

Exercise

Perform data engineering tasks

Section 1 Exercise 1

10/28/2020



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/sdsdataeng> and download the exercise data ZIP file.

Note: The complete URL to the exercise data file is <https://trainingservices.maps.arcgis.com/home/item.html?id=031bfbfd5814411bac3da0fbe6f35dcb>.

- c Extract the files to a folder on your local computer, saving the files 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 2.6.

Note: This test uses a third-party executable file. If you prefer not to run this test due to security reasons, you can review the Common Questions or see ArcGIS Pro Help: [ArcGIS Pro 2.6 system requirements](#).

- a Go to the [Can You Run It? test](#).
- 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, check the Common Questions to find links to complete the recommended updates, and then run the test again.

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 graphics card requirements, see ArcGIS Pro Help: [ArcGIS Pro 2.6 system requirements \(video and graphics adapter requirements\)](#).

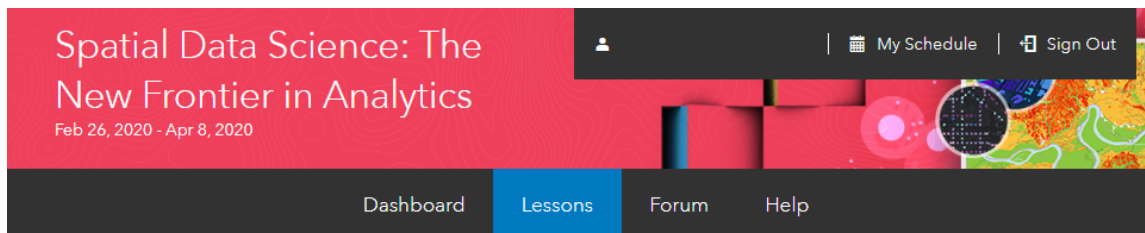
- e If your computer meets the requirements, save the report.

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

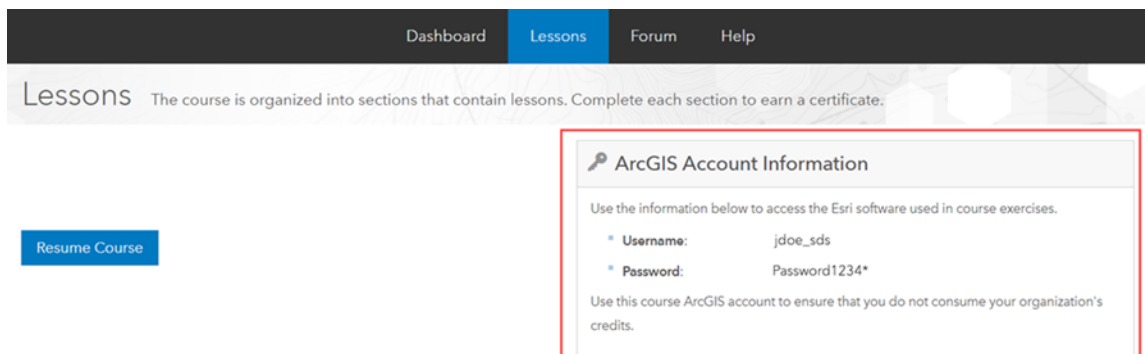
Step 3: Locate your course account to install ArcGIS Pro

This MOOC uses ArcGIS Pro 2.6. First, you will visit the MOOC home page to locate your course account user name and password. Then, you will install ArcGIS Pro 2.6 from ArcGIS Online.

- a On the MOOC home page, next to Dashboard, click Lessons.



- b Under Lessons, locate your ArcGIS Account Information.



This information is your course ArcGIS account user name and password. You will use these credentials to download ArcGIS Pro and complete all the MOOC exercises. 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 or you can always return to the Lessons tab to locate your credentials.

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.

If you already have ArcGIS Pro 2.6 installed, you can skip the remaining actions and move to the next step.


- c Open a new web browser in private or incognito mode.

Note: To learn how to enable private browsing, go to <https://bit.ly/howtobrowse>.

- d In the address bar, type **www.arcgis.com** and press Enter.



