How Predictive Modelers Should use Data to Tell Data Stories

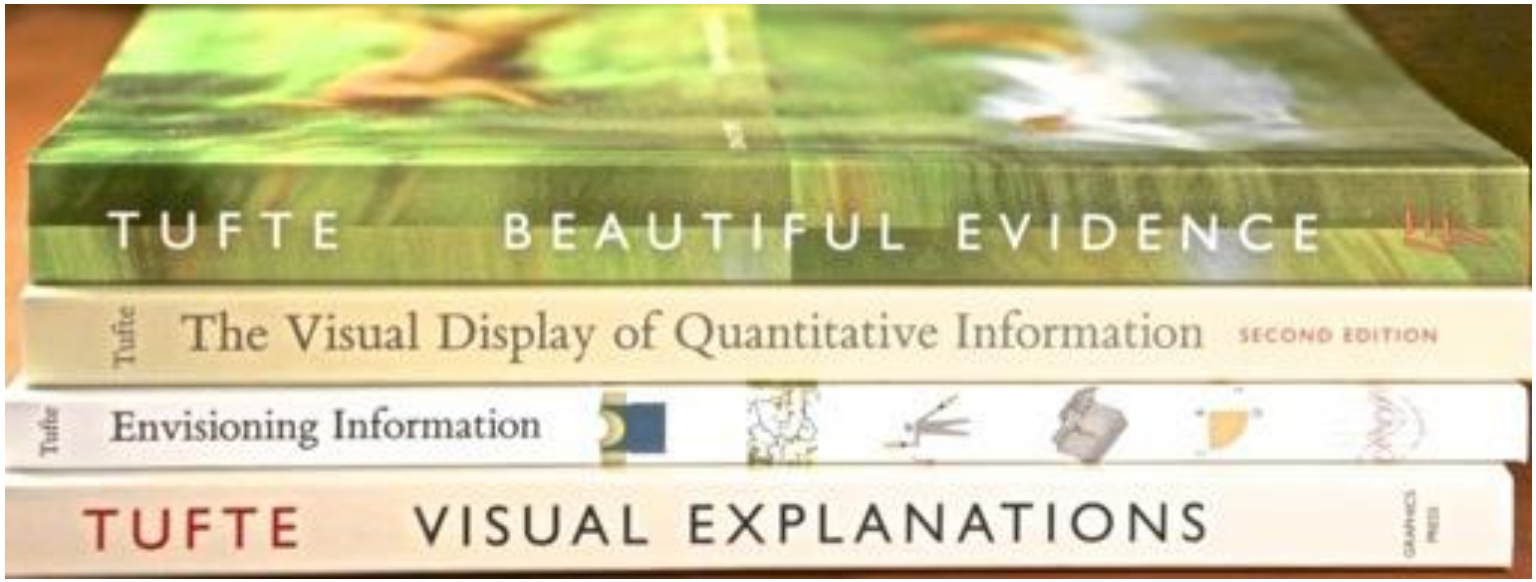
Dean Abbott
Co-Founder and Chief Data Scientist, SmarterHQ
President, Abbott Analytics
Twitter: @deanabb

smarterHQ

Telling Data Stories is Not Just Visualization

© Abbott Analytics, 2014-2018

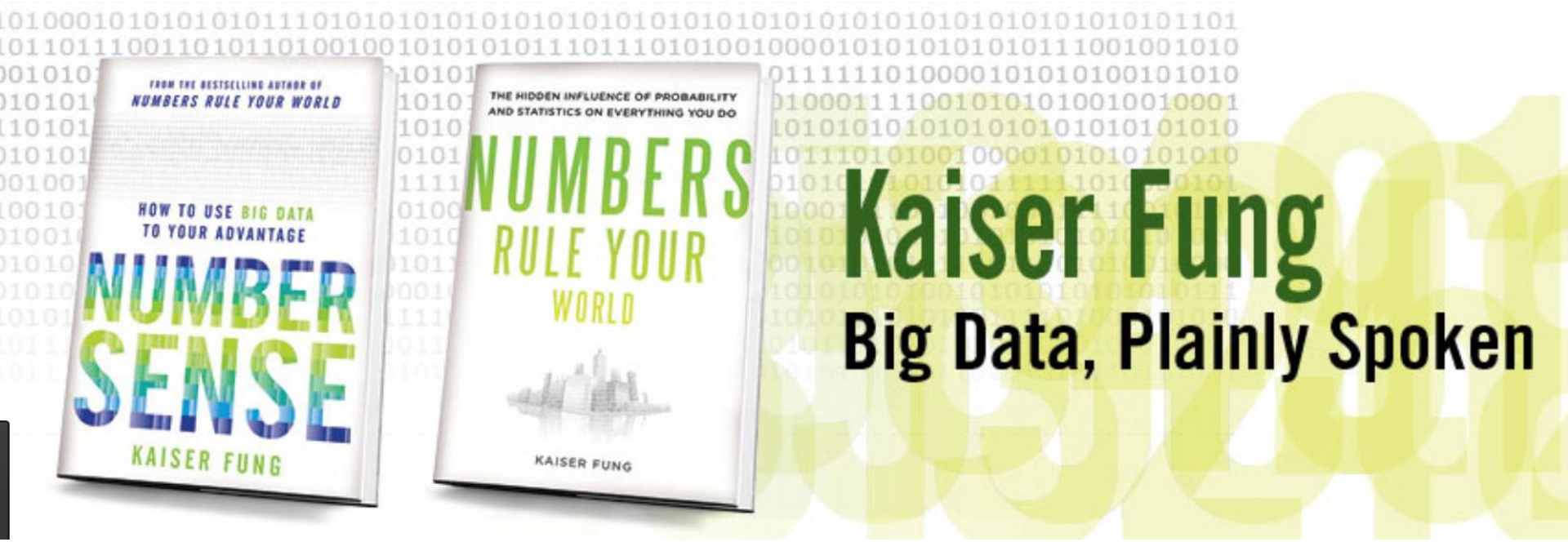
If you want in-depth
Data Visualization Thought Leadership



https://www.edwardtufte.com/tufte/

© Abbott Analytics, 2014-2018

If you want in-depth
Data Visualization Thought Leadership



http://www.kaiserfung.com/blogs/

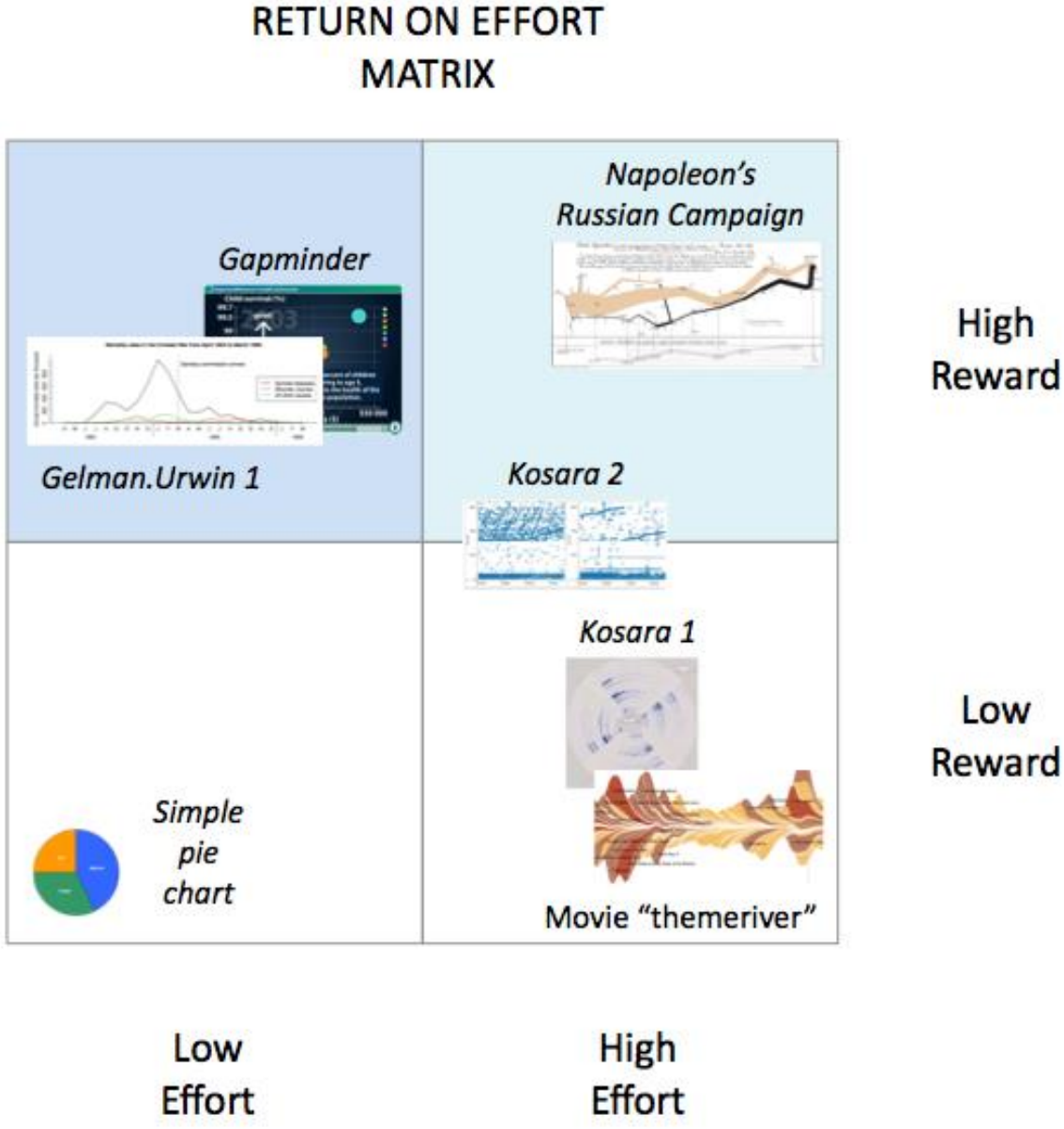
© Abbott Analytics, 2014-2018

Not All Visualization is Created Equally!

Fung: “Even acknowledged ‘perfect’ statistical graphics can require a high degree of effort.

[sometimes] the reward justifies the level of effort, and the return on effort is high.”

https://statisticsforum.wordpress.com/2011/07/31/one-difference-between-statistical-graphics-and-infoviz-is-the-return-on-effort/



© Abbott Analytics, 2014-2018

We Don't Have Time to Create Masterpieces

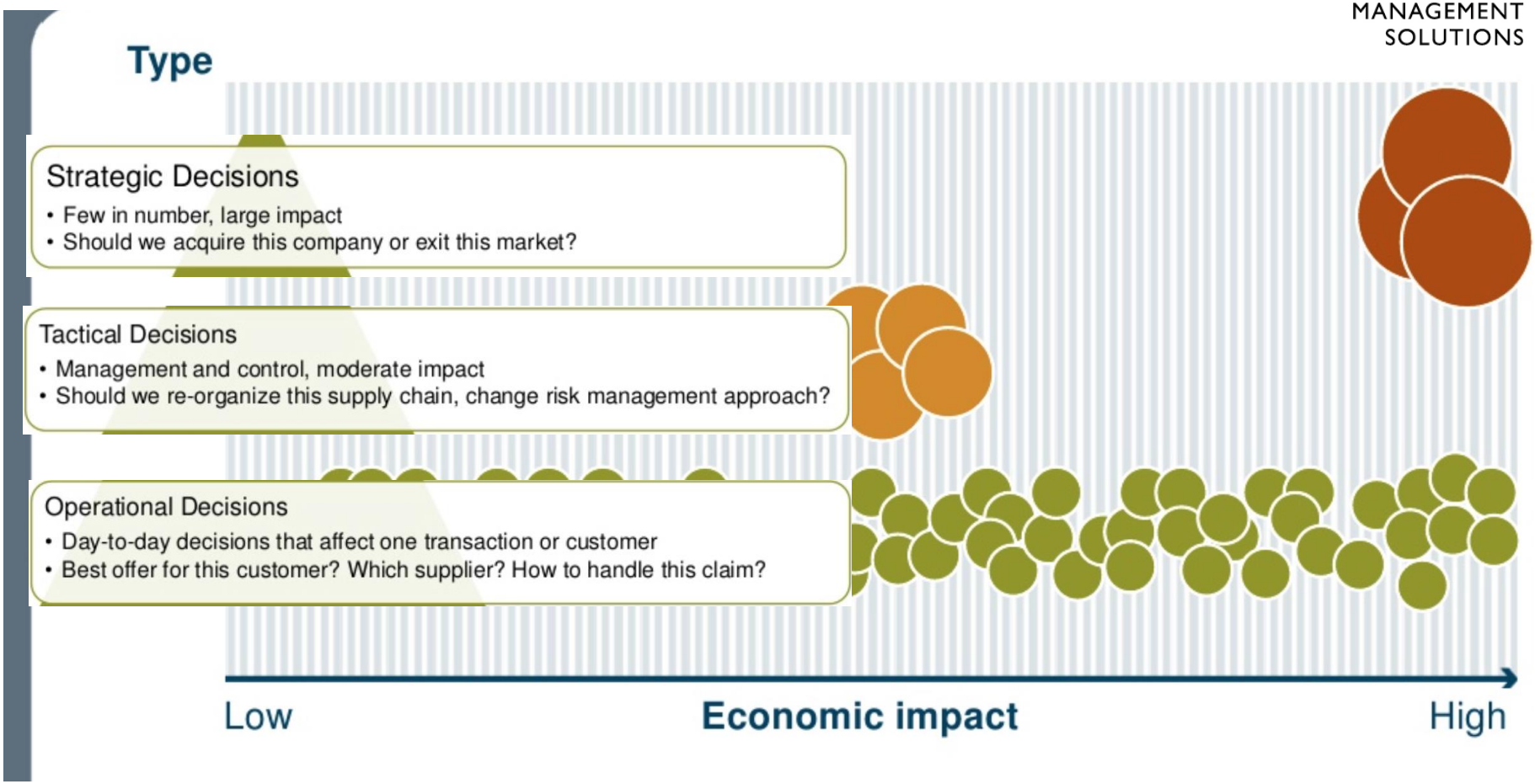
https://blog.udacity.com/2015/01/15-data-visualizations-will-blow-mind.html

https://en.wikipedia.org/wiki/Hans_Rosling

© Abbott Analytics, 2014-2018

James Taylor:
Importance of (Frequent) Operational Decisions

DECISION MANAGEMENT SOLUTIONS



https://www.slideshare.net/jamet123/10-best-practices-in-operational-analytics-6871966

smarterHQ

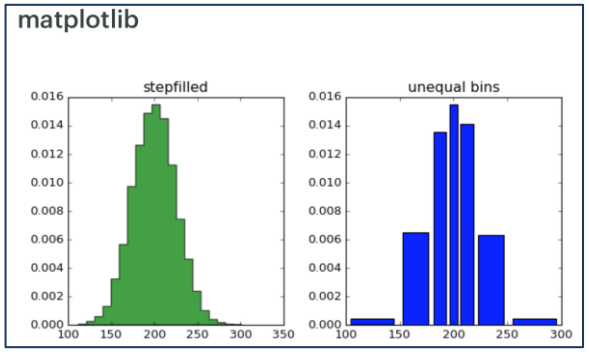
© Abbott Analytics, 2014-2018

Principle 1: Know Your Building Blocks (And Use the Good Ones!)

