Country House Price Index (HPI)

BI and data-science solution for analyses and visualization

LEO ANUCHIN, CKDU36
anuchin.leonid@gmail.com
for Business Intelligence Laboratory
[Semester: Spring 2023/2024]

2024/05/12

Link to GitHub:
https://github.com/Leo4CDP/HPI_INDEX
Link to video:
https://drive.google.com/file/d/18V4iQqaYlJ5G4yNyZr8_TDl9waUI_4Xv/view?usp=drive_link
Link to Google Colab:
https://colab.research.google.com/drive/1FVyBKiqb4eGsSNfORimxIGstJJig7Ec7?usp=drive_link

OUTLINE

Aim of the work: the work is to introduce BI and data science methods to work with house price index (HPI) data on the country level in chosen EU countries.

Data sources: API - Eurostat data (7 indicators).
Main indicators to be displayed: HPI growth, key macroeconomic indicators

Report features: during the project the database of indicators was gathered with use of API + ETL data transformations. The data was saved in database (SQL), and presented via Power BI.

Notebook: ML methods revealed the dependence of HPI from other macroeconomic indicators (p-value), and there was an attempt to forecasting its dynamics (ARIMA+regressors).

USED APPLICATIONS

TASK	TECHNOLOGY
ETL (Extract, transform, load)	
1. Data parsing via API	python script
2. ETL engine	VS – SSIS + python scripts
3. Database(s), special data storage solutions	MS SQL (Server name: MSI, database: CKDU36)
Reporting engine	
4. Reporting engine	PowerBI Desktop
5. Unique data visualization	tables, diagrams, geo-spatial data
Data final preparation and analyses	
6. Technologies used for data analysis/data science functions	Google Colab – Jupiter Notebook python libraries: pandas, sklearn, mathplotlib, ipywidgets

WHAT IS HPI?

HPI - House Price Index. Indicator of house price trends. The HPI also functions as an analytical tool for estimating changes in the rates of mortgage defaults, prepayments, and housing affordability. [Investopedia, https://www.investopedia.com/terms/h/house-price-index-hpi.asp]

This **indicator shows the state of the residential real estate market at the country level** (however, in different parts of the country - cities and regions, the situation may vary).

The indicator is useful for **assess the economic situation**, the pace of its development, **monitoring the affordability and effectiveness of investments in residential real estate**.

WHAT OTHER INDICATORS WE USE AND WHY?

KEY MACROECONOMIC INDICATORS:

- **I10GDP** – GDP growth data (index, quarterly)
- **I05HICP_T** – Inflation (customer price index) data (index, monthly – recalculated to quarterly)
- **I05HICP_R** – Housing rent prices data (index, monthly – recalculated to quarterly)
- **UNEMPLOYMENT** – Level of unemployment (15-72,%, quarterly)
- **I10LABOUR**– Spendings on labour (index, quarterly)

KEY NATIONAL INDICATORS:

- **POPULATION** – Total population (people, annual)

The indicators are partly based on report of BIS Determinants of house prices in central and eastern Europe (September 2007, ISSN 1682-7678), https://www.bis.org/publ/work236.pdf. The report considers this factors among connected to HPI.

ETL PROCESS

Link to GitHub:
https://github.com/Leo4CDP/HPI_INDEX

PREPARATION

1. Work folder is C:\SriptsHPI

2. When first start the project, **firstly the SQL table - CKDU36.sql is deployed**.

3. Some steps (python scripts) **need connection to Python program**. No extra packages needed.

4. The first launch occurs using the **EurostatDownloads.csv** database. The file is used to monitor current updates and store API request codes.

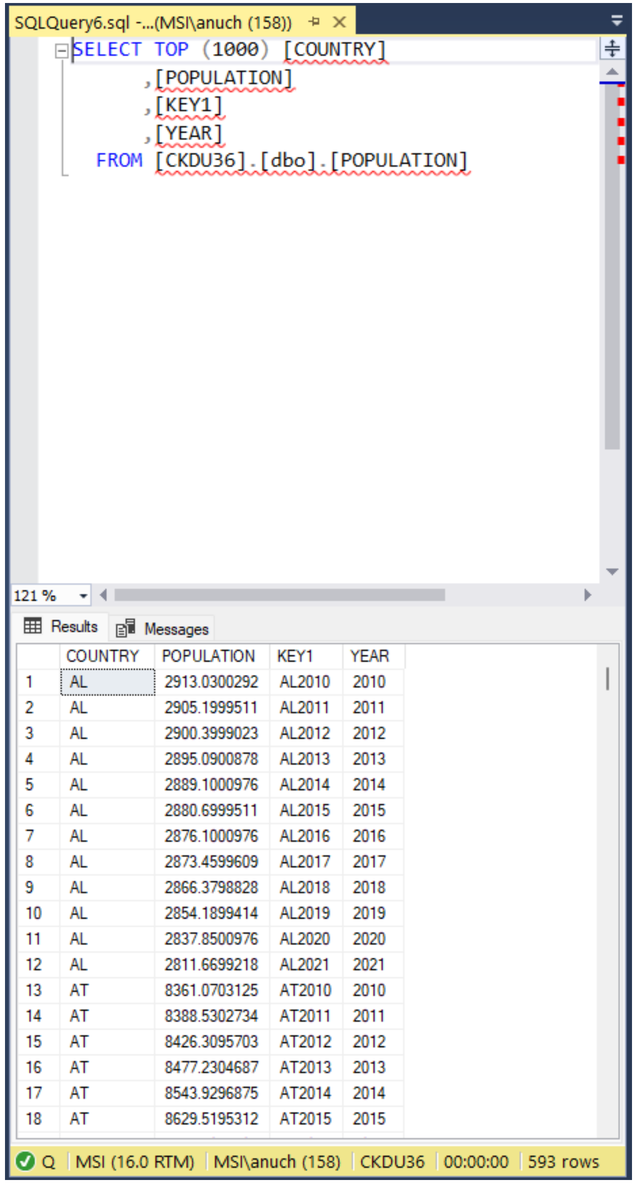
STEPS DESCRIPTION

#	Step	Description
1	NEW VERSION CHECK SCRIPT	SCRIPT1_Updates_checker.py is used. It checks the last updates on Eurostat. If there are new updates it marks EurostatDownloads.csv rows with "1". *If no updates are found if finishes SSIS (With Error). This may be handled by Send Email Task*
2	DOWNLOAD UPDATES	SCRIPT2_xmldownload.py is used. The xml files are downloaded and stored to xml12 folder. This script used codes from EurostatDownloads.csv (flags + API keys) to facilitate download process (<1 MB instead of 50 MB)
3	CLEAN_DOWNLOAD_MARK	SCRIPT2_1_CLEAN6.py is used. The script changes "1" in EurostatDownloads.csv to "0" to prevent unnecessary updates.
4	PARSE XML TO CSV	SCRIPT3_xml_parser.py is used. It parses XML files to CSV on pre-elaborated chema.
5	SQL TR	truncate table STATDATA; truncate table I15HPI; truncate table UNEMPLOYMENT; truncate table I10LABOUR; truncate table POPULATION; truncate table I05HICP_T; truncate table I05HICP_R
6	IMPORT 5 CSV TO SQL SERVER	Import first 5 CSV to SQL (separate tables) They don't need further processing on this stage.
7	JOIN 3 TABLES ON BASE OF GDP SQL TABLE	We sort the rows (on KEY1) and using OLE BD JOIN the tables.
8	CLEAN SQL NA	In resulted STATDATA table we clean all rows with N/A (empty data)

STEPS DESCRIPTION

#	Step	Description
9	CALCULATE NUMBER OF ROWS PER COUNTRY	We calculate number of rows per country. It is needed not to have countries that don’t have enough information in Eurostat. For example the UK stopped providing data science 2020. Greece doesn’t provide HPI. And etc.
10	KICK COUNTRIES WITH LOW ROWS	We kick countries with less then 56 rows (based on #9 results).
11	DICTIONAY TO TRANSFER MONTHLY CSV 2 QUATERLY CSV	Now, we can create dictionary of periods (YYYY-QX) and dictionary of countries. It may be used to parse xml of CPI and Rent prices data. These dictionaries help to avoid any mistakes on step 12 of process.
12	PARSE XML TO CSV 1	SCRIPT4_M2Q.py is used. It parses CPI and Rent prices data from xml to csv format, based on Dictionaries from step 11. It also re-calculates monthly data to quarterly date.
13	IMPORT 2 CSV TO SQL SERVER 1	Now we send csv received on step 12 to SQL tables.
14	JOIN 2 NEW TABLES TO GDP SQL TABLE	We sort the rows (on KEY1) and using OLE BD JOIN the tables.
15	CSV for Notebook	This step is export of STATDATA table to CSV4NOTEBOOK.csv. This file is needed to Google Colab (forecast)

