

# Web Data in Official Statistics: Process, Challenges, Solutions - Online real estate Webinar

Dominik Dąbrowski, Klaudia Peszat

**Trusted Smart Statistics – Web Intelligence Network**



**Web Intelligence**  
Network



**Funded by  
the European Union**

# Outline

1. Background
2. Introduction to web scraping
3. Data collection
4. Data quality



**Web Intelligence**  
Network



**Funded by**  
**the European Union**

# Background

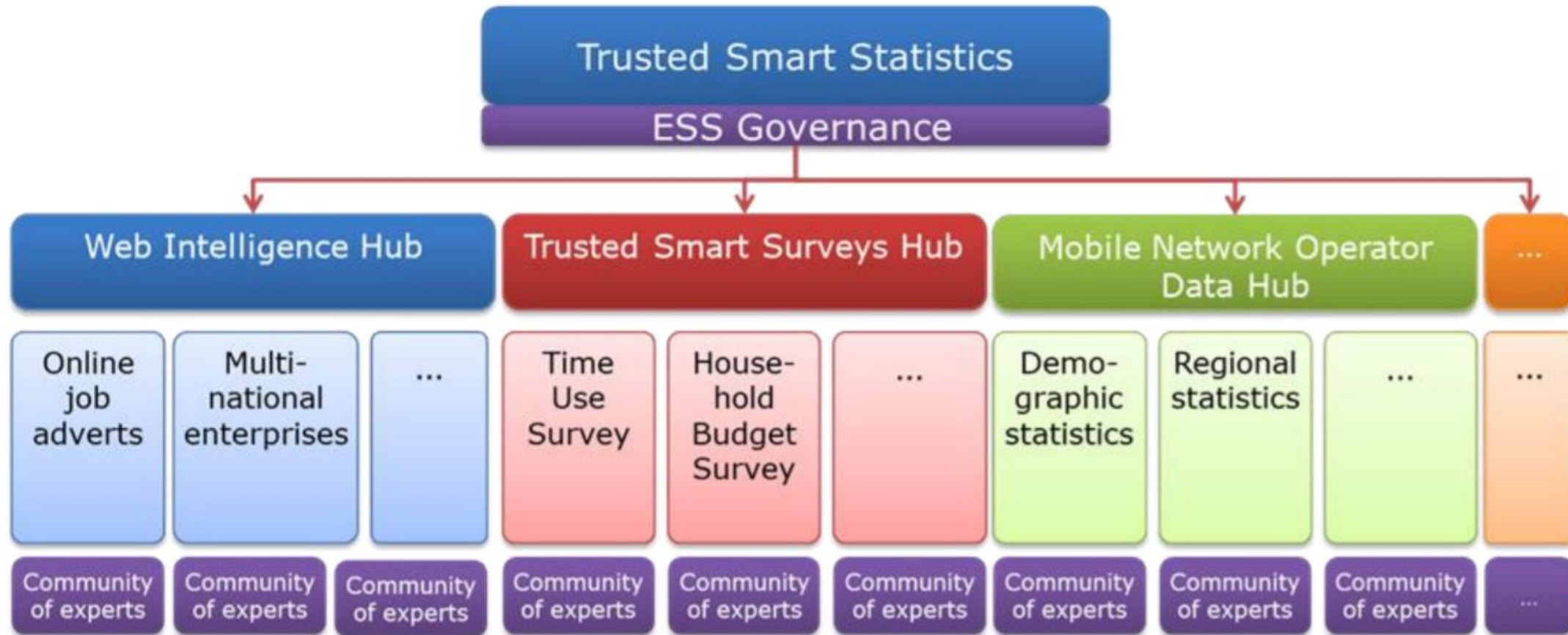


**Web Intelligence**  
Network



**Funded by**  
**the European Union**

# Web data in the TSS landscape



# Potential use-cases in the Web Intelligence Hub

## Towards production of official statistics

Online Job Advertisements

Online Based Enterprise Characteristics

## New use-cases

Real estate market

Construction activities

Online prices of household appliances and audio-visual, photographic and information processing equipment

Tourism statistics

Business register quality enhancement

Faster Economic Indicators



# Potential use-cases in the Web Intelligence Hub

## Towards production of official statistics

Online Job Advertisements

Online Based Enterprise Characteristics

## New use-cases

Real estate market

Construction activities

Online prices of household appliances and audio-visual, photographic and information processing equipment

