



# SK쉴더스 2차 프로젝트

팀명 : 오케스트라  
팀원 : 박진호, 홍민택, 김성언, 장혜림

# 목차

## Episode 1 - Regression

1. EDA
2. Feature Engineering
3. Modeling

## Episode 3 - Classification

1. EDA
2. Feature Engineering
3. Modeling

## 결론

최종 SCORE  
정리

# Episode 1 - Regression













Playground Series - Season 3, Episode 1

Tabular Regression with the California Housing Dataset

k

Kaggle

689 teams

60	▲ 66	Revanth		0.55526	12	1mo
61	▲ 5	Adeyinka Michael Sotunde		0.55528	10	1mo
62	▲ 6	离岛姬的北方酱		0.55528	3	1mo
63	▲ 6	Vitaliy Shpak		0.55528	10	1mo
64	▲ 6	Prabhanjan Jadhav		0.55528	10	1mo
65	▲ 7	kojimar		0.55528	10	1mo
66	▲ 5	Saumil Agrawal		0.55529	7	1mo
67	▼ 12	podsys		0.55530	6	1mo
68	▼ 34	Vinay Ethiraj		0.55530	4	1mo
69	▼ 34	Pietro Maldini		0.55531	6	1mo
70	▲ 25	Jan Niklas Ottow		0.55537	12	1mo
71	▼ 14	Suraj Dengale		0.55537	19	1mo

1차 목표 : 10% 이내



# Episode 1 - Data

	A	B	C	D	E	F	G	H	I	J	K
1	id	MedInc	HouseAge	AveRooms	AveBedrm	Population	AveOccup	Latitude	Longitude	MedHouseVal	
2	0	2.3859	15	3.82716	1.1121	1280	2.486989	34.6	-120.12	0.98	
3	1	3.7188	17	6.013373	1.054217	1504	3.813084	38.69	-121.22	0.946	
4	2	4.775	27	6.535604	1.103175	1061	2.464602	34.71	-120.45	1.576	
5	3	2.4138	16	3.350203	0.965432	1255	2.089286	32.66	-117.09	1.336	
6	4	3.75	52	4.284404	1.069246	1793	1.60479	37.8	-122.41	4.5	

```
1 train.shape, test.shape, submission.shape, original_df.shape
```

```
((37137, 9), (24759, 8), (24759, 2), (20640, 9))
```

- 원본 37137개 + 외부 20640개 총 57777개의 train data 사용
- 독립변수 8개
- 종속변수 1개

## Episode 1 - 평가지표

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2}$$

0.55530 ↓

목표

RMSE 지표는 크기가 0에 가까울수록 성능이 좋다고 판단

# EDA

- MedInc - 주택 블록 내 가구의 중위소득
  - HouseAge - 블록 내에 있는 주택의 나이
  - AveRooms - 블록 내의 평균 방 수
  - AveBedrms - 한 블록 내의 평균 침실 수
  - Population - 한 블록 내에 거주하는 총 인구 수
  - AveOccup - 평균 가구원 수
  - Longitude - 집이 서쪽으로 얼마나 떨어져 있는지를 나타내는 척도(낮을수록 서쪽)
  - Latitude - 집이 북쪽으로 얼마나 떨어져 있는지를 나타내는 척도(높을수록 북쪽)
  - MedHouseValue - 블록 내 가구의 중위 수 주택 가격
- 독립변수 8개
  - 종속변수 1개

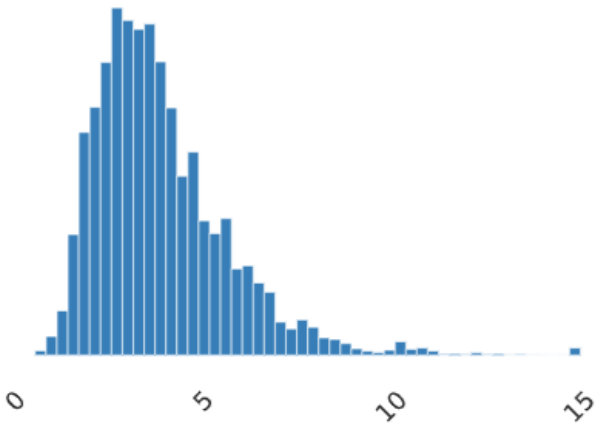
# EDA - MedInc

## MedInc

Real number ( $\mathbb{R}$ )

Distinct	12310
Distinct (%)	33.1%
Missing	0
Missing (%)	0.0%
Infinite	0
Infinite (%)	0.0%
Mean	3.8510294

Minimum	0.4999
Maximum	15.0001
Zeros	0
Zeros (%)	0.0%
Negative	0
Negative (%)	0.0%
Memory size	580.3 KiB



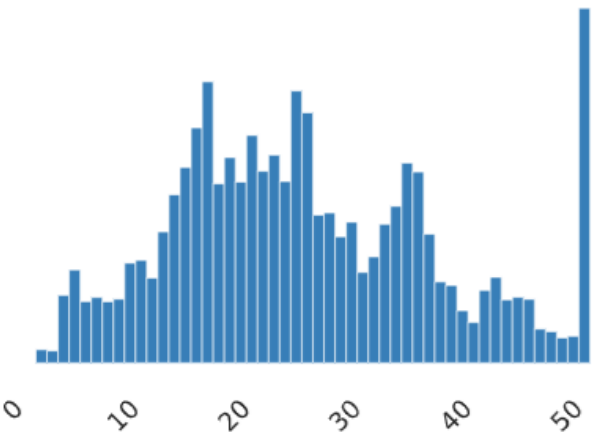
# EDA - HouseAge

## HouseAge

Real number ( $\mathbb{R}$ )

Distinct	51
Distinct (%)	0.1%
Missing	0
Missing (%)	0.0%
Infinite	0
Infinite (%)	0.0%
Mean	26.057005

Minimum	2
Maximum	52
Zeros	0
Zeros (%)	0.0%
Negative	0
Negative (%)	0.0%
Memory size	580.3 KiB





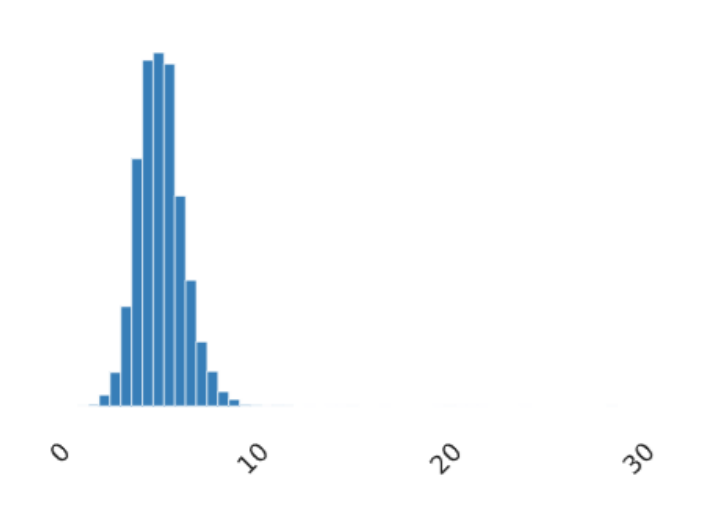
# EDA - AveRooms

## AveRooms

Real number ( $\mathbb{R}$ )

Distinct	22069
Distinct (%)	59.4%
Missing	0
Missing (%)	0.0%
Infinite	0
Infinite (%)	0.0%
Mean	5.1631237

Minimum	0.85106383
Maximum	28.837607
Zeros	0
Zeros (%)	0.0%
Negative	0
Negative (%)	0.0%
Memory size	580.3 KiB



# EDA - AveBedrms

## AveBedrms

Real number ( $\mathbb{R}$ )

Distinct	14066
Distinct (%)	37.9%
Missing	0
Missing (%)	0.0%
Infinite	0
Infinite (%)	0.0%
Mean	1.0622043

Minimum	0.5
Maximum	5.8731809
Zeros	0
Zeros (%)	0.0%
Negative	0
Negative (%)	0.0%
Memory size	580.3 KiB



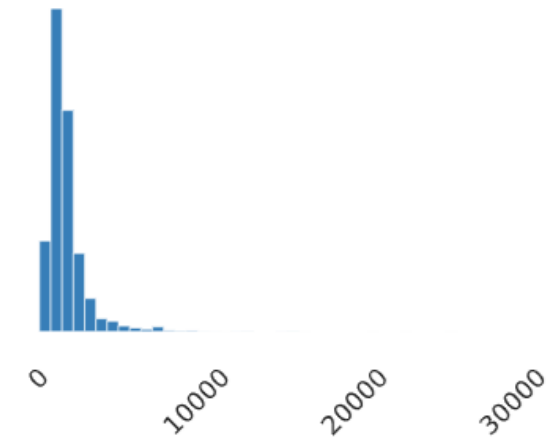
# EDA - Population

## Population

Real number ( $\mathbb{R}$ )

<b>Distinct</b>	3694
<b>Distinct (%)</b>	9.9%
<b>Missing</b>	0
<b>Missing (%)</b>	0.0%
<b>Infinite</b>	0
<b>Infinite (%)</b>	0.0%
<b>Mean</b>	1660.7789

<b>Minimum</b>	3
<b>Maximum</b>	35682
<b>Zeros</b>	0
<b>Zeros (%)</b>	0.0%
<b>Negative</b>	0
<b>Negative (%)</b>	0.0%
<b>Memory size</b>	580.3 KiB



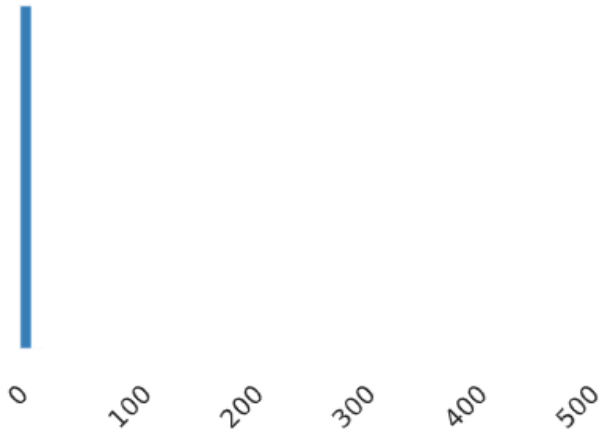
# EDA - AveOccup

## AveOccup

Real number ( $\mathbb{R}$ )

Distinct	21078
Distinct (%)	56.8%
Missing	0
Missing (%)	0.0%
Infinite	0
Infinite (%)	0.0%
Mean	2.831243

Minimum	0.95
Maximum	502.99061
Zeros	0
Zeros (%)	0.0%
Negative	0
Negative (%)	0.0%
Memory size	580.3 KiB



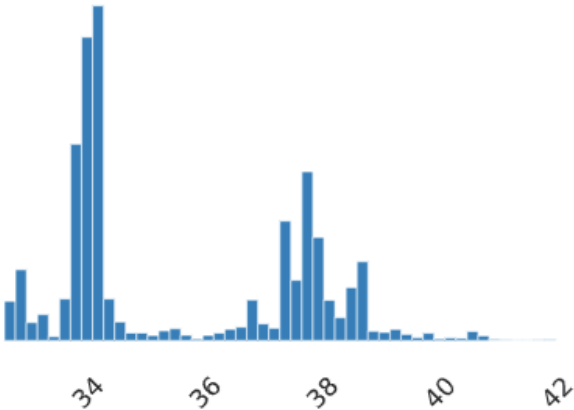
# EDA - Latitude

## Latitude

Real number ( $\mathbb{R}$ )

Distinct	791
Distinct (%)	2.1%
Missing	0
Missing (%)	0.0%
Infinite	0
Infinite (%)	0.0%
Mean	35.57003

Minimum	32.55
Maximum	41.95
Zeros	0
Zeros (%)	0.0%
Negative	0
Negative (%)	0.0%
Memory size	580.3 KiB



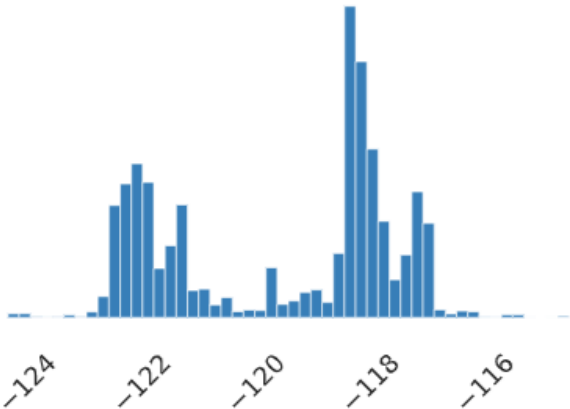
# EDA - Longitude

## Longitude

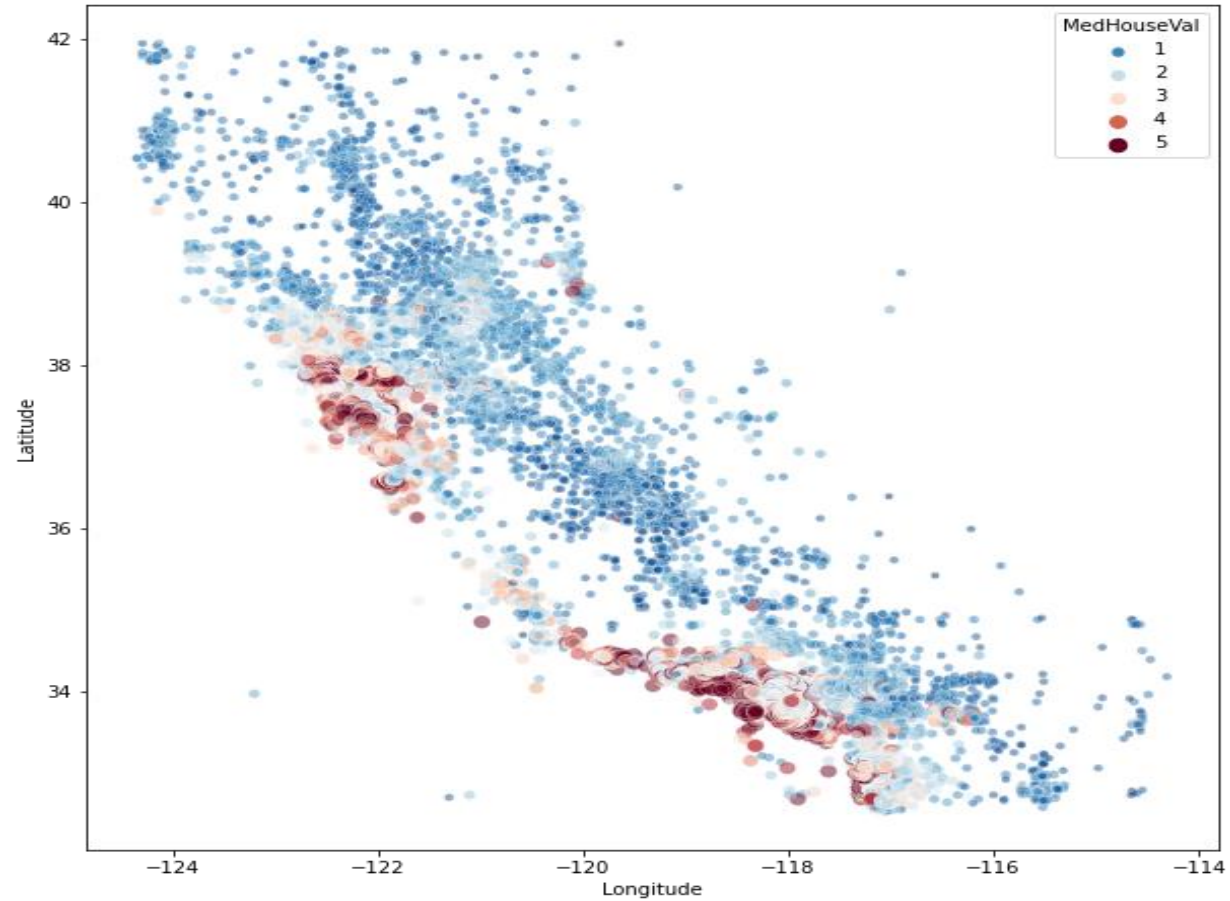
Real number ( $\mathbb{R}$ )

