데이터로 특정 행동을 보인 집단 간 특성 차이 이해하기

#세그멘테이션 #코호트특성

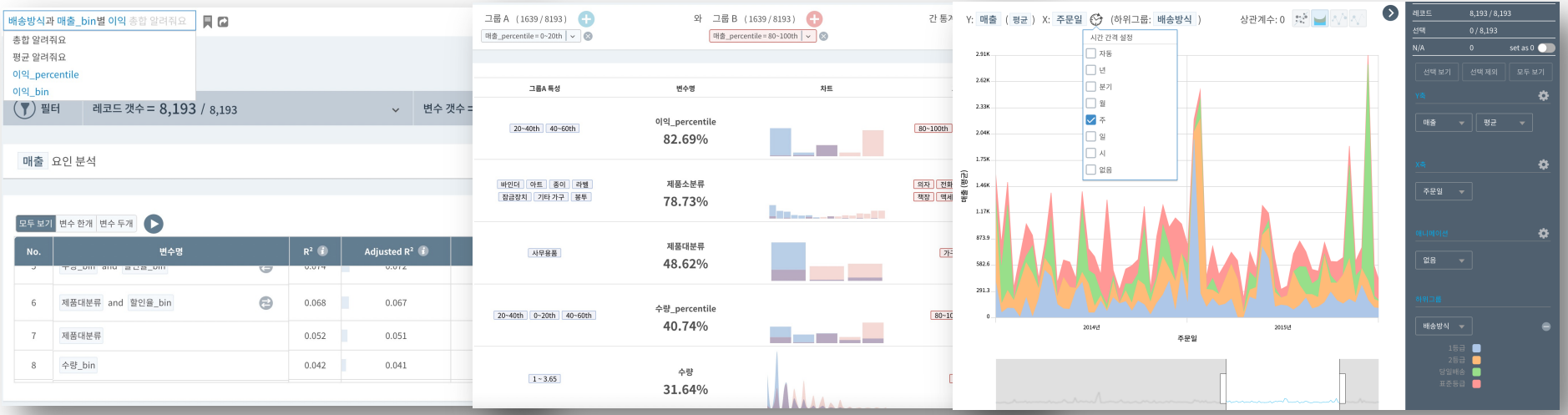
양 승 준 / sidney.yang@idk2.co.kr

다정한 데이터 도구, HEARTCOUNT

- HEARTCOUNT(하트카운트)는 비전문가도 쉽게 엑셀 데이터셋을 업로드하여 시각화하고 분석할 수 있는 SaaS 솔루션입니다.
- Google 계정만 있다면, 홈페이지에서 바로 사용을 시작할 수 있어요!

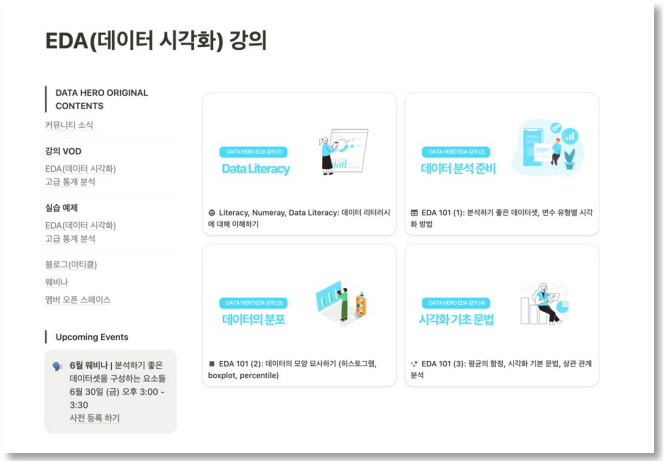
다정한 데이터 도구, HEARTCOUNT

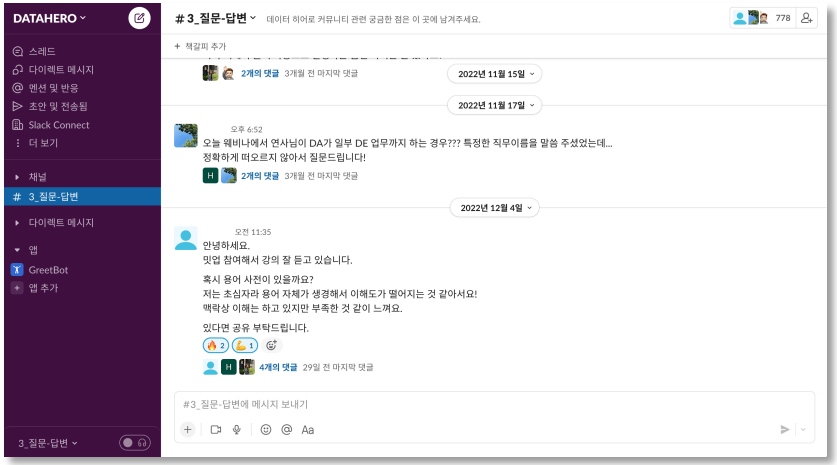
- 특장점으로는 '개별 레코드 수준의 시각화, '파생 변수 자동 생성', '패턴 자동 발견', '자연어 검색 + 설명' 등이 있습니다.



