Project Title : Employee Attrition Prediction Using Py Spark

Student name: Khalid Hassan Suliman .

Roll number:- 03

Registration number:12200160.

Project Description :-

Employee attrition when employees leave an organization—is a major challenge for businesses. High turnover increases recruitment costs, disrupts workflow, and reduces productivity.
This project focuses on building a **machine learning model using Py Spark** to predict whether an employee is likely to leave the company based on job-related and demographic factors.

The system uses a **Decision Tree Classifier** to analyse patterns in attributes such as salary, age, experience, department, and job role. Visualization tools are also used to explore trends and provide insights into attrition patterns.

Problem Statement:-

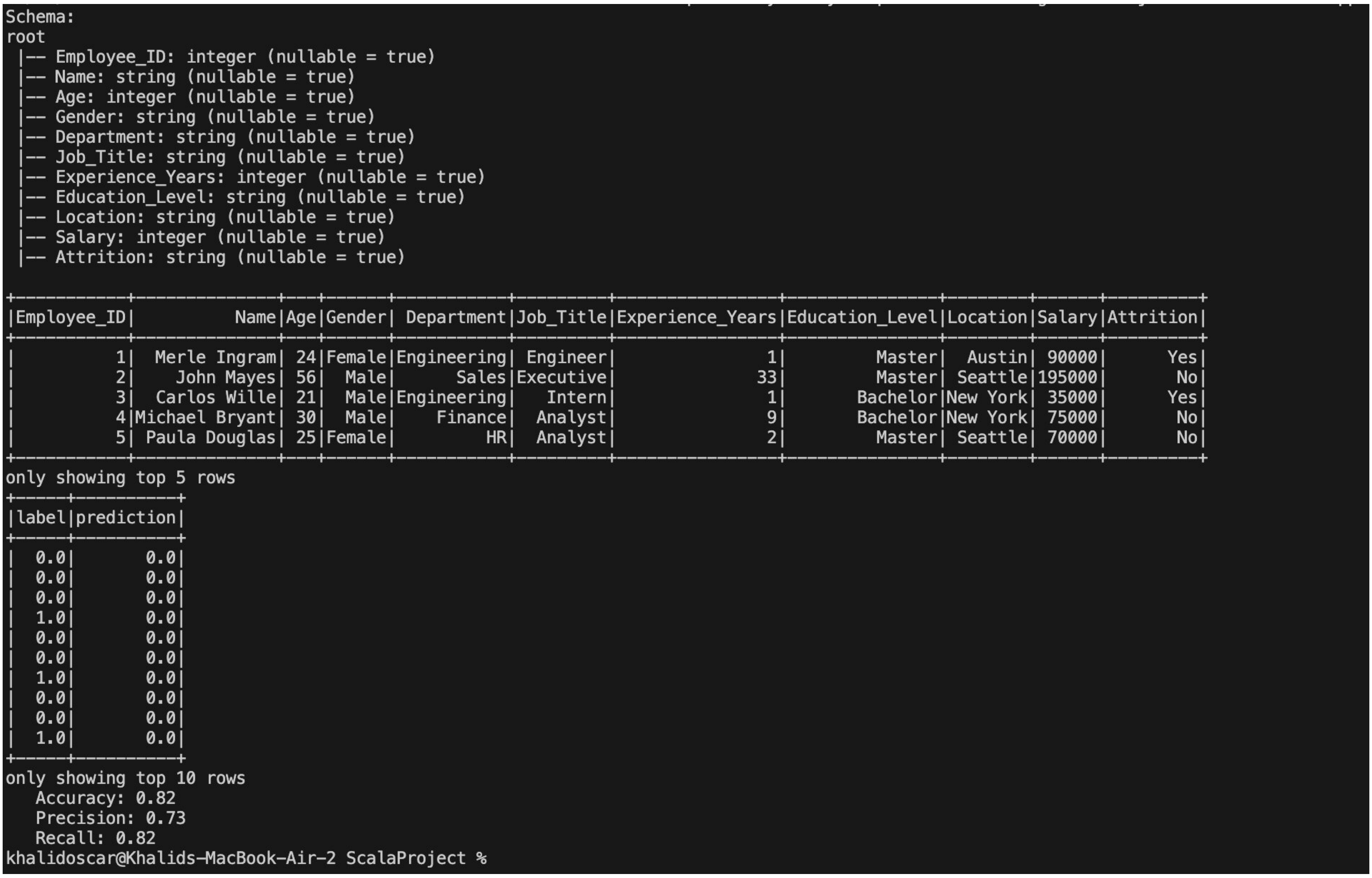
Employee attrition is a critical challenge for organizations, directly affecting productivity, recruitment costs, and overall organizational stability. Understanding the factors that influence employee turnover enables companies to make informed decisions to improve workplace satisfaction and retention. This project aims to **develop a predictive model that determines whether an employee is likely to leave the organization**, based on key attributes such as **job satisfaction**, **salary level**, **work experience**, and other relevant demographic and professional factors.

