Multidimensional Visualization

Maneesh Agrawala

CS 294-10: Visualization
Spring 2011

Last Time:
Exploratory Data Analysis

Topics

- Exploratory Data Analysis
  - Data Diagnostics
  - Graphical Methods
  - Data Transformation
- Confirmatory Data Analysis
  - Statistical Hypothesis Testing

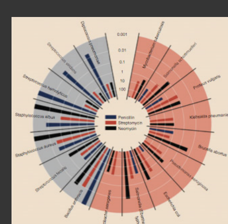
Exploratory Analysis:
Effectiveness of Antibiotics

What questions might we ask?

Table 1: Burtin's data.

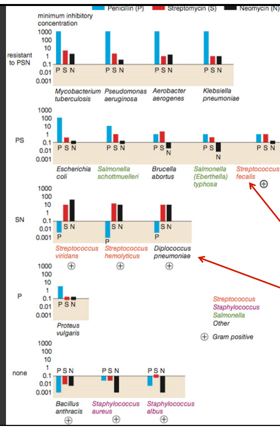
Bacteria	Antibiotic			Gram Staining
	Penicillin	Streptomycin	Neomycin	
Aerobacter *aerogenes*	870	1	1.6	negative
Brucella *abortus*	1	2	0.02	negative
Brucella *anthracis*	0.001	0.01	0.007	positive
Diplococcus *pneumoniae*	0.005	11	10	positive
Escherichia *coli*	100	0.4	0.1	negative
Klebsiella *pneumoniae*	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

Will Burtin, 1951



Bacteria	Antibiotic			Gram stain
	Penicillin	Streptomycin	Neomycin	
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?



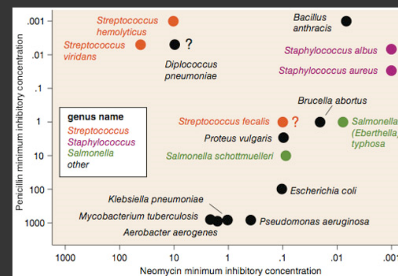
How do the bacteria group with respect to antibiotic resistance?

Not a streptococcus! (realized ~30 yrs later)

Really a streptococcus! (realized ~20 yrs later)

Wainer & Lysen

American Scientist, 2009



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

Wainer & Lysen

American Scientist, 2009

Common Data Transformations

Normalize	$y_i / \Sigma_i\, y_i$	(among others)
Log	log y	
Power	$y^{1/k}$	
Box-Cox Transform	$(y^\lambda - 1) / \lambda$	if $\lambda \neq 0$
	log y	if $\lambda = 0$
Binning	e.g., histograms	
Grouping	e.g., merge categories	

Often performed to aid comparison (% or scale difference) or better approx. normal distribution

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

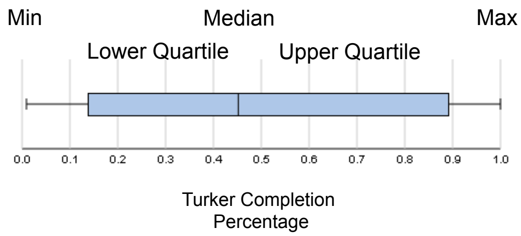
Exploratory Analysis: Participation on Amazon's Mechanical Turk

The Data Set (~200 rows)

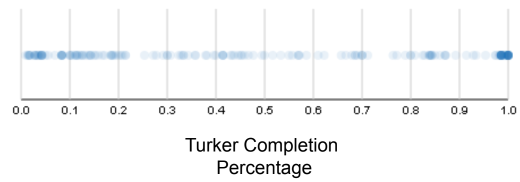
Turker ID	String
Avg. Completion Rate	Number [0,1]

Collected in 2009 by Heer & Bostock.

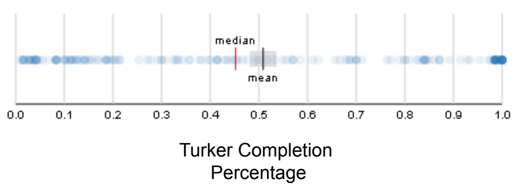
What questions might we ask of the data?
What charts might provide insight?



