Name	ID
حمزة حسين يوسف عمران	22011501
ياسر اشرف محمد	22010409
معاذ مصطفى عبدالحميد مصطفى	22010263
أحمد حسين حسن دويدار	20225926030
عمر محمد عبدالقادر مبروك	22010370

First we create a connection and a new database:

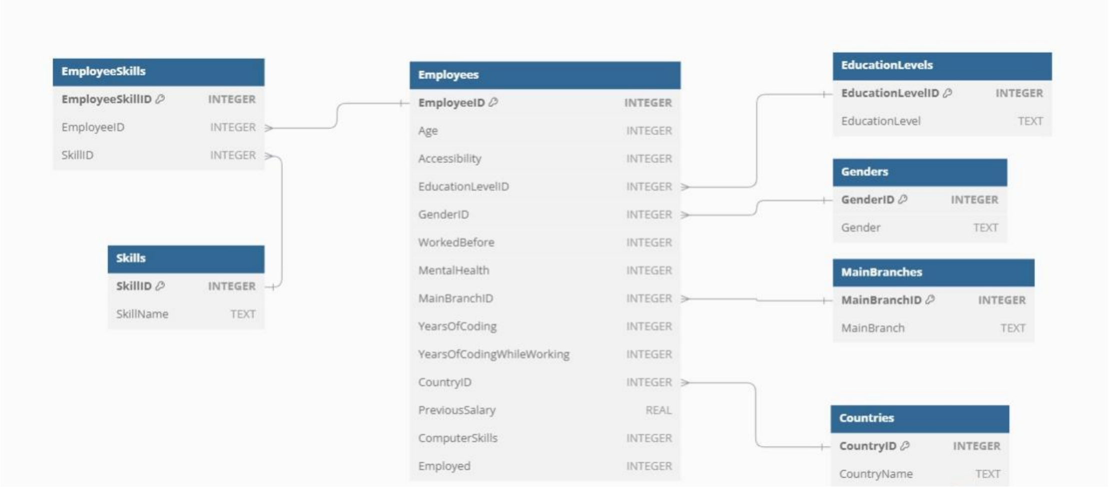
Then we have defined the cursor to do queries on the DB

Second following steps is for data migration from CSV file into DB:

1. We have created table with the same columns names for the CSV file

2. Then we have added the rows into the new created table

3. Then we have take columns and normalized the big table into more than one table and added relations between them



4. And have cleaned the Skills since it was all together so we got the unique values and inserted them into the Skills table then connected each employee with his corresponding skill in the EmployeeSkills Table

Then we create the following function for Manual HR System:

1. Add Employee
2. Update Employee
3. Delete Employee
4. Retrieve Employee

and we have give example on each function.

Then we have put the data from the database into Single Data:

Frame simulating the Market real work so we are able to analysis it flexibly.

We have imported:

1. Numpy
2. Pandas
3. Matplotlib
4. Seaborn
5. Train test split from sklearn.model selection
6. Logistic regression
7. Metrics for accuracy like accuracy score, confusion matrix
8. LabelEncoder
9. OneHotEncoder
10. StandardScalar
11. MinMaxScaler

In the EDA:

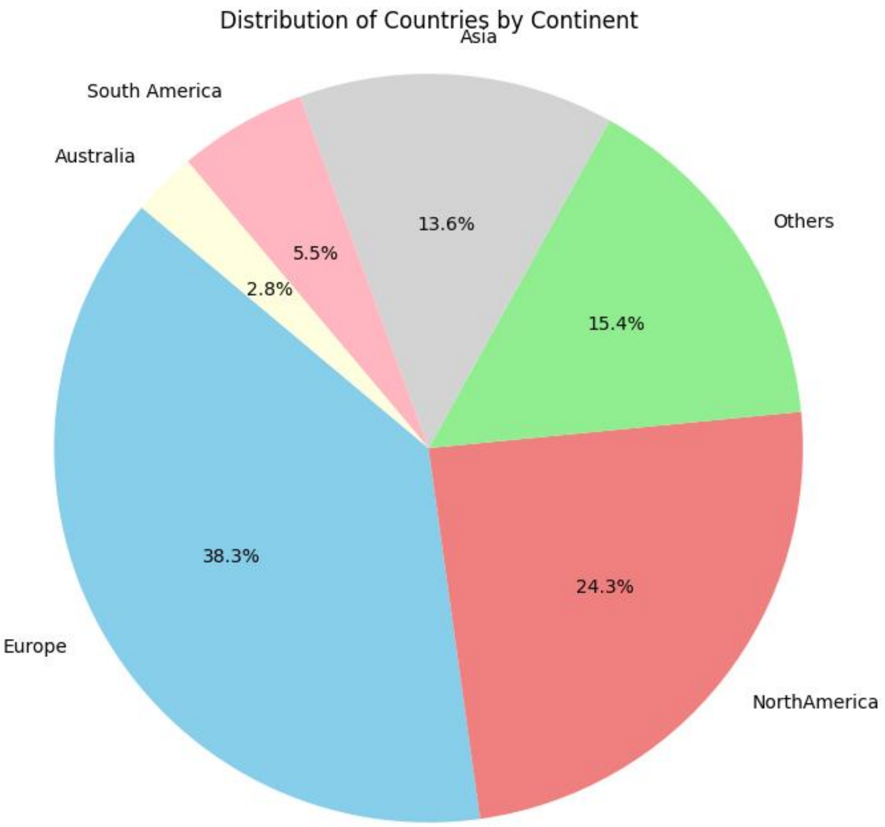
1. Then we checked df using head() built in function and get details using info() built in function and using describe() and dtypes