- e Click Sign In.
- f Under ArcGIS Login, copy and paste or type your course ArcGIS user name and password.

Sign in with 





ArcGIS login ^

☐ Keep me signed in

[Sign In](#)

[Forgot username?](#) or [Forgot password?](#)

Your ArcGIS organization's URL v

No account? [Create an account](#)

[Privacy](#)

- g Click Sign In.

The first time that you sign in, you may be asked to change your password and to set a security question.

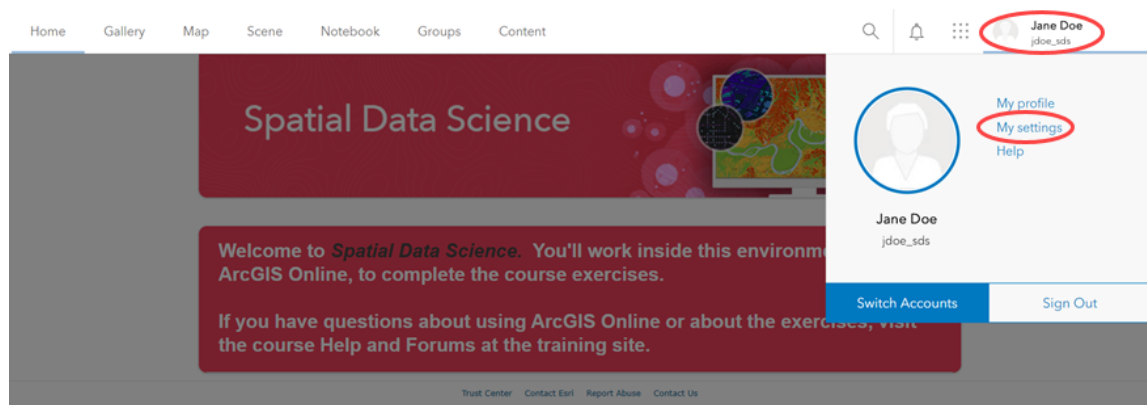
- h If necessary, follow the on-screen instructions to change your password.
- i Follow the on-screen instructions to set your security question.

The screenshot shows a web form titled "Security Question and Answer" with the Esri logo. The text reads: "A security question has not been set for your account. Setting a security question and answer allows you to reset your password if needed. Choose a question from the drop down menu below and enter your answer in the input box provided." Below this, there is a "Security Question:" label, a dropdown menu with "Select one" as the placeholder, an "Answer:" label, and a text input box. At the bottom is a blue "OK" button.

Note: An automated email will be sent to the email address associated with the account telling you that your account was recently modified. No action is required.

After you set your security question, you will see the home page of the MOOC organization.

- j** In the upper-right corner, click your account, and then click My Settings.









- k** On the left side of the page, under My Settings, click the Licenses tab.

- I Under Licensed Products, locate ArcGIS Pro.
- m To the right of the software name, click Download.

Licensed products ● Add-on license


Q Search licensed products


License

- >  ArcGIS Pro extensions
- >  Essential Apps Bundle
- >  Field Apps Bundle
- >  Office Apps Bundle
-  ArcGIS Pro ↓ Download
-  ArcGIS Runtime Standard

The Download window opens.

Download ✕


ArcGIS Pro
English (Version 2.6) ⌵
↓ Download

When your download is complete, start the installation program.
[View the installation process overview](#) 

- > File details
- > Need additional ArcGIS Pro downloads?

Note: You can run ArcGIS Pro in a different language by clicking the down arrow next to English (Version 2.6) and choosing a different supported language. 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.

n Click Download.

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

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

p Follow the installation instructions, accepting all defaults.

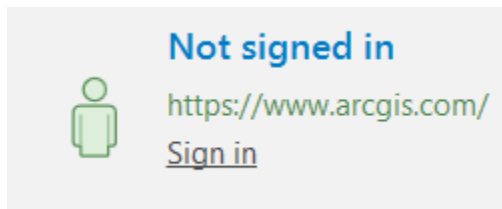
q When you are finished installing ArcGIS Pro, close the incognito browser window.