학습하고 소통하는 공간, DATA HERO

- 데이터의 기초부터 실전까지, 전용 페이지에서 무료로 학습 가능
- 하트카운트팀은 물론 다양한 실무자들과의 소통 공간
- 다양한 집중 교육 캠프, 오프라인 밋업 등 이벤트





의료보험 청구비용 최적화를 위한 고객 세그먼테이션

- 데이터셋(sample insurance)은 아래 표와 같이 구성되어 있음
  - https://support.heartcount.io/community/learning/sample-dataset-sites

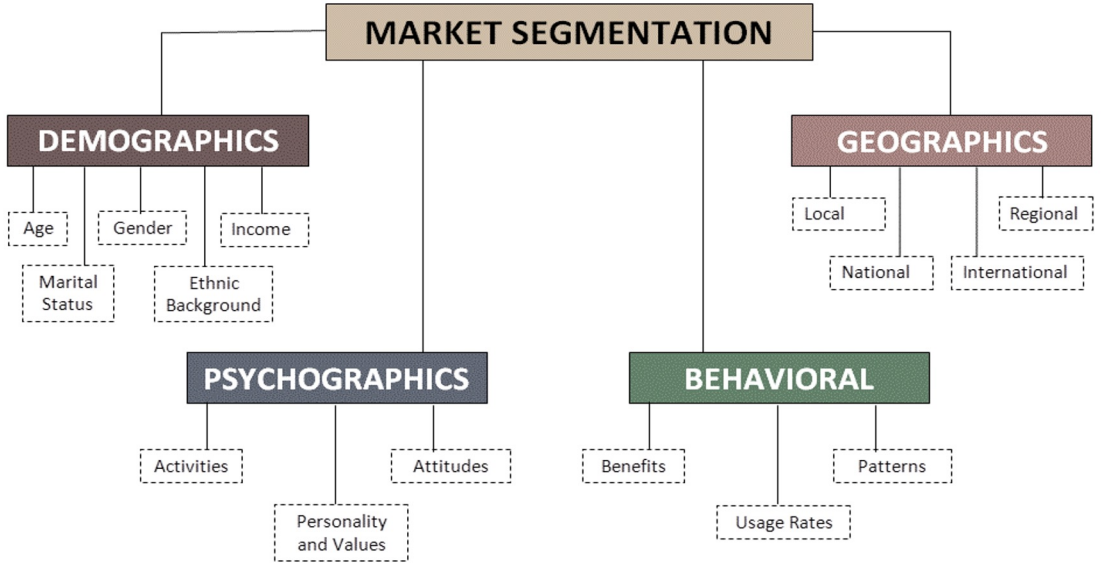
1. 의료 보험 청구액의 차이를 가져오는 주요 요인을 찾고
2. 보험 청구를 한 고객과 그렇지 않은 고객의 특성차이를 이해한 후
3. 보험 청구하지 않을 확률이 높은 고객 세그먼트를 찾고자 함

age	sex	bmi	steps.per.day	children	smoker	region	medical.cost	insurance.claim
19	female	27.9	3009	0	yes	southwest	16884.924	yes
18	male	33.77	3008	1	no	southeast	1725.5523	yes
28	male	33	3009	3	no	southeast	4449.462	no
33	male	22.705	10009	0	no	northwest	21984.47061	no
32	male	28.88	8010	0	no	northwest	3866.8552	yes
31	female	25.74	8005	0	no	southeast	3756.6216	no
46	female	33.44	3002	1	no	southeast	8240.5896	yes
37	female	27.74	8007	3	no	northwest	7281.5056	no
37	male	29.83	8002	2	no	northeast	6406.4107	no

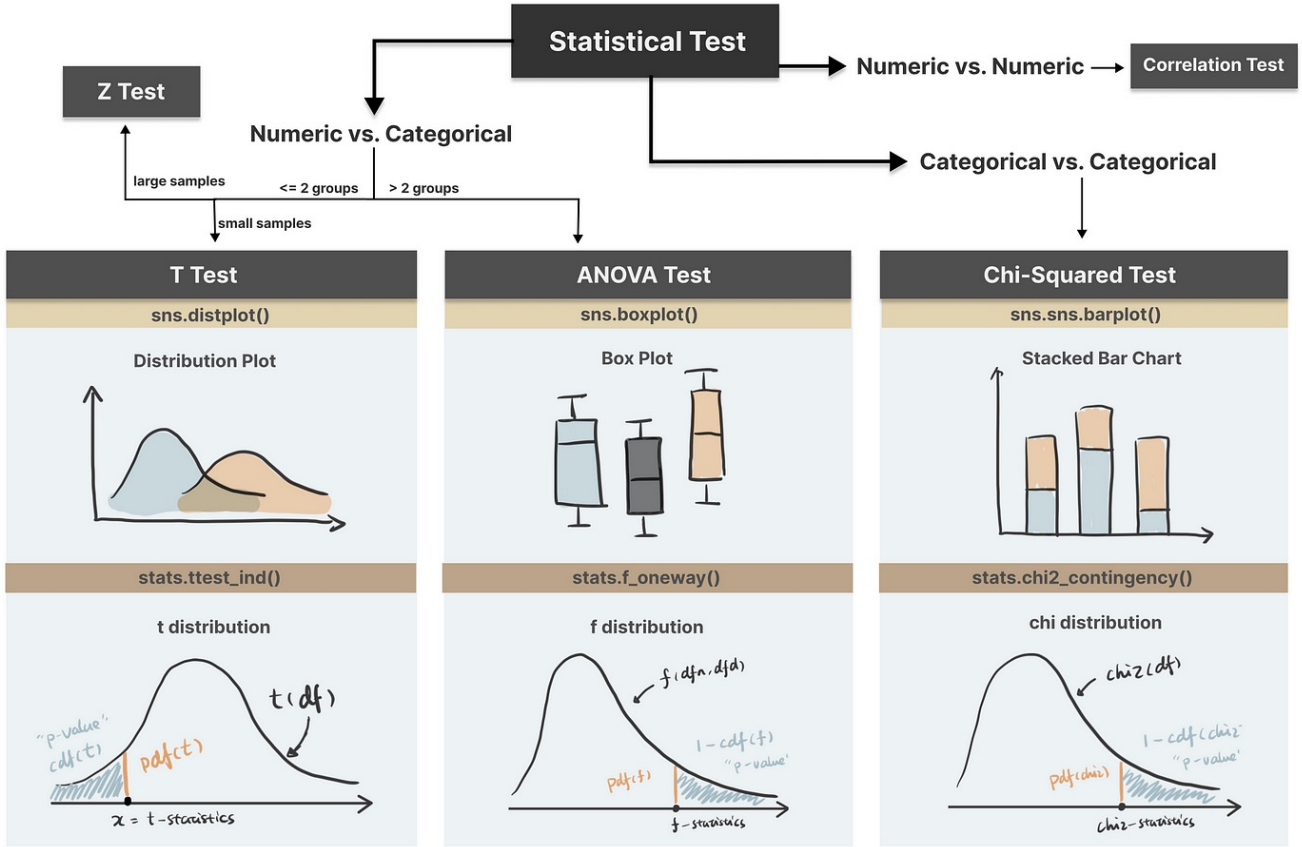
나이 / 성별 / BMI(체질량 지수) / 활동량(Steps per day) / 자녀수 / 흡연여부 / 지역 / 의료비 청구액 / 청구여부

