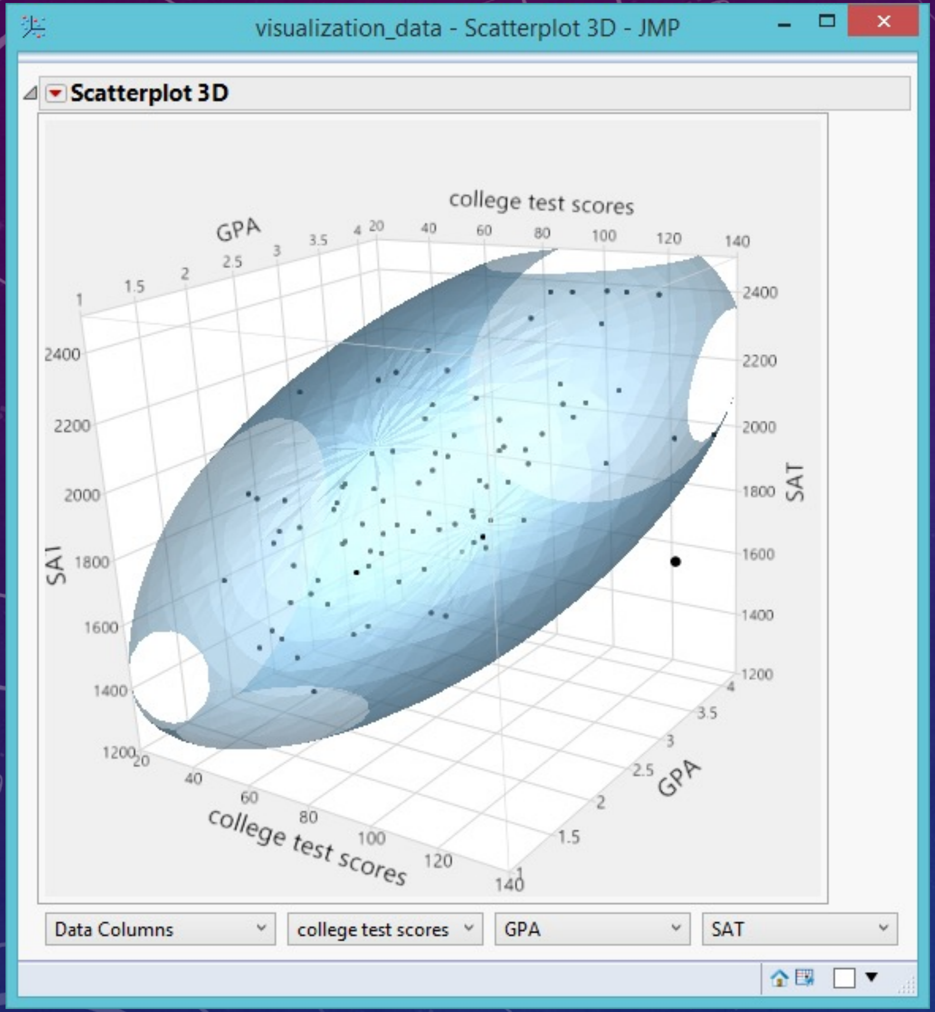


visualization_data - Scatterplot 3D - JMP

Scatterplot 3D

Data Columns | college test scores | GPA | SAT

DYNAMIC DATA VISUALIZATION FOR PATTERN SEEKING AND INSIGHT DISCOVERY

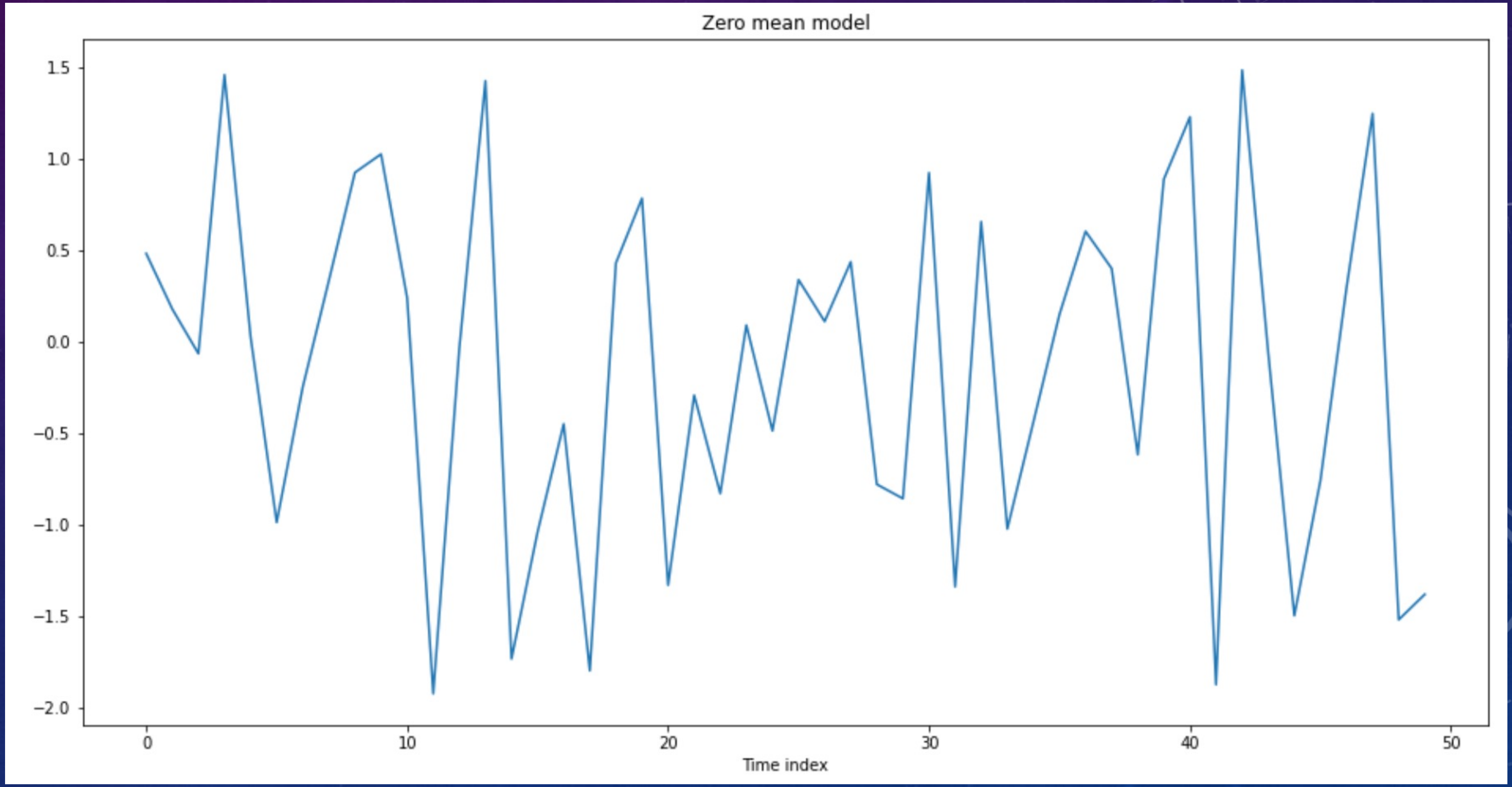
CHONG HO YU, PH.D., D. PHIL

2022 IDEAS GLOBAL A.I. CONFERENCE

LOS ANGELES, CA

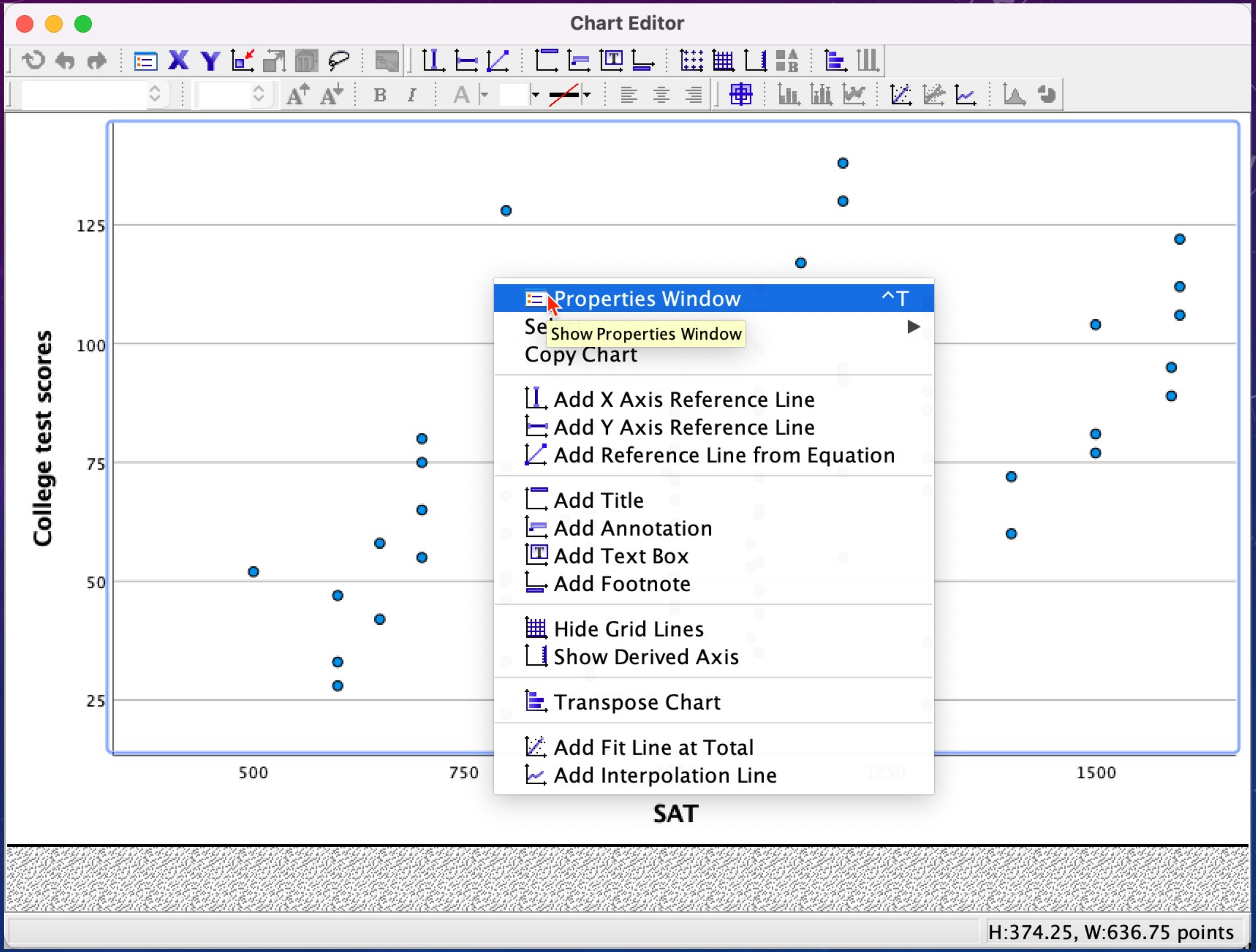
STATIC SYSTEM

- The output is frozen.
- The analyst cannot alter the results by adding new components to or removing existing elements from the output panel.
- What you see is what you get!



