Module 7: Post-data collection
Video 2: Summarizing, Visualizing, and Editing

Sharan Sharma

MANNHEIM
BUSINESS SCHOOL

Many activities…

1. Coding open-ended responses
2. Data preparation
3. Summarizing and Visualizing
4. Data editing
5. Imputation and Weighting
6. Disclosure control
7. Final processing, Documentation, and Dissemination

1. Missing data

- While monitored during data collection (→ Module 6, Video 4), should also be analyzed at this stage.

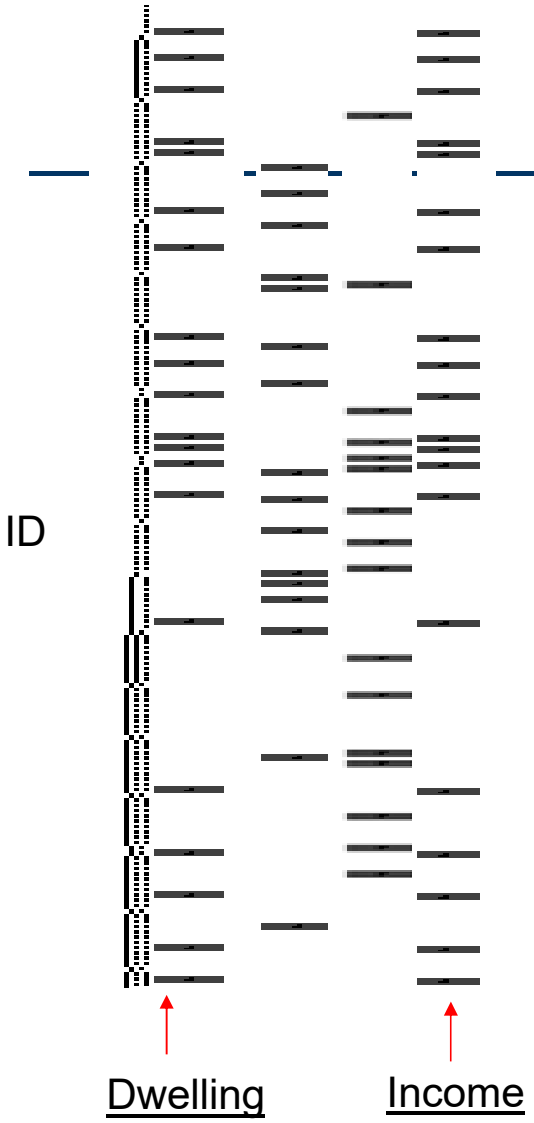
- Missing due to skip patterns (NA)? Or due to DK/RF? Or any other issue?
  - Useful to analyze these separately

- First compute % missing data for each variable.
  - Focus on variables that have a missingness rate > ___ %.
  - Compare with previous rounds in the case of panel or repeated cross-section surveys.

