Spatial Data Science MOOC

Exercise

Perform data engineering tasks

Section 1 Exercise 1

02/2020

esri

THE SCIENCE OF WHERE™

Perform data engineering tasks

Time to complete
90 minutes

Introduction

Data engineering is a fundamental part of every analysis. The term refers to the planning, preparation, and processing of data to make it more useful for analysis. It can include simple tasks like identifying and correcting imperfections in your data and calculating new fields. It can also include more complex tasks like reducing the dimensions of a multivariate dataset.

Data engineering also involves the process of geoenriching your data. Geoenrichment can include various tasks:

- Adding a spatial location to your data, referred to as geocoding
- Using other data sources to extract information and add, or enrich, these values to your dataset
- Calculating new fields that represent spatial characteristics, like the distance from a particular feature in a landscape

In this exercise, you will use ArcGIS Pro and ArcGIS Notebooks to perform data engineering tasks. These tasks will use the built-in tools available with these products as well as tools available by integrating open source libraries.

Exercise scenario

Because voting is voluntary in the United States, the level of voter participation (referred to as "voter turnout") has a significant impact on the election results and resulting public policy.

Modeling voter turnout, and understanding where low turnout is prevalent, can inform outreach efforts to increase voter participation. With the ultimate goal of predicting voter turnout, this exercise will focus on performing various data engineering tasks to prepare election result data for predictive analysis.

Step 1: Download the exercise data files

In this step, you will download the exercise data files.

a Open a new web browser tab or window.

b Go to https://bit.ly/2tNazj0 and download the exercise data ZIP file.

Note: The complete URL to the exercise data file is https://www.arcgis.com/home/item.html?id=1a3b235d44734d5d8d6ec756f26b38e3.

c Extract the files to a folder on your local computer, saving them in a location that you will remember.

Step 2: Confirm that your computer can run ArcGIS Pro

In this step, you will run a test to confirm that your computer can support ArcGIS Pro. Even if you have ArcGIS Pro installed, you should confirm that it can support ArcGIS Pro 2.5.

a Go to the system check link.

b Click the Can You Run It? button.

c Follow the steps to open and run the test.

The site generates a report that lists the minimum requirements and identifies if your machine meets these requirements.

d If your computer does not meet these requirements, use the provided links to complete the recommended updates, and then run the test again.

e If your computer meets the requirements, save the report.

Note: If your computer does not meet the requirements, you may need to use a different computer or update your graphics card. For more information about graphic card requirements, see ArcGIS Pro Help: ArcGIS Pro 2.5 system requirements.

The MOOC team may ask you to share the report if you need help in later ArcGIS Pro exercises.

Step 3: Install ArcGIS Pro

This MOOC uses ArcGIS Pro 2.5. If you already have ArcGIS Pro 2.5 installed, proceed to the next step. If you do not have ArcGIS Pro 2.5, complete this step.

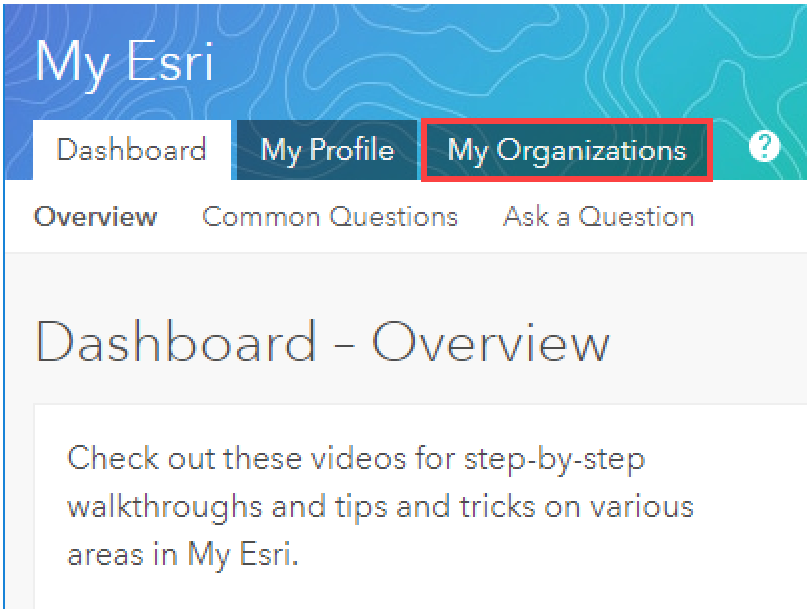
a Open a web browser and go to https://my.esri.com.

The My Esri page manages your account information, including access to software downloads.

b If necessary, sign in using your ArcGIS account.

Your ArcGIS account is the user name and password that you used to sign in to the Esri Training site and access the MOOC.

c At the top of the page, click the My Organizations tab.



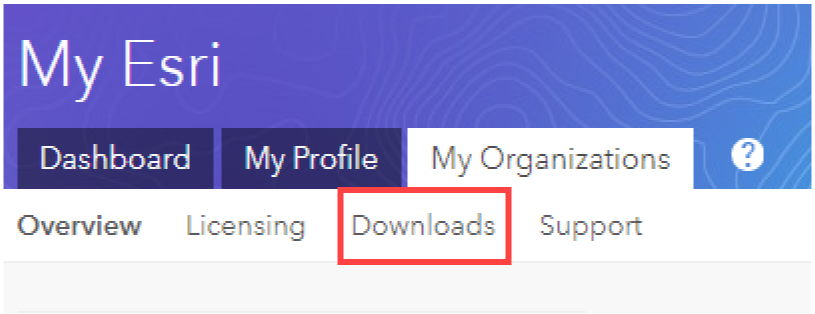
Dashboard – Overview

Check out these videos for step-by-step walkthroughs and tips and tricks on various areas in My Esri.

d Under My Esri / My Organizations / Overview, confirm that MOOC Program is the listed organization.

