|  |  |
| --- | --- |
|  | Data Analytics |

[Welfare and Inequality Analysis]

Hye-Jin Cho-Drugeon

Choh9323@gmail.com

June, 2023

**Table of content**

1. **Introduction**
2. **Data and data sources**
3. **Data collection**
4. **Data cleaning and Exploratory data analysis**
5. **Data base type selection**
6. **Entities. ERD**
7. **DATA Visualization**

**Introduction**

Business Use Case

Goal: This project studies intergenerational transfers such as tax and social security in the field of welfare and inequality for Euro-currency using countries with the OECD distribution data. It intends to explain the behavior of working age generation of 18-65 and old generation of above 65.

Domain: Economics and Statistics

Output

In Github [DAFT\_0410/module5 at main · chatlapin/DAFT\_0410 · GitHub](https://github.com/chatlapin/DAFT_0410/tree/main/module5)

Erdiagram\_OECD.drawio.pdf

Jira

OECD Data Cleaning and EDA.ipynb

OECD.sql

OECD.API31.ipynb

OECDRawData.ipynb

Project Toymodel.ipynb

UnitRootTestOECD meandi.ipynb

UnitRootTestOECDCI.ipynb

UnitRootTestOECDCPI.ipynb

UnitRootTestOECDP90P10.ipynb

UnitRootTestOECDTE.ipynb

UnitRootTestOECDtaxsecu.ipynb

UnitRootTestOECD transfer.ipynb

France.csv

Welfare\_DataVisualisation.xlsx

Plan

1. Planning of my project in Jira
2. Code in Python for Data collection and cleaning
3. ER Diagram
4. Data source and Meta Data
5. Database Script
6. Report (10 pages)
7. Slides

Jira <https://chatlapin.atlassian.net/jira/software/projects/CHAT/boards/1>

**Data and data sources**

**Une image contenant texte, Appareils électroniques, capture d’écran, Page web

Description générée automatiquement**

RAW data from: <https://www.oecd.org/social/income-distribution-database.htm#:~:text=The%20OECD%20Income%20Distribution%20Database%20provides%20information%20on%20the%20equivalised,households%20on%20a%20comparable%20basis>.

**Data collection**

Income components, disposable, market and primary income

From Term Reference of OECD in 2017-2018 (ref: [IDD-ToR.pdf (oecd.org)](https://www.oecd.org/els/soc/IDD-ToR.pdf), OECD project on the distribution of household incomes studies the relationship between welfare and inequality. Income distributions refer to a particular year, which should be indicated in the Excel spreadsheet “Metadata”. All income components should be reported on an annual basis and in nominal prices. Five main components of household disposable income are identified in the OECD questionnaire:

-E: employee income, including wages and salaries, cash bonuses and gratuities, commissions and tips, directors’ fees, profit sharing bonuses and other forms of profit-related pay, shares offered as part of employee remuneration, free and subsidized goods and services from an employer, severance and termination pay.1 Sick pay paid by social security should also be included.

-KI 2: capital and property income, including income from financial assets (net of expenses), income from non-financial assets (net of expenses) and royalties. Regular receipts from voluntary individual private pension plans and life insurance schemes should also be included in this income component. In line with the 2011 Canberra Handbook, capital gains should not be included in KI.

-SEI 3: income from self-employment, including profits and losses from unincorporated enterprises, as well as goods produced for own consumption (net of the costs of inputs). [The inclusion of this latter variable aims to adjust the OECD income concept to the realities of middle-income countries (such as Brazil, South Africa and others), where subsistence agriculture represents a significant income source for people at the bottom of the distribution. Countries that do not collect information on this income item should indicate so in the metadata sheet of the OECD questions

-TRR: current transfers received, including transfers from social security (including accident and disability benefits, old-age cash benefits, unemployment benefits, maternity allowances, child and/or family allowances, all income-tested and means-tested benefits that are part of social assistance, including quasi-cash transfers given for a specific purpose such as food stamps); transfers from employment related social insurance; as well as cash transfers from both non-profit institutions and other households.

-• TRP: current transfers paid, including direct taxes on income and wealth, social security contributions paid by households, contributions to employment-related social insurance, current transfers paid to both other households and non-profit institutions. Taxes on realised capital gains should be excluded from wealth taxes when possible. [Values for transfers paid should be reported in the OECD questionnaire with a negative sign].

