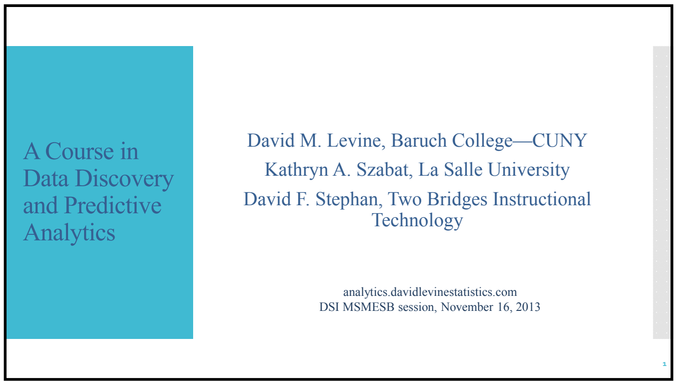


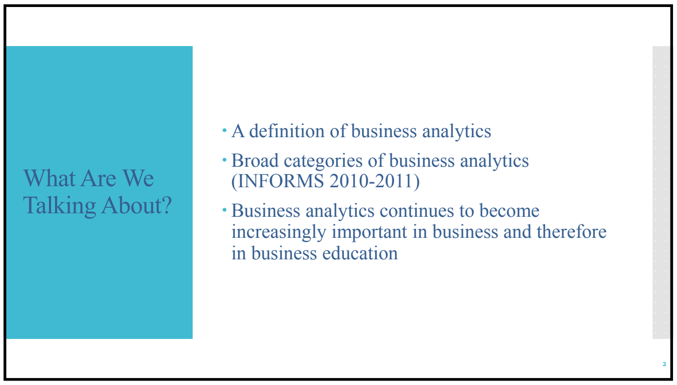
A Course in Data Discovery and Predictive Analytics

David M. Levine, Baruch College—CUNY

Kathryn A. Szabat, La Salle University

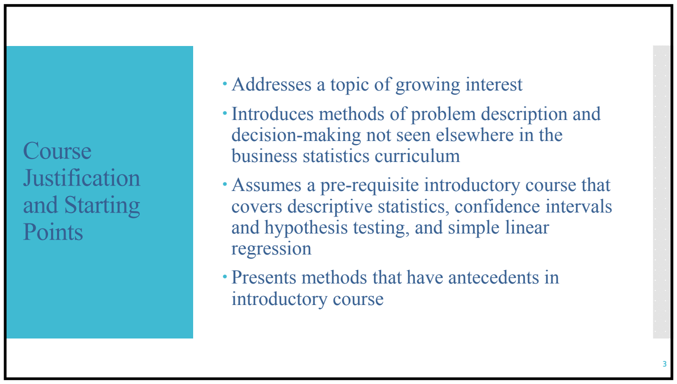
David F. Stephan, Two Bridges Instructional Technology

analytics.davidlevinestatistics.com
DSI MSMESB session, November 16, 2013



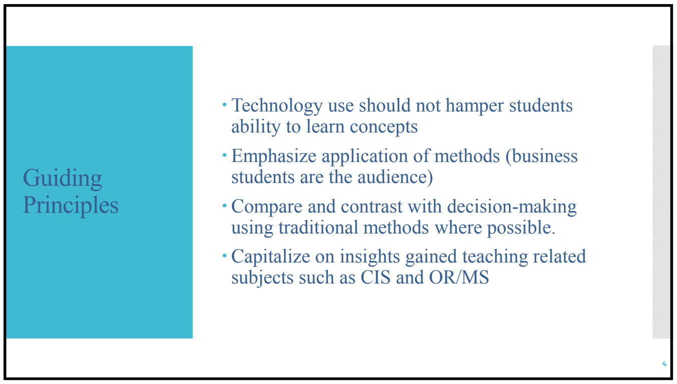
What Are We Talking About?

- A definition of business analytics
- Broad categories of business analytics (INFORMS 2010-2011)
- Business analytics continues to become increasingly important in business and therefore in business education



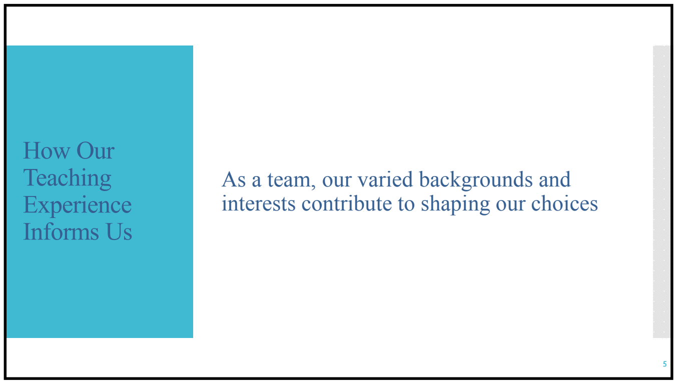
Course Justification and Starting Points

- Addresses a topic of growing interest
- Introduces methods of problem description and decision-making not seen elsewhere in the business statistics curriculum
- Assumes a pre-requisite introductory course that covers descriptive statistics, confidence intervals and hypothesis testing, and simple linear regression
- Presents methods that have antecedents in introductory course



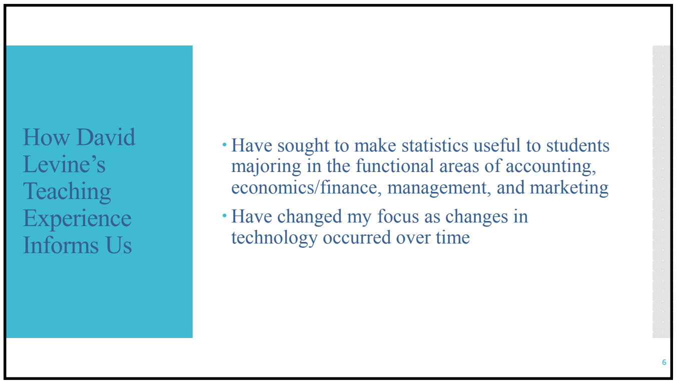
Guiding Principles

- Technology use should not hamper students ability to learn concepts
- Emphasize application of methods (business students are the audience)
- Compare and contrast with decision-making using traditional methods where possible.
- Capitalize on insights gained teaching related subjects such as CIS and OR/MS



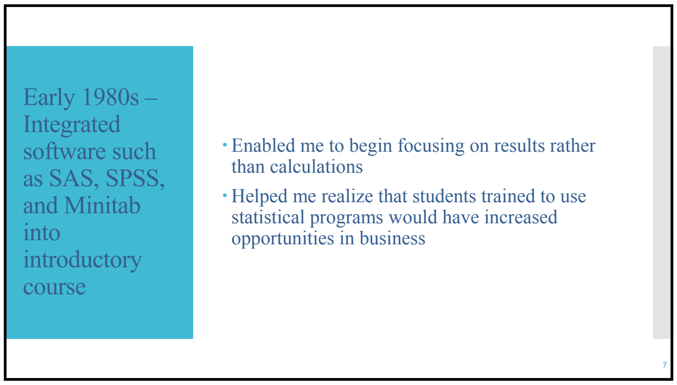
How Our Teaching Experience Informs Us

As a team, our varied backgrounds and interests contribute to shaping our choices



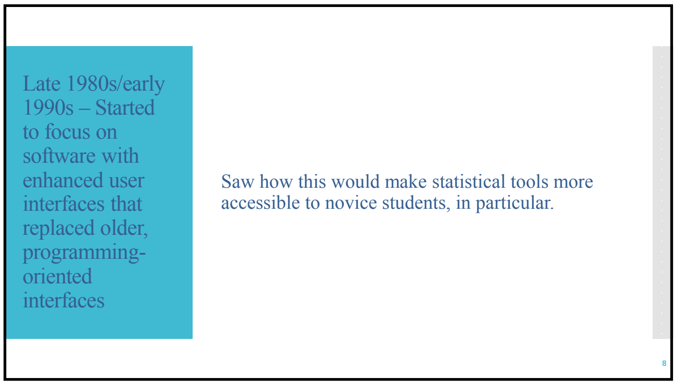
How David Levine’s Teaching Experience Informs Us

- Have sought to make statistics useful to students majoring in the functional areas of accounting, economics/finance, management, and marketing
- Have changed my focus as changes in technology occurred over time



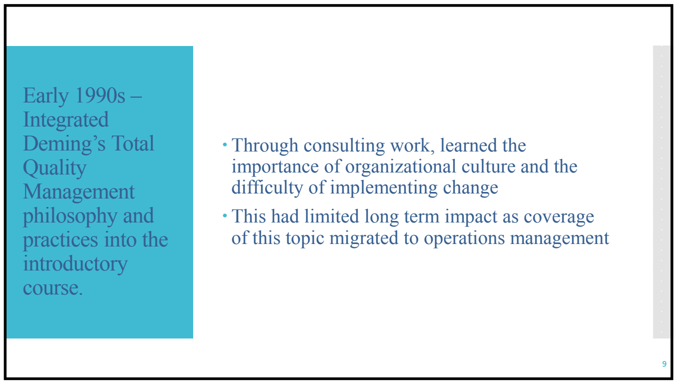
Early 1980s – Integrated software such as SAS, SPSS, and Minitab into introductory course

- Enabled me to begin focusing on results rather than calculations
- Helped me realize that students trained to use statistical programs would have increased opportunities in business



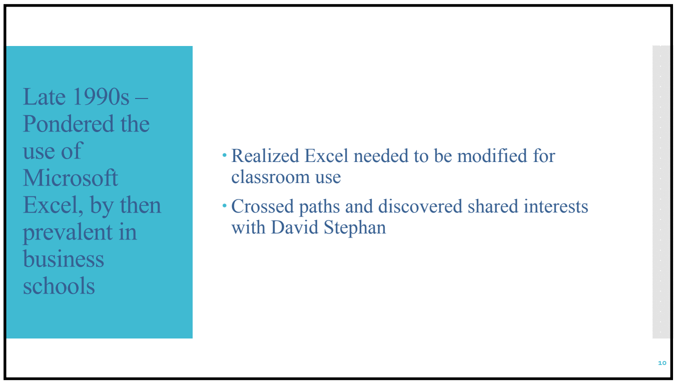
Late 1980s/early 1990s – Started to focus on software with enhanced user interfaces that replaced older, programming-oriented interfaces

Saw how this would make statistical tools more accessible to novice students, in particular.



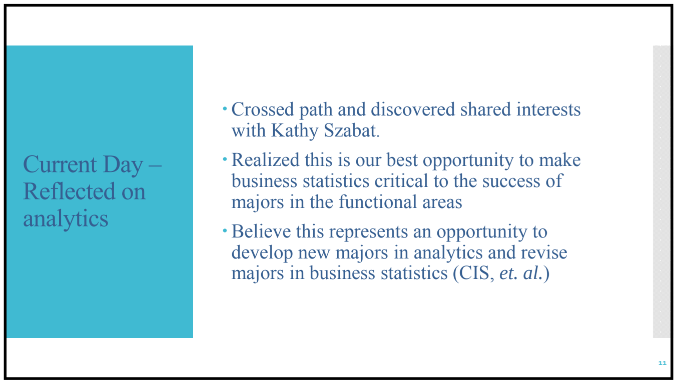
Early 1990s – Integrated Deming's Total Quality Management philosophy and practices into the introductory course.

- Through consulting work, learned the importance of organizational culture and the difficulty of implementing change
- This had limited long term impact as coverage of this topic migrated to operations management



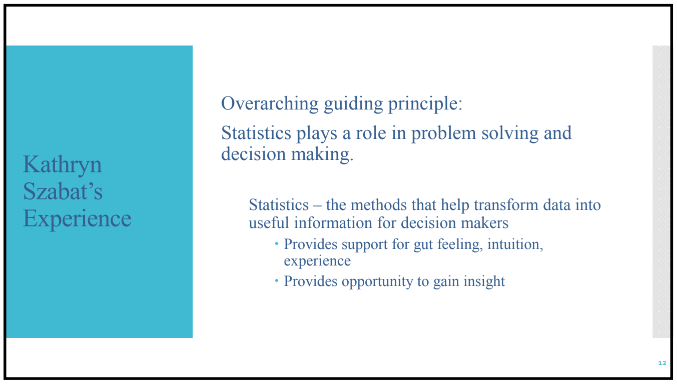
Late 1990s – Pondered the use of Microsoft Excel, by then prevalent in business schools

- Realized Excel needed to be modified for classroom use
- Crossed paths and discovered shared interests with David Stephan



Current Day – Reflected on analytics

- Crossed path and discovered shared interests with Kathy Szabat.
- Realized this is our best opportunity to make business statistics critical to the success of majors in the functional areas
- Believe this represents an opportunity to develop new majors in analytics and revise majors in business statistics (CIS, *et. al.*)



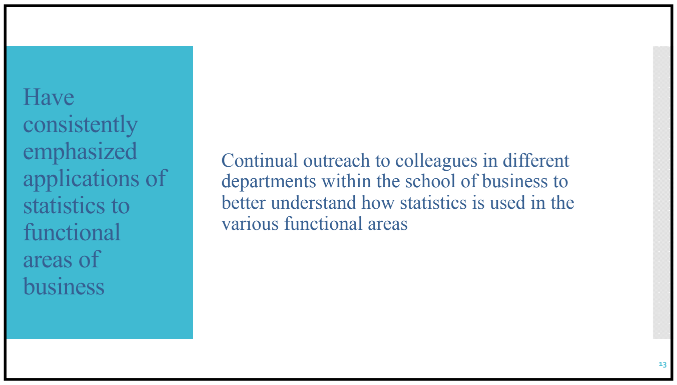
Kathryn Szabat's Experience

Overarching guiding principle:

Statistics plays a role in problem solving and decision making.

Statistics – the methods that help transform data into useful information for decision makers

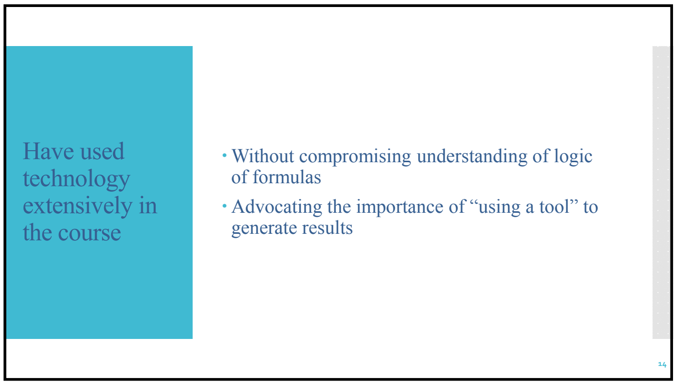
- Provides support for gut feeling, intuition, experience
- Provides opportunity to gain insight



Have consistently emphasized applications of statistics to functional areas of business

Continual outreach to colleagues in different departments within the school of business to better understand how statistics is used in the various functional areas

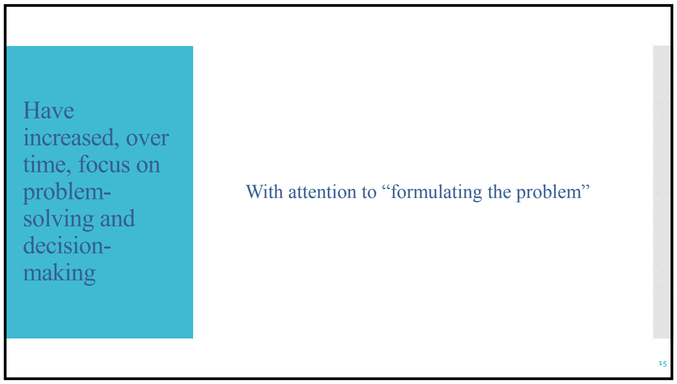
13



Have used technology extensively in the course

- Without compromising understanding of logic of formulas
- Advocating the importance of "using a tool" to generate results

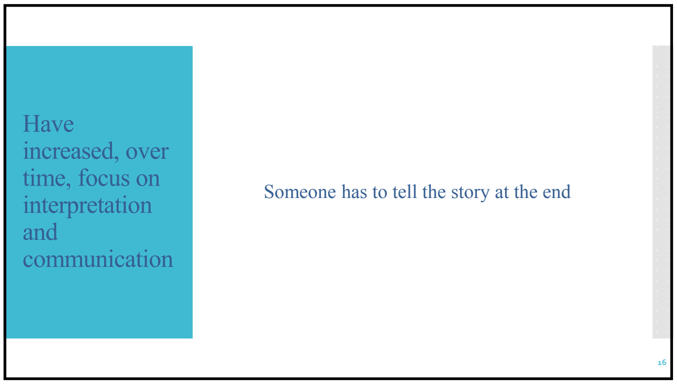
14



Have increased, over time, focus on problem-solving and decision-making

With attention to "formulating the problem"

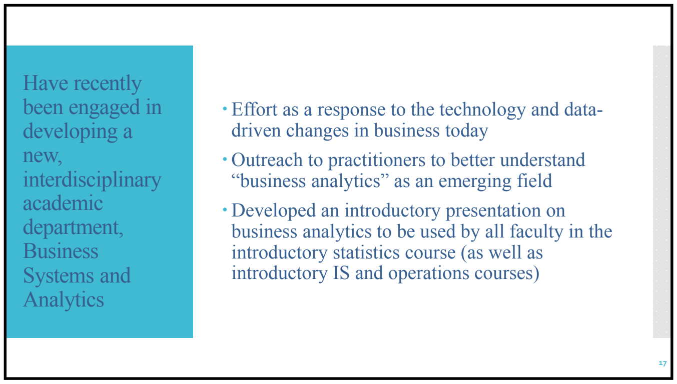
15



Have increased, over time, focus on interpretation and communication

Someone has to tell the story at the end

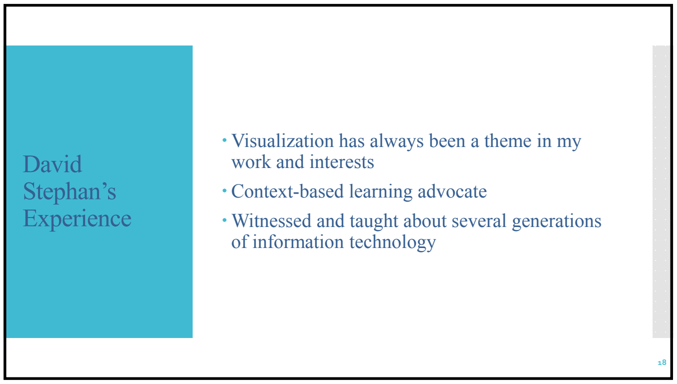
16



Have recently been engaged in developing a new, interdisciplinary academic department, Business Systems and Analytics

- Effort as a response to the technology and data-driven changes in business today
- Outreach to practitioners to better understand "business analytics" as an emerging field
- Developed an introductory presentation on business analytics to be used by all faculty in the introductory statistics course (as well as introductory IS and operations courses)

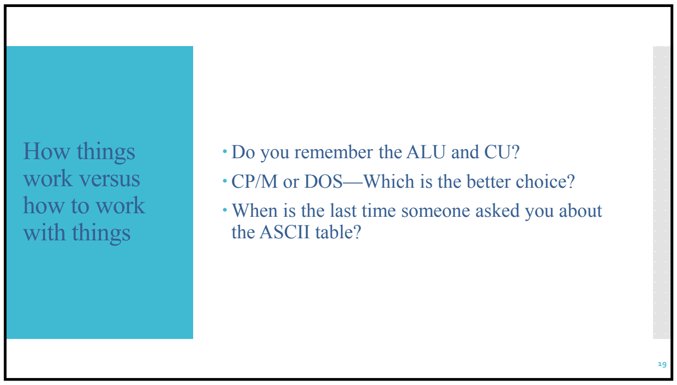
17



David Stephan's Experience

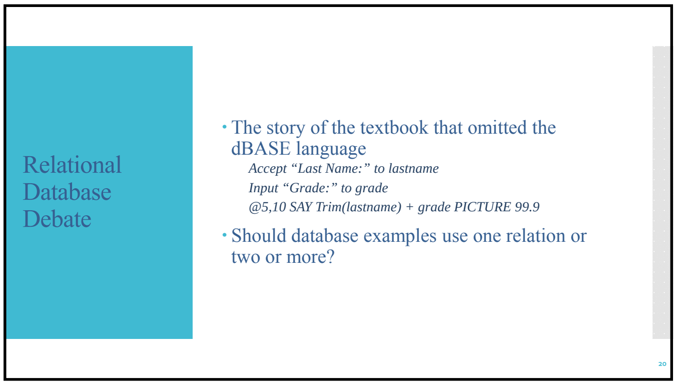
- Visualization has always been a theme in my work and interests
- Context-based learning advocate
- Witnessed and taught about several generations of information technology

18



How things work versus how to work with things

- Do you remember the ALU and CU?
- CP/M or DOS—Which is the better choice?
- When is the last time someone asked you about the ASCII table?