Distinct	755
Distinct (%)	2.0%
Missing	0
Missing (%)	0.0%
Infinite	0
Infinite (%)	0.0%
Mean	-119.55433

Minimum	-124.35
Maximum	-114.55
Zeros	0
Zeros (%)	0.0%
Negative	37137
Negative (%)	100.0%
Memory size	580.3 KiB



# EDA – Lat & Long



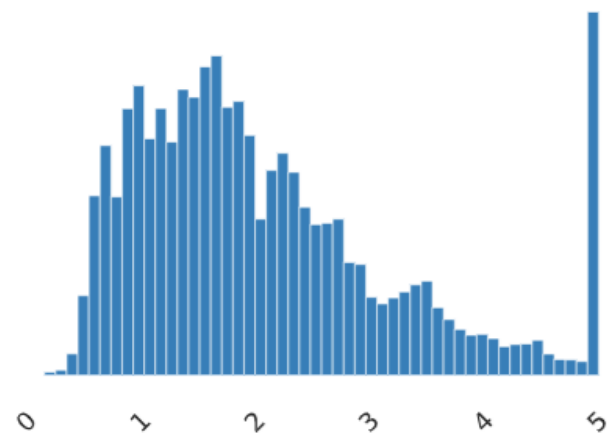
# EDA - 종속변수

MedHouseVal

Real number ( $\mathbb{R}$ )

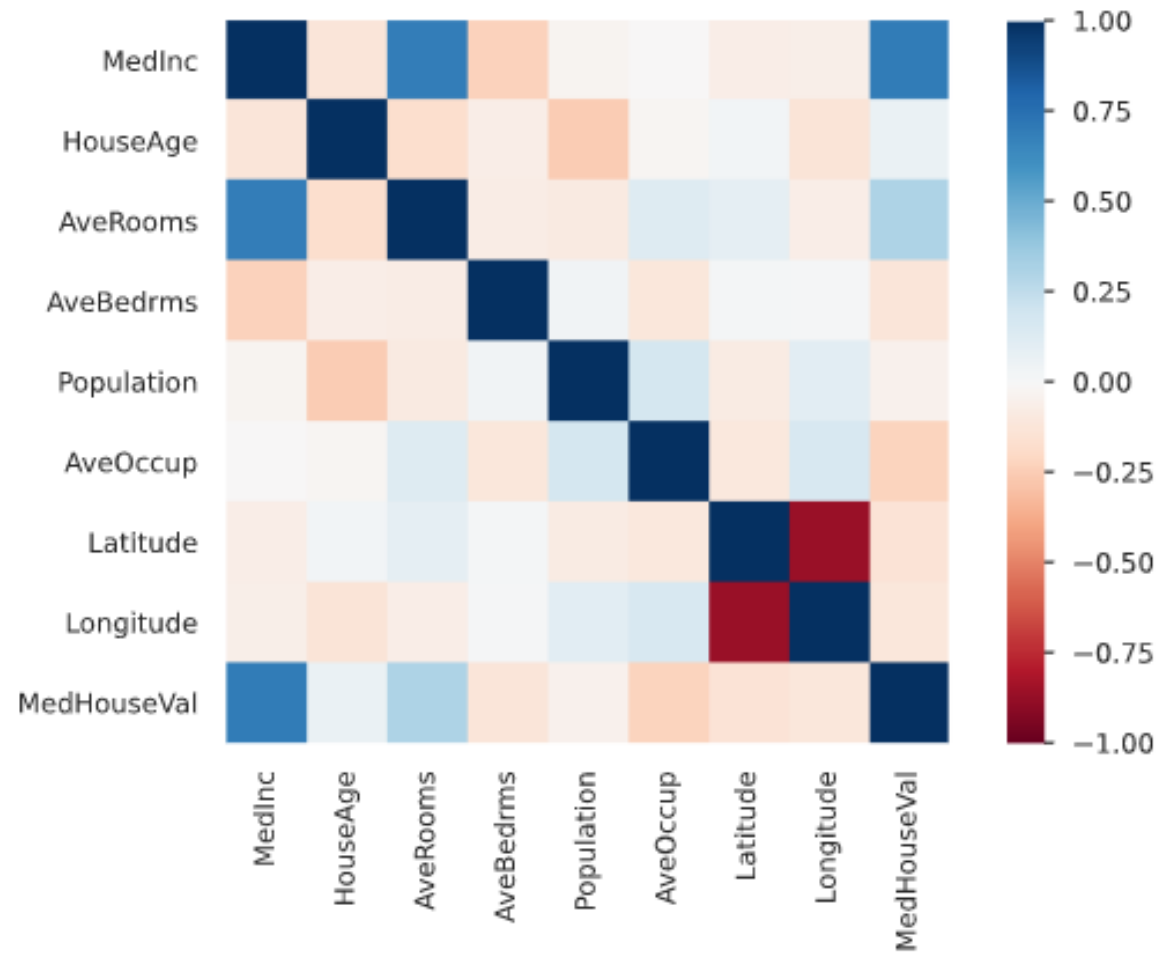
Distinct	3723
Distinct (%)	10.0%
Missing	0
Missing (%)	0.0%
Infinite	0
Infinite (%)	0.0%
Mean	2.0797513

Minimum	0.14999
Maximum	5.00001
Zeros	0
Zeros (%)	0.0%
Negative	0
Negative (%)	0.0%
Memory size	580.3 KiB





# EDA - Heatmap



# EDA - correlation

	MedInc	HouseAge	AveRooms	AveBedrms	Population	AveOccup	Latitude	Longitude	MedHouseVal
MedInc	1.000	-0.121	0.690	-0.227	-0.028	-0.007	-0.069	-0.057	0.697
HouseAge	-0.121	1.000	-0.170	-0.065	-0.255	-0.018	0.025	-0.129	0.063
AveRooms	0.690	-0.170	1.000	-0.076	-0.088	0.129	0.093	-0.066	0.310
AveBedrms	-0.227	-0.065	-0.076	1.000	0.038	-0.116	0.017	0.015	-0.121
Population	-0.028	-0.255	-0.088	0.038	1.000	0.177	-0.082	0.109	-0.042
AveOccup	-0.007	-0.018	0.129	-0.116	0.177	1.000	-0.107	0.161	-0.225
Latitude	-0.069	0.025	0.093	0.017	-0.082	-0.107	1.000	-0.865	-0.133
Longitude	-0.057	-0.129	-0.066	0.015	0.109	0.161	-0.865	1.000	-0.113
MedHouseVal	0.697	0.063	0.310	-0.121	-0.042	-0.225	-0.133	-0.113	1.000

