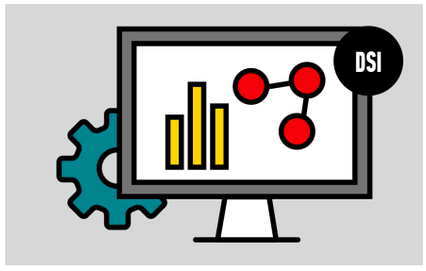
Attrition Prediction

(resignation prediction)

Atigon Hongchumpol / Mar 1, 2023



GENERAL ASSEMBLY

Problem Statement

Problem statement

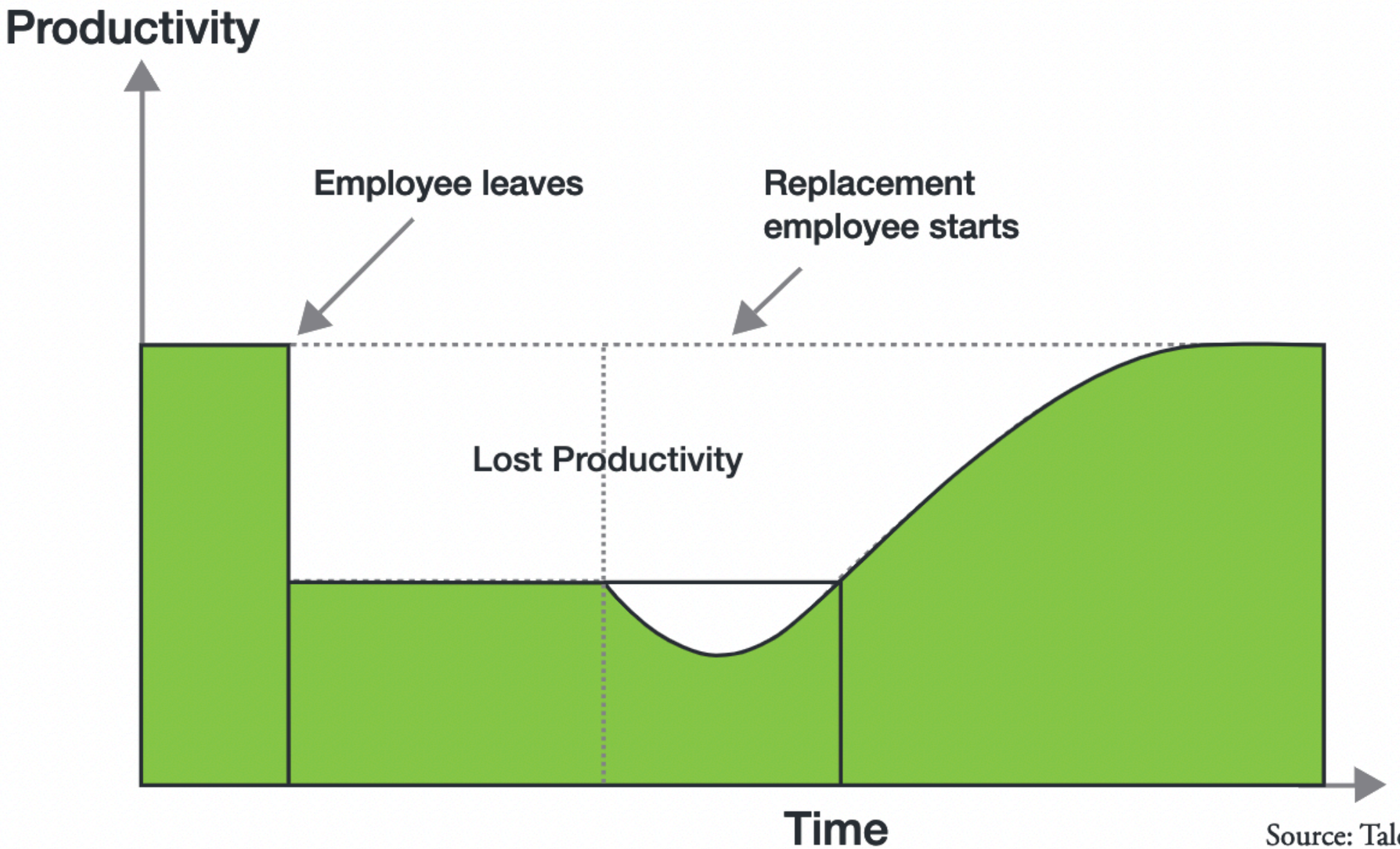
- Employee attrition is a significant challenge for many organizations, leading to increased costs and decreased productivity.

- The sample company has experienced a high employee turnover rate, and we want to develop an attrition prediction model that can identify employees at risk of leaving the company.

- This project aims to develop a predictive model that can accurately identify employees who are likely to leave, allowing the company to take proactive measures to retain valuable employees and reduce the negative impact of attrition on the organization.

- Expected 20% F1-score improvement from baseline score.

Productivity Losses Tied to Turnover



Source: Taleo Research

Data collection & Cleaning

Data collection & Cleaning

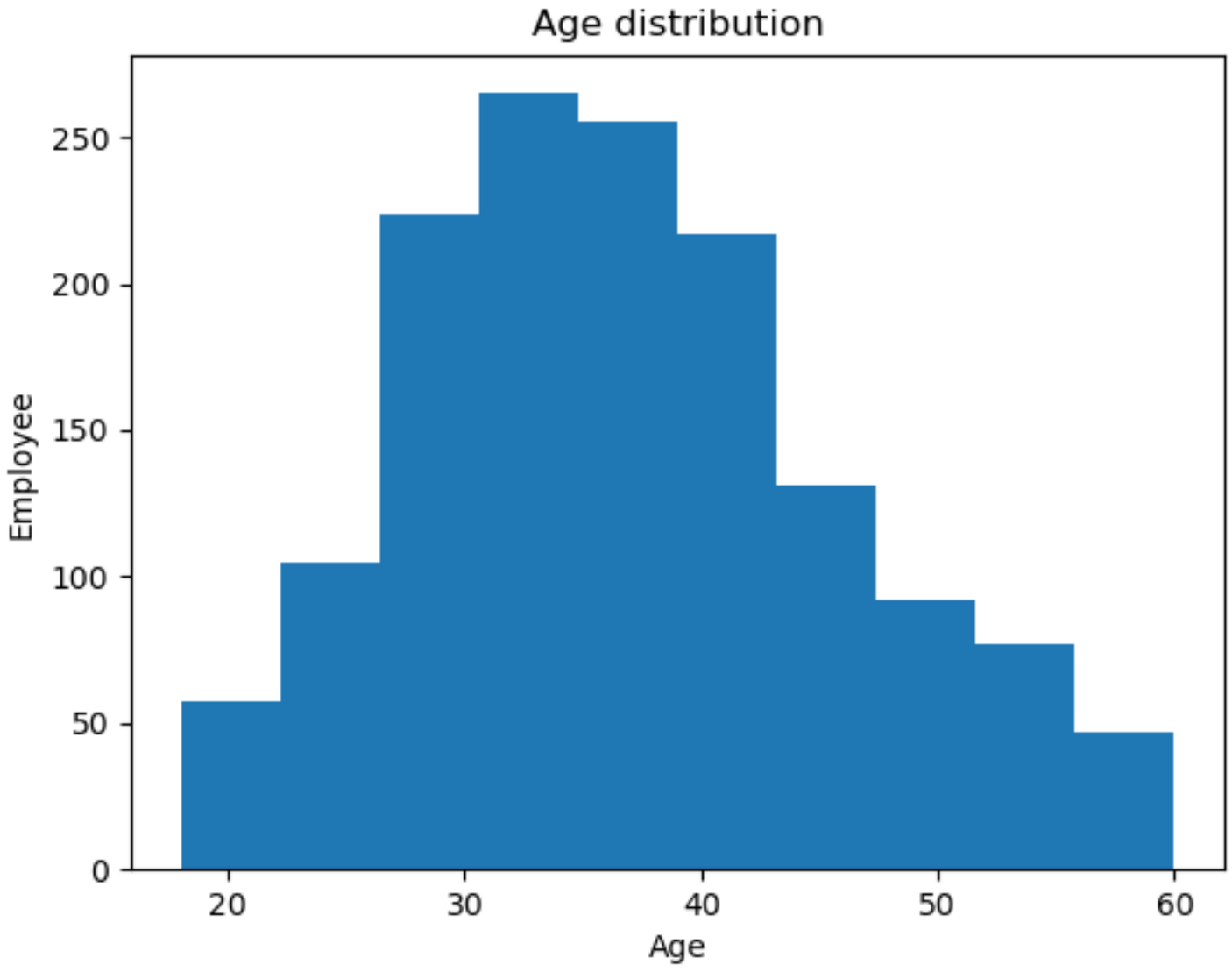
- The dataset was obtained from Kaggle, a popular platform for data science competitions and data analysis projects. The dataset was provided by Pavansubhash and is publicly available for download on the Kaggle website.
- The dataset contains 1470 rows and 35 columns, each row representing each employee.

