BAE Lab Bigdata Analytics & Engineering 부산대학교 PUSAN NATIONAL UNIVERSITY

데이터 시각화
(Data Visualization)

Hyerim Bae
Department of Industrial Engineering, Pusan National University
hrbae@pusan.ac.kr

Contents

한 장의 사진이 천 마디 말보다 낫다.

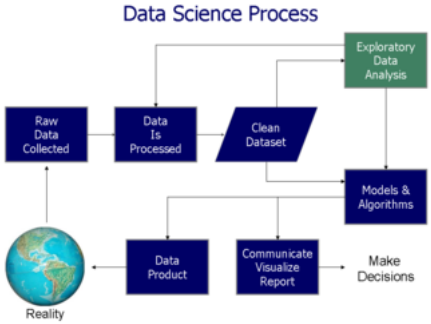
탐색적 데이터 분석(Explorative Data Analysis, EDA)

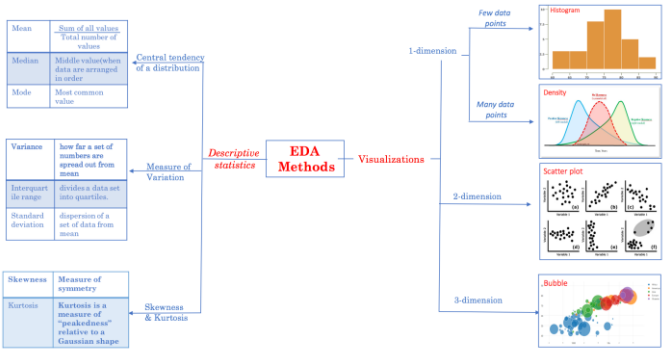
1. **예상되는 관계가** 데이터에 실재로 존재하는지 확인하고, 계획된 분석을 검증

2. 고려해야 할 데이터에서 **예상치 못한 관계를 찾아** 계획된 분석의 일부 변경 사항을 제안할 수 있음

3. 해당 데이터가 내포하는 실제 프로세스를 비즈니스 관계자가 제대로 이해하고 있는지 확인하고, **데이터 기반 통찰력을 제공**

4. 주어진 **문제에 대한 맥락을** 제공함

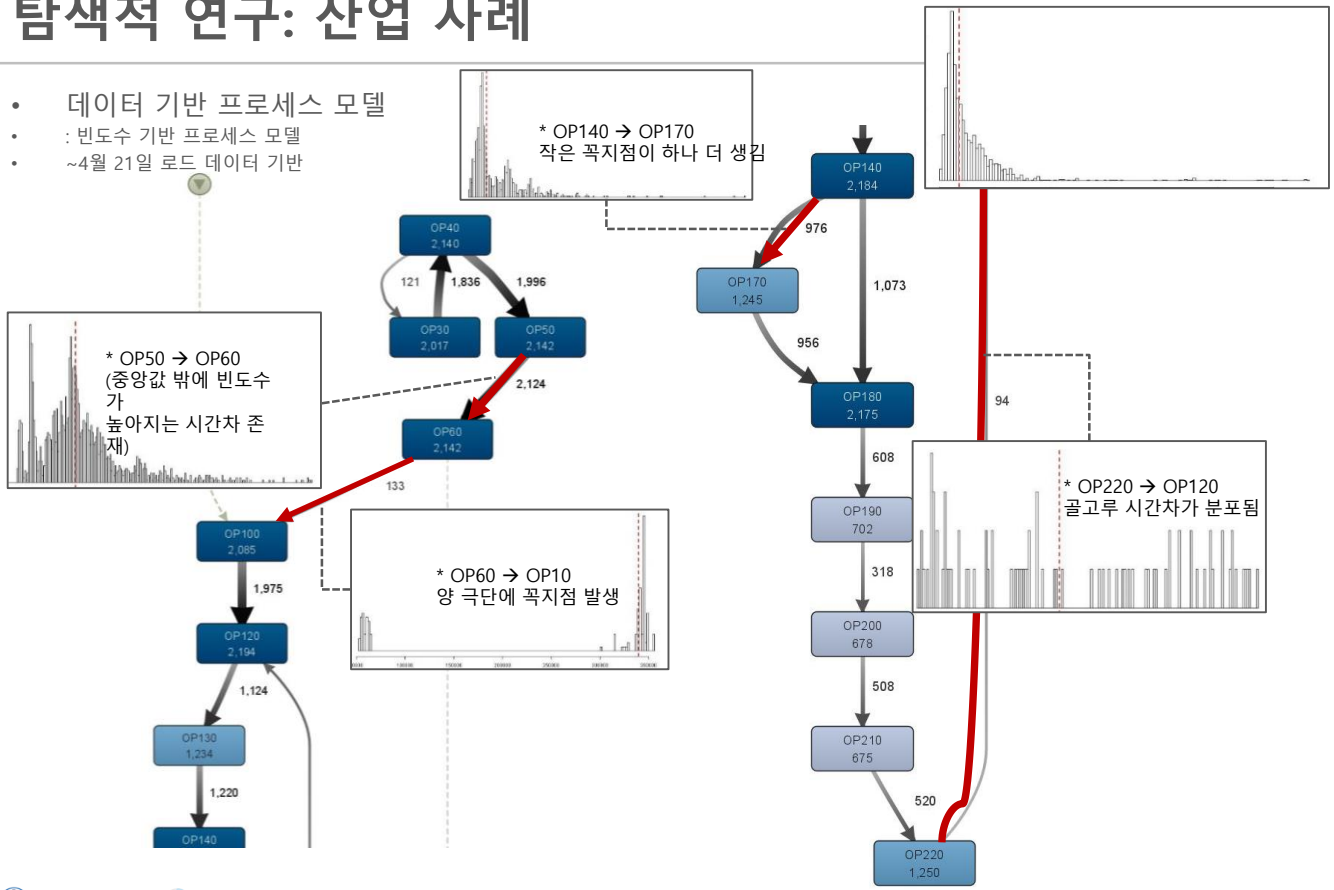




* 나머지 작업(OP) 간 시간 분포는 다음과 유사한 꼴을 보임
(중앙값을 중심으로 점차적으로 빈도수가 감소하는 꼴)

탐색적 연구: 산업 사례

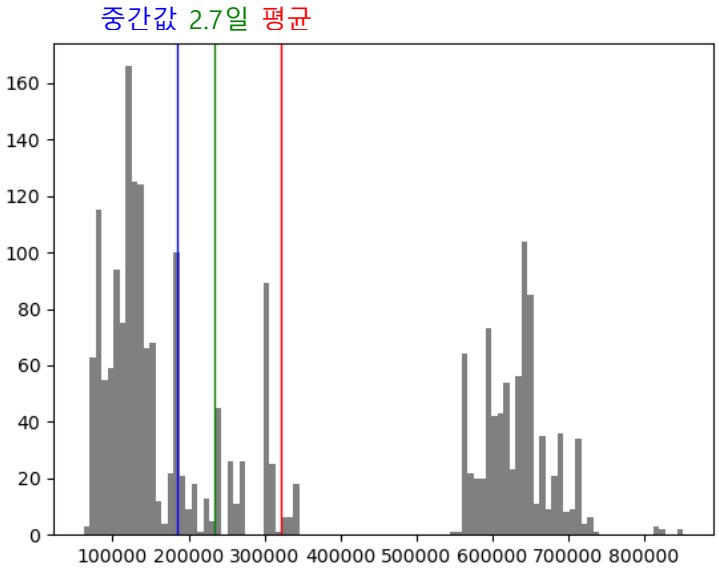
- 데이터 기반 프로세스 모델
- : 빈도수 기반 프로세스 모델
- ~4월 21일 로드 데이터 기반



