# DataBerries

Kainaat Amjid

Nazgul Rakhimzhanova

# Finding Data

## Datasets

<table>
<tr><td colspan="3"><b>DATASET 01</b> ONLINE DONATIONS MADE ON PLATFORM GOFUNDME</td></tr>
<tr><td colspan="3"><b>Data release date</b> July 10, 2019</td></tr>
<tr><td colspan="3"><b>Content</b></td></tr>
<tr><td colspan="3">Anonymised data about donations made on the platform during the 2012 - 2016 years (exact dates are not included with respect to the privacy of donation campaigns). Each row represents one donation made, data also includes the donor's gender, reason to donate, amount of donation made, average of donations made by others, which was visible during the donation period, empathy level of donors and some other interesting information. Dataset consists of file A - donation data, file B - donation reasons interview data.</td></tr>
<tr><td><b>Volume</b></td><td>File A 558067 rows, 9 columns<br>File B 305 rows, 15 columns</td><td><b>Date</b>: 2012 - 2016<br><b>Country</b>: USA</td></tr>
<tr><td colspan="3"><b>By</b>: Sisco, Matthew Ryan; Weber, Elke U - research paper authors.<br><b>From</b>: The publicly accessible data was downloaded from the GoFundMe website during May and June of 2016.<br><b>Method:</b> Automated, from the environment of GoFundMe platform, more details are not provided. Data collected during the interview of 305 respondents.</td></tr>
<tr><td><b>The resource link(s)</b></td><td colspan="2">Data files https://doi.org/10.7916/d8-cckc-3f61<br>Related research paper ttps://www.nature.com/articles/s41467-019-11852-z</td></tr>
<tr><td><b>Dataset found by</b></td><td colspan="2"><b>Nazgul K. Rakhimzhanova,</b><br>using the general google search engine https://www.google.fr/<br>Collected on October 10, 2022</td></tr>
<tr><td><b>Data availability</b></td><td colspan="2">No licence required, open to the public.</td></tr>
</table>

<table>
<tr><td colspan="3"><b>DATASET 02</b> GENERAL SOCIAL SURVEY CYCLE 33: GIVING, VOLUNTEERING AND PARTICIPATING</td></tr>
<tr><td colspan="3"><b>Data release date</b> January 26, 2021</td></tr>
<tr><td colspan="3"><b>Content</b></td></tr>
<tr><td colspan="3">Anonymised data of the survey performed by the GSS program (General SOcial Survey) among Canada population. Dataset covers detailed information about respondents' formal volunteering, informal volunteering and donations made to different categories of charity organisations, including the demographic information such as household size, marital status, education level, gender, income level, religious preferences, etc. Amount of donations made, hours of volunteering, type of volunteering activities and reasons to donate or not to donate are also covered by survey data.</td></tr>
<tr><td><b>Volume</b></td><td>16530 rows, 955 columns</td><td><b>Date</b>: 2018<br><b>Country</b>: Canada</td></tr>
<tr><td colspan="3"><b>By</b>: Statistics Canada (Survey manager Patric Fournier-Savard, patric.fournier-savard@canada.ca )<br><b>From</b>: Online survey platform of GSS.</td></tr>
</table>

| | |
|---|---|
| **Method:** Online questionnaire, pre-questionnaire interview, mailing letters. | |
| **The resource link(s)** | Data files: https://abacus.library.ubc.ca/dataset.xhtml?persistentId=hdl:11272.1/AB2/GBFDYG<br>Survey description https://www23.statcan.gc.ca/imdb/p2SV.pl?Function=getSurvey&Id=796234 |
| **Dataset found by** | **Nazgul K. Rakhimzhanova,**<br>using the dataset search google search engine. https://datasetsearch.research.google.com/<br>Collected on October 12, 2022 |
| **Data availability** | Statistics Canada Open Licence |