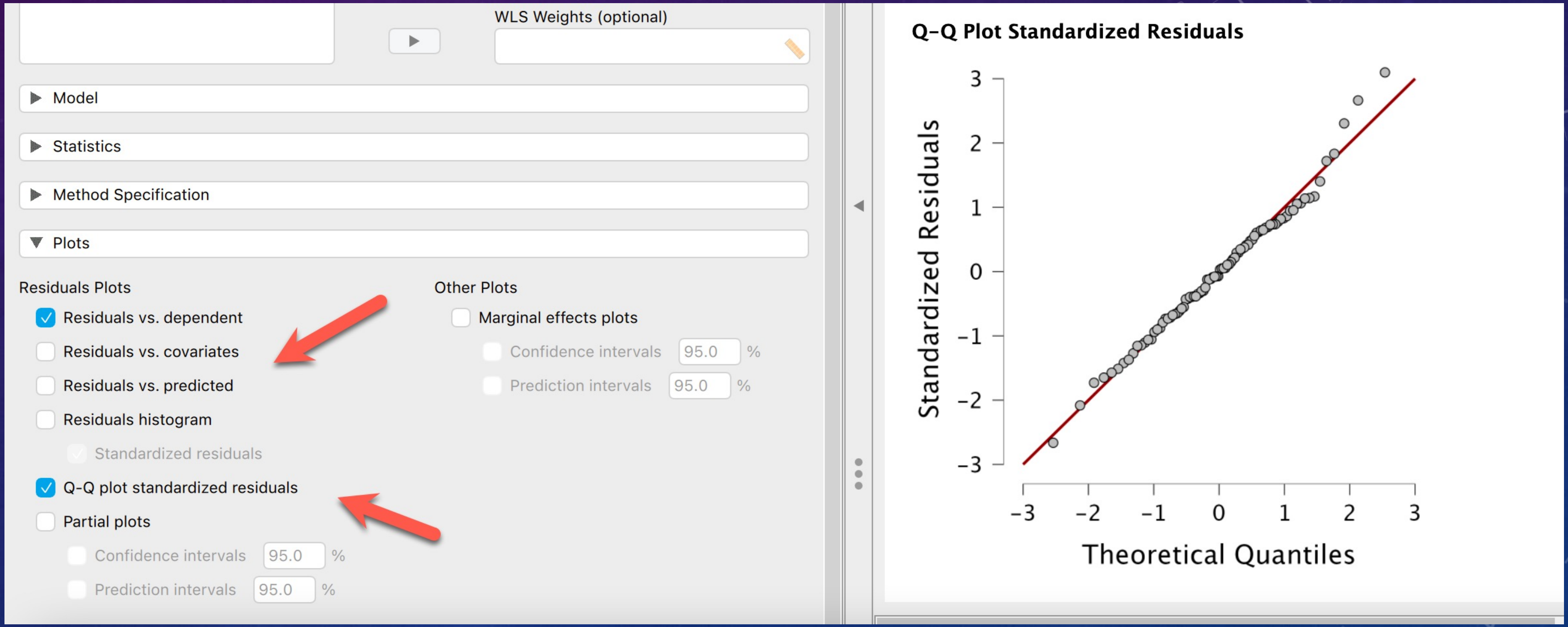
STATIC SYSTEM

- Some systems allow the user to change the appearance of the display only, such as changing the color of dots or adding a title, but this trivial change does not lead to insightful discovery.



SEMI-DYNAMIC SYSTEM

- A semi-dynamic system allows one-way manipulation only.
- Using this set up the analyst can change the options in the input panel to alter the output, but not the other way around.



WLS Weights (optional)

▶ Model

▶ Statistics

▶ Method Specification

▼ Plots

Residuals Plots

☑ Residuals vs. dependent

□ Residuals vs. covariates

□ Residuals vs. predicted

□ Residuals histogram

☑ Standardized residuals

☑ Q-Q plot standardized residuals

□ Partial plots

□ Confidence intervals 95.0 %

□ Prediction intervals 95.0 %

Other Plots

□ Marginal effects plots

□ Confidence intervals 95.0 %

□ Prediction intervals 95.0 %

FULLY DYNAMIC VISUALIZATION

In a fully dynamic system all objects, including the data table and the output panels, are inter-linked. In this configuration, manipulating any variable or observation in any object can trigger corresponding changes in all other objects.

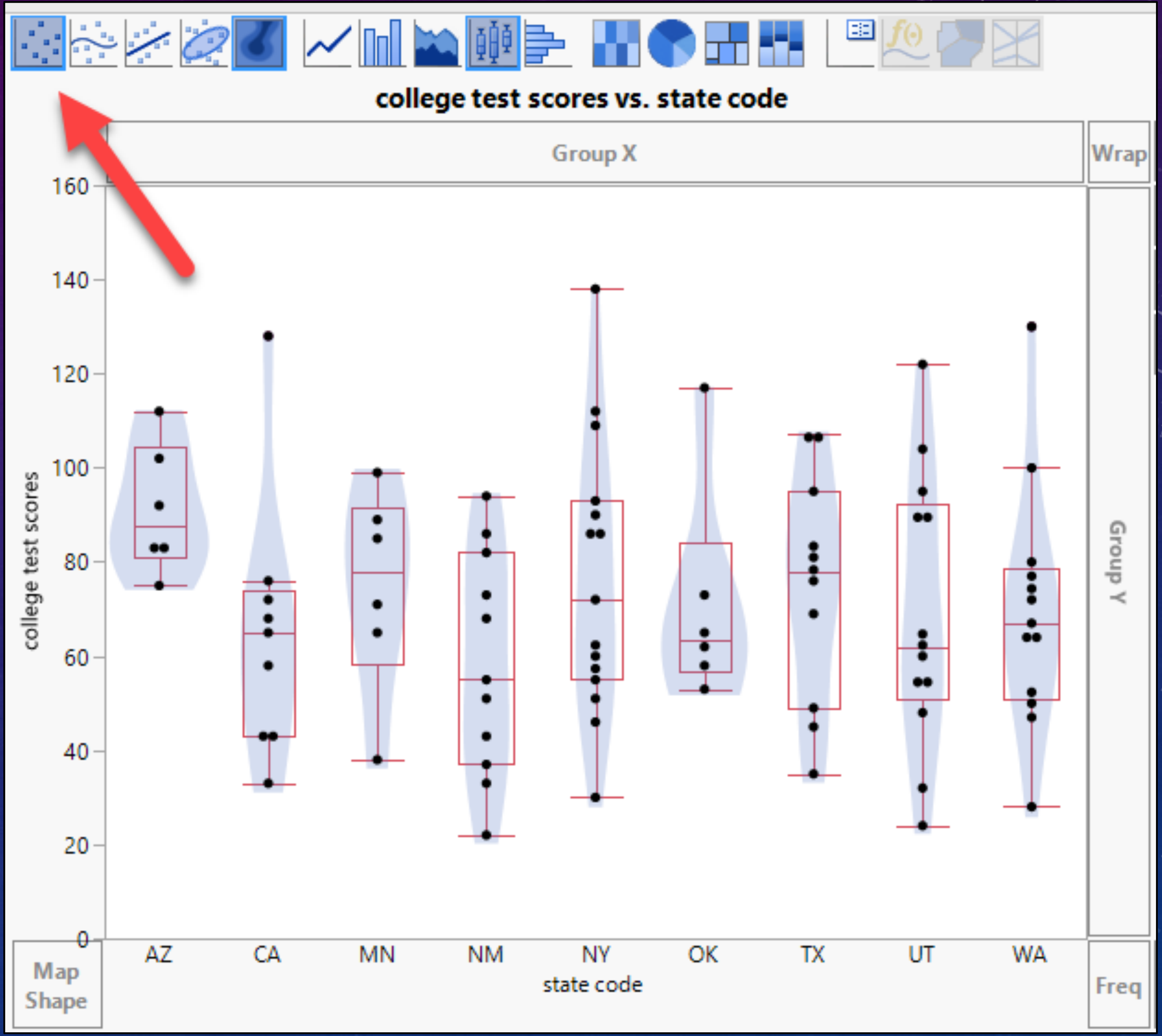
Examples: SAS Viya/Visual Analytics, JMP Pro, Tableau (Salesforce), RGL in R.

WHY DO WE NEED DYNAMIC VISUALIZATION?

- For seeking the pattern of the data, data mining and visualization should be an iterative and interactive process, not a single shot.
- A fully dynamic or semi-dynamic software system allows the analyst to explore the data by asking “what-if” questions.
- For instance, “how will the regression slope change if I remove one, three, or more outliers based on 99%, 95%, or 90% density ellipses?” “Will the relationship between X and Y remains the same if a third variable is added to the model?”

