University of Exeter

Analytic Workflow and Managing and Cleaning Data

Week 02-BEM2031

Term2: 2023/24

University of Exeter

Today:

- Analyse and identify the components of a **data pipeline**
- List the different stages of **data cleaning** and model development
- Appraise potential issues with data and evaluate which tools are best suited to solve these issues

University of Exeter

Advantages:

Large community of resources

Free and open source with a large number of statistics-related libraries

Excellent visualisations

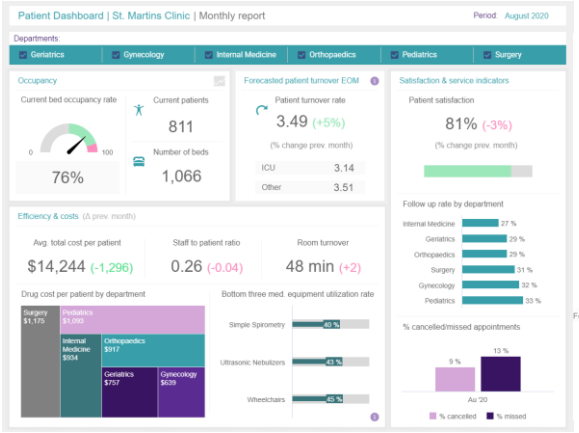
Disadvantages:

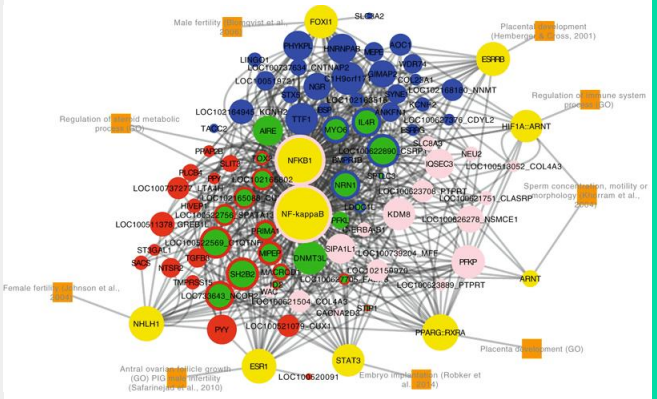
Speed and memory management

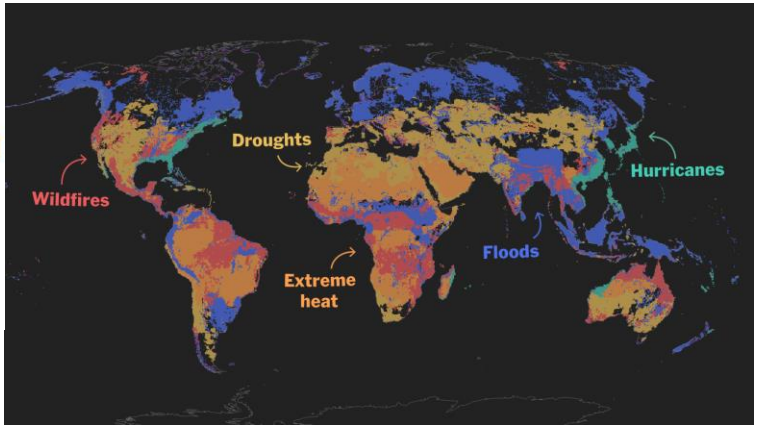
Why R?

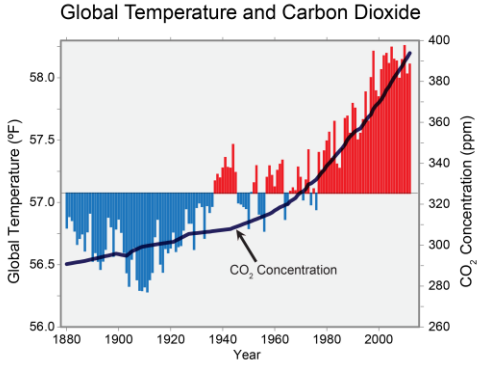
"Knowledge of programming fundamentals certainly helps when adding R to your toolbox, but I wouldn't say it's required to get started.

I wouldn't even say R is for programmers. It's best suited for people that have data-oriented problems they're trying to solve, regardless of their programming aptitude."



