Human Resources Dataset Analysis

Group C

Under Supervision Of:

Prof. Dr/ Wael Mahrous

Presented By:

Rashad Hazem Aly	1112542203	Leader
Mariem Korashi Khalaf	1112760217	Presenter
Basem Barakat Allah	1113635223	Member
Rasha Gaafar Osman	1123640976	Member
Mohamed AbdElAlim	1112535391	Member
Esraa Magdy Elsaid	1110329116	Member

What is HR analytics?

*HR analytics (usually referred to as people analytics or workforce analytics) involves gathering, analysing, and reporting HR data to drive business results.
*HR analytics enables organizations to better understand of their workforce, make decisions based on data, test the effectiveness of HR policies and interventions and measure the impact of a range of HR metrics, ultimately improving overall business performance to as much as a 25% rise in business productivity, a 50% decrease in attrition rates, and 80% increase in recruiting efficiency.
*In other words, HR analytics is a data-driven approach to Human Resources Management.

HR Database:

HR database is a system where you store and manage data on your company's employees. HR databases can be used to track a variety of information, including HR metrics, which give the HR team insights for better decision-making.

It may include:

Employee's name, Address, Contact info., Job title, Who they report to, Annual leaves, Benefits, Absenteeism, Skills, Salaries, Training and Development Programs

HR Dataset:

Employees Table

Employ	FirstNa	LastNa	Gender	Age	Busine	Departi	DistanceFromHom	State	Ethnici	Educati	EducationFiel	JobRol	MaritalSt	Salary	StockOptionLe	OverTime	HireDate	Attritio	YearsAtCompany	YearsInMostRecentRole	YearsSinceLastPromotic	YearsWithCurrManager
3012-1A4	Leonelle	Simco	Female	30	Some Tr	Sales	27	IL	White	5	Marketing	Sales Ex	Divorced	102059	1	No	03/01/2012	No	10	4	9	7
CBCB-9	Leonerd	Aland	Male	38	Some Tr	Sales	23	CA	White	4	Marketing	Sales Ex	Single	157718	0	Yes	04/01/2012	No	10	6	10	0
95D7-1C	Ahmed	Sykes	Male	43	Some Tr	Human F	29	CA	Asian or	4	Marketing	HR Busi	Married	309964	1	No	04/01/2012	No	10	6	10	8
47A0-55	Ermentr	Berrie	Non-Bin	39	Some Tr	Technol	12	IL	White	3	Computer Scien	Engineer	Married	293132	0	No	05/01/2012	No	10	10	10	0
42CC-04	Stace	Savege	Female	29	Some Tr	Human F	29	CA	White	2	Technical Degre	Recruite	Single	49606	0	No	05/01/2012	Yes	6	1	1	6
C219-6C	Clerkcla	Hinkins	Male	34	Some Tr	Sales	30	NY	Mixed or	2	Marketing	Sales Ex	Divorced	133468	1	No	05/01/2012	No	10	3	7	9
D906-BE	Uta	Melmar	Female	42	No Trav	Technol	45	NY	Black or	3	Information Syst	Engineer	Married	259284	1	No	09/01/2012	No	10	2	6	6
3C7D-8E	Joyan	Brason	Female	40	Some Tr	Sales	3	CA	Native H	2	Other	Sales Ex	Divorced	104426	1	No	11/01/2012	No	10	3	4	6
3D71-8D	Alix	Blazejew	Male	38	Some Tr	Sales	20	IL	Black or	4	Marketing	Sales Ex	Married	147098	1	No	11/01/2012	No	10	5	8	2
5476-CA	Kayley	Snoad	Female	31	Frequent	Technol	4	NY	Native H	2	Information Syst	Data Sci	Single	69747	0	No	12/01/2012	Yes	6	5	5	1
73CF-49	Hannis	Waslin	Female	32	Some Tr	Technol	42	CA	White	4	Computer Scien	Data Sci	Single	102022	0	No	12/01/2012	No	10	4	5	8
277A-AE	Annabel	Pablos	Female	35	Some Tr	Technol	8	NY	Other	4	Information Syst	Machine	Single	272175	0	No	12/01/2012	No	10	7	8	2
8BAB-B	Torey	Abram	Male	38	Some Tr	Sales	35	NY	Asian or	3	Marketing	Manager	Single	340229	0	Yes	13/01/2012	No	10	7	9	3
111D-E5E	Edna	Alison	Non-Bin	37	Some Tr	Technol	3	IL	White	3	Computer Scien	Software	Divorced	48395	1	No	15/01/2012	No	10	2	2	4
97F4-0B	Vernen	Powner	Male	33	Some Tr	Technol	4	NY	Mixed or	4	Other	Senior S	Single	97126	0	No	15/01/2012	No	10	8	10	2
5C03-10(	Willetta	Lurriman	Female	42	Some Tr	Technol	21	IL	Black or	3	Information Syst	Engineer	Married	316208	1	No	17/01/2012	No	10	8	8	7
BD1B-53	Wendall	Dryden	Male	43	Some Tr	Sales	27	CA	Mixed or	3	Marketing	Sales Ex	Single	128885	0	Yes	17/01/2012	No	10	1	10	2
DFA9-9!	Cale	Holston	Male	43	No Trav	Sales	34	NY	White	4	Marketing	Sales Ex	Married	108315	2	No	17/01/2012	No	10	9	10	10
ED73-FC	Ernaline	Napolior	Female	45	Frequent	Technol	19	NY	Asian or	1	Computer Scien	Software	Married	136521	1	No	18/01/2012	No	10	3	6	1
C6EC-FI	Charlena	Severwri	Female	38	Some Tr	Sales	1	CA	White	1	Economics	Sales Ex	Single	151141	0	No	19/01/2012	No	10	3	6	9
D7EE-5E	Zsa zsa	Evered	Female	39	Frequent	Sales	17	CA	White	2	Technical Degre	Sales Ex	Married	107863	1	Yes	19/01/2012	Yes	8	8	8	5
C395-8C	Curcio	Franek	Male	33	Some Tr	Human F	3	CA	Mixed or	1	Technical Degre	Recruite	Divorced	53616	1	Yes	19/01/2012	No	10	2	3	7
E348-E1	Burnaby	Guillet	Male	36	No Trav	Technol	36	IL	White	2	Information Syst	Software	Divorced	61298	1	Yes	19/01/2012	No	10	3	9	3
B3AF-7E	Elvira	Ianelli	Female	45	Some Tr	Human F	34	NY	Black or	2	Human Resourc	Recruite	Divorced	54132	1	Yes	20/01/2012	No	10	10	10	10
469A-81	Baxie	Rising	Male	30	Some Tr	Technol	36	CA	America	4	Information Syst	Engineer	Married	328415	0	No	20/01/2012	No	10	1	10	1
9E22-62	Gifford	Poynser	Non-Bin	48	Some Tr	Technol	37	CA	White	3	Computer Scien	Machine	Single	145337	0	No	21/01/2012	No	10	10	10	0
D5DA-3!	Rickey	Shere	Male	33	Some Tr	Sales	41	CA	Asian or	2	Marketing	Sales Ex	Married	71201	2	No	24/01/2012	No	10	8	10	4
BFF3-A.	Collen	Sedman	Female	31	No Trav	Human F	25	NY	America	2	Marketing	Recruite	Divorced	55682	1	No	25/01/2012	No	10	0	2	0
C0BE-FI	Bertram	Dolemar	Male	40	Some Tr	Sales	35	NY	White	2	Marketing	Sales Re	Married	63455	1	Yes	26/01/2012	No	10	4	8	7
D565-28	Bessie	Bellson	Female	47	Frequent	Sales	21	CA	White	3	Economics	Sales Ex	Single	65626	0	No	27/01/2012	No	10	2	5	1
00D4-DI	Joyce	Goor	Female	30	Frequent	Technol	44	CA	Black or	1	Computer Scien	Software	Single	68508	0	Yes	28/01/2012	Yes	5	4	4	4
7FFD-C!	Claresta	Impy	Female	32	Some Tr	Technol	22	IL	White	2	Other	Data Sci	Married	75821	1	No	30/01/2012	No	10	1	5	1
07B2-DE	Rossie	Everleigh	Male	44	Some Tr	Sales	40	NY	White	2	Economics	Manager	Divorced	285620	1	No	30/01/2012	No	10	6	8	4
1749-81A	Elora	Bentjens	Female	48	Some Tr	Technol	17	CA	Black or	3	Information Syst	Software	Single	65860	0	No	01/02/2012	No	10	6	7	1
FF14-A4	Koenraa	Nannizzi	Male	47	Some Tr	Technol	41	NY	America	1	Computer Scien	Data Sci	Married	40786	0	No	01/02/2012	No	10	3	5	5
3CD6-5E	Dorise	Klishin	Female	31	Some Tr	Technol	20	CA	White	2	Computer Scien	Software	Single	59697	0	Yes	02/02/2012	Yes	7	5	6	0
4FC2-A4	Gayle	Riseley	Female	32	No Trav	Technol	12	NY	Black or	3	Computer Scien	Engineer	Married	316725	1	No	02/02/2012	No	10	6	10	2
9AAB-D	Staci	Leith	Female	28	Some Tr	Technol	14	IL	White	2	Business Studie	Software	Divorced	105984	2	Yes	04/02/2012	No	10	7	9	10
427A-8C	Pasquale	Abrehea	Male	32	Some Tr	Technol	15	CA	Mixed or	3	Other	Analytics	Married	393294	1	No	04/02/2012	No	10	2	2	2
A923-9E	Alaine	Hinrichs	Female	41	Frequent	Technol	7	CA	Black or	1	Business Studie	Machine	Single	107008	0	No	05/02/2012	No	10	9	10	9
5C61-8F	Dyana	Gallie	Female	34	Some Tr	Technol	5	CA	White	1	Information Syst	Data Sci	Single	127432	0	Yes	06/02/2012	No	10	8	9	9
CF2F-8C	Lindy	Rawstor	Male	45	Some Tr	Sales	20	CA	White	2	Marketing	Sales Ex	Married	76693	1	No	06/02/2012	No	10	2	9	10
40A9-8E	Joannes	McFadd	Female	39	Some Tr	Technol	31	CA	White	3	Information Syst	Software	Single	90017	0	Yes	07/02/2012	No	10	7	7	3
2E72-4B	Grace	Gohier	Male	44	Some Tr	Sales	21	CA	White	3	Marketing	Manager	Married	426142	0	No	07/02/2012	No	10	4	5	4

Education Level Table

EducationI	EducationLevel
1	No Formal Qualifications
2	High School
3	Bachelors
4	Masters
5	Doctorate

Rating Level Table

RatingID	RatingLevel
1	Unacceptable
2	Needs Improvement
3	Meets Expectation
4	Exceeds Expectation
5	Above and Beyond

Satisfied Level Table

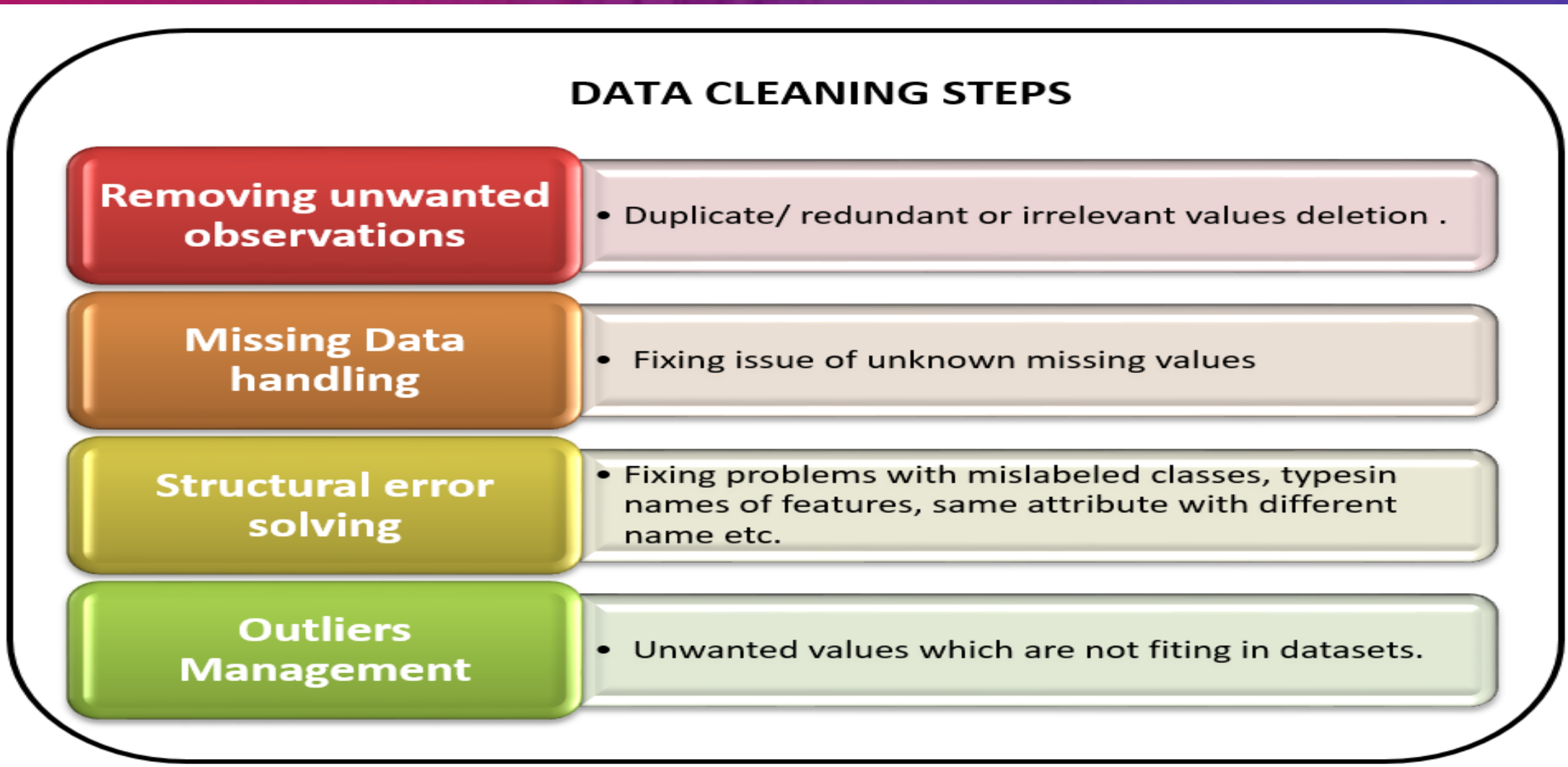
SatisfactionID	SatisfactionLevel
1	Very Dissatisfied
2	Dissatisfied
3	Neutral
4	Satisfied
5	Very Satisfied

Performance Level Table

Performance	EmployeeI	ReviewDat	EnvironmentSatis	JobSatisfac	RelationshipSa	TrainingOpportunitiesWithi	TrainingOpportunitiesTake	WorkLifeBalar	SelfRating	ManagerRating
PR01	79F7-78EC	01/02/2013	5	4	5	1	0	4	4	4
PR02	B61E-0F26	01/03/2013	5	4	4	1	3	4	4	3
PR03	F5E3-48BB	01/03/2013	3	4	5	3	2	3	5	4
PR04	0678-748A	01/04/2013	5	3	2	2	0	2	3	2
PR05	541F-3E19	01/04/2013	5	2	3	1	0	4	4	3
PR06	F93E-BDEF	01/04/2013	3	3	2	2	0	4	4	4
PR07	9E7A-1F70	01/08/2013	3	4	5	2	1	5	4	3
PR08	05ED-92F1	01/10/2013	4	5	4	1	1	3	3	2
PR09	F72D-261D	01/10/2013	4	5	2	1	1	4	5	4
PR10	774E-685D	01/11/2013	5	4	3	2	3	4	5	4
PR100	B013-7D0C	04/10/2013	4	3	3	2	0	4	3	3
PR1000	528C-3E0D	3/16/2016	4	4	2	2	2	4	5	5
PR1001	D077-169C	3/17/2016	3	5	3	2	2	3	5	5
PR1002	9727-BC84	3/18/2016	4	3	3	2	2	2	4	3
PR1003	DA8E-9496	3/18/2016	3	5	4	1	0	5	5	5
PR1004	DEC5-9319	3/18/2016	3	4	3	2	3	2	4	4
PR1005	88B8-EB84	3/19/2016	3	4	2	3	1	4	5	5
PR1006	9C57-828C	3/19/2016	5	4	2	1	1	2	3	3
PR1007	E1B4-9AA1	3/22/2016	5	4	3	3	2	3	4	3
PR1008	3CD6-5587	3/23/2016	5	4	2	2	0	4	4	3
PR1009	BAFA-86DF	3/23/2016	3	3	4	2	1	2	3	3
PR101	152E-8DB1	04/12/2013	5	2	5	1	0	5	5	4

Analyzing the Data:

DATA CLEANING



Importing Necessary Libraries For Data Cleaning

```python
import pandas  as pd
import plotly.express as px
import numpy as np
from matplotlib import pyplot as plt
%matplotlib inline
import seaborn as sns
pd.options.display.max_columns=None
pd.options.display.max_rows=None
pd.options.display.float_format= '{:,.2f}'.format
```

Recalling Data From Its Previous Format (CSV) And Storing It In A Custom Data Frame For Each File

```python
df_employee=pd.read_csv('Employee.csv')
df_RatingLevel=pd.read_csv('RatingLevel.csv')
df_satisfiedLevel=pd.read_csv('SatisfiedLevel.csv')
df_EducationLevel=pd.read_csv('EducationLevel.csv')
df_PerformanceRating=pd.read_csv('PerformanceRating.csv')
```

Displaying Information About Employees Table

```python
df_employee.info()
```

```
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
 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
...
 21   YearsSinceLastPromotion   1470 non-null   int64
 22   YearsWithCurrManager      1470 non-null   int64
dtypes: int64(9), object(14)
```

(No empty values, but HireDate is object Not DateTime)

Ensuring No Empty Data Or Nulls In Employees Table

```python
df_employee.isnull().sum()
```

```
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
```

No nulls found

Solving All Problems In Employees Table

```python
df_employee=df_employee.drop_duplicates(subset="EmployeeID")
df_employee.HireDate=pd.to_datetime(df_employee.HireDate)
df_employee['Full Name']=df_employee.FirstName+' '+df_employee.LastName
```

Python

```python
df_employee.drop("FirstName",axis=1,inplace=True)
df_employee.drop("LastName",axis=1,inplace=True)
```

Python

- Ensured deleting all duplicates in EmployeeID.
- Converted the hireDate type from object to datetime.
- Collected the first and last name in a table named full name.
- Removed fristName and LastName columns.

Displaying Employees Table After Cleaning

```python
df_employee.info()
```

```
Data columns (total 22 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   EmployeeID                1470 non-null   object
 1   Gender                    1470 non-null   object
 2   Age                       1470 non-null   int64
 3   BusinessTravel            1470 non-null   object
 4   Department                1470 non-null   object
 5   DistanceFromHome (KM)     1470 non-null   int64
 6   State                     1470 non-null   object
 7   Ethnicity                 1470 non-null   object
 8   Education                 1470 non-null   int64
 9   EducationField            1470 non-null   object
 10  JobRole                   1470 non-null   object
 11  MaritalStatus             1470 non-null   object
 12  Salary                    1470 non-null   int64
 13  StockOptionLevel          1470 non-null   int64
 14  OverTime                  1470 non-null   object
 15  HireDate                  1470 non-null   datetime64[ns]
 16  Attrition                 1470 non-null   object
 17  YearsAtCompany            1470 non-null   int64
 18  YearsInMostRecentRole     1470 non-null   int64
 19  YearsSinceLastPromotion   1470 non-null   int64
 20  YearsWithCurrManager      1470 non-null   int64
 21  Full Name                 1470 non-null   object
dtypes: datetime64[ns](1), int64(9), object(12)
memory usage: 252.8+ KB
```

Ensured everything is well.

Looking For Any Outlayers In All Tables

```python
df_employee.describe()
```

Python

	Age	DistanceFromHome (KM)	Education	Salary	StockOptionLevel	HireDate	YearsAtCompany	YearsInM
count	1,470.00	1,470.00	1,470.00	1,470.00	1,470.00	1470	1,470.00	
mean	28.99	22.50	2.91	112,956.50	0.79	2017-07-05 14:50:26.938775296	4.56	
min	18.00	1.00	1.00	20,387.00	0.00	2012-01-03 00:00:00	0.00	
25%	23.00	12.00	2.00	43,580.50	0.00	2014-07-10 06:00:00	2.00	
50%	26.00	22.00	3.00	71,199.50	1.00	2017-09-02 00:00:00	4.00	
75%	34.00	33.00	4.00	142,055.75	1.00	2020-05-14 18:00:00	7.00	
max	51.00	45.00	5.00	547,204.00	3.00	2022-12-31 00:00:00	10.00	
std	7.99	12.81	1.02	103,342.89	0.85	NaN	3.29	

Looking For Any Outlayers In All Tables (Autoclean Library)

```python
from AutoClean import AutoClean
import pandas as pd
resulant = pd.read_csv("./Employee.csv")
pipeline = AutoClean(resulant)
x = pipeline.output
m = resulant.isnull().sum()
print(m)
mr = resulant[resulant.isnull().any(axis=1)]
print(mr)
```

```
13-10-2024 21:31:10.80 - INFO - Started validation of input parameters...
13-10-2024 21:31:10.80 - INFO - Completed validation of input parameters
13-10-2024 21:31:10.80 - INFO - Started handling of duplicates... Method: "AUTO"
13-10-2024 21:31:10.82 - DEBUG - 0 missing values found
13-10-2024 21:31:10.82 - INFO - Completed handling of duplicates in 0.014074 seconds
13-10-2024 21:31:10.82 - INFO - Started handling of missing values...
13-10-2024 21:31:10.82 - DEBUG - 0 missing values found
13-10-2024 21:31:10.82 - INFO - Completed handling of missing values in 0.002094 seconds
13-10-2024 21:31:10.82 - INFO - Started handling of outliers... Method: "WINZ"
13-10-2024 21:31:10.82 - DEBUG - Outlier imputation of 1 value(s) succeeded for feature "Age"
13-10-2024 21:31:11.03 - DEBUG - Outlier imputation of 124 value(s) succeeded for feature "Salary"
13-10-2024 21:31:11.19 - DEBUG - Outlier imputation of 85 value(s) succeeded for feature "StockOptionLevel"
13-10-2024 21:31:11.19 - INFO - Completed handling of outliers in 0.376335 seconds
13-10-2024 21:31:11.19 - INFO - Started conversion of DATETIME features... Granularity: s
13-10-2024 21:31:11.27 - DEBUG - Conversion to DATETIME succeeded for feature "HireDate"
13-10-2024 21:31:11.28 - INFO - Completed conversion of DATETIME features in 0.0769 seconds
13-10-2024 21:31:11.28 - INFO - Started encoding categorical features... Method: "AUTO"
13-10-2024 21:31:11.28 - DEBUG - Encoding to ONEHOT succeeded for feature "Gender"
13-10-2024 21:31:11.28 - DEBUG - Encoding skipped for feature "LastName"
13-10-2024 21:31:11.28 - DEBUG - Encoding skipped for feature "EmployeeID"
13-10-2024 21:31:11.28 - DEBUG - Encoding skipped for feature "FirstName"
13-10-2024 21:31:11.30 - DEBUG - Encoding to ONEHOT succeeded for feature "OverTime"
13-10-2024 21:31:11.30 - DEBUG - Encoding to ONEHOT succeeded for feature "Department"
13-10-2024 21:31:11.30 - DEBUG - Encoding to ONEHOT succeeded for feature "Ethnicity"
13-10-2024 21:31:11.30 - DEBUG - Encoding to ONEHOT succeeded for feature "Attrition"
13-10-2024 21:31:11.30 - DEBUG - Encoding to ONEHOT succeeded for feature "BusinessTravel"
```

Those outlayers had explanation and did not affect the results.

