The Role of Data Visualization in Big Data Analytics

Chong Ho (Alex) Yu, Ph.D., D. Phil.

Professor of Behavioral and Applied Science

Azusa Pacific University

Paper Presented at 2020 IM Data Conference

imdata.
Innovative Methods with Data Science & AI

Potential problems in Big Data Analytics/Data Science

- Over-rely on or pre-maturely hand over our judgment to AI and machine learning (Black Box) without examining (exploring) the data pattern carefully.
- Over-plotting: two many observations
- Curse of dimensionality: too many variables

imdata.
Innovative Methods with Data Science & AI

Data mining and Exploratory Data Analysis (EDA)

- Data mining is not entirely new; it is an extension of EDA; EDA is the precursor of data mining.
- A philosophy about how data analysis should be carried out, rather than being a fixed set of techniques.
- Pattern-seeking
- Skepticism (detective spirit): John Tukey (not Turkey) suggested explore the data in as many ways as possible until a plausible story of the data emerges (Like a detective).

EDA and data mining

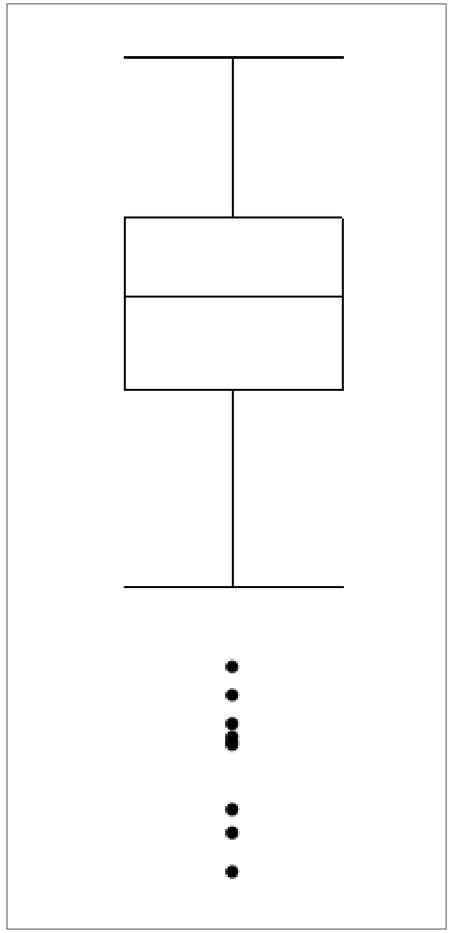
- Same:
  - Data mining is an extension of EDA: it inherits the exploratory spirit; don't start with a preconceived hypothesis or theory.
  - Both heavily rely on data visualization.
- Difference:
  - DM: Use machine learning and resampling
  - DM: More robust
  - DM: Deal with much bigger sample size

Elements of EDA: 4Rs

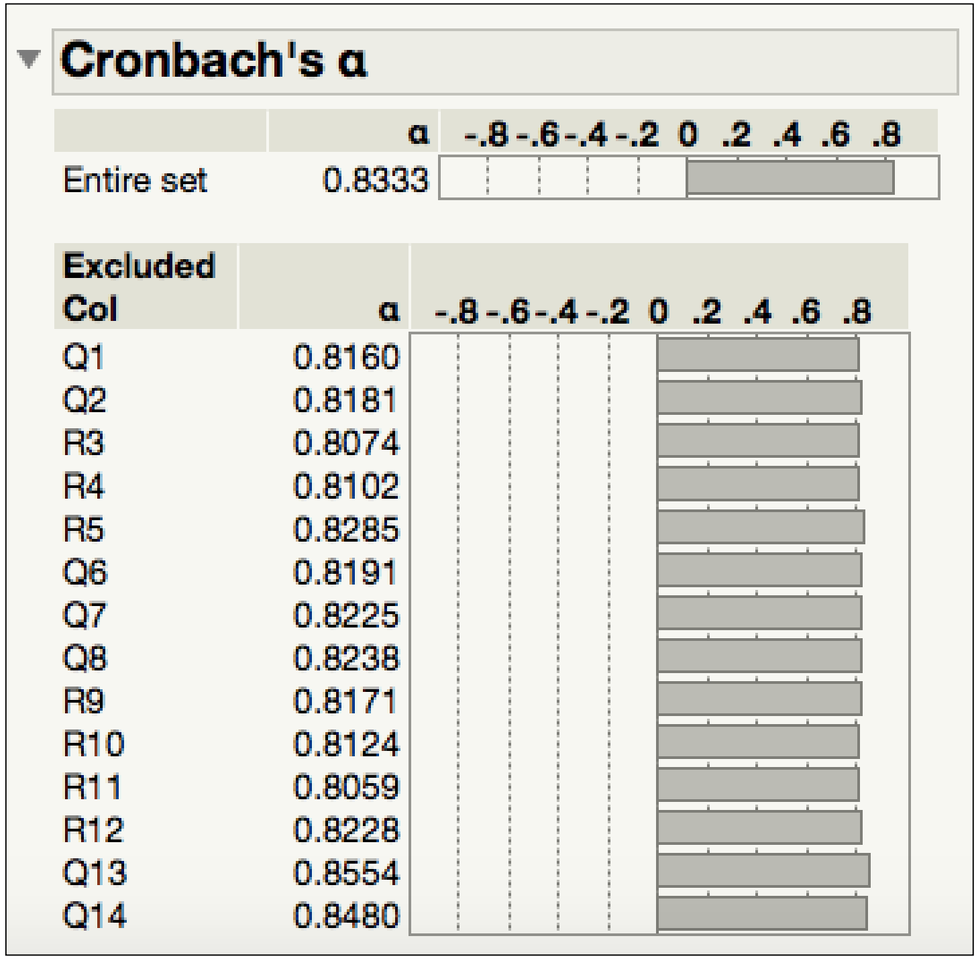
- Velleman & Hoaglin (1981):
  - Residual analysis
  - Re-expression (data transformation)
  - Resistant
  - Display (revelation, data visualization)

Data visualization

- One of the great inventions of graphical techniques by John Tukey is the **boxplot**.
  - It is resistant against extreme cases (use the median)
  - It can easily spot outliers.
  - It can check distributional assumption using a quick 5-point summary.



Data visualization



Cronbach's α

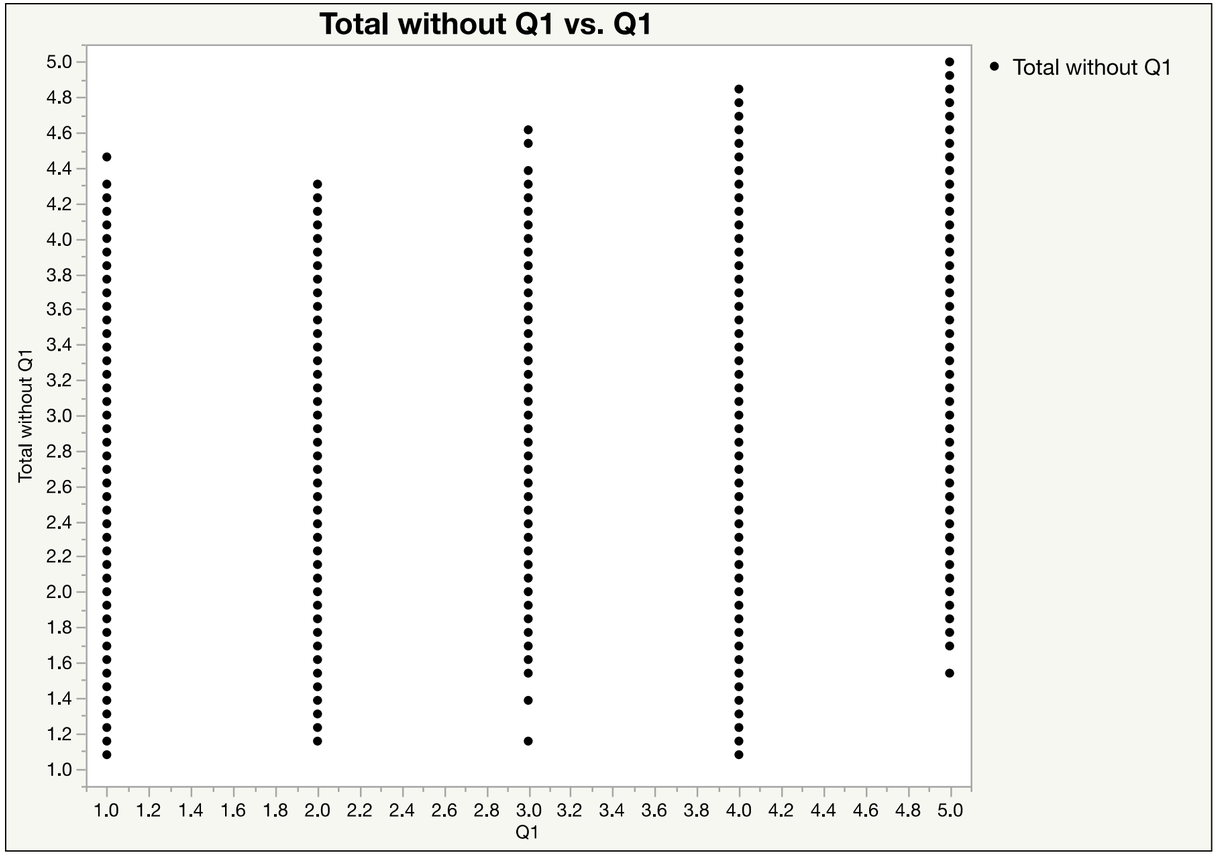
	α	-.8 -.6 -.4 -.2 0 .2 .4 .6 .8
Entire set	0.8333	

Excluded Col	α	-.8 -.6 -.4 -.2 0 .2 .4 .6 .8
Q1	0.8160	
Q2	0.8181	
R3	0.8074	
R4	0.8102	
R5	0.8285	
Q6	0.8191	
Q7	0.8225	
Q8	0.8238	
R9	0.8171	
R10	0.8124	
R11	0.8059	
R12	0.8228	
Q13	0.8554	
Q14	0.8480	

- Boxplot can be used in median-smoothing when there are too any data.
- These are 14,819 responses to the Consideration of Future Consequences Scale (CFCS).
- When the *n* is huge, even random numbers might appear to be good.
- To demonstrate the problem, Q13 and Q14 are randomly generated.
- The Cronbach Alpha looks good! No bad items!

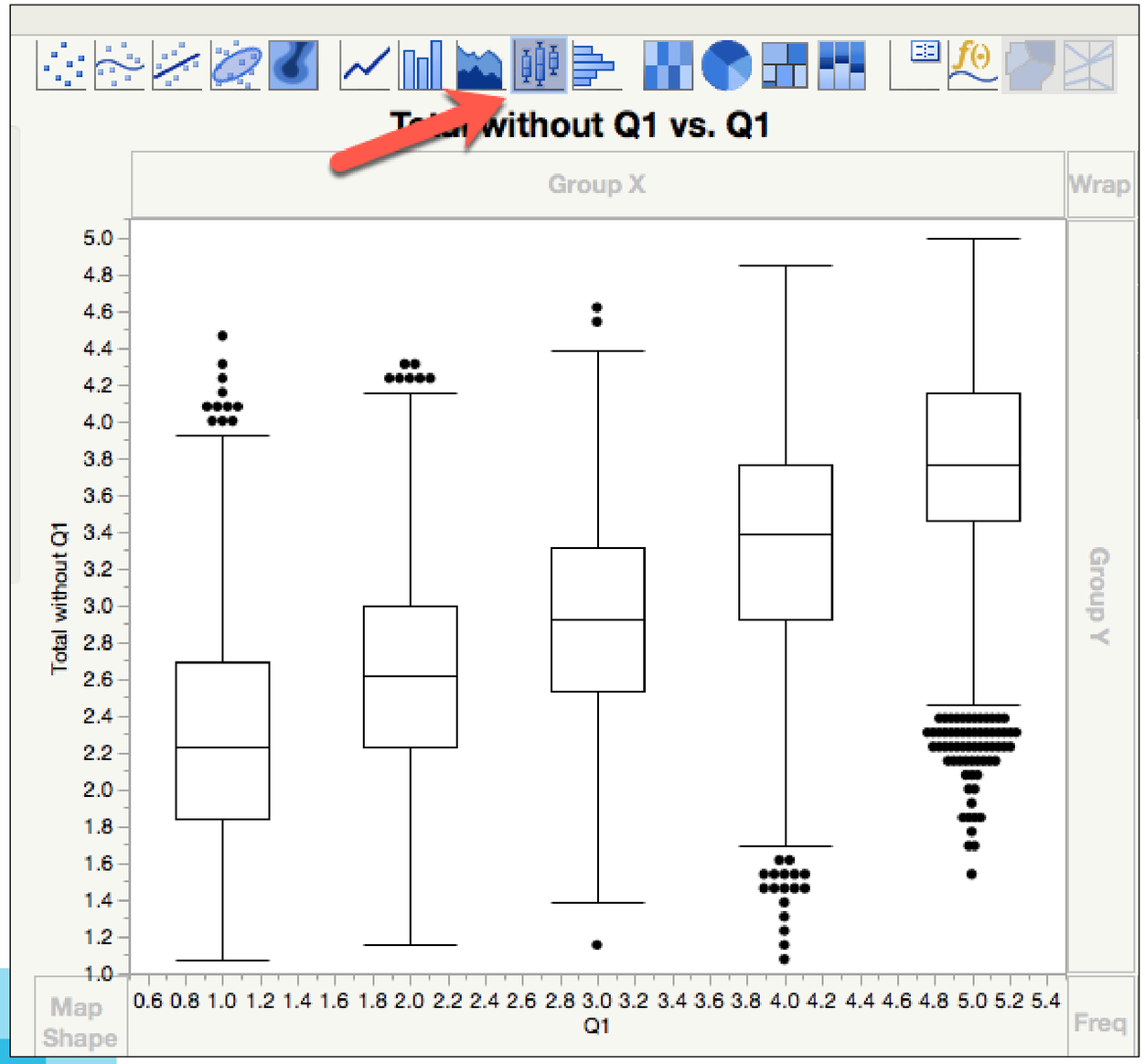
imdata.
Innovative Methods with Data Science & AI

Data visualization



- Data visualization can distinguish patterns from noise.
- I want to know whether Q1 is strongly correlated with the total scale (Q2-Q14).
- Item-total correlation
- If the response pattern of Q1 is correlated with the total, then they are internally consistent.
- Too many data points!

