E+ ADSEE

JobMarket Signalling
3. Data.Selection

Dr. Alan Berg

ADSEE

Co-funded by the
Erasmus+ Programme
of the European Union

Overview

In this section we review an opportunistic source of data, job vacancies and an example of the dimensions associated with the data.

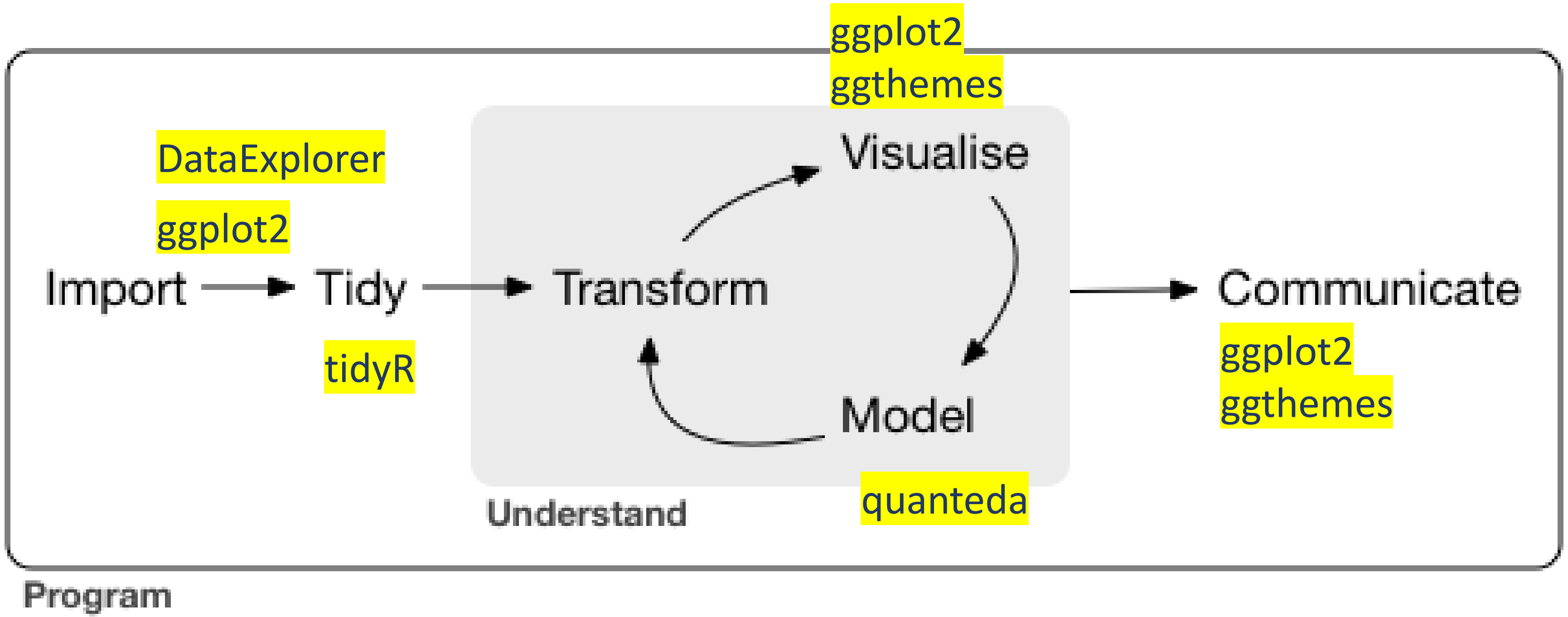
The coding section is describes building a reproducible data pipeline for the Job descriptions.

In the sample data set development module we will look data which is freely available on the Internet that would enrich the current data set. Enrichment is an iterative process which may change the details of the business problem you wish to solve.

Why these methods?

Because we want to:
Process data through a consistent, reproducible and minimalistic pipeline.