세그먼테이션

- breaking down data by different dimensions (e.g. age, gender, income, location, interest, etc.) **to identify homogeneous sub-groups**
- the motivation behind segmentation is **to identify and act on the key factors driving the majority of business outcomes (sales, conversions)**



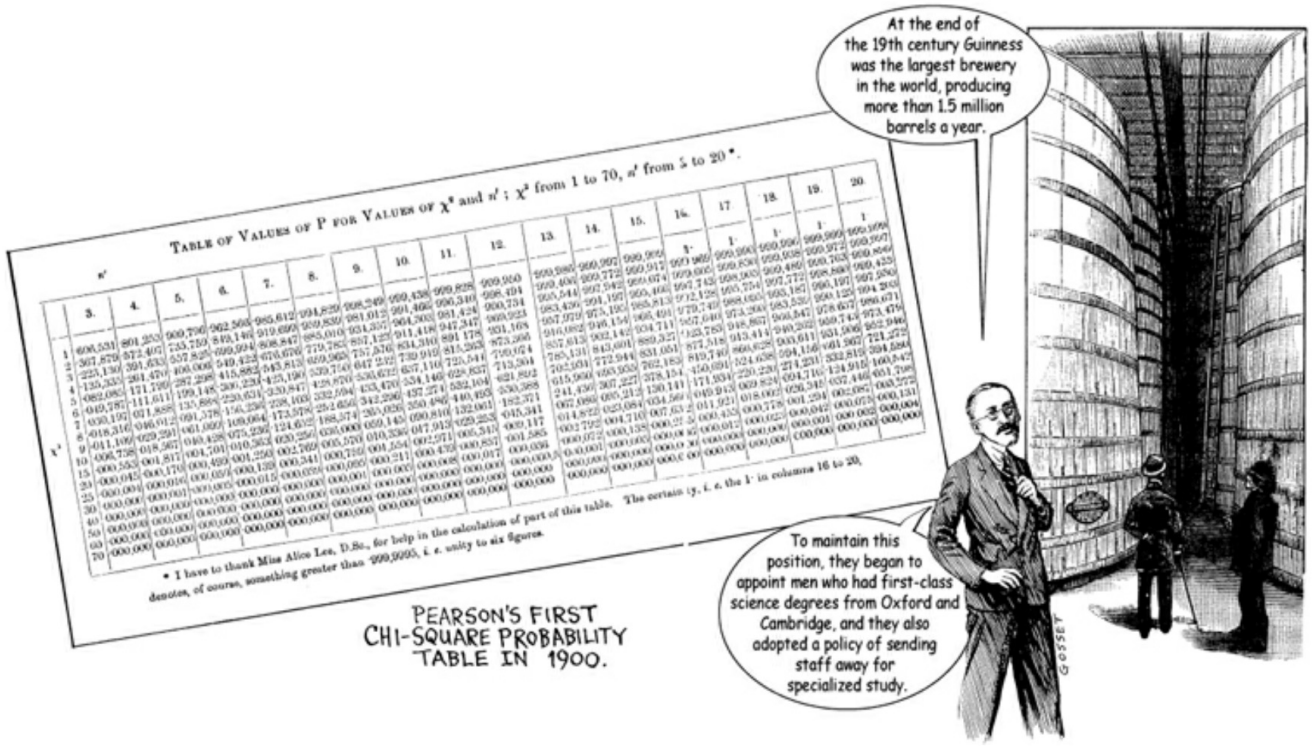
가설 검증 – 집단 간 유의미한 차이 / 관계



visit www.visual-design.net for step by step guide

A Statistical Test for the Guinness Brewery

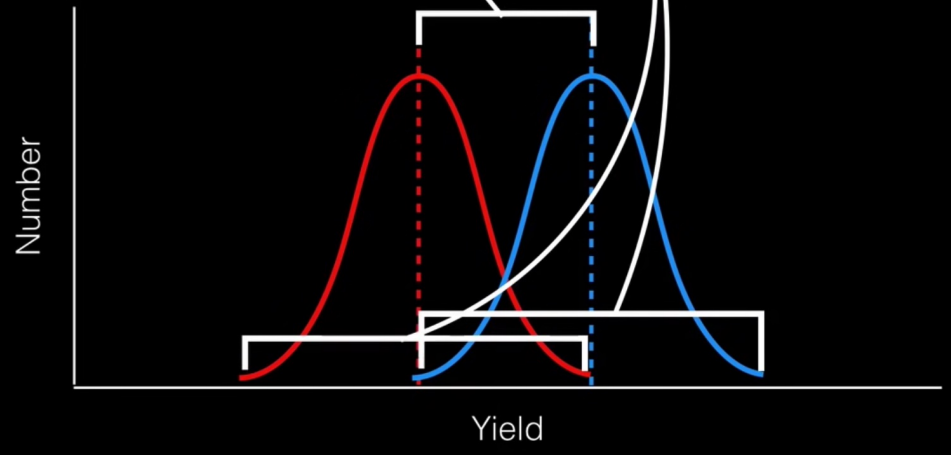
The first statistical quality control test for industry was devised by the statistician and chemist William Sealy Gosset, a master brewer at Guinness in the early years of the 20th century. Since Gosset was bound by his appointment not to publish under his own name (perhaps because Guinness didn't want rival breweries to know that they were training some of their scientific staff in statistical theory), he adopted the pseudonym "Student". This was standard practice at Guinness: Gosset's statistical assistant, Edward M. Somerfield, used "Alumnus" in his publications.




Some of the brewers undertook chemical analyses to try to identify and quantify the qualities that made hops and barley good brewing material ...
