Method of Processing and Analysing Web Scraped Tourism Data

New data sources in tourism statistics

Łukasz Zadorożny (GUS)

23 February 2023

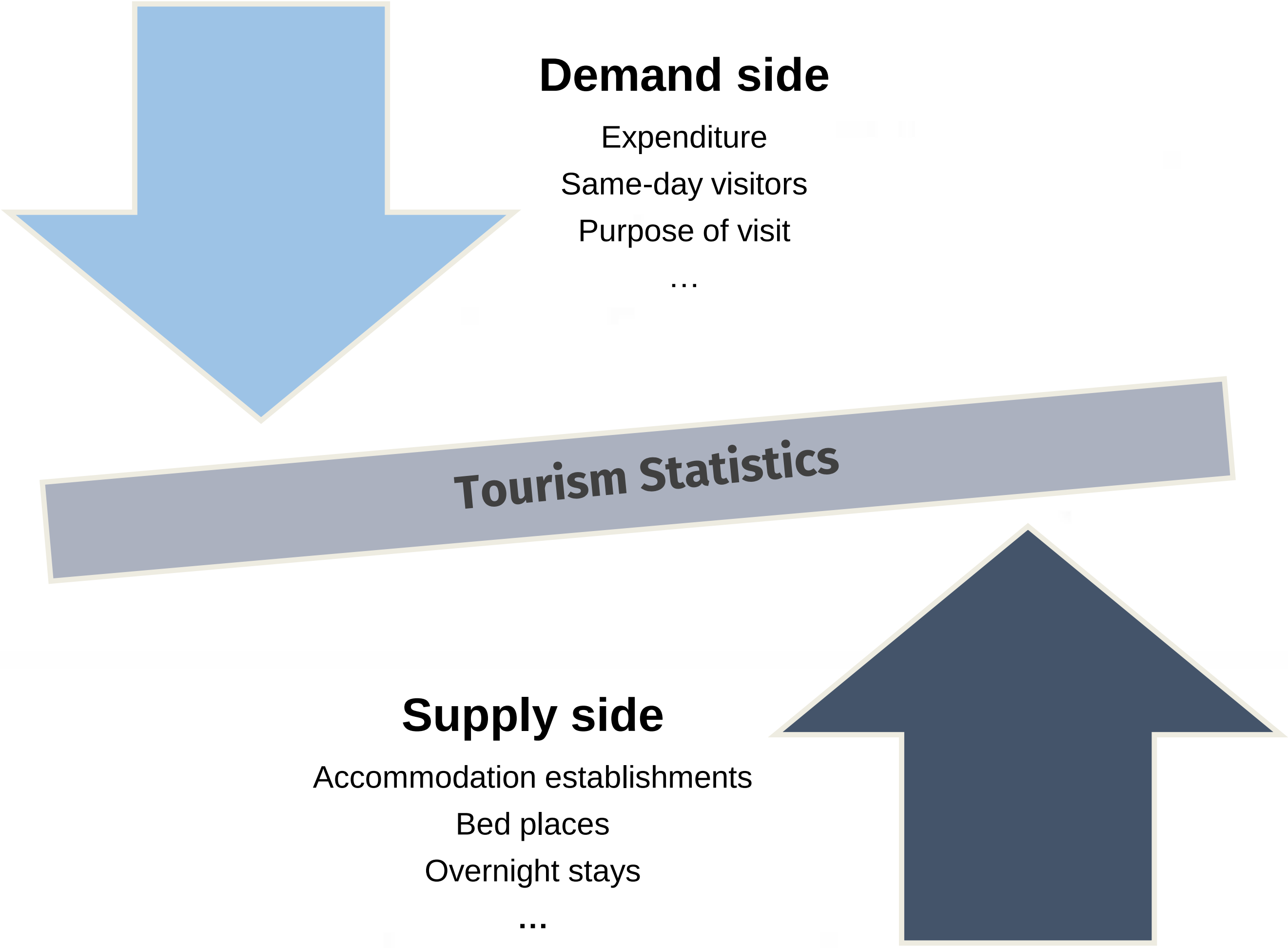
Trusted Smart Statistics – Web Intelligence Network

Grant Agreement: 101035829

Web Intelligence Network

Funded by the European Union

Tourism statistics



Demand side

Expenditure

Same-day visitors

Purpose of visit

…

Supply side

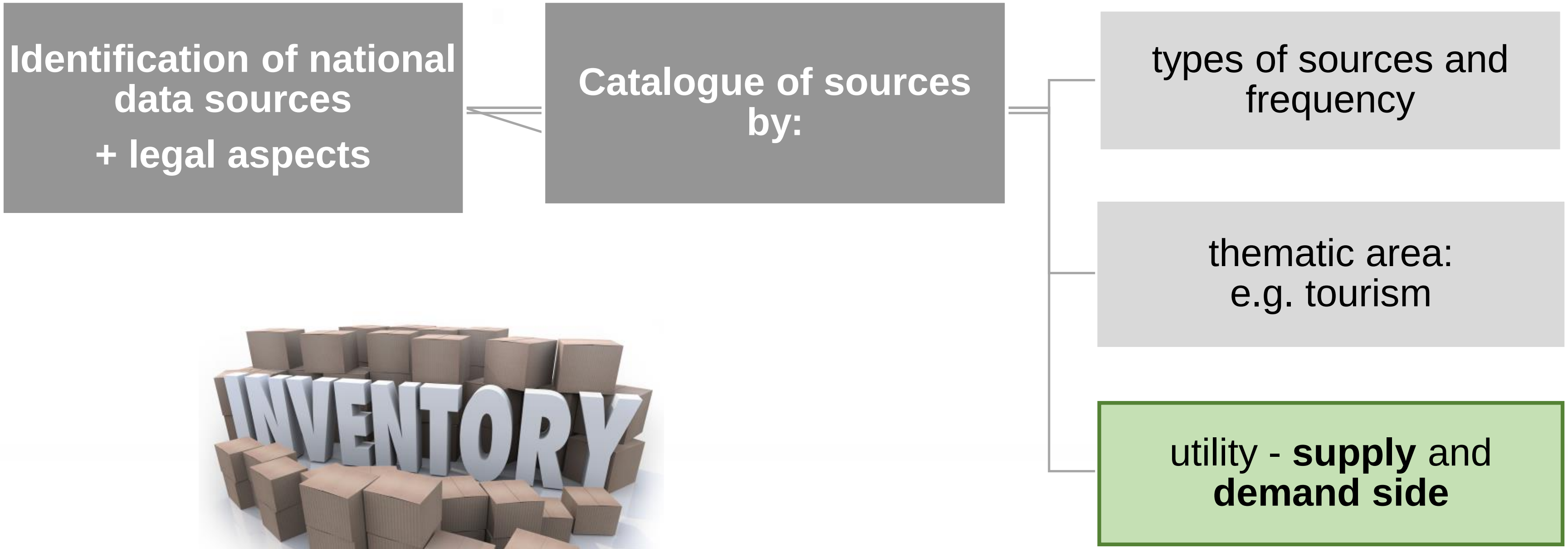
Accommodation establishments

Bed places

Overnight stays

…

External sources in tourism statistics
Inventory of data sources



Identification of national data sources

+ legal aspects

Catalogue of sources by:

types of sources and frequency

thematic area: e.g. tourism

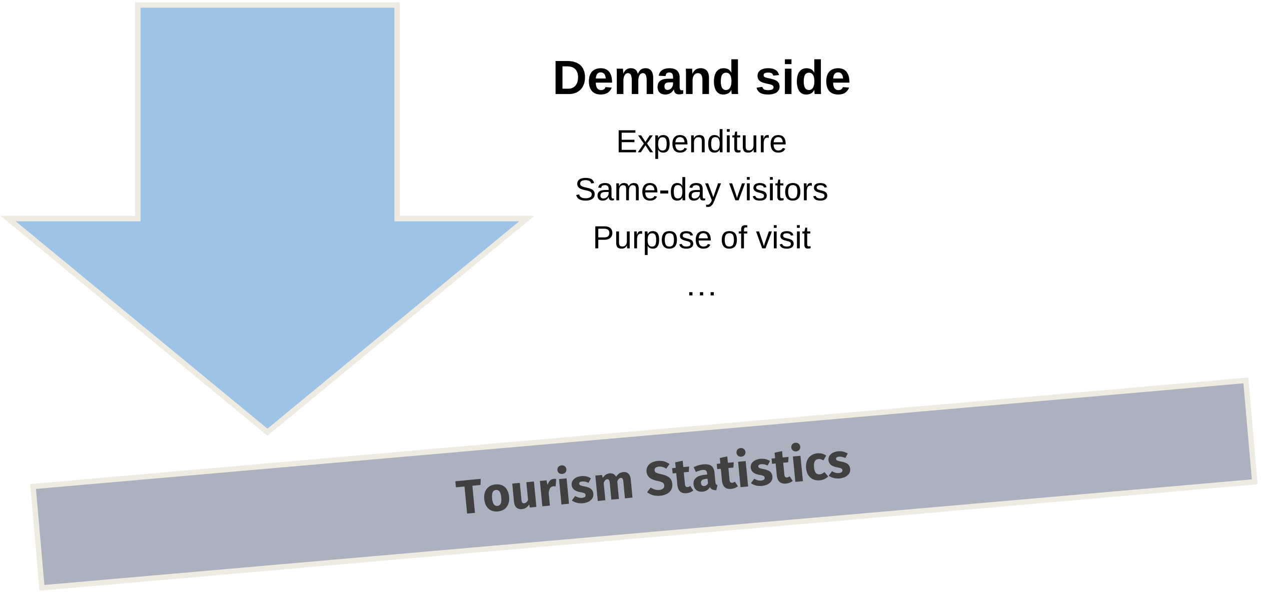
utility - **supply** and **demand side**

External sources in tourism statistics
Identification and analysis of websites

Portal name	Type	Relevance
Booking.com	Accommodation facilities	- data for the accommodation survey (new facilities) - data for trips survey (price per overnight stay)
Hotels.com		
Airbnb.com		
Tripadvisor.pl	Catering establishments	- data for estimating the costs of foreign and domestic trips
Trip.com	Portal related to aircraft flights	
Seatguru.com		Estimating tourist trips and their costs based on data on ticket prices, seats, aircraft model, etc.
Skyscanner.net		
Expedia.com		
Numbeo.com	Prices of goods and services	Living/trip cost estimation
Expatistan.com		

Funded by the European Union

External data in the demand side of tourism



Demand side

Expenditure

Same-day visitors

Purpose of visit

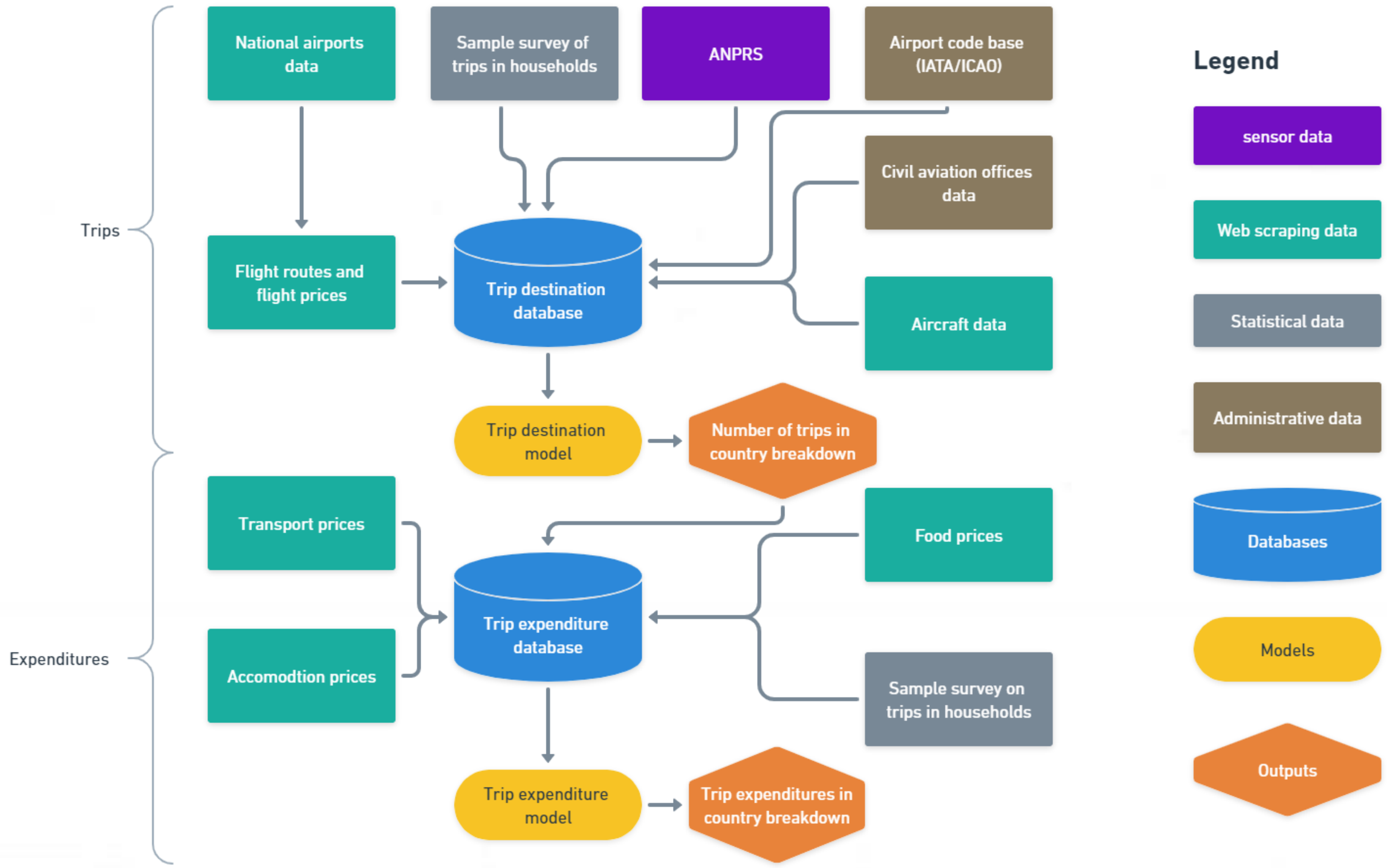
…

Tourism Statistics

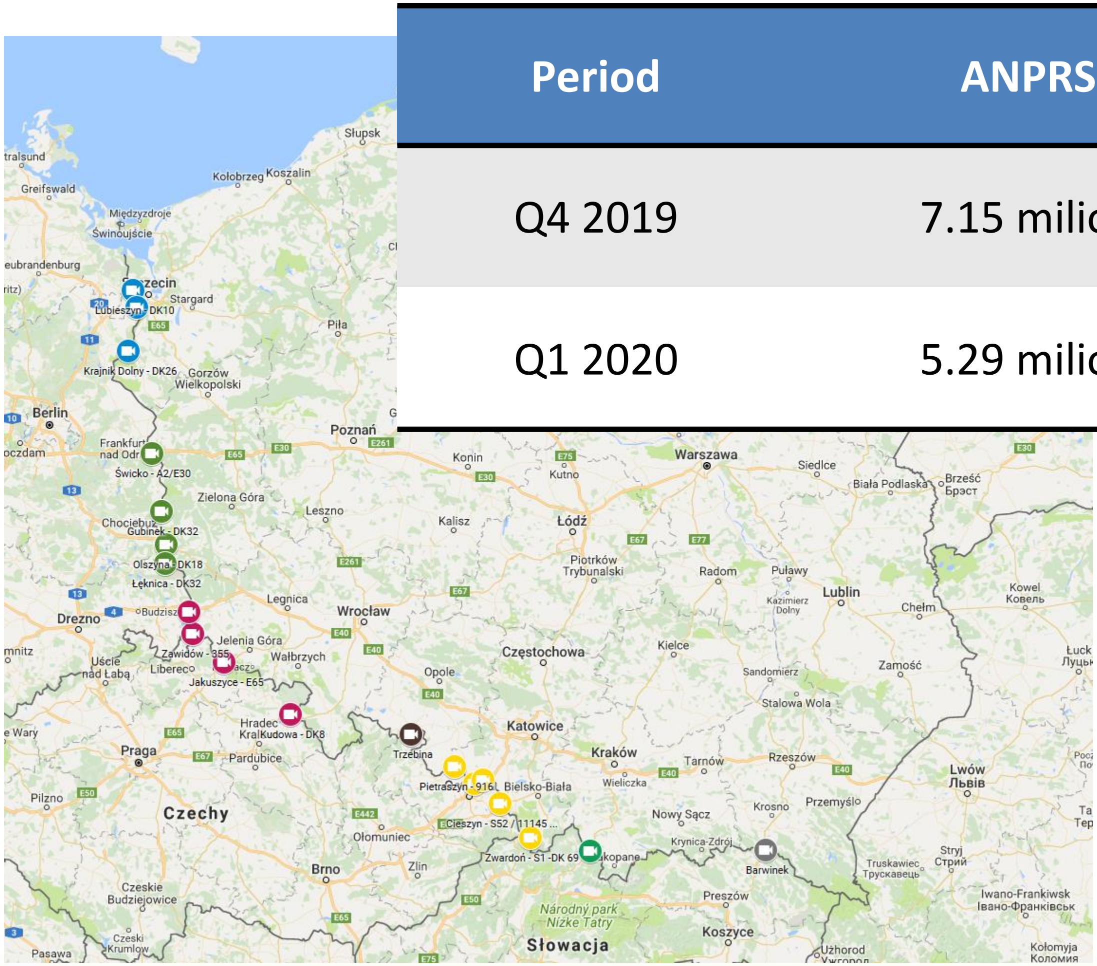
Web Intelligence Network

Funded by the European Union

External data in the demand side of tourism


Trips
Expenditures
National airports data
Sample survey of trips in households
ANPRS
Airport code base (IATA/ICAO)
Civil aviation offices data
Flight routes and flight prices
Trip destination database
Aircraft data
Trip destination model
Number of trips in country breakdown
Transport prices
Food prices
Trip expenditure database
Accomodtion prices
Sample survey on trips in households
Trip expenditure model
Trip expenditures in country breakdown
Legend
sensor data
Web scraping data
Statistical data
Administrative data
Databases
Models
Outputs

External data in the demand side of tourism



Period	ANPRS	Traffic intensity survey
Q4 2019	7.15 milion	7.17 milion
Q1 2020	5.29 milion	5.60 milion

ANPRS PL - aggregated hourly data on the number of vehicles broken down by traffic directions and vehicle type at 20 measurement points at the internal border of the EU.