탐색적 연구: 산업 사례

- 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)

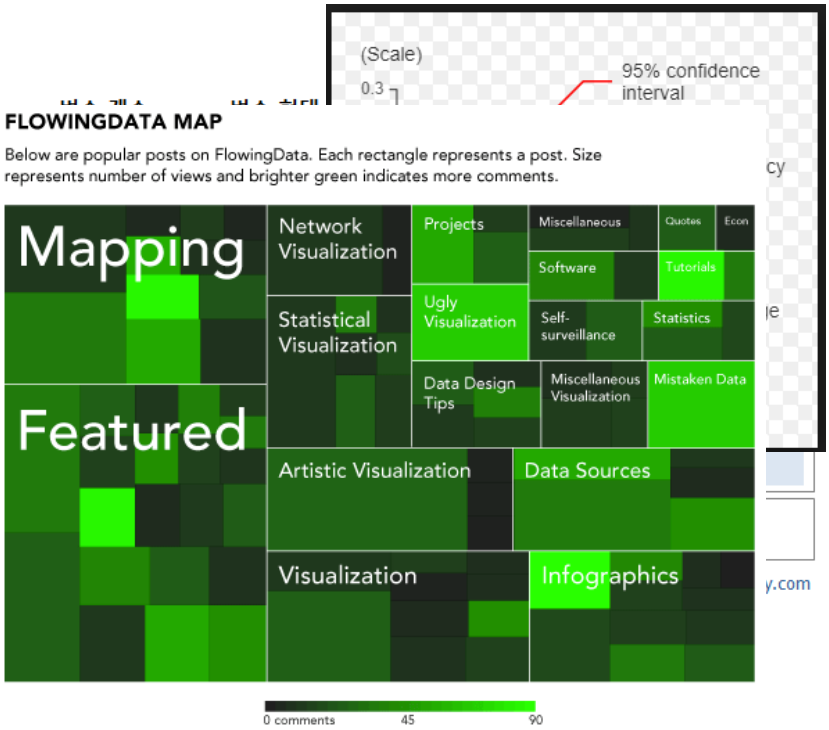
OP60 → OP100 평균 3.6일



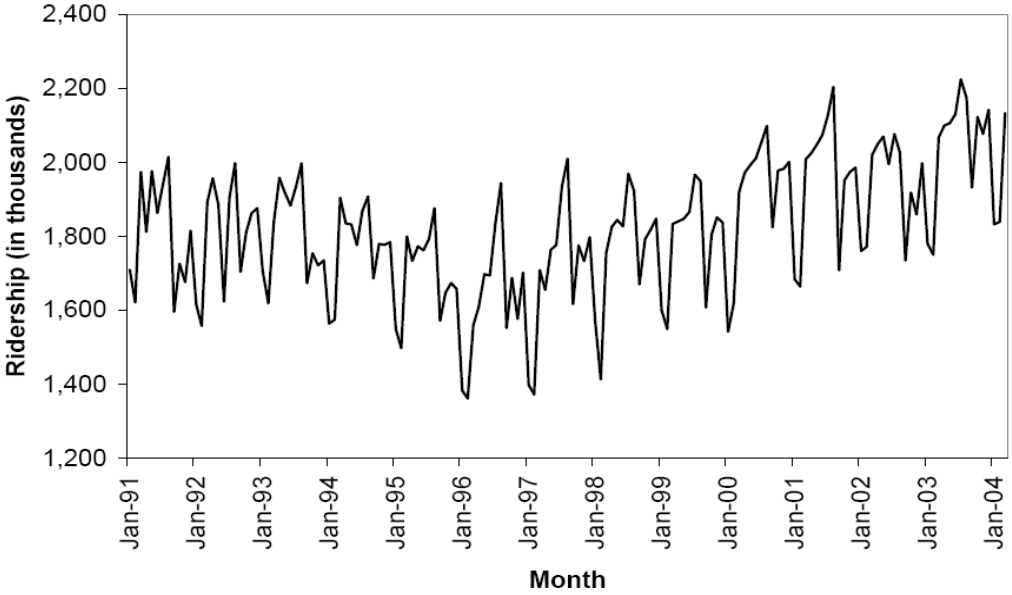
데이터 탐색을 위한 그래프

기본 그래프
- 선 그래프
- 막대 그래프
- 산점도

분포를 나타내는 그래프
- Boxplots
- 히스토그램



시계열 데이터 표현 – 선 그래프



Boston Housing 데이터 설명

CRIM: 도시 별 1인당 범죄율

ZN: 25,000 평방피트를 초과하는 거주지역의 비율

INDUS: 비소매상업지역이 점유하고 있는 토지 비율

CHAS: 찰스 강에 대한 더미 변수(강의 경계에 위치한 경우 1, 아니면 0)

NOX: 10ppm 당 농축 일산화질소

RM: 주택 1가구당 평균 방의 개수

AGE: 1940년 이전에 건축된 소유주택의 비율

DIS: 5개의 보스턴 직업센터까지의 접근성 지수

RAD: 방사형 도로까지의 접근성 지수

TAX: 10,000 달러 당 재산세율

PTRATIO: 도시 별 학생/교사 비율

B: 도시 별 흑인의 비율

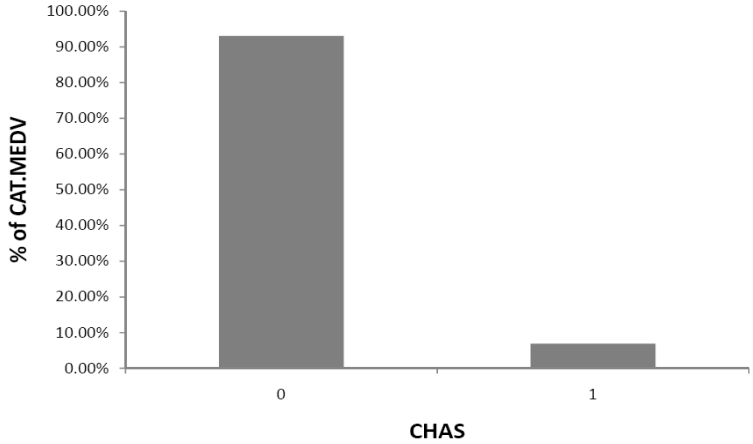
LSTAT: 하위계층 비율

MEDV: 본인 소유 주택 가격(중앙값, 단위 : $1,000)

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

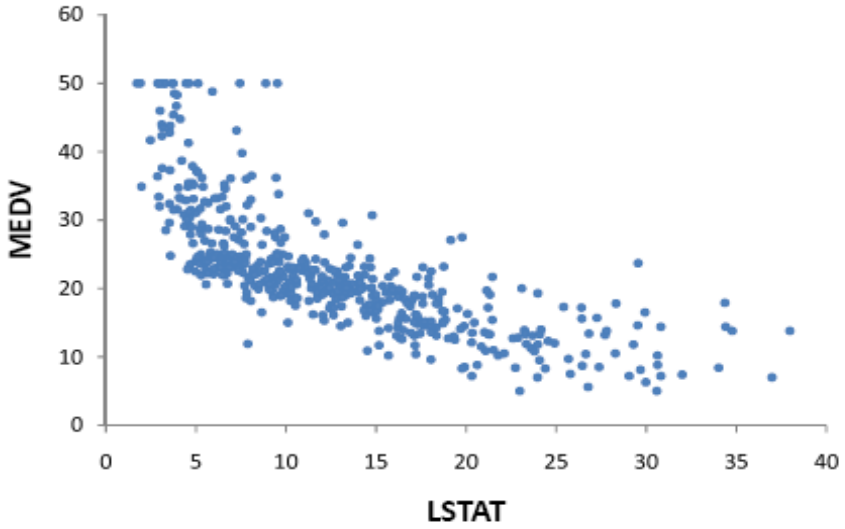
범주형 변수 표현 – Bar Chart

전체 지역의 95%가 찰스강과 접하지 않음



산점도(Scatterplot)