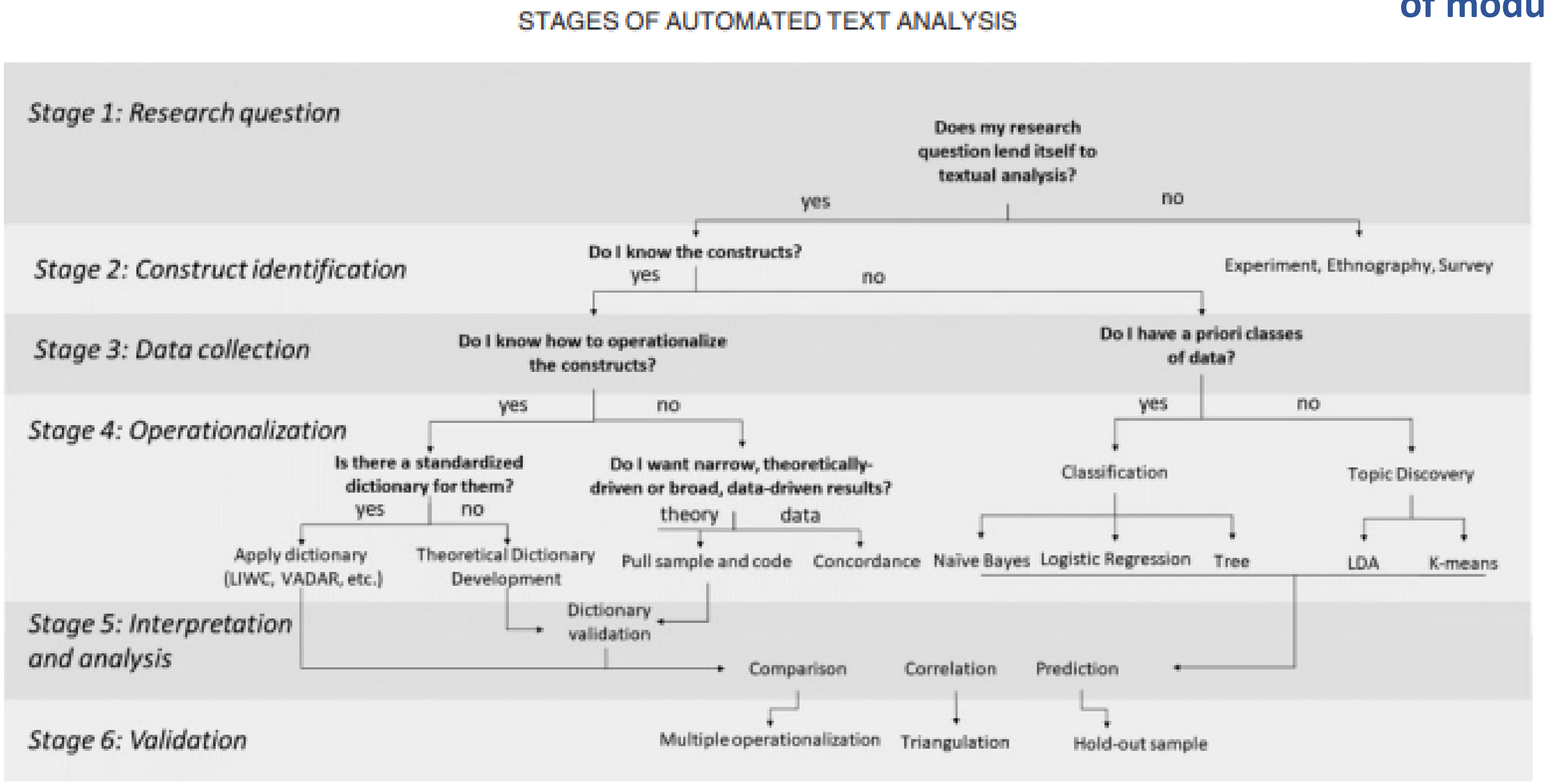
R for DataScience



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

Workflows

Will expand on this process over a number of modules



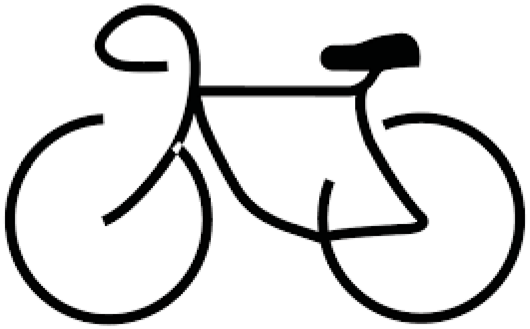
https://www.researchgate.net/publication/324495873_Automated_Text_Analysis_for_Consumer_Research

Example Research questions

Research question	Text	Linguistic aspect	Source
Dictionary-based—Comparison			
How does temporal and spatial distance affect emotions after a tragic event?	Twitter	Semantic	Doré et al. 2015
How do power and affiliation vary by political ideology?	Transcripts (chatrooms, State of the Union), news websites	Semantic	Fetterman et al. 2015
What explains representational gender bias in the media?	Newspapers	Phatic	Shor et al. 2015
How does personal pronoun use in firm-customer interactions impact customer attitude?	Transcripts	Pragmatic	Packard, Moore, and McFerran 2016
Why don't major crises like oil spills provoke broad changes in public discourse concerning the systemic risks inherent to a carbon-dependent economy?	Newspaper articles	Semantic	Humphreys and Thompson 2014
Do people modify warmth to appear competent (and vice versa) when doing impression management?	Emails	Semantic	Holoien and Fiske 2013
Does social hierarchy affect language use? In what ways?	Emails	Pragmatic	Kacewicz et al. 2014
Do Christians and atheists vary in their language use?	Twitter	Semantic	Ritter et al. 2013
How does someone's communication style change based on private versus public communication?	Facebook wall posts and private messages	Semantic, pragmatic	Bazarova 2012
How do letters to shareholders differ in a period of economic growth versus recession?	Letters to shareholders	Semantic	Pollach 2012
Are people with the same linguistic style more likely to form a romantic relationship?	Transcripts, instant messages	Stylistic, pragmatic	Ireland et al. 2011
How does happiness change throughout the lifecycle?	Personal blogs	Semantic	Mogilner et al. 2011
Dictionary-based—Correlation			
Do depressed patients use more self-focused language?	Written essays	Semantic	Brockmeyer et al. 2015

Phatic
relating to, or being speech used for social or emotive purposes rather than for communicating information

Exercise



- Look at the appendix of https://www.researchgate.net/publication/324495873_Automated_Text_Analysis_for_Consumer_Research for the type of research questions that you can ask
- Try and write 5 research questions within your context

Background

- Data is costly and dirty
- Data is complex
- The dimensions available in the UK dataset
- Data selection depends on the problem you wish to solve

Data Selection

names(sample)

"PositionAdID" "PositionID" "**JobTitle**" "City"
"**PostalCode**" "SalaryTypeID"

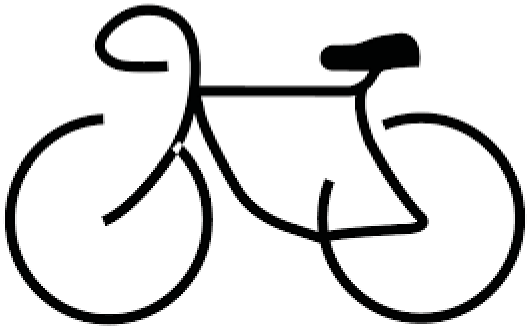
"SalaryFrom" "SalaryTo" "CareerLevelID" "**EducationLevelID**"
"DateActive" "DateExpires"

"JobAttributeIDs" "JobCategoryIDs" "OccupationIDs" "IndustryIDs"
"Keywords" "**JobBody**"