2. Then we got that there is no nulls
   And there is 4 duplicated rows so we dropped it

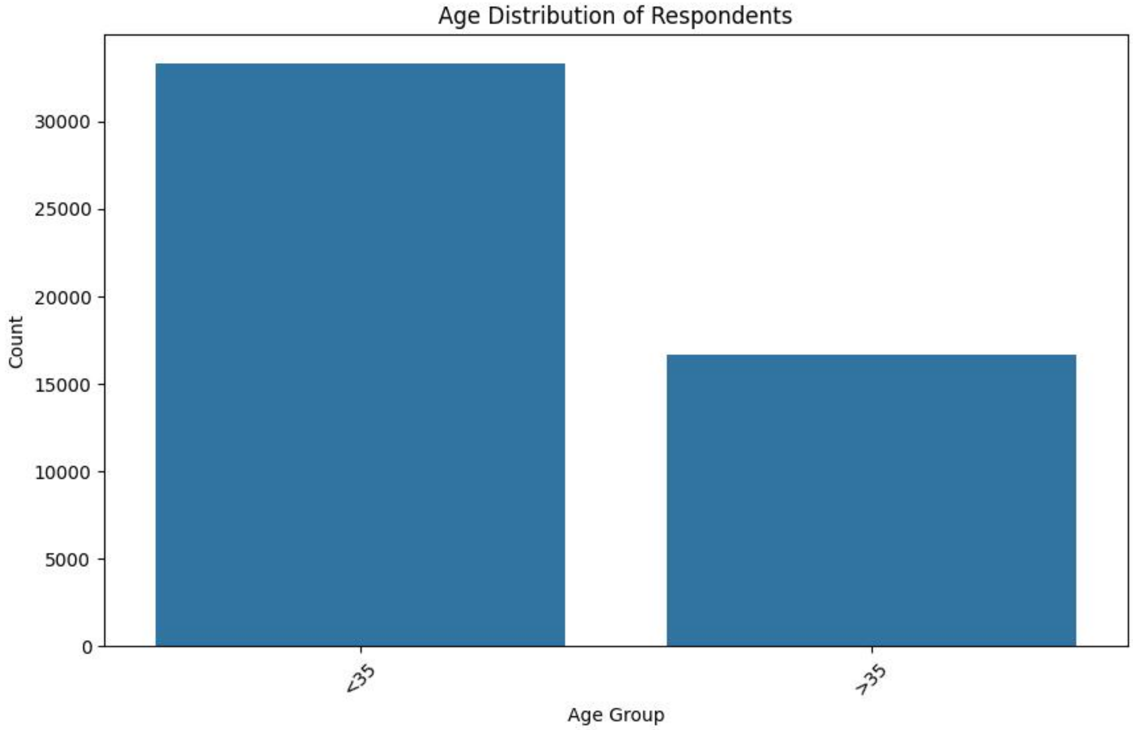
3. Then we got insights about how much unique value each column has

4. Then we made a function to assign each country to a continent and add the column Continent to the df

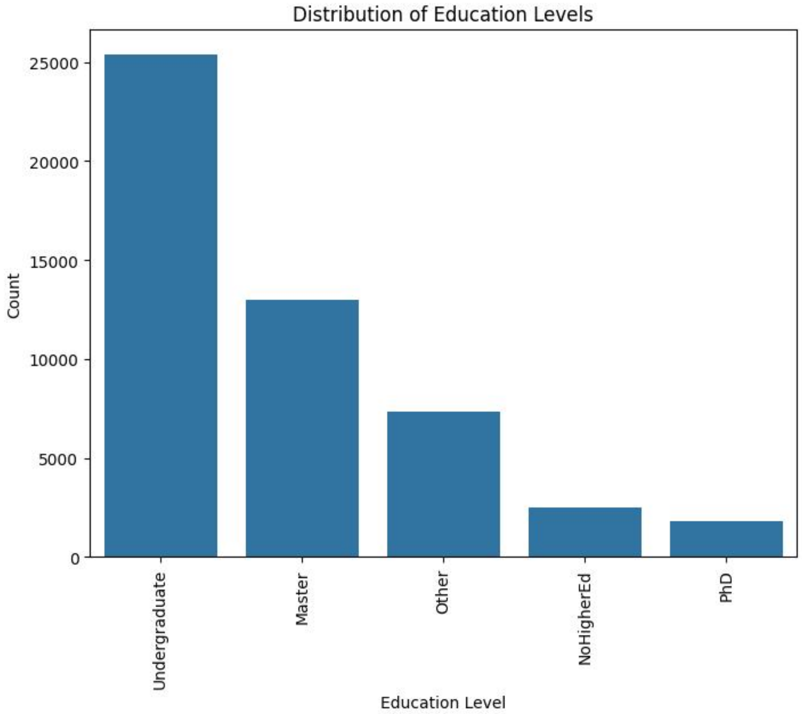
5. Then we made pie plot for the percentage of applicants from each continent



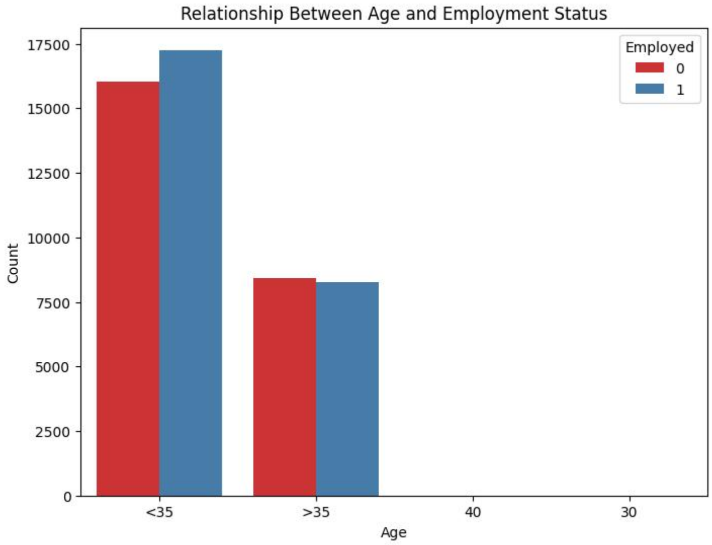
6. then we made a count plot to see the distribution of age and it shows that almost 2/3 of the applicants are over 35 years



