Workshop 01 - ETL

Manuela Mayorga Rojas
Universidad Autónoma de Occidente
March, 2024

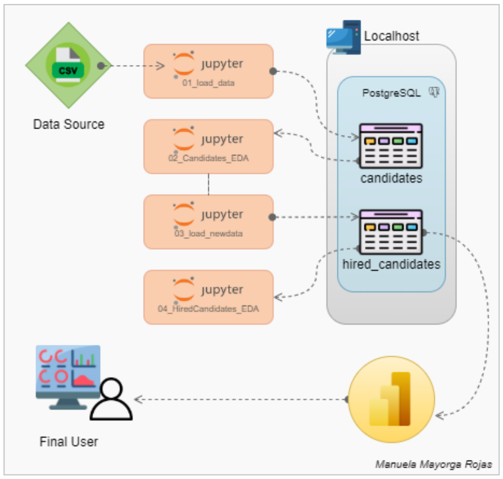
Introduction

The beginning of this work involved the analysis and manipulation of data contained in a CSV file with 50,000 rows about participant information in selection processes, generated randomly. Throughout this process, specific technologies were used as per the work instructions, including Python, Jupyter Notebook, and PostgreSQL as the relational database management system (chosen based on personal preference).

In this report, the comprehensive data migration process is detailed with a particular emphasis on the use of SQLAlchemy to manage data in the database. Additionally, the creation of visualizations in Power BI will be explored, addressing various types of charts such as pie charts, horizontal and vertical bars, as well as line charts.

Objectives:

- Migration of data from the CSV file to the PostgreSQL database using SQLAlchemy.
- Analyze and manipulate the data stored in the database.
- Create Power BI visualizations:
  - Hires by technology.
  - Hires by year.
  - Hires by seniority (level of experience).
  - Hires by country over years (USA, Brazil, Colombia, and Ecuador).



Project Steps:

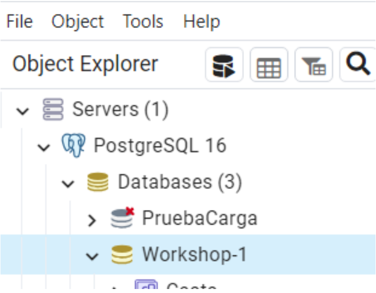
1. Database connection.
2. Exploratory data analysis.
3. Power BI Graphics.
4. Conclusions.

1. Database Connection

The main objective at this point is to establish an efficient connection with a PostgreSQL database, create tables, and import and load data from a CSV file. This begins with the definition of a function called ‘config_loader()’, responsible for establishing the connection with the database. This function uses a JSON-format configuration file named ‘db_settings.json’ to retrieve necessary parameters such as user, password, host, port, and database name.

In the project context, a crucial phase has been carried out related to the creation and management of tables in a PostgreSQL database. The implementation focuses on two main tables: ‘candidates’ and ‘hiredcandidates’, each designed to fulfill specific functions and house relevant information derived from data analysis. This division into two distinct tables allows for more efficient organization and subsequent data manipulation in a more specialized way.

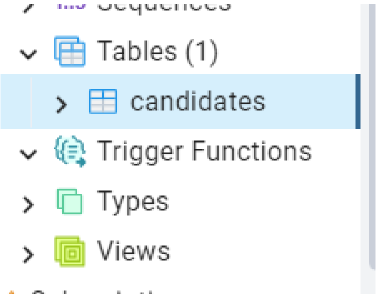
To start you need to have the database created in Postgress, in this case it is called Workshop-1.



