Dancing with the data

Chong Ho Yu (Alex)
Ph.D. (Psychometrics)
Ph.D. (Philosophy of science)
Associate Professor of Psychology
Azusa Pacific University
cyu@apu.edu

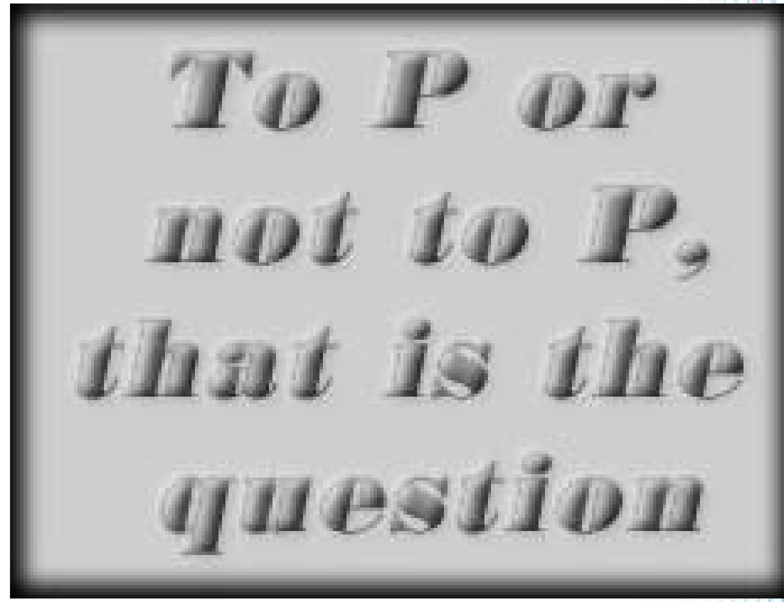
Seminar presented at
Center for Information Technology in Education
University of Hong Kong
July 15, 2014

Agenda

- Shortcomings and misuse of conventional hypothesis testing
- Rationale for and misconceptions of exploratory data analysis and data visualization
- Visualization techniques from 2 to 5 dimensions
- Future trend: multi-panel visualization to go beyond 5 dimensions

Shortcoming of conventional approach

- Over-reliance on hypothesis testing/confirmatory data analysis (CDA) and *p* values.
- The logic of hypothesis testing is: Given that the null hypothesis is true how likely we can observe the data in the **long run**? P(D|H)?
- What we really want to know is: Given the data what is the best theory to explain the data **no matter whether the event can be repeated** : P(H|D)?



Affirming the consequent

- P(D|H) <> P(H|D): “If H then D" does not logically imply "if D then H".

- If the theory/model/hypothesis is correct, it implies that we could observe Phenomenon X or Data X.
- X is observed.
- Hence, the theory is correct.

Affirming the consequent

- If George Washington was assassinated, then he is dead.
- George Washington is dead.
- Therefore George Washington was assassinated.

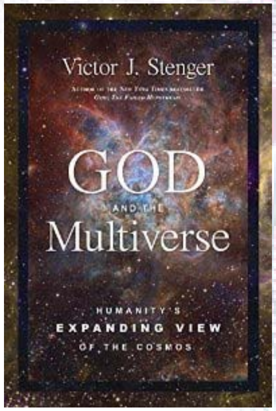
- If it rains, the ground is wet.
- The ground is wet.
- It must rain.

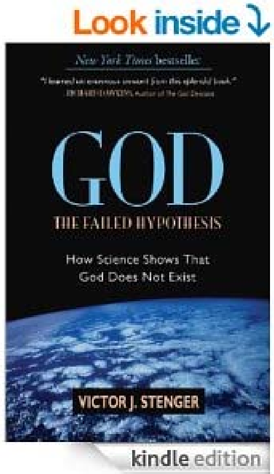
Can we “prove” or “disprove”?

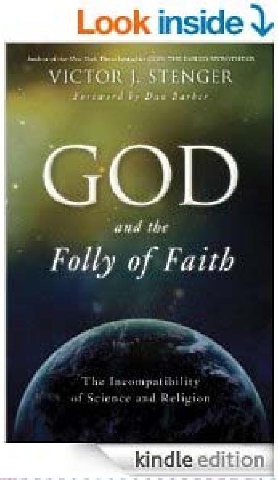
- Hypothesis testing or confirmatory data analysis (CDA):
  - Start with a strong theory/model/hypothesis
  - Collect data to see whether the data match the model.
  - If they fit each other, did you “prove” the theory?
  - If they don't, did you “disprove” it?
  - At most you can say whether the data and the model fit each other. In philosophy it is called “empirical adequacy.”

God: Failed hypothesis

- Prominent physicist Victor Stenger:
- “Our bones lose minerals after age thirty, making them susceptible to fracture and osteoporosis. Our rib cage does not fully enclose and protect most internal organs. Our muscles atrophy. Our leg veins become enlarged and twisted, leading to varicose veins. Our joints wear out as their lubricants thin. Our retinas are prone to detachment. The male prostate enlarges, squeezing and obstructing urine flow.”
- Hence, there is no intelligent designer.







Logical fallacy

- Hypothesis: If there is a God or intelligent designer, he is able to design a well-structured body. To prove the existence of God, we look for such data: P(D|H)
- No such data: Our bones start losing minerals after 30, and there are other flaws, and thus God is a “failed” hypothesis.
- You will see what you are looking for.
- But there are other **alternate explanations** that can fit the data.
- e.g. God did not make our body last forever, and thus dis-integration and aging is part of the design.

Common mistakes about p values

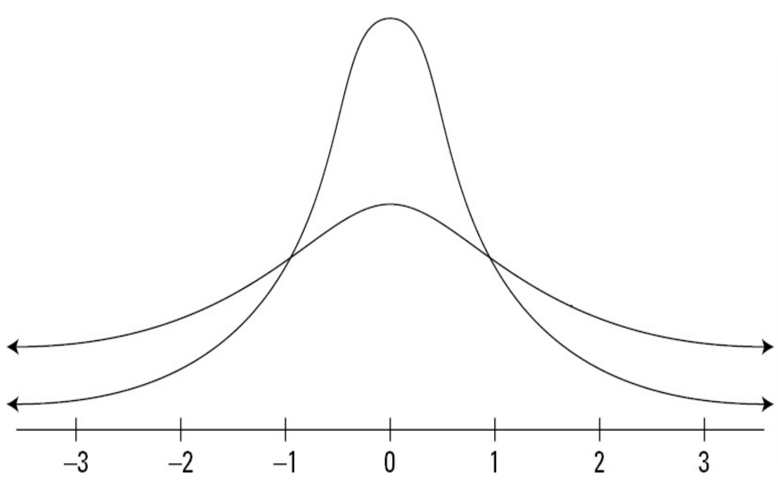
These multidimensional items were summed and proved to form a consistent scale of Lived Poverty (Cronbach's Alpha = 0.70), demonstrating a high level of internal consistency. The index is also strongly correlated at macro-level with both per capita GDP (in PPP) (Pearson R = 0.884, P = .000, N = 120) and the UNDP Human Development Index (R = 0.673, P = .000, N = 123), suggesting high levels of external validity.

The scatter plots presented in Figures 12.1 and 12.2 illustrate the macro-level relationship between the Lived Poverty Index and the distribution of religious values and practices across the 128 nations where complete data is available, without any prior controls. The results confirm that the Lived Poverty Index was indeed strongly correlated with religious values (R = 0.541, P = .000, N = 128); hence, some of the poorest

The comparison with religious practices, illustrated in Figure 12.2, shows a similar and almost equally strong relationship; thus, without any controls, the Lived Poverty Index proved to be a significant predictor of participation in religious services (R = .497, P = .000, N = 127). Again, the least developed nations, such as Chad, Uganda, Togo, and Rwanda, clustered together in the top right-hand quadrant, contain the poorest and most religious countries. By contrast, affluent Scandinavian and West European Protestant societies reported the lowest church attendance, along with Australia, New Zealand, and Canada. The United States is generally viewed as a deviant case, in that it is a rich country with higher church attendance than other affluent societies. But in a broader comparative perspective provided here, U.S. levels of religious participation are much closer to those found in Italy, Switzerland, and Portugal than to many other countries with low levels of economic development.

Can *p* be .000?

- P = probability that the statistics can be observed in the long run.
- "Long run" is expressed in terms of **sampling distributions**, in which sampling, in theory, is repeated infinitely.
- The two tails never touch down the x-axes.



- In an open universe anything has a remote probability.

Can *p* be .000?

Intraclass Correlation Coefficient

	Intraclass Correlation[b]	95% Confidence Interval		F Test with True Value 0			
		Lower Bound	Upper Bound	Value	df1	df2	Sig
Single Measures	.190[a]	-.017	.626	14.246	9	9	.000
Average Measures	.320[c]	-.035	.770	14.246	9	9	.000

Two-way mixed effects model where people effects are random and measures effects are fixed.

a. The estimator is the same, whether the interaction effect is present or not.

b. Type A intraclass correlation coefficients using an absolute agreement definition.

c. This estimate is computed assuming the interaction effect is absent, because it is not estimable otherwise.

- If p = 0.000, then it means there is no chance for such event to happen. Does it make any sense?
- When the p value is too small, SAS uses the e-notation and JMP reports it as p < .001, but SPSS shows it as .000.