두 연속형(수치형) 변수 간의 관계를 표현

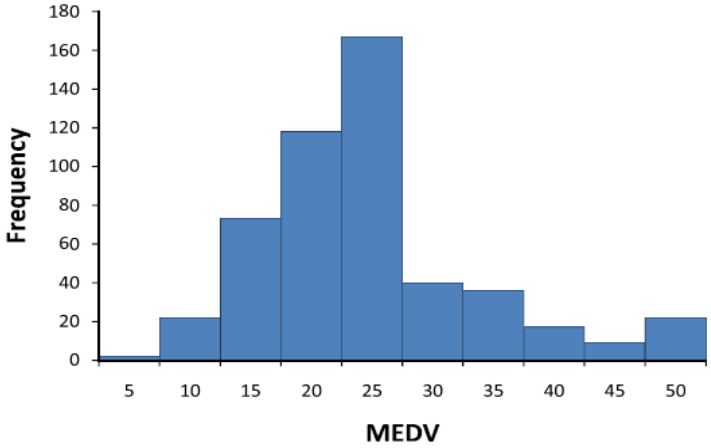


분포도(Distribution Plots)

- 범주형 데이터 셋 내에서 발생하는 값들이 “얼마나 많은지”를 시각적으로 표현
- 또는, 연속형(수치형) 데이터 셋의 경우, 특정 범위 내에 얼마나 많은 값이 분포하는지 시각적으로 표현

히스토그램(Histograms)

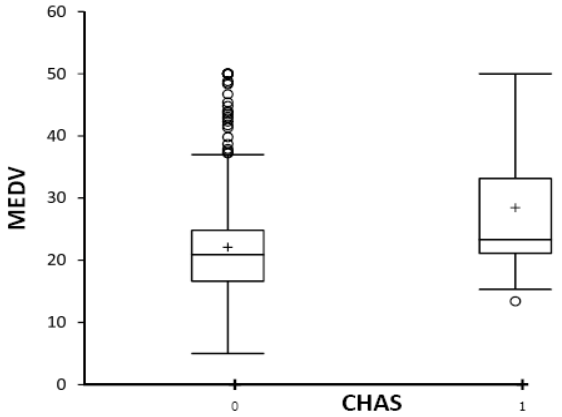
히스토그램은 출력변수의 분포를 표현함(ex. 집 값의 중앙값)



박스플롯(Boxplots)

Boston Housing 예제:
찰스강에 대한 더미변수 값 분포 표현

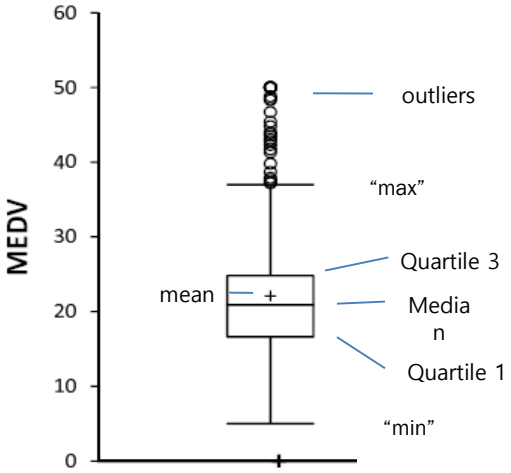
- Boxplot은 범주형 변수의 하위그룹을 비교하는데 유용함



Box Plot

- 이상치 기준(상한): Q3+1.5(Q3-Q1) 보다 큰 값

- 상한(max) = 정상 범주의 최대값
- 하한(min) = 정상 범주의 최소값은 소프트웨어 마다 다를 수 있음

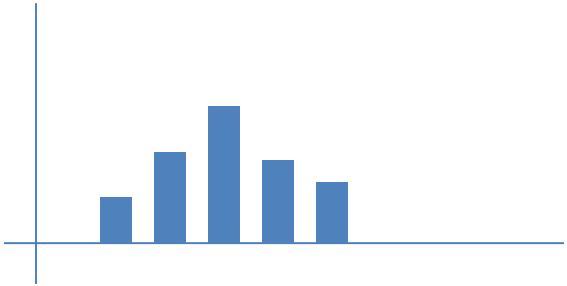


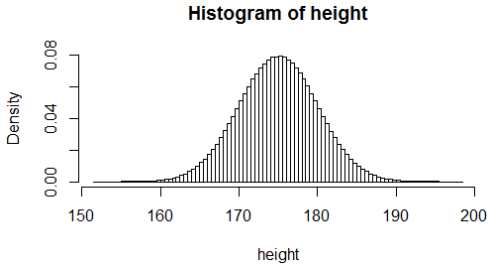
정규분포(Normal Distribution)

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

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





히트맵(Heat Maps)