So, they were not deleted.

Ensuring No Empty Data Or Nulls In Performance Rating Table

```python
df_PerformanceRating.isnull().sum()
```

```
PerformanceID                0
EmployeeID                   0
ReviewDate                   0
EnvironmentSatisfaction      0
JobSatisfaction              0
RelationshipSatisfaction     0
TrainingOpportunitiesWithinYear    0
TrainingOpportunitiesTaken   0
WorkLifeBalance              0
SelfRating                   0
ManagerRating                0
dtype: int64
```

No nulls found

Solving All Problems In Performance Rating Table

```python
df_PerformanceRating=df_PerformanceRating.drop_duplicates(subset="PerformanceID")
df_PerformanceRating. ReviewDate=pd.to_datetime(df_PerformanceRating. ReviewDate)
df_PerformanceRating.info()
```

Python

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 6709 entries, 0 to 6708
Data columns (total 11 columns):
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   PerformanceID                   6709 non-null   object
 1   EmployeeID                      6709 non-null   object
 2   ReviewDate                      6709 non-null   datetime64[ns]
 3   EnvironmentSatisfaction         6709 non-null   int64
 4   JobSatisfaction                 6709 non-null   int64
 5   RelationshipSatisfaction        6709 non-null   int64
 6   TrainingOpportunitiesWithinYear 6709 non-null   int64
 7   TrainingOpportunitiesTaken      6709 non-null   int64
 8   WorkLifeBalance                 6709 non-null   int64
 9   SelfRating                      6709 non-null   int64
 10  ManagerRating                   6709 non-null   int64
dtypes: datetime64[ns](1), int64(8), object(2)
memory usage: 576.7+ KB
```

- **Ensured deleting all duplicates in PerformanceID.**
- **Converted the reviewDate type from object to datetime.**

Ensuring No Empty Data Or Nulls In SatisfiedLevel Table And Changing The Columns Names

```python
df_satisfiedLevel.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5 entries, 0 to 4
Data columns (total 2 columns):
 #   Column            Non-Null Count  Dtype 
---  ------            --------------  ----- 
 0   SatisfactionID    5 non-null      int64 
 1   SatisfactionLevel  5 non-null      object
dtypes: int64(1), object(1)
memory usage: 212.0+ bytes
```

No nulls found

```python
df_satisfiedLevel=df_satisfiedLevel.rename(columns={'SatisfactionID':'SatisfactionLevelId'})
df_satisfiedLevel=df_satisfiedLevel.rename(columns={'SatisfactionLevel':'SatisfactionLevelName'})
```

Ensuring No Empty Data Or Nulls In EducationLevel And RatingLevel Table

```python
df_EducationLevel.info()
```

Python

```python
df_RatingLevel.info()
```

Python

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5 entries, 0 to 4
Data columns (total 2 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   EducationLevelID  5 non-null     int64
 1   EducationLevel    5 non-null     object
dtypes: int64(1), object(1)
memory usage: 212.0+ bytes
```

No nulls found

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5 entries, 0 to 4
Data columns (total 2 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   RatingID     5 non-null      int64
 1   RatingLevel  5 non-null      object
dtypes: int64(1), object(1)
memory usage: 212.0+ bytes
```

Show Sample From Dataframe

```python
df_employee.sample()
```

Python

```python
df_employee.describe()
```

Python

	Age	DistanceFromHome (KM)	Education	Salary	StockOptionLevel	HireDate	YearsAtCompany	YearsIn
count	1,470.00	1,470.00	1,470.00	1,470.00	1,470.00	1470	1,470.00	
mean	28.99	22.50	2.91	112,956.50	0.79	2017-07-05 14:50:26.938775296	4.56	
min	18.00	1.00	1.00	20,387.00	0.00	2012-01-03 00:00:00	0.00	
25%	23.00	12.00	2.00	43,580.50	0.00	2014-07-10 06:00:00	2.00	
50%	26.00	22.00	3.00	71,199.50	1.00	2017-09-02 00:00:00	4.00	
75%	34.00	33.00	4.00	142,055.75	1.00	2020-05-14 18:00:00	7.00	
max	51.00	45.00	5.00	547,204.00	3.00	2022-12-31 00:00:00	10.00	
std	7.99	12.81	1.02	103,342.89	0.85	NaN	3.29	

DATA QUALITY

```python
import pandas as pd
# Merge the two datasets on EmployeeID if not already merged
merged_data = df_employee.merge(df_PerformanceRating, on='EmployeeID')

# Convert the 'HireDate' and 'ReviewDate' to datetime format
merged_data['HireDate'] = pd.to_datetime(merged_data['HireDate'])
merged_data['ReviewDate'] = pd.to_datetime(merged_data['ReviewDate'])

# Calculate the number of employees where ReviewDate is after HireDate
employees_review_after_hire = merged_data[merged_data['ReviewDate'] < merged_data['HireDate']]
distinct_count = employees_review_after_hire['EmployeeID'].nunique()
# Count the number of such employees
num_employees_review_after_hire = distinct_count

print(f"Number of employees where the review date is after the hire date: {num_employees_review_after_hire}")
```

✓ 0.1s

```
Number of employees where the review date is after the hire date: 289
```

Some employees were found to have review dates in dates before their hiring dates ------ Did No Change

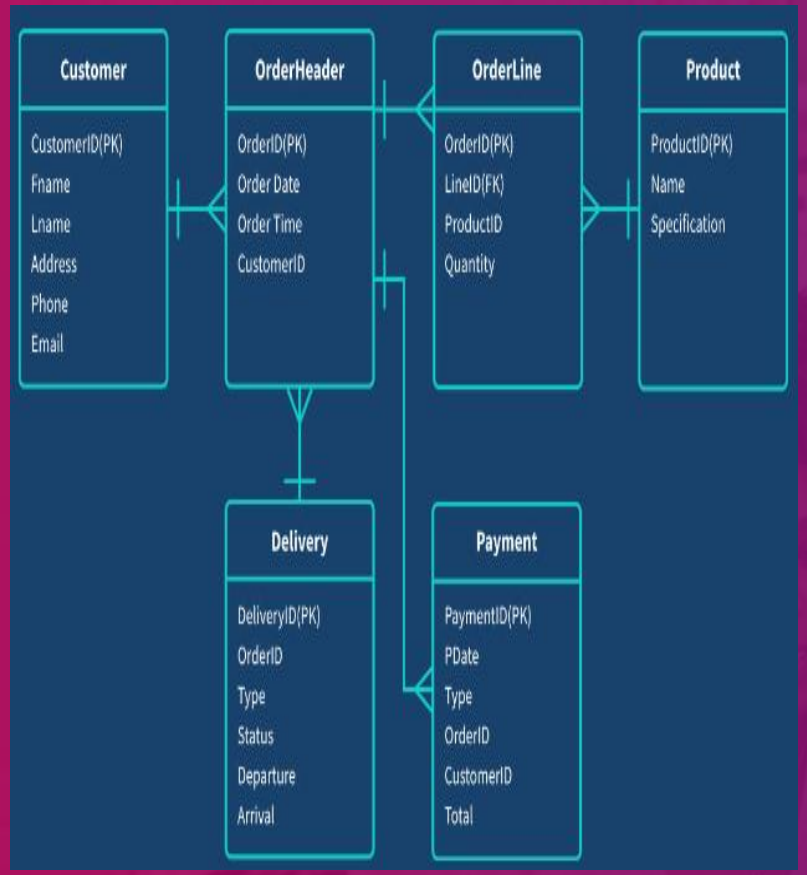
```
education_values = df_employee['EducationField'].unique()

# Display the unique values
print("Unique values in the 'Education' field:")
for value in education_values:
    print(value)
```

✓ 0.0s

```
Unique values in the 'Education' field:
Marketing
Marketing
Computer Science
Technical Degree
Information Systems
Other
Economics
Human Resources
Business Studies
```

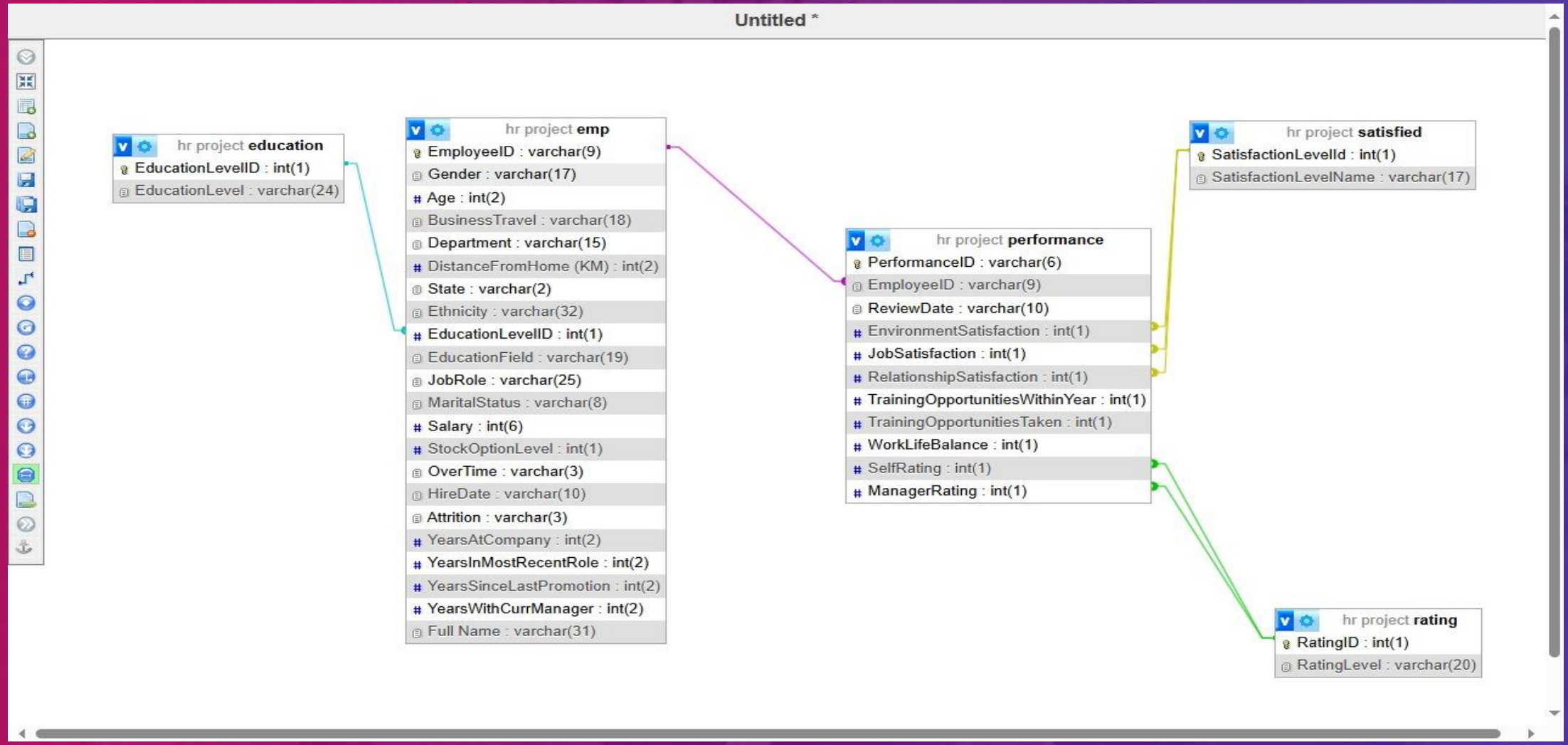
- **Two education value fields were found to be marketing without any significant difference ----- Did no change.**
- **The Gender column in Employee Table had 2 items beside male and female and both of them could affect the total no. of males and females if known---- No change was done.**
- **It was unethical to put a column for Ethnicity as it was of no value for analytical purposes.**
- **There were many outlayers between employees' salaries-------- No change was done----- as we suggested it may be due to difference in levels of employees, as between general manager and senior staff and normal employees**
- **More sources of data were needed for better understanding, analysing and consequent decision making.**

DATA MODELING

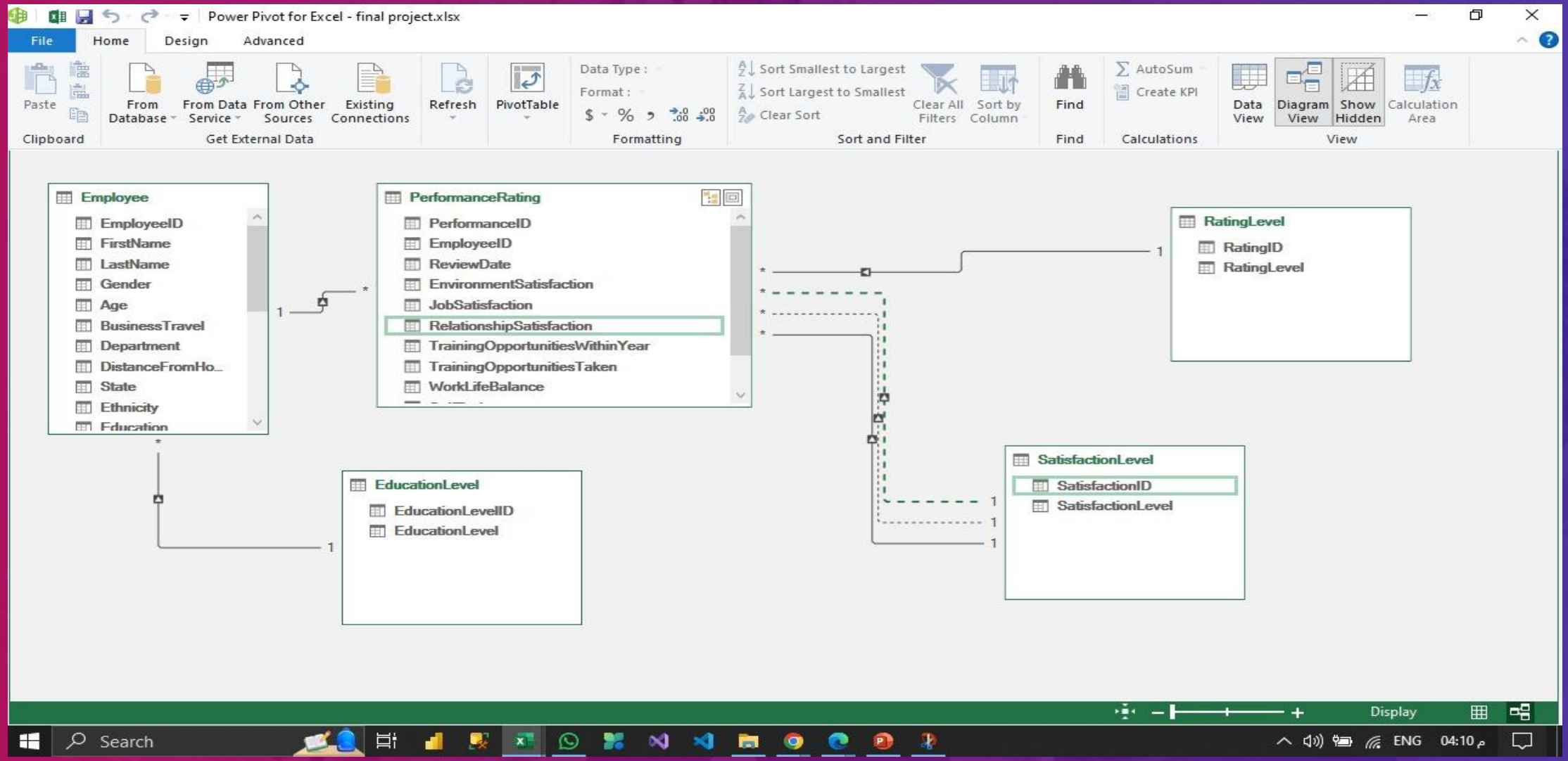


- Data modelling is the process of creating a diagram that represents your data system and defines the structure, attributes, and relationships of your data entities.
- Data modelling organizes and simplifies data in a way that makes it easy to understand, manage, and query, while also ensuring data integrity and consistency. Data models inform about data architecture, database design, and restructuring legacy systems.

Data Modeling Using SQL



Data Modeling Using Excel



Determining Data Analysis Questions and Answering Them

1. **How many employees are in the company?**

```
SELECT COUNT(*) AS "Total Employees" FROM employee;
```

Extra options

Total Employees
1470

2. **What is the average age of employees?**

```
SELECT ROUND(AVG(employee.Age),0) AS "Avarge of age " FROM employee;
```

Profiling [Edit inline] [Edit] [Explain SQL] [Create PHP code] [Refresh

Show all | Number of rows: 25

Extra options

Avarge of age
29

3. **What is the average salary of employees by department?**

```
SELECT Department, AVG(Salary) AS AvgSalary FROM Employee GROUP BY Department;
```

Profiling [Ed

Show all | Number of rows: 25 Filter rows: Search this table

Extra options

	Department	Avg Salary
Edit Copy Delete	Human Resources	119698.8095
Edit Copy Delete	Sales	119117.6099
Edit Copy Delete	Technology	109655.1228

4. What is the gender distribution among employees?

```
SELECT Gender, COUNT(*) AS TotalEmployees FROM Employee GROUP BY Gender;
```

Profili

Show all | Number of rows: 25 Filter rows: Search this table

Extra options

Gender	TotalEmployees
Female	675
Male	651
Non-Binary	124
Prefer Not To Say	20

5. How many employees have been with the company for more than 5 years?

```
SELECT COUNT(*) AS EmployeesOver5Years FROM Employee WHERE YearsAtCompany > 5;
```

Profiling [Edit

Extra options

EmployeesOver5Years
587

6. What is the distribution of employees by job title?

```
SELECT JobRole, COUNT(*) AS TotalEmployees FROM Employee GROUP BY JobRole;
```

Profiling [Edit inline] [Edit] [Explain SQL] [Create PHP code] [Refresh]

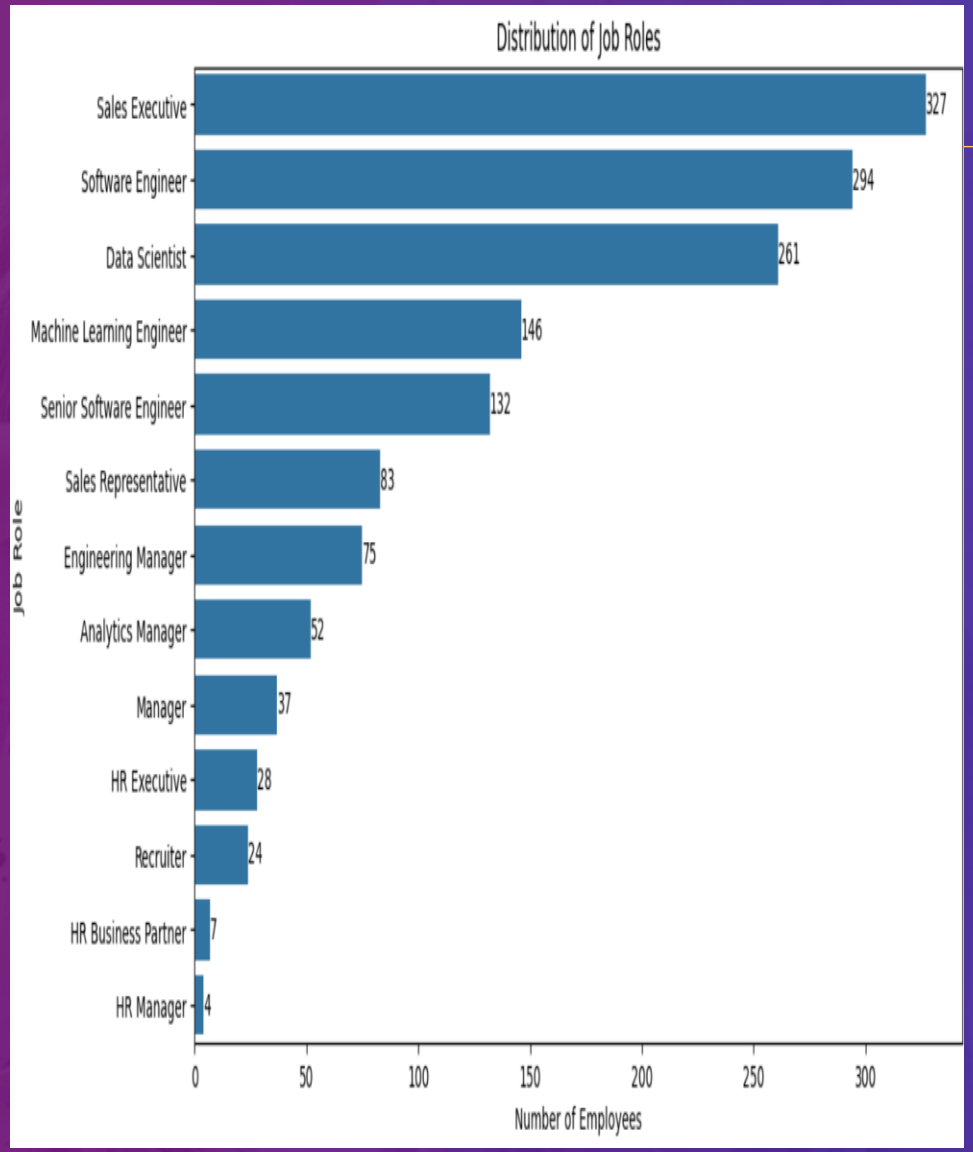
Show all | Number of rows: 25 Filter rows: Search this table

Extra options

JobRole	TotalEmployees
Analytics Manager	52
Data Scientist	261
Engineering Manager	75
HR Business Partner	7
HR Executive	28
HR Manager	4
Machine Learning Engineer	146
Manager	37
Recruiter	24
Sales Executive	327
Sales Representative	83
Senior Software Engineer	132
Software Engineer	294

```python
import matplotlib.pyplot as plt
import seaborn as sns

# Your existing code
plt.figure(figsize=(12, 6))
ax = sns.countplot(data=df_employee, y='JobRole', order=df_employee['JobRole'].value_counts().index)
plt.title('Distribution of Job Roles')
plt.xlabel('Number of Employees')
plt.ylabel('Job Role')

# Adding the numbers at the end of each bar
for container in ax.containers:
    ax.bar_label(container, fmt='%d')  # '%d' for integer formatting

plt.show()
```

Python



7. What is the distribution of employees by age group?