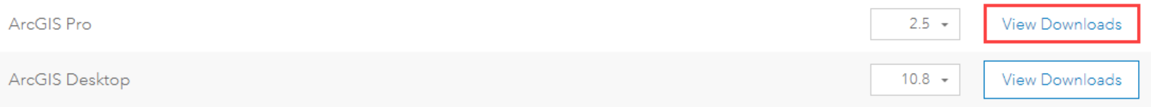
Note: If you do not see Downloads, click the My Profile tab, and then under Connected Organizations, select MOOC Program.

e From the My Organizations tab, click Downloads.



Note: If you do not see Downloads, you may be signed in to My Esri with the wrong account. Sign out of the site and sign back in with your ArcGIS account.

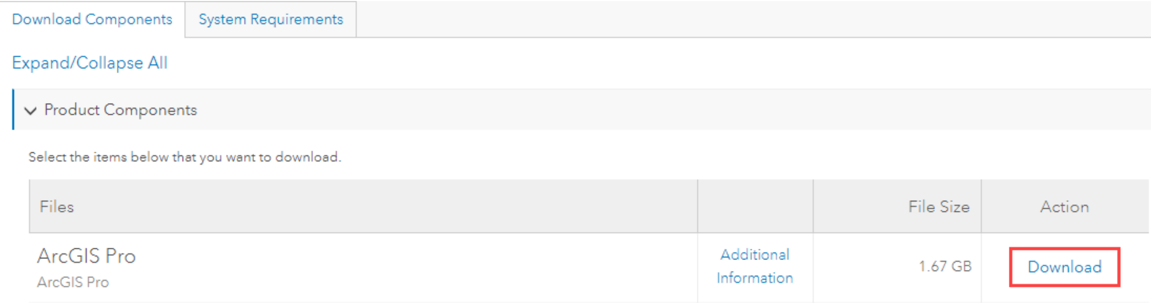
f For ArcGIS Pro 2.5, click View Downloads.



ArcGIS Pro	2.5	View Downloads
ArcGIS Desktop	10.8	View Downloads

Note: You can run ArcGIS Pro in a different language by installing a language pack. Keep in mind that this course is taught in English, which means that all screen shots and exercises will use the English version of ArcGIS Pro.

g From the Download Components tab, to the right of ArcGIS Pro, click Download.



Select the items below that you want to download.

Files		File Size	Action
ArcGIS Pro ArcGIS Pro	Additional Information	1.67 GB	Download

If the default download location does not have enough space, you can change the location by following the steps in this link.

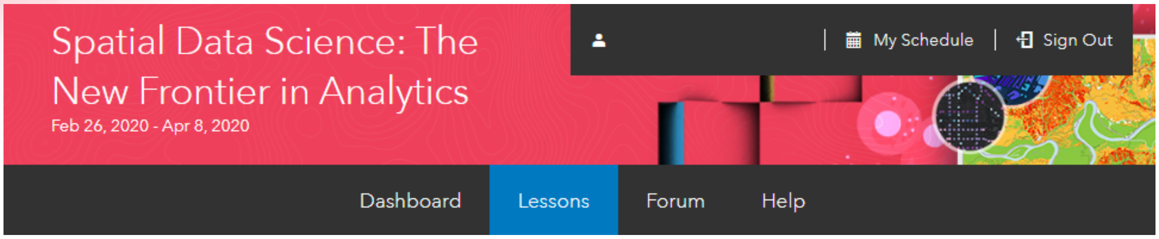
h After the download completes, double-click the .exe file.

i Follow the installation instructions, accepting all defaults.

Step 4: Sign in to ArcGIS Pro

You used your own ArcGIS account to install ArcGIS Pro in the previous step. The MOOC provides a separate ArcGIS account (user name and password) that you will need to use to license ArcGIS Pro and access other software applications used throughout the MOOC exercises.

a Near the top of this page, under Spatial Data Science: The New Frontier in Analytics, click Lessons.



Spatial Data Science: The New Frontier in Analytics

Feb 26, 2020 - Apr 8, 2020

The course ArcGIS account user name and password are listed under Lessons. The user name for this account ends with _sds (for example, jdoe_sds). You may want to write down the user name and password for quick reference.

Note: If you registered in the last few hours, your account may not be ready. Refresh the page in an hour or so to determine if your account is available.

b If necessary, start ArcGIS Pro.

c Sign in using the provided ArcGIS account.

Step 5: Open an ArcGIS Pro project

a In ArcGIS Pro, under Open, click Open Another Project.

b In the Open Project dialog box, browse to the Data Engineering and Visualization folder that you saved on your computer.

c Click Data Engineering and Visualization.aprx.

d Click OK.

Your ArcGIS Pro project opens to a gray reference map, called a basemap. You can zoom and pan this map to different areas of the world. Because you are preparing United States election data, it is currently focused on the contiguous United States.