# Feature Engineering

## Latitude & Longitude 파생 변수

1. 카운티
2. 주요 도시와의 거리
3. 해안선과의 거리
4. 클러스터링
5. PCA & UMAP
6. rotate 인코딩

## AveOccup 파생 변수

1. 가구원 한명당 방 개수
2. 주택밀집도

# Feature Engineering - AveOccup 파생변수

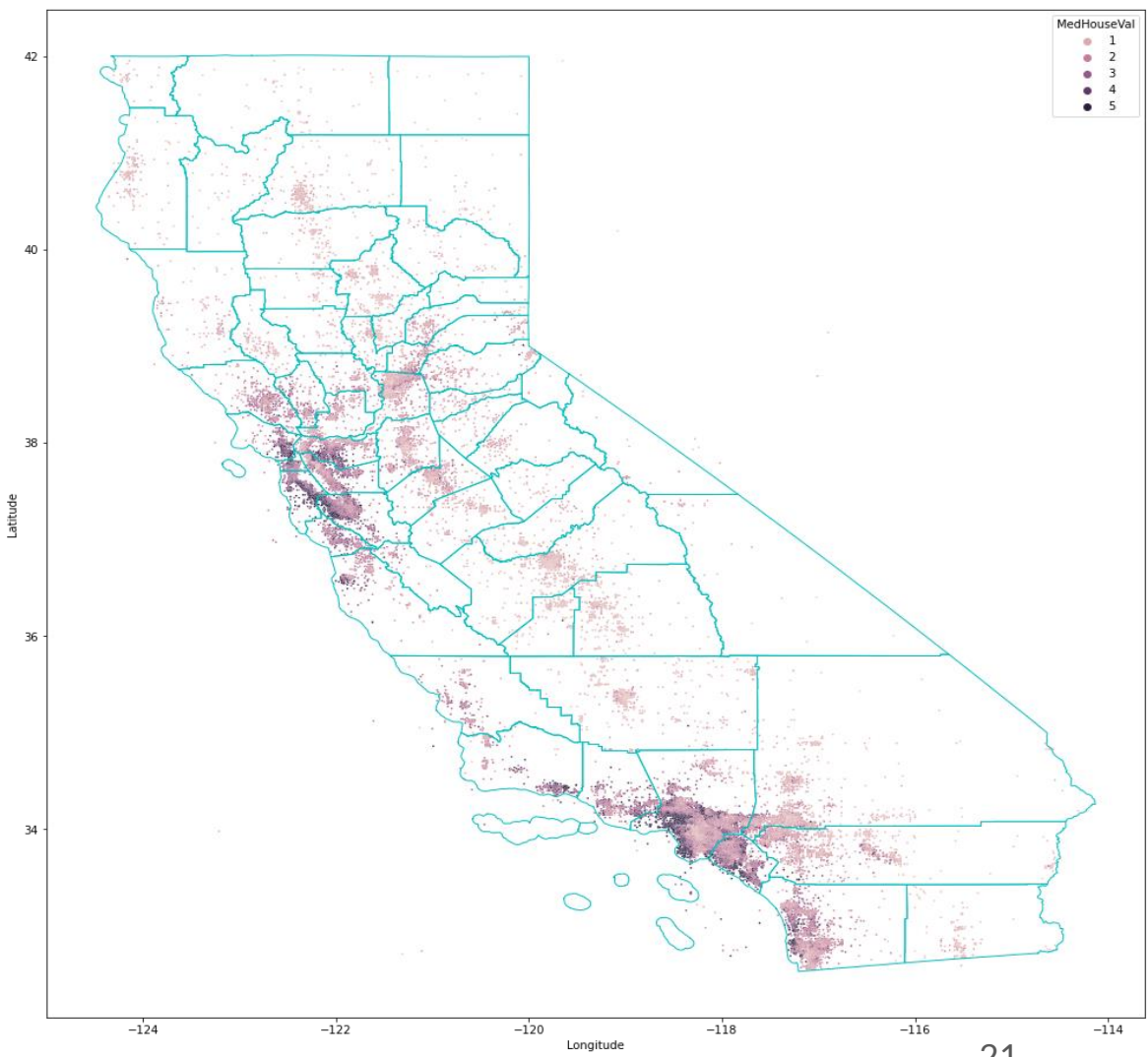
가구원 한명당 방 개수

$$\text{AveRooms\_per\_Occup} = \frac{\text{Averooms}}{\text{AveOccup}}$$

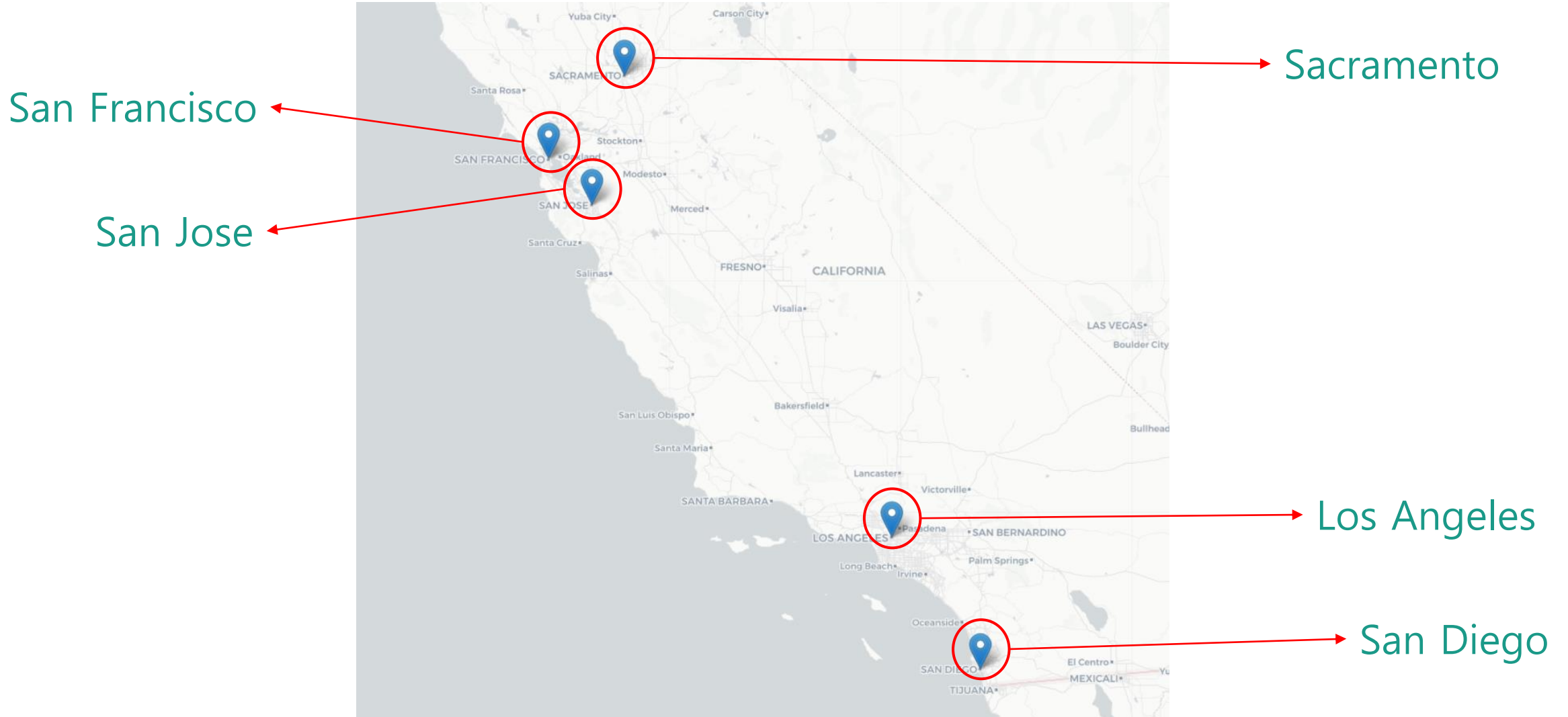
주택밀집도

$$\text{Density} = \frac{\text{Population}}{\text{AveOccup}}$$

# Feature Engineering - 카운티



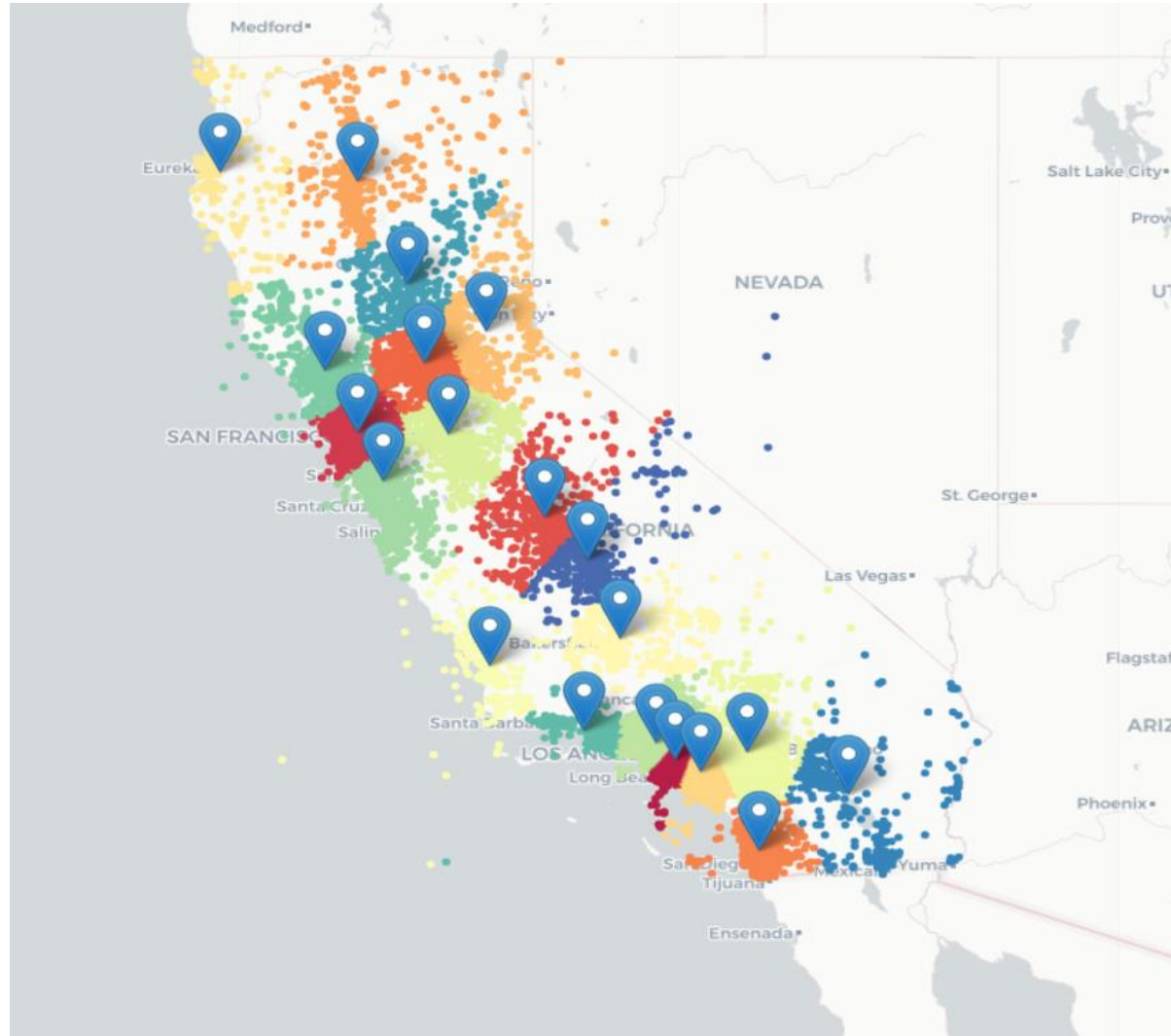
# Feature Engineering - 주요 도시와의 거리



# Feature Engineering - 해안선과의 거리

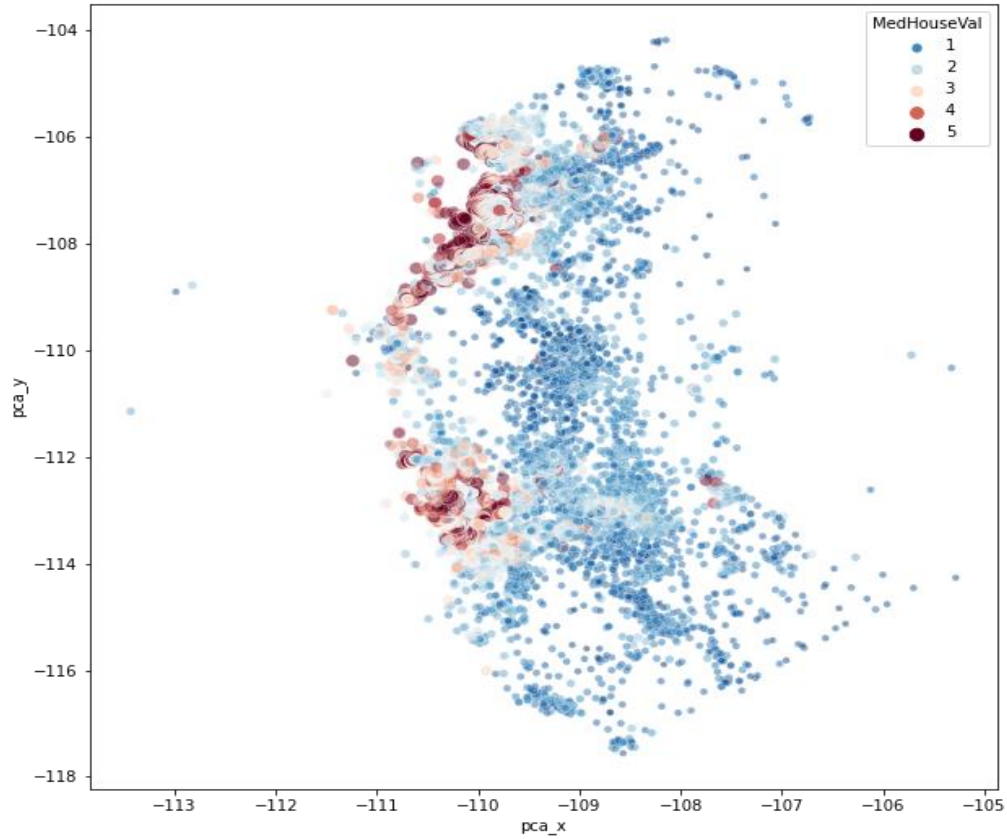


# Feature Engineering - 클러스터링

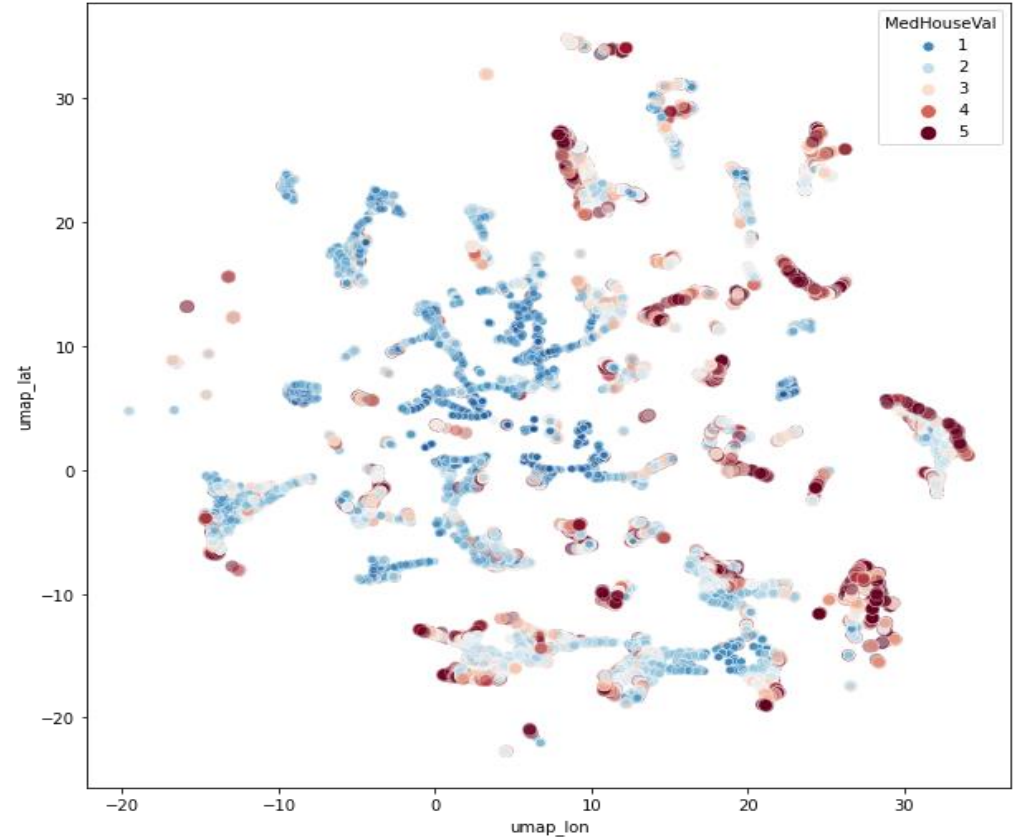




# Feature Engineering - PCA, UMAP

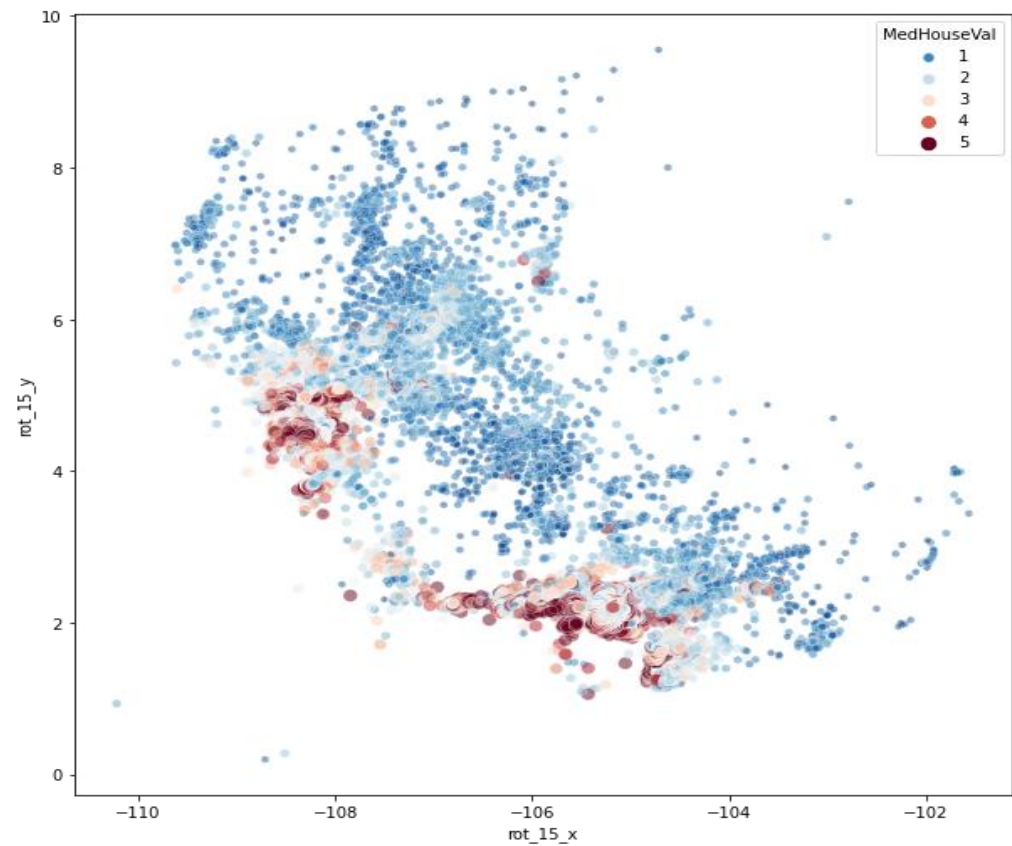


PCA

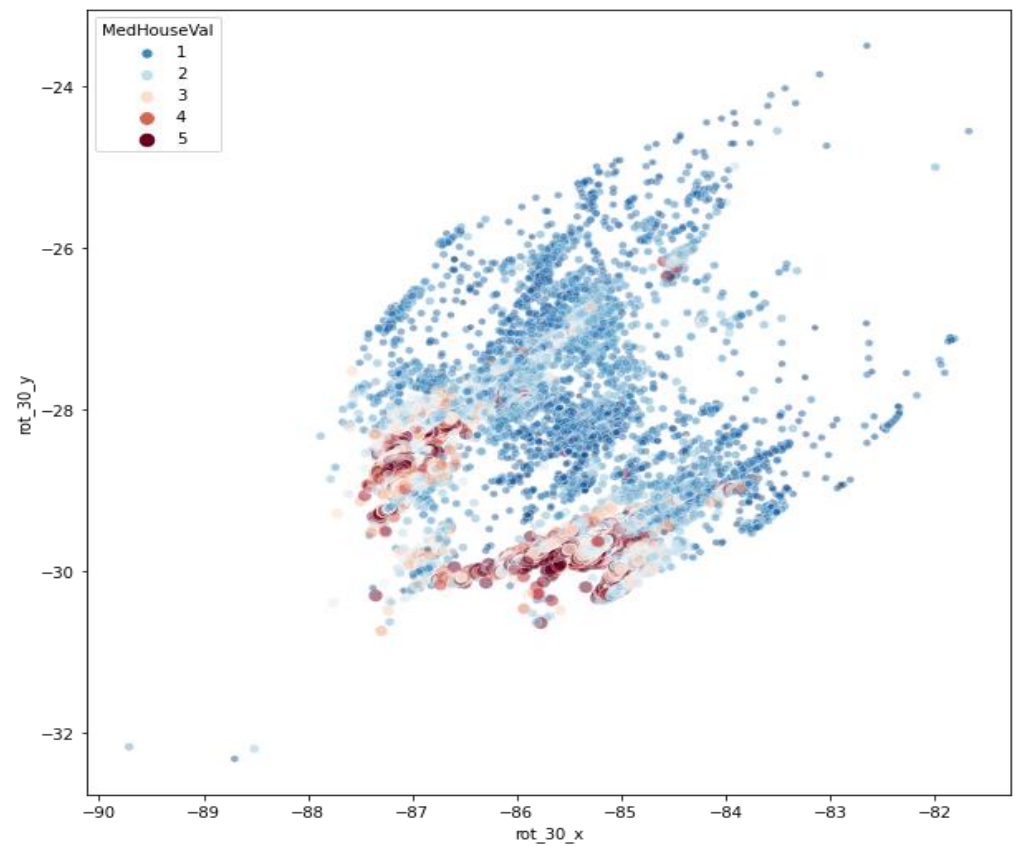


UMAP

# Feature Engineering - rotate



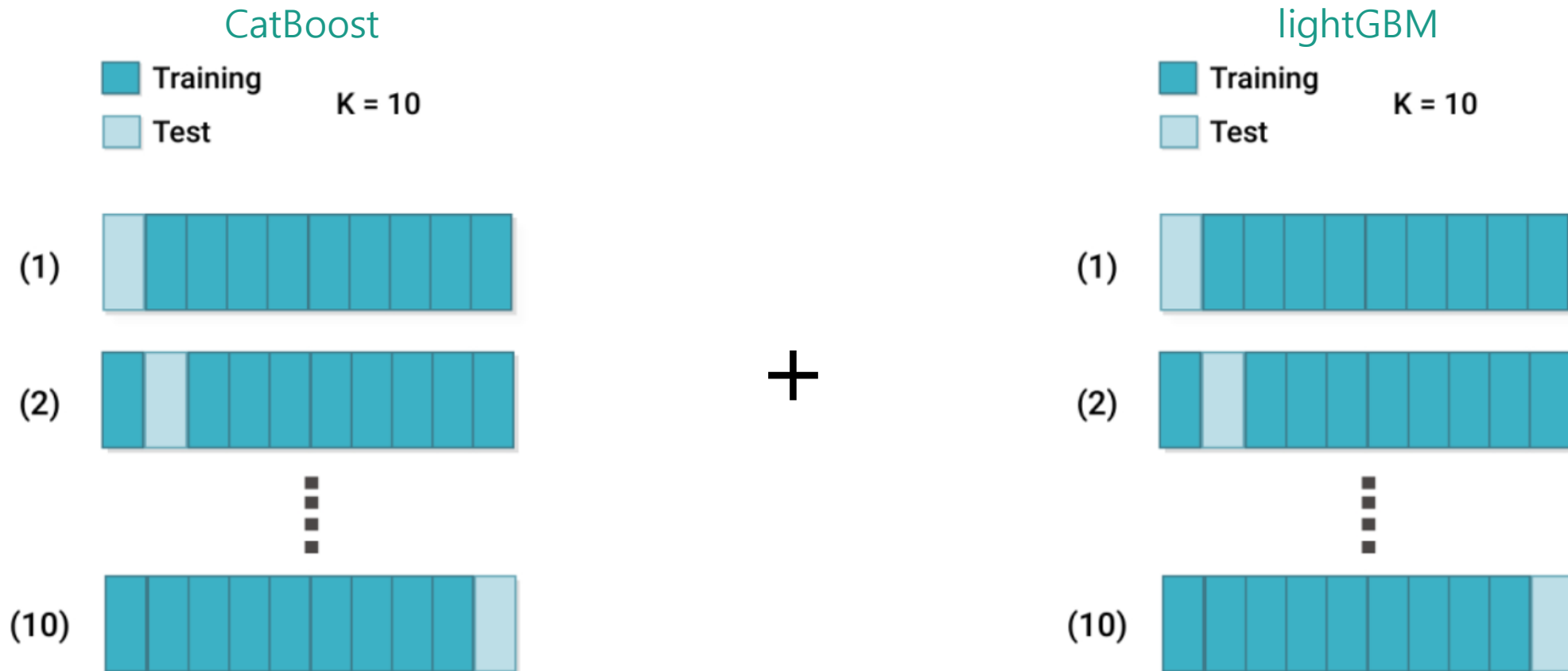
15°



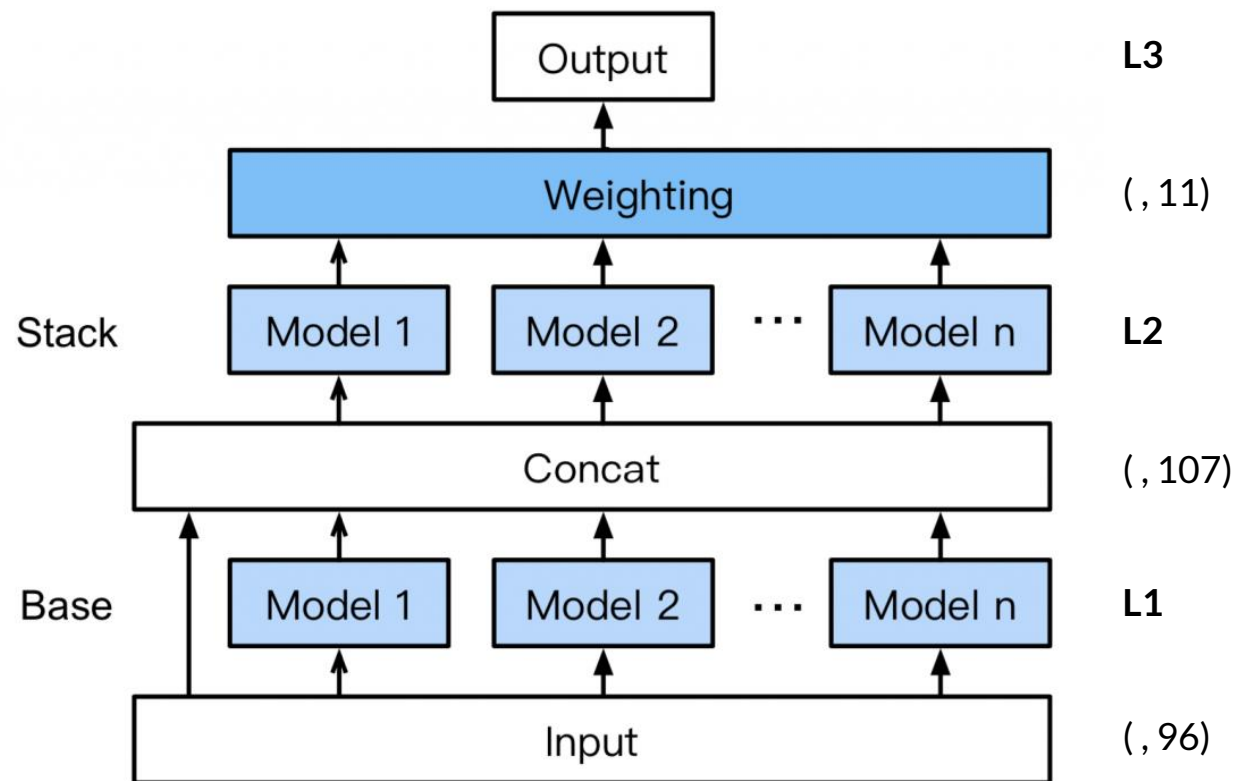
30°

# Modeling - Blending

블렌딩(  $\frac{1}{2}$ \*lightGBM의 예측값평균 +  $\frac{1}{2}$ \*CatBoost의 예측값평균 )



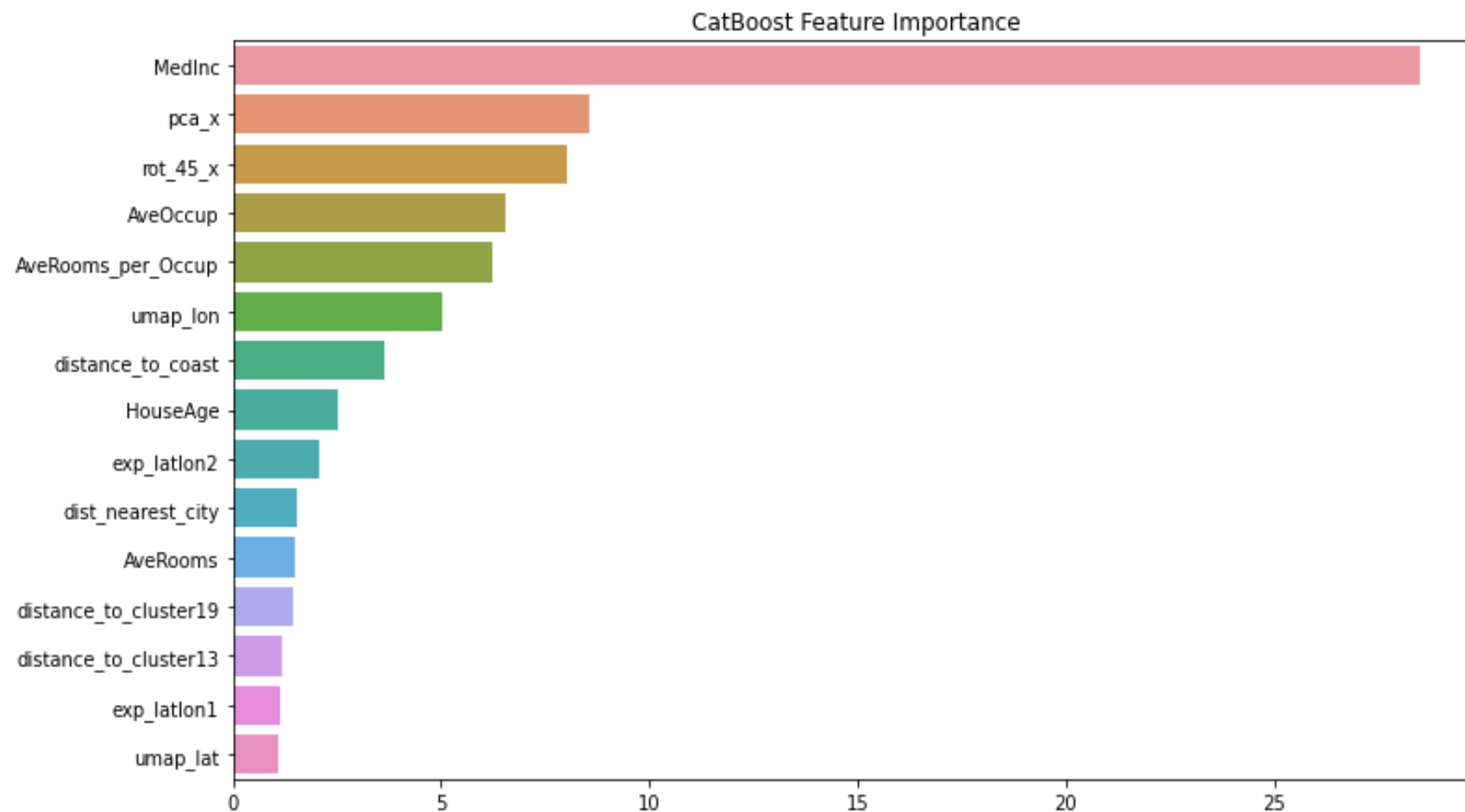
# Modeling - AutoGluon



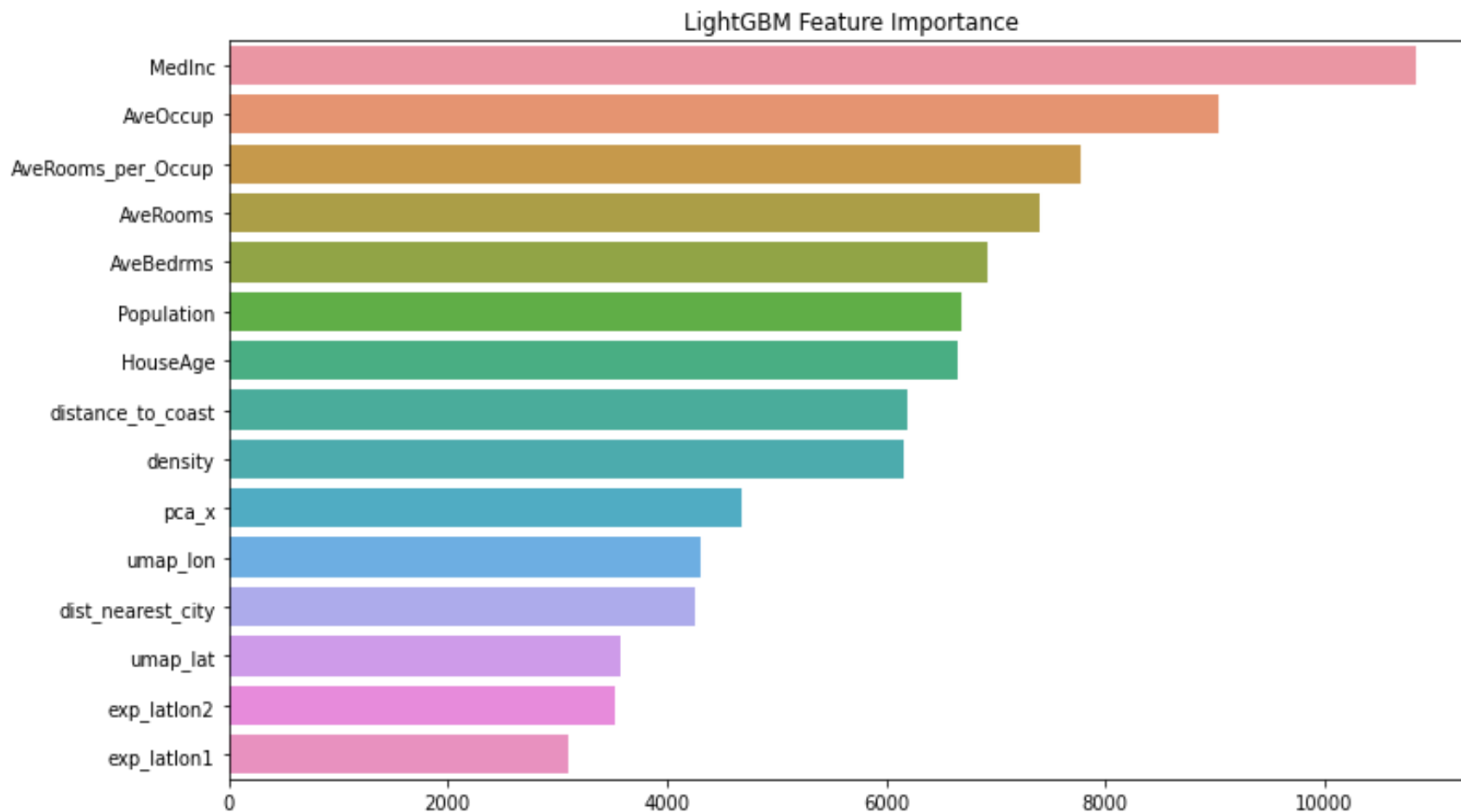
# Modeling - AutoGluon

	model	score_val	pred_time_val	fit_time	pred_time_val_marginal	fit_time_marginal	stack_level	can_infer	fit_order
0	WeightedEnsemble_L3	-0.501168	237.865424	6260.143441	0.001516	1.211894	3	True	22
1	NeuralNetFastAI_BAG_L2	-0.502204	225.927121	4172.562569	1.267741	474.260080	2	True	18
2	LightGBMXT_BAG_L2	-0.503352	225.135074	3732.709613	0.475695	34.407124	2	True	13