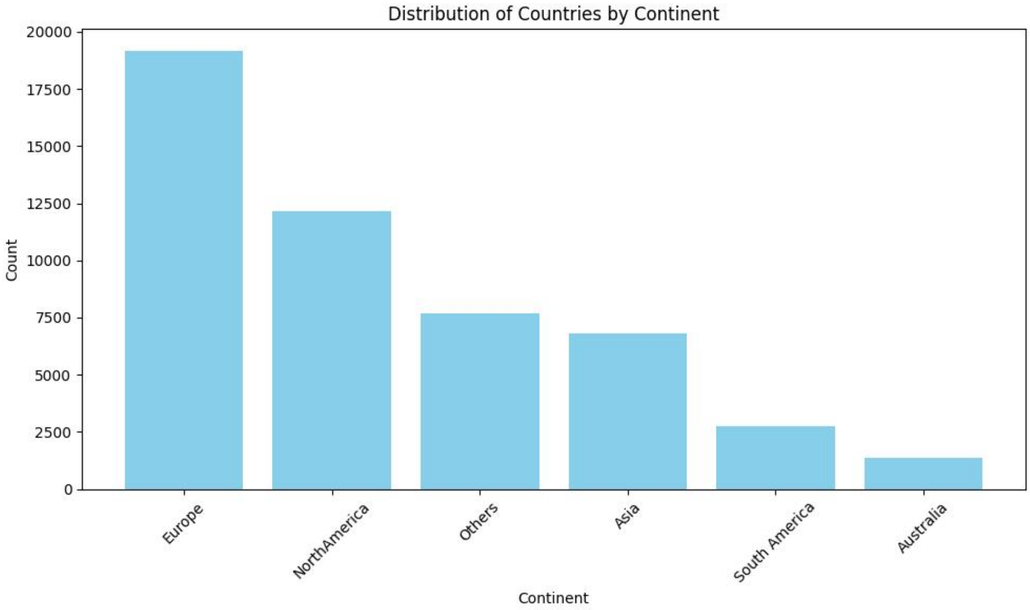
7. Then we showed the distribution of education levels and it shows that undergraduate are the most by 2 times the master then in 2nd place the master and the last is PhD



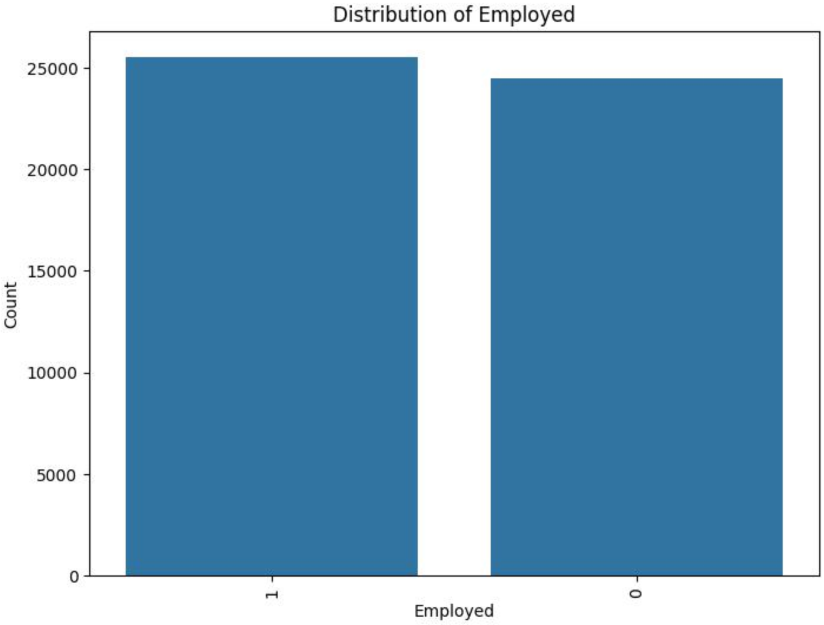
8. Then we have seen relation between age and employment in count plot and it shows that half of each groups (greater than 35 and smaller than 35) are employed



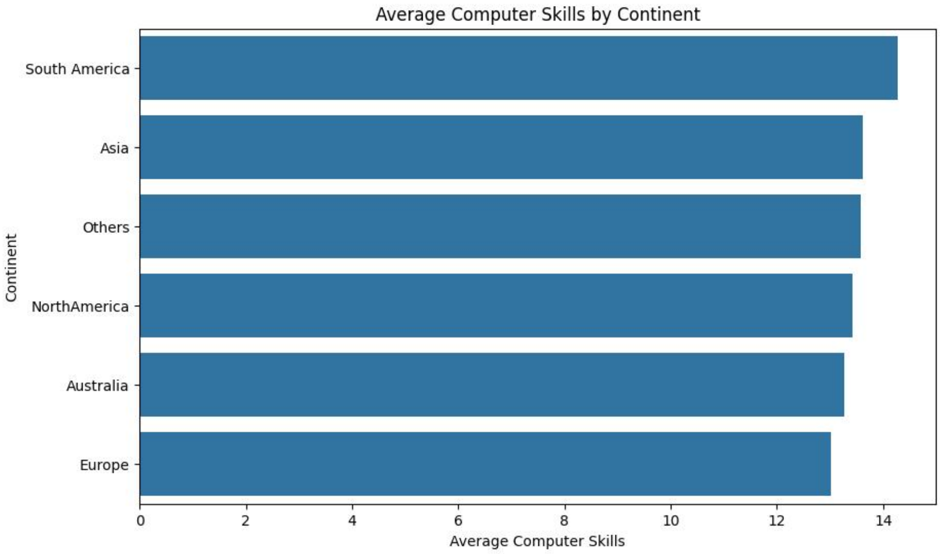
9. In the bar plot of distribution of countries by continent we recognized that most of them are from Europe then North America and the least are from Australia which make sense cause Australia is the smallest continent by population and size



