Week 01
R You Ready?

POLI3148. Data Science in Politics and Public Administration

Dr. Haohan Chen HKU-PPA

Today

- About Me
- About the course (syllabus)
- *R* Software Setup

About Me

About Me

Dr. CHEN Haohan 陳昊瀚

Interested in …

- Political communication
- Computational methods

My Research

Chinese context

- Understanding the ideological spectrum of Chinese intellectuals using cultural product reviews
- Using audio of news program to infer power dynamic and policy agenda

US context

- Perception on Covid-19 among the US public using Twitter data
- Monitoring political polarization in the US using Twitter data

How We Work Together Going Forward

- I would love to get to know each one of you in person.
- Language: English, Mandarin, Cantonese
- How we communicate
  - **Office hours**: Calendly appointment system
  - **CampusWire** Course Forum
- Appointments outside office hours possible. Email me.
- If I do not reply to your email within two days, kindly send me a nudge.

About **DaSPPA**

To discuss

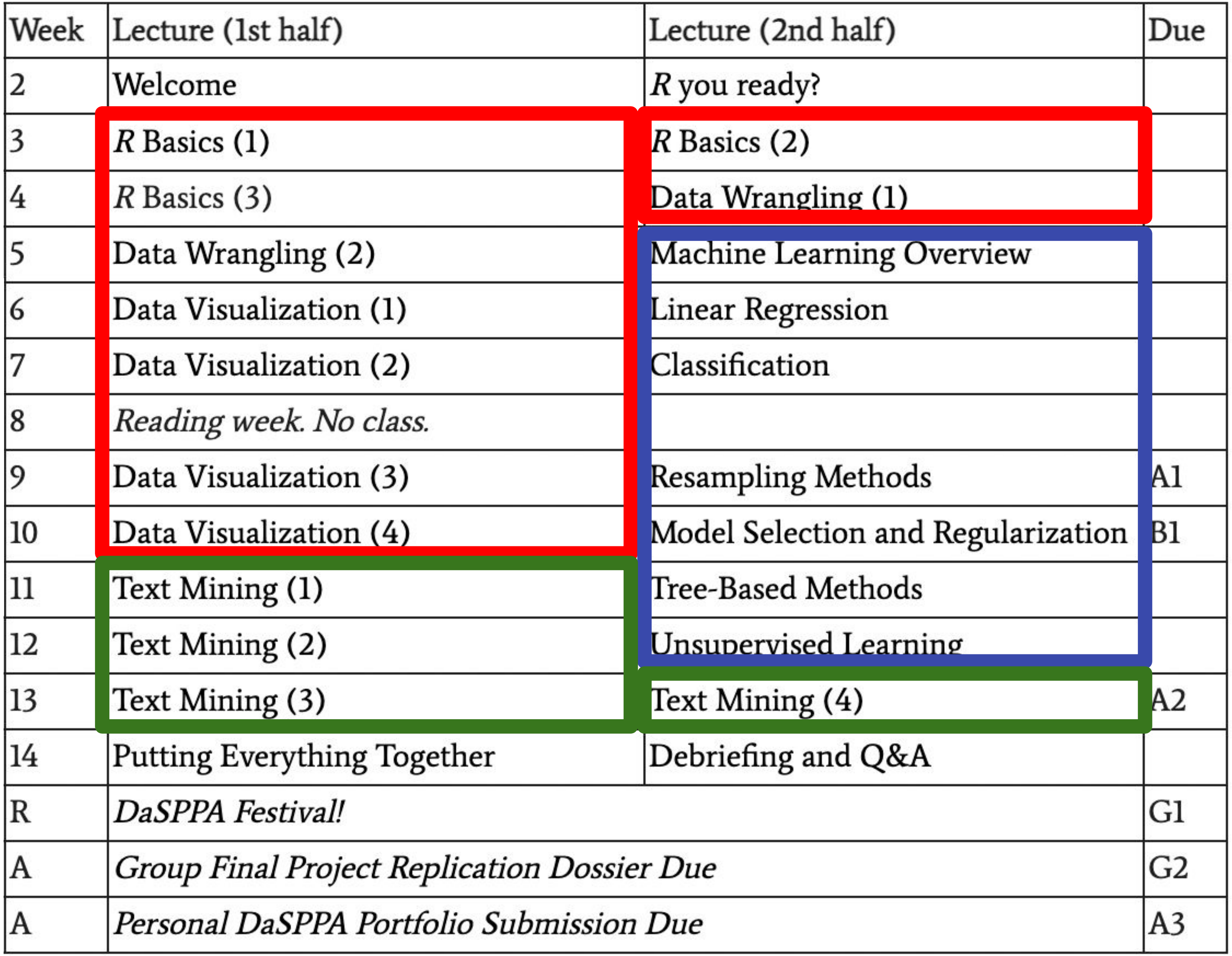
- Topics
- Readings
- Output and Assessment

Topics

R **+ Data**

Model

Text Mining



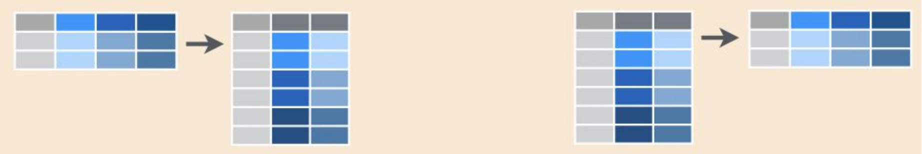
Week	Lecture (1st half)	Lecture (2nd half)	Due
2	Welcome	*R* you ready?	
3	*R* Basics (1)	*R* Basics (2)	
4	*R* Basics (3)	Data Wrangling (1)	
5	Data Wrangling (2)	Machine Learning Overview	
6	Data Visualization (1)	Linear Regression	
7	Data Visualization (2)	Classification	
8	*Reading week. No class.*		
9	Data Visualization (3)	Resampling Methods	A1
10	Data Visualization (4)	Model Selection and Regularization	B1
11	Text Mining (1)	Tree-Based Methods	
12	Text Mining (2)	Unsupervised Learning	
13	Text Mining (3)	Text Mining (4)	A2
14	Putting Everything Together	Debriefing and Q&A	
R	*DaSPPA Festival!*		G1
A	*Group Final Project Replication Dossier Due*		G2
A	*Personal DaSPPA Portfolio Submission Due*		A3