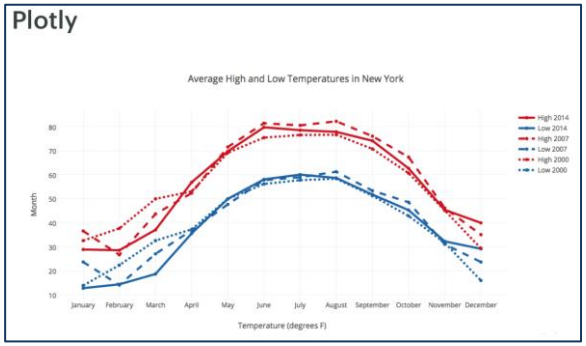
© Abbott Analytics, 2014-2018

Most of us Create These…

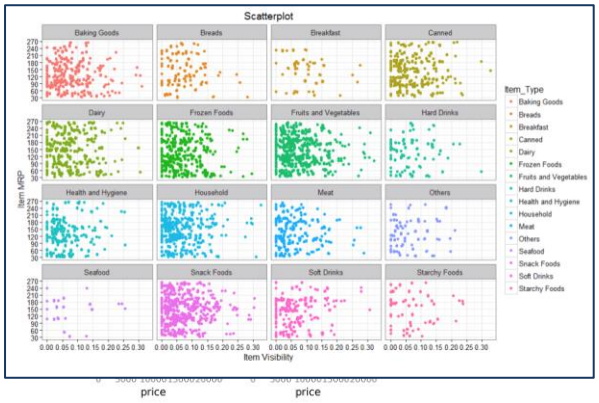
matplotlib



Plotly

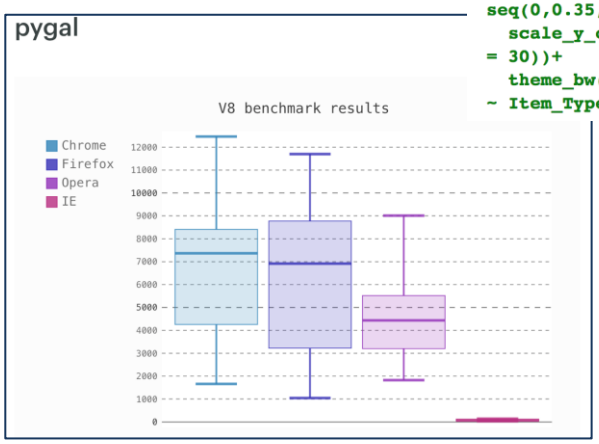


ggplot




```
import numpy as np
import matplotlib.pyplot as plt

np.random.seed(0)

mu = 200
sigma = 25
x = np.random.normal(mu, sigma, size=100)

fig, (ax0, ax1) = plt.subplots(ncols=2, figsize=(8, 4))

ax0.hist(x, 20, normed=1, histtype='stepfilled', facecolor='g', alpha=0.75)
ax0.set_title('stepfilled')

# Create a histogram by providing the bin edges (unequally spaced).
bins = [100, 150, 180, 195, 205, 220, 250, 300]
ax1.hist(x, bins, normed=1, histtype='bar', rwidth=0.8)
ax1.set_title('unequal bins')

fig.tight_layout()
plt.show()
```

pygal

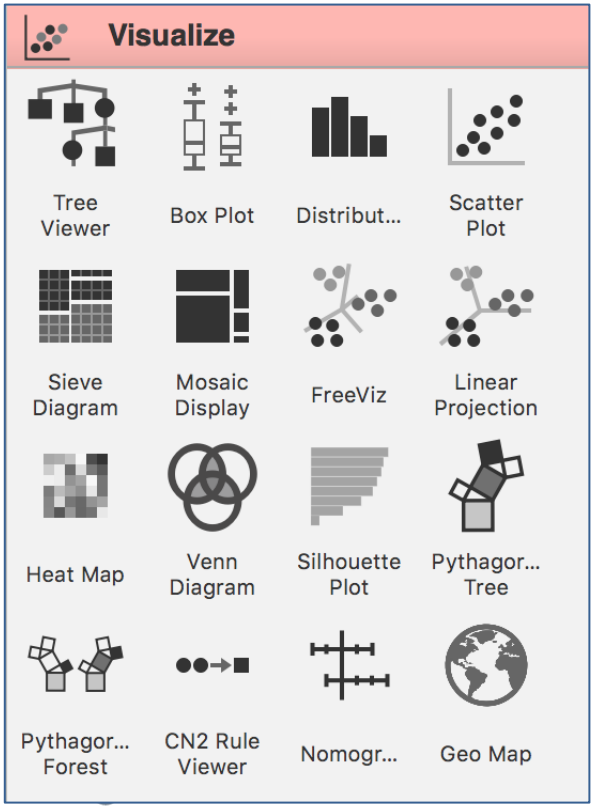


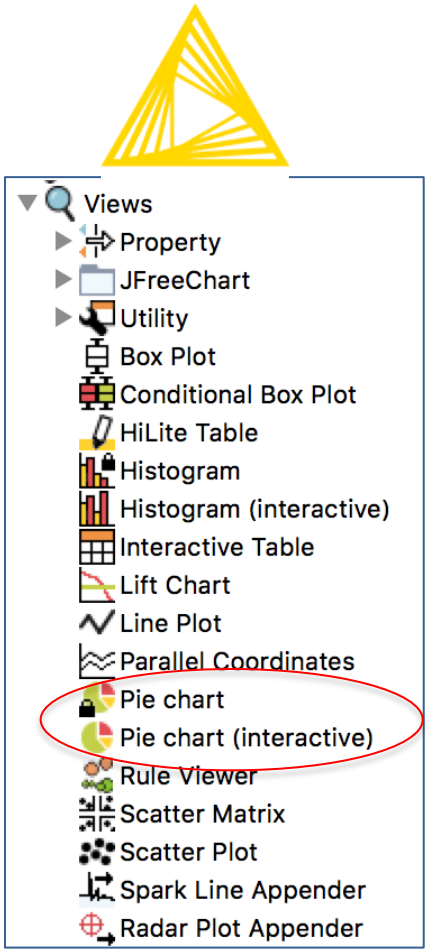

```
ggplot(train, aes(Item_Visibility, Item_MRP)) +
geom_point(aes(color = Item_Type)) +
  scale_x_continuous("Item Visibility", breaks =
seq(0,0.35,0.05))+
  scale_y_continuous("Item MRP", breaks = seq(0,270,by
= 30))+
  theme_bw() + labs(title="Scatterplot") + facet_wrap(
~ Item_Type)
```

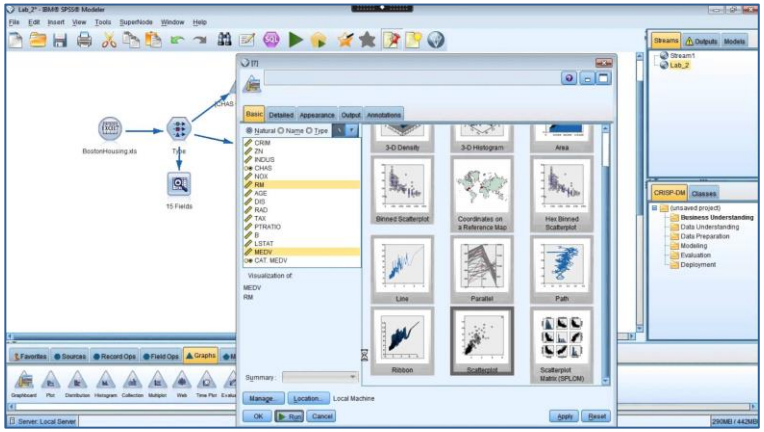
smarterHQ

© Abbott Analytics, 2014-2018

Or Use These…

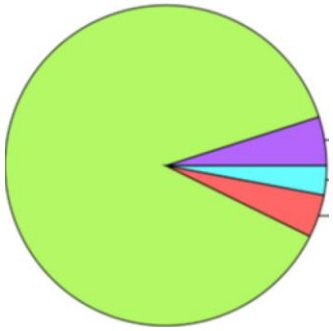






© Abbott Analytics, 2014-2018

Pie Charts

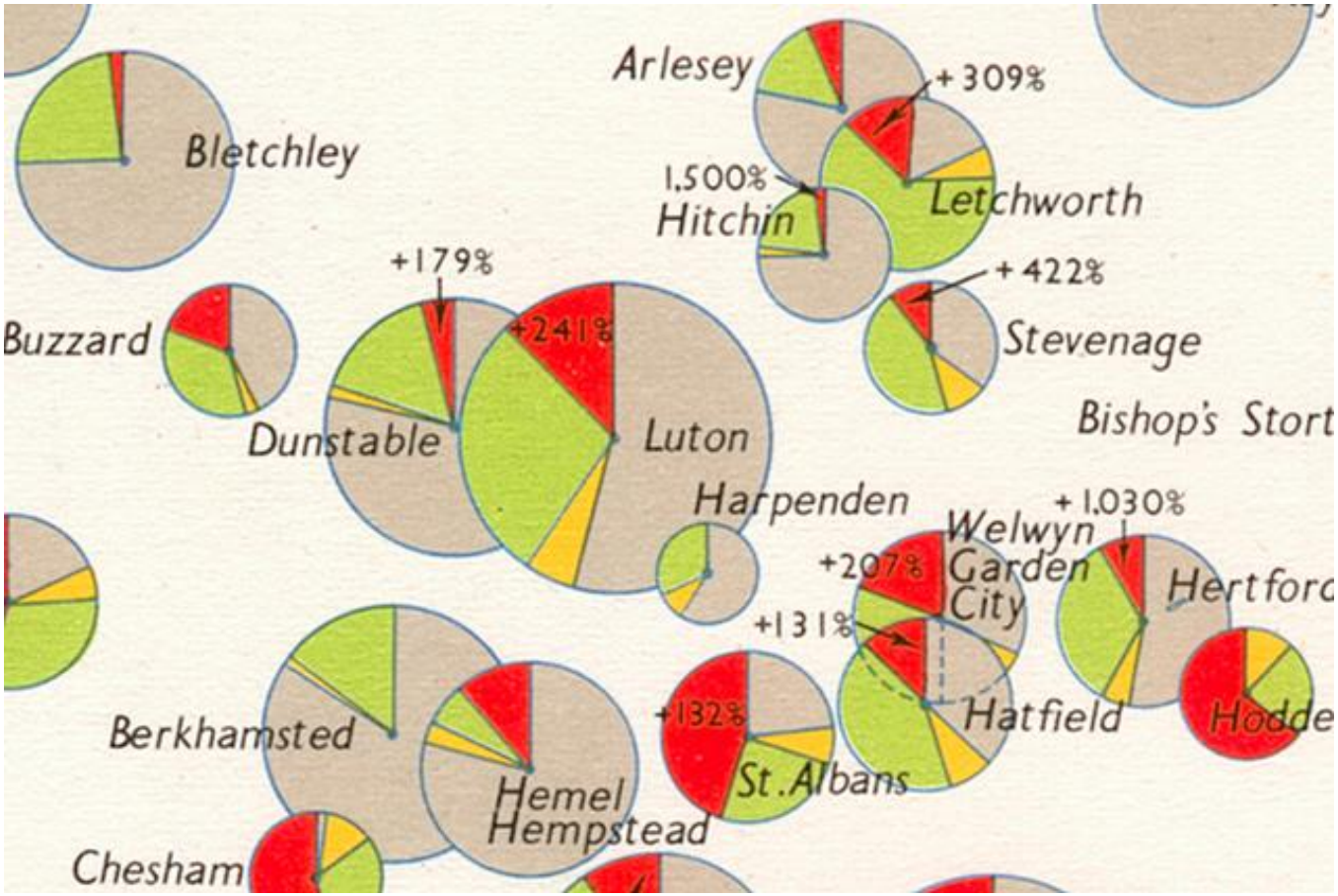


- Opinions of experts
  - Edward Tufte: “One of the prevailing orthodoxies of this forum - one to which I whole-heartedly subscribe - is that **pie charts are bad and that the only thing worse than one pie chart is lots of them**.”

