Exercise 1: Perform data engineering tasks

How can I print an exercise to PDF format?

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 Notebooks and the Data Engineering view in ArcGIS Pro to perform data engineering tasks. These tasks will use the built-in tools that are available with these products as well as tools that are available by integrating open-source libraries. Using ArcGIS Notebooks allows you to document and share the steps you take to prepare your data for an analysis. You will then have transparent and reproducible research or analysis.

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, in this exercise, you will focus on performing various data engineering tasks to prepare election result data for predictive analysis.

The data for this section is obtained from the Harvard Dataverse and the United States Census Bureau. The voter turnout dataset from Harvard Dataverse has vote totals from each U.S. county for U.S. presidential elections from 2000 to 2020.

Note: The exercises in this course include View Result links. Click these links to confirm that your results match what is expected.

Estimated completion time in minutes: 90 minutes

Expand all steps | Collapse all steps

- Step 1: Download the exercise data files

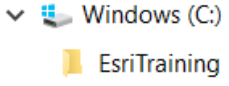
In this step, you will download the exercise data files.

a Open a new web browser tab or window.

b Go to https://links.esri.com/Section1_Data 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=9487775690064b159099d152ad04eec5.

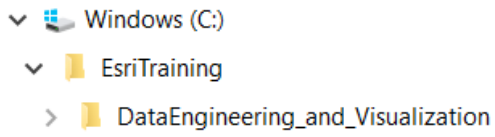
c Create a folder on your local computer and name it **EsriTraining**.



*Step 1c***: Download the exercise data files.*

Throughout this course, you will save all your data to this folder. When you create the folder, do not include any spaces or special characters in the folder name.

d Extract the exercise data files to the EsriTraining folder on your local computer.



*Step 1d***: Download the exercise data files.*

e After you extract the folder, confirm that the data files are stored in the DataEngineering_and_Visualization folder.

f Leave the DataEngineering_and_Visualization folder open.

You downloaded and extracted the exercise data files that you will need to complete the first section of the MOOC.

- 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 your computer can support ArcGIS Pro 3.1.

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 go to ArcGIS Pro Help: ArcGIS Pro 3.1 system requirements.

a Go to Can your computer run ArcGIS Pro 3.0 and 3.1?.

b Click Run Tech Check.

c Follow the steps to open and run the test.

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

d Save the report.

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

Note: If the report says that you need to update Microsoft .NET, the next step of this exercise, which is titled *Install Microsoft .NET Desktop Runtime*, will guide you through the process.

e If your computer does not meet these requirements, check the Common Questions to find links to complete any other 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, go to ArcGIS Pro Help: ArcGIS Pro 3.1 system requirements > Hardware requirements.

f If your computer meets the requirements, continue to the step that is titled *Locate your course account to install ArcGIS Pro*.

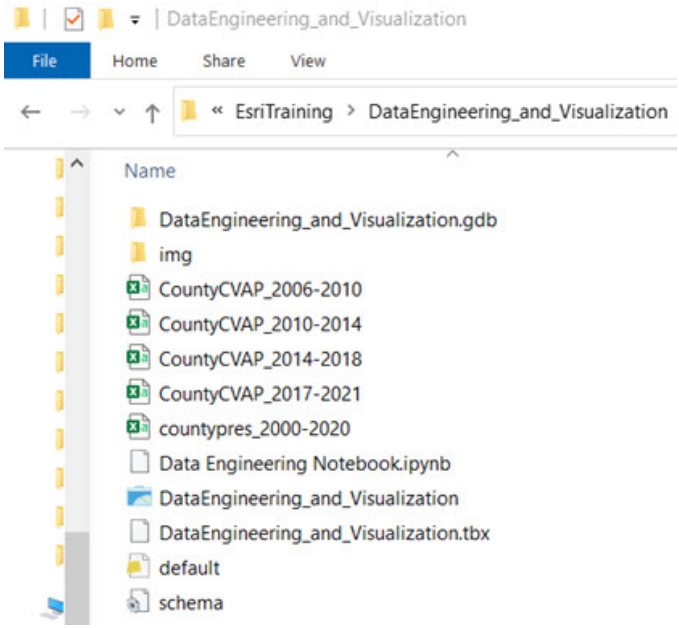
You ran and saved a report that told you whether your computer can support ArcGIS Pro 3.1.

- Step 3: Install Microsoft .NET Desktop Runtime

ArcGIS Pro 3.1 is built on .NET 6.0, Microsoft's latest edition of .NET that has long-term support. Moving to this version of .NET positions Esri and other ArcGIS Pro developers well for future development and enhancements. Because certain third-party components may start to be compatible only with .NET 6.0, it is best to use the most updated software framework.

Before you can install ArcGIS Pro 3.1 to use in the MOOC exercises, you must update your system to use .NET 6.0.

a In File Explorer, navigate to the DataEngineering_and_Visualization folder.



*Step 5a***: Explore an ArcGIS Pro project.*

The DataEngineering_and_Visualization folder shows all the data files that you need to complete both exercises in this section. You will open the ArcGIS Pro project from File Explorer.

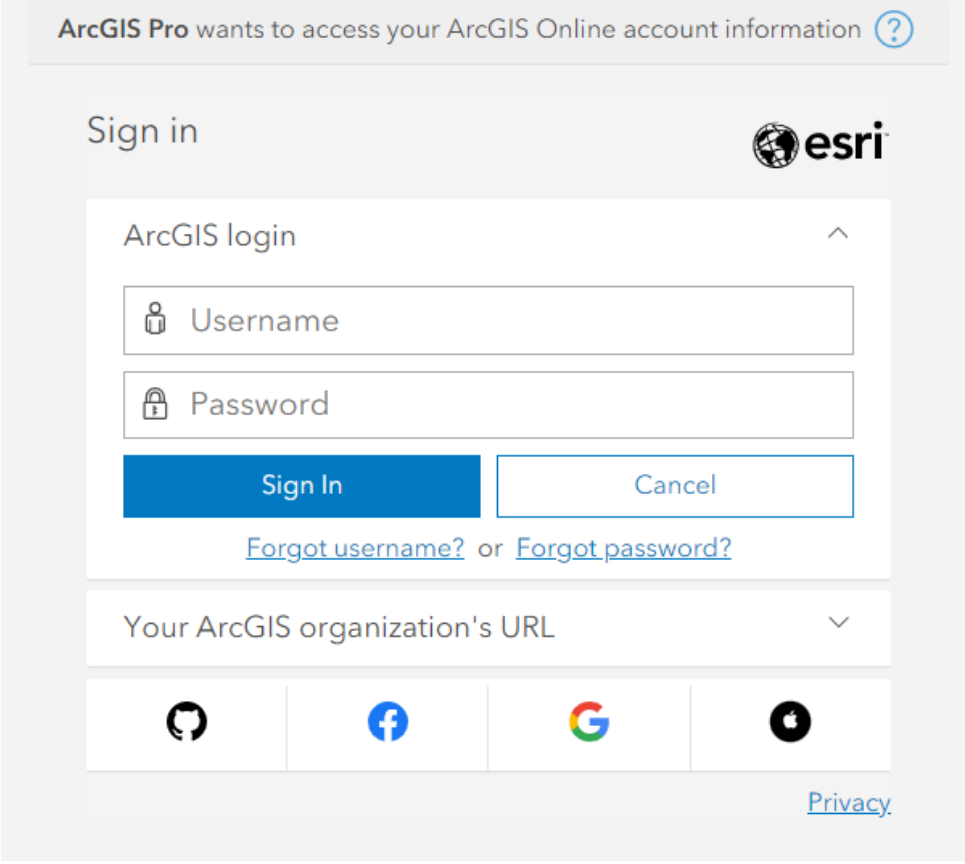
b Double-click the DataEngineering_and_Visualization ArcGIS project file.

