# Python and Viya Step-by-Step Tutorial: Exploring Medicare (CMS) Part D Opioid Data

Last update: June 10, 2020

## CONTENTS



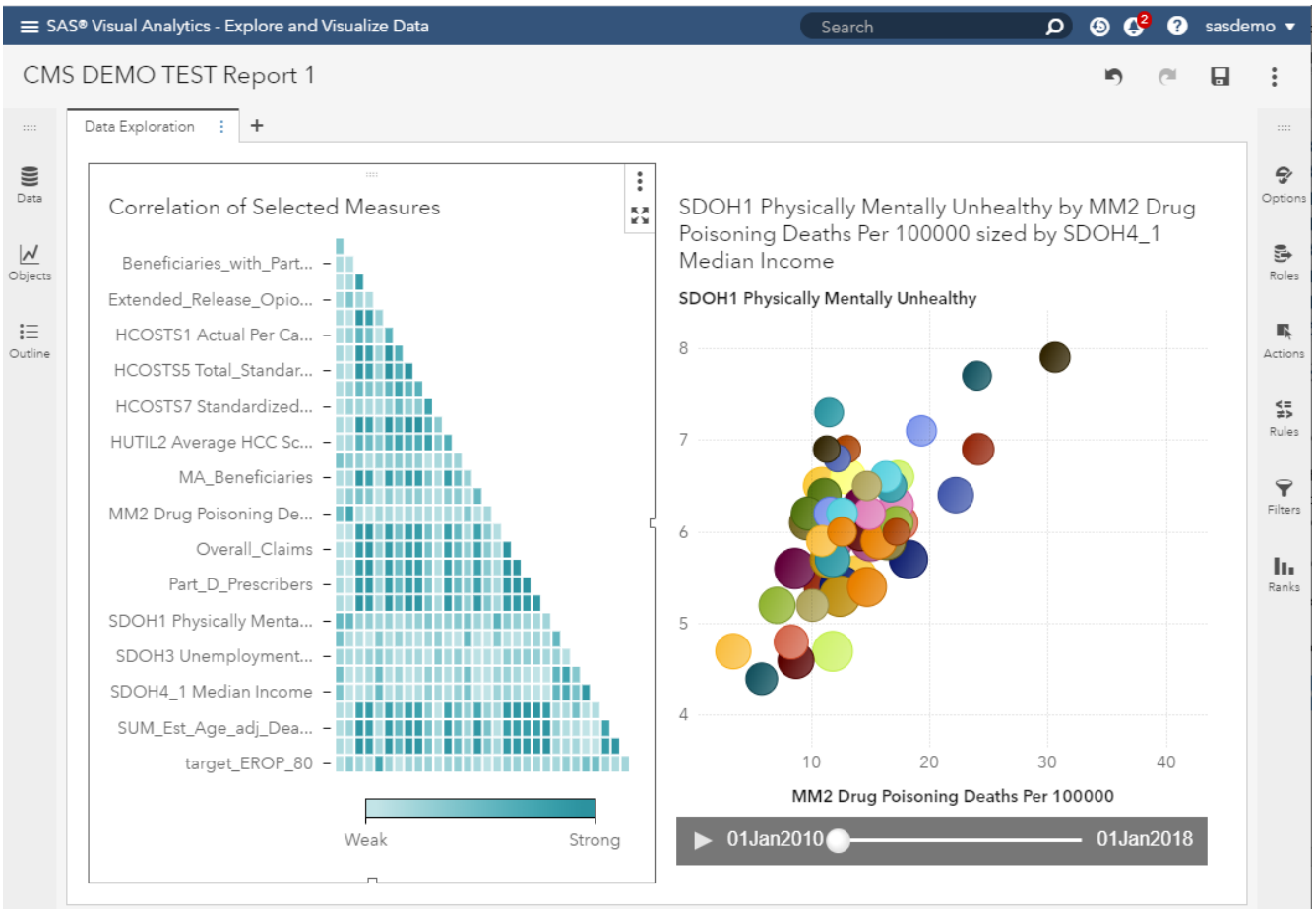
## PURPOSE

This tutorial is meant to jump start one's familiarity with SAS Studio and Viya and SAS Visual Statistics. SAS Viya provides several visualizations to help users gain insights into their data. In this tutorial, we examine CMS Part D Opioid Prescribing Rate data in the United States.

Data and Documentation is available at:

https://github.com/sasgovernment

Click on the folder link "Step-by-Step". The ultimate goal is to produce an exploration as follows:

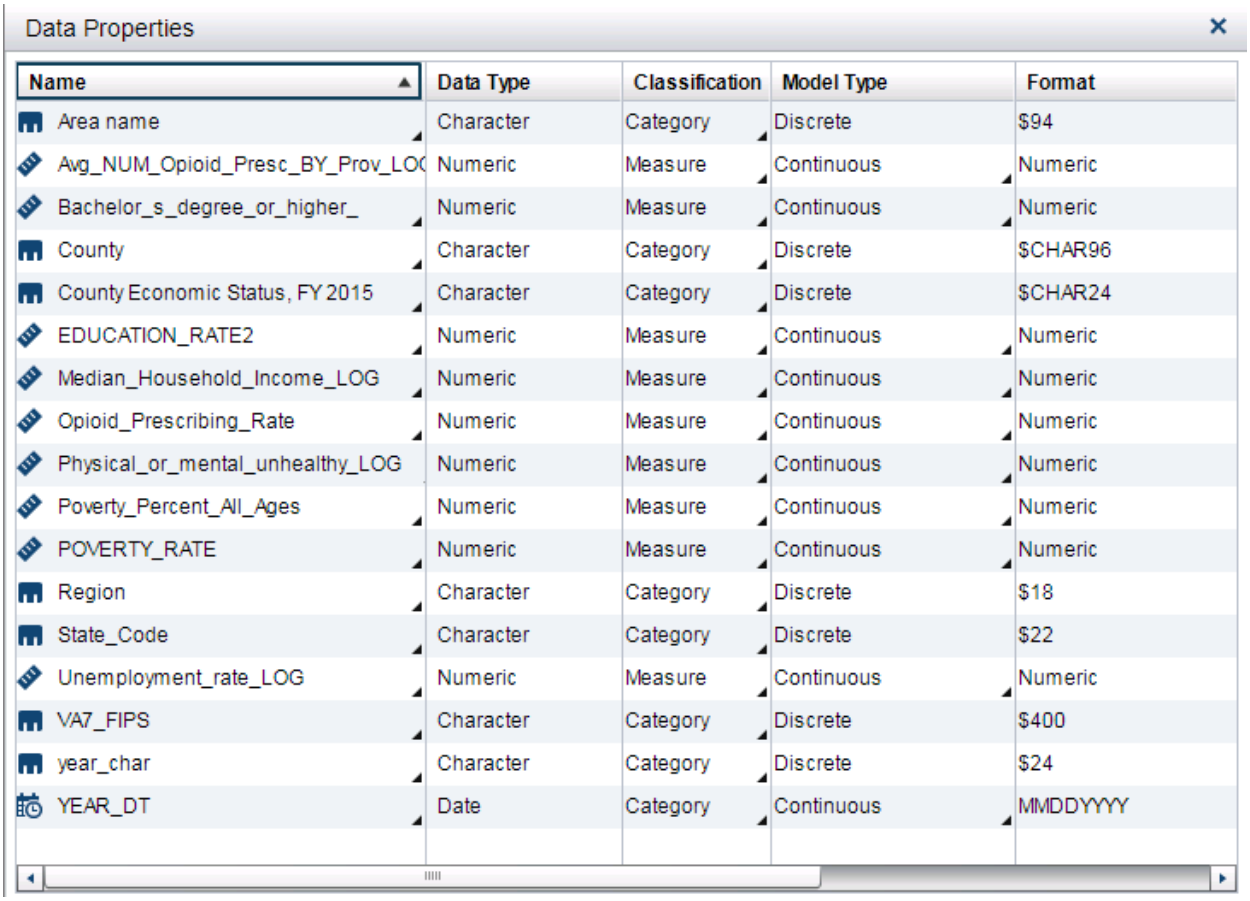



## DATA AND ANALYTICS QUESTION

The data is derived from Medicare Part D and includes primarily two sources:

- CCW - The CMS Chronic Conditions Data Warehouse (CCW) provides researchers with Medicare and Medicaid beneficiary, claims, and assessment data linked by beneficiary across the continuum of care. In the past, researchers analyzing data files were required to perform extensive analysis related to beneficiary matching, deduplication, and merging of the files in preparation for their study analysis. With the CCW data, this preliminary linkage work is already accomplished and delivered as part of the data files sent to researchers. The Chronic Conditions Data Warehouse (CCW) is a research database designed to make Medicare, Medicaid, Assessments, and Part D Prescription Drug Event data more readily available to support research designed to improve the quality of care and reduce costs and utilization.
- BRFSS - The Behavioral Risk Factor Surveillance System (BRFSS) is a state-based system of telephone health surveys that collects information on health risk behaviors, preventive health practices, and health care access primarily related to chronic disease and injury. The survey was established in 1984. Data are collected monthly in all 50 states, Puerto Rico, the U.S. Virgin islands, and Guam.

Key indicators include:



Data Properties

| Name | Data Type | Classification | Model Type | Format |
|---|---|---|---|---|
| Area name | Character | Category | Discrete | $94 |
| Avg_NUM_Opioid_Presc_BY_Prov_LOG | Numeric | Measure | Continuous | Numeric |
| Bachelor_s_degree_or_higher_ | Numeric | Measure | Continuous | Numeric |
| County | Character | Category | Discrete | $CHAR96 |
| County Economic Status, FY 2015 | Character | Category | Discrete | $CHAR24 |
| EDUCATION_RATE2 | Numeric | Measure | Continuous | Numeric |
| Median_Household_Income_LOG | Numeric | Measure | Continuous | Numeric |
| Opioid_Prescribing_Rate | Numeric | Measure | Continuous | Numeric |
| Physical_or_mental_unhealthy_LOG | Numeric | Measure | Continuous | Numeric |
| Poverty_Percent_All_Ages | Numeric | Measure | Continuous | Numeric |
| POVERTY_RATE | Numeric | Measure | Continuous | Numeric |
| Region | Character | Category | Discrete | $18 |
| State_Code | Character | Category | Discrete | $22 |
| Unemployment_rate_LOG | Numeric | Measure | Continuous | Numeric |
| VA7_FIPS | Character | Category | Discrete | $400 |
| year_char | Character | Category | Discrete | $24 |
| YEAR_DT | Date | Category | Continuous | MMDDYYYY |

Name of the dataset and data variables:

- cms_sdoh_wide_data3.sas7bdat.
- It can be retrieved using the Data Management Flow or by selecting the dataset in the Visual Analytics application itself.

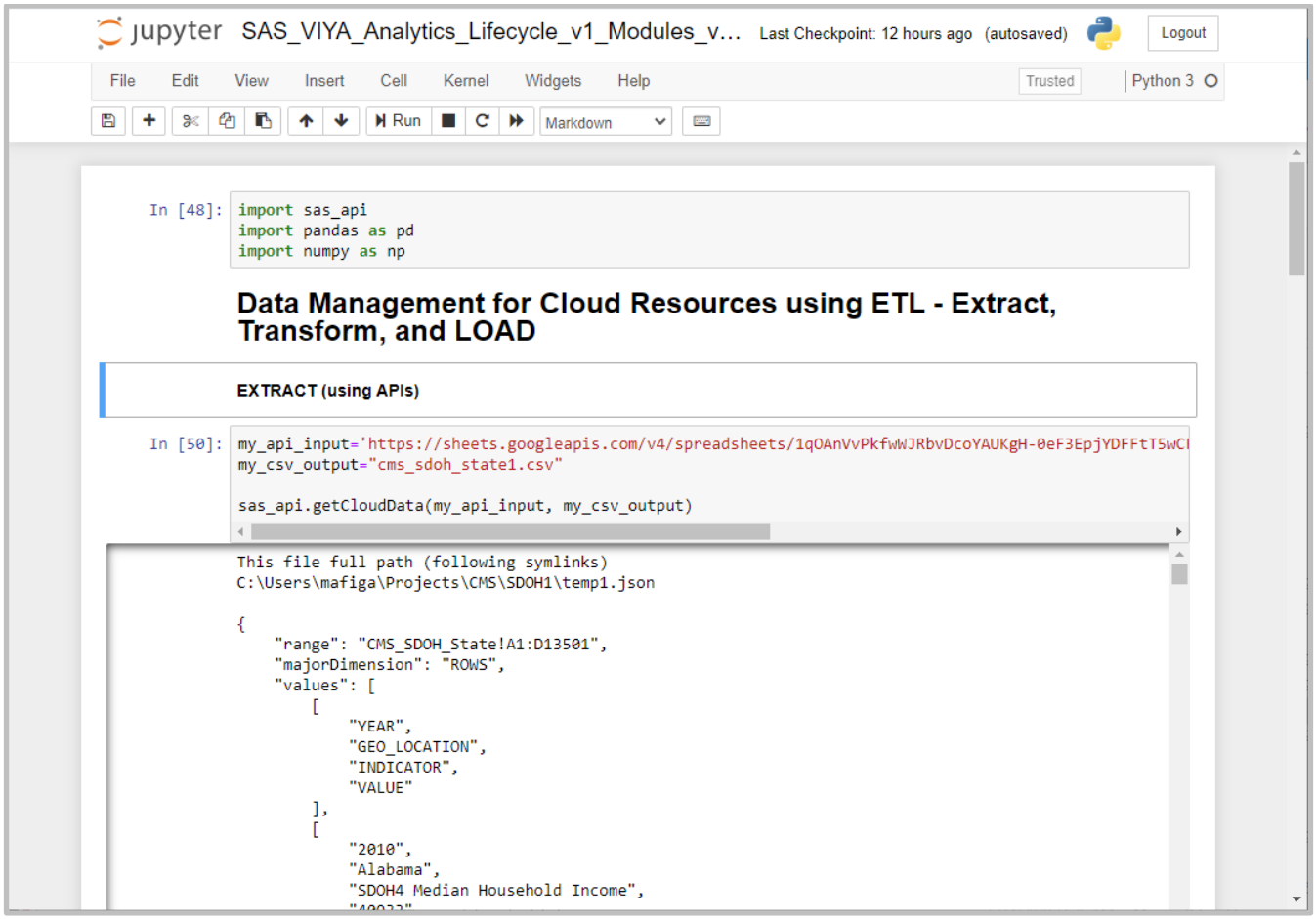
The Analytic question that we will pursue is:

- How can Jupyter Notebooks and Viya and SAS Visual Statistics be rapidly deployed to develop a strategy to help understand Medicare Part D program vulnerabilities related to the opioid crisis? How can we target specific socio-economic groupings?

# DATA MANAGEMENT

## STEP 0: GETTING STARTED

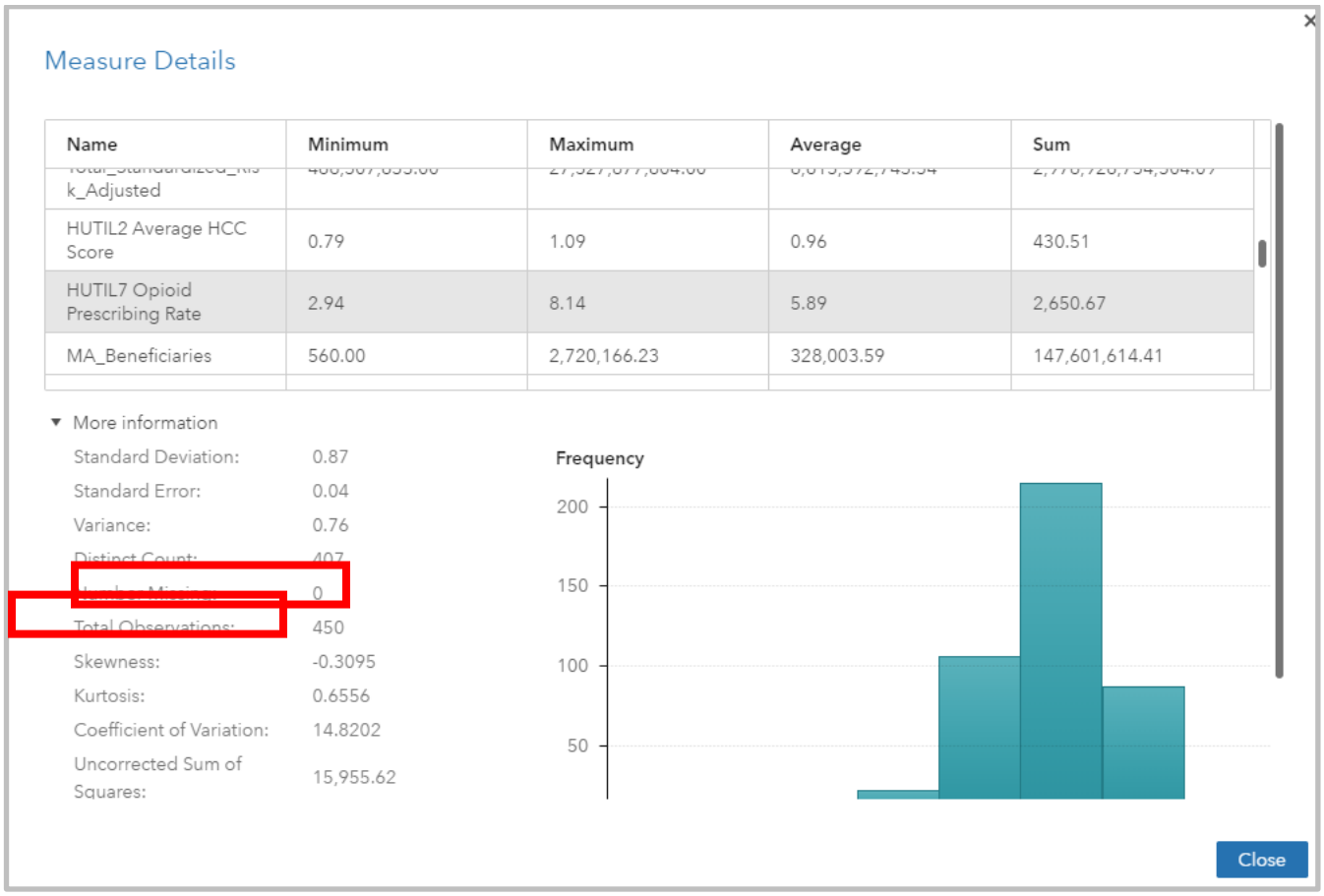
Data Management is the process of extracting, transforming, and loading the data. In our example, we are starting with a developed analytic dataset. We have already performed much of the initial analytic file creation including merging from the different data sources. We begin this process with the dataset as it exists in the link provided above. Our steps follow this process in the Jupyter notebook:




Once the file has been loaded, we begin Data Management tasks including assessing variables to see if the data is complete and if variables to be used for analysis are of a normal distribution.

These features are available by clicking on the down arrow in the left pane:


We also know that some of the statistical tests we will be running assume a normal distribution of the variables. SAS Studio and Viya makes this process easy with the "Measure Details" after selecting the down-arrow as shown above. You will get something like this for each variable to assess the completeness of the variables:




Notice the "Number Missing" for Opioid Prescribing Rate.