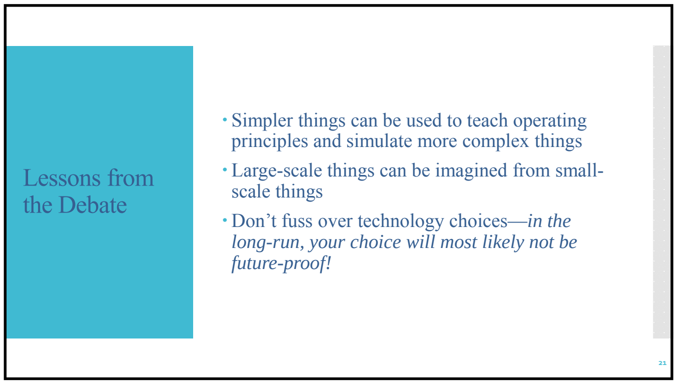
Relational Database Debate

- The story of the textbook that omitted the dBASE language

  Accept "Last Name:" to lastname

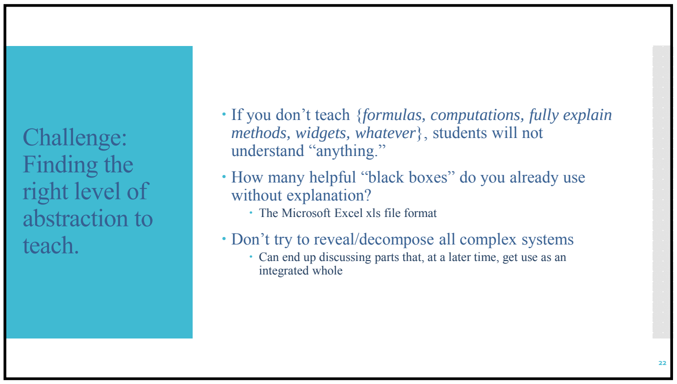
  Input "Grade:" to grade

  @5,10 SAY Trim(lastname) + grade PICTURE 99.9
- Should database examples use one relation or two or more?



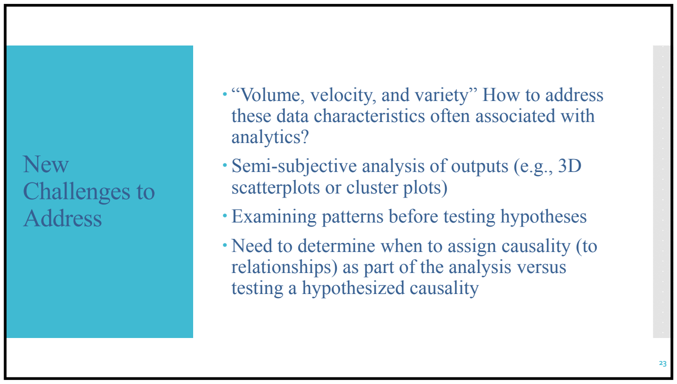
Lessons from the Debate

- Simpler things can be used to teach operating principles and simulate more complex things
- Large-scale things can be imagined from small-scale things
- Don't fuss over technology choices—*in the long-run, your choice will most likely not be future-proof!*



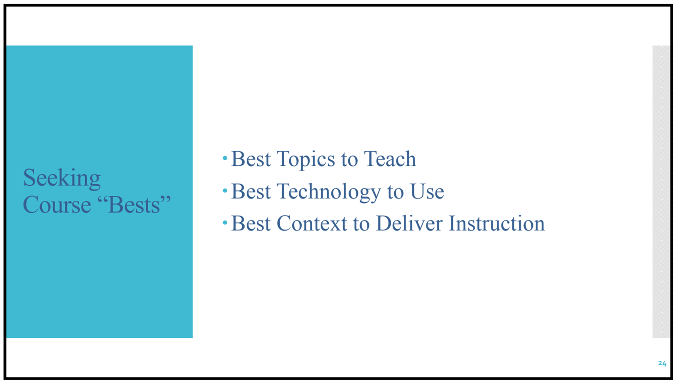
Challenge: Finding the right level of abstraction to teach.

- If you don't teach {*formulas, computations, fully explain methods, widgets, whatever*}, students will not understand "anything."
- How many helpful "black boxes" do you already use without explanation?
  - The Microsoft Excel xls file format
- Don't try to reveal/decompose all complex systems
  - Can end up discussing parts that, at a later time, get use as an integrated whole



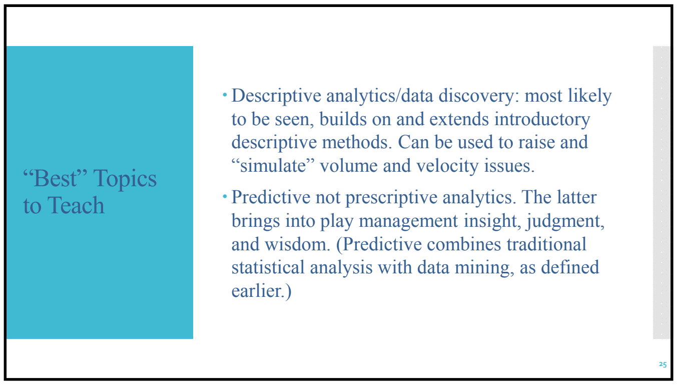
New Challenges to Address

- "Volume, velocity, and variety" How to address these data characteristics often associated with analytics?
- Semi-subjective analysis of outputs (e.g., 3D scatterplots or cluster plots)
- Examining patterns before testing hypotheses
- Need to determine when to assign causality (to relationships) as part of the analysis versus testing a hypothesized causality



Seeking Course "Bests"

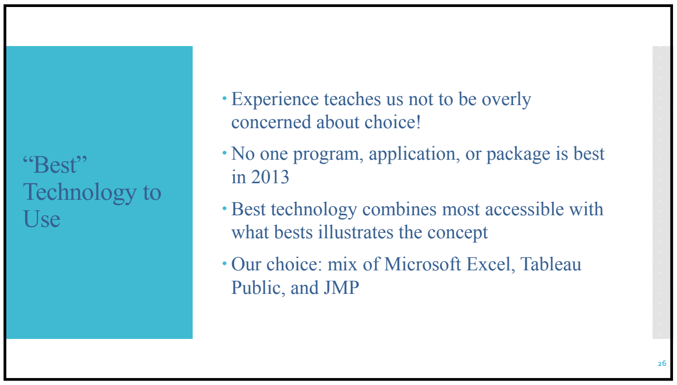
- Best Topics to Teach
- Best Technology to Use
- Best Context to Deliver Instruction



"Best" Topics to Teach

- Descriptive analytics/data discovery: most likely to be seen, builds on and extends introductory descriptive methods. Can be used to raise and "simulate" volume and velocity issues.
- Predictive not prescriptive analytics. The latter brings into play management insight, judgment, and wisdom. (Predictive combines traditional statistical analysis with data mining, as defined earlier.)

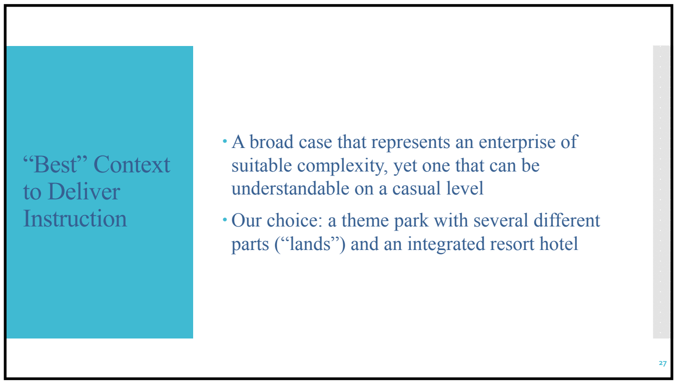
25



"Best" Technology to Use

- Experience teaches us not to be overly concerned about choice!
- No one program, application, or package is best in 2013
- Best technology combines most accessible with what bests illustrates the concept
- Our choice: mix of Microsoft Excel, Tableau Public, and JMP

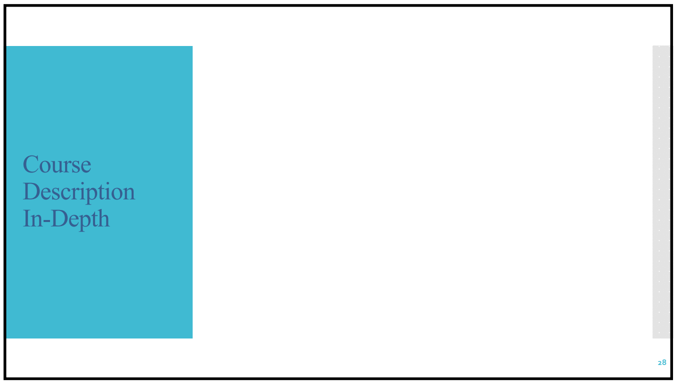
26



"Best" Context to Deliver Instruction

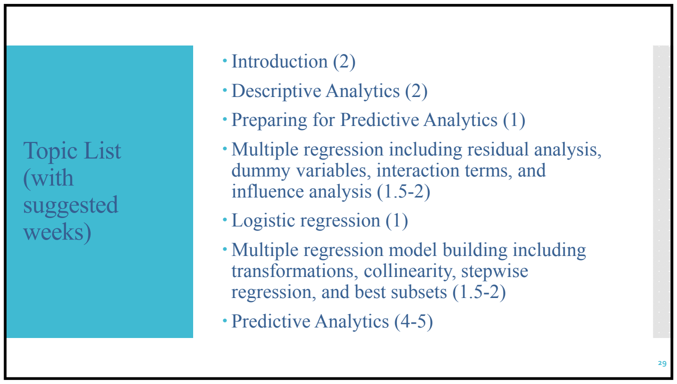
- A broad case that represents an enterprise of suitable complexity, yet one that can be understandable on a casual level
- Our choice: a theme park with several different parts ("lands") and an integrated resort hotel

27



Course Description In-Depth

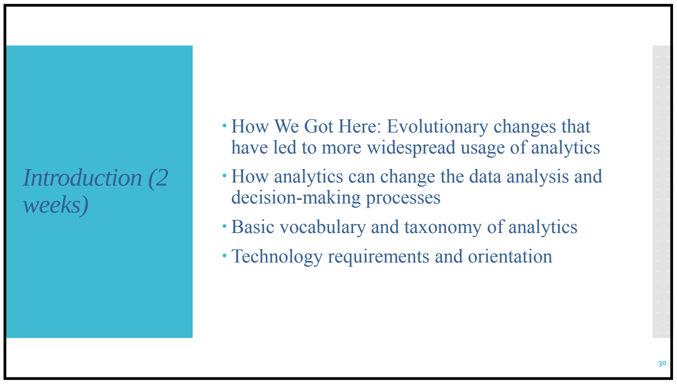
28



Topic List (with suggested weeks)

- Introduction (2)
- Descriptive Analytics (2)
- Preparing for Predictive Analytics (1)
- Multiple regression including residual analysis, dummy variables, interaction terms, and influence analysis (1.5-2)
- Logistic regression (1)
- Multiple regression model building including transformations, collinearity, stepwise regression, and best subsets (1.5-2)
- Predictive Analytics (4-5)

29



Introduction (2 weeks)

- How We Got Here: Evolutionary changes that have led to more widespread usage of analytics
- How analytics can change the data analysis and decision-making processes
- Basic vocabulary and taxonomy of analytics
- Technology requirements and orientation

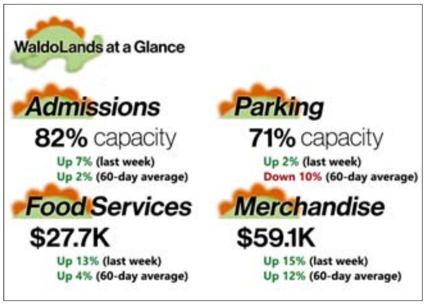
30

Descriptive Analytics (2 weeks)

- Summarizing volume and velocity
- “Sexiness” versus usefulness issue
- Levels of summary: drill down, levels of hierarchy, and subsetting
- Information design principles that inform descriptive methods

Summarizing volume and velocity: Dashboards

Provide information about the current status of a business or business activity in a form easy to comprehend and review.

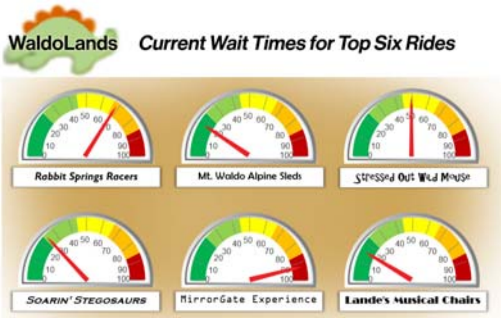


Sexiness versus usefulness: Gauges vs. bullet graphs

Example: combining a numerical measure with a categorical group

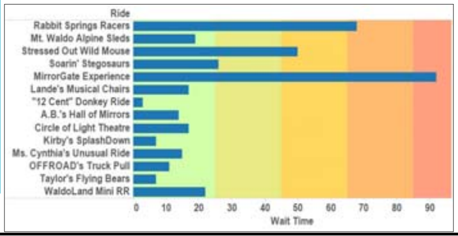
- Which one looks more “sexy,” appealing, interesting, *etc.*?
- Which one best facilitates comparisons?
- What if the answers to the two questions are different?

Sexiness versus usefulness: Gauges vs. bullet graphs



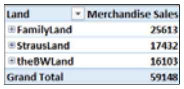
Sexiness versus usefulness: Gauges vs. bullet graphs

- Which one looks more “sexy,” appealing, interesting, *etc*.?
- Which one best facilitates comparisons?
- What if the answers to the two questions are different?

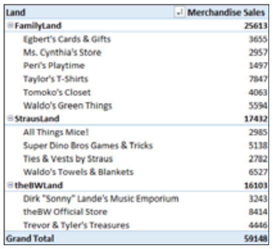


Levels of summary: drill down, levels of hierarchy, and subsetting

Drill-down sequence example (using Excel)



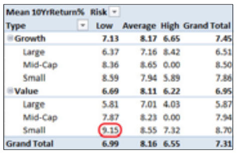
Land	Merchandise Sales
⊞ FamilyLand	25613
⊞ StrausLand	17432
⊞ theBWLand	16103
Grand Total	59148



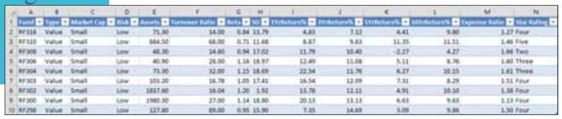
Land	Merchandise Sales
⊟ FamilyLand	25613
Egbert's Cards & Gifts	3655
Ms. Cynthia's Store	2957
Peri's Playtime	1497
Taylor's T-Shirts	7847
Tomoko's Closet	4063
Waldo's Green Things	5594
⊟ StrausLand	17432
All Things Mice!	2985
Super Dino Bros Games & Tricks	5138
Ties & Vests by Straus	2782
Waldo's Towels & Blankets	6527
⊟ theBWLand	16103
Dirk "Sonny" Lande's Music Emporium	3243
theBW Official Store	8414
Trevor & Tyler's Treasures	4446
Grand Total	59148

Levels of summary: drill down, levels of hierarchy, and subsetting

Financial example showing another level of drill-down



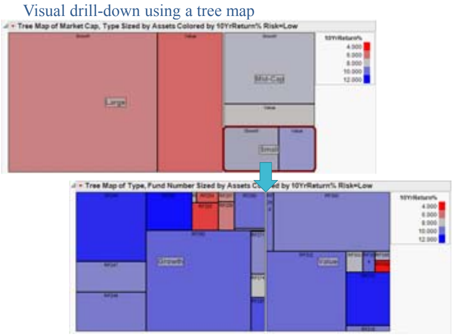
Mean 10YrReturn%	Risk			
Type	Low	Average	High	Grand Total
Growth	**7.13**	**8.17**	**6.65**	**7.45**
Large	6.37	7.16	8.42	6.51
Mid-Cap	8.36	8.65	0.00	8.50
Small	8.59	7.94	5.89	7.86
Value	**6.69**	**8.11**	**6.22**	**6.95**
Large	5.81	7.01	4.03	5.87
Mid-Cap	7.87	8.23	0.00	7.94
Small	9.15	8.55	7.32	8.70
Grand Total	**6.99**	**8.16**	**6.55**	**7.31**



	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	Fund	Type	Market Cap	Risk	Assets	Turnover Ratio	Beta	SD	1YrReturn%	3YrReturn%	5YrReturn%	10YrReturn%	Expense Ratio	Star Rating
2	RF316	Value	Small	Low	71.30	14.00	0.84	13.79	4.83	7.12	4.41	9.80	1.27	Four
3	RF310	Value	Small	Low	664.50	68.00	0.71	11.68	8.87	9.63	11.35	11.51	1.46	Five
4	RF308	Value	Small	Low	48.30	14.60	0.94	17.02	11.79	10.40	-2.27	4.27	1.66	Two
5	RF306	Value	Small	Low	40.90	28.00	1.16	18.97	12.49	11.08	5.11	8.76	1.60	Three
6	RF304	Value	Small	Low	73.30	32.00	1.15	18.69	22.54	11.76	6.27	10.15	1.61	Three
7	RF303	Value	Small	Low	101.20	16.78	1.05	17.41	16.54	12.09	7.31	8.29	1.51	Four
8	RF302	Value	Small	Low	1837.60	16.04	1.20	1.92	13.78	12.11	4.91	10.10	1.38	Four
9	RF300	Value	Small	Low	1980.30	27.00	1.14	18.80	20.13	13.13	6.63	9.63	1.13	Four
10	RF298	Value	Small	Low	127.80	89.00	0.95	15.90	7.15	14.69	3.09	9.86	1.50	Four

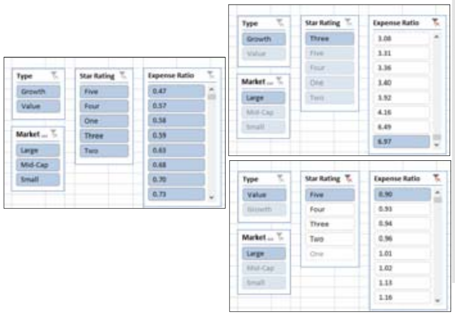
Levels of summary: drill down, levels of hierarchy, and subsetting

Visual drill-down using a tree map



Levels of summary: drill down, levels of hierarchy, and subsetting

Subsetting using "slicers" (Excel)



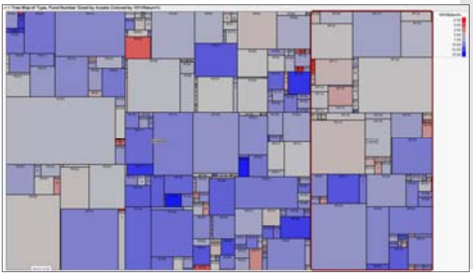
Information design principles

- Fostering efficient and effective communication and understanding
- Provide context for data in a compact presentation
- Add additional "dimensions" of data
- Misuse raises issues beyond "typical" statistical concerns: visual perception, artistic considerations

Does this tree map provide context for data in a compact presentation?