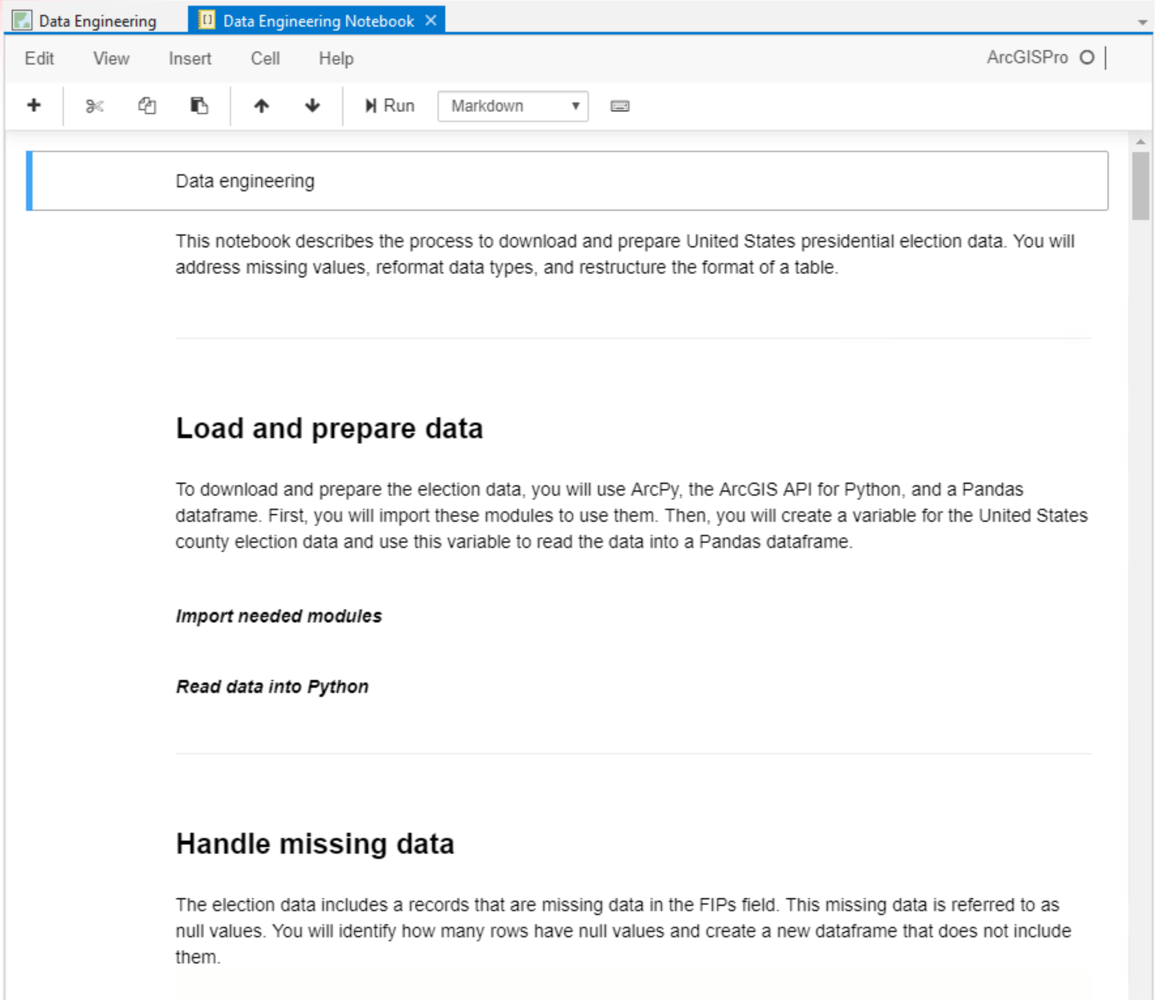
To the left of the map is the Contents pane; the Contents pane lists the layers that have been added to the map. To the right of the map is the Catalog pane; the Catalog pane lists the items associated with this ArcGIS Pro package—Maps, Toolboxes, Notebooks, Databases, Styles, Folders, and Locations. To learn more about the ArcGIS Pro interface, see ArcGIS Pro Help: ArcGIS Pro user interface, and to learn more about ArcGIS Pro projects, see ArcGIS Pro Help: Projects in ArcGIS Pro.

Step 6: Open an ArcGIS Notebook

This exercise uses ArcGIS Notebooks in ArcGIS Pro. ArcGIS Notebooks are built from Jupyter Notebooks, which structure content using cells. Cells can contain executable Python code (code cells) or explanatory text and media (markdown cells). In this step, you will open the ArcGIS Notebook used in this exercise.

a In the Catalog pane, expand Notebooks.

b Right-click Data Engineering Notebook.ipynb and choose Open Notebook.



Data engineering

This notebook describes the process to download and prepare United States presidential election data. You will address missing values, reformat data types, and restructure the format of a table.

Load and prepare data

To download and prepare the election data, you will use ArcPy, the ArcGIS API for Python, and a Pandas dataframe. First, you will import these modules to use them. Then, you will create a variable for the United States county election data and use this variable to read the data into a Pandas dataframe.