    https://www.edwardtufte.com/bboard/q-and-a-fetch-msg?msg_id=00018S

© Abbott Analytics, 2014-2018

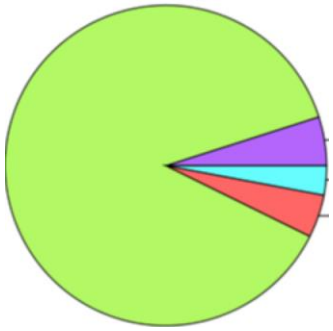
Tufte Example



https://www.edwardtufte.com/bboard/q-and-a-fetch-msg?msg_id=00018S

© Abbott Analytics, 2014-2018

Pie Charts

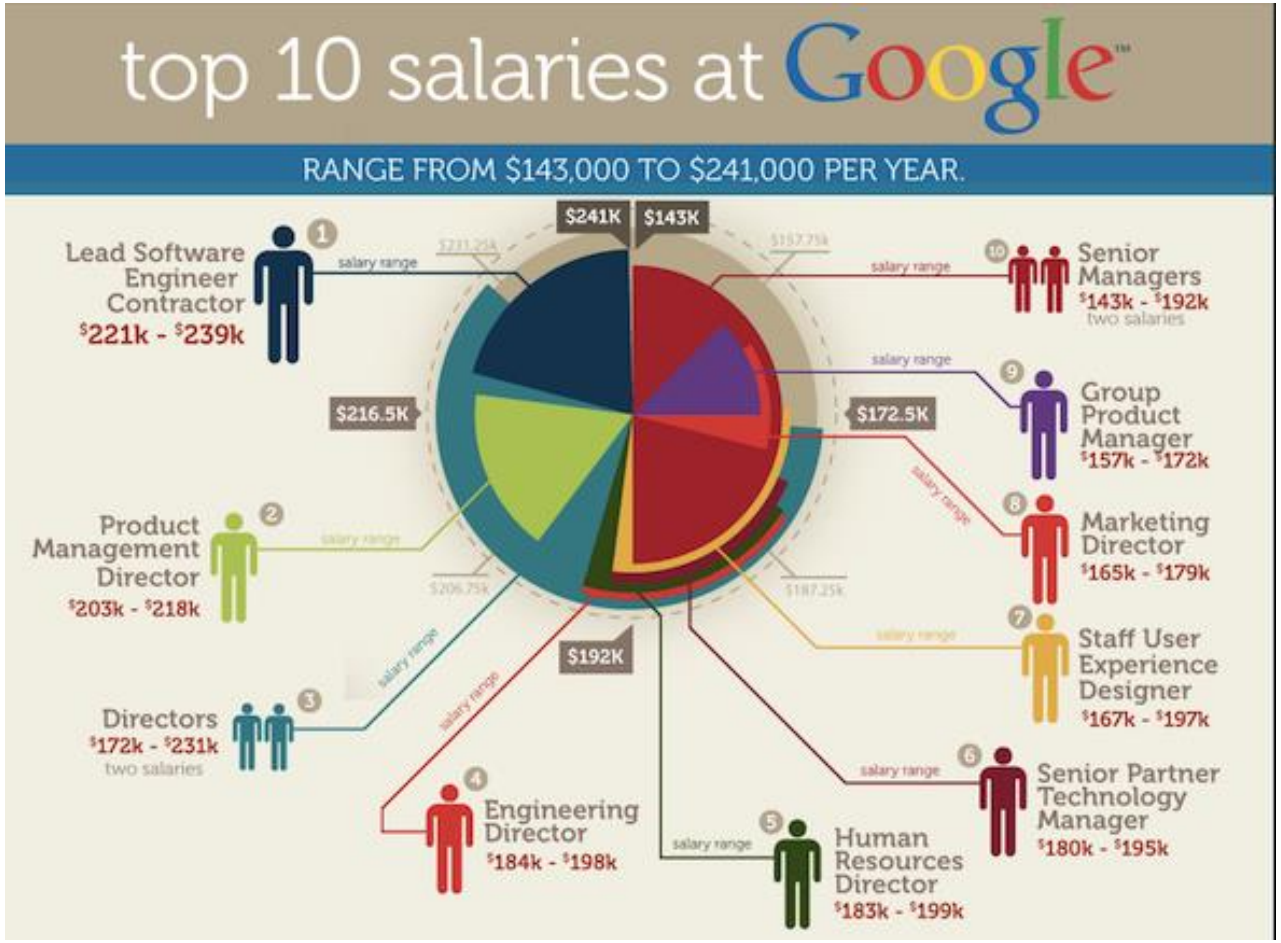


- Opinions of experts
  - Kaiser Fung: “The world would be a better place if **pie charts were banned**”

    http://junkcharts.typepad.com/junk_charts/2011/10/the-massive-burden-of-pie-charts.html

© Abbott Analytics, 2014-2018

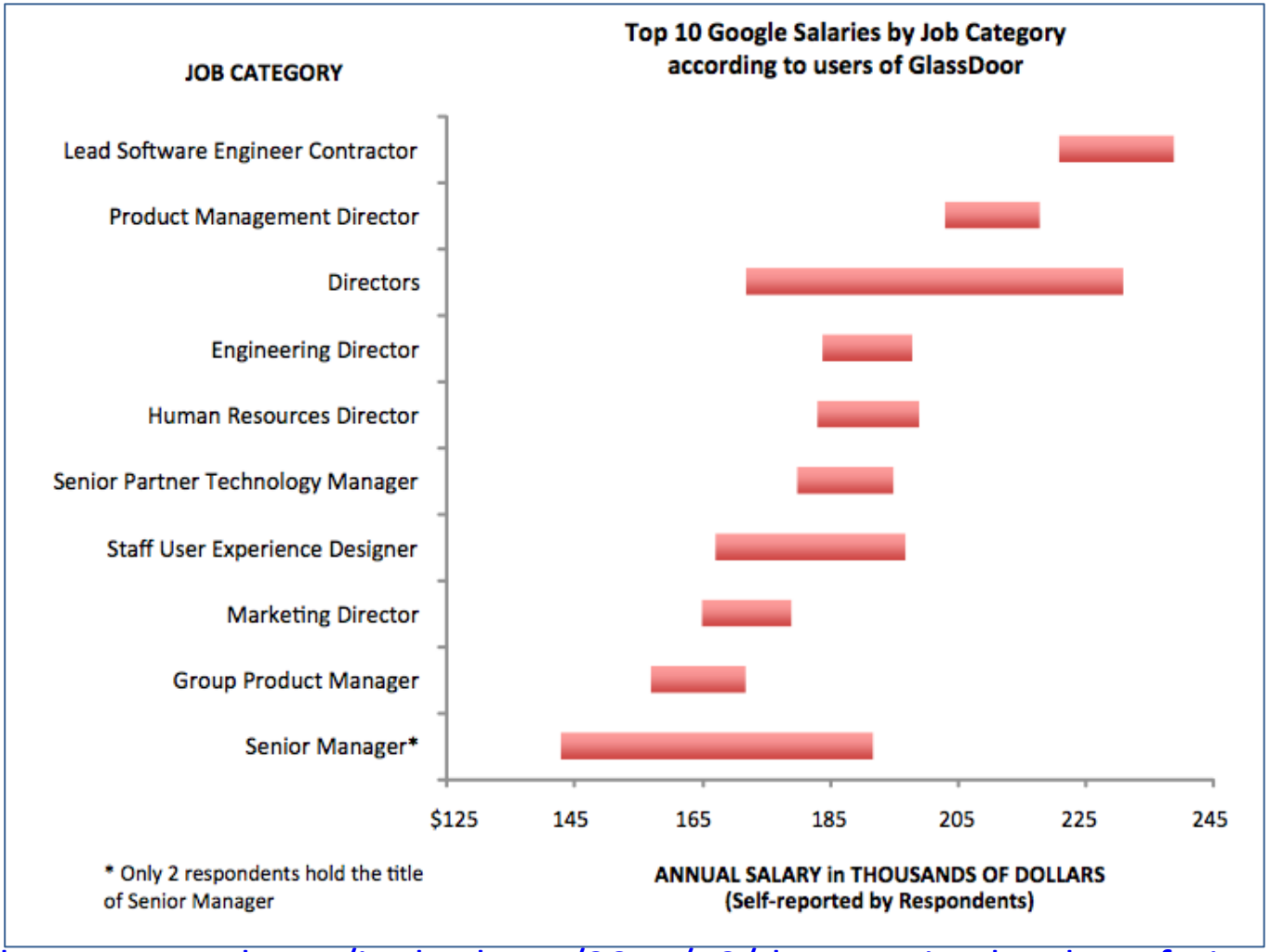
Sample Pie chart story from Fung



http://junkcharts.typepad.com/junk_charts/2011/10/the-massive-burden-of-pie-charts.html

© Abbott Analytics, 2014-2018

Information From Last Pie Chart Clearer as a Bar Chart

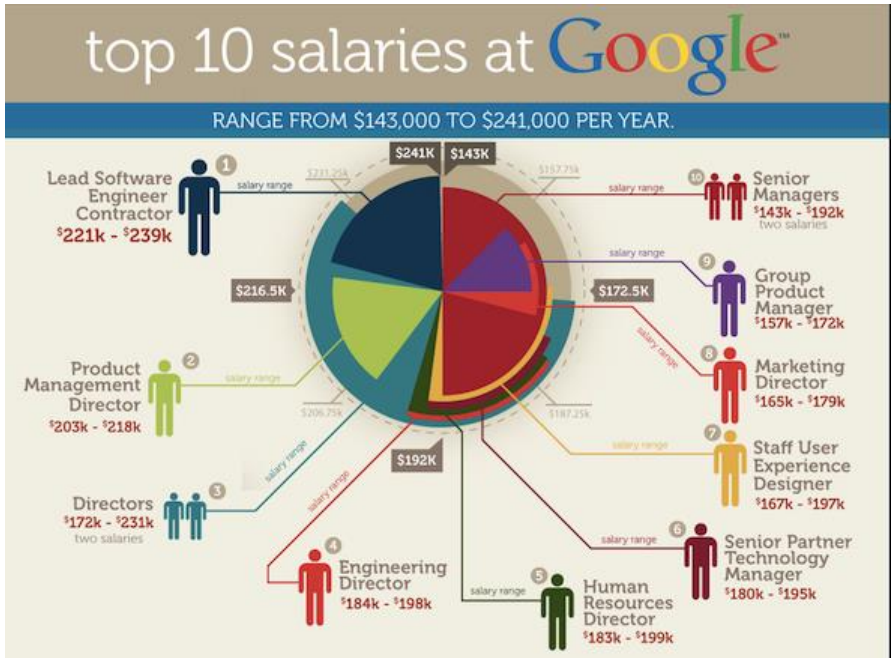


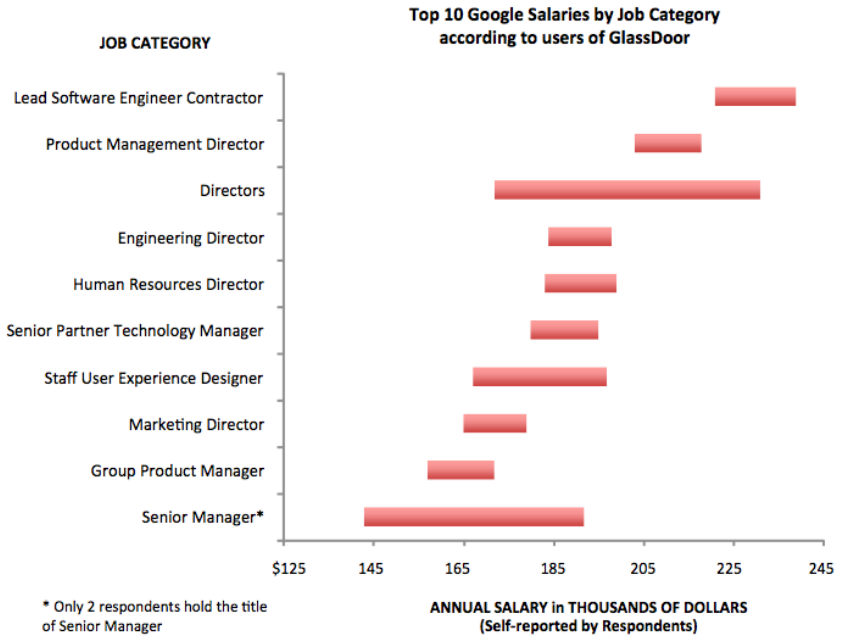
http://junkcharts.typepad.com/junk_charts/2011/10/the-massive-burden-of-pie-charts.html

smarterHQ

© Abbott Analytics, 2014-2018

Bringing A Tufte Concept into this Deck: Comparisons are Best Made on the Same Page



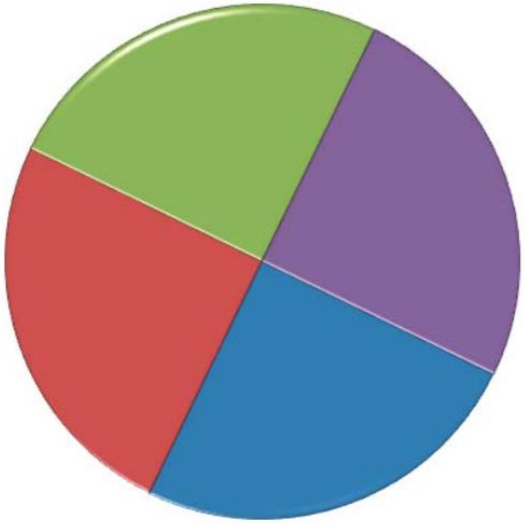


http://junkcharts.typepad.com/junk_charts/2011/10/the-massive-burden-of-pie-charts.html