Classes in programming languages

- Character
- Date
- Integer
- Factor
- Data.frame
- List
- Tibble

Exercise



Data Catalogue

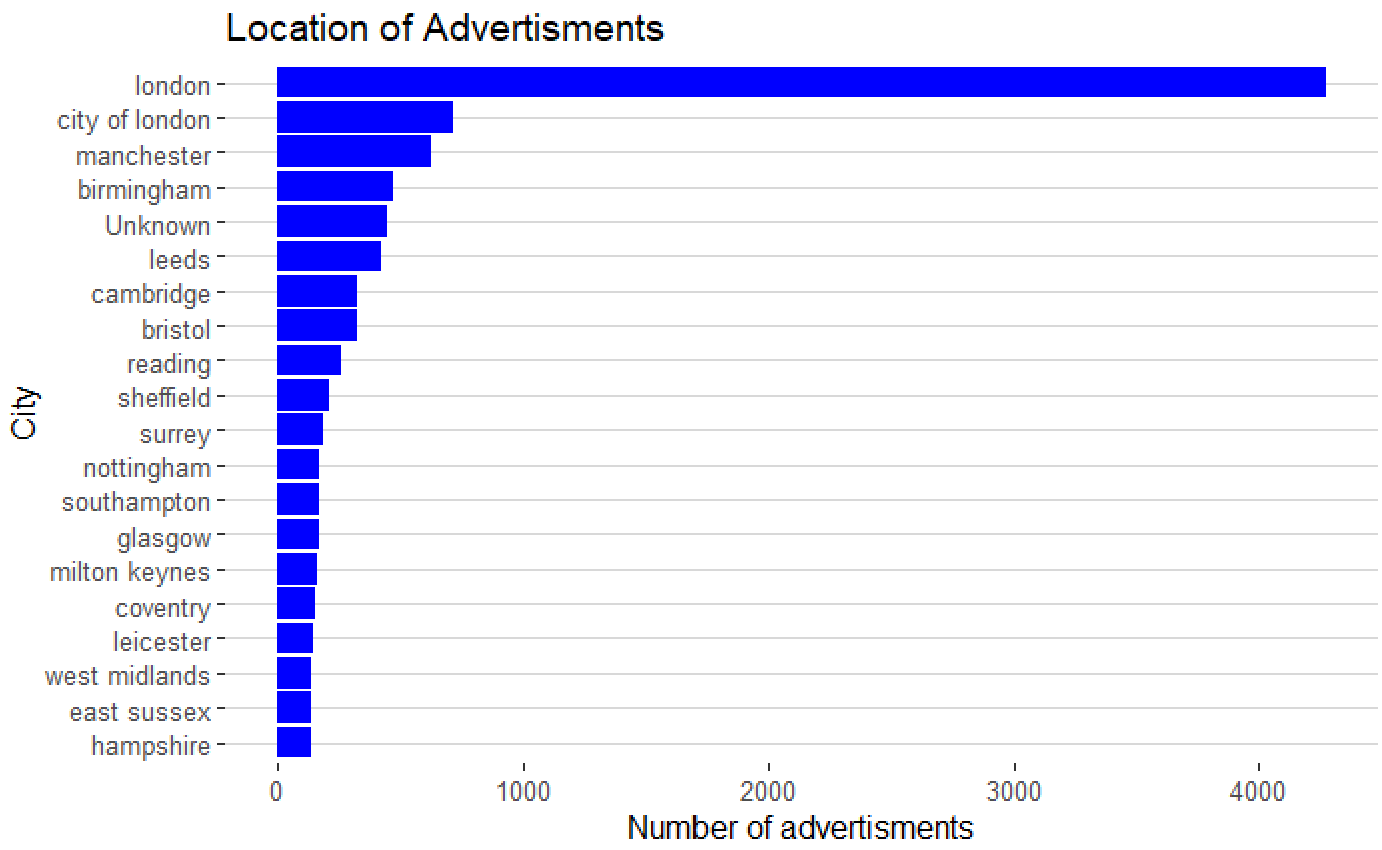
Data is not always clean and sometimes you need to guess the data catalogue:

1. Without visiting the Internet: Review the dataset sample and write a description of each column.
2. Now search for the column names on the Internet and improve on your descriptions
3. For each column label if you think the data is privacy sensitive with justification.

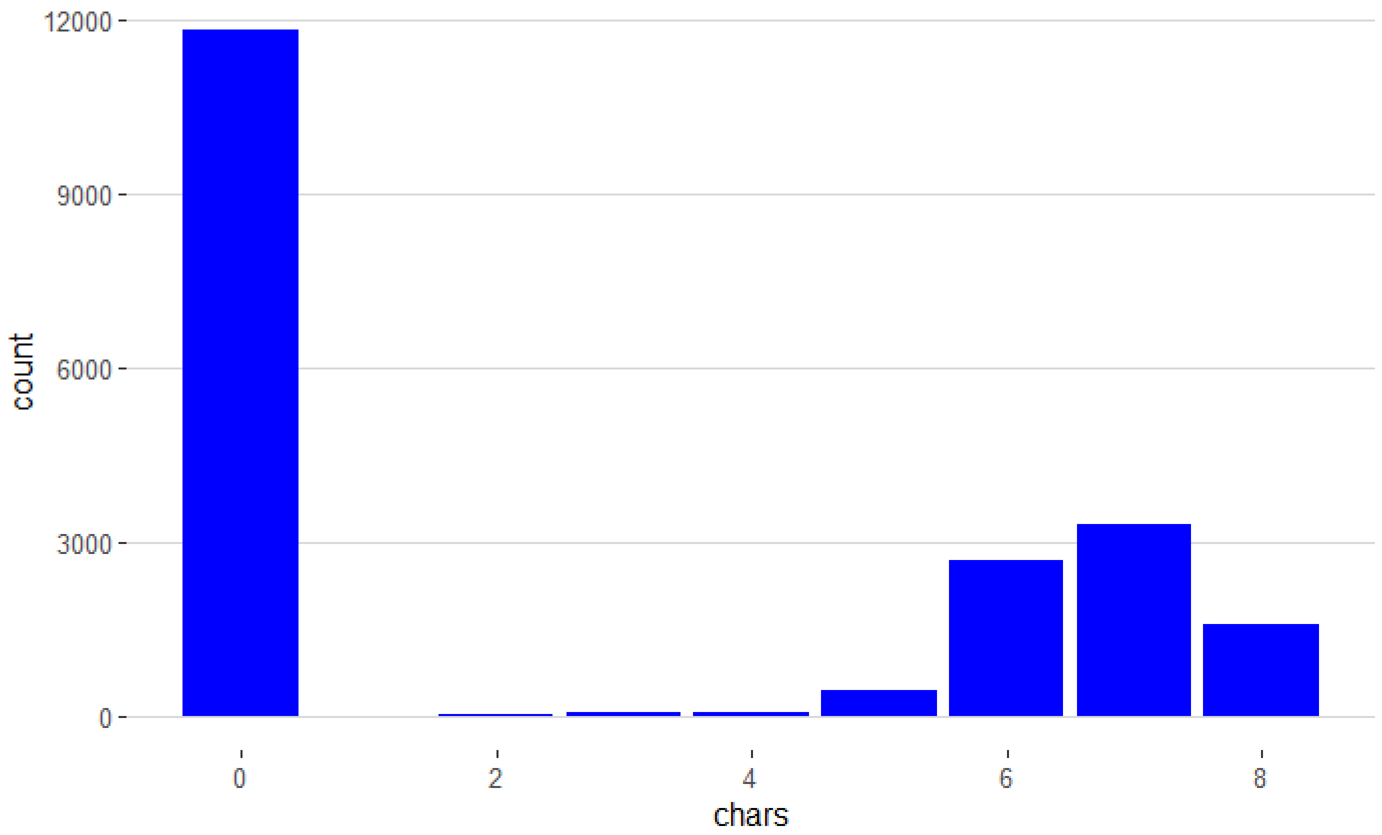
	City	Freq
	<fctr>	<int>
716	London	4110
286	City of London	713
784	Manchester	614
116	Birmingham	461
675	Leeds	413
216	Cambridge	321
170	Bristol	313
1		308
988	Reading	257
1064	Sheffield	206

