HR Analytics

- HR Data를 활용한 퇴사 예측 모델 구현 -

멋쟁이 사자처럼 AI SCHOOL 8기

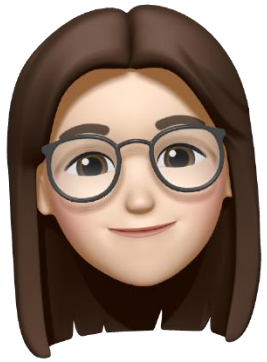
7조 죽어도 못 보내

김조은, 임승민, 조세연, 차은서

목차

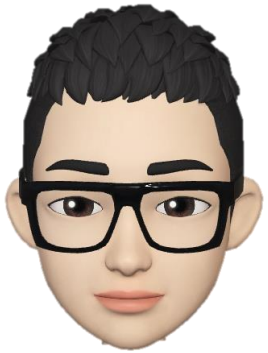
- 팀원 소개
- 주제 선정
- 데이터 개요
- 데이터 분석
- 활용방안

팀원소개



김조은

깃헙관리
머신러닝 고도화
코드취합

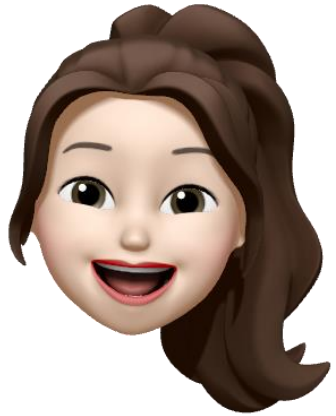


임승민

머신러닝 고도화
설문지 제작

조세연

PM
태블로 대시보드 제작
발표자



차은서

태블로 대시보드 제작
PPT 제작
설문지 제작

공통 역할 : 도메인 조사, 전처리, EDA, 발표 자료 제작

01 주제선정

프로젝트 배경

이직을 RESPECT, MZ를 중심으로 대퇴사 시대

"평생 직장? 옛말이죠" 퇴사 결심하는 20·30세대

프로젝트 배경

직원 1명 당 채용비용 1300만원, 교육비용 6000만원

https://www.saramin.co.kr/zf_user/help/live/view?idx=108748&listType=news

https://post.naver.com/viewer/postView.nhn?memberNo=24090434&volumeNo=6470527

프로젝트 배경

HR Analytics: HR 문제, 데이터로 해결

퇴사 예측 인공지능과 HR 시각화 대시보드

[HR테크의 진화] 회사 그만둘 직원, 95% 정확도로 예측

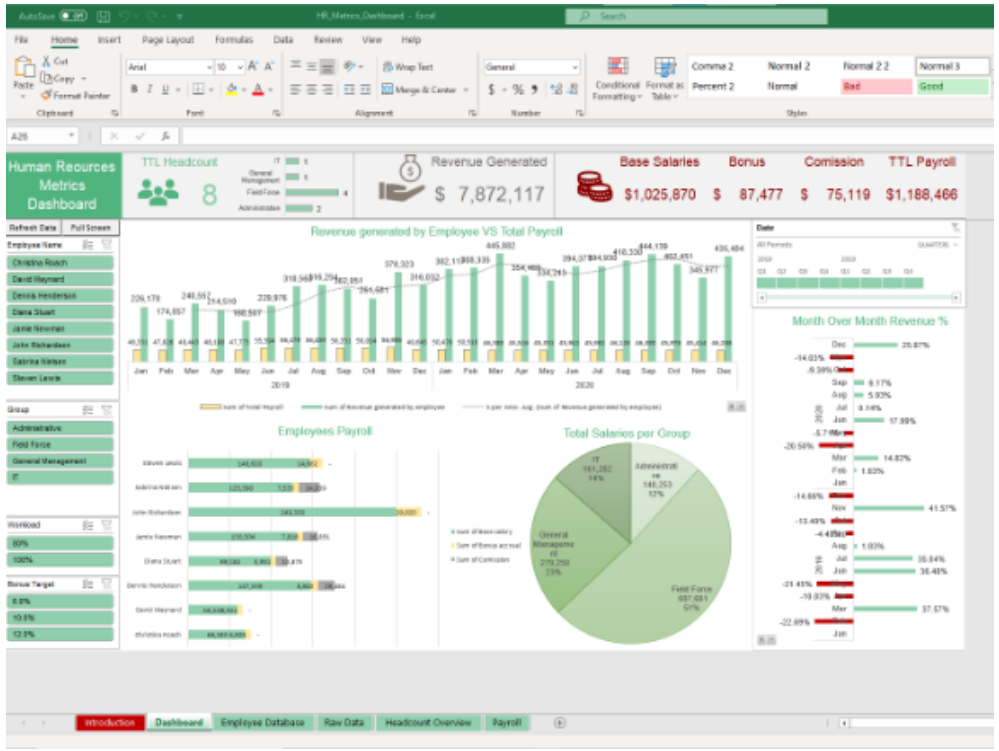
입력 2019-04-22 06:01

배준호 기자 baejh94@etoday.co.kr

퇴사 예측해 대체 인력 선제적 고용…IBM, 3400억 비용 절감

▲지니 로메티 IBM 최고경영자(CEO). 그는 최근 미국 CNBC방송이 개최한 행사에서 인공지능(AI)으로 퇴사할 직원의 95%를 미리 예측할 수 있다고 강조했다. AP뉴시스

https://www.etoday.co.kr/news/view/1747355



https://www.simplesheets.co/hr-metrics-dashboard

프로젝트 목표

- ✔ 탐색적 데이터 분석(EDA)을 통해 퇴사 여부 예측 머신러닝 모델 구축
- ✔ 퇴사 원인을 탐색적으로 파악할 수 있는 시각화 대시보드 제작

=> HR 부서의 인력 계획 의사결정을 위한 인사이트를 제공하자!

02 데이터 개요

데이터 개요

IBM HR Analytics Employee Attrition & Performance

Predict attrition of your valuable employees

https://www.kaggle.com/datasets/pavansubhasht/ibm-hr-analytics-attrition-dataset

구분	설명
소개	IBM에서 만든 가상의 HR Data
형식	csv
개수	1개
shape	(1470, 34)
전처리	불필요한 컬럼 제거, 연령대 및 연차 파생변수 생성