# P < .0001 → Significant?

Linear Fit

Linear Fit

Q4 = 2.8134904 + 0.4335653*A6

Summary of Fit

RSquare	0.141374
RSquare Adj	0.134718
Root Mean Square Error	1.123388
Mean of Response	5.183206
Observations (or Sum Wgts)	131

Lack Of Fit

Analysis of Variance

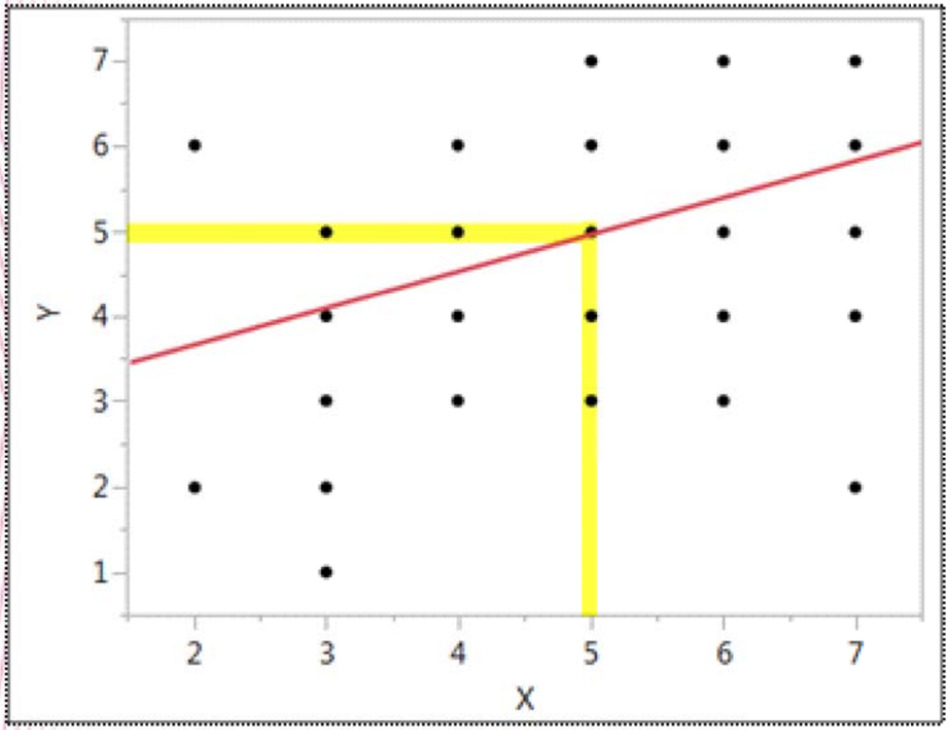
Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	26.80493	26.8049	21.2400
Error	129	162.79813	1.2620	Prob > F
C. Total	130	189.60305		<.0001*

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	2.8134904	0.523468	5.37	<.0001*
A6	0.4335653	0.094076	4.61	<.0001*

- In this simple regression model X is used to predict Y.
- P < .0001, significant! You may shout, "Allelujah!"
- Do you think it is a good model?

Significant: How rare the event is



- If my score on X is 5, the regression model predicts that my score on Y is also 5.
- Actually, it could be 3, 4, 5, 6, or 7.
- Five of out of seven! This "predictive" model is usefulness!
- Lesson: the p value can fool you!!!

A picture is worth a thousand *p* values

- In 1989, when Kenneth Rothman started the *Journal of Epidemiology*, he discouraged over–reliance on *p* values. However, the earth is round. When he left his position in 2001, the journal reverted to the *p*–value tradition.
- In *A Picture is Worth a Thousand p Values*, Loftus observed that many journal editors do not accept the results reported in mere graphical form. Test statistics must be provided for the consideration of publication. Loftus asserted that hypothesis testing ignores two important issues:
  - What is the **pattern** of population means over conditions?
  - What are the **magnitudes** of various variability measures?

What should be done?

- Reverse the logic.
- What people are doing now: starting with a single hypothesis and then computing the p value based on one sample: P(D|H)
- We should ask: given the pattern of the data, what is the best explanation out of many alternate theories (**inference to the best explanation**) using resampling, exploratory data analysis, data visualization, data mining: P(H|D)
- Today we focus on data visualization

Common misconceptions about EDA and data mining (DM)

- "It is fishing": Actually DM avoids fishing and capitalization on chance (over-fitting) by resampling (e.g. **cross-validation**).
- "There is no theory": Both EDA and CDA have some theories. CDA has a strong theory (e.g. Victor Stenger: There is no God) whereas EDA/DM has a weak theory.
- In EDA/DM when you select certain potential factors into the analysis, you have some rough ideas. But you **let the data speak for themselves**.

Common misconceptions about EDA and data mining (DM)

- "DM and EDA are based on **pattern recognition** of the data at hand. It cannot address the probability in the long run"
- Induction in the long run is based on the assumption that **the future must resemble the past**. Read David Humes, Nelson Goodman, and Nissam Taleb.
- Some events **are not repeatable** (Big bang).
- It is more realistic to make inferences based on the current patterns to the **near future**.

Titanic survivors

- After the disaster, people asked: What types of people tend to survive?

Whole Model Test

Model	-LogLikelihood	DF	ChiSquare	Prob>ChiSq
Difference	279.6978	5	559.3956	<.0001*
Full	1105.0306			
Reduced	1384.7284			

RSquare (U)	0.2020
AICc	2222.1
BIC	2256.24
Observations (or Sum Wgts)	2201

Measure	Training	Definition
Entropy RSquare	0.2020	1-Loglike(model)/Loglike(0)
Generalized RSquare	0.3135	(1-(L(0)/L(model))^(2/n))/(1-L(0)^(2/n))
Mean -Log p	0.5021	∑ -Log(ρ[j])/n
RMSE	0.4026	$\sqrt{\sum(y[j]-\rho[j])^2/n}$
Mean Abs Dev	0.3253	∑ |y[j]-ρ[j]|/n
Misclassification Rate	0.2217	∑ (ρ[j]≠ρMax)/n
N	2201	n

Lack Of Fit

Source	DF	-LogLikelihood	ChiSquare
Lack Of Fit	8	56.2833	112.5666
Saturated	13	1048.7473	Prob>ChiSq
Fitted	5	1105.0306	<.0001*

Parameter Estimates

Term	Estimate	Std Error	ChiSquare	Prob>ChiSq
Intercept	-0.4511951	0.1272927	12.56	0.0004*
Class[crew]	-0.0557072	0.0908828	0.38	0.5399
Class[first]	-0.9133833	0.1102375	68.65	<.0001*
Class[second]	0.10471162	0.1180263	0.79	0.3750
Age[adult]	0.53077119	0.1220129	18.92	<.0001*
Sex[female]	-1.2100302	0.0702051	297.07	<.0001*

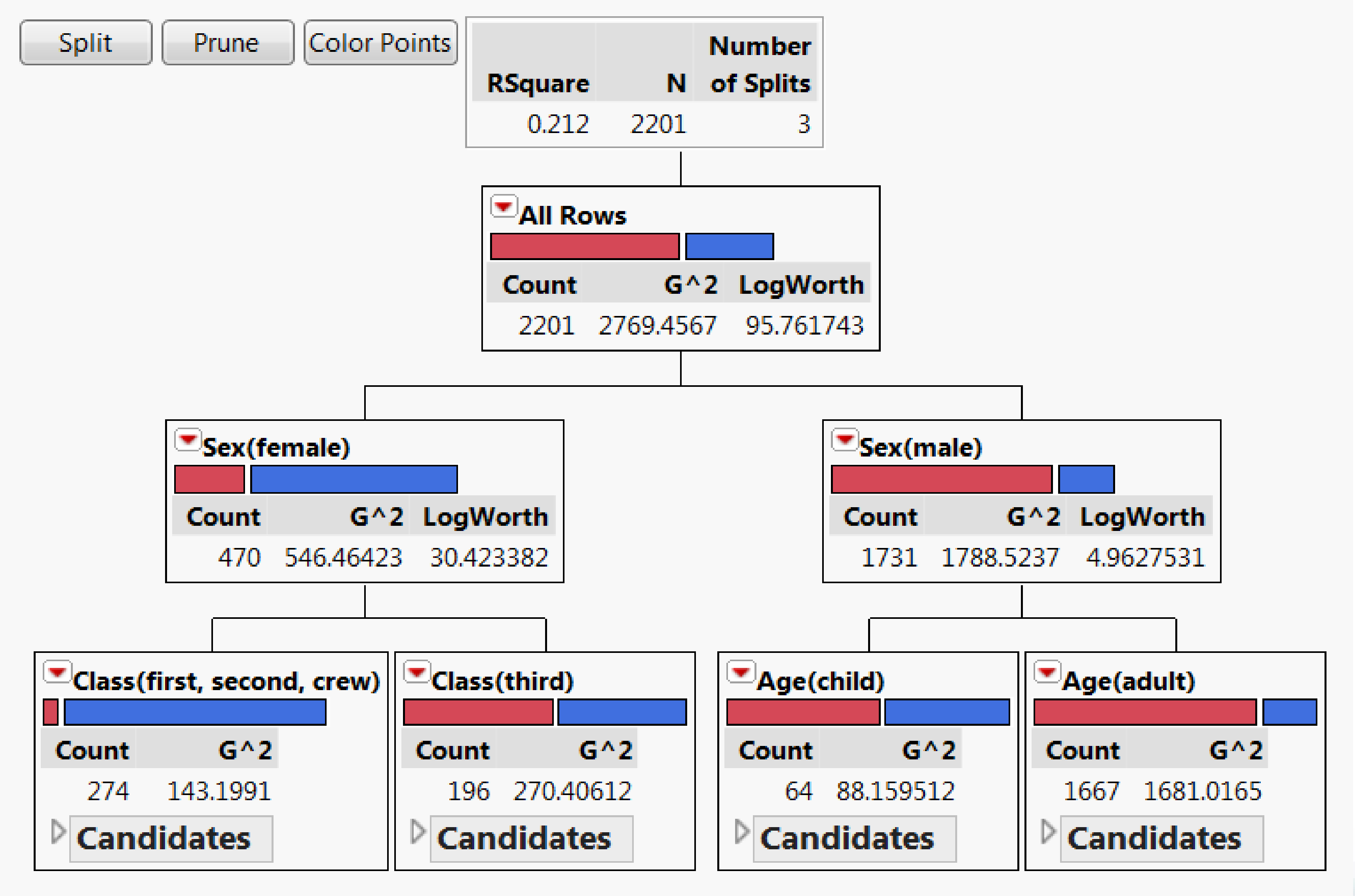
For log odds of no/yes

Covariance of Estimates

Effect Likelihood Ratio Tests

Source	Nparm	DF	L-R ChiSquare	Prob>ChiSq
Class	3	3	119.03384	<.0001*
Age	1	1	18.8517145	<.0001*
Sex	1	1	352.911216	<.0001*