	City <fctr>	Freq <int>
632	london	4276
249	city of london	719
693	manchester	628
106	birmingham	470
1125	Unknown	448
600	leeds	425
192	cambridge	323
152	bristol	322
869	reading	259
938	sheffield	211

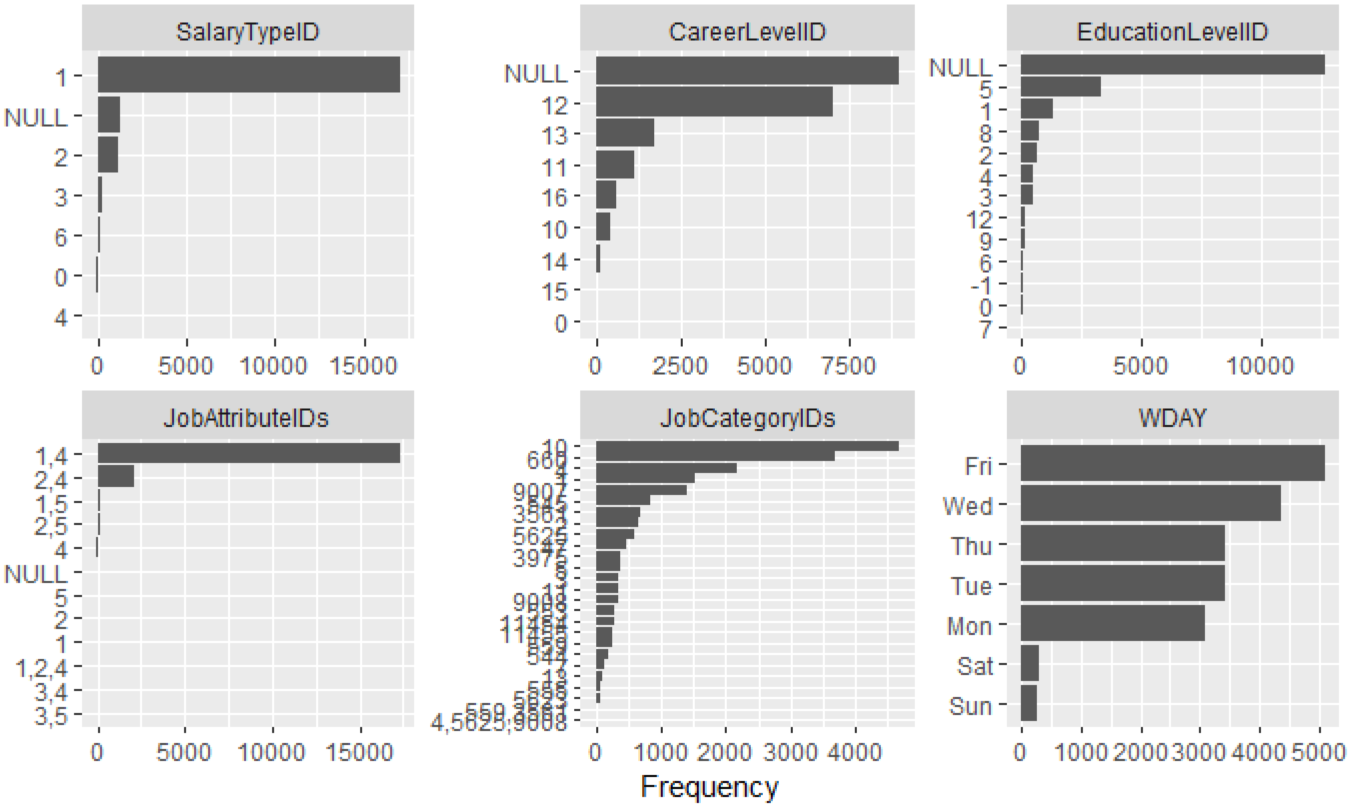
Visualize, visualize, visualize



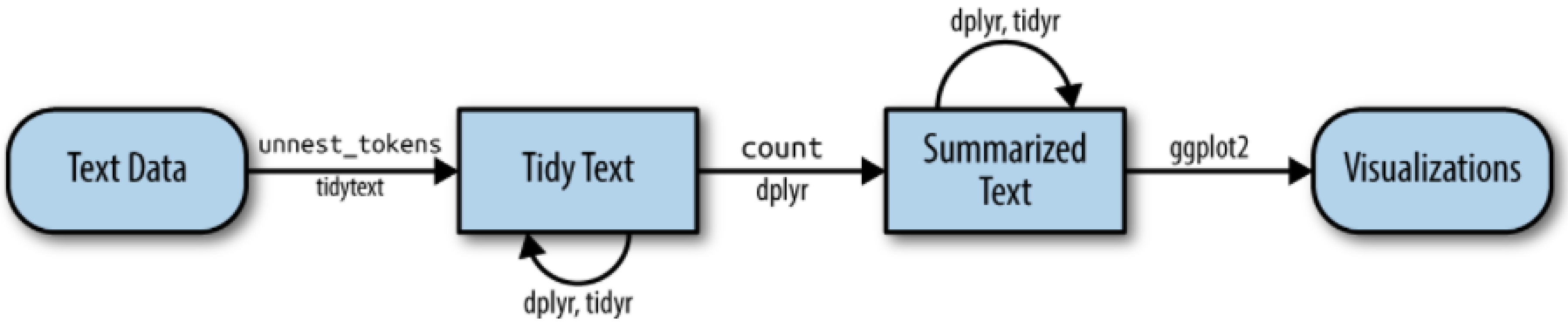
