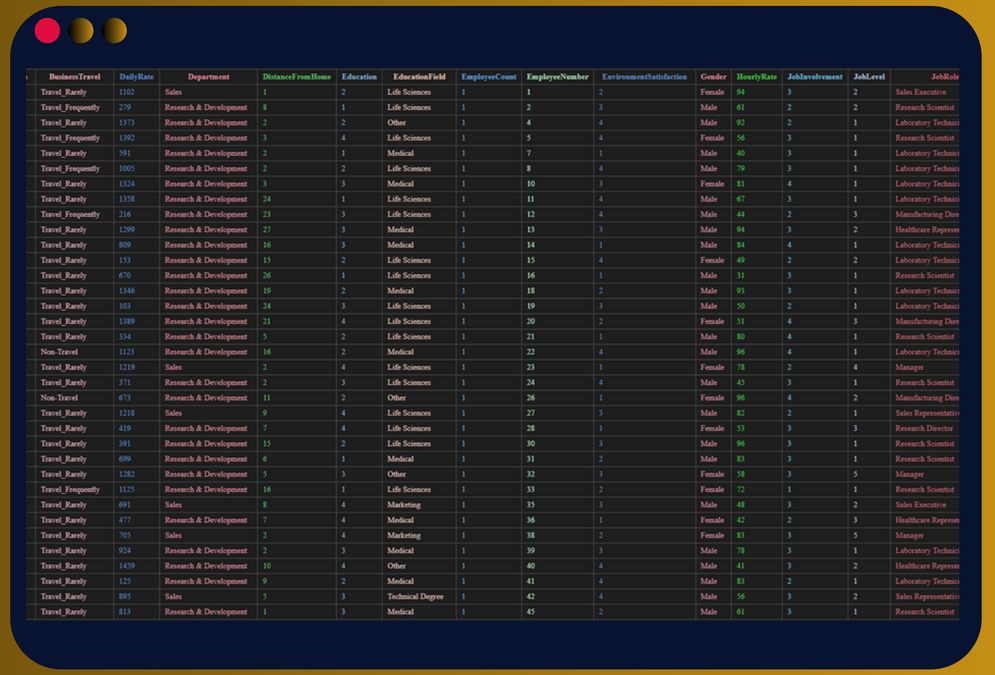
PROJECT TITLE

Understanding Employee Attrition

Dataset: IBM HR Analytics Employee Attrition & Performance

- *Rows: 1,470 employees*
- *Columns: 35 HR-related features*

Languages Used:
Python (Pandas) • SQL (Sqlite)



BusinessTravel	DailyRate	Department	DistanceFromHome	Education	EducationField	EmployeeCount	EmployeeNumber	EnvironmentSatisfaction	Gender	HourlyRate	JobInvolvement	JobLevel	JobRole
Travel_Rarely	1102	Sales	1	2	Life Sciences	1	1	2	Female	94	3	2	Sales Executive
Travel_Frequently	279	Research & Development	8	1	Life Sciences	1	2	3	Male	61	2	2	Research Scientist
Travel_Rarely	1373	Research & Development	2	2	Other	1	4	4	Male	92	2	1	Laboratory Technici
Travel_Frequently	1392	Research & Development	3	4	Life Sciences	1	5	4	Female	56	3	1	Research Scientist
Travel_Rarely	591	Research & Development	2	1	Medical	1	7	1	Male	40	3	1	Laboratory Technici
Travel_Frequently	1005	Research & Development	2	2	Life Sciences	1	8	4	Male	79	3	1	Laboratory Technici
Travel_Rarely	1324	Research & Development	3	3	Medical	1	10	3	Female	81	4	1	Laboratory Technici
Travel_Rarely	1358	Research & Development	24	1	Life Sciences	1	11	4	Male	67	3	1	Laboratory Technici
Travel_Frequently	216	Research & Development	23	3	Life Sciences	1	12	4	Male	44	2	3	Manufacturing Dire
Travel_Rarely	1299	Research & Development	27	3	Medical	1	13	3	Male	94	3	2	Healthcare Represe
Travel_Rarely	809	Research & Development	16	3	Medical	1	14	1	Male	84	4	1	Laboratory Technici
Travel_Rarely	153	Research & Development	15	2	Life Sciences	1	15	4	Female	49	2	2	Laboratory Technici
Travel_Rarely	670	Research & Development	26	1	Life Sciences	1	16	1	Male	31	3	1	Research Scientist
Travel_Rarely	1346	Research & Development	19	2	Medical	1	18	2	Male	93	3	1	Laboratory Technici
Travel_Rarely	103	Research & Development	24	3	Life Sciences	1	19	3	Male	50	2	1	Laboratory Technici
Travel_Rarely	1389	Research & Development	21	4	Life Sciences	1	20	2	Female	51	4	3	Manufacturing Dire
Travel_Rarely	334	Research & Development	5	2	Life Sciences	1	21	1	Male	80	4	1	Research Scientist
Non-Travel	1123	Research & Development	16	2	Medical	1	22	4	Male	96	4	1	Laboratory Technici
Travel_Rarely	1219	Sales	2	4	Life Sciences	1	23	1	Female	78	2	4	Manager
Travel_Rarely	371	Research & Development	2	3	Life Sciences	1	24	4	Male	45	3	1	Research Scientist
Non-Travel	673	Research & Development	11	2	Other	1	26	1	Female	96	4	2	Manufacturing Dire
Travel_Rarely	1218	Sales	9	4	Life Sciences	1	27	3	Male	82	2	1	Sales Representativ
Travel_Rarely	419	Research & Development	7	4	Life Sciences	1	28	1	Female	53	3	3	Research Director
Travel_Rarely	391	Research & Development	15	2	Life Sciences	1	30	3	Male	96	3	1	Research Scientist
Travel_Rarely	699	Research & Development	6	1	Medical	1	31	2	Male	83	3	1	Research Scientist
Travel_Rarely	1282	Research & Development	5	3	Other	1	32	3	Female	58	3	5	Manager
Travel_Frequently	1125	Research & Development	16	1	Life Sciences	1	33	2	Female	72	1	1	Research Scientist
Travel_Rarely	691	Sales	8	4	Marketing	1	35	3	Male	48	3	2	Sales Executive
Travel_Rarely	477	Research & Development	7	4	Medical	1	36	1	Female	42	2	3	Healthcare Represe
Travel_Rarely	705	Sales	2	4	Marketing	1	38	2	Female	83	3	5	Manager
Travel_Rarely	924	Research & Development	2	3	Medical	1	39	3	Male	78	3	1	Laboratory Technici
Travel_Rarely	1459	Research & Development	10	4	Other	1	40	4	Male	41	3	2	Healthcare Represe
Travel_Rarely	125	Research & Development	9	2	Medical	1	41	4	Male	83	2	1	Laboratory Technici
Travel_Rarely	895	Sales	5	3	Technical Degree	1	42	4	Male	56	3	2	Sales Representativ
Travel_Rarely	813	Research & Development	1	3	Medical	1	45	2	Male	61	3	1	Research Scientist

Presented By

SIMBIAT MUSA

MAY 2025

OVERVIEW

Objective:

To explore the factors that contribute to employee attrition and help HR teams develop better retention strategies.