imdata.
Innovative Methods with Data Science & AI

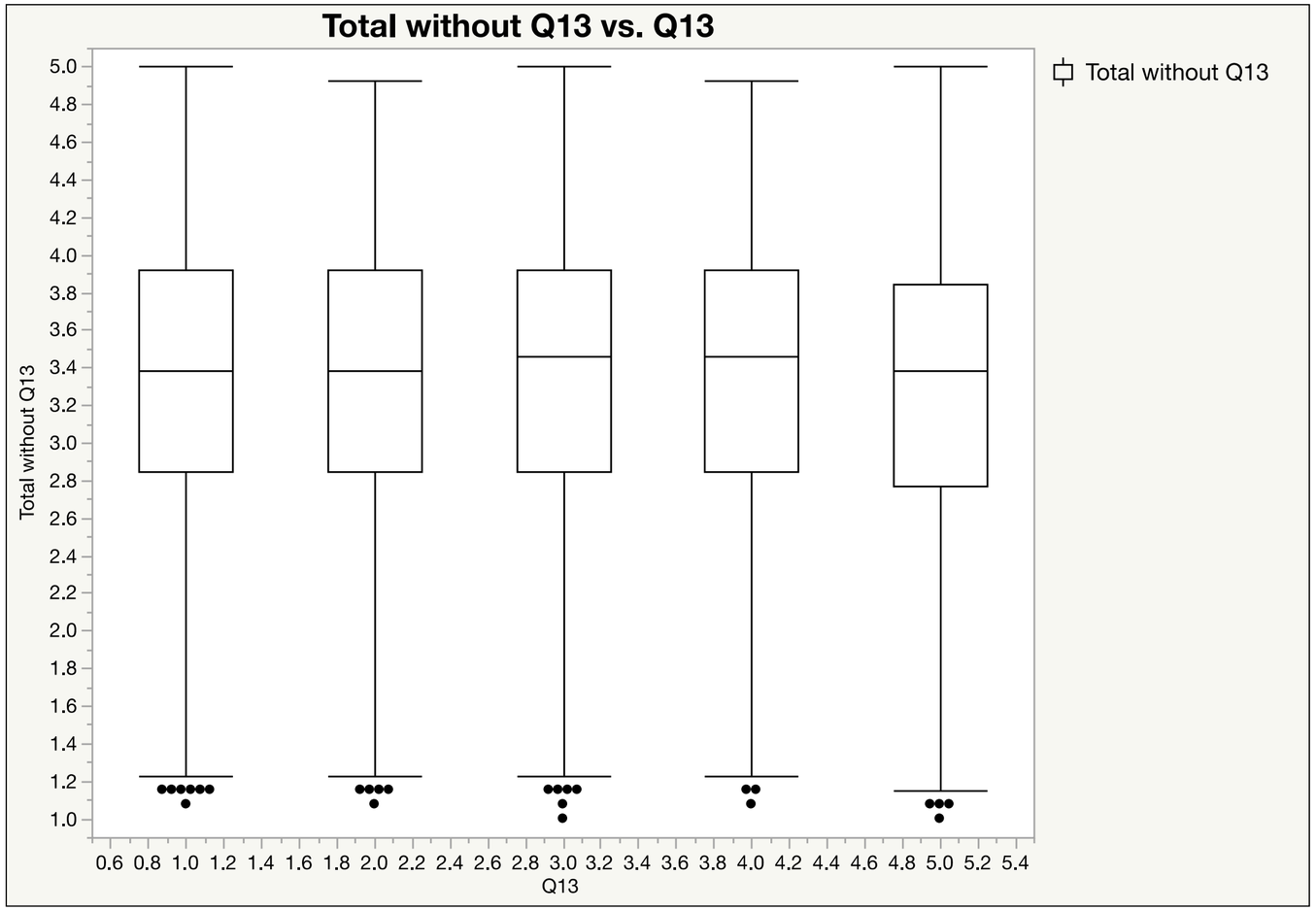


Data visualization

- Change the display option to Boxplot.
- Now the median of the total by each category of Q1 is shown.
- There is a positive relationship between Q1 and all other items.

imdata.
Innovative Methods with Data Science & AI

Data visualization



- When I use median – smoothing to check the association between Q13 and total without Q13., it was found that there is no relationship.
- The number can fool your mind, but the graph cannot fool your eyes.

imdata.
Innovative Methods with Data Science & AI

What is data visualization?

- The process of exploring and displaying data in a manner that builds a visual analogy in the service of researcher insight and learning.
- A common ground shared by EDA and data mining: The data visualizer should explore the data in as many ways as possible until a plausible story of the data emerges.
- For the purpose of exploration the tool should be interactive and dynamic.

Static vs. dynamic

- Static
  - What you see is what you get. The graph is frozen.
  - After the graph is made, you cannot manipulate the graph (changing the background color or the line width is not considered “data manipulation” because it cannot reveal any insight about the data)
- Dynamic
  - The data table and different graphic panels are linked. Changing one would change all others.
  - You can manipulate the graph to explore the data through different perspectives.

Dynamic graph

- Allow you to ask what-if questions:
  - What if I remove outliers?
  - Auto-refresh the model
- Combine multiple dimensions to see a holistic picture: Switch variables for comparison
- Divide and conquer: Localize the subgroups
- Example: COVID19 by US States

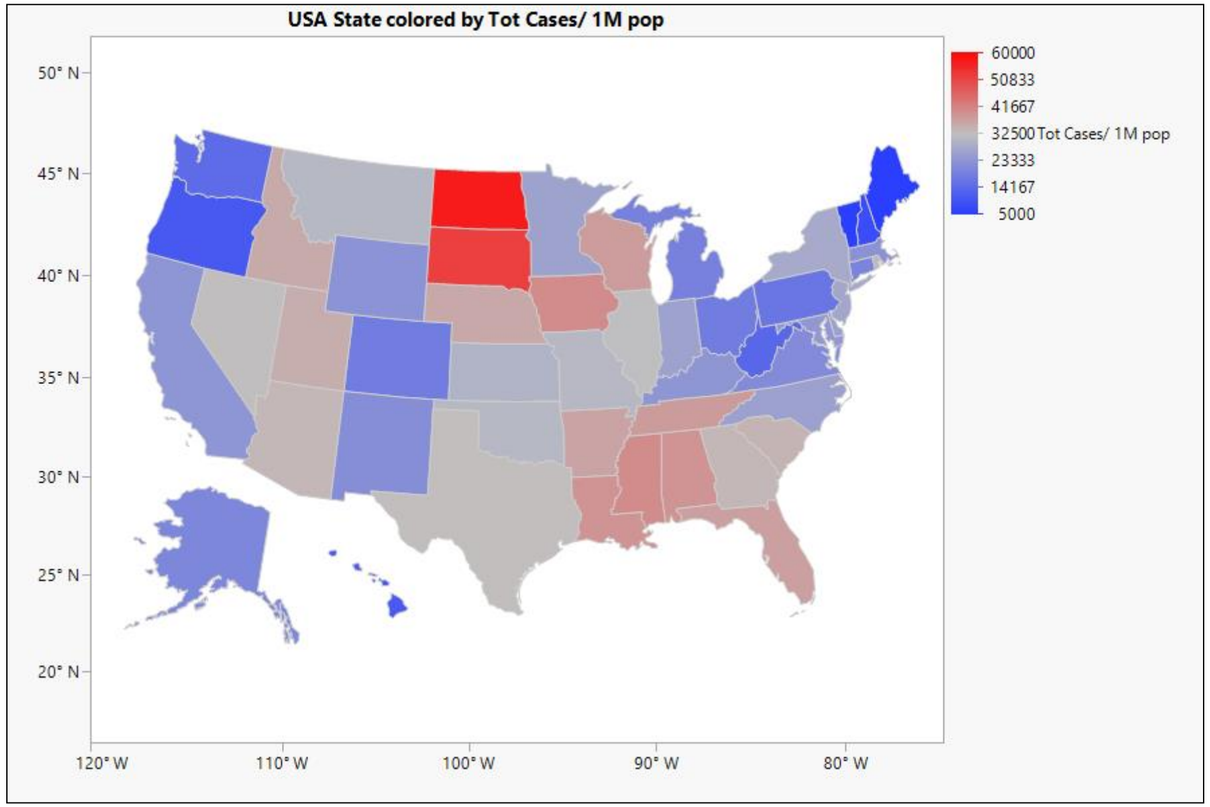
Dynamic graph

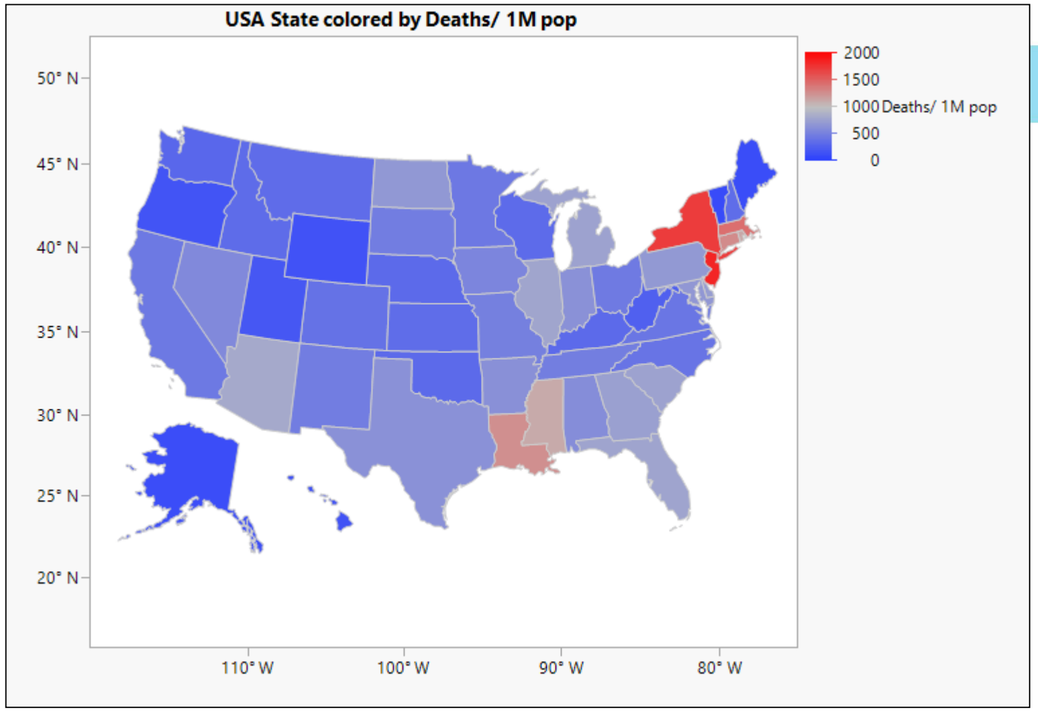
- Example: COVID19 by US States
- I am not endorsing or opposing any political party; I just show the data as is.
- The **Cook Partisan Voting Index (PVI)** is a measurement of how strongly a United States congressional district or states leans toward the Democratic or Republican Party compared to the national average.

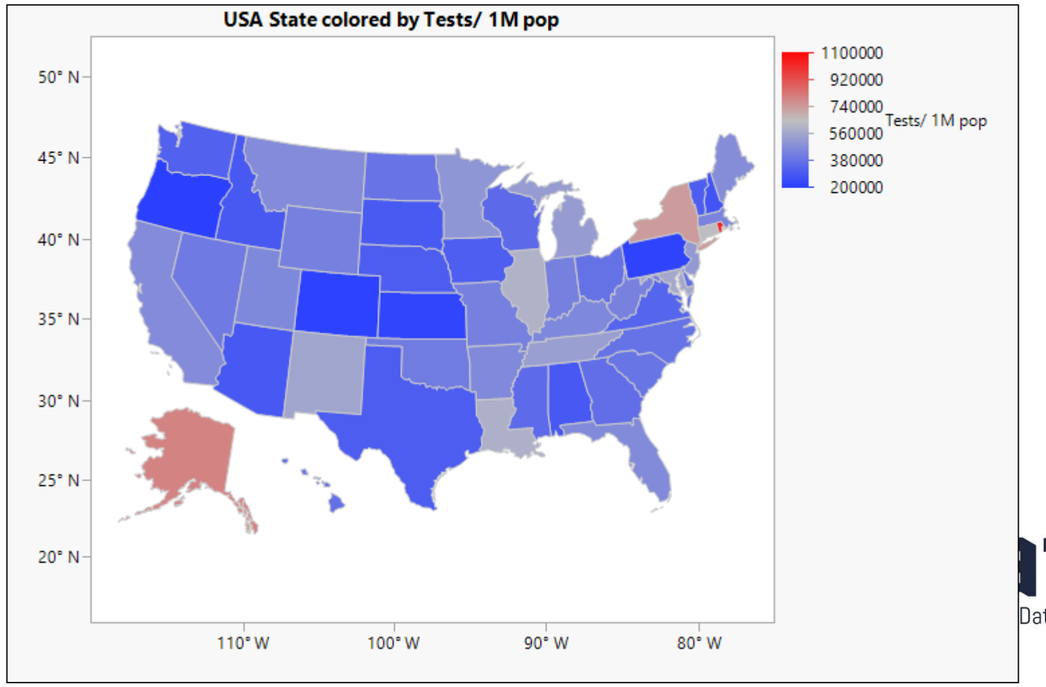
PVI	Rescaled PVI	Lean
R+19	19	Lean towards Republican
R+19	19	Lean towards Republican
R+17	17	Lean towards Republican
R+15	15	Lean towards Republican
R+15	15	Lean towards Republican
R+14	14	Lean towards Republican
R+14	14	Lean towards Republican
R+14	14	Lean towards Republican
R+14	14	Lean towards Republican
R+13	13	Lean towards Republican
R+11	11	Lean towards Republican
R+11	11	Lean towards Republican
R+9	9	Lean towards Republican
R+9	9	Lean towards Republican
R+9	9	Lean towards Republican
R+9	9	Lean towards Republican
R+8	8	Lean towards Republican
R+8	8	Lean towards Republican
R+5	5	Lean towards Republican
R+5	5	Lean towards Republican
R+3	3	Lean towards Republican
R+3	3	Lean towards Republican
R+3	3	Lean towards Republican
R+2	2	Lean towards Republican
Even	0	Even
Even	0	Even
Even	0	Even
D+1	-1	Lean towards Democratic
D+1	-1	Lean towards Democratic
D+1	-1	Lean towards Democratic
D+1	-1	Lean towards Democratic
D+1	-1	Lean towards Democratic
D+3	-3	Lean towards Democratic

imdata.
Innovative Methods with Data Science & AI

Separate maps:

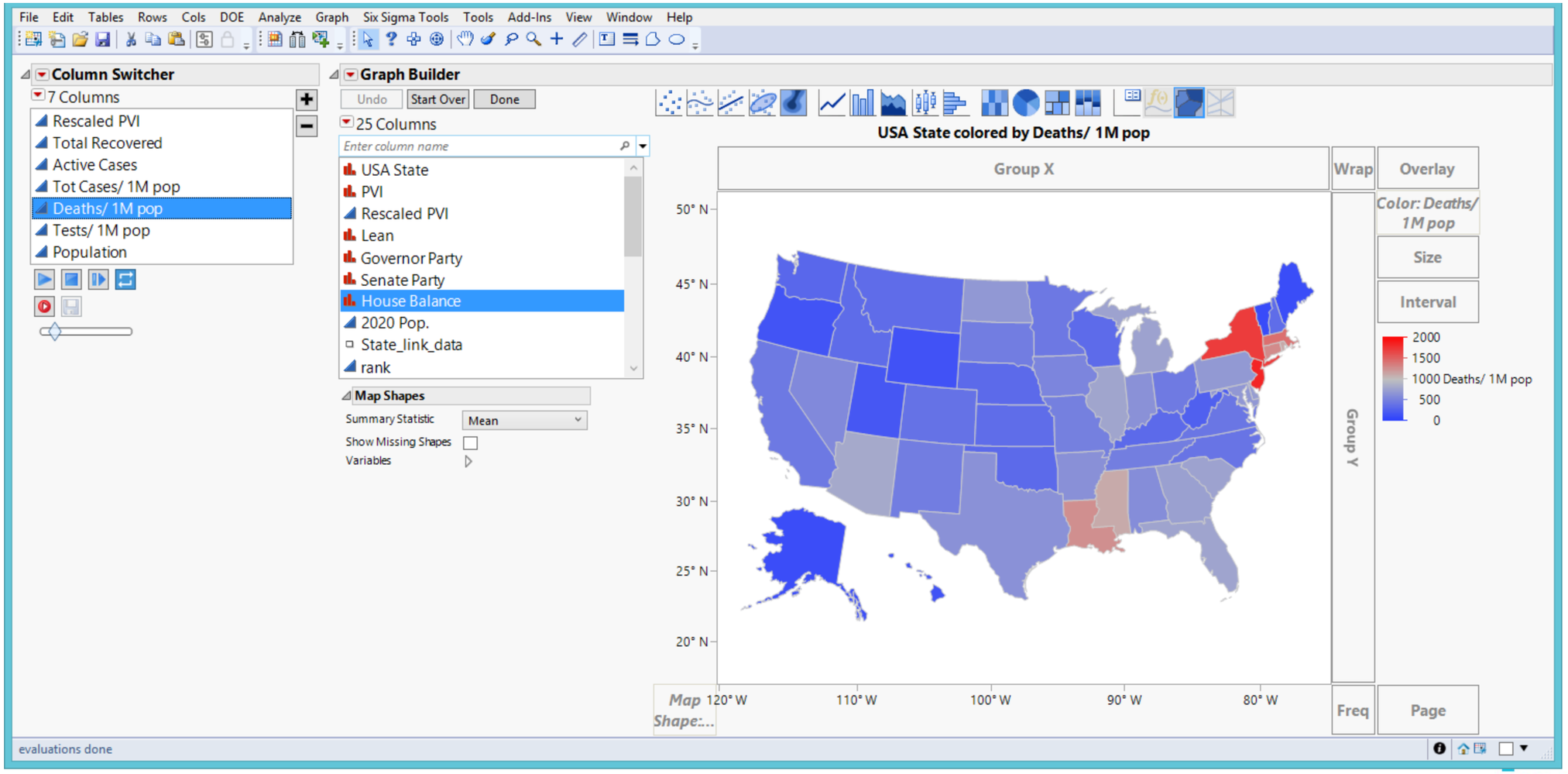






Data Science & AI

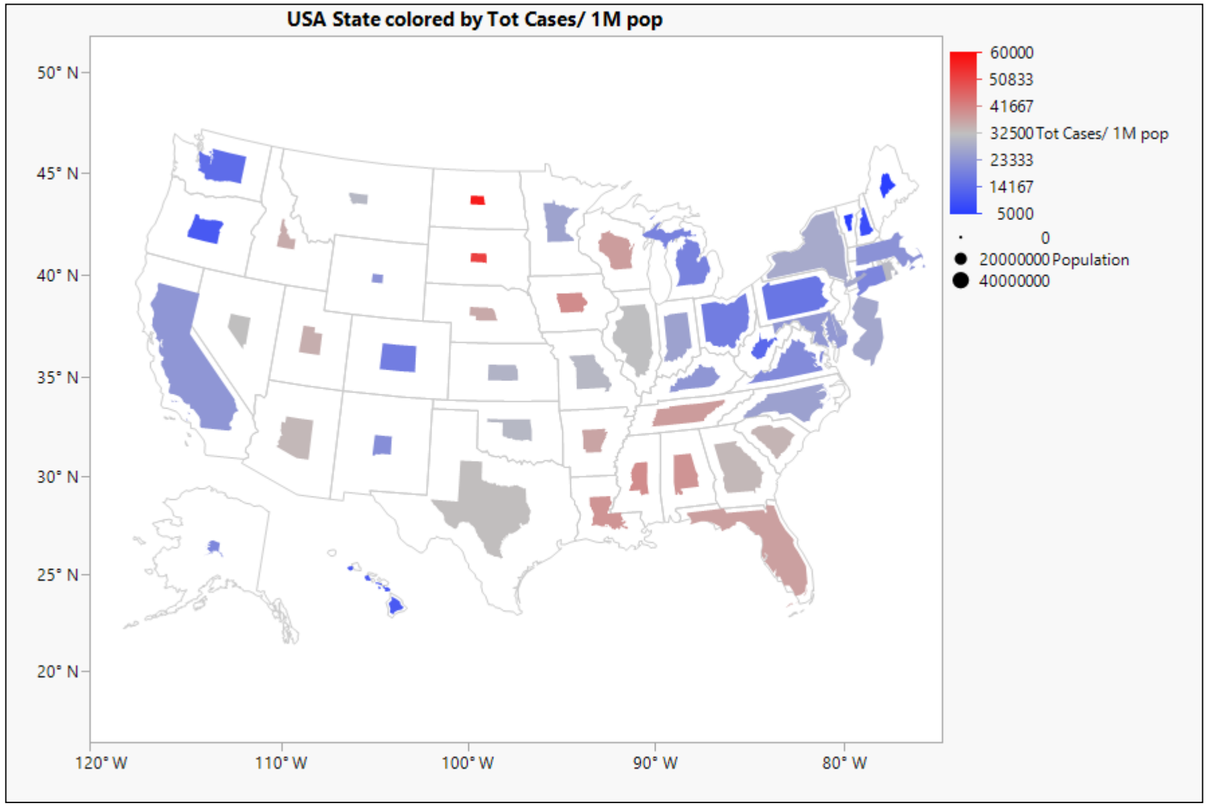
Switch variables



imdata.
Innovative Methods with Data Science & AI

Multiple variables

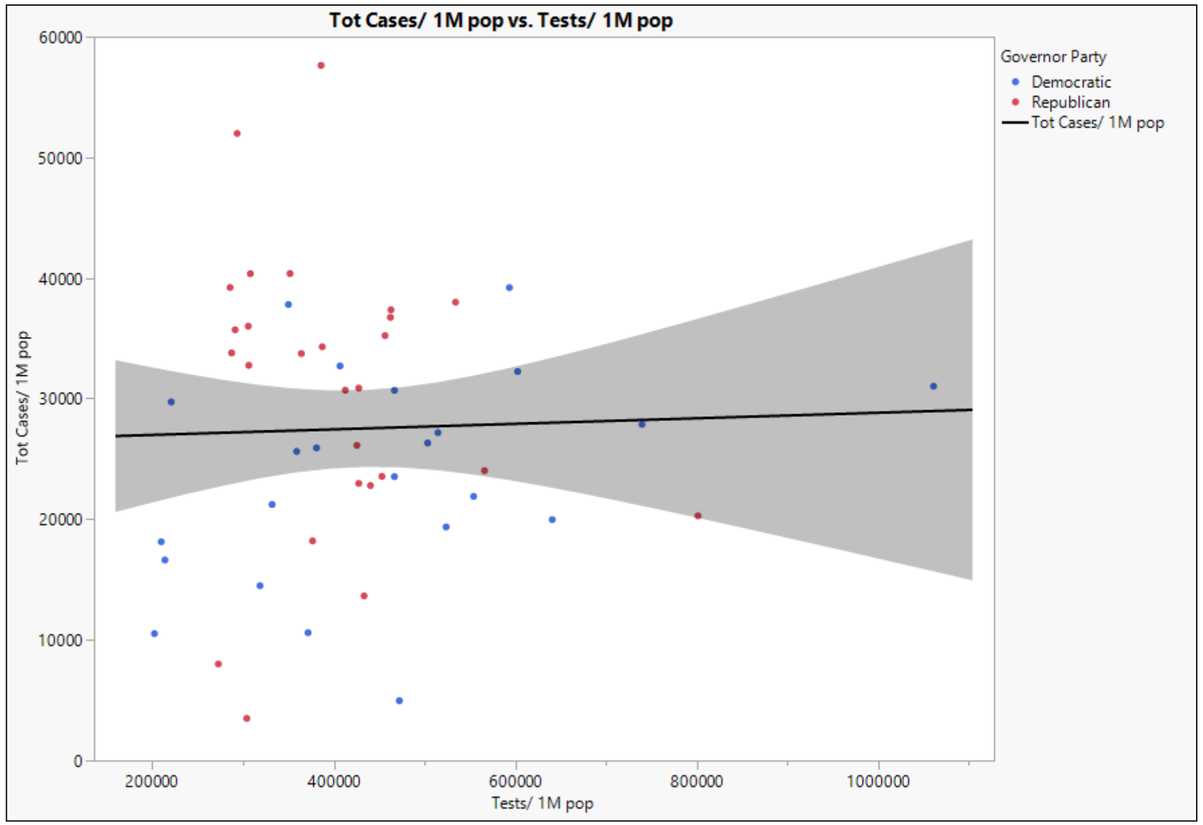
- Using “size” or adding more dimensions into the map is confusing.
- Difficult to see multiple dimensions concurrently
- Curse of dimensionality



imdata.
Innovative Methods with Data Science & AI

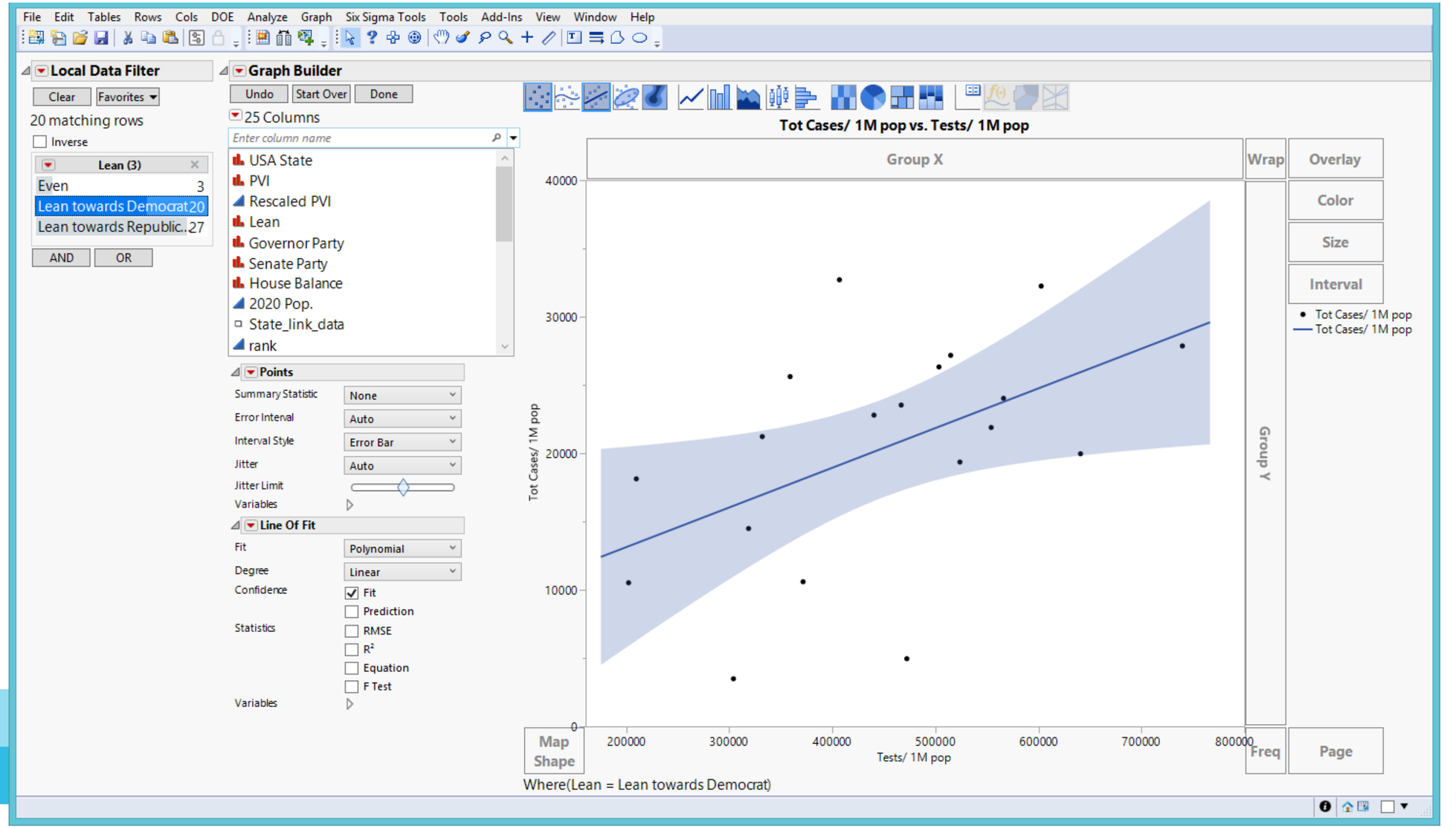
Detect association

- More tests? → more confirmed cases?
- A flat regression line
- No relationship nationwide



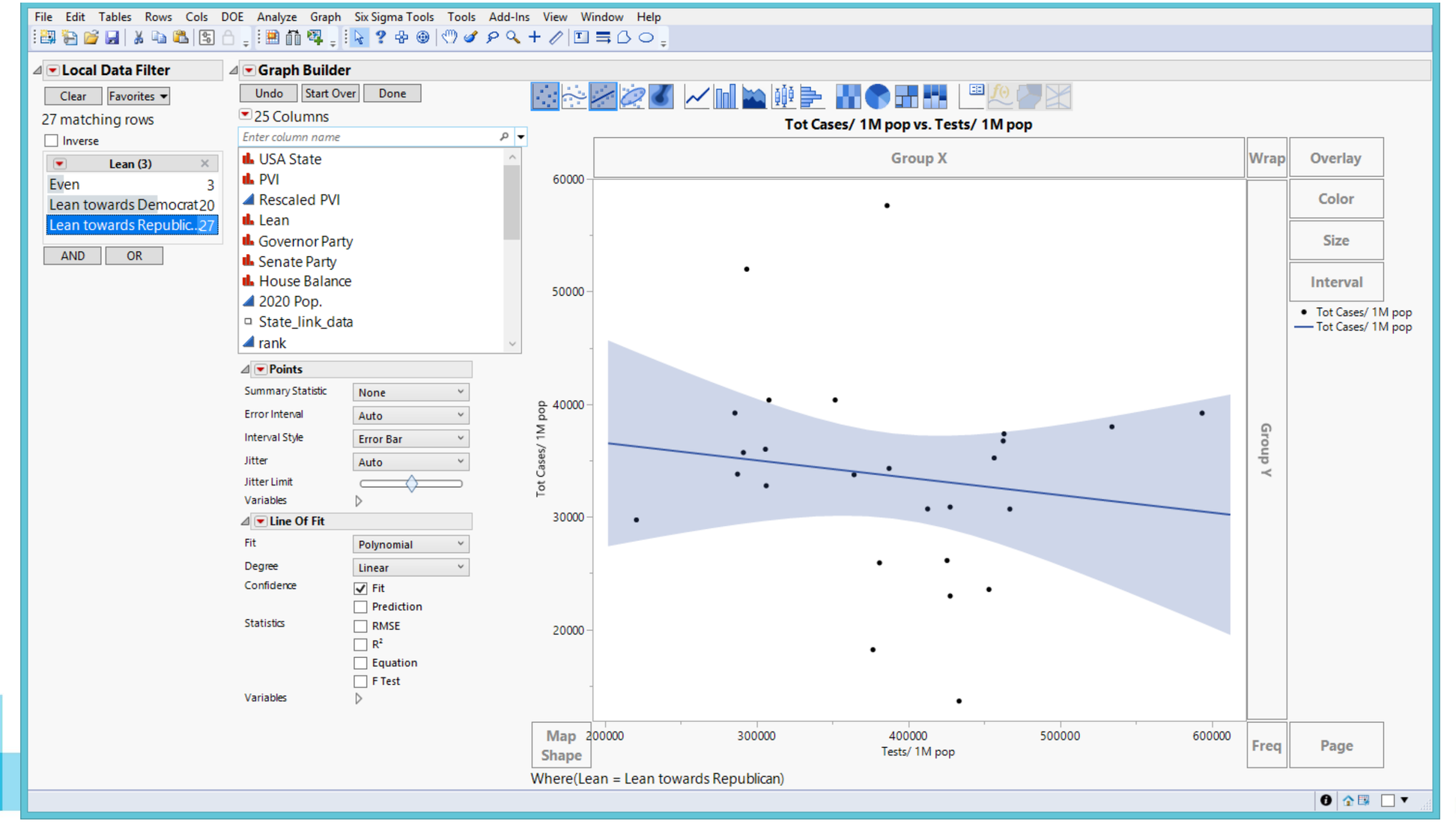
imdata.
Innovative Methods with Data Science & AI

