

[WELFARE AND INEQUALITY ANALYSIS]

Hye-Jin Cho-Druegon
Choh9323@gmail.com

This presentation file is only for RNCP certification. It contains the preliminary data work.
It is not intended to be distributed elsewhere. Thank you.

ROADMAP

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

INTRO

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

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



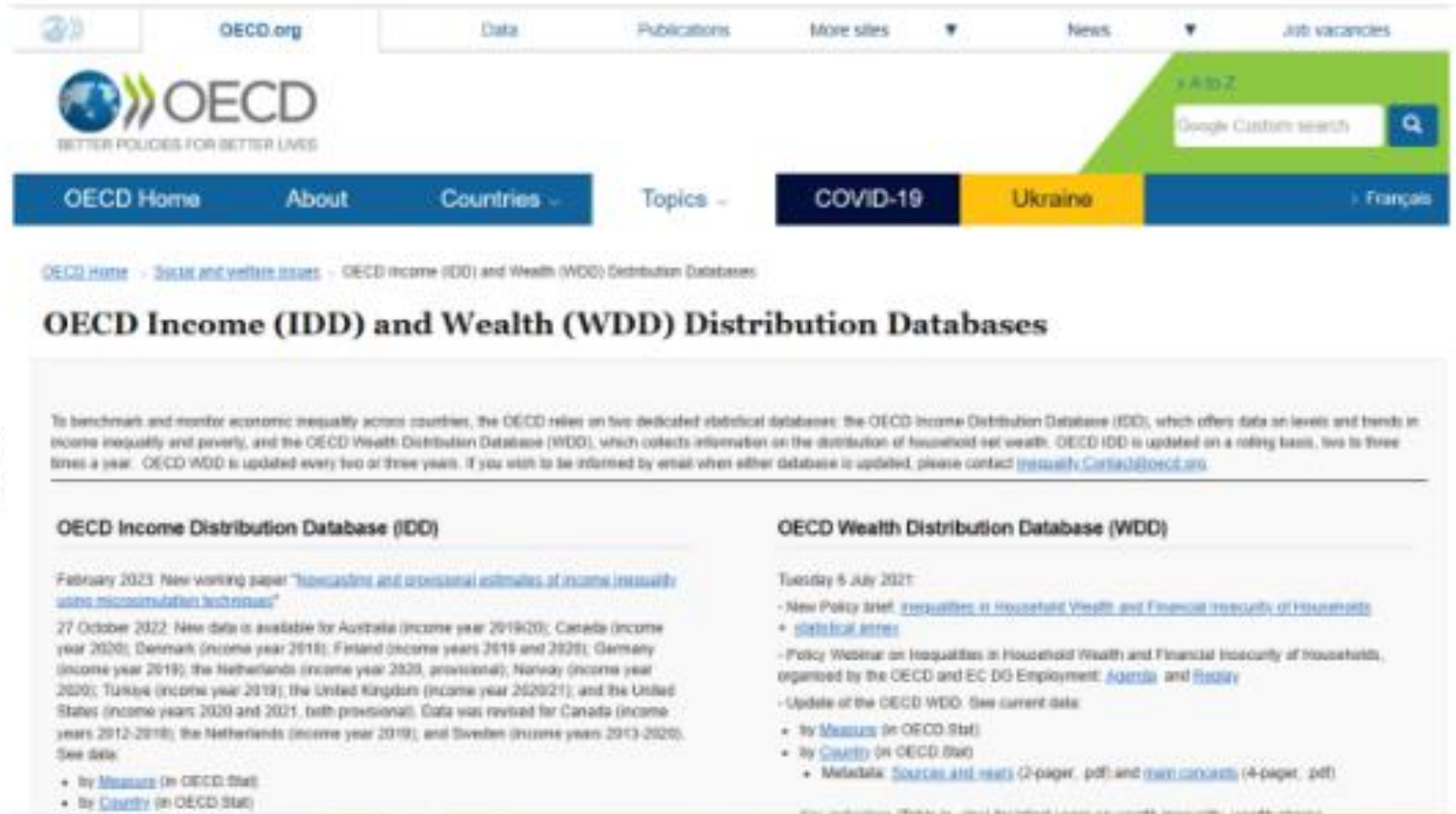
SOLUTION

Output

In Github [DAFT 0410/module5 at main · chatlapin/DAFT 0410 · GitHub](#)

<pre>#starter Jira Datadescription.pdf #ER, Data Cleaning, EDA Erdiagram_OECD.drawio.pdf OECD Data Cleaning and EDA.ipynb OECDRawData.ipynb #Data, cleaned (by variables) Clean_data_pivotedCI.csv Clean_data_pivotedCPI.csv Clean_data_pivotedMDI.csv Clean_data_pivotedP90.csv Clean_data_pivotedTE.csv Clean_data_pivotedTaxsecu.csv Clean_data_pivotedtransfer.csv Clean_data_pivotedCPI.csv #Data, pivoted (by variables) data_pivotedCI.csv data_pivotedCPI.csv data_pivotedMDI.csv data_pivotedP90.csv data_pivotedTE.csv data_pivotedTaxsecu.csv data_pivotedtransfer.csv data_pivotedCPI.csv</pre>	<pre>#SQL, API, Python Toymodel (France), Python ML OECD.sql OECD.API31.ipynb Project Toymodel.ipynb ProjectBig.ipynb #Euro-using country stationary test results UnitRootTestOECD meandi.ipynb UnitRootTestOECDCI.ipynb UnitRootTestOECD CPI.ipynb UnitRootTestOECDP90P10.ipynb UnitRootTestOECDTE.ipynb UnitRootTestOECDtaxsecu.ipynb UnitRootTestOECD transfer.ipynb #Rawdata (by world, France, Euro-using country group, each variable) France.csv Monde (1).csv Welfare(1).csv Welfare.csv # Data Visualisation Welfare_DataVisualisation.xlsx</pre>
---	---

