Workshop 1 - Juan Camilo Vargas - 2216273

1. **Migration of the dataset to the database:**
   To transfer the data from candidates.csv, it was taken to the Postgresql database using the psycopg2 library. Initially, in the script, the connection to the database called ‘connect_postgres()’ was established. The host information was included (using "local"), as well as the username and password. Once the connection was established, a cursor was generated to allow browsing and modification of the set of rows in the database.

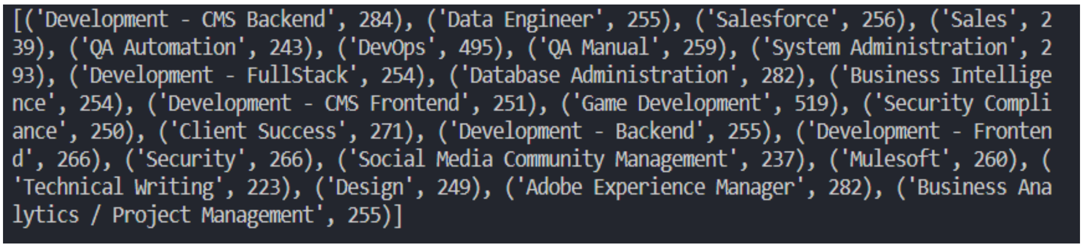
   Then i created the table for Postgres called ‘Candidates’ with the following columns and their respective data type:

   first_name: varchar
   last_name: varchar
   email: varchar
   application_date: date default current_date
   country: varchar
   yoe: integer
   seniority: varchar
   technology: varchar
   code_challenge_score: integer
   technical_interview_score: integer

   Once we have generated the table in PostgreSQL, we will proceed to import the dataset into the database by putting the exact location of the dataset using the with open() command and then cursor.copy_from() function to copy the dataset into the ‘candidates’ table with separator ‘;’ to match the exact format.

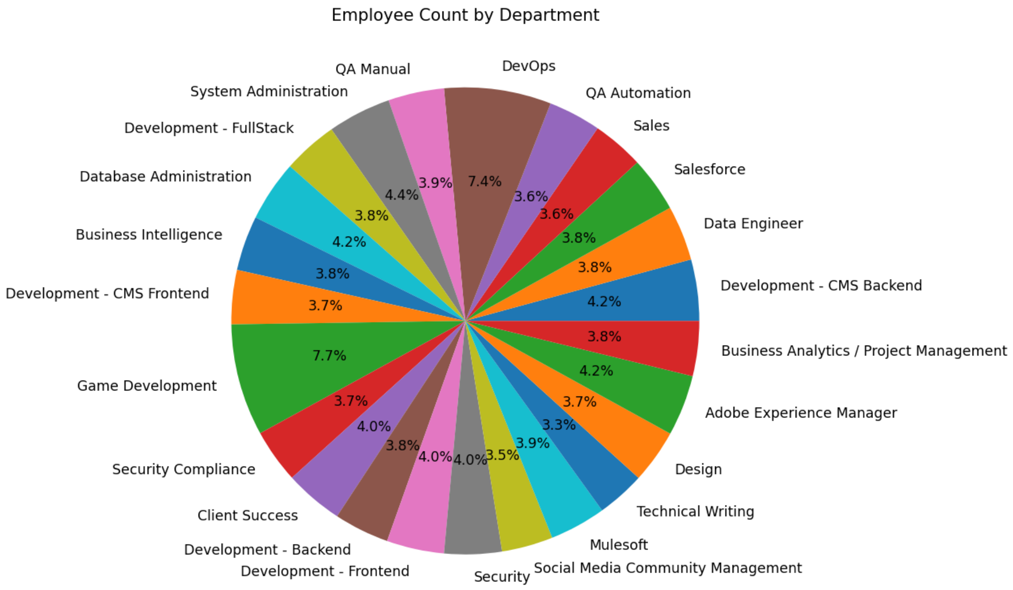
2. **Data Analysis**
   a. **Hires by Technology (Pie Plot):**
      Now for our first task, we need to select all the hired employees by ‘technology’, what I did was to select the 'technology' column, then make a count (COUNT) to all the rows of the dataset and finally make the filter of the hired employees, where I used the WHERE clause and put the columns 'code_challenge_score' and 'technical_interview_score' that the value was greater than 7.