Local filter



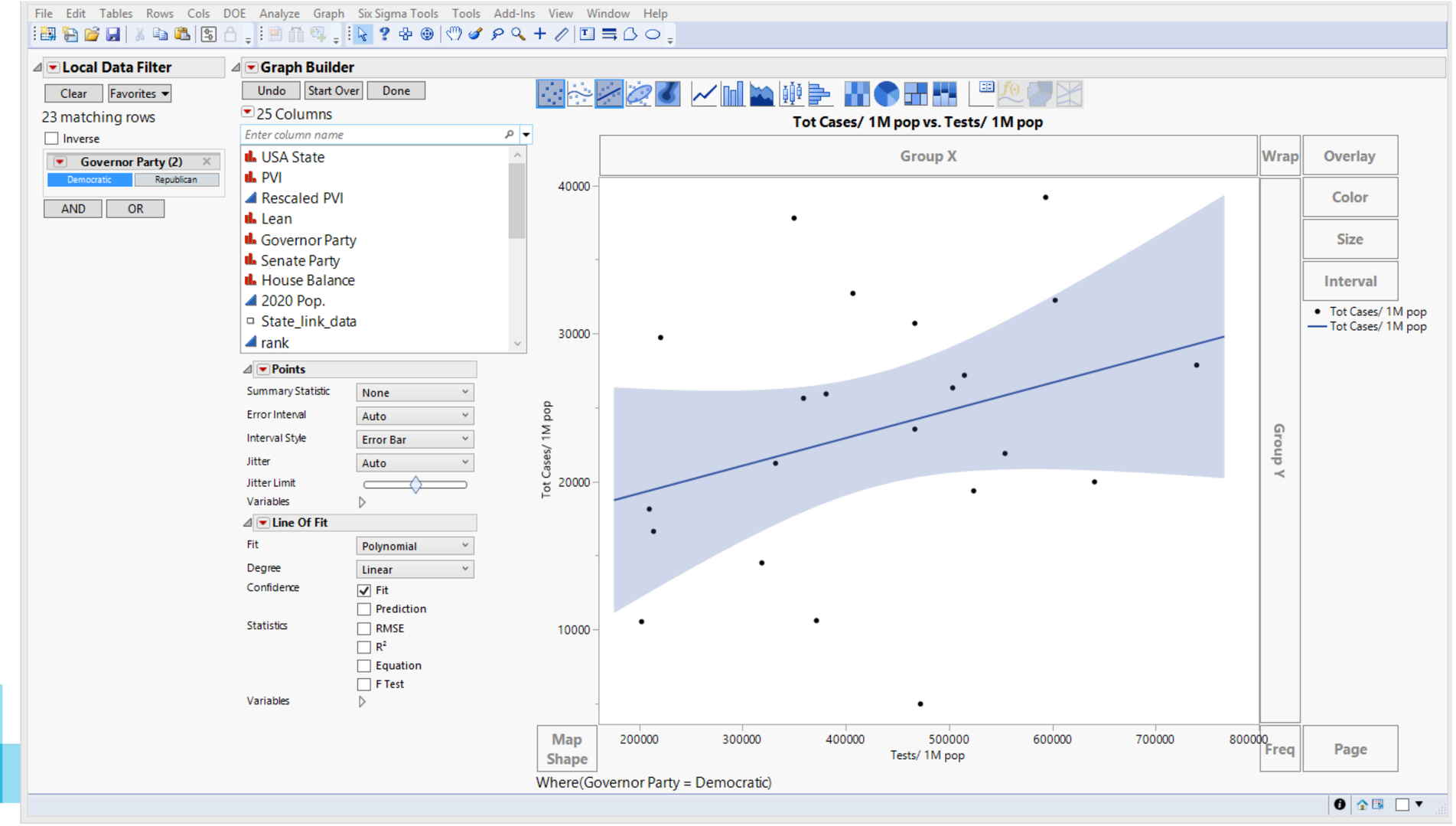
imdata.
Innovative Methods with Data Science & AI

Local filter



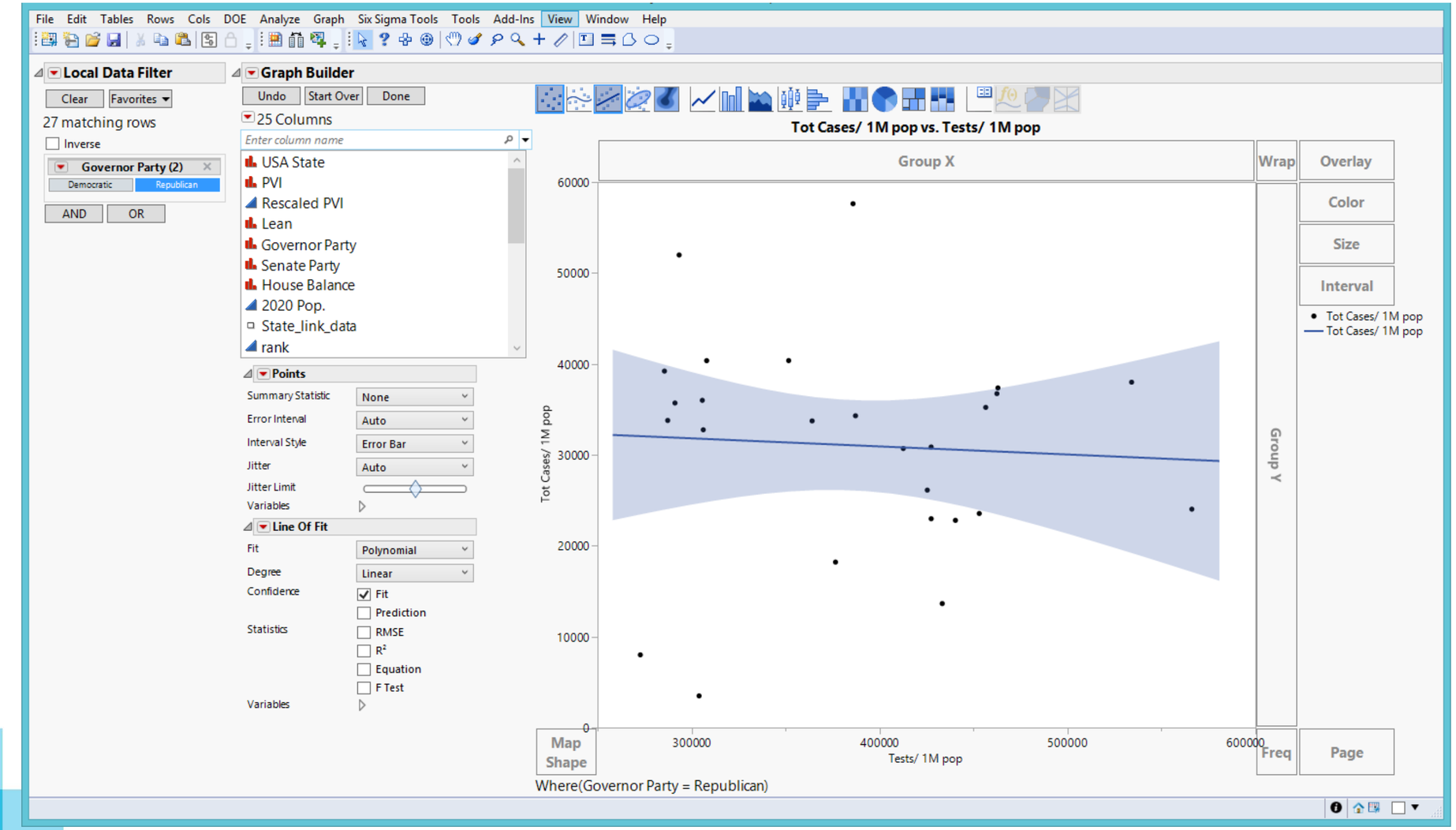
imdata.
Innovative Methods with Data Science & AI

Local filter



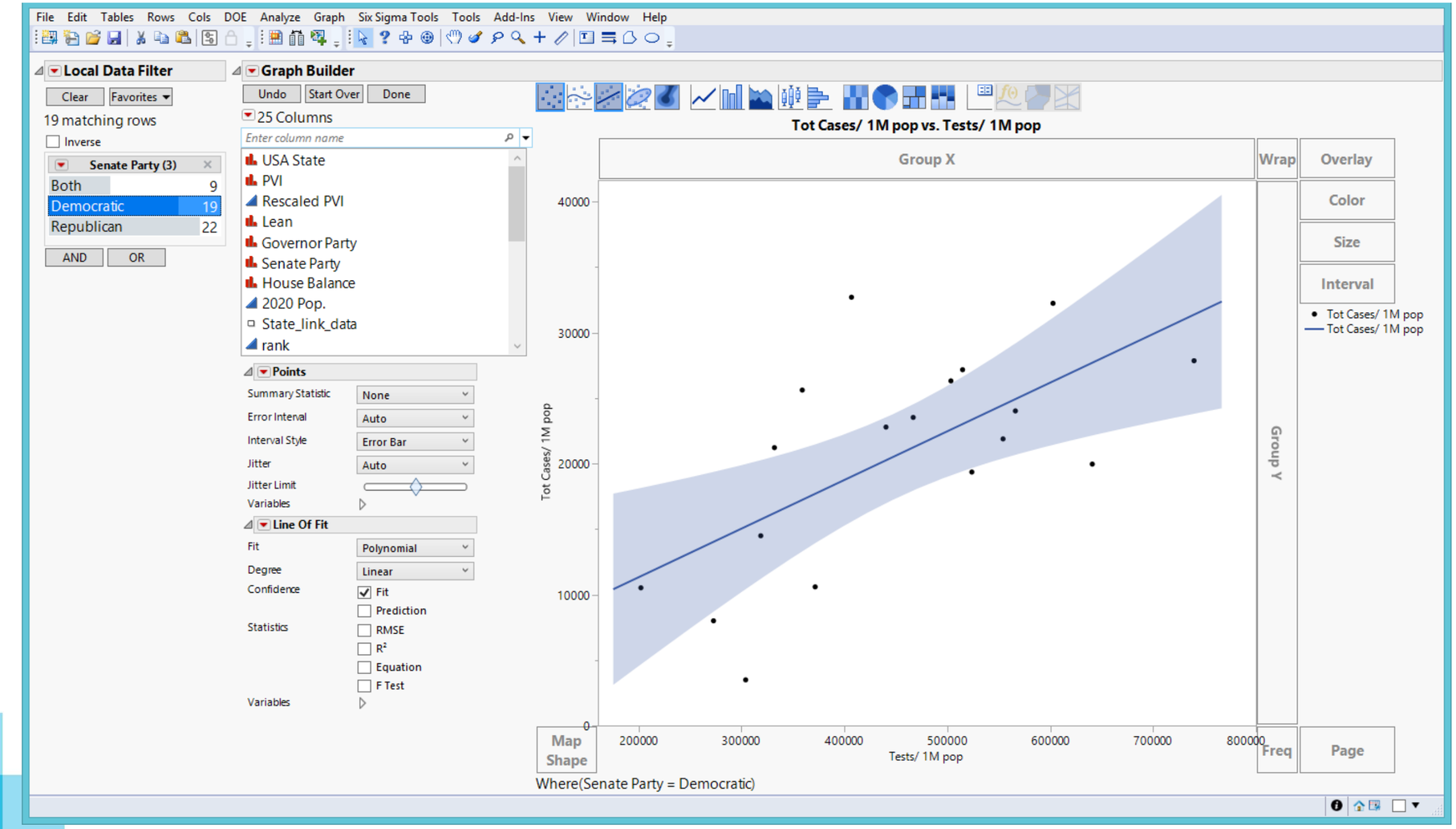
imdata.
Innovative Methods with Data Science & AI

Local filter



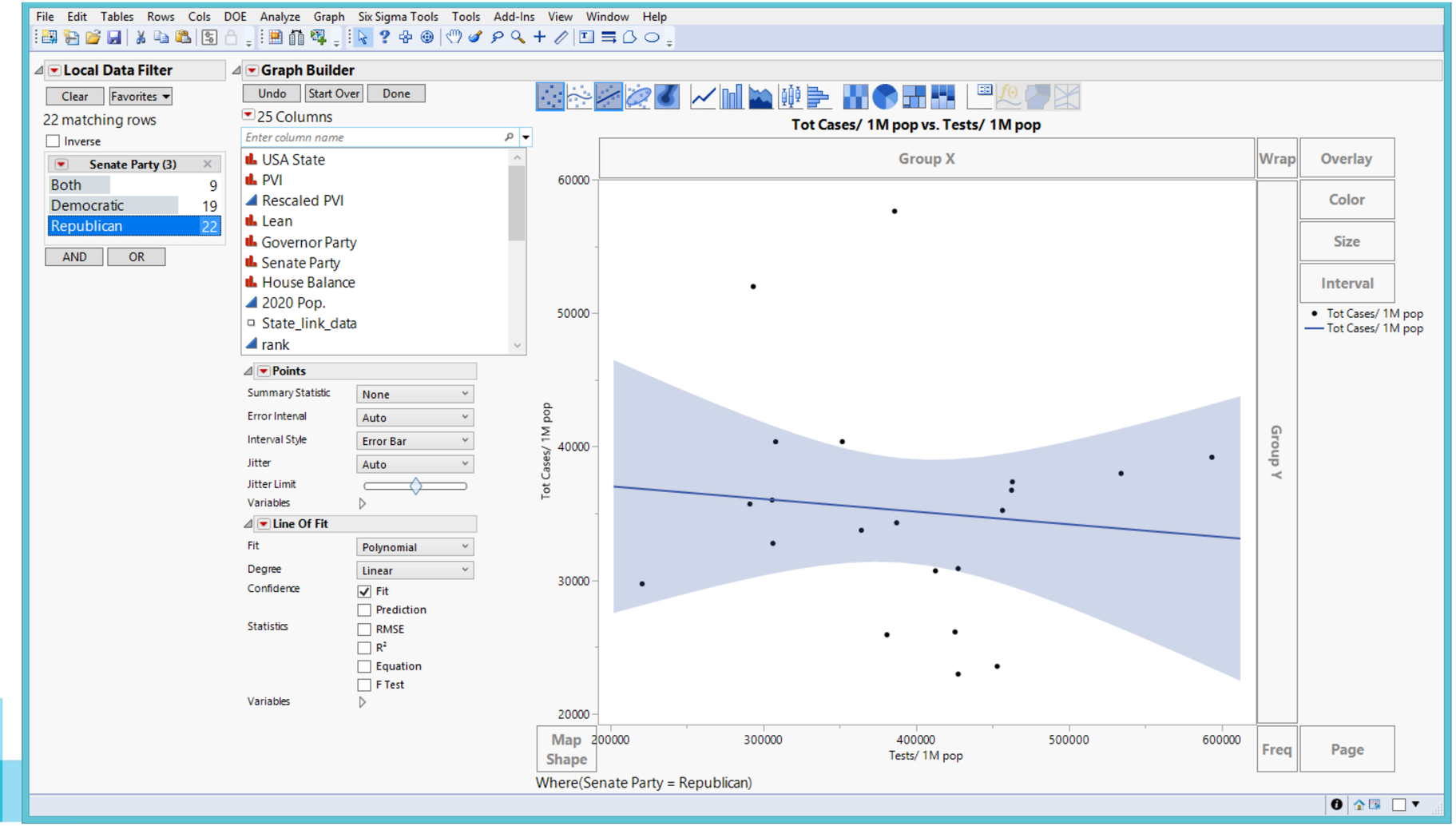
imdata.
Innovative Methods with Data Science & AI

