MIS 0855 – Data Science (Section 005) – Fall 2016
In-Class Exercise (Week 12) – Integrating Datasets

Objective: Analyze two data sets at the same time by combining them within Tableau.

Learning Outcomes:

- Identify common data between data sets that allow them to be connected.

- Generate a common field that facilitates connection by software such as Tableau.

- Analyze data from two different data sets once they are combined.

In this exercise, you'll be working with two data sets:

- 2012 Presidential Election Results by Congressional District (435 rows, House of Representatives only) adapted from the Daily Kos website. This provides the percentage of the vote given to Romney and Obama for each congressional district. It also has a field that lists who won that district.

- The demographic profiles of each current Congressperson (535 rows, House of Representatives and Senate) from the Measure of America project, part of the Social Science Research Council. The data set includes the political party, gender, race, and education level of the elected official (there's other data there too).

By combining these data sets, we can find out if there appear to be relationships between the demographics of the district-elected representative and how that district voted in the 2012 Presidential election.

Keep in mind that correlation does not always imply causation! When we see something that looks like a relationship, it doesn't necessarily mean that we understand the cause, if even if it's just a coincidence. But it still is interesting to look…

Part 1: Take a look at the data sets

1) Download the two data sets (2012 Presidential Election Results by District.xlsx and Portrait 113th Congress.xlsx) and save them to your computer. Remember where you saved them!

2) Open the "2012 Presidential Election Results by District" file in Excel and look at the data. You'll see an entry for each Congressional District (i.e., AZ-1, AZ-2, AZ-3…). Each state has at least one district, depending on the size of the population. It contains the percentage of the vote for Obama and Romney – it won't add up to 100% because there are always third-party and write-in candidates.

   You'll also see State and DistrictNo split into separate columns. We need to do this so we can do cool mapping things with Tableau later.

3) Now open the "Portrait 113th Congress" file in Excel and look at the data. Here you see a list of every elected representative and their demographic information.

Look at row 10 (the first row of data). Notice that DISTRICT (IF HOUSE) is just a number, instead of AL-1, like it was represented in the election results file. These different formats for district will make it impossible for Tableau to connect the data later – it won't be able to figure out that "Alabama 1" is the same as "AL-1." So we'll need to fix this before we do our analysis.

4) Close both Excel files.

Part 2: Create a common field to combine the data

As we've stated previously, the "2012 Presidential Election Results by District" data set represents districts this way:

	A	
1	District	
2	AK-1	
3	AL-1	
4	AL-2	
5	AL-3	
6	AL-4	
7	AL-5	
8	AL-6	

While the "Portrait 113th Congress" file represents districts this way:

STATE	DISTRICT (IF HOUSE)
Alabama	1
Alabama	2
Alabama	3
Alabama	4
Alabama	5
Alabama	6

We need to create an additional data column in one of the files that represents districts in the same way. We need a single column to do the matching, so we're going to modify the "Portrait" file to add an additional column with a single district label.

1) Open the "Portrait 113th Congress" file in Excel.

2) Note that there is a "State Lookup" tab. Click on that and you'll see abbreviations for all the states, listed in alphabetical order.

3) First, we will create a column with the correct state abbreviation for each row.
   Go back to the "Data" tab and scroll to column M. In cell M9, type "STATEABBR"

4) In cell M10, type the following formula:

   =VLOOKUP(B10,StateLookup!$A$1:$B$50,2)

   Remember, this means that it is using the value in B10 (the name of the state) to find the correct abbreviation, that the lookup table is in the StateLookup tab (StateLookups!$A$2:$B$50), and that the second column of that lookup table contains the two-letter state abbreviation (2).

5) You'll now see "AL" appear as the cell value.

6) Now we will create a column that combines the state abbreviation with the district number. In Cell N10, type "DISTRICTCODE"

7) In Cell N11, type the following formula:

   =CONCATENATE(M10,"-",C10)

   This builds a string of characters based on what's inside the parentheses. So here, we are taking the state abbreviation (M10), adding a dash ("-"), and then adding the district number (C10).