1.1. Creation of the First Table ‘candidates’ in the Database

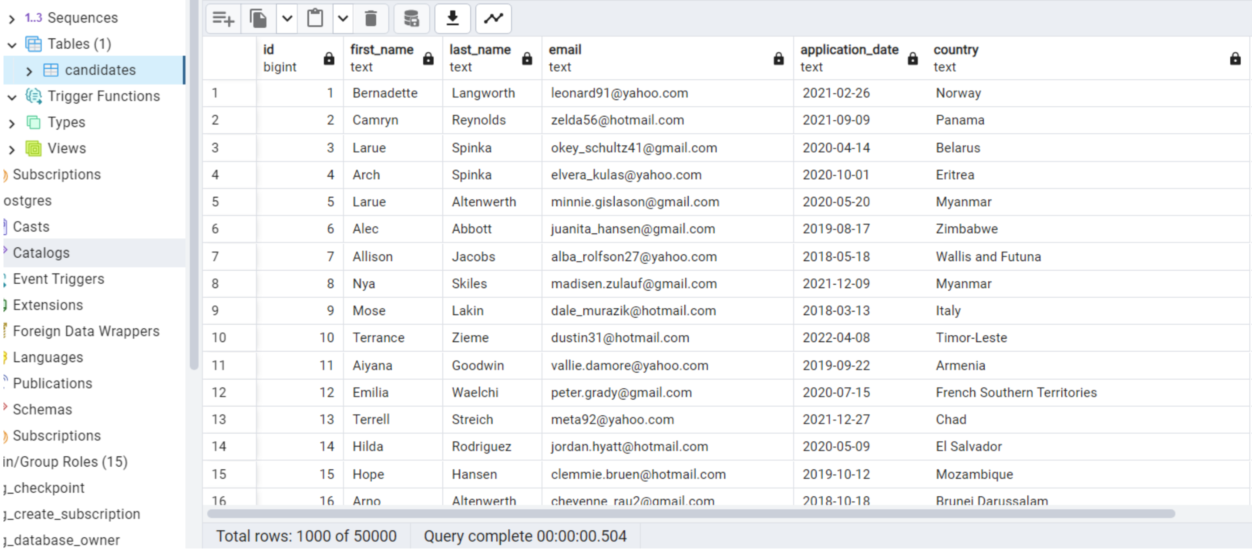
The ‘candidates’ table contains raw data obtained from a CSV file. This table is characterized by having the original information and being able to perform various data manipulation and analysis operations. The creation of this table is carried out using the SQLAlchemy library.

Subsequently, SQLAlchemy is used for the creation of the ‘candidates’ table. Declarative schema migration is performed through a class, simplifying table manipulation without the need for direct SQL queries, enabling efficient schema migration and easy interaction with the relational database.



1.2. Data Manipulation and Loading

Next, data manipulation and loading from a CSV file are performed. A class 'Processor' is used to manage file reading and operations such as correcting column names and inserting unique identifiers. The integration of SQLAlchemy is particularly evident in the use of the 'to_sql' function, which efficiently loads data into the 'candidates' table of the PostgreSQL database. This approach confirms the versatility of SQLAlchemy in working seamlessly with data manipulation operations in Python.



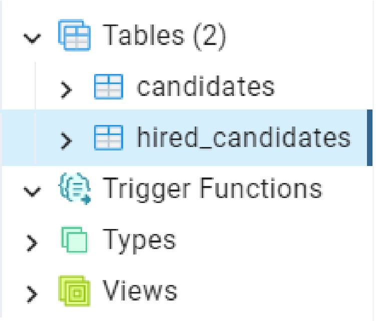
	id bigint	first_name text	last_name text	email text	application_date text	country text
1	1	Bernadette	Langworth	leonard91@yahoo.com	2021-02-26	Norway
2	2	Camryn	Reynolds	zelda56@hotmail.com	2021-09-09	Panama
3	3	Larue	Spinka	okey_schultz41@gmail.com	2020-04-14	Belarus
4	4	Arch	Spinka	elvera_kulas@yahoo.com	2020-10-01	Eritrea
5	5	Larue	Altenwerth	minnie.gislason@gmail.com	2020-05-20	Myanmar
6	6	Alec	Abbott	juanita_hansen@gmail.com	2019-08-17	Zimbabwe
7	7	Allison	Jacobs	alba_rolfson27@yahoo.com	2018-05-18	Wallis and Futuna
8	8	Nya	Skiles	madisen.zulauf@gmail.com	2021-12-09	Myanmar
9	9	Mose	Lakin	dale_murazik@hotmail.com	2018-03-13	Italy
10	10	Terrance	Zieme	dustin31@hotmail.com	2022-04-08	Timor-Leste
11	11	Aiyana	Goodwin	vallie.damore@yahoo.com	2019-09-22	Armenia
12	12	Emilia	Waelchi	peter.grady@gmail.com	2020-07-15	French Southern Territories
13	13	Terrell	Streich	meta92@yahoo.com	2021-12-27	Chad
14	14	Hilda	Rodriguez	jordan.hyatt@hotmail.com	2020-05-09	El Salvador
15	15	Hope	Hansen	clemmie.bruen@hotmail.com	2019-10-12	Mozambique
16	16	Arno	Altenwerth	cheyenne_rau2@gmail.com	2018-10-18	Brunei Darussalam

Total rows: 1000 of 50000 Query complete 00:00:00.504

The use of SQLAlchemy in this case not only simplifies database-related operations but also enhances code security, readability, and portability. Additionally, it underscores the effectiveness of SQLAlchemy as an essential tool for efficient data handling in Python applications interacting with relational databases.

