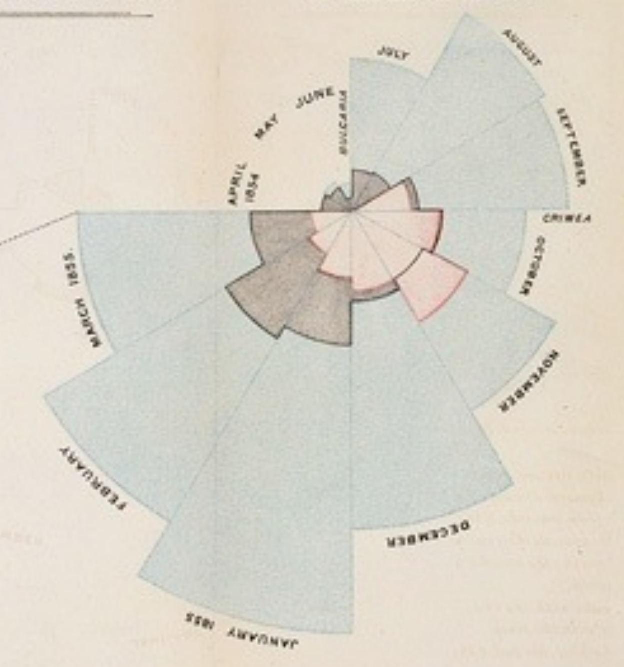
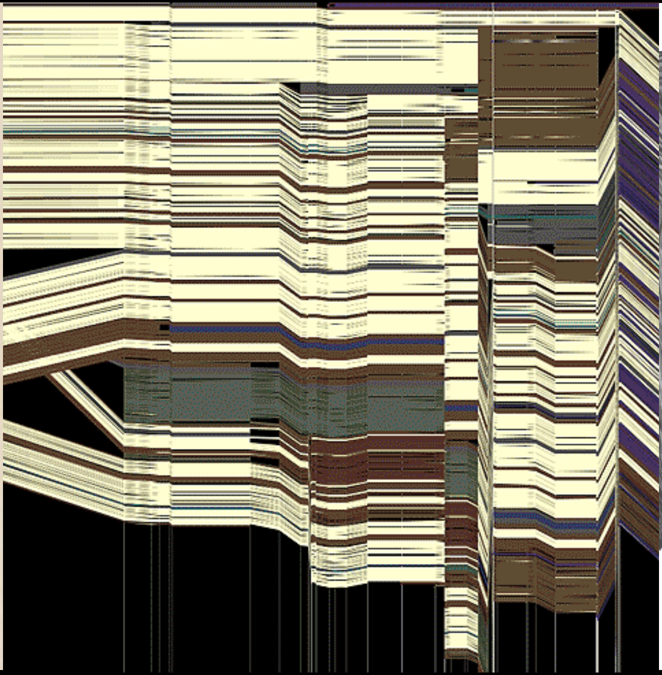
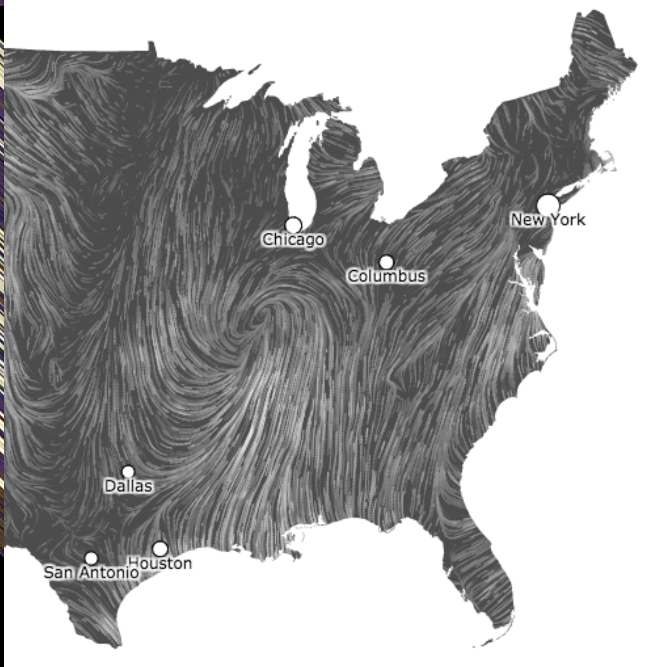
CSE 412 - Intro to Data Visualization

Exploratory Data Analysis





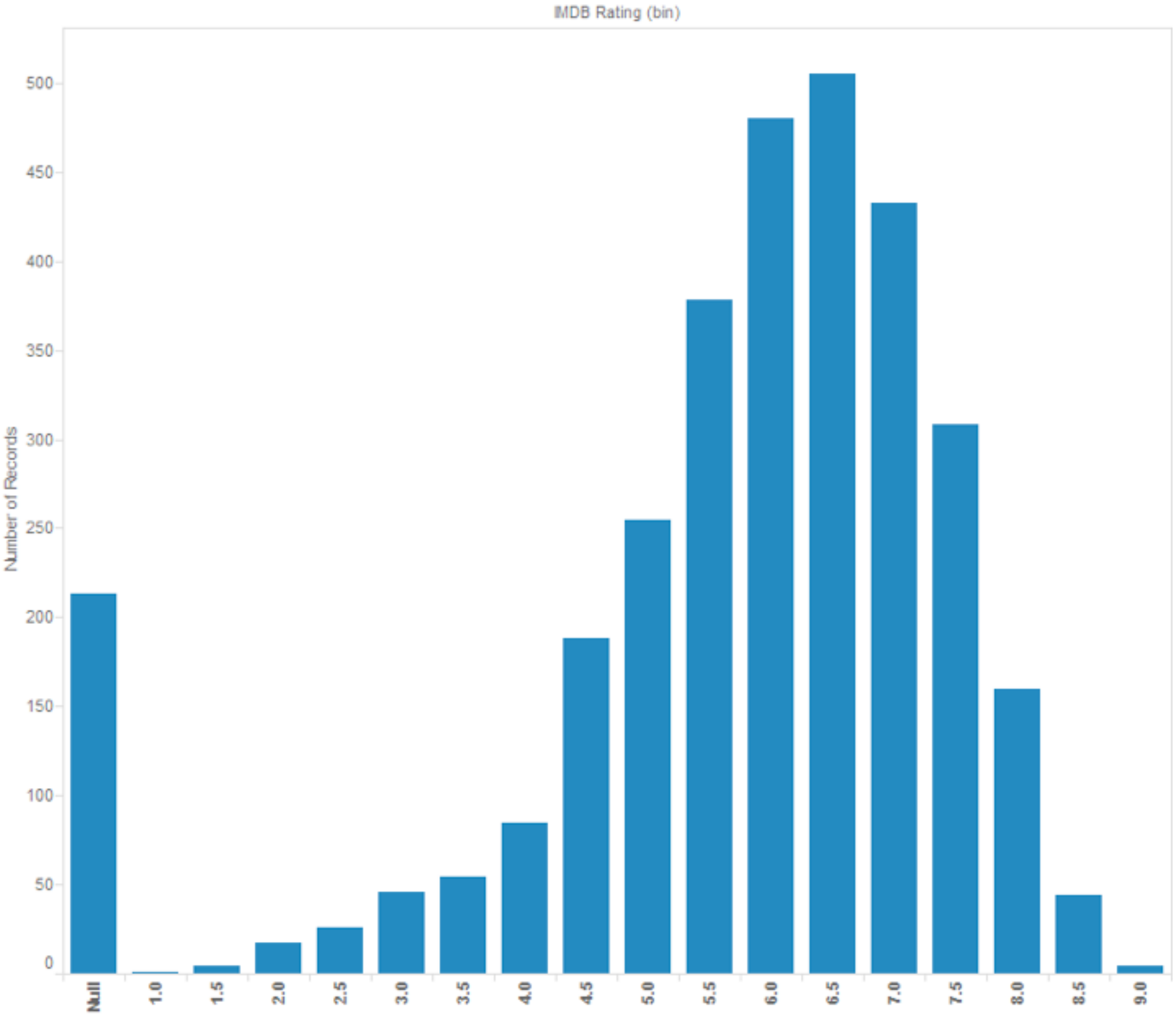


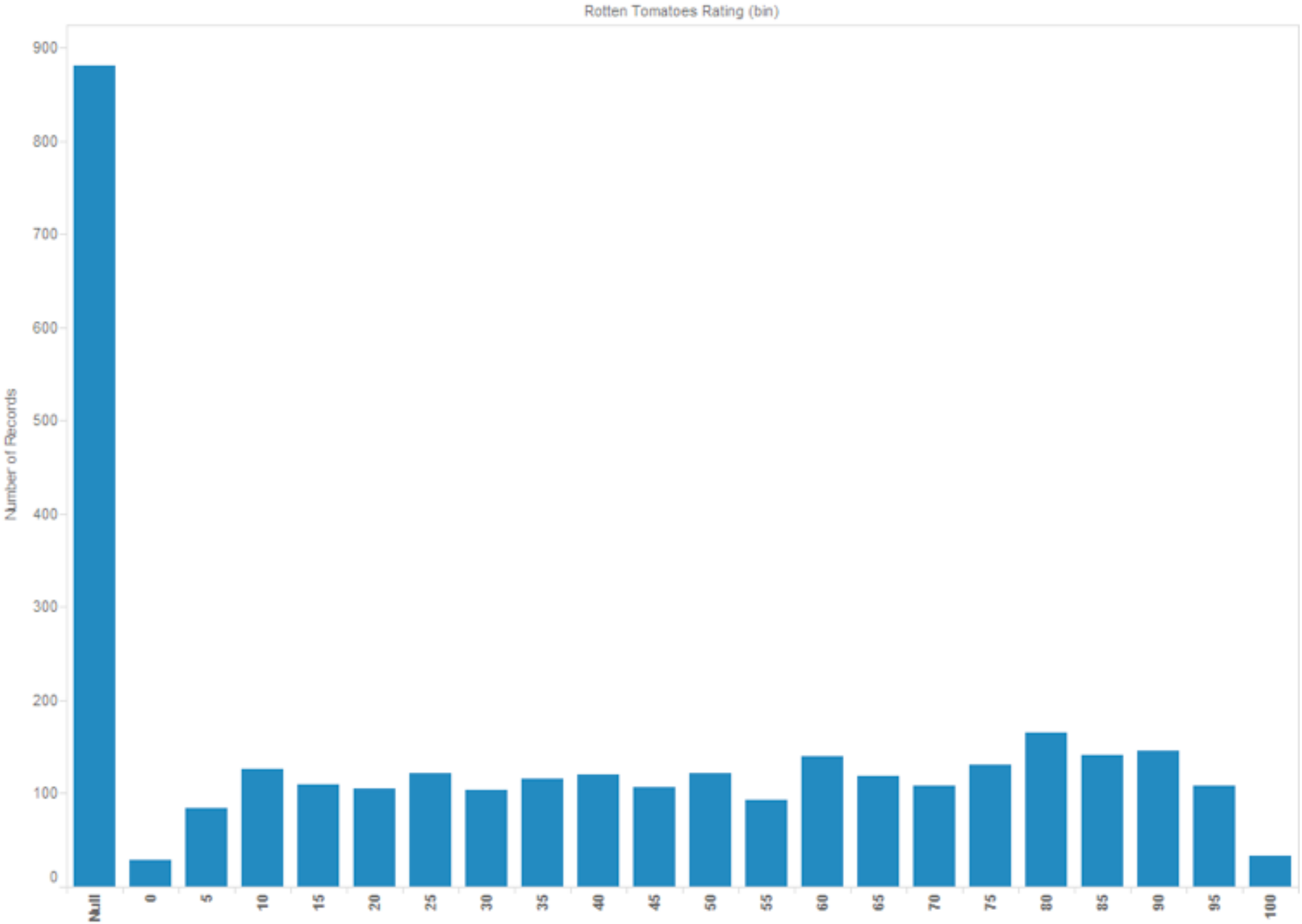
Jane Hoffswell University of Washington

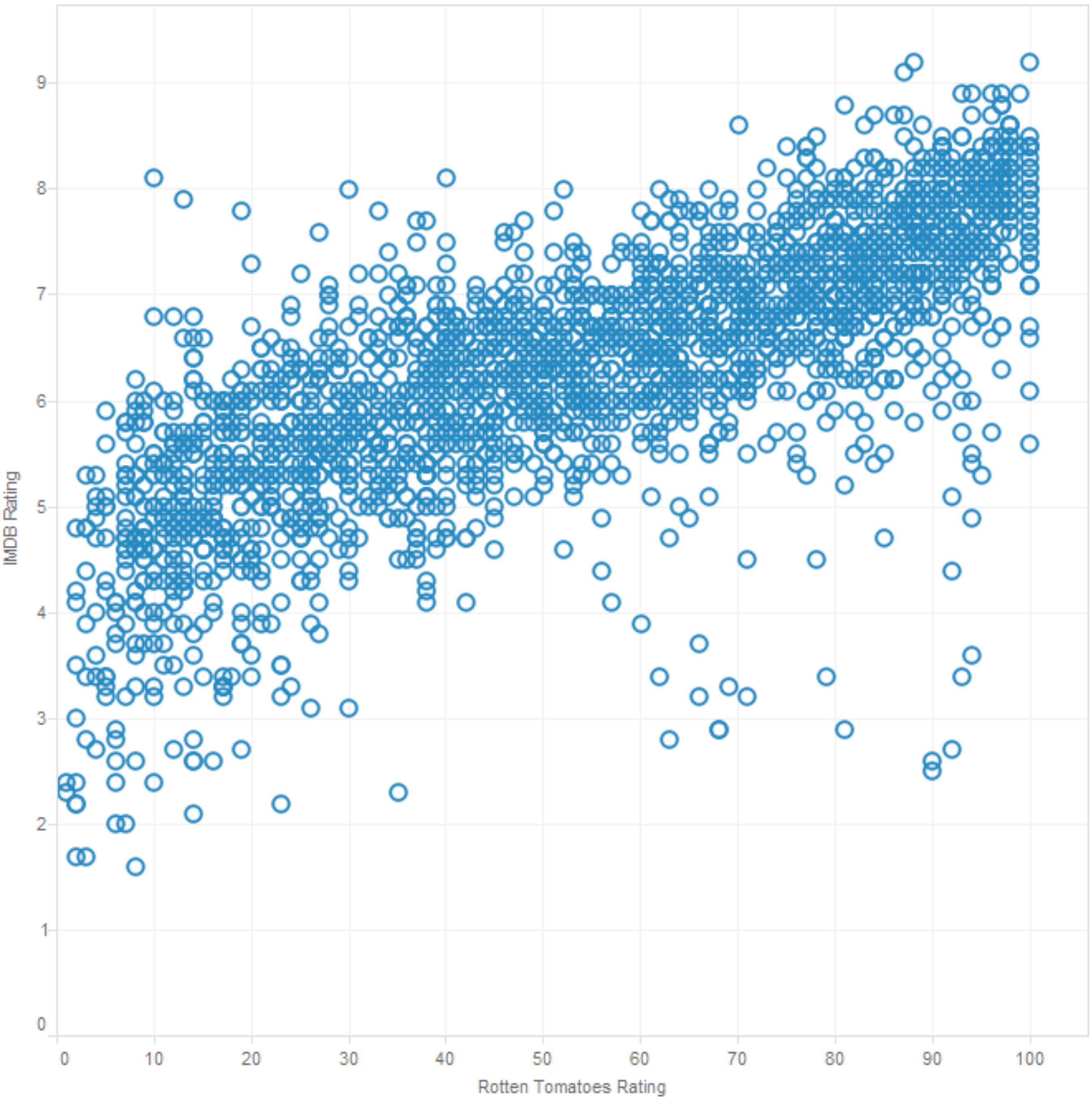
Analysis Example: Motion Pictures Data

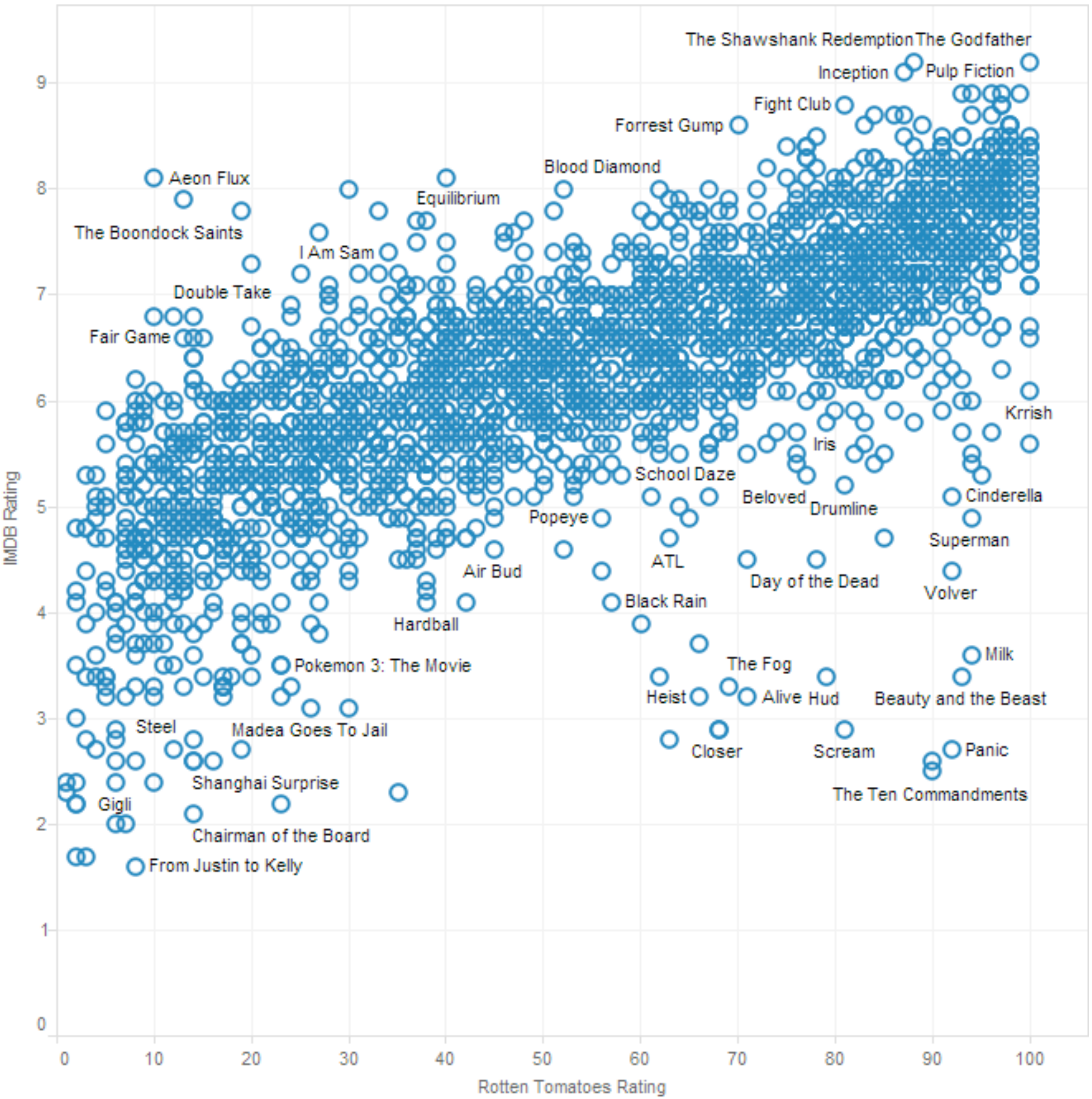
Motion Pictures Data

Title	String (N)
IMDB Rating	Number (Q)
Rotten Tomatoes Rating	Number (Q)
MPAA Rating	String (O)
Release Date	Date (T)

