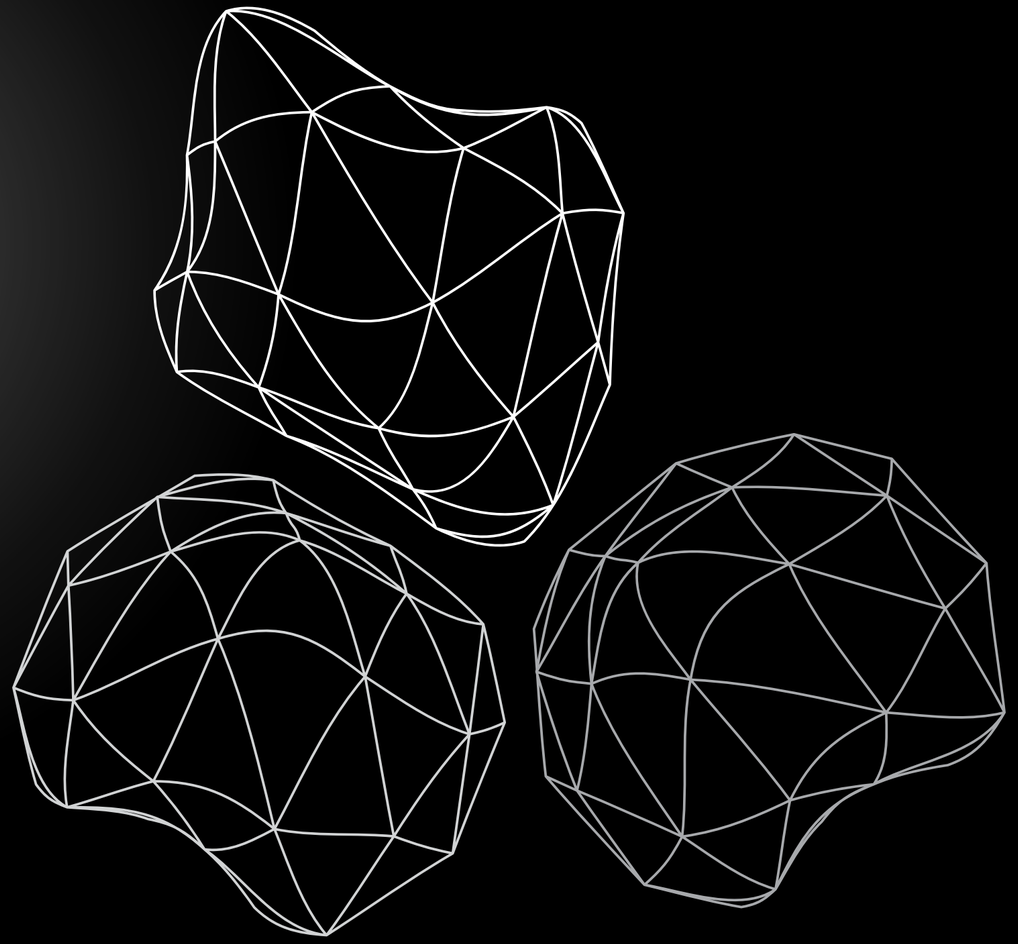
DIGITAL EGYPT PIONEERS INITIATIVE

Employee Retention & Performance Analysis (2022)

PRESENTATION



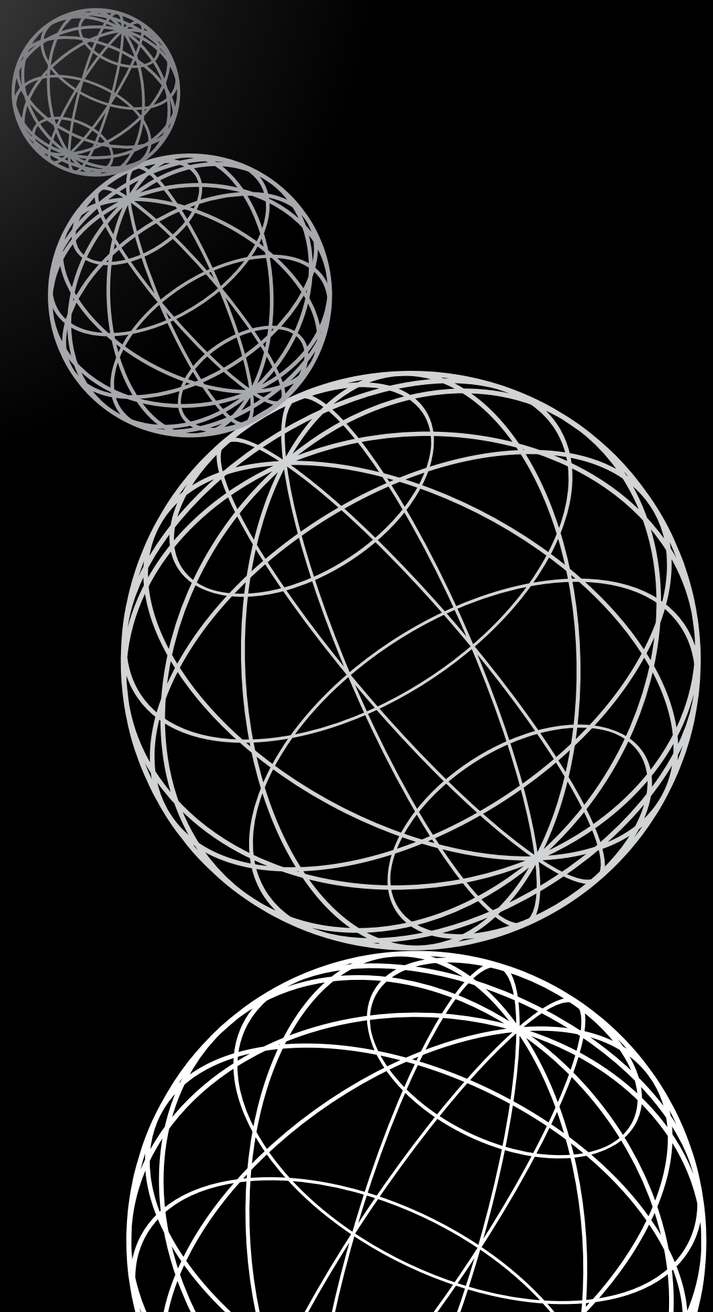
TEAM MEMBERS

YOUSSEF AHMED

SALAH MAHMOUD

MOHAMED KHALED

TABLE OF CONTENT



ABOUT HR DATASET

The analysis is based on two key datasets:

- EmployeeDim: Includes EmployeeID, HireDate, YearsAtCompany, Attrition, Gender, MaritalStatus, Age, YearsWithCurrentManager.
- PerformanceRatingFact: Includes JobSatisfactionID, PerformanceRatingID, WorkLifeBalanceID, EnvironmentSatisfactionID, RelationshipSatisfactionID, SelfRatingID, ReviewDate.

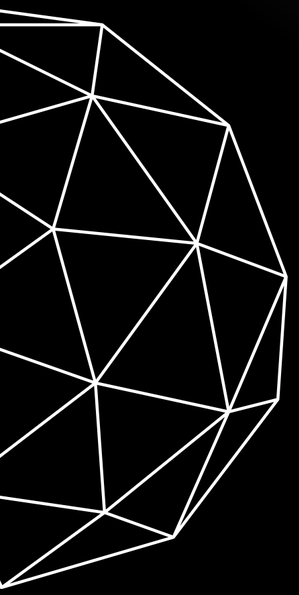
Data scope: Employee records and performance reviews up to 2022.

