V. 데이터 시각화

제 1장. 시각화 인사이트 프로세스

발표자:김은영
oneshot0229@gmail.com

시각화 인사이트 프로세스

1. 시각화 인사이트 프로세스의 의미
2. 탐색(1단계)
3. 분석(2단계)
4. 활용(3단계)

시각화 인사이트 프로세스의 의미

- **‘인사이트(insight, 통찰)’**

- 예리한 관찰력으로 사물을 환희 꿰뚫어 봄

- ‘정보, 인과관계, 본질, 이해’와 밀접한 관련

STEP1 검색어 트렌드 조회 | STEP2 사용자 데이터 융합 | STEP3 데이터 공유하기

STEP1

네이버에서 얼마나 많이 검색되는지 궁금한 주제가 있으신가요?

궁금한 주제어를 설정하고, 하위 주제어에 해당하는 검색어를 콤마(,)로 구분 입력해 주세요.
입력한 단어의 추이를 하나로 합산하여 해당 주제가 네이버에서 얼마나 검색되는지 관련 데이터를 제공합니다.

예) 주제어 캠핑 : 캠핑, Camping, 캠핑용품, 겨울캠핑, 캠핑장, 글램핑, 오토캠핑, 캠핑카, 텐트, 캠핑요리

전체선택 초기화

대상 ● 네이버 통합 검색어 ○ 네이버 쇼핑 검색어

기간 전체 | 1개월 | 3개월 | 1년 | 직접입력 | 2007 01 ~ 2017 02

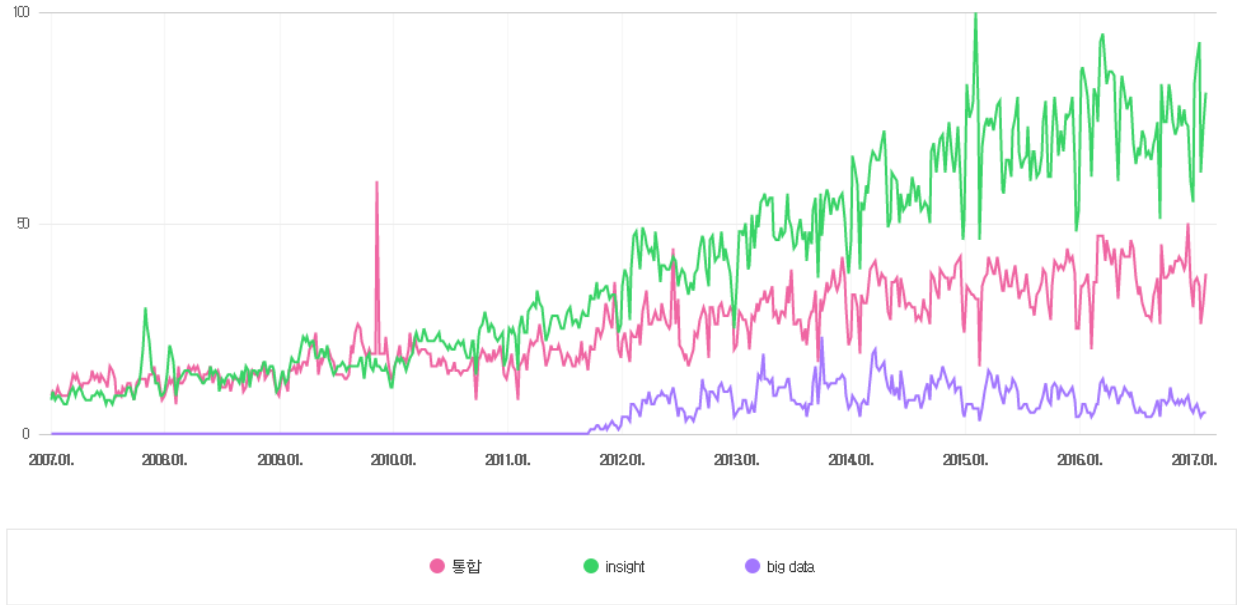
주제어1 통합 / 통합

주제어2 insight / insight

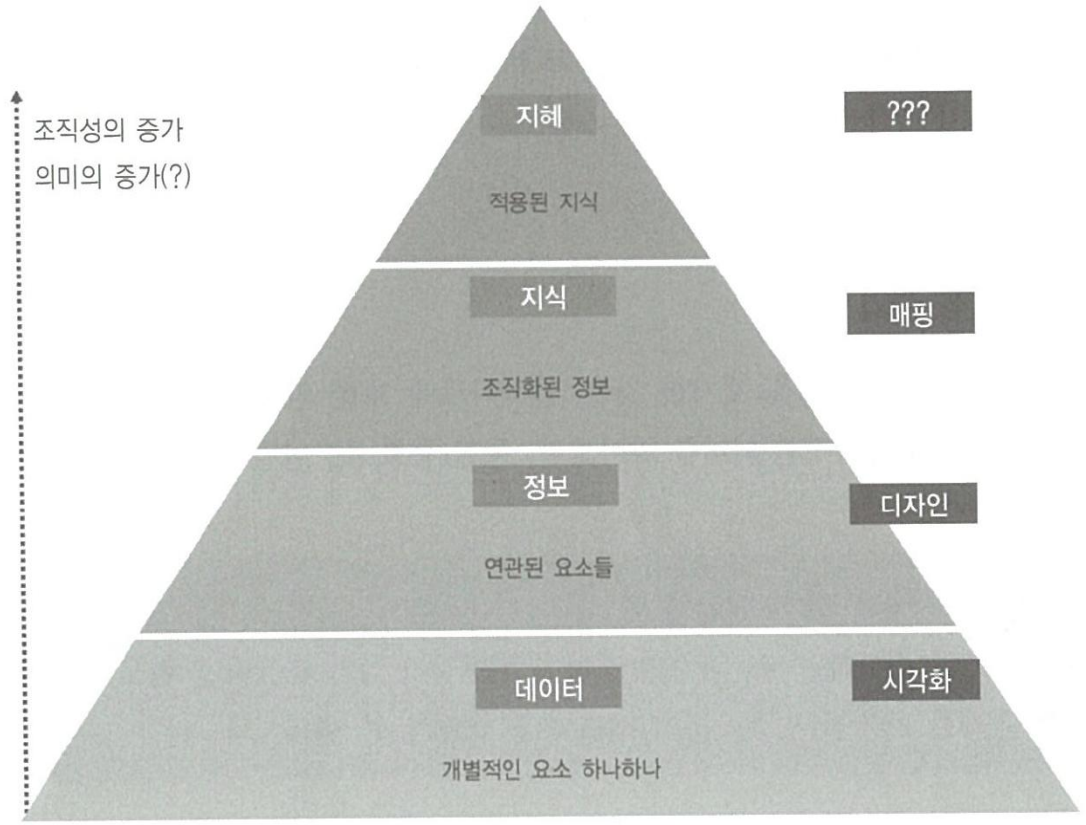
주제어3 주제어 3 입력 / 주제어 3에 해당하는 모든 검색어를 콤마(,)로 구분하여 최대 20개

주제는 최대 3개 까지 설정 가능하며, 한 주제당 최대 20개의 검색어를 추가할 수 있습니다.

조회하기

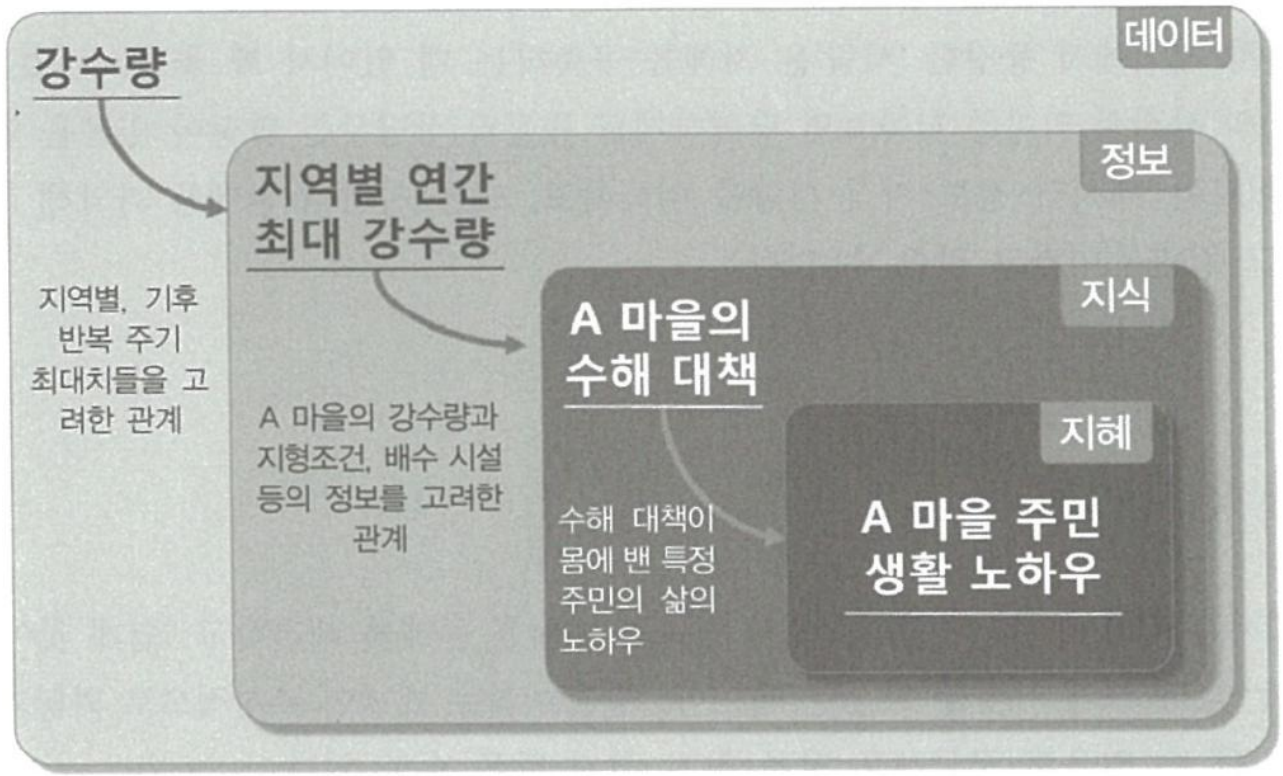


시각이해의 계층도
(Hierarchy of Visual Understanding)



✓ 데이비드 멕켄들레스는 시각이해의 계층도를 통해 데이터, 정보, 지식, 지혜 사이의 계층적관계를 시각적으로 표현

-개별 데이터간의 연결고리를 찾아서 관계가 생성될 때 정보
-정보가 보다 상위계념에서 (특히 인간과의 삶과) 관계를 맺고 조직화 되었을 때 지식
-지식들이 개인의 경험, 사고, 감정의 체계와 결합되고 관계를 맺고 구조화 될때, 지혜 (개인화한 지식)