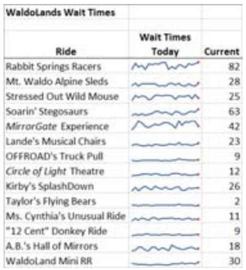
Add additional "dimensions" of data?

Tree Map of Retirement Fund Assets Colored by 10-Year Return Percentage, By Fund Type (JMP)

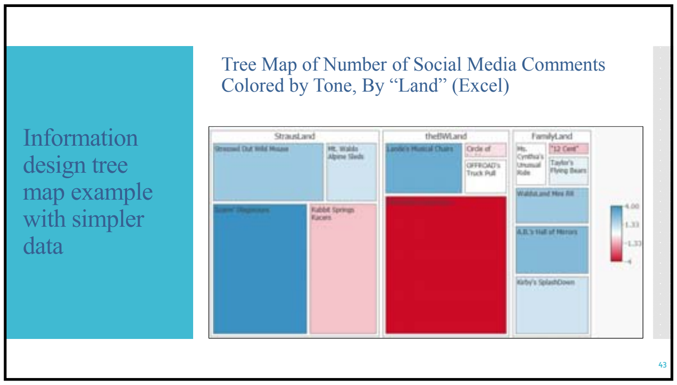


Does this table provide context for data in a compact presentation?

Sparklines example (Excel)



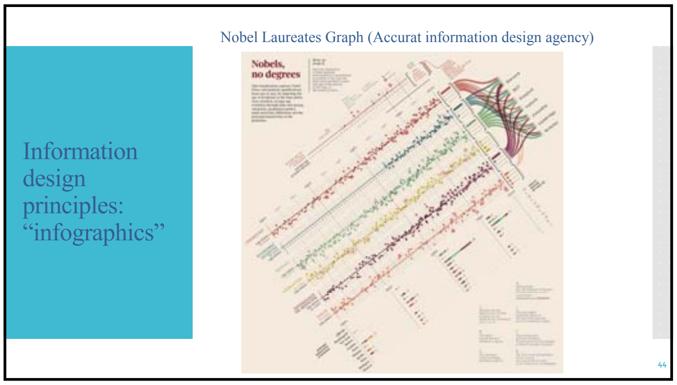
WaldoLands Wait Times		
Ride	Wait Times Today	Current
Rabbit Springs Racers		82
Mt. Waldo Alpine Sleds		28
Stressed Out Wild Mouse		25
Soarin' Stegosaurs		63
MirrorGate Experience		42
Lande's Musical Chairs		23
OFFROAD's Truck Pull		9
Circle of Light Theatre		12
Kirby's SplashDown		26
Taylor's Flying Bears		2
Ms. Cynthia's Unusual Ride		11
"12 Cent" Donkey Ride		9
A.B.'s Hall of Mirrors		18
WaldoLand Mini RR		30



Information design tree map example with simpler data

Tree Map of Number of Social Media Comments Colored by Tone, By "Land" (Excel)

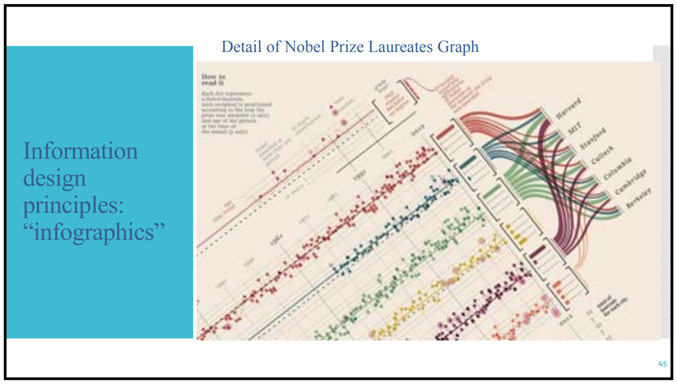
43



Information design principles: "infographics"

Nobel Laureates Graph (Accurat information design agency)

44



Information design principles: "infographics"

Detail of Nobel Prize Laureates Graph

45

Preparing for Predictive Analytics (1 week)

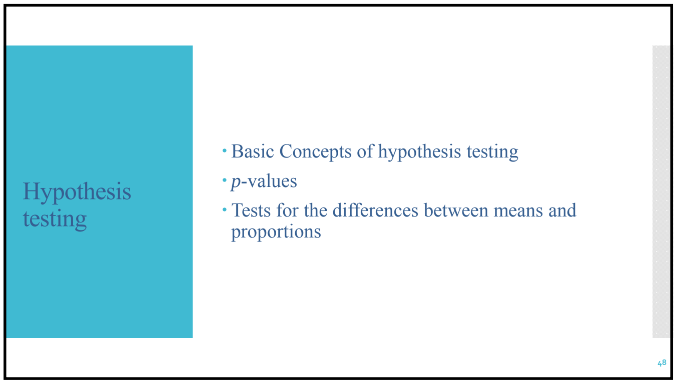
- Confidence intervals
- Hypothesis testing
- Simple linear regression

46

Confidence intervals

- Normal distribution
- Sampling distributions
- Confidence intervals for the mean and proportion

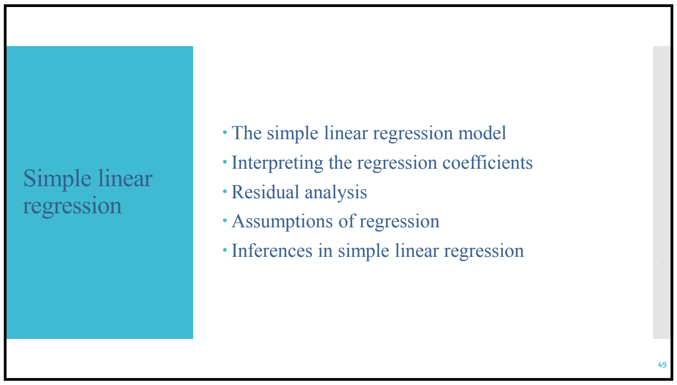
47



Hypothesis testing

- Basic Concepts of hypothesis testing
- *p*-values
- Tests for the differences between means and proportions

48



Simple linear regression

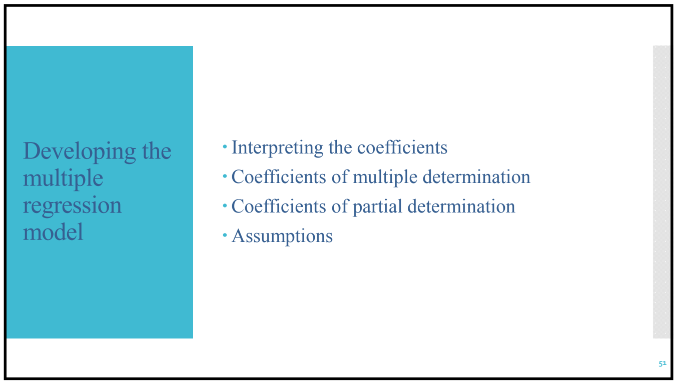
- The simple linear regression model
- Interpreting the regression coefficients
- Residual analysis
- Assumptions of regression
- Inferences in simple linear regression

49

Multiple Regression (1.5-2 weeks)

- Developing the multiple regression model
- Inference in multiple regression
- Residual analysis
- Dummy variables
- Interaction terms
- Influence analysis

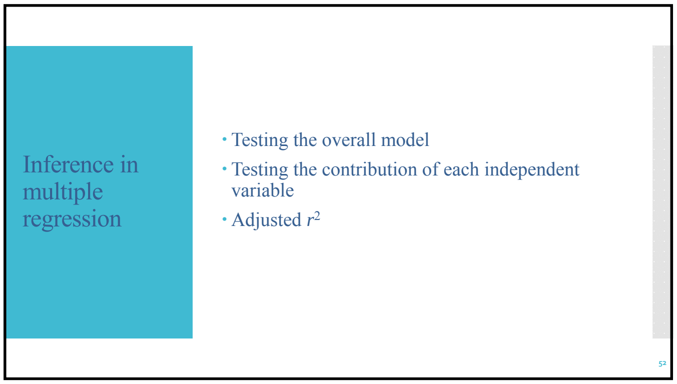
50



Developing the multiple regression model

- Interpreting the coefficients
- Coefficients of multiple determination
- Coefficients of partial determination
- Assumptions

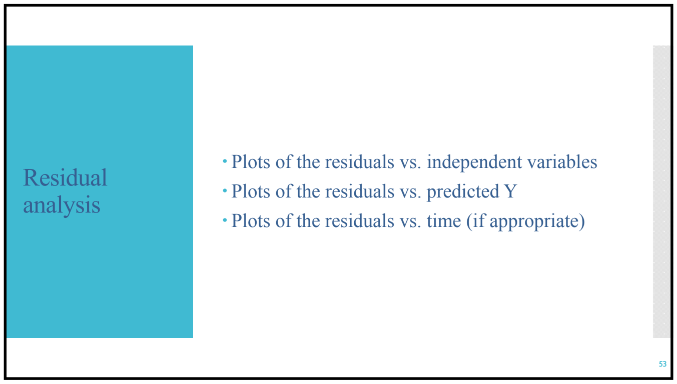
51



Inference in multiple regression

- Testing the overall model
- Testing the contribution of each independent variable
- Adjusted r^2

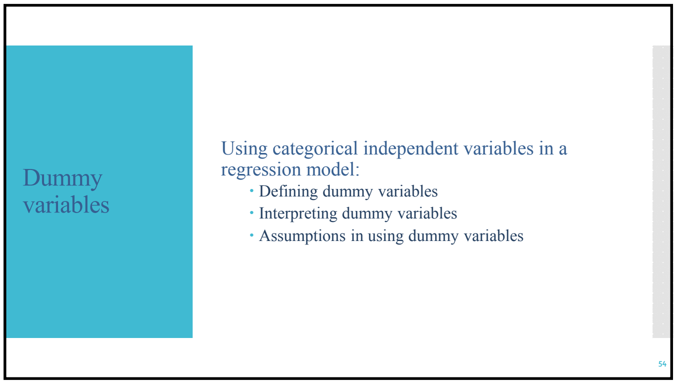
52



Residual analysis

- Plots of the residuals vs. independent variables
- Plots of the residuals vs. predicted Y
- Plots of the residuals vs. time (if appropriate)

53



Dummy variables

Using categorical independent variables in a regression model:

- Defining dummy variables
- Interpreting dummy variables
- Assumptions in using dummy variables

54

Interaction terms

- What they are
- Why they are sometimes necessary
- Interpreting interaction terms

Influence analysis

Examining the effect of individual observations on the regression model

- Hat matrix elements h_i
- Studentized deleted residuals t_i
- Cook's Distance statistic D_i

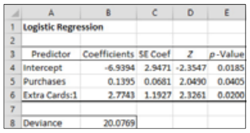
Logistic regression (1 week)

Predicting a categorical dependent variable

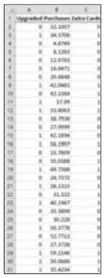
- Cannot use least squares regression
- Odds ratio
- Logistic regression model
- Predicting probability of an event of interest
- Deviance statistic
- Wald statistic

Logistic regression example using an Excel add-in

"Predicting the likelihood of upgrading to a premium credit card based on the monthly purchase amount and whether the account has multiple cards"



	A	B	C	D	E
1	**Logistic Regression**				
2					
3	Predictor	Coefficients	SE Coef	*Z*	*p*-Value
4	Intercept	-6.9394	2.9471	-2.3547	0.0185
5	Purchases	0.1395	0.0681	2.0490	0.0405
6	Extra Cards:1	2.7743	1.1927	2.3261	0.0200
7					
8	Deviance	20.0769			



	A	B	C
1	Upgraded	Purchases	Extra Cards
2	0	32.1007	0
3	1	34.3706	1
4	0	4.8749	0
5	0	8.1263	0
6	0	12.9783	0
7	0	16.0471	0
8	0	20.6648	0
9	1	42.0483	1
10	0	42.2264	1
11	1	37.99	1
12	1	53.6063	1
13	0	38.7936	0
14	0	27.9999	0
15	1	42.1694	0
16	1	56.1997	1
17	0	23.7609	0
18	0	35.0388	1
19	1	49.7388	1
20	0	24.7372	0
21	1	26.1023	1
22	0	31.322	1
23	1	40.1967	1
24	0	35.3899	0
25	0	30.228	0
26	1	50.3778	0
27	0	52.7713	0
28	0	27.3728	0
29	1	59.2346	1
30	1	50.0686	1
31	1	53.4234	1

Multiple Regression Model Building (1.5-2 weeks)

- Transformations
- Collinearity
- Stepwise regression
- Best subsets regression

Transformations

- Purposes
- Square root transformations
- Logarithmic transformations

Collinearity

- Effect on the regression model
- Measuring the variance inflationary factor (*VIF*)
- Dealing with collinear independent variables

Stepwise regression

- History
- How it works
- Limitations
- Use in an era of big data

Best subsets regression

- How it works
- Advantages and disadvantages vs. stepwise regression
- Mallows C_p statistic

Predictive Analytics (4-5 weeks)

	METHOD FOR			
METHOD	Prediction	Classification	Clustering	Association
Classification and regression trees (1-1.5 weeks)	●	●		
Neural networks (1-1.5 weeks)	●	●	●	
Cluster analysis (1 week)			●	
Multidimensional scaling (1week)		●		●

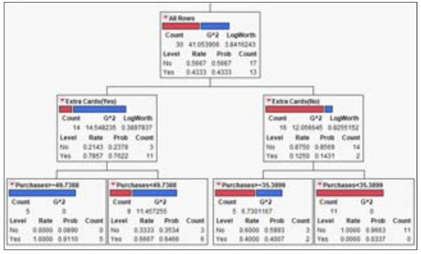
Classification and regression trees

Decision trees that split data into groups based on the values of independent or explanatory (*X*) variables.

- Not affected by the distribution of the variables
- Splitting determines which values of a specific independent variable are useful in predicting the dependent (*Y*) variable present
- Using a *categorical* dependent *Y* variable results in a *classification tree*
- Using a *numerical* dependent *Y* variable results in a *regression tree*
- Rules for splitting the tree
- Pruning back a tree
- If possible, divide data into training sample and validation sample

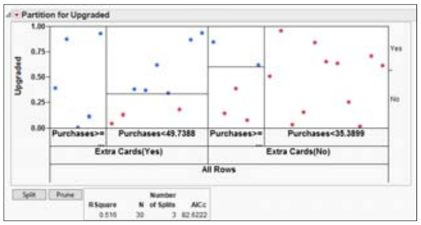
Classification tree example

"Predicting the likelihood of upgrading to a premium credit card based on the monthly purchase amount and whether the account has multiple cards" (same example used in logistic regression)



Classification tree example

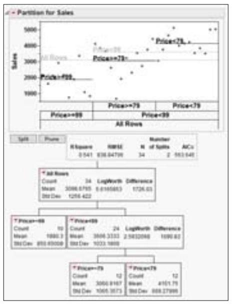
"Predicting the likelihood of upgrading to a premium credit card based on the monthly purchase amount and whether the account has multiple cards" (same example used in logistic regression)



RSquare	N	Number of Splits	AICc
0.516	30	3	32.6222

Regression tree example

"Predicting sales of energy bars based on price and promotion expenses" (could be multiple regression example, too)



RSquare	RMSE	N	Number of Splits	AICc
0.541	838.84706	34	2	563.646

All Rows

Count	34	LogWorth	Difference
Mean	3098.6765	6.6165883	1726.03
Std Dev	1256.422		

Price>=99

Count	10
Mean	1880.3
Std Dev	850.65008

Price<99

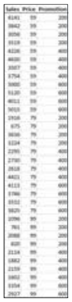
Count	24	LogWorth	Difference
Mean	3606.3333	2.5832096	1090.83
Std Dev	1033.1669		

Price>=79

Count	12
Mean	3060.9167
Std Dev	1065.3573

Price<79

Count	12
Mean	4151.75
Std Dev	658.27998



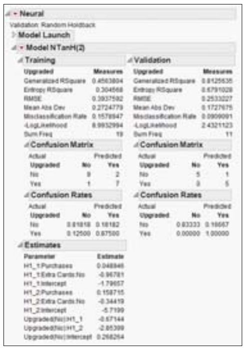
Sales	Price	Promotion
4141	59	200
3842	59	200
3056	59	200
3519	59	200
4226	59	400
4630	59	400
3507	59	400
3754	59	400
5000	59	600
5120	59	600
4011	59	600
5015	59	600
1916	79	200
675	79	200
3636	79	200
3224	79	200
2295	79	400
2730	79	400
2618	79	400
4421	79	400
4113	79	600
3746	79	600
3532	79	600
3825	79	600
1096	99	200
761	99	200
2088	99	200
820	99	200
2114	99	400
1882	99	400
2159	99	400
1602	99	400
3354	99	600
2927	99	600

Neural nets

- Constructs models from patterns and relationships uncovered in data
- Computations that begin with *inputs* and end with *outputs*
- Uses a hyperbolic tangent function
- Divide data into training sample and validation sample

Neural net example 1

"Predicting the likelihood of upgrading to a premium credit card based on the monthly purchase amount and whether the account has multiple cards" (same example used for logistic regression and classification tree)



Validation: Random Holdback

Model NTanH(2)

Training

Upgraded	Measures
Generalized RSquare	0.4563804
Entropy RSquare	0.304568
RMSE	0.3937592
Mean Abs Dev	0.2724779
Misclassification Rate	0.1578947
-LogLikelihood	8.9932994
Sum Freq	19

Confusion Matrix

Actual Upgraded	Predicted No	Yes
No	9	2
Yes	1	7

Confusion Rates

Actual Upgraded	Predicted No	Yes
No	0.81818	0.18182
Yes	0.12500	0.87500

Validation

Upgraded	Measures
Generalized RSquare	0.8125635
Entropy RSquare	0.6791028
RMSE	0.2533227
Mean Abs Dev	0.1727675
Misclassification Rate	0.0909091
-LogLikelihood	2.4321123
Sum Freq	11

Confusion Matrix

Actual Upgraded	Predicted No	Yes
No	5	1
Yes	0	5

Confusion Rates

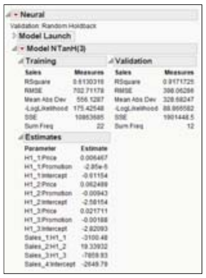
Actual Upgraded	Predicted No	Yes
No	0.83333	0.16667
Yes	0.00000	1.00000

Estimates

Parameter	Estimate
H1_1:Purchases	0.048846
H1_1:Extra Cards:No	-0.96781
H1_1:Intercept	-1.79657
H1_2:Purchases	0.158715
H1_2:Extra Cards:No	-0.34419
H1_2:Intercept	-5.7199
Upgraded(No):H1_1	-0.67144
Upgraded(No):H1_2	-2.85399
Upgraded(No):Intercept	0.268264

Neural net example 2

"Predicting sales of energy bars based on price and promotion expenses" (same example used in regression tree)



Validation: Random Holdback

Model NTanH(3)

Training

Sales	Measures
RSquare	0.6130316
RMSE	702.71178
Mean Abs Dev	556.1287
-LogLikelihood	175.42548
SSE	10863685
Sum Freq	22

Validation

Sales	Measures
RSquare	0.9171725
RMSE	398.06298
Mean Abs Dev	328.68247
-LogLikelihood	88.866582
SSE	1901448.5
Sum Freq	12

Estimates

Parameter	Estimate
H1_1:Price	0.006467
H1_1:Promotion	-2.85e-5
H1_1:Intercept	-0.61154
H1_2:Price	0.062489
H1_2:Promotion	-0.00943
H1_2:Intercept	-2.58154
H1_3:Price	0.021711
H1_3:Promotion	-0.00188
H1_3:Intercept	-2.82093
Sales_1:H1_1	-3100.48
Sales_2:H1_2	19.33932
Sales_3:H1_3	-7859.83
Sales_4:Intercept	-2649.79

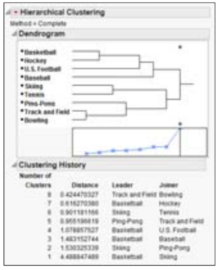
Cluster analysis

Classifies data into a sequence of groupings such that objects in each group are more alike other objects in their group than they are to objects found in other groups.