| | | |
|---|---|---|
| **DATASET 03** DATASET - MORAL FOUNDATIONS THEORY AND THE PSYCHOLOGY OF CHARITABLE GIVING | | |
| **Data release date** June 18, 2018, **updated** April 19, 2020 | | |
| **Content** | | |
| Donator's demographic information as age, gender, social status, education and moral foundations, beliefs and reasons to participate in charity. Anonymised. | | |
| **Volume** | 985 rows, 27 columns | **Date**: 2018<br>**Country**: Sweden |
| **By**: Daniel Västfjäll, Department of Behavioural Sciences and Learning, Linköping University, Linköping, Sweden.<br>**From**: Online questionnaires.<br>**Method:** Interview and quota sampling afterwards. | | |
| **The resource link(s)** | Data files: https://osf.io/49ehm , https://osf.io/mcwv7/?view_only=<br>Related research paper https://journals.sagepub.com/doi/10.1002/per.2256 | |
| **Dataset found by** | **Kainaat Amjid,**<br>using the dataset search google search engine. https://datasetsearch.research.google.com/<br>Collected on October 10, 2022 | |
| **Data availability** | Creative Commons Attribution Licence, open access | |

| | | |
|---|---|---|
| **DATASET 04** VOLUNTEER ACTIVITIES SURVEY 2018 | | |
| **Data release date** February 03, 2021 | | |
| **Content** | | |
| Anonymised data that includes: types of volunteer activities, reasons for volunteer activity, time spent on volunteer activity, and monetary value for volunteer activity. | | |
| **Volume** | 2832 rows, 43 columns | **Date**: June - August, 2018<br>**Country**: South Africa |

<table>
<tr><td colspan="2"><b>By</b>: Statistics South Africa, the national statistical agency of South Africa<br><b>From</b>: Online questionnaires.<br><b>Method:</b> National Interview and sampling.</td></tr>
<tr><td><b>The resource link(s)</b></td><td>Data files:<br>https://www.datafirst.uct.ac.za/dataportal/index.php/catalog/850/data-dictionary<br>Metadata:<br>https://www.datafirst.uct.ac.za/dataportal/index.php/catalog/850/study-description</td></tr>
<tr><td><b>Dataset found by</b></td><td><b>Kainaat Amjid,</b><br>using the dataset search google search engine. https://datasetsearch.research.google.com/<br>Collected on October 16, 2022</td></tr>
<tr><td><b>Data availability</b></td><td>Creative Commons Attribution 4.0 International Licence, open access.</td></tr>
</table>

Some other interesting dataset we came across, which worth mentioning:

- Anonymised donation transactions' data provided by MasterCard for USA, 2016-2017. https://nccs.urban.org/project/mastercard-donation-insights#data
- The annual NYC Volunteers Count report is the City's largest scan of residents volunteering at organisations, 2018. https://data.cityofnewyork.us/Social-Services/2019-Volunteers-Count-Report-Neighborhoods/72r6-mtgs

## Rethinking the Research question

Our initial research questions were as:

1. What is the trend of charity organisations' (NPO, NGO etc) targets in different countries for the last decade?
2. What are the preferred ways( through church, government, NGO, beggars, websites, events etc) and demographics of donors in different countries?
3. How transparent charity organisations are in different countries, is there relation between transparency and donors' trust to an organisation?

By examining the data we have collected we decided to work with the dataset provided by the Canadian Statistic Agency - **DATASET 02 GENERAL SOCIAL SURVEY CYCLE 33: GIVING, VOLUNTEERING AND PARTICIPATING.** As this dataset contains perfectly detailed information about the population engaging in charity activities like formal volunteering, informal volunteering and giving. The dataset also includes enough information about population demographics like age group, gender, education, marital status, social status, religious and spiritual beliefs, networking. In this survey data one also can find reasons for participating in charity as well as reasons for not participating in charity.

Taking into account the data we are working with, we narrowed our research question to:

**What is the population's behaviour of participation in charity across provinces of Canada?**

Our potential end result could be an interactive map (we are studying this topic now), where by clicking on the province area a User can see information about :

- What is the preferred activity of participation in offline and online charity: volunteering or donation?
- What are the preferred ways of giving donations - mail, online, shopping centre stands, etc?

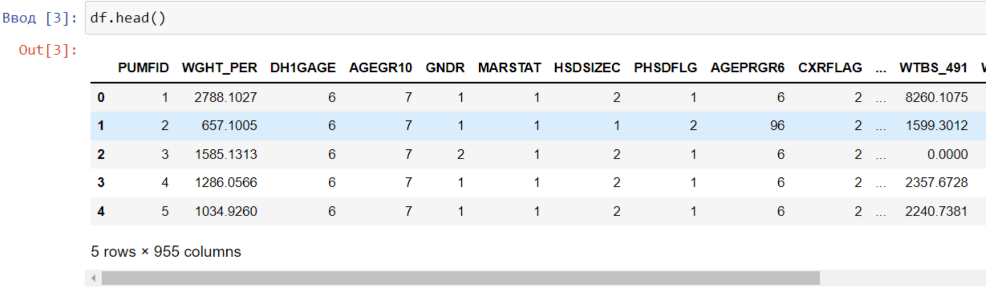
- If this province supports international charity organisations more or local ones?
- What are the top popular categories of charity organisations for this province?
- Is there a difference in reasons for participating according to the gender, education, age group and religious beliefs?
- How do people find charity organisations - internet, email, word of mouth, advertisements?
- What is the average amount of donations?