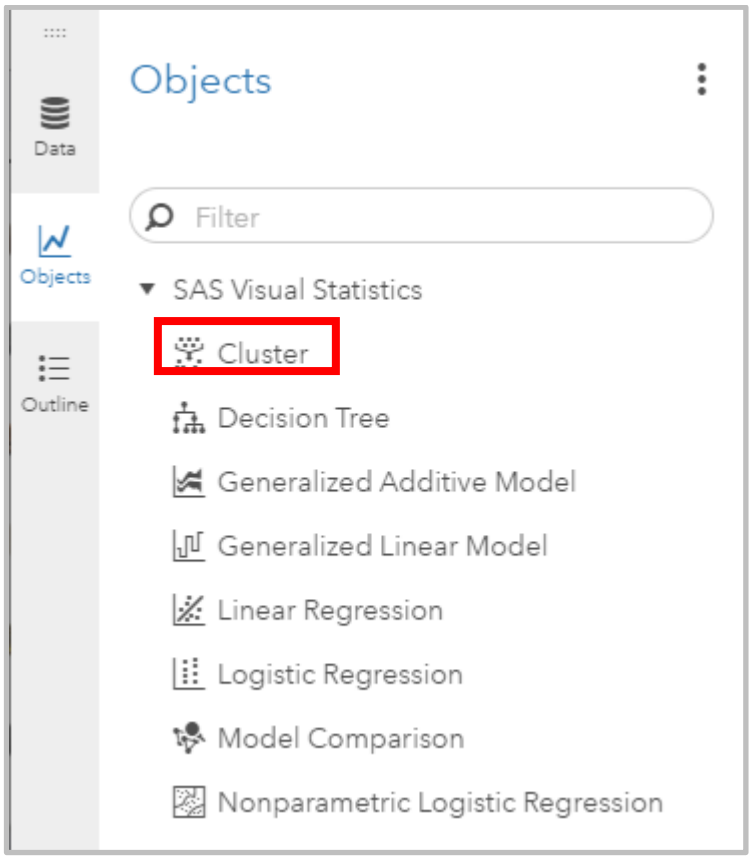
## VISUALIZATIONS

### STEP 1: GETTING STARTED WITH A CLUSTER EXPLORATION

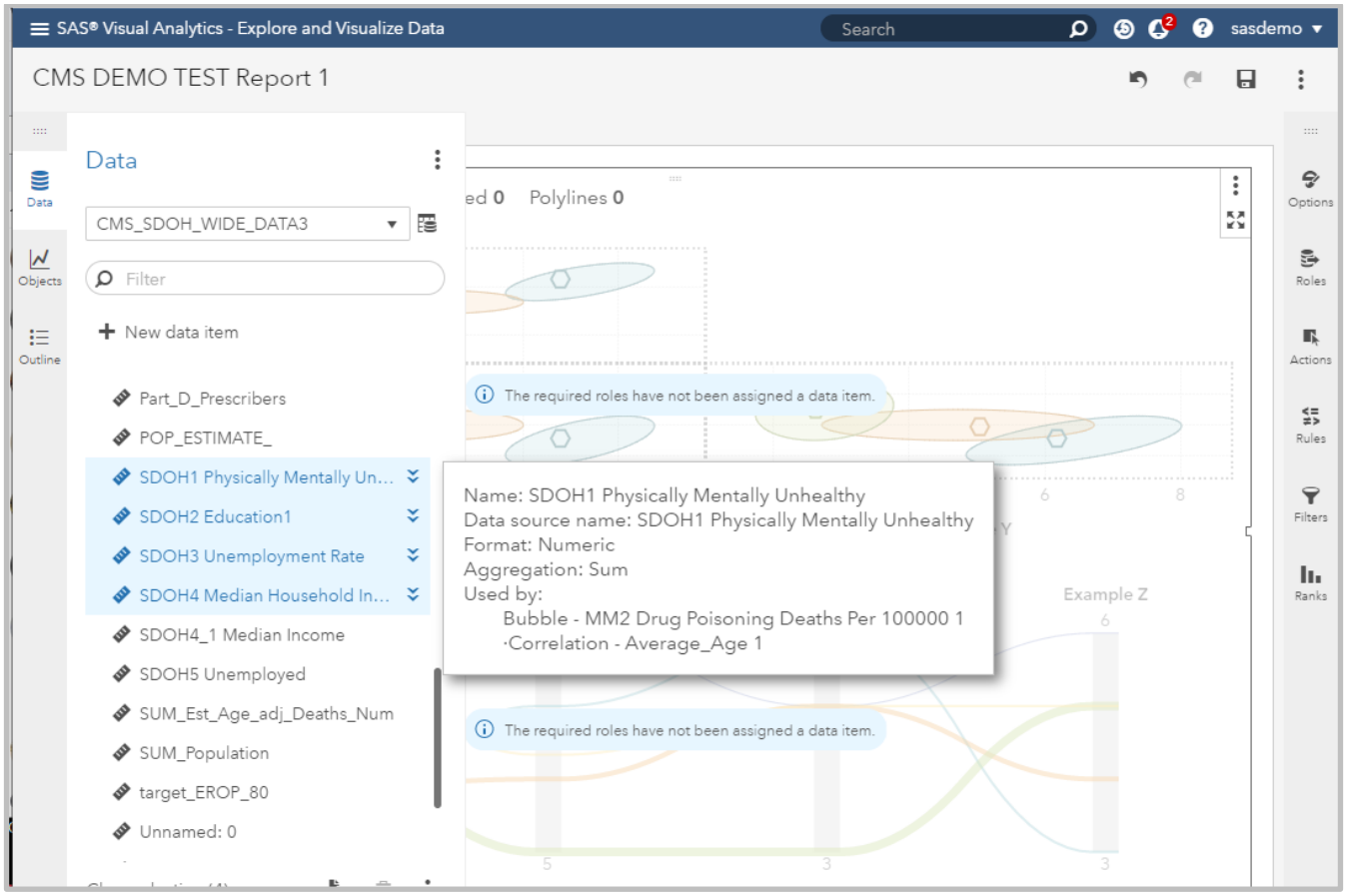
Now that the data has been loaded, we can begin by exploring it. Let's begin with a K-Means cluster which is an unsupervised maching learning algorithm. Clustering is a method of data segmentation that puts observations into

groups that are suggested by the data. The observations in each cluster tend to be similar in some measurable way, and observations in different clusters tend to be dissimilar. Observations are assigned to exactly one cluster. From the clustering analysis, you can generate a cluster ID variable to use in other explorations.

