FIT3152 Data analytics– Lecture 4

Assignment 1, questions answered

The data science industry

Data science workflow models

Dirty, and Tidy data

Transforming data: recoding, extracting subsets

Drawing a heatmap.

Advertising: Ideation Experience

Looking for some practical experience in ideation? Wanting to add experience with Miro to your CV? Interested in supporting a not for profit? Get hands on in the Ideas for Social Good Workshop!

Delivered by Hackathons International with mentors from Communiteer (a crowdsourcing platform that connects, engages and mobilises volunteers), you will be tasked with generating ideas and solutions around leveraging skills-based volunteering to become workforce ready.

Date: Tuesday, 30 March
Time: 5-7pm
Location: Via Zoom

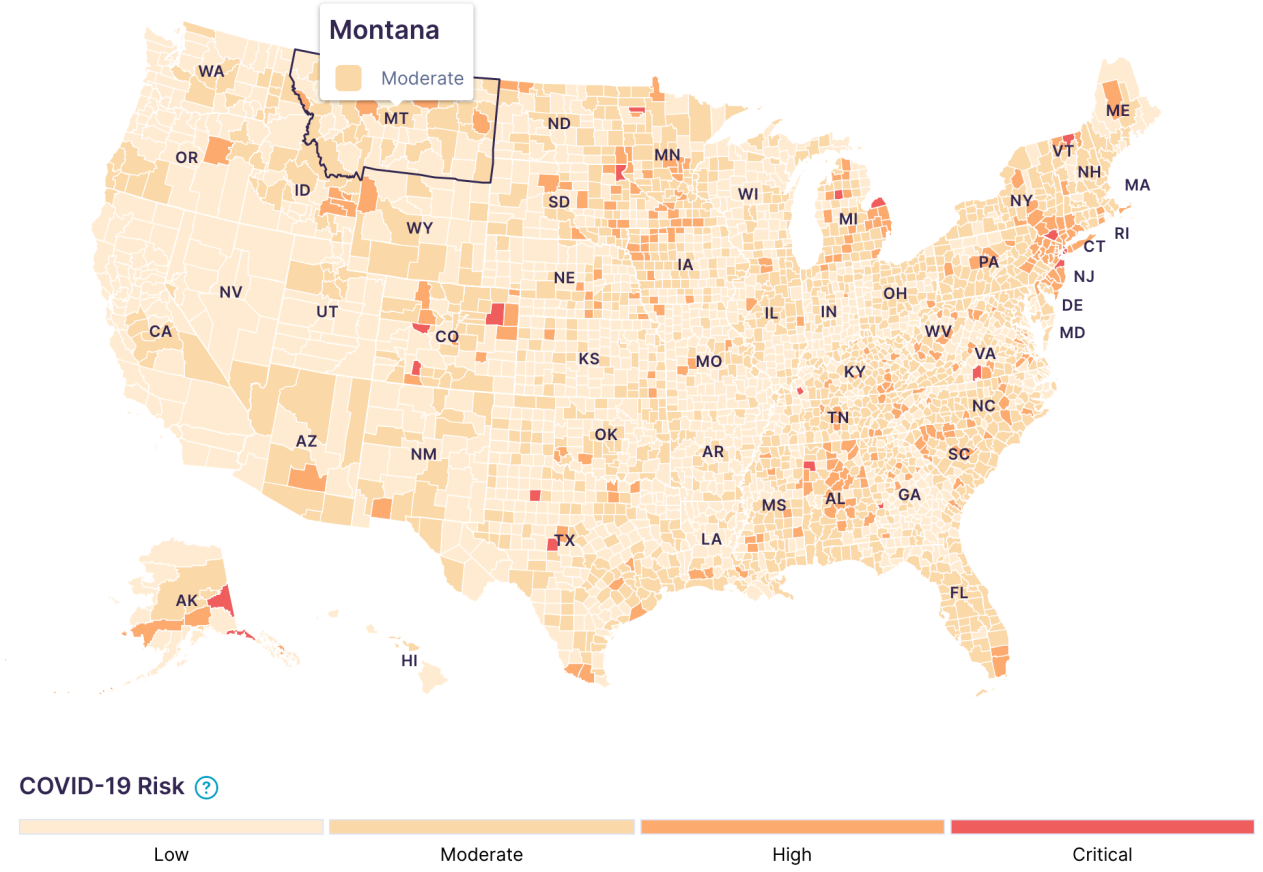
Register here

https://events.humanitix.com/ideas-for-social-good-ideation-workshop...

Week-by-week

Week Starting	Lecture	Topic	Tutorial	A1	A2
2/3/21	1	Intro to Data Science, review of basic statistics using R	...		
9/3/21	2	Exploring data using graphics in R	T1		
16/3/21	3	Data manipulation in R	T2	Released	
23/3/21	4	Data Science methodologies, dirty/clean/tidy data, data manipulation	T3		
30/3/21	5	Network analysis	T4		
6/4/21		Mid-semester Break			
13/4/21	6	Regression modelling	T5		
20/4/21	7	Classification using decision trees	T6	Submitted	
27/4/21	8	Naïve Bayes, evaluating classifiers	T7		Released
4/5/21	9	Ensemble methods, artificial neural networks	T8		
11/5/21	10	Clustering	T9		
18/5/21	11	Text analysis	T10		Submitted
25/5/21	12	Review of course, Exam preparation	T11		

Taking the pandemic's temperature



https://healthweather.us/covid-map?mode=covid_risk&resolution=county

Taking the pandemic's temperature

For years Singh's company, Kinsa Health, had tracked fevers using data from its network of thousands of smartphone-connected thermometers.

As the potential scope of the covid-19 outbreak became clear, Singh subtracted the typical cold and flu numbers gathered in years past from the company's graphs."What's left over are unusual fevers, and we saw hot spots across the country," Singh says. "We observed six years of data and there'd been hot spots, but nothing like we were seeing in early March."

https://www.technologyreview.com/

Assignment 1

Assignment 1

FIT3152 Data analytics: Assignment 1

This assignment is worth 20% of your final marks in FIT3152. Due: Friday 23rd April 2021.

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:

Assignment 1

a. Analyse activity and language on the forum over time. Some starting points:

- Describe your data: How active are participants, and are there periods where this increases or decreases? Is there a trend over time?
- Looking at the linguistic variables, do these change over time? Is there a relationship between variables?

b. Analyse the language used by groups. Some starting points:

- Threads indicate groups of participants communicating on the same topic. Describe the threads present in your data.
- By analysing the linguistic variables for all or some of the threads, is it possible to see a difference in the language used by different groups?
- Does the language used within threads (or between threads) change over time? How consistent or variable is the language used within threads?

Assignment 1

c. Challenge: Social networks online. We can think of participants posting to the same thread at similar times (for example during the same month) as forming a social network. When these participants also post to other threads over the same period, their social network extends.

- Can you define, graph and describe the social network that exists at a particular point in time, for example over one month? How does this change in the following months?
- Note: you only need to analyse a small portion of the social network over a short time period. We will cover social network analysis in Lecture 5.

d. Reflection on your investigation. What did you first investigate? How did you then modify your research based on the results of your first investigation?

- Using one of the data science methodologies in Lecture 4, illustrate your research process.

Assignment 1

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

ThreadID	AuthorID	Date	Time	WC	Analytic	Clout	Authentic	Tone	WPS	i	we	you	they	number	affect	posemo	negemo	anx
659289	193537	24/11/2009	5:36	53	82.26	71.43	25.14	25.77	26.5	0	1.89	0	3.77	3.77	3.77	1.89	1.89	0
432269	136196	26/11/2007	23:42	216	25.71	94.73	45.81	33.77	24	1.85	6.48	0.46	5.09	0.46	6.02	3.24	2.78	0
572531	170305	17/02/2009	7:31	136	31.61	67.04	28.81	79.41	13.6	3.68	0	5.15	2.94	0.74	9.56	5.88	2.94	0.74
230003	32359	7/09/2005	21:25	29	39.74	91.6	3.81	85.87	14.5	3.45	0	6.9	0	6.9	3.45	3.45	0	0
459059	47875	19/02/2008	5:23	108	80.75	60.95	23.51	88.52	13.5	2.78	0	0	0	0.93	9.26	6.48	2.78	0
635953	181593	28/09/2009	8:40	86	64.98	45.37	57.24	1	43	1.16	0	0	5.81	3.49	3.49	0	3.49	0
235116	51993	29/09/2005	15:59	49	33.33	20.71	13.15	25.77	16.33	6.12	0	0	2.04	0	8.16	4.08	4.08	0
593767	169459	23/04/2009	19:21	368	85.91	63.82	19.13	7.15	24.53	1.36	2.17	0	0.54	0.54	5.43	1.9	3.53	0.54
532649	248548	25/12/2011	8:28	13	92.84	50	1	25.77	13	0	0	0	0	61.54	0	0	0	0
517685	65	20/02/2005	10:50	65	91.21	62.1	33.6	81.28	13	7.69	0	0	0	0	9.23	6.15	3.08	0
588291	158329	23/04/2009	23:40	265	55.7	73.95	45.85	11.21	44.17	1.89	1.13	0.38	3.4	5.66	3.4	1.13	2.26	0
29936	194	25/07/2002	4:29	106	80.44	80.2	20.42	98.46	15.14	1.89	0	4.72	0	0.94	7.55	6.6	0.94	0.94
199787	47875	20/05/2005	16:48	160	94.48	73.4	2.07	5.64	22.86	1.25	0	0	0	5.62	8.12	3.12	5	1.88
545552	143229	24/11/2008	23:39	33	79.25	18.16	98.01	80.64	8.25	6.06	0	0	0	3.03	3.03	3.03	0	0
303058	88912	25/07/2006	23:57	244	44.21	65.92	33.49	7.09	27.11	2.87	0.82	0.41	4.51	1.64	6.56	2.46	4.1	0
772248	75628	16/01/2011	2:24	108	39.91	57.35	45.81	25.77	13.5	5.56	0	2.78	0	0.93	1.85	0.93	0.93	0
761807	227011	4/12/2010	23:48	104	73.9	57.63	74.76	62.24	34.67	0.96	0	2.88	3.85	2.88	5.77	3.85	1.92	0
110837	34501	24/01/2004	2:53	49	90.62	20.71	46.05	1	24.5	2.04	0	0	0	0	6.12	0	6.12	0
636255	180475	3/09/2009	22:25	2	92.84	99	1	99	2	0	0	0	0	0	50	50	0	0
178736	43291	18/01/2005	2:40	75	69.57	92.87	1	1	15	0	0	2.67	6.67	0	10.67	1.33	9.33	1.33
275754	-1	6/03/2006	18:01	56	92.84	70.4	41.07	6.15	18.67	1.79	0	1.79	0	1.79	1.79	0	1.79	0
833308	231141	21/09/2011	21:39	32	78.67	82.58	74.76	25.77	16	0	0	6.25	0	0	0	0	0	0
642657	180098	13/11/2009	16:34	13	92.84	6.21	99	1	13	23.08	0	0	0	0	7.69	0	7.69	7.69
365246	116735	17/02/2007	9:48	48	49.05	33.83	62.53	1	48	2.08	0	2.08	2.08	0	10.42	2.08	8.33	4.17
279233	84070	21/03/2006	1:59	51	77.76	50	66.34	25.77	51	3.92	0	1.96	0	1.96	7.84	3.92	3.92	0
300539	-1	8/06/2006	22:43	24	49.05	33.83	23.51	92.4	6	8.33	0	0	4.17	8.33	4.17	4.17	0	0
277955	32925	14/03/2006	23:45	87	55.99	78.96	62.98	3.63	43.5	0	0	1.15	4.6	2.3	2.3	0	2.3	1.15
90325	32485	25/09/2003	3:30	48	94.65	79.76	3.9	25.77	12	0	0	0	2.08	2.08	12.5	6.25	6.25	0
321495	90627	12/09/2006	1:40	42	40.66	68.29	37.24	70.57	21	4.76	4.76	2.38	2.38	0	2.38	2.38	0	0
281667	79878	28/03/2006	2:45	60	32.98	56.63	65.14	1.03	20	1.67	1.67	0	3.33	0	3.33	0	3.33	0
294983	75902	21/05/2006	0:07	60	56.15	25.24	32.84	25.77	60	3.33	0	0	0	0	6.67	3.33	3.33	0
397699	125170	21/06/2007	21:41	34	92.84	92.92	14.7	25.77	17	0	2.94	2.94	0	0	5.88	2.94	2.94	0
313191	101368	30/07/2006	17:53	25	81.4	2.31	43.37	25.77	25	0	0	0	0	12	0	0	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...