- Hint

  Do not open the files with file names that end with .gdb or .tbx.

c Sign in to ArcGIS Pro with the provided course ArcGIS account username that ends in *_sds*.

Note: The course ArcGIS account username and password are listed on the MOOC home page under Lessons. If you are already signed in to ArcGIS Pro with a different account, in the top-right corner, click your username. Then click Sign Out. Click the Not Signed In link and then click Sign In.

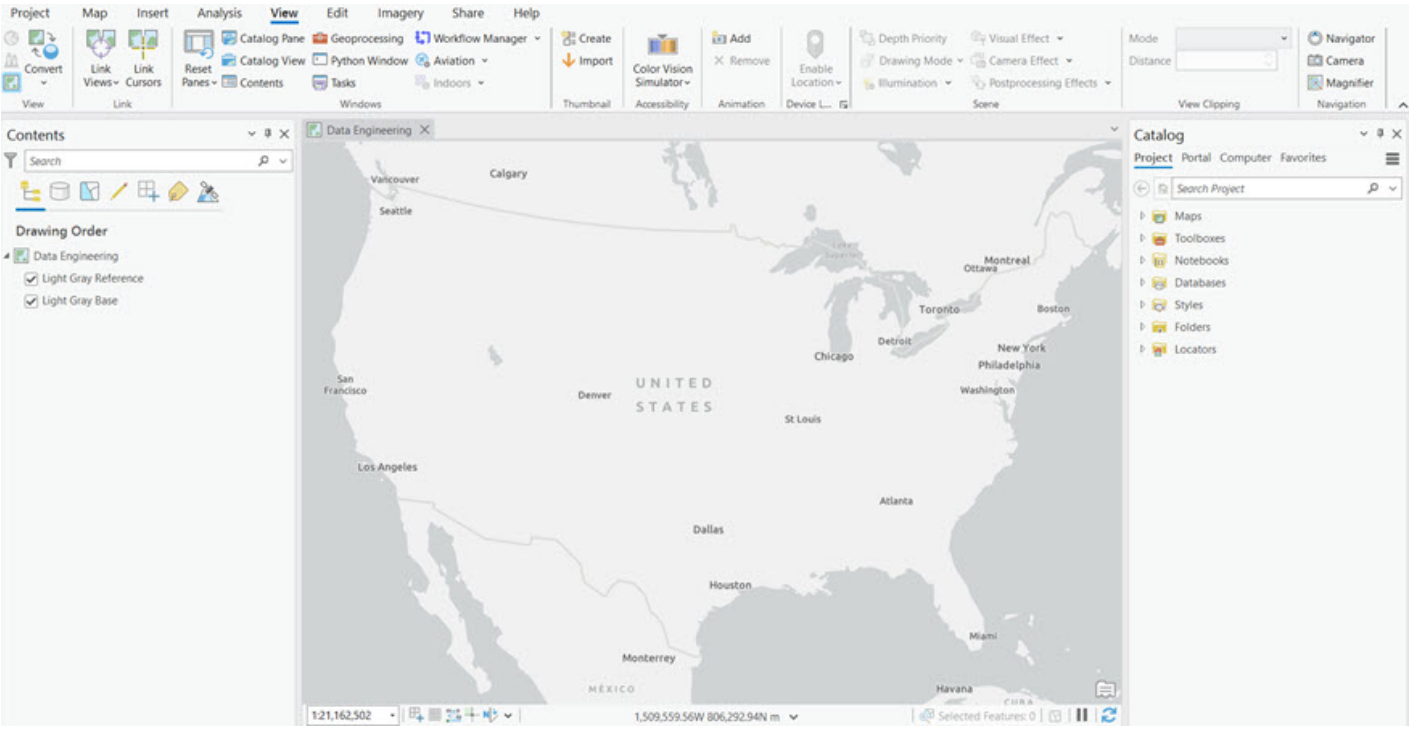


*Step 5c***: Explore an ArcGIS Pro project.*

You signed in to ArcGIS Pro with your MOOC ArcGIS account credentials. Next, you will explore the DataEngineering_and_Visualization ArcGIS Pro project that you downloaded previously. After you signed in, the ArcGIS Pro project opened to show the Data Engineering map.

d On the ribbon, click the View tab.

e In the Windows group, click Reset Panes and choose Reset Panes For Mapping (Default).



*Step 5e***: Explore an ArcGIS Pro project.*

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

Above 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 by using the mouse wheel or by using the Fixed Zoom In button or Fixed Zoom Out button in the Navigate group.

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

To learn more about the ArcGIS Pro interface, go to ArcGIS Pro Help: ArcGIS Pro user interface. To learn more about ArcGIS Pro projects, go to ArcGIS Pro Help: Projects in ArcGIS Pro.

f On the ribbon, click the Map tab.