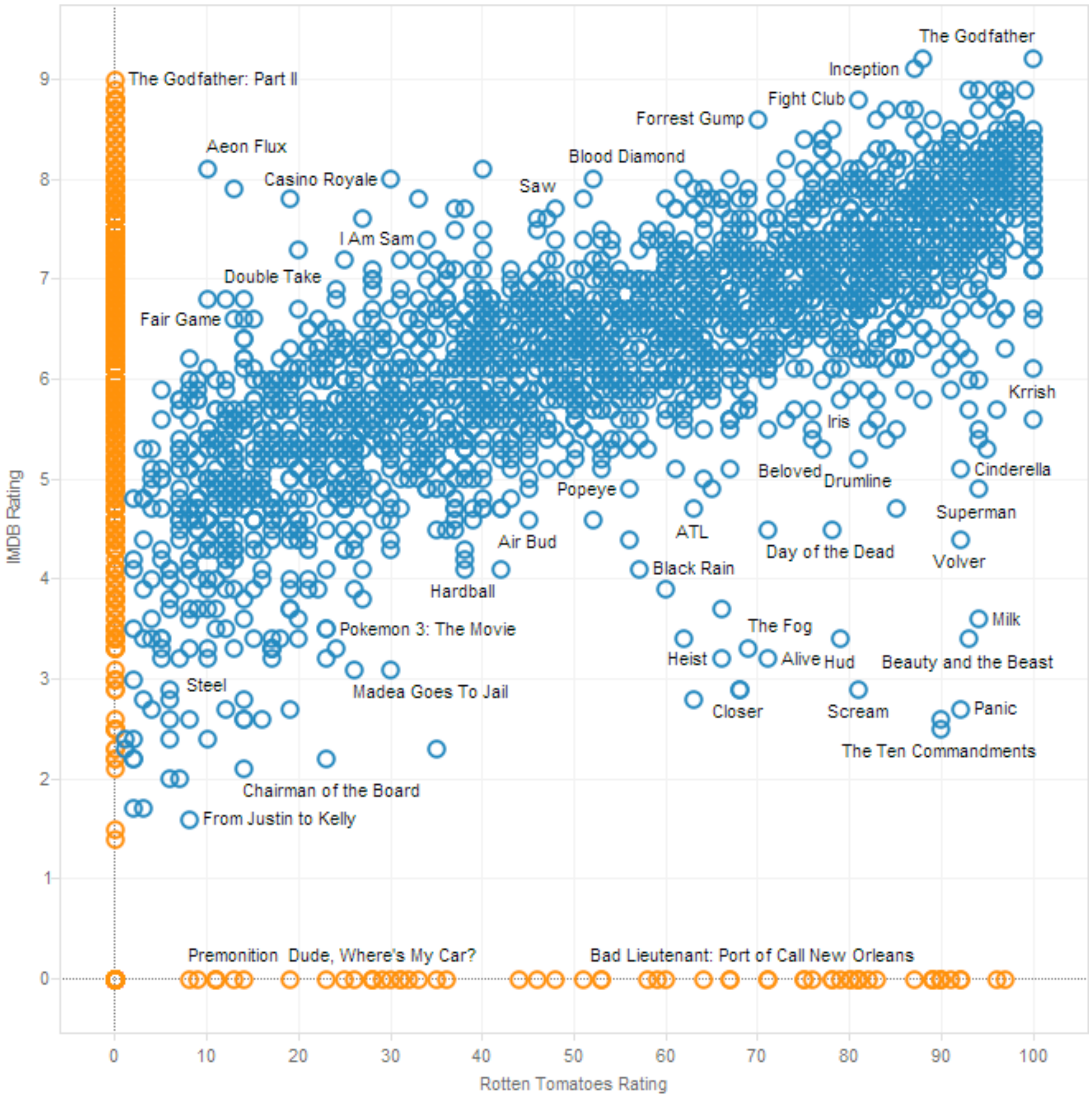

IMDB Rating (bin)
Number of Records
500
450
400
350
300
250
200
150
100
50
0
Null
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
5.5
6.0
6.5
7.0
7.5
8.0
8.5
9.0

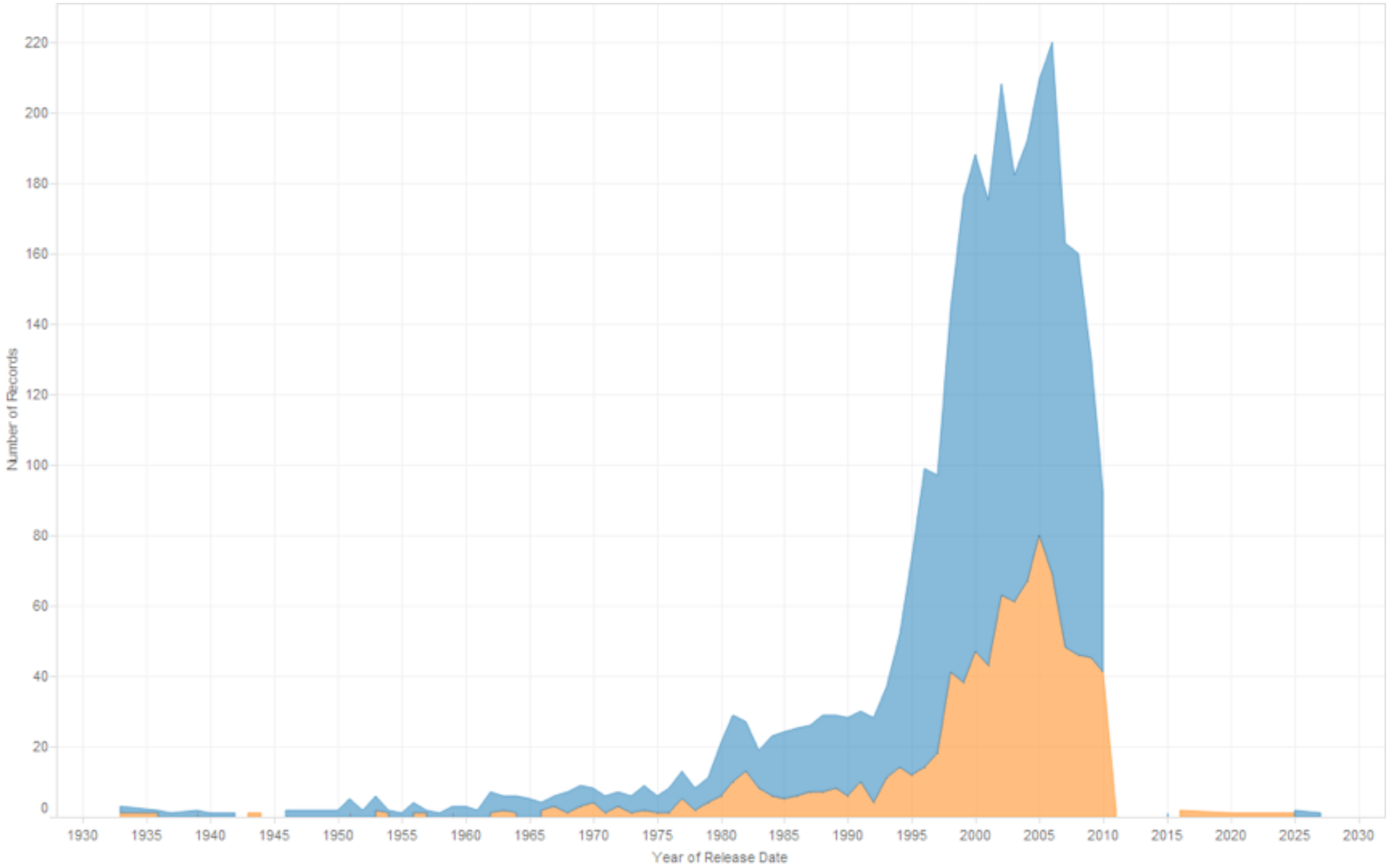

Rotten Tomatoes Rating (bin)
Number of Records
900
800
700
600
500
400
300
200
100
0
Null
0
5
10
15
20
25
30
35
40
45
50
55
60
65
70
75
80
85
90
95
100

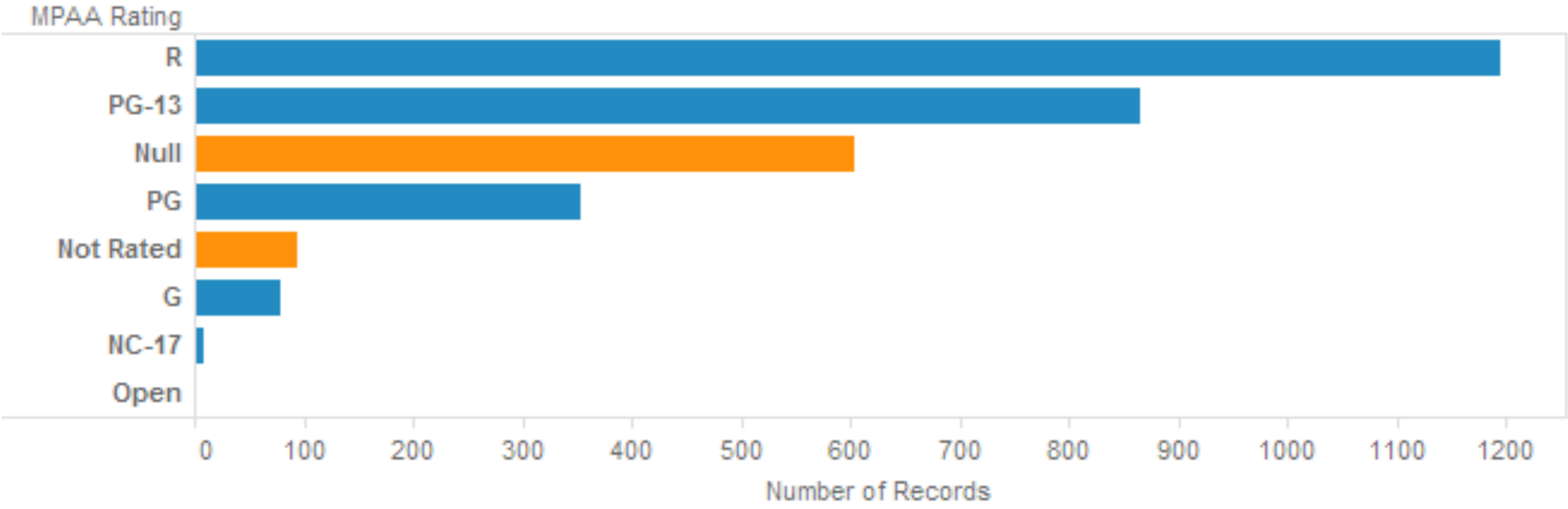

IMDB Rating
Rotten Tomatoes Rating
0
1
2
3
4
5
6
7
8
9
0
10
20
30
40
50
60
70
80
90
100

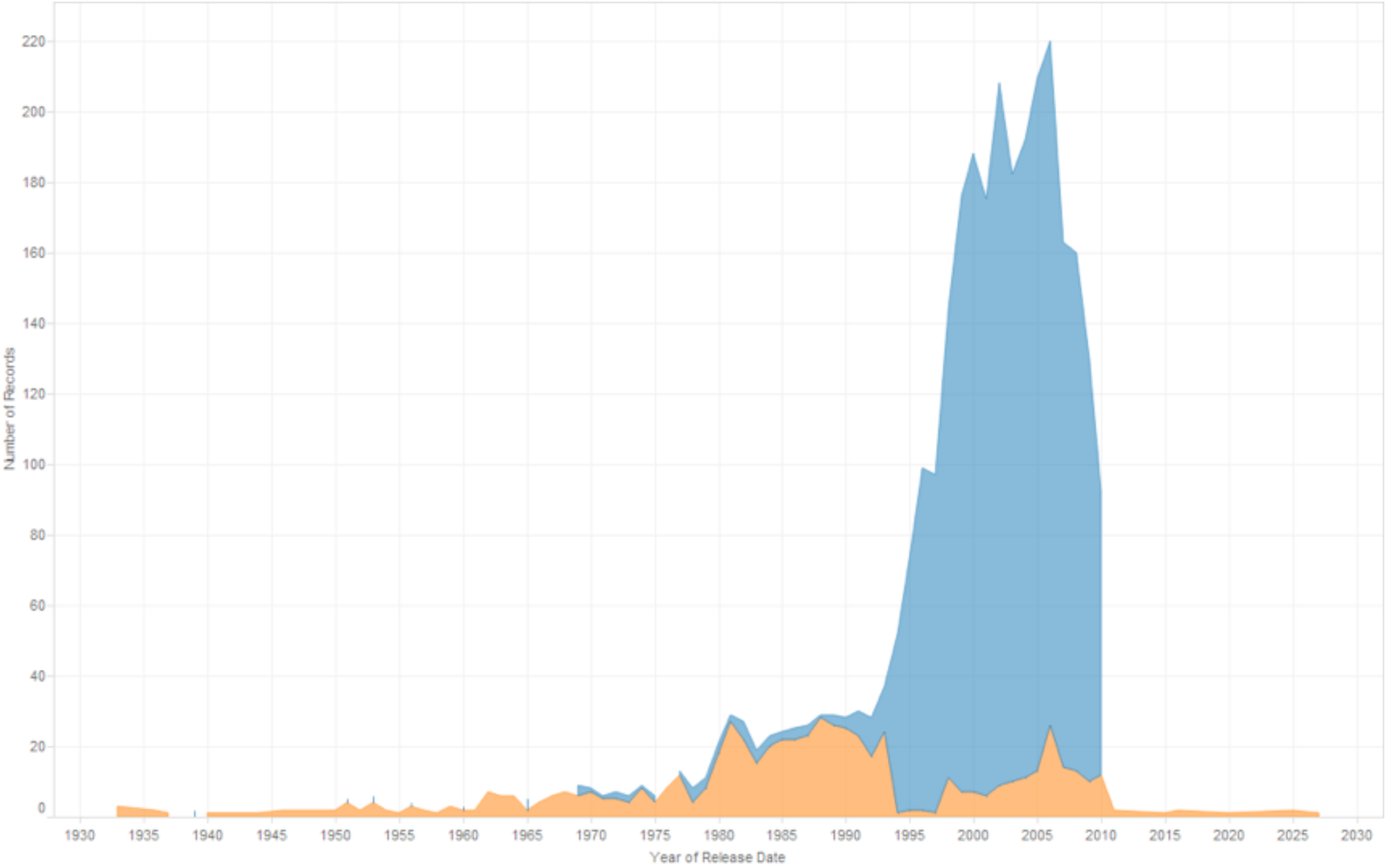


The Shawshank Redemption
The Godfather
Inception
Pulp Fiction
Fight Club
Forrest Gump
Blood Diamond
Aeon Flux
Equilibrium
The Boondock Saints
I Am Sam
Double Take
Fair Game
Krrish
Iris
School Daze
Beloved
Drumline
Cinderella
Popeye
Superman
ATL
Air Bud
Day of the Dead
Volver
Black Rain
Hardball
Pokemon 3: The Movie
The Fog
Milk
Heist
Alive
Hud
Beauty and the Beast
Steel
Madea Goes To Jail
Closer
Scream
Panic
Shanghai Surprise
The Ten Commandments
Gigli
Chairman of the Board
From Justin to Kelly
IMDB Rating
0
1
2
3
4
5
6
7
8
9
Rotten Tomatoes Rating
0
10
20
30
40
50
60
70
80
90
100


The Godfather
Inception
Fight Club
Forrest Gump
The Godfather: Part II
Aeon Flux
Casino Royale
Blood Diamond
Saw
I Am Sam
Double Take
Fair Game
Krrish
Iris
Beloved
Drumline
Cinderella
Popeye
Superman
ATL
Air Bud
Day of the Dead
Black Rain
Volver
Hardball
Milk
Pokemon 3: The Movie
The Fog
Heist
Alive
Hud
Beauty and the Beast
Steel
Madea Goes To Jail
Closer
Scream
Panic
The Ten Commandments
Chairman of the Board
From Justin to Kelly
Premonition
Dude, Where's My Car?
Bad Lieutenant: Port of Call New Orleans
IMDB Rating
0
1
2
3
4
5
6
7
8
9
Rotten Tomatoes Rating
0
10
20
30
40
50
60
70
80
90
100


Number of Records
220
200
180
160
140
120
100
80
60
40
20
0
1930
1935
1940
1945
1950
1955
1960
1965
1970
1975
1980
1985
1990
1995
2000
2005
2010
2015
2020
2025
2030
Year of Release Date


MPAA Rating
R
PG-13
Null
PG
Not Rated
G
NC-17
Open
0
100
200
300
400
500
600
700
800
900
1000
1100
1200
Number of Records


Number of Records
220
200
180
160
140
120
100
80
60
40
20
0
1930
1935
1940
1945
1950
1955
1960
1965
1970
1975
1980
1985
1990
1995
2000
2005
2010
2015
2020
2025
2030
Year of Release Date

Lesson: Exercise Skepticism

Check **data quality** and your **assumptions**.

Start with **univariate summaries**, then start to consider **relationships among variables**.

Avoid premature fixation!

Analysis Example: Antibiotic Effectiveness

Data Set: Antibiotic Effectiveness

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

Collected prior to 1951.

What questions might we ask?

