Data Mining: Employee Attrition

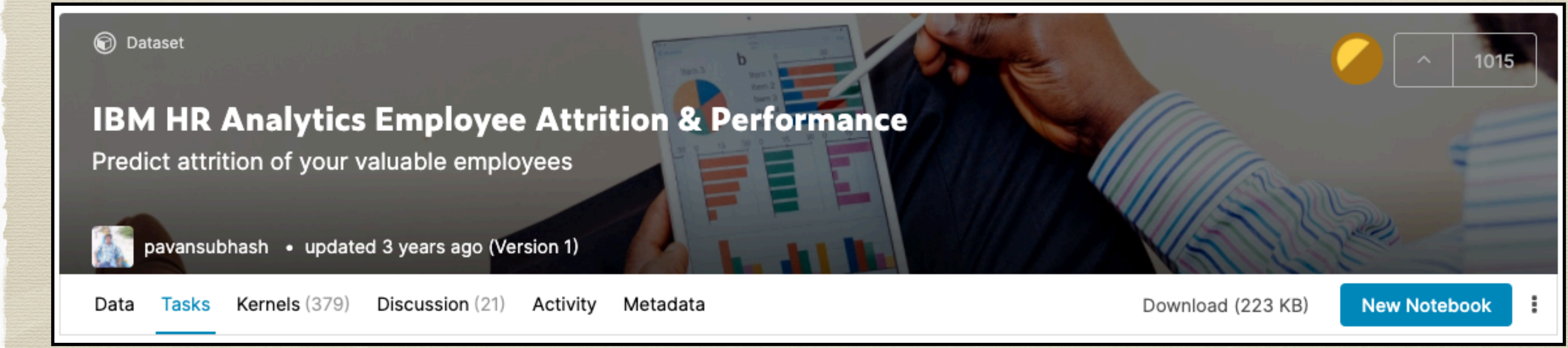
Tutorial
TA: Tzu-Heng Huang
Email: zihenghı@gmail.com
Course Instructor: Man-Kwan Shan

Contents

1. Introduction
2. Overview
3. Data Table
4. Data Description
5. EDA (Explore Data Analysis)
6. Data Cleaning & Feature Engineering
7. Data Splitting & Model Learning
8. Model Evaluation

Introduction

- git clone https://github.com/zihengh1/DM2020.git
- pip install -r requirement.txt
- Kaggle is a data science platform for model learning and data analysis.
- Data Source: https://www.kaggle.com/pavansubhasht/ibm-hr-analytics-attrition-dataset
- A classification problem to predict who will stay in the company.
- Attributes: 27, Data: 1470

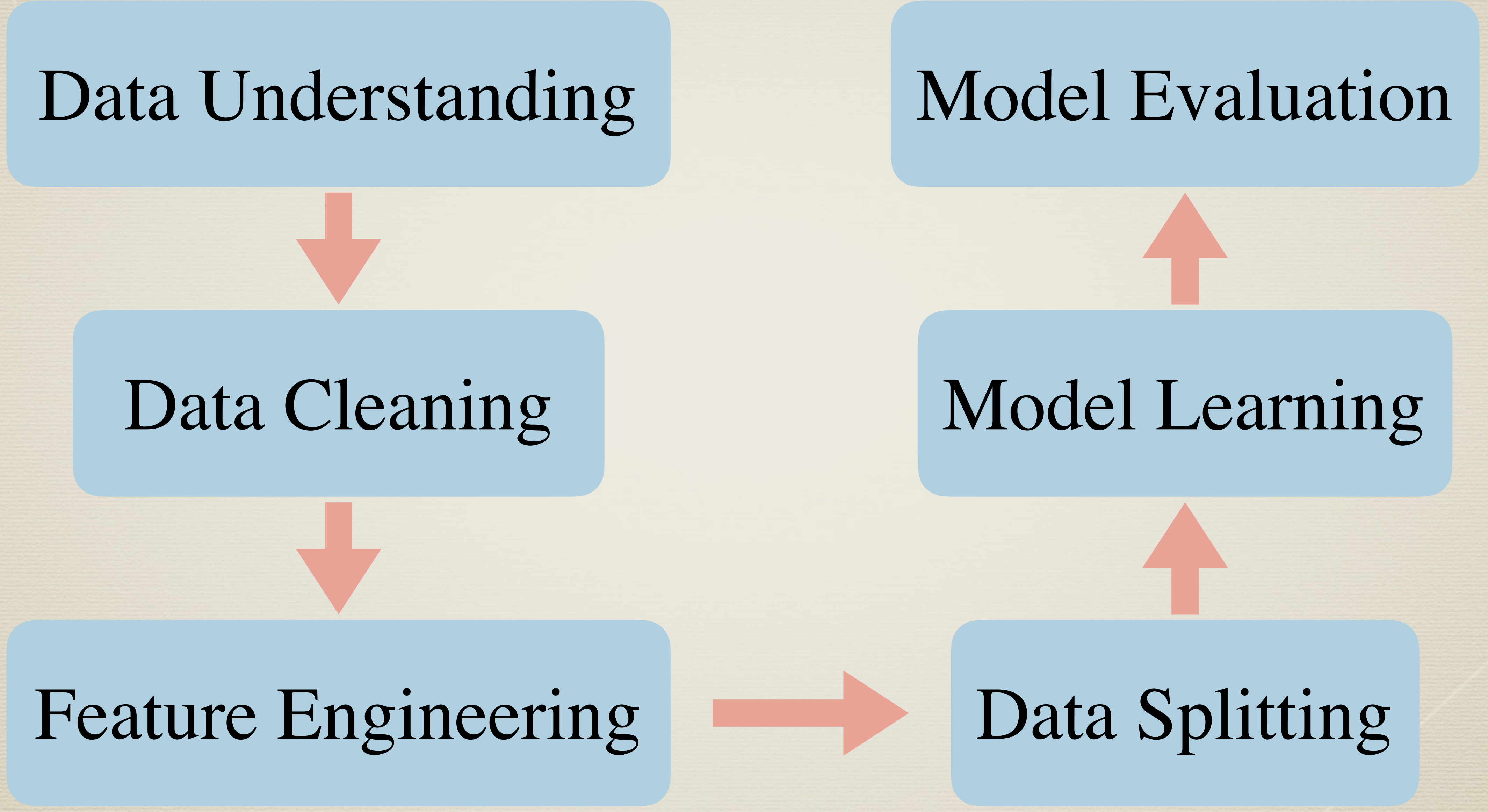


IBM HR Analytics Employee Attrition & Performance

Predict attrition of your valuable employees

pavansubhash • updated 3 years ago (Version 1)

