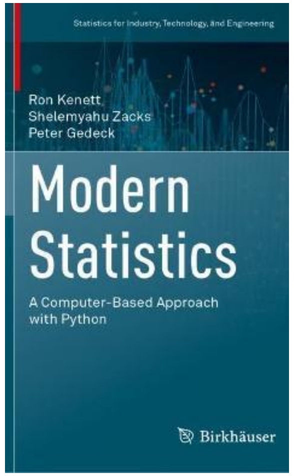
A Biomed Data Analyst Training Program

Data visualization

Professor Ron S. Kenett

Chapter 1
Analyzing Variability: Descriptive Statistics



Preview The chapter focuses on statistical variability and various methods of analyzing random data. Random results of experiments are illustrated with distinction between deterministic and random components of variability. The difference between accuracy and precision is explained. Frequency distributions are defined to represent random phenomena. Various characteristics of location and dispersion of frequency distributions are defined. The elements of exploratory data analysis are presented.

```
steelrod[26:] = steelrod[26:] - 3

ax = steelrod.plot(y='STEELROD', style='.', color='black')
ax.set_xlabel('Index')
ax.set_ylabel('Steel rod Length')
ax.hlines(y=steelrod[:26].mean(), xmin=0, xmax=26)
ax.hlines(y=steelrod[26:].mean(), xmin=26, xmax=len(steelrod))
plt.show()
```

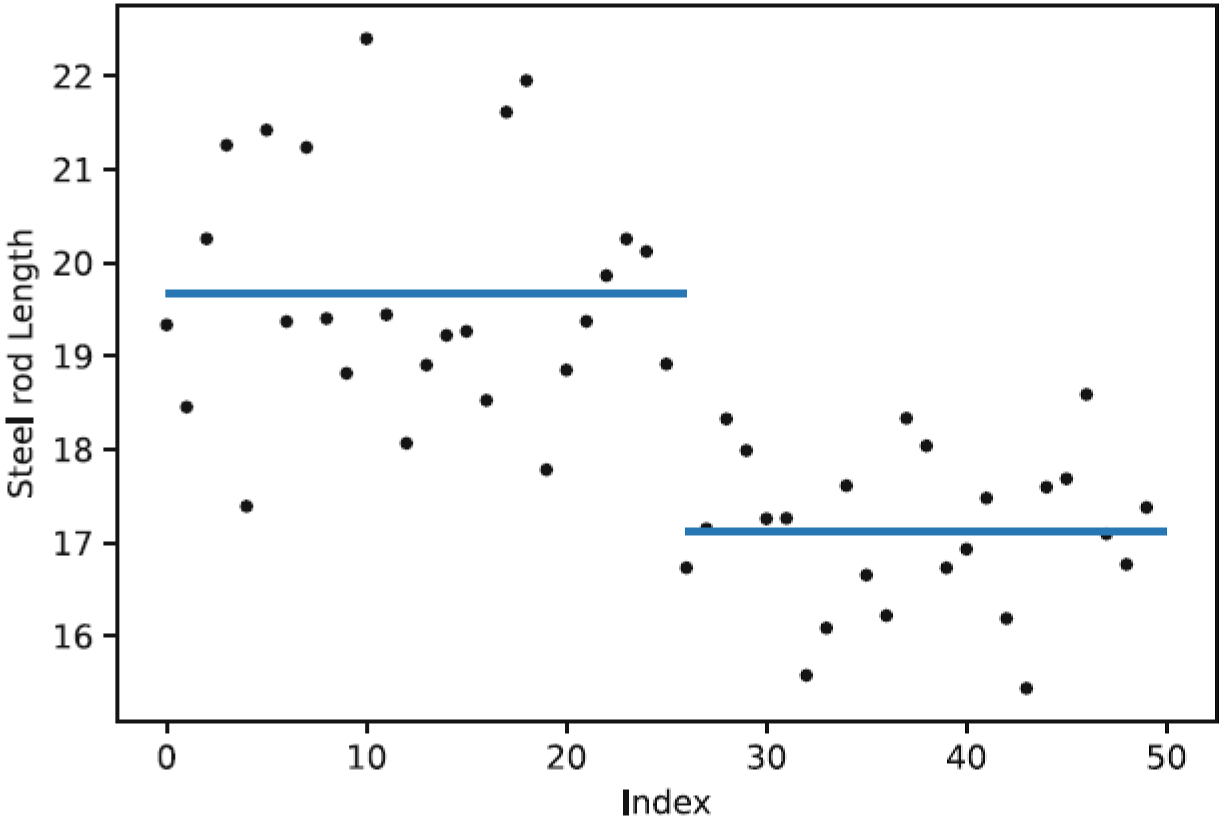


Fig. 1.2 Level shift after the first 25 observations

```
from scipy.stats import beta, norm

x = np.linspace(-3, 3, 200)
df = pd.DataFrame({'x': x,
                   'steep': beta(8, 8, loc=-3, scale=6).pdf(x),
                   'flat': beta(2.5, 2.5, loc=-3, scale=6).pdf(x),
                   'normal': norm().pdf(x),
                  })
ax = df.plot.line(x='x', y='steep', legend=False, color='black')
df.plot.line(x='x', y='normal', legend=False, color='black',
linestyle='--', ax=ax)
df.plot.line(x='x', y='flat', legend=False, color='black',
linestyle='-.', ax=ax)
ax.set_ylabel('y')
ax.text(0.5, 0.5, 'Steep')
ax.text(1.0, 0.35, 'Normal')
ax.text(2.0, 0.2, 'Flat')
plt.show()
```

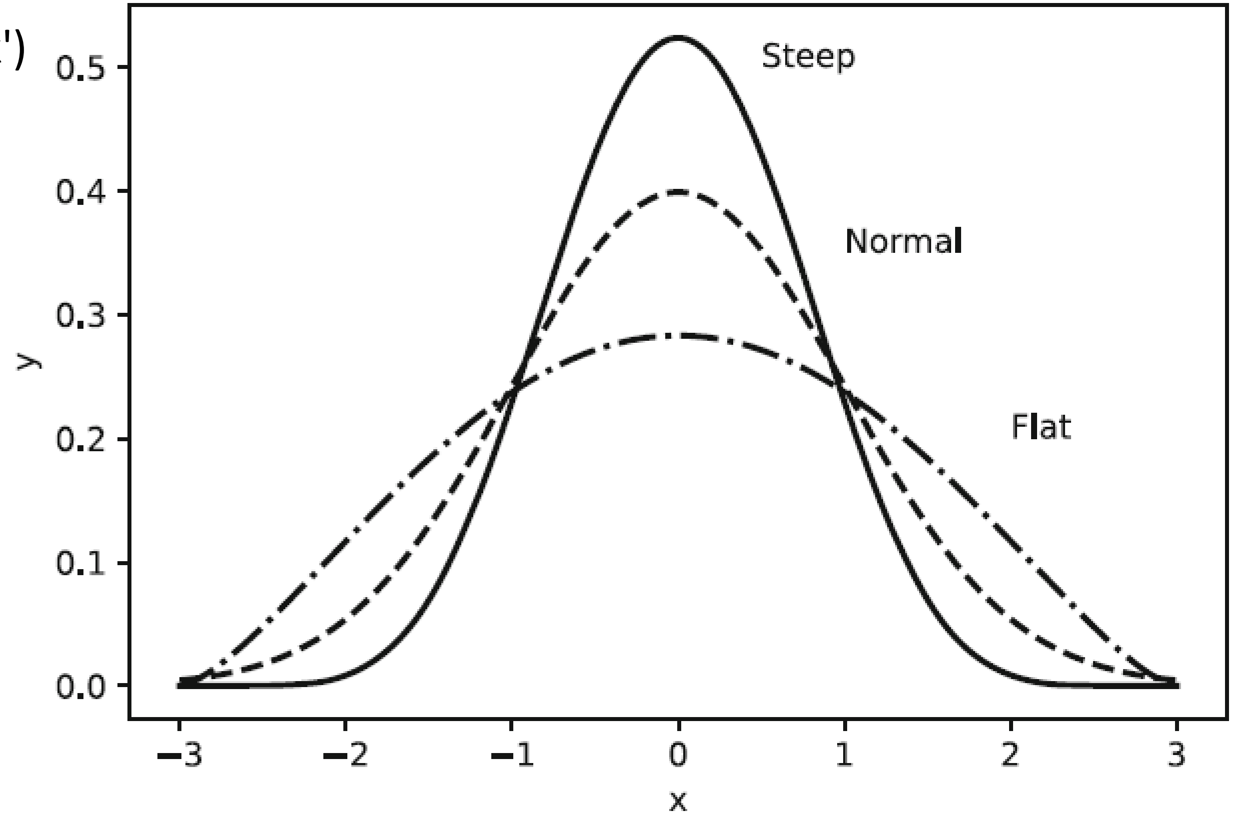


Fig. 1.13 Normal, steep, and flat distributions

```
X = mistat.load_data('YARNSTRG')
ax = X.plot.hist(bins=8, color='white', edgecolor='black',
legend=False, density=True)
X.plot.density(bw_method=0.2, ax=ax, color='black')
ax.set_xlabel('Log yarn strength')
plt.show()
```

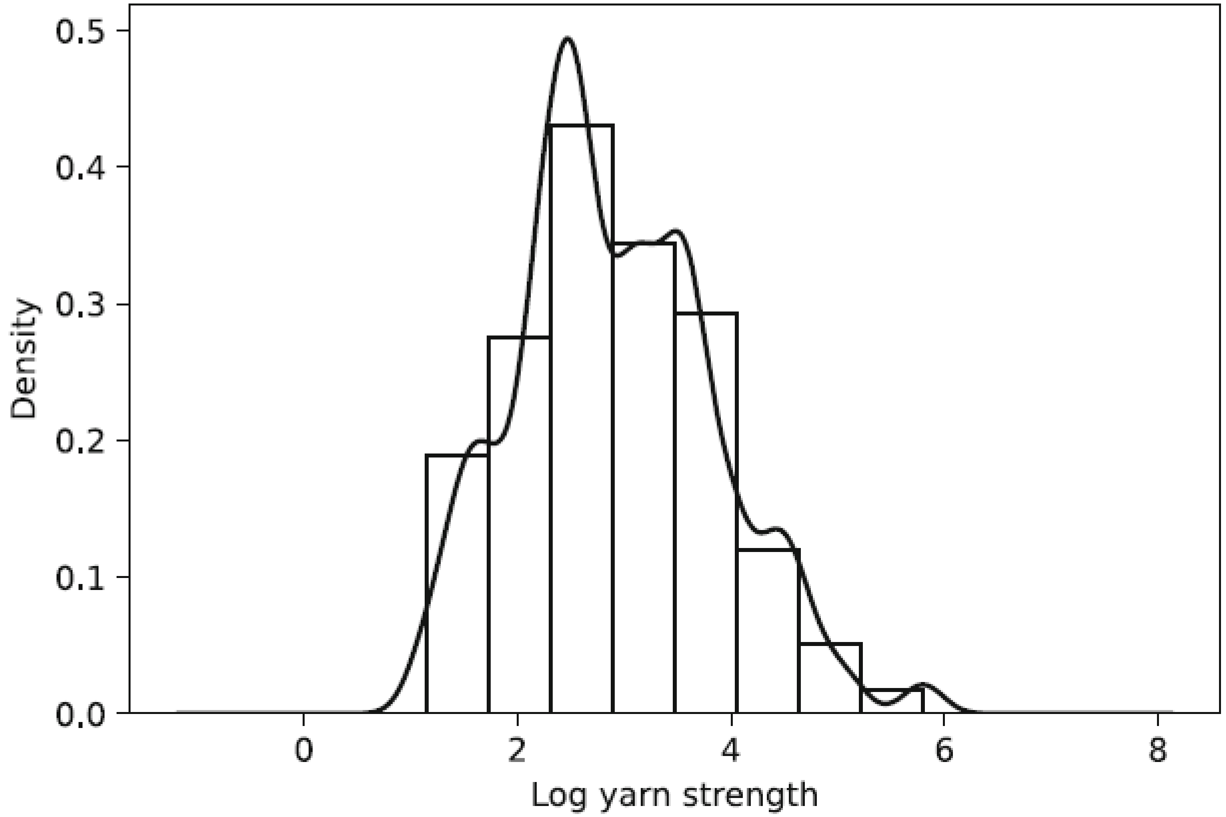
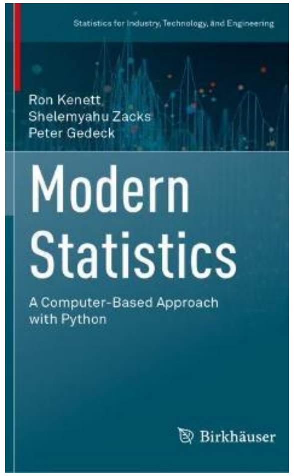


Fig. 1.14 Comparison of histogram and density plot for the log yarn strength datasets

Chapter 4
Variability in Several Dimensions and Regression Models



Preview When surveys or experiments are performed, measurements are usually taken on several characteristics of the observation elements in the sample. In such cases we have multivariate observations, and the statistical methods which are used to analyze the relationships between the values observed on different variables are called multivariate methods. In this chapter we introduce some of these methods. In particular, we focus attention on graphical methods, linear regression methods, and the analysis of contingency tables. The linear regression methods explore the linear relationship between a variable of interest and a set of variables, by which we try to predict the values of the variable of interest. Contingency tables analysis studies the association between qualitative (categorical) variables, on which we cannot apply the usual regression methods.

```
# The following command would be sufficient to create the scatterplot matrix
# matplotlib has however a problem with scaling xDev
sns.pairplot(place[['xDev', 'yDev', 'tDev']], markers=".",
          plot_kws={'facecolors': 'none', 'edgecolor': 'black'},
          diag_kws={'color': 'grey'})

#def panelPlot(x, y, **kwargs):
#    plt.scatter(x, y, **kwargs,
#              facecolors='none', edgecolor='black', s=20)
#    dx = 0.05*(max(x) - min(x))
#    plt.xlim(min(x)-dx, max(x) + dx)
#    dy = 0.05*(max(y) - min(y))
#    plt.ylim(min(y)-dy, max(y) + dy)
#g = sns.PairGrid(place[['xDev', 'yDev', 'tDev']])
#g = g.map_offdiag(panelPlot)
plt.show()
```

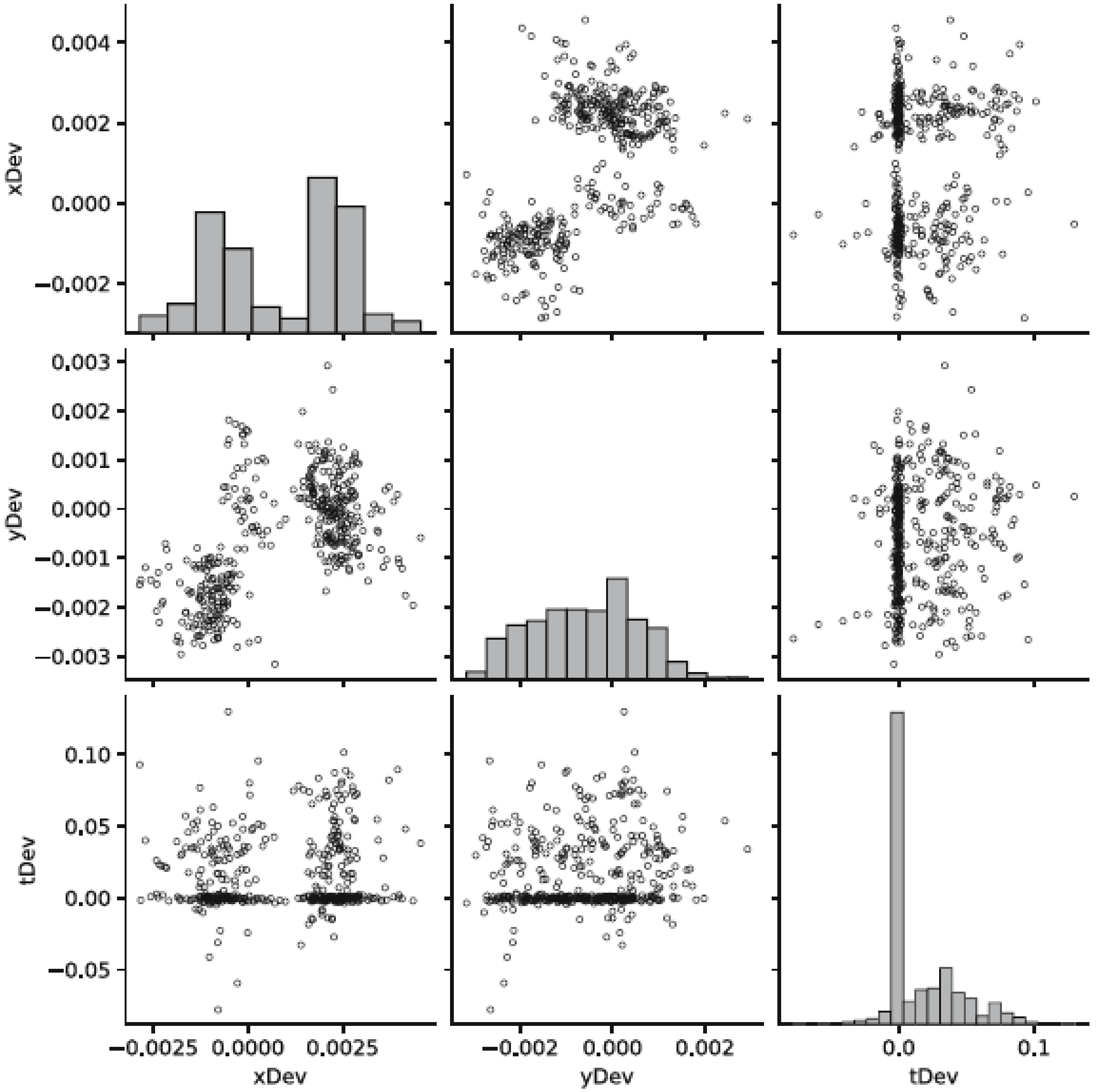


Fig. 4.2 Scatterplot matrix

```
# create visualization
ax = car_US.plot.scatter(x='turn', y='mpg', color='gray',
marker='o')
car_Asia.plot.scatter(x='turn', y='mpg', ax=ax, color='gray',
marker='^')

car_combined = car_combined.sort_values(['turn'])
ax.plot(car_combined['turn'],
model_US.predict(car_combined),
      color='gray', linestyle='--')
ax.plot(car_combined['turn'],
model_Asia.predict(car_combined),
      color='gray', linestyle=':')
ax.plot(car_combined['turn'],
model_simple.predict(car_combined),
      color='black', linestyle='-')
plt.show()
```

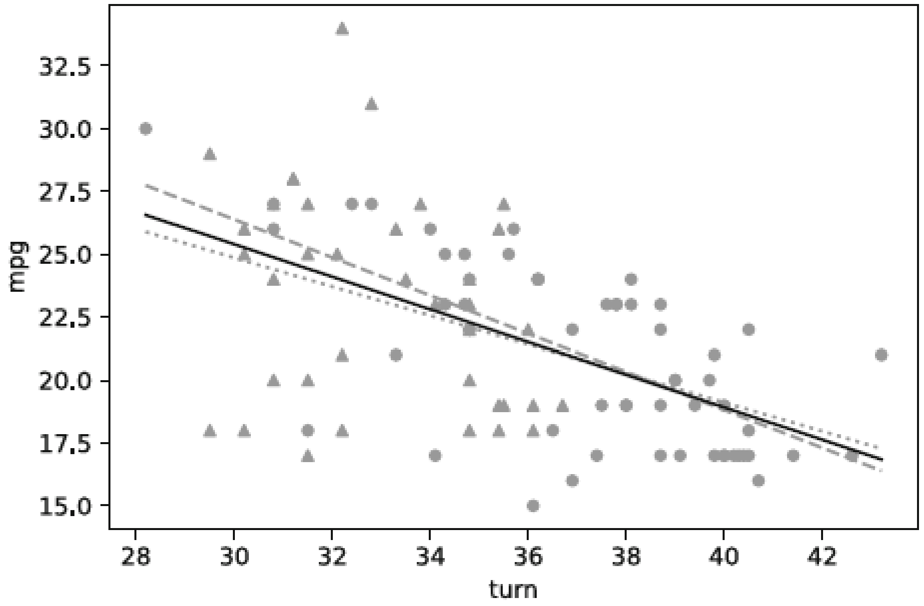


Fig. 4.16 Linear regression analysis for US (filled circle, dashed line) and Japanese cars (filled triangle, dotted line). The solid line is the linear regression of the combined data set


Baggage

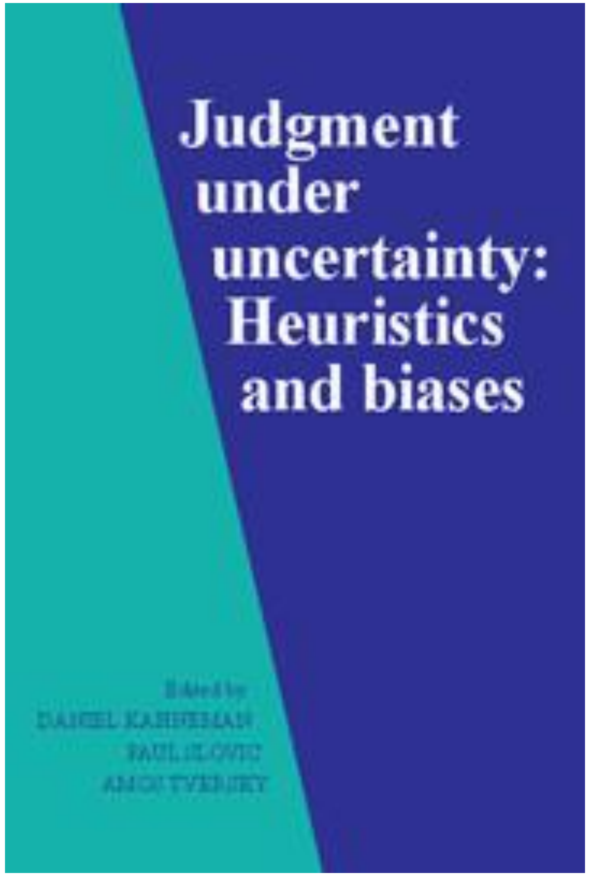
Presenting uncertainty in data

- Capability of human mind for solving complex problems is limited compared with the size of problems
- Lack of objectively rational behaviour in real world. Cognitive illusions.
- Use of simple “rules of thumb” to simplify decision making
- Heuristics can be helpful, but can also lead to biases, especially in uncertain situations where probabilities are encountered

Presenting uncertainty in data

- “Nothing is certain”
- In many situations, decisions have to be based on probabilities
- Interpretation of probabilities is sometimes not straightforward
- Appropriate presentation can help to make the right decisions

Presenting uncertainty in data



- formulating the problem:
  - probabilities vs. frequencies
  - the framing effect
  - the anchoring effect
- underweighting base rates
- hindsight and confirmation bias
- belief persistence: Primacy and inertia effect
- group conformity and decision regret

Conditional probabilities

- Breast cancer screening. The facts:
  - Probability that a woman aged 40-50 has breast cancer = 0.8%
  - If a woman has breast cancer, probability of positive test = 90%
  - If a woman does not have breast cancer, prob. of positive test=7%
- Imagine a woman with a positive test.

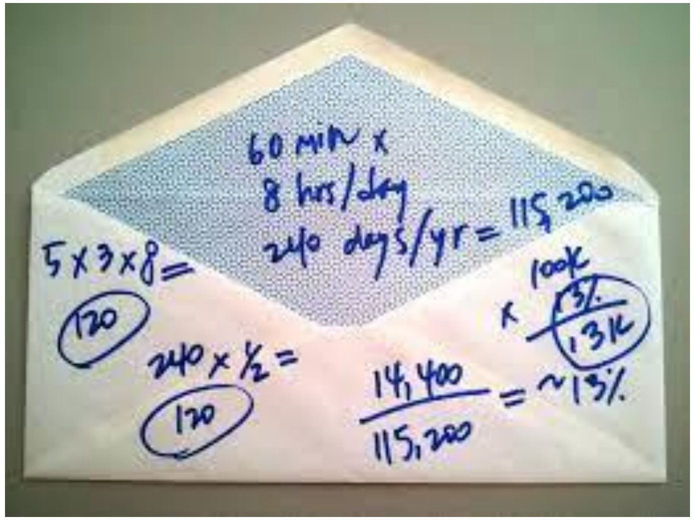
  What is the probability, that she actually has breast cancer?

