ETL project

FECHA: Friday, 21 July, 2023

1. Introduction

You have been hired by an NGO (a non-governmental organization; you choose the area of work, below you will understand :?) to enhance their Data Warehouse by incorporating public data from the internet. Your role is to identify, cleanse, and store the relevant data to facilitate better decision-making within the NGO.

2. What are the main objectives in this project?

- Cleaning, transforming and organizing the provided CSV file named *ETL_project-postal_codes.csv*. The resulting table should have the following columns:

    - id: An integer starting from 1001.
    - cp: Postal code.
    - provincia: Province.
    - ciudad: City.
    - provincia_local: Province name in the local language.
    - ciudad_local: City name in the local language.

    The fields 'provincia' and 'provincia_local' can be found together in the 'provincia' field of the original .csv separated by a slash ('/'). The same applies to the 'ciudad' and 'ciudad_local' fields.

    Attention should be given to ensuring the cleanliness and accuracy of the data. All fields need to be extracted from the .csv file. For example, if the raw data is

provincia	poblacion	Código Postal
Araba/Álava	Vitoria-Gasteiz	1001

    then, your standarized data would look like this:

id	cp	provincia	provincia_local	ciudad	ciudad_local
1003	1001	Álava	Araba	Vitoria-Gasteiz	Vitoria-Gasteiz

    Structure all the information into a pandas dataframe.

- Focus on the city of Girona and retrieve the population data for all cities within that region ordered by year. You will need to send a request using headers and endpoints to fetch the required information. This data is publicly available on the website of the National Institute of Statistics (INE) at

  https://servicios.ine.es/wstempus/js/es/DATOS_TABLA/33791?tip=AM

  Hint: The data of interest is labeled with the code "MUN" which can be found as an element within the "Metadata" list (see 0 in picture below). Filter desired data with this field.

  You need to utilize the JSON format datasource to create a table with the following columns:

  - id: A unique alphanumeric value serving as the primary key. Use the field labeled as 'COD' (see 1 in picture below).
  - population: 'Municipio' name. Use the field 'Nombre' (see 2 in picture below).
  - origin: Origin of the population censuses. Use the field 'Nombre' (see 3 in picture below).
  - years: Each column representing the population data for each year. Extract the year from the date as the key (see 4 in picture below) and the amount of population from the field 'Valor' (see 5 in picture below).

  Once you have structured the data, organize all the JSON information into a pandas dataframe. Check for duplicate information and draw conclusions!

  Observation: The number of columns (features) in the resulting data frame will depend on the data returned by the request. There will be as many columns with years as there is population information for that year.

  Important: All columns, without exception, need to be extracted from the JSON file that is returned by a request.

  Good practices: Use "try" and "except" to check whether all fields have been found and handle possible errors.

  Below you can find a screenshot of the JSON file from the INE, with hints indicating where you can find the relevant fields from which you will extract the necessary information:

```
{'COD': 'PC5662094', 1
 'Nombre': 'Total. Agullana. Total. Dato base. ',
 'T3_Unidad': 'Personas',
 'T3_Escala': ' ',
 'MetaData': [{'Id': 451,
   'Variable': {'Id': 18, 'Nombre': 'Sexo', 'Codigo': ''},
   'Nombre': 'Total',
   'Codigo': ''},
  {'Id': 4503,
   'Variable': {'Id': 19, 'Nombre': 'Municipios', 'Codigo': 'MUN'},
   'Nombre': 'Agullana', 2
   'Codigo': '17001'},
  {'Id': 16420,
   'Variable': {'Id': 431, 'Nombre': 'Países y Continentes', 'Codigo': ''},
   'Nombre': 'Total', 3
   'Codigo': ''},
  {'Id': 72,
   'Variable': {'Id': 3, 'Nombre': 'Tipo de dato', 'Codigo': ''},
   'Nombre': 'Dato base',
   'Codigo': ''}],
 'Data': [{'Fecha': '2022-01-01T00:00:00.000+01:00', 4
   'T3_TipoDato': 'Definitivo',
   'T3_Periodo': '1 de enero de',
   'Anyo': 2022,
   'Valor': 903.0}, 5
  {'Fecha': '2021-01-01T00:00:00.000+01:00', 4
   'T3_TipoDato': 'Definitivo',
   'T3_Periodo': '1 de enero de',
   'Anyo': 2021,
   'Valor': 885.0}, 5
```