```
SELECT CASE WHEN Age BETWEEN 20 AND 29 THEN '20-29' WHEN Age BETWEEN 30 AND 39 THEN '30-39' WHEN Age BETWEEN 40 AND 49 THEN '40-49' ELSE '50+'
END AS AgeGroup, COUNT(*) AS TotalEmployees FROM Employee GROUP BY CASE WHEN Age BETWEEN 20 AND 29 THEN '20-29' WHEN Age BETWEEN 30 AND 39 THEN
'30-39' WHEN Age BETWEEN 40 AND 49 THEN '40-49' ELSE '50+' END;
```

Profiling [Edit inline] [Edit] [Explain SQL] [Create PHP code] [Refresh]

Show all | Number of rows: 25 Filter rows: Search this table

Extra options

AgeGroup	TotalEmployees
20-29	874
30-39	289
40-49	219
50+	68

8. What is the number of employees in each education level?

```
SELECT education.EducationLevel,COUNT(emp.EmployeeID)AS "Number of Employee"
FROM emp INNER JOIN education ON emp.EducationLevelID =
education.EducationLevelID GROUP BY education.EducationLevel;
```

Profiling [Edit inline] [Edit] [Explain SQL] [Create PHP code] [Refresh]

Show all | Number of rows: 25

Extra options

EducationLevel	Number of Employee
Bachelors	572
Doctorate	48
High School	282
Masters	398
No Formal Qualifications	170

9. What is the distribution of education levels across departments?

```
SELECT employee.Department,educationlevel.EducationLevel,COUNT(employee.EmployeeID)AS "number of emp" FROM employee INNER JOIN educationlevel on
employee.Education=educationlevel.EducationLevelID GROUP by employee.Department,educationlevel.EducationLevel;
```

Profiling [Edit inline] [Edit] [Explain SQL] [Create PHP code] [Refresh]

Show all | Number of rows: 25 Filter rows: Search this table

Extra options

Department	EducationLevel	number of emp
Human Resources	Bachelors	27
Human Resources	Doctorate	3
Human Resources	High School	13
Human Resources	Masters	15
Human Resources	No Formal Qualifications	5
Sales	Bachelors	166
Sales	Doctorate	15
Sales	High School	87
Sales	Masters	128
Sales	No Formal Qualifications	50
Technology	Bachelors	379
Technology	Doctorate	30
Technology	High School	182
Technology	Masters	256
Technology	No Formal Qualifications	115

10. What is the average performance rating of employees based on their education level?

```
SELECT educationlevel.EducationLevel, AVG(performancerating.ManagerRating) AS AvgManagerRating,
AVG(performancerating.SelfRating) AS AvgSelfRating FROM performancerating JOIN employee ON
performancerating.EmployeeID = employee.EmployeeID JOIN educationlevel ON employee.Education =
educationlevel.EducationLevelID GROUP BY educationlevel.EducationLevel;
```

Profiling [Edit inline] [Edit] [Explain SQL] [Create PHP

Show all | Number of rows: 25 Filter rows: Search this table

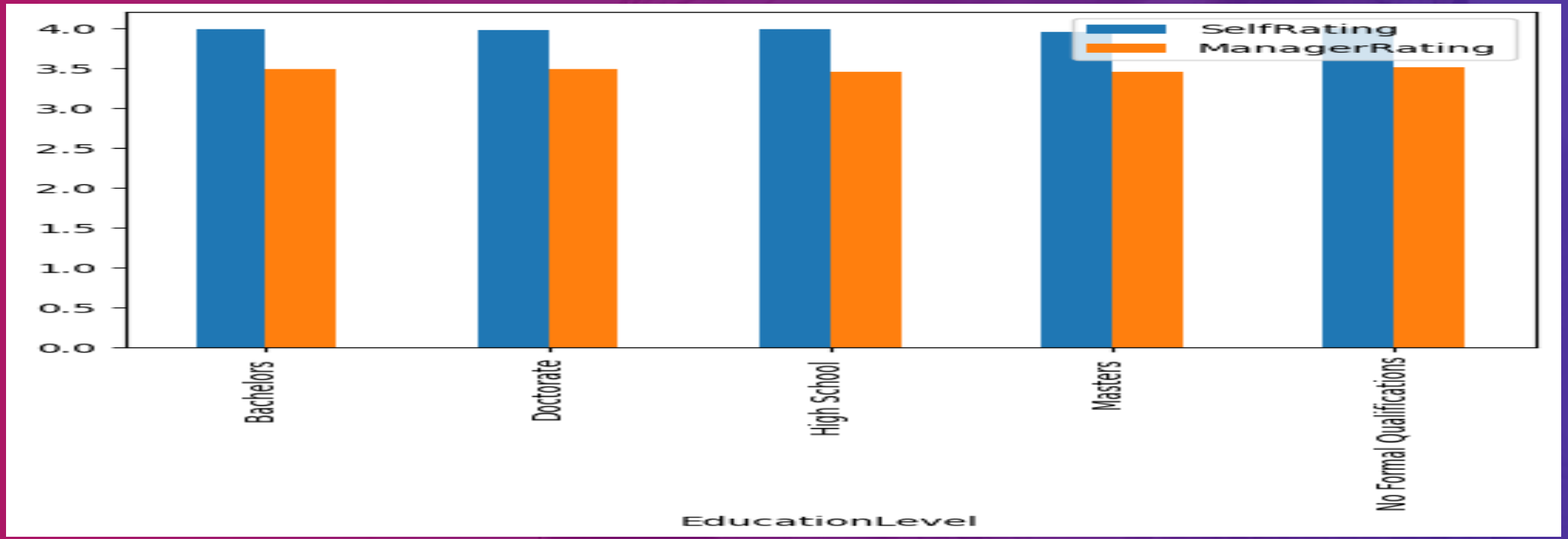
Extra options

EducationLevel	AvgManagerRating	AvgSelfRating
Bachelors	3.4833	3.9924
Doctorate	3.4882	3.9716
High School	3.4567	3.9904
Masters	3.4531	3.9582
No Formal Qualifications	3.5060	4.0024

```python
merged_df = df_PerformanceRating.merge(df_employee, on='EmployeeID').merge(df_EducationLevel, on='EducationLevelID')
education_performance = merged_df.groupby('EducationLevel')[['SelfRating', 'ManagerRating']].mean().reset_index()
education_performance.plot(x='EducationLevel', kind='bar')
```

Python

```
<Axes: xlabel='EducationLevel'>
```



11. in Each Job role what is the number of Employee with attrition

```
SELECT employee.JobRole,COUNT(employee.EmployeeID) AS "number of Emp" FROM employee
WHERE employee.Attrition="Yes" GROUP by employee.JobRole ORDER by `number of Emp`;
```

Profiling [Edit inline] [Edit] [Explain SQL] [Create PHP code] [R

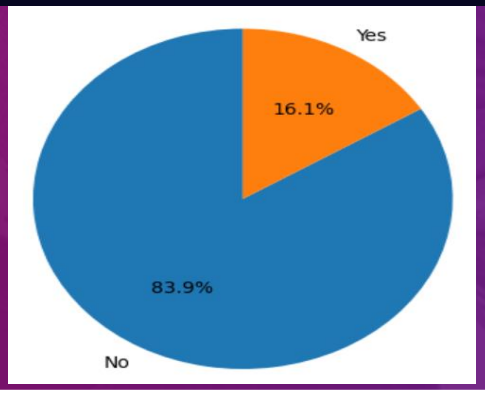
Show all | Number of rows: 25 Filter rows: Search this table

Extra options

JobRole	number of Emp ▲ 1
Engineering Manager	2
Manager	2
Analytics Manager	3
HR Executive	3
Recruiter	9
Senior Software Engineer	9
Machine Learning Engineer	10
Sales Representative	33
Software Engineer	47
Sales Executive	57
Data Scientist	62

What is the percentage of attrition among employees?

```python
Attrition_counts = df_employee['Attrition'].value_counts()
plt.pie(x=Attrition_counts.values, labels=Attrition_counts.index, autopct='%1.1f%%', startangle=90)
plt.show()
```



12. What is the avg salary grouped by the education field?

```
SELECT EducationField,SUM(Salary)AS "sum of salary" FROM employee GROUP BY
EducationField ORDER BY `sum of salary` DESC;
```

Profiling [Edit inline] [Edit] [Explain SQL] [Create PHP cod

Show all | Number of rows: 25 Filter rows: Search this table

Extra options

	EducationField	sum of salary
Edit Copy Delete	Computer Science	48115757
Edit Copy Delete	Information Systems	41520135
Edit Copy Delete	Marketing	40390223
Edit Copy Delete	Economics	11334205
Edit Copy Delete	Business Studies	9250227
Edit Copy Delete	Other	7900032
Edit Copy Delete	Human Resources	3930278
Edit Copy Delete	Technical Degree	3605195

13. What is the distribution of employees' education levels by gender?

```
SELECT educationlevel.EducationLevel,Gender,COUNT(employee.EmployeeID) AS 'number of emp 'FROM employee INNER JOIN educationlevel on employee.Education=educationlevel.EducationLevelID GROUP BY
educationlevel.EducationLevel,Gender;
```

Profiling [Edit inline] [Edit] [Explain SQL] [Create PHP code] [Refresh]

Show all | Number of rows: 25 Filter rows: Search this table

Extra options

EducationLevel	Gender	number of emp
Bachelors	Female	269
Bachelors	Male	254
Bachelors	Non-Binary	43
Bachelors	Prefer Not To Say	6
Doctorate	Female	21
Doctorate	Male	23
Doctorate	Non-Binary	4
High School	Female	136
High School	Male	119
High School	Non-Binary	23
High School	Prefer Not To Say	4
Masters	Female	164
Masters	Male	190
Masters	Non-Binary	37
Masters	Prefer Not To Say	7
No Formal Qualifications	Female	85
No Formal Qualifications	Male	65
No Formal Qualifications	Non-Binary	17
No Formal Qualifications	Prefer Not To Say	3

14. Which department has the highest average manager performance rating?

```
SELECT Department, AVG(ManagerRating) AS AvgPerformanceRating FROM employee e JOIN
performancerating p ON e.EmployeeID = p.EmployeeID GROUP BY Department ORDER BY
AvgPerformanceRating DESC;
```

Profiling [Edit inline] [Edit] [Explain SQL] [Create PHP code] [R

Show all | Number of rows: 25 Filter rows: Search this table

Extra options

Department	AvgPerformanceRating ▾ 1
Technology	3.4874
Sales	3.4500
Human Resources	3.4422

15. Is there a significant difference in performance ratings between Genders of employees?

```
SELECT e.Gender, AVG(p.ManagerRating) AS AvgPerformanceRating FROM Employee e JOIN
PerformanceRating p ON e.EmployeeID = p.EmployeeID GROUP BY e.Gender;
```

Profiling [Edit inline] [Edit] [Explain SQL] [Create PHP code] [R

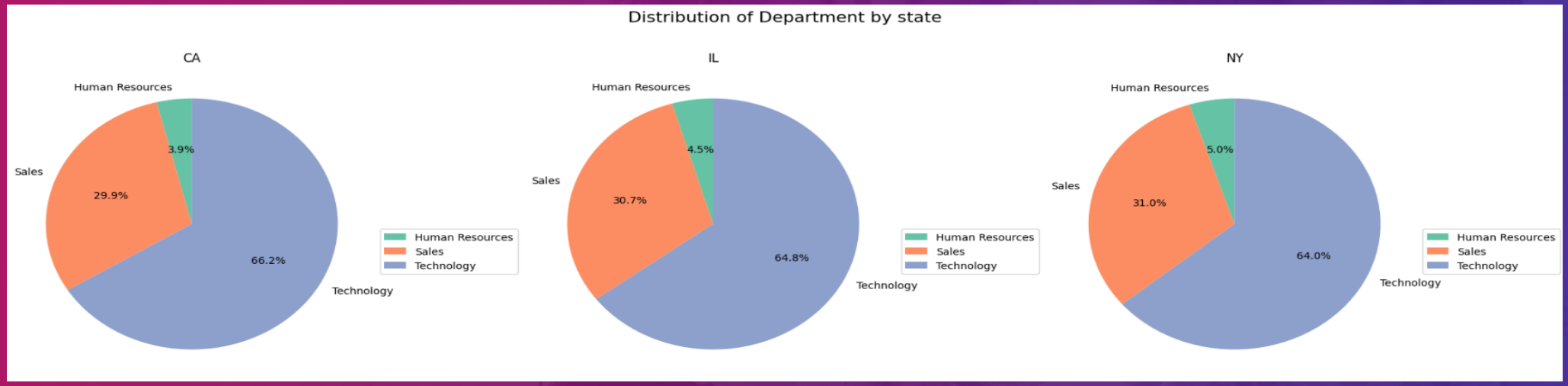
Show all | Number of rows: 25 Filter rows: Search this table

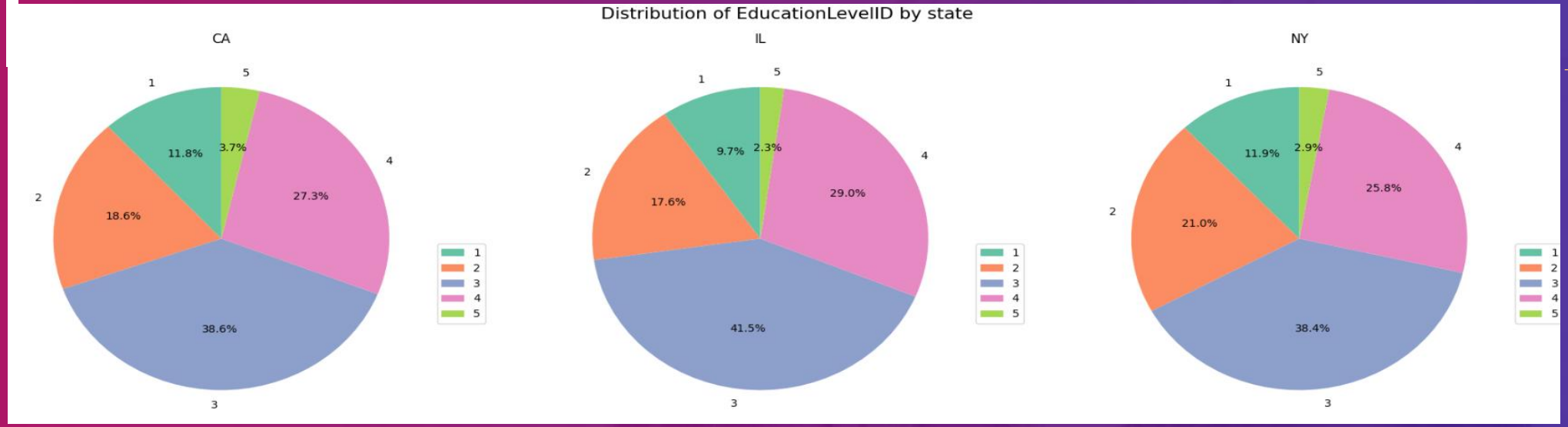
Extra options

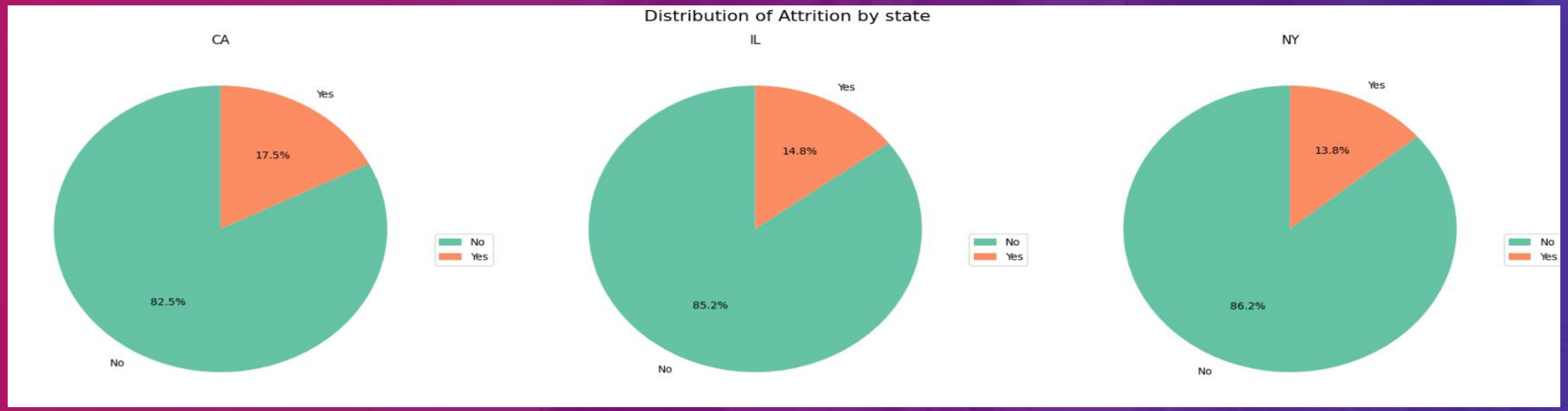
Gender	AvgPerformanceRating
Female	3.4866
Male	3.4609
Non-Binary	3.4822
Prefer Not To Say	3.3279

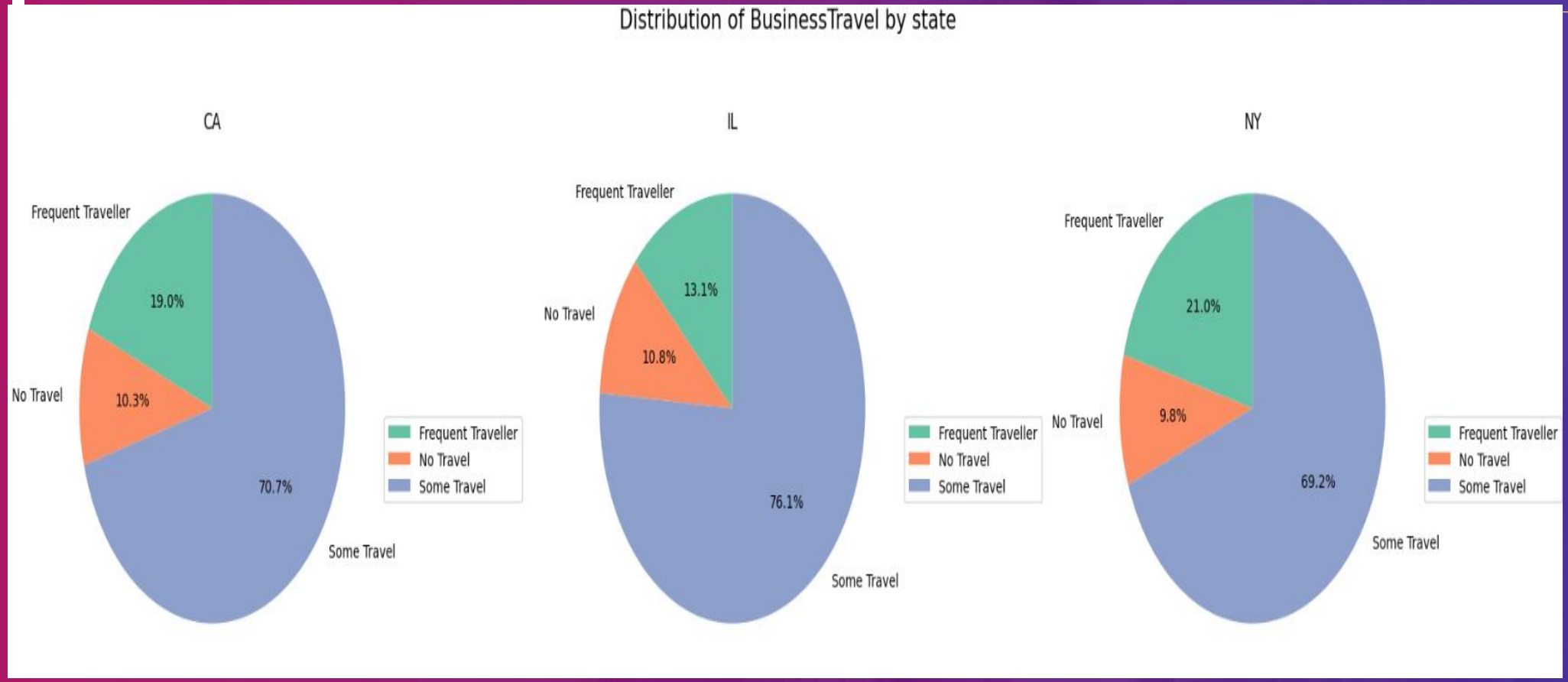
What Is The Effect Of The State On Other Parameters?

