



Data Visualization

Hyerim Bae

Department of Industrial Engineering, Pusan National University

hrbae@pusan.ac.kr

Contents

01

Explorative study

02

Multi-dimensional Vis.

03

Applications

한 장의 사진이 천 마디 말보다 낫다.



Explorative Data Analysis

1. Verify expected relationships

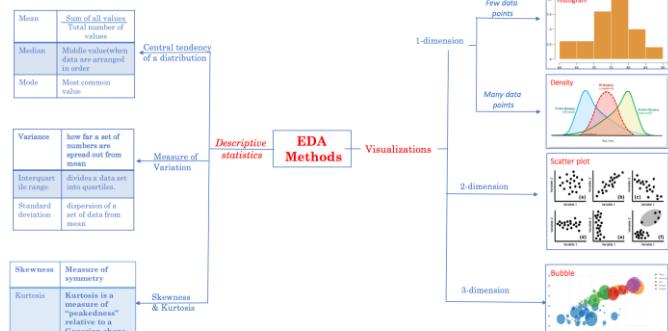
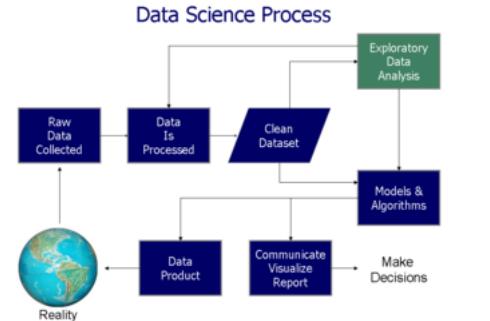
actually exist in the data, thus formulating and validating planned techniques of analysis.

2. To find some unexpected structure

in the data that must be taken into account, thereby suggesting some changes in the planned analysis.

3. Deliver data-driven insights to business stakeholders by confirming they are asking the right questions and not biasing the investigation with their assumptions.

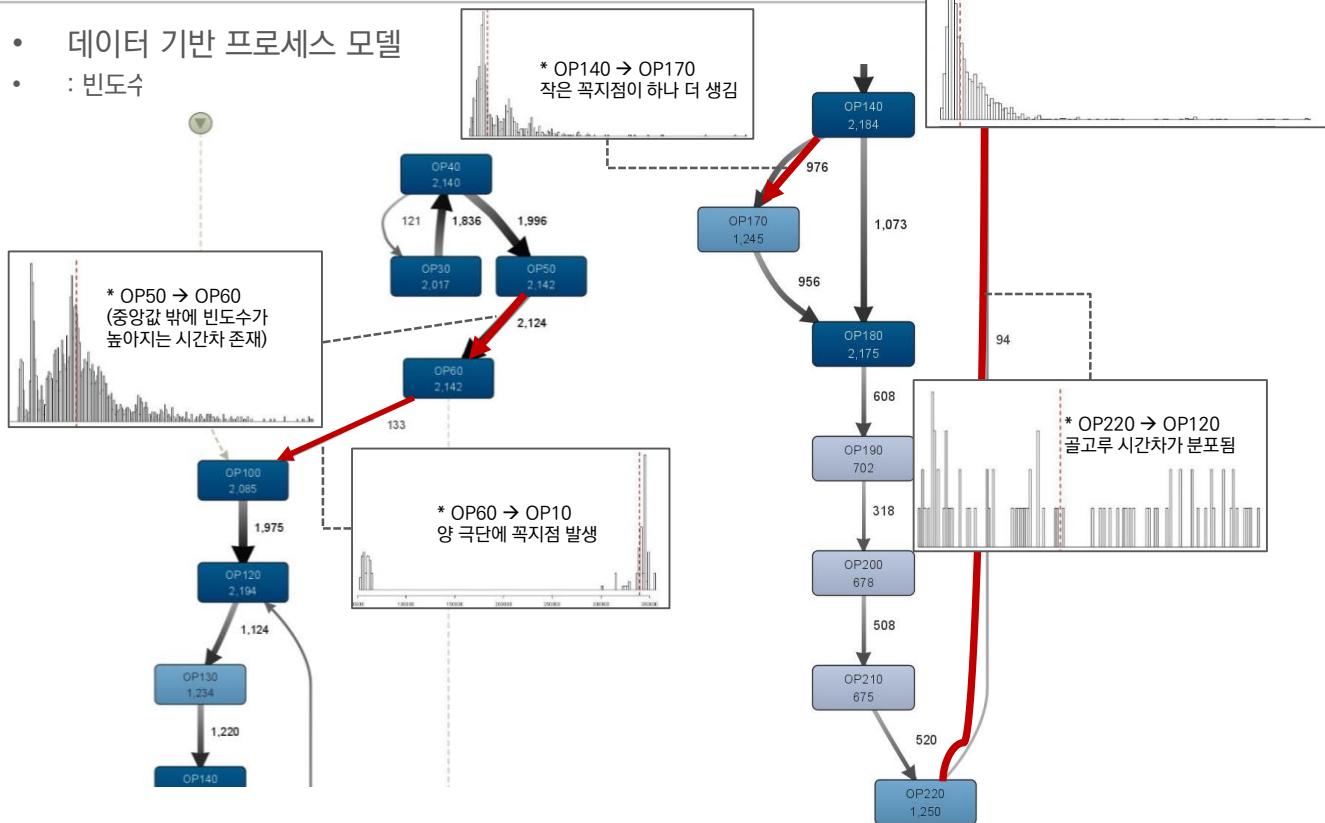
4. Provide the context around the problem to make sure the potential value of the data scientist's output can be maximized.



* 나머지 작업(OP) 간 시간 분포는 다음과 유사한 꼴을 보임
(중앙값을 중심으로 점차적으로 빈도수가 감소하는 꼴)

Explorative study: Industrial Example

- 데이터 기반 프로세스 모델
- : 빈도수



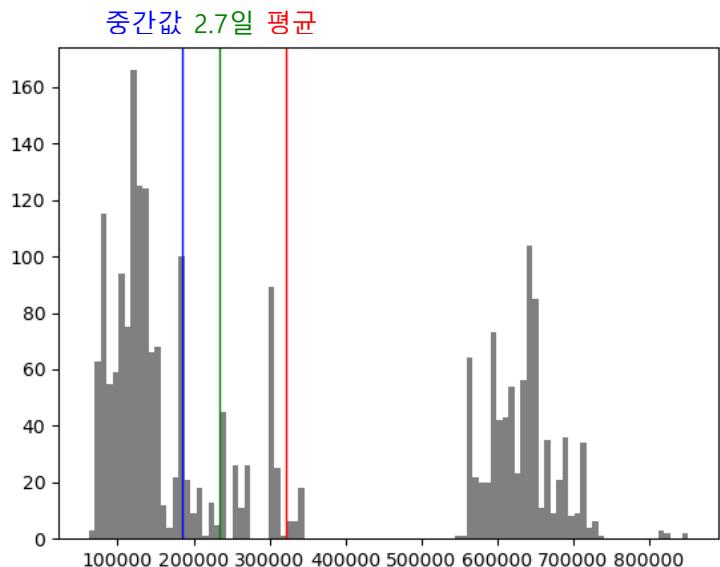
Continued

- OP30~OP220
 - 기간 내 OP30 ~ OP220을 모두 거친 제품은 2260개

- 생산에 걸린 시간:

- 최소: 17시간 7분
 - 최대: 9일 20시간
 - 평균: 3.7일
 - 중간값: 51.4시간 (약 2.1일)