© Abbott Analytics, 2014-2018

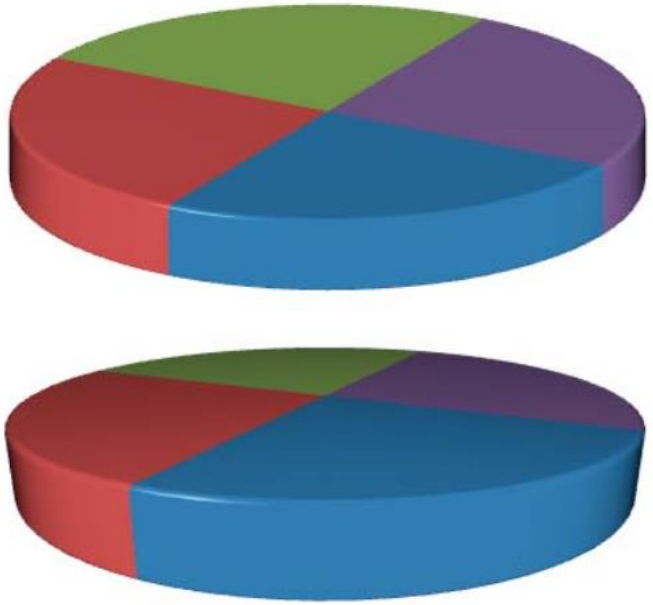
A Bad Idea: 3-D Pie Charts

"**the only thing worse than a pie chart is a 3-D pie chart**" – Dean Abbott



Color	Size
Green	25%
Blue	25%
Red	25%
Purple	25%

What is the relative size of the green wedge?
What is the relative size of the blue wedge?

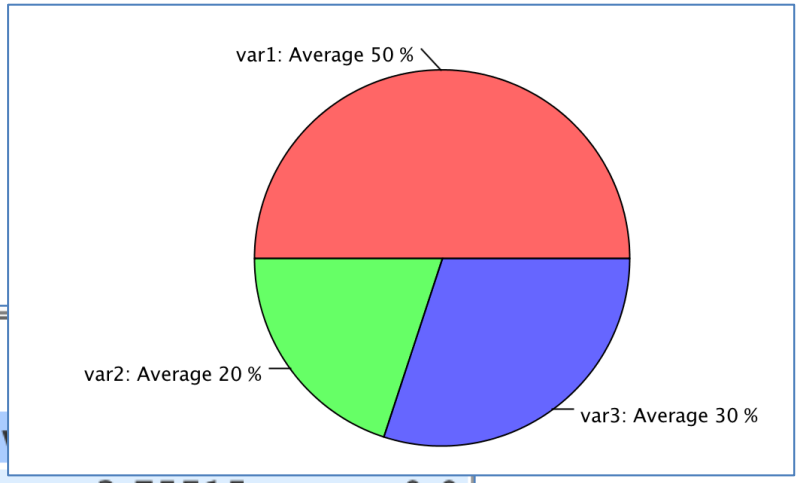


smarterHQ

© Abbott Analytics, 2014-2018

Pie Charts to Show Relative Proportion (that sum to 1)

The message: var1 has the most influence in the regression model predictions



Statistics on Linear Regression

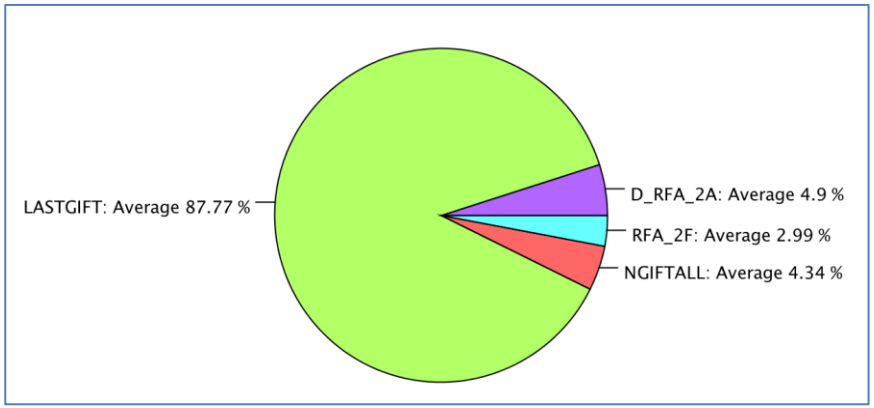
Variable	Coeff.	Std. Err.	t-v	
var1	0.5	1.33E-16	3.75E15	0.0
var2	0.2	1.34E-16	1.49E15	0.0
var3	0.3	1.34E-16	2.24E15	0.0
Intercept	-1.09E-13	4.68E-15	-23.3392	0.0

Multiple R-Squared: 1
Adjusted R-Squared: 1

© Abbott Analytics, 2014-2018

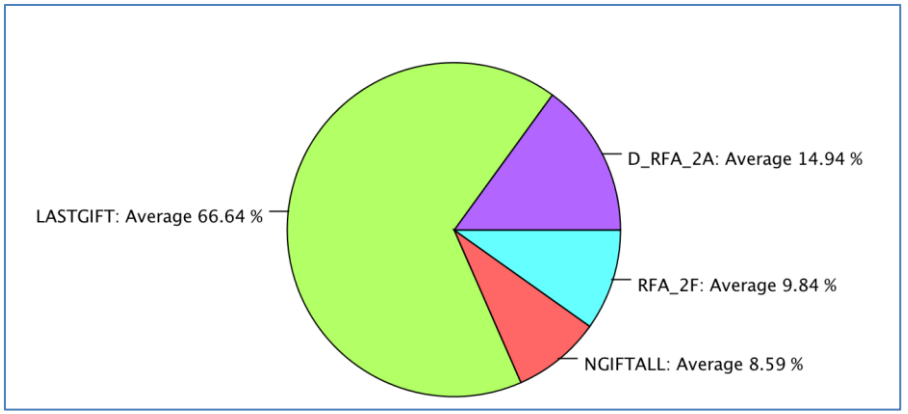
Pie Chart to Compare Multiple Measures (But One Wedge)

Coefficient



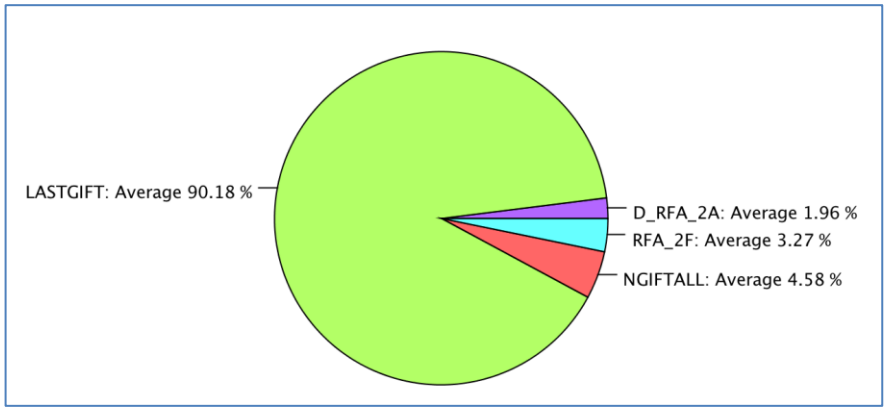
Use abs() for calculations

t-Proportion

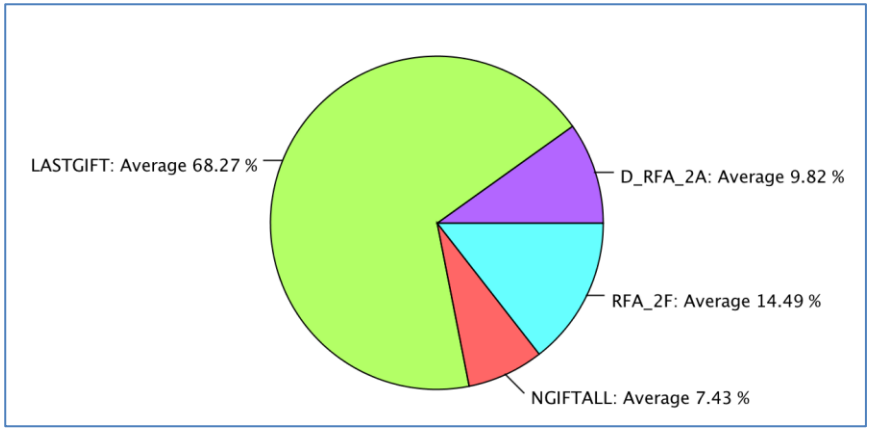


Use abs() for t-proportion calculations

Direct Proportion

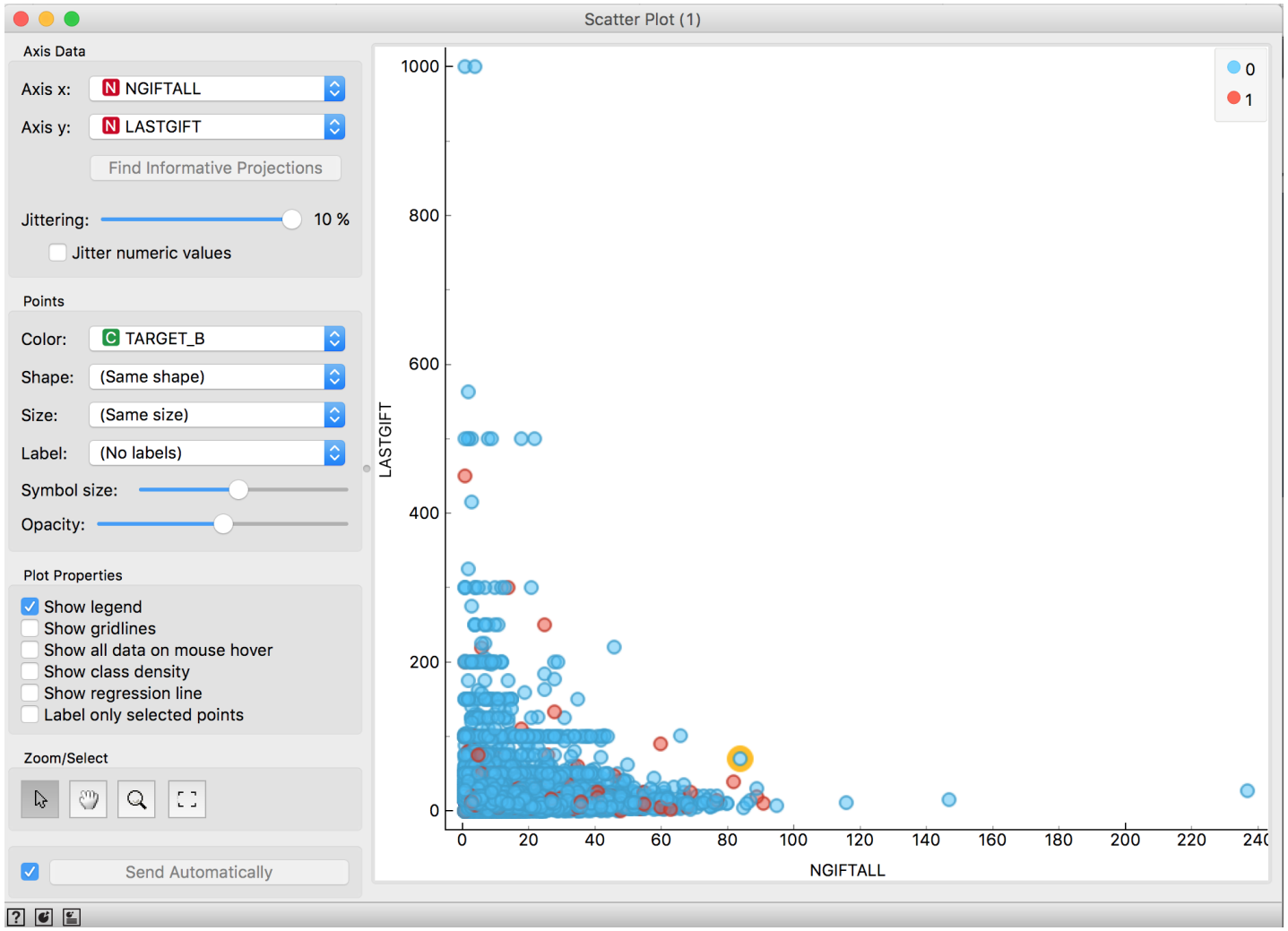


Input Shuffling Proportion



© Abbott Analytics, 2014-2018

Many Scatterplots end up looking like this: Skewed Distributions Create Lots of Empty Space