Visualize, visualize, visualize



Be lazy: Automate and visualize

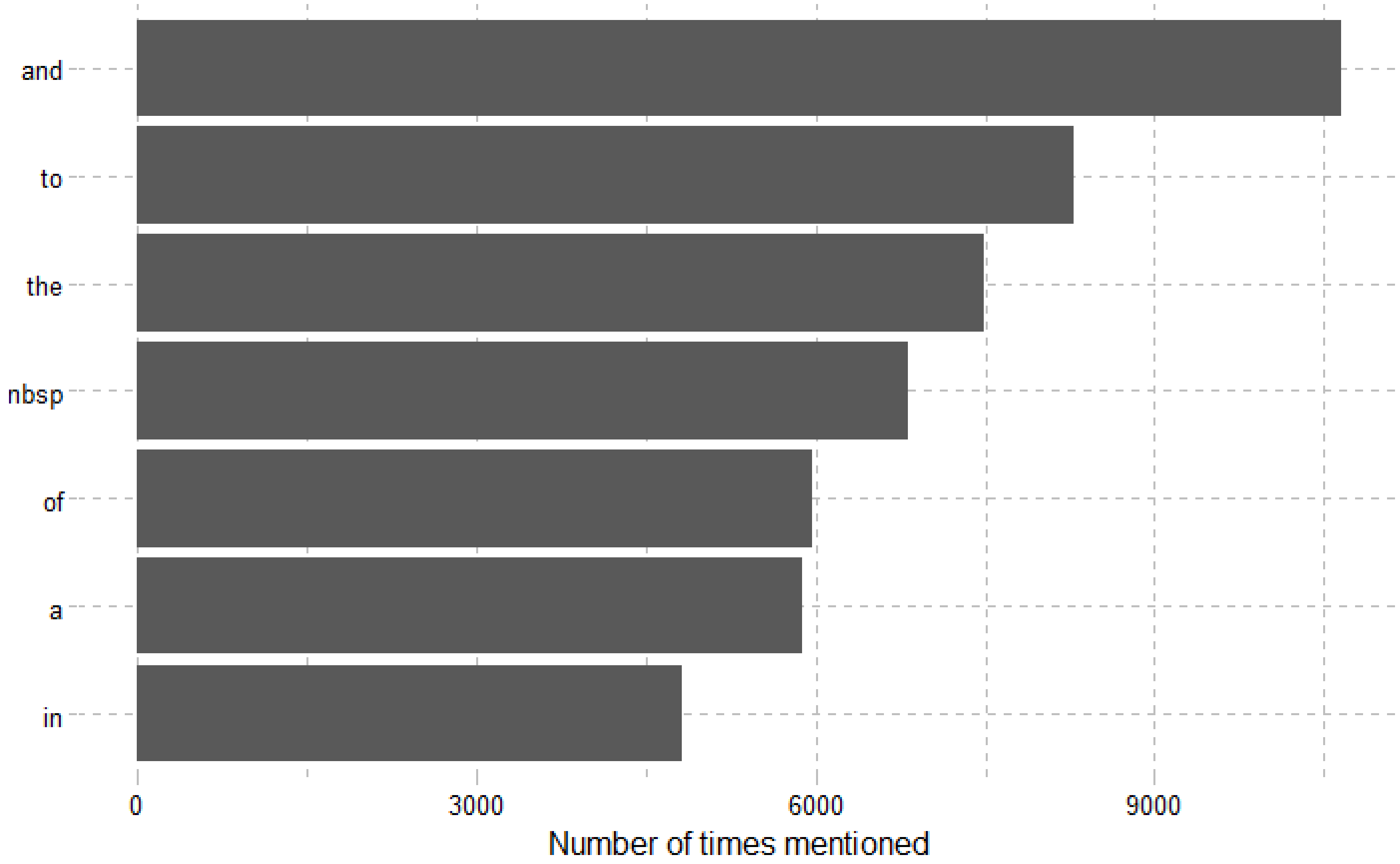


Pipeline



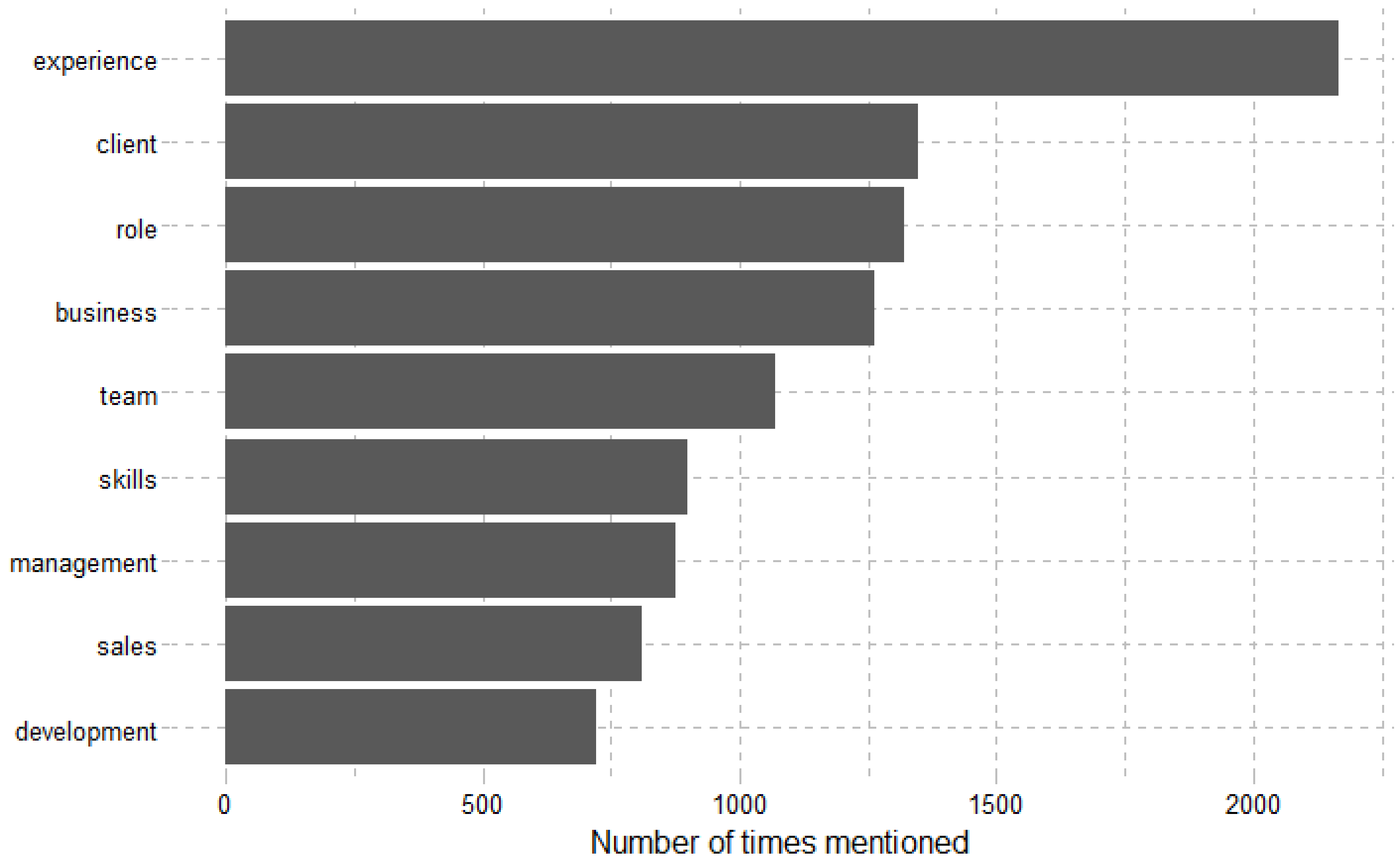
Text Data
unnest_tokens
tidytext
Tidy Text
dplyr, tidyr
count
dplyr
dplyr, tidyr
Summarized Text
ggplot2
Visualizations

