S.	BROOME, John Hen
481	Fourth edition, revis	London	1875	William Macintosh	[The World in which I live, and my place	A., E. S.	BROOME, John Hen
519		London	1872	The Author	Lagonells. By the author of Darmayne (F	A., F. E.	ASHLEY, Florence Er
667		pp. 40. G. Bryan & Co: Oxford, 1898			The Coming of Spring, and other poems	A., J.|A., J.	ANDREWS, J. - Write
874		London]	1676		A Warning to the inhabitants of England	ReMaÊ¿.	ADAMS, Mary.
1143		London	1679		A Satyr against Vertue. (A poem: suppos	A., T.	OLDHAM, John.
1280		Coventry	1802	Printed by J. Turner	An Account of the many and great Loan		CARTE, Samuel.|JAC
1808		Christiania	1859		Erindringer som Bidrag til Norges Histor	AALL, Jacob.	AALL, J. C.|LANGE, C
1905		Firenze	1888		Gli Studi storici in terra d'Otranto ... Fra	AAR, Ermanno - pse	S., L. G. D.|SIMONE,
1929		Amsterdam	1839, 38-54		De Aardbol. Magazijn van hedendaagsch		WITKAMP, Pieter Ha
2836		Savona	1897		Cronache Savonesi dal 1500 al 1570 ... A	ABATE, Giovanni Ag	ASSERETO, Giovann
2854		London	1865	E. Moxon & Co.	See-Saw; a novel ... Edited [or rather, w	ABATI, Francesco.	READE, William Win
2956		Paris	1860-63		Gel odel sie d'une partie de la Haute E	ABBADIE, Antoine T	RADAU, Rodolphe.
2957		Paris	1873		[With eleven maps.]	ABBADIE, Antoine T	RADAU, Rodolphe.
3017	Nueva edicion, anot	Puerto-Rico	1866		[Historia geogral fica, civil y politica de	ABBAD Y LASIERRA,	ACOSTA Y CALBO, Jo
3131		New York	1899	W. Abbatt	The Crisis of the Revolution, being the s	ABBATT, William.	ANDREI , John - Ma
4598		Hull	1814	The Author	Peace: a lyric poem. [With prefatory ad	ABBOTT, Thomas Ea	WRANGHAM, Franc
4884		London	1820	J. Hatchard & Son	Abdallah; or, The Arabian Martyr: a Chr		BARHAM, Thomas F
4976	[Another edition.] A	Oxonii	1800	J. Cooke, etc.	[Abdollatiphi HistoriÃ¦ Ã†gypti compen		WHITE, Joseph - Car
5382		London	1847, 48 [1846-48]	Punch Office	The Comic History of England ... With ...	A'BECKETT, Gilbert A	LEECH, John - Artist
5385	[Another edition.] Il	London	[1897?]	Bradbury, Agnew &	[The comic history of England ... With tv	A'BECKETT, Gilbert A	LEECH, John - Artist
5389	[Another edition.]	London	[1897?]	Bradbury, Agnew &	[The Comic History of Rome ... Illustrate	A'BECKETT, Gilbert A	LEECH, John - Artist
5432		Milano	1893		Signa: opera in tre atti [founded on the	A'BECKETT, Gilbert A	MAZZUCATO, Giova
6036		London	1805	C. & R. Baldwin	The Venetian Outlaw, a drama in three		ELLISTON, Robert W
6821		Aberdeen	1837	J. Davidson & Co.	Description of the Coast between Aberd		DUNCAN, William -

Dirty data

From: A Taxonomy of Dirty Data

Today large corporations are constructing enterprise data warehouses from disparate data sources in order to run enterprise-wide data analysis applications, including decision support systems, multidimensional online analytical applications, data mining, and customer relationship management systems. A major problem that is only beginning to be recognized is that the data in data sources are often "dirty". Broadly, dirty data include missing data, wrong data, and non-standard representations of the same data. The results of analyzing a database/data warehouse of dirty data can be damaging and at best be unreliable. In this paper, a comprehensive classification of dirty data is developed for use as a framework for understanding how dirty data arise, manifest themselves, and may be cleansed to ensure proper construction of data warehouses and accurate data analysis. The impact of dirty data on data mining is also explored.

https://link.springer.com/article/10.1023/A:1021564703268

Dirty data

Data in the real world is dirty, it can be:

- Incorrect:
- Inaccurate
- Incomplete
- Duplicate
- Violate business rules
- Inconsistent
- Non-integrated
- ...