smarterHQ

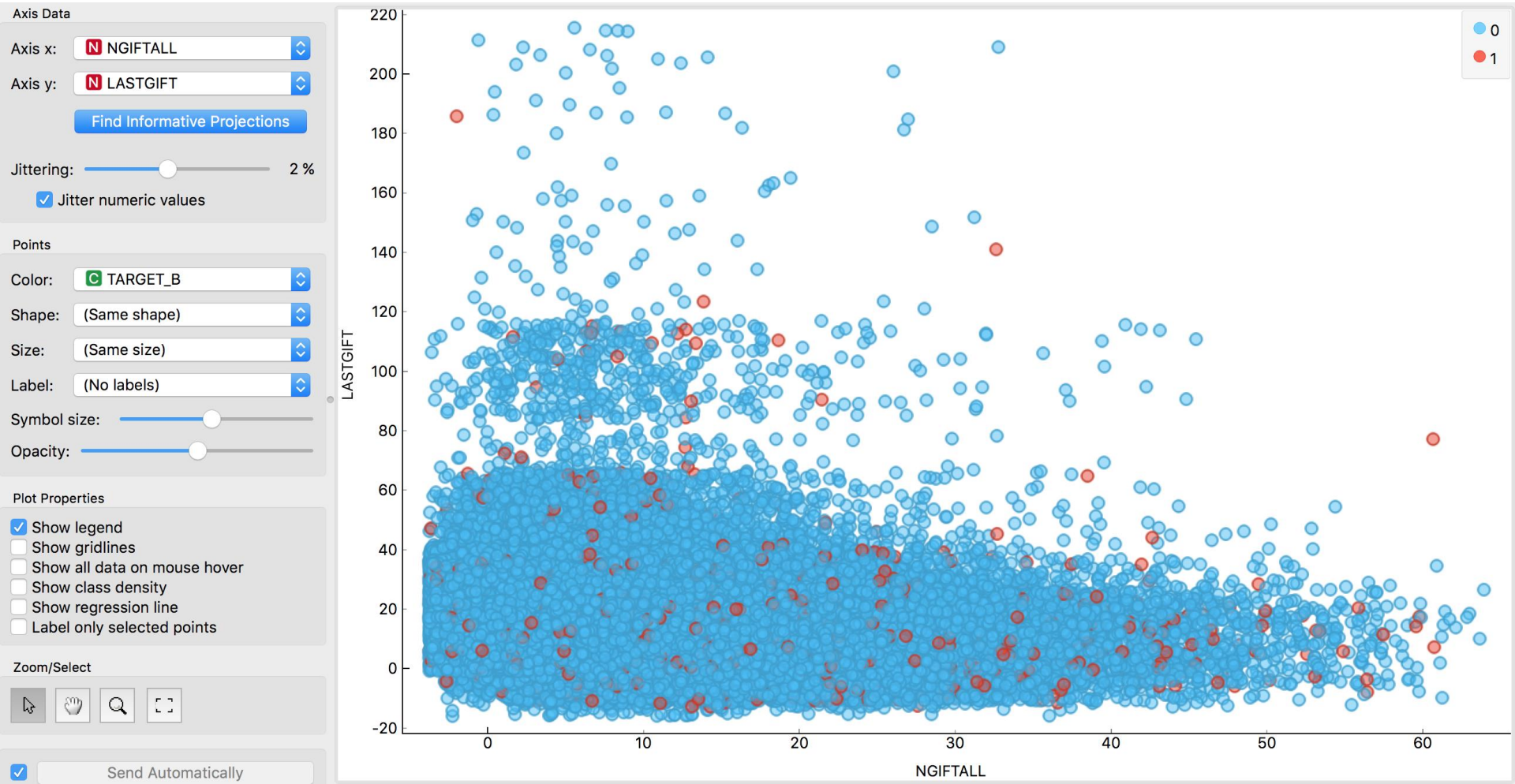
© Abbott Analytics, 2014-2018

Normalized Axes Make the Space More Dense, but at the Expense of Interpretability

smarterHQ

© Abbott Analytics, 2014-2018

Truncated Axes Make the Space Fuller and Intuitive, but Could Miss Interesting Patterns at the Edges



© Abbott Analytics, 2014-2018

Principle #2: Tell the Story You Need to Tell: But Be Transparent about the Bias

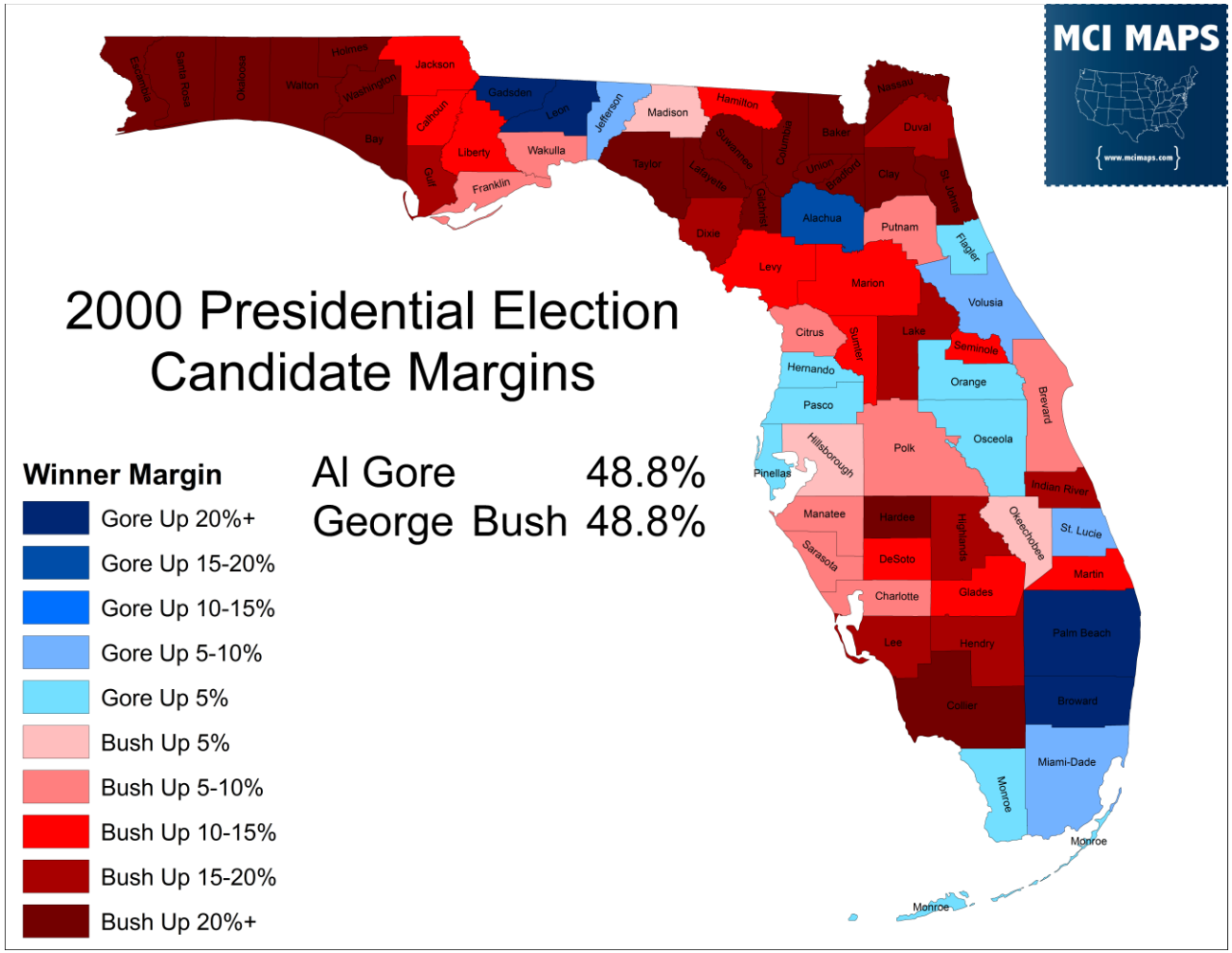
smarterHQ

© Abbott Analytics, 2014-2018

Think about the data, not the politics
What is the point of the visualization?

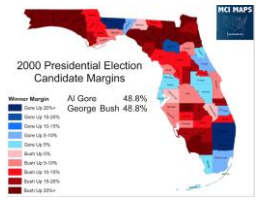
© Abbott Analytics, 2014-2018

Do We Really Want to Go Back to the 2000 Election?



© Abbott Analytics, 2014-2018

The Classic Scatterplot of FL Results: Can you Guess Which Point is Palm Beach Cty?



 # Buchanan Votes By County

3,500
3,250
3,000
2,750
2,500
2,250
2,000
1,750
1,500
1,250
1,000
750
500
250
0

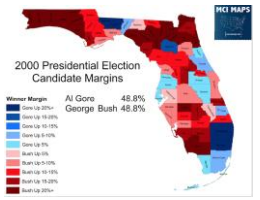
0 25,000 50,000 75,000 100,000 125,000 150,000 175,000 200,000 225,000 250,000 275,000 300,000

 # Bush Votes by County

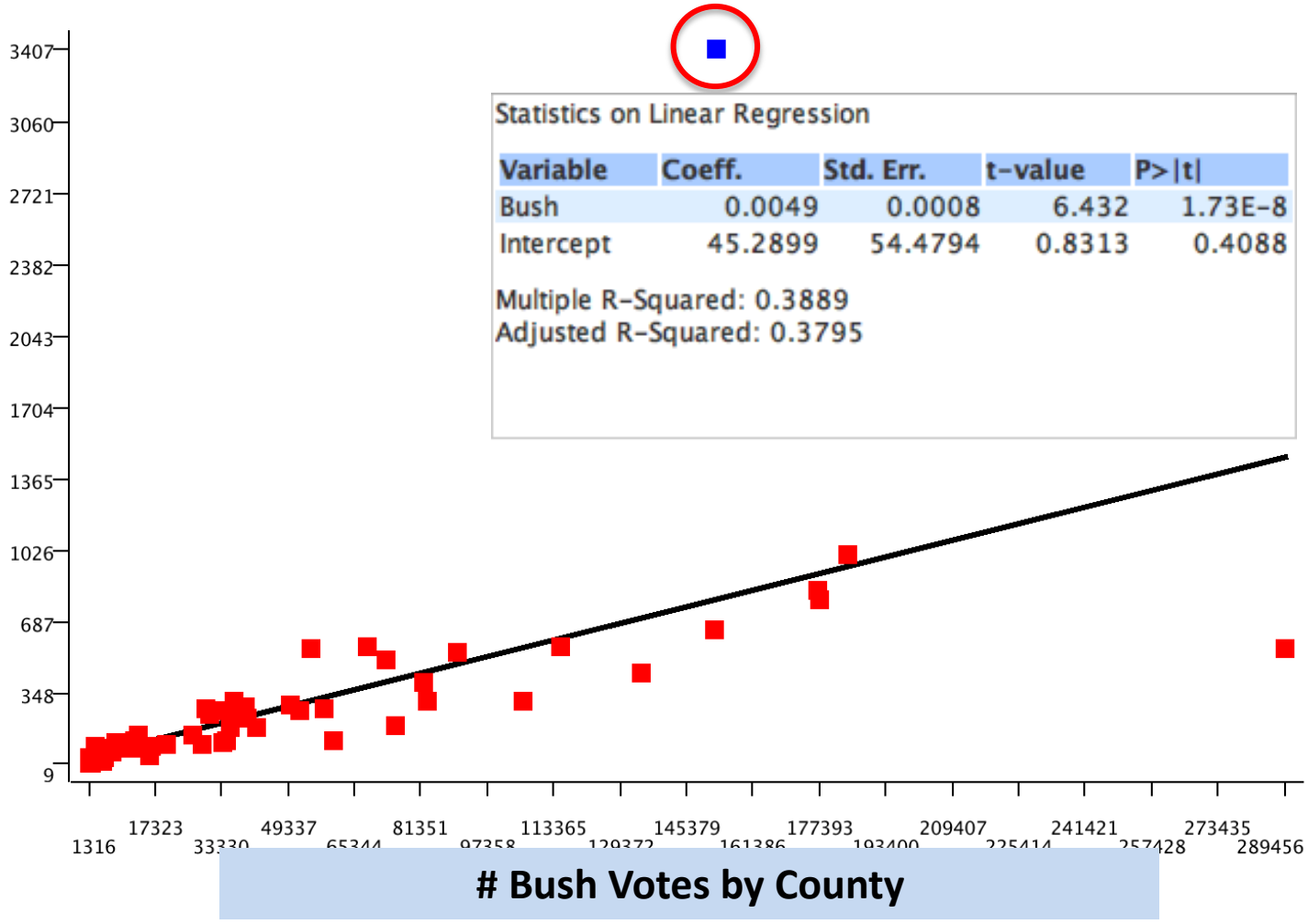
smarterHQ

© Abbott Analytics, 2014-2018

How Bad is this Outlier?