Decision tree



Leaf report

Leaf Report

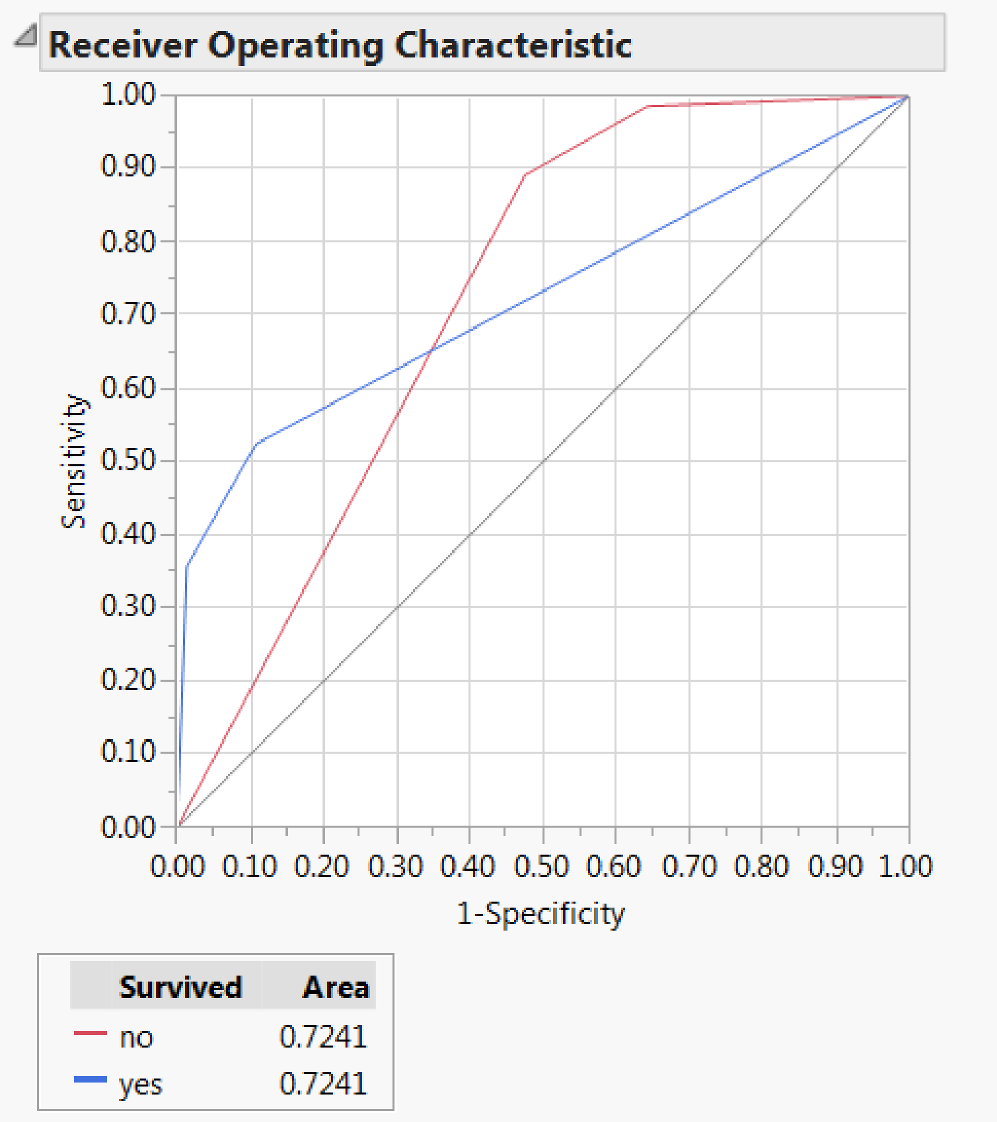
Response Prob

Leaf Label	no	yes
Sex(female)&Class(first, second, crew)	0.0750	0.9250
Sex(female)&Class(third)	0.5413	0.4587
Sex(male)&Age(child)	0.5490	0.4510
Sex(male)&Age(adult)	0.7972	0.2028

Response Counts

Leaf Label	no	yes
Sex(female)&Class(first, second, crew)	20	254
Sex(female)&Class(third)	106	90
Sex(male)&Age(child)	35	29
Sex(male)&Age(adult)	1329	338

ROC curves and AUC



4 possible outcomes:

- true positive (TP)
- false positive (FP)
- false negative (FN)
- true negative (TN).

Sensitivity = TP/(TP+FN)

1 - Specificity = 1 - [TN/(FP+TN)]