히트맵에 나타나는 색깔을 통해 정보 전달

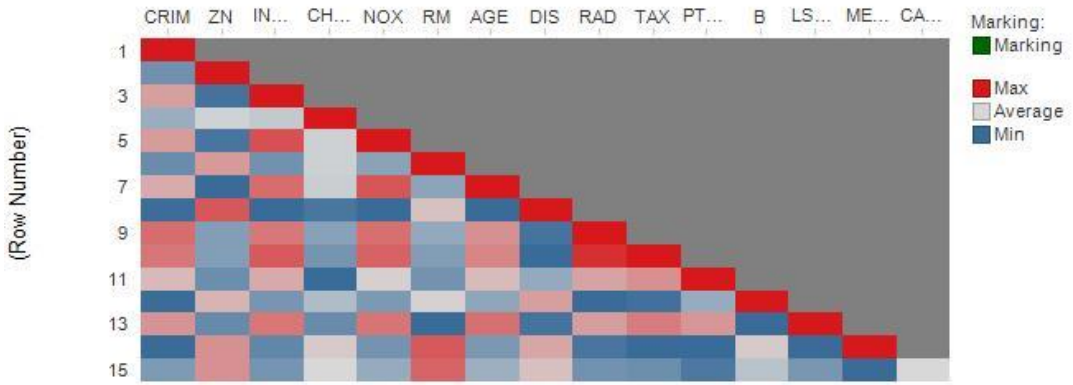
데이터 마이닝에서 변수 간 상관관계 및 결측치를 시각화 하는데 사용 됨

상관관계를 표현하기 위한 히트맵(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



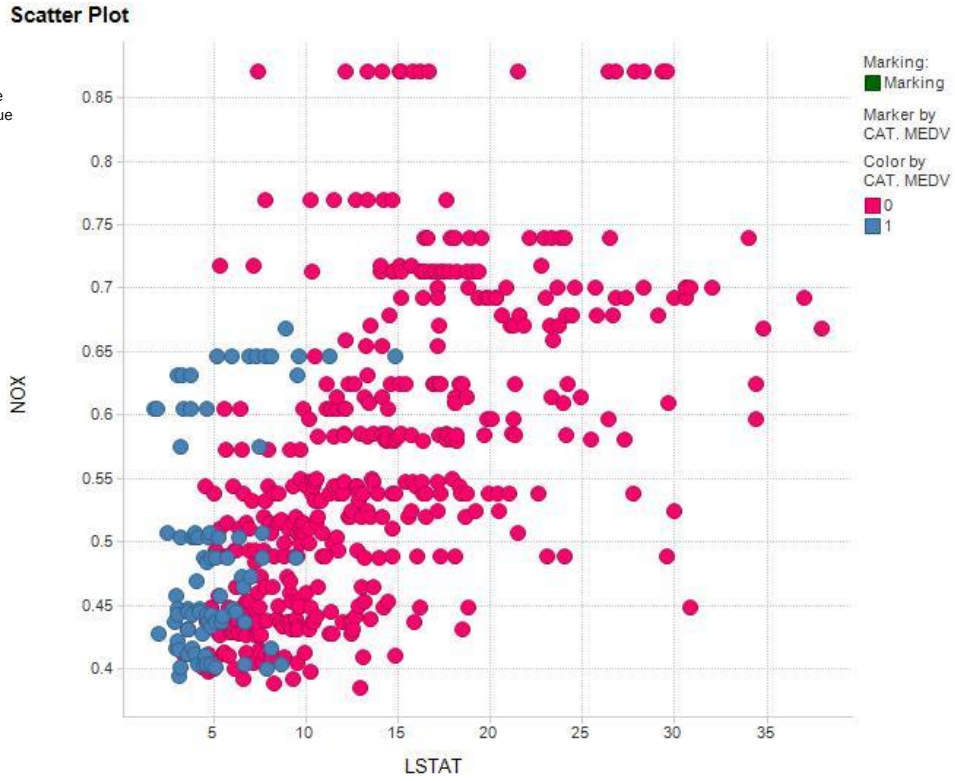
In Spotfire

다차원 시각화

색상을 활용한 산점도

Boston Housing

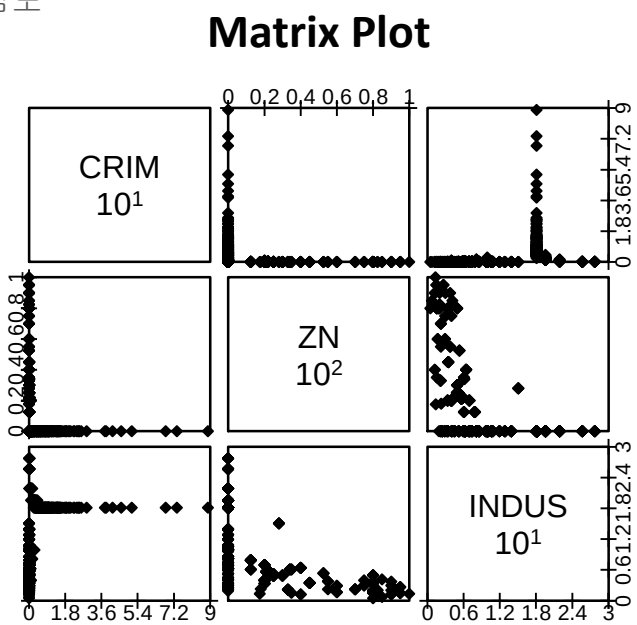
NOX vs. LSTAT
Red = low median value
Blue = high median value



매트릭스 플롯(Matrix Plot)

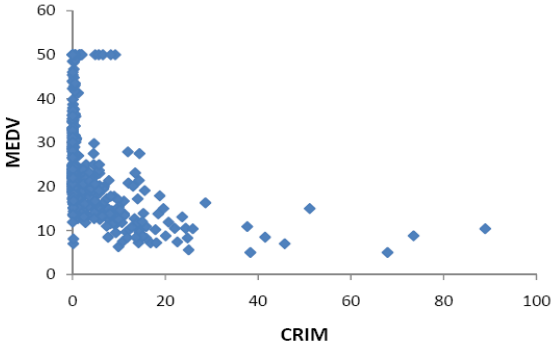
각 변수 쌍에 대한 산점도 표현

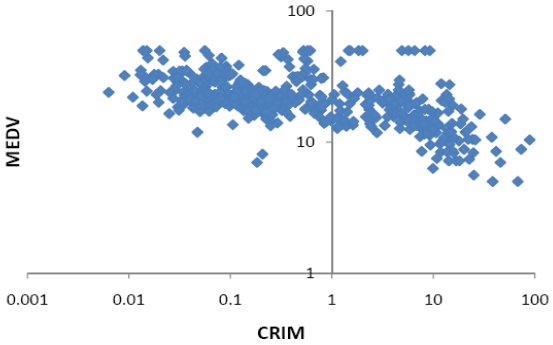
Example: Boston Housing 데이터 내의 변수 간 산점도

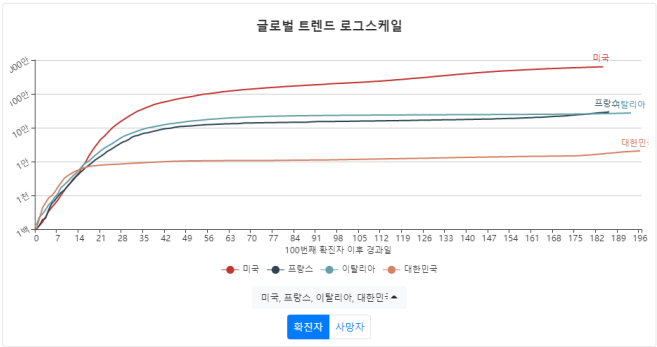


Log 함수를 활용한 Rescaling(재조정)