While relevance and data availability for the sub-components of current transfers will vary across countries (depending on the structure of their social protection system and on features of their micro-data), this more detailed breakdown allows better reflecting the situation of countries with an important employment-related pension pillar.

• In the case of current transfers received (TRR):

-TRRSS: current transfers received from social security.

− TRRER: current transfers received from employment-related social insurance schemes (e.g. occupational pensions), where such schemes meet at least one of the following conditions: i) participation is obligatory; ii) the scheme is collective; and iii) the employer makes a contribution on behalf of an employee. 4

-− TRROT: current transfers received from non-profit institutions and other private households, e.g. alimonies.

• In the case of current transfers paid (TRP):

TA: direct taxes on income and wealth paid by households (net of refunds), as well as contributions paid by households to public social security schemes.

− TRPER: contributions paid by households to employment-related social insurance schemes (as defined above). − TRPOT: current transfers paid by households to non-profit institutions and other households, e.g. alimonies.

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

Description générée automatiquement

Essential equations related to this study are:

1. Disposable income from primary income, market income, gross income (ref: [IDD-ToR.pdf (oecd.org)](https://www.oecd.org/els/soc/IDD-ToR.pdf). P.5)

Une image contenant texte, capture d’écran, Police

Description générée automatiquement

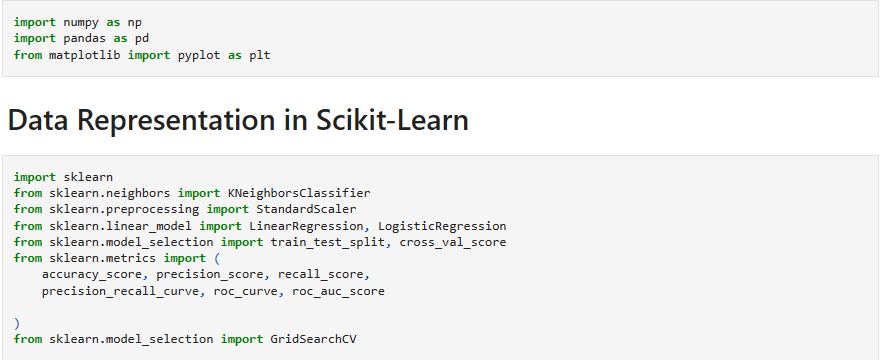
1. Gini index (ref: [IDD-ToR.pdf (oecd.org)](https://www.oecd.org/els/soc/IDD-ToR.pdf). P.8)

Une image contenant texte, capture d’écran, Police, ligne

Description générée automatiquement

**Data cleaning and Exploratory data analysis**

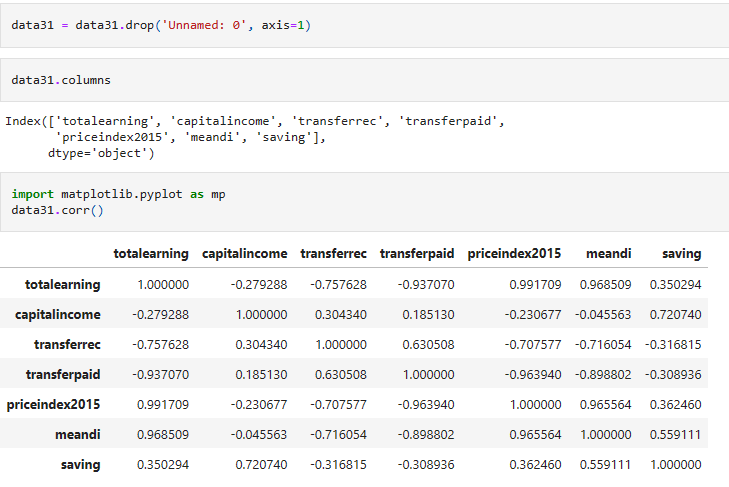
Starting to import all necessary libraries, This code imports necessary modules for data analysis and machine learning, including pandas for data manipulation, matplotlib and seaborn for data visualization, re for regular expressions, numpy for numerical computing, and scikit-learn for machine learning. • The **%matplotlib inline** command sets the backend of matplotlib to the 'inline' backend, which allows the plots to be displayed in the Jupyter notebook. • The **sns.set()** command sets the default style of the plots to the Seaborn style. • The code then imports the training and test data from CSV files using pandas' **read\_csv()** function and stores them in dataframes **df\_train** and **df\_test**, respectively.

