Northeastern University
DS2001 – Programming with Data — Social Science Practicum
Fall 2019, Mattsson & McCabe

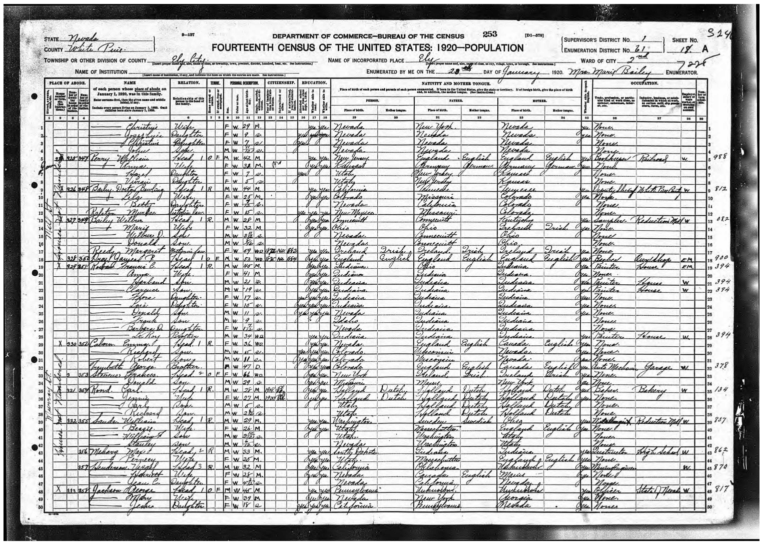
Practicum 3: Big Data

In this week's practicum we will be asking Python to process administrative data for us. Specifically, we will count how many persons of different demographic groups live in our little town and compute Labor Force Participation Rates. We will be working with a manageable-sized subset of data, but keep an eye out for pieces of code that would be extensible to larger datasets.

1 Demographics and Tax Collection in Pahrump, Nevada 1920

Data collection for public administration is, one could say, a progenitor of Data Science. Many of the basic tasks of public administration, such as counting people and calculating taxes, involve collecting and summarizing data. Collecting accurate administrative data is an important component of conducting accurate analyses, which allow the powers that be to make informed decisions. Censuses are a specific kind of administrative data collection, where every effort is made to collect the same basic information from every resident of a polity. Censuses have been conducted since ancient times, notably in Babylon, Persia, and Rome.

The US Census is conducted every 10 years and is used in many aspects of public administration. By law, US Census individual records (ie. 'microdata') are kept behind lock and key for 72 years. After this they become available for demographic, genealogical, and other study. We will use 1920 US Census data to characterize the population of Pahrump, Nevada at that point in time.


STATE Nevada
COUNTY White Pine
9–137
DEPARTMENT OF COMMERCE—BUREAU OF THE CENSUS 253 [D1–578]
FOURTEENTH CENSUS OF THE UNITED STATES: 1920—POPULATION
SUPERVISOR'S DISTRICT NO. 1
ENUMERATION DISTRICT NO. 61
SHEET NO. 18 A
TOWNSHIP OR OTHER DIVISION OF COUNTY Ely City
NAME OF INCORPORATED PLACE Ely
WARD OF CITY 2nd
NAME OF INSTITUTION
ENUMERATED BY ME ON THE 28th DAY OF January 1920. Mrs. Marie Bailey ENUMERATOR.

1.1 Download

Please download `pr03.zip` from the practicum website (or Blackboard), unzip it, and move the directory/folder to where you want it in your file system. Use Atom's `File > "Add Project Folder..."` to open this folder and view `census.py`.

1.2 View 1920 US Census Data

First, we must load the dataset. You should see `from data import data_header, data_as_list_of_lists` at the top of your file. This tells `census.py` to load the custom functions from `data.py` named `data_header` and `data_as_list_of_lists`. These functions take as an argument the name of the file that you want to load. Please use `‘nvcensus-1920-pahrump.csv’`.

