HR ANALYSIS CASE STUDY

Sicheng Yang
DSI
Oct 20

github.com/kkarzyang/Presentation

INTRO

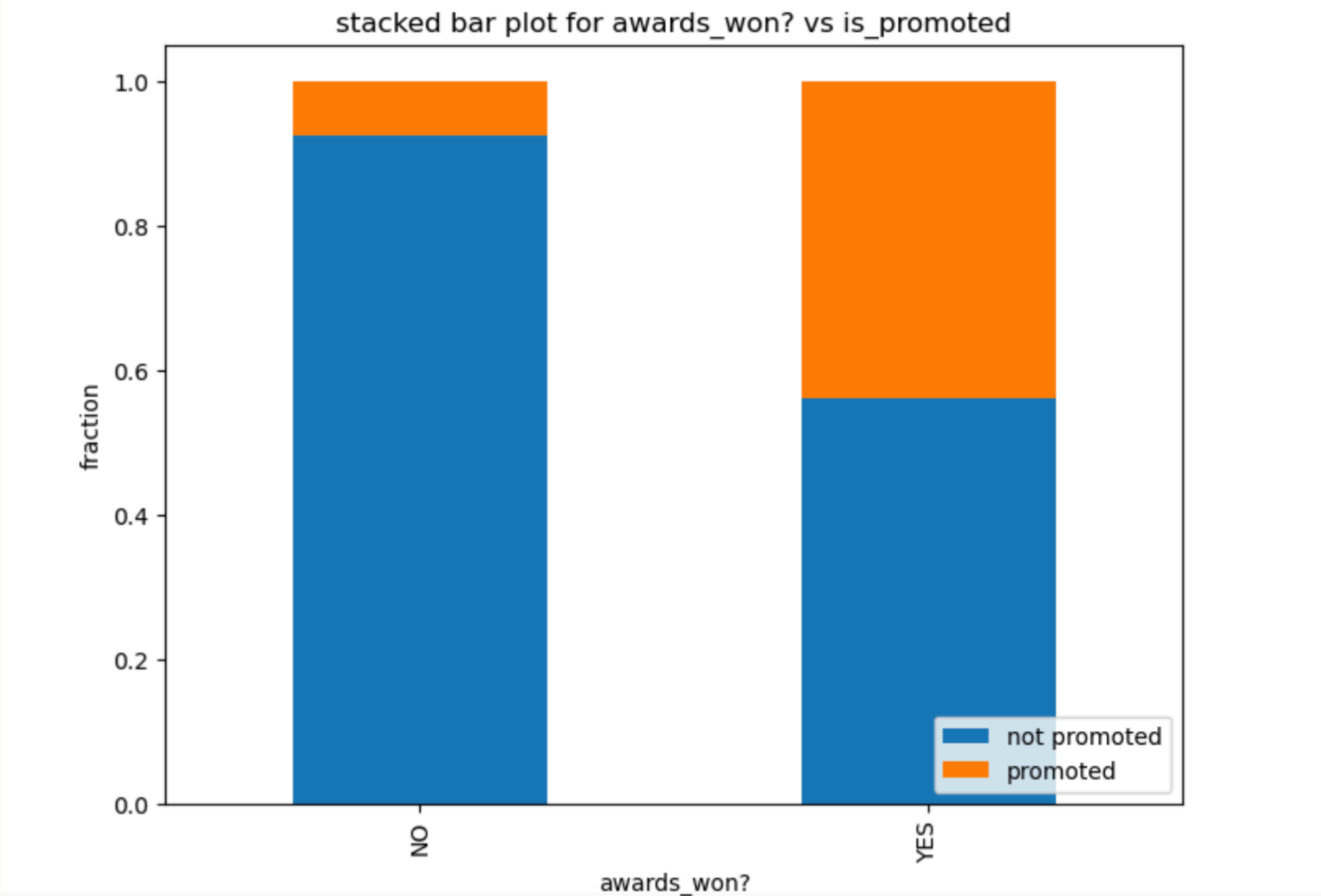
- Predict promotion or not
- Sociology research, junior employees, current students
- Classification problem
- Kaggle
- Collected from HR datasets (by shivan kumar)