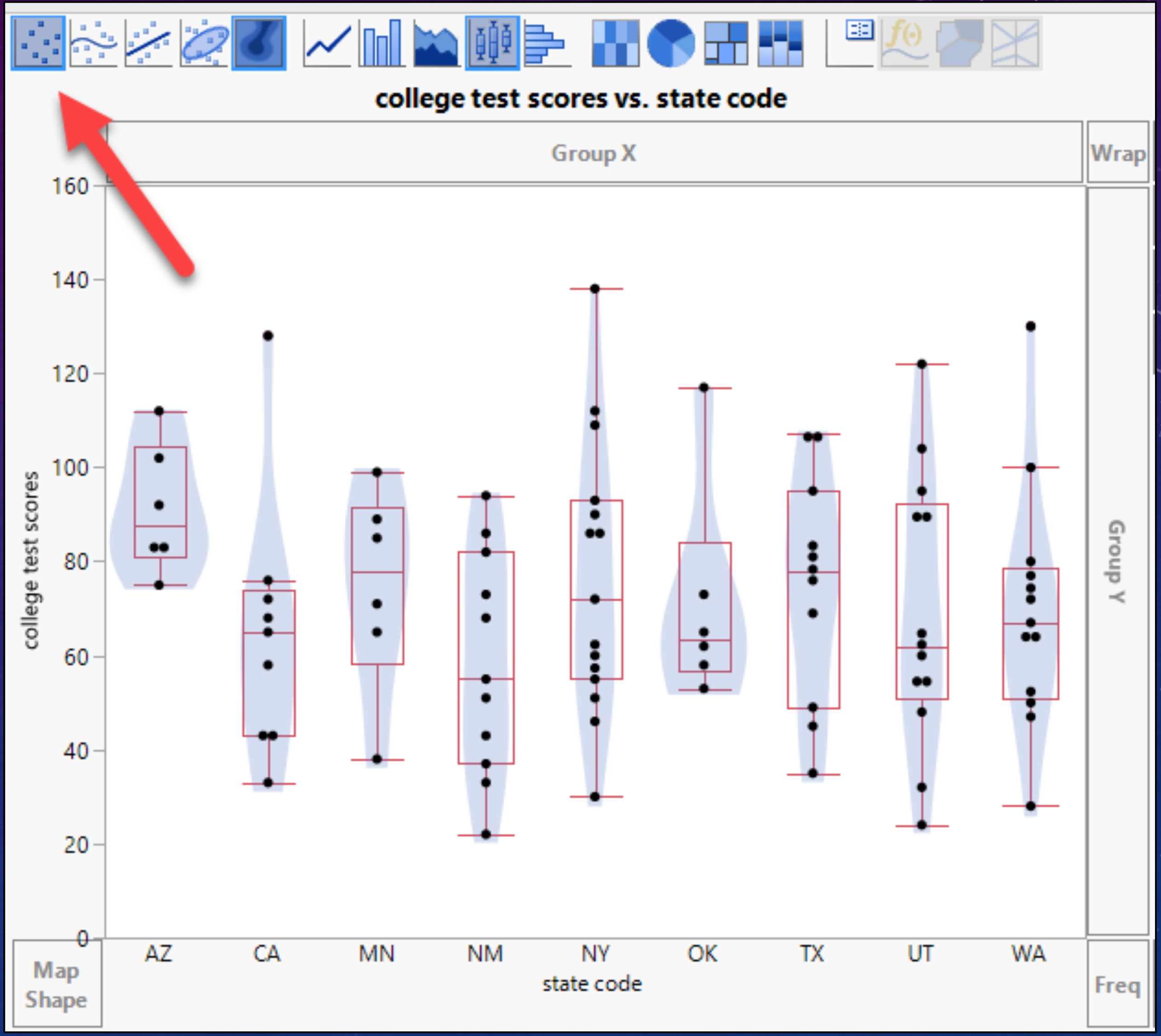
SUPERIMPOSITION

- Explore the data by changing the display options.
- You can even superimpose multiple chart types into one display by dragging the graph icon into the canvas.



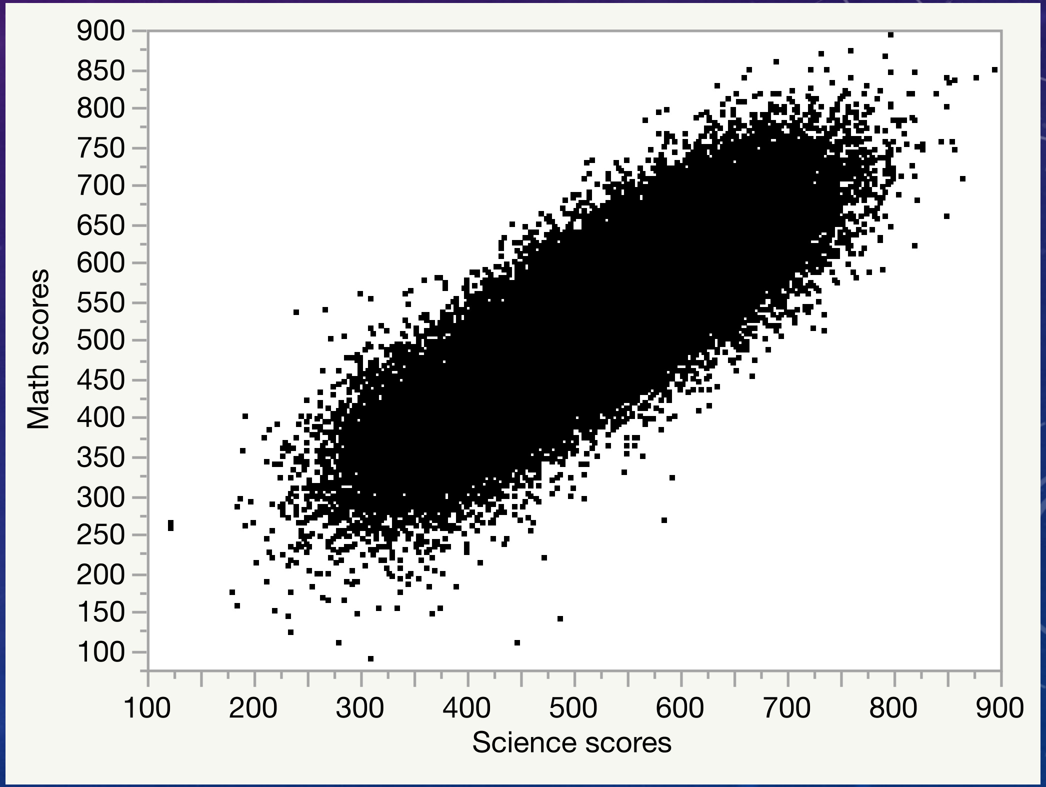
SUPERIMPOSITION

- Holistic view.
- The boxplots show the quantile info.
- The violin plots show the distribution.
- The dot plots expose outliers.