Data cleaning

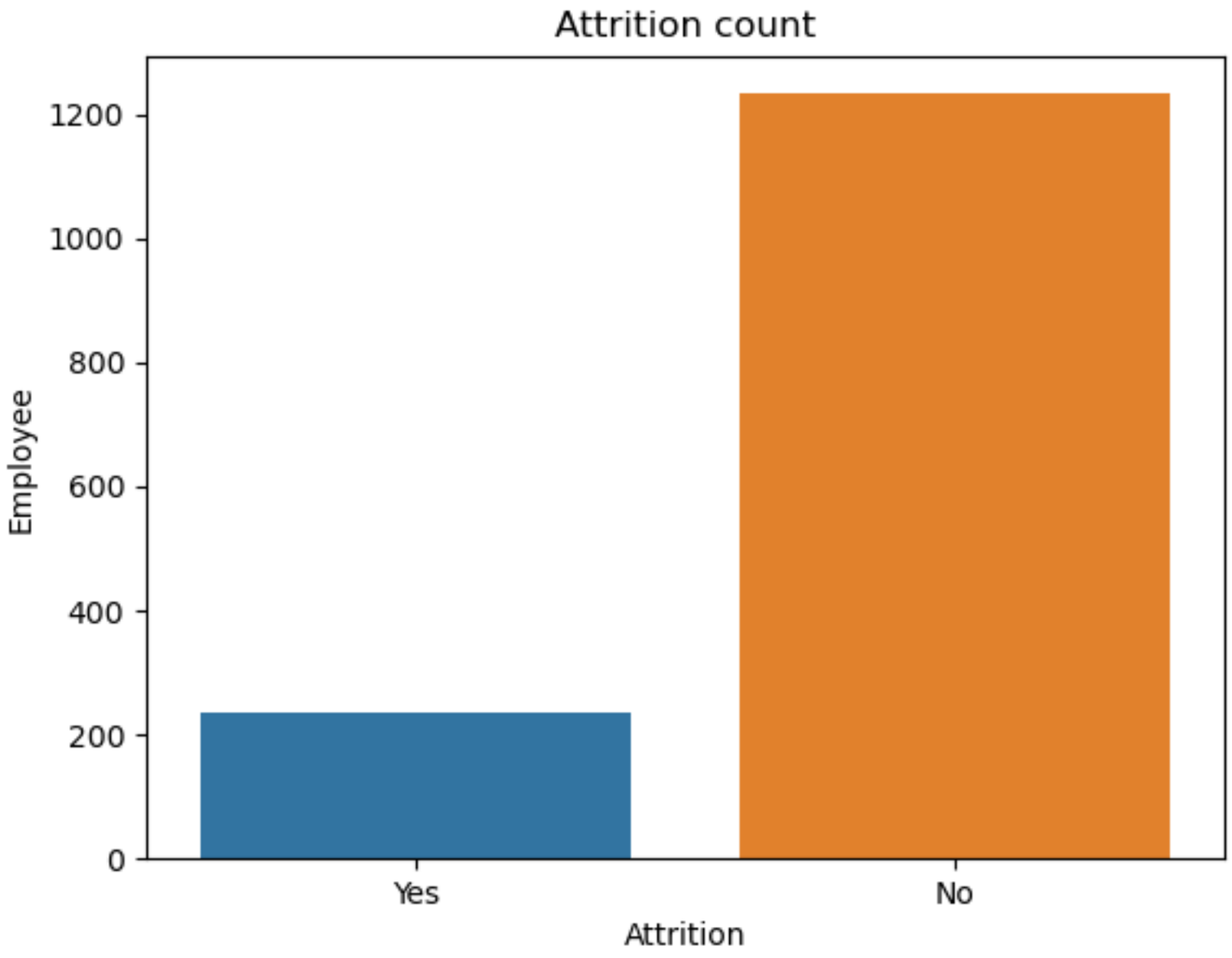
- This dataset is clean and ready for analysis, as it does not contain any null values or missing data. Therefore, no data cleaning is required prior to analysis.