... such as the "rub" of the hops or the "texture" of barley that might be "milky" or "steely".

$$\frac{\text{signal}}{\text{noise}} = \frac{\text{difference between group means}}{\text{variability of groups}} = \frac{\left|\overline{x_1} - \overline{x_2}\right|}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

t-value



Field 1	Field 2
15.2	15.9
15.3	15.9
16.0	15.2
15.8	16.6
15.6	15.2
14.9	15.8
15.0	15.8
15.4	16.2
15.6	15.6
15.7	15.6
15.5	15.8
15.2	15.5
15.5	15.5
15.1	15.5
15.3	14.9
15.0	15.9

t-value = 2.3

$df = n_1 + n_2 - 2$

H_0

t-test

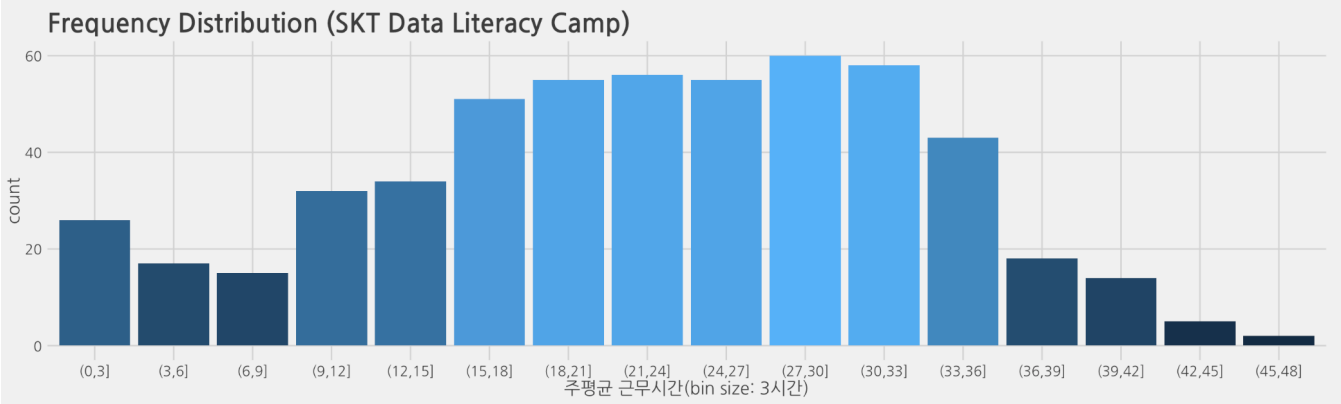
Don't Reject

Reject

Degrees of Freedom	p=0.05	p=0.025	p=0.01
1	12.71	25.45	63.66
2	4.30	6.20	9.92
3	3.18	4.17	5.84
4	2.78	3.50	4.60
5	2.57	3.16	4.03
6	2.45	2.97	3.71
7	2.36	2.84	3.50
8	2.31	2.75	3.36
9	2.26	2.68	3.25
10	2.23	2.63	3.17
11	2.20	2.59	3.11
12	2.18	2.56	3.05
13	2.16	2.53	3.01
14	2.14	2.51	2.98
15	2.13	2.49	2.95
16	2.12	2.47	2.92
17	2.11	2.46	2.90
18	2.10	2.44	2.88
19	2.09	2.43	2.86
20	2.09	2.42	2.84
21	2.08	2.41	2.83
22	2.07	2.41	2.82
23	2.07	2.40	2.81
24	2.06	2.39	2.80
25	2.06	2.38	2.79
26	2.06	2.38	2.78
27	2.05	2.37	2.77
28	2.05	2.37	2.76
29	2.04	2.36	2.76
30	2.04	2.36	2.75
40	2.02	2.33	2.70
60	2.00	2.30	2.66
120	1.98	2.27	2.62