OP60 → OP100 평균 3.6일



Graphs for Data Exploration

Basic Plots

Line Graphs

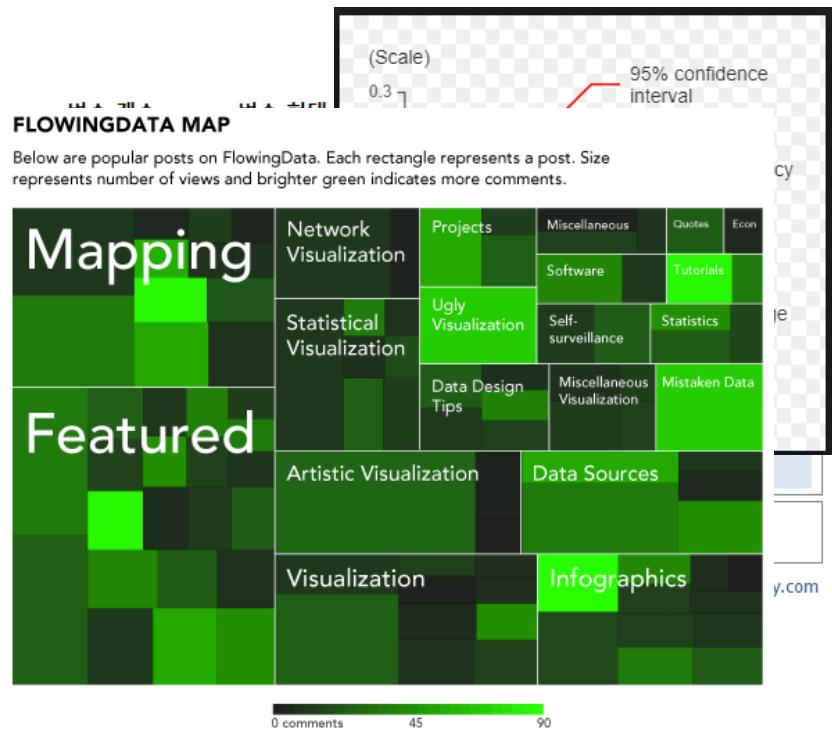
Bar Charts

Scatterplots

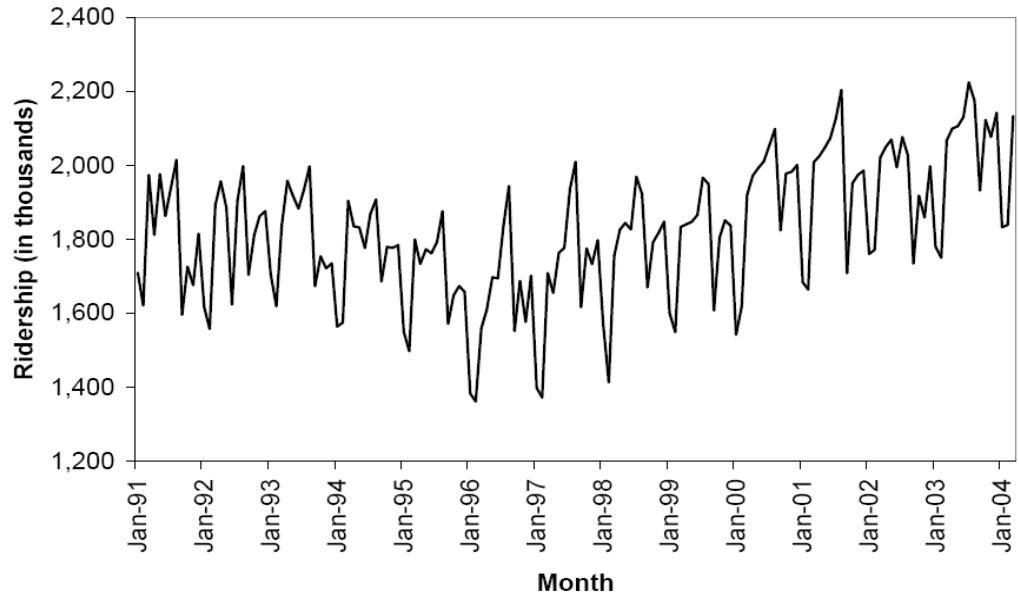
Distribution Plots

Boxplots

Histograms



Line Graph for Time Series

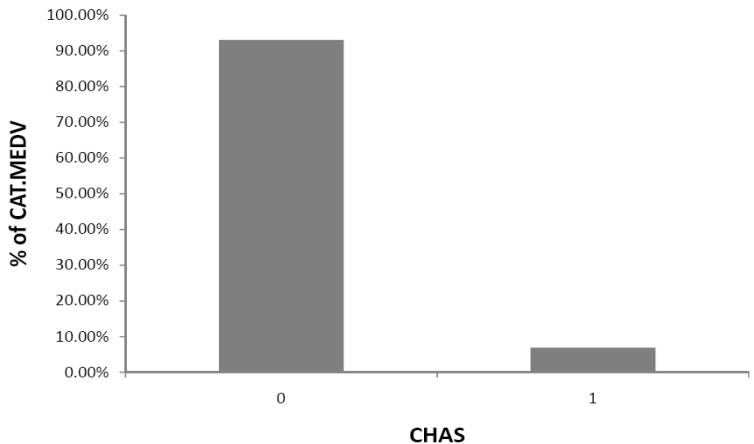


CRIM	per capita crime rate by town
ZN	proportion of residential land zoned for lots over 25,000 sq.ft.
INDUS	proportion of non-retail business acres per town.
CHAS	Charles River dummy variable (1 if tract bounds river; 0 otherwise)
NOX	nitric oxides concentration (parts per 10 million)
RM	average number of rooms per dwelling
AGE	proportion of owner-occupied units built prior to 1940
DIS	weighted distances to five Boston employment centres
RAD	index of accessibility to radial highways
TAX	full-value property-tax rate per \$10,000
PTRATIO	pupil-teacher ratio by town
B	$1000(Bk - 0.63)^2$ where Bk is the proportion of blacks by town
LSTAT	% lower status of the population
MEDV	Median value of owner-occupied homes in \$1000

Bar Chart for Categorical Variable

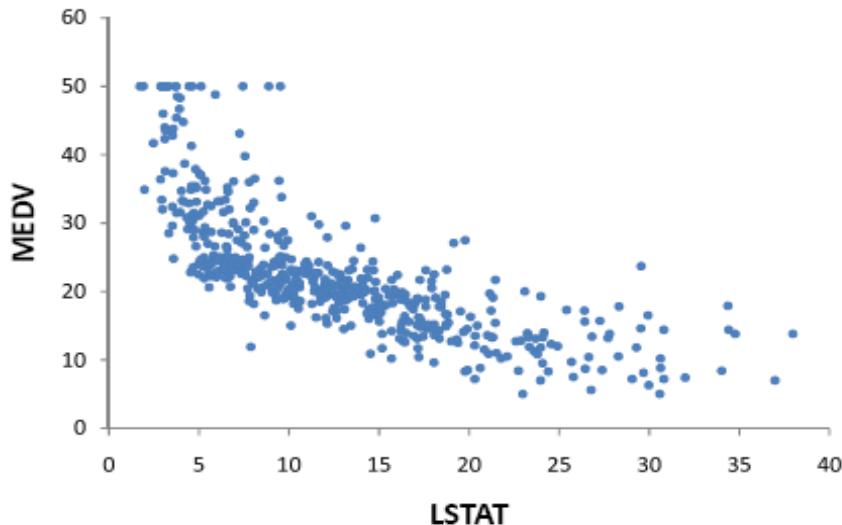
95% of tracts do not border Charles River

Excel can confuse: y-axis is actually “% of records that have a value for CATMEDV” (i.e., “% of all records”)