From here, we would pinpoint that information about reasons for participating in charity and the common ways of finding the charity organisation could be very useful for those, who are planning to promote charity campaigns.

**Why did we modify our initial research question?** - Our initial goal was to create some sort of a "population's behavioural profile" regarding their participation in charity for different countries to answer questions like, for example, "If educated women from Europe are more likely to support environmental charity organisations compared to educated women from South America?". This would be helpful for people who are planning to run and promote charity campaigns to target their audience more efficiently. But the obstacles we have encountered are:
**The first** and very serious one - we could collect data, as in many cases governmental statistics agencies release already aggregated data without possibility to do exploratory analysis on it, and not so many agencies perform detailed surveys of donations.
**Second**, some countries don't even provide publicly accessible data, or they don't provide an English version for it. For example, we found charity and donation data for Chile provided by a statistical agency, but it was a paid dataset and cost around 200 US dollars.
**Third**, some collected data couldn't be compared as they cover different time periods, like 2003 and 2016, this gap is too significant to perform comparative analysis.

# Data Cleaning

The dataset values were encoded, as follows:

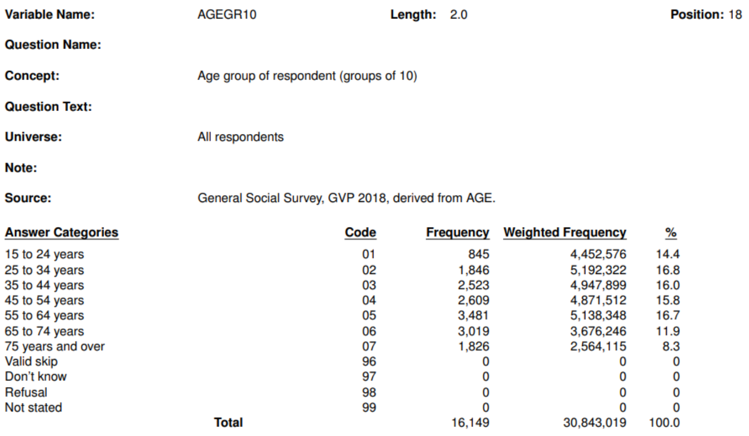


```
Ввод [3]: df.head()
Out[3]:
```

| | PUMFID | WGHT_PER | DH1GAGE | AGEGR10 | GNDR | MARSTAT | HSDSIZEC | PHSDFLG | AGEPRGR6 | CXRFLAG | ... | WTBS_491 | V |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 1 | 2788.1027 | 6 | 7 | 1 | 1 | 2 | 1 | 6 | 2 | ... | 8260.1075 | |
| **1** | 2 | 657.1005 | 6 | 7 | 1 | 1 | 1 | 2 | 96 | 2 | ... | 1599.3012 | |
| **2** | 3 | 1585.1313 | 6 | 7 | 2 | 1 | 2 | 1 | 6 | 2 | ... | 0.0000 | |
| **3** | 4 | 1286.0566 | 6 | 7 | 1 | 1 | 2 | 1 | 6 | 2 | ... | 2357.6728 | |
| **4** | 5 | 1034.9260 | 6 | 7 | 1 | 1 | 2 | 1 | 6 | 2 | ... | 2240.7381 | |

5 rows × 955 columns

Column values were not making sense. So we followed the official document(the link is provided above) with a description to decode the data.