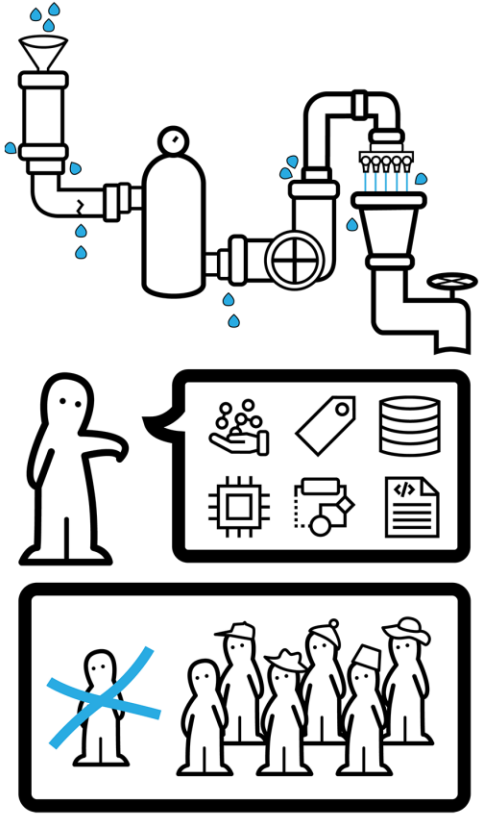




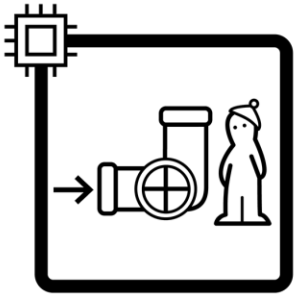


University of Exeter

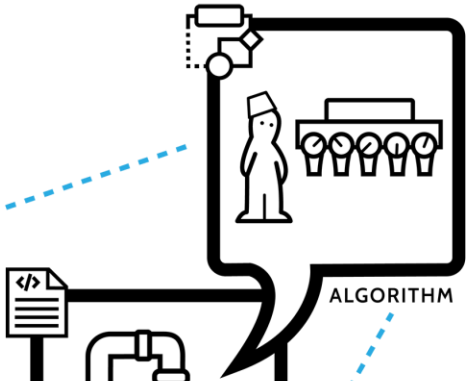
DATA PIPELINE



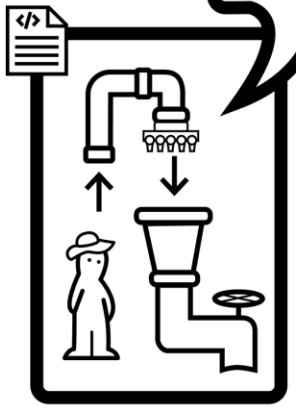
DATA COLLECTION



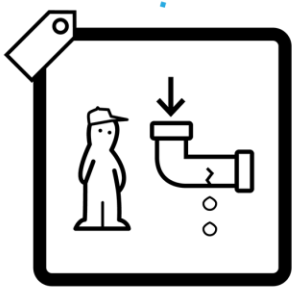
PROCESSING SUBSTRATE



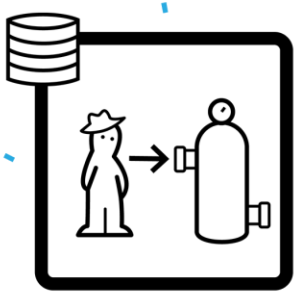
ALGORITHM



CODE

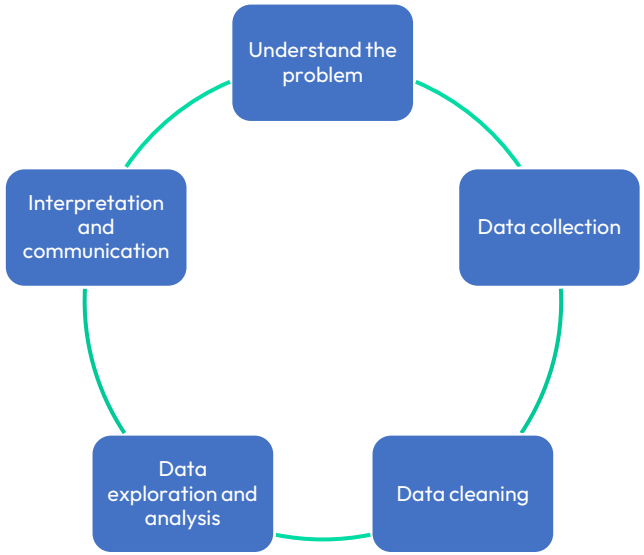


METADATA & CLEANING



STORAGE & RETRIEVAL ARCHITECTURE

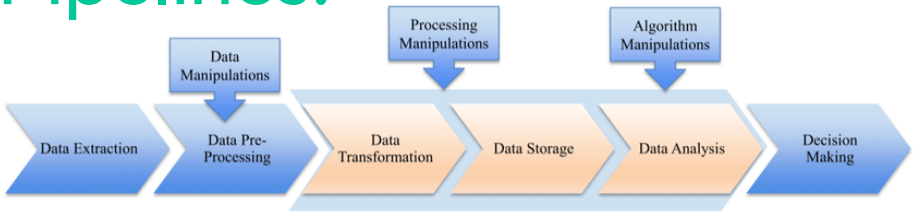
OUTPUT

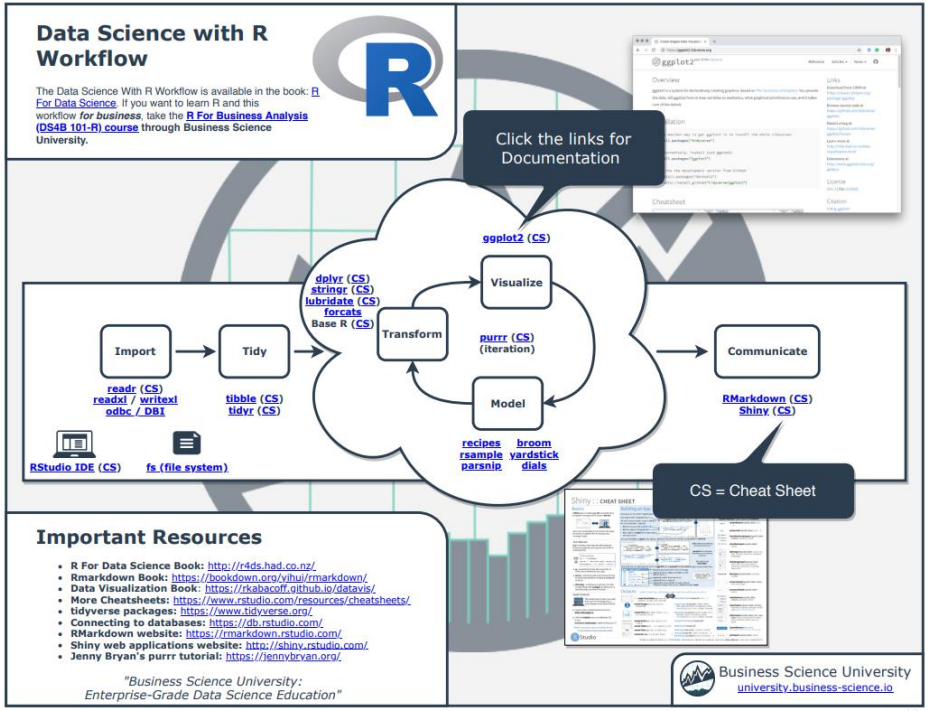


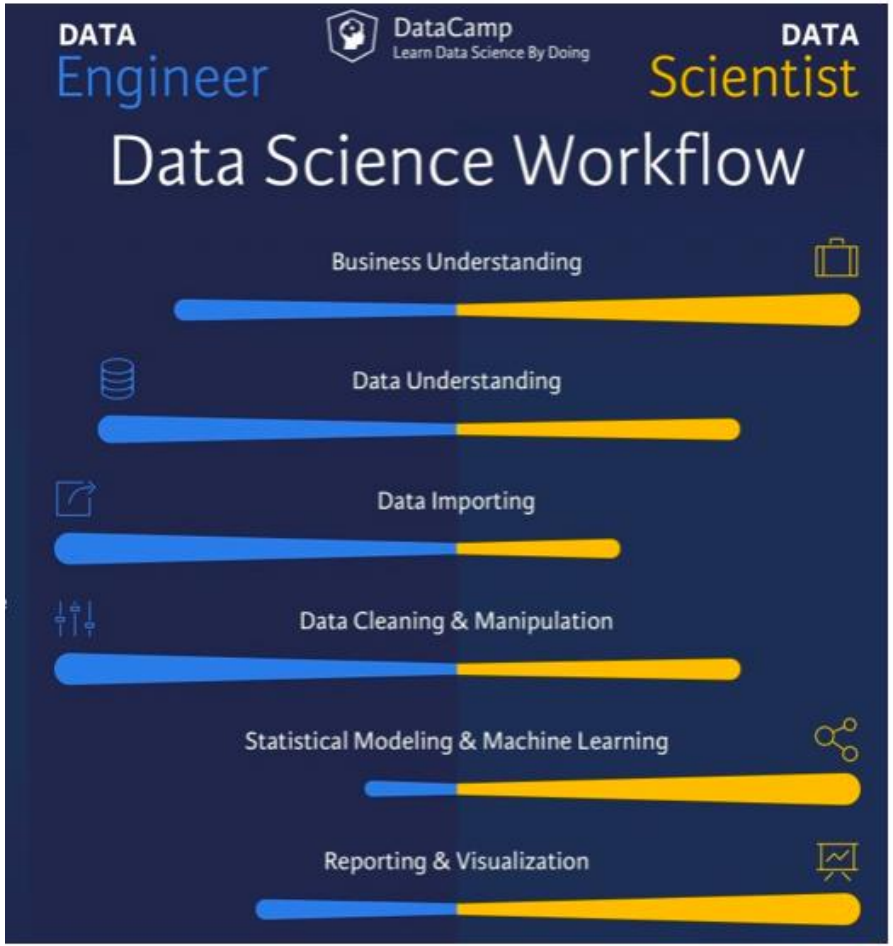
- **Problem definition:** Identifying and understanding business problem or opportunity
- **Data collection:** Gathering and understanding the necessary data from various sources
- **Data Cleaning and Preprocessing**: Refining the data to ensure quality and relevance
- **Data Exploration and Analysis:** Investigating the data for patterns and insights
- **Model building:** Developing descriptive/predictive/prescriptive models
- **Validation and testing:** Assessing the model's performance and accuracy
- **Interpretation and communication**
- **Deployment:** Implementing the model in a real-world setting (or further iteration – back round the loop)