Step 4: Sign in to ArcGIS Pro

In this step, you will use the course ArcGIS account to sign in to ArcGIS Pro. You will need to use your course ArcGIS account to license ArcGIS Pro and access other software applications used throughout the MOOC exercises.

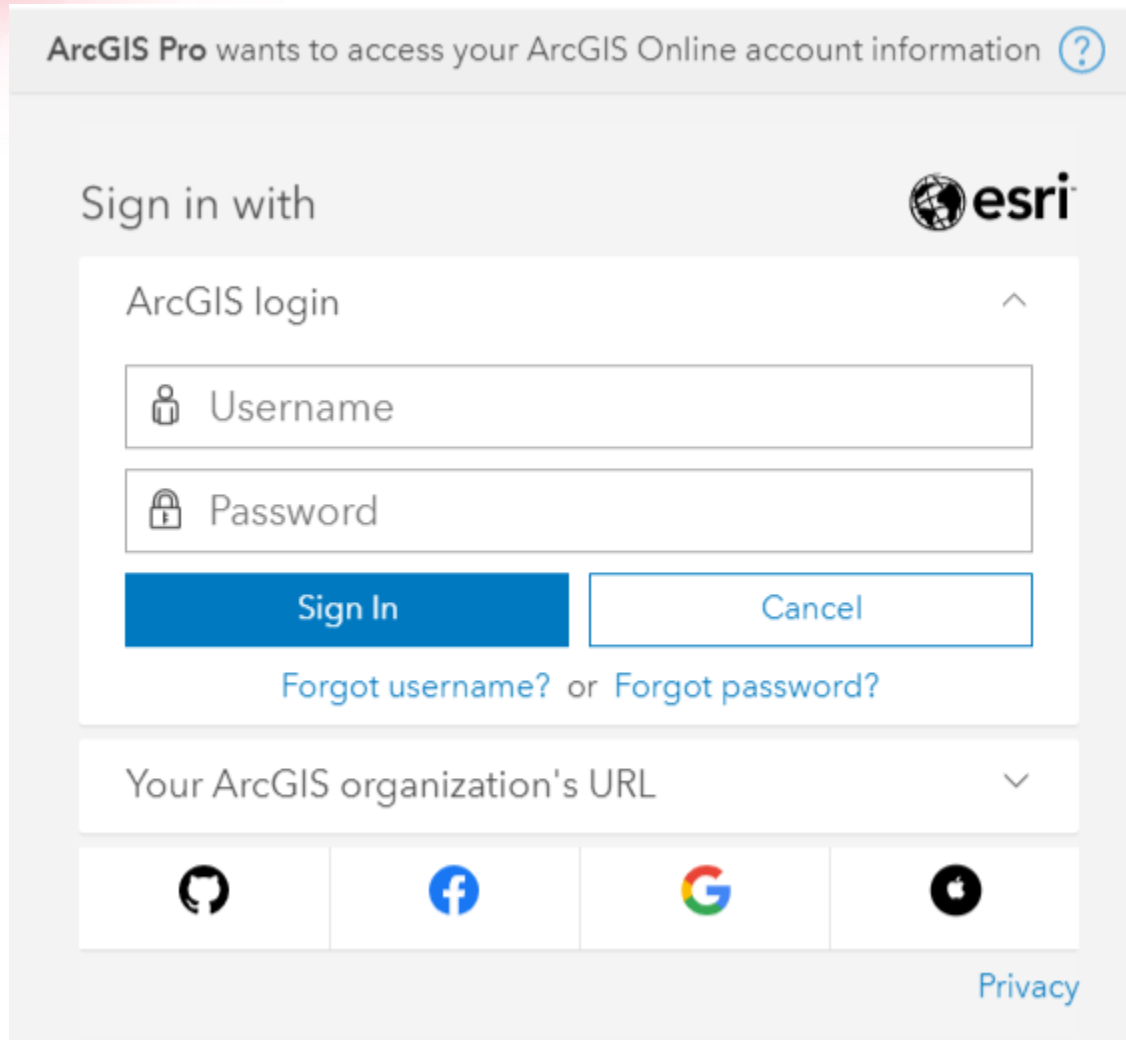
a If necessary, start ArcGIS Pro.

b In the top-right corner of ArcGIS Pro, click Sign In.




c If you are already signed in to ArcGIS Pro with a different account, click Sign Out, and then click Sign In.


d Sign in using the provided course ArcGIS account that ends in _sds.