Local filter



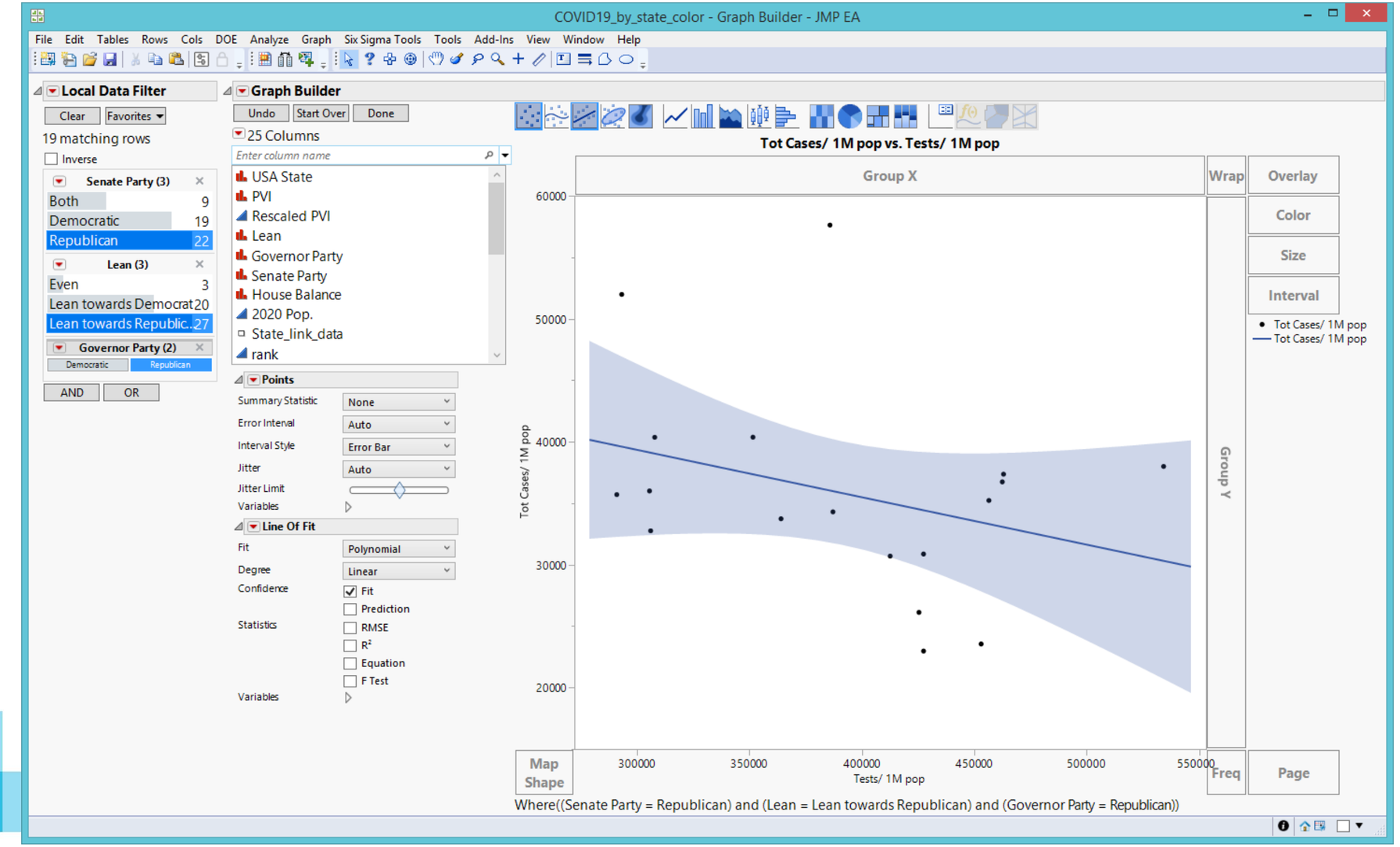
imdata.
Innovative Methods with Data Science & AI

Local filter



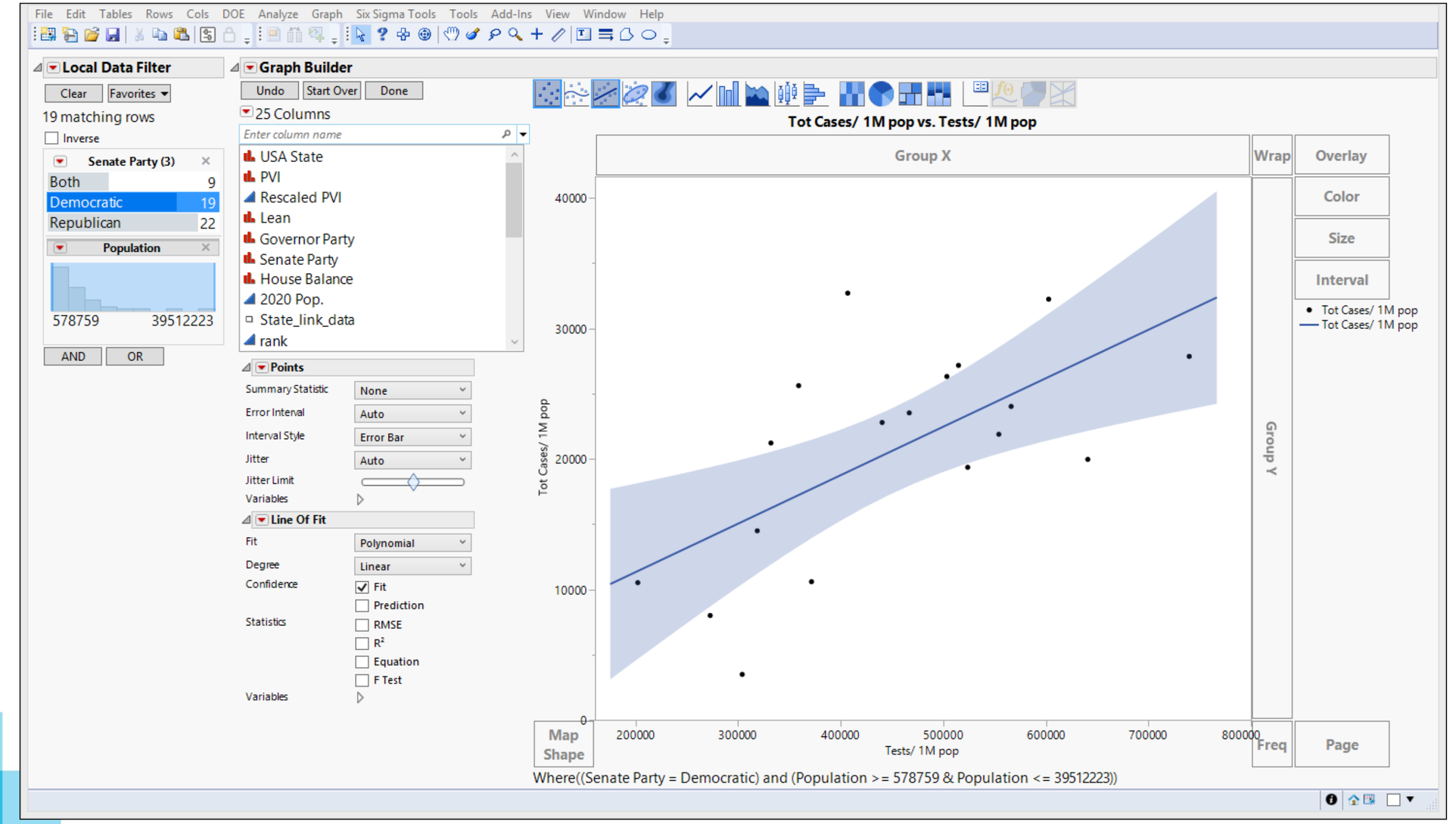
imdata.
Innovative Methods with Data Science & AI

Two or more filters



imdata.
Innovative Methods with Data Science & AI

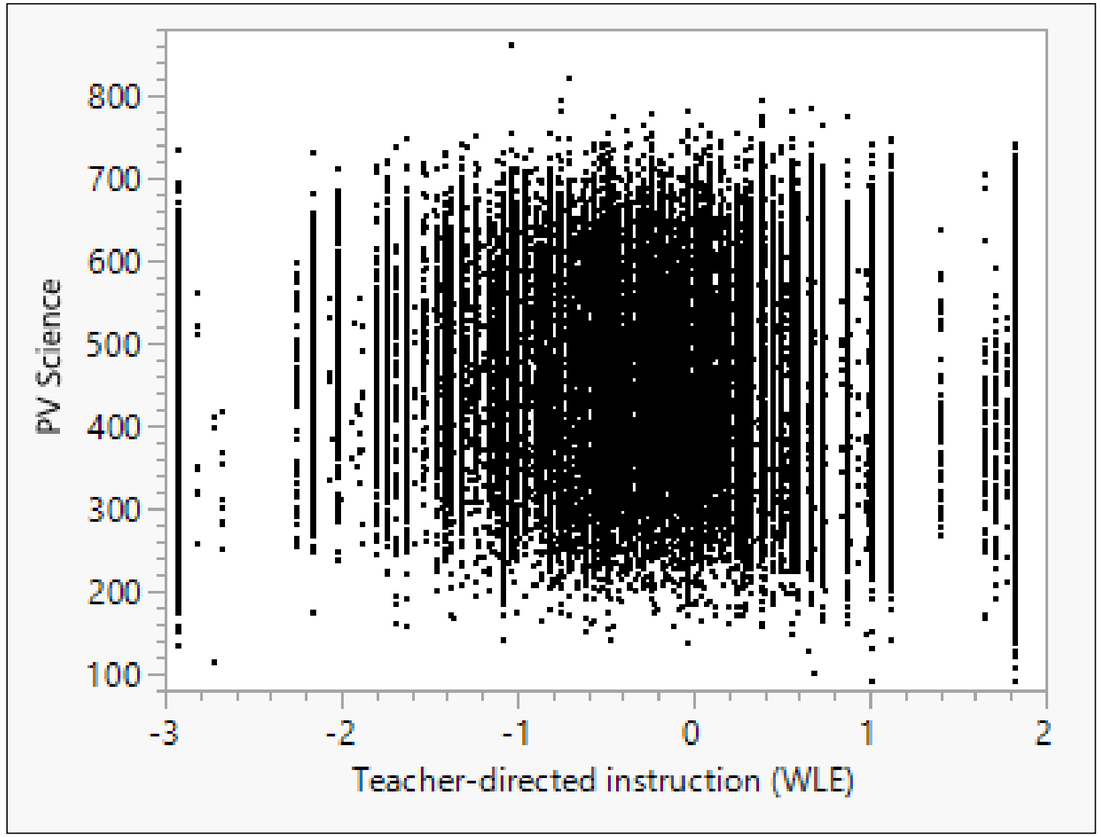
Two or more filters



imdata.
Innovative Methods with Data Science & AI

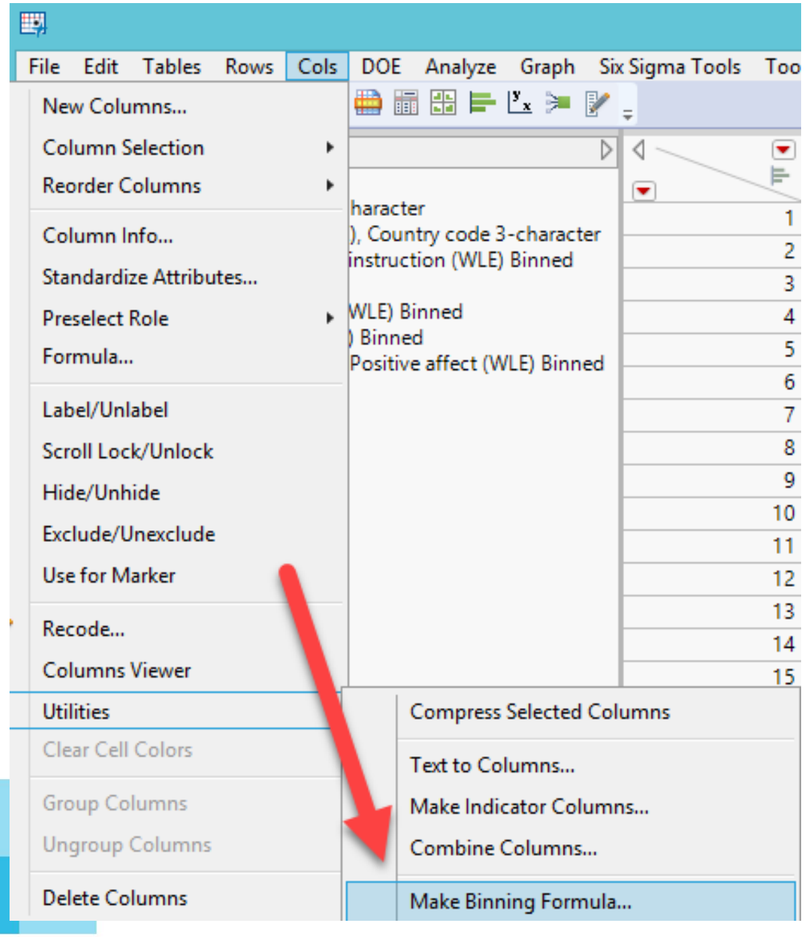
Less is more: Binning can be helpful!

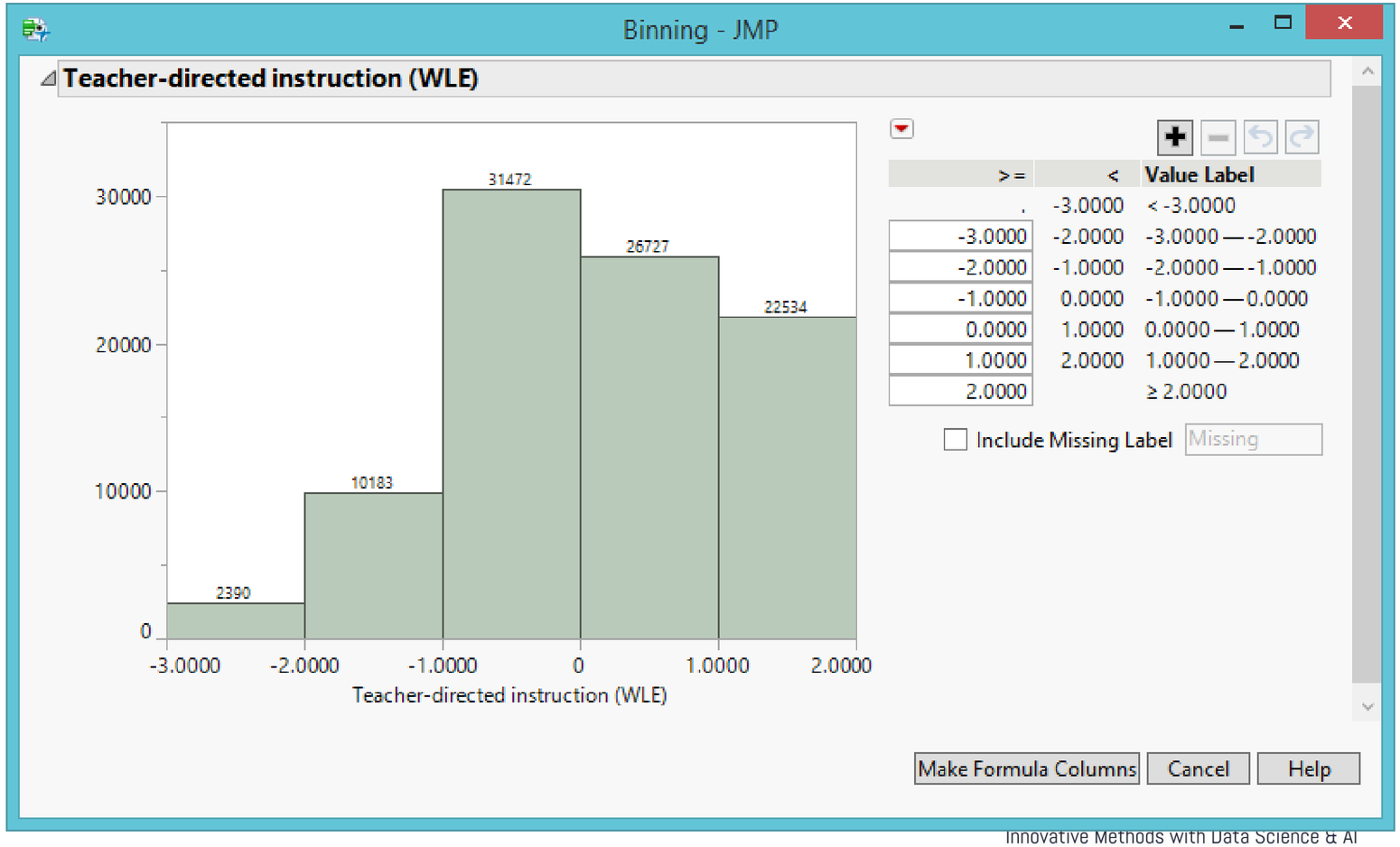
- Noise reduction
- 2018 PISA data
- WLE: Weighted Likelihood Estimates
- PV: Plausible values (test score)
- Too many data! Overplotting! It obscures us from seeing the relationship between science test performance and teacher-directed instruction.