University of Exeter

Pipelines:


Data Extraction
Data Pre-Processing
Data Manipulations
Data Transformation
Processing Manipulations
Data Storage
Data Analysis
Algorithm Manipulations
Decision Making

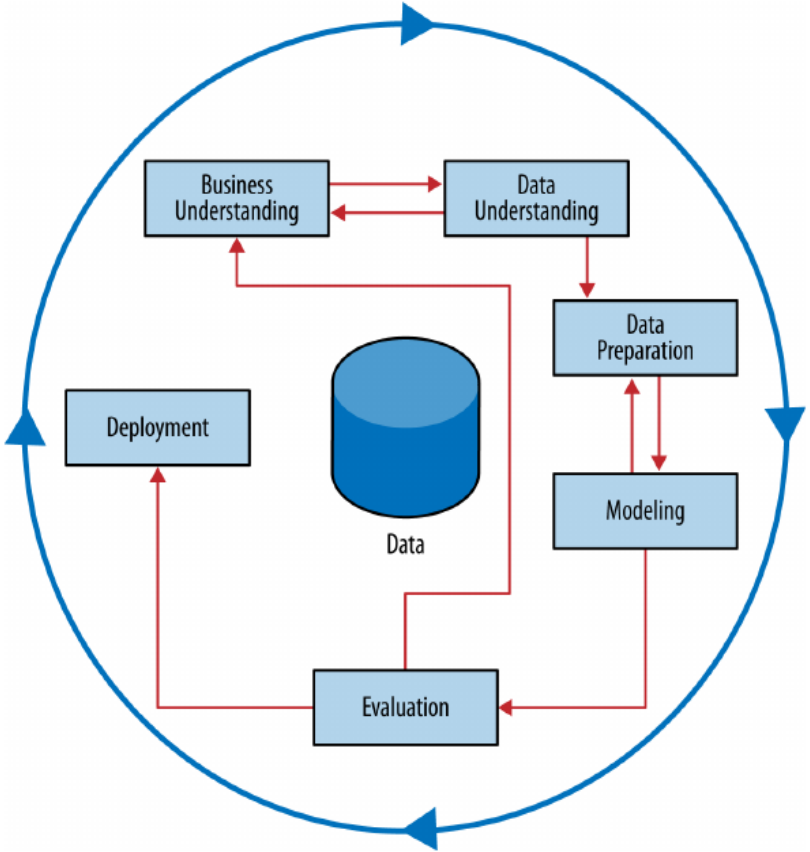

Data Science with R Workflow
The Data Science With R Workflow is available in the book: R For Data Science. If you want to learn R and this workflow for business, take the R For Business Analysis (DS4B 101-R) course through Business Science University.
Click the links for Documentation
ggplot2 (CS)
dplyr (CS)
stringr (CS)
lubridate (CS)
forcats
Base R (CS)
Visualize
Transform
purrr (CS)
(iteration)
Model
Import
Tidy
Communicate
readr (CS)
readxl / writexl
odbc / DBI
tibble (CS)
tidyr (CS)
RMarkdown (CS)
Shiny (CS)
RStudio IDE (CS)
fs (file system)
recipes
rsample
parsnip
broom
yardstick
dials
CS = Cheat Sheet
Important Resources
R For Data Science Book: http://r4ds.had.co.nz/
Rmarkdown Book: https://bookdown.org/yihui/rmarkdown/
Data Visualization Book: https://rkabacoff.github.io/datavis/
More Cheatsheets: https://www.rstudio.com/resources/cheatsheets/
tidyverse packages: https://www.tidyverse.org/
Connecting to databases: https://db.rstudio.com/
RMarkdown website: https://rmarkdown.rstudio.com/
Shiny web applications website: http://shiny.rstudio.com/
Jenny Bryan's purrr tutorial: https://jennybryan.org/
"Business Science University:
Enterprise-Grade Data Science Education"
Business Science University
university.business-science.io