Scatterplot

Displays relationship between two numerical variables



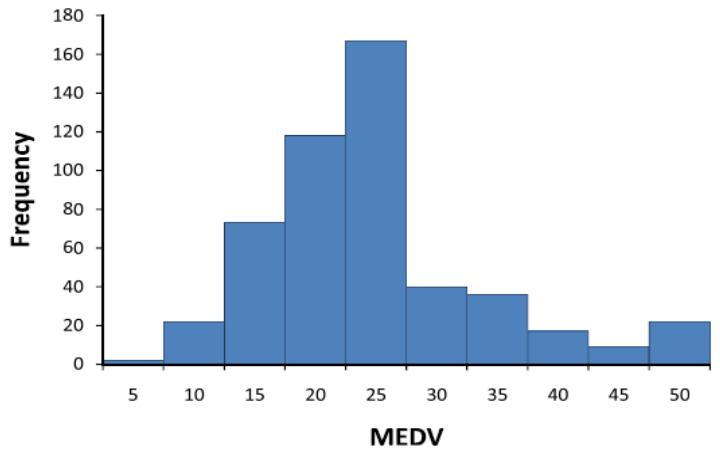
Distribution Plots

- Display “how many” of each value occur in a data set
- Or, for continuous data or data with many possible values, “how many” values are in each of a series of ranges or “bins”

Histograms

Histogram shows the distribution of the outcome variable (median house value)

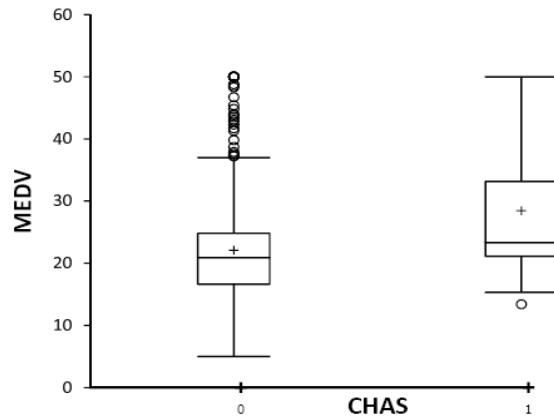
Boston Housing example:



Boxplots

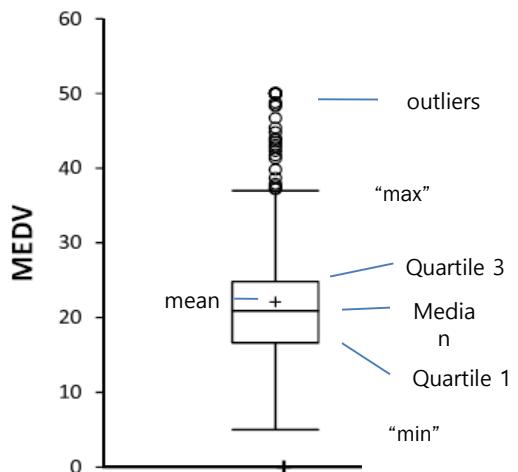
Boston Housing Example: Display distribution of outcome variable (MEDV) for neighborhoods on Charles river (1) and not on Charles river (0)

- Side-by-side boxplots are useful for comparing subgroups



Box Plot

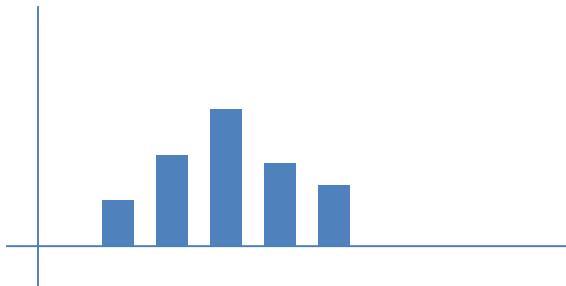
- Top outliers defined as those above $Q3 + 1.5(Q3 - Q1)$.
- “max” = maximum of non-outliers
- Analogous definitions for bottom outliers and for “min”
- Details may differ across software



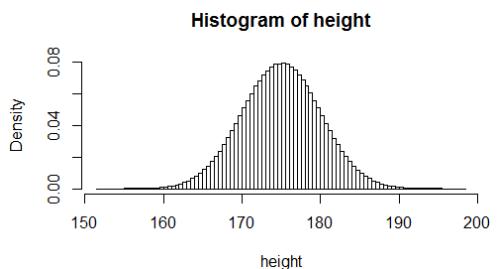
Normal Distribution

170	178	171	168	173	178	171	174	170	170	175
170	169	166	162	170	171	175	175	171	171	170
172	179	164	170	181	178	180	177	166	169	168
165	163	175	166	178	165	168	167	177	168	177
174	174	176	179	169	173	167	170	173	170	162

계급구간(cm)	도수
161.5 이상 165.5 미만	6
165.5 이상 169.5 미만	12
169.5 이상 173.5 미만	18
173.5 이상 177.5 미만	11
177.5 이상 181.5 미만	8
합계	55



$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-m)^2}{2\sigma^2}}$$



Heat Maps

Color conveys information

In data mining, used to visualize

Correlations

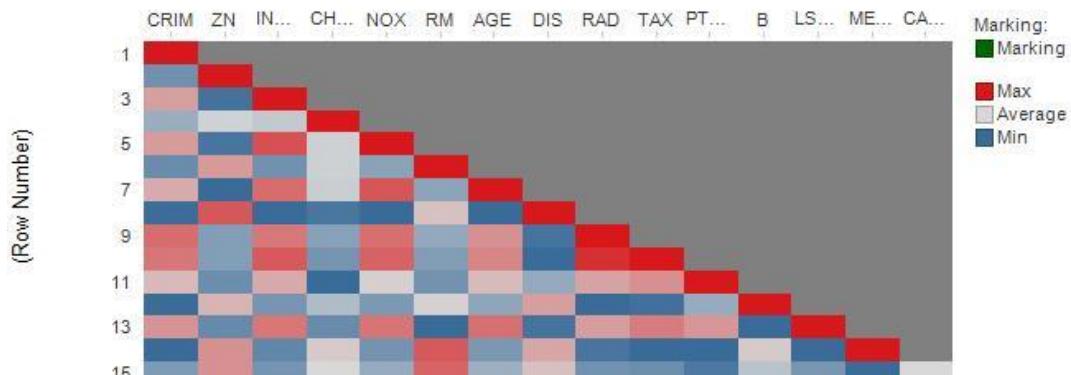
Missing Data

Heatmap to highlight correlations (Boston Housing)

	CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX	PTRATIO	B	LSTAT	MEDV
CRIM														
ZN	-0.20													
INDUS	0.41	-0.53												
CHAS	-0.06	-0.04	0.06											
NOX	0.42	-0.52	0.76	0.09										
RM	-0.22	0.31	-0.39	0.09	-0.30									
AGE	0.35	-0.57	0.64	0.09	0.73	-0.24								
DIS	-0.38	0.66	-0.71	-0.10	-0.77	0.21	-0.75							
RAD	0.63	-0.31	0.60	-0.01	0.61	-0.21	0.46	-0.49						
TAX	0.58	-0.31	0.72	-0.04	0.67	-0.29	0.51	-0.53	0.91					
PTRATIO	0.29	-0.39	0.38	-0.12	0.19	-0.36	0.26	-0.23	0.46	0.46				
B	-0.39	0.18	-0.36	0.05	-0.38	0.13	-0.27	0.29	-0.44	-0.44	-0.18			
LSTAT	0.46	-0.41	0.60	-0.05	0.59	-0.61	0.60	-0.50	0.49	0.54	0.37	-0.37		
MEDV	-0.39	0.36	-0.48	0.18	-0.43	0.70	-0.38	0.25	-0.38	-0.47	-0.51	0.33	-0.74	

In Excel
(using
conditional
formatting)