Raw



and
to
the
nbsp
of
a
in
0
3000
6000
9000
Number of times mentioned

Stop Words removed



Stemming

word <chr>	n <int>
experi	2165
client	1348
role	1322
busi	1264
team	1069
skill	900
manag	875
sale	811
develop	722
support	695

1-10 of 10 rows

Recipe

```
library(SnowballC)
my.freq.10 <- head(my.freq,n=10)
my.freq.10 %>% mutate(word = wordStem(word))
```

Bigrams

word <chr>	n <int>
ms word	222
word format	214
track record	181
business development	167
communication skills	166
cv nexusjobs.com	143
customer service	140
digital marketing	131
financial services	119
sql server	118

1-10 of 10 rows

Not covered

- Lemmentisation
- Removing stop words that are Parts Of Speech (noun, verb, adjective)
- Removing slang
- Disambiguation: Sitting on a bank, bank charges
- Context of the whole document or sentence

Lemmentisation

https://en.wikipedia.org/wiki/Lemmatisation

- In many languages, words appear in several *inflected* forms. For example, in English, the verb 'to walk' may appear as 'walk', 'walked', 'walks' or 'walking'. The base form, 'walk', that one might look up in a dictionary, is called the *lemma* for the word. The association of the base form with a part of speech is often called a *lexeme* of the word.
- Lemmatisation is closely related to stemming. The difference is that a stemmer operates on a single word *without* knowledge of the context, and therefore cannot discriminate between words which have different meanings depending on part of speech. However, stemmers are typically easier to implement and run faster.

- The word **"better"** has **"good"** as its lemma. This link is missed by stemming, as it requires a dictionary look-up.
- The word "walk" is the base form for the word "walking", and hence this is matched in both stemming and lemmatisation.

Named Entity Extraction

	doc_id	sentence_id	term_id	term	entity
1	1	1	1	Ik	O
2	1	1	1	heet	O
3	1	1	1	Karel	B-PER
4	1	1	1	je	O
5	1	1	1	kan	O
6	1	1	1	me	O
7	1	1	1	bereiken	O
8	1	1	1	op	O
9	1	1	1	paul@duchanel.be	B-EMAIL
10	1	1	1	of	O
11	1	1	1	www.duchanel.be	B-URL

Pipeline – Start Simple

```
# Look at bigrams via adding the parameters token = "ngrams", n = 2
my.words.bi <- my.job %>%  unnest_tokens(word, text, token = "ngrams", n = 2)
my.freq.bi <- my.words.bi %>%
  count(word, sort = TRUE)
```

Overtime improve

```
my.freq.bi.cleaned <-  my.freq.bi %>% separate(word, c("word1", "word2"), sep = " ") %>%
  filter(!word1 %in% stop_words$word) %>%
  filter(!word2 %in% stop_words$word) %>%
  unite(word,word1, word2, sep = " ")
```

Reading the literature helps

Studying the UK Job Market During the COVID-19 Crisis with Online Job Ads - Rudy Arthur

In particular we will look at

- The number of job postings by date.
- Time series of vacancies by sector.
- Time series of vacancies by geographic region.
- The distribution of salary; type of contract (full time, part time, contract) and mode of work (permanent or temporary) before and after the COVID crisis hit in 2020.

Data selection

Depends on the Research Question

- Aggregations
- Opportunistic
- Exploratory Data Analysis
- Feature selection in machine Learning followed by expert review

Confounders

Why does the Government sources react slower than companies? Is this an actionable signal. If so a signal for whom?

Can we look at the difference in wording between advertisers?

How would you design a feedback cycle to improve your analysis?

