직원 미래 예측

Employee Future Prediction

컴퓨터공학과 2017108248 고지훈

목차

1 데이터셋

데이터셋

	Education	JoiningYear	City	PaymentTier	Age	Gender	EverBenched	ExperienceInCurrentDomain	LeaveOrNot
0	Bachelors	2017	Bangalore	3	34	Male	No	0	0
1	Bachelors	2013	Pune	1	28	Female	No	3	1
2	Bachelors	2014	New Delhi	3	38	Female	No	2	0
3	Masters	2016	Bangalore	3	27	Male	No	5	1
4	Masters	2017	Pune	3	24	Male	Yes	2	1

Education : 취득 학위

JoiningYear : 입사일

City : 거주 지역

PaymentTier : 소득분위

Age : 나이

Gender : 성별

EverBenched : 1개월 이상 프로젝트에 참여하지 않은 지 여부

ExperienceInCurrentDomain : 경력기간

LeaveOrNot : 2년 안에 퇴사예정

데이터셋

```
print("총 데이터량:",df.size)
print("직원수:",df.shape[0])
```

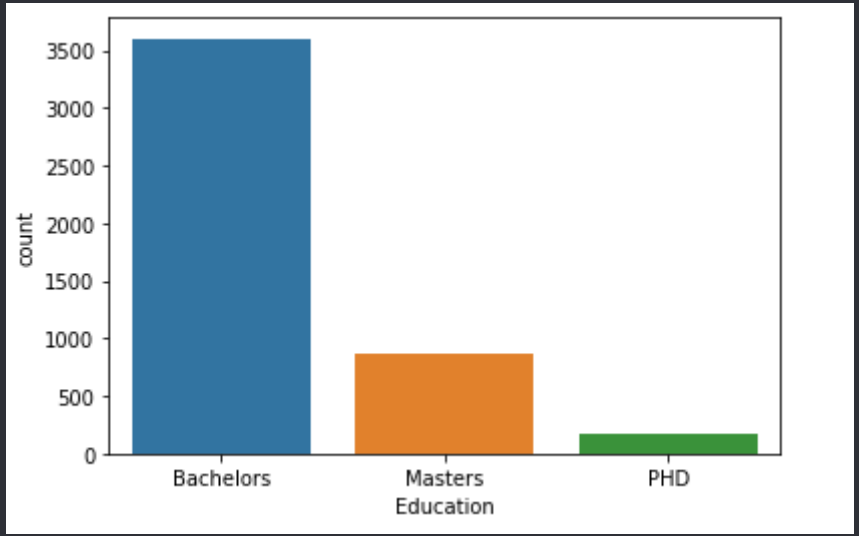
```
총 데이터량: 41877
직원수: 4653
```

```
df.dtypes
```

```
Education                   object
JoiningYear                  int64
City                        object
PaymentTier                  int64
Age                          int64
Gender                      object
EverBenched                 object
ExperienceInCurrentDomain    int64
LeaveOrNot                   int64
dtype: object
```

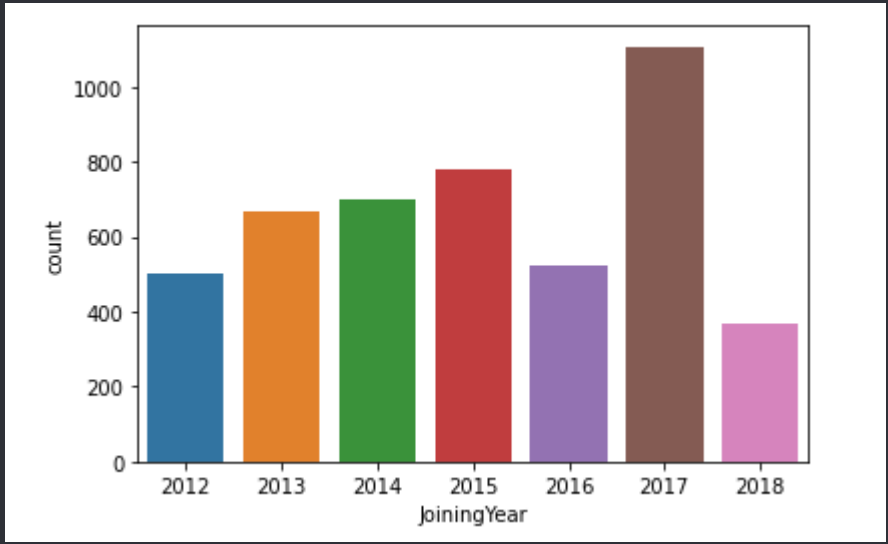
2 데이터 분석

데이터 분석(학위)