R + Data: R Basics

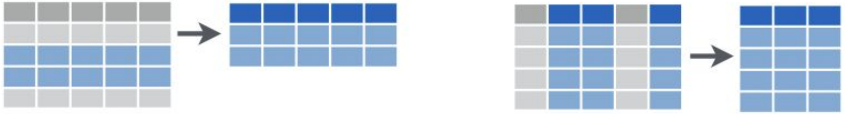
- Basic knowledge of R and Rstudio
- Assume no prior experience with R
- But if you have prior experience, it will be a good review

R + Data: Data Wrangling

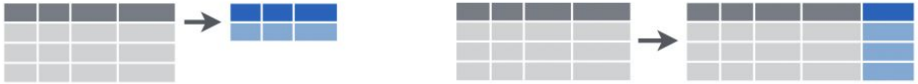
Reshape



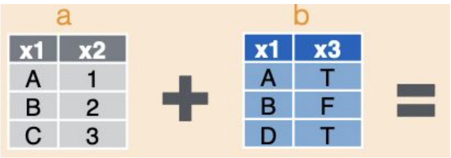
Subset rows and columns



Summarize and mutate variables



Combine datasets



a

x1	x2
A	1
B	2
C	3

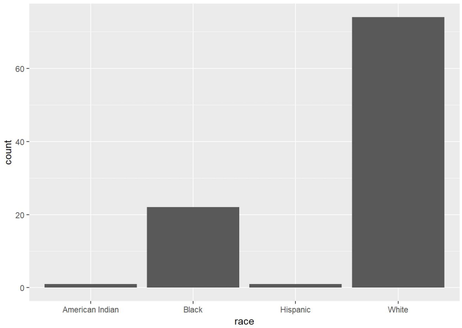
+

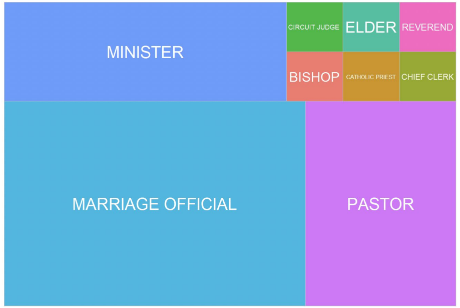
b

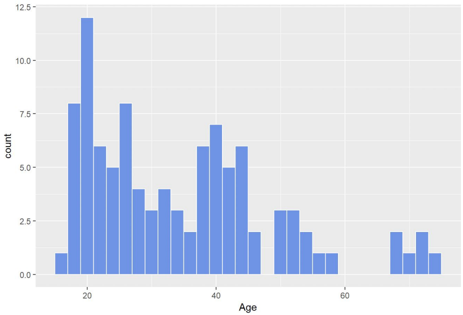
x1	x3
A	T
B	F
D	T

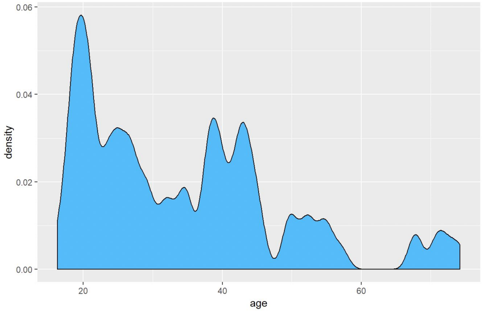
=

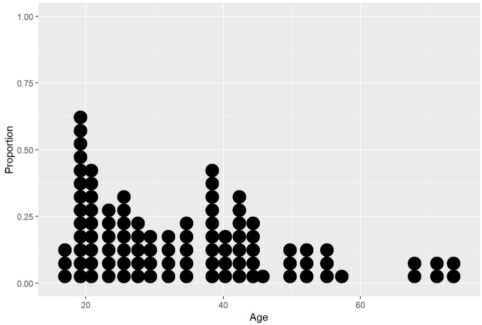
R + Data: Visualize one variable



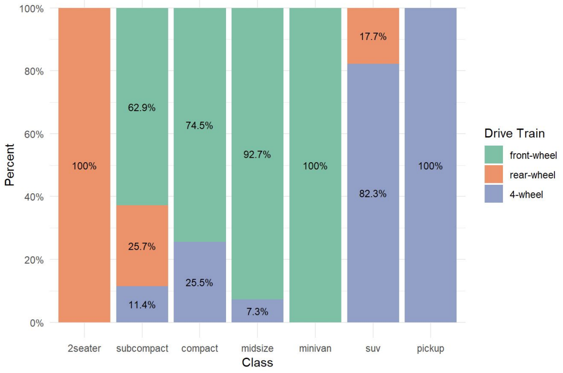


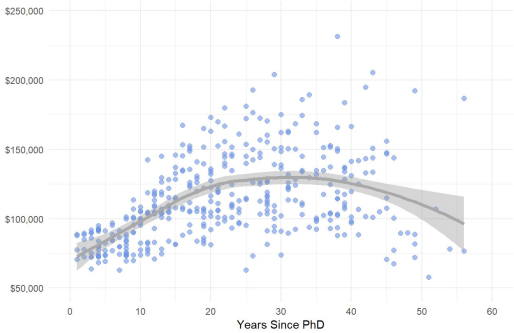


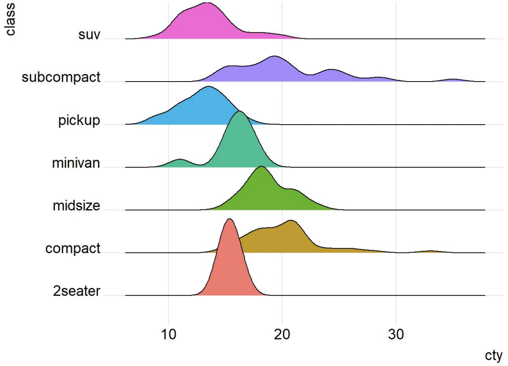


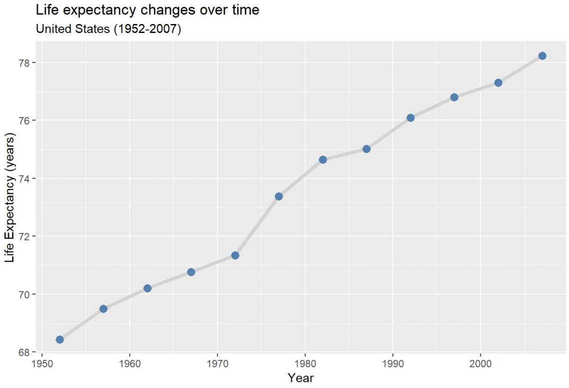