Dirty data

Incorrect data:

- For data to be correct (valid), its values must adhere to its domain (valid values). E.g. a month must be in the range of 1-12, or a person's age must be less than 130.

Inaccurate data:

- A data value can be correct without being accurate. For example, the state code "VIC" and the city name "Sydney" are both correct, but when used together (such as Sydney, VIC), the state code is wrong because Sydney is in NSW.
- **Note**: this data is also **inconsistent**.

Dirty data

Business rule violations:

- Another type of inaccurate data value is one that violates business rules. For example, a start date should always precede a finish date.

Inconsistent data:

- Uncontrolled data redundancy results in inconsistencies. For example: a customer name may be recorded on three different databases as: Mary Smith, Maria Louise Smith, and Mary L. Smith.

Dirty data

Incomplete data:

- During system requirements definition, we rarely gather the data requirements from down-stream information consumers (e.g. marketing department). If we build a system for the lending department of a bank, the users of that department will most likely list: Initial Loan Amount, Monthly Payment Amount and Interest Rate as some of the most critical data elements. However, the most important data elements for users of the marketing department are probably: Gender, Customer code or Postcode that might not be captured at all or only haphazardly.

Dirty data

Non-integrated data:

- Most organisations store data redundantly and inconsistently across many systems, which were never designed with integration or analytics in mind.
- Primary keys often don't match or are not unique and in some cases, they don't even exist. For example, customer data may exist on two or more outsourced systems under different customer numbers with different spellings of the customer name and even different phone numbers or addresses.

Source: Bernie Kruger

Tidy data

From: Tidy data

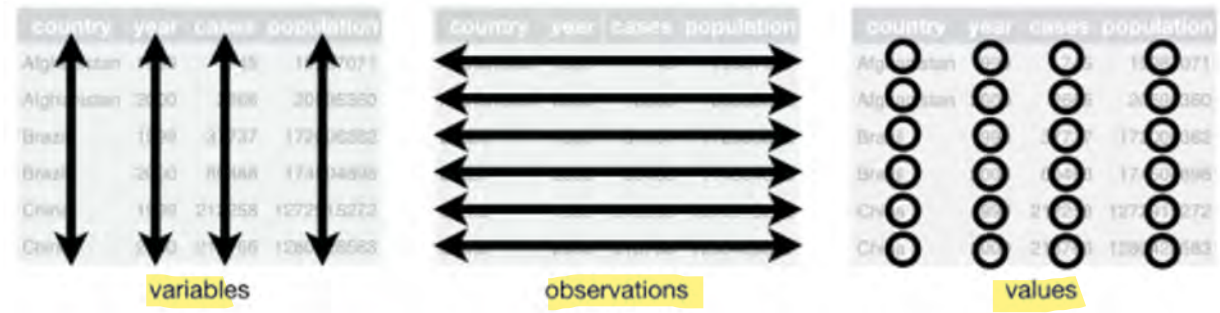
A huge amount of effort is spent cleaning data to get it ready for analysis, but there has been little research on how to make data cleaning as easy and effective as possible. This paper tackles a small, but important, component of data cleaning: data tidying. Tidy datasets are easy to manipulate, model and visualize, and have a specific structure: each variable is a column, each observation is a row, and each type of observational unit is a table. This framework makes it easy to tidy messy datasets because only a small set of tools are needed to deal with a wide range of un-tidy datasets. This structure also makes it easier to develop tidy tools for data analysis, tools that both input and output tidy datasets. The advantages of a consistent data structure and matching tools are demonstrated with a case study free from mundane data manipulation chores.

http://www.jstatsoft.org/v59/i10/paper

Tidy data

Tidy data seeks a consistent format that has:

- Each variable in its own column.
- Each observation in its own row.
- Each value in its own cell.