- Hierarchical clustering
- *k*-means clustering
- Distance measures
- Types of linkage between clusters

Cluster analysis example

"Perception of sports based on a survey of these attributes: movement speed, rules, team orientation, amount of contact"



Number of Clusters	Distance	Leader	Joiner
8	0.424470327	Track and Field	Bowling
7	0.616270380	Basketball	Hockey
6	0.901181166	Skiing	Tennis
5	0.955196619	Ping-Pong	Track and Field
4	1.078857527	Basketball	U.S. Football
3	1.483152744	Basketball	Baseball
2	1.530325339	Skiing	Ping-Pong
1	4.488847489	Basketball	Skiing

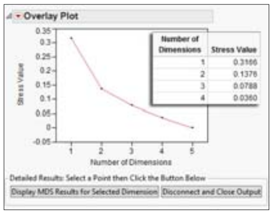
Multi-dimensional scaling

Visualizes objects in a two or more dimensional space, or map, with the goal of discovering patterns of similarities or dissimilarities among the objects.

- Types of multidimensional scaling
- Distance measures
- Stress statistic – measure of fit
- Challenge in interpreting dimensions

Multi-dimensional scaling example using JMP add-in

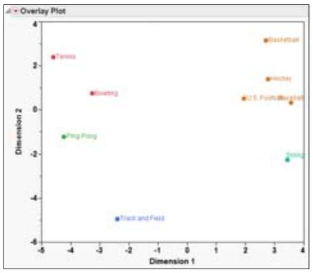
"Perception of sports based on a survey of these attributes: movement speed, rules, team orientation, amount of contact"



Number of Dimensions	Stress Value
1	0.3166
2	0.1376
3	0.0788
4	0.0360

Multi-dimensional scaling example using JMP add-in

"Perception of sports based on a survey of these attributes: movement speed, rules, team orientation, amount of contact"



Software Resources

- Microsoft Excel (latest versions equipped Apps for Office)
  - Good for selected dashboard elements (treemap, gauges, sparklines) and illustrating drill-down (with PivotTables) and subsetting (with Slicers)
  - Extend with third-party add-ins to perform logistic regression
- Tableau Public (web-based, free download)
  - Good for descriptive analytics (bullet graph, treemaps)
  - Drag-and-drop interface that can be taught in minutes
  - "Premium" version (not free) extends utility of software to many other methods, although this server-based version is more geared to business
- JMP
  - Many displays have drill-down built into them
  - Good for regression trees, neural nets, cluster analysis, and multidimensional scaling (with additional free add-in)
  - Requires SAS or R for some processing; user interface contains some quirks for new and casual users (most of which could be eliminated through the use of custom add-ins)
  - Future versions promise additional capabilities.

Can I Incorporate Any of This Into the Introductory Course?

- Could add some of the descriptive analytics into the introductory course
  - Drill down and subsetting
  - Perhaps one graph that summarize volume and velocity
  - Show-and-tell to illustrate information design and/or "sexiness" versus usefulness issue
- Could add binary logistic regression if your course covers multiple regression and mentions binary logistic regression, but this will not be feasible in most cases
- "Funny, you should ask that question…."

References

- Berenson, M. L., D. M. Levine, and K. A. Szabat. *Basic Business Statistics 13th edition*. Upper Saddle River: Pearson Education, forthcoming January 2014.
- Breiman, L., J. Friedman, C. J. Stone, and R. A. Olshen. *Classification and Regression Trees*. London: Chapman and Hall, 1984.
- Cox, T. F., and M. A. Cox. *Multidimensional Scaling, Second edition*. Boca Raton, FL: CRC Press, 2010.
- Everitt, B. S., S. Landau, and M. Leese. *Cluster Analysis, Fifth edition*. New York: John Wiley, 2011.
- Few, S. *Information Dashboard Design: Displaying Data for At-a-Glance Monitoring, Second edition*. Burlingame, CA: Analytics Press, 2013.
- Hakimpoor, H., K. Arshad, H. Tat, N. Khani, and M. Rahmandoust. “Artificial Neural Network Application in Management.” *World Applied Sciences Journal*, 2011, 14(7): 1008–1019.
- R. Klimberg, and B. D. McCullough. *Fundamentals of Predictive Analytics with JMP*. Cary, NC: SAS Press. 2013
- Lindoff, G., and M. Berry. *Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management*. Hoboken, NJ: Wiley Publishing, Inc., 2011.
- Loh, W. Y. “Fifty years of classification and regression trees.” *International Statistical Review*, 2013, in press
- Tufte, E. *Beautiful Evidence*. Cheshire, CT: Graphics Press, 2006.

79

Further Information or Contact

- Contact us at analytics@davidlevinestatistics.com
- Visit analytics.davidlevinestatistics.com for
  - Today’s slides including references
  - A preview of some of our current work in this area
  - *Coming soon* WaldoLands.com
- Look for our (*very* occasional) tweets using #AnalyticsEducation

80

A Course in Data Discovery and Predictive Analytics

David M. Levine, Baruch College—CUNY

Kathryn A. Szabat, La Salle University

David F. Stephan, Two Bridges Instructional Technology

analytics.davidlevinestatistics.com
DSI MSMESB session, November 16, 2013

What Are We Talking About?

- A definition of business analytics
- Broad categories of business analytics (INFORMS 2010-2011)
- Business analytics continues to become increasingly important in business and therefore in business education

Course Justification and Starting Points

- Addresses a topic of growing interest
- Introduces methods of problem description and decision-making not seen elsewhere in the business statistics curriculum
- Assumes a pre-requisite introductory course that covers descriptive statistics, confidence intervals and hypothesis testing, and simple linear regression
- Presents methods that have antecedents in introductory course

Guiding Principles

- Technology use should not hamper students ability to learn concepts
- Emphasize application of methods (business students are the audience)
- Compare and contrast with decision-making using traditional methods where possible.
- Capitalize on insights gained teaching related subjects such as CIS and OR/MS

How Our Teaching Experience Informs Us

As a team, our varied backgrounds and interests contribute to shaping our choices

How David Levine's Teaching Experience Informs Us

- Have sought to make statistics useful to students majoring in the functional areas of accounting, economics/finance, management, and marketing
- Have changed my focus as changes in technology occurred over time

Early 1980s – Integrated software such as SAS, SPSS, and Minitab into introductory course

- Enabled me to begin focusing on results rather than calculations
- Helped me realize that students trained to use statistical programs would have increased opportunities in business

Late 1980s/early 1990s – Started to focus on software with enhanced user interfaces that replaced older, programming-oriented interfaces

Saw how this would make statistical tools more accessible to novice students, in particular.

Early 1990s – Integrated Deming’s Total Quality Management philosophy and practices into the introductory course.

- Through consulting work, learned the importance of organizational culture and the difficulty of implementing change
- This had limited long term impact as coverage of this topic migrated to operations management

Late 1990s – Pondered the use of Microsoft Excel, by then prevalent in business schools

- Realized Excel needed to be modified for classroom use
- Crossed paths and discovered shared interests with David Stephan

Current Day – Reflected on analytics

- Crossed path and discovered shared interests with Kathy Szabat.
- Realized this is our best opportunity to make business statistics critical to the success of majors in the functional areas
- Believe this represents an opportunity to develop new majors in analytics and revise majors in business statistics (CIS, *et. al.*)

Kathryn Szabat's Experience

Overarching guiding principle:

Statistics plays a role in problem solving and decision making.

Statistics – the methods that help transform data into useful information for decision makers

- Provides support for gut feeling, intuition, experience
- Provides opportunity to gain insight

Have consistently emphasized applications of statistics to functional areas of business

Continual outreach to colleagues in different departments within the school of business to better understand how statistics is used in the various functional areas

Have used technology extensively in the course

- Without compromising understanding of logic of formulas
- Advocating the importance of “using a tool” to generate results

Have increased, over time, focus on problem-solving and decision-making

With attention to “formulating the problem”

Have increased, over time, focus on interpretation and communication

Someone has to tell the story at the end

Have recently been engaged in developing a new, interdisciplinary academic department, Business Systems and Analytics

- Effort as a response to the technology and data-driven changes in business today
- Outreach to practitioners to better understand "business analytics" as an emerging field
- Developed an introductory presentation on business analytics to be used by all faculty in the introductory statistics course (as well as introductory IS and operations courses)

David Stephan’s Experience

- Visualization has always been a theme in my work and interests
- Context-based learning advocate
- Witnessed and taught about several generations of information technology

How things work versus how to work with things

- Do you remember the ALU and CU?
- CP/M or DOS—Which is the better choice?
- When is the last time someone asked you about the ASCII table?

Relational Database Debate

- The story of the textbook that omitted the dBASE language

```
Accept “Last Name:” to lastname
Input “Grade:” to grade
@5,10 SAY Trim(lastname) + grade PICTURE 99.9
```

- Should database examples use one relation or two or more?

Lessons from the Debate

- Simpler things can be used to teach operating principles and simulate more complex things
- Large-scale things can be imagined from small-scale things
- Don’t fuss over technology choices—*in the long-run, your choice will most likely not be future-proof!*

Challenge: Finding the right level of abstraction to teach.

- If you don’t teach {*formulas, computations, fully explain methods, widgets, whatever*}, students will not understand “anything.”
- How many helpful “black boxes” do you already use without explanation?
  - The Microsoft Excel xls file format
- Don’t try to reveal/decompose all complex systems
  - Can end up discussing parts that, at a later time, get use as an integrated whole

New Challenges to Address

- "Volume, velocity, and variety" How to address these data characteristics often associated with analytics?
- Semi-subjective analysis of outputs (e.g., 3D scatterplots or cluster plots)
- Examining patterns before testing hypotheses
- Need to determine when to assign causality (to relationships) as part of the analysis versus testing a hypothesized causality

Seeking Course “Bests”

- Best Topics to Teach
- Best Technology to Use
- Best Context to Deliver Instruction

“Best” Topics to Teach

- Descriptive analytics/data discovery: most likely to be seen, builds on and extends introductory descriptive methods. Can be used to raise and “simulate” volume and velocity issues.
- Predictive not prescriptive analytics. The latter brings into play management insight, judgment, and wisdom. (Predictive combines traditional statistical analysis with data mining, as defined earlier.)

“Best” Technology to Use

- Experience teaches us not to be overly concerned about choice!
- No one program, application, or package is best in 2013
- Best technology combines most accessible with what bests illustrates the concept
- Our choice: mix of Microsoft Excel, Tableau Public, and JMP

“Best” Context to Deliver Instruction

- A broad case that represents an enterprise of suitable complexity, yet one that can be understandable on a casual level
- Our choice: a theme park with several different parts (“lands”) and an integrated resort hotel

Course Description In-Depth

Topic List (with suggested weeks)

- Introduction (2)
- Descriptive Analytics (2)
- Preparing for Predictive Analytics (1)
- Multiple regression including residual analysis, dummy variables, interaction terms, and influence analysis (1.5-2)
- Logistic regression (1)
- Multiple regression model building including transformations, collinearity, stepwise regression, and best subsets (1.5-2)
- Predictive Analytics (4-5)

Introduction (2 weeks)

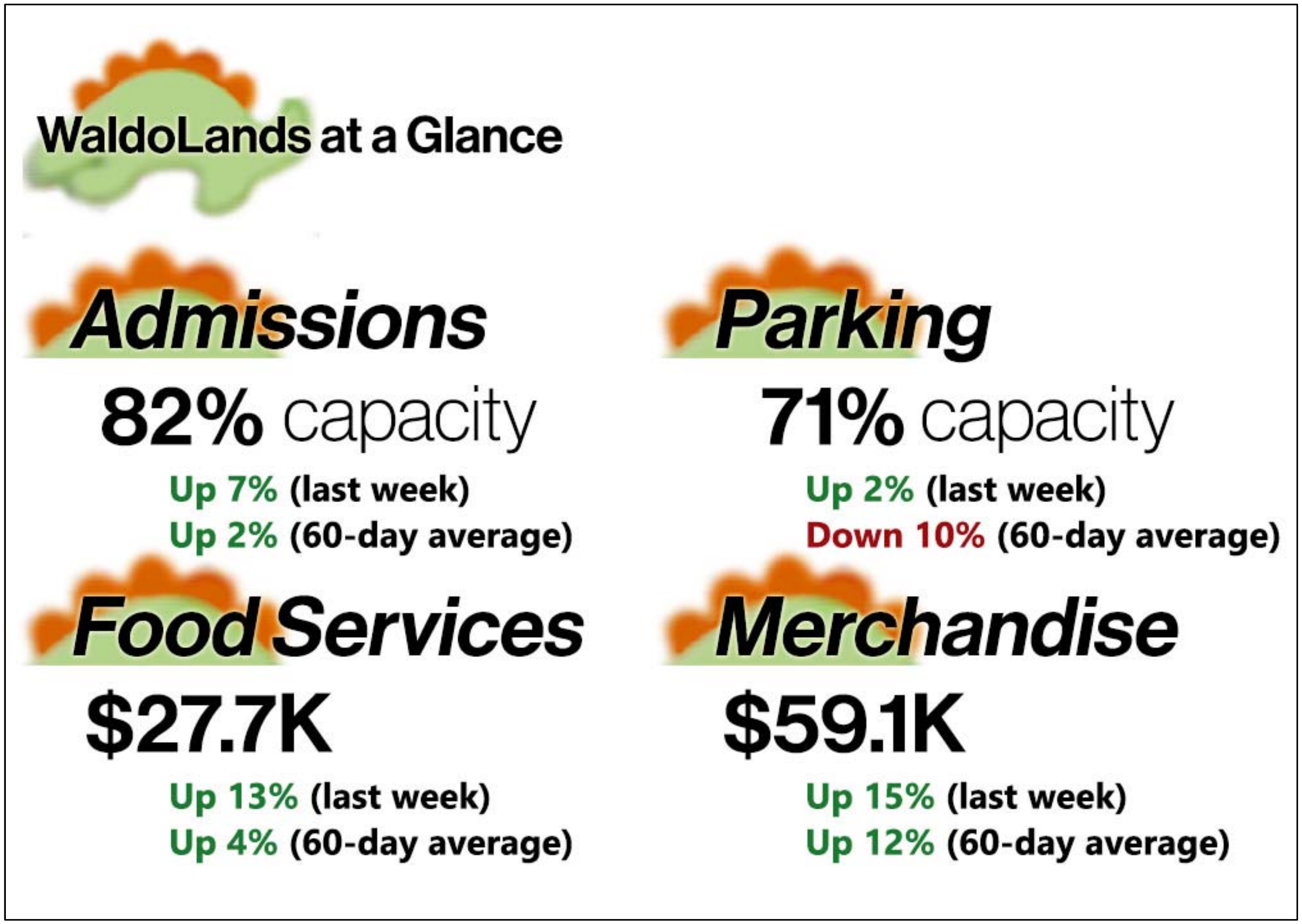
- How We Got Here: Evolutionary changes that have led to more widespread usage of analytics
- How analytics can change the data analysis and decision-making processes
- Basic vocabulary and taxonomy of analytics
- Technology requirements and orientation

Descriptive Analytics (2 weeks)

- Summarizing volume and velocity
- “Sexiness” versus usefulness issue
- Levels of summary: drill down, levels of hierarchy, and subsetting
- Information design principles that inform descriptive methods

Summarizing volume and velocity: Dashboards

Provide information about the current status of a business or business activity in a form easy to comprehend and review.