KWIC government vs Non Government Sources

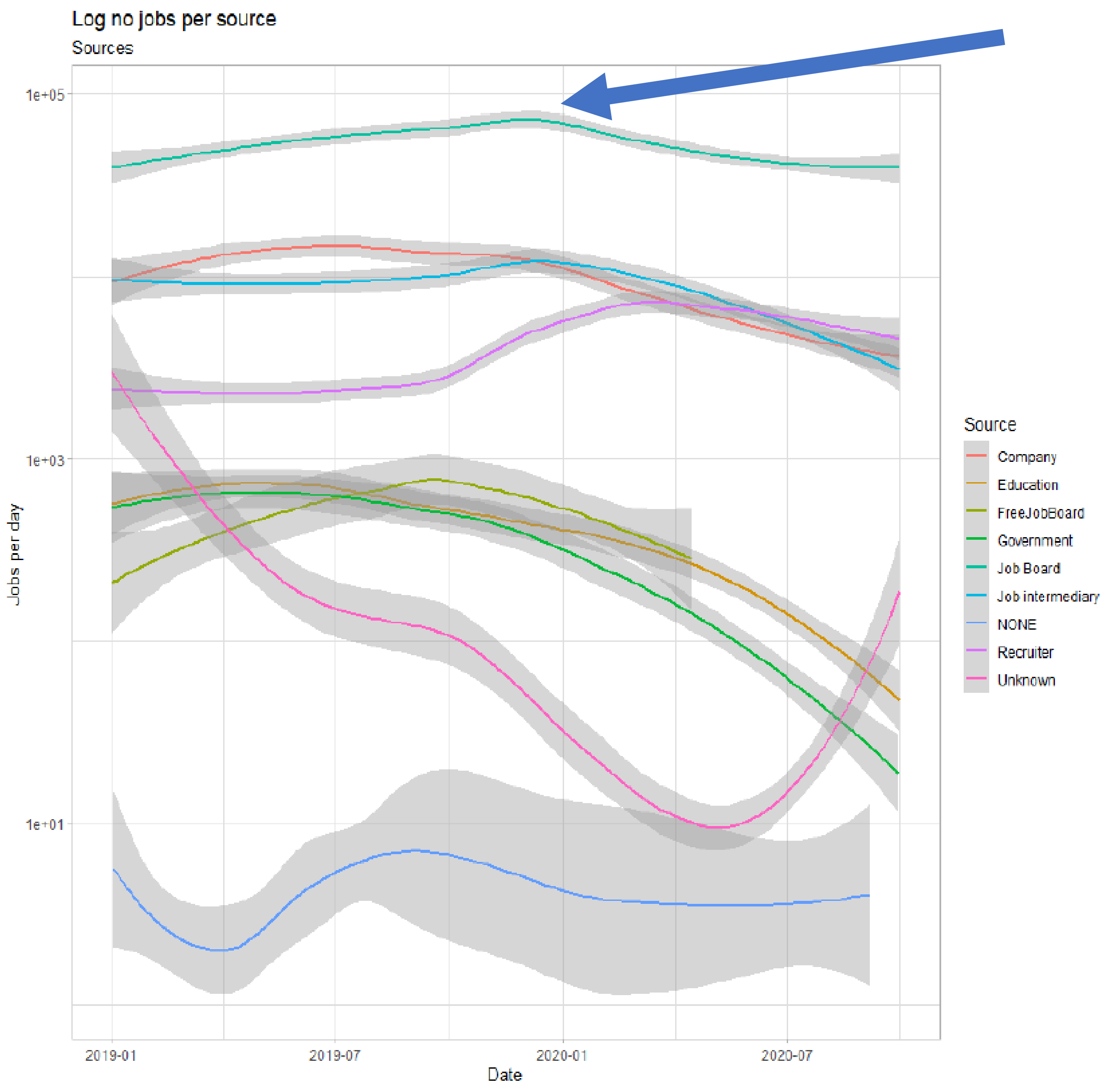
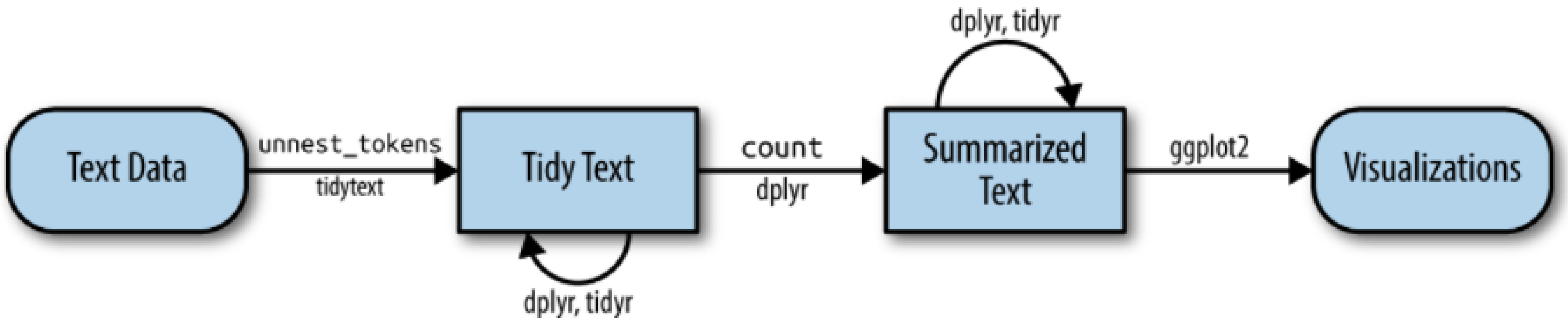


Figure 6: *Sources of Job advertisements*

Pipeline


Text Data
unnest_tokens
tidytext
Tidy Text
dplyr, tidyr
count
dplyr
dplyr, tidyr
Summarized Text
ggplot2
Visualizations

Reading the literature helps

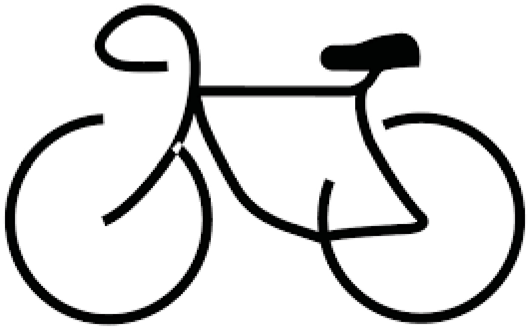
Studying the UK Job Market During the COVID-19 Crisis with Online Job Ads

- Rudy Arthur

In particular we will look at

- The number of job postings by date.
- Time series of vacancies by sector.
- Time series of vacancies by geographic region.
- The distribution of salary; type of contract (full time, part time, contract) and mode of work (permanent or temporary) before and after the COVID crisis hit in 2020.

Exercise



What is in the current literature?

Visit **Google scholar** and look for term "**Job Market Intelligence**." Review only for papers written in the last year. List the title, a brief summary of the abstracts and write a note if the paper is useful for finding Signals.

Note: If you have a particular research interest please consider refining the search term to match.