03 데이터 분석

데이터 분석 개요

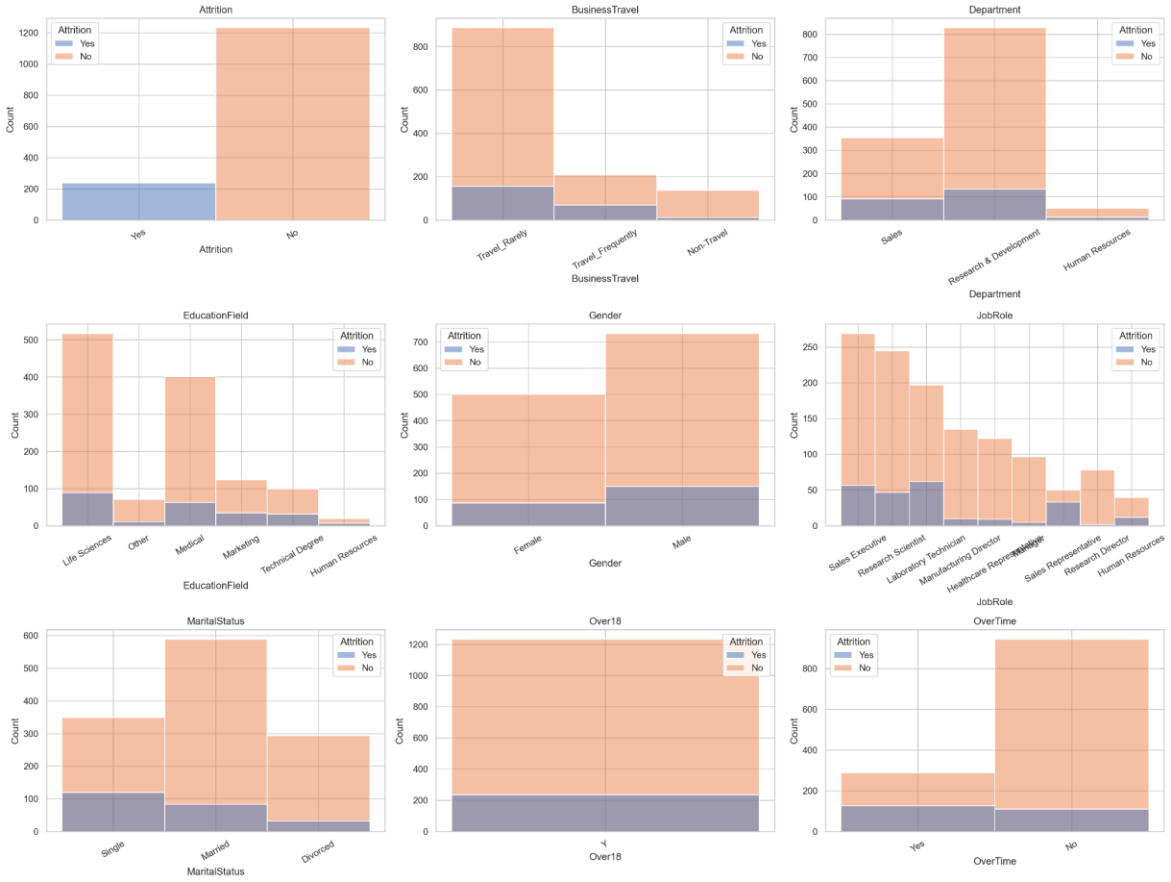
EDA

주요 피처 선정

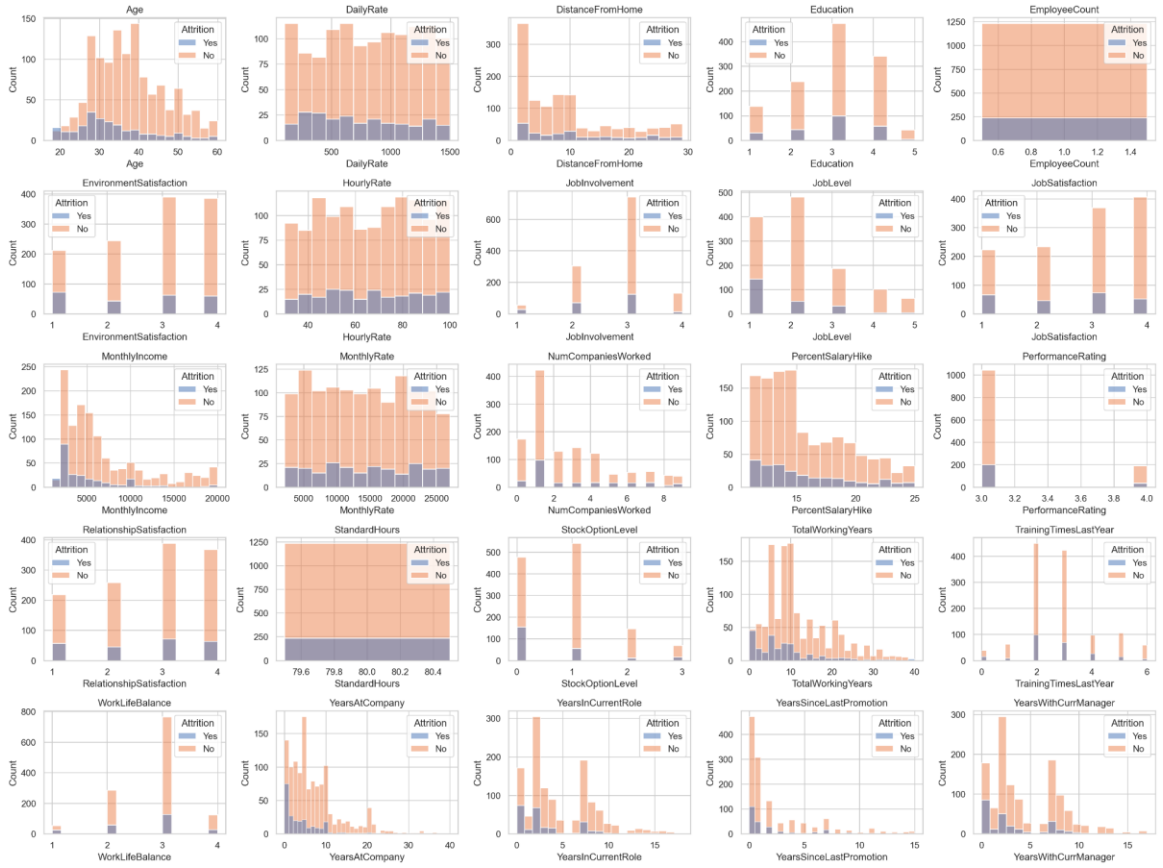
머신러닝
시각화 대시보드
설문지

1. EDA

전체 데이터 확인

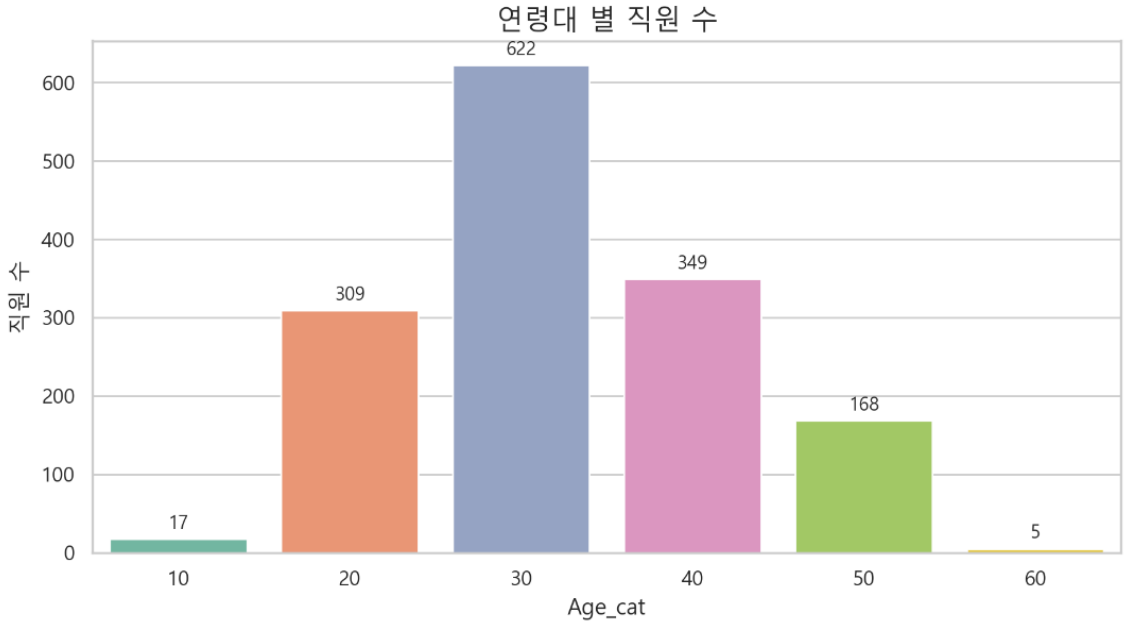


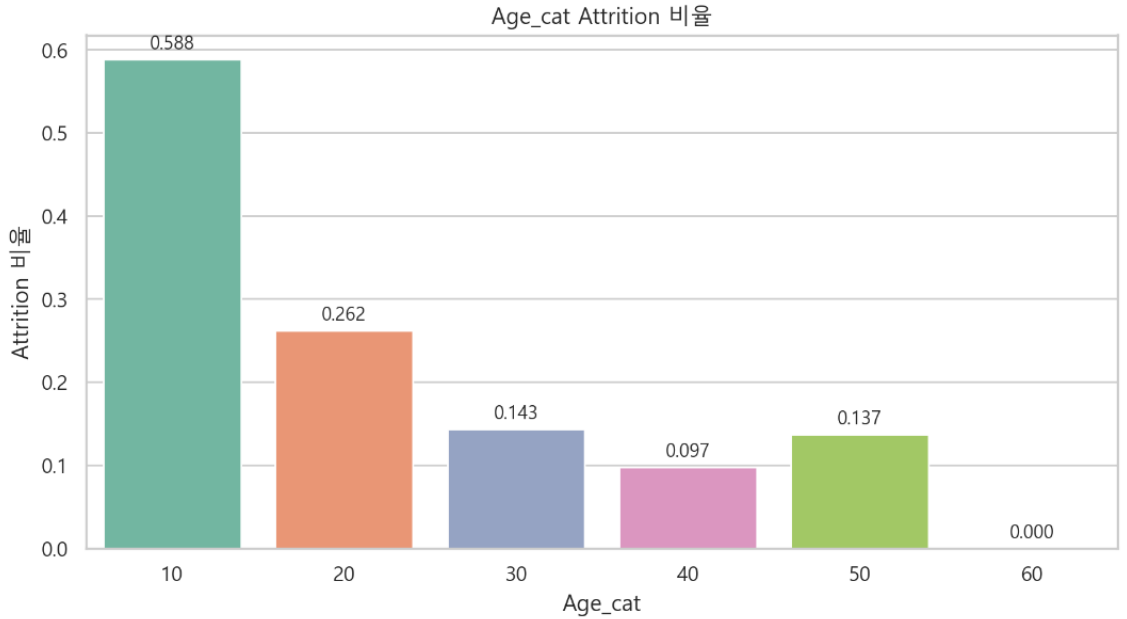
범주형 변수



수치형 변수
(만족도 변수 포함)

1. EDA





연령대 별 직원 수 순위

1위: **30대**

2위: 40대

3위: **20대**

연령대 별 퇴사비율 순위

1위: 10대

2위: **20대**

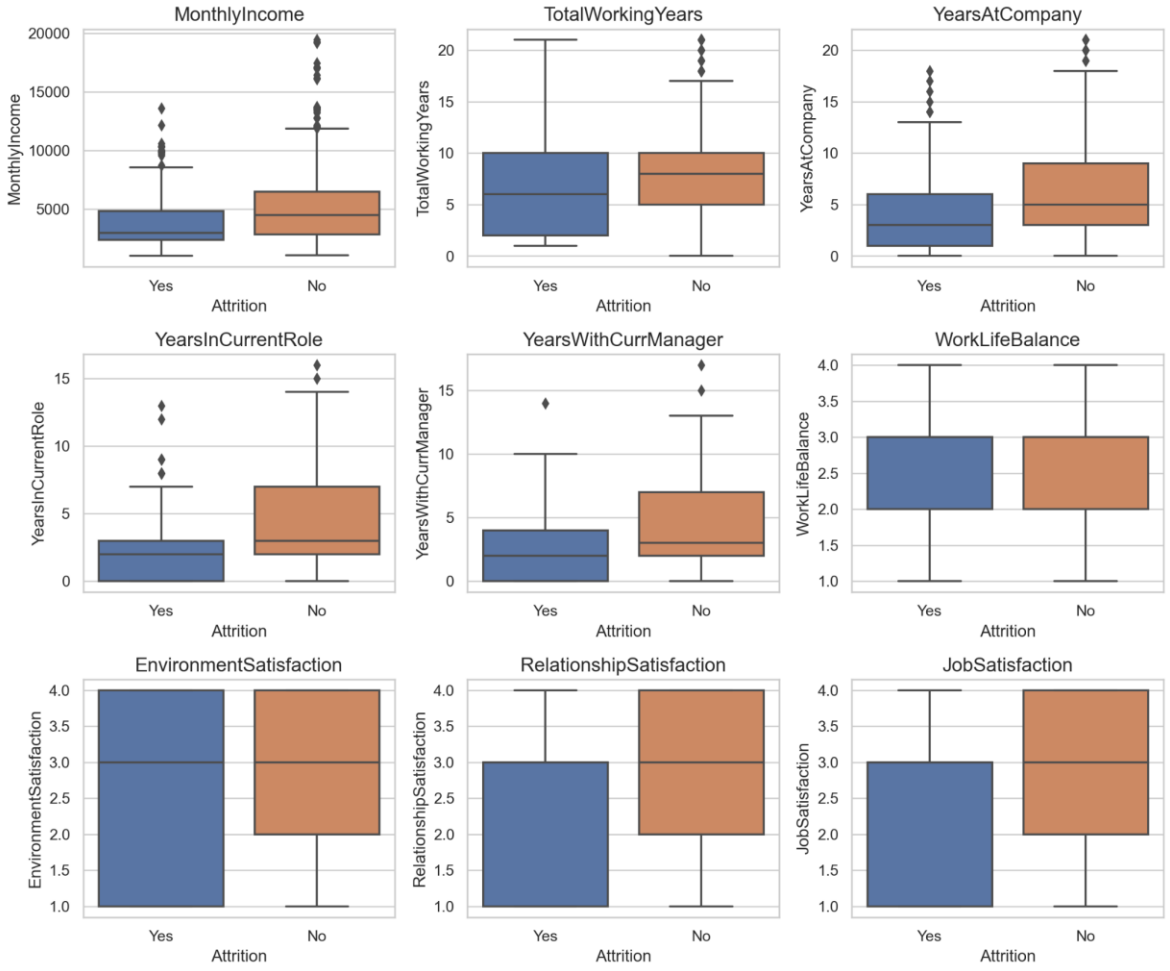
3위: **30대**

→ 가장 직원수가 많고 퇴사율도 높은 나이대인 20~30대를 중심으로 분석

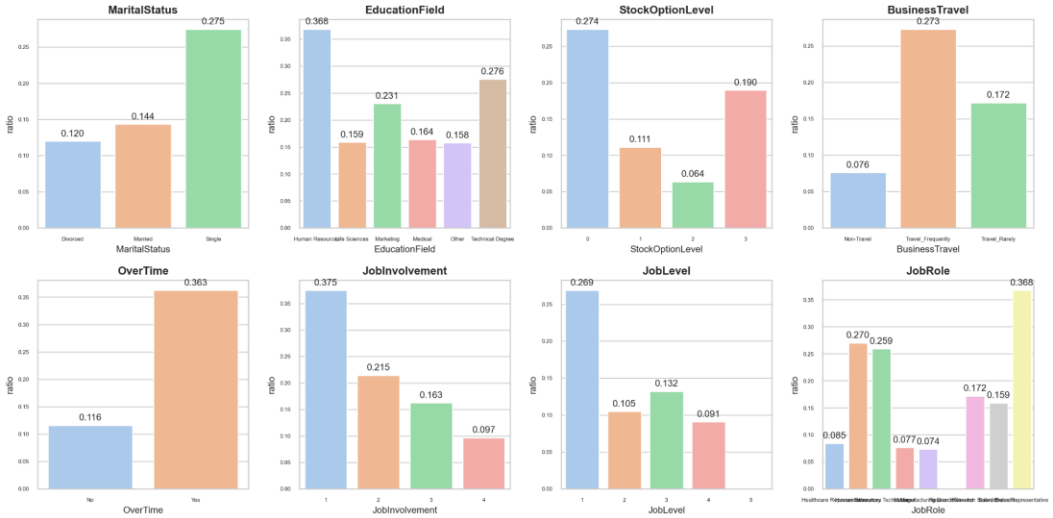
1. EDA

〈2030세대만 포함한 데이터 기준〉

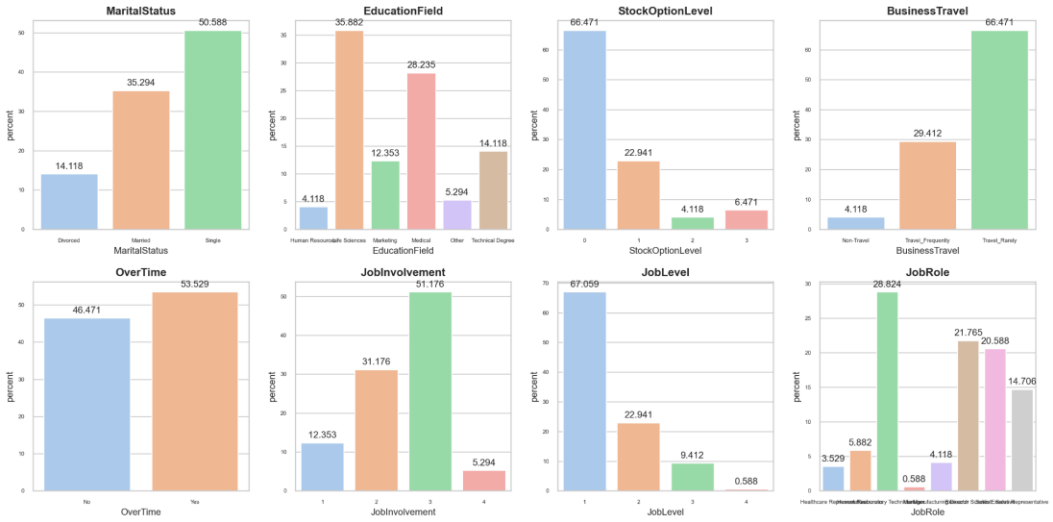
퇴사 여부에 따른 수치형 변수 분포



각 요소별 퇴직 비율(범주형 변수)