DATA Engineer
DataCamp
Learn Data Science By Doing
DATA Scientist
Data Science Workflow
Business Understanding
Data Understanding
Data Importing
Data Cleaning & Manipulation
Statistical Modeling & Machine Learning
Reporting & Visualization

CRISP-DM

Cross Industry Standard Process for Data Mining - **CRISP-DM**

- Business Understanding
- Data Understanding
- Data Preparation
- Modelling
- Evaluation
- Deployment



Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). CRISP-DM 1.0: Step-by-step data mining guide. SPSS inc, 9(13), 1-73 CRISPMWP-1104-1C.qxd (exeter.ac.uk)

Schröer, C., Kruse, F., & Gómez, J. M. (2021). A systematic literature review on applying CRISP-DM process model. *Procedia Computer Science*, *181*, 526-534.

University of Exeter

CRISP-DM

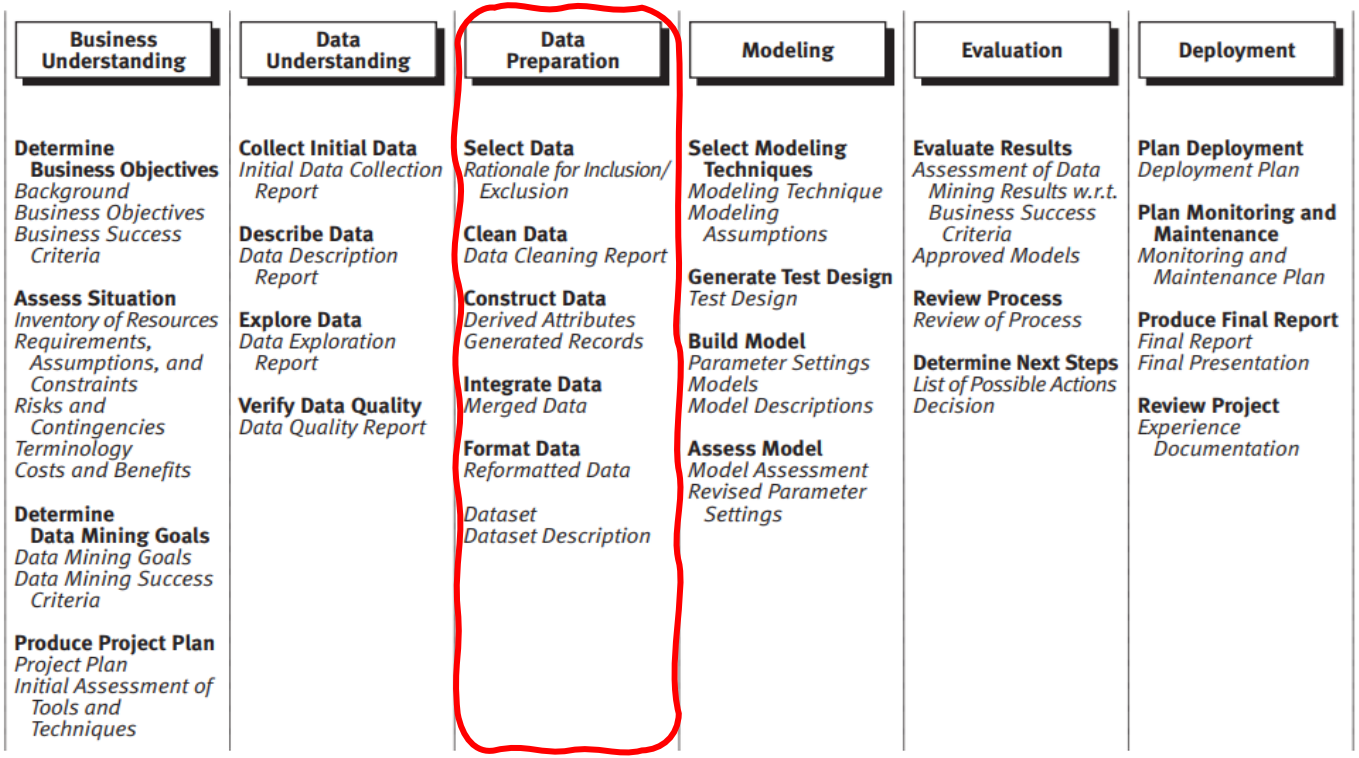
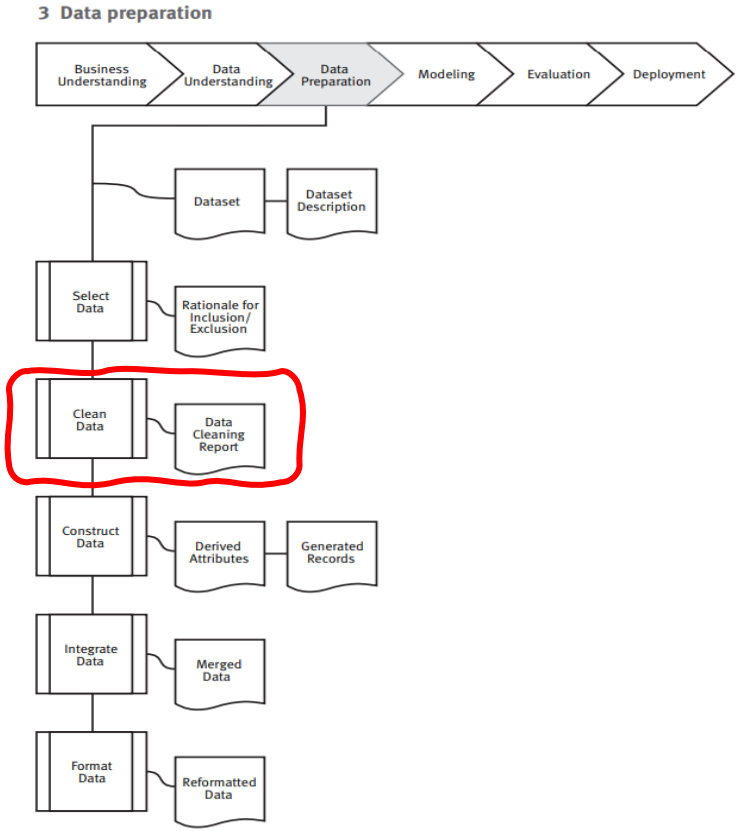