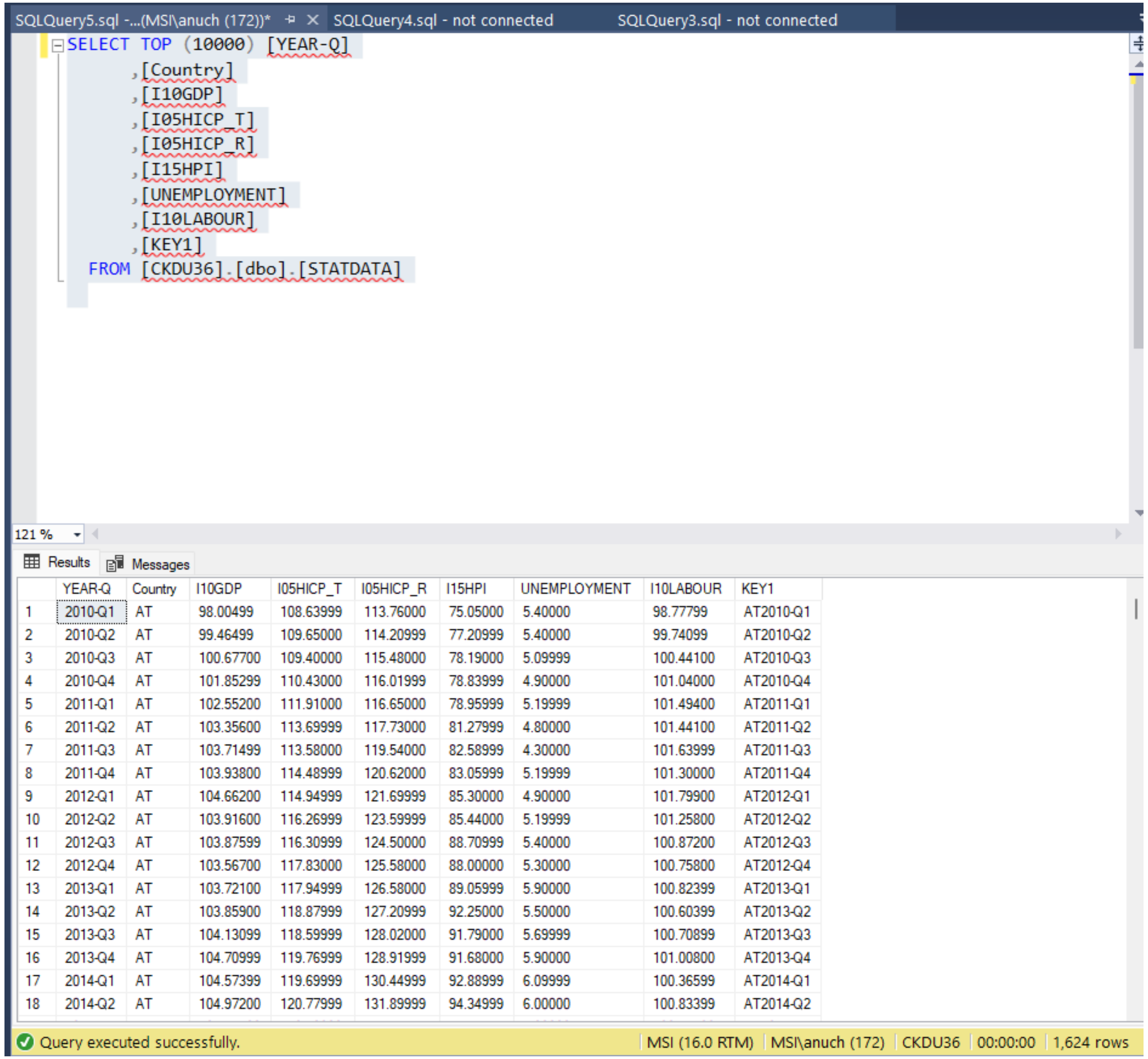
RESULTS OF ETL PROCESS



```sql
SELECT TOP (1000) [COUNTRY]
      ,[POPULATION]
      ,[KEY1]
      ,[YEAR]
  FROM [CKDU36].[dbo].[POPULATION]
```

	COUNTRY	POPULATION	KEY1	YEAR
1	AL	2913.0300292	AL2010	2010
2	AL	2905.1999511	AL2011	2011
3	AL	2900.3999023	AL2012	2012
4	AL	2895.0900878	AL2013	2013
5	AL	2889.1000976	AL2014	2014
6	AL	2880.6999511	AL2015	2015
7	AL	2876.1000976	AL2016	2016
8	AL	2873.4599609	AL2017	2017
9	AL	2866.3798828	AL2018	2018
10	AL	2854.1899414	AL2019	2019
11	AL	2837.8500976	AL2020	2020
12	AL	2811.6699218	AL2021	2021
13	AT	8361.0703125	AT2010	2010
14	AT	8388.5302734	AT2011	2011
15	AT	8426.3095703	AT2012	2012
16	AT	8477.2304687	AT2013	2013
17	AT	8543.9296875	AT2014	2014
18	AT	8629.5195312	AT2015	2015

✓ Q | MSI (16.0 RTM) | MSI\anuch (158) | CKDU36 | 00:00:00 | 593 rows


SQLQuery5.sql -...(MSI\anuch (172))*
SQLQuery4.sql - not connected
SQLQuery3.sql - not connected
121 %
Results
Messages

```sql
SELECT TOP (10000) [YEAR-Q]
      ,[Country]
      ,[I10GDP]
      ,[I05HICP_T]
      ,[I05HICP_R]
      ,[I15HPI]
      ,[UNEMPLOYMENT]
      ,[I10LABOUR]
      ,[KEY1]
  FROM [CKDU36].[dbo].[STATDATA]
```

	YEAR-Q	Country	I10GDP	I05HICP_T	I05HICP_R	I15HPI	UNEMPLOYMENT	I10LABOUR	KEY1
1	2010-Q1	AT	98.00499	108.63999	113.76000	75.05000	5.40000	98.77799	AT2010-Q1
2	2010-Q2	AT	99.46499	109.65000	114.20999	77.20999	5.40000	99.74099	AT2010-Q2
3	2010-Q3	AT	100.67700	109.40000	115.48000	78.19000	5.09999	100.44100	AT2010-Q3
4	2010-Q4	AT	101.85299	110.43000	116.01999	78.83999	4.90000	101.04000	AT2010-Q4
5	2011-Q1	AT	102.55200	111.91000	116.65000	78.95999	5.19999	101.49400	AT2011-Q1
6	2011-Q2	AT	103.35600	113.69999	117.73000	81.27999	4.80000	101.44100	AT2011-Q2
7	2011-Q3	AT	103.71499	113.58000	119.54000	82.58999	4.30000	101.63999	AT2011-Q3
8	2011-Q4	AT	103.93800	114.48999	120.62000	83.05999	5.19999	101.30000	AT2011-Q4
9	2012-Q1	AT	104.66200	114.94999	121.69999	85.30000	4.90000	101.79900	AT2012-Q1
10	2012-Q2	AT	103.91600	116.26999	123.59999	85.44000	5.19999	101.25800	AT2012-Q2
11	2012-Q3	AT	103.87599	116.30999	124.50000	88.70999	5.40000	100.87200	AT2012-Q3
12	2012-Q4	AT	103.56700	117.83000	125.58000	88.00000	5.30000	100.75800	AT2012-Q4
13	2013-Q1	AT	103.72100	117.94999	126.58000	89.05999	5.90000	100.82399	AT2013-Q1
14	2013-Q2	AT	103.85900	118.87999	127.20999	92.25000	5.50000	100.60399	AT2013-Q2
15	2013-Q3	AT	104.13099	118.59999	128.02000	91.79000	5.69999	100.70899	AT2013-Q3
16	2013-Q4	AT	104.70999	119.76999	128.91999	91.68000	5.90000	101.00800	AT2013-Q4
17	2014-Q1	AT	104.57399	119.69999	130.44999	92.88999	6.09999	100.36599	AT2014-Q1
18	2014-Q2	AT	104.97200	120.77999	131.89999	94.34999	6.00000	100.83399	AT2014-Q2

✓ Query executed successfully. | MSI (16.0 RTM) | MSI\anuch (172) | CKDU36 | 00:00:00 | 1,624 rows

Both tables are accessible via Power BI.

