Introduction to data visualization in Power BI

Written by Carolina M. Siniscalchi (Data Sciences Coordinator, Libraries)

Last updated on: 04/03/2024

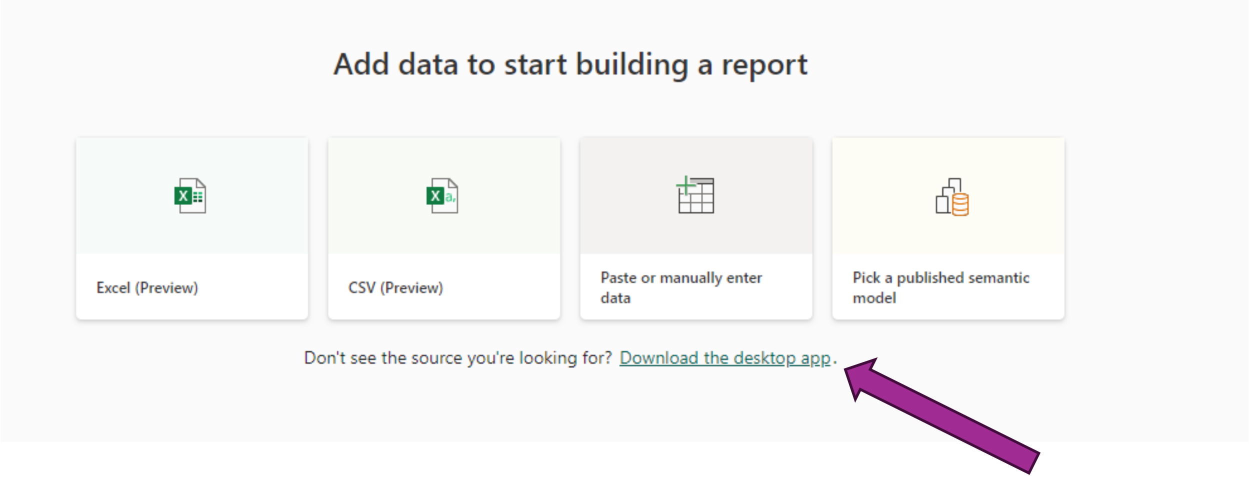
How to install it:

Go to https://www.microsoft365.com/. Sign in using your MSU netID and password. On the left side menu, click on "Apps".

On the menu cards, click on "Power BI". It will load an online version of the program.

Click on the top left green button "New report".

Under the central text "Add data to start building a report", there's a link to download and install the program on your computer.



Add data to start building a report
Excel (Preview)
CSV (Preview)
Paste or manually enter data
Pick a published semantic model
Don't see the source you're looking for? Download the desktop app.

A pop-up window will appear, asking your permission to open the Microsoft App Store. Click "Open store". You might need to do the dual authentication again. Install the program following the prompts.

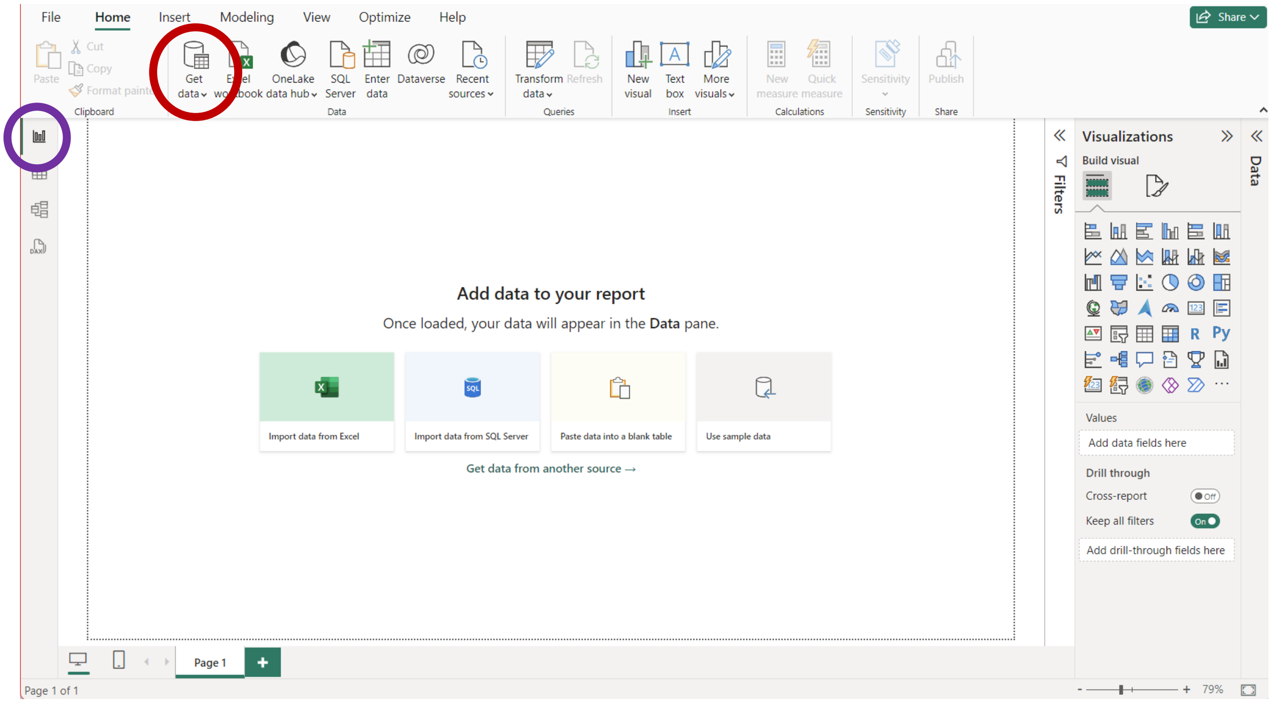
Open the program.

Starting the tutorial:

The essential feature of Power BI is the **report**. Each report can have multiple **pages**. Think of the report page as the canvas where you will show your data. To get started, click on "Report" on the "New" tab.

The new file will open in "Report view", identified by the small graph symbol on the left side menu (circled in purple).

Report view is where we will place all the plots and visualizations that we will build. Before starting plotting data, we need to load some data into the project. To do that, we will click on "Get data" (circled in red).



Add data to your report

Once loaded, your data will appear in the **Data** pane.

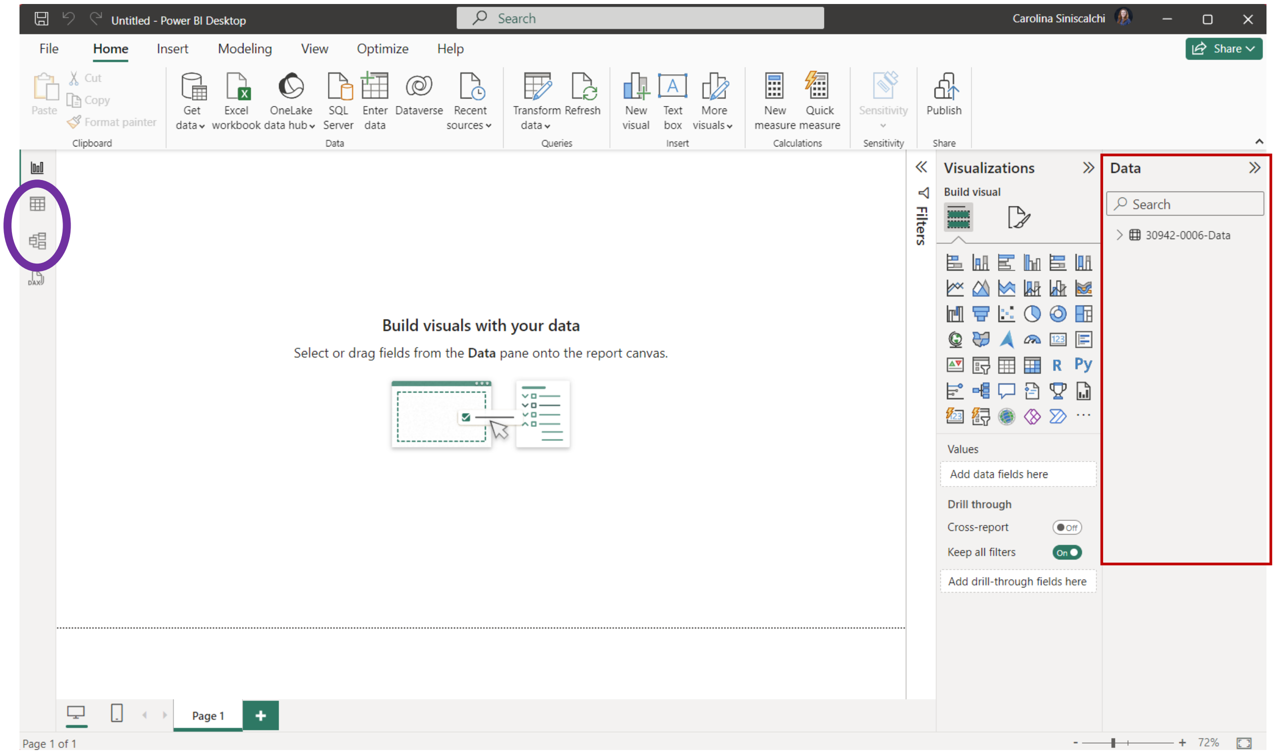
Get data from another source →

This button will show multiple options. For this tutorial, we will choose "Text/CSV". CSV (comma separated value) is a common and flexible file format, used to store data in tabular form. It's widely used across different statistical analysis and data visualization software.

For this tutorial, we will use a dataset that is publicly available on the ICPSR website (Inter-university Consortium for Political and Social Research). This dataset is part of the American House Survey, which is conducted every two years to generate information about housing situation across the country (see full citations and references at the end of this document). MSU Libraries is a member of ICPSR, so anyone from the university can create an account to use the website using their MSU NetId email.

In the tutorial, we will use the files referring to the 2009 survey carried out in New Orleans. The full data package downloaded from ICPSR has 8 subfolders, each referring to a part of the survey. Inside each folder, we find a .csv datafile and a pdf containing the codebook for the data file. The codebook contains important information about the variables in the datafile, like how they are named, if they are numerical, how they are coded, etc. Depending on how the data files are formatted, it is near impossible to understand them without the codebook, so make sure to keep it. We will work with 2 of the 8 files: 0002 (Journey to Work) and 0006 (Demographics).

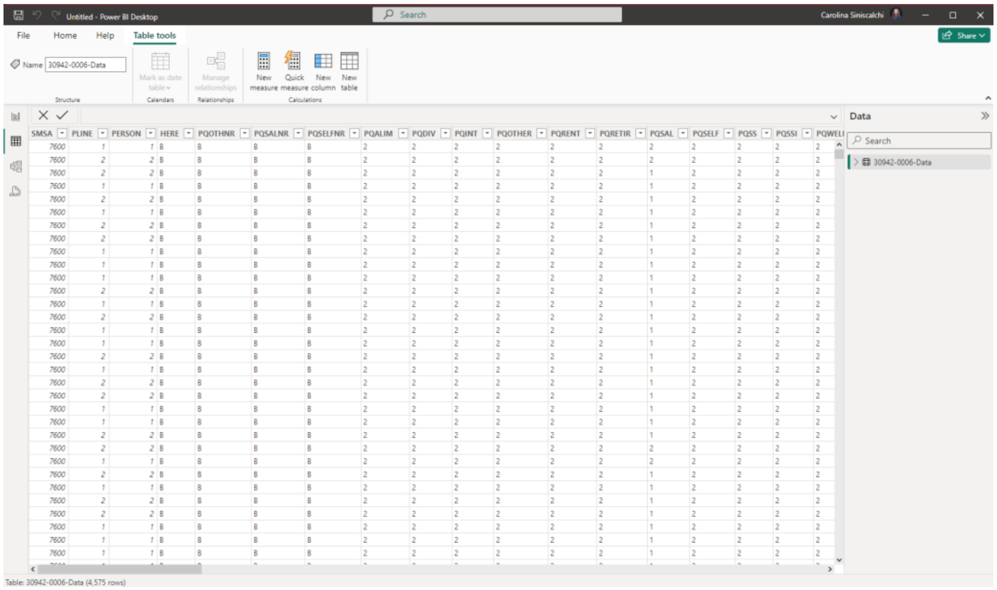
When clicking on "Get data > Text/CSV", a navigation window will pop up. Just navigate to the folder where you placed the downloaded files and select the .csv for folder 0006. A window showing a preview of the data will appear, and you will have the option to either load or transform the data. For now, let's just load it. Ignore any errors that might show for now. From here, we can see a couple of changes in our report. A tab called "Data" appeared on the right side, with our file (highlighted with a red square). If we click on the symbols highlighted by the purple circle, we can switch between Report, Table and Model view. Table shows the data that is loaded, Model shows the relationships build between different data files. This is very handy when aggregating across several files, and we will come back to it soon.



Build visuals with your data

Select or drag fields from the **Data** pane onto the report canvas.

Let's take a closer look at the data on Table view.