WaldoLands at a Glance

Admissions

82% capacity

Up 7% (last week)
Up 2% (60-day average)

Parking

71% capacity

Up 2% (last week)
Down 10% (60-day average)

Food Services

$27.7K

Up 13% (last week)
Up 4% (60-day average)

Merchandise

$59.1K

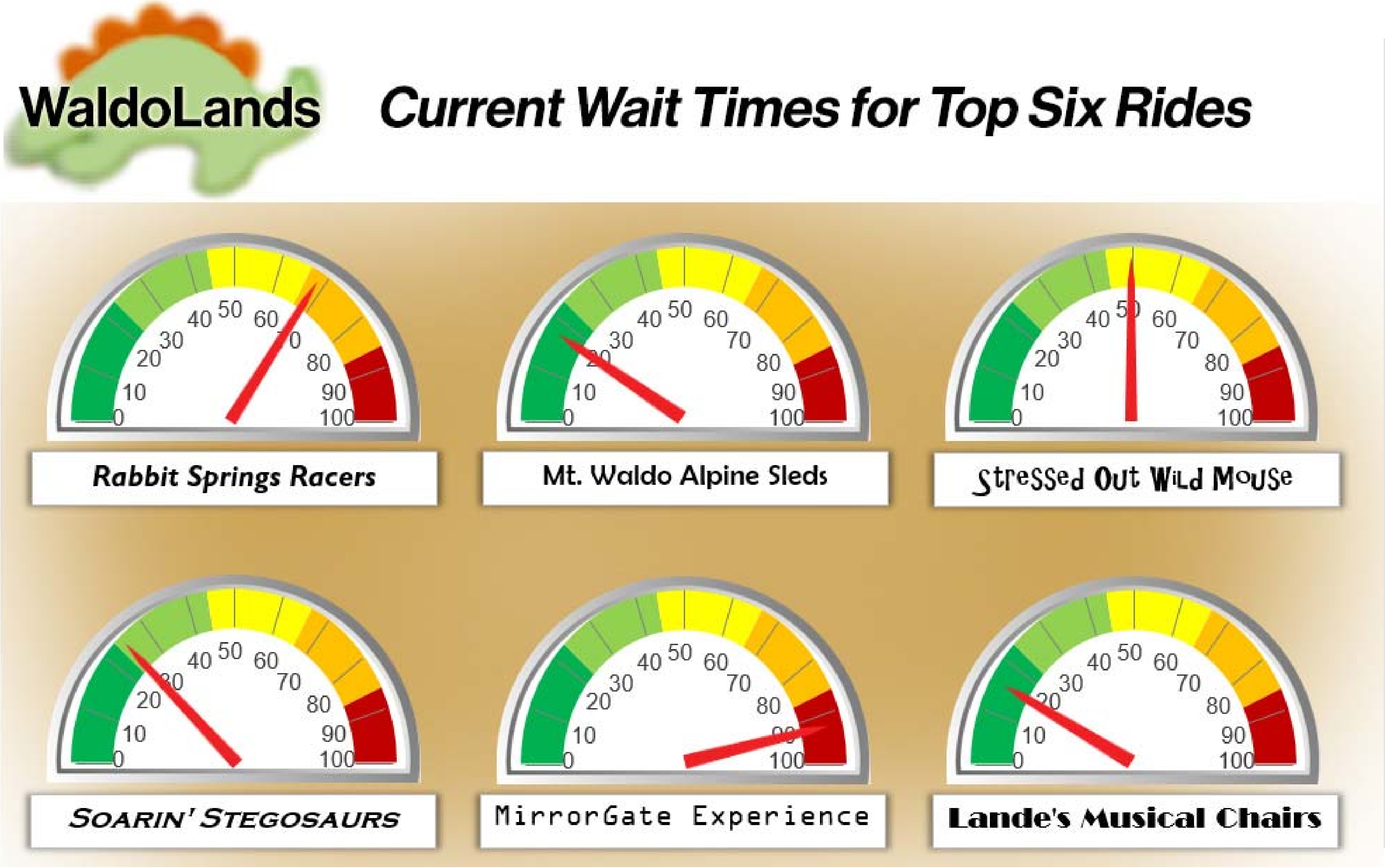
Up 15% (last week)
Up 12% (60-day average)

Sexiness versus usefulness: Gauges vs. bullet graphs

Example: combining a numerical measure with a categorical group

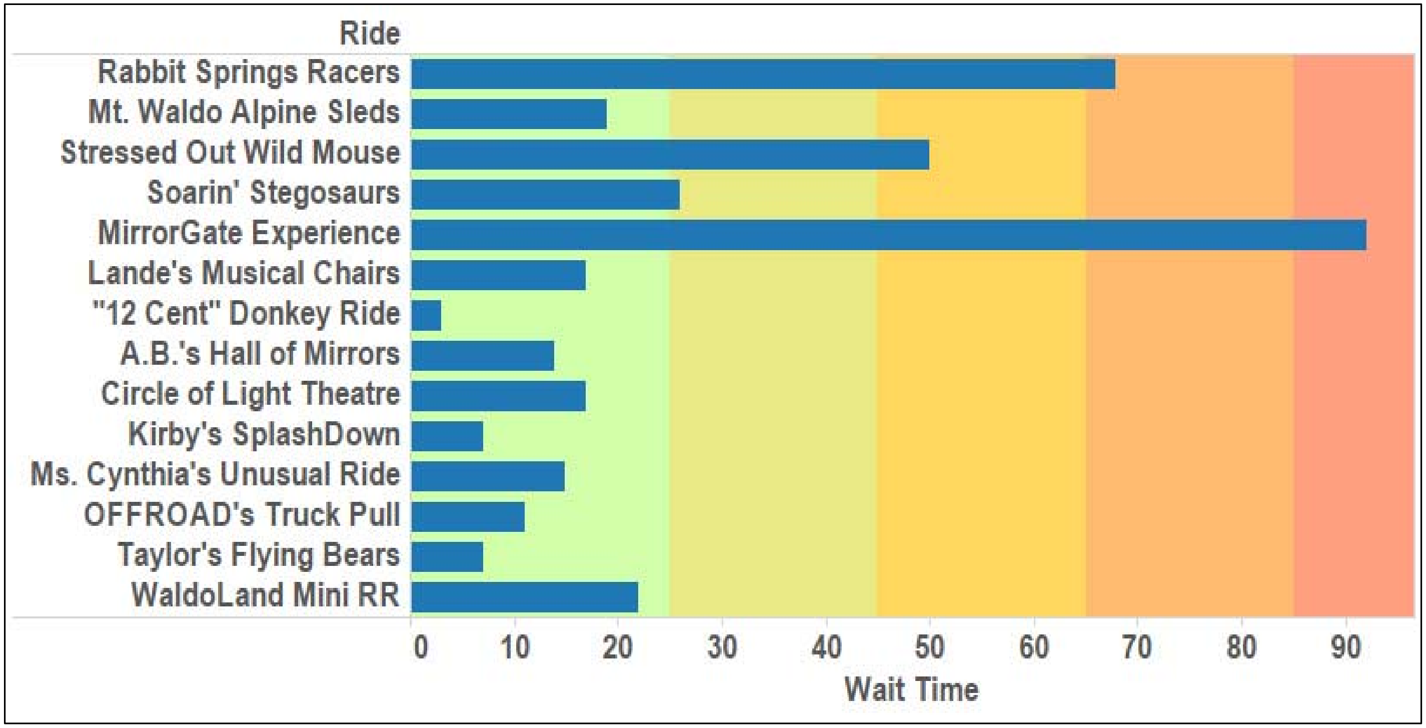
- Which one looks more “sexy,” appealing, interesting, *etc*.?
- Which one best facilitates comparisons?
- What if the answers to the two questions are different?

Sexiness versus usefulness: Gauges vs. bullet graphs



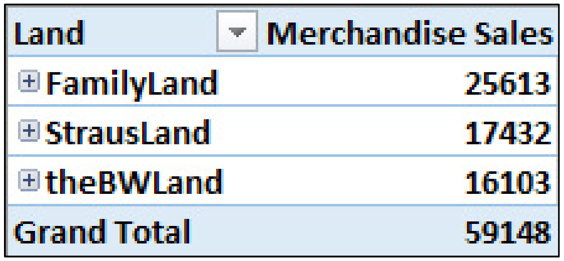
Sexiness versus usefulness: Gauges vs. bullet graphs

- Which one looks more “sexy,” appealing, interesting, *etc*.?
- Which one best facilitates comparisons?
- What if the answers to the two questions are different?

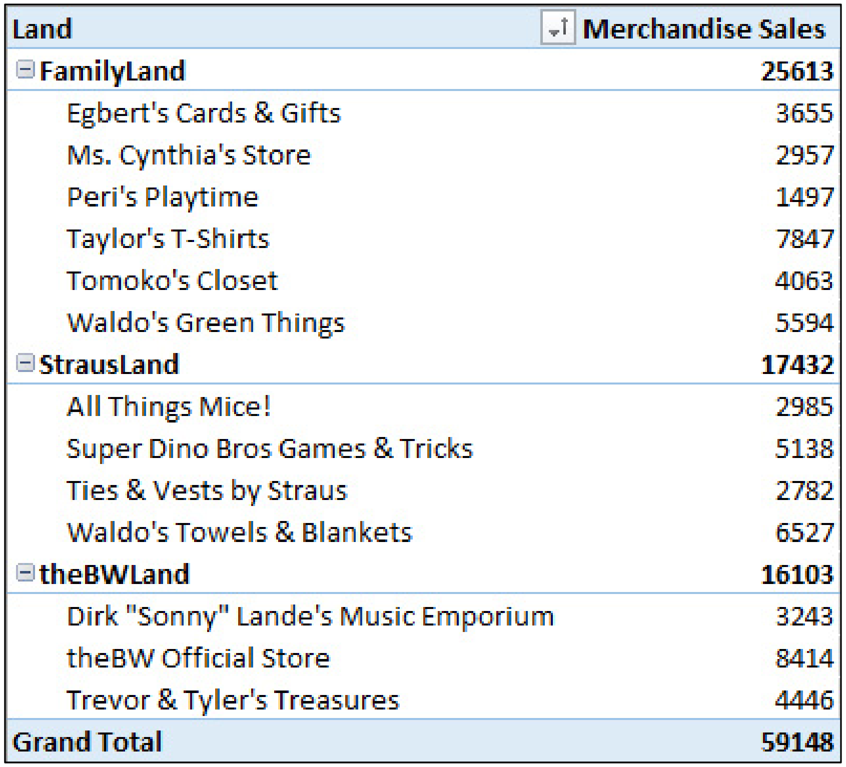


Levels of summary: drill down, levels of hierarchy, and subsetting

Drill-down sequence example (using Excel)



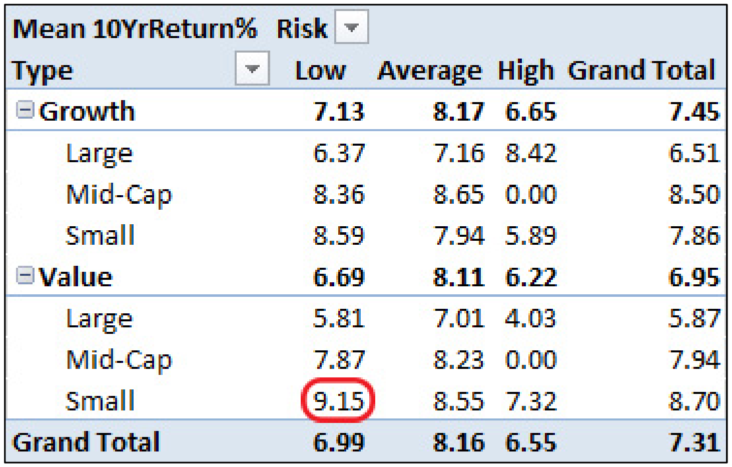
Land	Merchandise Sales
⊞ FamilyLand	25613
⊞ StrausLand	17432
⊞ theBWLand	16103
Grand Total	59148



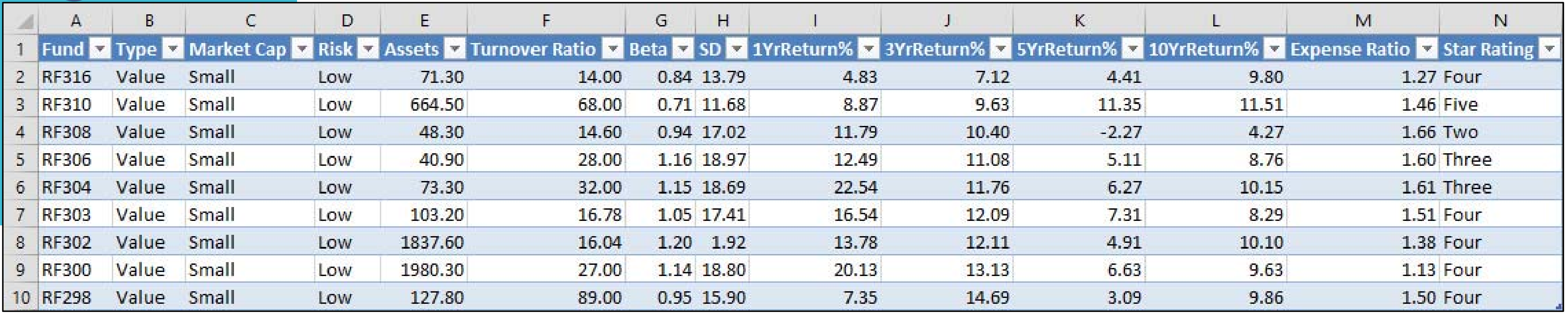
Land	Merchandise Sales
⊟ **FamilyLand**	**25613**
Egbert's Cards & Gifts	3655
Ms. Cynthia's Store	2957
Peri's Playtime	1497
Taylor's T-Shirts	7847
Tomoko's Closet	4063
Waldo's Green Things	5594
⊟ **StrausLand**	**17432**
All Things Mice!	2985
Super Dino Bros Games & Tricks	5138
Ties & Vests by Straus	2782
Waldo's Towels & Blankets	6527
⊟ **theBWLand**	**16103**
Dirk "Sonny" Lande's Music Emporium	3243
theBW Official Store	8414
Trevor & Tyler's Treasures	4446
Grand Total	**59148**

Levels of summary: drill down, levels of hierarchy, and subsetting

Financial example showing another level of drill-down



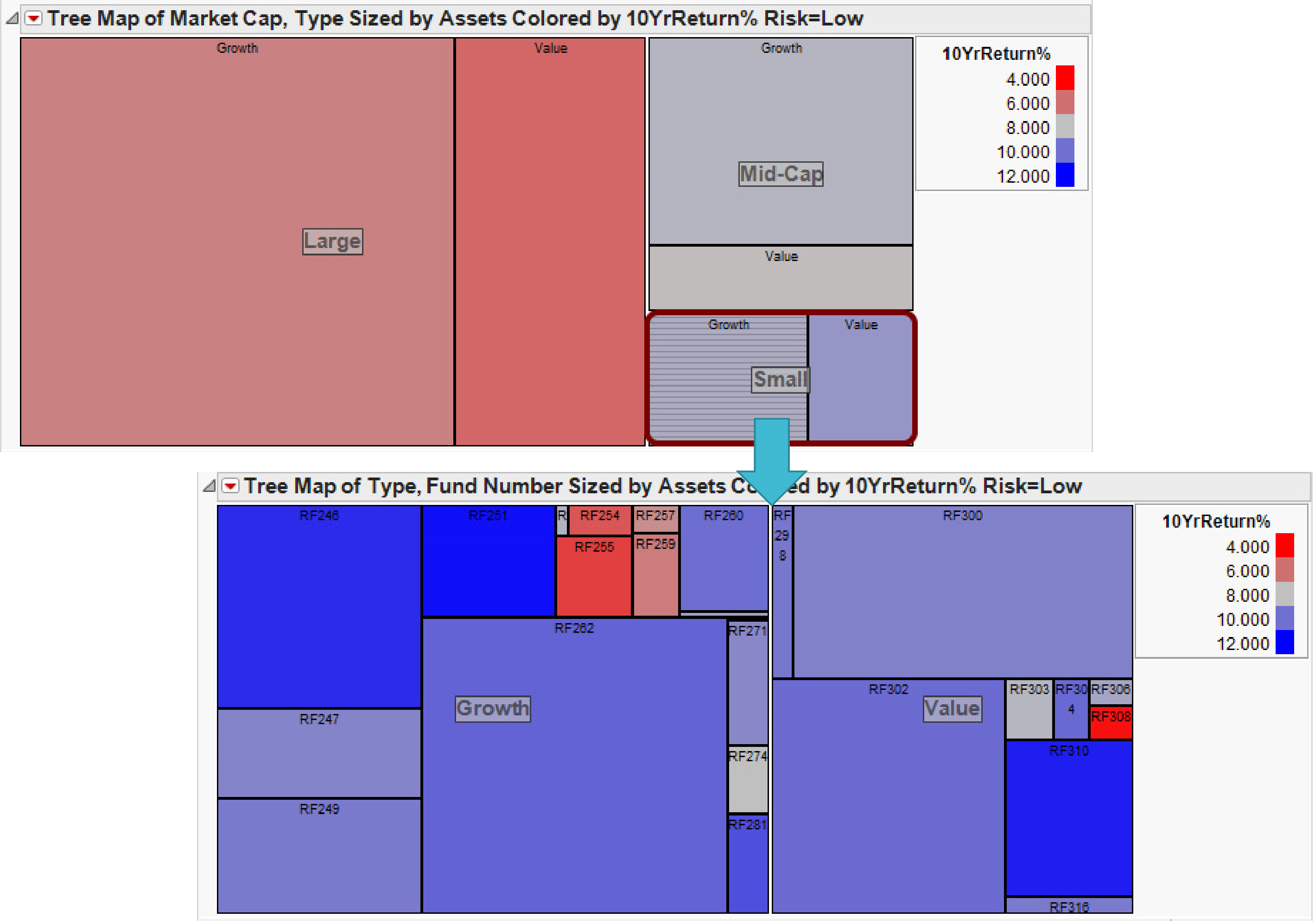
Mean 10YrReturn%	Risk			
Type	Low	Average	High	Grand Total
Growth	**7.13**	**8.17**	**6.65**	**7.45**
Large	6.37	7.16	8.42	6.51
Mid-Cap	8.36	8.65	0.00	8.50
Small	8.59	7.94	5.89	7.86
Value	**6.69**	**8.11**	**6.22**	**6.95**
Large	5.81	7.01	4.03	5.87
Mid-Cap	7.87	8.23	0.00	7.94
Small	9.15	8.55	7.32	8.70
Grand Total	**6.99**	**8.16**	**6.55**	**7.31**



	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	Fund	Type	Market Cap	Risk	Assets	Turnover Ratio	Beta	SD	1YrReturn%	3YrReturn%	5YrReturn%	10YrReturn%	Expense Ratio	Star Rating
2	RF316	Value	Small	Low	71.30	14.00	0.84	13.79	4.83	7.12	4.41	9.80	1.27	Four
3	RF310	Value	Small	Low	664.50	68.00	0.71	11.68	8.87	9.63	11.35	11.51	1.46	Five
4	RF308	Value	Small	Low	48.30	14.60	0.94	17.02	11.79	10.40	-2.27	4.27	1.66	Two
5	RF306	Value	Small	Low	40.90	28.00	1.16	18.97	12.49	11.08	5.11	8.76	1.60	Three
6	RF304	Value	Small	Low	73.30	32.00	1.15	18.69	22.54	11.76	6.27	10.15	1.61	Three
7	RF303	Value	Small	Low	103.20	16.78	1.05	17.41	16.54	12.09	7.31	8.29	1.51	Four
8	RF302	Value	Small	Low	1837.60	16.04	1.20	1.92	13.78	12.11	4.91	10.10	1.38	Four
9	RF300	Value	Small	Low	1980.30	27.00	1.14	18.80	20.13	13.13	6.63	9.63	1.13	Four
10	RF298	Value	Small	Low	127.80	89.00	0.95	15.90	7.35	14.69	3.09	9.86	1.50	Four

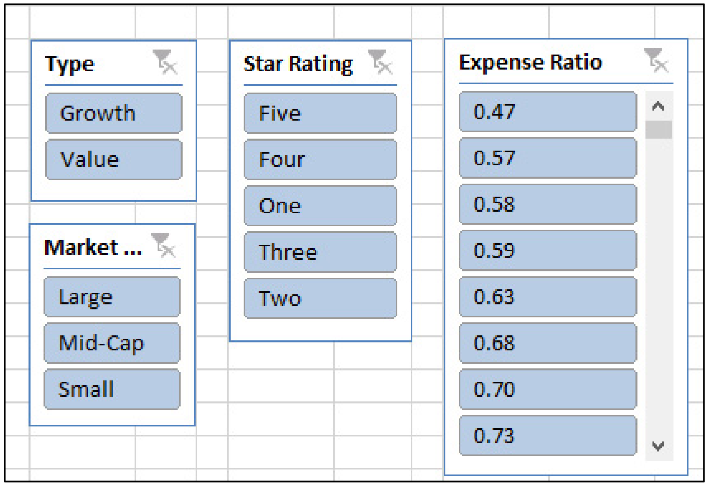
Levels of summary: drill down, levels of hierarchy, and subsetting

Visual drill-down using a tree map



Subsetting using “slicers” (Excel)