The data may be exported for different needs (for example in csv format to Jupiter Notebook)

POWER BI

Link to GitHub:
https://github.com/Leo4CDP/HPI_INDEX

ABOUT BI

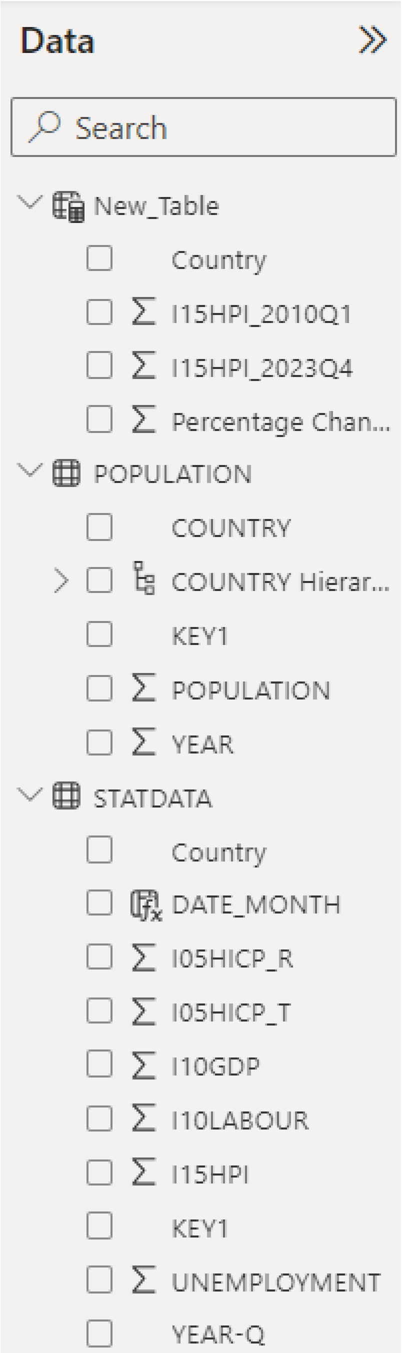
Power BI connect to MS SQL.

1. It connects to STATDATA table (with all indicators in quarterly format.

As supplementary Power BI changes XXXX-QY format to date format to build diagrams (New column).

2. It connects POPULATION database and is able to show population changes (annual basis).

3. New_Table was created to check the HPI change since 2010. This table is used to build geospatial map.



Data

Search

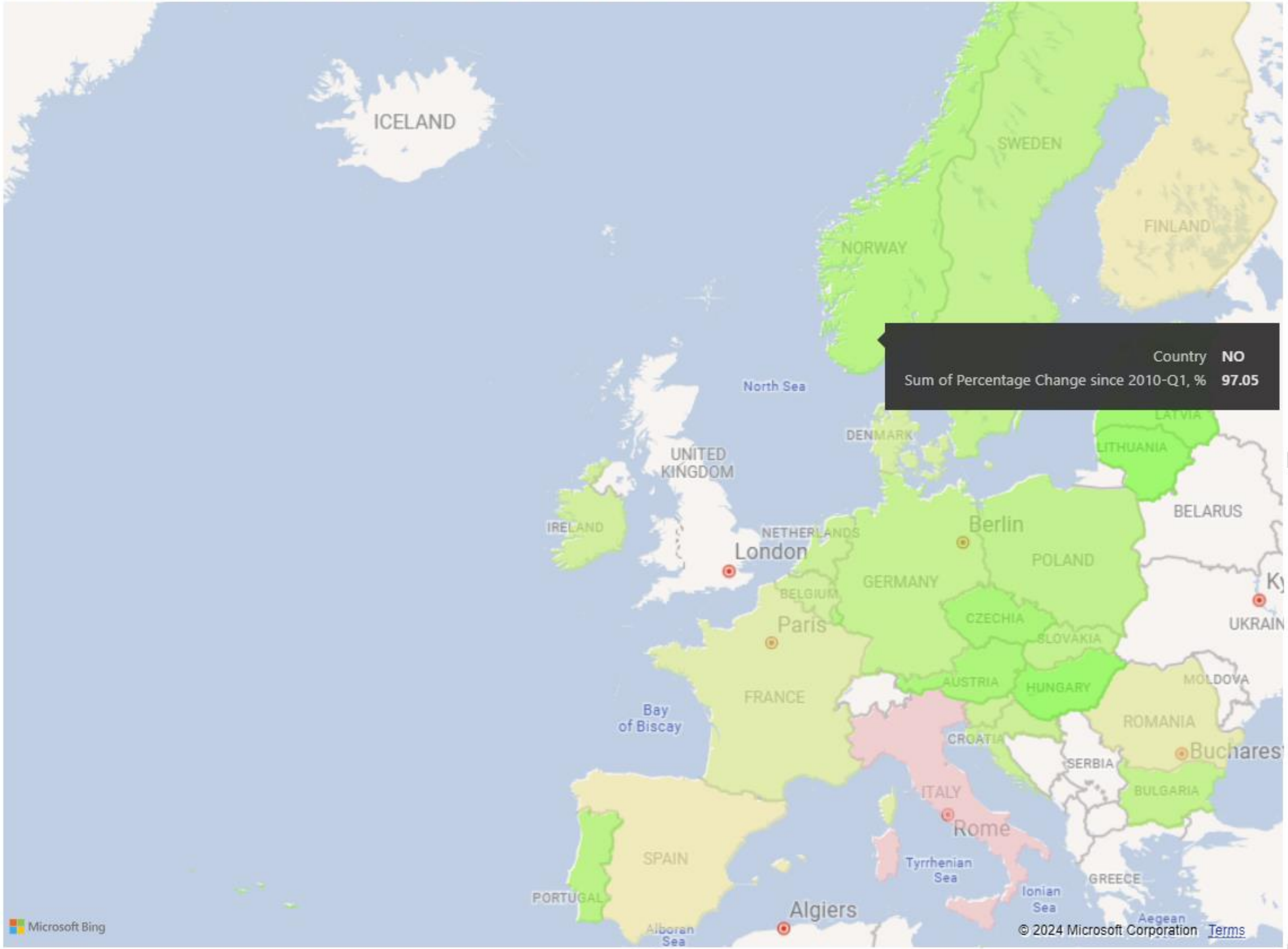
- New_Table
  - □ Country
  - □ Σ I15HPI_2010Q1
  - □ Σ I15HPI_2023Q4
  - □ Σ Percentage Chan...
- POPULATION
  - □ COUNTRY
  - □ COUNTRY Hierar...
  - □ KEY1
  - □ Σ POPULATION
  - □ Σ YEAR
- STATDATA
  - □ Country
  - □ DATE_MONTH
  - □ Σ I05HICP_R
  - □ Σ I05HICP_T
  - □ Σ I10GDP
  - □ Σ I10LABOUR
  - □ Σ I15HPI
  - □ KEY1
  - □ Σ UNEMPLOYMENT
  - □ YEAR-Q

BI TITLE

REPORT ON HPI AND CONNECTED MACROECONOMIC INDICATORS (EUROSTAT DATA, 2010-2023)