Overview



Data Understanding

Data Cleaning

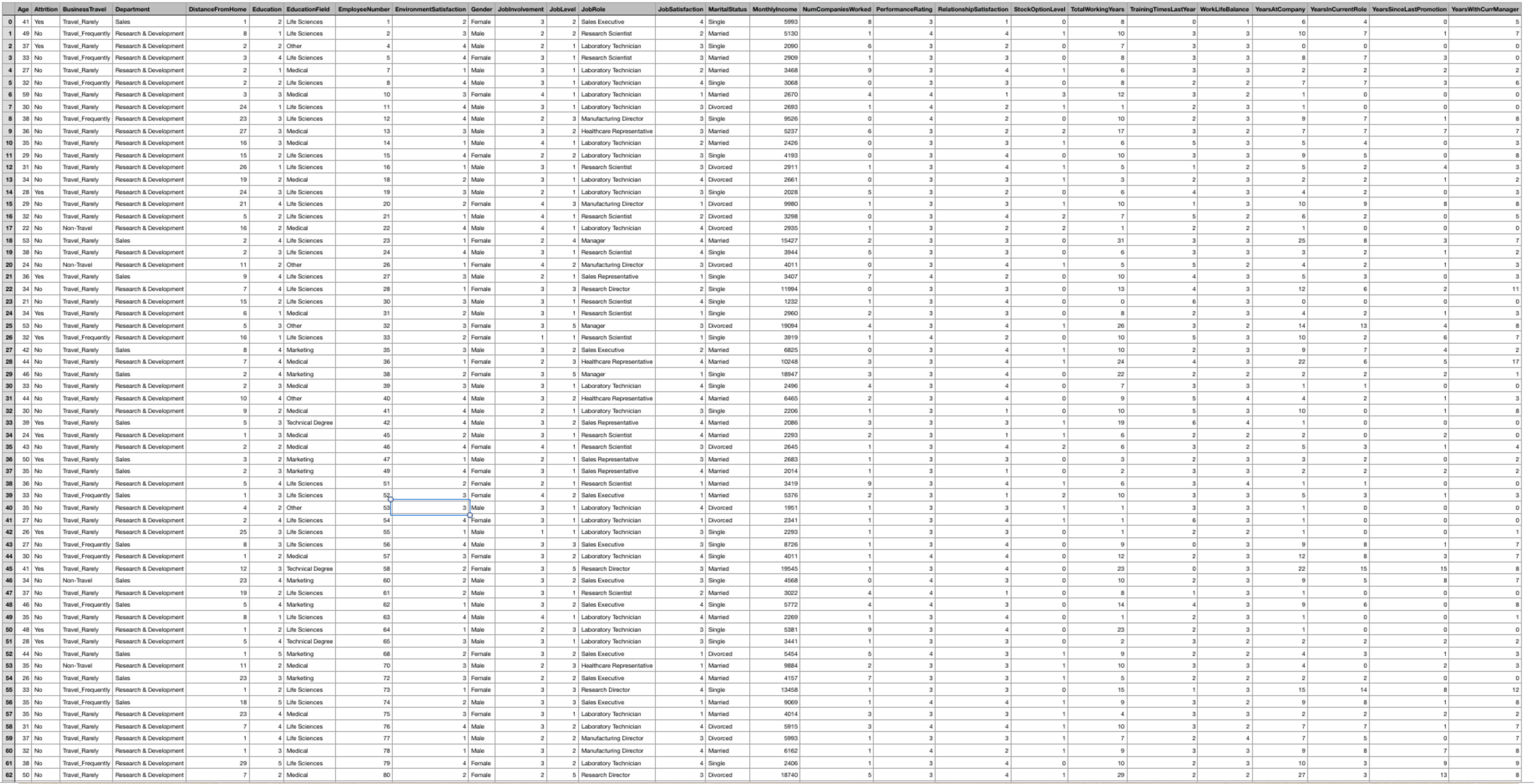
Feature Engineering

Data Splitting

Model Learning

Model Evaluation

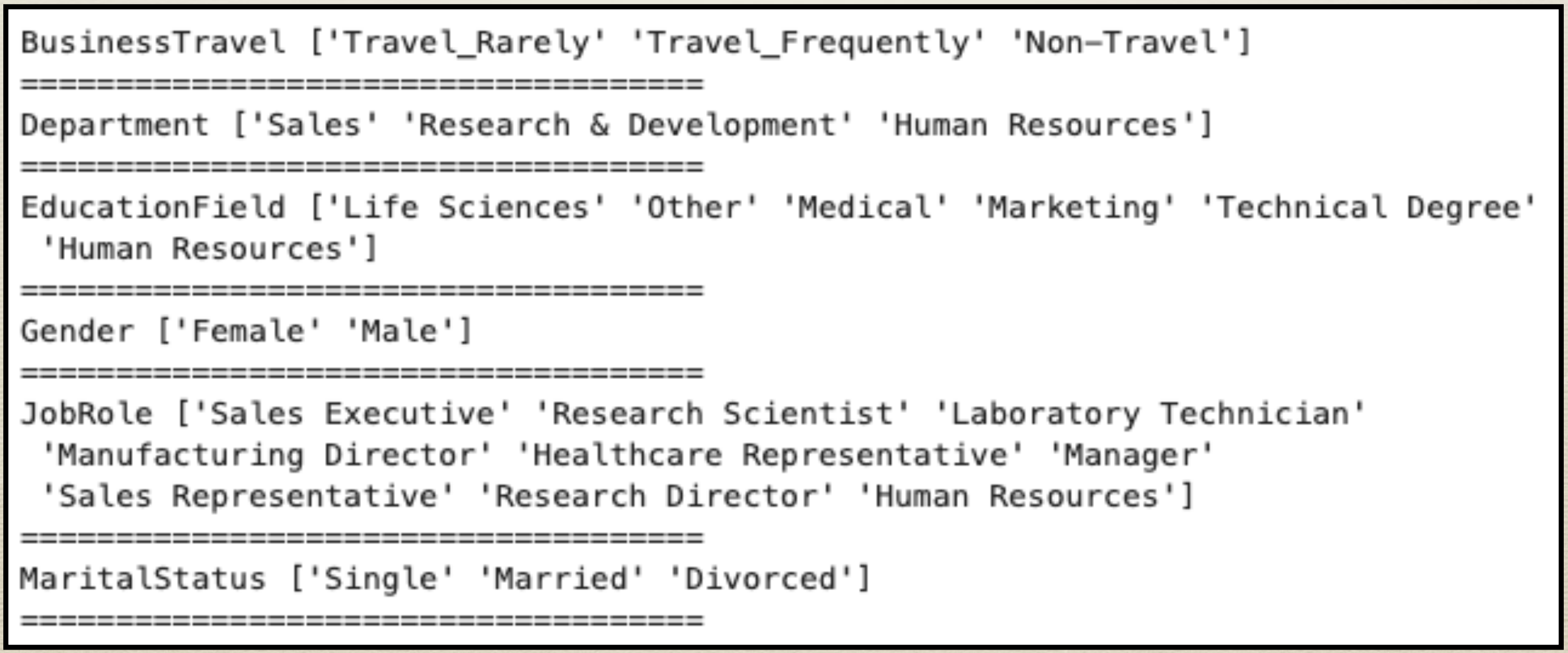
Data Table



	Age	Attrition	BusinessTravel	Department	DistanceFromHome	Education	EducationField	EmployeeNumber	EnvironmentSatisfaction	Gender	JobInvolvement	JobLevel	JobRole	JobSatisfaction	MaritalStatus	MonthlyIncome	NumCompaniesWorked	PerformanceRating	RelationshipSatisfaction	StockOptionLevel	TotalWorkingYears	TrainingTimesLastYear	WorkLifeBalance	YearsAtCompany	YearsInCurrentRole	YearsSinceLastPromotion	YearsWithCurrManager
0	41	Yes	Travel_Rarely	Sales	1	2	Life Sciences	1	2	Female	3	2	Sales Executive	4	Single	5993	8	3	1	0	8	0	1	6	4	0	5
1	49	No	Travel_Frequently	Research & Development	8	1	Life Sciences	2	3	Male	2	2	Research Scientist	2	Married	5130	1	4	4	1	10	3	3	10	7	1	7
2	37	Yes	Travel_Rarely	Research & Development	2	2	Other	4	4	Male	2	1	Laboratory Technician	3	Single	2090	6	3	2	0	7	3	3	0	0	0	0
3	33	No	Travel_Frequently	Research & Development	3	4	Life Sciences	5	4	Female	3	1	Research Scientist	3	Married	2909	1	3	3	0	8	3	3	8	7	3	0
4	27	No	Travel_Rarely	Research & Development	2	1	Medical	7	1	Male	3	1	Laboratory Technician	2	Married	3468	9	3	4	1	6	3	3	2	2	2	2
5	32	No	Travel_Frequently	Research & Development	2	2	Life Sciences	8	4	Male	3	1	Laboratory Technician	4	Single	3068	0	3	3	0	8	2	2	7	7	3	6
6	59	No	Travel_Rarely	Research & Development	3	3	Medical	10	3	Female	4	1	Laboratory Technician	1	Married	2670	4	4	1	3	12	3	2	1	0	0	0
7	30	No	Travel_Rarely	Research & Development	24	1	Life Sciences	11	4	Male	3	1	Laboratory Technician	3	Divorced	2693	1	4	2	1	1	2	3	1	0	0	0
8	38	No	Travel_Frequently	Research & Development	23	3	Life Sciences	12	4	Male	2	3	Manufacturing Director	3	Single	9526	0	4	2	0	10	2	3	9	7	1	8
9	36	No	Travel_Rarely	Research & Development	27	3	Medical	13	3	Male	3	2	Healthcare Representative	3	Married	5237	6	3	2	2	17	3	2	7	7	7	7
10	35	No	Travel_Rarely	Research & Development	16	3	Medical	14	1	Male	4	1	Laboratory Technician	2	Married	2426	0	3	3	1	6	5	3	5	4	0	3
11	29	No	Travel_Rarely	Research & Development	15	2	Life Sciences	15	4	Female	2	2	Laboratory Technician	3	Single	4193	0	3	4	0	10	3	3	9	5	0	8
12	31	No	Travel_Rarely	Research & Development	26	1	Life Sciences	16	1	Male	3	1	Research Scientist	3	Divorced	2911	1	3	4	1	5	1	2	5	2	4	3
13	34	No	Travel_Rarely	Research & Development	19	2	Medical	18	2	Male	3	1	Laboratory Technician	4	Divorced	2661	0	3	3	1	3	2	3	2	2	1	2
14	28	Yes	Travel_Rarely	Research & Development	24	3	Life Sciences	19	3	Male	2	1	Laboratory Technician	3	Single	2028	5	3	2	0	6	4	3	4	2	0	3
15	29	No	Travel_Rarely	Research & Development	21	4	Life Sciences	20	2	Female	4	3	Manufacturing Director	1	Divorced	9980	1	3	3	1	10	1	3	10	9	8	8
16	32	No	Travel_Rarely	Research & Development	5	2	Life Sciences	21	1	Male	4	1	Research Scientist	2	Divorced	3298	0	3	4	2	7	5	2	6	2	0	5
17	22	No	Non-Travel	Research & Development	16	2	Medical	22	4	Male	4	1	Laboratory Technician	4	Divorced	2935	1	3	2	2	1	2	2	1	0	0	0
18	53	No	Travel_Rarely	Sales	2	4	Life Sciences	23	1	Female	2	4	Manager	4	Married	15427	2	3	3	0	31	3	3	25	8	3	7
19	38	No	Travel_Rarely	Research & Development	2	3	Life Sciences	24	4	Male	3	1	Research Scientist	4	Single	3944	5	3	3	0	6	3	3	3	2	1	2
20	24	No	Non-Travel	Research & Development	11	2	Other	26	1	Female	4	2	Manufacturing Director	3	Divorced	4011	0	3	4	1	5	5	2	4	2	1	3
21	36	Yes	Travel_Rarely	Sales	9	4	Life Sciences	27	3	Male	2	1	Sales Representative	1	Single	3407	7	4	2	0	10	4	3	5	3	0	3
22	34	No	Travel_Rarely	Research & Development	7	4	Life Sciences	28	1	Female	3	3	Research Director	2	Single	11994	0	3	3	0	13	4	3	12	6	2	11
23	21	No	Travel_Rarely	Research & Development	15	2	Life Sciences	30	3	Male	3	1	Research Scientist	4	Single	1232	1	3	4	0	0	6	3	0	0	0	0
24	34	Yes	Travel_Rarely	Research & Development	6	1	Medical	31	2	Male	3	1	Research Scientist	1	Single	2960	2	3	3	0	8	2	3	4	2	1	3
25	53	No	Travel_Rarely	Research & Development	5	3	Other	32	3	Female	3	5	Manager	3	Divorced	19094	4	3	4	1	26	3	2	14	13	4	8
26	32	Yes	Travel_Frequently	Research & Development	16	1	Life Sciences	33	2	Female	1	1	Research Scientist	1	Single	3919	1	4	2	0	10	5	3	10	2	6	7
27	42	No	Travel_Rarely	Sales	8	4	Marketing	35	3	Male	3	2	Sales Executive	2	Married	6825	0	3	4	1	10	2	3	9	7	4	2
28	44	No	Travel_Rarely	Research & Development	7	4	Medical	36	1	Female	2	3	Healthcare Representative	4	Married	10248	3	3	4	1	24	4	3	22	6	5	17
29	46	No	Travel_Rarely	Sales	2	4	Marketing	38	2	Female	3	5	Manager	1	Single	18947	3	3	4	0	22	2	2	2	2	2	1