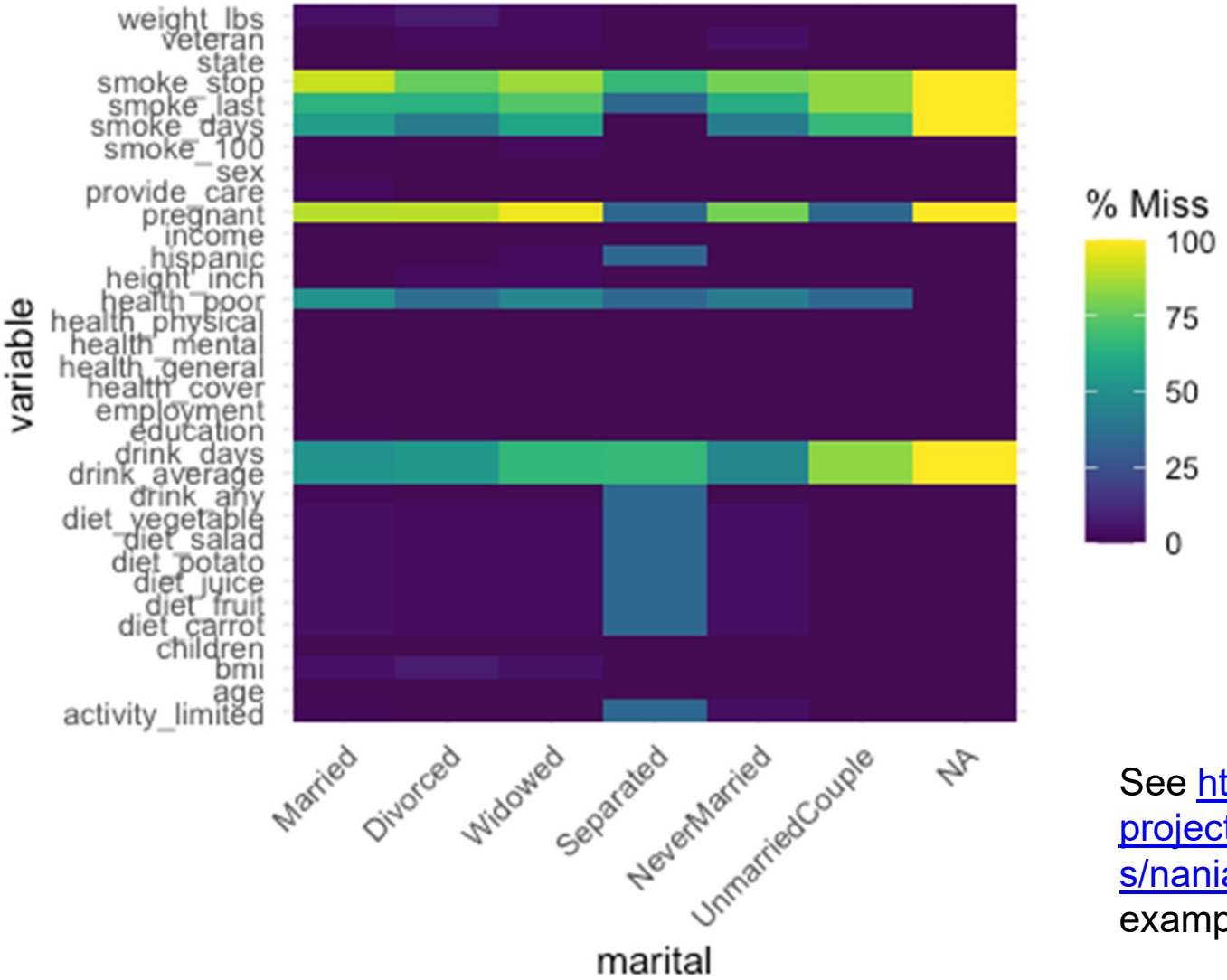
Missing data…

- Also look at *patterns* of missingness. Do we see any apparent problems?
  - e.g., All missing values for a variable occur in PSU 10.
  - Are there missing values when there should be none? e.g., Question should be asked to the respondent (not skipped) but showing as NA.

- Visualize missingness



- Correlation between cases which record missingness on dwelling and income variables.
  - Some IDs tend to have more missingness – what are the characteristics of these IDs?
- This visualization can be hard with 100s of variables ; best to analyze by some sensible variable groups.
- Many possible visualizations:



See https://cran.r-project.org/web/packages/naniar/vignettes/naniar-visualisation.html for other useful examples.

2. Get a sense of the aggregate data.

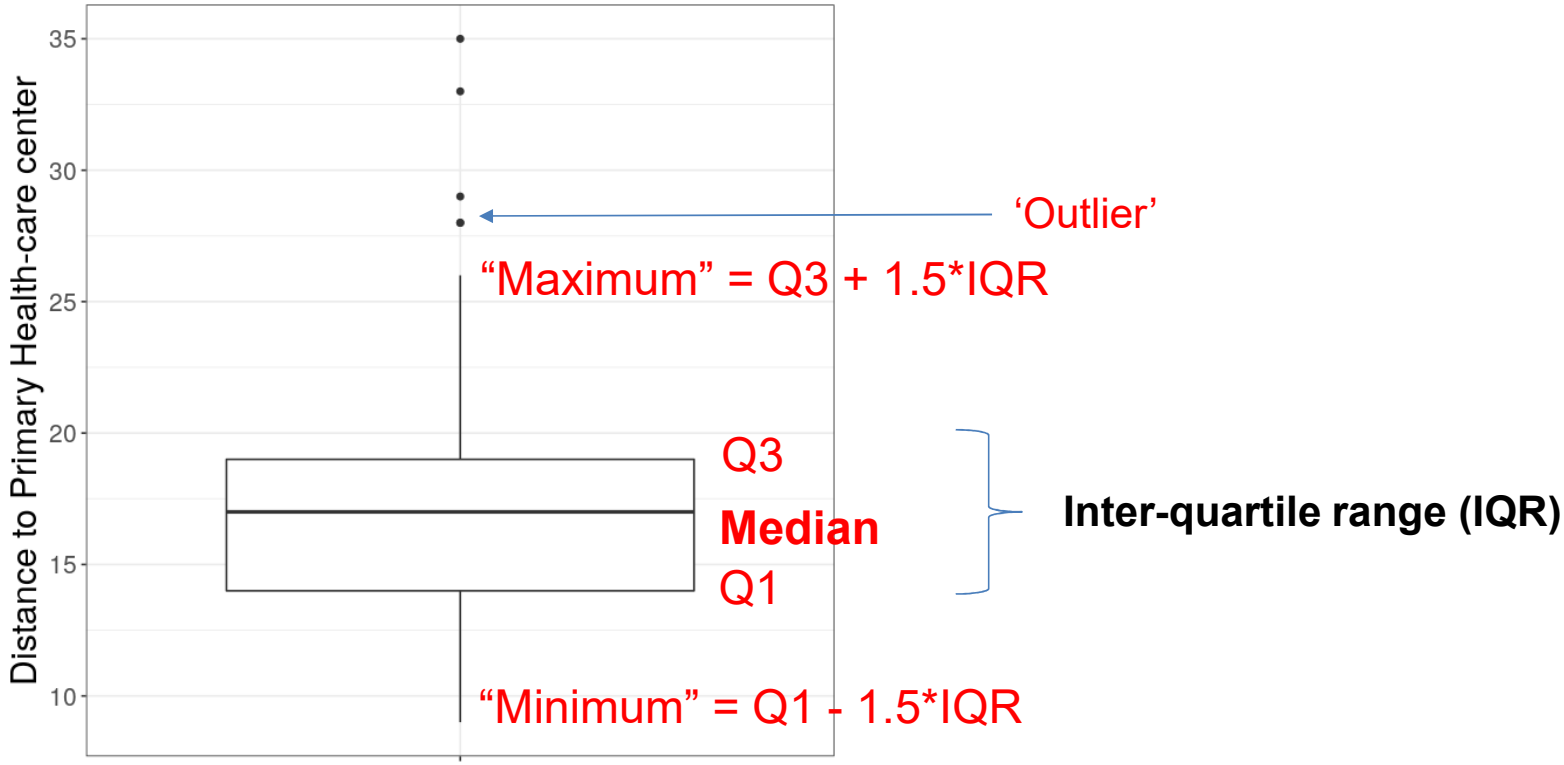
Summary tables: Five-number summary

	Distance to Primary health care center
Minimum	9
1st Quartile (Q1)	14
Median	17
3rd Quartile (Q3)	19
Max	35
Mean	16.9