ArcGIS Pro wants to access your ArcGIS Online account information ?

Sign in with 

ArcGIS login ^





 Username

 Password

Sign In Cancel

[Forgot username?](#) or [Forgot password?](#)

Your ArcGIS organization's URL v

[Privacy](#)

Note: The course ArcGIS account user name and password are listed on the MOOC home page under Lessons. The user name for this account ends with _sds (for example, jdoe_sds).

- e Click Sign In.

Step 5: Open an ArcGIS Pro project




- a In the bottom-left corner of the ArcGIS Pro Start page, click Open Another Project.

Note: If you have configured ArcGIS Pro to start without a project template or with a default project, you will not see the Start page. On the Project tab, click Open, and then click Open Another Project.



- b** In the Open Project dialog box, browse to the DataEngineering_and_Visualization folder that you saved on your computer.
- c** Click DataEngineering_and_Visualization.aprx to select it.
- d** Click OK.



Your ArcGIS Pro project opens to a gray reference map, called a basemap. Because you are preparing United States election data, it is currently focused on the contiguous United States.

At the top of the map is the ArcGIS Pro ribbon. ArcGIS Pro uses this horizontal ribbon to display and organize functionality into a series of tabs. On the Map tab is the Navigate group, which provides the tools that you need to navigate the map. The default tool is the Explore tool , which you can use to pan and zoom in and out of maps. To explore different areas of the world on this basemap, pan the map by clicking your mouse and holding down the button while you move the map. When you pan a map with the mouse, the pointer becomes a hand. Zoom in or out of the map using the mouse wheel or by using the Fixed Zoom In  and Fixed Zoom Out  buttons in the Navigate group.

To the left of the map is the Contents pane, which lists the layers that have been added to the map. To the right of the map is the Catalog pane, which lists the items associated with this ArcGIS Pro package—Maps, Toolboxes, Notebooks, Databases, Styles, Folders, and Locations.