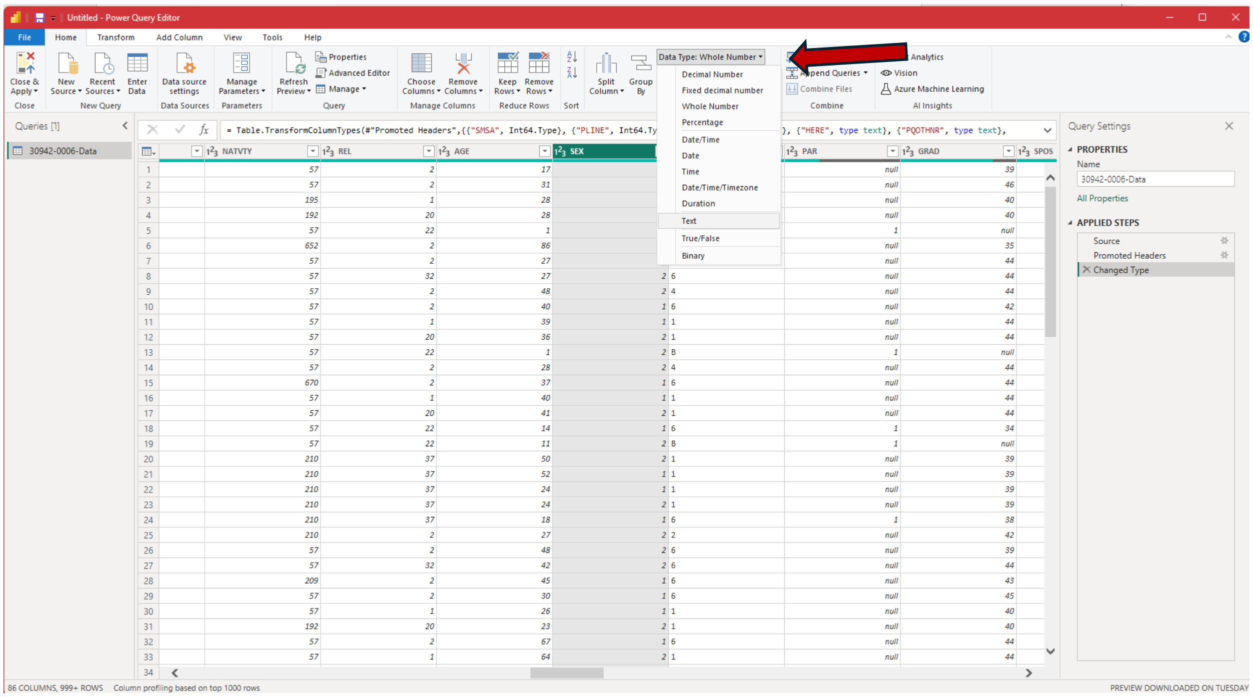
Name 30942-0006-Data

SMSA	PLINE	PERSON	HERE	PQOTHNR	PQSALNR	PQSELFNR	PQALIM	PQDIV	PQINT	PQOTHER	PQRENT	PQRETIR	PQSAL	PQSELF	PQSS	PQSSI
7600	1	1	B	B	B	B	2	2	2	2	2	2	2	2	2	2
7600	2	2	B	B	B	B	2	2	2	2	2	2	2	2	2	2
7600	2	2	B	B	B	B	2	2	2	2	2	2	1	2	2	2
7600	1	1	B	B	B	B	2	2	2	2	2	2	1	2	2	2
7600	2	2	B	B	B	B	2	2	2	2	2	2	1	2	2	2
7600	1	1	B	B	B	B	2	2	2	2	2	2	1	2	2	2
7600	2	2	B	B	B	B	2	2	2	2	2	2	1	2	2	2
7600	2	2	B	B	B	B	2	2	2	2	2	2	1	2	2	2
7600	1	1	B	B	B	B	2	2	2	2	2	2	1	2	2	2
7600	1	1	B	B	B	B	2	2	2	2	2	2	1	2	2	2
7600	1	1	B	B	B	B	2	2	2	2	2	2	1	2	2	2
7600	2	2	B	B	B	B	2	2	2	2	2	2	1	2	2	2
7600	1	1	B	B	B	B	2	2	2	2	2	2	1	2	2	2
7600	2	2	B	B	B	B	2	2	2	2	2	2	1	2	2	2
7600	1	1	B	B	B	B	2	2	2	2	2	2	1	2	2	2
7600	1	1	B	B	B	B	2	2	2	2	2	2	1	2	2	2
7600	2	2	B	B	B	B	2	2	2	2	2	2	1	2	2	2
7600	1	1	B	B	B	B	2	2	2	2	2	2	1	2	2	2
7600	2	2	B	B	B	B	2	2	2	2	2	2	1	2	2	2
7600	2	2	B	B	B	B	2	2	2	2	2	2	1	2	2	2
7600	1	1	B	B	B	B	2	2	2	2	2	2	1	2	2	2
7600	1	1	B	B	B	B	2	2	2	2	2	2	1	2	2	2
7600	2	2	B	B	B	B	2	2	2	2	2	2	1	2	2	2
7600	2	2	B	B	B	B	2	2	2	2	2	2	2	2	2	2
7600	1	1	B	B	B	B	2	2	2	2	2	2	2	2	2	2
7600	2	2	B	B	B	B	2	2	2	2	2	2	1	2	2	2
7600	1	1	B	B	B	B	2	2	2	2	2	2	1	2	2	2
7600	2	2	B	B	B	B	2	2	2	2	2	2	1	2	2	2
7600	2	2	B	B	B	B	2	2	2	2	2	2	1	2	2	2
7600	1	1	B	B	B	B	2	2	2	2	2	2	1	2	2	2
7600	1	1	B	B	B	B	2	2	2	2	2	2	1	2	2	2
7600	1	1	B	B	B	B	2	2	2	2	2	2	1	2	2	2

Table: 30942-0006-Data (4,575 rows)

As you can see, we need the codebook to understand the variable names (top row), and how they are coded. To start playing with this data, we will first look at the distribution of male and female individuals of different ages in this population. We will use 3 variables: CONTROL (the unique ID given to each respondent in this survey), SEX (coded as male [1] and female [2]) and AGE. Because gender is coded with numbers, we need to check and edit this variable on the table so it can be more easily understood. We will do this using the Power Query Editor. We will click on Home > Transform Data > Transform Data. A new Power Query window will appear with our data.

We will navigate to the column labeled "SEX" and click on it. The first step is to change the data type on this column from "Whole number" to "Text".

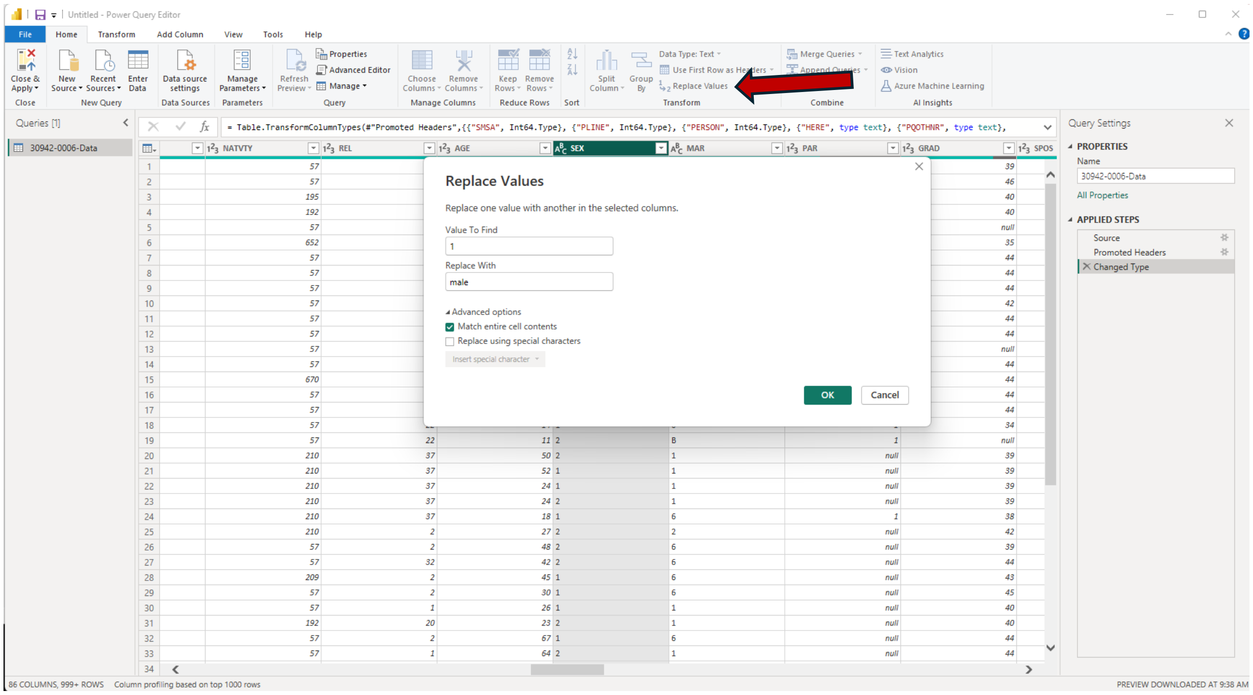


```
= Table.TransformColumnTypes(#"Promoted Headers",{{"SMSA", Int64.Type}, {"PLINE", Int64.Ty }, {"HERE", type text}, {"PQOTHNR", type text},
```

	NATVTY	REL	AGE	SEX		PAR	GRAD	SPOS
1	57	2	17			null	39	
2	57	2	31			null	46	
3	195	1	28			null	40	
4	192	20	28			null	40	
5	57	22	1			1	null	
6	652	2	86			null	35	
7	57	2	27			null	44	
8	57	32	27	2	6	null	44	
9	57	2	48	2	4	null	44	
10	57	2	40	1	6	null	42	
11	57	1	39	1	1	null	44	
12	57	20	36	2	1	null	44	
13	57	22	1	2	B	1	null	
14	57	2	28	2	4	null	44	
15	670	2	37	1	6	null	44	
16	57	1	40	1	1	null	44	
17	57	20	41	2	1	null	44	
18	57	22	14	1	6	1	34	
19	57	22	11	2	B	1	null	
20	210	37	50	2	1	null	39	
21	210	37	52	1	1	null	39	
22	210	37	24	1	1	null	39	
23	210	37	24	2	1	null	39	
24	210	37	18	1	6	1	38	
25	210	2	27	2	2	null	42	
26	57	2	48	2	6	null	39	
27	57	32	42	2	6	null	44	
28	209	2	45	1	6	null	43	
29	57	2	30	1	6	null	45	
30	57	1	26	1	1	null	40	
31	192	20	23	2	1	null	40	
32	57	2	67	1	6	null	44	
33	57	1	64	2	1	null	44	
34								

86 COLUMNS, 999+ ROWS Column profiling based on top 1000 rows PREVIEW DOWNLOADED ON TUESDAY

Now, we'll click on "Replace values (2 drop bars down from Data Type). A window will pop up where we will type the data that you want to replace. We'll replace 1 for "male" and 2 for "female" (we checked these values on the codebook). Make sure to click in Advanced options and mark "Match entire cell contents". Here this won't make a difference, but it is important when replacing numbers with more than one digit.



```
= Table.TransformColumnTypes(#"Promoted Headers",{{"SMSA", Int64.Type}, {"PLINE", Int64.Type}, {"PERSON", Int64.Type}, {"HERE", type text}, {"PQOTHNR", type text},
```

		NATVTY	REL	AGE	SEX	MAR	PAR	GRAD	SPOS
1		57						39	
2		57						46	
3		195						40	
4		192						40	
5		57						null	
6		652						35	
7		57						44	
8		57						44	
9		57						44	
10		57						42	
11		57						44	
12		57						44	
13		57						null	
14		57						44	
15		670						44	
16		57						44	
17		57						44	
18		57						34	
19		57	22	11	2	B	1	null	
20		210	37	50	2	1	null	39	
21		210	37	52	1	1	null	39	
22		210	37	24	1	1	null	39	
23		210	37	24	2	1	null	39	
24		210	37	18	1	6	1	38	
25		210	2	27	2	2	null	42	
26		57	2	48	2	6	null	39	
27		57	32	42	2	6	null	44	
28		209	2	45	1	6	null	43	
29		57	2	30	1	6	null	45	
30		57	1	26	1	1	null	40	
31		192	20	23	2	1	null	40	
32		57	2	67	1	6	null	44	
33		57	1	64	2	1	null	44	
34									

Replace Values

Replace one value with another in the selected columns.