Tourism statistics

Business register quality enhancement

Faster Economic Indicators



**Web Intelligence**  
Network



**Funded by**  
**the European Union**

# Characteristics of real estate market

6 partners

Web scraping and data providers

Different applications

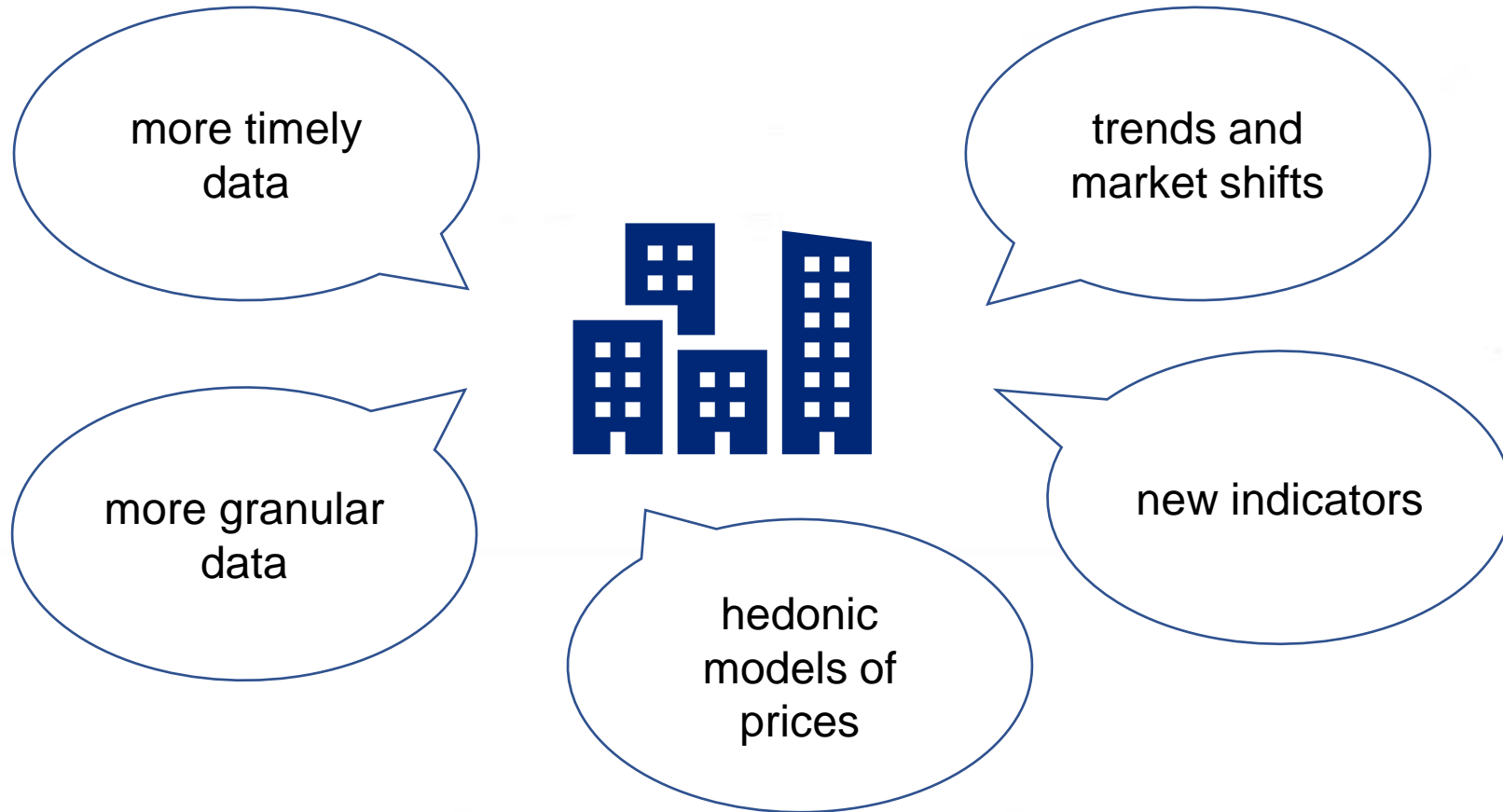


**Web Intelligence**  
Network



**Funded by**  
**the European Union**

# Application of web data in real estate statistics





# Statistics Poland's approach

- Lack of reference source for rental market in Poland
- Seeking for new data sources
- Exploration of web data from real estate portals



**Web Intelligence**  
Network



**Funded by  
the European Union**

# Introduction to web scraping



**Web Intelligence**  
Network



**Funded by  
the European Union**

**Slido**



**Web Intelligence**  
Network



**Funded by**  
**the European Union**

Multiple-choice poll

Have you ever heard about web scraping?

0 2 9

Yes



No



slido



**Web Intelligence**  
Network



**Funded by**  
**the European Union**

Multiple-choice poll

Have you ever tried to scrape some data from any website?