3	CatBoost_BAG_L2	-0.503742	224.822473	3782.600763	0.163093	84.298274	2	True	16
4	WeightedEnsemble_L2	-0.503891	219.871460	2517.521238	0.001176	1.089221	2	True	12
5	LightGBM_BAG_L2	-0.504282	224.997869	3731.434330	0.338489	33.131840	2	True	14
6	XGBoost_BAG_L2	-0.505004	225.089326	3765.672679	0.429946	67.370190	2	True	19
7	LightGBMLarge_BAG_L2	-0.505563	225.288923	3778.230351	0.629543	79.927862	2	True	21
8	ExtraTreesMSE_BAG_L2	-0.506888	228.784428	3936.402956	4.125049	238.100467	2	True	17
9	NeuralNetTorch_BAG_L2	-0.508388	225.233922	4033.557570	0.574542	335.255081	2	True	20
10	CatBoost_BAG_L1	-0.509790	0.264964	881.342007	0.264964	881.342007	1	True	6
11	RandomForestMSE_BAG_L2	-0.510260	230.198299	4945.312470	5.538919	1247.009981	2	True	15
12	LightGBMXT_BAG_L1	-0.510658	6.172180	206.595175	6.172180	206.595175	1	True	3
13	LightGBM_BAG_L1	-0.513476	3.707168	130.143349	3.707168	130.143349	1	True	4
14	LightGBMLarge_BAG_L1	-0.514150	2.416143	160.073250	2.416143	160.073250	1	True	11
15	ExtraTreesMSE_BAG_L1	-0.516010	4.583923	131.297368	4.583923	131.297368	1	True	7
16	XGBoost_BAG_L1	-0.516452	1.326900	271.015400	1.326900	271.015400	1	True	9
17	RandomForestMSE_BAG_L1	-0.527948	3.586012	715.034906	3.586012	715.034906	1	True	5
18	NeuralNetTorch_BAG_L1	-0.539319	0.499840	735.455241	0.499840	735.455241	1	True	10
19	NeuralNetFastAI_BAG_L1	-0.547167	1.203084	466.835567	1.203084	466.835567	1	True	8
20	KNeighborsUnif_BAG_L1	-0.693860	99.869215	0.385943	99.869215	0.385943	1	True	1
21	KNeighborsDist_BAG_L1	-0.718217	101.029951	0.124284	101.029951	0.124284	1	True	2
Number of models trained: 22									