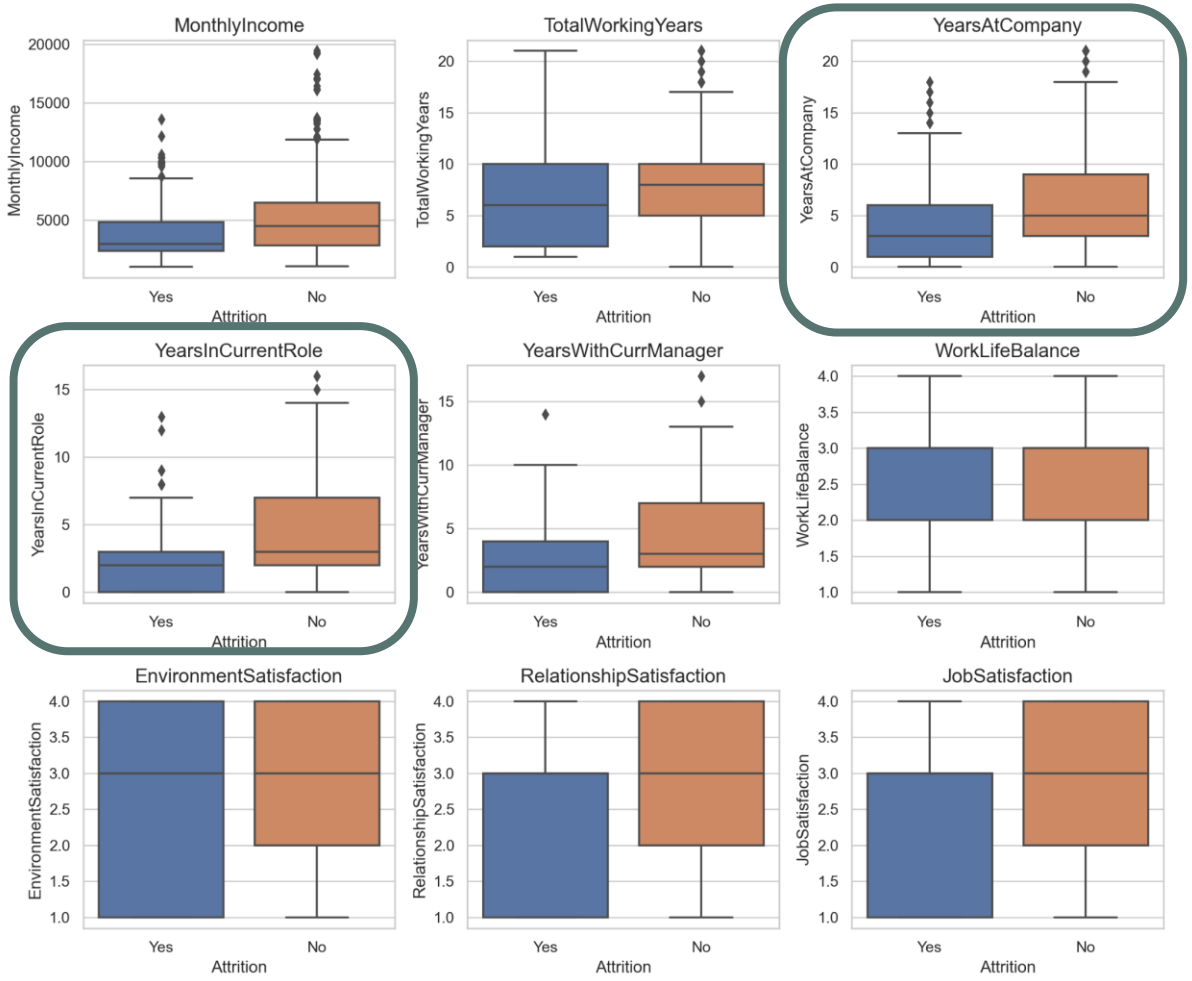
퇴직자 중 각 요소 비율(범주형 변수)



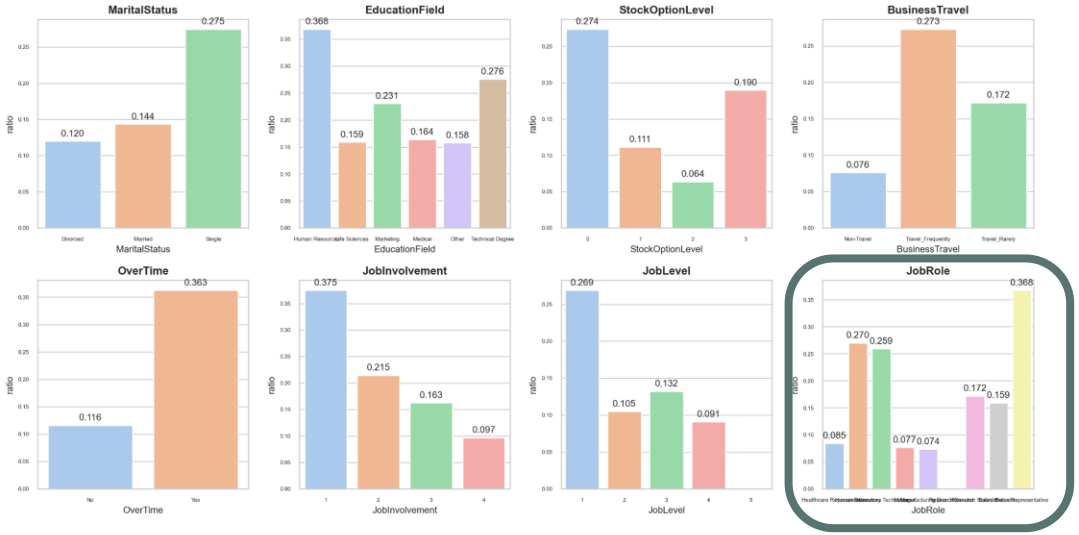
1. EDA

〈2030세대만 포함한 데이터 기준〉

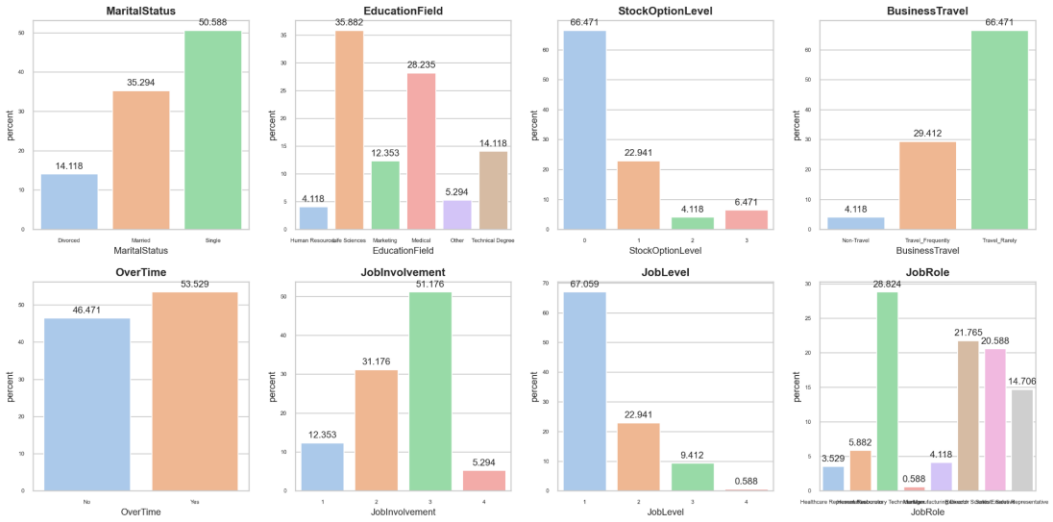
퇴사 여부에 따른 수치형 변수 분포



각 요소별 퇴직 비율(범주형 변수)

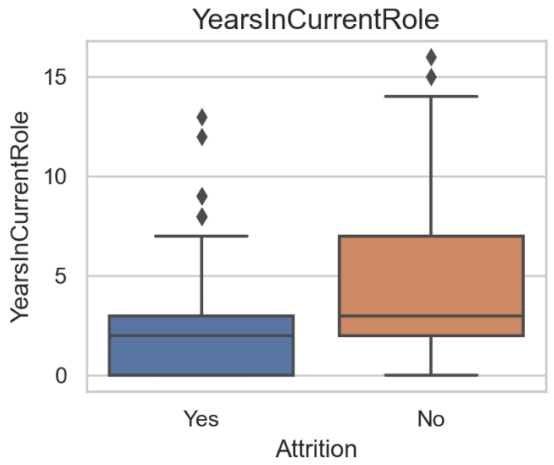


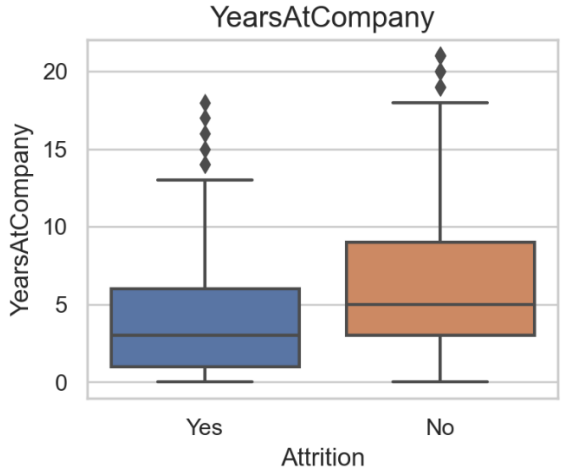
퇴직자 중 각 요소 비율(범주형 변수)

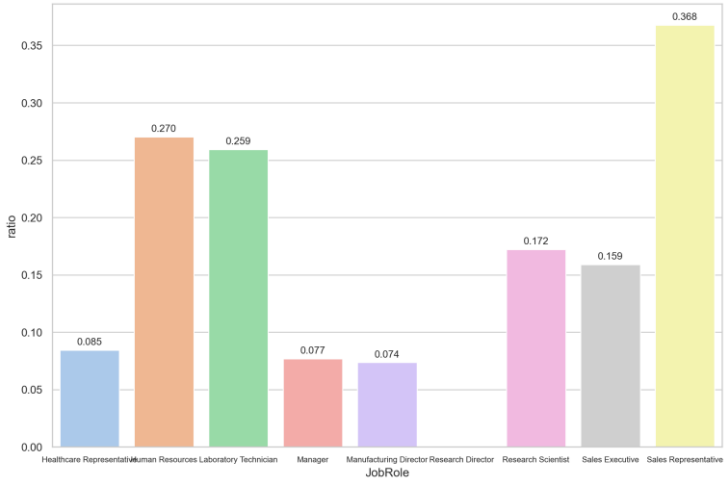


1. EDA

2030의 특이점을 나타내는 3가지 그래프를 중심으로 집중 분석







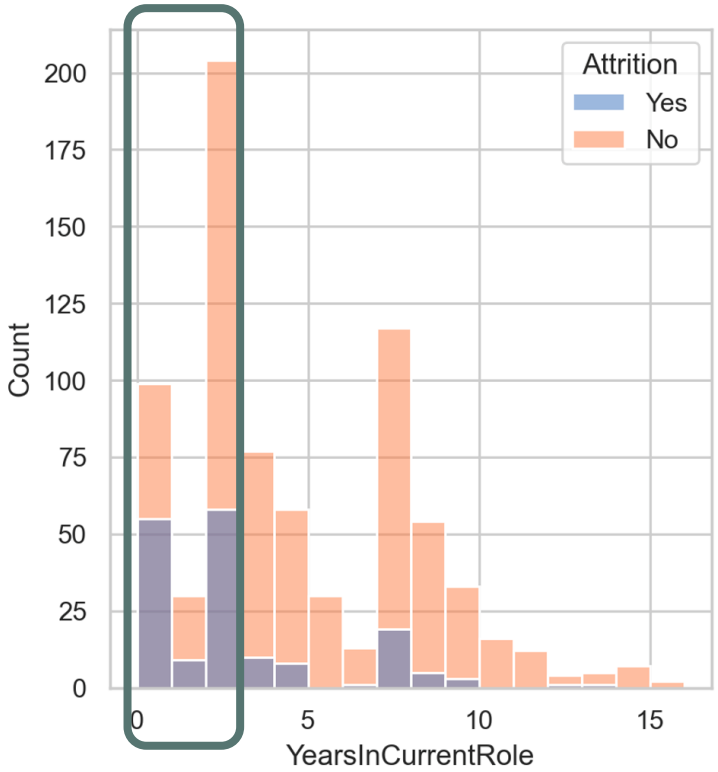
📌 2030 퇴직자는
현재 직무로 일한 기간이
왜 짧을까?

