

# 데이터로 특정 행동을 보인 집단 간 특성 차이 이해하기

#세그멘테이션 #코호트특성

양 승 준 / [sidney.yang@idk2.co.kr](mailto:sidney.yang@idk2.co.kr)



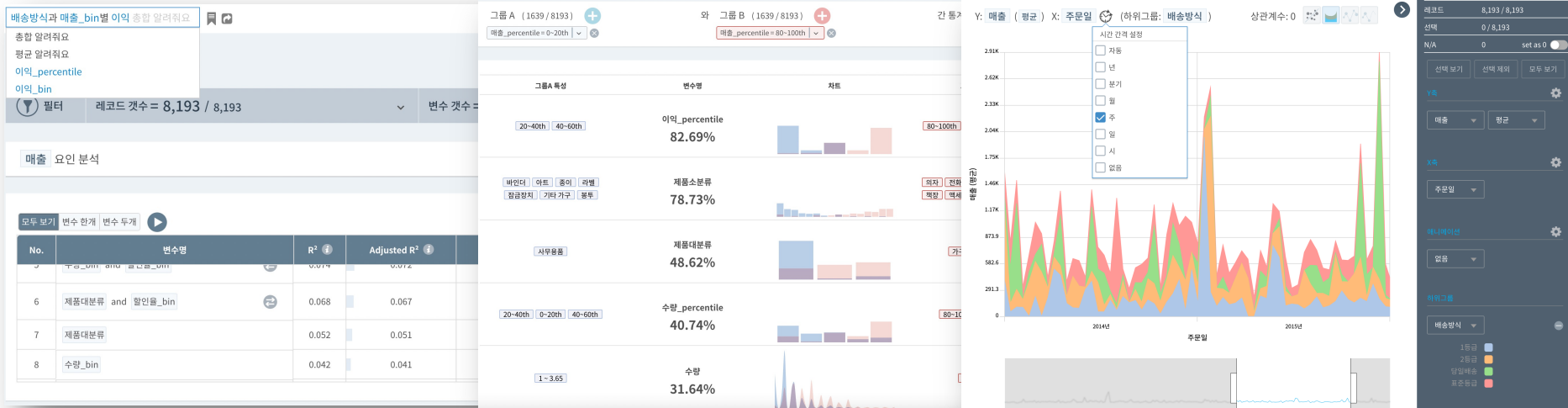
# 다정한 데이터 도구, HEARTCOUNT

- 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

- 특징점으로는 '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

### EDA(데이터 시각화) 강의

**DATA HERO ORIGINAL CONTENTS**

커뮤니티 소식

**강의 VOD**

EDA(데이터 시각화) 고급 통계 분석

**실습 예제**

EDA(데이터 시각화) 고급 통계 분석

**블로그(아티클)**

데이터나  
웹버 오픈 스테이스

**Upcoming Events**

6월 웨버나 | 분석하기 좋은 데이터셋을 구성하는 요소를 6월 30일 (금) 오후 3:00 - 3:30 사진 등록 하기

**DATA HERO ORIGINAL CONTENTS**

**Data Literacy**

© Literacy, Numeracy, Data Literacy: 데이터 리터러시 에 대해 이해하기

**DATA HERO ORIGINAL CONTENTS**

**데이터 분석 준비**

EDA 101 (1): 분석하기 좋은 데이터셋, 변수 유형별 시각화 방법

**DATA HERO ORIGINAL CONTENTS**

**데이터의 분포**

EDA 101 (2): 데이터의 모양 묘사하기 (히스토그램, boxplot, percentile)

**DATA HERO ORIGINAL CONTENTS**

**시각화 기초 문법**

EDA 101 (3): 평균의 함정, 시각화 기본 문법, 상관계수 분석

DATAHERO

- 📄 스케드
- 📄 다이렉트 메시지
- 📄 편성 및 반응
- 📄 초안 및 전송됨
- 📄 Slack Connect
- 📄 더 보기

채널

# 3. 질문-답변

다이렉트 메시지

앱

GreetBot

추가

# 3. 질문-답변 데이터 히어로 커뮤니티 관련 궁금한 점은 이곳에 남겨주세요.

적갈피 추가

2개의 댓글 3개월 전 마지막 댓글

2022년 11월 15일

2개의 댓글 3개월 전 마지막 댓글

2022년 11월 17일

오후 6:52

오늘 웨버나에서 연사님이 DA가 일부 DE 업무까지 하는 경우??? 특정한 직무이름을 말씀 주셨었는데... 정확하게 떠오르지 않아서 질문드립니다!

2개의 댓글 3개월 전 마지막 댓글

오전 11:05

안녕하세요.  
밋업 참여해서 강의 잘 듣고 있습니다.  
혹시 물어 사정이 있으신가요?  
저는 초심자라 물어 자체가 생경해서 이해도가 떨어지는 것 같아서요!  
백학상 이해는 하고 있지만 부족한 것 같이 느껴요.  
있다면 공유 부탁드립니다.