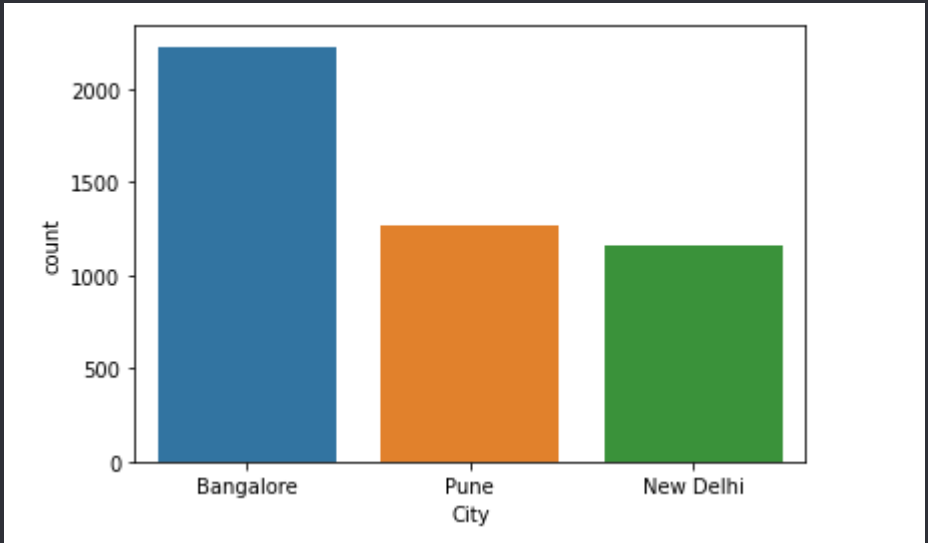
대부분의 직원이 학사 학위를 수료한것을 알수 있다.

데이터 분석(입사일)



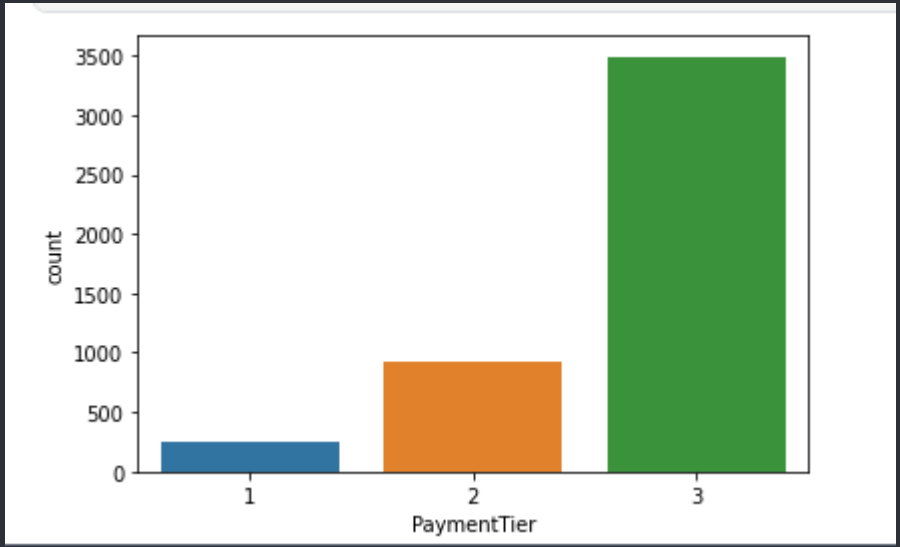
2017년에 가장 많이 입사한 것을 알 수 있다.

데이터 분석(지역)