# feature importance - CatBoost








# feature importance - LightGBM



# Result

#	△	Team	Members	Score	Entries	Last	Solution
1	▲ 6	Kirderf		0.55224	25	1mo	
2	▲ 2	jcerpent		0.55274	14	1mo	
3	▲ 5	NY		0.55292	27	1mo	

**AutoGluon**  
Score: 0.55204

19	▼ 8	stplgk		0.55428	12	1mo	
20	▲ 5	Siddharth Dhawan		0.55430	25	1mo	
21	▼ 12	identity_stolen		0.55436	7	1mo	
22	▲ 18	stefgina		0.55444	5	1mo	
23	▲ 7	Caleb Emelike		0.55449	12	1mo	

**Blending**  
Score: 0.55435



# Episode 3 - Classification

Predict attrition of employees



1차 목표 : 10% 이내

66	▲ 121	Jon Blanchard		0.89655
67	▲ 92	Mahyar Arani		0.89654



# Episode 3 - Data

## 데이터 개수

Train	Test	추가(ibm)
1667	1119	1470

## Train Data Features

Age	PercentSalaryHike	Education
YearsSinceLastPromotion	OverTime	EducationField
YearsInCurrentRole	Over18	YearsWithCurrManager
YearsAtCompany	NumCompaniesWorked	EmployeeCount
WorkLifeBalance	MonthlyRate	Gender
TrainingTimesLastYear	MonthlyIncome	HourlyRate
TotalWorkingYears	MaritalStatus	JobInvolvement
StockOptionLevel	BusinessTravel	JobLevel
StandardHours	DailyRate	JobRole
RelationshipSatisfaction	Department	JobSatisfaction
PerformanceRating	DistanceFromHome	EnvironmentSatisfaction

Attrition

- 독립변수 33개
- 종속변수 1개

# Episode 3 - Data 전처리

## 추가 데이터 전처리

'Attrition' (0, 1) 변환

train에는 없는 'EmployeeNumber' feature 제거

Attrition
Yes
No

35 columns

EmployeeNumber
1
2



Attrition
1
0

34 columns

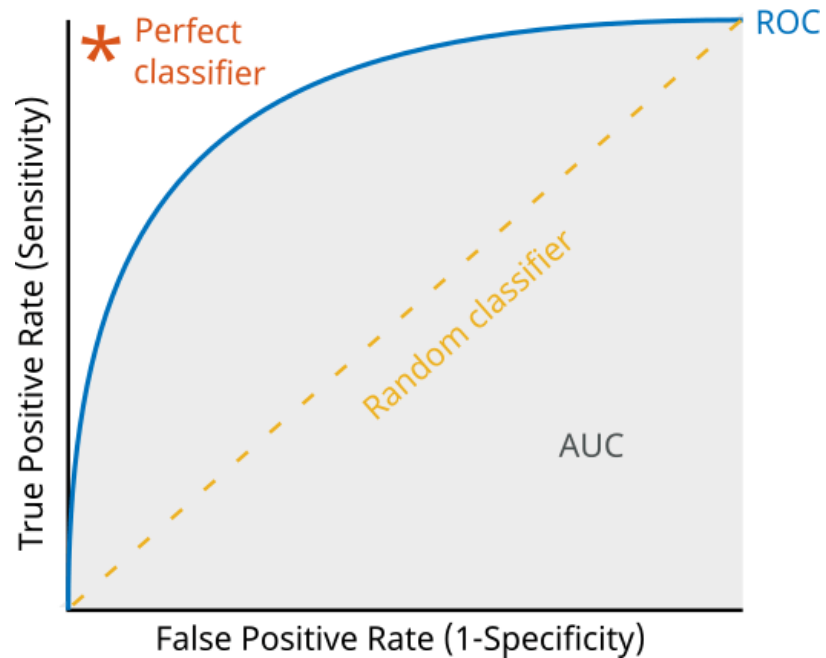
## 데이터 결합

train + ibm = 1677 + 1470 = 3147

## 고유값 1개 feature 제거

feature_name	type	결측값수	고유값수
EmployeeCount	int64	0	1
StandardHours	int64	0	1
Over18	object	0	1

## Episode 3 - 평가지표



0.89655 ↑

목표

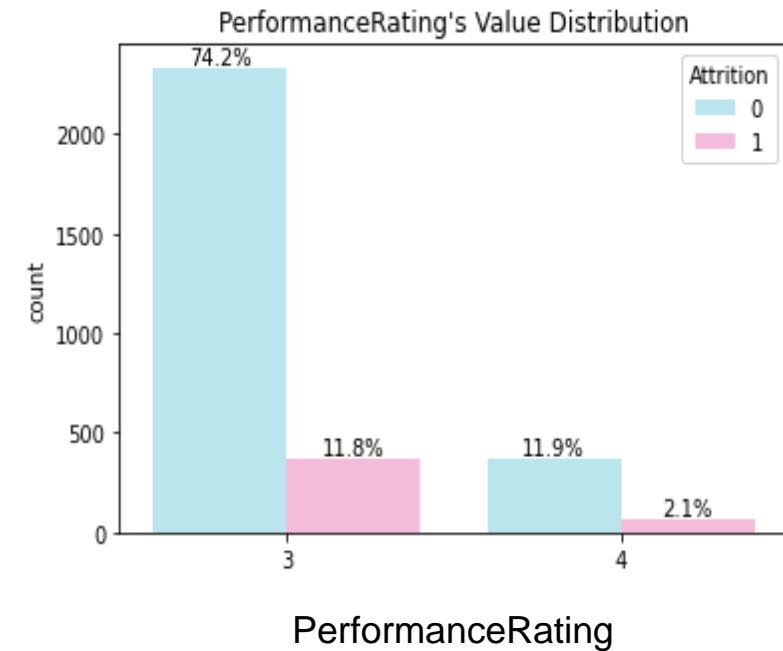
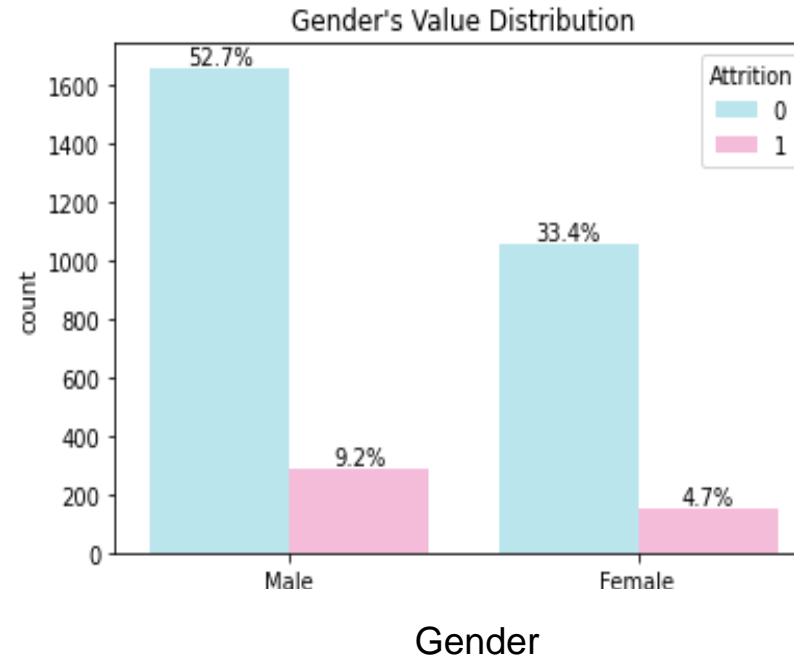
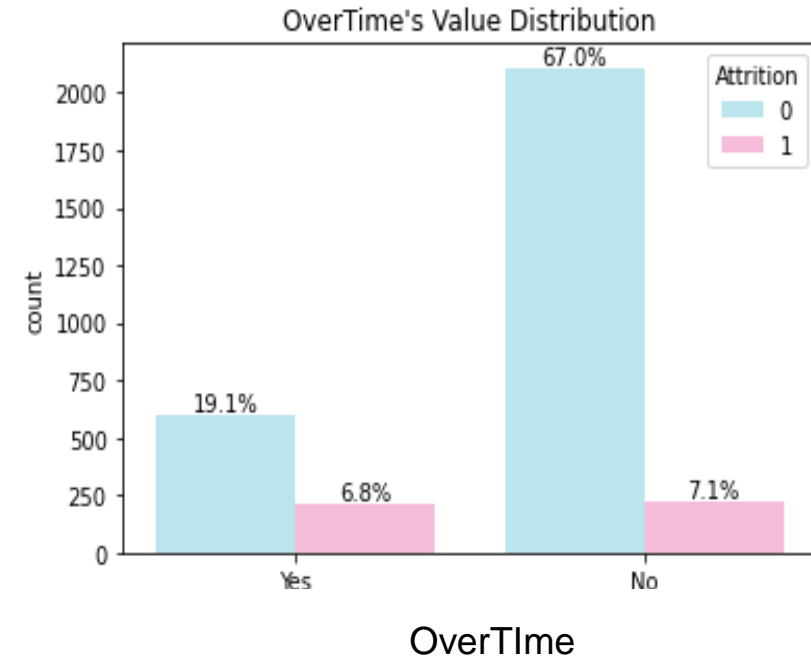
ROC 아래 AUC 면적이 넓을수록 클래스를 구별하는 모델의 성능이 우수

# EDA - 이진형 데이터

- OverTime : 초과 근무 (Yes, No)
- Gender : 성별 (Male, Female)
- PerformanceRating : 성과 등급 (3, 4)

feature_name	type	결측값수	고유값수	샘플값 0	샘플값 1	샘플값 2
OverTime	object	0	2	Yes	No	No
Gender	object	0	2	Male	Male	Male
PerformanceRating	int64	0	2	3	3	3

# EDA - Data visualization

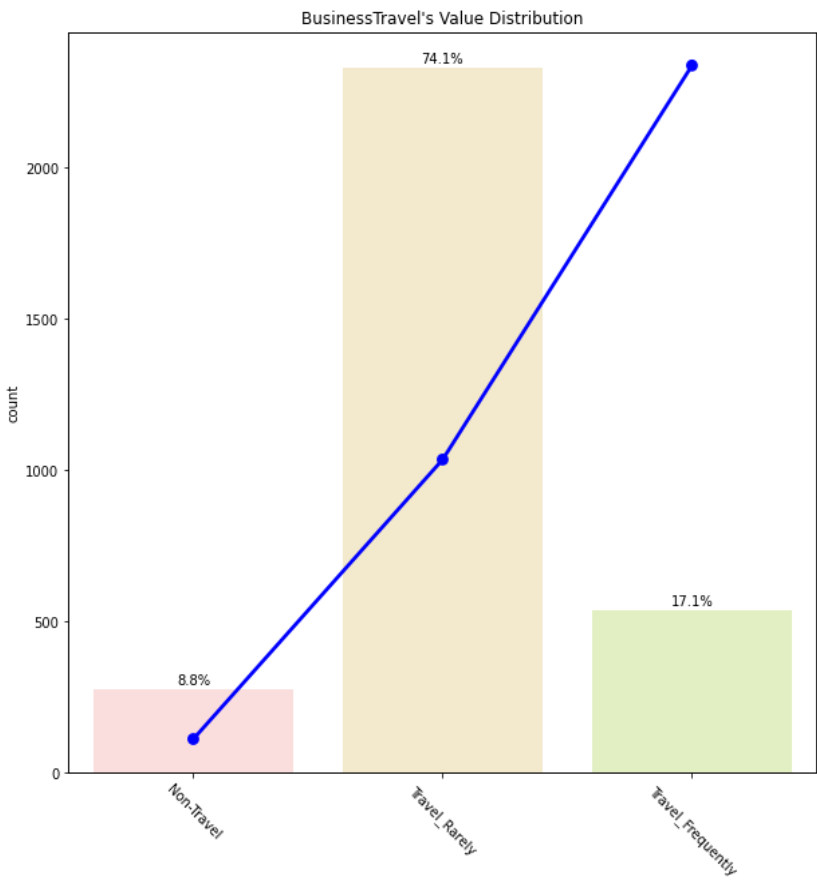


# EDA - 명목형 데이터

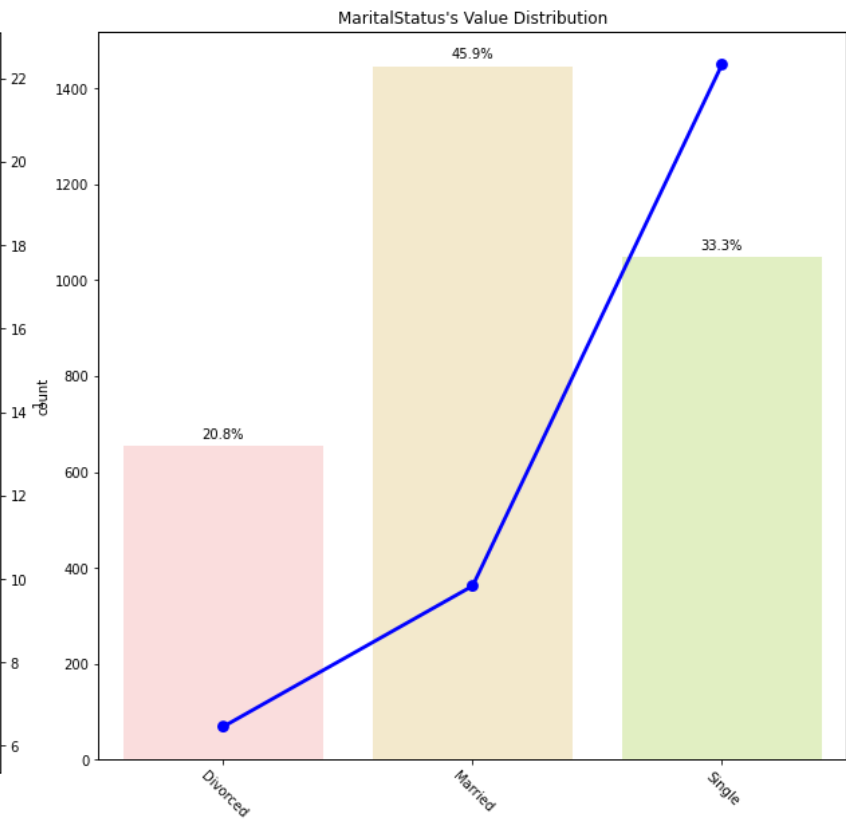
- BusinessTravel : 출장빈도
- Department : 부서
- MaritalStatus : 결혼여부
- EducationField : 교육분야
- JobRole : 직무

feature_name	type	결측값수	고유값수	샘플값 0	샘플값 1	샘플값 2
BusinessTravel	object	0	3	Travel_Frequently	Travel_Rarely	Travel_Rarely
Department	object	0	3	Research & Development	Sales	Sales
MaritalStatus	object	0	3	Married	Married	Divorced
EducationField	object	0	6	Medical	Other	Marketing
JobRole	object	0	9	Laboratory Technician	Sales Representative	Sales Executive