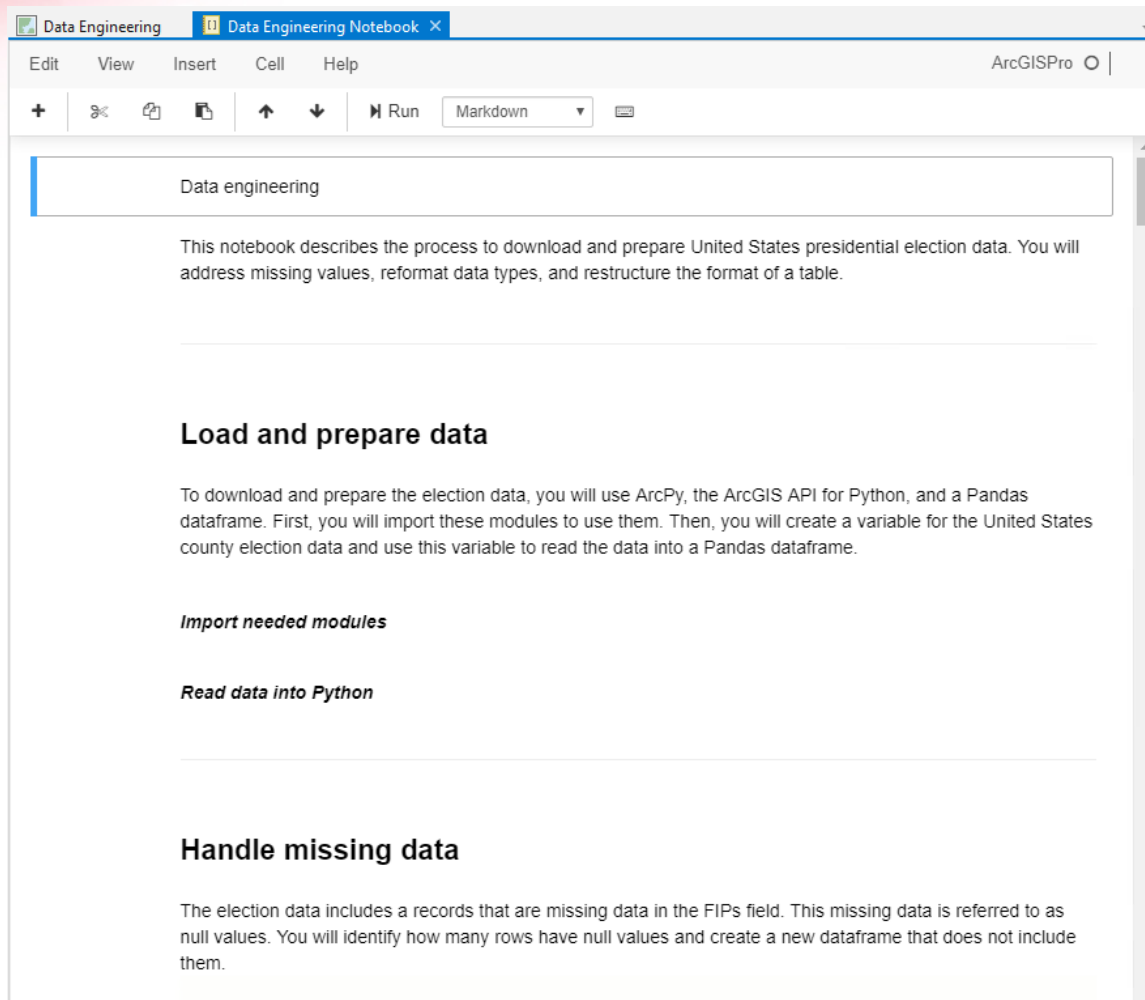
If you do not see the Contents or Catalog panes, from the View tab, in the Windows group, either click Contents  or Catalog Pane .

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.

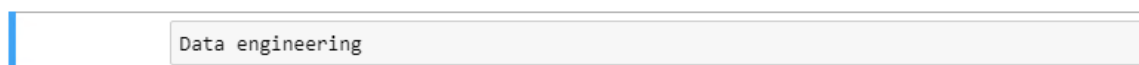


A notebook will open in the ArcGIS Pro project. The first few cells in this notebook are markdown cells used to explain the exercise.

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.



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 Notebooks toolbar, click Run .

Data engineering

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

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 Notebooks 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 Notebooks toolbar, click Run.

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**.

Import needed modules

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


f Click Run.

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 Notebooks 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, the letter *r* before the dataset to create a relative path, and enclose the dataset with quotation marks.*