4개의 댓글 29일 전 마지막 댓글

#3. 질문-답변에 메시지 보내기

# 의료보험 청구비용 최적화를 위한 고객 세그먼테이션

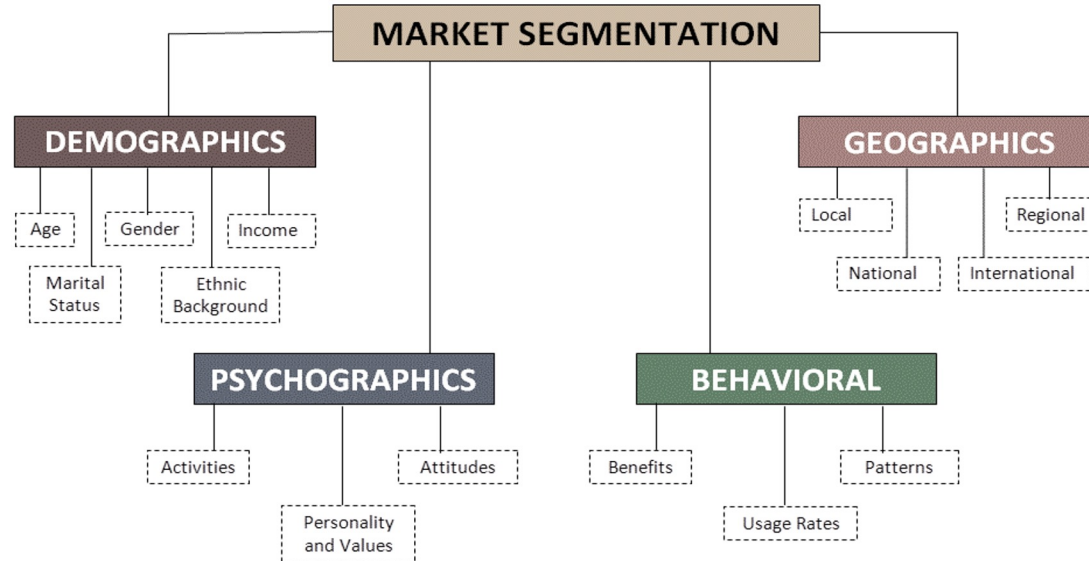
- 데이터셋(sample insurance)은 아래 표와 같이 구성되어 있음
    - <https://support.heartcount.io/community/learning/sample-dataset-sites>
1. 의료 보험 청구액의 차이를 가져오는 주요 요인을 찾고
  2. 보험 청구를 한 고객과 그렇지 않은 고객의 특성차이를 이해한 후
  3. 보험 청구하지 않을 확률이 높은 고객 세그먼트를 찾고자 함

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

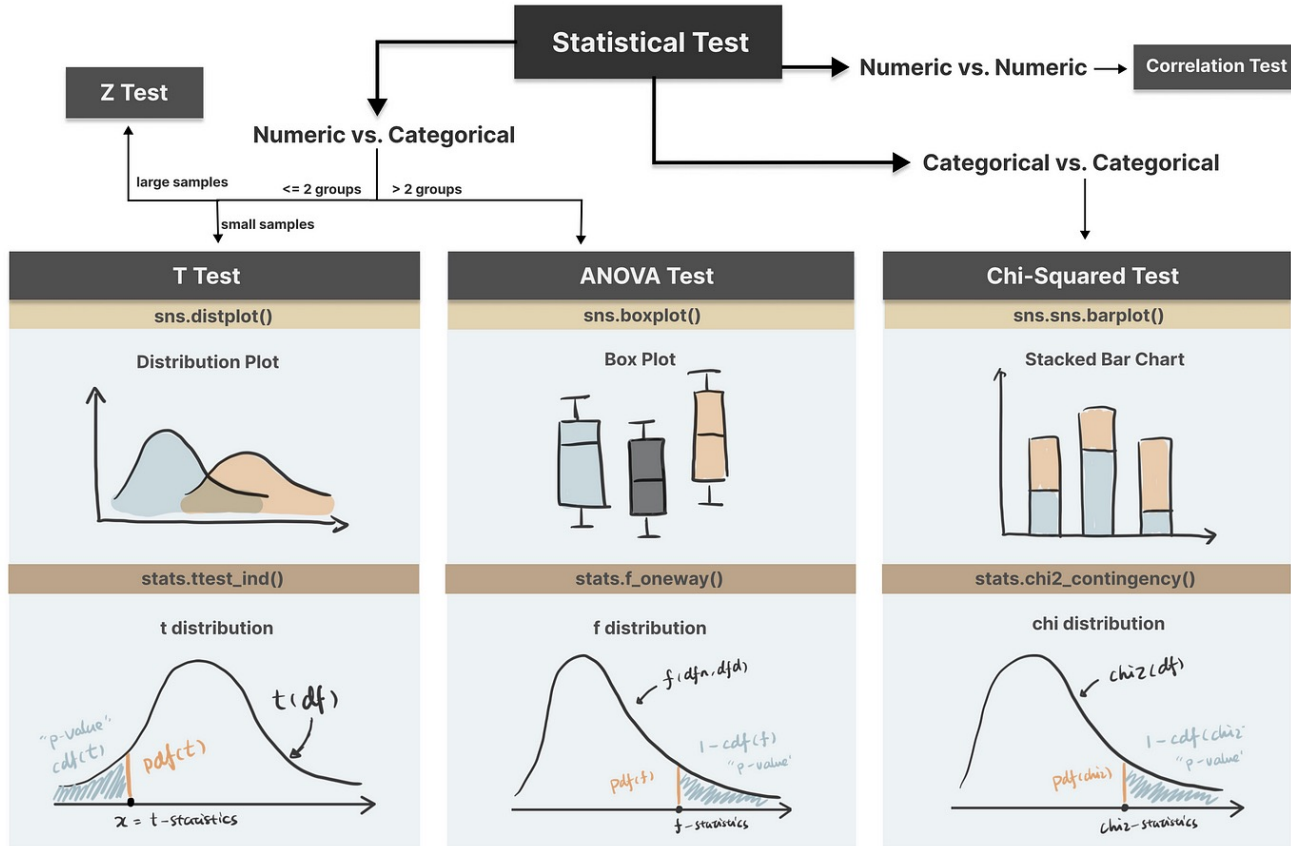
나이 / 성별 / BMI(체질량 지수) / 활동량(Steps per day) / 자녀수 / 흡연여부 / 지역 / 의료비 청구액 / 청구여부