Import needed modules

Read data into Python

Handle missing data

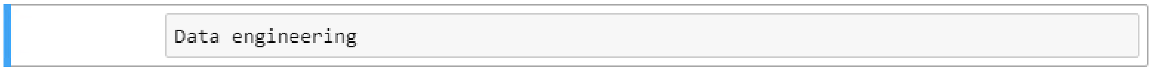
The election data includes a records that are missing data in the FIPs field. This missing data is referred to as null values. You will identify how many rows have null values and create a new dataframe that does not include them.

A notebook will open in the ArcGIS Pro project. The first few cells in this notebook are markdown cells used to explain the exercise.

Step 7: Modify a markdown cell

You will use this notebook to complete most of the exercise. In this step, you will learn how to use the markdown cells in the notebook.

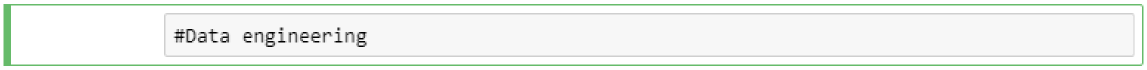
a In the notebook, double-click the first markdown cell titled Data Engineering.



```
Data engineering
```

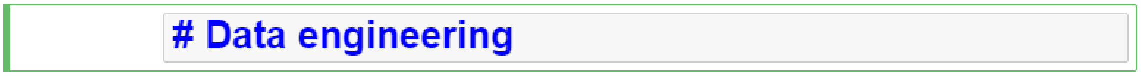
Markdown cells use hashtags to determine the size and format of the explanatory text.

b In front of Data Engineering, type a hashtag (**#**).



```
#Data engineering
```

c Add a space between the hashtag and the word Data Engineering.



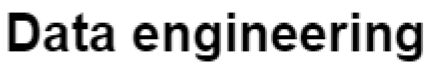
```
# Data engineering
```

The text font style and size change to make it appear more like a heading.

Note: Adding additional hashtags will decrease the size of the font. If you are familiar with HTML, you can think of this as switching between header tags (<h1>, <h2>, <h3>). Be sure to maintain a space between the hashtag and your text; otherwise, the font style and size will appear as regular text.

d From the ArcGIS Notebook Toolbar, click the Run button ⏭.

Note: Alternatively, you can select the cell and press Shift + Enter on your keyboard.



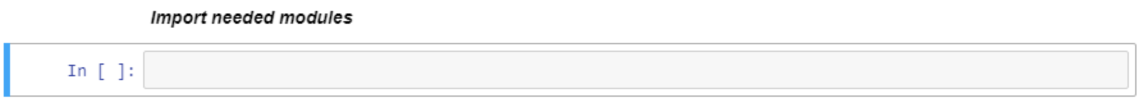
Data engineering

Running a markdown cell will apply the formatting that you have indicated in the cell. Similarly, running a code cell will execute the code that you have written in the cell.

Step 8: Import Python modules

a Click the markdown cell titled Import Needed Modules.

b From the ArcGIS Notebook Toolbar, click the Insert Cell Below button ✚.



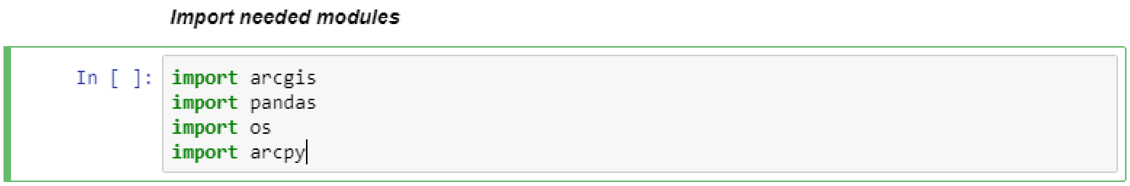
Import needed modules

```
In [ ]:
```

A code cell is added under the markdown cell. You will use this cell to import the Python modules required to complete this exercise.

c Use the **import** syntax to import the following Python modules:

- **arcgis**
- **pandas**
- **os**
- **arcpy**

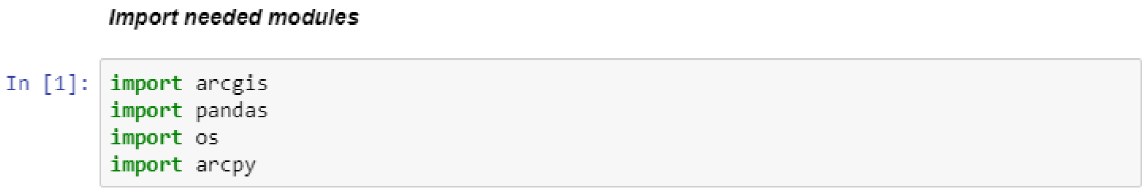


Import needed modules

```
In [ ]: import arcgis
        import pandas
        import os
        import arcpy
```

This code cell will call the modules from the ArcGIS Pro conda environment. To the left of the code cell is blue text with brackets. When you run a code cell, an asterisk appears in the brackets to indicate that the cell is running. When the cell is complete, the asterisk is replaced with a number.

d From the ArcGIS Notebook Toolbar, click the Run button.