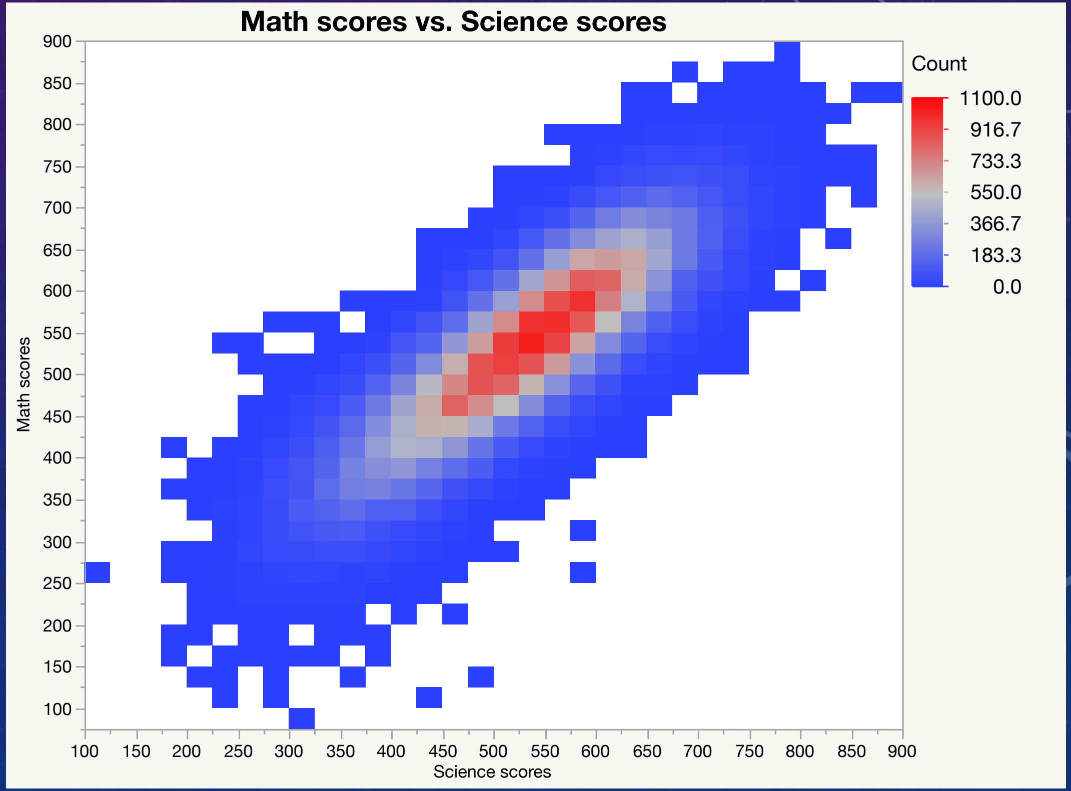
OVERPLOTTING

- Use PISA2015
- *n* = 54,978
- A big cloud: Overplotting



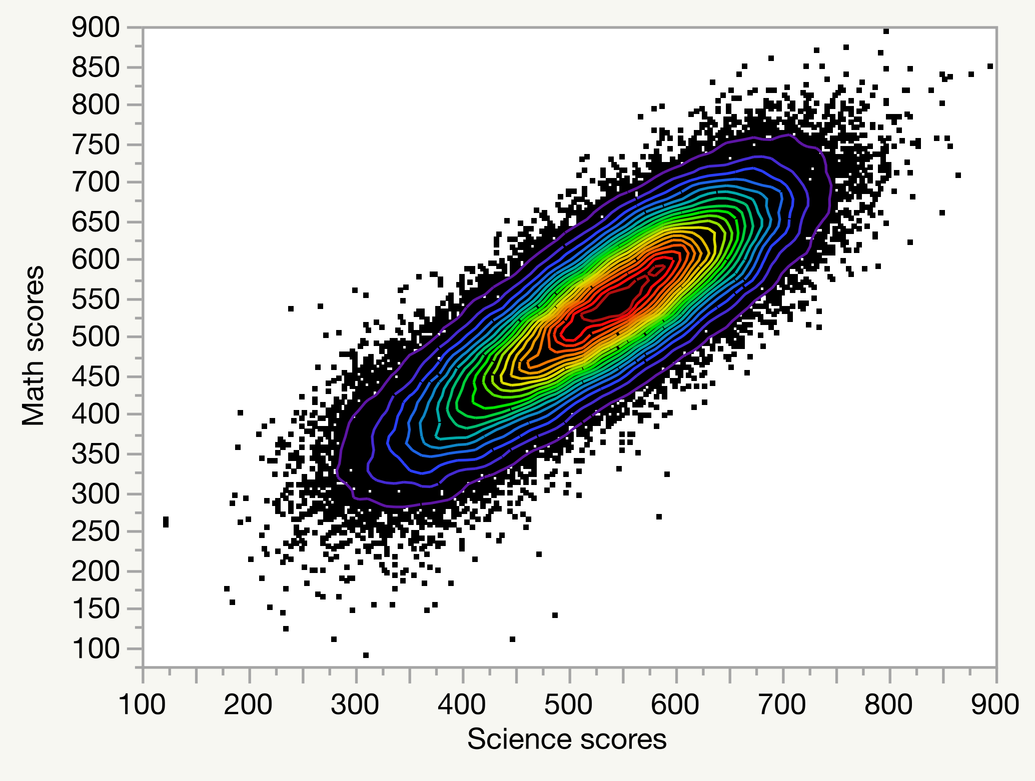
OVERPLOTTING

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



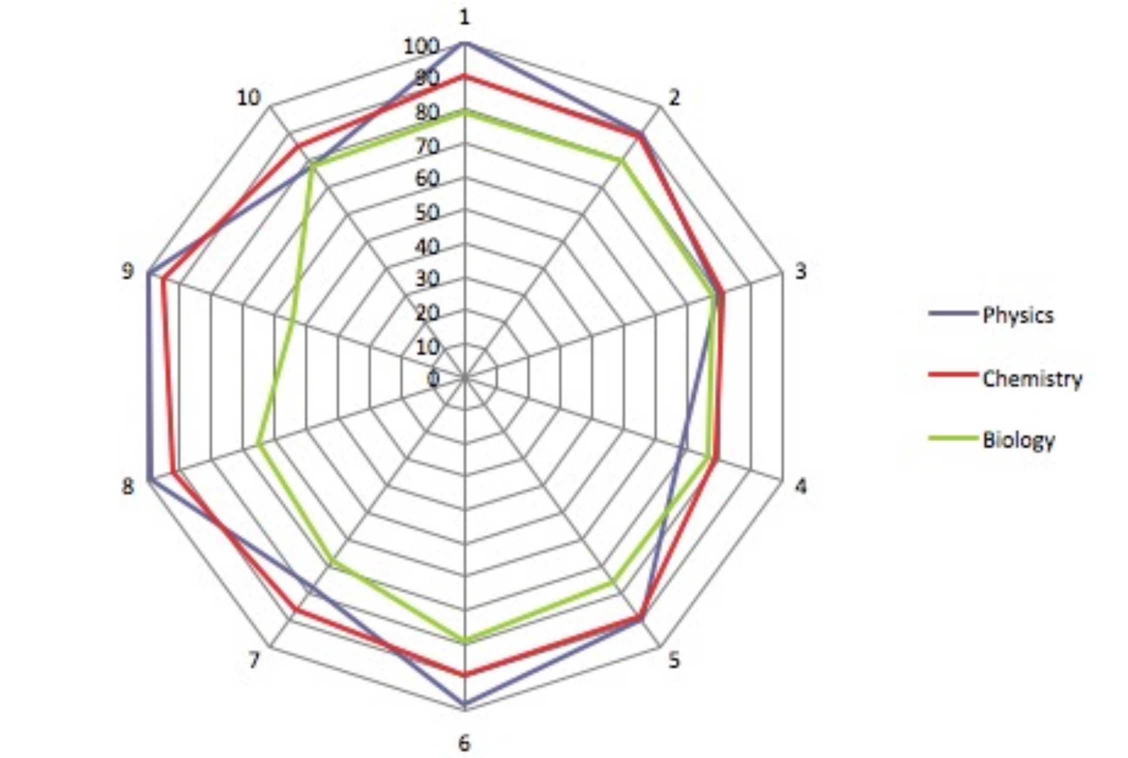
NONPARAMETRIC BIVARIATE DENSITY

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



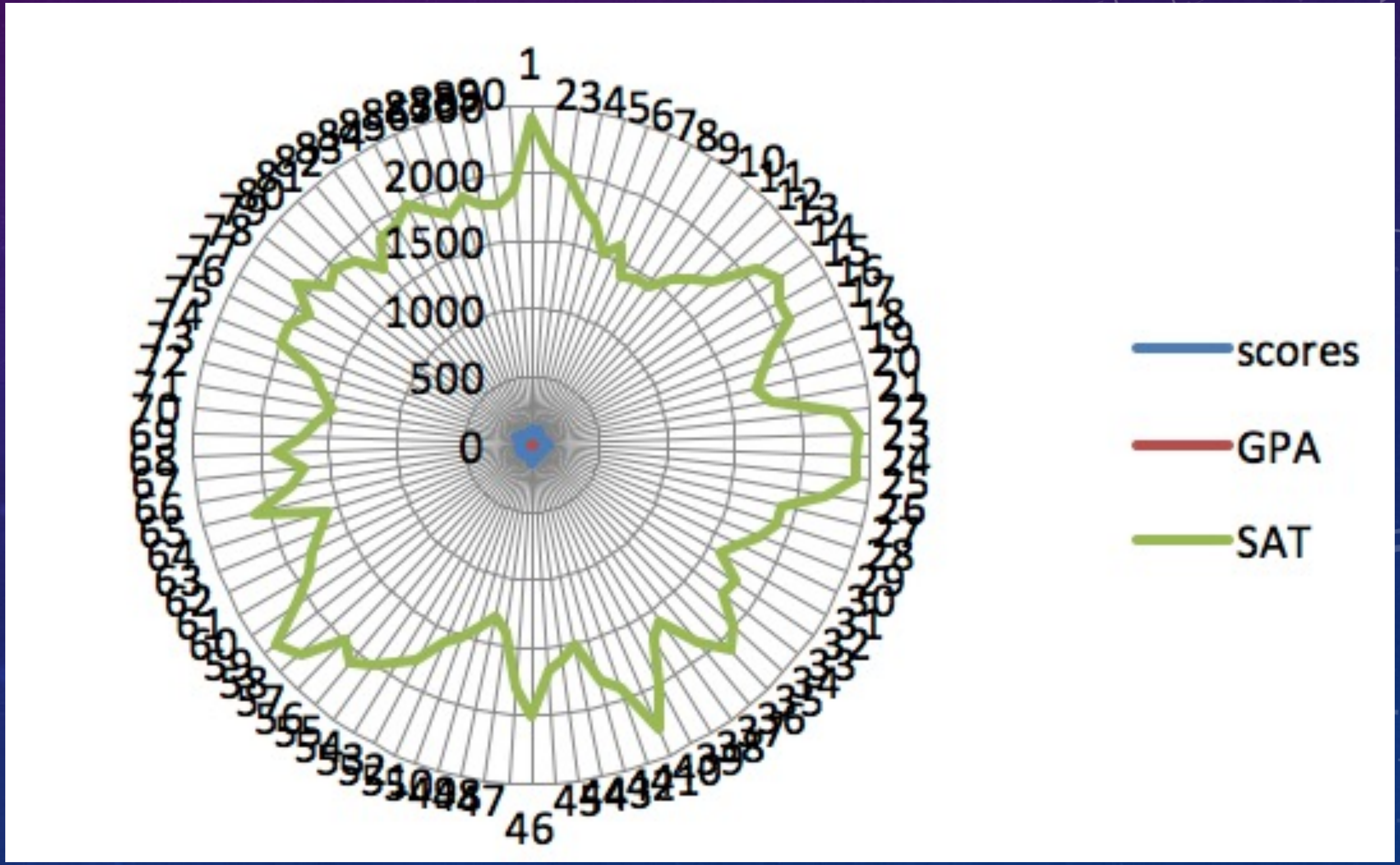
RADAR GRAPH

- When you have a small data set or summary data (no over-plotting), you can use Excel (e.g. Radar graph).
- The biology scores are the lowest and are not correlated to either physics or chemistry scores, but physics and chemistry scores are fairly good predictors of each other.
- This visualization approach is applicable to chemistry, toxicology, market research, and health care research.



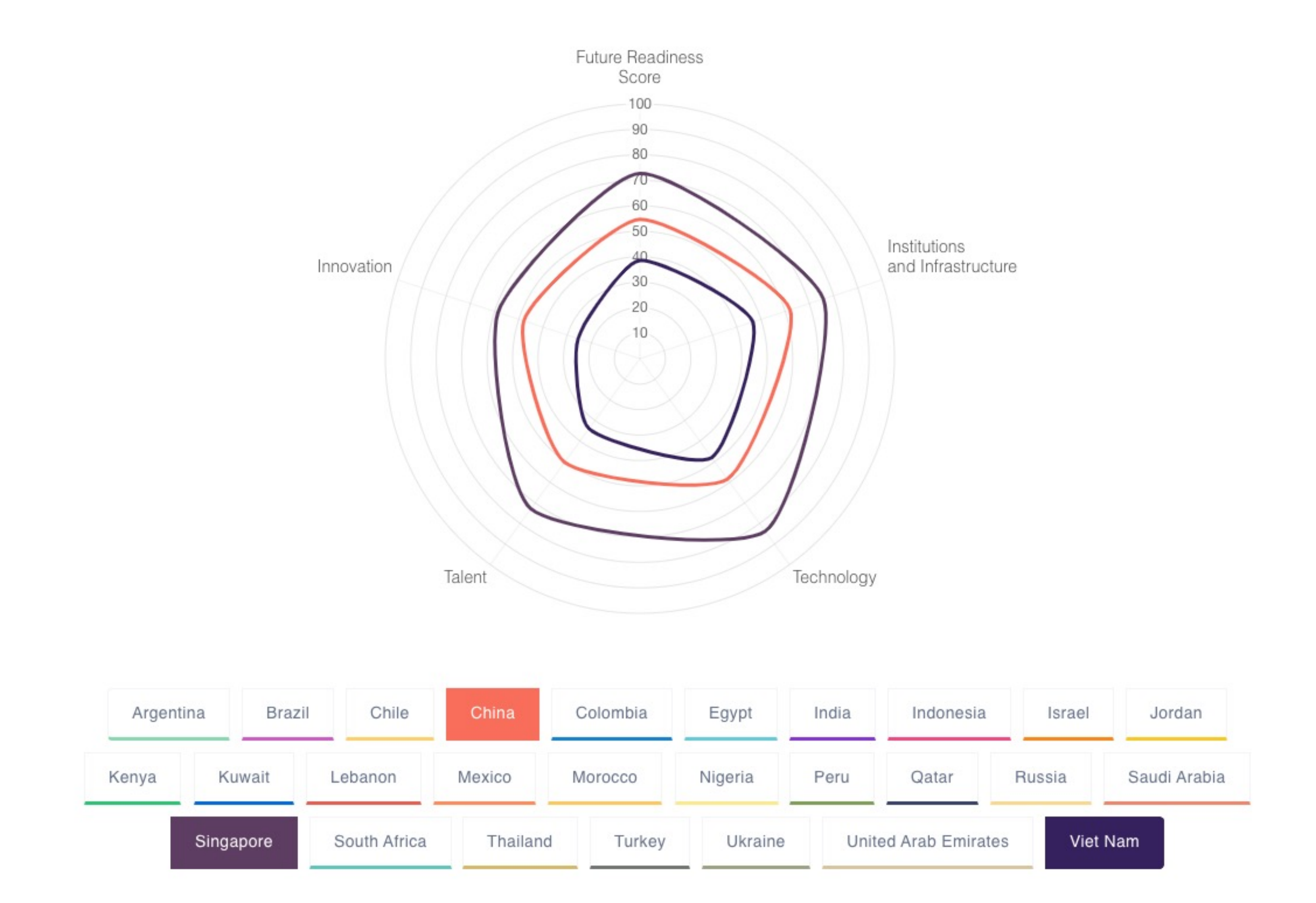
RADAR GRAPH: LIMITATION

- When there are too many variables, levels, and observations, the data pattern will be concealed.
- When the measurement scales are vastly different, it is very difficult, to display a meaningful result using a radar plot.



