강의 계획 및 평가 방법

주차	날짜	강의 내용	과제	대면/비대면	평가
1	03/06	강의 소개		Online	
2	03/13	데이터 마이닝 절차		A704	
3	**03/20**	**데이터 탐색 및 시각화**		**B224**	
4	03/27	차원 축소	과제 1 (10%)	Online	
5	04/03	예측성능 평가		Online	
6	04/10	다중 선형 회귀분석		A704	
7	04/17	중간 프로젝트 발표		B224	30%
8	04/24	k-최근접이웃 알고리즘 나이브 베이즈 분류	과제 2 (10%)	Online	
9	05/01	분류와 회귀 나무		Online	
10	05/08	로지스틱 회귀분석		Online	
11	05/15	신경망		A704	
12	05/22	판별 분석		Online	
13	05/29	연관 규칙		Online	
14	06/05	군집 분석		A704	
15	06/12	기말 프로젝트 발표		B224	40%

Data Mining for Business Analytics

Ch. 03 Data Visualization

2023.03.20.

Contents

3.1 Introduction

3.2 Data Examples

3.3 Basic Charts: Bar Charts, Line Graphs, and Scatter Plots

3.4 Multidimensional Visualization

3.5 Specialized Visualization

3.6 Summary: Major Visualizations and Operations, by Data Mining Goal

3.1 Introduction

데이터 시각화 용도

- 주로 데이터 전처리 단계에서 사용
- Data cleansing: 틀린 수치들, 결측값, 중복 행, 중복 열 등을 찾음
- Variable derivation and selection: 어떤 변수들을 분석에 포함할지 그리고 어떤 변수가 불필요한지 결정하는 데 사용
- Determining appropriate bin size: 구간의 크기가 적당한지 수치형 변수들의 구간화가 필요한지 여부 / 데이터 차원 축소의 일부로 범주를 결합하는 역할
- Expensive data collection: 데이터를 더 수집해야 하고 그 수집비용이 크다면 유용한 변수나 측정치를 미리 선별하는데 사용

3.2 Data Examples

Ex 1: Boston Housing Data

- 보스턴의 인구조사 구역에서 측정된 여러가지 정보(예: 범죄율, 학생-교사 비율 등)로 구성된 14개 변수
- CAT.MEDV: 주택 가격의 중앙값(MEDV)을 3만 달러 기준 “고”와 “저” 두 가지 범주로 변환한 변수

Possible Task

- 결과변수가 주택가격 중앙값(MEDV)인 지도 예측 문제
- 결과변수가 이진변수 CAT.MEDV인 지도 분류 문제
- 인구조사 구역을 군집화하는 것이 목적인 비지도 문제

✓ MEDV와 CAT.MEDV는 위 세가지 어떠한 문제에서도 함께 사용되지 않음

변수명	변수 내역
CRIM	범죄율
ZN	25,000 평방피트 기준 거주지 비율
INDUS	비소매업종 점유 구역 비율
CHAS	찰스강 인접 여부 (1: 인접, 0: 비인접)
NOX	10ppm 당 일산화 질소 농도
RM	거주자의 평균 방의 개수
AGE	1940년 이전에 건축된 주택에 사는 비율
DIS	보스톤 5대 상업지구와의 거리
RAD	고속도로 진입 용이성 정도
TAX	10,000달러 당 재산세율
PTRATIO	학생 대 교사 비율
LSTAT	저소득층 비율
MEDV	주택가격의 중앙값(단위: $1,000)
✓CAT.MEDV	주택가격의 중앙값이 3만 달러 이상 여부(1: 이상, 0: 이하)

3.2 Data Examples

Ex 1: Boston Housing Data

CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX	PTRATIO	LSTAT	MEDV	CAT_MEDV
0.00632	18.0	2.31	0	0.538	6.575	65.2	4.0900	1	296	15.3	4.98	24.0	0
0.02731	0.0	7.07	0	0.469	6.421	78.9	4.9671	2	242	17.8	9.14	21.6	0
0.02729	0.0	7.07	0	0.469	7.185	61.1	4.9671	2	242	17.8	4.03	34.7	1
0.03237	0.0	2.18	0	0.458	6.998	45.8	6.0622	3	222	18.7	2.94	33.4	1
0.06905	0.0	2.18	0	0.458	7.147	54.2	6.0622	3	222	18.7	5.33	36.2	1
0.02985	0.0	2.18	0	0.458	6.430	58.7	6.0622	3	222	18.7	5.21	28.7	0
0.08829	12.5	7.87	0	0.524	6.012	66.6	5.5605	5	311	15.2	12.43	22.9	0
0.14455	12.5	7.87	0	0.524	6.172	96.1	5.9505	5	311	15.2	19.15	27.1	0
0.21124	12.5	7.87	0	0.524	5.631	100.0	6.0821	5	311	15.2	29.93	16.5	0

변수명	변수 내역
CRIM	범죄율
ZN	25,000 평방피트 기준 거주지 비율
INDUS	비소매업종 점유 구역 비율
CHAS	찰스강 인접 여부 (1: 인접, 0: 비인접)
NOX	10ppm 당 일산화 질소 농도
RM	거주자의 평균 방의 개수
AGE	1940년 이전에 건축된 주택에 사는 비율

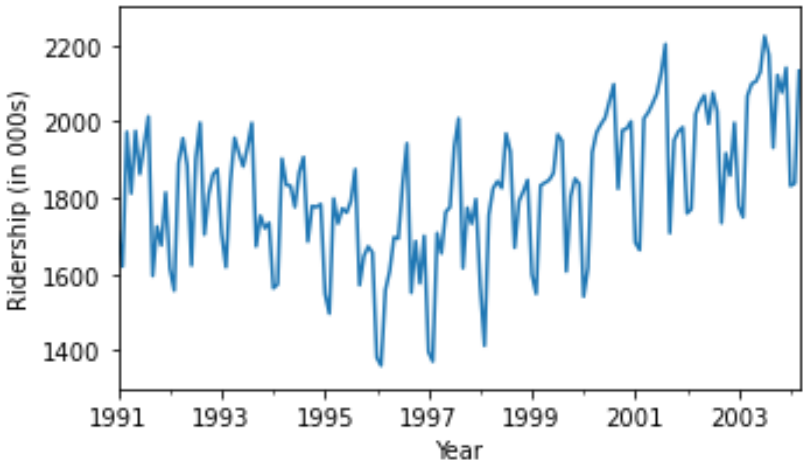
변수명	변수 내역
DIS	보스톤 5대 상업지구와의 거리
RAD	고속도로 진입 용이성 정도
TAX	10,000달러 당 재산세율
PTRATIO	학생 대 교사 비율
LSTAT	저소득층 비율
MEDV	주택가격의 중앙값(단위: $1,000)
CAT.MEDV	주택가격의 중앙값이 3만 달러 이상 여부(1: 이상, 0: 이하)

