HR Analytics

Celsiya A

Catherine Mariana Philips

PROBLEM STATEMENT

- ✓ The given dataset has a total of 14,999 records and 10 attributes.
- ✓ We have been **assigned the task of predicting which employees will churn out from the company.**
- ✓ We will also have to **recommend special plans or strategies which will help to retain the employees** which in turn will help the company to grow bigger.
- ✓ We will be using Python to predict whether the employee will leave or stay the company.

VARIABLE DESCRIPTION

Variables	Description
satisfactoryLevel	Scores given by the employees, scaling 0 to 1
lastEvaluation	Last evaluation points given, scaling 0 to 1
numberOfProjects	Number of projects involved
avgMonthlyHours	Average monthly hours
timeSpent.company	Time spent at the company, in years
workAccident	Whether he/she had a work accident
left	if the employee is about to leave or not, about to leave(serving notice period) – 1 and 0 otherwise
promotionInLast5years	Whether he/she had a promotion in the last 5 years
dept	Department he/she belongs to
Salary	Salary as high, medium or low

- ✓ We will first import the required libraries.
    - OS: To change the working directory
    - Numpy: To perform numerical operations
    - Pandas: To work with dataframes
- ✓ The dataset is imported from the required path.
- ✓ There are no null values in the dataset.
- ✓ Out of the 14999 rows in the dataset, there are 3008 records that are duplicated. These duplicates are dropped

UNDERSTANDING THE DATA

```
RangeIndex: 14999 entries, 0 to 14998
Data columns (total 10 columns):
 #   Column                 Non-Null Count  Dtype
---  ------                 --------------  -----
 0   satisfactoryLevel      14999 non-null  float64
 1   lastEvaluation         14999 non-null  float64
 2   numberOfProjects       14999 non-null  int64
 3   avgMonthlyHours        14999 non-null  int64
 4   timeSpent.company      14999 non-null  int64
 5   workAccident           14999 non-null  int64
 6   left                   14999 non-null  int64
 7   promotionInLast5years  14999 non-null  int64
 8   dept                   14999 non-null  object
 9   salary                 14999 non-null  object
dtypes: float64(2), int64(6), object(2)
memory usage: 1.1+ MB
```

SUMMARY- Numerical Variables

	satisfactoryLevel	lastEvaluation	avgMonthlyHours
count	9653.000000	9653.000000	9653.000000
mean	0.642706	0.717291	199.993681
std	0.234450	0.166164	47.815262
min	0.090000	0.360000	96.000000
25%	0.500000	0.570000	159.000000
50%	0.670000	0.720000	199.000000
75%	0.830000	0.860000	242.000000
max	1.000000	1.000000	310.000000

SUMMARY- Categorical Variables

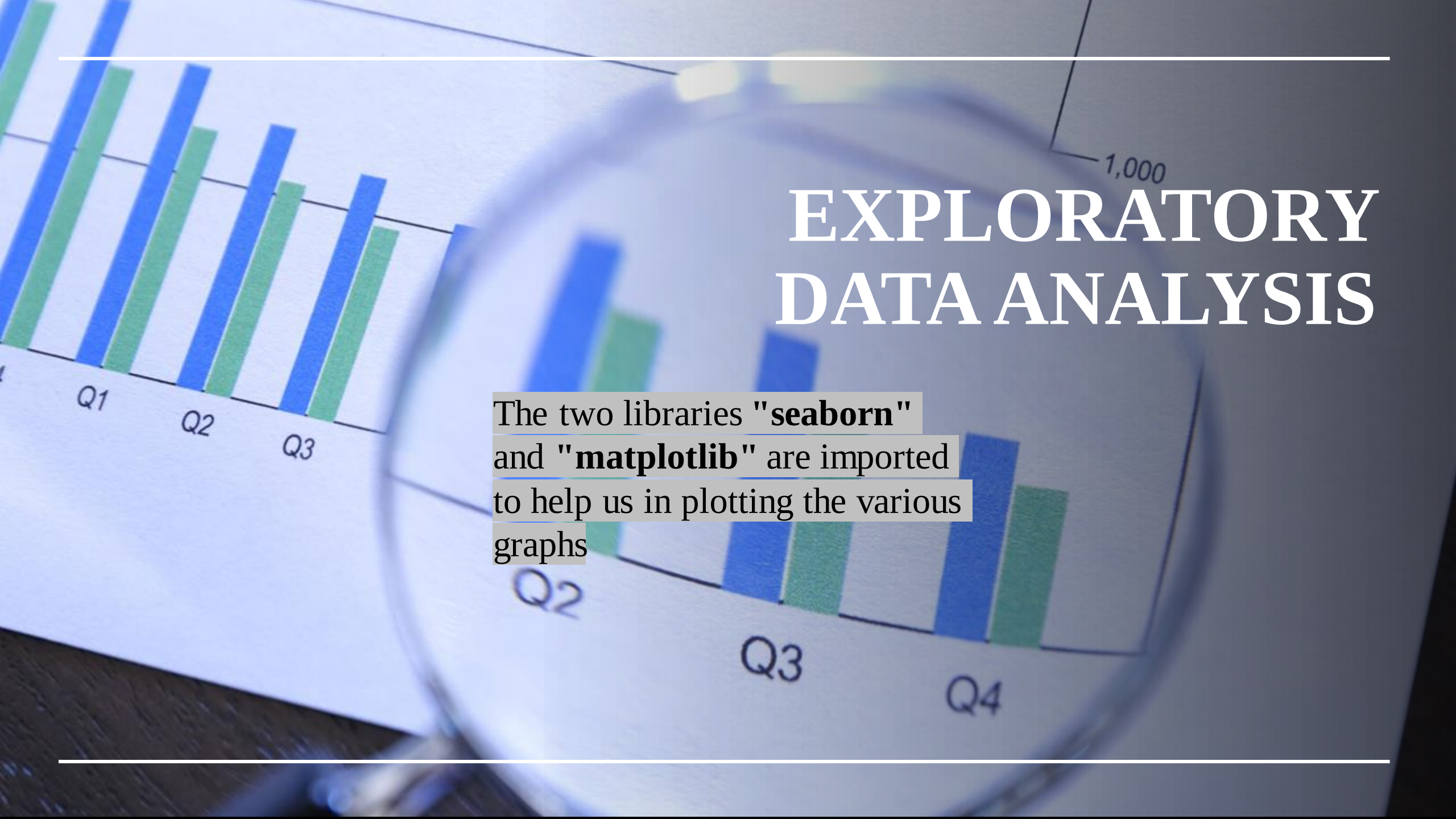
```
Number of Projects
4    4365
3    4055
5    2761
2    2388
6    1174
7     256
Name: numberOfProjects, dtype: int64
-----------------------------------
Time Spent at the Company
3     6443
2     3244
4     2557
5     1473
6      718
10     214
7      188
8      162
Name: timeSpent.company, dtype: int64
```

```
Work Accident
0    12830
1     2169
Name: workAccident, dtype: int64
-----------------------------------
Left
0    11428
1     3571
Name: left, dtype: int64
-----------------------------------
Promotion in the last 5 years
0    14680
1      319
Name: promotionInLast5years, dtype: int64
```

```
Department
sales          2532
technical      1899
support        1515
IT              774
RandD           616
marketing       526
accounting      517
hr              508
product_mng     508
management      258
Name: dept, dtype: int64
----------------------------
Salary
low       4594
medium    4302
high       757
Name: salary, dtype: int64
```

