Micro-Datathon #3 Instructions
Data Quality

Data Access, Result submission and communication

Data Access - Data.World

Step 1. Create data.world Account
All data you will need for this micro datathon is available on data.world. Please signup on https://data.world/ to get an account.

Step 2. Finding the project with datasets.

The data.world project Micro-datathon #3 can be accessed here.
(Earlier micro-datathon #1, Geospatial, can be accessed here, and micro-datathon #2, Statistics Canada Census, can be accessed here)

Sharing Results and Recipes

We would like participants to share their Recipes, code, modified datasets and any other information that can help us improve this data as follows:

Step 1. Adding your results summary to the projects insights page

1. We would encourage all participants to add a summary of their results as an insight page on the data.world project's insight page.

2. Once on the Click on the add new insights button to create your own insight page.

3. You can add details and interesting plots here. If you have modified any files provided for this datathon i.e data cleaning and think that the changes will improve the data, please share your improved datasets via data.world and include the link in your insights page.

4. Please name the insight page as your first and last name (first_last) so that we know who has posted the results and have a way of contacting you.

Step 2. Storing related files on DataForGood's NextCloud.

If you plan to share code or any other files, please store them in our Nextcloud. Please send an email to dataops-YYC@dataforgood.ca to get credentials for login. Once you get the login details, please create a

folder with the same name as your insight's page (First_Last name) and store all the files in it. Please make sure you add details about these files on the insight page.

Communication via Slack

For conversations, questions and access requests we recommend using the Data For Good Alberta slack channel **# yyc-micro-datathon**

If you are not yet a member of this Slack workspace please use the link below to join: Data For Good Alberta Invite Link

Recommended Software

There are many tools that can be used to perform data cleaning. A brief list is as follows:

1. Excel (yes, you can use it too!)
2. Trifacta
3. Power Query
4. OpenRefine

At the meetup we will be giving a demo for OpenRefine.

Overview of Data.World Datasets

Cleanup will focus on

- Organization names
- Domains (picklists, keywords) related to Homelessness, etc.
- Addresses

These Data.World datasets and tables have been provided in the Challenge exercises, you can access these from our Micro-Datathon #3 Data Quality project. Pay attention, as you open these datasets, for accompanying data models, definitions, etc.

Micro-datathon #3 Dataset

CHF Agency Program
This is a listing of agencies and programs, sourced from the Calgary Homeless Foundation. It includes agencies/programs that CHF funds or for which they provide IT functionality via the HMIS app.

- Agency_Name
- Program_Name
- Program_Type

Micro-Datathon #1 Dataset Getting to know Geospatial

(see Micro-Datathon #1, for data model)

Organization
This is Data for Good Calgary's Organization table (a work in progress)

- Organization (name), orgid
- Address columns (Tip: Accuracy and Address_Type hint at the accuracy and/or type of address)

Organization_Google_Place_Search_Term

- keyword (i.e. search term)

Organization_Calgary_Foundation_Issue

- Key_Issue, Sub_Key_Issue

Exploring Homelessness focus on Calgary

2013-2020-emergency-shelter-occupancy
A listing of shelter stays from Alberta Open Data

- Organization (name), Shelter (name)

Calgary Charities List from CRA and Alberta nonprofit list

CalgaryQualifiedDonees_GiftAmounts
Canadian Charitable-Donation Recipients (aka Donees)

- Donee_name (Tip: Donee_bn is "unique identifier", can help you group similar names for the same organization)
  000041948RR CAMPUS CALGARY / CITY OF CALGARY PARKS #59 INGLEWOOD BIRD SA
  000041948RR CAMPUS CALGARY/CITY OF CALGARY PARKS #59 INGLEWOOD BIRD SANC

AlbertaGrants
Alberta Open Data on Alberta Grant recipients.
Tip: not all grant recipients have a business number...

- Legal_entity_name

CRA domains (query)

- CRA_charity_type, CRA_category

Alberta Grants domain lists (query)

- BUName, Ministry

Data Cleansing - Train of Thought

For a good example of aligning multiple data sources, that shows sources of data + work done / decisions made / problems encountered, see Paula Jennings' notes on charity list cleanup, using CRA Charity and Alberta Grants data sets.

Recipes

Keeping track of data cleansing steps, tools can do this for you automatically!
See this blog for an example of a recipe:
https://mycreativecontradiction.wpcomstaging.com/2018/09/28/tools-to-make-data-wrangling-faster/

Challenges

Note: Please keep track of your steps, as you do each exercise.