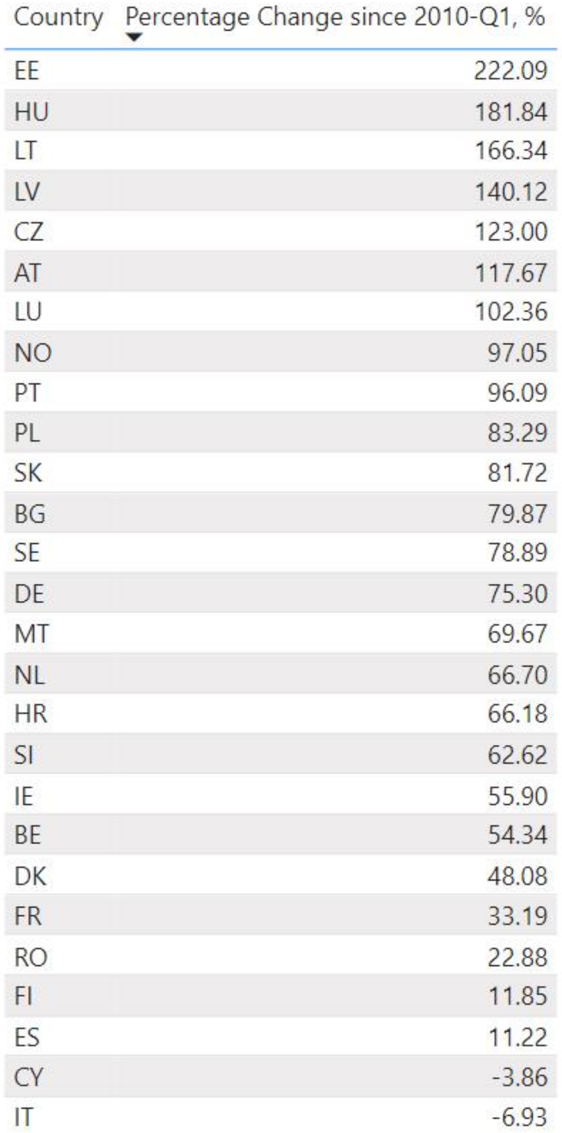
COUNTRY REPORT (1)

Change of House prices in different countries since 2010-Q1



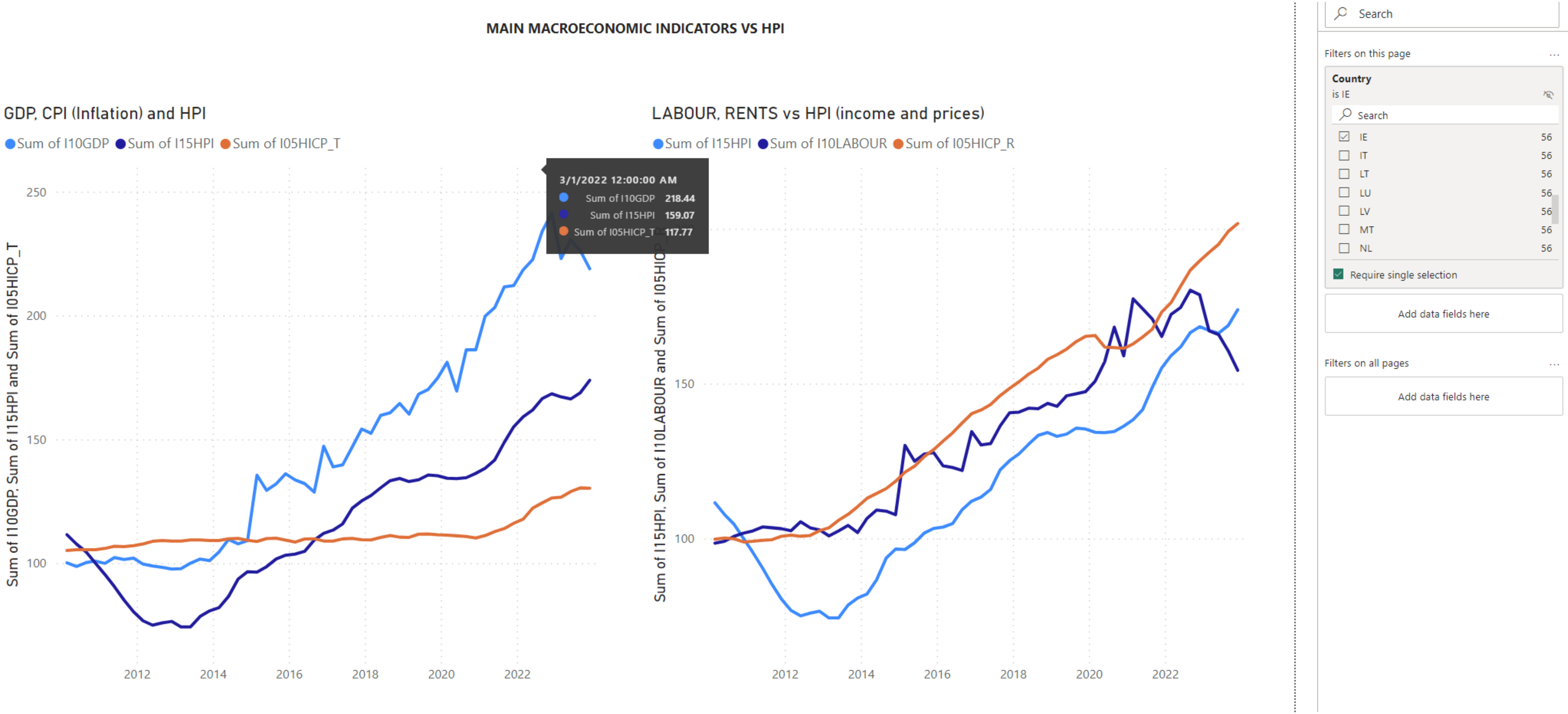
Country NO

Sum of Percentage Change since 2010-Q1, % 97.05



Country	Percentage Change since 2010-Q1, %
EE	222.09
HU	181.84
LT	166.34
LV	140.12
CZ	123.00
AT	117.67
LU	102.36
NO	97.05
PT	96.09
PL	83.29
SK	81.72
BG	79.87
SE	78.89
DE	75.30
MT	69.67
NL	66.70
HR	66.18
SI	62.62
IE	55.90
BE	54.34
DK	48.08
FR	33.19
RO	22.88
FI	11.85
ES	11.22
CY	-3.86
IT	-6.93

COUNTRY REPORT (2)



MAIN MACROECONOMIC INDICATORS VS HPI

GDP, CPI (Inflation) and HPI

● Sum of I10GDP ● Sum of I15HPI ● Sum of I05HICP_T
3/1/2022 12:00:00 AM
● Sum of I10GDP 218.44
● Sum of I15HPI 159.07
● Sum of I05HICP_T 117.77
250
200
150
100
Sum of I10GDP, Sum of I15HPI and Sum of I05HICP_T
2012
2014
2016
2018
2020
2022

LABOUR, RENTS vs HPI (income and prices)

● Sum of I15HPI ● Sum of I10LABOUR ● Sum of I05HICP_R
150
100
Sum of I15HPI, Sum of I10LABOUR and Sum of I05HICP_R
2012
2014
2016
2018
2020
2022