Import needed modules

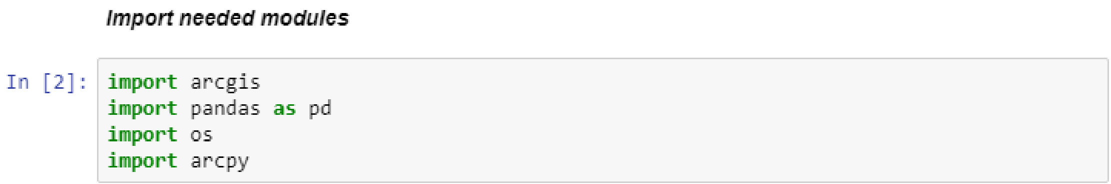
```
In [1]: import arcgis
        import pandas
        import os
        import arcpy
```

The number 1 appears in the brackets to indicate that the cell has been executed, which means that the modules were successfully loaded.

You will use the pandas module quite often in this exercise. Instead of typing `pandas` each time, you will shorten `pandas` to `pd`.

e Modify the line of code that says `import pandas` to say **import pandas as pd**.

f Click the Run button.



Import needed modules

```
In [2]: import arcgis
        import pandas as pd
        import os
        import arcpy
```

You used pd as a variable. A variable is a name that references an object. The object could be a dataset or, in this case, a Python module. You could have shortened `pandas` to any variable name. You used `pd` because it is the most common local name for pandas. The remaining code cells will use `pd` when using pandas functionality.

Step 9: Create a Pandas DataFrame

Next, you will use the pandas functionality to create a data frame. A Pandas DataFrame is a tabular data structure of columns and rows. The columns are referred to as the attributes, or attribute fields, and the rows are referred to as the records.

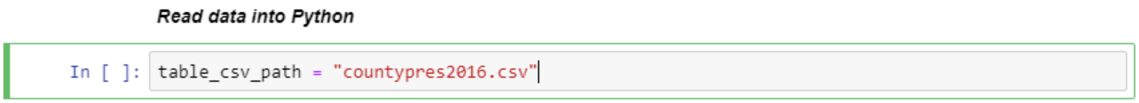
To create a data frame, your first step is to define a variable for the dataset.

a Click the markdown cell titled Read Data Into Python.

b From the ArcGIS Notebook Toolbar, click the Insert Cell Below button + .

c Create a variable called **table_csv_path** for the **countypres2016.csv** dataset.

Hint: Remember to add an equal sign (=) after the variable and enclose the dataset with quotation marks.



Read data into Python

```
In [ ]: table_csv_path = "countypres2016.csv"
```

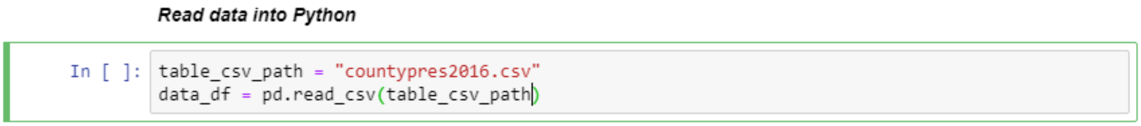
By defining this variable, you can use `table_csv_path` throughout the script to refer to the county election dataset (countypres2016.csv).

d On your keyboard, press Enter to start a new line of code.

You will use the Pandas `read` function to load the county election dataset into the data frame.

e In the code cell, create a variable called **data_df**.

f Add the **pd.read_csv** function with **table_csv_path** as the input parameter.

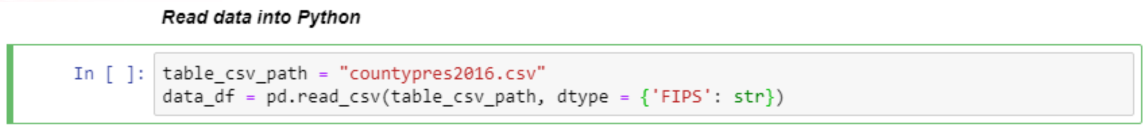


Read data into Python

```
In [ ]: table_csv_path = "countypres2016.csv"
        data_df = pd.read_csv(table_csv_path)
```

You want to specify that the FIPS attribute field in this data frame will be a text, or string value. You will use the `dtype` parameter to specify this field type.

g After `table_csv_path`, add a comma and type **dtype = {'FIPS': str}**.



Read data into Python

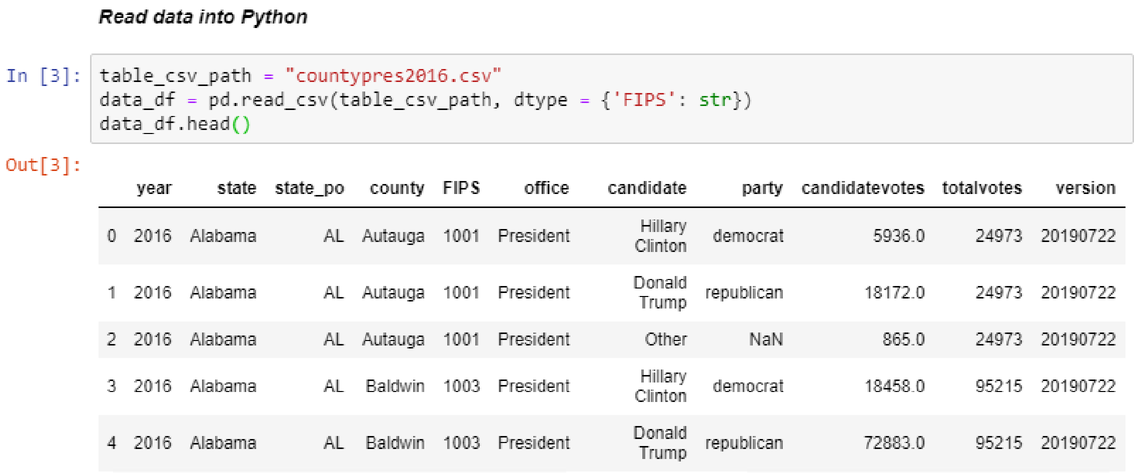
```
In [ ]: table_csv_path = "countypres2016.csv"
        data_df = pd.read_csv(table_csv_path, dtype = {'FIPS': str})
```

h On your keyboard, press Enter to start a new line of code.