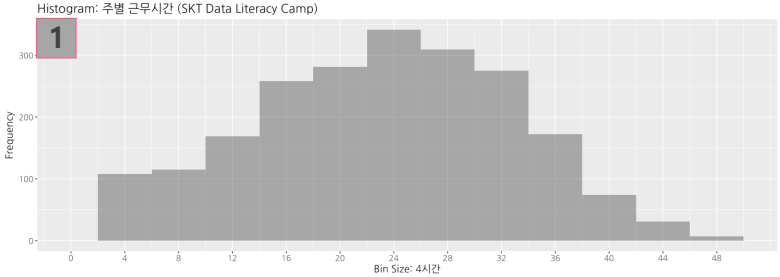
Histogram vs. Frequency Distribution Table

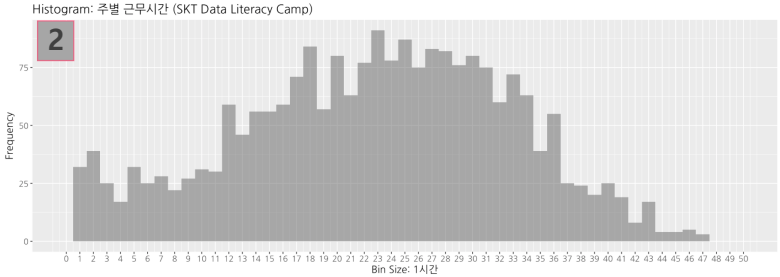
히스토그램과 도수분포표

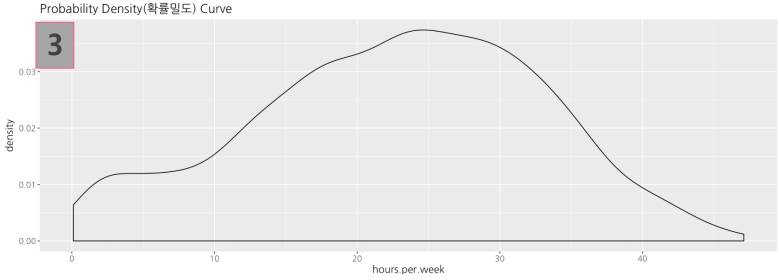


계급	(0,3]	(3,6]	(6,9]	(9,12]	(12,15]	(15,18]	(18,21]	(21,24]	(24,27]	(27,30]	(30,33]	(33,36]	(36,39]	(39,42]	(42,45]	(45,48]
빈도	26.0	17.0	15.0	32.0	34.0	51.0	55.0	56.0	55.0	60.0	58.0	43.0	18.0	14.0	5.0	2.0
누적 빈도	26.0	43.0	58.0	90.0	124.0	175.0	230.0	286.0	341.0	401.0	459.0	502.0	520.0	534.0	539.0	541.0
비율	4.8	3.1	2.8	5.9	6.3	9.4	10.2	10.4	10.2	11.1	10.7	7.9	3.3	2.6	0.9	0.4
누적 비율	4.8	7.9	10.7	16.6	22.9	32.3	42.5	52.9	63.1	74.2	84.9	92.8	96.1	98.7	99.6	100.0

Histogram vs. Density Plot







1 Histogram – Bin Size: 4시간

- 히스토그램: 도수(빈도)의 분포[도수분포표]를 차트로 표현한 것
- 계급: X축에 표현된 변수의 구간[4시간]