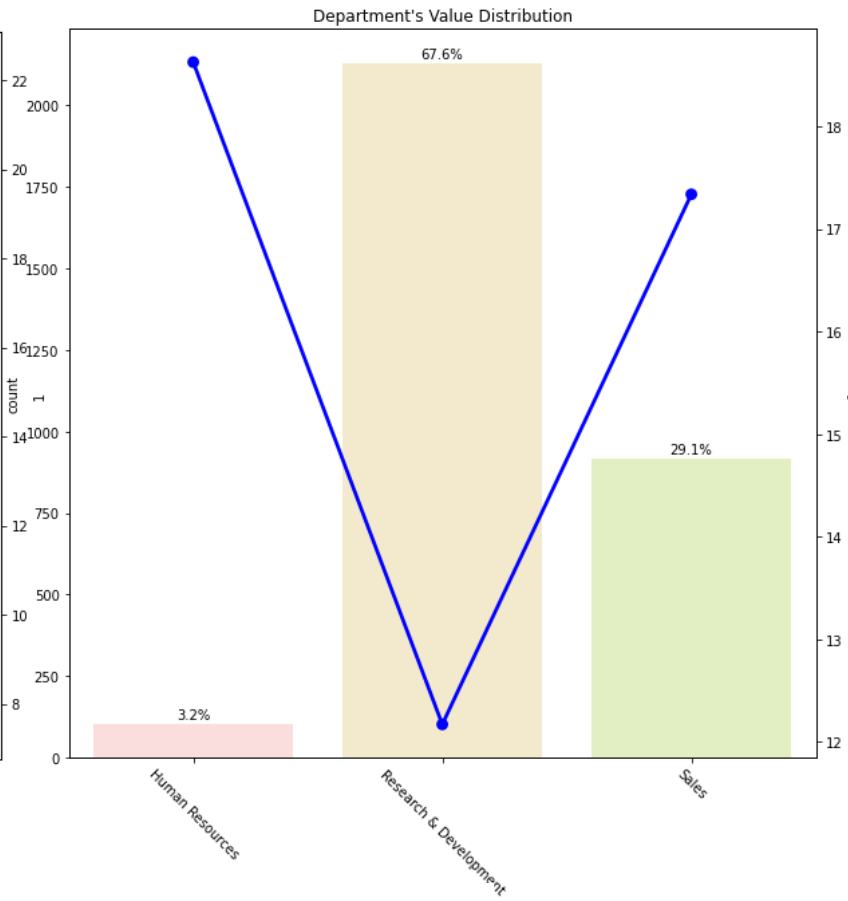
# EDA - Data visualization



BusinessTravel



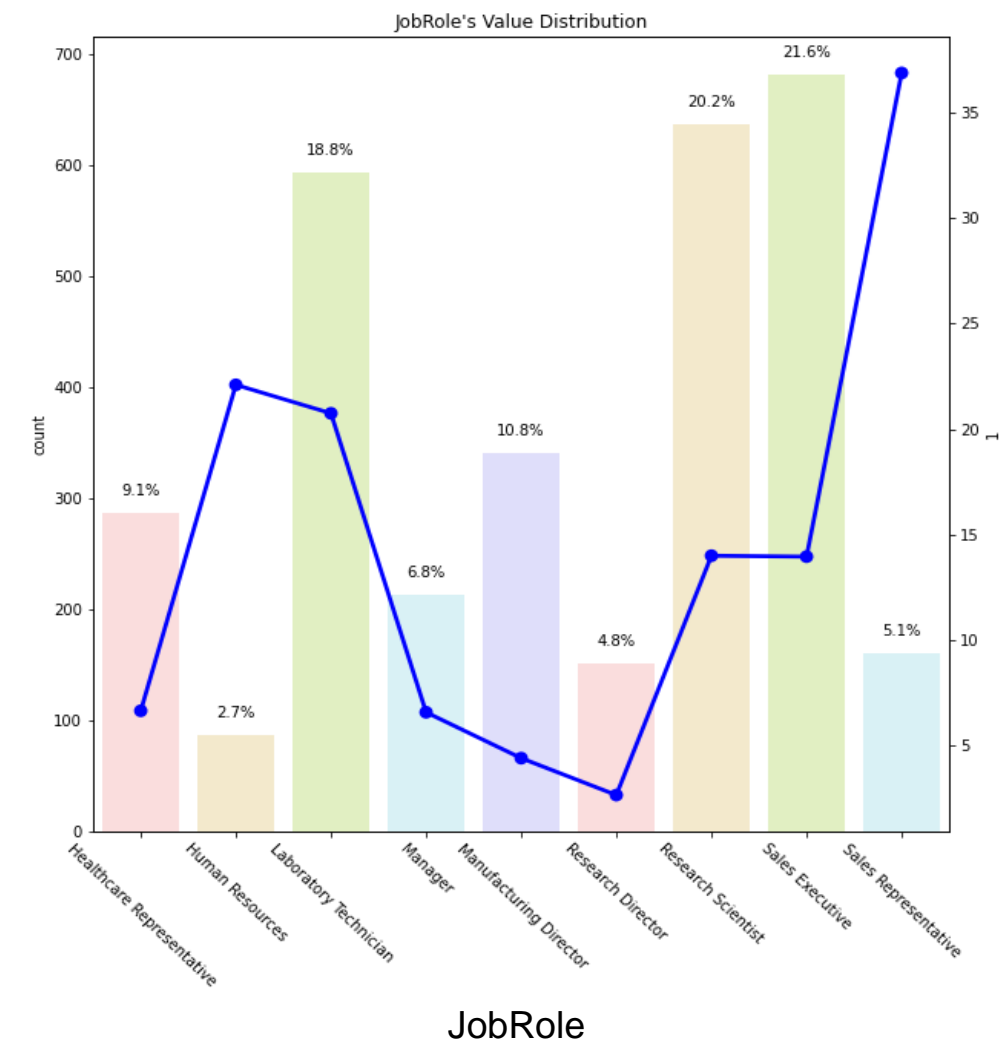
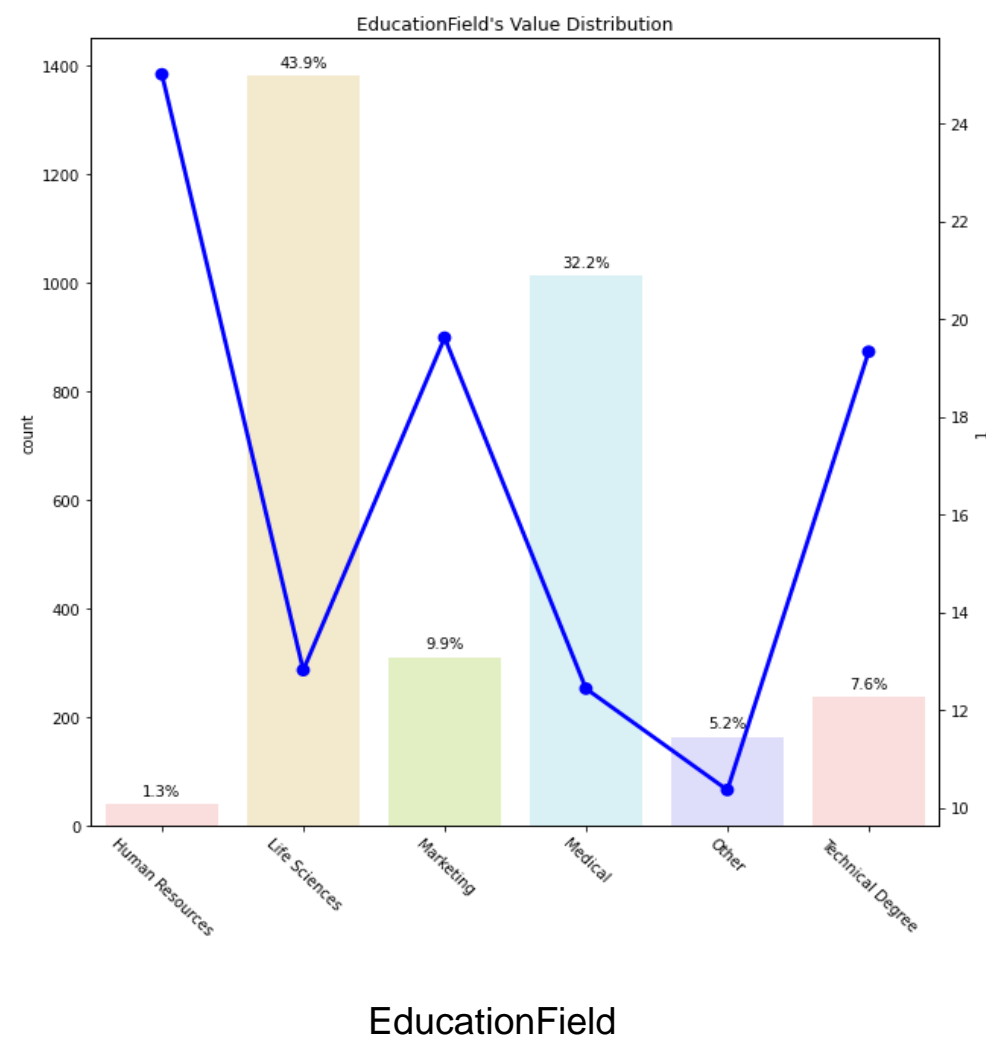
MaritalStatus