**Variable Name:** AGEGR10 **Length:** 2.0 **Position:** 18

**Question Name:**

**Concept:** Age group of respondent (groups of 10)

**Question Text:**

**Universe:** All respondents

**Note:**

**Source:** General Social Survey, GVP 2018, derived from AGE.

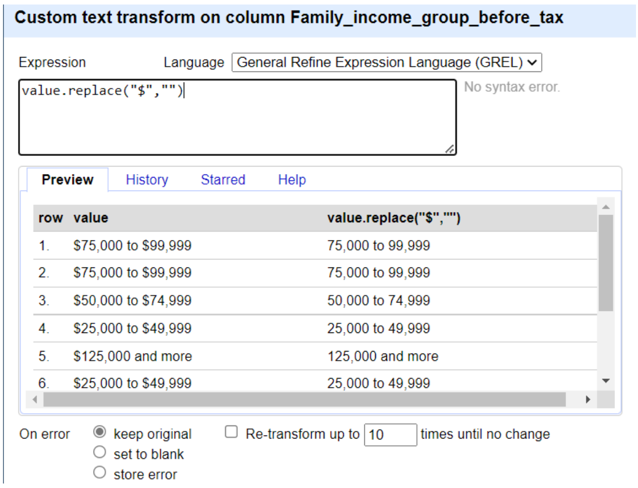
| Answer Categories | | Code | Frequency | Weighted Frequency | % |
|---|---|---|---|---|---|
| 15 to 24 years | | 01 | 845 | 4,452,576 | 14.4 |
| 25 to 34 years | | 02 | 1,846 | 5,192,322 | 16.8 |
| 35 to 44 years | | 03 | 2,523 | 4,947,899 | 16.0 |
| 45 to 54 years | | 04 | 2,609 | 4,871,512 | 15.8 |
| 55 to 64 years | | 05 | 3,481 | 5,138,348 | 16.7 |
| 65 to 74 years | | 06 | 3,019 | 3,676,246 | 11.9 |
| 75 years and over | | 07 | 1,826 | 2,564,115 | 8.3 |
| Valid skip | | 96 | 0 | 0 | 0 |
| Don't know | | 97 | 0 | 0 | 0 |
| Refusal | | 98 | 0 | 0 | 0 |
| Not stated | | 99 | 0 | 0 | 0 |
| | Total | | 16,149 | 30,843,019 | 100.0 |

And from 955 columns we have filtered out 119 columns, with the most interesting data for our further analysis and visualisation, as responded demographics, charity activity preferences and ranking.

Our data initially can be considered as clean, but we had to deal with specific values of survey questions as “Valid skip” and “Not stated”. “Valid skip” can be interpreted as 0 value, while “Not stated” could be interpreted as a missing value - although we weren’t sure and replaced “Not stated” values as “Valid skip”.

Because we cannot delete the row, as the person might have responded to other questions.

If the value in the numeric column is less than 0, it means they skipped the question. We changed the type of 85 columns to numeric in OpenRefine. Because their values were supposed to be numeric, but were loaded as text. We removed the $ sign from Tax and Income columns.
These are some facets to show there are no missing values or abnormal values.



**Custom text transform on column Family_income_group_before_tax**

Expression Language: General Refine Expression Language (GREL)