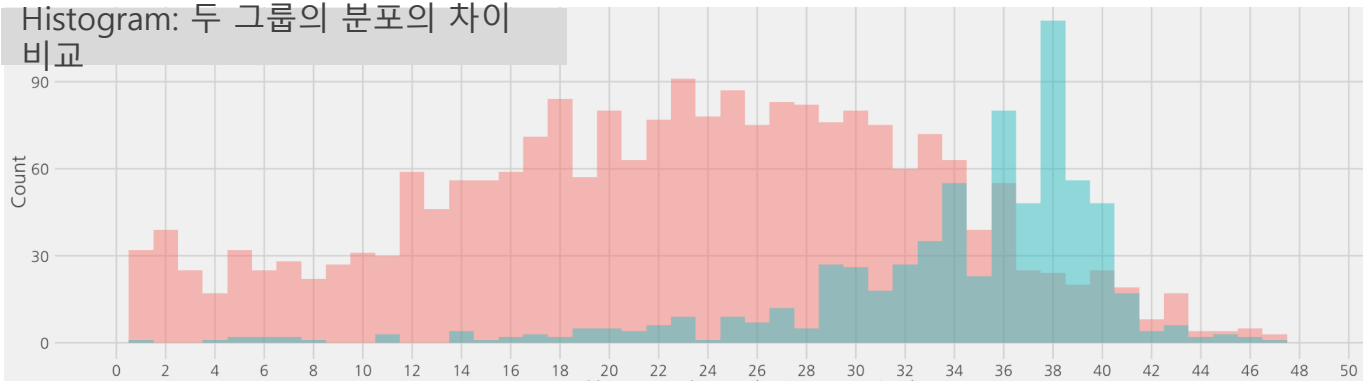
2 Histogram – Bin Size: 1시간

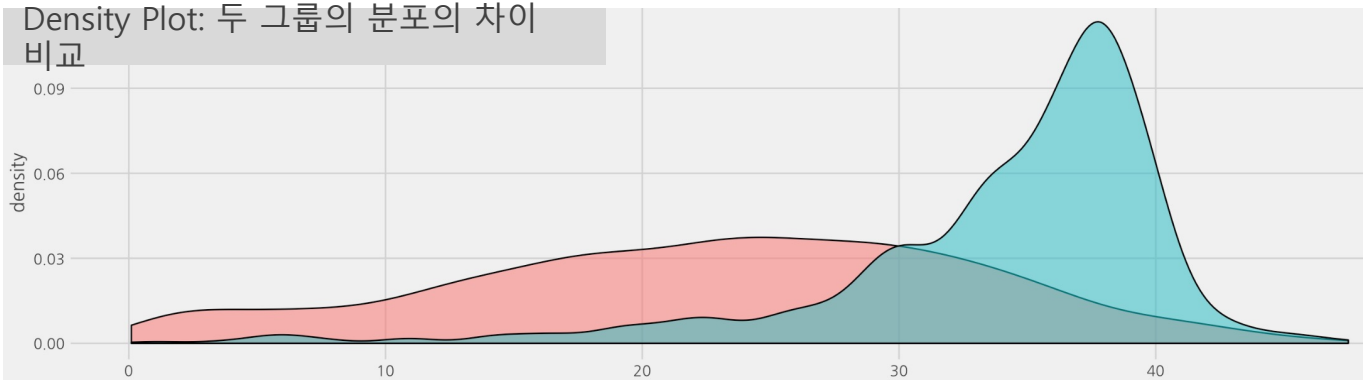
- X축 변수 구간의 크기(Bin Size)를 4시간에서 1시간으로 조정하였음

3 Probability Density Curve

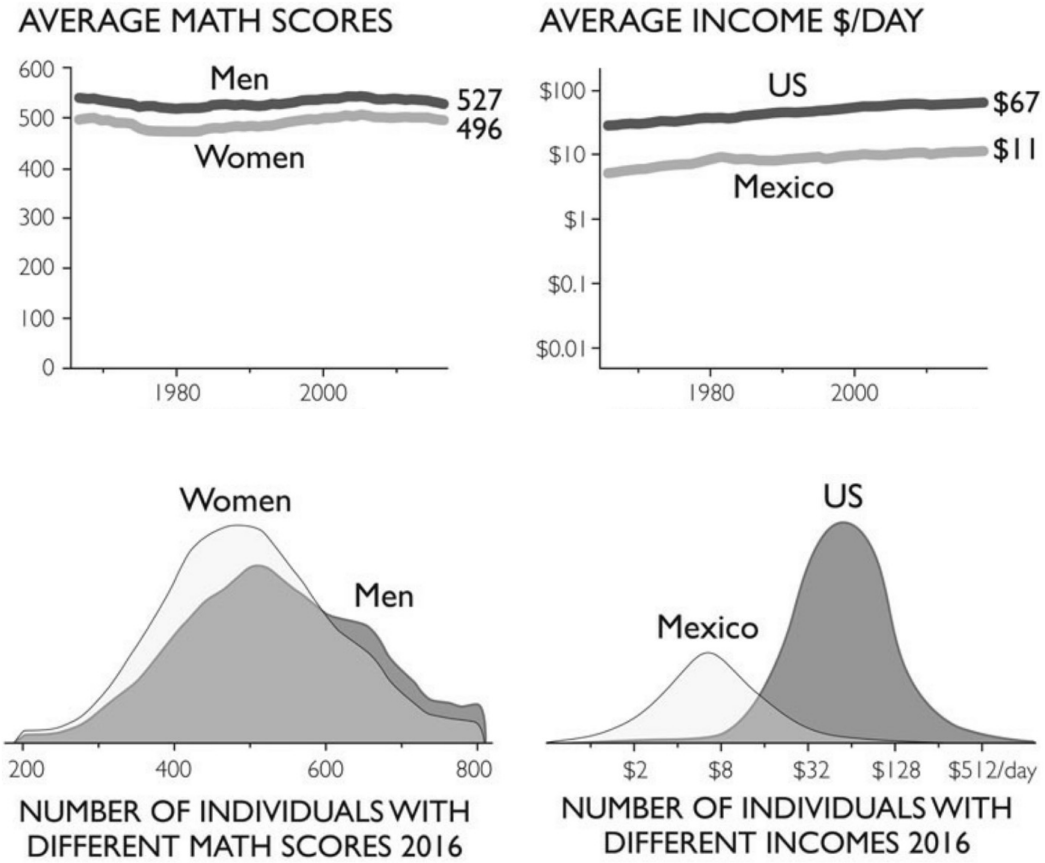
- 확률밀도: X가 연속형 변수일 경우 X값과 이에 대응하는 확률을 나타낸 그래프
- 좌측에서 X가 10~20시간 사이의 값을 가질 확률은 해당 구간의 면적과 동일함

