



Data for Good and PyData Calgary Jan28 Collaboration on Statistics Canada Data Tables Instructions

Data for Good and PyData Calgary Jan28 Collaboration on Statistics Canada Data Tables Instructions	1
Data Access, Result submission and communication	1
Data Access - Data.World	1
Sharing Results	2
Communication via Slack	3
Recommended Software	3
Reference Documents	4
Challenges	4
Basic	4
Intermediate (or Beginner Python)	5
Intermediate (Power BI based - no Python programming knowledge)	7
Advanced	10

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.

Previous Micro datathon related pages can be found here:

The data.world project Micro-datathon #2 can be accessed [here](#).

- It will probably be good to review the PDF document here called “Canada Census Information Sheet (Data for Good)” because it has background information in it and detail that was referenced but not included in this package.
- It has Micro datathon #2 Challenge Instruction sheet which you might like to review. It is called “Micro-Datathon #2 Instruction” and is a PDF.

Micro-datathon #1, Geospatial, can be accessed [here](#) and may be useful if you want to use shape files and geotopo files to use in your analysis.

Sharing Results

We would like participants to share their 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

This challenge is about accessing StatsCan data with Python. It assumed that if you use Python you are already familiar with package combinations you require for your regular analysis practice.

Power BI does allow us to load and visualize data using Python. If you are not familiar with programming and want to analyze the data in this challenge, then a pbix starter file can be used. This file uses the stats_can package and python reads directly from data.world urls and you can review how Power BI embeds the python script.

If you don't have access to PowerBI or are working on a Mac and are unfamiliar with Python, here are other options that were also mentioned during Micro-datathon #1. You can explore statscan and download data or download the files from data.world.

- Tableau (<https://public.tableau.com/s/>)
- QGIS (<https://qgis.org/en/site/>)
- R (<https://www.r-project.org/about.html>), Rstudio (<https://rstudio.com/products/rstudio/>), [leaflet package](#)
- Python (<https://www.python.org/>), [leaflet and folium](#)
- Jupyter notebooks (<https://jupyter.org/>)

Reference Documents

Helpful reference files can be found in our [Data.World Census Project](#), under project files.

- [Census Information Sheet](#) (assembled by Data for Good YYC, work in progress)
- Census Profile Metadata (assembled by Data for Good YYC, work in progress)

StatCan presentation by IanPreston, at Data for Good Jan 28, 2021 meeting

- Slidepack from Ian (link pending)
- Recording from Data for Good/PyData meeting (link pending)

Challenges

There are different levels of learning objectives in this micro-datathon. Start with the basic tasks and build your skills to the more advanced tasks.

Also feel free to make your own maps, based on the data we've provided (and any other data you want to link in). *Bonus if the visualizations relate to our Fall 2020 Datathon theme of homelessness.*

Basic

These tasks to help you find tables to build on and relate to previous data used in Micro Datathon #2.

- Open the StatCan / CMHC report on [Core Housing](#). Read the report (1-pager). Did you notice the 2016 maple leaf icon on the top right of the page's body? This means it is produced for the Census Profile. The core housing copy in data.world is called Core Housing Need (CSD).xlsx and the "clean" version has the proper column headers).

There are over 9000 Data Tables we can possibly use with the Python Package. The goal is to find some file that will help to get more usable information from Core Housing.

- Use the filter selections and keywords to select a few tables from the main Data interface on this page - <https://www150.statcan.gc.ca/n1/en/type/data?MM=1> . Think about the following:
 - What is the geographic level associated with The Core Housing file that was used and supplied in data.worlds and used in Micro datathon #2. If

were not previously involved in these Challenges - the link to download is and view the dataset is here -

https://www12.statcan.gc.ca/census-recensement/2016/dp-pd/chn-biml/CSD_SDR.csv . Should you try and match this geography level for comparison?