SAMPLE DATA

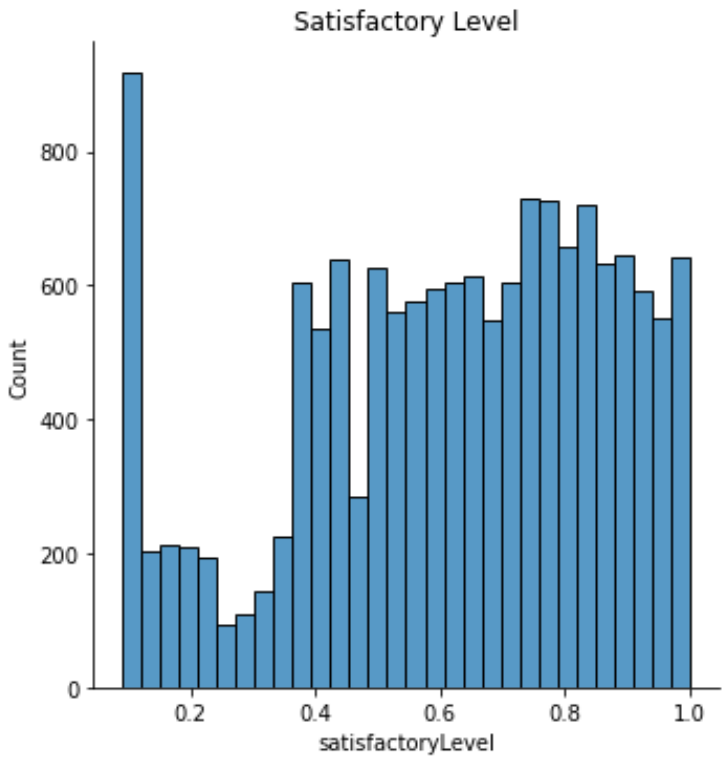
	satisfactoryLevel	lastEvaluation	numberOfProjects	avgMonthlyHours	timeSpent.company	workAccident	left	promotionInLast5years	dept	salary
0	0.38	0.53	2	157	3	0	1	0	sales	low
1	0.80	0.86	5	262	6	0	1	0	sales	medium
2	0.11	0.88	7	272	4	0	1	0	sales	medium
3	0.37	0.52	2	159	3	0	1	0	sales	low
4	0.41	0.50	2	153	3	0	1	0	sales	low



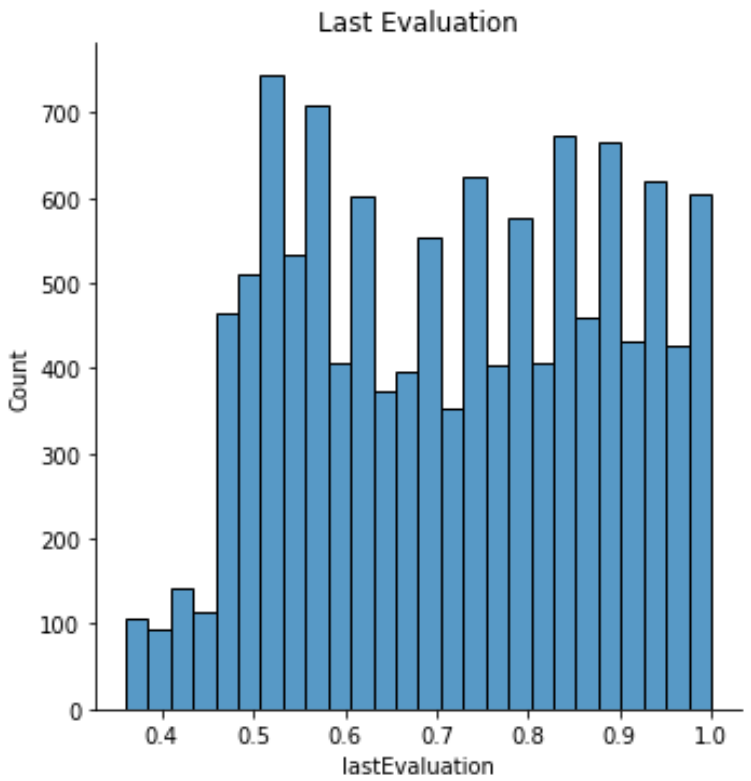
EXPLORATORY DATA ANALYSIS

The two libraries **"seaborn"** and **"matplotlib"** are imported to help us in plotting the various graphs