Data source



The screenshot shows the OECD.org website. The top navigation bar includes links for Data, Publications, More sites, News, and Job vacancies. The main header features the OECD logo and a search bar. Below the header, there are tabs for OECD Home, About, Countries, Topics, COVID-19, and Ukraine. The page title is 'OECD Income (IDD) and Wealth (WDD) Distribution Databases'. The main content area is divided into two columns. The left column is titled 'OECD Income Distribution Database (IDD)' and contains a paragraph about the database, a list of recent updates, and links to 'by Measure' and 'by Country'. The right column is titled 'OECD Wealth Distribution Database (WDD)' and contains a paragraph about the database, a list of recent updates, and links to 'by Measure' and 'by Country'.

OECD.org

Data Publications More sites News Job vacancies

OECD
BETTER POLICIES FOR BETTER LIVES

OECD Home About Countries Topics COVID-19 Ukraine Français

OECD Home - Social and welfare issues - OECD income (IDD) and Wealth (WDD) Distribution Databases

OECD Income (IDD) and Wealth (WDD) Distribution Databases

To benchmark and monitor economic inequality across countries, the OECD relies on two dedicated statistical databases: the OECD Income Distribution Database (IDD), which offers data on levels and trends in income inequality and poverty, and the OECD Wealth Distribution Database (WDD), which collects information on the distribution of household net wealth. OECD IDD is updated on a rolling basis, two to three times a year. OECD WDD is updated every two or three years. If you wish to be informed by email when either database is updated, please contact research.contact@oecd.org.

OECD Income Distribution Database (IDD)

February 2023: New working paper "Household and regional estimates of income inequality using microsimulation techniques"

27 October 2022: New data is available for Australia (income year 2019/20); Canada (income year 2020); Denmark (income year 2018); Finland (income years 2019 and 2020); Germany (income year 2019); the Netherlands (income year 2020, provisional); Norway (income year 2020); Turkey (income year 2019); the United Kingdom (income year 2020/21); and the United States (income years 2020 and 2021, both provisional). Data was revised for Canada (income years 2012-2018); the Netherlands (income year 2019); and Sweden (income years 2013-2020). See data:

- by [Measure](#) (in OECD Stat)
- by [Country](#) (in OECD Stat)

OECD Wealth Distribution Database (WDD)

Tuesday 5 July 2021:

- New Policy brief: [Inequalities in Household Wealth and Financial Insecurity of Households](#)
- [Statistical annex](#)
- Policy Webinar on Inequalities in Household Wealth and Financial Insecurity of households, organised by the OECD and EC DG Employment: [Agenda](#) and [Recording](#)
- Update of the OECD WDD: See current data
- by [Measure](#) (in OECD Stat)
- by [Country](#) (in OECD Stat)
 - Metadata: [Sources and notes](#) (2-page, pdf) and [Data contacts](#) (4-page, pdf)

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.>


```
df.sample(4)
```

	LOCATION	Country	MEASURE	Measure	AGE	Age group	DEFINITION	Definition	METHOD	Methodology	...	Year
19190	ESP	Spain	PVTAA5	Age group 51-65: Poverty rate after taxes and ...	TOT	Total population	CURRENT	Current definition	METH2012	New income definition since 2012	...	2011
18845	NOR	Norway	GINIG	Gini (gross income, before taxes)	OLD	Retirement age population: above 65	CURRENT	Current definition	METH2012	New income definition since 2012	...	2020
15389	POL	Poland	TRTOTCTOTAL	Current transfers received from non-profit ins...	WA	Working age population: 18-65	CURRENT	Current definition	METH2012	New income definition since 2012	...	2017

SAMPLING

MACHINE LEARNING: SUPERVISED LEARNING

TOY MODEL: FRANCE 2012_2019 DESCRIPTIVE STATISTICS

	totalearning	capitalincome	transferrec	transferpaid	priceindex2015	meandi	saving
count	8.000000	8.000000	8.000000	8.00000	8.000000	8.000000	8.00000
mean	21618.750000	2343.750000	5708.750000	-5001.25000	100.843026	26693.750000	2343.06875
std	1254.232121	189.581306	49.117207	313.16073	1.909861	903.199669	186.82237
min	20240.000000	2070.000000	5640.000000	-5470.00000	98.605000	25850.000000	2156.60000
25%	20760.000000	2272.500000	5667.500000	-5115.00000	99.835420	26142.500000	2192.83250
50%	21250.000000	2345.000000	5715.000000	-4930.00000	100.090835	26280.000000	2284.43000
75%	22265.000000	2402.500000	5742.500000	-4822.50000	101.685025	27020.000000	2463.00500
max	23790.000000	2710.000000	5780.000000	-4590.00000	104.232500	28390.000000	2617.00000