Read data into Python

```
In [ ]: table_csv_path = r"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 = r"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 a space, and then type **dtype = {'FIPS': str}**.

Read data into Python

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

- h 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 = r"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	democratic	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	democratic	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.

Before moving to the next step in this PDF, you must execute the rest of the notebook in ArcGIS Pro.

- l To execute the rest of the steps in the notebook, select each cell and either click Run  or press Shift + Enter on your keyboard.
- m Review the step as you run each cell.



You must run each cell in the notebook before proceeding to the next step.

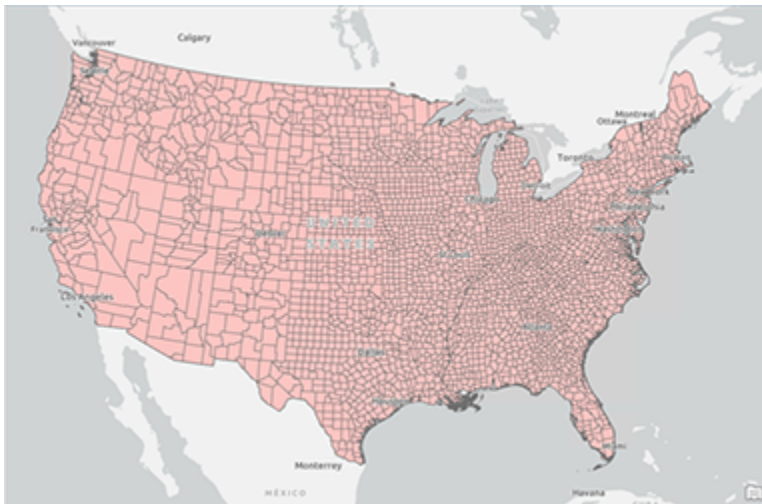
- n When you finish running each cell in the notebook, click Save.
- o Return to the PDF to continue with the rest of the steps.

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 familiarize you with the ArcGIS Notebooks 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

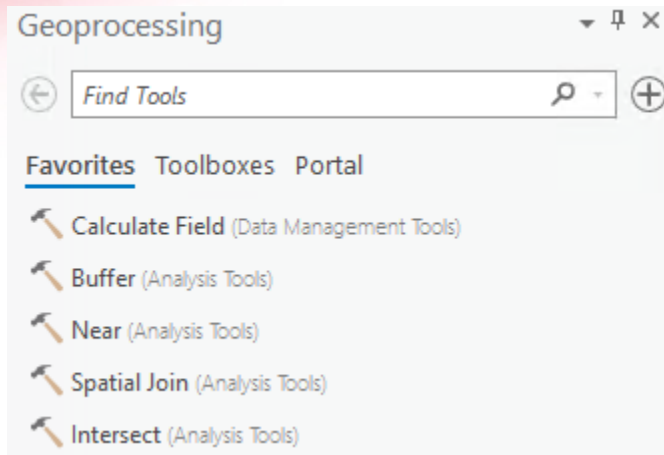
Geoenrichment will use the location of your data to add demographic variables as attributes to your feature class. Geoenrichment can be performed using ArcPy in a notebook, but the Enrich tool in ArcGIS Pro allows you to explore potential variables that you would like to add to the feature class.

- a In ArcGIS Pro, click the Data Engineering map tab.



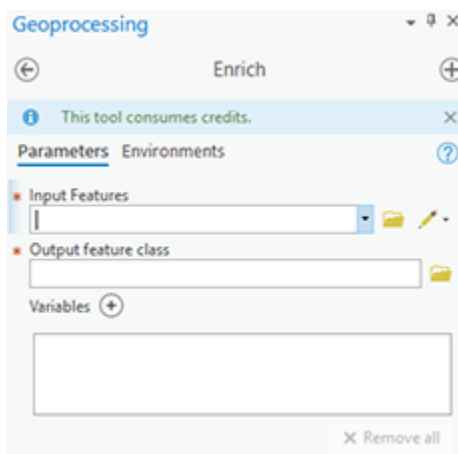
The feature class that you created in the notebook has been added to the map. The color of the data will vary every time it is added to the map.

- b From the Analysis tab, in the Geoprocessing group, click Tools.



The Geoprocessing pane opens. This 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.



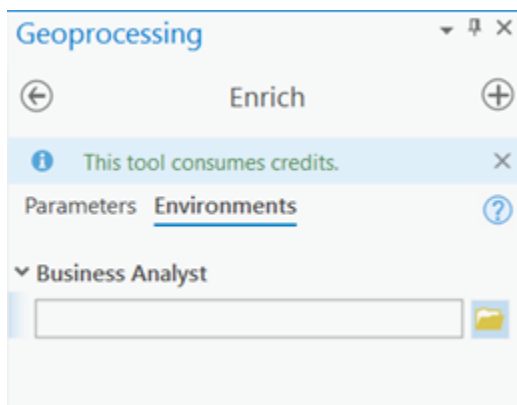
The Enrich tool opens in the Geoprocessing pane and lists the parameters required to run the tool. Parameters define the values used to run the tool and its underlying algorithms. To run the Enrich tool, you will need to define the input feature class, a name for the output feature class, and the variables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