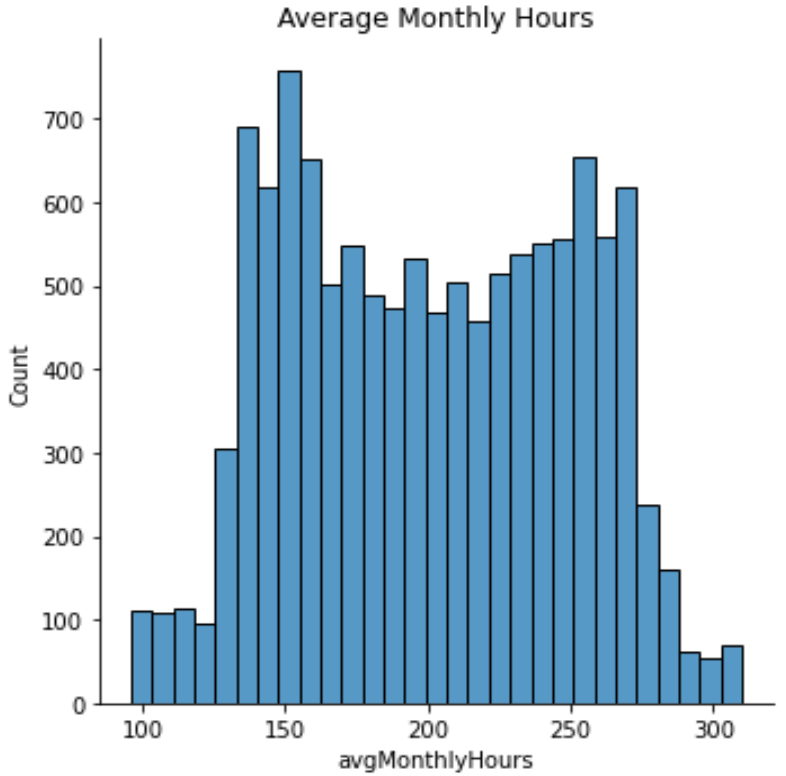
DISTRIBUTIONS



SATISFACTORY LEVEL

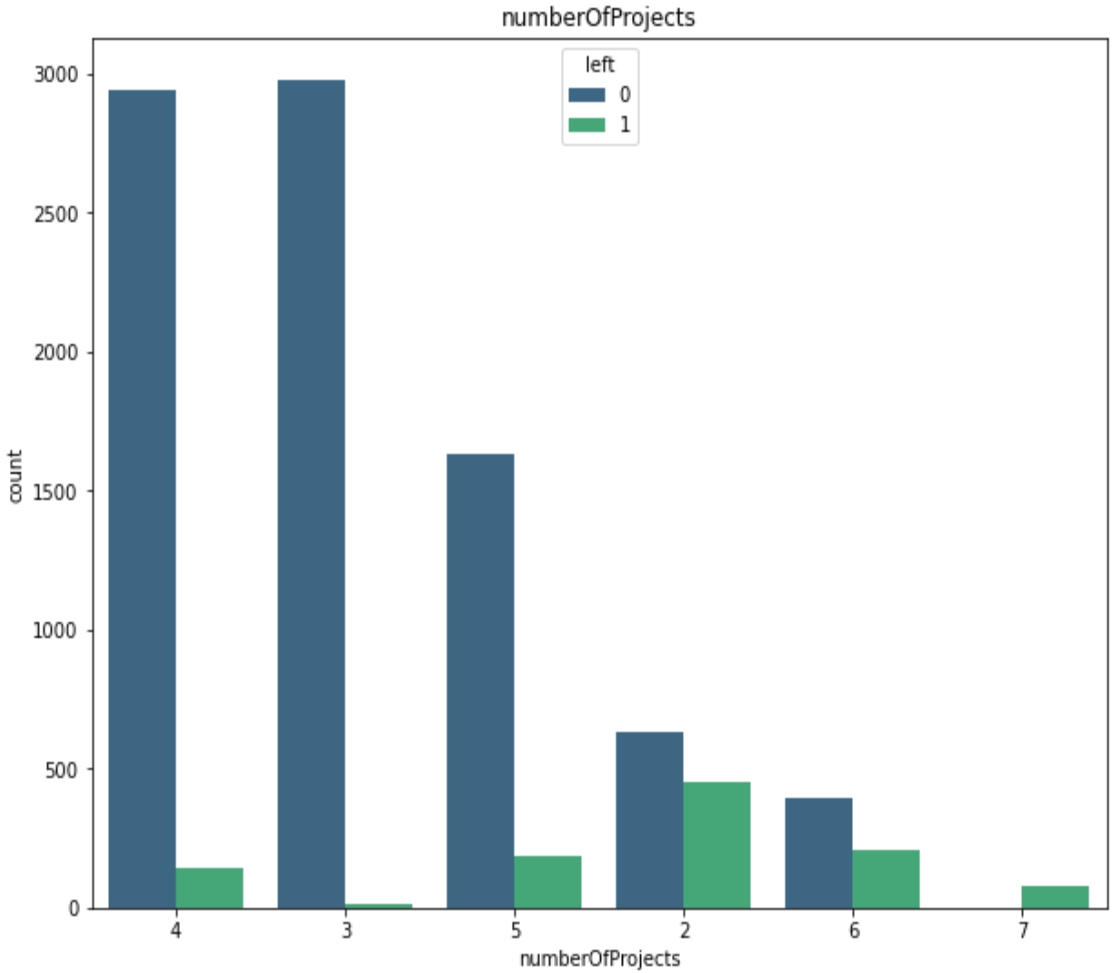


LAST EVALUATION

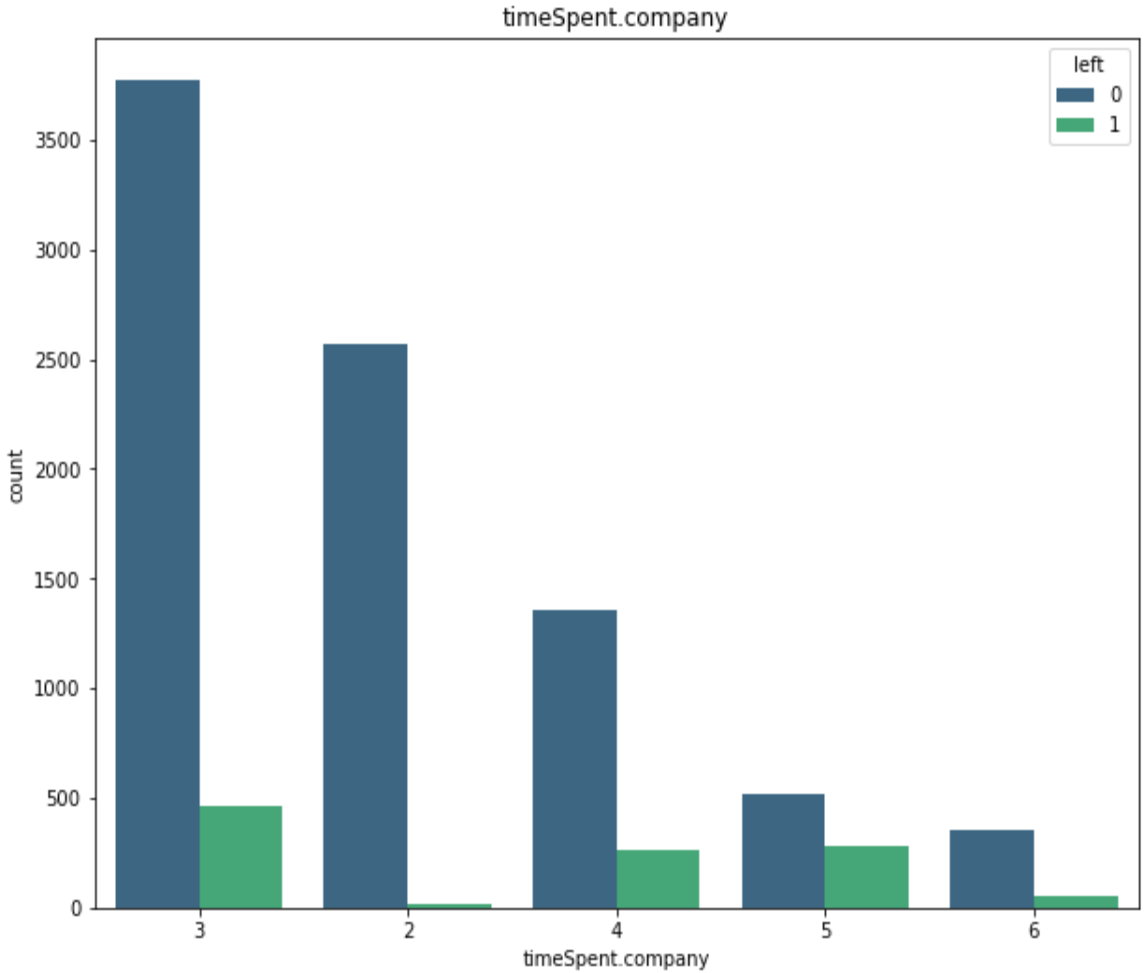


AVERAGE MONTHLY HOURS

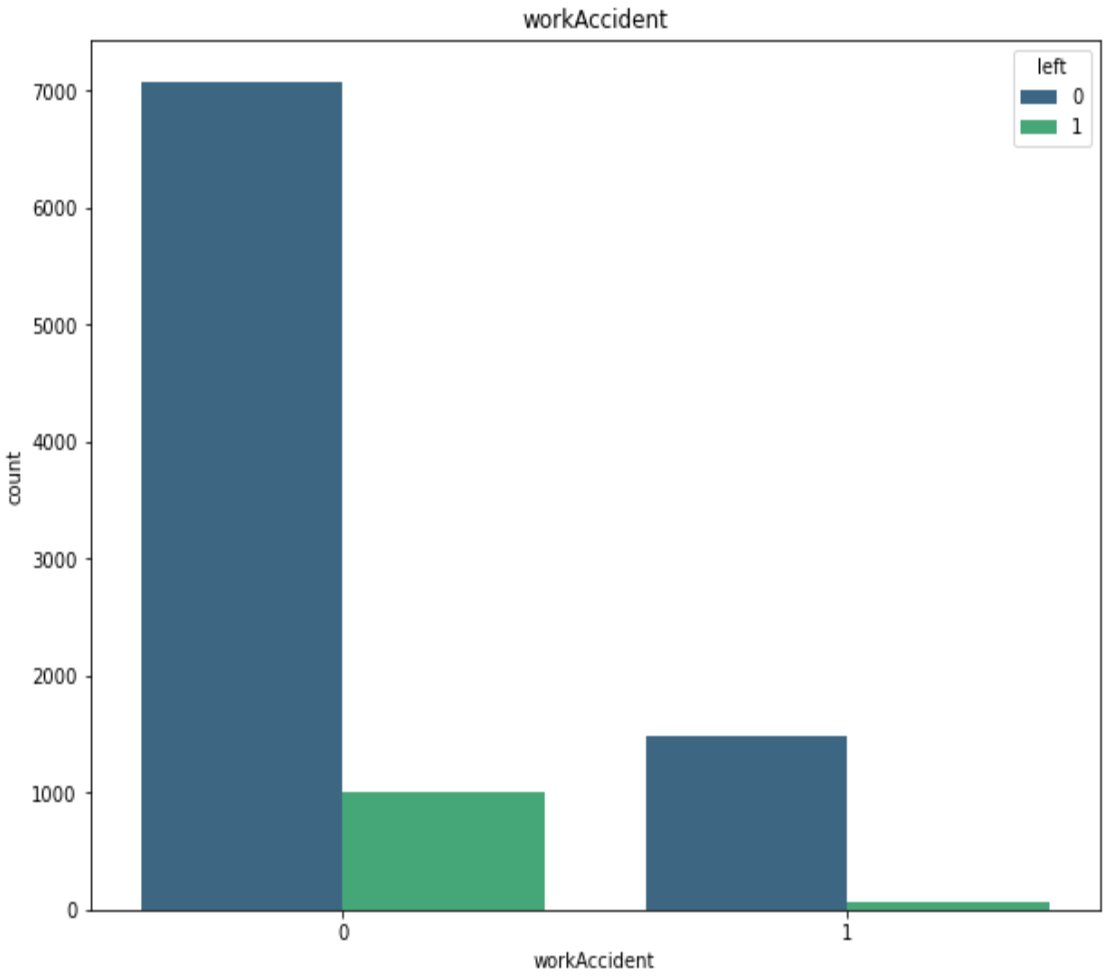
NUMBER OF PROJECTS



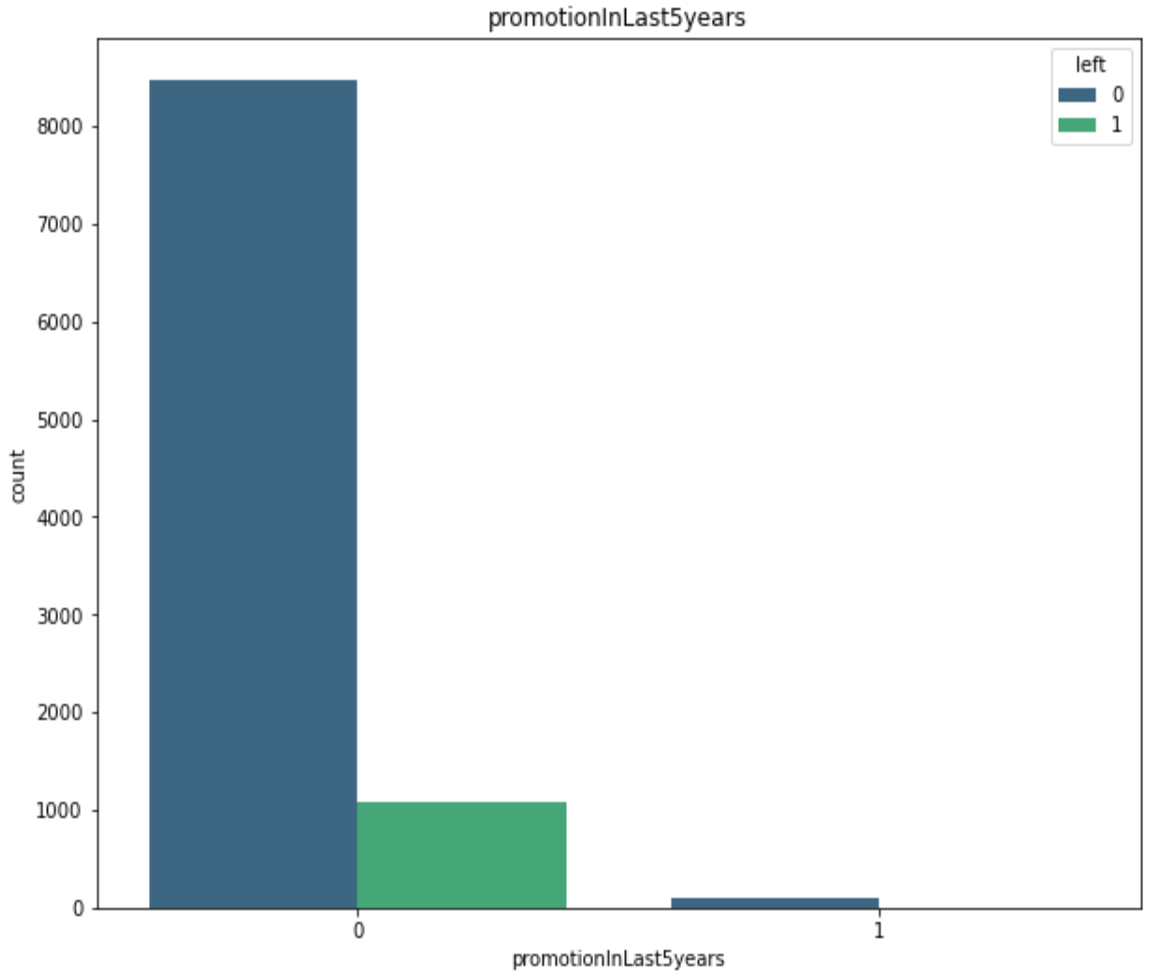