8) You'll now see "AL-1" appear as the cell value.

9) Now, carefully select Cells M10 through N544 (both columns!).

10) On the HOME tab, under Editing, select Fill/Down:
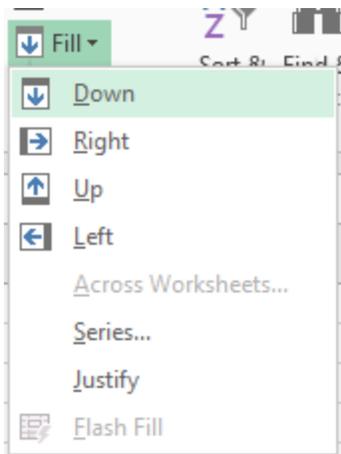
11) You'll now see values for STATEABBR and DISTRICTCODE all the way down to row 544:

WI	WI-3
WI	WI-4
WI	WI-5
WI	WI-6
WI	WI-7
WI	WI-8
WI	WI-...
WI	WI-...
WY	WY-1
WY	WY-...
WY	WY-...

12) We'll use the data in DISTRICTCODE (Column N) later to connect the two Excel workbooks, since now this looks exactly the same as "District" in the Election Results file.

13) Make sure you save the file!

Part 3: Start Tableau and open the data files

1) Start Tableau.

2) Click on "Microsoft Excel" under "To a file."

3) Navigate to the location where the data file "2012 Presidential Election Results by District" is stored and select it.

4) You'll see a list of Excel worksheets at the left side of your screen. These are all the sheets contained within the workbook. Drag the "Results By District" sheet to the workspace:

5) Click "Sheet 1" to Go to Worksheet.

6) Now let's connect to the second data file. Go to the Data menu and select "New Data Source"

Data	Worksheet	Dashboard	Story	Analysis	Map	Format	Server	Wi
New Data Source							Ctrl+D	
Paste Data							Ctrl+V	
Refresh All Extracts								
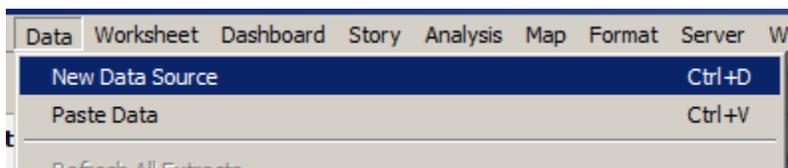
7) Click on "Excel."

8) This time, open the "Portrait 113th Congress" file.

9) Drag the "Data" worksheet to the workspace and click "Sheet 1."

10) You'll now see two data sources at the top left of your Tableau window:
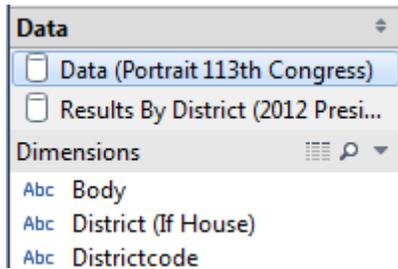
Part 3: Connect the data sources

We've opened both files, but they still are not connected. We know, however, that "Districtcode" in the "Portrait 113th Congress" file and "District" in the "2012 Presidential Election Results by District" file are in the same format (i.e., AL-1, AZ-3, PA-5).

We can use these fields with common data (Districtcode and District) to connect the data so we can use data from both sources in our analysis.