- Two benefits: consistency, exploits R's vector nature

https://www.jstatsoft.org/v59/i10/paper

Tidy data

Which of the two tables below would make it easier to evaluate two treatments (a and b)?

	treatmenta	treatmentb
John Smith	—	2
Jane Doe	16	11
Mary Johnson	3	1

Table 1: Typical presentation dataset.

	John Smith	Jane Doe	Mary Johnson
treatmenta	—	16	3
treatmentb	2	11	1

Table 2: The same data as in Table 1 but structured differently.

https://www.jstatsoft.org/v59/i10/paper

Tidy data

This data format is preferred as each observation is in a separate row, indexed by level (treatment).

person	treatment	result
John Smith	a	—
Jane Doe	a	16
Mary Johnson	a	3
John Smith	b	2
Jane Doe	b	11
Mary Johnson	b	1

Table 3: The same data as in Table 1 but with variables in columns and observations in rows.

http://www.jstatsoft.org/v59/i10/paper

Tidy data

How would you tidy the following?

country	year	m014	m1524	m2534	m3544	m4554	m5564	m65	mu	f014
AD	2000	0	0	1	0	0	0	0	—	—
AE	2000	2	4	4	6	5	12	10	—	3
AF	2000	52	228	183	149	129	94	80	—	93
AG	2000	0	0	0	0	0	0	1	—	1
AL	2000	2	19	21	14	24	19	16	—	3
AM	2000	2	152	130	131	63	26	21	—	1
AN	2000	0	0	1	2	0	0	0	—	0
AO	2000	186	999	1003	912	482	312	194	—	247
AR	2000	97	278	594	402	419	368	330	—	121
AS	2000	—	—	—	—	1	1	—	—	—

Table 9: Original TB dataset. Corresponding to each 'm' column for males, there is also an 'f' column for females, `f1524`, `f2534` and so on.

http://www.jstatsoft.org/v59/i10/paper

tidyr

tidyr is a package to help create tidy data, and is part of the "Tidyverse", see https://tidyr.tidyverse.org/

Recall, tidy data objectives:

- Every column is a variable.
- Every row is an observation.
- Every cell is a single value.

Cheat sheet: https://github.com/rstudio/cheatsheets…pdf

```
> install.packages("tidyverse") # or
> install.packages("tidyr")
```

tidyr

Example, reformat the iris data set into a single data column (long) format where “Species” and the “Attribute” are factors and data is in a single column:

```
> rm(list = ls())
> library(“tidyr”)
> niris = iris %>% pivot_longer(cols = 1:4, names_to = "Attribute", values_to = "Value")
```

tidyr

Long format:

> head(niris)

```
# A tibble: 6 × 3
  Species Attribute    Value
  <fct>   <chr>        <dbl>
1 setosa  Sepal.Length   5.1
2 setosa  Sepal.Width    3.5
3 setosa  Petal.Length   1.4
4 setosa  Petal.Width    0.2
5 setosa  Sepal.Length   4.9
6 setosa  Sepal.Width    3
```

Many other formatting options are possible…

Tidy data

There are many other actions that may be needed to clean and tidy data sets, including:

- Replacing missing values
- Standardisation
- Normalisation
- ...

We'll cover some of these throughout the course...

Transforming data

Data sets can be transformed in many ways. Too many to cover comprehensively.

We will look at some methods for organising, and reducing the size of a data set to isolate the key data required to answer a specific question.

Transforming data

4 Challenges, we will:

- Recode data by creating a new index
- Extract a subset of data based on values in a range.
- Extract a subset of data based on values in a second data frame.
- Display the effect of two variables on a third using a Heatmap.
- Note: there are many ways to achieve these transformations so be prepared to try alternative methods and packages...