📌 2030 퇴직자는
현 회사에서 근무한
기간이 왜 짧을까?

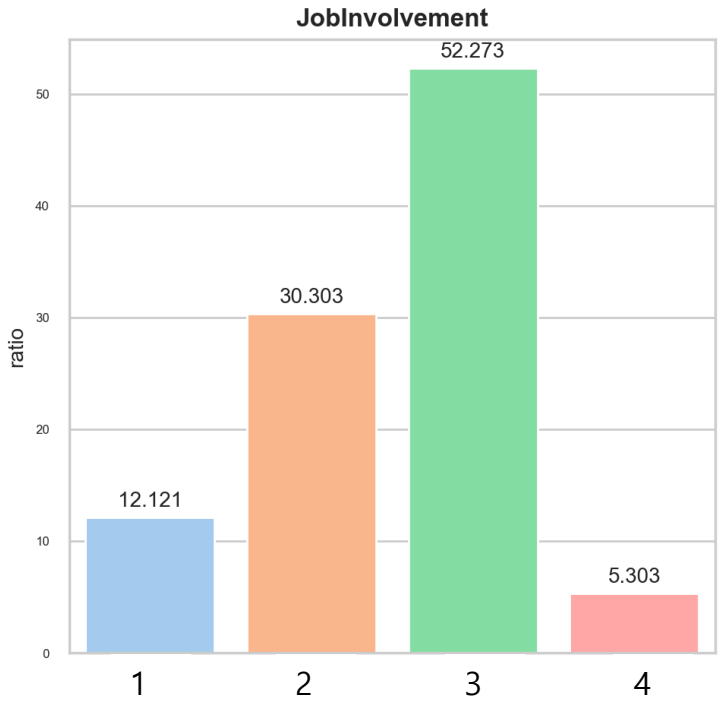
📌 2030 퇴사율이 높은 직무들은
왜 퇴사율이 높을까?

1. EDA

01) 현재 직무로 일한 기간이 짧은 사람들이 퇴직을 많이 한다.



2030, 현 직무 3년 이하, 퇴직자들의 업무 기여도

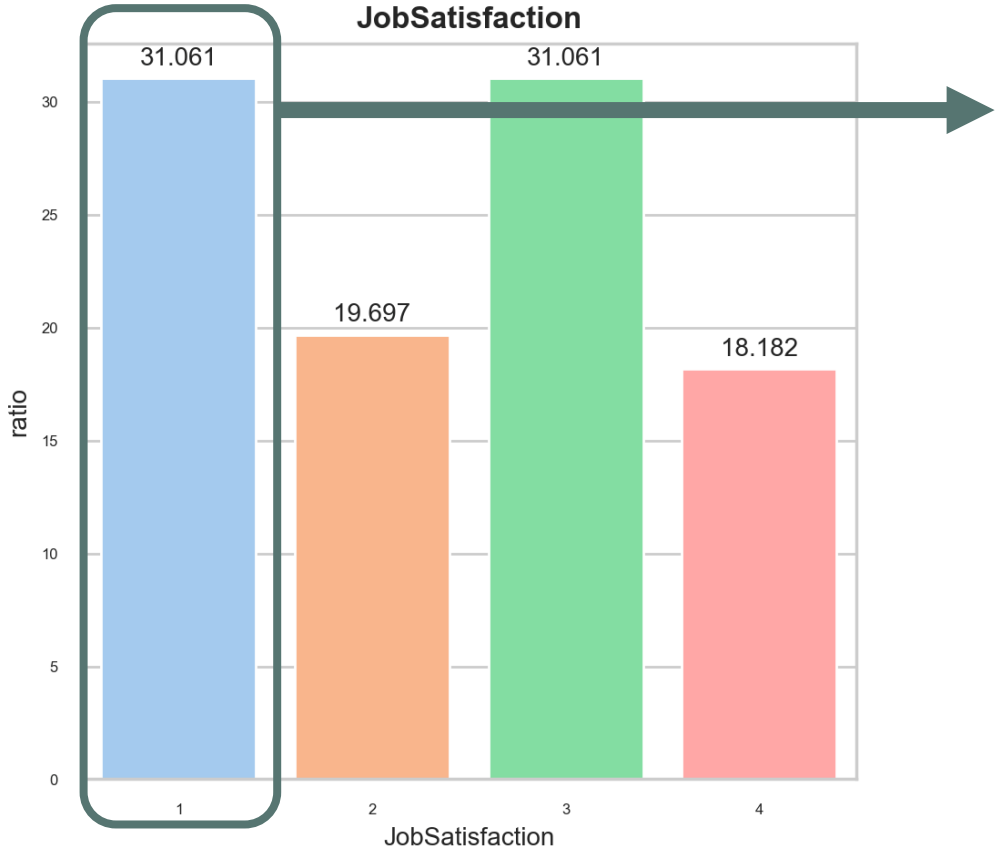


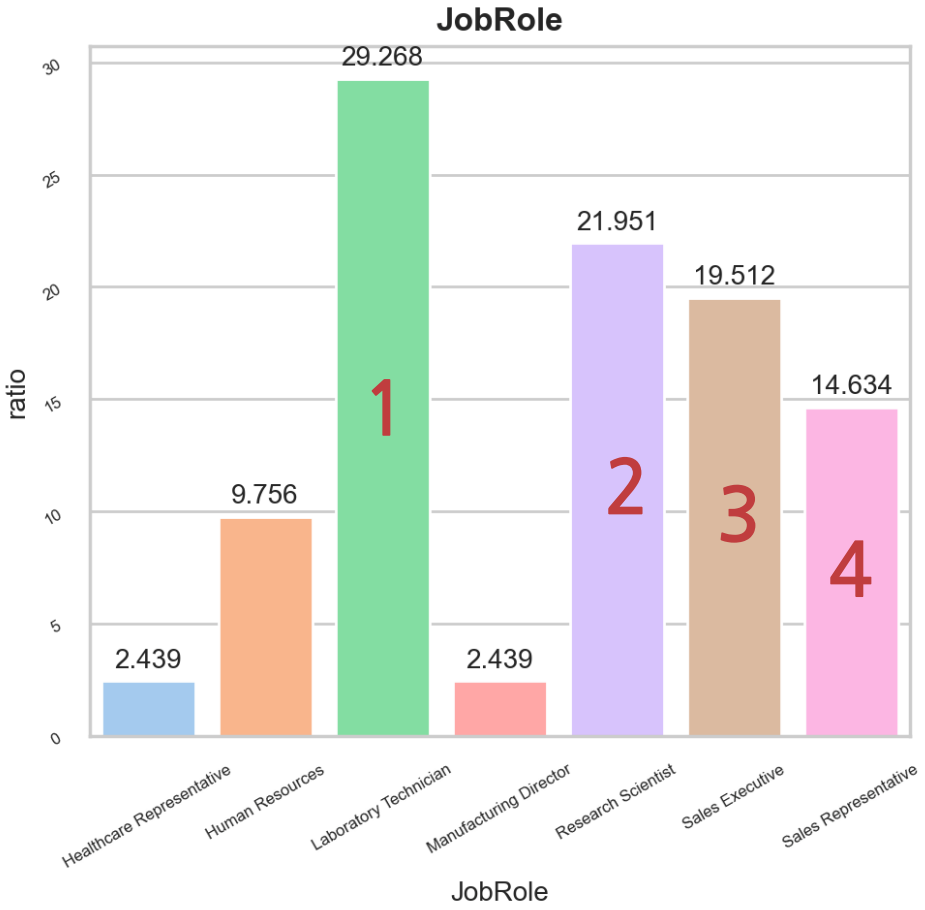
3년 이하로 일한 사람들, 업무 기여도가 높은데 많이 퇴사하는 이유?

```
df_2030_3yinrole
= df_2030[df_2030['YearsInCurrentRole']<4]
```