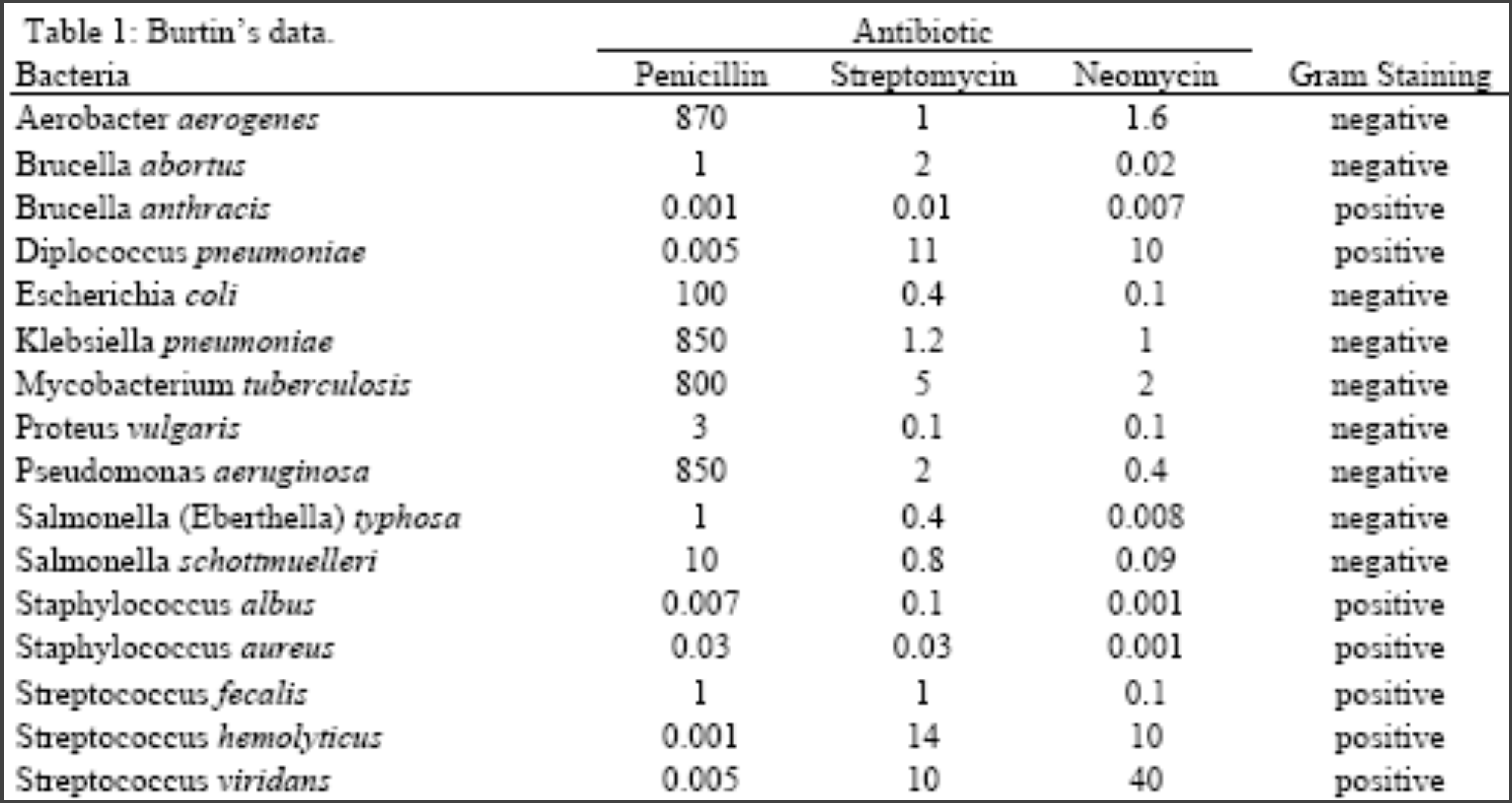
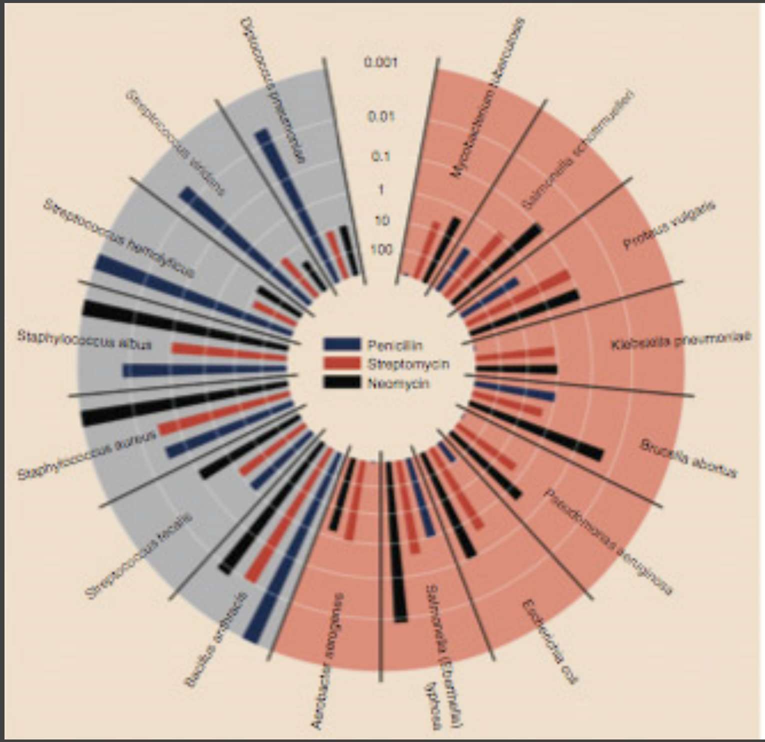
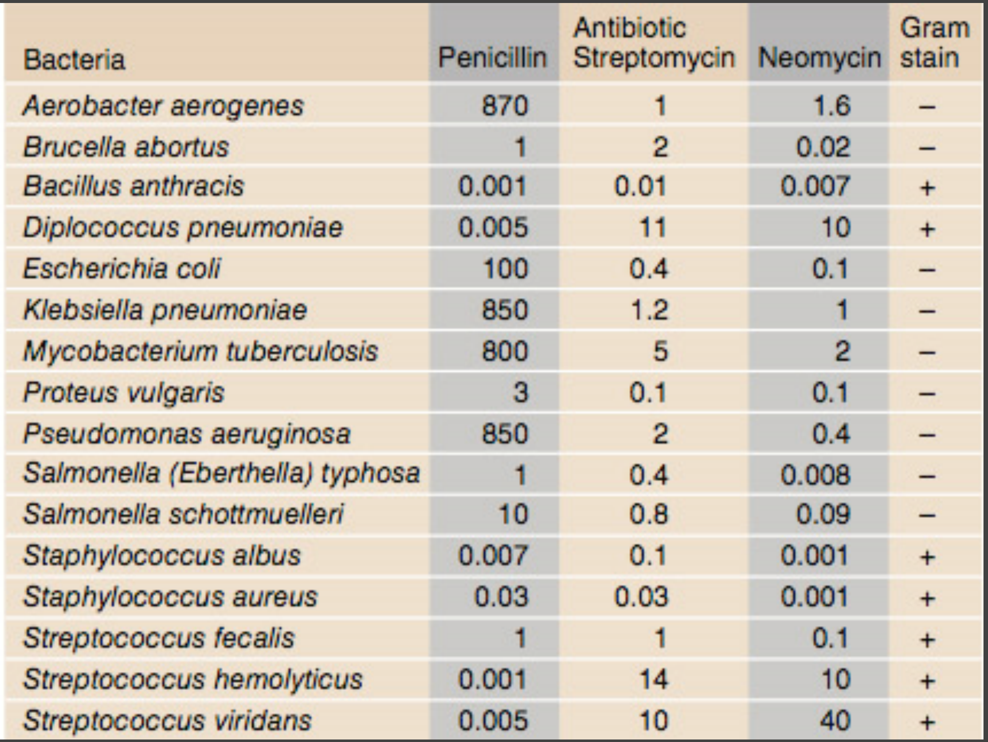


Table 1: Burtin's data.

	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

How do the drugs compare?

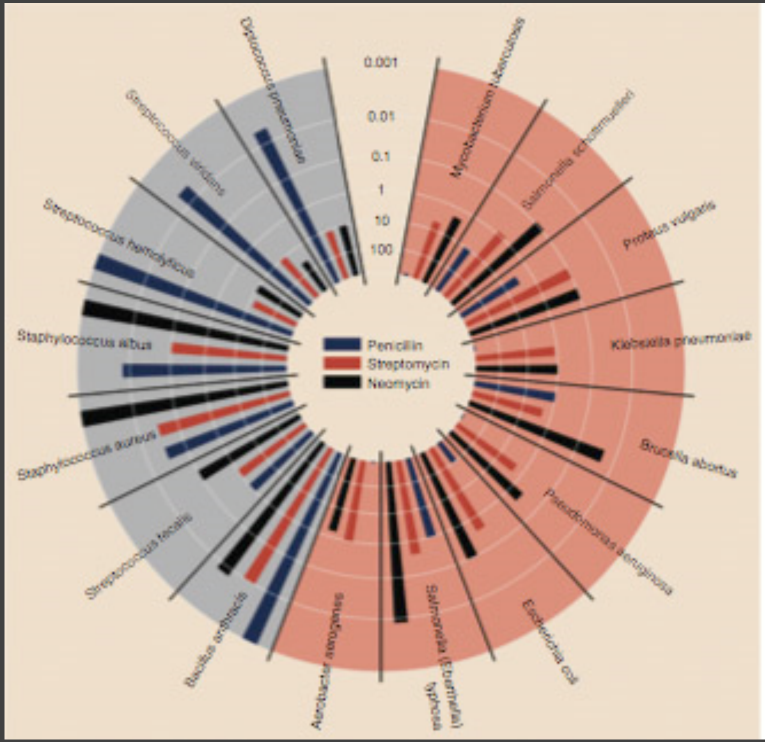


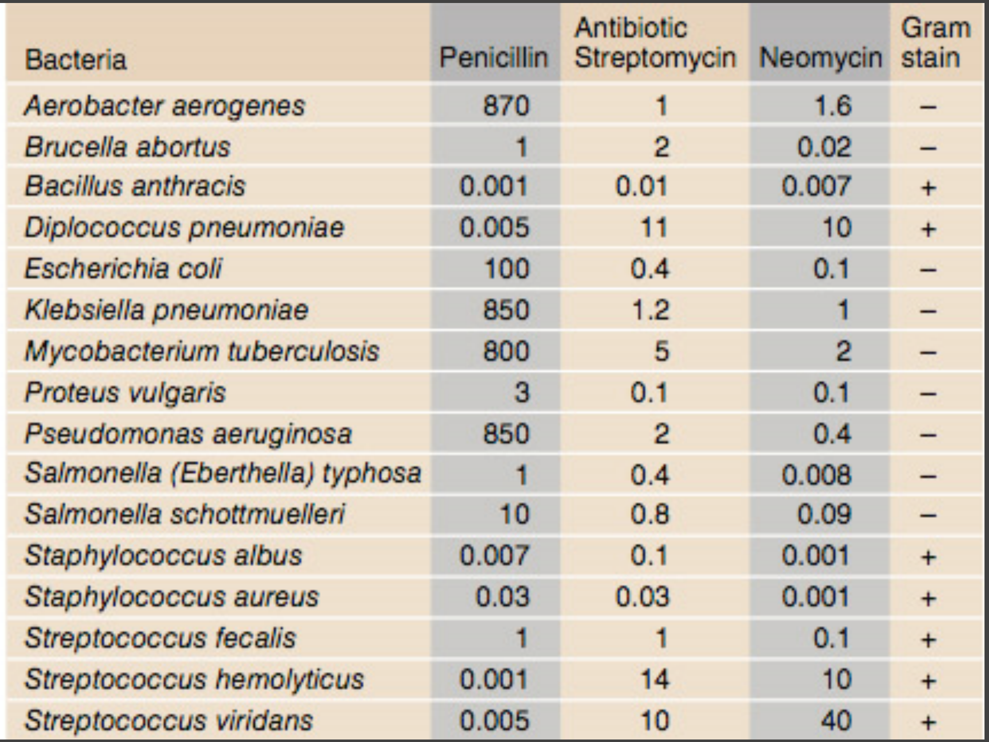


Bacteria	Penicillin	Antibiotic Streptomycin	Neomycin	Gram stain
Aerobacter aerogenes	870	1	1.6	–
Brucella abortus	1	2	0.02	–
Bacillus anthracis	0.001	0.01	0.007	+
Diplococcus pneumoniae	0.005	11	10	+
Escherichia coli	100	0.4	0.1	–
Klebsiella pneumoniae	850	1.2	1	–
Mycobacterium tuberculosis	800	5	2	–
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	+

Original graphic by Will Burtin, 1951

How do the drugs compare?





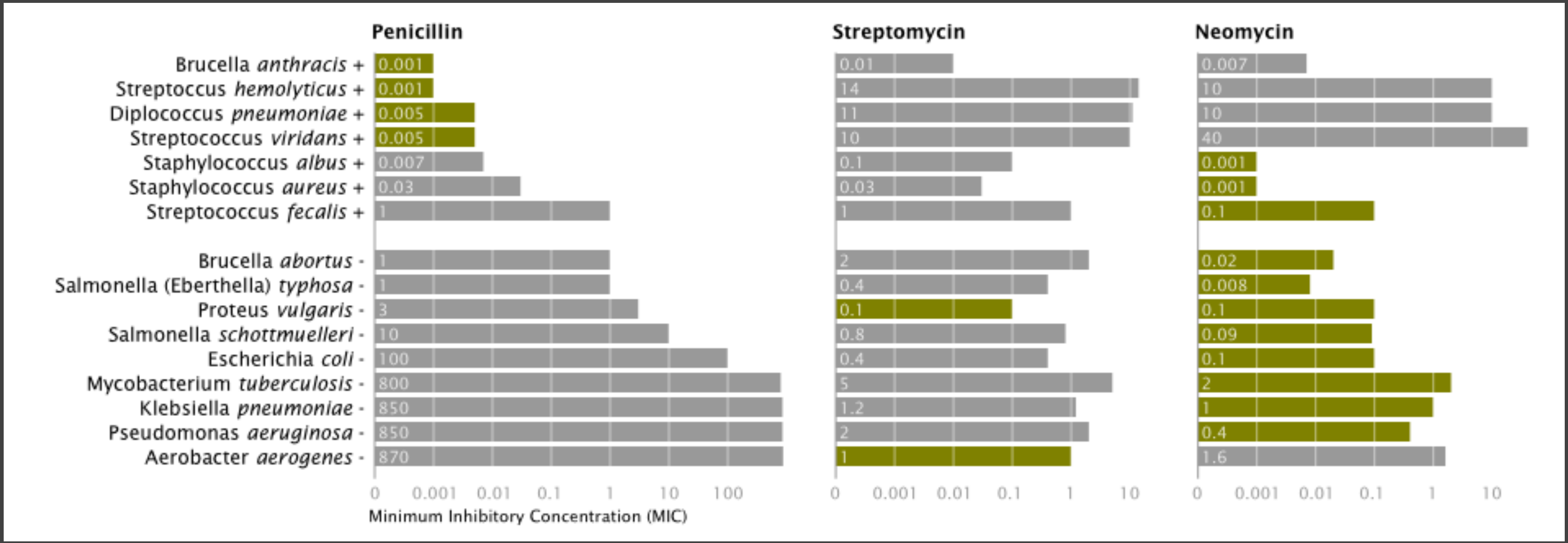
Bacteria	Penicillin	Antibiotic Streptomycin	Neomycin	Gram stain
Aerobacter aerogenes	870	1	1.6	–
Brucella abortus	1	2	0.02	–
Bacillus anthracis	0.001	0.01	0.007	+
Diplococcus pneumoniae	0.005	11	10	+
Escherichia coli	100	0.4	0.1	–
Klebsiella pneumoniae	850	1.2	1	–
Mycobacterium tuberculosis	800	5	2	–
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

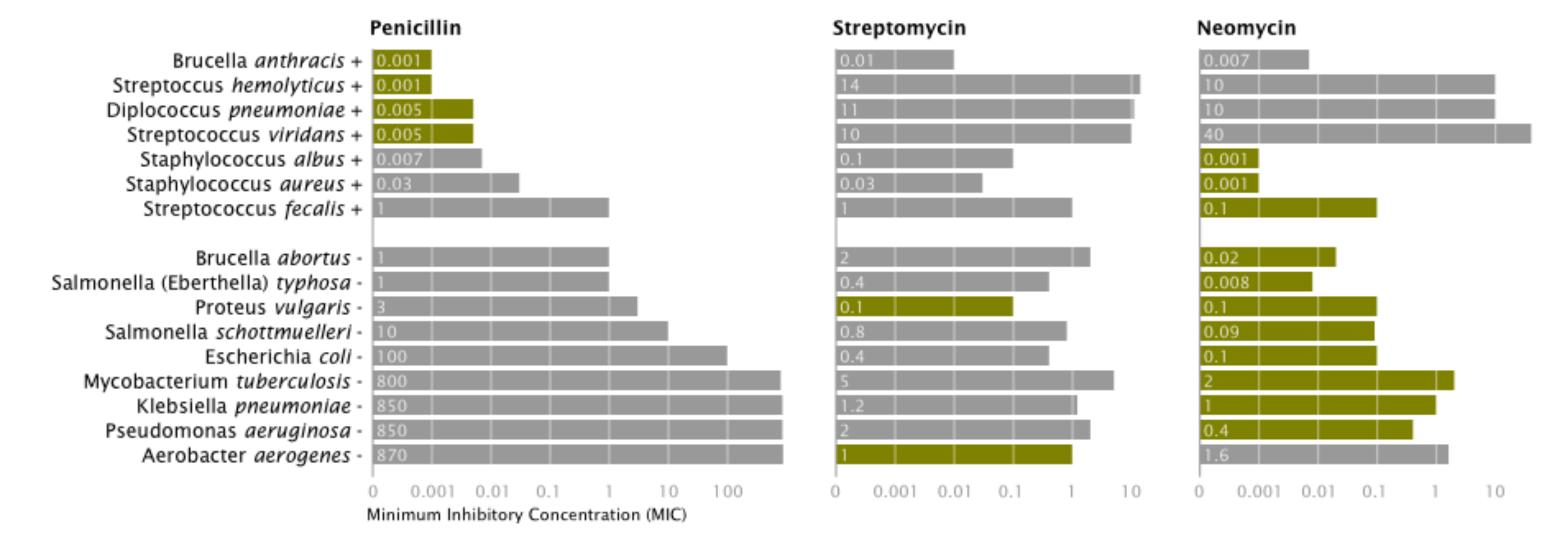
Background Color: Gram Staining

How do the drugs compare?



Mike Bostock
Stanford CS448B, Winter 2009

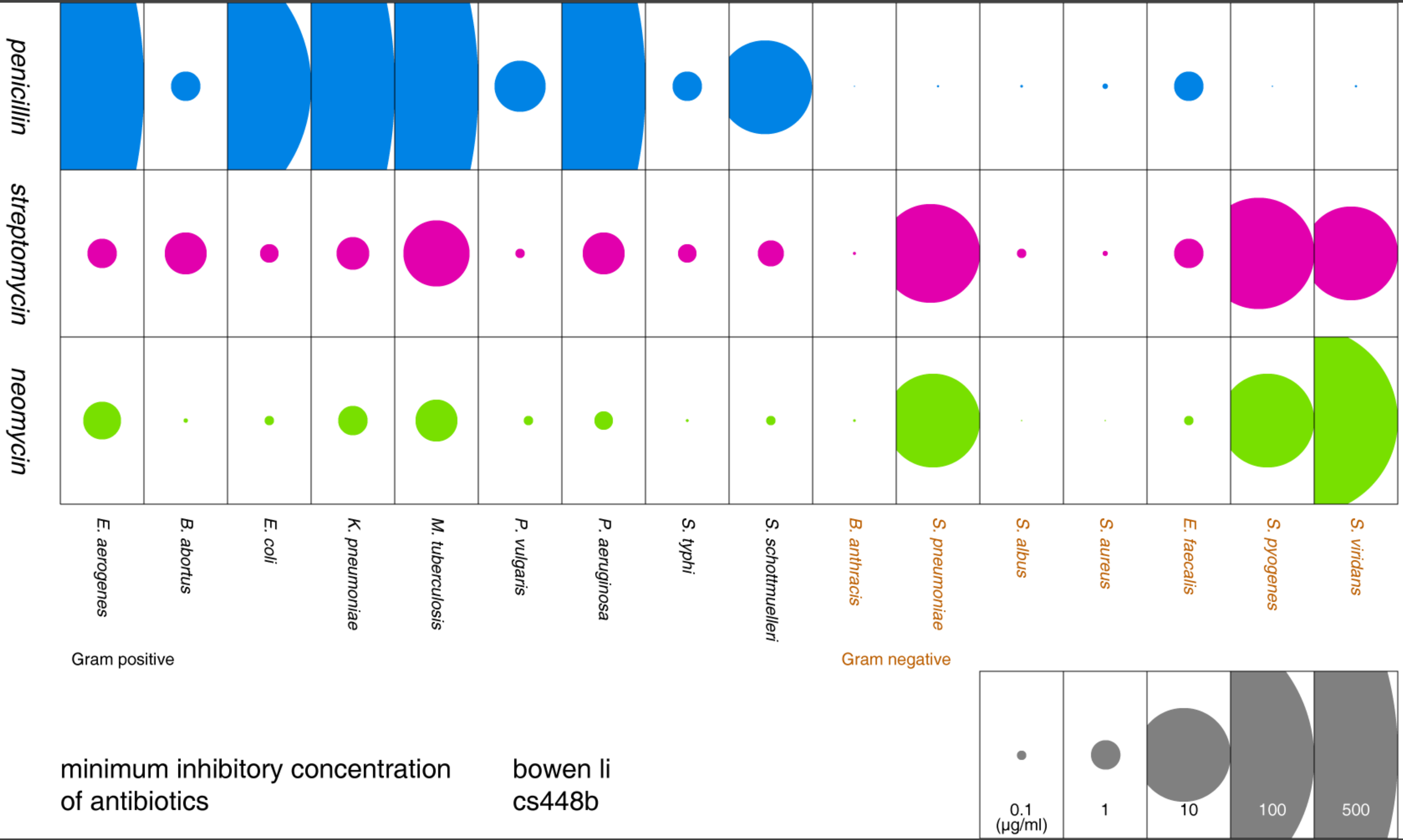
How do the drugs compare?