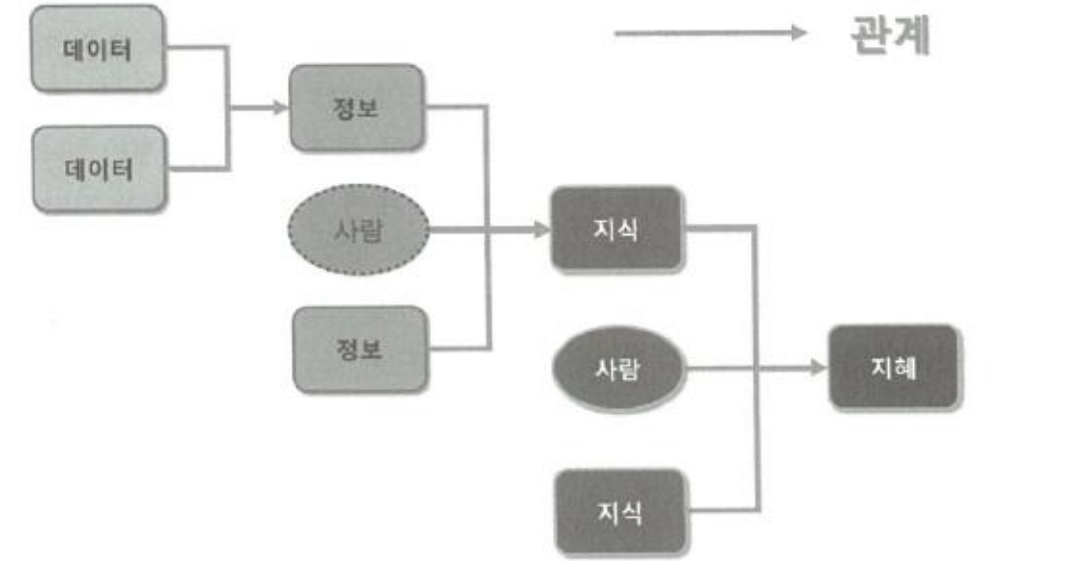


[그림 V-1-3] 강수량 데이터가 지혜로 발전하는 단계

개별강수량 (데이터)이 공간과 시간이라는 관계를 고려해 **지역별 연간 최대 강수량**이라는 **정보**가 형성

이정보가 "A 마을"이라는 특정 지역의 다른 성격의 정보가 결합되면 **그 마을의 수해대책 매뉴얼과 같은 지식**이 형성

주민별로 생활방식, 가치관이 다르기 때문에 수해로 인한 농경지 패해 최소화 노화우를 내놓을 주민도 있음(**지식의 개인화→지혜**)



[그림 V-1-4] 데이터, 정보, 지식, 지혜, 사람 사이의 관계

즉, 상위 개념으로 발전하기 위해서 중요한 요소: '관계', '사람'

시각화와 인사이트

- 대상(데이터, 정보, 지식) 사이에 숨어있는 관계를 찾아 융합된 상위개념을 발견 하는 것

삼찰(관찰, 성찰, 통찰)의 상관관계
관찰: 외부세계
성찰: 내면세계
통찰: 외부와 내면세계 사이

통찰을 추출하는 전 과정(시각화)
1단계:탐색
2단계:분석
3단계:활용

[그림 V-1-6] 시각화 인사이트 프로세스의 전체 개요

예시: 당일 할인 숙박권 판매 서비스의 최적화

- 서비스개요

당일 땡처리 호텔 숙박권을 판매하는 모바일 앱서비스

;이용자의 위치 기반으로 특정지역의 상품목록 조회, 원하는 상품 세부조회 및 결제, 이용후기와 평점 남겨 다른 사용자가 참고케 함

- 시각화인사이트 프로세스의 목표

어떤 상품을 소싱하고 어떻게 전시하

느냐가 매출규모 결정

[표 V-1-1] 서비스 데이터의 일부

구분	항목명	데이터형	설명
상품 전시 데이터	전시일	날짜형 YYYY-MM-DD	해당 상품을 이용자들이 조회할 수 있게 전시한 날짜. 날짜의 기준은 자정. 즉, 당일 0시부터 익일 0시 이전까지. 만약 동일 상품이 2일 이상에 걸쳐 전시되었을 경우 각 날짜별 데이터 집합 모두를 저장
	전시 위치	정수형	서비스 화면에서 상품이 전시된 위치. 첫 목록 화면의 최상단이 0, 아래로 내려갈수록 1씩 증가
	전시 상품 코드	문자형	전시된 상품의 고유 코드명
상품 조회 데이터	조회 시각	날짜형+시간형 YYYY-MM-DD hh:mm:ss	이용자가 전시 상품의 상세 화면을 살펴본 시각
	조회 상품 코드	문자형	조회된 상품의 고유 코드명
상품 결제 데이터	결제 시각	날짜형+시간형	이용자가 상품을 결제한 시각
	결제 금액	정수형	상품의 결제 금액
	결제 상품 코드	문자형	결제 상품의 고유 코드명

1단계:탐색

1. 사용 가능한 데이터의 확인

데이터의 명세화(어떤 데이터를 시용할 수 있는지): 차원(값이 측정된 기준)과 측정값
데이터구성원리: 이벤트 기록으로서 접근 (‘순간 동시접속자 수’, ‘일일 액티브 이용자수’)
객체지향 관점에서의 접근 (6하 원칙에 의거; object, class, method)

2. 연결고리의 확인

공통 요소 찾기
공통 요소로 변환하기:
시간데이터의 변환, 공간데이터의 변환, 일정한 규칙을 가진 분류형 데이터의 변환
탐색범위의 설정

3. 관계의 탐색

이상값 처리
차원과 측정값 유형에 따른 관계 파악 시각화
잘라보고 달리보기
내려다보고 올려다보기
척도의 조정

이벤트 기록으로서 접근

[표 V-1-2] 서비스 접속 로그 데이터 테이블 사례

Login_ID	Login_Time	Channel
honggildong	2013-12-01 14:23:12	android
choonhyang	2013-12-01 14:30:43	chrome
honggildong	2013-12-01 23:11:10	ios
honggildong	2013-12-02 11:31:09	chrome

2013-12-01
순간 동시 접속자 수: 2
일일 액티브 이용자수: 2

[표 V-1-1] 서비스 데이터의 일부

구분	항목명	데이터형	설명
상품 전시 데이터	전시일	날짜형 YYYY-MM-DD	해당 상품을 이용자들이 조회할 수 있게 전시한 날짜. 날짜의 기준은 자정. 즉, 당일 0시부터 익일 0시 이전까지. 만약 동일 상품이 2일 이상에 걸쳐 전시되었을 경우 각 날짜별 데이터 집합 모두를 저장
	전시 위치	정수형	서비스 화면에서 상품이 전시된 위치. 첫 목록 화면의 최상단이 0, 아래로 내려갈수록 1씩 증가
	전시 상품 코드	문자형	전시된 상품의 고유 코드명
상품 조회 데이터	조회 시각	날짜형+시간형 YYYY-MM-DD hh:mm:ss	이용자가 전시 상품의 상세 화면을 살펴본 시각
	조회 상품 코드	문자형	조회된 상품의 고유 코드명
상품 결제 데이터	결제 시각	날짜형+시간형	이용자가 상품을 결제한 시각
	결제 금액	정수형	상품의 결제 금액
	결제 상품 코드	문자형	결제 상품의 고유 코드명

[표 V-1-4] 보완된 서비스 로그 데이터 테이블

구분	항목명	데이터형	설명
class 상품 데이터	상품 코드	문자형	개별 상품의 고유 식별 코드
	생성일	날짜형 YYYY-MM-DD	상품이 구성된 날짜. 한 번 만들어진 상품은 상황에 따라 전시하거나 내림으로써 여러 차례에 걸쳐 전시될 수 있음
	호텔 명	문자형	호텔의 명칭
	호텔 지역 (도·시)	문자형	호텔이 위치한 지역 구분. 서울특별시, 강원도 등의 도·시 단위
	호텔 지역 (구·군)	문자형	호텔이 위치한 지역 구분. 강남구, 양양군 등 구·군 단위
	호텔 등급	문자형	특1·특2·1·2·3 등의 등급
	방 구성	문자형	(킹베드, 퀸베드) + (싱글, 더블, 트윈)의 조합으로 된 방 구성 데이터
	정가	정수형	정상 가격
	할인가	정수형	할인 판매하는 가격
상품전시 데이터	전시일	날짜형 YYYY-MM-DD	해당 상품을 이용자들이 조회할 수 있게 전시한 날짜. 날짜의 기준은 자정. 즉, 당일 0시부터 익일 0시 이전까지. 만약 동일 상품이 2일 이상에 걸쳐 전시되었을 경우 각 날짜별 데이터 집합 모두를 저장
	전시 위치	정수형	서비스 화면에서 상품이 전시된 위치. 첫 목록 화면의 최상단이 0, 아래로 내려갈수록 1씩 증가
	전시 상품 코드	문자형	전시된 상품의 고유 코드명

구분	항목명	데이터형	설명
상품조회 데이터	조회 시각	날짜형+시간형 YYYY-MM-DD hh:mm:ss	이용자가 전시된 해당 상품의 상품 상세 화면을 살펴본 시각
	조회 정렬조건	문자형	기본 전시 상태, 가격 또는 인기도 기준 정렬 등 현재 상품이 조회된 조건값
	조회 실행지점	문자형	조회한 시점의 실시간 모바일 기기 위치 정보값. 위도와 경도 쌍으로 기록됨
	사진조회 횟수	정수형	상세 설명 화면에서 추가 조작을 통해 제공된 호텔 관련 사진들을 본 전체 횟수
	조회 상품 코드	문자형	조회된 상품의 고유 코드명
상품결제 데이터	결제 시각	날짜형+시간형	이용자가 상품을 결제한 시각
	결제실행지점	문자형	결제를 실행한 시점의 실시간 모바일 위치 정보값. 위도와 경도 쌍으로 기록됨
	결제 금액	정수형	상품이 결제된 금액
	결제 상품 코드	문자형	결제된 상품의 고유 코드명
상품리뷰 데이터	리뷰 시각	날짜형+시간형	이용자가 상품을 리뷰한 시각
	리뷰 실행 지점	문자형	리뷰 실행 시점의 실시간 모바일 기기 위치 정보값. 위도와 경도 쌍으로 기록됨
	리뷰 종류	문자형	댓글과 평점 중에서 하나
	댓글 내용	문자형	댓글 전체 텍스트. 만약 종류가 평점인 경우는 null
	평점	정수형	매긴 평점. 1~5점 사이의 정수형. 만약 종류가 댓글이면 0
	리뷰 상품 코드	문자형	리뷰된 상품의 고유 코드명

하나의 완결된 오브젝트를 구분하는 대표값

상품이라는 클래스는 호텔등급과 같은 구체적인 속성값을 가짐.

상품(class)은 전시, 조회, 결제, 리뷰라는 행위(method)를 가지며 그 행위는 그에 따른 속성값을 가짐.

1단계:탐색

1. **사용 가능한 데이터의 확인**
   데이터의 명세화(어떤 데이터를 시용할 수 있는지): 차원(값이 측정된 기준)과 측정값
   데이터구성원리: 이벤트 기록으로서 접근 (‘순간 동시접속자 수’, ‘일일 액티브 이용자수’)
   객체지향 관점에서의 접근 (6하 원칙에 의거; object, class, method)

2. **연결고리의 확인**
   공통 요소 찾기
   공통 요소로 변환하기:
   시간데이터의 변환, **공간데이터의 변환**, **일정한 규칙을 가진 분류형 데이터의 변환**
   탐색범위의 설정

3. **관계의 탐색**
   이상값 처리
   차원과 측정값 유형에 따른 관계 파악 시각화
   잘라보고 달리보기
   내려다보고 올려다보기
   척도의 조정