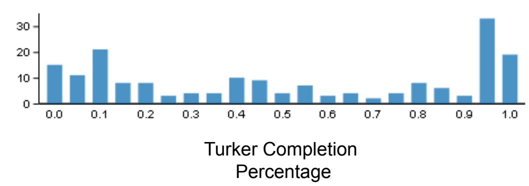
Box (and Whiskers) Plot



Dot Plot (with transparency to indicate overlap)



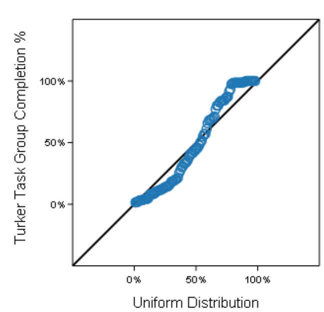
Dot Plot w/ Reference Lines



Histogram (binned counts)

```
 0 | 1 1 1 2 2 2 2 3 3 3 3 3 3 4 4 4 4 4 4 4 4 4 5 6 7 8 8 8 8 8 8 9
 1 | 0 0 0 0 1 1 1 1 2 2 3 3 3 3 4 4 4 4 5 5 6 7 7 8 9 9 9 9 9
 2 | 0 0 1 1 1 5 7 8 9
 3 | 0 0 1 2 3 3 3 4 6 6 8 8
 4 | 0 0 1 1 1 1 3 3 4 5 5 5 6 7 8 9
 5 | 0 2 3 5 6 7 7 7 9
 6 | 1 2 6 7 8 9 9 9
 7 | 0 0 0 1 6 7 9
 8 | 0 0 1 2 3 4 4 4 4 4 4 4 5 6 7 7 7 9
 9 | 1 3 3 5 7 8 8 8 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9
10 | 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
```

Stem-and-Leaf Plot

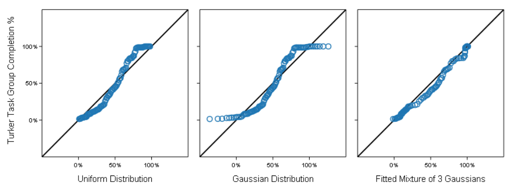


Used to compare two distributions; in this case, one actual and one theoretical.

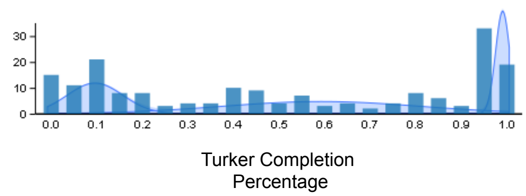
Plots the quantiles (here, the percentile values) against each other.

Similar distributions lie along the diagonal. If linearly related, values will lie along a line, but with potentially varying slope and intercept.

Quantile-Quantile Plot



Quantile-Quantile Plots



Histogram + Fitted Mixture of 3 Gaussians

Lessons

Even for "simple" data, a variety of graphics might provide insight. Again, tailor the choice of graphic to the questions being asked, but be open to surprises.

Graphics can be used to understand and help assess the quality of statistical models.

Premature commitment to a model and lack of verification can lead an analysis astray.

Confirmatory Data Analysis

Some Uses of Formal Statistics

What is the probability that the pattern I'm seeing might have arisen by chance?

With what parameters does the data best fit a given function? What is the goodness of fit?

How well do one (or more) data variables predict another?

…and many others.