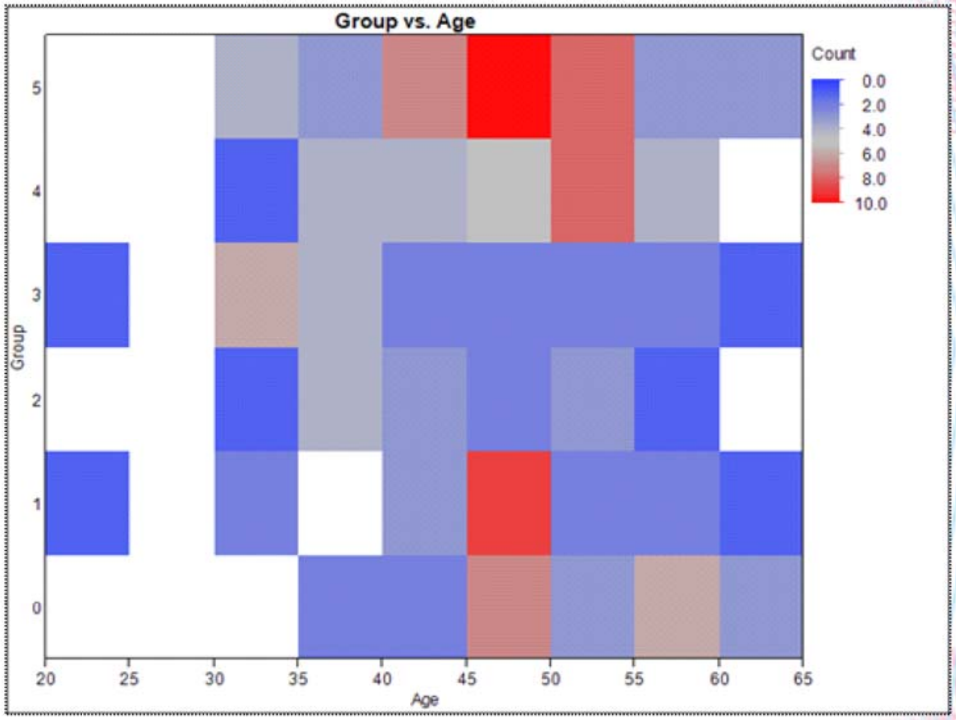
No model = 50% chance

Example: Logistic regression

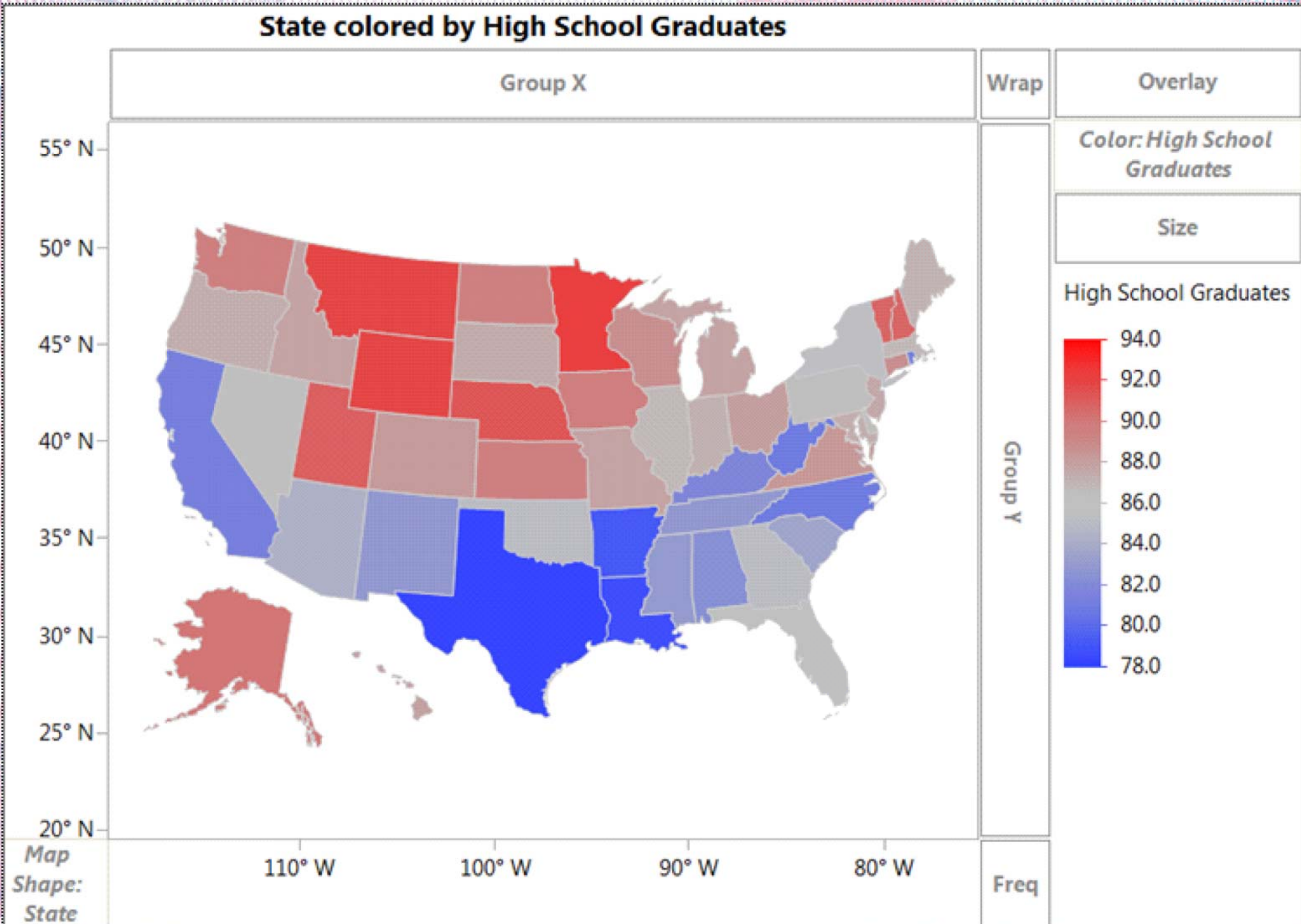
Parameter Estimates

Term	Estimate	Std Error	ChiSquare	Prob>ChiSq
Intercept[0]	-3.9568269	1.7597155	5.06	0.0245*
Age[0]	0.07073973	0.0351409	4.05	0.0441*
Intercept[1]	0.28606972	1.5540312	0.03	0.8539
Age[1]	-0.0202195	0.0334711	0.36	0.5458
Intercept[2]	0.83356399	1.7154638	0.24	0.6270
Age[2]	-0.0405404	0.0378312	1.15	0.2839
Intercept[3]	3.0865254	1.5202007	4.12	0.0423*
Age[3]	-0.0854613	0.0348684	6.01	0.0142*
Intercept[4]	-0.6470719	1.4798935	0.19	0.6619
Age[4]	0.00572117	0.0311456	0.03	0.8543

For log odds of 0/5, 1/5, 2/5, 3/5, 4/5



- Aged between 45 and 50 → in group 1 and 5.


State colored by High School Graduates
Group X
Wrap
Overlay
Color: High School Graduates
Size
High School Graduates
94.0
92.0
90.0
88.0
86.0
84.0
82.0
80.0
78.0
55° N
50° N
45° N
40° N
35° N
30° N
25° N
20° N
110° W
100° W
90° W
80° W
Group Y
Map Shape: State
Freq

ANOVA and multiple comparison

Oneway Anova

Summary of Fit

Rsquare	0.296455
Adj Rsquare	0.244341
Root Mean Square Error	10.82538
Mean of Response	79.43333
Observations (or Sum Wgts)	30

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio	Prob > F
Group	2	1333.2667	666.633	5.6885	0.0087*
Error	27	3164.1000	117.189		
C. Total	29	4497.3667			

Means for Oneway Anova

Level	Number	Mean	Std Error	Lower 95%	Upper 95%
Classroom	10	70.2000	3.4233	63.176	77.224
Hybrid	10	85.7000	3.4233	78.676	92.724
Online	10	82.4000	3.4233	75.376	89.424

Std Error uses a pooled estimate of error variance

Means Comparisons

Comparisons for all pairs using Tukey-Kramer HSD

Confidence Quantile

q*	Alpha
2.47942	0.05

LSD Threshold Matrix

Abs(Dif)-HSD

	Hybrid	Online	Classroom
Hybrid	-12.004	-8.704	3.496
Online	-8.704	-12.004	0.196
Classroom	3.496	0.196	-12.004

Positive values show pairs of means that are significantly different.

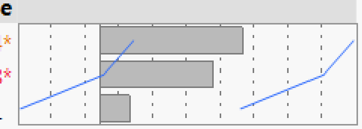
Connecting Letters Report

Level			Mean
Hybrid	A		85.700000
Online	A		82.400000
Classroom		B	70.200000

Levels not connected by same letter are significantly different.

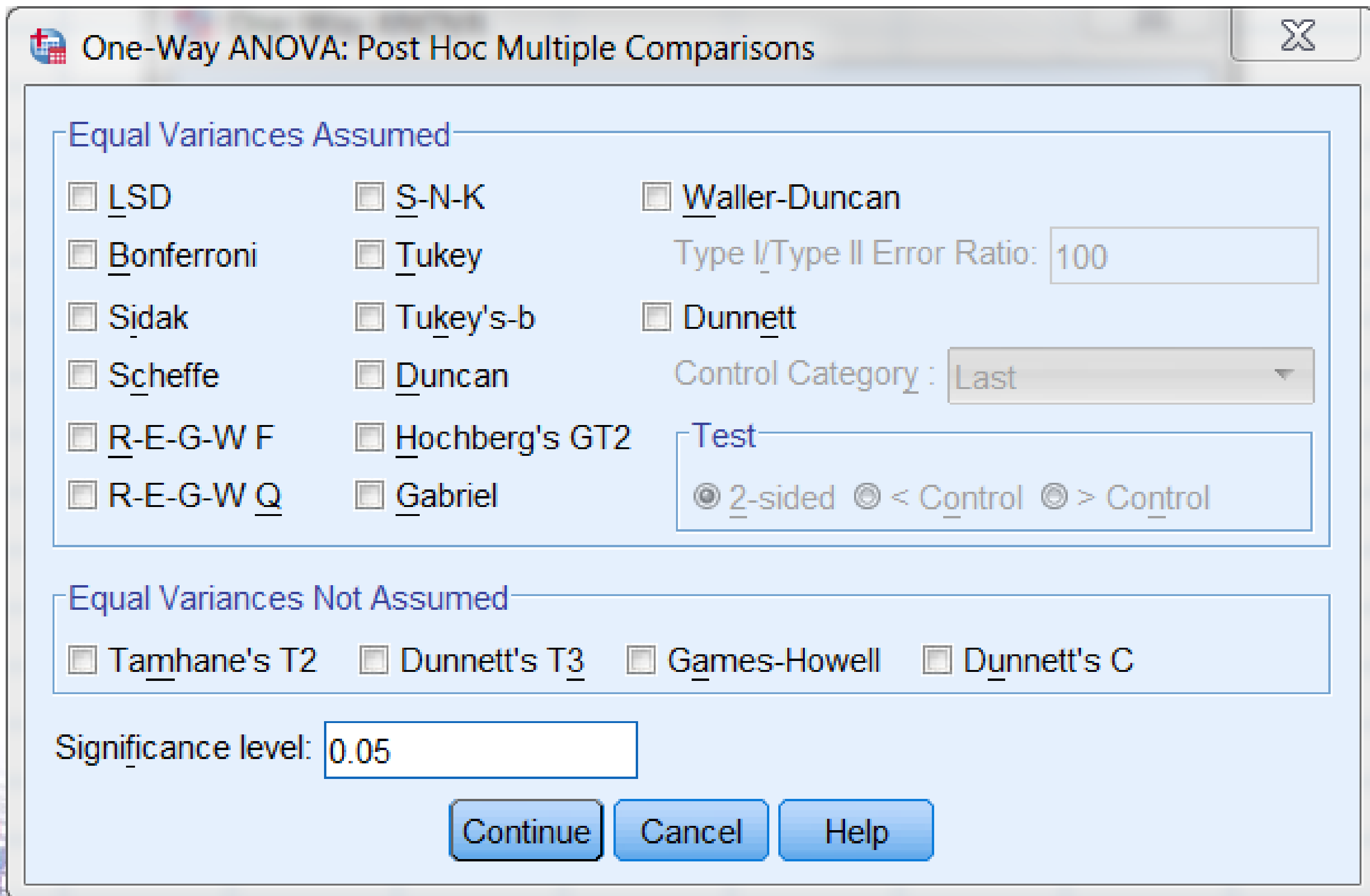
Ordered Differences Report

Level	- Level	Difference	Std Err Dif	Lower CL	Upper CL	p-Value
Hybrid	Classroom	15.50000	4.841258	3.49650	27.50350	0.0094*
Online	Classroom	12.20000	4.841258	0.19650	24.20350	0.0458*
Hybrid	Online	3.30000	4.841258	-8.70350	15.30350	0.7761