Edgar Anderson's Iris data

50 samples from 3 species:

- Iris setosa, – virginica, – versicolor

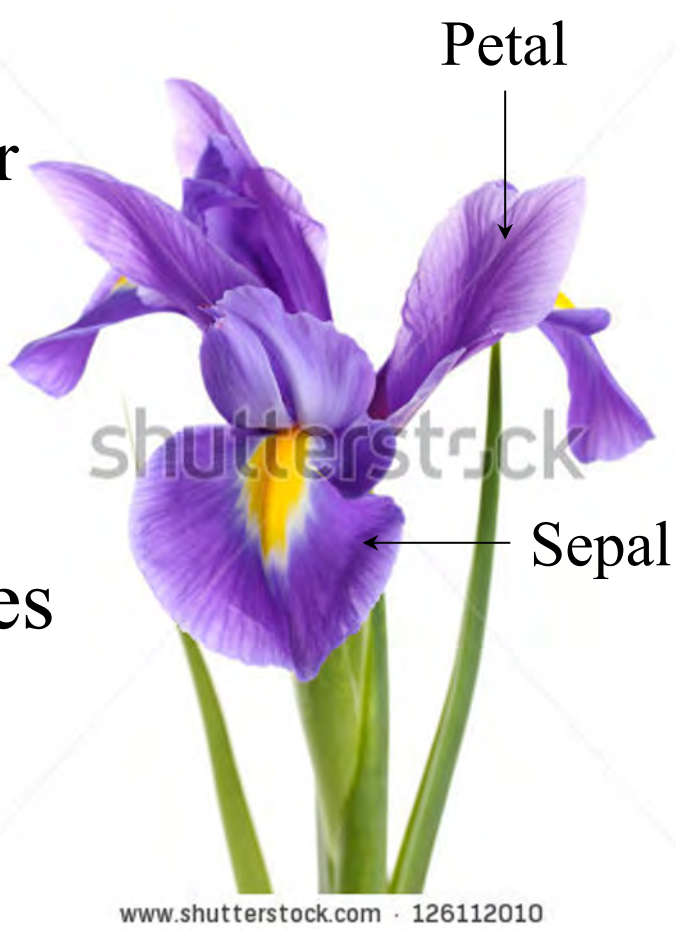
Four features measured:

- Sepal width and length
- Petal width and length

Is it possible to distinguish species using physical measurements?

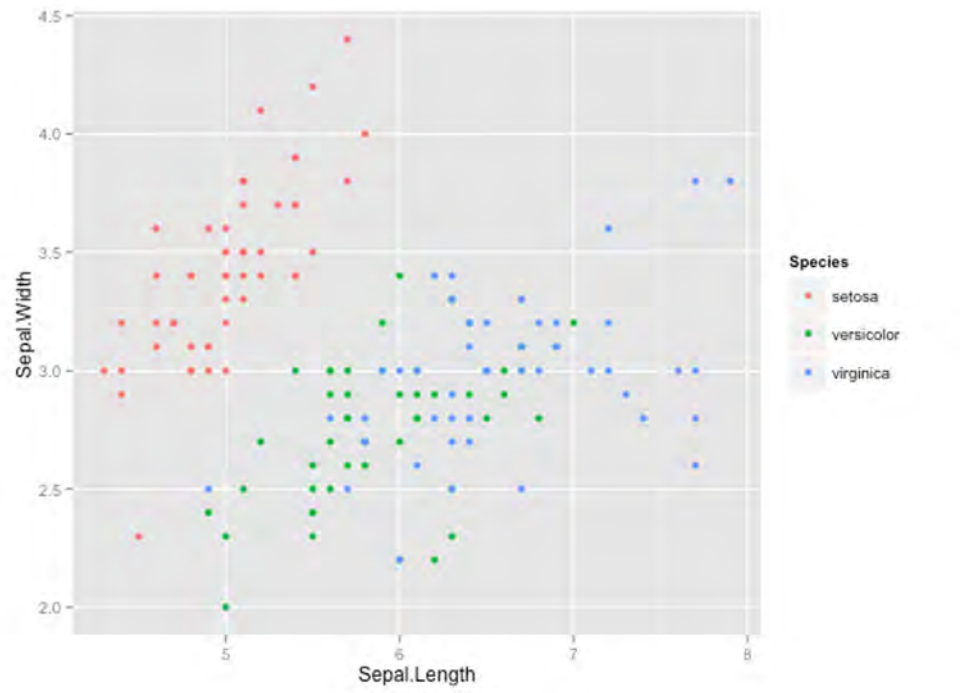
- Data is packaged with R: "iris"

http://en.wikipedia.org/wiki/Iris_flower_data_set



Challenge 1: recoding and indexing

Does Iris setosa have an average sepal width greater than I.versicolor and virginica combined?