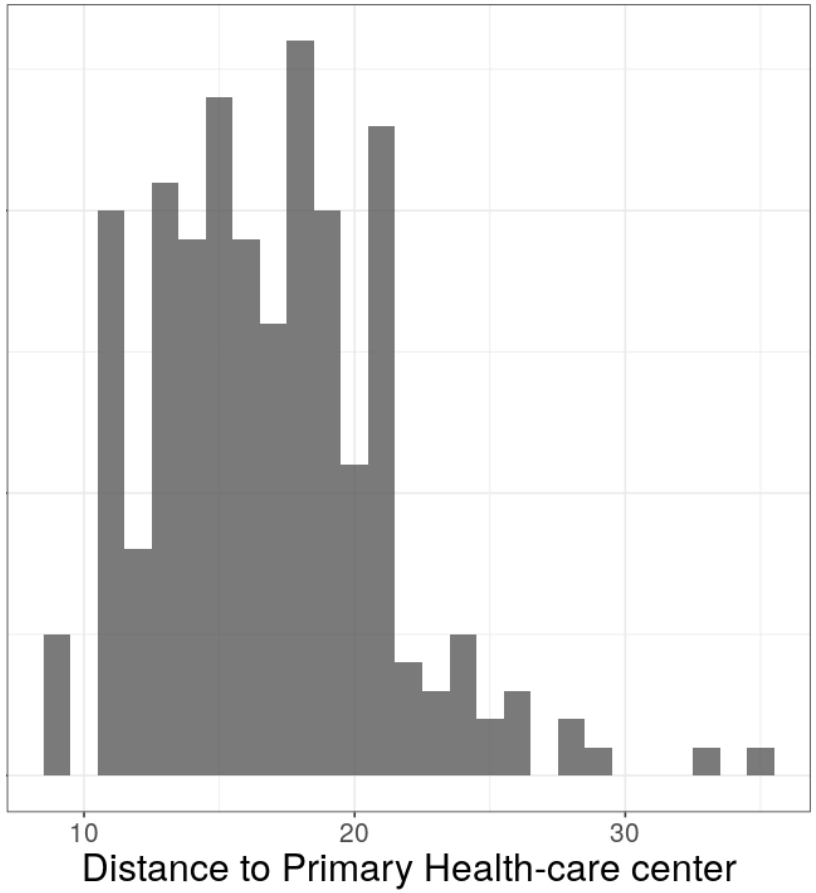
- Do these data make sense?
- How do they compare with previous survey rounds ?
- How do they compare with other benchmark data?

Data adapted from the 'mpg' dataset available in the ggplot2 package in R. Variables relabeled for the purposes of this presentation ('cty', 'hwy'). n = 234.