3.3 Basic Charts

[실습] Figure 3.1

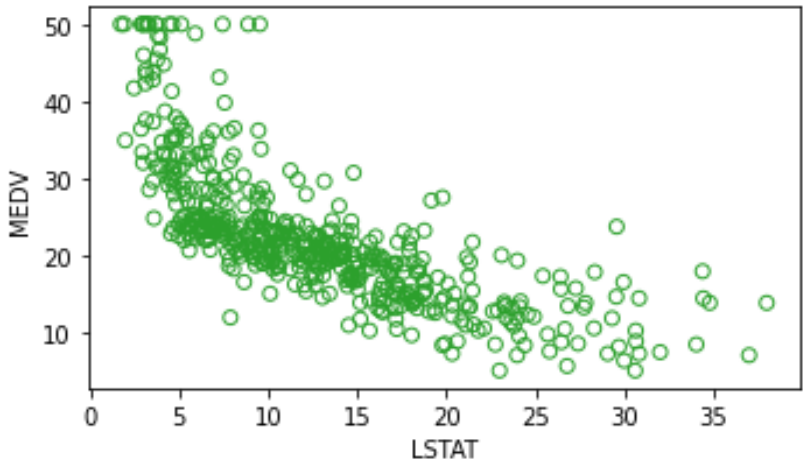
Line Graph

- 앰트랙의 월간 승객수의 시계열



Scatter Plot(산점도)

- 비지도학습의 경우 두 가지 수치형 변수 간의 정보 중복이나 군집 발견과 같은 연관성을 밝힐 수 있음



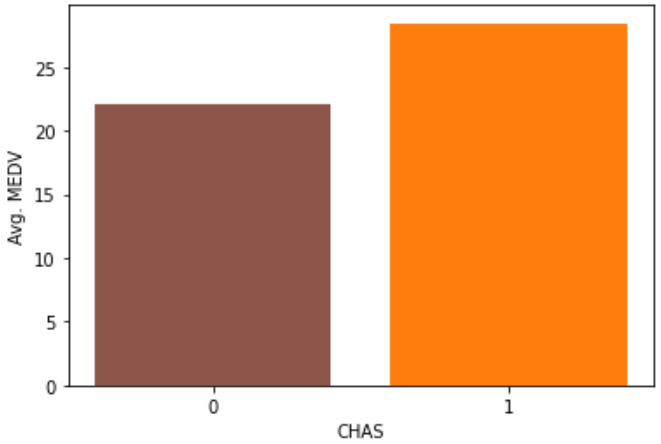
- ✓ y축: 결과변수(MEDV: 주택가격의 중앙값)
- ✓ x축: 예측변수(LSTAT: 저소득층 비율)

3.3 Basic Charts

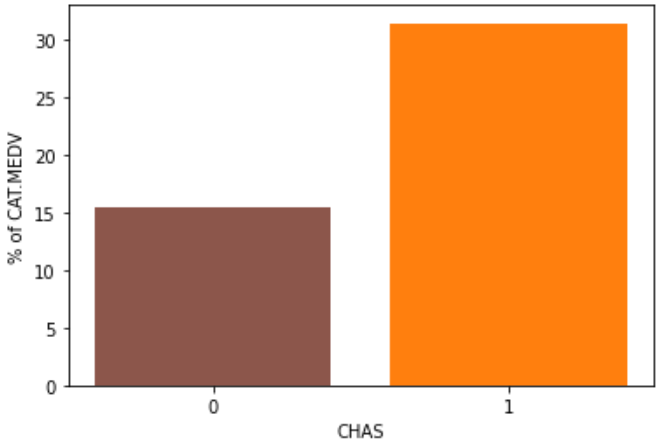
Bar Chart

- 평균, 개수, 비율과 같은 단일 통계값을 그룹별로 비교하는데 유용
- 막대: 각 집단, 서로 다른 막대는 다른 그룹
- 막대의 높이: 변수의 값
- y축: 수치형 결과 / x축: 범주형 예측변수

[수치형 변수] MEDV(주택가격)의 평균



[범주형 변수]
CAT.MEDV(주택가격이 3만 달러 이상인 구역의 비율)



[범주형 변수]
CHAS: 찰스강 인접 여부 (1: 인접, 0: 비인접)

3.3 Basic Charts

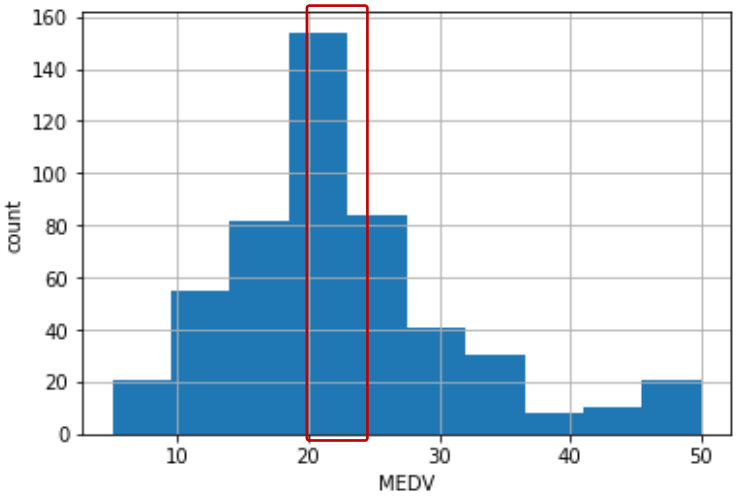
[실습] Figure 3.2

Distribution Plots: Boxplots and Histograms

- 수치형 변수의 전반적인 분포를 표시
- 박스플롯: 나란히 생성하여 하위그룹 간 비교 수행 또는 여러 개의 박스 플롯을 시간별로 생성하여 시간 변화에 따른 분포 관찰 가능
- 편향된 수치형 변수: 정규분포를 가정하는 분석방법(선형회귀, 판별분석)을 적용하려면, 반드시 변환(로그 스케일로 이동)을 수행해야 함