Assignment 1

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

Column	Brief Descriptor
ThreadID	Unique ID for each thread
AuthorID	Unique ID for each author
Date	Date
Time	Time
WC	Word count of the text of the post
Analytic	LIWC Summary (Analytical thinking)
Clout	LIWC Summary (Power, force, impact)
Authentic	LIWC Summary (Using an authentic tone of voice)
Tone	LIWC Summary (Emotional tone)
WPS	LIWC (Words per sentence)
i	LIWC ("I, me, mine" words) First person singular
we	LIWC ("We, us, our" words) First person plural
you	LIWC ("You" words) Second person
they	LIWC ("They" words) Third person plural
number	LIWC(Quantities and ranks)
affect	LIWC (Expressing sentiment)
posemo	LIWC (Positive emotions)
negemo	LIWC (Negative emotions)
anx	LIWC (Indicating anxiety)

Assignment 1

Submission. Due Friday 23rd April 2021 11:55pm GMT+10.
Suggested length: 6–8 A4 pages + appendix.
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. Submit your report as a single PDF with the file name *FirstnameSecondnameID.pdf* on Moodle.

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

Assessment criteria will include:

The quality of your analysis and description of your analytical process; Graphics and tables supporting your analysis; The quality of graphics used in the report. Justification of your findings and the degree of proof you can offer (for example statistical tests); Readability and quality of your written report; Insights gained from the data; Novelty of your approach.

Factors you should consider (starting points, not a complete list):

Techniques: summary/descriptive statistics, identification of important variables, networks, etc.
Major grouping variables: author, thread, date and/or time, or a combination of these.
Time window (days, weeks, months, years…); Subsets of the data to be analysed.
Graphics to communicate your analysis and insights (histograms, scatterplots, heat maps, time series are some basic starting points, but see https://datavizproject.com/ for inspiration.

Response to student questions

- Is a graph enough to answer a question such as "is there a relationship between variables?" or should we also be backing up our claims with statistical tests – e.g.: t-tests, size of confidence intervals, p-values etc.
  - > Descriptive stats/graphs/tables are ok for analysis. Statistical tests/models may earn additional marks.
- The suggested page length for this assignment is 6-8 pages long, how many pages should the reflection on investigation take up approximately?
  - > Half page, 1 page max.

Response to student questions

- Is it okay to go over the suggested page length?
  - > Yes, but be sensible.
- Is there a specific font/font size we have to use?
  - > Be sensible. 11/12 point is recommended.
- Should we consider model fitting such as decision trees, linear regressions, and kNN clustering to support our claims for relationships?
  - > These are not necessary but you could use regression as proof for a trend over time.

Response to student questions

- Do we submit R file with pdf?
  > Submit a single pdf with R code as an appendix.
- Are we allowed to use external libraries in R e.g. dplyr, tidyverse, etc?
  > Yes, note these in the report. Include in R script.
- Can we just write answers in the R script and submit the R Markup file.
  > Yes
- All code have to be text?
  > Yes, make it machine readable.

Response to student questions

- Can we have a rubric?
  - > See next slide
- Is clean/neat code going to be a part of the marking criteria in the assignments?
  - > Yes
- Does the code in the appendix need to be in a fixed-width font?
  - > Preferably

Response to student questions

- Rubric (20 Marks total)
  - > Understanding the data (2 Marks)
  - > Activity and language over time (3 Marks)
  - > Language used by groups (4 Marks)
  - > Social network analysis (2 Marks)
  - > Reflection (1 Mark)
  - > Quality of the analysis (2 Marks)
  - > R coding (2 Marks)
  - > Charts and tables (2 Marks)
  - > Quality of writing (2 Marks)

Data manipulation review questions

Please respond using Zoom chat...

Edgar Anderson's Iris data

50 samples from 3 species:

- Iris setosa, – virginica, – versicolor

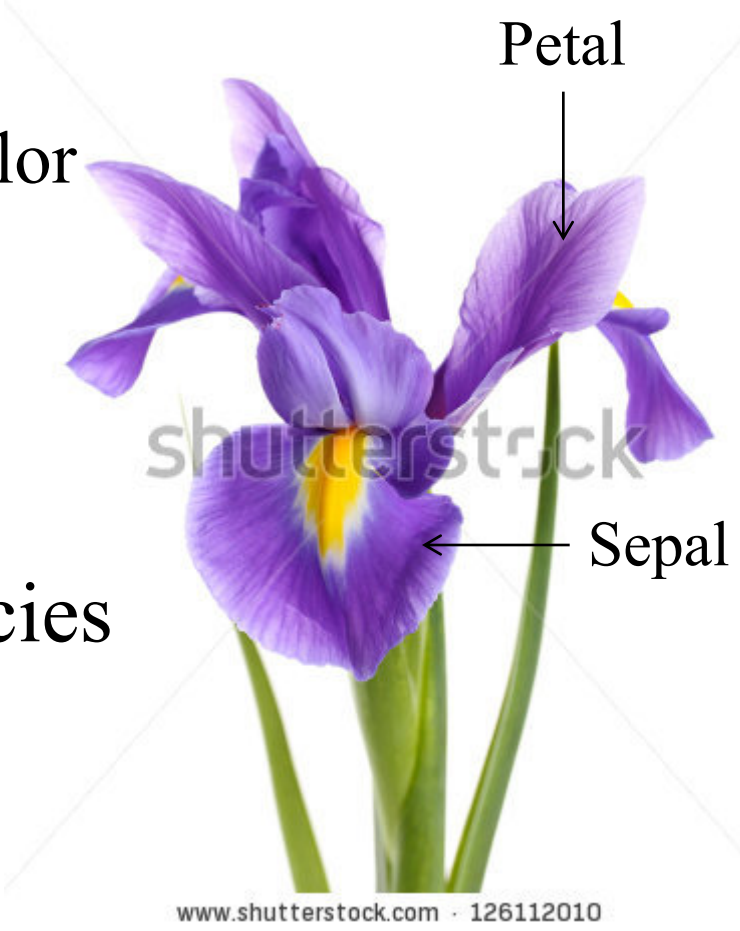
Four features measured:

- 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

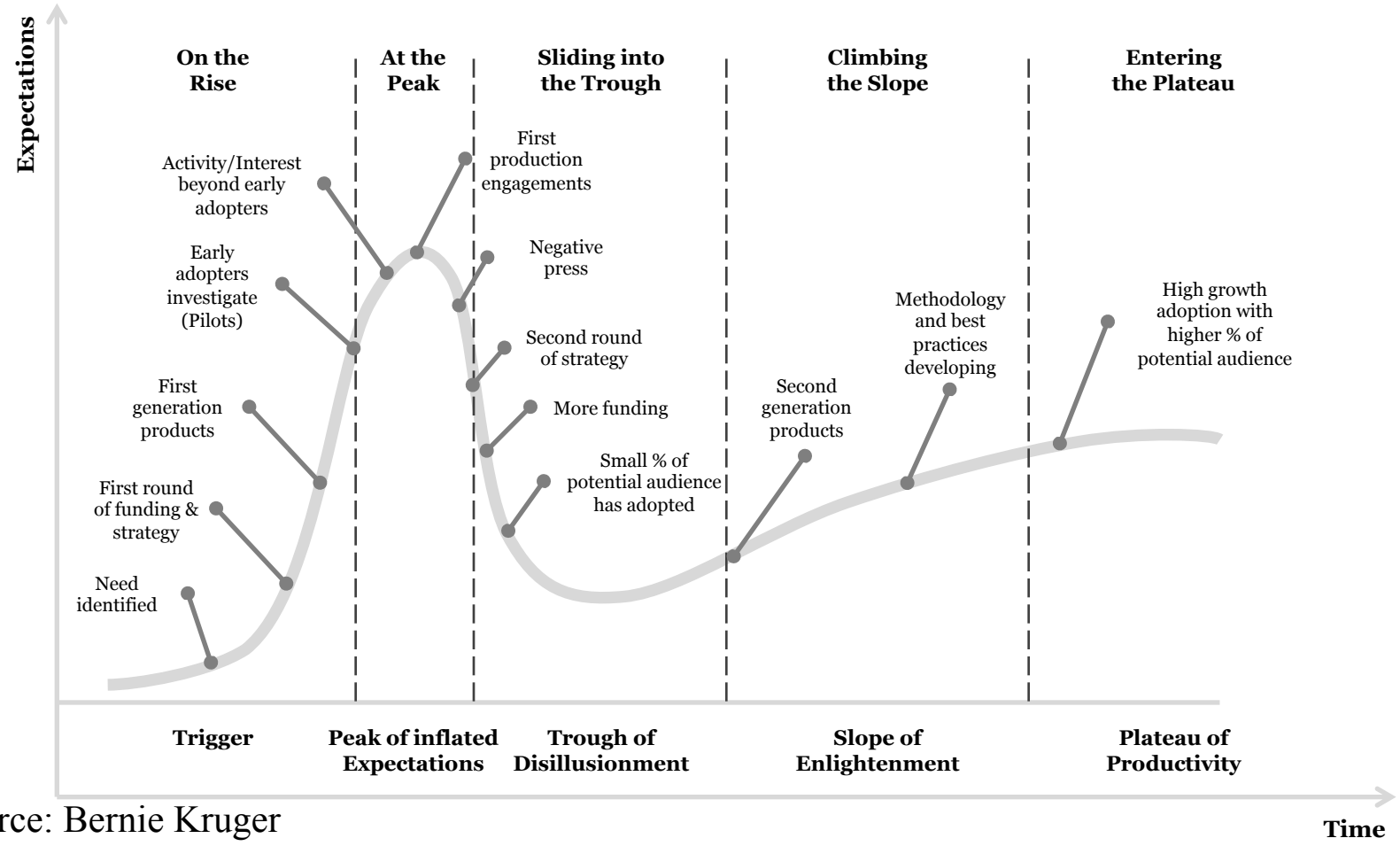
(d) Row data having longest petal by species

The data science industry

Some general thoughts, including those from a previous 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.

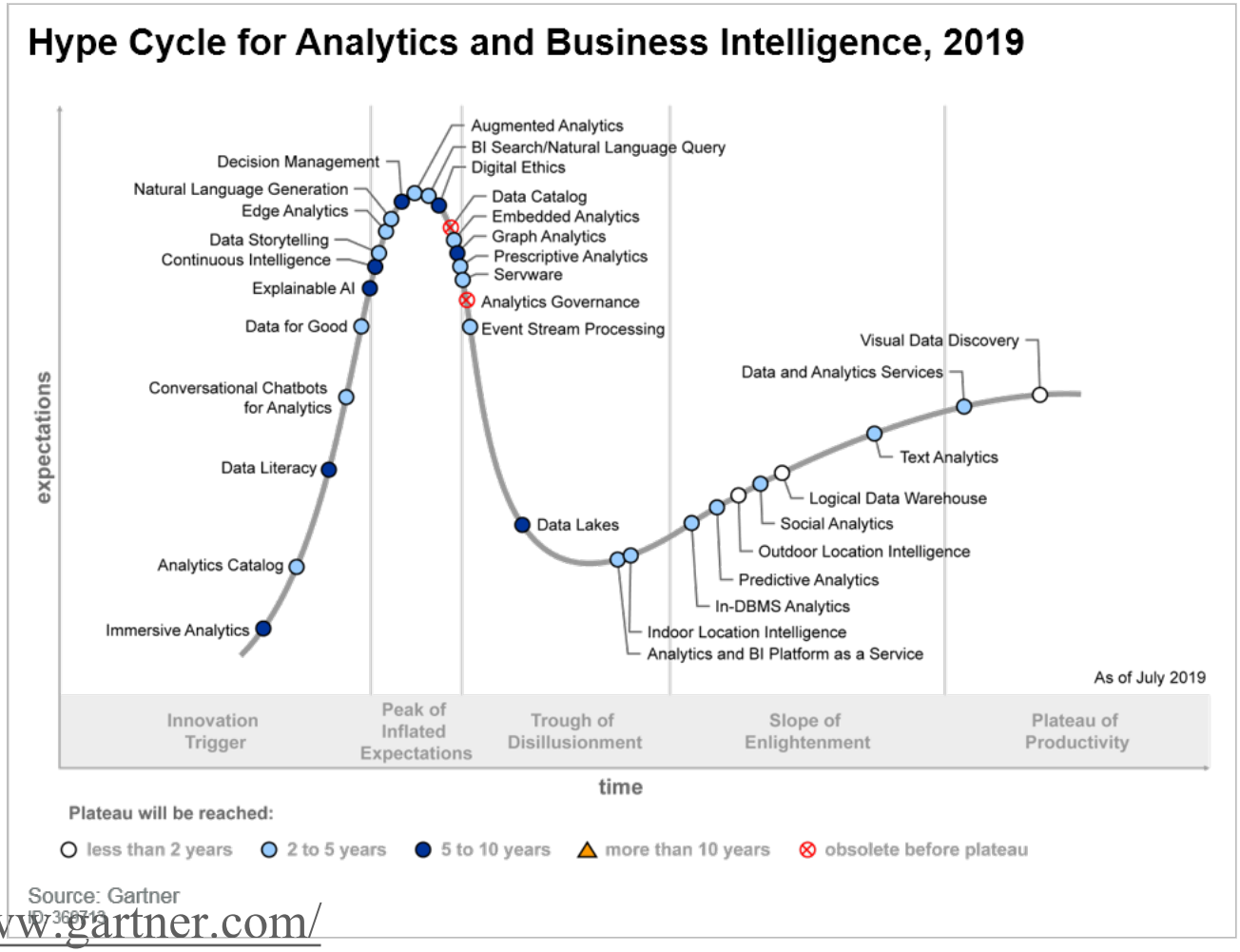
Trough of Disillusionment: Interest wanes as experiments and implementations fail to deliver. Producers of the technology shake out or fail. Investments continue only if the surviving providers improve their products to the satisfaction of early adopters.

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.

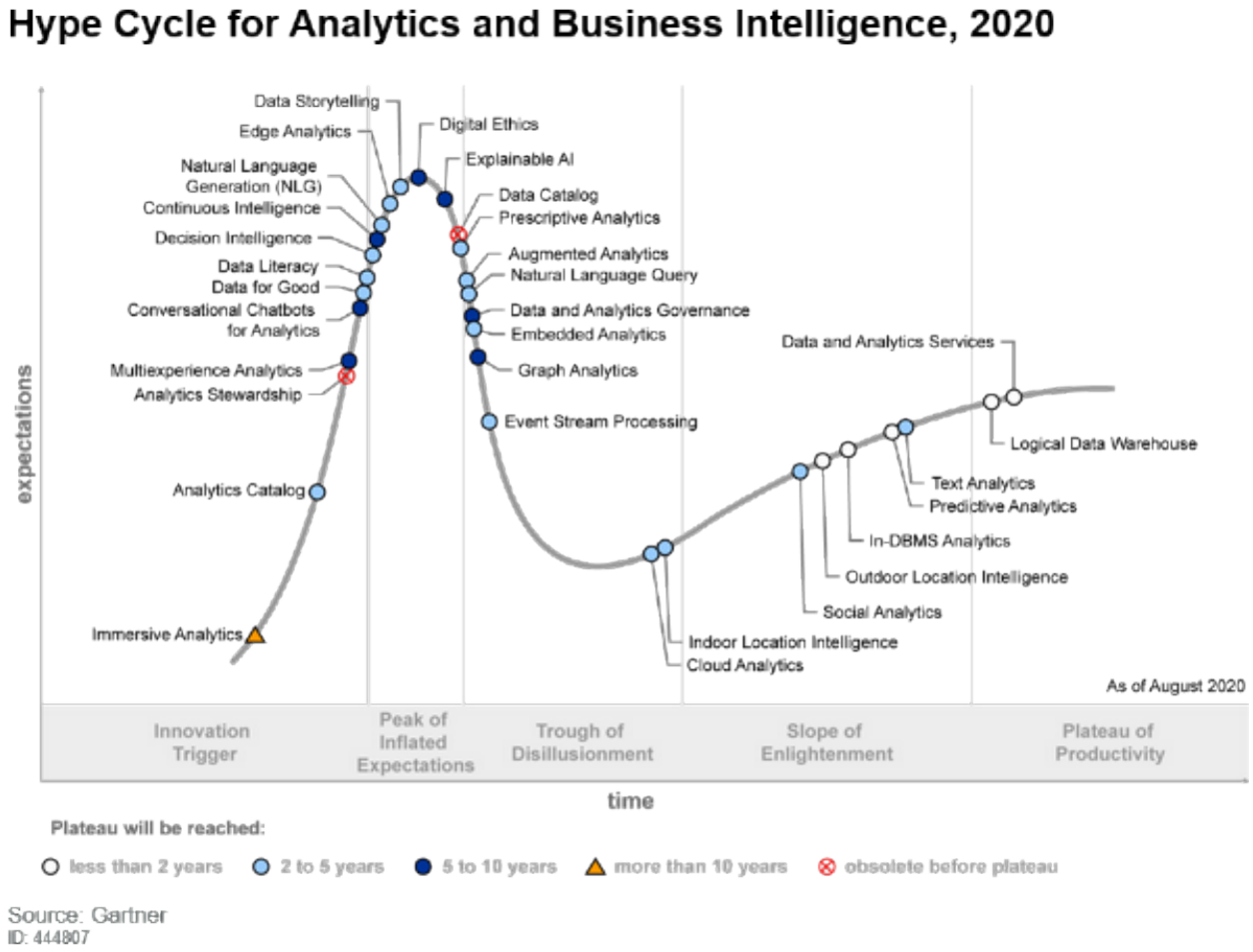
http://www.gartner.com/technology/research/methodologies/hype-cycle.jsp

Gartner Hype Cycle: Analytics & BI



https://www.gartner.com/

Gartner Hype Cycle: Analytics & BI



Source: Gartner
ID: 444807

Gartner Hype Cycle: Analytics & BI

On the rise:

- Data literacy, data for good, natural language generation, chatbots…

At the peak:

- Explainable AI, data storytelling, augmented analytics, natural language query, digital ethics, …

Sliding into the trough:

- Data lakes (disappeared 2020), analytics governance, analytics and BI as a service, …

Gartner Hype Cycle: Analytics & BI

They identify 5 key trends (from 2020):

- Augmented Analytics: using machine learning to automate data preparation and analytics…
- Digital Culture: data literacy, digital ethics, data-for-good…
- Relationship Analytics: growing use of graph, location and social network analysis...
- Decision Intelligence: capturing the factors leading to a decision...
- Operationalising and Scaling: building up analytics service for any part of the business that requires it...

