BSC LAB
Business & Service Computing

PNU Industrial Data Science

Data Visualization

Prof. Hyerim Bae

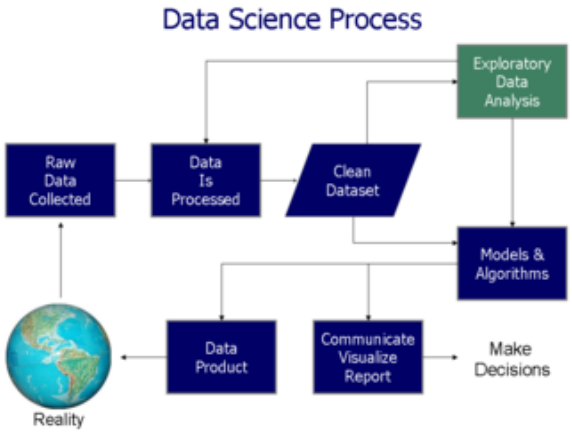
Contents

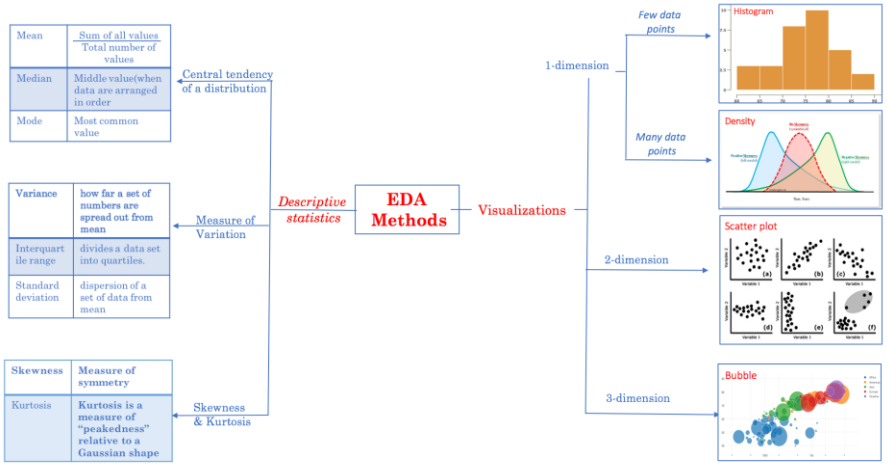
산업데이터과학은 산업현장에서 수집된 데이터를 분석하는데 필요한 기초 소양을 강의합니다.

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.

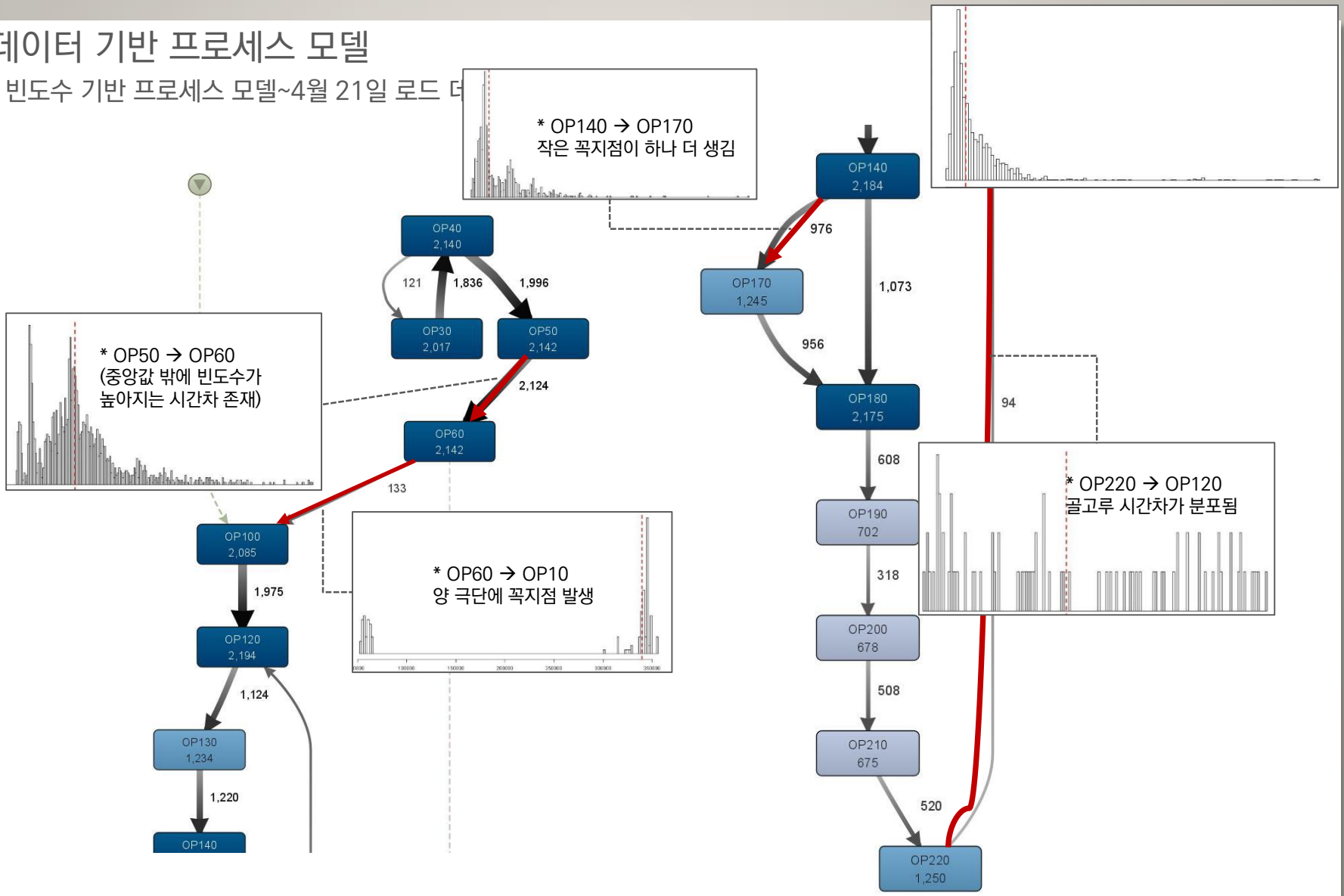




Explorative study: Industrial Example

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

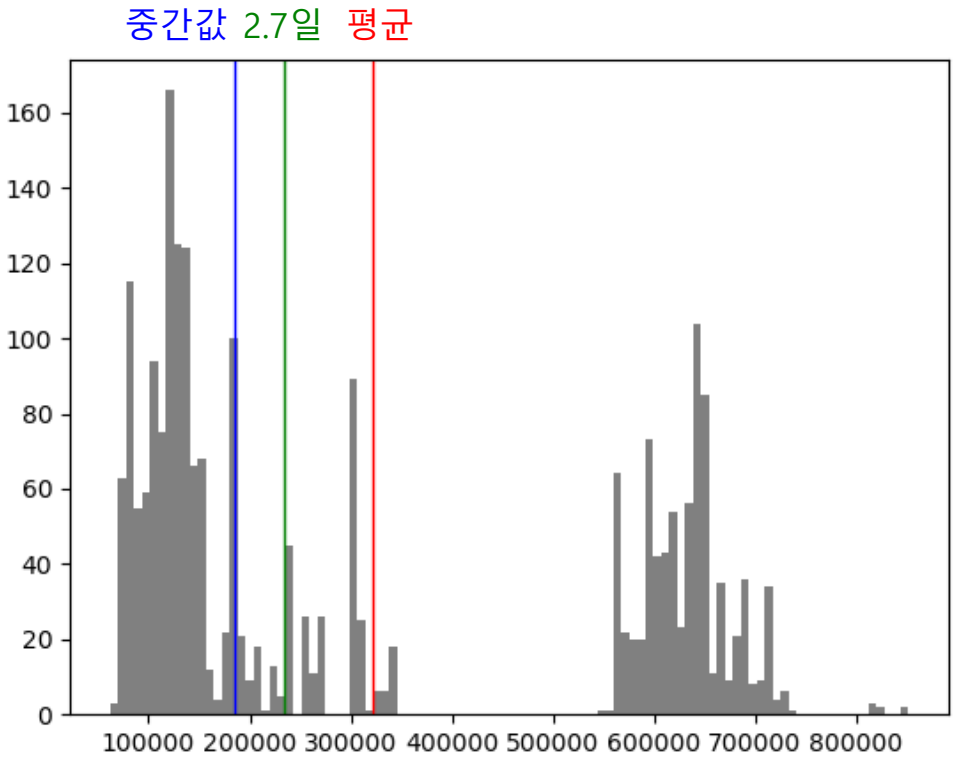
- 데이터 기반 프로세스 모델
- : 빈도수 기반 프로세스 모델~4월 21일 로드 더