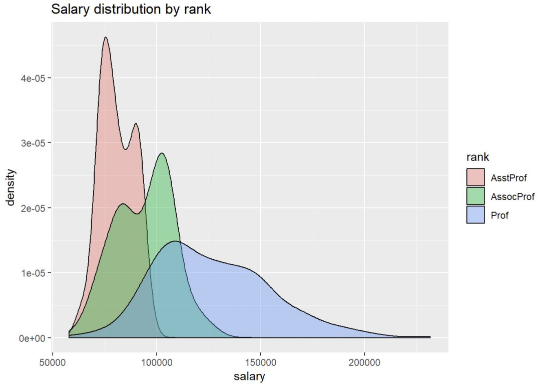
R + Data: Visualize two variables

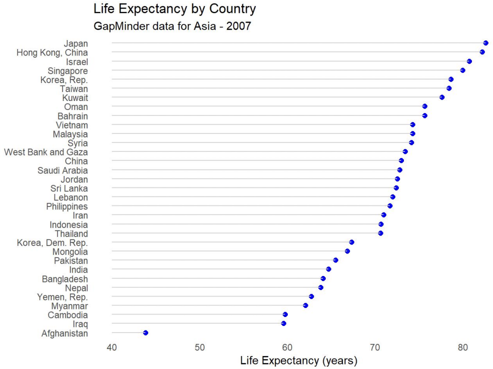




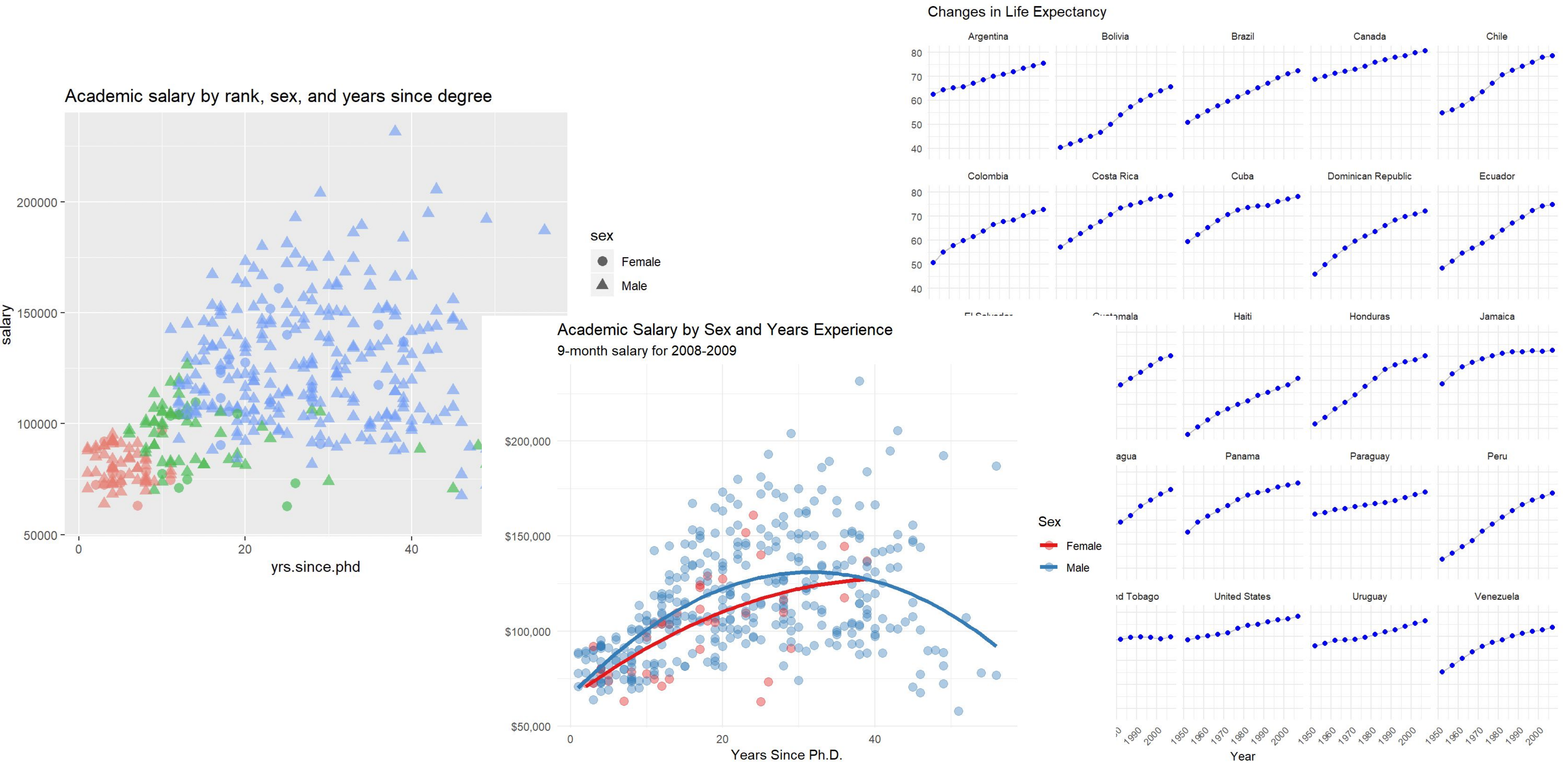








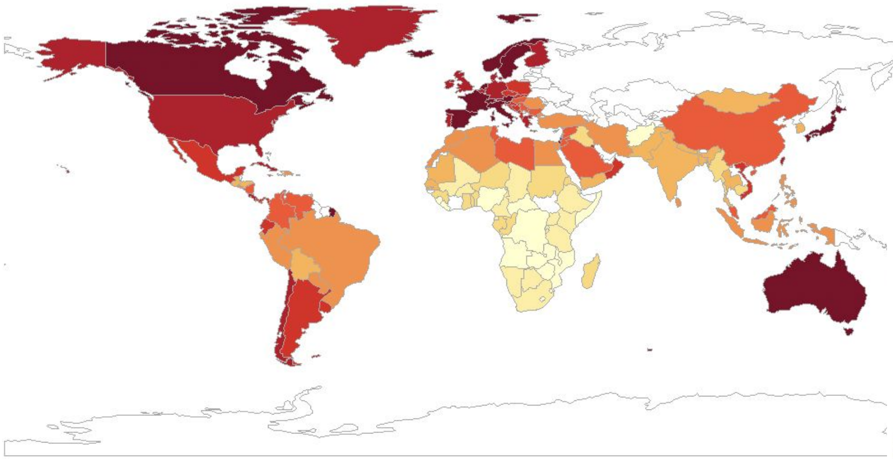
R + Data: Visualize multiple variables



R + Data: Visualize data wrt space and time

Life expectancy by country

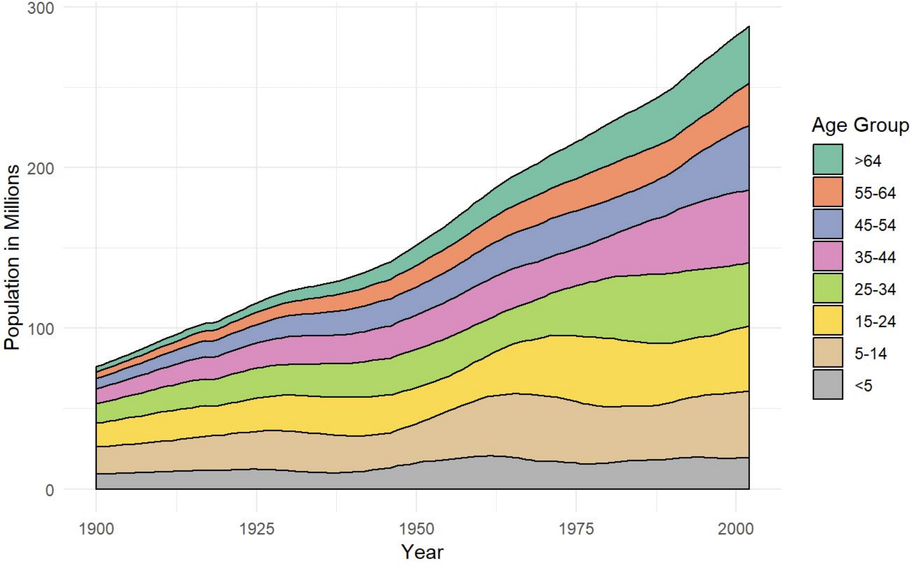
Gapminder 2007 data