Histogram

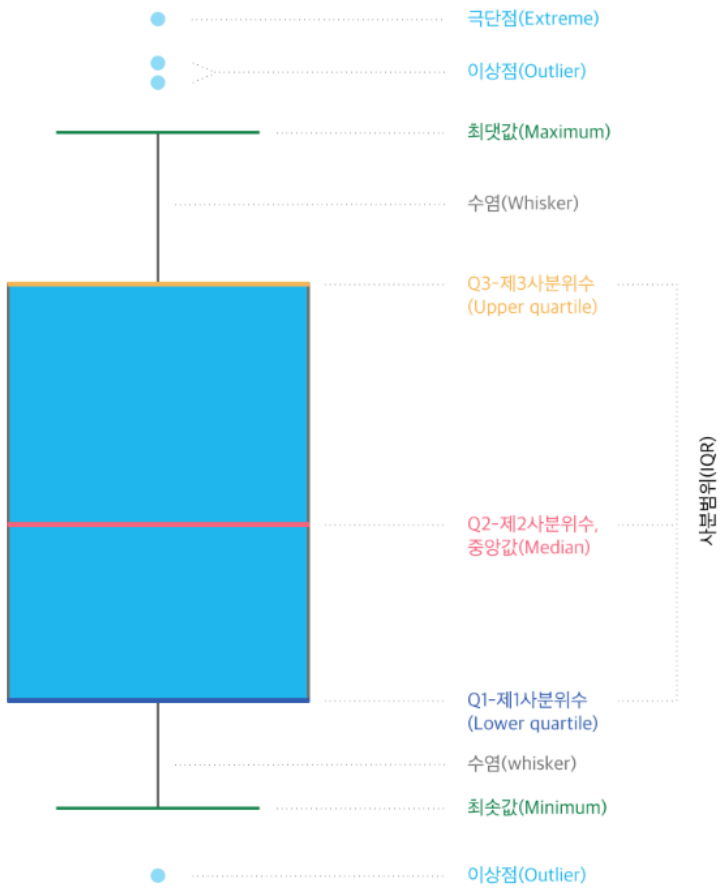
- 주택가격 중앙값(MEDV)이 2만 달러에서 2만 5천 달러인 구역이 150개 이상 있음



3.3 Basic Charts

Boxplots

- 데이터의 분포와 이상치를 동시에 보여주면서 서로 다른 데이터군 비교 가능
- 다섯 숫자 요약(Five-Number Summary)
  - ✓ 최솟값: (제 1사분위 - 1.5*IQR) 범위 내의 인접값
  - ✓ 제 1사분위 수(Q1): 중앙값 기준으로 하위 50% 중의 중앙값, 전체 데이터 중 하위 25%에 해당하는 값
  - ✓ 제 2사분위 수(Q2, 중앙값): 데이터의 정 가운데 순위에 해당하는 값.(관측치의 절반은 크거나 같고 나머지 절반은 작거나 같다.)
  - ✓ 제 3사분위 수(Q3): 중앙값 기준으로 상위 50% 중의 중앙값, 전체 데이터 중 상위 25%에 해당하는 값
  - ✓ 최댓값: (제 3사분위 + 1.5*IQR) 범위 내의 인접값

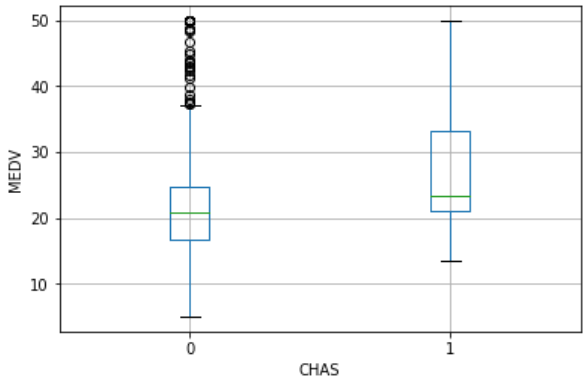


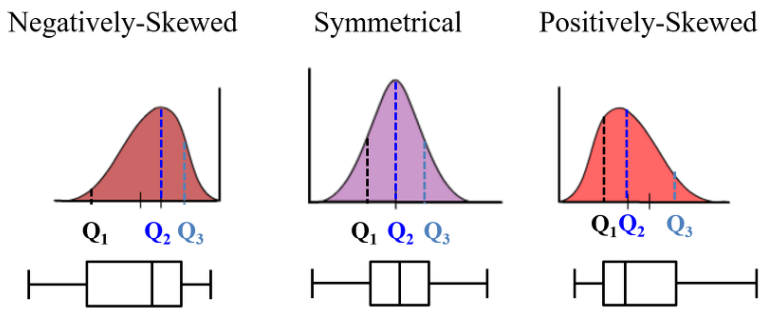
3.3 Basic Charts

[실습] Figure 3.2, 3.3

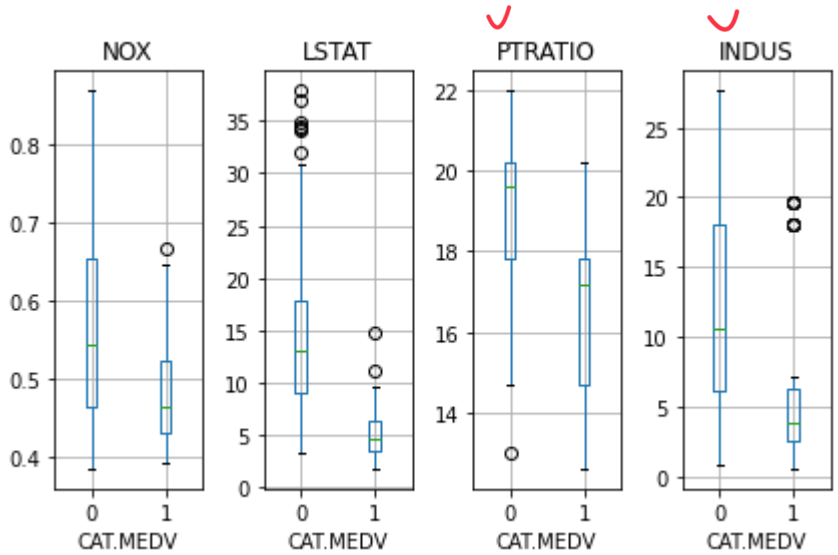
Boxplots

- 찰스강변 주택(1)의 평균 MEDV가 그 외 구역(0) 주택가격 보다 높을 뿐 아니라 모든 측정값이 더 높음





- 수치형 변수의 잠재성 평가 가능
- CAT.MEDV에 대한 4개 수치형 변수의 효과
- 가장 큰 간극을 가진 PTRATIO와 INDUS는 유용한 예측변수가 될 수 있음