Value To Find: 1

Replace With: male

Advanced options

☑ Match entire cell contents

□ Replace using special characters

Insert special character

OK Cancel

PROPERTIES

Name: 30942-0006-Data

All Properties

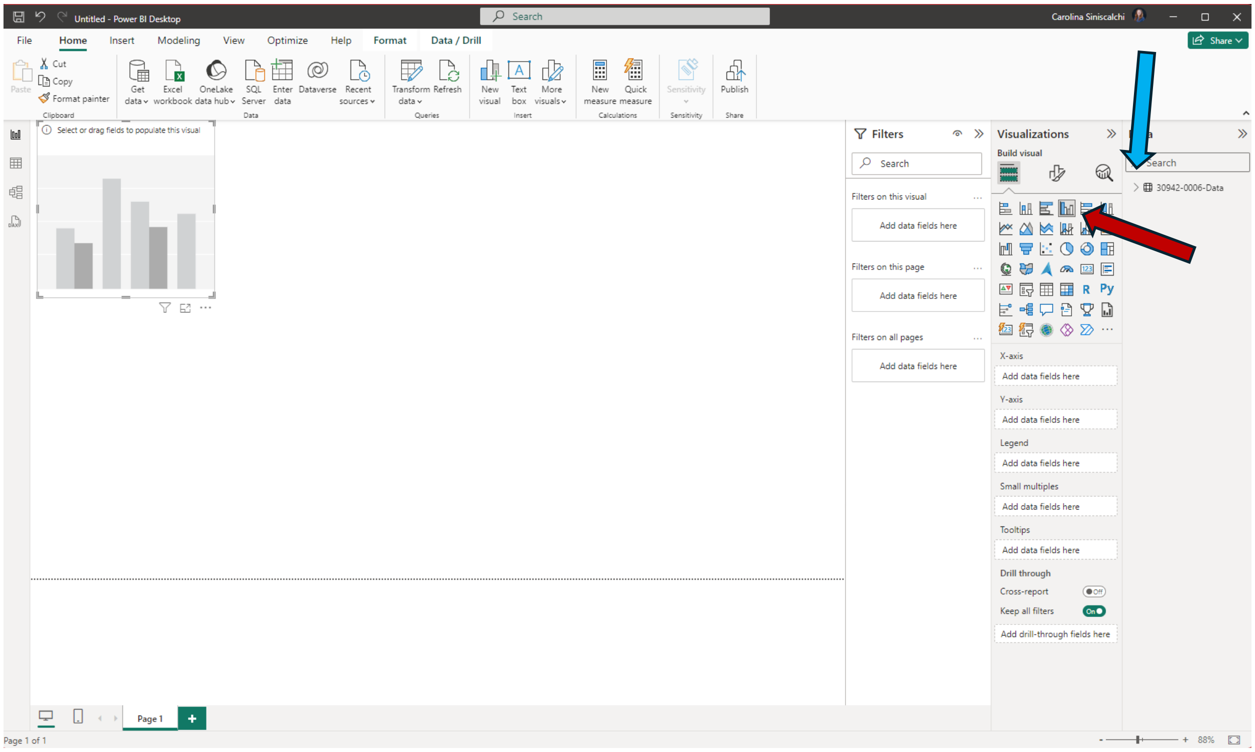
APPLIED STEPS

- Source
- Promoted Headers
- Changed Type

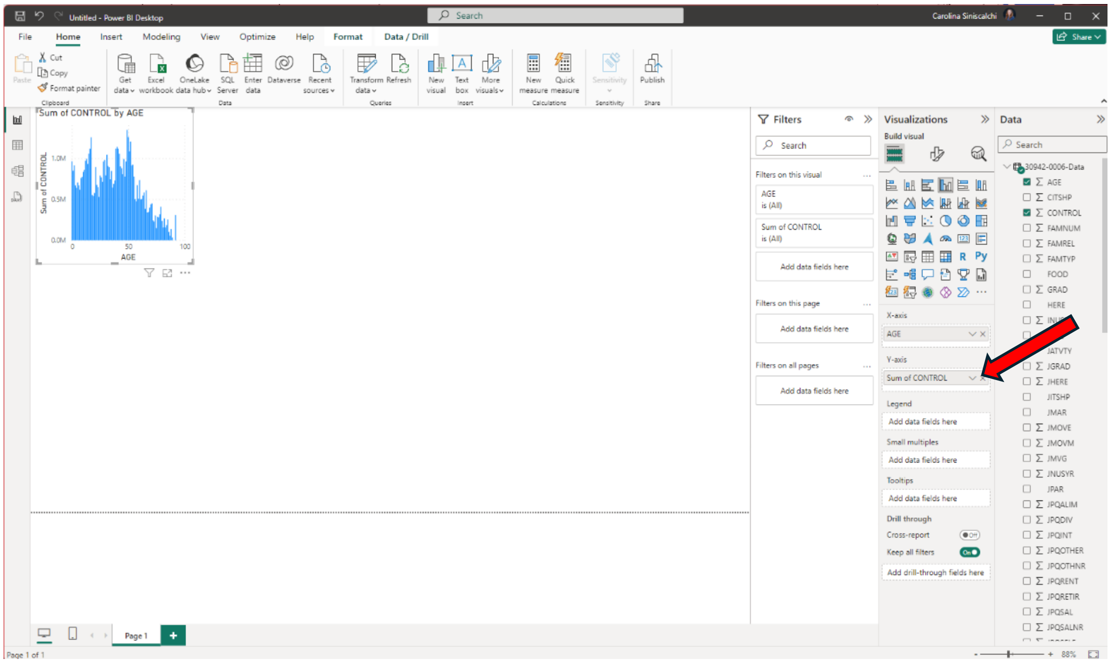
86 COLUMNS, 999+ ROWS Column profiling based on top 1000 rows

PREVIEW DOWNLOADED AT 9:38 AM

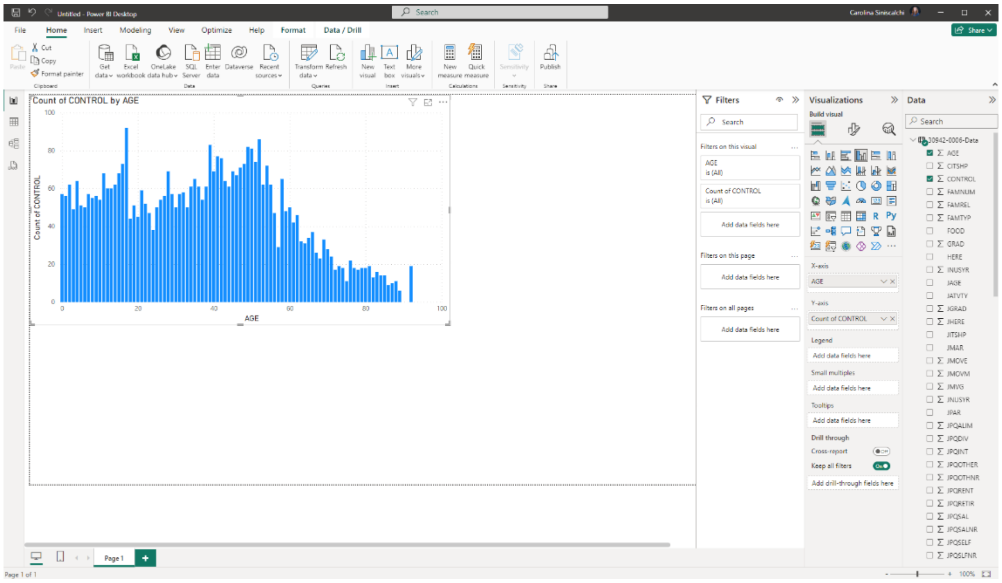
Now we click in the the Close & Apply button on the left corner, click in “Apply”, a few errors will show, but we will ignore them for now. Then click Close. We are back to our report, and navigate to Report view. We will start by building a visual. In the Visualizations tab on the right, we will click on Clustered column chart (red arrow). Now we need to add data to this visual. This is done on the data fields on the visualizations tab. On the Data tab, we can click on the arrow on the left of the data file name to expand it (blue arrow).



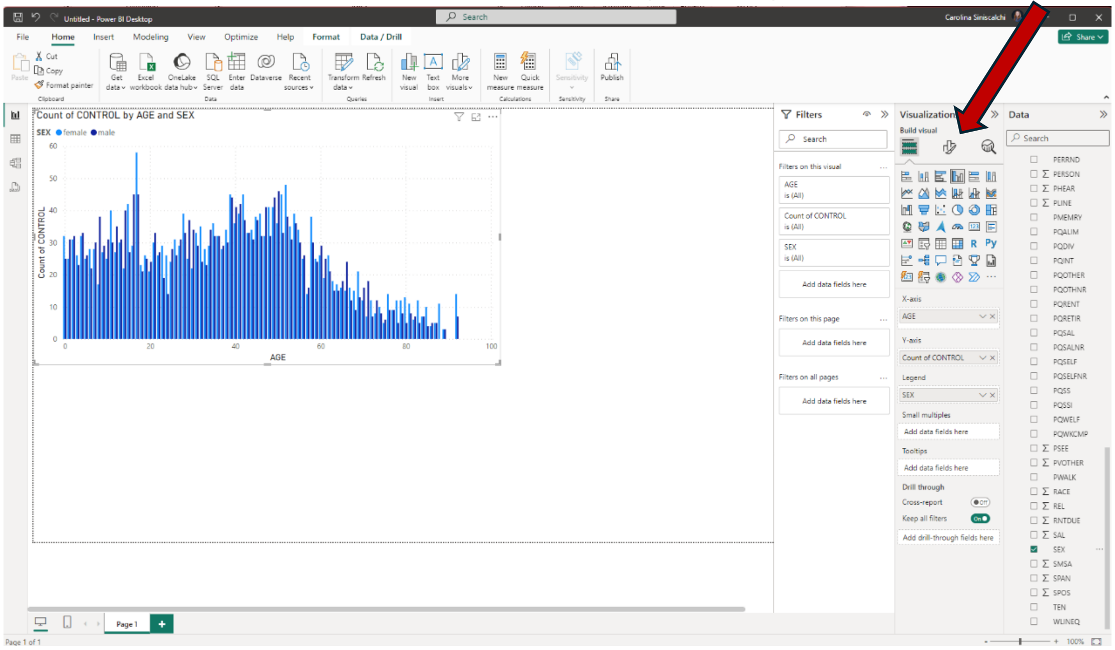
We want to plot a bar chart (here called column chart) of the number of people of each age. Age will be on the X axis, and the number of people, here represented by the variable CONTROL, on the Y axis. We will simply drag those variables from the Data tab to the Visualizations tab.



Our visual is now populated, but there's something weird on the Y-axis, the numbers are into the millions. If we look at the data fields, we'll see that "Sum of CONTROL" is showing. This is because the program is interpreting the IDs as whole numbers and using automatic aggregations to calculate the sums. To avoid that, we can either change the data type to text, or just change the aggregation. To do that, just click on the down arrow on the Y-axis box and choose "Count (Distinct)" (red arrow above). This will treat each distinct number in that column as a different individual. We can also increase the size of the visual by clicking and dragging on the corner of the visual. Much better now!

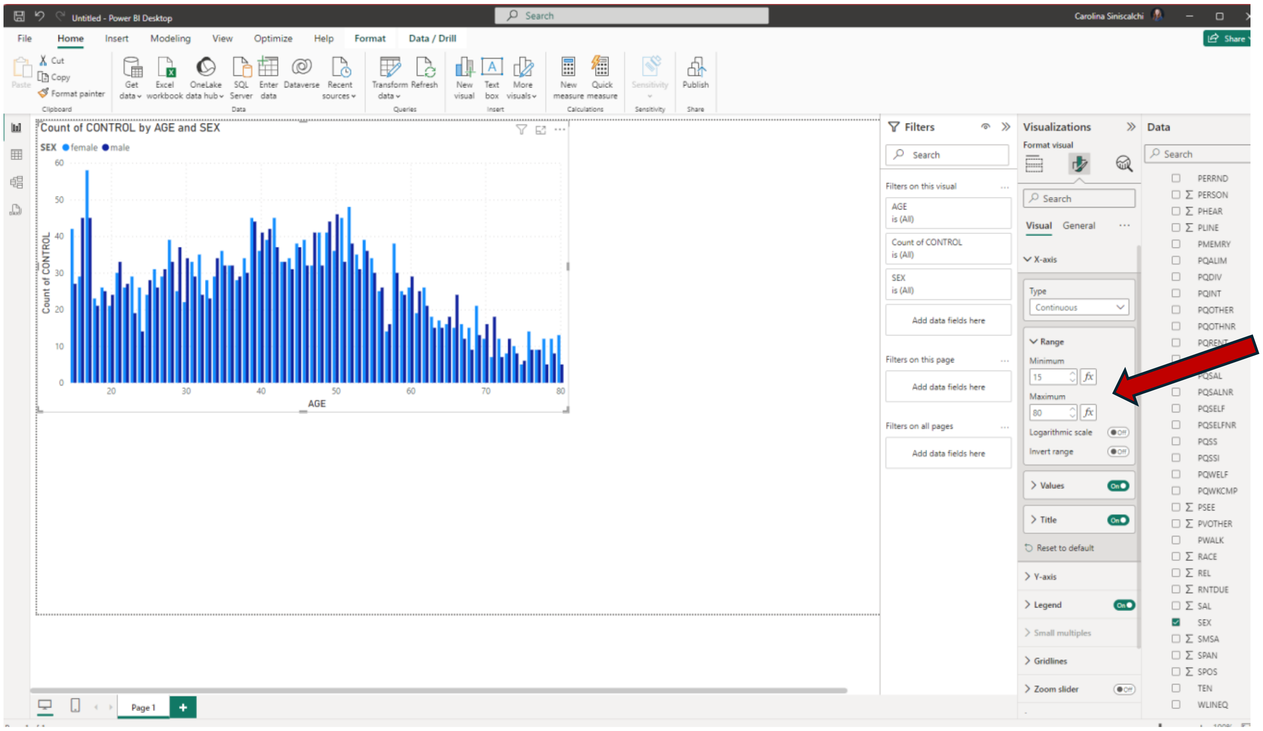


Now, we want to further break down our data by gender. To do that, we will drag the SEX variable and drop it in the Legend data field. We see now that we have two bars for each gender, and a legend appeared automatically. This plot is a bit crowded, so let's make some changes to make it better. To do that, we will use the Format visual tab (red arrow).