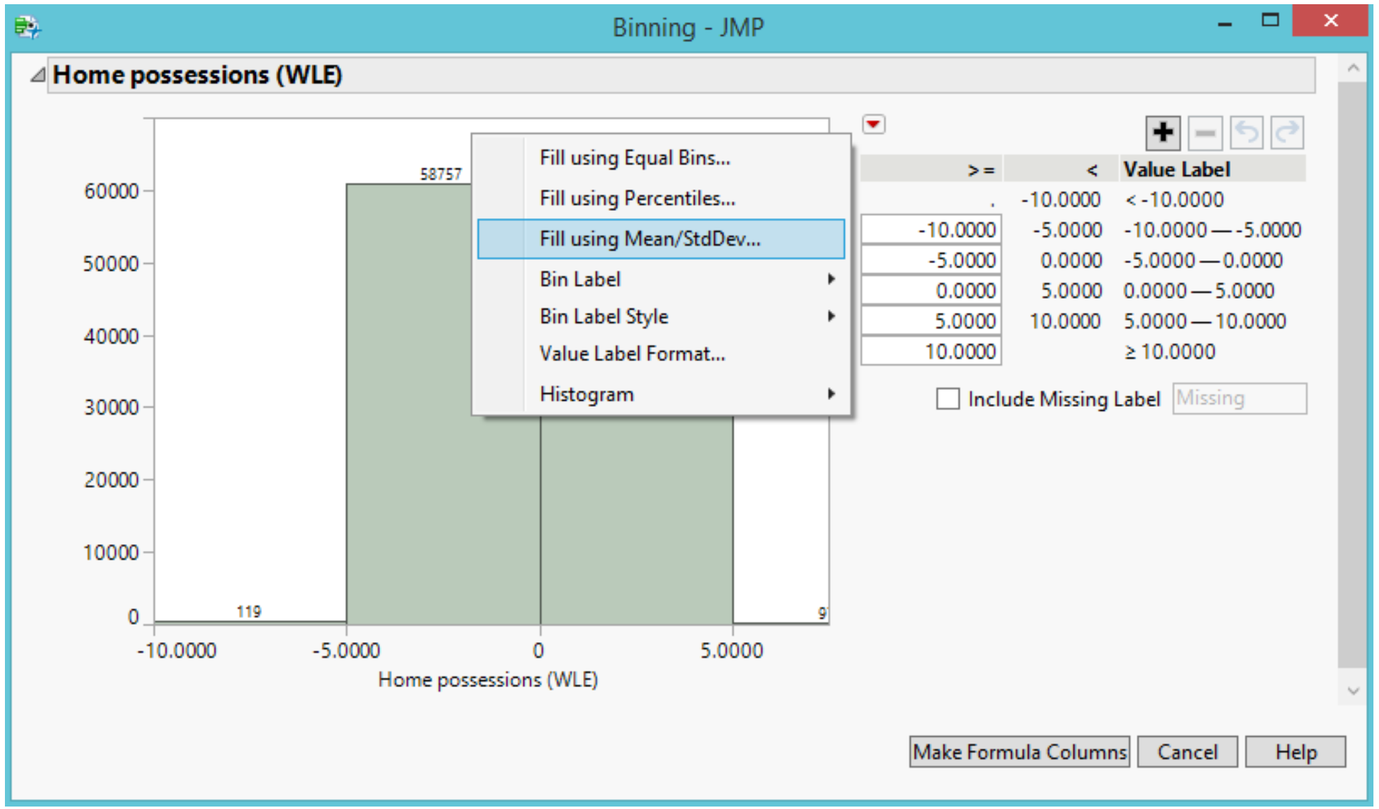
imdata.
Innovative Methods with Data Science & AI

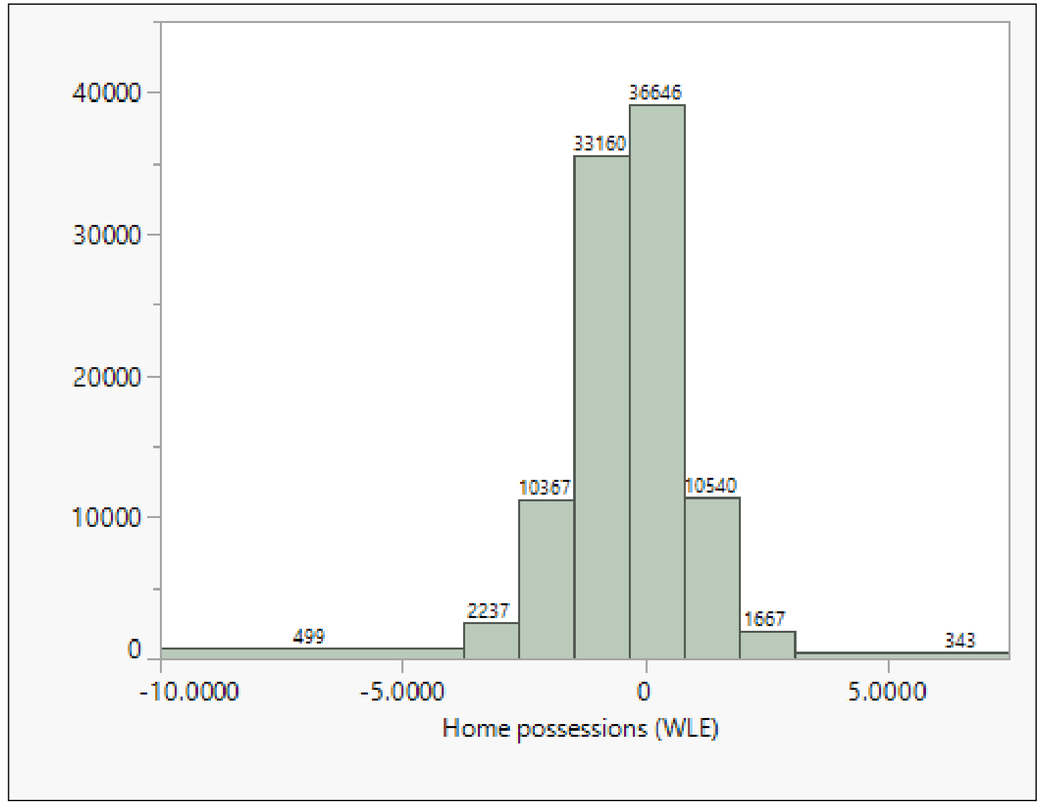
Select the variable and classify the data into different bins





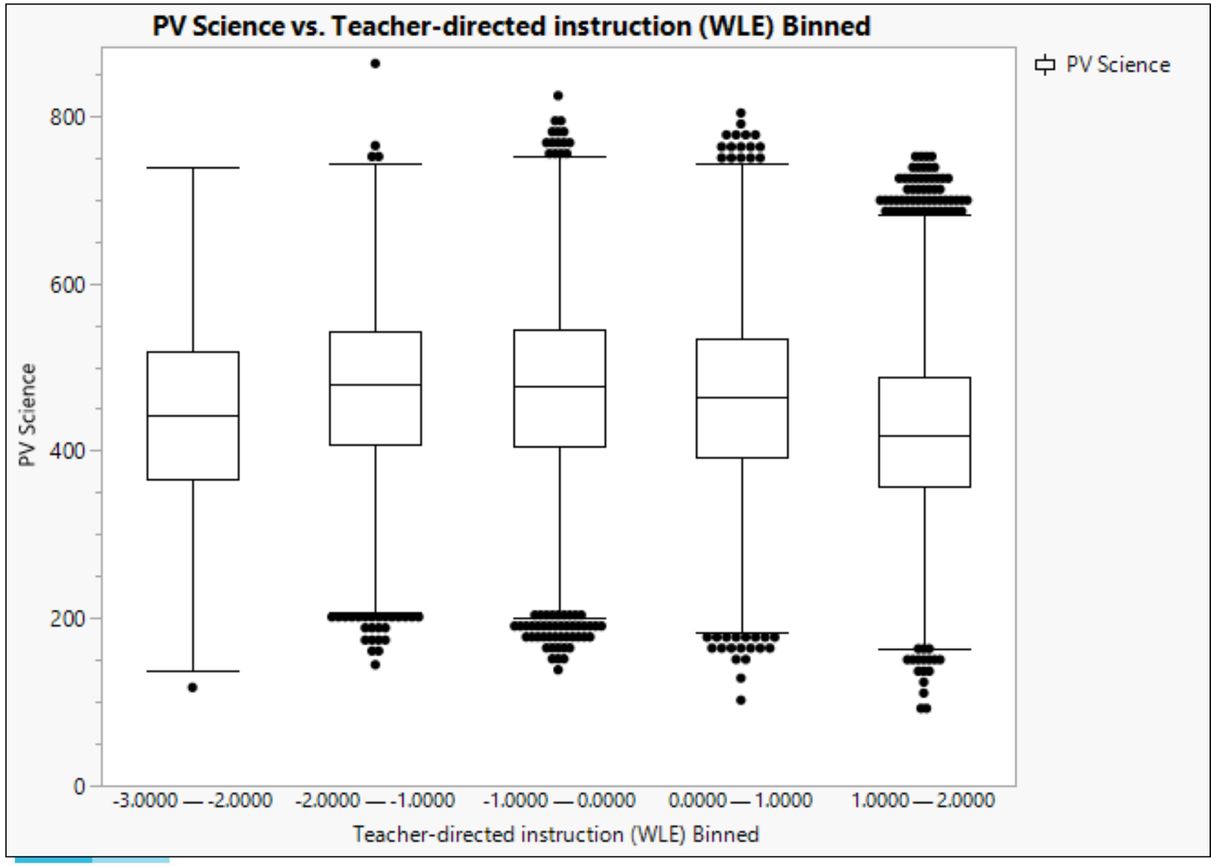
Select the variable and classify the data into different bins

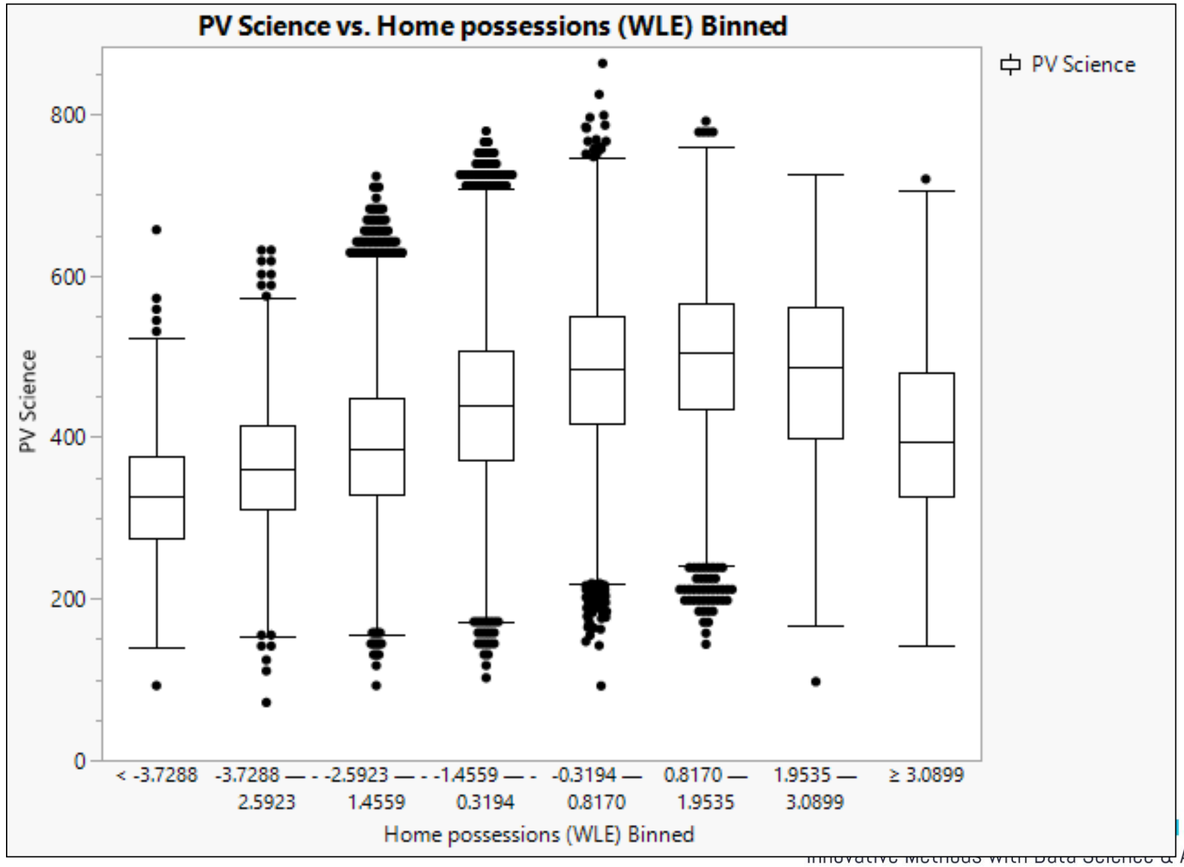




imdata.
Innovative Methods with Data Science & AI

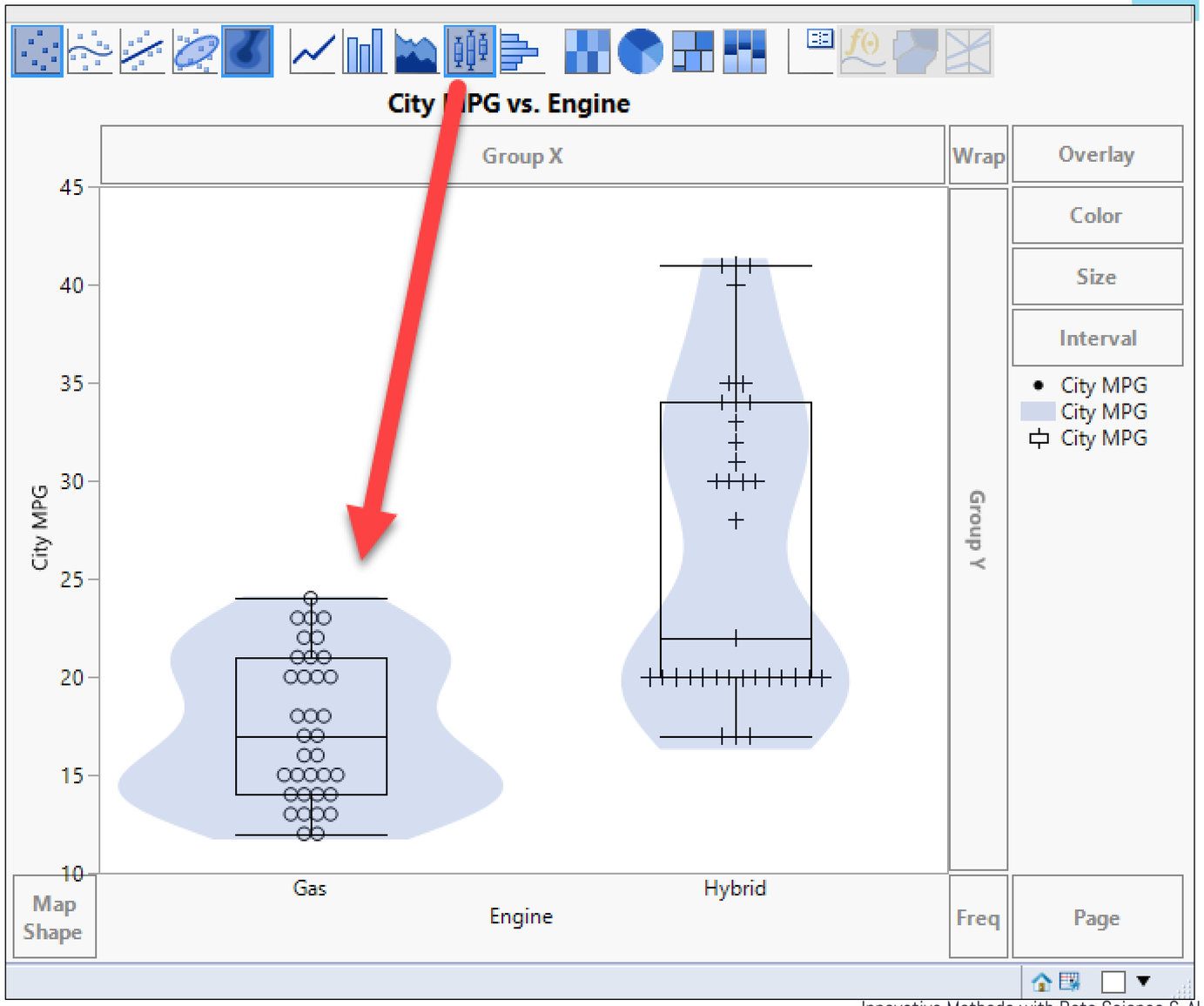
Binning and median smoothing: Non-linear pattern emerges!