```
def plot_pie_chart_by_State(feature):
    colors = ['#66c2a5','#fc8d62','#8da0cb','#e78ac3','#a6d854']
    fig, axes = plt.subplots(1, 3, figsize=(20, 6), subplot_kw=dict(aspect="equal"))
    grouped_data = df_employee.groupby([feature, 'State']).size().unstack().T
    for ax, (gender, data) in zip(axes, grouped_data.iterrows()):
        data.plot(kind='pie', ax=ax, autopct='%1.1f%%', startangle=90, colors=colors)
        ax.set_title(gender)
        ax.set_ylabel('')  # Hide the y-label
        ax.legend(loc='best', bbox_to_anchor=(1, 0.5))
    plt.suptitle(f'Distribution of {feature} by state', fontsize=16)
    plt.tight_layout()
    plt.show()

for feature in ['Department', 'EducationLevelID', 'Attrition', 'BusinessTravel']:
    plot_pie_chart_by_State(feature)
```




Distribution of EducationLevelID by state
CA
IL
NY
1
2
3
4
5
11.8%
18.6%
38.6%
27.3%
3.7%
9.7%
17.6%
41.5%
29.0%
2.3%
11.9%
21.0%
38.4%
25.8%
2.9%

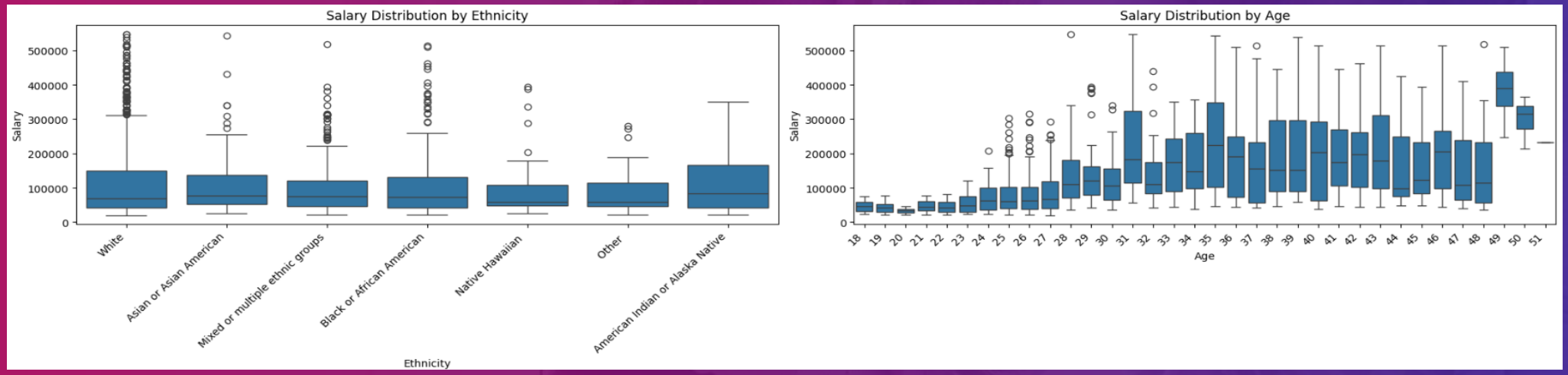

Distribution of Attrition by state
CA
IL
NY
No
Yes
82.5%
17.5%
85.2%
14.8%
86.2%
13.8%

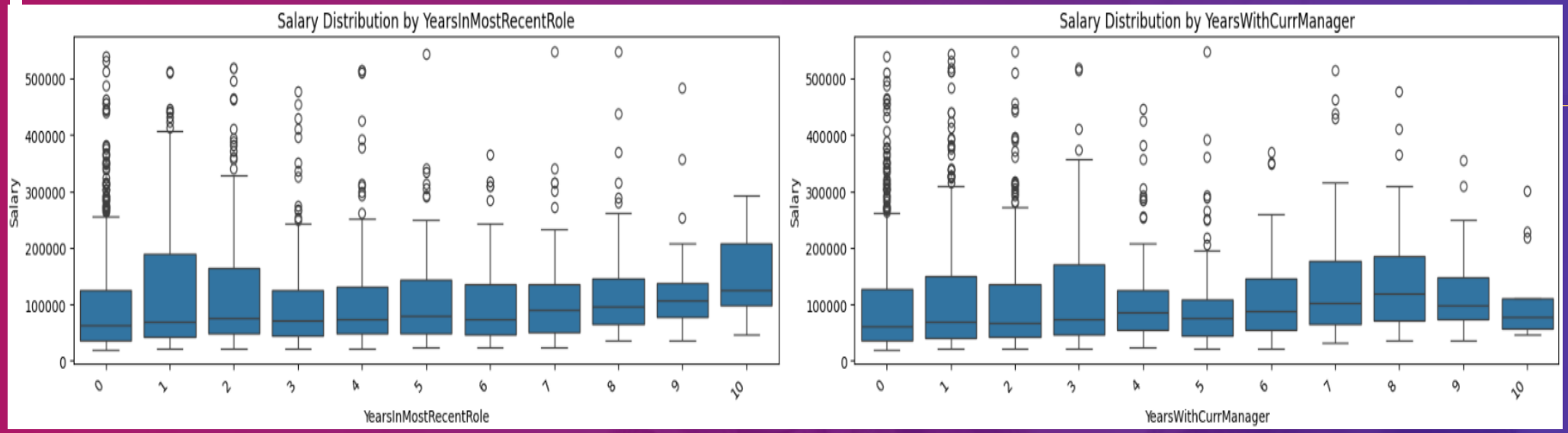

Distribution of BusinessTravel by state
CA
Frequent Traveller
19.0%
No Travel
10.3%
70.7%
Some Travel
Frequent Traveller
No Travel
Some Travel
IL
Frequent Traveller
13.1%
No Travel
10.8%
76.1%
Some Travel
Frequent Traveller
No Travel
Some Travel
NY
Frequent Traveller
21.0%
No Travel
9.8%
69.2%
Some Travel
Frequent Traveller
No Travel
Some Travel

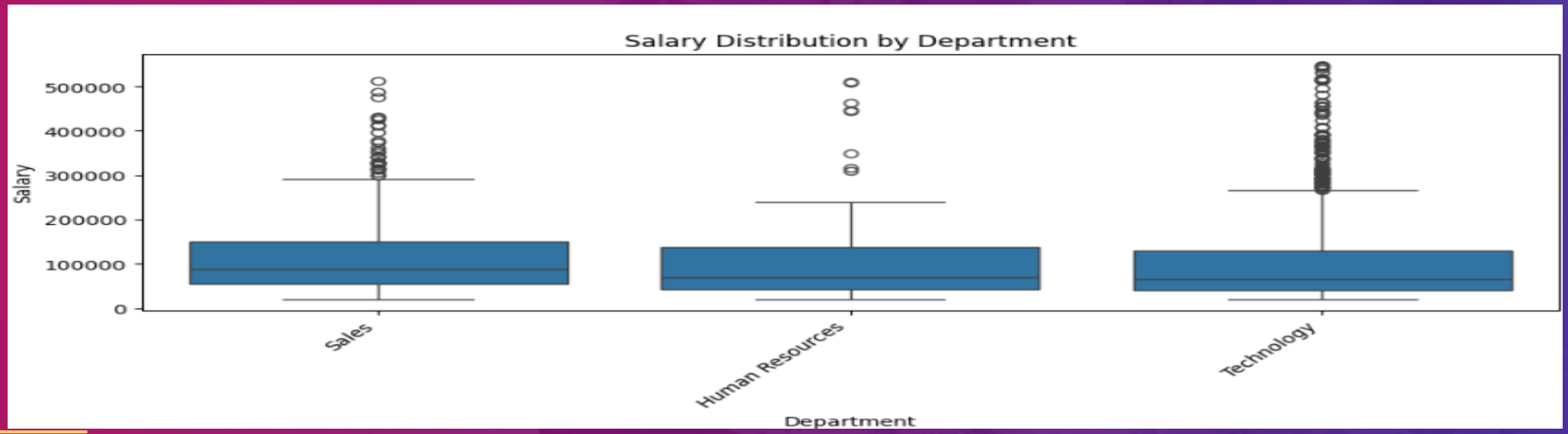
What Affects Employees Salaries?

```python
features_to_compare = ['Ethnicity', 'Age', 'YearsInMostRecentRole', 'YearsWithCurrManager', 'Department']
plt.figure(figsize=(20, 15))

for i, feature in enumerate(features_to_compare, 1):
    plt.subplot(3, 2, i)  # Adjusted to 3 rows, 2 columns to fit 5 plots
    sns.boxplot(x=feature, y='Salary', data=df_employee)
    plt.title(f'Salary Distribution by {feature}')
    plt.xticks(rotation=45, ha='right')
    plt.xlabel(feature)
    plt.ylabel('Salary')

plt.tight_layout()
plt.show()
```

Python




Salary Distribution by YearsInMostRecentRole
Salary
YearsInMostRecentRole
Salary Distribution by YearsWithCurrManager
Salary
YearsWithCurrManager


Salary Distribution by Department
Salary
Sales
Human Resources
Technology
Department

12. What is the avg salary grouped by the education field?

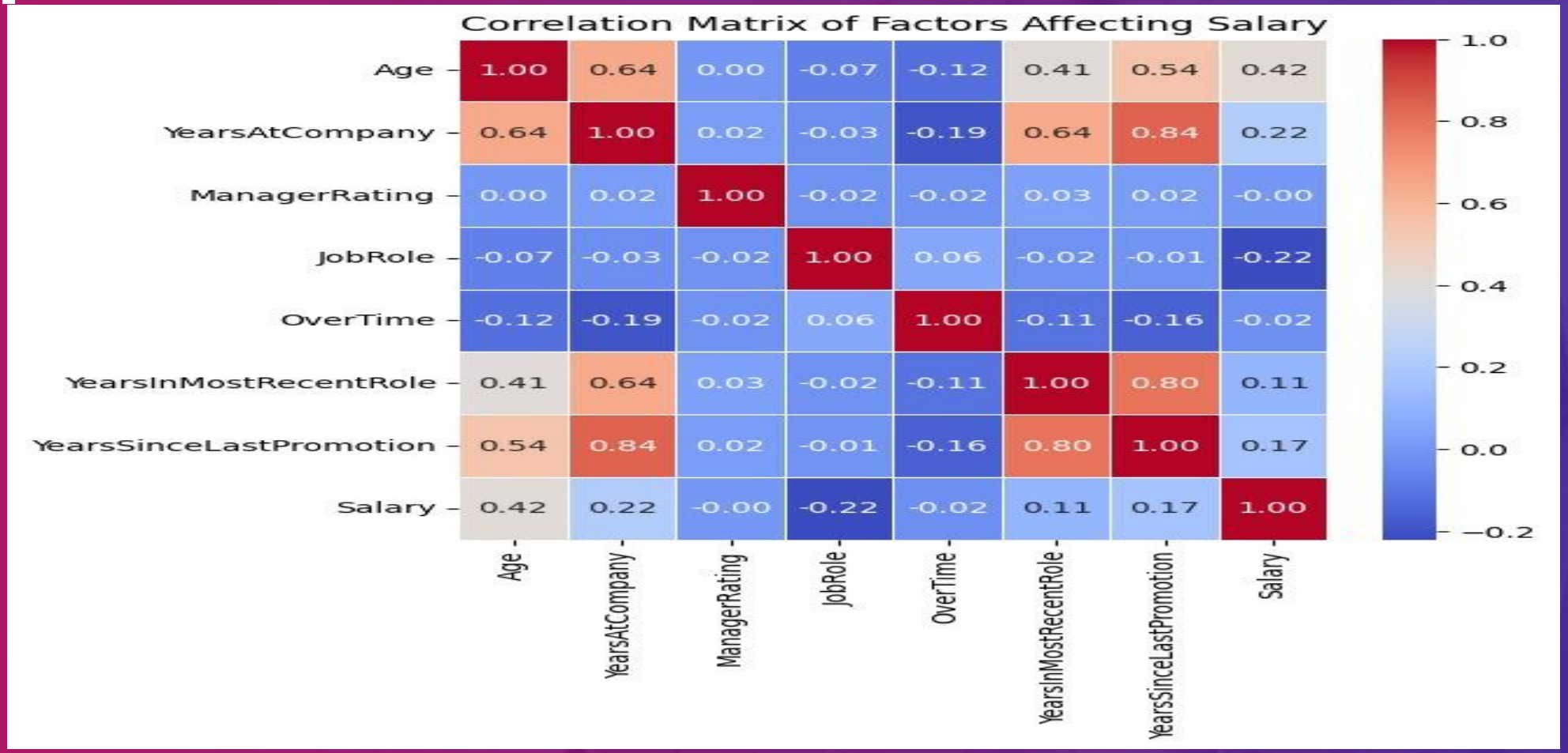
```
SELECT emp.EducationField,AVG(emp.Salary) AS"Avg of salary" FROM emp GROUP BY
emp.EducationField ORDER BY `Avg of salary` DESC;
```

Profiling [Edit inline] [Edit] [Explain SQL] [Create PHP code] [Refresh

Show all | Number of rows: 25

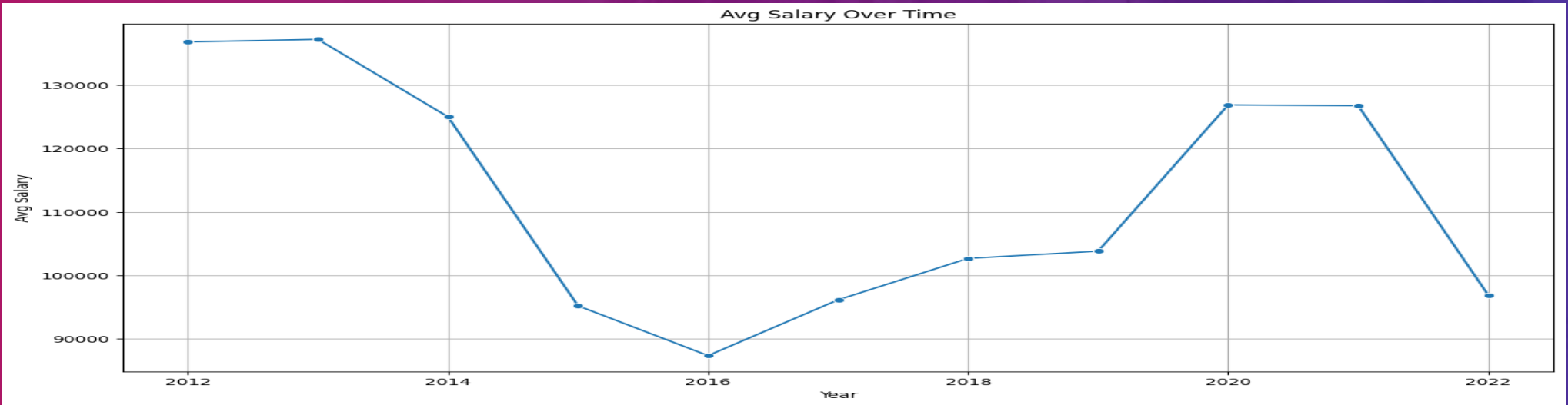
Extra options

	EducationField	Avg of salary
Edit Copy Delete	Human Resources	145565.8519
Edit Copy Delete	Marketing	124277.6092
Edit Copy Delete	Information Systems	114380.5372
Edit Copy Delete	Economics	112219.8515
Edit Copy Delete	Computer Science	109353.9932
Edit Copy Delete	Business Studies	98406.6702
Edit Copy Delete	Other	96341.8537
Edit Copy Delete	Technical Degree	94873.5526


Correlation Matrix of Factors Affecting Salary
Age
YearsAtCompany
ManagerRating
JobRole
OverTime
YearsInMostRecentRole
YearsSinceLastPromotion
Salary
1.00 0.64 0.00 -0.07 -0.12 0.41 0.54 0.42
0.64 1.00 0.02 -0.03 -0.19 0.64 0.84 0.22
0.00 0.02 1.00 -0.02 -0.02 0.03 0.02 -0.00
-0.07 -0.03 -0.02 1.00 0.06 -0.02 -0.01 -0.22
-0.12 -0.19 -0.02 0.06 1.00 -0.11 -0.16 -0.02
0.41 0.64 0.03 -0.02 -0.11 1.00 0.80 0.11
0.54 0.84 0.02 -0.01 -0.16 0.80 1.00 0.17
0.42 0.22 -0.00 -0.22 -0.02 0.11 0.17 1.00
1.0
0.8
0.6
0.4
0.2
0.0
−0.2

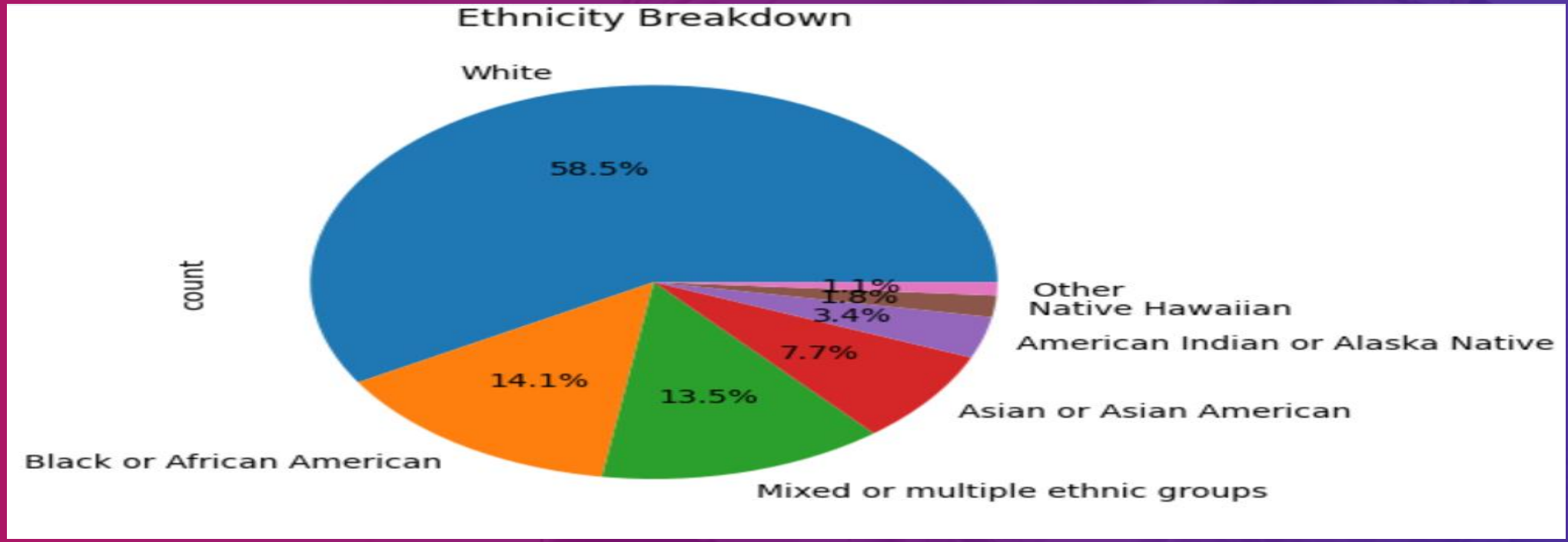
What Is The Average Of Salaries Overtime?

```
salary_over_time = df_employee.groupby(df_employee['HireDate'].dt.year).agg({
    'Salary': 'mean'
}).reset_index()
plt.figure(figsize=(14, 7))
sns.lineplot(x='HireDate', y='Salary', data=salary_over_time, marker='o')
plt.title('Avg Salary Over Time')
plt.xlabel('Year')
plt.ylabel('Avg Salary')
plt.grid(True)
plt.show()
```



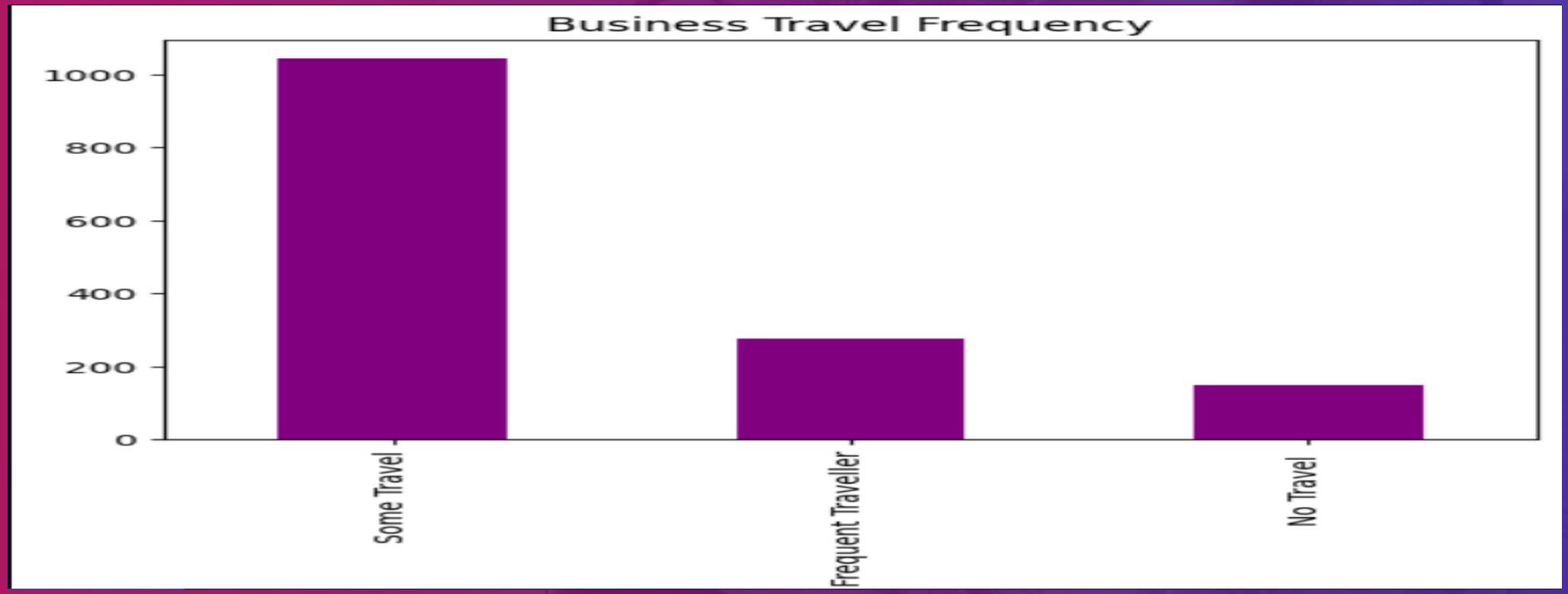
How Are Employees Distributed According To Ethnicity?