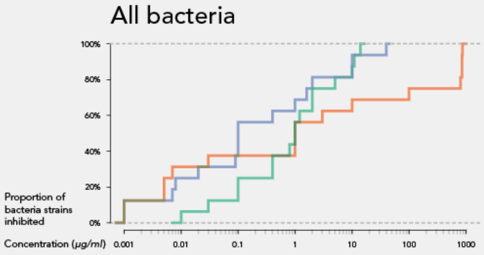
X-axis: Antibiotic | log(MIC)

Y-axis: Gram-Staining | Species

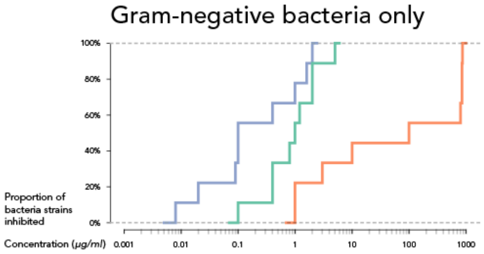
Color: Most-Effective?



Bowen Li
Stanford CS448B, Fall 2009

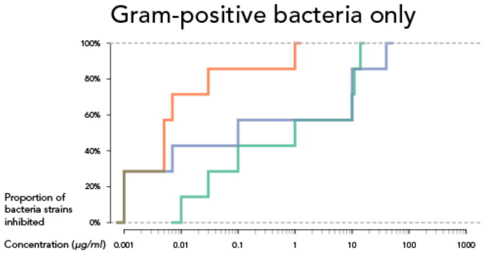


Streptomycin and Neomycin are more efficient broad-spectrum antibiotics than Penicilin.

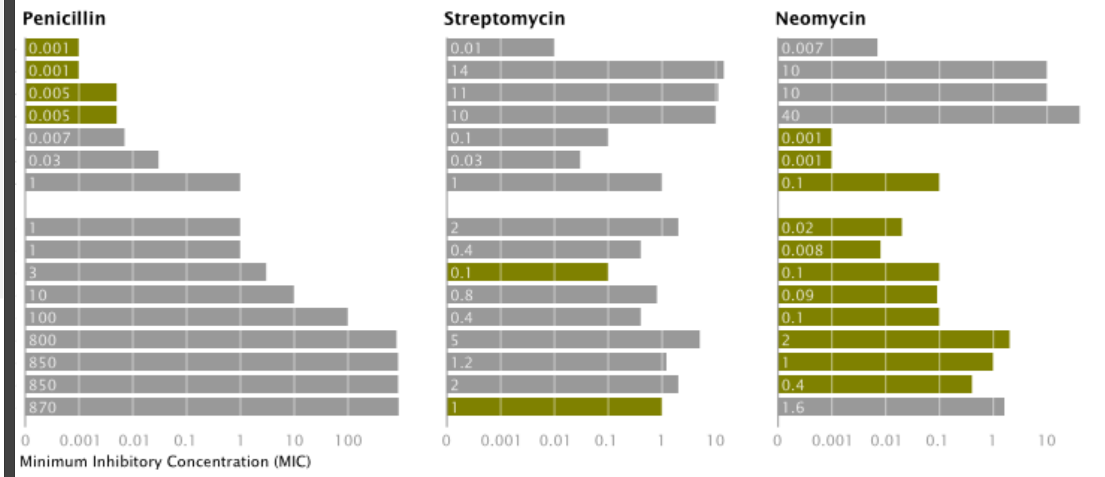


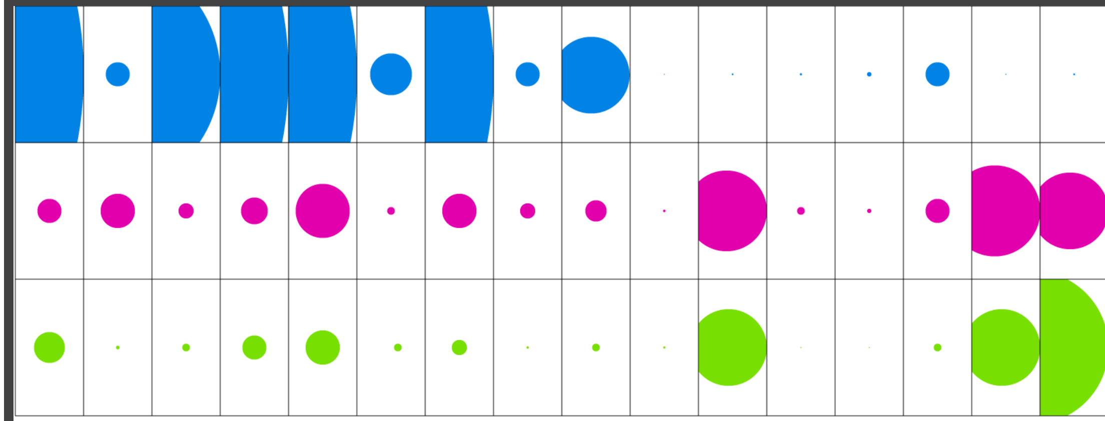
Neomycin and Streptomycin are more efficient against gram-negative bacteria, so can be used at a lower dosage here than above.

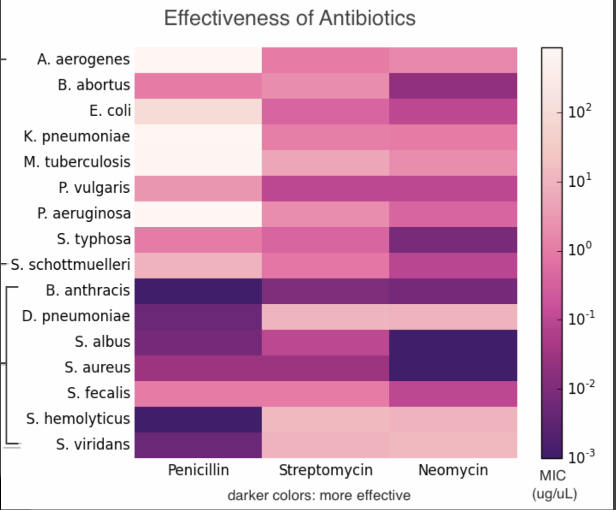
Gram staining quickly identifies bacteria as **Gram-negative** or **Gram-positive**, which can be used to find a more efficient antibiotic and dosage.

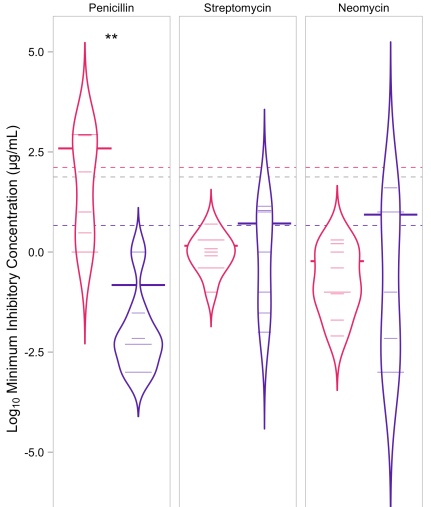


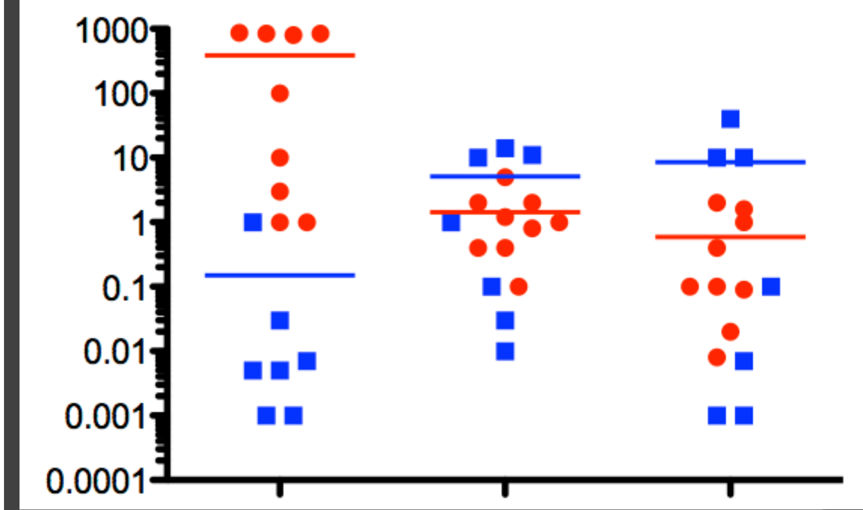
Penicilin is more efficient than either Streptomycin or Neomycin if the bacteria is known to be gram-positive.



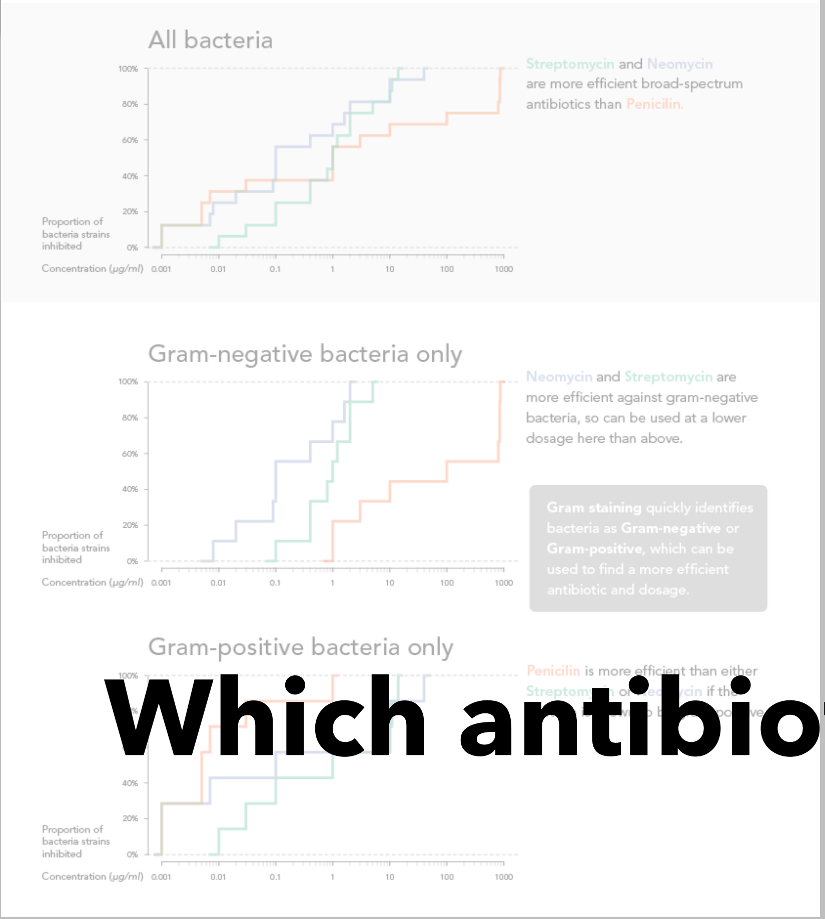








Which antibiotic should one use?

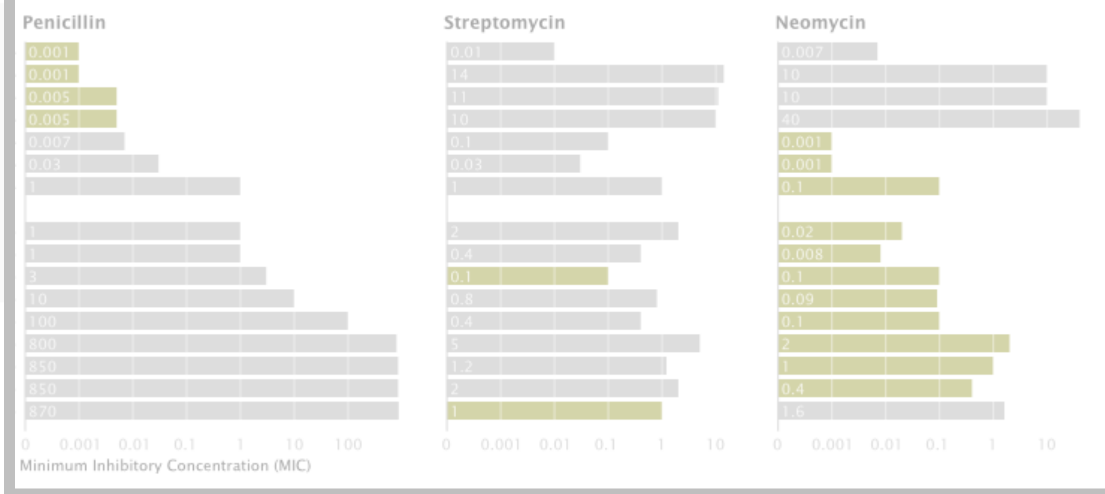


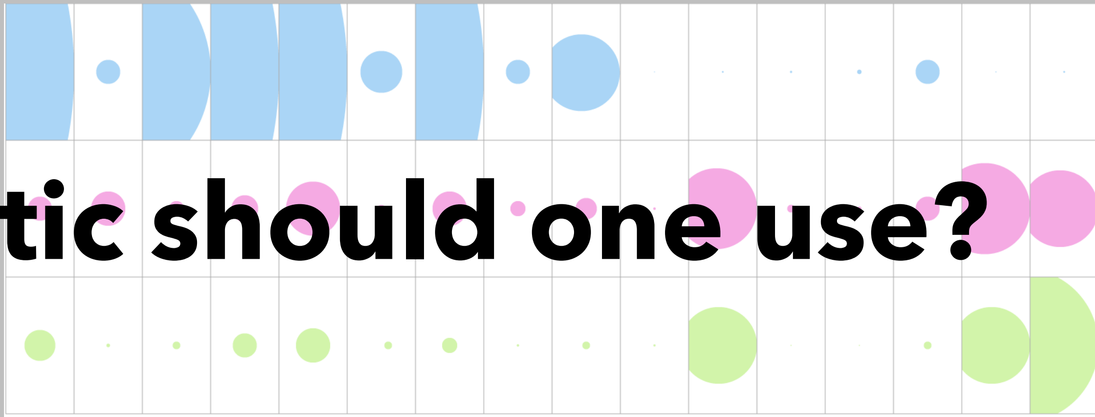
Streptomycin and Neomycin are more efficient broad-spectrum antibiotics than Penicilin.

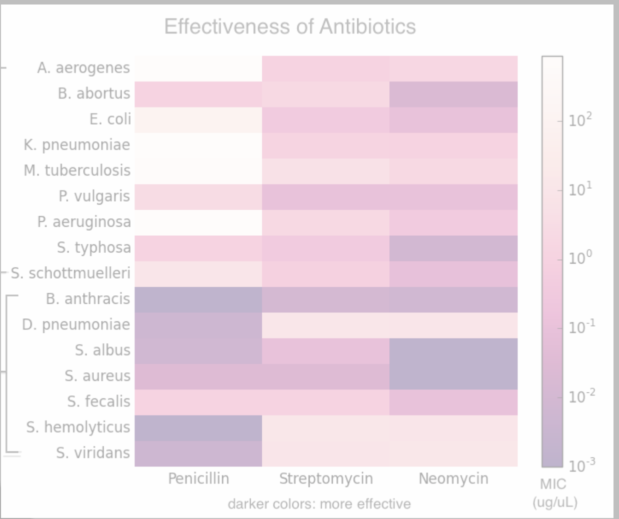
Neomycin and Streptomycin are more efficient against gram-negative bacteria, so can be used at a lower dosage here than above.

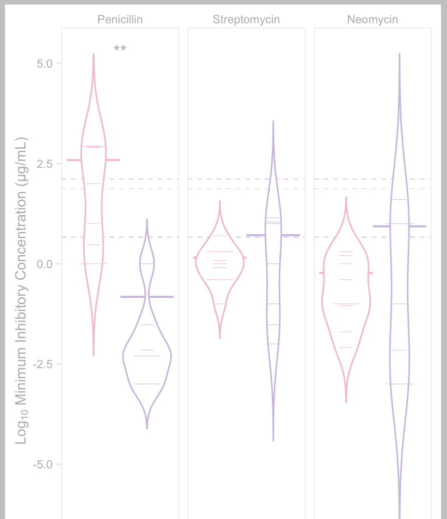
Gram staining quickly identifies bacteria as Gram-negative or Gram-positive, which can be used to find a more efficient antibiotic and dosage.