Business Questions Analysed:

1. Is there a difference in attrition between departments?
2. Do employees who work overtime leave more often?
3. How does job satisfaction relate to attrition?
4. Is there a pattern between years at company and attrition?
5. Do younger employees leave more often than older ones?
6. Is there a relationship between income and attrition?
7. Which job roles have the highest attrition rates?
8. Do people with more job involvement tend to stay?
9. How does work-life balance affect attrition?
10. Are employees who have worked at more companies more likely to leave?

Filtered Datasets

```
attrition_filtered = attrition_df[['Attrition', 'Age', 'Gender', 'MaritalStatus', 'JobRole', 'Department', 'D
                                   'BusinessTravel', 'OverTime', 'YearsAtCompany', 'YearsInCurrentRole',
                                   'JobSatisfaction', 'WorkLifeBalance', 'JobLevel', 'MonthlyIncome',
                                   'PercentSalaryHike', 'StockOptionLevel', 'PerformanceRating',
                                   'TotalWorkingYears', 'NumCompaniesWorked', 'YearsSinceLastPromotion',]]
# I will be using the following columns for my analysis
```

✓ 0.0s

```
attrition_filtered.head(10)
```

✓ 0.0s

	Attrition	Age	Gender	MaritalStatus	JobRole	Department	DistanceFromHome	BusinessTravel	OverTime
0	Yes	41	Female	Single	Sales Executive	Sales	1	Travel_Rarely	Yes
1	No	49	Male	Married	Research Scientist	Research & Development	8	Travel_Frequently	No
2	Yes	37	Male	Single	Laboratory Technician	Research & Development	2	Travel_Rarely	Yes
3	No	33	Female	Married	Research Scientist	Research & Development	3	Travel_Frequently	Yes
4	No	27	Male	Married	Laboratory Technician	Research & Development	2	Travel_Rarely	No

```
attrition_filtered.tail(5)
```

✓ 0.0s

	Attrition	Age	Gender	MaritalStatus	JobRole	Department	DistanceFromHome	BusinessTravel	OverTim
1465	No	36	Male	Married	Laboratory Technician	Research & Development	23	Travel_Frequently	N
1466	No	39	Male	Married	Healthcare Representative	Research & Development	6	Travel_Rarely	N
1467	No	27	Male	Married	Manufacturing Director	Research & Development	4	Travel_Rarely	Y
1468	No	49	Male	Married	Sales Executive	Sales	2	Travel_Frequently	N
1469	No	34	Male	Married	Laboratory Technician	Research & Development	8	Travel_Rarely	N

5 rows × 21 columns

```
attrition_filtered.shape
```

✓ 0.0s

```
(1470, 21)
```

GROUPED BY THEME

1. Target Variable (Attrition → Whether the employee left the company (Yes/No))
2. Demographics (Age, Gender, MaritalStatus)
3. Job-Related Info (JobRole, Department, BusinessTravel, DistanceFromHome, OverTime, YearsAtCompany, YearsInCurrentRole, JobSatisfaction (1 to 4), WorkLifeBalance (1 to 4), JobLevel, YearsSinceLastPromotion)

4. Compensation (MonthlyIncome, PercentSalaryHike, StockOptionLevel)
5. Performance & Tenure (PerformanceRating, TotalWorkingYears, NumCompaniesWorked)

```
attrition_filtered.describe()
# To check the ouliers and understand the spread of numeric data
```

✓ 0.0s

	Age	DistanceFromHome	YearsAtCompany	YearsInCurrentRole	JobSatisfaction	WorkLifeBalance	JobLevel	MonthlyIncome
count	1470.000000	1470.000000	1470.000000	1470.000000	1470.000000	1470.000000	1470.000000	1470.000000
mean	36.923810	9.192517	7.008163	4.229252	2.728571	2.761224	2.063946	6502.931293
std	9.135373	8.106864	6.126525	3.623137	1.102846	0.706476	1.106940	4707.956783
min	18.000000	1.000000	0.000000	0.000000	1.000000	1.000000	1.000000	1009.000000
25%	30.000000	2.000000	3.000000	2.000000	2.000000	2.000000	1.000000	2911.000000
50%	36.000000	7.000000	5.000000	3.000000	3.000000	3.000000	2.000000	4919.000000
75%	43.000000	14.000000	9.000000	7.000000	4.000000	3.000000	3.000000	8379.000000
max	60.000000	29.000000	40.000000	18.000000	4.000000	4.000000	5.000000	19999.000000

```
attrition_filtered.info()
# To identify which columns are categorical vs numerical, and which might need cleaning
```

✓ 0.0s

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1470 entries, 0 to 1469
Data columns (total 21 columns):
 #   Column                   Non-Null Count  Dtype
---  ------                   --------------  -----
 0   Attrition                1470 non-null   object
 1   Age                      1470 non-null   int64
 2   Gender                   1470 non-null   object
 3   MaritalStatus            1470 non-null   object
 4   JobRole                  1470 non-null   object
```

```
attrition_filtered.isnull().sum()
# There are no null values in the dataset
# I will now check for duplicates
```

✓ 0.0s

```
Attrition                 0
Age                       0
Gender                    0
MaritalStatus             0
JobRole                   0
Department                0
DistanceFromHome          0
BusinessTravel            0
OverTime                  0
YearsAtCompany            0
YearsInCurrentRole        0
JobSatisfaction           0
WorkLifeBalance           0
JobLevel                  0
MonthlyIncome             0
PercentSalaryHike         0
StockOptionLevel          0
PerformanceRating         0
TotalWorkingYears         0
NumCompaniesWorked        0
YearsSinceLastPromotion   0
dtype: int64
```

```
attrition_filtered.duplicated().sum()
# There are no duplicates in the dataset
# I will now check the data types of the columns
```

✓ 0.0s

```
0
```

Data Cleaning

EDA - UNIVARIATE/BIVARIATE

```python
from scipy.stats import chi2_contingency

# Contingency table for Attrition vs OverTime
table = pd.crosstab(attrition_filtered['Attrition'], attrition_filtered['OverTime'])

# Chi-square test
chi2, p, dof, expected = chi2_contingency(table)

print("Chi-Square:", chi2)
print("P-Value:", p)
```

✓ 0.0s

```
Chi-Square: 87.56429365828768
P-Value: 8.15842372153832e-21
```

```python
from scipy.stats import chi2_contingency

# Contingency table for Attrition vs Department
table2 = pd.crosstab(attrition_filtered['Attrition'], attrition_filtered['Department'