Search

Filters on this page

Country
is IE

Search

☑ IE	56
☐ IT	56
☐ LT	56
☐ LU	56
☐ LV	56
☐ MT	56
☐ NL	56

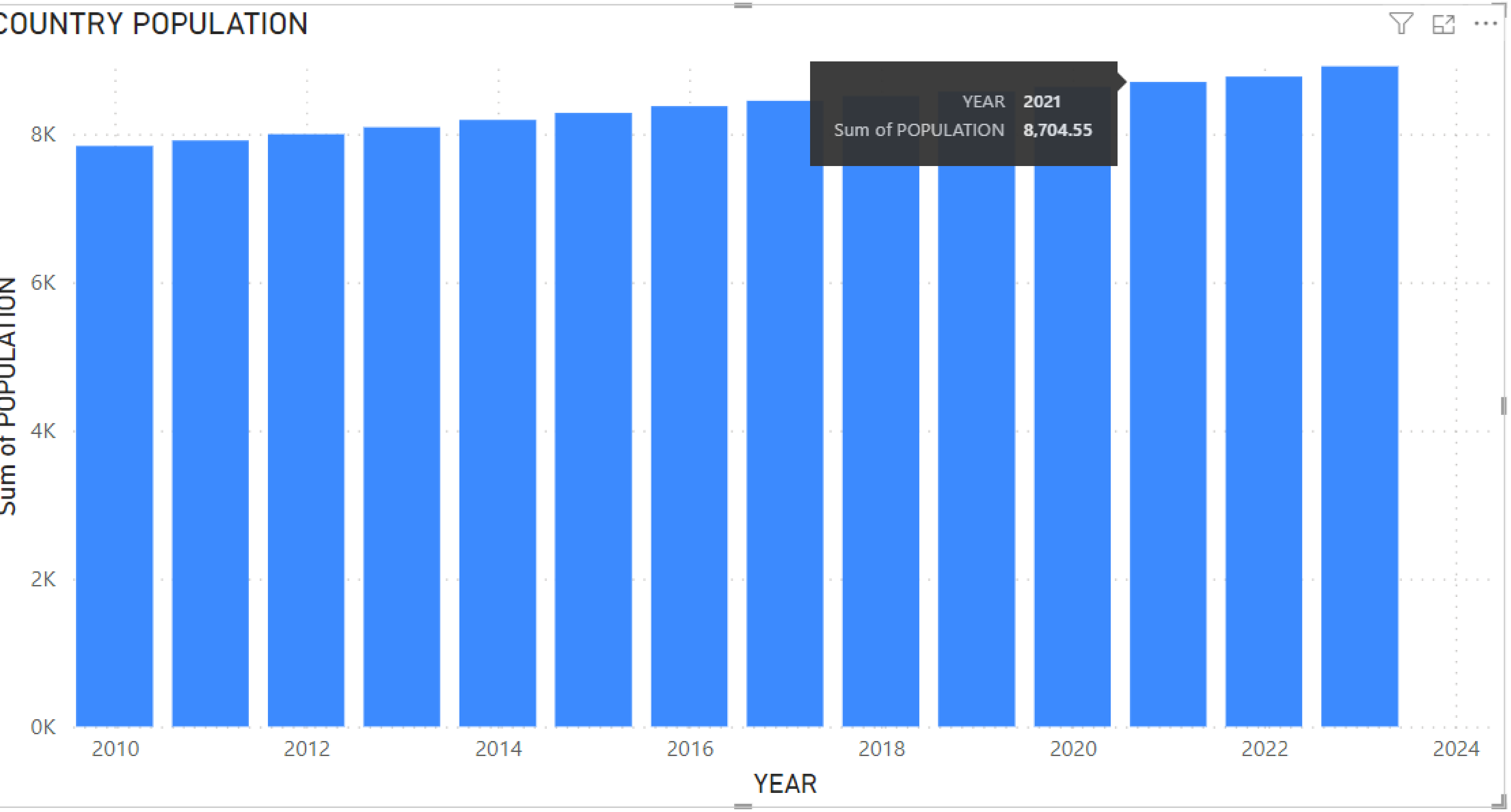
☑ Require single selection

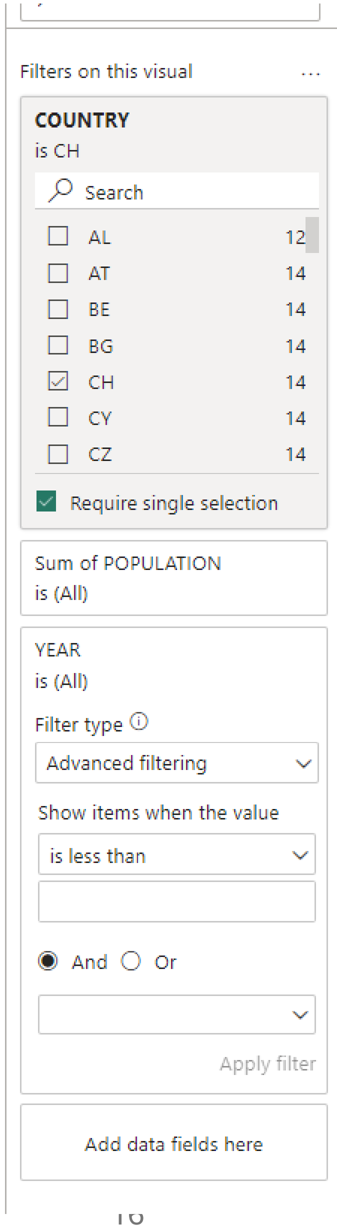
Add data fields here

Filters on all pages

Add data fields here

COUNTRY REPORT (3)





Filters on this visual

COUNTRY
is CH

Search

□	AL	12
□	AT	14
□	BE	14
□	BG	14
☑	CH	14
□	CY	14
□	CZ	14

☑ Require single selection

Sum of POPULATION
is (All)

YEAR
is (All)