Project code:-

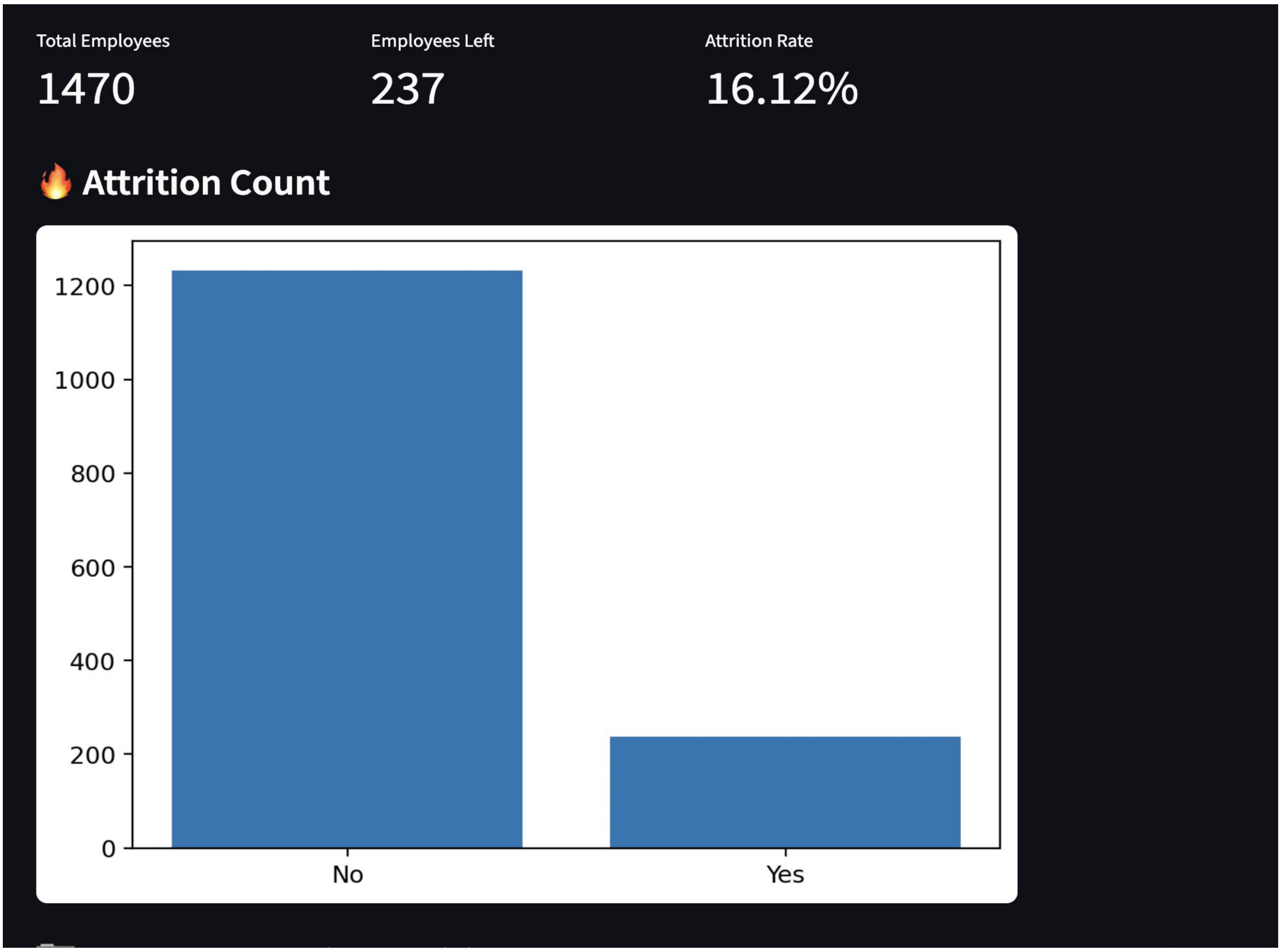
```
from pyspark.sql import SparkSession
from pyspark.sql.functions import col
from pyspark.ml.feature import StringIndexer, VectorAssembler
from pyspark.ml.classification import DecisionTreeClassifier
from pyspark.ml.evaluation import MulticlassClassificationEvaluator
from pyspark.ml import Pipeline
import matplotlib.pyplot as plt
import pandas as pd

spark =
SparkSession.builder.appName("EmployeeAttritionPrediction").getOrCreate()

data = spark.read.csv(
   "/Users/khalidoscar/Downloads/Employers_data.csv",
   header=True,
   inferSchema=True
)

data = data.dropna()

categorical_cols = ["Gender", "Department", "Job_Title", "Location",
"Education_Level", "Attrition"]

indexers = [
   StringIndexer(inputCol=col_name, outputCol=col_name + "Index")
   for col_name in categorical_cols
]
```

```
pipeline = Pipeline(stages=indexers)
data = pipeline.fit(data).transform(data)

data = data.withColumnRenamed("AttritionIndex", "label")

feature_cols = ["Age", "Experience_Years", "Salary",
                "GenderIndex", "DepartmentIndex", "Job_TitleIndex",
                "LocationIndex", "Education_LevelIndex"]

assembler = VectorAssembler(inputCols=feature_cols, outputCol="features")
final_data = assembler.transform(data).select("features", "label")

train_data, test_data = final_data.randomSplit([0.7, 0.3], seed=42)

dt = DecisionTreeClassifier(labelCol="label", featuresCol="features")
model = dt.fit(train_data)

predictions = model.transform(test_data)
