IBM HR Analytics Employee Attrition & Performance

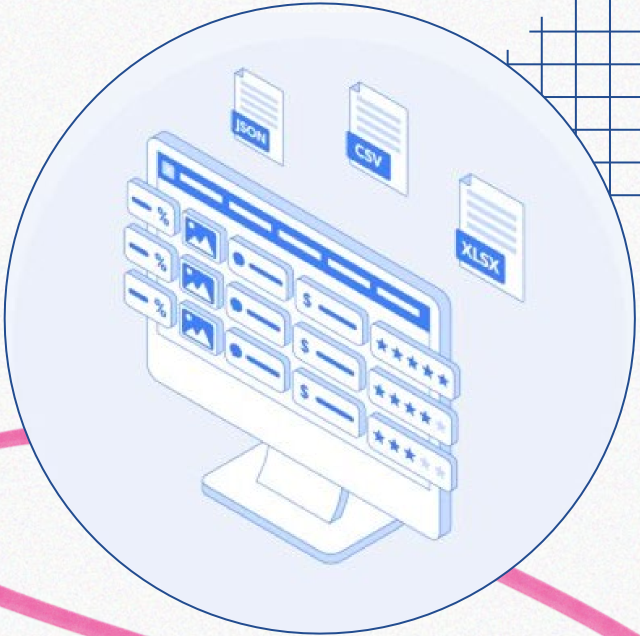
STAT 206

Cindy Miao, Haiying Lin, Susie Liang

Table of contents

01

About Dataset



Variables

Categorical (9)

- Attrition
- BusinessTravel
- Department
- EducationField
- Gender
- JobRole
- MaritalStatus
- Over18
- OverTime…

Numerical (26)

- Age
- DailyRate
- DistanceFromHome
- Education
- EmployeeCount
- EmployeeNumber
- EnvironmentSatisfaction
- HourlyRate
- JobInvolvement
- JobLevel…

```
IBM_Employee = CSV.read("WA_Fn-UseC_-HR-Employee-Attrition.csv", DataFrame, stringtype = String)
```

1470×35 DataFrame *1445 rows omitted*

Row	Age	Attrition	BusinessTravel	DailyRate	Department	DistanceFromHome	Education	EducationField	EmployeeCount	EmployeeNumber	Enviro
	Int64	String	String	Int64	String	Int64	Int64	String	Int64	Int64	Int64
1	41	Yes	Travel_Rarely	1102	Sales	1	2	Life Sciences	1	1	
2	49	No	Travel_Frequently	279	Research & Development	8	1	Life Sciences	1	2	
3	37	Yes	Travel_Rarely	1373	Research & Development	2	2	Other	1	4	
4	33	No	Travel_Frequently	1392	Research & Development	3	4	Life Sciences	1	5	
5	27	No	Travel_Rarely	591	Research & Development	2	1	Medical	1	7	
6	32	No	Travel_Frequently	1005	Research & Development	2	2	Life Sciences	1	8	
7	59	No	Travel_Rarely	1324	Research & Development	3	3	Medical	1	10	
8	30	No	Travel_Rarely	1358	Research & Development	24	1	Life Sciences	1	11	

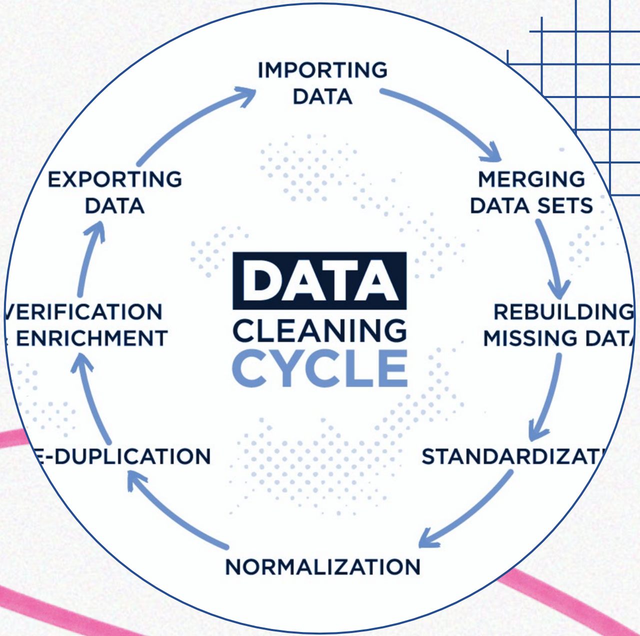
02

Problem Statement

Business Question

Based on the IBM Employee information records, analyze the factors most affect the employee attrition.

We aim to discover the relationship between employee's personal information and performance record with their attrition status. This analysis can help us get strategies to enhance understanding of employee performances, and potentially reducing overall attrition rates. This insight is crucial for creating a supportive work environment that encourages employees to stay.



03

Data Cleaning

- removed the unnecessary categories: *"EmployeeCount", "EmployeeNumber", "Over18", and "StandardHours"*

Data Cleaning

```
# remove the unnecessary categories: EmployeeCount, EmployeeNumber,Over18, and StandardHours, left 31 variables
IBM_Employee = select(IBM_Employee, Not([:EmployeeCount, :EmployeeNumber, :Over18, :StandardHours]))
```

1470×31 DataFrame

1445 rows omitted

Row	Age	Attrition	BusinessTravel	DailyRate	Department	DistanceFromHome	Education	EducationField	EnvironmentSatisfaction	Gender	HourlyRa
	Int64	String	String	Int64	String	Int64	Int64	String	Int64	String	Int64
1	41	Yes	Travel_Rarely	1102	Sales	1	2	Life Sciences	2	Female	
2	49	No	Travel_Frequently	279	Research & Development	8	1	Life Sciences	3	Male	
3	37	Yes	Travel_Rarely	1373	Research & Development	2	2	Other	4	Male	
4	33	No	Travel_Frequently	1392	Research & Development	3	4	Life Sciences	4	Female	
5	27	No	Travel_Rarely	591	Research & Development	2	1	Medical	1	Male	
6	32	No	Travel_Frequently	1005	Research & Development	2	2	Life Sciences	4	Male	
7	59	No	Travel_Rarely	1324	Research & Development	3	3	Medical	3	Female	
8	30	No	Travel_Rarely	1358	Research & Development	24	1	Life Sciences	4	Male	