External data in the demand side of tourism
Flight data – web scraping

Flight data

Booking.com

flight number

number of stops

type of aircraft

flight price

flight duration

Skyscanner.net

flight number

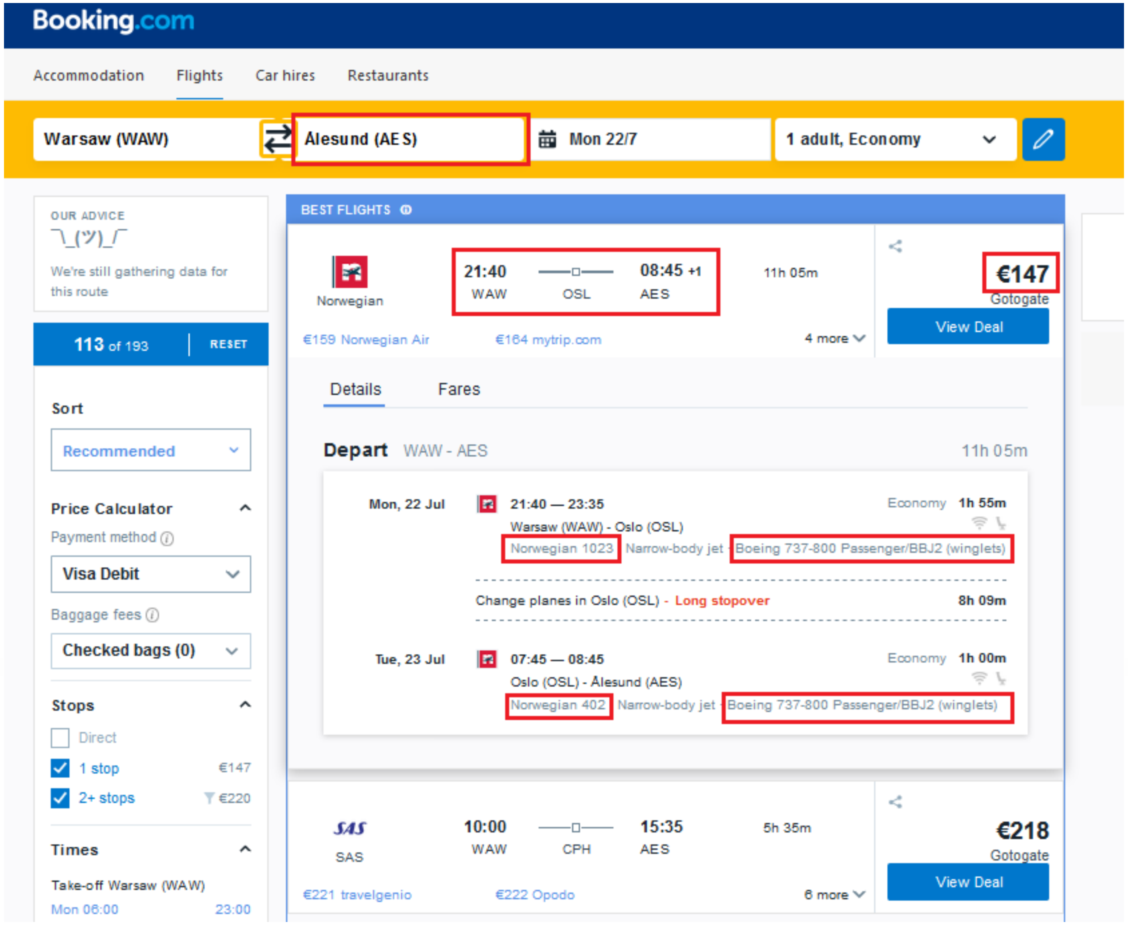
direct destination

indirect destination

Seatguru.com

number of seats in the plane

External data in the demand side of tourism
Flight data – web scraping



Warsaw (WAW) | Alesund (AES) | Mon 22/7 | 1 adult, Economy

OUR ADVICE
¯_(ツ)_/¯
We're still gathering data for this route

113 of 193 | RESET

Sort
Recommended

Price Calculator
Payment method
Visa Debit
Baggage fees
Checked bags (0)

Stops
- ☐ Direct
- ☑ 1 stop €147
- ☑ 2+ stops €220

Times
Take-off Warsaw (WAW)
Mon 06:00 23:00

BEST FLIGHTS

Norwegian | 21:40 WAW — OSL — 08:45 +1 AES | 11h 05m | €147 Gotogate | View Deal
€159 Norwegian Air | €164 mytrip.com | 4 more

Details | Fares

Depart WAW - AES 11h 05m

Mon, 22 Jul | 21:40 — 23:35 | Economy 1h 55m
Warsaw (WAW) - Oslo (OSL)
Norwegian 1023 Narrow-body jet · Boeing 737-800 Passenger/BBJ2 (winglets)

Change planes in Oslo (OSL) - Long stopover 8h 09m

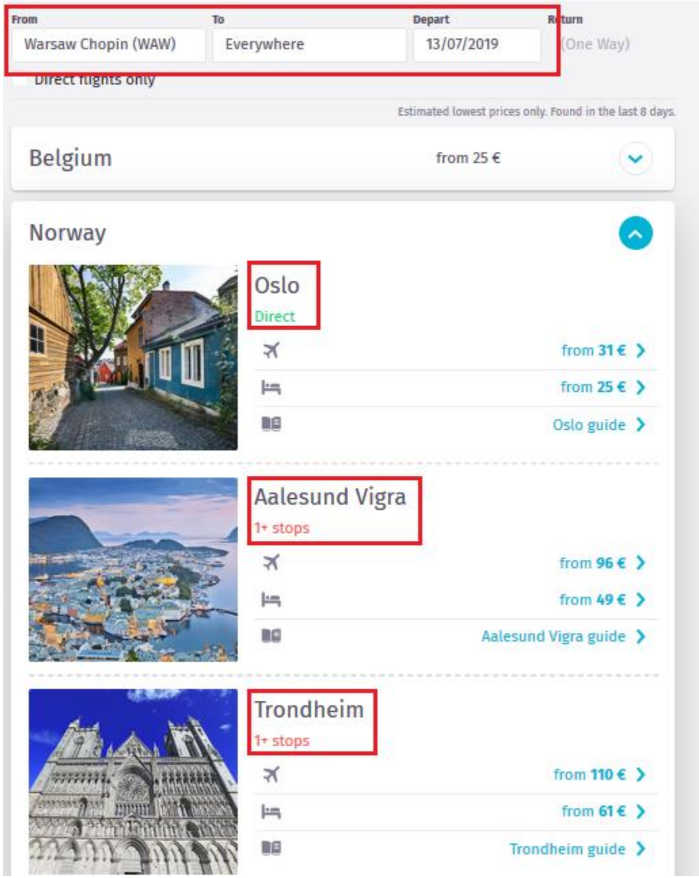
Tue, 23 Jul | 07:45 — 08:45 | Economy 1h 00m
Oslo (OSL) - Ålesund (AES)
Norwegian 402 Narrow-body jet · Boeing 737-800 Passenger/BBJ2 (winglets)

SAS | 10:00 WAW — CPH — 15:35 AES | 5h 35m | €218 Gotogate | View Deal
€221 travelgenio | €222 Opodo | 6 more

Web Intelligence Network

Funded by the European Union

External data in the demand side of tourism
Flight data – web scraping



From	To	Depart	Return
Warsaw Chopin (WAW)	Everywhere	13/07/2019	(One Way)

□ Direct flights only

Estimated lowest prices only. Found in the last 8 days.

Belgium from 25 €

Norway

Oslo
Direct
✈ from 31 €
🛏 from 25 €
Oslo guide

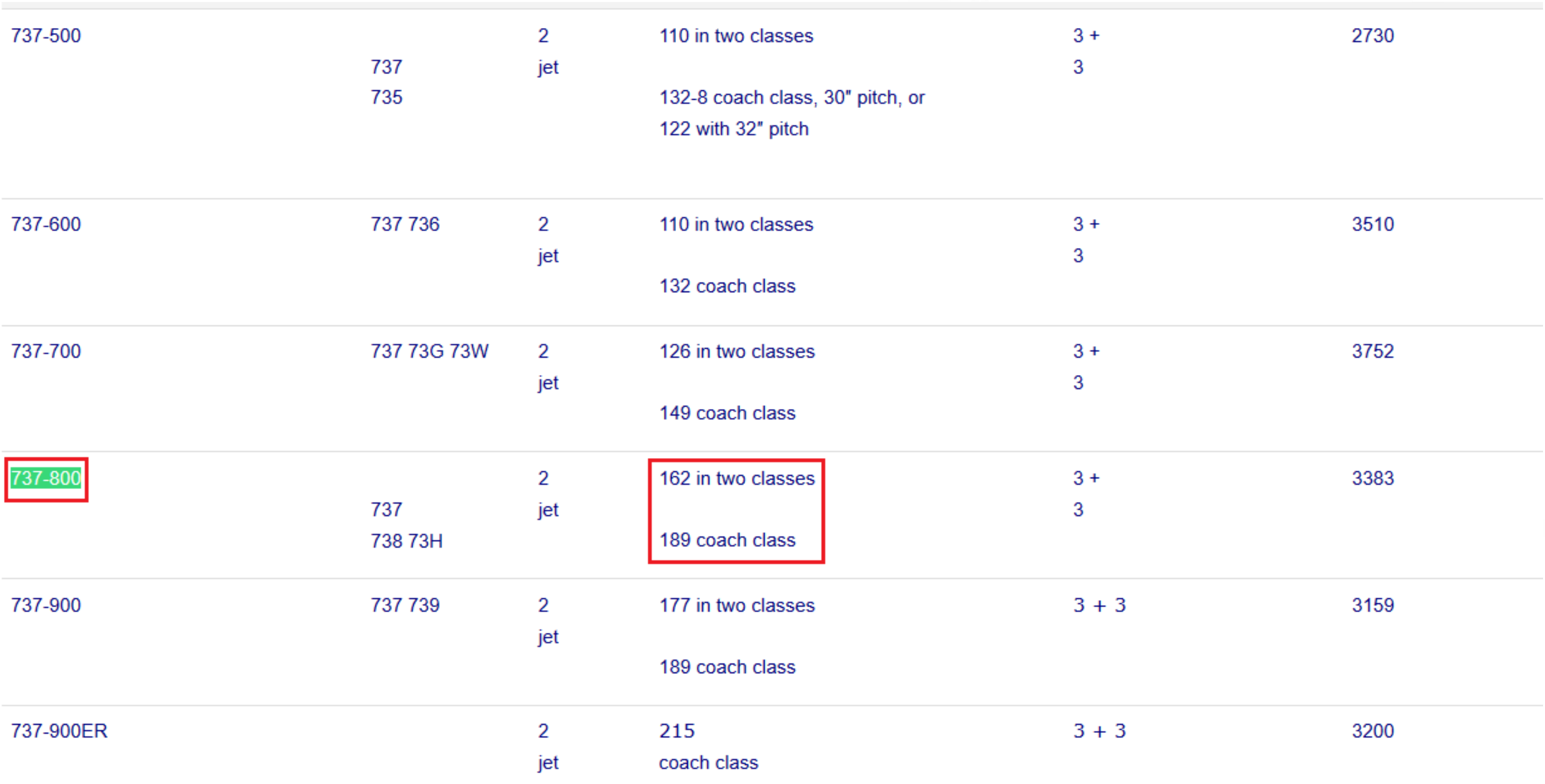
Aalesund Vigra
1+ stops
✈ from 96 €
🛏 from 49 €
Aalesund Vigra guide

Trondheim
1+ stops
✈ from 110 €
🛏 from 61 €
Trondheim guide

Funded by the European Union

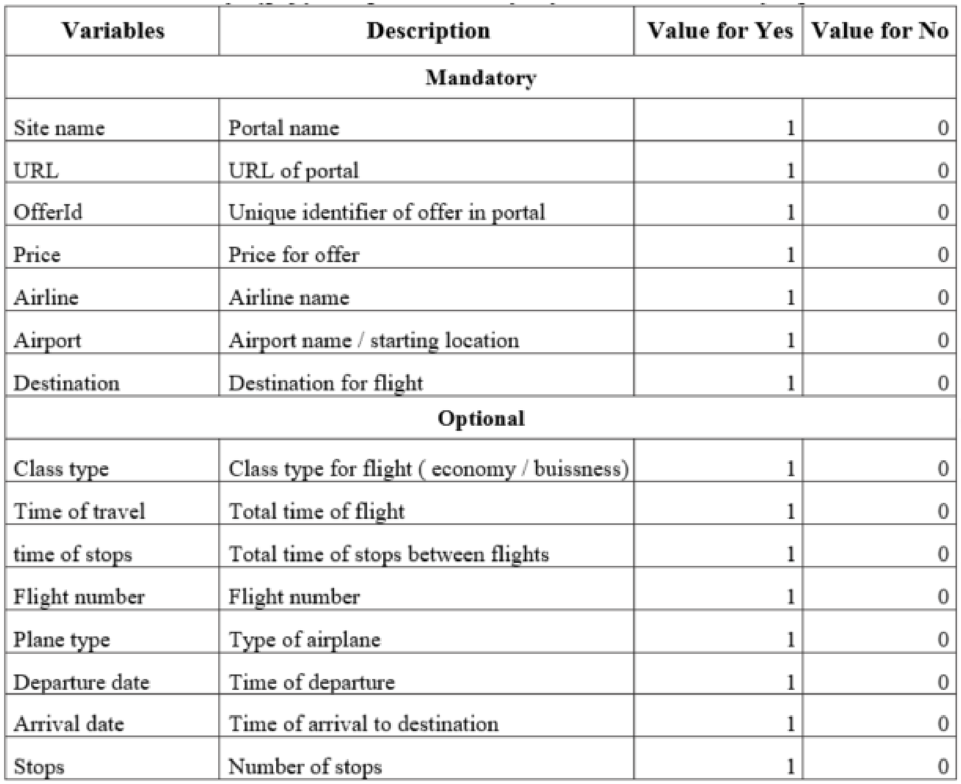
External data in the demand side of tourism
Flight data – web scraping

SeatGuru
by Tripadvisor



737-500	737 735	2 jet	110 in two classes 132-8 coach class, 30" pitch, or 122 with 32" pitch	3 + 3	2730
737-600	737 736	2 jet	110 in two classes 132 coach class	3 + 3	3510
737-700	737 73G 73W	2 jet	126 in two classes 149 coach class	3 + 3	3752
737-800	737 738 73H	2 jet	162 in two classes 189 coach class	3 + 3	3383
737-900	737 739	2 jet	177 in two classes 189 coach class	3 + 3	3159
737-900ER		2 jet	215 coach class	3 + 3	3200

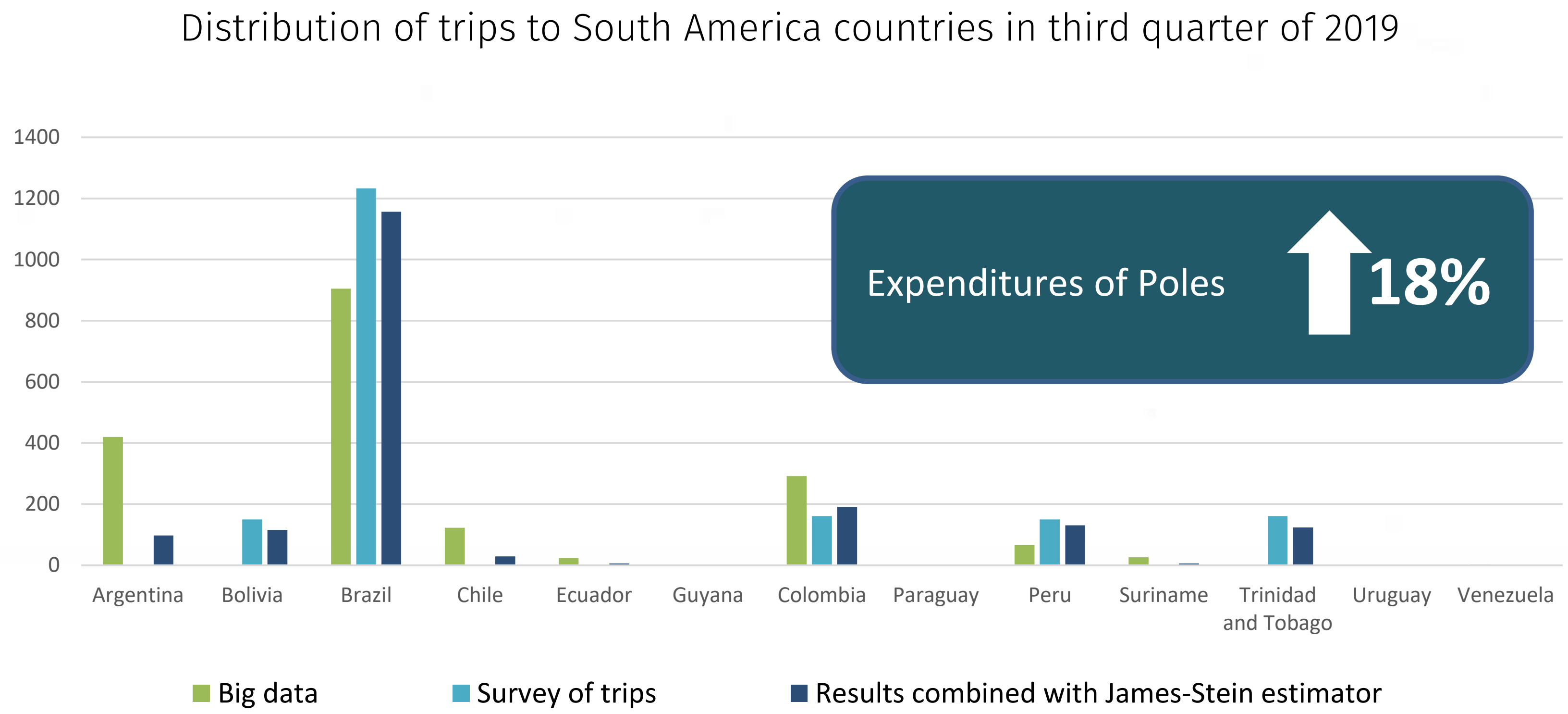
External data in the demand side of tourism
Flight data – web scraping



Variables	Description	Value for Yes	Value for No
Mandatory			
Site name	Portal name	1	0
URL	URL of portal	1	0
OfferId	Unique identifier of offer in portal	1	0
Price	Price for offer	1	0
Airline	Airline name	1	0
Airport	Airport name / starting location	1	0
Destination	Destination for flight	1	0
Optional			
Class type	Class type for flight (economy / buissness)	1	0
Time of travel	Total time of flight	1	0
time of stops	Total time of stops between flights	1	0
Flight number	Flight number	1	0
Plane type	Type of airplane	1	0
Departure date	Time of departure	1	0
Arrival date	Time of arrival to destination	1	0
Stops	Number of stops	1	0