Levels of summary: drill down, levels of hierarchy, and subsetting

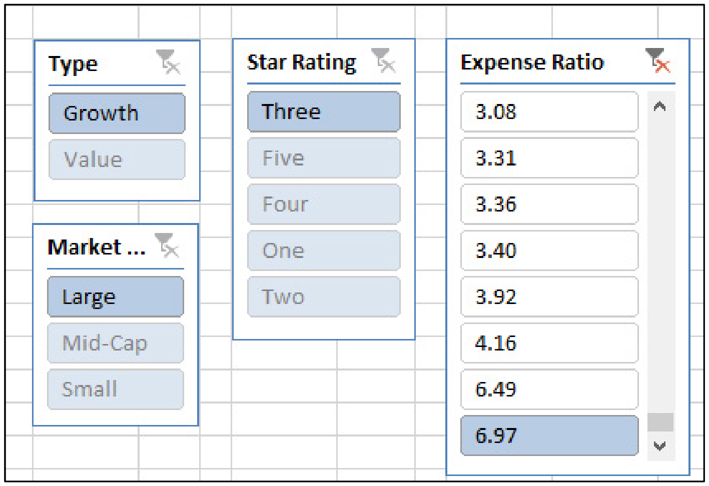


Type
Growth
Value

Market ...
Large
Mid-Cap
Small

Star Rating
Five
Four
One
Three
Two

Expense Ratio
0.47
0.57
0.58
0.59
0.63
0.68
0.70
0.73

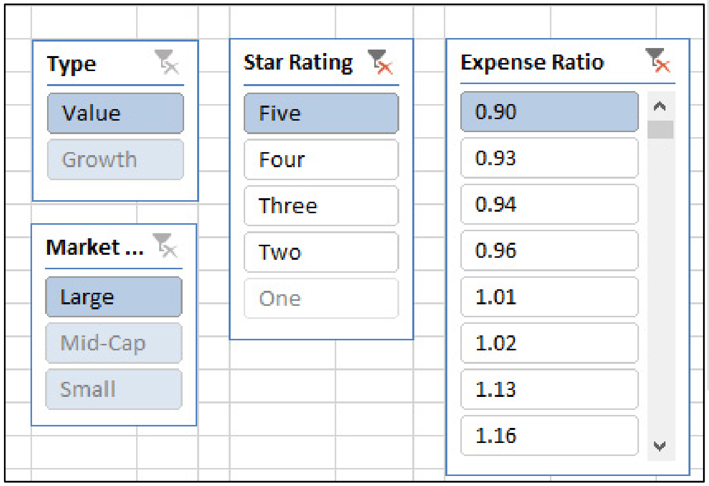


Type
Growth
Value

Market ...
Large
Mid-Cap
Small

Star Rating
Three
Five
Four
One
Two

Expense Ratio
3.08
3.31
3.36
3.40
3.92
4.16
6.49
6.97



Type
Value
Growth

Market ...
Large
Mid-Cap
Small

Star Rating
Five
Four
Three
Two
One

Expense Ratio
0.90
0.93
0.94
0.96
1.01
1.02
1.13
1.16

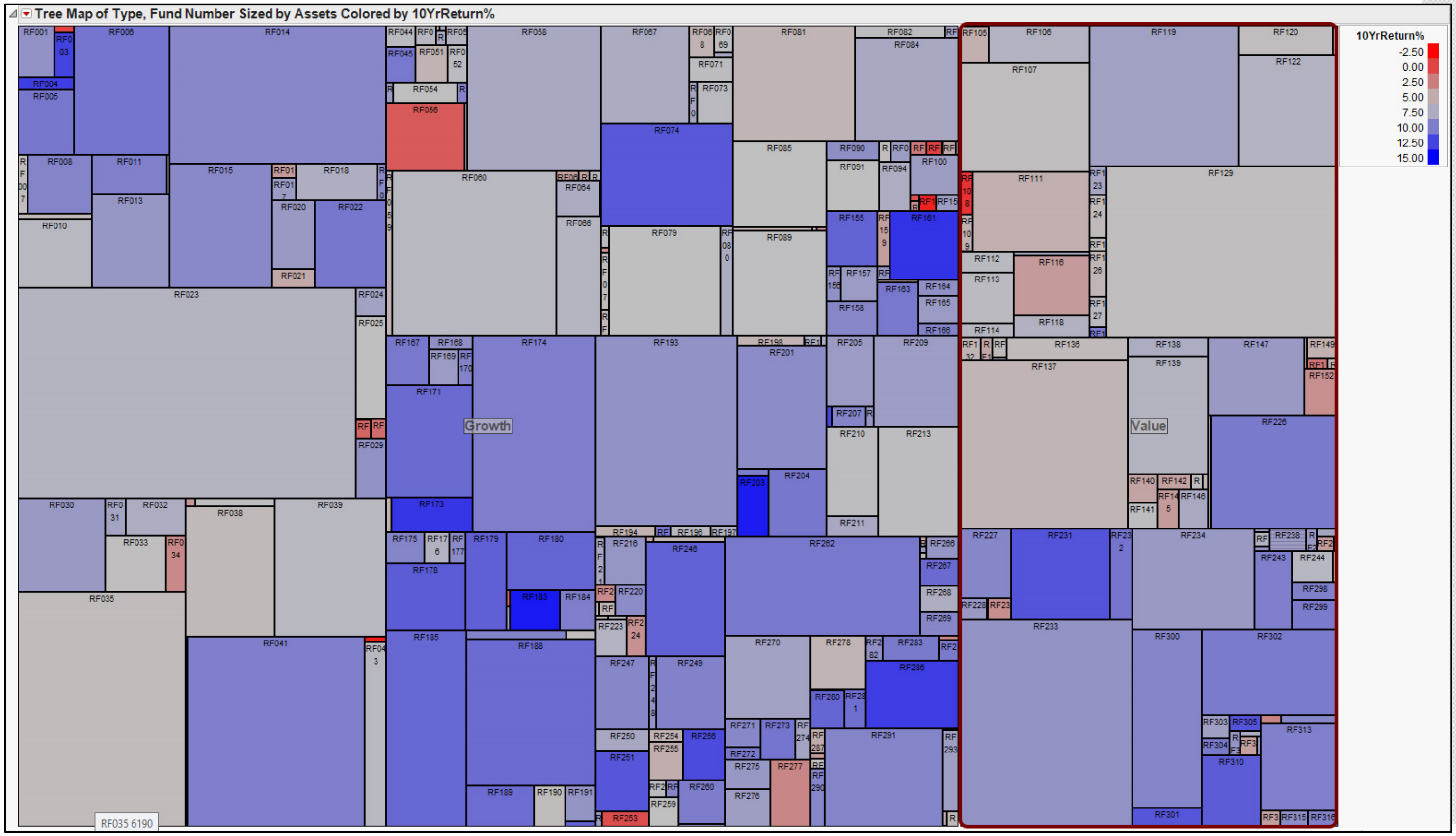
Information design principles

- Fostering efficient and effective communication and understanding
- Provide context for data in a compact presentation
- Add additional “dimensions” of data
- Misuse raises issues beyond “typical” statistical concerns: visual perception, artistic considerations

Does this tree map provide context for data in a compact presentation?

Add additional "dimensions" of data?

Tree Map of Retirement Fund Assets Colored by 10-Year Return Percentage, By Fund Type (JMP)

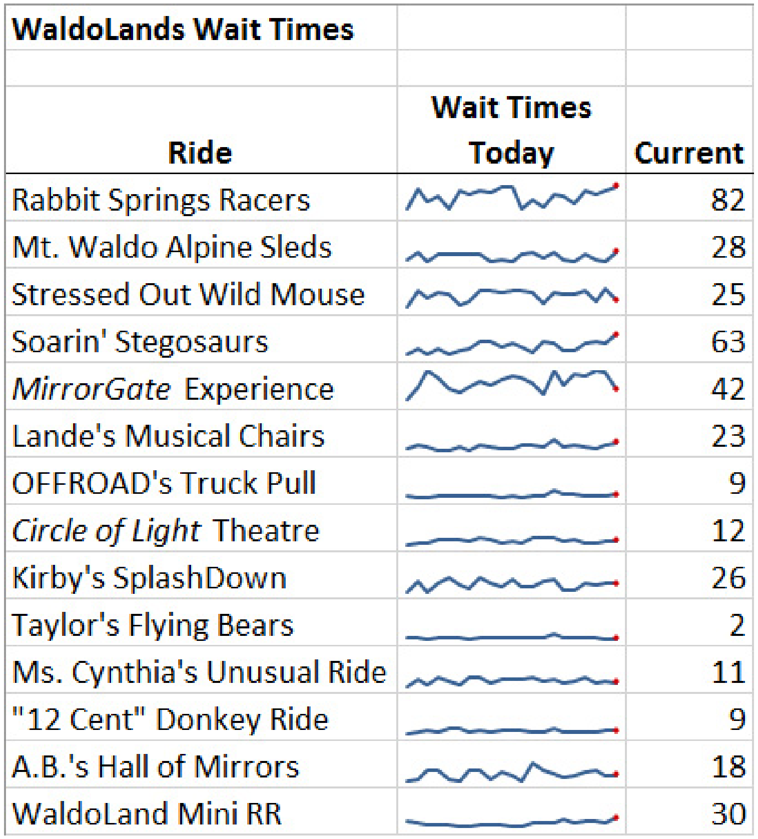


GROWTH FUNDS

VALUE FUNDS

Does this table provide context for data in a compact presentation?

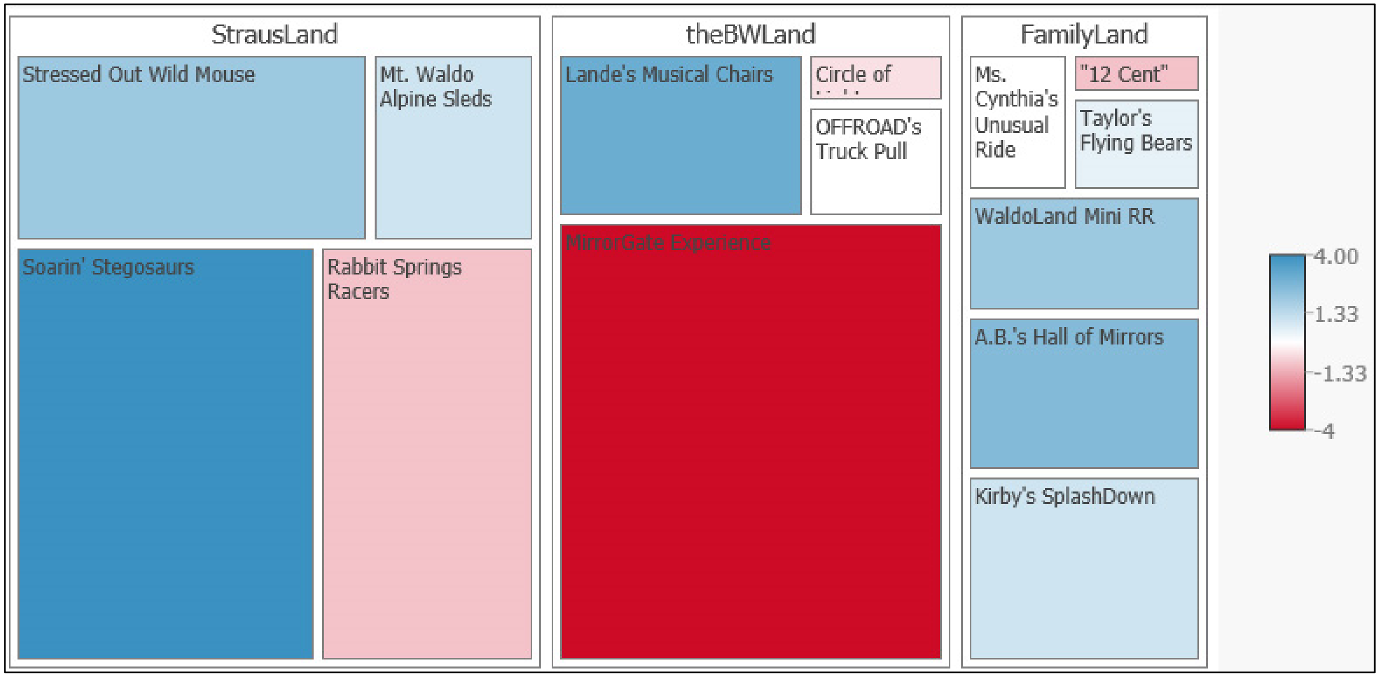
Sparklines example (Excel)



WaldoLands Wait Times		
Ride	**Wait Times Today**	**Current**
Rabbit Springs Racers		82
Mt. Waldo Alpine Sleds		28
Stressed Out Wild Mouse		25
Soarin' Stegosaurs		63
MirrorGate Experience		42
Lande's Musical Chairs		23
OFFROAD's Truck Pull		9
Circle of Light Theatre		12
Kirby's SplashDown		26
Taylor's Flying Bears		2
Ms. Cynthia's Unusual Ride		11
"12 Cent" Donkey Ride		9
A.B.'s Hall of Mirrors		18
WaldoLand Mini RR		30

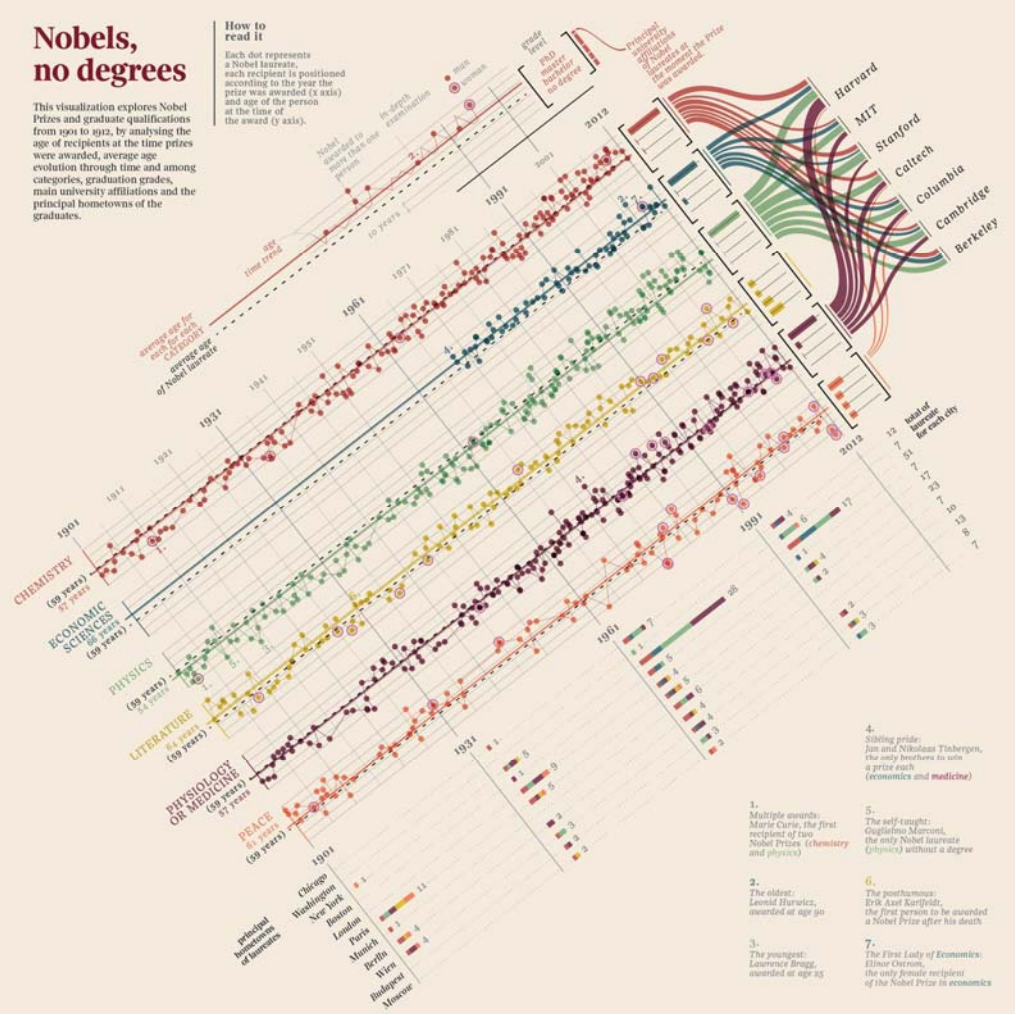
Information design tree map example with simpler data

Tree Map of Number of Social Media Comments Colored by Tone, By “Land” (Excel)



Information design principles: “infographics”

Nobel Laureates Graph (Accurat information design agency)



Nobels, no degrees

This visualization explores Nobel Prizes and graduate qualifications from 1901 to 1912, by analysing the age of recipients at the time prizes were awarded, average age evolution through time and among categories, graduation grades, main university affiliations and the principal hometowns of the graduates.

How to read it

Each dot represents a Nobel laureate, each recipient is positioned according to the year the prize was awarded (x axis) and age of the person at the time of the award (y axis).

1. Multiple awards: Marie Curie, the first recipient of two Nobel Prizes (chemistry and physics)

2. The oldest: Leonid Hurwicz, awarded at age 90

3. The youngest: Laurence Bragg, awarded at age 25

4. Sibling pride: Jan and Nikolaas Tinbergen, the only brothers to win a prize each (economics and medicine)

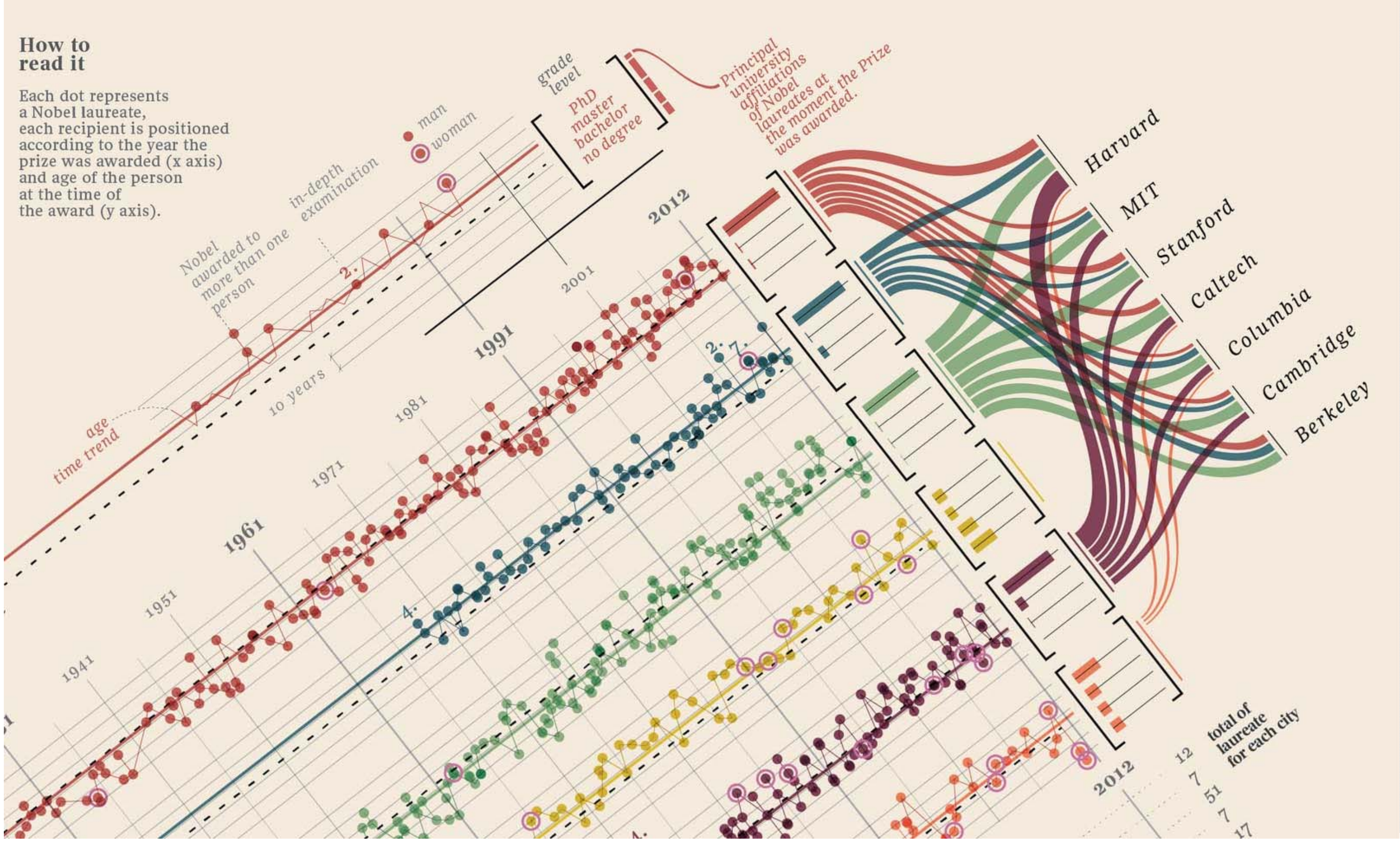
5. The self-taught: Guglielmo Marconi, the only Nobel laureate (physics) without a degree

6. The posthumous: Erik Axel Karlfeldt, the first person to be awarded a Nobel Prize after his death

7. The First Lady of Economics: Elinor Ostrom, the only female recipient of the Nobel Prize in economics

Information design principles: “infographics”

Detail of Nobel Prize Laureates Graph



How to read it

Each dot represents a Nobel laureate, each recipient is positioned according to the year the prize was awarded (x axis) and age of the person at the time of the award (y axis).

