SAS Visual Analytics Step-by-Step Tutorial: Exploring Medicare (CMS) Part D Opioid Data

Last update: January 17, 2018

Contents

[Purpose 1](#_Toc504407524)

[Data and analytics question 2](#_Toc504407525)

[Data management 4](#_Toc504407526)

[Step 1: Getting started 4](#_Toc504407527)

[Visualizations 4](#_Toc504407528)

[Step 1: Getting started 4](#_Toc504407529)

[Step 2: Creating a correlation matrix 5](#_Toc504407530)

[Step 3: Creating and Examining a linear regression 7](#_Toc504407531)

[Step 4: Examining and creating a bubble plot 8](#_Toc504407532)

[Step 5: Creating a box plot 11](#_Toc504407533)

[Conclusions 13](#_Toc504407534)

[Definitions 13](#_Toc504407535)

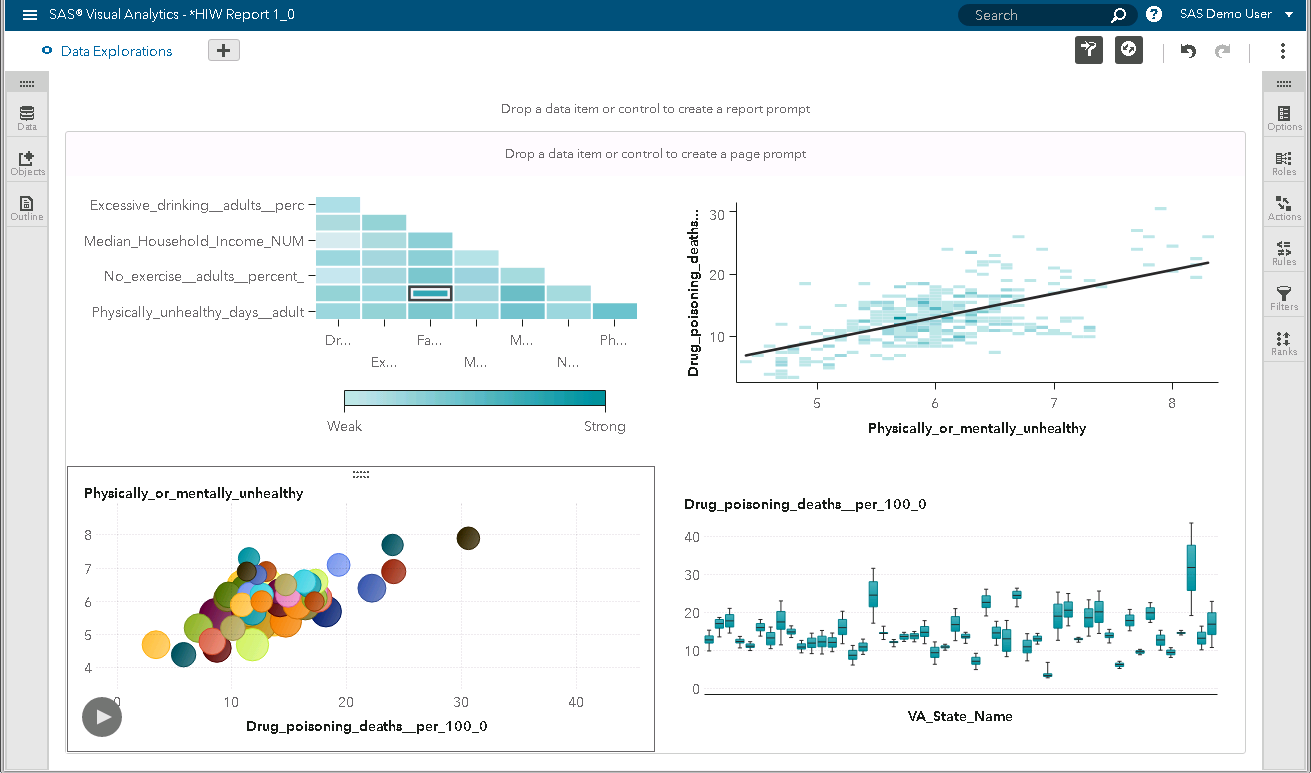
# Purpose

This tutorial is meant to jump start one’s familiarity with SAS Visual Analytics and SAS Visual Statistics. SAS Visual Analytics 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>

The ultimate goal is to produce an exploration as follows:



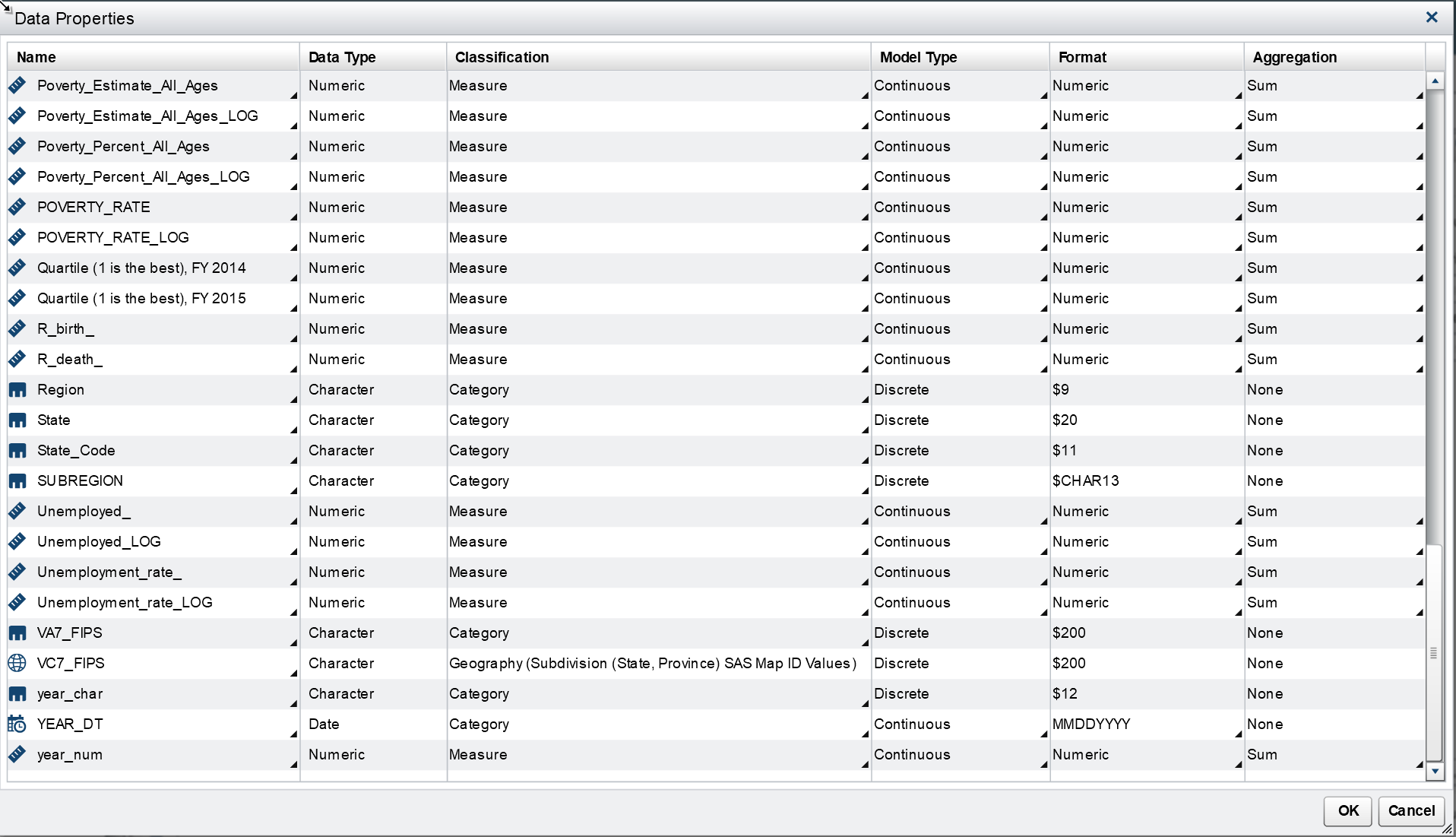
# Data and analytics question

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.
* NPPES - The National Plan and Provider Enumeration System is responsible for processing new NPI applications and processing changes of information for previously enumerated providers. Per the Centers for Medicare and Medicaid Services (CMS), the NPI Enumerator is not permitted to provide guidance on regulatory policy issues. These issues include but are not limited to questions related to subparting, sole proprietorship, and determining who is required or not required to obtain an NPI.
* 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:

* Opioid Prescribing Rate
* Part D Opioid Prescribers
* Opioid Claims
* Household Median Income
* Physically or Mentally Unhealthy Days
* Education Rate
* Unemployment Rate
* Average Opioid Prescription Per Provider
* In addition, the dataset contains other variables including:



Name of the dataset and data variables:

* SGF\_ARC\_2015I\_WIDE\_REGION2.sas7bdat.
* It can me retrieved using the Data Management Flow or by selecting the dataset in the Visual Analytics application itself.