1. EDA

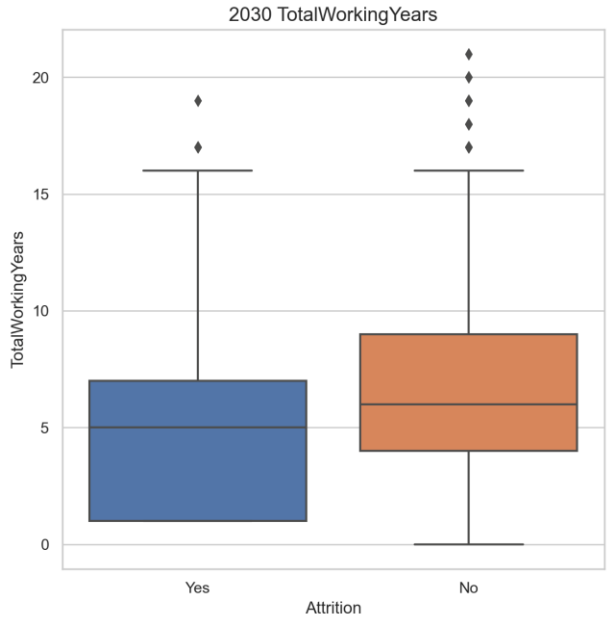
1. 직무가 만족스럽지 못하다: 기술, 판매직

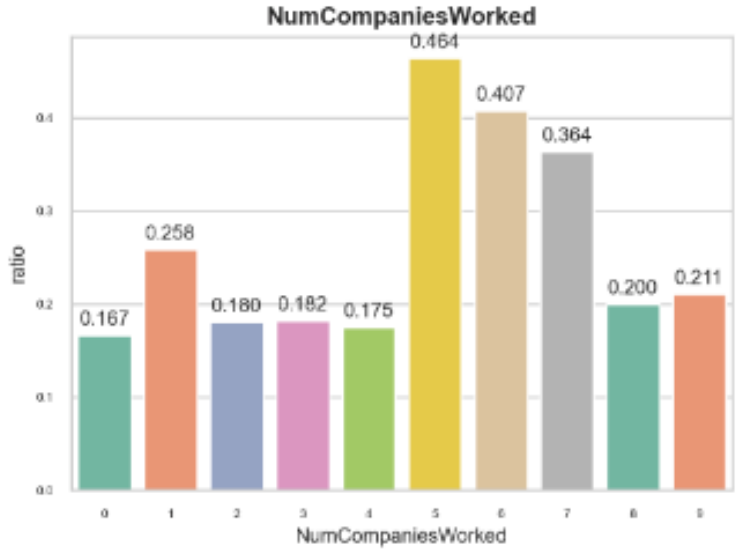


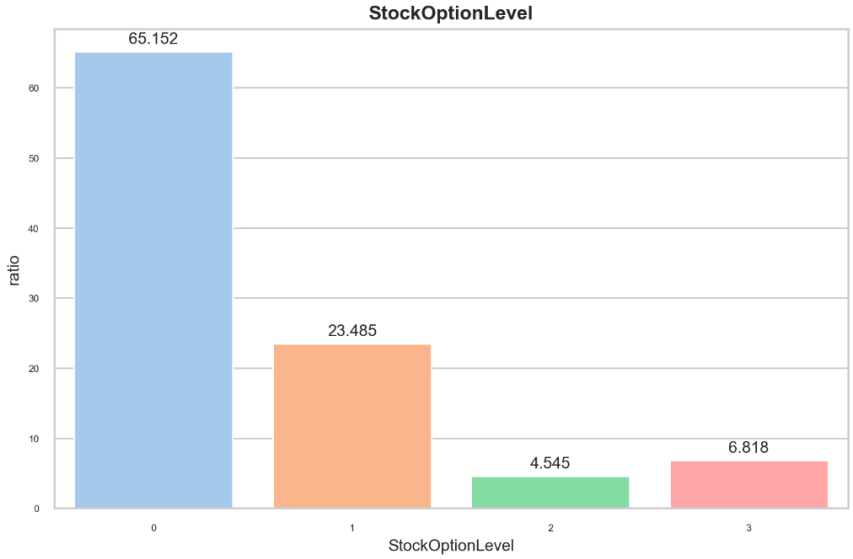


1. EDA

2. 회사를 쉽게 옮기는 사회 초년생



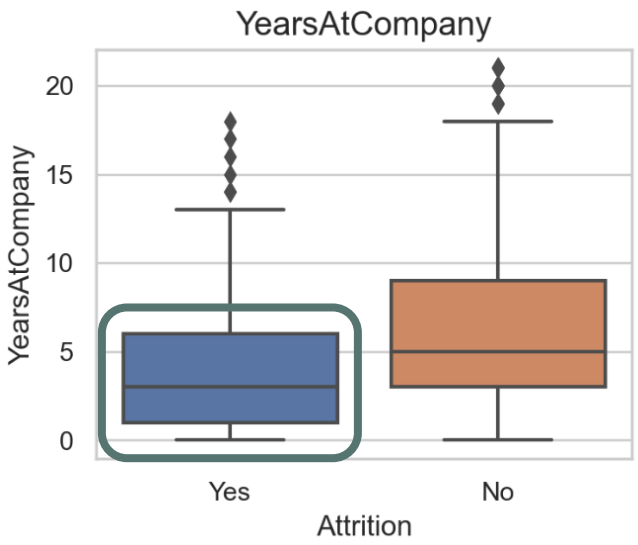




📌 **회사에 소속감을 잘 못 느끼는 2030의 특징**

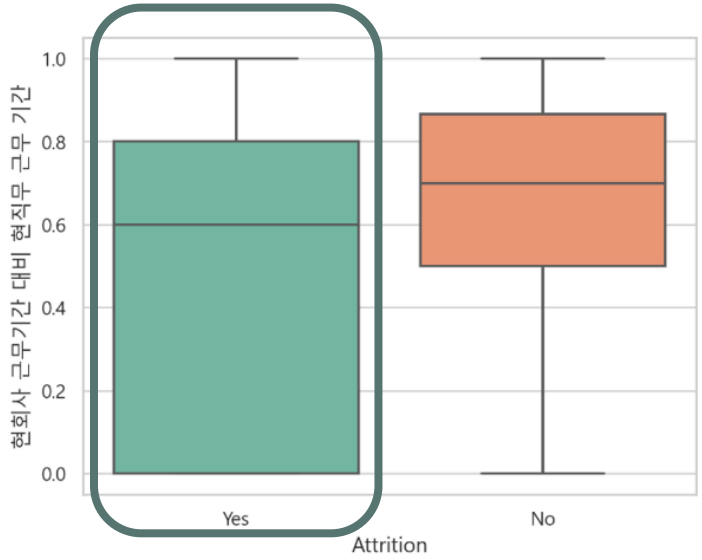
1. EDA

02) 퇴직자가 재직자보다 근속연수가 짧은 이유



df_2030['현회사 근무기간 대비 현직무 근무 기간']

$$= \frac{\text{df\_2030['YearsInCurrentRole']}}{\text{df\_2030['YearsAtCompany']}}$$



직무 변동에 적응을 못한 것이 원인?

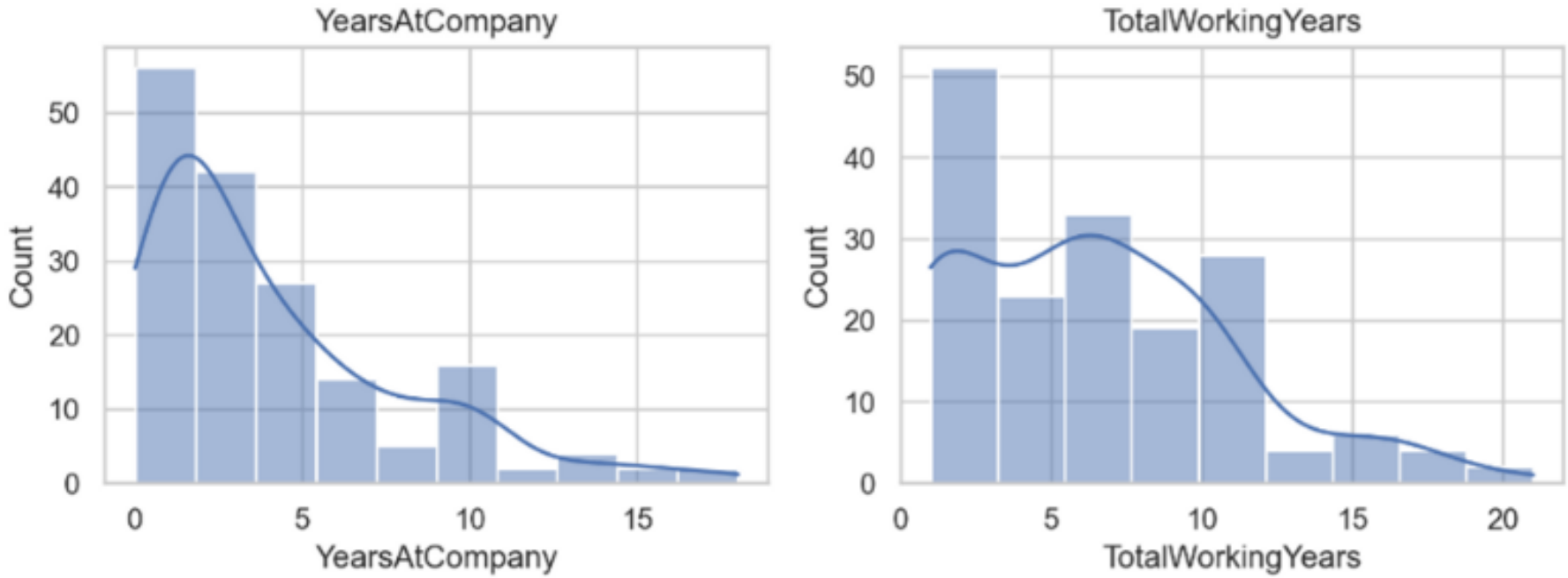
📌 **확인결과,**

최근에 직무 이동을 하고 곧바로 퇴사한 것으로 추정됨.

직무 이동 후 적응을 돕는 조치 필요

1. EDA

02) 퇴직자가 재직자보다 근속연수가 짧은 이유



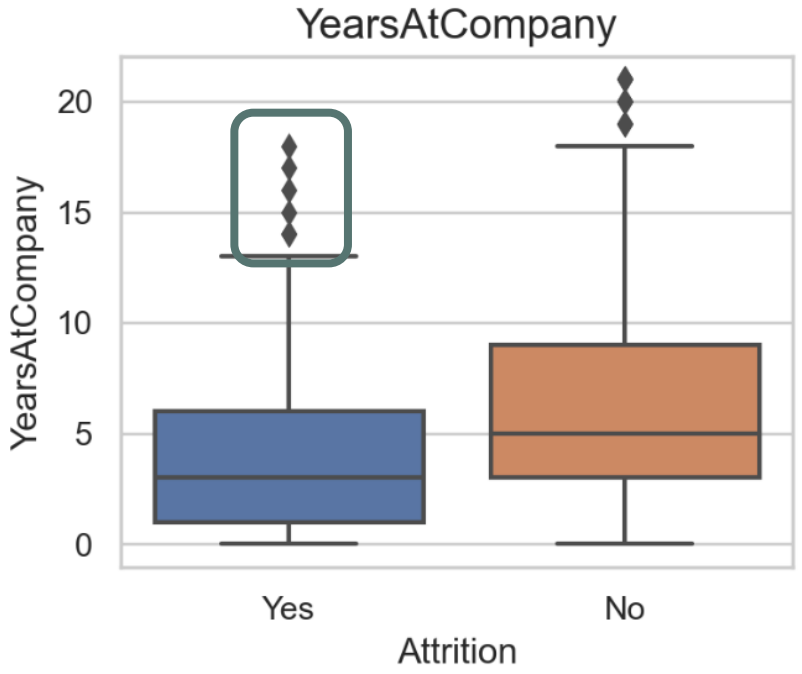
📌 2030 퇴직자의 TotalWorkingYears는 완만하게 분포된 것에 비해 YearsAtCompany는 왼쪽으로 치우쳐져 있음

📌 무경력 신입 뿐만 아니라 경력자의 조기 퇴사 문제도 주목해야 함

📌 경력직 2030이 전문적인 커리어를 쌓을 수 있도록 개선 필요

1. EDA

02) 퇴직자가 재직자보다 근속연수가 짧은 이유 - (이상치: 근속연수가 오래되었음에도 퇴사를 한 이유)



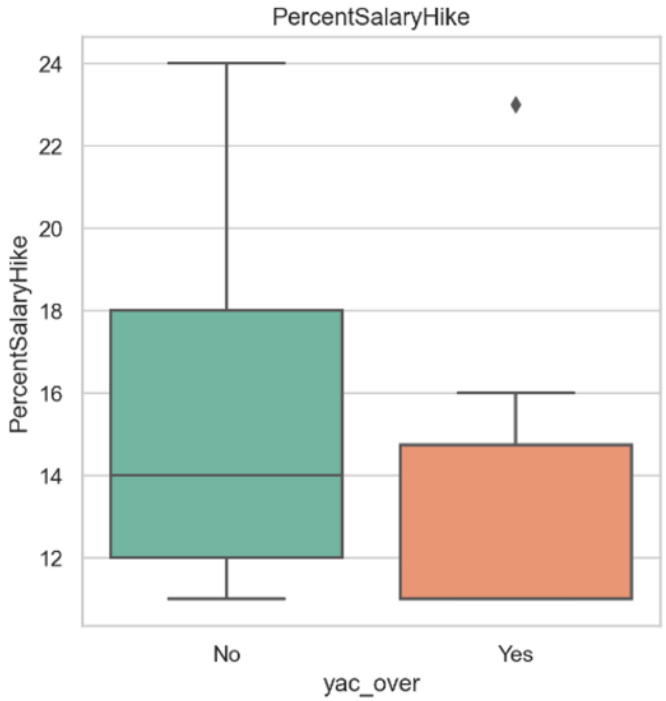

```
# 2030 연령대 & 퇴사자 & 근속연수 13.5년 이상인 직원들의 첫 입사 나이 추정
df_2030_yac_over['Age'] - df_2030_yac_over['YearsAtCompany']
```

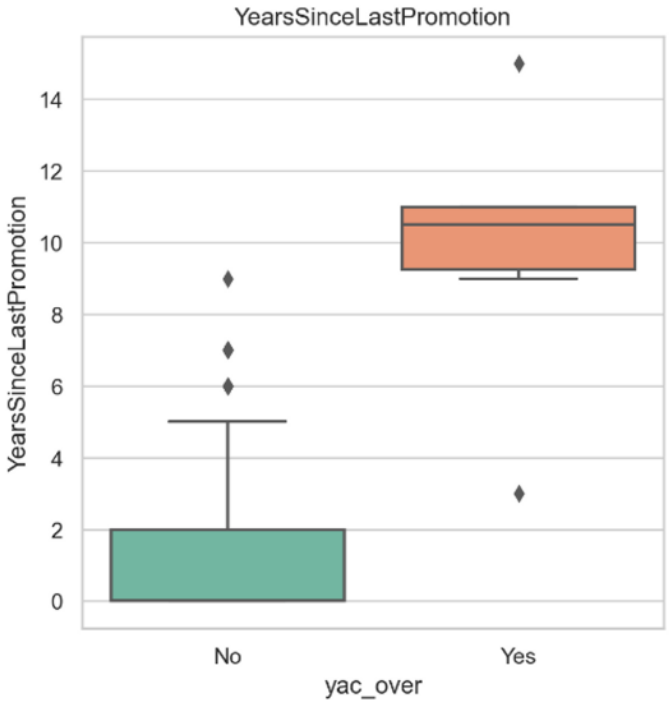
```
EmployeeNumber
291      18
967      20
970      23
1042     19
1127     21
1489     19
dtype: int64
```

