Python Data Engineer Workshop Project Documentation

Student: Laura Ximena Reyes Arcila
Course: Data Engineering and Artificial Intelligence.
Instructor: Javier Alejandro Vergara
University: Autonoma de Occidente
Year: 2024

Introduction

This document describes the process and outcomes of the Python Data Engineer Workshop project. The main objective of this project was to develop an ETL (Extraction, Transformation, and Load) process using a dataset of job candidates,Identify which people were hired, and to create visualizations to analyze the data.

Objectives

The workshop aimed to achieve the following objectives:

Create a relational database and tables using a Python script.
Load transformed data from a CSV file into a PostgreSQL database.
Generate visualizations to analyze hires by technology, year, seniority, and country.
process
The project was completed through the following steps:

Step 1: Database Connection and Table Creation

Established a connection to a PostgreSQL database using the psycopg2 library and a JSON configuration file named finaldatabase.json.
Created a table named Candidates with appropriate columns and data types, including a boolean column IsHired to indicate if a candidate was hired based on their scores.

Step 2: Data Insertion

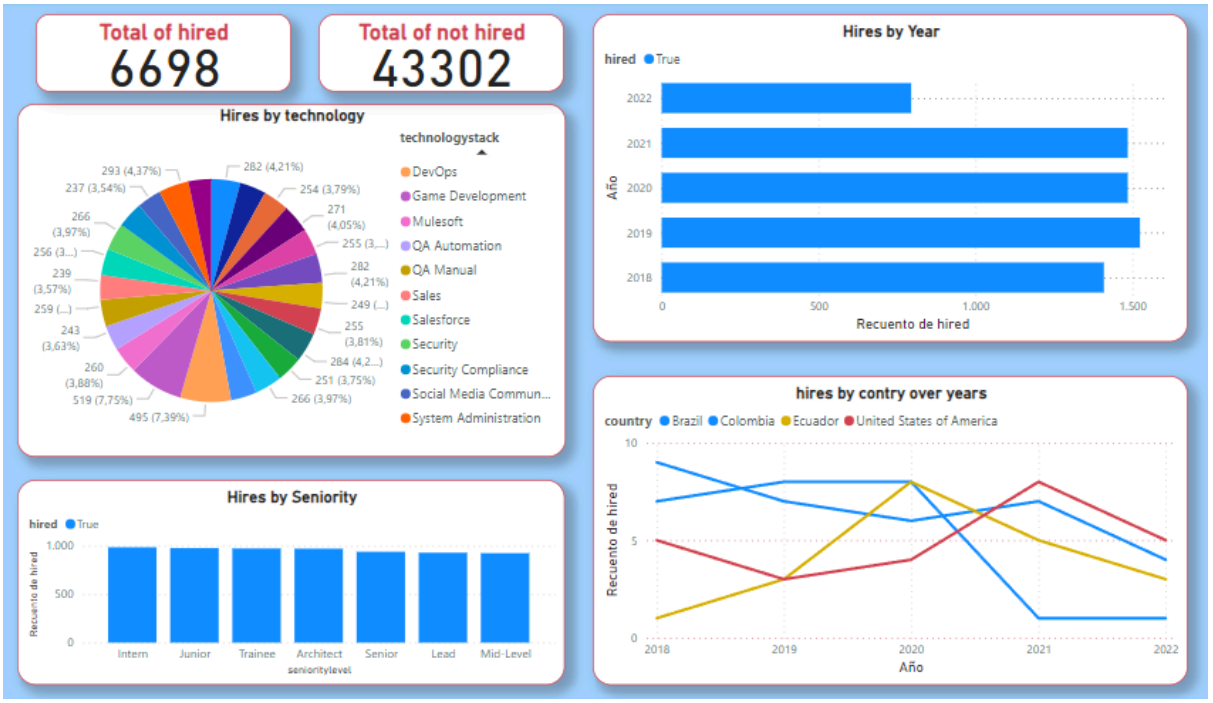
Loaded data from the candidates.csv file into a pandas DataFrame.
Inserted the data into the Candidates table, calculating the Hired status based on the Code Challenge Score and Technical Interview Score (both scores must be higher than 7 for a candidate to be considered hired).

Step 3: Exploratory Data Analysis (EDA)

Conducted an EDA to obtain a preliminary understanding of the data, examining the DataFrame's dimensions, unique values, data types, and summary statistics.

Visualizations and Analysis

The following visualizations were created using Power BI:



At the top, it's highlighted that 6,698 candidates have been hired compared to 43,302 who were not, revealing significant selectivity in the selection process.

Hires by Technology: Showed the distribution of hires across different technology categories.

We can observe that technologies with more hires have larger segments. For example, 'Game Development' and 'Mulesoft' stand out with a significant percentage, suggesting high demand for these profiles in the analyzed period. Similarly, other areas such as 'Development Operation,' 'Security,' and 'Social Media Community Management' also show a considerable number of hires.

Hires by Year : Illustrated the number of hires for each year from 2018 to 2022.

Hires by Seniority: Demonstrated the number of hires at different seniority levels within the organization.

categorizes hires based on their level of tenure, from interns to mid-level professionals. This visualization helps us understand which tenure levels are experiencing more employment and possibly reflect the company's focus on new talent or experienced professionals.

Hires by Country Over Years : Compared the number of hires from four countries (USA, Brazil, Colombia, and Ecuador) over a five-year period.

Finally, in the bottom right, the line graph 'Hires by Country Over the Years' compares hiring trends in four countries. It's an excellent way to identify regions experiencing growth or decline. We see a notable growth in Colombia and a decrease in the United

States, which could reflect changes in labor markets or the company's hiring strategies.

Conclusions

The Python Data Engineer Workshop project successfully demonstrated the ability to perform ETL processes, conduct exploratory data analysis, and create insightful visualizations. The findings from the analysis provided valuable insights into hiring trends and patterns, which can inform decision-making in recruitment and talent management.

NOTE: To go into more detail about the code, please go to the notebook finaltest.ipynb where you will find the functions explained step by step, how they work and their execution, thank you

REPOSITORY LINK: https://github.com/LauraXimenaReyes/Workshop1