*Note: **The Enrich tool uses credits.** You have been provided the necessary credits to complete this course. Make sure that you are signed in to ArcGIS Pro with the user name and password provided to you for the course (for example, jdoe_sds) to use the credits allocated for you to run this tool. To learn more about credits, see ArcGIS Online Help: [Understand credits](#).*

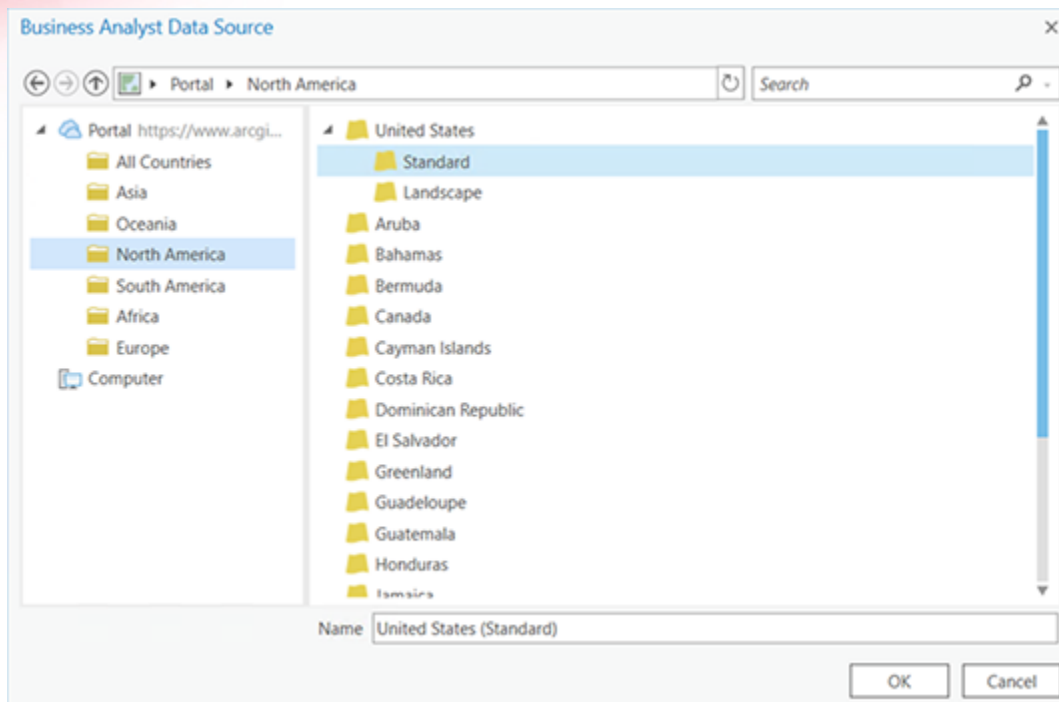
Step 11: Modify tool environment settings

First, you will set the environments of the Enrich tool to use demographic variables from the United States because the United States is our study area.

- a In the Geoprocessing pane, click the Environments tab.



- b For Business Analyst, click the Browse button .
- c In the Business Analyst Data Source dialog box, from the left pane, under Portal, click North America.
- d From the options that display under United States, click Standard to select it.




- e Click OK.

You have now set the region to select demographic variables from the United States.

- f In the Geoprocessing pane, click the Parameters tab.

Step 12: Geoenrich the data

Now you will define the parameters necessary to run the Enrich tool.

- a For Input Features, click the Browse button .
- b In the Input Features dialog box, double-click DataEngineering_and_Visualization.gdb.
- c Select County_elections_pres_2016_final and click OK.

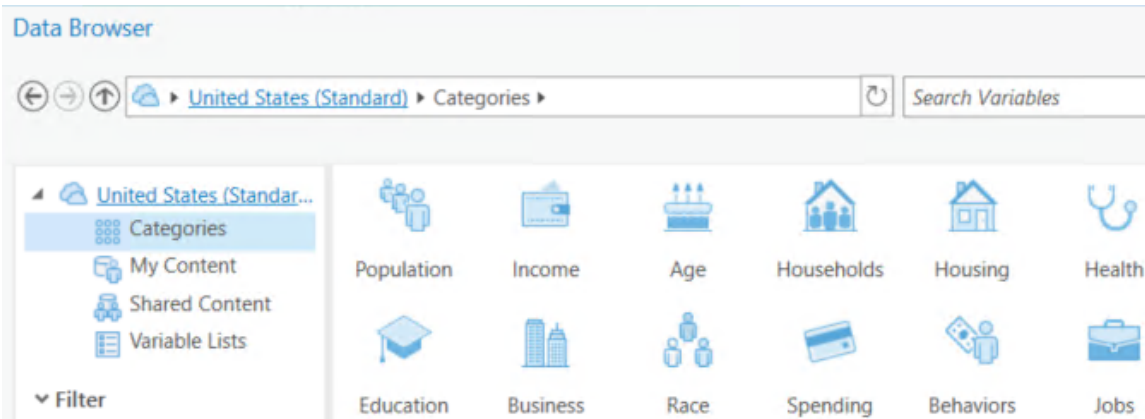
Note: If you do not see County_elections_pres_2016_final, return to the Create A Pandas DataFrame step and verify that you have executed each cell in the notebook.

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 For Output Feature Class, replace the current text with **CountyElections2016Enrich**.

Note: This parameter represents a file path that leads to the ArcGIS Pro project's file geodatabase (DataEngineering_and_Visualization.gdb). In ArcGIS Pro, the Current Workspace environment defaults to the project's default 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 Data Browser dialog box, in the Search Variables field, type **2020 Median Age** and press Enter.


- g If necessary, expand 2020 Age: 5 Year Increments (Esri).