📌 분석 타겟 직원들은 10대 후반, 20대 초반에 입사해 이 회사에 10년 넘게 다닌 사람들

1. EDA

02) 퇴직자가 재직자보다 근속연수가 짧은 이유 - (이상치: 근속연수가 오래되었음에도 퇴사를 한 이유)





📌 근속연수가 낮은 사람들보다 높은 사람들이 작년대비 연봉상승률이 낮았음

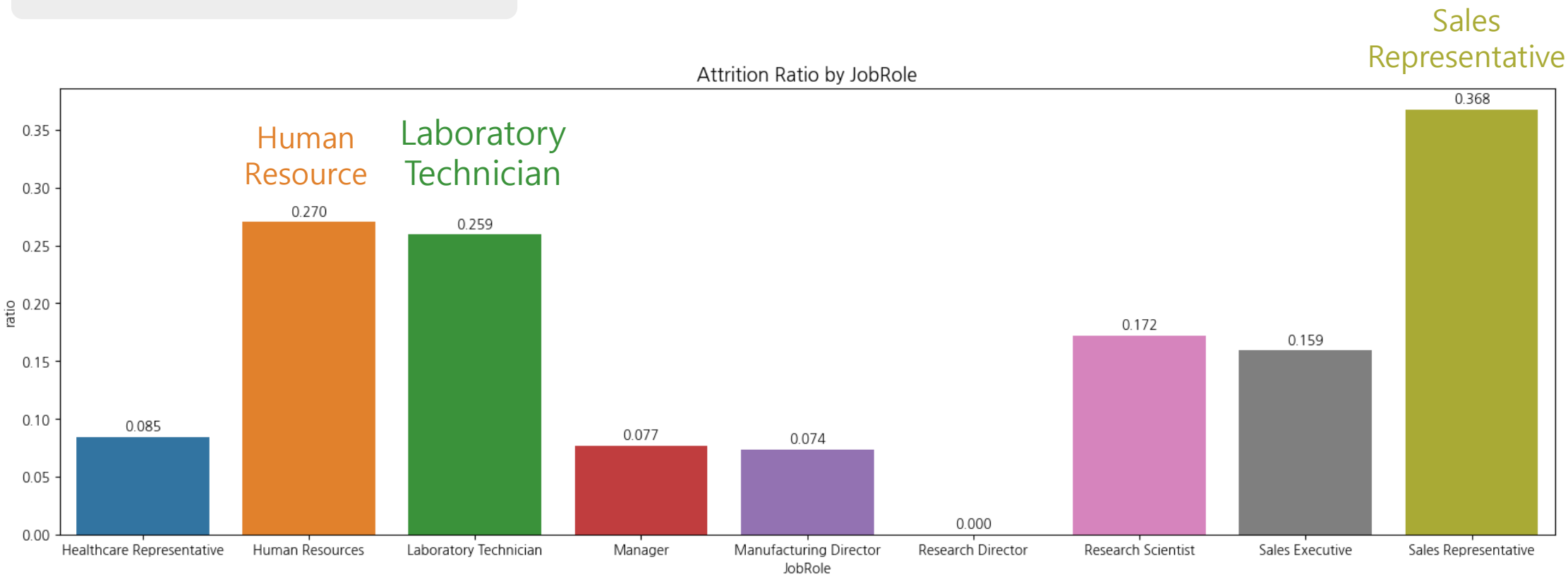
→ 연봉상승률에 불만족해서 퇴사?

📌 근속연수가 낮은 사람들보다 높은 사람들이 승진을 한지 오래됨

→ 승진을 못해서 퇴사?

1. EDA

3) 퇴사율 Top3 부서의 퇴사 이유

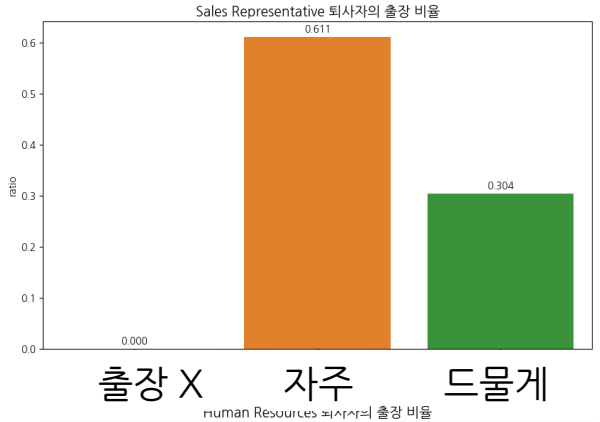
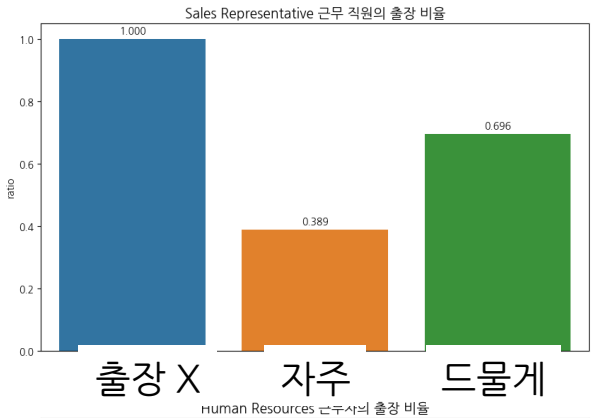
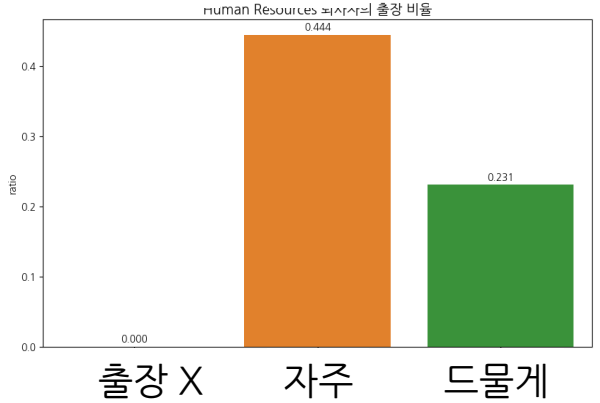
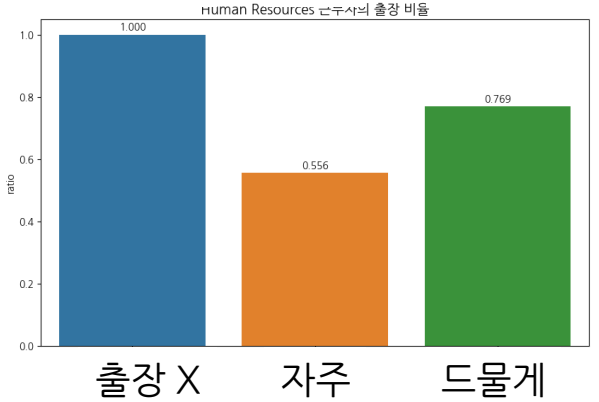
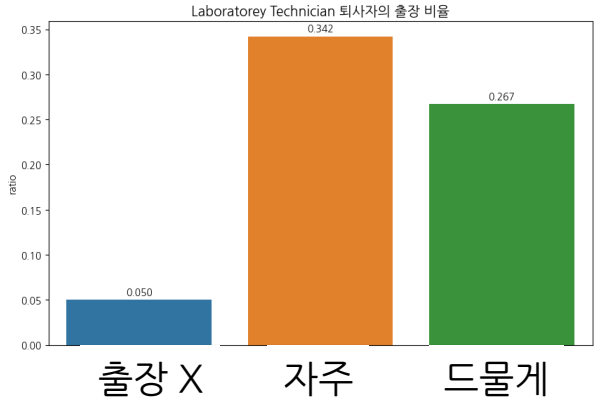
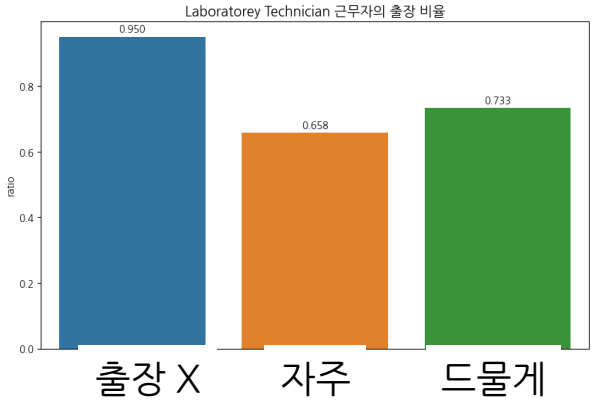


📌 Sales Representative, Human Resources, Laboratory Technician 📌

1. EDA

1. 잦은 출장

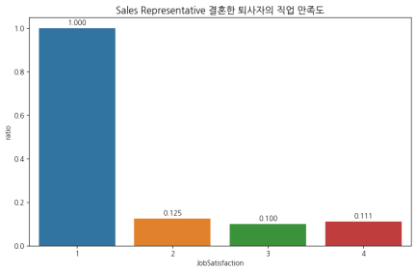
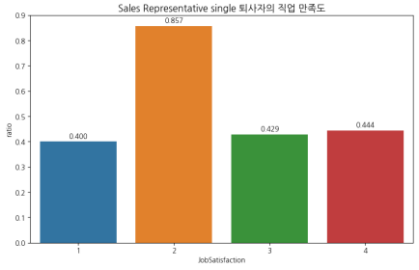
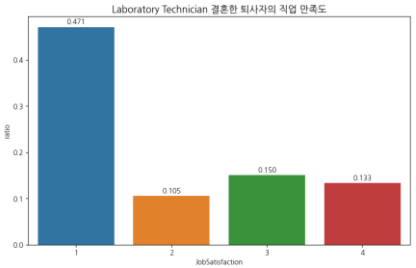
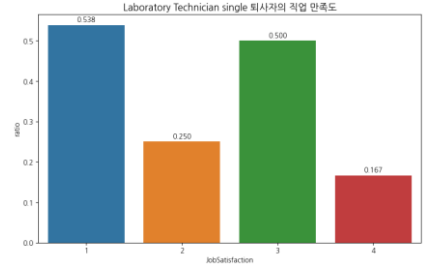
📌 퇴사자들의 출장비율을 볼 때,
퇴사율 Top3 직무 모두
Travel Frequently 비율이 높음.

	퇴사자	재직자
Sales Representative		
Human Resource		
Laboratory Technician		

1. EDA

2. 결혼여부와 직무만족도

1점 2점 3점 4점

	기혼 퇴사자 직업 만족도	미혼 퇴사자 직업 만족도
Sales Representative		
Human Resources		HR 부서는 미혼 퇴사자 없음
Laboratory Technician		

2. 주요 피처 선정

통계 분석 : attrition과 상관관계가 강한 변수를 찾으려 함

✔ 수치형 변수와 attrition : t-test(attrition이 이진 변수이므로) 검정

✔ 범주형 변수와 attrition : 언더샘플링/오버샘플링 후 카이제곱 검정

PCA도 추가로 진행

	t-test (df_raw)	t-test (df_under)	t-test (df_over)	PCA (df_raw)
p-value 0.05 이하 유의미한 변수	Age, MonthlyIncome, TotalWorkingYears, TrainingTimesLastYear, YearsAtCompany, YesrsInCurrentRole, YesrswithCurrentManager	Age, DistanceFromHome, MonthlyIncome, TotalWorkingYears, TrainingTimesLastYear, YearsAtCompany, YearsInCurrentRole, YearsWithCurrManager	Age DistanceFromHome MonthlyIncome PercentSalaryHike TotalWorkingYears TrainingTimesLastYear YearsAtCompany YearsInCurrentRole YearsSinceLastPromotion YearsWithCurrManager	PC1 : TotalWorkingYears (0.396864), YearsAtCompany, JobLevel, MonthlyIncome, YearsInCurrentRole, YearsWithCurrManager, Age, YearsSinceLastP PC2: PercentSalaryHike, PerformanceRating, NumCompaniesWorked

	chi-square (df_raw)	chi-square(df_under)	chi-square (df_over)	PCA (df_raw)
p-value 0.05 이하 유의미한 변수	EnvironmentSatisfaction JobInvolvement Joblevel JobSatisfaction StockOptionLevel WorkLifeBalance BusinessTravel Department EducationField JobRole MaritalStatus Overtime	EnvironmentSatisfaction JobInvolvement Joblevel JobSatisfaction StockOptionLevel WorkLifeBalance BusinessTravel EducationField JobRole MaritalStatus Overtime	Education EnvironmentSatisfaction JobInvolvement Joblevel JobSatisfaction PerformanceRating RelationshipSatisfaction StockOptionalLevel WorkLifeBalance BusinessTravel Department EducationField Gender JobRole MaritalStatus OverTime	PerformanceRating