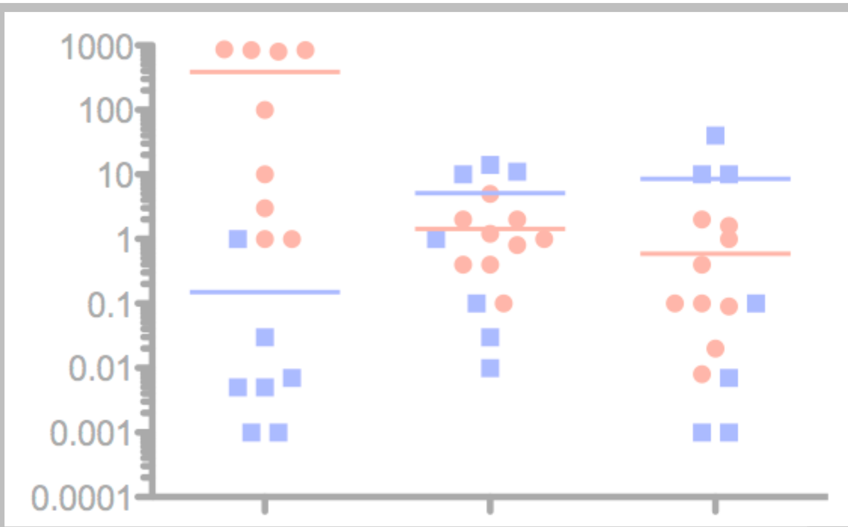
Penicilin is more efficient than either Streptom or eomycin if the



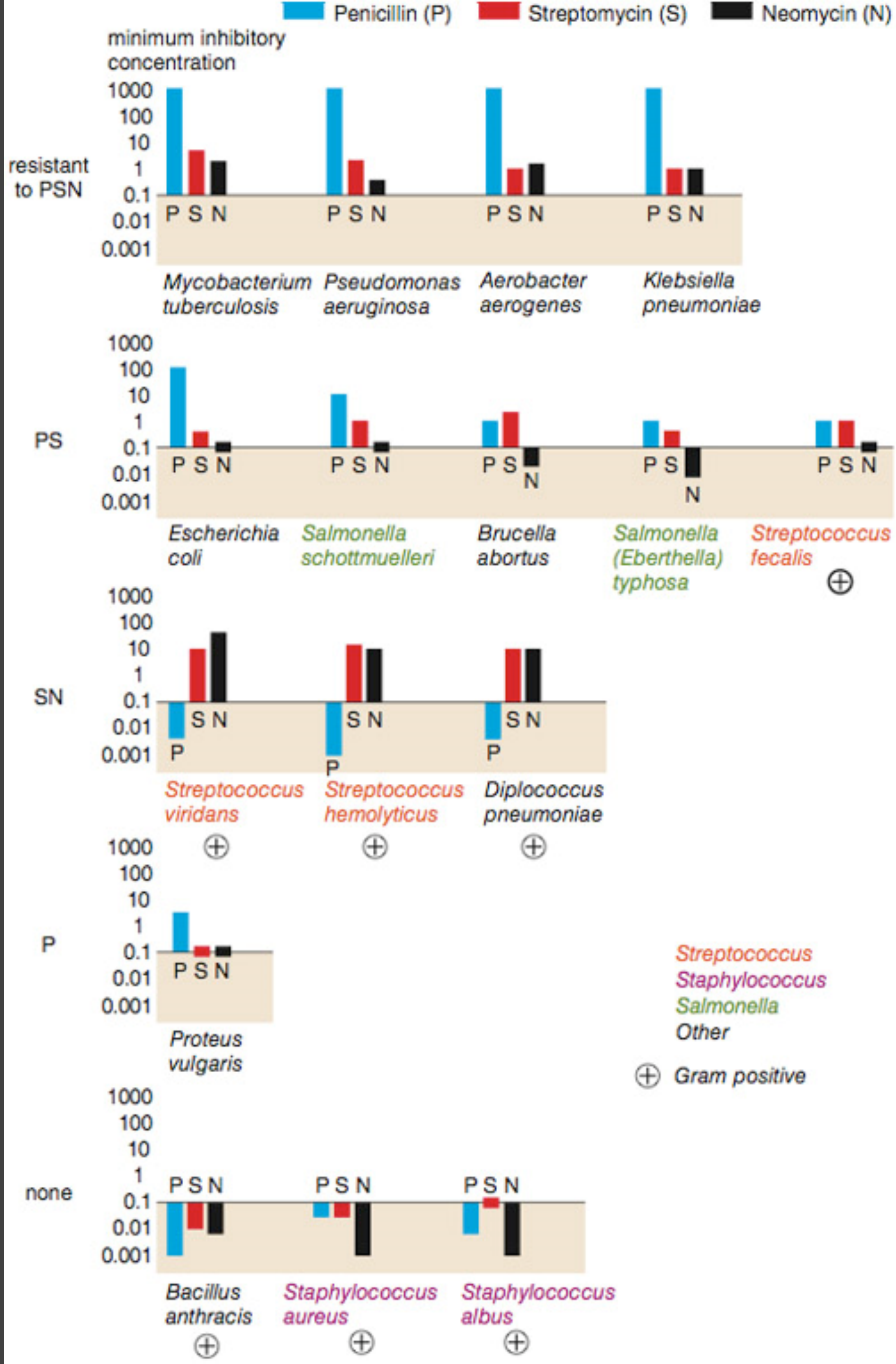






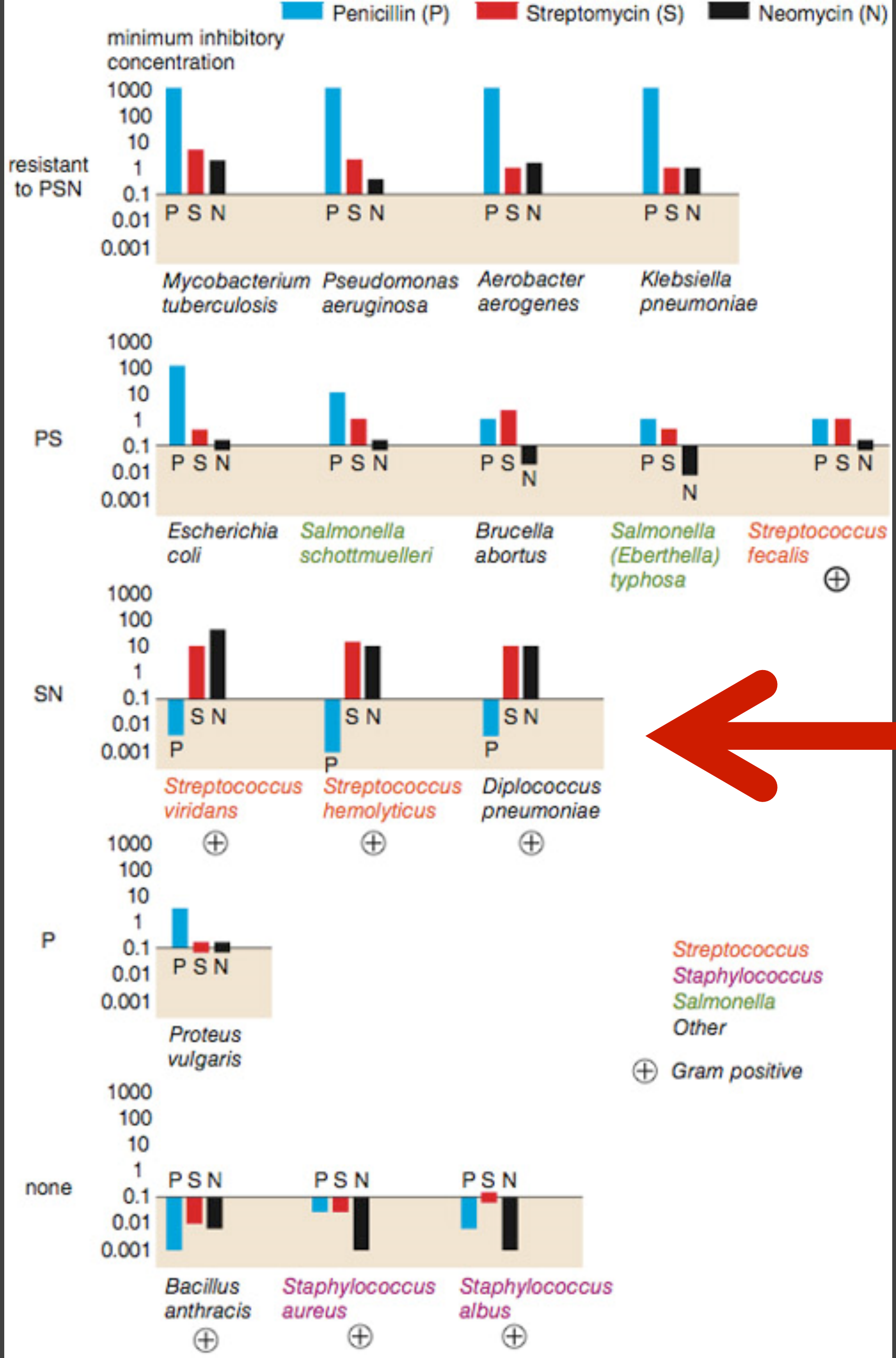


Do the bacteria group by antibiotic resistance?



Do the bacteria group by antibiotic resistance?

Wainer & Lysen
American Scientist, 2009

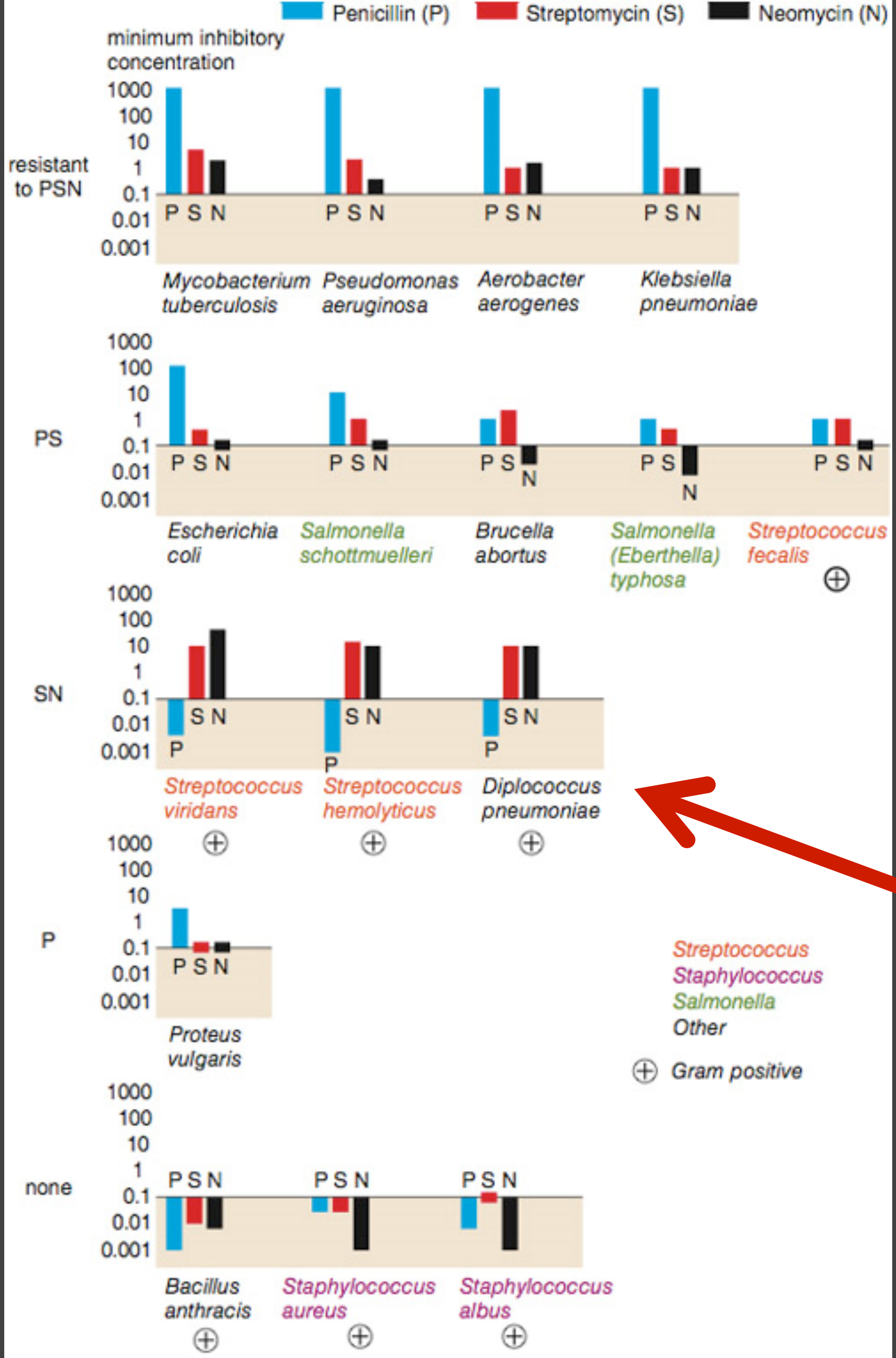


Do the bacteria group by antibiotic resistance?

Wainer & Lysen

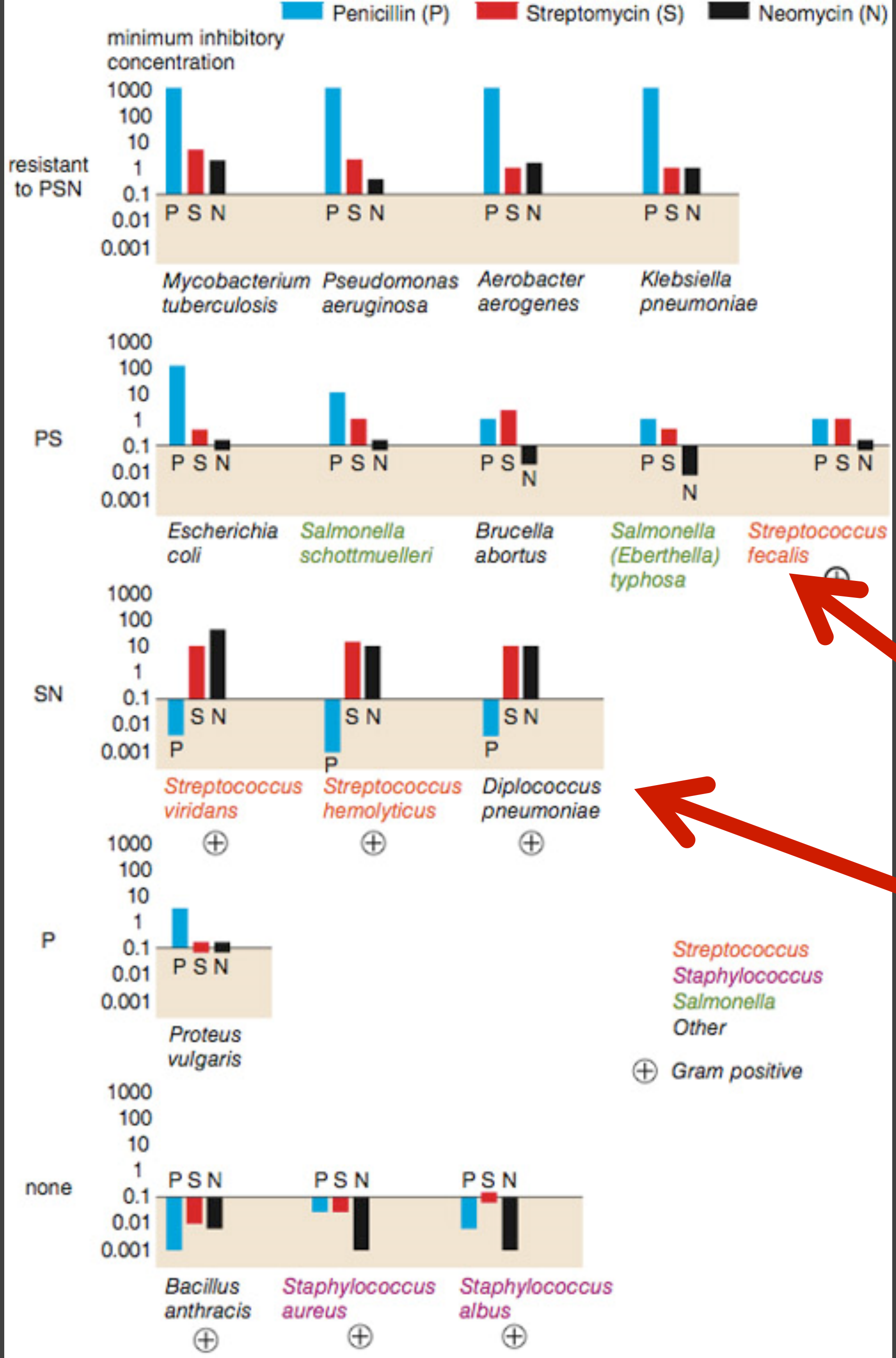
American Scientist, 2009

Do the bacteria group by antibiotic resistance?



Really a streptococcus!
(realized ~20 yrs later)

Wainer & Lysen
American Scientist, 2009



Do the bacteria group by antibiotic resistance?

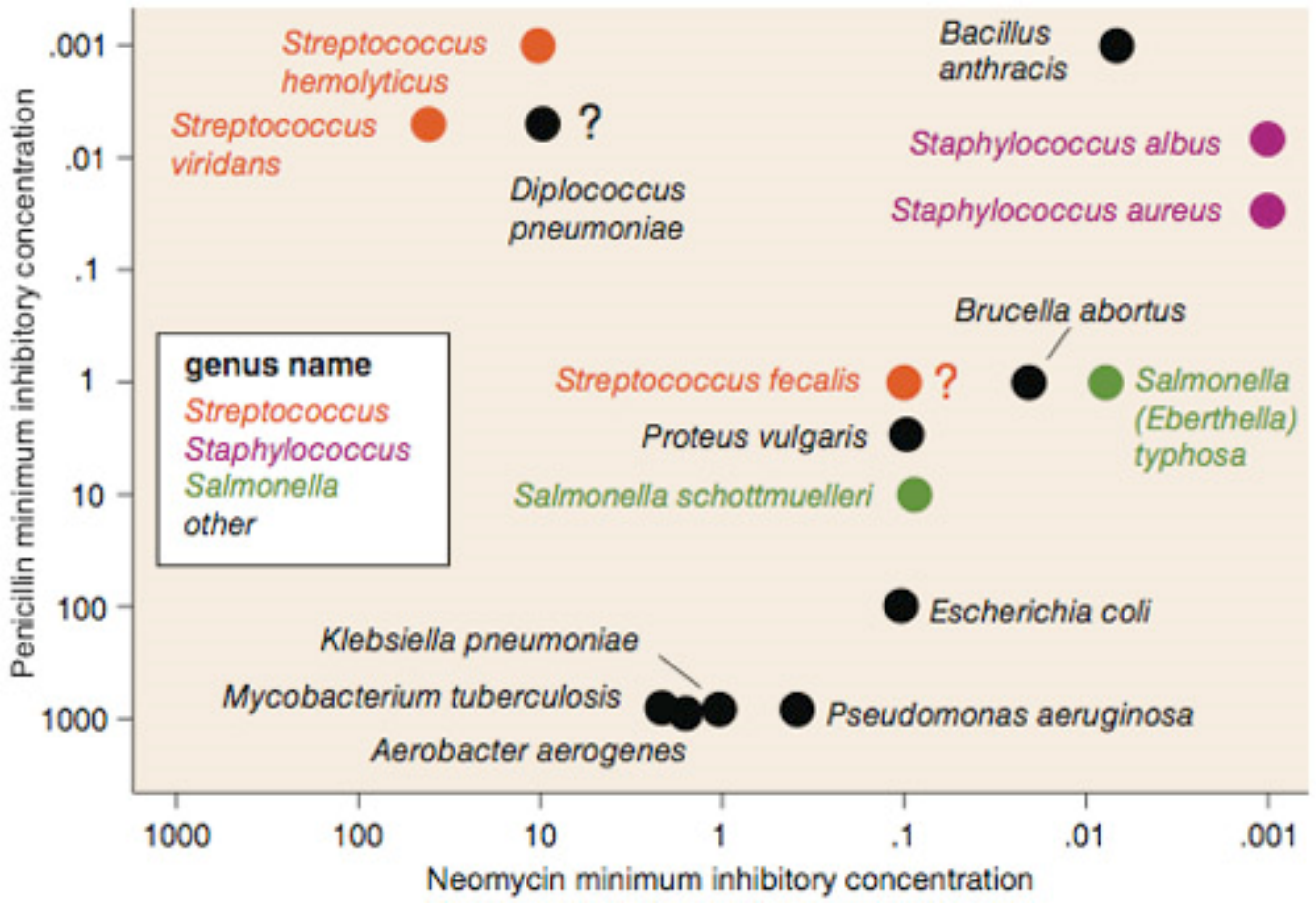
Not a streptococcus!
(realized ~30 yrs later)

Really a streptococcus!
(realized ~20 yrs later)

Wainer & Lysen
American Scientist, 2009

Do the bacteria group by resistance?