30	33	No	Travel_Rarely	Research & Development	2	3	Medical	39	3	Male	3	1	Laboratory Technician	4	Single	2496	4	3	4	0	7	3	3	1	1	0	0
31	44	No	Travel_Rarely	Research & Development	10	4	Other	40	4	Male	3	2	Healthcare Representative	4	Married	6465	2	3	4	0	9	5	4	4	2	1	3
32	30	No	Travel_Rarely	Research & Development	9	2	Medical	41	4	Male	2	1	Laboratory Technician	3	Single	2206	1	3	1	0	10	5	3	10	0	1	8
33	39	Yes	Travel_Rarely	Sales	5	3	Technical Degree	42	4	Male	3	2	Sales Representative	4	Married	2086	3	3	3	1	19	6	4	1	0	0	0
34	24	Yes	Travel_Rarely	Research & Development	1	3	Medical	45	2	Male	3	1	Research Scientist	4	Married	2293	2	3	1	1	6	2	2	2	0	2	0
35	43	No	Travel_Rarely	Research & Development	2	2	Medical	46	4	Female	4	1	Research Scientist	3	Divorced	2645	1	3	4	2	6	3	2	5	3	1	4
36	50	Yes	Travel_Rarely	Sales	3	2	Marketing	47	1	Male	2	1	Sales Representative	3	Married	2683	1	3	3	0	3	2	3	3	2	0	2
37	35	No	Travel_Rarely	Sales	2	3	Marketing	49	4	Female	3	1	Sales Representative	4	Married	2014	1	3	1	0	2	3	3	2	2	2	2
38	36	No	Travel_Rarely	Research & Development	5	4	Life Sciences	51	2	Female	2	1	Research Scientist	1	Married	3419	9	3	4	1	6	3	4	1	1	0	0
39	33	No	Travel_Frequently	Sales	1	3	Life Sciences	52	3	Female	4	2	Sales Executive	1	Married	5376	2	3	1	2	10	3	3	5	3	1	3
40	35	No	Travel_Rarely	Research & Development	4	2	Other	53	3	Male	3	1	Laboratory Technician	4	Divorced	1951	1	3	3	1	1	3	3	1	0	0	0
41	27	No	Travel_Rarely	Research & Development	2	4	Life Sciences	54	4	Female	3	1	Laboratory Technician	1	Divorced	2341	1	3	4	1	1	6	3	1	0	0	0
42	26	Yes	Travel_Rarely	Research & Development	25	3	Life Sciences	55	1	Male	1	1	Laboratory Technician	3	Single	2293	1	3	3	0	1	2	2	1	0	0	1
43	27	No	Travel_Frequently	Sales	8	3	Life Sciences	56	4	Male	3	3	Sales Executive	3	Single	8726	1	3	4	0	9	0	3	9	8	1	7
44	30	No	Travel_Frequently	Research & Development	1	2	Medical	57	3	Female	3	2	Laboratory Technician	4	Single	4011	1	4	4	0	12	2	3	12	8	3	7
45	41	Yes	Travel_Rarely	Research & Development	12	3	Technical Degree	58	2	Female	3	5	Research Director	3	Married	19545	1	3	4	0	23	0	3	22	15	15	8
46	34	No	Non-Travel	Sales	23	4	Marketing	60	2	Male	3	2	Sales Executive	3	Single	4568	0	4	3	0	10	2	3	9	5	8	7
47	37	No	Travel_Rarely	Research & Development	19	2	Life Sciences	61	2	Male	3	1	Research Scientist	2	Married	3022	4	4	1	0	8	1	3	1	0	0	0
48	46	No	Travel_Frequently	Sales	5	4	Marketing	62	1	Male	3	2	Sales Executive	4	Single	5772	4	4	3	0	14	4	3	9	6	0	8
49	35	No	Travel_Rarely	Research & Development	8	1	Life Sciences	63	4	Male	4	1	Laboratory Technician	4	Married	2269	1	3	4	0	1	2	3	1	0	0	1
50	48	Yes	Travel_Rarely	Research & Development	1	2	Life Sciences	64	1	Male	2	3	Laboratory Technician	3	Single	5381	9	3	4	0	23	2	3	1	0	0	0
51	28	Yes	Travel_Rarely	Research & Development	5	4	Technical Degree	65	3	Male	3	1	Laboratory Technician	3	Single	3441	1	3	3	0	2	3	2	2	2	2	2
52	44	No	Travel_Rarely	Sales	1	5	Marketing	68	2	Female	3	2	Sales Executive	1	Divorced	5454	5	4	3	1	9	2	2	4	3	1	3
53	35	No	Non-Travel	Research & Development	11	2	Medical	70	3	Male	2	3	Healthcare Representative	1	Married	9884	2	3	3	1	10	3	3	4	0	2	3
54	26	No	Travel_Rarely	Sales	23	3	Marketing	72	3	Female	2	2	Sales Executive	4	Married	4157	7	3	3	1	5	2	2	2	2	0	0
55	33	No	Travel_Frequently	Research & Development	1	2	Life Sciences	73	1	Female	3	3	Research Director	4	Single	13458	1	3	3	0	15	1	3	15	14	8	12
56	35	No	Travel_Frequently	Sales	18	5	Life Sciences	74	2	Male	3	3	Sales Executive	1	Married	9069	1	4	4	1	9	3	2	9	8	1	8
57	35	No	Travel_Rarely	Research & Development	23	4	Medical	75	3	Female	3	1	Laboratory Technician	1	Married	4014	3	3	3	1	4	3	3	2	2	2	2
58	31	No	Travel_Rarely	Research & Development	7	4	Life Sciences	76	4	Male	3	2	Laboratory Technician	4	Divorced	5915	3	4	4	1	10	3	2	7	7	1	7
59	37	No	Travel_Rarely	Research & Development	1	4	Life Sciences	77	1	Male	2	2	Manufacturing Director	3	Divorced	5993	1	3	3	1	7	2	4	7	5	0	7
60	32	No	Travel_Rarely	Research & Development	1	3	Medical	78	1	Male	3	2	Manufacturing Director	4	Married	6162	1	4	2	1	9	3	3	9	8	7	8
61	38	No	Travel_Frequently	Research & Development	29	5	Life Sciences	79	4	Female	3	2	Laboratory Technician	4	Single	2406	1	3	4	0	10	2	3	10	3	9	9
62	50	No	Travel_Rarely	Research & Development	7	2	Medical	80	2	Female	2	5	Research Director	3	Divorced	18740	5	3	4	1	29	2	2	27	3	13	8

Data Description (Categorical)

- Attrition (Bool): True / False
- BusinessTravel: 是否常常出差
- Department: 隸屬部門
- EducationField: 教育背景（科系）
- Gender: 性別
- JobRole: 職位
- MaritalStatus: 婚姻狀態