****

(1) Toy model: France

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

Description générée automatiquement

****

The data with key variables in France have variance with different range of mean values. Except for paid transfers, variables are greater than 49. For price index in 2015, the index is defined as percentage but mean values are 100 which is ideal for accommodating current market prices.

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

Description générée automatiquement

Une image contenant texte, capture d’écran, Police

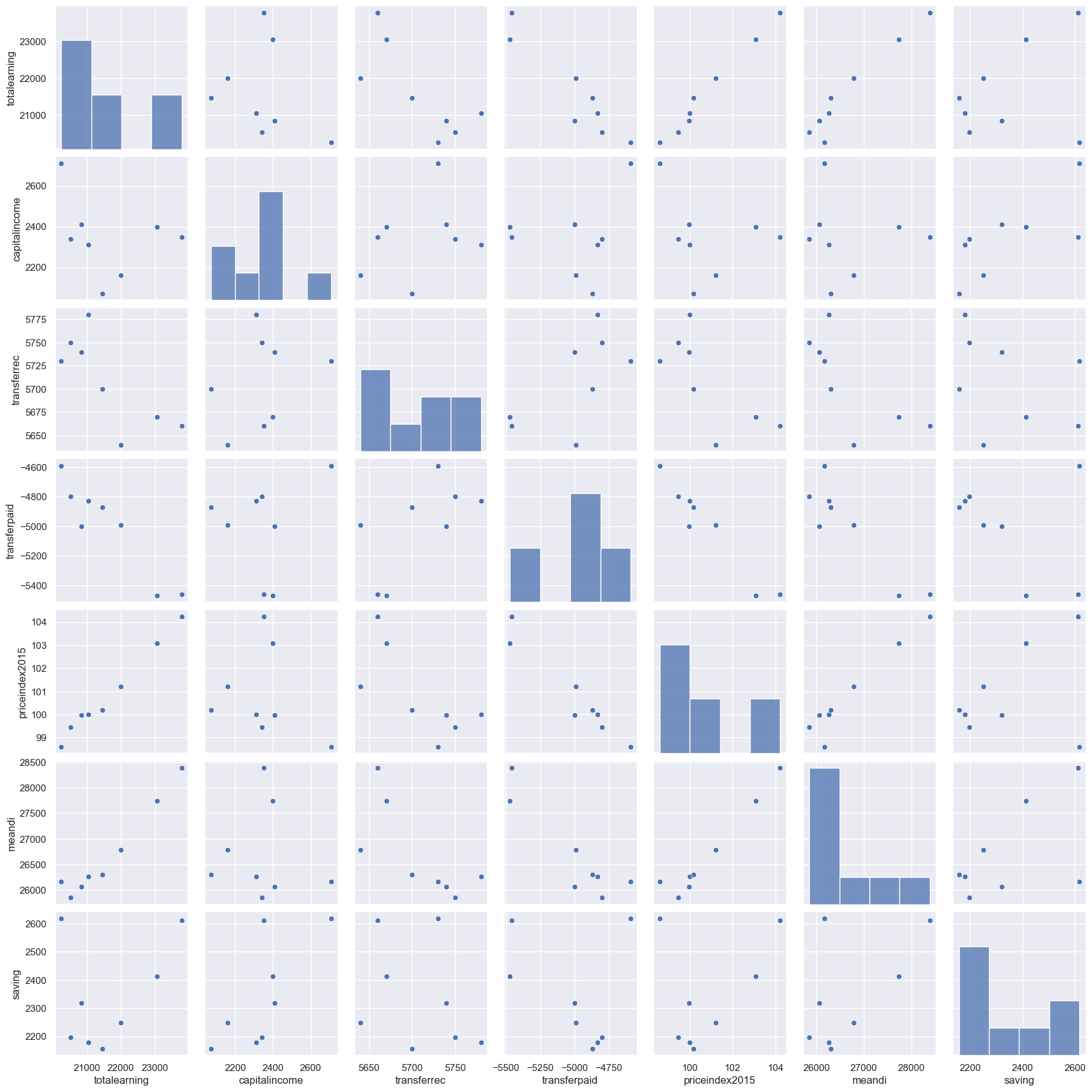
Description générée automatiquement

Une image contenant texte, diagramme, capture d’écran, Rectangle

Description générée automatiquement

Une image contenant texte, Police, capture d’écran, ligne

Description générée automatiquement



I may compare the heatmap with or without the index as percentage as below:

With price index 2015:

Une image contenant texte, capture d’écran, Police, ligne

Description générée automatiquement

Une image contenant capture d’écran, carré, Rectangle, texte

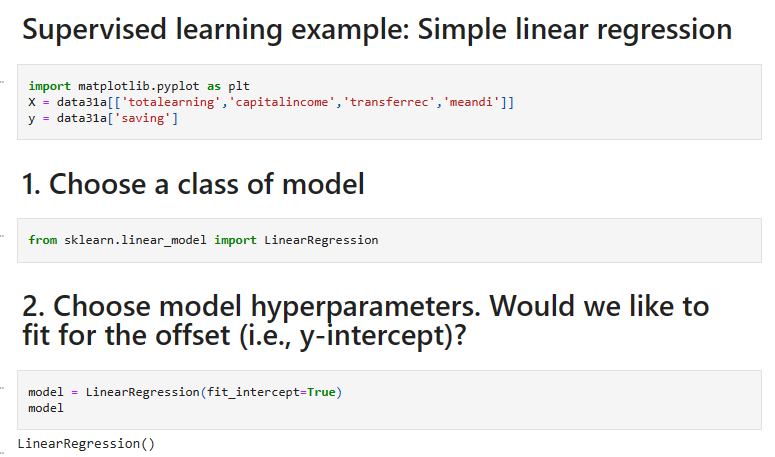
Description générée automatiquement

Without price index 2015

Une image contenant texte, capture d’écran, logiciel, Icône d’ordinateur

Description générée automatiquement

**Machine Learning (ML)**



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

Description générée automatiquement

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

Description générée automatiquement

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

Description générée automatiquement

**Data base type selection**

World data from a toy model of France

Summary: Stationary Test by Dickey-Fuller test: Except for Capital Income, variables are stationary.

|  |  |  |  |
| --- | --- | --- | --- |
| data31['saving'] | data31['meandi'] | data31['priceindex2015'] | data31['transferpaid'] |
| ADF Statistic: -2.09  Critial Values:  1%, -6.05  Critial Values:  5%, -3.93  Critial Values:  10%, -2.99  p-value: 0.25  Stationary | ADF Statistic: 1.32  Critial Values:  1%, -6.05  Critial Values:  5%, -3.93  Critial Values:  10%, -2.99  p-value: 1.00  Stationary | ADF Statistic: -0.44  Critial Values:  1%, -6.05  Critial Values:  5%, -3.93  Critial Values:  10%, -2.99  p-value: 0.90  Stationary | ADF Statistic: -0.72  Critial Values:  1%, -4.94  Critial Values:  5%, -3.48  Critial Values:  10%, -2.84  p-value: 0.84  Stationary |

|  |  |  |
| --- | --- | --- |
| data31['transferrec'] | data31['capitalincome'] | data31['totalearning'] |
| ADF Statistic: -0.93  Critial Values:  1%, -4.94  Critial Values:  5%, -3.48  Critial Values:  10%, -2.84  p-value: 0.78  Stationary | ADF Statistic: -8.90  Critial Values:  1%, -6.05  Critial Values:  5%, -3.93  Critial Values:  10%, -2.99  p-value: 0.00  Stationary  p-value: 0.00  Non-Stationary | ADF Statistic: 1.85  Critial Values:  1%, -5.35  Critial Values:  5%, -3.65  Critial Values:  10%, -2.90  p-value: 1.00  Stationary |

**SQL or No SQL**

For analyzing entity relation, SQL is useful to check which kind of joints are possibly applied. For cross-country data, it is not easy to make left-join or outer join, the reason is the number of row and variables are reduced due to duplicated elements. When we did pivot in Python, it’s the delicate method from the same reason.