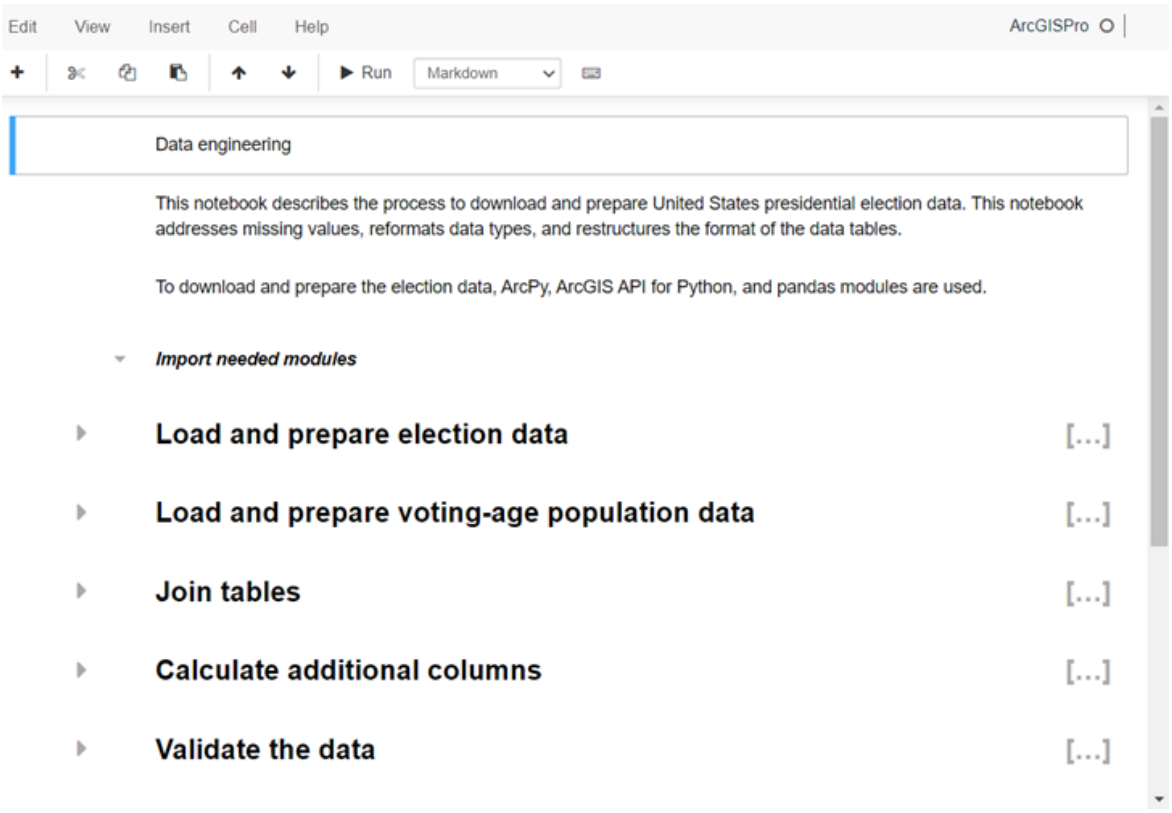
You explored an ArcGIS Pro project. Next, you will open an ArcGIS notebook.

- Step 6: Open an ArcGIS notebook

This exercise uses ArcGIS Notebooks in ArcGIS Pro. The ArcGIS Notebooks interface is built on top of Jupyter Notebook, which structures content using cells. Code cells contain executable Python code, and Markdown cells contain explanatory text and media. In this step, you will open the ArcGIS notebook that is used in this exercise.

a In the Catalog pane, expand Notebooks.

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



*Step 6b***: Open an ArcGIS notebook.*

A notebook opened in the ArcGIS Pro project. You will use this notebook to complete most of this exercise.

- Step 7: Modify a Markdown cell

The first few cells in this notebook are Markdown cells that help to explain 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 that is titled Data Engineering.

*Step 7a***: Modify a Markdown cell.*

Markdown cells use hashtags to determine the size and format of the explanatory text. Adding additional hashtags will decrease the size of the font. If you are familiar with HTML, you can think of this action 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.

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

*Step 7b***: Modify a Markdown cell.*

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

*Step 7c***: Modify a Markdown cell.*

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

d From the ArcGIS Notebooks toolbar, click Run ▶.

*Step 7d***: Modify a Markdown cell.*

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

In this step, you will import the necessary Python modules to execute the cells in the notebook. A Python module is a file that contains Python definitions and statements. A module can define functions, classes, and variables, and it can include runnable code. You will use the `import` statement to import the modules.

a Click the Markdown cell that is titled Import Needed Modules, to select the cell.

b From the ArcGIS Notebooks toolbar, click the Insert Cell Below button ✚.

*Step 8b***: Import Python modules.*

A Code cell is added under the Markdown cell. You will use this cell to import the Python modules that are required to complete this exercise.

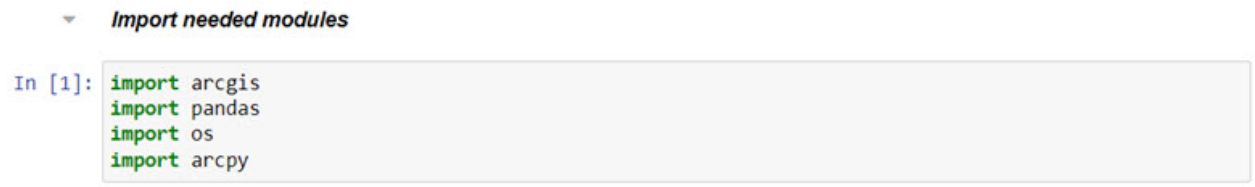
c Use the `import` syntax to import the following Python modules, pressing Enter on your keyboard after each line:

- `arcgis`
- `pandas`
- `os`
- `arcpy`

*Step 8c***: Import Python modules.*

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 inside the brackets to indicate that the cell is running. When the cell has completed running, the asterisk is replaced with a number.

d From the ArcGIS Notebooks toolbar, click Run.



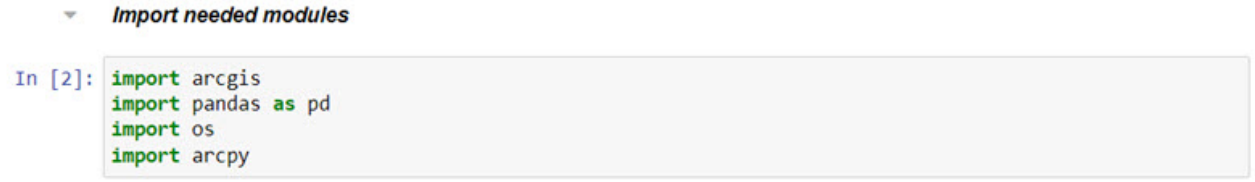
*Step 8d***: Import Python modules.*

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 Run.



*Step 8f***: Import Python modules.*

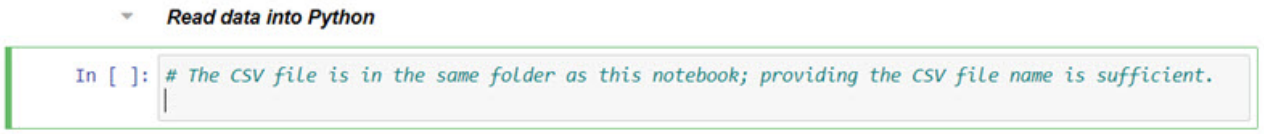
You used `pd` as a variable. A variable is a name that references an object. The object could be a dataset or, in this case, a Python module. You could have shortened `pandas` to any variable name. You used `pd` because it is the most common local name for pandas. The remaining Code cells will use `pd`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

- Step 9: Create a pandas data frame

Next, you will use the pandas functionality to create a data frame. A pandas data frame 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 gray arrow to the left of the Load And Prepare Election Data section to expand the section.

b Click the Markdown cell under Read Data Into Python.