- You can retrieve the population data for all regions where JSON or HTML formats are available from datos.gob.es. You can access the data catalog at

  https://datos.gob.es/es/catalogo

  You will find an option to download the complete catalog of available data. By searching for the previously used URL (for Girona) within the catalog, you can understand the desired data format. This will allow you to download the data for all provinces, for example.

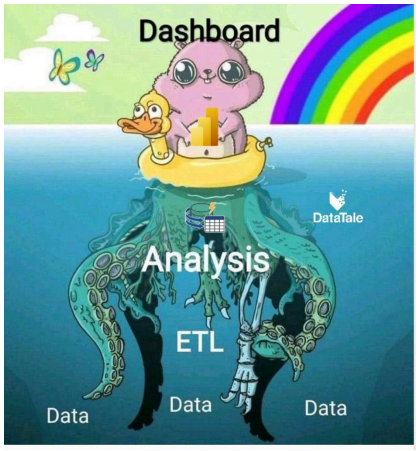
  Write a Python script that retrieves the desired information from all relevant pages within the catalog. The purpose of this script is closely related to the task in the following item.

Important: Once you have a clear understanding of the data structure, you can create a process to automate the retrieval of data from all URLs in the same format as in the previous exercise.

Hint: The file for Girona has the following identifier within the title field: `"[en]Population by sex, municipalities and nationality..."`

Observation: Which data is useful for your NGO?

- Create a Power BI Dashboard with a proposal of possible insights using relevant information obtained from data. It is not mandatory to use the data retrieved earlier, as long as your algorithm is automated to obtain interesting data from different links. **Remember:**



3. Project organization

To successfully complete this exercise, it is crucial to pay close attention to the data formats and data keys. Thoroughly investigating the data sources is necessary to understand the data structure and how to extract the desired information.

In addition, it is important to create a document (Notebook) that provides a **detailed explanation** of how the current project is **organized**. This document should be updated throughout the lifespan of the project.

The documents should include, at a minimum, the following:

- Description of the sources used: This should include information about the tables, column data formats, data types, and relationships between the tables used in the project.
- XLS/CSV output of the result of each exercise: This should include the final result of each exercise in an XLS/CSV format, allowing for easy analysis and review.

- Notebook with the commands used and the displayed results: The notebook should contain the commands executed during the project, along with the corresponding results. To limit the output of dataframes, the .head() method can be used.

- Final conclusion: This section should provide a comprehensive summary of the key learnings gained from completing this exercise.

By following these guidelines, you can effectively organize your project and ensure that important information is documented for future reference.

4. Requirements

It is crucial to include comments in the Python code. All code, including comments, must be written in English. The project should be implemented using a single Python notebook.

5. Resources

You will be using publicly available data sources such as datos.gob.es.

Please refer to the theory about the requests library for further guidance.

6. Presentation

To effectively conclude and evaluate your project, please adhere to the following guidelines:

- Submit the link to a **GitHub repository**, which should include all documents and additional files, for review by the tribunal prior to your presentation. Adhere to the deadlines set **in Google Classroom**; all materials must be submitted **a day before the defense of your work**. The tribunal will review the last commit registered on GitHub before the deadline. While you may update your project with future versions, only the last version submitted before the deadline will be considered for evaluation. It should serve as a concise yet comprehensive exposition for the evaluators, clearly explaining how you addressed and resolved the principal tasks.

- Present a **20/25-minute talk** with your Squad, highlighting the project's key aspects. The presentation time must be equally distributed among all Squad members. Choose to showcase the notebook directly or use slides with essential code snippets. Although the Dashboard is important, center the presentation in the code needed for extracting and transforming data. Describe the main aspects and insights of the project. The presentation of the Dashboard must be short and should aim to tell a compelling story about the project (both on technical aspects and insights obtained for your NGO).