Exploratory Data Analysis

EDA

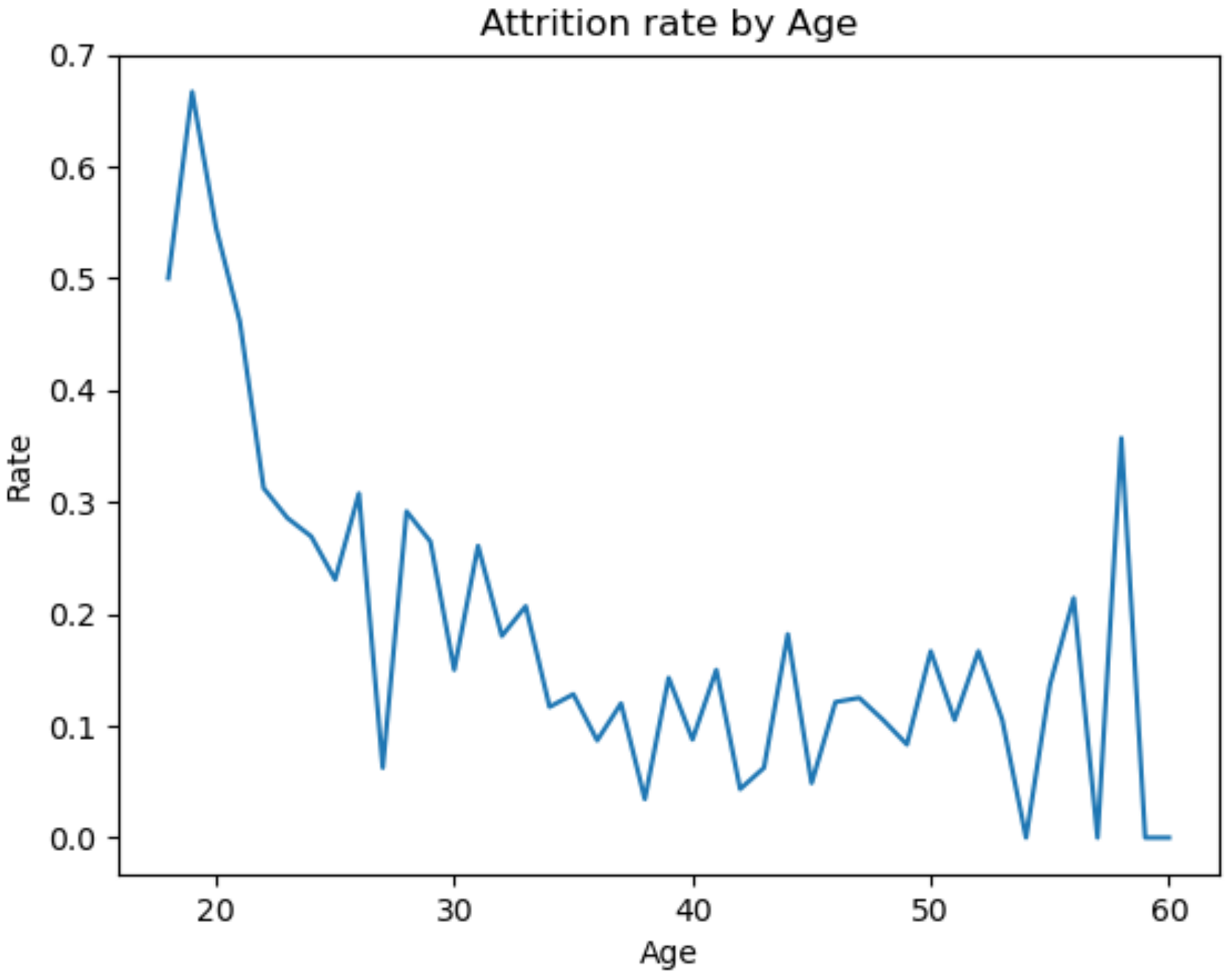

Age distribution
Employee
250
200
150
100
50
0
20
30
40
50
60
Age

EDA

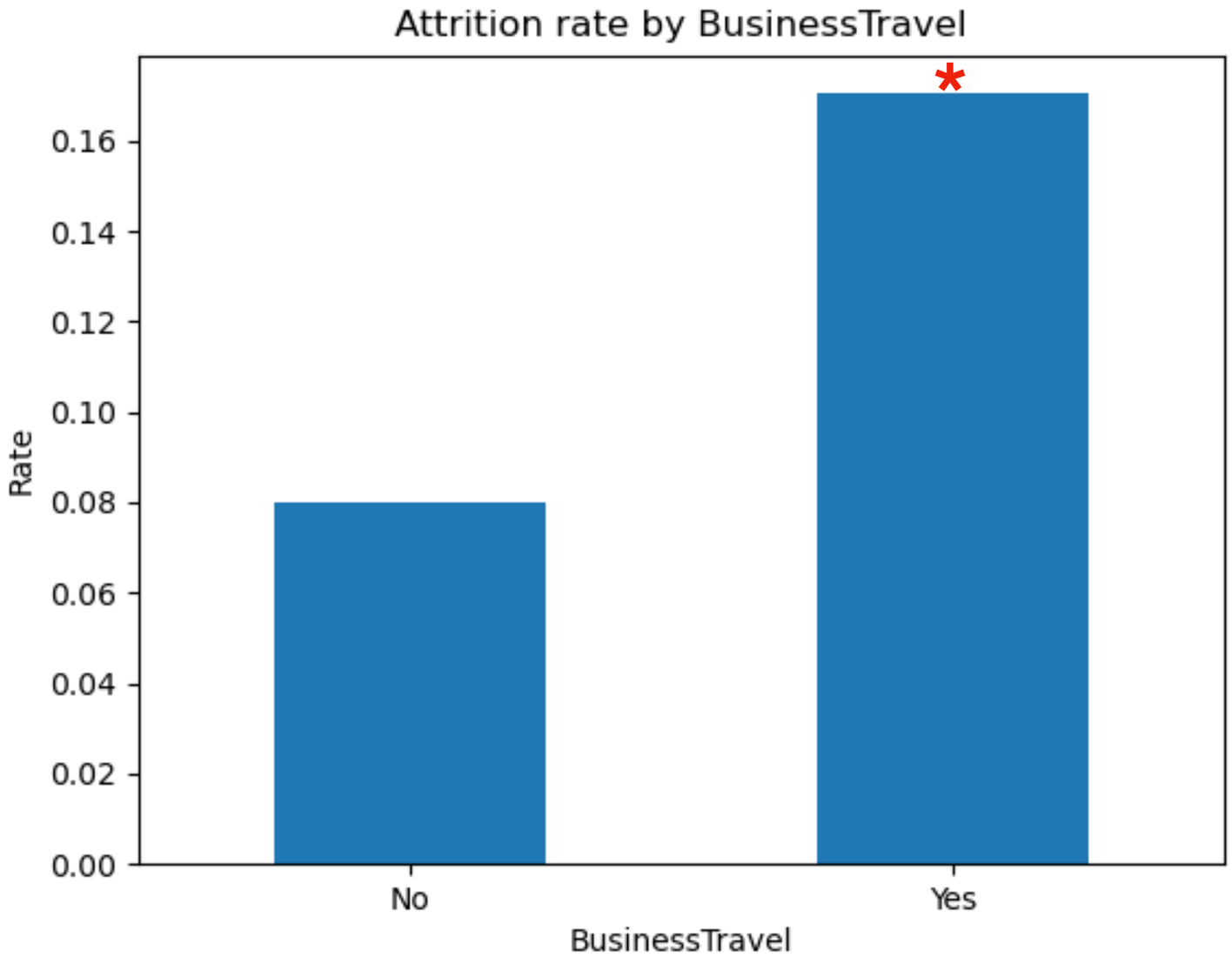

Attrition count
1200
1000
800
600
400
200
0
Employee
Yes
No
Attrition

16% **84%**

EDA

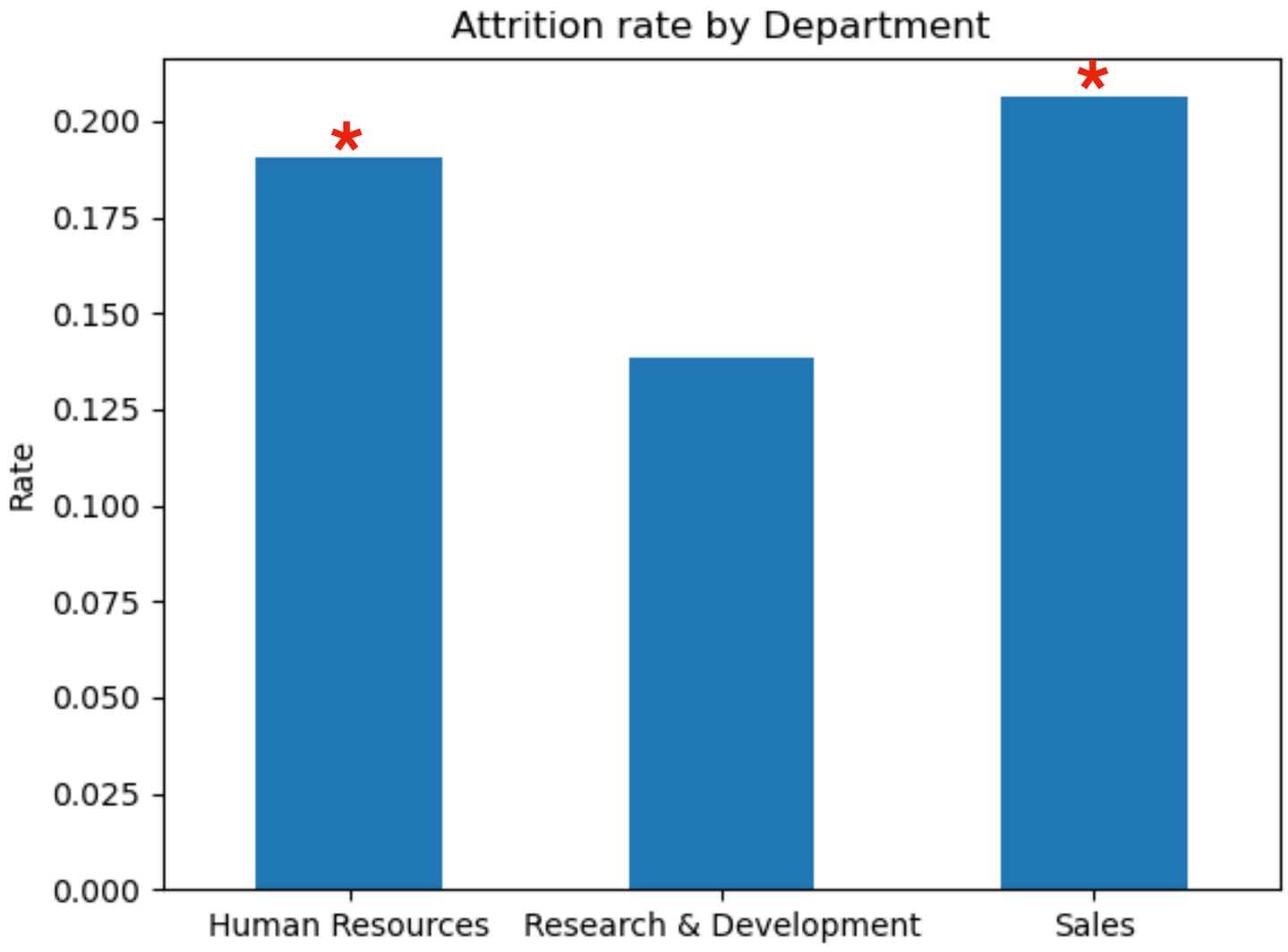

Attrition rate by Age
Rate
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
20
30
40
50
60
Age

EDA

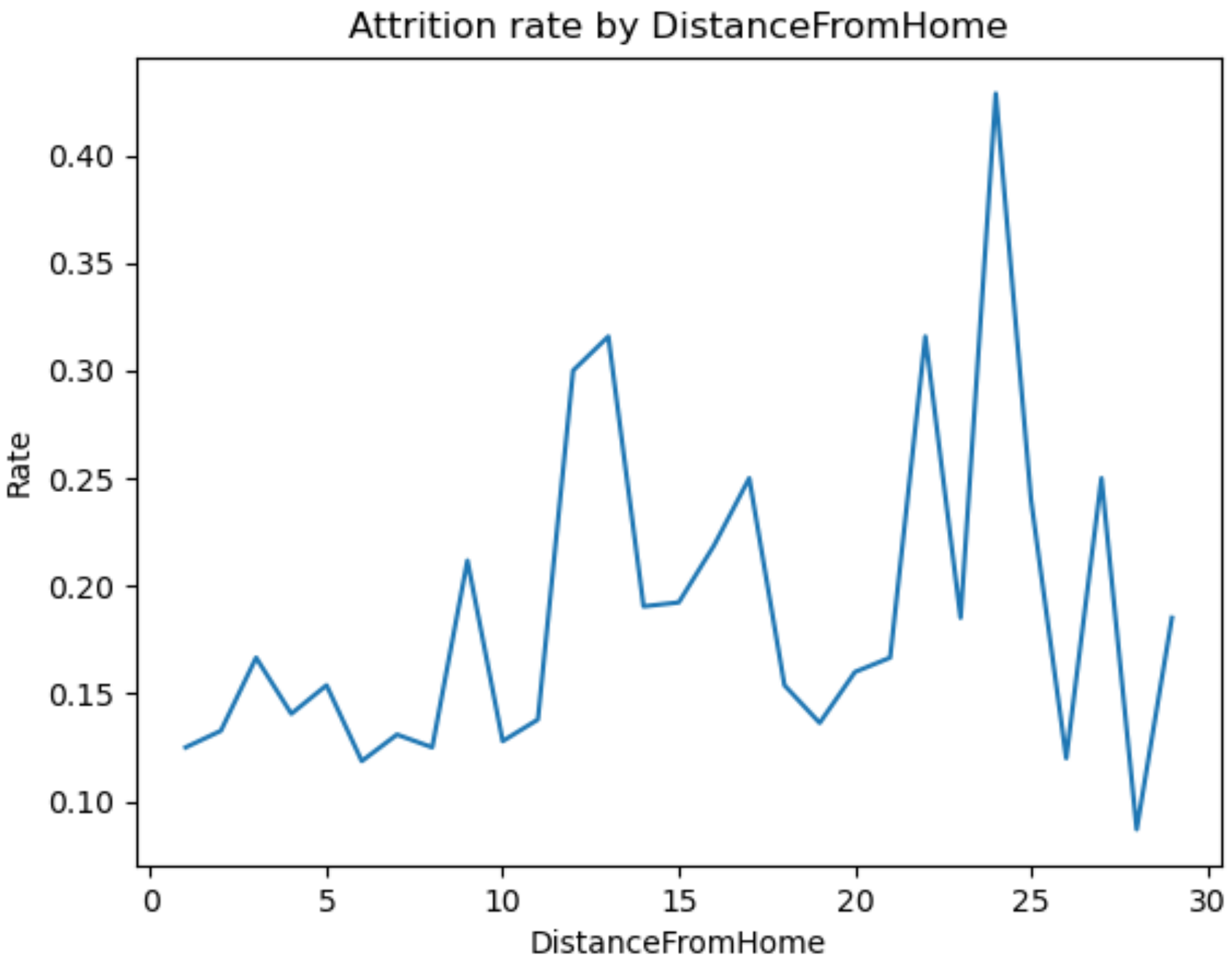

Attrition rate by BusinessTravel
Rate
0.16
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
No
Yes
BusinessTravel