# 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](http://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.

TABLE OF VALUES OF P FOR VALUES OF  $\chi^2$  AND  $n'$ ;  $\chi^2$  FROM 1 TO 70,  $n'$  FROM 5 TO 20\*

$n'$	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	0.000539	0.012538	0.024296	0.037566	0.051424	0.065861	0.080861	0.096419	0.112529	0.129194	0.146409	0.164170	0.182373	0.200914	0.219788	0.238991	0.258519	0.278359
2	0.077070	0.157053	0.236540	0.314534	0.390028	0.463016	0.533493	0.601454	0.666894	0.729808	0.789192	0.845052	0.897384	0.945194	0.988480	1.027248	1.061494	1.091225
3	0.274280	0.486915	0.673640	0.828264	0.954704	1.057008	1.139248	1.205408	1.258464	1.300416	1.333168	1.358616	1.376752	1.387472	1.391680	1.389376	1.380544	1.365184
4	0.539894	0.813760	1.048864	1.240000	1.391136	1.506272	1.589408	1.644544	1.684688	1.713840	1.734000	1.747160	1.754320	1.757480	1.757640	1.754800	1.748960	1.739120
5	0.854188	1.218264	1.548336	1.838400	2.082464	2.284528	2.447600	2.575680	2.672768	2.743856	2.791944	2.820032	2.831120	2.828208	2.813296	2.788384	2.753472	2.709560
6	1.179072	1.648128	2.078192	2.464256	2.800320	3.082384	3.315456	3.503536	3.650624	3.761712	3.831800	3.864888	3.873976	3.861064	3.828152	3.775240	3.702328	3.610416
7	1.513956	2.068032	2.598104	3.090176	3.548240	3.967304	4.342368	4.668432	4.940504	5.163576	5.332640	5.452704	5.520768	5.541832	5.517904	5.469968	5.408032	5.333104
8	1.858840	2.508916	3.139000	3.635072	4.093136	4.509200	4.879264	5.199328	5.465392	5.673456	5.829520	5.939584	5.999648	6.014712	6.000784	5.968848	5.918912	5.853984
9	2.213724	2.948800	3.578884	4.170960	4.729036	5.248104	5.723168	6.149232	6.521296	6.835360	7.087424	7.273488	7.400552	7.474616	7.500680	7.488744	7.448808	7.383880
10	2.578608	3.398684	4.028768	4.610844	5.158920	5.667984	6.133048	6.549112	6.911176	7.215240	7.457304	7.633368	7.740432	7.784496	7.770560	7.730624	7.665696	7.585768
11	2.953492	3.858568	4.488652	5.070728	5.618804	6.127868	6.592932	7.009000	7.371064	7.675128	7.917192	8.093256	8.200320	8.244384	8.230448	8.180512	8.115584	8.035656
12	3.338376	4.328452	4.958536	5.540612	6.088688	6.597752	7.062816	7.478880	7.840944	8.145008	8.397072	8.583136	8.700200	8.744264	8.720328	8.660392	8.585464	8.495536
13	3.733260	4.808336	5.438420	6.020496	6.568572	7.077636	7.542700	7.958764	8.316828	8.610892	8.842956	9.013020	9.110084	9.144148	9.110212	9.050276	8.975348	8.885420
14	4.138144	5.298220	5.928304	6.510380	7.058456	7.567520	8.032584	8.448648	8.806712	9.100776	9.332840	9.502904	9.600968	9.625032	9.581096	9.511160	9.421232	9.321304
15	4.553028	5.803104	6.433188	7.015264	7.563340	8.072404	8.537468	8.953532	9.311596	9.605660	9.837724	10.007788	10.105852	10.130916	10.086980	10.017044	9.927116	9.827188
16	4.977912	6.317988	6.948072	7.530148	8.078224	8.587288	9.052352	9.468416	9.826480	10.120544	10.352608	10.522672	10.620736	10.645800	10.591864	10.511928	10.412000	10.302072
17	5.412796	6.842872	7.472956	8.055032	8.603108	9.112172	9.577236	9.993300	10.351364	10.645428	10.877492	11.047556	11.145620	11.170684	11.126748	11.036812	10.936884	10.826956
18	5.857680	7.387756	8.017840	8.600916	9.148992	9.658056	10.123120	10.539184	10.907248	11.211312	11.453376	11.623440	11.711504	11.726568	11.672632	11.562696	11.442768	11.312840
19	6.312564	7.932640	8.562724	9.145800	9.693876	10.202940	10.668004	11.084068	11.452132	11.766196	12.018260	12.200324	12.308388	12.333452	12.279516	12.159588	12.029660	11.889732
20	6.777448	8.497524	9.127608	9.710684	10.258760	10.767824	11.232888	11.648952	12.017016	12.321080	12.573144	12.755208	12.853272	12.878336	12.824400	12.704472	12.574544	12.434616

