Exploratory Data Analysis

Maneesh Agrawala

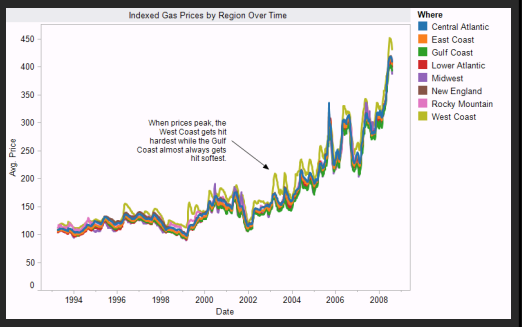
CS 448B: Visualization
Fall 2020

1

A2: Exploratory Data Analysis

Use Tableau to formulate & answer questions

First steps
- Step 1: Pick domain & data
- Step 2: Pose questions
- Step 3: Profile data
- Iterate as needed



Create visualizations
- Interact with data
- Refine questions

Author a report
- Screenshots of most insightful views (10+)
- Include titles and captions for each view

Due before class on Oct 6, 2020

2

Exploratory Data Analysis

3

The Rise of Statistics (1900-1950s)

Rise of **formal methods** in statistics and social science – Fisher, Pearson, …

Little innovation in graphical methods

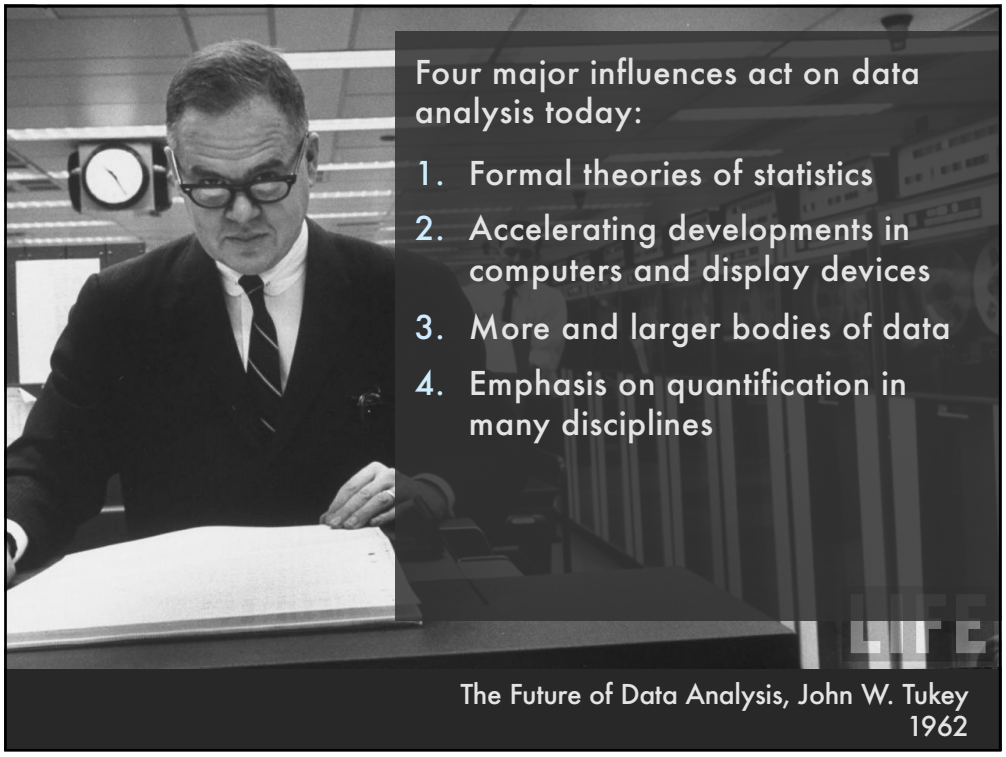
A period of **application** and **popularization**

