CENSUS 2023

DATA PROCESSING

DAVID CLARKE

- Career Software Developer
- Microsoft stack
- Pascal, C, C++, C#, JavaScript, Java, SQL, Python
- My Honesty Box www.myhonestybox.co.nz
- Zebra Crossing www.zebracrossing.co.nz

BACKGROUND

- 2018 Census not entirely successful
- Census Development squad dispersed
- August 2022 new team assembled
- Census date 3 March 2023

OUTLINE

- Inputs and Outputs
- Tools and frameworks
- Classifications
- Data Sourcing and Imputation
- Specifications
- Demo

INPUT AND OUTPUTS

Inputs

- Response store with individual and dwelling responses as JSON via HTTP
- Admin and historical data SQL Server
- Location data SQL Server

INPUTS AND OUTPUTS

Outputs

- Individuals table ~4,900,000 rows, 292 columns
- Dwellings table ~2,040,000 rows, 74 columns
- Household table ~1,900,000 rows, 23 columns
- Family table ~1,170,000 rows, 28 columns
- Extended Family table ~121,000 rows, 8 columns
- Supporting “Extra” tables including intermediate and input columns

INPUTS AND OUTPUTS

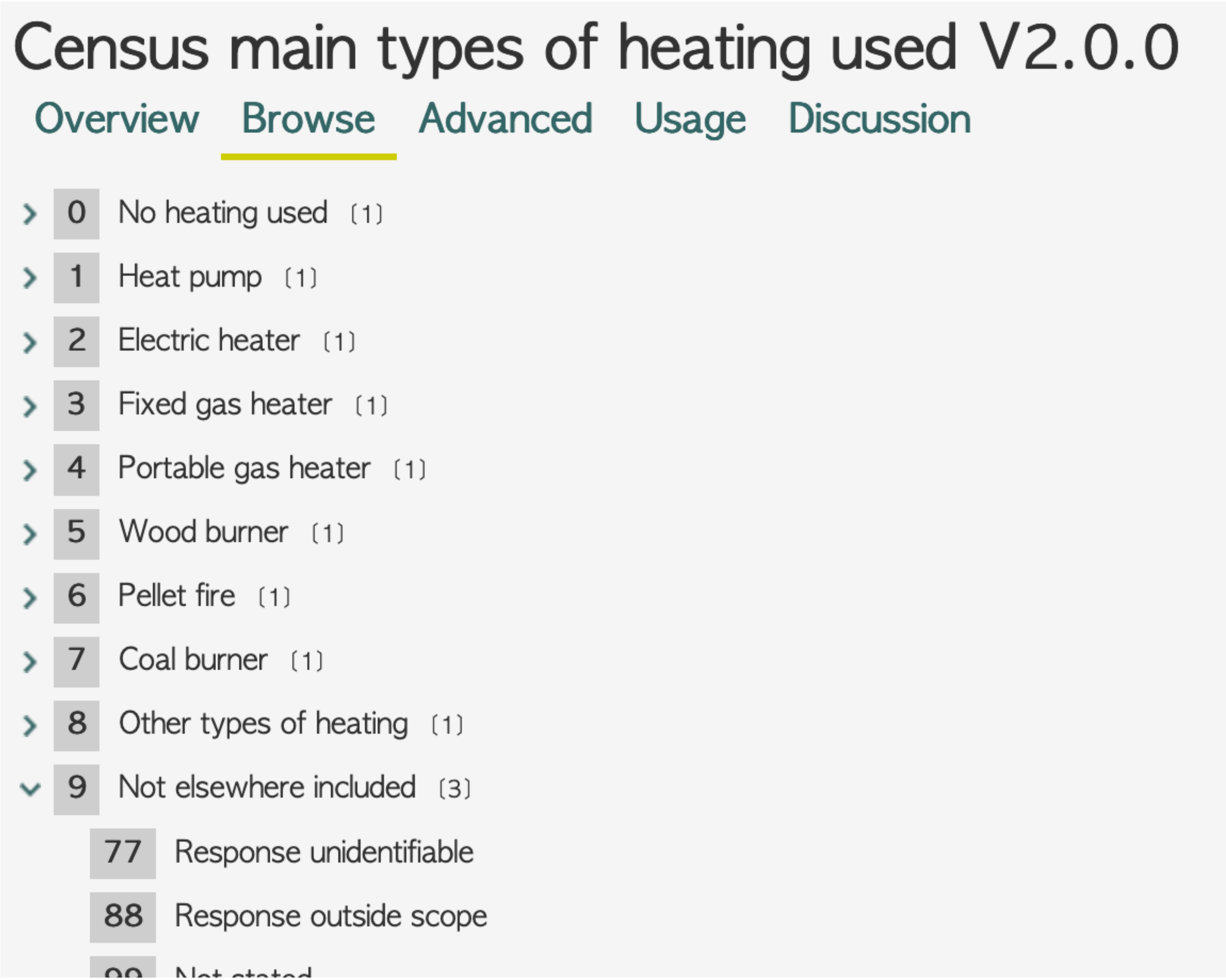
- All data treated as strings
- 280 specs/processing modules
- Census Processing takes 60+ hours to complete