MACHINE LEARNING: SUPERVISED LEARNING

TOY MODEL: FRANCE 2012_2019

```
model.coef_
```

```
array([32.25850999, 86.42024677])
```

```
model.intercept_
```

```
0.007303861711072557
```

5. Predict labels for unknown data

```
# new instances where we do not know the answer
Xfit, _ = make_regression(n_samples=3, n_features=2, noise=0.1, random_state=1)
# make a prediction
yfit = model.predict(Xfit)
# show the inputs and predicted outputs
for i in range(len(Xnew)):
    print("X=%s, Predicted=%s" % (Xfit[i], yfit[i]))
```

```
X=[-1.07296862 -0.52817175], Predicted=-80.2497983168563
```

```
X=[-0.61175641  1.62434536], Predicted=120.649280643451
```

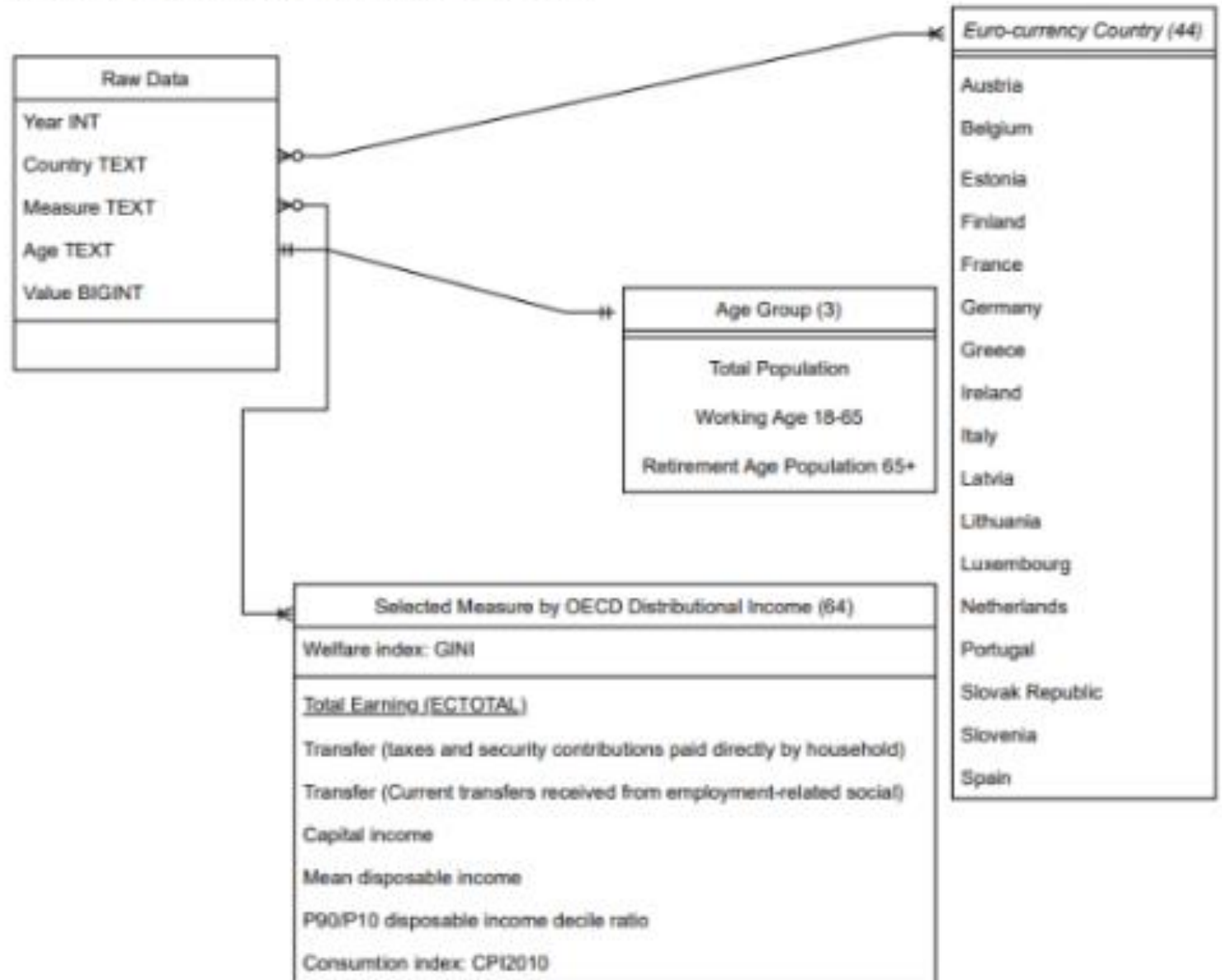
```
X=[-2.3015387  0.86540763], Predicted=0.5518357031231957
```

```
yfit = model.predict(Xfit)
```


ENTITIES AND ERD

```
Raw Data. Shape(77182, 21)
(#Number) : Unique at Raw Data)
```

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





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.