The cluster visualization is available from the lef pane:



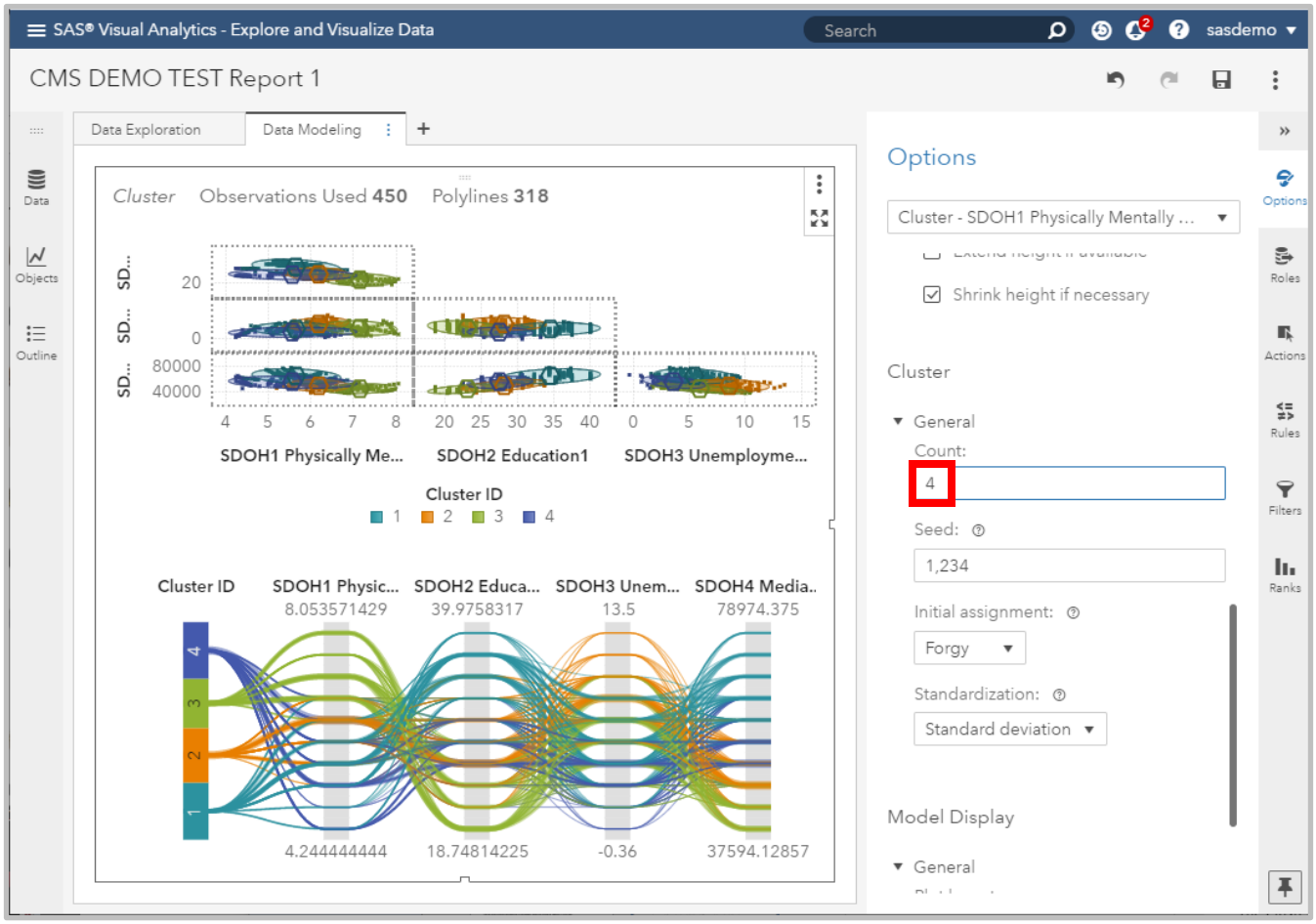

At this point, the graphical user interface should look like this:




The left pane contains the measure in the dataset. Select the following measure in this order:

- SDOH1 Physical_or_mental_unhealthy_LOG
- SDOH2 EDUCATION_RATE2
- SDOH3 Unemployment_rate_LOG
- SDOH4 Median_Household_Income_LOG

Then, drag and drop them into the middle pange or canvas. On the far right pane change the “Properties” to 4. This is shown here (as is the output from the cluster):