*Step 9b***: Create a pandas data frame.*

c Under the green explanation text beginning with *The CSV file . . .* , create a variable that is called **`elections_data_path = "countypres_2000-2020.csv"`** for the CSV dataset.

*Step 9c***: Create a pandas data frame.*

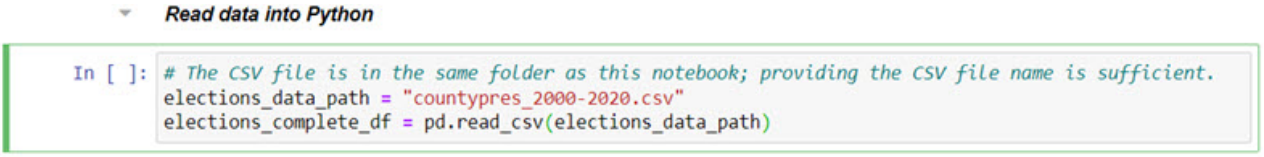
Adding an equal sign (=) after the variable followed by the countypres_2000-2020.csv dataset name enclosed within quotation marks defines this variable. You can now use the `elections_data_path` variable throughout the script to refer to the county election dataset (countypres_2000-2020.csv).

d 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 that is called **`elections_complete_df`**.

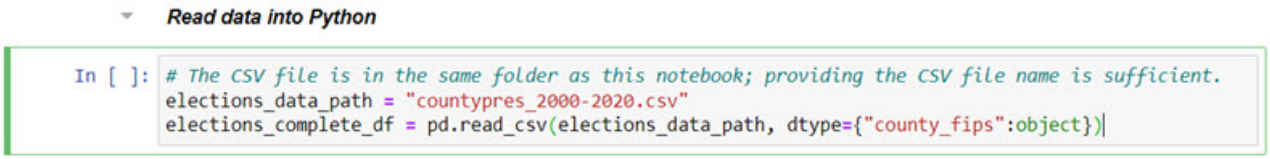
f Add the **`pd.read_csv`** function with **`elections_data_path`** as the input parameter.



*Step 9f***: Create a pandas data frame.*

You want to specify that the county_fips attribute field in this data frame will be an object. You will use the `dtype` parameter to specify this field's data type.

g After `elections_data_path`, add a comma and a space, and then type **`dtype={"county_fips":object}`**.

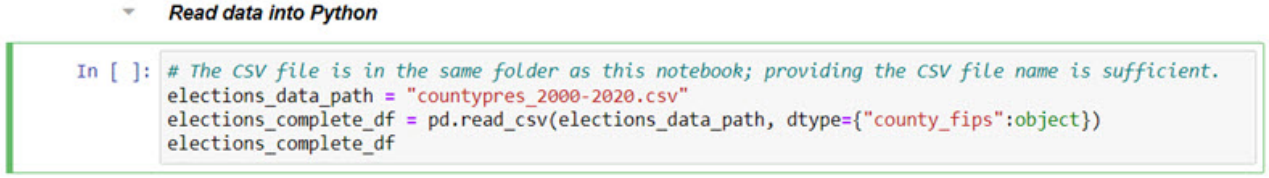


*Step 9g***: Create a pandas data frame.*

h Press Enter to start a new line of code.

You want to confirm that the dataset loaded properly before moving on in the notebook.

i In the Code cell, type **`elections_complete_df`**.



*Step 9i***: Create a pandas data frame.*

j Run the Code cell.

Read data into Python

```
In [3]: # The file is in the same folder as this notebook, so providing the file name is sufficient
elections_data_path = "countypres_2000-2020.csv"
elections_complete_df = pd.read_csv(elections_data_path, dtype={"county_fips":object})
elections_complete_df
```

Out[3]:

	year	state	state_po	county_name	county_fips	office	candidate	party	candidatevotes	to
0	2000	ALABAMA	AL	AUTAUGA	01001	US PRESIDENT	AL GORE	DEMOCRAT	4942	
1	2000	ALABAMA	AL	AUTAUGA	01001	US PRESIDENT	GEORGE W. BUSH	REPUBLICAN	11993	
2	2000	ALABAMA	AL	AUTAUGA	01001	US PRESIDENT	RALPH NADER	GREEN	160	
3	2000	ALABAMA	AL	AUTAUGA	01001	US PRESIDENT	OTHER	OTHER	113	
4	2000	ALABAMA	AL	BALDWIN	01003	US PRESIDENT	AL GORE	DEMOCRAT	13997	

*Step 9j***: Create a pandas data frame.*

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 exercise, you must expand each section and execute the rest of the steps in the notebook in ArcGIS Pro.

l Expand each section and select each cell and either click Run ▶ or press Shift + Enter on your keyboard.

m Review the step as you run each cell.

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

n After you have finished running all cells in the notebook, save and close the Data Engineering Notebook.

o Return to the exercise to continue with the rest of the steps.

Note: Although you did not write 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 the relevant Python syntax. The notebook can also act as sample code that you can reference for data engineering tasks.

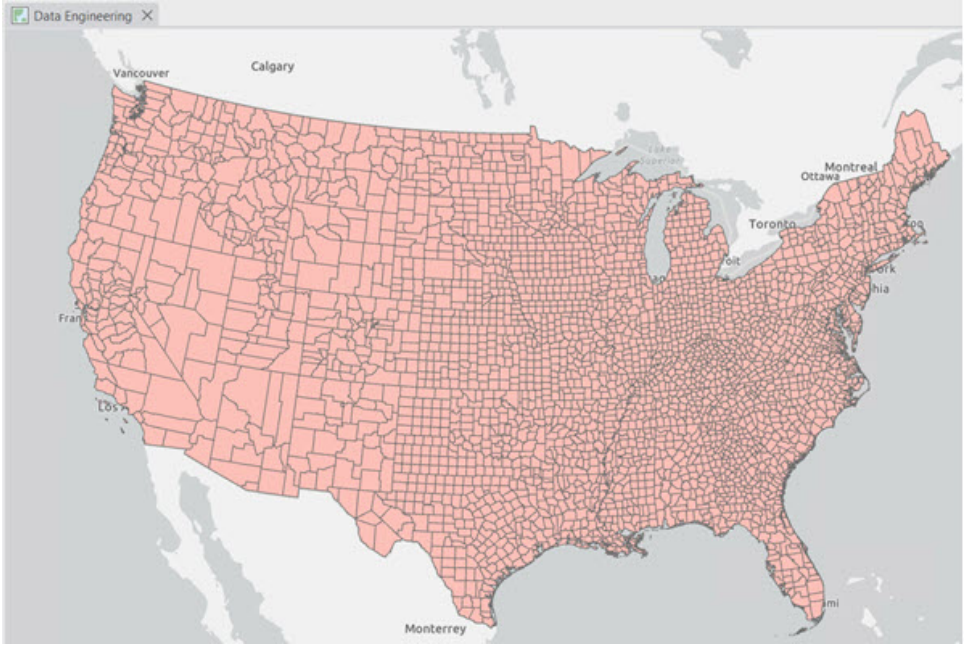
- Step 10: Modify environment settings

Before continuing with your data engineering task, you will set the geoprocessing environments. Environments are additional settings that affect geoprocessing tools and provide a powerful way to ensure that geoprocessing is performed in a controlled environment.

a In the Catalog pane, expand Databases and expand DataEngineering_and_Visualization.gdb.

b Right-click the County_Elections_Pres layer and choose Add To Current Map.