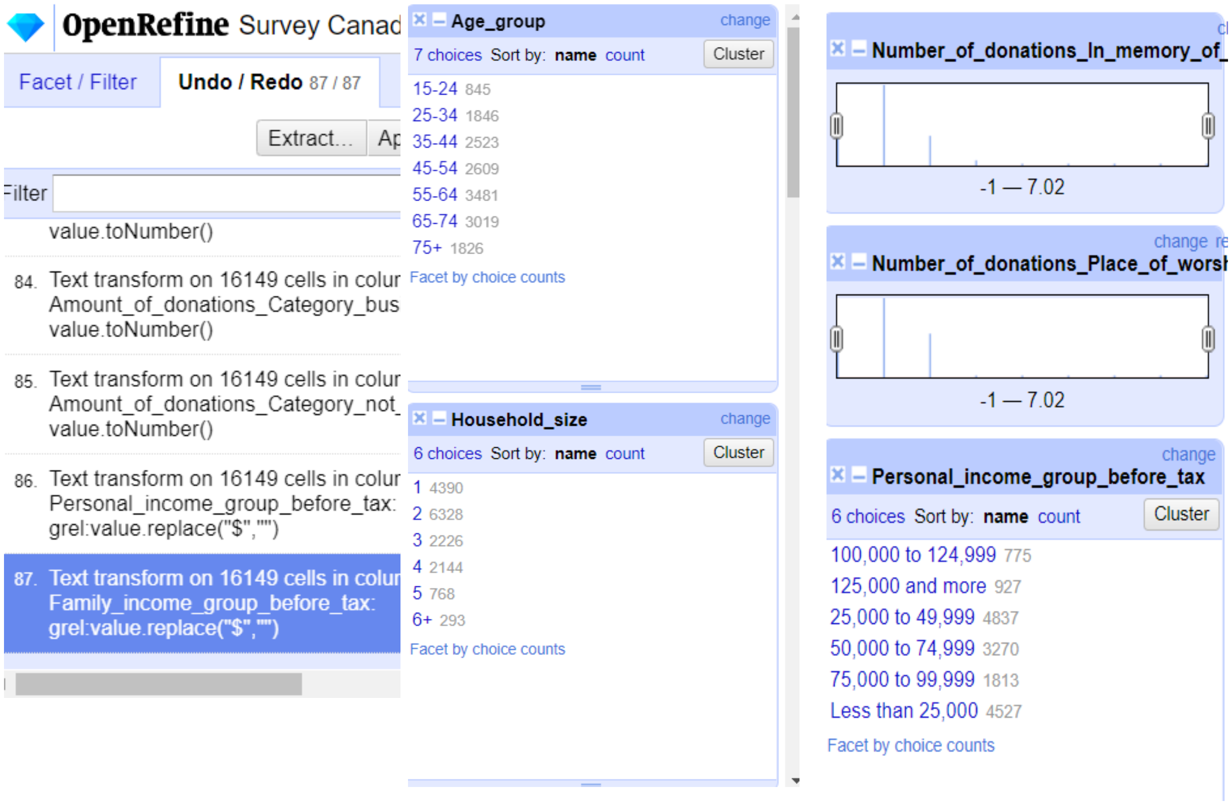
```
value.replace("$","")
```

No syntax error.

| row | value | value.replace("$","") |
|---|---|---|
| 1. | $75,000 to $99,999 | 75,000 to 99,999 |
| 2. | $75,000 to $99,999 | 75,000 to 99,999 |
| 3. | $50,000 to $74,999 | 50,000 to 74,999 |
| 4. | $25,000 to $49,999 | 25,000 to 49,999 |
| 5. | $125,000 and more | 125,000 and more |
| 6. | $25,000 to $49,999 | 25,000 to 49,999 |

On error ● keep original ○ set to blank ○ store error

□ Re-transform up to 10 times until no change



value.toNumber()

84. Text transform on 16149 cells in colur
Amount_of_donations_Category_bus
value.toNumber()

85. Text transform on 16149 cells in colur
Amount_of_donations_Category_not_
value.toNumber()

86. Text transform on 16149 cells in colur
Personal_income_group_before_tax:
grel:value.replace("$","")

87. Text transform on 16149 cells in colur
Family_income_group_before_tax:
grel:value.replace("$","")

**Age_group** change
7 choices Sort by: name count Cluster

| | |
|---|---|
| 15-24 | 845 |
| 25-34 | 1846 |
| 35-44 | 2523 |
| 45-54 | 2609 |
| 55-64 | 3481 |
| 65-74 | 3019 |
| 75+ | 1826 |

Facet by choice counts

**Household_size** change
6 choices Sort by: name count Cluster

| | |
|---|---|
| 1 | 4390 |
| 2 | 6328 |
| 3 | 2226 |
| 4 | 2144 |
| 5 | 768 |
| 6+ | 293 |

Facet by choice counts

**Number_of_donations_In_memory_of_**

-1 — 7.02

change re
**Number_of_donations_Place_of_worsh**

-1 — 7.02

change
**Personal_income_group_before_tax**
6 choices Sort by: name count Cluster

| | |
|---|---|
| 100,000 to 124,999 | 775 |
| 125,000 and more | 927 |
| 25,000 to 49,999 | 4837 |
| 50,000 to 74,999 | 3270 |
| 75,000 to 99,999 | 1813 |
| Less than 25,000 | 4527 |

Facet by choice counts

Link to clean dataset

We decided to use Google Drive storage as our main storage.
The link to the data folder and code notebook.
https://drive.google.com/drive/folders/15L-irXsCi01ZG73wu2ti3S7H3b2wmwEN?usp=sharing

We have 3 notebooks of code, but couldn't merge it. We have shared one notebook with code. Which shows how we decoded the names of 119 columns and values of those columns.