2X

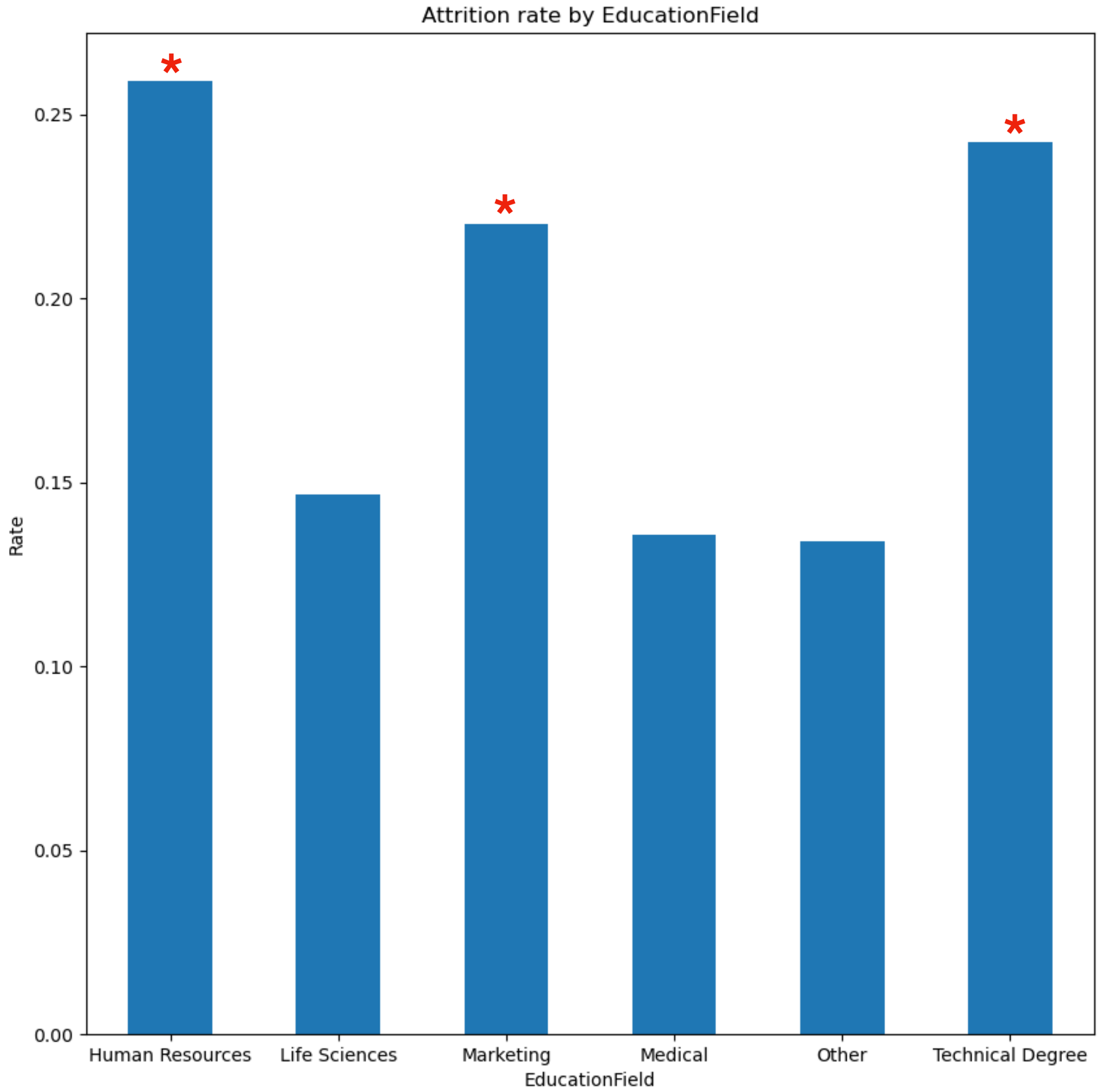
EDA


Attrition rate by Department
Rate
0.200
0.175
0.150
0.125
0.100
0.075
0.050
0.025
0.000
*
*
Human Resources
Research & Development
Sales

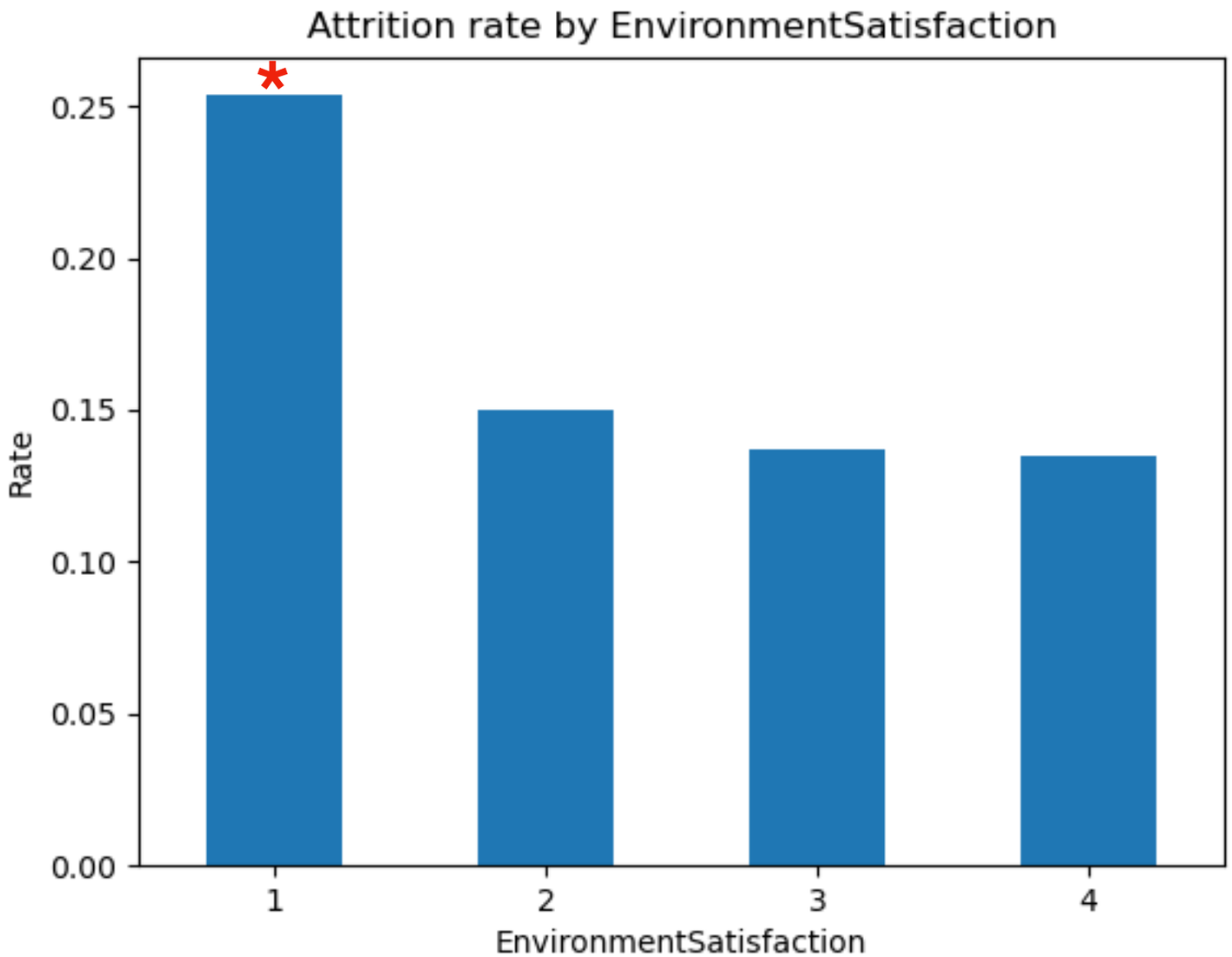
EDA


Attrition rate by DistanceFromHome
Rate
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0
5
10
15
20
25
30
DistanceFromHome