      We can get the results from the query by using ‘results = cursor.fetchall()’:



```
[('Development - CMS Backend', 284), ('Data Engineer', 255), ('Salesforce', 256), ('Sales', 2
39), ('QA Automation', 243), ('DevOps', 495), ('QA Manual', 259), ('System Administration', 2
93), ('Development - FullStack', 254), ('Database Administration', 282), ('Business Intellige
nce', 254), ('Development - CMS Frontend', 251), ('Game Development', 519), ('Security Compli
ance', 250), ('Client Success', 271), ('Development - Backend', 255), ('Development - Fronten
d', 266), ('Security', 266), ('Social Media Community Management', 237), ('Mulesoft', 260), (
'Technical Writing', 223), ('Design', 249), ('Adobe Experience Manager', 282), ('Business Ana
lytics / Project Management', 255)]
```

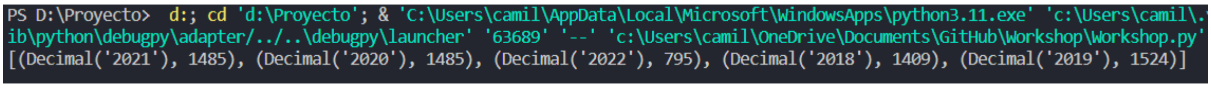
Now according to these results, we now plot:



Here for example we can see that the technology with more employees is ‘Game Development’ with 7.7% of the hired employees.

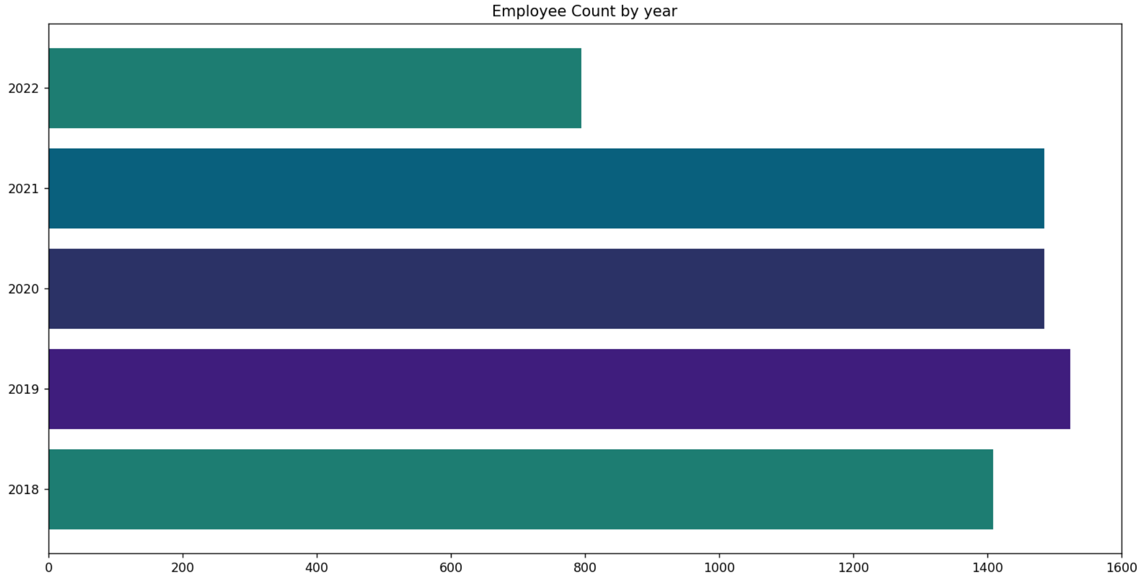
b. Hires by Year (Horizontal Bar Plot):

Now we need to select the hired employees by year, for this we will follow the same structure for the first query, but in this case we select the 'technology' column, then make a count (COUNT) to all the rows of the dataset and also select the ‘application_date’ .One of the first challenges of this query is to obtain the year, since 'application_date' has year, month and day, for this we need to use the EXTRACT clause and then the data we want which would be 'year' and for simplicity we call it Year. Finally make the filter of the hired employees, where I used the WHERE clause and put the columns 'code_challenge_score' and 'technical_interview_score' that the value was greater than 7.