TOOLS AND FRAMEWORKS

- Python 3.8 and a small C++ module
- Pandas and Numpy everywhere
- Behave and pytest
- Mypy
- Poetry
- Visual Studio Code
- Python Notebooks
- Azure Data Studio/SQL Server Management Studio

CLASSIFICATIONS

- Aria https://aria.stats.govt.nz/aria/
- E.g. main types of heating used
- qcyomC7wufic8HZJ

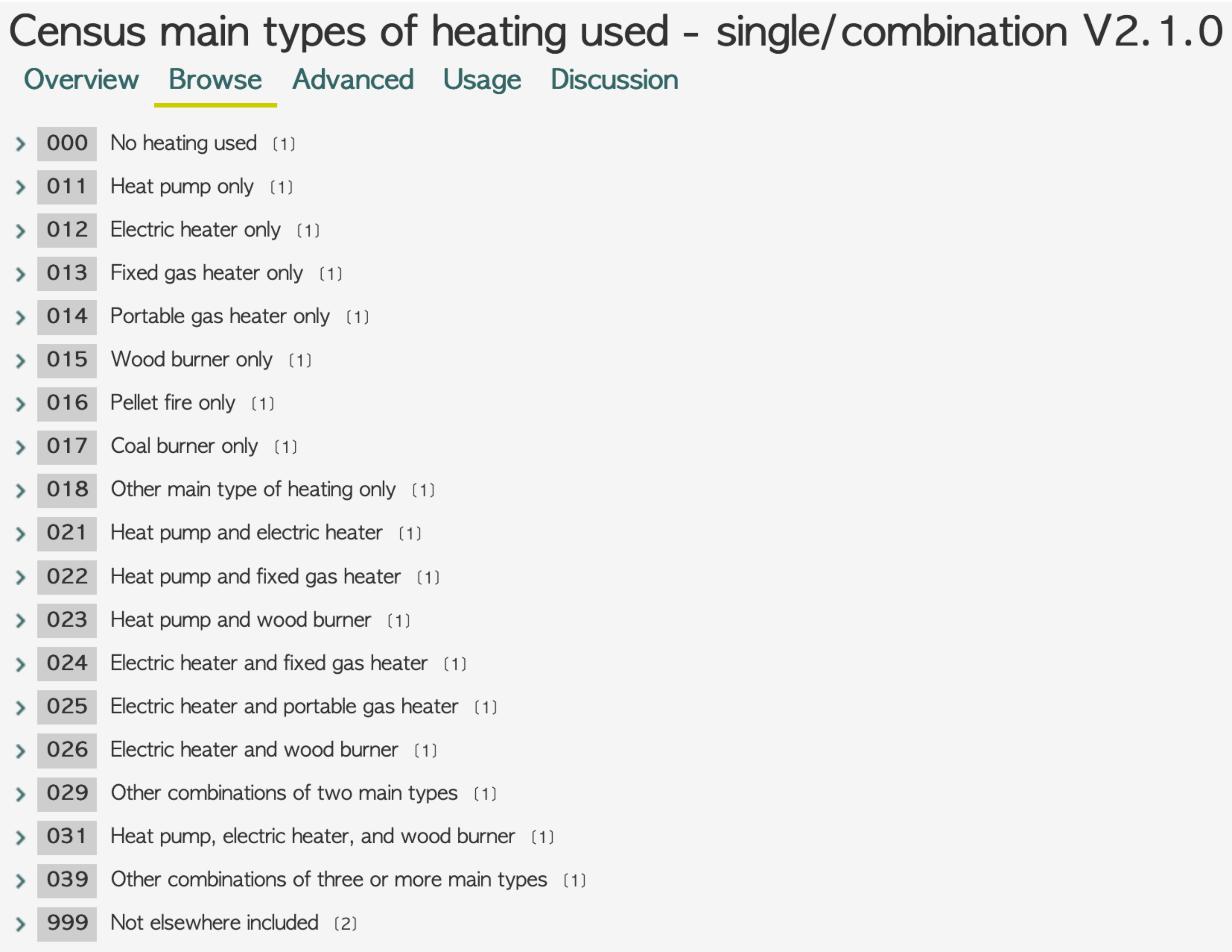


Census main types of heating used V2.0.0

0	No heating used	(1)
1	Heat pump	(1)
2	Electric heater	(1)