Graphical methods enter textbooks, curricula, and **mainstream use**

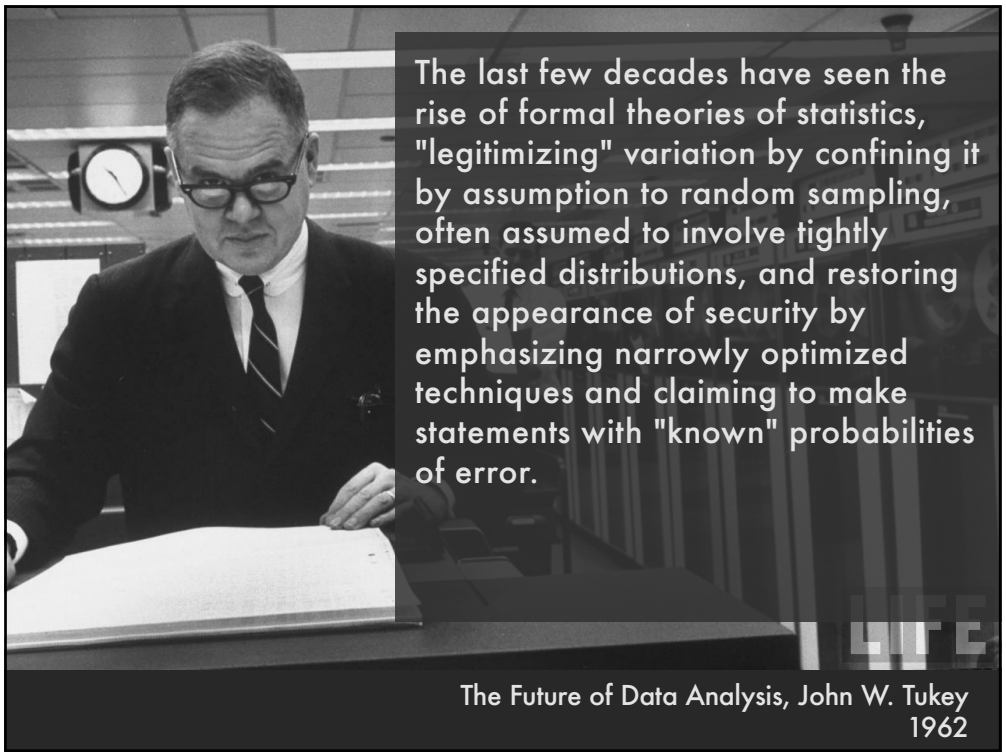
4


LIFE

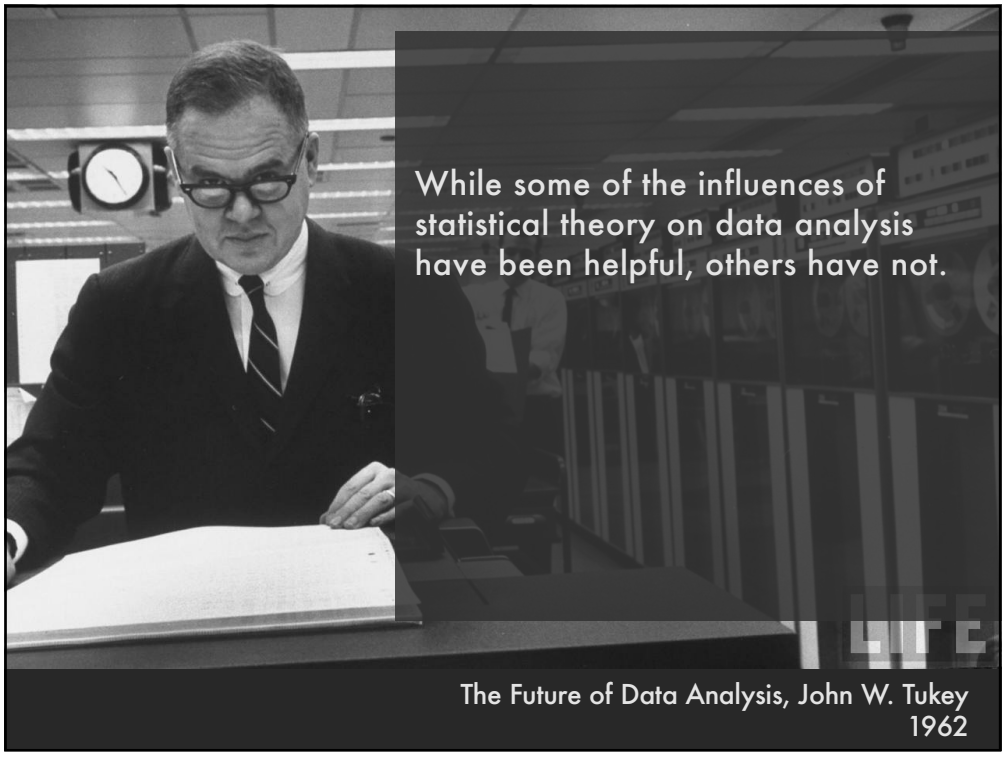
5


Four major influences act on data analysis today:
1. Formal theories of statistics
2. Accelerating developments in computers and display devices
3. More and larger bodies of data
4. Emphasis on quantification in many disciplines
LIFE
The Future of Data Analysis, John W. Tukey
1962

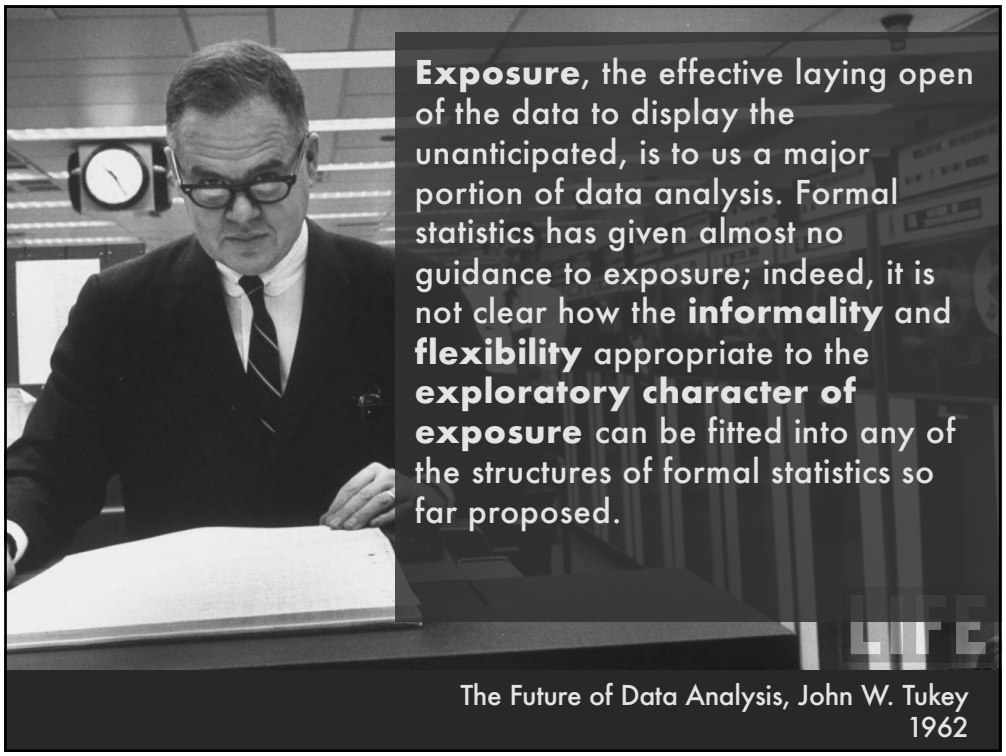
6


The last few decades have seen the rise of formal theories of statistics, "legitimizing" variation by confining it by assumption to random sampling, often assumed to involve tightly specified distributions, and restoring the appearance of security by emphasizing narrowly optimized techniques and claiming to make statements with "known" probabilities of error.
The Future of Data Analysis, John W. Tukey
1962

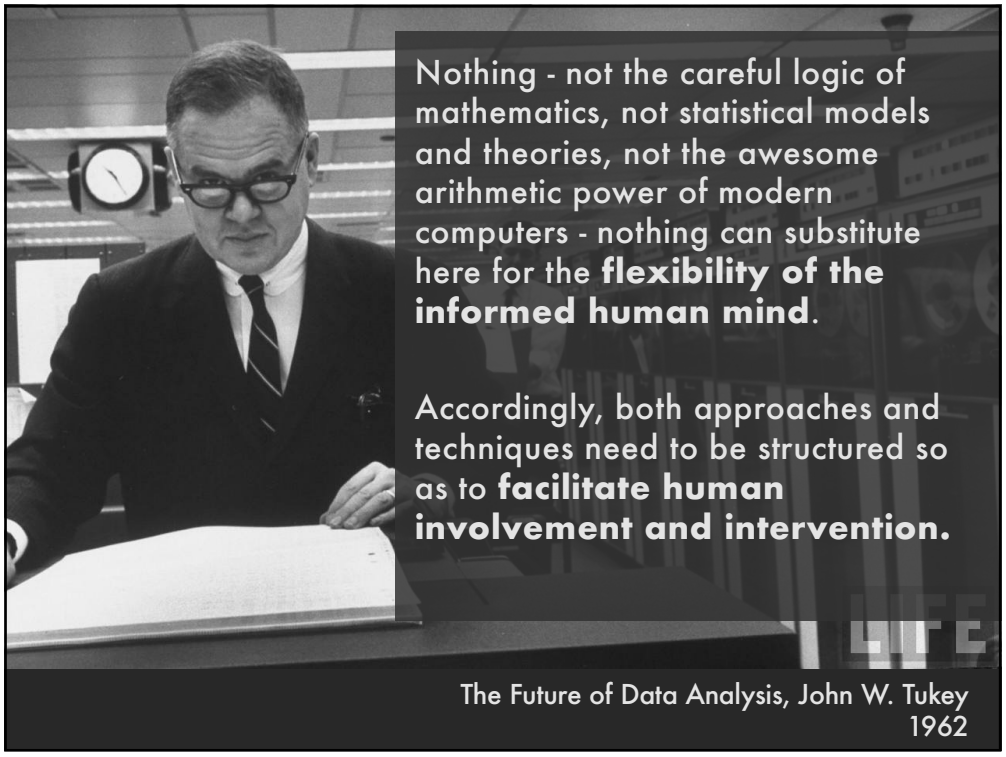
7


While some of the influences of statistical theory on data analysis have been helpful, others have not.
The Future of Data Analysis, John W. Tukey
1962

8


Exposure, the effective laying open of the data to display the unanticipated, is to us a major portion of data analysis. Formal statistics has given almost no guidance to exposure; indeed, it is not clear how the **informality** and **flexibility** appropriate to the **exploratory character of exposure** can be fitted into any of the structures of formal statistics so far proposed.
The Future of Data Analysis, John W. Tukey
1962

9


Nothing - not the careful logic of mathematics, not statistical models and theories, not the awesome arithmetic power of modern computers - nothing can substitute here for the **flexibility of the informed human mind**.
Accordingly, both approaches and techniques need to be structured so as to **facilitate human involvement and intervention.**
The Future of Data Analysis, John W. Tukey
1962