1) Go to the Data menu and select "Edit Relationships…"
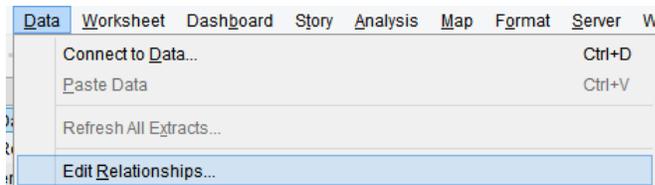
2) Select "Custom" and then click the Add… button
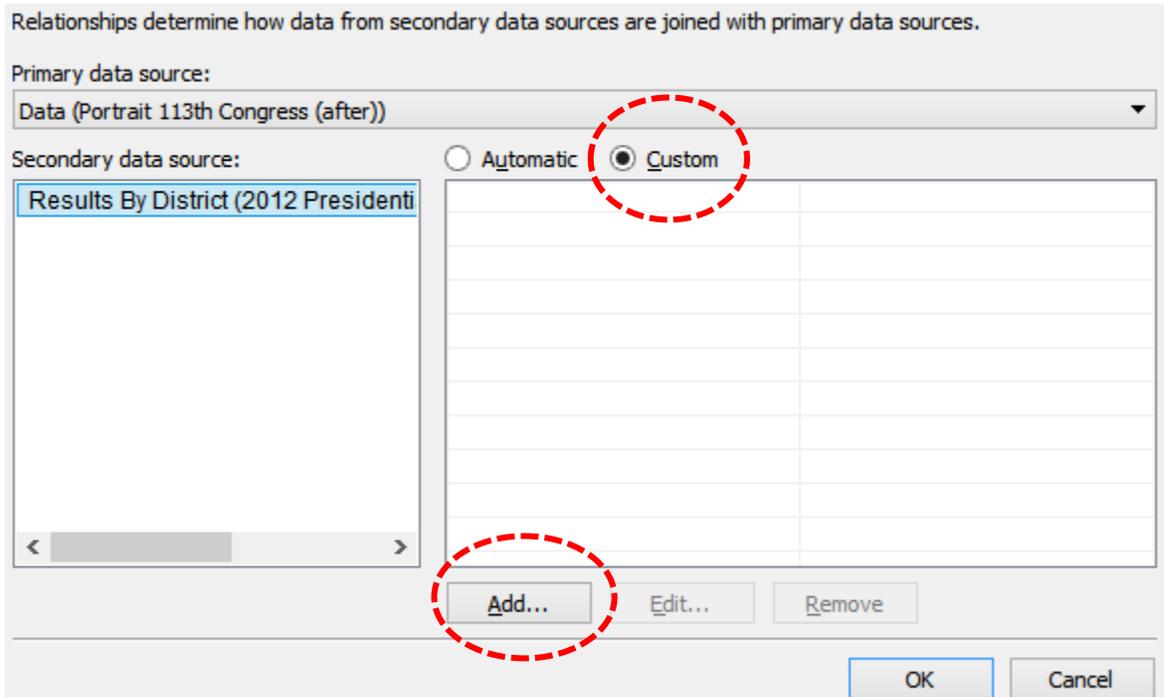
3) Select District and District code so that they are both highlighted, like this:
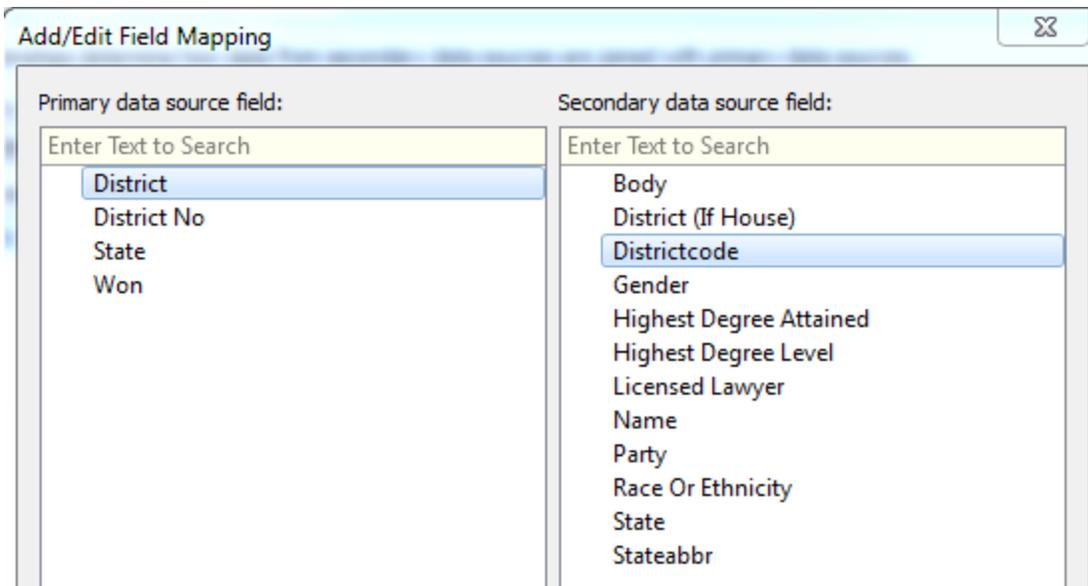
Then click OK.