- What subjects would help focus on compatible Data Tables for analysis? Select those?
- What table ids are we looking for if we are using the Python Package. See if you can find the list on the concordance list (<https://www.statcan.gc.ca/eng/developers/concordance>) and then you know it is accessible from the package.
- Go through a similar process to help identify other datasets to pull with the Python StatsCan package and will help enhance analysis for the other data sets that were used in the Micro Datathon #2.
 - *Tables:* 98-400-X2016019 (if you do not use data.world you can find it here <https://www150.statcan.gc.ca/n1/en/catalogue/98-400-X2016019>) (data.world file is this Collective Dwellings 98-400-X2016019.csv)

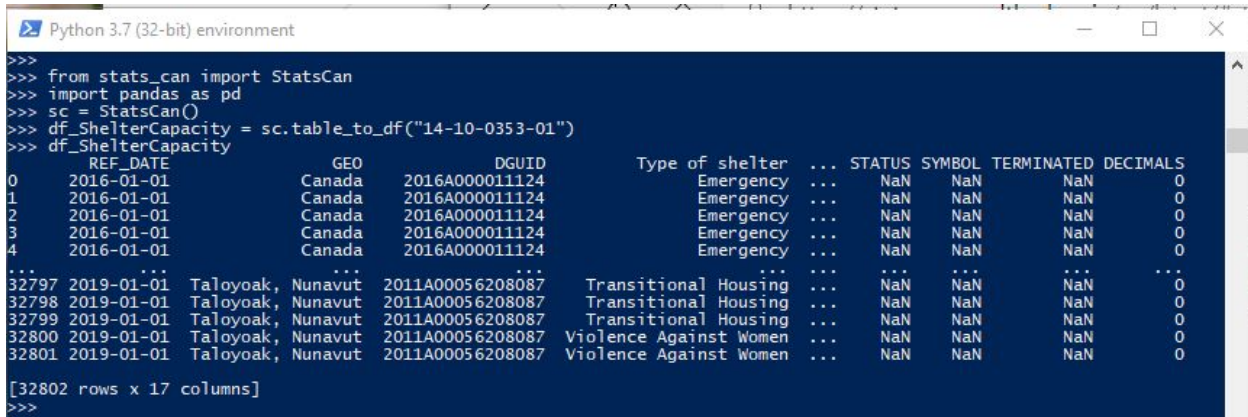
Description: This table presents Type of Collective Dwelling and Collective Dwellings Occupied by Usual Residents and Population in Collective Dwellings of Canada, Provinces and Territories.
 - Here is the third file we used previously and you can download from the url in this citation - Statistics Canada. [Table 14-10-0353-01 Homeless shelter capacity, bed and shelter counts for emergency shelters, transitional housing and violence against women shelters for Canada and provinces, Employment and Social Development Canada annual \(number\)](#) . Or if using data.world, it has this name.
- Do you think Table 14-10-0353-01 can be downloaded with the python package? Choose a vector in this dataset and look it up using the lookup page on StatsCan - <https://www150.statcan.gc.ca/t1/tbl1/en/sbv.action> . Do you need to use the entire Vector string in this case? Can you find how many vectors exist for Calgary Shelters? What is the use of retrieving data by vector?

Intermediate (or Beginner Python)

- If you are familiar with Python and have your IDE setup then with referencing the stats_can web page and using the examples of code below get the data from table [14-10-0353-01](#)

```
from stats_can import StatsCan
import pandas as pd
sc = StatsCan()
df_ShelterCapacity = sc.table_to_df("14-10-0353-01")
```

Using the Python Shell in Power shell on a windows environment thongs should look like this:



```
>>> from stats_can import StatsCan
>>> import pandas as pd
>>> sc = StatsCan()
>>> df_ShelterCapacity = sc.table_to_df("14-10-0353-01")
>>> df_ShelterCapacity
```