Univariate displays: Boxplot

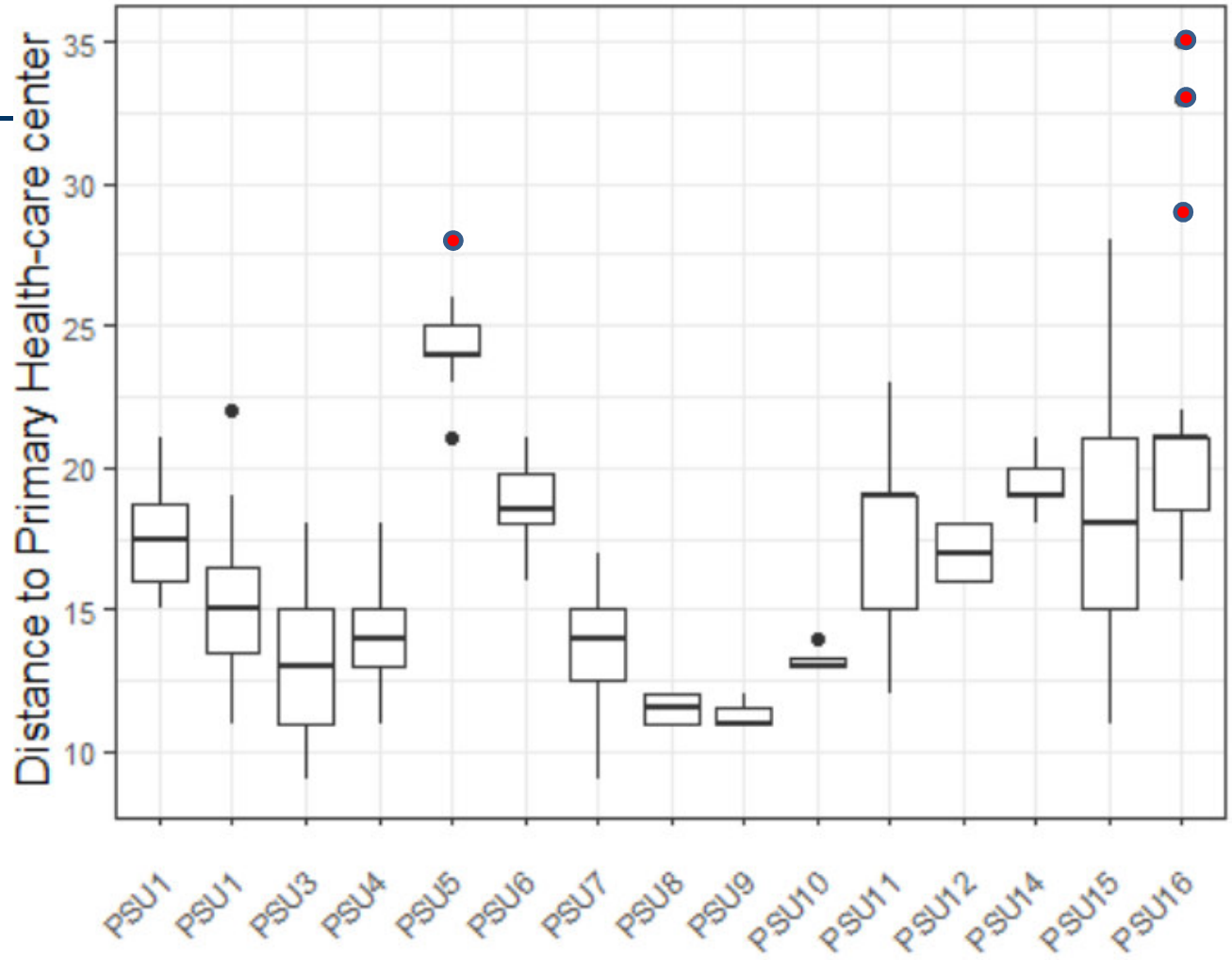


Univariate displays: Histograms or Density plots



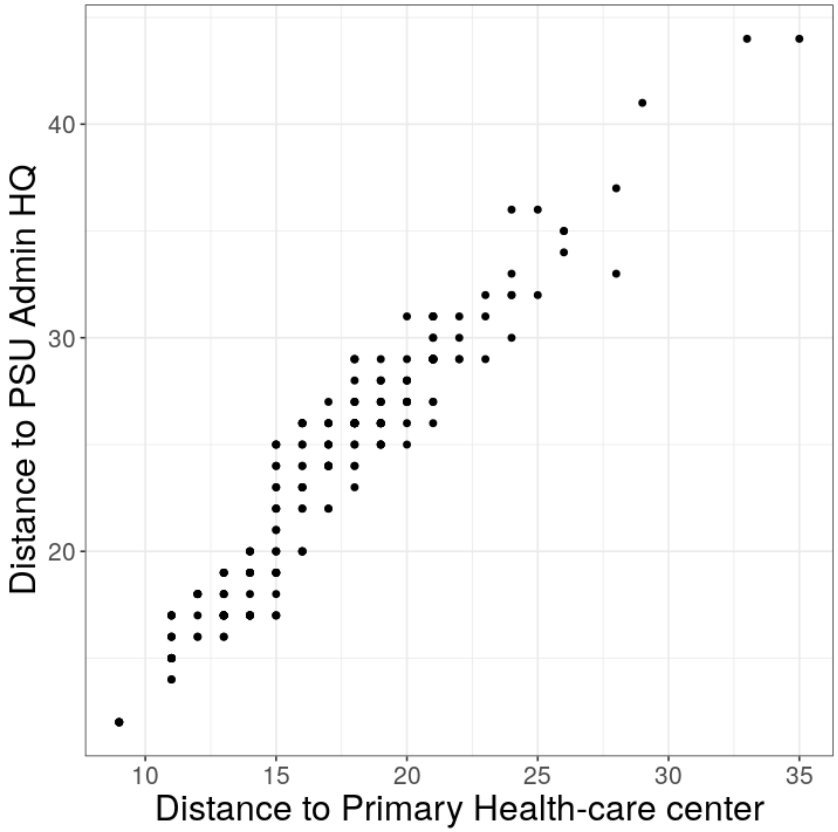
If you are also interested in seeing the shape of the distribution

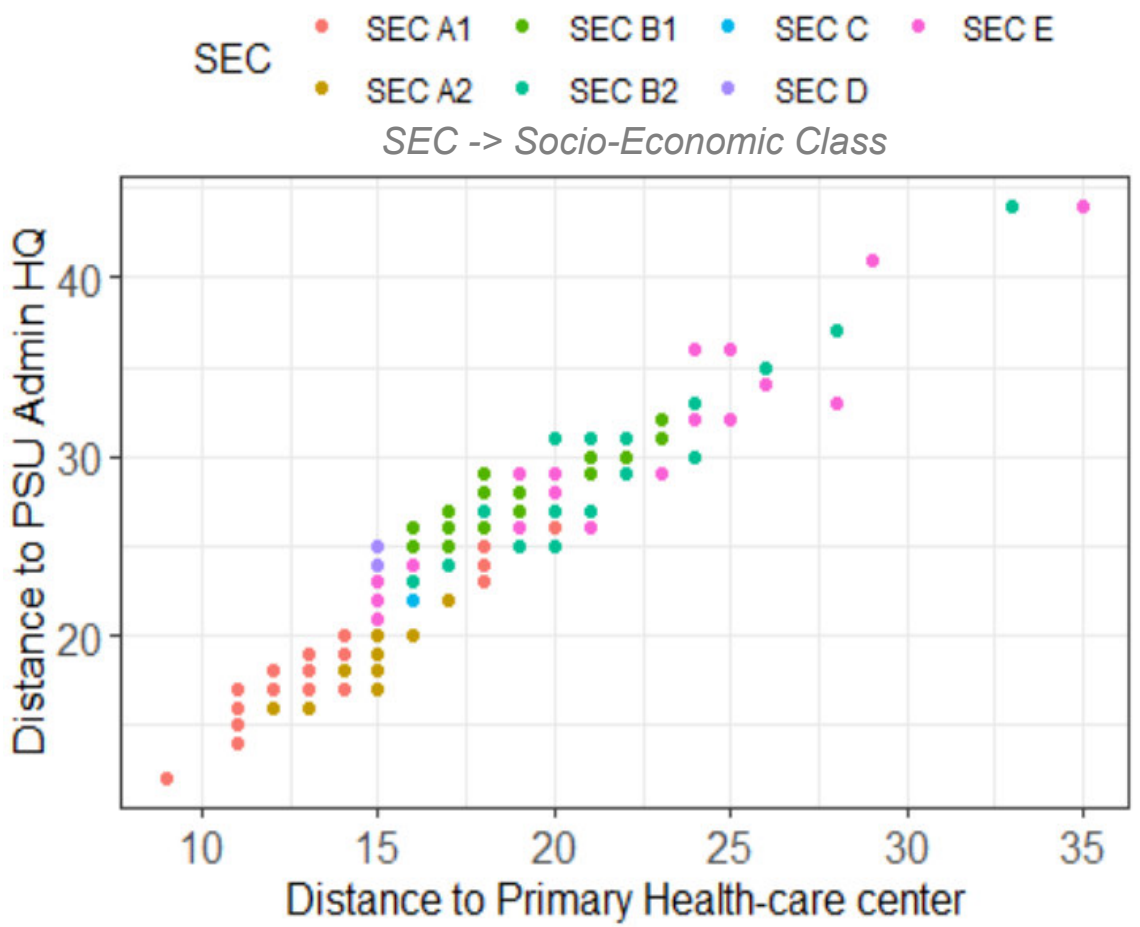
Where are those large values coming from?