X-axis
AGE
Y-axis
Count of CONTROL
Legend
SEX
Filters on this visual
AGE is (All)
Count of CONTROL is (All)
SEX is (All)

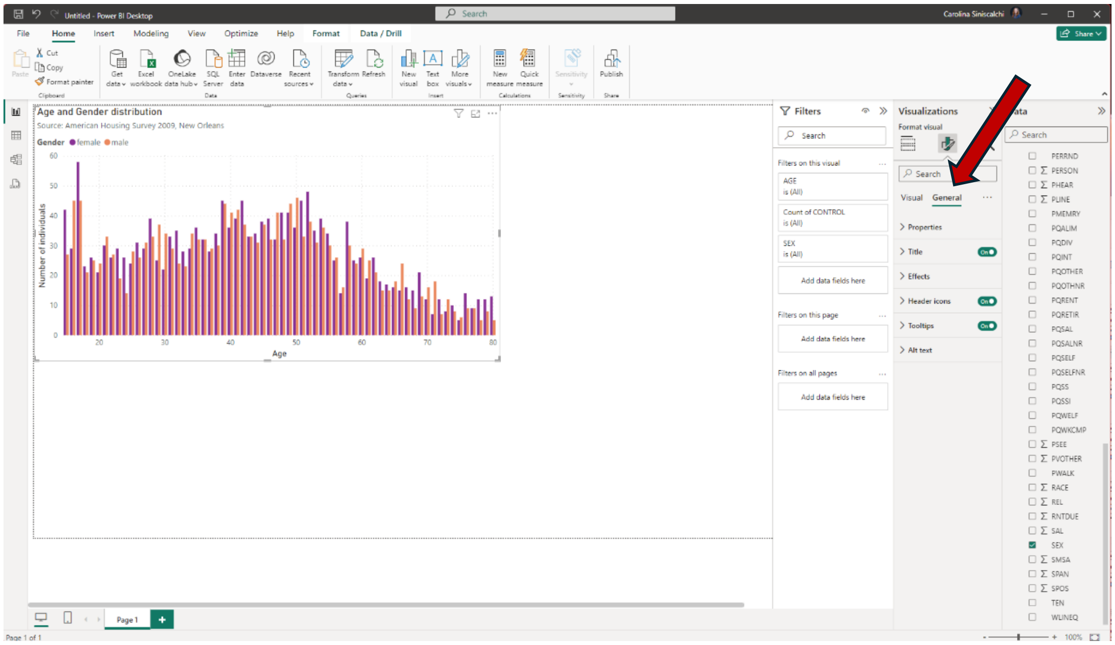
One thing we can do is restrict the range on the X axis. On the Format tab, we will click on the arrow besides X-axis to expand this tab. The Range is set up to Auto, so we will change it to a minimum of 15 and a maximum of 80. This makes it look slightly better.



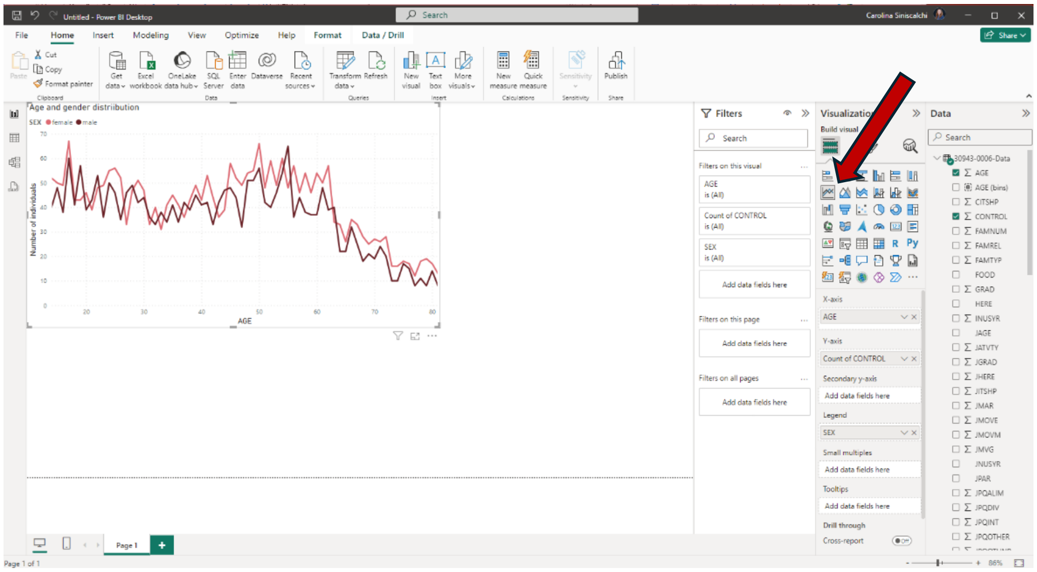
X-axis
Type: Continuous
Range
Minimum: 15
Maximum: 80
Logarithmic scale: Off
Invert range: Off
Values: On
Title: On
Legend: On
Zoom slider: Off

We can also change other features in the Format tab, such as the axis labels, bar colors, graph title, etc. These options are all available, but sometimes are hard to find. Under each axis drop menus, we can change the fonts and titles. On the Columns menu, we can choose different colors, by changing the series under "Apply settings to". On the Legend menu, we can change the position, title, etc.

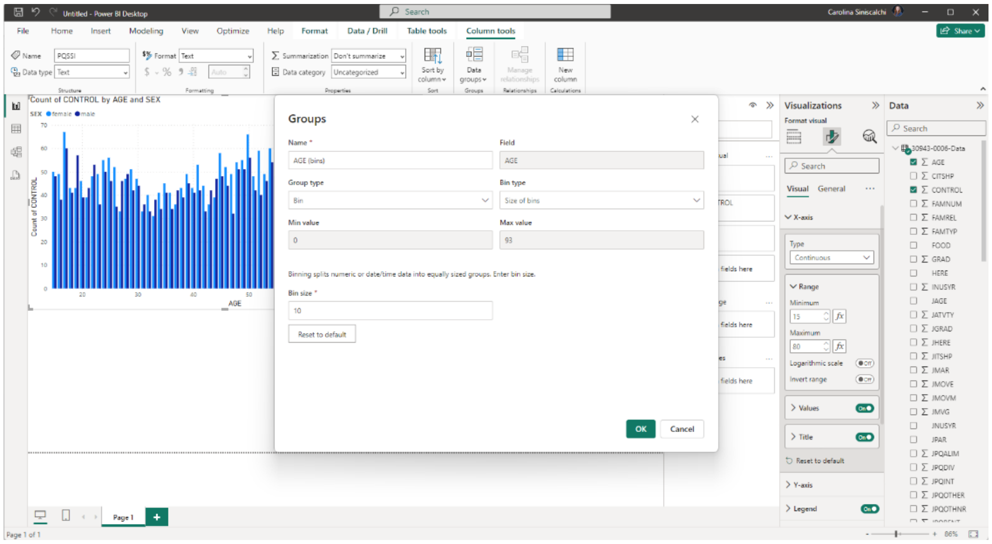
One tricky thing is that to change the label of the graph itself, we have to move to the General tab of the Format visual tab (red arrow). We can also add a Subtitle.



We can easily change the type of visual that we are showing too. Because age is a continuous variable, we could choose to represent this as a line graph, for example. To do that, we go back to the Build visual tab and click on the Line chart icon (red arrow).



Now let's add a second visual, showing the average salary by age. To make this visual clearer, we will create a derived variable from the AGE column, creating 10-year bins for Age. We can do that by right clicking the AGE field in the Data tab to the right, we will then click in "New group". A window will pop up, already with the bin calculation. We will choose a bin size of 10.



Groups

Name *: AGE (bins)

Field: AGE

Group type: Bin

Bin type: Size of bins

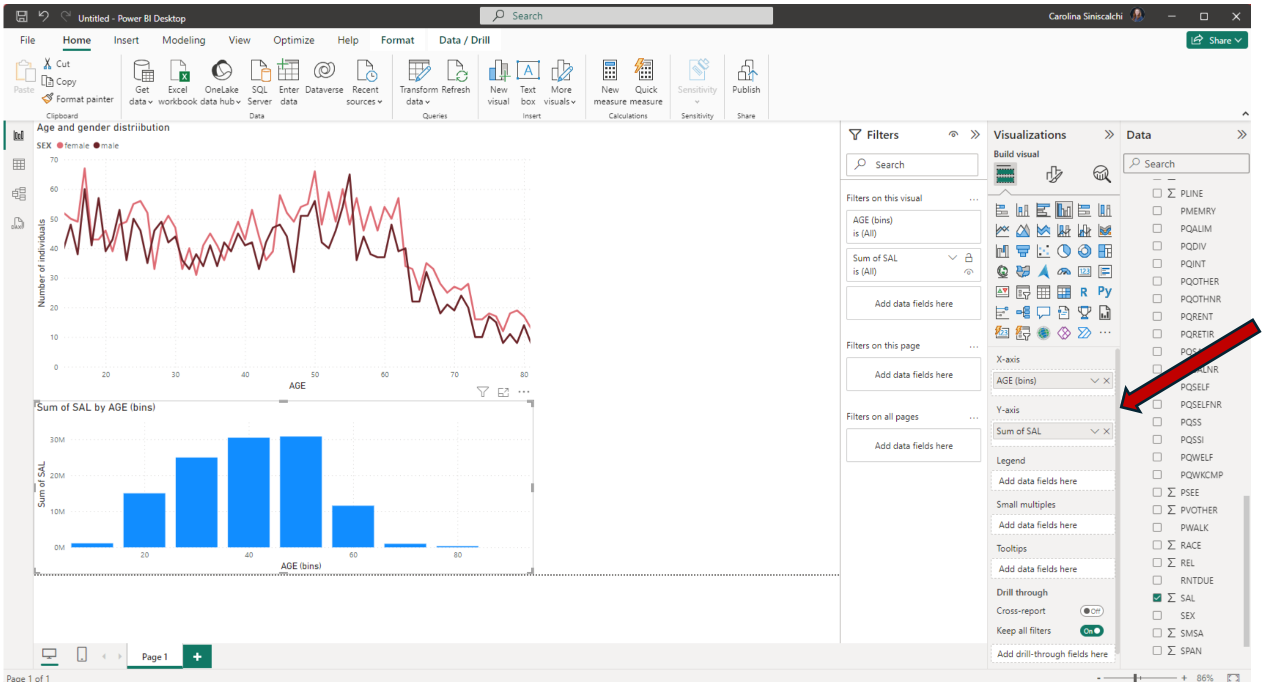
Min value: 0

Max value: 93

Binning splits numeric or date/time data into equally sized groups. Enter bin size.