Heat Map



In Spotfire



Multidimensional Visualization

Scatterplot with color added

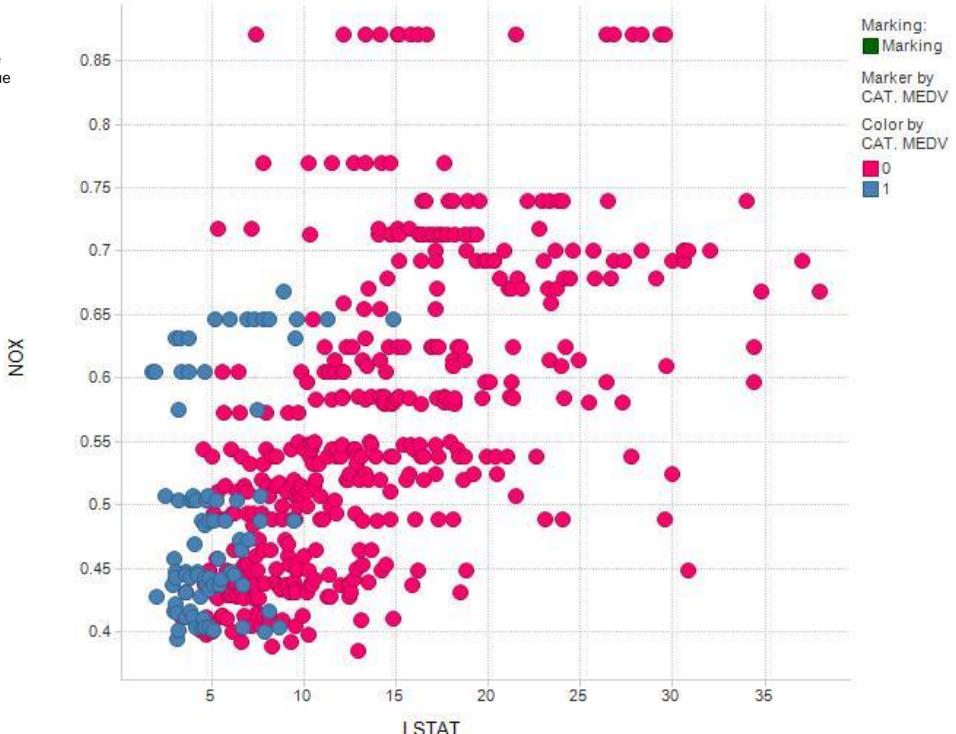
Scatter Plot

Boston Housing

NOX vs. LSTAT

Red = low median value

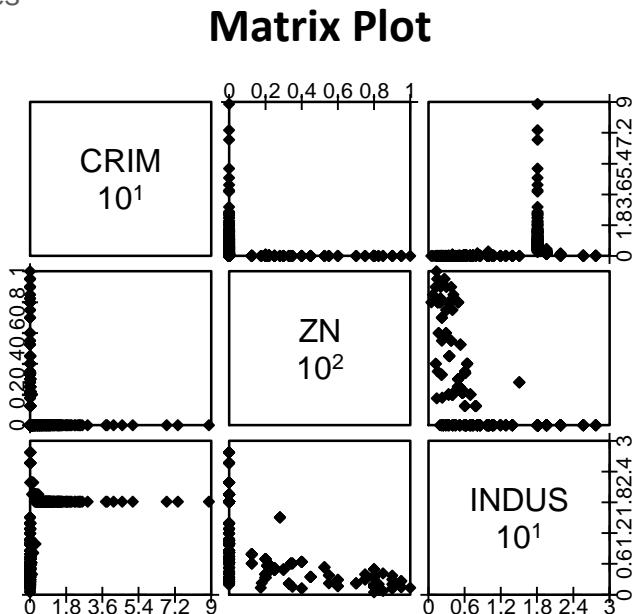
Blue = high median value



Matrix Plot

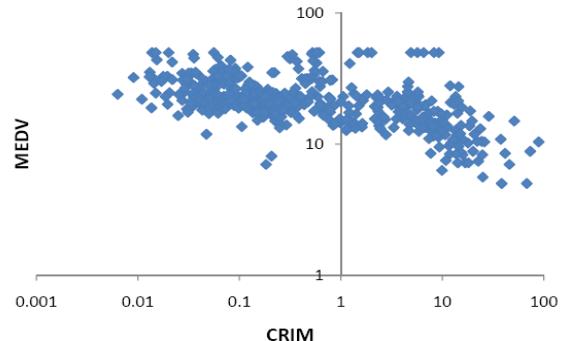
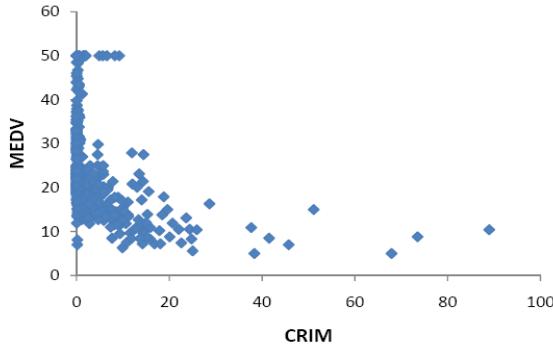
Shows scatterplots for variable pairs