1.3. Creation of the 'hired_candidates' Table

The second table, named 'hired_candidates' is an extension of the 'candidates' table and has been created to house specific information related to the hiring of candidates. Two new columns are introduced: 'hired' and 'category_of_technology,' providing a structure to classify candidates based on their hiring status and technological specialization, respectively.



1.3.1. 'hired' Column

In the structure of the 'hired_candidates' table, an additional column called 'hired' is incorporated. The inclusion of this column is based on specific instructions defined in the work guidelines. According to these specifications, a candidate is considered hired if both their 'code challenge score' and 'technical interview score' are equal to or higher than 7.

To implement this logic, a class named ‘Processor’ was designed with a specific function called ‘update_hired_column()’. Through this function, the ‘hired’ column in the ‘hired_candidates’ table is updated, evaluating the scores of the ‘code challenge’ and ‘technical interview’. If both scores are equal to or higher than 7, the value 1 is assigned to the ‘hired’ column, indicating that the candidate has been hired. Otherwise, the value 0 is assigned.

1.3.2. ‘category_of_technology’ Column

A categorization strategy has been applied to the technologies contained in the ‘hired_candidates’ table. This decision is based on observations during Exploratory Data Analysis (EDA), where 24 different technologies were identified, an amount that would be excessive for effectively representing the instruction to create a pie chart.

The implementation of this categorization was carried out through a specific function in the ‘Processor’ class called ‘technology_category()’. In this process, each technology is assigned to a predefined category, simplifying the graphical representation of technologies in the final analysis.

The categorization was as follows:

- **Development:**
  - Development, CMS Backend, Development, CMS Frontend, Development, FullStack, Development, Backend, Development, DevOps and Game Development

  Encompasses various specializations in software development, ranging from content management to the complete development of applications.

- **QA and Testing:**
  - QA Manual and QA Automation

  Includes quality assurance activities, covering both manual and automated testing to ensure software reliability.

- **Data and Analytics:**
  - Data Engineer, Business Analytics / Project Management, Business Intelligence and Database Administration

  Encompasses roles related to data management and analysis, from data engineers to business analysis professionals.

- **Sales and Marketing:**
  - Sales, Client Success and Social Media Community Management

  Includes roles in sales, customer success, marketing, as well as community management on social networks.

- **System Administration and Security:**
  - System Administration, Security and Security Compliance

  Groups functions related to system administration and the implementation of security and compliance measures.

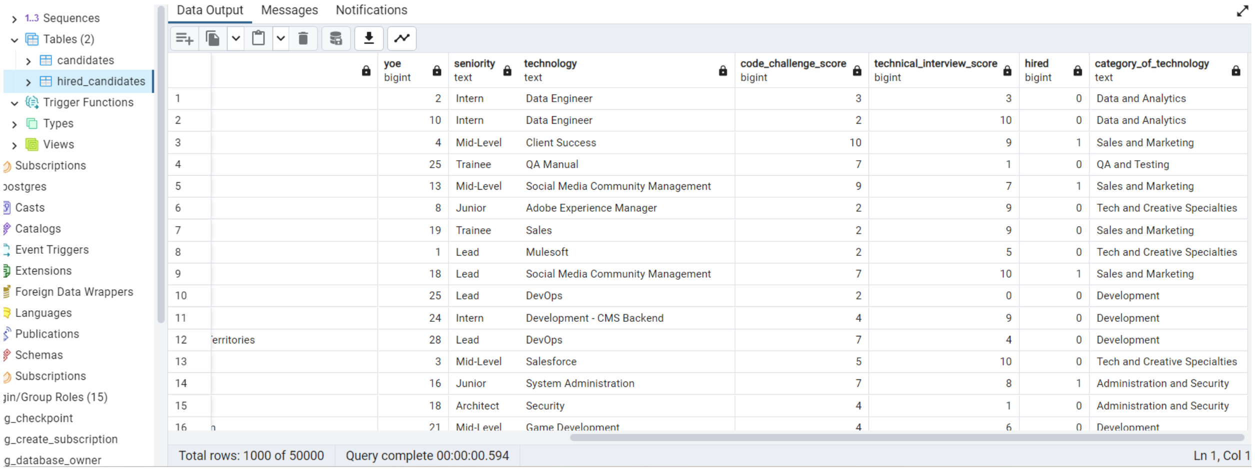
- **Tech and Creative Specialties:**
  - Mulesoft, Adobe Experience Manager, Design, Technical Writing and Salesforce

  Includes technological and creative specialties, such as system integration, content management, design, and technical writing.