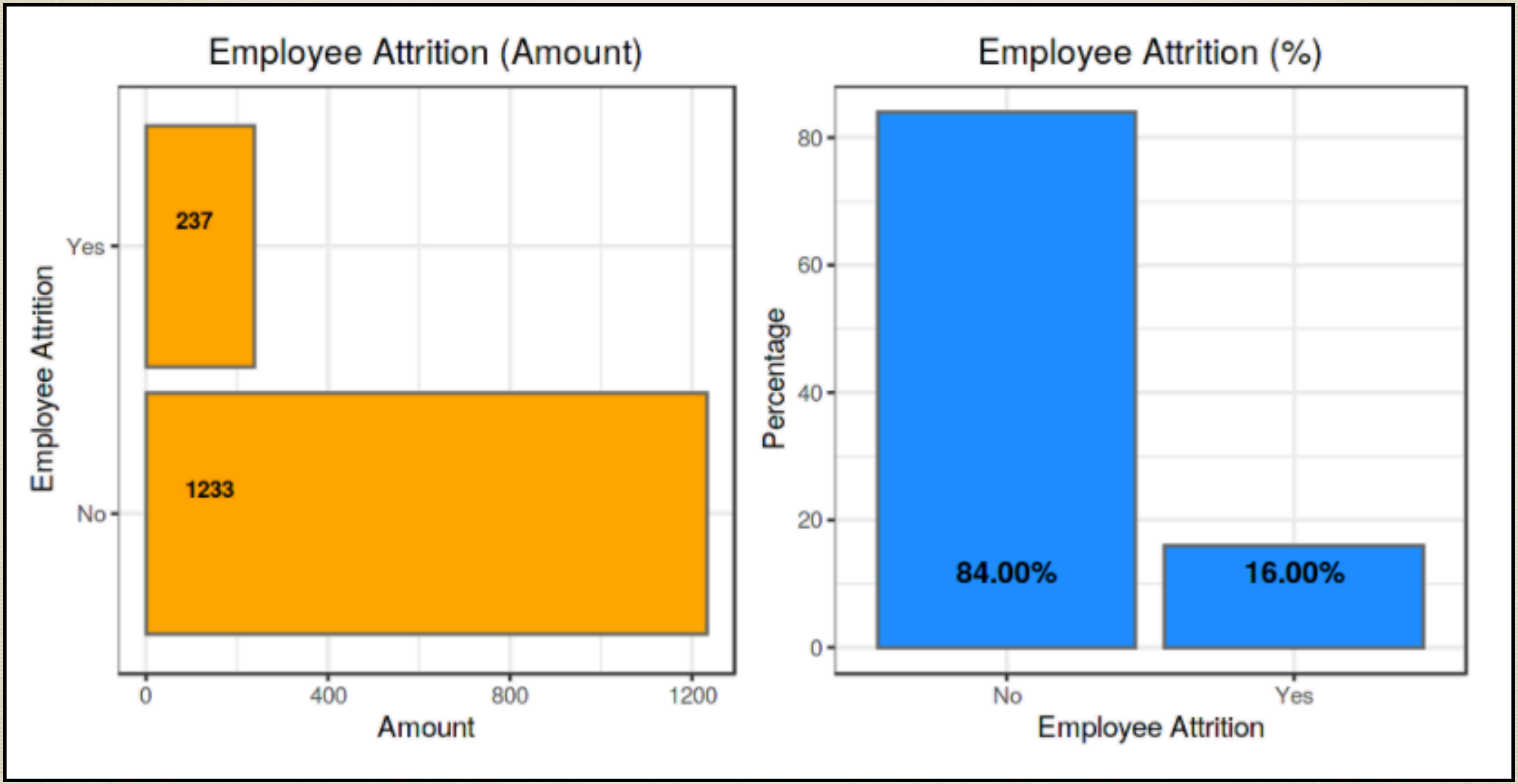


```
BusinessTravel ['Travel_Rarely' 'Travel_Frequently' 'Non-Travel']
===================================
Department ['Sales' 'Research & Development' 'Human Resources']
===================================
EducationField ['Life Sciences' 'Other' 'Medical' 'Marketing' 'Technical Degree'
 'Human Resources']
===================================
Gender ['Female' 'Male']
===================================
JobRole ['Sales Executive' 'Research Scientist' 'Laboratory Technician'
 'Manufacturing Director' 'Healthcare Representative' 'Manager'
 'Sales Representative' 'Research Director' 'Human Resources']
===================================
MaritalStatus ['Single' 'Married' 'Divorced']
===================================
```

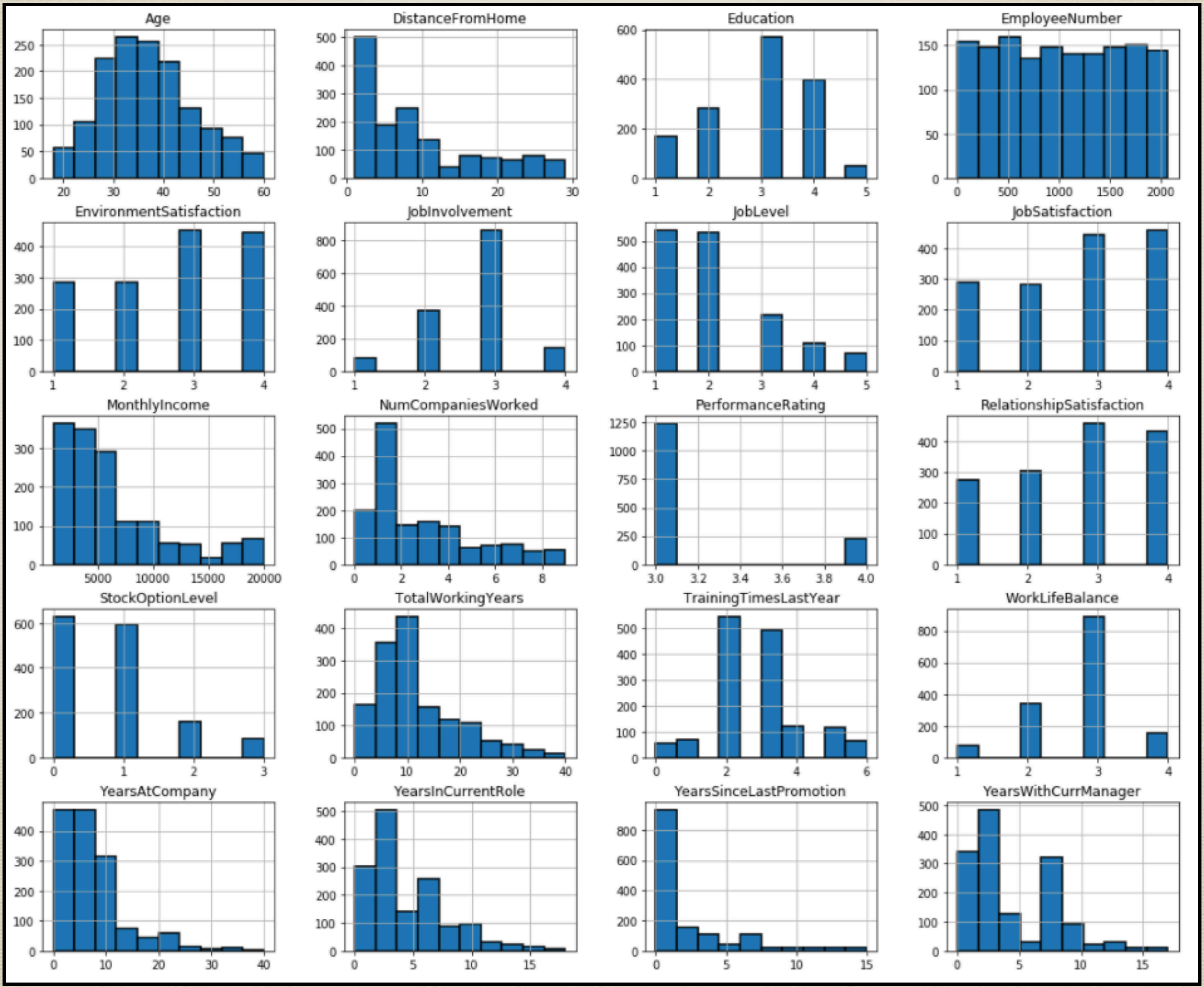
Data Description (Numeric)

- Age: 年齡
- DistanceFromHome: 公司離家距離
- Education: 'Below College’: 0, ‘College’: 1, ’Bachelor’: 2, ’Master’: 3, ‘Doctor’: 4
- EmployeeNumber: 公司人數
- EnvironmentSatisfaction: 對工作環境滿意度
- JobInvolvement: 工作參與度
- JobLevel: 職位等級
- JobSatisfaction: 對工作滿意度
- MonthlyIncome: 月收入
- NumCompaniesWorked: 曾經工作過的公司數量
- PerformanceRating: 工作表現
- RelationshipSatisfaction: 感情表現
- StockOptionLevel: 股票選擇權
- TotalWorkingYears: 總工作年數
- TrainingTimesLastYear: 去年訓練次數
- WorkLifeBalance: 工作與生活平衡滿意度
- YearsAtCompany: 待在公司年數
- YearsInCurrentRole: 待在此職位年數
- YearsSinceLastPromotion: 距離上次升職年數
- YearsWithCurrManager: 與現任經理共事年數