4) You'll return to the previous dialog:
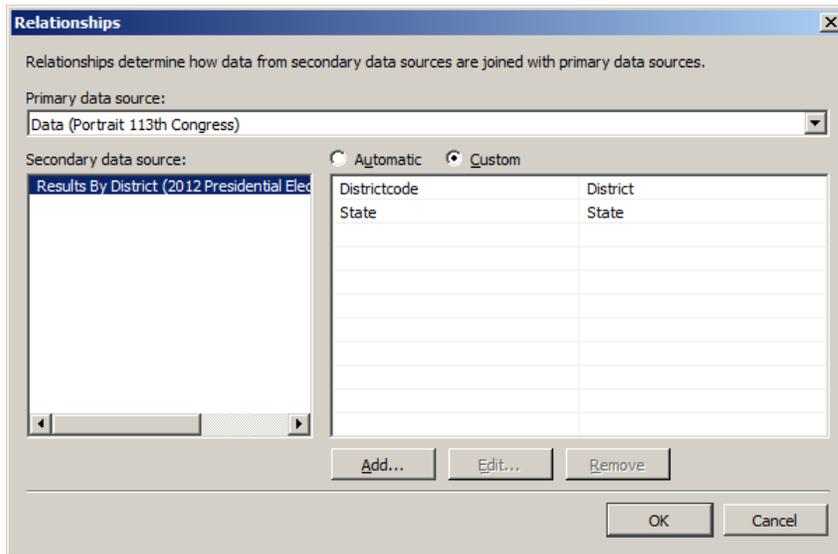
5) Remove the State→State relationship by clicking on that row and then clicking Remove.
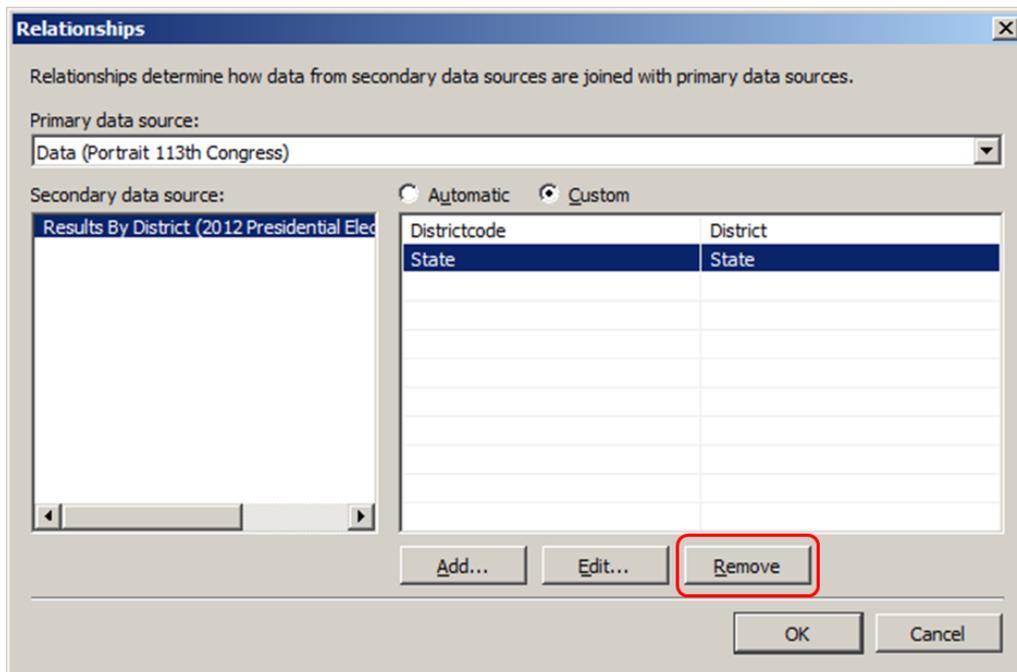
6) Click OK.