- *changed Attrition to factors "Yes" = 1 & "No" = 0*

```
IBM_Employee.Attrition = map(x -> x == "Yes" ? 1 : 0, IBM_Employee.Attrition)
first(IBM_Employee, 5)
```

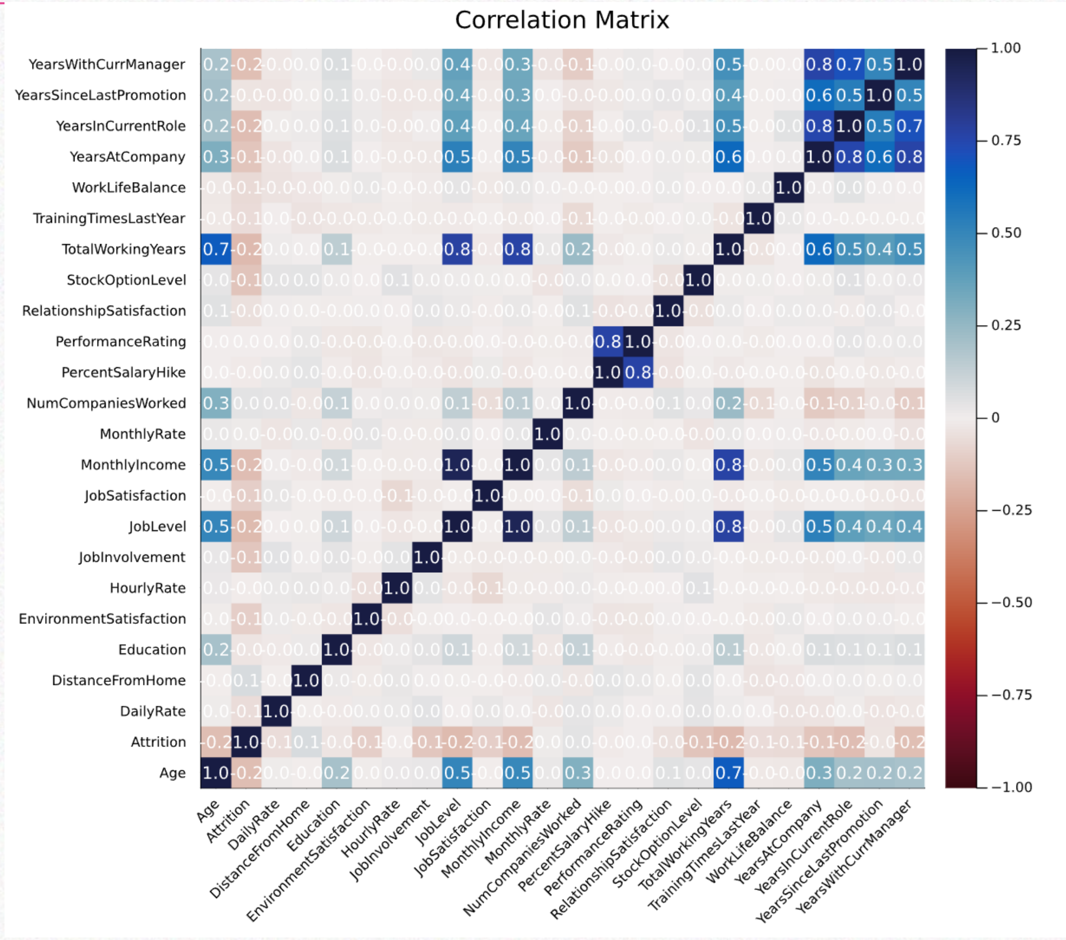
✓ 0.0s

5×31 DataFrame

Row	Age	Attrition	BusinessTravel	DailyRate	Department	DistanceFromHome	Education	EducationField
	Int64	Int64	String	Int64	String	Int64	Int64	String
1	41	1	Travel_Rarely	1102	Sales	1	2	Life Sciences
2	49	0	Travel_Frequently	279	Research & Development	8	1	Life Sciences
3	37	1	Travel_Rarely	1373	Research & Development	2	2	Other
4	33	0	Travel_Frequently	1392	Research & Development	3	4	Life Sciences
5	27	0	Travel_Rarely	591	Research & Development	2	1	Medical