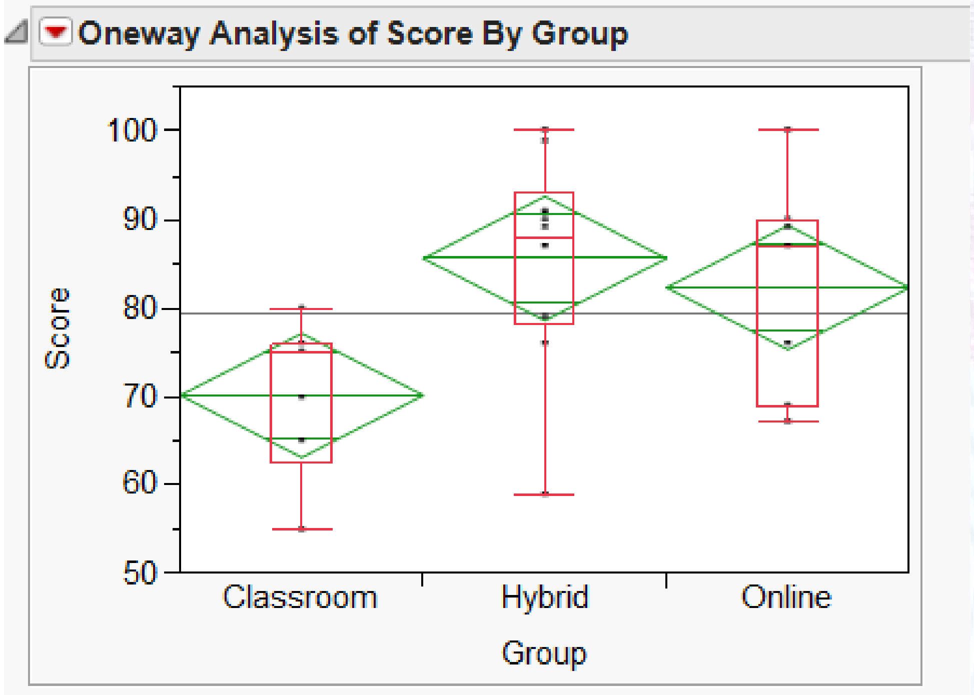


SPSS Post hoc multiple comparison

- In SPSS you have 18 options. When I was a graduate student, I took a course on it.



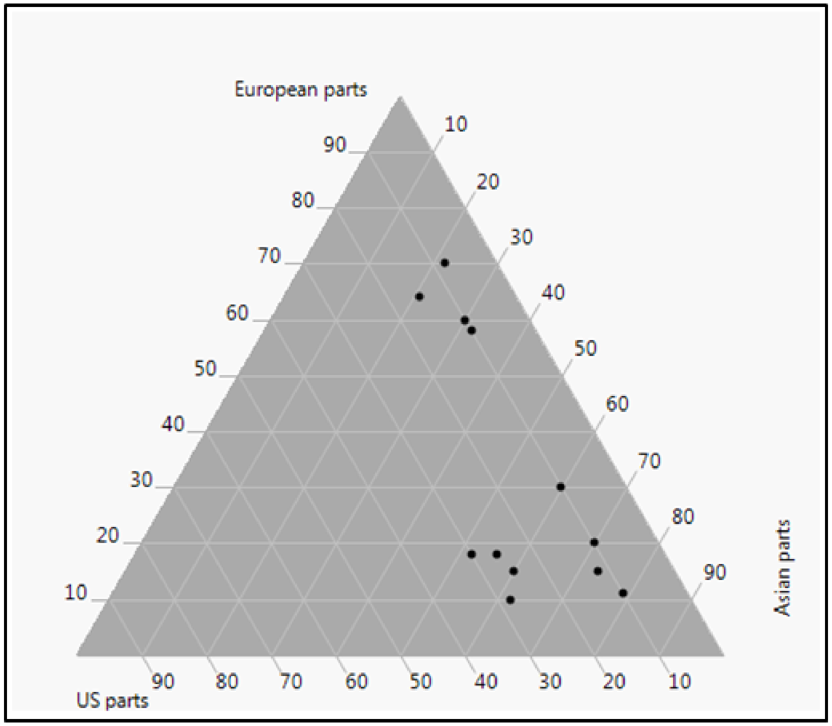
Diamond plot



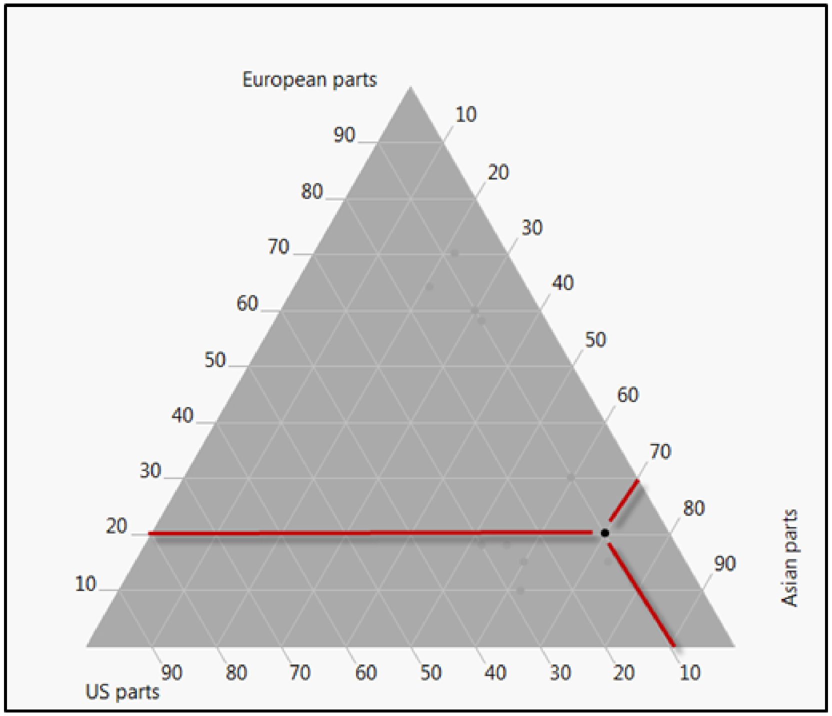
- Grand sample mean: horizontal black line
- Group means: horizontal line inside each diamond.
- Confidence intervals: The top of the diamond is the upper bound while the bottom is the lower bound.
- Quantile: boxplot

Ternary plot: Clustering and Profiling

- In the era of globalization, how can we define what a USA company is? One argue that if you buy a Korean Kia, you may help reducing the trade deficit.