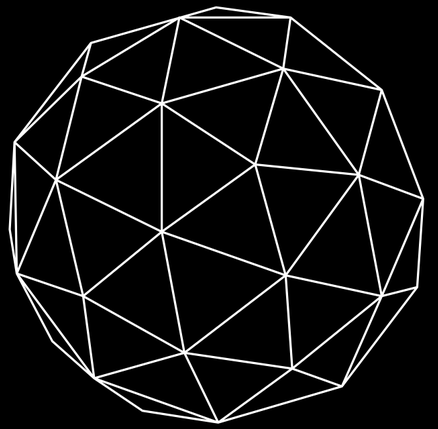
DATA MODELING & CLEANING

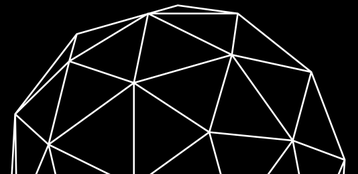
- Importing libraries
- Loading data
- Understanding data structure
- Handling missing data
- Handling Dublicates
- Data Type conversion
- Outlier Detection and Removal
- Relationship Establishment
- Data Source Integration

Loading Data & import libraries

```python
import pandas as pd

employee_df = pd.read_csv(r"C:\Users\dell\Downloads\Employee.csv")
performance_rating_df = pd.read_csv(r"C:\Users\dell\Downloads\PerformanceRating.csv")

print("Employee Data:")
print(employee_df.head())
print("\nPerformance Rating Data:")
print(performance_rating_df.head())
```







Data tybe conversion & outliers detection

أنواع البيانات بعد التحويل:

EmployeeID	object
FirstName	object
LastName	object
Gender	object
Age	int64
BusinessTravel	object
Department	object
DistanceFromHome (KM)	int64
State	object
Ethnicity	object
Education	int64
EducationField	object
JobRole	object
MaritalStatus	object
Salary	int64
StockOptionLevel	int64
OverTime	object
HireDate	datetime64[ns]
Attrition	object
YearsAtCompany	int64
YearsInMostRecentRole	int64
YearsSinceLastPromotion	int64
YearsWithCurrManager	int64

dtype: object

إحصائيات Age:

count	1470.000000
mean	28.989796
std	7.993055
min	18.000000
25%	23.000000
50%	26.000000
75%	34.000000
max	51.000000

Name: Age, dtype: float64

عدد القيم الشاذة في Age: 0

إحصائيات Salary:

count	1470.000000
mean	112956.497959
std	103342.889222
min	20387.000000
25%	43580.500000
50%	71199.500000
75%	142055.750000
max	547204.000000

Name: Salary, dtype: float64

عدد القيم الشاذة في Salary: 0

Handling missing data & dublicates

```
عدد القيم المفقوده بعد المعالجة:
EmployeeID                  0
FirstName                   0
LastName                    0
Gender                      0
Age                         0
BusinessTravel              0
Department                  0
DistanceFromHome (KM)       0
State                       0
Ethnicity                   0
Education                   0
EducationField              0
JobRole                     0
MaritalStatus               0
Salary                      0
StockOptionLevel            0
OverTime                    0
HireDate                    0
Attrition                   0
YearsAtCompany              0
YearsInMostRecentRole       0
YearsSinceLastPromotion     0
YearsWithCurrManager        0
dtype: int64
```

```
EmployeeID: 0 عدد الصفوف المكررة بناءً على
عدد الصفوف بعد حذف التكرارات: 1470
```

```
القيم الفريدة في Gender:
['Female' 'Male' 'Non-Binary' 'Prefer Not To Say']
القيم الفريدة في Gender بعد التنظيف:
['female' 'male' 'non-binary' 'prefer not to say']

القيم الفريدة في MaritalStatus:
['Divorced' 'Single' 'Married']
القيم الفريدة في MaritalStatus بعد التنظيف:
['divorced' 'single' 'married']

القيم الفريدة في Department:
['Sales' 'Human Resources' 'Technology']

القيم الفريدة في JobRole:
['Sales Executive' 'HR Business Partner' 'Engineering Manager' 'Recruiter'
 'Data Scientist' 'Machine Learning Engineer' 'Manager'
 'Software Engineer' 'Senior Software Engineer' 'Sales Representative'
 'Analytics Manager' 'HR Executive' 'HR Manager']

القيم الفريدة في Education:
[5 4 3 2 1]
```

Understanding data structure

Employee Data:

	EmployeeID	FirstName	LastName	Gender	Age	BusinessTravel	\
0	3012-1A41	Leonelle	Simco	Female	30	Some Travel	
1	CBCB-9C9D	Leonerd	Aland	Male	38	Some Travel	
2	95D7-1CE9	Ahmed	Sykes	Male	43	Some Travel	
3	47A0-559B	Ermentrude	Berrie	Non-Binary	39	Some Travel	
4	42CC-040A	Stace	Savege	Female	29	Some Travel	

	Department	DistanceFromHome (KM)	State	Ethnicity	...	\
0	Sales	27	IL	White	...	
1	Sales	23	CA	White	...	
2	Human Resources	29	CA	Asian or Asian American	...	
3	Technology	12	IL	White	...	
4	Human Resources	29	CA	White	...	

	MaritalStatus	Salary	StockOptionLevel	OverTime	HireDate	Attrition	\
0	Divorced	102059	1	No	2012-01-03	No	
1	Single	157718	0	Yes	2012-01-04	No	
2	Married	309964	1	No	2012-01-04	No	
3	Married	293132	0	No	2012-01-05	No	
4	Single	49606	0	No	2012-01-05	Yes	

Understanding data structure

Performance Rating Data:

	PerformanceID	EmployeeID	ReviewDate	EnvironmentSatisfaction	\
0	PR01	79F7-78EC	1/2/2013	5	
1	PR02	B61E-0F26	1/3/2013	5	
2	PR03	F5E3-48BB	1/3/2013	3	
3	PR04	0678-748A	1/4/2013	5	
4	PR05	541F-3E19	1/4/2013	5	

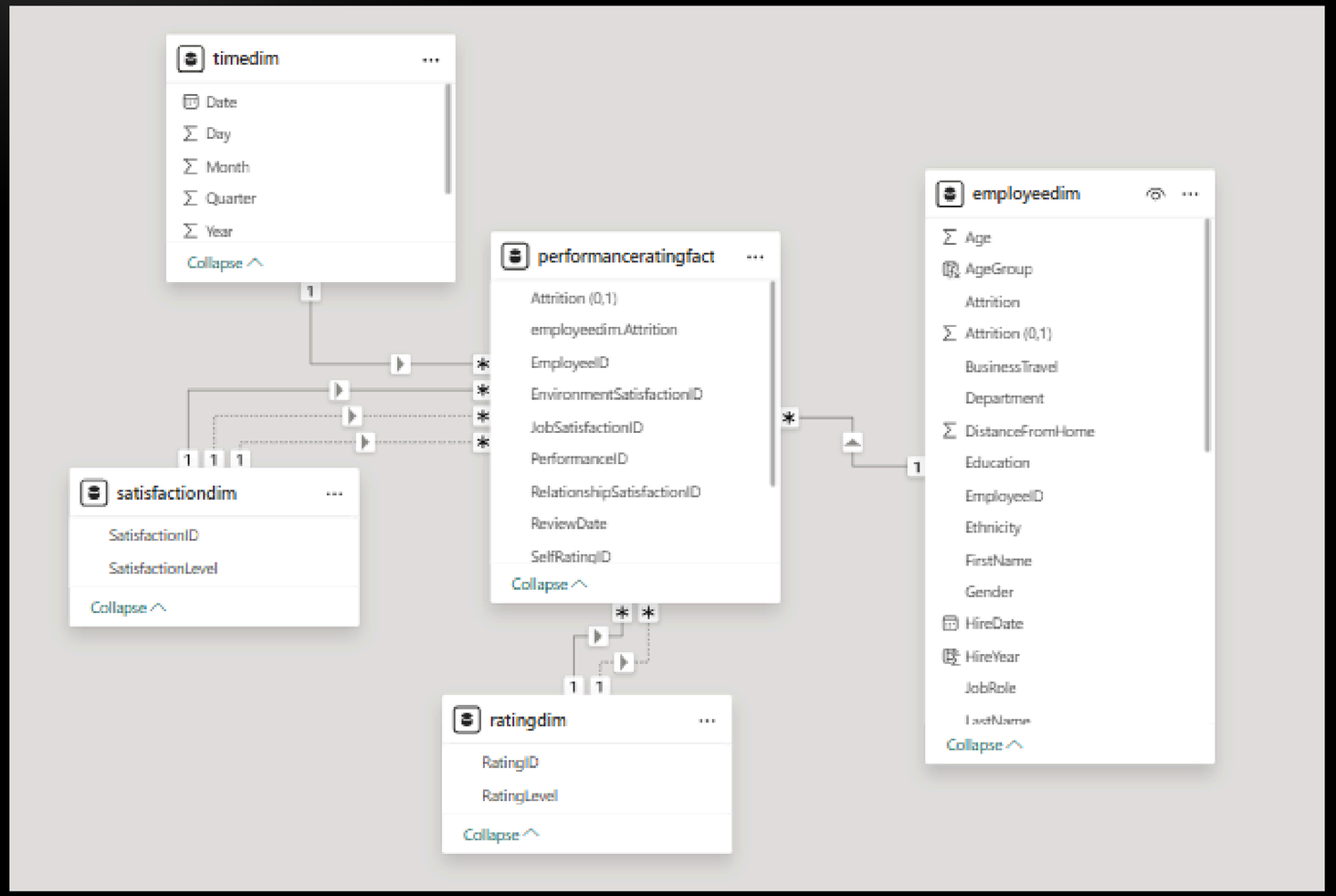
	JobSatisfaction	RelationshipSatisfaction	TrainingOpportunitiesWithinYear	\
0	4	5	1	
1	4	4	1	
2	4	5	3	
3	3	2	2	
4	2	3	1	