035

Yes



No



slido

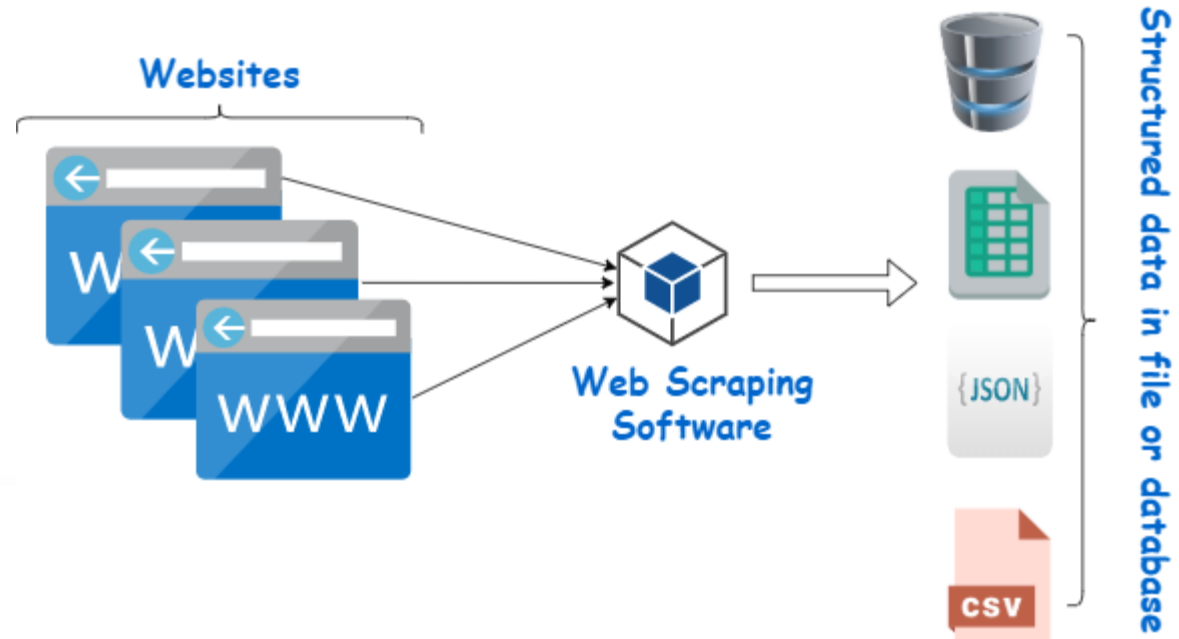


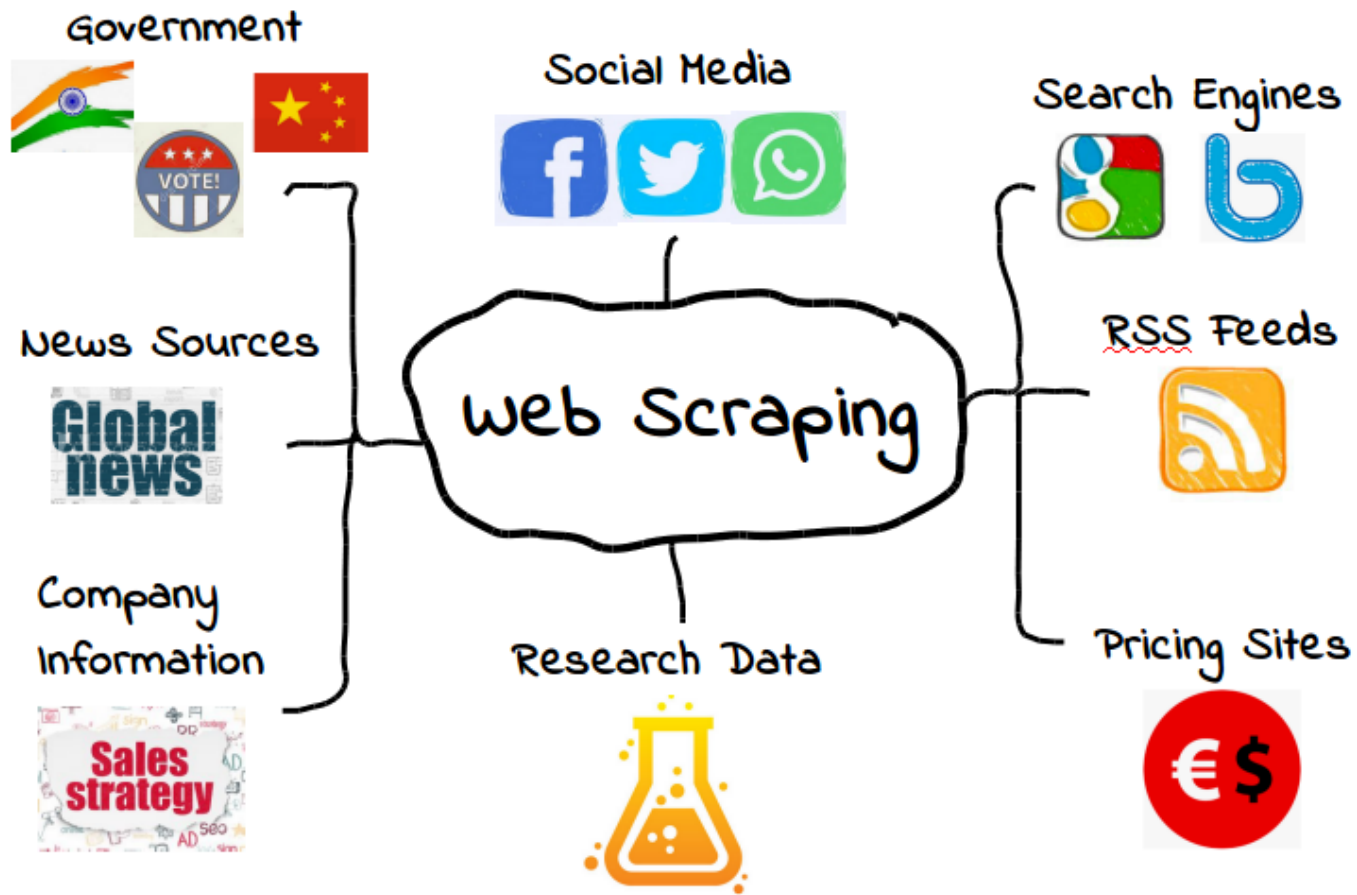
**Web Intelligence**  
Network



**Funded by  
the European Union**

# Web scraping, web harvesting, data extraction







**Crawling**

**Scraping**



**Web Intelligence**  
Network



**Funded by**  
**the European Union**



# Is Web Scraping Legal ?



LinkedIn Vs HiQ [[LINK](#)]

Facebook Vs Power Ventures [[LINK](#)]

French Data Protection Authority [[LINK](#)]

<https://stat.gov.pl/robots.txt>

<https://www.amazon.com/robots.txt>

What is  
Robots.txt?



**Web Intelligence**  
Network



**Funded by  
the European Union**

# Data collection



**Web Intelligence**  
Network



**Funded by  
the European Union**

# Criteria for assessing websites

1	Checklist for quality assessment of web data sources				
2	Use case	Use-case 1: Characteristics of the real estate market			
3	Partner	GUS			
4	Version: 2021-09-17				
5	Web Data Source	[redacted].pl	[redacted].pl	[redacted].pl	[redacted].pl
6	Score (range from 0 to 100)	82	80	89	75
7	<b>Stop criteria</b> <i>If at least one Stop criteria has a value of 1 then the web source is rejected</i>				
8	Captcha	0	0	0	0
9	Robots blocking	0	0	0	0
10	CDN	0	0	0	0
11	<b>Minimal criteria</b> <i>All present criteria should have a value of 1, otherwise the web source is rejected</i>				
12	List of pages	1	1	1	1
13	Filter criteria	1	1	1	1
14	GET HTTP method	1	1	1	1
15	Up to date content	1	1	1	1
16	Number of ads >10000	1	1	1	0
17	Structured description	1	1	1	1
18	<b>Additional criteria</b> <i>These criteria increase the viability of the web data source</i>				
19	Specific time period filter	1	0	1	0
20	robots.txt	1	0	1	1
21	Multilanguage	0	0	1	0
22	Aggregator	0	0	0	0



# Landscaping – criteria for assessing websites

- Accessibility
- Data quality
- Data structure
- Legal and ethical considerations

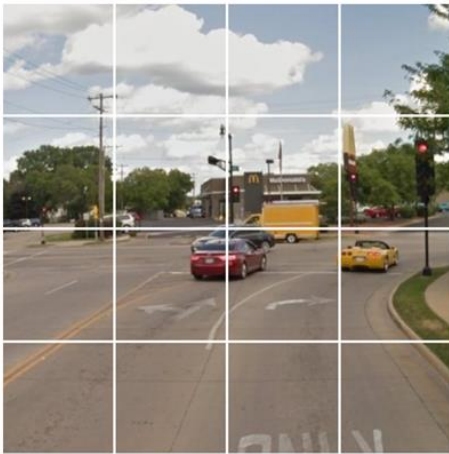





# Landscaping - selection of web data sources

- Significance of the web source (number of offers, share in the market)
- Technical restrictions (e.g. robots blocking, robots.txt, etc.)
- Mechanisms of navigating on the page and page's structure
- Completeness of the information
- Up to date content




Select all squares with  
**traffic lights**  
If there are none, click skip



   [SKIP](#)

Please check the box below to proceed.

I'm not a robot



reCAPTCHA  
Privacy - Terms



Type the characters above:

[Go](#)

okta



```
sakanal_blocked.ipynb > M+empty cell
+ Code + Markdown | ▶ Run All ≡ Clear All Outputs ↻ Restart | 📄 Variables ☰ Outline ...

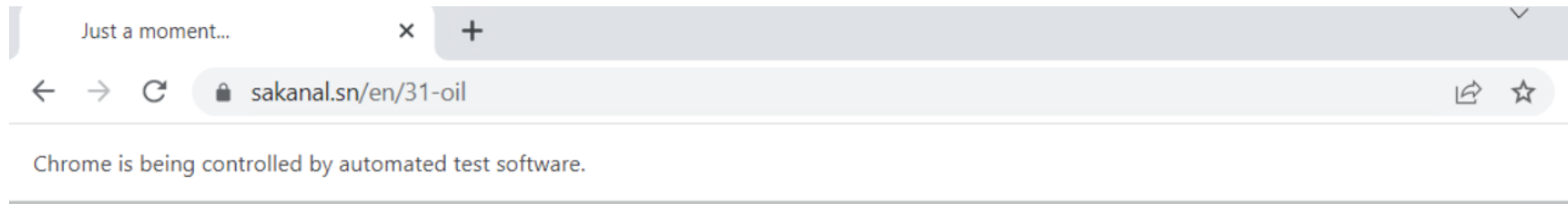
import requests
from bs4 import BeautifulSoup

homepage = 'https://sakanal.sn/en/31-oil'
a = requests.get(homepage)
page = BeautifulSoup(a.text, 'html.parser')
print(page)

[ ]

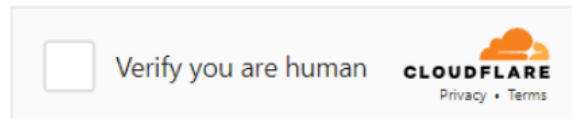
<div id="challenge-error-title">
<div class="h2">
<span class="icon-wrapper">
<div class="heading-icon warning-icon"></div>
</span>
<span id="challenge-error-text">
    Enable JavaScript and cookies to continue
</span>
</div>
</div>
...
</script>
</body>
</html>
```





# sakanal.sn

## Checking if the site connection is secure



sakanal.sn needs to review the security of your connection before proceeding.