- Solution:
  - p(disease) = 0.008
  - p(pos|disease) = 0.90
  - p(pos| no disease) = 0.07
  - p(disease|pos) = $\dfrac{\text{p(disease) * p(pos|disease)}}{\text{p(disease) * p(pos|disease) + p(no disease) * p(pos| no disease)}}$

    = 0.09

Frequency formulation

- Breast cancer screening. The facts:
  - Probability that a woman aged 40-50 has breast cancer = 0.8%
  - If a woman has breast cancer, probability of positive test = 90%
  - If a woman does not have breast cancer, prob. of positive test=7%



- Back of envelope solution:

1000 women

8: disease

992: no disease

7: positive

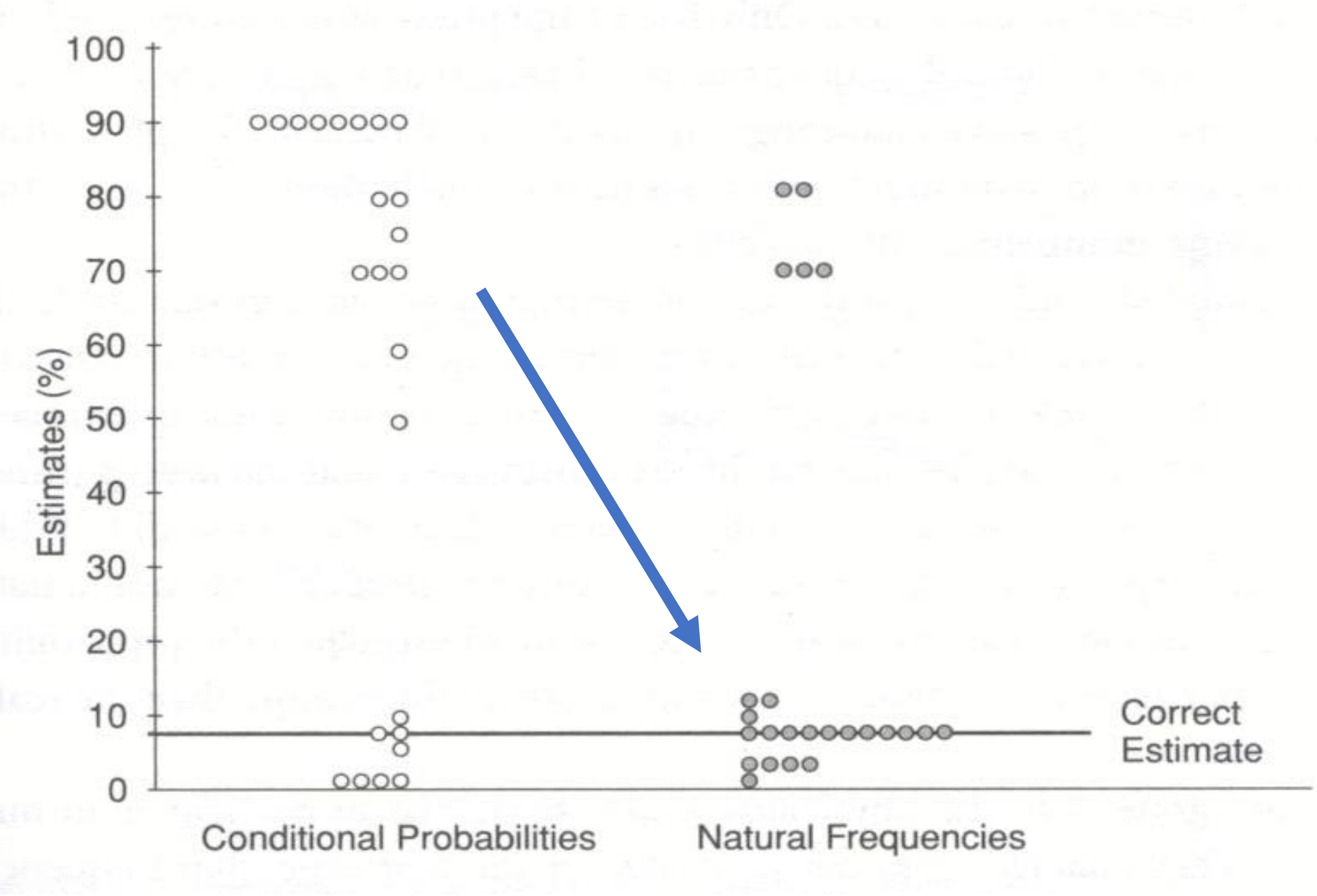
1: negative

69: positive

923: negative

p(disease | pos) = 7 / (7+69) = 0.09

Probabilities vs. frequencies



Estimated chances of breast cancer, given a positive screening mammogram (Gigerenzer, 2002)

The framing effect

- The way a problem (or forecast) is formulated can affect a decision
- Imagine that London faces an unusual disease that is expected to kill 600 people.

Two alternative programs to combat disease:

- Program A: 200 people will be saved

- Program B: 1/3 probability 600 saved, 2/3 probability nobody saved

Tests indicate that 72% would select program A (risk-averse)

- Slightly changed wording:

- Program C: 400 people will die

- Program D: 1/3 prob. that nobody will die, 2/3 prob. that 600 will die

Tests indicate that 78% would select program D (risk-taking)

The framing effect in real life

- Professionals, experienced in decision-making, are still affected
- E.g., information for doctors:

  - mortality rate of 7% within 5 years -> hesitant to recommend

  - survival rate after 5 years of 93% -> more inclined to recommend

- For weather predictions this suggests different response to forecasts expressed as likelihood of drought or non-likelihood of wet conditions

- E.g., different response to: 30% chance of drought and
  70% chance of normal or wet conditions

- Worded vs. numerical forecast:
  - 11% judge forecast "rain is likely" as poor if it did not rain
  - 37% judge forecast "70% chance of rain" as poor if it did not rain although they associate the word "likely" with probability of 70%

Test your knowledge of history

- What are the last three digits of your phone number?
- Add 400 to this number
- Do you think Attila the Hun was defeated in Europe before or after that year?
- In what year would you guess Attila the Hun was defeated?
- The correct answer is: A.D. 451

Range of initial anchor	Average estimate
400 – 599	629
600 – 799	680
800 – 999	789
1000 – 1199	885
1200 – 1399	988

Underweighting base rates

- Imagine a climate model (with 90% accuracy) predicts drought
- Historically, there is 10% chance of drought
- What is the chance that drought will occur in next season?
- Solution:

100 seasons

- 10: drought
  - 9: drought FC
  - 1: no-drought FC
- 90: no drought
  - 81: no-drought FC
  - 9: drought FC

p(drought | drought FC) = 9 / (9+9) = 0.50

Underweighting base rates

- Imagine a climate model (with 90% accuracy) predicts drought
- Historically, there is 10% chance of drought
- What is the chance that drought will occur in next season?

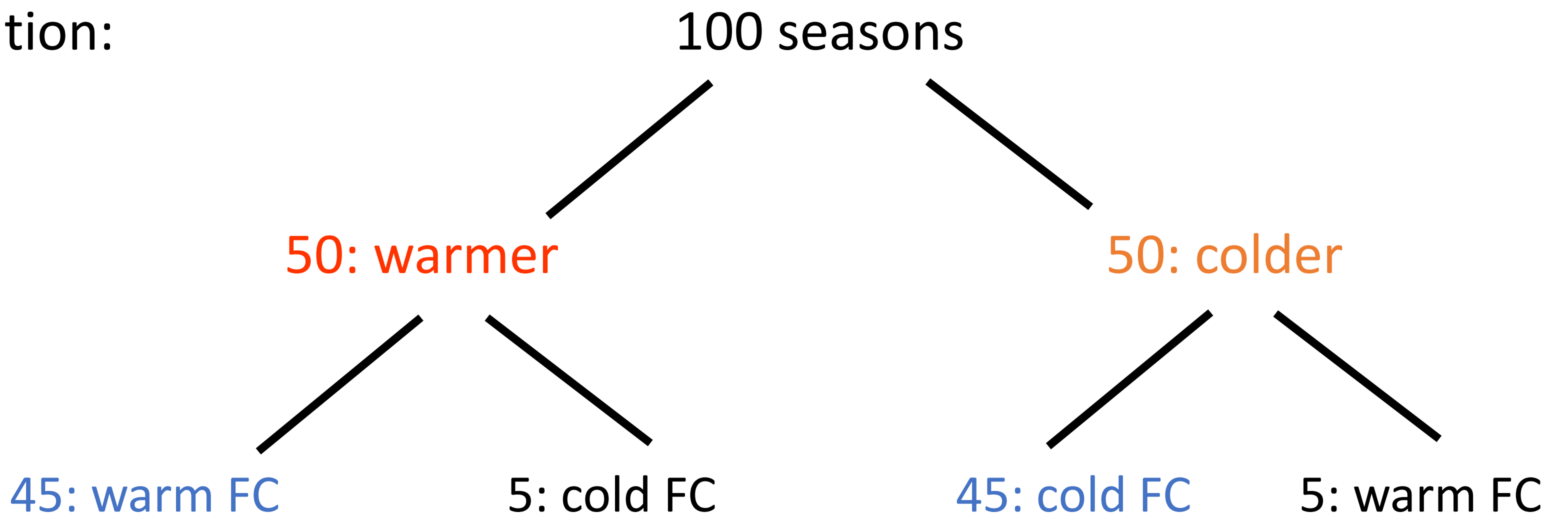
Challenge to convince user that

- Model was correct 90% of time
- the probability of a drought next season was only 50%

For equally likely events, accuracy translates into probabilities

Underweighting base rates

- Imagine a climate model (with 90% accuracy) predicts warmer than normal conditions
- There is a 50% chance of above normal
- What is the chance that warmer than normal conditions will occur?
- Solution:



$$p(\text{warmer} \mid \text{warm FC}) = 45 / (45+5) = 0.90$$

Hindsight and confirmation bias

Men mark where they hit, and not where they miss. (Jevons, 1958)

- After finding out whether or not an event occurred, individuals tend to overestimate the degree to which they would have predicted the correct outcome

- Reported outcomes seem less surprising in hindsight than in foresight

- Example: El Nino 1997 regarded as “stunning success”, although only one model was reported in the March 1997 NOAA Long-Lead Forecast Bulletin predicting more than slight warming. Some of the very poor forecasts simply ignored in hindsight
- Considerable evidence that people tend to ignore (and not search for) disconfirming information of any hypothesis

- Introduce “double-blind test” for model assessment, if posible

Belief persistence

- Primacy and inertia also tend to weight evidence inaccurately.

- People tend to weight more heavily evidence presented first, e.g. persons described as:
  - intelligent, industrious, impulsive, critical, stubborn, envious

  are more favourable perceived than persons described as
  - envious, stubborn, critical, impulsive, industrious, intelligent

- Inertia may lead people to ignore evidence that contradicts their prior belief (e.g. that a particular forecast system produces useful forecasts)

- Forecast producers may not recognise the disparity of model predictions, and instead rely too heavily on a forecast that supports their intuitive understanding of the current state of climate

Strategies to reduce cognitive illusions

- Recognition that decision-making is inherently biased

- Understanding how written forecasts, and numerical probability forecasts are interpreted by potential users

- Try to reduce impact of cognitive illusions by
  - ➢ encouraging forecaster groups to de-bias forecasts by e.g. reducing overconfidence or hindsight bias
  - ➢ taking care that media reports and forecasts do not cause anchoring to extreme events (e.g. El Nino 82/83)
  - ➢ taking care in wording forecasts to avoid framing
  - ➢ avoid "intuitive" approach when combining forecasts, objective approaches exist and are more successful
  - ➢ ensuring that base-rates are not ignored
  - ➢ using additional visual aids to convey real levels of skill

Checklists

https://fs.blog/before-you-make-that-big-decision/

A Simple Checklist to Improve Decisions

We owe thanks to the publishing industry. Their ability to take a concept and fill an entire category with a shotgun approach is the reason that more people are talking about biases.

Unfortunately, talk alone will not eliminate them but it is possible to take steps to counteract them. Reducing biases can make a huge difference in the quality of any decision and it is easier than you think.

In a recent article for Harvard Business Review, Daniel Kahneman (and others) describe **a simple way to detect bias and minimize its effects in the most common type of decisions people make**: determining whether to accept, reject, or pass on a recommendation.

Checklists

https://chemistry-europe.onlinelibrary.wiley.com/doi/full/10.1002/ansa.202000159

Analytical Science Advances
Chemistry Europe
European Chemical Societies Publishing

Research Article | Open Access

Helping reviewers assess statistical analysis: A case study from analytic methods

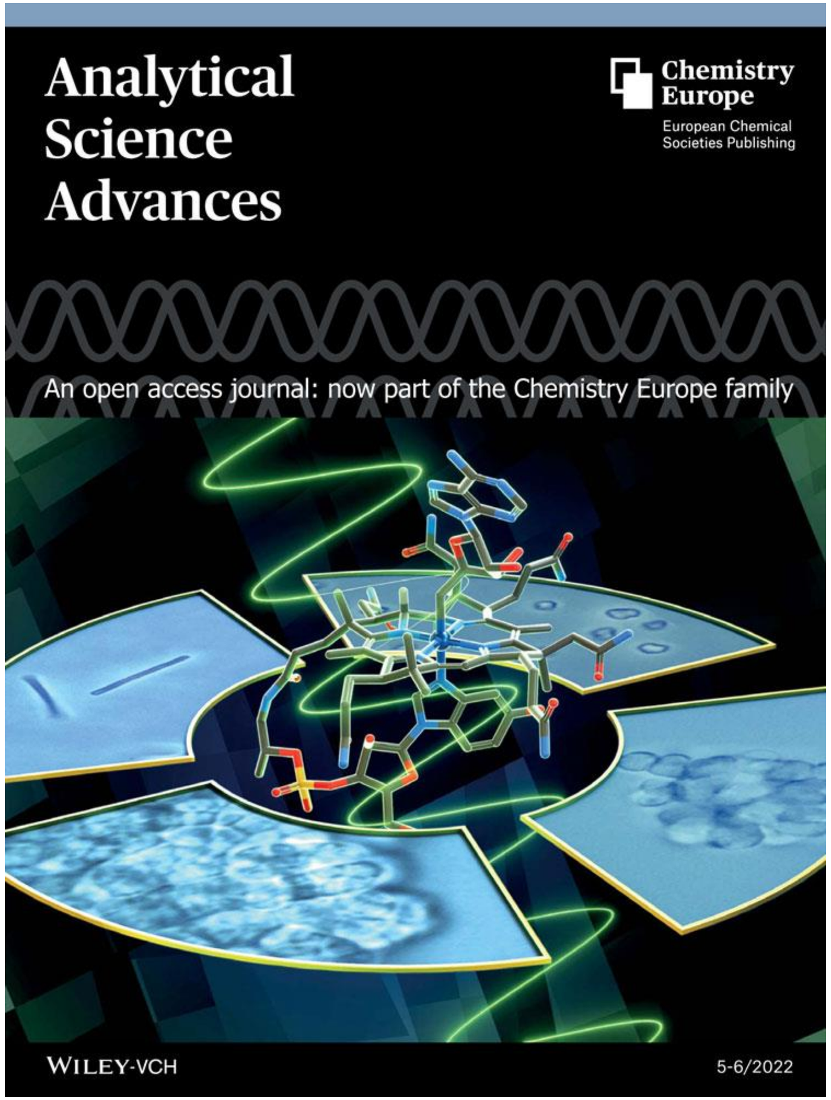
Ron S. Kenett, Bernard G. Francq

First published: 16 June 2022 | https://doi.org/10.1002/ansa.202000159

SECTIONS PDF TOOLS SHARE

Abstract

Analytic methods development, like many other disciplines, relies on experimentation and data analysis. Determining the contribution of a paper or report on a study incorporating data analysis is typically left to the reviewer's experience and good sense, without reliance on structured guidelines. This is amplified by the growing role of machine learning driven analysis, where results are based on computer intensive algorithm applications. The evaluation of a predictive model where cross validation was used to fit its parameters adds challenges to the evaluation of regression models, where the estimates can be easily reproduced. This lack of structure to support reviews increases uncertainty and variability in reviews. In this paper, aspects of statistical assessment are considered. We provide checklists for reviewers of applied statistics work with a focus on analytic method development. The checklist covers six aspects relevant to a review of statistical analysis, namely: (1) study design, (2) algorithmic and inferential methods in frequentism analysis, (3) Bayesian methods in Bayesian analysis (if relevant), (4) selective inference aspects, (5) severe testing properties and (6) presentation of findings. We provide a brief overview of these elements providing references for a more elaborate treatment. The robustness analysis of an analytical method is used to illustrate how an improvement can be achieved in response to questions in the checklist. The paper is aimed at both engineers and seasoned researchers.



Checklists

https://chemistry-europe.onlinelibrary.wiley.com/doi/full/10.1002/ansa.202000159

TABLE 1 Questions for reviewing statistical analysis in applied research

Part	Questions
1. Study design	1.1 Is the experimental set up clearly presented? 1.2 Have aliasing and power consideration been taken into account? 1.3 Is there reference to blocking, split plots and randomization? 1.4 Was an IRB required, and if so, was it obtained? (if relevant) 1.5 Are there any data ethics issues to consider?
2. Algorithmic and inferential methods	2.1 Are the algorithmic and inferential methods uses clearly stated? 2.2 Is the analysis aiming at estimation, predictive or explanatory goals? 2.3 Are data and code available to replicate the analysis? 2.4 Are outcomes of inferential analysis properly interpreted?
3. Bayesian analysis	3.1 Are prior distributions justified using prior experience or data? 3.2 What are the Bayesian methods used in the analysis? 3.3 How are Bayes factors interpreted?
4. Selective inference	4.1 Has the study been pre-registered? 4.2 Have any false discovery rate corrections been made? 4.3 Is the presentation of findings affected by selective inference?
5. Severe testing	5.1 Have the findings been tested with an option of failing the test? 5.2 Is the study a first or is it replicating previous studies? 5.3 Have probabilism, performance and probativeness criteria been considered? 5.4 What type of model is used in the analysis: primary models, experimental models or and data models? 5.5 If used, how are confidence interval (CI) interpreted?
6. Presentation of findings	6.1 How are the research findings presented? 6.2 Have the research findings been generalized? 6.3 Are there any causality arguments presented? 6.4 In a causal study, are there issues of endogeneity (reverse-causation)?

Checklists

https://chemistry-europe.onlinelibrary.wiley.com/doi/full/10.1002/ansa.202000159

TABLE 2 Checklist for analytic methods

Analytic method element	Description and question (Q)
Precision	This requirement makes sure that method variability is only a small proportion of the specifications range (upper specification limit – lower specification limit). This is also called gage reproducibility and repeatability (GR&R). *Q: Does the study address precision? How?*
Selectivity	Determination of impurities to monitor at each production step and specification of design methods that adequately discriminate the relative proportions of each impurity. *Q: Does the study address selectivity? How?*
Sensitivity	The achievement with the method of effective process control, by accurately reflecting changes in CQA's that are important relative to the specification limits. *Q: Does the study address sensitivity? How?*
Method Design Intent	Identification and specification of the analytical method performance *Q: Is the method design intent stated?*
Method Design Selection	Approach to the selection of the method work conditions to achieve the design intent *Q: Is the study design described?*
Method Control	Establishment and definition of appropriate controls for the components with the largest contributions to performance variability. *Q: Is the application of the method discussed?*
Method Control Validation	Demonstration of acceptable method performance with robust and effective controls. *Q: Is the method validation demonstrated?*
Method robustness	Testing robustness of analytical methods involves evaluating the influence of small changes in the operating conditions. *Q: Is the method robustness evaluated?*
Method ruggedness	Ruggedness testing identifies the degree of reproducibility of test results obtained by the analysis of the same sample under various normal test conditions such as different laboratories, analysts, and instruments *Q: Is the method ruggedness evaluated?*

Statistical Journal of the IAOS 32 (2016) 11–19
DOI 10.3233/SJI-160967
IOS Press

Helping authors and reviewers ask the right questions: The InfoQ framework for reviewing applied research

Ron S. Kenett[a,*] and Galit Shmueli[b]
[a]*University of Turin, Italy and KPA Group, Raanana, Israel*
[b]*Institute of Service Science, National Tsing Hua University, Hsinchu, Taiwan*

Abstract. Reviewers play a critical role in the publication process, the hallmark of scientific advancement. Yet, in many journals, determining the contribution of a paper is left to the reviewer's experience and good sense without providing structured guidelines. This lack of guidance to authors and reviewers increases uncertainty and variability in the usefulness of reviews. We propose an approach, based on the Information Quality (InfoQ) framework, that provides guideline scaffolding for the review process of applied research papers submitted for publication in scientific journals.

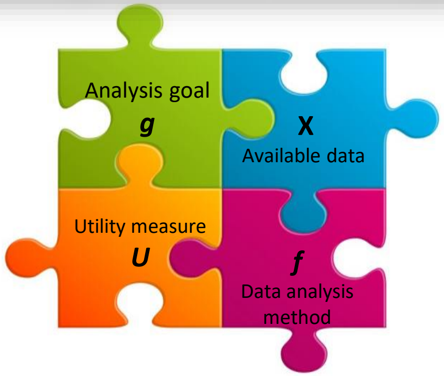
Keywords: Information quality, publication, empirical study, data analysis, reviewing guidelines

Table 2
InfoQ questionnaire for reviewing an empirical research paper or study

Dimension	Questions
1. Data Resolution	1.1 Is the data scale used aligned with the stated goal? 1.2 How reliable and precise are the measuring devices or data sources? 1.3 Is the data analysis suitable for the data aggregation level?
2. Data Structure	2.1 Is the type of the data used aligned with the stated goal? 2.2 Are data integrity details (corrupted/missing values) described and handled appropriately? 2.3 Are the analysis methods suitable for the data structure?
3. Data Integration	3.1 Are the data integrated from multiple sources? If so, what is the credibility of each source? 3.2 How is the integration done? Are there linkage issues that lead to dropping crucial information? 3.3 Does the data integration add value in terms of the stated goal? 3.4 Does the data integration cause any privacy or confidentiality concerns?
4. Temporal Relevance	4.1 Considering the data collection, data analysis and deployment stages, is any of them time-sensitive? 4.2 Does the time gap between data collection and analysis cause any concern? 4.3 Is the time gap between the data collection and analysis and the intended use of the model (e.g., in terms of policy recommendations) of any concern?
5. Chronology of Data & Goal	5.1 If the stated goal is predictive, are all the predictor variables expected to be available at the time of prediction? 5.2 If the stated goal is causal, do the causal variables precede the effects? 5.3 In a causal study, are there issues of endogeneity (reverse-causation)?
6. Generalizability	6.1 Is the stated goal statistical or scientific generalizability? 6.2 For statistical generalizability in the case of inference, does the paper answer the question "What population does the sample represent?" 6.3 For generalizability in the case of a stated predictive goal (predicting the values of new observations; forecasting future values), are the results generalizable to the to-be-predicted data? 6.4 Does the paper provide sufficient detail for the type of needed reproducibility and/or repeatability, and/or replicability?
7. Operationalization	Construct operationalization: 7.1 Are the measured variables themselves of interest to the study goal, or is their underlying construct? 7.2 What are the justifications for the choice of variables? Strength of operationalizing results: 7.3 Who can be affected (positively or negatively) by the research findings? 7.4 What can the affected parties do about it?
8. Communication	8.1 Is the exposition of the goal, data and analysis clear? 8.2 Is the exposition level appropriate for the readership of this journal? 8.3 Are there any confusing details or statements that might lead to confusion or misunderstanding?