	TrainingOpportunitiesTaken	WorkLifeBalance	SelfRating	ManagerRating
0	0	4	4	4
1	3	4	4	3
2	2	3	5	4
3	0	2	3	2
4	0	4	4	3

Understanding data structure

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1470 entries, 0 to 1469
Data columns (total 23 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   EmployeeID                1470 non-null   object
 1   FirstName                 1470 non-null   object
 2   LastName                  1470 non-null   object
 3   Gender                    1470 non-null   object
 4   Age                       1470 non-null   int64
 5   BusinessTravel            1470 non-null   object
 6   Department                1470 non-null   object
 7   DistanceFromHome (KM)     1470 non-null   int64
 8   State                     1470 non-null   object
 9   Ethnicity                 1470 non-null   object
 10  Education                 1470 non-null   int64
 11  EducationField            1470 non-null   object
 12  JobRole                   1470 non-null   object
 13  MaritalStatus             1470 non-null   object
 14  Salary                    1470 non-null   int64
 15  StockOptionLevel          1470 non-null   int64
 16  OverTime                  1470 non-null   object
 17  HireDate                  1470 non-null   object
 18  Attrition                 1470 non-null   object
 19  YearsAtCompany            1470 non-null   int64
 20  YearsInMostRecentRole     1470 non-null   int64
 21  YearsSinceLastPromotion   1470 non-null   int64
 22  YearsWithCurrManager      1470 non-null   int64
dtypes: int64(9), object(14)
memory usage: 264.3+ KB
```

Final Schema



DIGITAL EGYPT PIONEERS INITIATIVE

DATA ANALYSIS

Data Analysis

```sql
---------------------------------- Relationship Between JobRole and Avg Job Satisfac
WITH LatestRating AS (
    SELECT
        p.EmployeeID,
        p.JobSatisfactionID,
        ROW_NUMBER() OVER (PARTITION BY p.EmployeeID ORDER BY t.Date DESC) AS rn
    FROM
        PerformanceRatingFact p
    JOIN
        TimeDim t ON p.ReviewDate = t.Date  -- تغيرها من TimeID إلى Date
)
SELECT
    e.JobRole,
    e.Department,
    COUNT(CASE WHEN e.Attrition = 'Yes' THEN 1 END) AS AttritionCount
FROM
    EmployeeDim e
JOIN
    LatestRating lr ON e.EmployeeID = lr.EmployeeID
WHERE
    lr.rn = 1
GROUP BY
    e.JobRole,
    e.Department
ORDER BY
    e.Department,
    e.JobRole;
```

WITH LatestRating AS (SELECT p.EmployeeID, p.JobSatisf | Enter a SQL expression to

	JobRole	Department	AttritionCount
1	hr business partner	human resources	0
2	hr executive	human resources	1
3	hr manager	human resources	0
4	recruiter	human resources	6
5	manager	sales	1
6	sales executive	sales	47
7	sales representative	sales	24
8	analytics manager	technology	3
9	data scientist	technology	46
10	engineering manage	technology	1
11	machine learning er	technology	6
12	sales executive	technology	0
13	senior software engi	technology	6
14	software engineer	technology	35

Data Analysis

```sql
---------------------------------------------------------- Relationship Between Av
WITH LatestRating AS (
    SELECT
        p.EmployeeID,
        ROW_NUMBER() OVER (PARTITION BY p.EmployeeID ORDER BY t.Year DESC) AS rn
    FROM
        PerformanceRatingFact p
    JOIN
        TimeDim t ON p.ReviewDate = t.Date
)
SELECT
    e.Department,
    AVG(e.Salary) AS AvgSalary,
    COUNT(CASE WHEN e.Attrition = 'Yes' THEN 1 END) AS AttritionCount
FROM
    EmployeeDim e
JOIN
    LatestRating lr ON e.EmployeeID = lr.EmployeeID
WHERE
    lr.rn = 1
GROUP BY
    e.Department
ORDER BY
    AvgSalary DESC;
```

	Department	AvgSalary	AttritionCount
1	sales	124,919.8275862069	72
2	technology	113,379.9107373868	97
3	human resources	112,843.875	7

Data Analysis

```sql
---------------------------------------------------------- How Many Training Oppor
WITH LatestRating AS (
    SELECT
        p.EmployeeID,
        p.TrainingOpportunitiesTaken,
        ROW_NUMBER() OVER (PARTITION BY p.EmployeeID ORDER BY t.Year DESC) AS rn
    FROM
        PerformanceRatingFact p
    JOIN
        TimeDim t ON p.ReviewDate = t.Date
)
SELECT
    e.Education,
    AVG(CAST(lr.TrainingOpportunitiesTaken AS FLOAT)) AS AvgTrainingTaken,
    COUNT(DISTINCT e.EmployeeID) AS EmployeeCount,
    COUNT(CASE WHEN e.Attrition = 'Yes' THEN 1 END) AS AttritionCount
FROM
    EmployeeDim e
JOIN
    LatestRating lr ON e.EmployeeID = lr.EmployeeID
WHERE
    lr.rn = 1
GROUP BY
    e.Education
ORDER BY
    e.Education;
```

WITH LatestRating AS (SELECT p.EmployeeID, p.TrainingO

Enter a SQL expression to filter results (use Ctrl+Space)

Grid

Text

	Education	AvgTrainingTaken	EmployeeCount	AttritionCount
1	1	0.9782608696	138	25
2	2	0.8744588745	231	32
3	3	0.9454148472	458	78
4	4	1.0695364238	302	36
5	5	1.1	40	5

Data Analysis

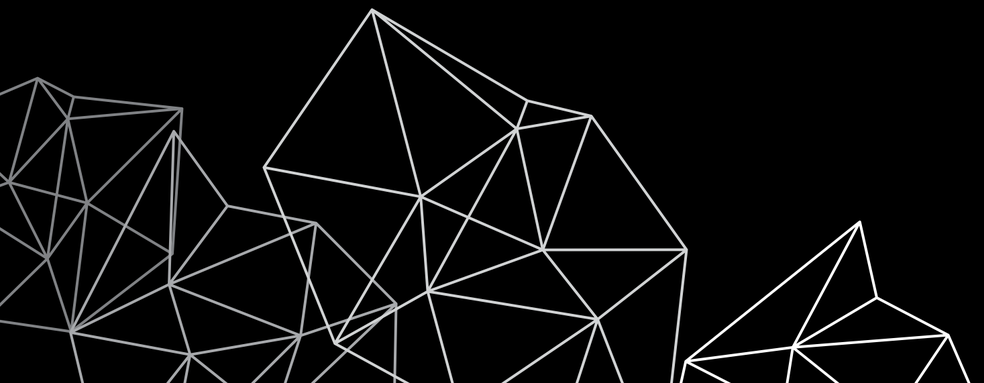
```sql
---------------------------------- Does Distance From Home Affect Attrition?
SELECT
    CASE
        WHEN e.DistanceFromHome <= 10 THEN '0-10 miles'
        WHEN e.DistanceFromHome <= 20 THEN '11-20 miles'
        WHEN e.DistanceFromHome <= 30 THEN '21-30 miles'
        ELSE 'Over 30 miles'
    END AS DistanceGroup,
    COUNT(CASE WHEN e.Attrition = 'Yes' THEN 1 END) AS AttritionCount,
    COUNT(*) AS TotalEmployees,
    (COUNT(CASE WHEN e.Attrition = 'Yes' THEN 1 END) * 100.0 / COUNT(*)) AS AttritionRate
FROM
    EmployeeDim e
GROUP BY
    CASE
        WHEN e.DistanceFromHome <= 10 THEN '0-10 miles'
        WHEN e.DistanceFromHome <= 20 THEN '11-20 miles'
        WHEN e.DistanceFromHome <= 30 THEN '21-30 miles'
        ELSE 'Over 30 miles'
    END
ORDER BY
    DistanceGroup;
```

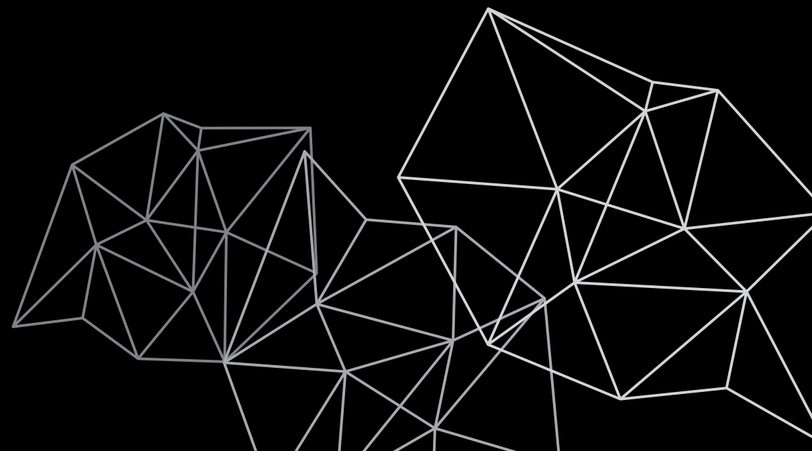
SELECT CASE WHEN e.DistanceFromHome <= 10 THEN '(| Enter a SQL expression to filter results (use Ctrl+Spa

	DistanceGroup	AttritionCount	TotalEmployees	AttritionRate
1	0-10 miles	59	333	17.71772
2	11-20 miles	55	345	15.94203
3	21-30 miles	49	339	14.45428
4	Over 30 miles	74	453	16.33554

DATA FORECASTING

- Importing libraries
- Ai models used
- Data prediction





Importing libraries & Loading data

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from sklearn.linear_model import LinearRegression, LogisticRegression
from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score
```