This additional column, ‘category_of_technology’, enriches the ‘hired_candidates’ table by providing a more generalized and meaningful classification of the technologies used by the hired candidates. It is important to highlight that the process of creating, manipulating, and loading data into the ‘hired_candidates’ table was conducted in the same way as mentioned in points 1.1. and 1.2.

1.4. Data Loading and Updating

The process continues with the loading and updating of data in the ‘hired_candidates’ table. The ‘Processor’ class is utilized to manage CSV file reading, insertion of unique identifiers, updating the ‘hired’ column, and technological categorization. Subsequently, the data is loaded into the ‘hired_candidates’ table of the PostgreSQL database.



		yoe bigint	seniority text	technology text	code_challenge_score bigint	technical_interview_score bigint	hired bigint	category_of_technology text
1		2	Intern	Data Engineer	3	3	0	Data and Analytics
2		10	Intern	Data Engineer	2	10	0	Data and Analytics
3		4	Mid-Level	Client Success	10	9	1	Sales and Marketing
4		25	Trainee	QA Manual	7	1	0	QA and Testing
5		13	Mid-Level	Social Media Community Management	9	7	1	Sales and Marketing
6		8	Junior	Adobe Experience Manager	2	9	0	Tech and Creative Specialties
7		19	Trainee	Sales	2	9	0	Sales and Marketing
8		1	Lead	Mulesoft	2	5	0	Tech and Creative Specialties
9		18	Lead	Social Media Community Management	7	10	1	Sales and Marketing
10		25	Lead	DevOps	2	0	0	Development
11		24	Intern	Development - CMS Backend	4	9	0	Development
12	ʻerritories	28	Lead	DevOps	7	4	0	Development
13		3	Mid-Level	Salesforce	5	10	0	Tech and Creative Specialties
14		16	Junior	System Administration	7	8	1	Administration and Security
15		18	Architect	Security	4	1	0	Administration and Security
16	n	21	Mid-Level	Game Development	4	6	0	Development

Total rows: 1000 of 50000 Query complete 00:00:00.594 Ln 1, Col 1

2. Exploratory Data Analysis (EDA)

2.1 EDA made to ‘candidates’

In this phase of the work, an Exploratory Data Analysis (EDA) is carried out to understand and contextualize the information contained in the ’candidates’ table. It begins by importing essential libraries for data analysis and visualization, including matplotlib, pandas, and seaborn. Additionally, a connection to the PostgreSQL database is established using SQLAlchemy.

Data Retrieval from the Database

SQLAlchemy is used to query the ‘candidates’ table and load the data into a pandas DataFrame. This operation provides an initial insight into the structure and content of the table.

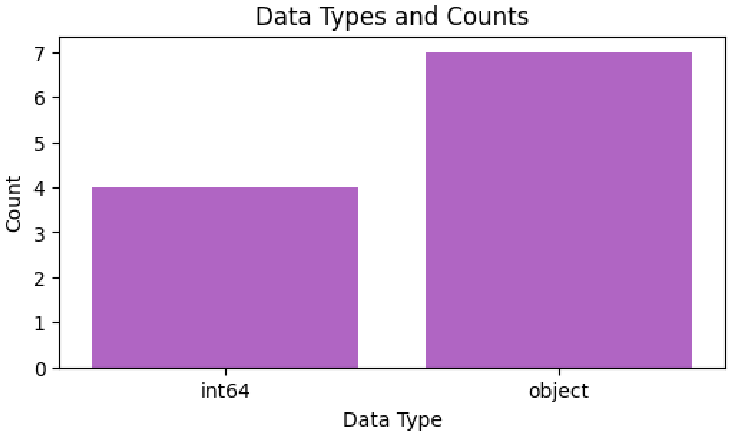
A count of rows and columns is made where we can see that the dataset has 5000 rows and 11 columns. It is important to highlight that an id was added here to have a better structure in the database. This initial exploration is essential to understand the structure and types of data present in the table.

A detailed summary of the columns present in the ‘candidates’ table is generated, providing valuable information about the types of data, the presence of null values and the number of unique values in each column.

Column Name	Data Type	Null Values	Unique Values
id	int64	0	50000
first_name	object	0	3007
last_name	object	0	474
email	object	0	49833
application_date	object	0	1646
country	object	0	244
yoe	int64	0	31
seniority	object	0	7
technology	object	0	24
code_challenge_score	int64	0	11
technical_interview_score	int64	0	11

It is evident that there are no null values, two data types, and the count of unique values per column. In the columns 'code_challenge_score' and 'technical_interview_score', it is mentioned that there are 11 unique values because the numerical scale in these columns, representing scores, ranges from 0 to 10.