	REF_DATE	GEO	DGUID	Type of shelter	STATUS	SYMBOL	TERMINATED	DECIMALS
0	2016-01-01	Canada	2016A000011124	Emergency	...	NaN	NaN	0
1	2016-01-01	Canada	2016A000011124	Emergency	...	NaN	NaN	0
2	2016-01-01	Canada	2016A000011124	Emergency	...	NaN	NaN	0
3	2016-01-01	Canada	2016A000011124	Emergency	...	NaN	NaN	0
4	2016-01-01	Canada	2016A000011124	Emergency	...	NaN	NaN	0
...
32797	2019-01-01	Taloyoak, Nunavut	2011A00056208087	Transitional Housing	...	NaN	NaN	0
32798	2019-01-01	Taloyoak, Nunavut	2011A00056208087	Transitional Housing	...	NaN	NaN	0
32799	2019-01-01	Taloyoak, Nunavut	2011A00056208087	Transitional Housing	...	NaN	NaN	0
32800	2019-01-01	Taloyoak, Nunavut	2011A00056208087	Violence Against Women	...	NaN	NaN	0
32801	2019-01-01	Taloyoak, Nunavut	2011A00056208087	Violence Against Women	...	NaN	NaN	0

```
[32802 rows x 17 columns]
>>>
```

- Read another table using the Python stats_can package. You can read in ones you have found from the previous section or read tables 18-10-0004-12 and 18-10-025-01.
 - How would you cite this data?
 - What data is there for Alberta?
- For the data on data.world we used in Micro Datathon #2 you can read the files another way. To read the excel files you might have to install xlrd and openpyxl Python packages.

```
import pandas as pd
df_collectiveDwellings =
pd.read_csv("https://query.data.world/s/perdspko2k7xbfbkq2opfnu2b5ujyy")
df_coreneed =
pd.read_excel('https://query.data.world/s/soju7zvubclfvipd7elthxuhlrq5q2',
engine='openpyxl')
```

- This is how it should look. You will have to use your favourite packages in Python to explore and plot the data you would like to visualize. Or alternatively use another tool to explore the data and then use Python.

```
Python 3.7 (32-bit) environment
>>> import pandas as pd
>>> df_collectiveDwellings = pd.read_csv("https://query.data.world/s/perdspko2k7xbfbkq2opfnu2b5ujyy")
>>> df_coreneed = pd.read_excel('https://query.data.world/s/soju7zvubclfvipd7elthxuhlrq5q2', engine='openpyxl')
>>> df_collectiveDwellings
  CENSUS_YEAR  GEO_CODE (POR)  ...  Collective dwellings occupied by usual residents  Population in collective dwellings
0         2016         11124  ...                               27780                               685480
1         2016         11124  ...                               15195                               509220
2         2016         11124  ...                               540                               17680
3         2016         11124  ...                               2090                               168205
4         2016         11124  ...                               3045                               171405
...
219        2016          62  ...                               20                               45
220        2016          62  ...                               5                                10
221        2016          62  ...                               0                                0
222        2016          62  ...                               0                                0
223        2016          62  ...                               5                                5
[224 rows x 12 columns]
>>> df_coreneed
      Geographic code  ...  Unsuitable housing not applicable for core need, 2016
0             1001105  ...                               0.0
1             1001113  ...                               0.0
2             1001120  ...                               0.0
3             1001124  ...                               0.0
4             1001126  ...                               0.0
...
4676  Source : Statistique Canada, Recensement de la...  ...          NaN
4677              NaN  ...          NaN
4678  Comment citer : Statistique Canada. 2017. Beso...  ...          NaN
4679  Produit no 98-509-X2016001 au catalogue de Sta...  ...          NaN
4680  http://www12.statcan.gc.ca/census-recensement/...  ...          NaN
[4681 rows x 37 columns]
```

- Can the new table data be combined with any of the datasets from Micro-Datathon #2? Please share any insights, frustrations, visualizations and programming with the community (see [Sharing Results](#))
- If you read and found other tables to integrate or decided to analyze new data related to these subject areas - please cite and share.

Intermediate (Power BI based - no Python programming knowledge)

This section is for people not versed in Python programming to learn how the Python Stats_Can library is used to bring in a data source and gives you the opportunity to concentrate on analyzing data and making visualizations utilizing the tool and adding additional visualizations using Python programming if you wish to do so.

Start by setting up your Python environment to use with Power BI. The latest Power BI is best used with **3.7 Python** because of release differences with the numpy library. Please follow this resource [Run Python Scripts in Power BI Desktop - Power BI | Microsoft Docs](#) . Here is a list of supported runtime and packages if you already have an environment installed and want to troubleshoot - [Learn which Python packages are supported - Power BI | Microsoft Docs](#) .