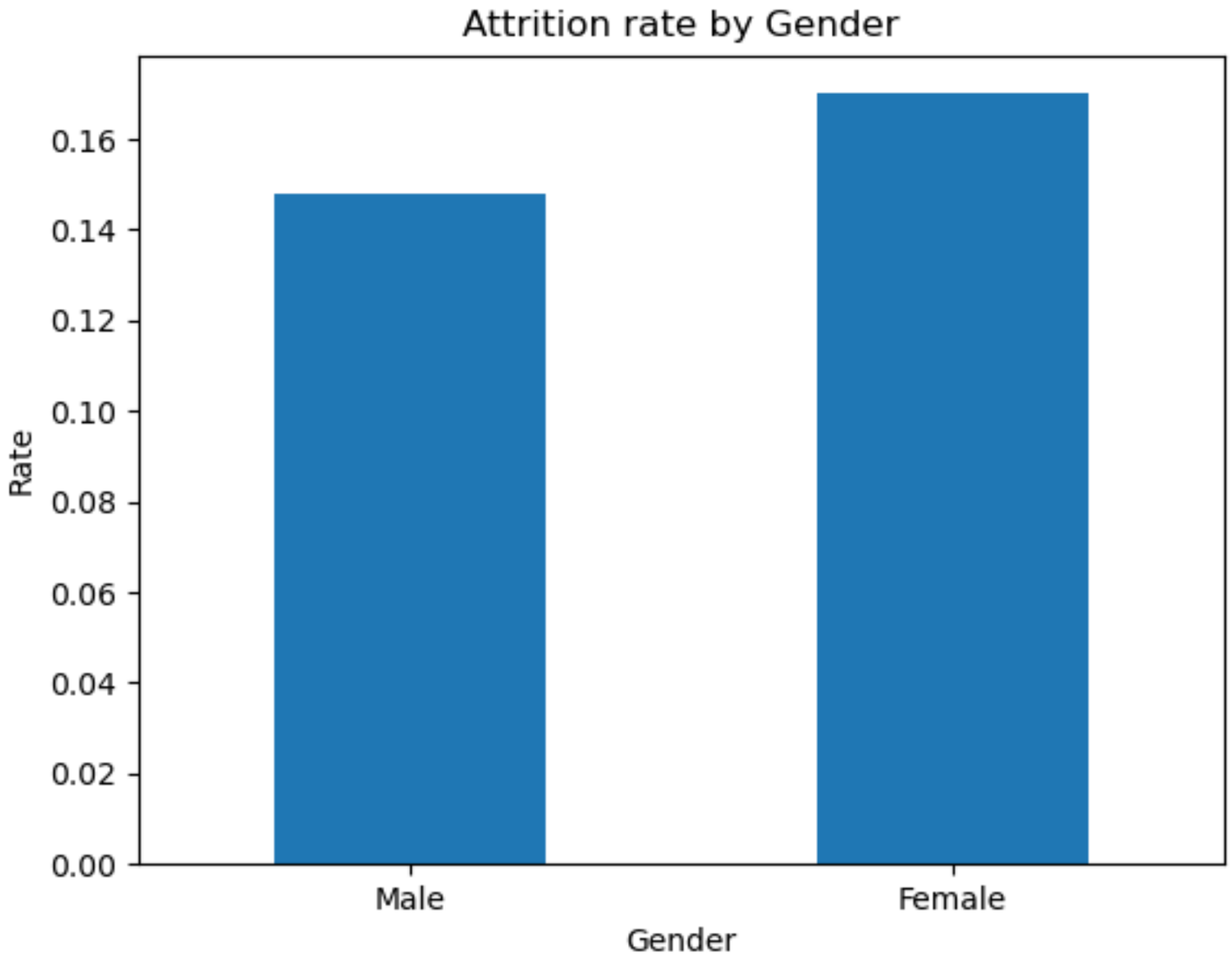
EDA


Attrition rate by EducationField
0.25
0.20
0.15
0.10
0.05
0.00
Rate
*
*
*
Human Resources
Life Sciences
Marketing
Medical
Other
Technical Degree
EducationField

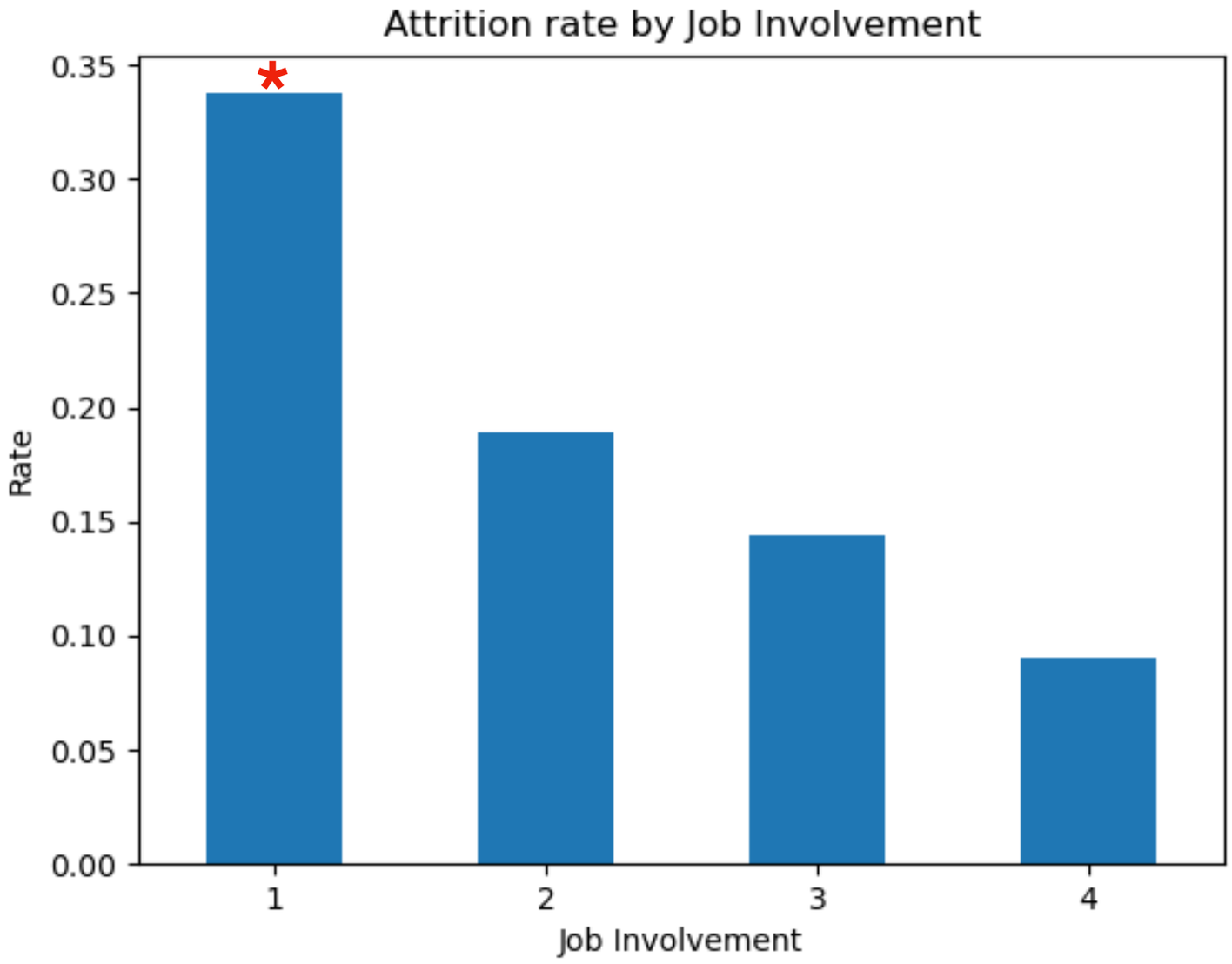
EDA


Attrition rate by EnvironmentSatisfaction
*
0.25
0.20
0.15
0.10
0.05
0.00
Rate
1
2
3
4
EnvironmentSatisfaction

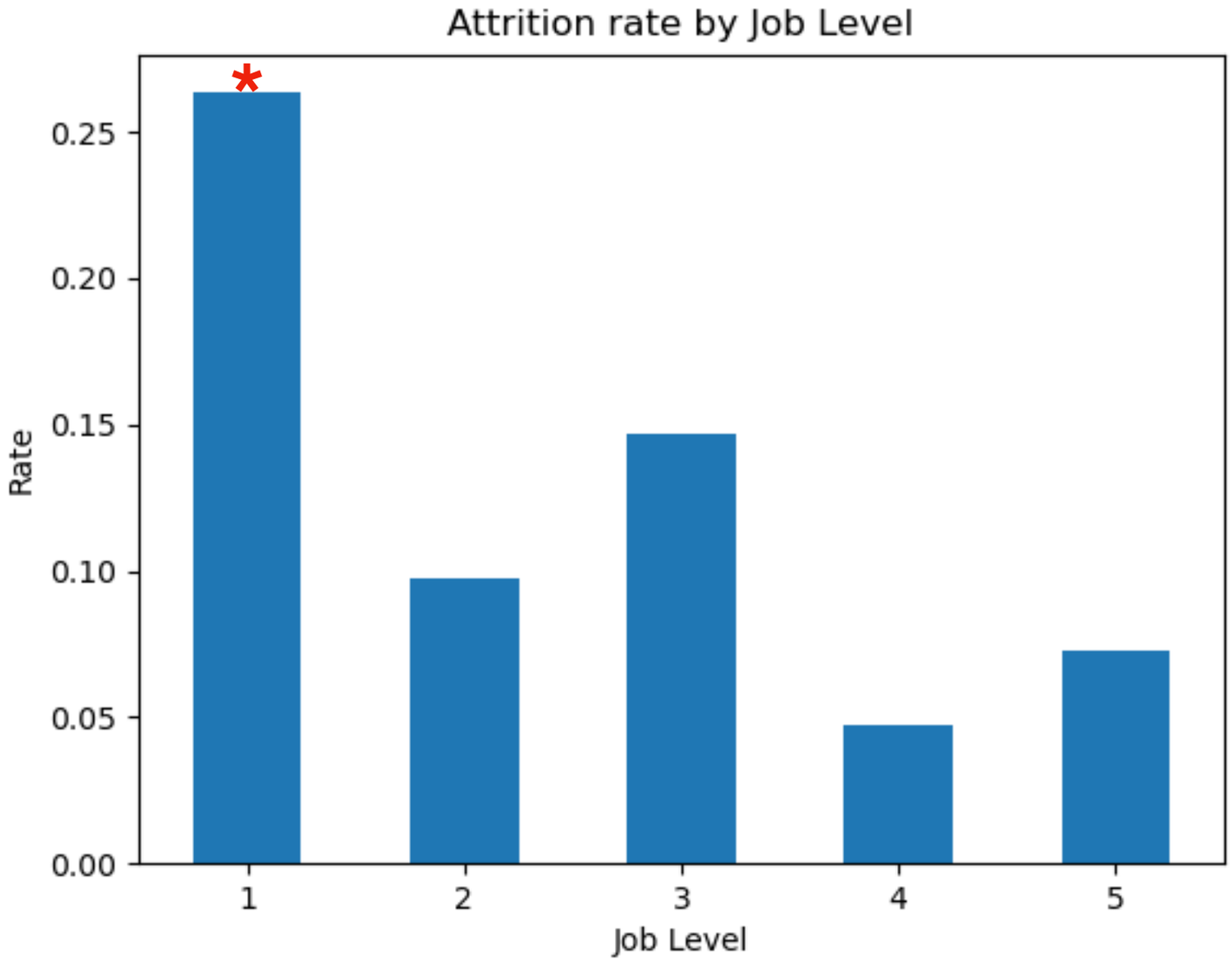
EDA


Attrition rate by Gender
Rate
0.16
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
Male
Female
Gender

