Une image contenant Graphique, Police, capture d’écran, logo

Description générée automatiquement

Stéphanie Tuyêt-Hông Pragassam

IRONHACK | DATA ANALYTICS | APRIL 2024

*Understanding Socio-Economic Dynamics in France:*

Insights into Population, Establishments, Salaries, and Gender Equality

Table of content

[Introduction 2](#_Toc164704646)

[Project Management: 3](#_Toc164704647)

[High level plan: 3](#_Toc164704648)

[Project Planning: 3](#_Toc164704649)

[I. Data sources & Data collection 4](#_Toc164704650)

[1. Data from flat files 4](#_Toc164704651)

[2. Data from API 5](#_Toc164704652)

[3. Data from web scraping 5](#_Toc164704653)

[4. Data from Big Query 5](#_Toc164704654)

[II. Data cleaning and Exploratory data analysis 5](#_Toc164704655)

[III. Database type selection 5](#_Toc164704656)

[IV. Entities. ERD 5](#_Toc164704657)

[V. API 5](#_Toc164704658)

# Introduction

In today's data-driven world advanced analytics, exploring large datasets offers unparalleled opportunities to understand societal complexities. This project dives into the intricate socio-economic fabric of France, leveraging datasets provided by the French National Institute of Statistics and Economic Studies (INSEE). Through an analysis of population demographics, establishment distributions, salary structures, and job categories across French departments and regions, we seek to uncover patterns, trends, and correlations that offer valuable insights into the socio-economic landscape of France.

With access to INSEE datasets, our project aims to address fundamental questions about resource allocation, employment trends, and gender dynamics in the French labor market. By employing rigorous data analysis, our goal is to provide valuable insights to inform decision-making and drive socio-economic progress.

As we embark on this journey, we cannot ignore the issue of gender wage disparities and the underrepresentation of women in managerial positions in France. By examining salary data and analyzing the presence of women in leadership roles, we aim to shed light on persistent inequalities in the labor market. Through our analysis, we hope to contribute to discussions on gender equality and advocate for measures to promote women's advancement in the workplace.

Furthermore, our project delves into job categories to understand employment patterns and socio-economic stratification in France. By examining the distribution of individuals across different job categories and exploring the representation of women in managerial positions within these categories, we seek to uncover patterns of occupational mobility and socio-economic disparities. Our analysis aims to provide insights into the composition of the French workforce and the challenges faced by women in accessing leadership roles. From demographic shifts and employment patterns to income distributions and regional disparities, our analysis endeavors to paint a comprehensive picture of the socio-economic dynamics shaping French society.

Embark on a journey through the intricate web of data, revealing hidden insights and advancing towards a more equitable and inclusive future for all in France.

The goals of this project are to:

* Explore demographic trends and disparities in France
* Analyze salary distributions across different demographics and job categories
* Investigate the representation of women in managerial positions
* Identify correlations between population demographics and establishment data

# Project Management :

## High level plan :

* Research about project topic
* Data collection
* Project scope
* Project planning in Trello
* Exploratory data analysis in Python
  + data wrangling,
  + data cleaning
  + data visualization
* Selection and creation of a database using MySQL
* Adding data to database and create Entity Relationship Diagram
* Data manipulation in SQL
* Exposing data via API
* Visualization insights in Tableau – for presentation
* Train and test models – for presentation

## Project Planning :

Trello board to manage daily project tasks:

Une image contenant texte, capture d’écran, logiciel, Page web

Description générée automatiquement

# Data sources & Data collection

## Data from flat files

We sourced our dataset from reliable sources, including the National Institute of Statistics and Economic Studies (INSEE) and the data.gouv.fr platform.

Here's a summary of our data collection:

1. **Number of active establishments as of the end of 2020 grouped into 5 major sectors** :
   * This dataset provides information on the number of active establishments categorized into five major sectors. It offers insights into the distribution of establishments across different sectors at the end of the year 2020.
   * <https://www.insee.fr/fr/statistiques/4991205>
   * ‘etablissement.csv’
2. **Average net hourly wage by socio-professional category, gender, and age in 2020** :
   * This dataset offers details on the average net hourly wage based on socio-professional category, gender, and age in the year 2020. It enables analysis of wage disparities across various socio-demographic groups.
   * <https://www.insee.fr/fr/statistiques/2021266>
   * ‘salaire\_par\_commune.csv’
3. **Population aged 15 years and older by gender, age, and socio-professional category** :
   * This dataset presents demographic information on the population aged 15 years and older, segmented by gender, age, and socio-professional category. It allows for a comprehensive examination of the demographic composition and socio-economic characteristics of the population.
   * <https://www.insee.fr/fr/statistiques/7631680?sommaire=7632456>
   * ‘population.csv’
4. **Communes de France - Base of postal codes** :
   * This dataset, provided by La Poste, includes the names of regions, departments, and municipalities in lowercase letters, along with their respective INSEE codes. It serves as a valuable resource for geographic analysis and mapping.
   * <https://www.data.gouv.fr/fr/datasets/communes-de-france-base-des-codes-postaux/>
   * ‘communes-departement-region.csv’