Filter type
Advanced filtering

Show items when the value
is less than

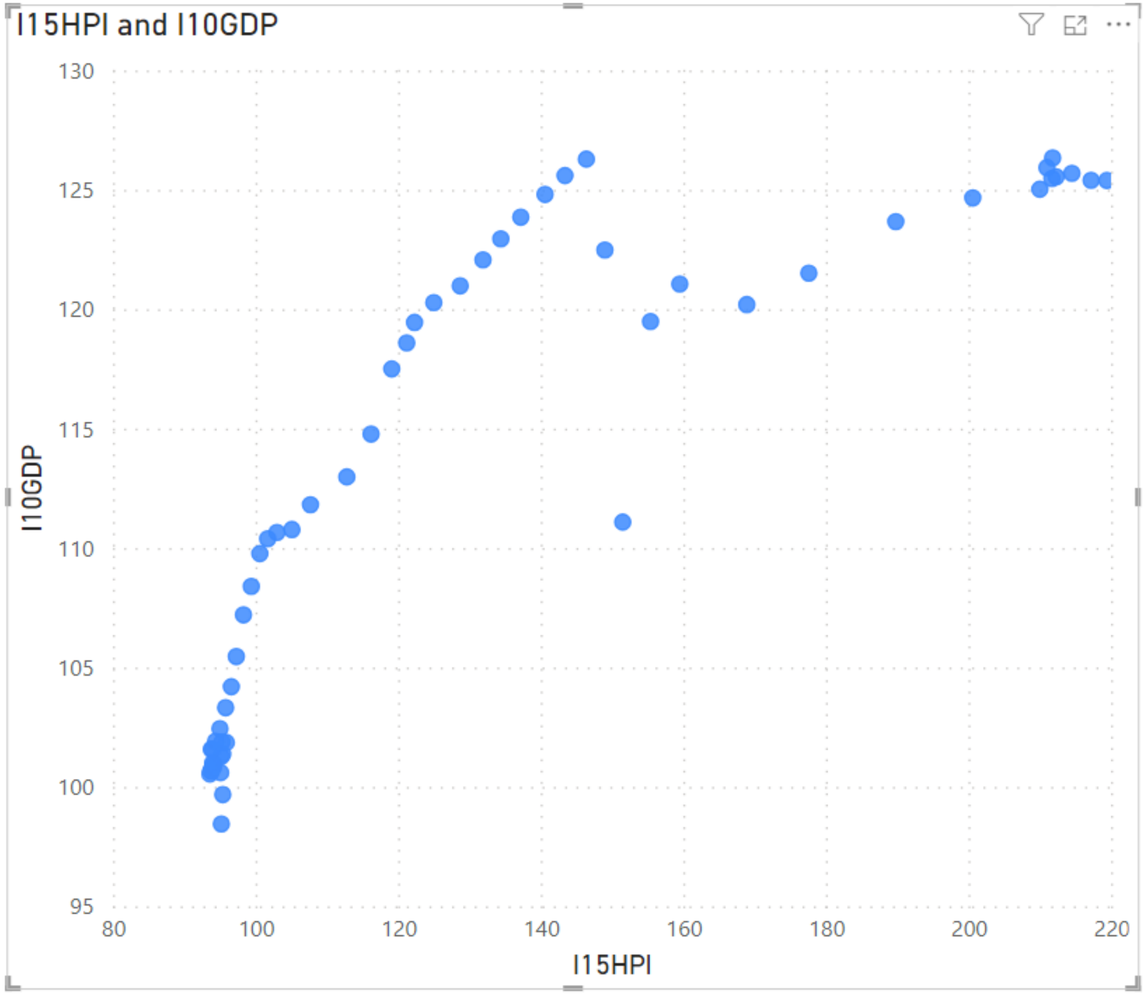
● And ○ Or

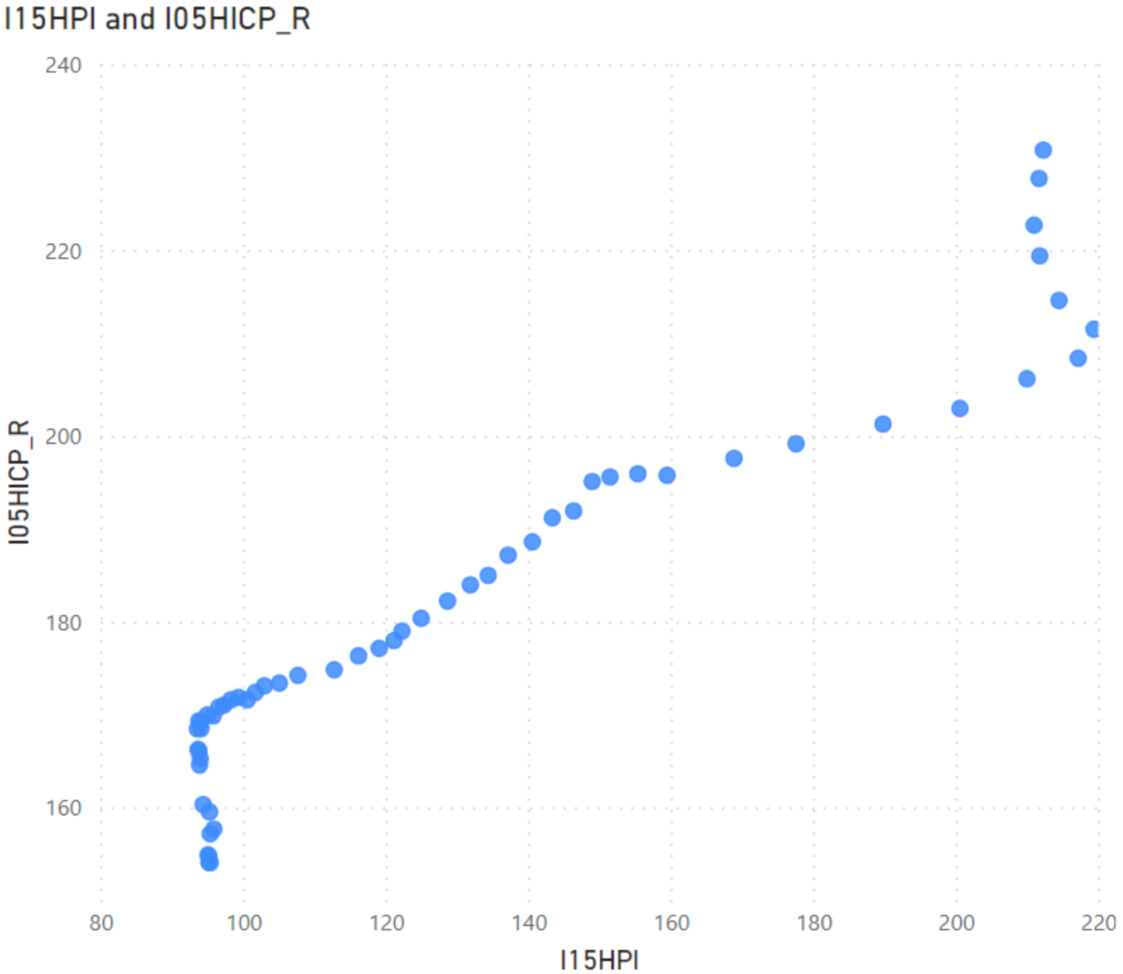
Apply filter

Add data fields here

COUNTRY REPORT (4)

SCATTER CHART: RELATIVITY OF SOME INDICATORS CHANGE





FORECASTING
(FIRST ATTEMPT)

(Jupiter Notebook python)

Link to Google Colab:
https://colab.research.google.com/drive/1FVyBKiqb4eGsSNfORimxIGstJJig7Ec7?usp=drive_link

Link to CSV:
https://drive.google.com/file/d/1qd3XR6Gz8ACCRpjY0QBW7n1WpTIrZv3h/view?usp=drive_link

ABOUT HPI FORECASTING

WHY IS IT IMPORTANT TO FORECAST HPI:

- The HPI measures the cost and affordability of housing for residents. Its prediction allows making decisions that affect the availability of demand (mortgages, preferential programs) and supply (construction volumes) in the medium term.
- the indicator lags behind other macroeconomic indicators - it is prepared six months later than data on inflation and economic growth.

INPUTS

To solve the problem of HPI forecasting at national level, we will use a previously generated data array (based on Eurostat data). **We use CSV4NOTEBOOK.csv – one of products of ETL process. It is to be put to Content folder of Notebook.**

A preliminary analysis of the data showed that the following **restrictions have a significant impact on the results of predictions**:

1) the **economic crisis of 2008**, its effects are observed until the end of 2013-2015. Intervals before 2010 were completely excluded from analyses. (~Dummy-variable);

2) **Covid effect**. In almost all countries, the macroeconomic indicators studied (for example, GDP, inflation) responded to the pandemic faster than real estate prices, which also somewhat reduced the quality of predictions.

MODEL DESCRIPTION

Preliminary tests showed that **ARIMA (1,1,1) + regressors** and **Linear regression (=ARIMA (0,0,0))** show the most alike results to the expected. The sample for each country **(with ipywidgets)** is divided on 2 parts (parameters may be adjusted according to macroeconomic situation in national economic):

- test data (after 2014-01-01)
- validation data (after 2021-01-01)

Used forecasting models*:

- ARIMA (1,1,1) with regressors
- ARIMA (0,0,0) with regressors (=linear regression)

Regressors:

- **I10GDP** – GDP growth data (index)
- **I05HICP_T** – Inflation (customer price index) data (index)
- **I05HICP_R** – Housing rent prices data (index)
- **UNEMPLOYMENT** – Level of unemployment (15-72,%)
- **I10LABOUR**– Spendings on labour (index)

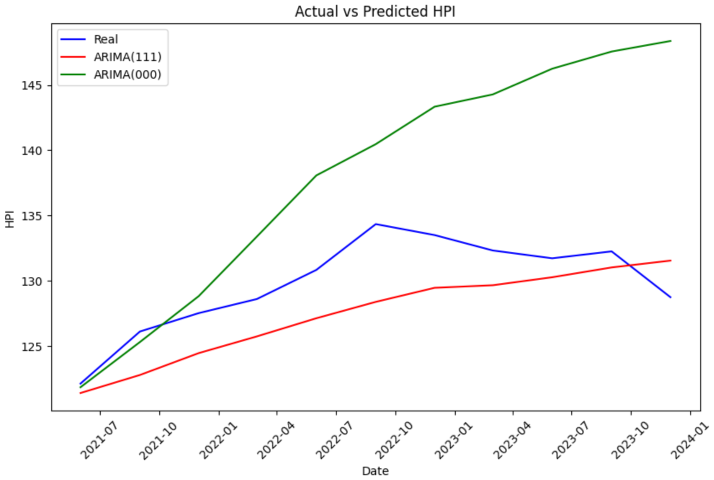
Jupiter Notebook calculates MSE for each model, and shows p-value of every regressor.

* Test on neuro networks models (Keras) showed worser results, probably because of limited sample rows.
** Population indicator was excluded from analyses on this stage, as for some countries (it shows jumps and falls due to statistical methodology and updates on census results)

Examples of predictions: COVID-effect

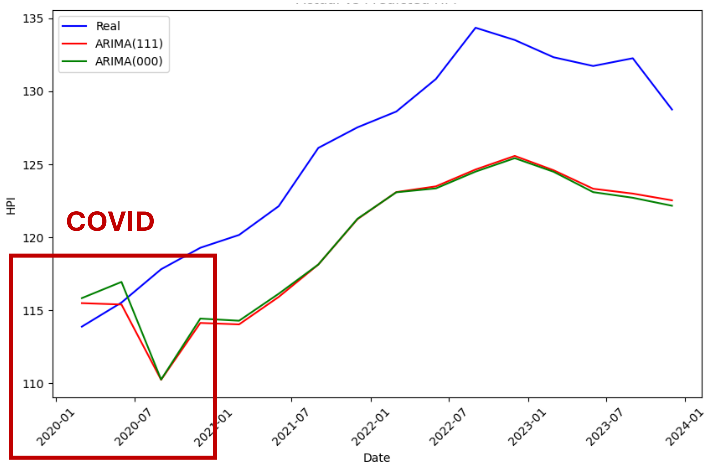
Economic and historic factors influence forecast results. For example, **if we start forecast since 2020-01-01 Hungary macro indicators (GDP, CPI, unemployment and etc.) suffered less than in France, and for Hungary model shows better quality.** But if we take **later date (2021-06-01)**, we see, that **model based on macroeconomic regressors expected better HPI growth, than it occurred in fact.**