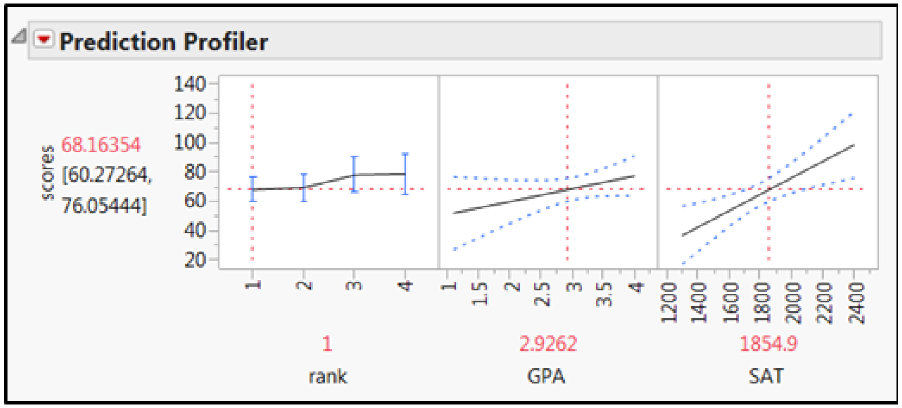


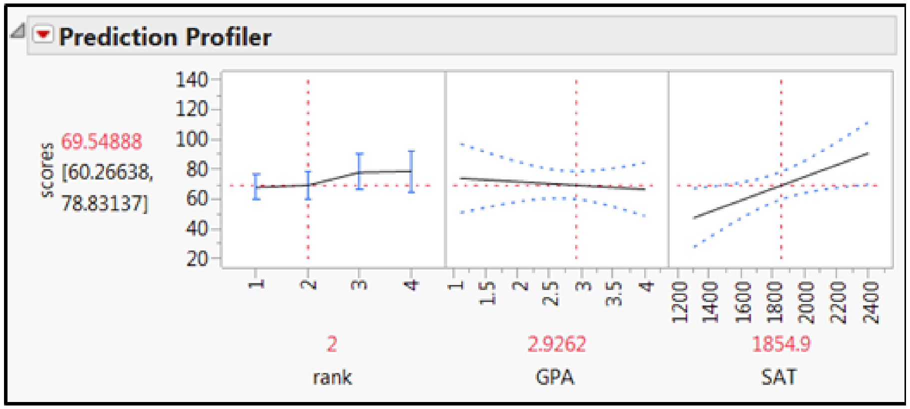
Showing all data

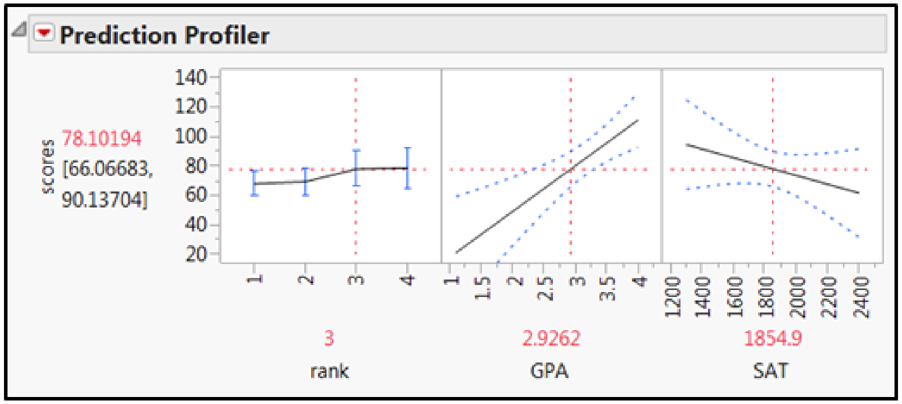


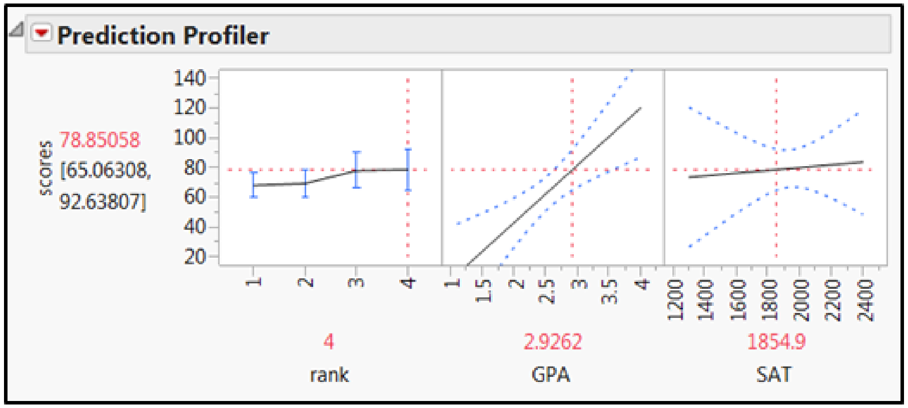
Showing GM: 10% from US, 20% from Europe, and 70% from Asia.