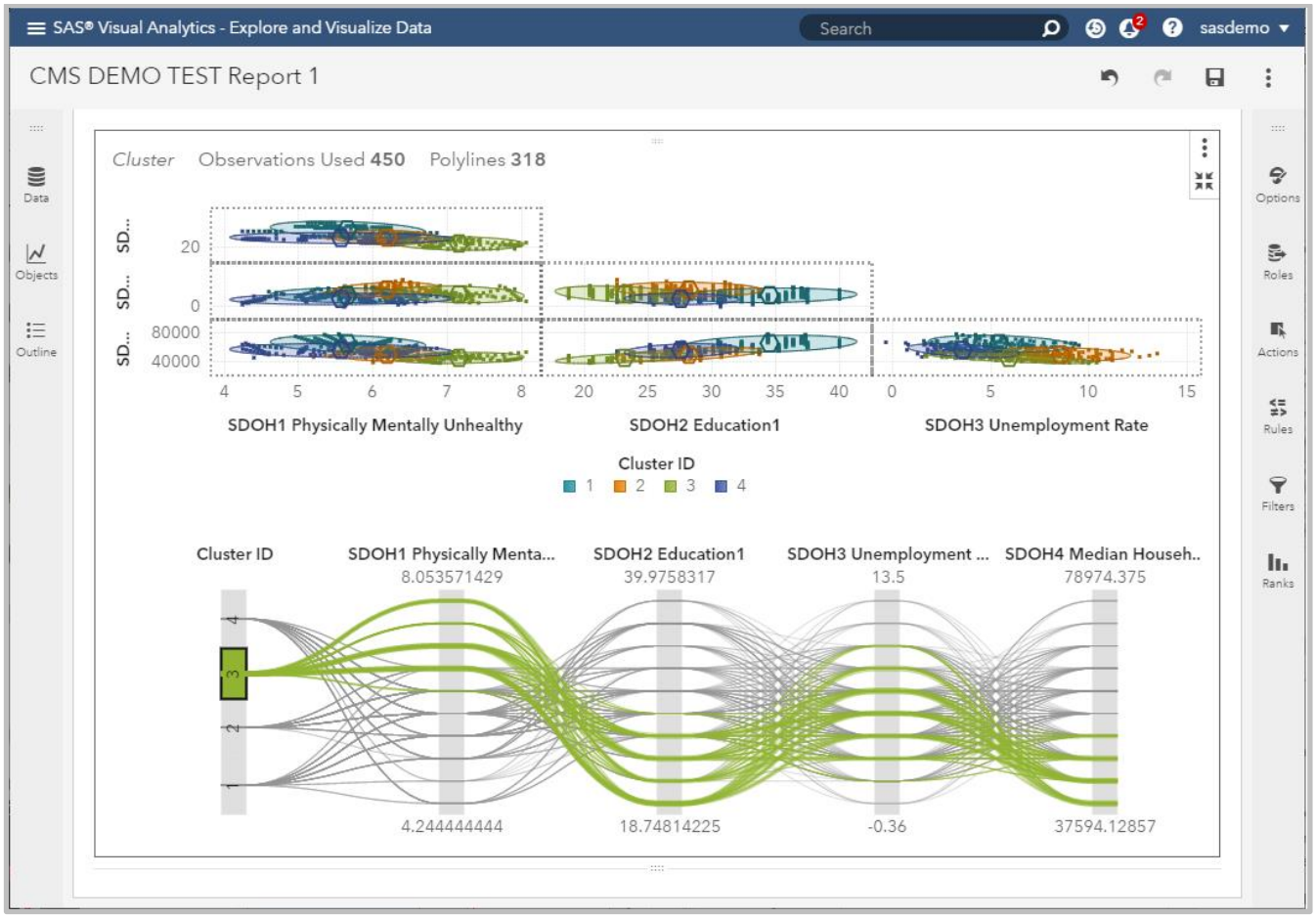

Notice that there are two section to the cluster: the cluster matrix, and the parallel coordinates plot. In the upper-right hand corner of the the parallel coordinates plot, click on the following icon: 

The Cluster Matrix displays a two-dimensional projection of each cluster onto a specified number of effect pairs. These projections are useful for spotting cluster similarities and differences within the plotted effect pairs.

We will focus our attention on the parallel coordinates plot, however. The Parallel Coordinates plot shows patterns in the data and clusters. In this plot, the cluster ID is on the far left, and each variable is a column with its binned range of values displayed vertically. Color-coded polylines are drawn from each cluster and show which range of values the cluster contains for every variable.

You should now see this:


Click on “Cluster ID” 0 in the far left (blue) to see this:




This represents counties with the following attributes:

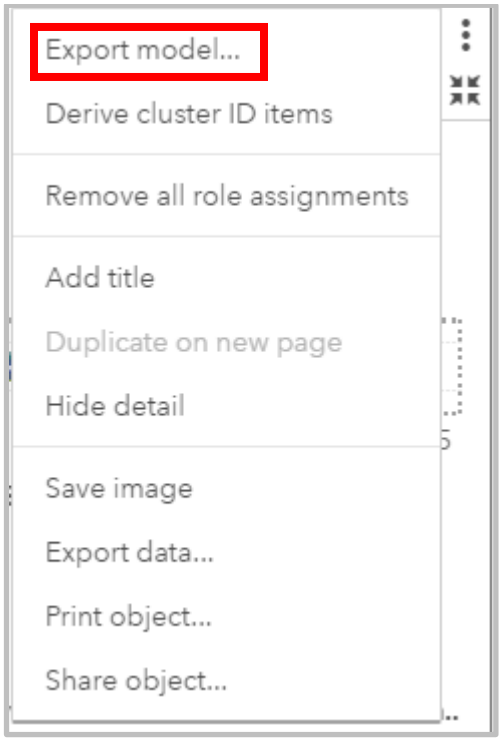
- Physical_or_mental_unhealthy_LOG (HIGH)
- EDUCATION_RATE2 (LOW)
- Unemployment_rate_LOG (HIGH)
- Median_Household_Income_LOG (LOW)

Next we will export the score code for the cluster and update our dataset with it.

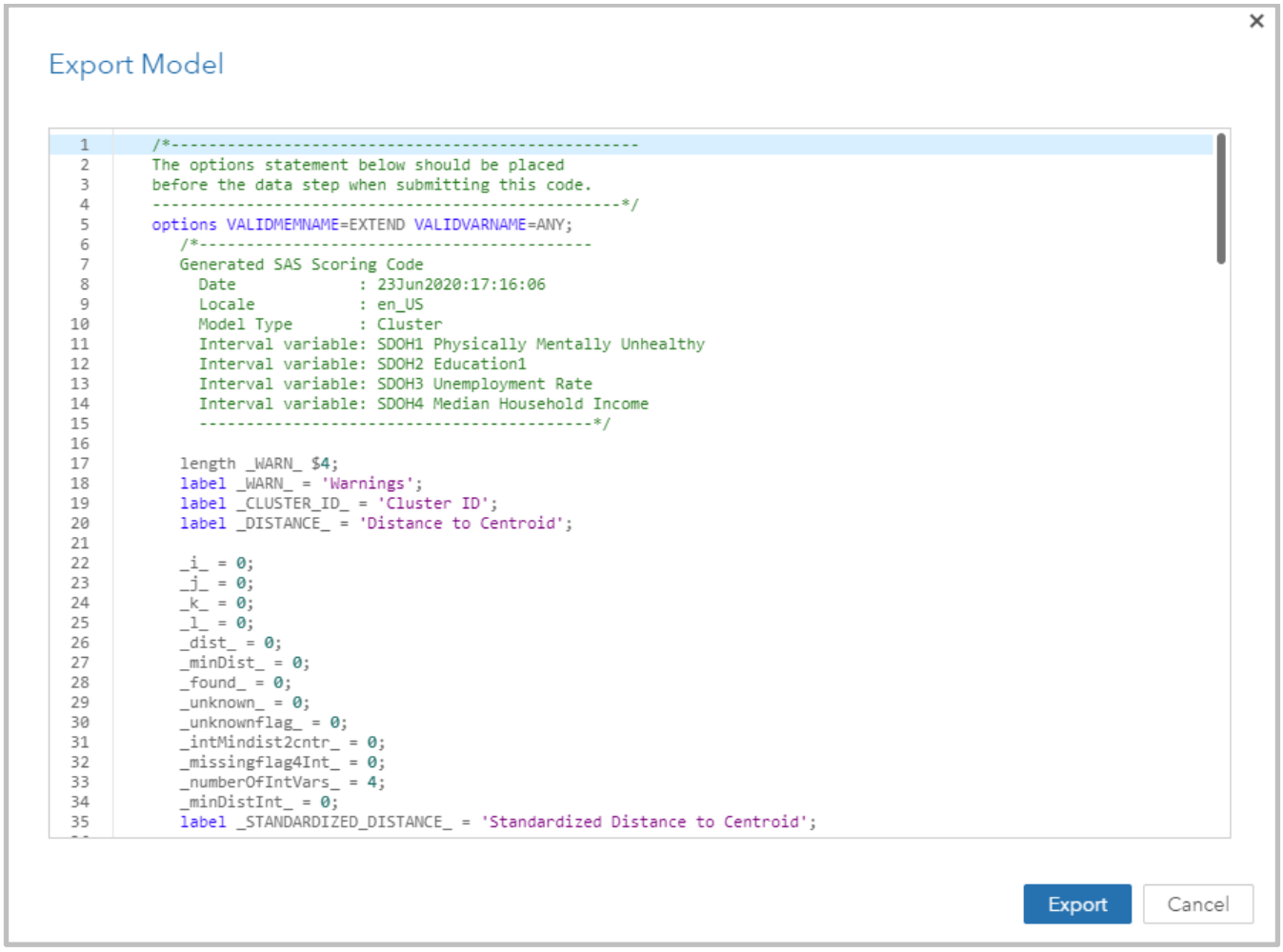
## STEP 2: EXPORTING SCORE CODE AND SCORING

To score a dataset you will first need to export the score code. From the top of the cluster visualization, select the down arrow.