Information Quality

The potential of a particular dataset to achieve a particular goal using a given empirical analysis method



InfoQ(f,X,g) = U(f(X|g))

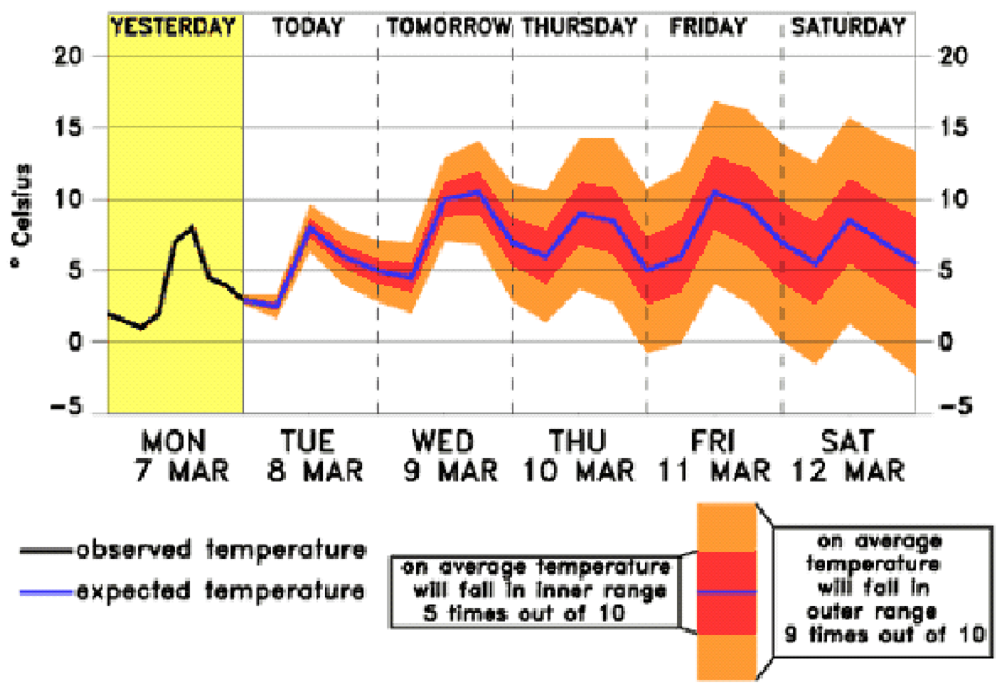
g	**A specific analysis goal**
X	**The available dataset**
f	**An empirical analysis method**
U	**A utility measure**

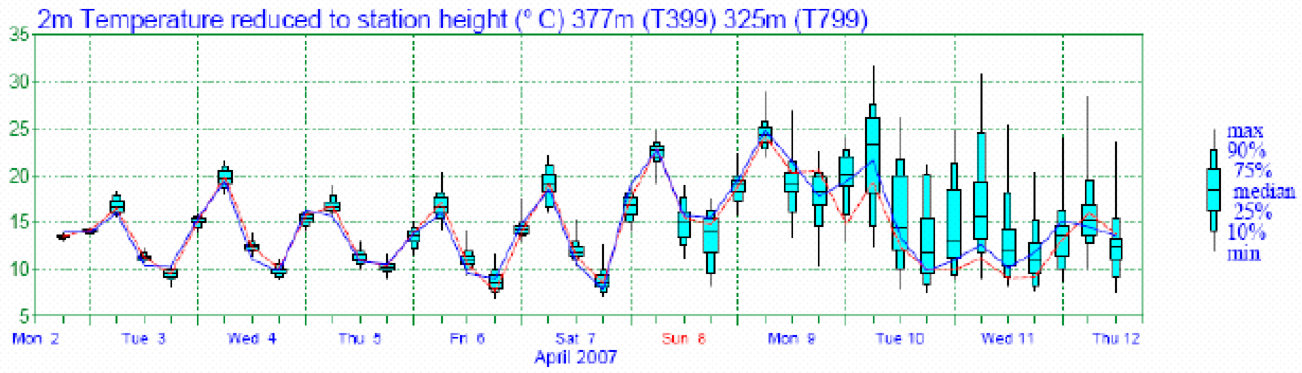
What

How

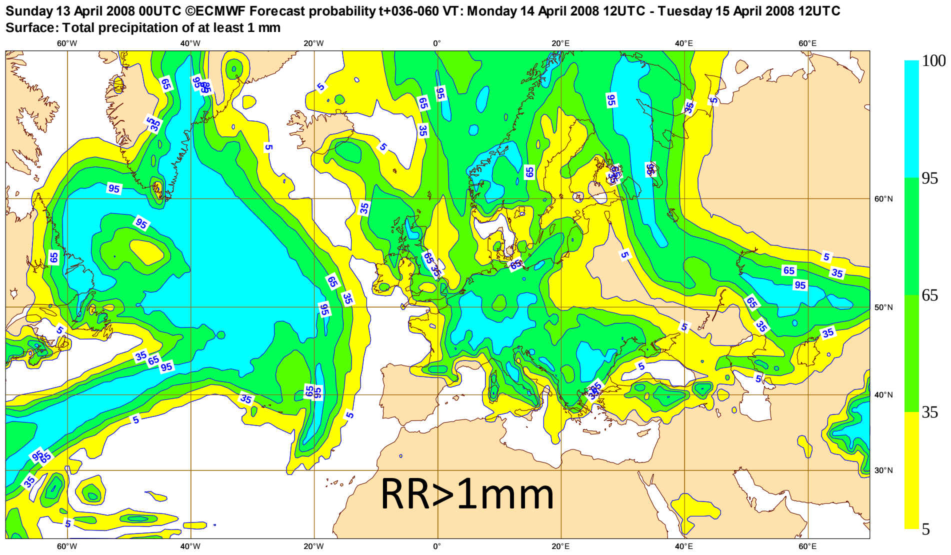
1. Data resolution
2. Data structure
3. Data integration
4. Temporal relevance
5. Chronology of data and goal
6. Generalizability
7. Operationalization
8. **Communication**

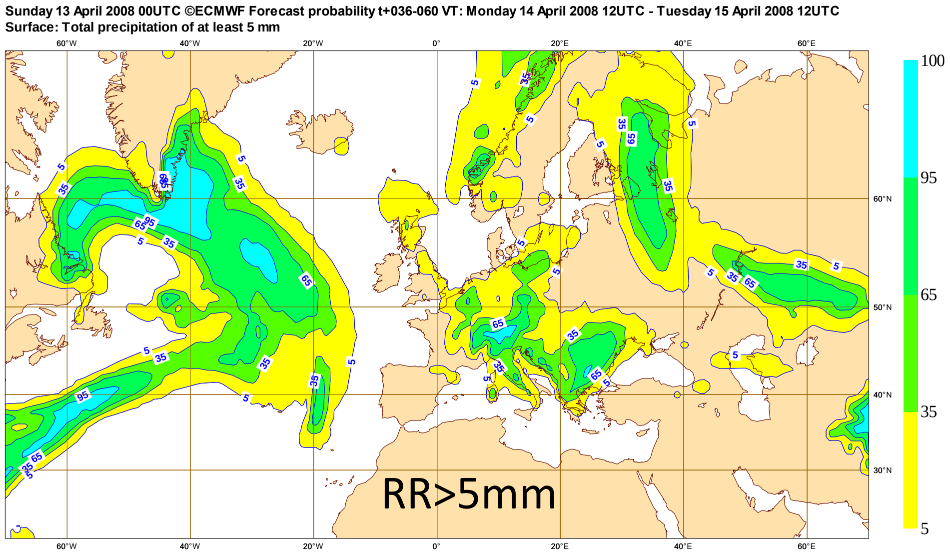
Visualization of time series

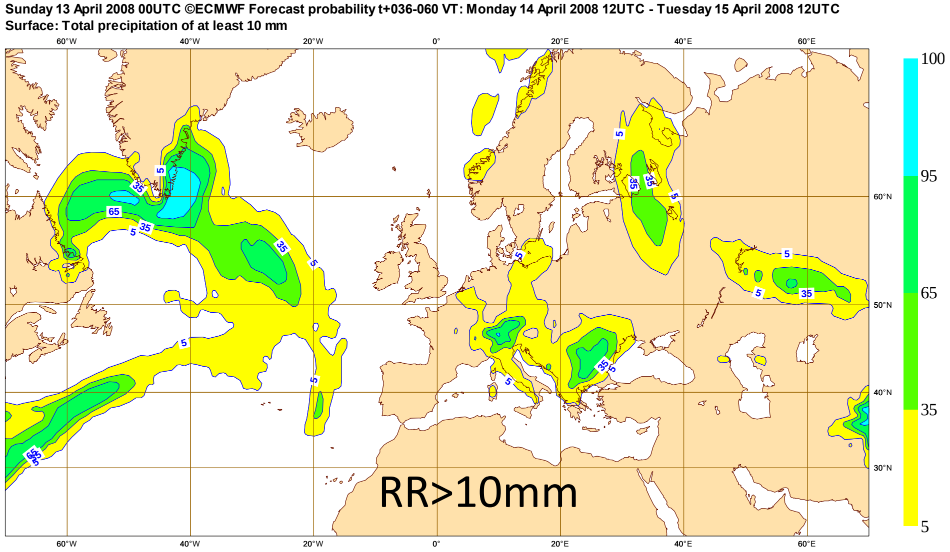


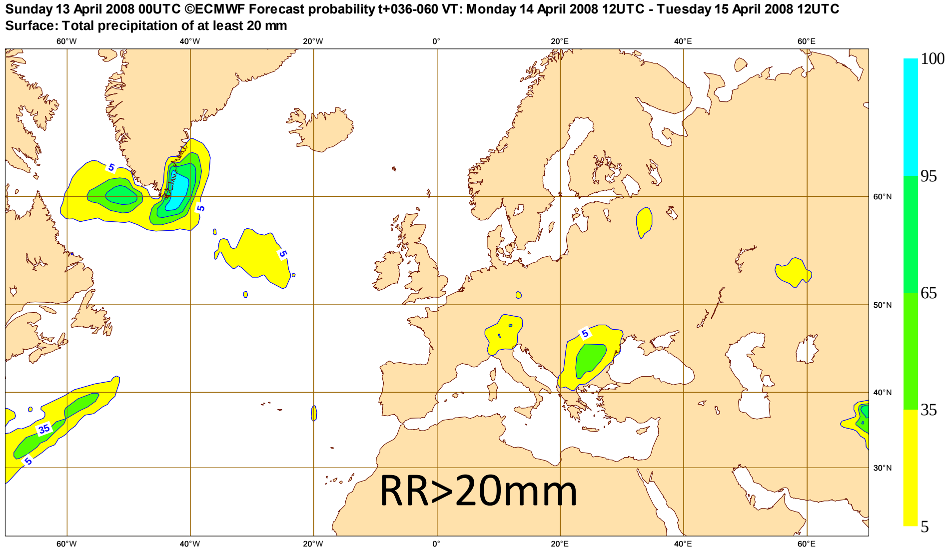


Probability Maps









Communication Checklist

1. Why was this work done?
2. For whom was it done?
3. To whom do you want to communicate information about the work?
4. Why would they be interested?
5. What information for what audiences?
6. Who may benefit from the work?
7. Who originated it?

Communication Checklist

8. What exchange, style and content of memoranda were needed to clarify the purpose of the project?
9. What communication measures were needed to establish high quality and timely data collection?
10. What support was needed from colleagues or specialists?
11. What progress memoranda and reports were written and for whom?

Tables

- Right justify numbers in tables;
- Line up decimal points in columns;
- Round numbers so that the two most effective digits are visible;
- Avoid distortion of the information in the data;
- Add rows and column averages or total where these are appropriate and may help;
- Consider re-ordering rows and/or columns to make the table clearer;
- Consider transposing the table;
- Give attention to the spacing and layout of the table.

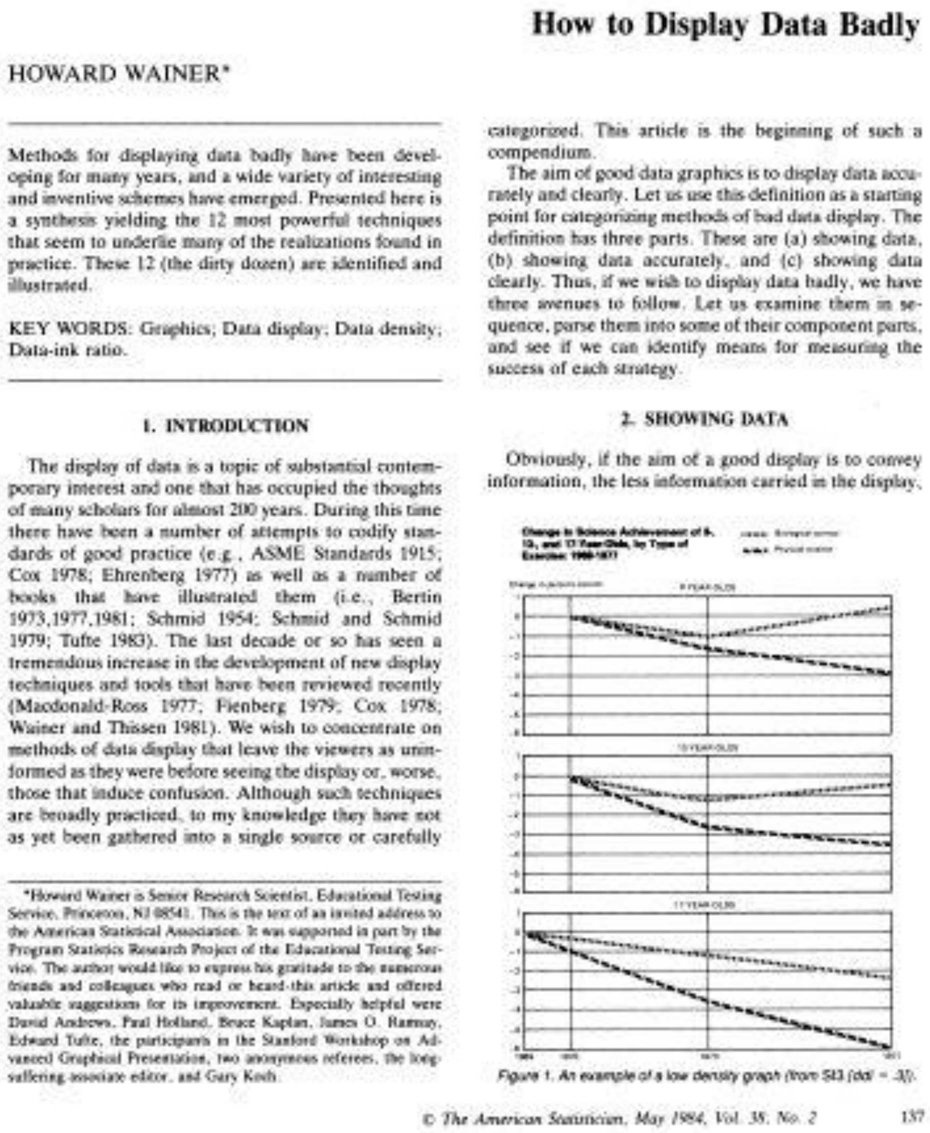
Graphics

- Use graphs when the shape of the data, such as trends or groups, are more important than exact values;
- Be sure that the graphic shows the data, so that you persuade the reader to think about the substance rather than the methodology or graphic design;
- Design the graphic so that it encourages the reader's eye to compare different pieces of data;
- Reveal the data at several levels of detail, from a broad overview to the fine structure
- Give every graph a clear, self-explanatory title

Graphics

- State all measurement units;
- Choose scales on graphs carefully;
- Label axes clearly;
- Avoid chart junk;
- Improve by trial-and-error since you rarely get the graphic right first time;
- Beware of the graphic artist who aims to beautify the image but fails to elucidate the data. So insist on checking the figures after the artist has done the work.
- Beware of misleading scales.

"How to Display Data Badly" (1984) H. Wainer, *The American Statistician*, vol 38, pp 137-147.



How to Display Data Badly

HOWARD WAINER*

Methods for displaying data badly have been developing for many years, and a wide variety of interesting and inventive schemes have emerged. Presented here is a synthesis yielding the 12 most powerful techniques that seem to underlie many of the realizations found in practice. These 12 (the dirty dozen) are identified and illustrated.

KEY WORDS: Graphics; Data display; Data density; Data-ink ratio.

1. INTRODUCTION

The display of data is a topic of substantial contemporary interest and one that has occupied the thoughts of many scholars for almost 200 years. During this time there have been a number of attempts to codify standards of good practice (e.g., ASME Standards 1915; Cox 1978; Ehrenberg 1977) as well as a number of books that have illustrated them (i.e., Bertin 1973,1977,1981; Schmid 1954; Schmid and Schmid 1979; Tufte 1983). The last decade or so has seen a tremendous increase in the development of new display techniques and tools that have been reviewed recently (Macdonald-Ross 1977; Fienberg 1979; Cox 1978; Wainer and Thissen 1981). We wish to concentrate on methods of data display that leave the viewers as uninformed as they were before seeing the display or, worse, those that induce confusion. Although such techniques are broadly practiced, to my knowledge they have not as yet been gathered into a single source or carefully categorized. This article is the beginning of such a compendium.

The aim of good data graphics is to display data accurately and clearly. Let us use this definition as a starting point for categorizing methods of bad data display. The definition has three parts. These are (a) showing data, (b) showing data accurately, and (c) showing data clearly. Thus, if we wish to display data badly, we have three avenues to follow. Let us examine them in sequence, parse them into some of their component parts, and see if we can identify means for measuring the success of each strategy.

2. SHOWING DATA

Obviously, if the aim of a good display is to convey information, the less information carried in the display,

Figure 1. An example of a low density graph (from S13 (ddi = .3)).

*Howard Wainer is Senior Research Scientist, Educational Testing Service, Princeton, NJ 08541. This is the text of an invited address to the American Statistical Association. It was supported in part by the Program Statistics Research Project of the Educational Testing Service. The author would like to express his gratitude to the numerous friends and colleagues who read or heard this article and offered valuable suggestions for its improvement. Especially helpful were David Andrews, Paul Holland, Bruce Kaplan, James O. Ramsay, Edward Tufte, the participants in the Stanford Workshop on Advanced Graphical Presentation, two anonymous referees, the long-suffering associate editor, and Gary Koch.

© The American Statistician, May 1984, Vol. 38, No. 2 137

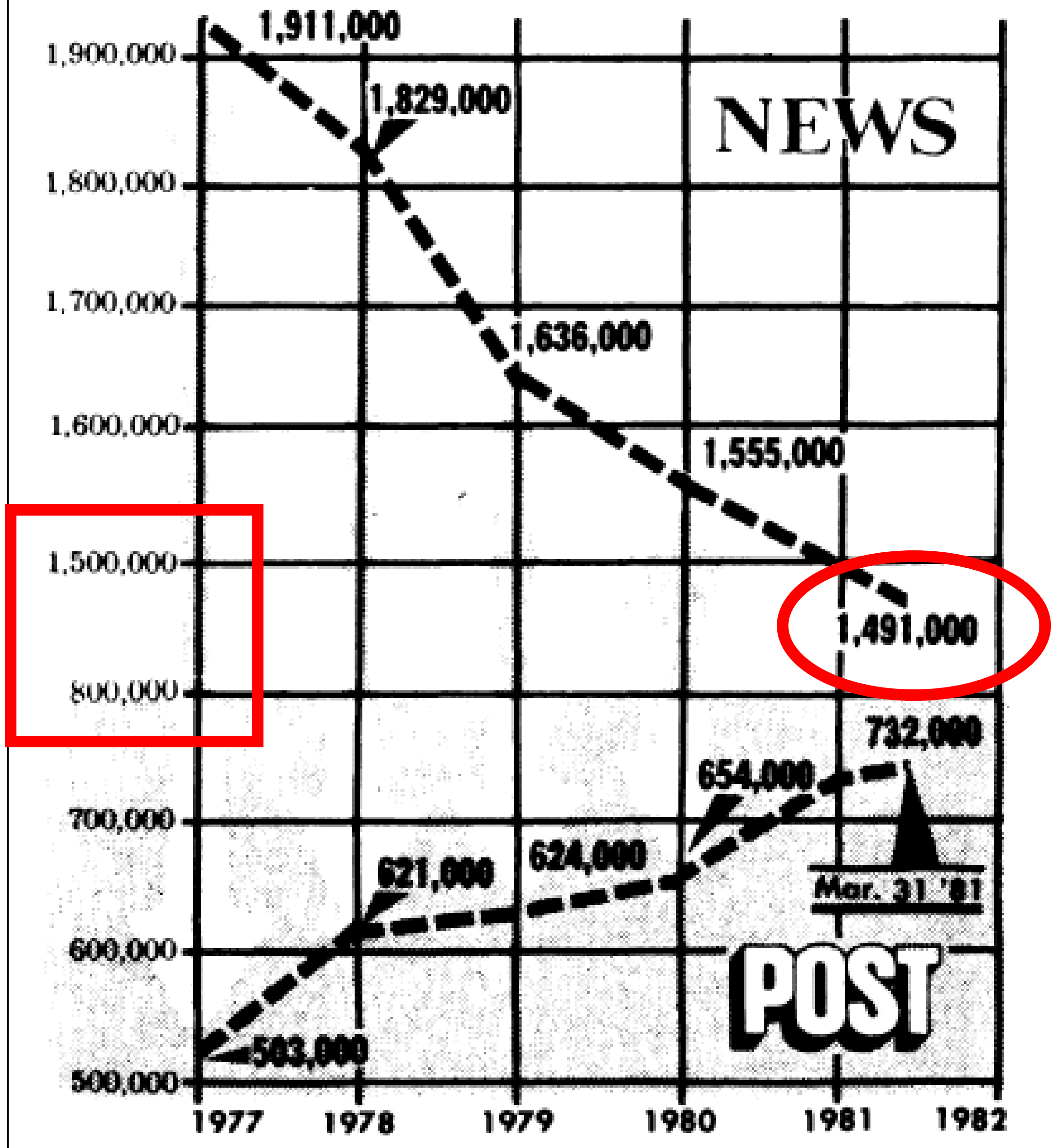
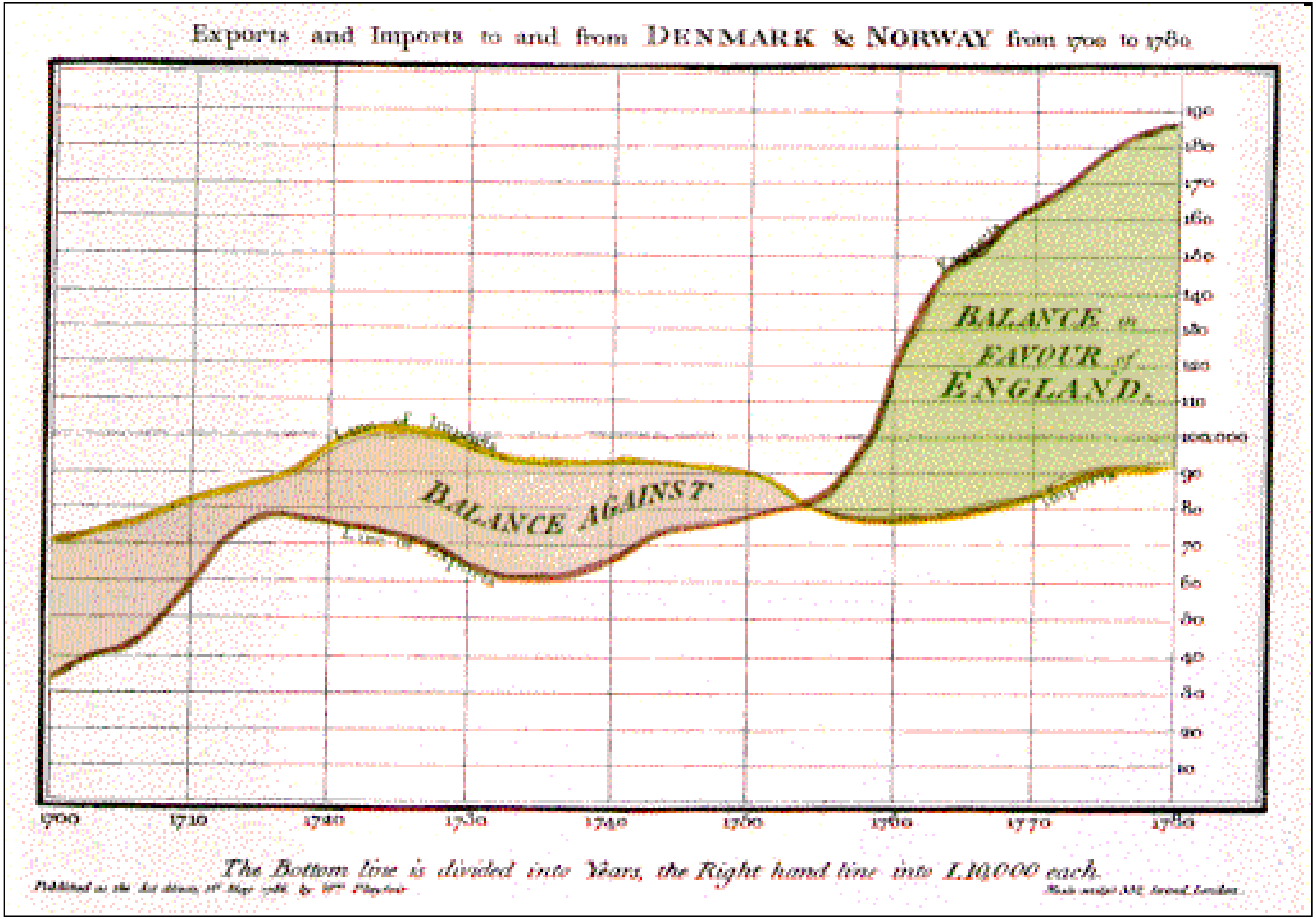
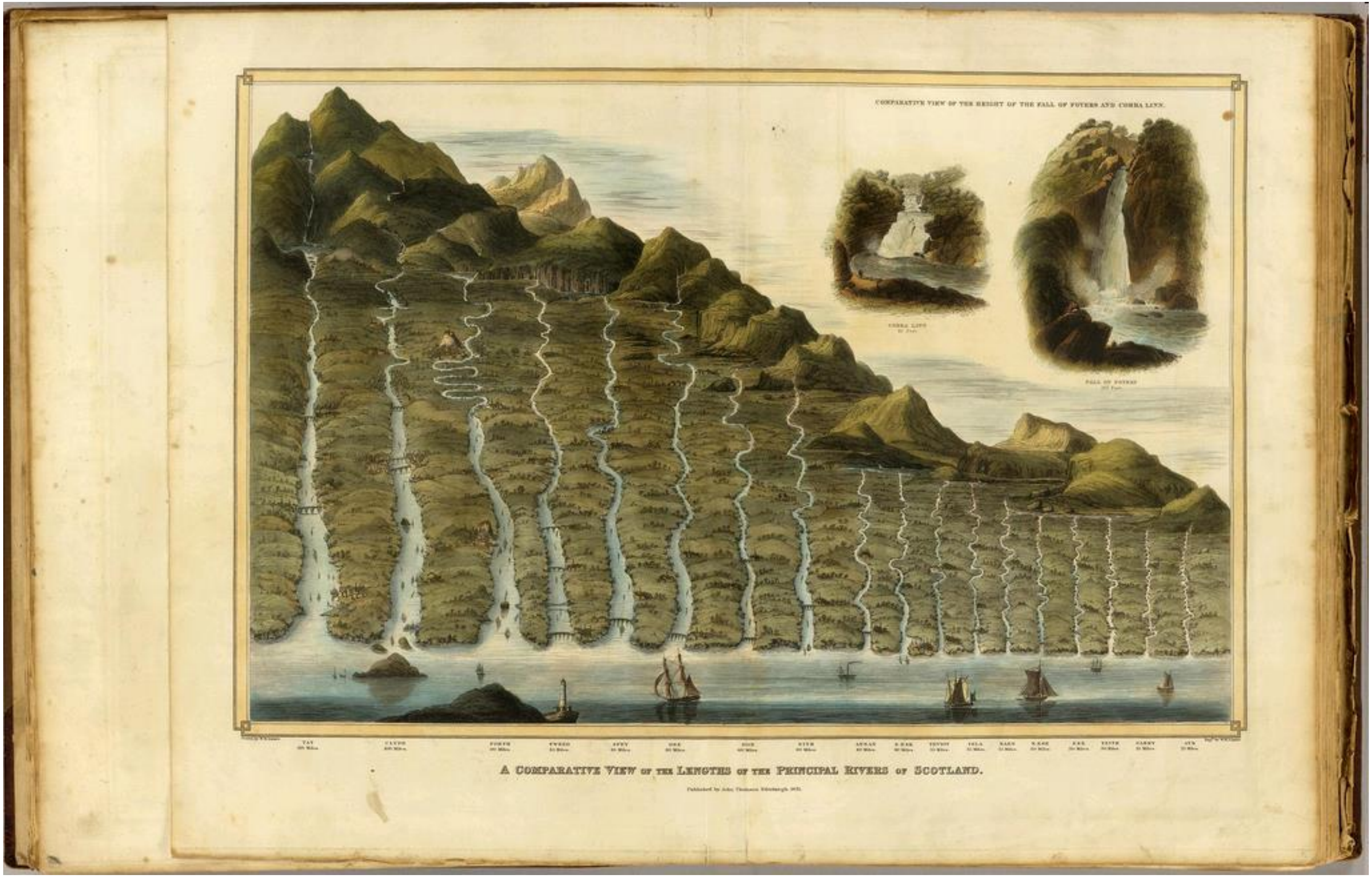


Figure 12. Changing scale in mid-axis to make large differences small (© 1981, New York Post).