Example: Heights by Gender

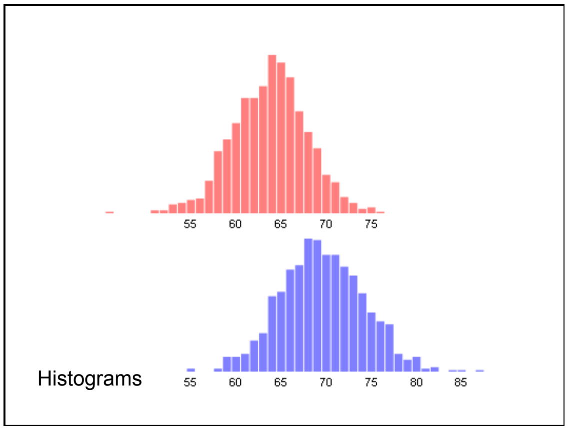
Gender	Male / Female
Height (in)	Number

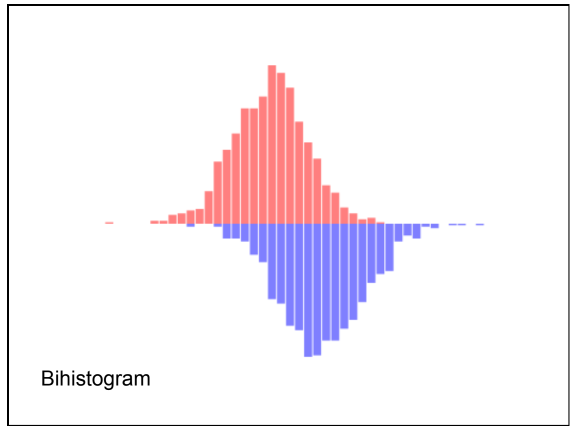
$\mu_m = 69.4$ $\sigma_m = 4.69$ $N_m = 1000$

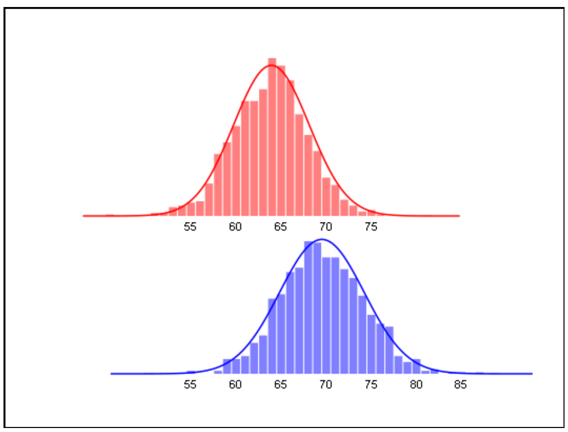
$\mu_f = 63.8$ $\sigma_f = 4.18$ $N_f = 1000$

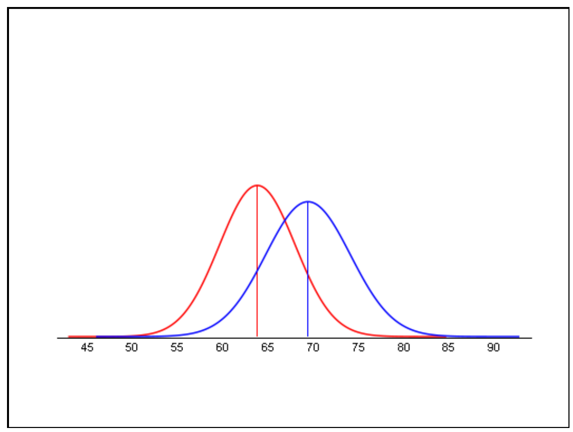
Is this difference in heights significant?

In other words: assuming no true difference, what is the prob. that our data is due to chance?









Formulating a Hypothesis

Null Hypothesis (H_0): $\mu_m = \mu_f$ (population)
Alternate Hypothesis (H_a): $\mu_m \neq \mu_f$ (population)

A statistical hypothesis test assesses the likelihood of the null hypothesis.

What is the probability of sampling the observed data assuming population means are equal?

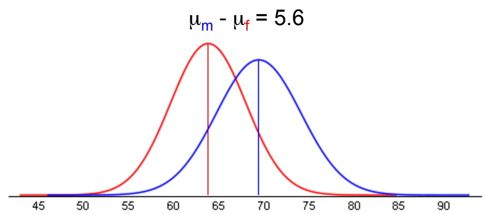
This is called the *p* value.

Testing Procedure

Compute a test statistic. This is a number that in essence summarizes the difference.

Compute test statistic

$$Z = \frac{\mu_m - \mu_f}{\sqrt{\sigma^2_m / N_m + \sigma^2_f / N_f}}$$



Testing Procedure

Compute a test statistic. This is a number that in essence summarizes the difference.

The possible values of this statistic come from a known probability distribution.

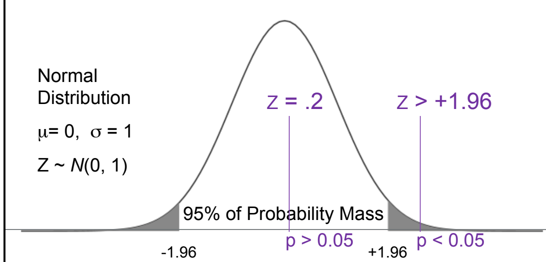
According to this distribution, look up the probability of seeing a value meeting or exceeding the test statistic. This is the *p* value.