Example: scatterplots for 3 Boston Housing variables



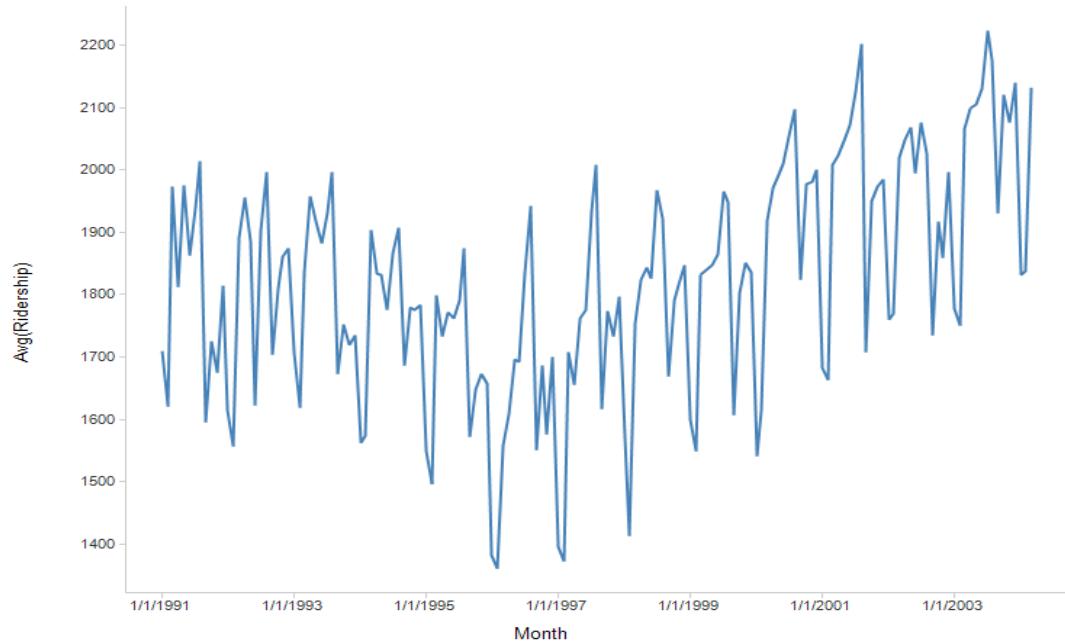
Rescaling to log scale

- “uncrowds” the data

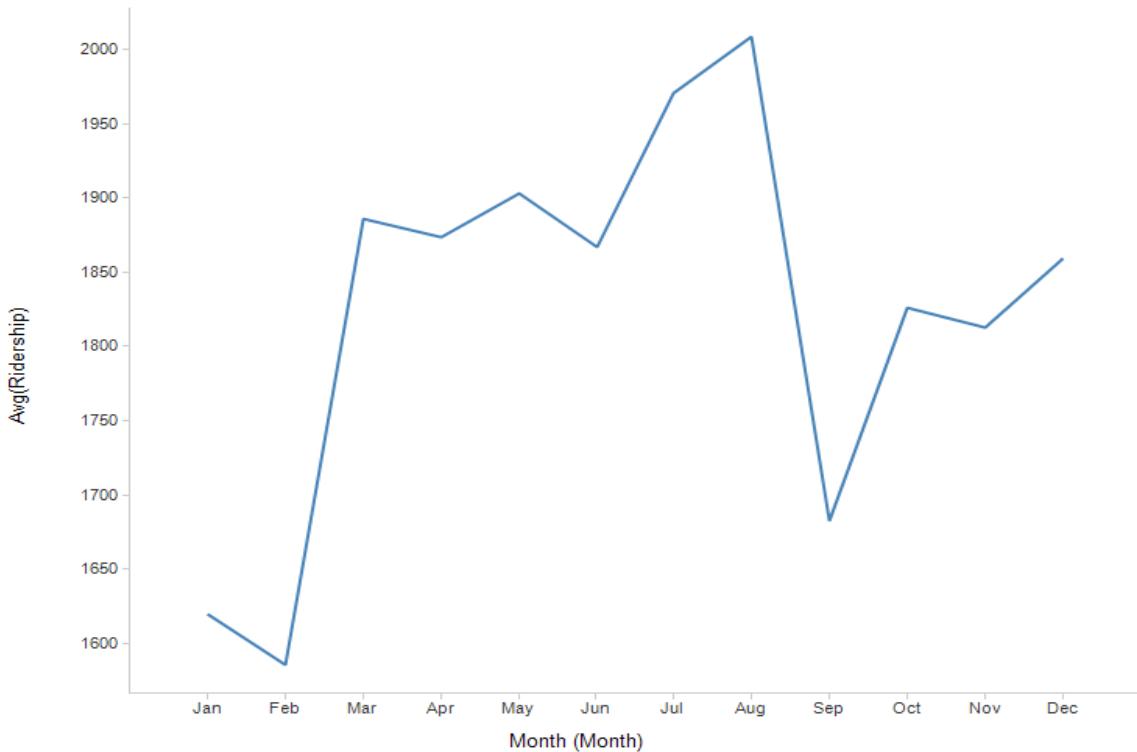


Aggregation

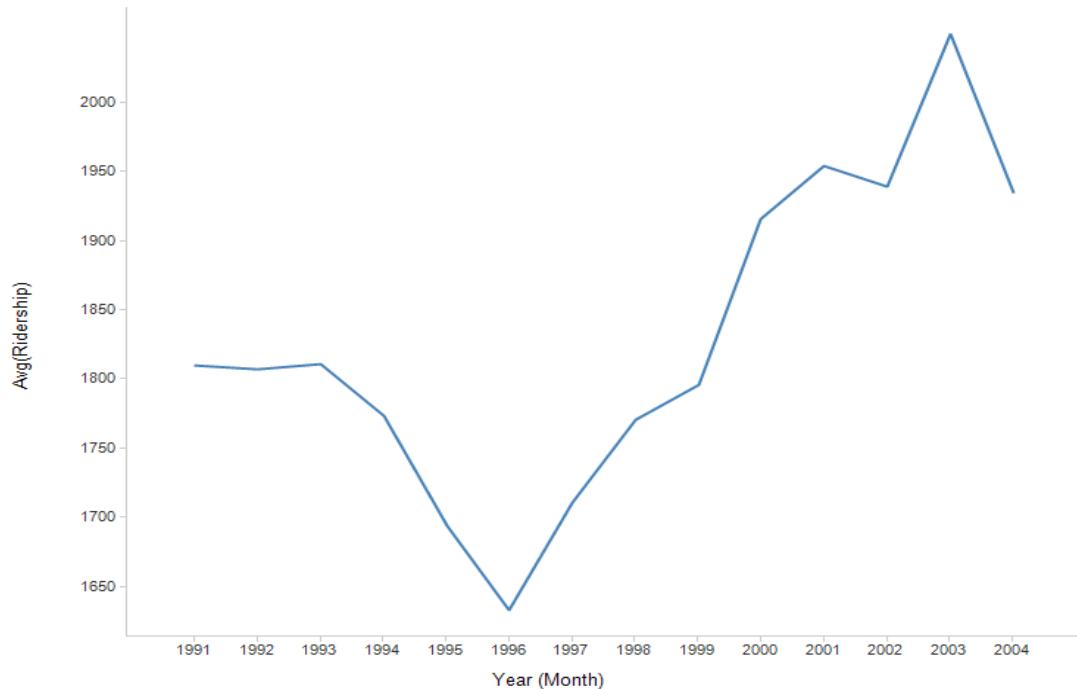
Amtrak Ridership – Monthly Data



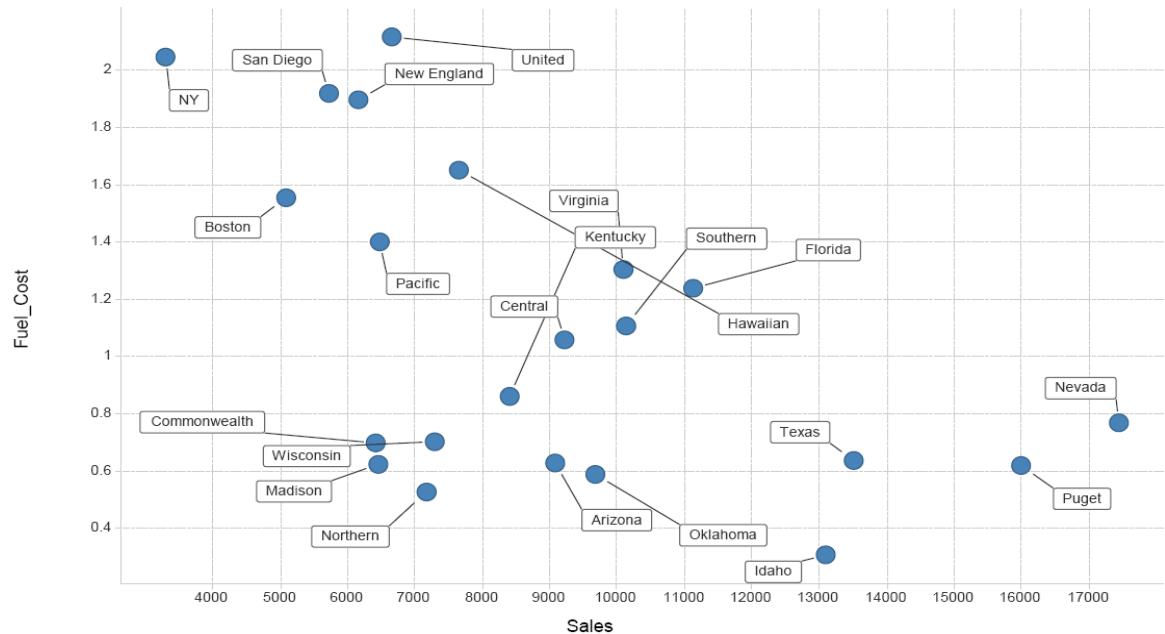
Aggregation – Monthly Average



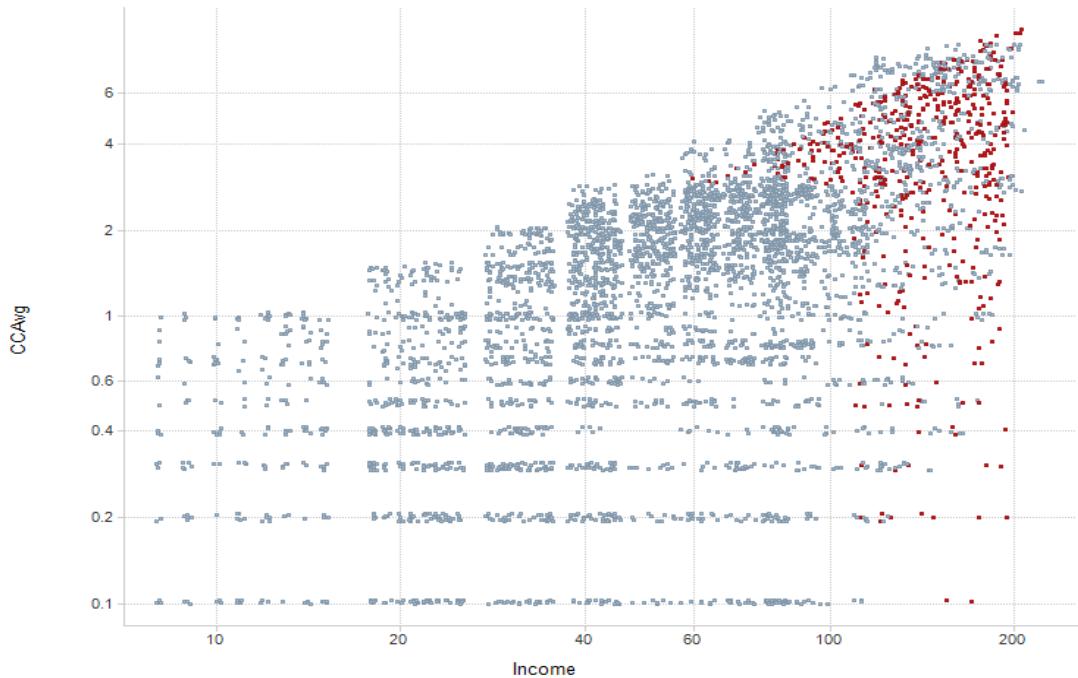
Aggregation – Yearly Average