Department



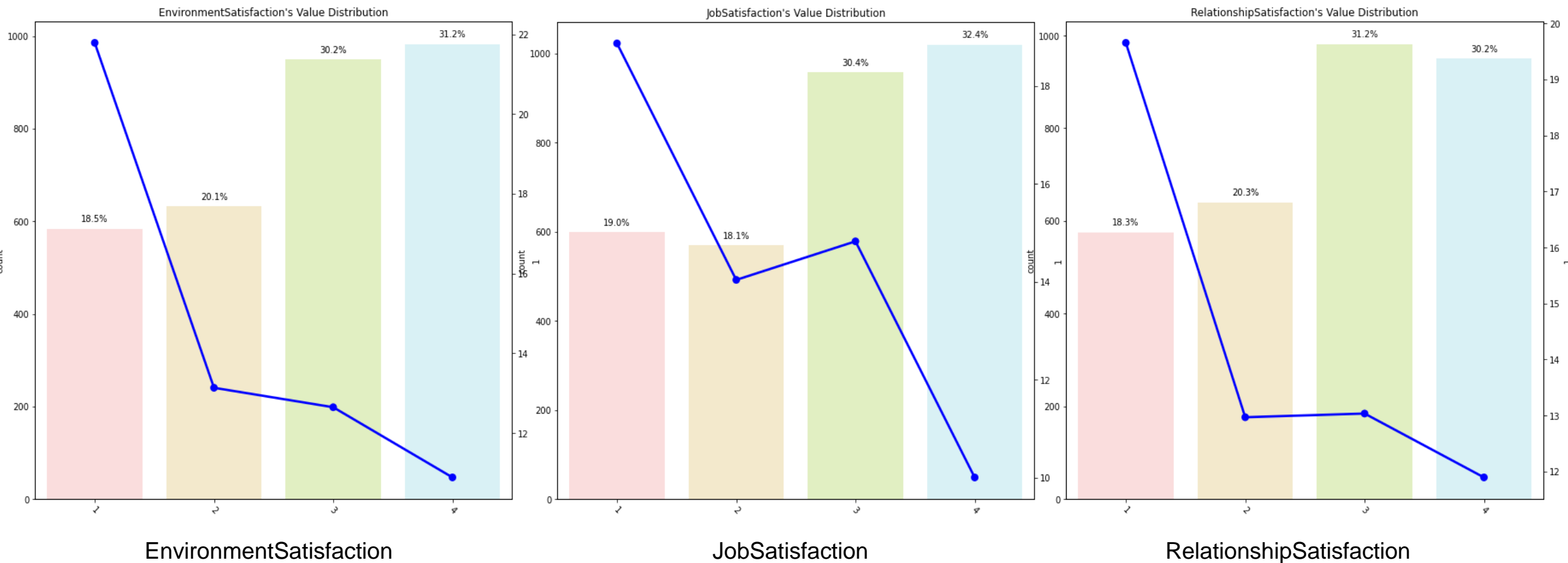
# EDA - Data visualization



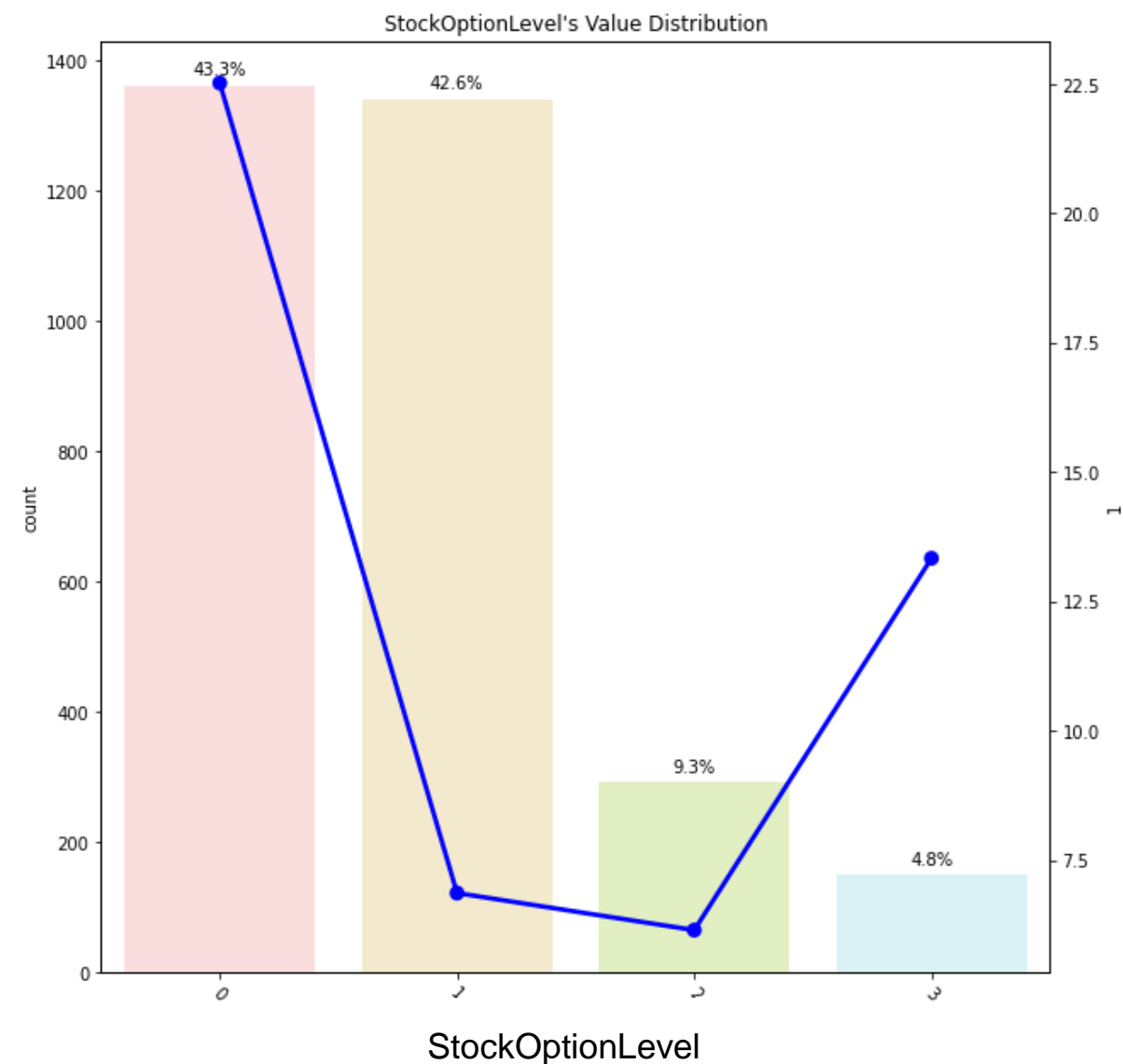
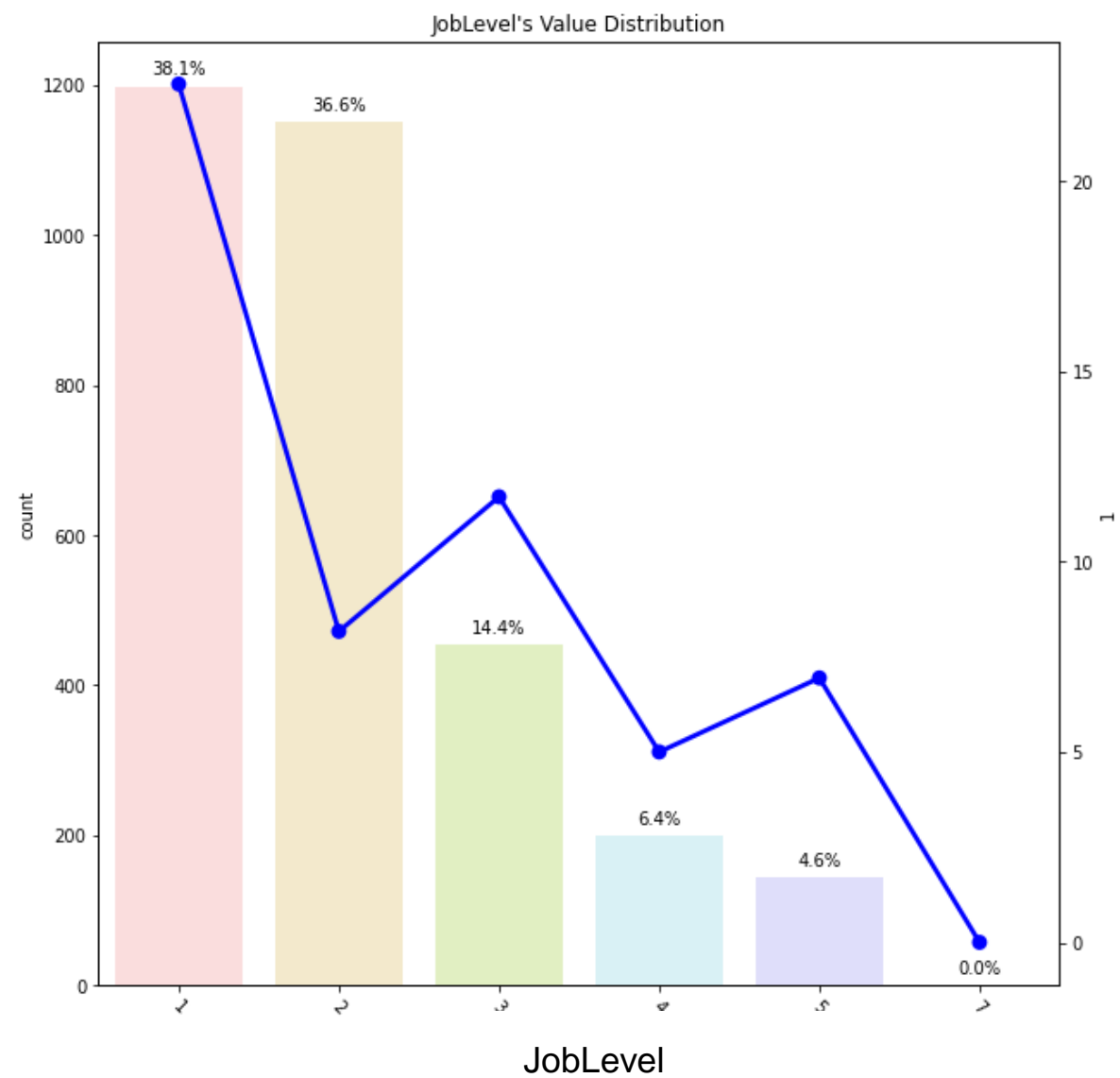
# EDA - 순서형 데이터

feature_name	type	결측값수	고유값수	샘플값 0	샘플값 1	샘플값 2
EnvironmentSatisfaction	int64	0	4	4	1	3
JobSatisfaction	int64	0	4	4	1	4
JobLevel	int64	0	5	1	1	2
RelationshipSatisfaction	int64	0	4	2	4	4
StockOptionLevel	int64	0	4	1	1	2
JobInvolvement	int64	0	4	3	3	3
WorkLifeBalance	int64	0	4	3	3	3

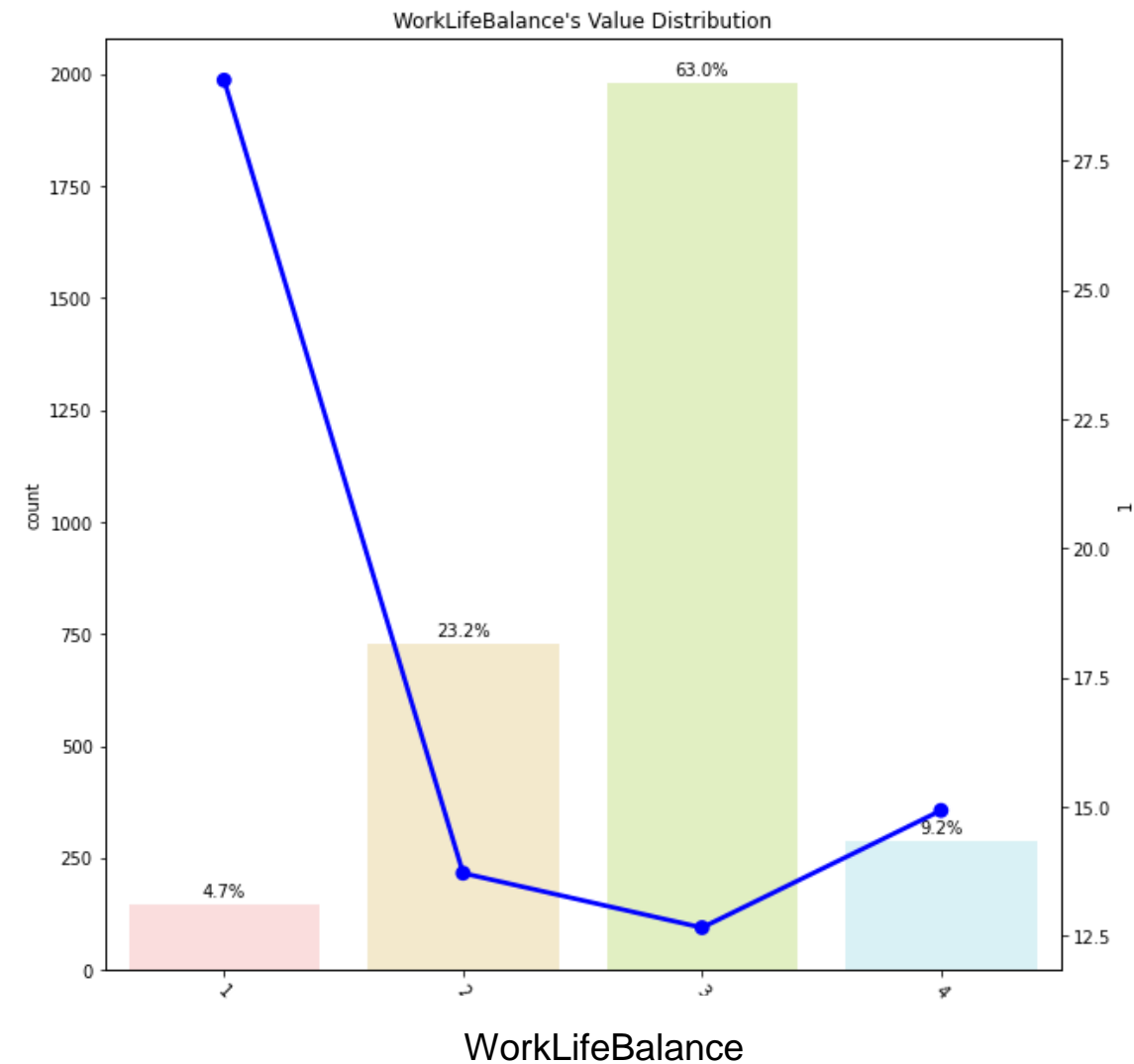
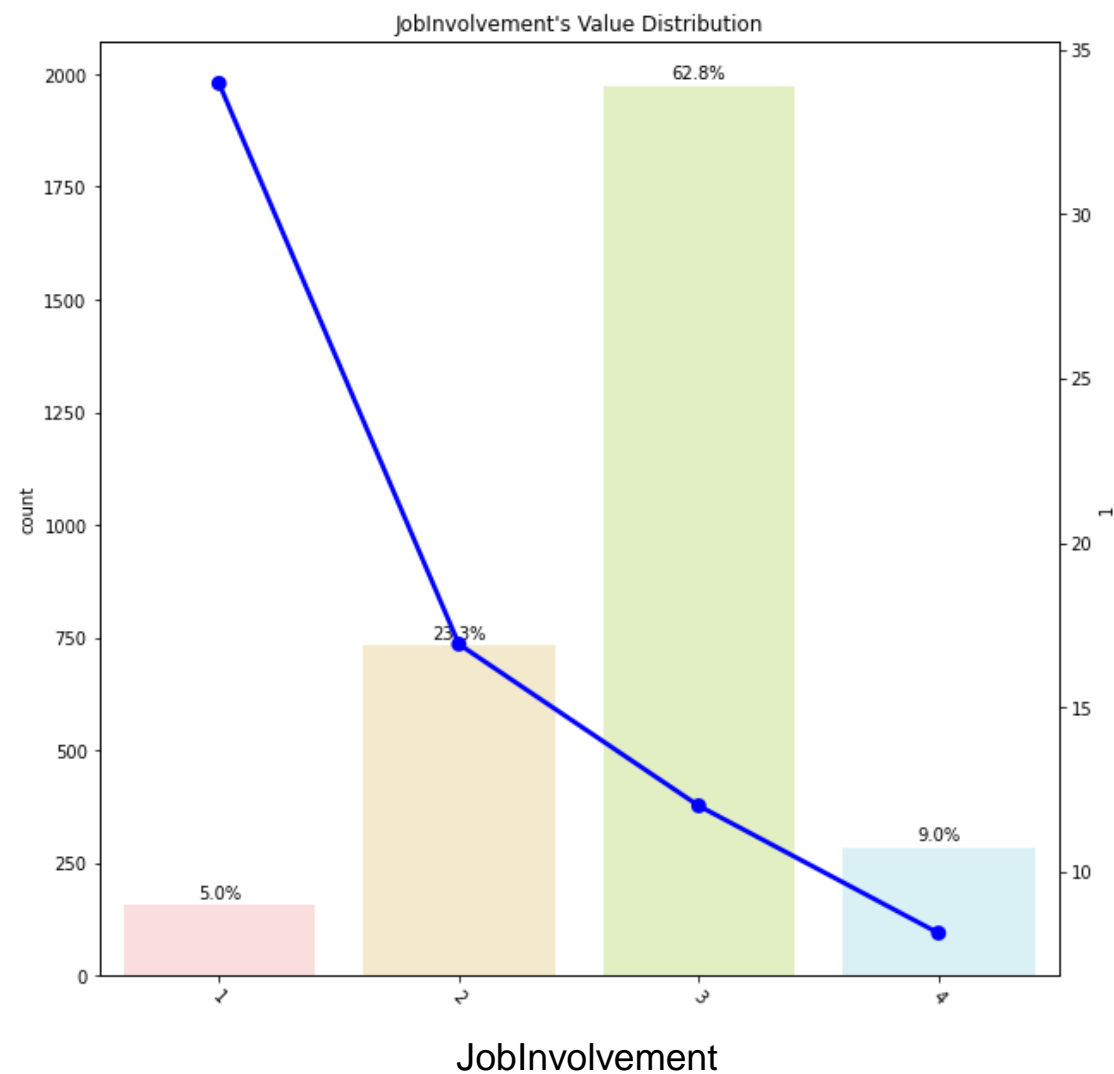
# EDA - Data visualization



# EDA - Data visualization



# EDA - Data visualization



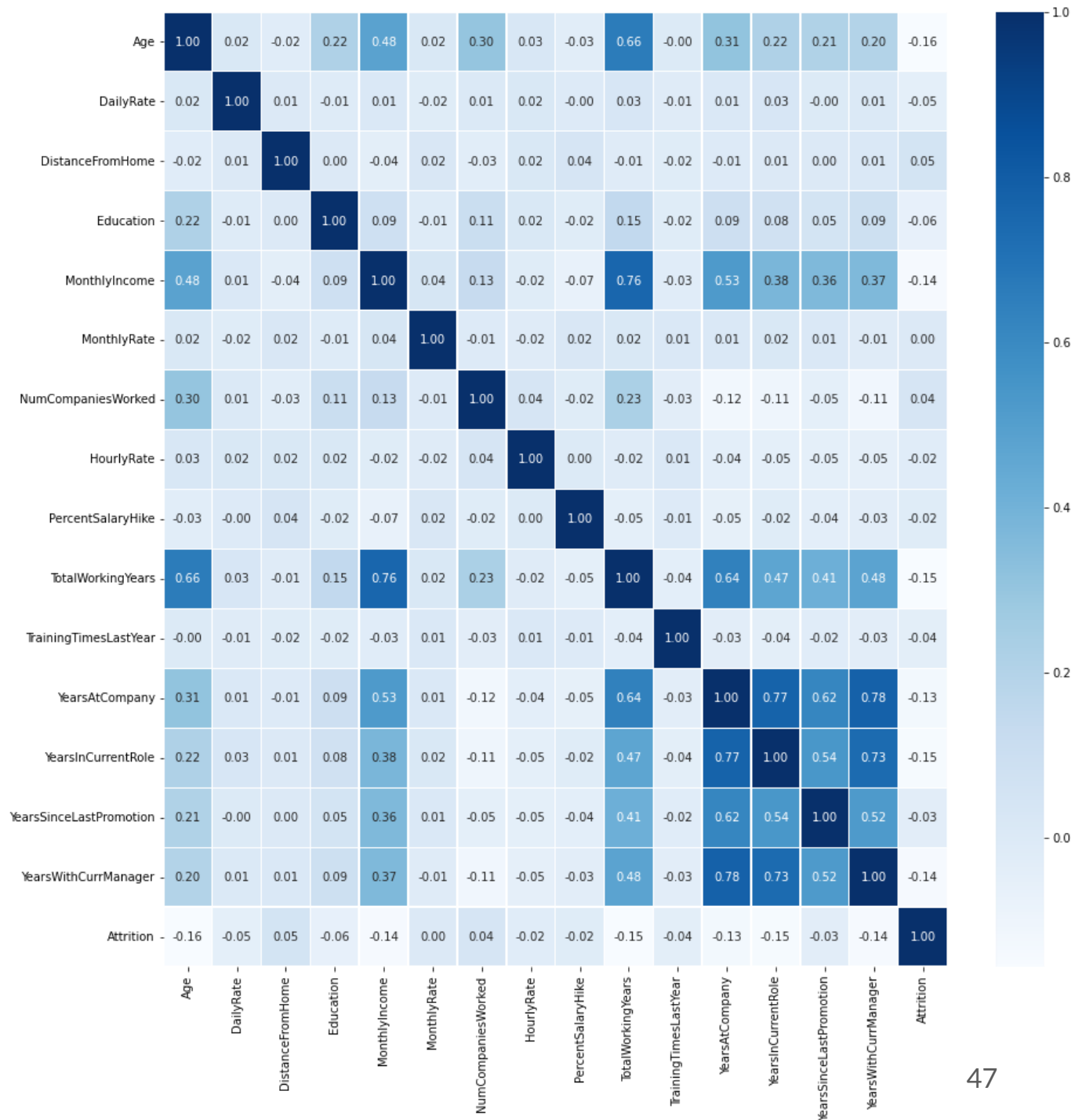
# EDA - 수치형 데이터

	type	결측값수	고유값수	샘플값 0	샘플값 1	샘플값 2
feature_name						
Age	int64	0	43	36	35	32
DailyRate	int64	0	901	599	921	718
DistanceFromHome	int64	0	29	24	8	26
Education	int64	0	6	3	3	3
MonthlyIncome	int64	0	1383	2596	2899	4627
MonthlyRate	int64	0	1447	5099	10778	16495
NumCompaniesWorked	int64	0	10	1	1	0
HourlyRate	int64	0	71	42	46	80
PercentSalaryHike	int64	0	15	13	17	17
TotalWorkingYears	int64	0	41	10	4	4
TrainingTimesLastYear	int64	0	7	2	3	3
YearsAtCompany	int64	0	38	10	4	3
YearsInCurrentRole	int64	0	19	0	2	2
YearsSinceLastPromotion	int64	0	16	7	0	1
YearsWithCurrManager	int64	0	18	8	3	2