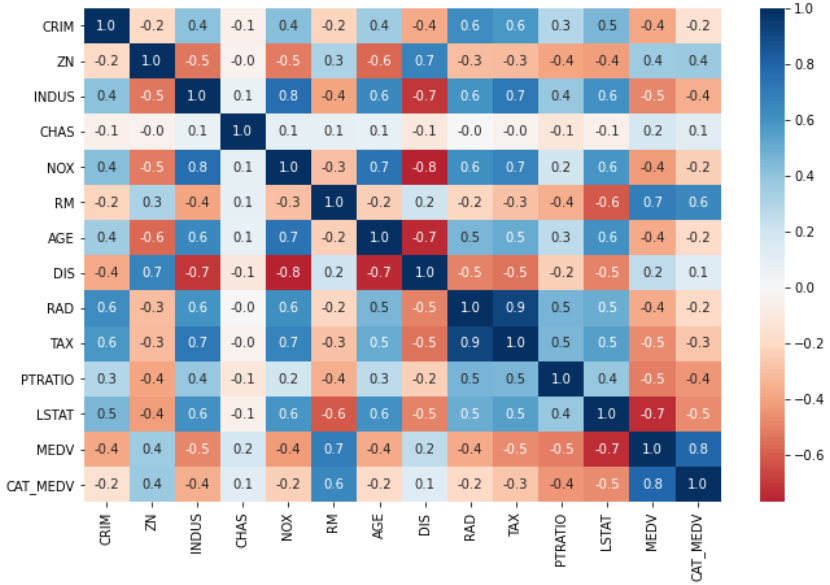
3.3 Basic Charts

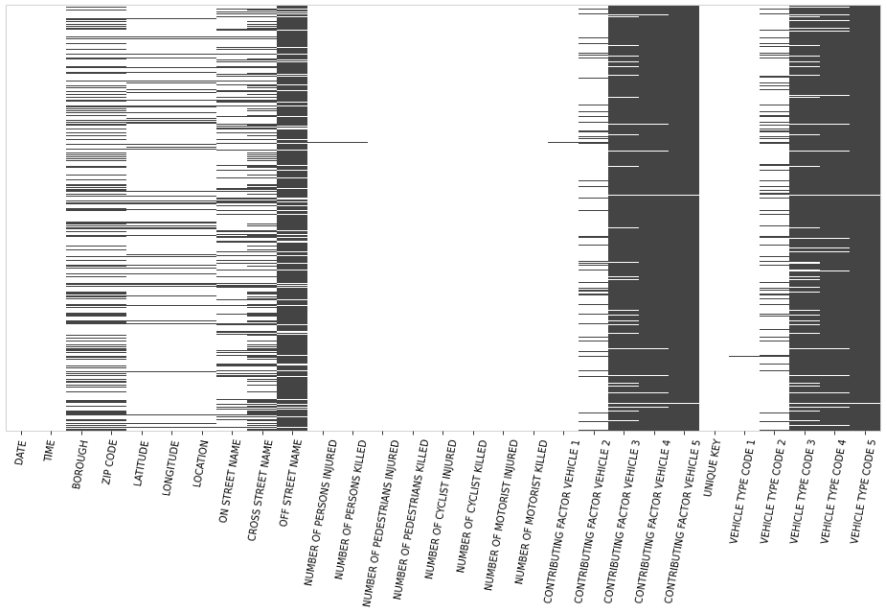
[실습] Figure 3.4, 3.5

Heatmaps: Visualizing Correlations and Missing Values

- 상관관계
  - ✓ p개의 변수에 대한 상관관계 표는 p열과 p행으로 표현
  - ✓ 색상의 변화(blue/red)로 상관관계 표현

- 결측값 heatmap
  - ✓ 결측값(회색): 1 / 나머지: 0
  - ✓ 데이터 파일의 결측 정도와 양을 시각화



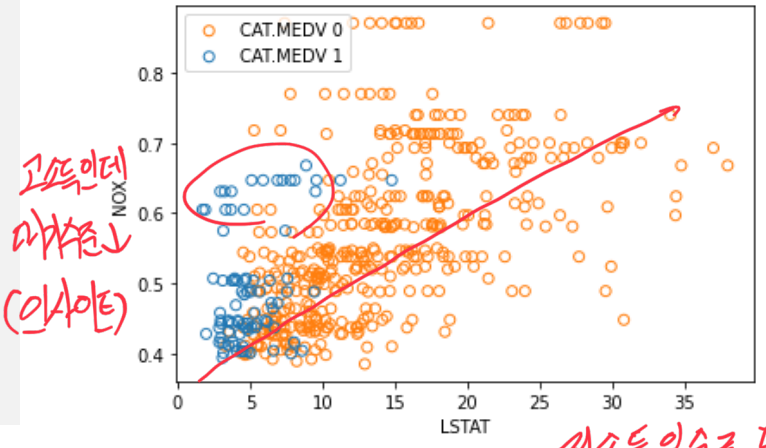


3.4 Multidimensional Visualization

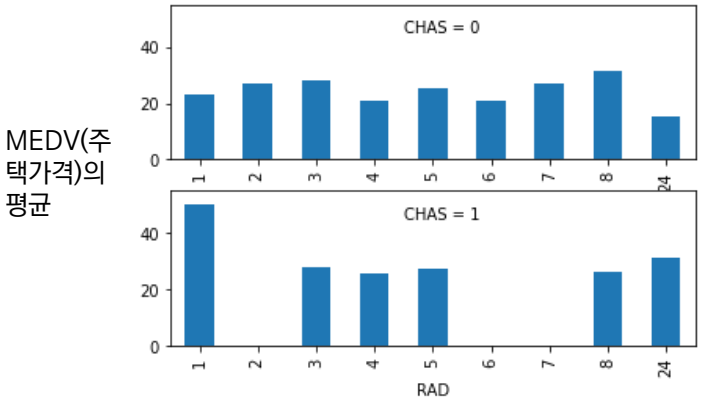
[실습] Figure 3.6

Adding Variables: Color, Size, Shape, Multiple Panels, and Animation

- 예측의 관점에서 컬러-코딩은 수치형 결과변수(y축)와 수치형 예측변수(x축) 사이의 관계 탐색 가능
- 컬러-코딩(Color-coding)된 산점도는 상호작용하는 조건을 만들 필요가 있는지 평가(예를 들어 NOX(일산화질소 농도)와 LSTAT(저소득층 비율)의 관계가 3만 달러 이상의 고급주택(1)과 아닌 주택(0)을 대비했을 때 다른가?)



- RAD(고속도로 진입 용이성)에 따른 평균 MEDV의 막대차트가 CHAS(찰스강 인접여부(1: 인접, 0: 비인접))에 의해 두 개의 패널로 분리
- RAD=1일 때, 현저한 차이 / RAD=2, 6, 7에는 강변주택이 전혀 없음 → RAD 범주의 축소 고려



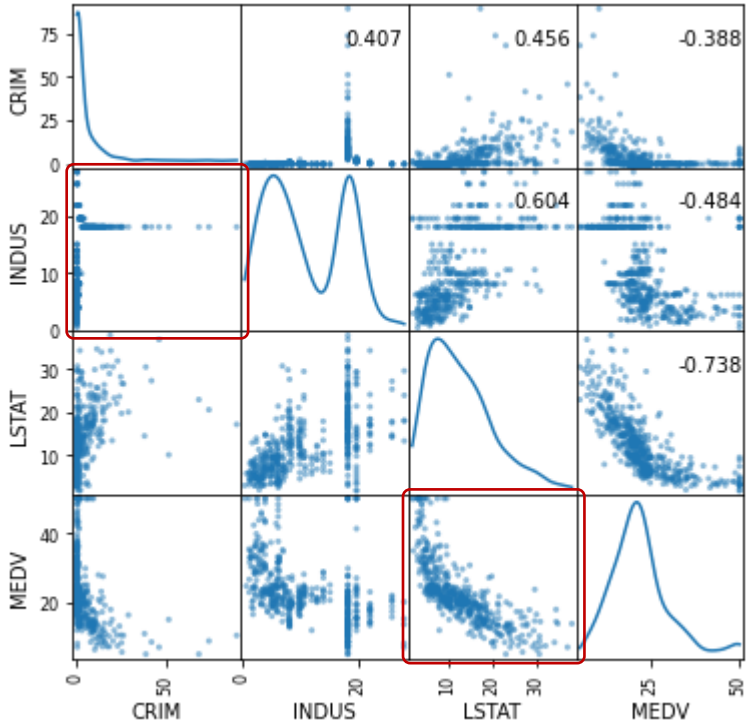
3.4 Multidimensional Visualization

[실습] Figure 3.7

Adding Variables: Color, Size, Shape, Multiple Panels, and Animation

Scatter Plot Matrix(산점도 매트릭스)