TIME SPENT AT COMPANY



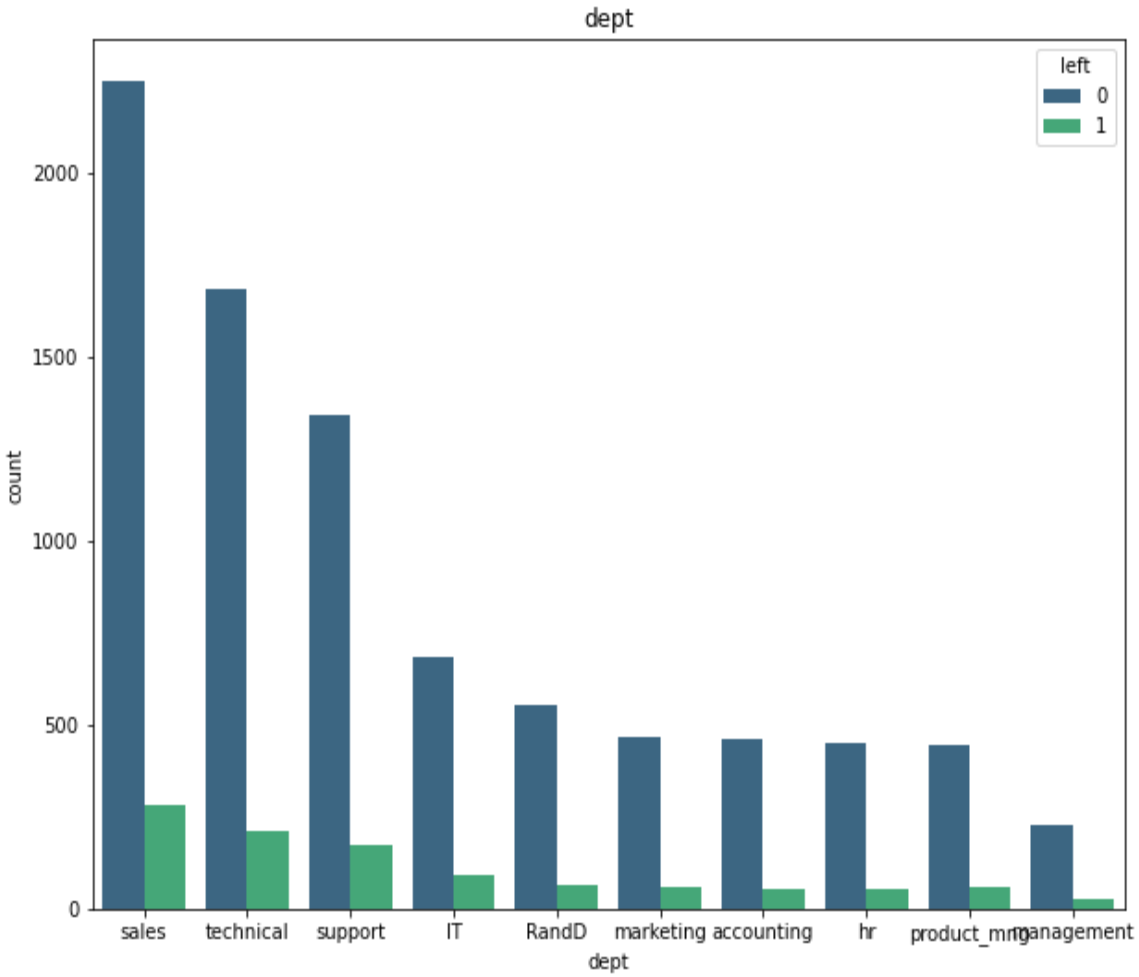
WORK ACCIDENT



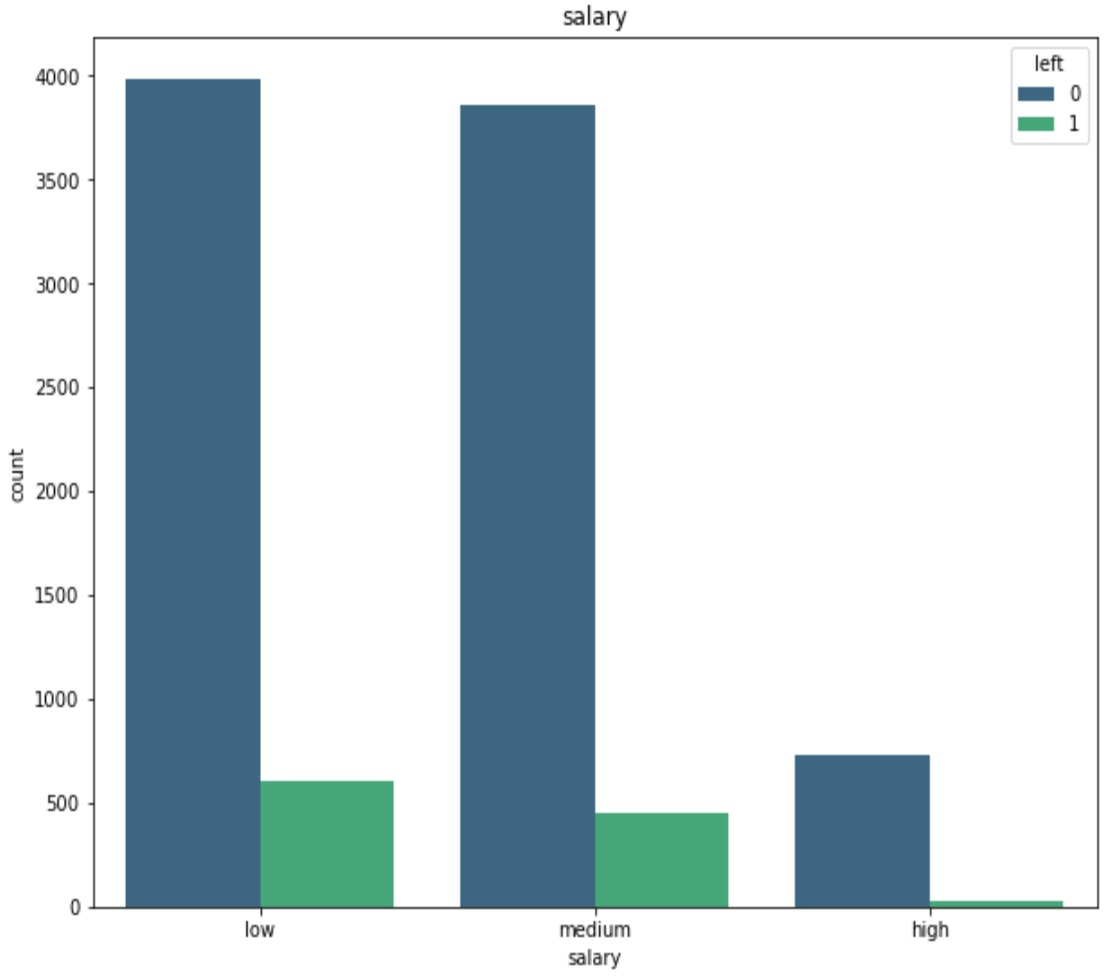
PROMOTION IN LAST 5 YEARS



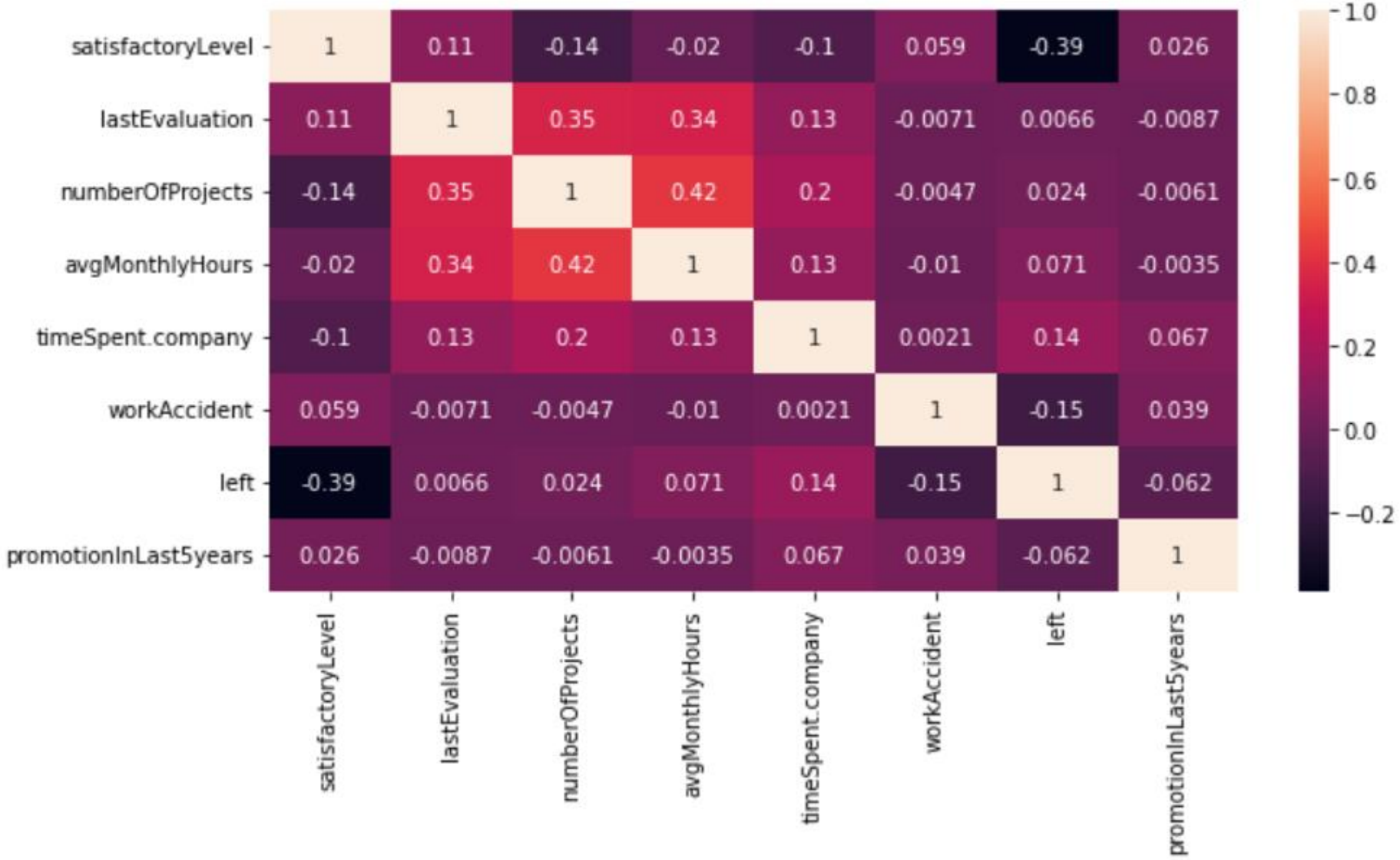
DEPARTMENT



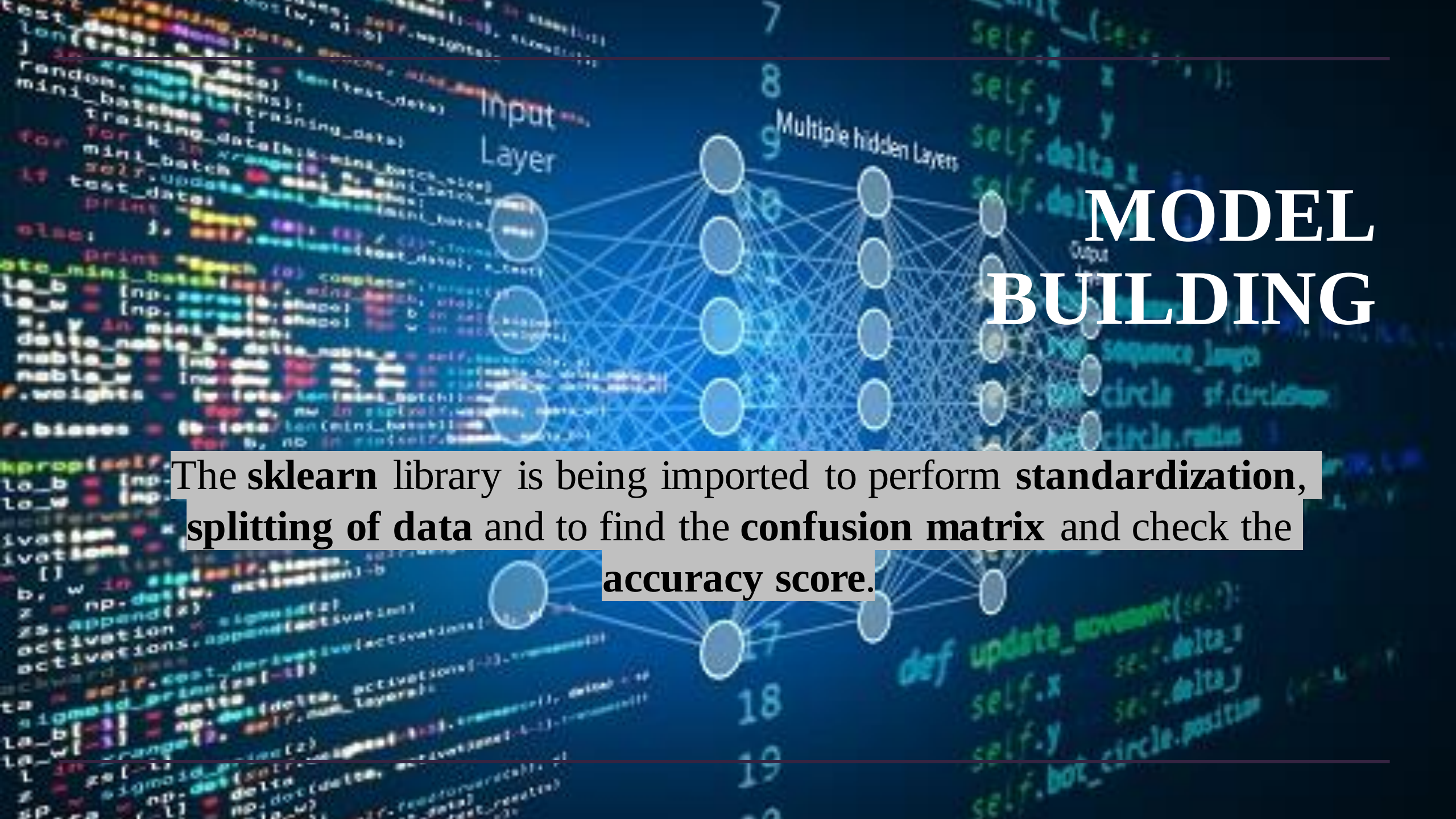
SALARY



CORRELATION



MODEL BUILDING