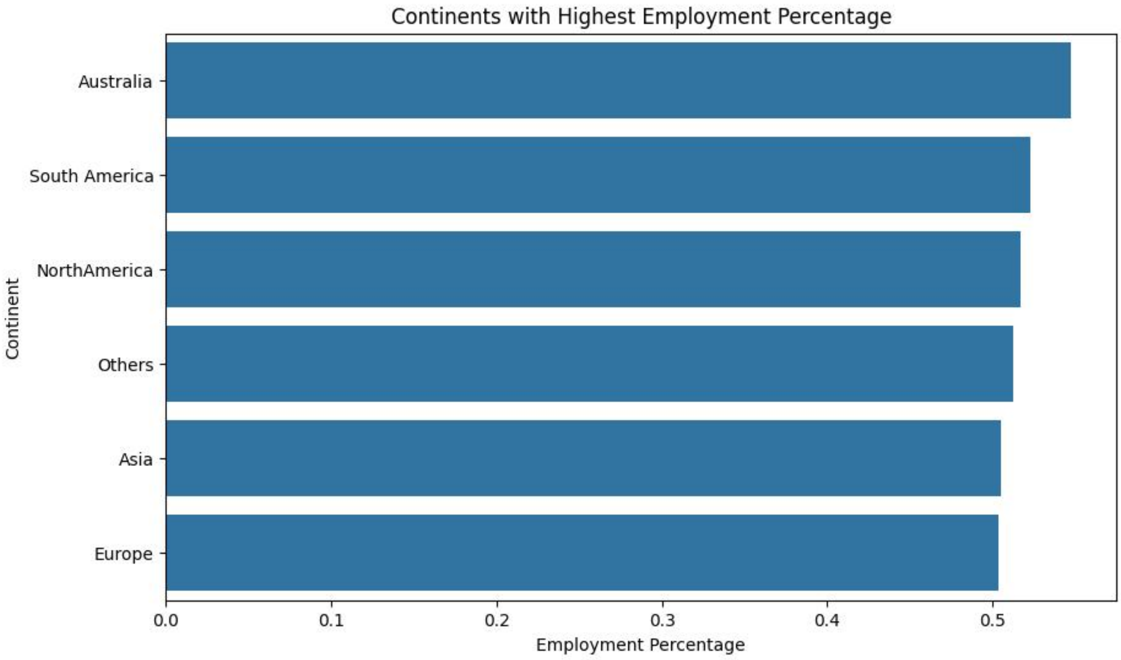
10. Then we checked the distribution of employed and figured out they are almost equal with a little higher count for employed



11. Then we visualized the average computer skills by continent and we got that south America is the most then Asia and in general all of them are quite not different regarding south America has a big increment and Europe is the least in avg number of skills



12. Then we showed the continent with higher employment percentage and we figured out that Australia is the most then south America and North America and in general they don’t have big different

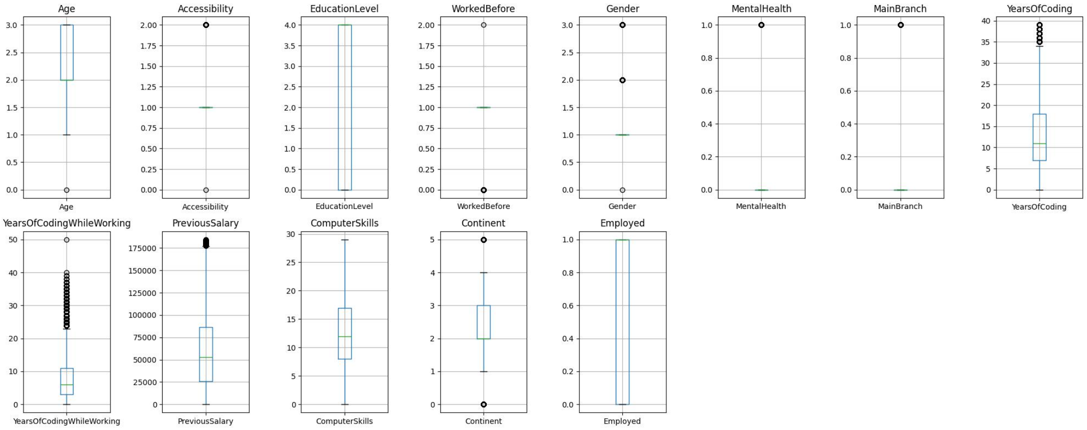


13. Then we have made a copy of df and altered it by changing the categorical columns like 'Age', 'Accessibility', 'EducationLevel', 'WorkedBefore','Gender', 'MentalHealth', 'MainBranch', 'Continent' into numerical values

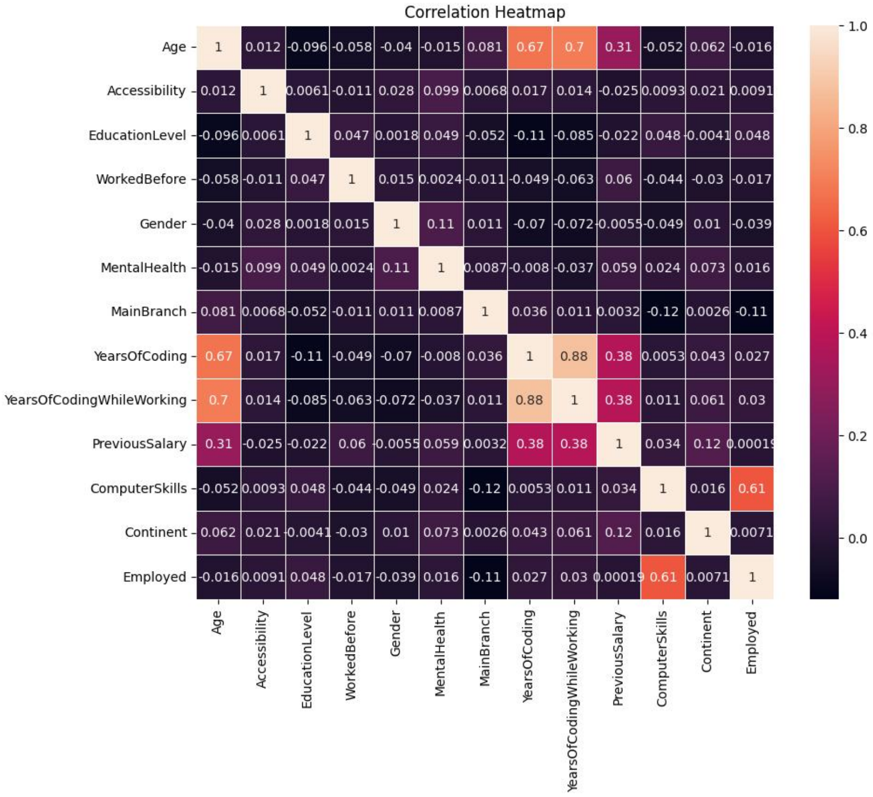
14. Then we dropped the unnecessary columns like skills and country