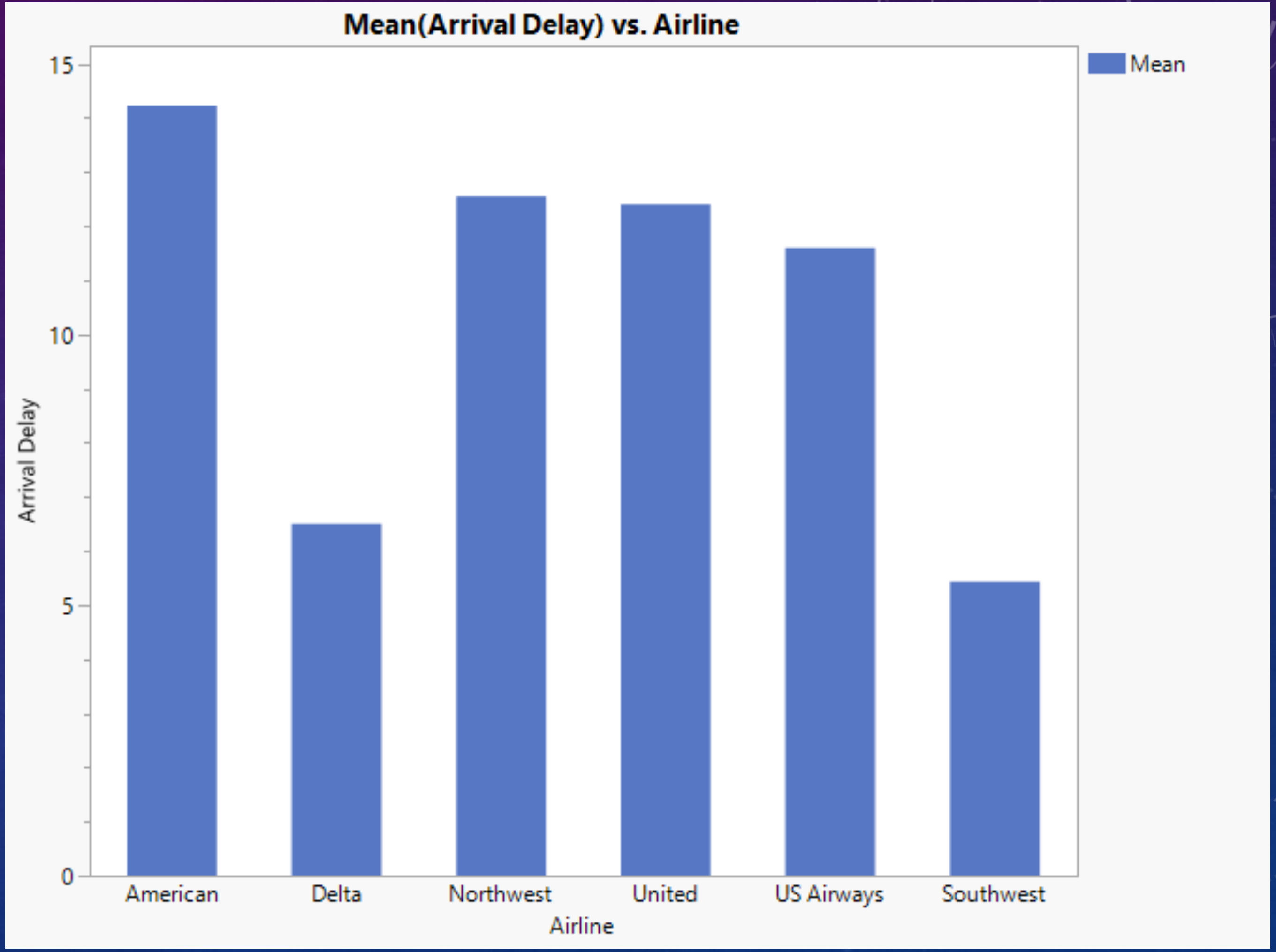
RADAR GRAPH

- Google uses it for comparing Future Readiness Scores across countries.
- When there are many countries (observations), it necessitates a dynamic system.
- https://www.portulansinstitutefrei.com/2021/report



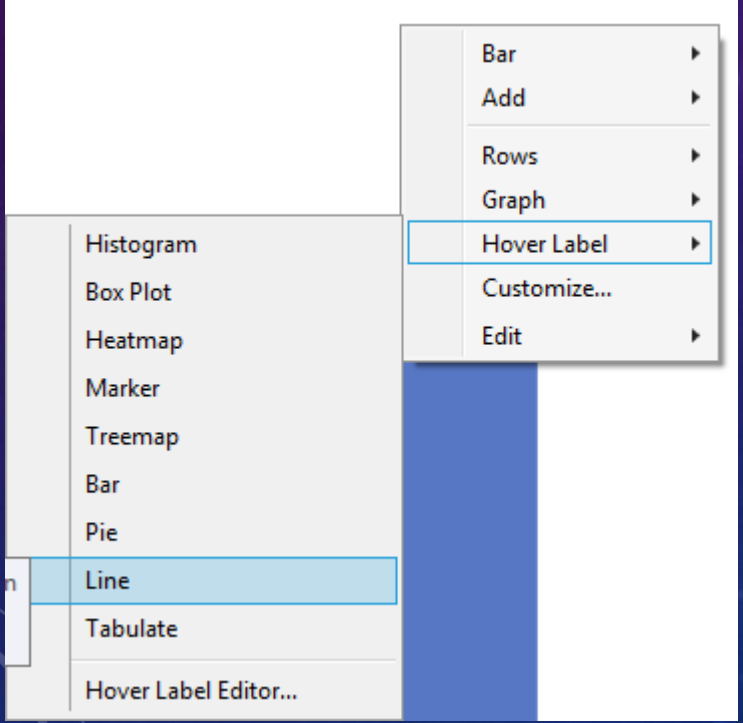
DRILL DOWN

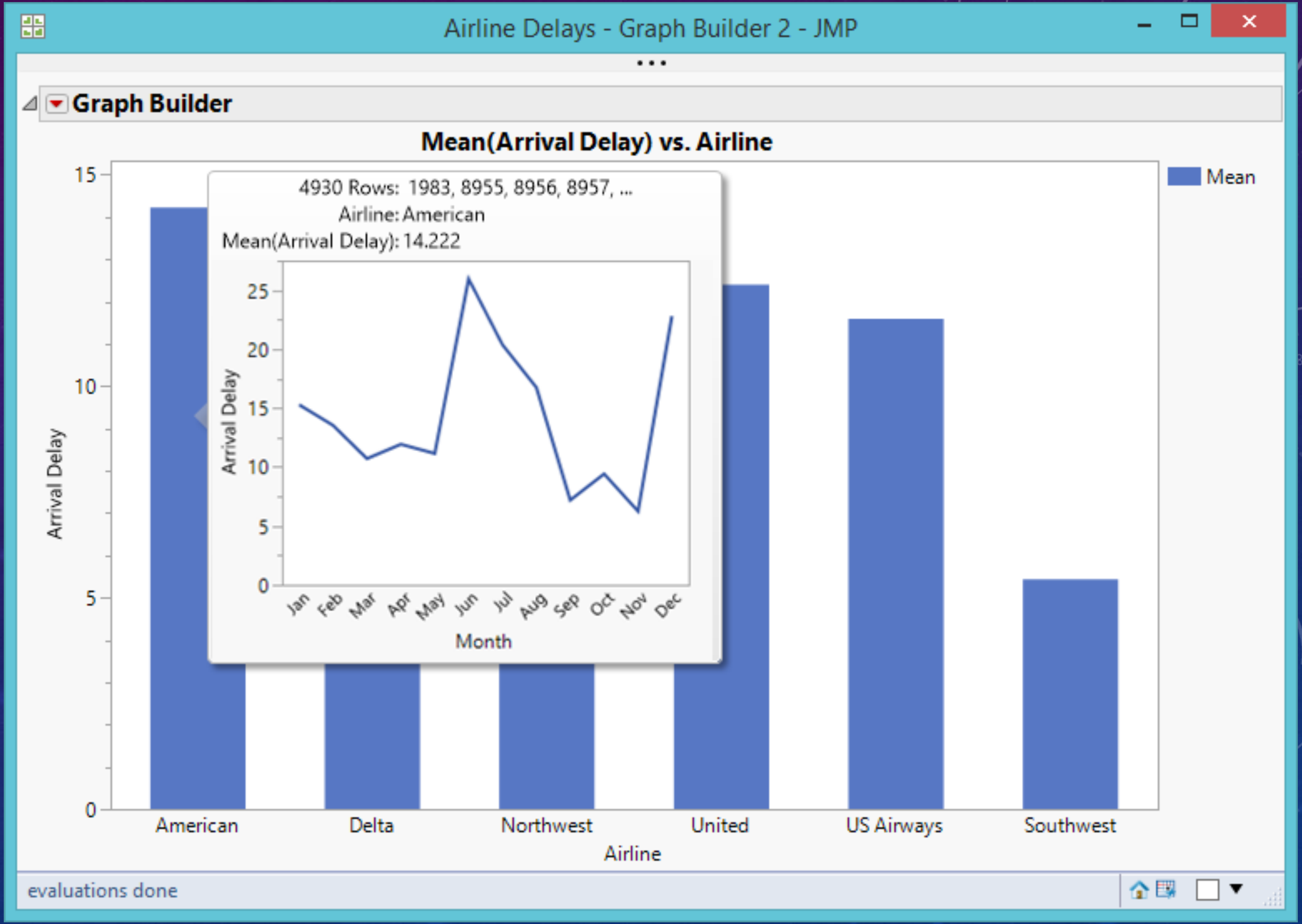
- Seeing four or more dimensions simultaneously may cause cognitive overload. You can localize the information by seeing one particular subset at one time.
- Use drill down to dig into the next level of detail if a particular piece of data is interesting
- American Airlines has the longest delay. What is going on?