3. 머신러닝

퇴사자 예측 ML 모델 만들기

데이터 전처리

파생변수 생성 & 컬럼 제거

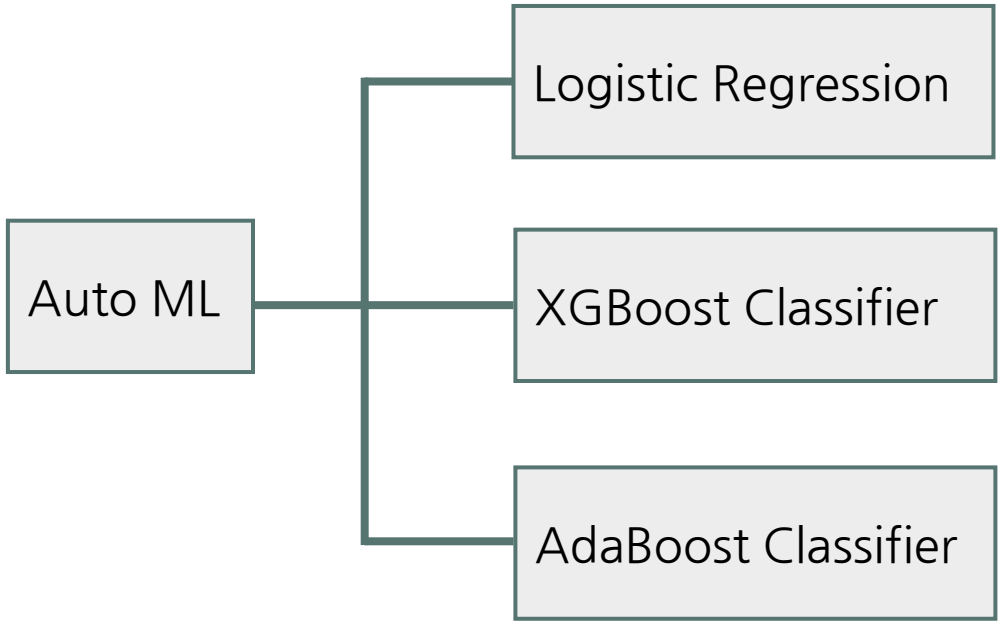
standard-scaling(수치형 피처)

one-hot-encoding(범주형 피처)

데이터 병합 및 train-test set 분리

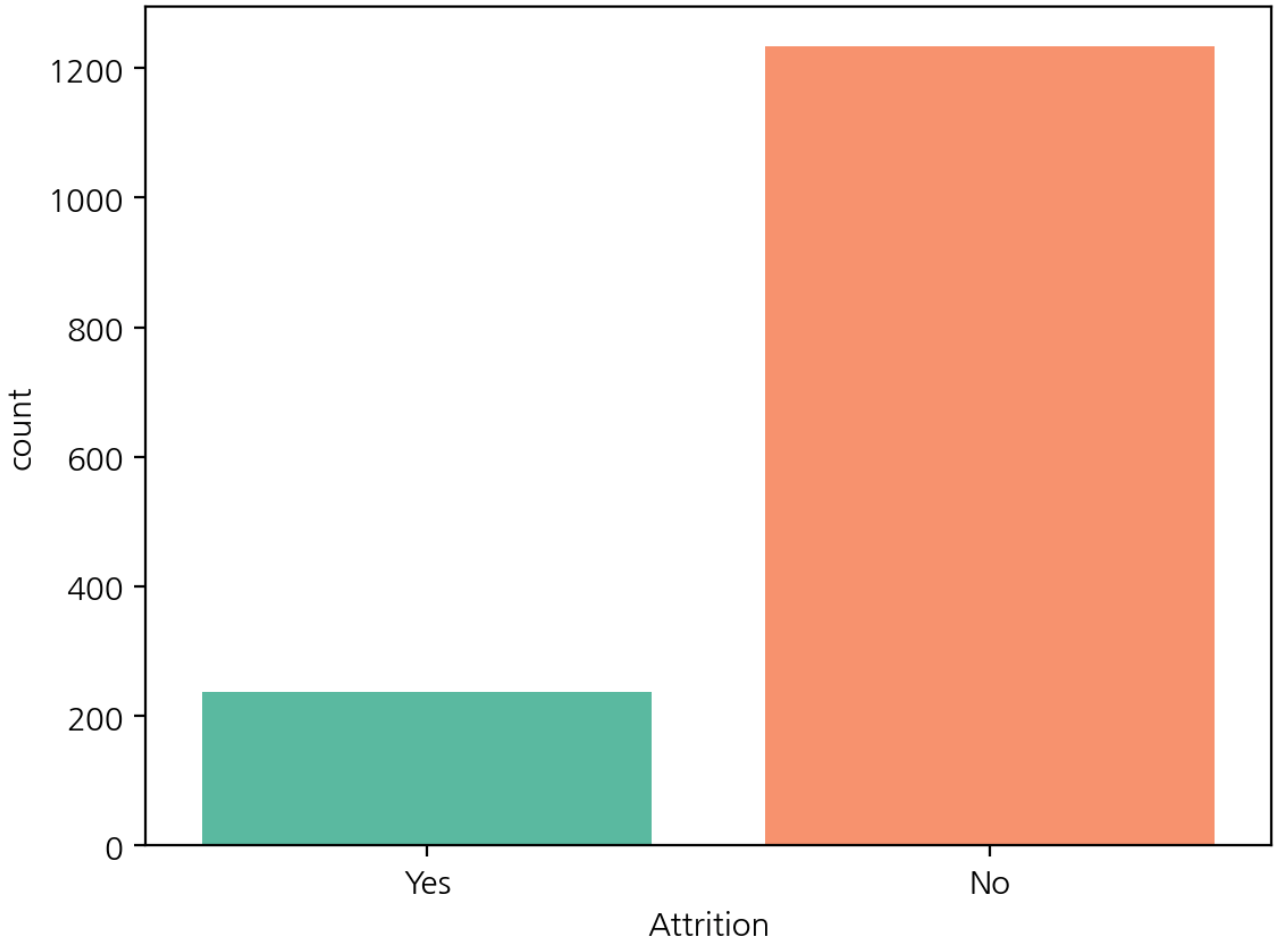
ADASYN 오버샘플링(train set)

모델링



3. 머신러닝

데이터 문제: 데이터 불균형



3. 머신러닝

모델 선정 & 평가지표 선정

SMOTE 적용 결과
정확도 : 0.69, 정밀도 : 0.23, 재현율 : 0.56
f1-score : 0.32, auc : 0.63

	precision	recall	f1-score	support
0	0.91	0.71	0.80	255
1	0.23	0.56	0.32	39
accuracy			0.69	294
macro avg	0.57	0.63	0.56	294
weighted avg	0.82	0.69	0.73	294

ADASYN 적용 결과
정확도 : 0.70, 정밀도 : 0.25, 재현율 : 0.62
f1-score : 0.36, auc : 0.67

	precision	recall	f1-score	support
0	0.92	0.72	0.81	255
1	0.25	0.62	0.36	39
accuracy			0.70	294
macro avg	0.59	0.67	0.58	294
weighted avg	0.83	0.70	0.75	294

오버샘플링 기법

ADASYN(Adaptive Synthetic Sampling)

- SMOTE의 업그레이드 버전
- ADASYN은 소수 클래스 데이터 포인트의 밀도를 기준으로 가상 데이터 포인트를 생성하기 때문에, SMOTE보다 데이터 분포의 차이를 보다 잘 보완 가능
- 따라서, 더 자연스러운 데이터 분포를 생성하고 모델이 학습하는 특성을 보다 실제 데이터에 가깝게 만들어 줌

성능 확인

- 기본 피처(df_raw) + 로지스틱 기본 모델 사용
- 성능 확인 결과, ADASYN에서 성능이 조금 향상됨을 볼 수 있음

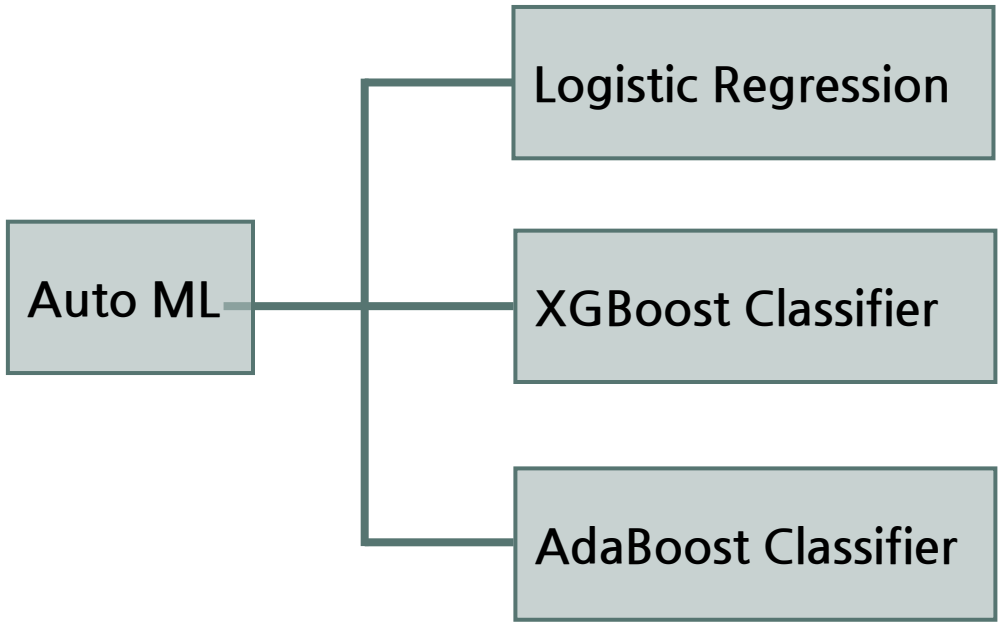
3. 머신러닝

퇴사자 예측 ML 모델 만들기

데이터 전처리

- 파생변수 생성 & 컬럼 제거
- standard-scaling(수치형 피처)
- one-hot-encoding(범주형 피처)
- 데이터 병합 및 train-test set 분리
- ADASYN 오버샘플링(train set)

모델링



3. 머신러닝

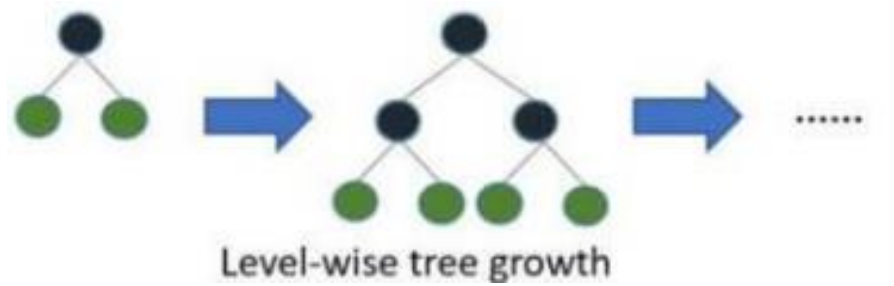
모델별 성능 비교

model	Over Sampling	
	Accuracy	f1_score
Logistic Regression	0.76	0.40
XGBoost Classifier	0.89	0.48
AdaBoost Classifier	0.85	0.47
Accuracy와 f1_score를 성능 지표로 사용한 이유 : 분류 문제 + 불균형 데이터		

