Documentation Workshop 01: ETL with Python and SQL

Made by: Isabella Pérez C. - 2230603

https://github.com/isabellaperezcav/workshop_01_ETL

1. Introduction

This project aims to migrate hiring data in the technology sector from a CSV file to a MySQL relational database, followed by generating interactive visualizations in Power BI to analyze hiring trends by technology, year, experience level, and country. The final result is a dashboard that provides key insights for companies and analysts to better understand the hiring landscape in the tech industry.

The project follows an ETL (Extract, Transform, Load) approach, where data is extracted from the CSV file, transformed for analysis, and loaded into a MySQL database. Subsequently, an exploratory data analysis (EDA) is performed, and advanced visualizations are generated to facilitate decision-making.

2. Project Objectives

- Automate the migration of data from a CSV file to a MySQL database.
- Conduct exploratory data analysis (EDA) to identify hiring patterns and trends.
- Create an interactive Power BI dashboard to facilitate data visualization and analysis.
- Provide detailed and structured documentation to serve as a guide for project replication and maintenance.

3. Technologies Used

The project uses a set of modern tools to ensure an efficient, scalable, and reproducible workflow:

- **Python 3.8+**: Main language for data processing and database migration.
- **Jupyter Notebook**: Interactive environment for development and exploratory data analysis (EDA).
- **MySQL 8.0+**: Relational database management system for storing processed data.
- **Pandas 1.3+**: Python library for data manipulation, cleaning, and transformation.
- **Matplotlib 3.4+ and Seaborn 0.11+**: Python libraries for generating static visualizations during EDA.
- **MySQL Connector/Python 8.0+**: Official connector for Python-MySQL integration.
- **Power BI Desktop**: Business intelligence tool for creating interactive dashboards and advanced visualizations.

4. Project Structure

The project is organized in a modular way to facilitate understanding and maintenance. Below is the directory structure and main files:

```
workshop_01_ETL/
├── 📂 config          # Future configurations
│   ├── requirements.txt   # Project dependencies
├── 📂 data            # Input data
│   ├── candidates.csv # Dataset with candidate information
├── 📂 DB_model        # Database model and loading scripts
│   ├── Hirings.sql    # Database and table creation script
│   ├── import_csv.py  # CSV to MySQL migration script
├── 📂 notebooks       # Exploratory data analysis (EDA)
│   ├── EDA_001.ipynb  # Data cleaning, visualization, and analysis
├── 📂 venv            # Python virtual environment
├── .gitignore         # Files to exclude from version control
```

```
├── db_connection.py   # MySQL connection module
├── [README.md]        # Project documentation
```

4.1 Main Files

- `import_csv.py`: Python script that automates data migration from the CSV file to the `Candidates` table in MySQL. Reads the CSV, normalizes columns, and inserts the data.
- `db_connection.py`: Reusable module that establishes and manages the connection to the MySQL database.
- `EDA_001.ipynb`: Jupyter Notebook containing exploratory data analysis, including cleaning, descriptive statistics, and preliminary visualizations.
- `.gitignore`: Excludes sensitive files (e.g., credentials, temporary files) from version control.
- `Hirings.sql`: SQL script to create the `Hirings` database and the `Candidates` table with its defined structure.
- `desafioETL.pbix`: Power BI file with the interactive dashboard displaying key visualizations.

5. Database

5.1 Database Structure

The **Hirings** database contains a table called `Candidates`, designed to store information about candidates and their hiring details. Its structure is as follows:

Field	Data Type	Description
`id`	INT AUTO_INCREMENT PK	Unique candidate identifier
`first_name`	VARCHAR(100)	Candidate's first name
`last_name`	VARCHAR(100)	Candidate's last name
`email`	VARCHAR(255)	Candidate's email
`application_date`	DATE	Application date

`country`	VARCHAR(50)	Candidate's country
`yoe`	INT	Years of experience
`seniority`	VARCHAR(50)	Experience level (e.g., Junior, Senior)
`technology`	VARCHAR(50)	Technology associated with hiring
`code_challenge_score`	FLOAT	Code challenge score
`technical_interview_score`	FLOAT	Technical interview score

5.2 Hiring Criteria

A candidate is considered **hired** if they meet the following criteria:

```
code_challenge_score >= 7 AND technical_interview_score >= 7
```

This criterion is applied in SQL queries and analysis to filter hired candidates.