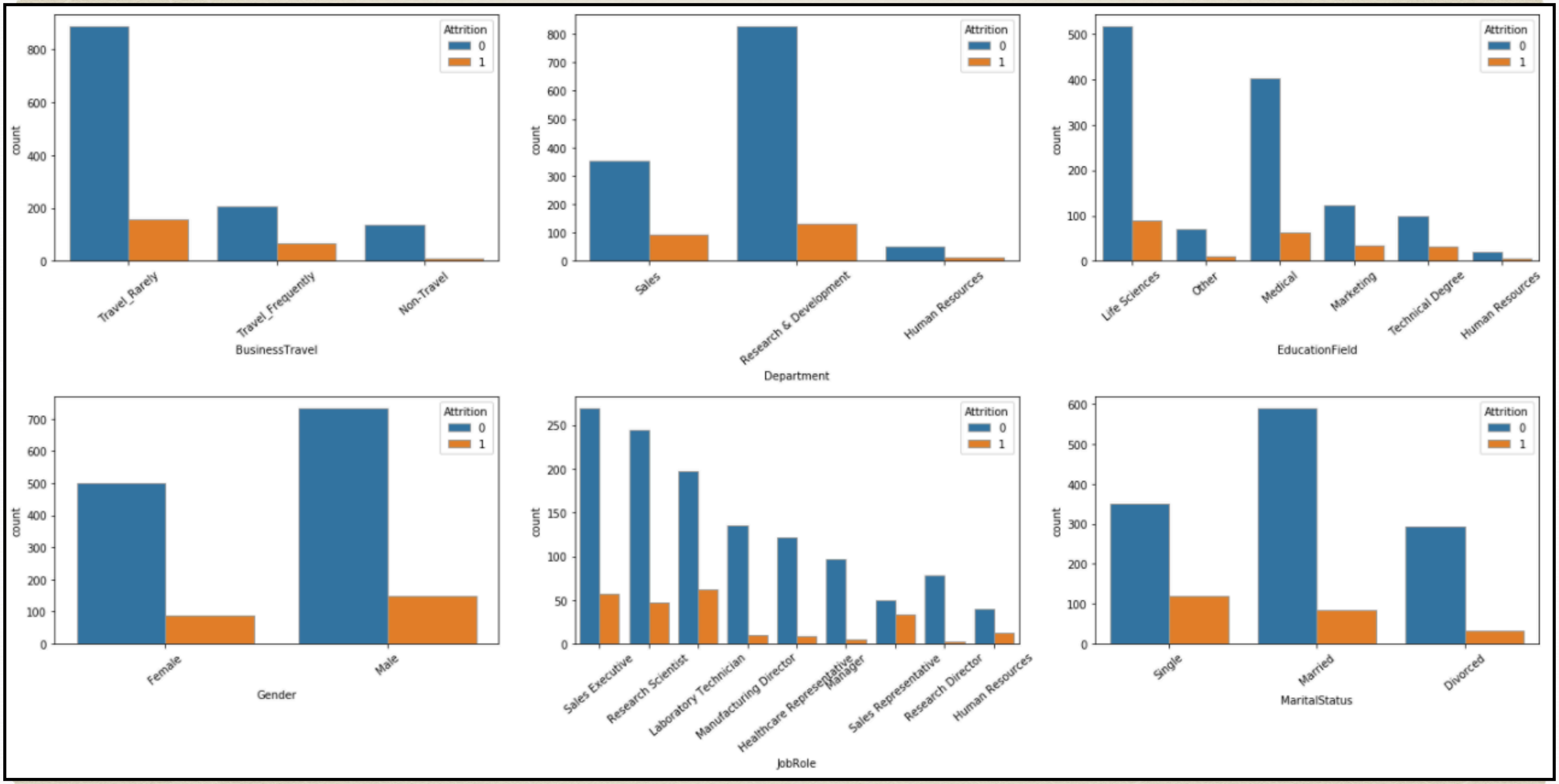
Data Imbalance

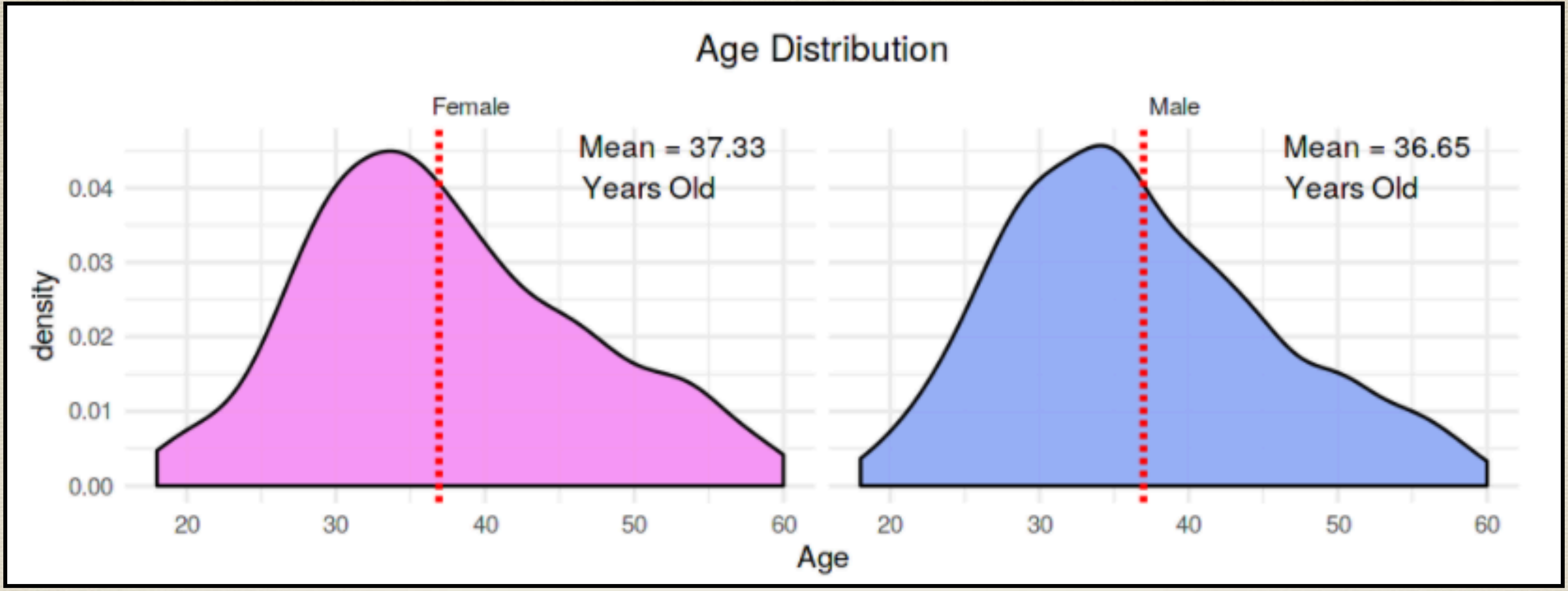


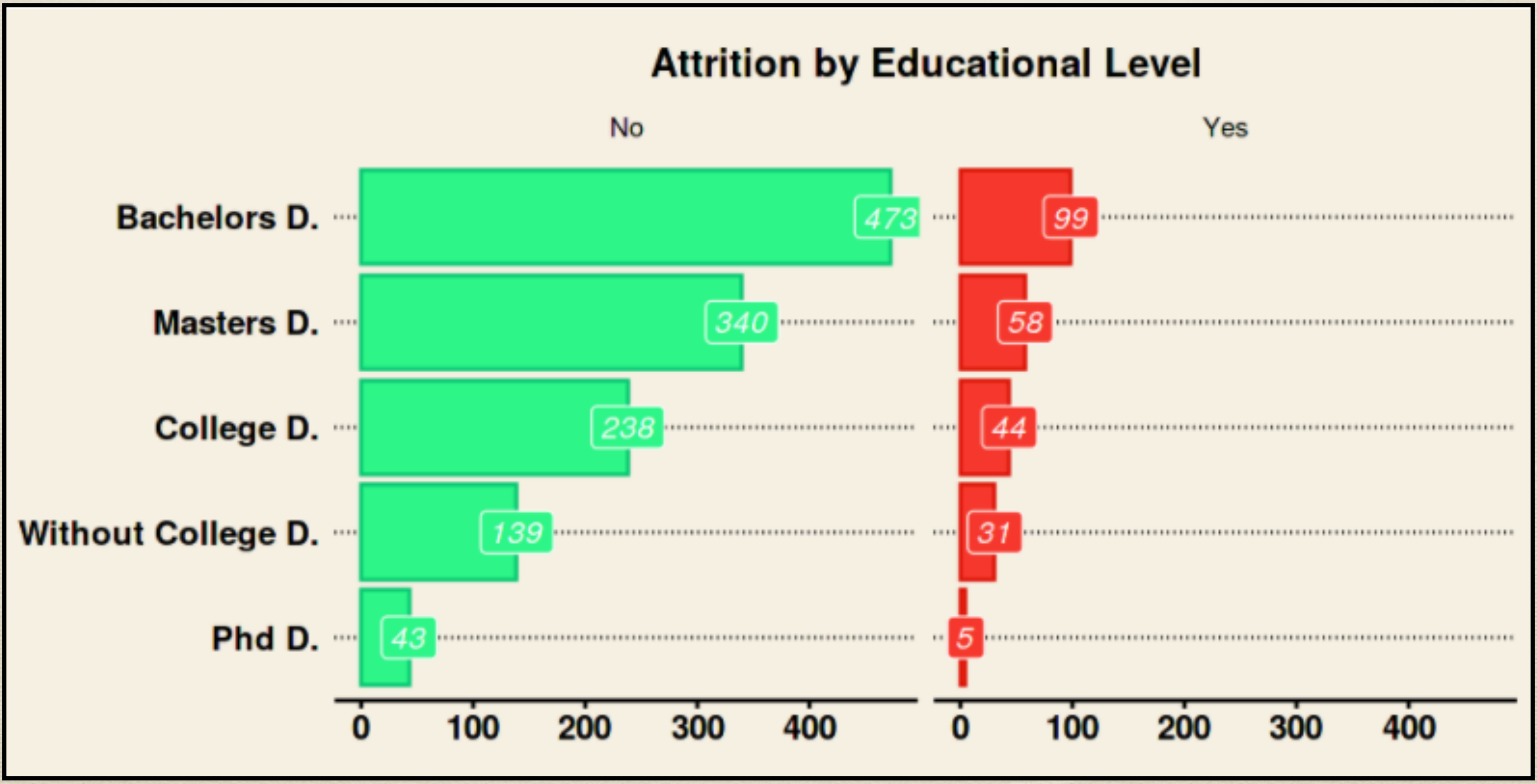
Exploration Data Analysis

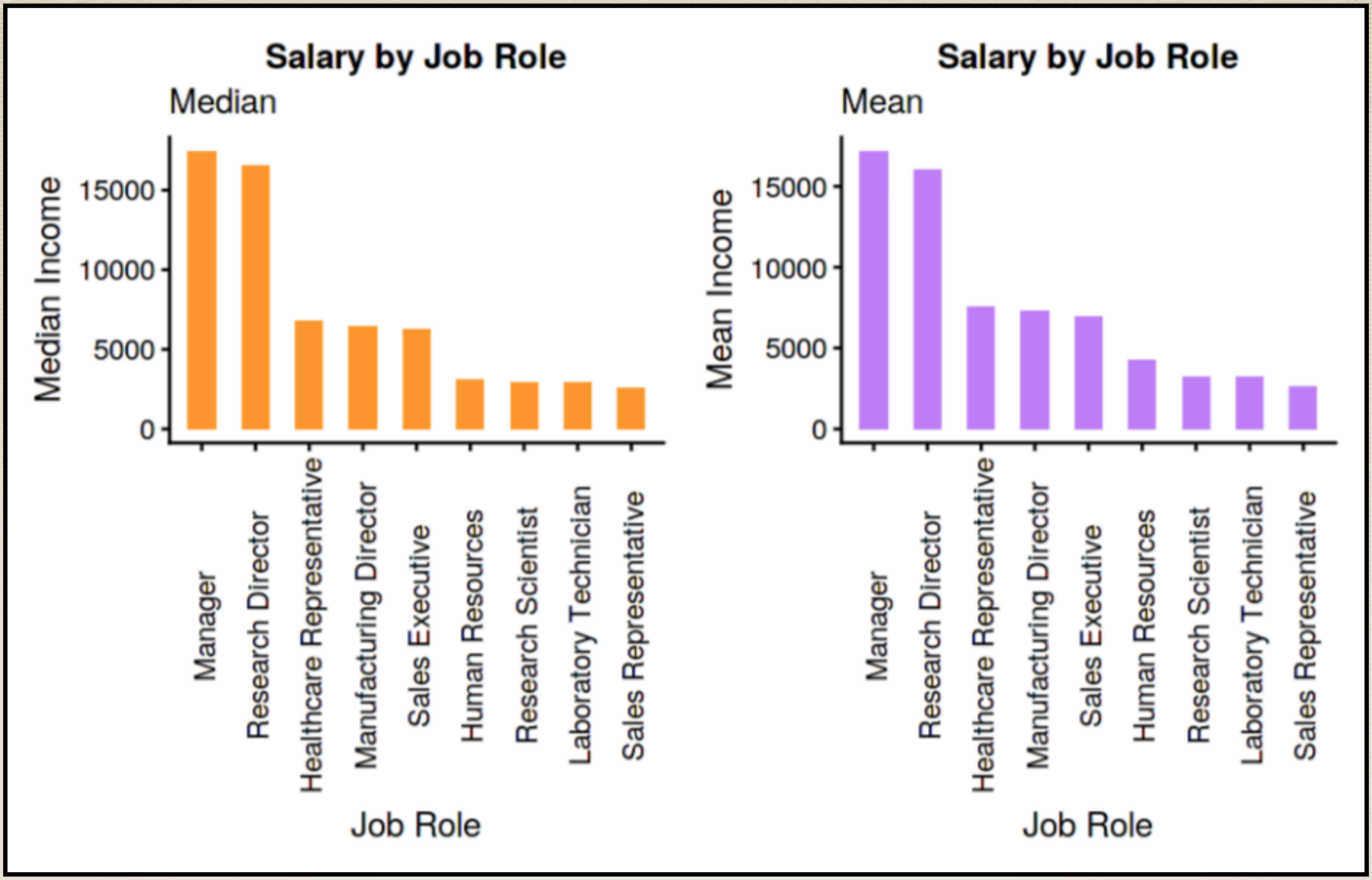