3	Fixed gas heater	(1)
4	Portable gas heater	(1)
5	Wood burner	(1)
6	Pellet fire	(1)
7	Coal burner	(1)
8	Other types of heating	(1)
9	Not elsewhere included	(3)
77	Response unidentifiable	
88	Response outside scope	
99	Not stated	

CLASSIFICATIONS

- Derivations
- **Census main types of heating used - single/combinationV2.1.0**
- kXrpegiKTUFysYWU



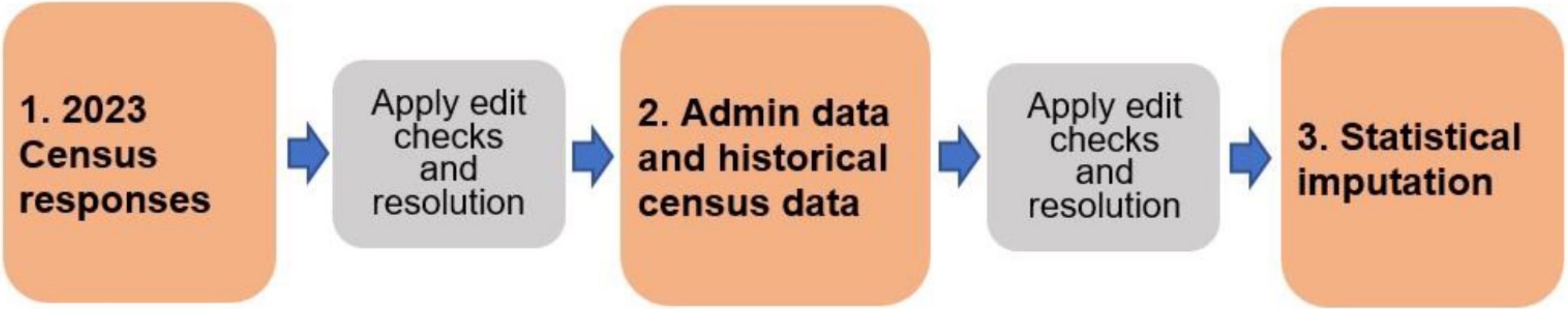
Census main types of heating used - single/combination V2.1.0