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



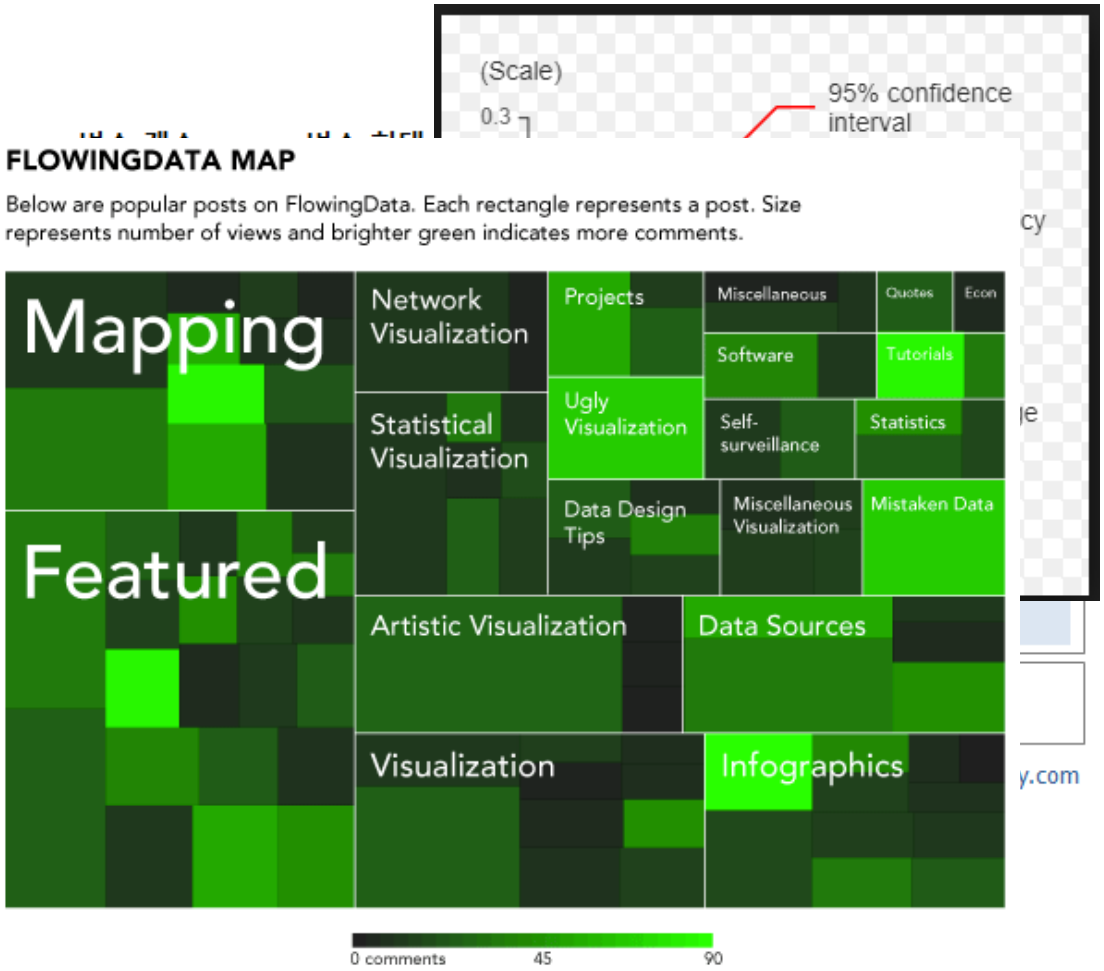
Graphs for Data Exploration

Basic Plots

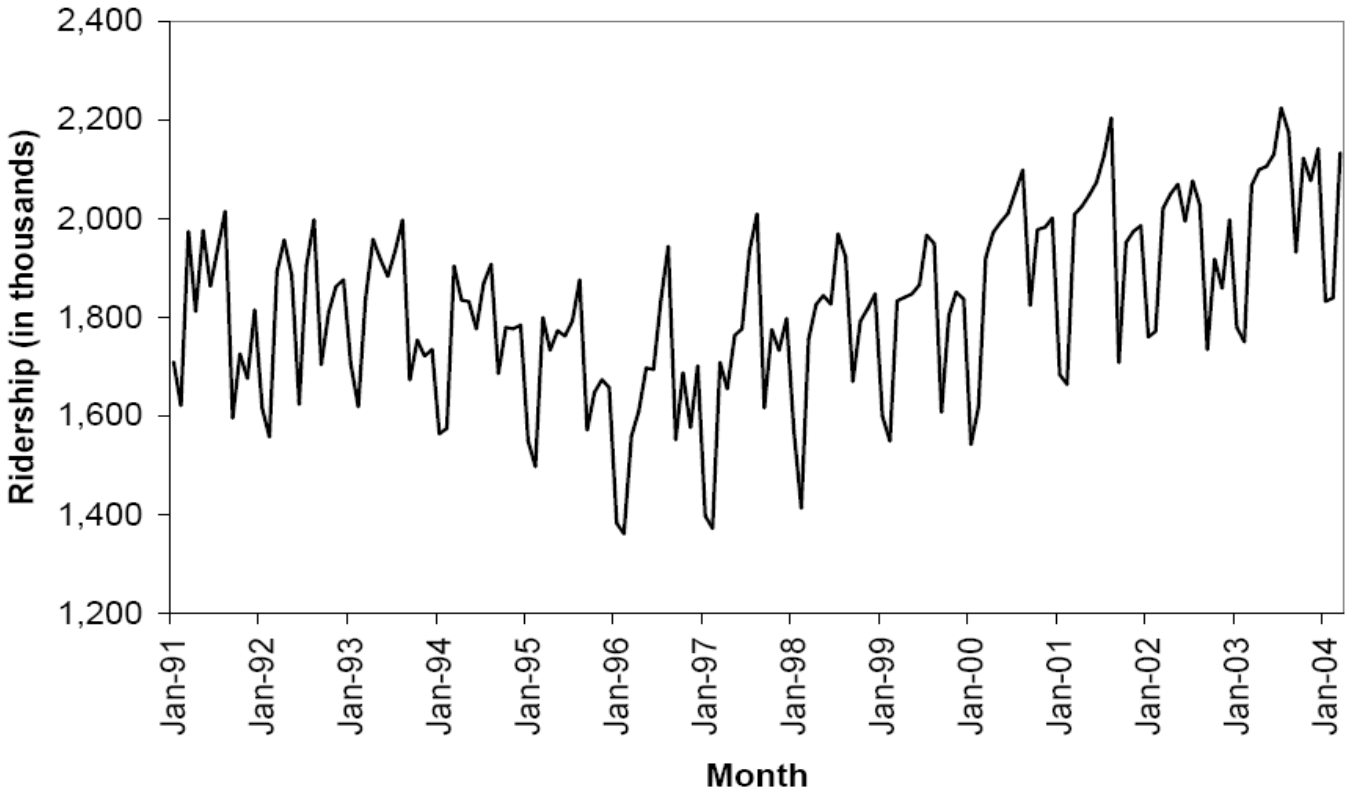
- Line Graphs
- Bar Charts
- Scatterplots

Distribution Plots

- Boxplots
- Histograms



Line Graph for Time Series

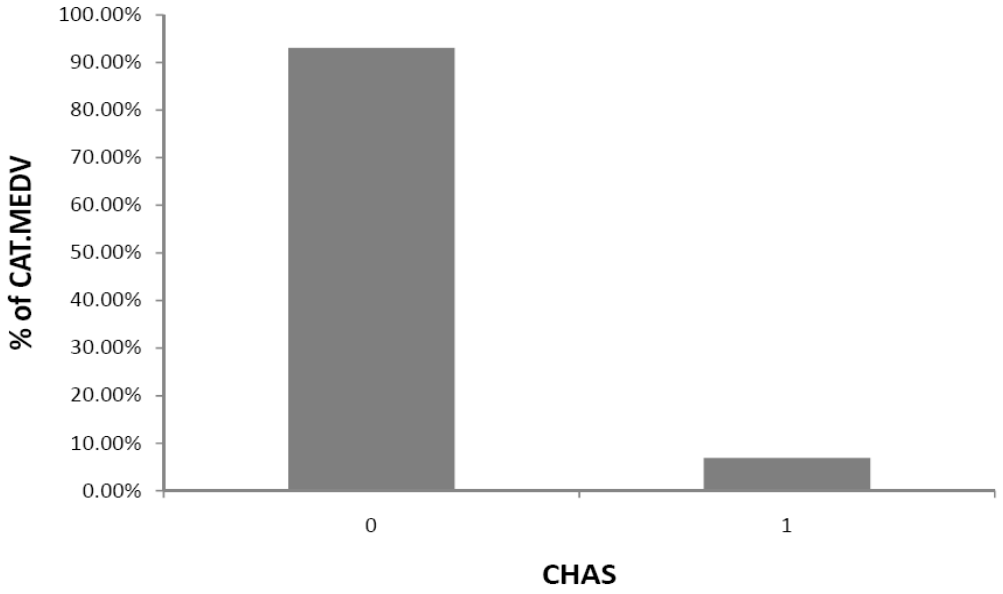


CRIM per capita crime rate by town
ZN proportion of residential land zoned for lots over 25,000 sq.ft.
INDUS proportion of non-retail business acres per town.
CHAS Charles River dummy variable (1 if tract bounds river; 0 otherwise)
NOX nitric oxides concentration (parts per 10 million)
RM average number 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

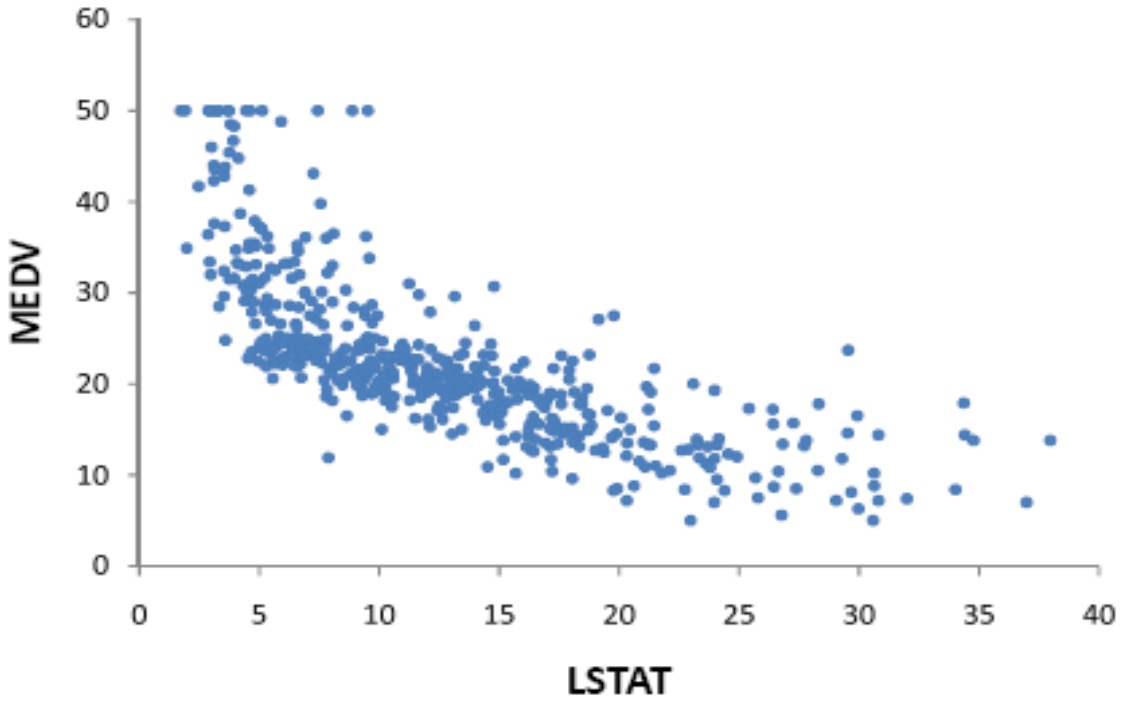
95% of tracts do not border Charles River

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



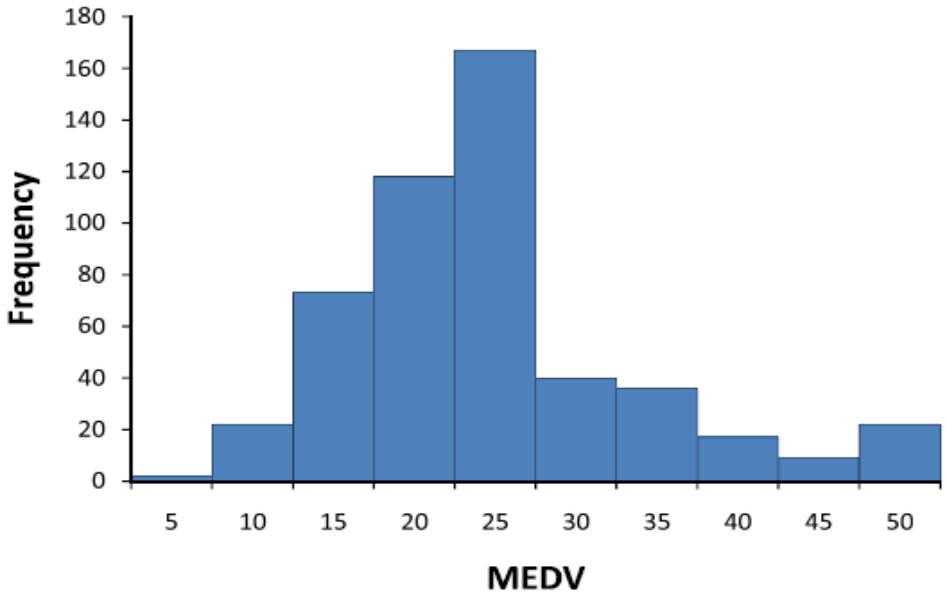
Distribution Plots

- 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)