William Playfair's trade-balance time-series chart, published in his Commercial and Political Atlas, 1786



A comparative view of the lengths of the principal rivers of Scotland with comparative view of the height of the falls of Foyers and Corra Linn (1831).

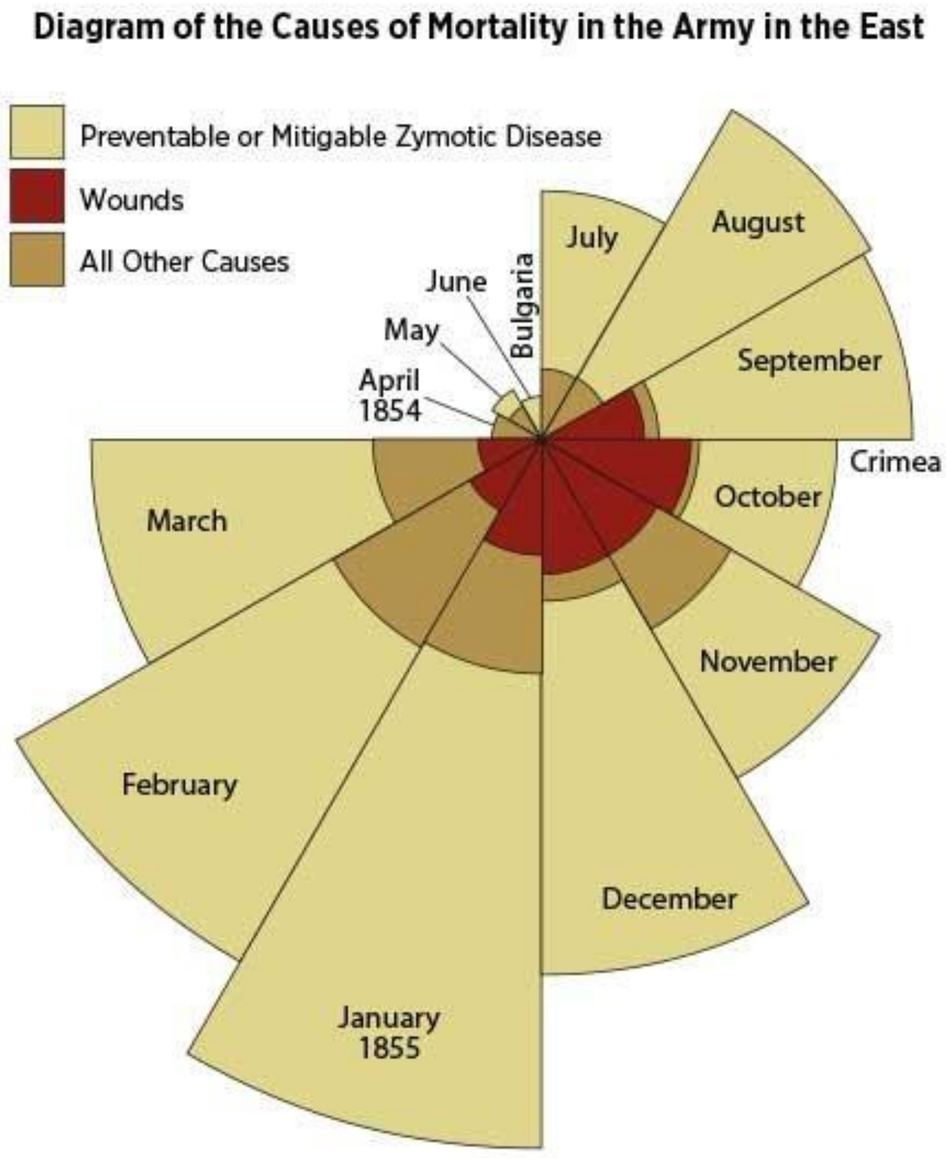


After witnessing deplorable sanitary conditions in the Crimea, Florence Nightingale wrote *Notes on Matters Affecting the Health, Efficiency and Hospital Administration of the British Army* (1858), including several graphs of her own design, which she called "Coxcombs". This figure makes it clear that far more deaths were attributable to non-battle causes ("preventable causes") than to battle-related causes

TABLE SHOWING the ESTIMATED AVERAGE MONTHLY STRENGTH of the ARMY; and the Deaths and Annual Rate of Mortality per 1,000, in each Month, from April, 1854, to March, 1856, (inclusive), in the Hospitals of the Army in the East.

Months	Estimated Average Monthly Strength of the Army.	Deaths.			Annual Rate of Mortality per 1,000.		
		Zymotic Diseases.	Wounds and Injuries.	All other Causes.	Zymotic Diseases.	Wounds and Injuries.	All other Causes.
1854 April	8,571	1	..	5	1·4	..	7·0
May	23,333	12	..	9	6·2	..	4·6
June..	28,333	11	..	6	4·7	..	2·5
July..	28,722	359	..	23	150·0	..	9·6
August	30,246	828	1	30	328·5	·4	11·9
September ..	30,290	788	81	70	312·2	32·1	27·7
October	30,643	503	132	128	197·0	51·7	50·1
November ..	29,736	844	287	106	340·6	115·8	42·8
December ..	32,779	1,725	114	131	631·5	41·7	48·0
1855 January	32,393	2,761	83	324	1022·8	30·7	120·0
February	30,919	2,120	42	361	822·8	16·3	140·1
March	30,107	1,205	32	172	480·3	12·8	68·6
April	32,252	477	48	57	177·5	17·9	21·2
May	35,473	508	49	37	171·8	16.6	12·5
June..	38,863	802	209	31	247·6	64·5	9·6
July	42,647	382	134	33	107·5	37·7	9·3
August	44,614	483	164	25	129·9	44·1	6·7
September ..	47,751	189	276	20	47·5	69·4	5 0
October	46,852	128	53	18	32·8	13·6	4·6
November ..	37,853	178	33	32	56·4	10·5	10·1
December ..	43,217	91	18	28	25·3	5·0	7·8
1856 January	44,212	42	2	48	11·4	·5	13·0
February	43,485	24	..	19	6·6	..	5·2
March	46,140	15	..	35	3·9	..	9·1

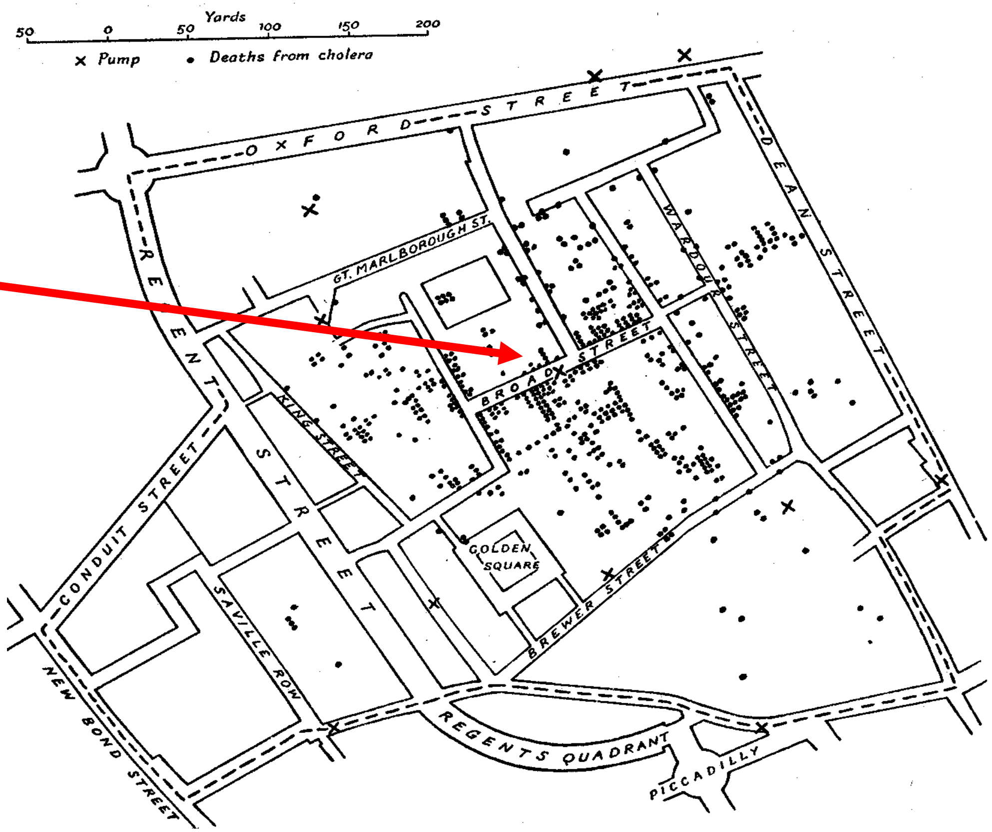
The Deaths under the head of "Wounds and Injuries," comprise the following causes:—Luxatio, Sub-Luxatio, Vulnus Sclopitorum, Vulnus Incisum, Contusio, Fractura, Ambustio, and Concussio Cerebri.



The black line acoss November 1854 marks the boundary of the deaths from all other causes during that month. In October 1854, the black coincides with the red.

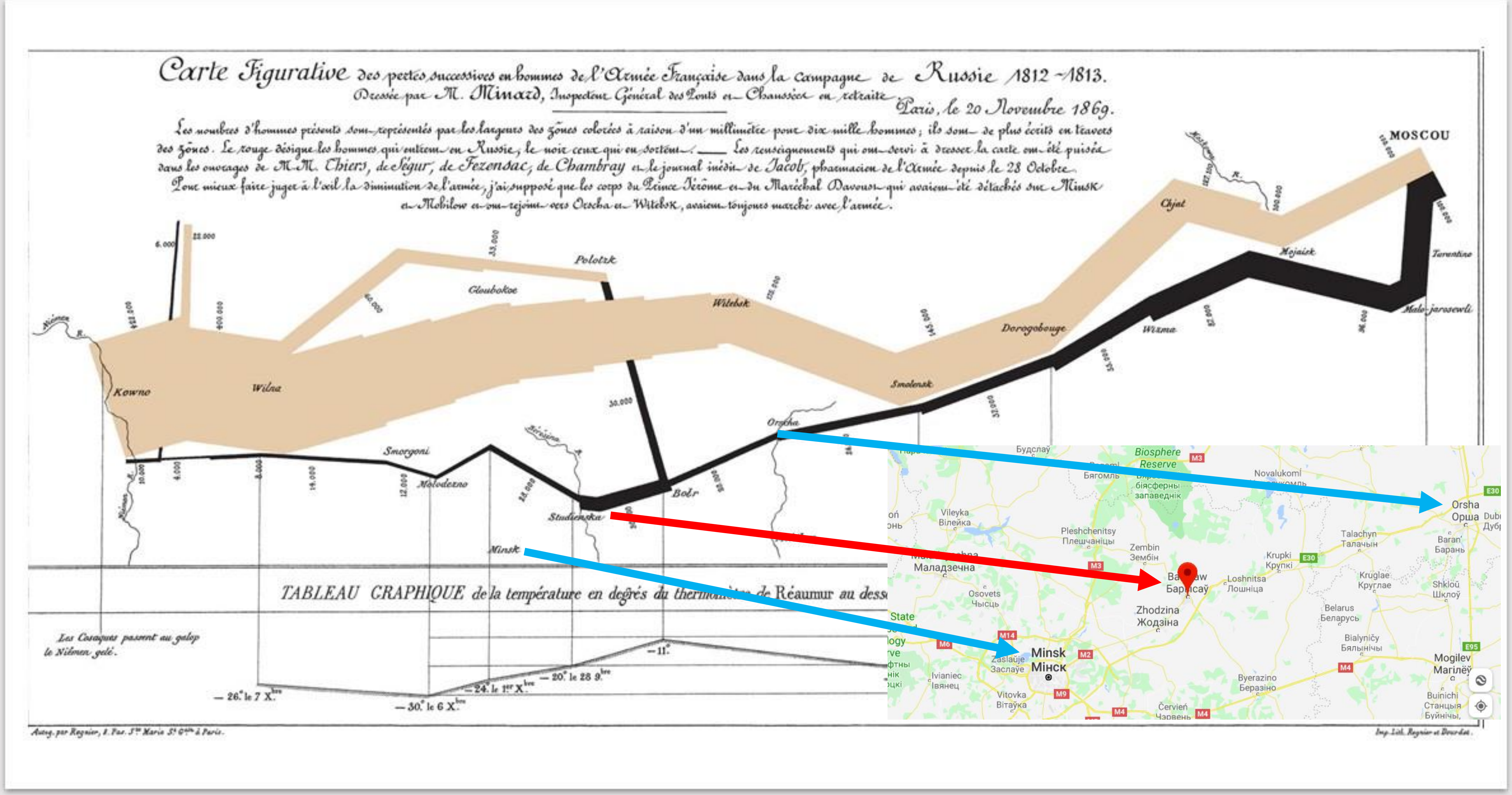
Florence Nightingale 1856

The cholera outbreak in Soho, England, in 1854.
John Snow (1813 –1858)

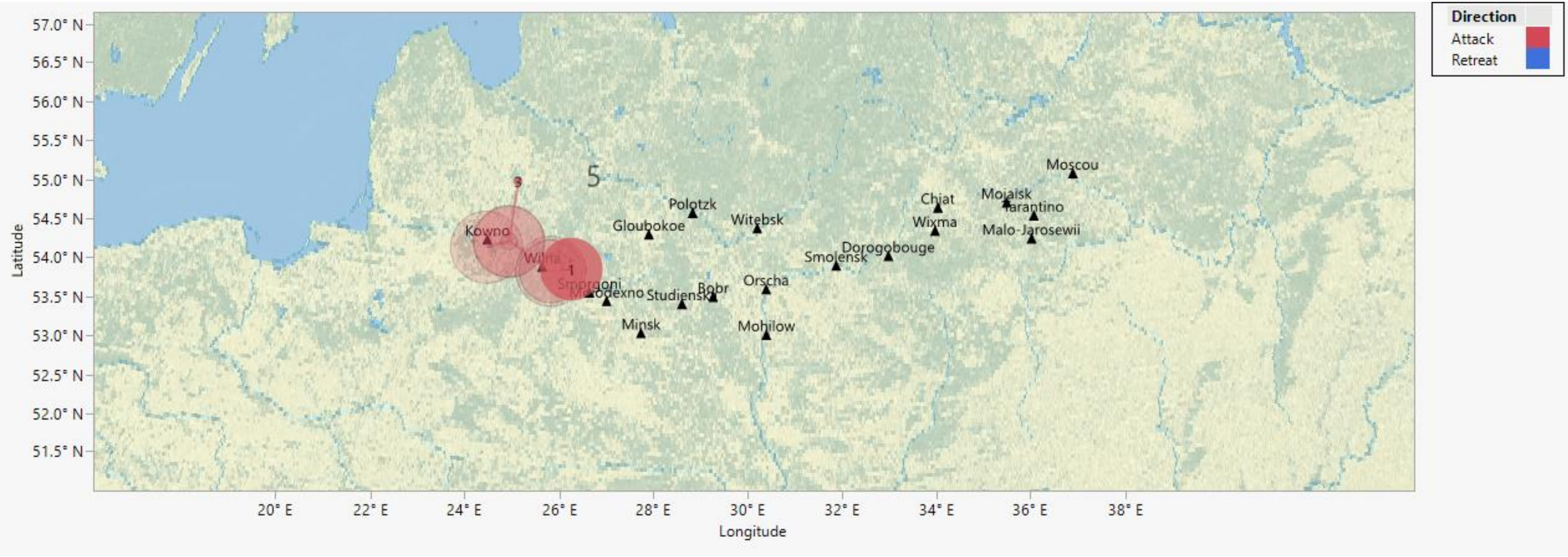


Hans Rosling

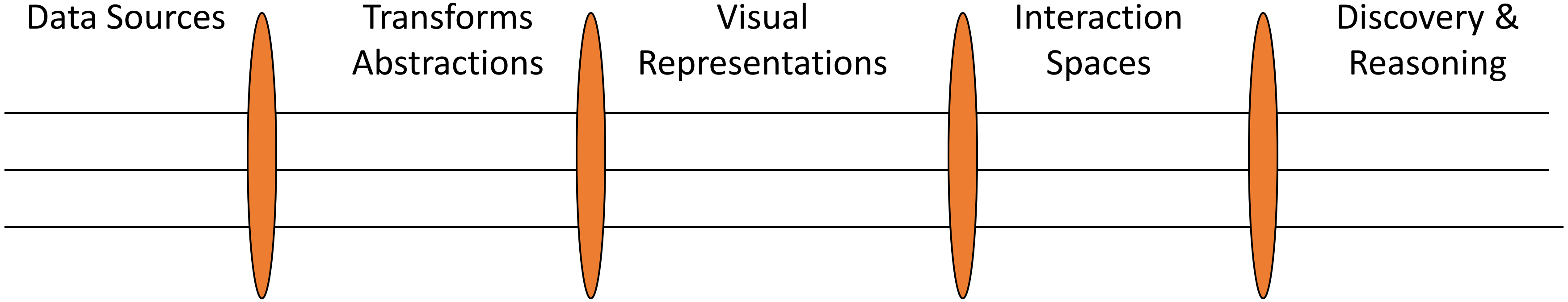
http://www.gapminder.org/videos/ted-talks/hans-rosling-ted-2006-debunking-myths-about-the-third-world/



Carte Figurative des pertes successives en hommes de l'Armée Française dans la campagne de Russie 1812–1813.
Dressée par M. Minard, Inspecteur Général des Ponts et Chaussées en retraite.
Paris, le 20 Novembre 1869.
Les nombres d'hommes présents sont représentés par les largeurs des zônes colorées à raison d'un millimètre pour dix mille hommes; ils sont de plus écrits en travers des zônes. Le rouge désigne les hommes qui entrent en Russie, le noir ceux qui en sortent. — Les renseignements qui ont servi à dresser la carte ont été puisés dans les ouvrages de M.M. Thiers, de Ségur, de Fezensac, de Chambray et le journal inédit de Jacob, pharmacien de l'Armée depuis le 28 Octobre.
Pour mieux faire juger à l'oeil la diminution de l'armée, j'ai supposé que les corps du Prince Jérôme et du Maréchal Davoust qui avaient été détachés sur Minsk et Mobilow et ont rejoint vers Orscha et Witebsk, avaient toujours marché avec l'armée.
MOSCOU
Chjat
Mojaisk
Tarantino
Malo-jarosewli
Wixma
Dorogobouge
Smolensk
Witebsk
Polotzk
Gloubokoe
Wilna
Kowno
Niemen
Smorgoni
Molodezno
Studienska
Botr
Minsk
Orscha
6.000
22.000
33.000
60.000
422.000
400.000
175.000
145.000
127.100
100.000
100.000
96.000
87.000
55.000
37.000
30.000
28.000
24.000
14.000
12.000
10.000
8.000
4.000
TABLEAU GRAPHIQUE de la température en degrés du thermomètre de Réaumur au dessous de zéro
Les Cosaques passent au galop le Niémen gelé.
– 11°
– 20° le 28 9bre
– 24° le 1er Xbre
– 30° le 6 Xbre
– 26° le 7 Xbre
Autog. par Regnier, 8. Pas. Ste Marie St Gain à Paris.
Imp. Lith. Regnier et Dourdet.
Orsha
Орша
Minsk
Мінск
Barysaw
Барысаў
Zhodzina
Жодзіна
Maladzyechna
Маладзечна
Vileyka
Вілейка
Pleshchenitsy
Плешчаніцы
Zembin
Зембін
Loshnitsa
Лошніца
Krupki
Крупкі
Talachyn
Талачын
Kruglae
Круглае
Shkloŭ
Шклоў
Belarus
Беларусь
Bialyničy
Бялынічы
Mogilev
Магілёў
Byerazino
Беразіно
Osovets
Чысць
Zaslaŭje
Заслаўе
Ivianiec
Івянец
Vitovka
Вітаўка
Červien
Чэрвень
Novalukoml
Budslaŭ
Будслаў
Biosphere Reserve
Baran'
Барань


Direction
Attack
Retreat
57.0° N
56.5° N
56.0° N
55.5° N
55.0° N
54.5° N
54.0° N
53.5° N
53.0° N
52.5° N
52.0° N
51.5° N
Latitude
20° E
22° E
24° E
26° E
28° E
30° E
32° E
34° E
36° E
38° E
Longitude
Moscou
Mojaisk
Tarantino
Chjat
Wixma
Malo-Jarosewii
Dorogobouge
Smolensk
Witebsk
Polotzk
Gloubokoe
Orscha
Bobr
Studienska
Molodexno
Smorgoni
Minsk
Mohilow
Kowno
Wilna
5

Visual Analytics



Data Sources

-Files
-Databases
-Numeric
-Nominal

-Quality
-Uncertainty
-Missing values

Transforms Abstractions

-Clustering
-Sampling
-Nominal to ordinal
-Dimension reduction

-Clutter reduction

Visual Representations

-Data (multiple)
-Statistics
-Structure (hierarchy)

-Data quality
-Abstraction quality
-Anomalies

Interaction Spaces

-Data
-Structure (hierarchy)