Why am I seeing this page? ▾



**Web Intelligence**  
Network

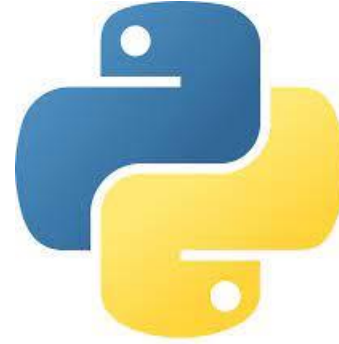


**Funded by  
the European Union**



# Preparing Dedicated Tools

- Scrapers (Python, R, Ruby)
- IDE (PyCharm, VS Code, Anaconda)



Slido



**Web Intelligence**  
Network



**Funded by  
the European Union**

Open text poll

### What programming language/software do you use for data processing, data extraction, data management?

(1/2)

0 2 7

- Python, R
- R
- Python
- SAS,R, Python
- www.codeium.com (free AI assistant)
- Python, sas
- R, Python
- SAS
- R
- R, python, vs code, rstudio, selenium
- R, python, nodejs
- R
- SQL
- DuckDB for local database
- SAS, R, Python
- python bs4 , sql server
- R
- R/SQL
- Python
- Python, Spyder and Jupyter Notebook
- Python
- R, Python, KNIME, Selenium, SAS
- SAS

slido

Open text poll

### What programming language/software do you use for data processing, data extraction, data management?

(2/2)

0 2 7

- R/MySQL/Python
- R, SQL
- dbt
- Python and R
- Python, R, SQL
- Python with codeium
- R, Python
- Python and R
- sas, R

slido



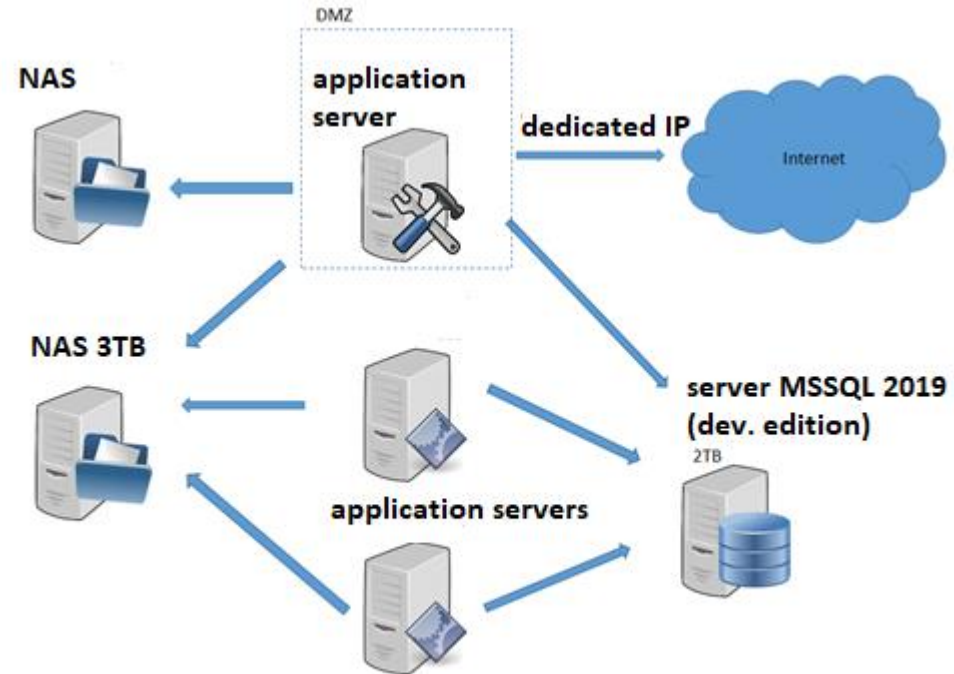
**Web Intelligence**  
Network



**Funded by**  
**the European Union**

# Preparing Environment

- Disk space
- Servers
- Databases



# Common Problems During Web Scraping

- CAPTCHA
- Cookies
- Dynamic content (with scrolling)
- Structure changes
- Blocking access after too many requests
- Cloudflare
- [Honeypot](#)
- Configure the environment for multiple users
- Up-to-date content (list of offers modified during scraping process)
- Server errors/updates/restarts



```
▼<urlset xmlns="http://www.sitemaps.org/schemas/sitemap/0.9" xmlns:xhtml="http://www.w3.org/1999/xhtml">
  ▼<url>
    <loc>https://www.otodom.pl</loc>
    <lastmod>2023-05-14T04:25:28+02:00</lastmod>
    <changefreq>daily</changefreq>
    <priority>0.9</priority>
  </url>
  ▼<url>
    <loc>https://www.otodom.pl/pl/oferta/2-rozkladowe-pokoje-ksiazat-pomorskich-ID4lsFj</loc>
    <lastmod>2023-05-14T03:55:22+02:00</lastmod>
    <changefreq>weekly</changefreq>
    <priority>0.3</priority>
  </url>
  ▼<url>
    <loc>https://www.otodom.pl/pl/oferta/gotowy-139m2-55m2-poddasza-5km-od-bemowa-metro-ID4lsFi</loc>
    <lastmod>2023-05-14T03:39:07+02:00</lastmod>
    <changefreq>weekly</changefreq>
    <priority>0.3</priority>
  </url>
  ▼<url>
    <loc>https://www.otodom.pl/pl/oferta/mieszkanie-gdansk-wrzeszcz-ID4lsFh</loc>
    <lastmod>2023-05-14T02:55:22+02:00</lastmod>
    <changefreq>weekly</changefreq>
    <priority>0.3</priority>
  </url>
  ▼<url>
    <loc>https://www.otodom.pl/pl/oferta/ta-dzialka-czeka-na-ciebie-domatowo-ID4lsFg</loc>
    <lastmod>2023-05-14T02:45:20+02:00</lastmod>
    <changefreq>weekly</changefreq>
    <priority>0.3</priority>
  </url>
  ▼<url>
    <loc>https://www.otodom.pl/pl/oferta/mieszkanie-z-ogrodkiem-3-pokoje-68-m2-ID4lsFf</loc>
    <lastmod>2023-05-14T02:41:09+02:00</lastmod>
    <changefreq>weekly</changefreq>
    <priority>0.3</priority>
  </url>
  ▼<url>
    <loc>https://www.otodom.pl/pl/oferta/dom-jednorodzinny-w-mlynisku-kolo-kalisza-ID4lsFe</loc>
    <lastmod>2023-05-14T02:41:10+02:00</lastmod>
    <changefreq>weekly</changefreq>
    <priority>0.3</priority>
  </url>
```

```
▼<url>
  <loc>https://www.otodom.pl/pl/oferta/-ID4lqdj</loc>
  <lastmod>2023-05-11T00:05:23+02:00</lastmod>
  <changefreq>weekly</changefreq>
  <priority>0.3</priority>
</url>
▼<url>
  <loc>https://www.otodom.pl/pl/oferta/kawalerka-morasko-ID4lqdi</loc>
  <lastmod>2023-05-11T00:14:47+02:00</lastmod>
  <changefreq>weekly</changefreq>
  <priority>0.3</priority>
</url>
▼<url>
  <loc>https://www.otodom.pl/pl/oferta/4-pokojowy-apartament-w-avangarden-ID4lqdh</loc>
  <lastmod>2023-05-11T12:08:10+02:00</lastmod>
  <changefreq>weekly</changefreq>
  <priority>0.3</priority>
</url>
▼<url>
  <loc>https://www.otodom.pl/pl/oferta/2-pokoje-saska-kepa-50-m2-ul-wandy-ID4lqdg</loc>
  <lastmod>2023-05-13T06:00:53+02:00</lastmod>
  <changefreq>weekly</changefreq>
  <priority>0.3</priority>
</url>
▼<url>
  <loc>https://www.otodom.pl/pl/oferta/dzialka-uslugowo-mieszkaniowa-o-pow-3-57ha-ID4lqdf</loc>
  <lastmod>2023-05-10T23:55:36+02:00</lastmod>
  <changefreq>weekly</changefreq>
  <priority>0.3</priority>
</url>
```