- 다중 패널 산점도 이용
- 각 열과 행: 각각 다른 변수에 상응
- 교차점: 가능한 모든 변수 조합의 산점도
- 비지도학습의 경우: 수치형 변수들 간의 연관성 분석, 이상치 탐지, 군집 식별
- 지도학습: 예측변수들 간의 쌍별 관련성을 평가하여 변수 변환과 변수 선택 도움
- e.g.) CRIM과 INDUS(비소매업종 점유 비율) 사이의 극단적인 편향성, MEDV와 LSTAT(저소득층 비율) 사이의 지수 관계



3.4 Multidimensional Visualization

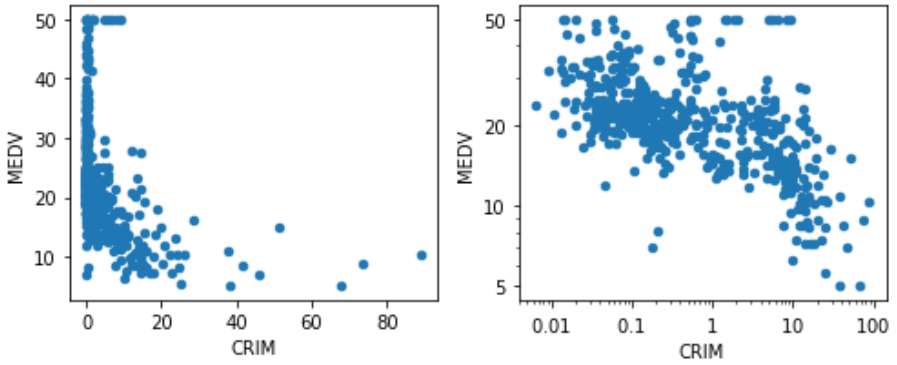
[실습] Figure 3.8

Manipulations: Rescaling, Aggregation and Hierarchies, Zooming, Filtering

Rescaling

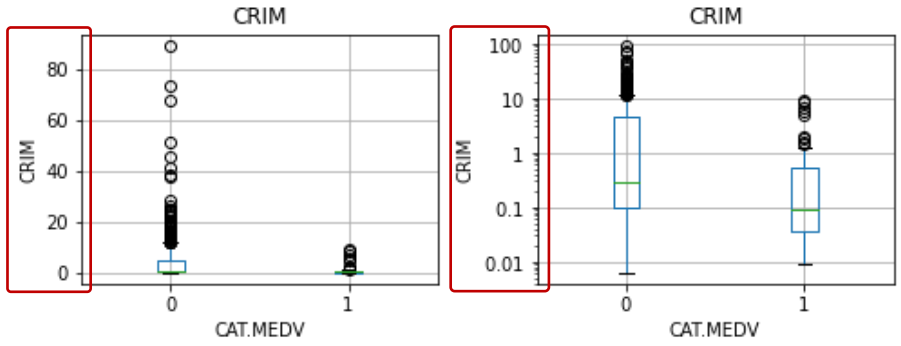
- 스케일 조절을 통해 y축 밀집 현상을 제거하고 MEDV와 CRIM 사이의 선형 관계(로그-로그 관계) 확인 가능

```
# Regular scale
housing_df.plot.scatter(x='CRIM', y='MEDV', ax=axes[0])
# log scale
ax = housing_df.plot.scatter(x='CRIM', y='MEDV',
logx=True, logy=True, ax=axes[1])
```



- x축 밀집 현상을 제거하고, 두 개의 박스플롯을 서로 비교하여 CAT.MEDV에 따른 CRIM의 분포 확인 가능

```
# log scale
ax = housing_df.boxplot(column='CRIM', by='CAT_MEDV',
ax=axes[1])
ax.set_yscale('log')
```



3.4 Multidimensional Visualization

[실습] Figure 3.9

Manipulations: Rescaling, Aggregation and Hierarchies, Zooming, Filtering

Aggregation and Hierarchies

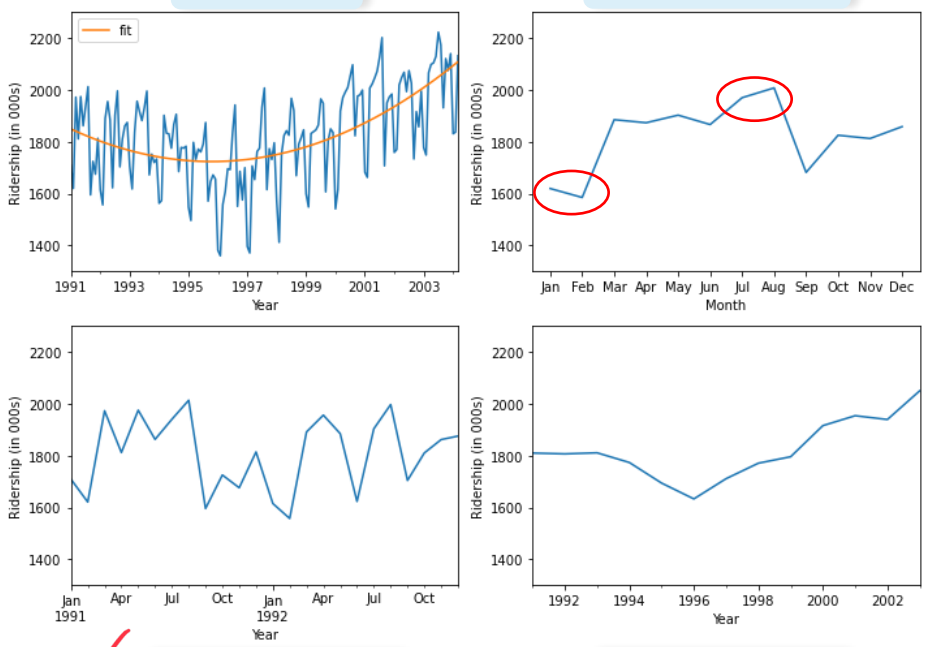
- 집계의 수준 변경
- Seasonal 집계(각 년, 월별)는 7-8월의 절정기와 1-2월의 침체기 판독 용이