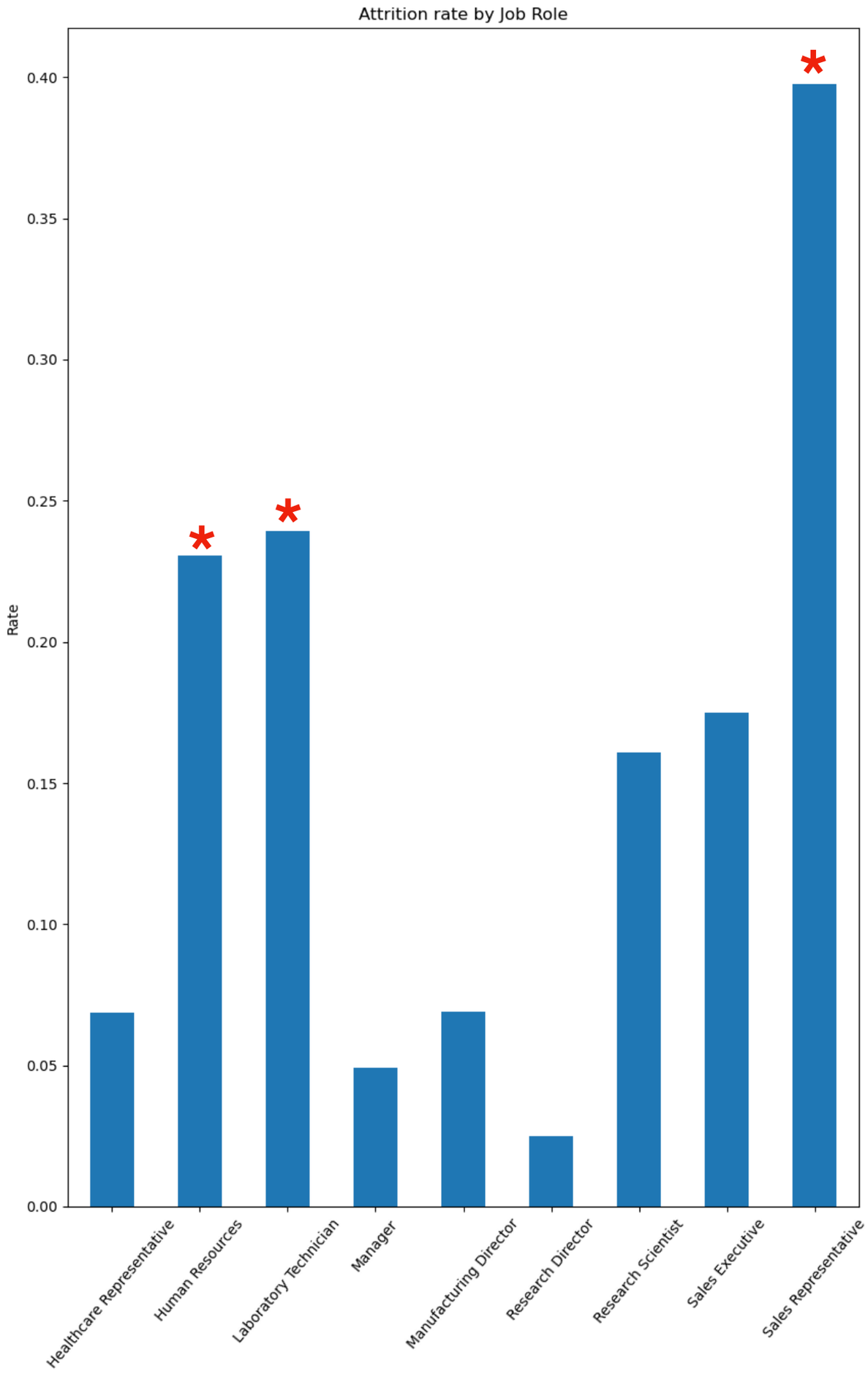
EDA


Attrition rate by Job Involvement
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
Rate
1
2
3
4
Job Involvement

EDA


Attrition rate by Job Level
*
0.25
0.20
0.15
0.10
0.05
0.00
Rate
1
2
3
4
5
Job Level

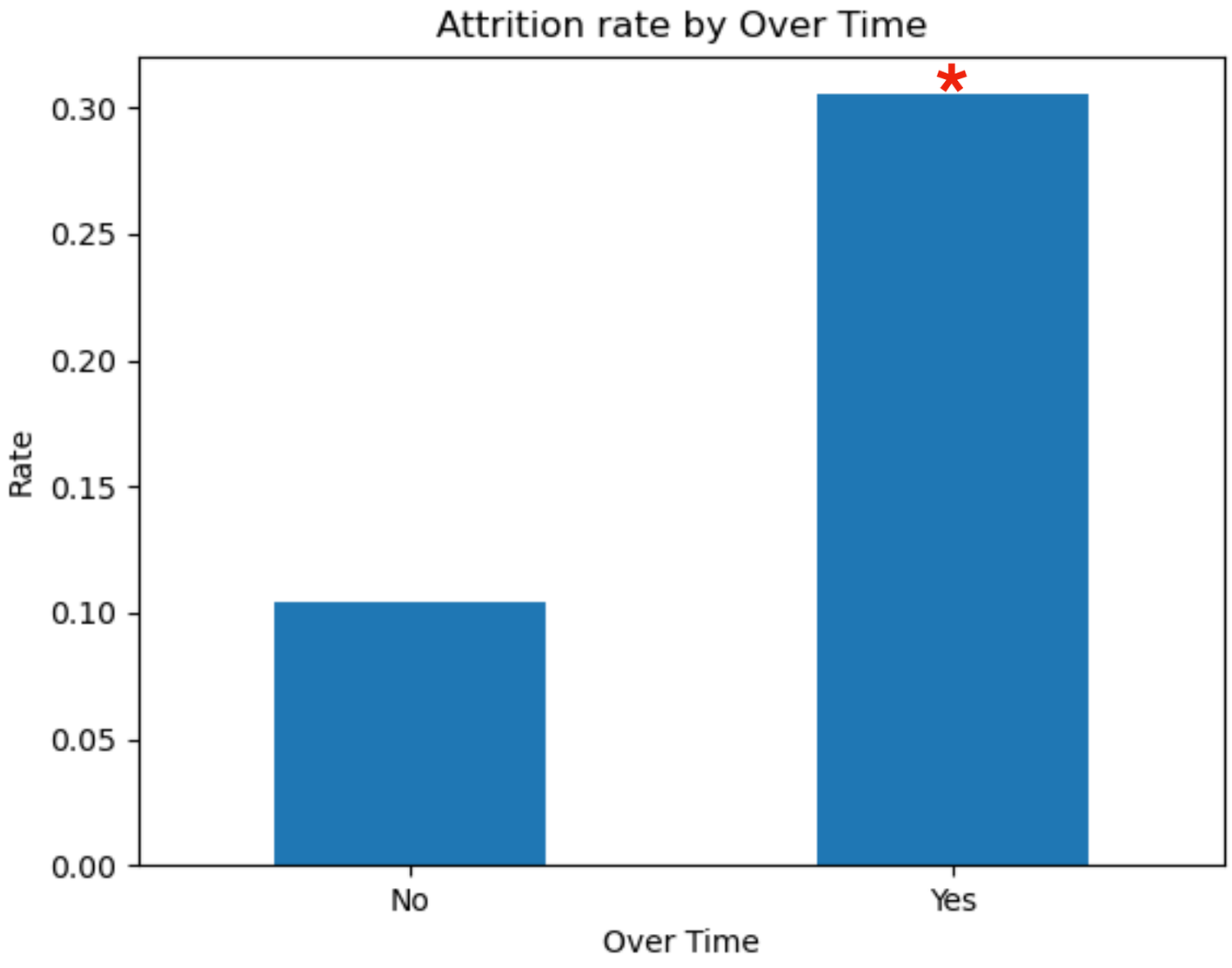
EDA


Attrition rate by Job Role
Rate
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
*
*
*
Healthcare Representative
Human Resources
Laboratory Technician
Manager
Manufacturing Director
Research Director
Research Scientist
Sales Executive
Sales Representative