Bar chart data types

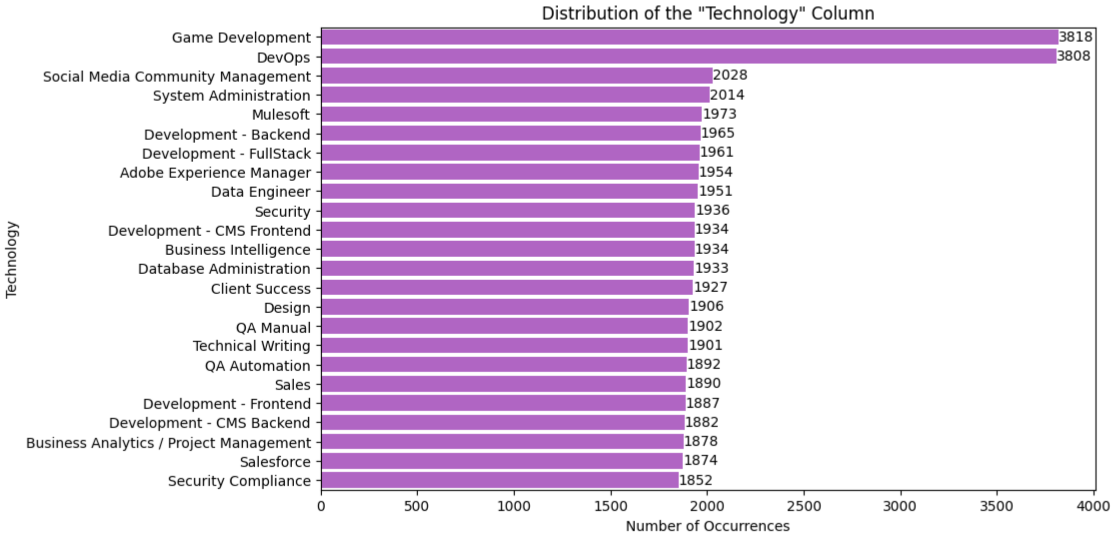


To evaluate the type of data, a bar chart was made, where it can be seen that there are 4 columns of integer type and 7 are object type.

Name Column	Count	Mean	Std	Min	25%	50%	75%	Max
id	50000.0	25000.50000	14433.901067	1.0	12500.75	25000.5	37500.25	50000.0
yoe	50000.0	15.28698	8.830652	0.0	8.00	15.0	23.00	30.0
code_challenge_score	50000.0	4.99640	3.166896	0.0	2.00	5.0	8.00	10.0
technical_interview_score	50000.0	5.00388	3.165082	0.0	2.00	5.0	8.00	10.0

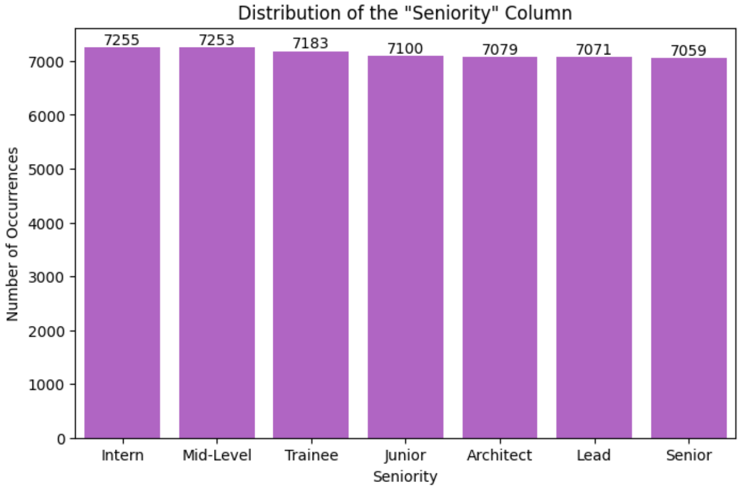
In the descriptive statistics, the following observations were made: In the 'yoe' column (years of experience), there is a mean of 15.28, indicating that, on average, candidates have approximately 15.29 years of work experience. The highest standard deviation in this column suggests greater variability in the years of experience among candidates. The 'code_challenge_score' and 'technical_interview_score' columns show scores ranging from 0 to 10. Both columns have a mean close to 5, indicating an equitable distribution of scores. The presence of minimum scores of 0 indicates that there are candidates who scored the lowest possible in both aspects. Conversely, some candidates achieved the highest score in both aspects.