Histogram vs. Density Plot



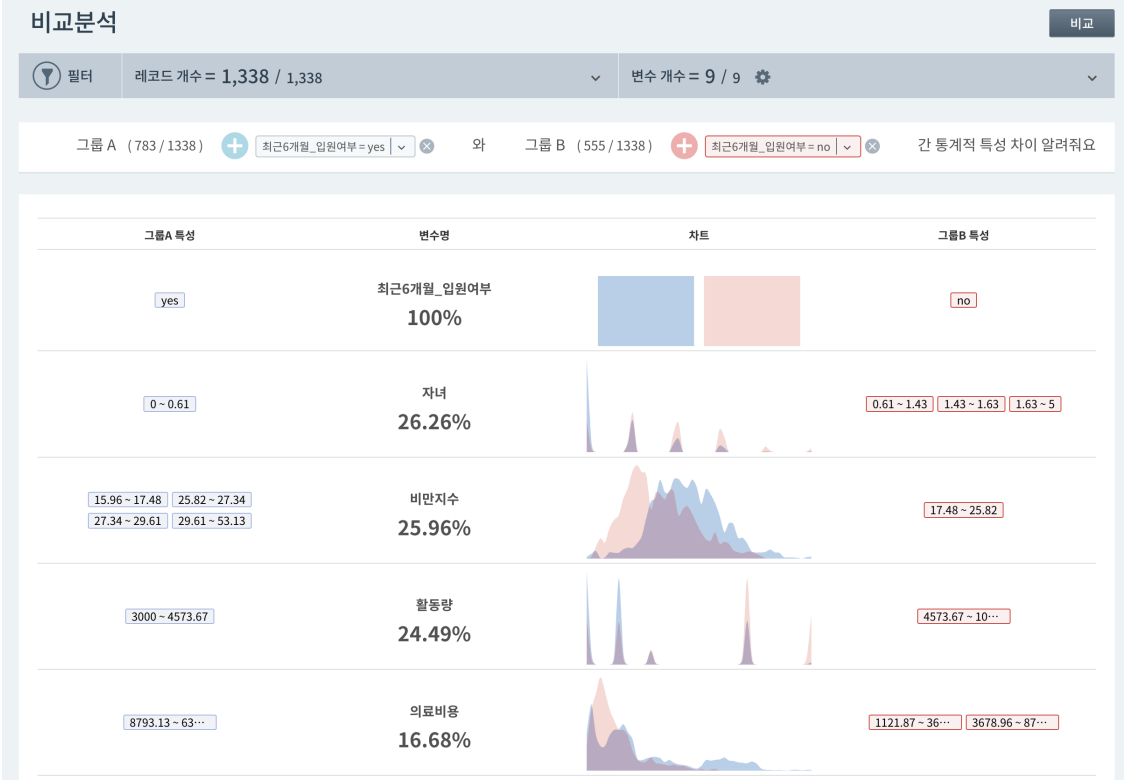


Comparison of Average vs. Distribution



최근 6개월 입원했던 고객 집단의 특성 이해

- 두 집단의 주요 특성 차이를 중요도(특성 차이의 크기) 순으로 보여주는 화면
- 세번째 줄. 입원했던 고객군의 비만지수가 전체적으로 높게 분포하고 있음



Decision Tree and Classification

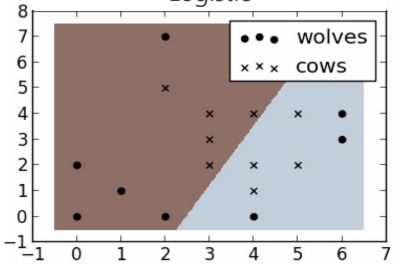
- Decision Tree: 의사결정트리; 대표적 Classification Model
- Classification: 서로 다른 집단을 구분하는 규칙(경계) 찾는 모형