10

Topics

Data Wrangling
Effectiveness of antibiotics
Intro to Tableau

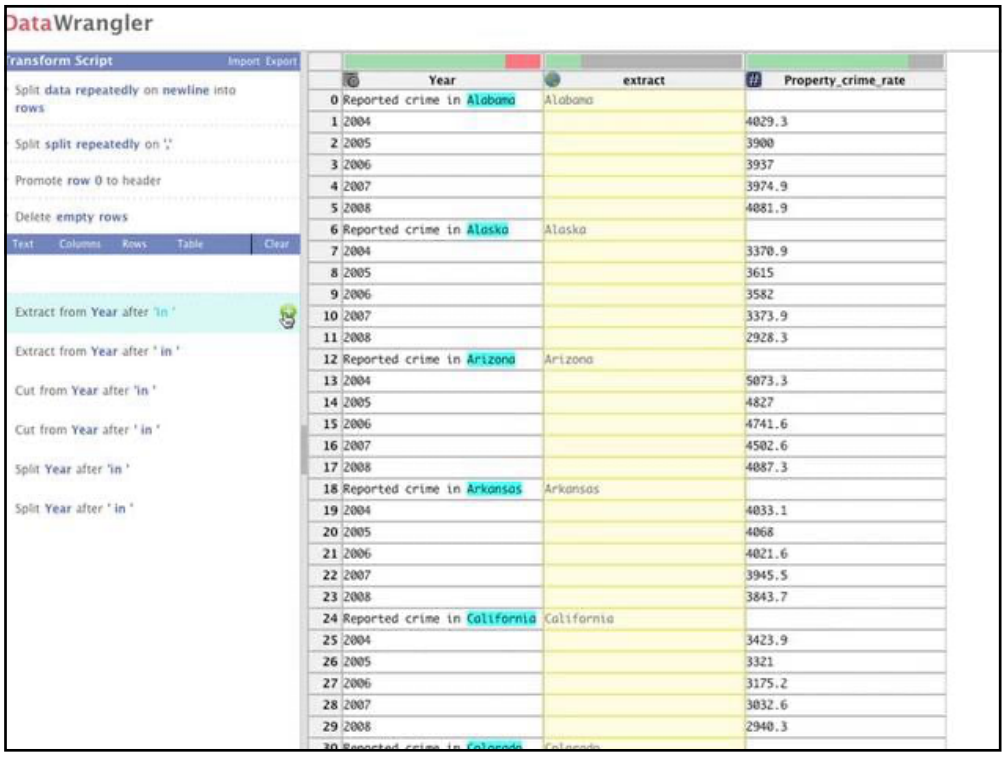
13

Data Wrangling

14

```
Bureau of Justice Statistics - Data Online
http://bjs.ojp.usdoj.gov/

Reported crime in Alabama

Year    Population      Property crime rate     Burglary rate   Larceny-theft rate      Motor vehicle theft rate
2004    4525375 4029.3  987     2732.4  309.9
2005    4548327 3900    955.8   2656    289
2006    4599030 3937    968.9   2645.1  322.9
2007    4627851 3974.9  980.2   2687    307.7
2008    4661900 4081.9  1080.7  2712.6  288.6

Reported crime in Alaska

Year    Population      Property crime rate     Burglary rate   Larceny-theft rate      Motor vehicle theft rate
2004    657755  3370.9  573.6   2456.7  340.6
2005    663253  3615    622.8   2601    391
2006    670053  3582    615.2   2588.5  378.3
2007    683478  3373.9  538.9   2480    355.1
2008    686293  2928.3  470.9   2219.9  237.5

Reported crime in Arizona

Year    Population      Property crime rate     Burglary rate   Larceny-theft rate      Motor vehicle theft rate
2004    5739879 5073.3  991     3118.7  963.5
2005    5953007 4827    946.2   2958    922
2006    6166318 4741.6  953     2874.1  914.4
2007    6338755 4502.6  935.4   2780.5  786.7
2008    6500180 4087.3  894.2   2605.3  587.8

Reported crime in Arkansas

Year    Population      Property crime rate     Burglary rate   Larceny-theft rate      Motor vehicle theft rate
2004    2750000 4033.1  1096.4  2699.7  237
2005    2775708 4068    1085.1  2720    262
2006    2810872 4021.6  1154.4  2596.7  270.4
2007    2834797 3945.5  1124.4  2574.6  246.5
2008    2855390 3843.7  1182.7  2433.4  227.6

Reported crime in California

Year    Population      Property crime rate     Burglary rate   Larceny-theft rate      Motor vehicle theft rate
2004    35842038        3423.9  686.1   2033.1  704.8
2005    36154147        3321    692.9   1915    712
2006    36457549        3175.2  676.9   1831.5  666.8
2007    36553215        3032.6  648.4   1784.1  600.2
2008    36756666        2940.3  646.8   1769.8  523.8

Reported crime in Colorado

Year    Population      Property crime rate     Burglary rate   Larceny-theft rate      Motor vehicle theft rate
2004    4601821 3918.5  717.3   2679.5  521.6
```

15



16

Data "Wrangling"

One often needs to manipulate data prior to analysis. Tasks include reformatting, cleaning, quality assessment, and integration

Some approaches:

- Writing custom scripts
- Manual manipulation in spreadsheets
- Trifacta Wrangler: http://trifacta.com/products/wrangler/
- Open Refine: http://openrefine.org

17

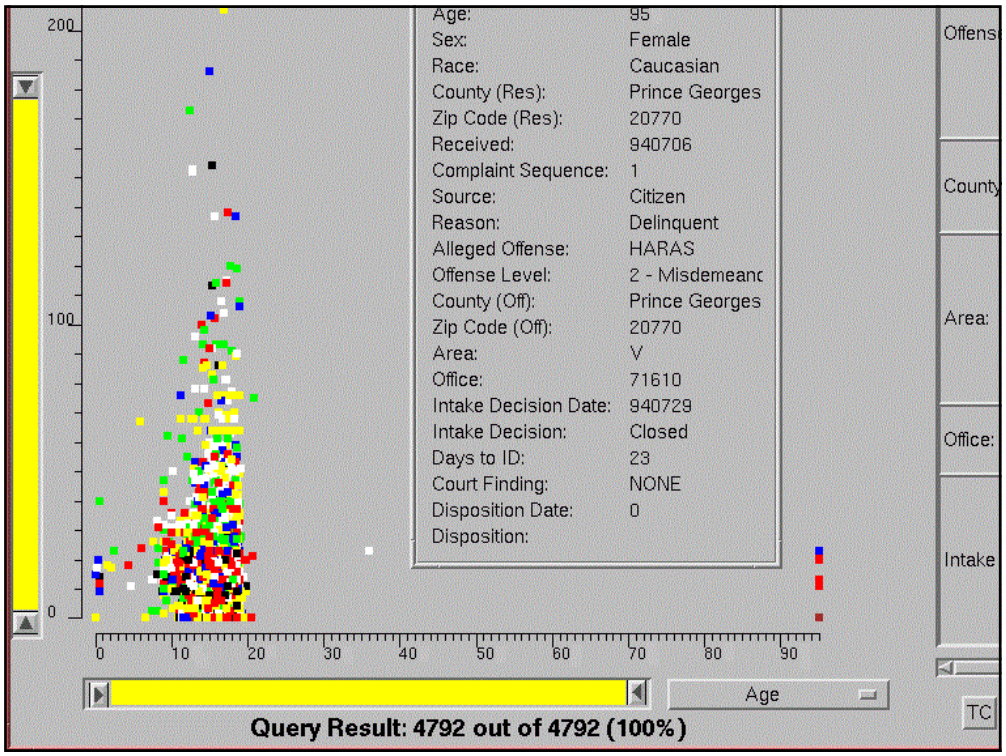
How to gauge the quality of a visualization?