공간데이터의 변환

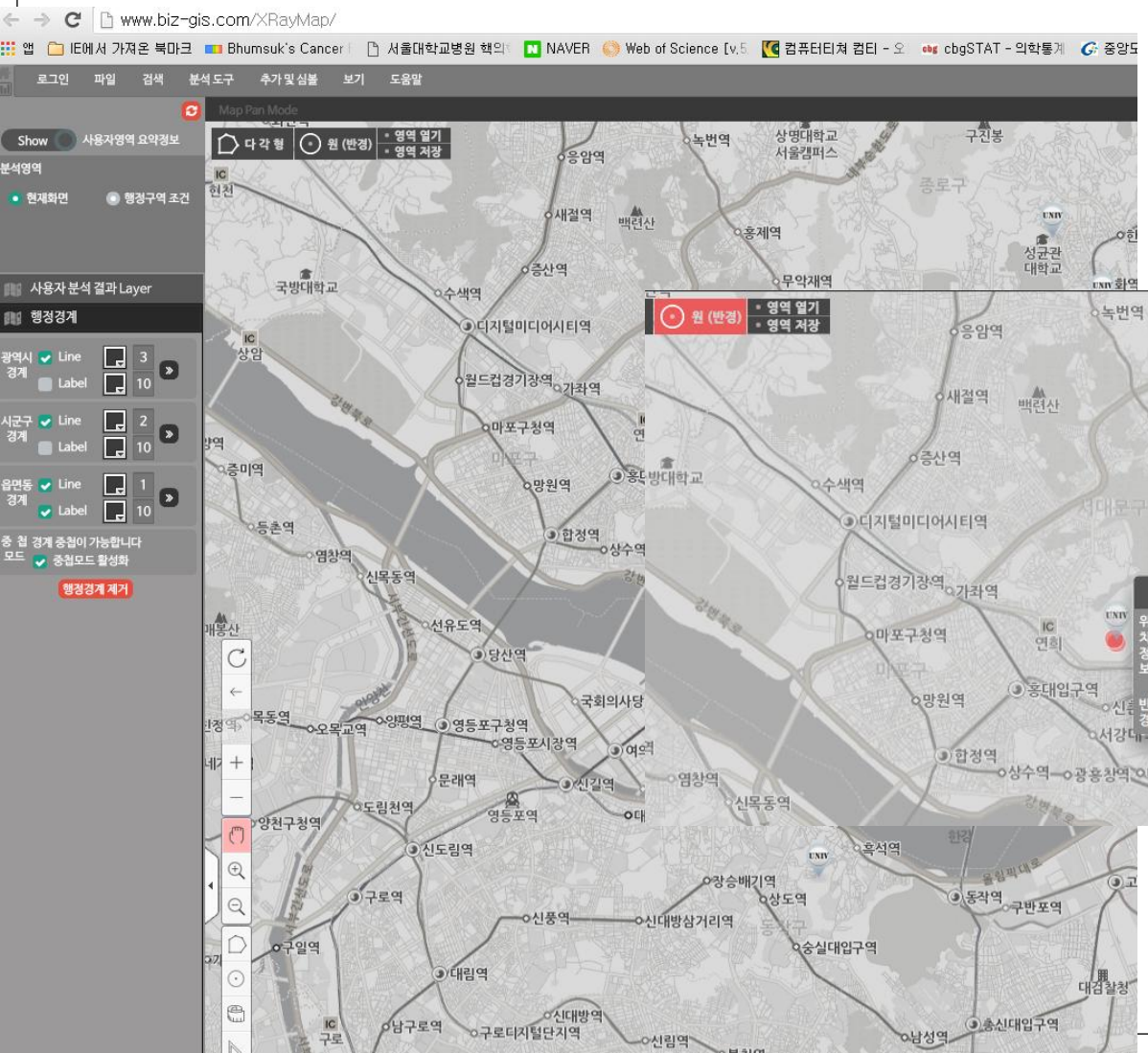
공간데이터 유형: 주소, 세분화 계층형 행정구역(시도군리 등), 좌표값

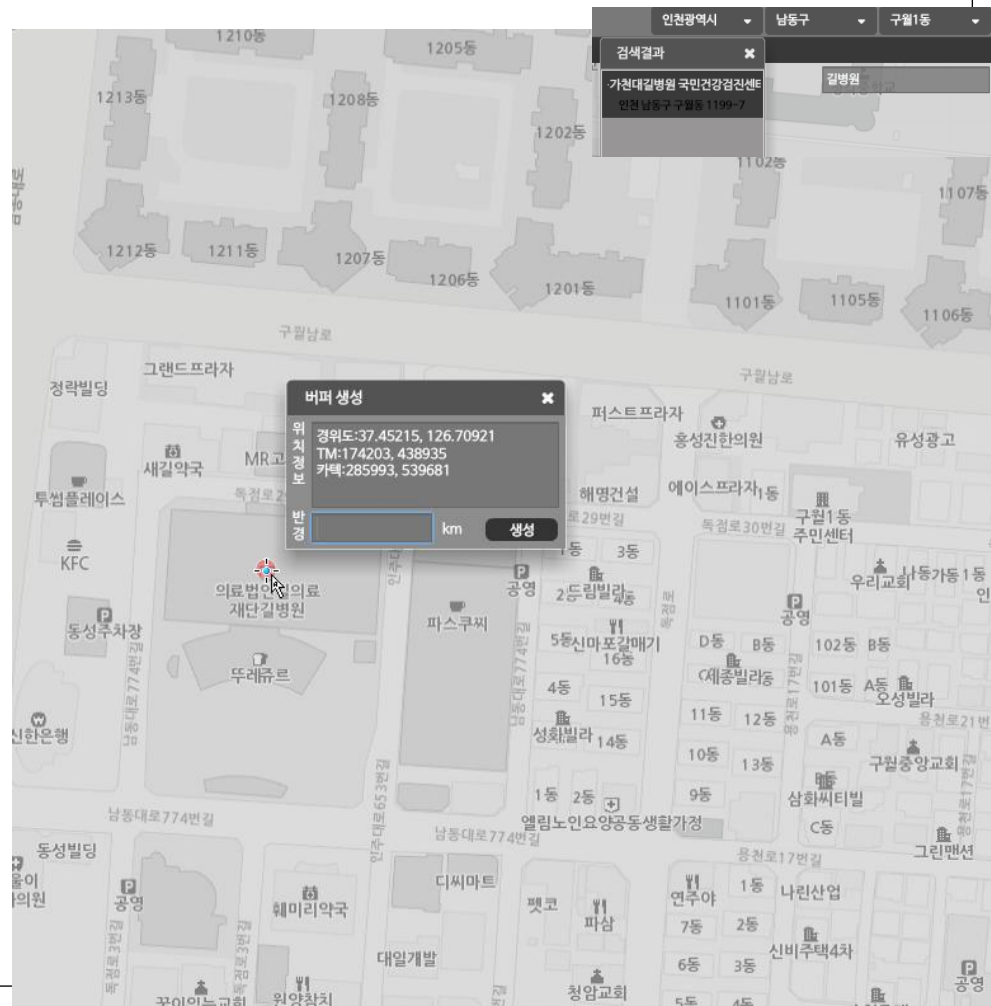
- ✓ 주소: 공백문자를 기준으로 분할 (엑셀의 '텍스트 나누기')
- ✓ 좌표계를 행정구역으로 변환 (지오코딩, geocoding)

ex) 비즈 GIS의 무료 웹 GIS 분석도구 X-ray (http://www.biz-gis.com/XRayMap/)

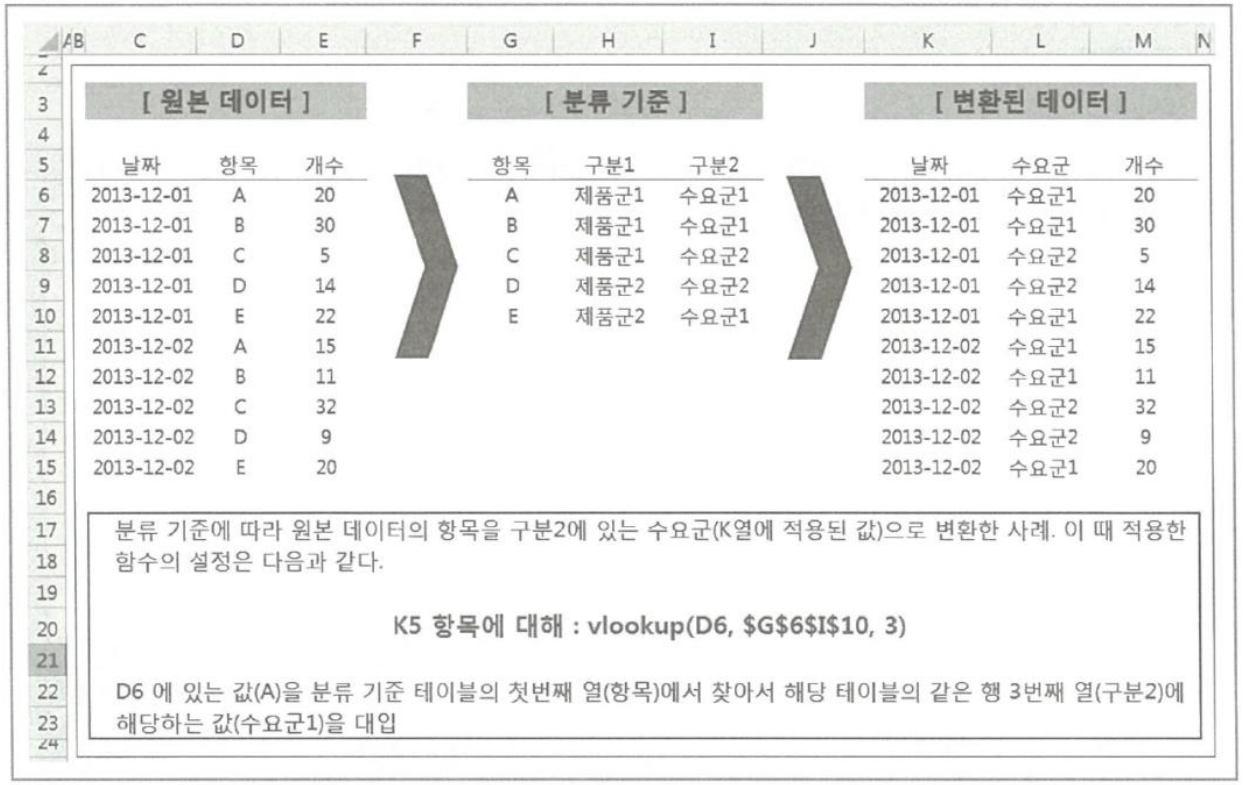
[표 V-1-5] 스프레드시트에서 기본으로 제공하는 문자열 처리 함수들

함수명	함수 사용 형태	함수 기능 설명
split	split(문자열, 구분자)	문자열을 구분 문자(공백이나 쉼표 등) 기준으로 뜯어서 제공
find	find(찾는 문자, 문자열)	문자열에서 찾는 문자가 맨 왼쪽으로부터 몇 번째에 있는지 숫자값 제공
left	left(문자열, 개수)	문자열의 맨 왼쪽부터 정해진 개수만큼의 문자열 제공
mid	mid(문자열, 시작 위치, 개수)	문자열의 시작 위치에서부터 정해진 개수만큼의 문자열 제공
split	split(문자열, 구분자)	문자열을 구분 문자(공백이나 쉼표 등) 기준으로 뜯어서 제공





분류형 데이터의 변환- 연결고리 탐색



[원본 데이터]

날짜	항목	개수
2013-12-01	A	20
2013-12-01	B	30
2013-12-01	C	5
2013-12-01	D	14
2013-12-01	E	22
2013-12-02	A	15
2013-12-02	B	11
2013-12-02	C	32
2013-12-02	D	9
2013-12-02	E	20

[분류 기준]

항목	구분1	구분2
A	제품군1	수요군1
B	제품군1	수요군1
C	제품군1	수요군2
D	제품군2	수요군2
E	제품군2	수요군1

[변환된 데이터]

날짜	수요군	개수
2013-12-01	수요군1	20
2013-12-01	수요군1	30
2013-12-01	수요군2	5
2013-12-01	수요군2	14
2013-12-01	수요군1	22
2013-12-02	수요군1	15
2013-12-02	수요군1	11
2013-12-02	수요군2	32
2013-12-02	수요군2	9
2013-12-02	수요군1	20

분류 기준에 따라 원본 데이터의 항목을 구분2에 있는 수요군(K열에 적용된 값)으로 변환한 사례. 이 때 적용한 함수의 설정은 다음과 같다.

