사회통계
변수 간 관계의 시각화

김현우, PhD[1]

[1] 충북대학교 사회학과 조교수

진행 순서

1. 두 변수 간 관계의 시각화: 산점도
2. 두 변수 간 관계의 시각화: 시계열 도표
3. 두 변수 간 관계의 시각화: 그 밖의 경우
4. 시각화에 관한 코멘트

두 변수 간 관계의 시각화: 산점도

두 변수 간 관계의 시각화: 산점도

산점도는 이변량/다변량 데이터 시각화 기법 가운데 가장 중요하다.

- 지금까지 단변량(univariate) 자료의 시각화를 다루었다면 지금부터는 이변량(bivariate) 또는 다변량(multivariate) 자료의 시각화에 대해 이야기한다.
- 모든 변수가 양적변수라면 그 관계를 살펴보기 위해 산점도(scatterplot)를 활용할 수 있다.

두 변수 간 관계의 시각화: 산점도

- 아래 자료는 수학과 영어에 대한 호감도를 평가한 것이다. 모두 4명을 조사하였으며 이들은 각 과목에 대해 매우 좋아하면 최대 3점, 매우 싫어하면 최소 -3점을 부여하였다(편의상 두 변수가 모두 양적변수라고 하자).

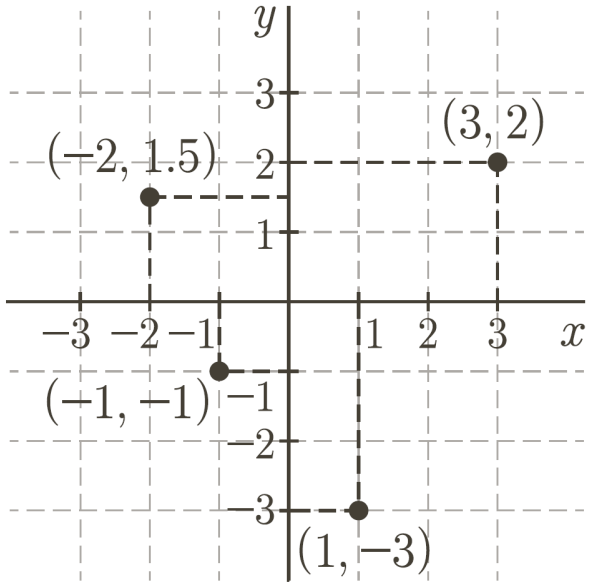
ID	MATH	ENG
1	-2	1.5
2	3	2
3	-1	-1
4	1	-3

- 자료 첫째 줄을 하나의 벡터(vector) $(-2, 1.5)$로 나타낼 수 있다.
- 전체 자료, 즉 수열(series)을 아래처럼 4개의 벡터로 나타낼 수 있다.

$$\{(-2, 1.5), (3, 2), (-1, -1), (1, -3)\}$$

두 변수 간 관계의 시각화: 산점도

- 4개의 벡터를 (X, Y)로 각각 파악한 뒤, 데카르트 좌표계(Cartesian coordinates) 위에서 4개의 점으로 찍을 수 있다.

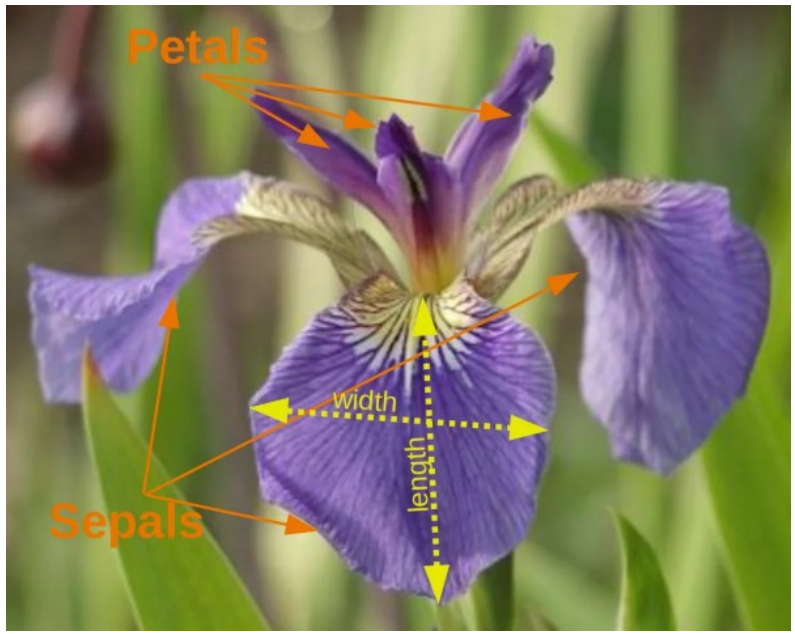


두 변수 간 관계의 시각화: 산점도

- 산점도란 결국 두 변수 X와 Y가 주어졌을 때 하나의 관측치(observation)를 하나의 벡터 (X, Y)로 파악한 다음, 이것들을 데카르트 좌표계 위에 뿌린 것이다.
- 산점도와 더불어 "여러 점들 사이의 추세를 나타내는 선", 즉 적합선(fitting line)을 그리면 두 변수 사이의 관계를 시각화할 수 있다.

두 변수 간 관계의 시각화: 산점도

예제 5. iris.csv는 세 종류의 붓꽃에 관하여 꽃잎(petal)과 꽃받침(sepal)의 길이(length)와 넓이(width)를 나타낸 것이다. 이 자료에서 두 변수 petal_length와 petal_width의 자료유형을 판단하고 그 관계를 적절히 시각화하시오.