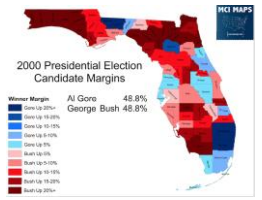
 # Buchanan Votes By County



smarterHQ

© Abbott Analytics, 2014-2018

Remove the Outlier, and Refit



 # Buchanan Votes By County

Palm Beach actual: 3407
Palm Beach predicted: 598

Statistics on Linear Regression

Variable	Coeff.	Std. Err.	t-value	P>|t|
Bush	0.0035	0.0003	13.9226	0.0
Intercept	65.5735	17.3304	3.7837	0.0003

Multiple R-Squared: 0.7518
Adjusted R-Squared: 0.7479

3407
3060
2721
2382
2043
1704
1365
1026
687
348
9

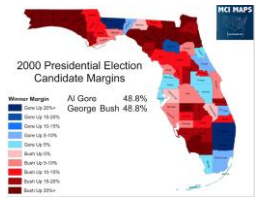
1316 17323 33330 49337 65344 81351 97358 113365 129372 145379 161386 177393 193400 209407 225414 241421 257428 273435 289456

 # Bush Votes by County

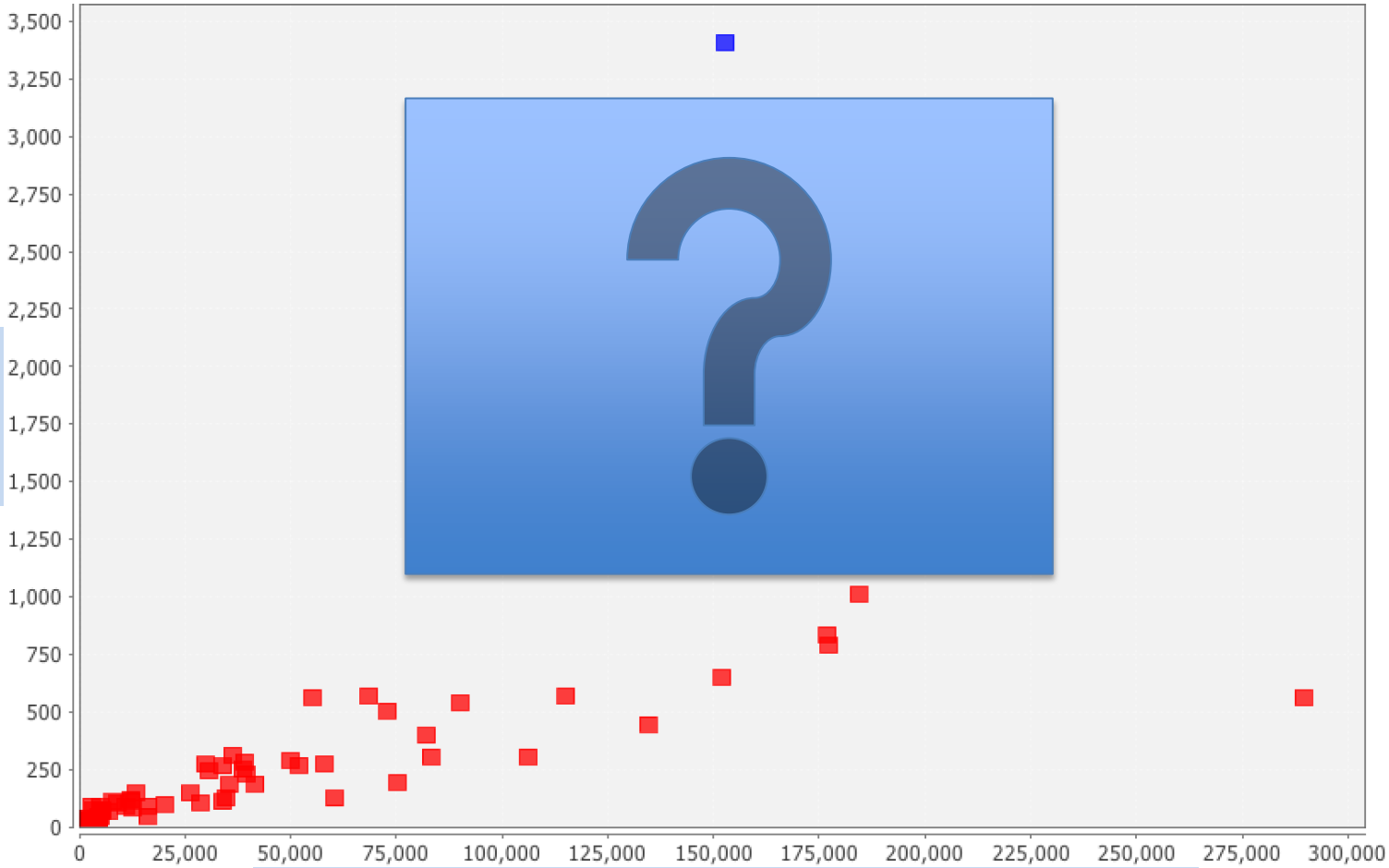
smarterHQ

© Abbott Analytics, 2014-2018

What Biases Could Exist in This Data? What Assumptions are We Making?



 # Buchanan Votes By County

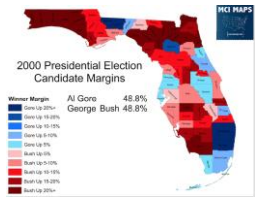


 # Bush Votes by County

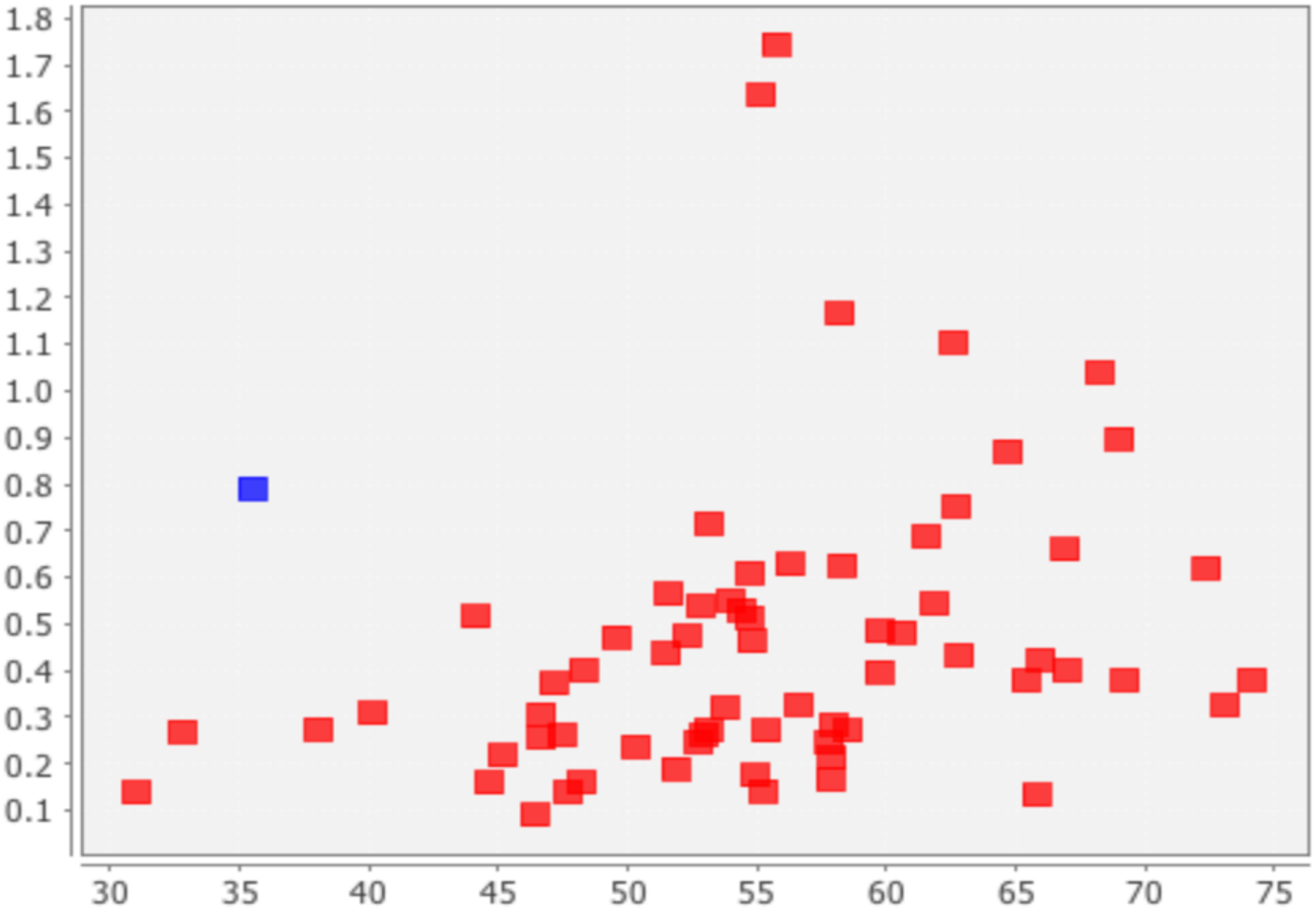
smarterHQ

© Abbott Analytics, 2014-2018

Normalizing Scales: Remove Magnitude Bias



% Buchanan Votes By County

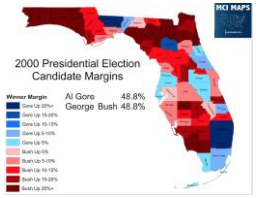


% Bush Votes by County

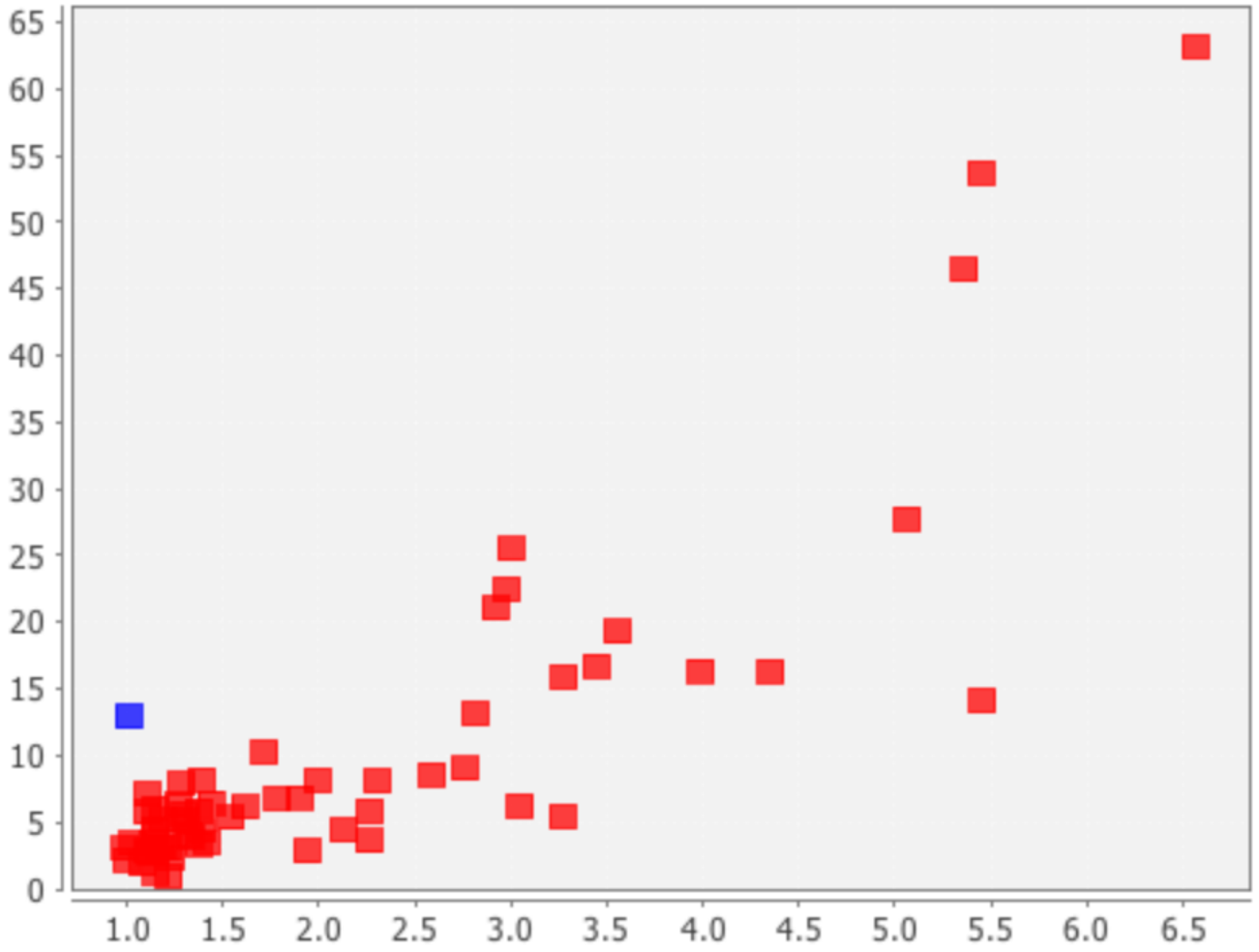
smarterHQ

© Abbott Analytics, 2014-2018

Adding “Expected” Results: Priors



% Buchanan Votes By County / % Reform Party Registrants

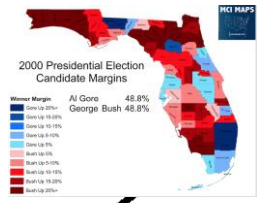


% Bush Votes by County / % Rep. Registrants

smarterHQ

© Abbott Analytics, 2014-2018

Refit



% Buchanan Votes By County / % Reform Party Registrants

Does Palm Beach Look Like An Outlier Now?