Classification: 집단을 구분하는 경계 찾기

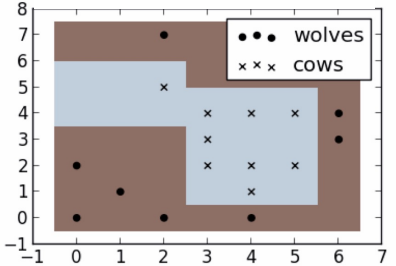
Linear Classifier

Non-Linear Classifier

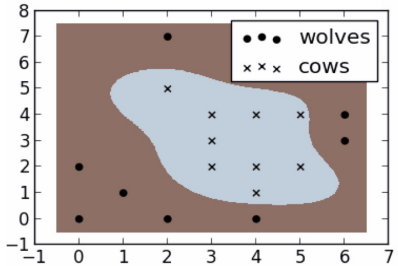
Logistic



Decision Tree



SVM

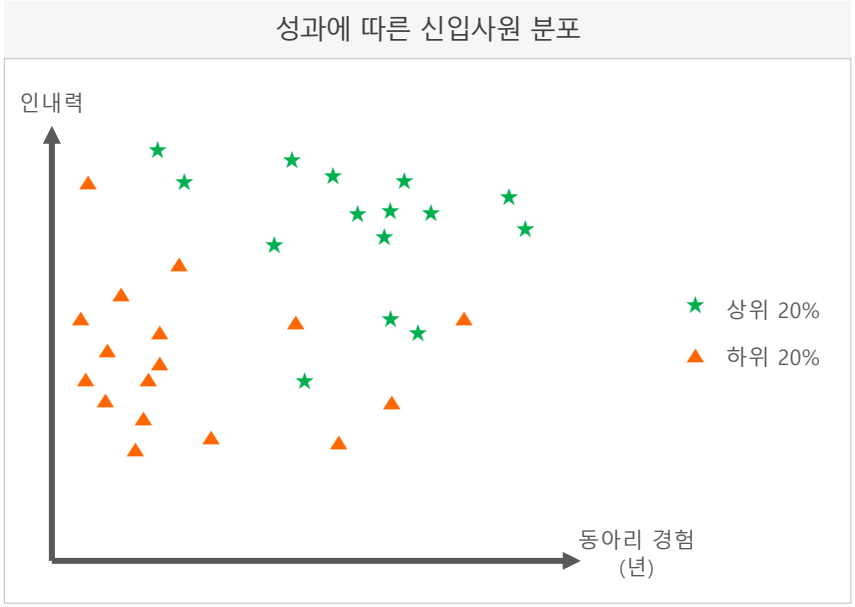


Decision Tree - Minimizing Entropy

Decision Tree Algorithm

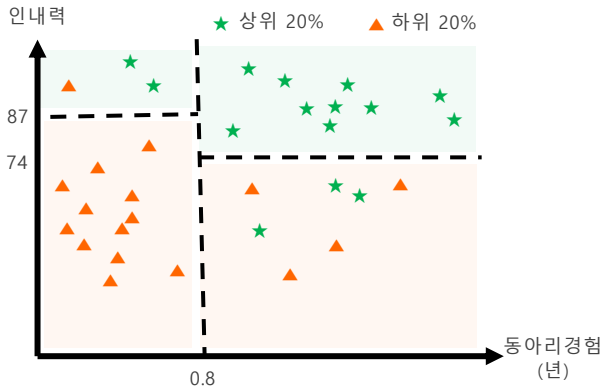
- Purity: *엔트로피를 최소화하도록(= 끼리끼리 모이도록) 공간 구획
- Homogeneity: 동질적 집단이 밀집한 세그먼트의 논리적 규칙 찾기

***엔트로피(Entropy): Measure of Impurity**

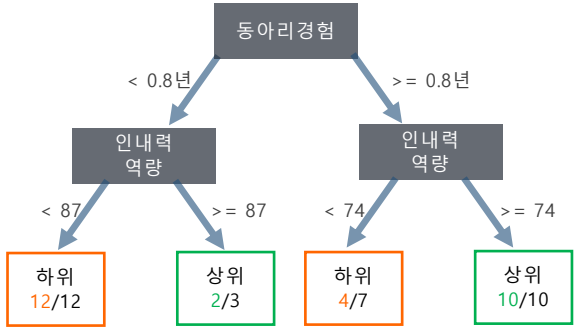


Decision Tree - Minimizing Entropy

성과에 따른
신입사원 분포



고성과 신입사원
분류모형 (의사결정트리)

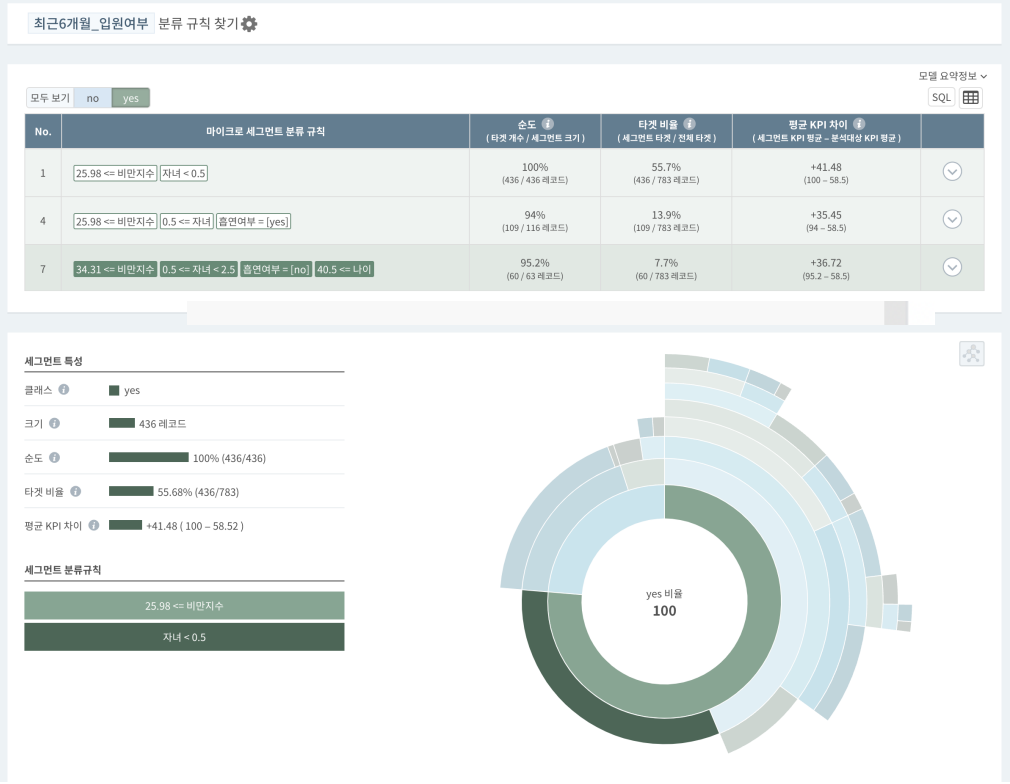


분류규칙

규칙 (Rule Set)	확률 (Probability)
IF (동아리경험 < 0.8년) and (인내력역량 < 87) then Class = 저성과 신입사원	100% (12/12)
IF (동아리경험 < 0.8년) and (인내력역량 >= 87) then Class = 고성과 신입사원	67% (2/3)
IF (동아리경험 >= 0.8년) and (인내력역량 < 74) then Class = 저성과 신입사원	57% (4/7)
IF (동아리경험 >= 0.8년) and (인내력역량 >= 74) then Class = 고성과 신입사원	100% (10/10)

보험 손해율 최적화를 위한 고객 세그먼트 규칙

- Decision Tree 알고리즘으로 특정 집단을 타게팅할 수 있는 세그먼트 규칙을 보여주는 화면
- 비만



실습 시간

EDA 도구: https://www.heartcount.io/login