Part 4: Create a chart using data from both sources

1) Click on the "Data (Portrait 113th Congress)" data source:
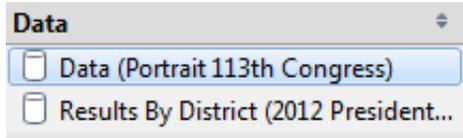
2) Drag the "Party" Dimension to the Columns shelf.

3) Click on the "Results by District (2012 Presidential Election Results…)" data source.

4) Drag the "Obama 2012" Measure to the Rows shelf.

5) You will see the following dialog:
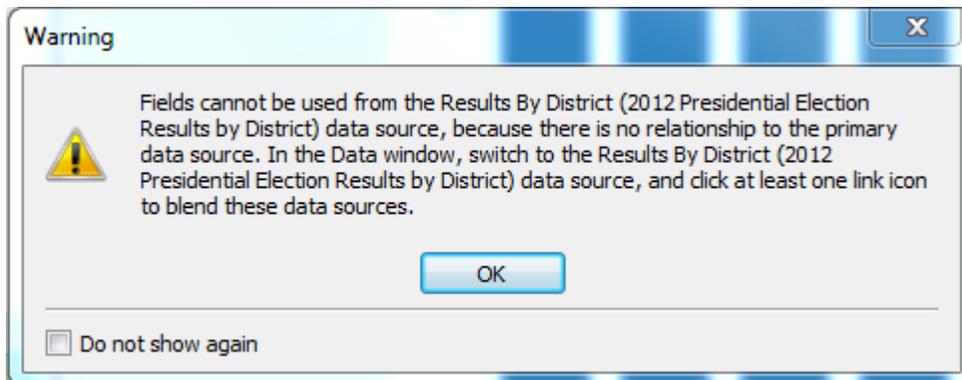
Click OK.

6) You'll now see the following at the top left of your Tableau window:
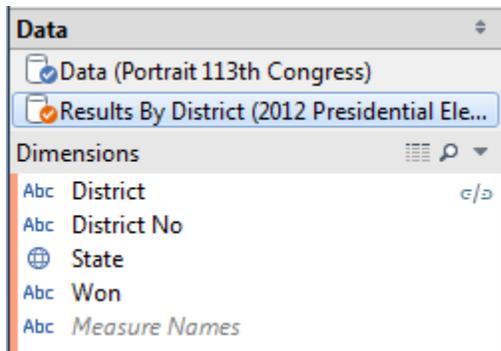
7) Click on the "broken link" (⊏/⊐) next to District. The link will change to a connected orange link and the chart will look like this:
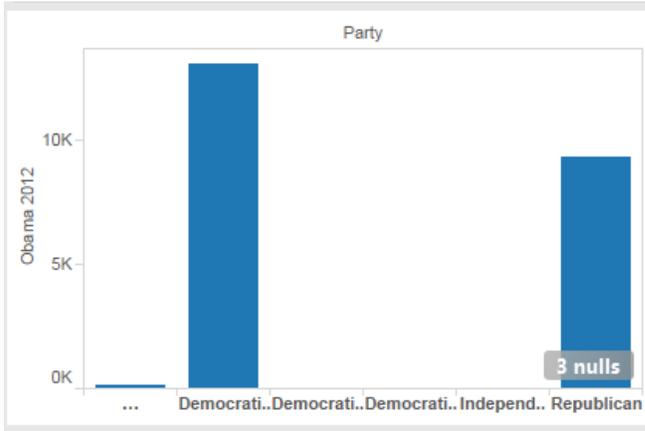
8) Now right-click on SUM(Obama 2012) in the Rows shelf and select Measure/Average.