# Lessons learned

On the organizational level:

- a need to control the vast amount of open-sourced programmes
- creating a common environment
- WIH - a possibility to use the platform for web scraping and working well with our systems



**Web Intelligence**  
Network



**Funded by**  
**the European Union**



# Lessons learned

On the level of the scraper:

- Focus on inspecting the page
- monitor how the data behaves in time series
- keep logs of the executed programs (you may need them in the future)
- if aggregating data on the level of a category, be sure all the products are properly categorized on the page
- keep monitoring/validating your data



# Data quality

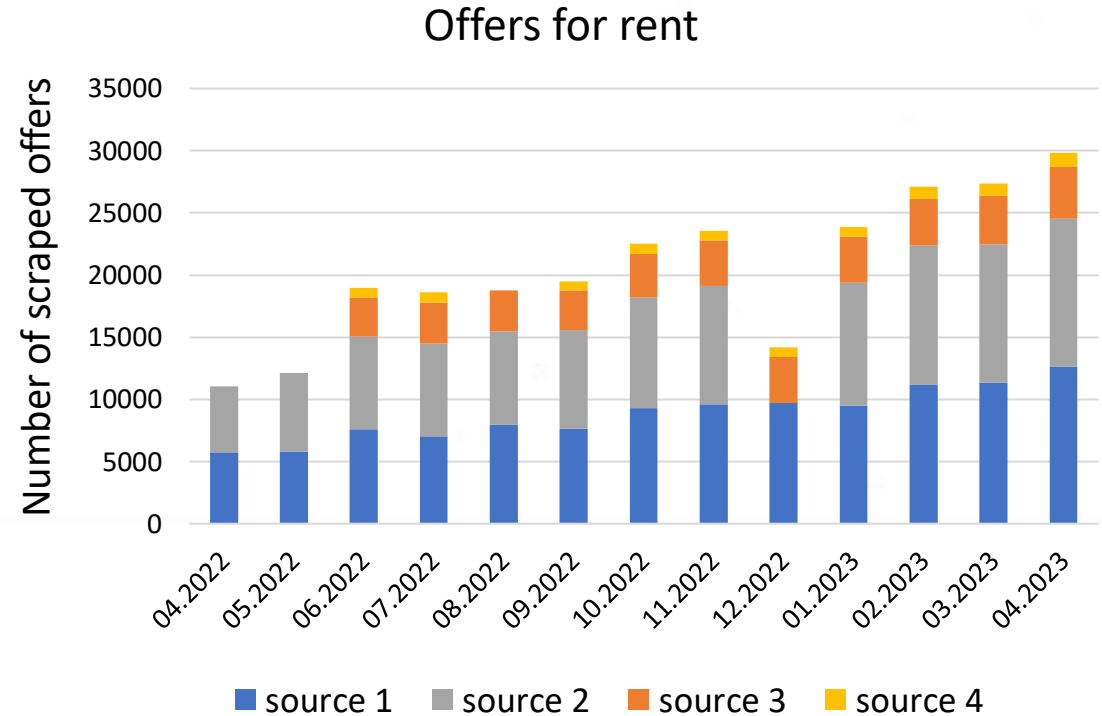


**Web Intelligence**  
Network



**Funded by  
the European Union**

# Stability of web data sources over time



# Preliminary quality assessment

Quality indicator	Definition	Value
Sample size	Number of offers on the website	102957
Unit non-response	Number of offers with negative code response	1440
Collected offers	Number of offers with positive code response (scraped)	101517
Missing values	Number of cells filled incorrectly or empty*	20364
Missing values %	Number of cells filled incorrectly or empty divided by number of all cells*	4,01%

\* in the mandatory variables

Source: Data source 1, offers for sale and rent, April 2023.



**Web Intelligence**  
Network



**Funded by**  
**the European Union**

# Missing data

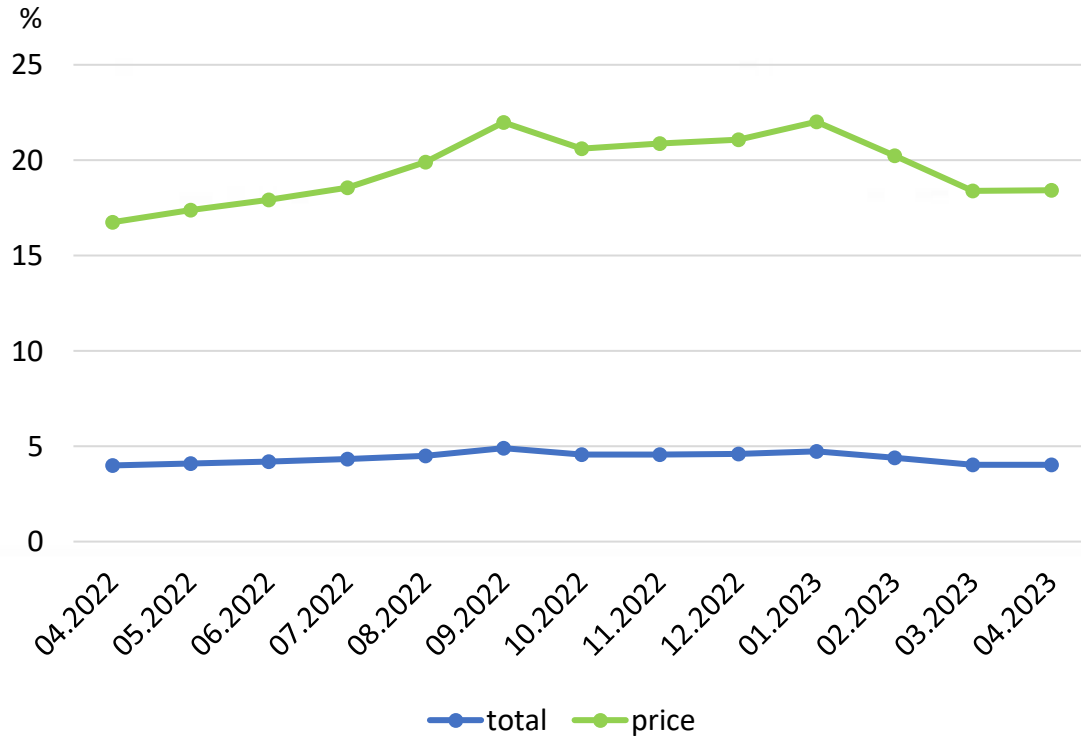
	<b>Total</b>	<b>Price</b>	<b>Area</b>	<b>Rooms</b>	<b>Floor</b>	<b>Location</b>
Source 1	4,01%	18,43%	0,08%	0,26%	1,29%	0,00%
Source 2	4,72%	19,41%	0,88%	1,17%	2,15%	0,00%
Source 3	0,18%	0,92%	0,00%	0,00%	100,0%	0,00%
Source 4	11,08%	0,01%	0,00%	27,34%	28,04%	0,00%