To the right of 2020 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 Click the 2020 Median Age variable.

- i Confirm that the hashtag (total count) is selected for 2020 Median Age.

- j Click OK.

- k Click the Add button  to add the next variable and repeat the previous steps in the Data Browser dialog box to search for and select the following variables:

- **2020 Per Capita Income** (Total Count)
- **2020 Pop Age 25+: High School/No Diploma** (Percent)
- **2020 Own A Selfie Stick** (Percent)

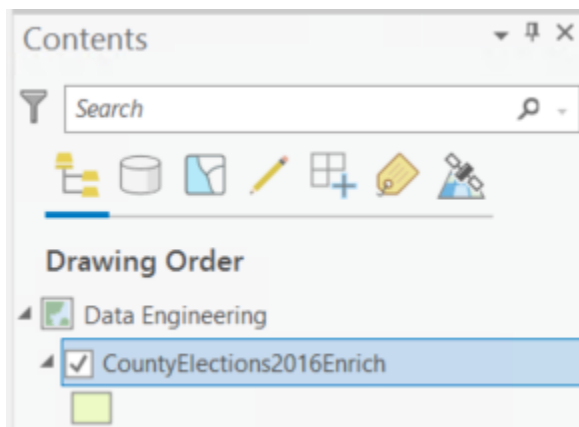
2020 Median Age	#	Index
2020 Per Capita Income	#	Index
2020 Pop Age 25+: High School/No Diploma	#	%
2020 Own a selfie stick	#	% Index

Note: If necessary, you can make any updates to icon selection directly in the Geoprocessing pane to ensure that your selections match the preceding graphic.

When you finish adding all four variables, you are ready to run the tool to enrich your data. It may take a few minutes for this tool to run.

i At the bottom of the Geoprocessing pane, click Run.

Note: You may notice a warning icon ⚠ next to Output Feature Class after you run the tool. This notification is not an issue, just a reminder that you will overwrite the layer that you previously created if you run the tool again without changing the output feature's name.



The CountyElections2016Enrich layer is added to the map.

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

n 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.

- o Close the attribute table.
- p If you would like to perform additional data engineering tasks, proceed to the optional stretch goal; otherwise, close the Data Engineering map and notebook tabs, save the project, and then 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.

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 DataEngineering_And_Visualization folder. The ArcGIS Pro Help: [Apply Symbology From Layer \(Data Management\)](#) documentation 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.

Note: Alaska does not have counties. Research its 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.

When you are finished, close the Data Engineering map and notebook tabs, save the project, and then exit ArcGIS Pro.