SQL OR NO SQL

```
#function 1: Group by
#show the table with selected Euro-using countries with measures.
SELECT *
FROM OECD."monde (1)"
GROUP BY Country;
```

```
#total generations
SELECT *
FROM OECD."monde (1)"
WHERE Age='TOT';
```

```
#average values for total generations (young and old)
SELECT avg('values')
FROM (select *
FROM OECD."monde (1)"
group by Age='TOT'
limit 5) summary;
```

```
# function 2: Order by
#working-ageing generations
SELECT *
FROM OECD."monde (1)"
WHERE Age='WA'
ORDER BY Country;
```

```
# function 3: UNION
SELECT *
FROM OECD."monde (1)"
WHERE Age='OLD'
UNION
SELECT *
FROM OECD."monde (1)"
WHERE Age='WA';
```

```
#(1) measure: GINI
SELECT *
FROM OECD."monde (1)"
WHERE Age='OLD'
and Measure='GINI';
```

```
#GROUP FUNCTIONS: MAX(4), MIN(5), AVG(6), SUM(7), COUNT(7)
SELECT MAX('values') as max,
MIN('values') as min,
AVG('values') as average,
SUM('values') as total,
COUNT('values') as NROWS_columns
FROM OECD."monde (1)"
WHERE Age='OLD'
and Measure='GINI';
```


DATA TYPE SELECTION

World data from a toy model of France

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

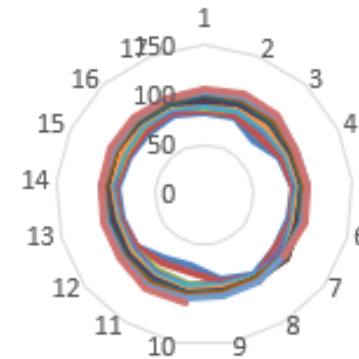
data31['saving']	data31['meandi']	data31['priceindex2015']	data31['transferpaid']
ADF Statistic: -2.09 Critical Values: 1%, -6.05 Critical Values: 5%, -3.93 Critical Values: 10%, -2.99 p-value: 0.25 Stationary	ADF Statistic: 1.32 Critical Values: 1%, -6.05 Critical Values: 5%, -3.93 Critical Values: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

DATA TYPE SELECTION

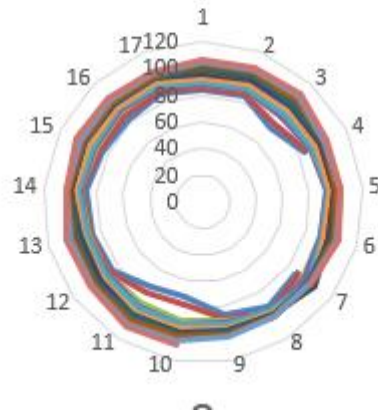
data31['transferrec']	data31['capitalincome']	data31['totalearning']
ADF Statistic: -0.93 Critical Values: 1%, -4.94 Critical Values: 5%, -3.48 Critical Values: 10%, -2.84 p-value: 0.78 Stationary	ADF Statistic: -8.90 Critical Values: 1%, -6.05 Critical Values: 5%, -3.93 Critical Values: 10%, -2.99 p-value: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