K5 항목에 대해 : vlookup(D6, G6I10, 3)

D6 에 있는 값(A)을 분류 기준 테이블의 첫번째 열(항목)에서 찾아서 해당 테이블의 같은 행 3번째 열(구분2)에 해당하는 값(수요군1)을 대입

[그림 V-1-9] vlookup 함수를 이용한 데이터 묶음 변환

분류형 데이터의 변환- 연결고리 탐색 예

[표 V-1-4] 보완된 서비스 로그 데이터 테이블

구분	항목명	데이터형	설명
상품 데이터	상품 코드	문자형	개별 상품의 고유 식별 코드
	생성일	날짜형 YYYY-MM-DD	상품이 구성된 날짜. 한 번 만들어진 상품은 상황에 따라 전시하거나 내림으로써 여러 차례에 걸쳐 전시될 수 있음
	호텔 명	문자형	호텔의 명칭
	호텔 지역 (도·시)	문자형	호텔이 위치한 지역 구분. 서울특별시, 강원도 등의 도·시 단위
	호텔 지역 (구·군)	문자형	호텔이 위치한 지역 구분. 강남구, 양양군 등 구·군 단위
	호텔 등급	문자형	특1·특2·1·2·3 등의 등급
	방 구성	문자형	(킹베드, 퀸베드) + (싱글, 더블, 트윈)의 조합으로 된 방 구성 데이터
	정가	정수형	정상 가격
	할인가	정수형	할인 판매하는 가격
상품전시 데이터	전시일	날짜형 YYYY-MM-DD	해당 상품을 이용자들이 조회할 수 있게 전시한 날짜. 날짜의 기준은 자정. 즉, 당일 0시부터 익일 0시 이전까지. 만약 동일 상품이 2일 이상에 걸쳐 전시되었을 경우 각 날짜별 데이터 집합 모두를 저장
	전시 위치	정수형	서비스 화면에서 상품이 전시된 위치. 첫 목록 화면의 최상단이 0, 아래로 내려갈수록 1씩 증가
	전시 상품 코드	문자형	전시된 상품의 고유 코드명

구분	항목명	데이터형	설명
상품조회 데이터	조회 시각	날짜형+시간형 YYYY-MM-DD hh:mm:ss	이용자가 전시된 해당 상품의 상품 상세 화면을 살펴본 시각
	조회 정렬조건	문자형	기본 전시 상태, 가격 또는 인기도 기준 정렬 등 현재 상품이 조회된 조건값
	조회 실행지점	문자형	조회한 시점의 실시간 모바일 기기 위치 정보값. 위도와 경도 쌍으로 기록됨
	사진조회 횟수	정수형	상세 설명 화면에서 추가 조작을 통해 제공된 호텔 관련 사진들을 본 전체 횟수
	조회 상품 코드	문자형	조회된 상품의 고유 코드명
상품결제 데이터	결제 시각	날짜형+시간형	이용자가 상품을 결제한 시각
	결제실행지점	문자형	결제를 실행한 시점의 실시간 모바일 위치 정보값. 위도와 경도 쌍으로 기록됨
	결제 금액	정수형	상품이 결제된 금액
	결제 상품 코드	문자형	결제된 상품의 고유 코드명
상품리뷰 데이터	리뷰 시각	날짜형+시간형	이용자가 상품을 리뷰한 시각
	리뷰 실행 지점	문자형	리뷰 실행 시점의 실시간 모바일 기기 위치 정보값. 위도와 경도 쌍으로 기록됨
	리뷰 종류	문자형	댓글과 평점 중에서 하나
	댓글 내용	문자형	댓글 전체 텍스트. 만약 종류가 평점인 경우는 null
	평점	정수형	매긴 평점. 1~5점 사이의 정수형. 만약 종류가 댓글이면 0
	리뷰 상품 코드	문자형	리뷰된 상품의 고유 코드명

[표 V-1-6] 서비스 로그 데이터 테이블에서 찾아내 변환한 공통 요소

연결고리	설명
상품코드	상품 속성 및 상품과 관련된 모든 액션들을 살펴보는 연결고리
날짜	상품의 생성, 전시, 조회, 판매, 리뷰라는 전체 서비스 사이클을 살펴볼 수 있는 연결고리
장소	상품(호텔)이 위치하는 행정관리상의 지역 및 조회·판매·리뷰되는 지역에 대한 연결고리. 모두 도·시, 구·군 단위로 값이 지정됨

탐색범위의 설정

- 모든 가능성을 열어두되 우선순위를 설정

  ✓개별데이터 내 ->전체 데이터 집합 내에서 탐색

  ✓측정값에 한 차원만 연결 -> 단계적으로 연결 차원 증가

  ✓같은 데이터 내 차원과 측정값을 맞바꾸어 보기

  ✓비주얼 인사이트 프로세스를 적용하여 목표에 관련됐을 법한 조합 우선
  (연결할수 있는 모든 조합을 살펴보는 것 보다는)

  ✓상식적으로 의미나 연계성이 없는 조합은 가급적 배제

[표 V-1-7] 당일 할인 숙박권 판매 서비스 최적화 인사이트 도출을 위한 탐색 범위 설정

구분	차원	측정값	살펴보려는 탐색의 세부 설명
상품 특성	호텔 지역, 호텔 등급, 전시일	할인가	전시 날짜별, 지역별, 등급별 호텔의 할인 가격 분포. 날짜는 일·월·분기로 살펴볼 수 있게 함
	호텔 지역, 호텔 등급	방구성	지역과 등급에 따라 방구성 종류가 어떤 패턴을 띠는지 방구성 종류들의 개수를 측정값으로 파악
	호텔 지역, 호텔 등급	상품코드	지역별·등급별로 어떤 호텔들이 어떤 밀집도로 분포하는지 상품코드의 개수로 살펴봄. 이때 지역은 도·시 및 구·군 단위로 계층을 나누어 살펴볼 수 있게 함
상품 조회	조회 시각, 호텔 지역, 방구성	조회 상품코드	특정 기간에 상품 조회가 발생한 모든 건에 대해 어떤 지역의 어떤 방구성의 상품들이 많이 조회됐는지 살펴봄
	조회 실행 지점, 호텔 지역	조회 상품코드	이용자들이 일반적으로 현재 위치에서 근처의 상품들에 대한 관심이 높은지를 탐색
	호텔 등급, 방구성, 조회 정렬조건	조회 상품코드	전 기간의 전체 조회 건에 대해 정렬 조건에 따라 실제 조회까지 이뤄진 상품은 어떤 등급과 방구성 특성을 지니는지 파악
상품 결제	결제 시각, 조회 시각	결제 상품코드	각각의 결제 건에 대해 조회 후 얼마나 빠른 시간 내에 결제가 이뤄지는지, 시간 딜레이 간격에 따른 전체 결제건 분포를 살펴봄
	호텔 지역, 호텔 등급, 결제 금액, 결제 시각	결제 상품코드	지역별·등급별로 일정 기간 내에 결제된 건수에 대해 언제 얼마의 금액대 상품들이 결제되는지에 대한 전체 패턴을 살펴봄

구분	차원	측정값	살펴보려는 탐색의 세부 설명
	호텔 등급, 방구성	결제 상품코드	호텔 등급별 방구성에 따라 매출 규모를 봄으로써 인기 등급별 방구성을 파악. 또한 가격을 추가로 넣어 봄으로써 방구성은 별다른 의미가 없고 가격 중심으로 의사결정을 하는가를 살펴봄
상품 리뷰	리뷰 시각, 리뷰 실행지점, 결제시각, 호텔 지역	리뷰 상품코드	상품을 이용한 고객들이 이용 후 언제 어디서 리뷰를 하는지에 대한 전체적인 패턴을 파악. 호텔 지역과 결제 시각을 같이 살펴봄으로써 리뷰 액션의 즉시성에 대해 탐색
	호텔 지역, 호텔명	평점	개별 호텔 및 호텔들의 지역들 집합 기준으로 평점 분포를 탐색
	호텔 지역, 호텔명	댓글 내용	개별 호텔 및 호텔들의 지역들 집합 기준으로 댓글들이 어떤 키워드와 감성 중심으로 분포되는지 파악
	할인가	평점, 댓글 내용	상품의 가격과 리뷰의 정성적인 내용 및 정량적인 평점이 어떤 관계가 있는지 탐색

1단계:탐색

1. 사용 가능한 데이터의 확인

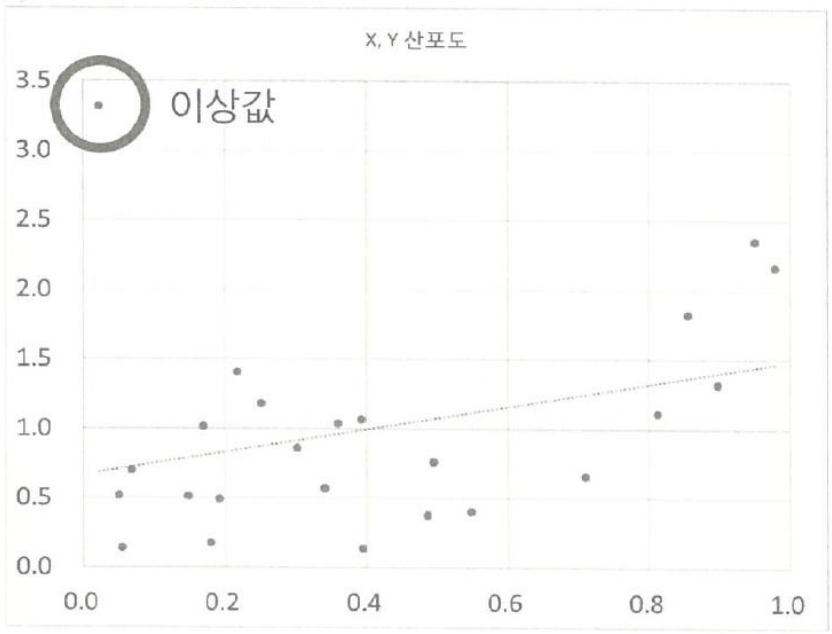
데이터의 명세화(어떤 데이터를 시용할 수 있는지): 차원(값이 측정된 기준)과 측정값
데이터구성원리: 이벤트 기록으로서 접근 (‘순간 동시접속자 수’, ‘일일 액티브 이용자수’)
객체지향 관점에서의 접근 (6하 원칙에 의거; object, class, method)

2. 연결고리의 확인

공통 요소 찾기
공통 요소로 변환하기:
시간데이터의 변환, 공간데이터의 변환, 일정한 규칙을 가진 분류형 데이터의 변환
탐색범위의 설정

3. 관계의 탐색 (본격적 탐색)

이상값 처리
차원과 측정값 유형에 따른 관계 파악 시각화
잘라보고 달리보기
내려다보고 올려다보기
척도의 조정



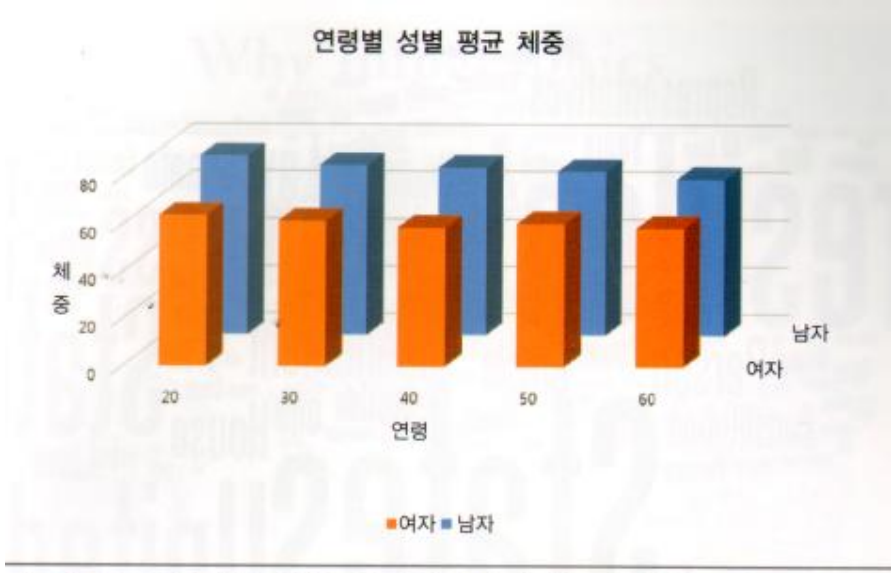
[그림 V-1-10] 산포도를 이용한 이상값 찾기

이상값(outlier): 다른 관측값과 동떨어진 값

- ✓ 측정 오류 (제거의 대상)
  예; 절대 0이나 10,000이상의 값이 나올수 없는 데이터인데 그런 값이 나온 경우
- ✓ 기록.관리 오류
  예; 정수형 데이터가 있어야 할 곳에 문자형 데이터
- ✓ 무언가 의미 있는 이유 (구체적으로 파고들어야 할 대상)