You will see a list of possible actions:



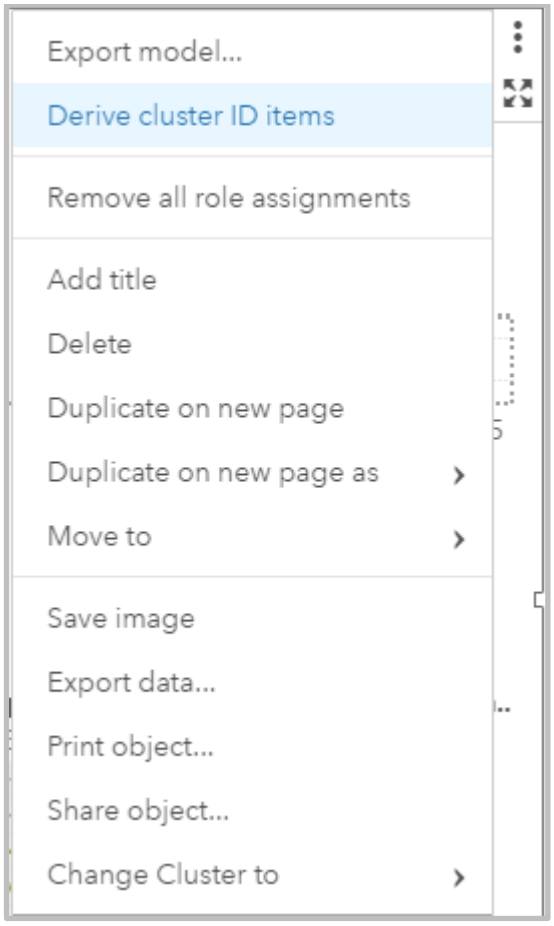

Select “Export Score code..” and place the score code on the server.




Next, you can use the “deployClusterModel” macro available in the “macros” folder in the GITHUB location:

https://github.com/sasgovernment

Another approach is to simply “Derive cluster ID items”:

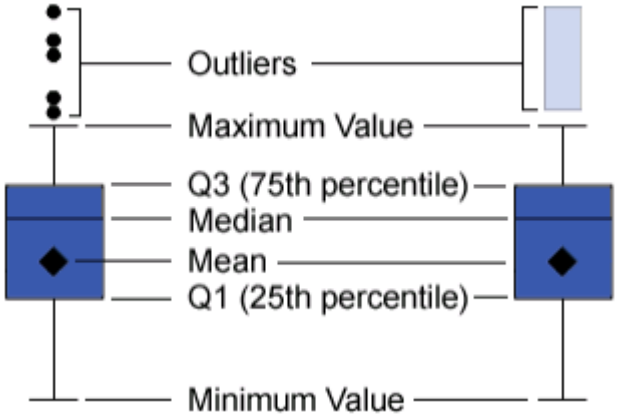



## STEP 3: CREATING A BOX PLOT WITH MEDICARE OPIOID PRESCRIBING RATE

The updated dataset with CLUSTERID is “cms_sdoh_wide_data3”.

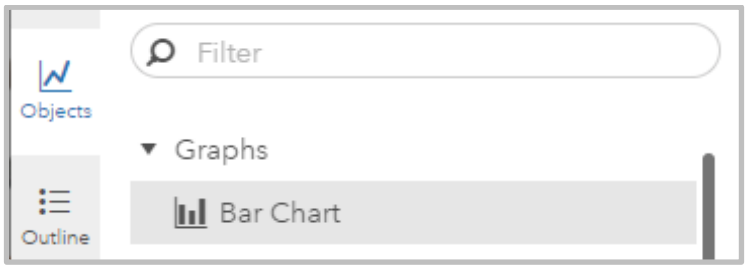
To understand how opioid_prescribing_rates may be distributed among our newly created clusters, we use a box plot.

A box plot displays the distribution of data values by using a rectangular box and lines called “whiskers.”




The bottom and top edges of the box indicate the interquartile range (IQR). That is, the range of values that are between the first and third quartiles (the 25th and 75$^{th}$ percentiles). The marker inside the box indicates the mean value. The line inside the box indicates the median value.

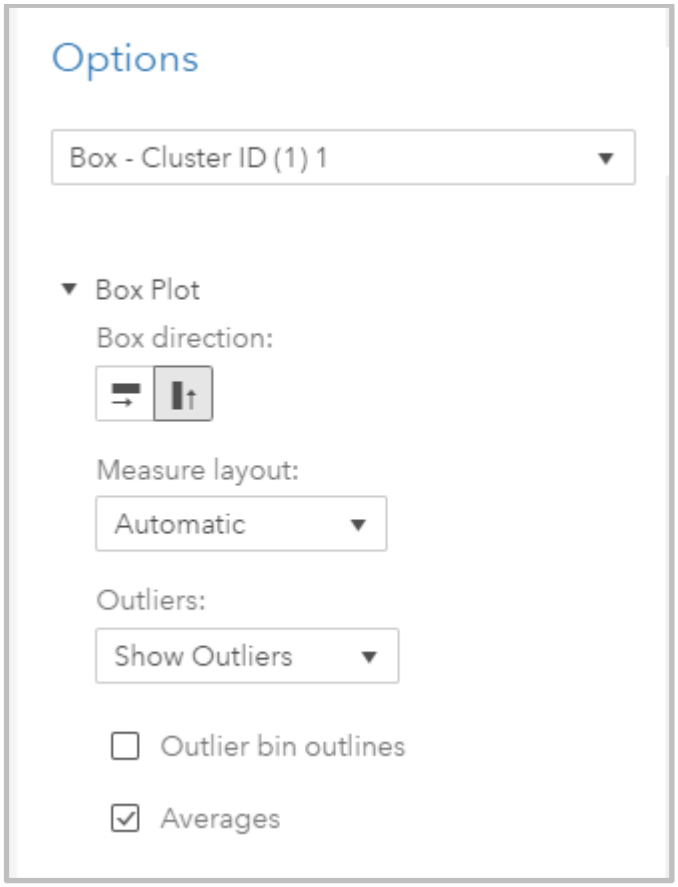
First, select Visualization→New. Then, select the following icon from the top:




You can then produce the box plot by selecting the following variables in “Roles”:


(Note that you have to convert ClusterID to a “Category” type to use it in the box plot.)