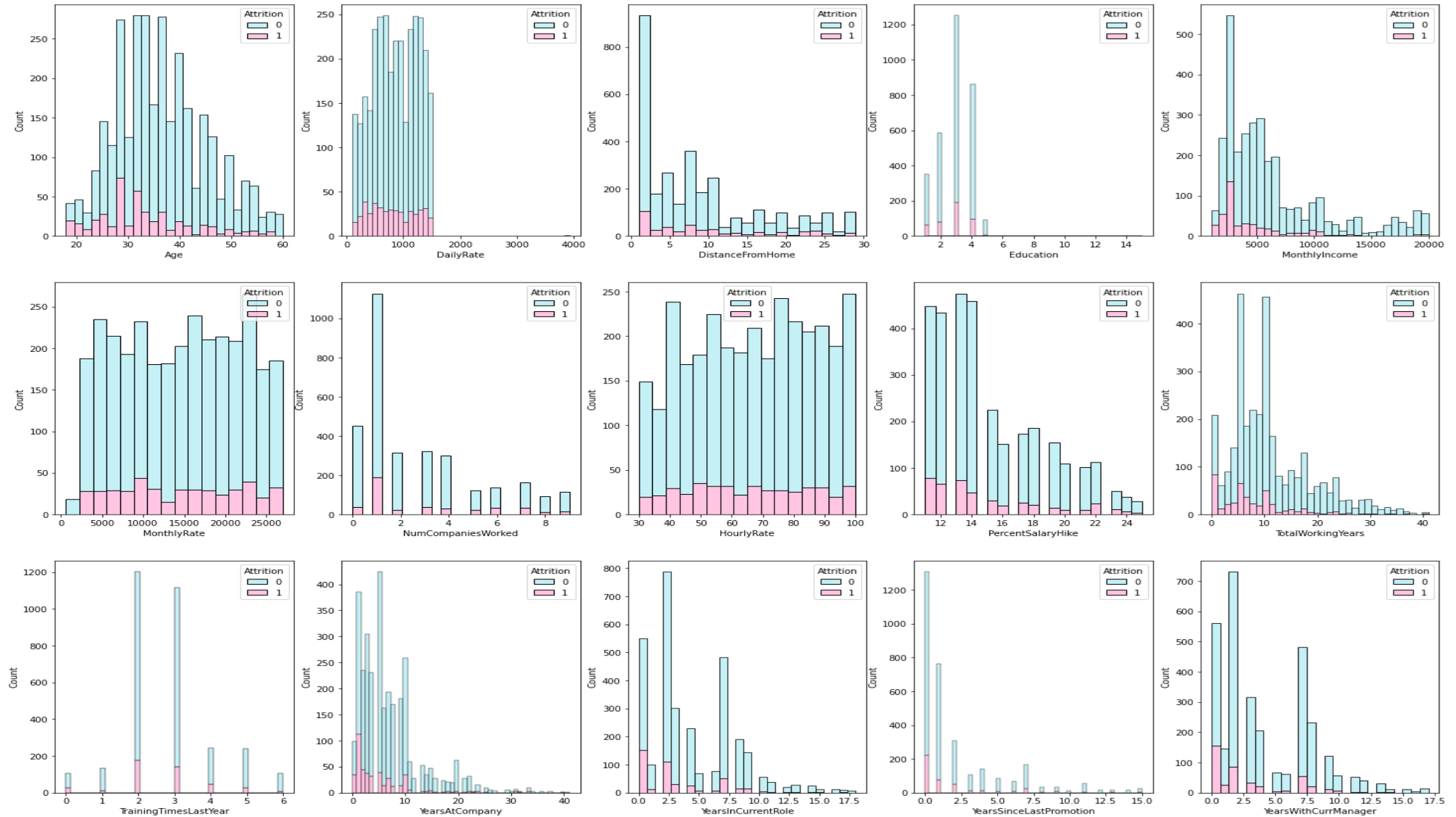
# EDA - Data visualization

## 히트맵으로 나타낸 상관관계

- Age랑 TotalWorkingYears는 0.65 로 상관
- TotalWorkingYears랑 MonthlyIncome 0.74로 상관
- TotalWorkingYears는 피쳐명에 Years가 앞에 붙음 피쳐들과 상관
- 피쳐명에 Years가 앞에 붙음 피쳐들끼리 상관관계



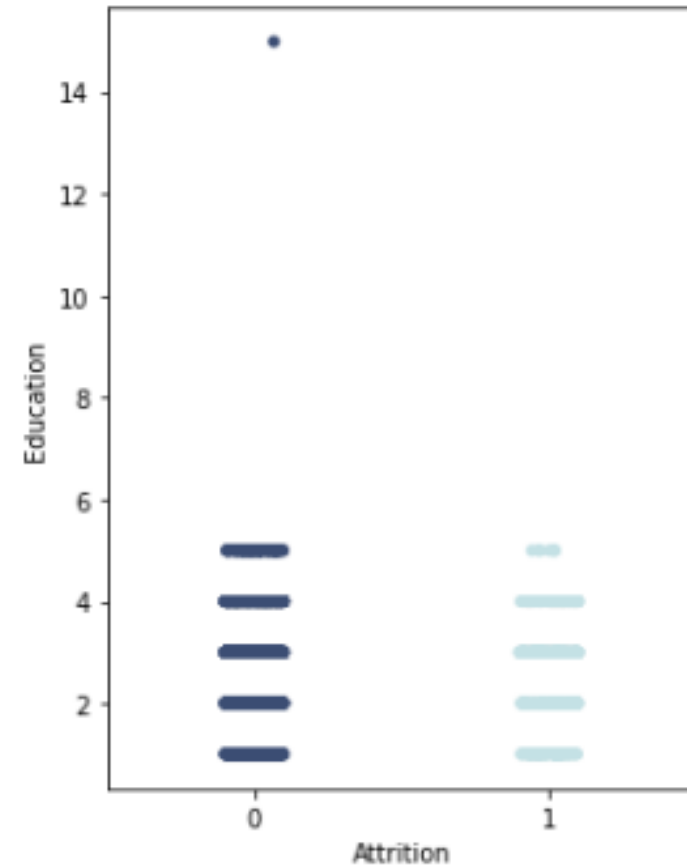
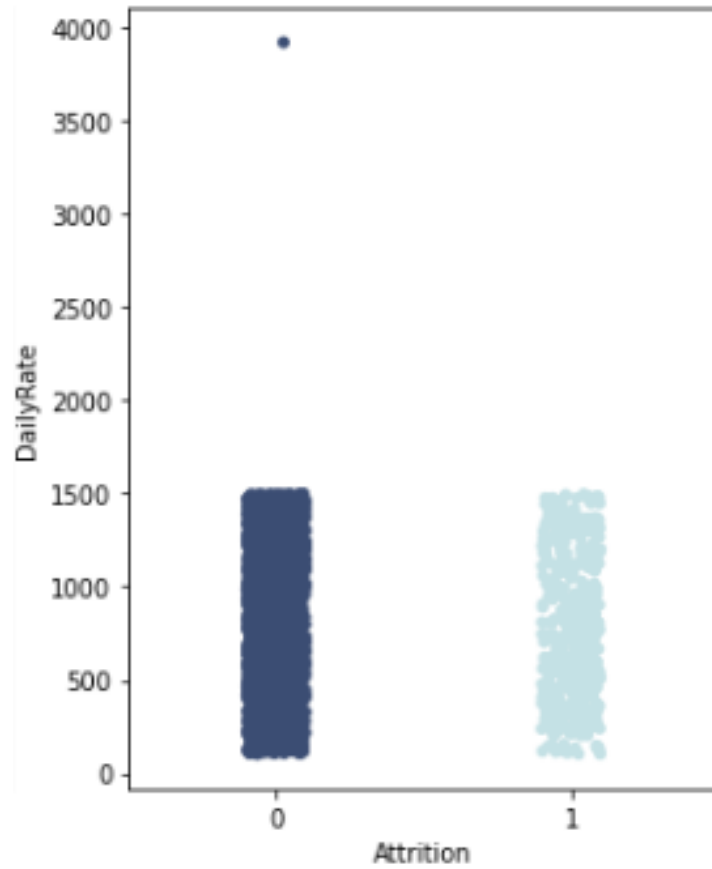
# EDA - Data visualization





# EDA - Data visualization

산점도



# Feature Engineering

## 인코딩

이진형 EDA 결과 - 0과 1로 변환

OverTime	Gender	PerformanceRating
Yes	Male	3
No	Male	3

↓

OverTime	Gender	PerformanceRating
1	1	1
0	1	1

명목형 EDA 결과

- 'BusinessTravel' 피쳐는 순서 지정.
- 'Non-Travel':0, 'Travel\_Rarely':1, 'Travel\_Frequently':2 순서대로 인코딩

BusinessTravel	BusinessTravel
Travel_Frequently	2
Travel_Rarely	1

\* 원-핫 인코딩 X => CatBoost는 범주형 데이터 인코딩을 자체적으로 해줌

# Feature Engineering

## 스케일링

### StandardScaler로 진행

(평균 0, 표준편차 1이 되도록 모든 값을 조정하여 변환)

#### 순서형 스케일링 결과

EnvironmentSatisfaction	1.156127
JobSatisfaction	1.125829
JobLevel	-0.941910
RelationshipSatisfaction	-0.679735
StockOptionLevel	0.301665
JobInvolvement	0.355669
WorkLifeBalance	0.346421

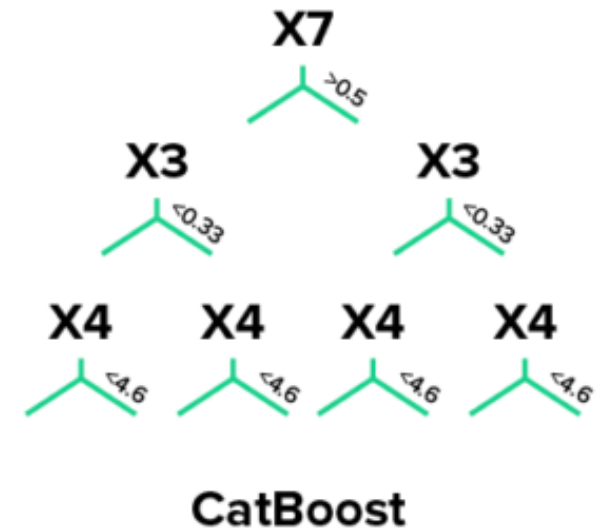
#### 수치형 스케일링 결과

Age	-0.051082
DailyRate	-0.647934
DistanceFromHome	1.893264
Education	0.076865
MonthlyIncome	-0.812007
MonthlyRate	-1.329491
NumCompaniesWorked	-0.654436
HourlyRate	-1.252502
PercentSalaryHike	-0.579114
TotalWorkingYears	-0.129550
TrainingTimesLastYear	-0.629423
YearsAtCompany	0.517586
YearsInCurrentRole	-1.161396
YearsSinceLastPromotion	1.572161
YearsWithCurrManager	1.076581

# Modeling



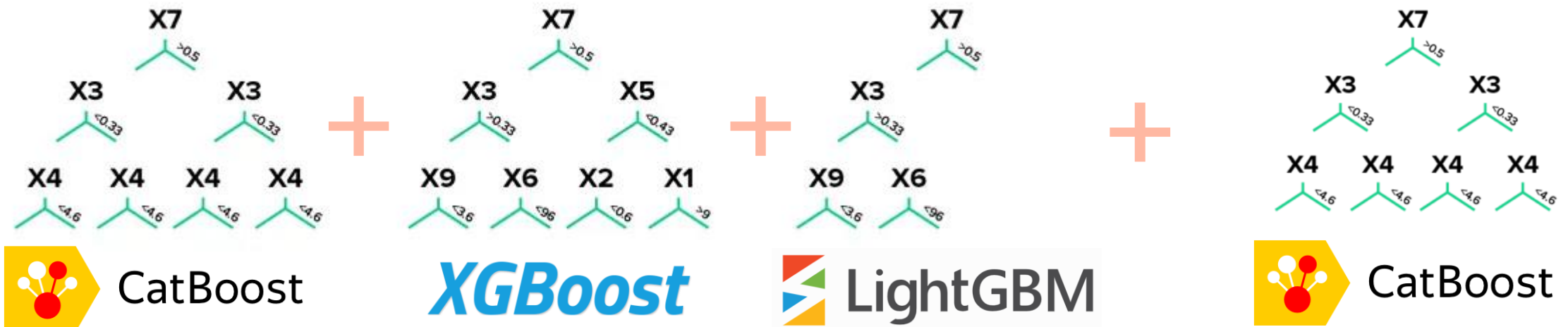
CatBoost





# Modeling


Blending : (XGBoost + LightGBM + CatBoost) + 생성한 모델(CatBoost) 예측값의 평균

Tree growth examples:





# Result

11	tharun_nayak14		0.90052	10	<b>CatBoost Single Model</b> Private Score : 0.90037
12	@kaggleqrdl		0.90029	18	

					<b>Blending Models</b>
1	Bill Cruise		0.90185		Private Score : 0.90407

## 최종 SCORE

메달	Episode	순위	Private Score
	1	1/689	0.55204
	2	3/770	0.90059
	3	1/665	0.90407
	6	41/629	102008.08412

# 정리

## 결과

Ensemble Technique을 이용하여 더 좋은 성능의 Model을 만듦  
다양한 Feature Engineering 기법을 이용하여 Model 성능을 올림  
Pipeline Optimization을 통해 Kaggle Score값을 올림

## 경험 및 고찰

시각화와 데이터 자료를 이용한 EDA 및 데이터 전처리 역량을 키움  
다양한 인코딩과 스케일링 기법을 통해 피쳐 엔지니어링을 이해 및 적용  
ML 알고리즘에 대해 이해하고, 실제로 적용하여 사용법에 익숙해짐  
최적화에 대한 이해 및 적용 방법 탐색 수행  
캐글 대회 경험



감사합니다

# Q&A