Scatter Plot with Labels (Utilities)



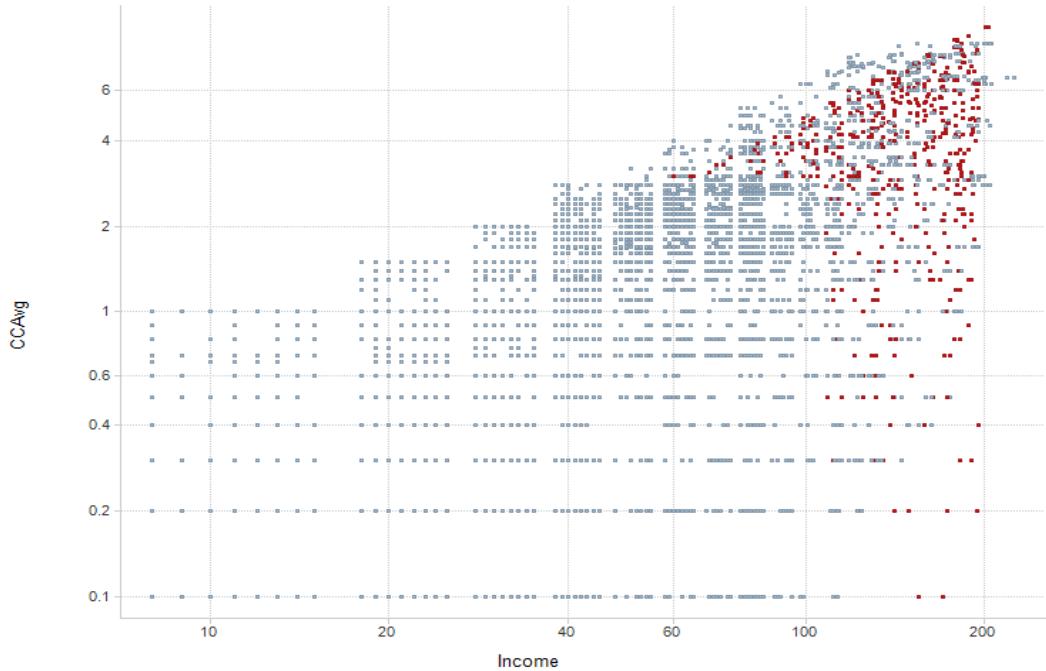
Scaling: Smaller markers, jittering, color contrast (Universal Bank; red = accept loan)



Jittering

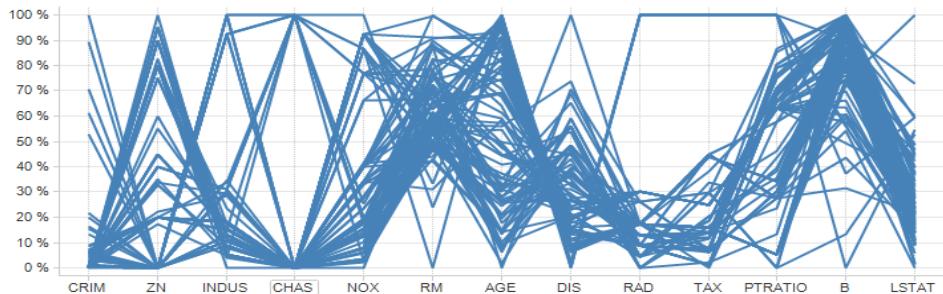
- Moving markers by a small random amount
- Uncrowds the data by allowing more markers to be seen

Without jittering (for comparison)

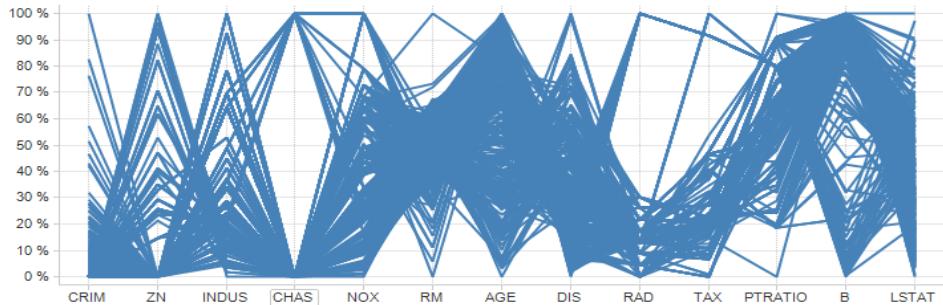


Parallel Coordinate Plot (Boston Housing)