04

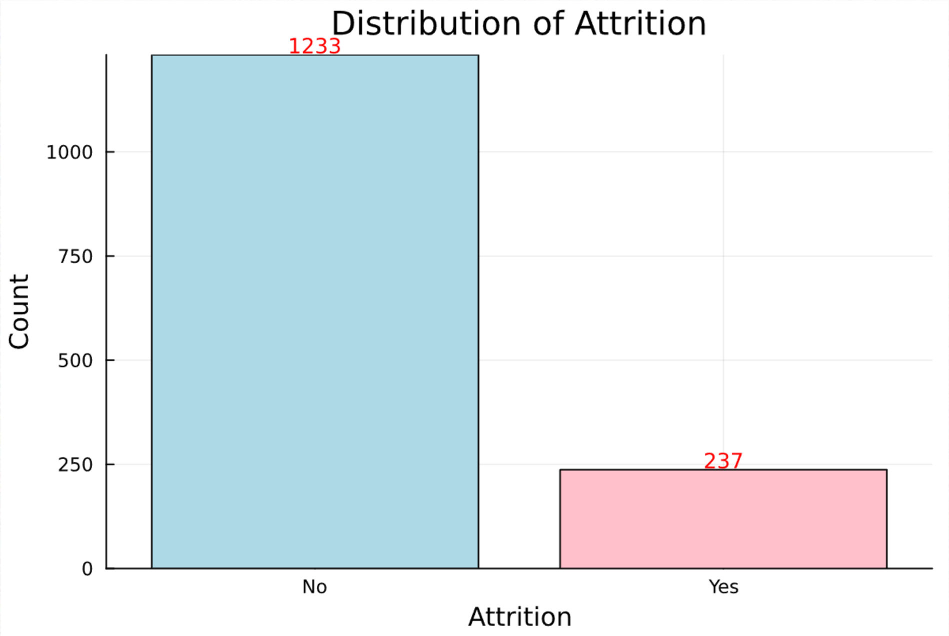
Data Exploration

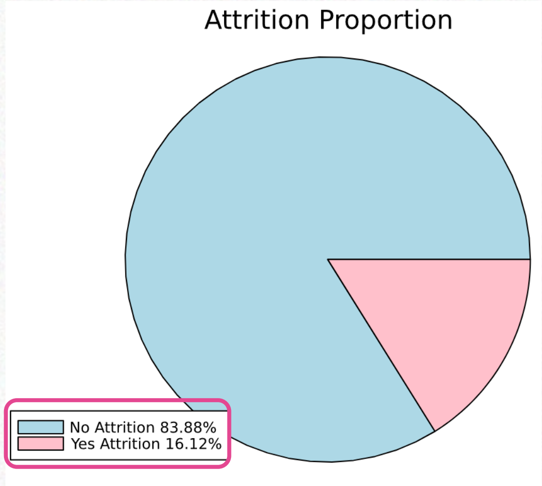


Numerical data

24 variables

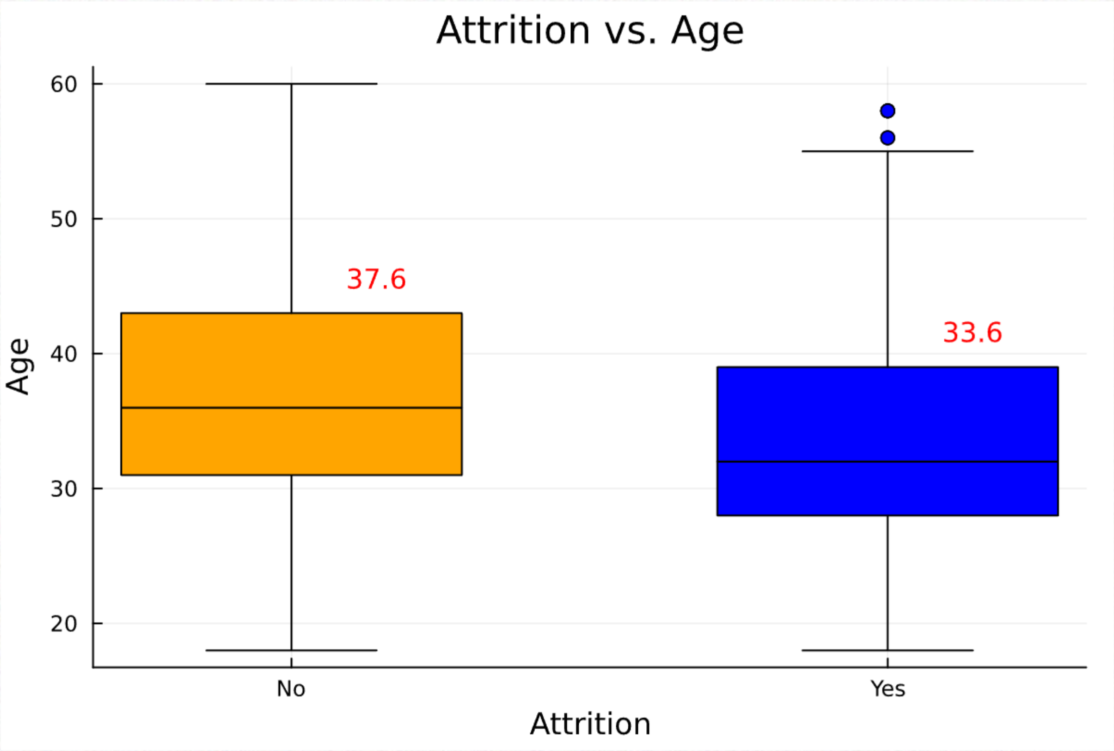
Dependent Variable —Attrition



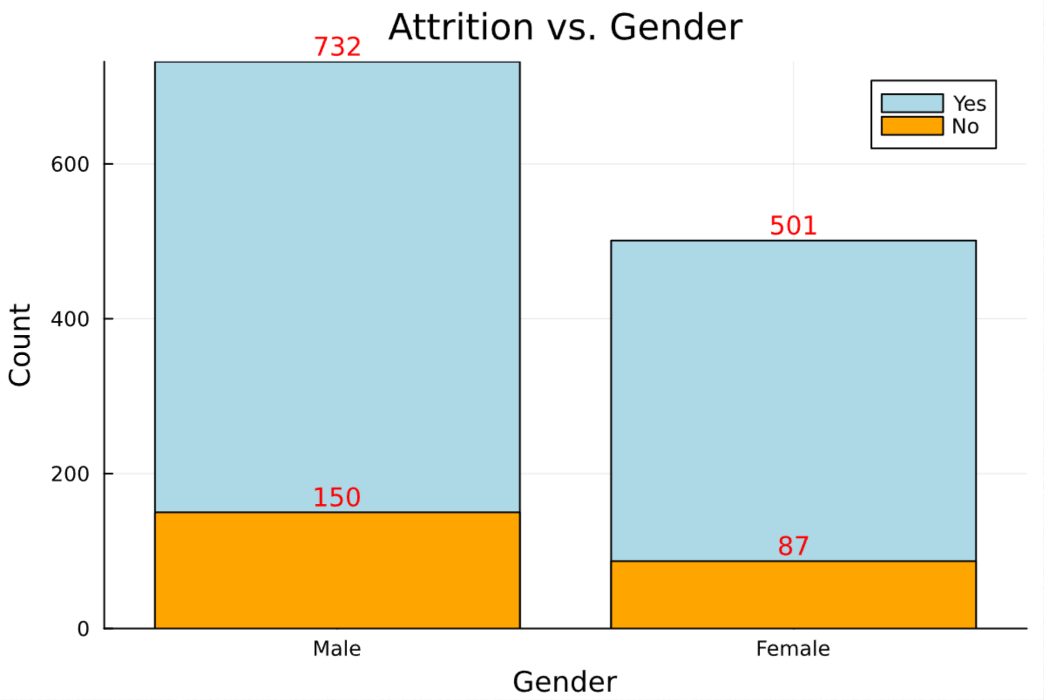


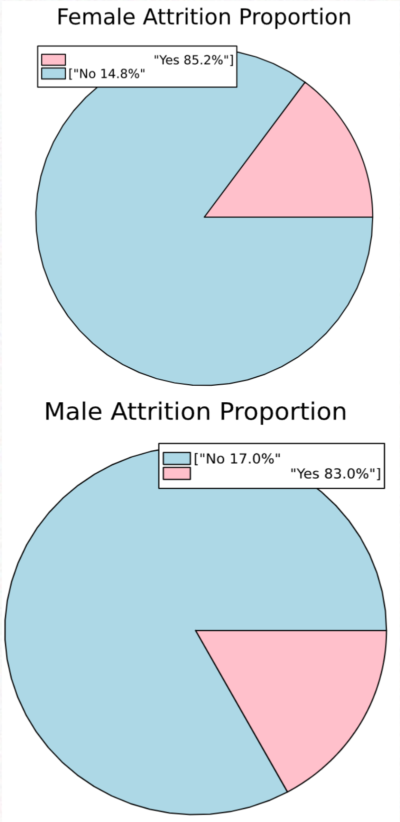
set Cut Off as 80% for prediction

Grouped Box Plot —Age

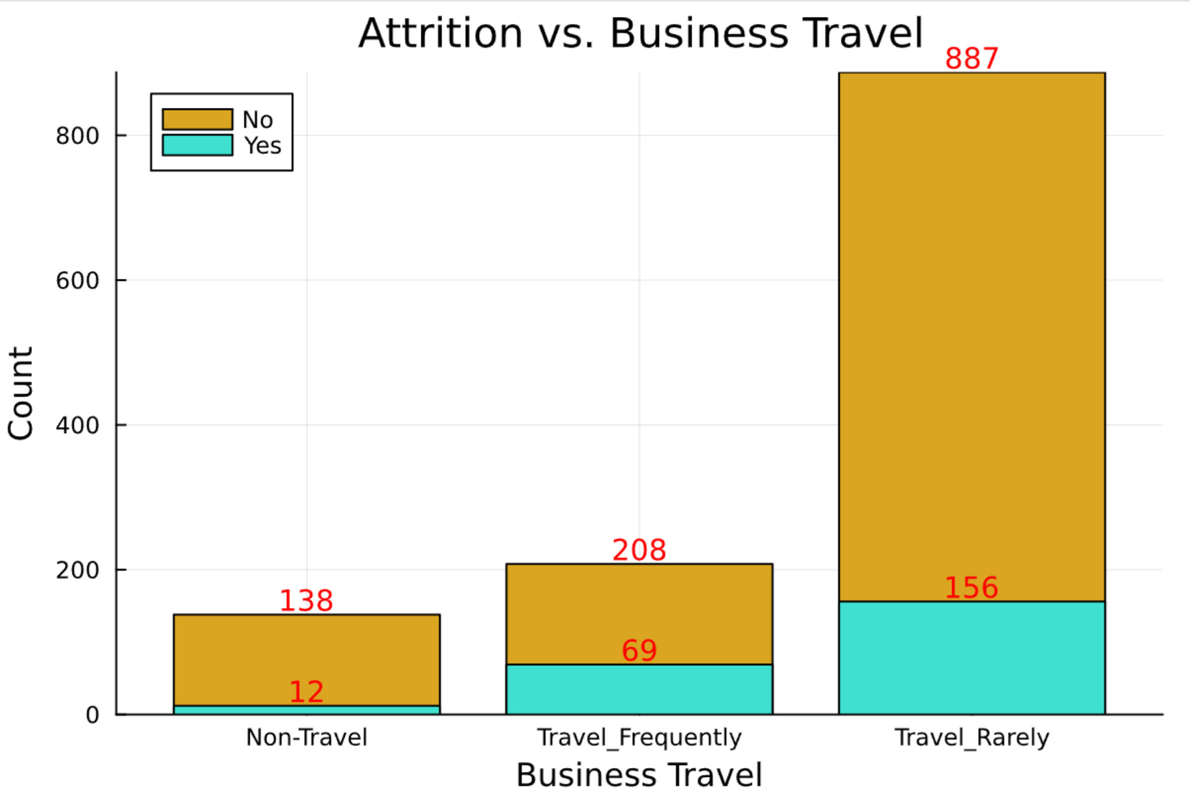


Grouped Bar & Pie Chart —Gender

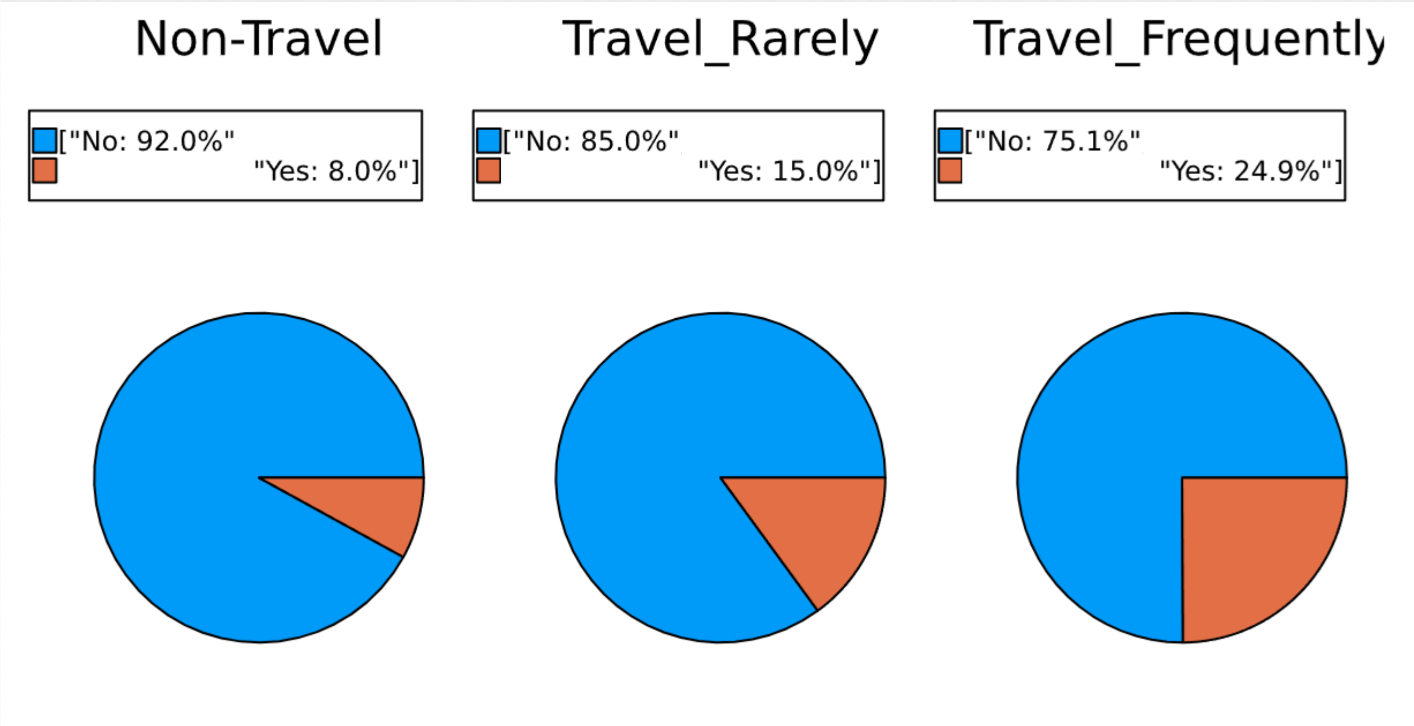




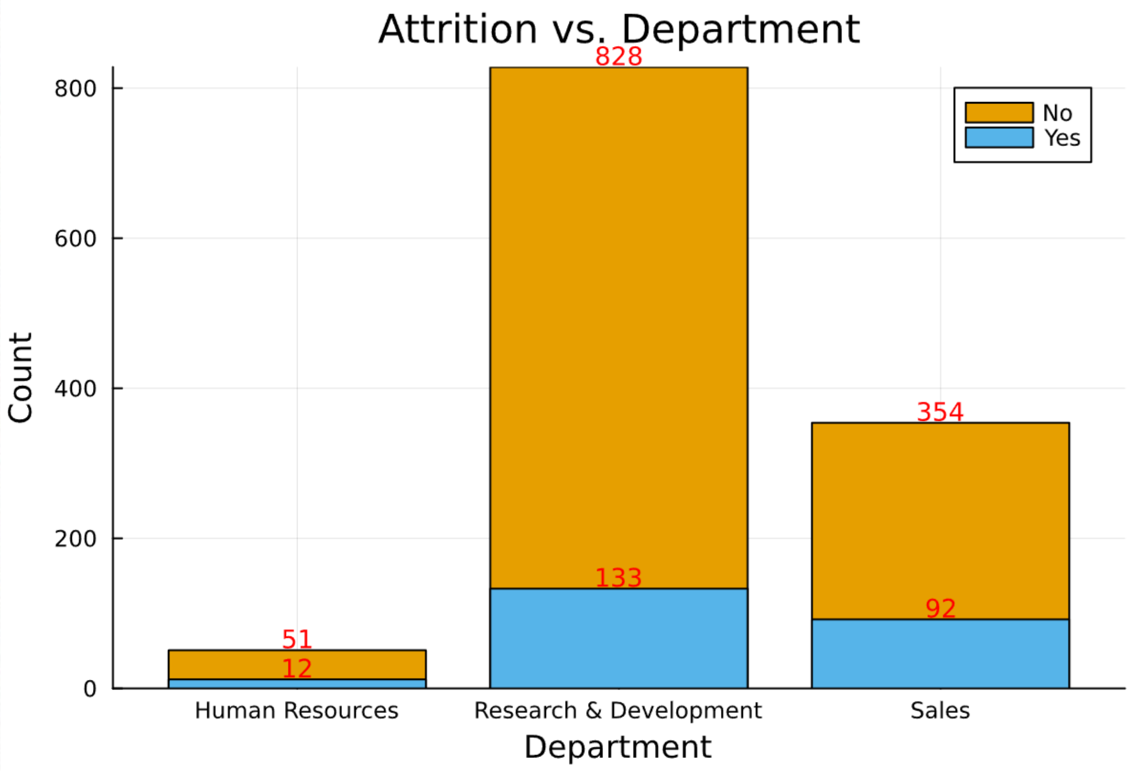
Grouped Bar Chart —Business Travel



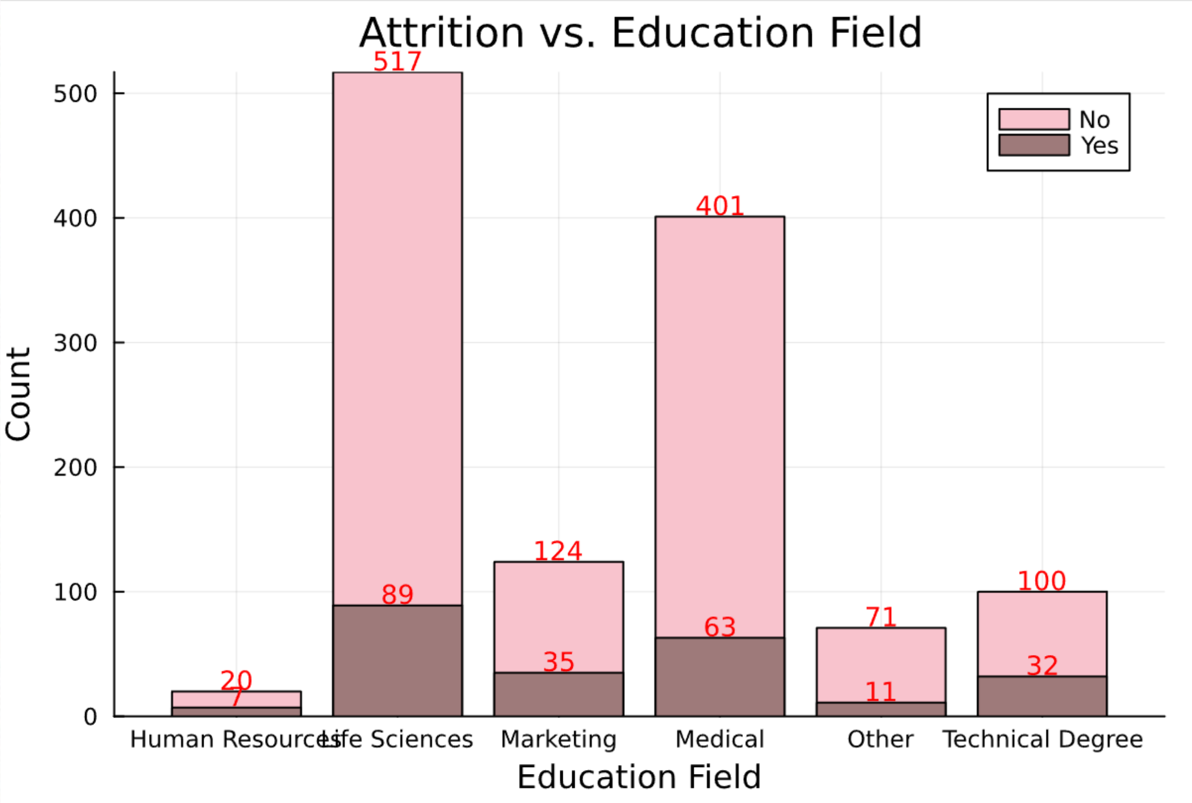
Grouped Pie Chart —Business Travel



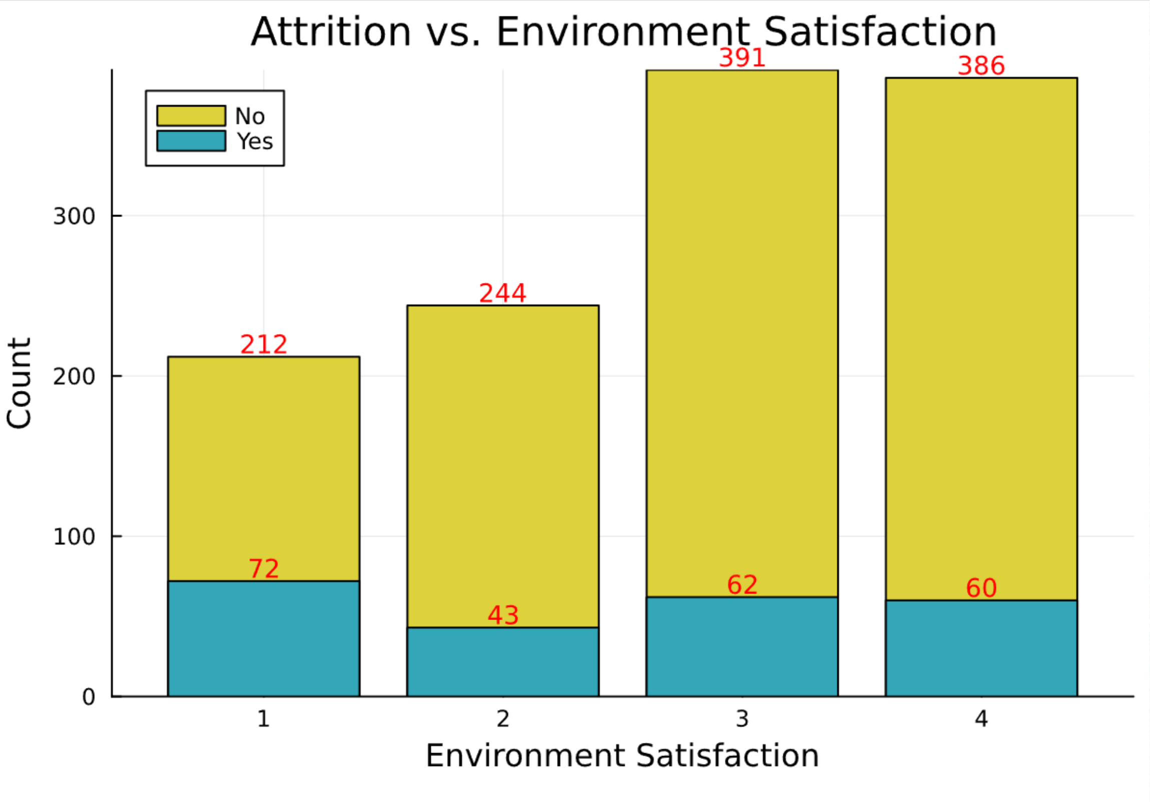
Grouped Bar Chart —Department



Grouped Bar Chart —Education Field



Grouped Bar Chart —Environment Satisfaction



05

Logistic Regression