Figure 3: Generic tasks (bold) and outputs (italic) of the CRISP-DM reference model

University of Exeter

Data Preparation

- Selecting
- Cleaning
- Constructing
- Integrating
- Formatting



University of Exeter

Data Preparation

A	B	C	D	E	F	G	H	I	J	K	L	M
LoanRange	BusinessName	Address	City	State	Zip	NAICSCode	BusinessType	RaceEthnicity	Gender	Veteran	NonProf	JobsRetained
a $5-10 million	A G EQUIPMENT COMPANY	3401 W ALBANY ST	BROKEN ARROV	OK	74012	333132	Corporation	Unanswered	Unanswered	Unanswered		470
a $5-10 million	ADVANTAGE ENERGY SERVICES LLC	1010 N MAIN ST	MCALESTER	OK	74501	211120	Limited Liability	Unanswered	Unanswered	Unanswered		115
a $5-10 million	AMERICAN PIPING INSPECTION, IN	17110 E Pine	TULSA	OK	74037	238210	Corporation	White	Male Owned	Non-Veteran		344
a $5-10 million	AXH HOLDINGS	2230 E 49TH ST	TULSA	OK	74105	332410	Limited Liability	Unanswered	Unanswered	Unanswered		409
a $5-10 million	B & H CONSTRUCTION, LLC	301 James Dean Dr	NORMAN	OK	73072	237110	Limited Liability	Unanswered	Unanswered	Unanswered		250
a $5-10 million	BECCO CONTRACTORS INC	13737 E 46TH ST N	TULSA	OK	74116	237310	Corporation	Unanswered	Unanswered	Unanswered		0
a $5-10 million	BENNETT CONSTRUCTION, INC.	525 CENTRAL PARK DR	OKLAHOMA CIT	OK	73105	236117	Corporation	White	Male Owned	Non-Veteran		429
a $5-10 million	BERENDSEN FLUID POWER, INC.	401 S Boston Ave, Ste	TULSA	OK	74103	423830	Corporation	Unanswered	Unanswered	Unanswered		280
a $5-10 million	BOFS MANAGEMENT LLC	210 Park Ave	OKLAHOMA CIT	OK	73102	213112	Limited Liability	Unanswered	Unanswered	Unanswered		494
a $5-10 million	CHC HOLDINGS	3105 S MERIDIAN AVE	OKLAHOMA CIT	OK	73119	621610	Limited Liability	Unanswered	Unanswered	Unanswered		386
a $5-10 million	COLLISION WORKS OF OKLAHOMA, LLC	3224 SE 29TH ST	OKLAHOMA CIT	OK	73115	811121	Limited Liability	Unanswered	Unanswered	Unanswered		449
a $5-10 million	COMPSOURCE MUTUAL INSURANCE COM	1901 N Walnut	OKLAHOMA CIT	OK	73105	524210	Corporation	Unanswered	Unanswered	Unanswered		289
a $5-10 million	COVERCRAFT INDUSTRIES, LLC	100 Enterprise B;vd	PAULS VALLEY	OK	73075	314999	Limited Liability	Unanswered	Female Owne	Unanswered		475
a $5-10 million	D & M CARRIERS LLC	8125 SW 15TH ST	OKLAHOMA CIT	OK	73128	484110	Limited Liability	Unanswered	Unanswered	Unanswered		495
a $5-10 million	DELAWARE RESOURCE GROUP OF OKLAH	3220 Quail Springs Par	OKLAHOMA CIT	OK	73134	336413	Corporation	American Indi	Male Owned	Unanswered		500
a $5-10 million	FUSION INDUSTRIES, LLC	700 NE 63RD ST	OKLAHOMA CIT	OK	73105	237130	Limited Liability	White	Male Owned	Non-Veteran		217