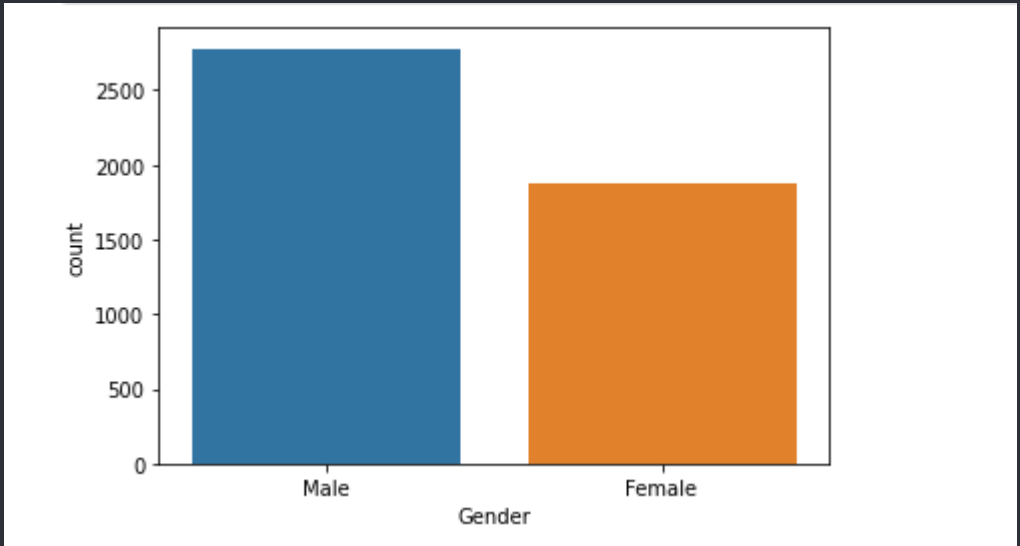
가장 많이 거주하는 지역이 Bangalore이라는 것을 알 수 있습니다.

데이터 분석(소득분위)



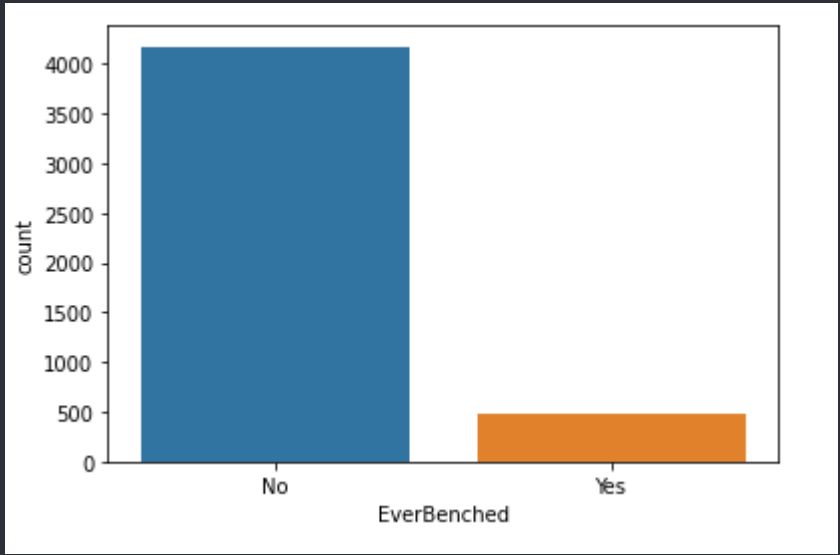
소득분위의 경우 3분위 즉 낮은 분위에 많이 인원이 분포한다는 것을 알 수 있다.

데이터 분석(성별)



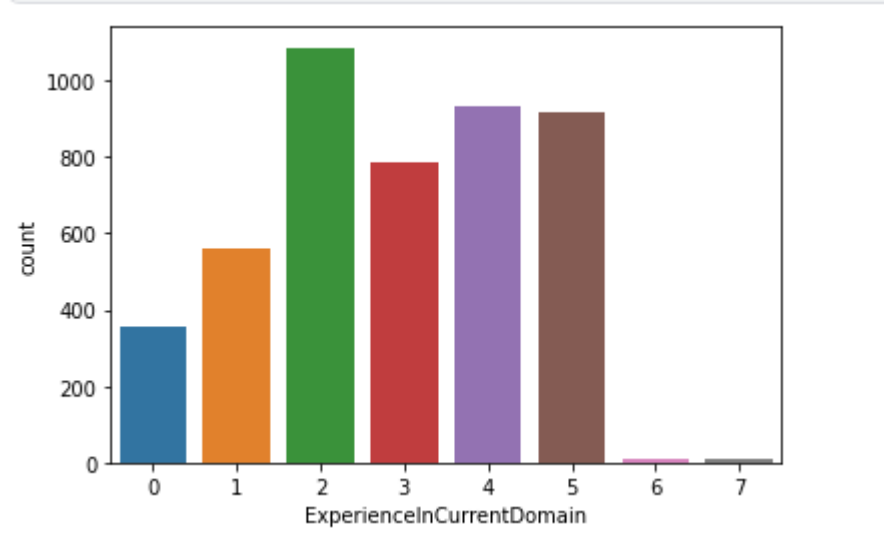
여성보다 남성이 더 많은 수를 차지하는 것을 알 수 있다.