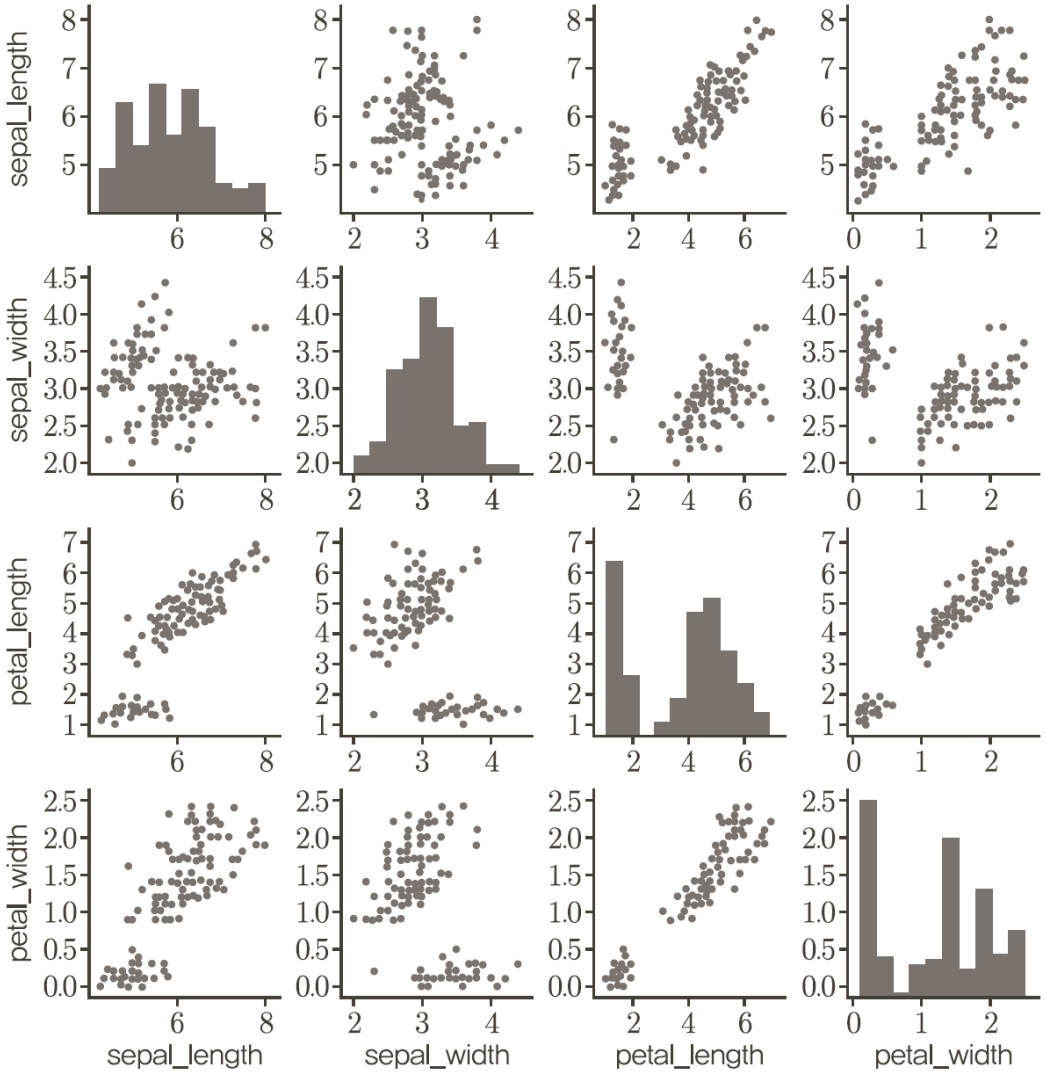


두 변수 간 관계의 시각화: 산점도

- 먼저 petal_length와 petal_width 두 변수를 살펴보자. 자료유형은 무엇인가?
- 두 변수의 columns를 하이라이트한 다음, [삽입]-[분산형 차트]를 클릭한다.
- 분산형 차트를 유심히 보고 맞는 것을 골라야 한다.
- 그래프를 꾸미되 다음에 주의하자.
  (1) 두 축 제목을 반드시 기입한다.
  (2) 축 서식에서 경계 최소최대값을 적절히 선택한다.
  (3) "추세선 삽입(R)"을 선택해 추세선(trend line)을 함께 그린다. 추세선은 적합선의 다른 이름이다.

두 변수 간 관계의 시각화: 산점도

- 모든 변수의 산점도를 행렬처럼 나타낼 수도 있다. 대각선은 히스토그램이다.



두 변수 간 관계의 시각화: 산점도

추세선을 수식으로 표현할 수 있다.

- 중학생 시절 당신은 알고 있었다: 직선을 아래의 수식으로 나타낼 수 있음을!