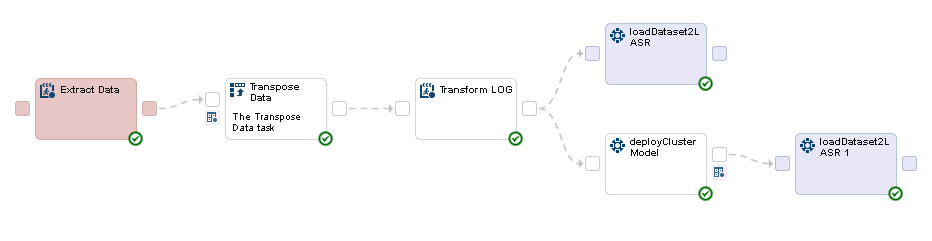
The Analytic question that we will pursue is:

* How can SAS Visual Analytics and SAS Visual Statistics be rapidly deployed to develop a strategy to help understand Medicare Part D program vulnerabilities related to the opioid crisis?

# Data management

## Step 1: 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:



We begin with loading the data from the link provided earlier in this document. After downloading the file, we [insert instructions on loading to LASR]

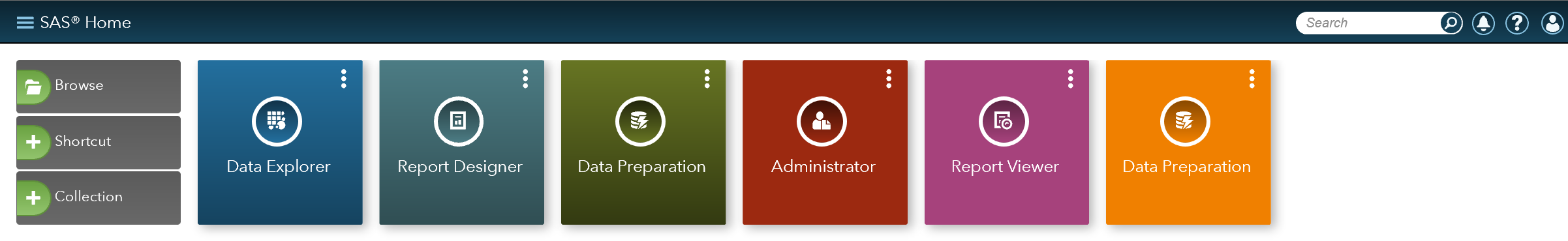
Once the file has been loaded, we begin Data Management tasks including assessing variables to see if the data is complete and if variables to be used for analysis are of a normal distribution. To assess the completeness of the variables, we do [insert process to validate dataset is complete]

We also know that some of the statistical tests we will be running assume a normal distribution of the variables. SAS Visual Analytics makes this process easy with the

# Visualizations

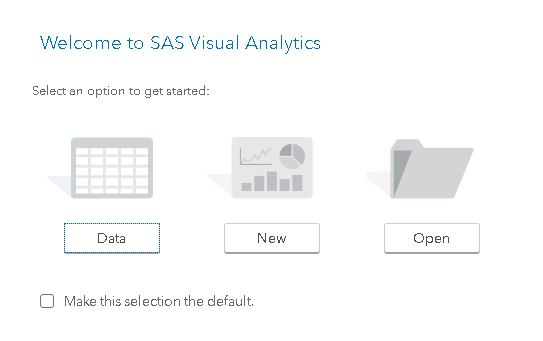
## Step 1: Getting started

After you sign in to SAS Visual Analytics using the standard sign-in window for SAS applications, you will see SAS Home (the home page). The home page enables you to create new content in SAS Visual Analytics. In addition, it enables you to access content that you and others have created.

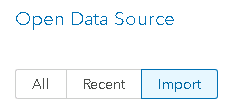


Select “Data Explorer”