-Spatial
-Temporal
-Quality

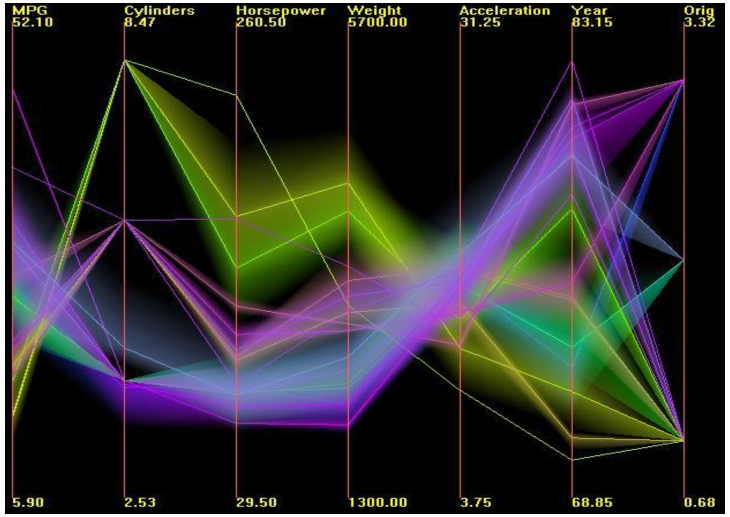
Discovery & Reasoning

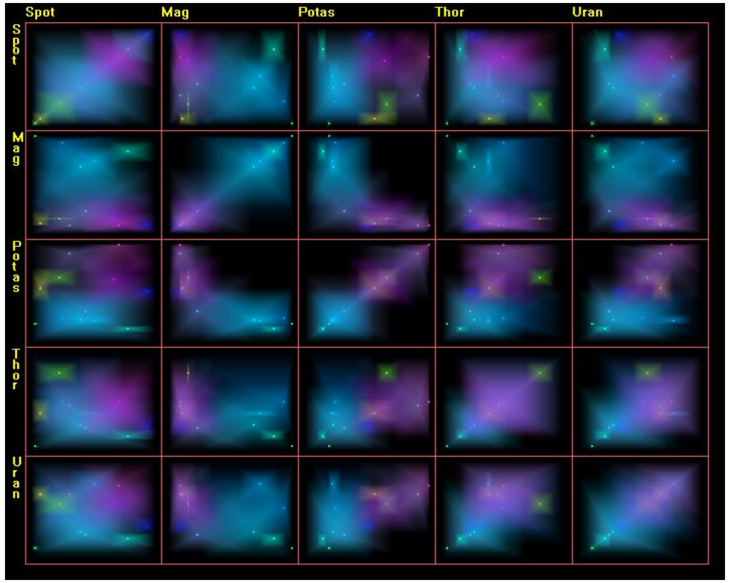
-Clusters
-Associations

-Trends
-Hypotheses
-Outliers

Multiresolution Visualization

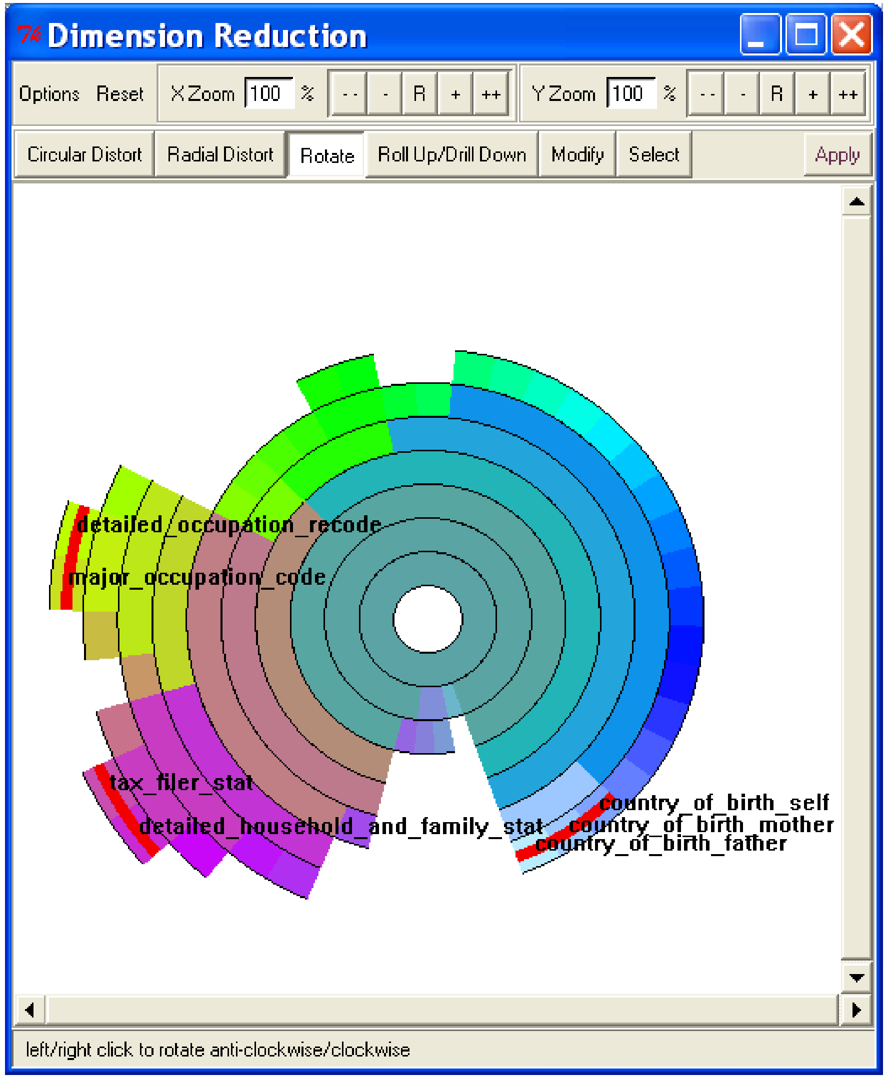
- For large datasets, visualizations quickly get cluttered
- Hierarchical clustering generates many levels of detail
- User can select areas of interest to view at full resolution while the rest of the data is shown via cluster centers and extents (shown as bands of variable opacity)





Dimension Reduction

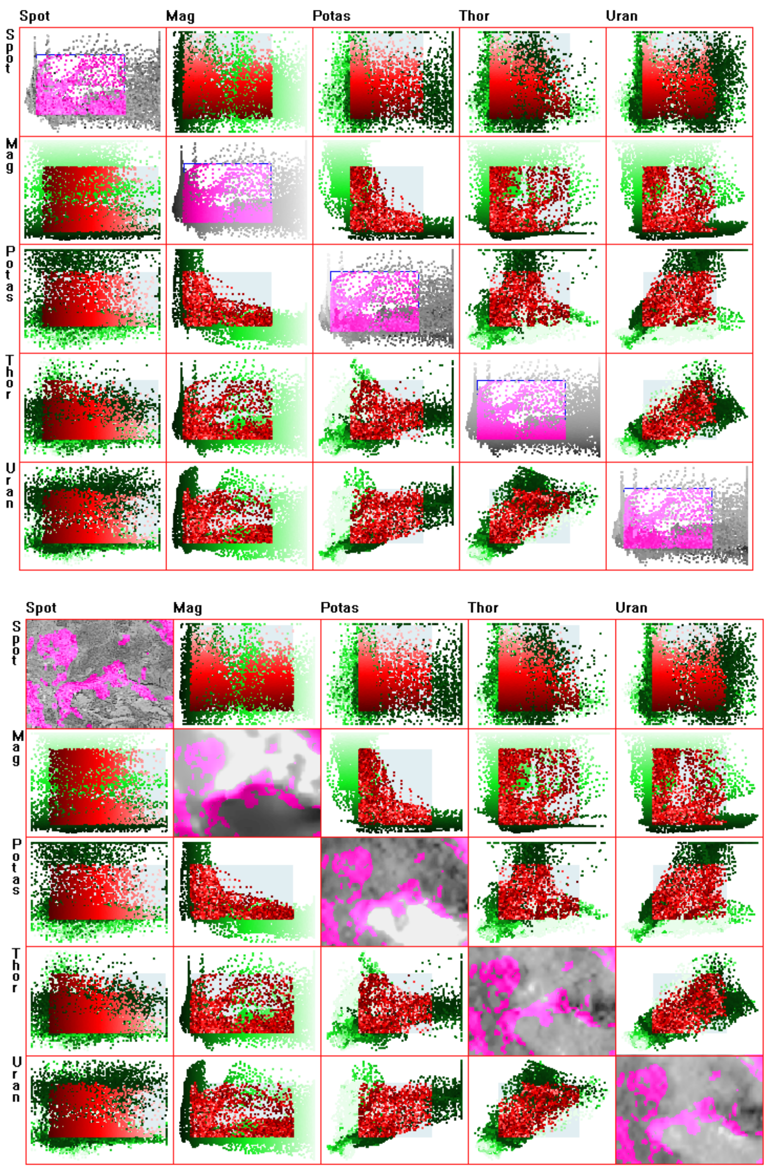
- Dimensions are hierarchically clustered based on similarity measures
- Hierarchy displayed using Inter Ring
- Users select clusters of dimensions or representative dimensions for detailed analysis



42 dimension census dataset.

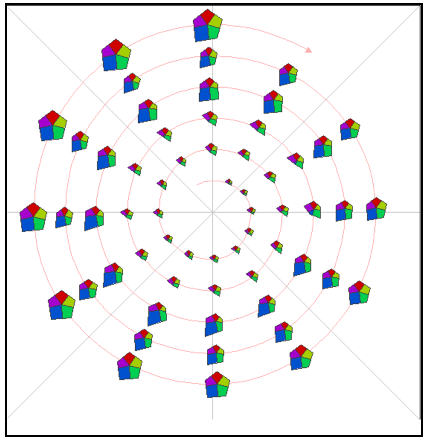
Linking Spatial and Non-Spatial

- Diagonal plots of scatterplot matrix can have numerous uses
- Example shows multispectral remote sensing data, 1 layer per diagonal plot
- User can select in either 2-D or parameter space and see corresponding elements in other views.

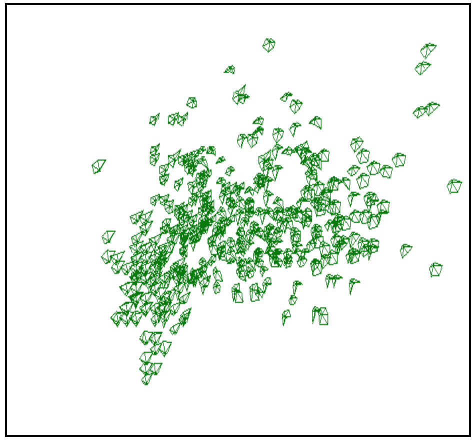


Layout Strategies

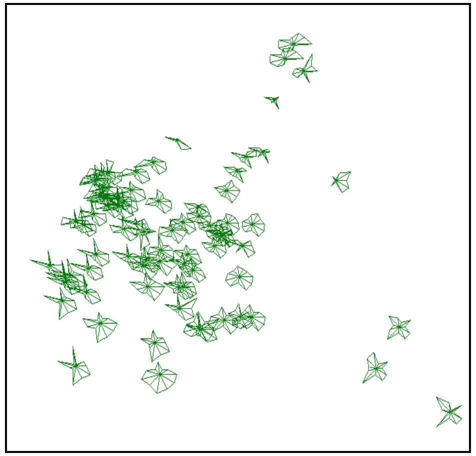
- Different layout strategies can reveal different patterns in the data
- Detecting, classifying, and measuring trends, outliers, repeated patterns, clusters, and correlations can be facilitated via specific layouts



Cyclic



Data Driven

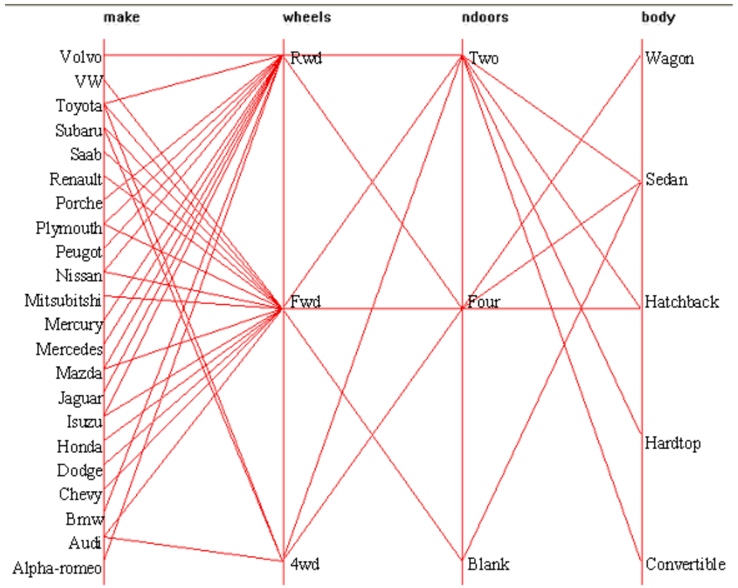


Principal Components

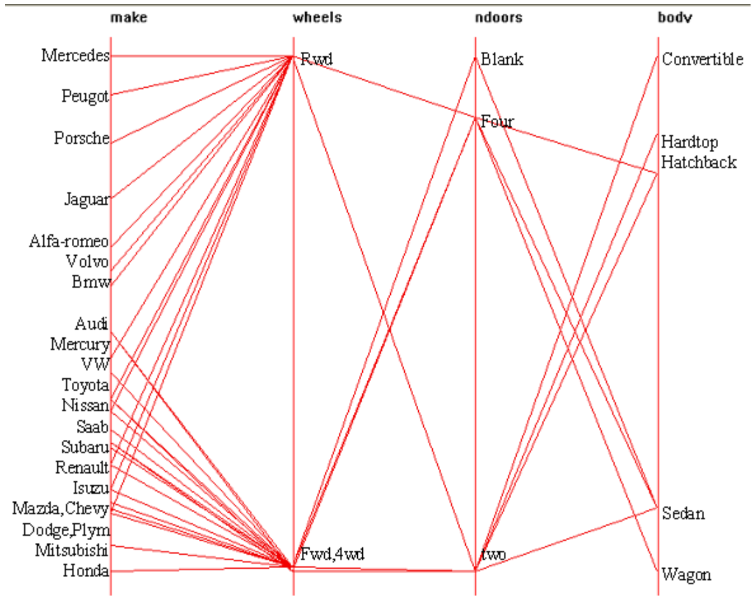
Order Driven

Visualizing Data with Parallel Plots

- Arbitrary assignment of non-numeric fields to numbers can lead to misinterpretation, lost patterns
- By looking at similarities in distributions across all dimensions, we can group values of a nominal variable with similar global characteristics
- Assignments used to convey order and relative distance



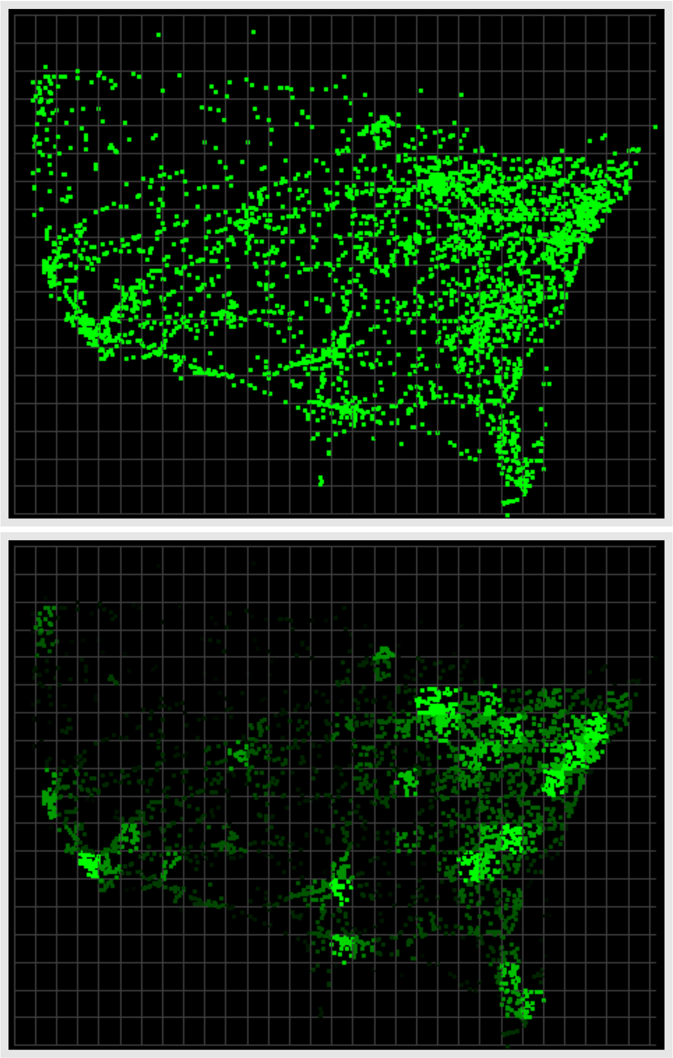
Original Assignment



Assignment after Correspondence Analysis

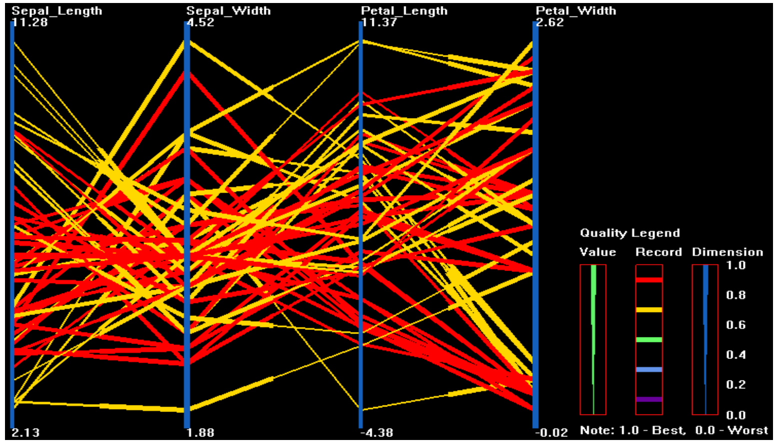
Visual Clutter Reduction

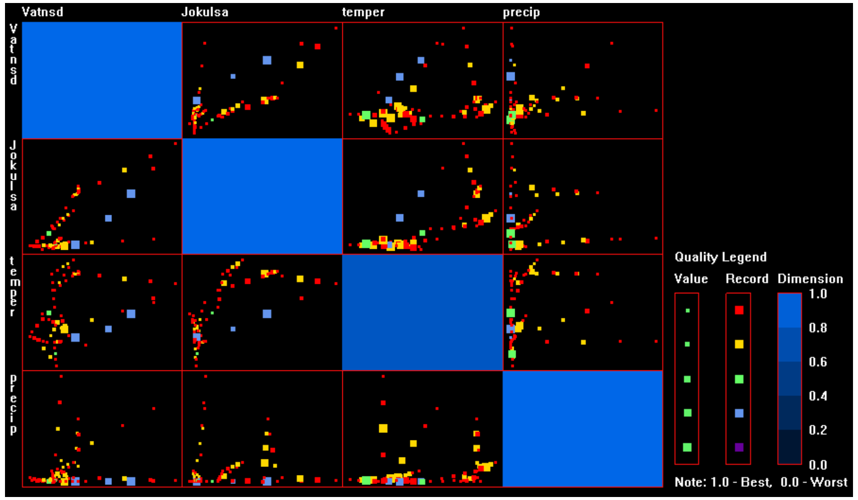
- In scenes with thousands of moving objects, there is need to reduce clutter
- Many strategies, including:
  - Information-preserving
  - Information-reducing
  - Visual remapping



Data Quality Visual Encoding

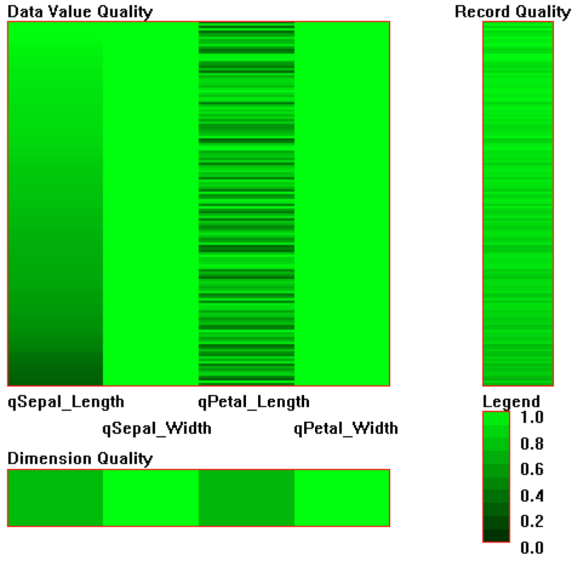
- Data quality refers to the degree of uncertainty of data
- Quality measures are visually encoded into existing visualizations
- This helps users focus on high quality data to draw reliable conclusions



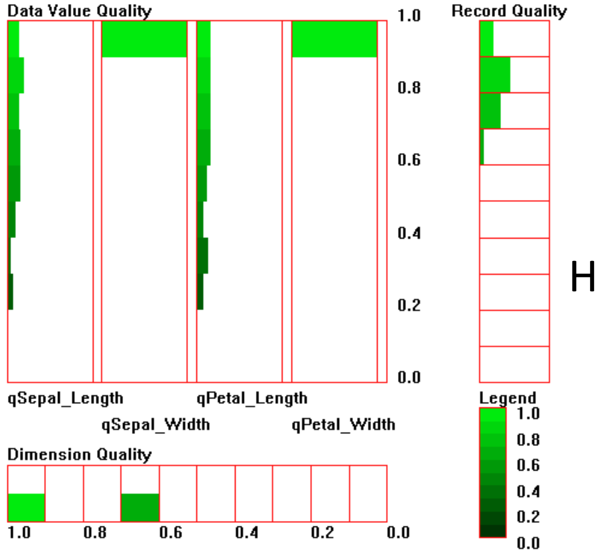


Quality Space Visualization

- Quality space is visualized separately to convey patterns in the data quality measures
- Records or dimensions can be ordered by quality to reveal structure and relations
- Stripe view shows individual data value quality; Histogram view shows summarization and distribution



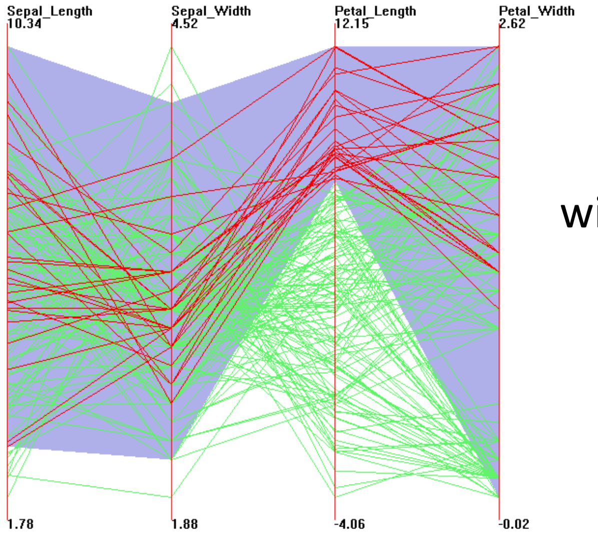
Stripe Quality Map



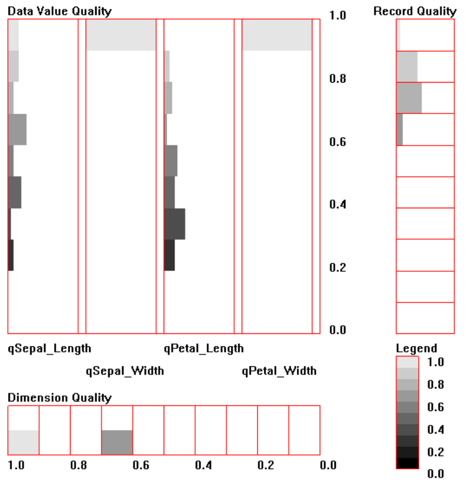
Histogram Quality Map

Interactions between Data Space and Quality Space

- Linking brush: When users select a subset in one space, the corresponding subset in the other space will be highlighted accordingly.
- Sample figures: The data points in the data space with high values in the third dimension are highlighted, then the distribution of quality measures for this subset is rendered in the quality map.