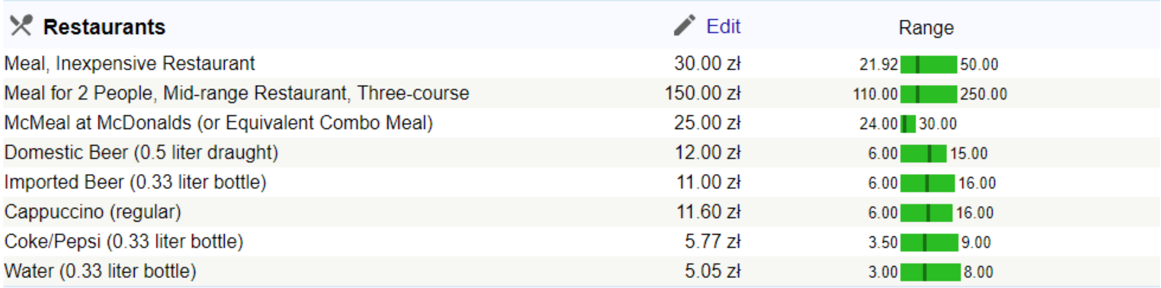
Funded by the European Union

External data in the demand side of tourism
Flight data



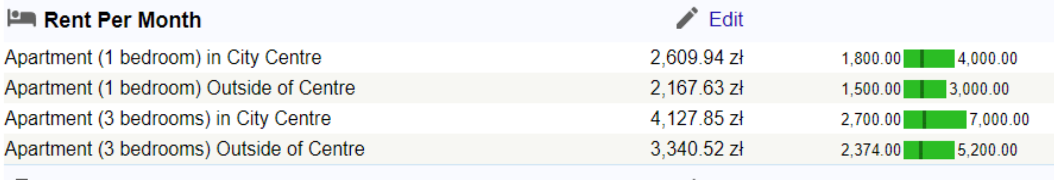
Expenditures of Poles ⬆ 18%

External data in the demand side of tourism
Costs of living/cost of trip

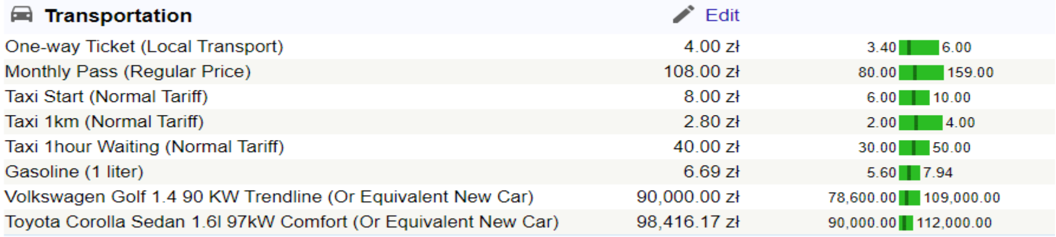


Restaurants	Edit	Range
Meal, Inexpensive Restaurant	30.00 zł	21.92 🟩 50.00
Meal for 2 People, Mid-range Restaurant, Three-course	150.00 zł	110.00 🟩 250.00
McMeal at McDonalds (or Equivalent Combo Meal)	25.00 zł	24.00 🟩 30.00
Domestic Beer (0.5 liter draught)	12.00 zł	6.00 🟩 15.00
Imported Beer (0.33 liter bottle)	11.00 zł	6.00 🟩 16.00
Cappuccino (regular)	11.60 zł	6.00 🟩 16.00
Coke/Pepsi (0.33 liter bottle)	5.77 zł	3.50 🟩 9.00
Water (0.33 liter bottle)	5.05 zł	3.00 🟩 8.00

NUMBEO

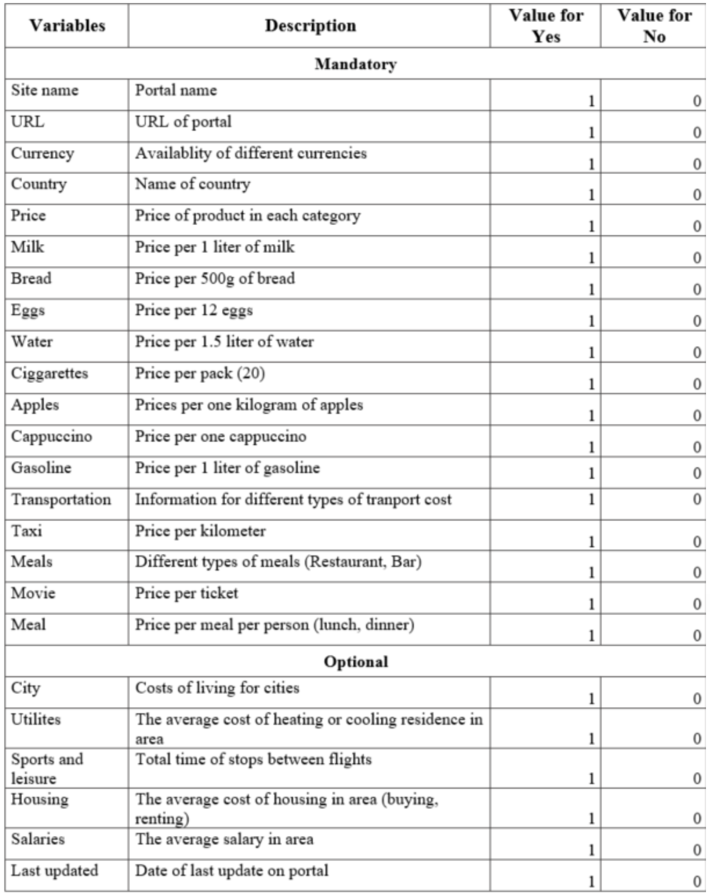


Rent Per Month	Edit	
Apartment (1 bedroom) in City Centre	2,609.94 zł	1,800.00 🟩 4,000.00
Apartment (1 bedroom) Outside of Centre	2,167.63 zł	1,500.00 🟩 3,000.00
Apartment (3 bedrooms) in City Centre	4,127.85 zł	2,700.00 🟩 7,000.00
Apartment (3 bedrooms) Outside of Centre	3,340.52 zł	2,374.00 🟩 5,200.00



Transportation	Edit	
One-way Ticket (Local Transport)	4.00 zł	3.40 🟩 6.00
Monthly Pass (Regular Price)	108.00 zł	80.00 🟩 159.00
Taxi Start (Normal Tariff)	8.00 zł	6.00 🟩 10.00
Taxi 1km (Normal Tariff)	2.80 zł	2.00 🟩 4.00
Taxi 1hour Waiting (Normal Tariff)	40.00 zł	30.00 🟩 50.00
Gasoline (1 liter)	6.69 zł	5.60 🟩 7.94
Volkswagen Golf 1.4 90 KW Trendline (Or Equivalent New Car)	90,000.00 zł	78,600.00 🟩 109,000.00
Toyota Corolla Sedan 1.6l 97kW Comfort (Or Equivalent New Car)	98,416.17 zł	90,000.00 🟩 112,000.00

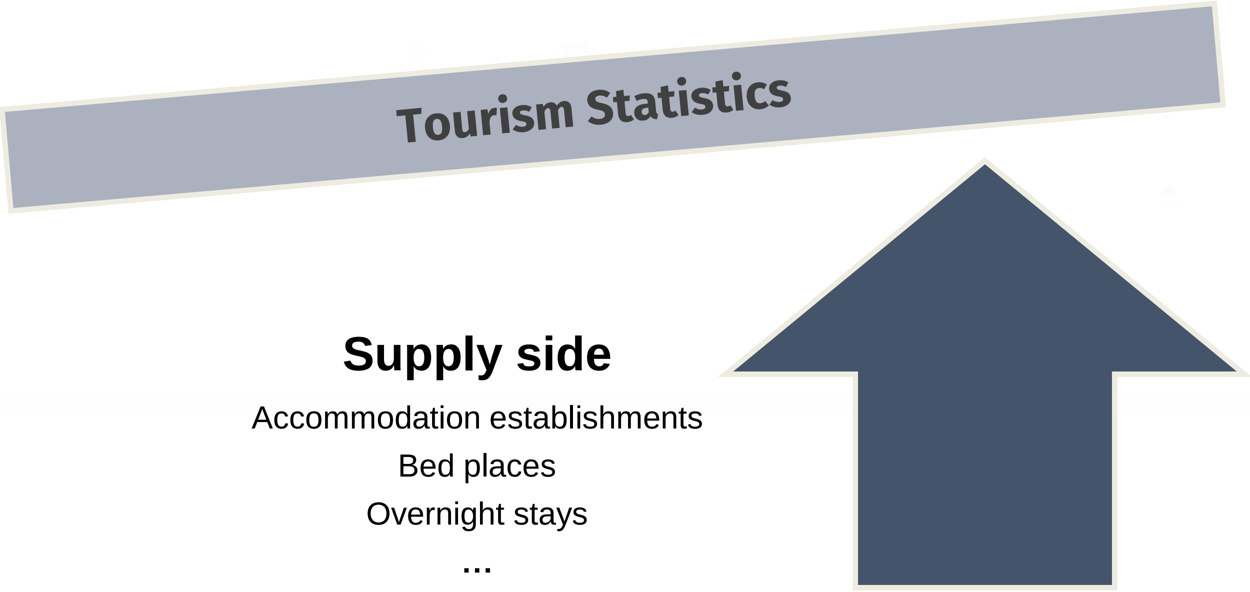
External data in the demand side of tourism
Costs of living/cost of trip



Variables	Description	Value for Yes	Value for No
Mandatory			
Site name	Portal name	1	0
URL	URL of portal	1	0
Currency	Availablity of different currencies	1	0
Country	Name of country	1	0
Price	Price of product in each category	1	0
Milk	Price per 1 liter of milk	1	0
Bread	Price per 500g of bread	1	0
Eggs	Price per 12 eggs	1	0
Water	Price per 1.5 liter of water	1	0
Ciggarettes	Price per pack (20)	1	0
Apples	Prices per one kilogram of apples	1	0
Cappuccino	Price per one cappuccino	1	0
Gasoline	Price per 1 liter of gasoline	1	0
Transportation	Information for different types of tranport cost	1	0
Taxi	Price per kilometer	1	0
Meals	Different types of meals (Restaurant, Bar)	1	0
Movie	Price per ticket	1	0
Meal	Price per meal per person (lunch, dinner)	1	0
Optional			
City	Costs of living for cities	1	0
Utilites	The average cost of heating or cooling residence in area	1	0
Sports and leisure	Total time of stops between flights	1	0
Housing	The average cost of housing in area (buying, renting)	1	0
Salaries	The average salary in area	1	0
Last updated	Date of last update on portal	1	0

Funded by the European Union

External data in the demand side of tourism



Tourism Statistics

Supply side

Accommodation establishments

Bed places

Overnight stays

...

External data in the supply side of tourism
Survey of tourist accommodation establishments

In European countries tourist accommodation establishments are surveyed regardless of the type of facility, owner and location, as well as establishments for other purposes (not related to tourism) that are temporarily used by tourists (e.g. student dormitories, sports and recreation centres).

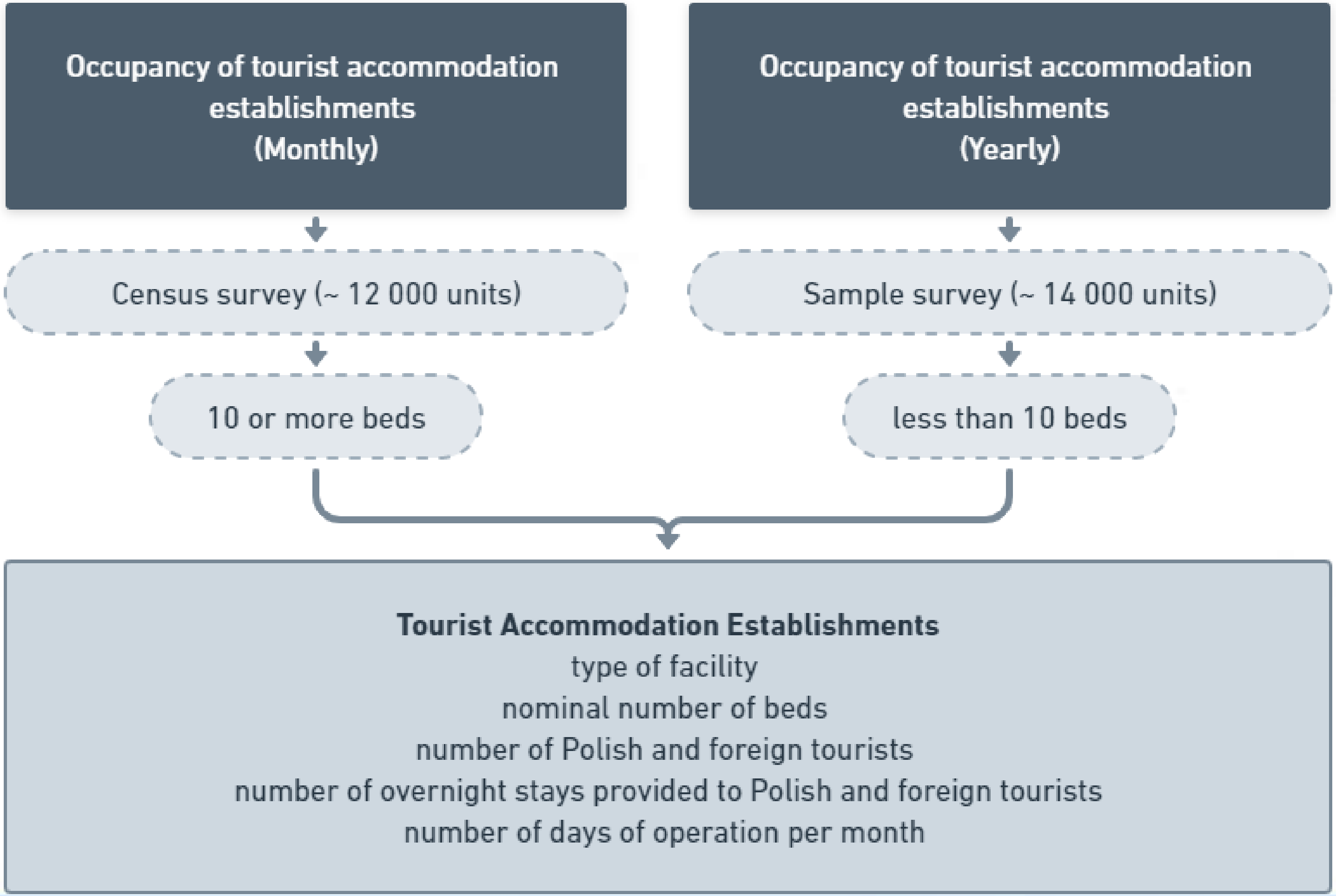
Tourist accommodation establishments classified into the following **NACE** activity groups:

55.1 – Hotels and similar accommodation

55.2 – Holiday and other short-stay accommodation

55.3 – Camping grounds, recreational vehicle parks and trailer parks

External data in the supply side of tourism
Survey of tourist accommodation establishments – case of Poland



Occupancy of tourist accommodation establishments (Monthly)

Census survey (~ 12 000 units)

10 or more beds

Occupancy of tourist accommodation establishments (Yearly)

Sample survey (~ 14 000 units)

less than 10 beds

Tourist Accommodation Establishments

type of facility

nominal number of beds

number of Polish and foreign tourists

number of overnight stays provided to Polish and foreign tourists

number of days of operation per month

Funded by the European Union

External data in the supply side of tourism
Survey of tourist accommodation establishments – case of Poland

Survey frame of accommodation establishments

Register of Hotels and similar accommodation

(Ministry of Sport and Tourism)

Booking platforms (Web scraping)

