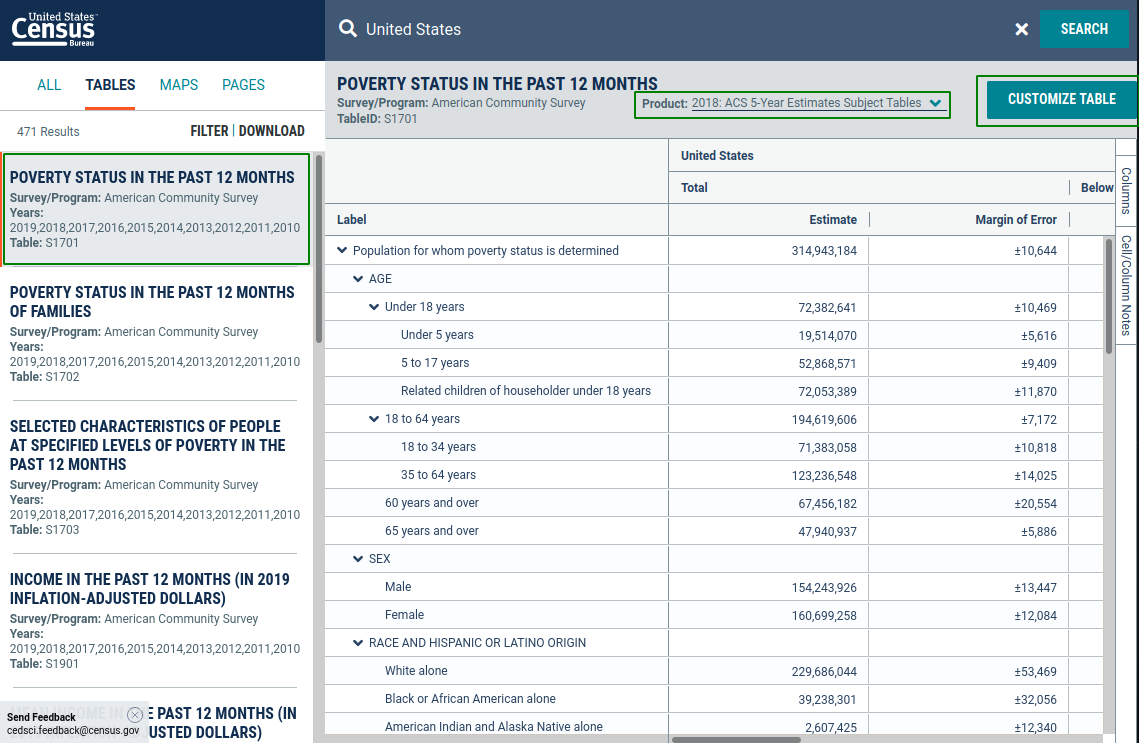
Census Data Gathering & Preparation

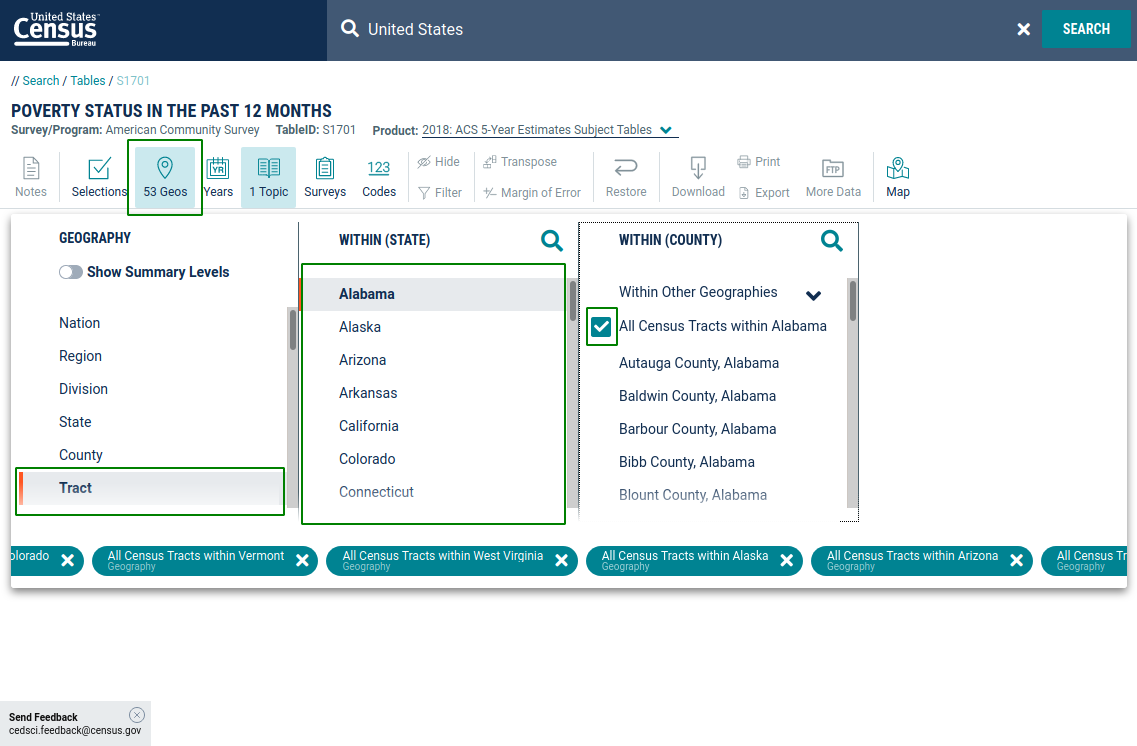
**Census Estimates Data**

The Census estimates data contains the bulk of information on which clustering will be performed. These are the Census 5-year American Community Survey Responses for 2018.

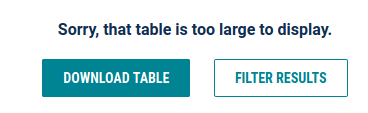
The census estimates data were found through tables prepared on data.census.gov. Tables prepared by the Census Bureau are selected on the left side of the “View Tables” interface. The product is set to the “2018: ACS 5-Year Estimates Data Profiles”. “CUSTOMIZE TABLE” is then selected:



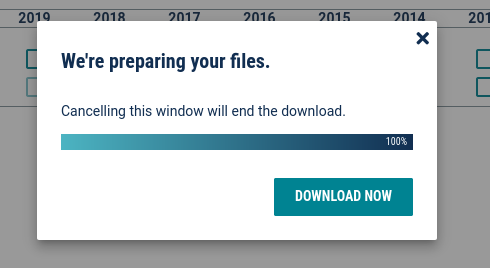
Next all “Tract” geographies under the “Geos” menu are selected for all states:



When all tracts are selected, the data will be too large to display, and an option to download the table will appear:



The download button sends a request to the server that collects all of the selected data and packages it into a file for download:



The zip can then be downloaded, and the csv can be extracted for consumption.

This selection process can be tedious if multiple tables are required.

The URL for this example is:

https://data.census.gov/cedsci/table?q=United%20States&g=0400000US01.140000,02.140000,04.140000,05.140000,06.140000,08.140000,09.140000,10.140000,11.140000,12.140000,13.140000,15.140000,16.140000,17.140000,18.140000,19.140000,20.140000,21.140000,22.140000,23.140000,24.140000,25.140000,26.140000,27.140000,28.140000,29.140000,30.140000,31.140000,32.140000,33.140000,34.140000,35.140000,36.140000,37.140000,38.140000,39.140000,40.140000,41.140000,42.140000,44.140000,45.140000,46.140000,47.140000,48.140000,49.140000,50.140000,51.140000,53.140000,54.140000,55.140000,56.140000,72.140000&tid=ACSST5Y2018.S1701&hidePreview=true

It was noted that:

* The “**g**” parameter in the above URL is sufficient to specify all of the census tract selections that were made manually in the first iteration. This can be leveraged to reduce time downloading data from additional tables.
* The table ID parameter “**tid**” could also be leveraged to substitute the ID for the ID of other tables of interest.