Principal university affiliations of Nobel laureates at the moment the Prize was awarded.

Preparing for Predictive Analytics (1 week)

- Confidence intervals
- Hypothesis testing
- Simple linear regression

Confidence intervals

- Normal distribution
- Sampling distributions
- Confidence intervals for the mean and proportion

Hypothesis testing

- Basic Concepts of hypothesis testing
- *p*-values
- Tests for the differences between means and proportions

Simple linear regression

- The simple linear regression model
- Interpreting the regression coefficients
- Residual analysis
- Assumptions of regression
- Inferences in simple linear regression

Multiple Regression (1.5-2 weeks)

- Developing the multiple regression model
- Inference in multiple regression
- Residual analysis
- Dummy variables
- Interaction terms
- Influence analysis

Developing the multiple regression model

- Interpreting the coefficients
- Coefficients of multiple determination
- Coefficients of partial determination
- Assumptions

Inference in multiple regression

- Testing the overall model
- Testing the contribution of each independent variable
- Adjusted r^2

Residual analysis

- Plots of the residuals vs. independent variables
- Plots of the residuals vs. predicted Y
- Plots of the residuals vs. time (if appropriate)

Dummy variables

Using categorical independent variables in a regression model:

- Defining dummy variables
- Interpreting dummy variables
- Assumptions in using dummy variables

Interaction terms

- What they are
- Why they are sometimes necessary
- Interpreting interaction terms

Influence analysis

Examining the effect of individual observations on the regression model

- Hat matrix elements h_i
- Studentized deleted residuals t_i
- Cook's Distance statistic D_i

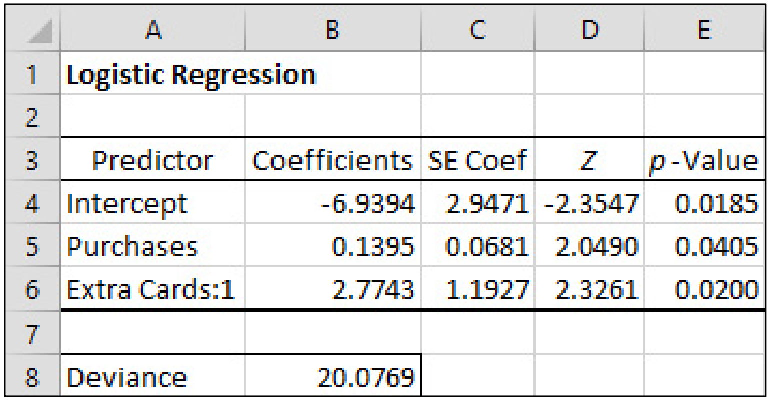
Logistic regression (1 week)

Predicting a categorical dependent variable

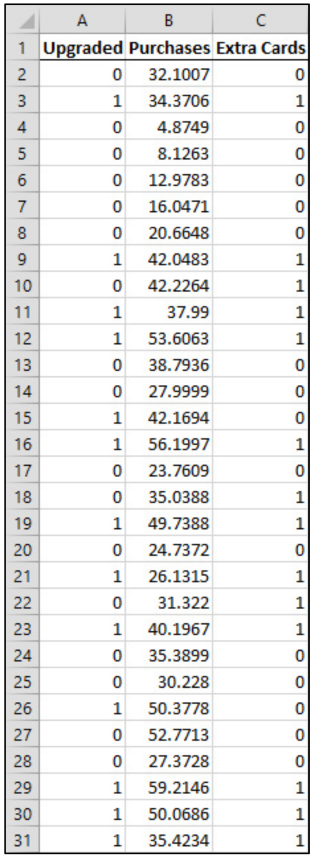
- Cannot use least squares regression
- Odds ratio
- Logistic regression model
- Predicting probability of an event of interest
- Deviance statistic
- Wald statistic

Logistic regression example using an Excel add-in

“Predicting the likelihood of upgrading to a premium credit card based on the monthly purchase amount and whether the account has multiple cards”



	A	B	C	D	E
1	**Logistic Regression**				
2					
3	Predictor	Coefficients	SE Coef	*Z*	*p*-Value
4	Intercept	-6.9394	2.9471	-2.3547	0.0185
5	Purchases	0.1395	0.0681	2.0490	0.0405
6	Extra Cards:1	2.7743	1.1927	2.3261	0.0200
7					
8	Deviance	20.0769			



	A	B	C
1	**Upgraded**	**Purchases**	**Extra Cards**
2	0	32.1007	0
3	1	34.3706	1
4	0	4.8749	0
5	0	8.1263	0
6	0	12.9783	0
7	0	16.0471	0
8	0	20.6648	0
9	1	42.0483	1
10	0	42.2264	1
11	1	37.99	1
12	1	53.6063	1
13	0	38.7936	0
14	0	27.9999	0
15	1	42.1694	0
16	1	56.1997	1
17	0	23.7609	0
18	0	35.0388	1
19	1	49.7388	1
20	0	24.7372	0
21	1	26.1315	1
22	0	31.322	1
23	1	40.1967	1
24	0	35.3899	0
25	0	30.228	0
26	1	50.3778	0
27	0	52.7713	0
28	0	27.3728	0
29	1	59.2146	1
30	1	50.0686	1
31	1	35.4234	1

Multiple Regression Model Building (1.5-2 weeks)

- Transformations
- Collinearity
- Stepwise regression
- Best subsets regression

Transformations

- Purposes
- Square root transformations
- Logarithmic transformations

Collinearity

- Effect on the regression model
- Measuring the variance inflationary factor (*VIF*)
- Dealing with collinear independent variables

Stepwise regression

- History
- How it works
- Limitations
- Use in an era of big data

Best subsets regression

- How it works
- Advantages and disadvantages vs. stepwise regression
- Mallows C_p statistic

Predictive Analytics (4-5 weeks)

	METHOD FOR			
METHOD	Prediction	Classification	Clustering	Association
Classification and regression trees (1-1.5 weeks)	●	●		
Neural networks (1-1.5 weeks)	●	●	●	
Cluster analysis (1 week)			●	
Multidimensional scaling (1week)		●		●

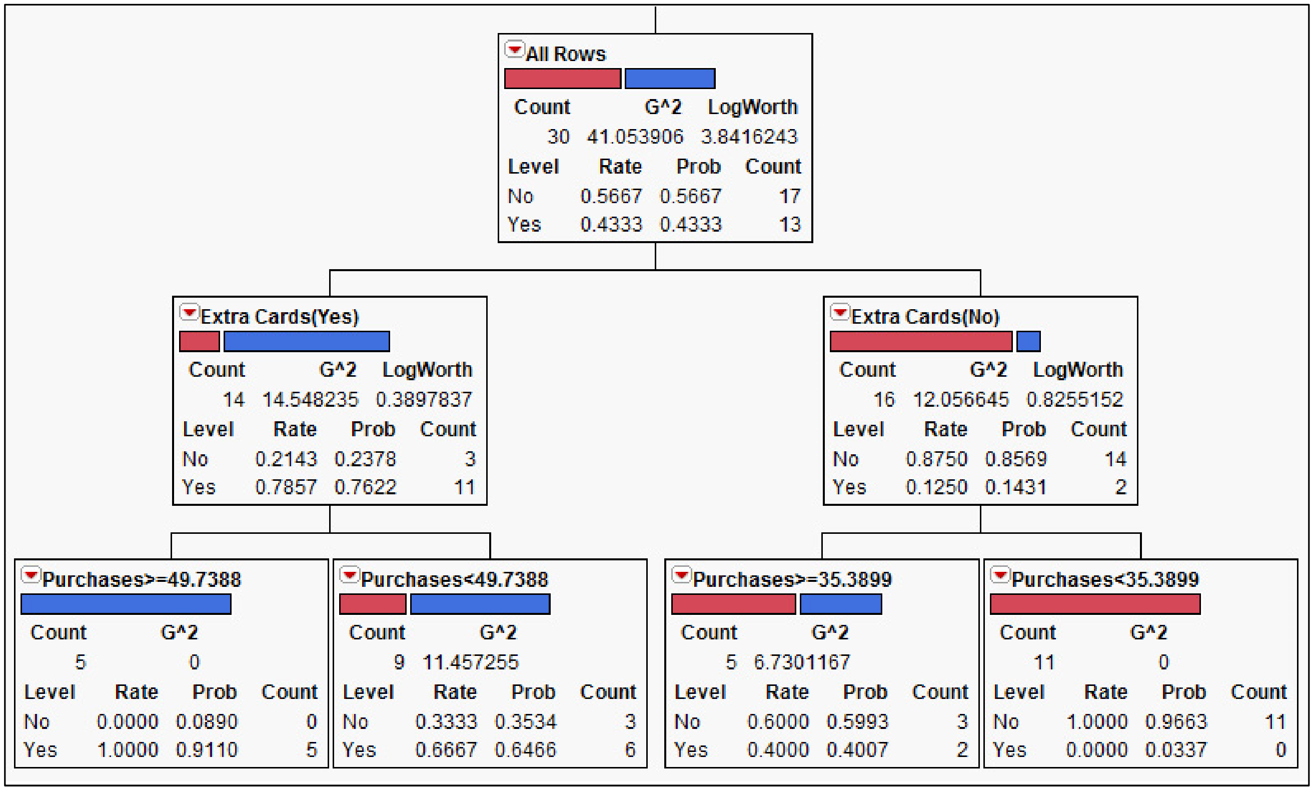
Classification and regression trees

Decision trees that split data into groups based on the values of independent or explanatory (*X*) variables.

- Not affected by the distribution of the variables
- Splitting determines which values of a specific independent variable are useful in predicting the dependent (*Y*) variable present
- Using a *categorical* dependent *Y* variable results in a *classification tree*
- Using a *numerical* dependent *Y* variable results in a *regression tree*
- Rules for splitting the tree
- Pruning back a tree
- If possible, divide data into training sample and validation sample

Classification tree example

"Predicting the likelihood of upgrading to a premium credit card based on the monthly purchase amount and whether the account has multiple cards" (same example used in logistic regression)



```
All Rows
Count        G^2   LogWorth
   30  41.053906  3.8416243
Level   Rate    Prob   Count
No     0.5667  0.5667     17
Yes    0.4333  0.4333     13
```

```
Extra Cards(Yes)
Count        G^2   LogWorth
   14  14.548235  0.3897837
Level   Rate    Prob   Count
No     0.2143  0.2378      3
Yes    0.7857  0.7622     11
```

```
Extra Cards(No)
Count        G^2   LogWorth
   16  12.056645  0.8255152
Level   Rate    Prob   Count
No     0.8750  0.8569     14
Yes    0.1250  0.1431      2
```

```
Purchases>=49.7388
Count        G^2
    5          0
Level   Rate    Prob   Count
No     0.0000  0.0890      0
Yes    1.0000  0.9110      5
```

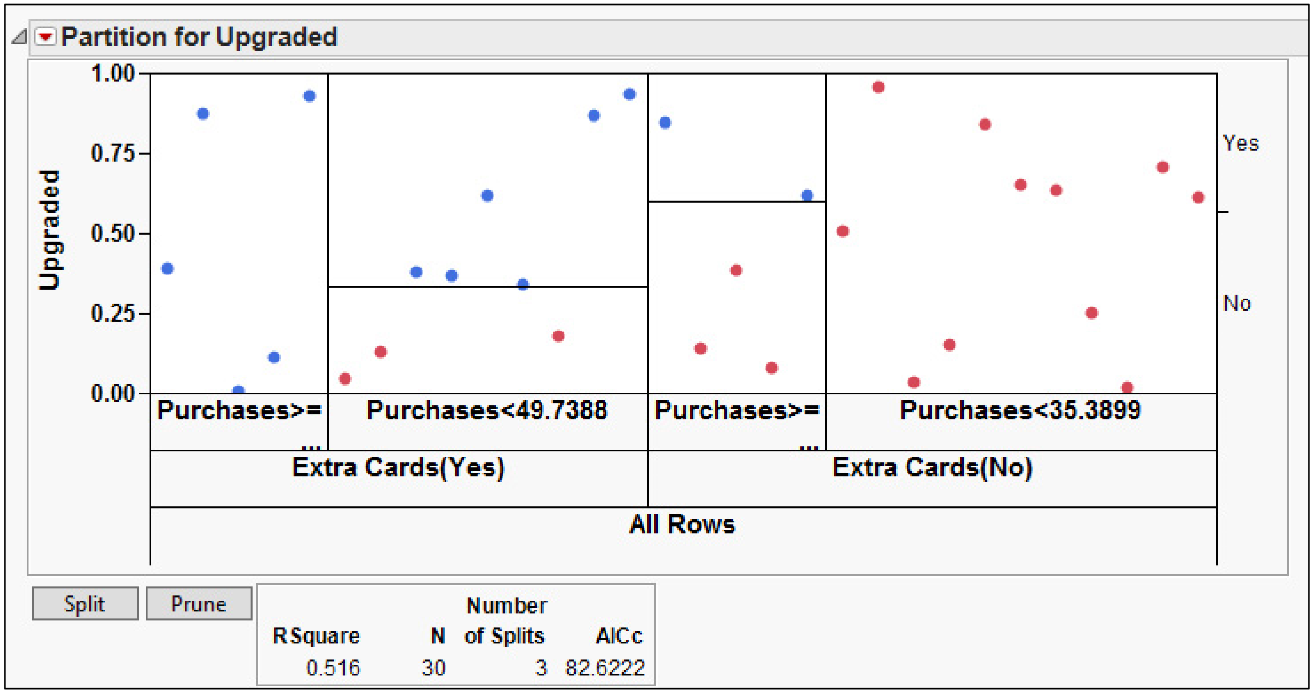
```
Purchases<49.7388
Count        G^2
    9  11.457255
Level   Rate    Prob   Count
No     0.3333  0.3534      3
Yes    0.6667  0.6466      6
```

```
Purchases>=35.3899
Count        G^2
    5  6.7301167
Level   Rate    Prob   Count
No     0.6000  0.5993      3
Yes    0.4000  0.4007      2
```

```
Purchases<35.3899
Count        G^2
   11          0
Level   Rate    Prob   Count
No     1.0000  0.9663     11
Yes    0.0000  0.0337      0
```

Classification tree example

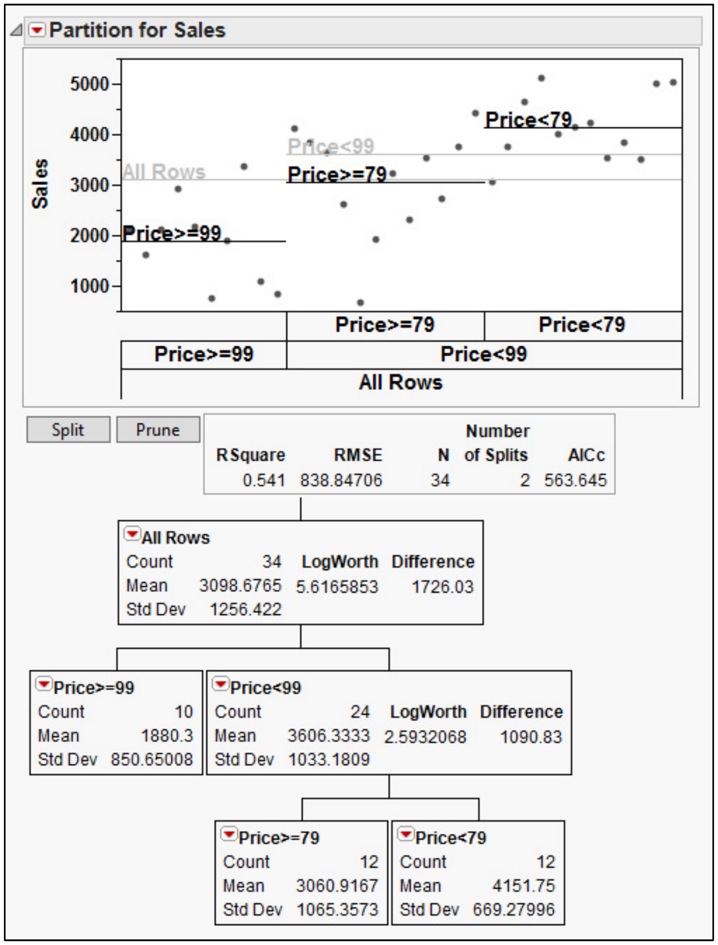
"Predicting the likelihood of upgrading to a premium credit card based on the monthly purchase amount and whether the account has multiple cards" (same example used in logistic regression)



RSquare	N	Number of Splits	AICc
0.516	30	3	82.6222

Regression tree example

"Predicting sales of energy bars based on price and promotion expenses" (could be multiple regression example, too)



RSquare	RMSE	N	Number of Splits	AICc
0.541	838.84706	34	2	563.645

All Rows

		LogWorth	Difference
Count	34		
Mean	3098.6765	5.6165853	1726.03
Std Dev	1256.422		

Price>=99

Count	10
Mean	1880.3
Std Dev	850.65008

Price<99

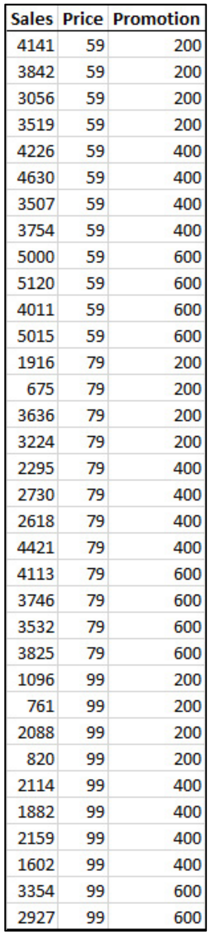
		LogWorth	Difference
Count	24		
Mean	3606.3333	2.5932068	1090.83
Std Dev	1033.1809		