source: https://www.gapı

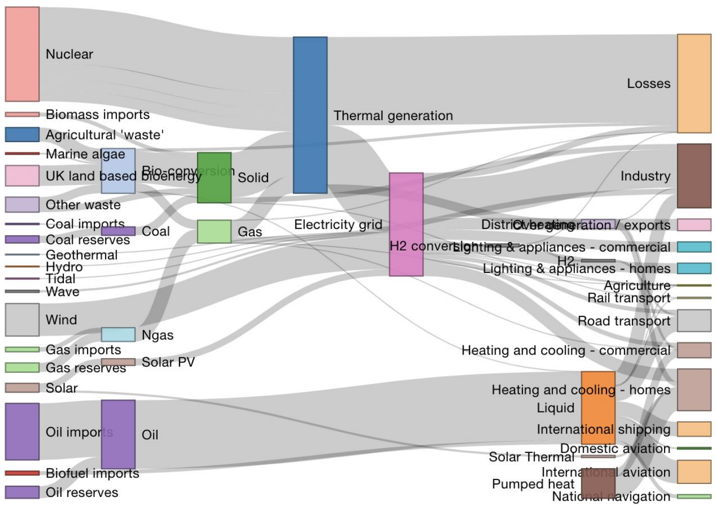
US Population by age

1900 to 2002

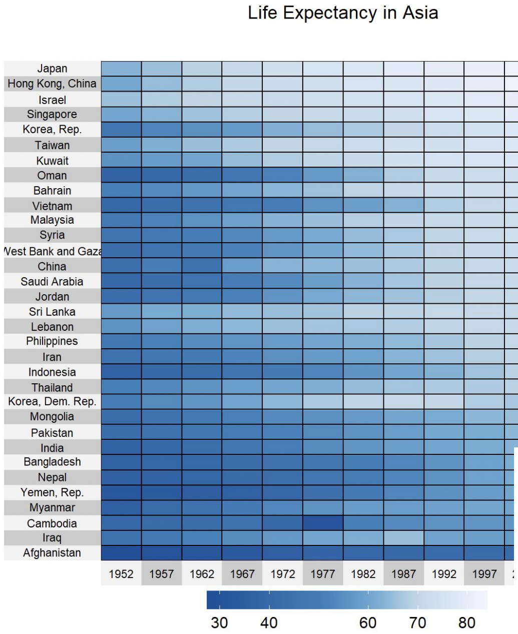


source: U.S. Census Bureau, 2003, HS-3

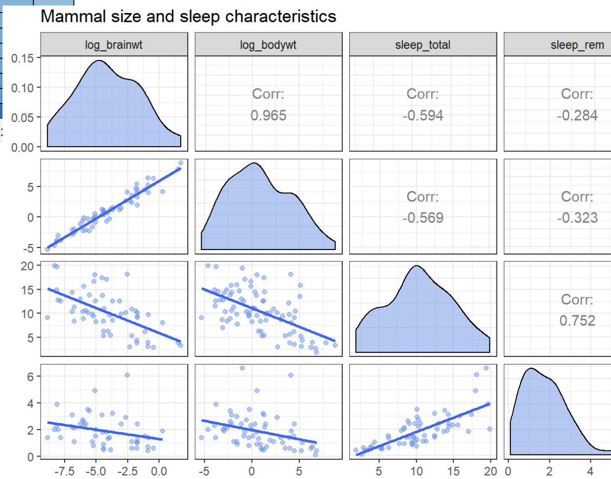
R + Data: Other cool visualization



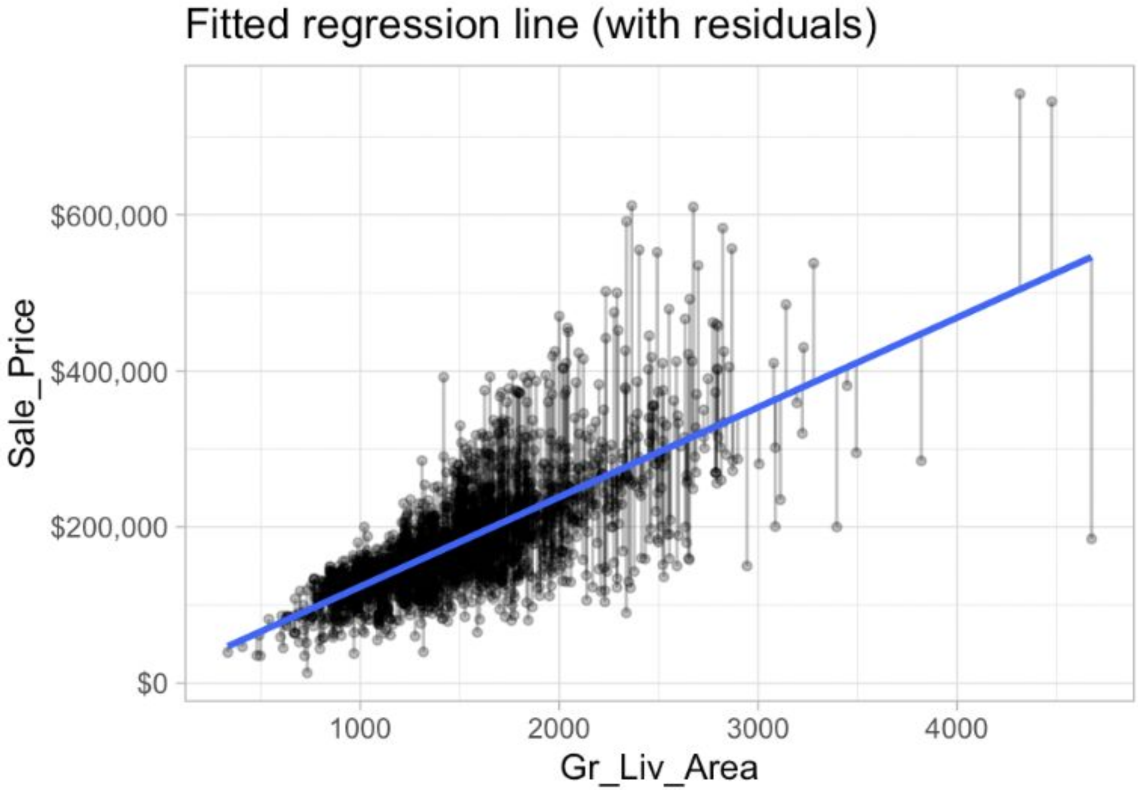
Life Expectancy in Asia



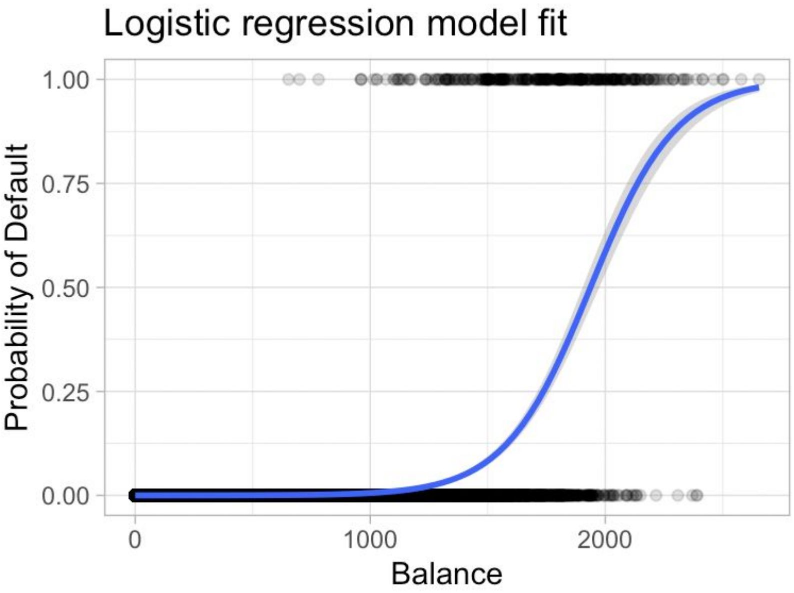
Mammal size and sleep characteristics