Bin size *: 10

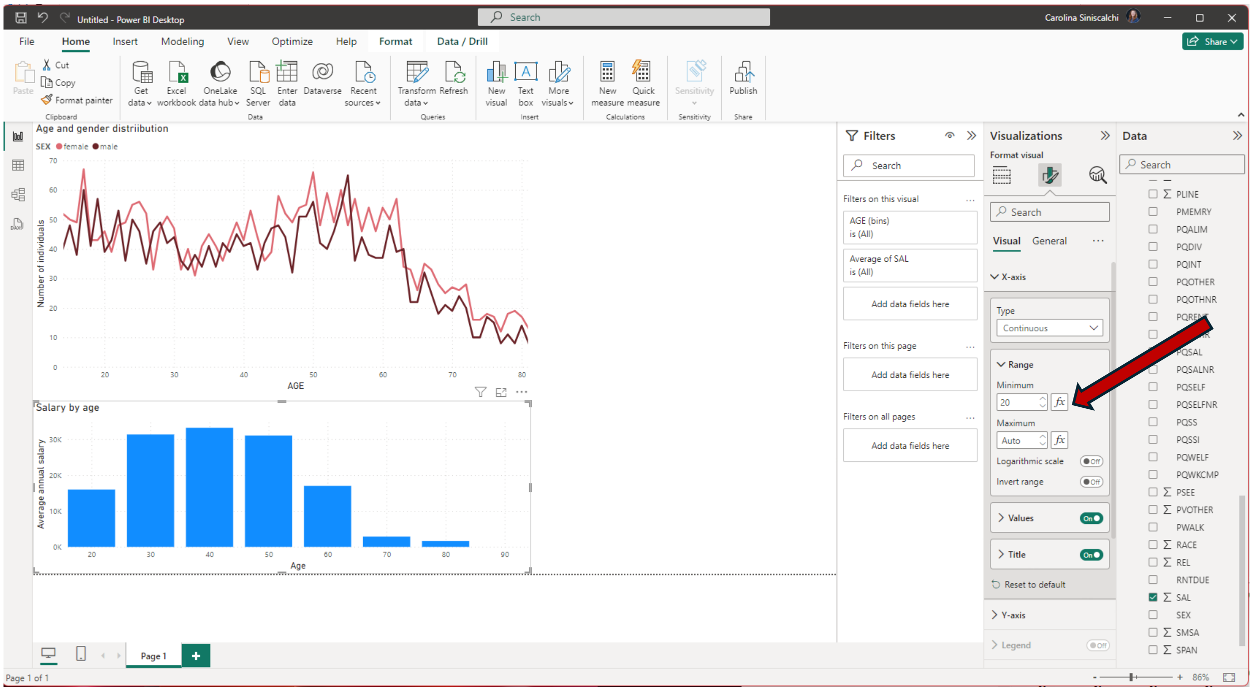
A new variable called AGE (bins) is now in our model. We can build our next visual. We go back to the Build visual tab, make sure to de-select the first visual, then click in Clustered column chart again. It will automatically select an area of the report, let's use it for now, but feel free to move it around if you prefer. We will then drag the AGE (bins) variable to the X-axis and the variable SAL (which represents annual salary) to the Y-axis.



X-axis: AGE (bins)

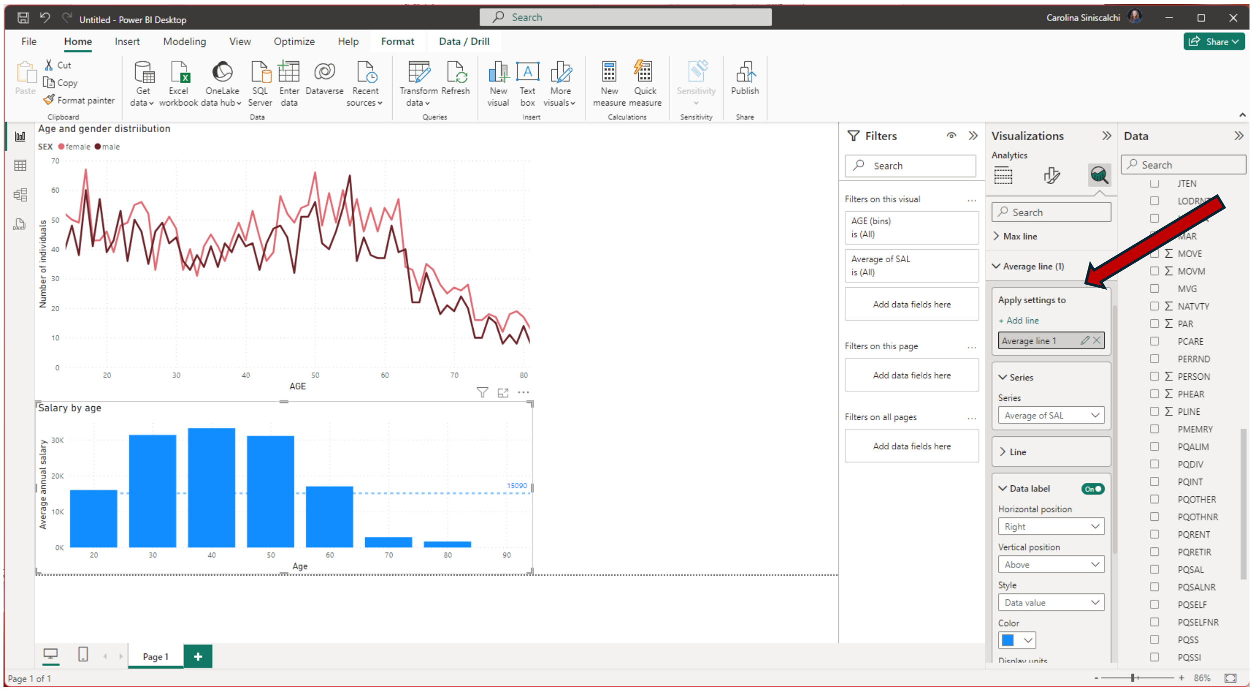
Y-axis: Sum of SAL

Again, we need to change the summarization of SAL. We click on the arrow and choose Average. Let's also remove the 0-10 bin, because it doesn't make much sense here. We go on the Format visual tab and change the minimum value of AGE to 20. Let's also change the data labels.



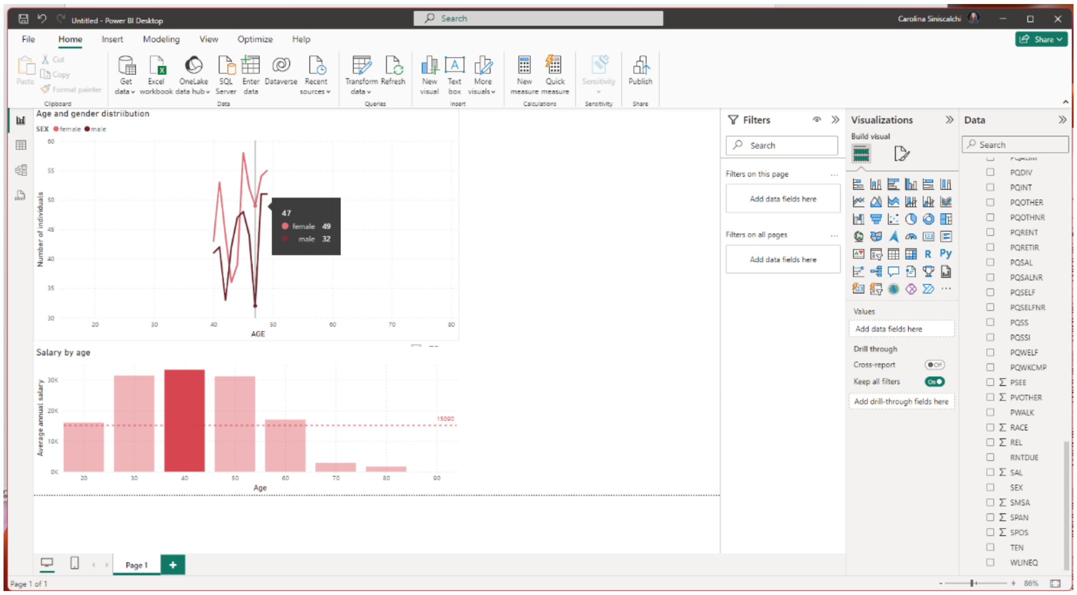
X-axis
Type: Continuous
Range
Minimum: 20
Maximum: Auto
Logarithmic scale: Off
Invert range: Off
Values: On
Title: On
Reset to default
Y-axis
Legend: Off

Clustered charts allow the inclusion of lines representing different measures. We can do that by moving to the Analytics tab in the Visualizations tab. Let's add an Average line, by clicking on the drop arrow. We only have one variable on the Y axis, so it automatically applies to Salary. We can also change color and style on this tab and add a label showing the average value.

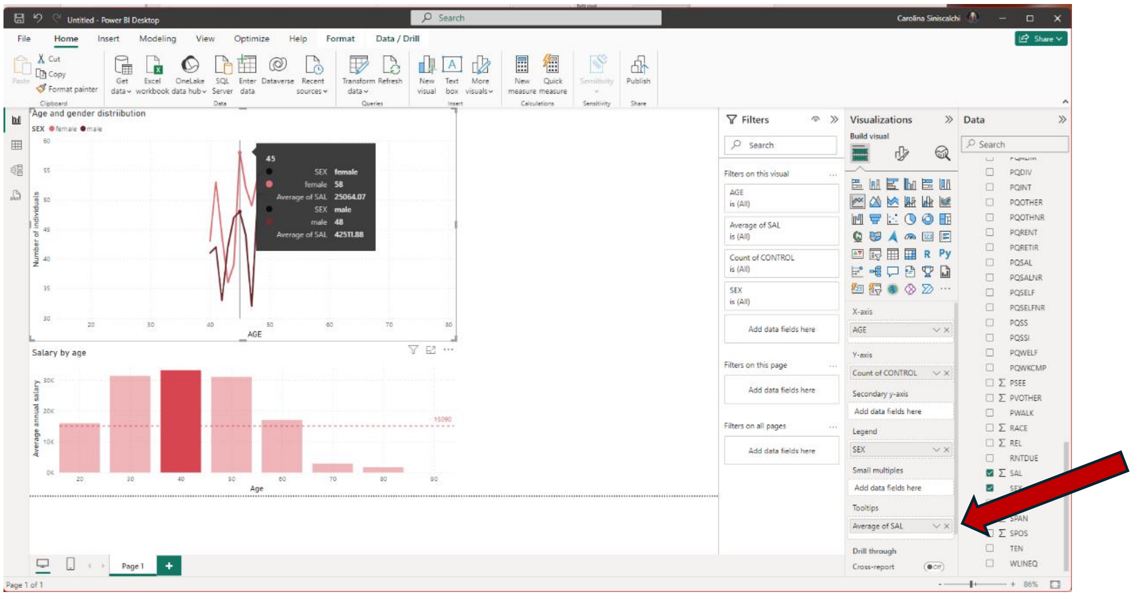


Max line
Average line (1)
Apply settings to
+ Add line
Average line 1
Series
Series: Average of SAL
Line
Data label: On
Horizontal position: Right
Vertical position: Above
Style: Data value
Color

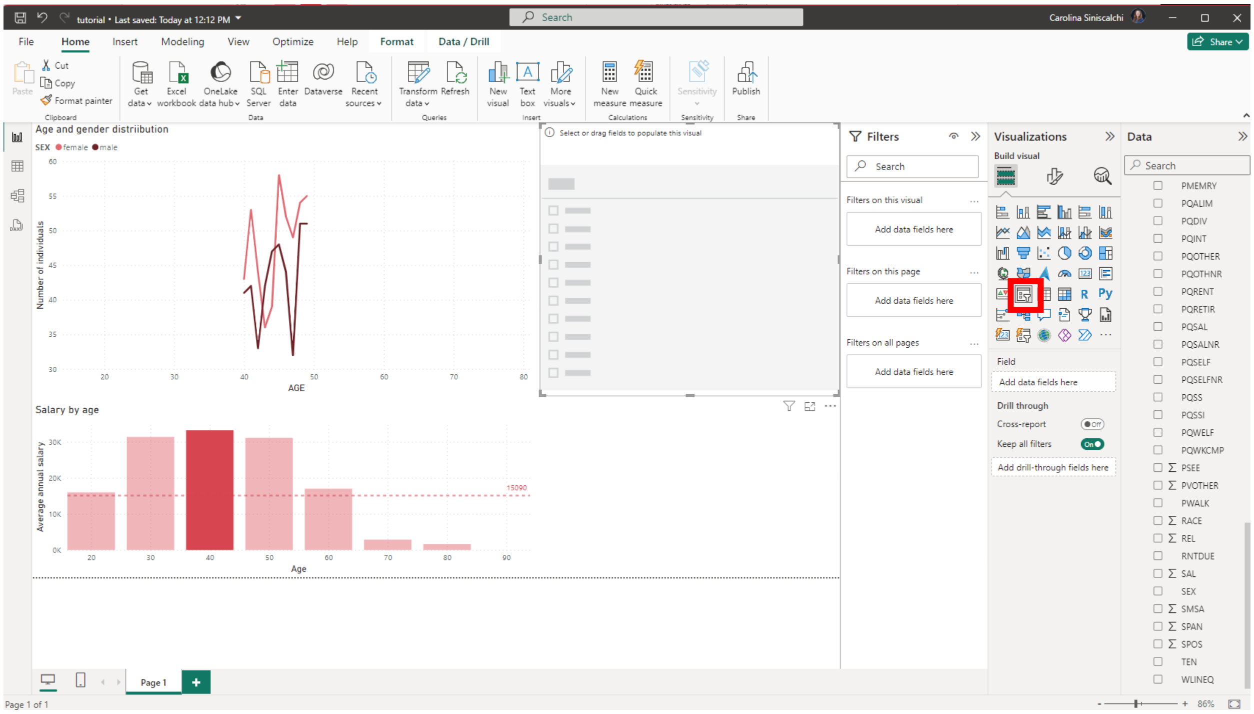
Now let's see some properties of this report. If we click in any of the bars on the second visual, the first visual will change. This is because visuals on a report that come from the same data are linked throughout.