DATA VISUALISATION – INTERGENERATIONAL TRANSFER

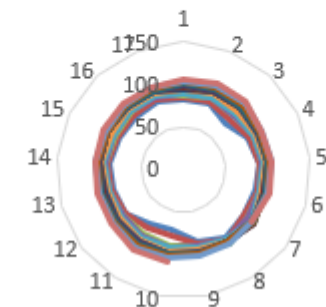
CPI2015 TOTAL



CPI2015 OLD



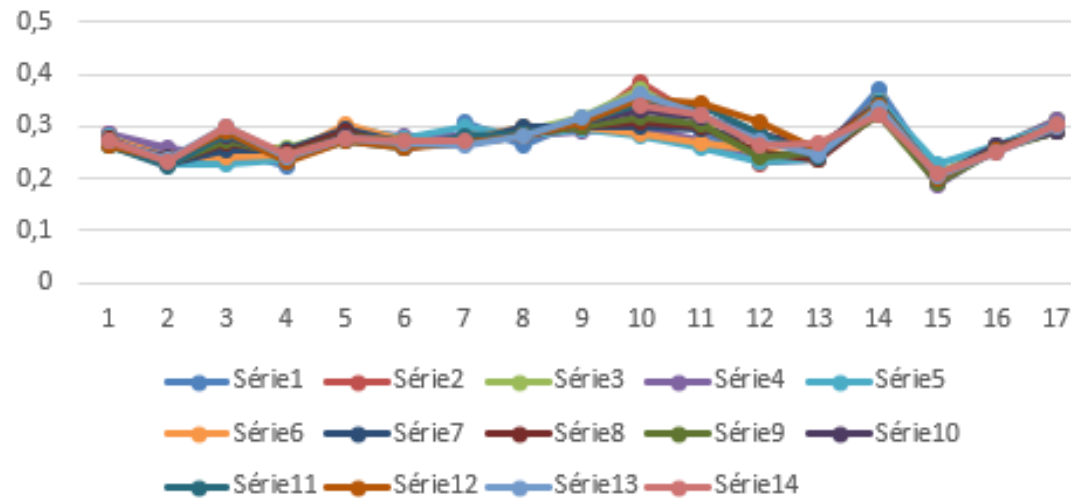
CPI2015 Working Age



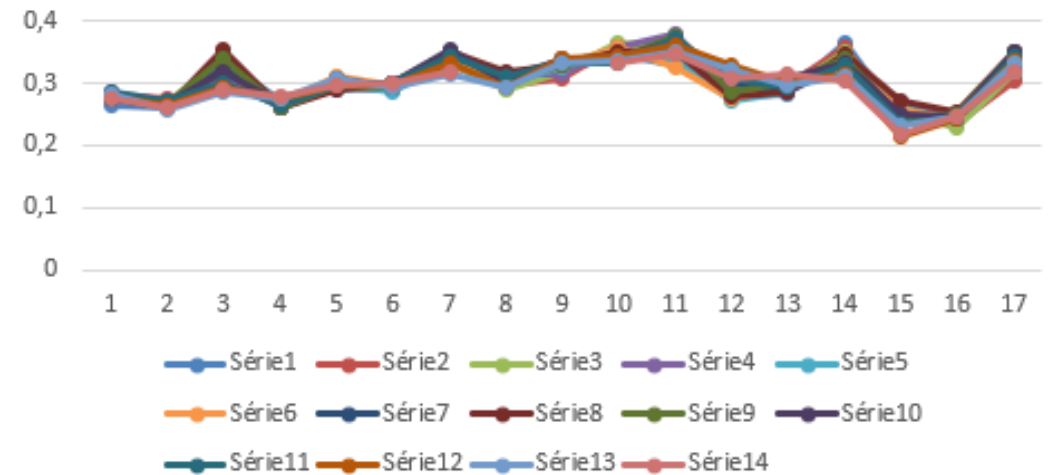
DATA VISUALISATION

INEQUALITY AND GINI INDEX

GINI (OLD) 2006-2019
ref: OECD

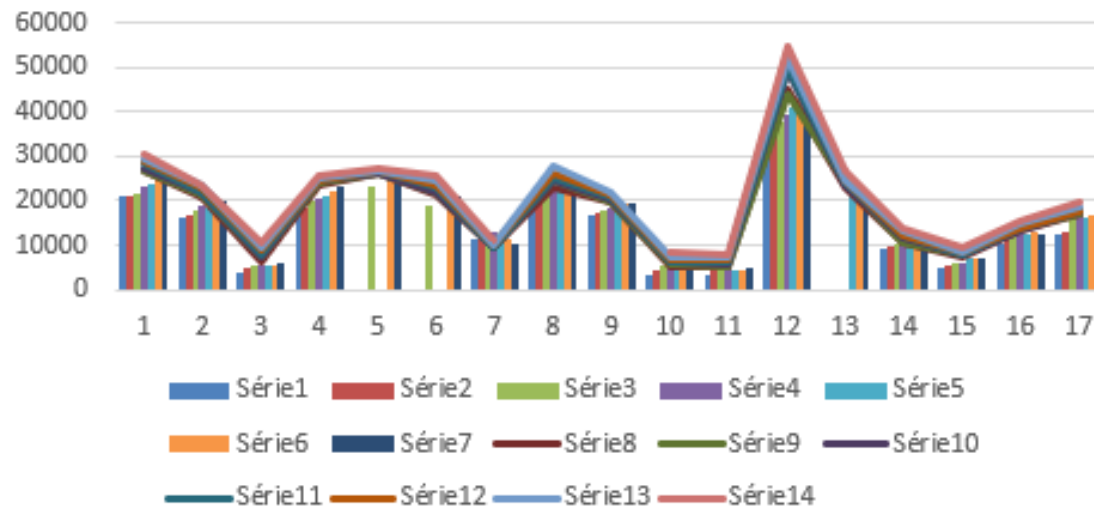


GINI WA (2006-2019)
ref: OECD

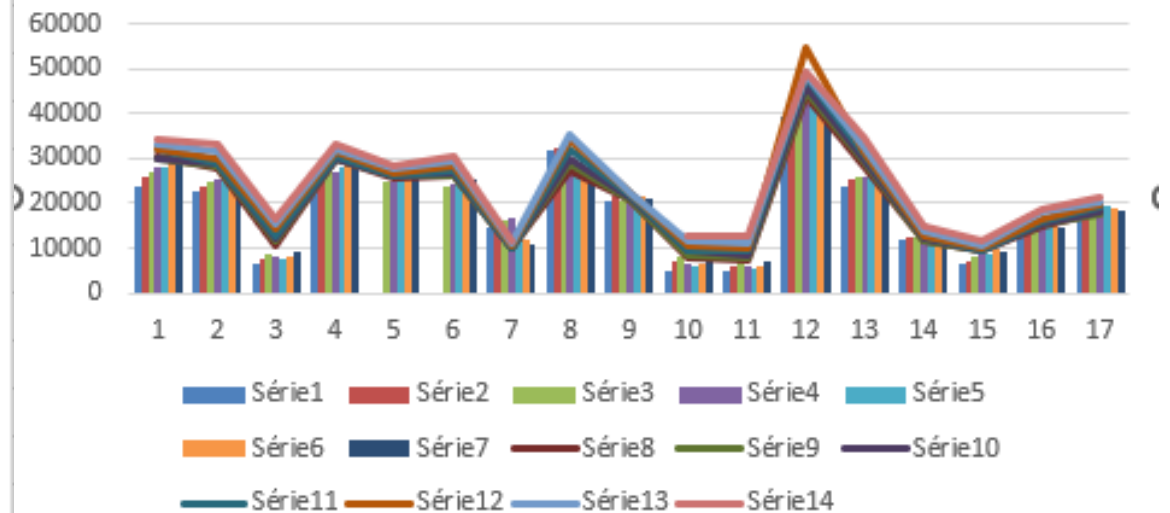


DATA VISUALISATION - MEAN DISPOSABLE INCOME: ESTIMATED AS THE GREATER AMOUNT AMONG VARIABLES. TARGET VARIABLE

Mean Disposable Income (OLD) 2006-2013