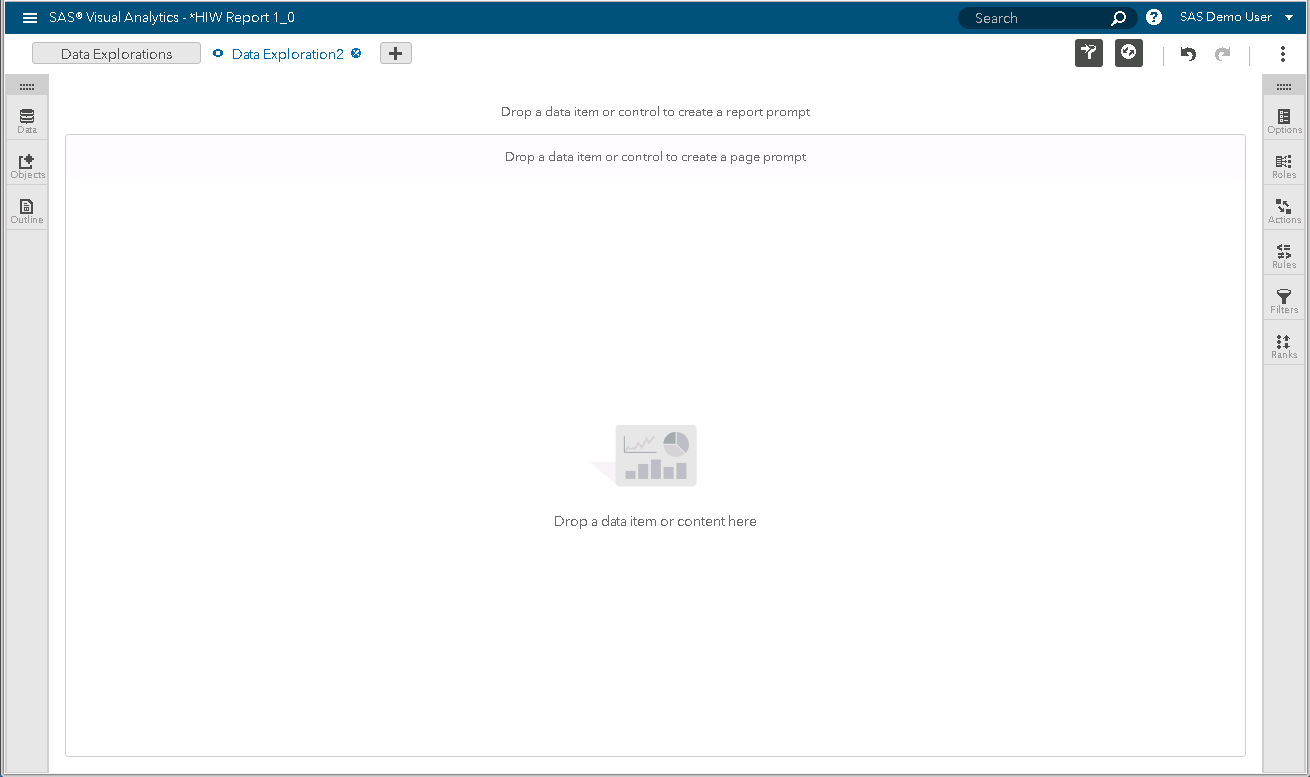
You will asked to “select an option to get started”. Choose “Data”:



Import “HIW\_DEMO\_DATA1” using the following GUI:



The following screen will appear:



The next section will allow us to gain some initial insights into the data.

## Step 2: Creating a correlation matrix

A correlation matrix displays the degree of correlation between multiple intersections of measures as a matrix of rectangular cells. Each cell in the matrix represents the intersection of two measures, and the color of the cell indicates the degree of correlation between those two measures.

The correlation values are calculated by using Pearson’s product-moment correlation coefficient. Correlation values are identified as weak, moderate, or strong as follows:

Weak

The absolute value is 0.3 or lower

Moderate

The absolute value is greater than 0.3 and less than or equal to 0.6

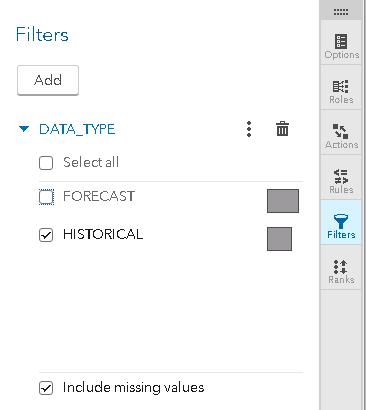
Strong

The absolute value is greater than 0.6

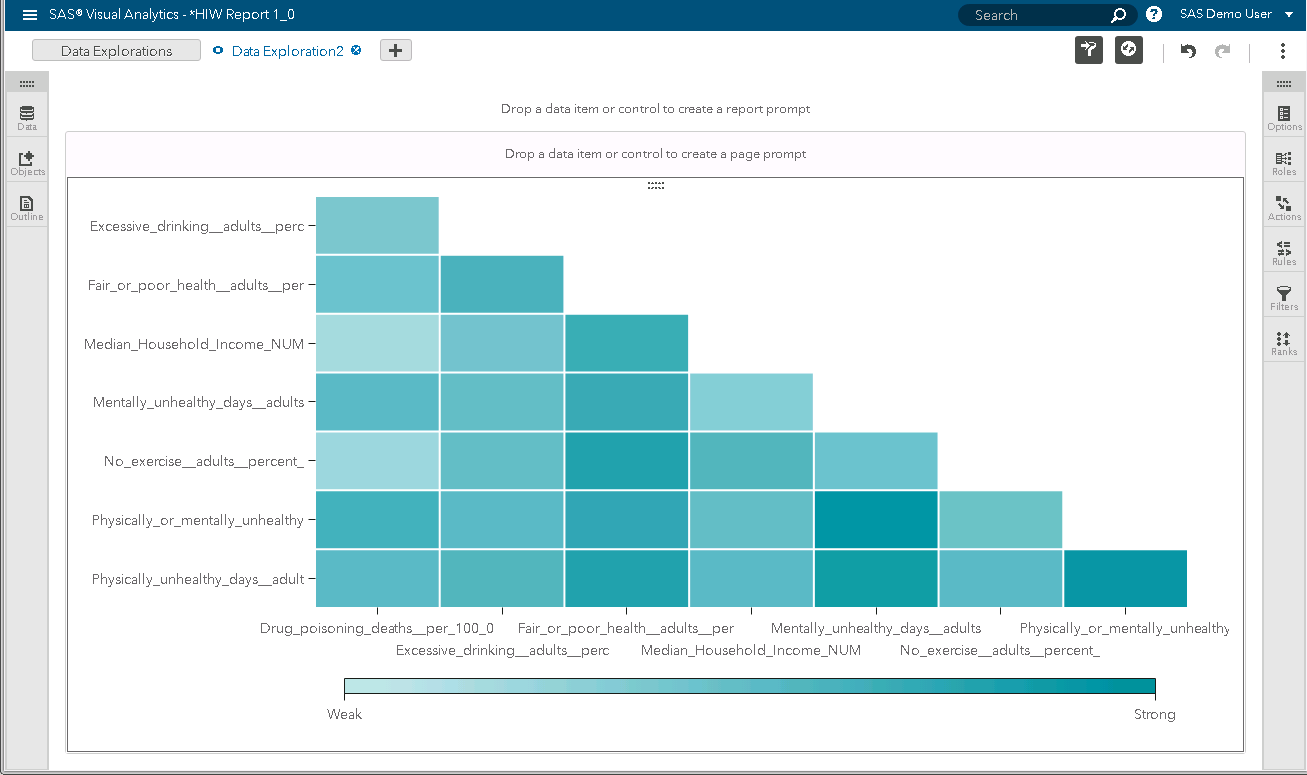
Choose a correlation matrix from the left pane:



Since we’re only interested in historical data, drag and drop the “DATA\_TYPE” variable in the Filter tab (far right) to examine only historical data:



Next select all “Measure” variables in Roles onto the canvas to produce the following:



Notice the strong correlation between “Drug Poisoning Deaths” and “Physically or Mentally Unhealthy Days”.

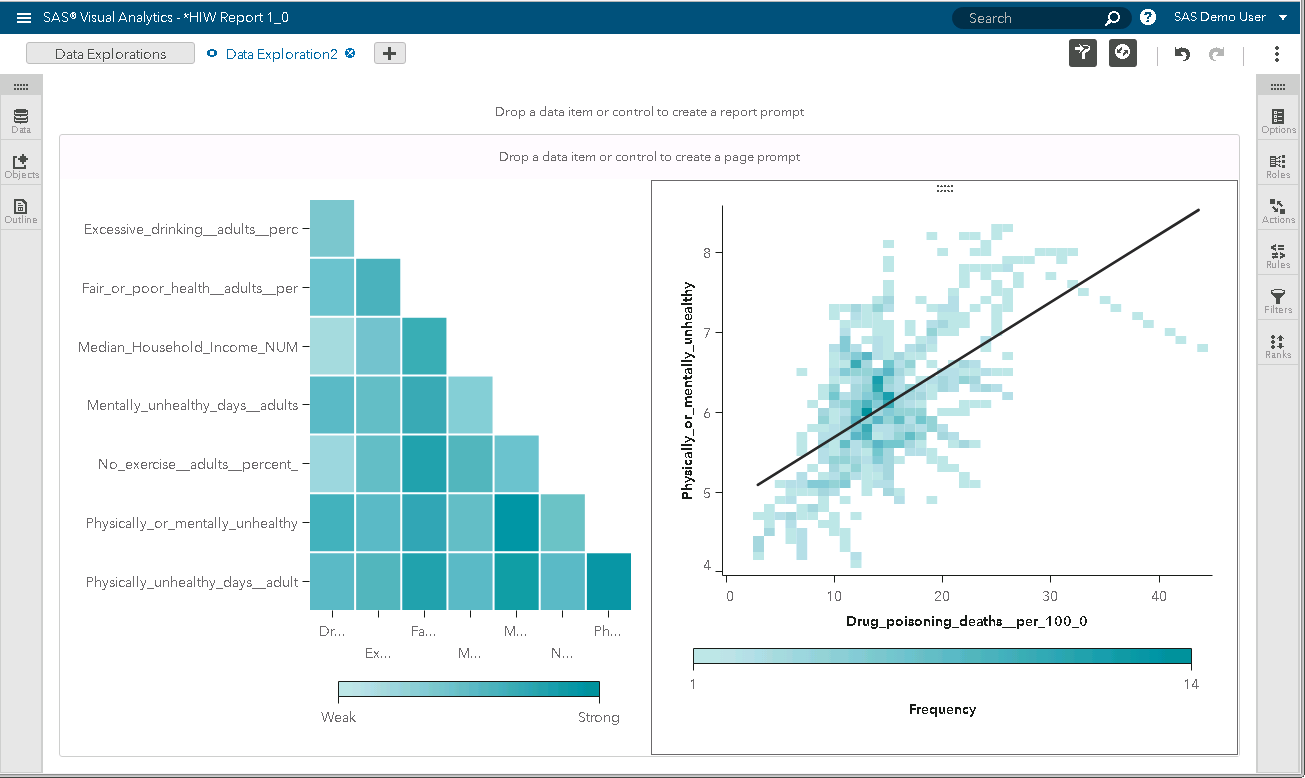
## Step 3: Creating and Examining a linear regression