Zooming and Panning

- 패닝: 줌 윈도우를 다른 영역으로 이동
- 전역 행동을 가정하는 방법(회귀모델)과 데이터 중심 방법(지수평활예측, k-최근접이웃분류) 중에서의 기법 선택에 활용

글로벌패턴

[철도 승객 시계열 데이터]



→ 트렌드를 읽기 쉬움

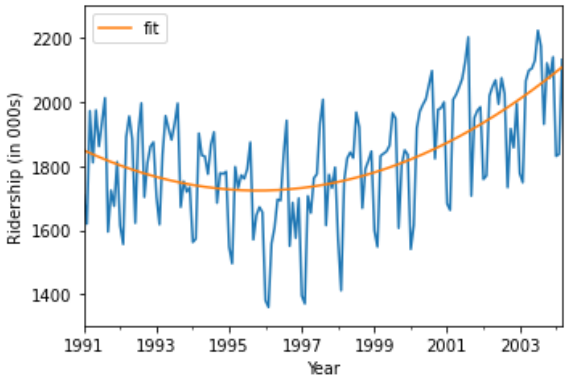
로컬패턴

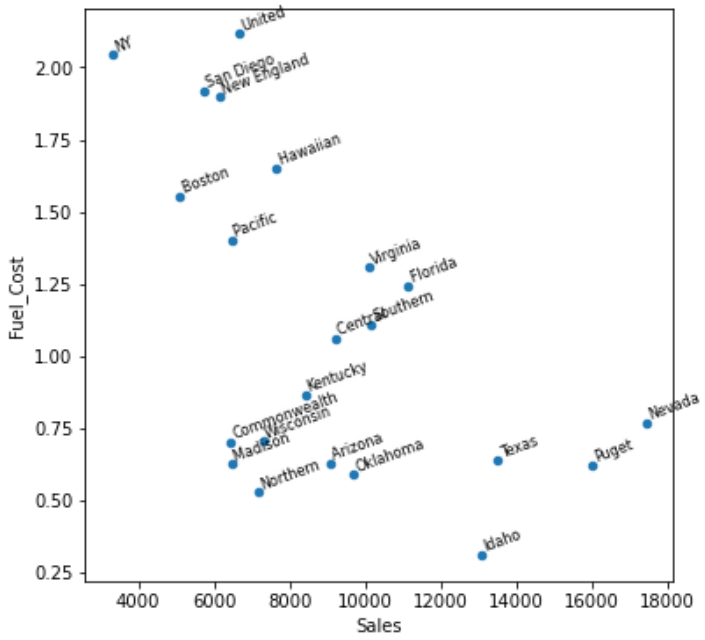
3.4 Multidimensional Visualization

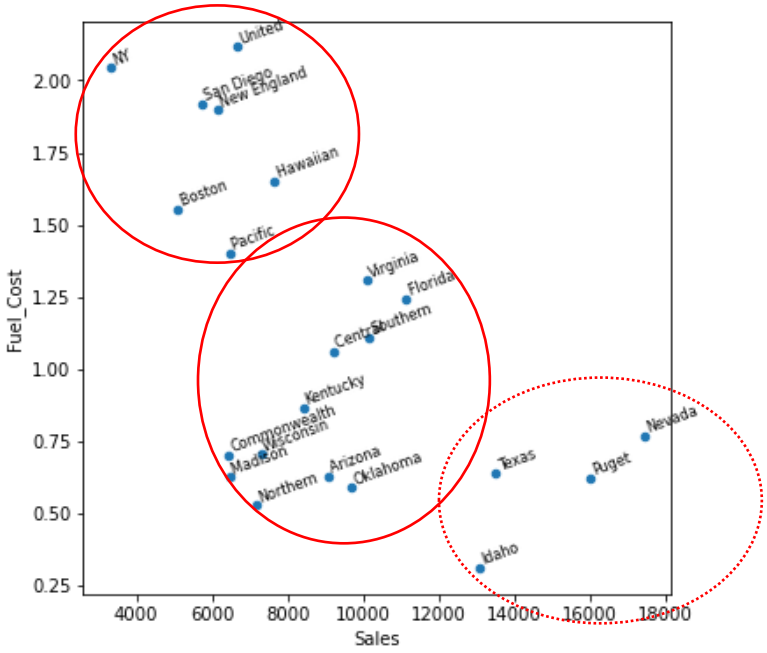
[실습] Figure 3.9, 3.10

Reference: Trend Lines and Labels

- 추세선: 참고사항을 제공하고, 패턴의 형태 평가에 활용
- 레이블: 인-플롯 레이블(In-plot label)의 사용은 이상치와 군집 분석에 활용







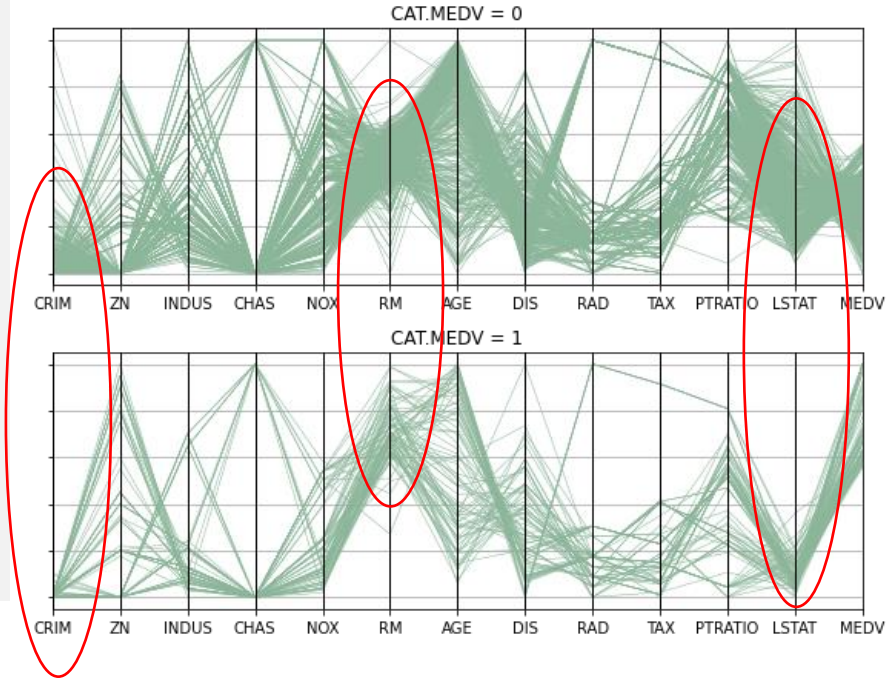
✓ 저연료비용 고매출

3.4 Multidimensional Visualization

[실습] Figure 3.12

Multivariate Plot: Parallel Coordinates Plot

- 다변량 차트: 평행좌표 차트
- 다변량 프로파일(Multivariate Profile): 각 변수에 한 개의 수직축, 개별 관측들은 각 수직축 위에 있는 관측값 연결
- 고가의 주택(1)은 저가의 주택(0) 보다 일관되게 낮은 CRIM(범죄율), 낮은 LSTAT(빈곤율), 그리고 높은 RM(침실 개수)를 보임
- 비지도학습 과제에 유용: 군집, 이상치, 변수 간의 중복된 정보 분별