9) Hold down the control key (CTRL) and click on the Democratic and Republican bars:
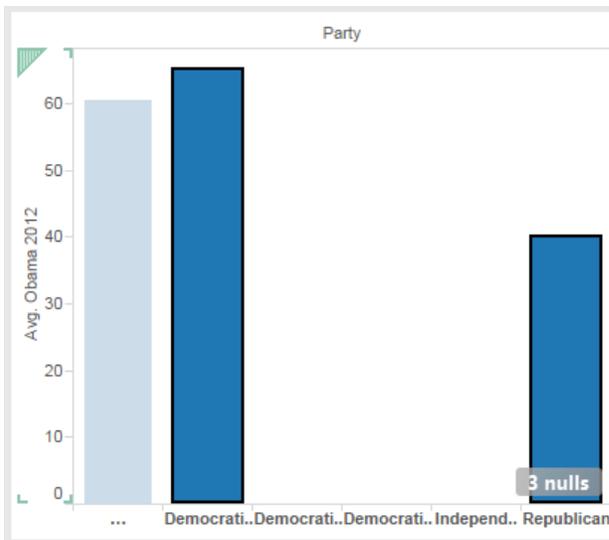
10) Hover your mouse over either of the highlighted bars and select "Keep Only."

11) Drag "Romney 2012" from Measures and place it next to AVG(Obama 2012) on the Rows shelf.

12) Right-click on SUM(Romney 2012) and change it to Average.

	Columns	Party			
	Rows	AVG(Obama 2012)		AVG(Romney 2012)	
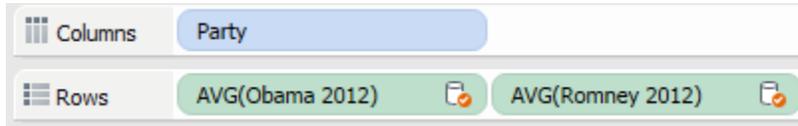
13) The result should look like this:
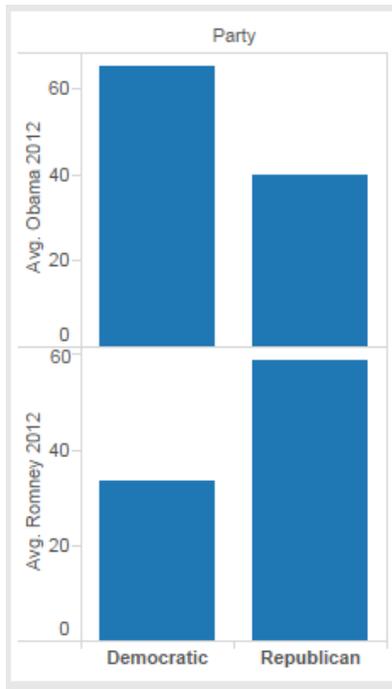
We learn that in congressional districts where the elected Representative is Democratic, Obama averaged 65% of the vote to Romney's 33%. In districts where the elected Representative is Republican, Romney averaged 59% of the vote to Obama's 40%.

We did it by combining election result data from the "2012 Presidential Election Results" worksheet with political party data from the "Portrait 113th Congress" worksheet.

14) Name the sheet "Rep Party and Election Results." Then save the workbook.

Part 5: Combine a calculated field in one data source with the original data from the other

1) Create a new worksheet. Name the worksheet "Rep Age and Election Results."

2) Click on the "Data (Portrait 113th Congress)" data source.

3) Create a calculated field by clicking on Analysis/Create Calculated Field.