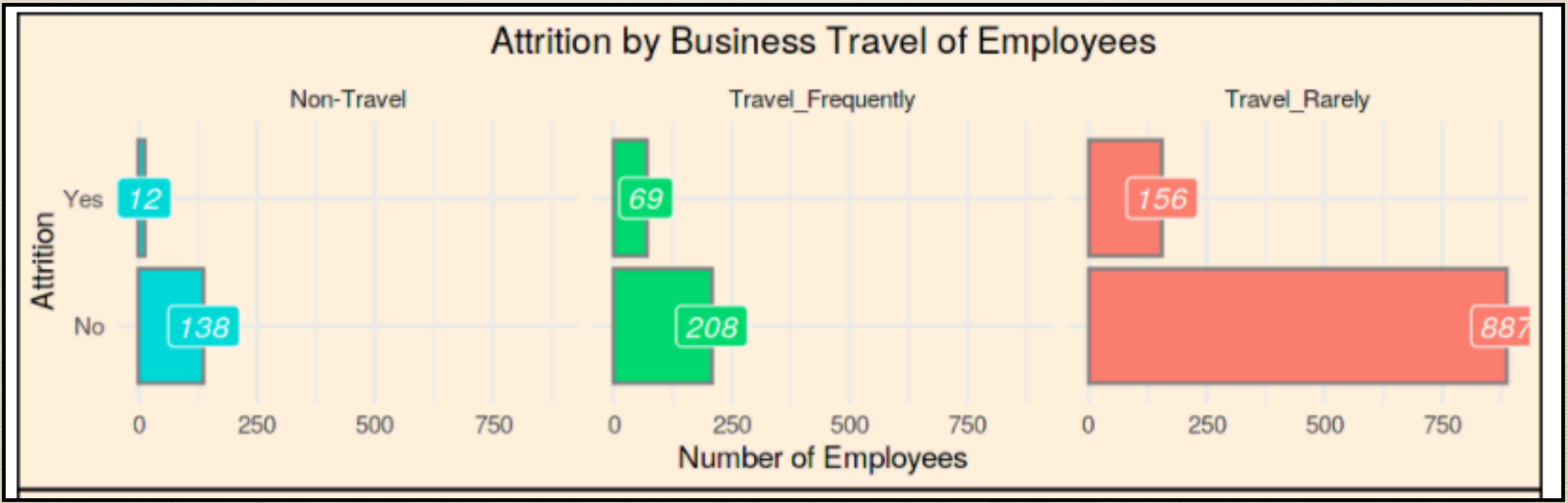
Exploration Data Analysis




Age Distribution
Female
Male
Mean = 37.33
Years Old
Mean = 36.65
Years Old
density
0.04
0.03
0.02
0.01
0.00
20
30
40
50
60
Age