3. 머신러닝

최종 선정 모델

XGBoost



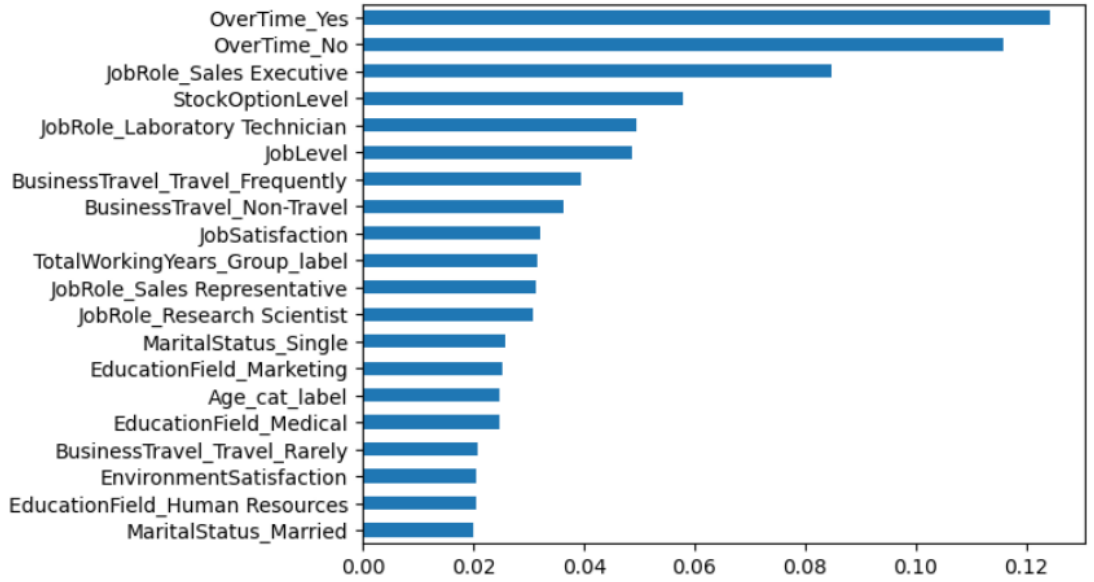
https://images.app.goo.gl/eyQd81GW2FdCWybE7

Hyper Parameter	
n_estimators	500
Learning_rate	0.01
Max_depth	8
colsample_bytree	0.9
reg_alpha	0.01
random_state	42

3. 머신러닝

데이터분석 ML 최종 모델

Final_ model	Accuracy	f1_score
XGBoost Classifier	0.88	0.55
데이터 불균형으로 인해 퇴사자에 대한 특성 학습 부족		



📌 최종적으로 선택된 모델의 중요변수는

OverTime, JobRole, StockOption, JobLevel, BusinessTravel 이다.

추가적인 HR Analytics 방법

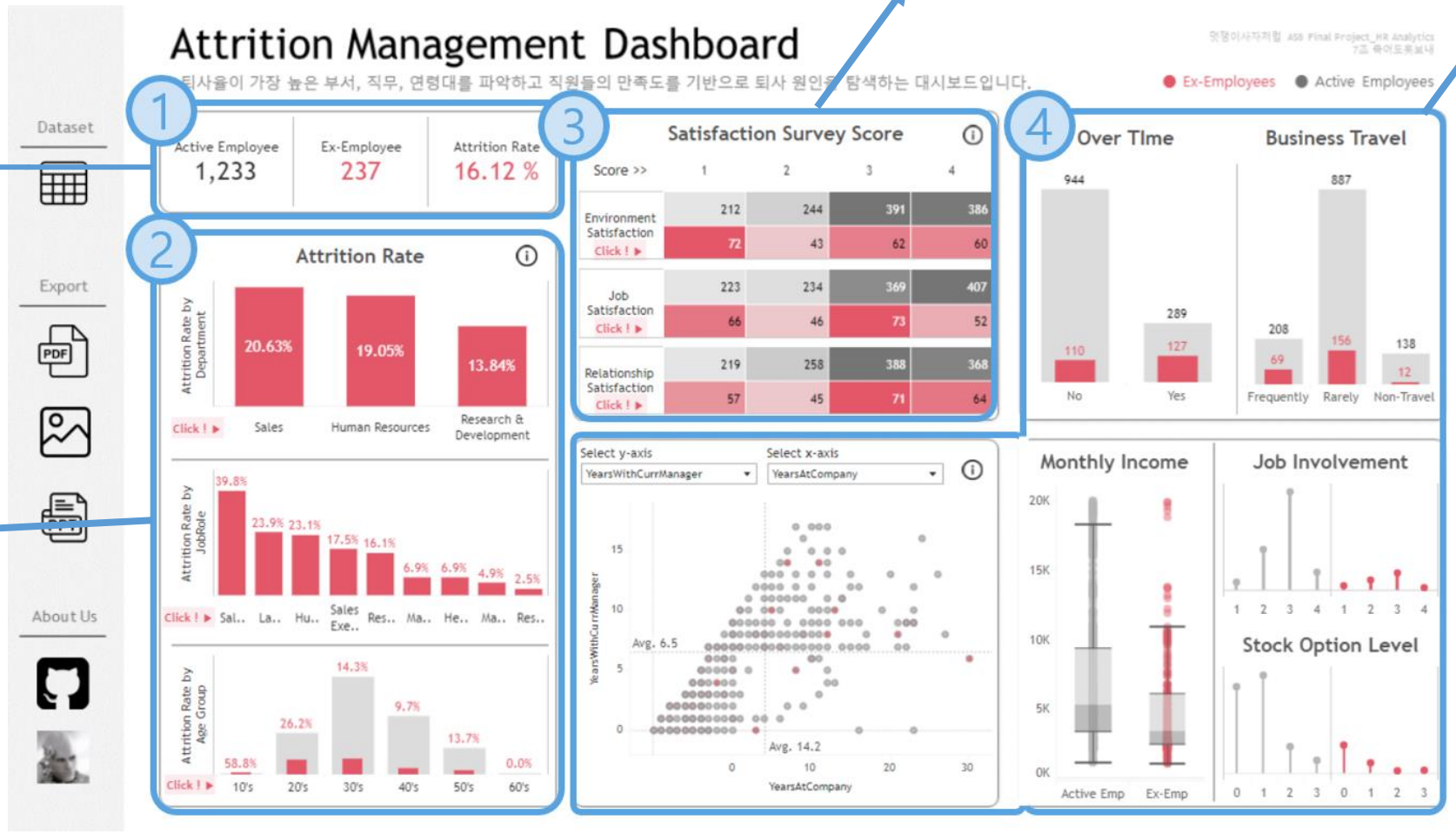
시각화 대시보드

만족도 점수

퇴사 주요 원인 변수들

주요 KPI

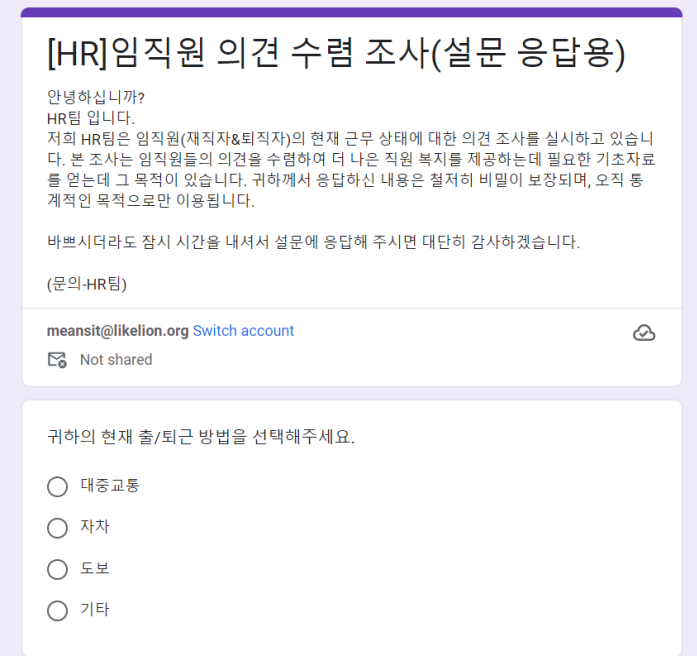
부서별, 직무별,
연령대별 퇴사율



https://public.tableau.com/app/profile/seyeon.cho/viz/Final_HR_Analytics_0505/DashBoard

추가적인 HR Analytics 방법

설문지



[HR]임직원 의견 수렴 조사(설문 응답용)

안녕하십니까?
HR팀 입니다.
저희 HR팀은 임직원(재직자&퇴직자)의 현재 근무 상태에 대한 의견 조사를 실시하고 있습니다. 본 조사는 임직원들의 의견을 수렴하여 더 나은 직원 복지를 제공하는데 필요한 기초자료를 얻는데 그 목적이 있습니다. 귀하께서 응답하신 내용은 철저히 비밀이 보장되며, 오직 통계적인 목적으로만 이용됩니다.

바쁘시더라도 잠시 시간을 내셔서 설문에 응답해 주시면 대단히 감사하겠습니다.

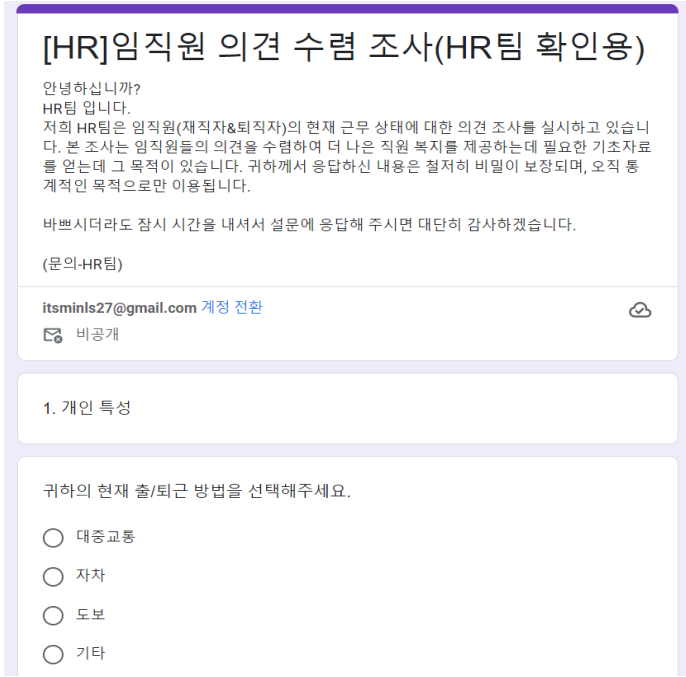
(문의-HR팀)

meansit@likelion.org Switch account

Not shared

귀하의 현재 출/퇴근 방법을 선택해주세요.

- 대중교통
- 자차
- 도보
- 기타



[HR]임직원 의견 수렴 조사(HR팀 확인용)

안녕하십니까?
HR팀 입니다.
저희 HR팀은 임직원(재직자&퇴직자)의 현재 근무 상태에 대한 의견 조사를 실시하고 있습니다. 본 조사는 임직원들의 의견을 수렴하여 더 나은 직원 복지를 제공하는데 필요한 기초자료를 얻는데 그 목적이 있습니다. 귀하께서 응답하신 내용은 철저히 비밀이 보장되며, 오직 통계적인 목적으로만 이용됩니다.

바쁘시더라도 잠시 시간을 내셔서 설문에 응답해 주시면 대단히 감사하겠습니다.

(문의-HR팀)

itsminls27@gmail.com 계정 전환

비공개

1. 개인 특성

귀하의 현재 출/퇴근 방법을 선택해주세요.

- 대중교통
- 자차
- 도보
- 기타

https://docs.google.com/forms/d/e/1FAIpQLSclUBOkG7sNrp7EBjQRe6S4A9YeK0lknzuhUDwhiVAv-2SsJQ/viewform

https://docs.google.com/forms/d/e/1FAIpQLSfiy6s-XkdqybmeIjLIXF5OUzxfir31tkYqMa0J9SeTqAvjyw/viewform

(예시) 조사방법

- 주기 : 분기별 1회 (연 4회)
- 일시 : 목요일 11시 or 14시
- 데이터 수집 방법 : 사내 메신저& 사내 메일

05 활용 방안

프로젝트 활용방안 제시

📌 HR Analytics의 ‘HR 데이터 수집 - 데이터 분석 - 퇴사여부 예측’ 프로세스를 제공

✔ HR Analytics에 특화된 설문지 항목과 최적의 설문 방법으로 양질의 HR 데이터를 수집 가능

✔ 탐색적 데이터 분석(EDA)과 태블로 대시보드를 활용하여 사내 직원들의 니즈와 컴플레인을 파악 및 개선

✔ 머신러닝 모델을 활용해 퇴사 여부를 예측하여 인재 유출을 방지 가능

감사합니다

멋쟁이 사자처럼 AI SCHOOL 8기

7조 죽어도 못 보내

김조은, 임승민, 조세연, 차은서