ref: OECD



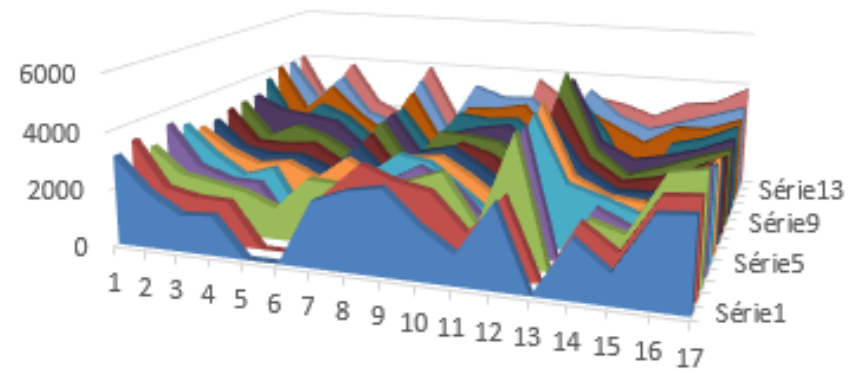
Mean Disposable Income (WA) 2006-2013
ref: OECD



DATA VISUALISATION

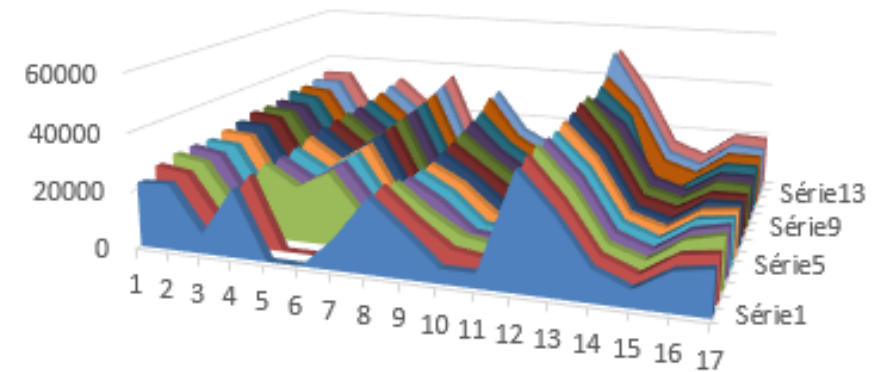
TOTAL YEARLY EARNING PER PERSON

Total Earning (OLD) 2006-2019
ref: OECD



Série1 Série2 Série3 Série4 Série5 Série6 Série7
Série8 Série9 Série10 Série11 Série12 Série13 Série14

Total Earning (Working AGE) 2006-2019
ref: OECD

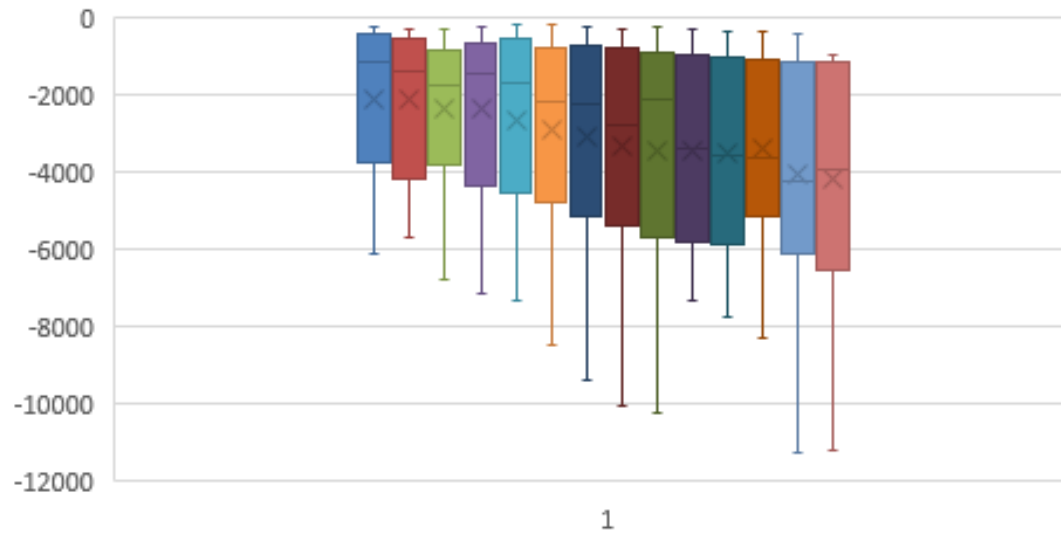


Série1 Série2 Série3 Série4 Série5 Série6 Série7
Série8 Série9 Série10 Série11 Série12 Série13 Série14

