Getting started with secondary data analysis:

Dr Ana Morales-Gomez

Introduction to analysing data about crime using R
Manchester
4-5 February 2020
Data and Crime

• Jack Maple and the “Charts of the Future”

• Steve Talley: How facial recognition can ruin your life

• Paul Zilly: Human versus Machine

Steve Talley: https://theintercept.com/2016/10/13/how-a-facial-recognition-mismatch-can-ruin-your-life/
Paul Zilly: https://www.sciencefocus.com/future-technology/can-an-algorithm-deliver-justice/
More about Data and Crime

Chapter Justice
Chapter Crime

Twitter: @FryRsquared
What is Research?

Adding a contribution to an existing body of knowledge
THE DATA SCIENCE PROCESS

Ask an interesting question
What is the specific GOAL?
What would you do if you had all the DATA?
What do you want to PREDICT or ESTIMATE?

GET the data
How were the data SAMPLED?
Which data are RELEVANT?
Are there PRIVACY issues?

EXPLORE the data
Plot the DATA
Are there ANOMALIES?
Are there PATTERNS?

MODEL the data
BUILD a model
FIT the model
VALIDATE the model

Communicate and Visualize the result
What did we LEARN?
Do the results make SENSE?
Can we tell a STORY?

© Joe Blitzstein And Hanspeter Pfister, Harvard Data Science Course available here: http://cs109.github.io/2015/
The data Science process (1): Ask an interesting question

✓ A topic of interest:
  ➢ Crime
  ➢ Health inequalities
  ➢ Pollution

✓ Specific goal
  1. Confidence in the Criminal Justice System in England
  2. Antisocial behaviour in Manchester

✓ What do you want to predict or estimate?
  ➢ National level estimates
  ➢ Local level indicators
  ➢ CJS as a whole or concentrate on Police, prisons, Sentencing?
The data Science process (2): Get the data

✓ Police recorded crime data
✓ CSEW: For England and Wales
✓ Scottish Crime Survey
✓ European Social Survey
✓ Others:
  ➢ Administrative data of prisons
  ➢ Administrative data sentencing council
The data Science process (2): Get the data

✓ Police recorded crime data:
  ✓ CSEW UK Data Service
  ➢ Coverage
    • Date range
    • Spatial units

➢ What data
  • Available for surveys
  • Open data may not have any

➢ Format
  • Depending on the source. UKDS: Stata, SPSS,
  • Excel
  • Text
The data Science process (3a): Explore the data

✓ What data do we have?
  ➢ Variables
    (name some variables)
  ➢ Type of data
    • Numeric?
    • Attribute (character)
  ➢ Is it ready to analyse?
    • Data cleaning
    • Manipulation
The data Science process (3b): Explore the data

✓ What data do we have?
   ➢ Variables
     (name some variables)

➢ Type of data
   • Numeric?
   • Attribute (character)

➢ Is it ready to analyse?
   • Data cleaning
   • Manipulation

✓ Descriptive statistics
   ➢ Central tendency measures
     • Any correlations?
     • Anomalies?

➢ Plot the data
   • Anomalies?
   • Patterns?

➢ More questions
   • Are the data enough for my RQ?
   • Do we need more data?
   • Is there more data?
   • Change RQs?
What is the best approach to understand the data we have?

**Depends on…**
- Our research questions
- Our data available

**Example:**
- Correlation to look for association of two variables
- Generalised Linear models /Regression based models for
  - Multiple linear regression (continuous outcome)
  - Logistic/Probit regression (binary outcome)
  - Ordinal regressions
  - Multilevel models (clusters and hierarchy dependence)
  - Longitudinal models (samples at different time points)

The Data Science process (4): **Model the data**
The data Science process (5): Communicate and visualise the results

✓ Visualise the results
- Tables
- Figures
- Plots
- Maps

✓ Communicate the results

➢ Know your audience
  - Effective
  - The right details for each audience
  - Academic ≠ Local Government officers

Source: Financial Times: https://ig.ft.com/sites/elections/2016/uk/eu-referendum/
Exploratory Data Analysis
Exploratory data analysis

Flowchart for data preparation

From R for data science

https://r4ds.had.co.nz/explore-intro.html
What is data?

- **Information**, especially **facts or numbers**, collected to be examined and considered and used to help decision-making, or information in an electronic form that can be stored and used by a computer (Cambridge dictionary)
  - Numeric
  - Images
  - Attributes (characters)
Describe the data

To Understand:

- Data availability,
- Types,
- Quality,
- Data complexity (i.e. nonlinearity, requires transformation, etc)

Guided by two types of questions (Grolemund and Wickham, 2016):

- What type of covariation occurs between my variables?
- What type of variation occurs within my variables?
How to describe the data (1)

✓ Distribution of numerical variables:
  ➢ Extreme values (outliers)
  ➢ Shape of the distribution
  ➢ Missing cases
  ➢ Unusual patterns

✓ Distribution of categorical variables
  ➢ Missing cases
  ➢ Odd values
  ➢ Unusual patterns
  ➢ Most common values

Figure: © Allison Horst
How to describe the data (2)

✓ **Central tendency measures**
  - Mean
  - Median
  - Mode

✓ **Measures of spread**
  - Variance and standard deviation
  - Range:
    - Interquartile range (IQR)

✓ **Visualisations**
  - Histograms, boxplots, bar plots, scatterplots
How to describe the data (3)

- **Mean (sample vs. population):**
  \[ \mu = \frac{\sum x}{N} \]
  - The "average" number; found by adding all data points and dividing by the number of data points

- **Median**
  - Middle value if odd number of values, or average of the middle two values otherwise

- **Mode**
  - Value that occurs most frequently in the data
    - Unimodal, bimodal, trimodal

Is the mean always the best central tendency measure?
The problem with the mean

“There are two pieces of bread. You eat two. I eat none. Average consumption: one bread per person.“

Nicanor Parra, (Anti)Poet, Mathematician and Physicist
More about visualisations

Two types:

1. Exploring and getting to know the data
   1. Assess the data: decide what to do next
   2. Accurate
   3. Internal, never reach the wider audience

2. Communication
   1. Present data and ideas
   2. Accurate: provide evidence
   3. Easy to understand
   4. Effective
   5. It would depend on the audience

Images: https://r4ds.had.co.nz/exploratory-data-analysis.html
Effective visualisations for communication

✓ Simple but effective (don’t over do it!)
  ✓ Easy to understand
✓ Use the right type of graph of figure
  ✓ Not a fit them all purpose graph
✓ Appropriate use of colours (colour blind people)
✓ Know your audience
Effective visualisations for communication

Chart Suggestions—A Thought-Starter

Comparison

What would you like to show?

Relationship

Distribution

Composition

https://extremepresentation.typepad.com/blog/files/choosing_a_good_chart.pdf
Effective visualisations for communication: Use the right display

✓ Comparisons:
  ➢ Bars
  ➢ lines
✓ Proportions
  ➢ Pie charts
  ➢ Stacked charts
✓ Trends over time
  ➢ Lines
  ➢ Scatterplots
✓ Distributions
  ➢ Density plots
  ➢ Histograms
✓ Correlations
  ➢ Scatterplots
Your turn
Questions

Ana Morales-Gomez

ana.morales@Manchester.ac.uk