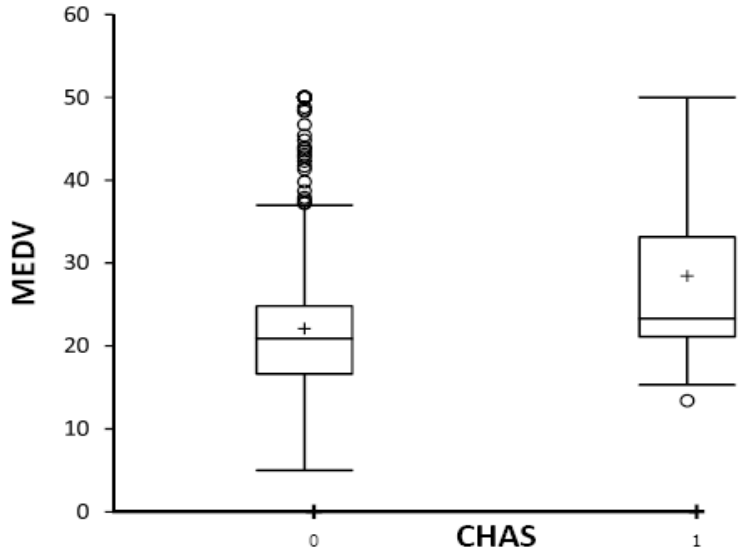
Boston Housing example:



Boxplots

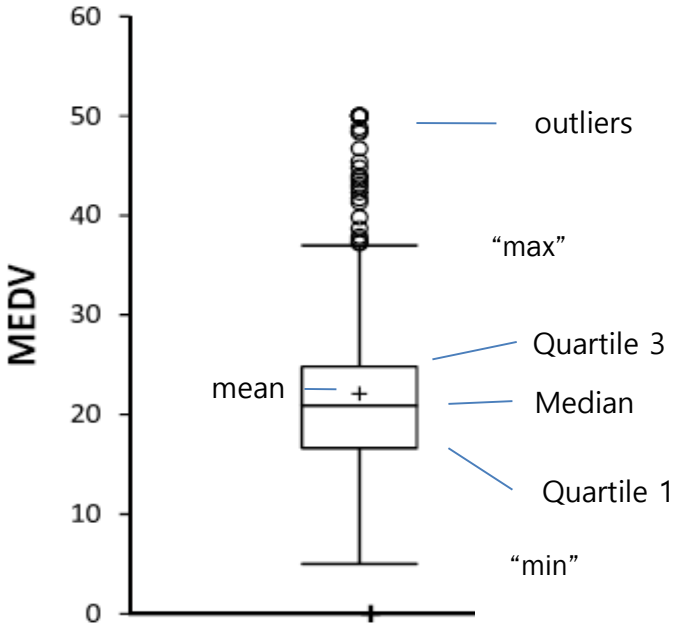
Boston Housing Example: Display distribution of outcome variable (MEDV) for neighborhoods on Charles river (1) and not on Charles river (0)

- 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



Heat Maps

Color conveys information

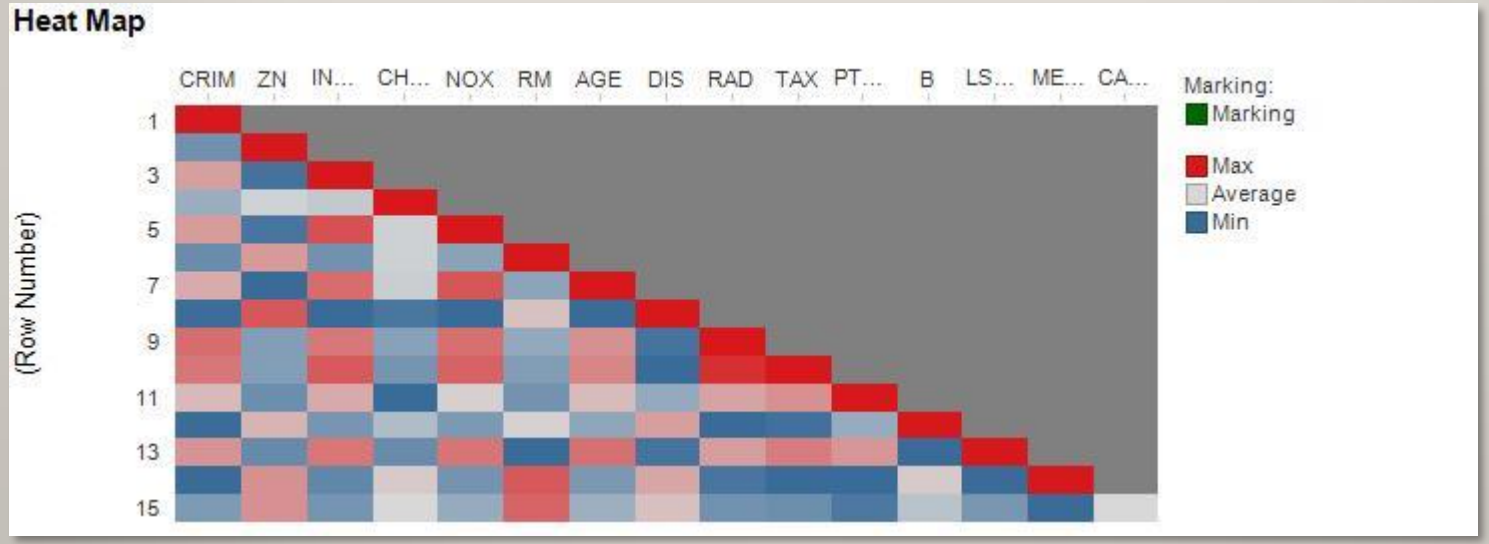
In data mining, used to visualize

- 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)