EDA

- IDs get dropped
- Features: 12
- Target Variable: *is_promoted*

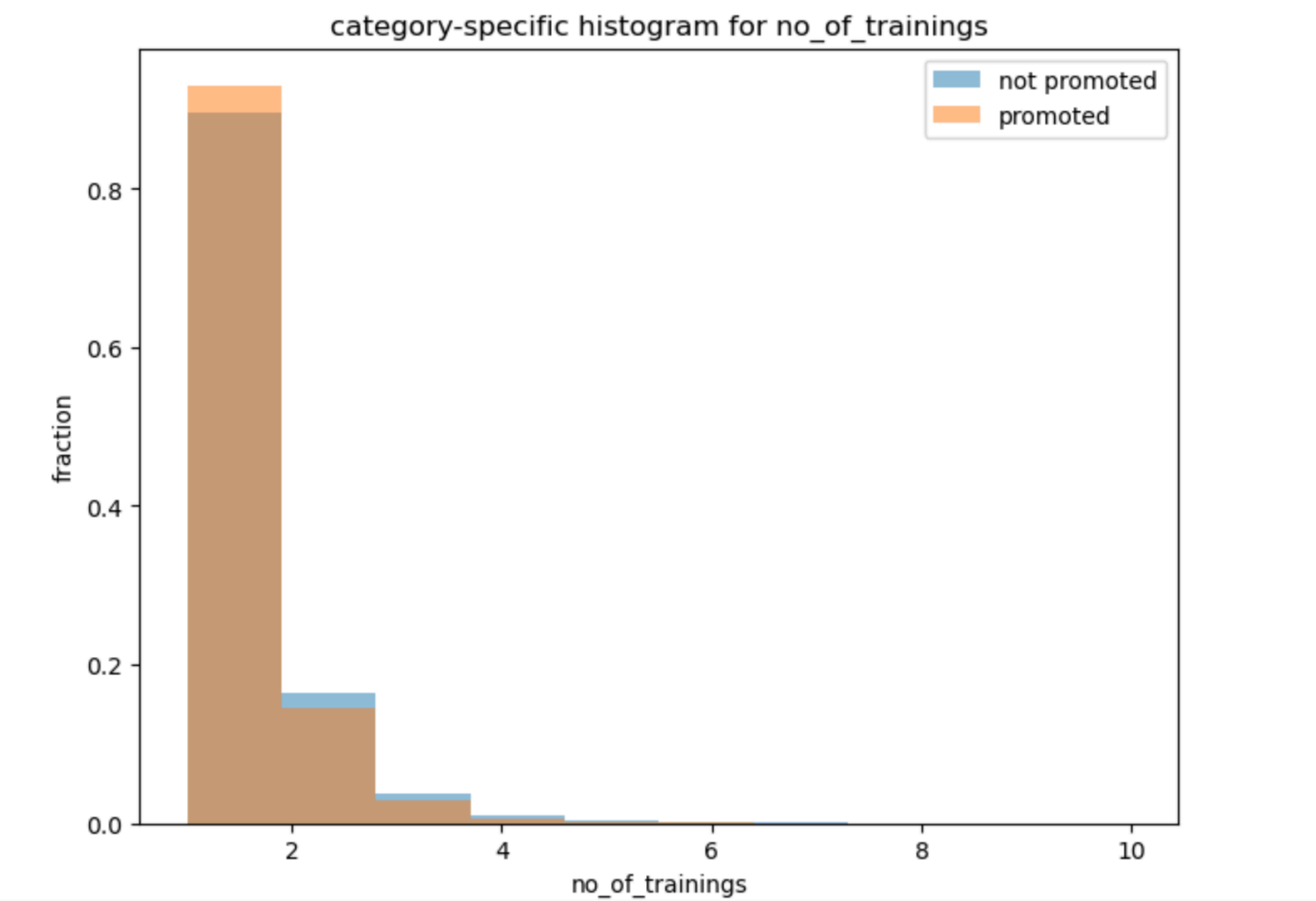
is_promoted	0	1
awards_won?		
0	0.923251	0.076749
1	0.559843	0.440157

Why?