predictions.select("label", "prediction").show(10)

evaluator = MulticlassClassificationEvaluator(labelCol="label",
predictionCol="prediction")

accuracy = evaluator.evaluate(predictions, {evaluator.metricName: "accuracy"})
precision = evaluator.evaluate(predictions, {evaluator.metricName:
"weightedPrecision"})
recall = evaluator.evaluate(predictions, {evaluator.metricName: "weightedRecall"})

print(f"Accuracy: {accuracy:.2f}")
print(f"Precision: {precision:.2f}")
print(f"Recall: {recall:.2f}")

attrition_trend = data.groupBy("Attrition").count().toPandas()
plt.figure(figsize=(6,4))
plt.bar(attrition_trend["Attrition"], attrition_trend["count"], color=["green", "red"])
plt.xlabel("Attrition")
plt.ylabel("Count")
plt.show()

dept_trend = data.groupBy("Department", "Attrition").count().toPandas()
dept_pivot = dept_trend.pivot(index="Department", columns="Attrition",
values="count").fillna(0)
dept_pivot.plot(kind="bar", figsize=(8,5))
plt.xlabel("Department")
plt.ylabel("Count")
plt.show()

age_df = data.select("Age").toPandas()
plt.hist(age_df["Age"], bins=15)
plt.xlabel("Age")
plt.ylabel("Frequency")
plt.show()

salary_attr = data.select("Salary", "Attrition").toPandas()
salary_attr.boxplot(by="Attrition", column=["Salary"])
plt.xlabel("Attrition (No / Yes)")
plt.ylabel("Salary")
plt.suptitle("")
plt.show()

exp_salary = data.select("Experience_Years", "Salary").toPandas()
plt.scatter(exp_salary["Experience_Years"], exp_salary["Salary"])
plt.xlabel("Experience (Years)")
plt.ylabel("Salary")
plt.show()
```