+ all types of facilities

+ frequently updated

- linking data with a statistical survey

Funded by the European Union

Nocowanie.pl

krok po kroku

Expedia

airbnb

Booking.com

eholiday.pl

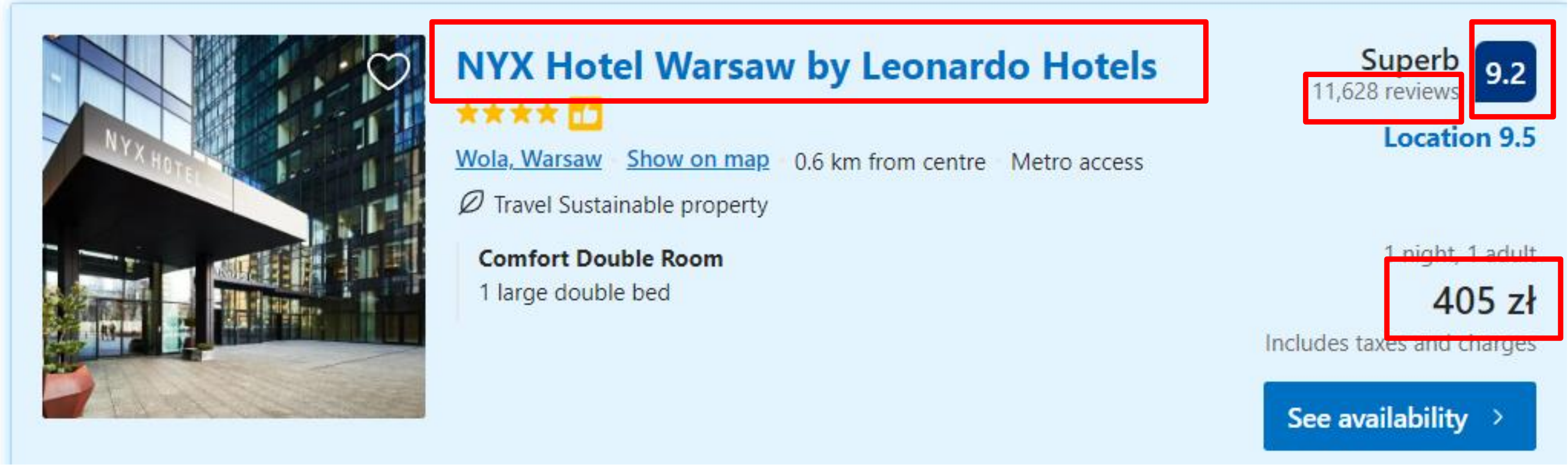
DOBRE NOCLEGI

Hotels.com™

OdkryjPolske.pl

Polski Portal Rezerwacyjny

Tripadvisor



NYX Hotel Warsaw by Leonardo Hotels

Wola, Warsaw Show on map 0.6 km from centre Metro access

Travel Sustainable property

Comfort Double Room

1 large double bed

Superb 9.2

11,628 reviews

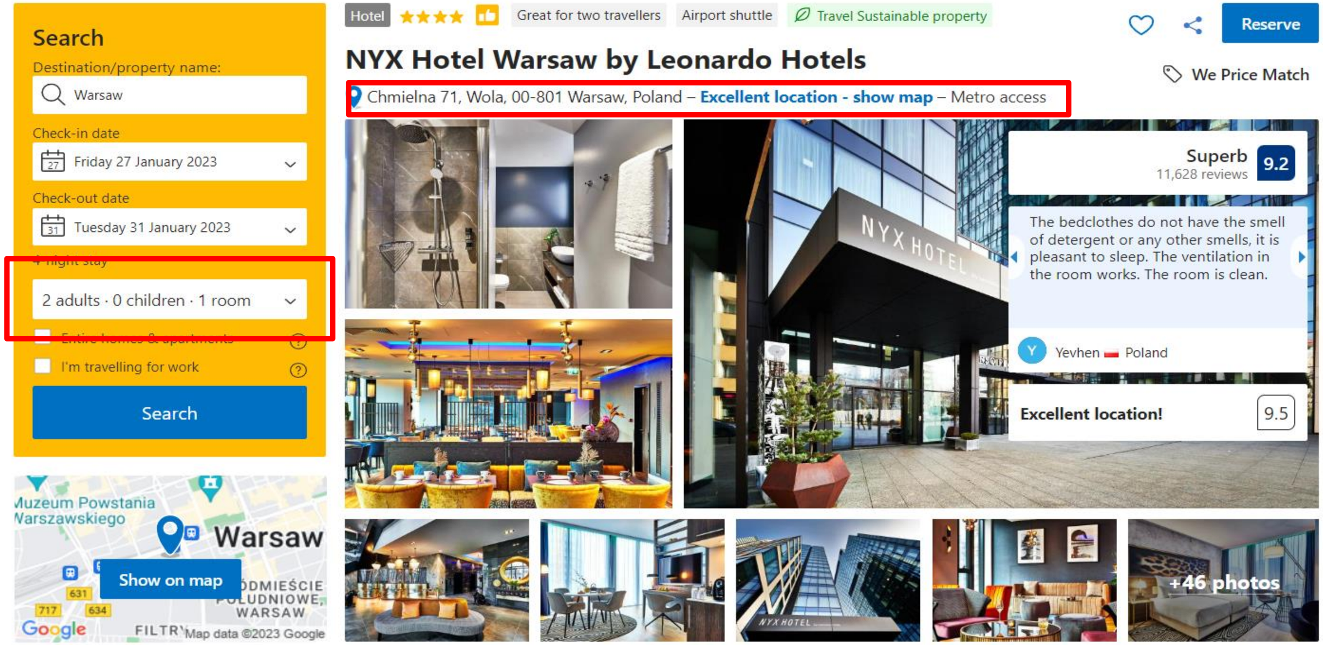
Location 9.5

1 night, 1 adult

405 zł

Includes taxes and charges

See availability ›



Search

Destination/property name:

Warsaw

Check-in date

Friday 27 January 2023

Check-out date

Tuesday 31 January 2023

4-night stay

2 adults · 0 children · 1 room

□ Entire homes & apartments

□ I'm travelling for work

Search

Muzeum Powstania Warszawskiego
Warsaw
Show on map
ŚRÓDMIEŚCIE POŁUDNIOWE, WARSAW
Google
Map data ©2023 Google

Hotel Great for two travellers Airport shuttle Travel Sustainable property

Reserve

NYX Hotel Warsaw by Leonardo Hotels

We Price Match

Chmielna 71, Wola, 00-801 Warsaw, Poland – Excellent location - show map – Metro access

Superb 9.2

11,628 reviews

The bedclothes do not have the smell of detergent or any other smells, it is pleasant to sleep. The ventilation in the room works. The room is clean.

Yevhen Poland

Excellent location! 9.5

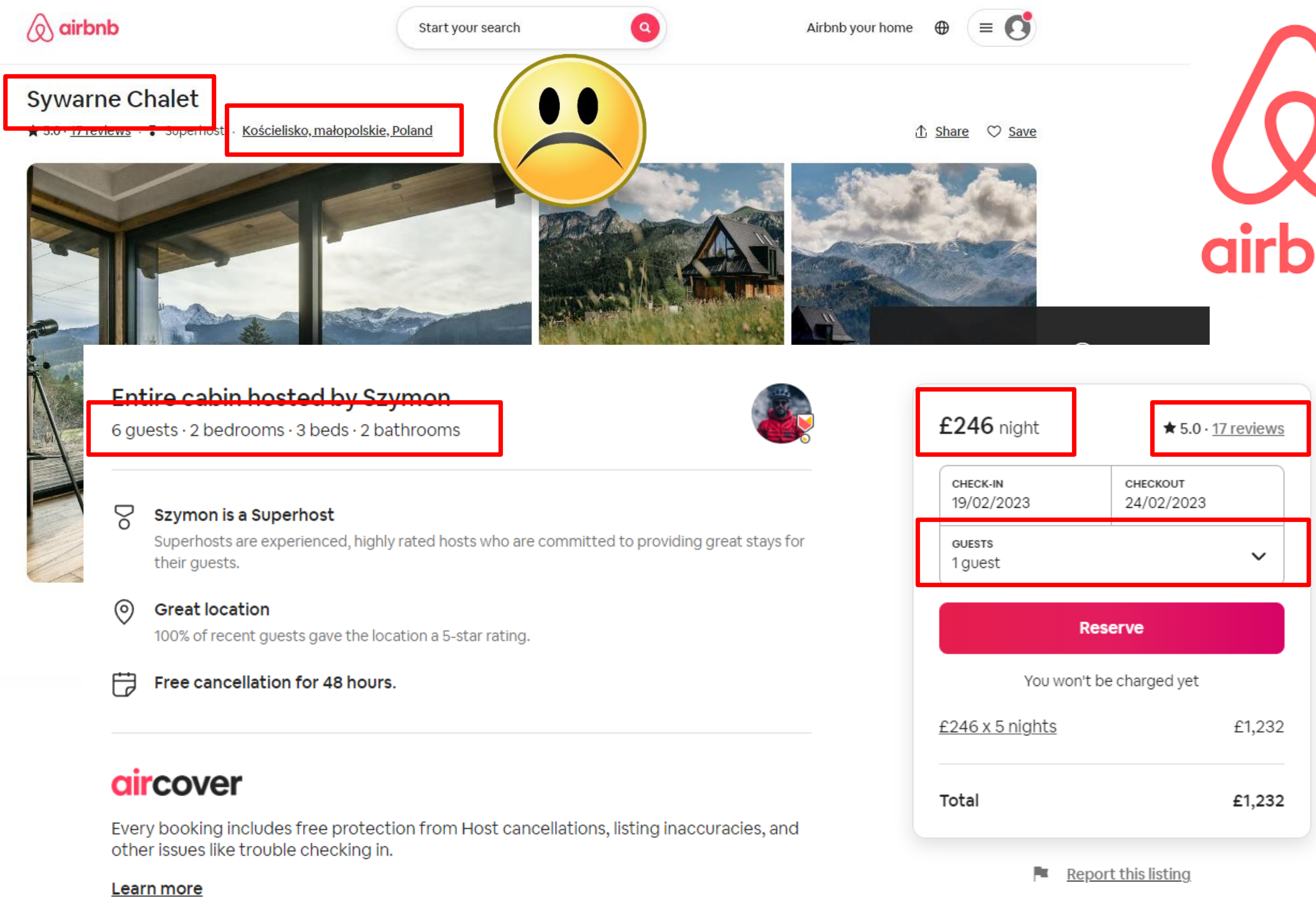
+46 photos

- Object name
- Type of facility,
- Exact address.
- Number of guests
- Price

- Number of reviews
- Facility rating
- Number of beds/rooms

Web Intelligence Network

Funded by the European Union



airbnb

Start your search

Airbnb your home

Sywarne Chalet

★ 5.0 · 17 reviews · Superhost · Kościelisko, małopolskie, Poland

Share Save

Entire cabin hosted by Szymon

6 guests · 2 bedrooms · 3 beds · 2 bathrooms

Szymon is a Superhost

Superhosts are experienced, highly rated hosts who are committed to providing great stays for their guests.

Great location

100% of recent guests gave the location a 5-star rating.

Free cancellation for 48 hours.

aircover

Every booking includes free protection from Host cancellations, listing inaccuracies, and other issues like trouble checking in.

Learn more

£246 night

★ 5.0 · 17 reviews

CHECK-IN 19/02/2023

CHECKOUT 24/02/2023

GUESTS 1 guest

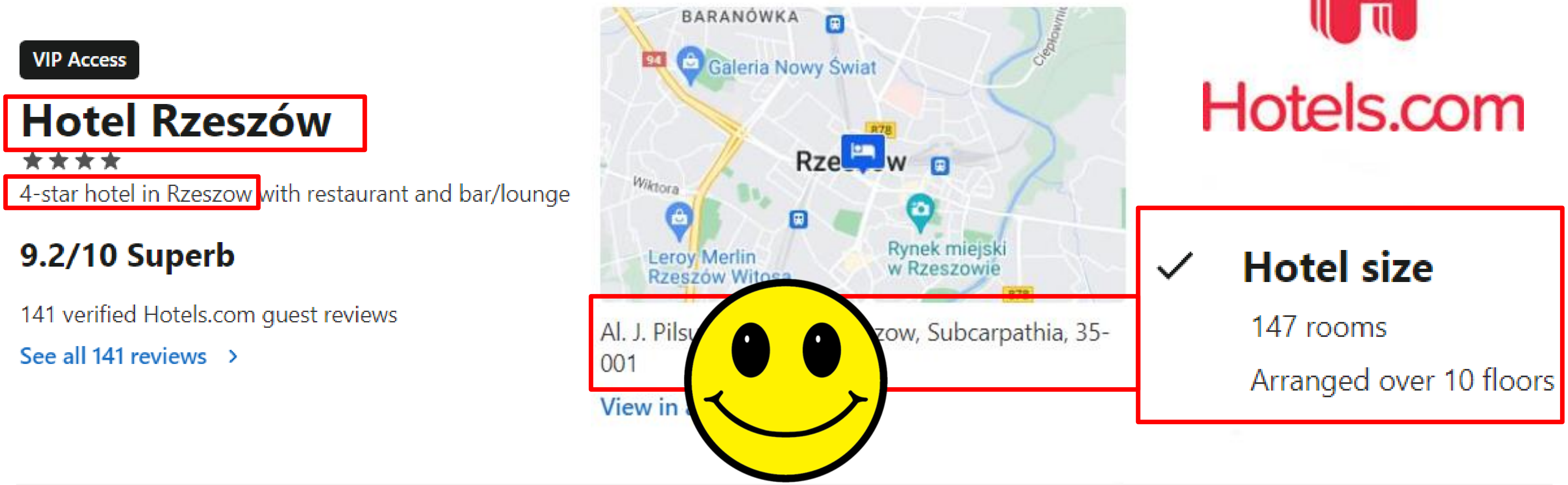
Reserve

You won't be charged yet

£246 x 5 nights	£1,232
Total	£1,232

Report this listing

Hotels.com


VIP Access

Hotel Rzeszów

★★★★

4-star hotel in Rzeszow with restaurant and bar/lounge

9.2/10 Superb

141 verified Hotels.com guest reviews

See all 141 reviews ›

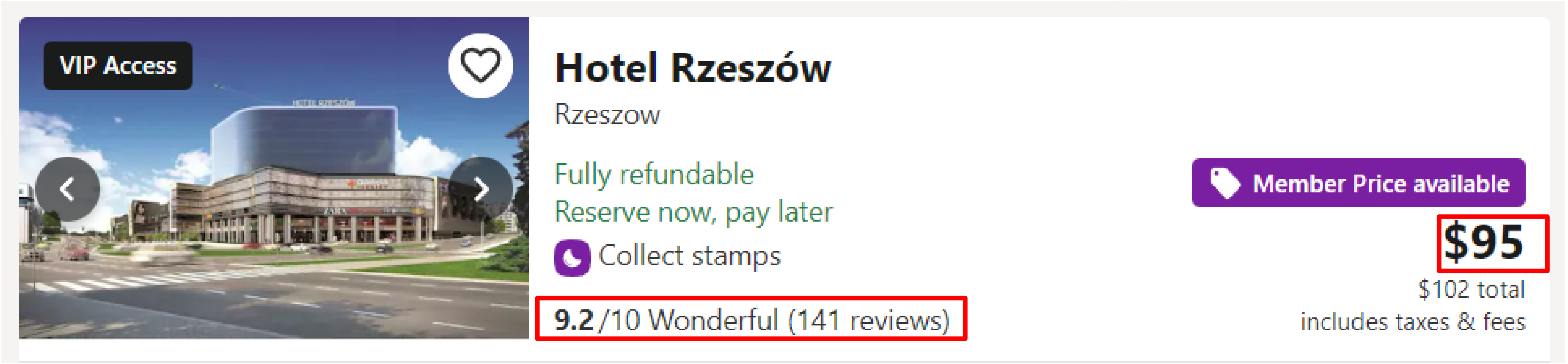
Al. J. Pilsu zow, Subcarpathia, 35-001

View in

✓ Hotel size

147 rooms

Arranged over 10 floors


VIP Access

Hotel Rzeszów

Rzeszow

Fully refundable

Reserve now, pay later

Collect stamps

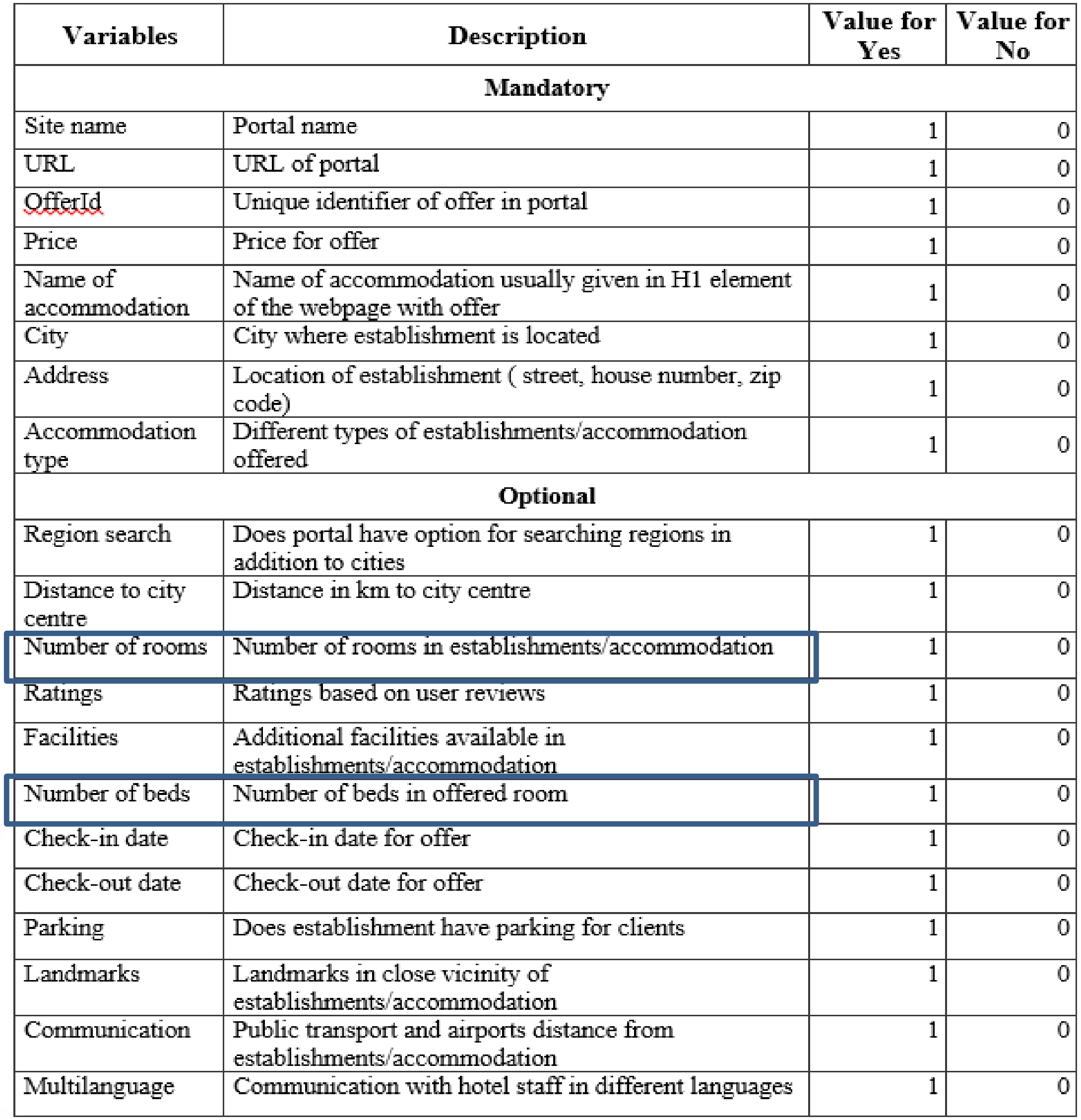
9.2/10 Wonderful (141 reviews)