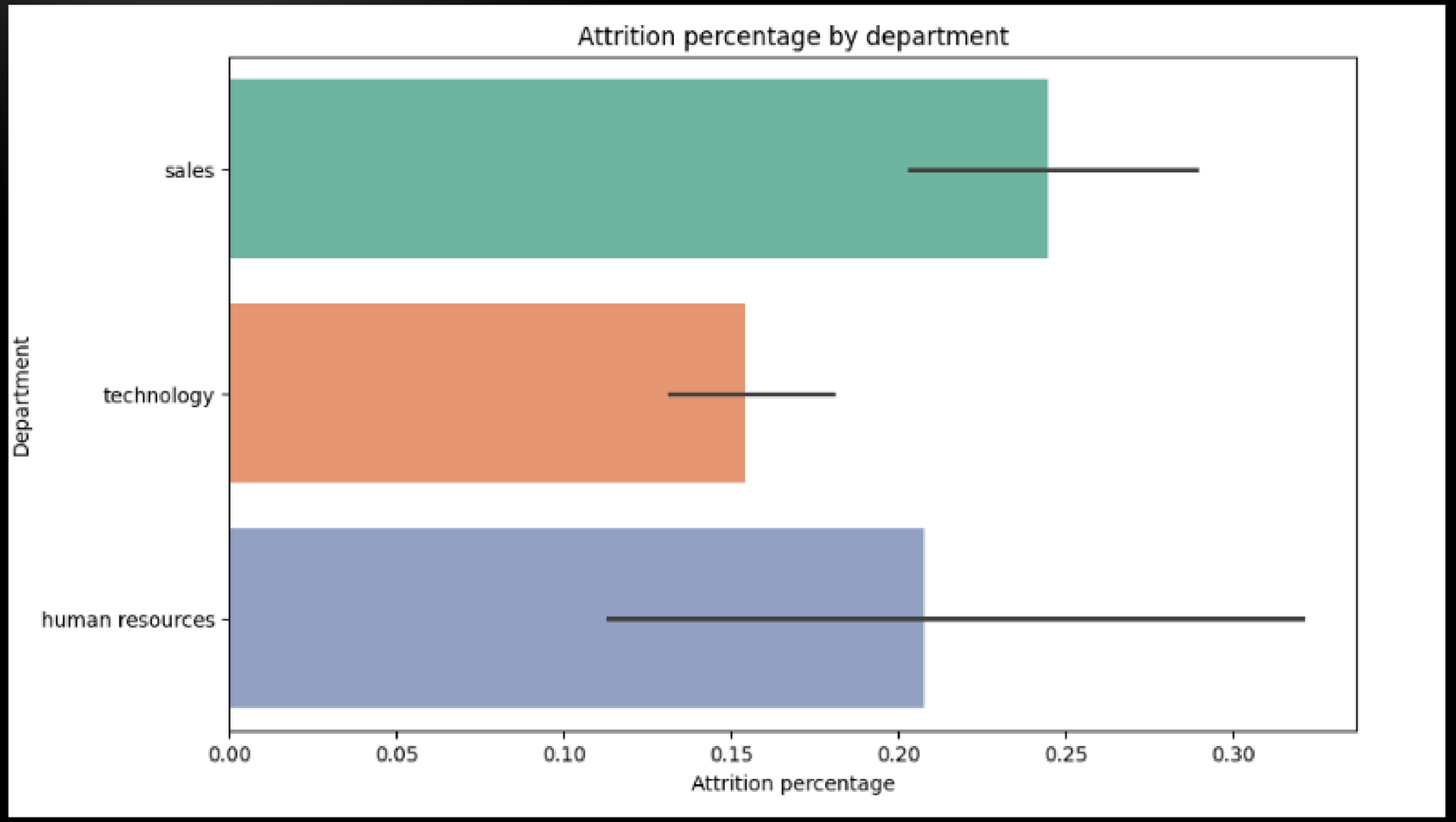
```
===== أول 5 صفوف من الجدول: employeedim =====
EmployeeID | FirstName | LastName | Gender | Age | BusinessTravel | DistanceFromHome | State | Ethnicity | Education | MaritalStatus | Salary | StockOpti
onLevel | OverTime | HireDate | Attrition | YearsAtCompany | YearsInMostRecentRole | YearsSinceLastPromotion | YearsWithCurrManager | Department | JobRol
e
----------------------------------------
001A-8F88 | Christy | Jumel | male | 22 | Some Travel | 40 | CA | White | 4 | married | 27763.0 | 0 | No | 2021-09-05 | No | 1 | 0 | 1 | 0 | technology |
software engineer
005C-E0FB | Fin | O'Halleghane | non-binary | 24 | Frequent Traveller | 17 | CA | White | 4 | married | 56155.0 | 1 | No | 2017-08-26 | No | 5 | 2 | 2 |
0 | sales | sales executive
00A3-2445 | Wyatt | Ziehm | male | 30 | Some Travel | 6 | CA | Black or African American | 2 | married | 126238.0 | 0 | No | 2012-03-08 | No | 10 | 3 | 6
| 6 | technology | machine learning engineer
00B0-F199 | Trueman | Jirasek | male | 23 | Some Travel | 35 | CA | White | 1 | married | 97824.0 | 1 | Yes | 2020-03-16 | Yes | 1 | 0 | 1 | 0 | sales |
sales executive
00D4-DD53 | Joyce | Goor | female | 30 | Frequent Traveller | 44 | CA | Black or African American | 1 | single | 68508.0 | 0 | Yes | 2012-01-28 | Yes | 5
| 4 | 4 | 4 | technology | software engineer

===== أول 5 صفوف من الجدول: performanceratingfact =====
PerformanceID | EmployeeID | EnvironmentSatisfactionID | JobSatisfactionID | RelationshipSatisfactionID | WorkLifeBalanceID | SelfRatingID | TrainingOppo
rtunitiesWithinYear | TrainingOpportunitiesTaken | ReviewDate
----------------------------------------
PR01 | 79F7-78EC | 5 | 4 | 5 | 4 | 4 | 1 | 0 | 2013-01-02
PR02 | B61E-0F26 | 5 | 4 | 4 | 4 | 4 | 1 | 3 | 2013-01-03
PR03 | F5E3-48BB | 3 | 4 | 5 | 3 | 5 | 3 | 2 | 2013-01-03
PR04 | 0678-748A | 5 | 3 | 2 | 2 | 3 | 2 | 0 | 2013-01-04
PR05 | 541F-3E19 | 5 | 2 | 3 | 4 | 4 | 1 | 0 | 2013-01-04
```