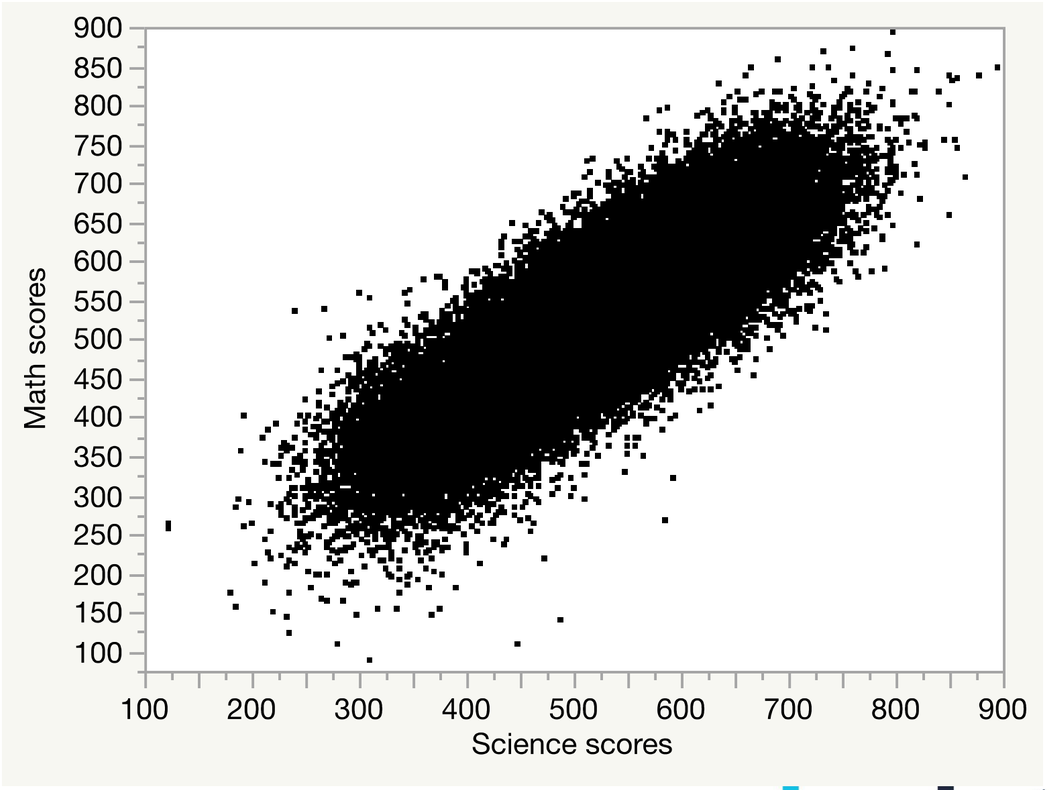
Violin plot and boxplot

- There are three displays in one: raw data, violin plot showing the distribution, and the boxplot showing the 5-point summary.



Revisit overplotting

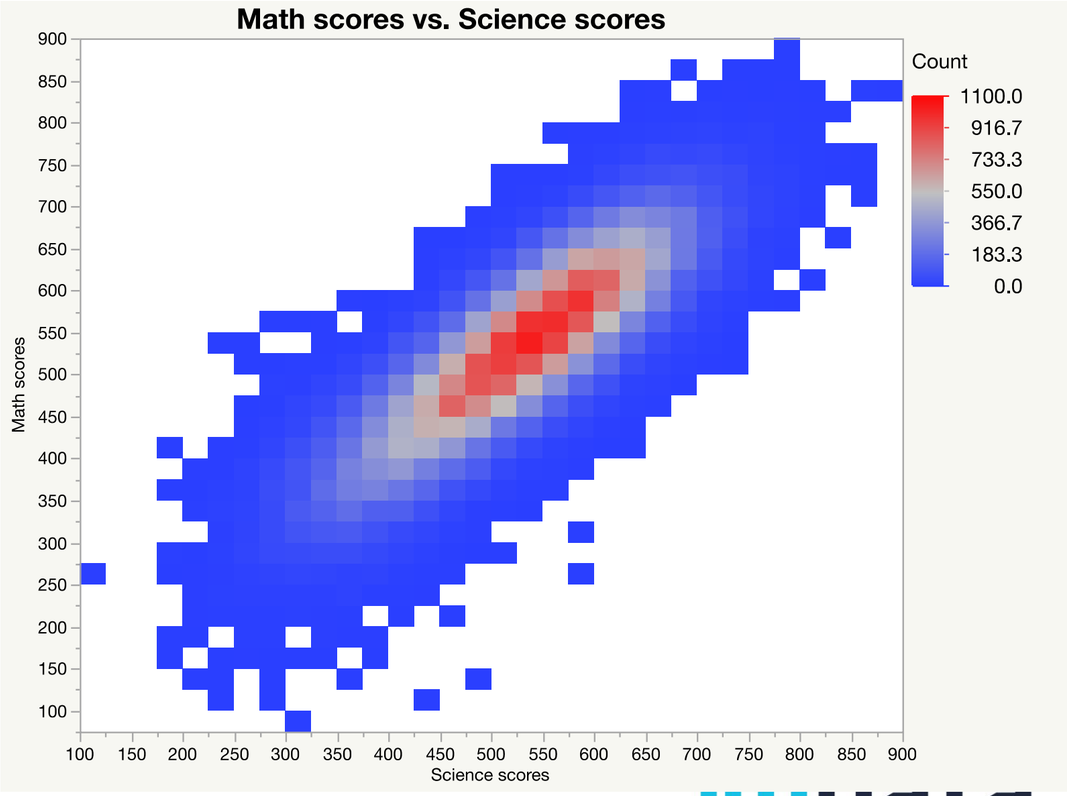
- 2015 PISA data
- *n* = 54,978
- A big cloud: Overplotting



imdata.
Innovative Methods with Data Science & AI

Overplotting

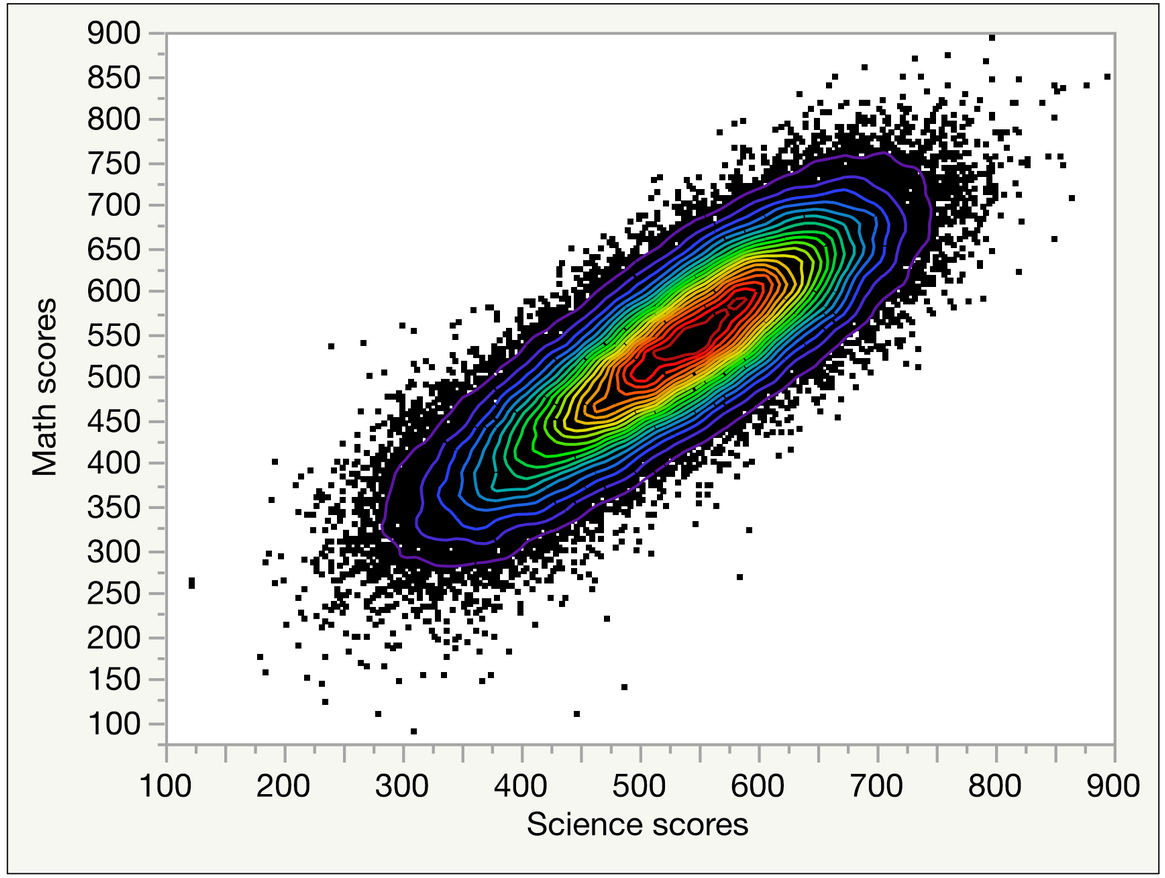
- One way to “see through” the cloud is using the heat map.
- Limitation:
  - Density is shown by colors only.
  - It hides the raw data.



Innovative Methods with Data Science & AI

Nonparametric Bivariate Density

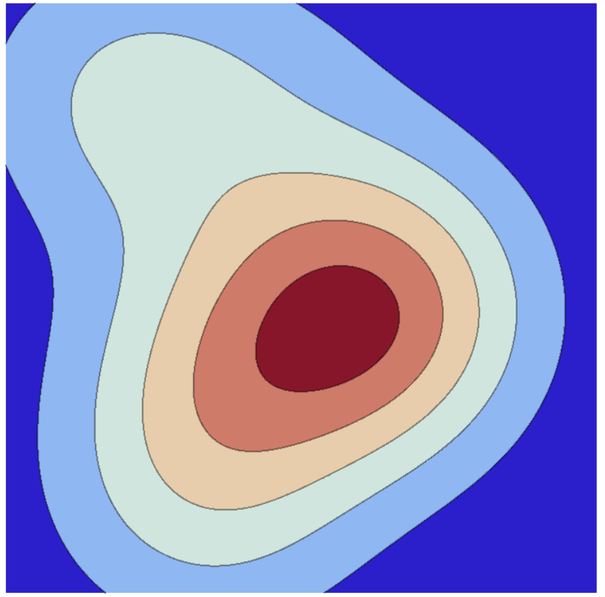
- The density of data points is represented by both colors and contour lines.
- Its interpretation is straight–forward.

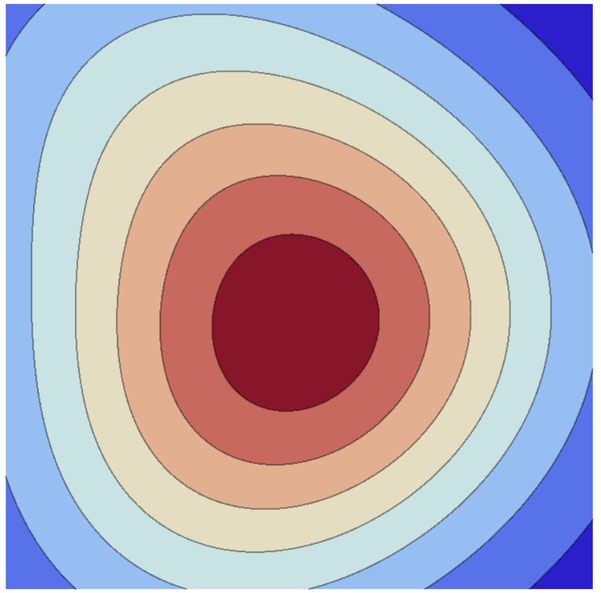


imdata.
Innovative Methods with Data Science & AI

Contour plot

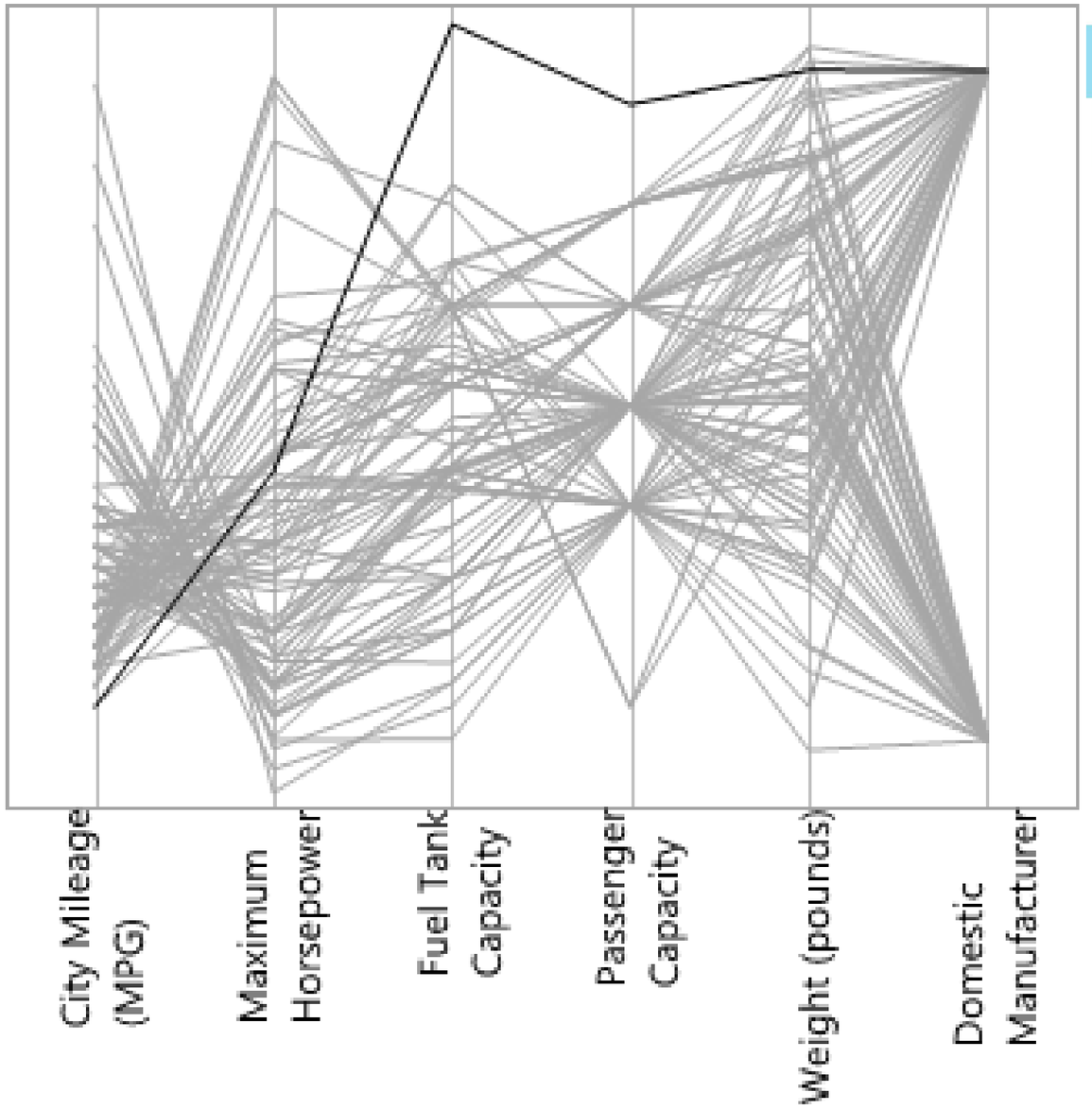
- How about using the contour lines without showing the raw data?
- It could be confusing and misleading.
- The appearance of a one-dimensional histogram is tied to its binwidth or bandwidth.
- By the same token, the bandheight of the isolines determines how a contour plot appears.
- The two contour plots are based on the same data and they use different bandheights!