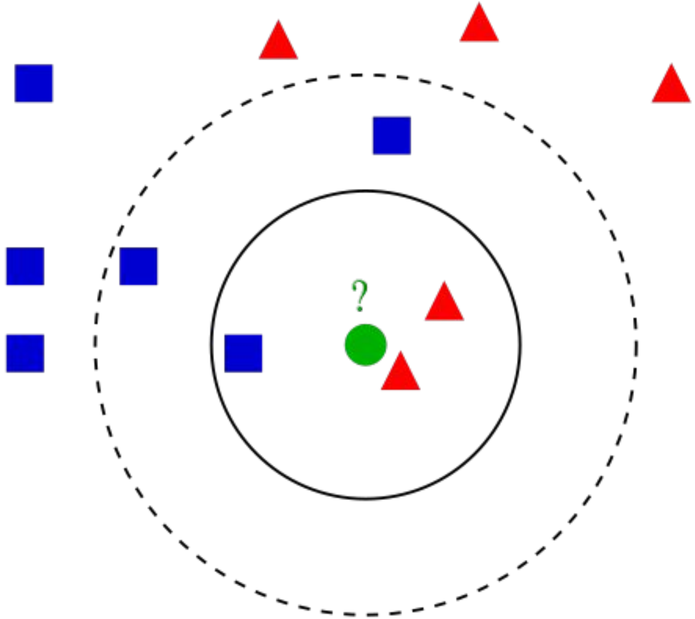


Model: Linear Regression



Model: Classification

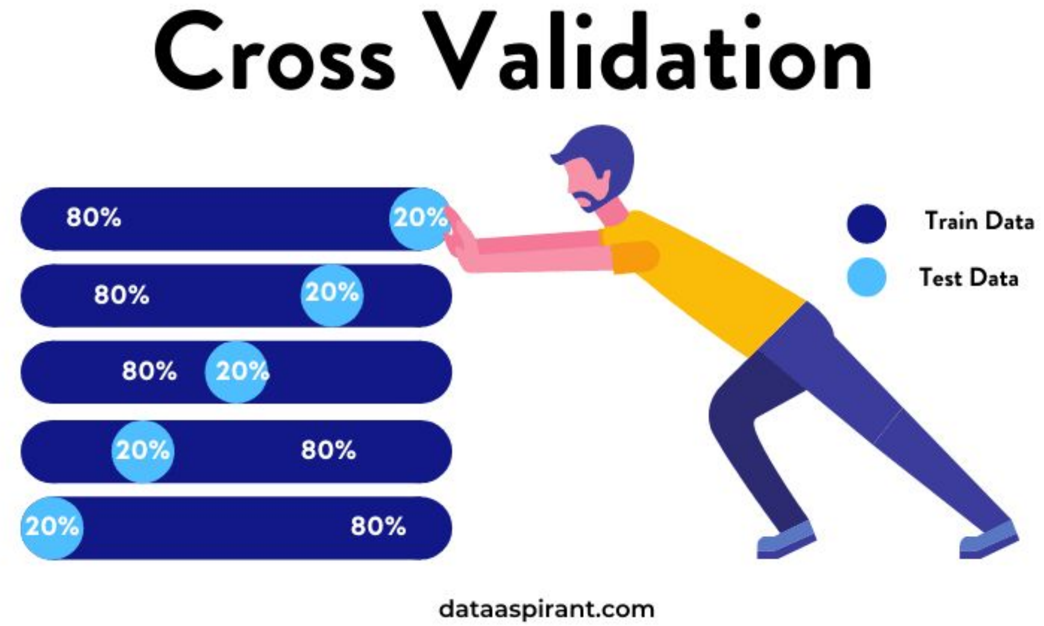




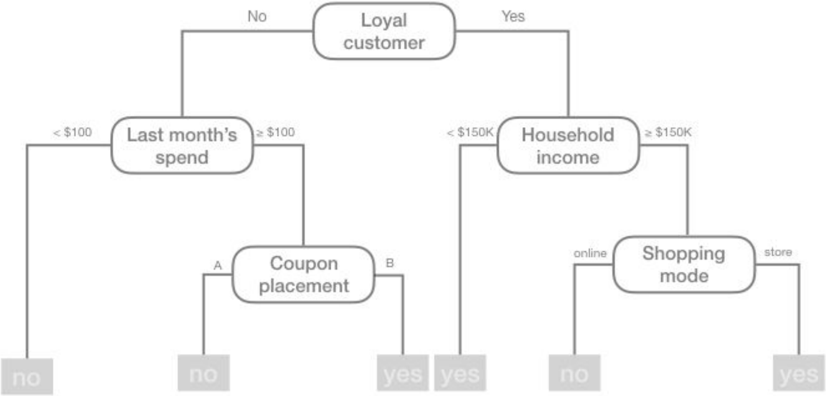
Model: Model Selection and Regularization

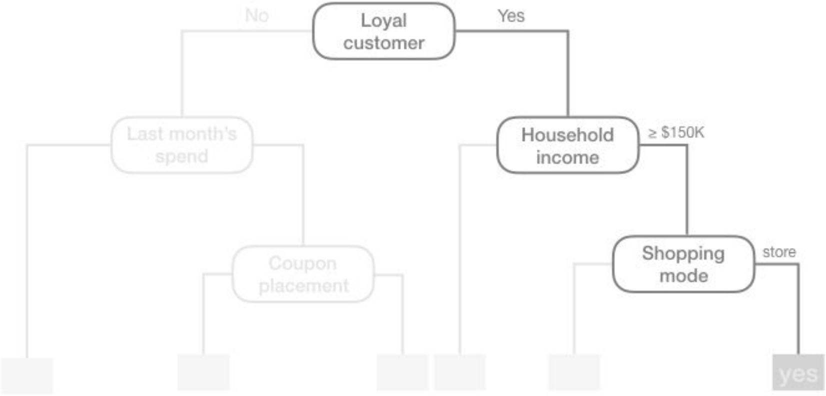
- Criteria to evaluate how “good” a machine learning model is
- Forms of Linear Regression and Logistic Regression when you have too many predictors/ independent variables