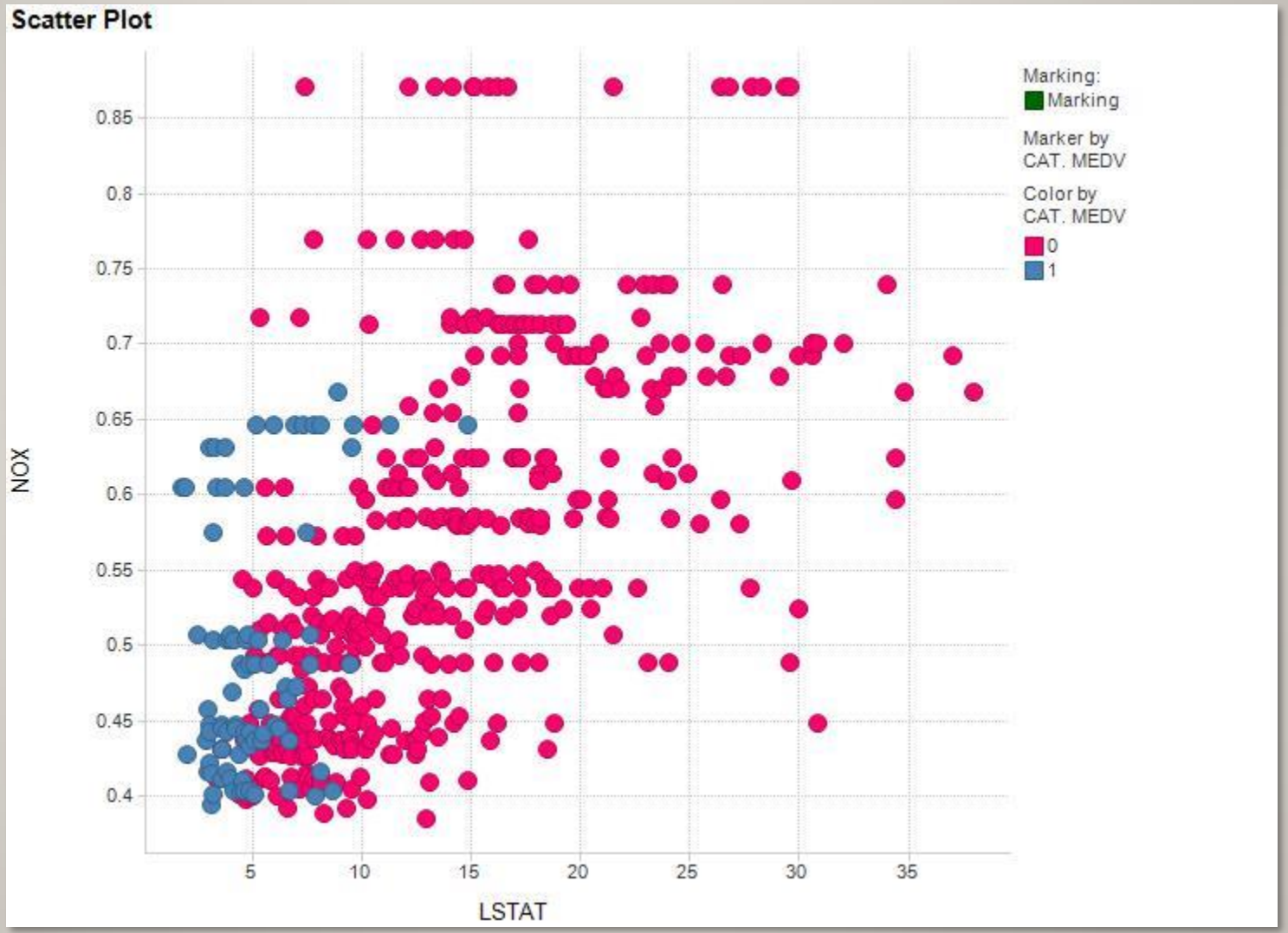
In Spotfire

Multidimensional Visualization

Scatterplot with color added

Boston Housing

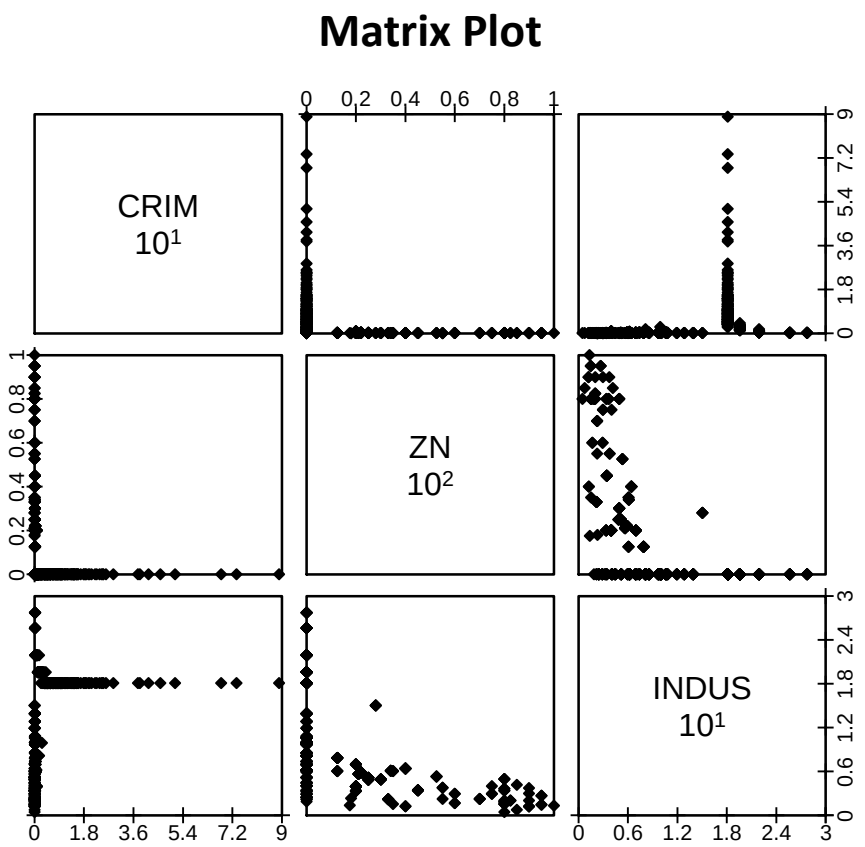
NOX vs. LSTAT
Red = low median value
Blue = high median value



Matrix Plot

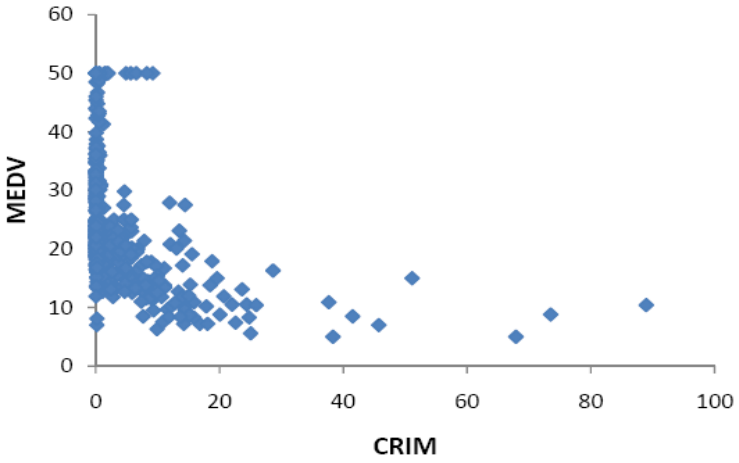
Shows scatterplots for variable pairs

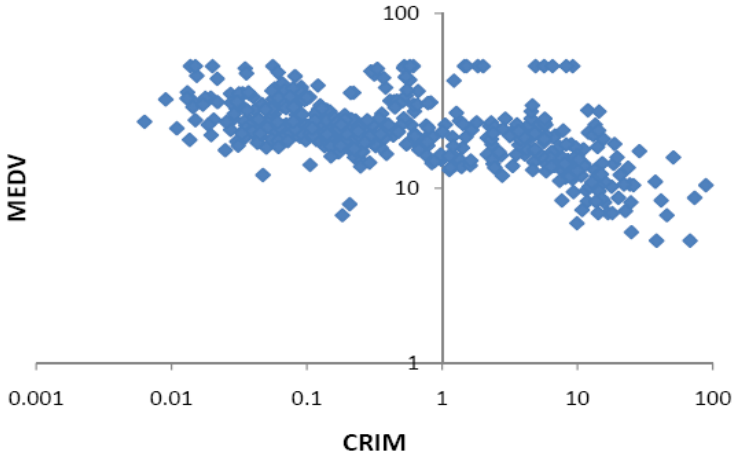
Example: scatterplots for 3 Boston Housing variables