6. Data Migration

The data migration process from the CSV file to MySQL is carried out using the `import_csv.py` script. The steps are as follows:

1. **CSV Load**: The `candidates.csv` file is read using Pandas.
2. **Column Normalization**: Columns are renamed and adjusted to match the `Candidates` table structure.
3. **MySQL Connection**: The `db_connection.py` module is used to establish the database connection.
4. **Data Insertion**: Data is inserted into the `Candidates` table using an optimized SQL query.
5. **Connection Closure**: Changes are committed, and the connection is closed.

6.1 Execution Example

Ensure the database is set up (see section 9.3) and run:

```
python scripts/import_csv.py
```

7. Exploratory Data Analysis (EDA)

The exploratory analysis is conducted in the `EDA_001.ipynb` file and covers:

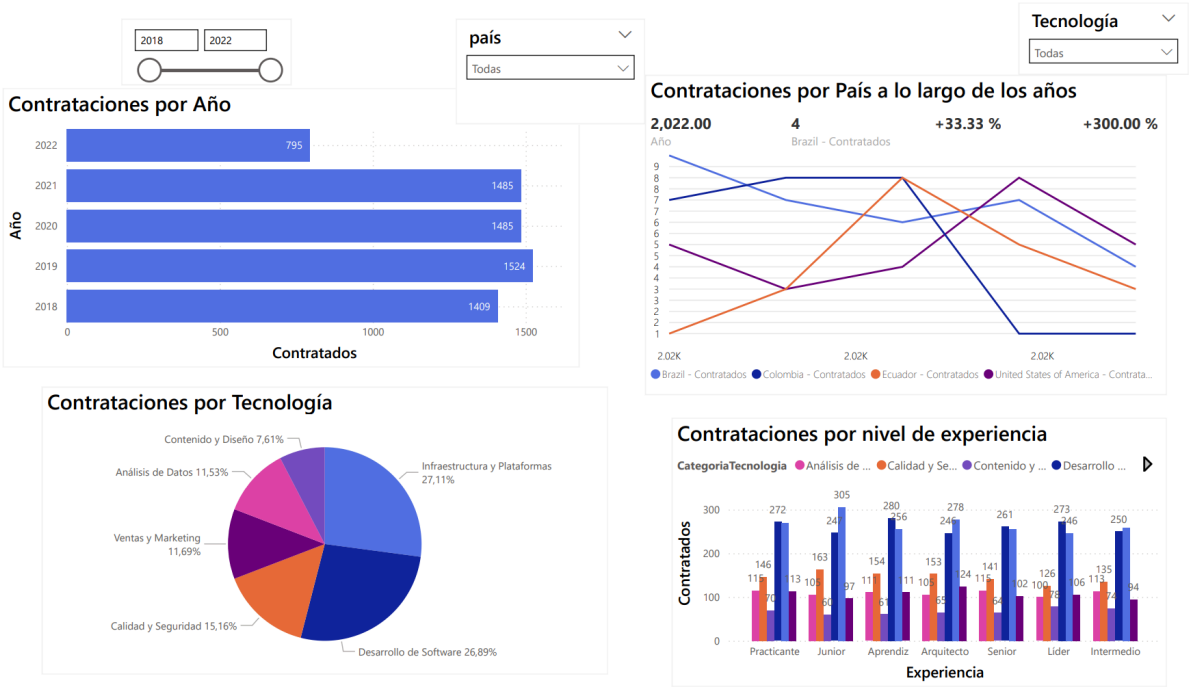
- **Data Cleaning**: Handling null values, duplicates, and outliers.
- **Descriptive Statistics**: Calculation of metrics such as mean, median, and standard deviation for numerical variables.
- **Preliminary Visualizations**: Graphs generated with Matplotlib and Seaborn to identify initial trends.