Do different drugs correlate?



Do the bacteria group by resistance?
Do different drugs correlate?

Wainer & Lysen
American Scientist, 2009

Lesson: Iterative Exploration

Exploratory Process

1 Construct graphics to address questions

2 Inspect “answer” and assess new questions

3 Repeat…

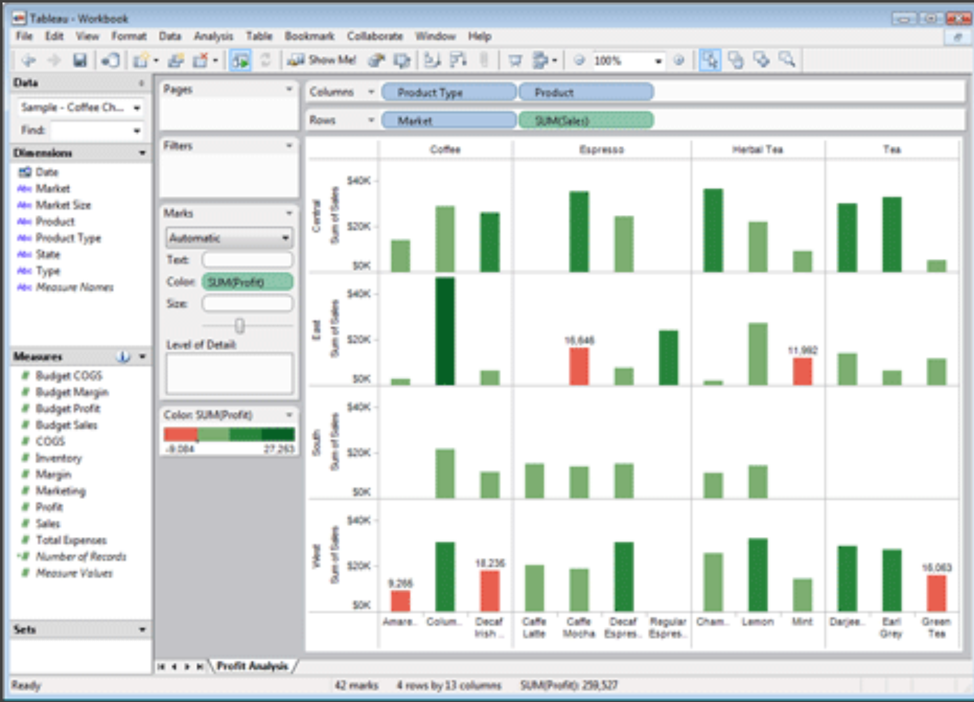
Transform data appropriately (e.g., invert, log)

Show data variation, not design variation [Tufte]

Administrivia

A2: Exploratory Data Analysis

Use visualization software to form & answer questions



Sample - Coffee Ch...

Columns: Product Type, Product

Rows: Market, SUM(Sales)

Automatic

Color: SUM(Profit)

Color: SUM(Profit)
-9,084 27,263

42 marks 4 rows by 13 columns SUM(Profit): 259,527

First steps:

Step 1: Pick domain & data

Step 2: Pose questions

Step 3: Profile the data

Iterate as needed

Create visualizations

Interact with data

Refine your questions

Author a report

Screenshots of most insightful views *(8+)*

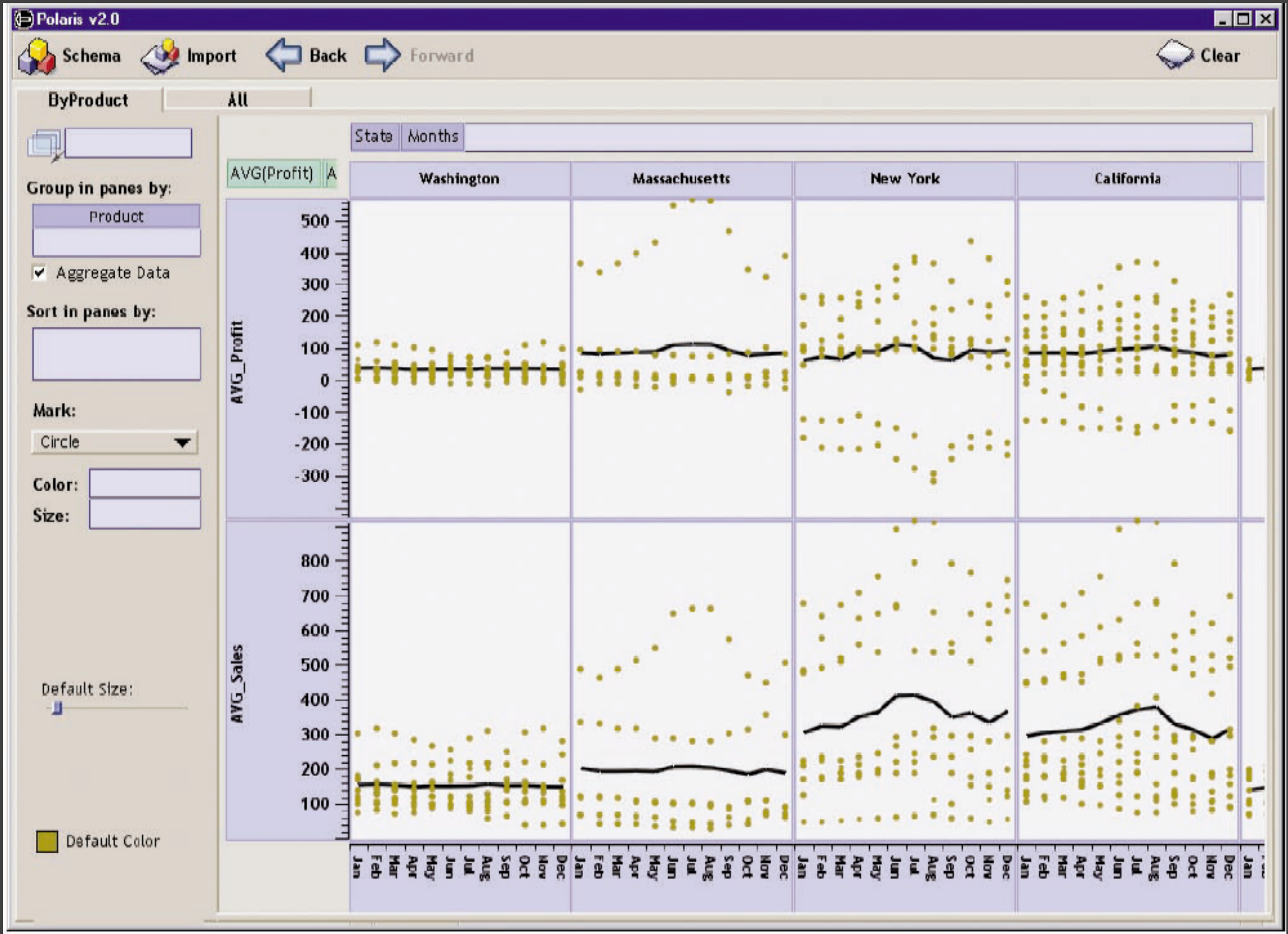
Include titles and captions for each view

Due by 11:59pm

Monday, Jan 25

Tableau / Polaris

Polaris [Stolte et al.]



Group in panes by: Product

☑ Aggregate Data

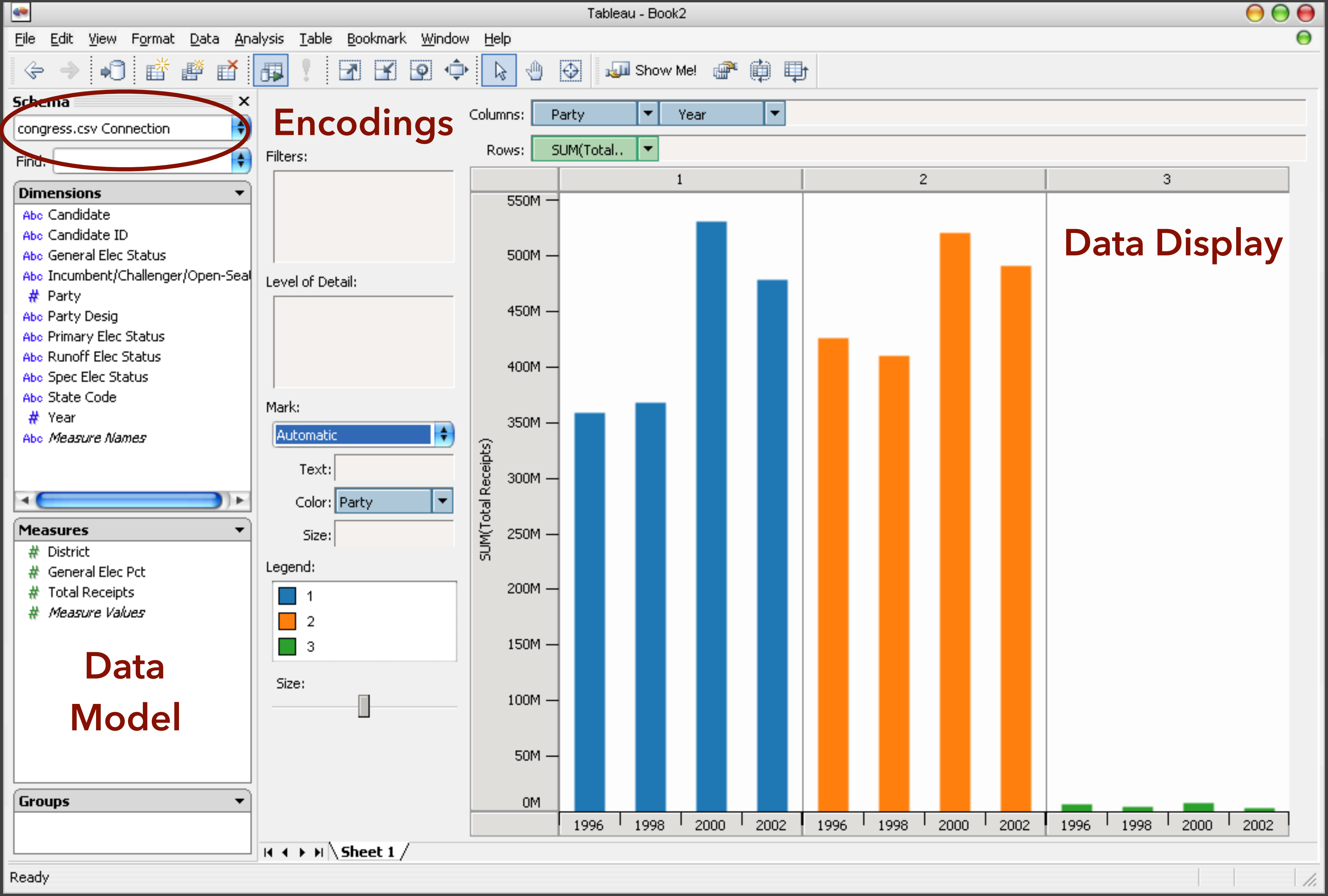
Sort in panes by:

Mark: Circle

Color:

Size:

Tableau



Encodings

Data Display

Data Model

Tableau - Book2

Schema: congress.csv Connection

Columns: Party, Year

Rows: SUM(Total..

Mark: Automatic

Color: Party

Dimensions: Candidate, Candidate ID, General Elec Status, Incumbent/Challenger/Open-Seat, Party, Party Desig, Primary Elec Status, Runoff Elec Status, Spec Elec Status, State Code, Year, *Measure Names*

Measures: District, General Elec Pct, Total Receipts, *Measure Values*

Legend:
- 🟦 1
- 🟧 2
- 🟩 3

Ready

Tableau / Polaris Approach

Insight: can simultaneously specify both database queries and visualization

Choose data, then visualization, not vice versa

Use smart defaults for visual encodings

Can also suggest encodings upon request

Tableau Demo

The dataset:

Federal Elections Commission Receipts (2012)

Every Congressional Candidate from 1996 to 2002

4 Election Cycles

9216 Candidacies

Dataset 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.

Hypotheses?

What might we learn from this data?

Hypotheses?

What might we learn from this data?

Correlation between receipts and winners?

Do receipts increase over time?

Which states spend the most?

Which party spends the most?

Margin of victory vs. amount spent?

Amount spent between competitors?

Tableau Demo

Specifying Table Configurations

Operands are the database fields
Each operand interpreted as a set {…}
Quantitative and Ordinal fields treated differently

Three operators:
concatenation (+)
cross product (x)
nest (/)

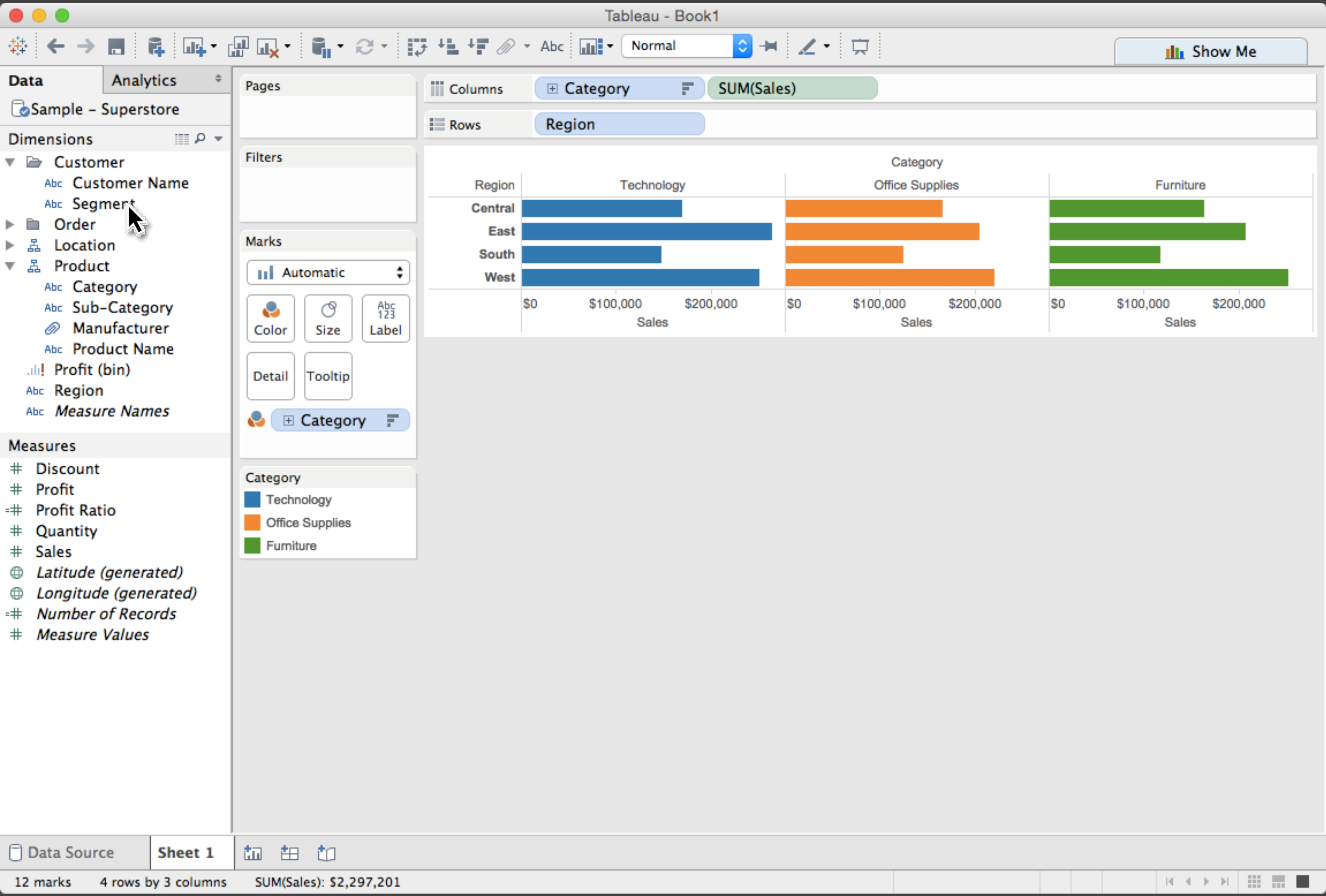


Tableau - Book1

Columns: Category, SUM(Sales)

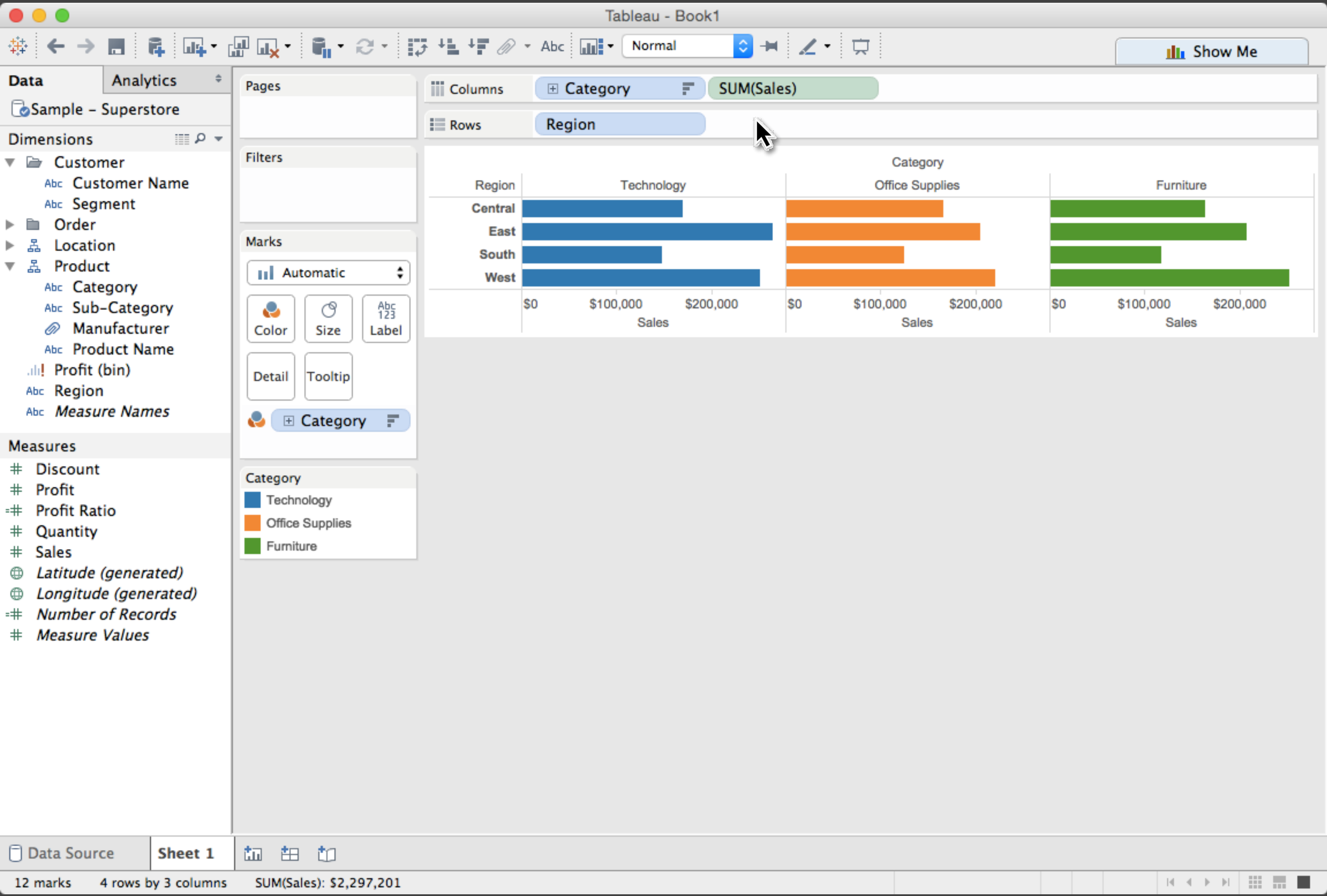
Rows: Region

Marks: Automatic

Color: Category

Category
- 🟦 Technology
- 🟧 Office Supplies
- 🟩 Furniture

12 marks 4 rows by 3 columns SUM(Sales): $2,297,201



Columns: Category, SUM(Sales)