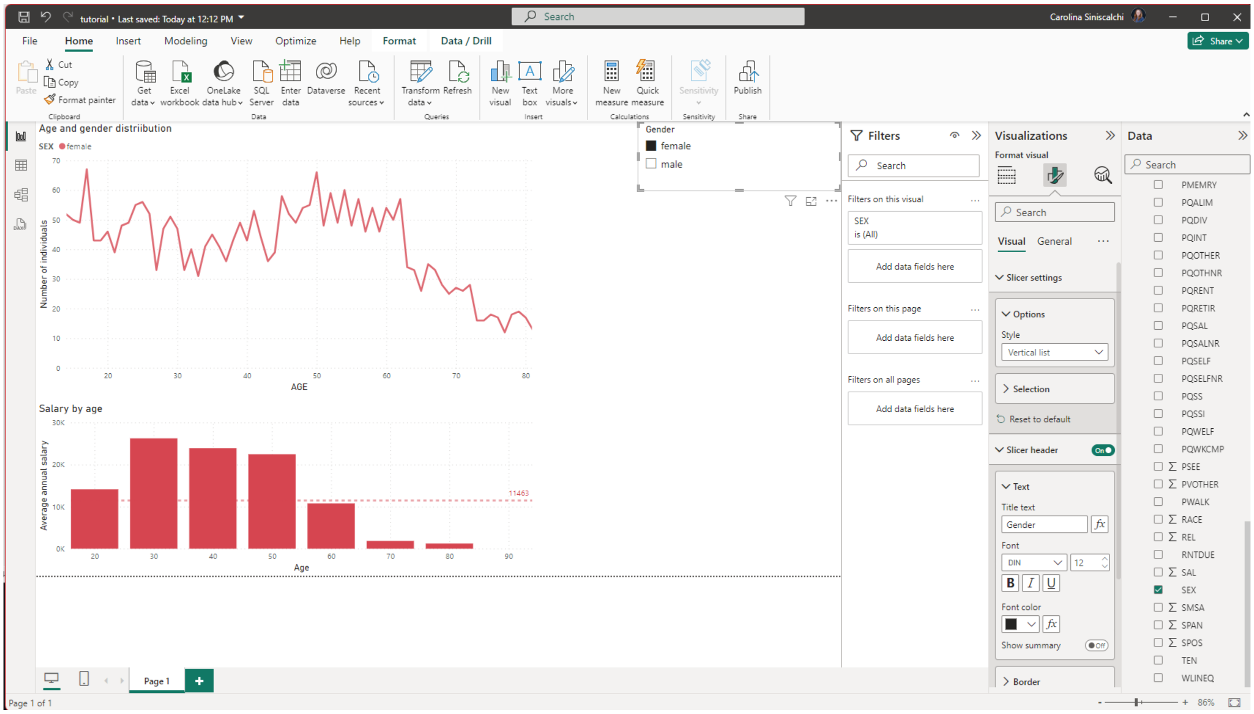
Another interesting feature is the tooltip. When we hover our arrow on a data point, it will show a black box containing the values for that point. It loads the variables that are in the visual automatically but can be enriched by dragging other variables and dropping them in the Tooltips field under Build visuals. Let's do that by dragging SAL into the Tooltips field of the first visual and changing the summarization to Average. The tooltip now also shows the average salary by gender for that specific age.



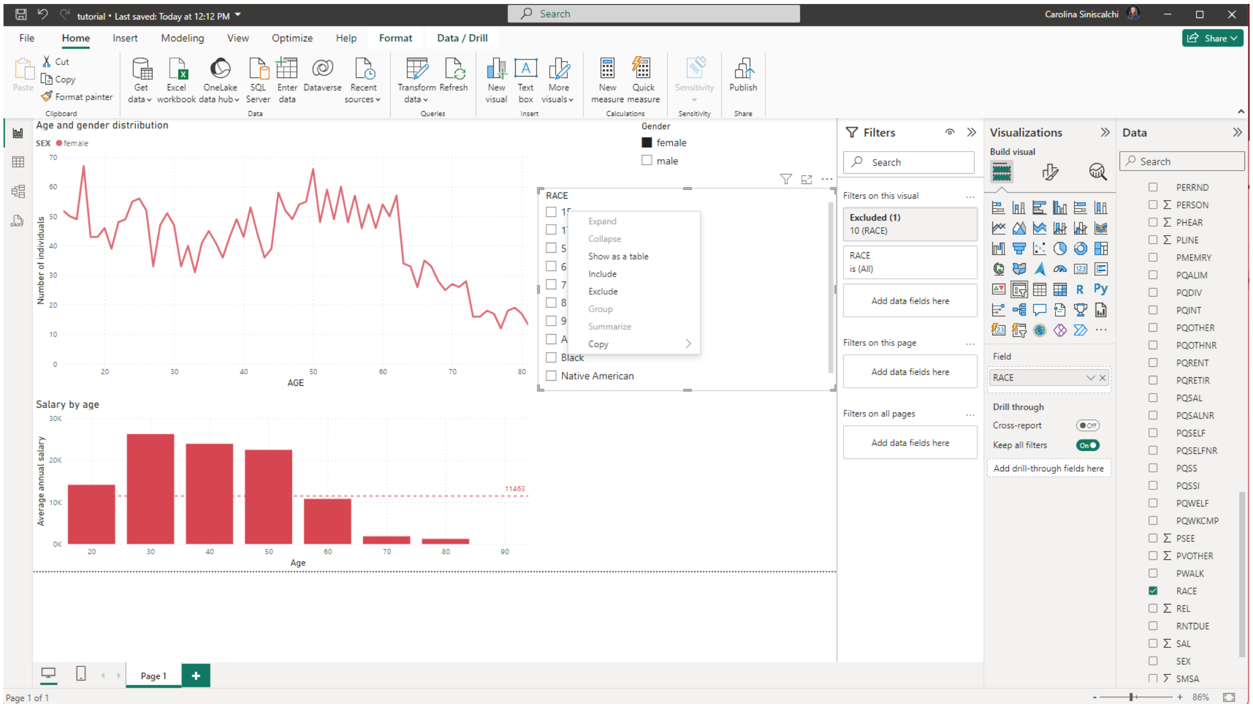
One powerful feature of Power BI is the ability to add interactive filters to reports. We will do this to our report, adding a global filter based on gender. To do that, we go back to the Build visual tab and click on Slicer (red square). Notice a new card appeared on the visual.



We'll drag the variable SEX to this card. When we do that, it immediately shows the two categories in this variable If we click in "female" all the visuals will show only the data pertaining to this category. In the Format tab, we can change the title and how the slicer card appears.

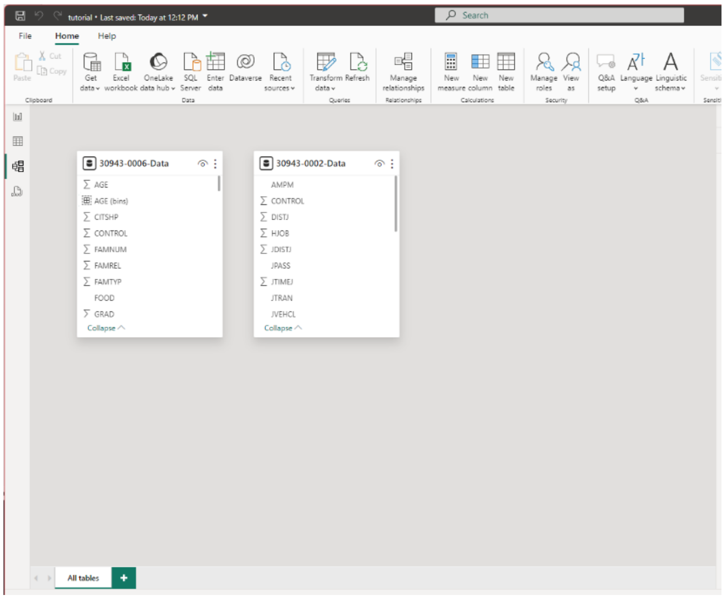


We can add a second slicer for Race, for example, but to do that, we first need to recode this variable. We will open Power Query again. According to the Code Book, there are 21 categories in this variable, but the first four account for more than 98% of the respondents. We will only recode those, which are 01=White, 02=Black, 03=Native American, 04=Asian. We will first change the data type to Text, then replace the values. Don't forget to mark "Match entire cell contents", as we have double figure numbers here. Apply and close power query. In the Report view, we create a new slicer, and drag RACE to this card. All the categories will show, even the ones we didn't re-code. We just need to right click each category that we want to exclude and click on Exclude.

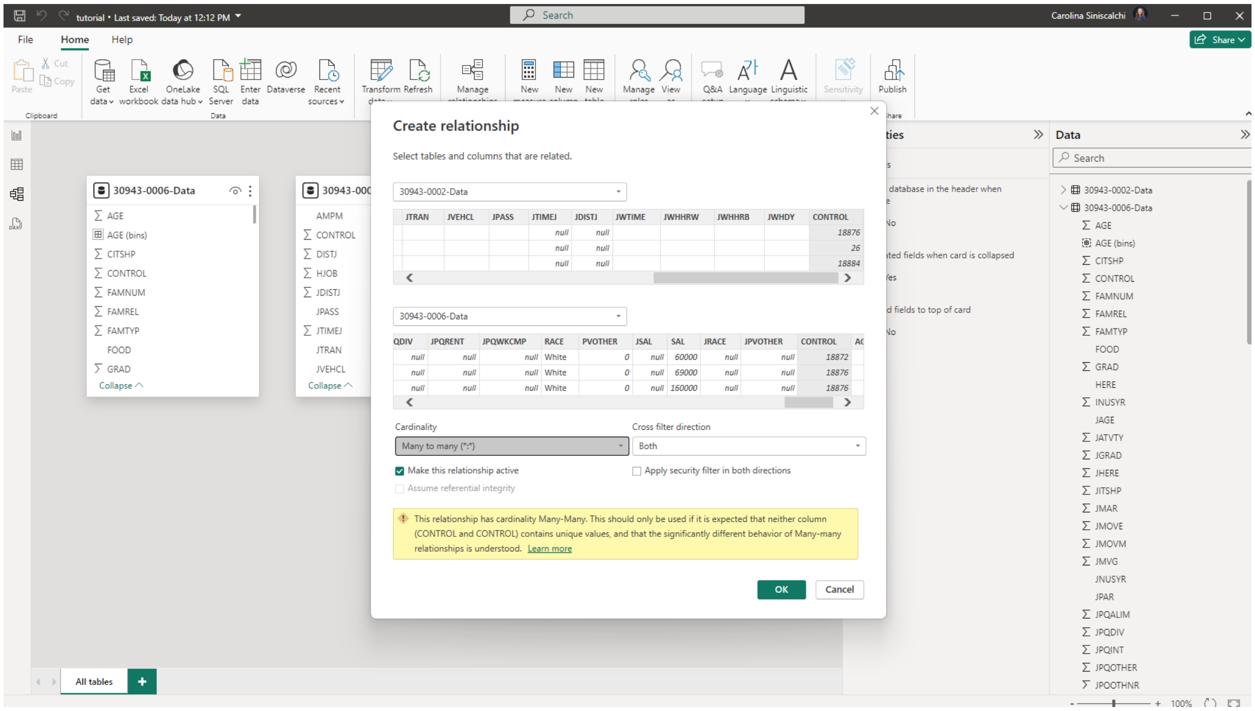


Now we can select categories in both cards at the same time and filter the data in two ways.

Now let’s add a second page to our report and bring another file from the American Housing Survey package. We will navigate to Model view (left side menu, third icon from the top). It should show one card, related to the data file that is already loaded. We go to Get data, select Text/CSV and navigate to our file 0002, then click load. A new card will appear for this new data file.



We will now use a variable shared by both files to link them, so we can grab variables from both files in our analyses. This variable in this case is CONTROL, which is the unique ID given to all participants. To link them, we click on CONTROL in the second card and drag it to the first. A window will pop up, where we can determine the directionality of the relationship, and how they relate. We will leave it as is and click Ok. A line will appear between both cards.



Create relationship

Select tables and columns that are related.