- 분포가 집중된 데이터

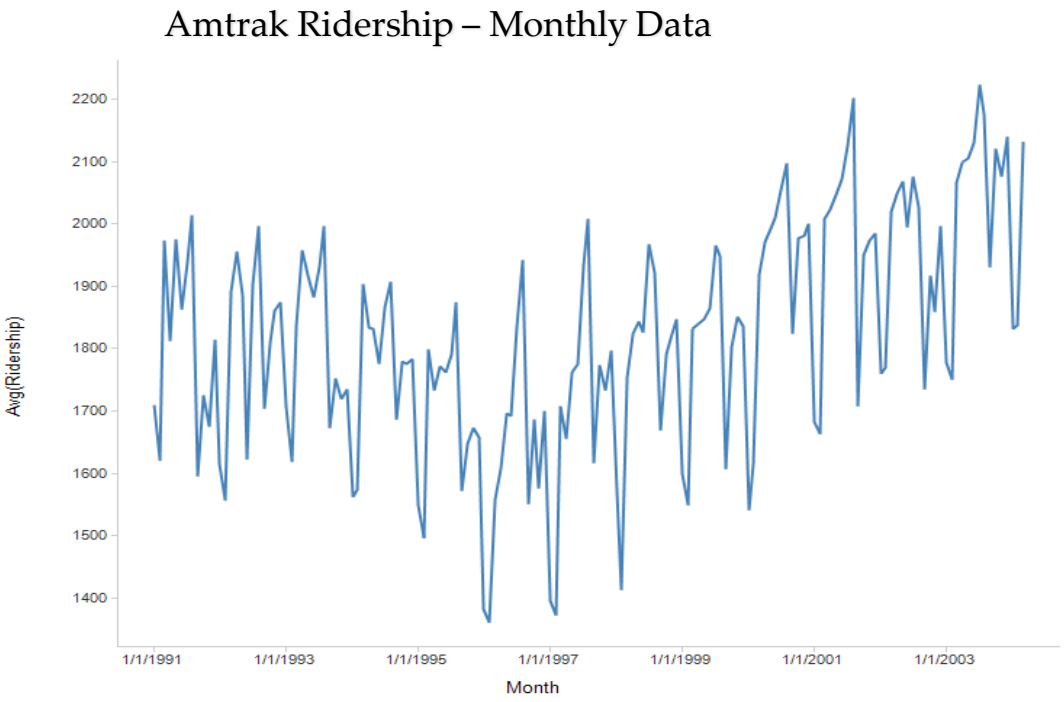






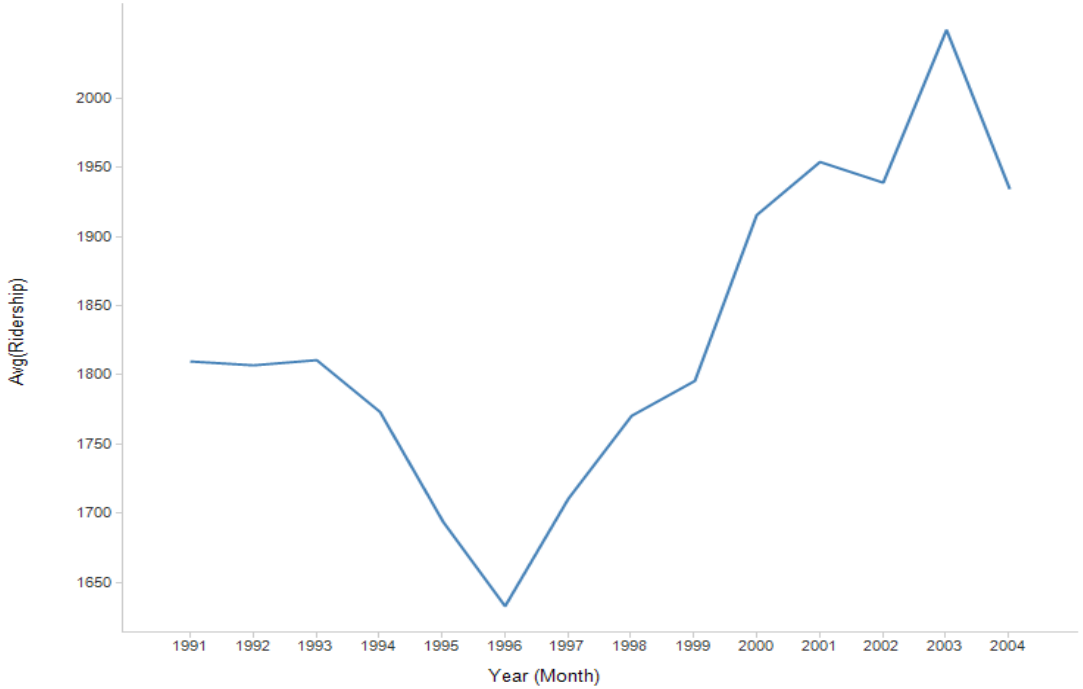
http://coronaboard.kr

Aggregation(집계)

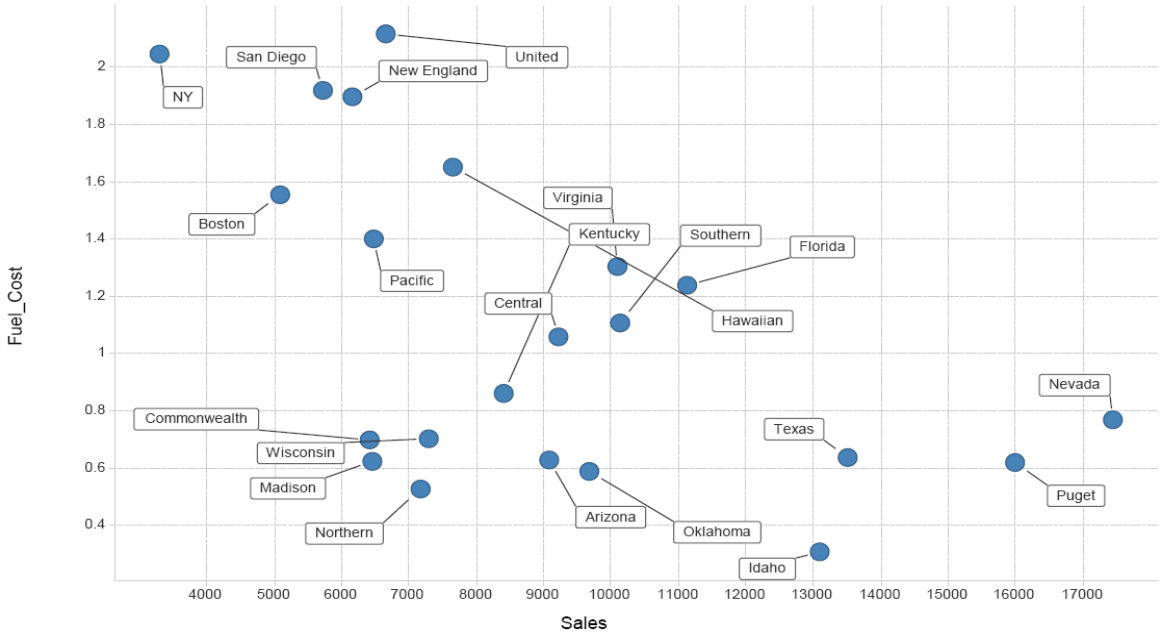


Aggregation – 월별 평균

Aggregation – 연도별 평균



레이블을 포함한 산점도

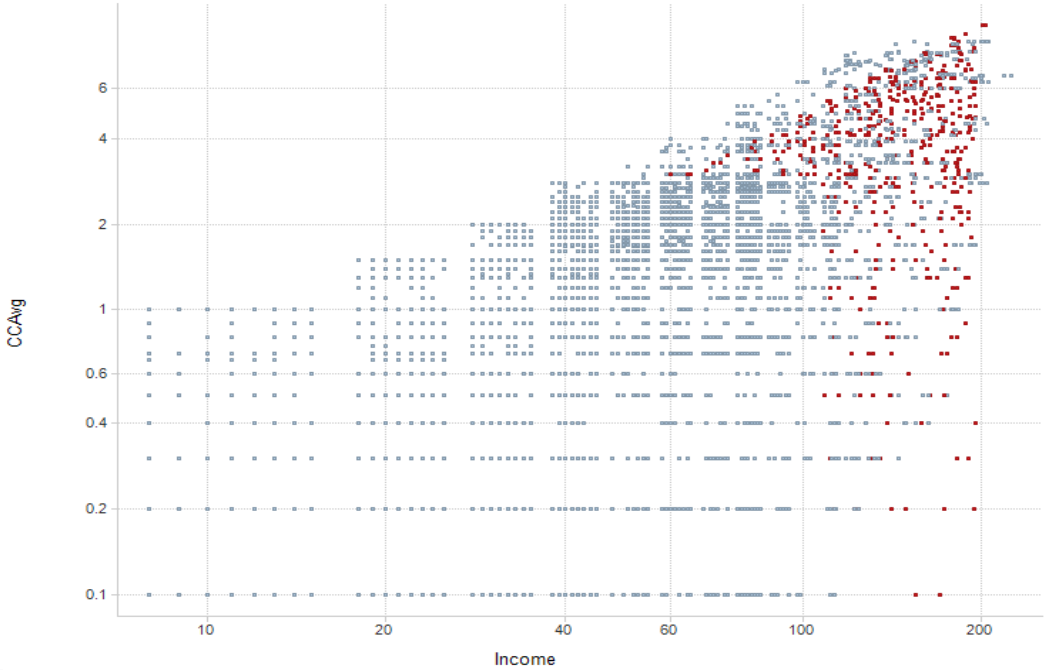

Fuel_Cost
Sales
2
1.8
1.6
1.4
1.2
1
0.8
0.6
0.4
4000
5000
6000
7000
8000
9000
10000
11000
12000
13000
14000
15000
16000
17000
San Diego
United
New England
NY
Boston
Virginia
Kentucky
Southern
Florida
Pacific
Central
Hawaiian
Nevada
Commonwealth
Texas
Wisconsin
Madison
Puget
Northern
Arizona
Oklahoma
Idaho

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

지터링(Jittering)