Member Price available

$95

$102 total

includes taxes & fees



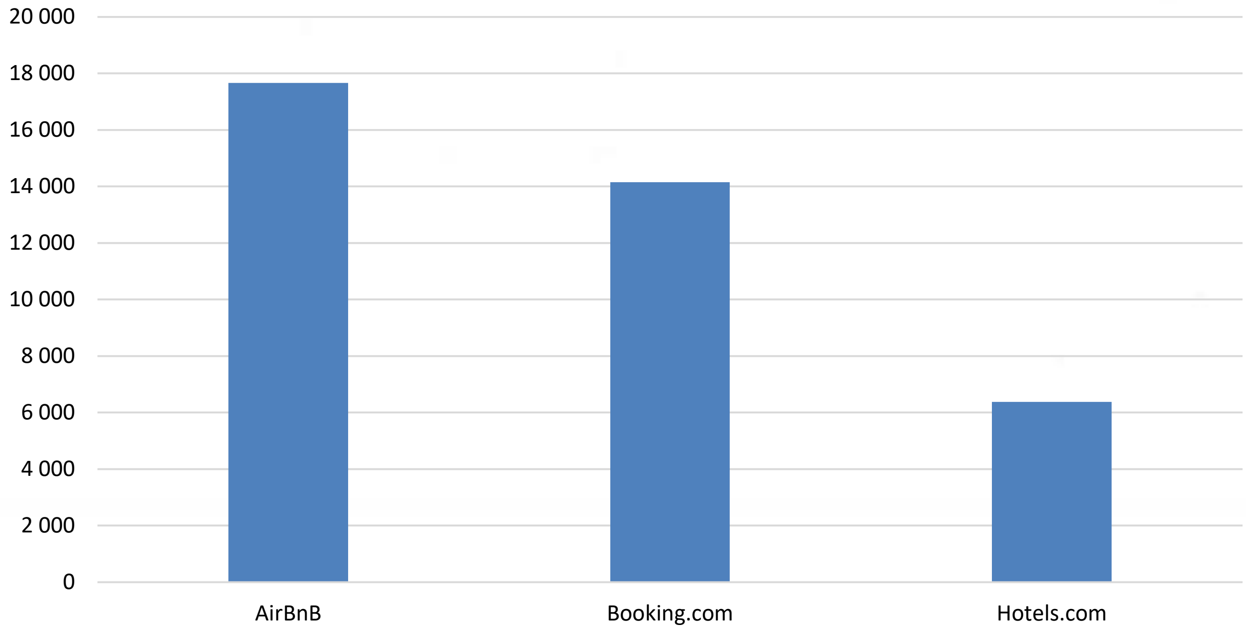
Variables	Description	Value for Yes	Value for No
Mandatory			
Site name	Portal name	1	0
URL	URL of portal	1	0
OfferId	Unique identifier of offer in portal	1	0
Price	Price for offer	1	0
Name of accommodation	Name of accommodation usually given in H1 element of the webpage with offer	1	0
City	City where establishment is located	1	0
Address	Location of establishment (street, house number, zip code)	1	0
Accommodation type	Different types of establishments/accommodation offered	1	0
Optional			
Region search	Does portal have option for searching regions in addition to cities	1	0
Distance to city centre	Distance in km to city centre	1	0
Number of rooms	Number of rooms in establishments/accommodation	1	0
Ratings	Ratings based on user reviews	1	0
Facilities	Additional facilities available in establishments/accommodation	1	0
Number of beds	Number of beds in offered room	1	0
Check-in date	Check-in date for offer	1	0
Check-out date	Check-out date for offer	1	0
Parking	Does establishment have parking for clients	1	0
Landmarks	Landmarks in close vicinity of establishments/accommodation	1	0
Communication	Public transport and airports distance from establishments/accommodation	1	0
Multilanguage	Communication with hotel staff in different languages	1	0

Hotels.com™

airbnb

External data in the supply side of tourism
Analysis of booking platforms

Number of establishments offering accommodation in Poland in 2022

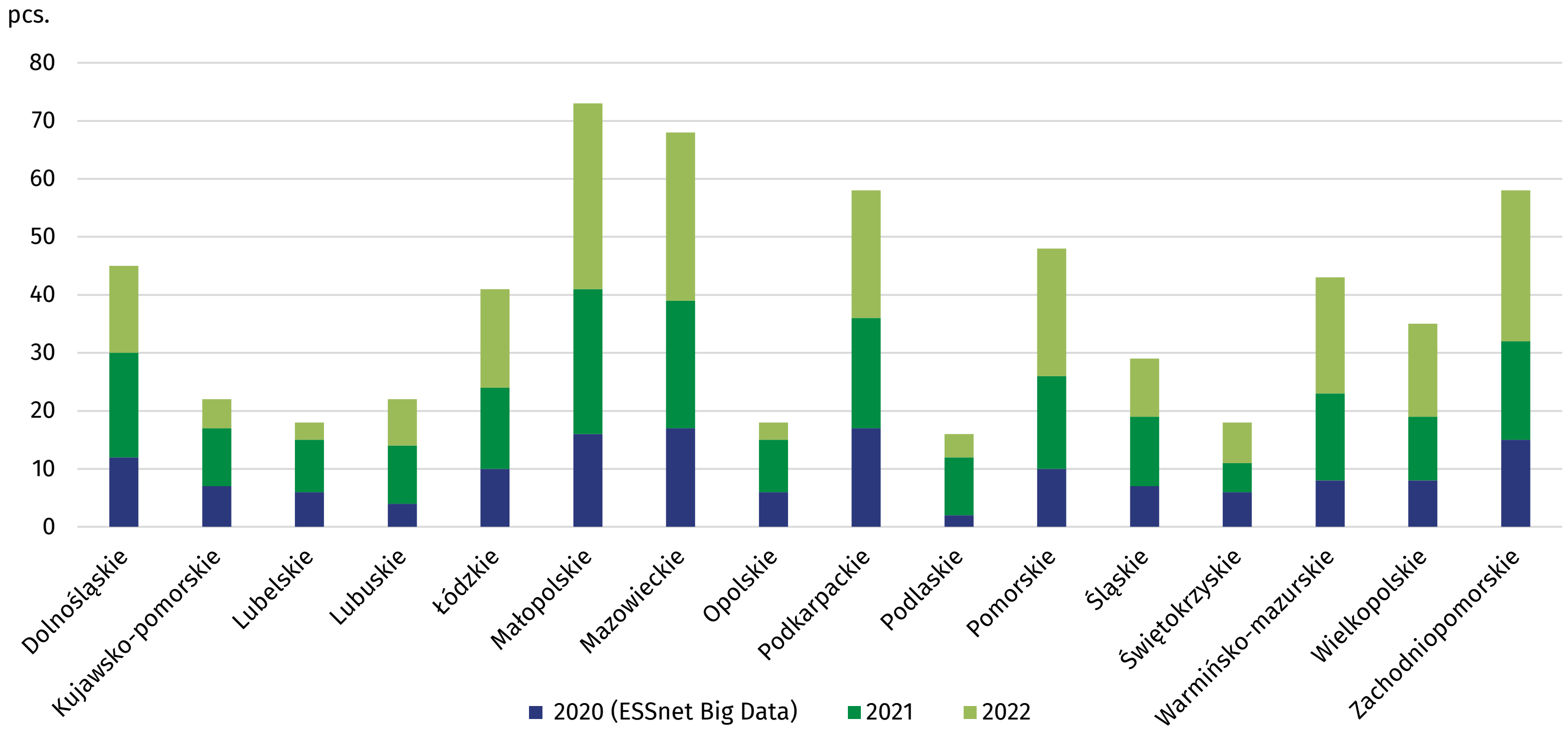


Web Intelligence Network

Funded by the European Union

External data in the supply side of tourism

Approximately 15,000 unique accommodation establishments per month are collected from Hotels.com and Booking.com. In 2022, 239 new accommodation establishments with 10 or more beds places were identified through web scraping.



Web Intelligence Network

Funded by the European Union

Method of Processing and Analysing Web Scraped Tourism Data

From web scraping to data combining

Piotr Szlachta (GUS)

23 February 2023

Trusted Smart Statistics – Web Intelligence Network

Grant Agreement: 101035829

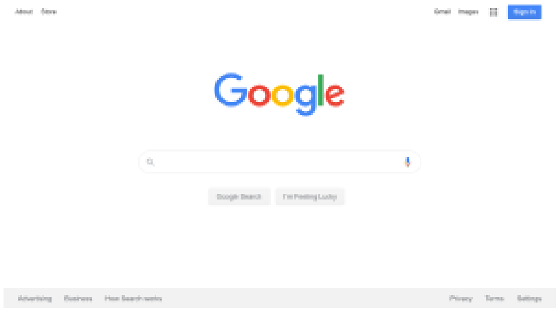
Web Intelligence Network

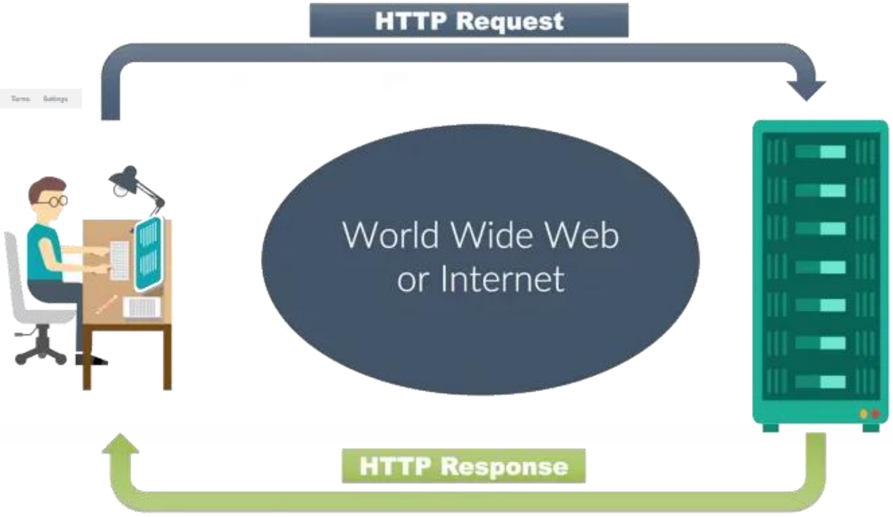
Funded by the European Union

Outline

- Basics of web portals structure
- Web scraping
- Data cleaning and data analysis
- Combining Data

Basics





HTTP Version

Status Code

Headers

Body

```
HTTP/1.1 200 OK
Content-Type: application/json; charset=utf-8
Server: Kestrel
X-Powered-By: ASP.NET
Date: Sun, 11 Feb 2018 18:34:00 GMT
Content-Length: 69

{
  "name":"Product",
  "category":"Appliances",
  "subcategory":"Microwaves“
}
```

Web Intelligence Network

Funded by the European Union

Basics

HTML - page content

CSS - defined visual appearance (e.g. font styles, paragraph styles, etc.)

Images – graphics

JS (JavaScript) - adding interactivity to a website

HTML
HTML + CSS
HTML + CSS
+ JAVASCRIPT

Funded by the European Union

Basics

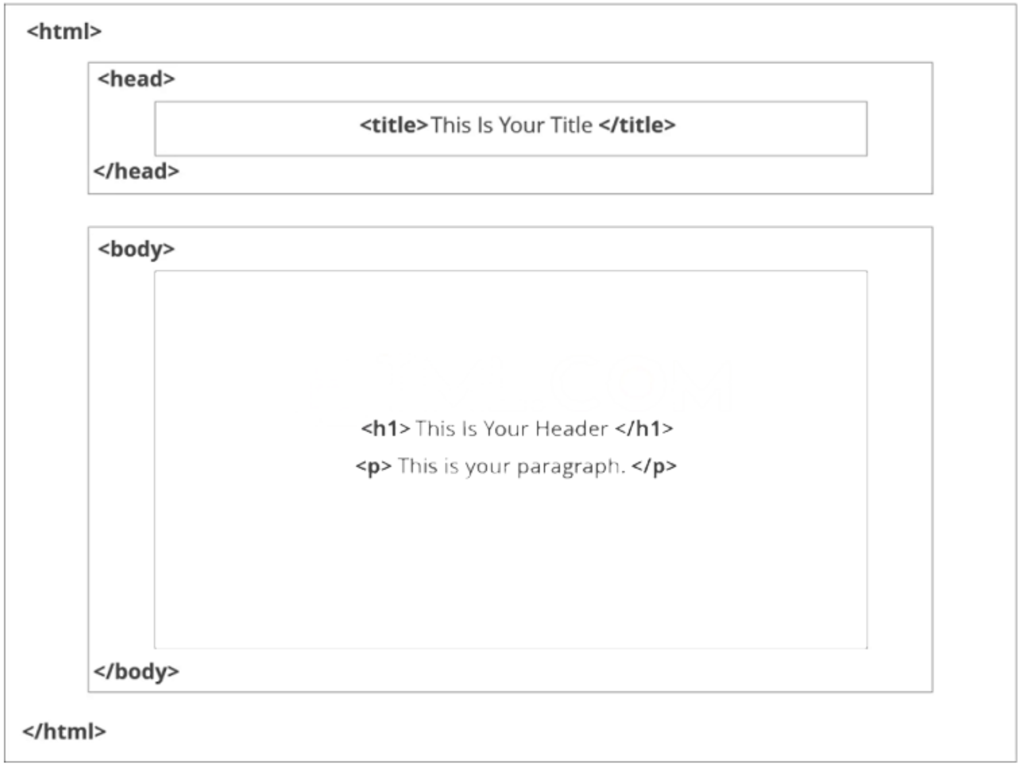
The basic page design includes:

<! DOCTYPE html> - defines the language in which the page is written (HTML 5)

<html> opening tag of the HTML code

<head> page metadata - information mainly for search engines and other computer programs

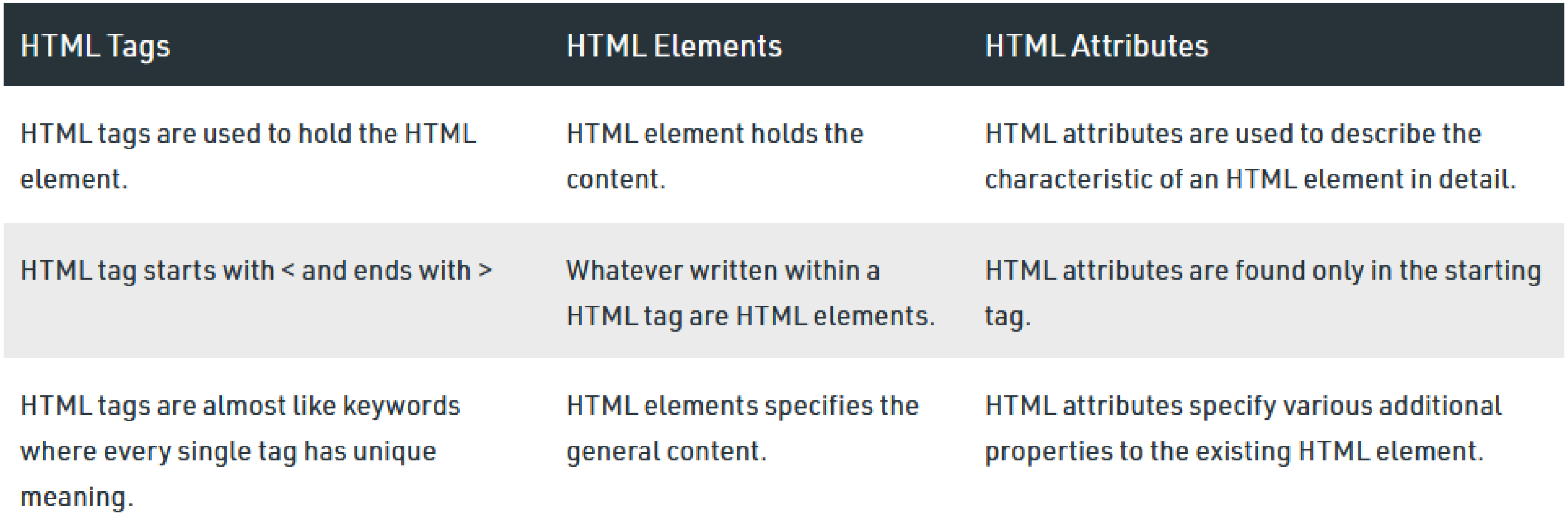
<body> content of the page



```
<html>
    <head>
        <title>This Is Your Title </title>
    </head>
    <body>
        <h1> This Is Your Header </h1>
        <p> This is your paragraph. </p>
    </body>
</html>
```

Basics

- Tags and attributes are the basis of HTML



HTML Tags	HTML Elements	HTML Attributes
HTML tags are used to hold the HTML element.	HTML element holds the content.	HTML attributes are used to describe the characteristic of an HTML element in detail.
HTML tag starts with < and ends with >	Whatever written within a HTML tag are HTML elements.	HTML attributes are found only in the starting tag.
HTML tags are almost like keywords where every single tag has unique meaning.	HTML elements specifies the general content.	HTML attributes specify various additional properties to the existing HTML element.

Source: https://www.geeksforgeeks.org/tags-vs-elements-vs-attributes-in-html/

Basics

```
<!DOCTYPE HTML
<html>
  <head>
    <meta name="TITLE
     <meta name="KEYWORDS
      <meta name="DESCRIPTION
       <link rel="stylesheet
        <script language="javas
        </head>
       <body bgcolor="#ffff
                   width="
```

Identifiers (id) and classes (class)

- optional and not all elements will have them
- an ***identifier*** can only be used once per page
- each element may have only one ***identifier***
- an element can have multiple ***classes***
- one ***class*** can be used for any number of elements on a page

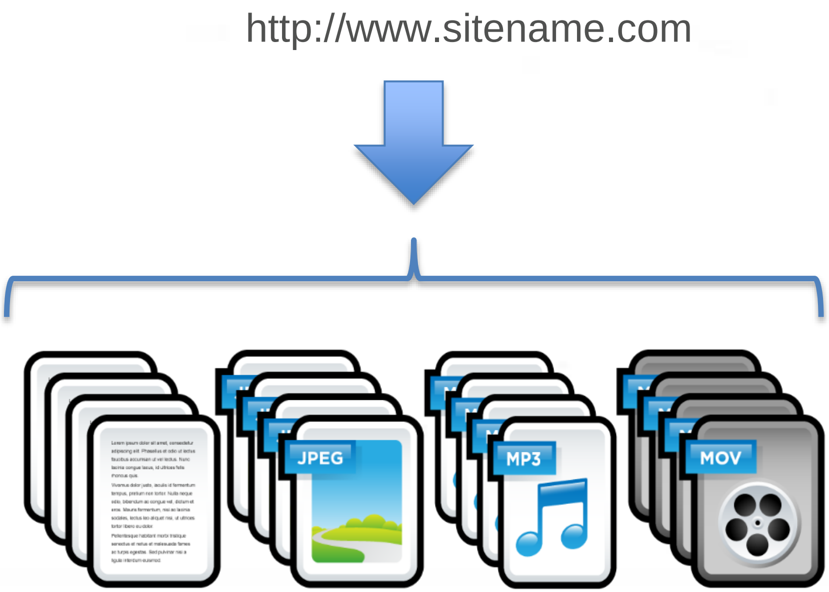
Web Intelligence Network

Funded by the European Union

Data extraction from web

Web crawling

- Locating information on the World Wide Web (WWW)
- Indexing all words in a document
- Adding them to the database
- Tracking all hyperlinks and indexes and adding this information to the database as well.