DATA VISUALISATION

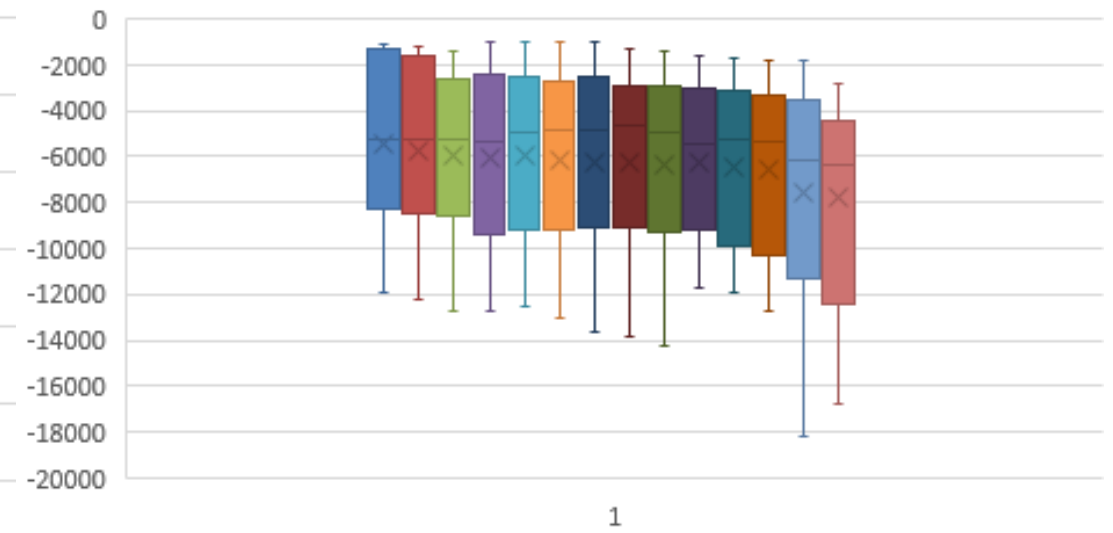
PAID TRANSFER (TAX, INSURANCE)

Paid Transfers (Tax, Social Security) OLD, 2006-2019
ref: OECD



-1864	-1976	-2030	-2143	-2028	-2091	-3751	-3684	-3694	-3786	-3739	-3800
-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------

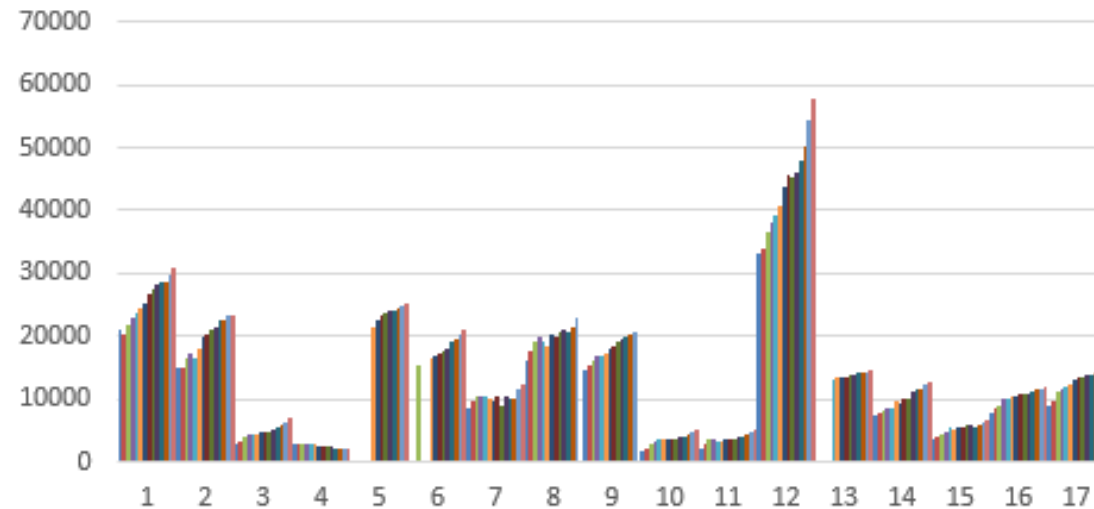
Paid Transfer (Tax, Social Security) WA 2006-2019
ref: OECD