→ 산포도로 이상값을 찾고 자료의 구조를 보고 판단

차원과 측정값 유형에 따른 관계파악 시각화



[그림 V-1-11] 차원 2개로 된 측정값 데이터를 시각화한 그래프의 예

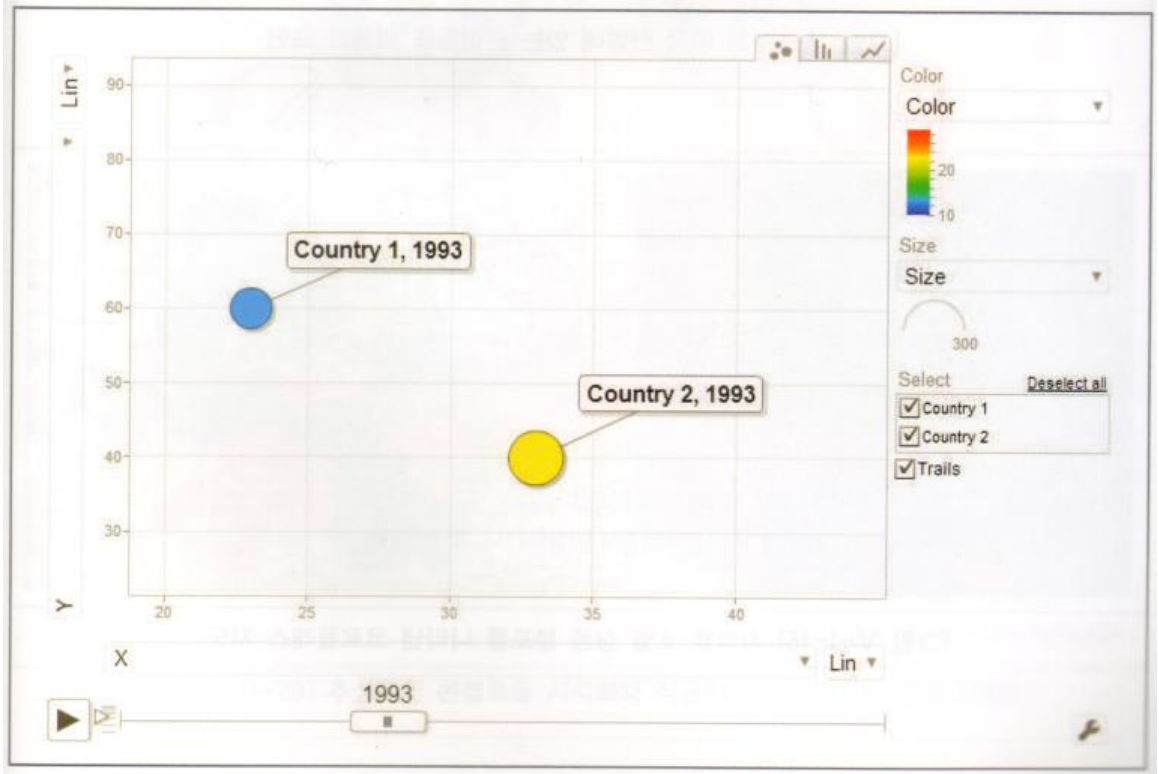
예) 나이와 성별에 따른 체중 (3차원그래프)

차원: 나이,성별
측정값: 체중

시간데이터의 관계탐색 (시간에 따른 패턴의 변화)

방법1: x축에 시간을 설정해 평면이나 공간상에 데이터를 뿌려 전개 모양을 보는 것

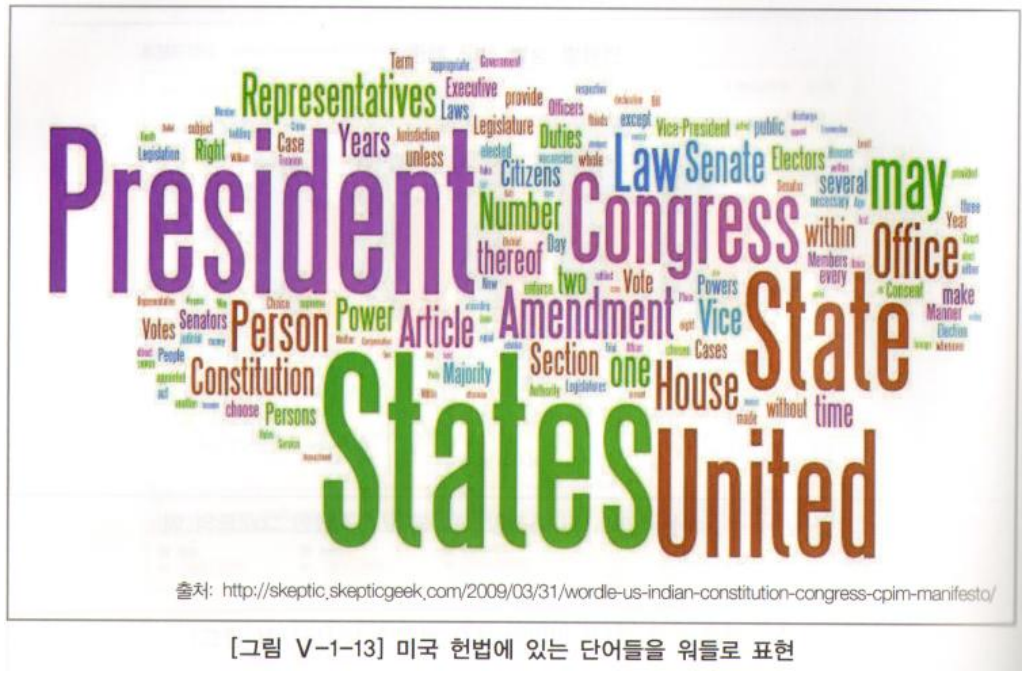
방법2: 모션차트 (motion chart)를 이용하여 시간흐름에 따라 다른 차원의 측정값이 어떻게 변화하는지 움직임으로 보여주는 동적 시각화 도구



[그림 V-1-12] 구글 스프레드시트에서 제공하는 모션 차트 사례

(V-1-12설명)연도, x, y, 나라 4가지 차원으로 구성 (측정값은 하나)
: 연속값인 x,y를 2차원 좌표계의 평면으로 구성,
나라는 한가지 구분값(country1 vs. 2)만 가지므로 색상으로 구분
시간차원을 별도의 축 (아래 슬라이드 바로 설정)
연속값인 측정값을 원의 면적으로 처리함.

비정형데이터의 관계탐색 (예: 텍스트)



[그림 V-1-13] 미국 헌법에 있는 단어들을 워들로 표현

워들(wordle): 문장들 안에 단어들의 빈도분포

텍스트 데이터에서 의미를 갖는 형태소 단위를 추출 → 빈도를 계산 → 빈도에 따라 색상과 크기를 결정 →서로 겹치지 않게 배치

잘라보고 달리보기

예) (연령·성별) 체중데이터

잘라보기 (slice)
20세~40세 남자 vs. 전체패턴

달리보기 (Dice)
성별을 고정(예 남자)한 상태로 연령과 체중 관계

Count of 숙소코드	열 레이블								
행 레이블	특1급	특2급	1급	2급	3급	게스트하우스	레지던스	부띠크	총합계
서울특별시	12	27	19	2	2	1	28	5	96
강원도	2	3	3						8
제주도	11	2	1	1					15
부산광역시			3	1					4
전라북도	1		1						2
경기도	1	2	3	1			1		8
인천광역시	1	3		1					5
경상북도			1						1
전라남도	1		3						4
충청북도			2						2
울산광역시	1								1
대구광역시	1								1
총합계	31	37	36	6	2	1	29	5	147

상품 패턴 분석용 피벗 테이블
(현재 설정 : 지역과 등급에 따른 일부 호텔 상품 분포)

[그림 V-1-14] 엑셀의 기본 피벗 테이블 기능

잘라보고 달리보기 손쉽게 할 수 있는 분석도구:

- 엑셀의 피벗테이블
- 구글 스프레드시트의 피벗 테이블 리포트 (Pivot Table Report)
- OLAP (Online Analytical Process)

c.f) 내려다보고 올려다보기 (계층형 구조)

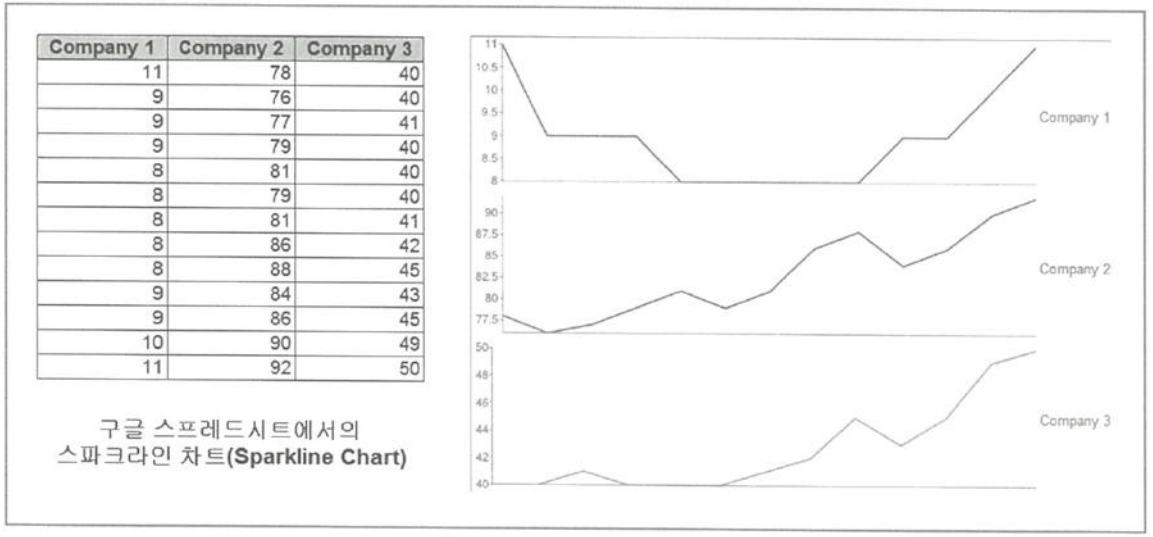
예) 시간데이터 일.주.월.분기.연 단위

척도(scale)의 조정

- 데이터의 최소.최대값 범위보다 그래프 축의 범위가 훨씬 넓으면 패턴이 제대로 나타나지 않음.
- 다른 계열의 데이터를 함께 표현시, 두 계열의 측정값 범위가 너무 다른 경우, 동일 평면에서 둘 중 하나의 패턴을 제대로 평가할 수 없음.