30943-0002-Data

JTRAN	JVEHCL	JPASS	JTIMEJ	JDISTJ	JWTIME	JWHHRW	JWHHRB	JWHDY	CONTROL
			null	null					18876
			null	null					26
			null	null					18884

30943-0006-Data

QDIV	JPQRENT	JPQWKCMP	RACE	PVOTHER	JSAL	SAL	JRACE	JPVOTHER	CONTROL	AG
null	null	null	White	0	null	60000	null	null	18872	
null	null	null	White	0	null	69000	null	null	18876	
null	null	null	White	0	null	160000	null	null	18876	

Cardinality: Many to many (*:*)

Cross filter direction: Both

☑ Make this relationship active

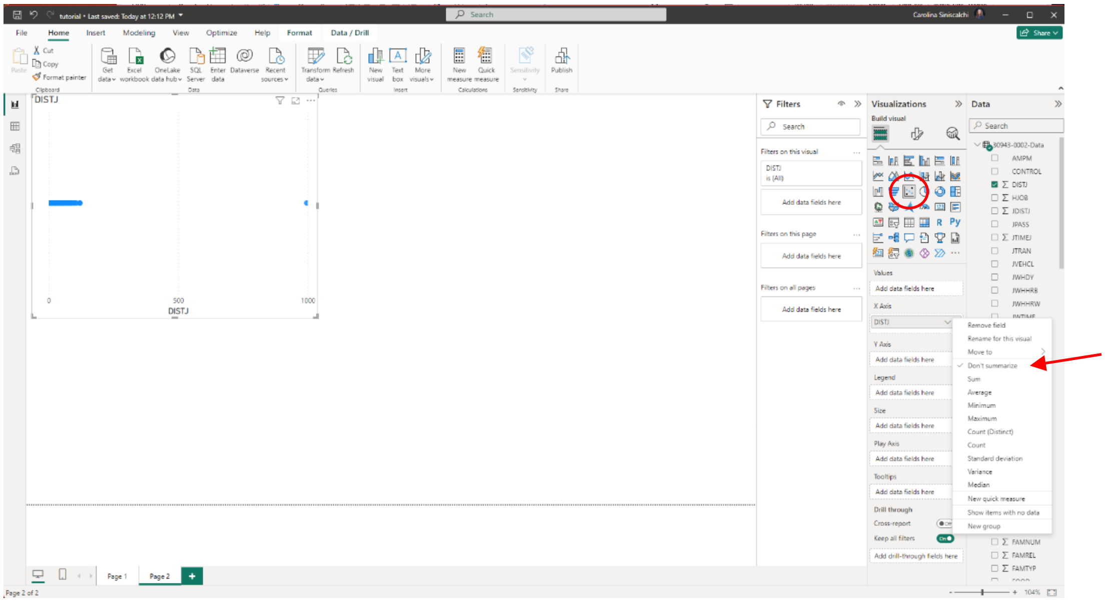
☐ Apply security filter in both directions

☐ Assume referential integrity

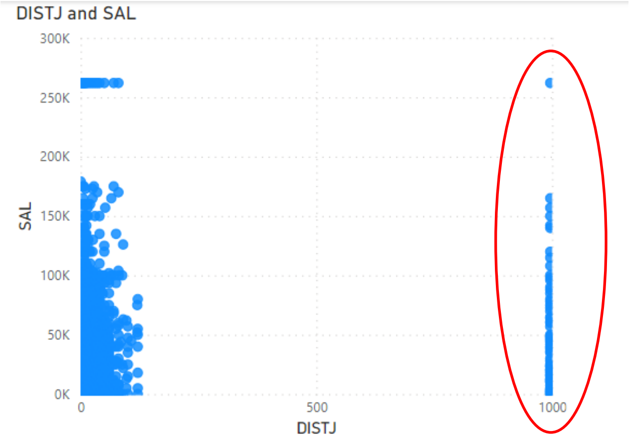
This relationship has cardinality Many-Many. This should only be used if it is expected that neither column (CONTROL and CONTROL) contains unique values, and that the significantly different behavior of Many-many relationships is understood. Learn more

OK Cancel

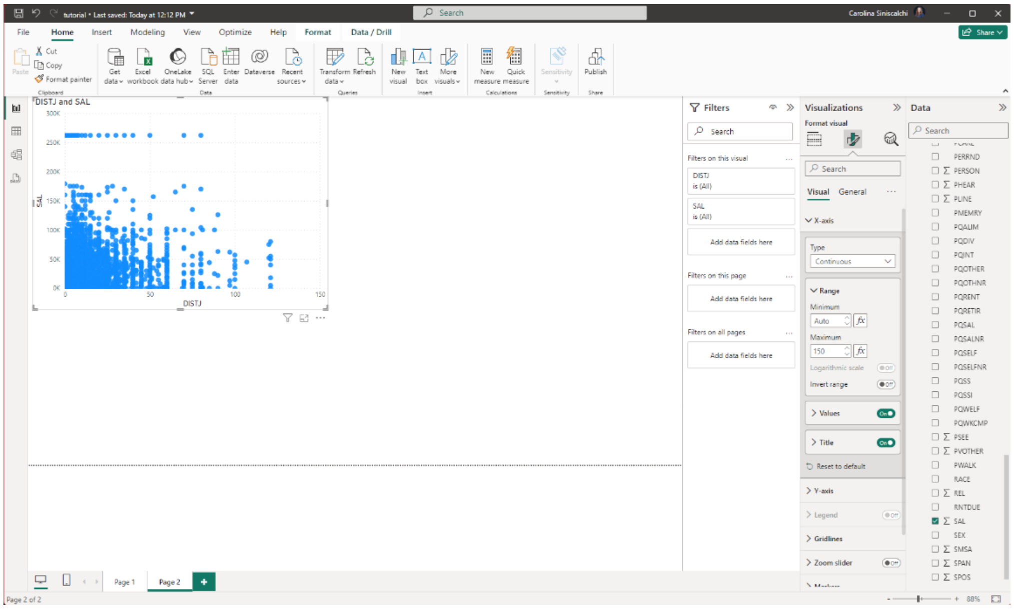
We can go back to Report view. At the bottom of the page, we can create a new page on the report by clicking in the green tab with the plus sign. If we look at the Data tab on the right side now, there are two data files available. Let's look at the relationship between salary (variable SAL) and the distance the individuals travel to work from home to work (in file 0002, variable DISTJ). We'll build a scatter plot this time. Click on the scatter plot icon (circled red). When the visual appears, we can load the X-axis. We'll drag the variable DISTJ from file 0002 to the X-axis field and click on "Don't summarize" (red arrow) .



Then, we'll drag the variable SAL from file 0006 to the Y-axis field. We'll choose "Don't summarize" too. We immediately see that there's a group of outlier values (red circle). Circling back to the codebook, with some inspection of several variables, we figure out that for people that reported as working from home, the distance and time to work variables received the value 996. We can safely remove these values from our plot.



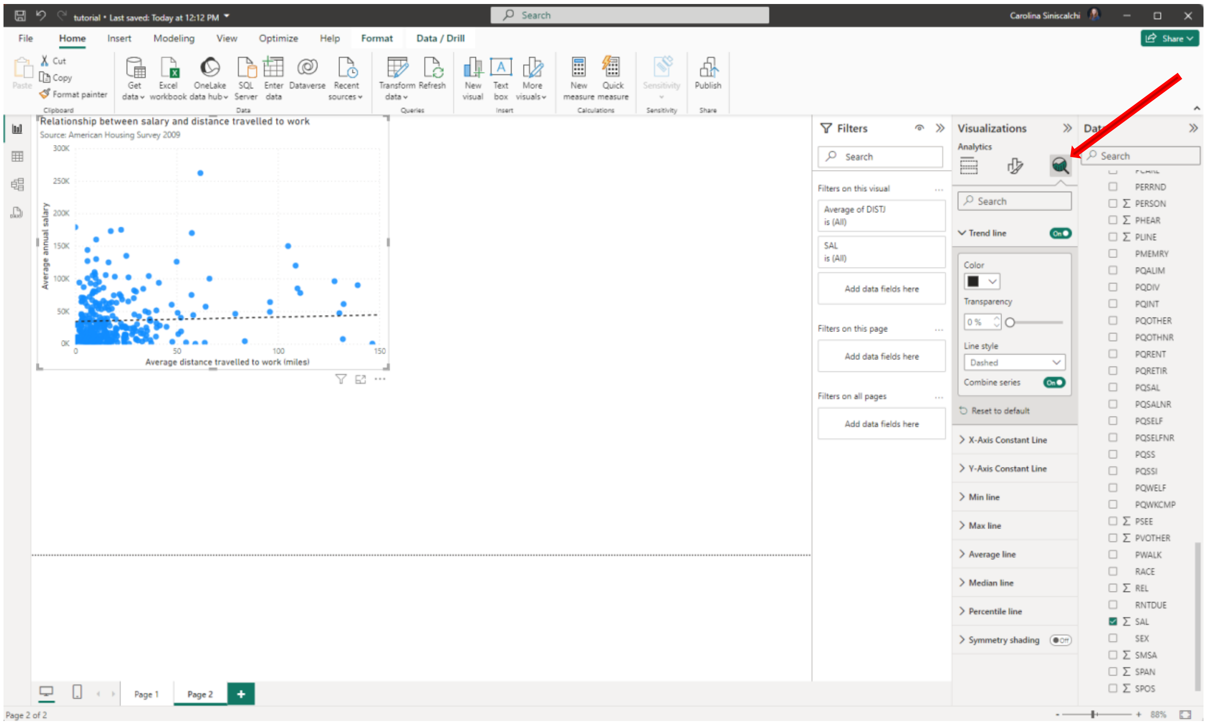
The easiest way to do this is to limit the maximum value of the X-axis, using the Format visual tab. It looks better now. We can also play with the summarization of the distance variable here, such as changing between no summarization and average. Average gives a cleaner look. We can also use the Format visual tab to fix the labels and title.



Filters on this visual
DISTJ is (All)
SAL is (All)

Visualizations — Format visual — Visual
X-axis
Type: Continuous
Range
Minimum: Auto
Maximum: 150
Logarithmic scale: Off
Invert range: Off
Values: On
Title: On

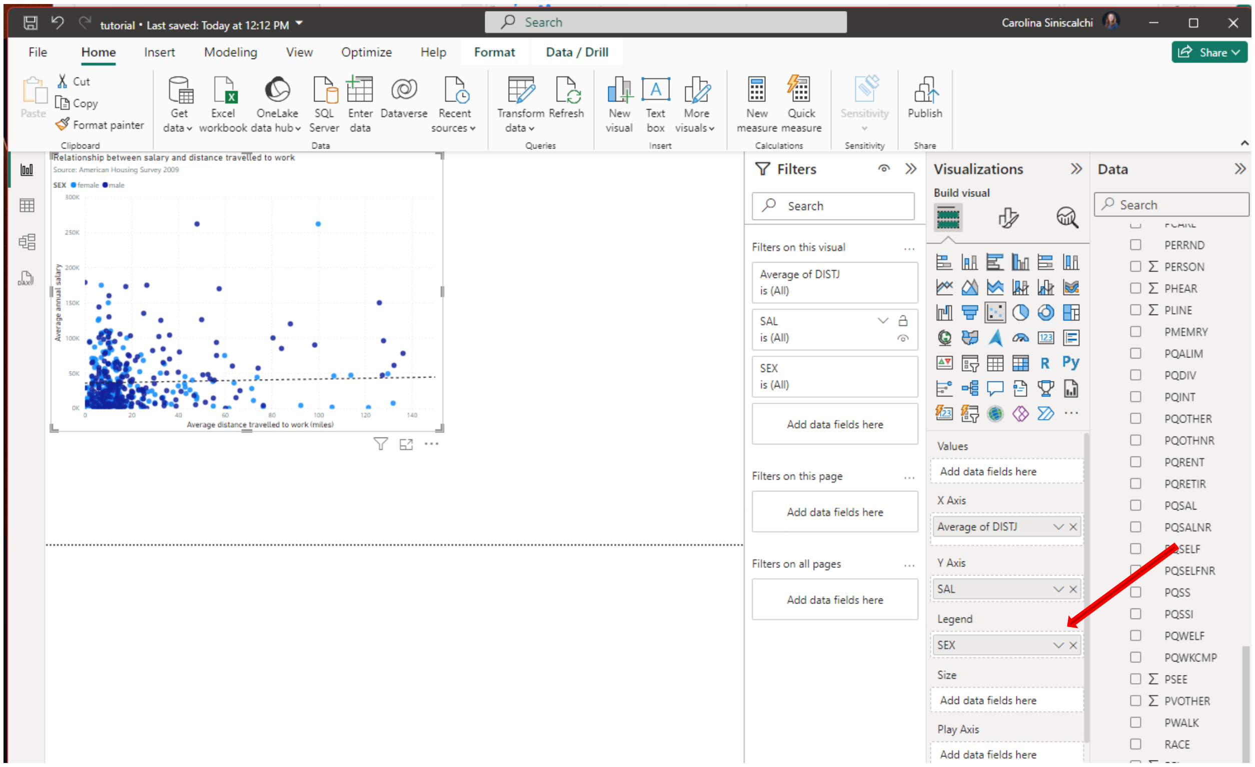
A common feature of scatter plots is a trend line, calculated through a linear equation. This can be easily added by turning on the Trend line feature in the Visualizations > Analytics tab (red arrow).



Filters on this visual
Average of DISTJ is (All)
SAL is (All)

Visualizations — Analytics
Trend line: On
Color: ■
Transparency: 0 %
Line style: Dashed
Combine series: On
Symmetry shading: Off

We can add even more information to this plot by adding another variable to the "Legend" field in the Build visual tab. We'll drag the variable SEX there. When we do that, the dots change color to represent the different categories.