AI models used

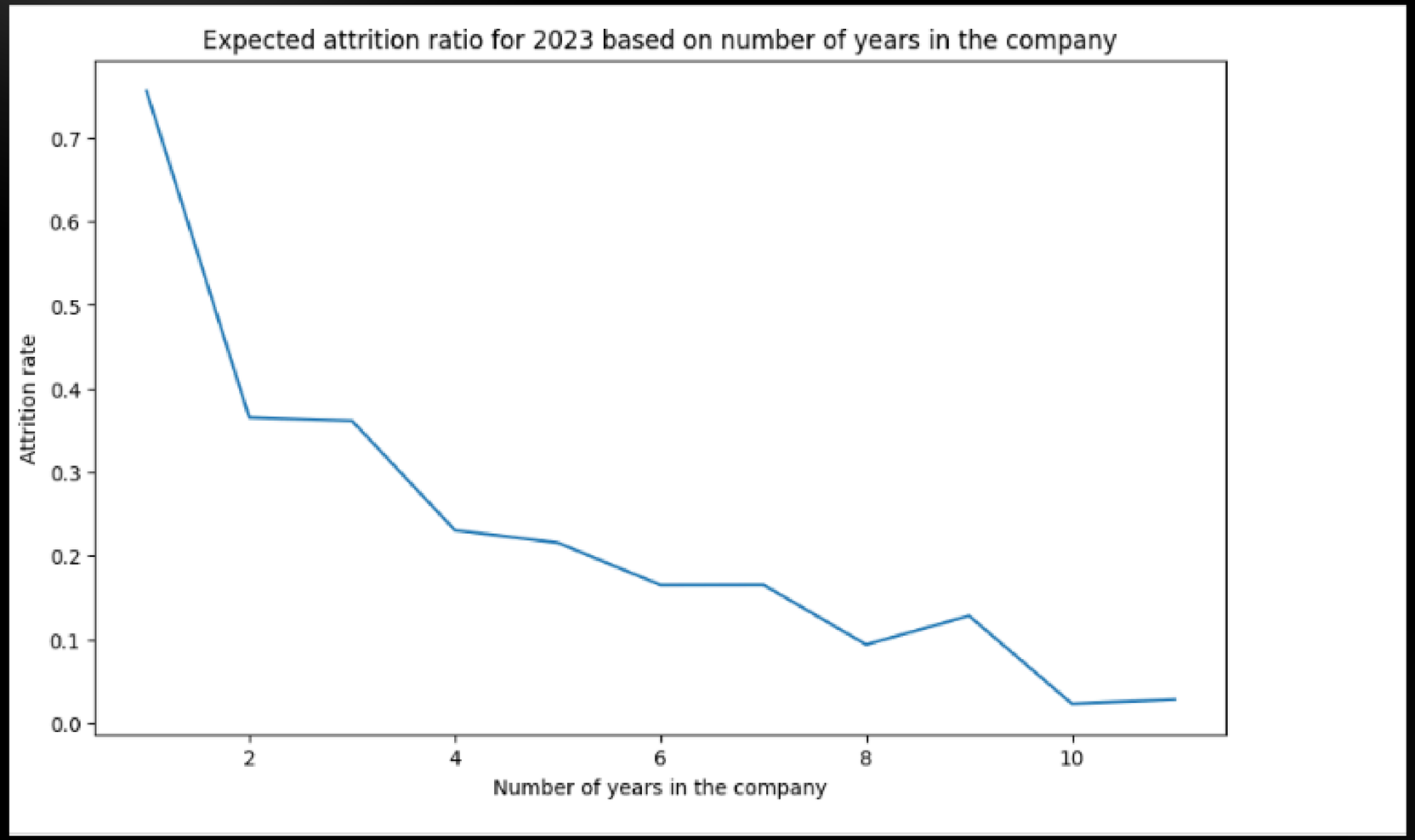
```python
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import accuracy_score

model = RandomForestClassifier(random_state=42)

model.fit(X_train, y_train)

y_pred = model.predict(X_test)

accuracy = accuracy_score(y_test, y_pred)
print(f"دقة النموذج: {accuracy:.2f}")
```

```python
from xgboost import XGBClassifier
from sklearn.metrics import accuracy_score, classification_report

scale_pos_weight = len(y_train[y_train == 0]) / len(y_train[y_train == 1])

model = XGBClassifier(scale_pos_weight=scale_pos_weight, max_depth=3, n_estimators=50, random_state=42)
model.fit(X_train, y_train)

y_pred = model.predict(X_test)
accuracy = accuracy_score(y_test, y_pred)
print(f"\nModel accuracy: {accuracy:.2f}")

print("\nPerformance report:")
print(classification_report(y_test, y_pred))
```