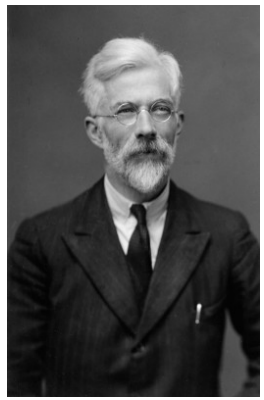
$$Y = b_0 + b_1 X$$

- 위 식에서 b_1은 직선의 기울기(slope)를 나타낸다.
- b_1이 양수(+)이면 직선은 우상향하고, 음수(-)이면 직선은 우하향한다.
- 위 식에서 b_0은 직선의 절편(intercept)을 나타낸다.
- 절편이란 이 직선과 y축이 만나는 자리를 의미한다.
- 당연히 $b_0 = 0$인 경우 이 직선은 원점(origin)을 통과한다.
- 엑셀에서는 "추세선 서식(F)"에서 "수식을 차트에 표시(E)"를 체크하면 수식이 나타난다.

두 변수 간 관계의 시각화: 산점도

이 데이터 iris.csv에는 나름의 역사가 있다.

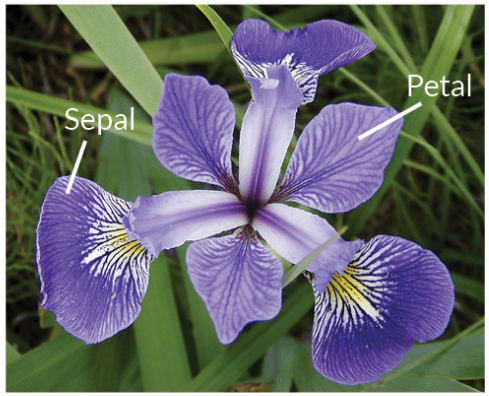
- 거의 백년 쯤 전 통계학자 Ronald Fisher가 처음 사용한 이래 매우 고전적인 자료로 여전히 다양한 목적에 활용되어 왔다.

Fisher, Ronald. A. 1936. "The Use of Multiple Measurements in Taxonomic Problems." Annual Eugenics 7(II): 179-188.

두 변수 간 관계의 시각화: 산점도

- 많은 교과서에서 통계 분석기법이나 알고리즘을 설명할 때 이 자료를 예제로 활용하면서, “(그 기법으로) 꽃잎과 꽃받침의 길이와 넓이로 붓꽃의 종류를 식별할 수 있을까?” 라는 질문을 던진다.



Iris Versicolor

Iris Setosa

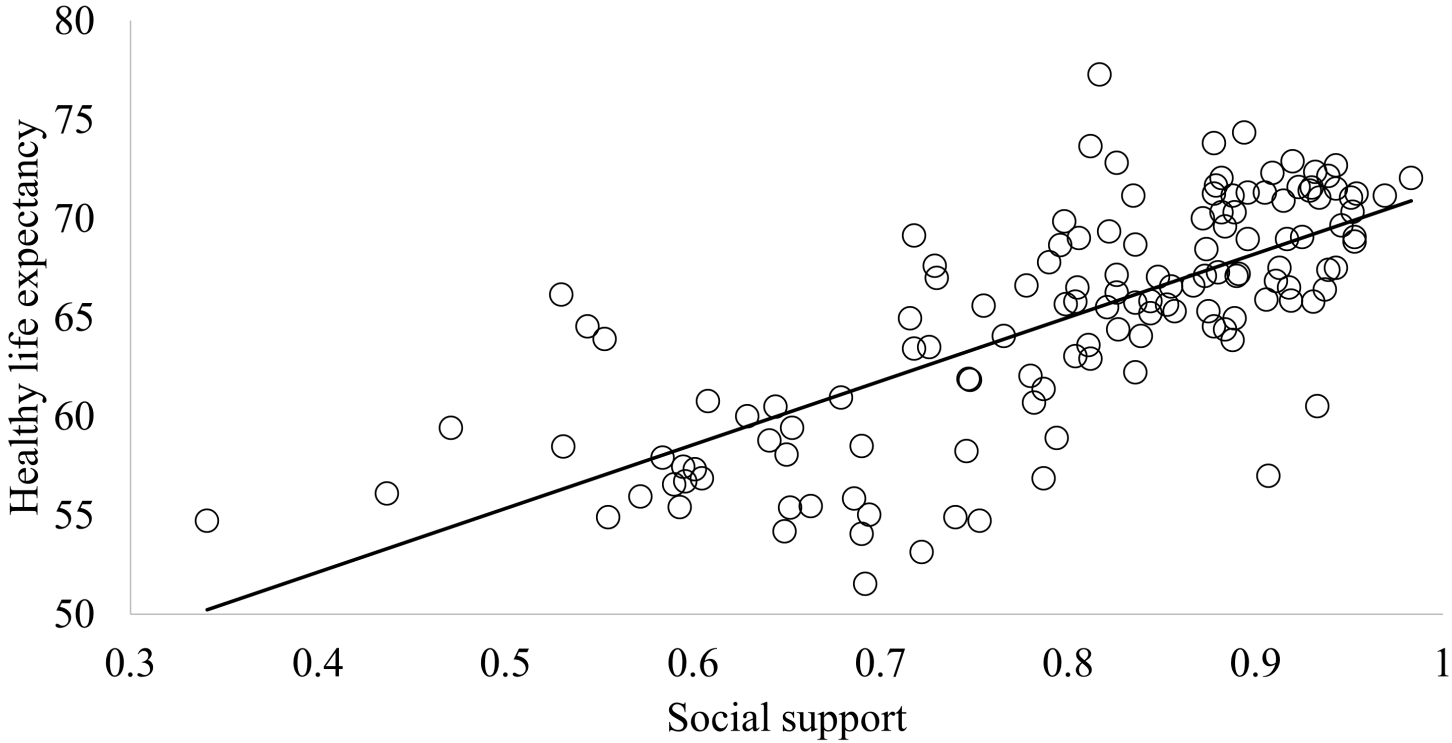
Iris Virginica

두 변수 간 관계의 시각화: 산점도

예제 6. WHR.csv를 불러와 국가별로 사회적 지지(social support)와 건강기대수명(healthy life expectancy) 간 연관성이 어떠한지 적절히 시각화하시오.