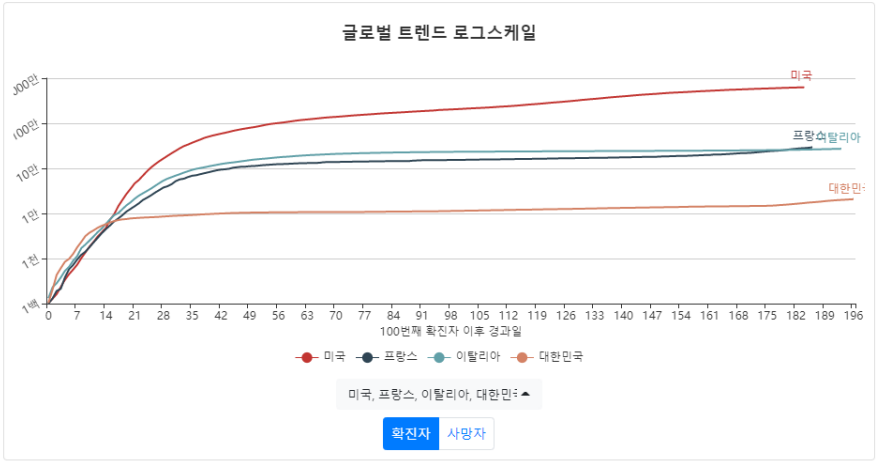


Rescaling to log scale

- "uncrowds" the data

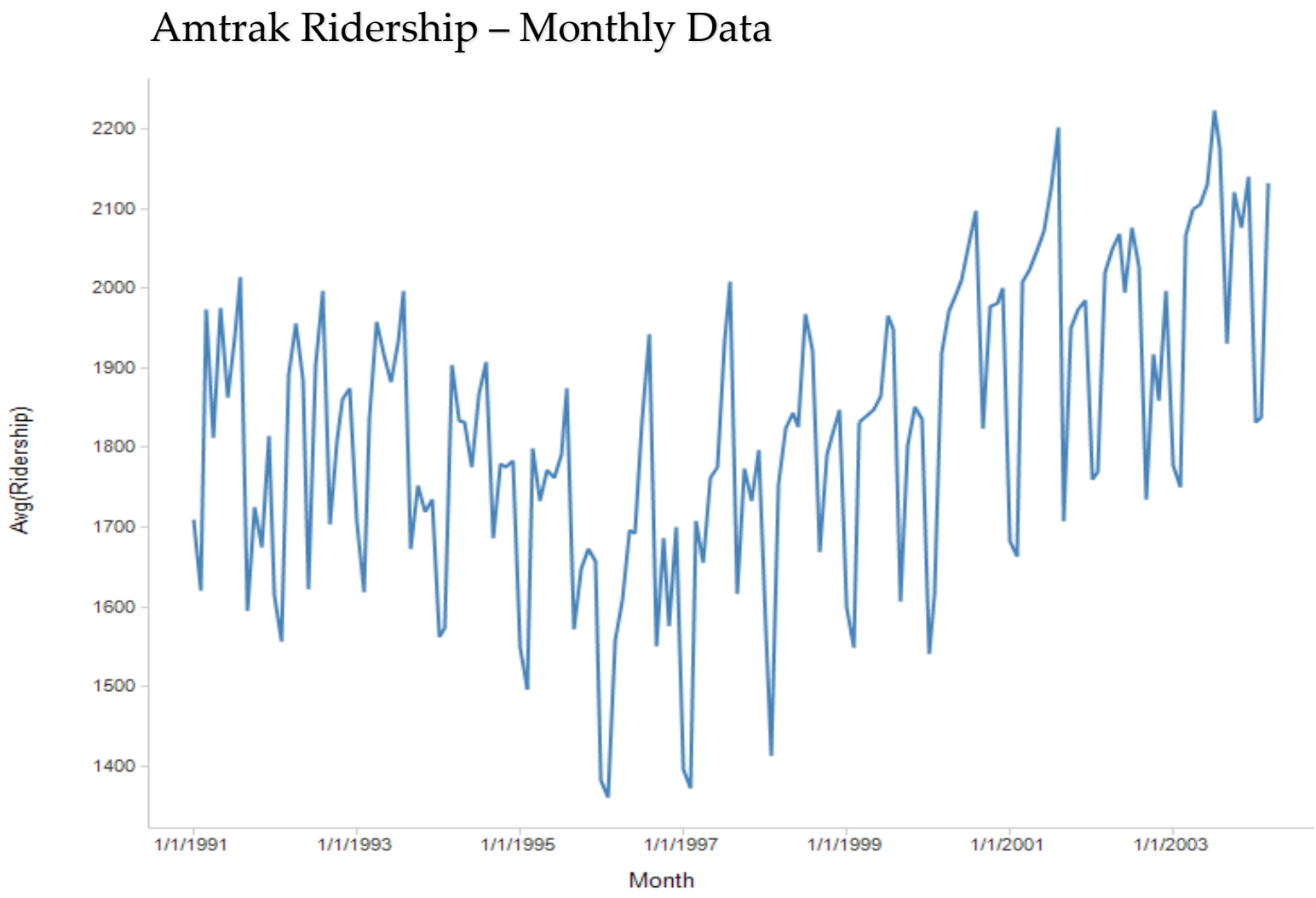




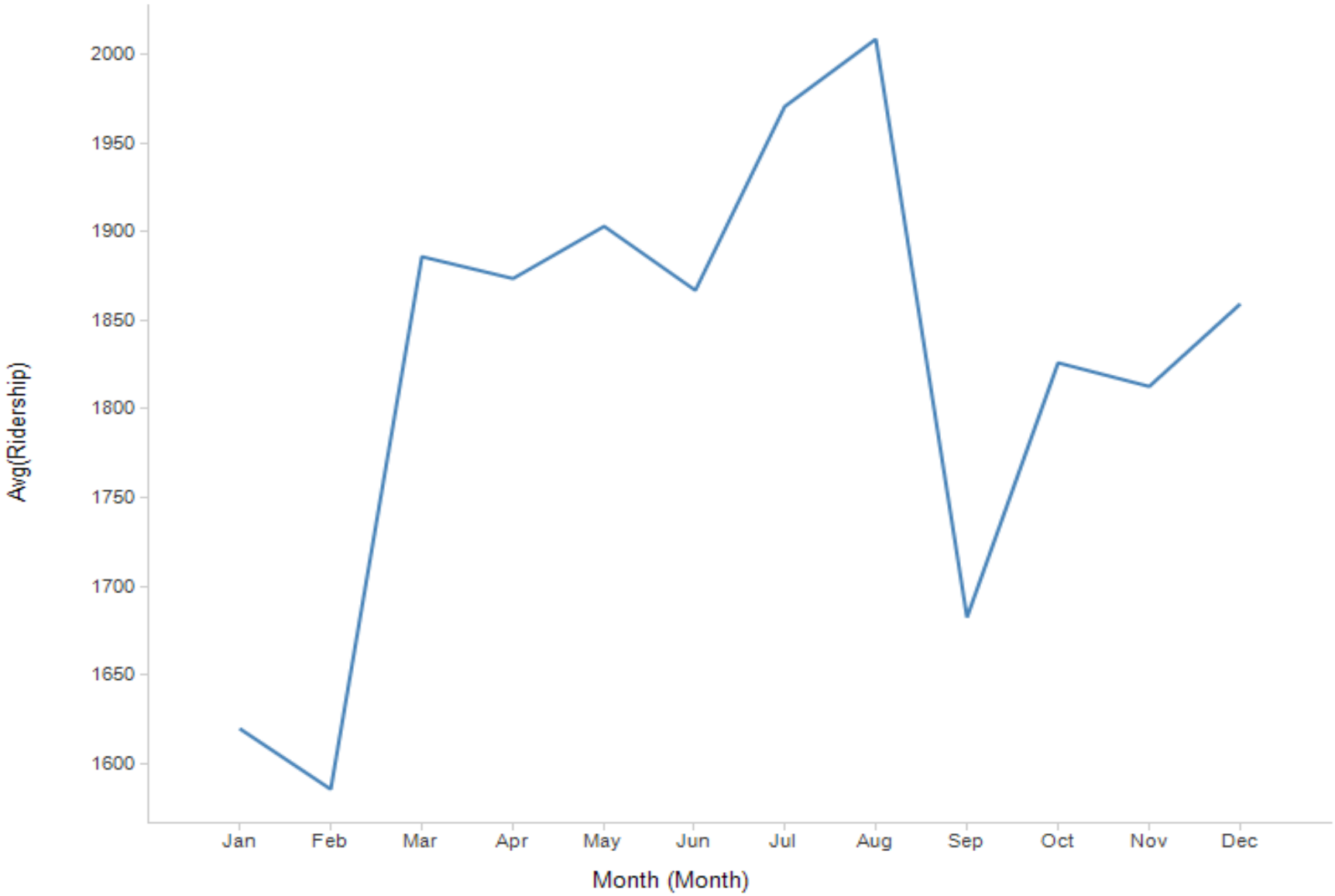


http://coronaboard.kr

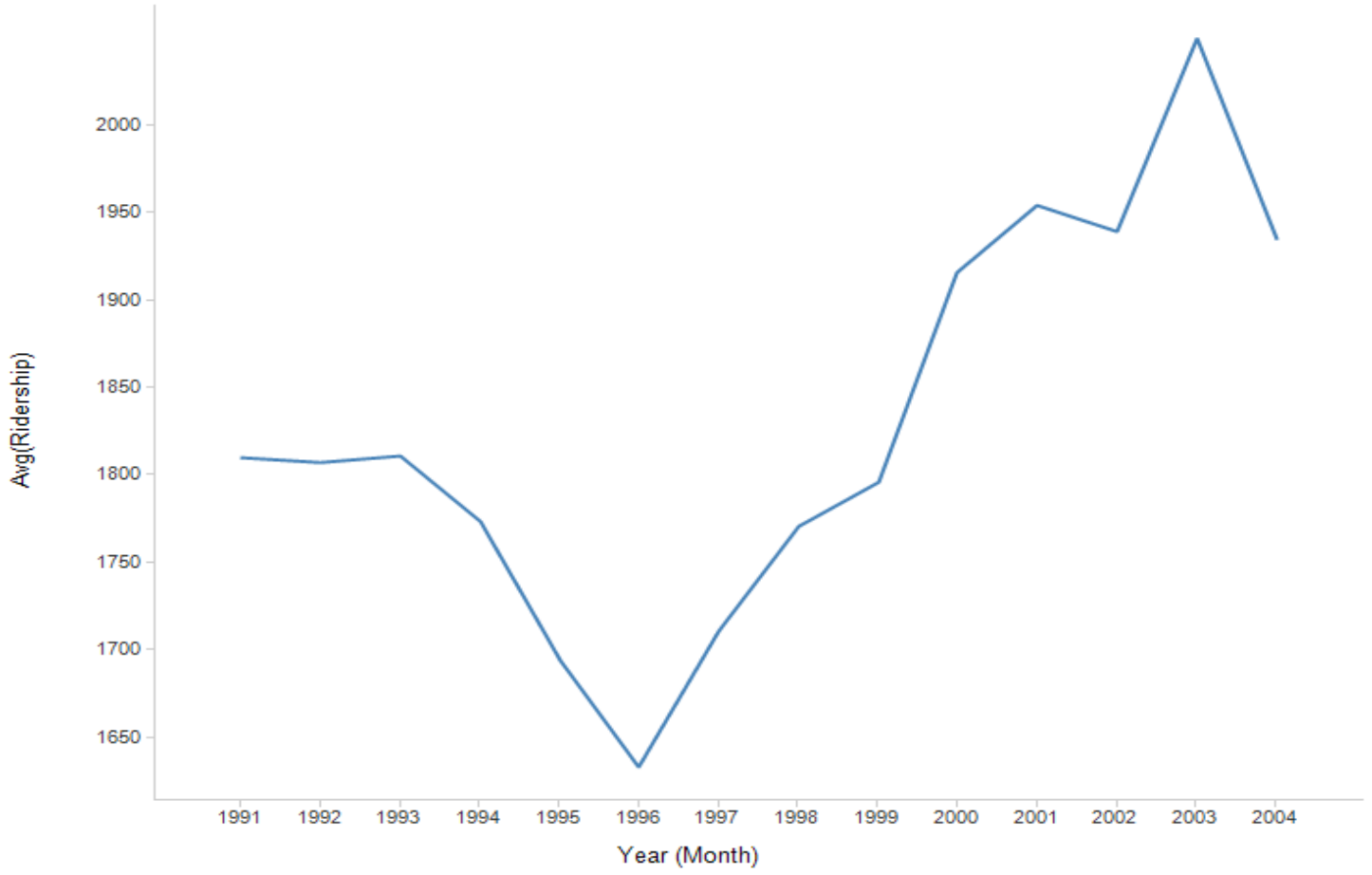
Aggregation



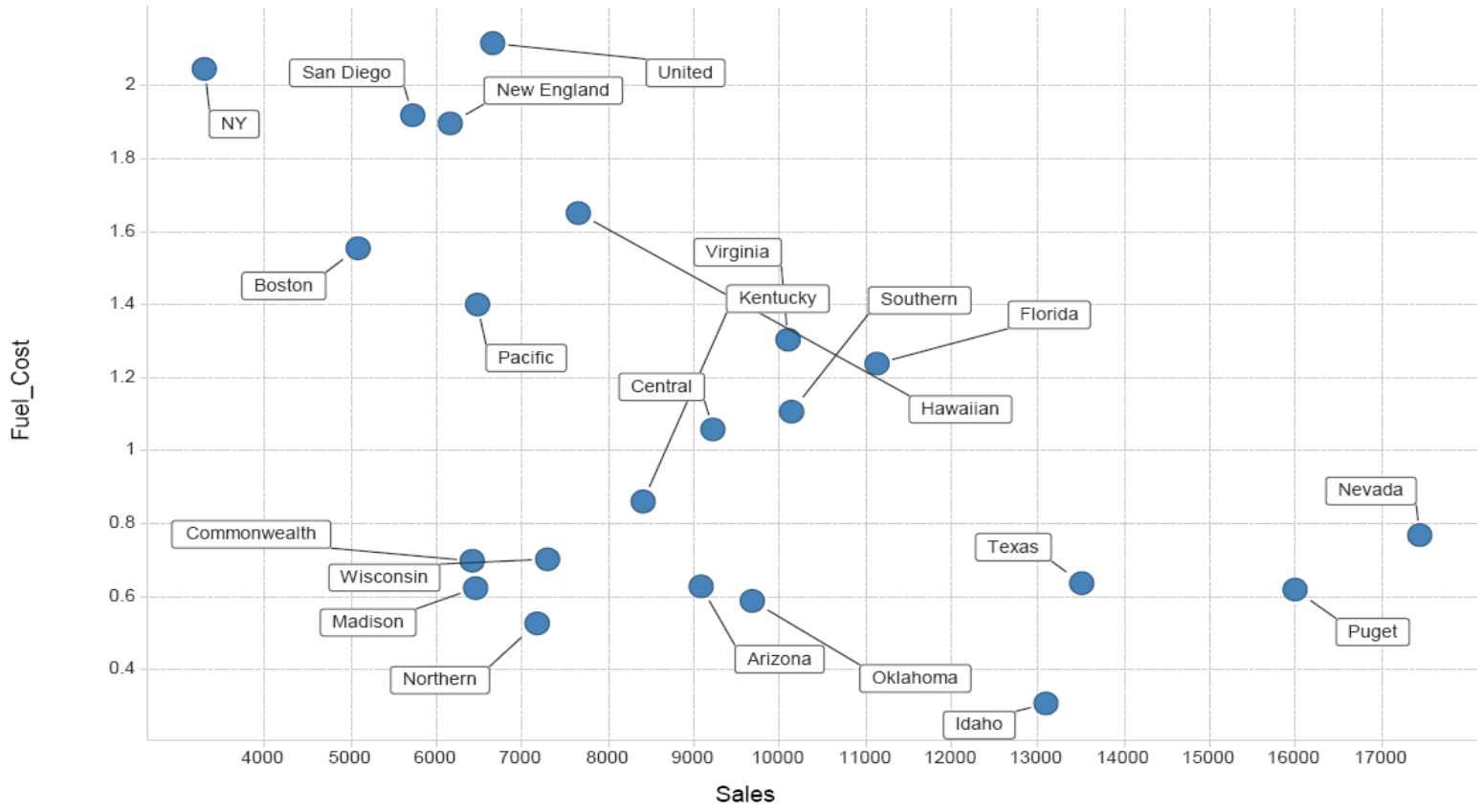
Aggregation – Monthly Average



Aggregation – Yearly Average



Scatter Plot with Labels (Utilities)