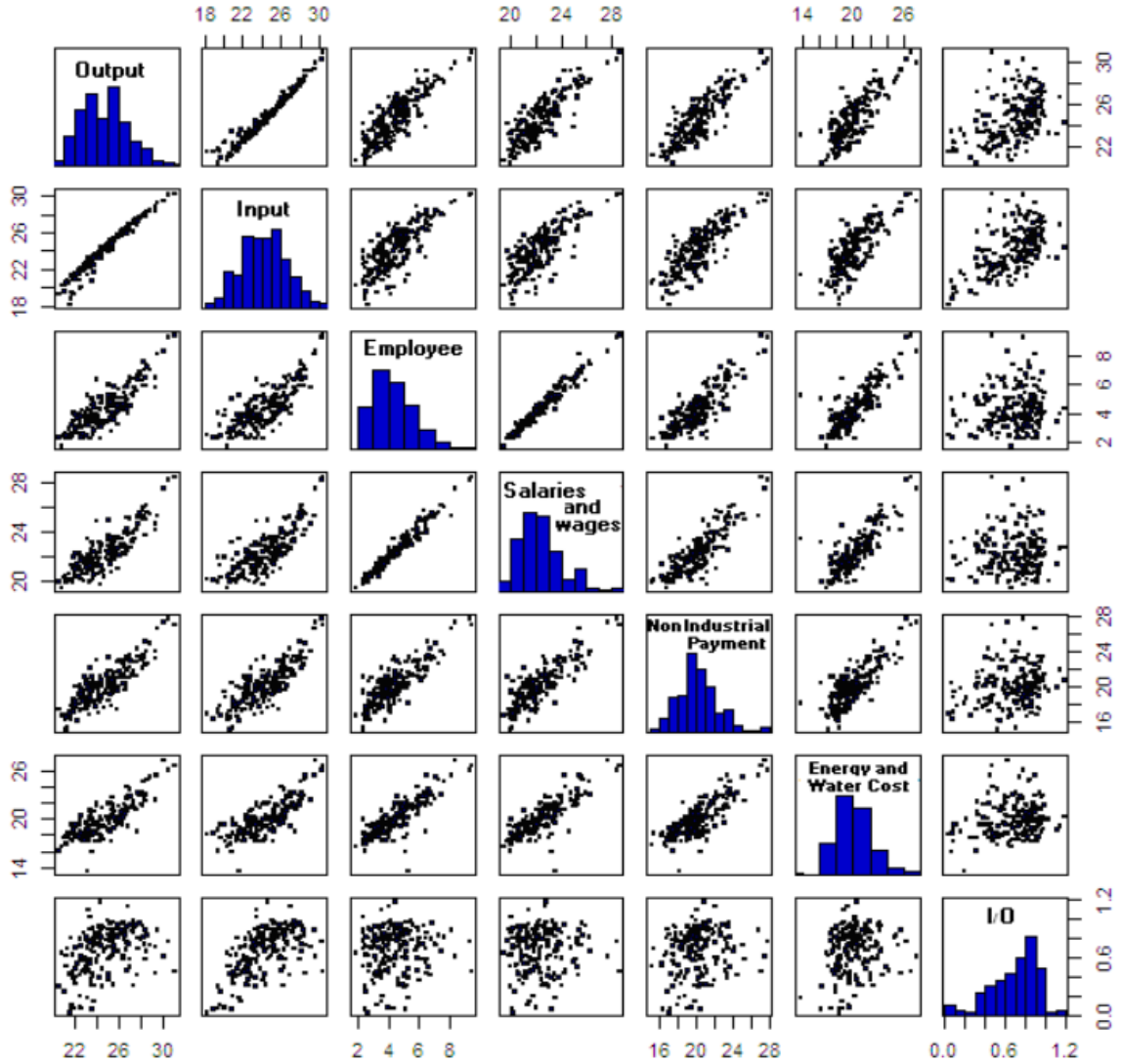

Distance to Primary Health-care center
35
30
25
20
15
10
PSU1
PSU1
PSU3
PSU4
PSU5
PSU6
PSU7
PSU8
PSU9
PSU10
PSU11
PSU12
PSU14
PSU15
PSU16

Bivariate displays: Scatter plot

Color can add information!







Jointly displaying Univariate and Bivariate distributions

Source: Ghahroodi et al 2015

Software for visualization

- Standard statistical software (e.g., SAS, STATA, SPSS) can produce these outputs.

- Open-source software like R have contributed a lot e.g. the ggplot2 package.

- Can also use web-based interactive visualizations. The plots in the earlier slides were made using: https://shiny.gmw.rug.nl/ggplotgui/

For categorical variables

- Categorical variables often used to split the dataset for analysis. But need to analyze them independently as well.

- Summarize/plot a distribution by category (bar plot). Compare with previous waves or external data, if possible. Anything surprising ? e.g. % graduates > % of those less than high-school.

- Contingency tables used when more than one categorical variable
  - SEC x Income categories
  - Any inconsistent combinations? e.g. cases which have “Highest level of education completed” = Bachelors but “Currently pursuing” = Higher secondary.