"The first sign that a visualization is good is that it shows you a problem in your data...

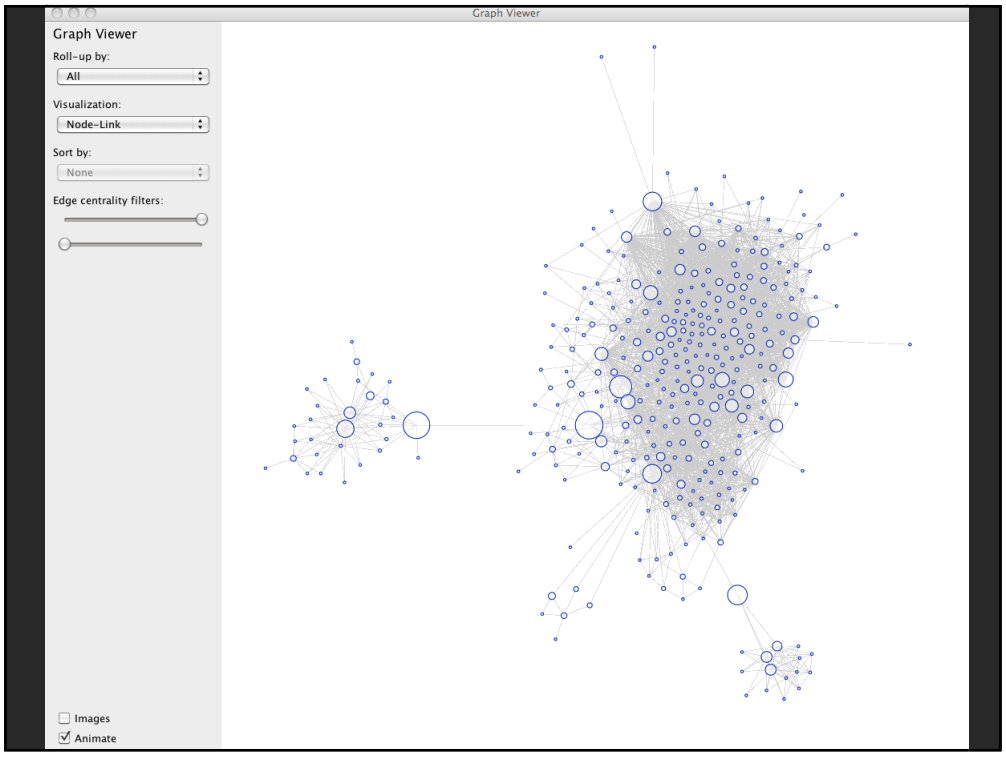
...every successful visualization that I've been involved with has had this stage where you realize, "Oh my God, this data is not what I thought it would be!" So already, you've discovered something."

- Martin Wattenberg

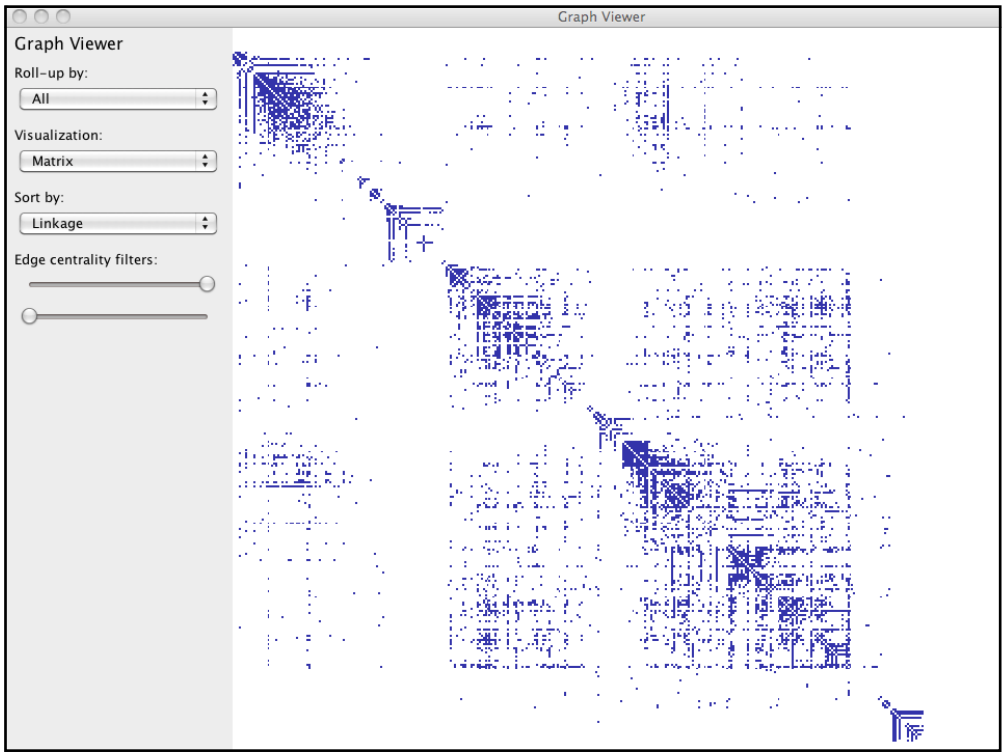
18


Age: 95
Sex: Female
Race: Caucasian
County (Res): Prince Georges
Zip Code (Res): 20770
Received: 940706
Complaint Sequence: 1
Source: Citizen
Reason: Delinquent
Alleged Offense: HARAS
Offense Level: 2 - Misdemeanc
County (Off): Prince Georges
Zip Code (Off): 20770
Area: V
Office: 71610
Intake Decision Date: 940729
Intake Decision: Closed
Days to ID: 23
Court Finding: NONE
Disposition Date: 0
Disposition:
200
100
0
0 10 20 30 40 50 60 70 80 90
Age
Offens
County
Area:
Office:
Intake
TC
Query Result: 4792 out of 4792 (100%)

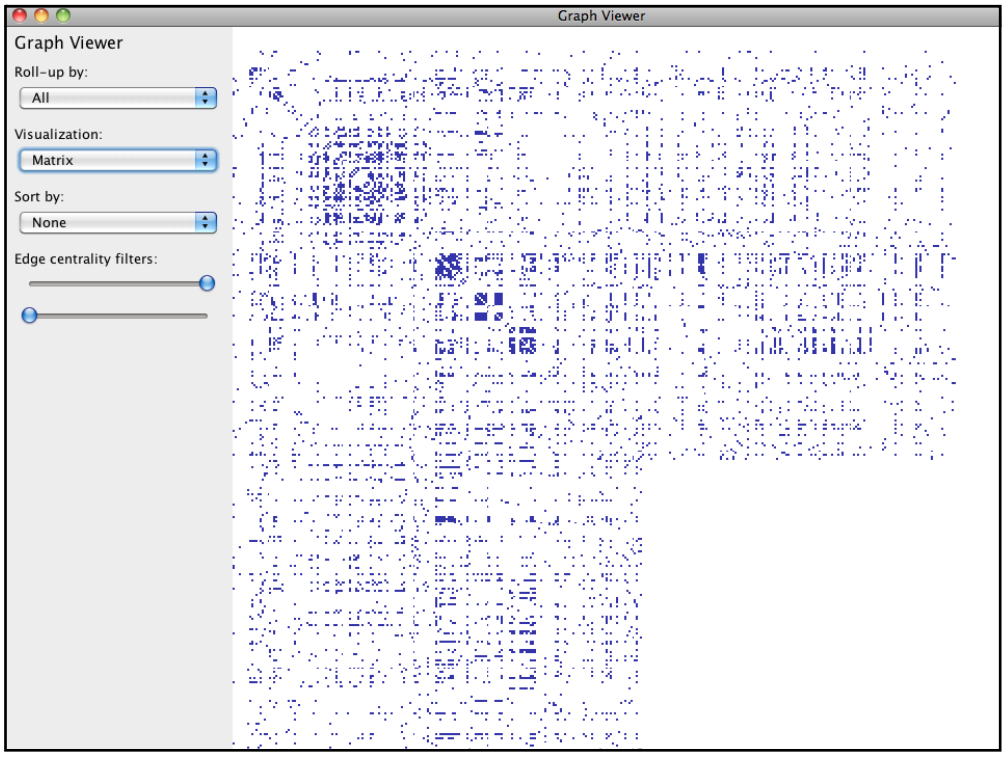
19


Graph Viewer
Graph Viewer
Roll-up by:
All
Visualization:
Node-Link
Sort by:
None
Edge centrality filters:
Images
Animate