# Chi-square test
chi2, p, dof, expected = chi2_contingency(table2)

print("Chi-Square:", chi2)
print("P-Value:", p)
```

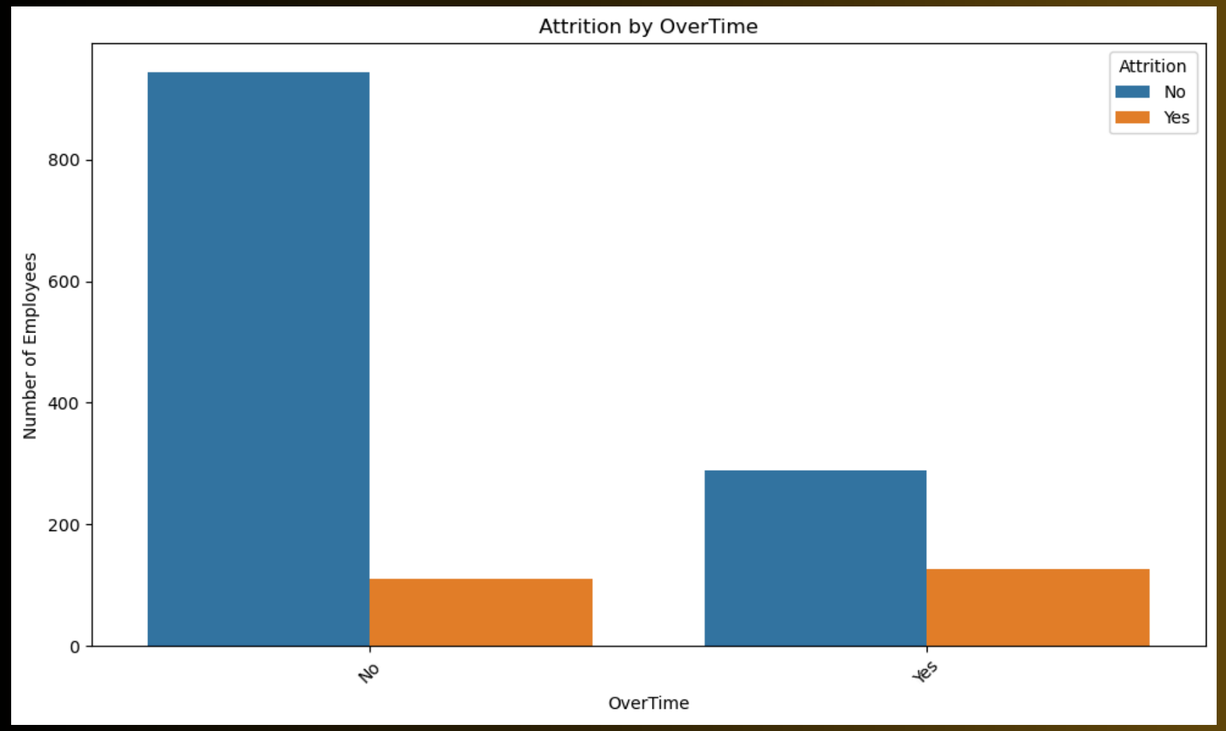
✓ 0.0s

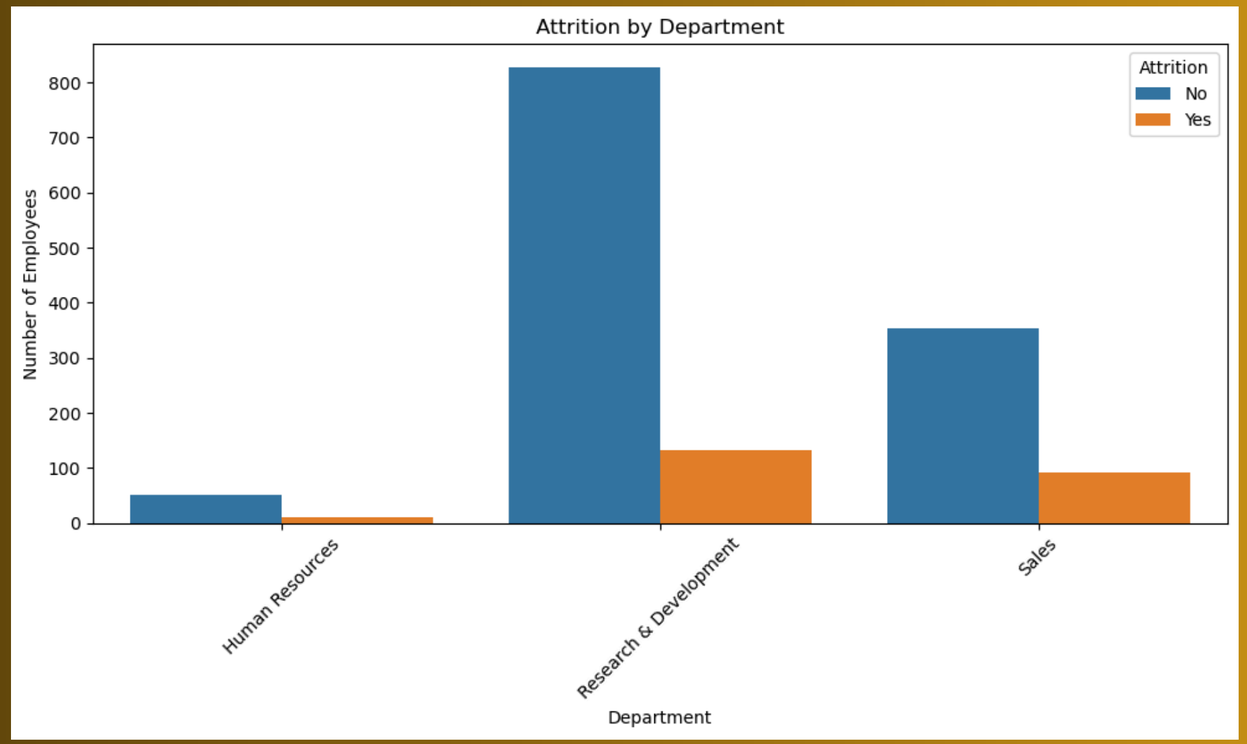
```
Chi-Square: 10.79600732241067
P-Value: 0.004525606574479633
```