데이터 분석(EverBenched)



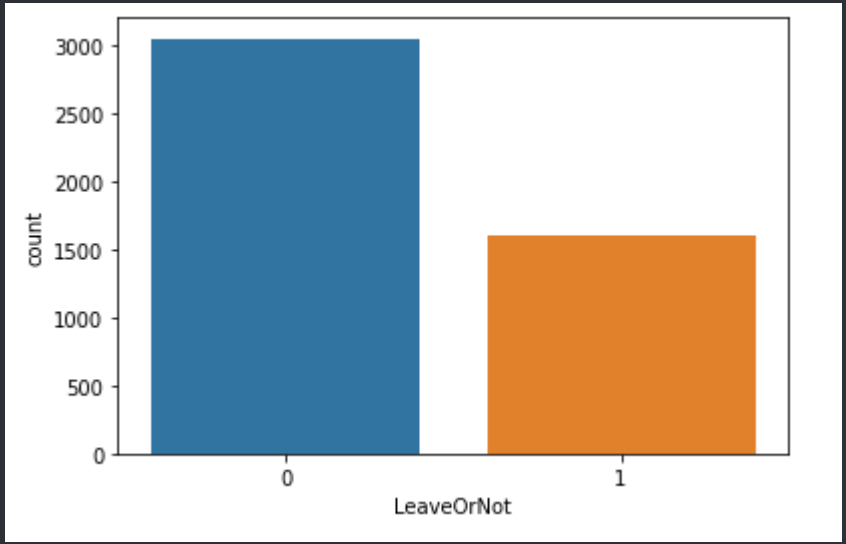
1개월 이상 일을 하지 않은 사람은 많이 없는 것으로 확인 할 수 있다.

데이터 분석(경력 기간)



직원의 경력이 대부분 2~5년 사이임을 알 수 있습니다.

데이터 분석(Leave Or Not)



대부분의 직원들이 2년 안에 퇴사예정자는 없는 것으로 확인 됩니다.

2 데이터 분석(EDA)

데이터 분석(EDA)

EDA란

정의

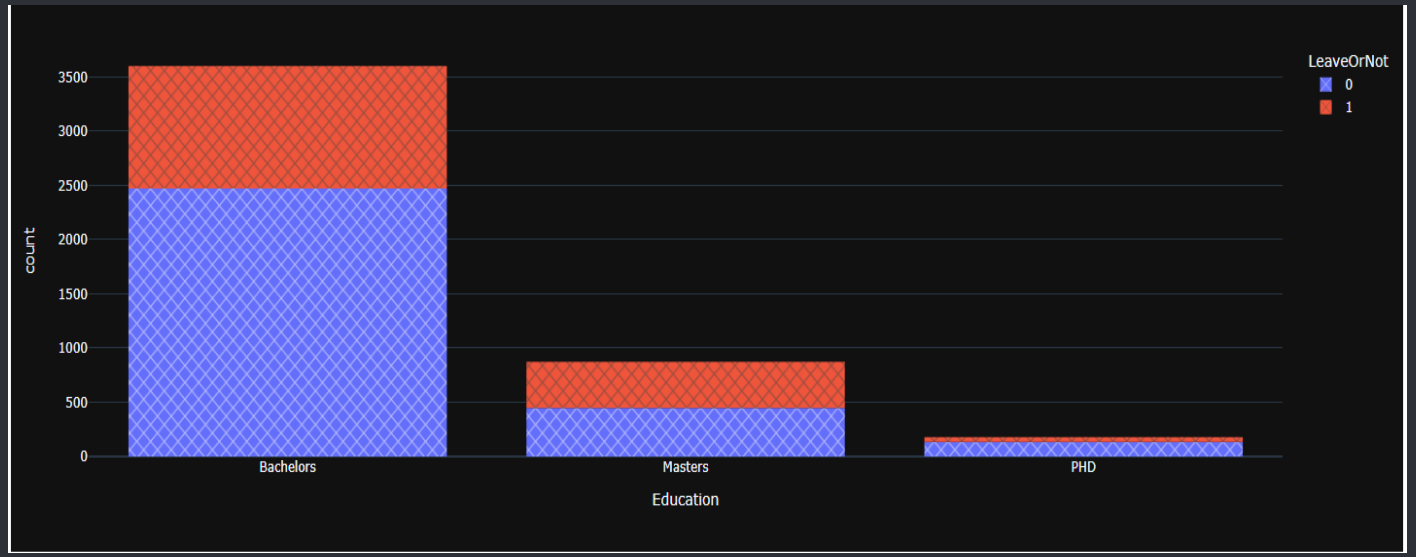
수집한 데이터가 들어왔을 때, 이를 다양한 각도에서 관찰하고 이해하는 과정