Logistic Regression —Full Model

```
#full model for 31 variables
fullmodel = @formula(Attrition ~ Age+BusinessTravel+DailyRate+Department+DistanceFromHome
    + Education+EducationField+EnvironmentSatisfaction+Gender+HourlyRate
    + JobInvolvement + JobLevel+ JobRole+JobSatisfaction+MaritalStatus+MonthlyIncome
    +MonthlyRate+NumCompaniesWorked+OverTime+PercentSalaryHike+PerformanceRating
    +RelationshipSatisfaction+StockOptionLevel+TotalWorkingYears+TrainingTimesLastYear+ WorkLifeBalance
    +YearsAtCompany+YearsInCurrentRole+YearsSinceLastPromotion+YearsWithCurrManager)

# Fit the logistic regression model
logit = glm(fullmodel, IBM_Employee, Binomial(), ProbitLink())
```

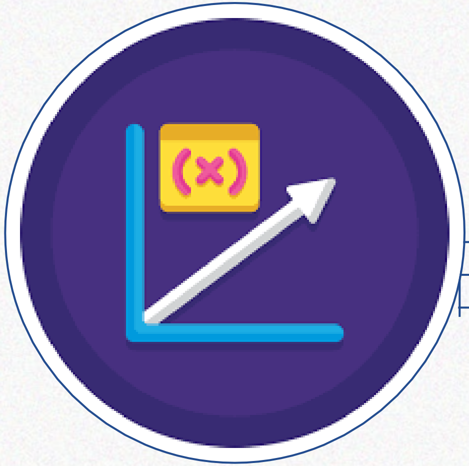
Logistic regression —Full model

Coefficients:

	Coef.	Std. Error	z	Pr(>|z|)	Lower 95%	Upper 95%
(Intercept)	-2.54191	14.8398	-0.17	0.8640	-31.6273	26.5435
Age	-0.0136459	0.007373	-1.85	0.0642	-0.0280968	0.000804874
BusinessTravel: Travel_Frequently	1.02638	0.218882	4.69	<1e-05	0.597377	1.45538
BusinessTravel: Travel_Rarely	0.591372	0.201567	2.93	0.0033	0.196309	0.986436
DailyRate	-0.000171212	0.000119246	-1.44	0.1511	-0.00040493	6.25058e-5
Department: Research & Development	3.33559	14.8188	0.23	0.8219	-25.7087	32.3799
Department: Sales	3.28092	14.8204	0.22	0.8248	-25.7666	32.3284
DistanceFromHome	0.0235123	0.00579719	4.06	<1e-04	0.01215	0.0348745
Education	0.00148223	0.0476734	0.03	0.9752	-0.0919559	0.0949204
EducationField: Life Sciences	-0.514246	0.452783	-1.14	0.2561	-1.40168	0.373193
EducationField: Marketing	-0.282085	0.478466	-0.59	0.5555	-1.21986	0.65569
EducationField: Medical	-0.532097	0.452461	-1.18	0.2396	-1.4189	0.35471
EducationField: Other	-0.508492	0.48489	-1.05	0.2943	-1.45886	0.441874
EducationField: Technical Degree	0.0565341	0.464226	0.12	0.9031	-0.853332	0.9664
EnvironmentSatisfaction	-0.233072	0.0446329	-5.22	<1e-06	-0.320551	-0.145594
Gender: Male	0.185489	0.0993257	1.87	0.0618	-0.00918555	0.380164
HourlyRate	-2.91035e-5	0.00238389	-0.01	0.9903	-0.00470144	0.00464324
JobInvolvement	-0.281007	0.0670137	-4.19	<1e-04	-0.412352	-0.149663
JobLevel	0.0200362	0.168681	0.12	0.9054	-0.310573	0.350645

Finalized Independent Variables

OverTime	JobInvolvement	NumCompanies Worked
JobRole	BusinessTravel	TotalWorking Years
Joblevel	WorkLifeBalance	EducationField
StockOption Level	DistanceFrom Home	
JobSatisfaction	Age	



Finally, we left with 13 independent variables.

Logistic Regression and Prediction

Split the data set in to train & test 70% : 30%

```
# Define the formula
fm = @formula(Attrition ~ OverTime + JobRole + JobLevel + StockOptionLevel + JobSatisfaction + JobInvolvement +
          BusinessTravel + WorkLifeBalance + DistanceFromHome + Age + NumCompaniesWorked + TotalWorkingYears + EducationField)

# Fit the logistic regression model
logit = glm(fm, train, Binomial(), ProbitLink())
```

```
# Predict the Attrition using test data
prediction = predict(logit, test)

# Converting probability score to classes
prediction_class = [if i < 0.8 0 else 1 end for i in prediction]
```

Evaluate the significant model

```
# Accuracy Score
accuracy_score = GLM.mean(prediction_df.correctly_classified)
```

```
0.854875283446712
```