Technology Graph



As mentioned earlier, a horizontal bar chart was created to observe the quantity of technologies. It reveals that the most popular technology is ‘Game Development’ with 3818 occurrences, followed by ‘DevOps’ with 3808, and ‘Social Media Community Management’ with 2028. Technologies with fewer occurrences include ‘Salesforce’ with 1874 and ‘Security Compliance’ with 1852. This information is relevant for answering questions such as: What technologies are the most popular among candidates? or Which technologies have the least demand?

Seniority Graph



As one of the charts requested in the instructions relates to ‘seniority’, I wanted to create a bar chart indicating the occurrence count for each category. It is evident that the 7 categories are very evenly distributed, as they hover around 7000 each.

2.2 EDA made to 'hired_candidates'

In this second Exploratory Data Analysis (EDA), the investigation was expanded to include two new key categories in the 'hired_candidates' table: 'hired' and 'category_of_technology.' The connection process and libraries used remained consistent with the previous analysis, ensuring coherence and comparability in the results.

The essential libraries continue to be used for data analysis and visualization. The connection to the PostgreSQL database is carried out using the same process employed in the first EDA, ensuring uniformity in data extraction. SQLAlchemy is used to query the 'hired_candidates' table and load the data into a pandas DataFrame, providing an initial insight into the structure and content of the table.

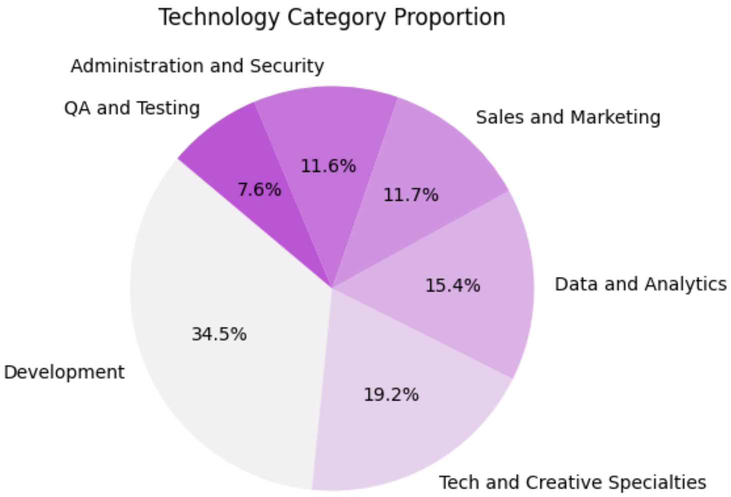
Column Name	Data Type	Null Values	Unique Values
id	int64	0	50000
first_name	object	0	3007
last_name	object	0	474
email	object	0	49833
application_date	object	0	1646
country	object	0	244
yoe	int64	0	31
seniority	object	0	7
technology	object	0	24
code_challenge_score	int64	0	11
technical_interview_score	int64	0	11
category_of_technology	object	0	6
hired	int64	0	2

As in the previous EDA we make a visualization of the data, but in this case we observe that we have 13 columns. In 'category_of_technology' we observe that some of the categories that have been created appear and in 'hired' we have a content of 0 and 1. And in this detailed summary of the table 'hired_candidates' where we see that there are no null values in any of the columns and that there are now 5 columns with the integer data type.

Column Name	Count	Mean	Std	Min	25%	50%	75%	Max
id	50000.0	25000.50000	14433.901067	1.0	12500.75	25000.5	37500.25	50000.0
yoe	50000.0	15.28698	8.830652	0.0	8.00	15.0	23.00	30.0
code_challenge_score	50000.0	4.99640	3.166896	0.0	2.00	5.0	8.00	10.0
technical_interview_score	50000.0	5.00388	3.165082	0.0	2.00	5.0	8.00	10.0
hired	50000.0	0.13396	0.340613	0.0	0.00	0.0	0.00	1.0

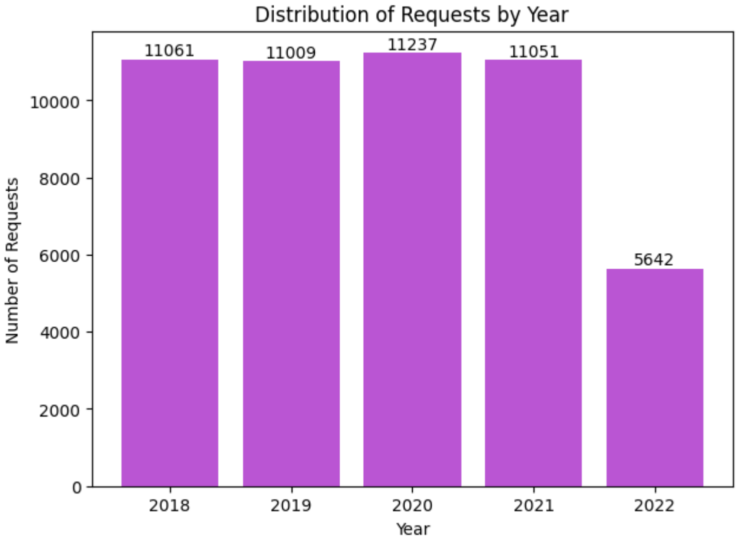
In the descriptive statistics, it was observed that the mean is 0.13396, indicating that, on average, approximately 13.4% of the candidates have been marked as 'hired'.