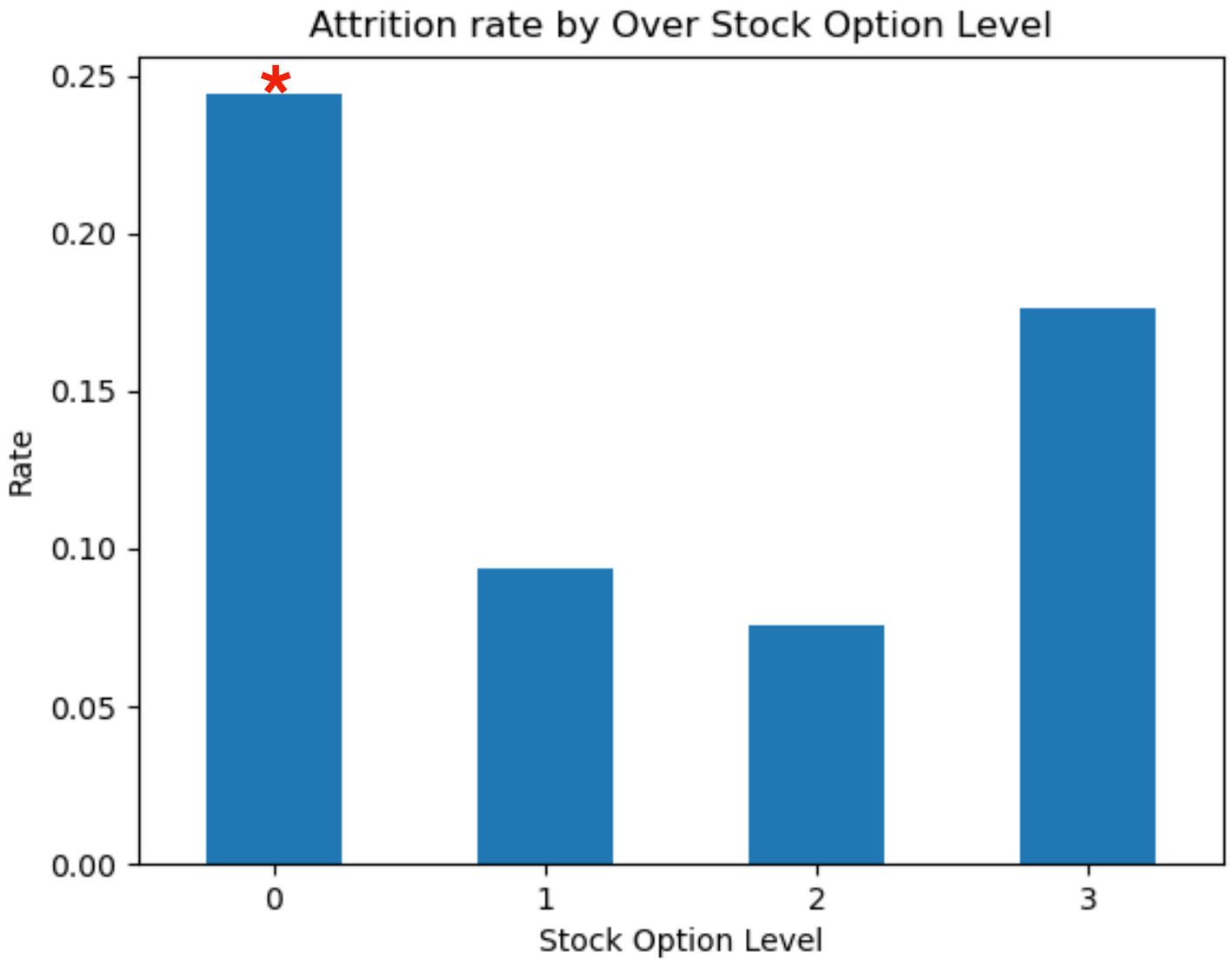
EDA

Attrition rate by Salary
Rate
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
2500
5000
7500
10000
12500
15000
17500
Salary

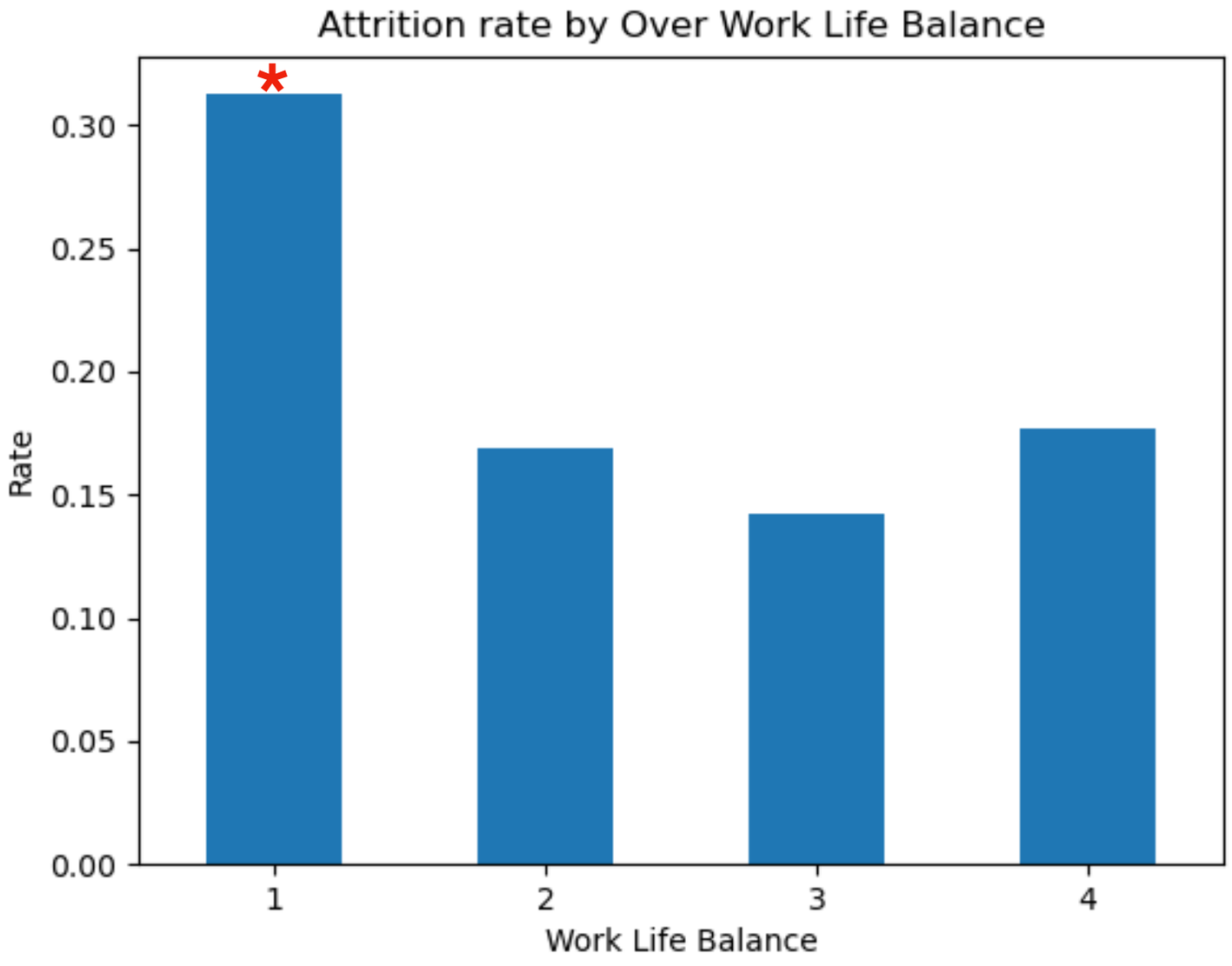
EDA


Attrition rate by Over Time
Rate
0.30
0.25
0.20
0.15
0.10
0.05
0.00
No
Yes
Over Time
*

EDA


Attrition rate by Over Stock Option Level
0.25
0.20
0.15
0.10
0.05
0.00
Rate
*
0
1
2
3
Stock Option Level

EDA


Attrition rate by Over Work Life Balance
*
0.30
0.25
0.20
0.15
0.10
0.05
0.00
Rate
1
2
3
4
Work Life Balance

Baseline model

Baseline model

	model	test_accuracy	test_recall	tn	fp	fn	tp	precision	F1-score
0	LogisticRegression	0.846154	0.055556	185	0	34	2	1.000000	0.105263
1	RandomForestClassifier	0.859729	0.194444	183	2	29	7	0.777778	0.311111
2	AdaBoostClassifier	0.855204	0.361111	176	9	23	13	0.590909	0.448276
3	KNeighborsClassifier	0.782805	0.027778	172	13	35	1	0.071429	0.040000

- The best baseline score came from the AdaBoost model
    - Recall 36%
    - Precision 59%
    - F1-score 45%
- Our goal is to improve F1-score by 20%

Feature selection model

Selected features

```
selected_features = [
    'Age',
    'BusinessTravel',
    'Department',
    'DistanceFromHome',
    'EducationField',
    'EnvironmentSatisfaction',
    'JobInvolvement',
    'JobLevel',
    'JobRole',
    'JobSatisfaction',
    'MaritalStatus',
    'salary_range',
    'NumCompaniesWorked',
    'num_comp_work',
    'OverTime',
    'RelationshipSatisfaction',
    'StockOptionLevel',
    'TotalWorkingYears',
    'TrainingTimesLastYear',
    'WorkLifeBalance',
    'YearsAtCompany',
    'YearsInCurrentRole',
    'YearsWithCurrManager'
]
```

Feature selection model

	model	test_accuracy	test_recall	tn	fp	fn	tp	precision	F1-score
0	LogisticRegression	0.855204	0.085714	186	0	32	3	1.000000	0.157895
1	RandomForestClassifier	0.850679	0.142857	183	3	30	5	0.625000	0.232558
2	AdaBoostClassifier	0.873303	0.428571	178	8	20	15	0.652174	0.517241
3	KNeighborsClassifier	0.832579	0.142857	179	7	30	5	0.416667	0.212766