Lookup probability of test statistic

Normal Distribution

$\mu = 0, \sigma = 1$

$Z \sim N(0, 1)$



Statistical Significance

The threshold at which we consider it safe (or reasonable?) to *reject the null hypothesis*.

If $p < 0.05$, we typically say that the observed effect or difference is statistically significant.

This means that there is a less than 5% chance that the observed data is due to chance.

Note that the choice of 0.05 is a somewhat arbitrary threshold (chosen by R. A. Fisher)

Common Statistical Methods

Question | Data Type | Parametric | Non-Parametric

Parametric: Assumes a particular distribution for the data -- usually normal, a.k.a. Gaussian.

Non-Parametric: Does not assume a distribution. Typically works on rank orders.

Common Statistical Methods

Question	Data Type	Parametric	Non-Parametric
Do data distributions have different "centers"? (aka "location" tests)	2 uni. dists > 2 uni. dists > 2 multi. dists	t-Test ANOVA MANOVA	Mann-Whitney U Kruskal-Wallis Median Test
Are observed counts significantly different?	Counts in categories		χ^2 (chi-squared)
Are two vars related?	2 variables	Pearson coeff.	Rank correl.
Do 1 (or more) variables predict another?	Continuous Binary	Linear regression Logistic regression	

Summary

Exploratory analysis may combine graphical methods, data transformations, and statistics.

Use questions to uncover more questions.

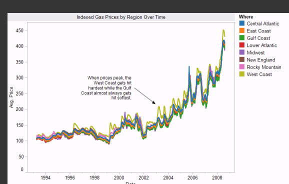
Formal methods may be used to confirm, sometimes on held-out or new data.

Announcements

Assignment 2: Exploratory Data Analysis

Use existing software to formulate & answer questions

First steps
- Step 1: Pick a domain
- Step 2: Pose questions
- Step 3: Profile data
- Iterate



Create visualizations
- Interact with data
- Refine your questions
- Tableau

Make wiki notebook
- Keep record of all steps you took to answer the questions

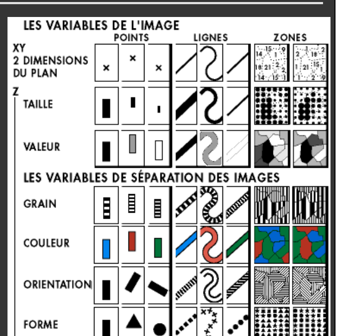
Due before class on Feb 14, 2011

Multidimensional Visualization

Visual Encoding Variables

Position
Length
Area
Volume
Value
Texture
Color
Orientation
Shape

~8 dimensions?



Small Multiples [from Wills 95]

how long in majors

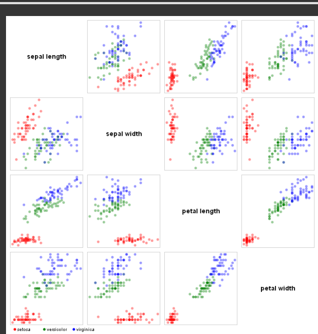
avg assists vs avg putouts (fielding ability)

distribution of positions played

select high salaries

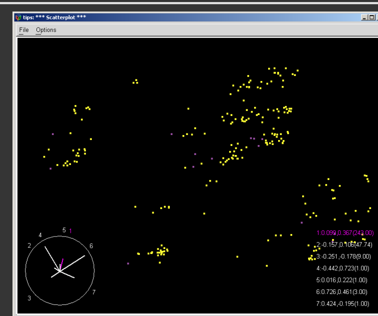
avg career HRs vs avg career hits (batting ability)

Scatterplot Matrix (SPLOM)

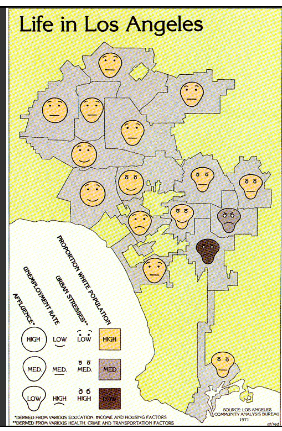


Scatter plots enabling pair-wise comparison of each data dimension.