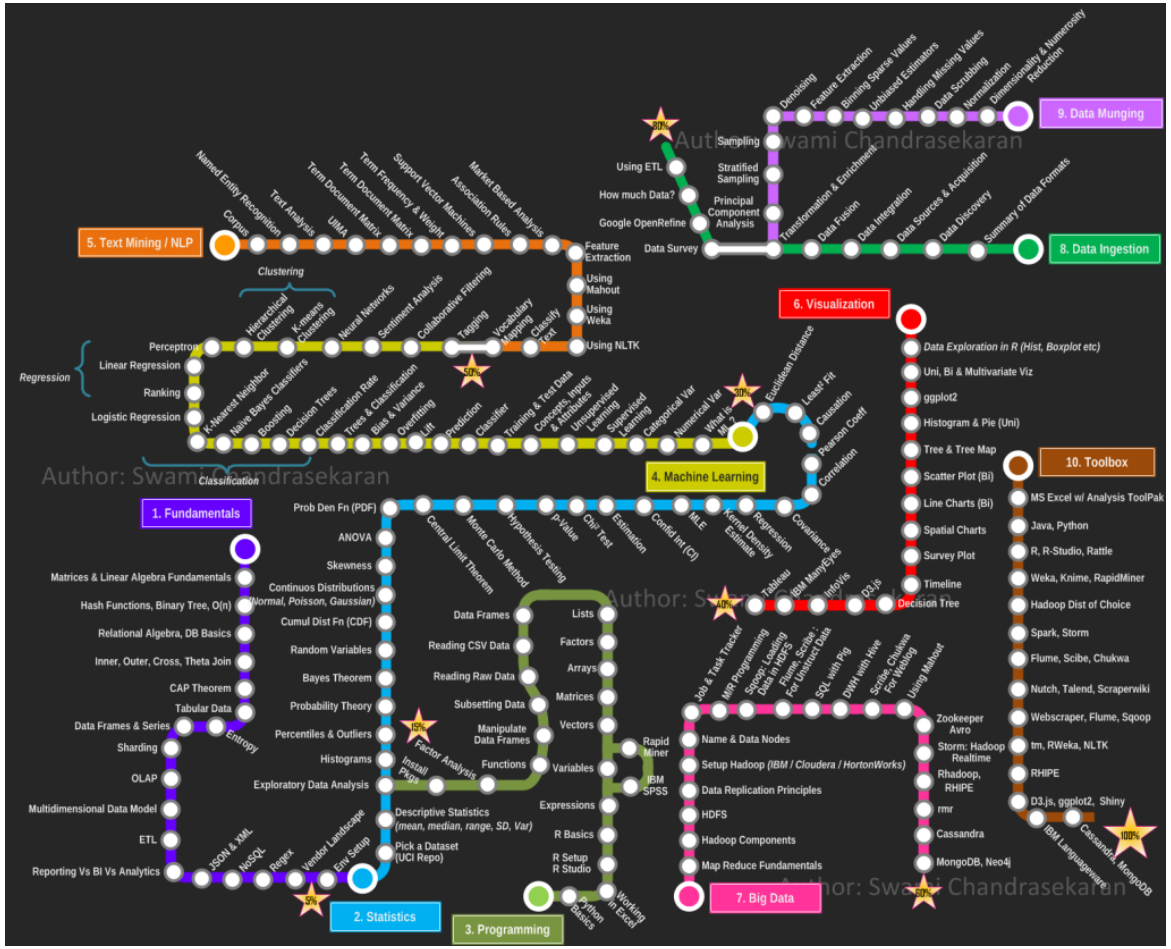
Data Science: Key Skills

From Swami Chandrasekaran:

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

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?

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)

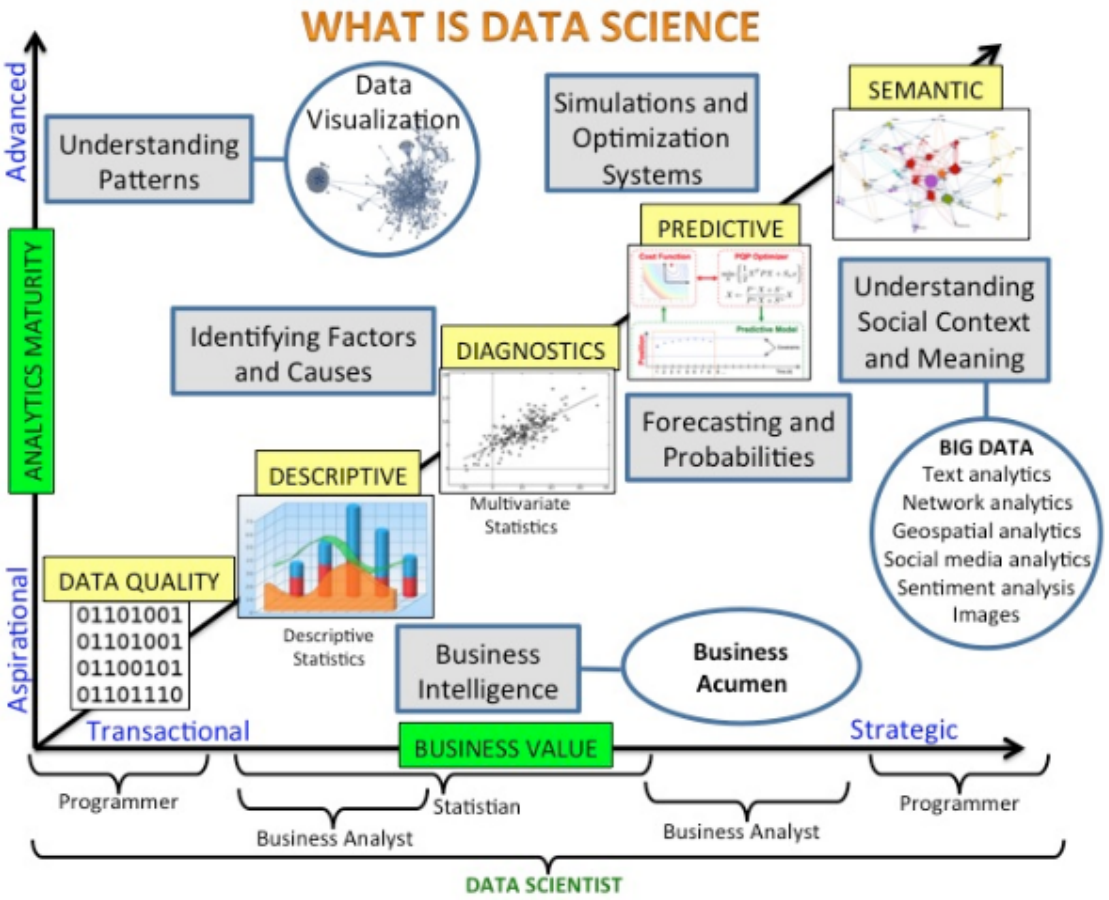
How data science is being used

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

Big Data Landscape

Matt Turck

VC at FirstMark

@mattturck

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

Resilience and Vibrancy: The 2020 Data & AI Landscape

https://mattturck.com/data2020/

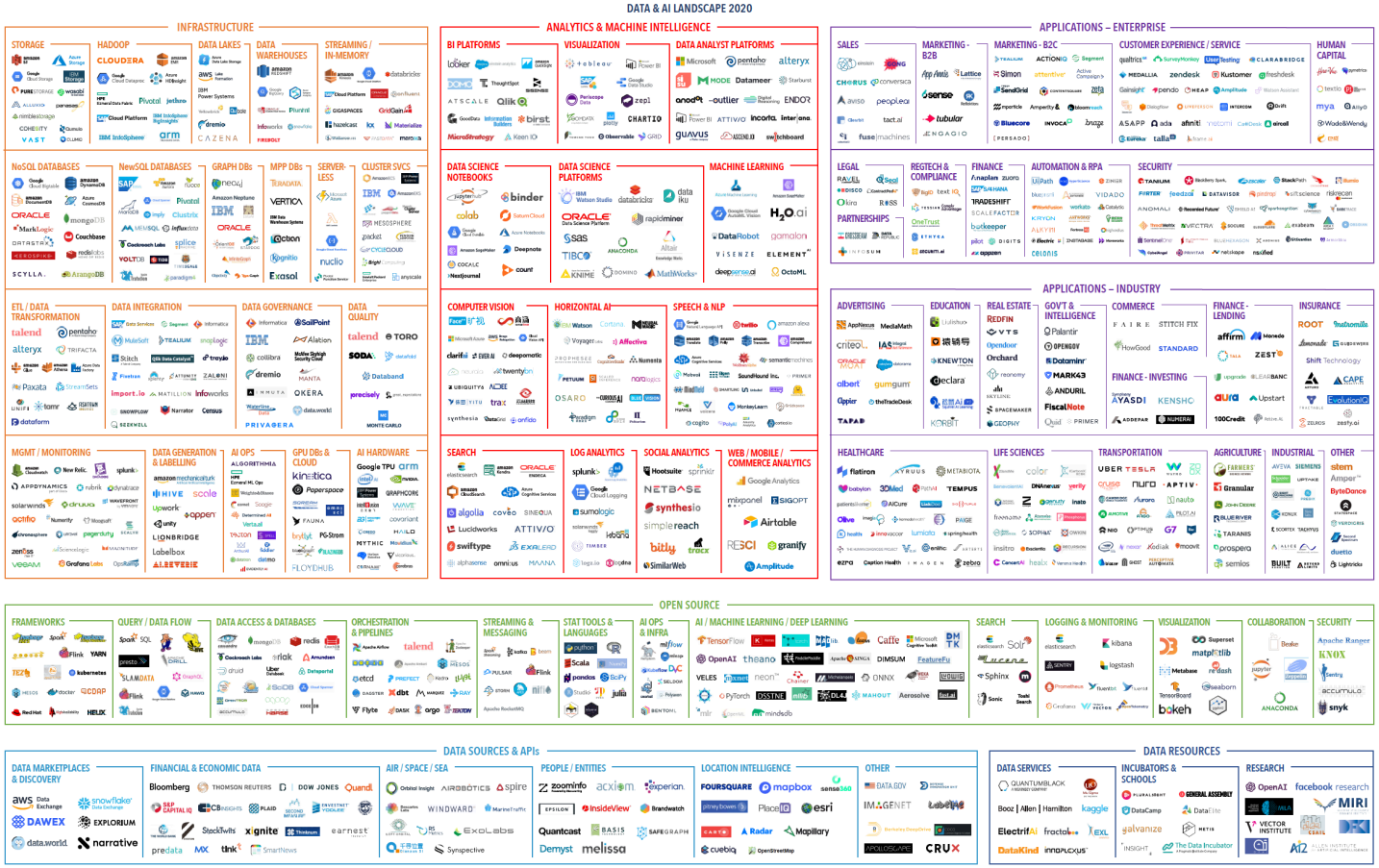
The Data & AI Landscape

In a year like no other in recent memory, the data ecosystem is showing ... exciting vibrancy.