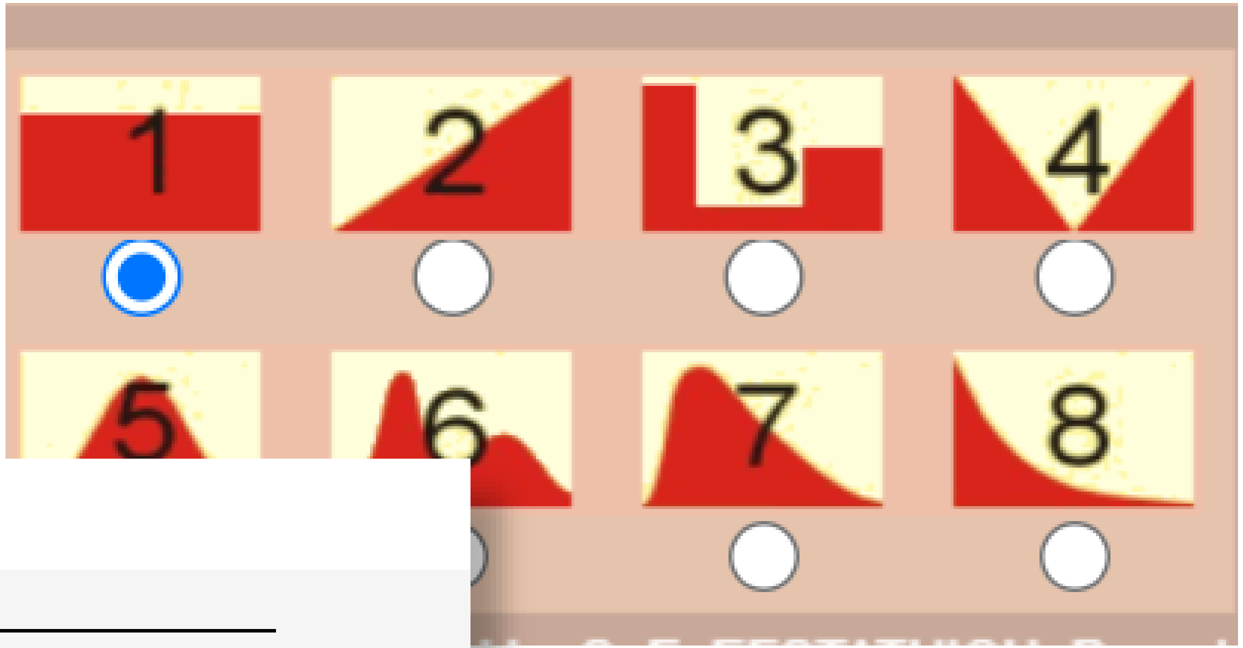


Data space
with highlighting

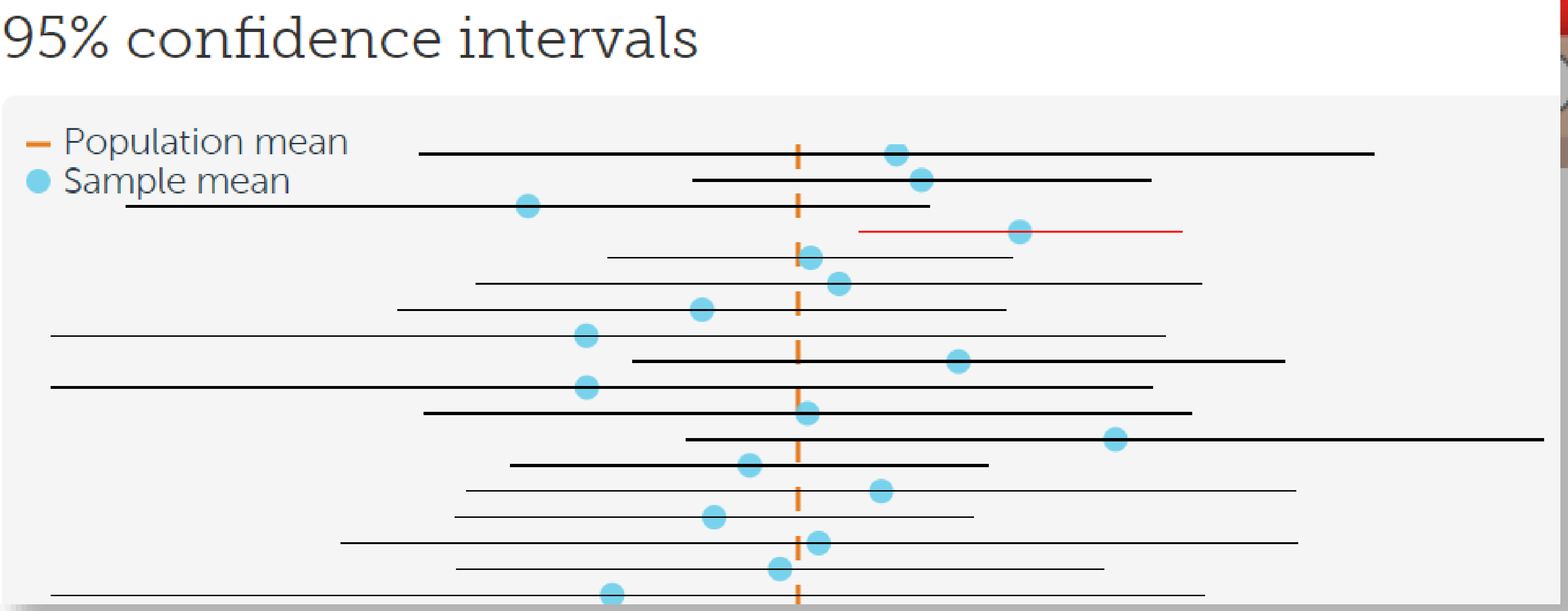


Linked
Quality space

http://195.134.76.37/applets/AppletCentralLimit/Appl_CentralLimit2.html

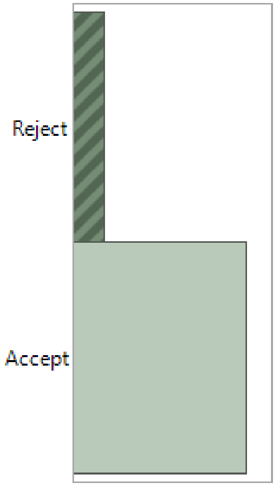


https://rpsychologist.com/d3/CI/



Distributions

Lot Acceptance

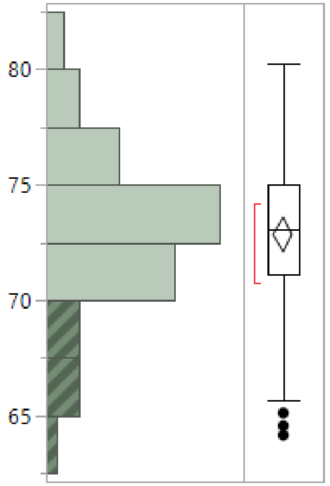


Frequencies

Level	Count	Prob
Accept	76	0.84444
Reject	14	0.15556
Total	90	1.00000

N Missing 0
2 Levels

Disso



Quantiles

100.0%	maximum	80.23
99.5%		80.23
97.5%		80.1025
90.0%		77.532
75.0%	quartile	74.9925
50.0%	median	73.05
25.0%	quartile	71.11
10.0%		67.844
2.5%		64.704
0.5%		64.15
0.0%	minimum	64.15

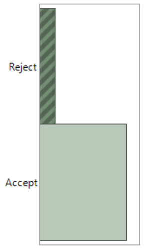
Summary Statistics

Mean	72.860556
Std Dev	3.5121345
Std Err Mean	0.3702115
Upper 95% Mean	73.596158
Lower 95% Mean	72.124953
N	90

Tablet Production.jmp

Distributions

Lot Acceptance

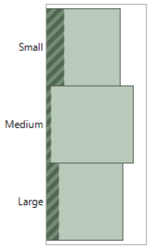


Frequencies

Level	Count	Prob
Accept	76	0.84444
Reject	14	0.15556
Total	90	1.00000

N Missing 0
2 Levels

API Particle Size

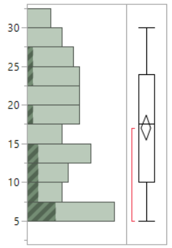


Frequencies

Level	Count	Prob
Large	29	0.32222
Medium	33	0.36667
Small	28	0.31111
Total	90	1.00000

N Missing 0
3 Levels

Mill Time



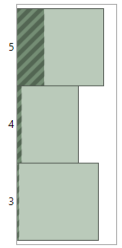
Quantiles

100.0%	maximum	30
99.5%		30
97.5%		30
90.0%		28
75.0%	quartile	24
50.0%	median	17.5
25.0%	quartile	10
10.0%		6.1
2.5%		5
0.5%		5
0.0%	minimum	5

Summary Statistics

Mean	17.011111
Std Dev	7.7669632
Std Err Mean	0.8187098
Upper 95% Mean	18.63787
Lower 95% Mean	15.384352
N	90

Screen Size

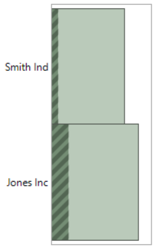


Frequencies

Level	Count	Prob
3	32	0.35556
4	24	0.26667
5	34	0.37778
Total	90	1.00000

N Missing 0
3 Levels

Mag. Stearate Supplier

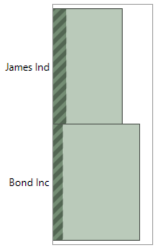


Frequencies

Level	Count	Prob
Jones Inc	49	0.54444
Smith Ind	41	0.45556
Total	90	1.00000

N Missing 0
2 Levels

Lactose Supplier

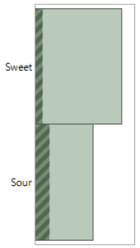


Frequencies

Level	Count	Prob
Bond Inc	50	0.55556
James Ind	40	0.44444
Total	90	1.00000

N Missing 0
2 Levels

Sugar Supplier

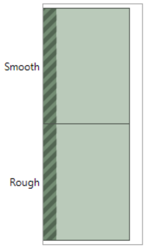


Frequencies

Level	Count	Prob
Sour	36	0.40000
Sweet	54	0.60000
Total	90	1.00000

N Missing 0
2 Levels

Talc Supplier

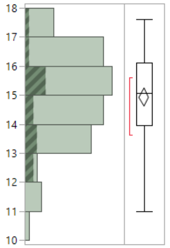


Frequencies

Level	Count	Prob
Rough	45	0.50000
Smooth	45	0.50000
Total	90	1.00000

N Missing 0
2 Levels

Blend Time



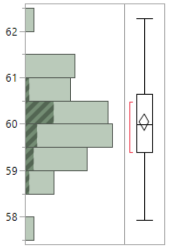
Quantiles

100.0%	maximum	17.61332799
99.5%		17.61332799
97.5%		17.52757757
90.0%		16.812313775
75.0%	quartile	16.09911638
50.0%	median	15.054428233
25.0%	quartile	13.953733668
10.0%		13.134801999
2.5%		11.298700455
0.5%		10.987701779
0.0%	minimum	10.987701779

Summary Statistics

Mean	14.922363
Std Dev	1.5031489
Std Err Mean	0.1584458
Upper 95% Mean	15.237192
Lower 95% Mean	14.607535
N	90

Blend Speed



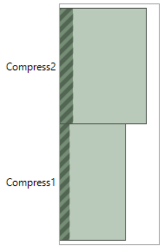
Quantiles

100.0%	maximum	62.270752055
99.5%		62.270752055
97.5%		61.934040863
90.0%		61.200184556
75.0%	quartile	60.64921934
50.0%	median	59.990524694
25.0%	quartile	59.397242304
10.0%		59.008650295
2.5%		58.150950737
0.5%		57.933175924
0.0%	minimum	57.933175924

Summary Statistics

Mean	60.044232
Std Dev	0.8474732
Std Err Mean	0.0893315
Upper 95% Mean	60.221732
Lower 95% Mean	59.866732
N	90

Compressor

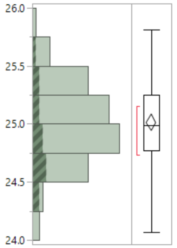


Frequencies

Level	Count	Prob
Compress1	39	0.43333
Compress2	51	0.56667
Total	90	1.00000

N Missing 0
2 Levels

Force



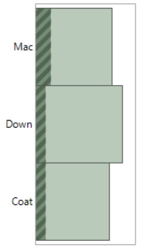
Quantiles

100.0%	maximum	25.809483784
99.5%		25.809483784
97.5%		25.629128071
90.0%		25.466161676
75.0%	quartile	25.251414744
50.0%	median	24.987746476
25.0%	quartile	24.769992107
10.0%		24.596742503
2.5%		24.253166051
0.5%		24.069607874
0.0%	minimum	24.069607874

Summary Statistics

Mean	25.012335
Std Dev	0.3341375
Std Err Mean	0.0352212
Upper 95% Mean	25.082318
Lower 95% Mean	24.942351
N	90

Coating Supplier

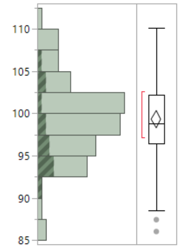


Frequencies

Level	Count	Prob
Coat	28	0.31111
Down	33	0.36667
Mac	29	0.32222
Total	90	1.00000

N Missing 0
3 Levels

Coating Viscosity

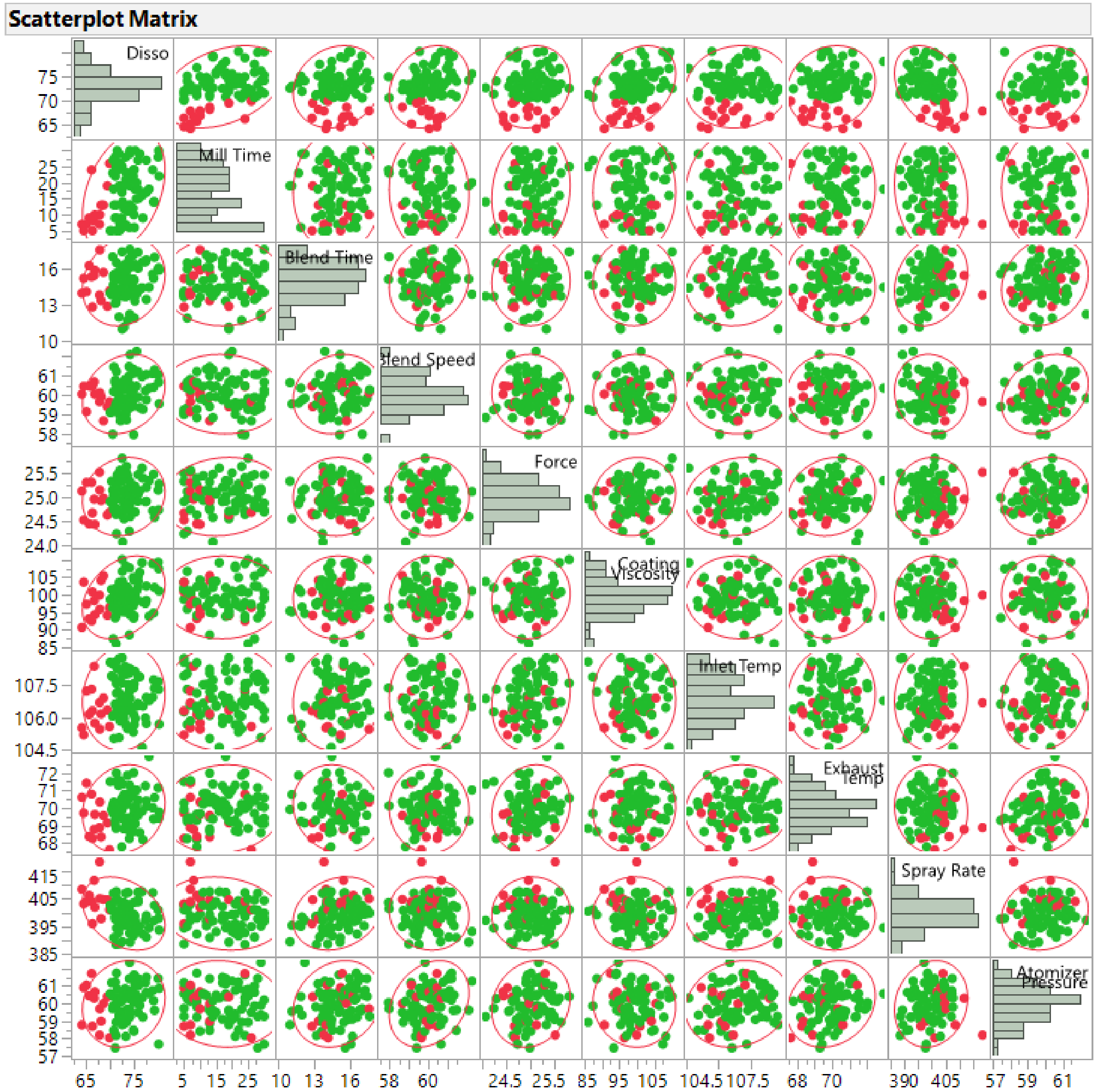


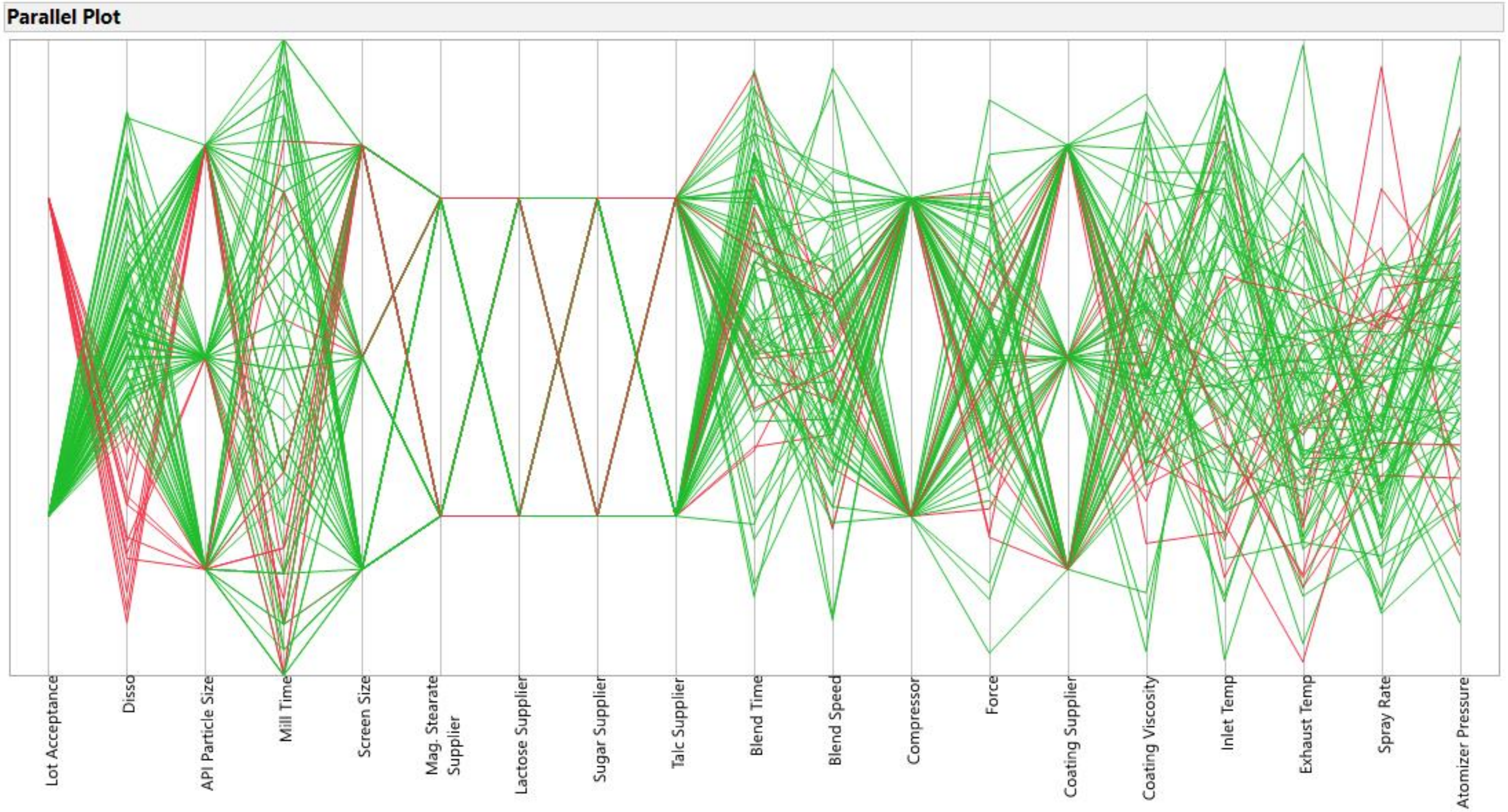
Quantiles

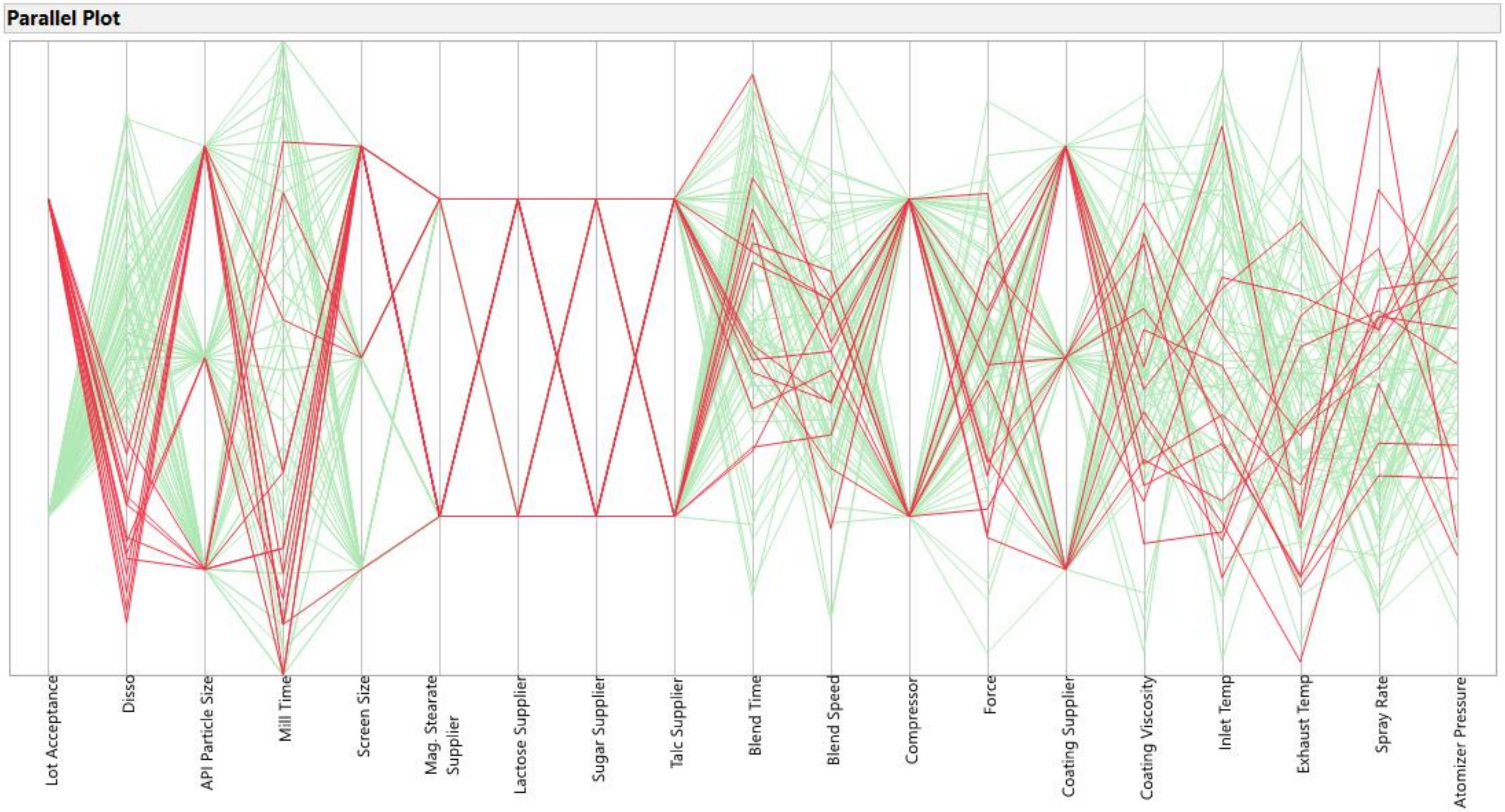
100.0%	maximum	110.12741925
99.5%		110.12741925
97.5%		109.3221013
90.0%		106.3656815
75.0%	quartile	102.18471735
50.0%	median	98.791015892
25.0%	quartile	96.401432091
10.0%		93.900898604
2.5%		87.737118076
0.5%		86.018346028
0.0%	minimum	86.018346028