21


Graph Viewer
Graph Viewer
Roll-up by:
All
Visualization:
Matrix
Sort by:
Linkage
Edge centrality filters:

22


Graph Viewer
Graph Viewer
Roll-up by:
All
Visualization:
Matrix
Sort by:
None
Edge centrality filters:

23

Visualize Friends by School?

Berkeley	|||||||||||||||||||||||||||||||||
Cornell	||||
Harvard	|||||||||
Harvard University	|||||||
Stanford	||||||||||||||||||||
Stanford University	||||||||||
UC Berkeley	||||||||||||||||||||||
UC Davis	||||||||||
Univ. of California at Berkeley	||||||||||||||||
Univ. of California, Berkeley	||||||||||||||||||
Univ. of California, Davis	|||

24

Data Quality Hurdles

Missing Data	no measurements, redacted, ...?
Erroneous Values	misspelling, outliers, ...?
Type Conversion	e.g., zip code to lat-lon
Entity Resolution	diff. values for the same thing?
Data Integration	effort/errors when combining data

LESSON: **Anticipate problems with your data. Many research problems around these issues!**

25

Analysis Example: Effectiveness of Antibiotics

35

Antibiotic Effectiveness: The Data

Genus of Bacteria	String
Species of Bacteria	String
Antibiotic Applied	String
Gram-Staining	Pos / Neg
Min. Inhibitory Concent. (g)	Number

Collected prior to 1951

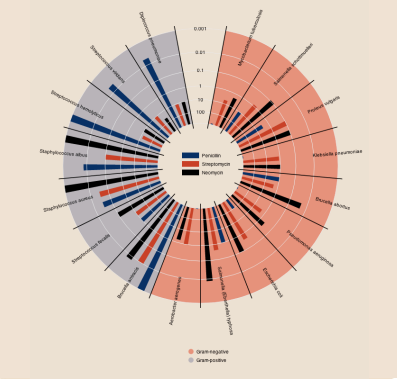
36

What questions might we ask?

Table 1: Burtin's data.	Antibiotic			
Bacteria	Penicillin	Streptomycin	Neomycin	Gram Staining
Aerobacter *aerogenes*	870	1	1.6	negative
Brucella *abortus*	1	2	0.02	negative
Brucella *anthracis*	0.001	0.01	0.007	positive
Diplococcus *pneumoniae*	0.005	11	10	positive
Escherichia *coli*	100	0.4	0.1	negative
Klebsiella *pneumoniae*	850	1.2	1	negative
Mycobacterium *tuberculosis*	800	5	2	negative
Proteus *vulgaris*	3	0.1	0.1	negative
Pseudomonas *aeruginosa*	850	2	0.4	negative
Salmonella (Eberthella) *typhosa*	1	0.4	0.008	negative
Salmonella *schottmuelleri*	10	0.8	0.09	negative
Staphylococcus *albus*	0.007	0.1	0.001	positive
Staphylococcus *aureus*	0.03	0.03	0.001	positive
Streptococcus *fecalis*	1	1	0.1	positive
Streptococcus *hemolyticus*	0.001	14	10	positive
Streptococcus *viridans*	0.005	10	40	positive

37

Will Burtin, 1951

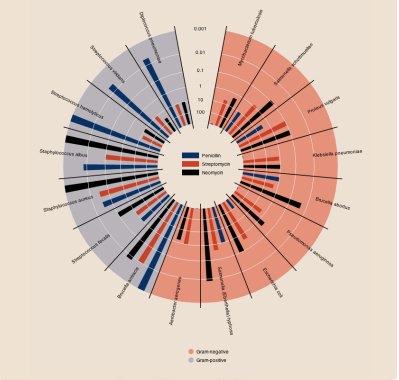


Bacteria	Penicillin	Antibiotic Streptomycin	Neomycin	Gram stain
Aerobacter aerogenes	870	1	1.6	–
Brucella abortus	1	2	0.02	–
Bacillus anthracis	0.001	0.01	0.007	+
Diplococcus pneumoniae	0.005	11	10	+
Escherichia coli	100	0.4	0.1	–
Klebsiella pneumoniae	850	1.2	1	–
Mycobacterium tuberculosis	800	5	2	–
Proteus vulgaris	3	0.1	0.1	–
Pseudomonas aeruginosa	850	2	0.4	–
Salmonella (Eberthella) typhosa	1	0.4	0.008	–
Salmonella schottmuelleri	10	0.8	0.09	–
Staphylococcus albus	0.007	0.1	0.001	+
Staphylococcus aureus	0.03	0.03	0.001	+
Streptococcus fecalis	1	1	0.1	+
Streptococcus hemolyticus	0.001	14	10	+
Streptococcus viridans	0.005	10	40	+

How do the drugs compare?

38

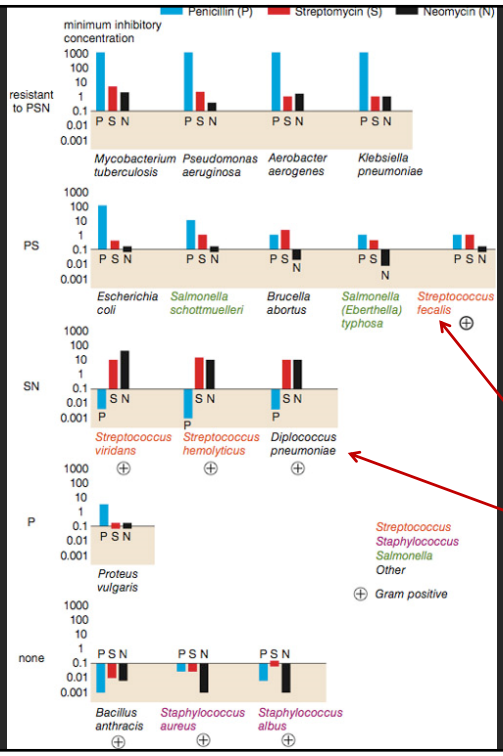
Will Burtin, 1951



Bacteria	Penicillin	Antibiotic Streptomycin	Neomycin	Gram stain
Aerobacter aerogenes	870	1	1.6	–
Brucella abortus	1	2	0.02	–
Bacillus anthracis	0.001	0.01	0.007	+
Diplococcus pneumoniae	0.005	11	10	+
Escherichia coli	100	0.4	0.1	–
Klebsiella pneumoniae	850	1.2	1	–
Mycobacterium tuberculosis	800	5	2	–
Proteus vulgaris	3	0.1	0.1	–
Pseudomonas aeruginosa	850	2	0.4	–
Salmonella (Eberthella) typhosa	1	0.4	0.008	–
Salmonella schottmuelleri	10	0.8	0.09	–
Staphylococcus albus	0.007	0.1	0.001	+
Staphylococcus aureus	0.03	0.03	0.001	+
Streptococcus fecalis	1	1	0.1	+
Streptococcus hemolyticus	0.001	14	10	+
Streptococcus viridans	0.005	10	40	+

Radius: 1/log(MIC)
Bar Color: Antibiotic
Background Color: Gram Staining

39



Do bacteria group by antibiotic resistance?