Rows: Region

Marks: Automatic

Color: Category

Category
- 🟦 Technology
- 🟧 Office Supplies
- 🟩 Furniture

12 marks   4 rows by 3 columns   SUM(Sales): $2,297,201

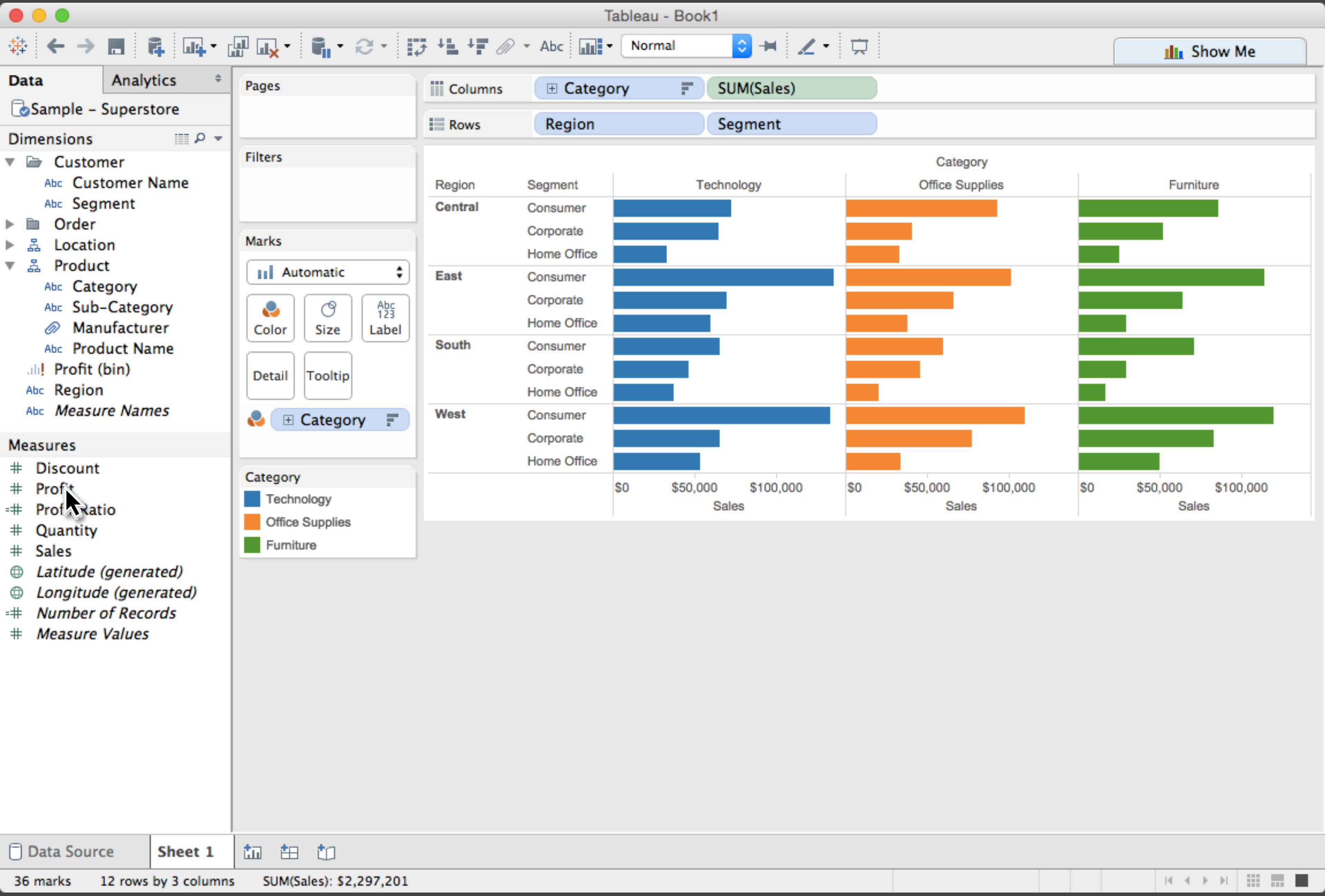


Tableau - Book1

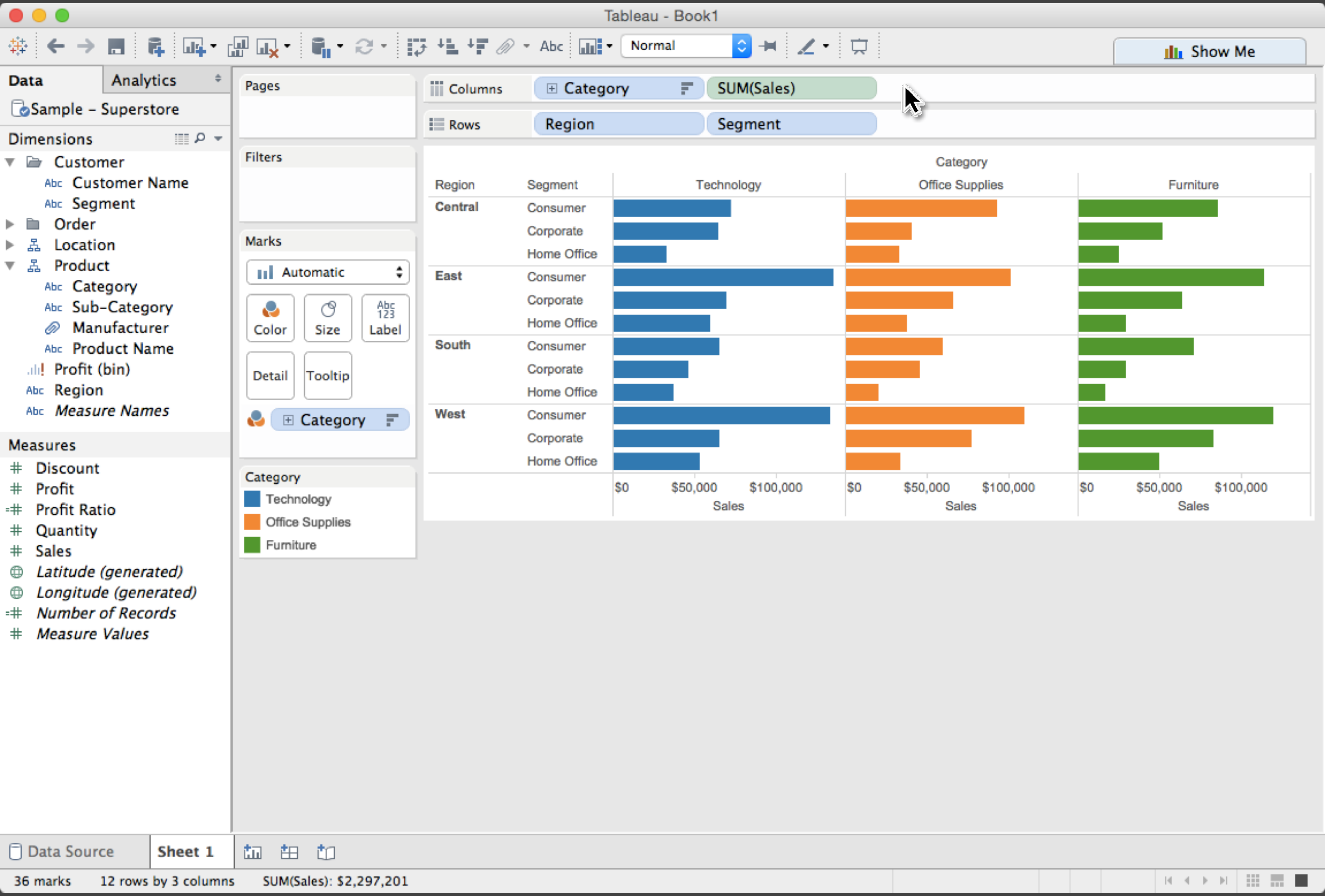
Columns: Category, SUM(Sales)

Rows: Region, Segment

Marks: Automatic

Color: Category

36 marks | 12 rows by 3 columns | SUM(Sales): $2,297,201



Columns: Category, SUM(Sales)

Rows: Region, Segment

36 marks 12 rows by 3 columns SUM(Sales): $2,297,201

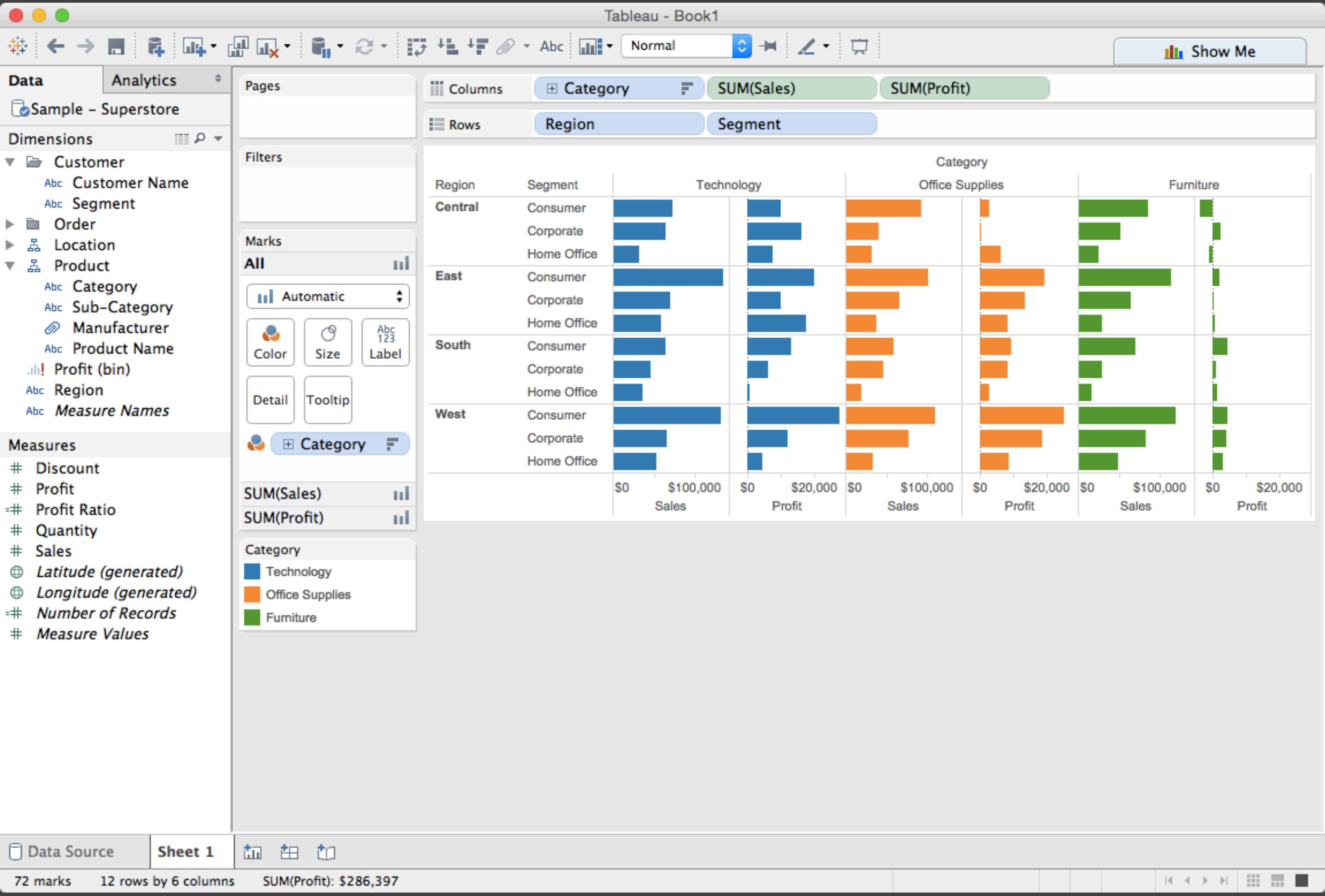


Tableau - Book1

Columns: Category | SUM(Sales) | SUM(Profit)

Rows: Region | Segment

Marks: All — Automatic

Color: Category

SUM(Sales)

SUM(Profit)

Category

🟦	Technology
🟧	Office Supplies
🟩	Furniture

72 marks   12 rows by 6 columns   SUM(Profit): $286,397

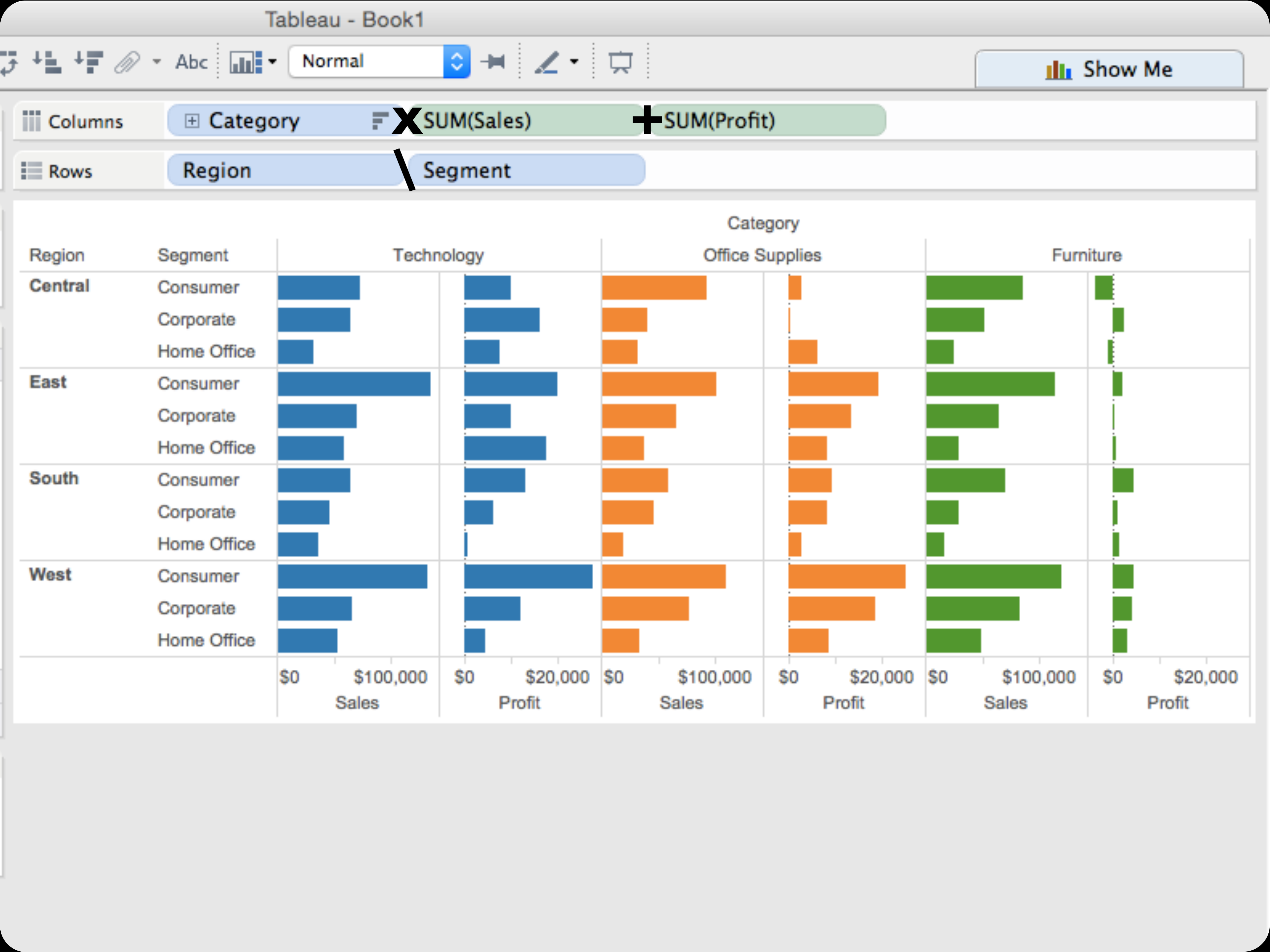


Tableau - Book1

Columns: Category X SUM(Sales) + SUM(Profit)

Rows: Region \ Segment

Table Algebra

The operators (+, x, /) and operands (O, Q) provide an *algebra* for tabular visualization.

Algebraic statements are then mapped to:

Visualizations - trellis plot partitions, visual encodings

Queries - selection, projection, group-by aggregation

In Tableau, users make statements via drag-and-drop

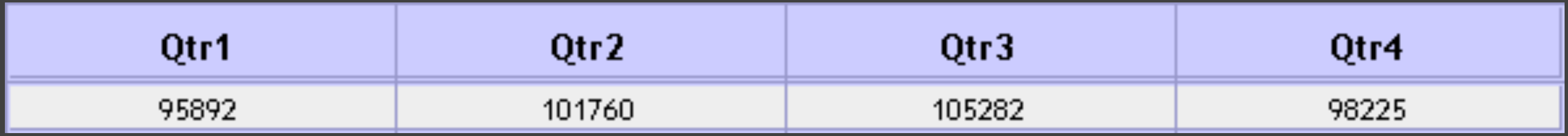
Note that this specifies operands *NOT* operators!

Operators are inferred by data type (O, Q)

Table Algebra: Operands

Ordinal fields: interpret domain as a set that partitions table into rows and columns.

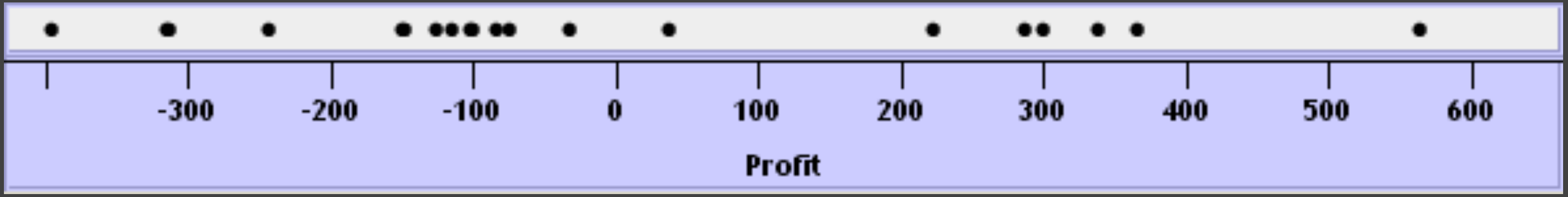
Quarter = {(Qtr1),(Qtr2),(Qtr3),(Qtr4)} ->



Qtr1	Qtr2	Qtr3	Qtr4
95892	101760	105282	98225

Quantitative fields: treat domain as single element set and encode spatially as axes.