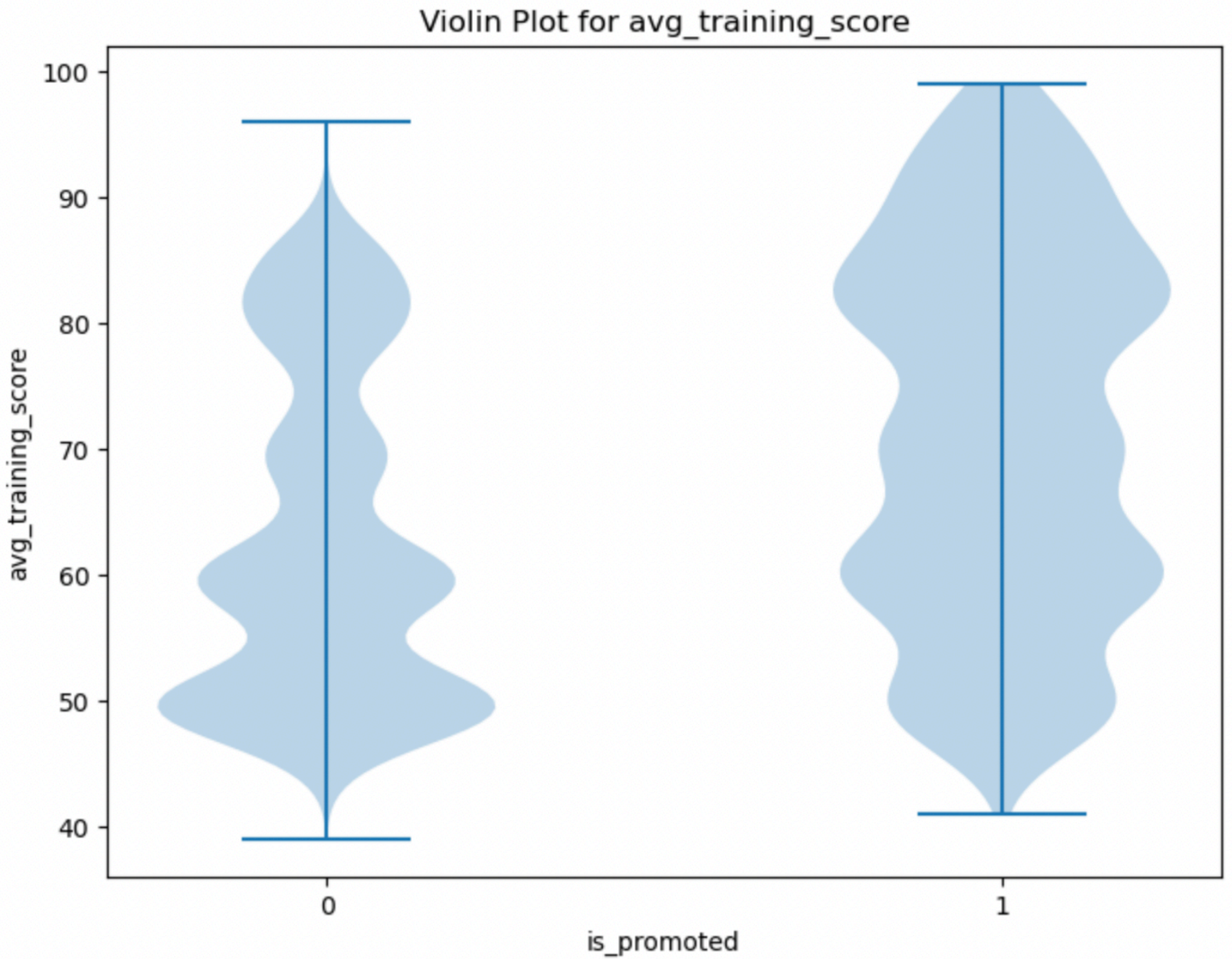
EDA

Why?


category-specific histogram for no_of_trainings
not promoted
promoted
fraction
0.0
0.2
0.4
0.6
0.8
2
4
6
8
10
no_of_trainings

EDA

Why?