a $5-10 million	GDH CONSULTING INC	6100 S YALE AVE	TULSA	OK	74136	561320	Corporation	Unanswered	Unanswered	Unanswered		500
a $5-10 million	GRIFFIN COMMUNICATIONS, LLC	7401 N Kelley Ave	OKLAHOMA CIT	OK	73111	515210	Limited Liability	Unanswered	Unanswered	Unanswered		434
a $5-10 million	HAC, INC	390 NW 36TH ST	OKLAHOMA CIT	OK	73105	445110	Subchapter S Co	Unanswered	Unanswered	Unanswered		500
a $5-10 million	IMAGENET CONSULTING LLC	913 N. Broadway Ave.	OKLAHOMA CIT	OK	73102	423430	Limited Liability	Unanswered	Unanswered	Unanswered		388
a $5-10 million	INCEED LLC	907 S DETROIT AVE	TULSA	OK	74120	561311	Limited Liability	White	Male Owned	Non-Veteran		478
a $5-10 million	INDUSTRIAL PIPING SPECIALISTS INC.	606 N. 145th East Ave	TULSA	OK	74116	237120	Subchapter S Co	White	Male Owned	Non-Veteran		421
a $5-10 million	LATSHAW DRILLING COMPANY, LLC	4500 S. 129th E Ave, S	TULSA	OK	74134	211120	Limited Liability	Unanswered	Unanswered	Unanswered		
a $5-10 million	LEGEND ENERGY SERVICES, LLC	5801 N BROADWAY EX	OKLAHOMA CIT	OK	73118	213112	Limited Liability	Unanswered	Unanswered	Unanswered		229
a $5-10 million	LIFE. CHURCH OPERATIONS, LLC	4600 E. 2nd Street	EDMOND	OK	73034	813110	Non-Profit Orgar	Unanswered	Unanswered	Unanswered	Y	451

Select data

Task

Select data

Decide on the data to be used for analysis. Criteria include relevance to the data mining goals, quality, and technical constraints such as limits on data volume or data types. Note that data selection covers selection of attributes (columns) as well as selection of records (rows) in a table.

Output

Rationale for inclusion/exclusion List the data to be included/excluded and the reasons for these decisions.

University of Exeter

Clean data

Task

Clean data

Raise the data quality to the level required by the selected analysis techniques. This may involve selection of clean subsets of the data, the insertion of suitable defaults, or more ambitious techniques such as the estimation of missing data by modelling.

Data cleaning report

Describe what decisions and actions were taken to address the data quality problems reported during the Verify Data Quality task of the Data Understanding phase. Transformations of the data for cleaning purposes and the possible impact on the analysis results should be considered.

Construct, Integrate, Format data

Deriving attributes:

e.g. A new column combining data from two other columns

Merging data:

e.g. joining together other datasets or tables, aggregating data

Format data:

e.g. Sorting, re-shaping, removing punctuation from headers etc.