```python
df_employee['Ethnicity'].value_counts().plot(kind='pie', autopct='%1.1f%%', title='Ethnicity Breakdown')
plt.show()
```



Which Departments Have More Business Travels

```python
df_employee['BusinessTravel'].value_counts().plot(kind='bar', title='Business Travel Frequency', color='purple');
```



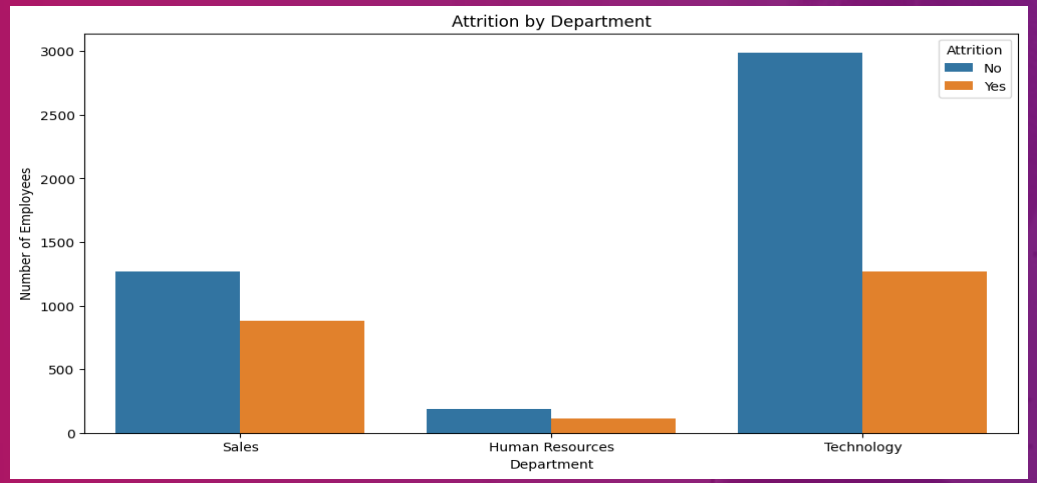
What Are The Main Causes Of Employees Attrition

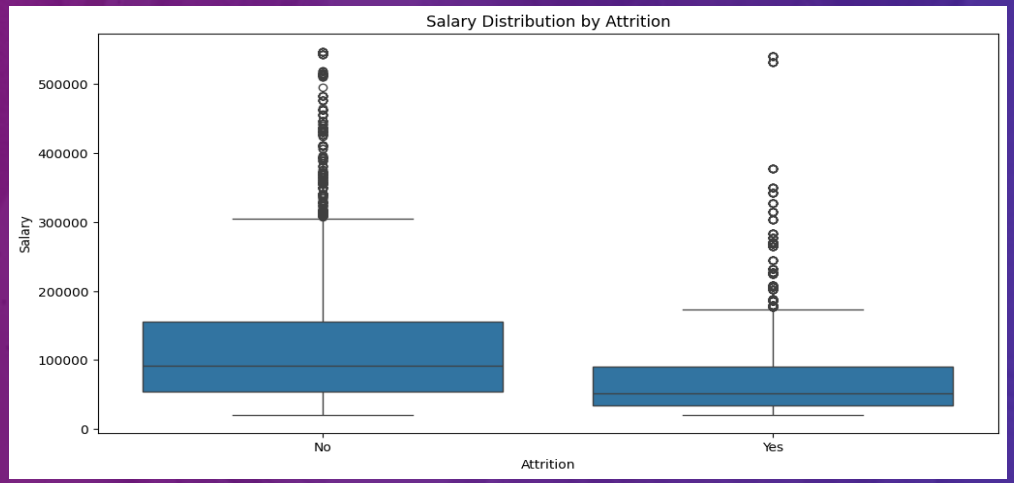
```python
plt.figure(figsize=(12, 6))
sns.boxplot(data=merged_data, x='Attrition', y='Salary')
plt.title('Salary Distribution by Attrition')
plt.xlabel('Attrition')
plt.ylabel('Salary')
plt.show()
```

Python

```python
plt.figure(figsize=(12, 6))
sns.countplot(data= merged_data, x='Department', hue='Attrition')
plt.title('Attrition by Department')
plt.xlabel('Department')
plt.ylabel('Number of Employees')
plt.legend(title='Attrition', loc='upper right')
plt.show()
```

Python



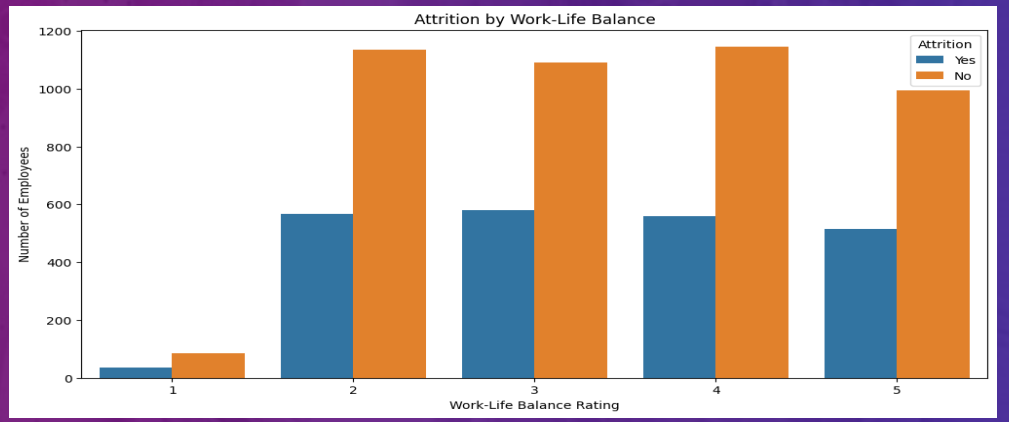


```python
# Plotting training opportunities vs attrition
plt.figure(figsize=(12, 6))
sns.countplot(data=merged_data, x='TrainingOpportunitiesTaken', hue='Attrition')
plt.title('Attrition by Training Opportunities Taken')
plt.xlabel('Training Opportunities Taken')
plt.ylabel('Number of Employees')
plt.legend(title='Attrition', loc='upper right')
plt.show()
```

Python

```python
plt.figure(figsize=(12, 6))
sns.countplot(data=merged_data, x='WorkLifeBalance', hue='Attrition')
plt.title('Attrition by Work-Life Balance')
plt.xlabel('Work-Life Balance Rating')
plt.ylabel('Number of Employees')
plt.legend(title='Attrition', loc='upper right')
plt.show()
```

Python



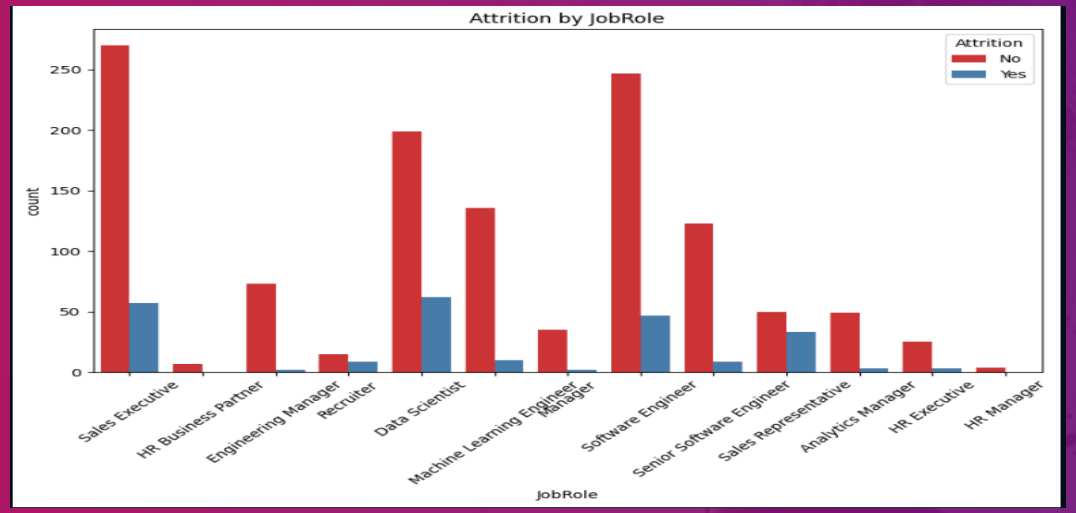
What Is the effect of JobRole & Age On Attrition?

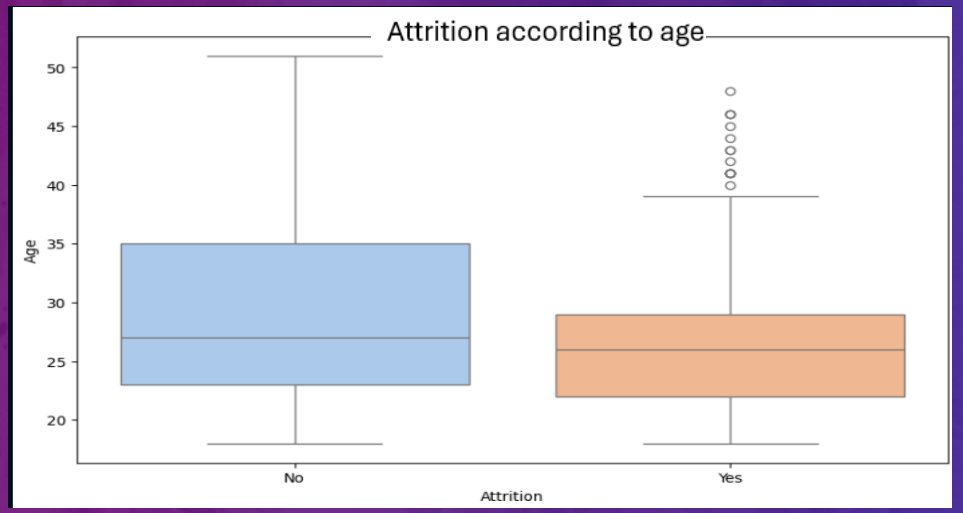
```python
plt.figure(figsize=(10, 6))
sns.countplot(x='JobRole', hue='Attrition', data=df_employee, palette='Set1')
plt.title('Attrition by JobRole')
plt.xticks(rotation=45)
plt.show()
```

Python

```python
plt.figure(figsize=(10, 6))
sns.boxplot(x='Attrition', y='Age', data=df_employee, palette='pastel')
plt.title('Attrition according to age')
plt.show()
```

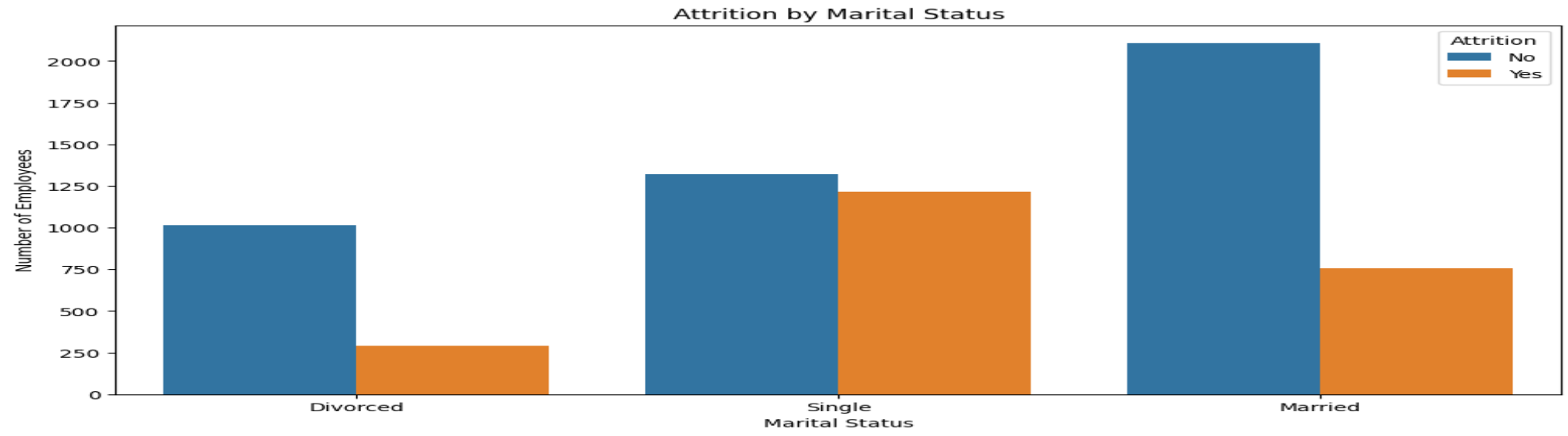
Python





What Is The Effect Of Marital Status On Attrition?

```python
# Plotting marital status vs. attrition
plt.figure(figsize=(12, 6))
sns.countplot(data=merged_data, x='MaritalStatus', hue='Attrition')
plt.title('Attrition by Marital Status')
plt.xlabel('Marital Status')
plt.ylabel('Number of Employees')
plt.legend(title='Attrition', loc='upper right')
plt.show()
```



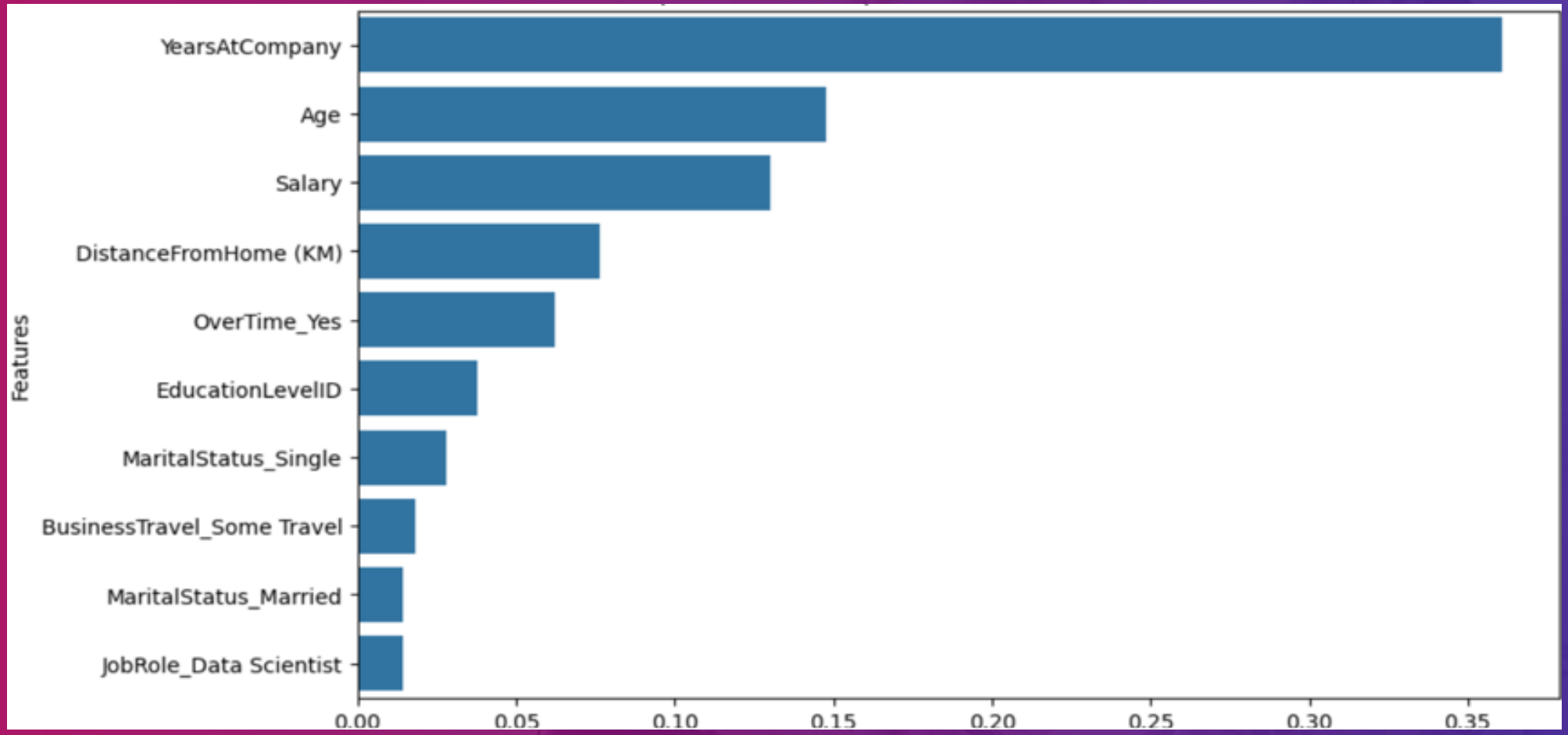
```python
average_age_leavers = merged_data[merged_data['Attrition'] == 'Yes']['Age'].mean()
print(f"Average age of employees who leave: {average_age_leavers:.2f} years")
```

Python

```
Average age of employees who leave: 26.63 years
```

Comment:

The Main Causes of Attrition are:



What Is The Percentage Of Employees With Stock Options?

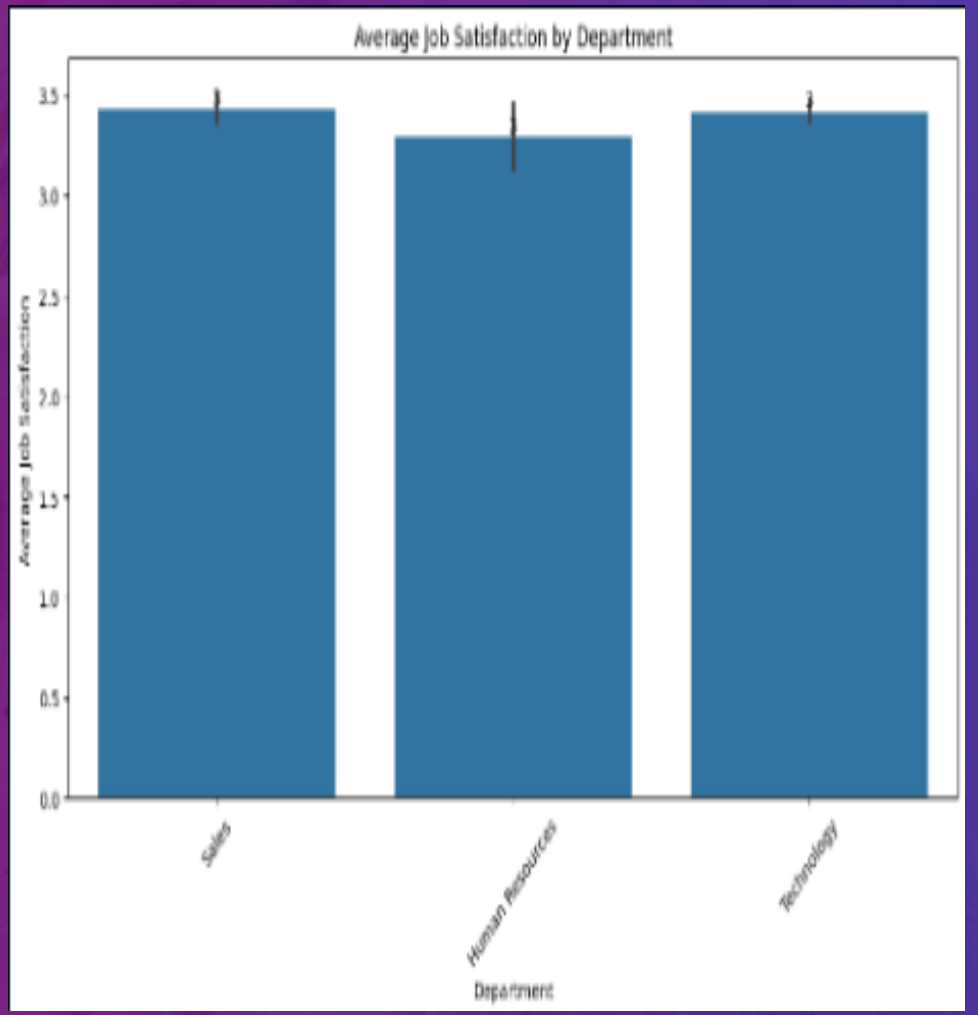
```python
# Calculate percentage of employees with stock options
stock_option_percentage = (merged_data['StockOptionLevel'].value_counts(normalize=True) * 100).get(1, 0)
print(f"Percentage of employees with stock options: {stock_option_percentage:.2f}%")
```

```
Percentage of employees with stock options: 37.50%
```

What Is The Average Job Satisfaction By Department?

```python
plt.figure(figsize=(12, 6))
avg_job_satisfaction = df_PerformanceRating.groupby('EmployeeID')['JobSatisfaction'].mean().reset_index()
merged_df = pd.merge(df_employee, avg_job_satisfaction, on='EmployeeID')
ax=sns.barplot(data=merged_df, x='Department', y='JobSatisfaction')
plt.title('Average Job Satisfaction by Department')
plt.xlabel('Department')
plt.ylabel('Average Job Satisfaction')
plt.xticks(rotation=45)
for container in ax.containers:
    ax.bar_label(container, fmt='%d')
plt.show()
```

Python



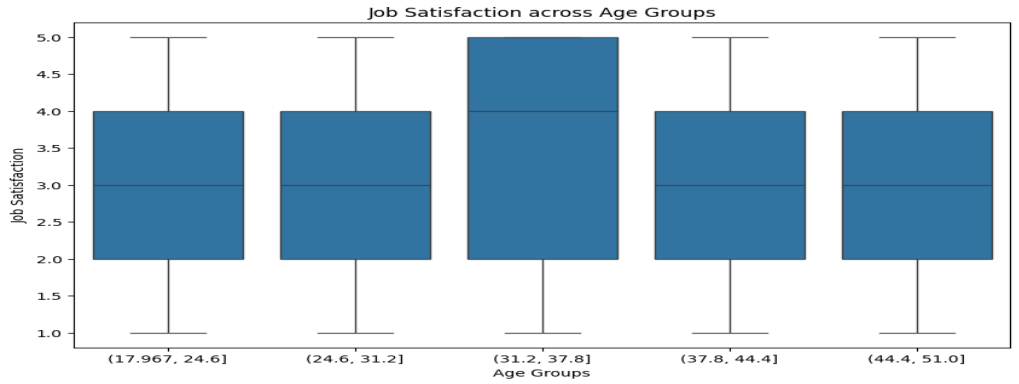
Which Ages Have More Job Satisfaction

```python
plt.figure(figsize=(10, 6))
sns.boxplot(data=merged_data, x=pd.cut(merged_data['Age'], bins=5), y='JobSatisfaction')
plt.title('Job Satisfaction across Age Groups')
plt.xlabel('Age Groups')
plt.ylabel('Job Satisfaction')
plt.show()
```

Python

```python
# Group by Age and calculate the average Job Satisfaction
age_job_satisfaction = merged_data.groupby('Age')['JobSatisfaction'].mean().reset_index()

# Plot a line chart for Age vs Average Job Satisfaction
plt.figure(figsize=(10, 6))
plt.plot(age_job_satisfaction['Age'], age_job_satisfaction['JobSatisfaction'], marker='o', color='b', label='Job Satisfaction')

# Add labels and title
plt.title('Job Satisfaction by Age', fontsize=14)
plt.xlabel('Age', fontsize=12)
plt.ylabel('Average Job Satisfaction', fontsize=12)
plt.grid(True)
plt.legend()

# Show the plot
plt.show()
```

Python



16.Avarage job Sataisfaction for each age group for all departments

```
SELECT emp.Department, emp.AgeGroup, AVG(pr.JobSatisfaction) AS "Avg JobSatisfaction" FROM employee AS emp JOIN performancerating AS pr ON emp.EmployeeID = pr.EmployeeID
GROUP BY Department, AgeGroup;
```

Profiling [Edit inline] [Edit] [Explain SQL] [Create PHP code] [Refresh]

Show all | Number of rows: 25 | Filter rows: Search this table

Extra options

Department	AgeGroup	Avg JobSatisfaction
Human Resources	[18-24]	3.1639
Human Resources	[25-31]	3.5203
Human Resources	[32-38]	3.3824
Human Resources	[39-45]	3.4067
Human Resources	[46-52]	3.5800
Sales	[18-24]	3.4021
Sales	[25-31]	3.4306
Sales	[32-38]	3.4854
Sales	[39-45]	3.3624
Sales	[46-52]	3.4545
Technology	[18-24]	3.4526
Technology	[25-31]	3.4088
Technology	[32-38]	3.4752
Technology	[39-45]	3.4625
Technology	[46-52]	3.3556

17.Avarage performance rating for all education levels

```
SELECT el.EducationLevel, (AVG(pr.SelfRating) + AVG(pr.ManagerRating)) / 2 AS "Avg PerformanceRating" FROM employee AS emp JOIN educationlevel AS el ON emp.Education =
el.EducationLevelID JOIN performancerating AS pr ON emp.EmployeeID = pr.EmployeeID GROUP BY el.EducationLevel ORDER BY `Avg PerformanceRating` DESC;
```

Profiling [Edit inline] [Edit] [Explain SQL] [Create PHP code] [Refresh]

Show all | Number of rows: 25 | Filter rows: Search this table

Extra options

EducationLevel	Avg PerformanceRating
No Formal Qualifications	3.75416667
Bachelors	3.73785118
Doctorate	3.72985782
High School	3.72353812
Masters	3.70561865

18. Average salary by Ethnicity

```
SELECT emp.Ethnicity ,AVG(emp.Salary) AS"Avg Salary by Ethnicity" FROM emp GROUP
BY emp.Ethnicity ORDER by `Avg Salary by Ethnicity` DESC;
```

Profiling [Edit inline] [Edit] [Explain SQL] [Create PHP code] [Refresh

Show all | Number of rows: 25

Extra options

	Ethnicity	Avg Salary by Ethnicity
Edit Copy Delete	White	115316.9349
Edit Copy Delete	Native Hawaiian	115274.1923
Edit Copy Delete	Black or African American	112176.8019
Edit Copy Delete	American Indian or Alaska Native	112037.2800
Edit Copy Delete	Asian or Asian American	109850.9646
Edit Copy Delete	Mixed or multiple ethnic groups	106132.8485
Edit Copy Delete	Other	101652.1250

19. Average job satisfaction by Ethnicity

Showing rows 0 - 6 (7 total, Query took 0.1265 seconds.)