Challenge 1:

To compare I.setosa against the other two species, we need to create a new index as a column that groups I.versicolor and virginica.

- Note: use the function “recode” from the “car” package
  - > niris = iris # clone iris data
  - > install.packages("car")
  - > library(car)

Challenge 1:

```
> ...
> niris$vvs = recode(niris$Species," 'versicolor' =
  '0';'virginica' = '0';'setosa' = '1' ")
> nirisprint(niris[c(1,51,101),]))
```

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species	vvs
1	5.1	3.5	1.4	0.2	setosa	1
51	7.0	3.2	4.7	1.4	versicolor	0
101	6.3	3.3	6.0	2.5	virginica	0

Challenge 1:

```
> ...
> t.test(niris$Sepal.Width~niris$vvs, alternative = "less")
```

```
    Welch Two Sample t-test
data:  niris$Sepal.Width by niris$vvs
t = -8.8121, df = 87.596, p-value = 5.177e-14
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
      -Inf -0.451108
sample estimates:
mean in group 0 mean in group 1
          2.872           3.428
```

Challenge 2: extracting subsets of data

The Dunnhumby data (Tutorial 2) records the sale date and amount spent by 20 customers.

Plot histograms of the six top-spending customers from June – December in the first year of the survey.

But first, a note on dates and times...

Dunnhumby: data

	customer_id	visit_date	visit_delta	visit_spend
1	40	04-04-10	NA	44.83
2	40	06-04-10	2	69.68
3	40	19-04-10	13	44.61
4	40	01-05-10	12	30.39
5	40	02-05-10	1	60.73
6	40	12-05-10	10	50.00
7	40	15-05-10	3	3.00
8	40	18-05-10	3	36.89
9	40	19-05-10	1	9.07
10	40	23-05-10	4	14.01
11	40	26-05-10	3	16.97
12	40	31-05-10	5	8.69
13	40	01-06-10	1	18.32
14	40	04-06-10	3	52.13
15	40	05-06-10	1	44.88

Without date conversion

Calculating minimums without date conversion:

> min.type <- by(DH, DH[1], function(df)
df[which.min(df[,2]),])

> do.call(rbind,min.type)

.	customer_id	visit_date	visit_delta	visit_spend
40	40	01-05-10	12	30.39
79	79	01-01-11	9	81.70
119	119	01-03-11	3	10.69
123	123	01-02-11	4	35.20
134	134	01-02-11	1	54.77

With date conversion

Calculating minimums using date conversion:

```
> min.type <- by(DH, DH[1], function(df)
  df[which.min(as.Date(df[,2],"%d-%m-%y")),])
> do.call(rbind,min.type)
```

Specify date format used.

.	customer_id	visit_date	visit_delta	visit_spend
40	40	04-04-10	NA	44.83
79	79	07-04-10	NA	150.87
119	119	01-04-10	NA	20.00
123	123	02-04-10	NA	66.94
134	134	01-04-10	NA	50.32

Challenge 2: extracting subsets of data

We are going to analyse the spending pattern of the six top-spending customers.

What are the steps we need to follow to analyse:

- By customer ID, ‘by hand’?
- Time period?
- Top spenders?

Extracting ‘by hand’

To study a particular customer, you can just create a subset of the original data by hand.

For example, to analyse Customer #40 only:

```
> DH40 = DH[(DH$customer_id == 40),]
> head(DH40)
# A tibble: 6 x 4
customer_id visit_date visit_delta visit_spend
<int>       <chr>    <int>       <dbl>
1  40    04-04-10        NA      44.83
2  40    06-04-10         2      69.68
3  ...
```

Challenge 2:

Setup environment (using script to set working directory):

```
> rm(list = ls()) # Empty the environment...
> library(readr)
> library(ggplot2)
> DH <- read_csv("Dunnhumby1-20.csv")
```

Challenge 2:

Calculate range over which data collected:

```
> # Find the earliest and latest dates recorded
> DH[which.min(as.Date(DH$visit_date,"%d-%m-%y")),]
> DH[which.max(as.Date(DH$visit_date,"%d-%m-%y")),]

...
1          119   01-04-10          NA          20
1          134   31-03-11           2       39.75
```

Challenge 2:

Extract sales from 1 June to 31 Dec 2010 as a new data frame DHX:

```
> DHX = DH[as.Date(DH$visit_date,"%d-%m-%y") >
  as.Date("31-05-10","%d-%m-%y"),]
> DHX = DHX[as.Date(DHX$visit_date,"%d-%m-%y") <
  as.Date("01-01-11","%d-%m-%y"),]
```

Challenge 2:

Create a table calculating total spend for each customer:

```
> attach(DHX)
> CustSpend = as.table(by(visit_spend, customer_id, sum))
> CustSpend
   40      79     119     123     134       ...
1668.64 2395.72  986.97 4333.02 4722.42 ...
```

Challenge 2:

Sort table, retain six top-spending customers, convert to data frame and tidy column names:

```
> CustSpend = sort(CustSpend, decreasing = TRUE)
> CustSpend = head(CustSpend, 6)

> CustSpend = as.data.frame(CustSpend)
> colnames(CustSpend) = c("customer_id", "amtspent")
```

Challenge 3:

Extract the top six customers from DHX data frame (using CustSpend data frame) and rename as DHX6:

```
> DHX6 = DHX[(DHX$customer_id %in%
  CustSpend$customer_id),]
```

Challenge 3:

CustSpend and DHX6 tables:

	customer_id	amtspent
1	140	4873.97
2	134	4722.42
3	123	4333.02
4	263	3120.58
5	254	3067.80
6	199	2737.57

	customer_id	visit_date	visit_delta	visit_spend
240	123	02-06-10	3	25.00
241	123	05-06-10	3	63.29
242	123	06-06-10	1	40.19
243	123	09-06-10	3	18.25
244	123	11-06-10	2	79.26
245	123	15-06-10	4	20.52
246	123	18-06-10	3	110.57
247	123	21-06-10	3	4.85
248	123	22-06-10	1	63.81
249	123	25-06-10	3	96.39
250	123	29-06-10	4	43.84
251	123	02-07-10	3	83.16
252	123	07-07-10	5	26.39

Challenge 3:

Now plot histogram

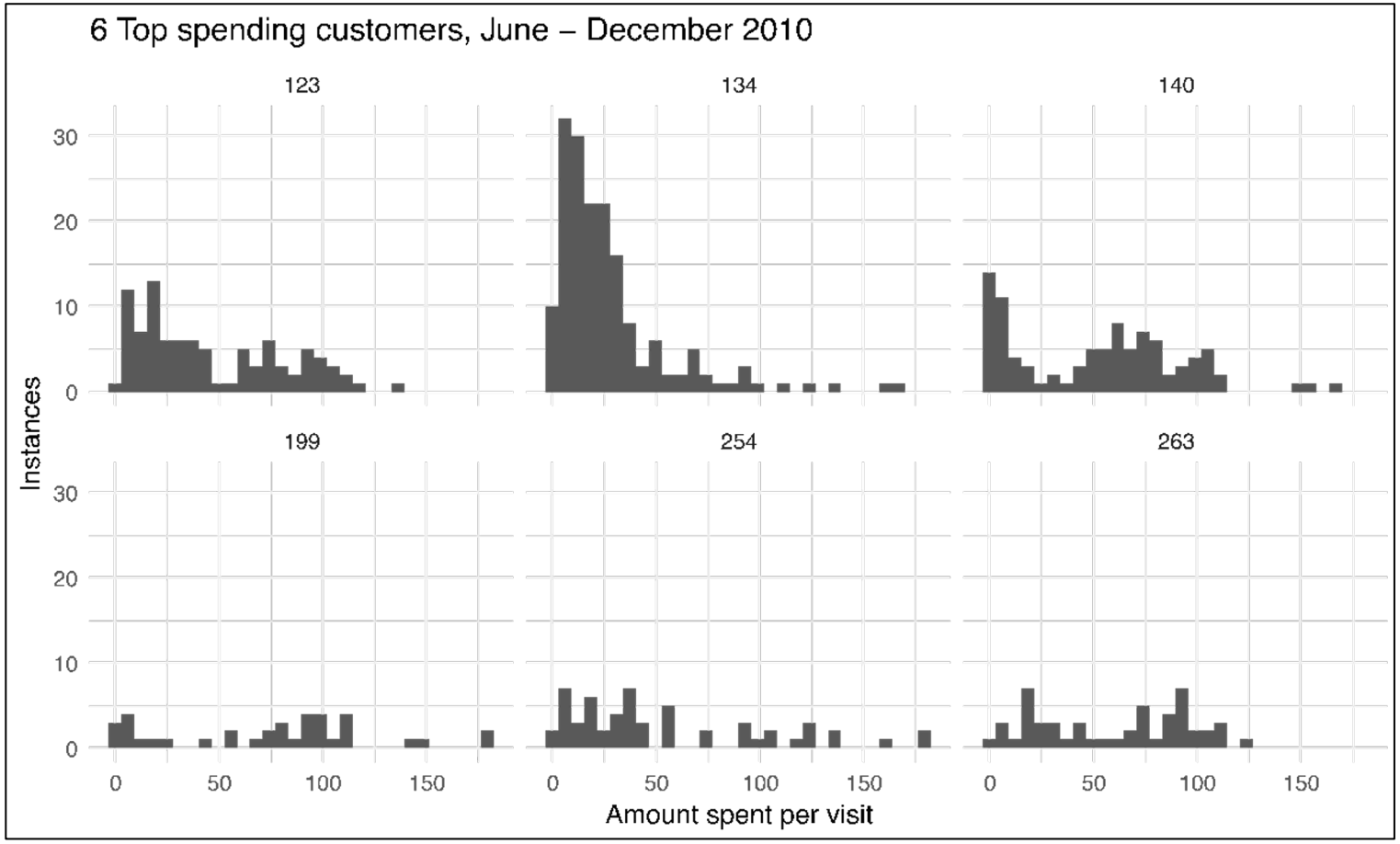
```
>   g  = ggplot(data = DHX6) +
>       geom_histogram(mapping = aes(x = visit_spend)) +
>       theme_minimal() +
>       ggtitle("6 Top spending customers...") +
>       xlab("Amount spent per visit") +
>       ylab("Instances") +
>       facet_wrap(~ customer_id, nrow = 2)
```

Challenge 3:

Save

```
> ggsave("Top 6 June-Dec 2010.pdf", g, width = 20,
  height = 12, unit = "cm")
```

Challenge 3:


6 Top spending customers, June – December 2010
123
134
140
199
254
263
Instances
30
20
10
0
0
50
100
150
Amount spent per visit

Challenge 4: two – way comparisons

Heatmaps are a useful graphic to observe the effect of two factors on a variable.

In this example, we will compare the number of visits made by each customer, by month.

Challenge 4:

Setup environment (using script to set working directory):

```
> rm(list = ls())
> library(readr)
> library(ggplot2)
> DH <- read_csv("Dunhumby1-20.csv")
```

Challenge 4:

To count visits by months, first create a separate "month" column:

```
> DH$tempdate = as.Date(DH$visit_date,"%d-%m-%y")
  # make date object
> DH$month = as.numeric(format(DH$tempdate,
  "%m")) # extract month
> DH$tempdate = NULL # delete temp column
```

Challenge 4:

Now count visits by ID and month:

```
> attach(DH)
> CustVisits = as.table(by(visit_spend, list(customer_id, month), length)) # make table
> CustVisits = as.data.frame(CustVisits) # convert to df
> colnames(CustVisits) = c("ID", "Month", "Visits")
> CustVisits$Month = as.numeric(CustVisits$Month) # make months numeric
```

Note: the use of a list.