Statistics on Linear Regression

Variable	Coeff.	Std. Err.	t-value	P>|t|
Bush v R	7.5024	0.6041	12.4199	0.0
Intercept	-5.9228	1.4998	-3.9492	0.0002

Multiple R-Squared: 0.7166
Adjusted R-Squared: 0.712

63.123
55.951
50.951
45.951
40.951
35.951
30.951
25.951
20.951
15.951
10.951
5.951
0.951

0.988 1.488 1.988 2.488 2.988 3.488 3.988 4.488 4.988 5.488 5.988 6.488 6.988 7.488 7.988 8.488 8.988 9.488

% Bush Votes by County / % Rep. Registrants

smarterHQ

© Abbott Analytics, 2014-2018

Principle 3: Tell the Data Story so the Hearer Can Understand

© Abbott Analytics, 2014-2018

Tell the Data Story
so Your Boss Can Understand

© Abbott Analytics, 2014-2018

YMCA Survey Data

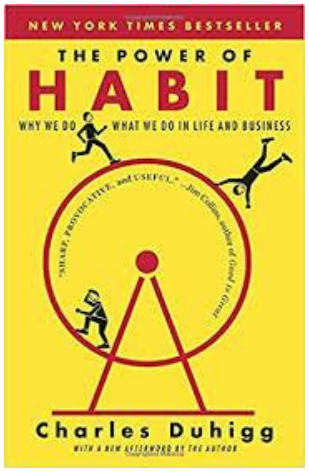
- 32,811 surveys, 57 questions
  - All survey questions in analysis had 4, 5, or 6 responses,
  - coded 1 for top/highest answer

- **Q1** - **Satisfaction = 1**: 31%

- Q48 - Recommend to Friend = 1: 54%

http://www.abbottanalytics.com/data-mining-case-study-2.php

http://media.salford-systems.com/presentation/A-More-Transparent-Interpretation-of-Health-Club-Surveys-YMCA.pdf



smarterHQ

© Abbott Analytics, 2014-2018

Regression Analysis Revealed Seven Question That Were the Most Important

- Empirical comparison: Factors as inputs vs. Top-loading question in factor as input
  - Top-loading or most interesting question on factor as representative of that factor produced slightly better models
  - Use of top-loading question makes final model more easily understood
  - This flies in the face of traditional theory, but worked better operationally
- Final regression model contained these fields:

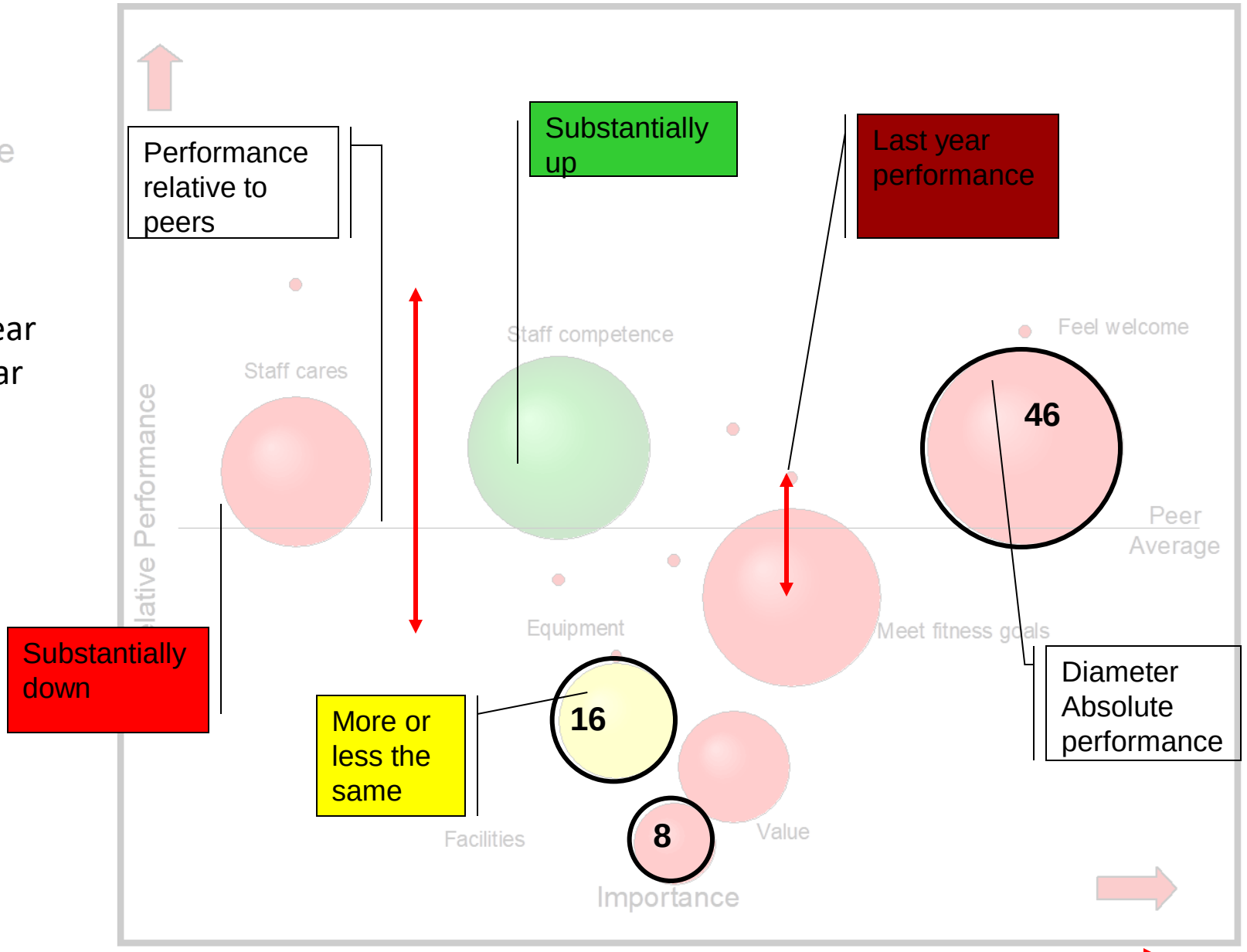
Database_Question_Number	Factor	Description
Q13	facilities 1	Facilities clean
Q18	equipment 1	Equipment maintained
Q22	value 1	Value for the Money
Q25	relationships 1	Feel Welcome
Q44	goals 1	Y Helps meet fitness goals
Q6	staff 1	Competent Staff
Q9	staff 2	Staff Cares about Well-being

© Abbott Analytics, 2014-2018

drivers of excellence

Current year vs. last year

- Better
- Same
- Worse



Importance

© Abbott Analytics, 2014-2018

I Know….Now

Kaiser Fung: “The **bubble chart is one of the most hopeless data graphics ever invented**. It is sometimes useful for conceptual charts but trying to express data with it is a lost cause.”

http://junkcharts.typepad.com/junk_charts/2013/03/blowing-the-whistle-at-bubble-charts.html