\* I have to thank Miss Alice Lee, D.Sc., for help in the calculation of part of this table. The certain  $xy$ , i. e. the 1's in columns 16 to 20, denotes, of course, something greater than 999,9995, i. e. unity to six figures.

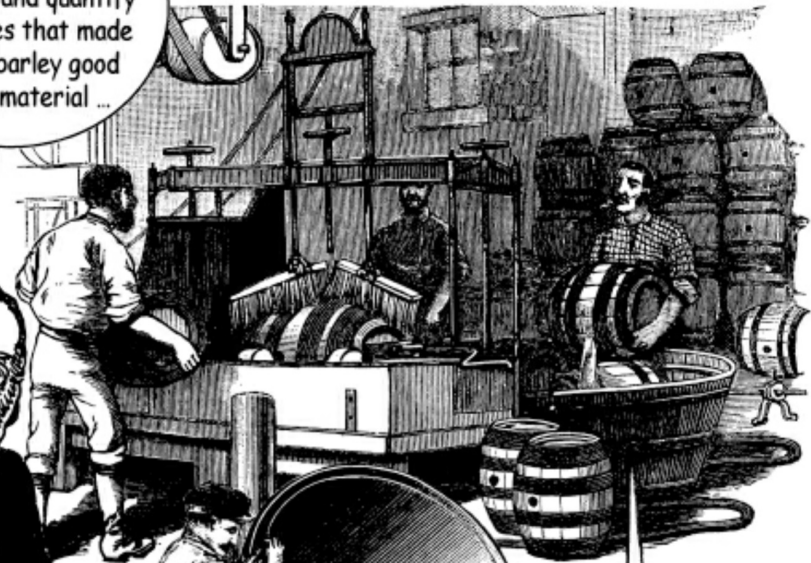
At the end of the 19th century Guinness was the largest brewery in the world, producing more than 1.5 million barrels a year.

To maintain this position, they began to appoint men who had first-class science degrees from Oxford and Cambridge, and they also adopted a policy of sending staff away for specialized study.

PEARSON'S FIRST CHI-SQUARE PROBABILITY TABLE IN 1900.



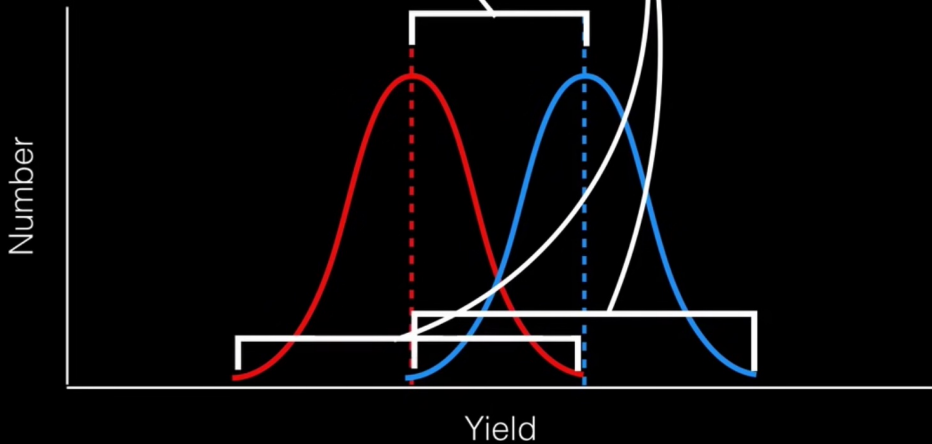
Some of the brewers undertook chemical analyses to try to identify and quantify the qualities that made hops and barley good brewing material ...