Source: Offers for sale and rent, April 2023.



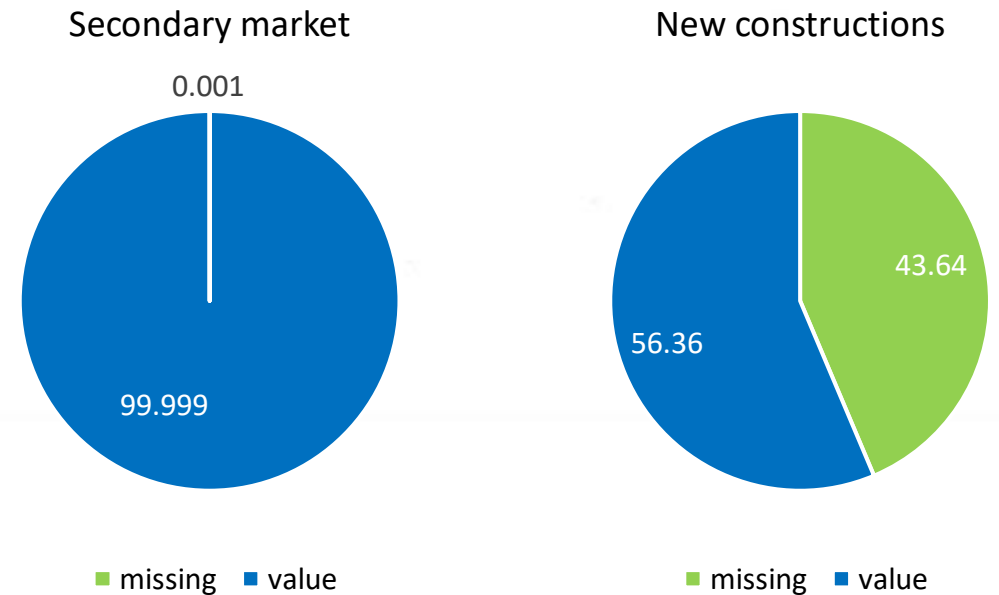
# Missing data

## Share of missing data by month



Source: Data source 1, offers for sale and rent.

## Average month share of missing values in the price variable by market type



Source: Data source 1, offers for sale.



# Missing data

- How to deal with it?
- Drop them or impute missing values - if so, what kind of method?



# Errors

- Extreme values (very high or very small) – area, price
- Building plots (lands) for sale
- Categorized variables (above the 8th floor)
- Multi-offers (apartments in new constructions)





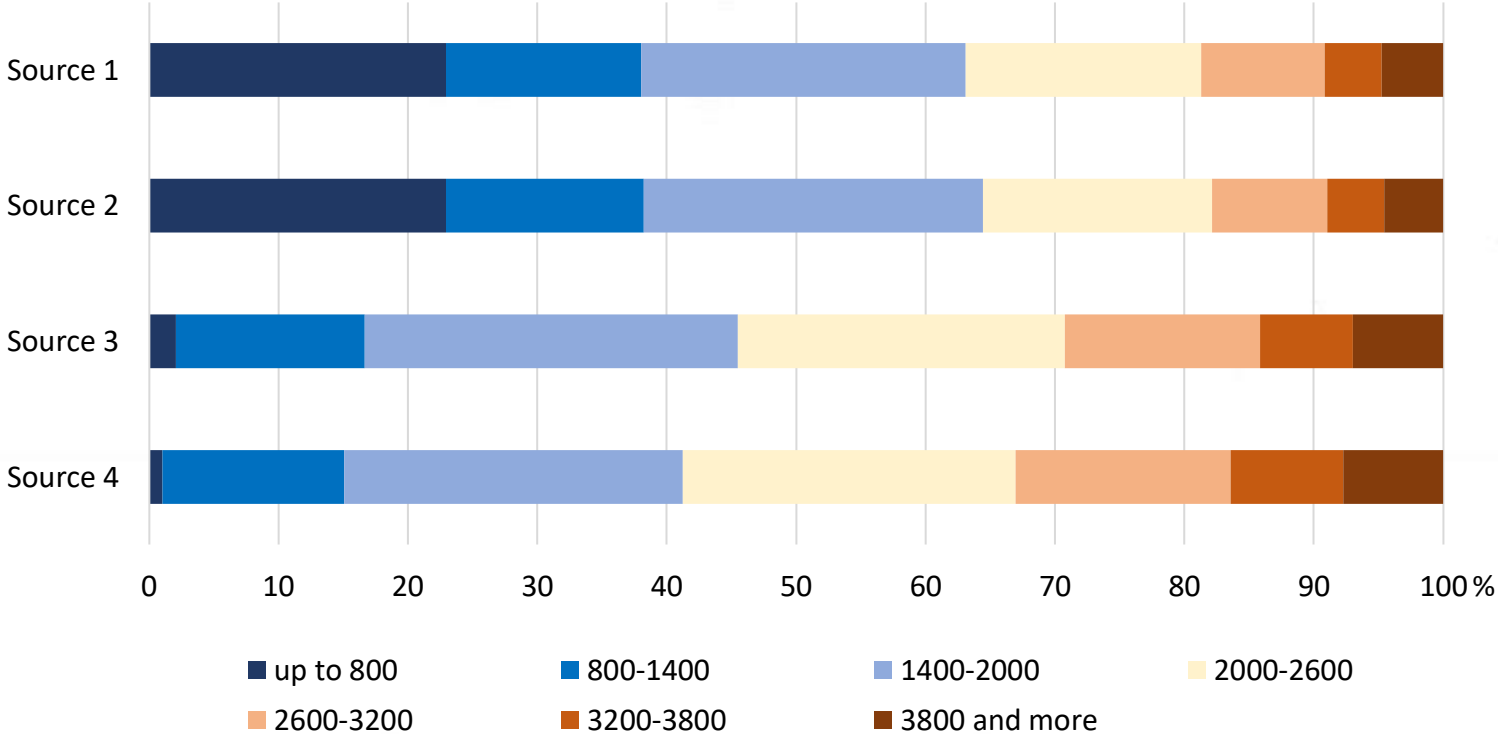
# Duplicates

- Only 0,5-2,0% duplicates within a monthly database (analysis by ID or URL)
- But the issue is more complex...
  - update of offer with new ID and different characteristics
  - overlap between portals
  - multi-offers in new buildings