FRANCE



FRANCE
train_start = "2013-01-01"
valid_start = "**2021-06-01**"

ARIMA 111 - MSE: 107.7782
ARIMA 000 - MSE: 10.3012

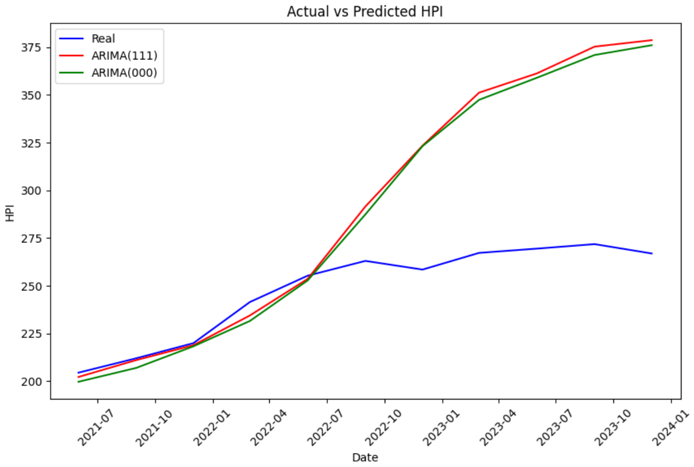


COVID

train_start = "2013-01-01"
valid_start = "2020-01-01"

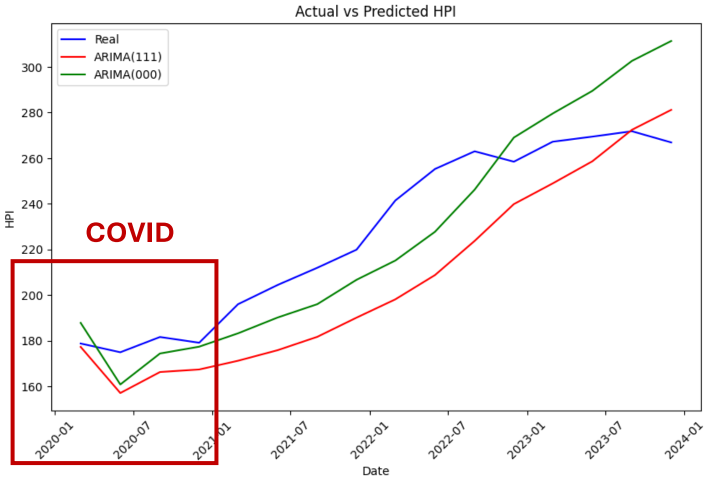
ARIMA 111 - MSE: 48.7467
ARIMA 000 - MSE: 47.6718

HUNGARY



train_start = "2013-01-01"
valid_start = " 2021-06-01"

ARIMA 111 - MSE: 3740.1201
ARIMA 000 - MSE: 3981.8389



COVID

train_start = "2013-01-01"
valid_start = "**2020-01-01**"

ARIMA 111 - MSE: 403.6217
ARIMA 000 - MSE: 656.9420

KEY PRELIMINARY FINDINGS

- The **model allows you to predict the level of HPI depending on other macroeconomic indicators at the national level**, even in conditions of significant turbulence in the macroeconomic situation currently observed. If we use available macroeconomic forecasts (like Central bank forecasts and etc, we may predict house price growth at national level).
- **Model quality results are different for each country**. The most strongly influenced indicators are those associated with a long test base - since countries unevenly emerged from the 2008 crisis (for example, for Romania the model gives the most accurate results if the test sample is taken from 2017; and for Hungary from 2015).
- The **model can be improved by adding additional data sources** (construction volumes, bank lending data) and clarifying HPI data based on real sector data. Also, it is advisable to analyze major economies to find conditions (test base, model parameters) under which the model shows greater accuracy.

FUTHER STEPS

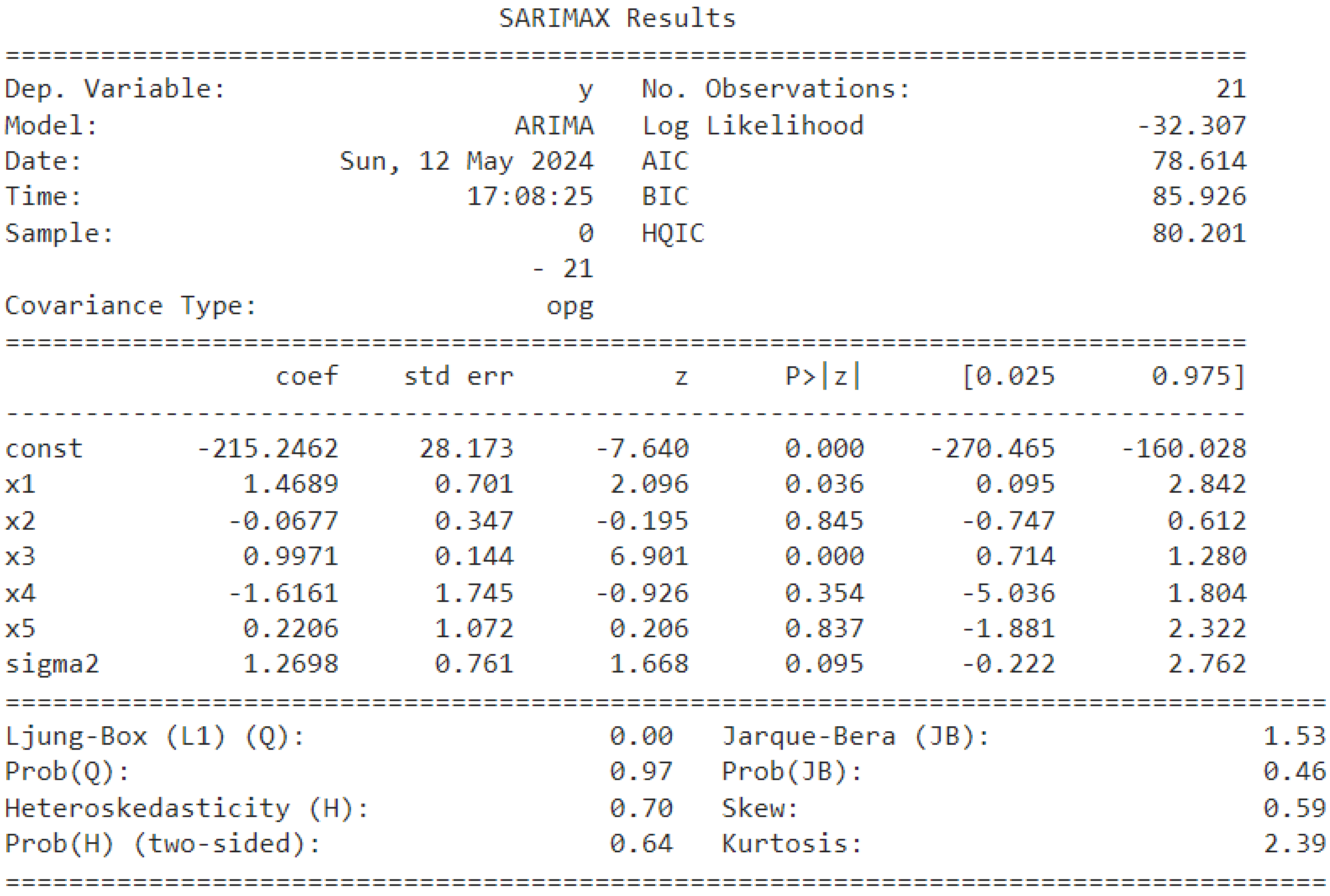
FUTHER STEPS

- Improving the forecasting model (selection of parameters for large national economies, expanding the list of parameters by including information on new construction)
- Testing ETL solution scripts over longer periods
- Scientific publication based on p-value and forecasting results

THANK YOU!