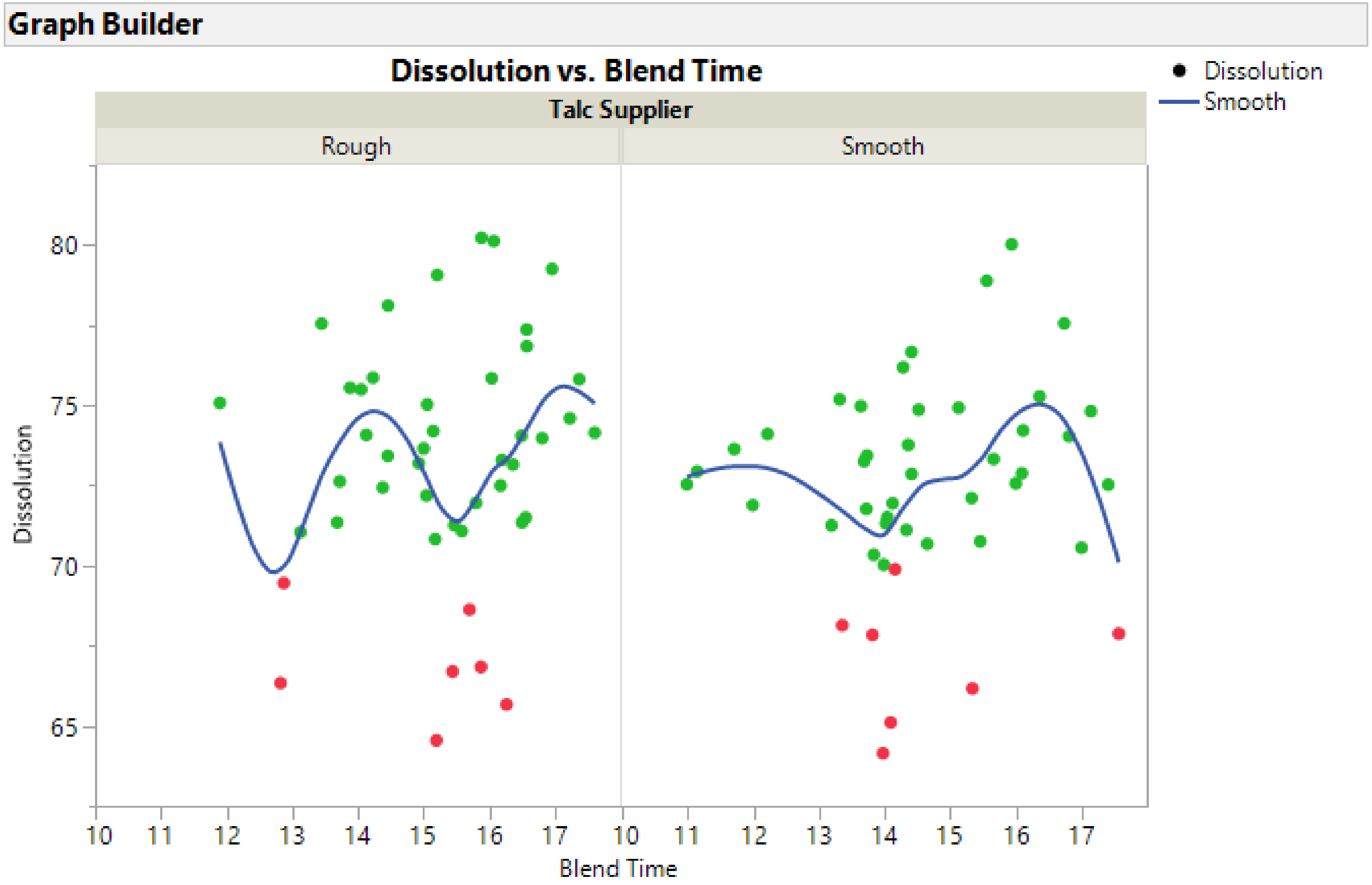
Summary Statistics

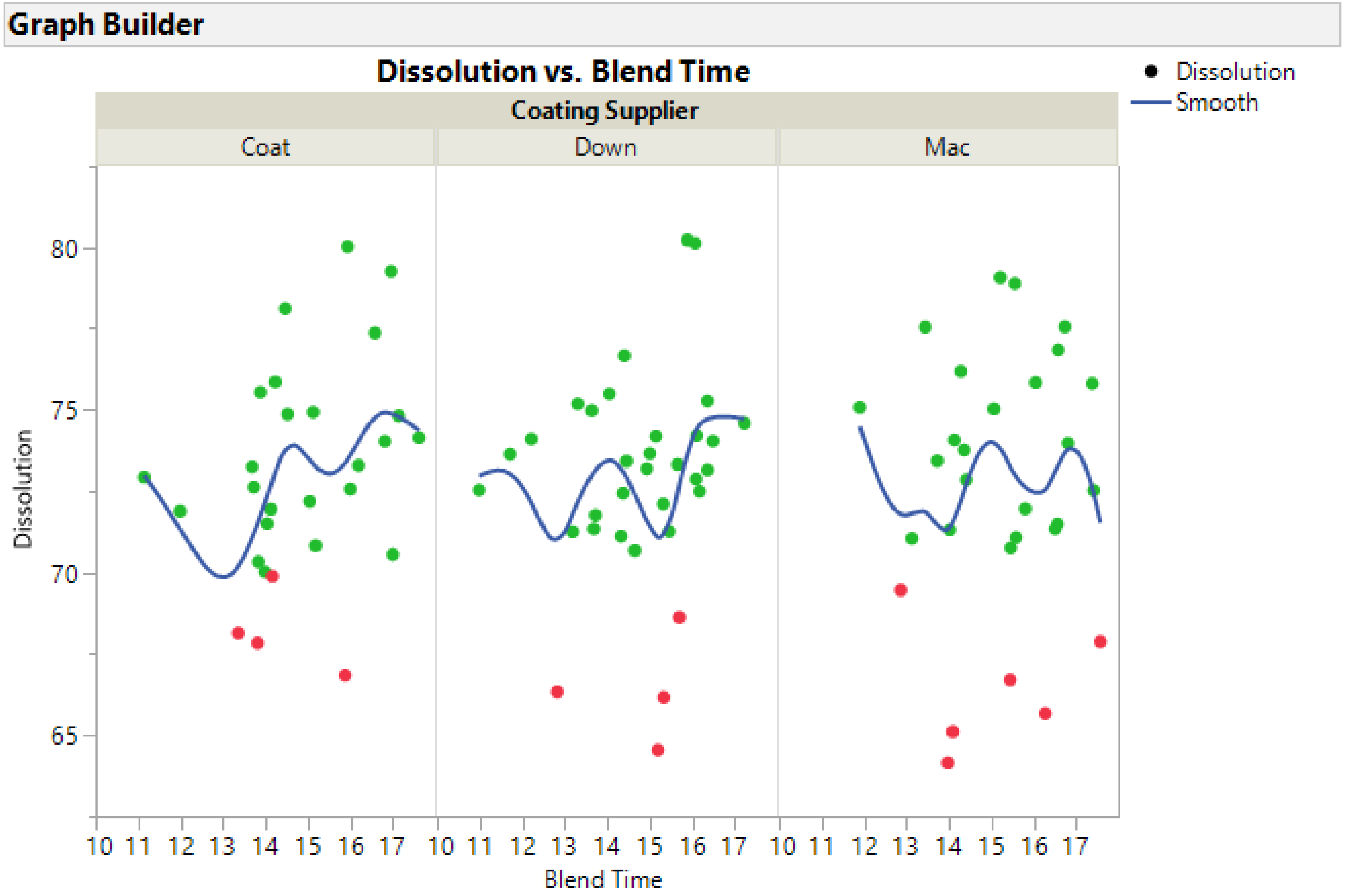
Mean	99.315625
Std Dev	4.8466816
Std Err Mean	0.5108851
Upper 95% Mean	100.33074
Lower 95% Mean	98.300507
N	90

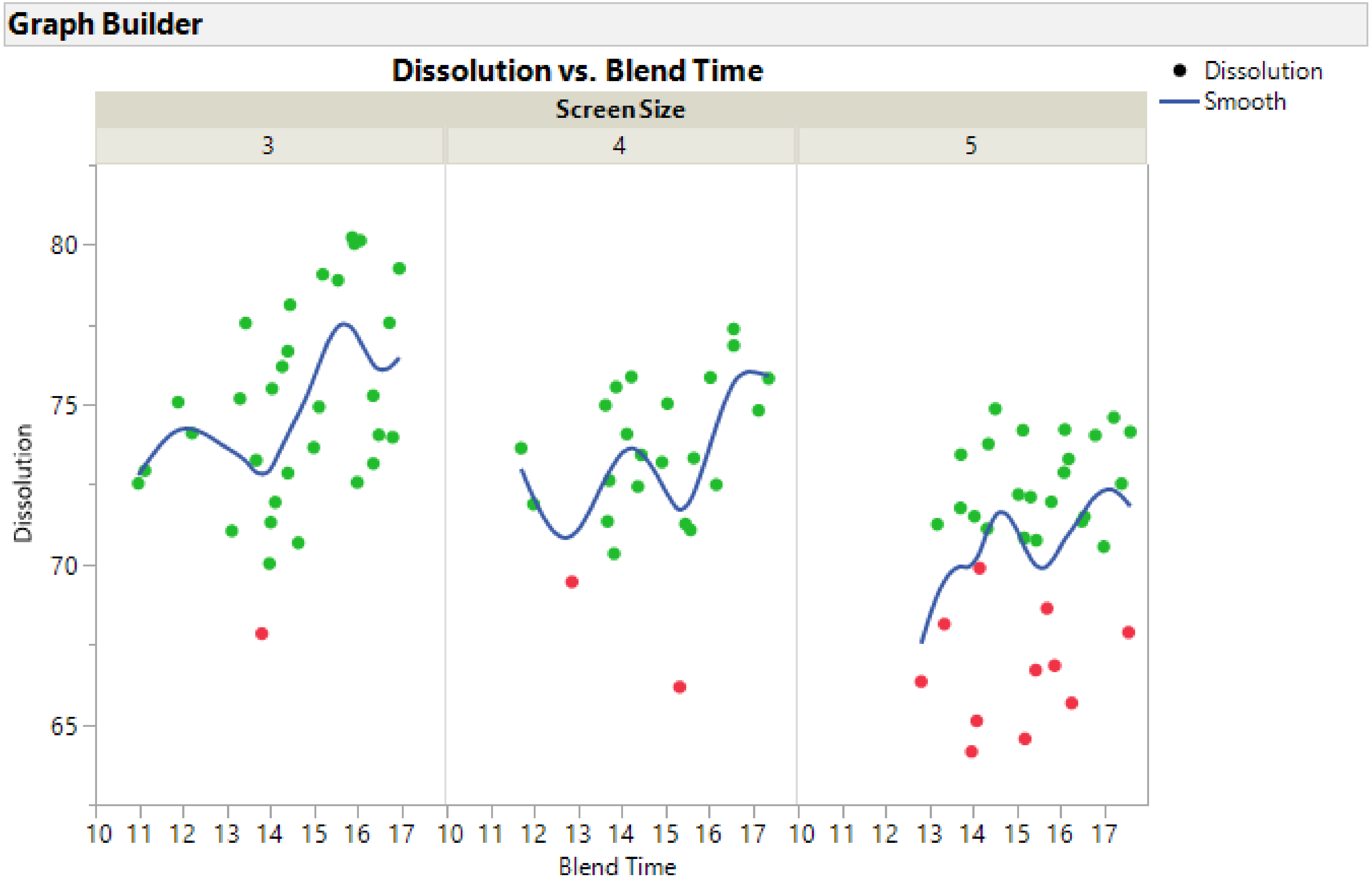

Scatterplot Matrix
Disso
Mill Time
Blend Time
Blend Speed
Force
Coating Viscosity
Inlet Temp
Exhaust Temp
Spray Rate
Atomizer Pressure


Parallel Plot
Lot Acceptance
Disso
API Particle Size
Mill Time
Screen Size
Mag. Stearate Supplier
Lactose Supplier
Sugar Supplier
Talc Supplier
Blend Time
Blend Speed
Compressor
Force
Coating Supplier
Coating Viscosity
Inlet Temp
Exhaust Temp
Spray Rate
Atomizer Pressure

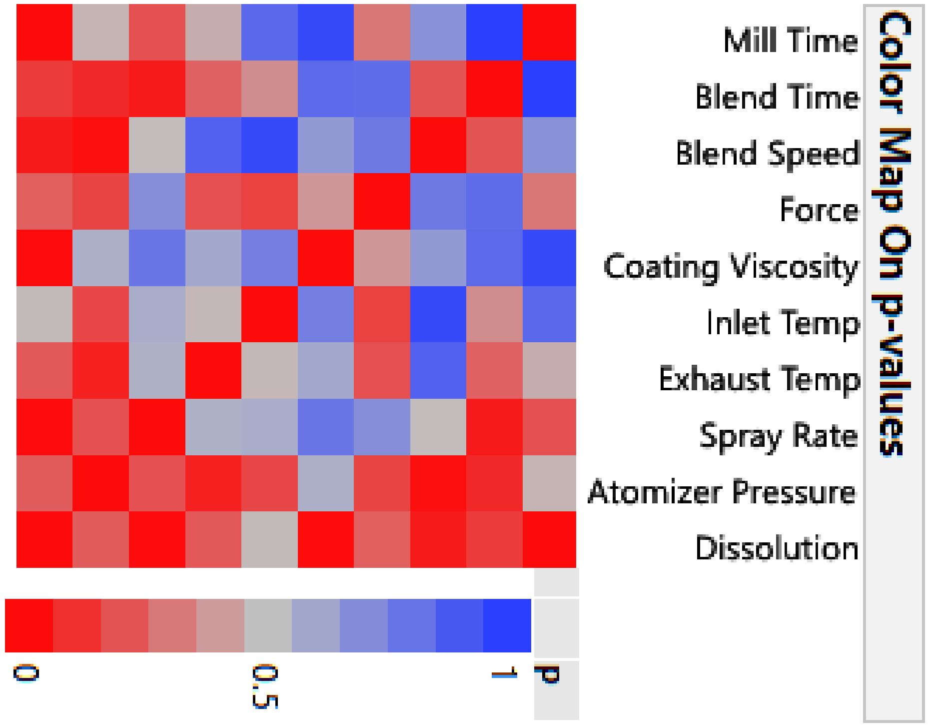

Parallel Plot
Lot Acceptance
Disso
API Particle Size
Mill Time
Screen Size
Mag. Stearate Supplier
Lactose Supplier
Sugar Supplier
Talc Supplier
Blend Time
Blend Speed
Compressor
Force
Coating Supplier
Coating Viscosity
Inlet Temp
Exhaust Temp
Spray Rate
Atomizer Pressure


Graph Builder
Dissolution vs. Blend Time
Dissolution
Smooth
Talc Supplier
Rough
Smooth
80
75
70
65
Dissolution
10
11
12
13
14
15
16
17
Blend Time


Graph Builder
Dissolution vs. Blend Time
Dissolution
Smooth
Coating Supplier
Coat
Down
Mac
80
75
70
65
Dissolution
10 11 12 13 14 15 16 17
Blend Time



Graph Builder
Dissolution vs. Blend Time
Dissolution
Smooth
Screen Size
3
4
5
80
75
70
65
Dissolution
10 11 12 13 14 15 16 17 10 11 12 13 14 15 16 17 10 11 12 13 14 15 16 17
Blend Time



Explore Outliers

Quantile Range Outliers

Outliers are values Q times the interquantile range past the lower and upper quantiles.

Tail Quantile 0.1

Q 3

Select columns and choose an action.

☐ Restrict search to integers

☐ Show only columns with outliers

Column	10% Quantile	90% Quantile	Low Threshold	High Threshold	Number of Outliers	Outliers (Count)
Mill Time	6.1	28	-59.6	93.7	0	
Blend Time	13.1348	16.8123	2.10227	27.8448	0	
Blend Speed	59.0087	61.2002	52.434	67.7748	0	
Force	24.5967	25.4662	21.9885	28.0744	0	
Coating Viscosity	93.9009	106.366	56.5066	143.76	0	
Inlet Temp	105.586	108.278	97.5089	116.356	0	
Exhaust Temp	68.6515	71.3634	60.5157	79.4991	0	
Spray Rate	392.541	407.083	348.914	450.71	0	
Atomizer Pressure	58.5073	61.2789	50.1927	69.5935	0	
Dissolution	67.844	77.532	38.78	106.596	0	

Correlations

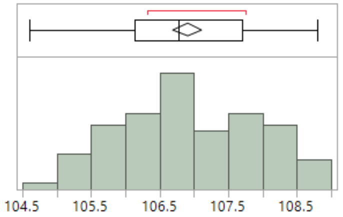
	Mill Time	Blend Time	Blend Speed	Force	Coating Viscosity	Inlet Temp	Exhaust Temp	Spray Rate	Atomizer Pressure	Dissolution
Mill Time	1.0000	0.0004	-0.0436	0.1116	0.0057	0.0217	0.0810	-0.1381	-0.0775	0.3638
Blend Time	0.0004	1.0000	0.1348	0.0241	0.0223	0.0977	-0.1257	0.2145	0.1841	0.1598
Blend Speed	-0.0436	0.1348	1.0000	-0.0301	0.0482	0.0059	-0.0184	-0.0745	0.2632	0.2143
Force	0.1116	0.0241	-0.0301	1.0000	0.0928	0.1535	0.1402	0.0421	0.1506	0.1271
Coating Viscosity	0.0057	0.0223	0.0482	0.0928	1.0000	-0.0331	0.0570	0.0287	-0.0615	0.3194
Inlet Temp	0.0217	0.0977	0.0059	0.1535	-0.0331	1.0000	0.0761	0.0603	0.1476	0.0755
Exhaust Temp	0.0810	-0.1257	-0.0184	0.1402	0.0570	0.0761	1.0000	-0.0628	0.1977	0.1327
Spray Rate	-0.1381	0.2145	-0.0745	0.0421	0.0287	0.0603	-0.0628	1.0000	0.1380	-0.3292
Atomizer Pressure	-0.0775	0.1841	0.2632	0.1506	-0.0615	0.1476	0.1977	0.1380	1.0000	0.1288
Dissolution	0.3638	0.1598	0.2143	0.1271	0.3194	0.0755	0.1327	-0.3292	0.1288	1.0000

The correlations are estimated by Row-wise method.

Inlet Temp	Exhaust Temp	Spray Rate	Atomizer Pressure
107.9	70.5	404.6	61.0
107.5	70.8	407.4	60.6
106.6	69.2	399.3	59.1
106.1	68.8	403.7	58.8
108.3	69.4	396.7	59.6
106.3	69.1	404.7	60.4
106.1	69.7	399.3	58.4
107.6	70.0	398.5	61.6
107.2	71.4	404.0	61.1
106.8	70.4	394.9	59.5
105.2	69.7	403.3	60.9
105.5	71.9	395.4	59.7
106.6	69.3	397.7	57.5
106.6	70.1	388.4	58.2
105.3	68.7	391.6	58.5
106.7	68.9	418.5	58.2
108.1	69.4	402.3	60.5
105.9	69.1	396.8	60.2
106.5	69.1	397.2	62.3
105.5	70.6	408.5	58.8
107.6	70.6	401.1	60.4
106.9	69.6	404.8	60.0
106.7	69.6	407.4	61.4
107.4	72.0	403.6	60.1
105.0	70.3	390.9	58.5
107.7	71.2	400.9	61.4

Distributions

Inlet Temp



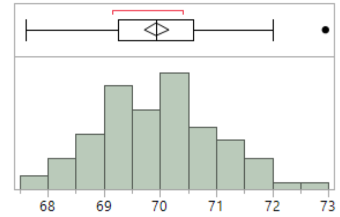
Quantiles

100.0%	maximum	108.79765542
99.5%		108.79765542
97.5%		108.7196325
90.0%		108.27844338
75.0%	quartile	107.69690533
50.0%	median	106.77968341
25.0%	quartile	106.13953111
10.0%		105.58605819
2.5%		105.030777
0.5%		104.60908689

Summary Statistics

Mean	106.9035
Std Dev	1.0021833
Std Err Mean	0.1056394
Upper 95% Mean	107.1134
Lower 95% Mean	106.6936
N	90

Exhaust Temp



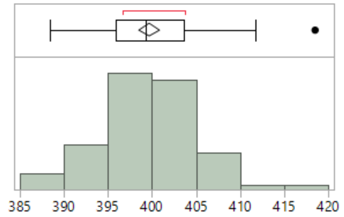
Quantiles

100.0%	maximum	72.948836456
99.5%		72.948836456
97.5%		72.008546738
90.0%		71.36338255
75.0%	quartile	70.590886383
50.0%	median	69.937330759
25.0%	quartile	69.258273575
10.0%		68.651473499
2.5%		67.88661075
0.5%		67.613981065
0.0%	minimum	67.613981065

Summary Statistics

Mean	69.937392
Std Dev	1.0361819
Std Err Mean	0.1092232
Upper 95% Mean	70.154416
Lower 95% Mean	69.720368
N	90

Spray Rate



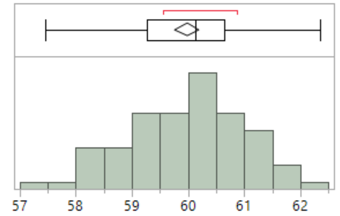
Quantiles

100.0%	maximum	418.49400196
99.5%		418.49400196
97.5%		410.88281454
90.0%		407.08319782
75.0%	quartile	403.66964498
50.0%	median	399.32559204
25.0%	quartile	395.85948298
10.0%		392.54095393
2.5%		388.78379884
0.5%		388.40679041
0.0%	minimum	388.40679041

Summary Statistics

Mean	399.70046
Std Dev	5.4666398
Std Err Mean	0.5762344
Upper 95% Mean	400.84543
Lower 95% Mean	398.5555
N	90

Atomizer Pressure

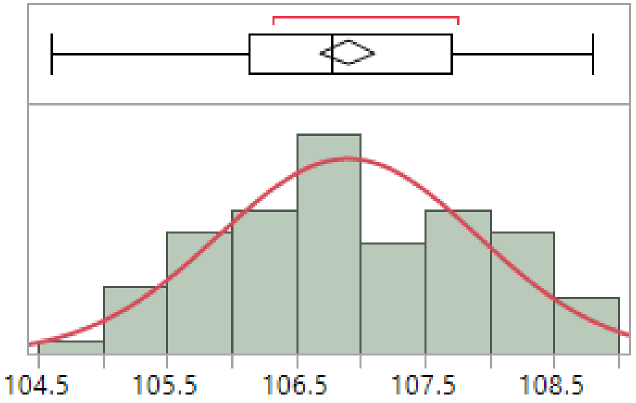


Quantiles

100.0%	maximum	62.349078059
99.5%		62.349078059
97.5%		61.736589899
90.0%		61.278894542
75.0%	quartile	60.642367179
50.0%	median	60.13028396
25.0%	quartile	59.256497573
10.0%		58.507347823
2.5%		57.7781621
0.5%		57.459007957
0.0%	minimum	57.459007957

Summary Statistics

Mean	59.980797
Std Dev	1.0137139
Std Err Mean	0.1068548
Upper 95% Mean	60.193115
Lower 95% Mean	59.768479
N	90



Normal(106.903,1.00218)

Compare Distributions

Show	Distribution	Number of Parameters	-2*LogLikelihood	AICc
☐	SHASH	4	248.053662	256.52425
☐	Normal 2 Mixture	5	247.568499	258.282785
☐	Gamma	2	254.794139	258.93207
☐	LogNormal	2	254.79609	258.934021
☑	Normal	2	254.801502	258.939433
☐	Johnson Sl	3	254.794544	261.073614
☐	GLog	3	254.795904	261.074974
☐	Johnson Su	4	254.795964	263.266553
☐	Normal 3 Mixture	8	247.982016	265.759794
☐	Weibull	2	262.445187	266.583118
☐	Extreme Value	2	262.445187	266.583118
☐	Exponential	1	1020.94678	1022.99223

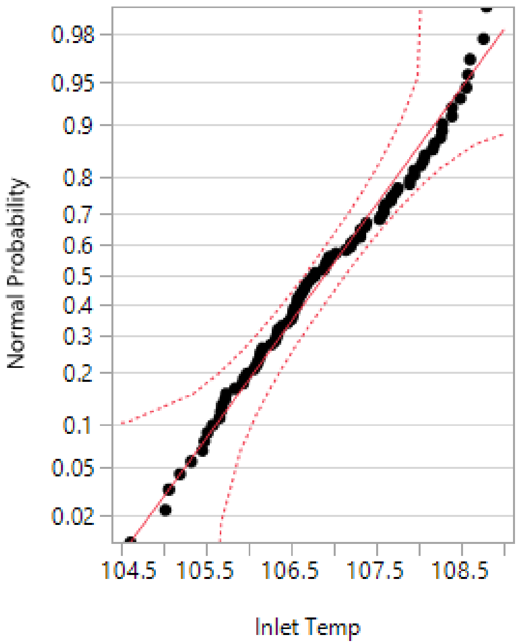
Fitted Normal

Parameter Estimates

Type	Parameter	Estimate	Lower 95%	Upper 95%
Location	μ	106.9035	106.6936	107.1134
Dispersion	σ	1.0021833	0.8741169	1.1745646

Measure	
-2*LogLikelihood	254.8015
AICc	258.93943
BIC	263.80112

Diagnostic Plot

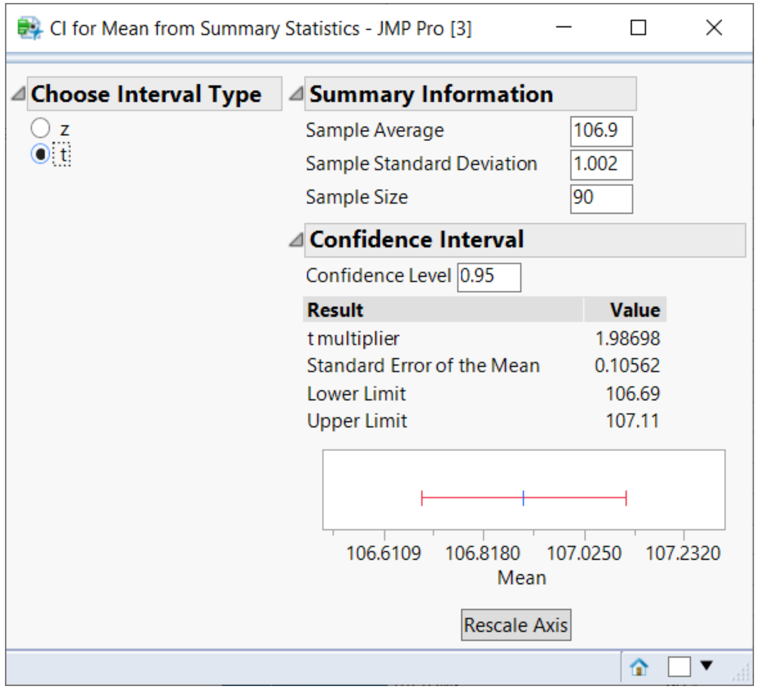


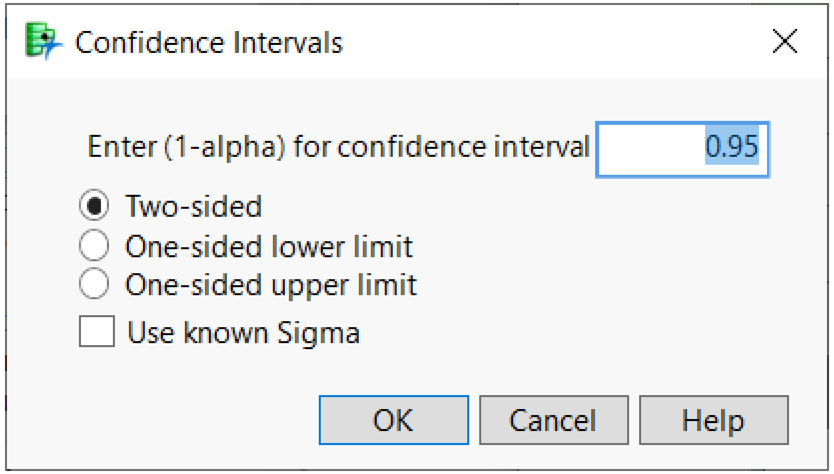
Goodness-of-Fit Test

Shapiro-Wilk W Test

W	Prob<W
0.980334	0.1906

Note: Ho = The data is from the Normal distribution. Small p-values reject Ho.