DRILL DOWN

- Mouse over “American” and a line chart pops up



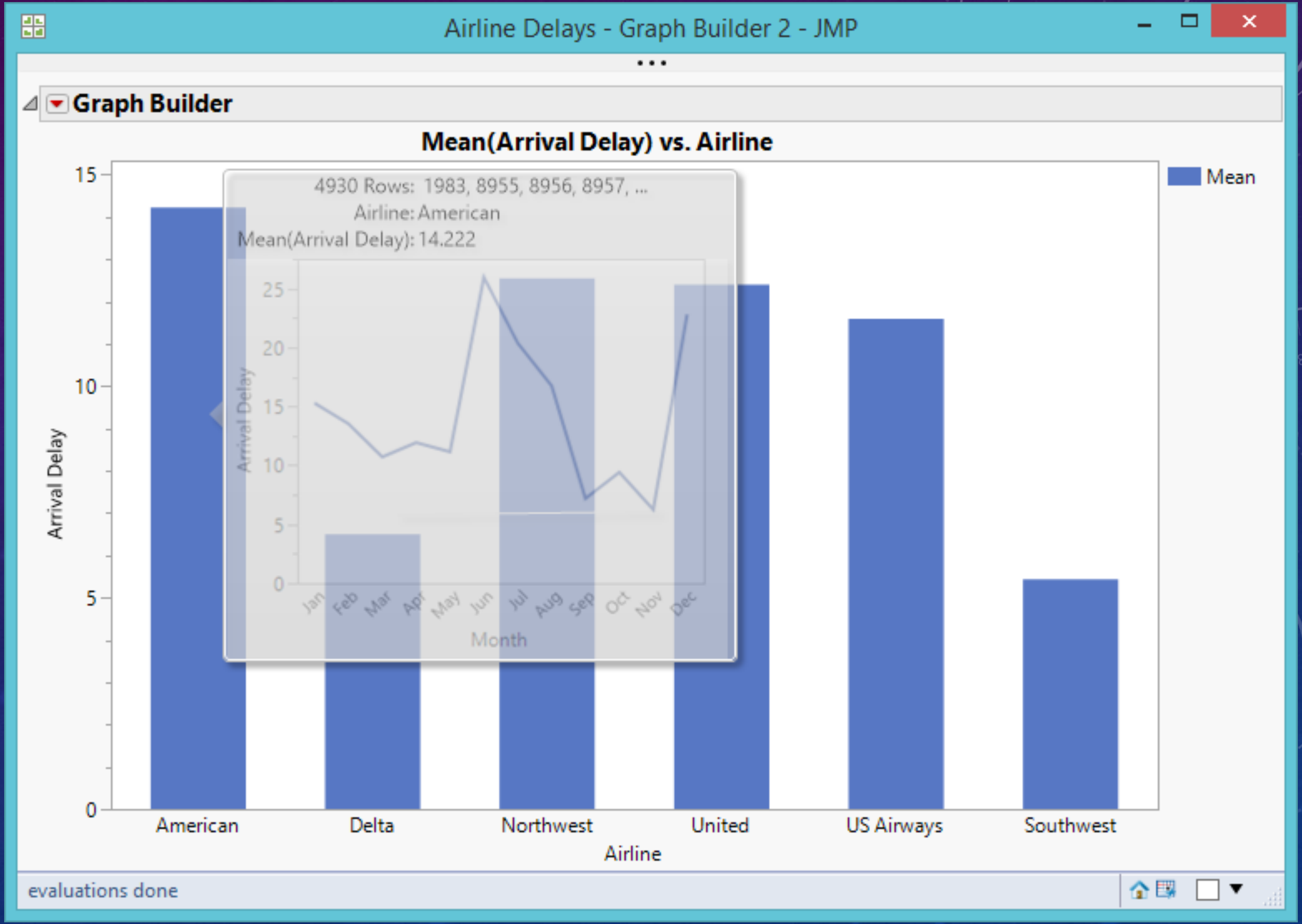


4930 Rows: 1983, 8955, 8956, 8957, ...
Airline: American
Mean(Arrival Delay): 14.222

evaluations done

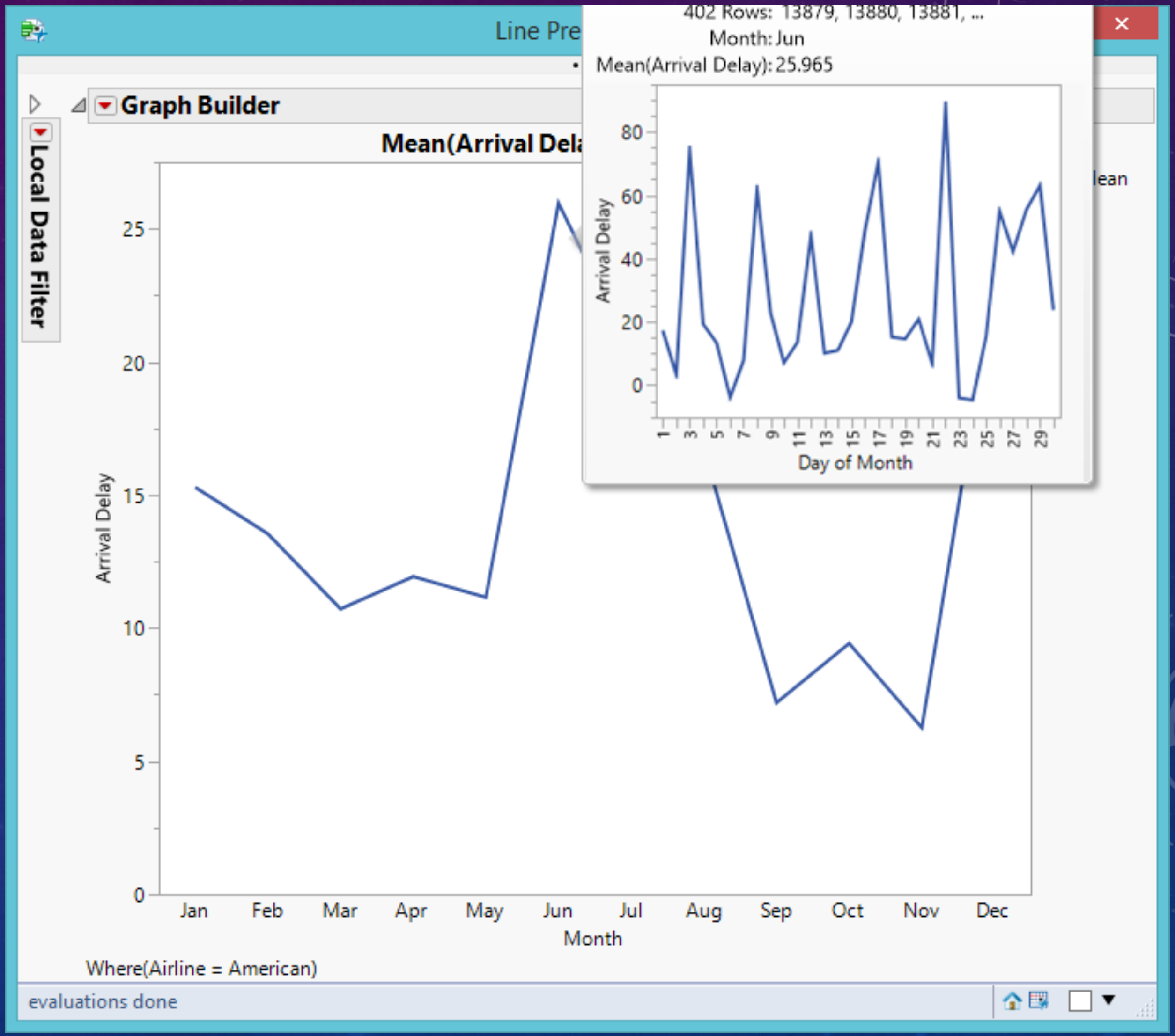
DRILL DOWN

- You can keep moving the mouse cursor to other airlines.
- If you click on the line chart, the graph will expand.



DRILL DOWN

- It seems that the worst is June.
- When you mouse over “June,” another line chart showing Day of Month pops up.
- Click on the new line chart and it will expand.
- The worst is day is June 22.
- Investigate what happened on that day!

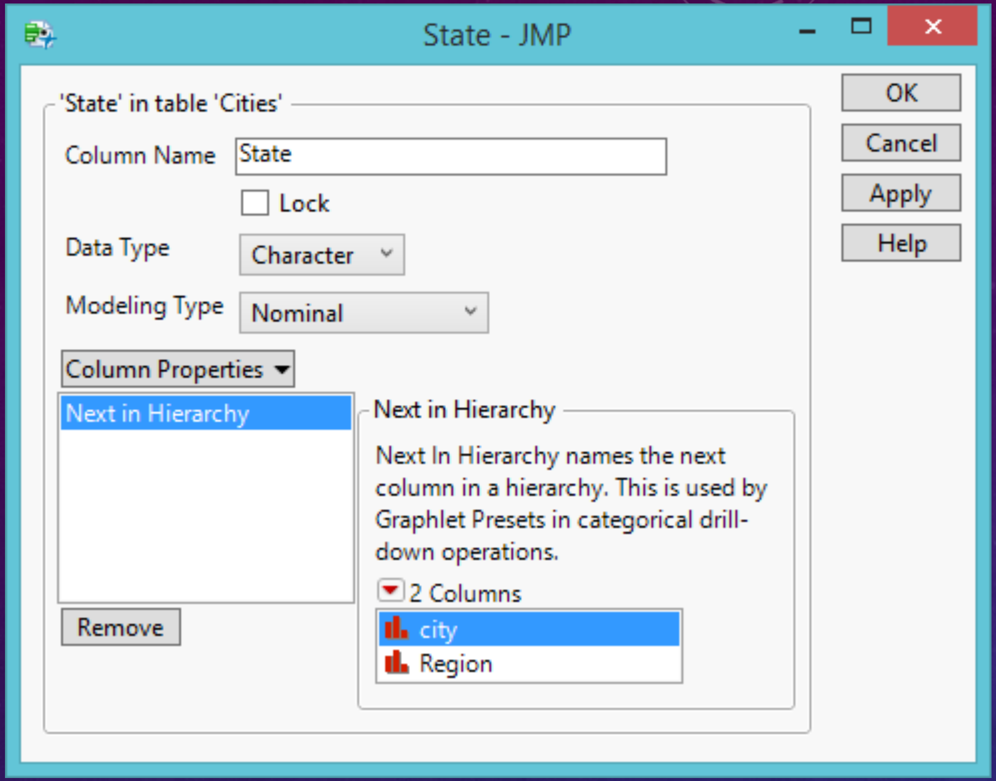


402 Rows: 13879, 13880, 13881, ...
Month: Jun
Mean(Arrival Delay): 25.965

Where(Airline = American)

evaluations done

DRILL DOWN



State - JMP

'State' in table 'Cities'