Can Parallel coordinate overcome the curse of dimensionality?

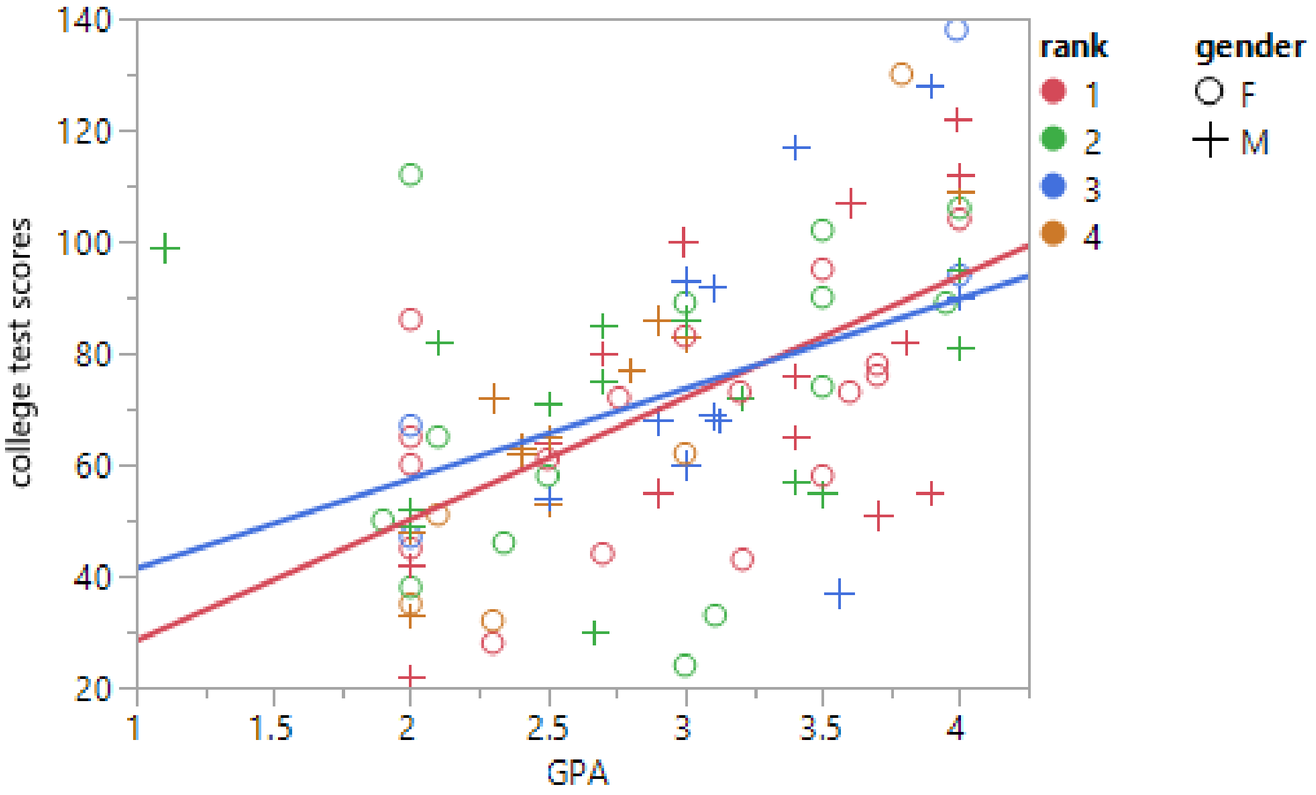
- By joining individual observations in parallel coordinate plot, you can look at multiple dimension.
- By you are facing the problem of over-plotting!



imdata.
Innovative Methods with Data Science & AI

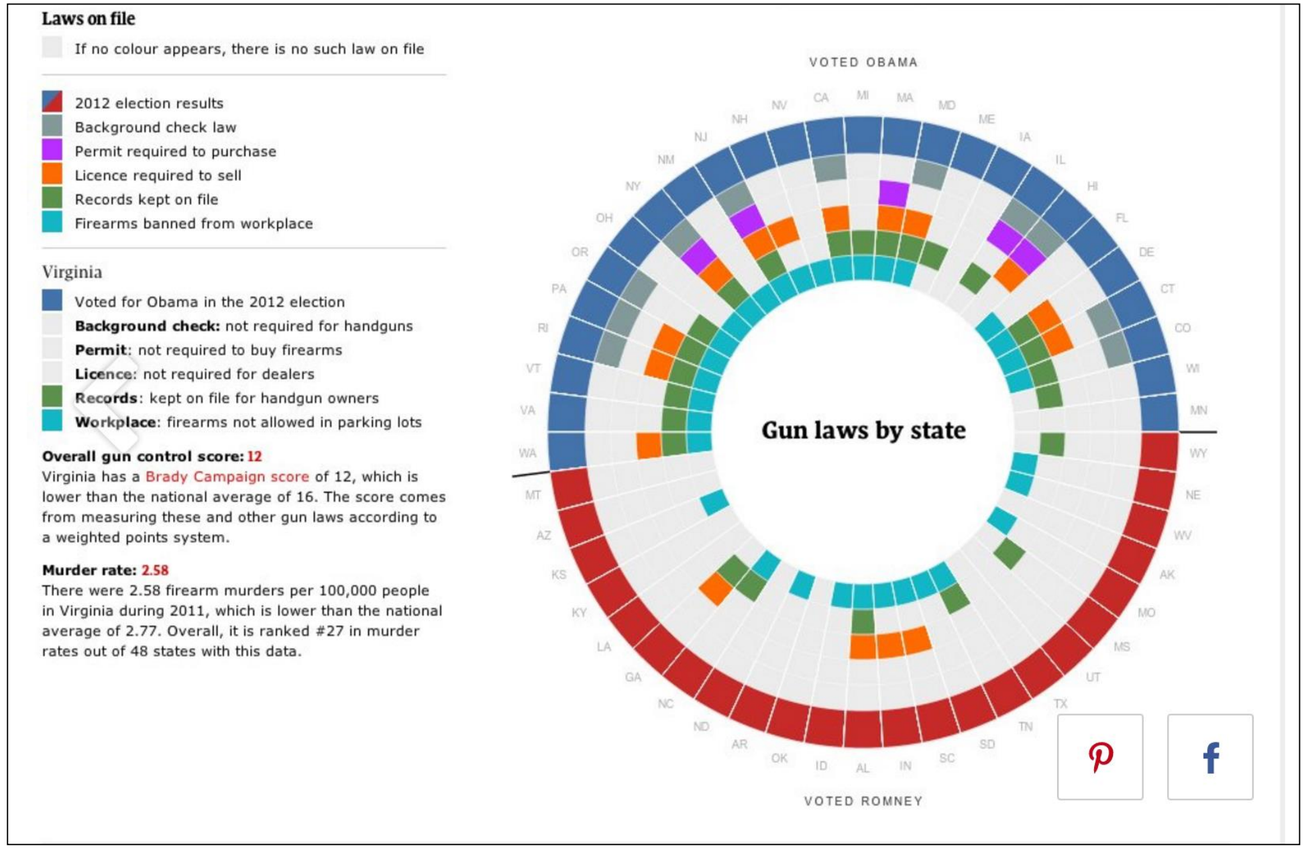
Curse of dimensionality

- Using colors, shape to represent different dimensions.
- Can be confusing.



imdata.
Innovative Methods with Data Science & AI

Extreme examples



imdata.
Innovative Methods with Data Science & AI

Conclusion

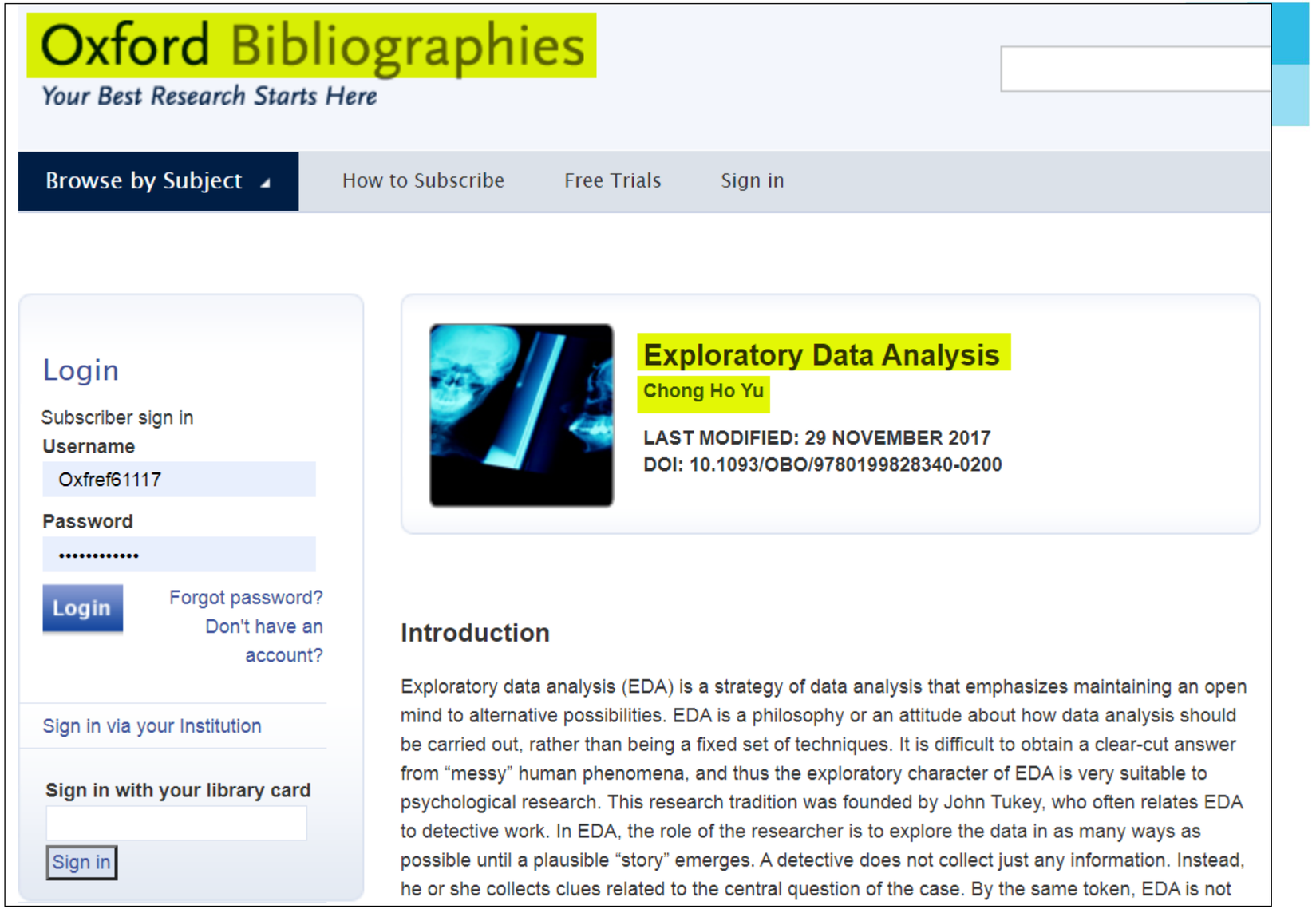
- Exploratory data analysis utilizing data visualization should work hand in hand with AI and machine learning.
- Data exploration requires dynamic, not static graphic.
- Overplotting can be overcome by using boxplot (median smoothing), heatmap, non-parametric contour…etc.
- Curse of dimensionality can be overcome by switching variables, local filters…etc.
- Too much data reduction is problematic (e.g. contour)
- No data reduction is also problematic (e.g. parallel coordinate plot)

imdata.
Innovative Methods with Data Science & AI

For more info

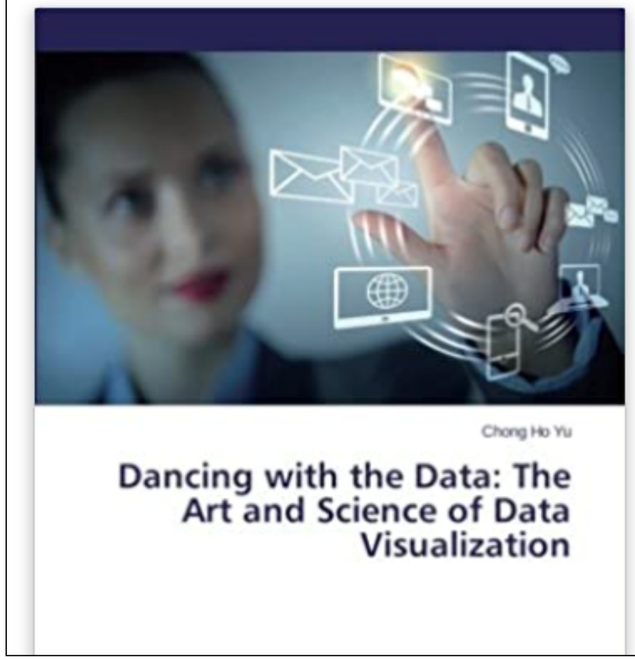
- Google search for

“Oxford,” “Exploratory data analysis” or/and my name.



imdata.
Innovative Methods with Data Science & AI

For more info



Dancing with the Data: The Art and Science of Data Visualization Paperback – June 19, 2014

by Chong Ho Yu (Author)

› See all formats and editions

Paperback
$88.00 ✓prime

1 Used from $86.78
11 New from $75.67

imdata.
Innovative Methods with Data Science & AI

Contact Info

- Chong Ho (Alex) Yu
- chonghoyu@gmail.com
- cyu@apu.edu
- www.creative-wisdom.com/pub/pub.html

imdata.
Innovative Methods with Data Science & AI