15. Then we have removed outliers in the following columns 'YearsCode', 'PreviousSalary,' and 'ComputerSkills'

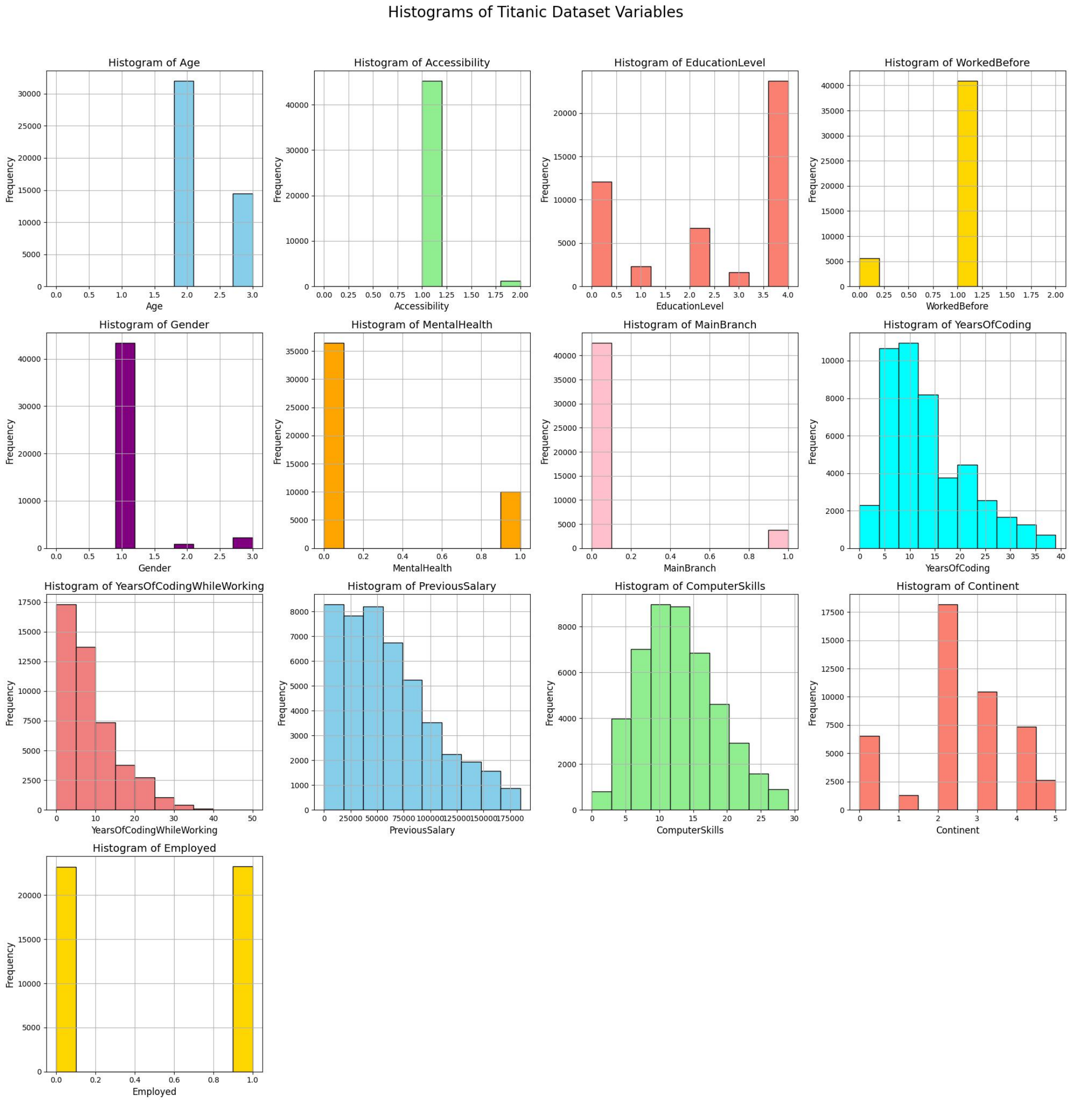
16. Then we made a box plot for each numerical column to see its distribution



17. Then we visualized correlation matrix using heatmap function in seaborn and there is big correlation between years of coding and age and years of coding while working and previous salary And we got that there is also big correlation between computerskills and employed which make sense



18. Then we visualized also the distribution of columns using histogram an the computer skills and years of coding were normal distribution and years of coding while working column and previous salary were chi squared distribution