Download the sample pbix file supplied in this new data.world dataset

<https://data.world/dataforgood/dfgpydata-collaboration-jan28-statistics-canada-data-tables>

called "from Micro2.pbix" . The visualizations in them are just for starters and to show how it is

possible to use the data files with minimal changes to the data. You can check the Transformation (Query) Editor to view what was done.

- Go to Query Editor and view in the advanced editor the way the shelter capacity table is read from data.world.

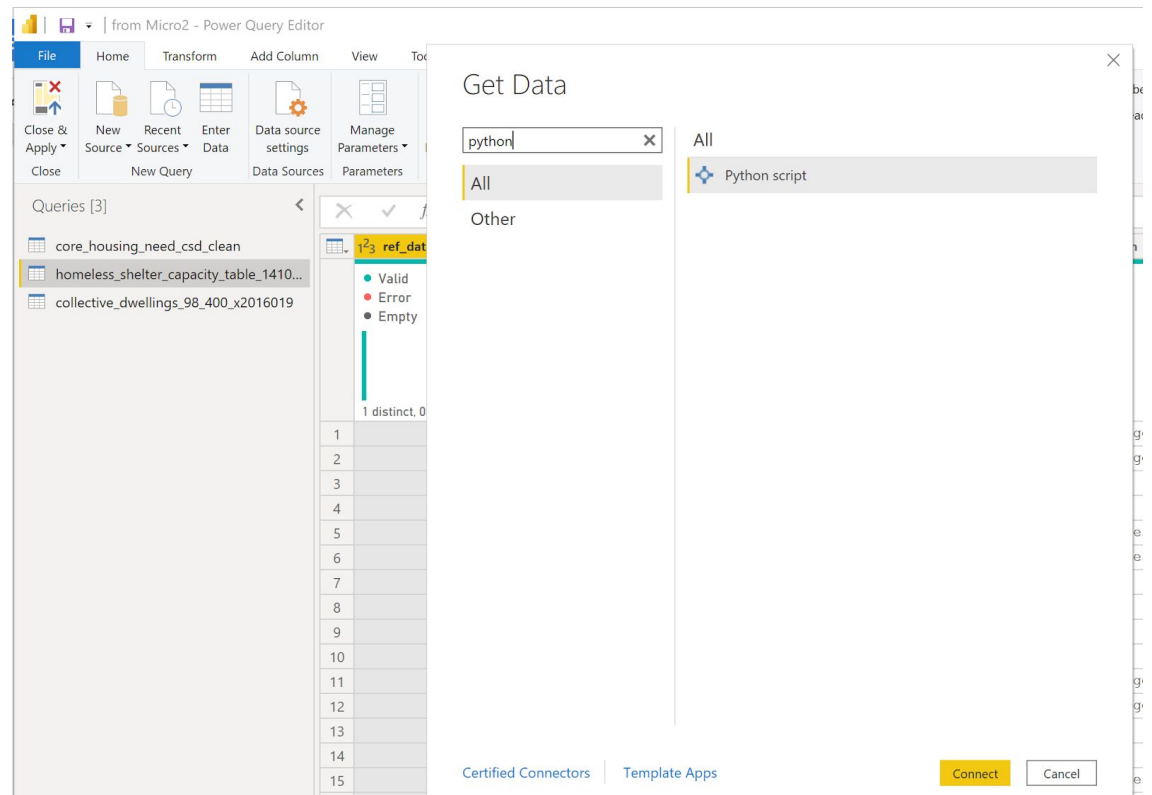
Advanced Editor

homeless_shelter_capacity_table_14100353

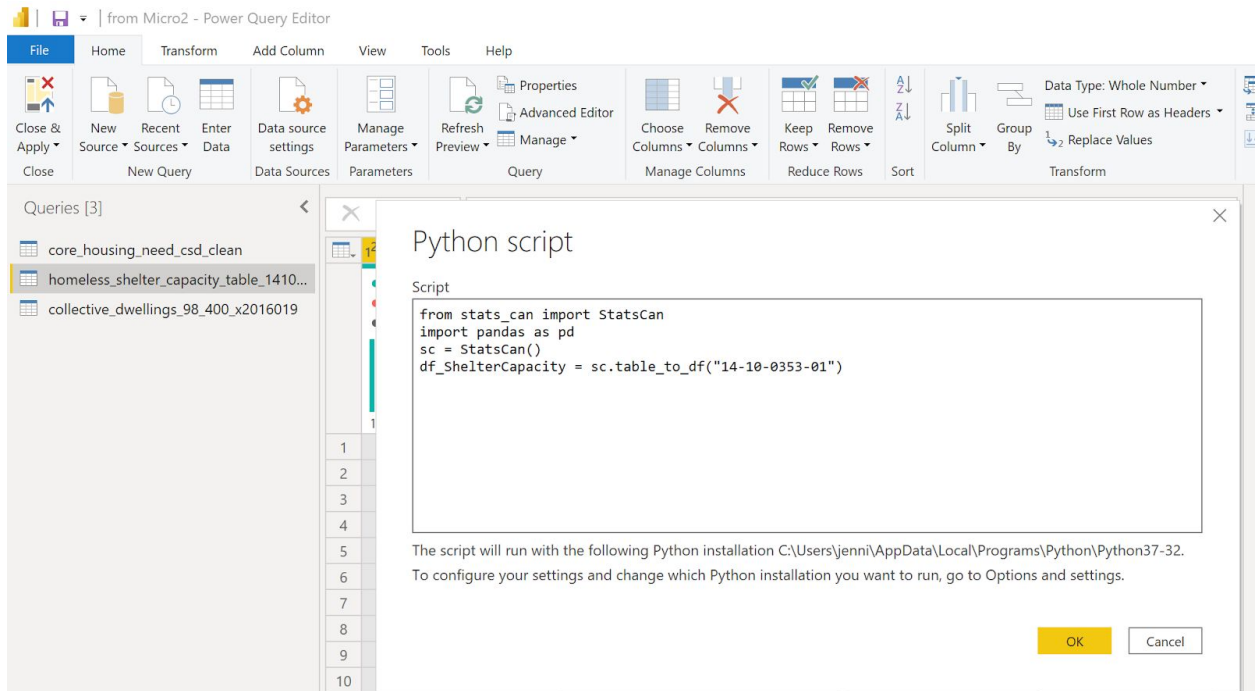
Display Options ?

```
let
    Source = DataWorld.Dataset("dataforgood", "canada-census-2016-homelessness", null),
    homeless_shelter_capacity_table_1 = Source{[tableId="homeless_shelter_capacity_table_14100353"]}[Data],
    #"Filtered Rows2" = Table.SelectRows(homeless_shelter_capacity_table_1, each ([geoid_derived] <> "#VALUE!")),
    #"Changed Type" = Table.TransformColumnTypes(#"Filtered Rows2",{{"geoid_derived", Int64.Type}}),
    #"Filtered Rows" = Table.SelectRows(#"Changed Type", each ([geoid_derived] <> null and [geoid_derived] <> 10 and [geoid_derived] <> 11 and
    #"Removed Columns" = Table.RemoveColumns(#"Filtered Rows",{"dguid", "geoid_derived"})
in
    #"Removed Columns"
```