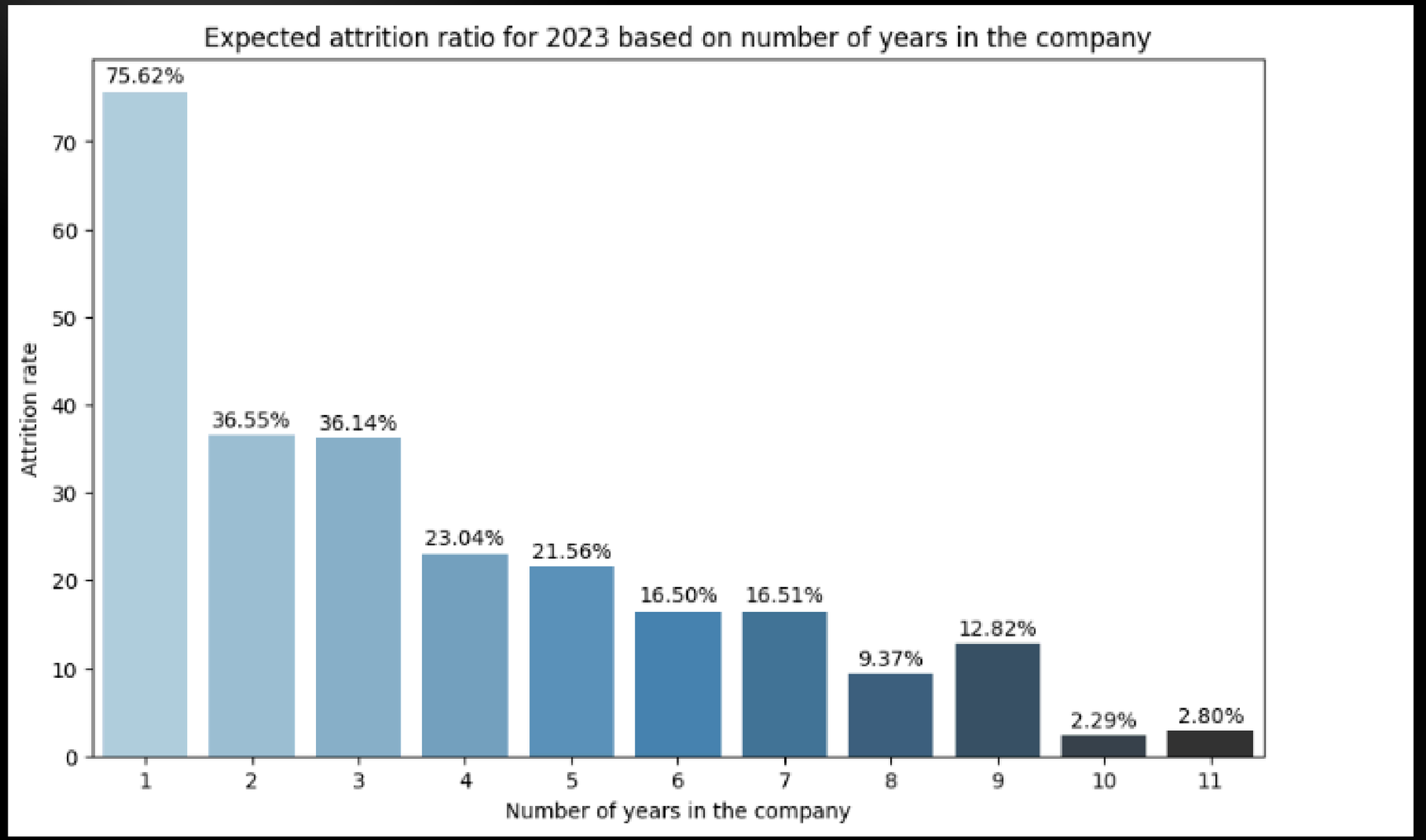
Data prediction



Data prediction



Data prediction

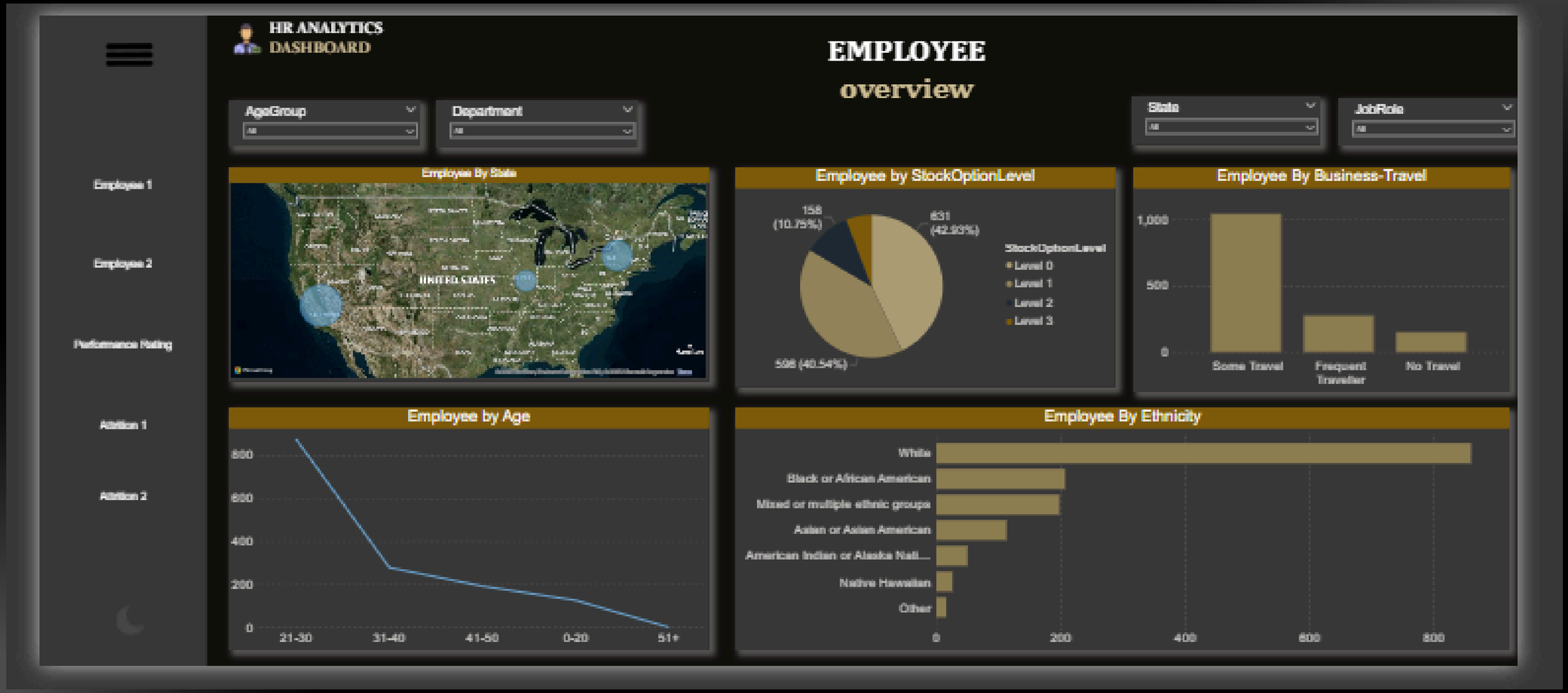


DIGITAL EGYPT PIONEERS INITIATIVE

FINAL DASHBOARD

Employee Overview

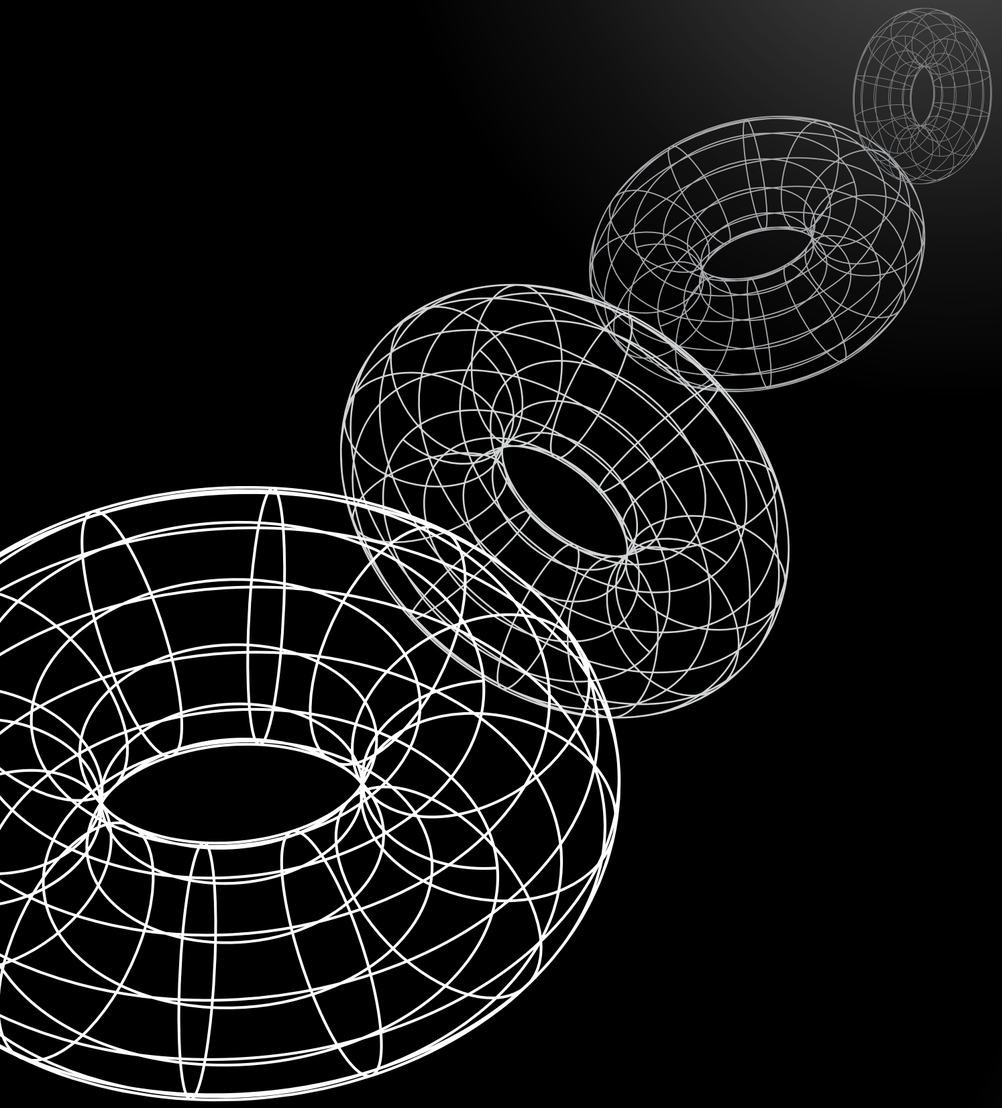
Employee Overview



Performance rating Overview

Attrition Overview

Attrition Overview



INSIGHTS & RECOMMENDATIONS

Insights & Recommendations

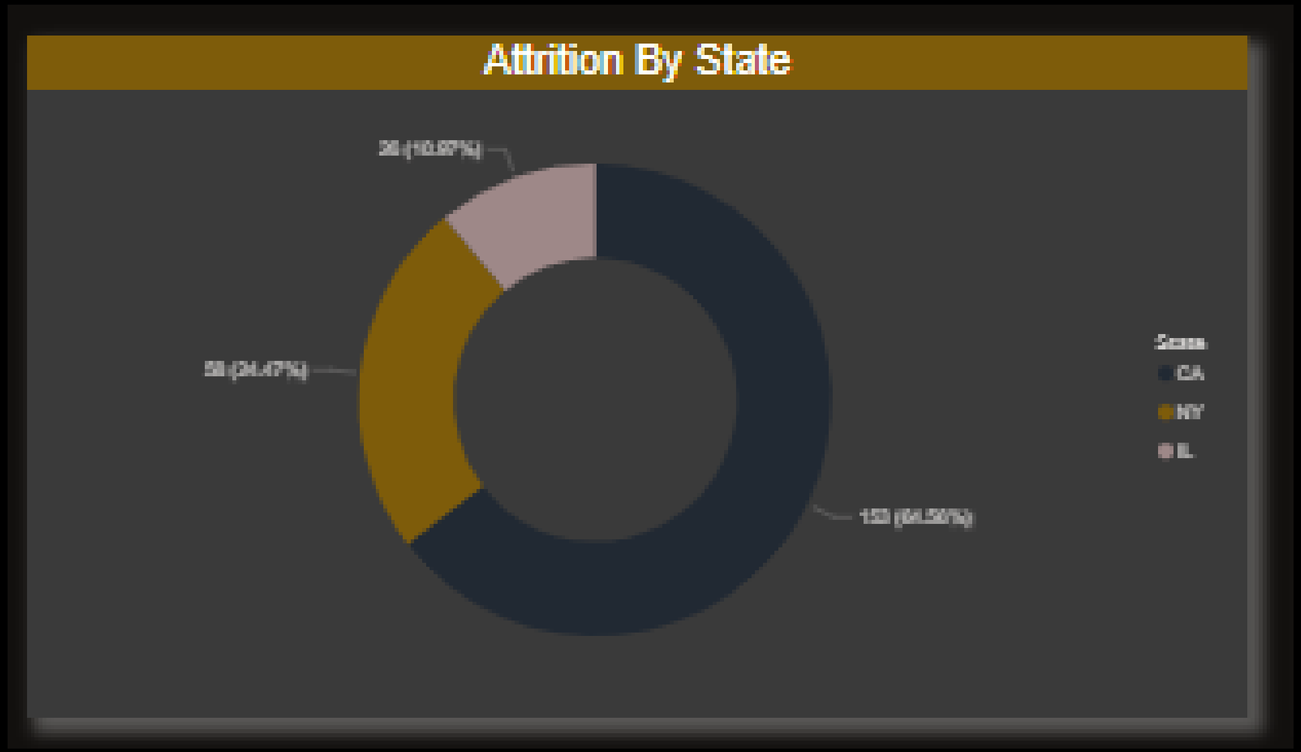
- SINCE MOST OF THE DEPARTURES ARE FROM CALIFORNIA, THIS SUGGESTS THERE ARE REASONS SPECIFIC TO THE REGION (SUCH AS HIGH COST OF LIVING OR WORKING CONDITIONS)

- THE COST OF LIVING IN CALIFORNIA IS VERY HIGH (LIKE LIVING IN LOS ANGELES OR SAN FRANCISCO), AND THIS MAY BE A REASON TO LEAVE.

RECOMMENDATION :

- CONDUCT EMPLOYEE SURVEYS IN CALIFORNIA TO UNDERSTAND CHALLENGES (SUCH AS SALARIES, TRANSPORTATION, OR WORK ENVIRONMENT).