... such as the "rub" of the hops or the "texture" of barley that might be "milky" or "steely".

$$\frac{\text{signal}}{\text{noise}} = \frac{\text{difference between group means}}{\text{variability of groups}} = \frac{|\bar{x}_1 - \bar{x}_2|}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

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

$$df = n_1 + n_2 - 2$$

Don't  
Reject

H<sub>0</sub>

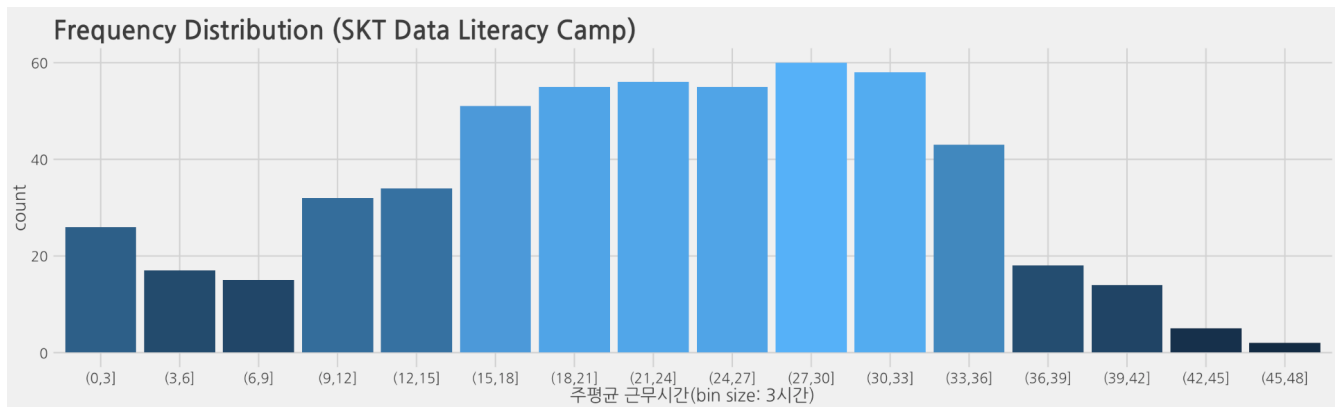
Degrees of Freedom	p=0.05	p=0.025	p=0.01
1	12.71	25.45	63.66
2	4.30	6.20	9.92
3	3.18	4.17	5.84
4	2.78	3.50	4.60
5	2.57	3.16	4.03
6	2.45	2.97	3.71
7	2.36	2.84	3.50
8	2.31	2.75	3.36
9	2.26	2.68	3.25
10	2.23	2.63	3.17
11	2.20	2.59	3.11
12	2.18	2.56	3.05
13	2.16	2.53	3.01
14	2.14	2.51	2.98
15	2.13	2.49	2.95
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

t-test

Reject

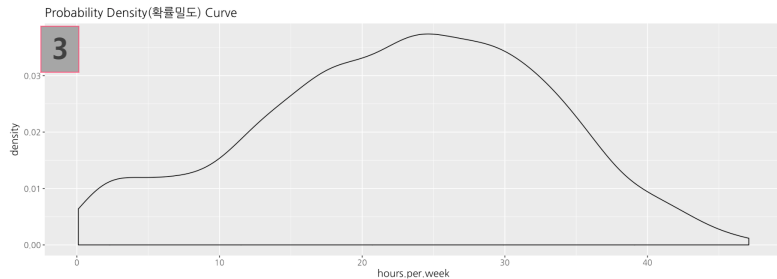
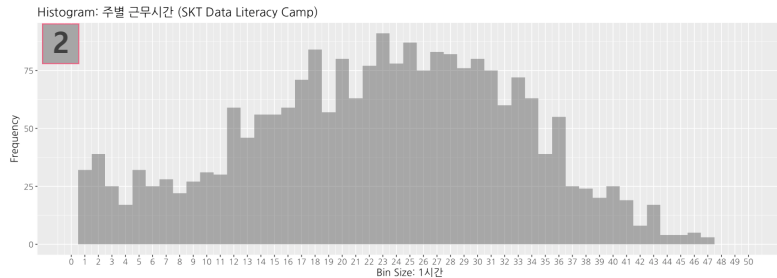
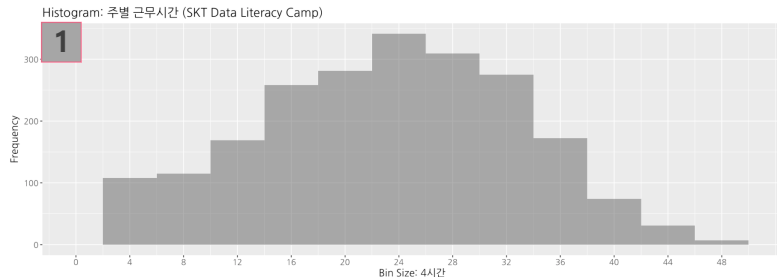
# Histogram vs. Frequency Distribution Table