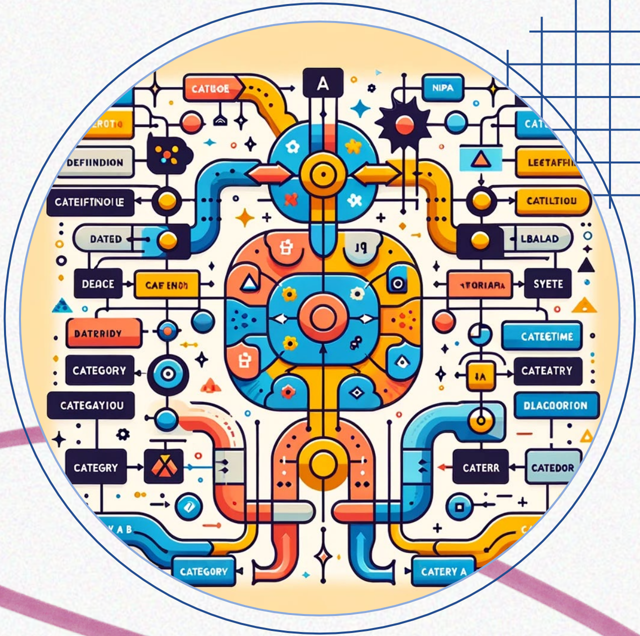
Confusion Matrix

Row	Class	Yes	No
	String	Int64	Int64
1	Yes	1	0
2	No	64	376

Accuracy: 85.48%

Sensitivity: 1.5%

Specificity: 100%



06

Classification

KNN Classifier, LDA,

Neural Network Classifier, Multinomial Classifier Models

```
IBM2 = select(IBM, Not(:Attrition))  # filter out Attrition for predictors

# change "Texual to Multiclass" for all categorical variables
coerce!(IBM2, Dict(:BusinessTravel => Multiclass, :Department => Multiclass, :EducationField => Multiclass,
                   :Gender => Multiclass, :JobRole => Multiclass, :MaritalStatus => Multiclass, :OverTime => Multiclass))
MLJ.schema(IBM2)
```

names	scitypes	types
Age	Count	Int64
BusinessTravel	Multiclass{3}	CategoricalValue{String, UInt32}
DailyRate	Count	Int64
Department	Multiclass{3}	CategoricalValue{String, UInt32}
DistanceFromHome	Count	Int64
Education	Count	Int64
EducationField	Multiclass{6}	CategoricalValue{String, UInt32}
EnvironmentSatisfaction	Count	Int64
Gender	Multiclass{2}	CategoricalValue{String, UInt32}
HourlyRate	Count	Int64
JobInvolvement	Count	Int64
JobLevel	Count	Int64
JobRole	Multiclass{9}	CategoricalValue{String, UInt32}
JobSatisfaction	Count	Int64
MaritalStatus	Multiclass{3}	CategoricalValue{String, UInt32}
MonthlyIncome	Count	Int64

Change Categorical Variables to **Multiclass** type

Create Machine Models

```julia
# Create machine using OneHotEncoder
mach = machine(OneHotEncoder(), IBM2) |> fit!
```

Age	Count	Int64
BusinessTravel__Non-Travel	Continuous	Float64
BusinessTravel__Travel_Frequently	Continuous	Float64
BusinessTravel__Travel_Rarely	Continuous	Float64
DailyRate	Count	Int64
Department__Human Resources	Continuous	Float64
Department__Research & Development	Continuous	Float64
Department__Sales	Continuous	Float64
DistanceFromHome	Count	Int64
Education	Count	Int64
EducationField__Human Resources	Continuous	Float64
EducationField__Life Sciences	Continuous	Float64
EducationField__Marketing	Continuous	Float64
EducationField__Medical	Continuous	Float64
EducationField__Other	Continuous	Float64
EducationField__Technical Degree	Continuous	Float64

BusinessTravel__Non-Travel	BusinessTravel__Travel_Frequently	BusinessTravel__Travel_Rarely	DailyRate	Department__Human Resources	Department__Research & Development
Float64	Float64	Float64	Int64	Float64	Float64
0.0	0.0	1.0	1102	0.0	0.0
0.0	1.0	0.0	279	0.0	1.0
0.0	0.0	1.0	1373	0.0	1.0
0.0	1.0	0.0	1392	0.0	1.0
0.0	0.0	1.0	591	0.0	1.0
0.0	1.0	0.0	1005	0.0	1.0
0.0	0.0	1.0	1324	0.0	1.0
0.0	0.0	1.0	1358	0.0	1.0
0.0	1.0	0.0	216	0.0	1.0
0.0	0.0	1.0	1299	0.0	1.0
0.0	0.0	1.0	809	0.0	1.0
0.0	0.0	1.0	153	0.0	1.0
0.0	0.0	1.0	670	0.0	1.0
⋮	⋮	⋮	⋮	⋮	⋮
0.0	0.0	1.0	287	0.0	1.0
0.0	0.0	1.0	1378	0.0	1.0
0.0	0.0	1.0	468	0.0	1.0

Confusion Matrix for each model

```
# ConfusionMatrix for each model
mat[1]  # KNNClassifier
```

	Ground Truth	
Predicted	Yes	No
Yes	11	14
No	61	355

```
mat[2]  # LDA
```

	Ground Truth	
Predicted	Yes	No
Yes	58	161
No	14	208

```
mat[3]  # NeuralNetworkClassifier
```

	Ground Truth	
Predicted	Yes	No
Yes	0	0
No	72	369

```
mat[4]  # MultinomialClassifier
```

	Ground Truth	
Predicted	Yes	No
Yes	0	0
No	72	369

Model Performances

```
# Perform Accuracy, Precision, Recall, F1 Results
results = DataFrame(
    Model = typeof.(model_list),
    Accuracy = acc,
    Precision = pre,
    Recall = rec,
    F1 = f1s
)
```

4×5 DataFrame

Row	Model	Accuracy	Precision	Recall	F1
	DataType	Float64	Float64	Float64	Float64
1	KNNClassifier	0.829932	0.646683	0.557419	0.565631
2	LDA	0.603175	0.600889	0.684621	0.551259
3	NeuralNetworkClassifier{Short, typeof(softmax), Adam, typeof(crossentropy)}	0.836735	0.418367	0.5	0.455556
4	MultinomialClassifier	0.836735	0.418367	0.5	0.455556

08

Conclusion & Discussion

5 Ways to Reduce Employee Attrition

Healthy organizations have an attrition rate of 10% or less

16% to 10%

- Decrease work overtime
- Improve employee Job satisfaction
- Improve employee Job Involvement
- Decrease business travel times
- Appropriately arrange work address

Q & A