필요한 이유

1. 데이터의 분포 및 값을 검토함으로써 데이터가 표현하는 현상을 더 잘 이해하고, 데이터에 대한 잠재적인 문제를 발견할 수 있습니다.

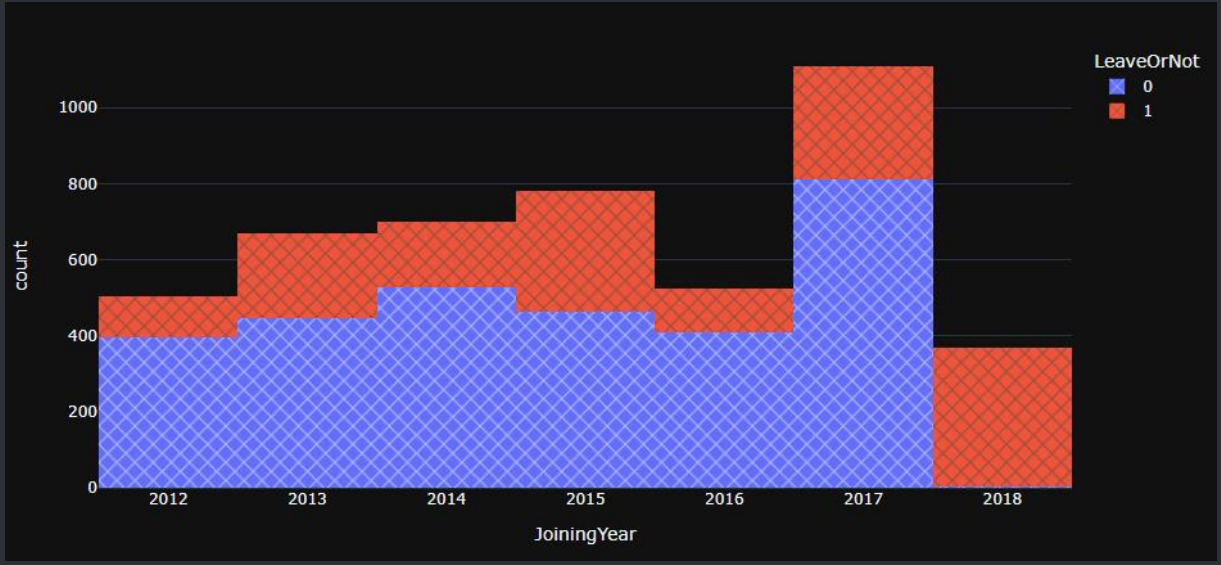
2. 다양한 각도에서 살펴보는 과정을 통해 문제 정의 단계에서 미쳐 발생하지 못했을 다양한 패턴을 발견 할 수 있습니다.

데이터 분석(학위 vs LeaveOrNot)