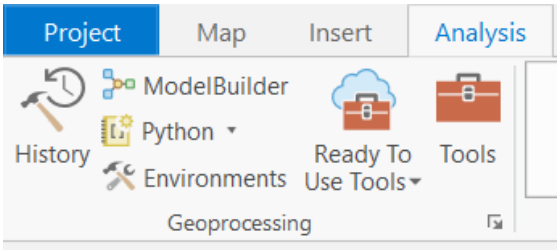
Note: If you do not see the layer, return to the *Create a pandas data frame* step and verify that you have executed each cell in the notebook.



*Step 10b***: Modify environment settings.*

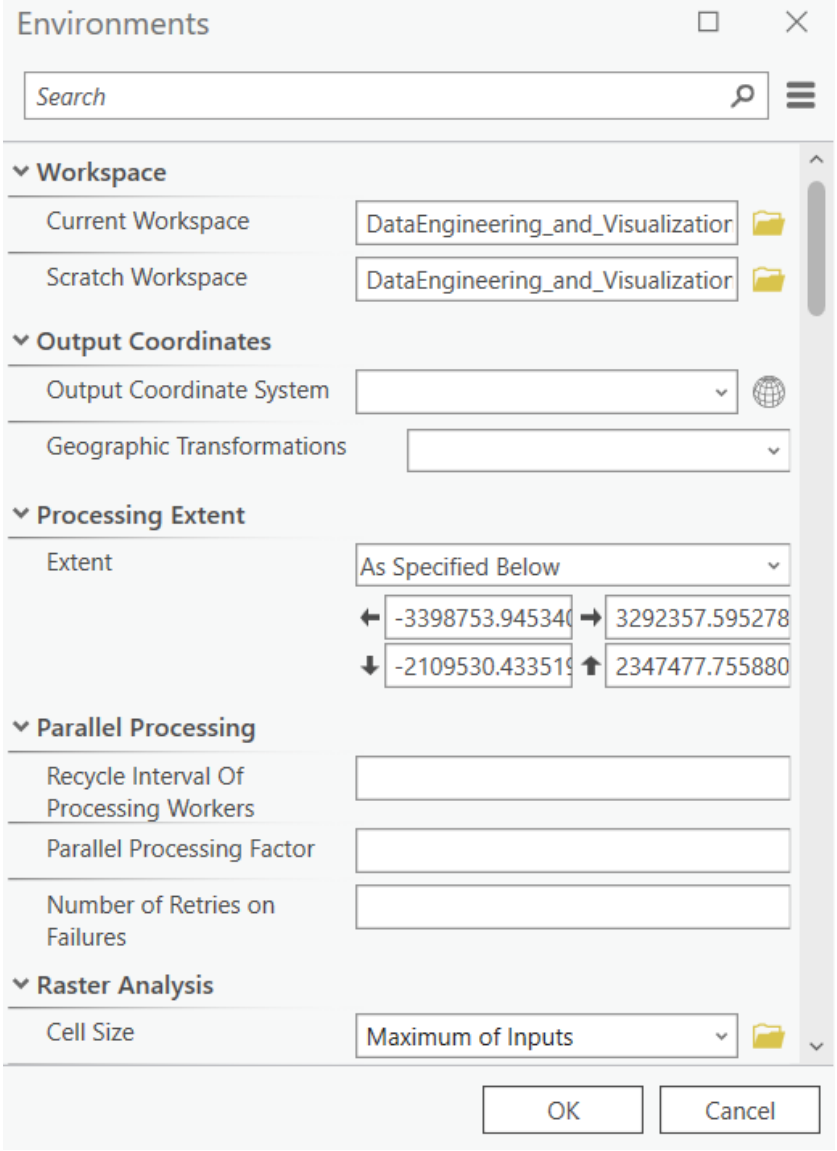
You added the feature class that you created by running all the cells in the Data Engineering Notebook. The color of the data will vary every time it is added to the map.

c On the ribbon, click the Analysis tab.



*Step 10c***: Modify environment settings.*

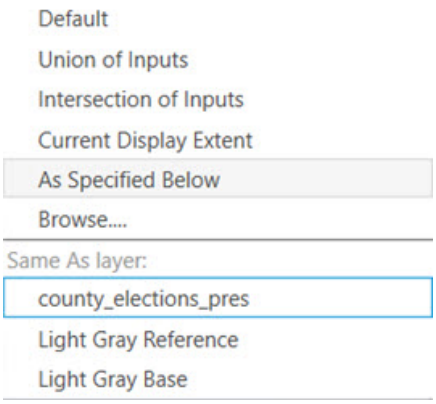
d In the Geoprocessing group, click Environments.



*Step 10d***: Modify environment settings.*

The Environments dialog box opens. Here, you can set parameters that apply to geoprocessing tools, such as the processing extent that limits processing to a specific geographic area, a coordinate system for all output geodatasets, or the cell size of output raster datasets.

e Under Processing Extent, click the Extent down arrow and choose the County_Elections_Pres layer.