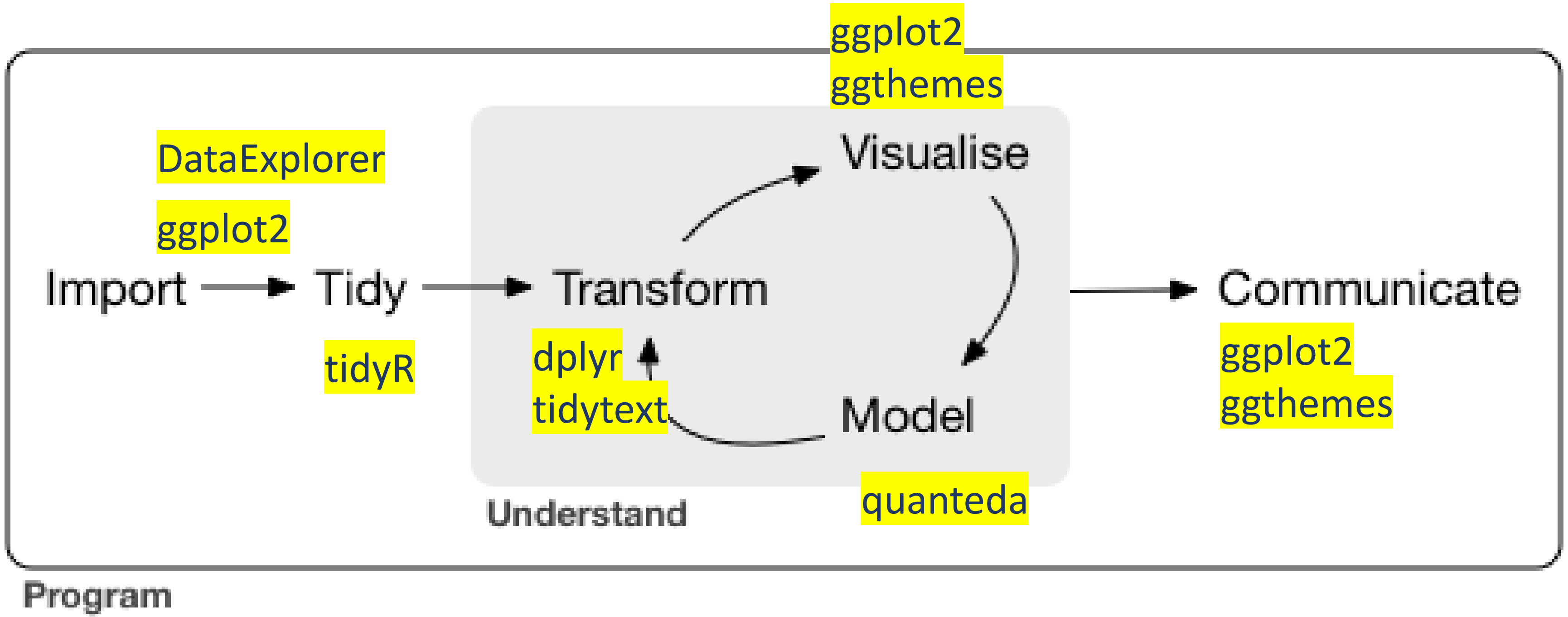
Review Code Book

```
sp(library(tidyverse))

# Load data
file.jobs <- "../../DATA/MonsterBoard-2013-n=20000.Rdata"
load(file.jobs)

# Custom dictionary, replace with your own
# In the topic notebook you have a recipe to divid into topics as well.
word <- c("young","age","race","disability","sexual","discriminate","ethnicity","families", "family","faith","abr
oad","barrier","creed","carer","home", "marital","she", "female","her","mother","minority","hate","care","support
ive","nurture","carer","local","helpful","social","parent","flexible","friendly")

custom.dic <-data.frame(word=word)
#custom.dic

# Place the job adverts into a tibble for easy manipulation
my.job <- tibble(Row = seq_along(sample$JobBody) , text = sample$JobBody)

# Load in stop words
data("stop_words")
stop_words <- rbind(stop_words,c("nbsp","Custom"))
stop_words <- rbind(stop_words,c(" ","Custom"))

# clean words
my.words <- my.job %>%  unnest_tokens(word, text) %>%  anti_join(stop_words)
my.words$EDU <- sample$EducationLevelID[my.words$Row]
my.words$EDU[my.words$EDU=="NULL"]<- 10
table(my.words$EDU)
```

Mentality

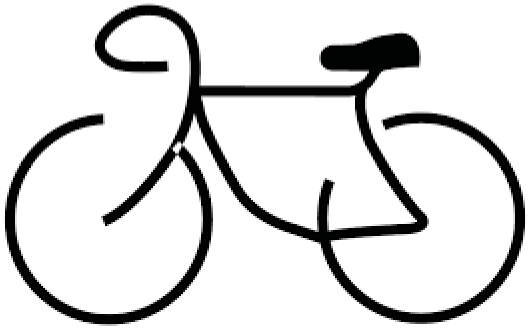
- **Be kind to yourself.** You do not need to understand every detail.
- Be patient.
- Learning does cost energy
- Read small chunks of code at a time
- Keep practicing reading code. Make notes on each function
- Small recipes can do a lot, so try and find those recipes
- The Internet is your friend
- If you want to learn R for the first time then try the following:
  - https://rafalab.github.io/dsbook/r-basics.html
  - https://bookdown.org/dli/rguide/

R for DataScience



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

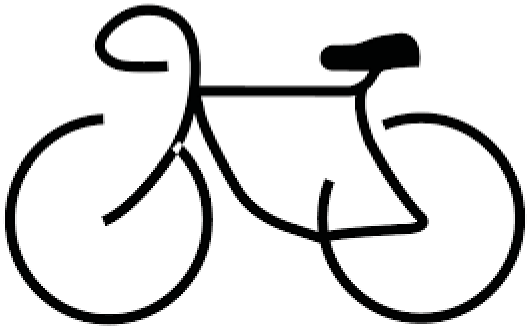
Exercise



Review the output from the notebook .nb.html file and write in your notes the following.

1. List the packages used and describe their purpose
2. List the new functions used and what they do
3. Describe any short recipes in your own words
4. Search the Internet for at least two links for similar examples

Exercise



Search the Internet for at least 3 text mining tools

- What are the techniques used by the tools?
- Which tool(s) are your favourite and why?
- What are the limitations of an application compared to a programming language?
- What are the limitations of a programming language compared to an application?
- What are the shared strengths?

Reading List

Review the links in the reading list for this module.
Write brief notes, answering the questions:

1. Do the links still work
2. Which links are relevant for you
3. Which links are not relevant for you
4. What further information would you of wished