When COVID hit the world a few months ago, an extended period of gloom seemed all but inevitable. Yet, ... “two years of digital transformation [occurred] in two months”. Cloud and data technologies (data infrastructure, machine learning / artificial intelligence, data driven applications) are at the heart of digital transformation. As a result, many companies in the data ecosystem have not just survived, but in fact thrived...

https://mattturck.com/data2020/

Data and AI Landscape



https://mattturck.com/data2020/

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

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)

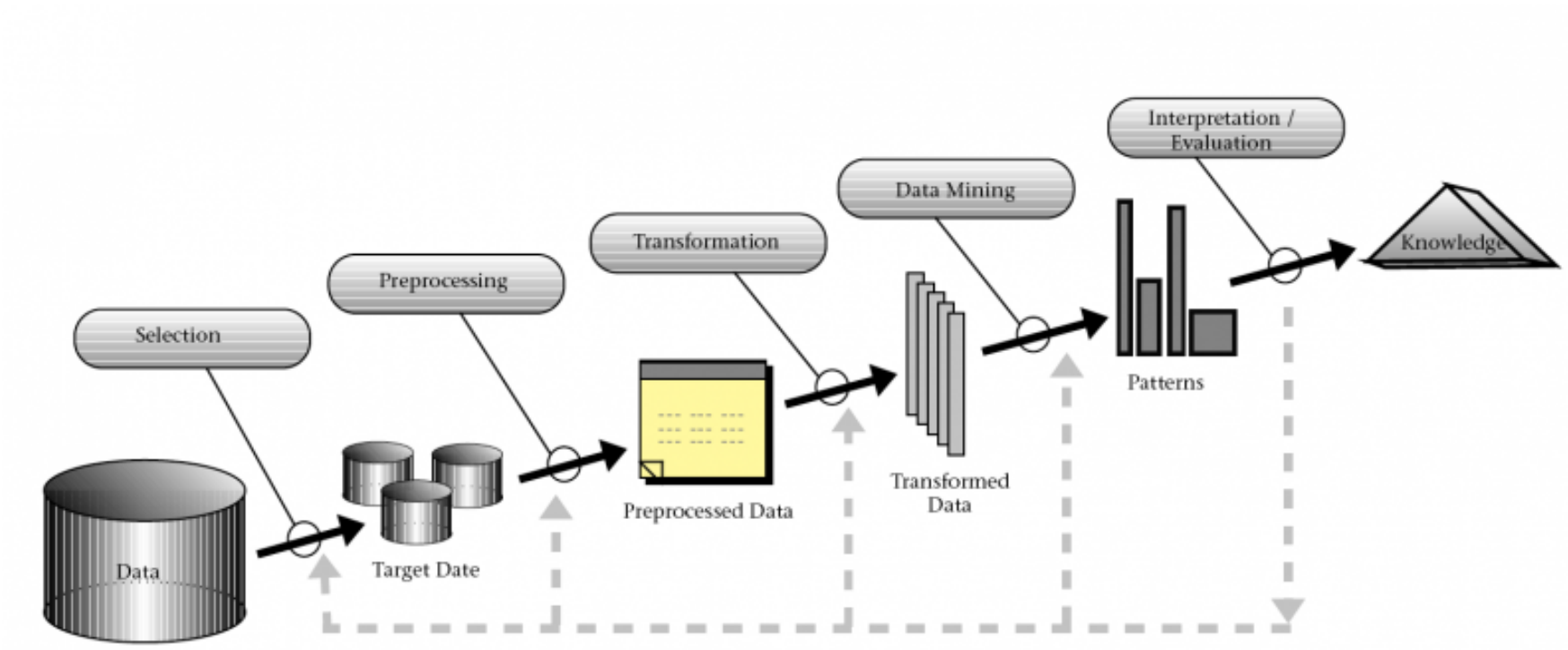
- Methodology for data mining processes proposed by the SAS Institute for the software package Enterprise Miner.

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)

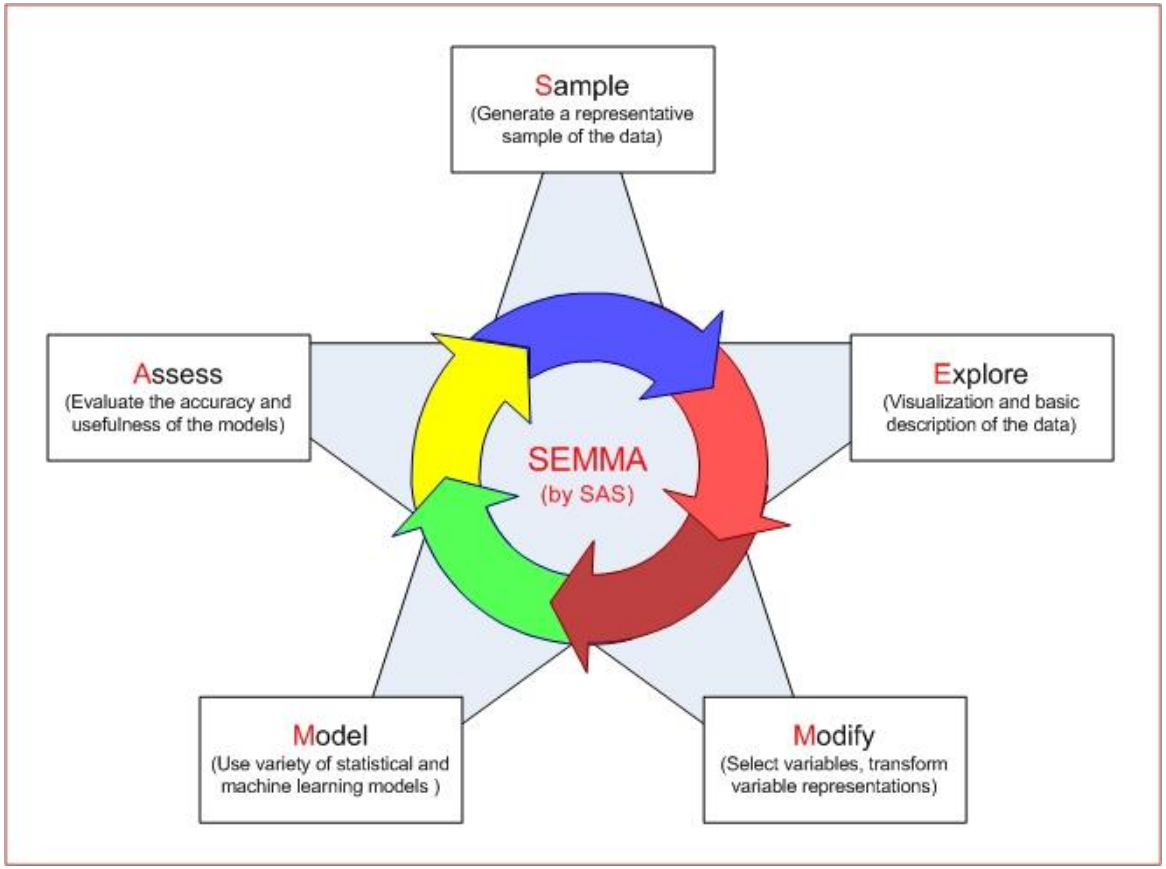
Source: Bernie Kruger

KDD



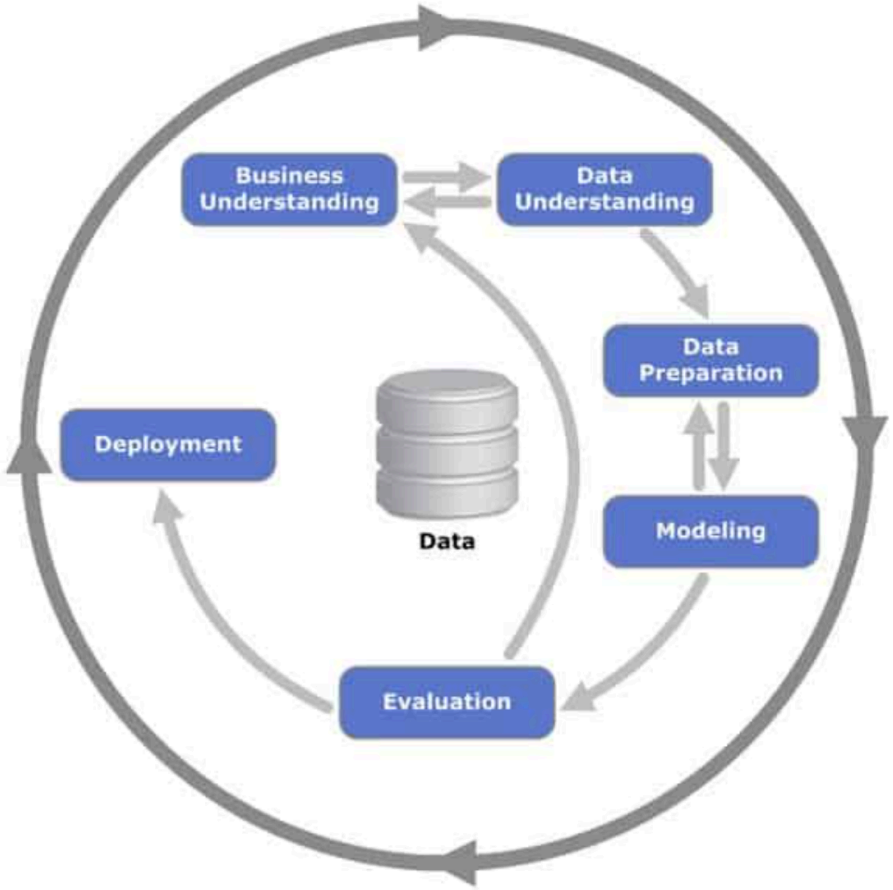
https://infovis-wiki.net/wiki/Knowledge_Discovery_in_Databases_(KDD)

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

Business understanding: Understanding the project objectives and requirements from a business perspective, then converting this knowledge into a data analysis problem definition and a preliminary plan.