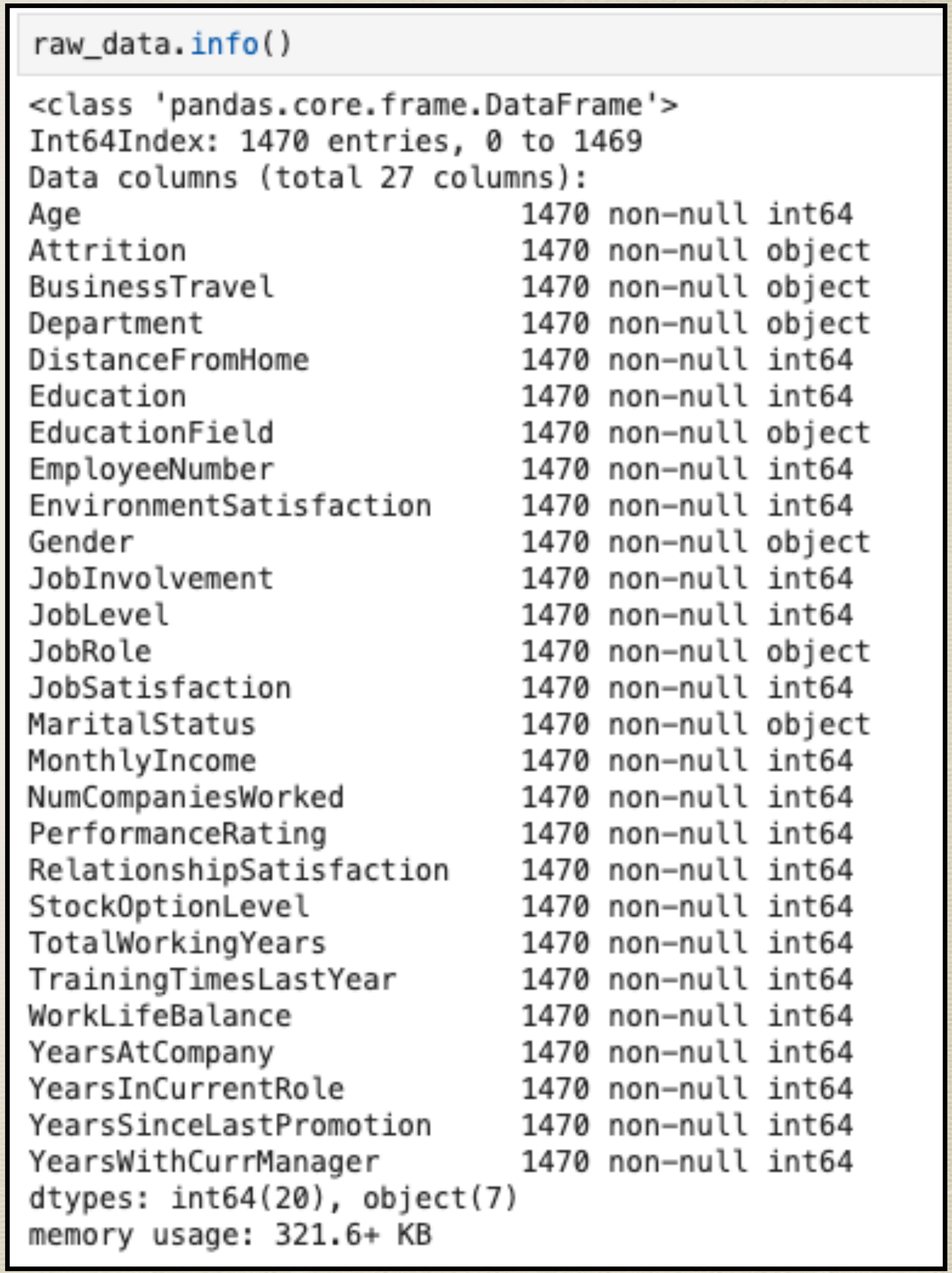

Attrition by Educational Level
No
Yes
Bachelors D.
Masters D.
College D.
Without College D.
Phd D.
473
340
238
139
43
99
58
44
31
5
0
100
200
300
400


Salary by Job Role
Median
Median Income
15000
10000
5000
0
Manager
Research Director
Healthcare Representative
Manufacturing Director
Sales Executive
Human Resources
Research Scientist
Laboratory Technician
Sales Representative
Job Role
Salary by Job Role
Mean
Mean Income
15000
10000
5000
0
Manager
Research Director
Healthcare Representative
Manufacturing Director
Sales Executive
Human Resources
Research Scientist
Laboratory Technician
Sales Representative
Job Role


Attrition by Business Travel of Employees
Non-Travel
Travel_Frequently
Travel_Rarely
Attrition
Yes
No
12
138
69
208
156
887
0
250
500
750
Number of Employees

Data Cleaning & Feature Engineering

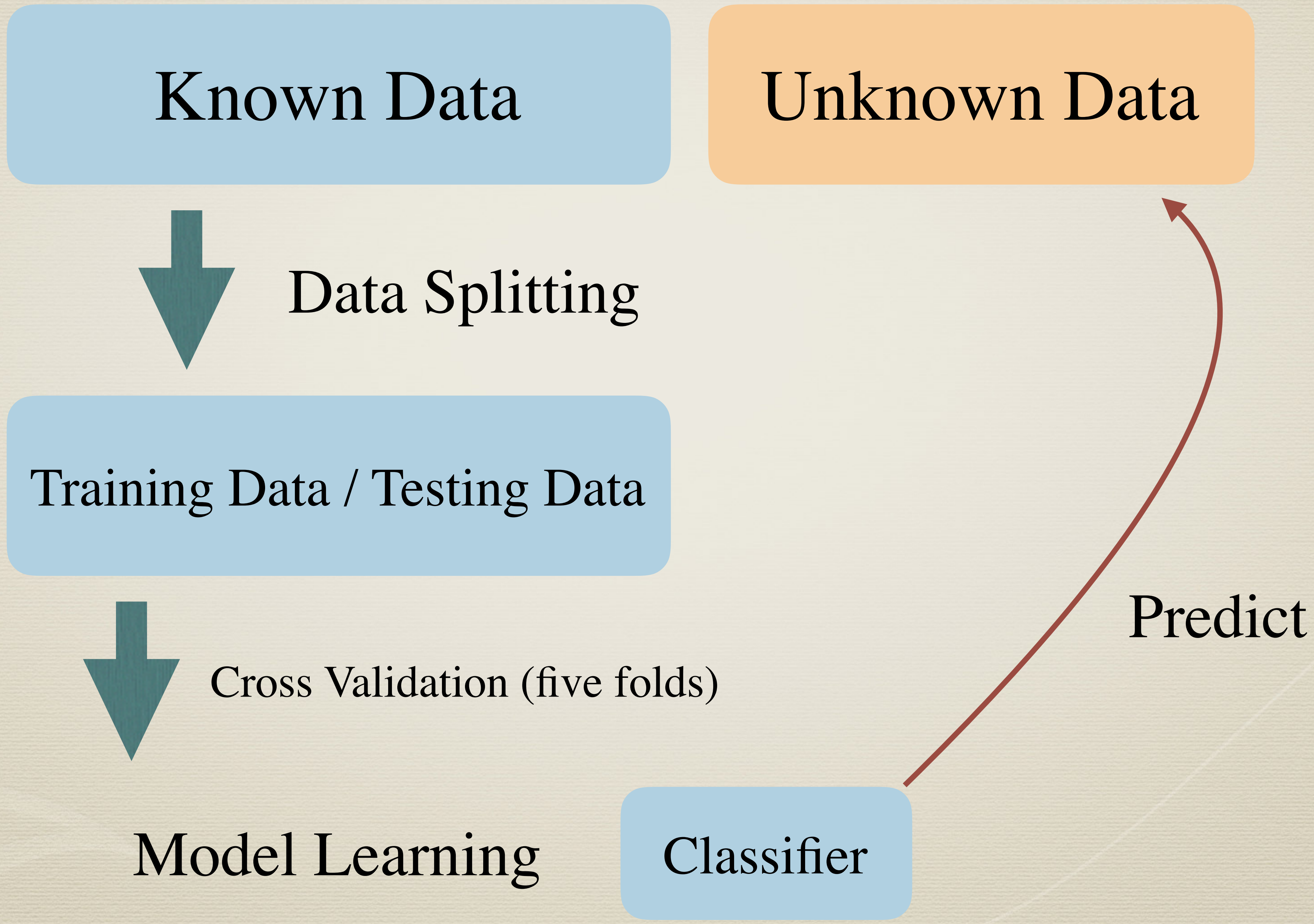
- 1. Check Missing Value and Data Type
- 2. For categorical value: Data Encoding (One-Hot Encoding)
- 3. For numerical value: Data Normalization (z-score scaling, min-max scaling)
- 4. For time stamp value: Extract year, month, date, or hour



```
raw_data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 1470 entries, 0 to 1469
Data columns (total 27 columns):
Age                         1470 non-null int64
Attrition                   1470 non-null object
BusinessTravel              1470 non-null object
Department                  1470 non-null object
DistanceFromHome            1470 non-null int64
Education                   1470 non-null int64
EducationField              1470 non-null object
EmployeeNumber              1470 non-null int64
EnvironmentSatisfaction     1470 non-null int64
Gender                      1470 non-null object
JobInvolvement              1470 non-null int64
JobLevel                    1470 non-null int64
JobRole                     1470 non-null object
JobSatisfaction             1470 non-null int64
MaritalStatus               1470 non-null object
MonthlyIncome               1470 non-null int64
NumCompaniesWorked          1470 non-null int64
PerformanceRating           1470 non-null int64
RelationshipSatisfaction    1470 non-null int64
StockOptionLevel            1470 non-null int64
TotalWorkingYears           1470 non-null int64
TrainingTimesLastYear       1470 non-null int64
WorkLifeBalance             1470 non-null int64
YearsAtCompany              1470 non-null int64
YearsInCurrentRole          1470 non-null int64
YearsSinceLastPromotion     1470 non-null int64
YearsWithCurrManager        1470 non-null int64
dtypes: int64(20), object(7)
memory usage: 321.6+ KB
```