## 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]

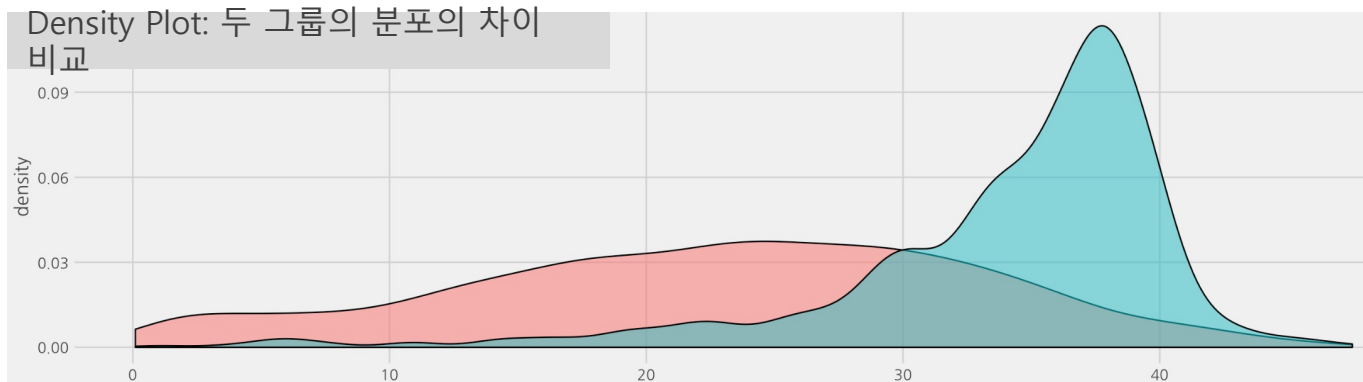
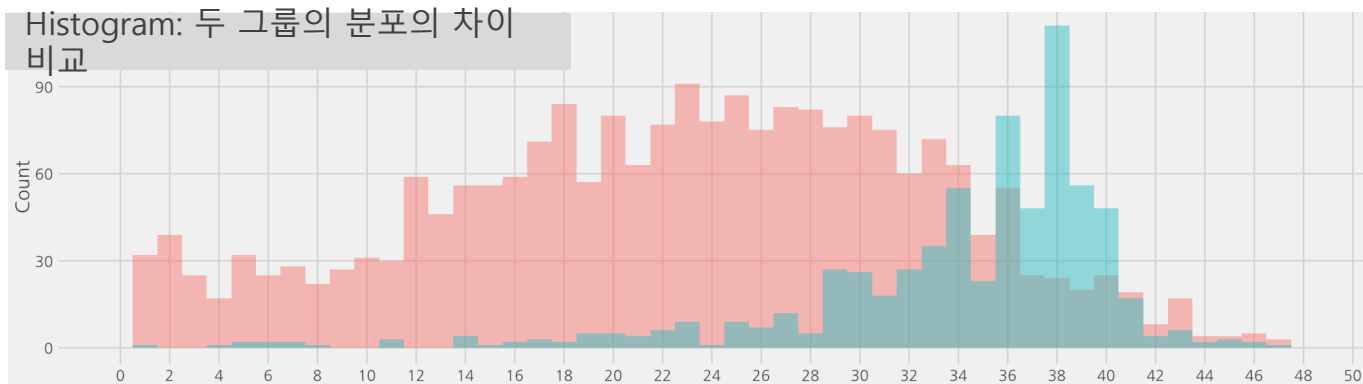
## 2 Histogram – Bin Size: 1시간

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

## 3 Probability Density Curve

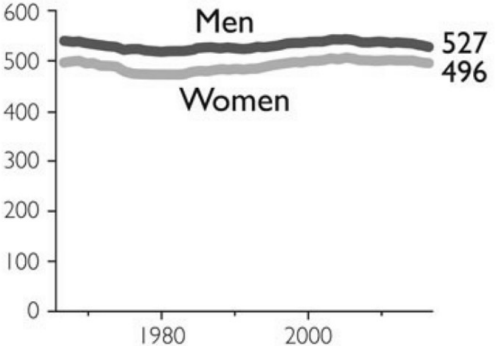
- 확률밀도: X가 연속형 변수일 경우 X값과 이에 대응하는 확률을 나타낸 그래프
- 좌측에서 X가 10~20시간 사이의 값을 가질 확률은 해당 구간의 면적과 동일함

# Histogram vs. Density Plot

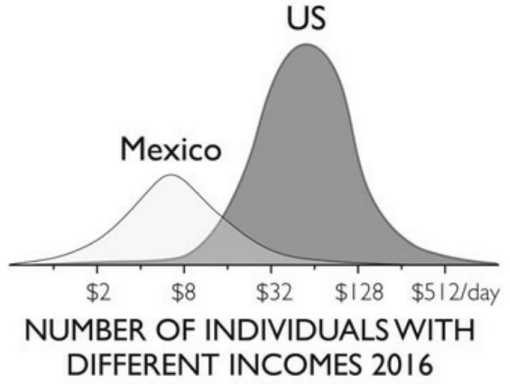
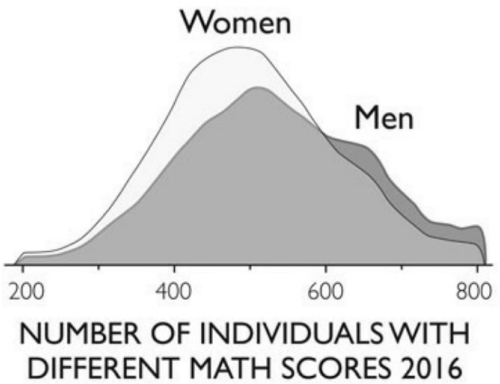
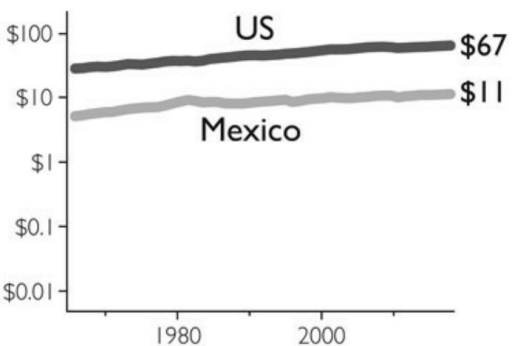