두 변수 간 관계의 시각화: 산점도

- 어떤 변수가 x축에 와야 하는지 미리 결정해야 한다(Why?).
- 적절한 x축과 y축의 표시 구간을 설정하자. 적합선도 그려야 한다.
- 시각화를 넘어 어떤 통계량이 두 관계의 연관성을 나타내기에 적절한지 고민해보자.
- 실제로 그 사회학적 의미를 해석해보자.



두 변수 간 관계의 시각화: 시계열 도표

두 변수 간 관계의 시각화: 시계열 도표

마지막으로 시간에 따라 변화하는 값들도 시각화 할 수 있다.

- 시계열 자료(time-series data)란 시간의 흐름에 따라 관측된 자료이다. 대표적인 예로 일별 주가지수나 연간 강수량 등을 생각해 볼 수 있다.
- 사회학에서는 시계열 자료를 거의 다루지 않지만 경제학에서는 시계열 자료를 분석하는 시계열 분석(time-series analysis)이 이미 엄청나게 발전했다. 특히 금융공학(financial engineering) 쪽에서 고도의 기법이 연구되고 있다.

두 변수 간 관계의 시각화: 시계열 도표

예제 7. NASDAQ.csv는 1971년에서 2022년 3월 23일까지 NASDAQ Composite 주가지수의 실제 자료이다. 조정종가지수(Adj. Close)의 시계열 도표를 작성하시오.

두 변수 간 관계의 시각화: 시계열 도표

- 시계열 도표에서 x축은 반드시 시간이다. y축은 나타내고자 하는 자료의 단위가 된다.
- 이것은 일간(daily) 데이터이지만, 개장하지 않는 휴일 등이 있으므로 간격(gap)이 존재한다.
- 불필요한 나머지 변수를 일단 삭제하고 Date와 Adj. Close 두 변수를 가까이 붙이자.
- 이 둘을 하이라이트하고 [메뉴]-[2차원 꺾은선형]을 고르자.
- 이런 종류의 그림은 종종 라인 차트(line chart)라고도 불리우며 시계열 자료에 대해 자주 쓰인다.
- 만약 자료의 관측치 사이가 선으로 이어져야 할 명확한 이유가 없다면 라인 차트로 표현해서는 안된다!

두 변수 간 관계의 시각화: 그 밖의 경우

두 변수 간 관계의 시각화: 그 밖의 경우

두 변수 중 하나는 양적변수가 아니라면 어떤 대안이 있을까?

- 만약 두 변수 모두 질적변수라면 어떻게 시각화할 것인가?
- (지난 주에 배웠듯) 두 질적변수의 관계는 교차표로 나타내 분석할 수 있다.
- 그런데 교차표를 일단 만들면 셀 안의 숫자는 비율/백분율이므로 다시 양적변수처럼 다룰 수 있게 된다(Why?).
- 그러므로 우리에게 질적변수 두 개가 주어졌더라도, 사실상 양적변수와 질적변수의 연관성을 살펴보게 된다!

두 변수 간 관계의 시각화: 그 밖의 경우

예제 8. 어떤 조사업체는 경쟁관계에 있는 휴대폰 브랜드 S와 I에 대한 선호도 조사를 A지역과 B지역에서 실시하였다. 총 200명의 조사 결과를 교차표로 정리하여 다음과 같이 행 표준화된 교차표를 얻었다. 이 교차표를 시각화하여 브랜드 선호도를 지역별로 비교하시오.