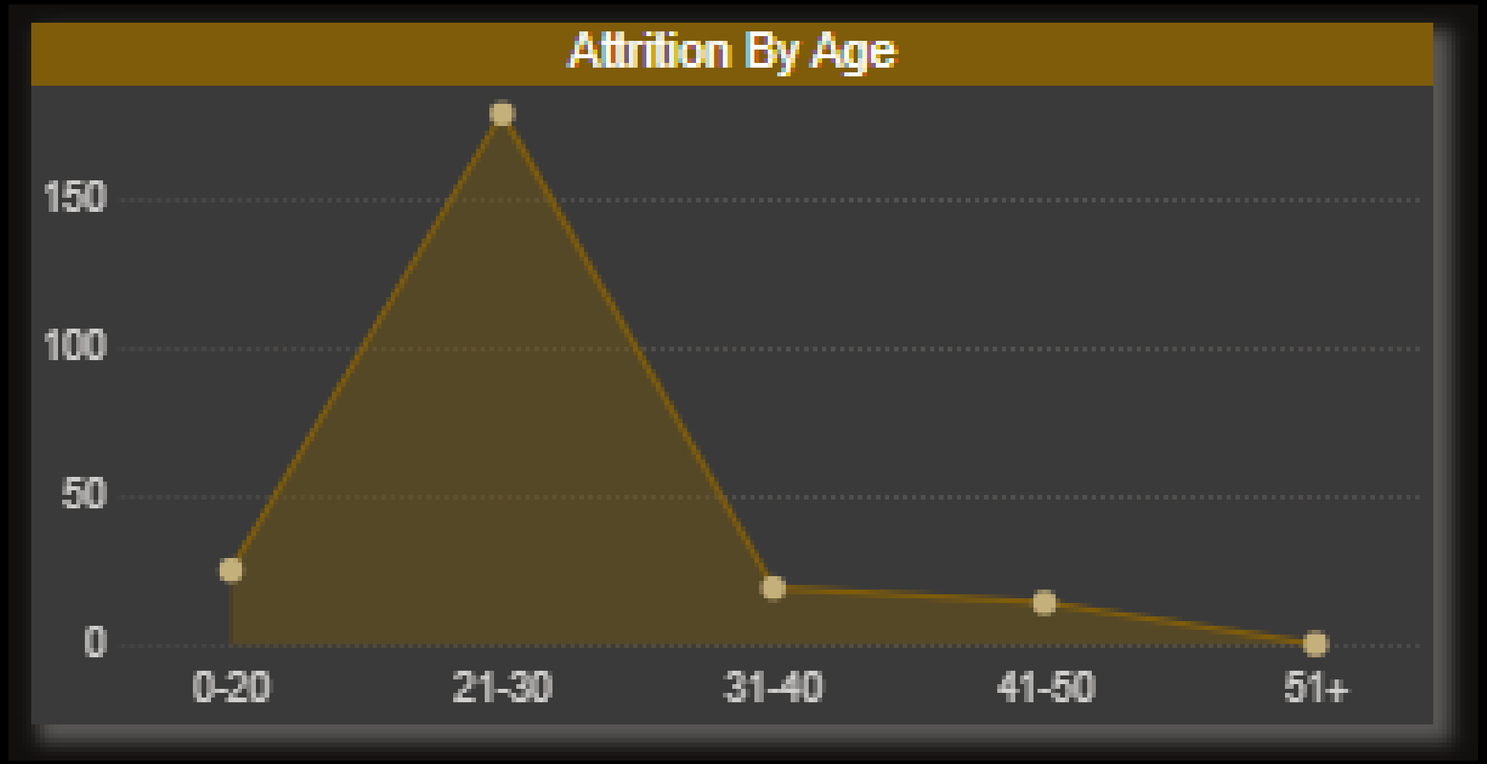
- OFFER FINANCIAL INCENTIVES (SUCH AS A HOUSING ALLOWANCE) OR REMOTE WORK OPTIONS TO EASE FINANCIAL STRESS.



Avg Salary	AVG distance from home
114.06K	22.50

Insights & Recommendations

- THE 21-30 AGE GROUP ACCOUNTS FOR THE HIGHEST ATTRITION RATE LIKELY DUE TO EARLY-CAREER CHALLENGES, SALARY EXPECTATIONS, OR WORK-LIFE BALANCE ISSUES



RECOMMENDATION :

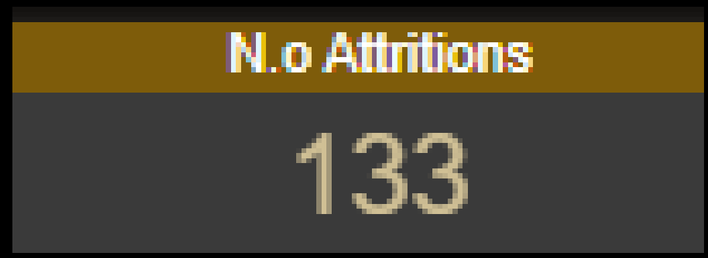
- DEVELOP TARGETED TRAINING PROGRAMS, OFFER CLEAR CAREER PATHS, COMPETITIVE INCENTIVES, AND FLEXIBLE WORK OPTIONS FOR THE 21-30 AGE GROUP

Insights & Recommendations

- THE TECHNOLOGY DEPARTMENT HAS THE HIGHEST ATTRITION RATE, SUGGESTING SPECIFIC CHALLENGES LIKE WORKLOAD, SKILL DEMANDS, OR LACK OF GROWTH OPPORTUNITIES.