University of Exeter

University of Exeter

Clean data

ID	Cust-Name	DOB	Home_ town	Postcode	Gender	Product Rating	Feedback
C1001	J.Smith	10/02/1981	London	LD1 2AB	M	5	excelent
C1002	Mary Williams	2.02/1981	Exeter	EX1 2AB		4	good
C1003	Emily	3/02/1980	Plymouth	PL1 2AB		5	excellent
C1004	Sara Roberts	1981-02-5	York		F	3	fair
C1005	Dave	15/02/1981			M		
C1006	June Ford	21/04/1981	Glasgow	GA1 2AB	F		good
C1007	Winter Phillips	11/12/1981	York	Y01 2AB	F	3	goOd
C1008	Julie	19/08/1981		PL2 2AB	F	2	Acceptab;le
C1009	Andi Smith	1 May 1980	Oxford	OX1 2AB		1	poor
C1010	Paula Penn	10/11/1981	Stafford	ST1 2AB	F	10	brilliant

Data Cleaning

- ID
- Variable type
- Extreme values
- Missing values
- Missing data
- Duplicate values
- Text processing – stop word removal, case conversion, special character removal

Data Preparation

- Derived attributes
- Merge with other datasets for extra insight
- Reformat

Re-shaping data

Chipset	Site A	Site B	Site C	Site D
Hello	15	8	30	27
Snapdragon	29	17	14	42
Dimensity	10	19	25	23

Wide

Chipset	Site	DBH (mill)
Hello	A	15
Hello	B	8
Hello	C	30
Hello	D	27
Snapdragon	A	29
Snapdragon	B	17
Snapdragon	C	14
Snapdragon	D	42
Dimensity	A	10
Dimensity	B	19
Dimensity	C	25
Dimensity	D	23

Long

University of Exeter

Basic data types in R:

- `numeric` - (10.5, 55, 787)
- `integer` - (1L, 55L, 100L, where the letter "L" declares this as an integer)
- `Character` (e.g postcodes, IDs, names)
- `factor` ('female', 'male')('excellent', 'okay', 'poor')
- `logical` (a.k.a. boolean) - (TRUE or FALSE)

We can use the `class()` function to check the data type of a variable

Or glimpse() to see the class of each column in a dataset

is.na() to find missing values

Useful commands for inspecting data in R

head: this by default prints the first 6 rows of the dataframe

tail: this by default prints the last 6 rows to the console

str: this prints the structure of your dataframe

dim: this by default prints the dimensions, that is, the number of rows and columns of your dataframe

colnames: this prints the names of the columns of your dataframe

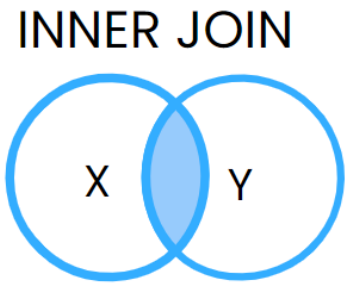
summary: this function provides summary statistics on the columns of the data frame

[How to merge in R (youtube.com)]

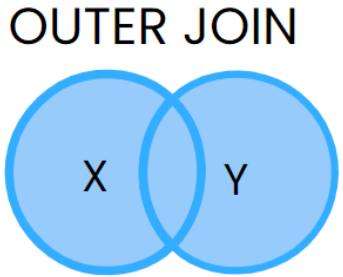
University of Exeter

Merging / joining

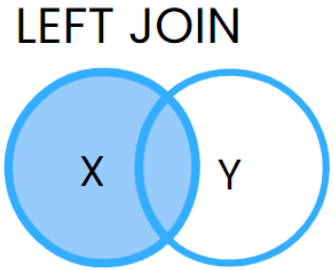
Inner join consists of merging two dataframes in one that contains the common elements of both, as described in the following illustration:



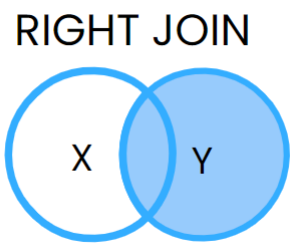
Full (outer) join merges all the columns of both data sets into one for all elements:



Left (outer) join consists of matching all the rows in the first data frame with the corresponding values on the second.



The right join in R is the opposite of the left outer join. In this case, the merge consists of joining all the rows in the second data frame with the corresponding on the first.



[How to merge data in R using R merge, dplyr, or data.table | InfoWorld]

University of Exeter

Tasks for next week: Data visualisation

Read: Storytelling with Data (pdf on ELE)

Watch: Turning bad charts into compelling stories Dominic Bohan | TEDxYouth@Singapore - YouTube

Listen: Data Is Personal with Evan Peck – Data Stories

Do the Rstudio primer/visualisation Posit Cloud

Download and save files for Week 3 workshop

University of Exeter

Any questions?

?