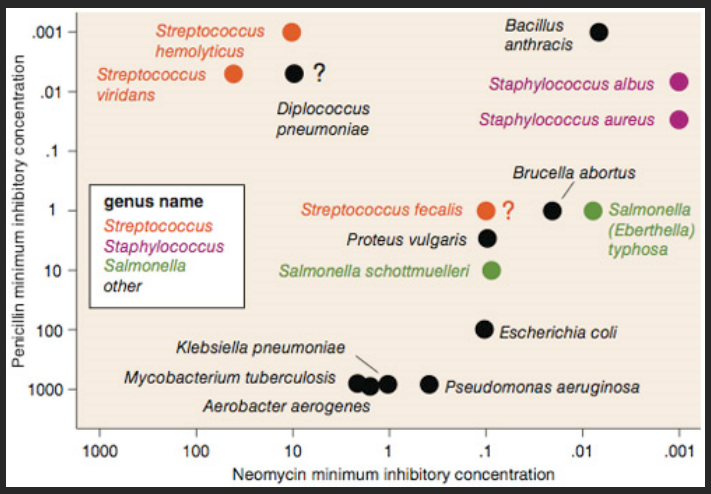
Not a streptococcus! (realized ~30 yrs later)

Really a streptococcus! (realized ~20 yrs later)

Wainer & Lysen

American Scientist, 2009

40



How do the bacteria group w.r.t. resistance? Do different drugs correlate?

Wainer & Lysen

American Scientist, 2009

41

Lessons

Exploratory Process

1. **Construct graphics to address questions**
2. **Inspect "answer" and assess new questions**
3. **Repeat!**

Transform the data appropriately (e.g., invert, log)

"Show data variation, not design variation"

-Tufte

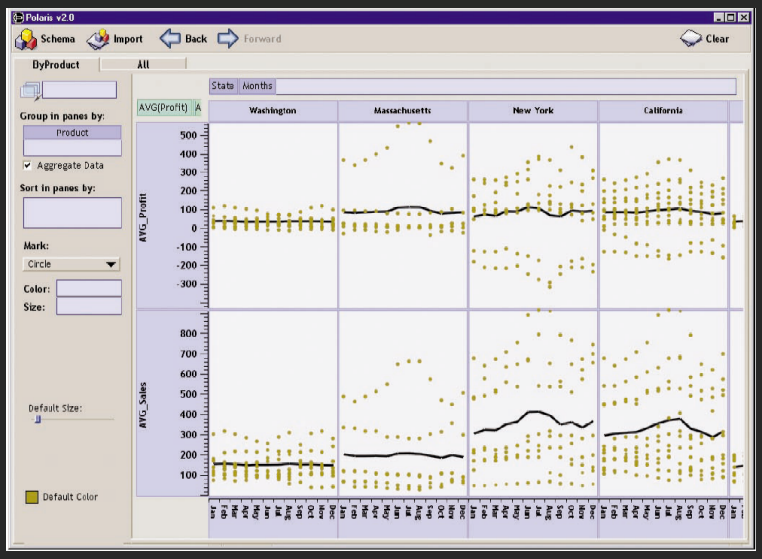
42

Tableau / Polaris

77

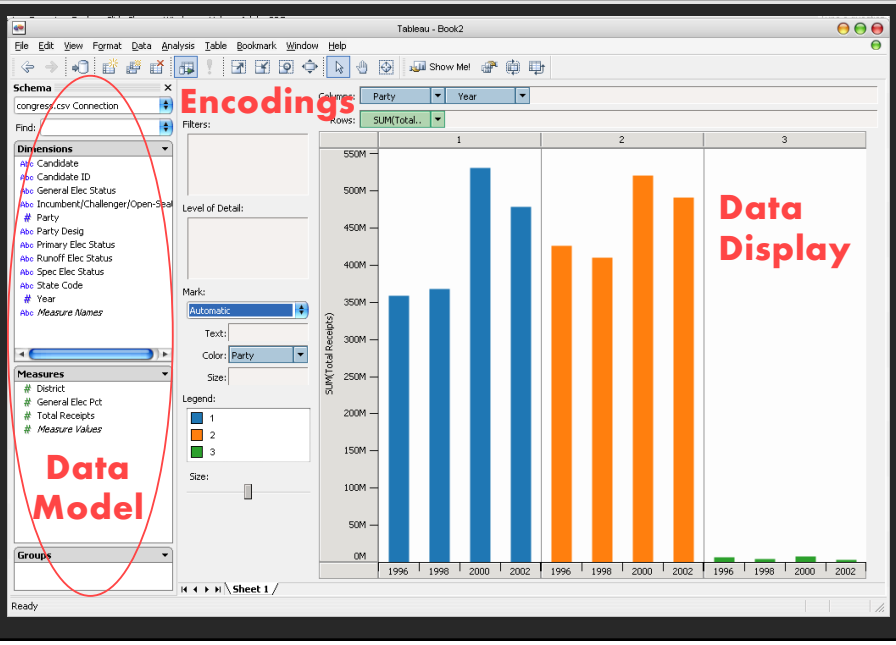
Tableau

Research at Stanford: “Polaris” by Stolte, Tang & Hanrahan.



78

Tableau



79

Polaris/Tableau Approach

Insight: simultaneously specify both database queries and visualization

Choose data, then visualization, not vice versa

Use smart defaults for visual encodings

Can also suggest more encodings upon request (ShowMe – Like APT)

80

Dataset

- Federal Elections Commission Receipts
- Every Congressional Candidate from 1996 to 2002
- 4 Election Cycles
- 9216 Candidacies

81

Data Set Schema

- Year (Qi)
- Candidate Code (N)
- Candidate Name (N)
- Incumbent / Challenger / Open-Seat (N)
- Party Code (N) [1=Dem,2=Rep,3=Other]
- Party Name (N)
- Total Receipts (Qr)
- State (N)
- District (N)

- This is a subset of the larger data set available from the FEC, but should be sufficient for the demo

82

Hypotheses?

What might we learn from this data?

83

Hypotheses?

What might we learn from this data?

- Have receipts increased over time?
- Do democrats or republicans spend more?
- Candidates from which state spend the most money?