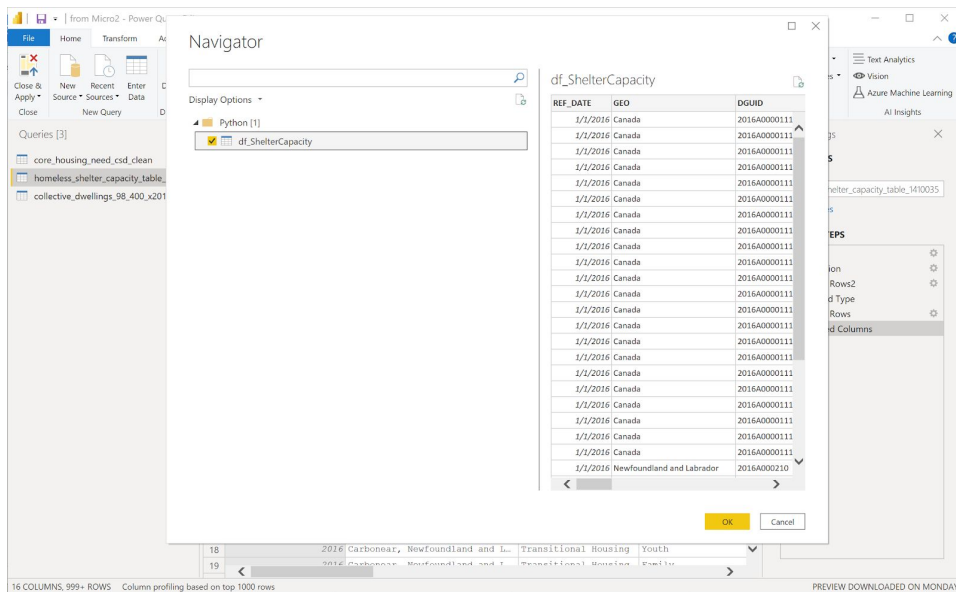
- Now read the same file via Python scripting:
 - Get new Data with Python Script



- Copy the code from above to read the same table directly from StatsCan



- When it has finished connecting, you can select the pandas dataframe



- If you like you can look at the Advanced editor and view how Power Query coded the read from StatsCan using the loaded Python packages (pandas and stats_can).

df_ShelterCapacity

Display Options

```
let
Source = Python.Execute("from stats_can import StatsCan; if import pandas as pd; if sc = StatsCan(); if df_ShelterCapacity = sc.table_to_df("14-10-0353-01")#(1f)",
df_ShelterCapacity1 = Source[["Name=df_ShelterCapacity"]][Value],
#"Changed Type" = Table.TransformColumnTypes(df_ShelterCapacity1,{{"REF_DATE", type date}, {"GEO", type text}, {"DGUID", type text}, {"Type of shelter", type text}, {"Target p
in
#"Changed Type"
```

- Read other tables you are interested in by repeating this step for them or by duplicating and copying the advanced editor script and see how things work.
- Use the Micro-Datathon #2 datasets and the new entries you have read to create dashboard/reports and visualizations.
- Please share any insights you have as per suggestions in the [Sharing Results](#) section.
- If you would like to try accessing a single Vector by selecting a Vector and using the code presented in [Intermediate \(or Beginner Python\)](#) . Is it useful to access Vector information via Power BI? Is it useful to access StatCan tables via Power BI utilizing the stats_can package?

Advanced

- There are vectors present in all the Table Data we read in the previous section. Pick a table to explore some Vectors related to Alberta and Calgary.

```
>>> list(df_ShelterCapacity)
['REF_DATE', 'GEO', 'DGUID', 'Type of shelter', 'Target population', 'Statistics', 'UOM', 'UOM_ID', 'SCALAR_FACTOR', 'SCALAR_ID', 'VECTOR', 'COORDINATE', 'VALUE', 'STATUS', 'SYMBOL', 'TERMINATED', 'DECIMALS']
>>> list(sc.table_to_df("18-10-0004-12"))
['REF_DATE', 'GEO', 'DGUID', 'Products and product groups', 'UOM', 'UOM_ID', 'SCALAR_FACTOR', 'SCALAR_ID', 'VECTOR', 'COORDINATE', 'VALUE', 'STATUS', 'SYMBOL', 'TERMINATED', 'DECIMALS']
>>> list(sc.table_to_df("181002501"))
['REF_DATE', 'GEO', 'DGUID', 'Client groups', 'UOM', 'UOM_ID', 'SCALAR_FACTOR', 'SCALAR_ID', 'VECTOR', 'COORDINATE', 'VALUE', 'STATUS', 'SYMBOL', 'TERMINATED', 'DECIMALS']
```