Challenge 4:

Data file is now in the format:

ID	Month	Visits
40	1	8
79	1	5
119	1	6
123	1	12
134	1	17
140	1	14
148	1	18
149	1	1
168	1	8
…	…	…

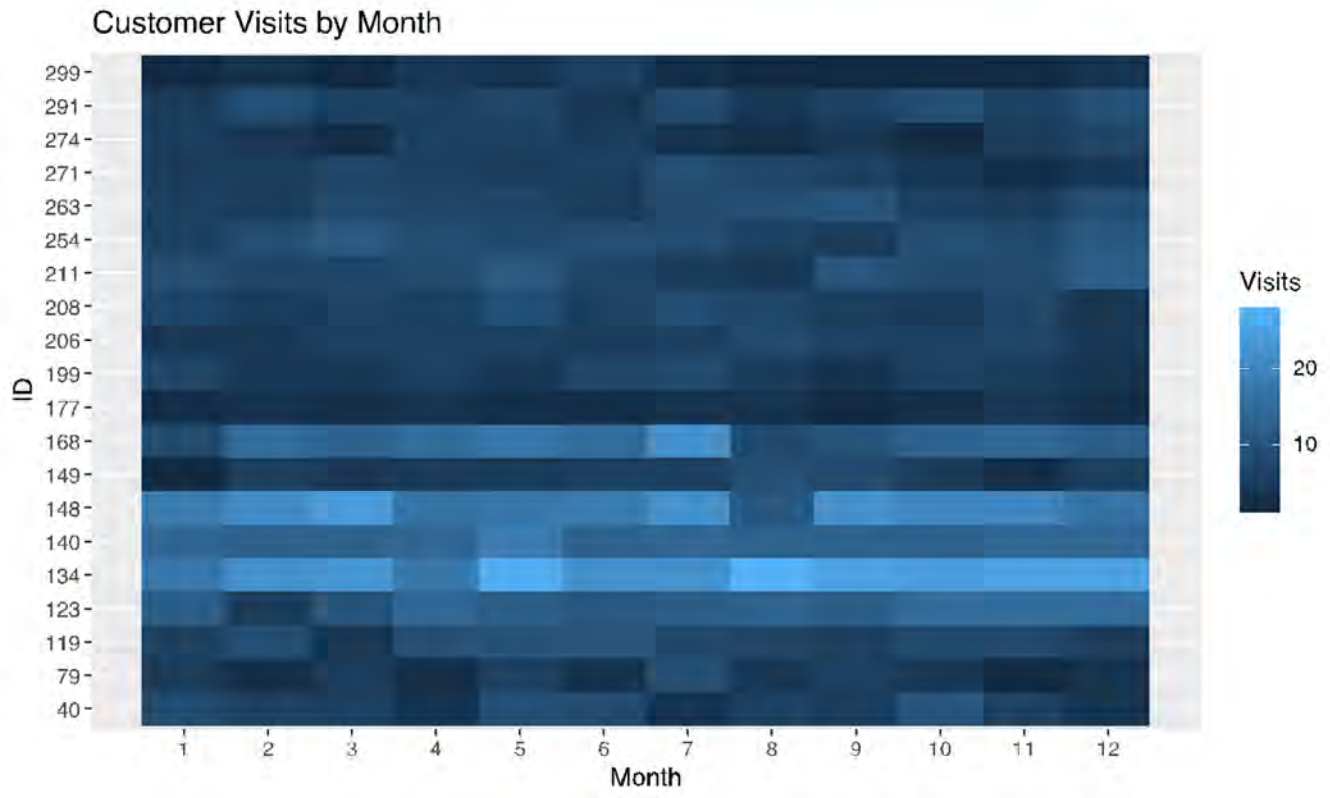
Challenge 4:

Plot and save. Set x breaks by hand.

```
> g = ggplot(data = CustVisits, aes(x = Month, y = ID))
> g = g + geom_tile(aes(fill = Visits))
> g = g + ggtitle("Customer Visits by Month")
> g = g + scale_x_continuous(breaks = c(1, 2, 3, 4, 5, 6, 7,
  8, 9, 10, 11, 12))
> g
> ggsave("Customer Visits by Month.pdf", g, width = 20,
  height = 12, unit = "cm")
```

Challenge 4:

The finished heat map (could be improved…)



Review questions - answers

1. C
2. B
3. B
4. B

Notes

Acknowledgement

- Material on the data science industry and data preparation was taken from or inspired by previous guest lectures by Mr Bernie Kruger.