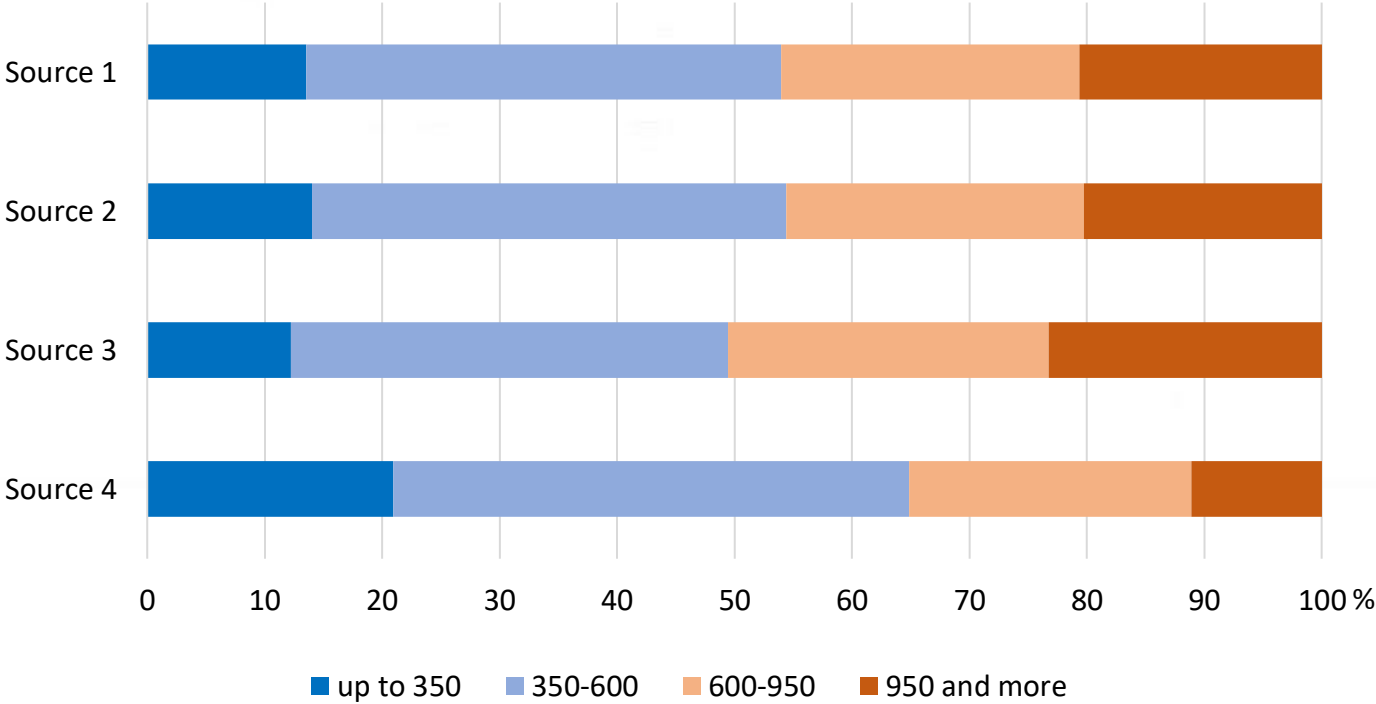
# Structure of variables

Share of offers for sale by price in EUR per m2

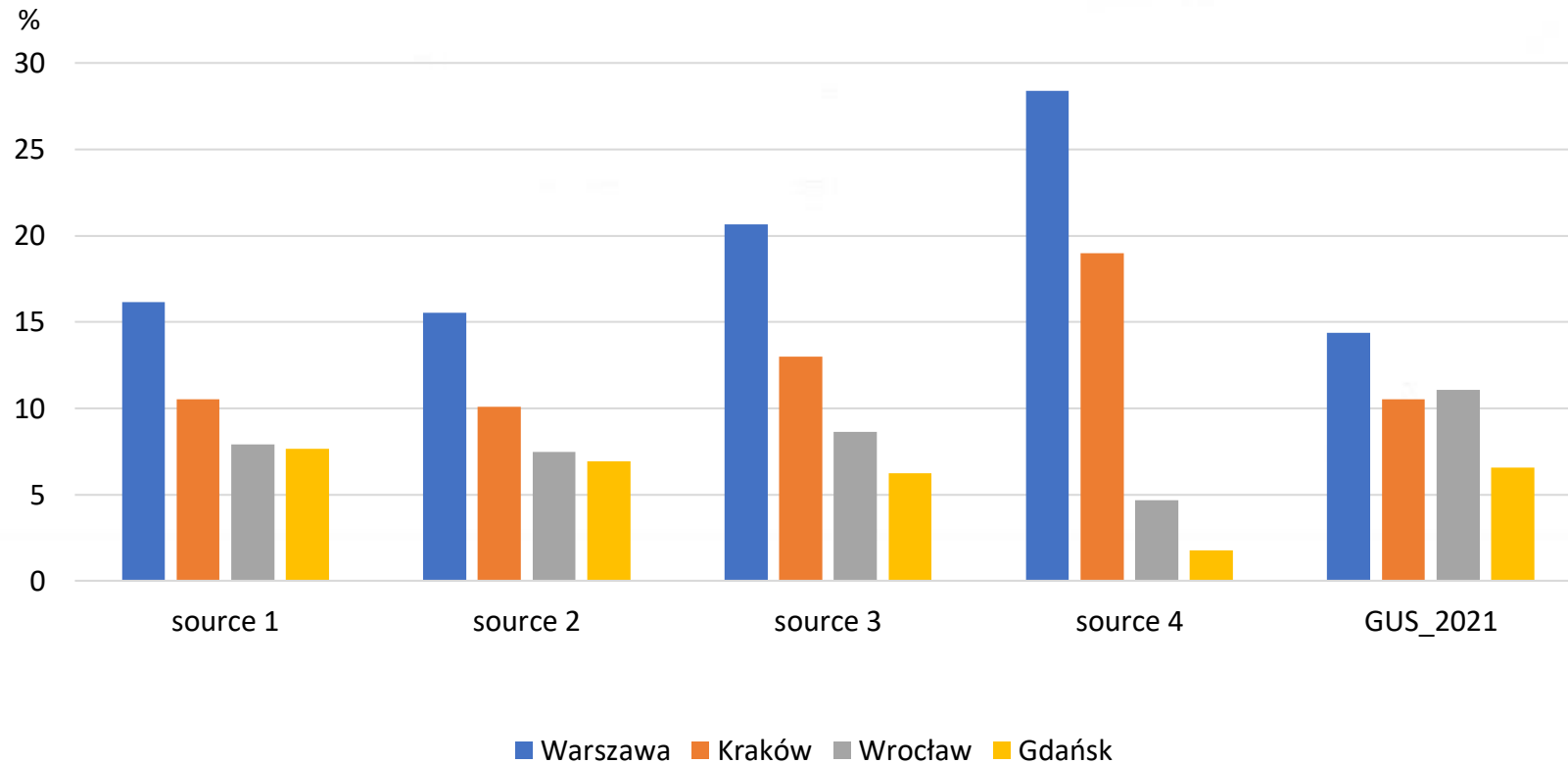


# Structure of variables

Share of offers for rent by price in EUR



# Territorial distribution



# Quality and methodology – challenges

- Duplication of offers (within a single portal, as well as across different portals)
- Multi-offers (apartments in new constructions)
- Missing values (e.g., price for apartments in new constructions)
  