000	No heating used	〔1〕
011	Heat pump only	〔1〕
012	Electric heater only	〔1〕
013	Fixed gas heater only	〔1〕
014	Portable gas heater only	〔1〕
015	Wood burner only	〔1〕
016	Pellet fire only	〔1〕
017	Coal burner only	〔1〕
018	Other main type of heating only	〔1〕
021	Heat pump and electric heater	〔1〕
022	Heat pump and fixed gas heater	〔1〕
023	Heat pump and wood burner	〔1〕
024	Electric heater and fixed gas heater	〔1〕
025	Electric heater and portable gas heater	〔1〕
026	Electric heater and wood burner	〔1〕
029	Other combinations of two main types	〔1〕
031	Heat pump, electric heater, and wood burner	〔1〕
039	Other combinations of three or more main types	〔1〕
999	Not elsewhere included	〔2〕

DATA SOURCING AND IMPUTATION

- Editing, data sourcing, and imputation: Planned approach for the 2023 Census
  - https://www.stats.govt.nz/assets/Methods/Editing-data-sourcing-and-imputation-Planned-approach-for-the-2023-Census/Editing-data-sourcing-and-imputation-Planned-approach-for-the-2023-Census-.pdf

Figure 1: Hierarchy of data sources



1. 2023 Census responses

Apply edit checks and resolution

2. Admin data and historical census data

Apply edit checks and resolution

3. Statistical imputation

DATA SOURCING AND IMPUTATION

2. Admin data and historical census data

Sources of information about the individual or dwelling other than a 2023 Census form:

- Historical (2013 or 2018) census responses – information provided from a previous census form. Only information from an actual response is included. Imputed or alternatively sourced values within those previous census files will not be considered as historical census information.
- Admin data – information taken from an admin data source, such as birth registrations or tax data. This may also include a small amount of information from other Stats NZ surveys.

DATA SOURCING AND IMPUTATION

3. Statistical imputation

- Within-household donor imputation – for example, in 2018, the person closest in age to the respondent in the respondent's usual residence household is selected as a donor. This is most likely to be used for cultural variables, such as ethnicity, Māori descent, religious affiliation, and language. Other implementations of within-household donor imputation could also be considered.
- Deterministic imputation – the attribute is derived from other available variables. In previous censuses, sex was imputed based on name, and Māori descent (electoral) was imputed based on iwi affiliation.
- Donor imputation – information is taken from a similar record, with donors found using nearest-neighbour imputation methodology (NIM).

DATA SOURCING AND IMPUTATION

Editing is used to identify and resolve erroneous and suspicious data. Common errors that are the subject of edits include:

- basic errors in form filling, such as ticking multiple boxes on a single-response question
- illegible or ambiguous marks
- inconsistent responses, such as a person saying they have no sources of income, and yet they had an income of $50,000
- suspicious responses, for example, if somebody says that they are 16 years old but have a master’s degree.

DATA SOURCING AND IMPUTATION

Donor imputation will be applied in CANCEIS (the CANadian Census Editing and

Editing, data sourcing, and imputation: Planned approach for the 2023 Census Imputation System).

CANCEIS is Stats NZ’s standard corporate tool for statistical imputation and is also used internationally by several statistical agencies

MANUAL INTERVENTION

- Input data exists
- Unable to code, Response Unidentifiable
- Data reviewed/updated by staff
- Minimise requirement for MI

SPECIFICATIONS

- Written by analysts
- Word documents
- Input variables and classifications
- Output variables and classifications
- Pseudo code processing instructions

SPECIFICATIONS

Census main types of heating used

Input variable	Classification
dwell_nbr	
d_heating_predefined	
d_heating_text	
d_heating_code_ics	

Output variable	Classification
d_heating_predefined_code	
d_heating_code_data_source	

DEMO