VCOSS Cleaning Notes 2020

Ben Cole

23 Sep 2020

Contents

Previous work **1**

1 Individual Charities Data **3**
1.1 Data Source . . . 3
1.2 Data_Cleaning.R script . . . 3
1.2.1 Recoding booleans . . . 3
1.2.2 Checking for Invalid ABNs . . . 3
1.2.3 Main Activity . . . 3
1.2.4 Victorian Charities . . . 3
1.2.5 Organisation size . . . 3
1.2.6 VCOSS Charity sizes . . . 4
1.2.7 Registration Status . . . 4
1.2.8 Inaccurate Income . . . 4
1.2.9 Inaccurate Expenditure . . . 4
1.2.10 Negative variables . . . 5
1.2.11 Creating files . . . 5

2 Group Charities **5**

3 Usage **5**

4 Web repository **5**

Previous work

The Victorian Council of Social Service (*VCOSS*) has produced reports in past years using ACNC data cleaned with proprietary software requiring subscriptions. The R language was chosen to clean the data sets for this new VCOSS report so that the cleaning procedures could be transparent and accessible without need for purchasing software.
Previous reports from VCOSS had not cleaned the 2018 ACNC data as it had not been released at the time. Considering the 2018 data needed to be cleaned and filtered to VCOSS guidelines, it was decided to clean the ACNC data from years 2014 - 2018 inclusive for consistency. Some

new decisions were made while cleaning the data that produced different data sets than previously used by VCOSS.

1 Individual Charities Data

1.1 Data Source

The ACNC data was sourced from **the data.gov.au website** for **2014**, **2015**, **2016**, **2017**, and **2018** at the beginning of the project.

1.2 Data_Cleaning.R script

As with any script of code, the *Data_Cleaning.R* script serves as a record of the cleaning procedures as well as a reusable+reproducible means for repetitive cleaning of the data.

1.2.1 Recoding booleans

Boolean columns in the data were stored with the values "y" and "n" throughout the years, so were recoded to the native R boolean format `TRUE` and `FALSE`.

1.2.2 Checking for Invalid ABNs

This is legacy code from previous work by the Future Social Service Institute and was included in initial stages of cleaning. Subsequent consultation with stakeholders led to the decision being made not to remove charities with invalid ABNs. The code is left in place, but the filter is not applied.

1.2.3 Main Activity

Selecting charities based on their reported *Main Activity* was the first filter applied by the script. If a charity reported that they performed any of the activities from the list below they were retained, and any charity that did not was removed.
The following activities varied in their naming structure throughout the years, so variants were deliberately used to catch them. For example, `Main Activity` in the 2015 data contained "Civic and advocacy activities", while the 2016 data contained *both* "Civic and advocacy activities" and "Advocacy and civic activities".

Aged care activities
Civic and advocacy activities
Economic, social, and community development
Emergency relief
Employment and training
Housing activities
Income support and maintenance
International activities
Law and legal services
Mental health and crisis intervention
Other education
Other health service delivery
Social services

1.2.4 Victorian Charities

If the charity listed their state as Victoria **and** that they operated in Victoria they were retained in the data. Any charity that didn't meet both these criteria were removed, eg if they were based in Victoria but did not operate in

1.2.5 Organisation size

The reported size of charities was inconsistent across years with some charities listing their size as "Samll". Charity size needed to be cleaned, so the first letter used as an indicator of size - S, M, and L. The original charity size variable was left untouched and a new variable `cleaned_charitysize` was created.
This cleaned charity size was then checked against the reported Total Gross Income at the following thresholds:

- Small - if Total Gross Income <= $250,000

- Medium - if Total Gross Income > $250,00 and <= $1,000,000
- Large - if Total Gross Income > 1,000,000

If a charity reported Total Gross Income incosistent with these thresholds they were removed from the dataset.

1.2.6 VCOSS Charity sizes

VCOSS also uses more granular definitions of charity sizes, so a new variable was created using Total Gross Income

- Extra small - if Total Gross Income < $50,000
- Small - if Total Gross Income >= $50,000 and < $250,000
- Medium - if Total Gross Income >= $250,000 and < $ 1m
- Large - if Total Gross Income >= $1m and < $10m
- Extra large - if Total Gross Income >= $10m and < $100m
- Extra extra large - if Total Gross Income >= $100m

1.2.7 Registration Status

A charity was removed if its registration status was any variant of “revoked”.

1.2.8 Inaccurate Income

In previous years charities with no Total Gross Income were removed from the datasets, but the decision was made not to take this step on consultation with stakeholders.
The sum of all income sources was checked against Total Gross Income. If the absolute difference between the income sources and Total Gross Income fell into the below criteria a charity was removed.
The sum of income fields used were:

- Government Grants
- Donations and Bequests
- All Other Revenue
- Revenue from Goods and Services
- Revenue from Investments
- Other Income

Inaccurate Income Criteria:

- If small charity and if absolute difference > $25,000
- If medium charity and if absolute difference > $100,000
- If large charity and if absolute difference > $1,000,000

1.2.9 Inaccurate Expenditure

Similar to income, absolute difference in expenditure data was also checked. If a charity’s absolute difference between expenses fields and reported Total Expenses fell into the below criteria they were removed from the data set.
The sum of expense fields used were:

- Employee Expenses
- Interest Expenses
- Grants and Donations Made for use in Australia
- Grants and Donations Made for use outside Australia
- All Other Expenses

Inaccurate Expenses Criteria:

- If small charity and if absolute difference > $25,000
- If medium charity and if absolute difference > $100,000
- If large charity and if absolute difference > $1,000,000

1.2.9.1 Average employee expenses

If a charity spent more than $300,000 per employee they were removed from the datasets. Employees included paid employees of all types as well as volunteers. If a charity reported 0 employees and employee expenses other than $0 they were removed.

1.2.10 Negative variables

Some numeric fields of data should not have values < 0 recorded. Any charity listing negative values in the following variables was removed.

- Staff Full Time
- Staff Part Time
- Staff Casual
- Staff Volunteers
- Government Grants
- Donations and Bequests
- All Other Revenue
- Other Income
- Employee Expenses
- Interest Expenses
- Grants and Donations Made for use in Australia
- Grants and Donations Made for use outside Australia
- All Other Expenses

If a charity reported Government Grants greater than their Total Gross Income they were removed from the data.

1.2.11 Creating files

After above cleaning was completed, the datasets were combined into one dataframe containing all years. The combined dataset and the datasets for each year were all written to .csv format before combining in MS Excel into a master file. No cleaning was performed in Excel, only aesthetic formatting.

2 Group Charities

The data sets on Group Reporting charities varied throughout the years and were not addressed in the data cleaning. Many reporting groups were included in the individual charities data for 2018 and retained if they met the cleaning criteria set out above.

3 Usage

The cleaned and filtered ACNC data was used for analyses informing VCOSS of the state and structure of community service charities operating in Victoria. Some reporting was produced in Tableau and some reporting was produced with R, the latter of which can be viewed at the git repository.

4 Web repository

A git repository was used for this project and will be left open to access. Please **click here** to view the web repository.