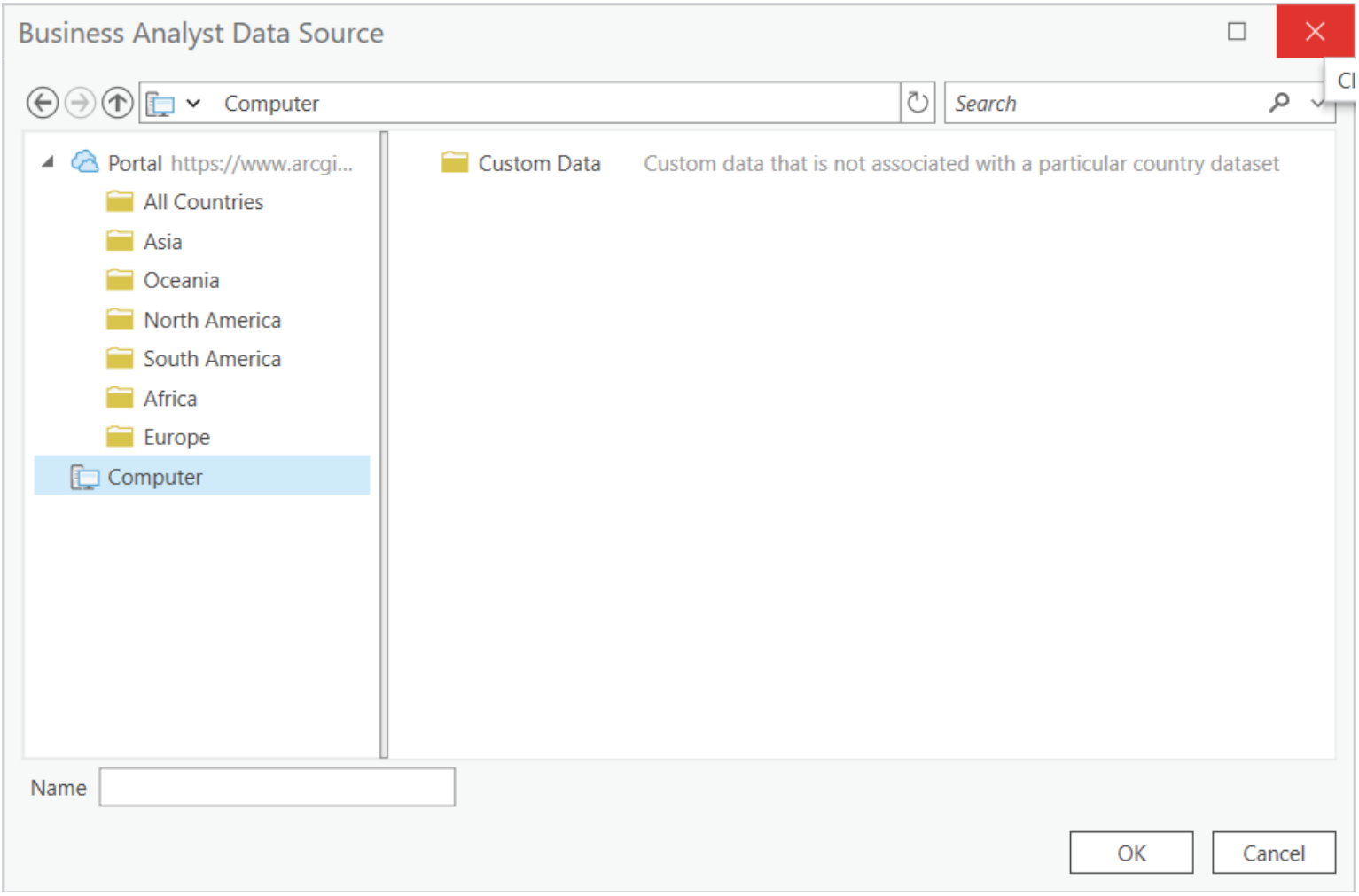


Note: Your extent will again be As Specified Below, but your extent figures now match the County_Elections_Pres layer.

Next, you will set the data source for the Enrich tool to use demographic variables from the United States because the United States is your study area.

f Scroll to the bottom of the Environments dialog box.

g Under Business Analyst, next to Data Source, click the Browse button .



*Step 10g***: Modify environment settings.*

The Business Analyst Data Source dialog box opens, which is where you can set the data source for geoenrichment to a specific country. You will set the data source to the United States.

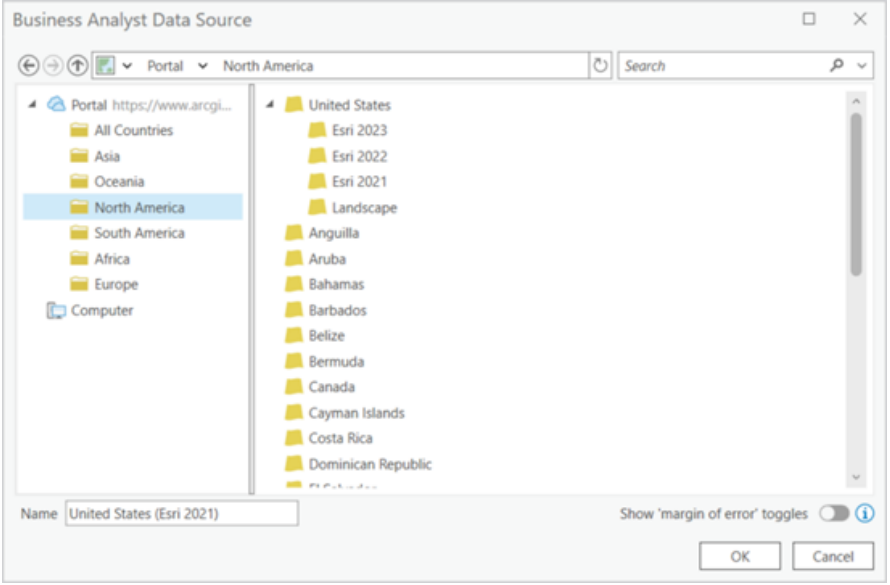
h In the Business Analyst Data Source dialog box, on the left side, under Portal, click All Countries.

i Scroll through the countries listed to see which countries and regions have demographic data available through Esri.

Esri's GeoEnrichment service enables you to query authoritative global data for more than 150 countries and regions. This extensive global data portfolio allows you to integrate global demographics, business, behavioral, environmental, and places datasets into your own data.

j On the left side of the dialog box, click North America.

k From the options that display under United States, click Esri 2022.



*Step 10k***: Modify environment settings.*

l Click OK.

Because your study area is the United States, you set the region to select demographic variables from the United States to geoenrich your data.

m In the Environments dialog box, click OK.

Note: For more information on demographic data from Esri, go to the Esri Location Data Resources web page.

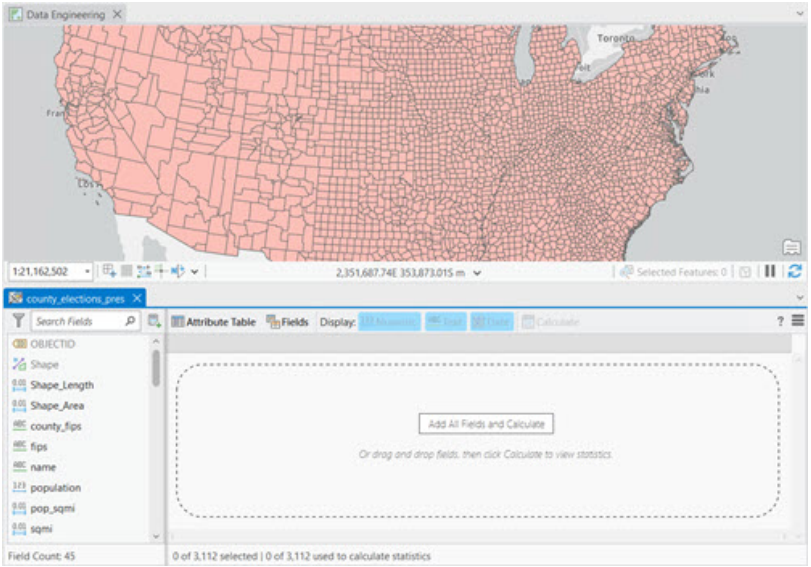
- Step 11: Open the Data Engineering view

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 Data Engineering view in ArcGIS Pro allows you to explore potential variables that you would like to add to the feature class.

a In the Contents pane, click the County_Elections_Pres layer, if necessary.

b On the ribbon, click the Analysis tab, if necessary.

c In the Workflows group, click Data Engineering.