Select the “Heat Map” visualization:



Use the variables “Physically or Mentally Unhealthy Days…” and “Drug Poisoning Deaths Per 100,000” as well as the option to overlay a linear regression line.

At this point, you will have this visualization:



SAS Visual Analytics has created a linear fit line from a linear regression algorithm. A linear fit line produces the straight line that best represents the relationship between two measures.

The slope of the line is important since this implies that as one variable increase so does the other.

With the insights gained from our analysis of “Drug Poisoning Deaths” and “Physically or Mentally Unhealthy Days” we can begin to create our first animated graph.

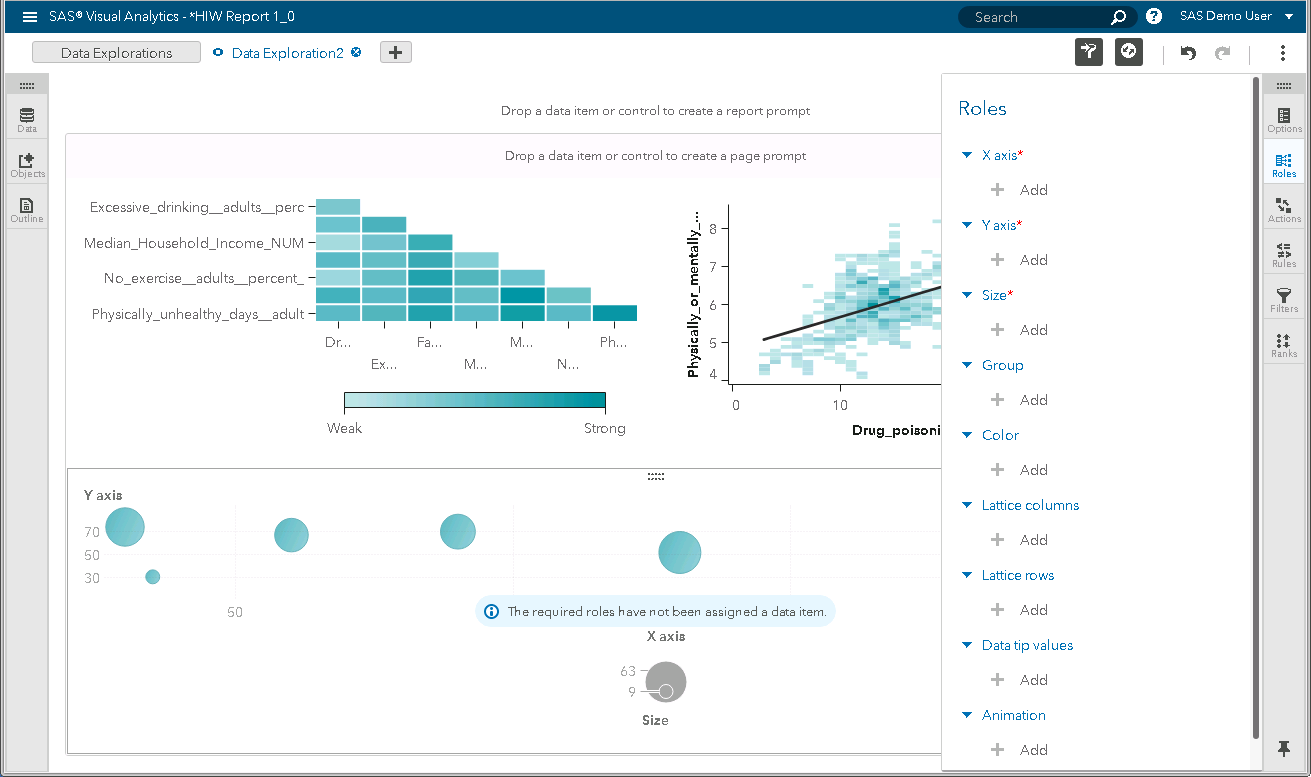
## Step 4: Examining and creating a bubble plot