→ 계열별로 최대값을 100으로 설정한 다음, 동일 계열 내의 다른 값들을 이 비율에 맞춰 변환하여 동일한 공간에서 각각의 패턴변화를 비교

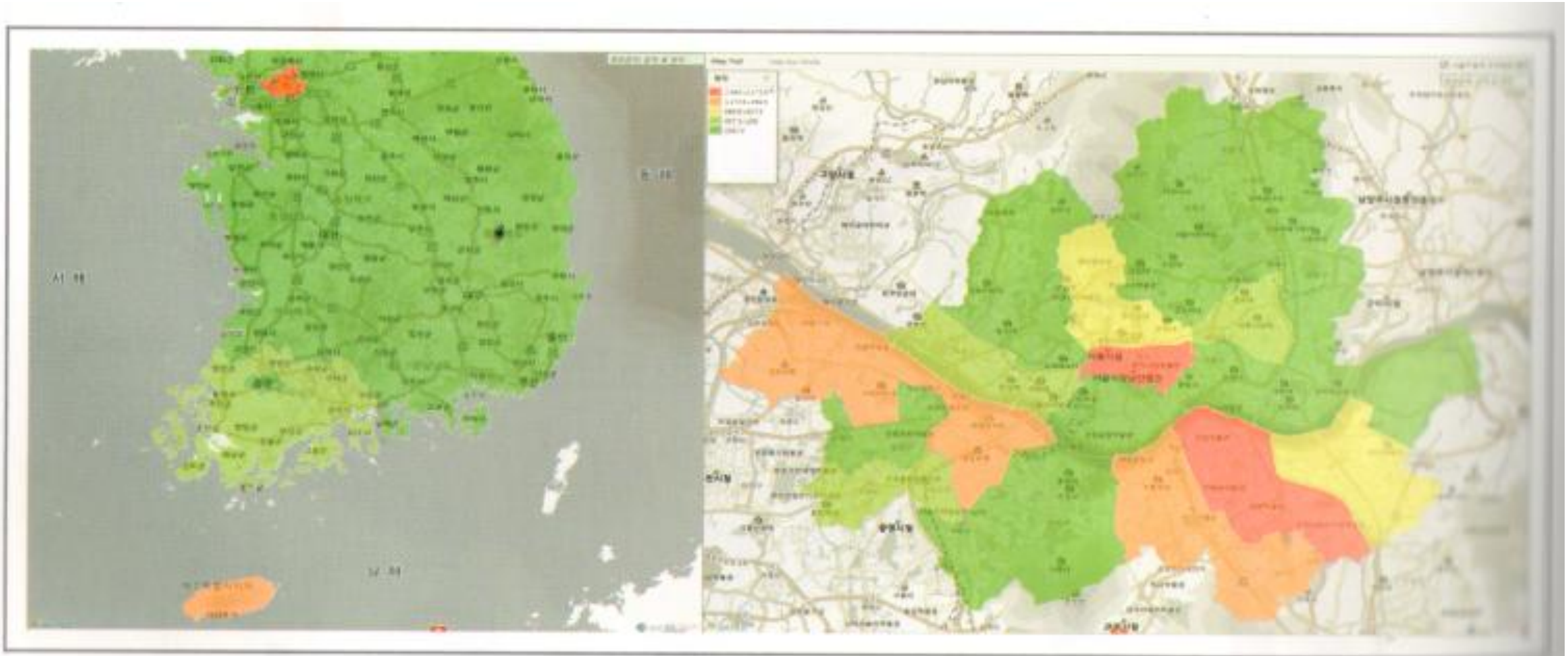
Ex) 스파크라인 차트 (Sparkline chart)



Company 1	Company 2	Company 3
11	78	40
9	76	40
9	77	41
9	79	40
8	81	40
8	79	40
8	81	41
8	86	42
8	88	45
9	84	43
9	86	45
10	90	49
11	92	50

[그림 V-1-15] 계열별로 다른 범위의 측정값을 스파크라인 차트로 표현해 패턴 비교

사례적용



[그림 V-1-16] 지역별 호텔 상품 분포를 나타낸 코로플레스 지도

1. 분석대상의 구체화

가. 2차 탐색

나. 분석 목표에 따른 분석기법

2. 분석과 시각화 도구

3. 지표 설정과 분석

가.지표의 기본 구조

나.지표 활용시 주의점

분석목표에 따른 분석기법

[표 V-1-8] 분석 목표에 따른 적합한 통계적 분석 기법

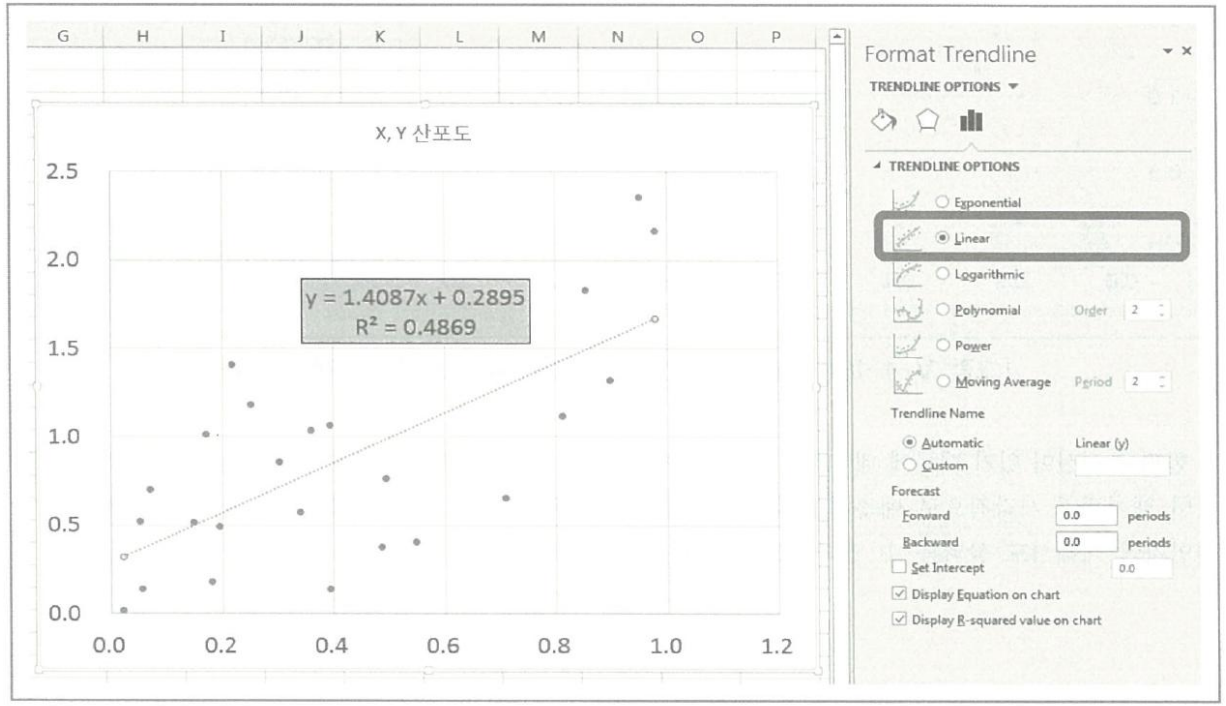
분석 목표	설명	통계적 분석 기법
평균에 대한 검정과 추정[61]	평균값에 대한 모델링	T검정
비율에 대한 검정과 추정	비율에 대한 모델링	직접확률계산법, F분포법
분할표의 검정	각각 2개 이상의 분류값을 지닌 2개 이상의 차원이 있고 그 결과로 하나의 측정값이 있을 때, 분류 조합에 따라 측정값에 유효한 차이가 발생하는지를 검정	카이제곱 검정, Fisher의 직접확률 검정, 맥네마의 검정, 잔차 분석
변수들 간의 상관관계의 강도 추출	독립적으로 움직이는 두 변수들 사이의 관계(상관관계)의 강도를 상관계수로 나타내어 표시함	상관분석
변수들 간의 선형/비선형 인과관계의 형태와 강도 추출	종속적으로 움직이는 두 개 이상의 변수들 사이의 관계(인과관계)의 강도를 결정 계수로 나타내고, 각 변수의 계수를 추정해 모델화함. 변수들은 연속적인 값일 수도 있고 분류값일 수도 있음	회귀분석, 다중회귀분석, 로지스틱회귀분석, 판별분석
어떤 결과에 영향을 미치는 요인들 사이의 관계와 핵심 요인의 선별	어떤 측정값에 변화 요인이 되는 값들이 세 개의 차원이라고 할 때, 각 차원들 중에 어떤 것이 측정값에 가장 큰 영향을 미치는지, 각 차원은 다른 차원의 영향력과 어느 정도 겹치는지 분석	요인분석, 주성분 분석
대상들을 여러 기준값들에 따라 분류하고, 다차원 공간에 배치	측정값과 차원들이 있을 때 차원들의 값을 기준으로 측정값들 사이의 거리를 계산해 적절하게 그룹을 짓고, 이 거리가 의미 있는 차원들로 축을 구성한 다차원 공간에 측정값들을 배치	군집 분석, 다차원척도법(MDS)
차원값들의 패턴이 비슷한 측정값과 그렇지 않은 측정값을 분류	예를 들어, 설문 항목에 대한 답변들의 패턴에 따라 비슷한 답변을 한 응답자와 그렇지 않은 응답자를 분류	대응분석
시간의 흐름에 따라 변하는 데이터를 분석할 수 있는 모델의 도출	시계열 데이터에 영향을 주는 요인을 추세요인, 계절요인, 순환요인, 불규칙요인으로 분해해서 시계열 데이터를 가장 잘 설명할 수 있는 모델을 만들고, 이 모델을 통해 미래에 대해서도 예측	시계열분석

1. 분석대상의 구체화
 가. 2차 탐색
 나. 분석 목표에 따른 분석기법

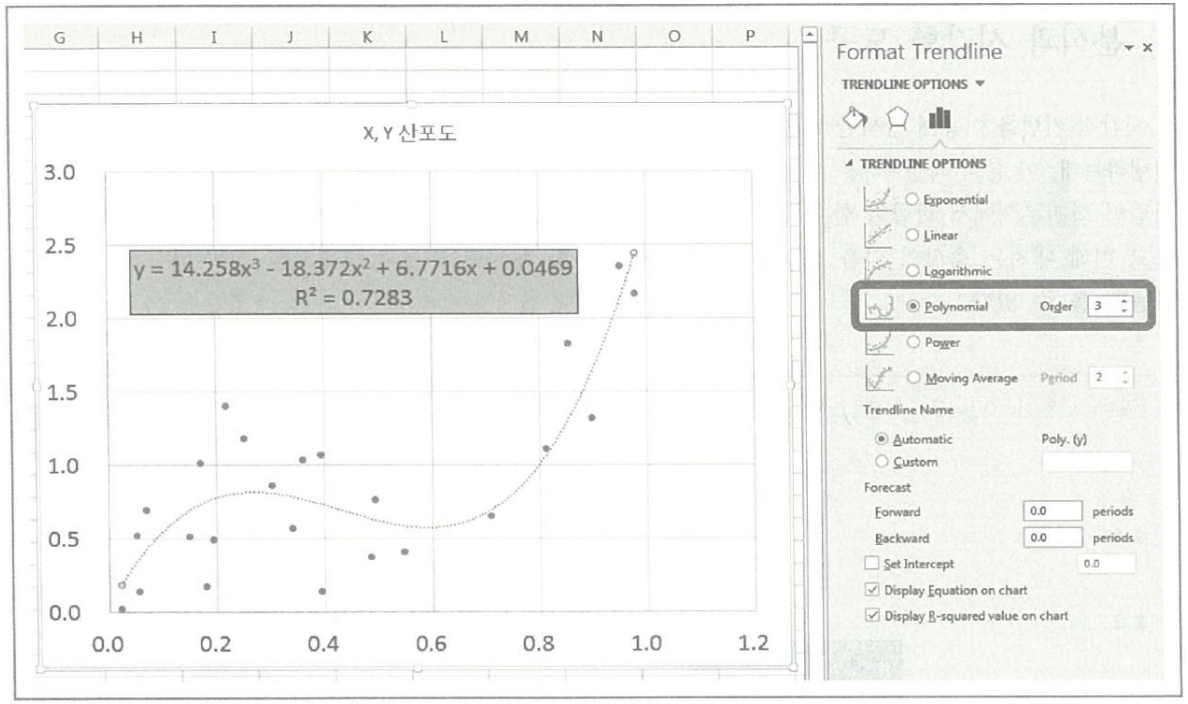
2. 분석과 시각화 도구

3. 지표 설정과 분석
 가. 지표의 기본 구조
 나. 지표 활용 시 주의점

추세선그래프 (수치에 보조적인 역할)