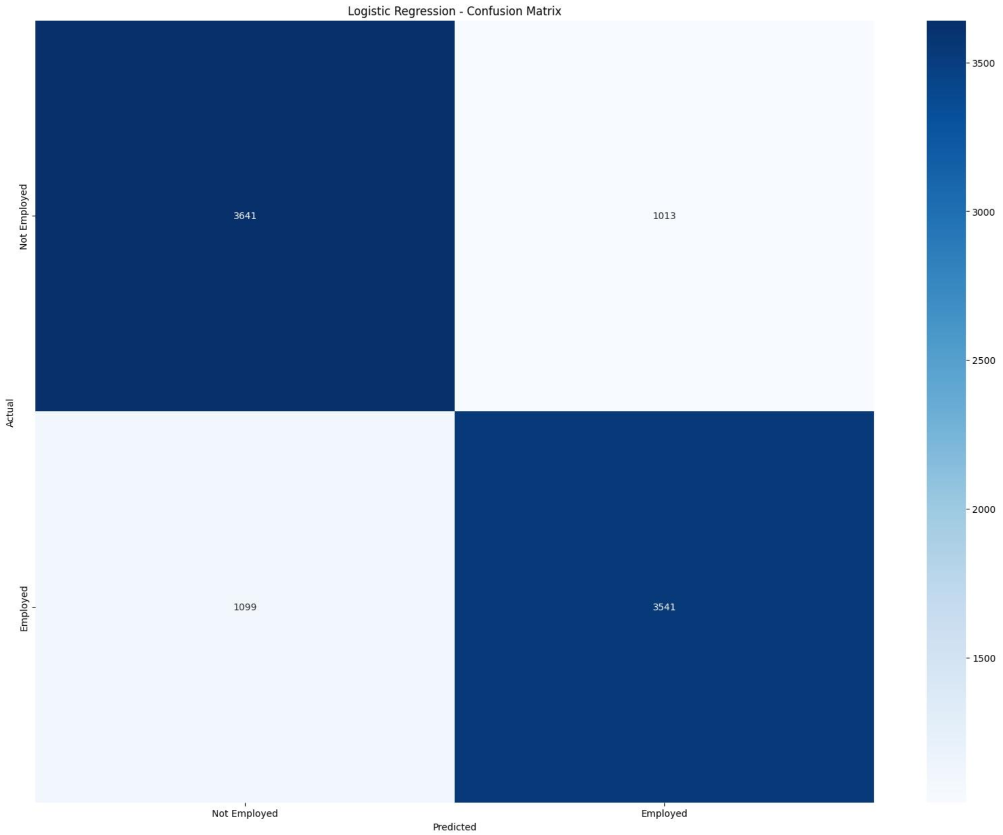
19. Then we have used ComputerSkills as input feature and the target is Employed column and we have split it into train and test

Then we done logistic regression model and fitted data and we showed the accuracy is 0.7727 which is acceptable for a support HR

In the decision tree:

we have tried different parameters like for each of the following max depth unlimited ,10,20,30 and min samples split 2,5,10 and min samples leaf 1,2,4 implementing all possible combinations w e got that the best params are
{'max_depth': None, 'min_samples_leaf': 1, 'min_samples_split': 2}
Having these measures also
F-1 Score : 0.7730794060684313
Precision Score : 0.7730794060684313
Recall Score : 0.7730794060684313

And this is the confusion matrix



In the random forest classifier:

we used these different combinations

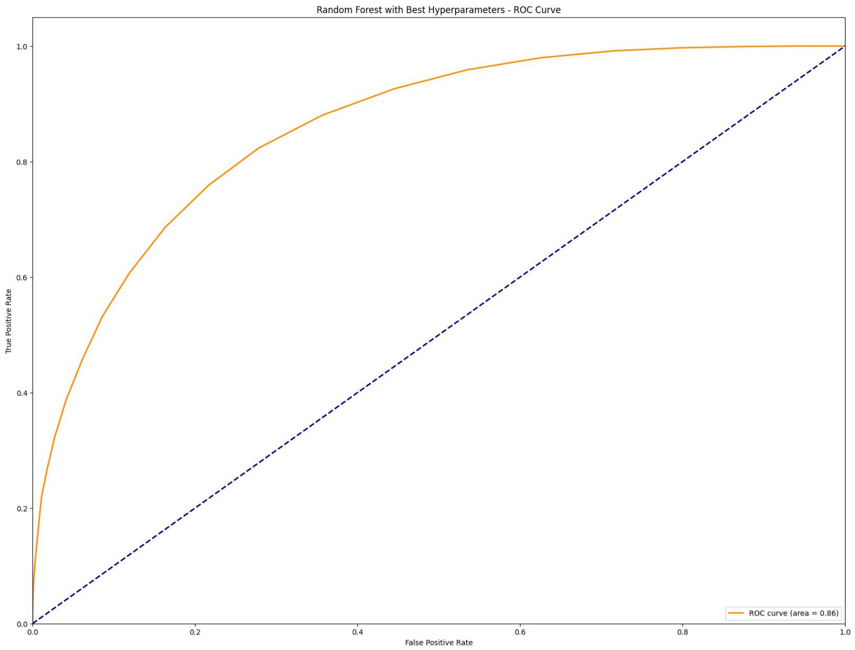
'n_estimators': [50, 100, 200],
'max_depth': [None, 10, 20, 30],
'min_samples_split': [2, 5, 10],

'min_samples_leaf': [1, 2, 4]

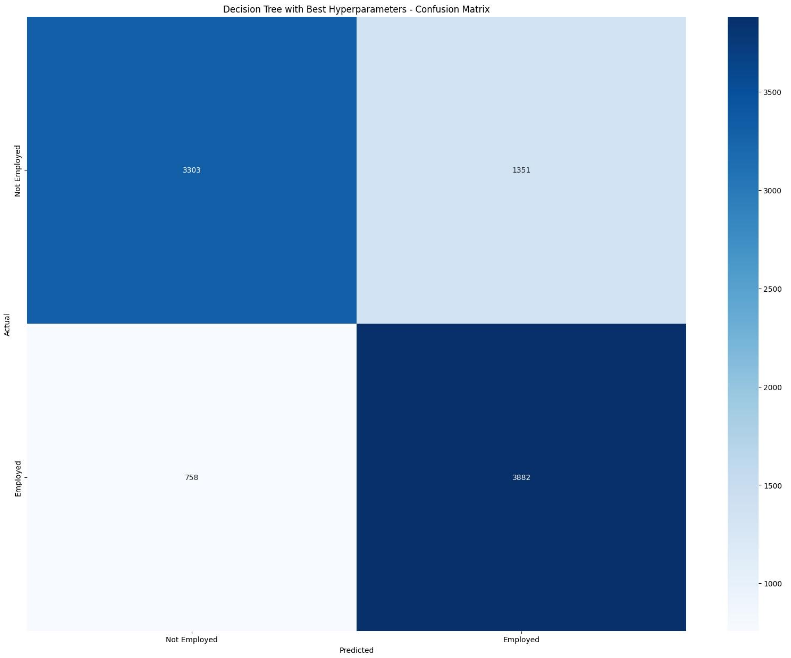
And the best combination of params is
{'max_depth': None,
'min_samples_leaf': 1,
'min_samples_split': 2,
'n_estimators': 100}
Having these measures