You should print out the dataset and confirm that it looks like census data. Use the table below to make any notes you think you might need about the dataset, such as what the options are in each column.

Table 1: 1920 US Census, Pahrump, Nevada

Year	**Subunit**	**FName**	**Surname**	**Age**	**Sex**	**Color**	**Profession**
1920	PAHRUMP PRECINCT	FRANK L.	CATE	44	M	W	MINER

Looping through lists In Python, the main use for lists is to loop through them. This means running the same snippet of code for every element in the given list, and this is useful for all sorts of things. You can print, count, subset, or even edit each entry in the same way using a very short and concise piece of code. The syntax you would use is called a “for loop” and it goes like this:

```
for ELEMENT in LIST:
    some_function_of(ELEMENT)
```

1.3 Population

We would like to know the population of Pahrump, NV in 1920 (ie. the number of records). Please print.

1.4 Birth of a Child

Congratulations Pahrump! `WILLIE` and `LUCY GOODWIN` just had a daughter! The town would like to keep their records updated, so please add little `ANNIE` to the list and confirm that the population rose by `1`.

1.5 Labor Force Participation

Along with the total population, we would also like to know the Labor Force Participation Rate. To calculate this, please find the working population, which is the number of residents who have a listed `Profession` (ie. one that is not `'NONE'`), and divide by the total population.

Please have your code print this number. To check your code, make sure this number makes sense from your understanding of the dataset.

Indexing into lists To get the `Profession` from the record of one of the respondents, you will need to figure out what index in the `record` (a list) this corresponds to. This is why it is useful to save/load/have also the `header` to any dataset you are using as a list. Indexing into `header[i]` gives you the column label for the value at `record[i]`.

1.6 Segmenting the Population

We would also like to know *who* is working. Please create four new datasets (ie. lists) for the following population subsets, and calculate the Labor Force Participation Rate for each.

Population subsets:

- Children – residents under the `Age` of `18`
- Elderly – residents who are `75` years of `Age` or older
- Women – adult residents whose `Sex` is `F`
- Men – adult residents whose `Sex` is `M`

Do these numbers make sense? Why or why not? Please print these numbers, and write your explanation as a comment in your code.

1.7 Submit Your Work!

Please submit your work to Blackboard.

If you used one file, you can upload it as-is.

If you used more, please first make a `zip` file of the folder within which you did your work.

- On MacOS, select your work folder, right-click and choose `Compress pr03`
- On Windows, select your work folder, right-click, `Send To`, `Compressed (zipped) folder`

1.8 Citations

James, Ronald M.; Fliess, Kenneth H.; Nystrom, Eric C., 2014, ”Nevada Census Microdata, 1860-1920”, https://doi.org/10.7910/DVN/27218, Harvard Dataverse, V2

`https://www.newworldencyclopedia.org/entry/Census`

`https://www.wikitree.com/photo/jpg/1920_United_States_Federal_Census_Nevada_White_Pine_County_Ely_Disrict_0061_p_34`

`https://twitter.com/PahrumpNV`

`https://taxfoundation.org/us-federal-individual-income-tax-rates-history-1913-2013-nominal-and-inflation-adjusted-brackets/`

2 Done Early?

2.1 Estimate of Tax Receipts

One of the age-old uses of the census is to estimate the tax base of the polity in questions. We would like to estimate the taxes that the town of Pahrump, NV could collect in 1920. See the table below for the expected contribution from each of the professions listed. Print your answer.

Profession	Salary	Tax Receipts
MINER	5000	240
FARMER	12000	1150
FARM LABORER	2600	104
OTHER PROFESSION	3200	128
NO PROFESSION	0	0

Is your code correct? How would you check? Would it be better to overestimate or underestimate?

2.2 Beyond Pahrump, NV

Within `data.py` there are code comments that would let you produce the corresponding dataset for any of Nevada’s precincts. See if you can edit this code to do this exercise for a larger precinct.

This handout was originally created by Carolina Mattsson and Stefan Mccabe, Fall 2019.