Data understanding: Initial data collection and activities to get familiar with the data, to identify data quality problems, to discover first insights into the data or to detect interesting subsets to form hypotheses for hidden information.

CRISP–DM

Data preparation: Activities required to construct the final dataset from the initial raw data. Data preparation tasks are likely to be performed repeatedly and not in any prescribed order.

Modelling: Modelling techniques are selected, applied and their parameters are calibrated. Some techniques have specific requirements on the form of data. Therefore, it is often necessary to step back to the data preparation phase.

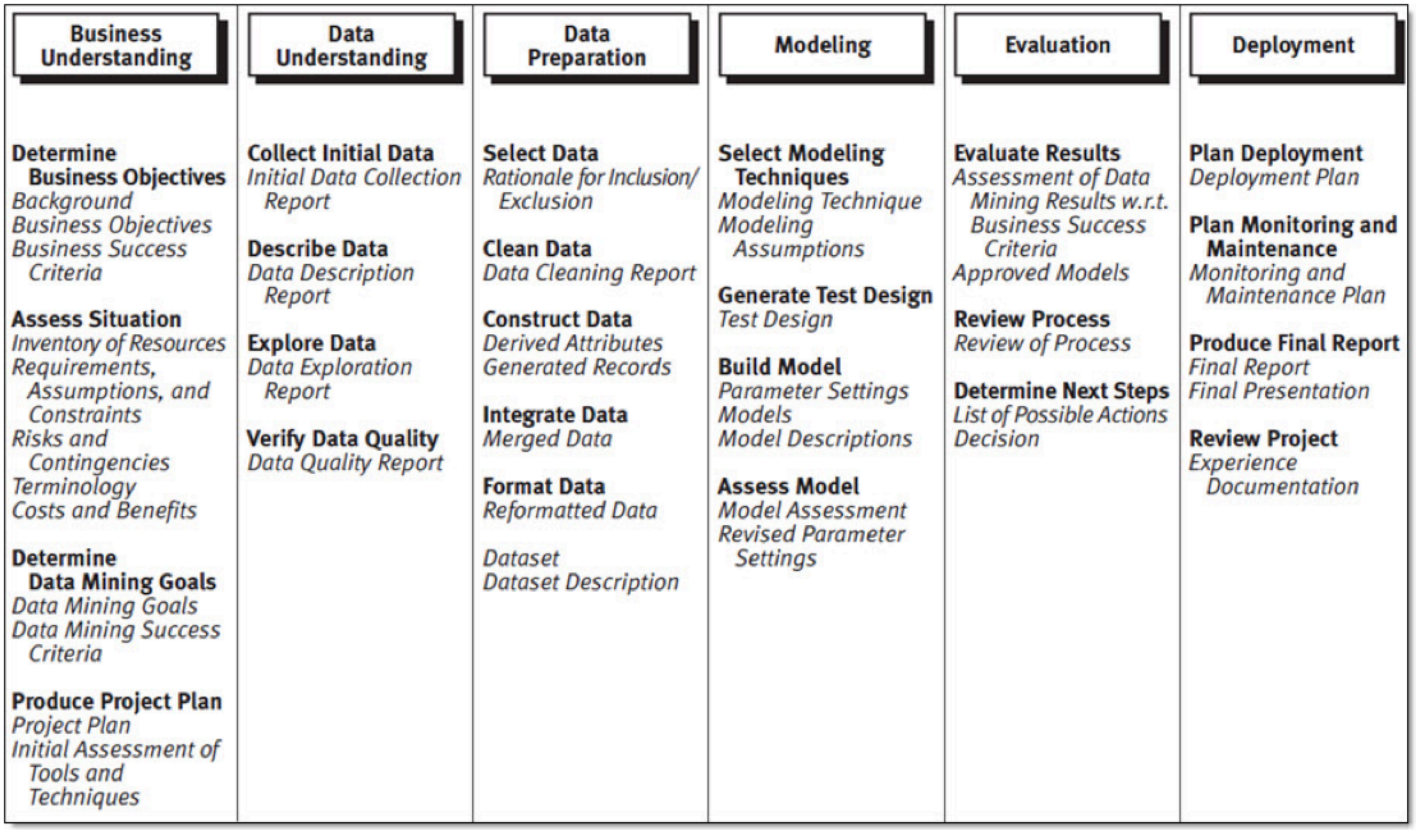
CRISP–DM

Evaluation: Before proceeding to final model deployment, it is important to evaluate the model more thoroughly and review the steps taken to build it to be certain that it properly achieves the business objectives. At the end of this phase, a decision should be reached on how to use the results.

Deployment: Model construction is generally not the end of the project. Even if the purpose of the model is to increase knowledge of the data, the knowledge gained will need to be organised and presented in a way that the customer can use.

Source: Bernie Kruger

CRISP–DM



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

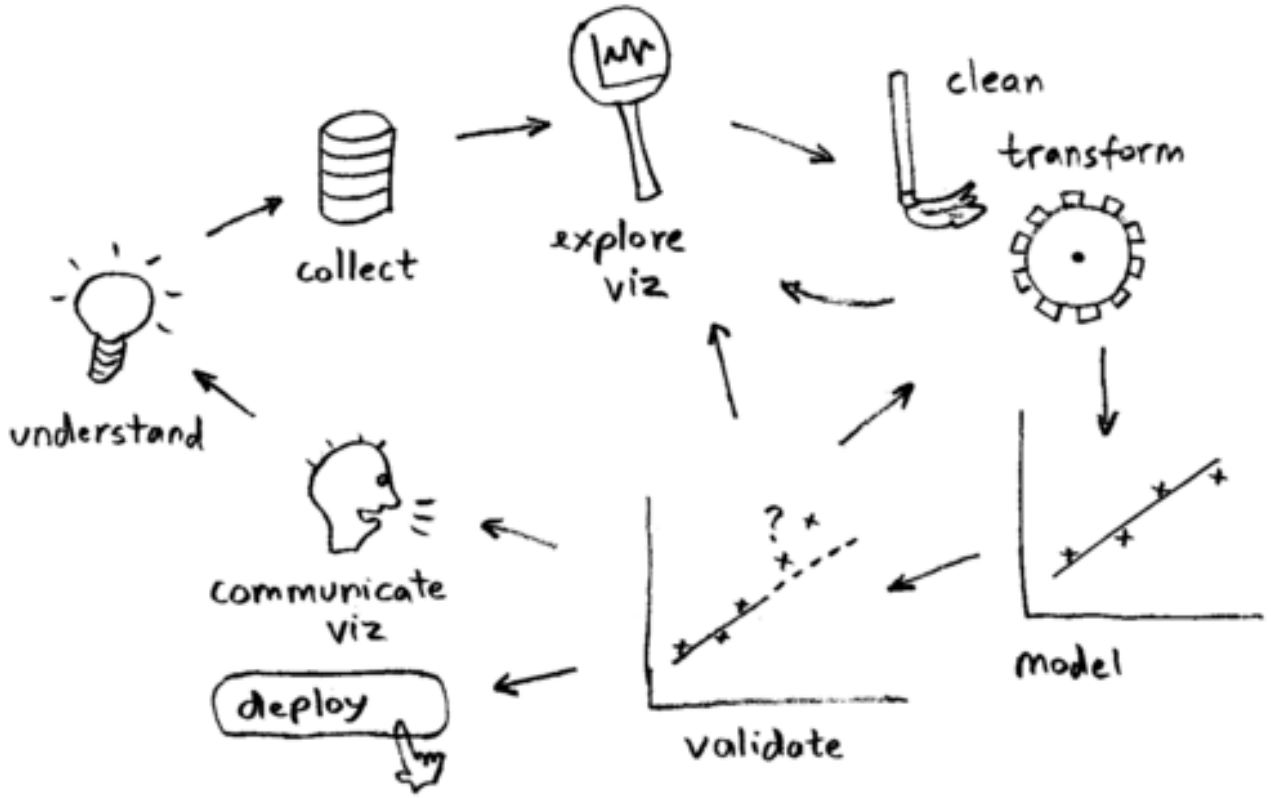
The data science workflow

What then is a typical methodology?

What is typical workflow?

Key elements of a typical workflow?