4) Call the field "RepAge" and use the formula:

YEAR(TODAY())-[YEAR OF BIRTH]

RepAge	Data (Portrait 113th Congress)	
YEAR(TODAY())-[YEAR OF BIRTH]		
The calculation is valid.	Apply	OK

This calculates the age of the Representative by subtracting the year of their birth from the current year.

5) Drag RepAge (under Measures) to the Columns shelf.

6) Click on the "Results by District (2012 Presidential Election Results…)" data source.

7) Drag "Romney 2012" (under Measures) to the Rows shelf. When you see the warning dialog, just like before, click OK.

8) Then click the broken link next to District. It will again turn orange.

Dimensions	
Abc District	ᴄ/ᴐ
Abc District No	
State	
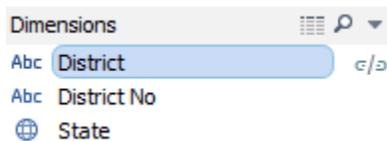
9) Right click on SUM(RepAge) and select Dimension. Do the same for SUM(Romney 2012).
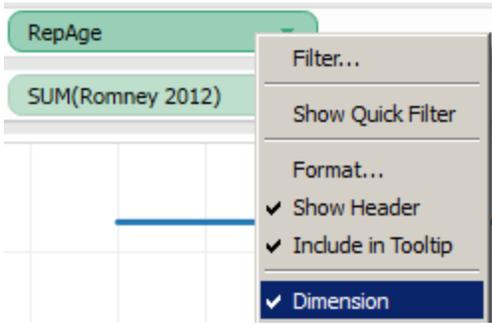
10) Right-click inside the scatterplot and click "Trend Lines/Show Trend Lines."
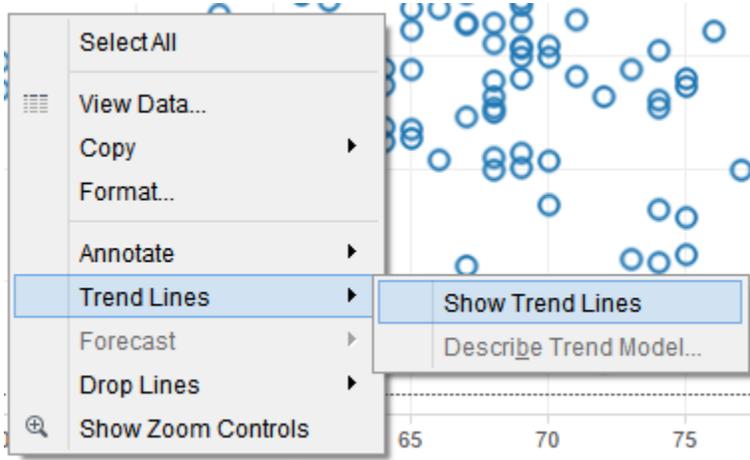
You'll see this
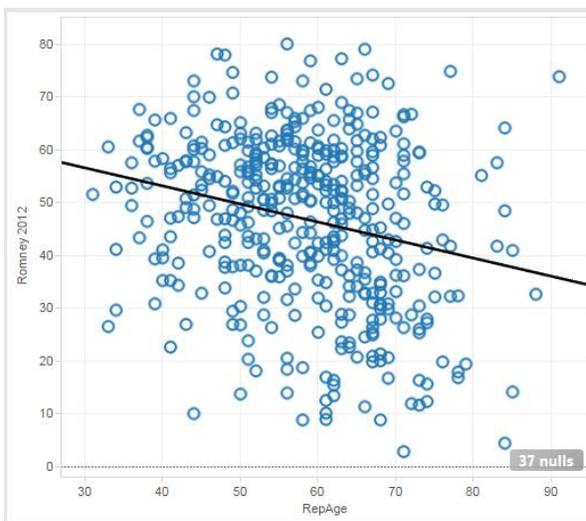
This implies a negative relationship between the age of the elected Representative and whether that district voted for Romney.