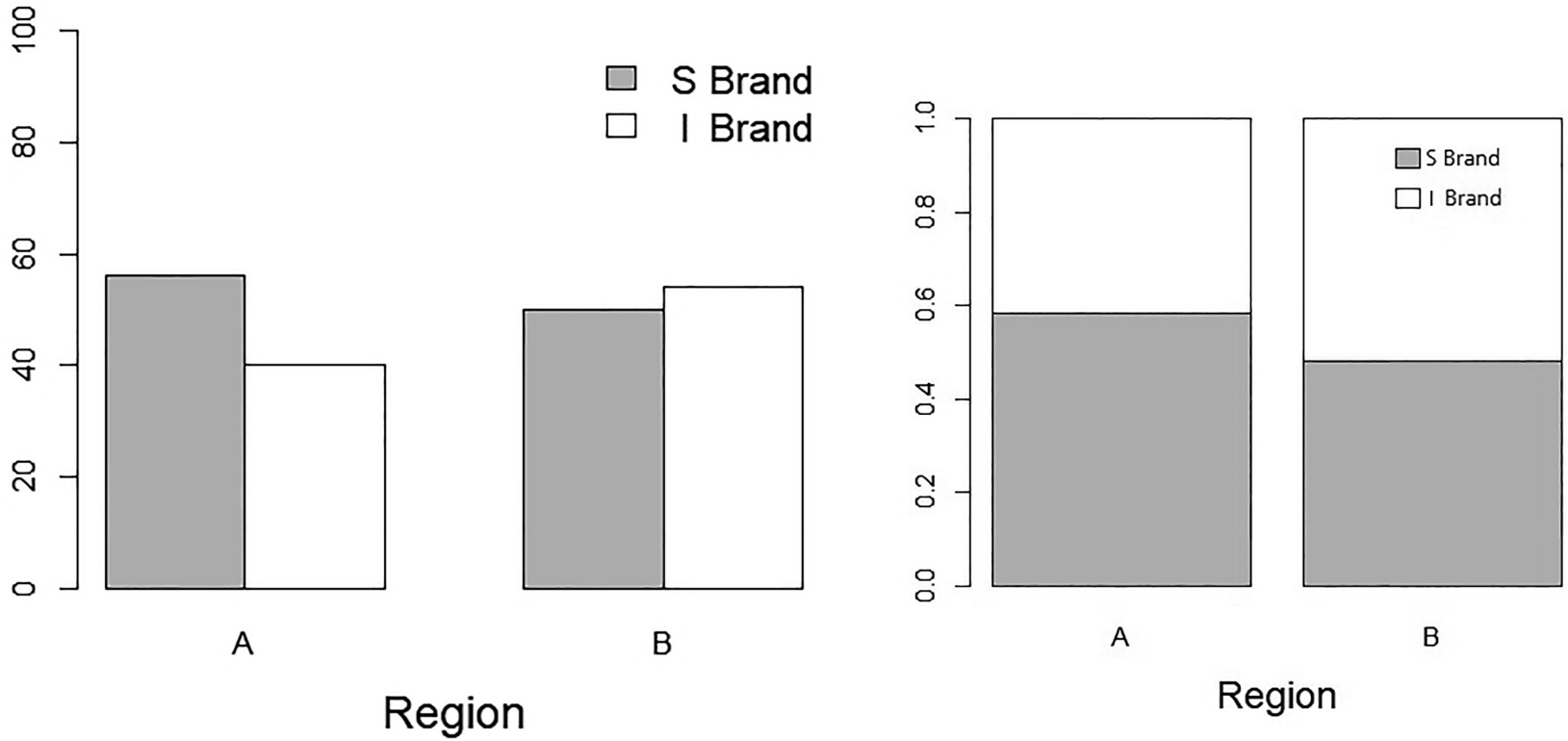
	S 브랜드	I 브랜드	행 합계
지역 A	0.58	0.42	1
지역 B	0.48	0.52	1
열 합계	0.53	0.47	1

두 변수 간 관계의 시각화: 그 밖의 경우

- 브랜드 선호도 비율 자체는 양적변수이다. 이것은 설문지의 문항(리커트 척도 또는 Y/N)과 별개의 문제임에 주의하자(Why?). 그리고 지역은 질적변수이다.
- 그러므로 양적변수와 질적변수의 관계를 분석하기 위해 양적변수의 그림을 질적변수별로 따로 그려서 비교하면 된다.
- 엑셀에서는 교차표 전체를 하이라이트한 다음, [삽입] 메뉴에서 [세로 막대형 차트 더 보기]를 눌러 신중하게 선택해야 한다(Why?).
- '묶은 세로 막대형' 그래프 두 개는 구체적으로 어떻게 다른지 고민해보자.

두 변수 간 관계의 시각화: 그 밖의 경우

- (1) 옆으로 나란히 비교하거나 (2) 상대비율을 비교할 수 있다.