[그림 V-1-17] 엑셀에서 선형 회귀분석 결과와 추세선



[그림 V-1-18] 선형함수 대신 3차원 함수를 적용한 회귀분석 결과

1. 분석대상의 구체화
가. 2차 탐색
나. 분석 목표에 따른 분석기법

2. 분석과 시각화 도구

3. 지표 설정과 분석 (예)강수확률지표)
가. 지표의 기본 구조 (예 X=A*B*C)
나. 지표 활용 시 주의점
1) 지표의 단위에 유의
2) 시각화 표현시 다른 데이터와 함께 적절히 표현될수 있는지 (척도의 문제)
3) 통계모형에 지표를 만들어낸 다른 변수와 (A, B, C)와 함께 들어갈 때 모델의 설명력이 과대 평가 (요인분석으로 설명력이 겹치는지를 확인)

3단계: 활용

1. 내부에서의 적용
2. 외부에서의 적용
3. 인사이트의 발전과 확장

내부에서 적용

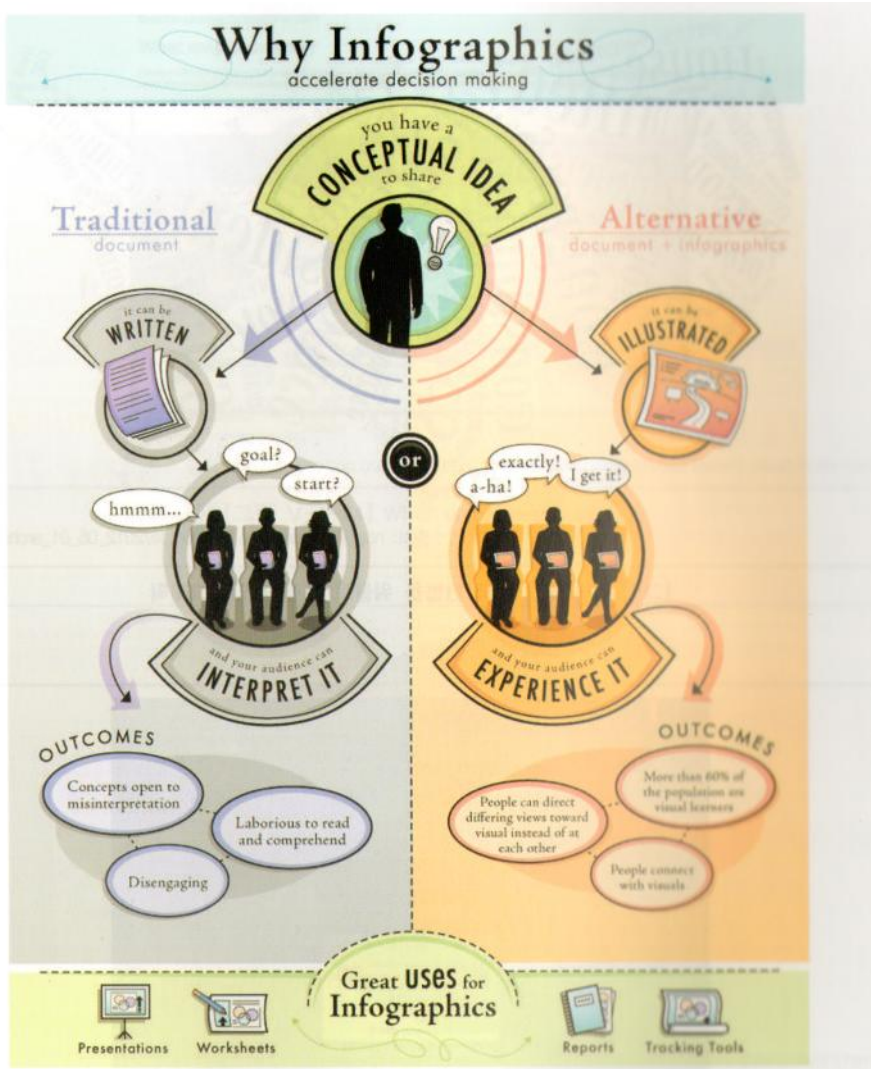
- 얻어진 통찰을 실행으로 옮기는 것

  -기존 문제 해결 방식이나 설명 모델의 수정
  -새로운 문제 해결 방식의 도입
  -새롭게 발견한 가능성에 대한 구체적인 탐색과 발전

외부에서 적용

- 발견한 통찰을 관련된 사람들에게 설명하거나 설득하는 과정

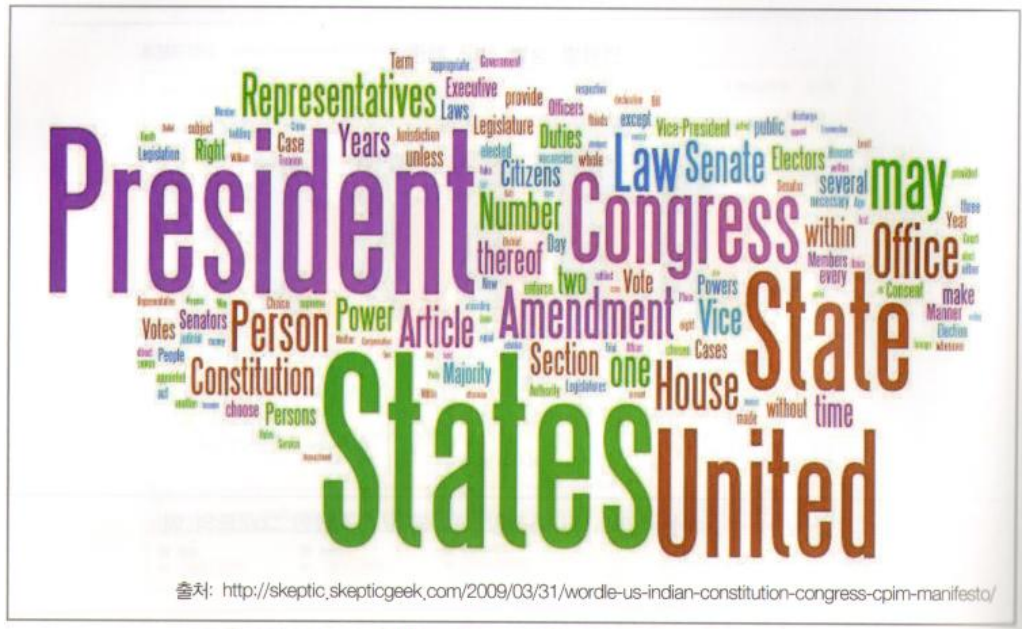
- 시각화한 그림이나 그래프 활용



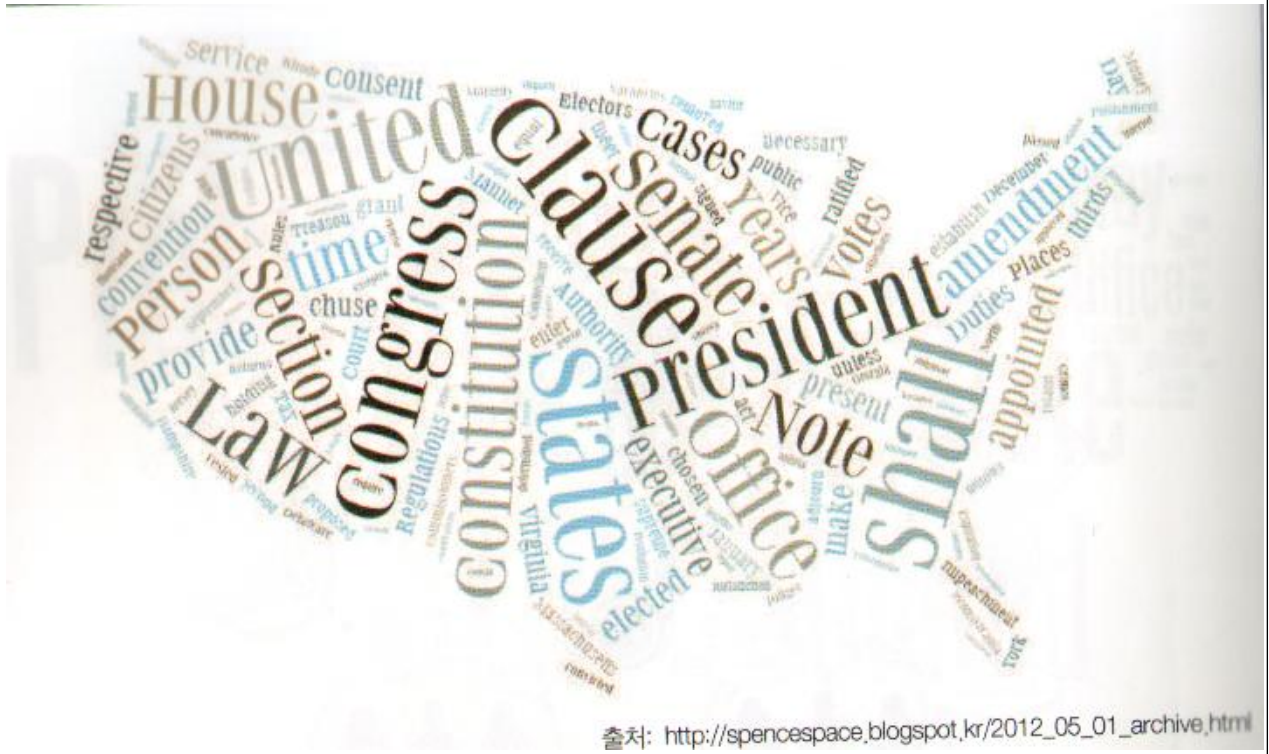
출처: http://visual.ly/why-infographics

[그림 V-1-20] 추상적인 콘셉트를 시각화해 직관적으로 보여주는 인포그래픽

3단계:활용



[그림 V-1-13] 미국 헌법에 있는 단어들을 워들로 표현



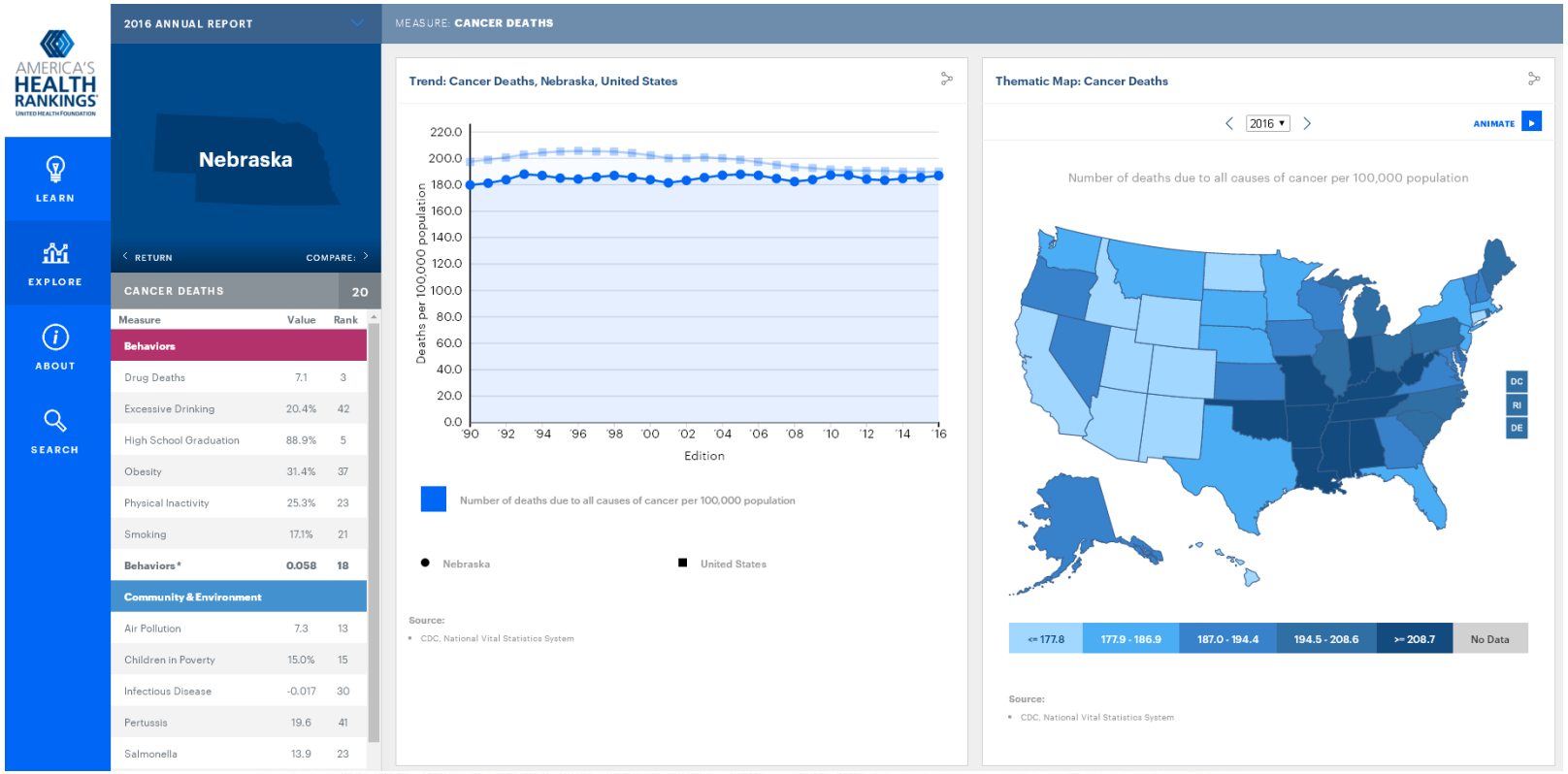
[V-1-21] 워들을 미국지도 디자인 틀에 맞춰 넣어 “미국”이라는 메시지를 강화