```python
cat = attrition_filtered.select_dtypes(exclude="number")
num = attrition_filtered.select_dtypes(include="number")
print(cat.columns)
print(num.columns)
```

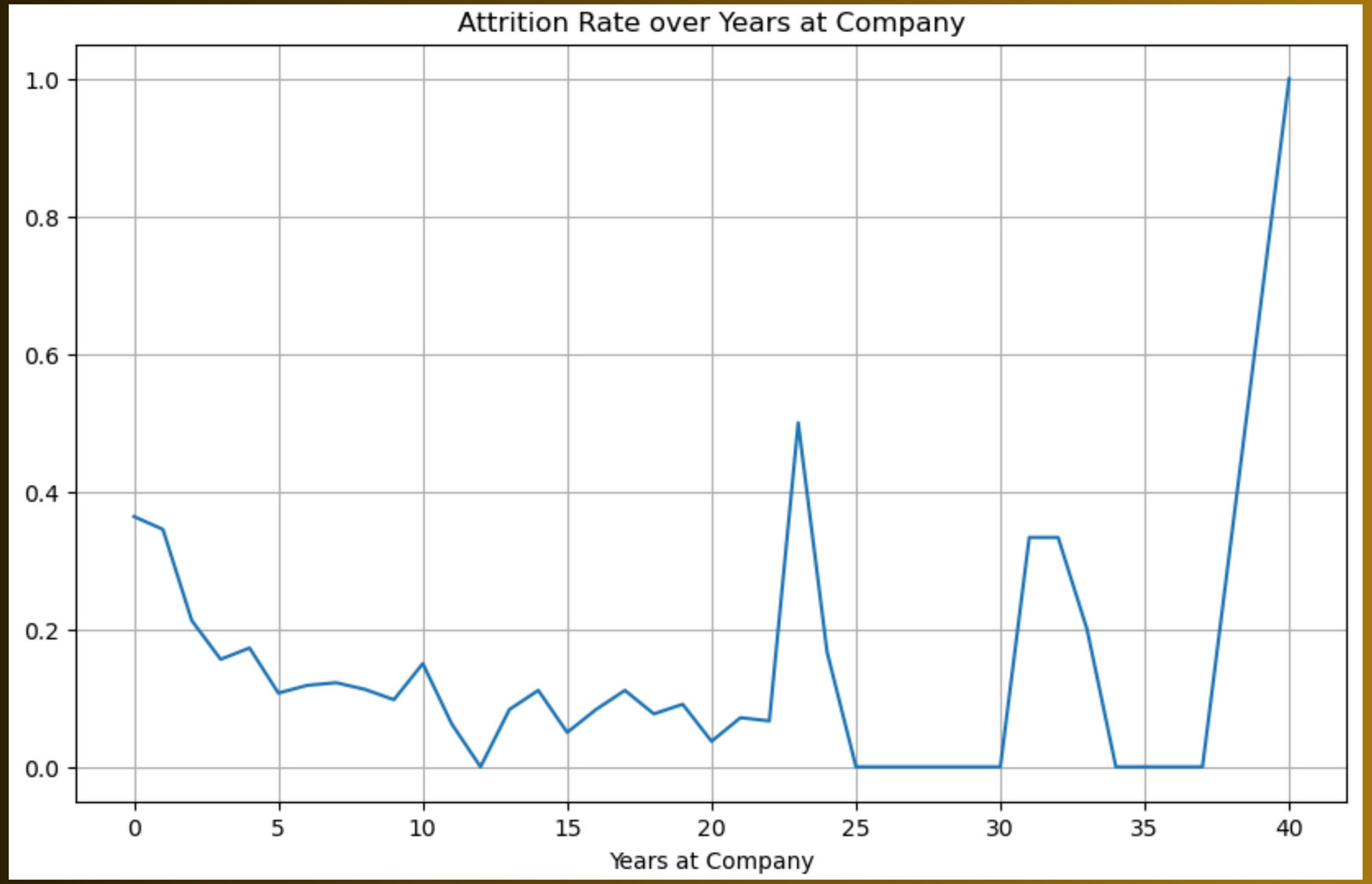
✓ 0.0s