Tableau Demo

84

Specifying Table Configurations

Operands are names of database fields

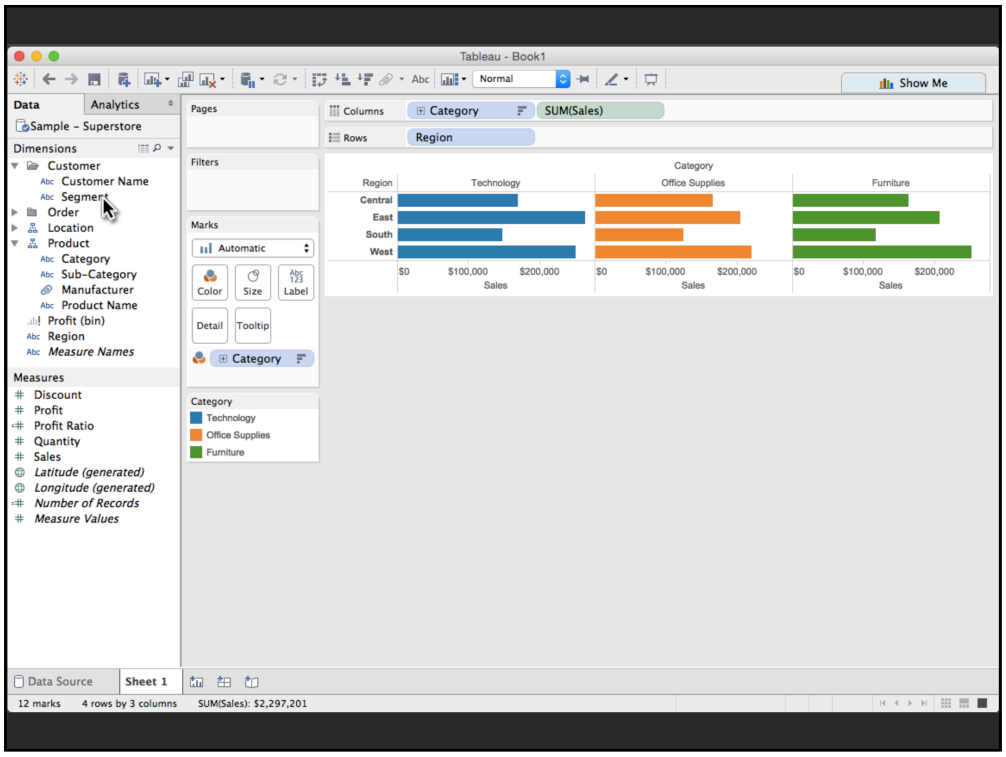
Each operand interpreted as a set {…}

Data is either O or Q and treated differently

Three operators:

- concatenation (+)
- cross product (x)
- nest (/)

85



86

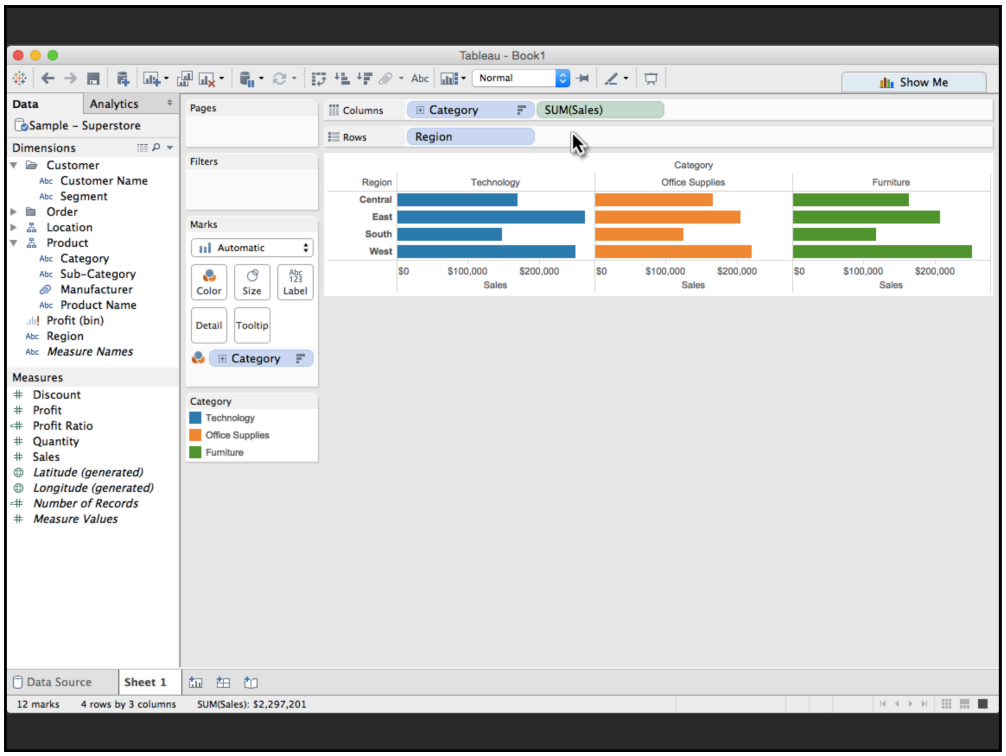


Tableau - Book1
Normal
Show Me
Data
Analytics
Sample – Superstore
Dimensions
Customer
Customer Name
Segment
Order
Location
Product
Category
Sub-Category
Manufacturer
Product Name
Profit (bin)
Region
Measure Names
Measures
Discount
Profit
Profit Ratio
Quantity
Sales
Latitude (generated)
Longitude (generated)
Number of Records
Measure Values
Pages
Filters
Marks
Automatic
Color
Size
Label
Detail
Tooltip
Category
Technology
Office Supplies
Furniture
Columns
Category
SUM(Sales)
Rows
Region
Central
East
South
West
$0
$100,000
$200,000
Sales
Data Source
Sheet 1
12 marks
4 rows by 3 columns
SUM(Sales): $2,297,201

87

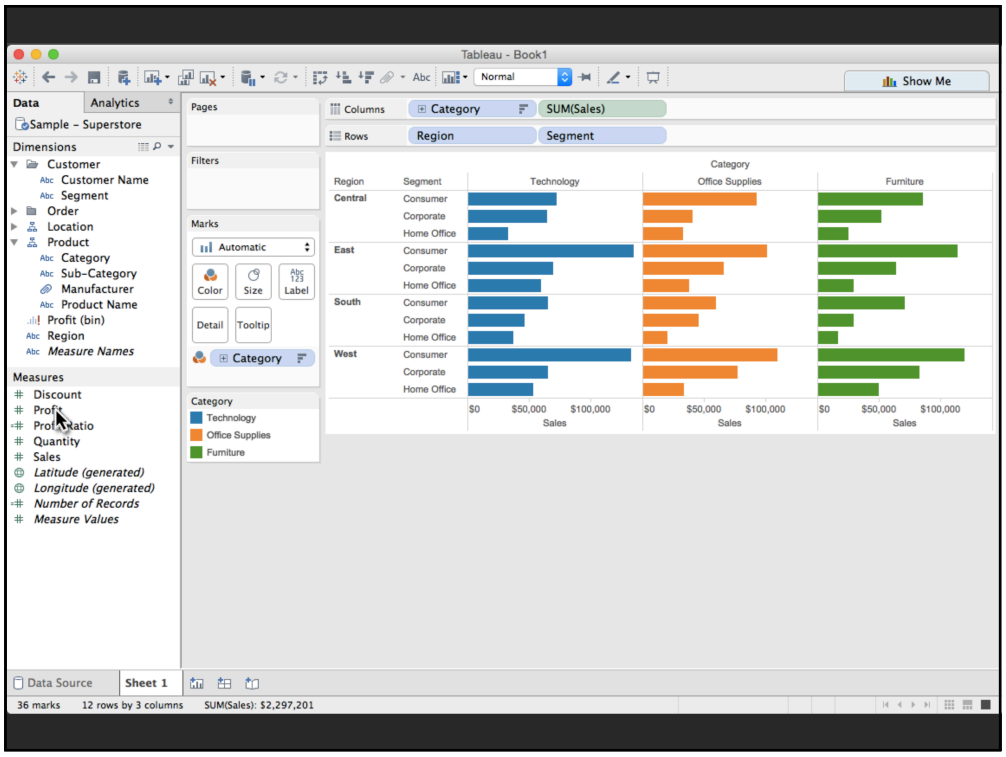


Tableau - Book1
Normal
Show Me
Data
Analytics
Sample – Superstore
Dimensions
Customer
Customer Name
Segment
Order
Location
Product
Category
Sub-Category
Manufacturer
Product Name
Profit (bin)
Region
Measure Names
Measures
Discount
Profit
Profit Ratio
Quantity
Sales
Latitude (generated)
Longitude (generated)
Number of Records
Measure Values
Pages
Filters
Marks
Automatic
Color
Size
Label
Detail
Tooltip
Category
Technology
Office Supplies
Furniture
Columns
Category
SUM(Sales)
Rows
Region
Segment
Central
East
South
West
Consumer
Corporate
Home Office
$0
$50,000
$100,000
Sales
Data Source
Sheet 1
36 marks
12 rows by 3 columns
SUM(Sales): $2,297,201