http://www.sitename.com

Web scraping

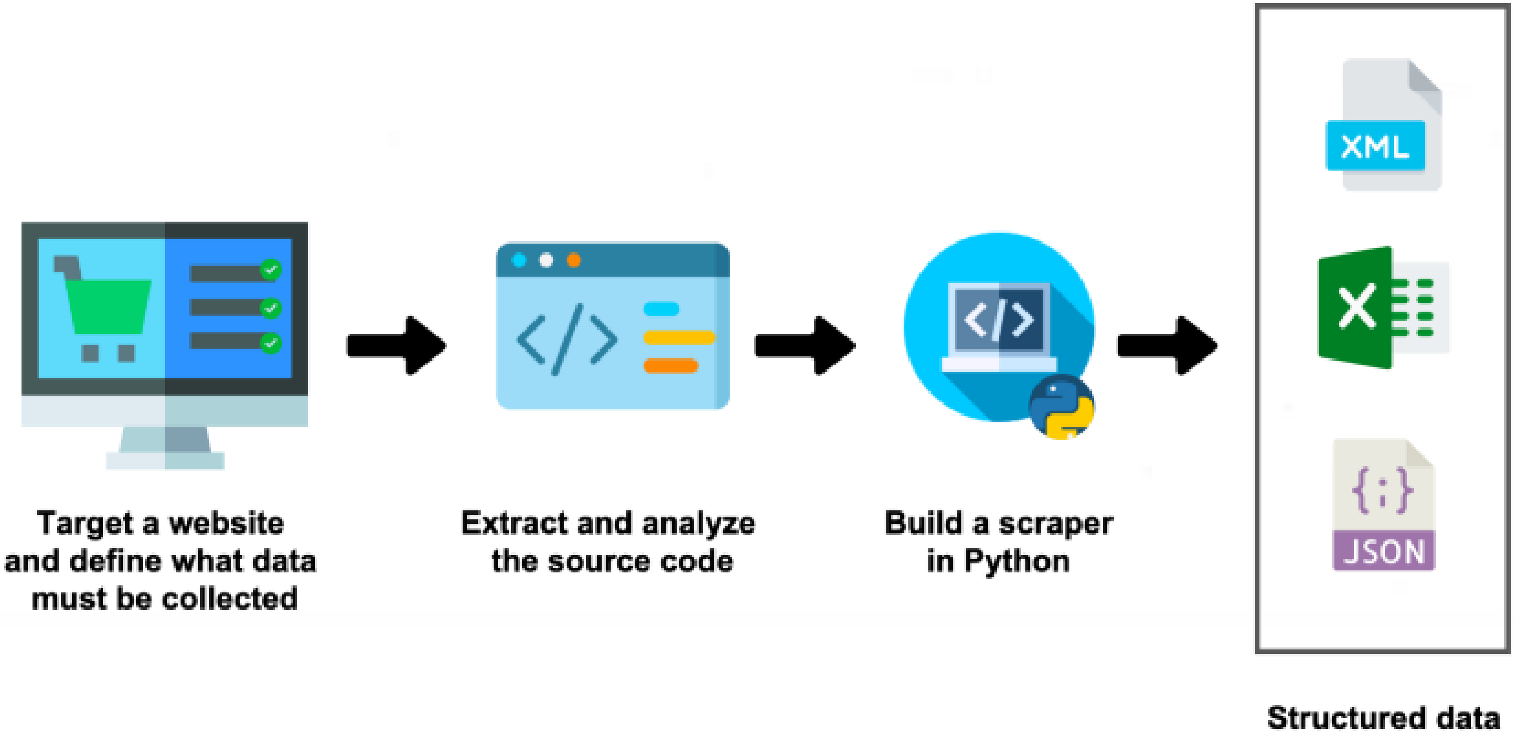
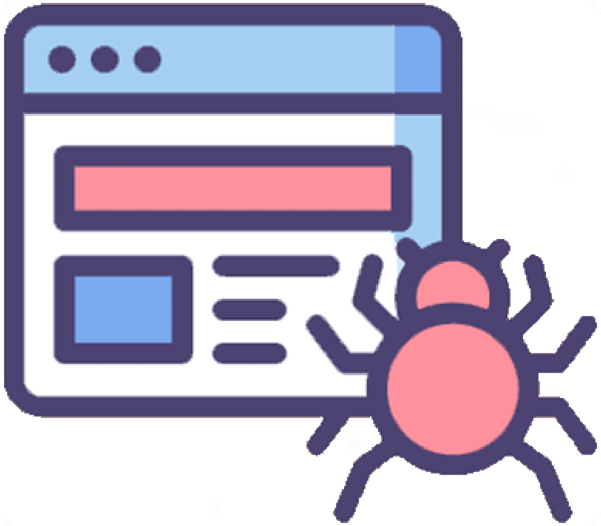

Target a website and define what data must be collected
Extract and analyze the source code
Build a scraper in Python
XML
JSON
Structured data

Image by the author: scraping workflow

Good practices



Characteristics of good scrapers:

- „introduces itself"
- does not affect the daily operations and functioning of the portals
- performed outside peak hours
- subsequent queries to the server using time intervals

Libraries

Request / Beautiful Soup

- Best for pages without JavaScript
- Easy to use
- Retrieval of HTML code

Selenium

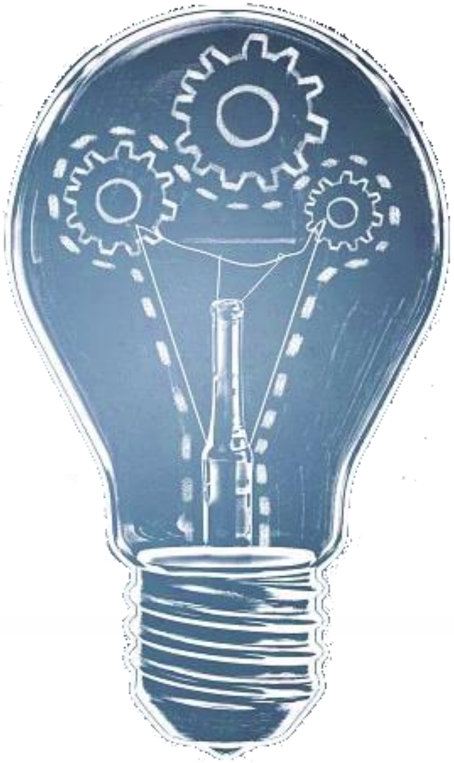
- Emulating user behaviour
- Sending forms with data
- Executing Javascript scripts

Web Intelligence Network

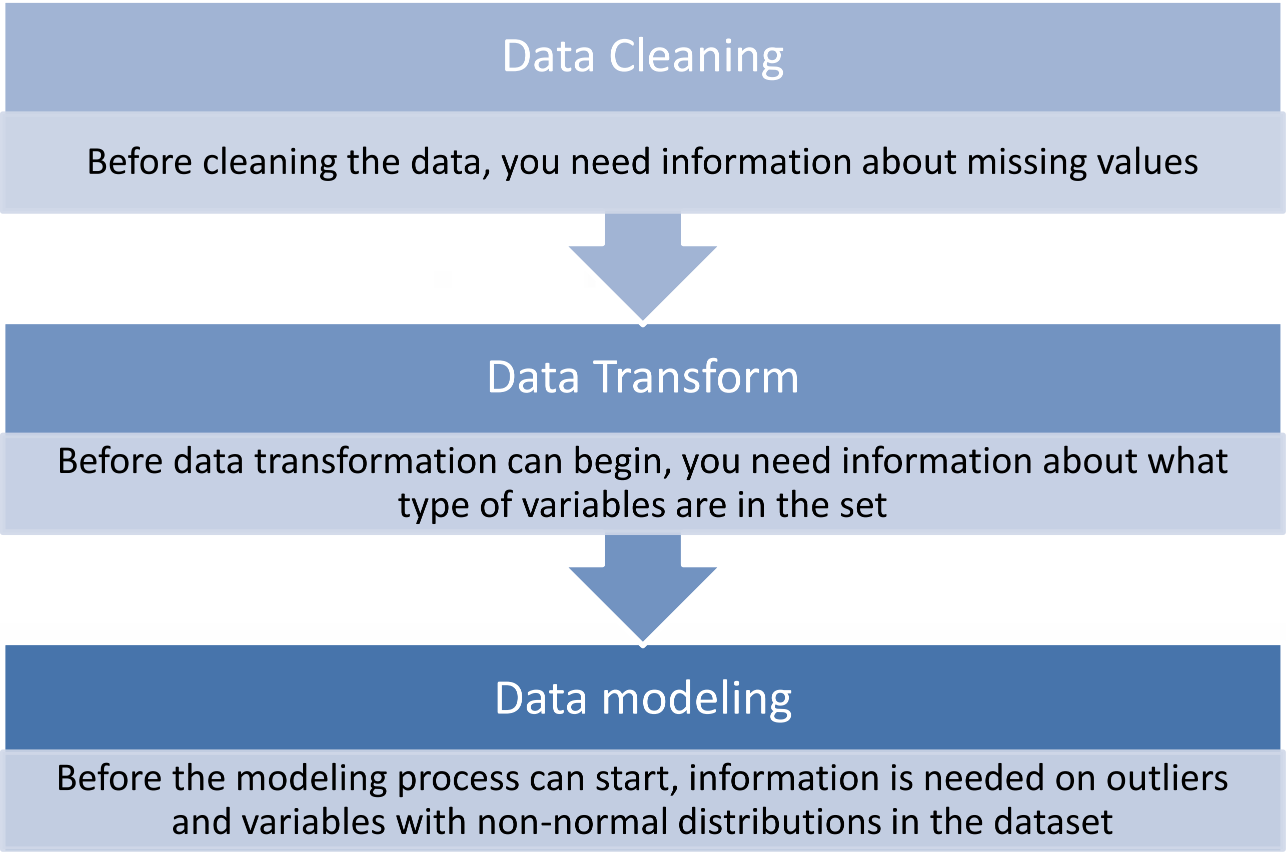
Funded by the European Union

Tourism portals

- Fewer and fewer portals do not contain JavaScript
- Increasing number of elements on portals is generated dynamically
- Fetching all variables requires more and more interaction with portal
- Portals change their structure several times a year



From data cleaning to data combining



Data Cleaning

Before cleaning the data, you need information about missing values

Data Transform

Before data transformation can begin, you need information about what type of variables are in the set

Data modeling

Before the modeling process can start, information is needed on outliers and variables with non-normal distributions in the dataset

8 steps

Funded by the European Union

Step 1. Look at the data

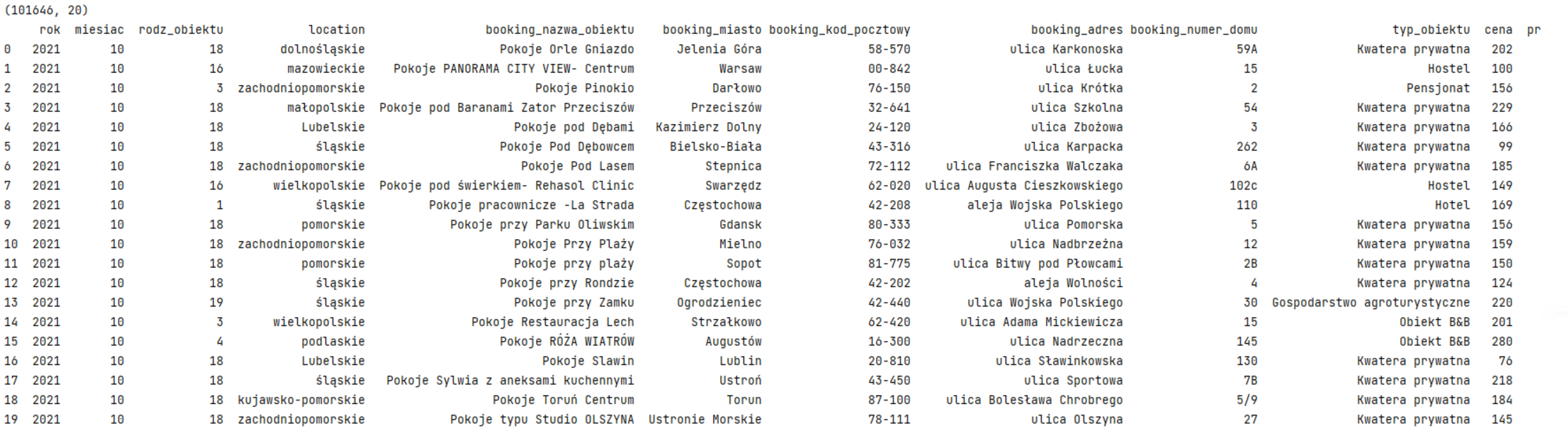
The first and basic step is to know the size of the set to be analysed. Both the number of observations and the number of variables that describe them should be checked.

WHY?

- Facilitates decision-making regarding the tools and hardware.
- Estimate the time consumption of the process.
- Understand the structure of the collection.
- Preliminary relevance of individual variables.

Step 1. Look at the data

- df.shape
- df.head(20)



```
(101646, 20)
     rok  miesiac  rodz_obiektu            location                 booking_nazwa_obiektu    booking_miasto  booking_kod_pocztowy                 booking_adres  booking_numer_domu                   typ_obiektu  cena  pr
0   2021       10            18        dolnośląskie                   Pokoje Orle Gniazdo      Jelenia Góra                58-570              ulica Karkonoska                 59A              Kwatera prywatna   202
1   2021       10            16         mazowieckie    Pokoje PANORAMA CITY VIEW- Centrum            Warsaw                00-842                   ulica Łucka                  15                        Hostel   100
2   2021       10             3  zachodniopomorskie                        Pokoje Pinokio           Darłowo                76-150                  ulica Krótka                   2                     Pensjonat   156
3   2021       10            18         małopolskie  Pokoje pod Baranami Zator Przeciszów        Przeciszów                32-641                 ulica Szkolna                  54              Kwatera prywatna   229
4   2021       10            18           Lubelskie                     Pokoje pod Dębami   Kazimierz Dolny                24-120                 ulica Zbożowa                   3              Kwatera prywatna   166
5   2021       10            18             śląskie                   Pokoje Pod Dębowcem     Bielsko-Biała                43-316                ulica Karpacka                 262              Kwatera prywatna    99
6   2021       10            18  zachodniopomorskie                      Pokoje Pod Lasem          Stepnica                72-112     ulica Franciszka Walczaka                  6A              Kwatera prywatna   185
7   2021       10            16       wielkopolskie  Pokoje pod świerkiem- Rehasol Clinic          Swarzędz                62-020  ulica Augusta Cieszkowskiego                102c                        Hostel   149
8   2021       10             1             śląskie         Pokoje pracownicze -La Strada       Częstochowa                42-208        aleja Wojska Polskiego                 110                         Hotel   169
9   2021       10            18           pomorskie            Pokoje przy Parku Oliwskim            Gdansk                80-333                ulica Pomorska                   5              Kwatera prywatna   156
10  2021       10            18  zachodniopomorskie                     Pokoje Przy Plaży            Mielno                76-032              ulica Nadbrzeżna                  12              Kwatera prywatna   159
11  2021       10            18           pomorskie                     Pokoje przy plaży             Sopot                81-775      ulica Bitwy pod Płowcami                  2B              Kwatera prywatna   150
12  2021       10            18             śląskie                   Pokoje przy Rondzie       Częstochowa                42-202                aleja Wolności                   4              Kwatera prywatna   124
13  2021       10            19             śląskie                     Pokoje przy Zamku      Ogrodzieniec                42-440        ulica Wojska Polskiego                  30  Gospodarstwo agroturystyczne   220
14  2021       10             3       wielkopolskie               Pokoje Restauracja Lech        Strzałkowo                62-420       ulica Adama Mickiewicza                  15                    Obiekt B&B   201
15  2021       10             4           podlaskie                   Pokoje RÓŻA WIATRÓW          Augustów                16-300              ulica Nadrzeczna                 145                    Obiekt B&B   280
16  2021       10            18           Lubelskie                         Pokoje Slawin            Lublin                20-810            ulica Sławinkowska                 130              Kwatera prywatna    76
17  2021       10            18             śląskie   Pokoje Sylwia z aneksami kuchennymi            Ustroń                43-450                ulica Sportowa                  7B              Kwatera prywatna   218
18  2021       10            18  kujawsko-pomorskie                  Pokoje Toruń Centrum             Torun                87-100     ulica Bolesława Chrobrego                 5/9              Kwatera prywatna   184
19  2021       10            18  zachodniopomorskie            Pokoje typu Studio OLSZYNA  Ustronie Morskie                78-111                 ulica Olszyna                  27              Kwatera prywatna   145
```

Step 2. Verify variable types

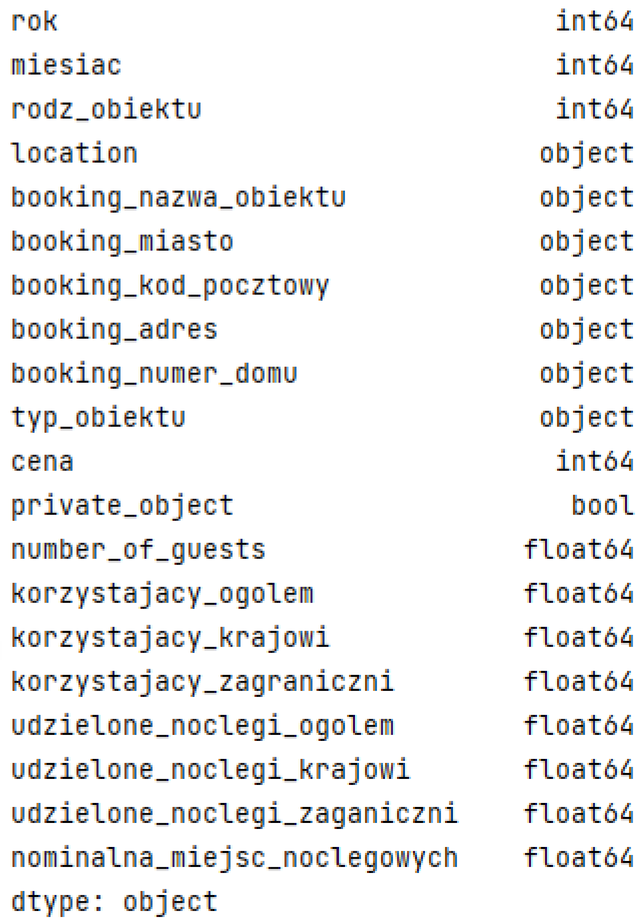
The previous step gives a general overview of what variables are present in the collection. You need to be sure of the types of each variable.

WHY?

- Verify the types of variables present in the collection (integer variables, floating point variables, categorical variables, logical type variables, dates).
- Fixing structural errors.
- Fixing type conversion and syntax errors.