- DQ tools can do this automatically, calling this a recipe.
- You can also track your approach manually, if you aren't using a DQ tool).

Your recipe is a valuable part of your data cleanup (it helps others know and validate your approach, if they choose to use your results.

Basic

Organizations

1. Align the agencies in CHF_Agency_Program with the organizations in Data for Good's Organization table
    a. Choose the 'best' name, so there's 1 name per organization

b. Build a cross-reference between CHF Agency to D4G Organization (1 name per organization, and cross-referencing the 2 tables)
You can look at these Tips, if you'd some help getting started

2. Repeat Simple #1, building a cross-ref between CHF Program Names and D4G Organization Names (where applicable, e.g. Brenda's Cottage)
3. Repeat Simple #1, but now align CHF's Agencies with the Alberta 2013_2020_Emergency_Shelter_Occupancy table's Organizations
4. Repeat Simple #3, for CHF's Programs and the Occupancy table's Shelters

Intermediate

1. Organizations

- Continuing in the vein of the Simple Challenges, but with more-challenging tables, align CHF's Agencies and Programs with
    - Alberta Grants
    - CRA Charity Donees (donee_name; donee_bn may also be helpful)

2. Domain Lists

Domain lists (aka categories, or pick lists) offer various DQ challenges. These include

- Overloading (several concepts, in a single list, e.g. if combining gender and age)
- Incompleteness (gaps between domain values)
- Overlapping (domain values that partially cover the same concept)

Goal - Review the various domains that may identify/categorize homelessness-related keywords, and

- Offer suggestions to improve them
- combine them into 1 or more lists, that are universal across all our tables. Build cross-references to the various pre-existing domain values, if possible.

- CHF Domains
    - Program_type
- Calgary Foundation Domains
    - Key_issue
    - Sub_key_issue
- Google Place Domains
    - Google_Keyword (i.e. search term)
- Alberta Grants Domain Lists

- Ministry
- BU Name

- CRA Domains
  - CRA_category
  - CRA_Charity_type

3. Addresses

Addresses are useful for geospatial work, but can pose several challenges. Three such challenges are

- is the address complete? (missing postal code, city, …) and
- Is it accurate? (the organization is at this address)
- What type of address is it? (e.g. mailing address, billing address, head office, functional site, other). In particular, useful to know if this is where services are provided (functional site)

1. Review addresses in D4G Organization,
   - E.g. address = Edmonton Trail, can you improve this?
   - Verify address is complete and accurate, identify address type, and indicate your confidence (H, M, L) as to the address being complete, accurate, and suitably typed
   - TIP: Accuracy and address_type are populated by Google Place, see API interface, referenced in Micro-Datathon #1, for details

2. CHF's Agency list
   - How many addresses can you provide, how complete are they, and can you identify the type of address? Again, indicate your confidence in the address DQ (H, M, L)

Advanced

Organizations

Repeat the Basic and Intermediate organization exercises, using Data for Good's Organization, instead of the smaller CHF Agency, as a starting point (bigger data sets = bigger challenges, does this cause you to alter your approach?)

Addresses

Some addresses are confidential, e.g. an address for a Women's Emergency Shelter.

- What options do we have for capturing information about this address.
- Is there other information we may want to add, to indicate that this address is publicly unavailable?

- Consider your answer, in terms of impact on reporting, and how different options could skew geospatial reports

Miscellaneous (Free Rein, Exploration)

Review other Data for Good data sets, can you find data quality problems? Can you produce a cleaner data set?

Tips for Simple Challenge #1

A good example of an organization with many spellings of their name (and thus, tricky to work with): SORCe (Safe Communities Opportunity and Resource Centre)

One possible approach to Simple Challenge #1

Input Tables

- CHF_Agency_Program
- Organization

Output Table: Create a table (or spreadsheet) with these column headings

- Clean_Organization_Name
- CHF_Agency_ID
- CHF_Agency_Name
- D4G_Organization_ID
- D4G_Organization_Name
- Comment (optional, can state why you chose the name you could)

Goals

- To have a single list of Organization names, that can be cross-referenced to both Data for Good's organization table and CHF's agency table

Tips

- You can produce an output table, using database tools / data.world, or you can do this work in a spreadsheet
- It's up to you, if you use tools to help you do this work, or simply do this manually
- You can add additional columns, to your table/spreadsheet, if they help you with your cleaning task

****

We welcome you to share your work, following the instructions above, with the rest of the Data for Good group!