```
SELECT emp.Ethnicity ,AVG(performance.JobSatisfaction) AS"Avg job satisfaction
by Ethnicity" FROM emp INNER JOIN performance ON
emp.EmployeeID=performance.EmployeeID GROUP BY emp.Ethnicity ORDER by `Avg job
satisfaction by Ethnicity` DESC;
```

Profiling [Edit inline] [Edit] [Explain SQL] [Create PHP code] [Refresh

Show all | Number of rows: 25

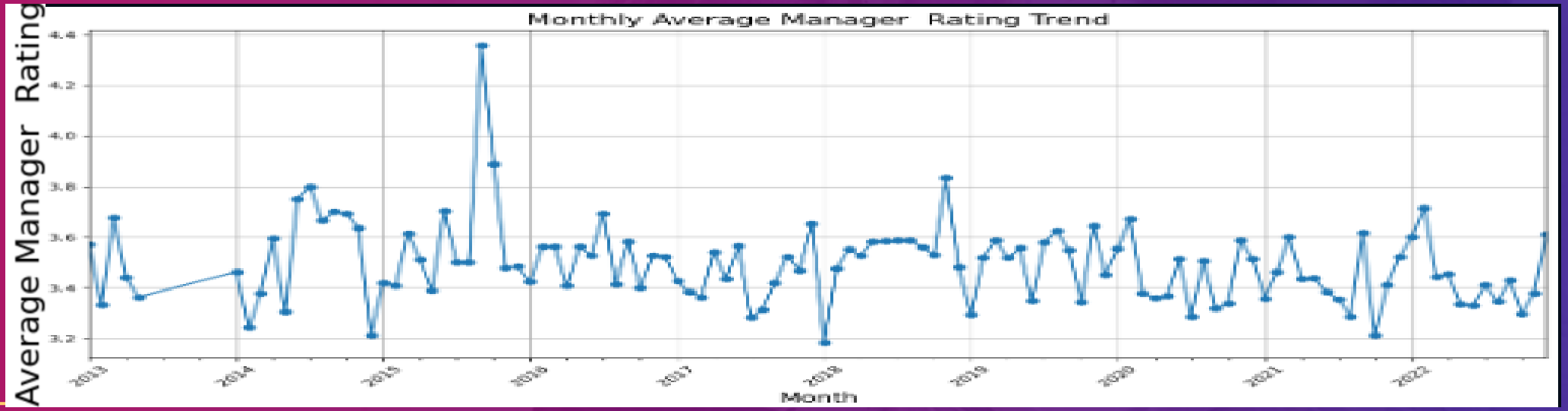
Extra options

Ethnicity	Avg job satisfaction by Ethnicity
American Indian or Alaska Native	3.5211
Black or African American	3.4631
Asian or Asian American	3.4275
Other	3.4253
White	3.4240
Mixed or multiple ethnic groups	3.4086
Native Hawaiian	3.3567

Monthly Average Manager Rating Trend

```python
merged_df['ReviewDate'] = pd.to_datetime(merged_df['ReviewDate'])
trend = merged_df.groupby(merged_df['ReviewDate'].dt.to_period('M'))['ManagerRating'].mean()
plt.figure(figsize=(12, 6))
trend.plot(marker='o')
plt.title('Monthly Average Manager  Rating Trend', fontsize=16)
plt.xlabel('Month', fontsize=14)
plt.ylabel('Average Self Rating', fontsize=14)
plt.xticks(rotation=45)
plt.grid()
plt.tight_layout()
plt.show()
```

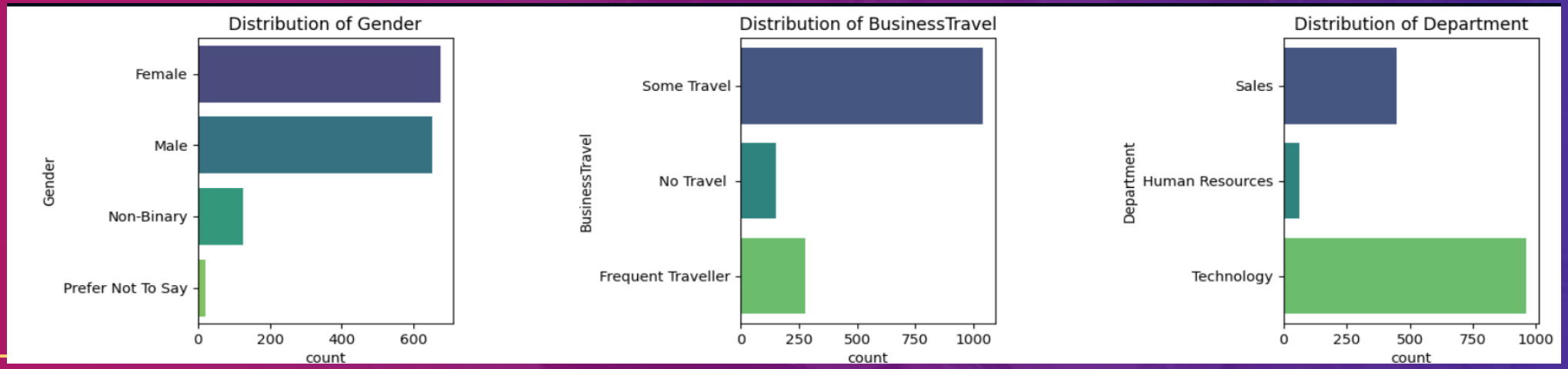
Python



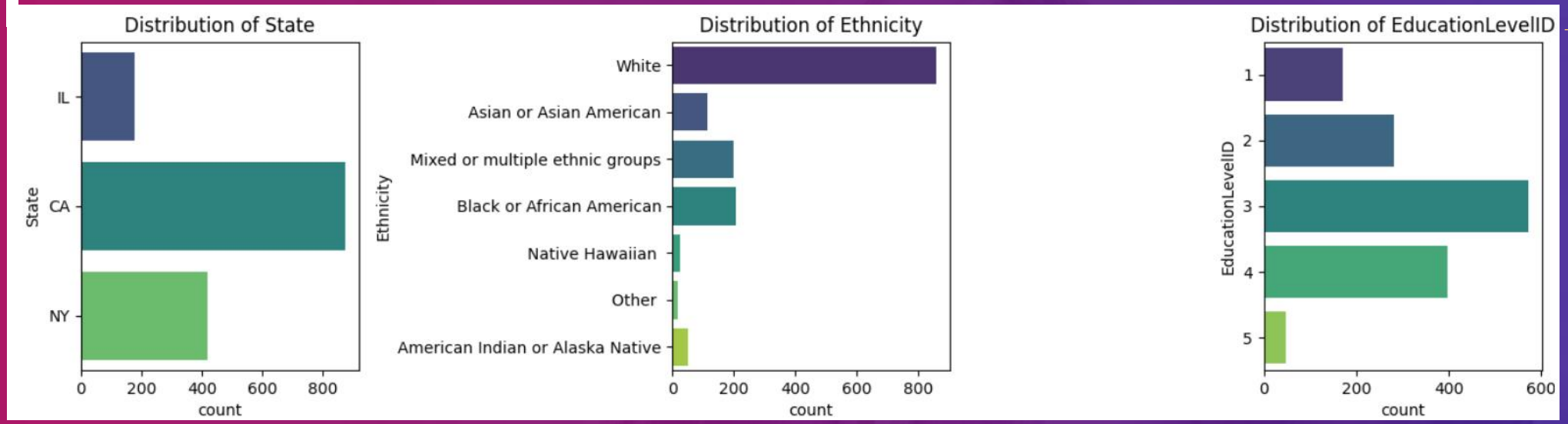
Plotting Count Plots For Categorical Columns

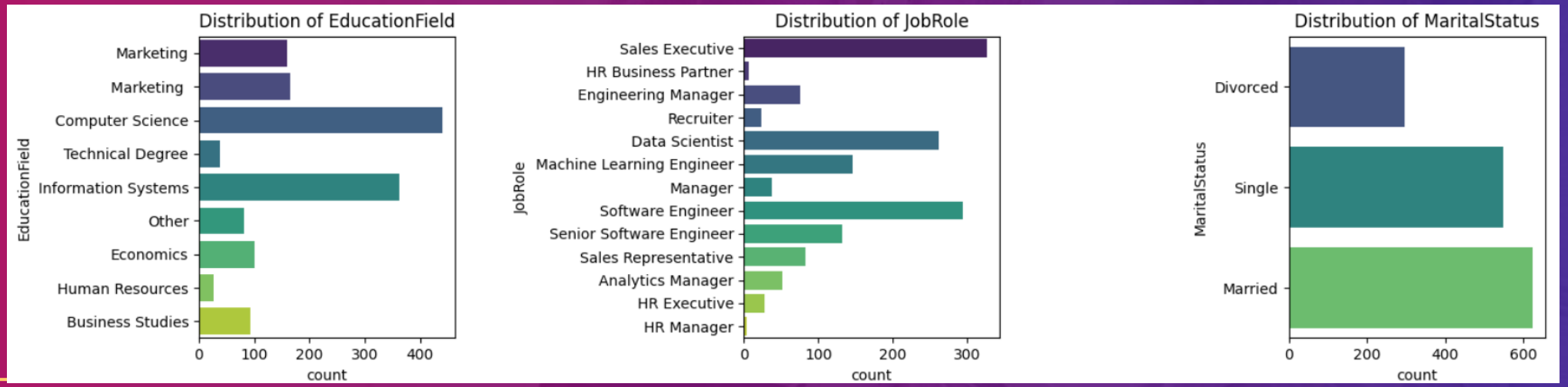
```python
categorical_cols = ['Gender', 'BusinessTravel', 'Department', 'State', 'Ethnicity', 'EducationLevelID', 'EducationField', 'JobRol

plt.figure(figsize=(15, 15))

for i, col in enumerate(categorical_cols, 1):
    plt.subplot(4, 3, i)
    sns.countplot(y = df_employee[col], palette='viridis')
    plt.title(f'Distribution of {col}')

plt.tight_layout()
plt.show()
```

Python

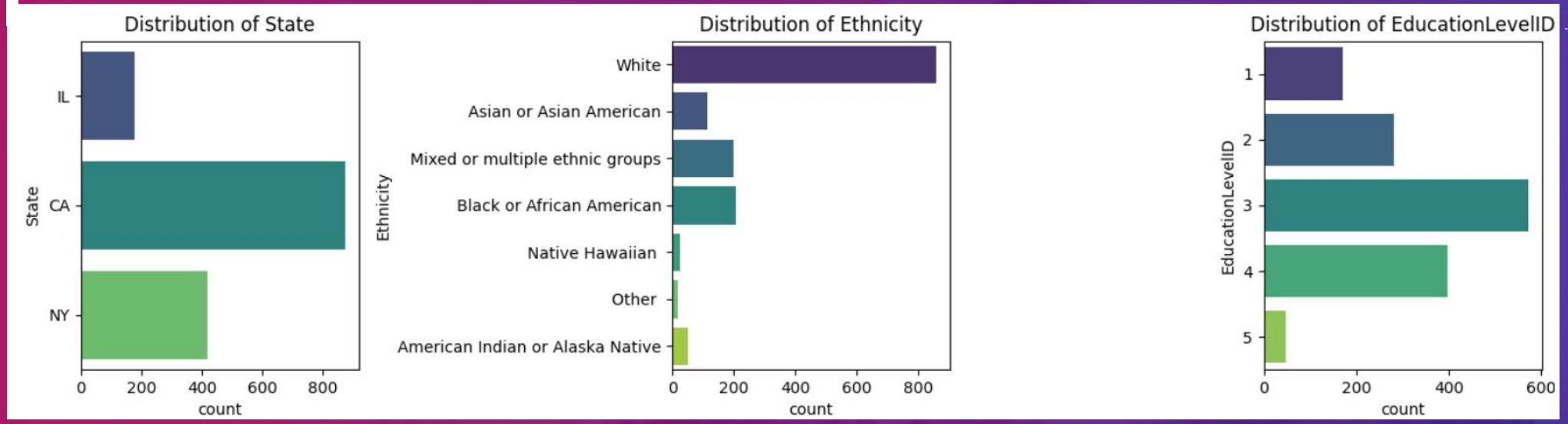


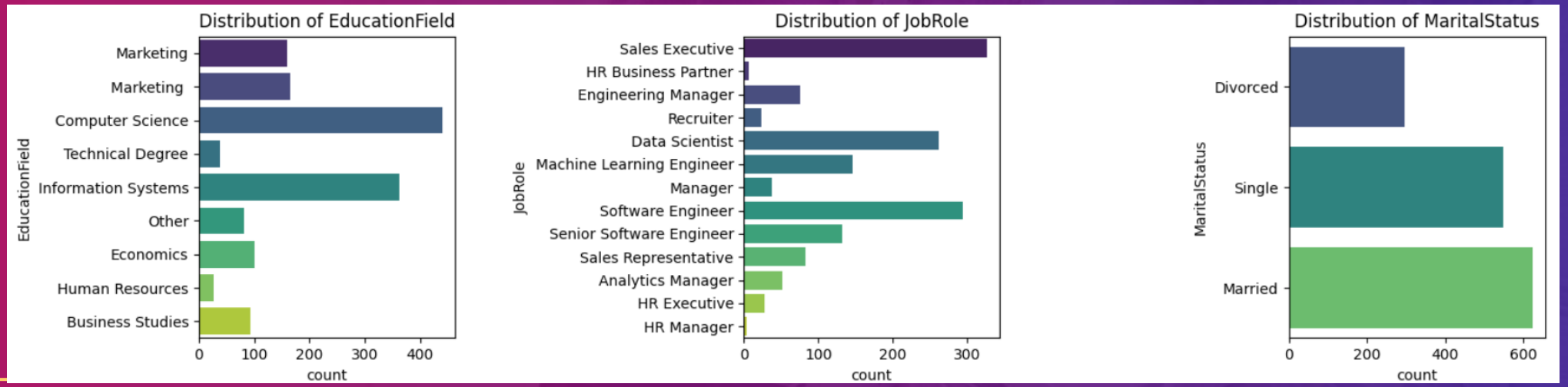
Ploting Count Plots For Categorical Columns



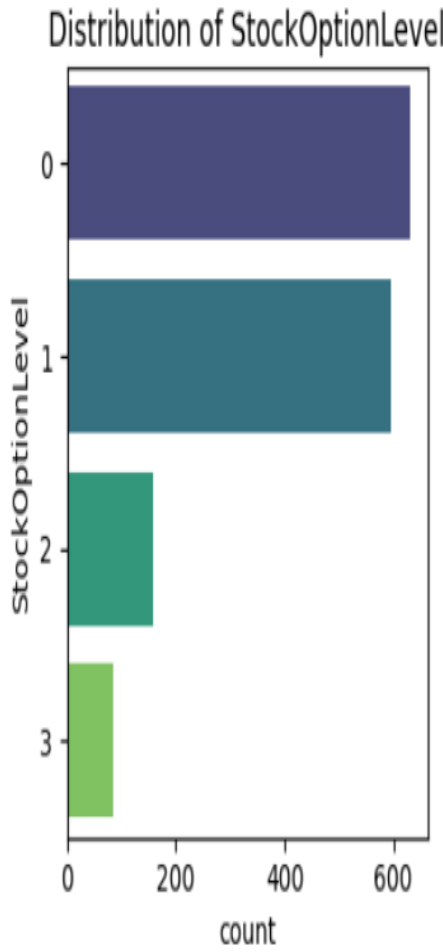


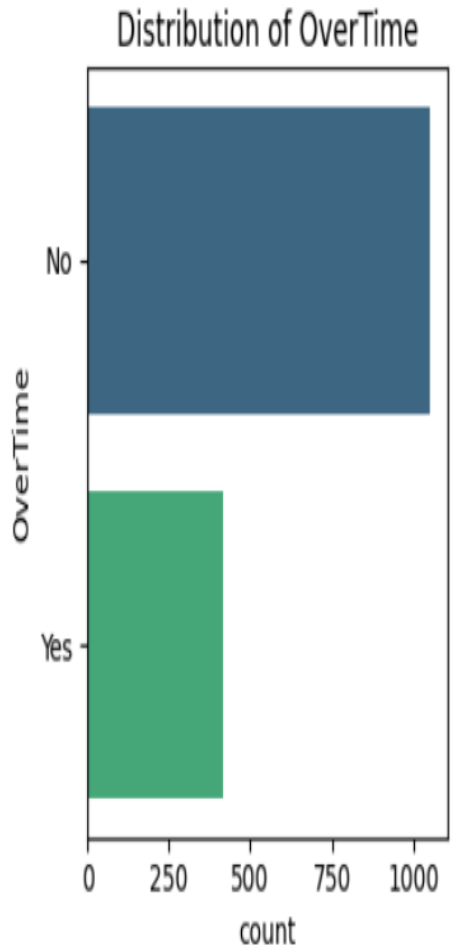
Plotting Count Plots For Categorical Columns

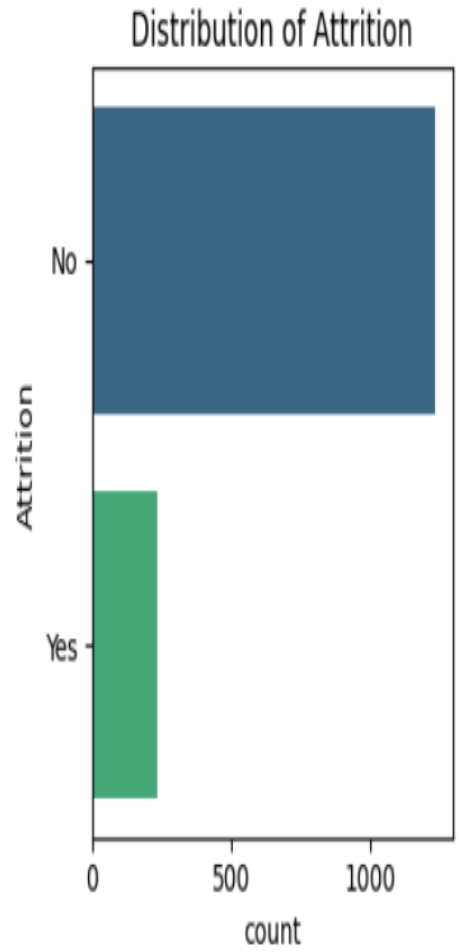




Ploting Count Plots For Categorical Columns



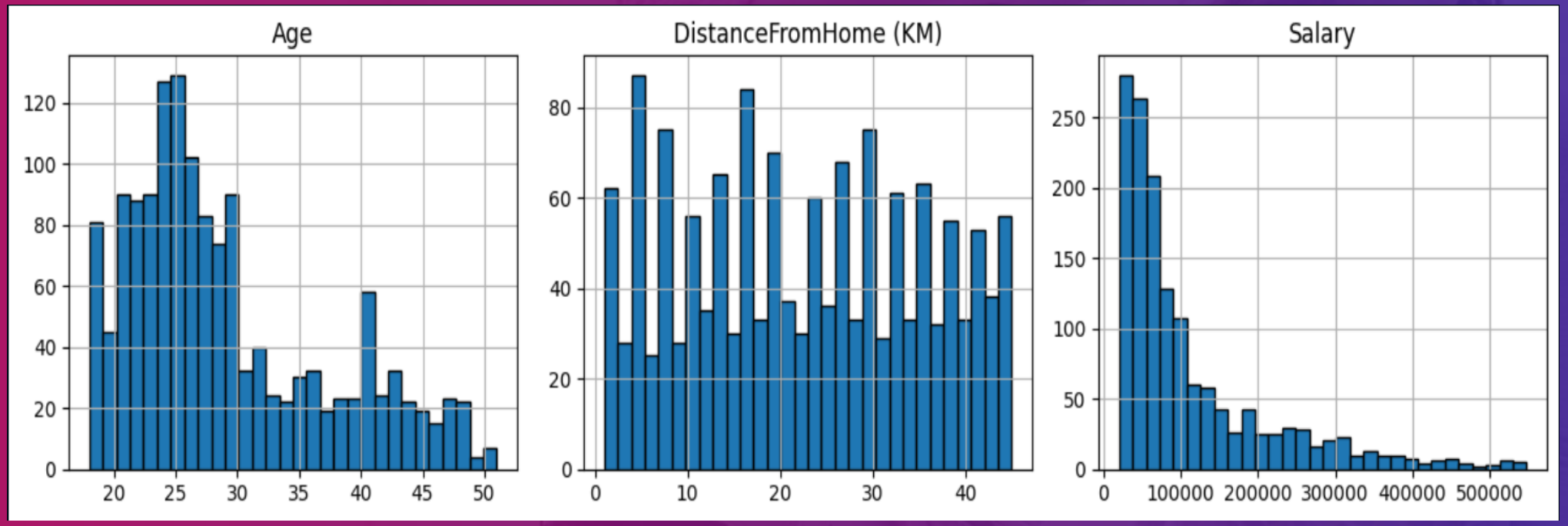


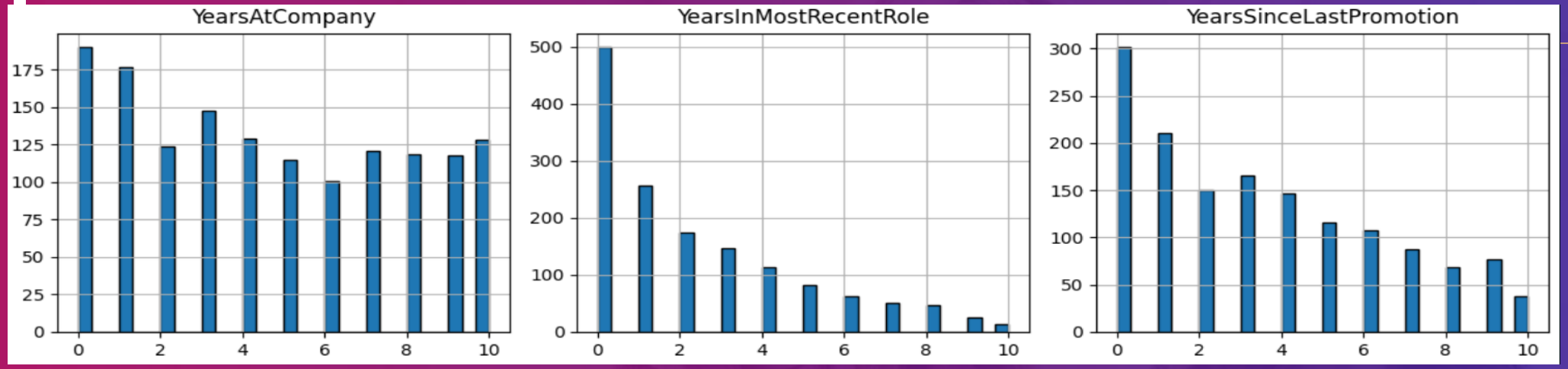


Plotting Count Plots For Numerical Columns