Prediction Profiler

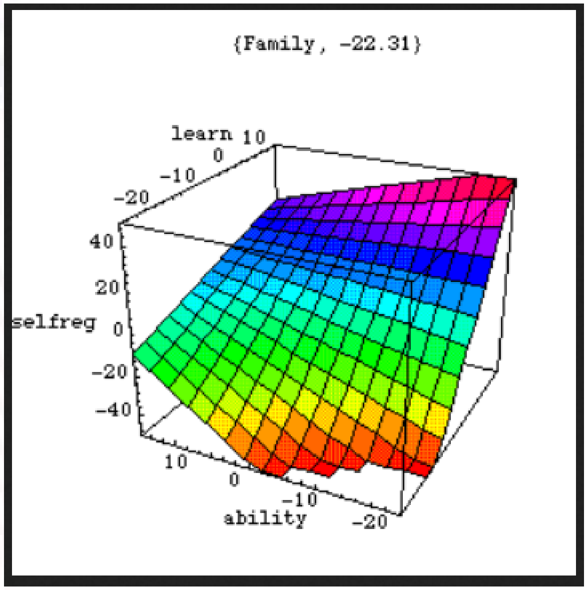


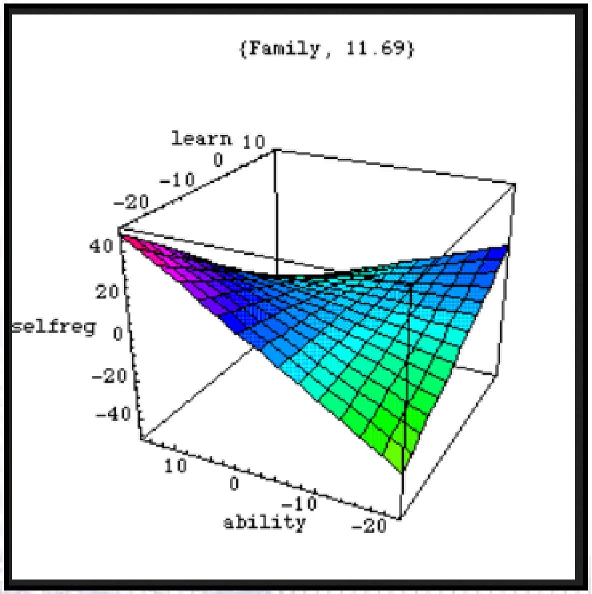






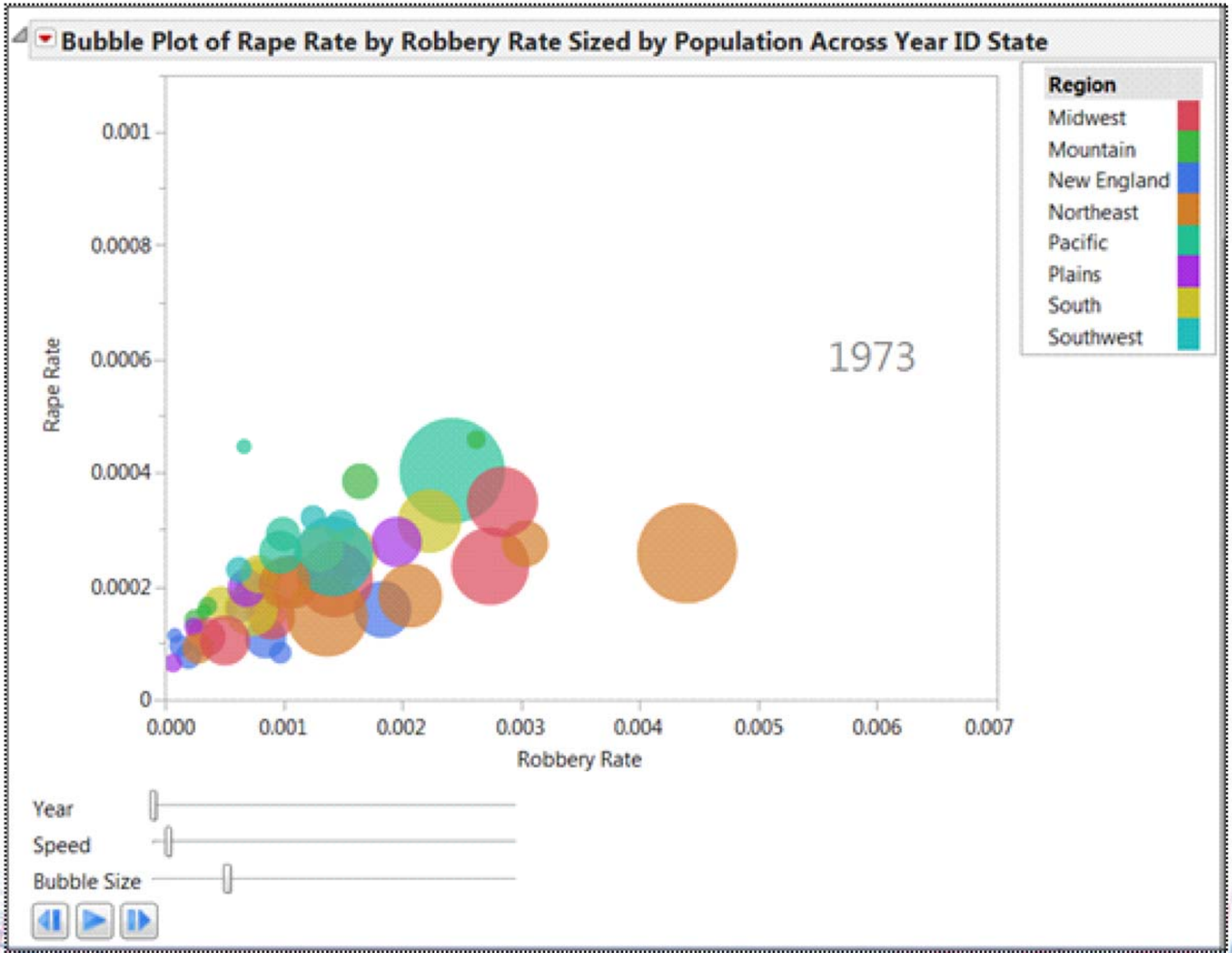
Dancing with three-way interaction

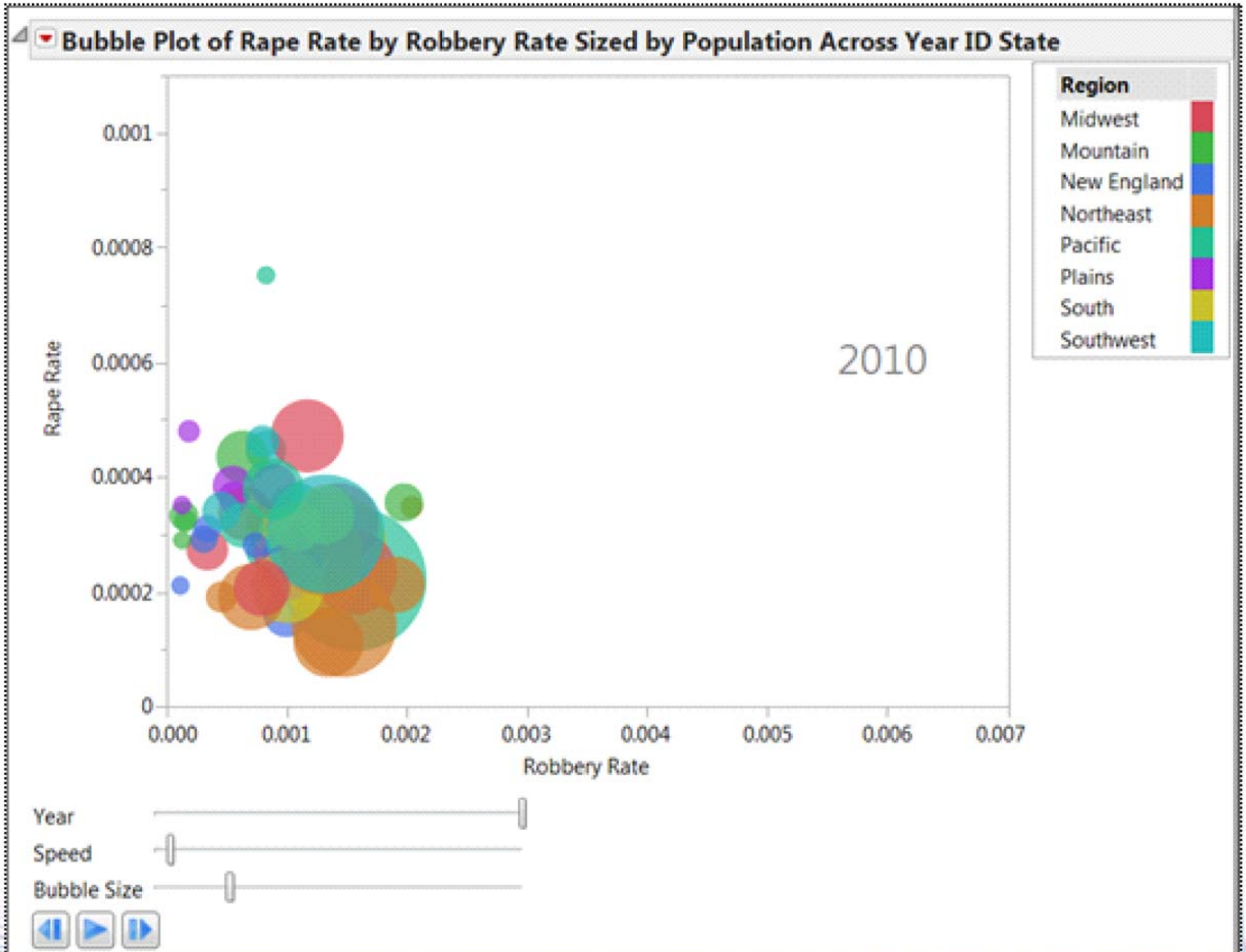




- Detecting and interpreting three-way interactions in regression may be very complicated. Using a mesh surface is much clearer.
- Interaction: the effect of X on Y is not consistent across all levels of A and B → regression lines vary
- If there is NO interaction, there should be no curving or dancing in the movie. Every frame should look the same.

How about five dimensions? Bubble plot



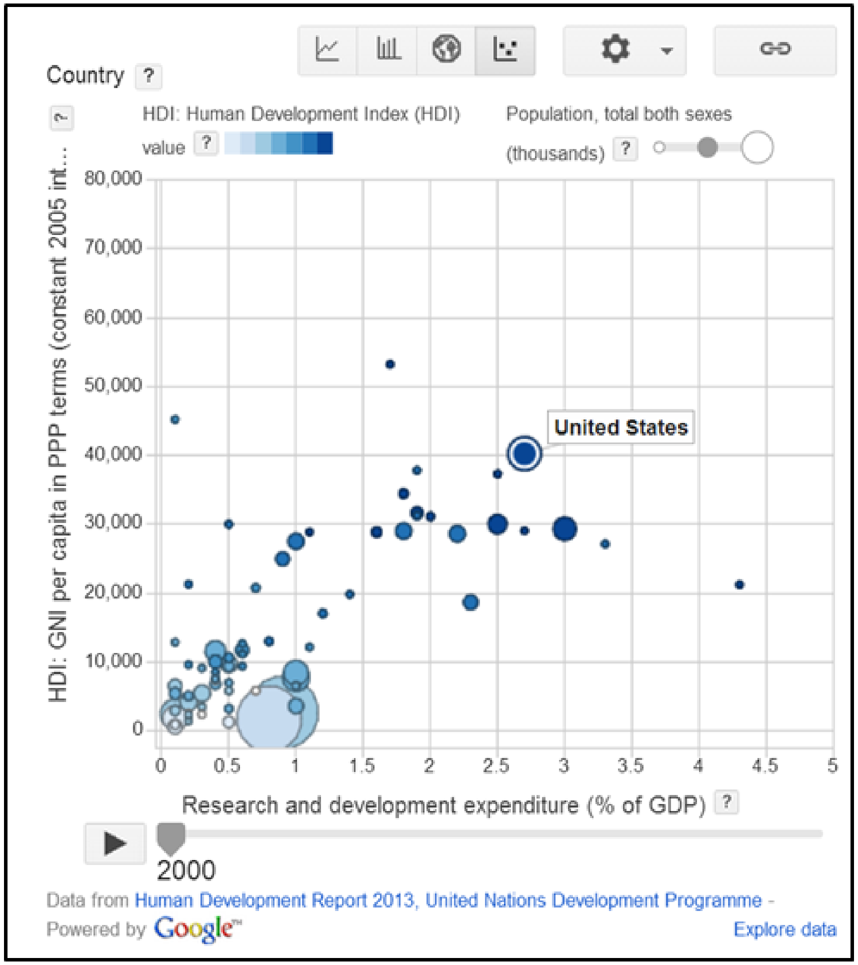


Bubble Plot of Rape Rate by Robbery Rate Sized by Population Across Year ID State
Region
Midwest
Mountain
New England
Northeast
Pacific
Plains
South
Southwest
2010
Rape Rate
0.001
0.0008
0.0006
0.0004
0.0002
0
0.000
0.001
0.002
0.003
0.004
0.005
0.006
0.007
Robbery Rate
Year
Speed
Bubble Size

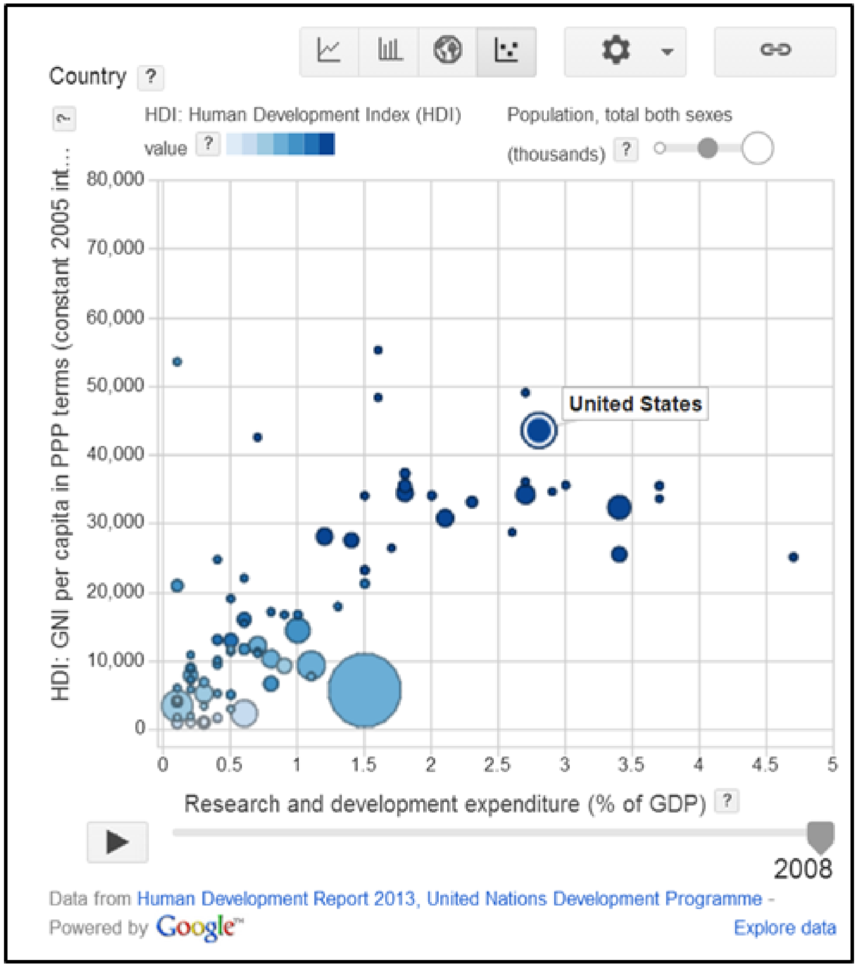
What the bubble dance tell you?

- In 1973 a strong association was found between the two crime rates, but 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.
- Big cities and Northeast are no longer the most dangerous places to live.