Output:-



```
Schema:
root
 |-- Employee_ID: integer (nullable = true)
 |-- Name: string (nullable = true)
 |-- Age: integer (nullable = true)
 |-- Gender: string (nullable = true)
 |-- Department: string (nullable = true)
 |-- Job_Title: string (nullable = true)
 |-- Experience_Years: integer (nullable = true)
 |-- Education_Level: string (nullable = true)
 |-- Location: string (nullable = true)
 |-- Salary: integer (nullable = true)
 |-- Attrition: string (nullable = true)

+-----------+--------------+---+------+-----------+---------+----------------+---------------+--------+------+---------+
|Employee_ID|          Name|Age|Gender| Department|Job_Title|Experience_Years|Education_Level|Location|Salary|Attrition|
+-----------+--------------+---+------+-----------+---------+----------------+---------------+--------+------+---------+
|          1|  Merle Ingram| 24|Female|Engineering| Engineer|               1|         Master|  Austin| 90000|      Yes|
|          2|    John Mayes| 56|  Male|      Sales|Executive|              33|         Master| Seattle|195000|       No|
|          3|  Carlos Wille| 21|  Male|Engineering|   Intern|               1|       Bachelor|New York| 35000|      Yes|
|          4|Michael Bryant| 30|  Male|    Finance|  Analyst|               9|       Bachelor|New York| 75000|       No|
|          5| Paula Douglas| 25|Female|         HR|  Analyst|               2|         Master| Seattle| 70000|       No|
+-----------+--------------+---+------+-----------+---------+----------------+---------------+--------+------+---------+
only showing top 5 rows
+-----+----------+
|label|prediction|
+-----+----------+
|  0.0|       0.0|
|  0.0|       0.0|
|  0.0|       0.0|
|  1.0|       0.0|
|  0.0|       0.0|
|  0.0|       0.0|
|  1.0|       0.0|
|  0.0|       0.0|
|  0.0|       0.0|
|  1.0|       0.0|
+-----+----------+
only showing top 10 rows
   Accuracy: 0.82
   Precision: 0.73
   Recall: 0.82
khalidoscar@Khalids-MacBook-Air-2 ScalaProject %
```