Profit = {(Profit[-410,650])} ->



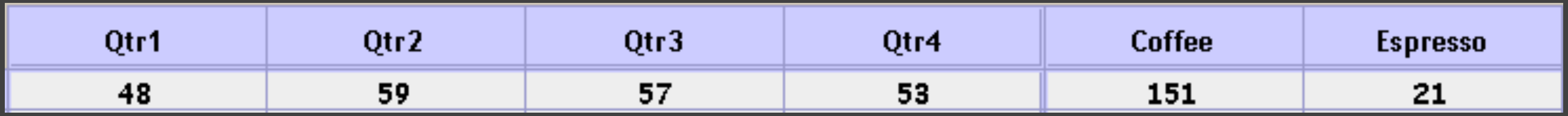
Concatenation (+) Operator

Ordered union of set interpretations

Quarter + Product Type

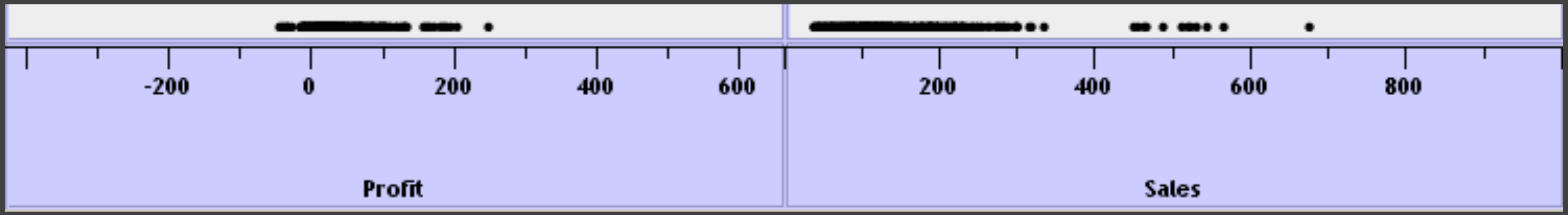
= {(Qtr1),(Qtr2),(Qtr3),(Qtr4)} + {(Coffee), (Espresso)}

= {(Qtr1),(Qtr2),(Qtr3),(Qtr4),(Coffee),(Espresso)}



Qtr1	Qtr2	Qtr3	Qtr4	Coffee	Espresso
48	59	57	53	151	21

Profit + Sales = {(Profit[-310,620]),(Sales[0,1000])}

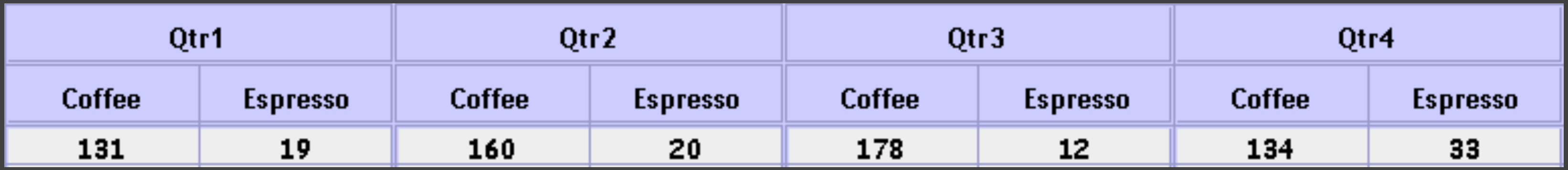


Cross (x) Operator

Cross-product of set interpretations

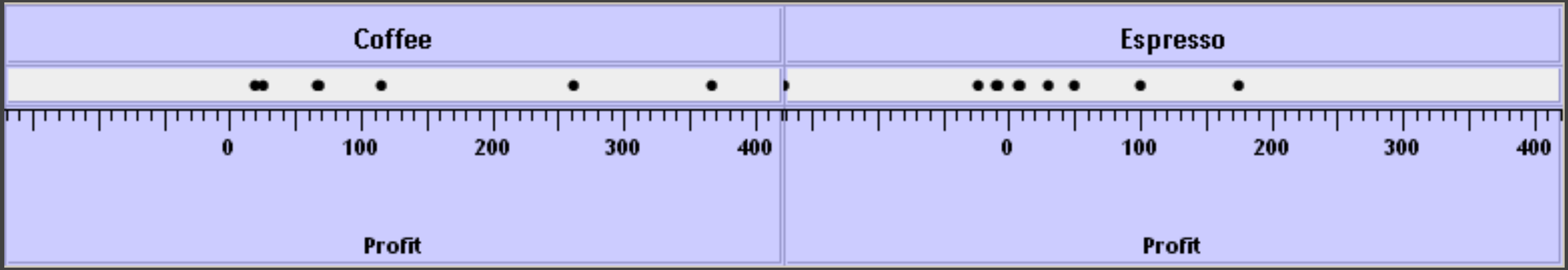
Quarter x Product Type =

{(Qtr1,Coffee), (Qtr1, Tea), (Qtr2, Coffee), (Qtr2, Tea), (Qtr3, Coffee), (Qtr3, Tea), (Qtr4, Coffee), (Qtr4,Tea)}



Qtr1		Qtr2		Qtr3		Qtr4	
Coffee	Espresso	Coffee	Espresso	Coffee	Espresso	Coffee	Espresso
131	19	160	20	178	12	134	33

Product Type x Profit =



Nest (/) Operator

Cross-product filtered by existing records

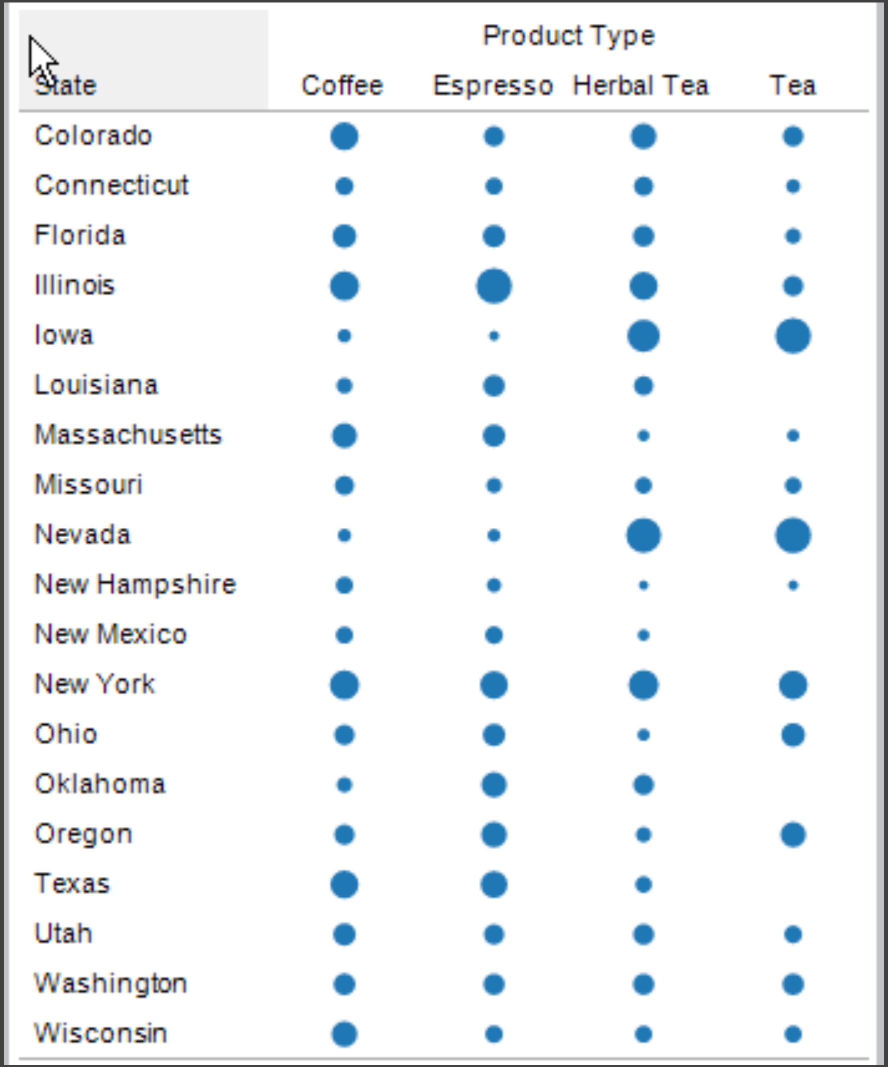
Quarter x Month ->

creates twelve entries for each quarter. i.e., (Qtr1, December)

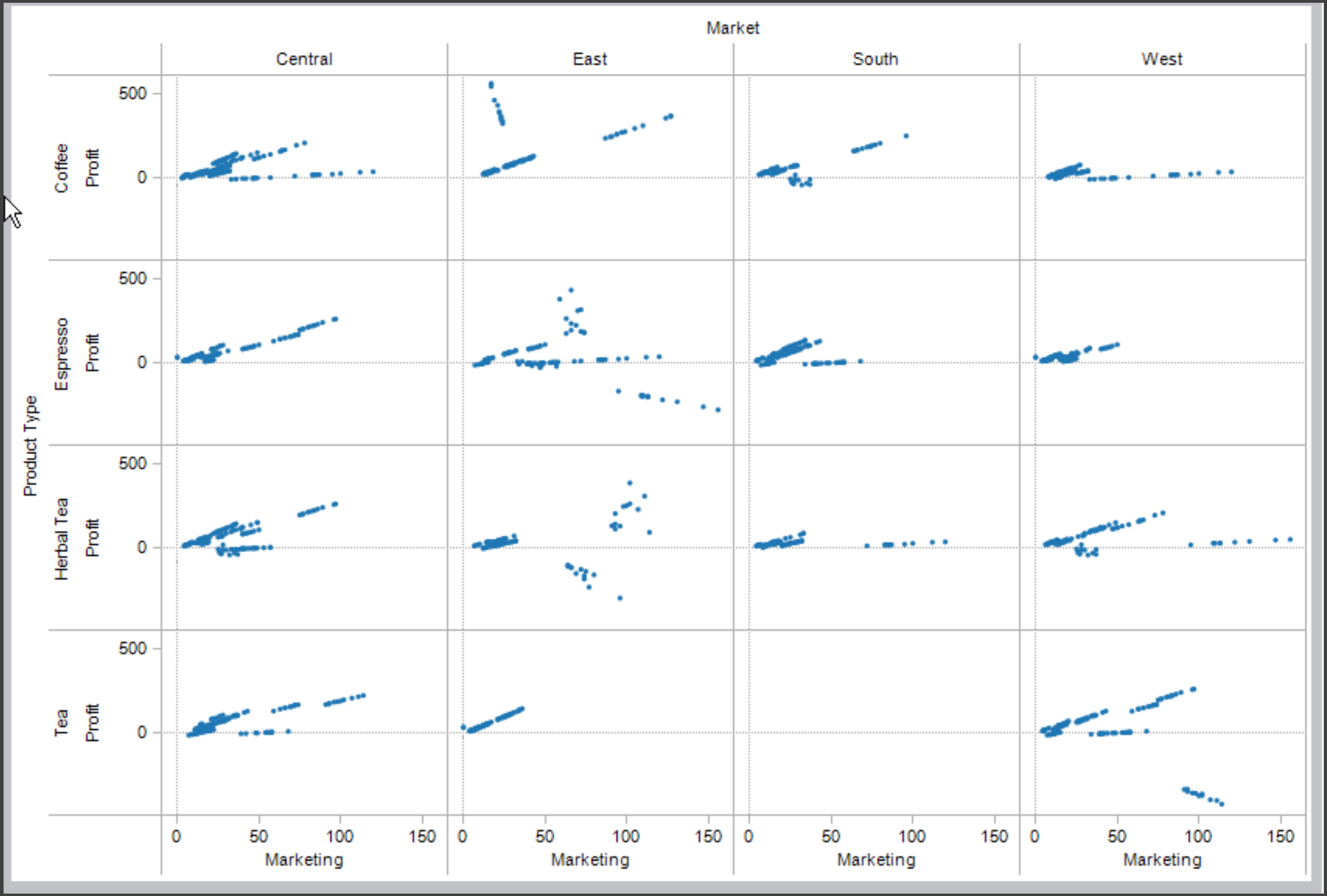
Quarter / Month ->

creates three entries per quarter based on tuples in database (not semantics)

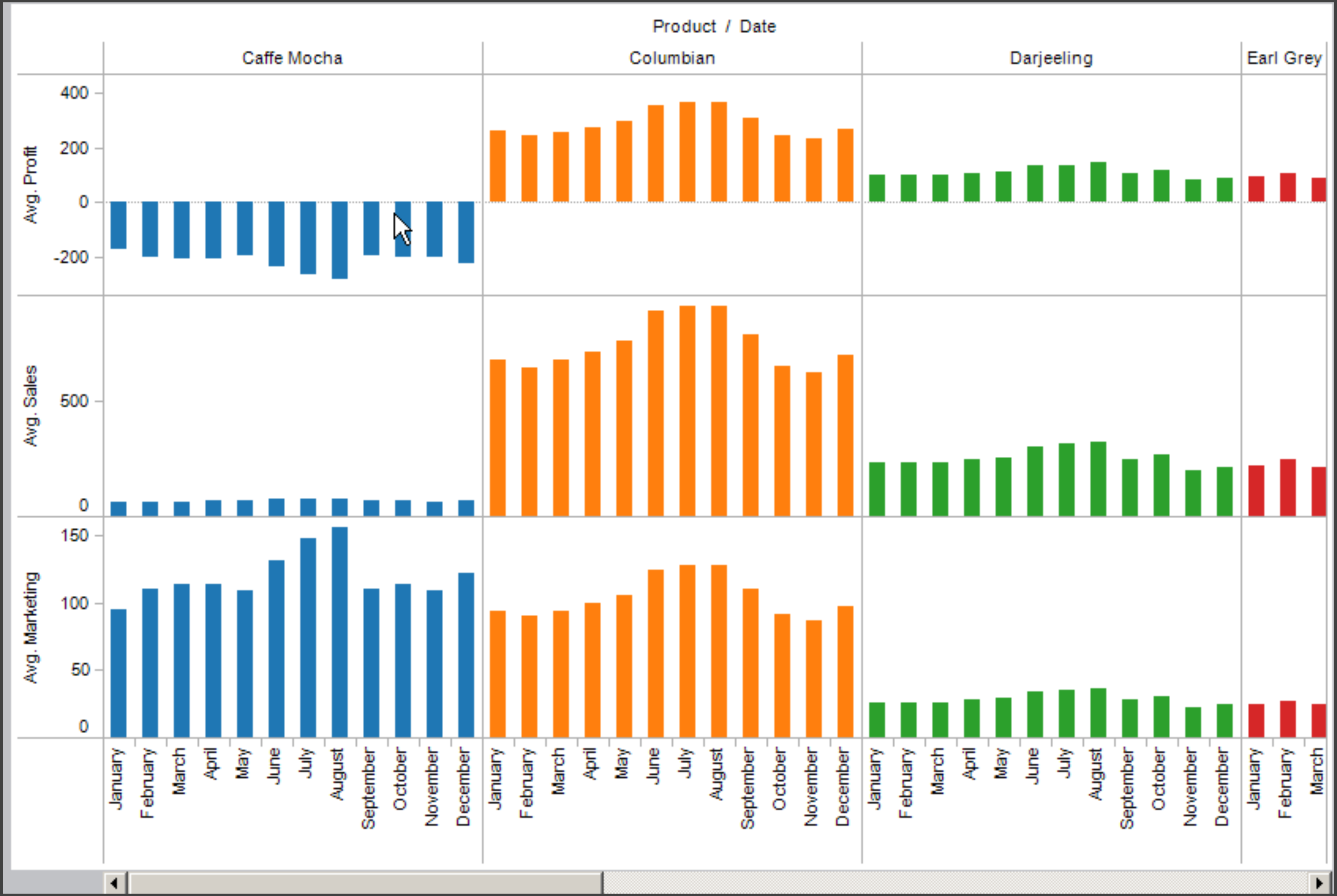
Ordinal-Ordinal



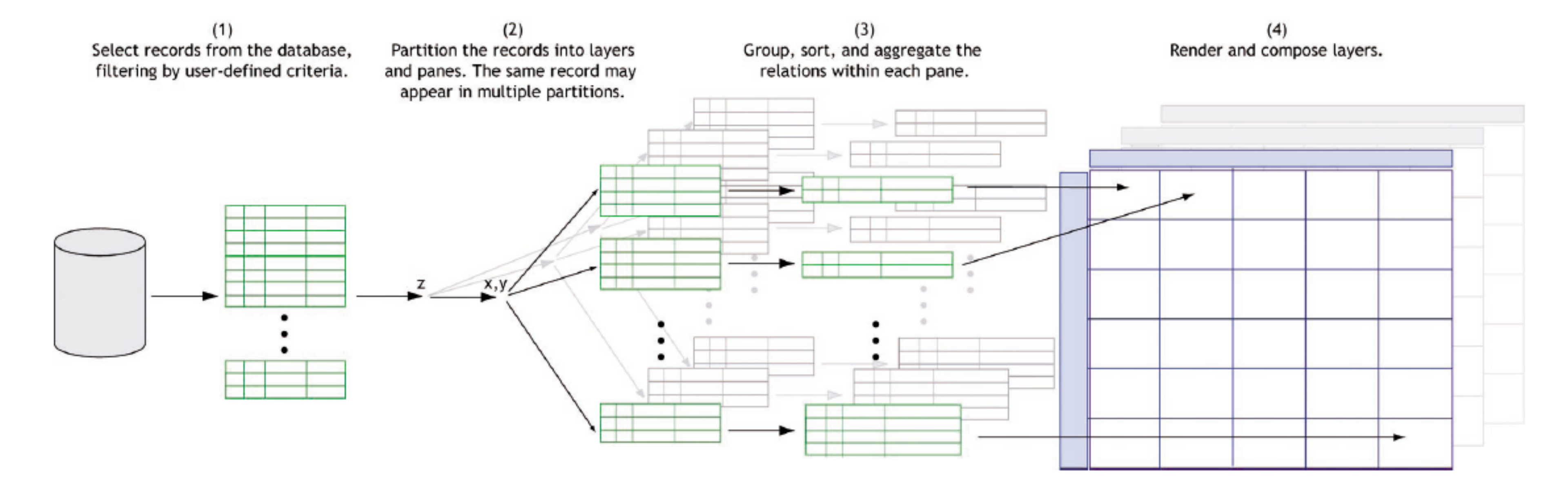
Quantitative-Quantitative



Ordinal-Quantitative



Querying the Database


(1)
Select records from the database, filtering by user-defined criteria.
(2)
Partition the records into layers and panes. The same record may appear in multiple partitions.
(3)
Group, sort, and aggregate the relations within each pane.
(4)
Render and compose layers.
z
x,y

Quiz Section: Tableau

Tomorrow, Thursday January 14th

Introduction and hands-on experience in Tableau
Come prepared with Tableau installed
See announcement on Ed for instructions

Up Next: Jane's Office Hour (link on Canvas)