# Comparison of Average vs. Distribution

AVERAGE MATH SCORES

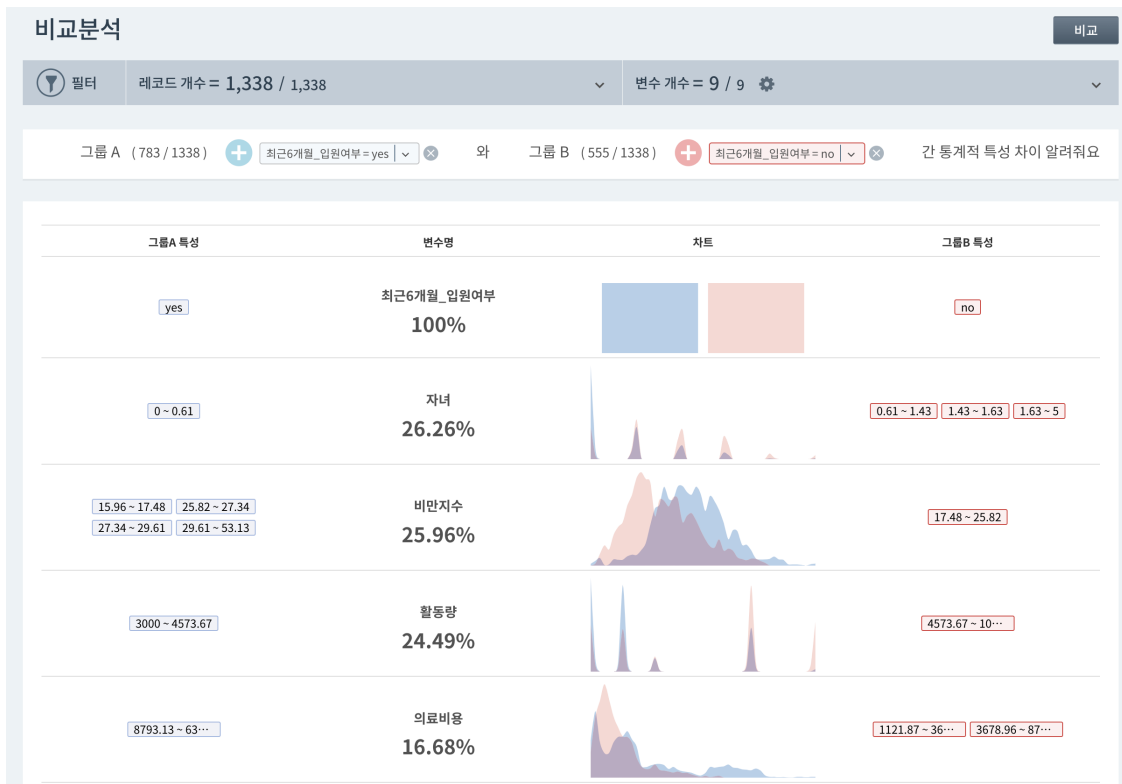


AVERAGE INCOME \$/DAY



# 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





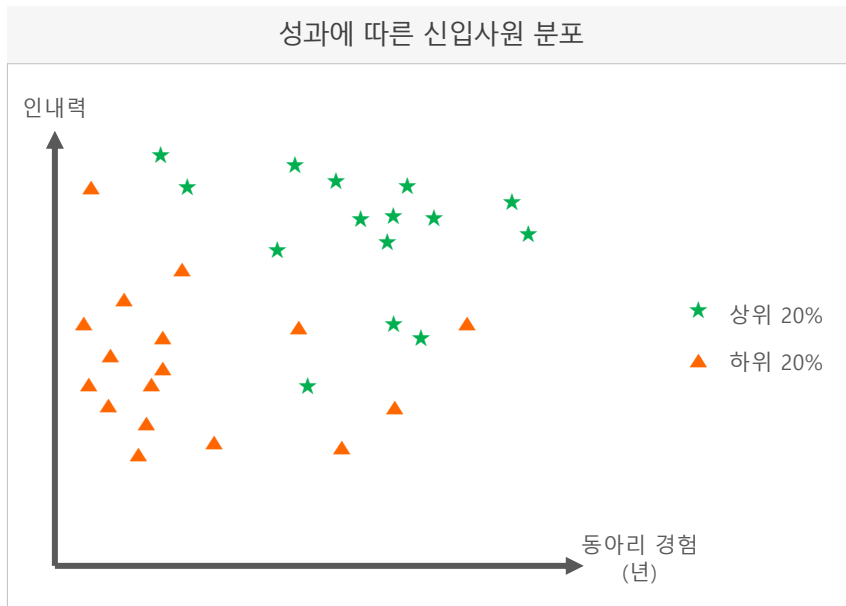


# Decision Tree - Minimizing Entropy

## Decision Tree Algorithm

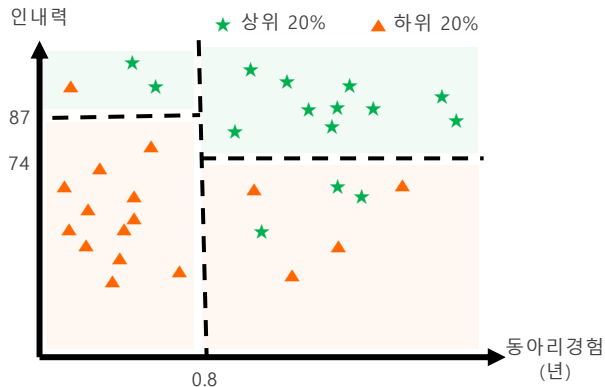
- Purity: \*엔트로피를 최소화하도록(= 끼리끼리 모이도록) 공간 구획
- Homogeneity: 동질적 집단이 밀집한 세그먼트의 논리적 규칙 찾기