Column Name: State

□ Lock

Data Type: Character

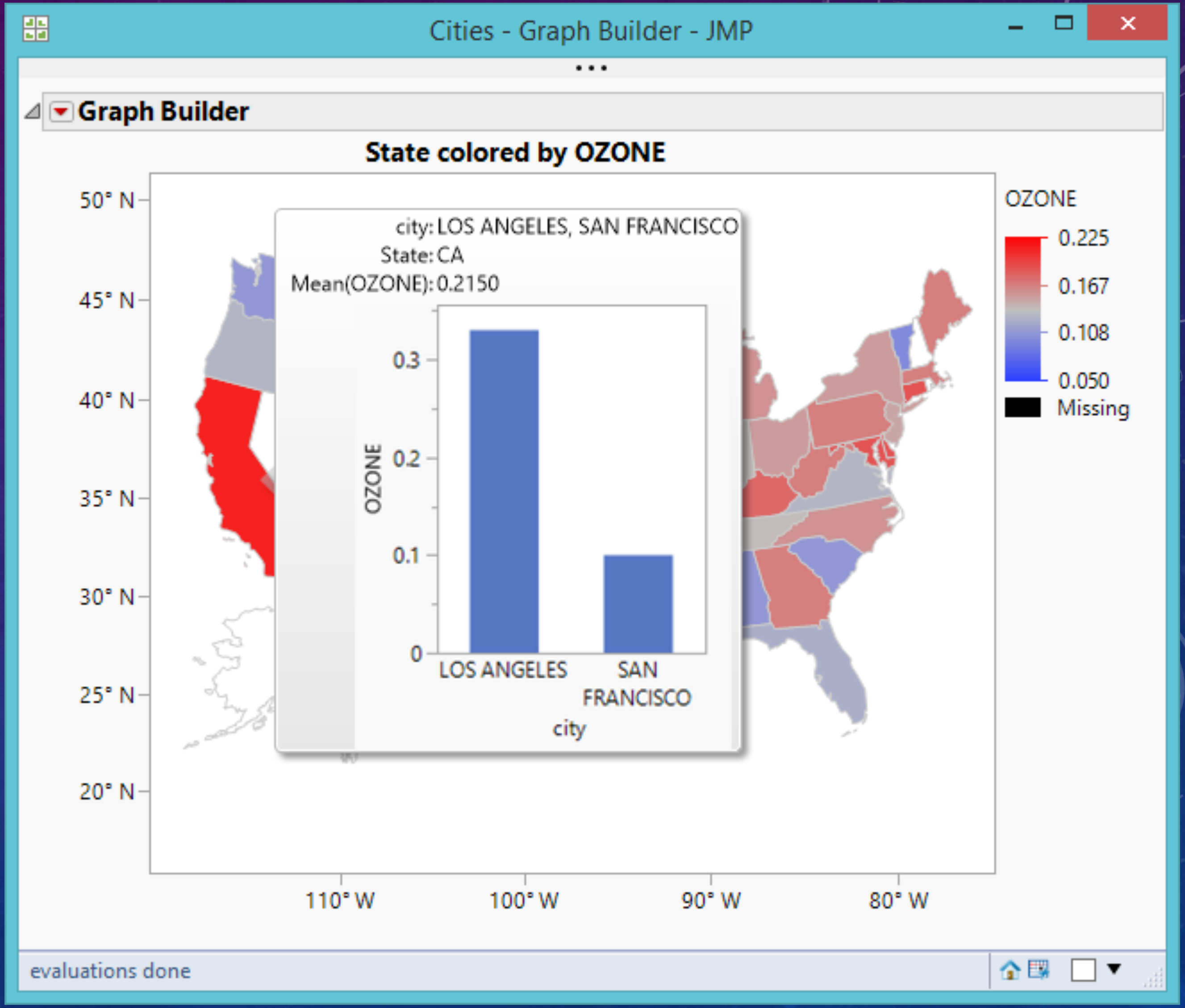
Modeling Type: Nominal

Column Properties

Next in Hierarchy

Next In Hierarchy names the next column in a hierarchy. This is used by Graphlet Presets in categorical drill-down operations.

2 Columns
- city
- Region



Cities - Graph Builder - JMP

Graph Builder

State colored by OZONE

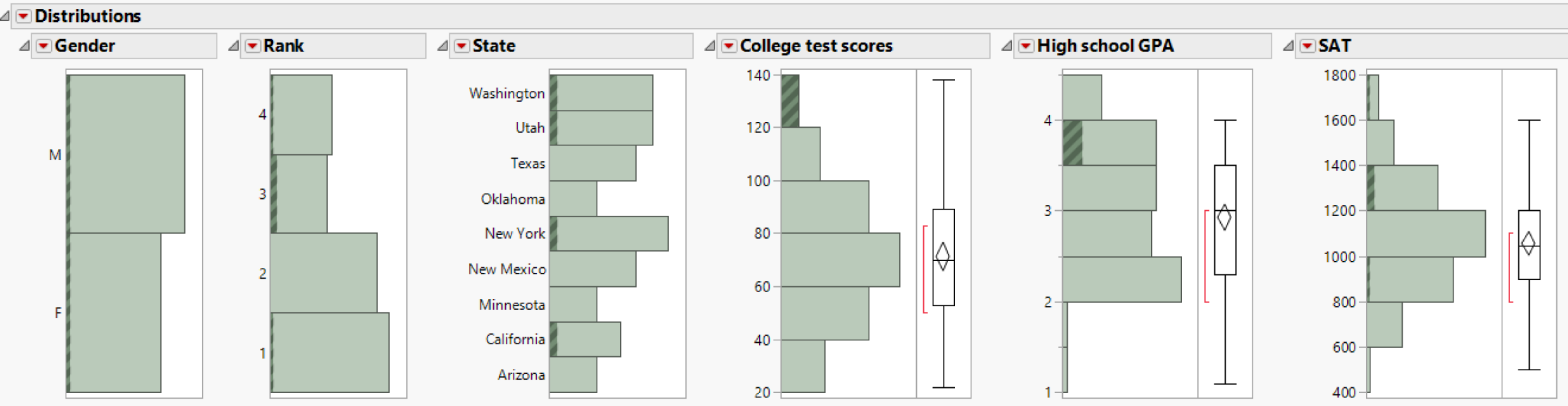
city: LOS ANGELES, SAN FRANCISCO
State: CA
Mean(OZONE): 0.2150

evaluations done

- This example shows a temporal order: Month, Day of Month, Day of Week.
- If you have a spatial data set, it could be: Country, State/Province, Region, County, City, Zip code.

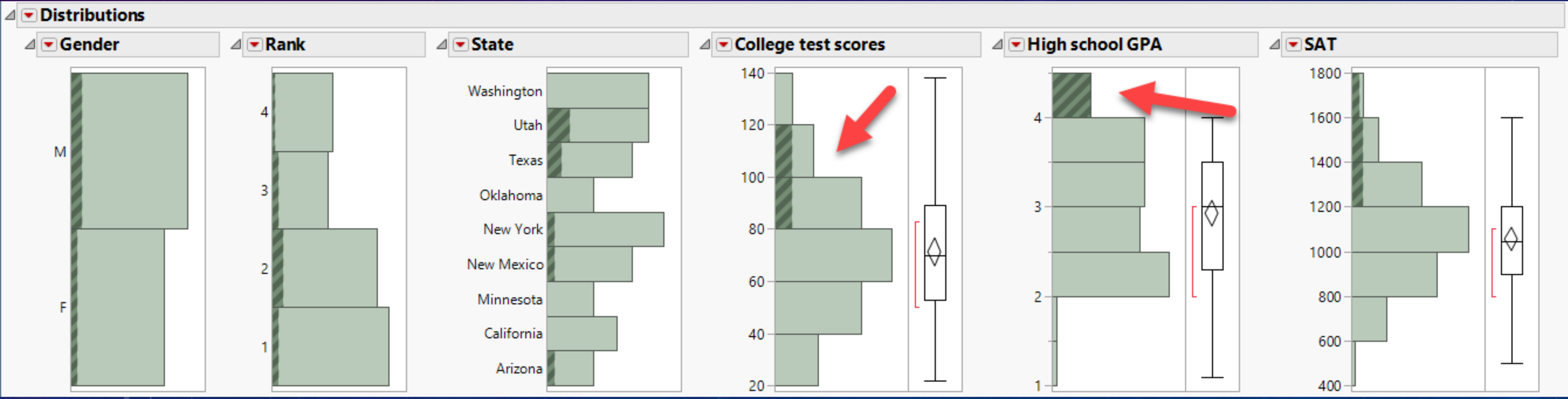
BEYOND 4-DIMENSION: LINKING AND BRUSHING

- What are the characteristics of top performers in the college test?
- They are from WA, UT, and CA.
- Their high school GPA is good but their SAT is not necessarily good.



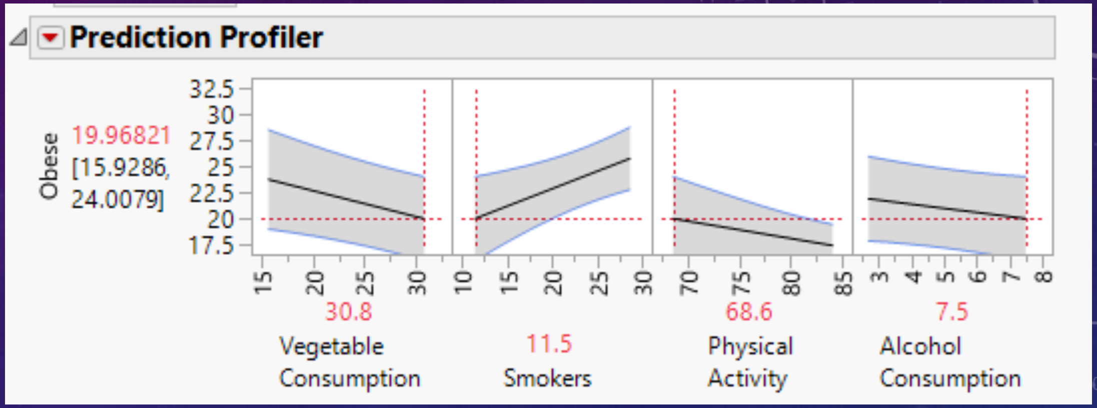
LINKING AND BRUSHING

- Interestingly, students whose high school GPA is perfect (4.0) are not the top performers in college.



PREDICTION PROFILER

- What would the obesity rate be if vegetable consumption is high, the smoker rate is low, physical activity is medium, and alcohol consumption is low?
- Ask "What if…." question?
- You can use profiler to go beyond 5 dimensions.
- http://www.creative-wisdom.com/teaching/551/Lecture_PowerPoint/Unit_5_multi-dimensional/profiler.html



Prediction Profiler

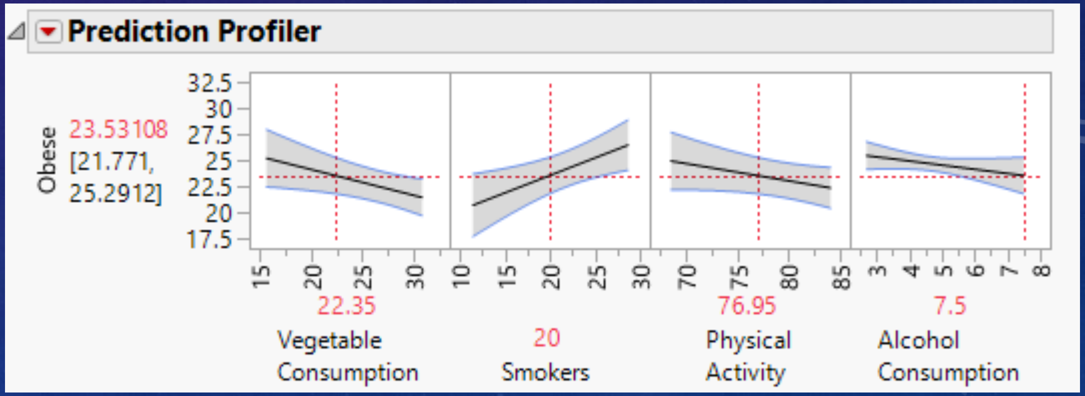
Obese 19.96821 [15.9286, 24.0079]

Vegetable Consumption 30.8

Smokers 11.5

Physical Activity 68.6

Alcohol Consumption 7.5



Prediction Profiler

Obese 23.53108 [21.771, 25.2912]