Confidence Intervals

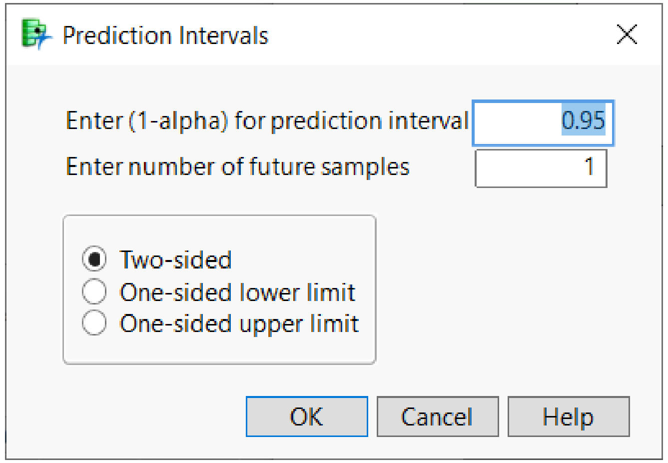
Parameter	Estimate	Lower CI	Upper CI	1-Alpha
Mean	106.9035	106.6254	107.1816	0.990
Std Dev	1.002183	0.838607	1.237307	0.990

Confidence Intervals

Parameter	Estimate	Lower CI	Upper CI	1-Alpha
Mean	106.9035	106.6936	107.1134	0.950
Std Dev	1.002183	0.874117	1.174565	0.950

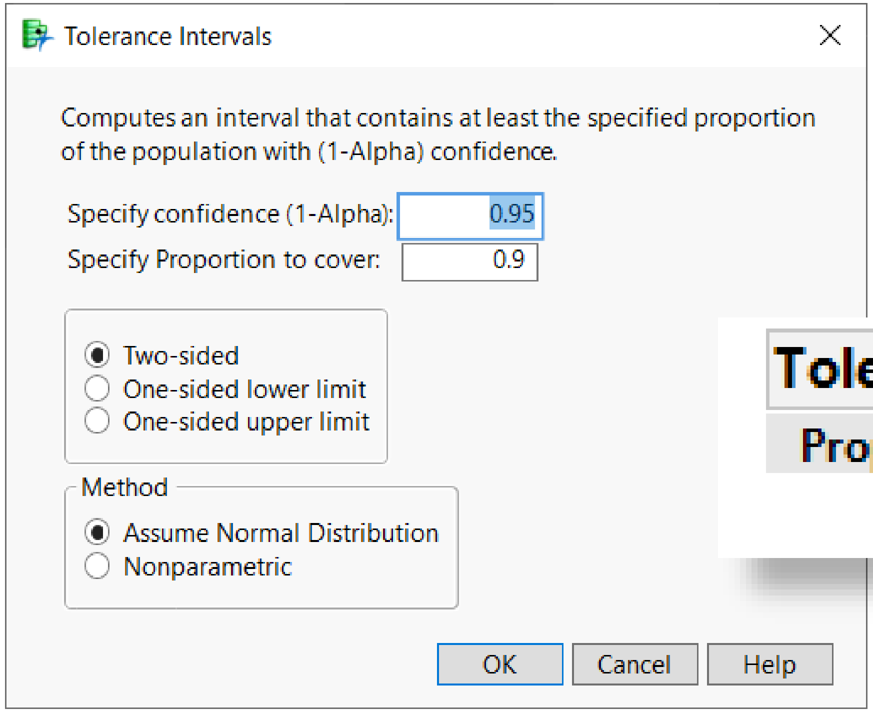
Confidence Intervals

Parameter	Estimate	Lower CI	Upper CI	1-Alpha
Mean	106.9035	106.7279	107.0791	0.900
Std Dev	1.002183	0.893286	1.14444	0.900



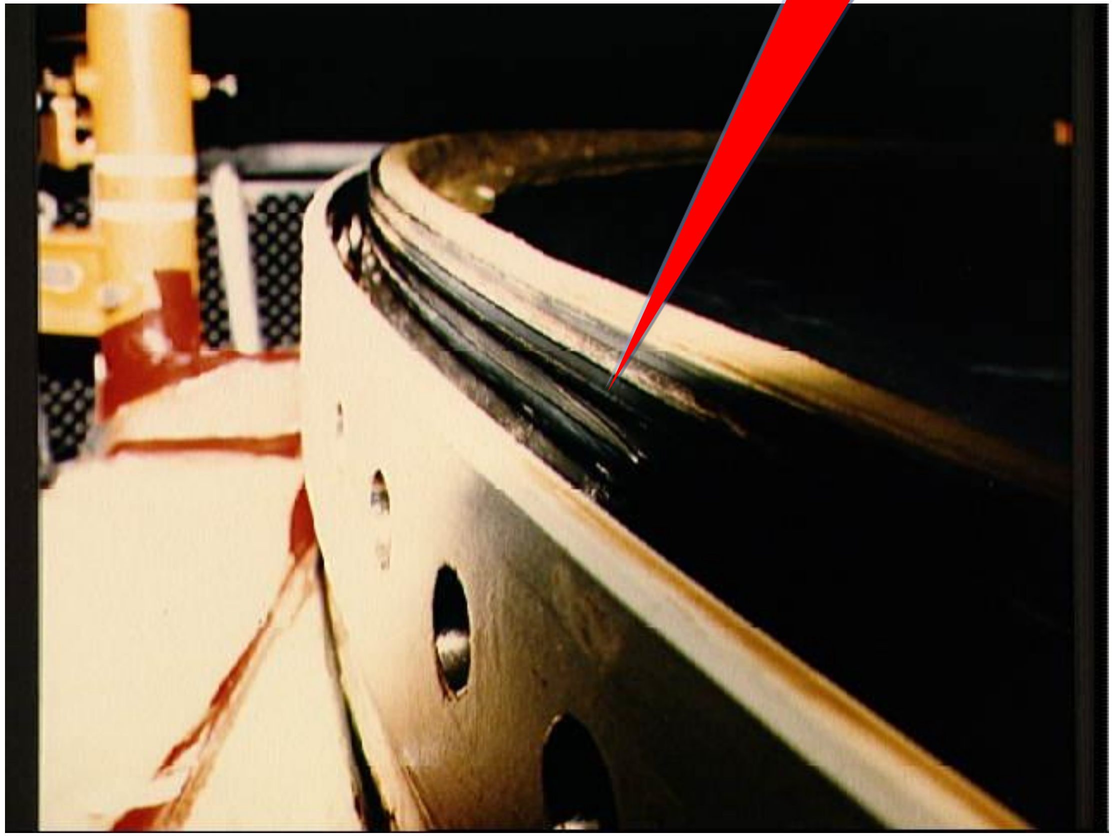
Prediction Interval				
Parameter	Future N	Lower PI	Upper PI	1-Alpha
Individual	1	104.9012	108.9058	0.950
Mean	1	104.9012	108.9058	0.950
Std Dev	1	.	.	0.950

Prediction Interval				
Parameter	Future N	Lower PI	Upper PI	1-Alpha
Individual	10	104.0025	109.8045	0.950
Mean	10	106.2397	107.5673	0.950
Std Dev	10	0.542555	1.506759	0.950



Tolerance Intervals			
Proportion	Lower TI	Upper TI	1-Alpha
0.900	105.0095	108.7975	0.950

The Challenger



Kenett, R. and Thyregod, P. (2006) Aspects of statistical consulting not taught by academia, *Statistica Neerlandica*, special issue on Industrial Statistics, 30, 3, pp. 396-412.

The Challenger

The US space shuttle Challenger was cheduled to take-off on January 28th, 1986, with seven crew members. Engineers from Morton Thiokol, manufacturers of the rocket motors, had been worried about problems with the O-ring seals. They feared that low temperatures greatly and adversely affected the ability of O-rings to create a seal on solid rocket booster joints.

On the night before the flight, the temperature predicted at launch time was 3° C, and the engineers expressed their concerns over the effect of the unseasonable cold weather on the O-rings and suggested to abort the flight.

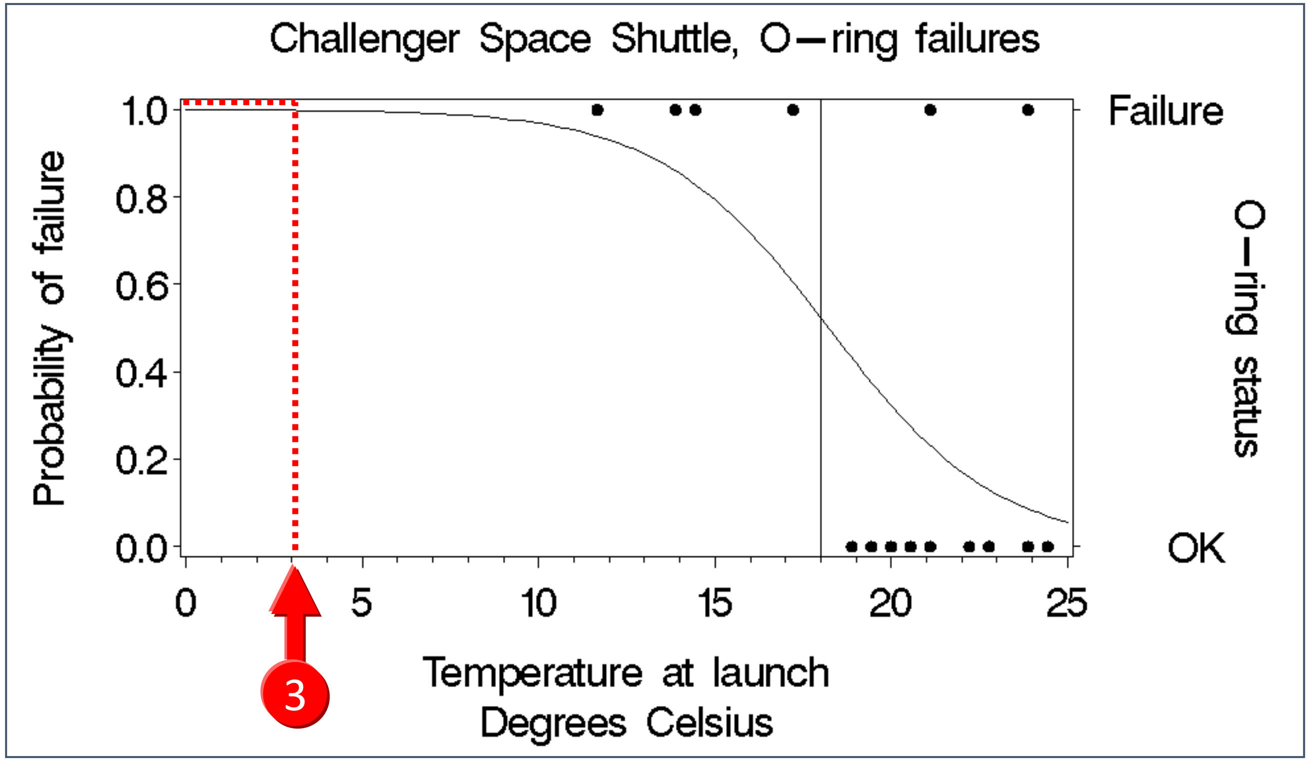
The Challenger

A telephone conference was held between NASA engineers and managers and Thiokol engineers and managers.

With short notice, the Thiokol engineers presented their case via 13 telefaxed charts and their commentary and argument.

However, they failed to convince the managers that temperature was a factor in O-ring performance or damage, and it was decided to **go ahead** with the launching.

Probability of O-ring Failure



Tablet Production - JMP Pro

File Edit Tables Rows Cols DOE Analyze Graph Tools Add-Ins View Window Help

Tablet Production

Reference Based on tablet produc

Control Chart and Distribution
Distribution
Multivariate
Oneway
Parallel Plot
Partition
Fit Model
Fit Model with Interactions
Fit Model with...action Profiles
Generalized Regression
Generalized R...educed Model
Fit Y by X of L... by Inlet Temp

Columns (21/1)

API Lot No
API Particle Size
Mill Time
Screen Size
Mag. Stearate Supplier
Lactose Supplier
Sugar Supplier
Talc Supplier
Blend Time
Blend Speed
Compressor
Force
Coating Supplier
Coating Viscosity

Rows

All rows	90
Selected	0
Excluded	0
Hidden	0
Labeled	0

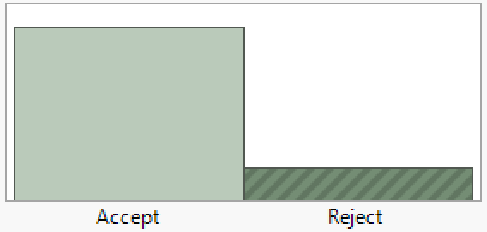
20/1 Cols

	API Particle Size	Mill Time	Screen Size	Mag. Stearate Supplier	Lactose Supplier	Sugar Supplier	Talc Supplier	Blend Time	Blend Speed	Compressor	Force	Co
1	Small	27	4	Smith Ind	James Ind	Sour	Rough	16.0	59.9	Compress2	25.5	Ma
2	Small	11	5	Jones Inc	James Ind	Sour	Smooth	14.4	59.8	Compress2	24.9	Ma
3	Small	20	4	Jones Inc	Bond Inc	Sour	Rough	14.5	60.8	Compress2	25.5	Do
4	Small	13	3	Smith Ind	Bond Inc	Sweet	Smooth	14.4	59.4	Compress1	24.8	Ma
5	Small	13	5	Smith Ind	James Ind	Sweet	Smooth	16.1	59.9	Compress2	25.3	Do
6	Small	19	4	Smith Ind	Bond Inc	Sweet	Rough	12.9	59.4	Compress2	24.6	Ma
7	Small	10	4	Jones Inc	Bond Inc	Sweet	Smooth	13.6	59.8	Compress2	25.0	Do
8	Small	24	4	Jones Inc	James Ind	Sour	Rough	15.1	61.1	Compress2	24.9	Ma
9	Small				ames Ind	Sour				ess1	25.3	Ma
10	Small				ond Inc	Sweet				ess2	25.4	Co
11	Small				ond Inc	Sweet				ess1	24.5	Do
12	Small				ames Ind	Sour				ess1	24.9	Do
13	Small				ames Ind	Sour				ess2	25.0	Ma
14	Small				ond Inc	Sour				ess1	24.6	Do
15	Small	22	5	Jones Inc	James Ind	Sweet				ess2	24.9	Co
16	Small	7	3	Jones Inc	James Ind	Sour				ess2	25.5	Co
17	Small	6	3	Jones Inc	James Ind	Sweet					25.1	Co
18	Small	30	3	Jones Inc	Bond Inc	Sweet	Smooth	16.4	61.2	Compress1	24.7	Do
19	Small	29	3	Smith Ind	Bond Inc	Sour	Smooth	12.2	59.8	Compress1	25.2	Do
20	Small	7	5	Jones Inc	Bond Inc	Sour	Smooth	14.0	60.0	Compress1	25.1	Ma
21	Small	25	5	Jones Inc	James Ind	Sour	Smooth	17.4	59.8	Compress1	25.8	Ma
22	Small	13	5	Jones Inc	Bond Inc	Sour	Rough	15.7	58.7	Compress2	25.0	Do
23	Small	18	4	Jones Inc	James Ind	Sweet	Rough	17.4	61.2	Compress2	24.9	Ma
24	Small	24	3	Smith Ind	Bond Inc	Sweet	Smooth	15.1	57.9	Compress2	25.0	Co
25	Small	13	5	Smith Ind	Bond Inc	Sour	Rough	15.2	61.5	Compress2	24.7	Co
26	Small	28	3	Jones Inc	Bond Inc	Sour	Smooth	15.9	61.1	Compress2	25.1	Co
27	Small	19	5	Smith Ind	James Ind	Sweet	Rough	15.8	60.2	Compress2	24.7	Ma
28	Small	9	5	Jones Inc	James Ind	Sour	Rough	16.3	60.5	Compress2	24.4	Ma

Tablet Production.jmp

Distributions

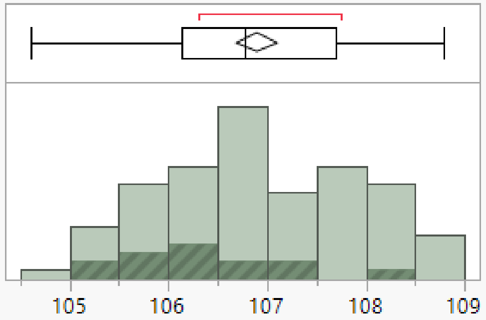
Lot Acceptance



Frequencies

Level	Count	Prob
Accept	76	0.84444
Reject	14	0.15556
Total	90	1.00000
N Missing	0	
2 Levels		

Inlet Temp



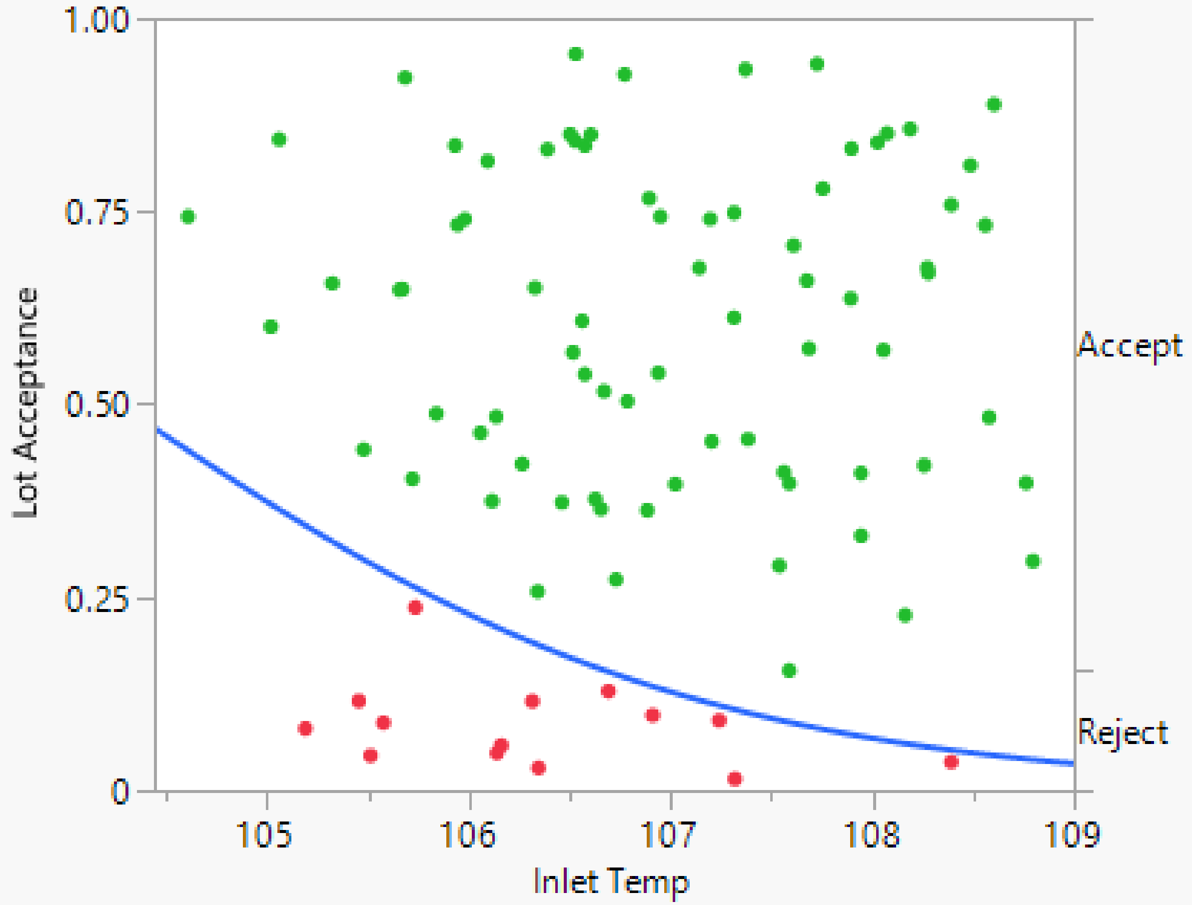
Quantiles

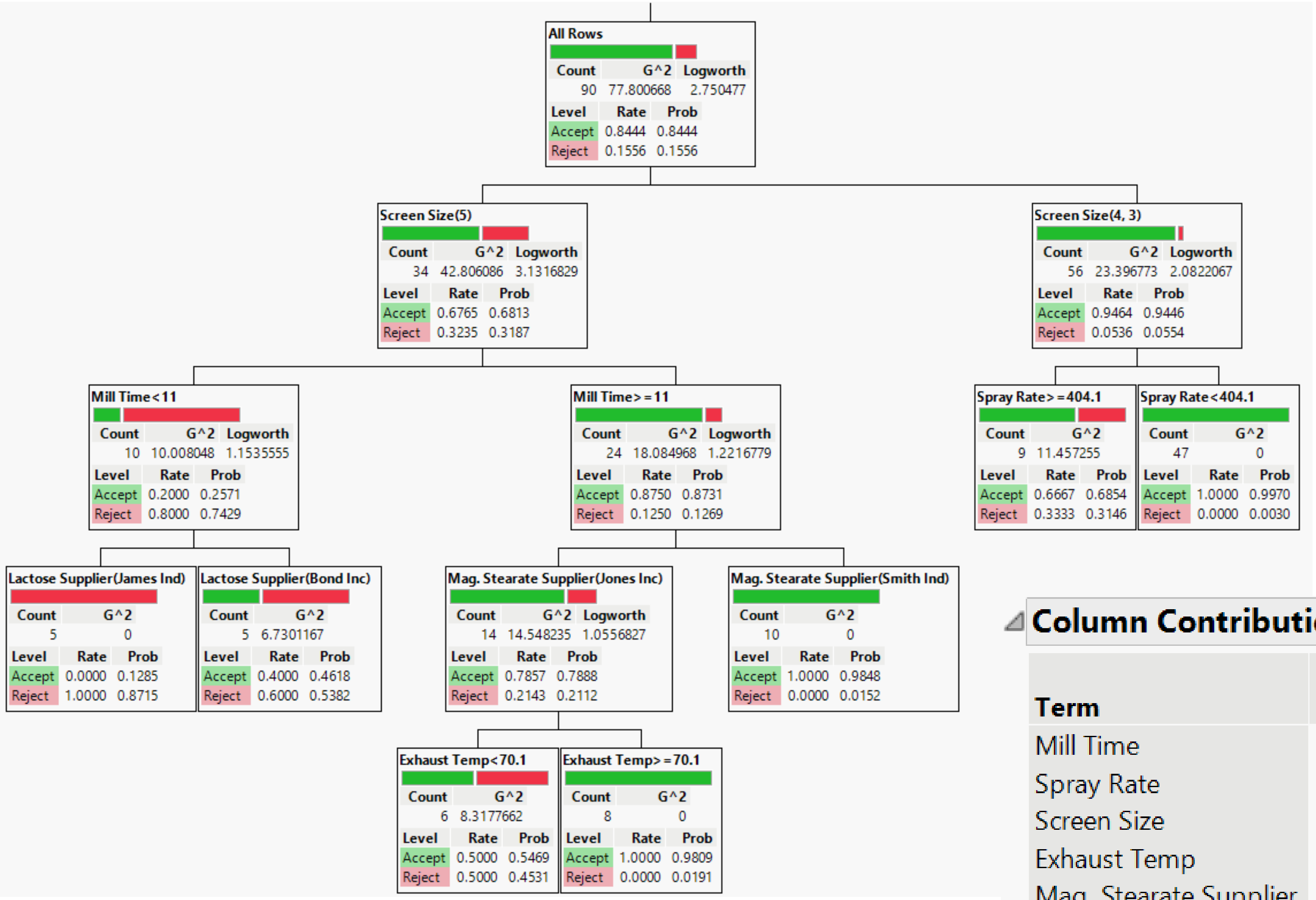
100.0%	maximum	108.79765542
99.5%		108.79765542
97.5%		108.7196325
90.0%		108.27844338
75.0%	quartile	107.69690533
50.0%	median	106.77968341
25.0%	quartile	106.13953111
10.0%		105.58605819
2.5%		105.030777
0.5%		104.60908689
0.0%	minimum	104.60908689

Summary Statistics

Mean	106.9035
Std Dev	1.0021833
Std Err Mean	0.1056394
Upper 95% Mean	107.1134
Lower 95% Mean	106.6936
N	90
N Missing	0

Logistic Fit of Lot Acceptance By Inlet Temp





Fit Details

Measure	Training	Definition
Entropy RSquare	0.6282	1-Loglike(model)/Loglike(0)
Generalized RSquare	0.7241	(1-(L(0)/L(model))^(2/n))/(1-L(0)^(2/n))
Mean -Log p	0.1607	∑ -Log(ρ[j])/n
RASE	0.2315	√ ∑(y[j]-ρ[j])²/n
Mean Abs Dev	0.1166	∑ |y[j]-ρ[j]|/n
Misclassification Rate	0.0889	∑ (ρ[j]≠ρMax)/n
N	90	n

Confusion Matrix

Training

Actual Lot Acceptance	Predicted Count Accept	Reject
Accept	74	2
Reject	6	8

Actual Lot Acceptance	Predicted Rate Accept	Reject
Accept	0.974	0.026
Reject	0.429	0.571

Column Contributions

Term	Number of Splits	G^2	Portion
Mill Time	1	14.7130695	0.2868
Spray Rate	1	11.9395178	0.2328
Screen Size	1	11.5978092	0.2261
Exhaust Temp	1	6.23046933	0.1215
Mag. Stearate Supplier	1	3.53673224	0.0689
Lactose Supplier	1	3.2779318	0.0639