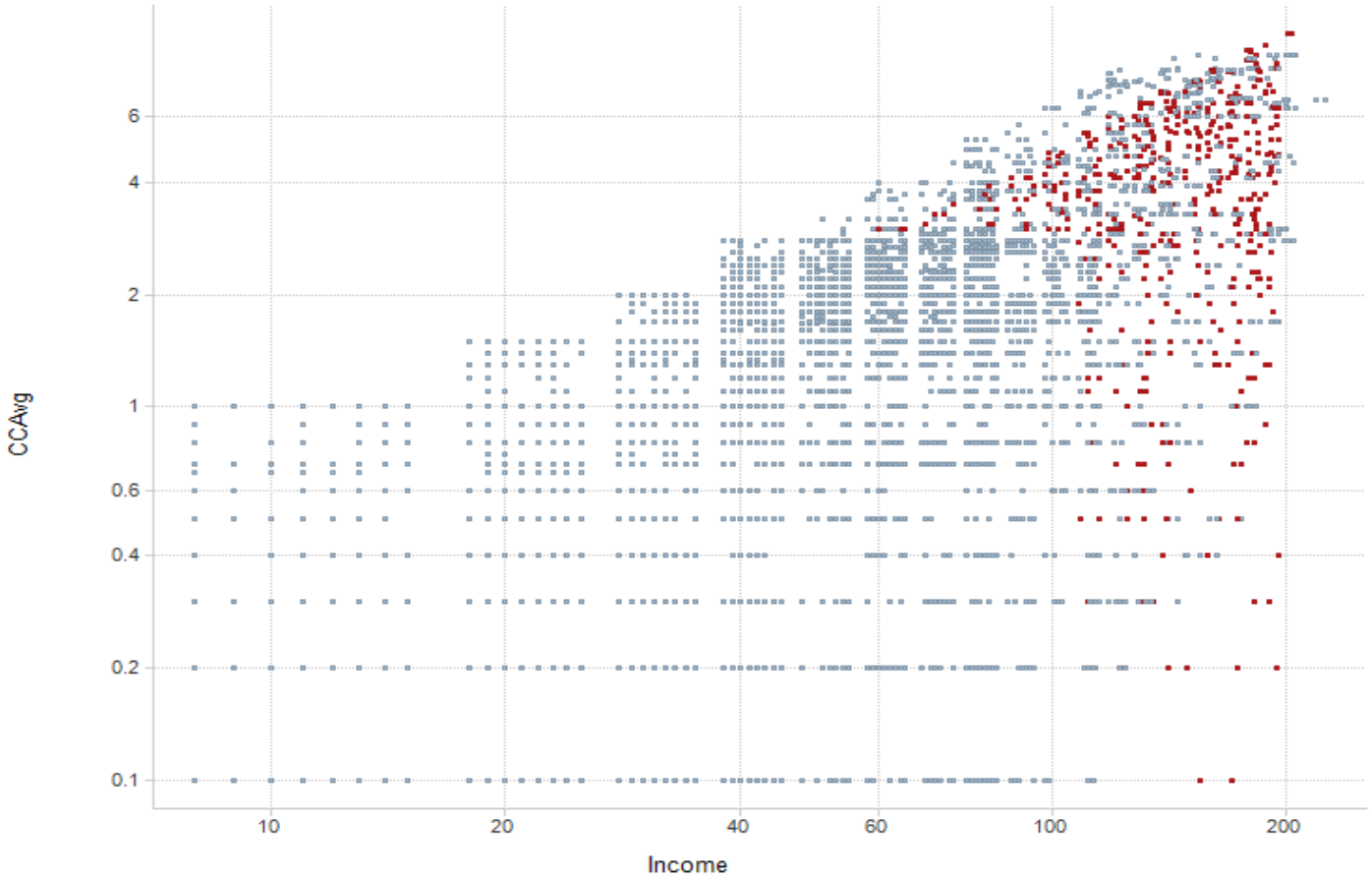


Scaling: Smaller markers, jittering, color contrast (Universal Bank; red = accept loan)

Jittering

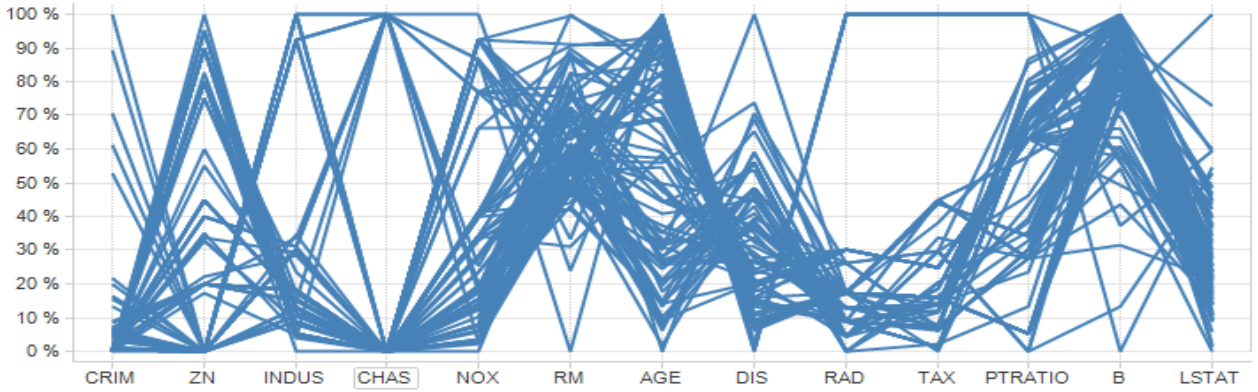
- Moving markers by a small random amount
- Uncrowds the data by allowing more markers to be seen

Without jittering (for comparison)

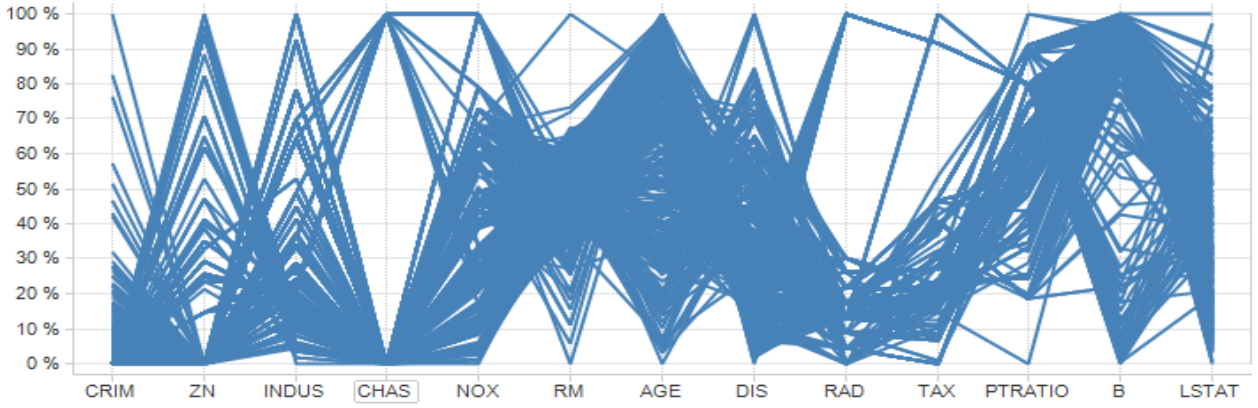


Parallel Coordinate Plot (Boston Housing)

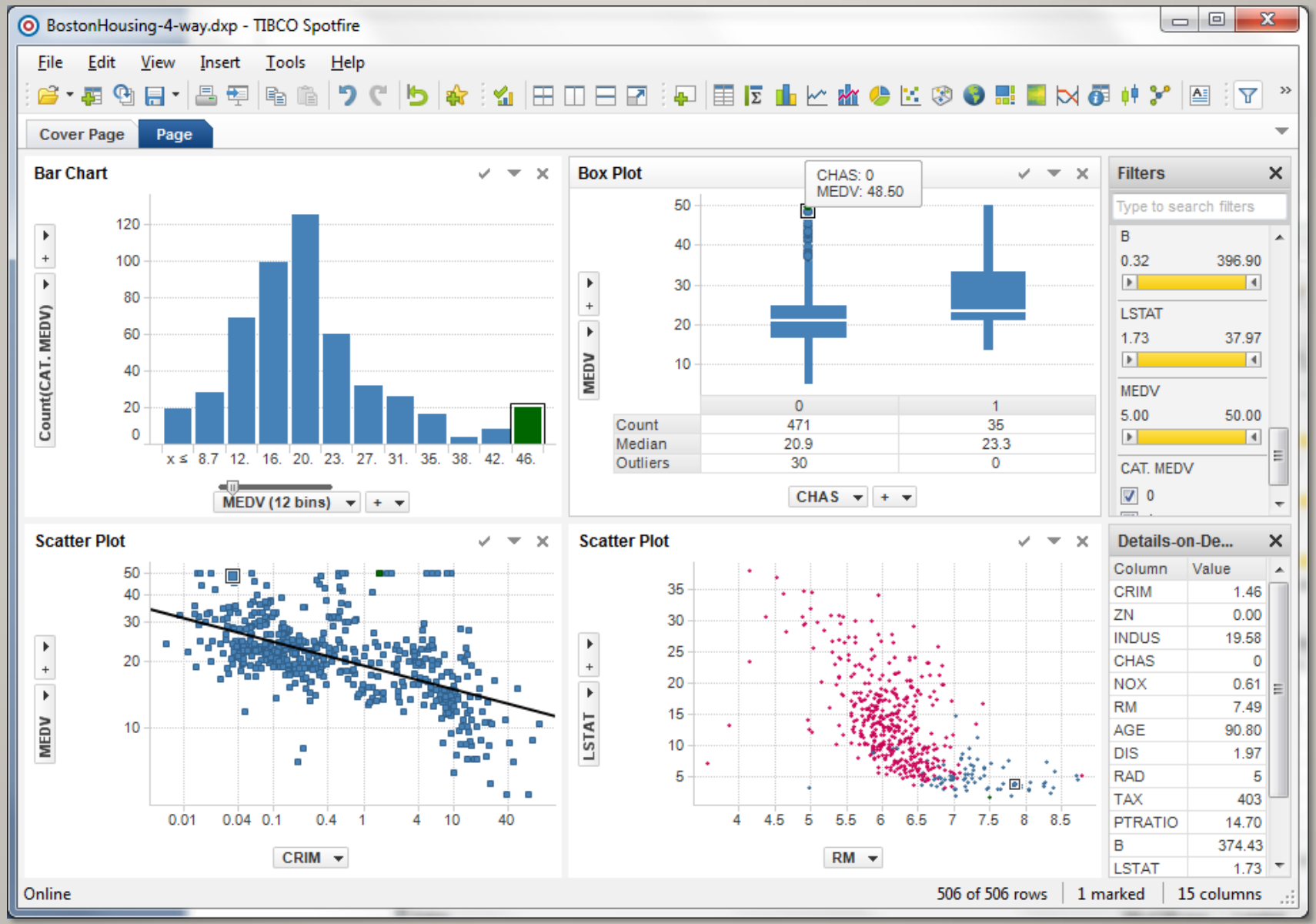
CATMEDV =1



CATMEDV =0



Linked plots
(same record is highlighted in each plot)