We specifically chose data from the year 2020 to maintain consistency throughout our dataset, as certain data for subsequent years was not available. This decision ensures coherence, reliability and completeness in our analysis and allows for meaningful comparisons across different variables and indicators.

## Data from API

The data collection process also involved extracting relevant information from the United Nations Statistics Division SDG API, specifically focusing on target ‘5.5’, which pertains to ensuring women's full and effective participation and equal opportunities for leadership at all levels of decision-making, and on the area code ‘250’ for France.

This API enables an access to data related to the proportion of women holding managerial positions in France. By leveraging this API, we can retrieve valuable insights into the representation of women in leadership roles. This dataset from the SDG API complements our other sources of data and contributes to a more comprehensive analysis of gender equality and women's empowerment in France.

* SDG API link : <https://unstats.un.org/sdgapi/swagger/#!/Target/V1SdgTargetDataGet>
* Target for the API : <https://sdgs.un.org/goals/goal5#targets_and_indicators>
* Data from API : ‘api\_data5.5\_raw.csv’

Une image contenant texte, capture d’écran, logiciel, nombre

Description générée automatiquement

Une image contenant texte, capture d’écran

Description générée automatiquement

## Data from web scraping

For the collection of data on professions and socio-professional categories (PCS), we utilized web scraping techniques to extract information from the Wikipedia page titled "Professions et catégories socioprofessionnelles en France." Specifically, we focused on retrieving data from the third table on the page, which provides the correspondence between socio-professional categories and socio-professional groups.

The professions and socio-professional categories (PCS) serve as a statistical classification system used to categorize occupations. This classification was established by the National Institute of Statistics and Economic Studies (INSEE) in 1982. By scraping this data from Wikipedia, we were able to obtain more insights into the classification of professions and socio-professional categories, which are essential for our analysis of various socio-economic factors in France.

* URL: <https://fr.wikipedia.org/wiki/Professions_et_cat%C3%A9gories_socioprofessionnelles_en_France>
* Data from webscraping : ‘cat\_job\_webscrap\_raw.csv’

Une image contenant texte, capture d’écran

Description générée automatiquement

## Data from Big Query

# Data cleaning and Exploratory data analysis

# Database type selection

# Entities. ERD

Une image contenant texte, capture d’écran, conception

Description générée automatiquement

# API

***Certifications professionnelles link to description:***

[**RNCP link**](https://www.francecompetences.fr/recherche/rncp/37827/#ancre3)

Explain the use case

Collect data. Explain why you choose this data, from these data sources

Clean data and execute exploratory data analysis (+visualization)

Choose the database type (compare several types and explain why)

Create an entity-relationship diagram (at least 4 entities)

Create a database (database, tables)

Add data to the database

Create 5 scripts showing the insights

Expose data via API

Prepare the 10 pages report

**Requirements & Deliverables**

Each student should upload all the project materials to Github.

You should deliver:

**Planning of your project in Trello/Jira,**

Code in Python for data collection and cleaning, api

ER model,

data sources and metadata,

database script,

**report (10 pages)**

slides.

The link to the Github repository and the report in pdf format should be sent on Tuesday, **23/04/2024 till 13 PM.**

**Presentation**

The presentation time limit is 20 minutes. You will have 20 minutes for Q&A.

The slides of your presentation must include the content listed below and a demo of your project:

Title of the project + Name

Objective / Context

* Planning
* Data Gathering
* Initial Exploratory Data Analysis?
* Data Wrangling/Cleaning
* ERD
* Database Schema
* Example SQL query (Just 1/2 on one slide if at all)
* API - resources exposed/endpoints - screenshot of documentation
* Viz
* ML Process / general overview
* Feature Selection / Feature Engineering
* Handling Imbalance?
* Parameter Hypertuning
* Evaluation - evaluation metric? comparisons
* Main results
* Highlights
* Challenges
* Next steps
* Demo (If applicable)

Possible  Data sources:

<https://opendata.paris.fr/pages/home/>

<https://data.gov.uk/>

<https://data.iledefrance.fr/pages/home-open-data/>