- Price of offer versus price of transaction
- Undefined population



# Any questions?



**Web Intelligence**  
Network



**Funded by  
the European Union**

# Thank you!

More information on:



CROS: [https://ec.europa.eu/eurostat/cros/WIN\\_en](https://ec.europa.eu/eurostat/cros/WIN_en)



Twitter: <https://twitter.com/EssnetWin>



LinkedIn: <https://www.linkedin.com/company/essnet-project-web-intelligence-network/>

Get in touch:

Dominik Dąbrowski: [d.dabrowski@stat.gov.pl](mailto:d.dabrowski@stat.gov.pl)

Klaudia Peszat: [k.peszat@stat.gov.pl](mailto:k.peszat@stat.gov.pl)

Dominika Nowak: [do.nowak@stat.gov.pl](mailto:do.nowak@stat.gov.pl) (project coordinator)



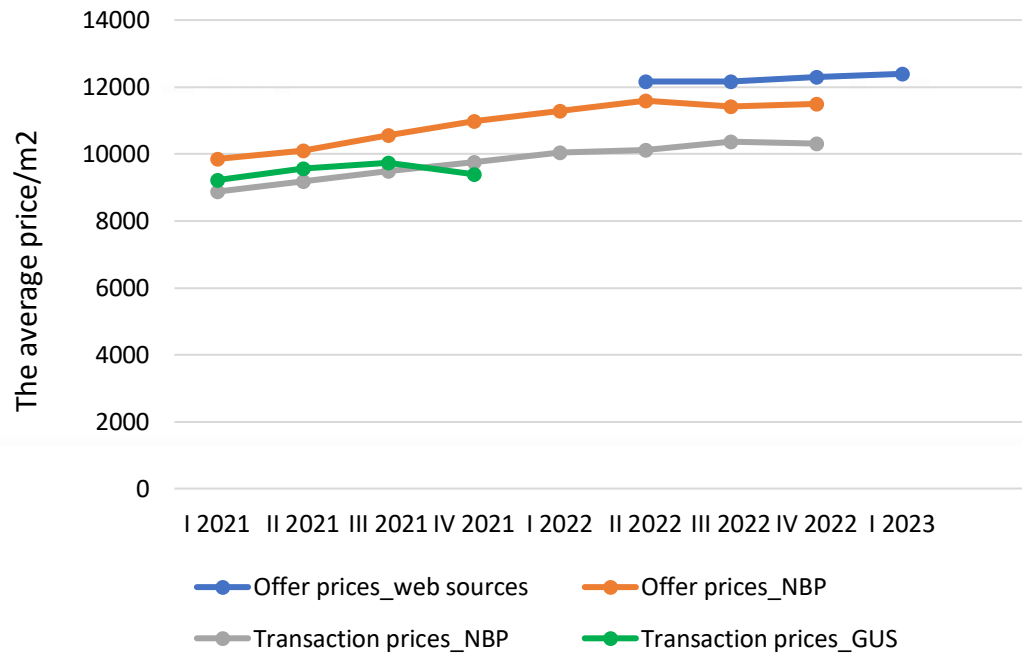
**Web Intelligence**  
Network



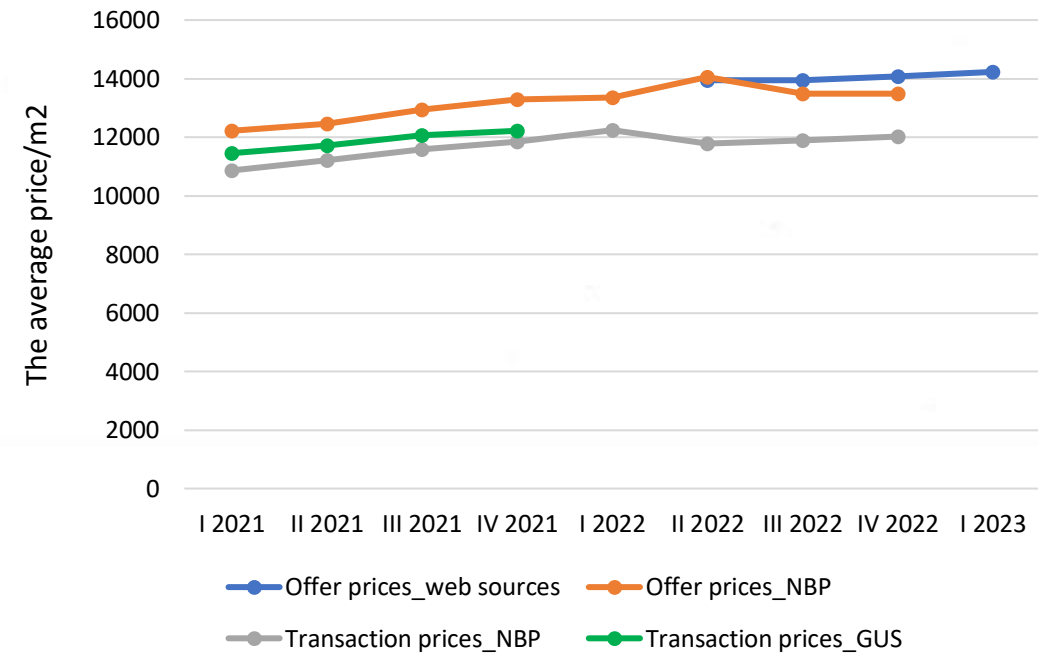
**Funded by**  
**the European Union**

# Comparison between web data and official statistics

## 7 biggest cities in Poland



## Warsaw





# Webinar Feedback Survey

- 12 Questions
  - A mixture of multiple choice and free text questions
- [www.smartsurvey.co.uk/s/WSTDP1/](http://www.smartsurvey.co.uk/s/WSTDP1/)



**Web Intelligence**  
Network



**Funded by**  
**the European Union**

Thank you for joining us today.  
Please remember to complete our  
feedback survey which has been  
sent via email.



**Web Intelligence**  
Network



**Funded by**  
**the European Union**