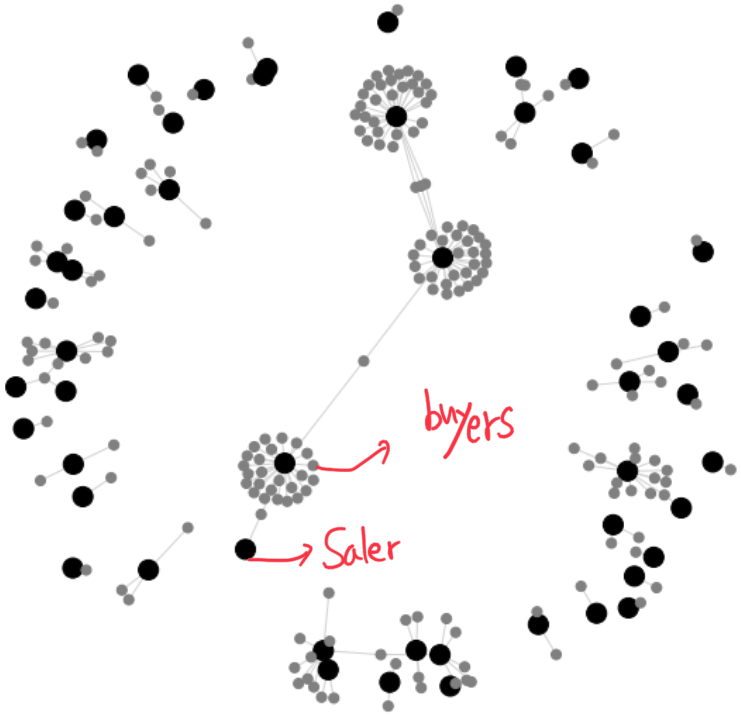
3.5 Specialized Visualization

[실습] Figure 3.14

Visualizing Networked Data

- 네트워크 다이아그램: 액터와 액터 간의 관계
- 노드(node): 행위자(소셜 네트워크 상의 사람들 또는 제품 네트워크 상의 제품들), 원으로 표현
- 연결선(edge): 노드들 간의 관계, 노드를 연결하는 선
- 예) 소셜 네트워크 상의 사용자들 – 서로 “친구” 관계인 사용자들
- 연관성 규칙 탐색에 용이
  - ✓ 노드: 품목 / 연결선: 함께 구매된 품목
  - ✓ “맥주와 기저귀” 조합이 연관성이 아주 높은 한 쌍의 노드로 표현됨

- 예) Ebay 판매자와 구매자 네트워크
- 선의 넓이: 경매 참가자가 판매자에게 가격을 제시한 회수
- 3-4명의 대형 판매자 존재



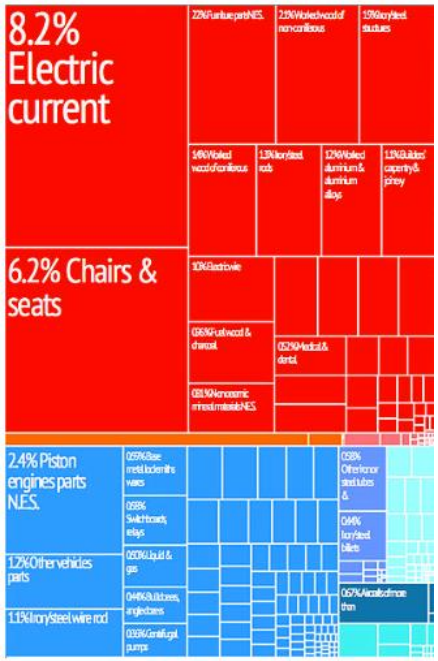
[Network plot Ebay sellers(black circles) and buyers(grey circles) of Swarovski beads]

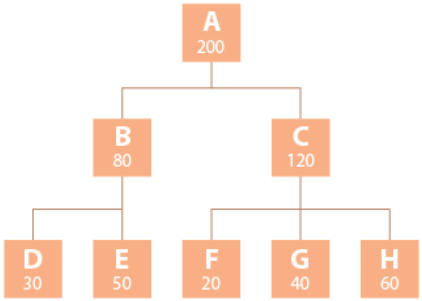
3.5 Specialized Visualization

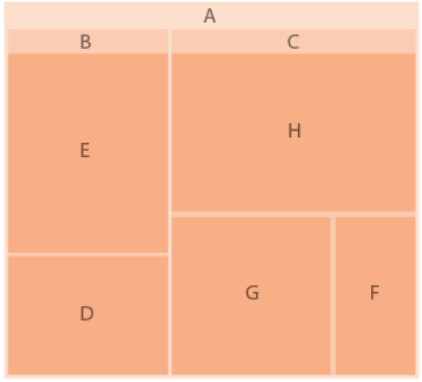
[실습] Figure 3.15

Visualizing Hierarchical Data: Treemap

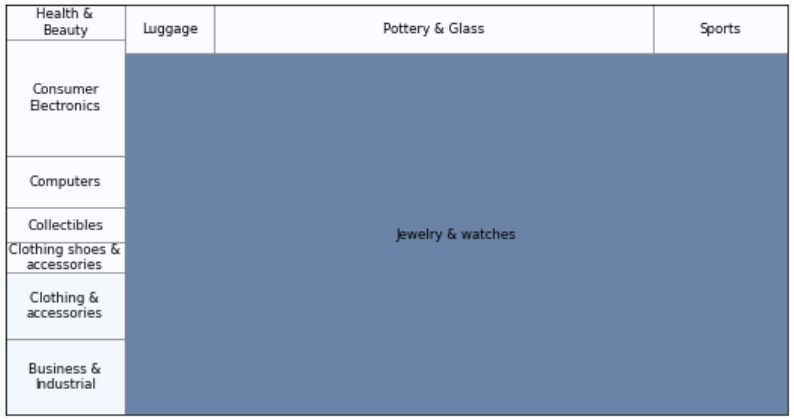
- 트리맵 계층구조의 하위 레벨은 직사각형 안에 들어있는 부속 직사각형 또는 아래 트리맵 처럼 표현