```python
numerical_cols = ['Age', 'DistanceFromHome (KM)', 'Salary', 'YearsAtCompany', 'YearsInMostRecentRole', 'YearsSinceLastPromotion',
df_employee[numerical_cols].hist(figsize=(12, 10), bins=30, edgecolor='black')
plt.tight_layout()
plt.show()
```

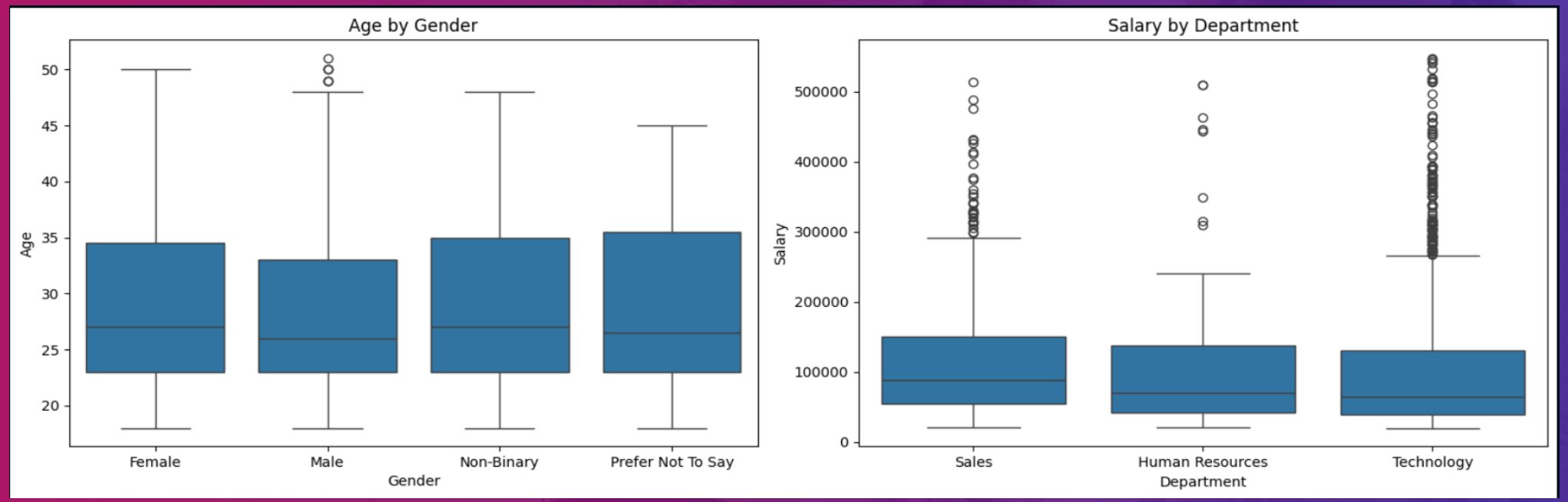
Python




YearsAtCompany
YearsInMostRecentRole
YearsSinceLastPromotion

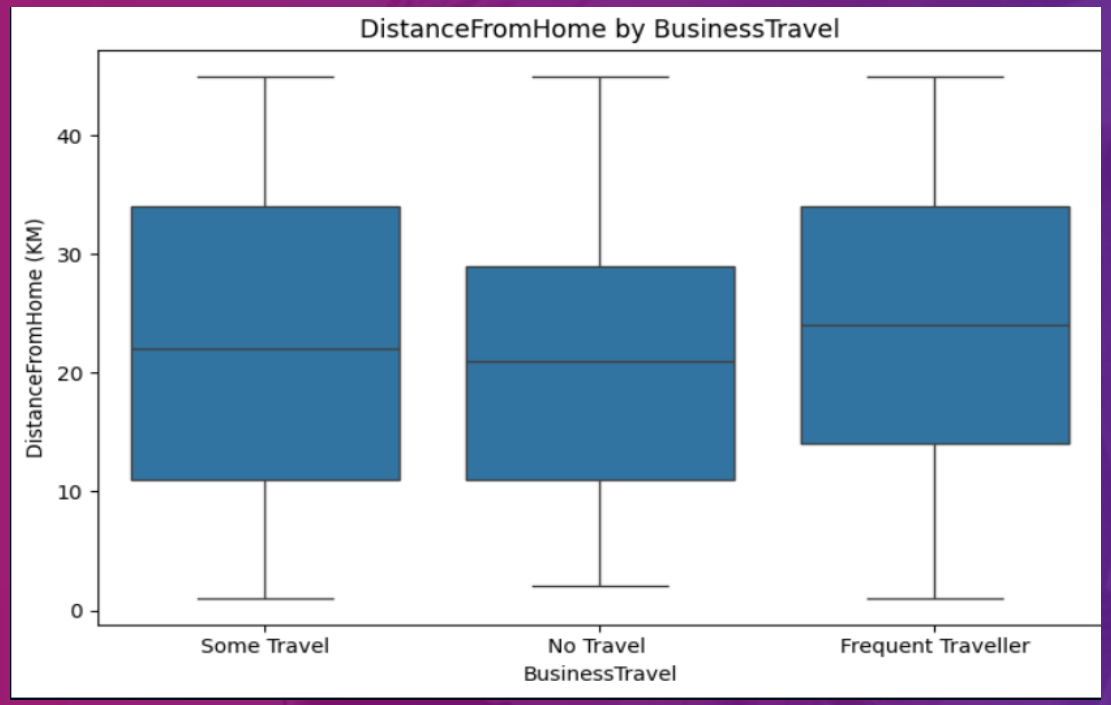

YearsWithCurrManager

```python
plt.figure(figsize=(15, 10))
# Age by Gender
plt.subplot(2, 2, 1)
sns.boxplot(x='Gender', y='Age', data=df_employee)
plt.title('Age by Gender')

# Salary by Department
plt.subplot(2, 2, 2)
sns.boxplot(x='Department', y='Salary', data=df_employee)
plt.title('Salary by Department')
```



```python
# DistanceFromHome by BusinessTravel
plt.subplot(2, 2, 3)
sns.boxplot(x='BusinessTravel', y='DistanceFromHome (KM)', data=df_employee)
plt.title('DistanceFromHome by BusinessTravel')
plt.tight_layout()
```

Python



FORECASTING

1- Predecting The Number of Employees that may have attrition in the company by the next year

Import libraries and merge 2 tables in one new table

```
import pandas as pd
from sklearn.ensemble import RandomForestClassifier
from sklearn.model_selection import train_test_split
from sklearn.metrics import classification_report
from sklearn.preprocessing import OneHotEncoder

# Assuming df_employee and df_PerformanceRating are your original DataFrames

# Merge the two DataFrames
merged_data = df_employee.merge(df_PerformanceRating, on='EmployeeID',how='inner')
```

```python
features = ['Age', 'JobRole', 'Salary', 'YearsAtCompany', 'WorkLifeBalance', 'JobSatisfaction']
X = merged_data[features]
y = merged_data['Attrition'].apply(lambda x: 1 if x == 'Yes' else 0)  # Convert Attrition Yes/No to binary
X = pd.get_dummies(X, columns=['JobRole'], drop_first=True)  # Convert JobRole to dummy variables
assert len(X) == len(y), "X and y must have the same number of samples."
X_train, X_test, y_train, y_test, employee_ids_train, employee_ids_test = train_test_split(
    X, y, merged_data['EmployeeID'], test_size=0.3, random_state=42
)
model = RandomForestClassifier(random_state=42)  # Set random_state for reproducibility
model.fit(X_train, y_train)
y_pred = model.predict(X_test)
print(classification_report(y_test, y_pred))
results = pd.DataFrame({
    'EmployeeID': employee_ids_test,   # The correct EmployeeID for each test prediction
    'Attrition_Prediction': y_pred     # The predictions (0 = stay, 1 = leave)
})
employees_to_leave = results[results['Attrition_Prediction'] == 1]
distinct_count = employees_to_leave['EmployeeID'].nunique()
print("Employees likely to leave:")
print(distinct_count)
```

```
...                   precision    recall  f1-score   support

           0       0.99      0.97      0.98      1356
           1       0.94      0.97      0.96       657

    accuracy                           0.97      2013
   macro avg       0.96      0.97      0.97      2013
weighted avg       0.97      0.97      0.97      2013

Employees likely to leave:
265
```

2- Predict the Avg Salary to Each Job role:

```python
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
features = ['Age', 'YearsAtCompany', 'JobSatisfaction', 'ManagerRating', 'JobRole']
X = merged_data[features]
y = merged_data['Salary']
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=42)
model = LinearRegression()
model.fit(X_train, y_train)
job_roles_columns = [col for col in X_test.columns if 'JobRole' in col]
job_roles_encoded = pd.get_dummies(job_roles, columns=['JobRole'], drop_first=True)
job_role_names = job_roles_encoded.columns
average_salary_by_job_role = X_test.groupby('JobRole')['PredictedSalary'].mean()
Print("متوسط الرواتب المتوقعة لكل دور وظيفي في السنة القادمة:")
print(average_salary_by_job_role)
```

Python

```
متوسط الرواتب المتوقعة لكل دور وظيفي في السنة القادمة:
JobRole
Data Scientist                  97,731.42
Engineering Manager            294,319.15
HR Business Partner            311,464.38
HR Executive                    96,212.39
HR Manager                     424,422.60
Machine Learning Engineer      131,475.11
Manager                        340,470.42
Recruiter                       40,073.01
Sales Executive                128,787.84
Sales Representative            40,446.79
Senior Software Engineer       131,410.38
Software Engineer               54,650.61
Name: PredictedSalary, dtype: float64
```

3- Which departments are likely to have the highest attrition rates next year?

```python
# Group data by Department and calculate attrition rate
attrition_by_department = merged_data.groupby('Department')['Attrition'].apply(lambda x: (x == 'Yes').mean())

# Sort by highest attrition rate
attrition_by_department = attrition_by_department.sort_values(ascending=False)

# Display results
print(attrition_by_department)
```

Python

```
Department
Sales              0.41
Human Resources    0.38
Technology         0.30
Name: Attrition, dtype: float64
```