Accuracy: 0.7730794060684313
Decision Tree - AUC with Best Hyperparameters: 0.8580
F-1 Score : 0.7730794060684313
Precision Score : 0.7730794060684313
Recall Score : 0.7730794060684313

Then we plotted the confusion matrix and calculate the AUC and Plotted the ROC curve

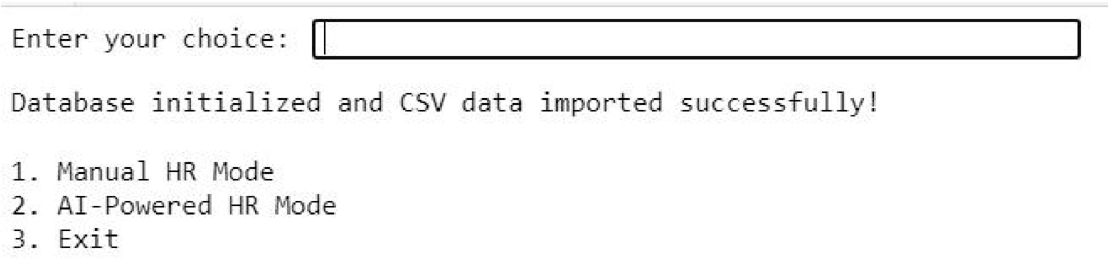


And this is the confusion matrix



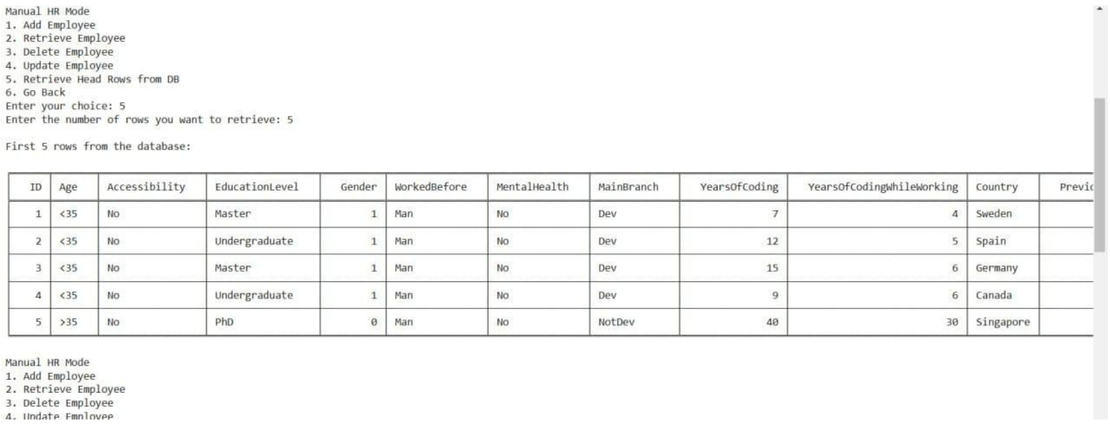
In the User Interface:

When we start the program



```
Enter your choice: 

Database initialized and CSV data imported successfully!

1. Manual HR Mode
2. AI-Powered HR Mode
3. Exit
```

If we choose Manual HR Mode it will show each option refers to i ts function by its name

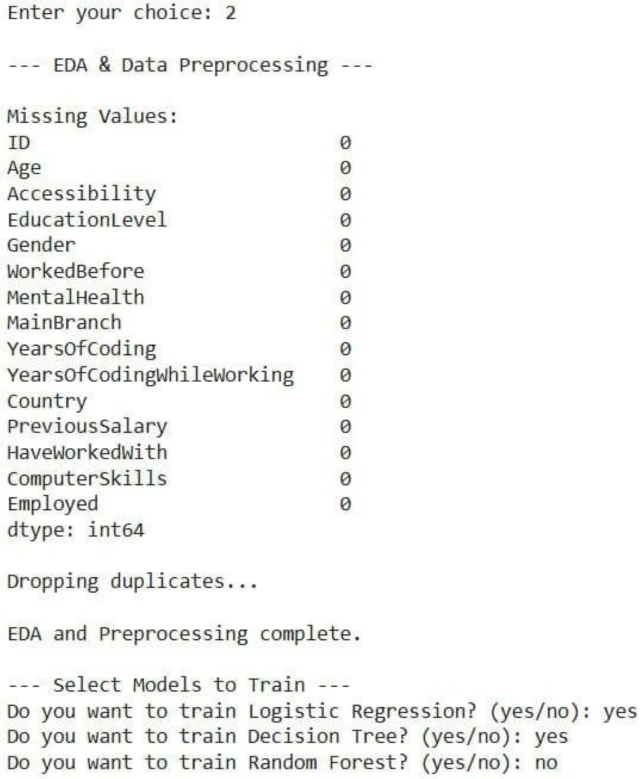


```
Manual HR Mode
1. Add Employee
2. Retrieve Employee
3. Delete Employee
4. Update Employee
5. Retrieve Head Rows from DB
6. Go Back
Enter your choice: 5
Enter the number of rows you want to retrieve: 5

First 5 rows from the database:
```

ID	Age	Accessibility	EducationLevel	Gender	WorkedBefore	MentalHealth	MainBranch	YearsOfCoding	YearsOfCodingWhileWorking	Country	Previc
1	<35	No	Master	1	Man	No	Dev	7	4	Sweden	
2	<35	No	Undergraduate	1	Man	No	Dev	12	5	Spain	
3	<35	No	Master	1	Man	No	Dev	15	6	Germany	
4	<35	No	Undergraduate	1	Man	No	Dev	9	6	Canada	
5	>35	No	PhD	0	Man	No	NotDev	40	30	Singapore	