Data Splitting & Model Learning



Known Data

Unknown Data

Data Splitting

Training Data / Testing Data

Cross Validation (five folds)

Predict

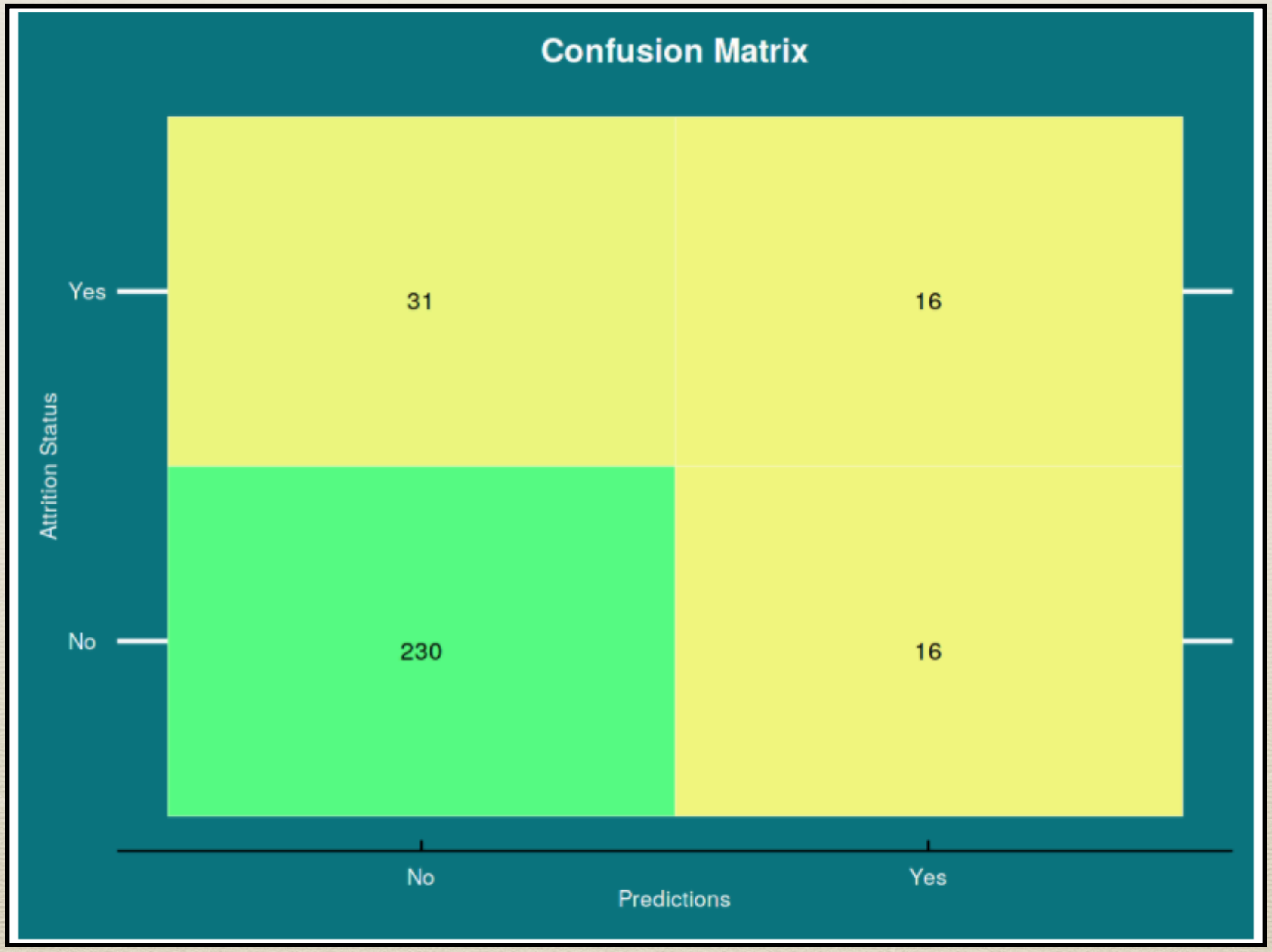
Model Learning

Classifier

Model Evaluation

- Classification Problem:
  Accuracy, Recall, Precision, F1-score, Confusion Matrix

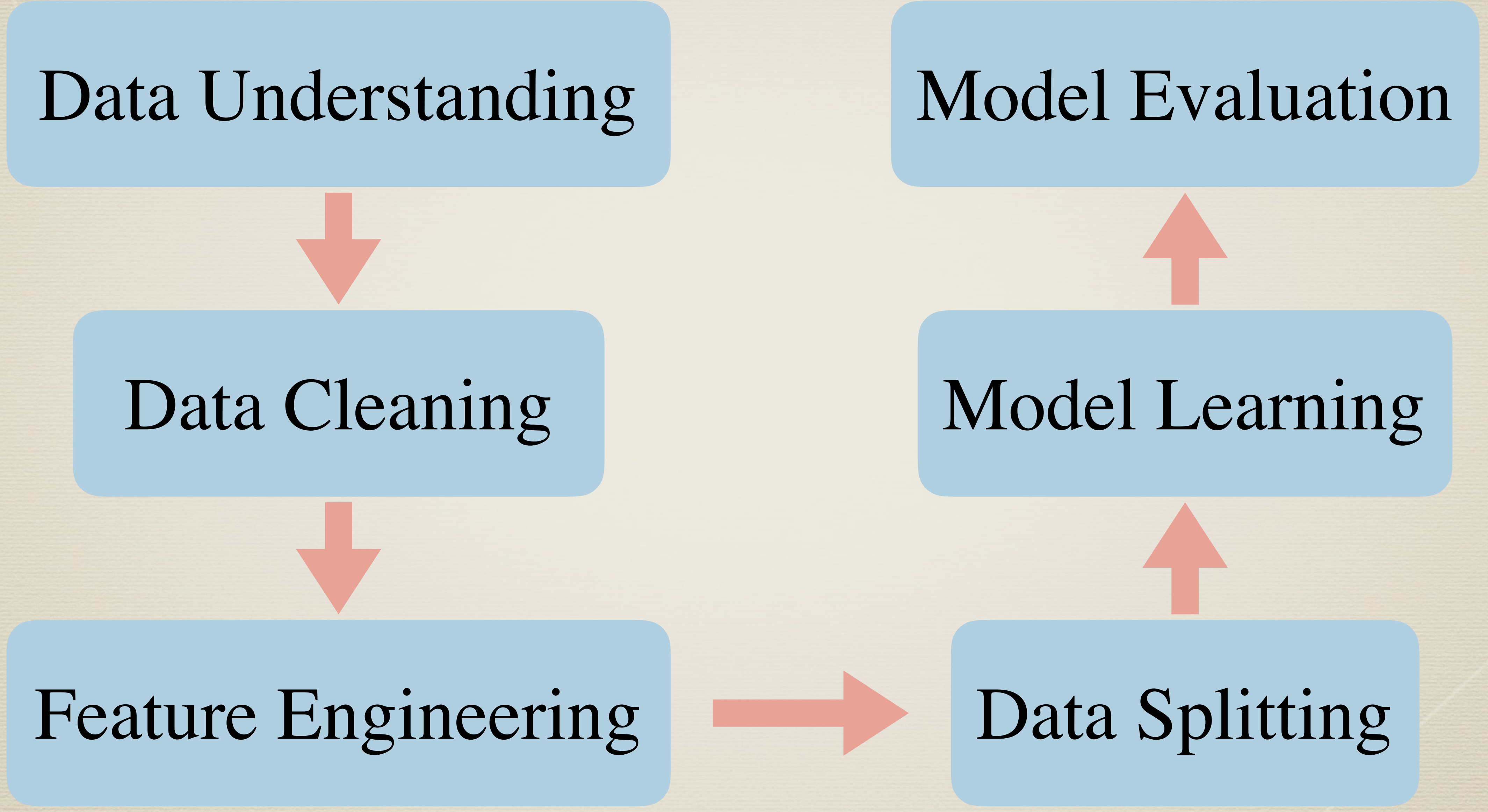
- Regression Problem:
  Error Rate, Mean Absolute Error, Mean Square Error



Confusion Matrix

Attrition Status	Predictions: No	Predictions: Yes
Yes	31	16
No	230	16

Overview



Data Understanding

Data Cleaning

Feature Engineering

Data Splitting

Model Learning

Model Evaluation