Total Employees

1470

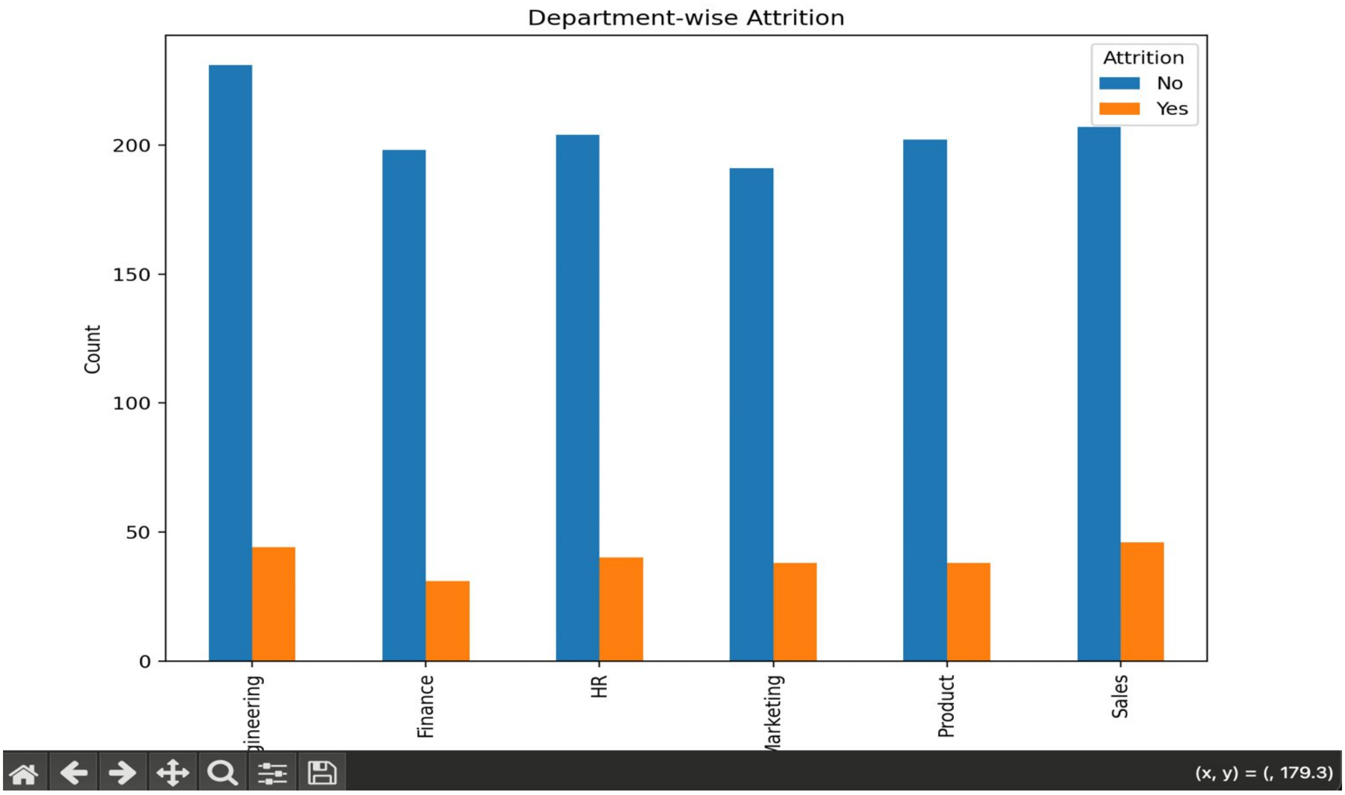
Employees Left

237

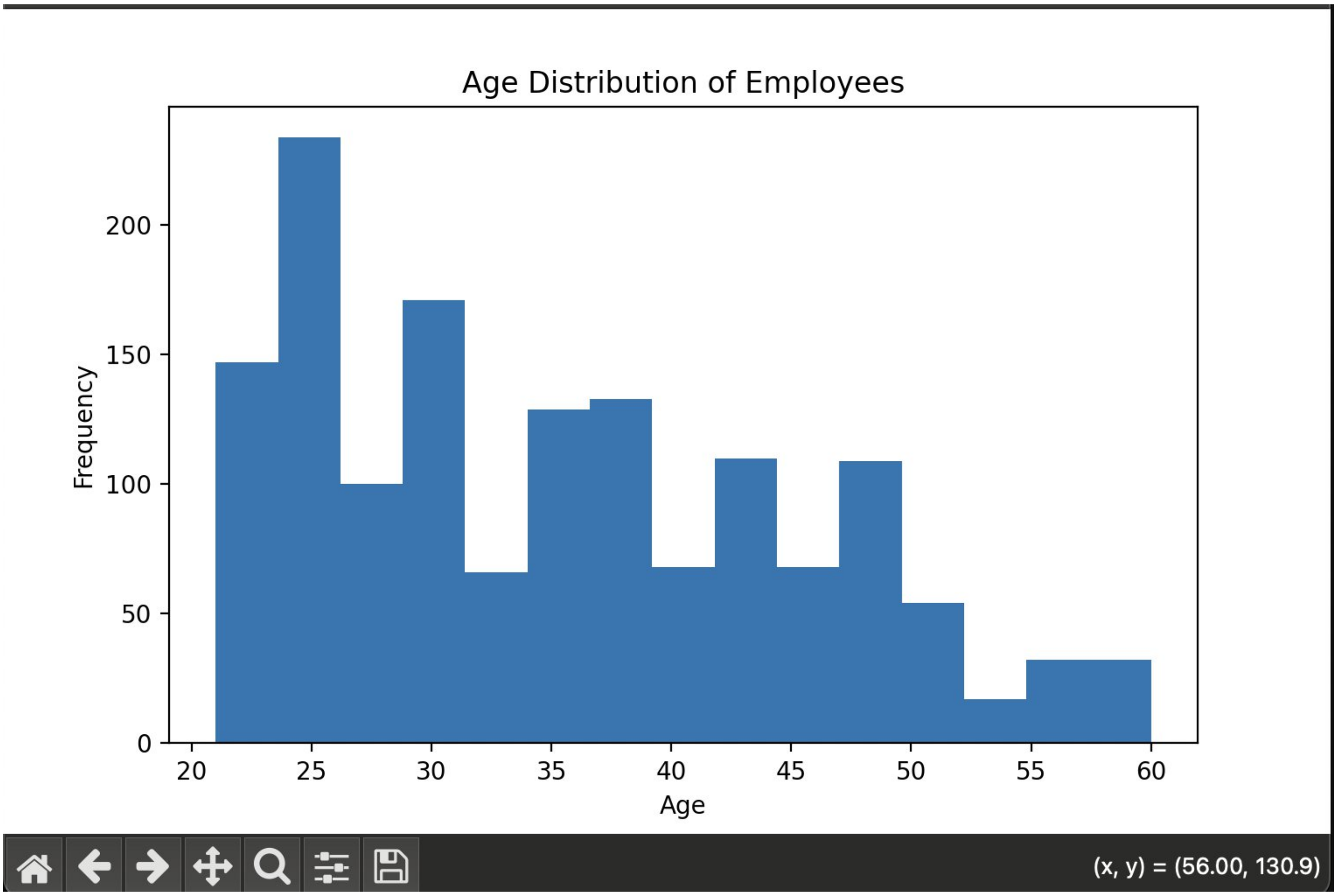