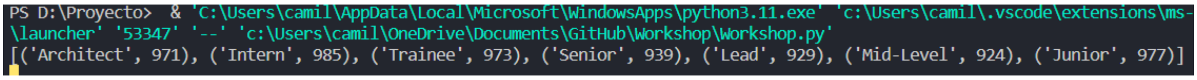
```
PS D:\Proyecto>  d:; cd 'd:\Proyecto'; & 'C:\Users\camil\AppData\Local\Microsoft\WindowsApps\python3.11.exe' 'c:\Users\camil\.
ib\python\debugpy\adapter/../..\debugpy\launcher' '63689' '--' 'c:\Users\camil\OneDrive\Documents\GitHub\Workshop\Workshop.py'
[(Decimal('2021'), 1485), (Decimal('2020'), 1485), (Decimal('2022'), 795), (Decimal('2018'), 1409), (Decimal('2019'), 1524)]
```

And now the plot:

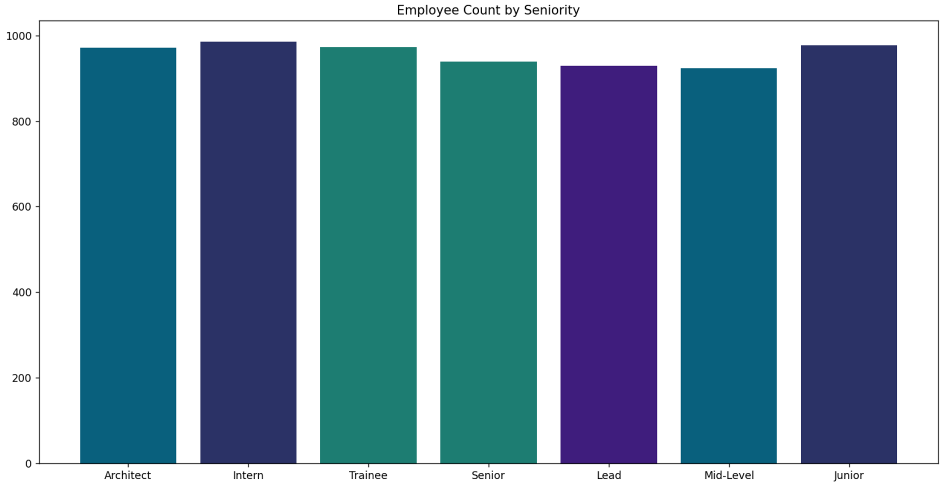


Here we can see that one of the most relevant data is the year 2022, which was the year in which the company hired the fewest personnel.

c. **Hires by Seniority (Bar Plot):**
Within this graph, each hiring is classified in 7 different levels according to their experience. It is noteworthy that the level with the highest number of hires corresponds to "Intern", which is coherent given that these are individuals who are in internships or work experience programs, and have a higher probability of being hired to gain experience.

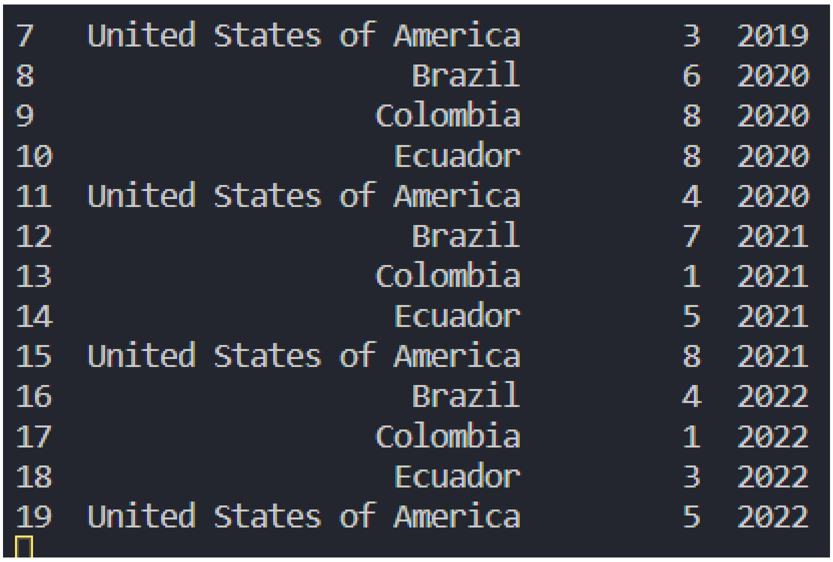


```
PS D:\Proyecto>  & 'C:\Users\camil\AppData\Local\Microsoft\WindowsApps\python3.11.exe' 'c:\Users\camil\.vscode\extensions\ms-
\launcher' '53347' '--' 'c:\Users\camil\OneDrive\Documents\GitHub\Workshop\Workshop.py'
[('Architect', 971), ('Intern', 985), ('Trainee', 973), ('Senior', 939), ('Lead', 929), ('Mid-Level', 924), ('Junior', 977)]
```