Technology Category pie chart



I performed a representation in a pie chart with 'category_of_technology' to assess the quantity related to these categories. It is evident that the Development category, which includes all technologies associated with development, has the highest score.

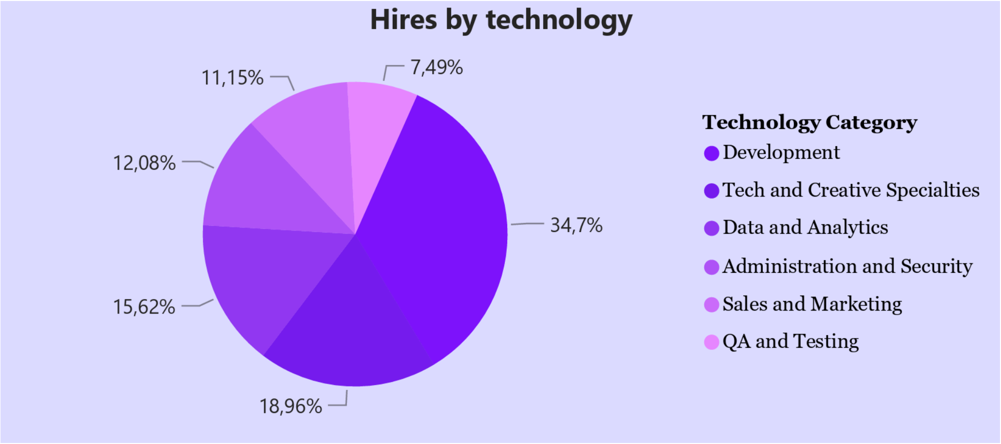
Requests by year graph



And finally, an analysis was conducted regarding the years, where it can be observed that the number of candidate applications between 2018 and 2021 is quite evenly distributed. However, in 2022, the number is below half. I counted the existing months in 2022, and there are only 7. Subsequently, an analysis was performed to determine until which date the information goes, and it precisely extends until July 4, 2022.

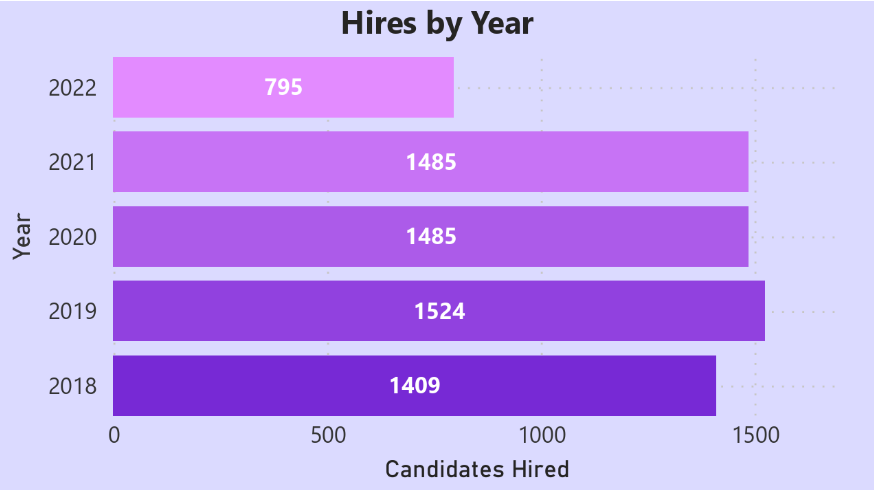
3. Charts

Hires by technology (pie chart)