You will use the Pandas `head` function to preview the first five records of the data frame, confirming that the dataset loaded properly.

i In the code cell, type **data_df.head()**.

j Run the code cell.



Read data into Python

```
In [3]: table_csv_path = "countypres2016.csv"
        data_df = pd.read_csv(table_csv_path, dtype = {'FIPS': str})
        data_df.head()
```

Out[3]:

	year	state	state_po	county	FIPS	office	candidate	party	candidatevotes	totalvotes	version
0	2016	Alabama	AL	Autauga	1001	President	Hillary Clinton	democrat	5936.0	24973	20190722
1	2016	Alabama	AL	Autauga	1001	President	Donald Trump	republican	18172.0	24973	20190722
2	2016	Alabama	AL	Autauga	1001	President	Other	NaN	865.0	24973	20190722
3	2016	Alabama	AL	Baldwin	1003	President	Hillary Clinton	democrat	18458.0	95215	20190722
4	2016	Alabama	AL	Baldwin	1003	President	Donald Trump	republican	72883.0	95215	20190722

You created a data frame for the county elections dataset that you will use to prepare, reformat, and geoenable your data.

k In ArcGIS Pro, from the Notebook tab, in the Notebook group, click Save.

l Execute the rest of the notebook and review each step as you execute each cell.

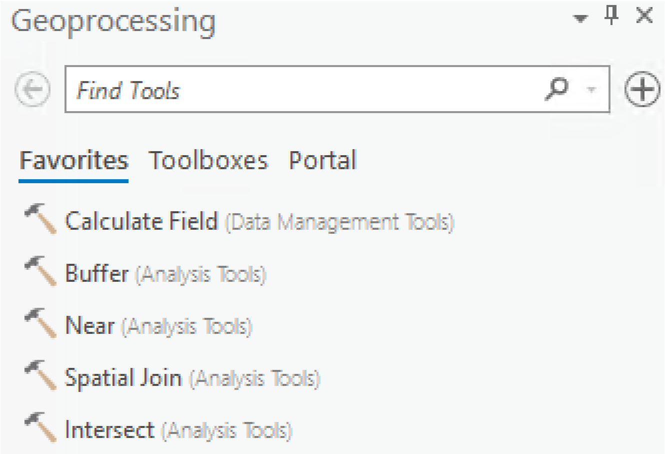
> ⚠ You must execute each cell in the notebook before proceeding to the next step.

Note: Although you are not writing all the Python code, it is recommended that you carefully look at the Python syntax and logic in each cell. Reviewing each cell can help you to familiarize yourself with the ArcGIS Notebook interface and learn Python syntax. The notebook can also act as sample code that you can reference for data engineering tasks.

Step 10: Open the Enrich tool

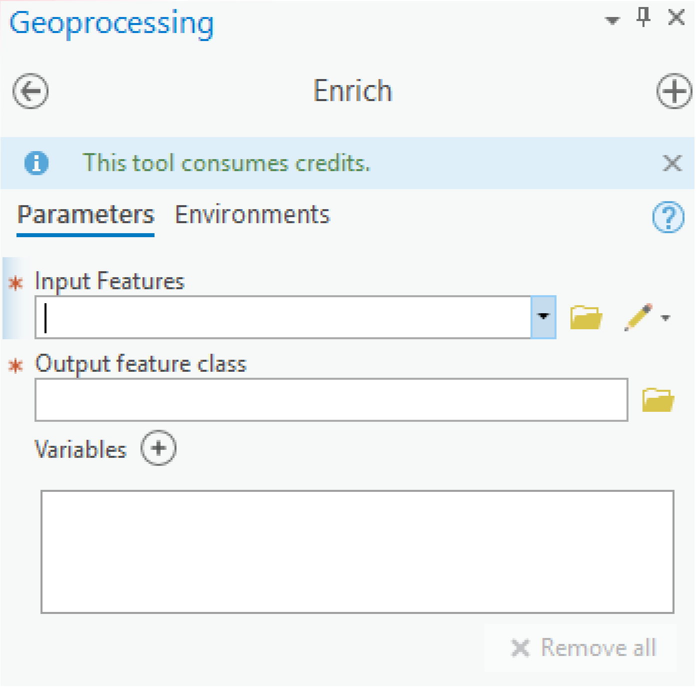
Geoenrichment will use the location of your data to add demographic variables as attributes to your feature class. Geoenrichment can be performed using ArcPy in a notebook, but the Enrich tool in ArcGIS Pro allows you to explore potential variables that you would like to add to the feature class.