```
Index(['Attrition', 'Gender', 'MaritalStatus', 'JobRole', 'Department',
       'BusinessTravel', 'OverTime'],
      dtype='object')
Index(['Age', 'DistanceFromHome', 'YearsAtCompany', 'YearsInCurrentRole'
       'JobSatisfaction', 'WorkLifeBalance', 'JobLevel', 'MonthlyIncome'
       'PercentSalaryHike', 'StockOptionLevel', 'PerformanceRating',
       'TotalWorkingYears', 'NumCompaniesWorked', 'YearsSinceLastPromoti
      dtype='object')
```

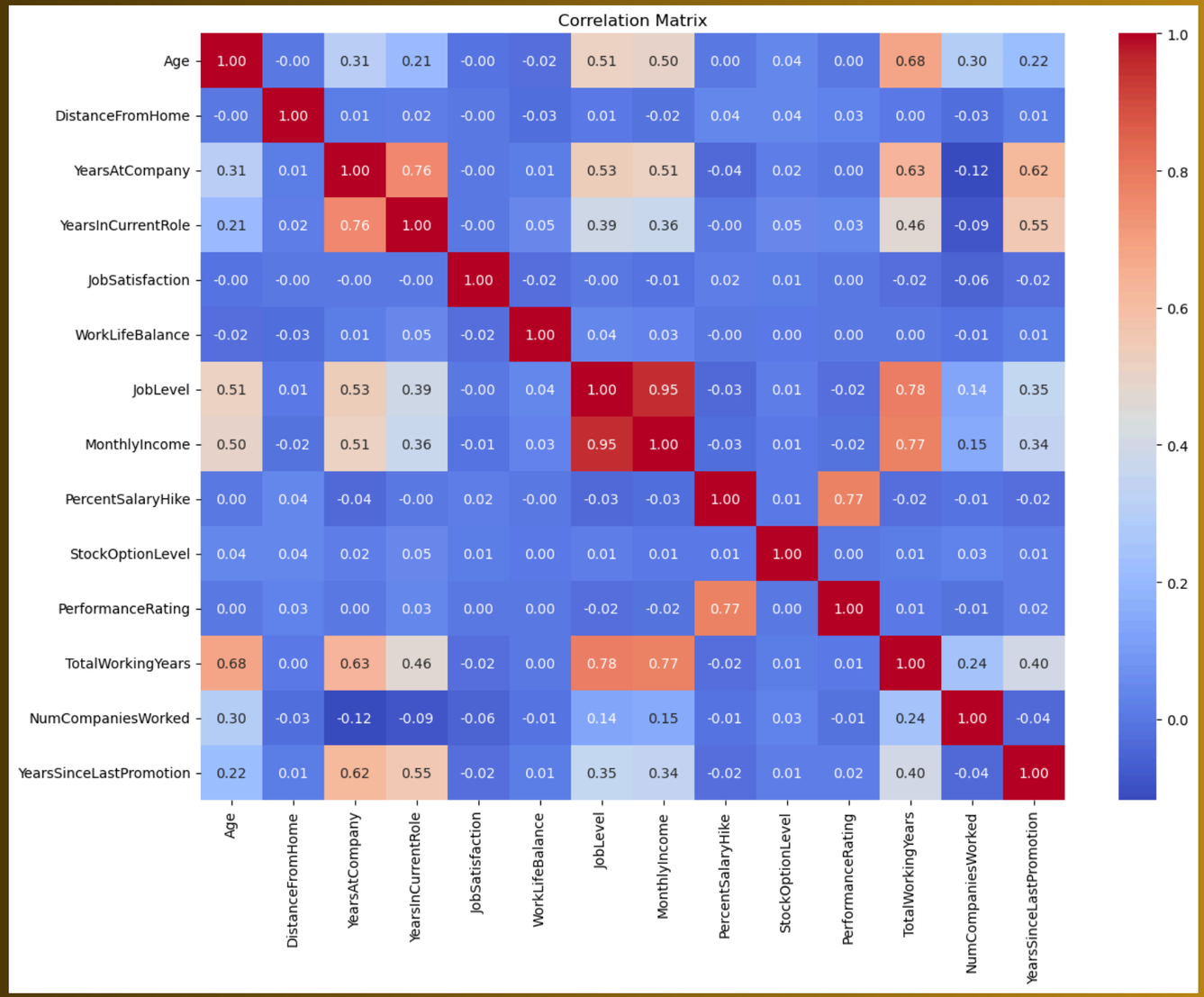


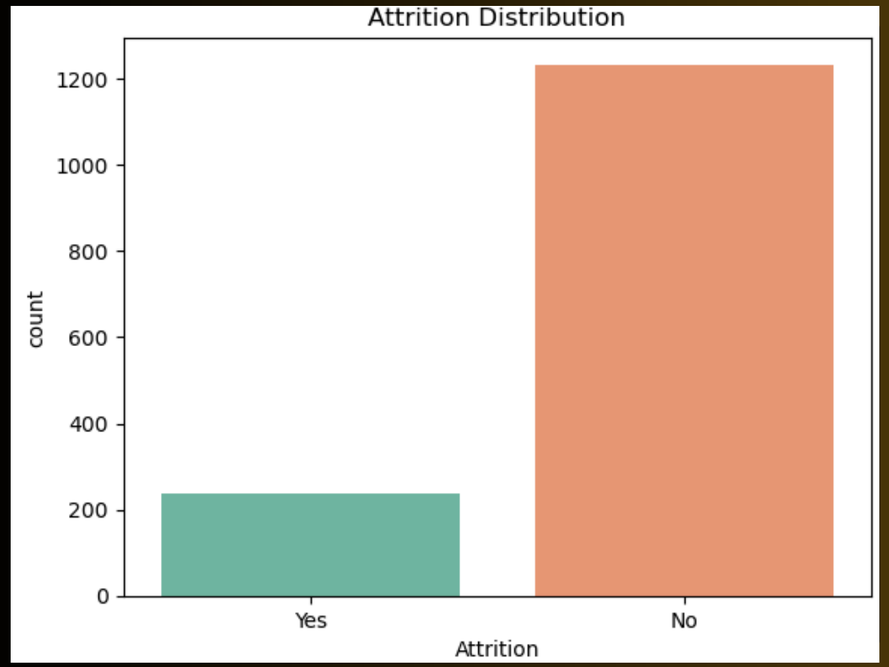


The Evolution of Attrition Rate

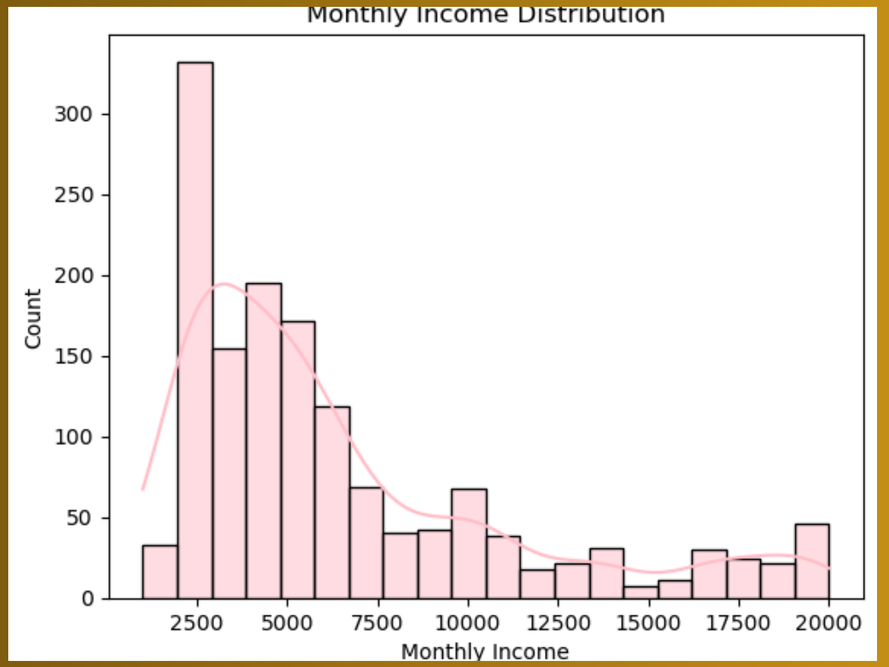


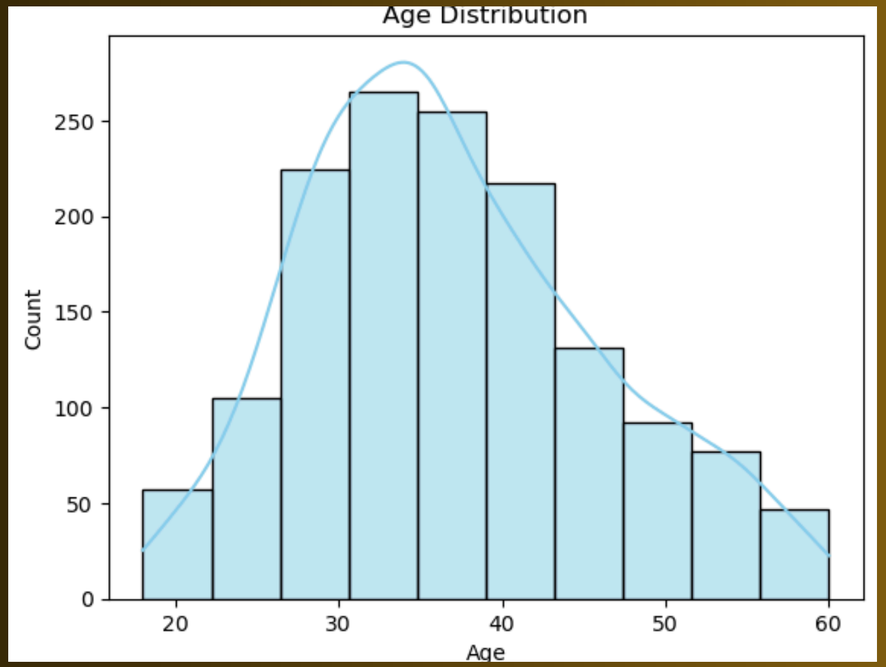
Correlation Matrix





The Countplot of Attrition which is the Target Variable shows that around 16% left the company, and 84% stayed.





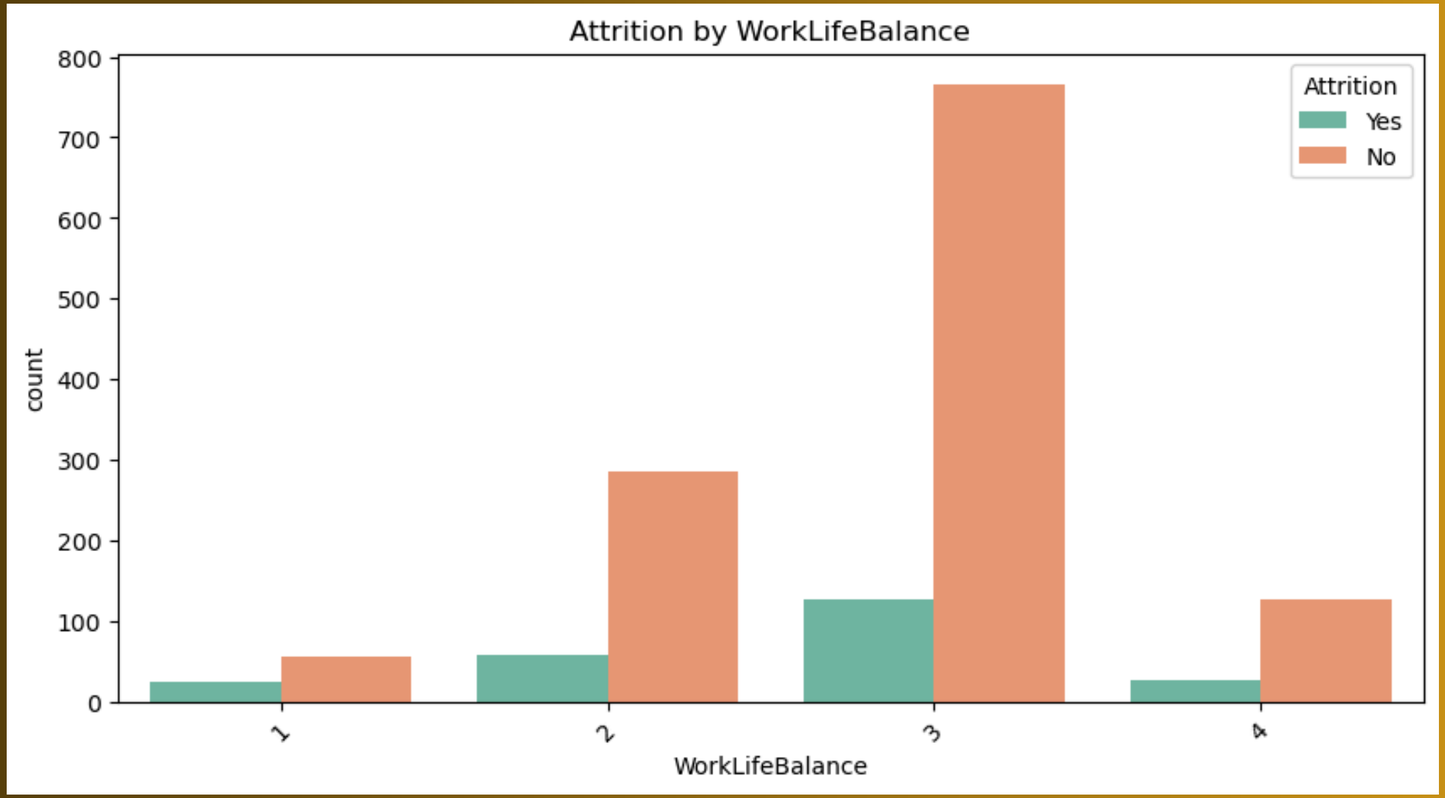
The histogram distribution of age above shows most employees are in their 30s to early 40s.

The histogram of monthly income distribution shows that income is right_skewed. A few employees earn much more.

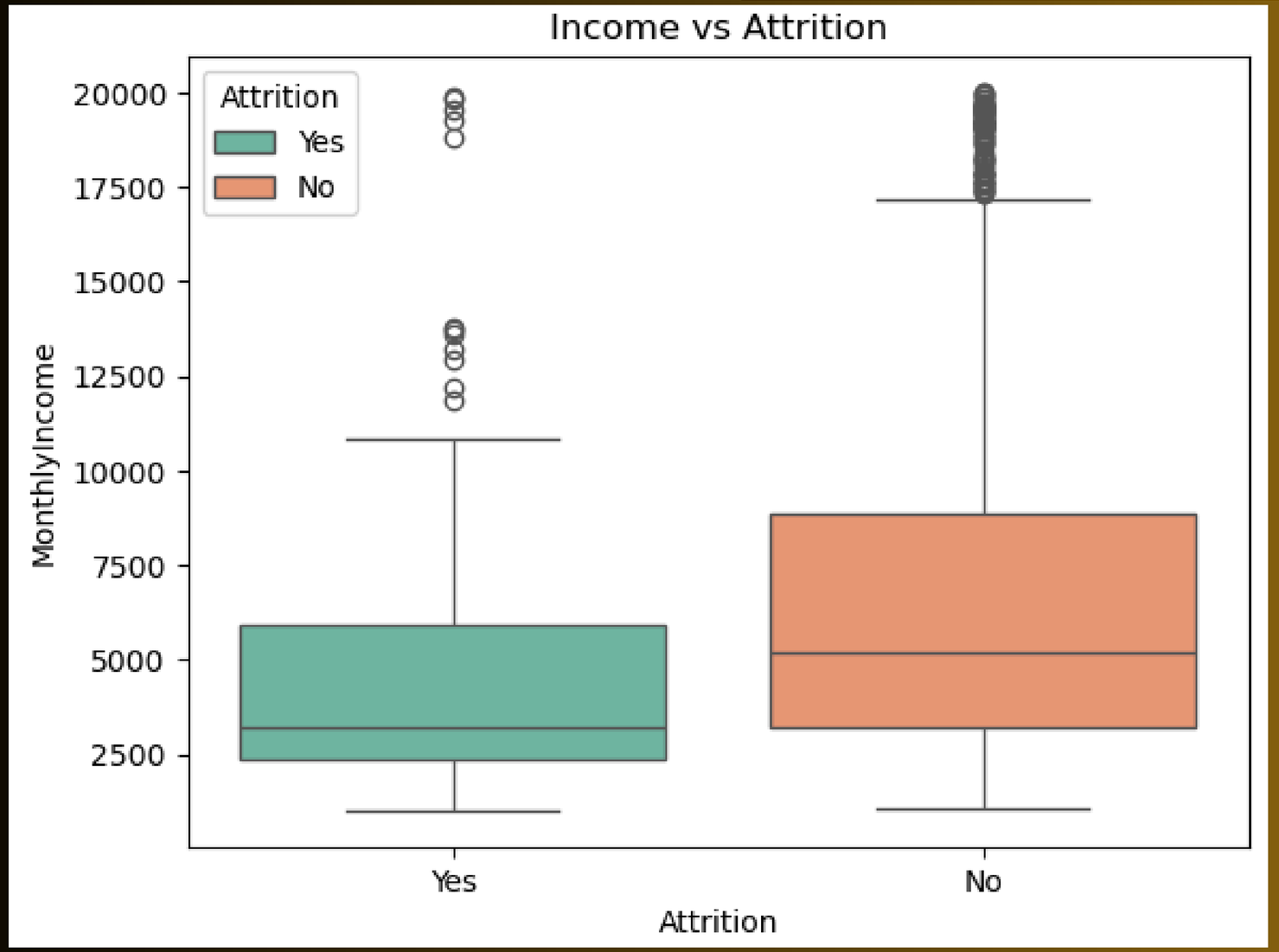
Some of the measured relationship with Attrition (PYTHON vs SQL)