Step 2. Verify variable types

- df.dtype



```
rok                                int64
miesiac                            int64
rodz_obiektu                       int64
location                          object
booking_nazwa_obiektu             object
booking_miasto                    object
booking_kod_pocztowy              object
booking_adres                     object
booking_numer_domu                object
typ_obiektu                       object
cena                               int64
private_object                      bool
number_of_guests                 float64
korzystajacy_ogolem              float64
korzystajacy_krajowi             float64
korzystajacy_zagraniczni         float64
udzielone_noclegi_ogolem         float64
udzielone_noclegi_krajowi        float64
udzielone_noclegi_zaganiczni     float64
nominalna_miejsc_noclegowych     float64
dtype: object
```

Funded by
the European Union

Step 3. Create data summary

A summary of the variables describing the dataset containing basic information about the numeric variables, such as:

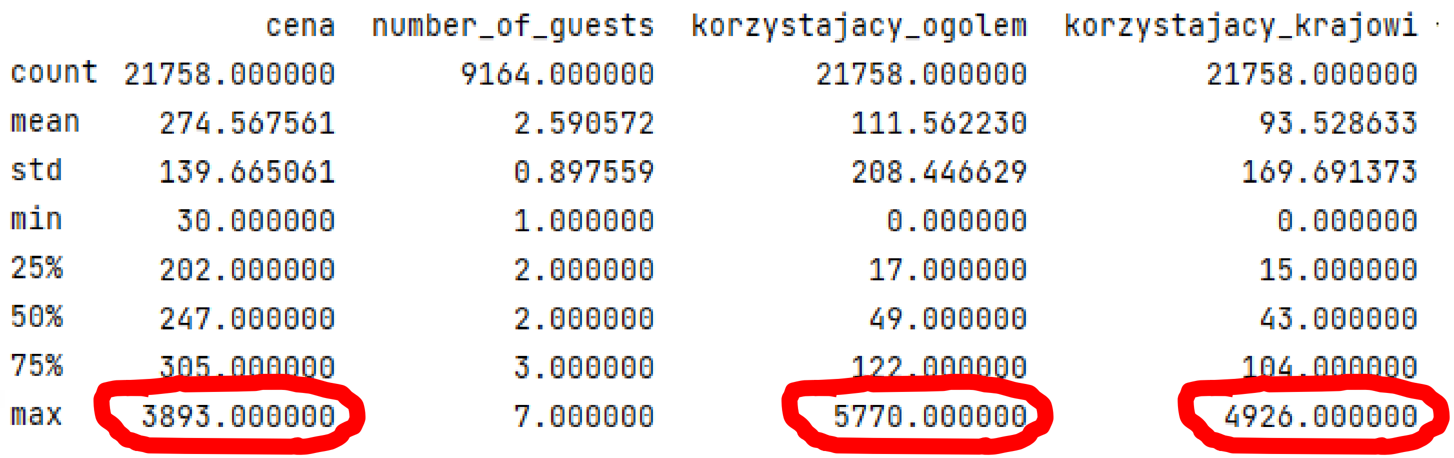
- Minimum and maximum values
- Mean and median
- Second (lower) quartile and third (upper) quartile
- Standard deviation.

WHY?

- Entry point for further analysis.
- Knowing which variables to keep an eye on in the next steps.

Step 3. Create data summary

- df.describe()



```
               cena  number_of_guests  korzystajacy_ogolem  korzystajacy_krajowi
count  21758.000000       9164.000000         21758.000000          21758.000000
mean     274.567561          2.590572           111.562230             93.528633
std      139.665061          0.897559           208.446629            169.691373
min       30.000000          1.000000             0.000000              0.000000
25%      202.000000          2.000000            17.000000             15.000000
50%      247.000000          2.000000            49.000000             43.000000
75%      305.000000          3.000000           122.000000            104.000000
max     3893.000000          7.000000          5770.000000           4926.000000
```

Step 4. Check the missing data

Create a summary focusing on finding missing values in the set. What variables contain missing values and what is the number of missing values. Remember that in some cases missing data can also be valuable information.

WHY?

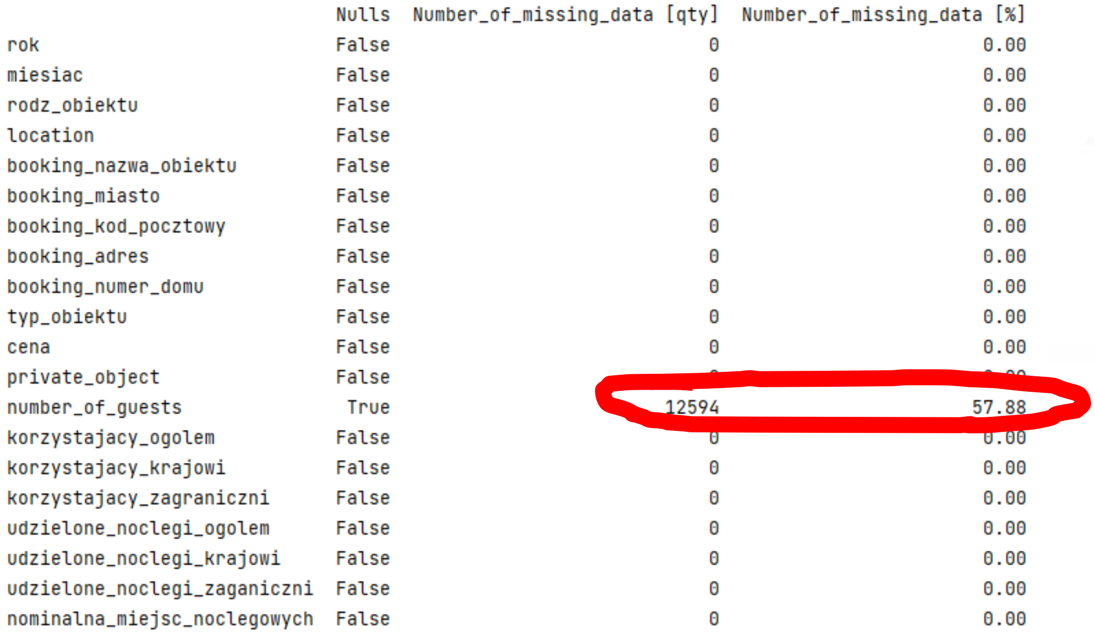
- Some algorithms are sensitive to missing values.
- Knowing how many missing values a variable contains makes it easier to decide whether to include it in the model.

Funded by the European Union

Step 4. Check the missing data

- df.isnull()

```
md_summary = pd.DataFrame(df.isnull().any(), columns=['Nulls'])
md_summary['Number_of_missing_data [qty]'] = pd.DataFrame(df.isnull().sum())
md_summary['Number_of_missing_data [%]'] = round((df.isnull().mean()*100),2)
```



```
                              Nulls  Number_of_missing_data [qty]  Number_of_missing_data [%]
rok                           False                             0                        0.00
miesiac                       False                             0                        0.00
rodz_obiektu                  False                             0                        0.00
location                      False                             0                        0.00
booking_nazwa_obiektu         False                             0                        0.00
booking_miasto                False                             0                        0.00
booking_kod_pocztowy          False                             0                        0.00
booking_adres                 False                             0                        0.00
booking_numer_domu            False                             0                        0.00
typ_obiektu                   False                             0                        0.00
cena                          False                             0                        0.00
private_object                False                             0                        0.00
number_of_guests               True                         12594                       57.88
korzystajacy_ogolem           False                             0                        0.00
korzystajacy_krajowi          False                             0                        0.00
korzystajacy_zagraniczni      False                             0                        0.00
udzielone_noclegi_ogolem      False                             0                        0.00
udzielone_noclegi_krajowi     False                             0                        0.00
udzielone_noclegi_zaganiczni  False                             0                        0.00
nominalna_miejsc_noclegowych  False                             0                        0.00
```

Web Intelligence Network

Funded by the European Union

Step 4a. Deal with missing data

Delete the missing data

Impute the missing data

Flag the missing data

Step 5. Check distribution of variables

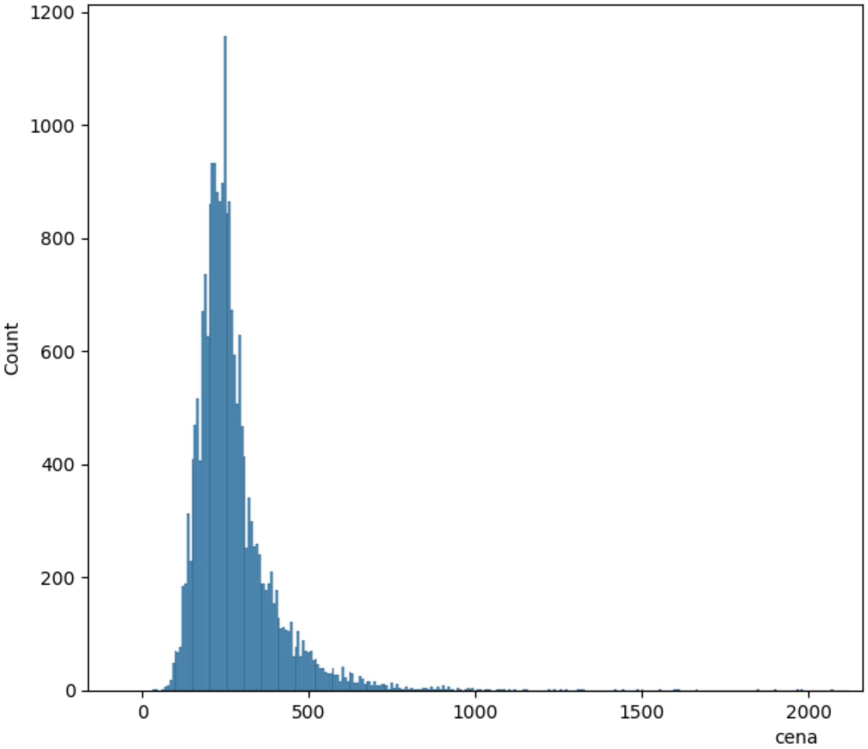
Calculate the values of each quartile and skewness. For each numerical variable, produce a histogram and try to recognise the distribution.

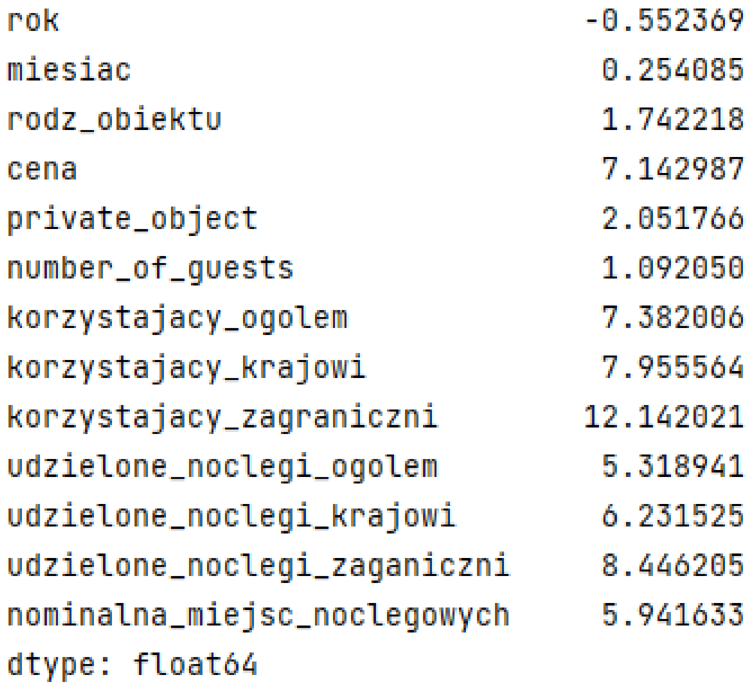
WHY?

- The conclusions can be used, for example, in the imputation of numerical variables.
- Some statistical techniques have assumptions about the distribution of the variables (e.g. in Pearson correlation it is desirable that the variables have a normal distribution).

Step 5. Check distribution of variables

- df.skew()





```
rok                               -0.552369
miesiac                            0.254085
rodz_obiektu                       1.742218
cena                               7.142987
private_object                     2.051766
number_of_guests                   1.092050
korzystajacy_ogolem                7.382006
korzystajacy_krajowi               7.955564
korzystajacy_zagraniczni          12.142021
udzielone_noclegi_ogolem           5.318941
udzielone_noclegi_krajowi          6.231525
udzielone_noclegi_zaganiczni       8.446205
nominalna_miejsc_noclegowych       5.941633
dtype: float64
```

Web Intelligence Network

Funded by the European Union

Step 6. Identify outlier observations

Outlier points are data points that are drastically different from others in the set. They can cause issues with certain types of data models and analyses. Removal of outliers can only occur if we are certain that they are wrong, e.g. if they are clearly caused by incorrect data entry.

WHY?

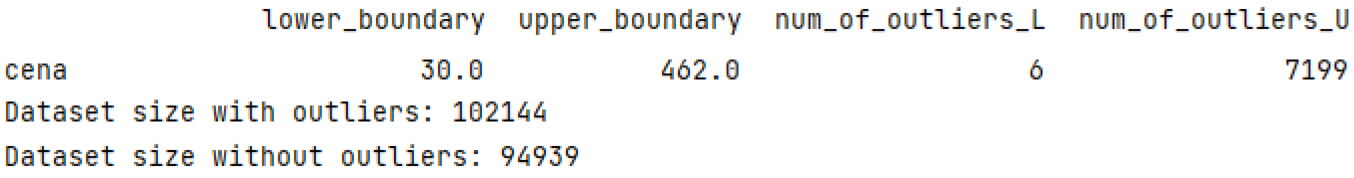
- Some algorithms are sensitive to the presence of outlier observations.
- Some methods used in statistics (e.g. Pearson correlation), are sensitive to outliers.

Funded by the European Union

Step 6. Identify outlier observations

```
q1 = df.quantile(0.25)
q3 = df.quantile(0.75)
iqr = q3-q1

low_boundary = (q1 - 1.5 * iqr)
upp_boundary = (q3 + 1.5 * iqr)
num_of_outliers_L = (df[iqr.index] < low_boundary).sum()
num_of_outliers_U = (df[iqr.index] > upp_boundary).sum()
```



```
                 lower_boundary  upper_boundary  num_of_outliers_L  num_of_outliers_U
cena                       30.0           462.0                  6               7199
Dataset size with outliers: 102144
Dataset size without outliers: 94939
```

Funded by
the European Union

Step 7. Check categorical variables

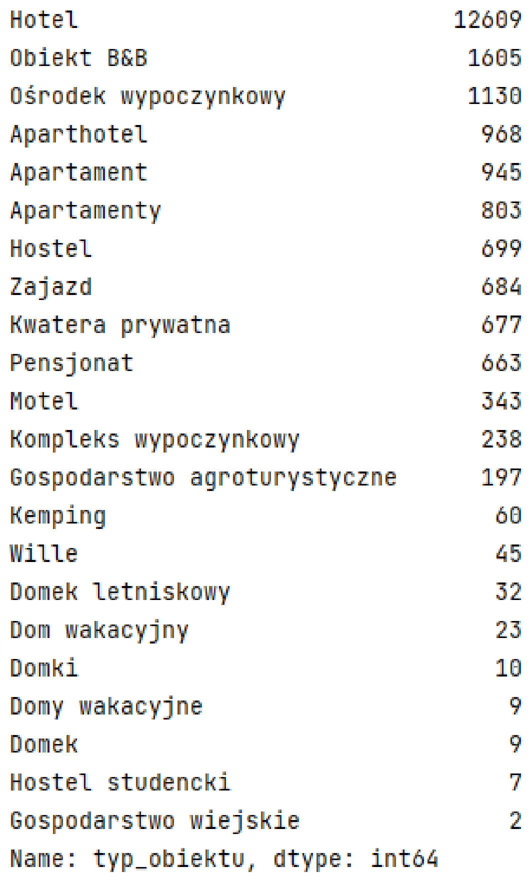
Check the counts of categorical variables. Summary should include the number of categorical variables, the number of categories included in each variable, the coverage of the set by each category and the percentage coverage of the set by each category.

WHY?

- Knowledge of whether the set is appropriately balanced.
- Knowing the coverage of the set by categories often allows you to focus on the most relevant ones.

Step 7. Check categorical variables

```
for col in df.select_dtypes(['object', 'category']):
    print(df[col].value_counts())
```



```
Hotel                           12609
Obiekt B&B                       1605
Ośrodek wypoczynkowy             1130
Aparthotel                        968
Apartament                        945
Apartamenty                       803
Hostel                            699
Zajazd                            684
Kwatera prywatna                  677
Pensjonat                         663
Motel                             343
Kompleks wypoczynkowy             238
Gospodarstwo agroturystyczne      197
Kemping                            60
Wille                              45
Domek letniskowy                   32
Dom wakacyjny                      23
Domki                              10
Domy wakacyjne                      9
Domek                               9
Hostel studencki                    7
Gospodarstwo wiejskie               2
Name: typ_obiektu, dtype: int64
```

Funded by the European Union

Step 8. Check correlation between variables

Verification of the levels of coefficients:

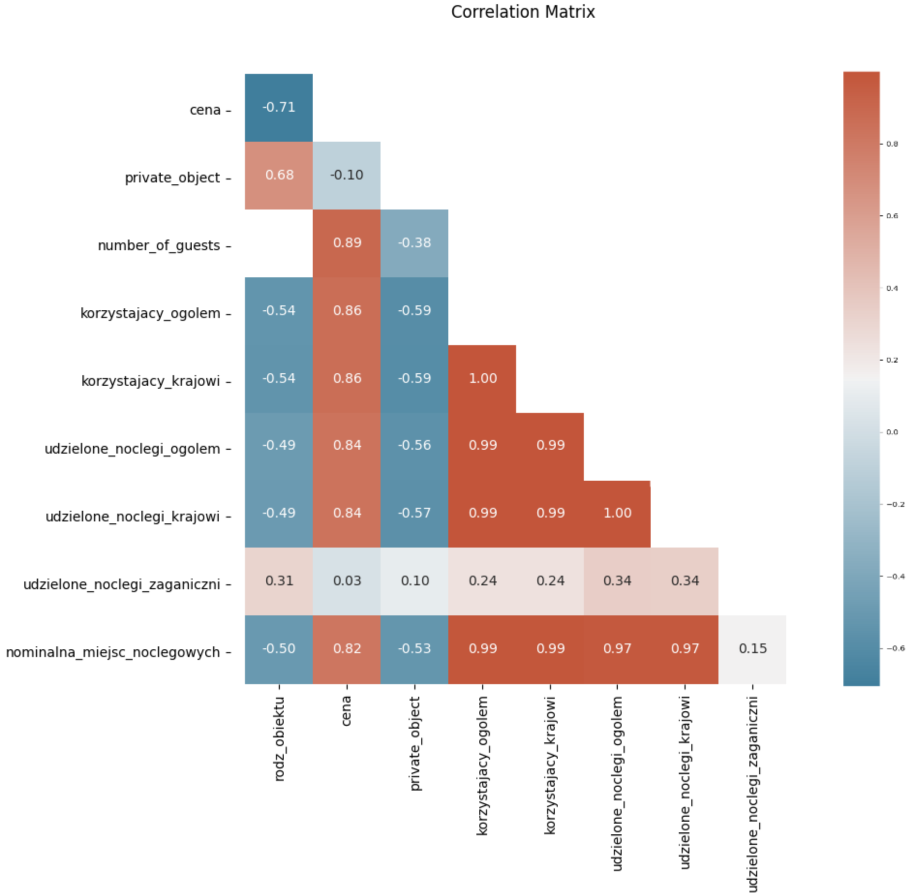
- Correlation between numerical variables.
- Correlations between categorical variables.
- Correlation between categorical and numerical variables.