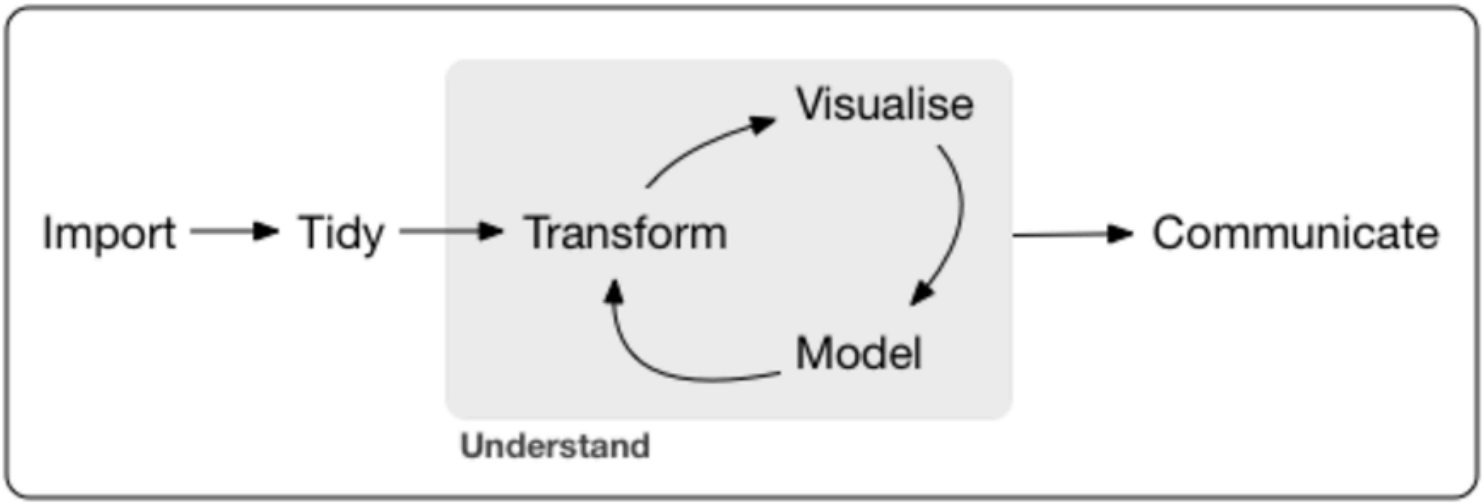
The data science workflow



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

The data science workflow

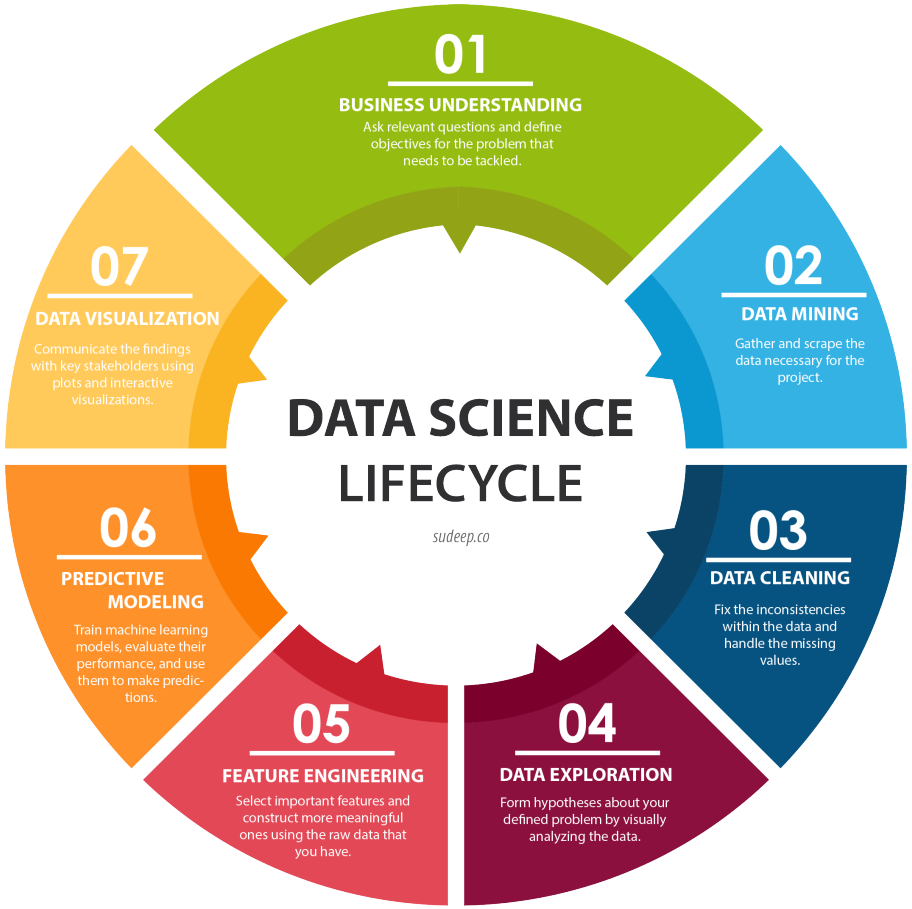
From: R for Data Science



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, Mar
218		London	1869	Bradbury, Evans & C	Love the Avenger. By the author of â€œ	A., A. A.	BLAZE DE BURY, Mar
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 Henr
481	Fourth edition, revis	London	1875	William Macintosh	[The World in which I live, and my place	A., E. S.	BROOME, John Henr
519		London	1872	The Author	Lagonells. By the author of Darmayne (F	A., F. E.	ASHLEY, Florence En
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, Giovanni
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, Franci
4884		London	1820	J. Hatchard & Son	Abdallah; or, The Arabian Martyr: a Chri		BARHAM, Thomas F
4976	[Another edition.] A	Oxonii	1800	J. Cooke, etc.	[Abdollatiphi HistoriÃ¦ Ã†gypti compen		WHITE, Joseph - Can
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, Giovar
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 - C

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, Mar
218		London	1869	Bradbury, Evans & C	Love the Avenger. By the author of â€œ	A., A. A.	BLAZE DE BURY, Mar
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 Henr
481	Fourth edition, revis	London	1875	William Macintosh	[The World in which I live, and my place	A., E. S.	BROOME, John Henr
519		London	1872	The Author	Lagonells. By the author of Darmayne (F	A., F. E.	ASHLEY, Florence En
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, Giovanni
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 geogra⍰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.	ANDRE⍰, John - Ma
4598		Hull	1814	The Author	Peace: a lyric poem. [With prefatory ad	ABBOTT, Thomas Ea	WRANGHAM, Franci
4884		London	1820	J. Hatchard & Son	Abdallah; or, The Arabian Martyr: a Chri		BARHAM, Thomas F
4976	[Another edition.] A	Oxonii	1800	J. Cooke, etc.	[Abdollatiphi HistoriÃ¦ Ã†gypti compen		WHITE, Joseph - Can
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, Giovar
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 - C

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.

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

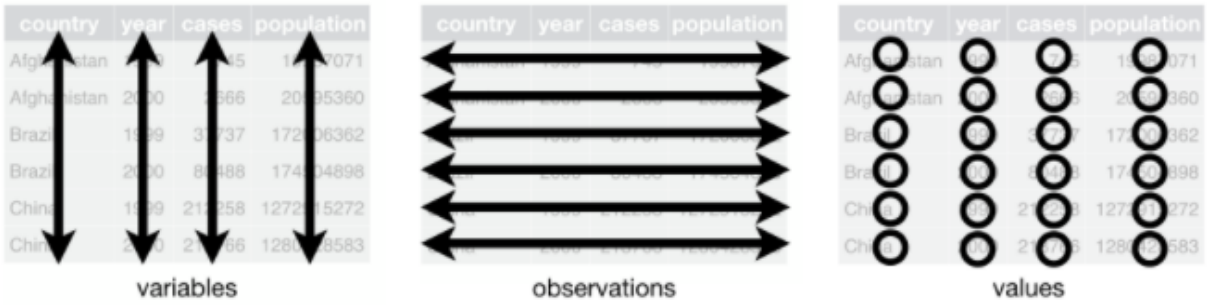
A huge amount of effort is spent cleaning data to get it ready for analysis, but there has been little research on how to make data cleaning as easy and effective as possible. This paper tackles a small, but important, component of data cleaning: data tidying. Tidy datasets are easy to manipulate, model 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

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 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 the type of transformations we look at today so be prepared to try alternative methods and packages...

Edgar Anderson's Iris data

50 samples from 3 species:

- Iris setosa, – virginica, – versicolor

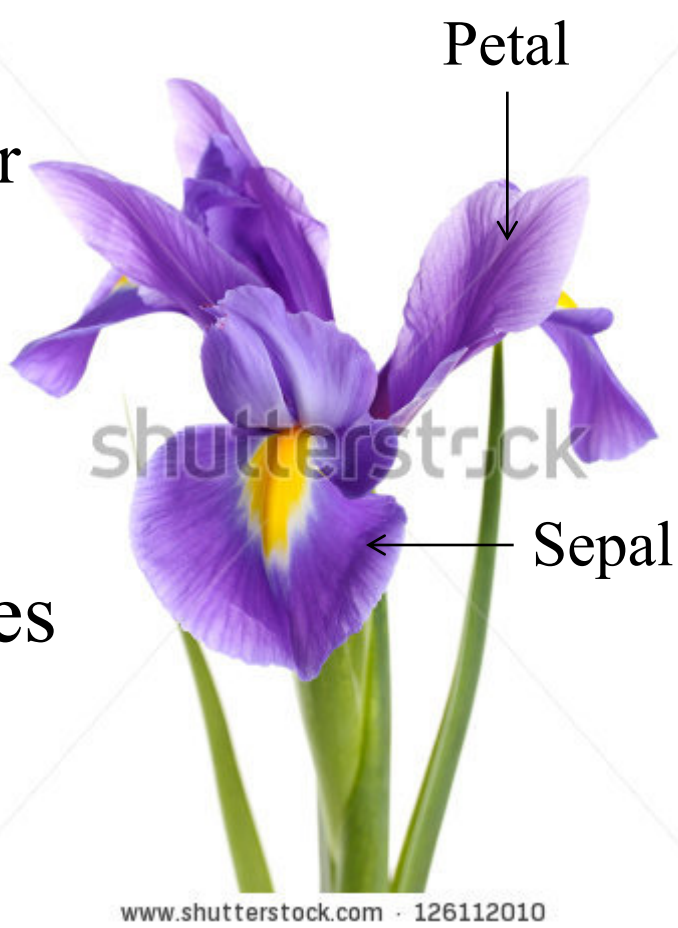
Four features measured:

- Sepal width and length
- Petal width and length

Is it possible to distinguish species using physical measurements?