Consistency checks

- Some checks automatically incorporated into the computer-based instrument.
  - Range checks e.g. Minor child's age < 18 years
  - Response type e.g. number of children must be an integer, an open-ended question on occupation must contain some text values, etc.
  - Logical checks (e.g. no. of years of marriage < age)

- But not possible to have all possible checks in-built. Need to run consistency checks after data collection…

Table 10.1: Range restriction rules for inconsistent and extreme values in the student file

Sequence	Description	SAS Code
1	Invalidate if number for an individual's weight is negative.	if (WB151Q01HA < 0) then WB151Q01HA=.I;
2	Invalidate if number for an individual's height is negative.	if (WB152Q01HA < 0) then WB152Q01HA=.I;
3	Invalidate if number of class periods per week in test language lessons (ST059Q01TA) is greater than 40.	if (ST059Q01TA > 40) then ST059Q01TA =.I;
4	Invalidate if number of class periods per week in maths (ST059Q02TA) is greater than 40.	if (ST059Q02TA > 40) then ST059Q02TA =.I;
5	Invalidate if number of class periods per week in science (ST059Q03TA) is greater than 40.	if (ST059Q03TA > 40) then ST059Q03TA =.I;
6	Invalidate if number of <class periods> per week in foreign language is greater than 40.	if (ST059Q04HA > 40) then ST059Q04HA= .I;
7	Invalidate if number of total class periods in a week (ST060Q01NA) is greater than 120 or less than 10	if (ST060Q01NA > 120 or ST060Q01NA < 10) and NOT MISSING(ST060Q01NA) then ST060Q01NA =.I;

Consistency checks example from Programme for International Student Assessment (2018).

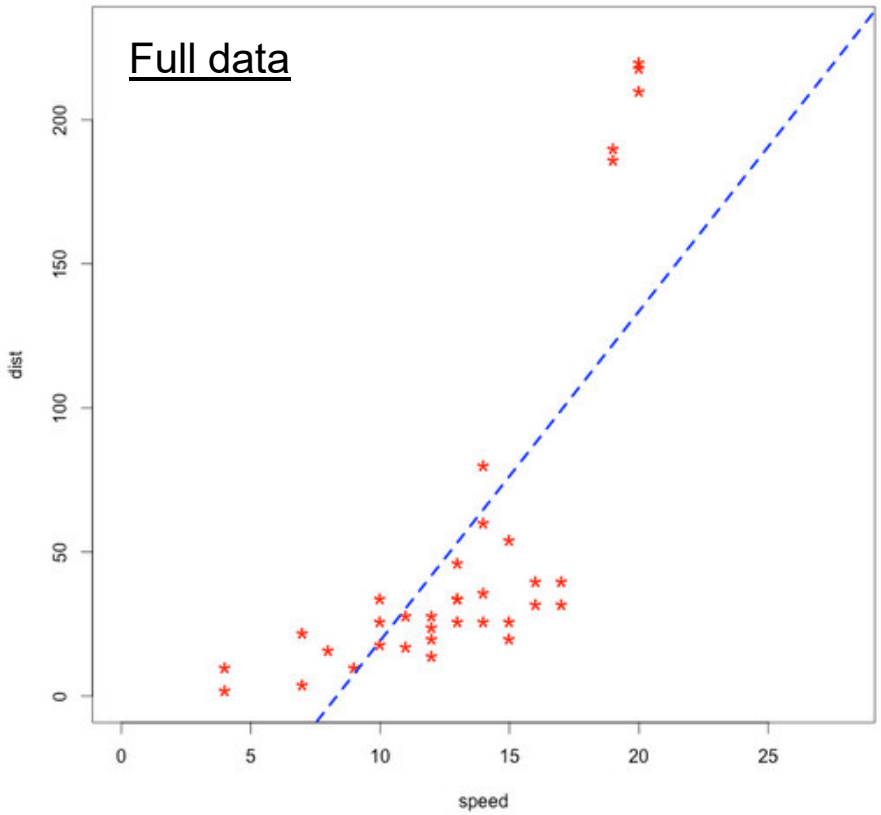
https://www.oecd.org/pisa/data/pisa2018technicalreport/PISA2018%20TecReport-Ch-10-Data-Management.pdf