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

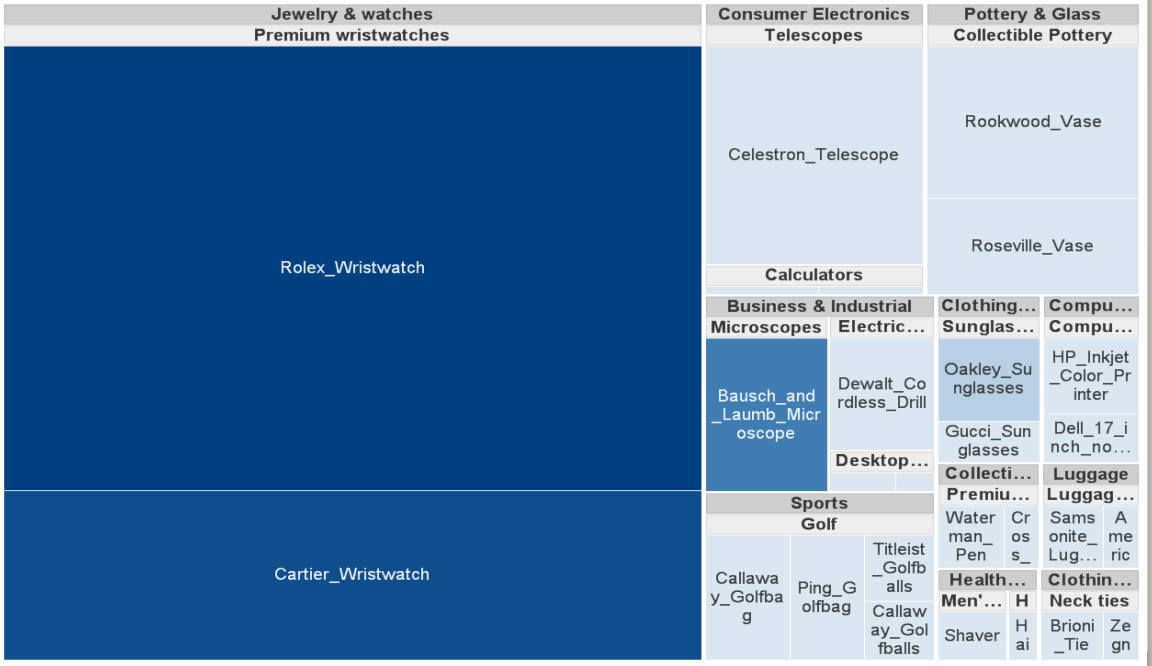
Circle size = # of transactions for the node

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

Arrows point from buyer to seller

Rectangle size = average closing price (=item value)

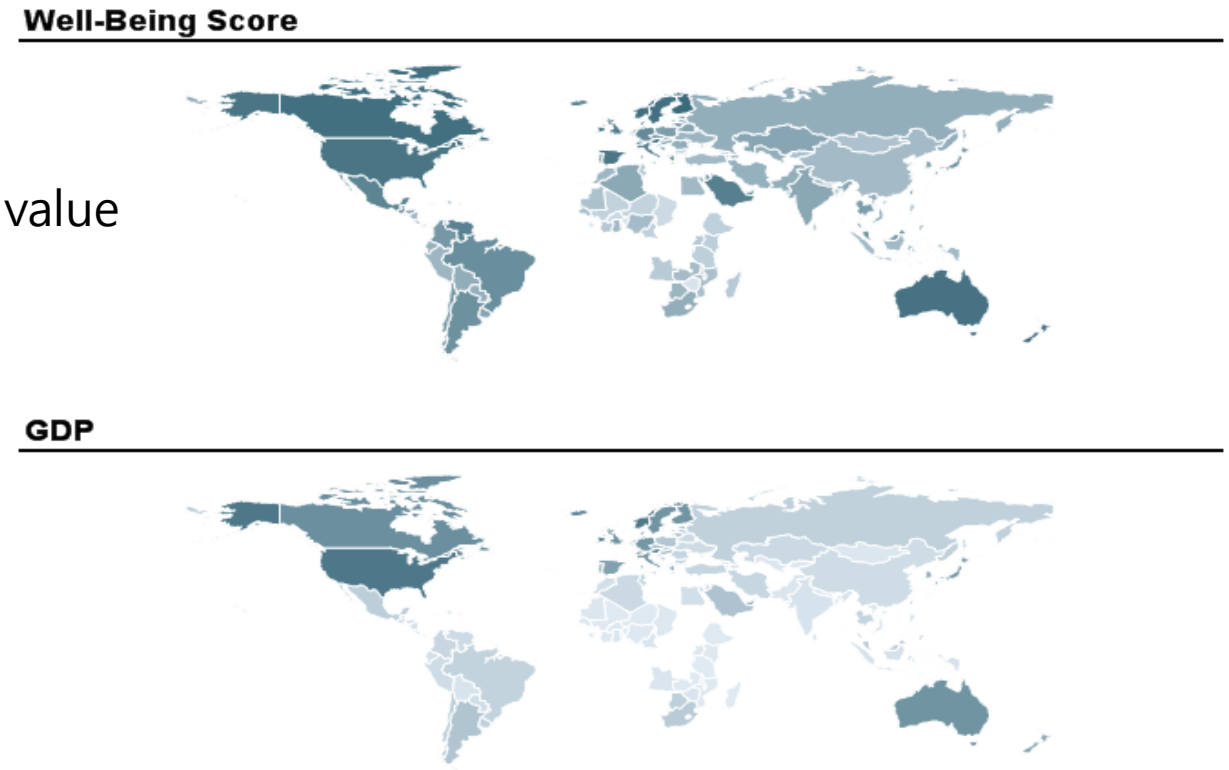
Color = % sellers with negative feedback (darker=more)



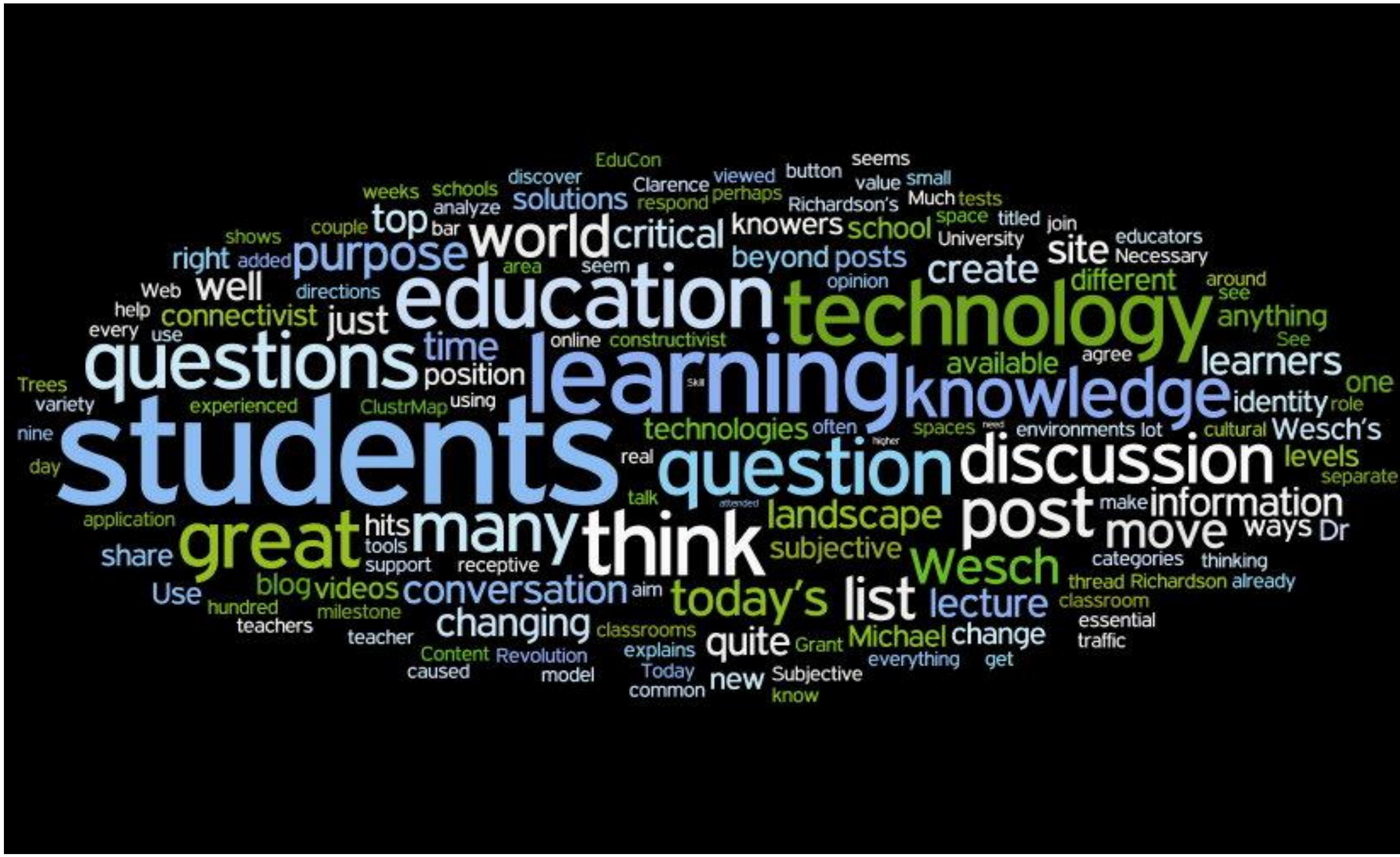
Map Chart

(Comparing countries' well-being with GDP)

Darker = higher value



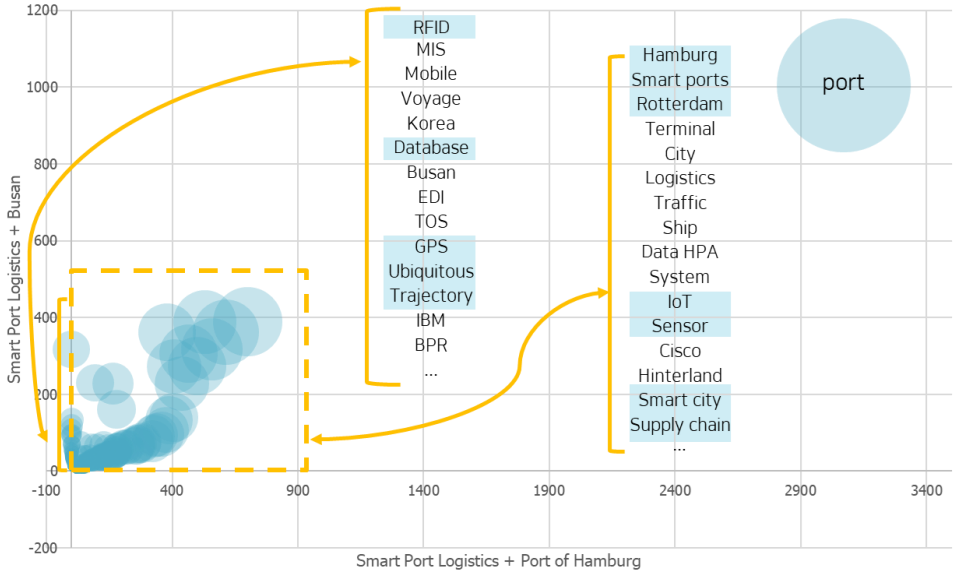
wordle


EduCon seems
weeks schools discover Clarence viewed button value small
analyze solutions respond perhaps Richardson's Much tests
shows couple top bar world critical knowers school space titled join
right added purpose area seem beyond posts University educators
Web well directions education technology create site Necessary
help connectivist just opinion different around see
every use time online constructivist anything See
questions learning available agree learners
Trees position knowledge one
variety experienced ClustrMap using identity role
nine students technologies often spaces environments lot cultural Wesch's
day real question discussion levels separate
application talk landscape post make information
share great hits many think subjective move ways Dr
tools support receptive Wesch categories thinking
Use blog videos conversation aim today's list lecture thread Richardson already
hundred milestone classroom
teachers changing classrooms essential
teacher Content Revolution explains quite Grant Michael change traffic
caused model Today new Subjective everything get
common know