```sql
-- 9. How does work-life balance affect attrition?
-- This shows the attrition count by work-life balance ratings.
SELECT WorkLifeBalance, Attrition, COUNT(*) AS BalanceCount
FROM EmployeeAttrition
GROUP BY WorkLifeBalance, Attrition
ORDER BY WorkLifeBalance;
```

	WorkLifeBalance	Attrition	BalanceCount
1	1	No	55
2	1	Yes	25
3	2	No	286
4	2	Yes	58
5	3	No	766
6	3	Yes	127
7	4	No	126
8	4	Yes	27

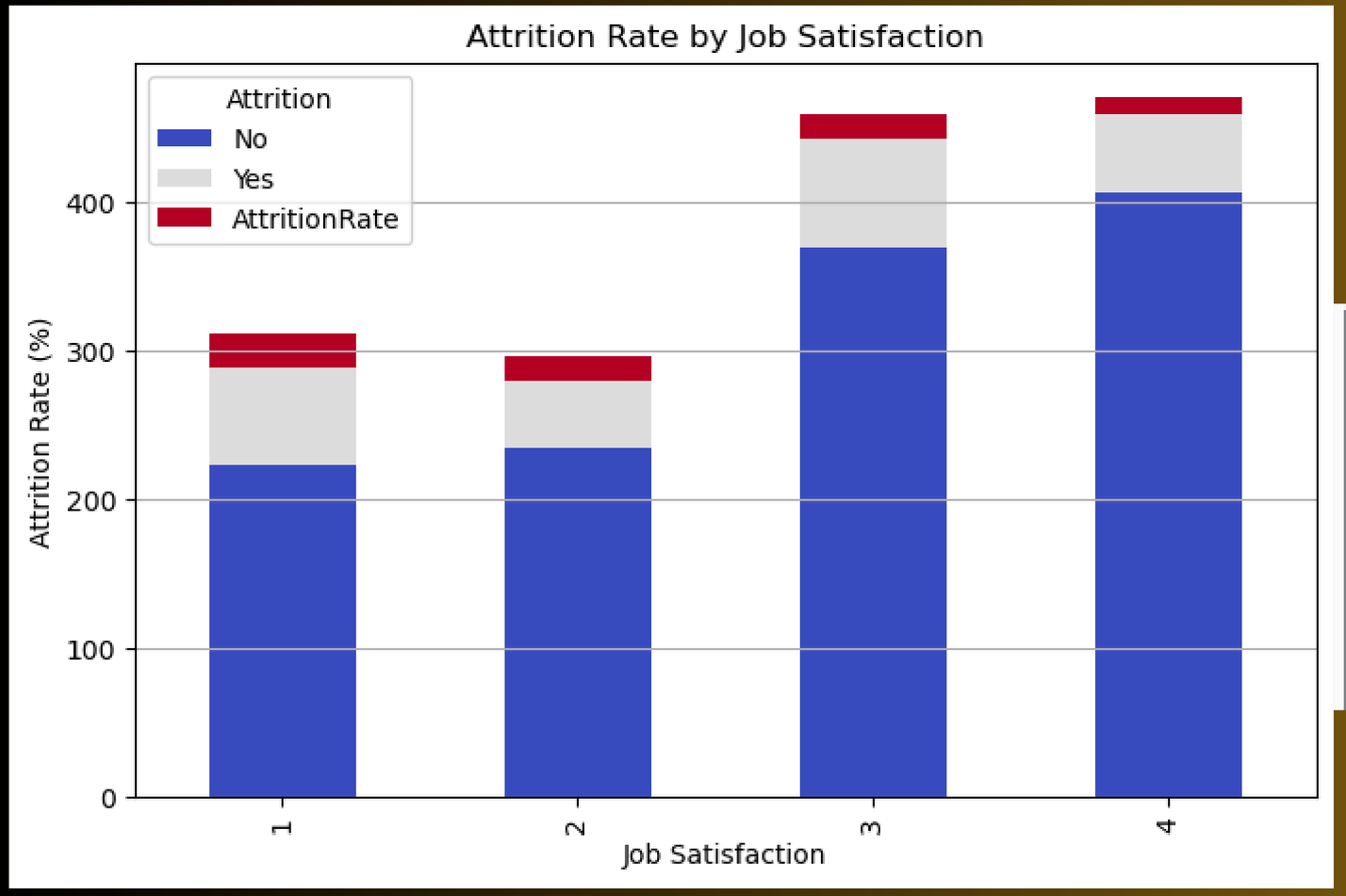