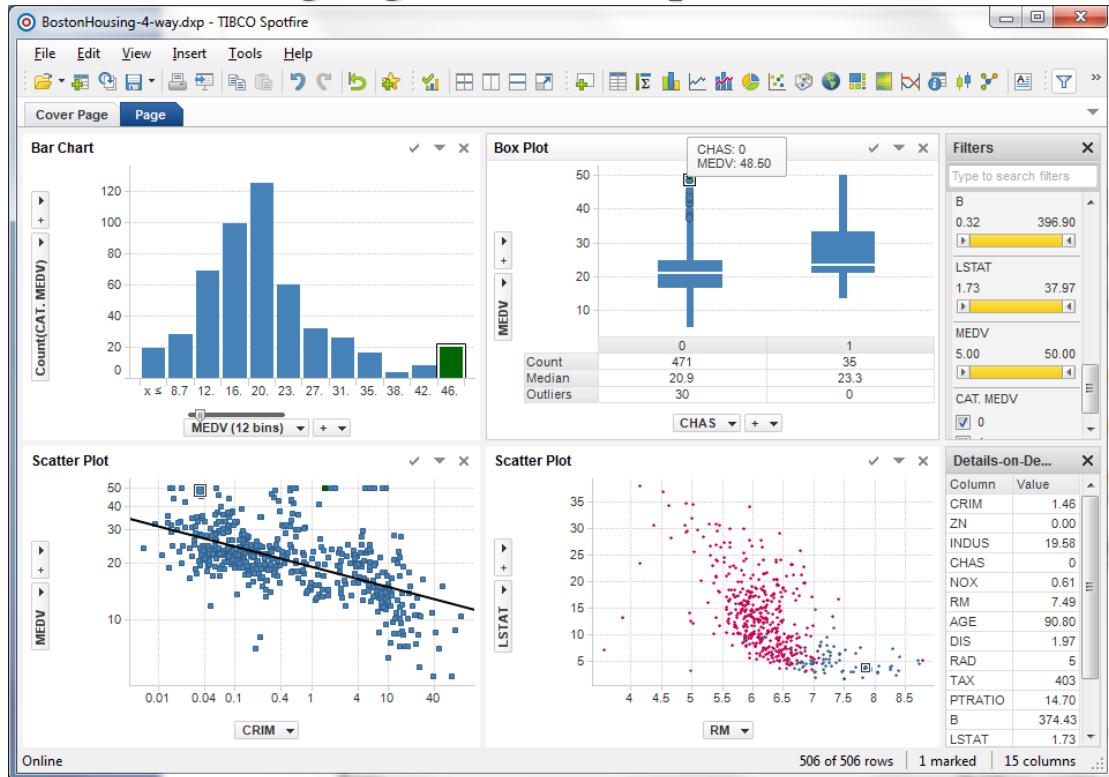
CATMEDV = 1



CATMEDV = 0



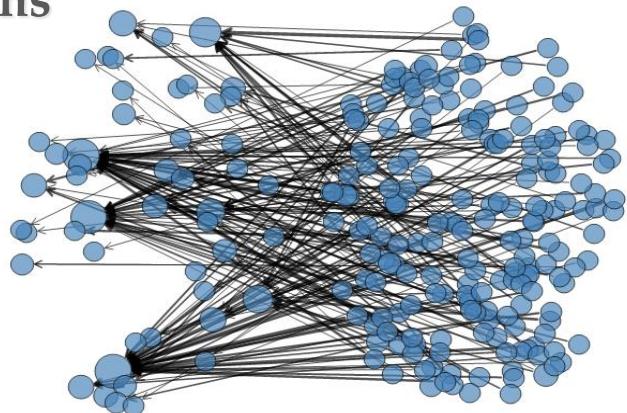
Linked plots (same record is highlighted in each plot)



Network Graph – eBay Auctions (sellers on left, buyers on right)

Circle size = # of transactions for the node

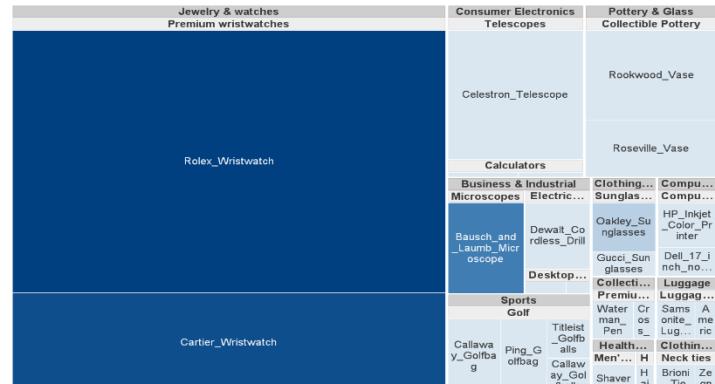
Line width =# of auctions for the buyer-seller pair



Arrows point from buyer to seller

Rectangle size =
average closing
price (=item
value)

Color = % sellers
with negative
feedback
(darker=more)



Map Chart

(Comparing countries' well-being with GDP)