WHY?

- To discover correlations between variables. Correlation information can be used, for example, at the variable transformation stage.
- On the basis of correlation analysis, the decision on the choice of variables for the model can be made.

Step 8. Check correlation between variables

- df.corr()



Correlation Matrix

	rodz_obiektu	cena	private_object	korzystajacy_ogolem	korzystajacy_krajowi	udzielone_noclegi_ogolem	udzielone_noclegi_krajowi	udzielone_noclegi_zaganiczni
cena	-0.71							
private_object	0.68	-0.10						
number_of_guests		0.89	-0.38					
korzystajacy_ogolem	-0.54	0.86	-0.59					
korzystajacy_krajowi	-0.54	0.86	-0.59	1.00				
udzielone_noclegi_ogolem	-0.49	0.84	-0.56	0.99	0.99			
udzielone_noclegi_krajowi	-0.49	0.84	-0.57	0.99	0.99	1.00		
udzielone_noclegi_zaganiczni	0.31	0.03	0.10	0.24	0.24	0.34	0.34	
nominalna_miejsc_noclegowych	-0.50	0.82	-0.53	0.99	0.99	0.97	0.97	0.15

Data cleaning and analysis done!

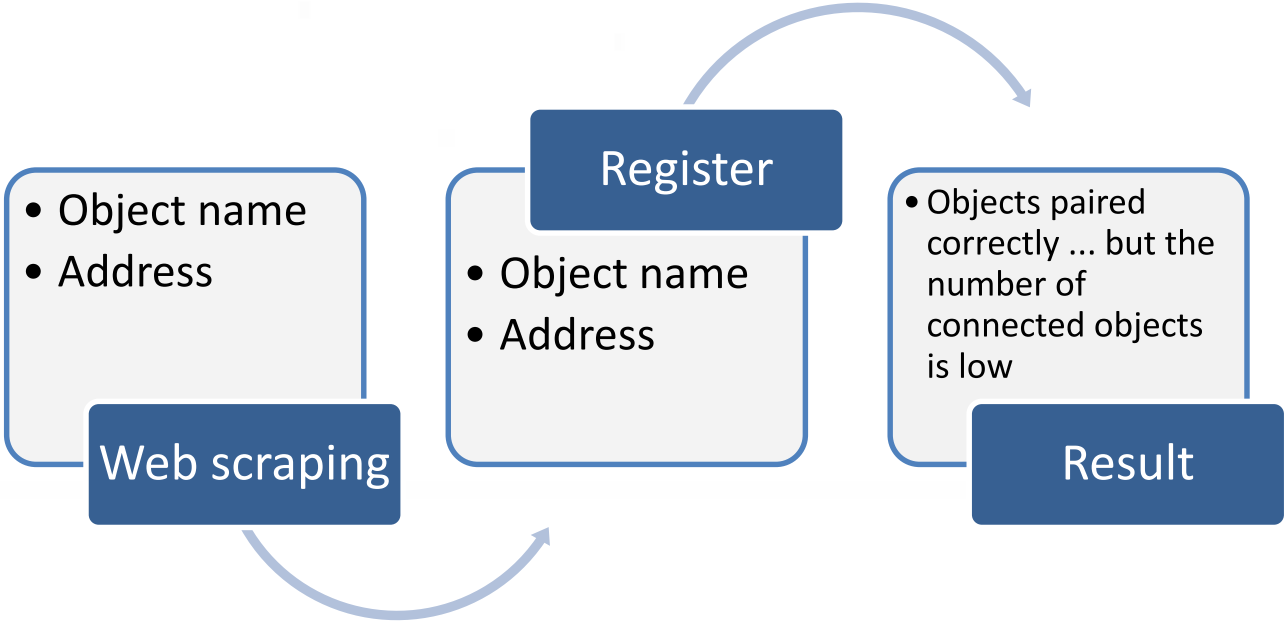
… what now?

Record linkage

How do we check which web scraping establishments we already have in our registers?

- Let’s use object name and/or address.

Record linkage



Web scraping

- Object name
- Address

Register

- Object name
- Address

Result

- Objects paired correctly ... but the number of connected objects is low

Funded by the European Union

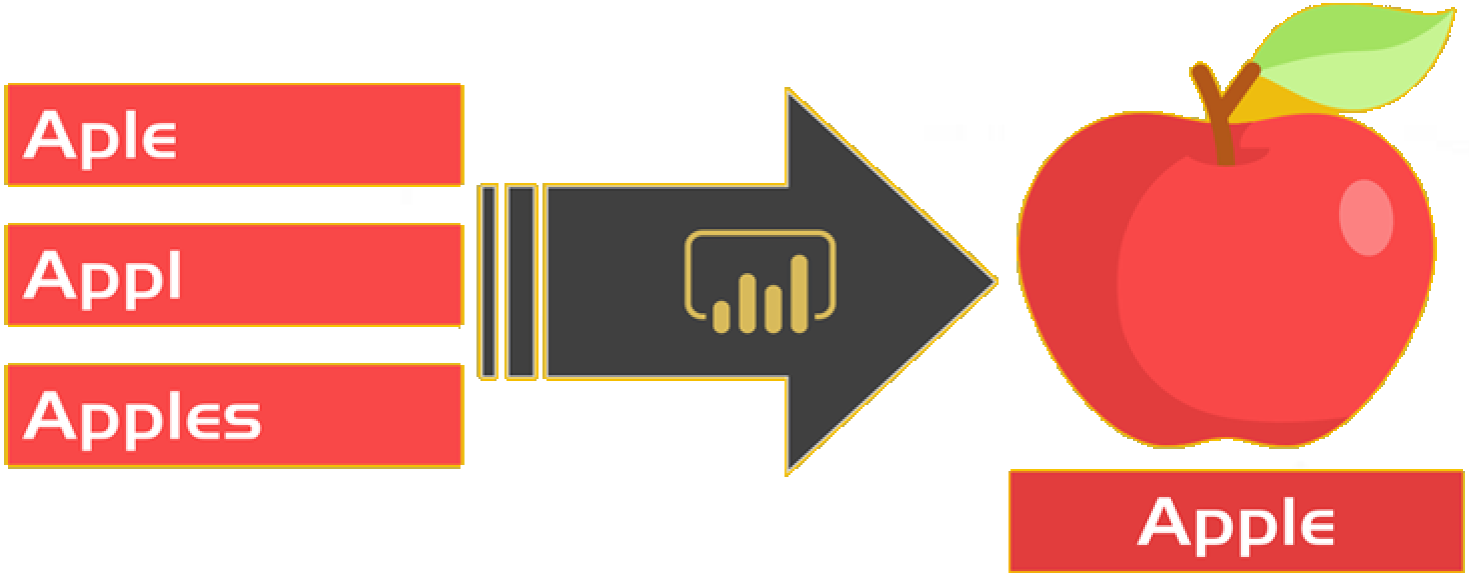
Record linkage

How do we combine objects that have similar but not identical names?

- Let us calculate the similarity between names and addresses in the register and the set from web scraping using for example, the Levenshtein, Jaro-Winkler or Jaccard formula. This is known as a fuzzy matching.

Funded by the European Union

Record linkage



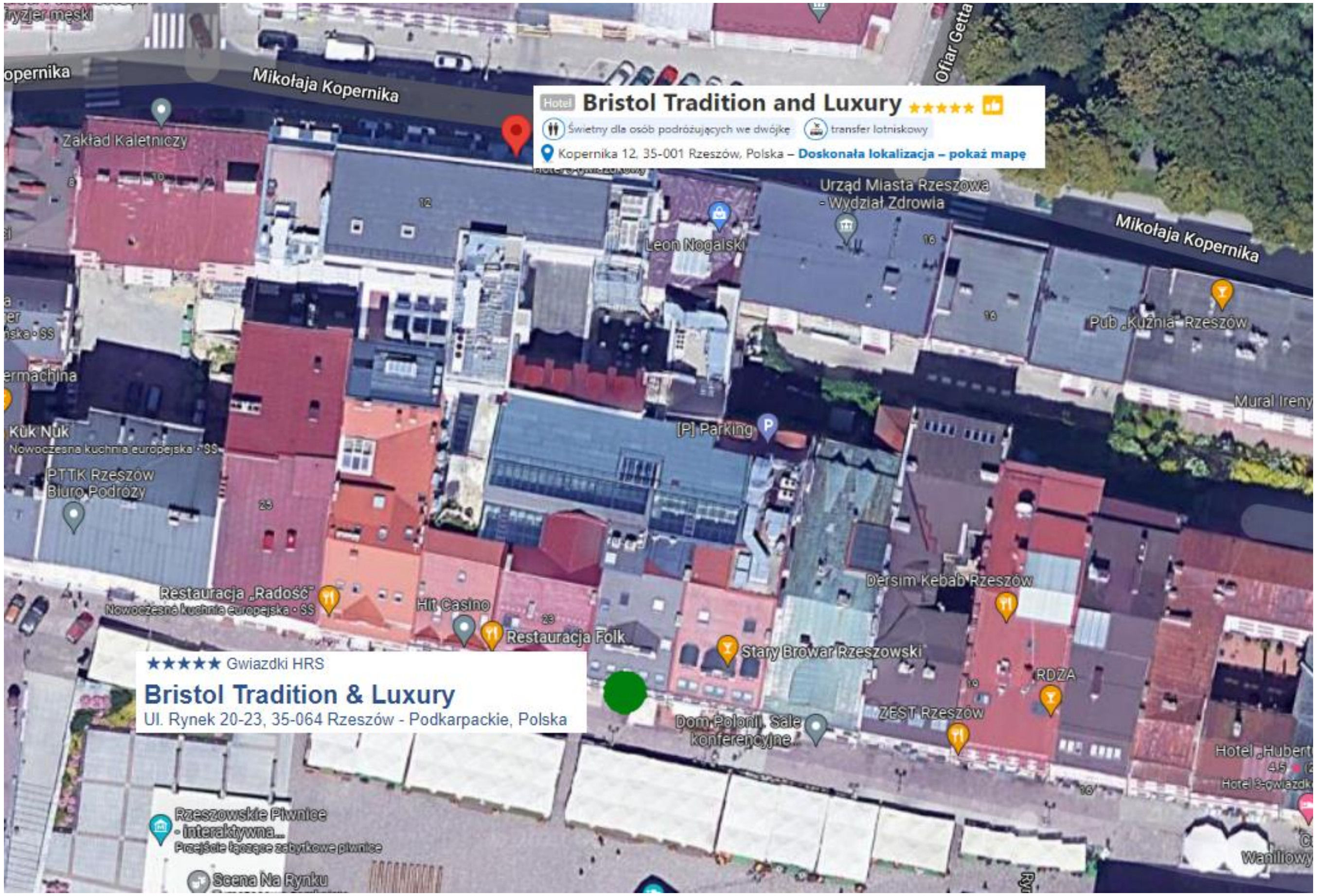
- A technique for finding strings of characters that match an approximate pattern.
- A fault-tolerant search that returns records even if the search term contains typos or extra/missing characters.

Record linkage

What about geographical coordinates?

- Let us calculate the distance between objects in the register and database from web scraping using, for example, the haversinus formula or the Vincentian formula.

Record linkage



Hotel Bristol Tradition and Luxury ★★★★★

Świetny dla osób podróżujących we dwójkę | transfer lotniskowy

Kopernika 12, 35-001 Rzeszów, Polska – Doskonała lokalizacja – pokaż mapę

★★★★★ Gwiazdki HRS

Bristol Tradition & Luxury

Ul. Rynek 20-23, 35-064 Rzeszów - Podkarpackie, Polska

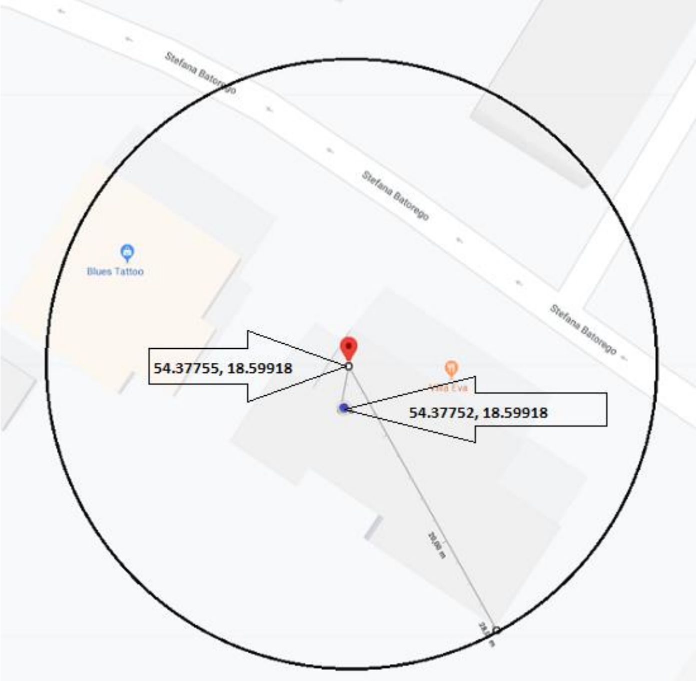
Web Intelligence Network

Funded by the European Union

Record linkage

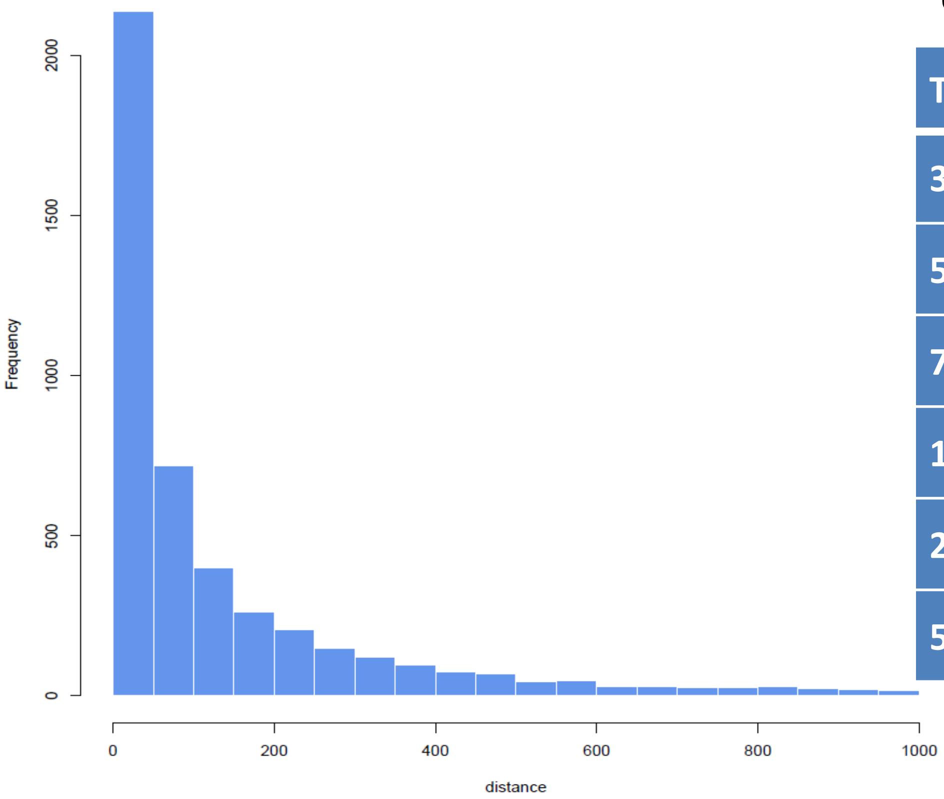
The distance-based approach can be applied in the following way:

- calculate the distance between all establishments
- for each establishment find all establishments within a threshold
- match the closest one



Record linkage

Minimum Haversine distance between scraped and registered accommodations within municipalities



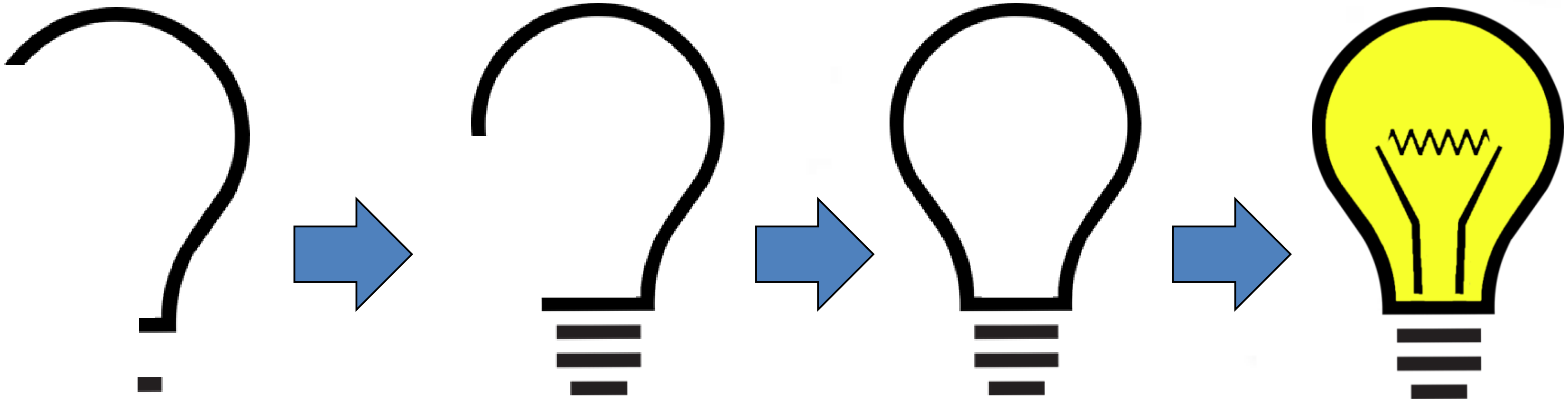
Quality measures of data linkage based on confusion matrix

Threshold	Precision	Sensitivity	Accuracy
30 m	1	0.5	0.82
50 m	1	0.52	0.82
70 m	1	0.55	0.83
100 m	1	0.52	0.8
200 m	1	0.64	0.87
500 m	0.97	0.6	0.81

Web Intelligence Network

Funded by the European Union

Conclusions



Record linkage?

Deterministic record linkage

Probabilistic (or fuzzy) record linkage

Distance-based linkage + objects' name

Funded by the European Union

Thank you for your attention!

Funded by the European Union