Attrition Rate

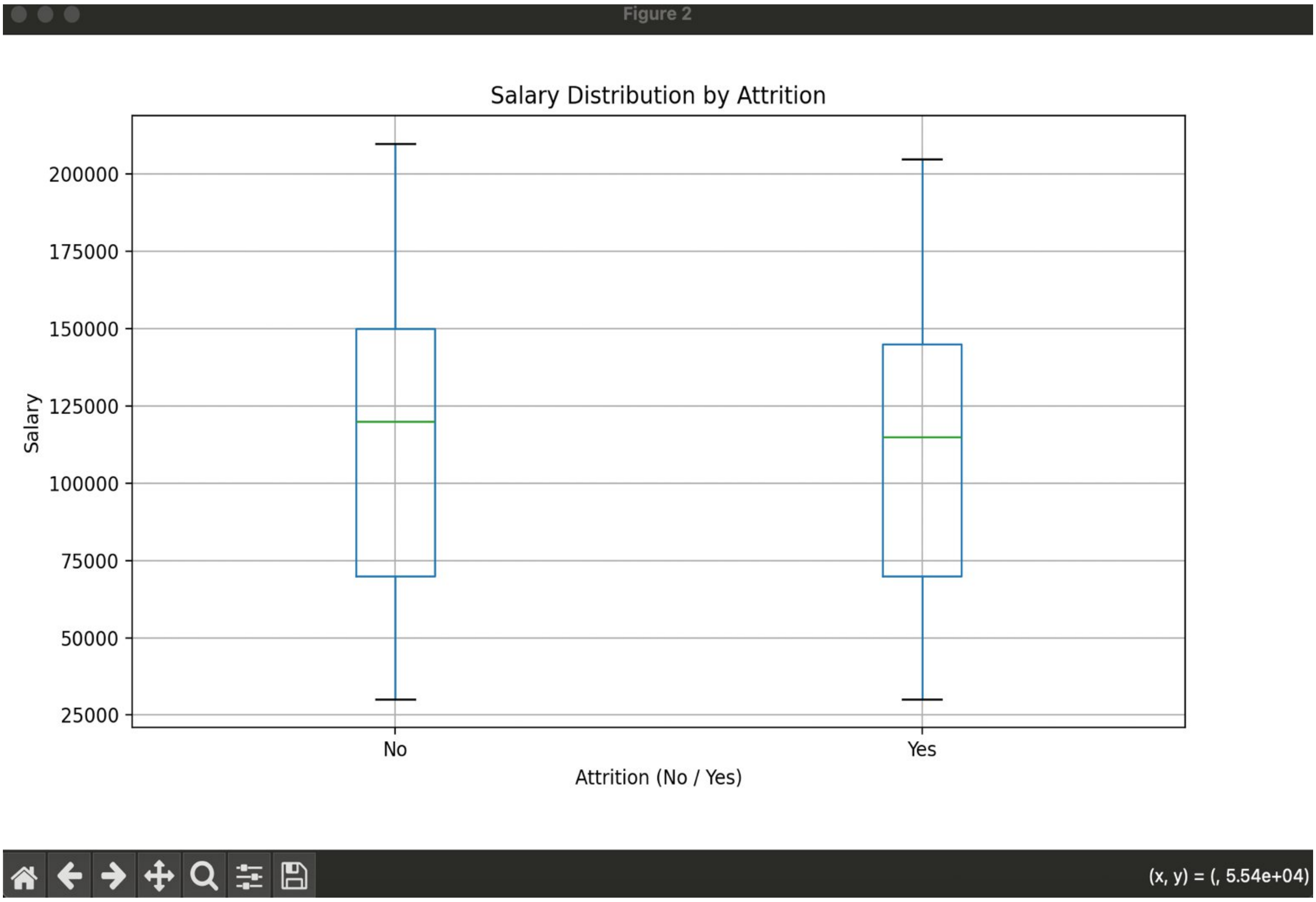
16.12%

🔥 Attrition Count



(x, y) = (, 179.3)





End of Project .