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 axes, 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.

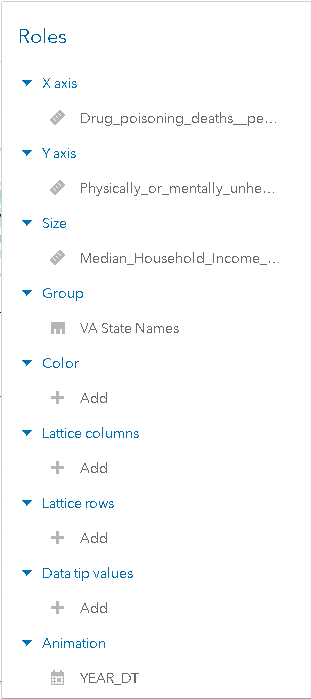
Now create a “Bubble Plot” icon from the left pane:



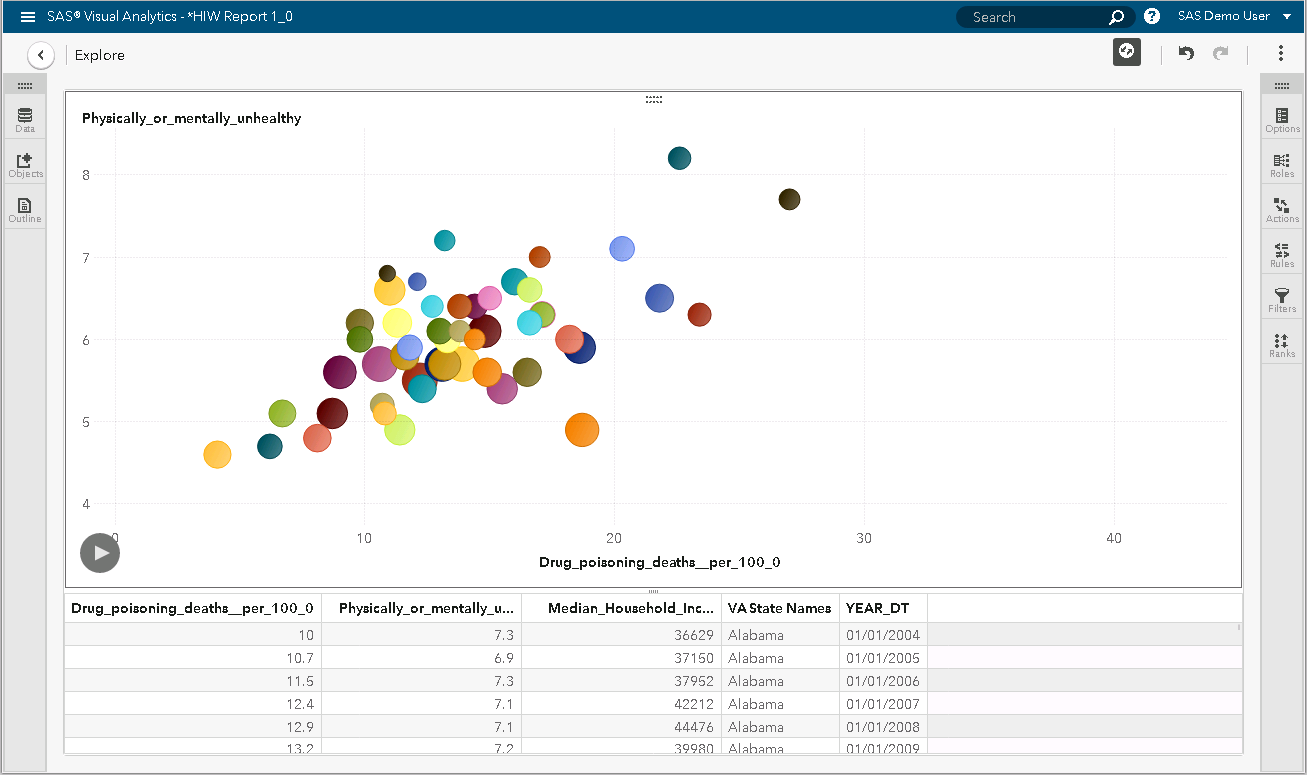
At this point, you should see the following to select X-axis and Y-axis coordinates:



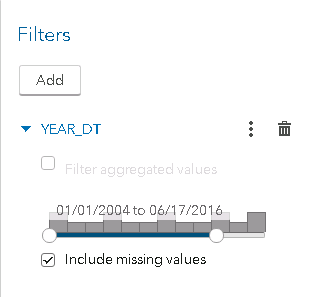
Select these variables in the “Roles” tab:



You will have produced the following:

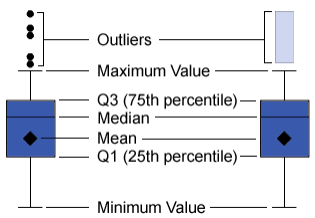


To view only historical data, you can add a filter, as follows:



## Step 5: Creating a box plot

Additionally, you can create a box plot to examine the distribution of “drug poisoning deaths” across US states.

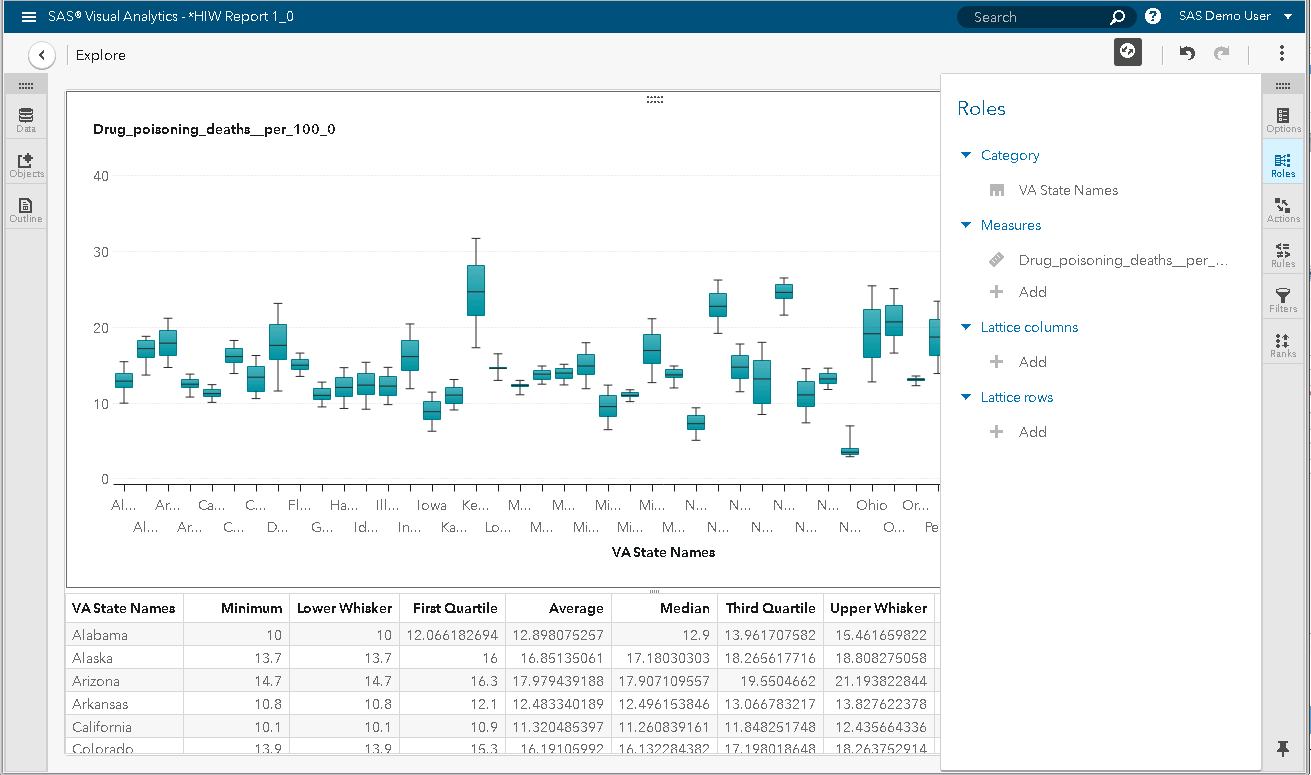
A box plot displays the distribution of data values by using a rectangular box and lines called “whiskers.”

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 75th 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 side pane:



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



# Conclusions

Using SAS Visual Analytics, 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 axes, 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.”