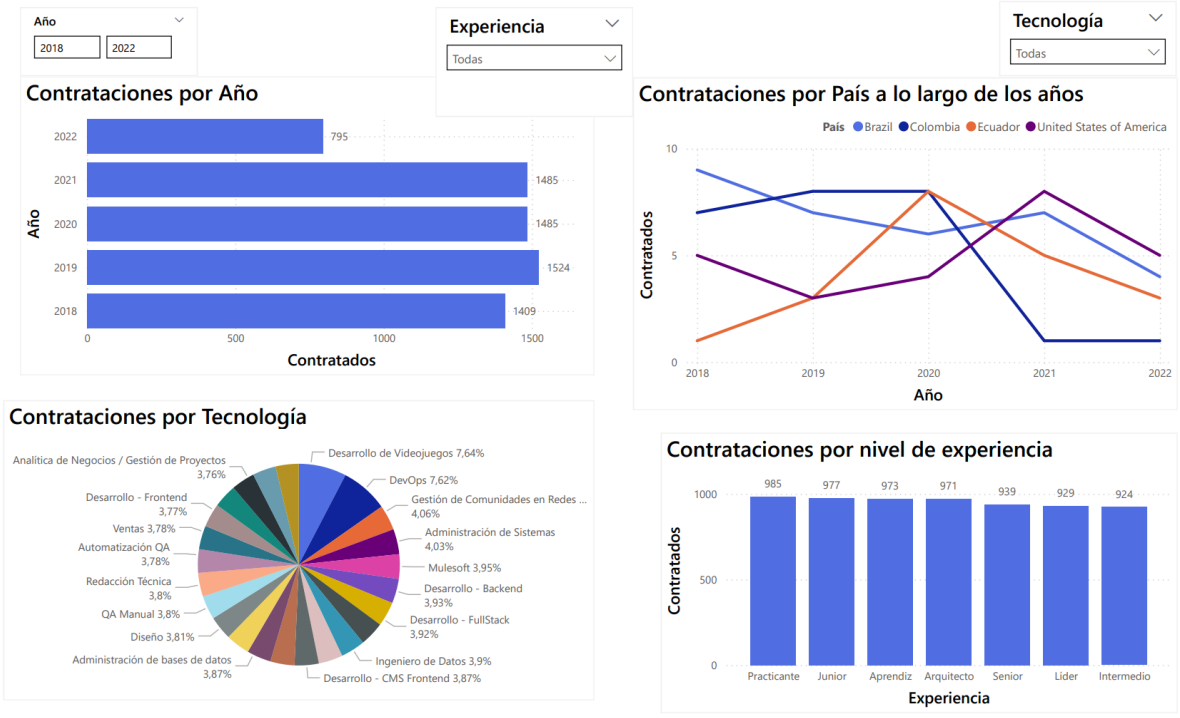
7.1 Key Findings

- Hiring rates decrease over time.
- Most candidates have less than 5 years of experience (Intern, Junior, or Trainee).
- The most in-demand technologies are **Game Development** and **DevOps**.

8. Data Visualization with Power BI

The Power BI dashboard (located in `desafioETL.pbix`) includes the following visualizations:





8.1 Hires by Year

- **Description**: Displays the number of hires per year.

- **Insight**: A declining trend in 2022, possibly reflecting labor market changes.

8.2 Hires by Country Over the Years

- **Description**: Analyzes hiring trends in Brazil, Colombia, Ecuador, and the USA.
- **Insight**: Brazil shows a post-2020 decline, while the USA experiences sustained growth.

8.3 Hires by Technology

- **Description**: Percentage distribution of hires by technology.
- **Insight**: **Game Development (7.64%)** and **DevOps (7.62%)** lead in demand.

8.4 Hires by Experience Level

- **Description**: Number of hires by seniority level.
- **Insight**: High hiring rates for **Intern, Junior, and Trainee**, indicating a focus on young talent.

8.5 Dashboard Access

- **Interactive**: Open `desafioETL.pbix` in Power BI Desktop.
- **Static**: View `DesafioETL.pdf` for a preview.

9. Setup and Execution

9.1 Prerequisites

- Python 3.8+.
- MySQL 8.0+ configured.
- Power BI Desktop installed (for interactive dashboard access).

9.2 Dependency Installation

Run:

```
pip install pandas mysql-connector...
```

9.3 Database Configuration

1. Create the database and table with:

```
mysql -u <user> -p < sql/Hirings.sql
```

1. Configure the credentials in `db_connection.py`.

9.4 Data Import

Run:

```
python scripts/import_csv.py
```

9.5 Visualization

Open `desafioETL.pbix` in Power BI or check `dashDesafioETL.pdf`.

10. Conclusions

This project demonstrates an effective ETL pipeline for migrating and analyzing technology hiring data. The combination of Python, MySQL, and Power BI enables in-depth and visually accessible analysis, highlighting trends such as the high demand for certain technologies and the hiring of young talent. It is a valuable resource for optimizing hiring strategies in the industry.