In total, eight Census tables were identified that contained data relevant to our project that contained potentially 1313 data fields for all census tracts that can be used for the clustering algorithm. Note that additional data will be added via amenities data that are obtained separately.

The table IDs that were utilized are:

|  |  |
| --- | --- |
| **tid** | **Description** |
| ACSST5Y2018.S0101 | Age & Sex |
| ACSST5Y2018.S1501 | Educational Attainment |
| ACSST5Y2018.S2301 | Employment Status |
| ACSST5Y2018.S1101 | Households and Families |
| ACSST5Y2018.S1701 | Income and Poverty |
| ACSST5Y2018.S2405 | Industry by Occupation |
| ACSDT5Y2018.B02001 | Race |
| ACSDP5Y2018.DP05 | ACS Demographic and Housing Estimates |

One can use the above URL and replace the tid parameter with these to download all of the same data.

The total compressed storage size of these are ~300 MB (~1,200 MB uncompressed)

All of the downloads are unzipped to their own folders. The code in the notebook “01\_process\_data.census.gov\_downloads\_\_v4.ipynb” is then used to read and process this data into a consumable format. Data fields were cleaned and converted to numeric where possible. Fields consisting fully of non-numeric information were dropped. See the notebook for additional details on the transformations that were performed.

Tables that are output for the data.census.gov data are “2018\_5yr\_cendatagov\_ESTIMATES\_v4.pkl”, which contains the numeric census estimate response data, and “2018\_5yr\_cendatagov\_ESTIMATES\_DD\_v4.pkl”, a data dictionary to map the column name codes to their descriptions. These are pickled pandas dataframes.

Version 4 selects a narrow range of fields than earlier versions that were found through feature selection of the earlier data. Essentially, columns that were very sparse or deemed too specific of a census category were considered to keep or drop. The columns in the v4 estimates table reflect the results of this selection.

Additional cleaning steps are highlighted in Census\_Data\_Cleanup\_Process.vsdx.

**Census Gazetteer File**

The 2018 Census Gazetteer file contains geography data necessary for this project. It links each census tract geoID to the tracts center latitude and longitude as well as estimates for area of land and water within each tract.

The direct download link to this file is https://www2.census.gov/geo/docs/maps-data/data/gazetteer/2018\_Gazetteer/2018\_Gaz\_tracts\_national.zip

Additional information can be found here: https://www.census.gov/geographies/reference-files/time-series/geo/gazetteer-files.2018.html

The Gazetteer data is prepared alongside the data.census.gov data in “01\_process\_data.census.gov\_downloads\_\_v4”. The prepared data is output to a pickle file (pandas dataframe) named “2018\_5yr\_cendatagov\_GAZ\_v4.pkl”. This combined with the other Census tables were too large to store in Github and were distributed to the team via SharePoint storage in Microsoft Teams. This table can be reproduced as needed given the links and the script above.

**Census Geometry Data (for Mapping)**

The geometries (e.g. shape files) for census tracts can be found at the following link: <https://github.com/loganpowell/census-geojson/tree/master/GeoJSON/500k/2018>. Each folder at that link represents a state, so all needed to be downloaded.

To download all of these, the following command was used:

svn checkout <https://github.com/loganpowell/census-geojson/trunk/GeoJSON/500k/2018>

These were prepared into a geoJSON format via code in the notebook “prepare\_census\_tract\_geojson.ipynb”.

The raw data that were used in this notebook from that repo are provided in /raw\_data/census\_tract\_geometries/2018/

The output of this preparation is a single, gzip compressed, utf-8 encoded, geoJSON file “all\_census\_tract\_shapes.json.gz”.