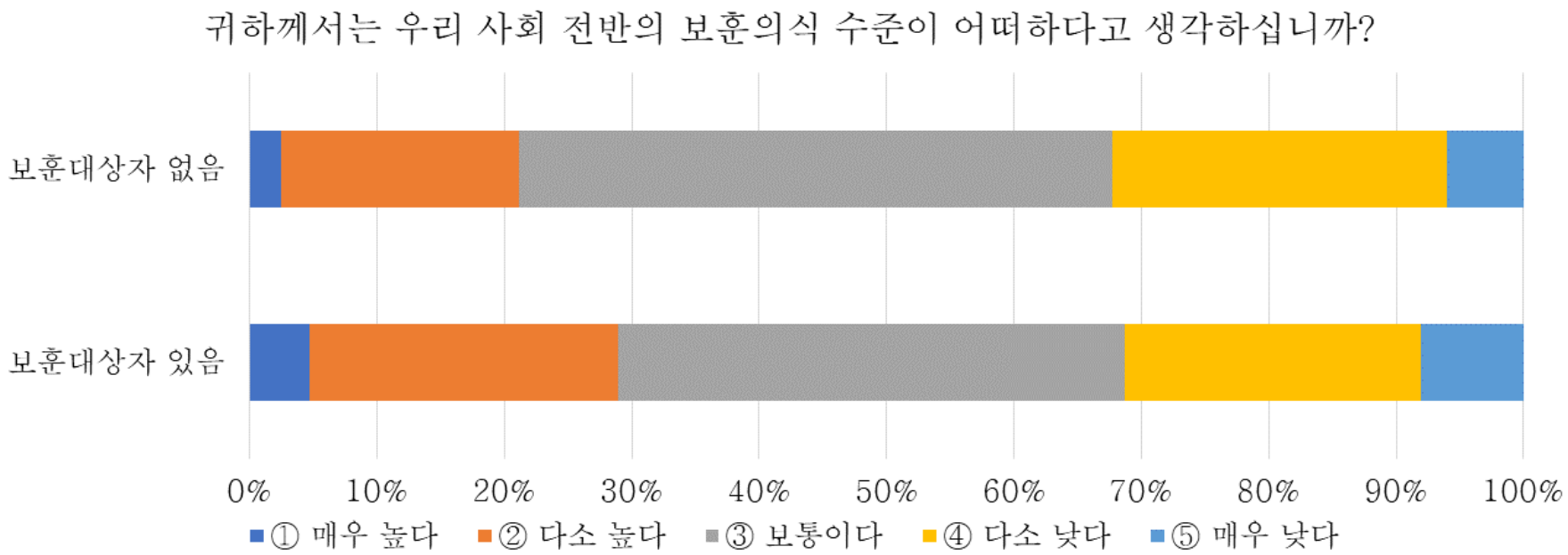


두 변수 간 관계의 시각화: 그 밖의 경우

예제 9. vpolicy.csv를 사용하여, 가족 중 보훈대상자 유무(DQ5)와 우리 사회 전반의 보훈의식(Q15) 사이의 연관성을 시각화하고 해석하시오.

두 변수 간 관계의 시각화: 그 밖의 경우

- 원자료(raw data)에서는 둘 다 질적변수이다. 그러나 연관성을 살펴보기 위해 일단 교차표를 만들고 나면, 표 안의 숫자는 양적변수처럼 시각화할 수 있다(Why?).



두 변수 간 관계의 시각화: 그 밖의 경우

꼭 막대 그래프가 아니라 상자-수염 도표를 활용할 수도 있다.

- 지난 주 우리는 양적변수와 질적변수의 연관성을 분석할 때, 양적변수에 대한 요약통계량을 나란히 구해 비교할 수 있다고 했다.
- 이와 마찬가지로 시각화할 수도 있다.

두 변수 간 관계의 시각화: 그 밖의 경우

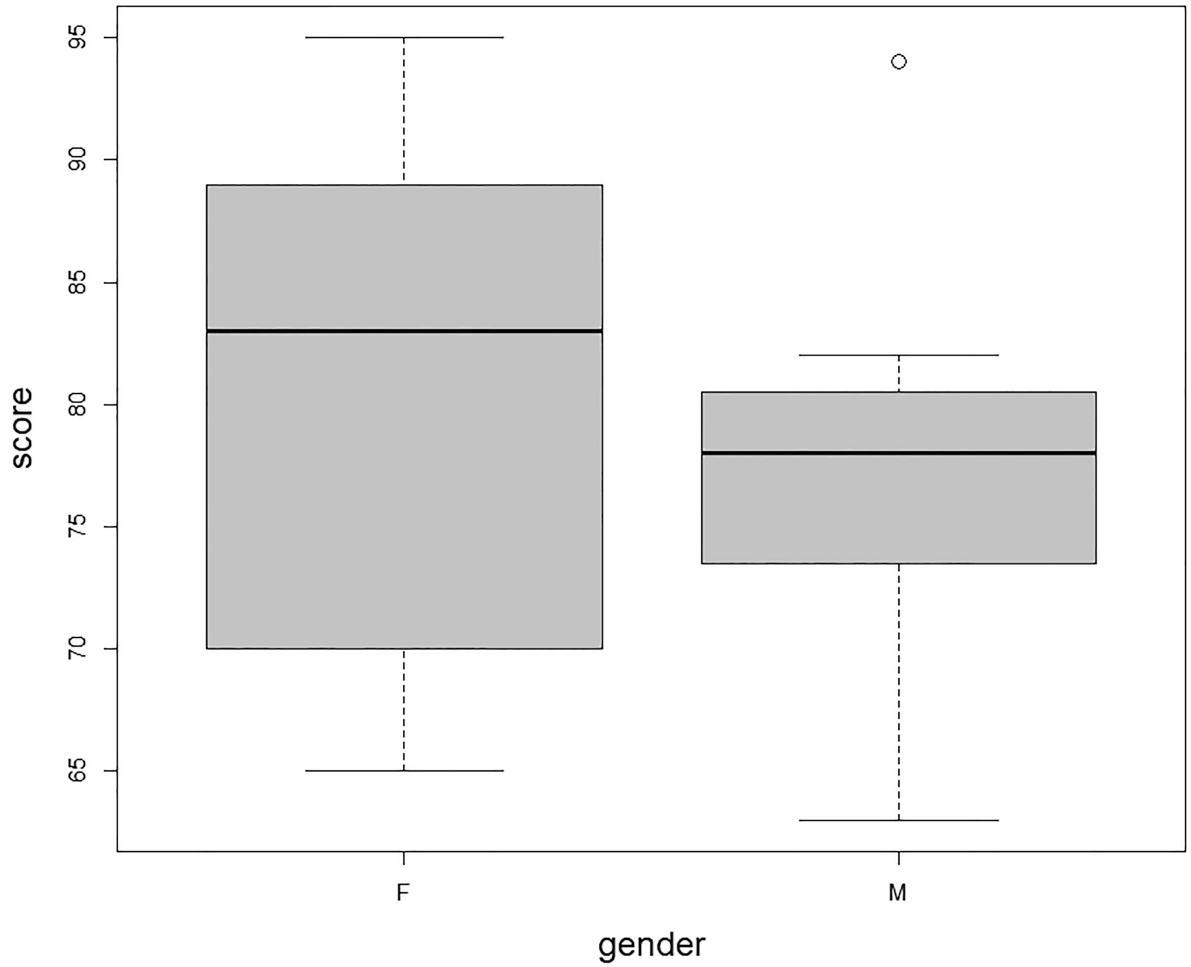
예제 10. math_score.csv는 남녀 학생 20명의 수학시험 점수를 나타낸 것이다. 두 변수의 연관성을 시각화하고 이를 해석하시오.