Well-Being Score

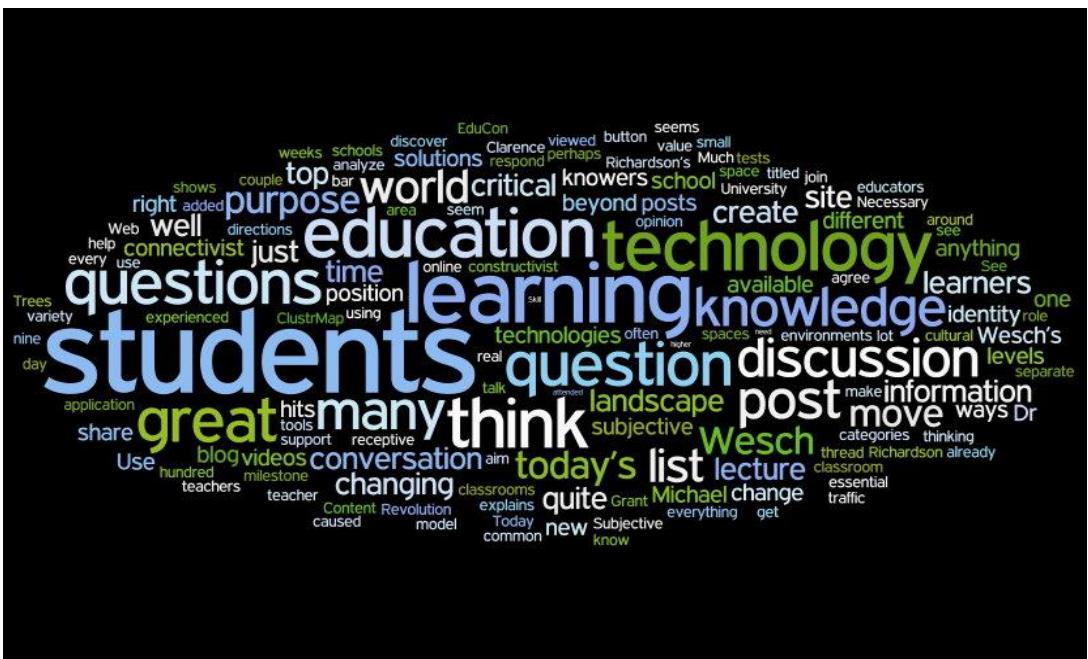


Darker = higher value

GDP



wordle

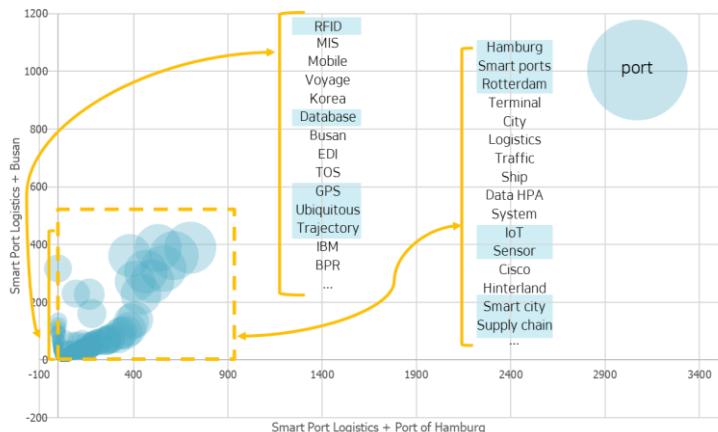


항만 불개미 해양

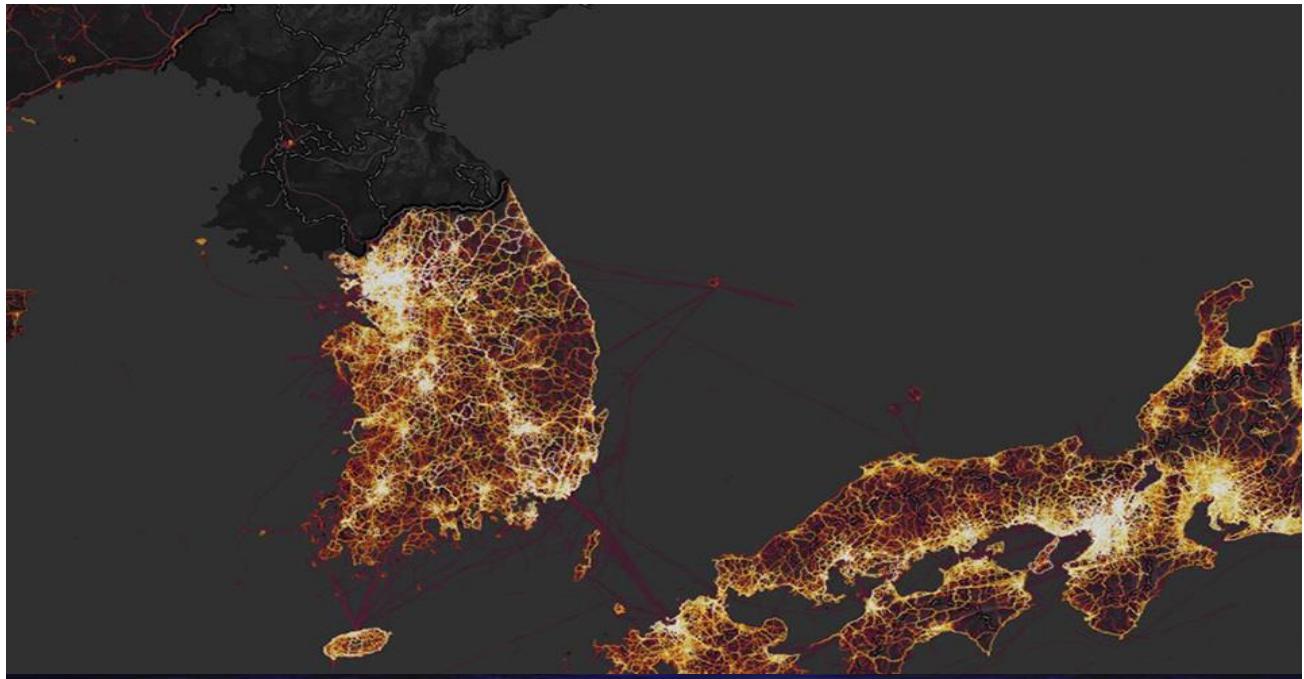
한진해운 아시아 세계 컨테이너
 현대상선 하역 일항보도자료
 환적화물 해수증기 수산물로벌 경쟁력을 높여 부두 정부해운
 터미널선박 사업공사
 주진 수출 지역
 물동량 운영
 선사 크루즈
 국내복항 처리

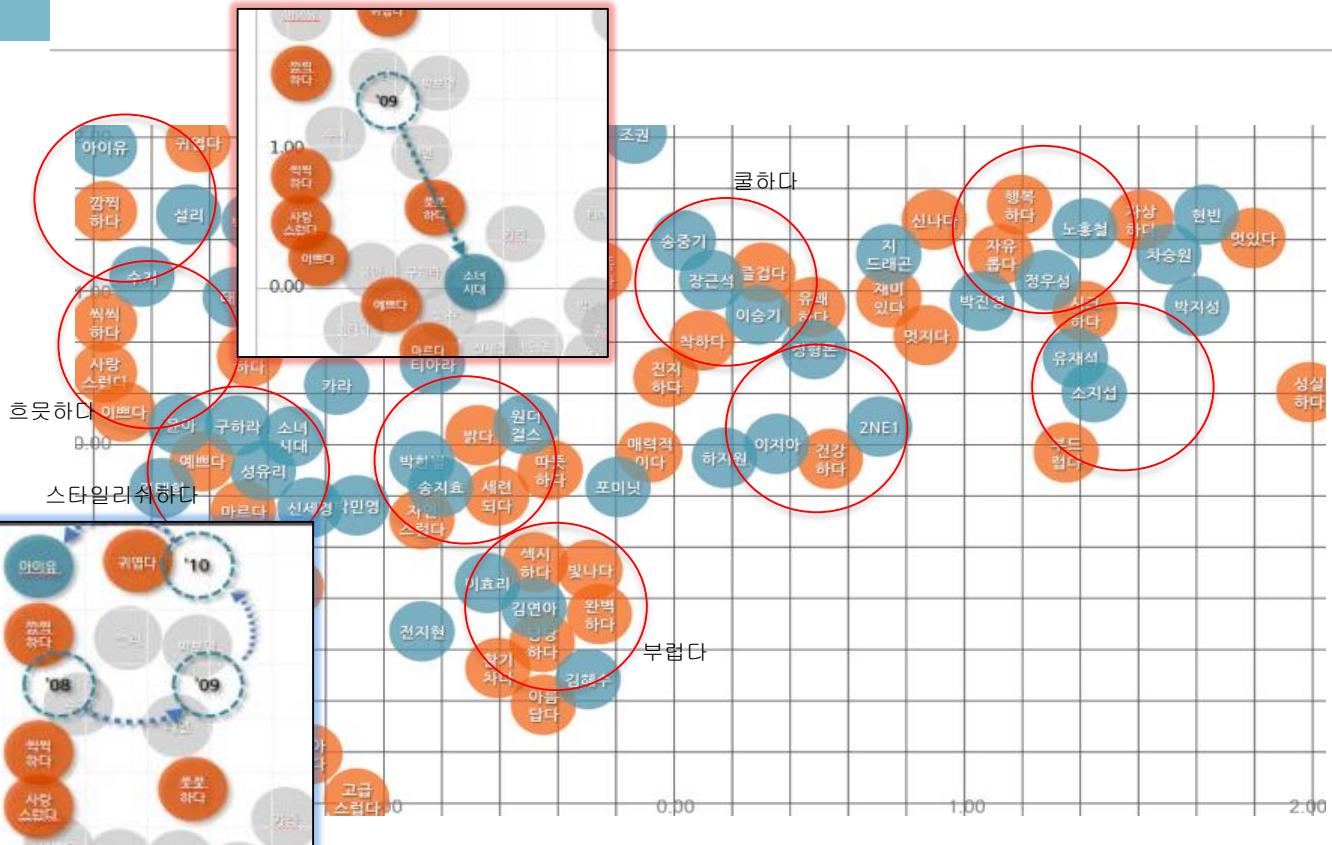
- 조사 기간 : 17.01.01 – 17.11.30 뉴스 기사 대상
- 주요 이슈
 - 살인 불개미 발견 (불개미, 발견, 안전, 부두)
 - 물동량 2000만TEU 기대 (물동량, 컨테이너, 처리, 증가)
 - 사드 보복에 따른 크루즈 이용객 급감 대책 마련 필요 +
크루즈 관광 활성화 (여객터미널, 크루즈, 북항)
- 기타 : 채용비리, 해수 온천, 한진해운 파산 여파

❖ Busan VS Hamburg



- Smart port logistics이라는 키워드에 Busan과 Hamburg를 추가하여 비교 분석
- Busan 키워드에서는 RFID, Database, GPS, Ubiquitous 등 이전 세대의 기술적 키워드가 도출
- Hamburg 키워드에서는 IoT, Smart ports, Rotterdam, Sensor, Smart city 등 4차 산업혁명의 기반 기술적 키워드가 도출





LineUp



Data visualization tools

- D3
- HighCharts
- Echarts
- Leaflet
- Vega
- Deck.gl
- Power BI
- Tableau
- FineReport

Visualization case

<https://informationisbeautiful.net/visualizations/covid-19-coronavirus-infographic-datapack/#activities>

<https://www.tableau.com/ko-kr/learn/articles/best-beautiful-data-visualization-examples>