The **sklearn** library is being imported to perform **standardization**, **splitting of data** and to find the **confusion matrix** and check the **accuracy score**.

LOGISTIC REGRESSION

BEFORE DROPPING COLUMNS

```
Confusion matrix for Logistic Regression
[[2462  108]
 [ 201  125]]
```

Accuracy Score: 0.893
Misclassified samples: 306

AFTER DROPPING COLUMNS

```
Confusion matrix for Logistic Regression
[[2460  110]
 [ 189  137]]
```

Accuracy Score: 0.897
Misclassified samples: 309

RANDOM FOREST

BEFORE DROPPING COLUMNS

```
Confusion matrix for Random Forest
[[2567    3]
 [  24  302]]
```

Accuracy Score: 0.990
Misclassified samples: 309

AFTER DROPPING COLUMNS

```
Confusion matrix for Random Forest
[[2567    3]
 [  23  303]]
```

Accuracy Score: 0.991
Misclassified samples: 309

K NEAREST NEIGHBORS

BEFORE DROPPING COLUMNS

```
Confusion matrix for KNN
[[2462  108]
 [ 201  125]]
```

Accuracy Score: 0.893
Misclassified samples: 307

AFTER DROPPING COLUMNS

```
Confusion matrix for KNN
[[2487   83]
 [  42  284]]
```

Accuracy Score: 0.956
Misclassified samples: 309

STRATEGIC RETENTION PLANS

- Salary
- Number of Projects
- Time spent at the company
- Promotion in the last 5 years

THANK YOU