Change the Properties to accommodate and show outliers and averages:




You will now have this visualization:


As mentioned in the E-poster (also provided in the GITHUB for reference),

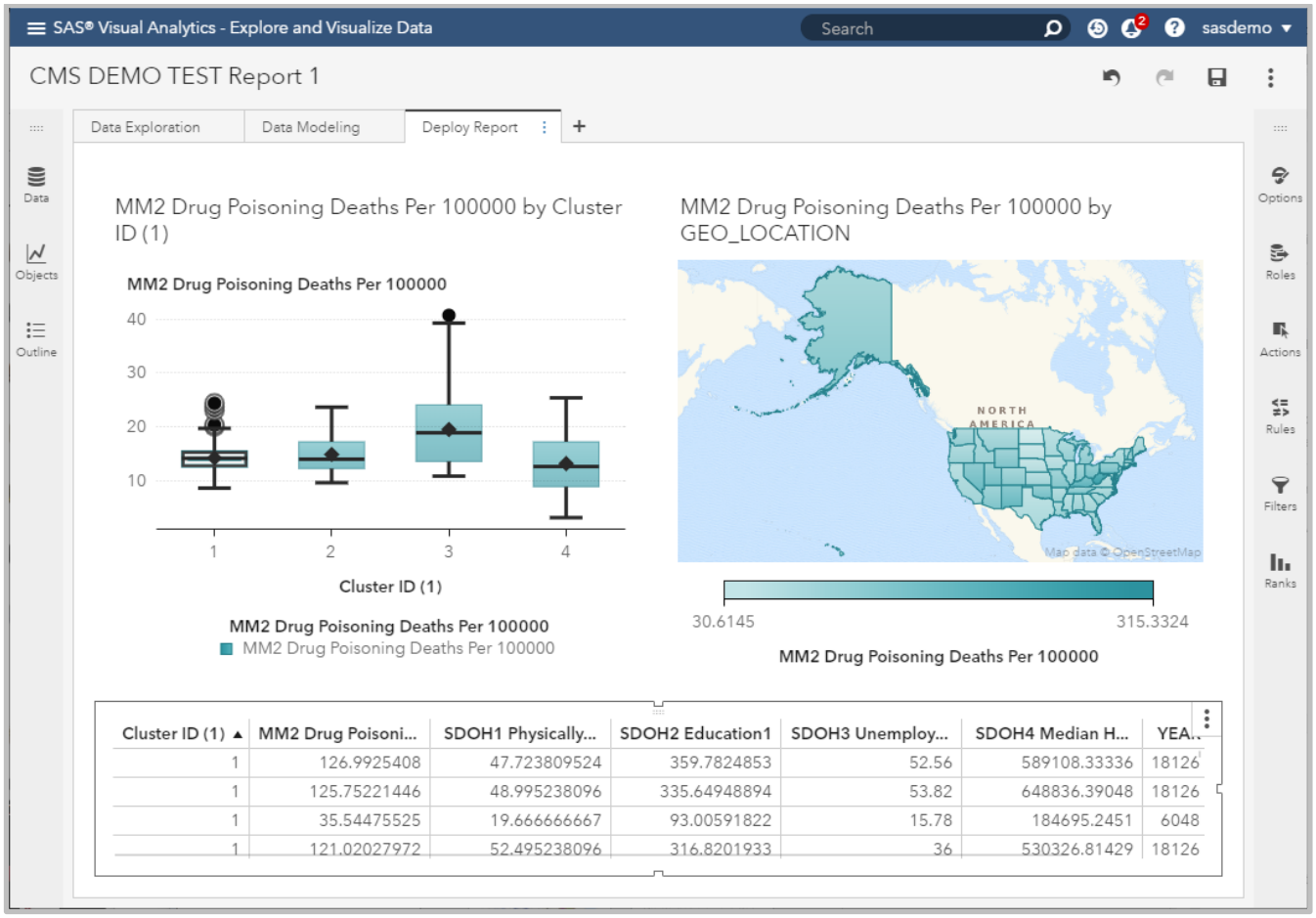
Interpretation of this finding could be that the Appalachian area does not have enough variance in socio-economic status or that the opioid epidemic in the Part D program does not differ between socio-economic groupings.

## STEP 4: DEPLOYING THE CLUSTER MODEL IN A REPORT

In a similar fashion as in Step 1, select the folloiwng variables to produce a new cluster:

- Opioid_Prescribing_Rate
- Median_Household_Income_LOG
- Avg_NUM_Opioid_Presc_BY_Prov_LOG
- Unemployment_rate_LOG

After following the instructions from Step 1 and with the new variables in this Step, you'll have something like this:




## CONCLUSIONS

Using Jupyter Notebooks (Python) and Viya, users can enhance the analytic power of their data, explore new data sources, investigate them, and create visualizations to uncover relevant patterns. Users can then easily share those visualizations in reports. In traditional reporting, the resulting output is well-defined up-front. That is, you know what you are looking at and what you need to convey. However, data discovery invites you to plumb the data, its characteristics, and its relationships. Then, when useful visualizations are created, you can incorporate those visualizations into reports that are available on a mobile device or in the viewer.

## DEFINITIONS

**Correlation** - Correlation is a measure of association between two variables. The strength of the relationship is described as a value between -1 and 1. The closer the value is to -1 or 1, the stronger the relationship. The closer the value is to 0, the weaker the relationship. The colors in the correlation matrix shows the relationship in absolute terms, either weak (0)

or strong (1, -1). The actual value of the correlation appears in the tooltip and the results table. Double click or exploring a cell in the matrix will allow you to see a plot of the regression line.

**Linear Regression** - A linear fit line is the straight line that best represents the relationship between two variables. If the points on the scatter plot are tightly clustered around the line, then it likely provides a good approximation for the relationship. If not, another fit line should be considered to represent the relationship. If outliers (points which are distant from the rest of data) are present, they can have a strong influence on the slope of the line, and those points should be examined more closely.

**Bubble Plots** - A bubble plot displays the values of at least three measures by using differently sized plot markers (bubbles) in a scatter plot. The values of two measures are represented by the position on the plot axies, and the value of the third measure is represented by the marker size. You can create animated bubble plots to display changing data over time.

**Box Plots** - A box plot displays the distribution of data values by using a rectangular box and lines called “whiskers.”