- The best production score came from the AdaBoost model
  - Recall 43% (baseline 36%)
  - Precision 65% (baseline 59%)
  - F1-score 52% (baseline 45%) **+15%**

Hyper Tuning &

Final Model

Best hyper tuning model

	model	test_accuracy	test_recall	tn	fp	fn	tp	precision	f1-score
0	AdaBoostClassifier	0.895928	0.4	184	2	21	14	0.875	0.54902

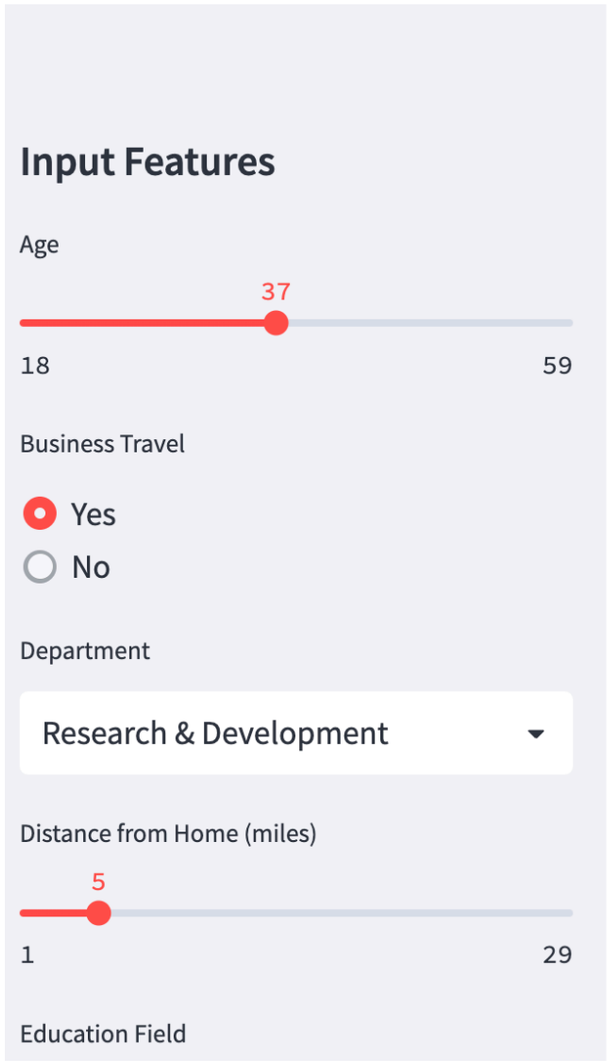
- The best production score came from the AdaBoost model
    - Recall 40% (baseline 36%)
    - Precision 88% (baseline 59%)
    - F1-Score 55% (baseline 45%) **+22%**

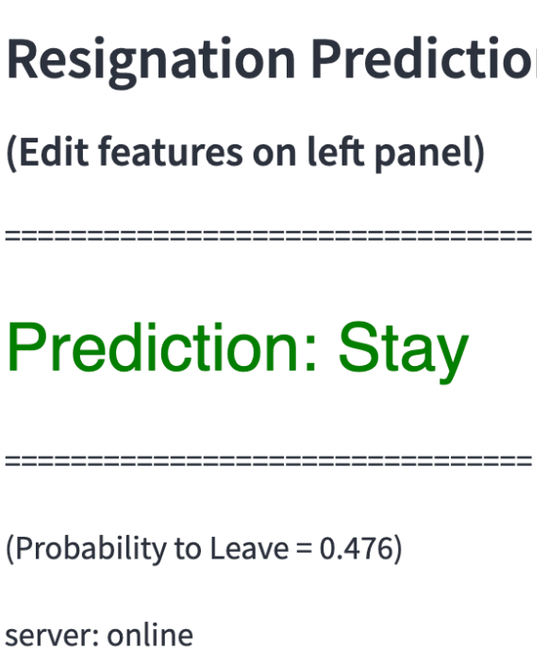
Streamlit

Streamlit

Online

- https://atigonh-resignation-streamlitattrition-app-lgwauk.streamlit.app/





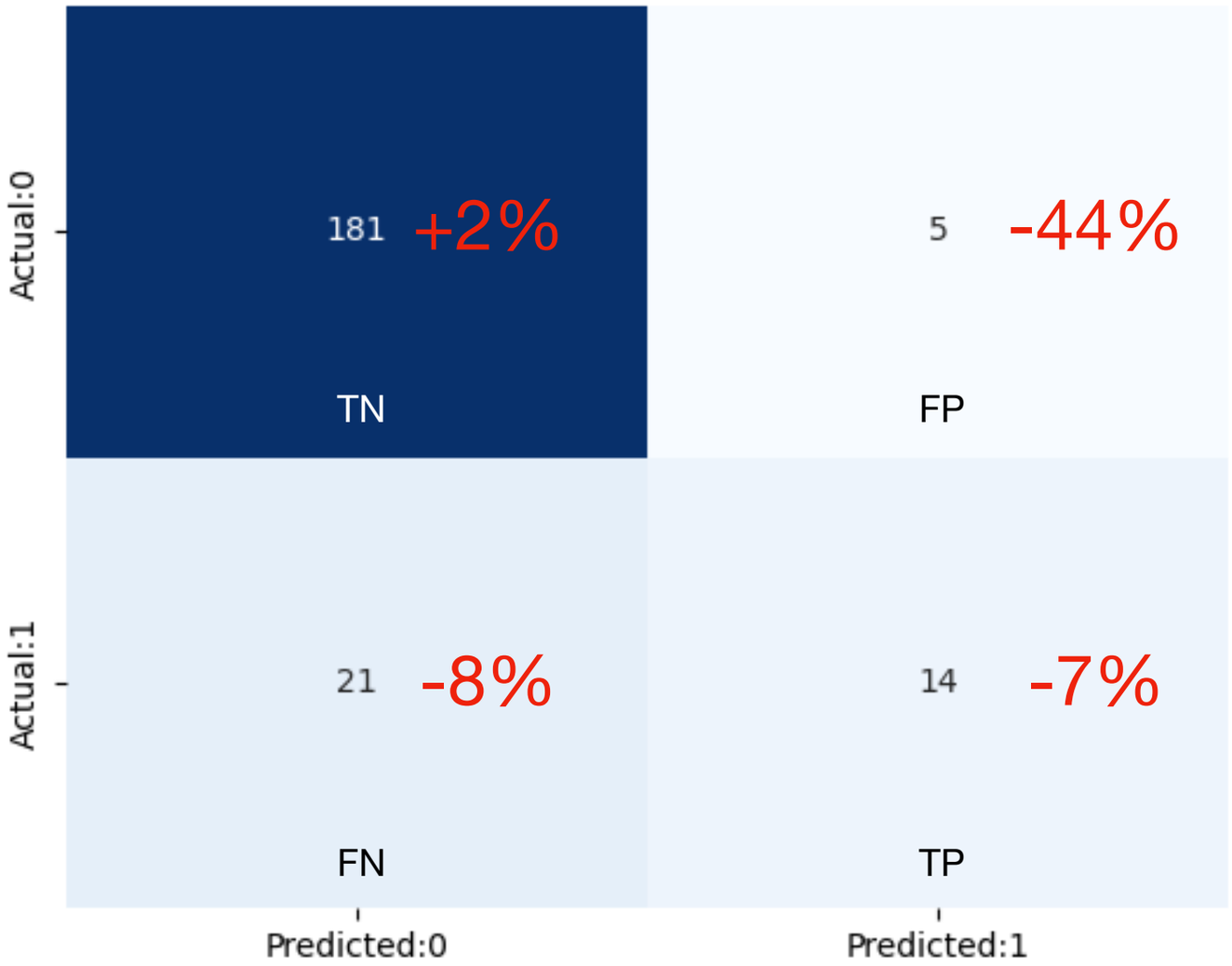
	0
Age	37
BusinessTravel	1
Department	Research & Development
DistanceFromHome	5
EducationField	Life Sciences
EnvironmentSatisfaction	3
JobInvolvement	3
JobLevel	2
JobRole	Sales Executive
JobSatisfaction	3
MaritalStatus	0
salary_range	6250
NumCompaniesWorked	2

Evaluation

Evaluation

Score type	Baseline	Production	% of improvement
Recall	0.36	0.40	11%
Precision	0.59	0.88	49%
F1-score	0.45	0.55	22%

Evaluation


Actual:0
181
+2%
TN
5
-44%
FP
Actual:1
21
-8%
FN
14
-7%
TP
Predicted:0
Predicted:1

Summary & Recommendations

Summary

- The model met the expectations by improving the F1-Score by 22%. (Expected 20%)
- Precision is 88% (if you guess 9 leave, 8 will be correct)

Recommendation

- Any future improvements should prioritize high precision and acceptable recall.
- To approach the employee, using this model carefully and with caution is important.

Thank you

Atigon Hongchumpol / Mar 1, 2023

- **Source**

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

- Image: https://silo.tips/download/hidden-roi-of-talent-acquisition-and-mobility

- **Image:** https://www.google.co.th/url?sa=i&url=https%3A%2F%2Fwww.pinterest.com%2Fpin%2F678917712545864019%2F&psig=AOvVaw32QiSD437M9wg29P3NnISQ&ust=1677346125350000&source=images&cd=vfe&ved=0CA8QjRxqFwoTCOjkvbHXrv0CFQAAAAAdAAAAABAE

- ChatGPT

- https://www.kaggle.com/code/otasra/eda-xgboost-randomforest-performance-tuning