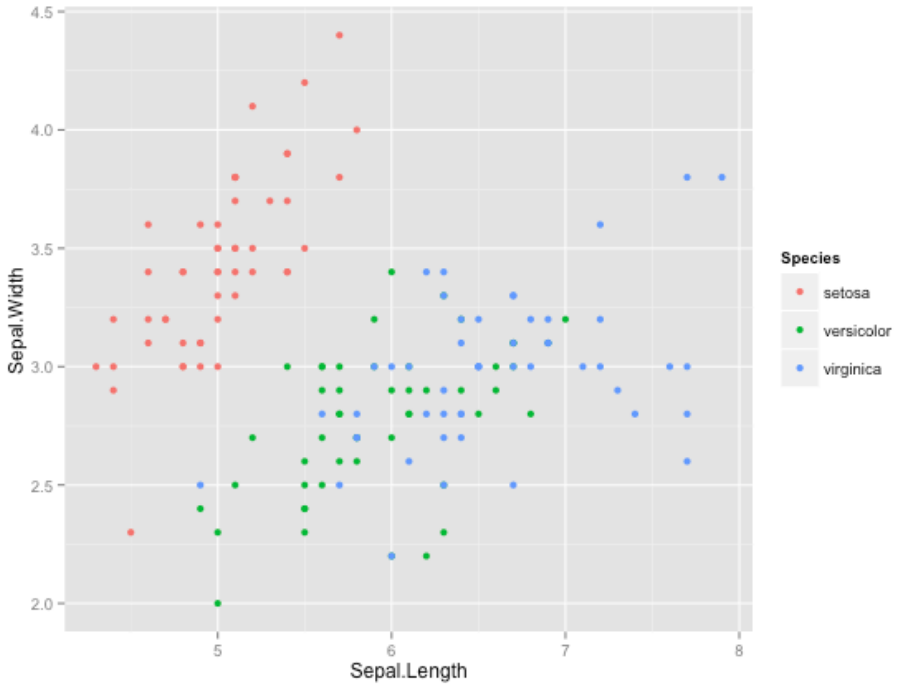
- Data is packaged with R: "iris"

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
40	4/04/10	NA	44.83
40	6/04/10	2	69.68
40	19/04/10	13	44.61
40	1/05/10	12	30.39
40	2/05/10	1	60.73
40	12/05/10	10	50
40	15/05/10	3	3
40	18/05/10	3	36.89
40	19/05/10	1	9.07
40	23/05/10	4	14.01
40	26/05/10	3	16.97
40	31/05/10	5	8.69
…	…	…	…

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 top six-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:

```
> 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
140	4873.97
134	4722.42
123	4333.02
263	3120.58
254	3067.8
199	2737.57

customer_id	visit_date	visit_delta	visit_spend
123	2/06/2010	3	25
123	5/06/2010	3	63.29
123	6/06/2010	1	40.19
123	9/06/2010	3	18.25
123	11/06/2010	2	79.26
123	15/06/2010	4	20.52
123	18/06/2010	3	110.57
123	21/06/2010	3	4.85
123	22/06/2010	1	63.81
123	25/06/2010	3	96.39
123	29/06/2010	4	43.84
…	…	…	…

Challenge 3:

Now plot histogram

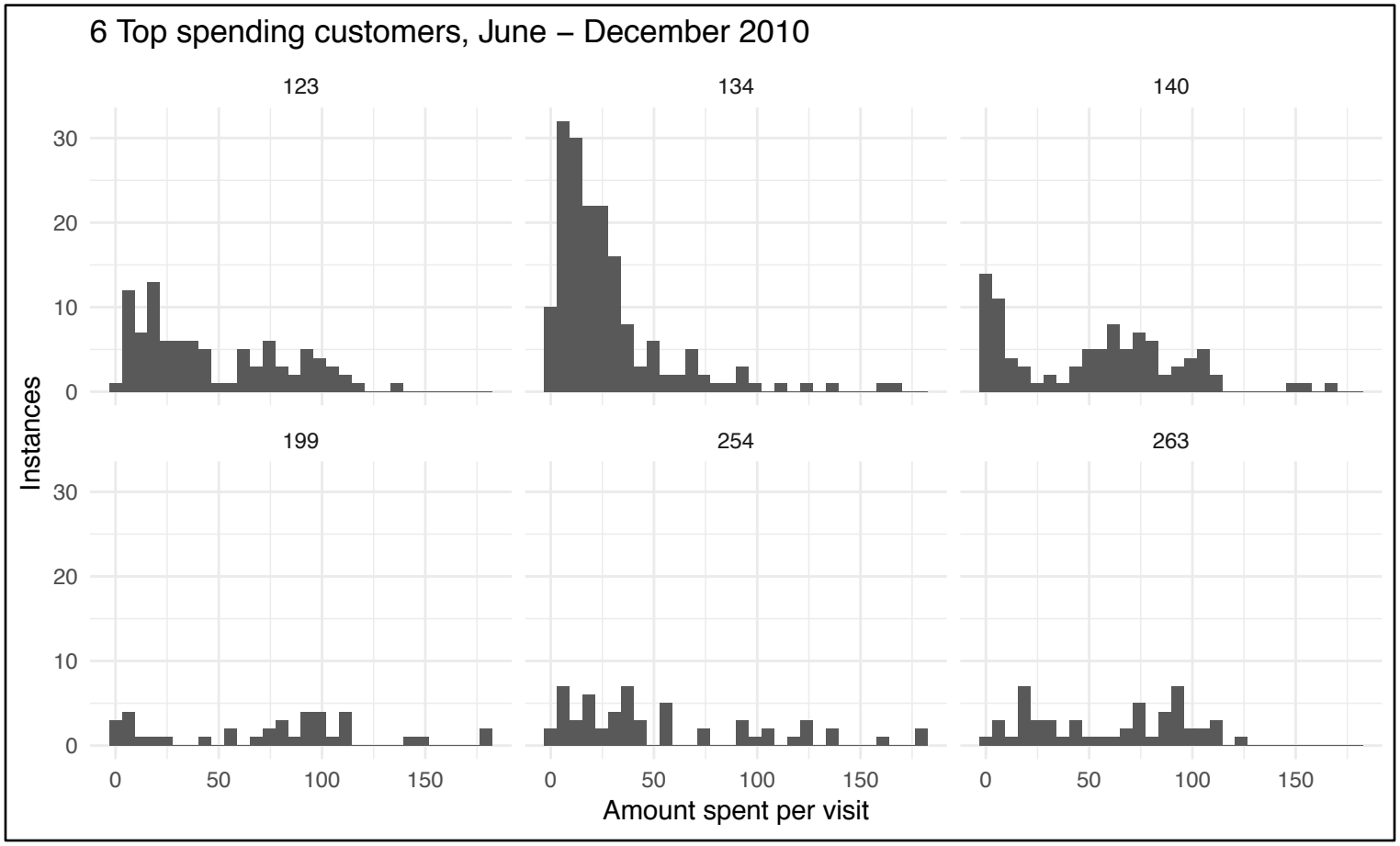
```
>   g  = ggplot(data = DHX6) +
>       geom_histogram(mapping = aes(x = visit_spend)) +
>       theme_minimal() +
>       ggtitle("6 Top spending customers...") +
>       xlab("Amount spent per visit") +
>       ylab("Instances") +
>       facet_wrap(~ customer_id, nrow = 2)
```

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
```

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
…	…	…

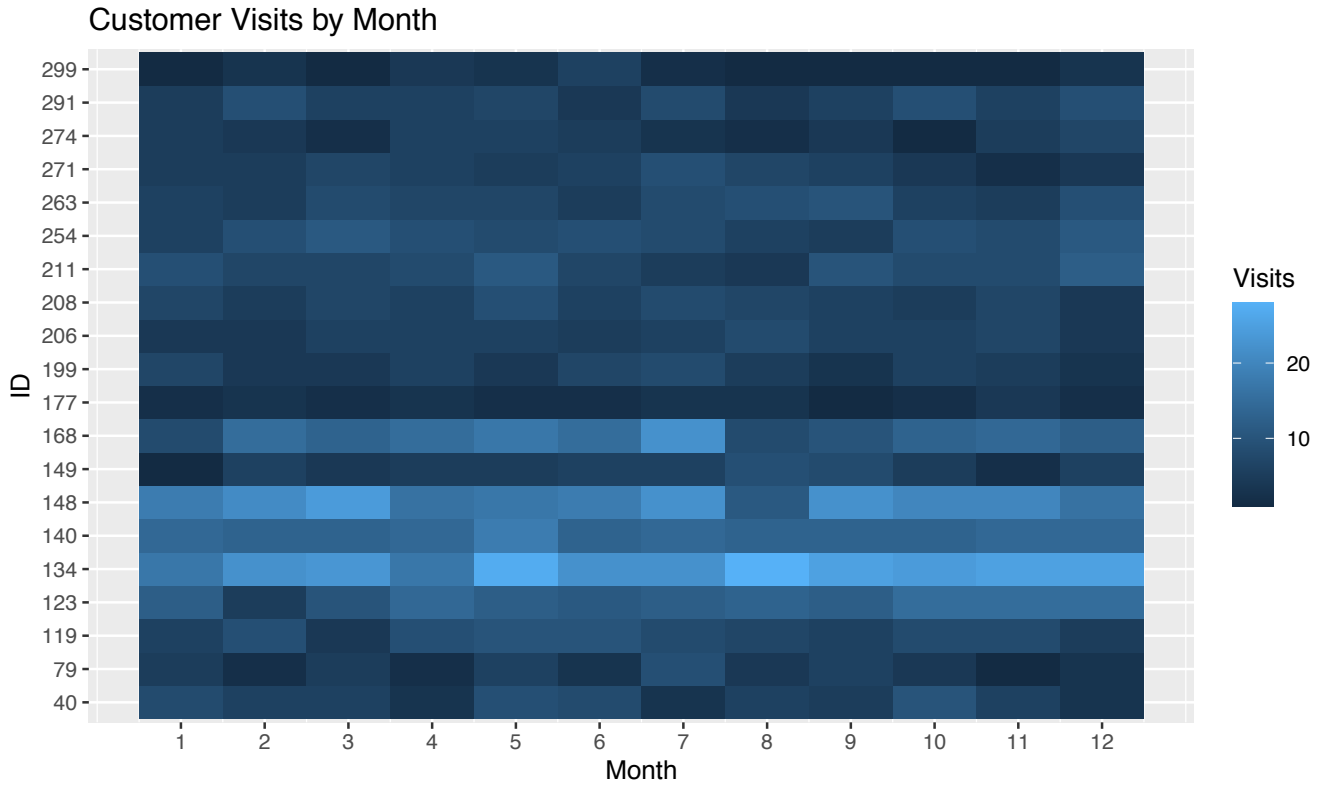
Challenge 4:

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…)



Notes

Acknowledgement

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

Reading

- R for Data Science, Chapters 1 and 2 as introductory material. Chapter 3 for notes on graphing with ggplot2. Chapter 12 for Tidy Data.

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

- A taxonomy of dirty data

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

- Tidy Data

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