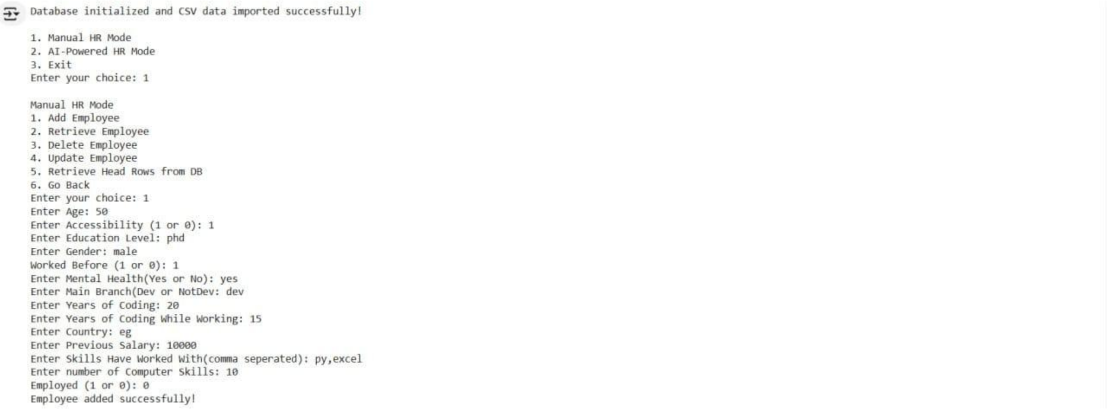
```
Manual HR Mode
1. Add Employee
2. Retrieve Employee
3. Delete Employee
4. Update Employee
```

If we choose option 2 which is AI_Powered HR Mode it will do EDA and then ask which model to apply

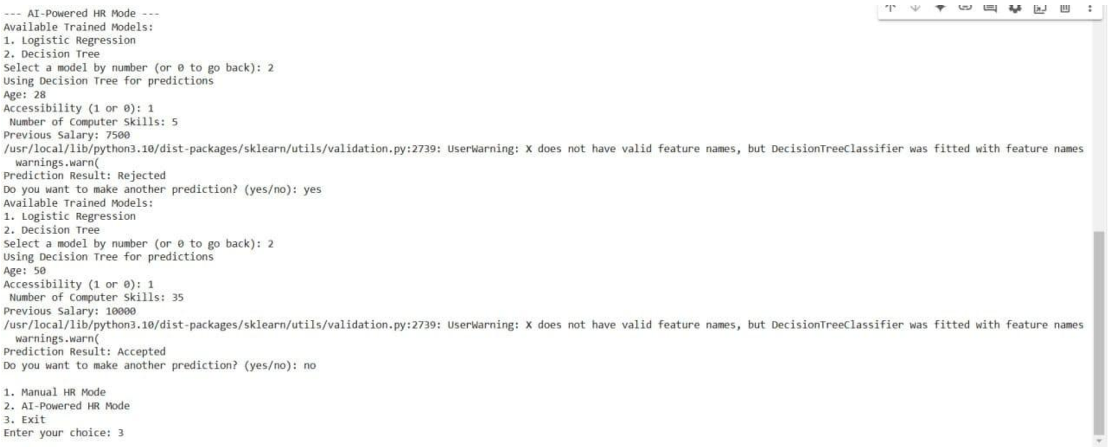


```
Enter your choice: 2

--- EDA & Data Preprocessing ---

Missing Values:
ID                           0
Age                          0
Accessibility                0
EducationLevel               0
Gender                       0
WorkedBefore                 0
MentalHealth                 0
MainBranch                   0
YearsOfCoding                0
YearsOfCodingWhileWorking    0
Country                      0
PreviousSalary               0
HaveWorkedWith               0
ComputerSkills               0
Employed                     0
dtype: int64

Dropping duplicates...

EDA and Preprocessing complete.

--- Select Models to Train ---
Do you want to train Logistic Regression? (yes/no): yes
Do you want to train Decision Tree? (yes/no): yes
Do you want to train Random Forest? (yes/no): no
```

And after choosing it will show this to take applicant row as parameters



```
Database initialized and CSV data imported successfully!

1. Manual HR Mode
2. AI-Powered HR Mode
3. Exit
Enter your choice: 1

Manual HR Mode
1. Add Employee
2. Retrieve Employee
3. Delete Employee
4. Update Employee
5. Retrieve Head Rows from DB
6. Go Back
Enter your choice: 1
Enter Age: 50
Enter Accessibility (1 or 0): 1
Enter Education Level: phd
Enter Gender: male
Worked Before (1 or 0): 1
Enter Mental Health(Yes or No): yes
Enter Main Branch(Dev or NotDev: dev
Enter Years of Coding: 20
Enter Years of Coding While Working: 15
Enter Country: eg
Enter Previous Salary: 10000
Enter Skills Have Worked With(comma seperated): py,excel
Enter number of Computer Skills: 10
Employed (1 or 0): 0
Employee added successfully!
```



```
--- AI-Powered HR Mode ---
Available Trained Models:
1. Logistic Regression
2. Decision Tree
Select a model by number (or 0 to go back): 2
Using Decision Tree for predictions
Age: 28
Accessibility (1 or 0): 1
 Number of Computer Skills: 5
Previous Salary: 7500
/usr/local/lib/python3.10/dist-packages/sklearn/utils/validation.py:2739: UserWarning: X does not have valid feature names, but DecisionTreeClassifier was fitted with feature names
  warnings.warn(
Prediction Result: Rejected
Do you want to make another prediction? (yes/no): yes
Available Trained Models:
1. Logistic Regression
2. Decision Tree
Select a model by number (or 0 to go back): 2
Using Decision Tree for predictions
Age: 50
Accessibility (1 or 0): 1
 Number of Computer Skills: 35
Previous Salary: 10000
/usr/local/lib/python3.10/dist-packages/sklearn/utils/validation.py:2739: UserWarning: X does not have valid feature names, but DecisionTreeClassifier was fitted with feature names
  warnings.warn(
Prediction Result: Accepted
Do you want to make another prediction? (yes/no): no

1. Manual HR Mode
2. AI-Powered HR Mode
3. Exit
Enter your choice: 3
```