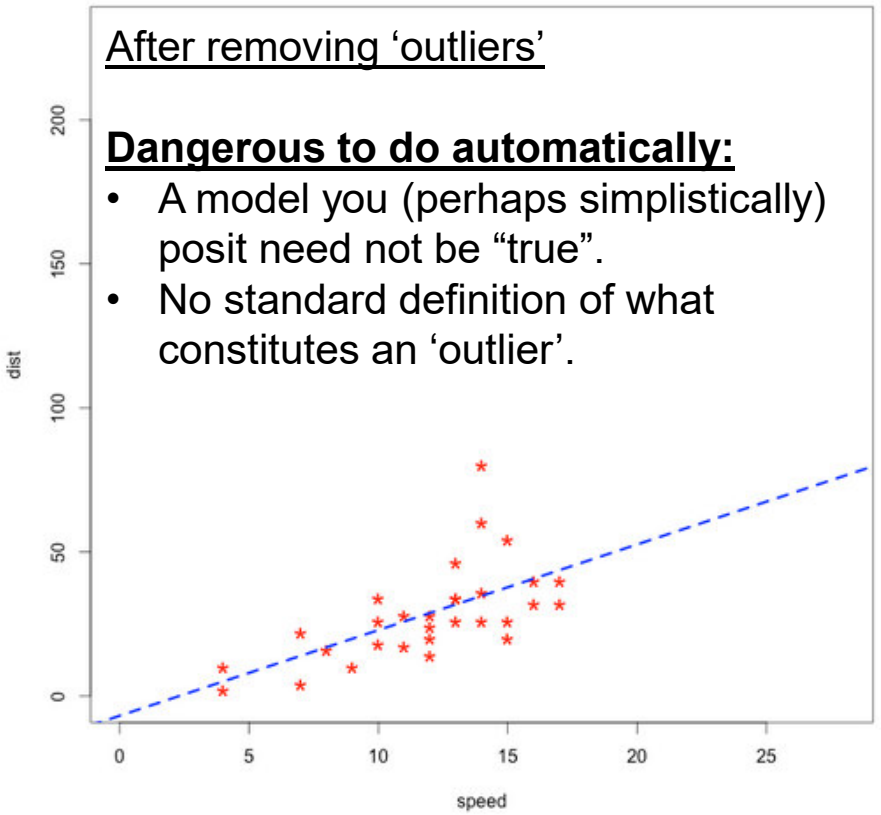
Many activities…

1. Coding open-ended responses
2. Data preparation
3. Summarizing and Visualizing
4. Data editing
5. Imputation and Weighting (covered in Module 2)
6. Disclosure control
7. Final processing, Documentation, and Dissemination

3. Data editing

- What should we do with those ‘extreme’ or inconsistent values we saw earlier?
- Are they plausible?
  - ‘Representative’ outliers (extreme but valid) vs ‘Non-representative’ outliers (errors)
- “numerically distant from the rest of the data”. But what data?
  - Numerous algorithms
- More formally: Data that doesn’t fit a model.





Plot source: r-statistics.co/Outlier-Treatment-With-R.html

Outlier detection – practical issues

1. No clear definition of even what an outlier is.
   - Aguinis et al (2013): Literature review of 46 methodological sources, 232 organizational science journal articles.

     Result: 14 definitions of outliers, 39 outlier detection techniques, and 20 different ways to manage detected outliers.
   - Whatever the definition applied, be consistent.

2. Skewed distributions
   - Can still use the Boxplot but with a Box-Cox transformation applied to the data.

Outlier detection – practical issues

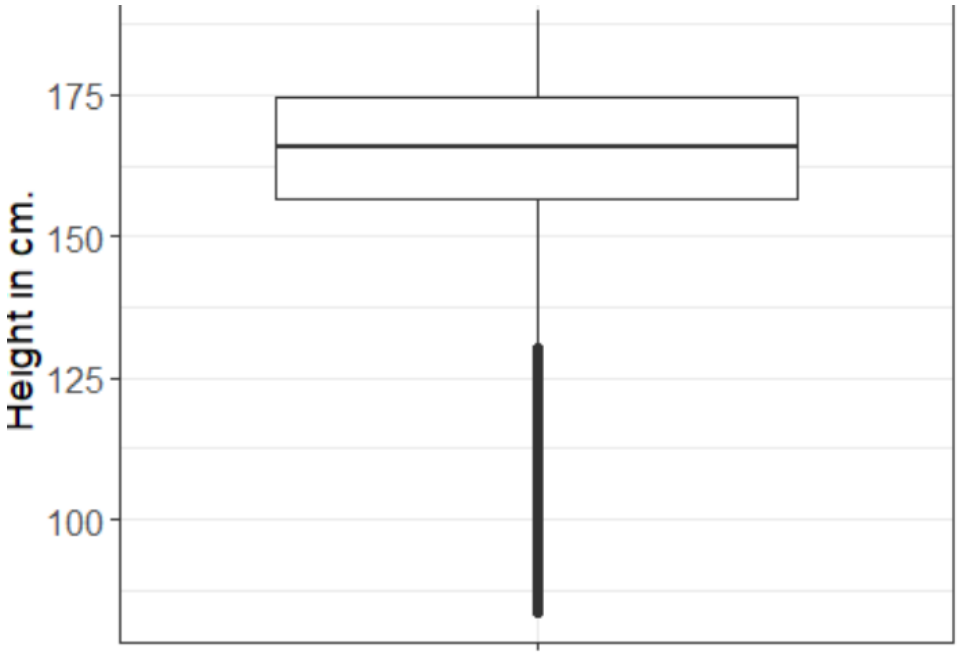
3. Many zeros.
   - Describe/plot without the zeros.
   - For very rare events, usual outlier detection processes may not be valid.

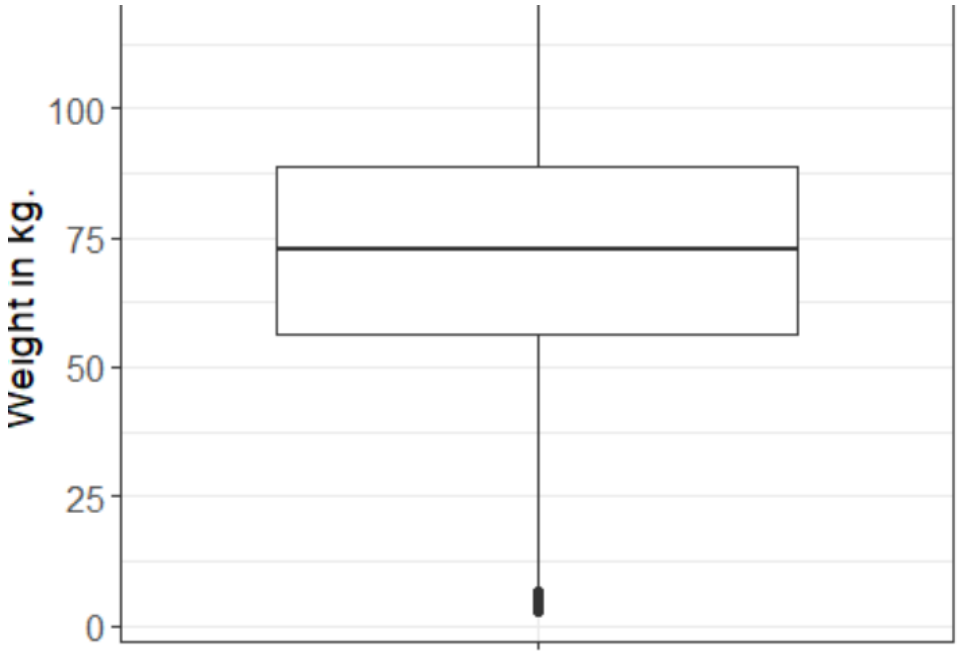
4. Unweighted or weighted?
   - Start with unweighted but do not ignore weighted outlier detection

5. Common mistake: Remove outliers. Redo analysis. Some more come up. Again delete. -> Data editing is itself a source of error.