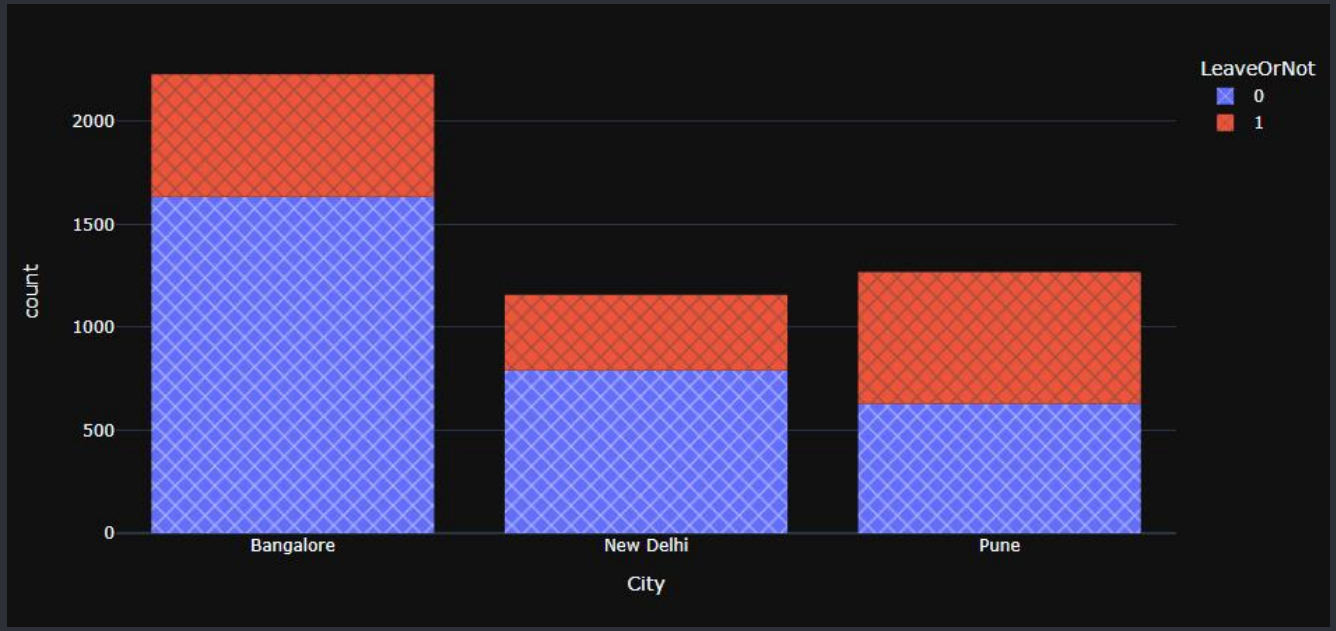
다른 학위에 비해 석사가 가장 퇴사율이 높은 것을 알 수 있습니다.

데이터 분석(JoiningYear vs LeaveOrNot)



2018년에 입사한 사람과 2015,2017년의 퇴사율이 매우 높다는 것을 알 수 있습니다.

데이터 분석(City vs LeaveOrNot)

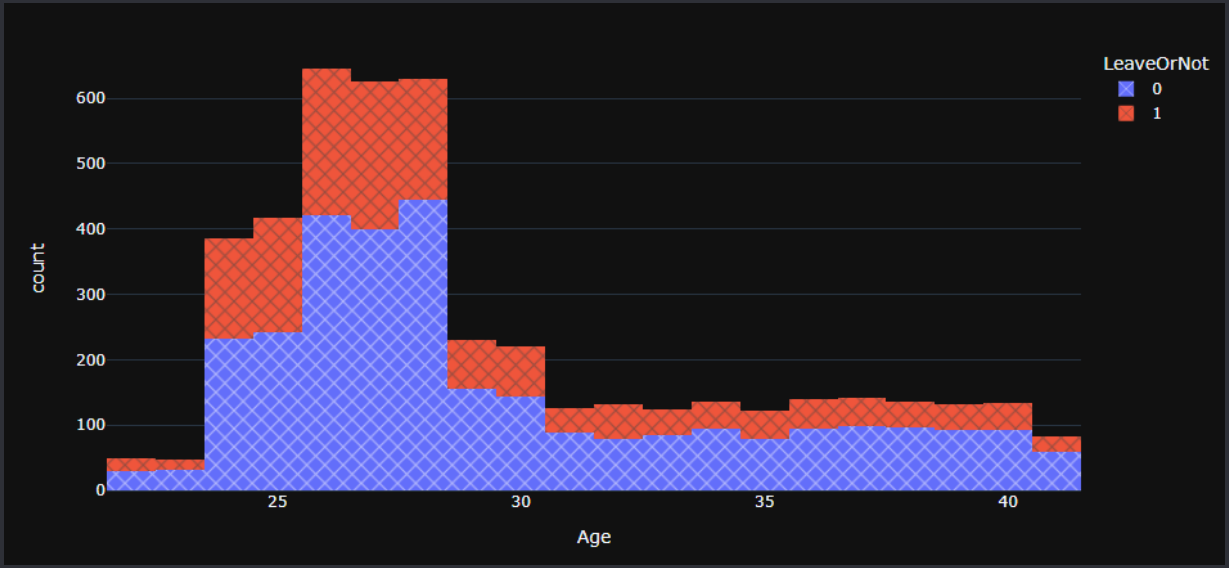


전체 도시 중 Pune가 퇴사율이 매우 높은 것을 알 수 있습니다.