Model: Resampling Methods

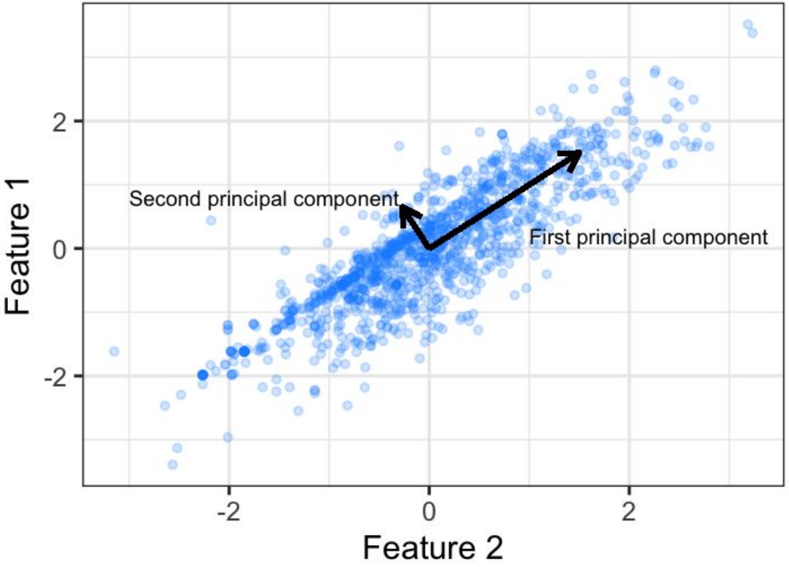


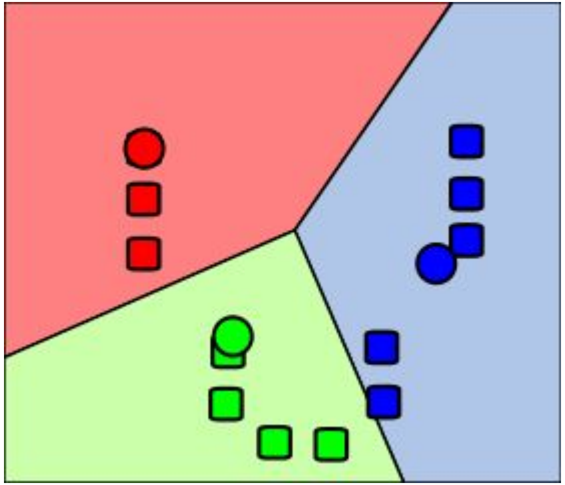
Model: Tree-Based Methods



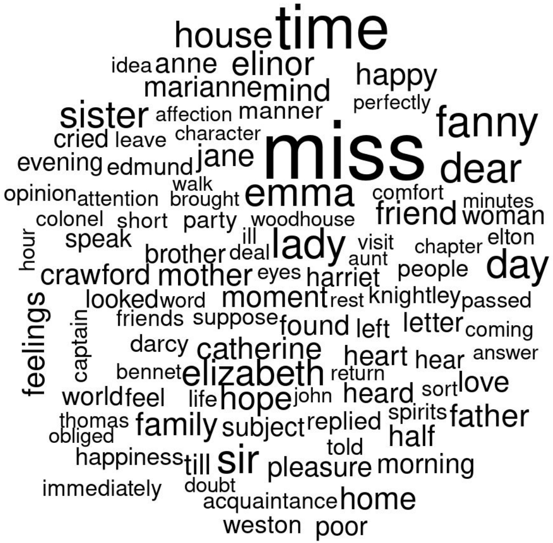


Model: Unsupervised Learning





Text Mining: Basics



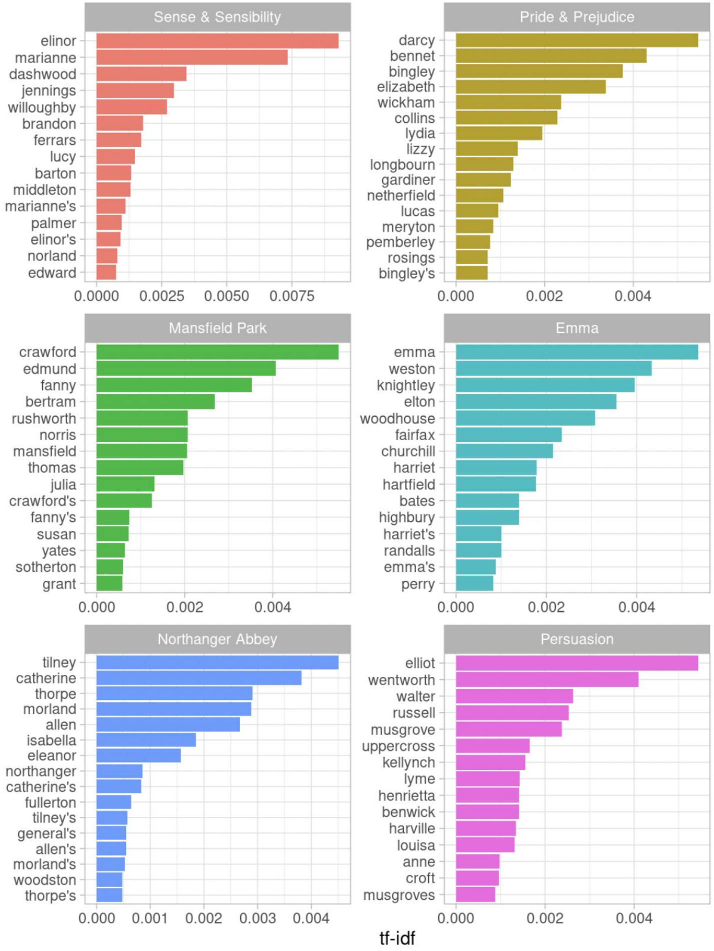


Figure 3.4: Highest tf-idf words in each Jane Austen novel

Text Mining:
Sentiment Analysis

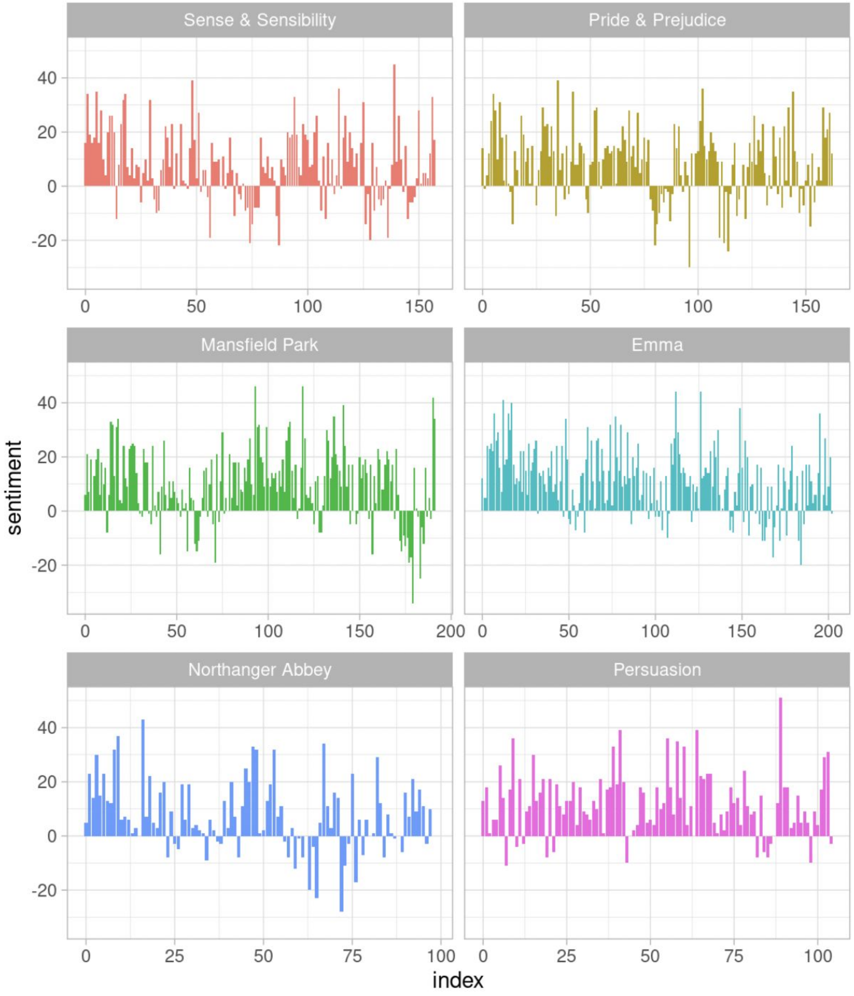
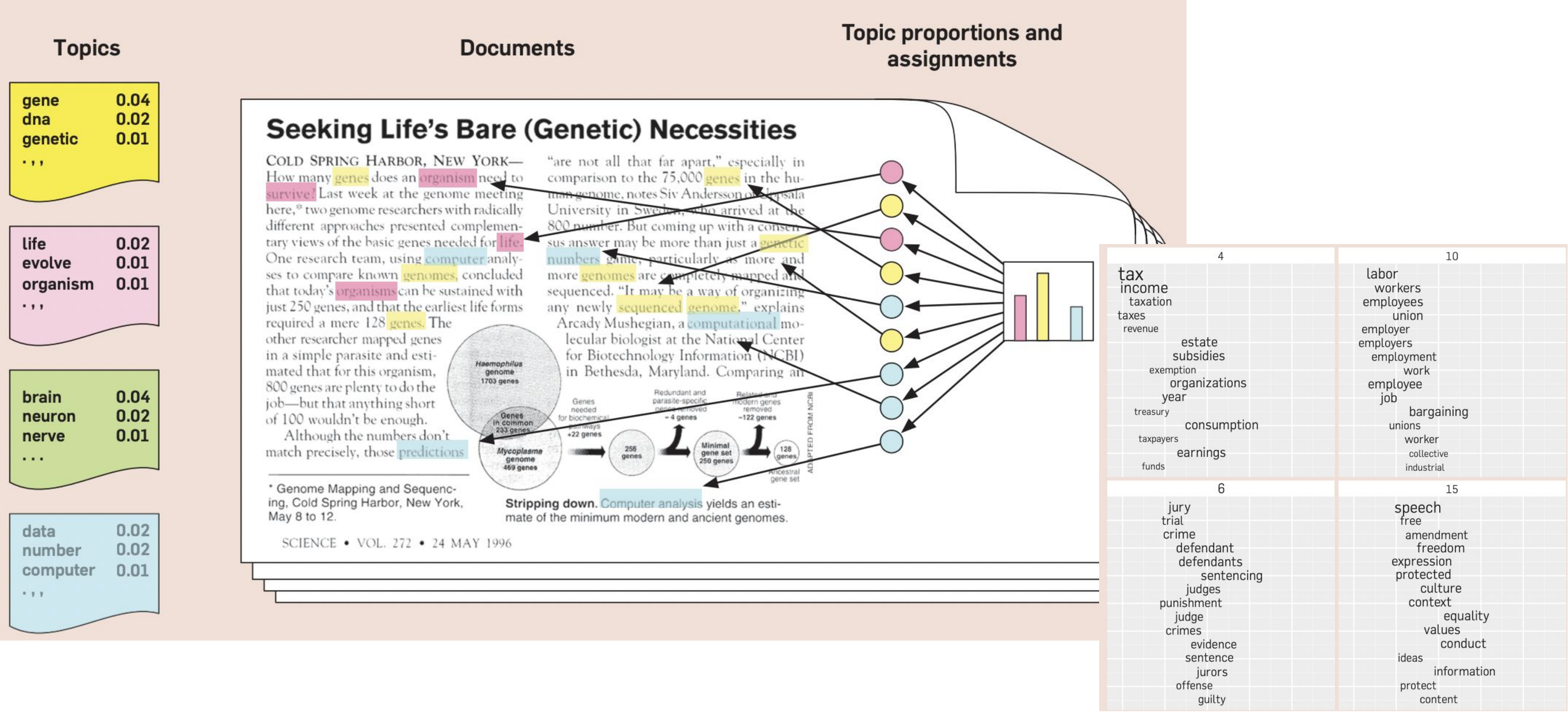


Figure 2.2: Sentiment through the narratives of Jane Austen’s novels

Text Mining:
Text Summarization and Information Extraction



Topics

gene	0.04
dna	0.02
genetic	0.01
. , ,	

life	0.02
evolve	0.01
organism	0.01
. , ,	

brain	0.04
neuron	0.02
nerve	0.01
. . .	

data	0.02
number	0.02
computer	0.01
. , ,	

Documents

Topic proportions and assignments

Seeking Life's Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK— How many genes does an organism need to survive? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today's organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

Although the numbers don't match precisely, those predictions "are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a genetic numbers game, particularly as more and more genomes are completely mapped and sequenced. "It may be a way of organizing any newly sequenced genome," explains Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an

Haemophilus genome 1703 genes