Average of DISTJ is (All)

SAL is (All)

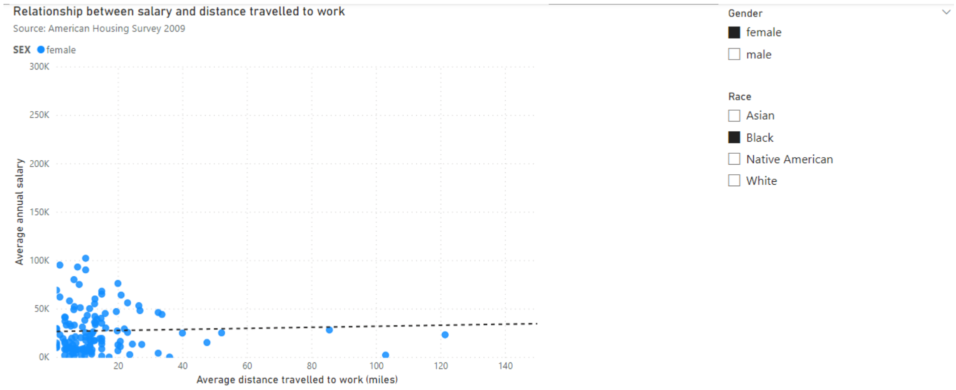
SEX is (All)

X Axis: Average of DISTJ

Y Axis: SAL

Legend: SEX

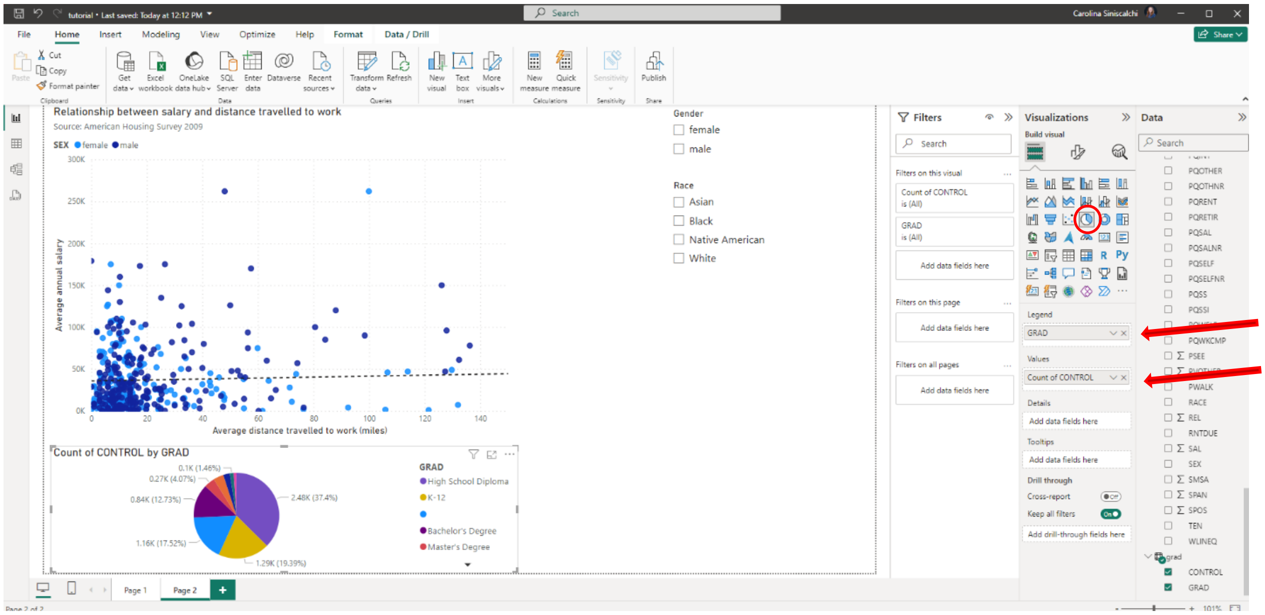
Because both files are linked, we can use the same slicer filters from the previous page here. We'll add them by going back to the first page, click both slicer cards, right click and select Copy > Copy visuals. Then we go to the second page and paste the visuals. A pop-up window will appear asking if you want to Sync the visuals between pages. Click on Sync. Now we can apply the slicers to this visual as well. Note that the trend line also changes when the slicers are applied.



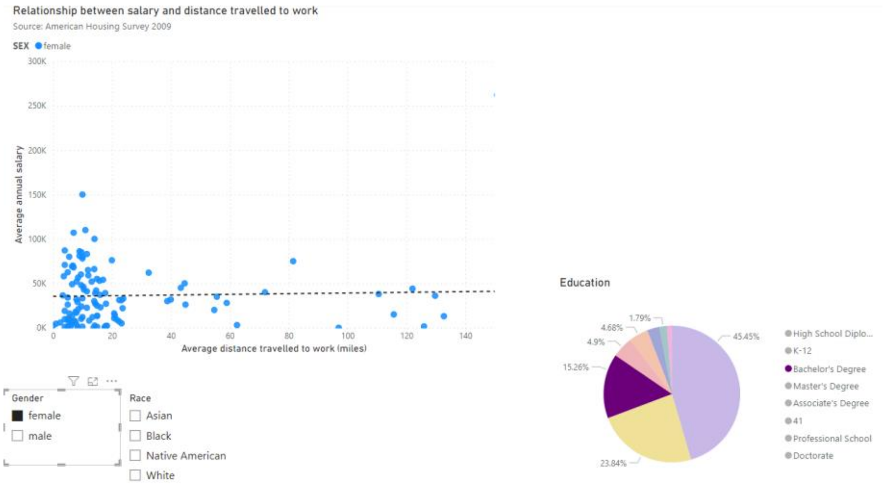
Gender
■ female
□ male

Race
□ Asian
■ Black
□ Native American
□ White

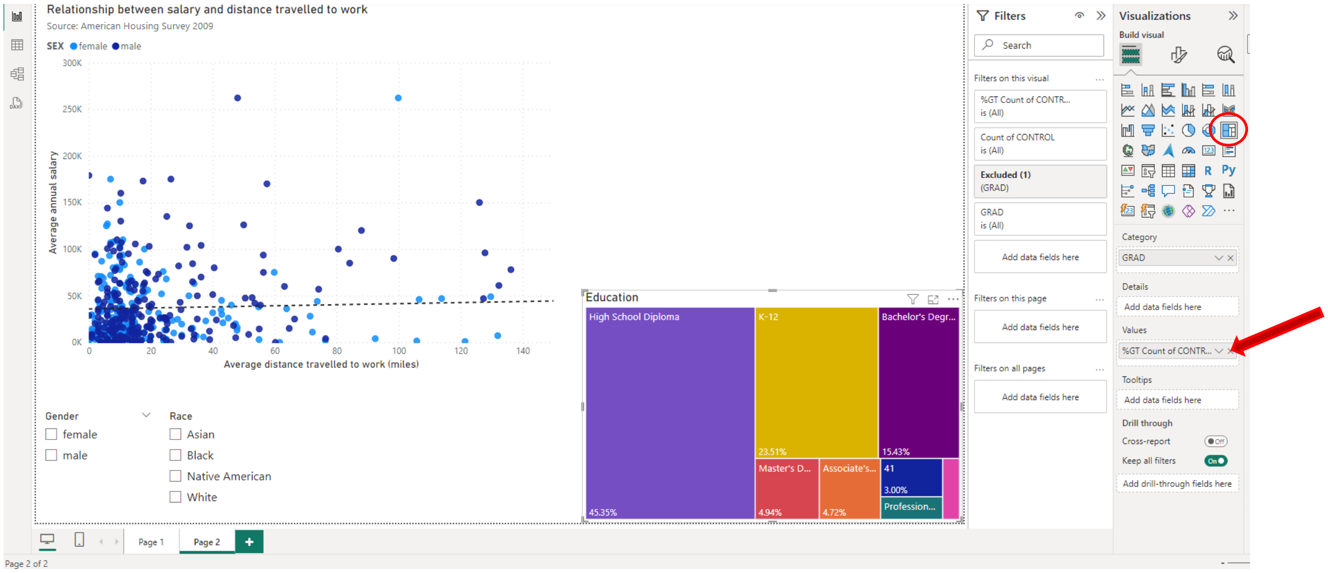
Now imagine we want to explore the relationship between the level of education achieved and journey to work. We can add another visual showing the proportion of people in each level of education that can act as a filter on our scatter plot. The variable that contains this information is on file 0006 and is called GRAD. This categorical variable is also coded with numbers, like RACE, which means we'll need to transform our data again. The variable has more than 20 categories, but to make it easier we'll aggregate some of the categories. To do that, we'll go to Power Query and do the same replace steps we did above. One option of visual that we can use is a Pie chart (red circle). We'll click on the icon on the Build visual table, and then drag the variable GRAD to the Legend field and CONTROL to the values. Make sure that CONTROL is summarized as Count (red arrows).



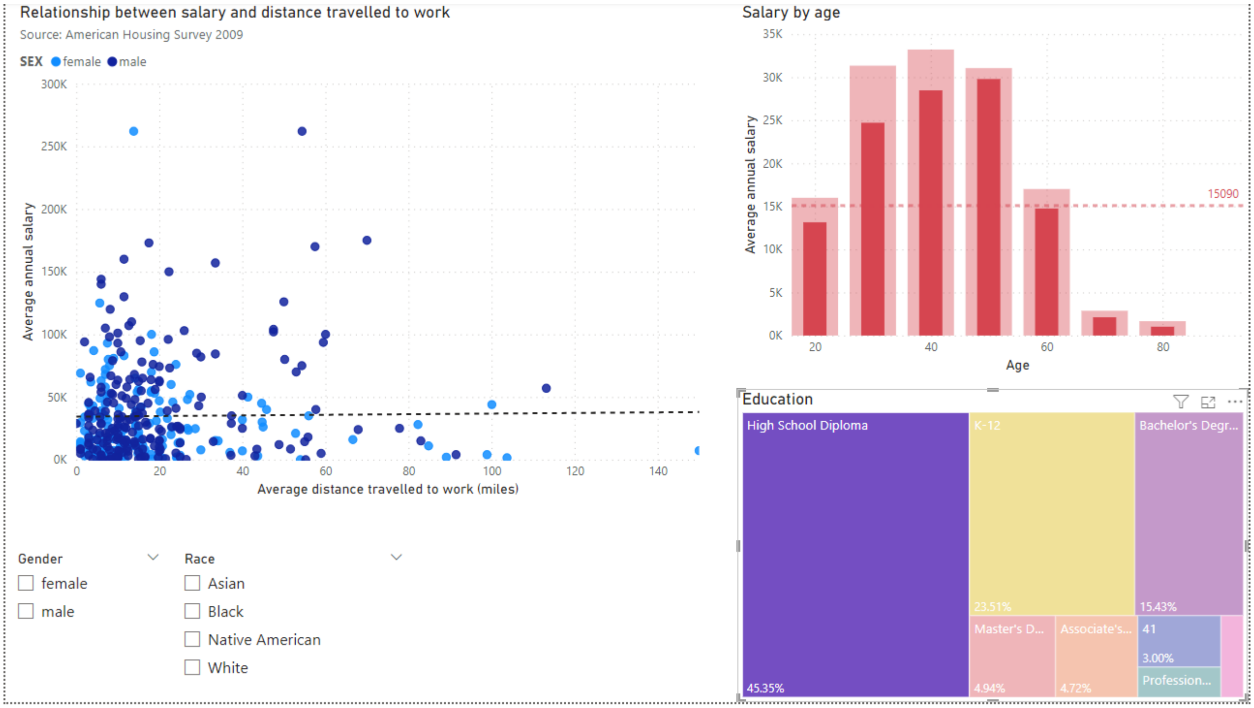
Note that the large blue slice appears blank on the visual legend. That's because these are empty values. To remove them, we can right click on the slice and click Exclude. In the Format visuals tab, we can change how the labels are displayed, the colors of the slices, etc. Notice that if we click in a specific slice, the scatter plot syncs up with the pie chart. The slicers also work across the visuals.



Another type of plot we could use here is a Treemap. We can switch between types by clicking on the visual area, then clicking on the Treemap icon (red circle) on the Build visual tab. To show the values in percentages, click on the down arrow on Values, then Show value as > Percentage of the grand total. Treemaps are more friendly than pie charts because we interpret squares volumes more easily than slices.



To make our visual even more complete, we could copy the salary by age plot from the other page. Again, click on it, Copy > Copy visual, then paste on page 2. You'll notice as you click around that all visuals are sync up on each other.



Finally, if you are done with your report, you can save it as a pbix file, which can be open on Power BI, or export it as a static PDF. To be able to share the interactive report as a web-based file, a paid subscription is needed, which we do not have access to right now.

Thank you for following this tutorial! I hope it is useful to you!

References:

United States. Bureau of the Census. American Housing Survey, 2009: New Orleans Data. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 2016-04-18. https://doi.org/10.3886/ICPSR30943.v1

American Housing Survey: https://www.census.gov/programs-surveys/ahs/about.html