```sql
-- 6. Is there a relationship between income and attrition?
-- This query is to check if monthly income correlates with employee attrition.
SELECT MonthlyIncome, Attrition, COUNT (*) AS IncomeCount
FROM EmployeeAttrition
GROUP BY MonthlyIncome
ORDER BY MonthlyIncome ASC;
```



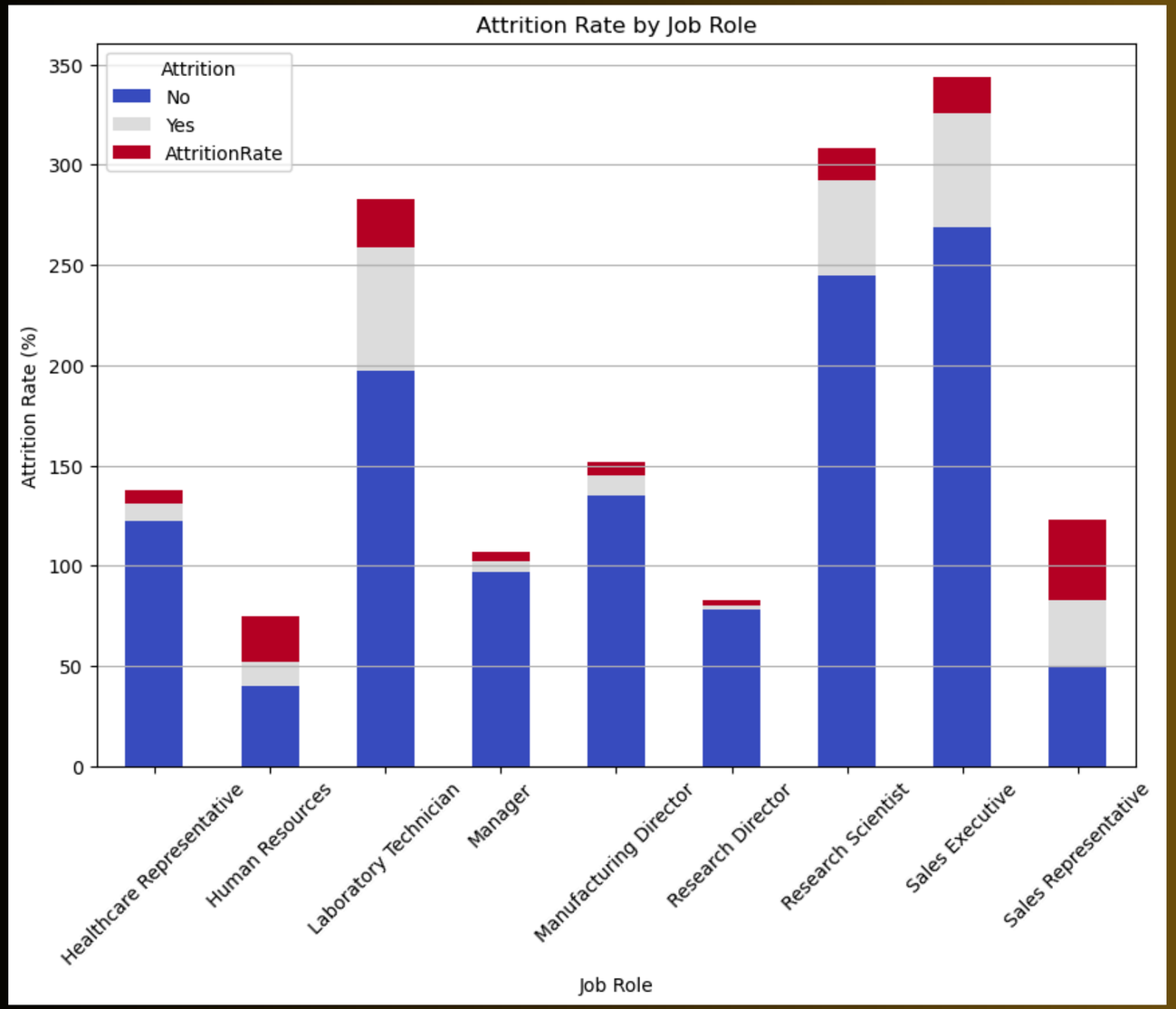
	MonthlyIncome	Attrition	IncomeCount
1	1009	Yes	1
2	1051	No	1
3	1052	No	1
4	1081	Yes	1
5	1091	Yes	1
6	1102	Yes	1
7	1118	Yes	1
8	1129	No	1
9	1200	No	1
10	1223	No	1
11	1232	No	1
12	1261	Yes	1
13	1274	No	1
14	1281	No	1
15	1359	Yes	1
16	1393	Yes	1
17	1416	Yes	1
18	1420	Yes	1
19	1483	No	1
20	1514	No	1
21	1555	Yes	1
22	1563	No	1
23	1569	Yes	1
24	1601	Yes	1
25	1611	No	1
26	1675	Yes	1