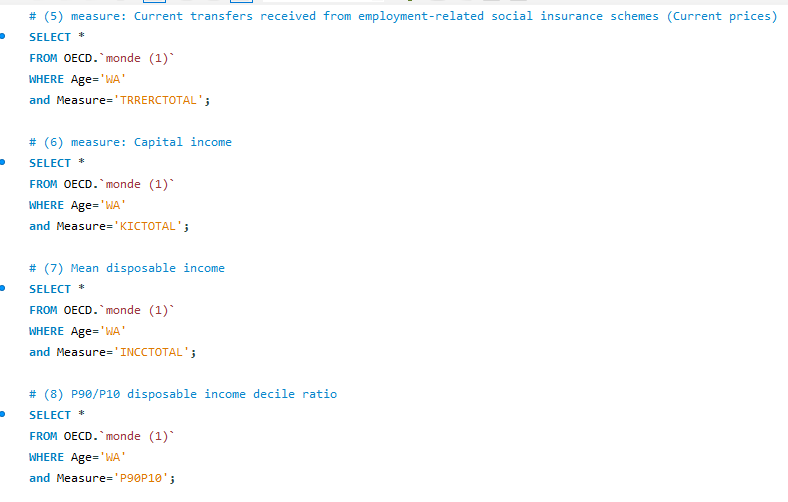
Though, when we use SQL, the entity relation of data is automatically captured. In addition, by using functions such as group by, order by, sum, avg, min, max, it was useful to check the data structure as below.

**Une image contenant texte, capture d’écran, Police

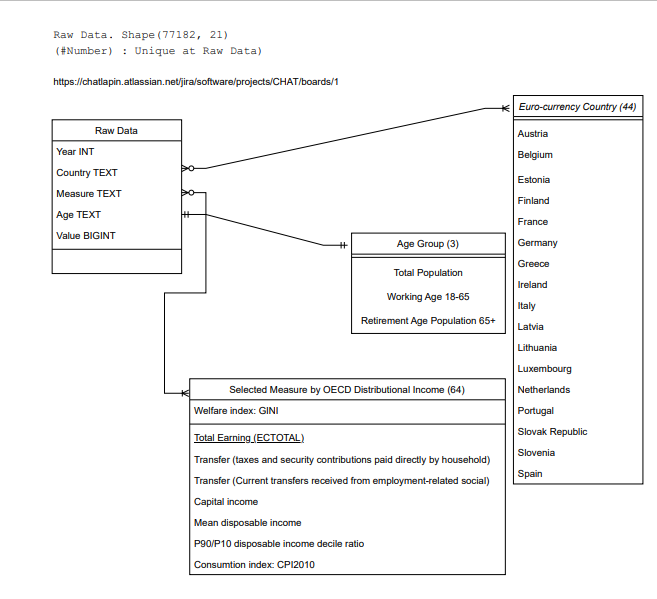
Description générée automatiquementUne image contenant texte, capture d’écran, Police

Description générée automatiquement**

Une image contenant texte, capture d’écran

Description générée automatiquement

**Entities. ERD**

****

**DATA Visualization**

The comparison of Old generation and Working Age (18-65) generation

Consumption: Consumer Price Index (CPI)

**Une image contenant texte, capture d’écran, cercle, diagramme

Description générée automatiquementUne image contenant texte, capture d’écran, diagramme, nombre

Description générée automatiquementUne image contenant texte, capture d’écran, cercle, diagramme

Description générée automatiquement**

Inequality: Gini Index

**Une image contenant texte, capture d’écran, Tracé, diagramme

Description générée automatiquement**Une image contenant texte, capture d’écran, Tracé

Description générée automatiquement

Mean Disposable Income: estimated as the greater amount among variables. Target Variable

Une image contenant texte, capture d’écran, Tracé, Police

Description générée automatiquementUne image contenant texte, capture d’écran, Tracé, Police

Description générée automatiquement

Total Yearly Earning per person

Une image contenant texte, capture d’écran, graphisme, Graphique

Description générée automatiquement

Une image contenant texte, capture d’écran, Police, graphisme

Description générée automatiquement

Paid Transfer (Tax, Insurance…)

Une image contenant texte, capture d’écran, Tracé, ligne

Description générée automatiquementUne image contenant texte, capture d’écran, Tracé, diagramme

Description générée automatiquement

Received Transfer

Une image contenant texte, capture d’écran, Tracé, ligne

Description générée automatiquement

Une image contenant texte, capture d’écran, gratte-ciel

Description générée automatiquement

The P90/P10 ratio compares the income at the 90th percentile to the one at the tenth percentile

Une image contenant texte, capture d’écran, Logiciel multimédia, logiciel

Description générée automatiquement