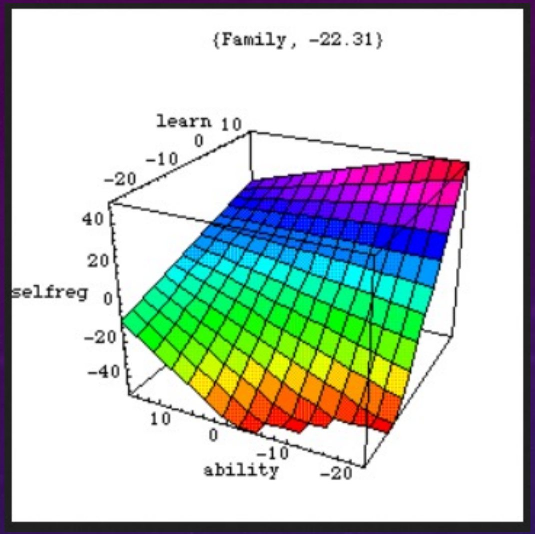
Vegetable Consumption 22.35

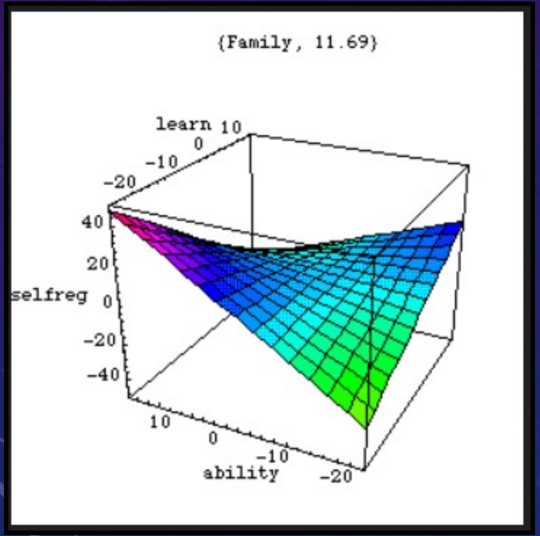
Smokers 20

Physical Activity 76.95

Alcohol Consumption 7.5

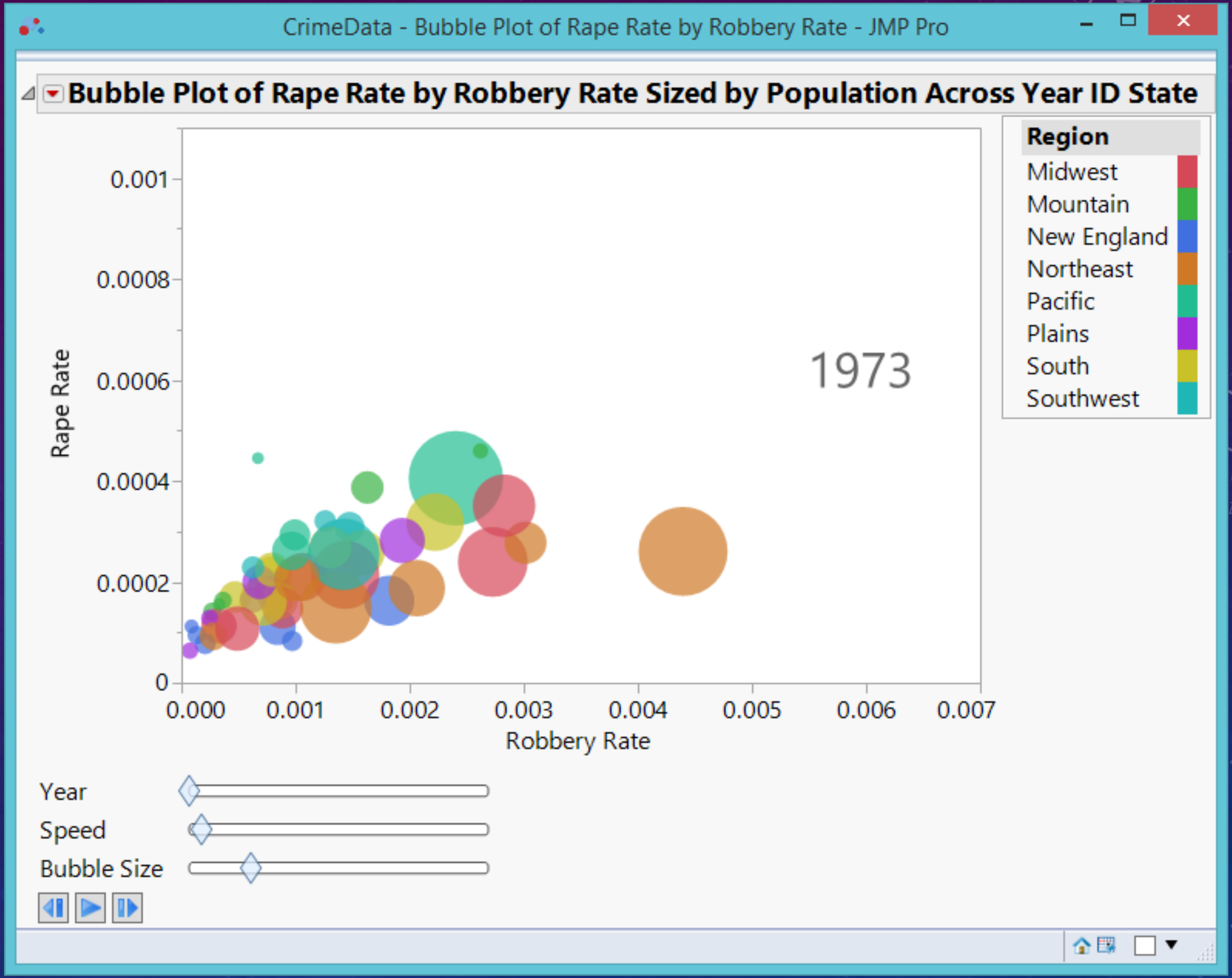
DANCING WITH 3-WAY INTERACTION





- In the three-way interaction, the fourth dimension is the temporal dimension (e.g. when d = 1, d = 2, d =3…etc.)
- Interaction: the effect of X on Y is not consistent across all levels of A and B → regression lines vary.
- If there is NO interaction, there should be no curving or dancing in the movie. Every frame should look the same.
- http://creative-wisdom.com/multimedia/regression.html

BUBBLE PLOT



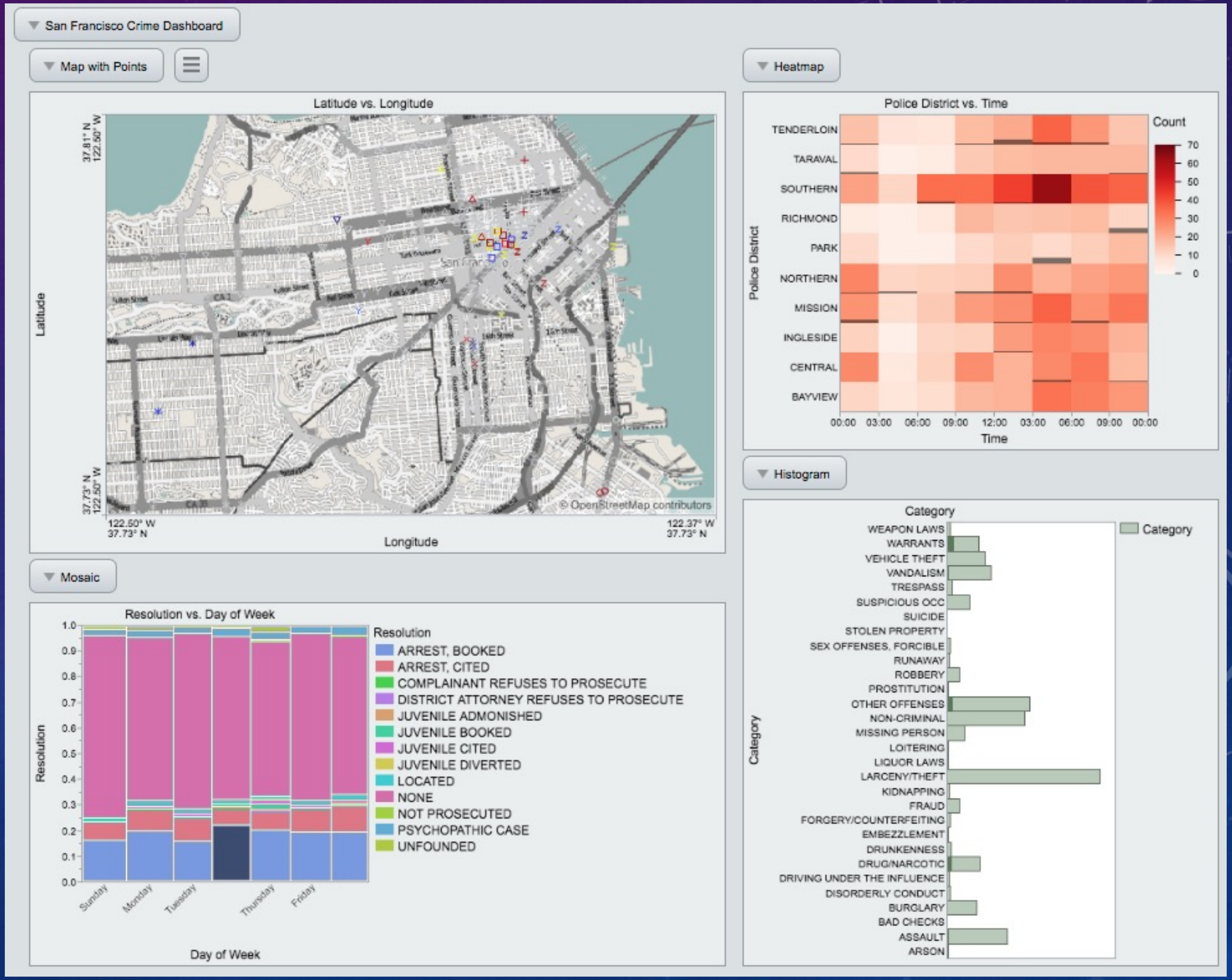
https://creative-wisdom.com/teaching/551/Lecture_PowerPoint/Unit_6_time-series/Bubble_Plot.html

WHAT THE BUBBLE DANCE (ANIMATION) TELL ME ABOUT THE OVERALL PATTERN?

- In 1973 a strong association was found between the two crime rates, but in 1993 their connection became weaker.
- In both years big cities with a large population size tended to suffer from higher crime rates, with the Northeast region being the worst.
- The US crime rate has been steadily declining since the 1990s. In 2010, the crime rates appear to be under control. The robbery rate and the rape rate seemed to be negatively correlated.
- In many years Alaska is the worst in sexual assault rate.
- Big cities and Northeast are no longer the most dangerous places to live.
- Dancing with the data!

DASHBOARD

- http://www.creative-wisdom.com/teaching/551/Lecture_PowerPoint/Unit_8_dashboard/JMP_files/Dashboard.htm
- More arrests (booked, released by paying bail) on Wednesday.
- Most crimes are concentrated on a the southern region of SF.
- More crimes happen in the late afternoon (3-6 PM).
- These salient characteristics pop up!



Q & A AND CONTACT INFO

- Chong Ho Yu, Ph.D., D. Phil.
- Email: cyu@apu.edu, chonghoyu@gmail.com
- Website: http://www.creative-wisdom.com/pub/pub.html
- Linkedin: https://www.linkedin.com/in/chong-ho-alex-yu-00b22547/