Price>=79

Count	12
Mean	3060.9167
Std Dev	1065.3573

Price<79

Count	12
Mean	4151.75
Std Dev	669.27996



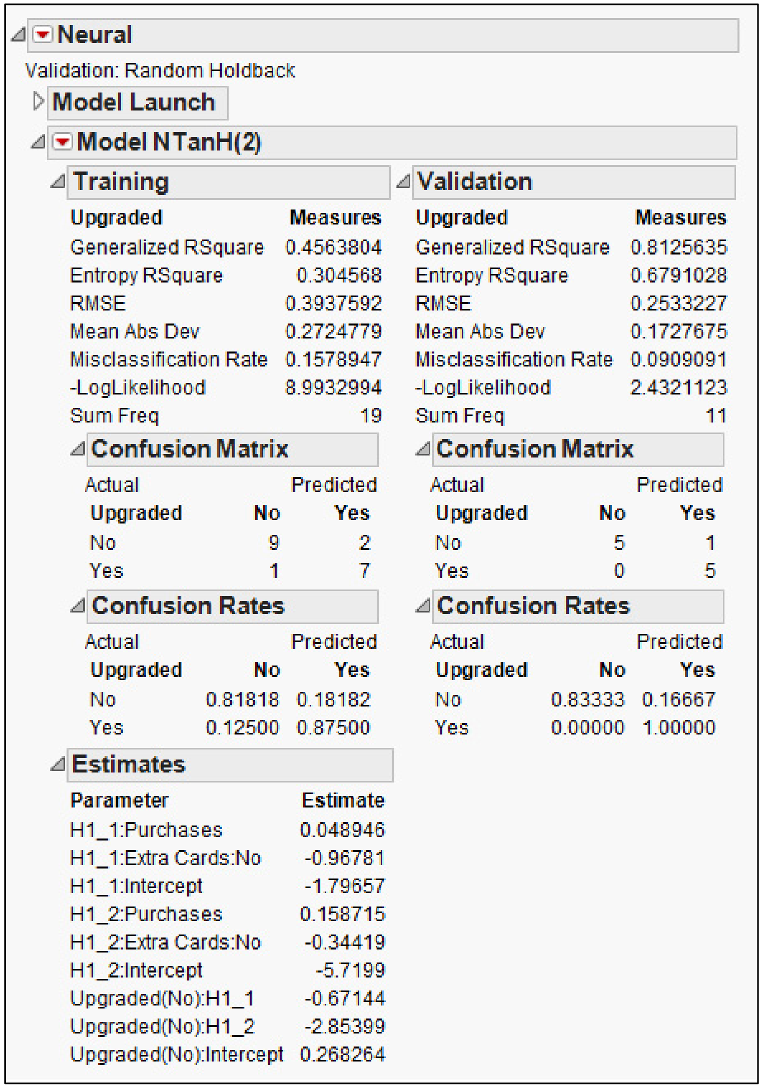
Sales	Price	Promotion
4141	59	200
3842	59	200
3056	59	200
3519	59	200
4226	59	400
4630	59	400
3507	59	400
3754	59	400
5000	59	600
5120	59	600
4011	59	600
5015	59	600
1916	79	200
675	79	200
3636	79	200
3224	79	200
2295	79	400
2730	79	400
2618	79	400
4421	79	400
4113	79	600
3746	79	600
3532	79	600
3825	79	600
1096	99	200
761	99	200
2088	99	200
820	99	200
2114	99	400
1882	99	400
2159	99	400
1602	99	400
3354	99	600
2927	99	600

Neural nets

- Constructs models from patterns and relationships uncovered in data
- Computations that begin with *inputs* and end with *outputs*
- Uses a hyperbolic tangent function
- Divide data into training sample and validation sample

Neural net example 1

"Predicting the likelihood of upgrading to a premium credit card based on the monthly purchase amount and whether the account has multiple cards" (same example used for logistic regression and classification tree)



Neural

Validation: Random Holdback

Model Launch

Model NTanH(2)

Training

Upgraded	Measures
Generalized RSquare	0.4563804
Entropy RSquare	0.304568
RMSE	0.3937592
Mean Abs Dev	0.2724779
Misclassification Rate	0.1578947
-LogLikelihood	8.9932994
Sum Freq	19

Confusion Matrix

Actual	Predicted	
Upgraded	No	Yes
No	9	2
Yes	1	7

Confusion Rates

Actual	Predicted	
Upgraded	No	Yes
No	0.81818	0.18182
Yes	0.12500	0.87500

Validation

Upgraded	Measures
Generalized RSquare	0.8125635
Entropy RSquare	0.6791028
RMSE	0.2533227
Mean Abs Dev	0.1727675
Misclassification Rate	0.0909091
-LogLikelihood	2.4321123
Sum Freq	11

Confusion Matrix

Actual	Predicted	
Upgraded	No	Yes
No	5	1
Yes	0	5

Confusion Rates

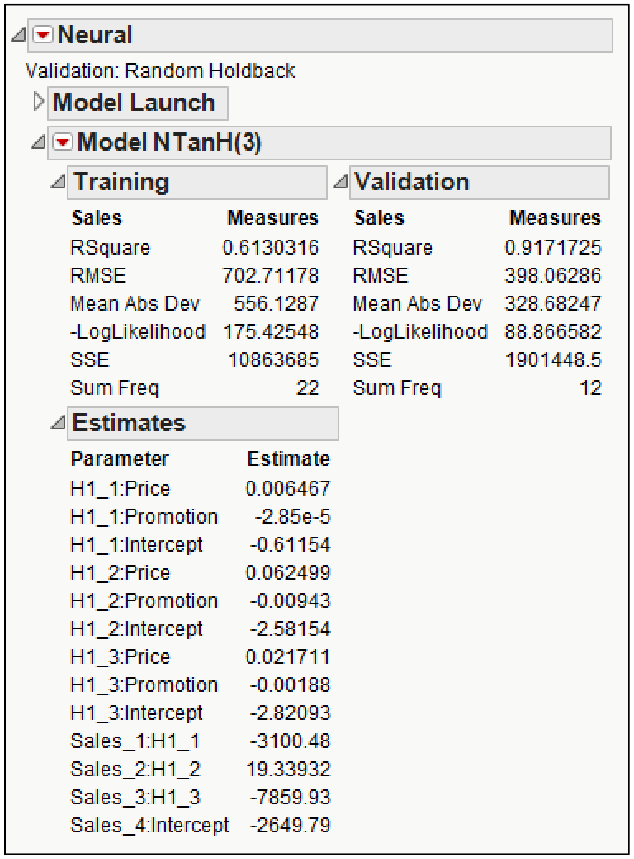
Actual	Predicted	
Upgraded	No	Yes
No	0.83333	0.16667
Yes	0.00000	1.00000

Estimates

Parameter	Estimate
H1_1:Purchases	0.048946
H1_1:Extra Cards:No	-0.96781
H1_1:Intercept	-1.79657
H1_2:Purchases	0.158715
H1_2:Extra Cards:No	-0.34419
H1_2:Intercept	-5.7199
Upgraded(No):H1_1	-0.67144
Upgraded(No):H1_2	-2.85399
Upgraded(No):Intercept	0.268264

Neural net example 2

"Predicting sales of energy bars based on price and promotion expenses" (same example used in regression tree)



Neural

Validation: Random Holdback

Model Launch

Model NTanH(3)

Training

Sales	Measures
RSquare	0.6130316
RMSE	702.71178
Mean Abs Dev	556.1287
-LogLikelihood	175.42548
SSE	10863685
Sum Freq	22

Validation

Sales	Measures
RSquare	0.9171725
RMSE	398.06286
Mean Abs Dev	328.68247
-LogLikelihood	88.866582
SSE	1901448.5
Sum Freq	12

Estimates

Parameter	Estimate
H1_1:Price	0.006467
H1_1:Promotion	-2.85e-5
H1_1:Intercept	-0.61154
H1_2:Price	0.062499
H1_2:Promotion	-0.00943
H1_2:Intercept	-2.58154
H1_3:Price	0.021711
H1_3:Promotion	-0.00188
H1_3:Intercept	-2.82093
Sales_1:H1_1	-3100.48
Sales_2:H1_2	19.33932
Sales_3:H1_3	-7859.93
Sales_4:Intercept	-2649.79

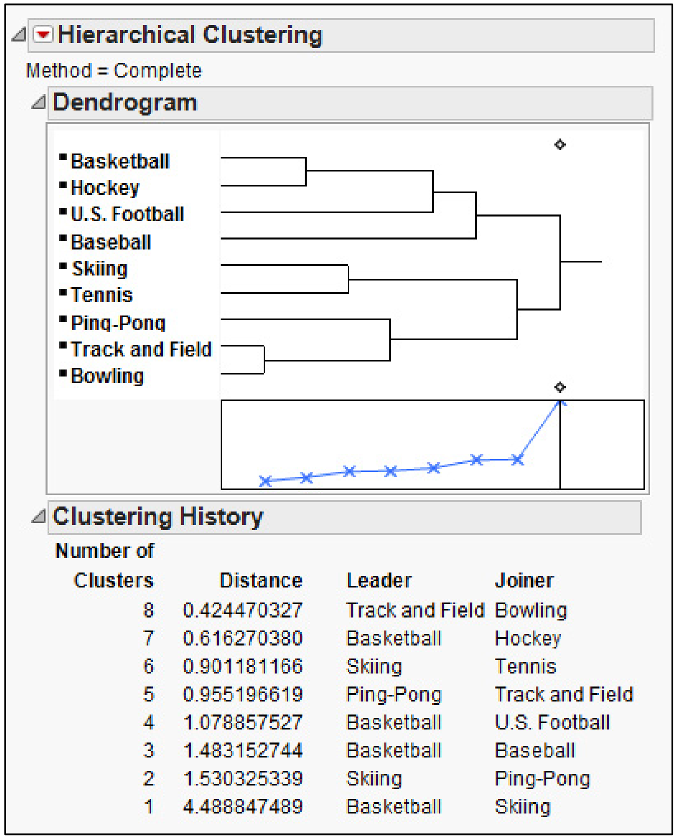
Cluster analysis

Classifies data into a sequence of groupings such that objects in each group are more alike other objects in their group than they are to objects found in other groups.

- Hierarchical clustering
- *k*-means clustering
- Distance measures
- Types of linkage between clusters

Cluster analysis example

“Perception of sports based on a survey of these attributes: movement speed, rules, team orientation, amount of contact”



Hierarchical Clustering

Method = Complete

Dendrogram

Clustering History

Number of Clusters	Distance	Leader	Joiner
8	0.424470327	Track and Field	Bowling
7	0.616270380	Basketball	Hockey
6	0.901181166	Skiing	Tennis
5	0.955196619	Ping-Pong	Track and Field
4	1.078857527	Basketball	U.S. Football
3	1.483152744	Basketball	Baseball
2	1.530325339	Skiing	Ping-Pong
1	4.488847489	Basketball	Skiing

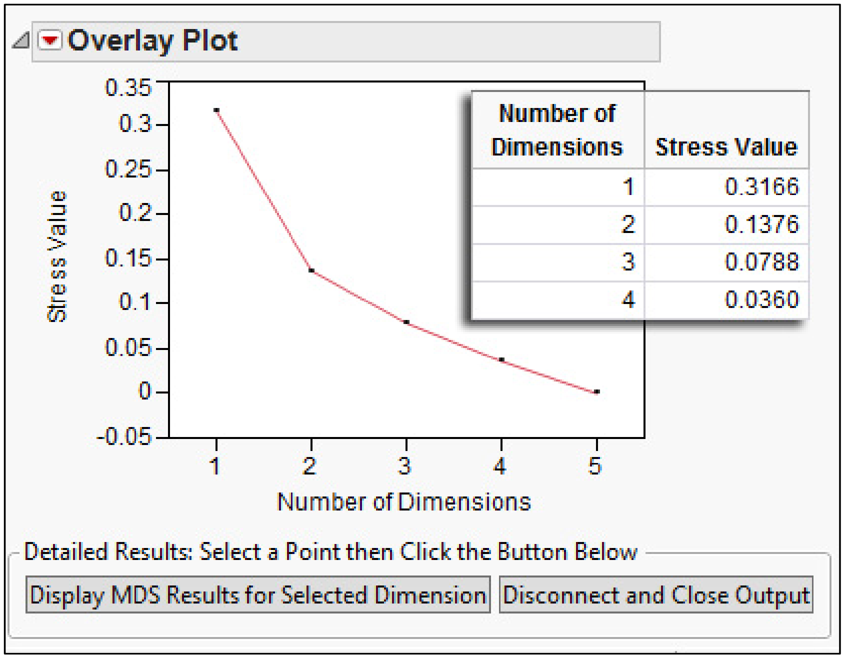
Multi-dimensional scaling

Visualizes objects in a two or more dimensional space, or map, with the goal of discovering patterns of similarities or dissimilarities among the objects.

- Types of multidimensional scaling
- Distance measures
- Stress statistic – measure of fit
- Challenge in interpreting dimensions

Multi-dimensional scaling example using JMP add-in

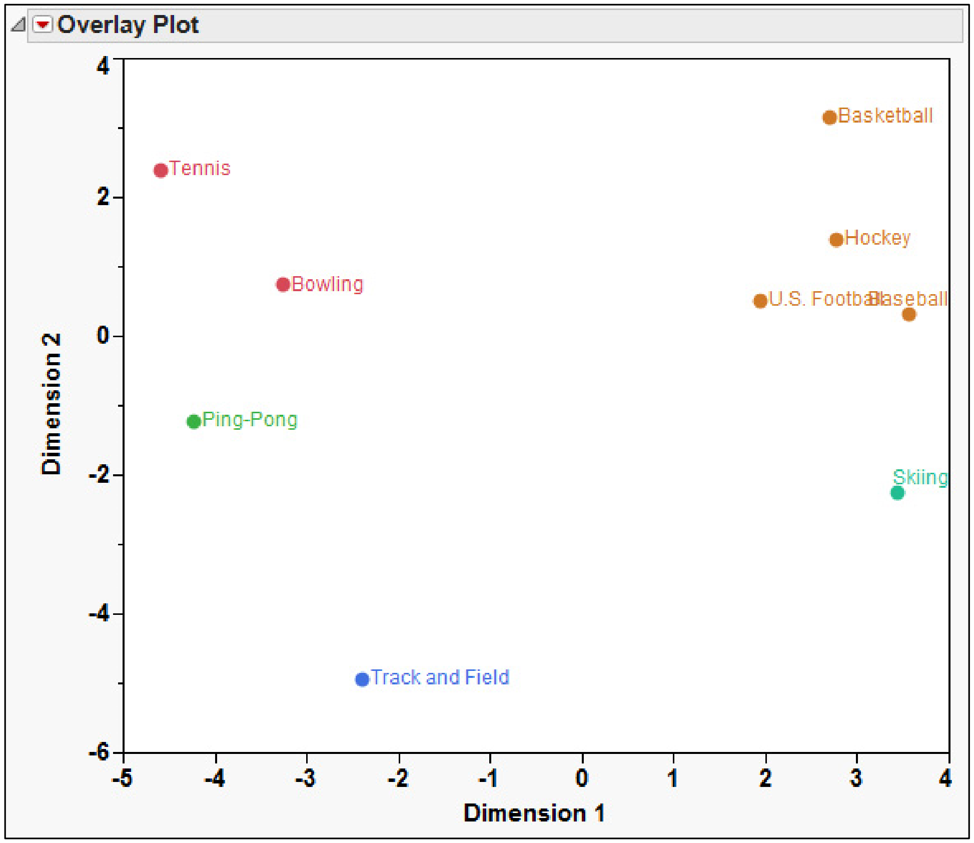
“Perception of sports based on a survey of these attributes: movement speed, rules, team orientation, amount of contact”


Overlay Plot
Stress Value
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
-0.05
1
2
3
4
5
Number of Dimensions
Detailed Results: Select a Point then Click the Button Below
Display MDS Results for Selected Dimension
Disconnect and Close Output

Number of Dimensions	Stress Value
1	0.3166
2	0.1376
3	0.0788
4	0.0360

Multi-dimensional scaling example using JMP add-in

“Perception of sports based on a survey of these attributes: movement speed, rules, team orientation, amount of contact”



Software Resources

- Microsoft Excel (latest versions equipped Apps for Office)
  - Good for selected dashboard elements (treemap, gauges, sparklines) and illustrating drill-down (with PivotTables) and subsetting (with Slicers)
  - Extend with third-party add-ins to perform logistic regression
- Tableau Public (web-based, free download)
  - Good for descriptive analytics (bullet graph, treemaps)
  - Drag-and-drop interface that can be taught in minutes
  - "Premium" version (not free) extends utility of software to many other methods, although this server-based version is more geared to business
- JMP
  - Many displays have drill-down built into them
  - Good for regression trees, neural nets, cluster analysis, and multidimensional scaling (with additional free add-in)
  - Requires SAS or R for some processing; user interface contains some quirks for new and casual users (most of which could be eliminated through the use of custom add-ins)
  - Future versions promise additional capabilities.

Can I Incorporate Any of This Into the Introductory Course?

- Could add some of the descriptive analytics into the introductory course
  - Drill down and subsetting
  - Perhaps one graph that summarize volume and velocity
  - Show-and-tell to illustrate information design and/or “sexiness” versus usefulness issue
- Could add binary logistic regression if your course covers multiple regression and mentions binary logistic regression, but this will not be feasible in most cases
- “Funny, you should ask that question….”

References

- Berenson, M. L., D. M. Levine, and K. A. Szabat. *Basic Business Statistics 13th edition*. Upper Saddle River: Pearson Education, forthcoming January 2014.
- Breiman, L., J. Friedman, C. J. Stone, and R. A. Olshen. *Classification and Regression Trees*. London: Chapman and Hall, 1984.
- Cox, T. F., and M. A. Cox. *Multidimensional Scaling, Second edition*. Boca Raton, FL: CRC Press, 2010.
- Everitt, B. S., S. Landau, and M. Leese. *Cluster Analysis, Fifth edition*. New York: John Wiley, 2011.
- Few, S. *Information Dashboard Design: Displaying Data for At-a-Glance Monitoring, Second edition*. Burlingame, CA: Analytics Press, 2013.
- Hakimpoor, H., K. Arshad, H. Tat, N. Khani, and M. Rahmandoust. "Artificial Neural Network Application in Management." *World Applied Sciences Journal*, 2011, 14(7): 1008–1019.
- R. Klimberg, and B. D. McCullough. *Fundamentals of Predictive Analytics with JMP*. Cary, NC: SAS Press. 2013
- Lindoff, G., and M. Berry. *Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management*. Hoboken, NJ: Wiley Publishing, Inc., 2011.
- Loh, W. Y. "Fifty years of classification and regression trees." *International Statistical Review*, 2013, in press
- Tufte, E. *Beautiful Evidence*. Cheshire, CT: Graphics Press, 2006.

Further Information or Contact

- Contact us at analytics@davidlevinestatistics.com
- Visit analytics.davidlevinestatistics.com for
  - Today's slides including references
  - A preview of some of our current work in this area
  - *Coming soon* WaldoLands.com
- Look for our (*very* occasional) tweets using #AnalyticsEducation