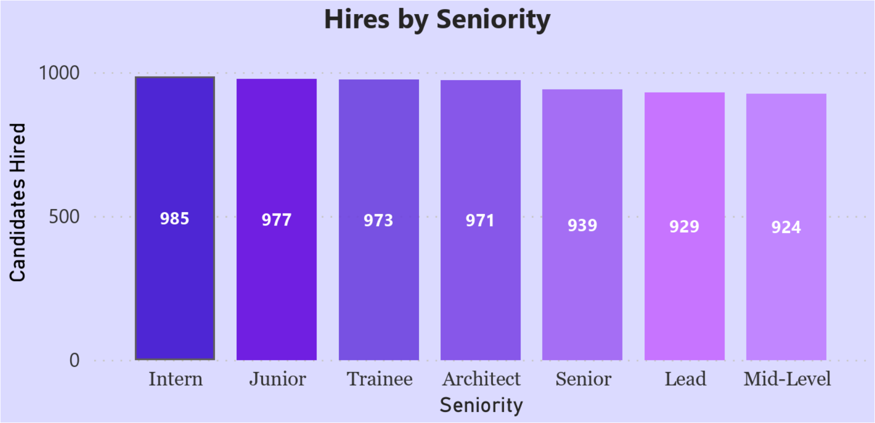
From this chart, we can analyze that the most frequent technology category is Development, accounting for 34.7% of hirings. This may be due to the grouping mentioned earlier, where Development encompasses a large number of technologies. Therefore, this number could be a result of that grouping, or it could indicate that this area requires a large number of people, hence having a higher number of hires. Another possibility is that candidates in the Development category perform exceptionally well in the tests, leading to a higher hiring rate.

Hires by year (horizontal bar chart)



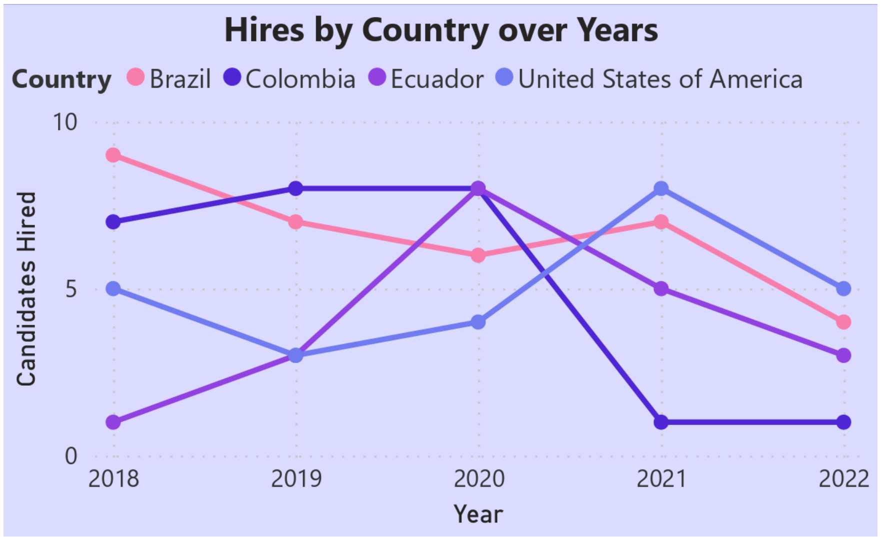
In this chart, it can be observed that the year with the highest number of hirings was 2019, with 1524 hirings, compared to the other years which maintain a more stable count, except for 2022, where there are only 795 hirings. This discrepancy in 2022 may be due to incomplete data, as mentioned earlier. The higher number in 2019 could suggest an expansion into new markets or increased hiring needs during that period.

Hires by seniority (bar chart)



In this chart, there is an observable decreasing trend in the number of hirings as the seniority level increases. The seniority levels with the highest number of hirings are ‘Intern’, while the seniority level with the lowest number of hirings is ‘Mid-Level’. This trend may be attributed to interns and juniors being more likely to be hired due to lower costs and less experience. It’s also possible that companies are actively seeking individuals for entry-level positions.

Hires by country over years (USA, Brazil, Colombia, and Ecuador)(multiline chart)



In this graph, there is an increasing trend in the number of hires in the United States, while there is also a decreasing trend in the number of hires in Colombia. This could be because the data for Colombia in 2022 is incomplete, making it difficult to draw definitive conclusions about hiring trends in Colombia due to this lack of information.

4. Conclusion

This work demonstrated the effectiveness of using Python, SQLAlchemy, and Power BI for migrating, manipulating, and visualizing data related to selection processes. The creation of specialized tables in PostgreSQL allowed for efficient information management, while Exploratory Data Analysis provided valuable insights into candidates and hires. The visualizations generated in Power BI offered a graphical representation of the data, facilitating the identification of trends and patterns.

In summary, this work underscored the importance of appropriate tools and efficient processes in data management, providing valuable information for decision-making in the field of personnel selection and hiring trend analysis.