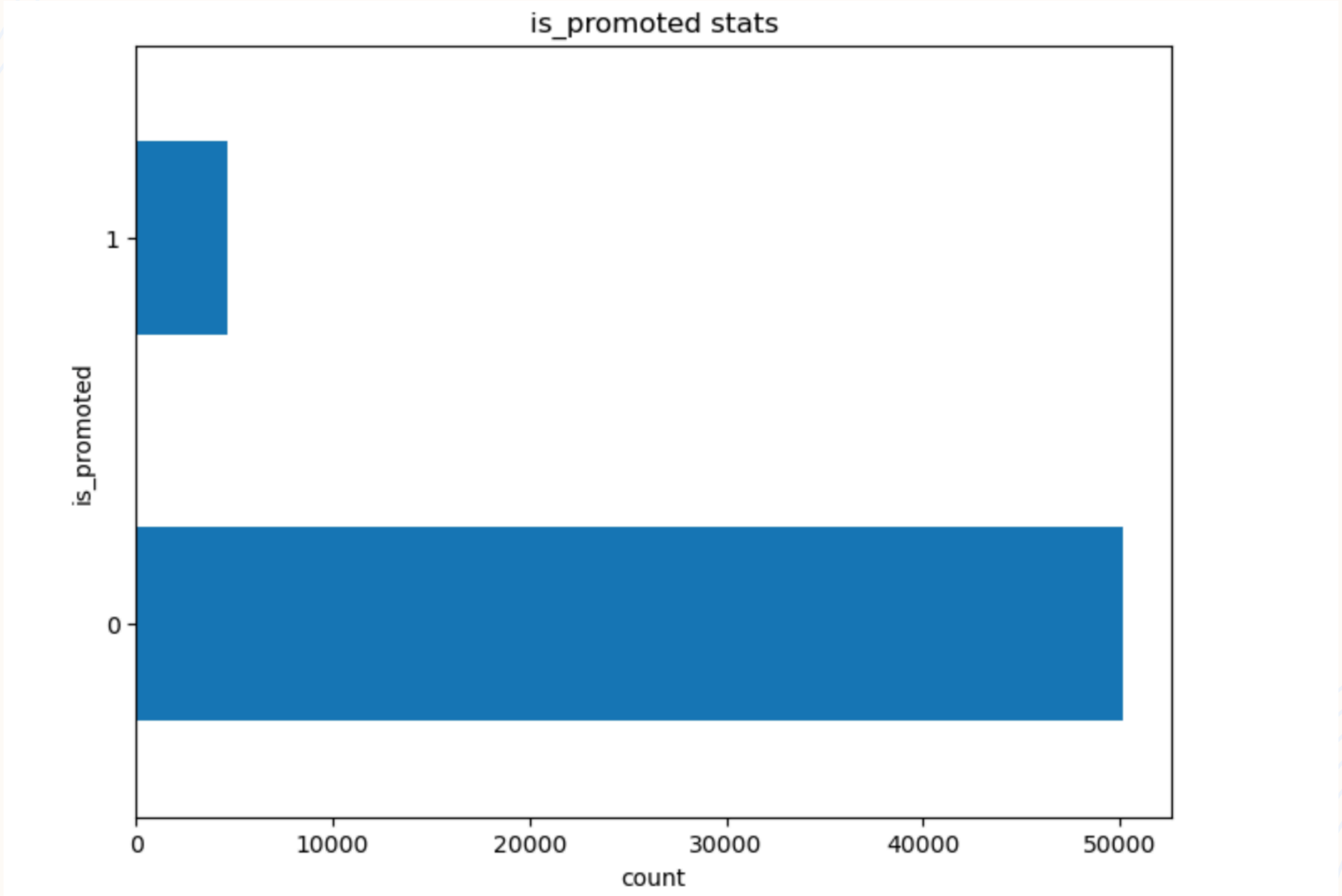

Violin Plot for avg_training_score
avg_training_score
100
90
80
70
60
50
40
0
1
is_promoted

SPLITTING

- HOW & WHY
- Data Dimension: 54808 * 13
- iid, train/validation/test set (80%/10%/10%)
- X and traget(y) variable
- Stratify

avg_training_score	is_promoted
49	0
60	0
50	0
50	0
73	0
...	...

	department	region	education	gender	recruitment_channel	no_of_trainings	age	previous_year_rating	length_of_service	KPIs_met >80%	awards_won?
0	Sales & Marketing	region_7	Master's & above	f	sourcing	1	35	5.0	8	1	0
1	Operations	region_22	Bachelor's	m	other	1	30	5.0	4	0	0
2	Sales & Marketing	region_19	Bachelor's	m	sourcing	1	34	3.0	7	0	0
3	Sales & Marketing	region_23	Bachelor's	m	other	2	39	1.0	10	0	0
4	Technology	region_26	Bachelor's	m	other	1	45	3.0	2	0	0
...	...	...	...	...	...	...	...	...	...	...	...
803	Technology	region_14	Bachelor's	m	sourcing	1	48	3.0	17	0	0


is_promoted stats
is_promoted
1
0
0
10000
20000
30000
40000
50000
count

PREPROCESSING

- Missing values

- Fraction of missing values in features:
- education: 0.043333
- previous_year_rating: 0.075218

- (43846, 12)
- (38949, 12)
- 4897 rows with missing values: 0.1116863567942343 6

PREPROCESSING

- WHAT & WHY

```
std_cols = ['no_of_trainings','length_of_service','avg_training_score']
cat_cols = ['department','region','gender','recruitment_channel','KPIs_met >80%','awards_won?']
ord_col1 = ['education']
ord_col2 = ['previous_year_rating']
minmax_cols = ['age']

ord_cat1 = [["NA","Below Secondary","Bachelor's","Master's & above"]]

ord_cat2 = [[0.0,1.0,2.0,3.0,4.0,5.0]]
```

PREPROCESSING

- WHAT & WHY

- Categorical features -> one-hot-encoding
- Ordinal features -> OrdinalEncoder (After filled in missing values)
- Numerical features with a well known range -> minmaxscaler
- Numerical features -> standard scaler

PREPROCESSING

- BEFORE & AFTER

```
print(df_train.shape)
print(X_train.shape)
```

```
(43846, 58)
(43846, 12)
```