Dimensional Projection



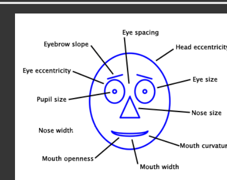
http://www.ggobi.org/



Chernoff Faces (1973)

Insight: We have evolved a sophisticated ability to interpret facial expression

Idea: Map data variables to facial features



Question: Do we process facial features in an uncorrelated way? (i.e., are they *separable*?)

This is just one example of nD "glyphs"

Visualizing Multiple Dimensions

Strategies
- **Avoid "over-encoding"**
- **Use space and small multiples intelligently**
- **Reduce the problem space**
- **Use interaction to generate *relevant* views**

There is rarely a single visualization that answers all questions. Instead, the ability to generate appropriate visualizations quickly is key

Tableau / Polaris

Tableau

Research at Stanford: “Polaris” by Stolte and Hanrahan.

Tableau

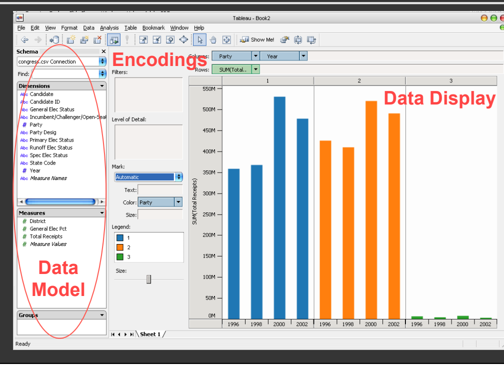


Tableau demo

The dataset:

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

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

Hypotheses?

What might we learn from this data?
Correlation between receipts and whether elected?
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?

Hypotheses?

What might we learn from this data?

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

Polaris/Tableau Approach

Insight: simultaneously specify both database queries and visualization

Choose data, then visualization, not vice versa

Use smart defaults for visual encodings

Recently: automate visualization design (ShowMe – Like APT)

Specifying Table Configurations

Operands are names of database fields

Each operand interpreted as a set {…}

Quantitative and Ordinal fields treated differently

Three operators:

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

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

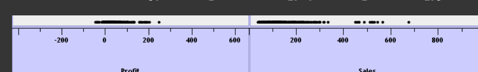
Quarter + Product Type

= {(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

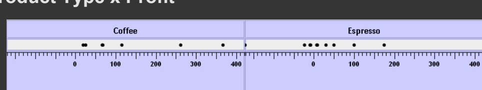
Cross-product of set interpretations

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

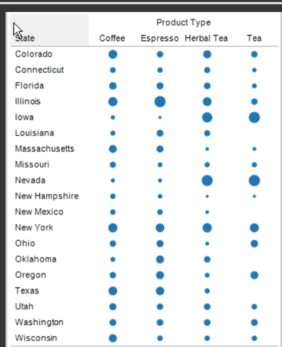
Quarter x Month

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

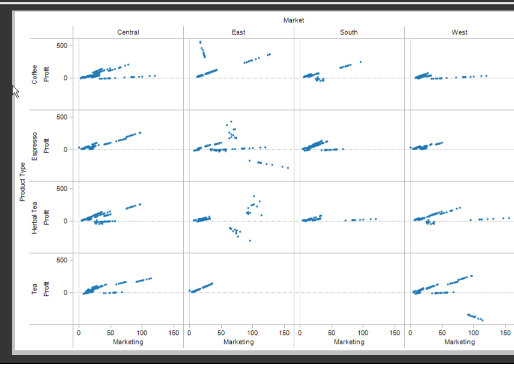
Quarter / Month

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

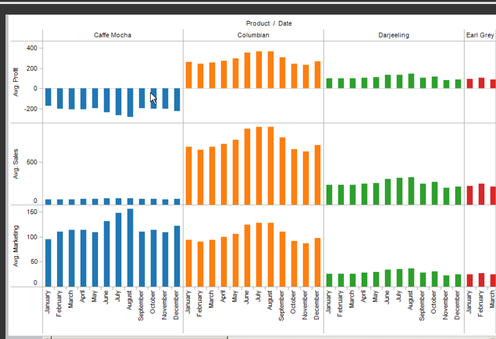
Ordinal - Ordinal



Quantitative - Quantitative



Ordinal - Quantitative



Querying the Database