```sql
-- 3. How does job satisfaction relate to attrition?
-- This query is to group employees by their job satisfaction level and attrition status.
SELECT JobSatisfaction, COUNT(*) AS Total, SUM(CASE WHEN Attrition = 'Yes' THEN 1 ELSE 0 END) AS AttritionCount, ROUND(100.0*SUM(CASE WHEN
FROM EmployeeAttrition
GROUP BY JobSatisfaction
ORDER BY AttritionRate ASC;
```



	JobSatisfaction	Total	AttritionCount	AttritionRate
1	4	459	52	11.33
2	2	280	46	16.43
3	3	442	73	16.52
4	1	289	66	22.84

```sql
-- 7. Which job roles have the highest attrition rates?
-- This is to calculate the attrition rate per job role.
SELECT JobRole, COUNT(*) AS Total, SUM(CASE WHEN Attrition = 'Yes' THEN 1 ELSE 0 END) AS AttritionCount, ROUND(100.0*SUM(CASE WHEN Attrition = 'Yes' THEN 1 ELSE 0 END) / COUNT(*),2) AS Attri
FROM EmployeeAttrition
GROUP BY JobRole
ORDER BY AttritionRate DESC;
```



	JobRole	Total	AttritionCount	AttritionRate
1	Sales Representative	83	33	39.76
2	Laboratory Technician	259	62	23.94
3	Human Resources	52	12	23.08
4	Sales Executive	326	57	17.48
5	Research Scientist	292	47	16.1
6	Manufacturing Director	145	10	6.9
7	Healthcare Representative	131	9	6.87
8	Manager	102	5	4.9
9	Research Director	80	2	2.5

KEY FINDINGS

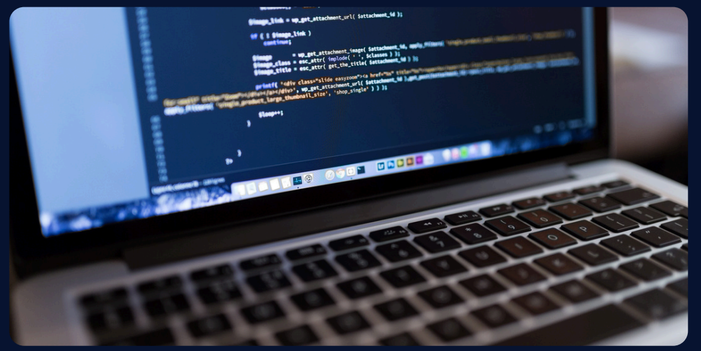
The analysis revealed that employees who work overtime are significantly more likely to leave the company, as confirmed by both chi-square testing and visual trends. Attrition is higher in certain departments, especially in Sales, indicating department-specific issues. Job satisfaction numbers proved inconclusive as seemingly satisfied employees still left pointing to possibility of job satisfaction surveys might not capture honest opinions. Attrition is also more common among younger employees, those with fewer years at the company, and those who have worked at multiple companies before. Work-life balance, job involvement, and income levels also show measurable impact on retention.

RECOMMENDATIONS

To reduce employee attrition, the organization should prioritize reducing overtime, as it strongly correlates with higher resignation rates. Improving job satisfaction through career growth opportunities, recognition, and engagement can help retain talent. Department-specific retention strategies, especially in Sales, should be implemented to address unique challenges. Younger employees and those with fewer years at the company need stronger onboarding and mentorship to increase loyalty. Additionally, promoting a healthy work-life balance and using data-driven attrition risk indicators will allow HR to proactively support at-risk employees and improve overall retention.

LEARNINGS

- I learned how to extract meaningful insights from the employee data by applying both SQL and Python-based analysis.
- I deepened my understanding on Python and SQL.
- I practiced applying Chi-Square tests to determine statistically significant relationships and used visualizations (line plots, bar charts, heatmaps) to clearly communicate findings.
- This end-to-end project reinforced the full data analysis lifecycle—from dataset cleaning to hypothesis testing and drawing actionable conclusions.



Thank You