두 변수 간 관계의 시각화: 그 밖의 경우

- 엑셀 스프레드시트 안에 남녀 수학 점수를 나란히 재배치하자.
- 새로 재배치된 자료를 모두 하이라이트하고 다음, [삽입]-[모든 차트 보기]에서 “상자 수염”을 고르면 된다.
- 하한선을 다소 조정해야 한다(Why?).

두 변수 간 관계의 시각화: 그 밖의 경우

- 극단치(outliers)를 엑셀에서 곧바로 파악하기란 쉽지 않다.



시각화에 관한 코멘트

시각화에 관한 코멘트

"A picture is worth a thousand words."

- 데이터 시각화(data visualization)는 최근 수 년 사이에 폭발적인 인기를 얻고 있다.
- 데이터분석 밖에서도 몇 가지 중요한 추세가 발견된다! 웹 디자인 시장이 폭발적으로 성장했고 모바일 웹 다지인 역시 지속적으로 성장하고 있다.
- 유튜브, 틱톡, 인스타그램 등의 인기에 힘입어 사진/그림/동영상 제작 및 편집 역시 중요한 스킬 중 하나가 되었다.
- 데이터와 관련해서는 데이터 저널리즘(Data Journalism)의 성장이 이 분야를 리드한 것으로 보인다.
- 미국에서는 데이터 시각화 분야만으로 학위과정이 따로 있다. 한국에서도 데이터사이언스 분야가 급성장하고 있기 때문에 자신의 역량으로 고려해 볼 만한 분야.

시각화에 관한 코멘트

시각화는 졸업한 다음에는 무엇에 쓰이나?

- 데이터 시각화는 (물론 사회학의 연구 목적에도 유용하지만) 여러분이 정부조직, 비영리단체, 복지기관, 교육기관, 사기업 등에 자리를 잡은 뒤에도 기획서 · 보고서 · 대중 강연 등에 실용적인 쓸모를 갖는다.
- 여러분의 한참 선배 세대는 오로지 글로써 의도를 전달했다. 하지만 여러분은 이미 멀티미디어 세대이고 여러분의 후배 세대는 아예 글을 읽지 않으려 할지도 모른다. 시각화는 점진적으로 필수적인 의사소통의 수단이 되어가고 있다.

시각화에 관한 코멘트

시각화를 프로처럼 제대로 활용하려면 거의 필수적으로 JavaScript를 습득해야 한다.

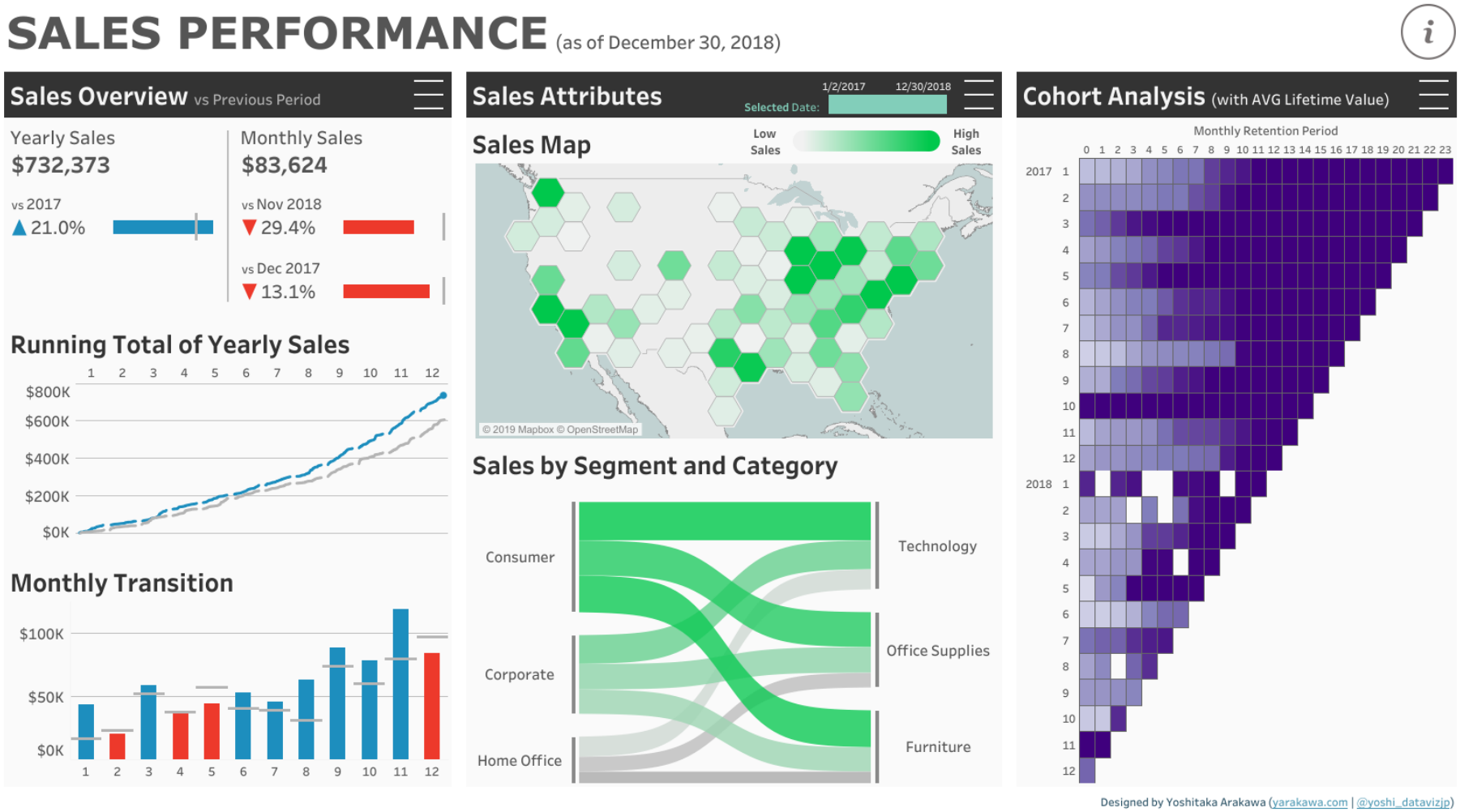
- Java 계열 자체가 플랫폼 독립적인 범용 스크립트 언어다.
- 세계 최대 규모의 개발자 커뮤니티. 국비지원 무료교육 프로그램, 취업 1순위 언어.
- 웹 프론트엔드(web front-end)를 디자인할 때 JavaScript를 모르고는 이야기가 안된다.
- 이 수업은 JavaScript를 다루지 않았다. 애초에 다룰 수도 없다. 별도로 한 학기 수업이 필요하기 때문이다. 게다가 나도 JavaScript를 모른다.
- 대규모 기관/조직에서는 대체로 JavaScript 전문가를 따로 고용하지만 상황이 꼭 그렇게 여의치 않을수도 있다.
- 대안적으로 다른 앱(app)을 사용하여 어느 정도 시각화를 흉내낼 수는 있고, BI 툴을 사용할 수도 있다.