데이터 분석(PaymentTier vs LeaveOrNot)

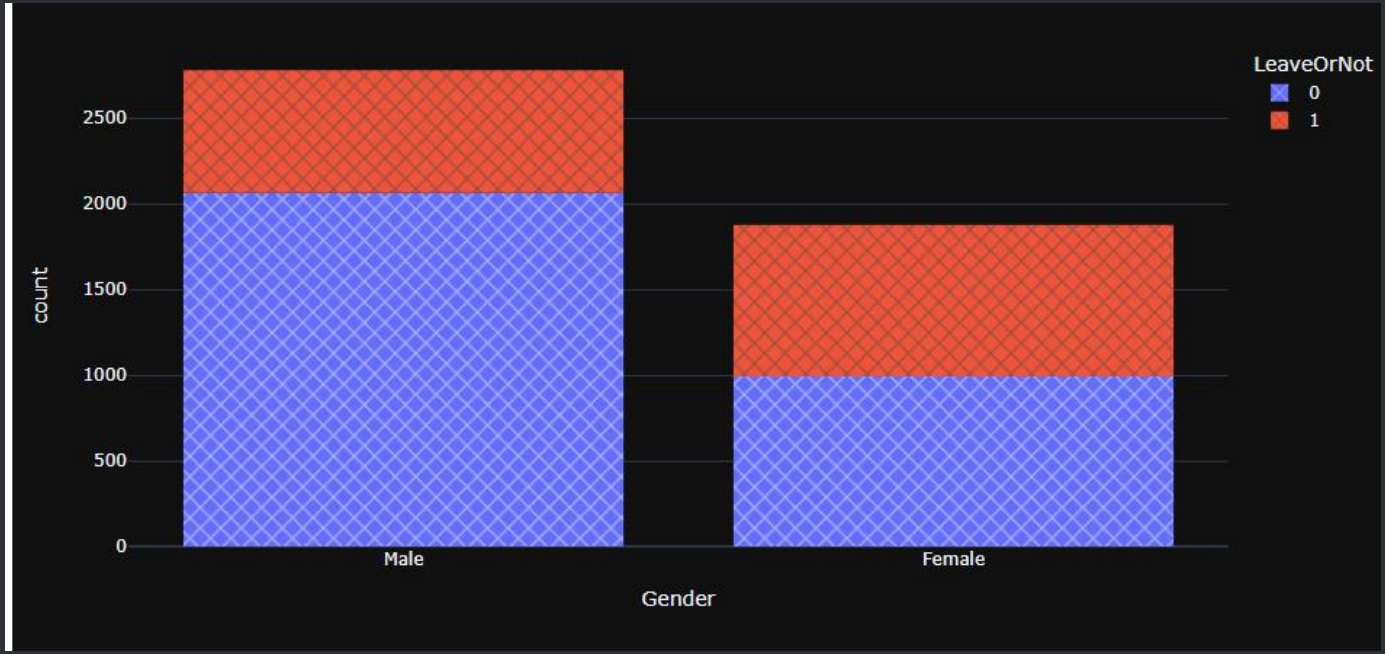
소득분위 중 2분위가 가장 퇴사율이 높은 것을 알 수 있습니다.

데이터 분석(Age vs LeaveOrNot)



23~28세가 가장 퇴사율이 높은 것을 확인 할 수 있습니다.

데이터 분석(Gender vs LeaveOrNot)



남성보다 여성의 퇴사율이 더 높은 것을 알 수 있습니다.

3 모델 학습

모델 학습(전처리)

```
df = pd.read_csv("../input/employee-future-prediction/Employee.csv")
df=pd.get_dummies(data=df,columns=['Education','City','Gender','EverBenched'],drop_first=True)
scaler = StandardScaler()
scaler.fit(df.drop('LeaveOrNot',axis = 1))
scaled_features = scaler.transform(df.drop('LeaveOrNot',axis = 1))
df_feat = pd.DataFrame(scaled_features,columns = ['JoiningYear', 'PaymentTier', 'Age', 'ExperienceInCurrentDomain',
        'Education_Masters', 'Education_PHD', 'City_New Delhi',
       'City_Pune', 'Gender_Male', 'EverBenched_Yes'])
df_feat.head()
```

```
X = df_feat
y = df['LeaveOrNot']
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=101)
```

문자로 되어 있는 값을 숫자로 인코딩과 표준화를 통하여 데이터의
범위를 같게 만들어 줍니다. 및 split를 통하여 테스트 데이터 및 학습 데이터를 나누었습니다.