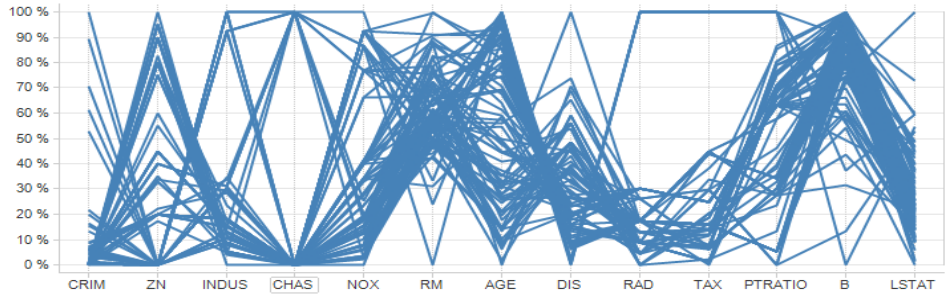
- 마커에 약간의 노이즈를 더하여 움직임
- 더 많은 마커를 볼 수 있도록 데이터를 퍼뜨림

지터링을 적용하지 않은 산점도(비교용)

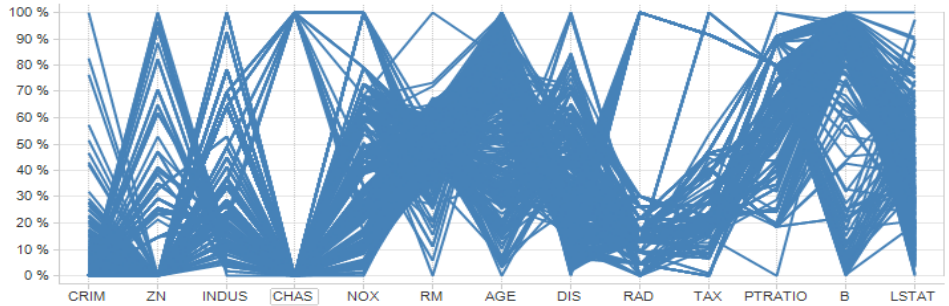


Parallel Coordinate Plot (Boston Housing)

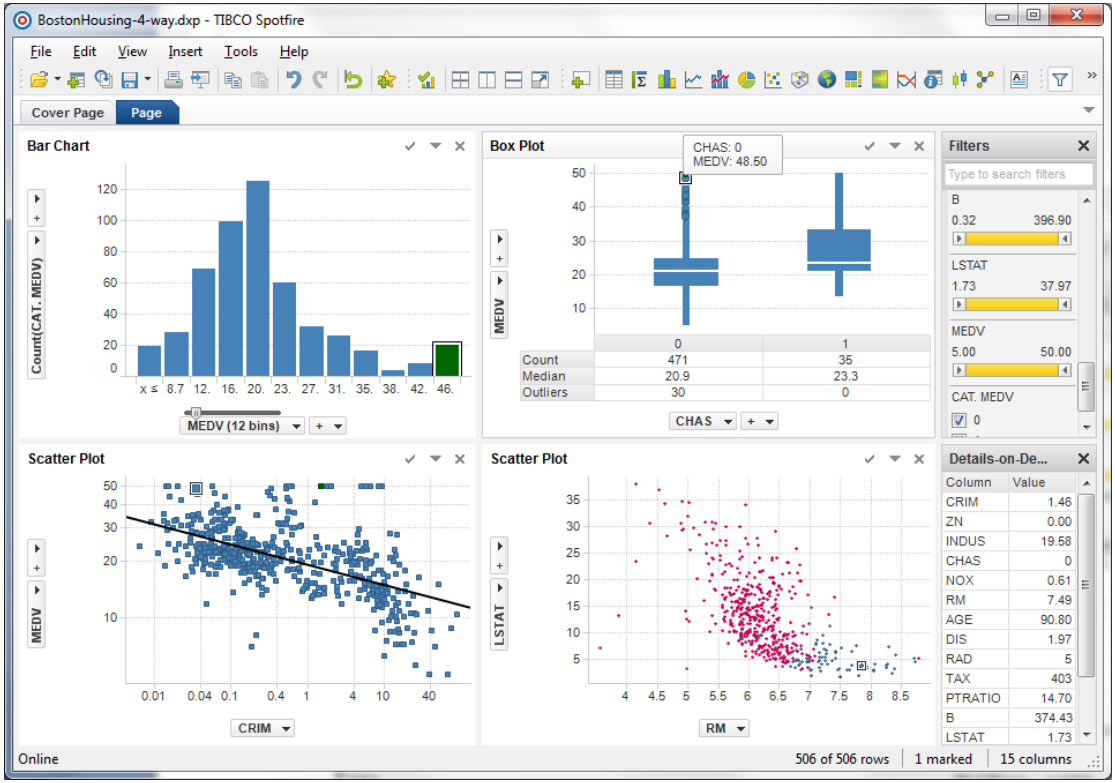
CATMEDV =1



CATMEDV =0



Linked plot

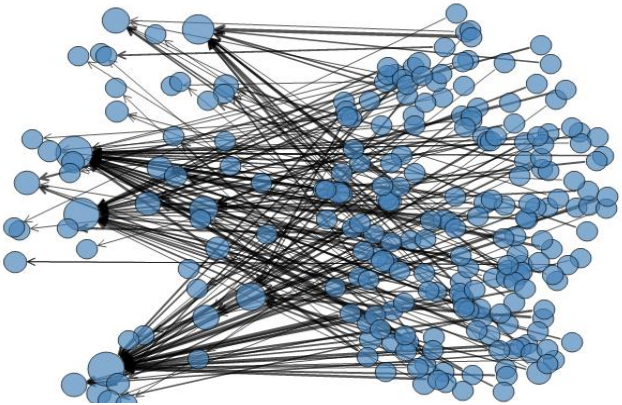

BostonHousing-4-way.dxp - TIBCO Spotfire
File Edit View Insert Tools Help
Cover Page
Page
Bar Chart
Count(CAT. MEDV)
MEDV (12 bins)
Box Plot
CHAS: 0
MEDV: 48.50
MEDV
CHAS
Count 471 35
Median 20.9 23.3
Outliers 30 0
Filters
Type to search filters
B 0.32 396.90
LSTAT 1.73 37.97
MEDV 5.00 50.00
CAT. MEDV
Scatter Plot
MEDV
CRIM
Scatter Plot
LSTAT
RM
Details-on-De...
Column Value
CRIM 1.46
ZN 0.00
INDUS 19.58
CHAS 0
NOX 0.61
RM 7.49
AGE 90.80
DIS 1.97
RAD 5
TAX 403
PTRATIO 14.70
B 374.43
LSTAT 1.73
Online
506 of 506 rows
1 marked
15 columns

네트워크 그래프– eBay Auctions 데이터(좌: 판매자, 우: 구매자)

원의 크기 = 각 노드에 대한 거래 수

선 두께= 구매자 – 판매자 쌍에 대한 경매 수

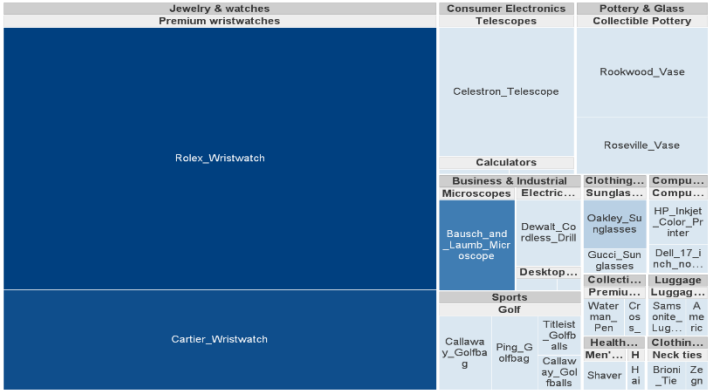
화살표: 구매자에서 판매자를 가리킴



사각형 크기 = 평균 종가

(물품의 가치)