시각화에 관한 코멘트

데이터 시각화는 경영지원의 핵심적인 도구 중 하나다.

- 경영지원 혹은 비즈니스 인텔리전스(Business Intelligence; BI) 차원에서 수많은 도구들이 출시되었다.
- 잘 알려진 툴 중에서는 (1) 태블로(Tableau), (2) 마이크로소프트의 Power BI, (3) Microstrategy 등이 있다. 대부분 고가의 툴이라 개인은 라이선스를 구매하기 어렵고 기관/조직에서 구매해야 한다.
- 다만 일부 프로그램은 기능제한부로 무료다. 어떤 프로그램은 대학생에게 특별히 무료로 풀려있다. 나도 대학(원)생 때는 무료로 사용했었다.
- 이런 것들은 주로 리포트(report) 또는 대시보드(dashboard)를 만들어 수집된 데이터의 요약 및 분석 결과를 실시간으로 조직 구성원과 공유하려는 목적을 갖는다.

시각화에 관한 코멘트


SALES PERFORMANCE (as of December 30, 2018)
Sales Overview vs Previous Period
Yearly Sales
$732,373
vs 2017
▲ 21.0%
Monthly Sales
$83,624
vs Nov 2018
▼ 29.4%
vs Dec 2017
▼ 13.1%
Running Total of Yearly Sales
$800K
$600K
$400K
$200K
$0K
Monthly Transition
$100K
$50K
$0K
Sales Attributes
1/2/2017
12/30/2018
Selected Date:
Sales Map
Low Sales
High Sales
© 2019 Mapbox © OpenStreetMap
Sales by Segment and Category
Consumer
Corporate
Home Office
Technology
Office Supplies
Furniture
Cohort Analysis (with AVG Lifetime Value)
Monthly Retention Period
2017
2018
Designed by Yoshitaka Arakawa (yarakawa.com | @yoshi_datavizjp)

시각화에 관한 코멘트

- 학술적인 측면에서는 다소 평가절하되고 있는 느낌(학자들은 대체로 그림보다는 글을 강조하기 때문)이다.
- 최근 데이터 저널리즘의 발전을 보면 앞으로 이 분야에서 지속적으로 전문가에 대한 수요가 있을 것임을 짐작케 한다.
- Google Analytics 등이 개인 또는 조직/기관이 운영하는 웹페이지의 방문자 데이터를 수집 및 분석하는데 중요한 도구로 급부상하였는데, 그 분석 결과를 어떻게 멋진 그래프로 정리하고 대시보드를 만들거나 리포트를 제작하는 것은 데이터 분석가 (Data Analyst)의 중요한 직무가 되었다.

시각화에 관한 코멘트

말할 필요도 없지만 이 수업에서 시각화를 전부 다룬게 아니다.

- 못 다룬 시각화 기법들, 예컨대 소셜 네트워크(social network)이나 애니메이션(animation) 뿐 아니라, AR/VR/XR 등 차세대 기법들은 중요하지 않아서 안 다룬게 아니다.
- "시각화"로 검색해서 이 주제에 관해 어떤 책들이 나와있는지를 쭉 살펴보면 대략적으로 이 분야가 어떤 느낌인지 파악할 수 있다.
- 나중에 조금 시간을 내서 큰 서점이나 도서관에 가보자. 인터넷 검색도 나쁘지 않지만 별로 체계적이진 않다.
- 다만 책을 보고 괜히 압도당하지는 말자. 배우지 않은 뒷부분을 미리 보면 겁을 먹기 쉽다. 막상 배우면 기초 정도는 생각보다 쉽게 한다.