\*엔트로피(Entropy): Measure of Impurity

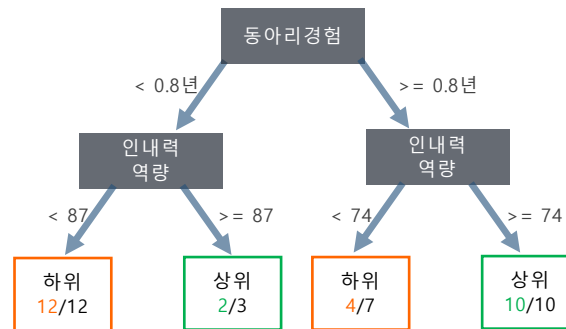


# Decision Tree - Minimizing Entropy

성과에 따른  
신입사원 분포



고성과 신입사원  
분류모형 (의사결정트리)

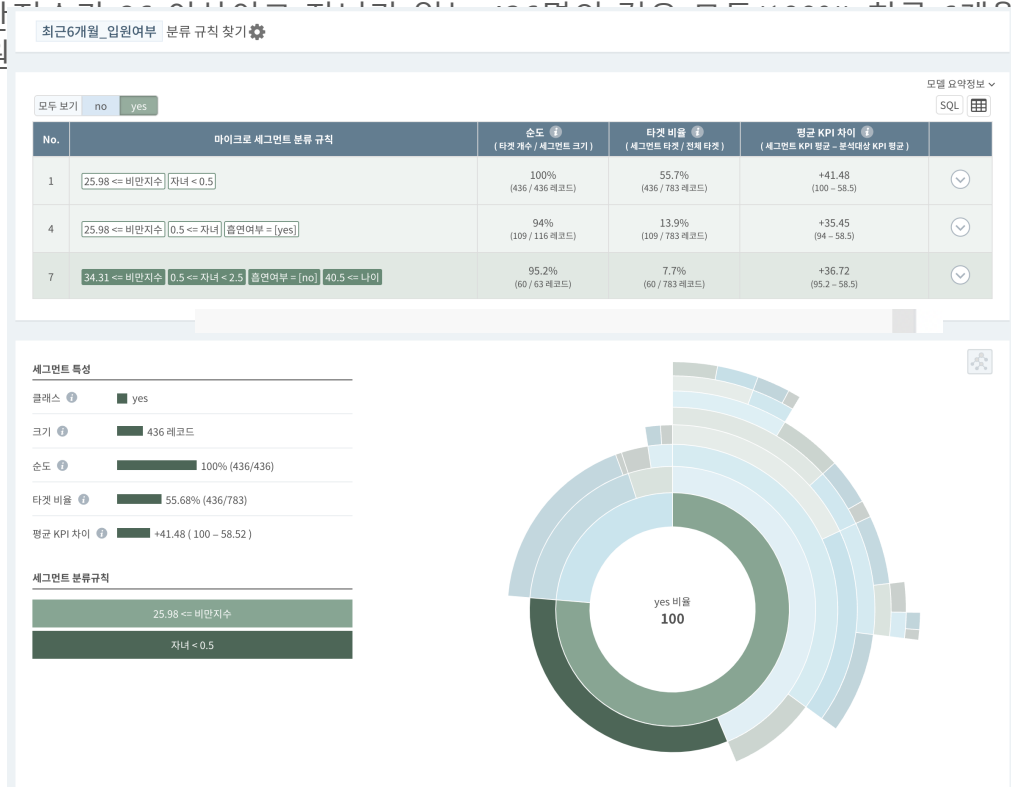


## 분류규칙

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 손해를 최적화를 위한 고객 세그먼트 규칙

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



이내



## 실습 시간

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