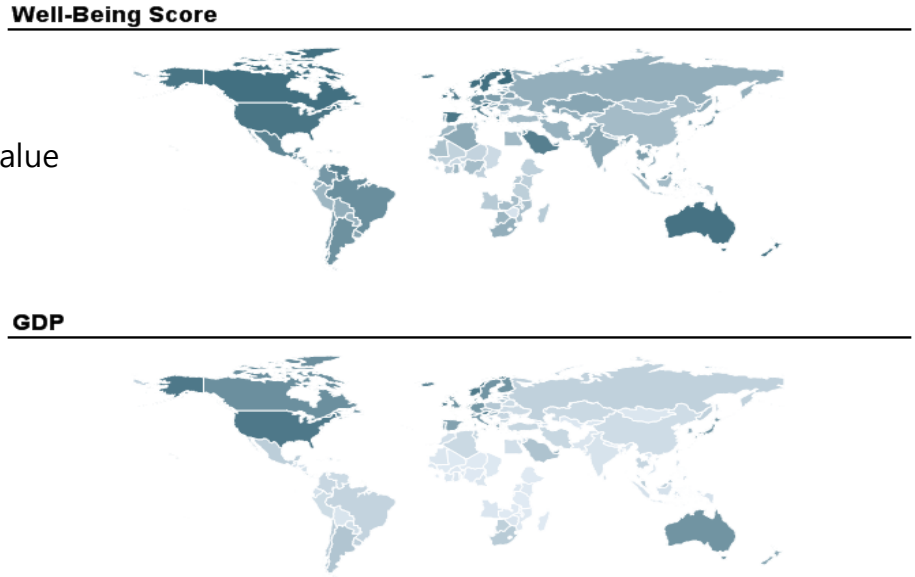
색상 = 부정적 피드백이 있는 판매자 비율(%) (darker=more)



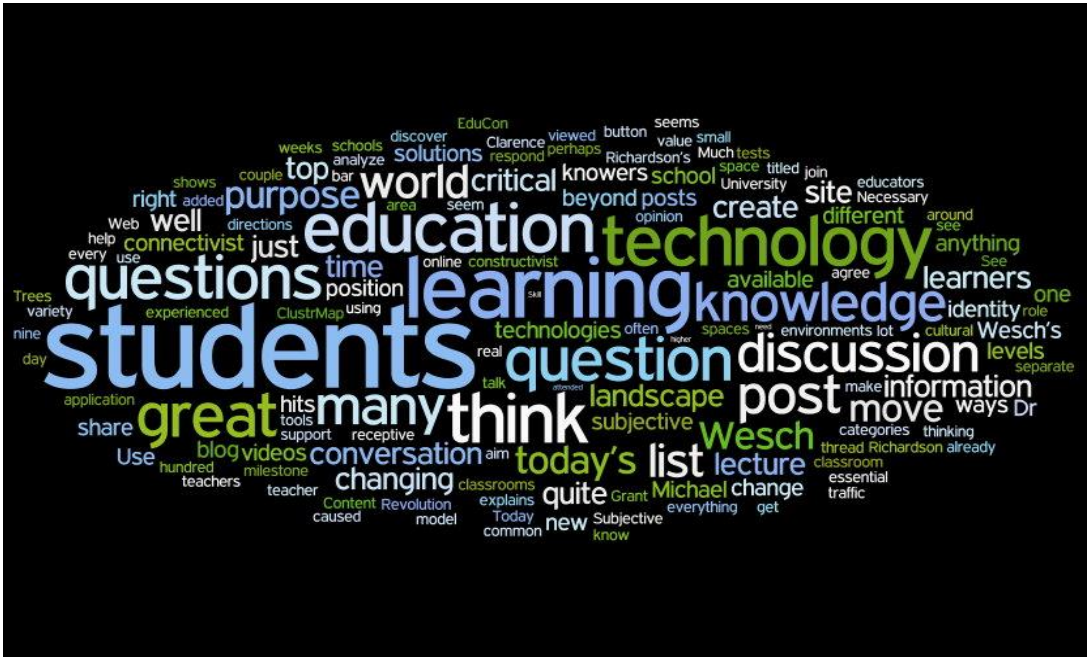
Map Chart

(Comparing countries' well-being with GDP)

Darker = higher value



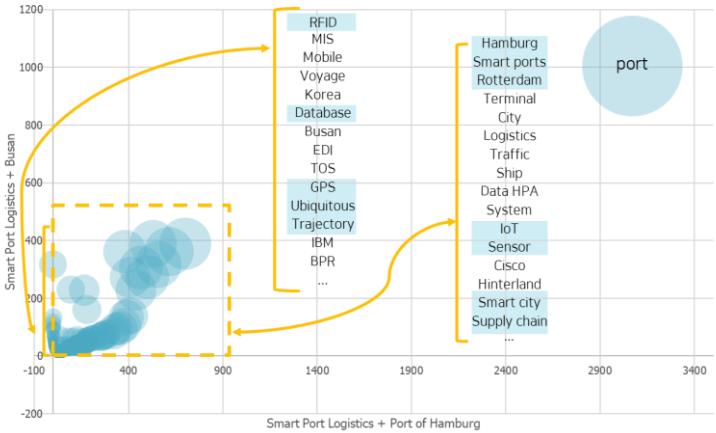
wordle


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

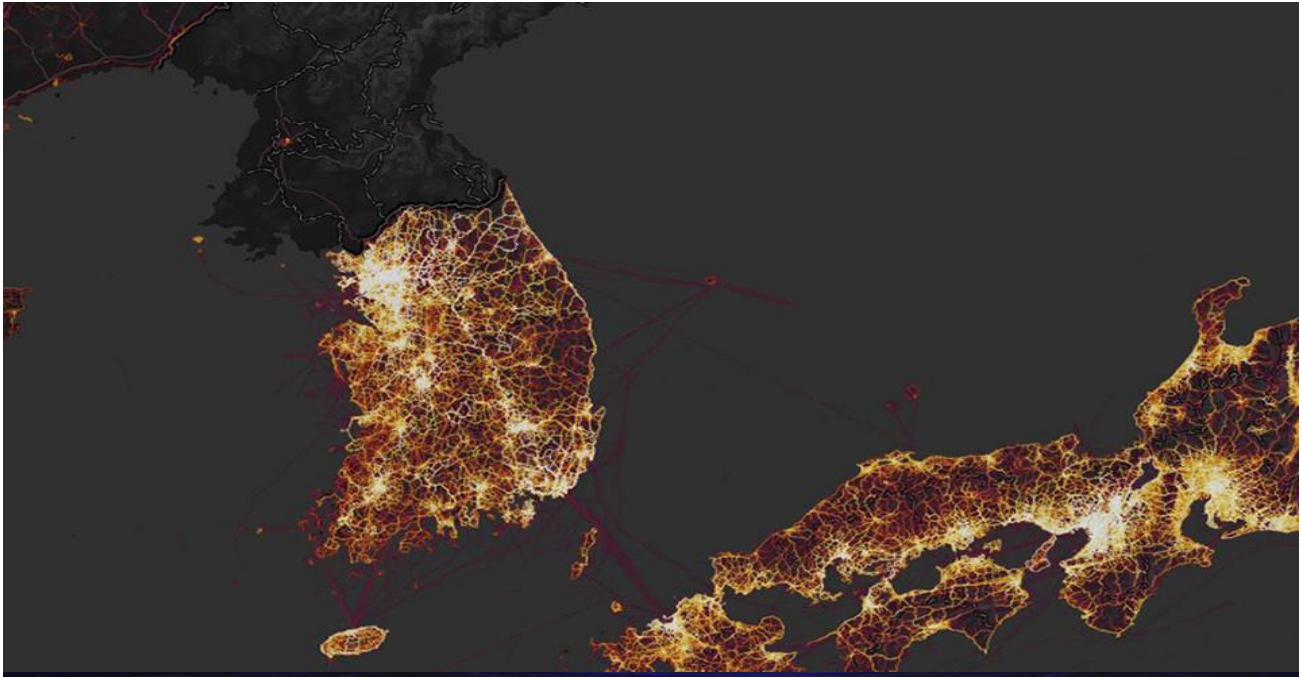
- 조사 기간 : 17.01.01 – 17.11.30 뉴스 기사 대상
- 주요 이슈
  1. 살인 불개미 발견 (불개미, 발견, 안전, 부두)
  2. 물동량 2000만TEU 기대 (물동량, 컨테이너, 처리,증가)
  3. 사드 보복에 따른 크루즈 이용객 급감 대책 마련 필요+ 크루즈 관광 활성화 (여객터미널, 크루즈, 북항)