- a In ArcGIS Pro, click the Data Engineering map tab.
- b From the Analysis tab, in the Geoprocessing group, click Tools.


Geoprocessing
Find Tools
Favorites
Toolboxes
Portal
Calculate Field (Data Management Tools)
Buffer (Analysis Tools)
Near (Analysis Tools)
Spatial Join (Analysis Tools)
Intersect (Analysis Tools)

The Geoprocessing pane opens. The Geoprocessing pane is used to browse or search for geoprocessing tools available with ArcGIS Pro.

- c In the Geoprocessing pane, click Toolboxes.
- d Expand Analysis Tools, and then expand Statistics.
- e Click Enrich.



The Enrich tool opens. In the Geoprocessing pane, the Enrich tool lists the parameters required to run the tool. Parameters define the values used to run the tool and its underlying algorithms. To run the Enrich tool, you will need to define the input feature class, a name for the output feature class, and the variables that will be added to the output feature class.

The Enrich tool uses credits. To learn more about credits, see ArcGIS Pro Help: Understand credits.

Step 11: Geoenrich the data

a Under Input Features, click the Browse button 📂.

b In the Input Features dialog box, double-click Data Engineering and Visualization.gdb.

c Click county_elections_pres_2016_final and click OK.

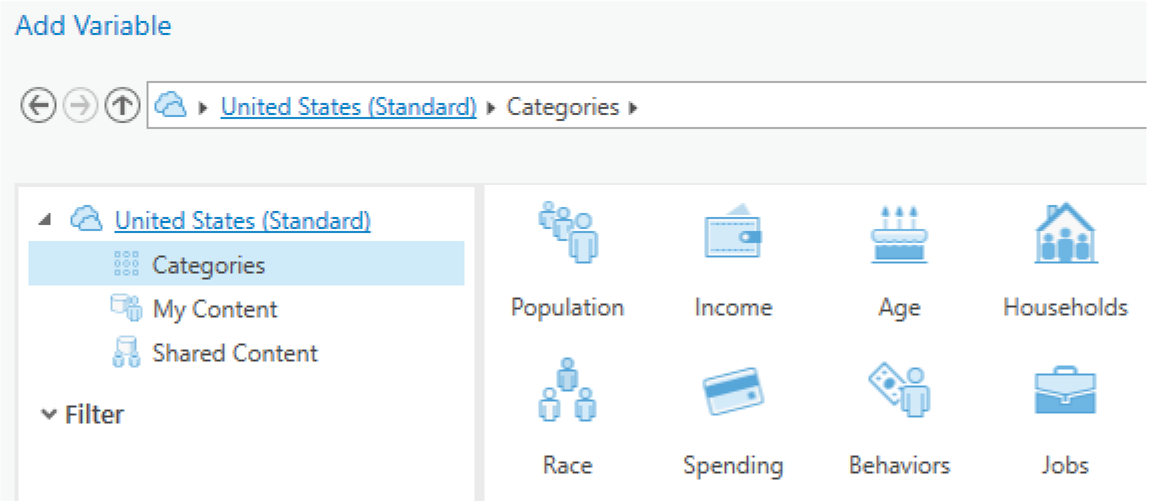
Note: If you do not see county_elections_pres_2016_final, return to the Create a Pandas DataFrame step and verify that you have executed each cell in the notebook.

The tool will automatically create an output feature class name that reflects the input. You can keep this name or modify it to be more meaningful for your analysis.

d Under Output Features, delete county_elections_pres_2016_Enrich, and then type **CountyElections2016Enrich**.

Note: This parameter represents a file path that leads to the ArcGIS Pro project's file geodatabase (Data Engineering and Visualization.gdb). Ensure that the output name that you enter is at the end of this pathname to indicate that the feature class should be stored inside the geodatabase.

e Next to Variables, click the Add button ⊕.



Esri provides various demographic variables that you can add to your data. You can also add variables that you created or that were shared with you.

f In the Add Variable dialog box, in the search field, type **2019 Median Age** and press Enter.

g If necessary, expand 2019 Age: 5 Year Increments (Esri) and click 2019 Median Age.

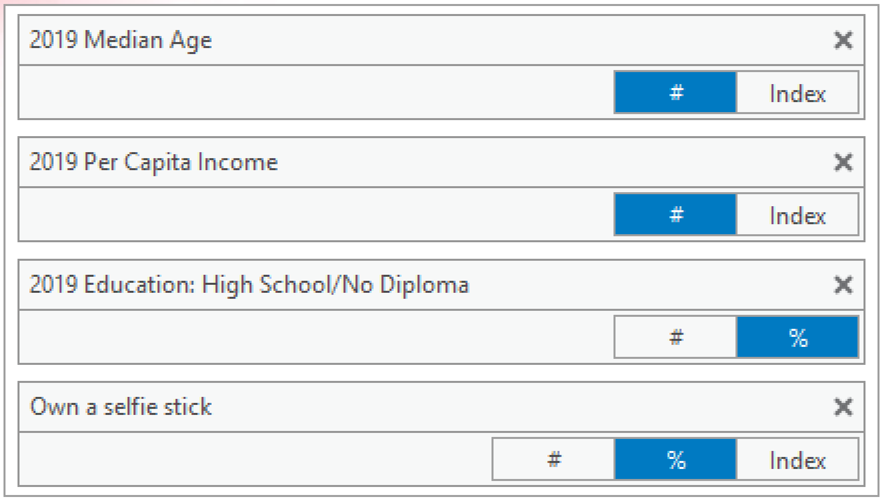
To the right of 2019 Median Age are a hashtag and the word Index. These icons, along with a percent sign icon, are used to specify if you want a total count (hashtag), index, or percentage (percent sign) of the variable.

h Confirm that the hashtag is selected.

i Repeat the previous steps to select the following variables:

- **2019 Per Capita Income** (Count)
- **2019 Education High School/No Diploma** (Percent)
- **Own A Selfie Stick** (Percent)

j Click OK.