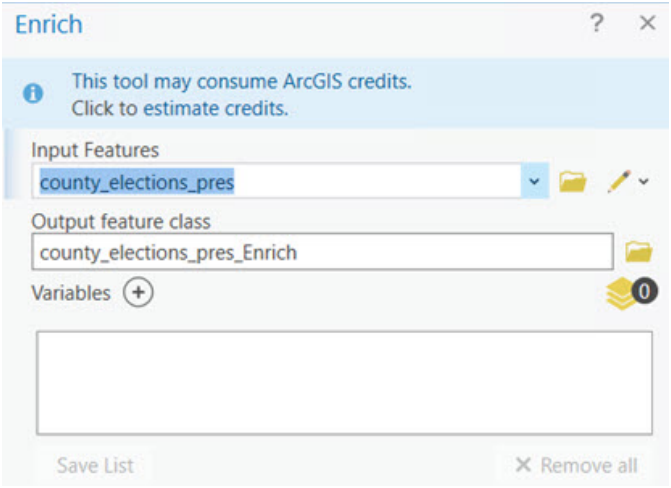


*Step 11c***: Open the Data Engineering view.*

The Data Engineering view opens in a dockable window that can be moved and docked in the same way that you dock maps, layouts, and attribute tables. In addition to the view, a Data Engineering contextual tab is available. The tab provides access to commands that are used for data engineering.

The Data Engineering view contains two panels: a fields panel and a statistics panel. The fields panel lists the fields in the layer that you used to open the view. The fields panel allows you to explore fields, change symbology, and produce charts for fields in the layer. The statistics panel allows you to explore the values and distribution of your data by viewing statistics and data quality metrics. The panel's statistics table is empty by default until you add fields from the fields panel.

d From the Data Engineering contextual tab, in the Tools group, click Integrate and choose Enrich.



*Step 11d***: Open the Data Engineering view.*

The Enrich dialog box opens.

The four tool galleries on the Data Engineering contextual tab (Clean, Construct, Integrate, and Format) each contain a subset of geoprocessing tools that can be used for data engineering tasks. You selected the Enrich tool, which enables you to add demographic variables as attributes to your feature class. The Enrich tool lists the parameters that are required to run the tool. Parameters define the values that are 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.

e Leave the Enrich dialog box open.

- Step 12: Explore geoenrichment options

You will review the workflow for geoenriching your data using the Enrich tool in the Data Engineering view.

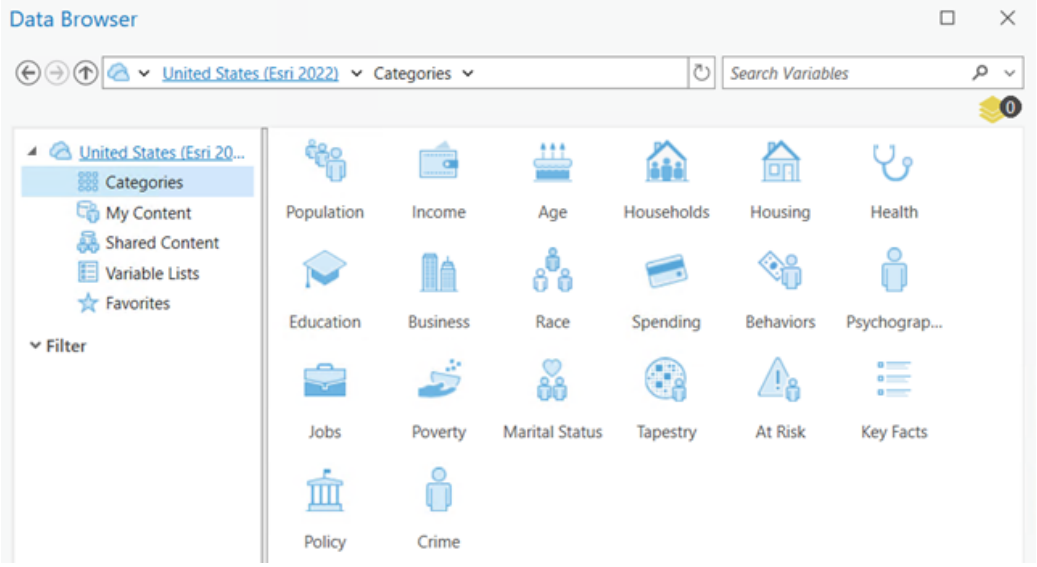
a Return to the Enrich dialog box, and confirm that the Input Features parameter is set to County_Elections_Pres.

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.

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

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.

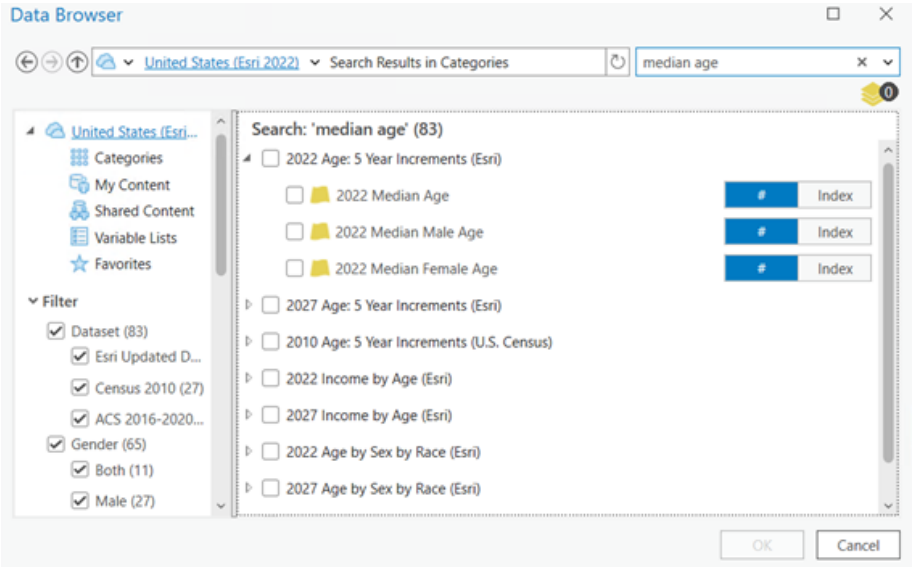
c Next to Variables, click the Add button ⊕.



*Step 12c***: Explore geoenrichment options.*

The Data Browser window is where you can explore the different demographic variables that are available for data enrichment. Esri provides various demographic variables that are regularly updated with the latest available data. For the United States, Esri also provides attributes from previous censuses (2000 and 2010) that are recalibrated with the most current census (2020) geography. You can quickly add various demographic variables to your data using the Enrich tool. You can also add variables that you created or that were shared with you.