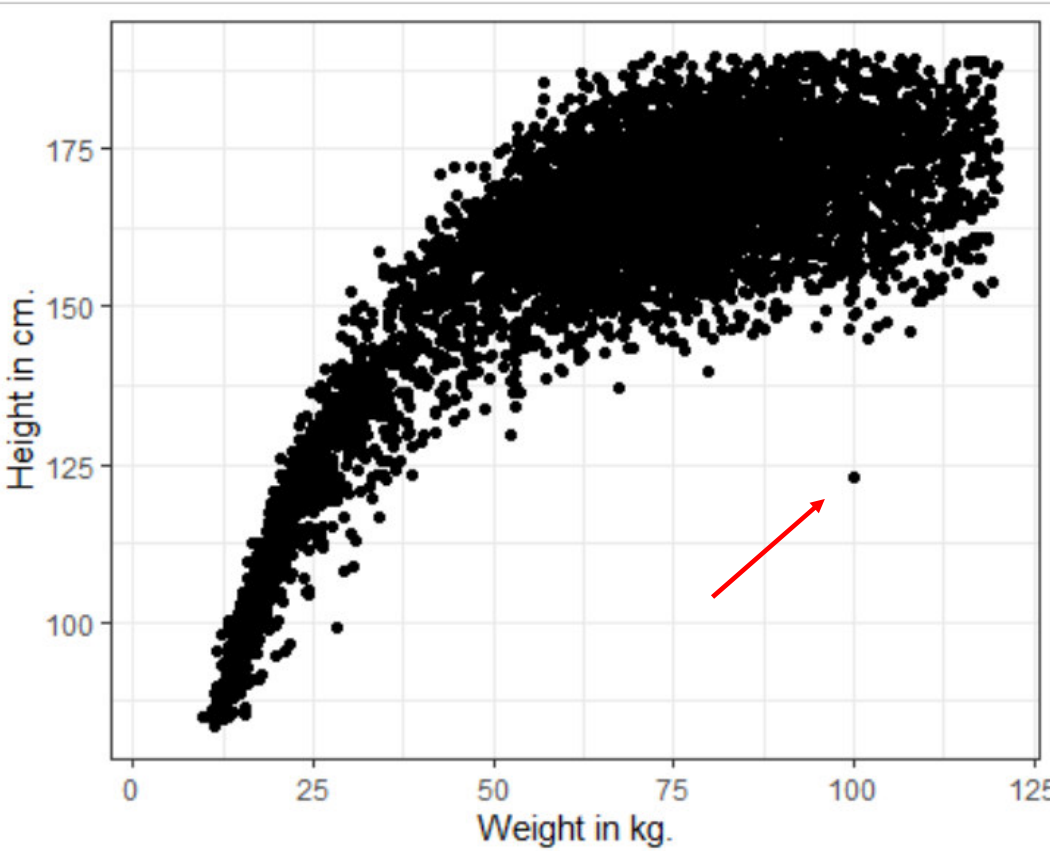
6. Do not stop at univariate analysis….

Univariate boxplots…





Bivariate scatter



Multivariate outliers

- Given that most surveys collect several (100s/ 1000s) of variables, can theoretically have 3D, 4D…
  - Need an automatic solution.

- Several multivariate outlier detection algorithms available e.g. Epidemic, BACON-EEM
  - Use robust estimation to prevent the ‘center’ of the data to itself be distorted by extreme values.
  - Also takes into account sampling weights.

For more detail, see Filzmoser et al (2016) & Todorov et al (2009)

What do we do with outliers or values that fail consistency checks?

1: Keep

If values result from a verified data entry error should we still retain? Can harm analysis.

In some cases, can call the respondent to confirm response.

For panel studies, also worth checking for past values provided by the respondent for the same variable/s.

What do we do with outliers or values that fail consistency checks?

1: Keep

2: Delete

- While automatic detection methods are needed for the typical survey, automatically deletion based on an ad-hoc threshold is not a good idea.
- Outliers need to be investigated.
- Often outliers are an informative part of the data…

A lesson from a non-survey setting

Why hadn't they discovered the phenomenon earlier? Unfortunately, the TOMS data analysis software had been programmed to flag and set aside data points that deviated greatly from expected measurements and so the initial measurements that should have set off alarms were simply overlooked. In short, the TOMS team failed to detect the ozone depletion years earlier because it was much more severe than scientists expected.

https://earthobservatory.nasa.gov/features/RemoteSensingAtmosphere/remote_sensing5.php

What do we do with outliers or values that fail consistency checks?

1: Keep

2: Delete

3: Winsorize/Statistically adjust

4: Impute

Winsorize/Statistically adjust

- Values above/below a cut-off are adjusted to that cut-off.
- Never a free lunch: You may be reducing variance but bias may increase (→ Module 1 for concepts of Bias and Variance).
- For skewed distributions, procedures such as Pareto tail modeling can be used. Values above the distribution-defined threshold are replaced by a predicted value.

What do we do with outliers or values that fail consistency checks?

1: Keep

2: Delete

3: Winsorize

4: Impute

Many activities…

1. Coding open-ended responses
2. Data preparation
3. Summarizing and Visualizing
4. Data editing
5. Imputation and Weighting (weighting covered in Module 2)
6. Disclosure control
7. Final processing, Documentation, and Dissemination

END OF Video 2