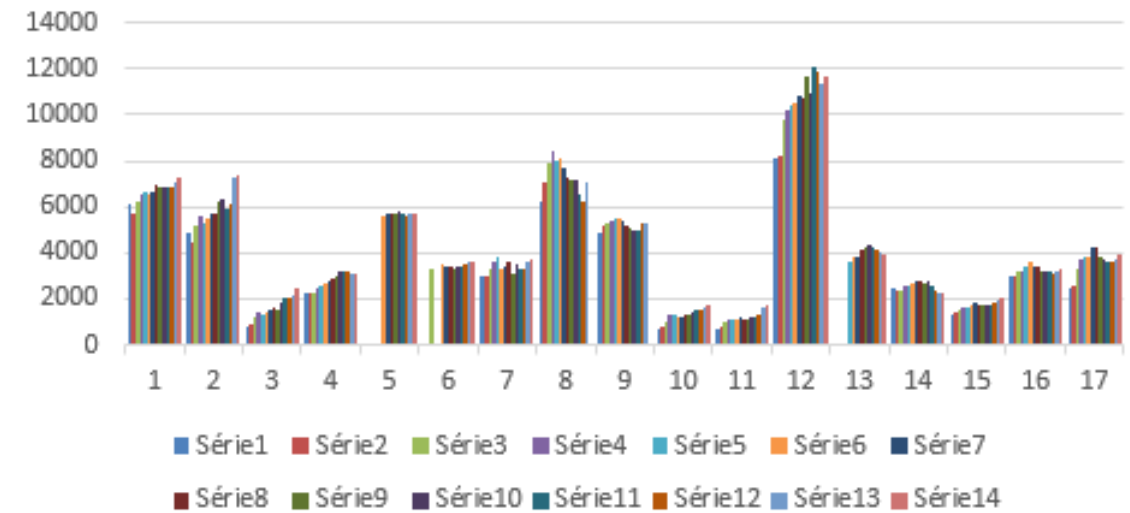
DATA VISUALISATION

RECEIVED TRANSFER (TAX, INSURANCE)

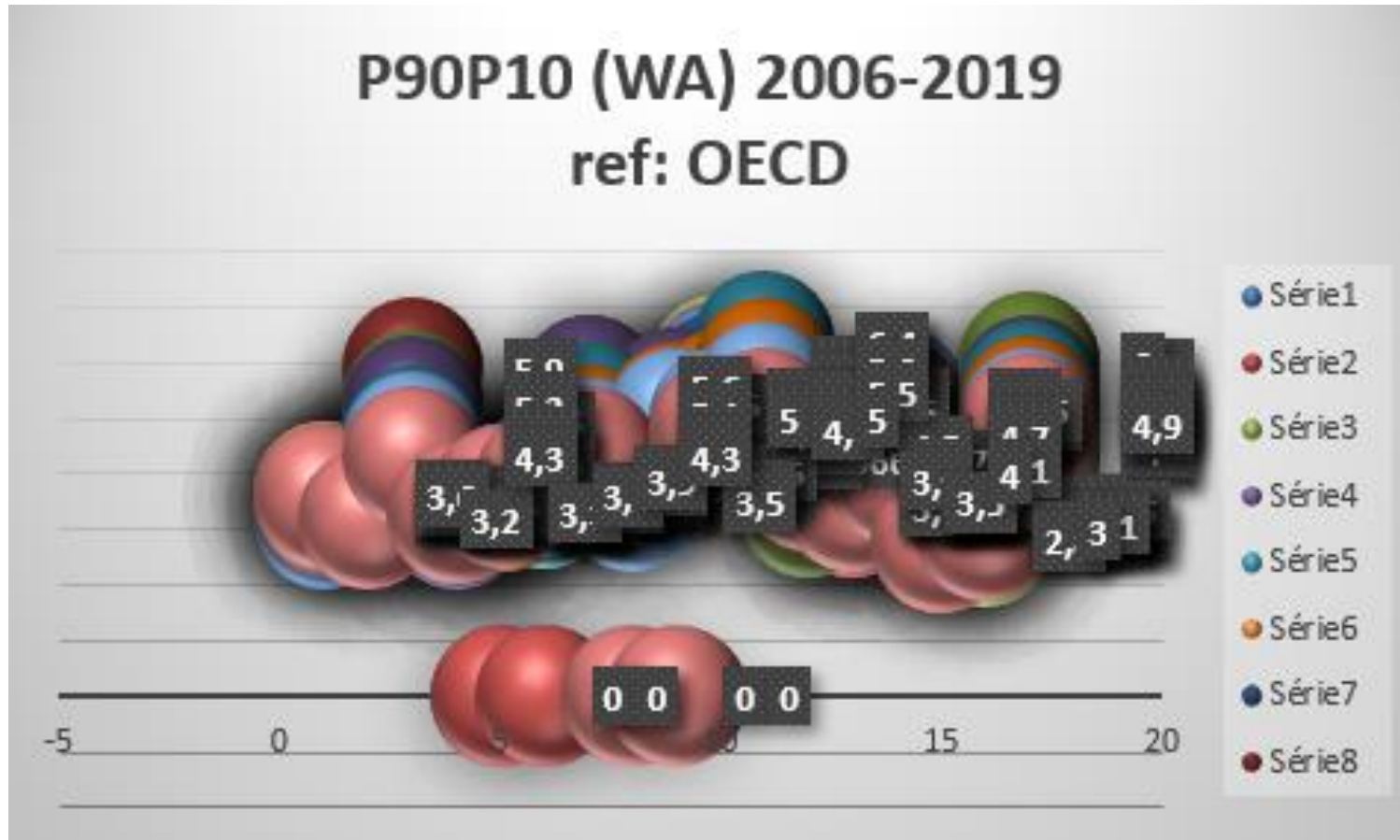
Received Transfer (OLD)2006-2019
ref: OECD



Received Transfer (WA) 2006-2019
ref: OECD



DATA VISUALISATION INEQUALITY FOR THE WORKING AGE GROUP



Thank you.