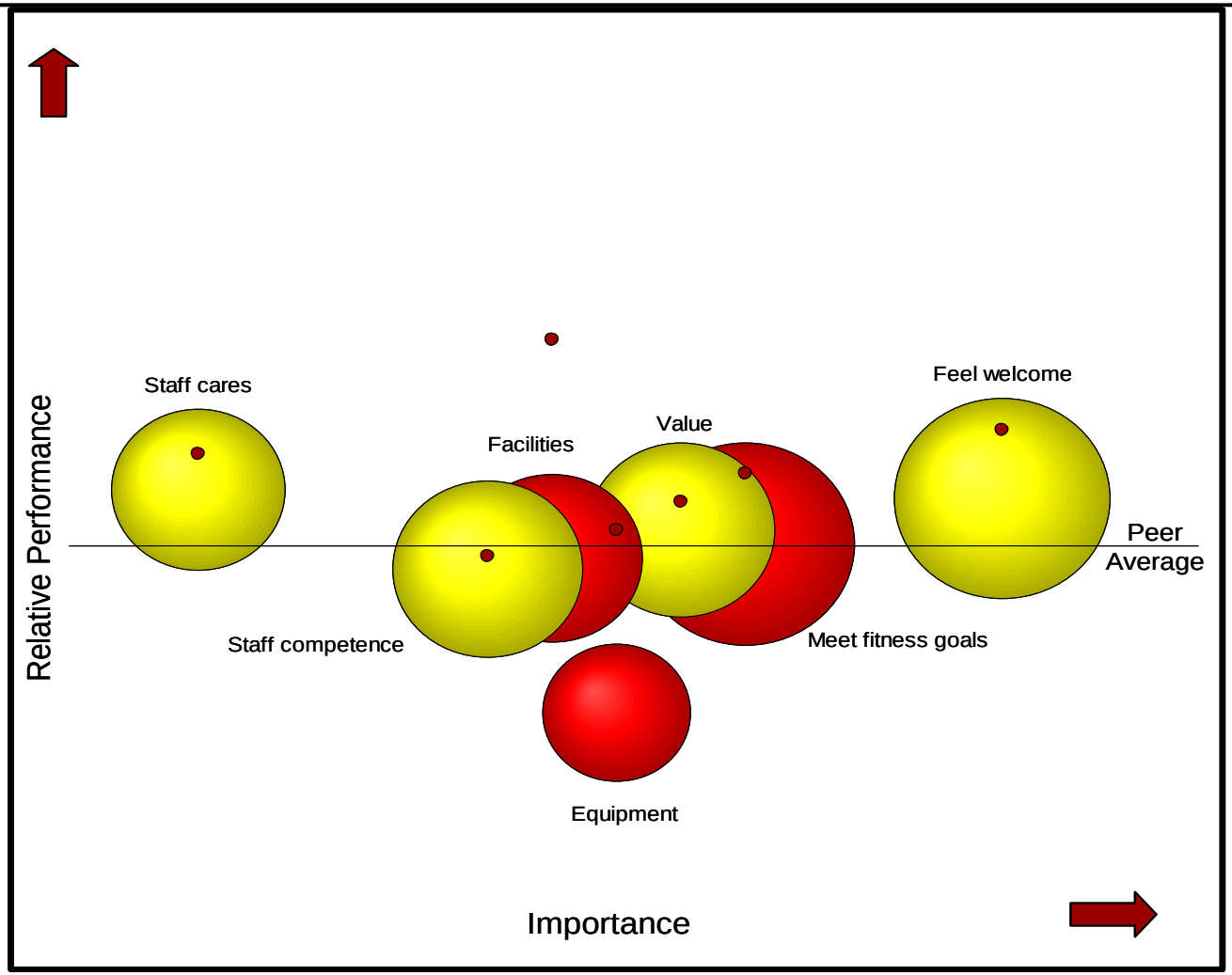
Photo © LoloStock/Shutterstock

© Abbott Analytics, 2014-2018

drivers of excellence

Current year vs. last year

- Better
- Same
- Worse



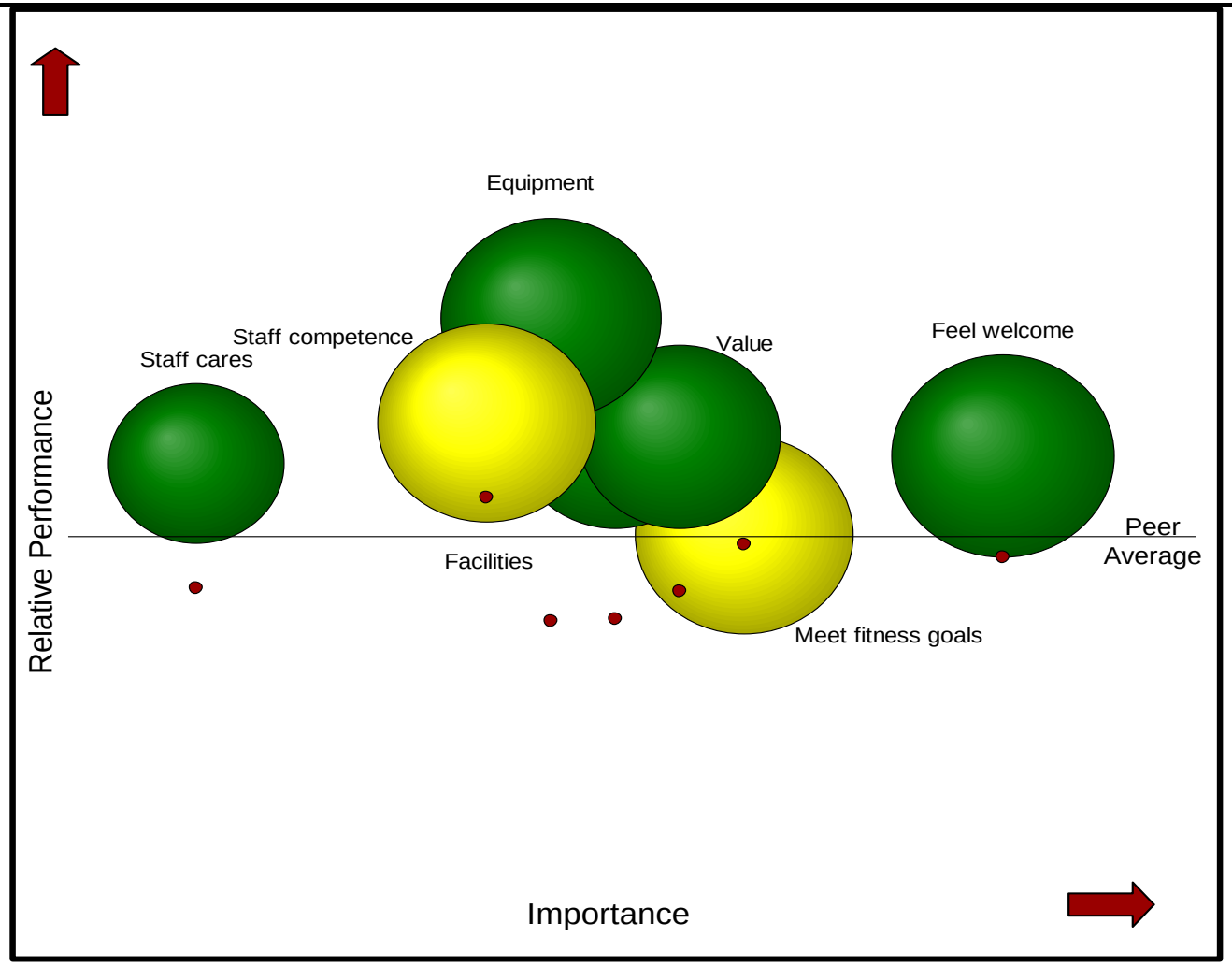
smarterHQ

© Abbott Analytics, 2014-2018

drivers of excellence

Current year vs. last year

- Better
- Same
- Worse



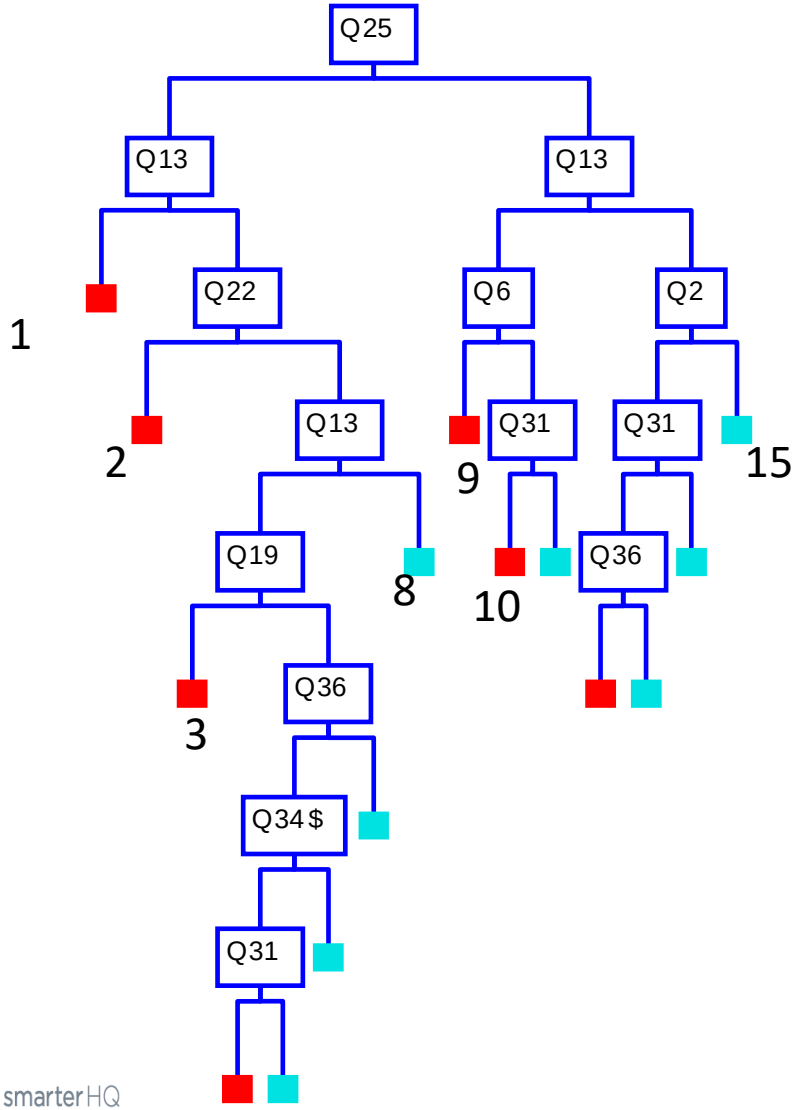
smarterHQ

© Abbott Analytics, 2014-2018

smarterHQ

© Abbott Analytics, 2014-2018

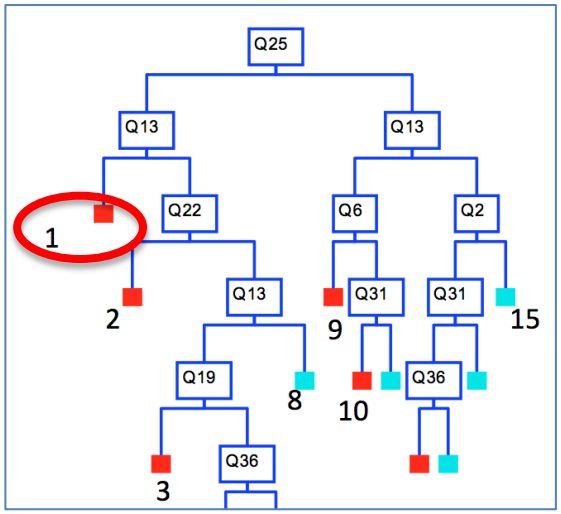
Another Approach to Telling the Data Story

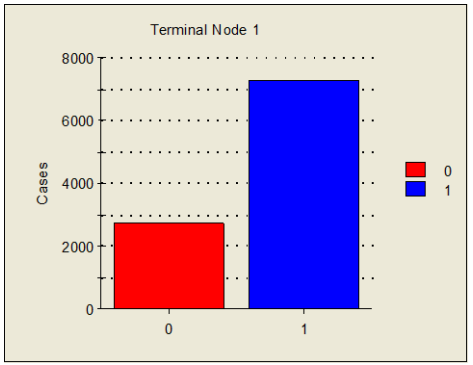


- Q25: Feel Welcome
  - Q13: Facilities are clean
- Q22: Value for Money
  - Q6: Staff Competent
  - Q2: Staff is Efficient
  - Q31: You are Loyal to the YMCA
  - Q36: Available Cardio Equipment

© Abbott Analytics, 2014-2018

Member Satisfaction Model: Key Rules





If strongly agree that **facilities are clean** and strongly agree that member **feels welcome**, then highly satisfied

~3/4 of members who match this profile are Highly Satisfied

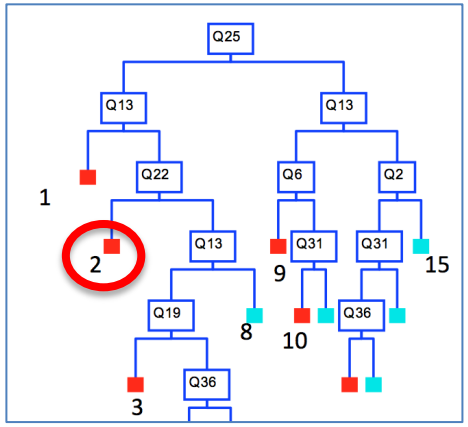
~1/2 of ALL Highly Satisfied Members Fit this Profile

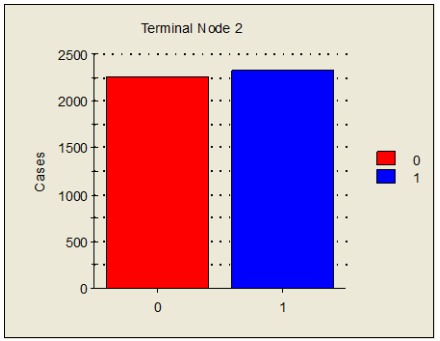
/*Rules for terminal node 1*/
Matches
- 10,014 surveys (20.8%),
- 7,289 highly satisfied (72.8%),
- 49% of all highly satisfied

RULE:
If (Q13 = 1 and Q25 = 1)
Then Satisfaction = 1
P(1) = 0.727881; Lift = 2.4

© Abbott Analytics, 2014-2018

Member Satisfaction Model: Key Rules





Even if don't strongly agree **facilities are clean**,
If strongly agree that **feel welcome** and
strongly agree Y is **value for money**,
then
highly satisfied

~1/2 of members who match this profile are Highly Satisfied

~1/6 of ALL Highly Satisfied Members Fit this Profile

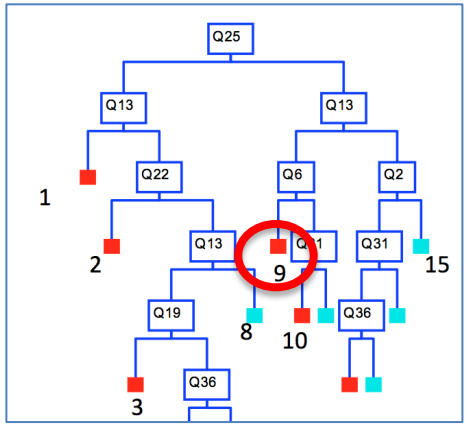
/*Rules for terminal node 2*/
Matches
- 4.578 surveys (9.5%),
- 2,317 highly satisfied (50.6%),
- 15.6% of all highly satisfied

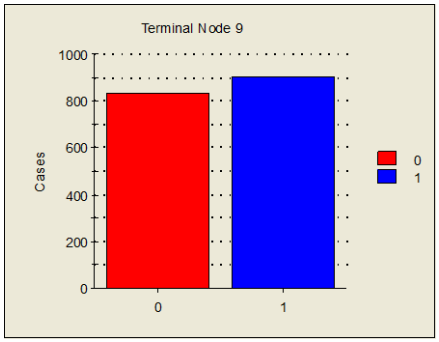
RULE:
If (Q13 <> 1 and
Q22 = 1 and Q25 = 1
Then Satisfaction = 1
P(1) = 0.506116; Lift = 1.6

smarterHQ

© Abbott Analytics, 2014-2018

Member Satisfaction Model: Key Rules





Even if don't strongly agree **feel welcome**,
If strongly agree that **facilities are clean**, and
strongly agree that **staff is competent**,
then
highly satisfied

~1/2 of members who match this profile are Highly Satisfied

~1 in 16 of ALL Highly Satisfied Members Fit this Profile

/*Rules for terminal node 9*/
Matches
- 1,739 surveys (3.6%),
- 904 highly satisfied (52.0%),
- 6.1% of all highly satisfied

RULE
IF (Q6 = 1 and
Q13 = 1 and Q25 <> 1
Then Satisfaction = 1
P(1) = 0.52; Lift = 1.7

smarterHQ

© Abbott Analytics, 2014-2018

Member Satisfaction Model: Key Rules

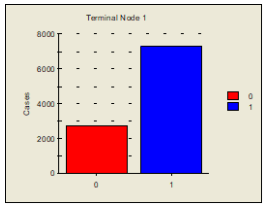
- facilities are clean and member feels welcome

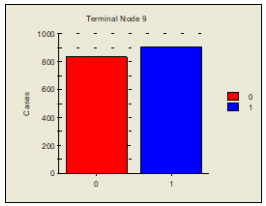
  +

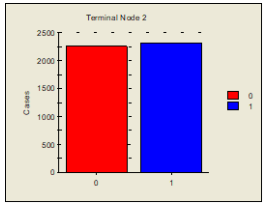
- even if don't feel welcome, if facilities are clean and staff is competent

  +

- even don't agree facilities are clean, if feel welcome and Y is value for money



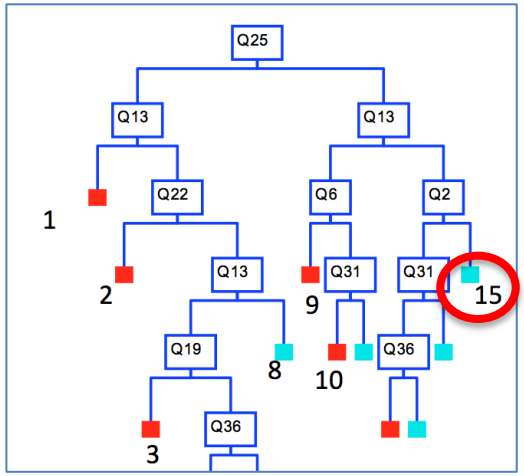


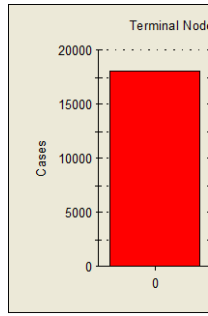


- Top 3 rules comprise ~ 3/4 of all highly satisfied members

© Abbott Analytics, 2014-2018

Member Satisfaction Model: Key Negative Rules





If don't strongly agree that **staff is efficient** and don't strongly agree that **feel welcome**, and don't strongly agree that the **facilities are clean**, then member isn't highly satisfied

~11 of 12 of members who match this profile are NOT Highly Satisfied

~6 in 10 of ALL NOT Highly Satisfied Members Fit this Profile

/*Rules fo
Matches
• 19,323 su
• 1,231 hig
• 8.3% of
• 58.2% of

RULE:
If (Q2 <>
Q25 <> 1
Then Not
P(1) = 0.0

smarterHQ

© Abbott Analytics, 2014-2018

Member Satisfaction Model: Not Highly Satisfied

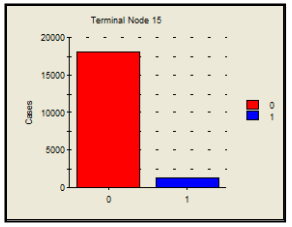
- If staff not efficient and don't feel welcome and facilities aren't clean

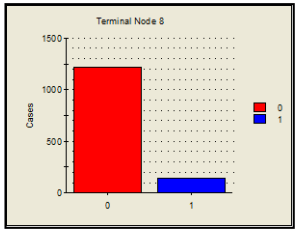
  +

- Even if feel welcome, if facilities are not clean and Y is not a good value for the money

  +

- Top 2 rules comprise ~ 2/3 of all NOT highly satisfied members





© Abbott Analytics, 2014-2018

Conclusions

- Know your software visualization building blocks
  - You will have to build most visualizations quickly, so become efficient with what you have

© Abbott Analytics, 2014-2018

Conclusions

- Know your software visualization building blocks
  - You will have to build most visualizations quickly, so become efficient with what you have
- Every visualization is biased
  - Bias isn't bad; be transparent about the bias
  - Different biases reveal complementary aspects of the data story

© Abbott Analytics, 2014-2018

Conclusions

- Know your software visualization building blocks
  - You will have to build most visualizations quickly, so become efficient with what you have
- Every visualization is biased
  - Bias isn't bad; be transparent about the bias
  - Different biases reveal complementary aspects of the data story
- Know your Audience
  - Tell a data story that the hearers understand

© Abbott Analytics, 2014-2018