- 예) Ebay 경매품의 트리맵
- 크기: 평균 낙찰가
- 색상의 강도: 부정적인 피드백을 받은 판매자의 비율

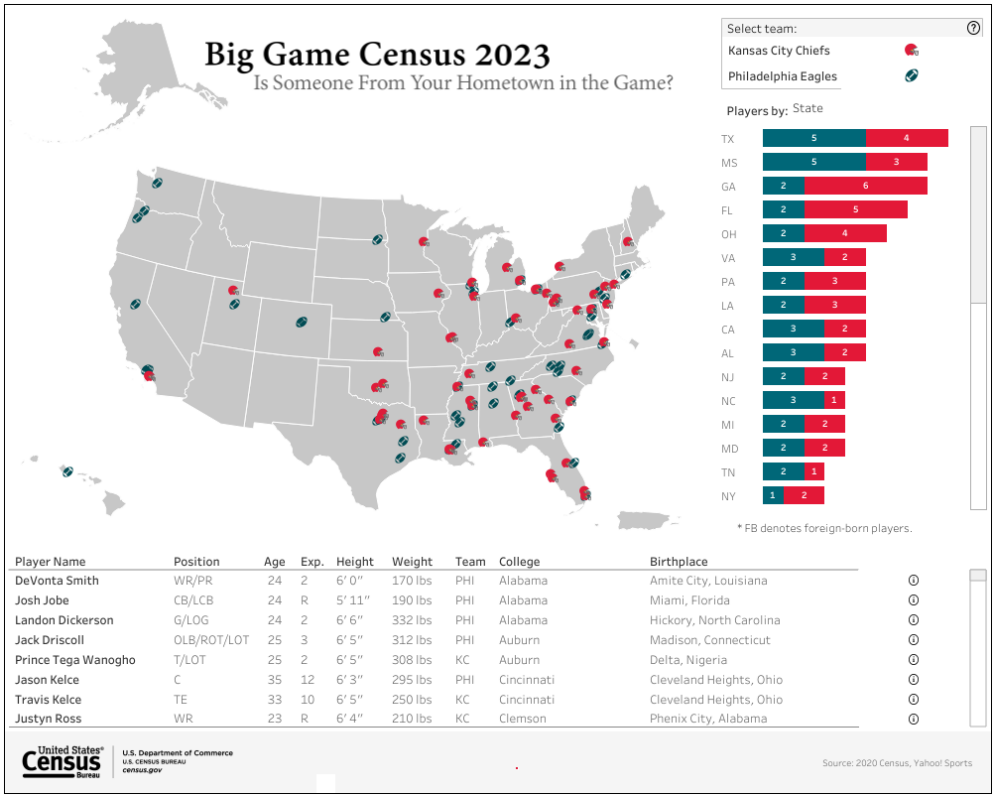


[상품분류로 분류된 Ebay 경매품의 트리맵]

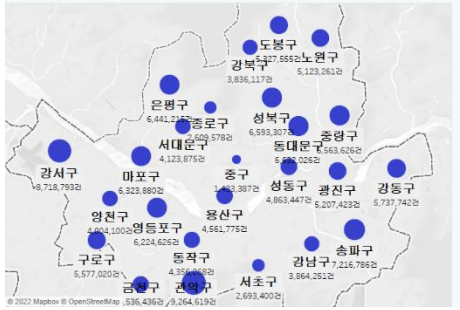
눈에 시각화를 많이 써

3.5 Specialized Visualization

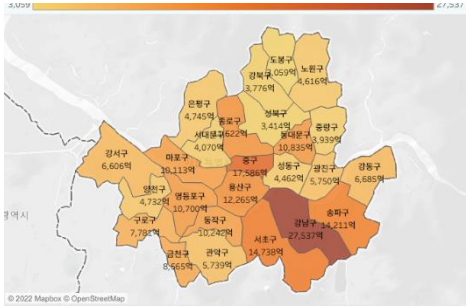
Visualizing Geographical Data: Map Charts



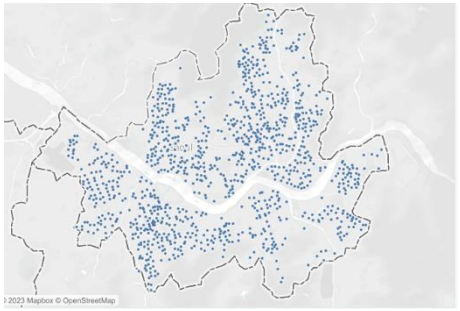
Bubble Map



Choropleth Map
(단계구분도)



Dot Map



Reference: “Tableau Public”, “https://jaydata.tistory.com”

3.6 Summary: Major Visualizations and Operations, by Data Mining Goal

Prediction

목적	시각화 방법
y축에 결과변수 배치	박스플롯, 막대차트, 산점도
결과변수와 범주형 예측변수 사이의 관계 탐색	병렬 박스플롯, 막대차트, 멀티패널
결과변수와 수치형 예측변수 사이의 관계 탐색	산점도
결과변수(그리고/또는 수치형 예측변수들)의 변환 필요성 결정	분포도(박스플롯, 히스토그램)
상호작용 조건의 필요성 탐색	산점도 분석(색상, 패널, 크기 추가)
상이한 행동양식의 데이터 탐색 및 글로벌 패턴과 로컬 패턴의 수준 평가	집계 수준과 확대/축소

3.6 Summary: Major Visualizations and Operations, by Data Mining Goal

Classification

목적	시각화 방법
결과변수와 범주형 예측변수의 관계 탐색	막대 차트
결과변수와 쌍별 수치형 예측변수의 관계 탐색	컬러코드된 산점도(색상: 결과변수)
결과변수와 수치형 예측변수 사이의 관계 탐색	병렬 박스플롯. 다수의 수치형 예측변수에 대해 박스플롯 시각화 수행. 가장 분리된 박스가 유용한 예측변수
결과변수(그리고/또는 수치형 예측변수들)의 변환 필요성 결정	분포도(박스플롯, 히스토그램)
상호작용 조건의 필요성 탐색	산점도 분석(색상, 패널, 크기 추가)
상이한 행동양식의 데이터 탐색 및 글로벌 패턴과 로컬 패턴의 수준 평가	집계 수준과 확대/축소

3.6 Summary: Major Visualizations and Operations, by Data Mining Goal

Time Series Forecasting

목적	시각화 방법
패턴의 종류 결정	다양한 시간으로 집계한 선 그래프 생성
단기 시계열 탐색 및 서로 다른 행동 양식을 보이는 데이터 영역 탐색	확대/축소/패닝
글로벌 패턴과 로컬 패턴 탐색	다양한 집계 수준 활용
시계열 데이터의 결측값 식별	결측값 heatmap
적절한 모델 선택	여러 유형의 추세선 겹쳐 보기

Unsupervised Learning

목적	시각화 방법
관측의 쌍별 관계와 군집 식별	산점도 매트릭스
상관관계표 검토	히트맵
다른 행동 양식을 보이는 데이터 영역 탐색	다양한 집계 수준, 확대/축소 활용
데이터 군집 탐색	평행좌표 차트