모델 학습(학습 및 평가)

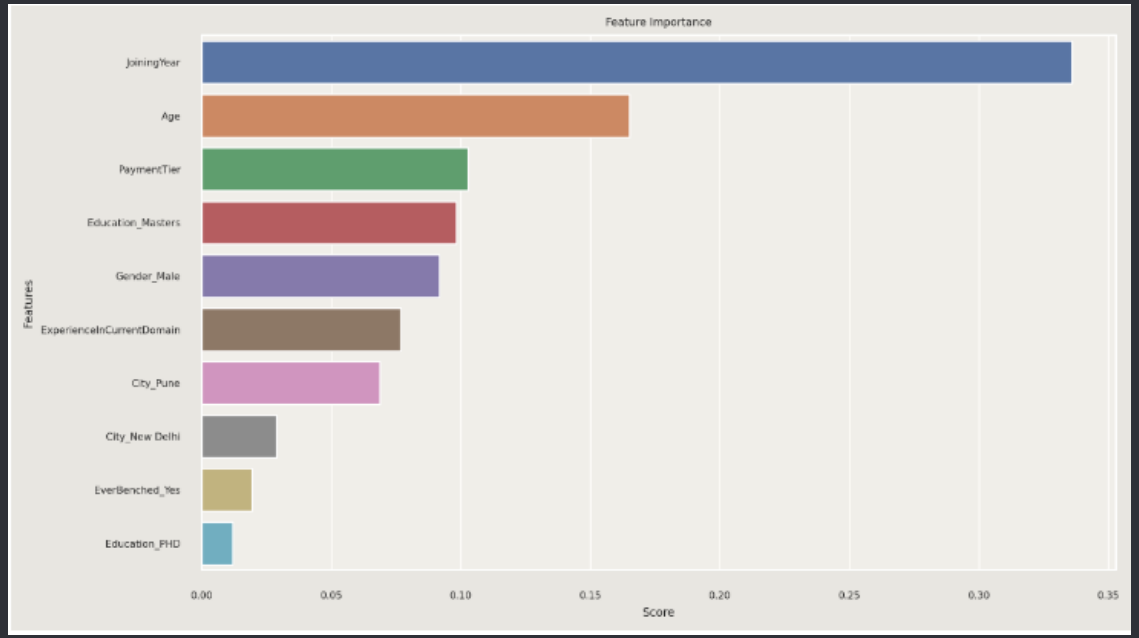
```
models=[SVC(),
        LogisticRegression(),
        dtc,KNeighborsClassifier(n_neighbors=3),
        RandomForestClassifier(n_estimators =20, random_state = 0),
        GridSearchCV(dtc, grid_params, cv = 5, n_jobs = -1, verbose = 0),
        AdaBoostClassifier(base_estimator = dtc, algorithm = 'SAMME.R', learning_rate = 0.001, n_estimators = 200),
       GradientBoostingClassifier(),
        CatBoostClassifier(iterations = 180, learning_rate = 0.1,verbose=0),
        XGBClassifier(booster = 'gblinear', learning_rate = 1, n_estimators = 10)]
for model in models:
    model.fit(X_train, y_train)
    y_pred = model.predict(X_test)
    score=accuracy_score(y_test, y_pred)
    print(type(model).__name__,":",score)
```

```
SVC : 0.8087392550143266
LogisticRegression : 0.7449856733524355
DecisionTreeClassifier : 0.7851002865329513
KNeighborsClassifier : 0.7958452722063037
RandomForestClassifier : 0.8130372492836676
GridSearchCV : 0.8359598853868195
AdaBoostClassifier : 0.7893982808022922
GradientBoostingClassifier : 0.8438395415472779
CatBoostClassifier : 0.8488538681948424
[05:53:44] WARNING: ../src/learner.cc:1095: Starti
e the old behavior.
XGBClassifier : 0.7449856733524355
```

Cat와 그라디언트 부스팅 분류기가 가장 우수한 결과를 나온것을 확인 할 수 있습니다.

4 결론

결론



퇴사율과 가장 연관이 높은 것은 입사일이라는 것을 알 수 있습니다.

5 느낀 점

느낀점

- 이번 발표를 준비하면서 내가 알는 것 보다 더 많은 학습 방법이 있다는 것을 알게 되었고 EDA를 통하여 데이터를 여러 방면으로 확인할 수 있다는 것을 알게 되어 좋은 경험이 되었다고 생각합니다.