UN Public Data Explorer

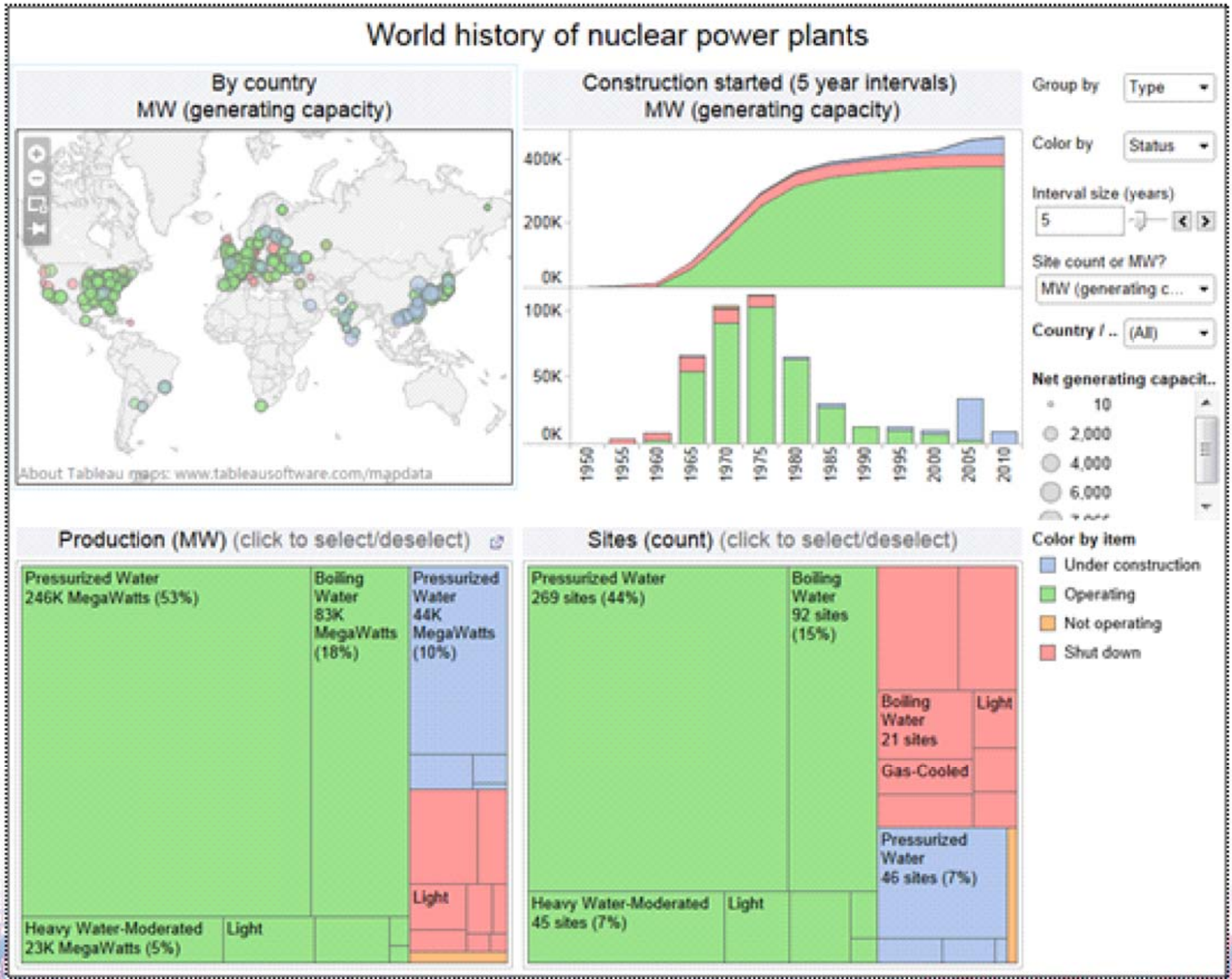


a. HDI data in 2000

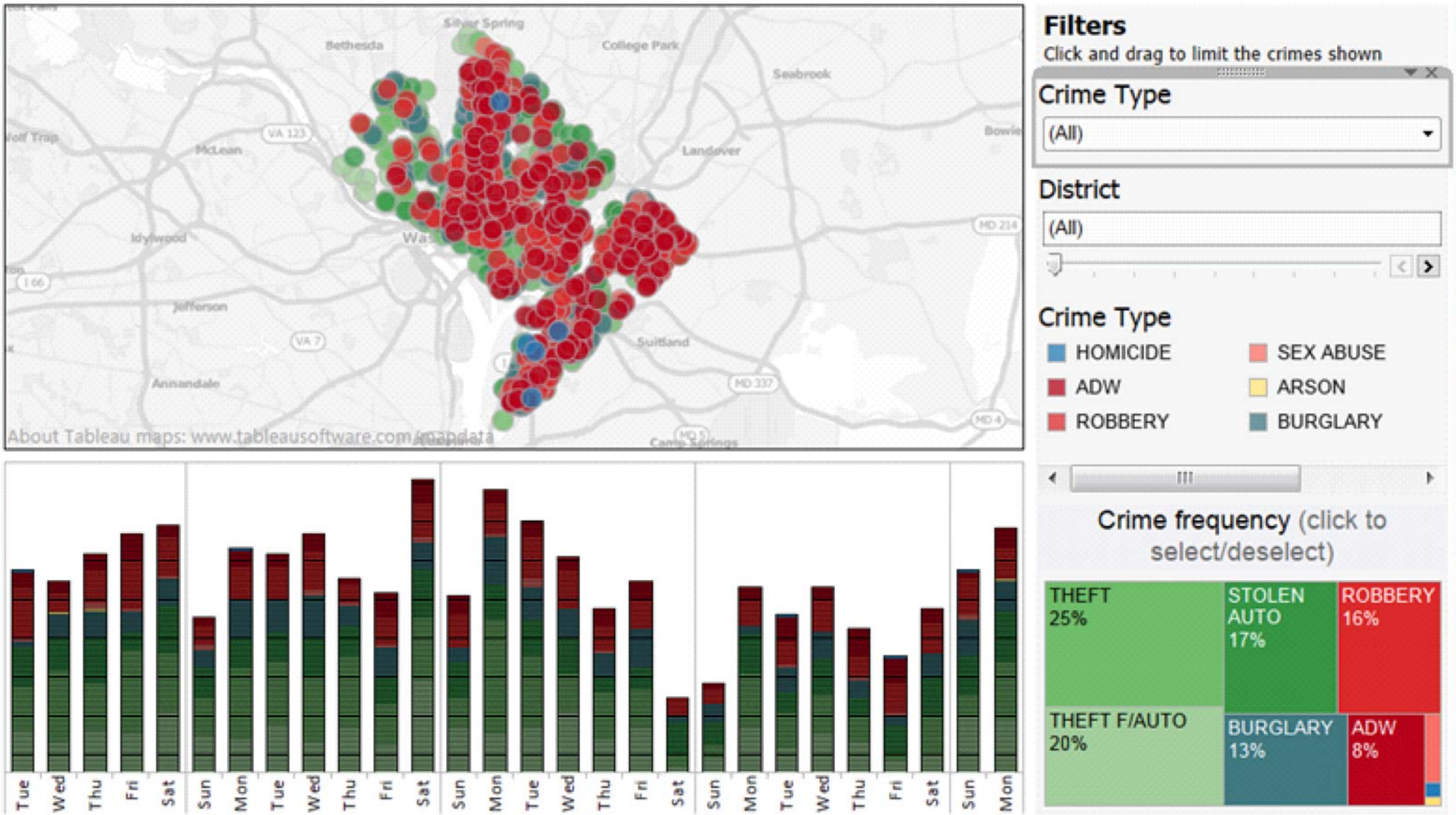


b. HDI in 2008

Tableau: Multi-panel visualization



District of Columbia Crimespotting



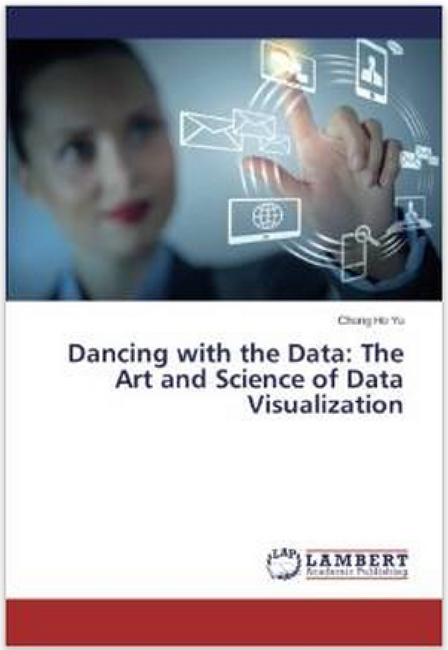
The contents are based upon

amazon Try Prime

Chong's Amazon.com | Today's Deals | Gift Cards | Sell | Help

Shop by **Department** | **Search** | All | dancing with the data | Go

Books | Advanced Search | New Releases | Best Sellers | The New York Times® Best Sellers | Children's Books | Textbooks | Textbook Rentals | Sell Us Your Books | Best Books of the Month



See this image

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

by Chong Ho Yu (Author)

Be the first to review this item

Paperback
$83.60

2 New from $83.60

Revealing the mysteries of a data set can be quite challenging. Data analysis is about discovering hidden patterns within the data and exploring the plausible "stories" that can explain those patterns. This book is about employing the art and science of data visualization. Beginning with an overview of what data visualization is and is not, the author introduces a plethora of graphing techniques that can be well-applied to both exploratory and confirmatory data analyses. Organized with a logical, sequential flow, the chapters move from one dimensional to multidimensional visualization examples, that include a rationale for the usefulness when applying them to specific data types. The overarching message to the reader is that, just like the appropriateness of a particular statistical test is tied to the research goal and the data structure, proper data visualization should align with data dimensionality and research objectives. While many texts that are devoted to data analysis focus on hypothesis testing or confirmatory data analysis, this book is a needed guide for creating meaningful and interpretable displays to depict data.

Read less

amazon Try Prime

Chong's Amazon.com | Today's Deals | Gift Cards | Sell | Help

Shop by **Department** | **Search** | All | Chong Ho Yu

Books | Advanced Search | New Releases | Best Sellers | The New York Times® Best Sellers | Children's Books | Textbooks | Textbook Rentals | Sell Us Your Books | E

Click to open expanded view

Share your own customer images

Publisher: learn how customers can search inside this book.

Philosophical Foundations of Quantitative Research Methodology [Paperback]

Chong Ho Yu (Author)

Be the first to review this item

List Price: ~~$38.99~~

Price: $29.95 & **FREE Shipping** on orders over $35. Details

You Save: $9.04 (23%)

In Stock.

Ships from and sold by **Amazon.com**. Gift-wrap available.

Want it tomorrow, July 15 to 91702? Order within **21 hrs 19 mins** and choose **Same-Day Delivery** at checkout.

17 new from $29.92 **10 used** from $26.85