설득은 정보의 전달뿐 아니라 자신의 의도를 상대방이 공감하고 행동의 변화를 이끌어야 하기 때문에, 상대방의 이성과 감성에 적절하게 호소 해야

→사람의 마음을 움직이는 디자인(design) 시각화 도구 이용

출처: http://www.onehourpersecond.com/

[그림 V-1-22] 유튜브의 '1초에 1시간' 캠페인 사이트



[그림 V-1-23] 미국의 건강 지표 사이트

3. 인사이트의 발전과 확장
인사이트는 꼬리에 꼬리를 무는 것과 같다 (‘connect the dots’)
이미 존재하는 다양한 자료와 정보를 접하는 과정이 중요

- 탑다운 vs. 보텀업
- 2차 잘라보기 • 달리 보기, 내려다보기 • 올려다보기
- 실시간 vs. 비실시간
- 지표의 운영
- 추가데이터의 필요성
- 시각화 오류 (Optical illusion)
- 사람의 문제

[그림 V-1-25] 도로 정체현황 표시판

- ✓ 정말 실시간으로 처리해야 하는 이슈인지 (주기적으로 살펴보아도 되는지)...비용분석이 필요
- ✓ 실시간으로 처리해야 하는 것은 긴급한 위기상황 처리를 위한 모니터링/경보시스템의 경우이고, 시각화가 굉장히 유용.

Summary

시각화 인사이트 프로세스

: 시각화를 통한 인사이트 (데이터,정보, 지식, 사람 사이의 관계파악을 통해 지혜도출) 도출

1단계 탐색

: 데이터 명세화, 이해.

데이터내 연결고리, 탐색범위설정 (잘라/달리보기, 내려다/올려다보기)

2단계 분석

: 수치적 모델, 특정 값으로 표현 (인터렉티브 그래프)

3단계 활용

: 설명 및 설득 (인포그래픽)

문제 1. 시각적 이해의 위계 구도 상에서 인간의 경험이 본격적으로 개입되는 단계는 무엇인가?

① 데이터
② 정보
③ 지식
④ 지혜

문제 2. 시각화 인사이트 프로세스에 대한 다음 설명 중 틀린 것 두 개는 무엇인가?

① 통찰을 얻기 위해 살펴봐야 할 대상은 외부와 내부(사람)의 두 가지다.
② 지혜는 개인화된 지식이다.
③ 분석 단계에서는 그래프를 이용하지 않고 수치분석 기법을 사용한다.
④ 통찰의 활용에는 프레젠테이션도 포함된다.

문제 3. 다음 데이터 예시 중 간편한 조작으로 계층형 구조를 만들 수 없는 것은 무엇인가?

① 경위도 데이터
② YYYY-MM-DD
③ 행정구역 데이터
④ 일반 텍스트 데이터

문제 4. 데이터를 명세화하기 위한 개념과 관계가 없는 것은 무엇인가?

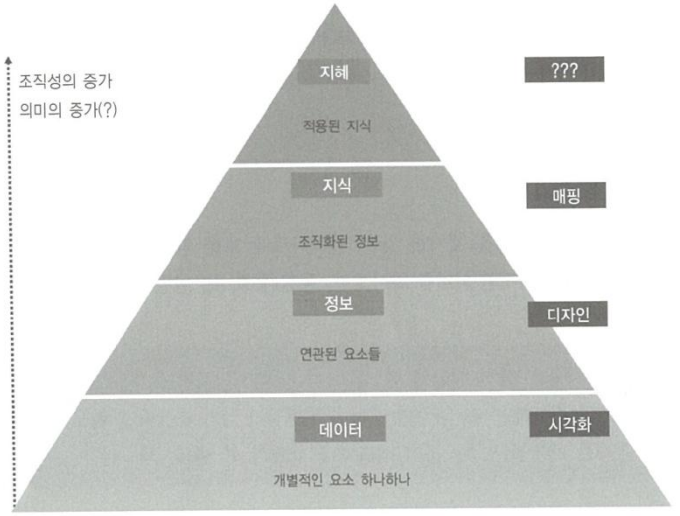
① 데이터형
② 로그 데이터
③ 클래스
④ 메소드

문제 5. 공간 데이터의 처리와 직접적인 관계가 없는 것 두 개는 무엇인가?

① 코로플레스 지도
② 지오코딩
③ 워들
④ vlookup 함수

문제 1. 시각적 이해의 위계 구도 상에서 인간의 경험이 본격적으로 개입되는 단계는 무엇인가?

① 데이터
② 정보
③ 지식
④ 지혜



지식은 단편적인 정보들을 조직화 할때 발생하며 이때부터 경험이 본격적 개입됨.

문제 2. 시각화 인사이트 프로세스에 대한 다음 설명 중 틀린 것 두 개는 무엇인가?

① 통찰을 얻기 위해 살펴봐야 할 대상은 외부와 내부(사람)의 두 가지다.
② 지혜는 개인화된 지식이다.
③ 분석 단계에서는 그래프를 이용하지 않고 수치분석 기법을 사용한다.
④ 통찰의 활용에는 프레젠테이션도 포함된다.

통찰을 얻기 위해 내구와 외부 및 그 사이를 연결하는 고리, 즉 세가지 관찰(삼찰)을 이용

문제 3. 다음 데이터 예시 중 간편한 조작으로 계층형 구조를 만들 수 없는 것은 무엇인가?

① 경위도 데이터
② YYYY-MM-DD
③ 행정구역 데이터
④ 일반 텍스트 데이터

문제 4. 데이터를 명세화하기 위한 개념과 관계가 없는 것은 무엇인가?

① 데이터형
② 로그 데이터
③ 클래스
④ 메소드

로그데이터는 명세화의 기본 대상이지 데이터형, 클레스, 메소드와 같이 명세화 할때 반드시 고려해야 하는 개념은 아님

문제 5. 공간 데이터의 처리와 직접적인 관계가 없는 것 두 개는 무엇인가?

① 코로플레스 지도
② 지오코딩
③ 워들
④ vlookup 함수

워들은 텍스트 데이터를 처리하는 시각화 기법.
Vlookup함수는 범용성 도구 (연결고리탐색-분류형 데이터 변환에 유리)

문제 6. 시간에 따른 다차원 데이터의 변화를 직관적으로 살펴볼 수 있는 기법은 무엇인가?

① 트리맵
② 모션차트
③ 산포도
④ 피벗테이블

문제 7. 결과에 영향을 미치는 요인들 사이의 관계와 핵심 요인을 선별하는 통계적 분석기법 두 개는 무엇인가?

① 요인분석
② 상관분석
③ 판별분석
④ 주성분 분석

문제 8. 분석 및 지표에 대한 다음 설명 중 맞는 것 두 개는 무엇인가?

① 빅데이터 분석에서는 표본 데이터에 기반한 분석 모델링이 중요하다.
② 결정 계수는 모델의 설명력을 의미한다.
③ 원본 데이터에서 여러 지표를 잘 추출하여 모델에 많이 반영할수록 설명력이 유의미하게 커진다.
④ 지표간의 설명력에 대한 효과는 요인분석을 통해서 확인해볼 수 있다.

문제 9. 통찰을 활용하는 다음 방법들 중 다른 세 개와 성격이 다른 하나는 무엇인가?

① 조기 경보 체계의 구축
② 설명력을 강화하는 변인의 추가
③ 설명과 설득을 위한 스토리텔링 콘텐츠 제작
④ 기존 모델에 대한 전면적인 재검토

문제 10. 인사이트의 발전과 확장에 대한 다음 설명들 중 틀린 것은 무엇인가?

① 범위와 관점이 잘 정해지지 않았을 때에는 보텀업 방식의 접근이 탑다운보다 낫다.
② 인사이트는 결국은 사람의 문제로 귀결된다.
③ 지표의 단점은 지표만 보다보면 지표의 변화에 영향을 미치는 요인을 쉽게 찾을 수 없다는 것이다.
④ 빅데이터 환경에서는 실시간으로 복잡한 분석을 하는 것이 바람직하다.

문제 6. 시간에 따른 다차원 데이터의 변화를 직관적으로 살펴볼 수 있는 기법은 무엇인가?

① 트리맵
② 모션차트
③ 산포도
④ 피벗테이블

문제 7. 결과에 영향을 미치는 요인들 사이의 관계와 핵심 요인을 선별하는 통계적 분석기법 두 개는 무엇인가?

① 요인분석
② 상관분석
③ 판별분석
④ 주성분 분석

상관분석은 독립적인 두 변수간의 상관관계 추출
판별분석은 종속적으로 움직이는 두 변수들간의 선형/비선형 인과 관계의 형태와 강도 추출

문제 8. 분석 및 지표에 대한 다음 설명 중 맞는 것 두 개는 무엇인가?

① 빅데이터 분석에서는 표본 데이터에 기반한 분석 모델링이 중요하다.
② 결정 계수는 모델의 설명력을 의미한다.
③ 원본 데이터에서 여러 지표를 잘 추출하여 모델에 많이 반영할수록 설명력이 유의미하게 커진다.
④ 지표간의 설명력에 대한 효과는 요인분석을 통해서 확인해볼 수 있다.

① 표본데이터->모수데이터
③ 지표는 여러 기본 변인의 집합으로 구성되기 때문에 지표는 많이 넣으면 모델의 설명력이 수치상 증가하지만 중복되어 과대평가된 설명력이지 유의미한 향상은 아니다.

문제 9. 통찰을 활용하는 다음 방법들 중 다른 세 개와 성격이 다른 하나는 무엇인가?

① 조기 경보 체계의 구축
② 설명력을 강화하는 변인의 추가
③ 설명과 설득을 위한 스토리텔링 콘텐츠 제작
④ 기존 모델에 대한 전면적인 재검토

문제 10. 인사이트의 발전과 확장에 대한 다음 설명들 중 틀린 것은 무엇인가?

① 범위와 관점이 잘 정해지지 않았을 때에는 보텀업 방식의 접근이 탑다운보다 낫다.
② 인사이트는 결국은 사람의 문제로 귀결된다.
③ 지표의 단점은 지표만 보다보면 지표의 변화에 영향을 미치는 요인을 쉽게 찾을 수 없다는 것이다.
④ 빅데이터 환경에서는 실시간으로 복잡한 분석을 하는 것이 바람직하다.