Genes in common 233 genes
Mycoplasma genome 469 genes
Genes needed for biochemical pathways +22 genes
256 genes
Redundant and parasite-specific genes removed −4 genes
Minimal gene set 250 genes
Related and modern genes removed −122 genes
128 genes
Ancestral gene set
ADAPTED FROM NCBI

Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

SCIENCE • VOL. 272 • 24 MAY 1996

4

tax
income
taxation
taxes
revenue
estate
subsidies
exemption
organizations
year
treasury
consumption
taxpayers
earnings
funds

10

labor
workers
employees
union
employer
employers
employment
work
employee
job
bargaining
unions
worker
collective
industrial

6

jury
trial
crime
defendant
defendants
sentencing
judges
punishment
judge
crimes
evidence
sentence
jurors
offense
guilty

15

speech
free
amendment
freedom
expression
protected
culture
context
equality
values
conduct
ideas
information
protect
content

Readings

- Most are hands-on materials
  - Read
  - Try the code yourself
  - Tweak the code and see what happens
- Expect familiarity with the reading materials before class
- Clarify and extend in class
- Strongly encourage review and taking notes after class

Output and Assessment

See the syllabus.

R Setup

In-class Exercise 1: Setup R

- Install R
- Install RStudio
- Open RStudio
- Run the following code in R Console

```
install.packages(“tidyverse”)
```

- Post a screenshot of your Rstudio interface on CampusWire