- What do the Vectors you chose represent?
- Use the Python library to retrieve a vector. As an example I have chosen "V113749287" and "V113749288".

```
>>> sc.vectors_to_df(["V113749287", "V113749288"])
      v113749287 v113749288
REF_DATE
2016-01-01      1.0      60.0
2017-01-01      1.0      60.0
2018-01-01      2.0     170.0
2019-01-01      2.0     180.0
>>>
```

- How would you use this output?
- Can you visualize data at a Census Tract level and gain insight? You can utilize and integrate the following bits of data:
 - The Core housing data also comes at a census tract granularity
https://www12.statcan.gc.ca/census-recensement/2016/dp-pd/chn-biml/CT_SR.csv
 - The shape files and topojson for Census Tracts can be found on data.world. The shape for reading with Python can be downloaded and unzipped and the topojson (.json file) can be used as a shape in Power BI. Remember census tract names look like floating point numbers and make sure they are read and stay defined as text or character. The “.00” and “.10” and endings like those must remain or else your data will not map.



The screenshot shows the data.world website interface. At the top, there's a search bar and the data.world logo. Below that, the main heading is 'Calgary Spatial Files (Alberta and Edmonton)' with the subtitle 'DATASET IN DATAFORGOOD'. There are tabs for 'Overview', 'Access' (with a red badge showing '1'), 'Discussion', 'Activity', and 'Settings'. Under the 'Overview' tab, there are two dataset entries:

- Alberta_CT2016_CensusTracts.json**: A topojson version for Power BI of Alberta_CT2016_Cens... Edit
- Alberta_CT2016_CensusTracts_WGS84.zip**: clean data Subset from lct_000a16a_e.zip RAW file an... Edit

-
- Read the table below to try and gain any insight. Or select another table you think is helpful to relate to core housing need, is at the census tract level and over Calgary.

Filter results by ⁱ

Remove "Housing" filter✕

Remove "Census tract" filter✕

Clear all

Sort by date ⁱ Apply

Show 10 entries Apply

- [Statistics Canada's Trust Centre](#)
- [How to access microdata](#)
- [Customized products and services](#)

Keyword(s)

Search title, description, etc.



Subject

☒ Housing (4)

- ☐ Dwelling characteristics (1)
- ☐ Housing and living arrangements (1)
- ☐ Housing costs and affordability (1)
- ☐ Other content related to Housing (1)

Geography

☒ Census tract (4)

- ☐ Census metropolitan area (3)
- ☐ Census agglomeration (3)
- ☐ Canada (2)
- ☐ Province or territory (2)
- ☐ Census division (2)

All (4)	Tables (3)	Profiles of a community or region (1)	Thematic maps (0)	Public use microdata (0)	Data Visualization (0)
---------	------------	---------------------------------------	-------------------	--------------------------	------------------------

1. [Income divergence index \(D-index\) by census tract](#)

Table: 11-10-0074-01

Geography: Census tract

Frequency: Occasional

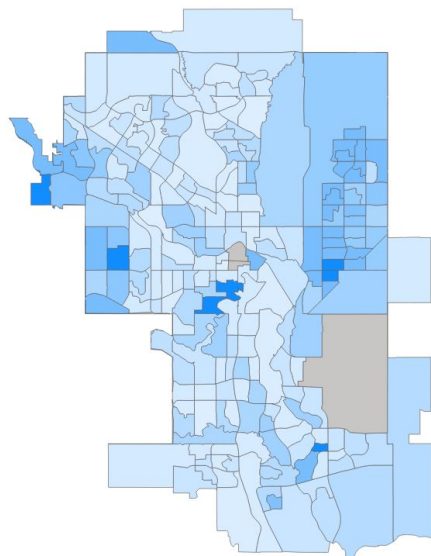
Description: The divergence index (D-index) describes the degree that families with different income levels are mixing together in neighbourhoods. It compares neighbourhood (census tract, CT)...

[More](#)

Release date: 2020-06-22

- Subset all data to Calgary and see if you can tell a story making whatever data connections you choose. Please improve upon this display of just the Divergence Index for 2017. Please share your results according to the suggestions here [Sharing Results](#) .

Divergence Index by CensusTractID





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