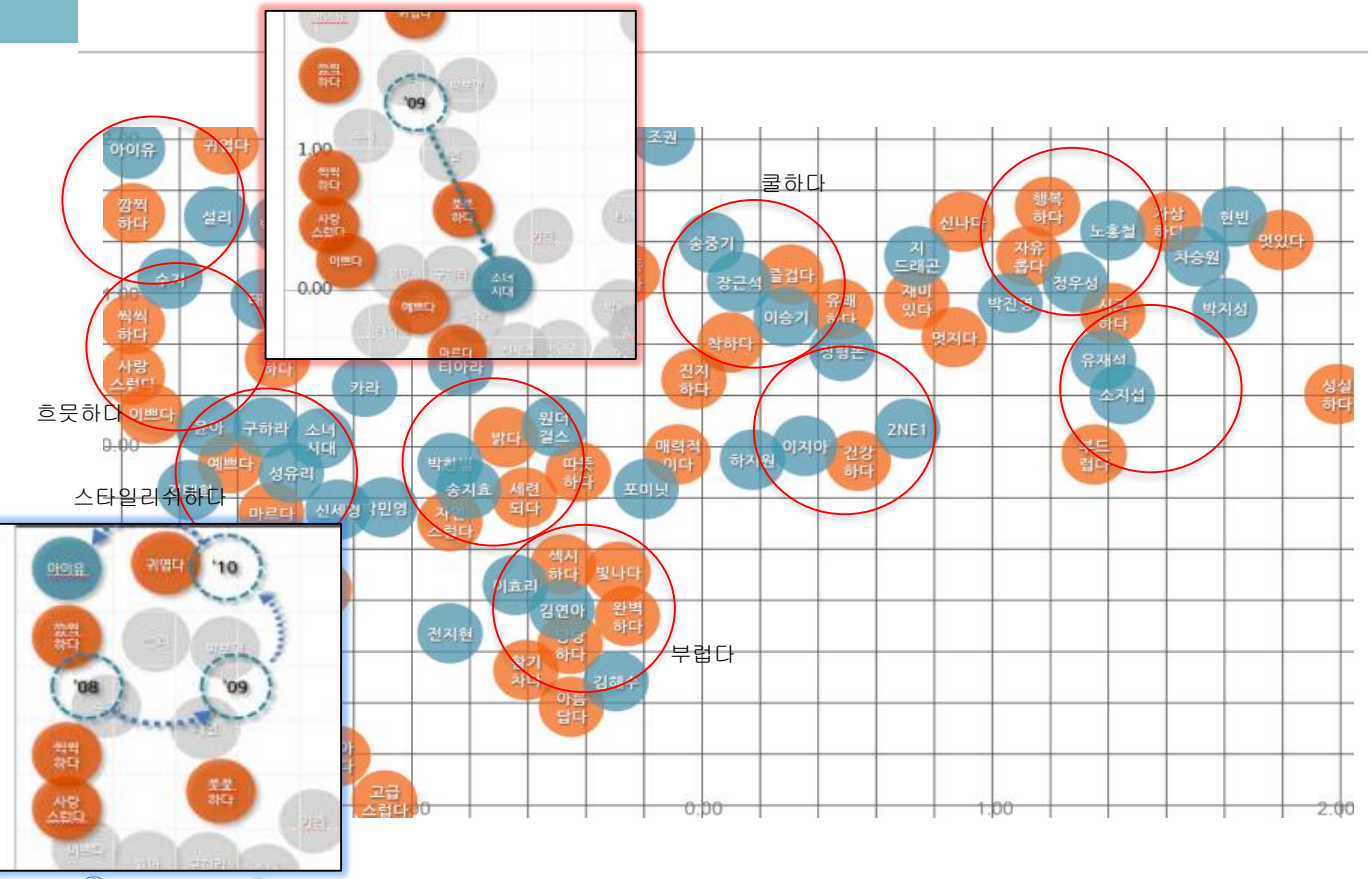
  기타 : 채용비리, 해수 온천, 한진해운 파산 여파

❖ Busan VS Hamburg

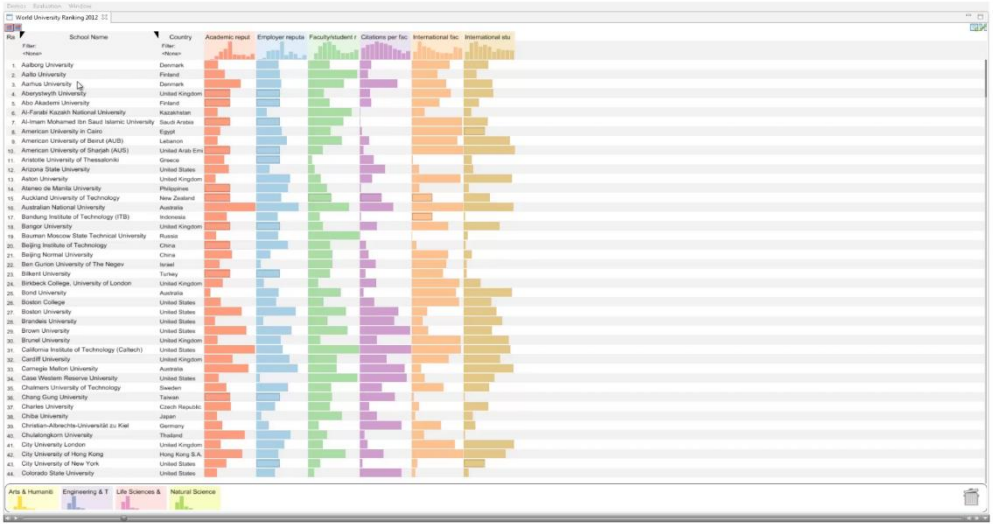


- Smart port logistics이라는 키워드에 Busan과 Hamburg를 추가하여 비교 분석
- Busan 키워드 에서는 RFID, Database, GPS, Ubiquitous 등 이전 세대의 기술적 키워드가 도출
- Hamburg 키워드 에서는 IoT, Smart ports, Rotterdam, Sensor, Smart city 등 4차 산업혁명의 기반 기술적 키워드가 도출




쿨하다
흐뭇하다
스타일리쉬하다
부럽다
'09
'10
'08
1.00
0.00
-1.00
2.00

LineUp


World University Ranking 2012
School Name
Country
Academic reput
Employer reputa
Faculty/student r
Citations per fac
International fac
International stu
Arts & Humaniti
Engineering & T
Life Sciences &
Natural Science

데이터 시각화 도구

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

시각화 예시

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

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