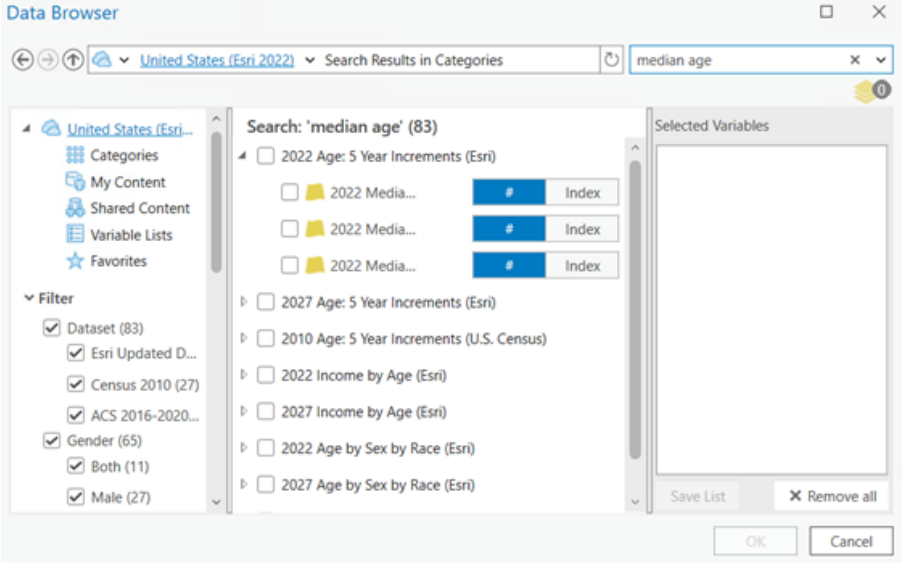
d In the Data Browser window, in the Search Variables field, type **Median Age** and press Enter.



*Step 12d***: Explore geoenrichment options.*

On the left, you have the option to filter the available variables so that you can easily focus your search. To the right of the Median Age variables, you see a hashtag and the word Index. For each variable, these icons, along with a percent sign icon, are used to specify whether you want a total count (hashtag), index, or percentage (percent sign) for the variable.

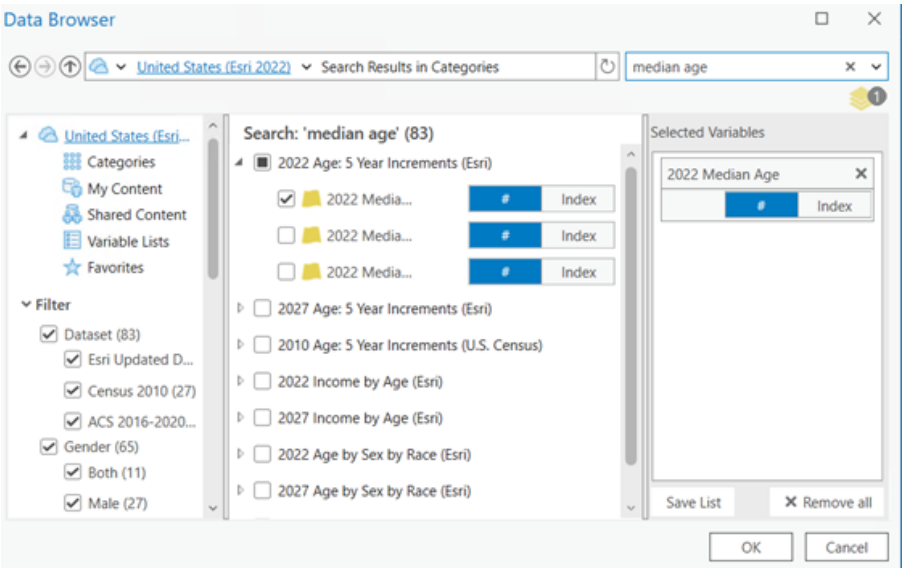
e Click the Show/Hide Details Panel button .



*Step 12e***: Explore geoenrichment options.*

The Details panel helps you keep track of the variables that you select. When a variable is selected, it is automatically listed in the Details panel.

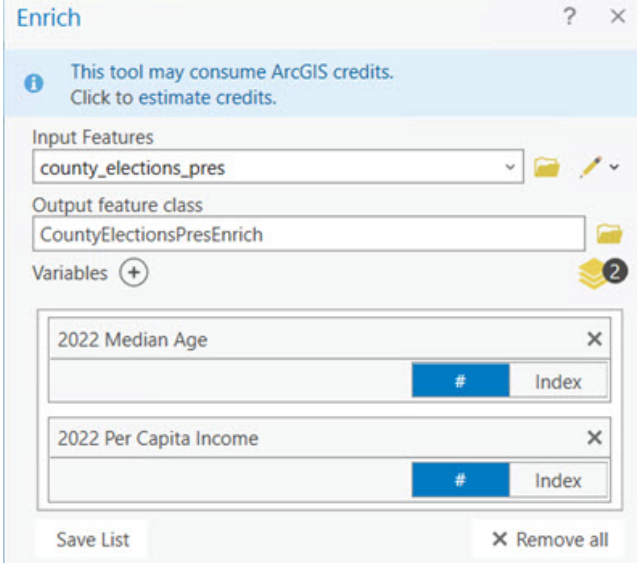
f Select the Median Age variable closest in time to 2020.



*Step 12f***: Explore geoenrichment options.*

g Search for and select the Per Capita Income variable closest in time to 2020.

h In the Data Browser window, click OK.



*Step 12h***: Explore geoenrichment options.*

The variables that you selected are added to the Variables parameter.

i At the top of the Enrich dialog box, click the Estimate Credits link.

The Enrich tool consumes ArcGIS credits when it is run. By clicking Estimate Credits, an estimate of the number of ArcGIS credits displays in the banner, as well as the number of available credits that you have. For this course, your ArcGIS Online organizational account is allocated 300 credits. You will not enrich the data because an enriched data layer has been provided for you in the next exercise.

j At the top of the Enrich dialog box, click the Close button ✕.

You explored the workflow for geoenriching your data using the Enrich tool.

k Under the map, on the Data Engineering view tab, click the Close button ✕.

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.

In the next exercise, 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 later in the MOOC.

l If you would like to perform additional data engineering tasks, proceed to the optional stretch goal. Otherwise, close the Data Engineering map, save the project, and then exit ArcGIS Pro.

- Step 13: 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:

a Identify and remove records with null `candidatevotes` values in the election data.

b Apply a symbology layer (default.lyrx) to the 2020 election turnout feature class (out_2020_fc_name).

The default.lyrx file 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 the syntax to use in your script.

c Determine how to incorporate Alaska into this analysis.