88

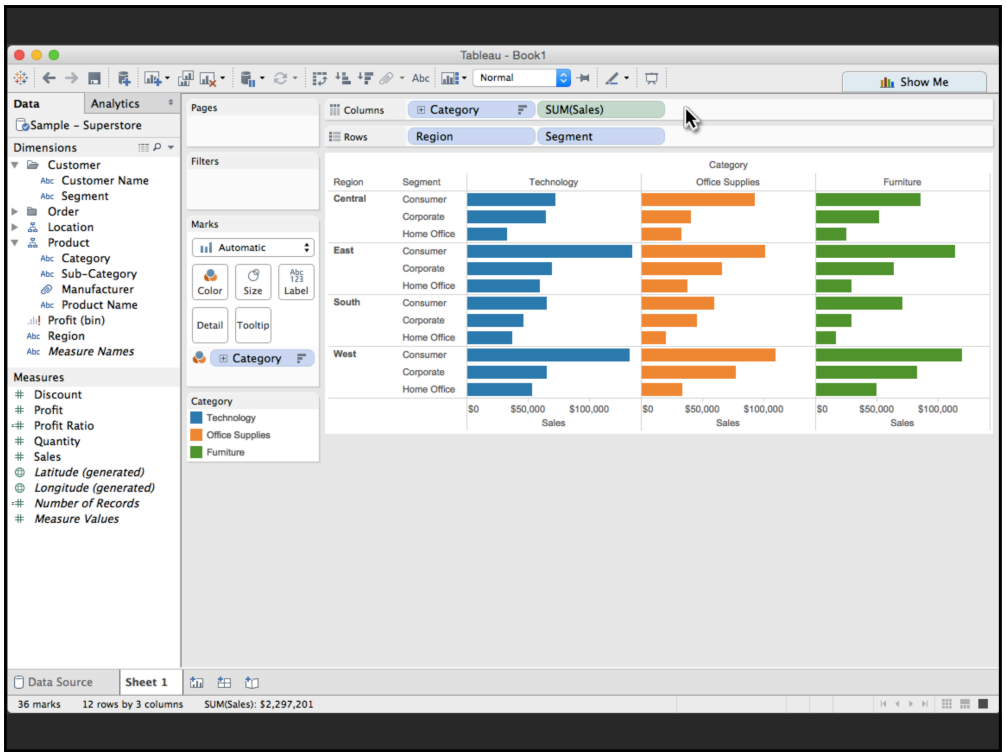


Tableau - Book1
Show Me
Data
Analytics
Sample – Superstore
Dimensions
Customer
Customer Name
Segment
Order
Location
Product
Category
Sub-Category
Manufacturer
Product Name
Profit (bin)
Region
Measure Names
Measures
Discount
Profit
Profit Ratio
Quantity
Sales
Latitude (generated)
Longitude (generated)
Number of Records
Measure Values
Pages
Filters
Marks
Automatic
Color
Size
Label
Detail
Tooltip
Category
Technology
Office Supplies
Furniture
Columns
Category
SUM(Sales)
Rows
Region
Segment
Central
East
South
West
Consumer
Corporate
Home Office
$0
$50,000
$100,000
Sales
Data Source
Sheet 1
36 marks
12 rows by 3 columns
SUM(Sales): $2,297,201

89

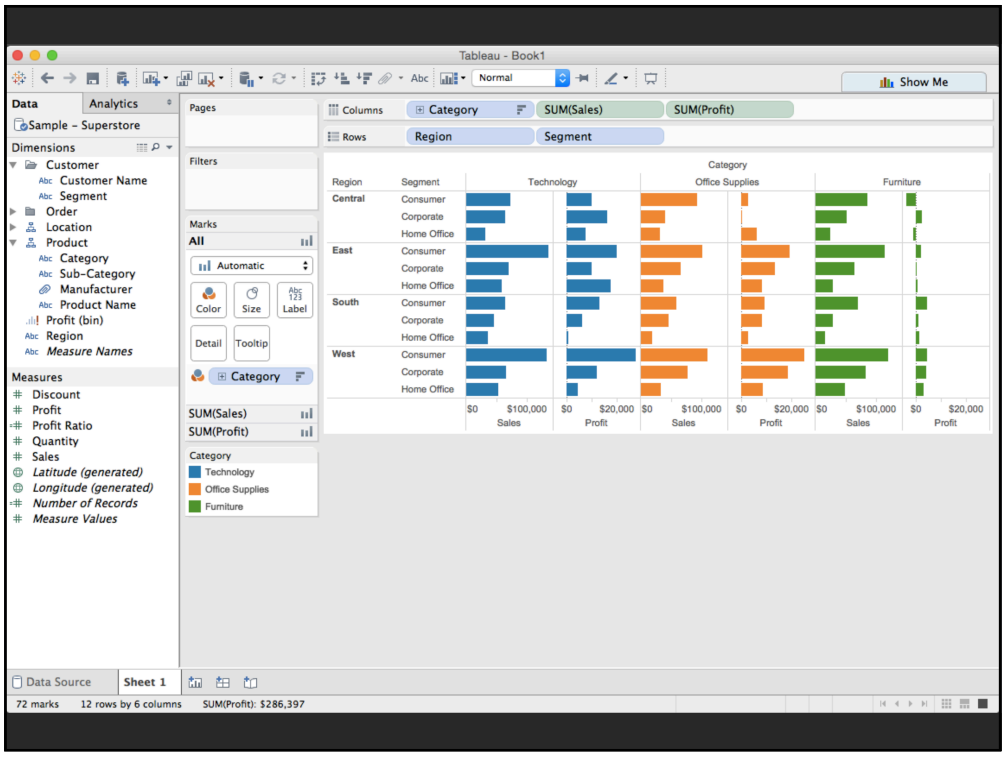


Tableau - Book1
Show Me
Data
Analytics
Sample – Superstore
Dimensions
Customer
Customer Name
Segment
Order
Location
Product
Category
Sub-Category
Manufacturer
Product Name
Profit (bin)
Region
Measure Names
Measures
Discount
Profit
Profit Ratio
Quantity
Sales
Latitude (generated)
Longitude (generated)
Number of Records
Measure Values
Pages
Filters
Marks
All
Automatic
Color
Size
Label
Detail
Tooltip
Category
SUM(Sales)
SUM(Profit)
Technology
Office Supplies
Furniture
Columns
Category
SUM(Sales)
SUM(Profit)
Rows
Region
Segment
Central
East
South
West
Consumer
Corporate
Home Office
$0
$100,000
$20,000
Sales
Profit
Data Source
Sheet 1
72 marks
12 rows by 6 columns
SUM(Profit): $286,397

90


Columns
Category
SUM(Sales)
SUM(Profit)
Rows
Region
Segment
Category
Region
Segment
Technology
Office Supplies
Furniture
Central
East
South
West
Consumer
Corporate
Home Office
$0
$100,000
$20,000
Sales
Profit

91