ANNEX: P-VALUE

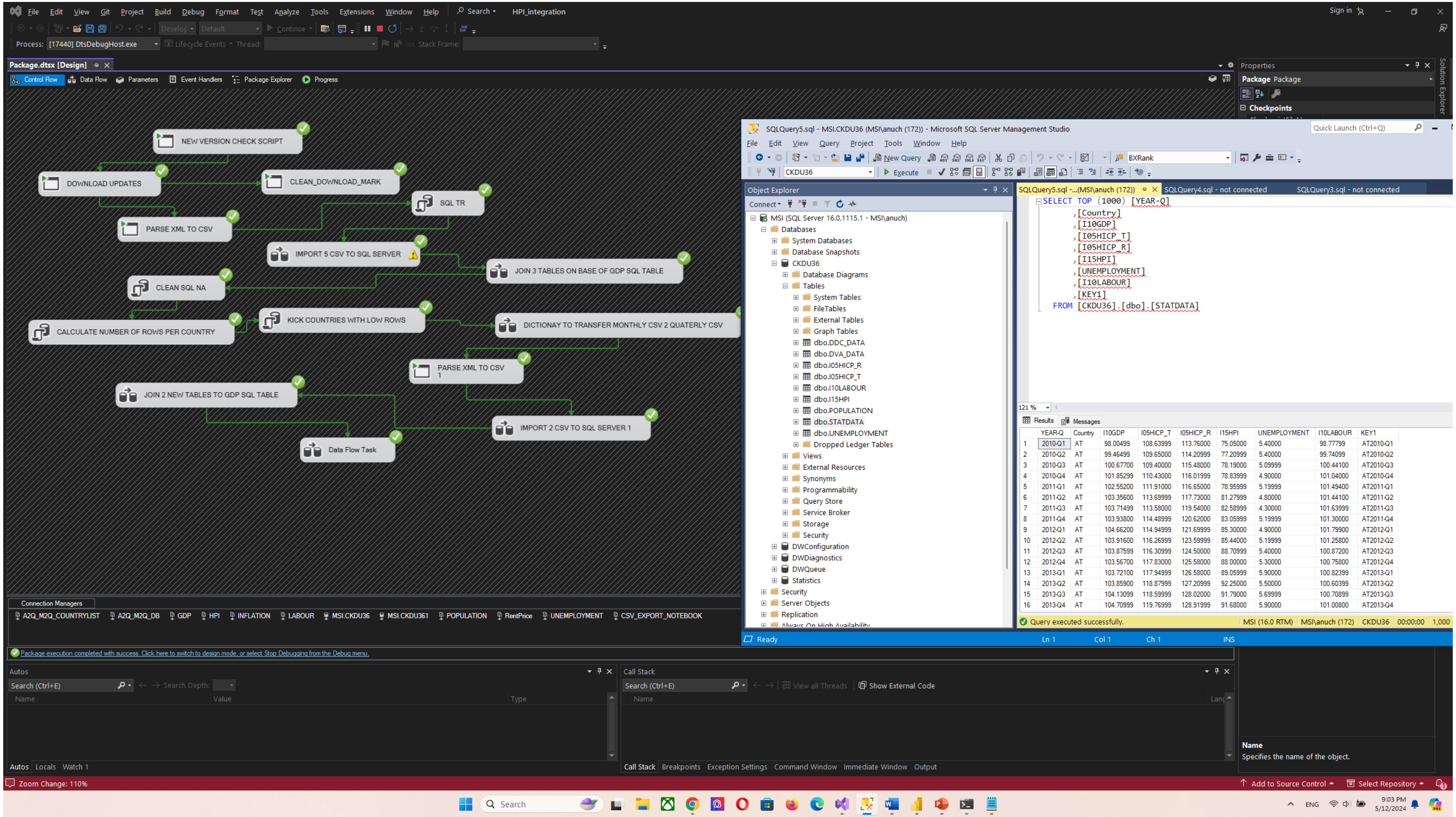
ARIMA(0,0,0)
print(model_fit_0.summary())



```
                               SARIMAX Results                                
==============================================================================
Dep. Variable:                      y   No. Observations:                   21
Model:                          ARIMA   Log Likelihood                 -32.307
Date:                Sun, 12 May 2024   AIC                             78.614
Time:                        17:08:25   BIC                             85.926
Sample:                             0   HQIC                            80.201
                                 - 21                                         
Covariance Type:                  opg                                         
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const       -215.2462     28.173     -7.640      0.000    -270.465    -160.028
x1             1.4689      0.701      2.096      0.036       0.095       2.842
x2            -0.0677      0.347     -0.195      0.845      -0.747       0.612
x3             0.9971      0.144      6.901      0.000       0.714       1.280
x4            -1.6161      1.745     -0.926      0.354      -5.036       1.804
x5             0.2206      1.072      0.206      0.837      -1.881       2.322
sigma2         1.2698      0.761      1.668      0.095      -0.222       2.762
===================================================================================
Ljung-Box (L1) (Q):                   0.00   Jarque-Bera (JB):                 1.53
Prob(Q):                              0.97   Prob(JB):                         0.46
Heteroskedasticity (H):               0.70   Skew:                             0.59
Prob(H) (two-sided):                  0.64   Kurtosis:                         2.39
===================================================================================
```

Regressors:
x1 - I10GDP
x2 - I05HICP_T)
x3 - I05HICP_R
x4 - UNEMPLOYMENT
x5 - I10LABOUR

ANNEX: ETL PROCESS + SQL FINAL TABLE



Package execution completed with success. Click here to switch to design mode, or select Stop Debugging from the Debug menu.

SQLQuery5.sql - MSI.CKDU36 (MSI\anuch (172)) - Microsoft SQL Server Management Studio

```
SELECT TOP (1000) [YEAR-Q]
      ,[Country]
      ,[I10GDP]
      ,[I05HICP_T]
      ,[I05HICP_R]
      ,[I15HPI]
      ,[UNEMPLOYMENT]
      ,[I10LABOUR]
      ,[KEY1]
  FROM [CKDU36].[dbo].[STATDATA]
```

	YEAR-Q	Country	I10GDP	I05HICP_T	I05HICP_R	I15HPI	UNEMPLOYMENT	I10LABOUR	KEY1
1	2010-Q1	AT	98.00499	108.63999	113.76000	75.05000	5.40000	98.77799	AT2010-Q1
2	2010-Q2	AT	99.46499	109.65000	114.20999	77.20999	5.40000	99.74099	AT2010-Q2
3	2010-Q3	AT	100.67700	109.40000	115.48000	78.19000	5.09999	100.44100	AT2010-Q3
4	2010-Q4	AT	101.85299	110.43000	116.01999	78.83999	4.90000	101.04000	AT2010-Q4
5	2011-Q1	AT	102.55200	111.91000	116.65000	78.95999	5.19999	101.49400	AT2011-Q1
6	2011-Q2	AT	103.35600	113.69999	117.73000	81.27999	4.80000	101.44100	AT2011-Q2
7	2011-Q3	AT	103.71499	113.58000	119.54000	82.58999	4.30000	101.63999	AT2011-Q3
8	2011-Q4	AT	103.93800	114.48999	120.62000	83.05999	5.19999	101.30000	AT2011-Q4
9	2012-Q1	AT	104.66200	114.94999	121.69999	85.30000	4.90000	101.79900	AT2012-Q1
10	2012-Q2	AT	103.91600	116.26999	123.59999	85.44000	5.19999	101.25800	AT2012-Q2
11	2012-Q3	AT	103.87599	116.30999	124.50000	88.70999	5.40000	100.87200	AT2012-Q3
12	2012-Q4	AT	103.56700	117.83000	125.58000	88.00000	5.30000	100.75800	AT2012-Q4
13	2013-Q1	AT	103.72100	117.94999	126.58000	89.05999	5.90000	100.82399	AT2013-Q1
14	2013-Q2	AT	103.85900	118.87999	127.20999	92.25000	5.50000	100.60399	AT2013-Q2
15	2013-Q3	AT	104.13099	118.59999	128.02000	91.79000	5.69999	100.70899	AT2013-Q3
16	2013-Q4	AT	104.70999	119.76999	128.91999	91.68000	5.90000	101.00800	AT2013-Q4

Query executed successfully. MSI (16.0 RTM) MSI\anuch (172) CKDU36 00:00:00 1,000