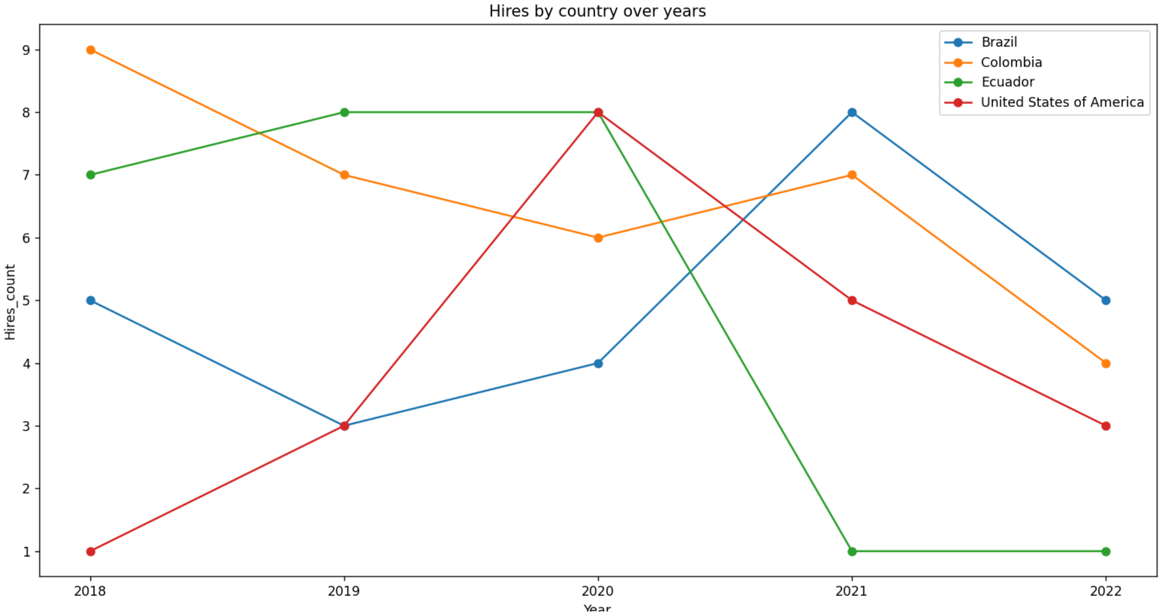
d. **Hires by Country (Multilineal Plot - Only Colombia, Brasil, Ecuador and the USA)**

In this graph we can see that it really varies a lot in all countries, they all have periods when they hire a lot of employees.



```
7   United States of America        3  2019
8                     Brazil        6  2020
9                   Colombia        8  2020
10                   Ecuador        8  2020
11  United States of America        4  2020
12                    Brazil        7  2021
13                  Colombia        1  2021
14                   Ecuador        5  2021
15  United States of America        8  2021
16                    Brazil        4  2022
17                  Colombia        1  2022
18                   Ecuador        3  2022
19  United States of America        5  2022
```

This query, like the others, has the same structure, but organized by country and year in order to be able to define the graph well.


Hires by country over years
Hires_count
Year
Brazil
Colombia
Ecuador
United States of America