2019 Median Age ✕

[✓ #] Index

2019 Per Capita Income ✕

[✓ #] Index

2019 Education: High School/No Diploma ✕

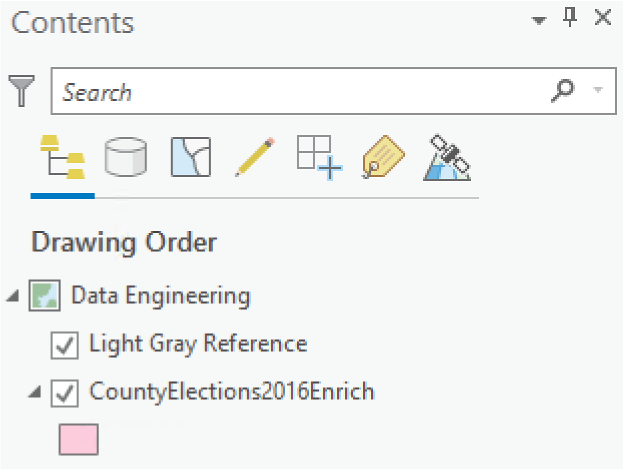
# [✓ %]

Own a selfie stick ✕

# [✓ %] Index

k Click Run.

Note: It may take a few minutes for this tool to run.



Drawing Order

◢ Data Engineering

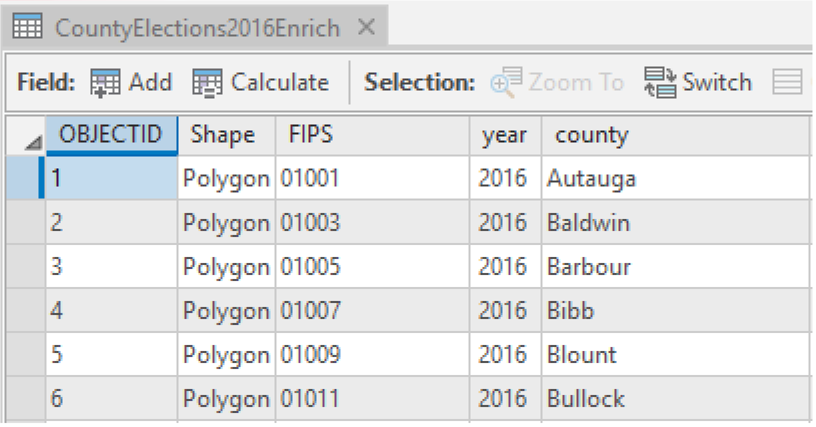
☑ Light Gray Reference

◢ ☑ CountyElections2016Enrich

The CountyElections2016Enrich layer is added to the map.

l In the Contents pane, right-click CountyElections2016Enrich and choose Attribute Table.

m Scroll the attribute table to review the data's attribute fields.



OBJECTID	Shape	FIPS	year	county
1	Polygon	01001	2016	Autauga
2	Polygon	01003	2016	Baldwin
3	Polygon	01005	2016	Barbour
4	Polygon	01007	2016	Bibb
5	Polygon	01009	2016	Blount
6	Polygon	01011	2016	Bullock

The attribute table includes the fields added in the initial data engineering steps as well as the fields added using the Enrich tool.

After completing various data engineering techniques, you cleaned and prepared the election data. Geoenabling and geoenriching the data provides demographic variables that you can use to model or predict voter turnout. You will use various visualization techniques to explore relationships between voter turnout and these variables. You will use this information to identify potential variables to use in your prediction model.

If you would like to perform additional data engineering tasks, proceed to the optional stretch goal; otherwise, save the project and exit ArcGIS Pro.

Stretch goal (Optional)

Throughout this course, you will see exercise stretch goals. These goals include ways that you can continue or enhance the work that you completed during the exercise.

Stretch goals are community-supported (meaning that your fellow MOOC participants can assist you with the steps to complete the stretch goal using the Lesson Forum), and they are a great opportunity to work together to learn together.

If you would like to continue engineering your data, you can modify the ArcGIS Notebook to include the following tasks:

1. Identify and remove records with null `candidatevotes` values in the election data.
2. Apply a symbology layer (default.lyrx) to the 2016 election turnout feature class (out_2016_fc_name).

   The default.lyrx is located in the Data Engineering and Visualization folder. The ArcGIS Pro Help: Apply Symbology From Layer (Data Management) describes the process of applying a symbology layer and includes syntax to use in your script.
3. Determine how to incorporate Alaska into this analysis.

Hint: Alaska does not have counties. Research their administrative and political subdivisions to determine how the data would need to be engineered to address this issue.

Use the Lesson Forum to post your questions, observations, and syntax examples. Be sure to include the **#stretch** hashtag in the posting title.