```
Department
Sales              0.41
Human Resources    0.38
Technology         0.30
Name: Attrition, dtype: float64
```

DATA VISUALIZATION

Human Resource Dashboard

Go to Employye Dashboard

Go toPerformce Rating Dashboard

Members Team:

1- Rashad Hazem Ali
2- Rasha Gaafar
3- Mariam Korashy
4- Mohamed Abd El Aleem mohamed
5- Esraa Eleraky
6- Basem Barakat-Allah Abdulqader Abdulkarim

Performance (2) | Performance | Main | Sheet 9 | Sheet 10 | Sheet 11 | Sheet 12 | Sheet 13 | Sheet 14 | Sheet 15 | Sheet 16 | Sheet 17 | Sheet 18 | Sheet 19 | Sheet 20 | Sheet 21

Employee Dashboard

Year of HireDate

(Multiple values)

Enter search text

- [ ] (All)
- [ ] 2012
- [x] 2013
- [ ] 2014
- [x] 2015
- [ ] 2016
- [ ] 2017
- [ ] 2018
- [ ] 2019
- [ ] 2020
- [ ] 2021
- [ ] 2022

Total Employee

263

Total Salary

30,746,522

Avg Salary

116,907

Hire Employee per years

Hiring by Each Department and Date

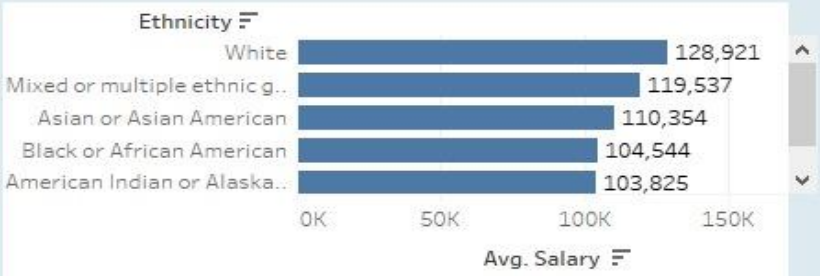
Attrition by Department

	Attrition	
Department	No	Yes
Human Resources	8	1
Sales	54	19
Technology	163	18

Avg Salary by Education Level

EducationLevel	Count of Employee	Avg. Salary
Bachelors	95	124,074
Doctorate	12	168,634
High School	63	109,701
Masters	61	115,546
No Formal Qualifications	32	93,011

Avarage Salary by Ethncity



Over Time by Department

Employee1 | Employee2 | Performance (2) | Performance | Main | Sheet 9 | Sheet 10 | Sheet 11 | Sheet 12 | Sheet 13 | Sheet 14 | Sheet 15 | Sheet 16 | Sheet 17 | Sheet 18

Employee Dashboard

Year of HireDate

(All)

Department

(All)

Select Gender

All

Total Employee

1,470

Total Salary

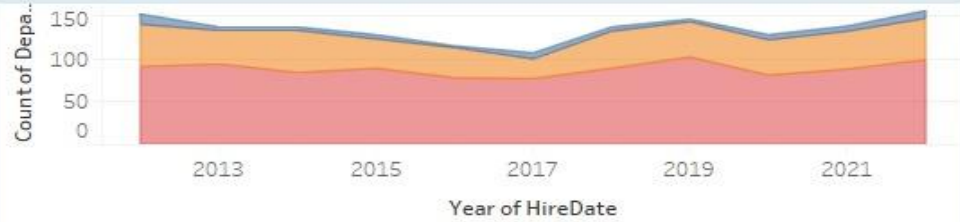
166,046,052

Avg Salary

112,956

Hire Employee per years

Hiring by Each Department and Date



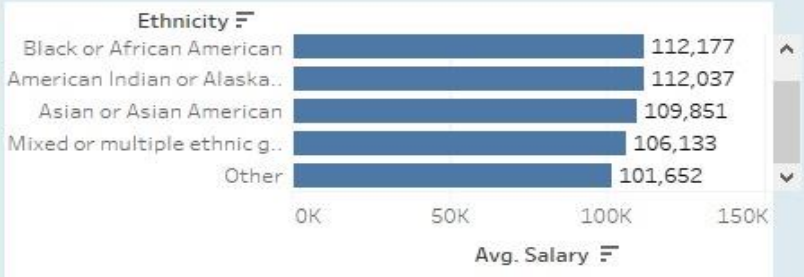
Attrition by Department

Department	Attrition	
	No	Yes
Human Resources	51	12
Sales	354	92
Technology	828	133

Avg Salary by Education Level

EducationLevel	Count of Employee	Avg. Salary
Bachelors	572	115,405
Doctorate	48	154,269
High School	282	105,181
Masters	398	117,641
No Formal Qualifications	170	94,983

Avarage Salary by Ethncity



Over Time by Department

Employee1 | Employee2 | Performance (2) | Performance | Main | Sheet 9 | Sheet 10 | Sheet 11 | Sheet 12 | Sheet 13 | Sheet 14 | Sheet 15 | Sheet 16 | Sheet 17 | Sheet 18

Employee Dashboard

EducationField

(All)

Select Gender

All

Year of HireDate

2018 — 2022

JobRole

(All)

Total Employee

564

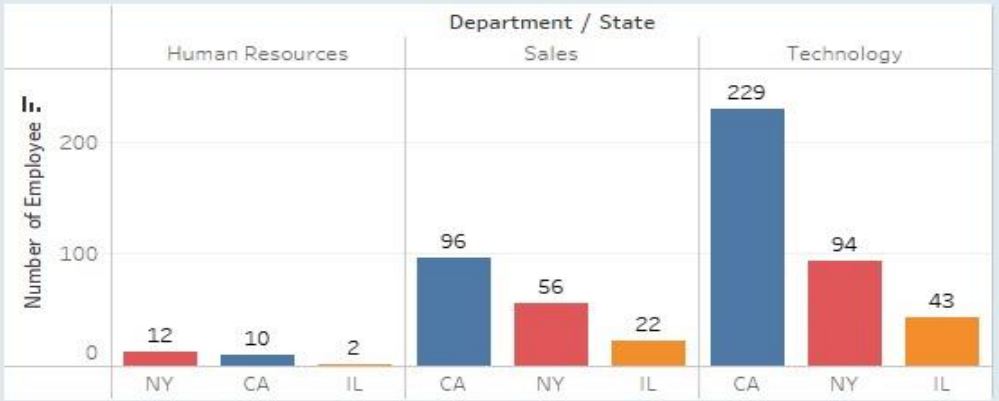
Total Salary

63,539,050

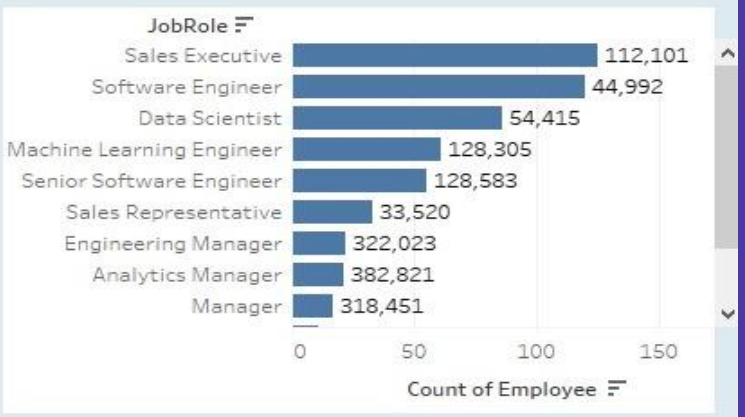
Avg Salary

112,658

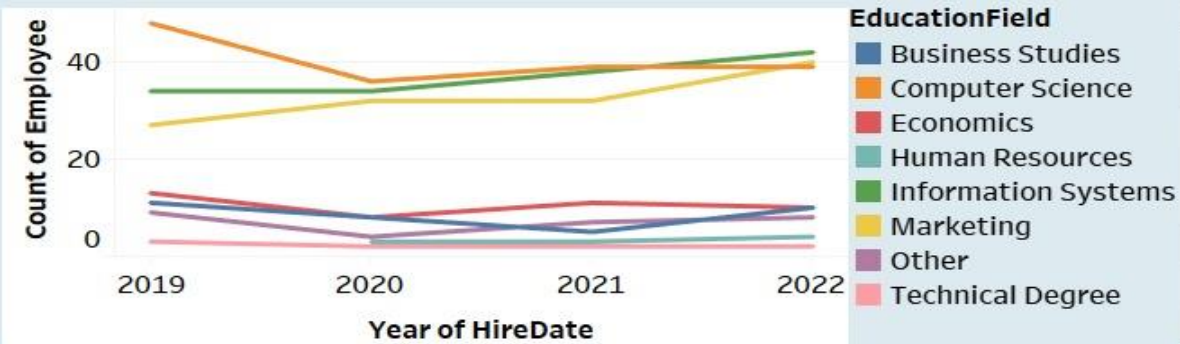
Num of Emp in Each state and department



Jop Role count



Education Field Hired by time



Salary by Each Education field

EducationField	
Business Studies	3,541,368
Computer Science	16,775,966
Economics	3,900,357
Human Resources	1,708,988
Information Systems	17,819,650
Marketing	16,357,520
Other	2,578,752
Technical Degree	856,449

Employee1 | Employee2 | Performance (2) | Performance | Main | Sheet 9 | Sheet 10 | Sheet 11 | Sheet 12 | Sheet 13 | Sheet 14 | Sheet 15 | Sheet 16 | Sheet 17 | Sheet 18 | S

Employee Dashboard

EducationField

(All)

Select Gender

All

Year of HireDate

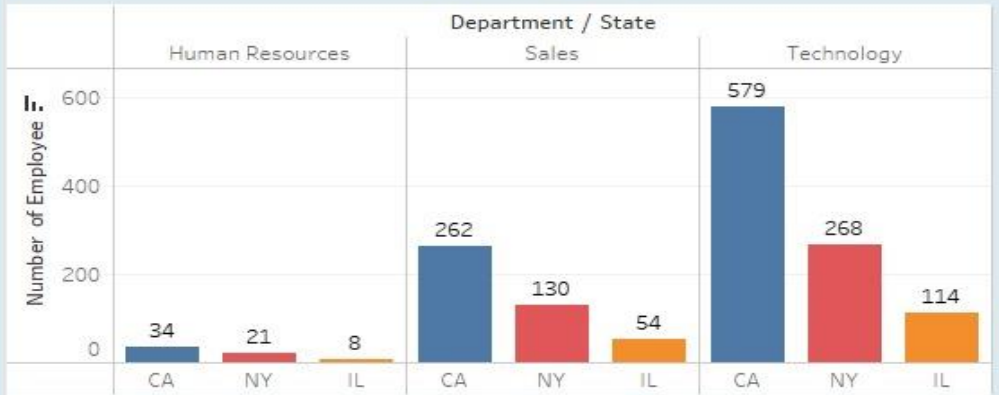
2012 — 2022

JobRole

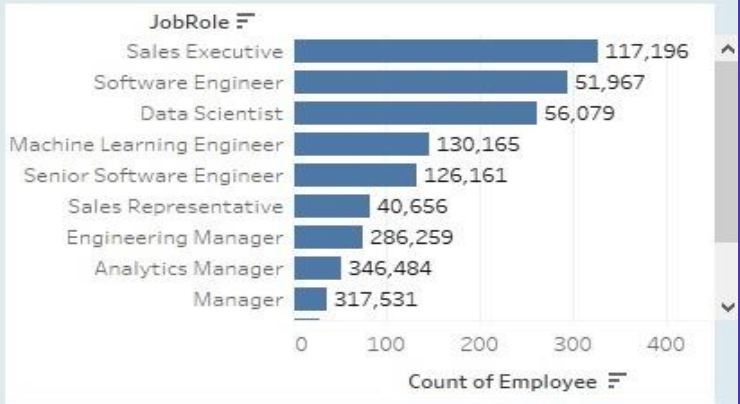
(All)

Total Employee	Total Salary	Avg Salary
1,470	166,046,052	112,956

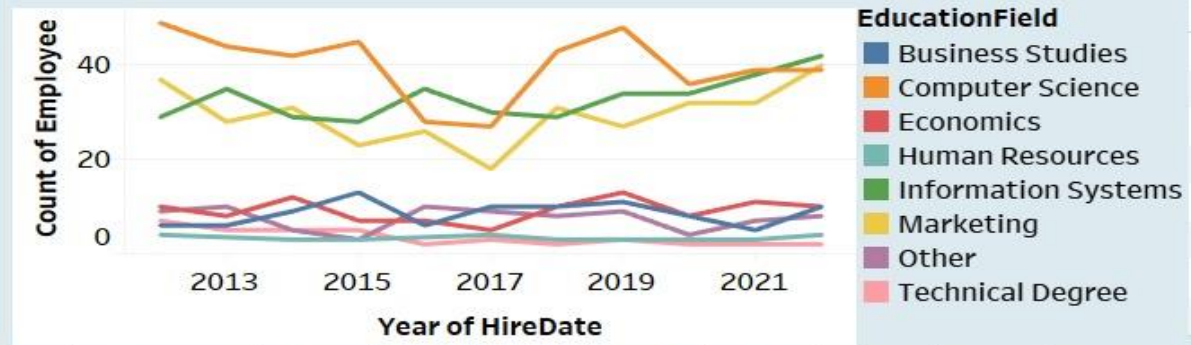
Num of Emp in Each state and department



Jop Role count



Education Field Hired by time



Salary by Each Education field

EducationField	
Business Studies	9,250,227
Computer Science	48,115,757
Economics	11,334,205
Human Resources	3,930,278
Information Systems	41,520,135
Marketing	40,390,223
Other	7,900,032
Technical Degree	3,605,195

Employee1 | Employee2 | Performance (2) | Performance | Main | Sheet 9 | Sheet 10 | Sheet 11 | Sheet 12 | Sheet 13 | Sheet 14 | Sheet 15 | Sheet 16 | Sheet 17 | Sheet 18 | S

Employee Dashboard

EducationField

(Multiple values)

Enter search text

- [ ] (All)
- [ ] Business Studies
- [x] Computer Science
- [ ] Economics
- [x] Human Resources
- [ ] Information Systems
- [ ] Marketing
- [ ] Other
- [ ] Technical Degree

(All)

Total Employee

467

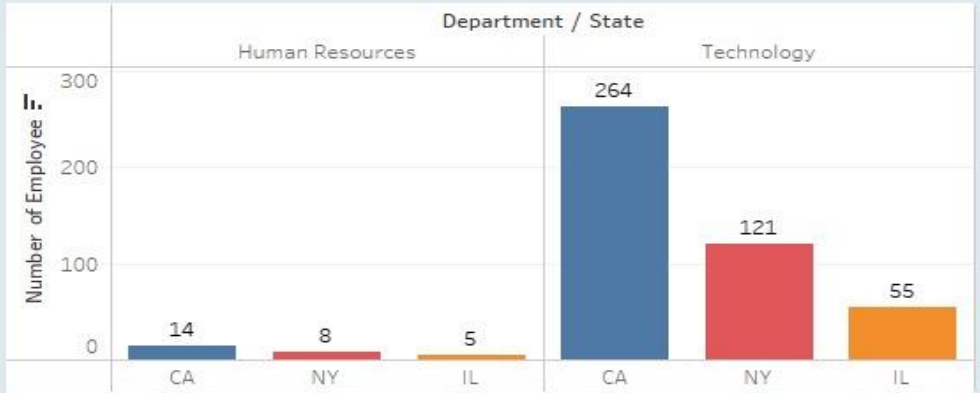
Total Salary

52,046,035

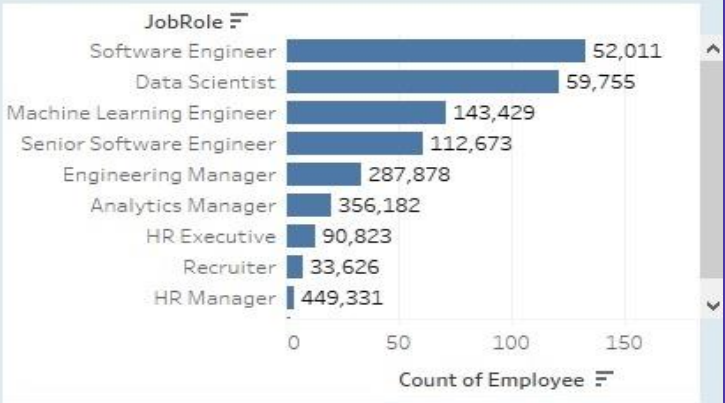
Avg Salary

111,448

Num of Emp in Each state and department



Jop Role count



Education Field Hired by time

Salary by Each Education field

EducationField	
Computer Science	48,115,757
Human Resources	3,930,278

Employee1 | **Employee2** | Performance (2) | Performance | Main | Sheet 9 | Sheet 10 | Sheet 11 | Sheet 12 | Sheet 13 | Sheet 14 | Sheet 15 | Sheet 16 | Sheet 17 | Sheet 18

Employee Dashboard

EducationField

(Multiple values)

Enter search text

- [ ] (All)
- [ ] Business Studies
- [x] Computer Science
- [ ] Economics
- [x] Human Resources
- [ ] Information Systems
- [ ] Marketing
- [ ] Other
- [ ] Technical Degree

(All)

Total Employee

467

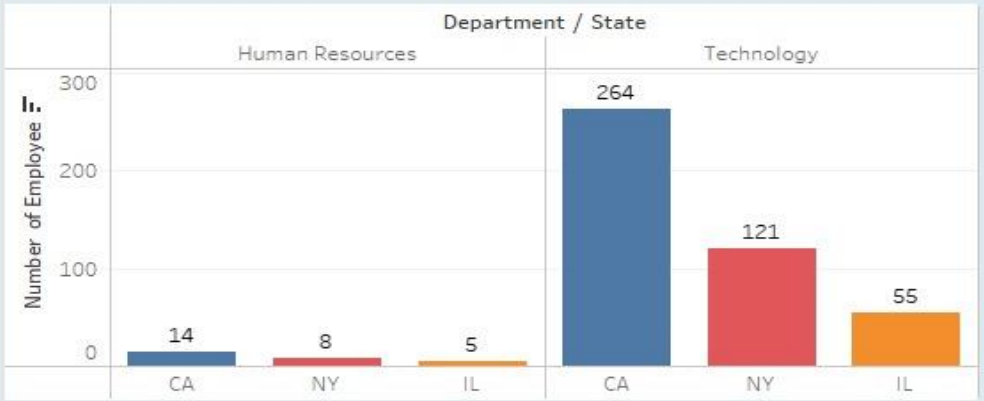
Total Salary

52,046,035

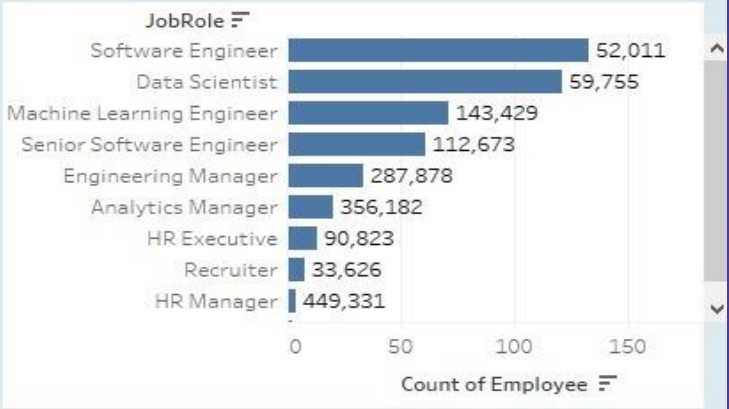
Avg Salary

111,448

Num of Emp in Each state and department



Jop Role count



Education Field Hired by time

Salary by Each Education field

EducationField	
Computer Science	48,115,757
Human Resources	3,930,278

Employee1 | Employee2 | Performance (2) | Performance | Main | Sheet 9 | Sheet 10 | Sheet 11 | Sheet 12 | Sheet 13 | Sheet 14 | Sheet 15 | Sheet 16 | Sheet 17 | Sheet 18

Performance rating

ManagerRating

(All)

Select Gender

All

JobSatisfaction Rating

(All)

Year of ReviewDate

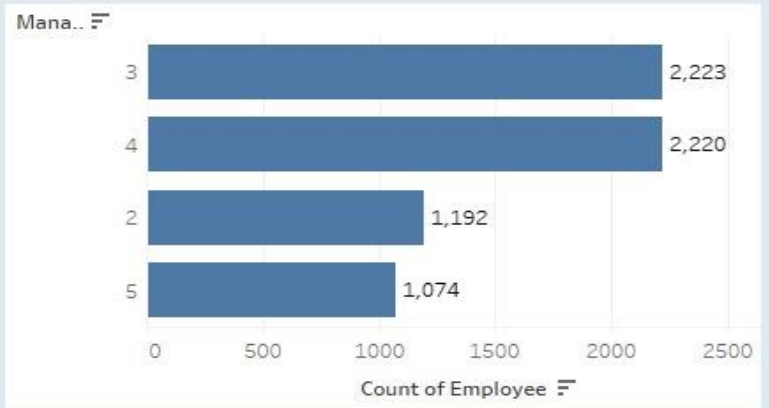
2013 — 2022

Total Employee	Total Salary	Avg Salary
1,470	166,046,052	112,956

Manager Rating For Each Department

ManagerRa..	Department	
2	Human Resources	67
	Sales	417
	Technology	708
3	Human Resources	87
	Sales	683
	Technology	1,453
4	Human Resources	97
	Sales	714
	Technology	1,409
5	Human Resources	52
	Sales	335
	Technology	687

Manager rating to all Employee



Job Satisfaction over years

Environment satisfaction by Ethnicity

	Ethnicity						
Environment Satisfactio..	American Ind..	Asian or Asi..	Black or Afr..	Mixed or mul..	Native Hawa..	Other	White
1	4	11	18	20	2	1	80
2	5	14	21	24	6	1	70
3	94	220	356	350	54	30	1,107
4	90	224	350	373	48	29	1,061
5	91	214	326	332	47	26	1,010

Employee1 | Employee2 | Performance (2) | **Performance** | Main | Sheet 9 | Sheet 10 | Sheet 11 | Sheet 12 | Sheet 13 | Sheet 14 | Sheet 15 | Sheet 16 | Sheet 17 | Sheet 18 | S

Performance rating

ManagerRating

(Multiple values)

- (All)
- 2
- 3
- 4
- 5

JobSatisfaction Rating

(All)

Year of ReviewDate

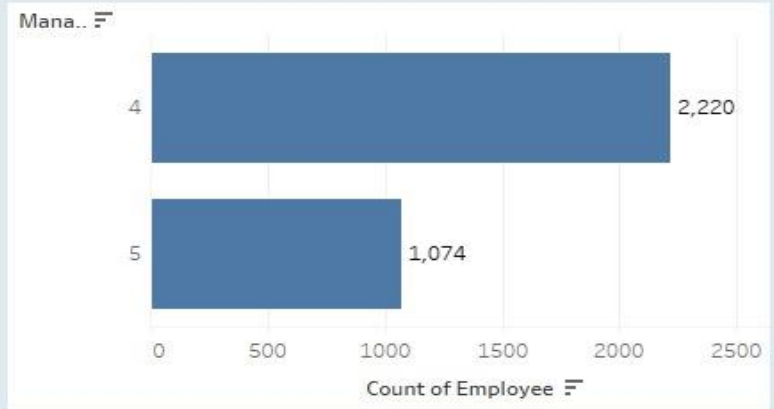
2013 — 2022

Total Employee	Total Salary	Avg Salary
1,121	125,226,438	111,710

Manager Rating For Each Department

ManagerRa..	Department	
4	Human Resources	97
	Sales	714
	Technology	1,409
5	Human Resources	52
	Sales	335
	Technology	687

Manager rating to all Employee



Job Satisfaction over years

Environment satisfaction by Ethnicity

	Ethnicity						
Environment Satisfactio..	American Ind..	Asian or Asi..	Black or Afr..	Mixed or mul..	Native Hawa..	Other	White
1		7	5	11	1	1	41
2	5	5	10	15	3	1	36
3	37	110	177	158	24	17	560
4	42	130	169	177	25	14	515
5	43	100	156	159	19	16	505

Employee1 | Employee2 | Performance (2) | **Performance** | Main | Sheet 9 | Sheet 10 | Sheet 11 | Sheet 12 | Sheet 13 | Sheet 14 | Sheet 15 | Sheet 16 | Sheet 17 | Sheet 18

Performance rating Dashboard

Gender

(All)

JobRole

(All)

EducationLevel

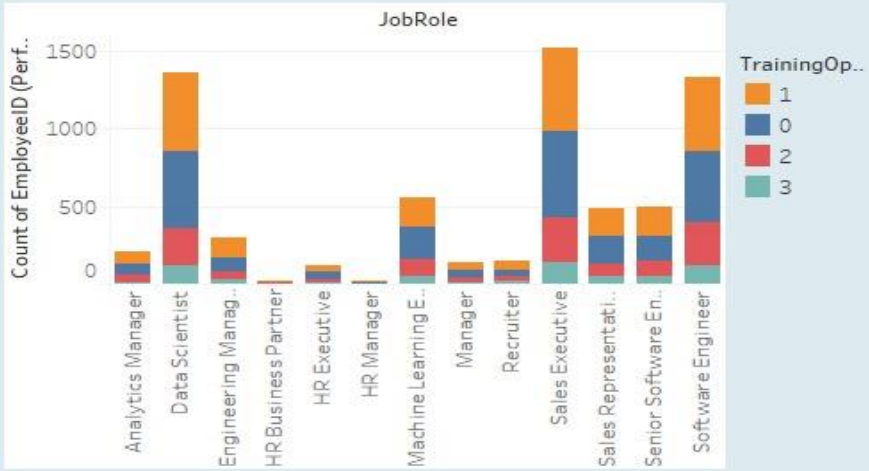
(All)

Total Employee	Total Salary	Avg Salary
1,470	166,046,052	112,956

Work life balance for Each Gender

	WorkLifeBalance (PerformanceRating U.csv)				
Gender	1	2	3	4	5
Female	52	783	762	831	663
Male	61	750	744	710	703
Non-Binary	8	151	153	145	132
Prefer Not To Say		18	11	20	12

Training opportunities token for Each job role



Manager Rating for Each Education level

Employee1 | Employee2 | **Performance (2)** | Performance | Main | Sheet 9 | Sheet 10 | Sheet 11 | Sheet 12 | Sheet 13 | Sheet 14 | Sheet 15 | Sheet 16 | Sheet 17 | Sheet 18

Performance rating Dashboard

Gender

(All)

JobRole

(All)

EducationLevel

(Multiple values)

Enter search text

- [ ] (All)
- [ ] Bachelors
- [x] Doctorate
- [x] High School
- [x] Masters
- [x] No Formal Qualifications

Total Employee

898

Total Salary

100,034,146

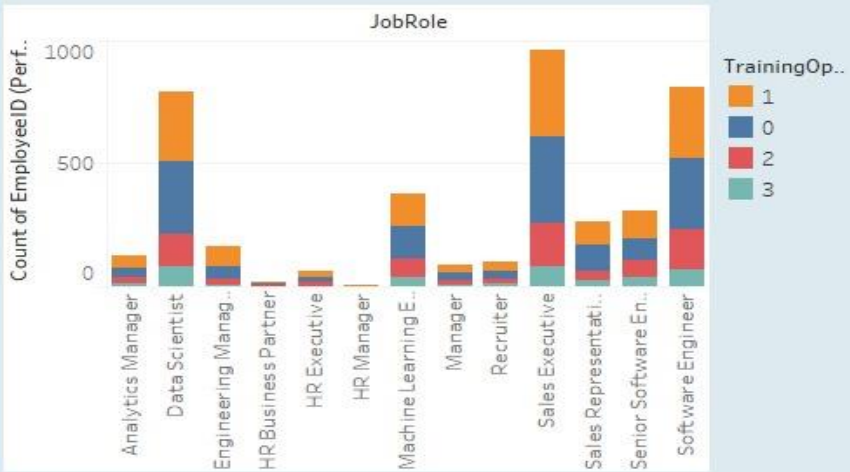
Avg Salary

111,397

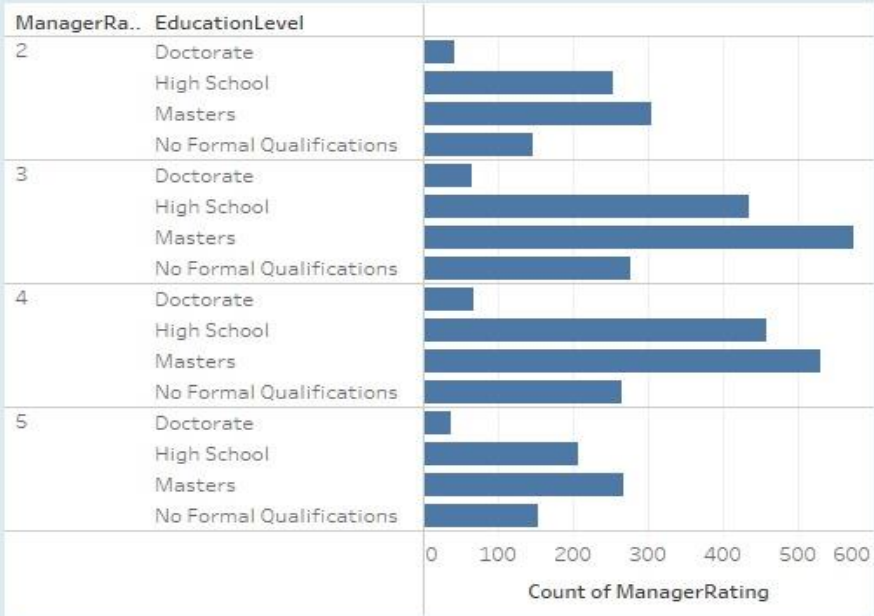
Work life balance for Each Gender

	WorkLifeBalance (PerformanceRating U.csv)				
Gender	1	2	3	4	5
Female	27	477	453	507	402
Male	36	462	446	446	418
Non-Binary	6	93	94	91	70
Prefer Not To Say		15	8	15	9

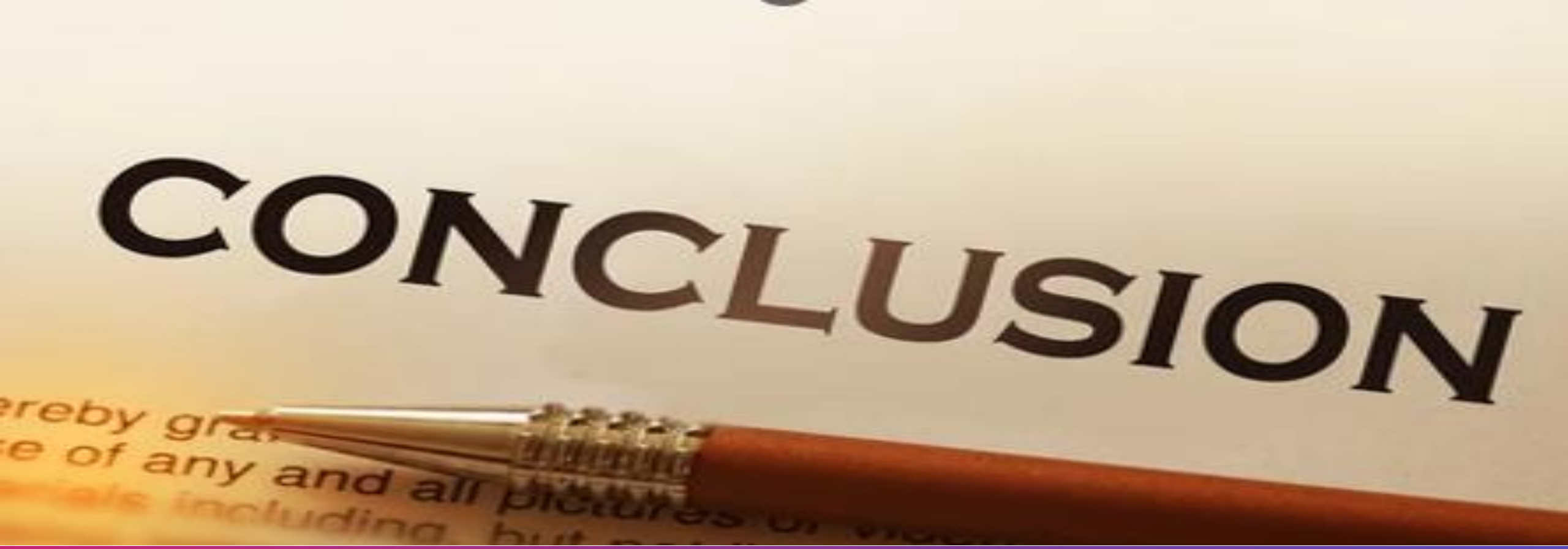
Training opportunities token for Each job role



Manager Rating for Each Education level



Employee1 | Employee2 | **Performance (2)** | Performance | Main | Sheet 9 | Sheet 10 | Sheet 11 | Sheet 12 | Sheet 13 | Sheet 14 | Sheet 15 | Sheet 16 | Sheet 17 | Sheet 18 | S



Please Listen to the Final Project Conclusion

THANK YOU!