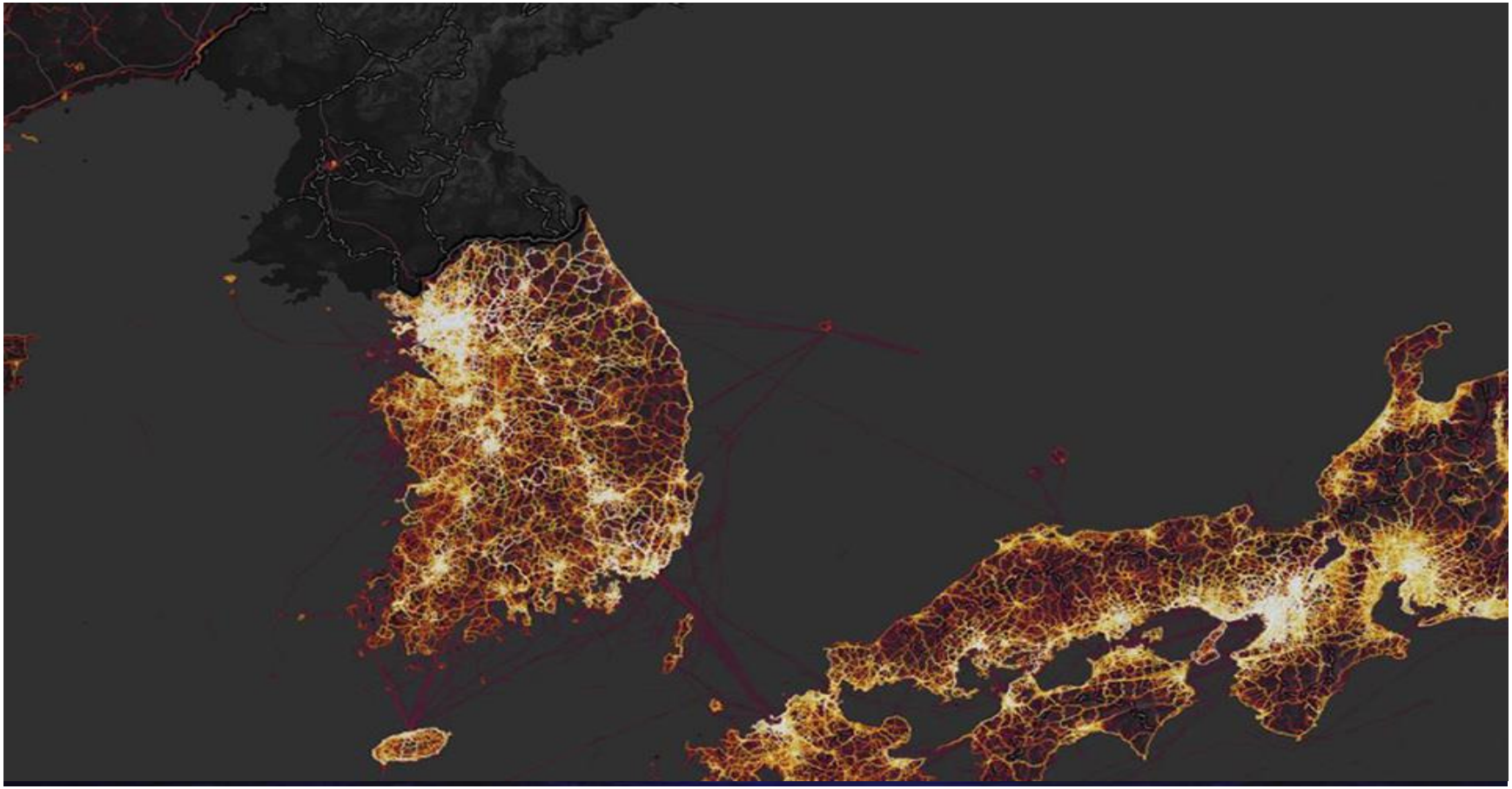
- 조사 기간 : 17.01.01 – 17.11.30 뉴스 기사 대상
- 주요 이슈
  1. 살인 불개미 발견 (불개미, 발견, 안전, 부두)
  2. 물동량 2000만TEU 기대 (물동량, 컨테이너, 처리,증가)
  3. 사드 보복에 따른 크루즈 이용객 급감 대책 마련 필요+ 크루즈 관광 활성화 (여객터미널, 크루즈, 북항)

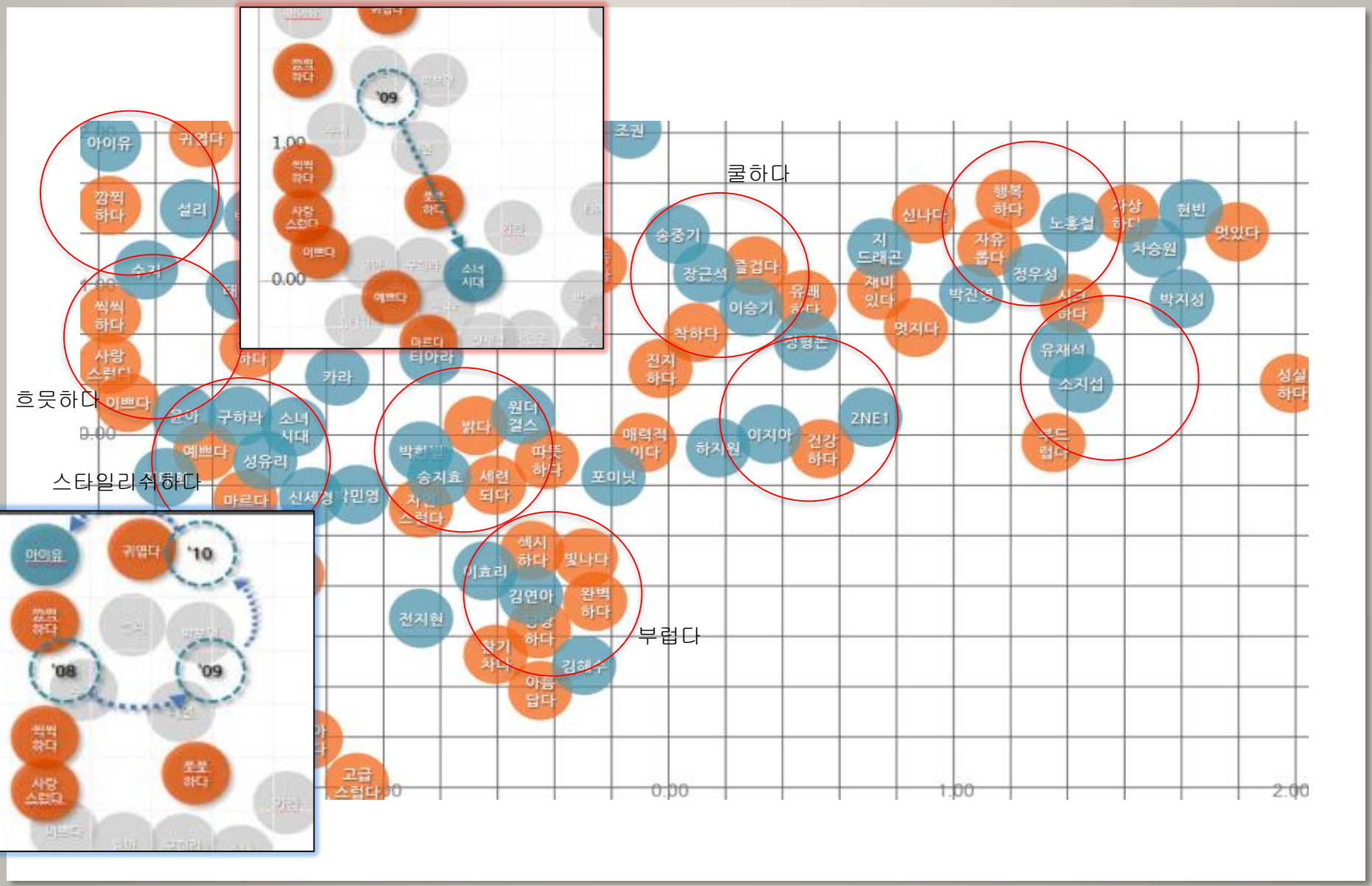
  기타 : 채용비리, 해수 온천, 한진해운 파산 여파

❖ Busan VS Hamburg



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





hrbae@pusan.ac.kr

LineUp


Demos Evaluation Window
World University Ranking 2012
Ra
School Name
Country
Academic reput
Employer reputa
Faculty/student r
Citations per fac
International fac
International stu
Filter: <None>
1. Aalborg University Denmark
2. Aalto University Finland
3. Aarhus University Denmark
4. Aberystwyth University United Kingdom
5. Abo Akademi University Finland
6. Al-Farabi Kazakh National University Kazakhstan
7. Al-Imam Mohamed ibn Saud Islamic University Saudi Arabia
8. American University in Cairo Egypt
9. American University of Beirut (AUB) Lebanon
10. American University of Sharjah (AUS) United Arab Emi
11. Aristotle University of Thessaloniki Greece
12. Arizona State University United States
13. Aston University United Kingdom
14. Ateneo de Manila University Philippines
15. Auckland University of Technology New Zealand
16. Australian National University Australia
17. Bandung Institute of Technology (ITB) Indonesia
18. Bangor University United Kingdom
19. Bauman Moscow State Technical University Russia
20. Beijing Institute of Technology China
21. Beijing Normal University China
22. Ben Gurion University of The Negev Israel
23. Bilkent University Turkey
24. Birkbeck College, University of London United Kingdom
25. Bond University Australia
26. Boston College United States
27. Boston University United States
28. Brandeis University United States
29. Brown University United States
30. Brunel University United Kingdom
31. California Institute of Technology (Caltech) United States
32. Cardiff University United Kingdom
33. Carnegie Mellon University Australia
34. Case Western Reserve University United States
35. Chalmers University of Technology Sweden
36. Chang Gung University Taiwan
37. Charles University Czech Republic
38. Chiba University Japan
39. Christian-Albrechts-Universität zu Kiel Germany
40. Chulalongkorn University Thailand
41. City University London United Kingdom
42. City University of Hong Kong Hong Kong S.A.
43. City University of New York United States
44. Colorado State University United States
Arts & Humaniti
Engineering & T
Life Sciences &
Natural Science

Data visualization tools

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

Prof. Hyerim Bae (hrbae@pusan.ac.kr)

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