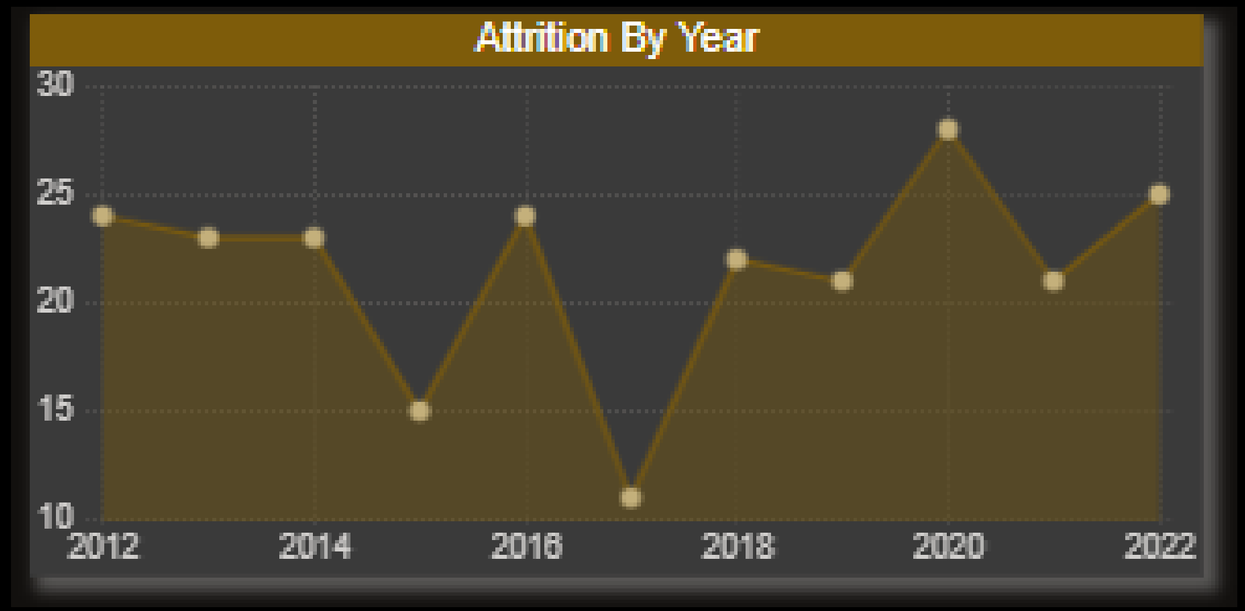
RECOMMENDATION :

- IMPLEMENT TARGETED RETENTION STRATEGIES FOR THE TECHNOLOGY DEPARTMENT, INCLUDING SKILL DEVELOPMENT PROGRAMS, WORKLOAD MANAGEMENT, AND COMPETITIVE



Insights & Recommendations

- THE HIGHEST ATTRITION OCCURRED IN 2020,WITH (28) EMPLOYEES LEAVING, LIKELY INFLUENCED BY THE UNCERTAINTY AND DISRUPTIONS CAUSED BY THE COVID-19 PANDEMIC



RECOMMENDATION :

- ENHANCE EMPLOYEE SUPPORT DURING CRISES LIKE COVID-19 BY OFFERING REMOTE WORK OPTIONS, MENTAL HEALTH RESOURCES, AND FINANCIAL STABILITY MEASURES TO REDUCE FUTURE ATTRITION.

Insights & Recommendations

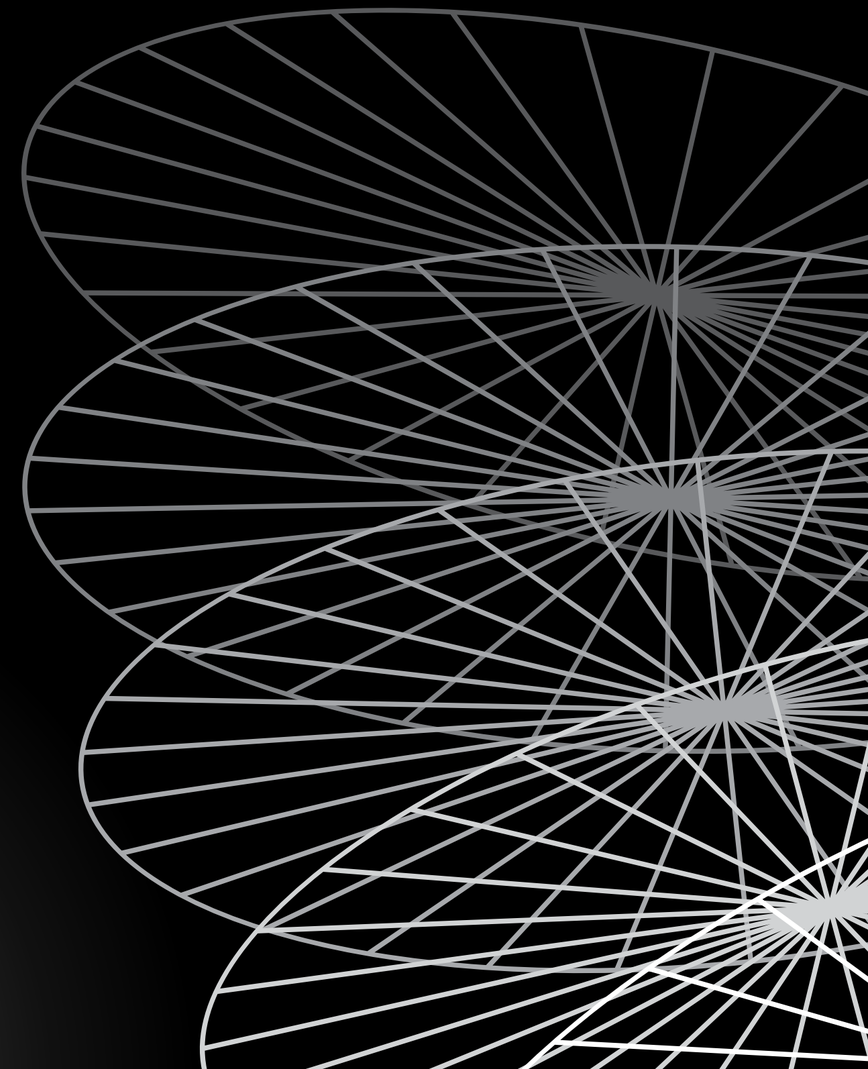
- THE VAST MAJORITY OF EMPLOYEES ARE DISSATISFIED WITH THEIR JOB ROLES, INDICATING A MISMATCH BETWEEN THEIR SKILLS OR INTERESTS AND THE NATURE OF THE WORK.

RECOMMENDATION :

- RE-EVALUATE JOB ROLE DISTRIBUTION WITH CUSTOMIZED TRAINING TO IMPROVE SKILLS, CONDUCT SURVEYS TO UNDERSTAND EMPLOYEE NEEDS AND MODIFY JOBS ACCORDINGLY

CONCLUSION

THIS PROJECT HAS PROVIDED A COMPREHENSIVE ANALYSIS OF EMPLOYEE RETENTION AND PERFORMANCE THROUGH A ROBUST POWER BI DASHBOARD, LEVERAGING DATA FROM EMPLOYEE AND PERFORMANCERATING UP TO 2022. KEY INSIGHTS REVEAL THAT THE HIGHEST ATTRITION OCCURRED IN 2020, LIKELY INFLUENCED BY THE COVID-19 PANDEMIC, WITH A SIGNIFICANT CONCENTRATION OF DEPARTURES AMONG THE 21-30 AGE GROUP AND THE TECHNOLOGY DEPARTMENT. ADDITIONALLY, A NOTABLE DISSATISFACTION WITH JOB ROLES AND LOW JOB SATISFACTION LEVELS HAVE EMERGED AS CRITICAL FACTORS DRIVING EMPLOYEE TURNOVER. THE ANALYSIS ALSO HIGHLIGHTS GENDER DISPARITIES IN TENURE WITH MANAGERS AND REGIONAL CHALLENGES, PARTICULARLY IN CALIFORNIA. THESE FINDINGS UNDERSCORE THE NEED FOR TARGETED INTERVENTIONS TO ADDRESS SPECIFIC PAIN POINTS, ENHANCE EMPLOYEE ENGAGEMENT, AND FOSTER A SUPPORTIVE WORK ENVIRONMENT. BY IMPLEMENTING THE RECOMMENDED STRATEGIES, THE ORGANIZATION CAN MITIGATE ATTRITION, IMPROVE PERFORMANCE, AND BUILD A MORE RESILIENT WORKFORCE FOR THE FUTURE.



THANK YOU