Adatminőség kiértékelő eszköz fejlesztése

Pósfay Benedek, konzulens: Gönczy László

2024. 05. 30.

Budapest University of Technology and Economics
Department of Measurement and Information Systems
ftsrg Research Group

Adatminőség fontossága ML területen

1. Modellezési pontosság javítása
   - Zaj csökkentése
   - Példa: energiafogyasztás pontosabb előrejelzése
2. Megbízható eredmények
   - Téves következtetések → rossz döntések
   - Példa: rossz minőségű ingatlanadatok → hamis kép a piacról
3. Hatékonyabb modellezés
   - Jobb minőségű adatok → gyorsabb modell tanítás
   - Kevesebb idő és erőforrás

Példa: kanadai energiafogyasztás előrejelzés

- Vancouveri háztartások
  - Óránkénti energiafelhasználási adatok
  - Következő óra fogyasztásának előrejelzése
- Összecsúszott fogyasztás
  - 12-24 órányi fogyasztás 1 óra alatt
  - **<0.1%** adatoknak
- Javítás után több, mint **50%-kal nőtt** a pontosság
  - Random Forest modell

Projekt célja

Általános **adatminőség kiértékelő** eszköz létrehozása

Python

- ML területen elterjedt
- Pandas
- Korábbi energiafogyasztási projekt alapjai

Great Expectations

- Adatminőség ellenőrzésének eszköze
- Bővíthető, testreszabható
- Automatikus validáció
- HTML jelentések generálása

Adatforrások kiválasztása

Madridi ingatlanhirdetés

- ~21 000 sor
- 40+ feature
  - Ár, terület, szobák száma, típusa

Lakossági energiafogyasztás

- ~600 000 sor
- 20+ feature
  - Fogyasztás, időjárás, ház metaadatok

Madridi ingatlanhirdetés 2

- ~900 sor
- 13 feature
  - Ár, terület, szobák száma, hirdetők adatai

Melbourne-i ingatlanhirdetés

- ~13 000 sor
- 20+ feature
  - Ár, terület, szobák száma, típusa

Kiértékelő algoritmus

Bemenet

- Adatforrás
- Oszlop metaadatok
  - Típusok
- Szigorúság
- (Exportálás részletessége)

Kiértékelő algoritmus

Bemenet

- Adatforrás
- Oszlop metaadatok
  - Típusok
- Szigorúság
- (Exportálás részletessége)

Minőség értékelése

- Érték 0-1 között
- Különböző szabályok teljesülésének súlyozott átlaga

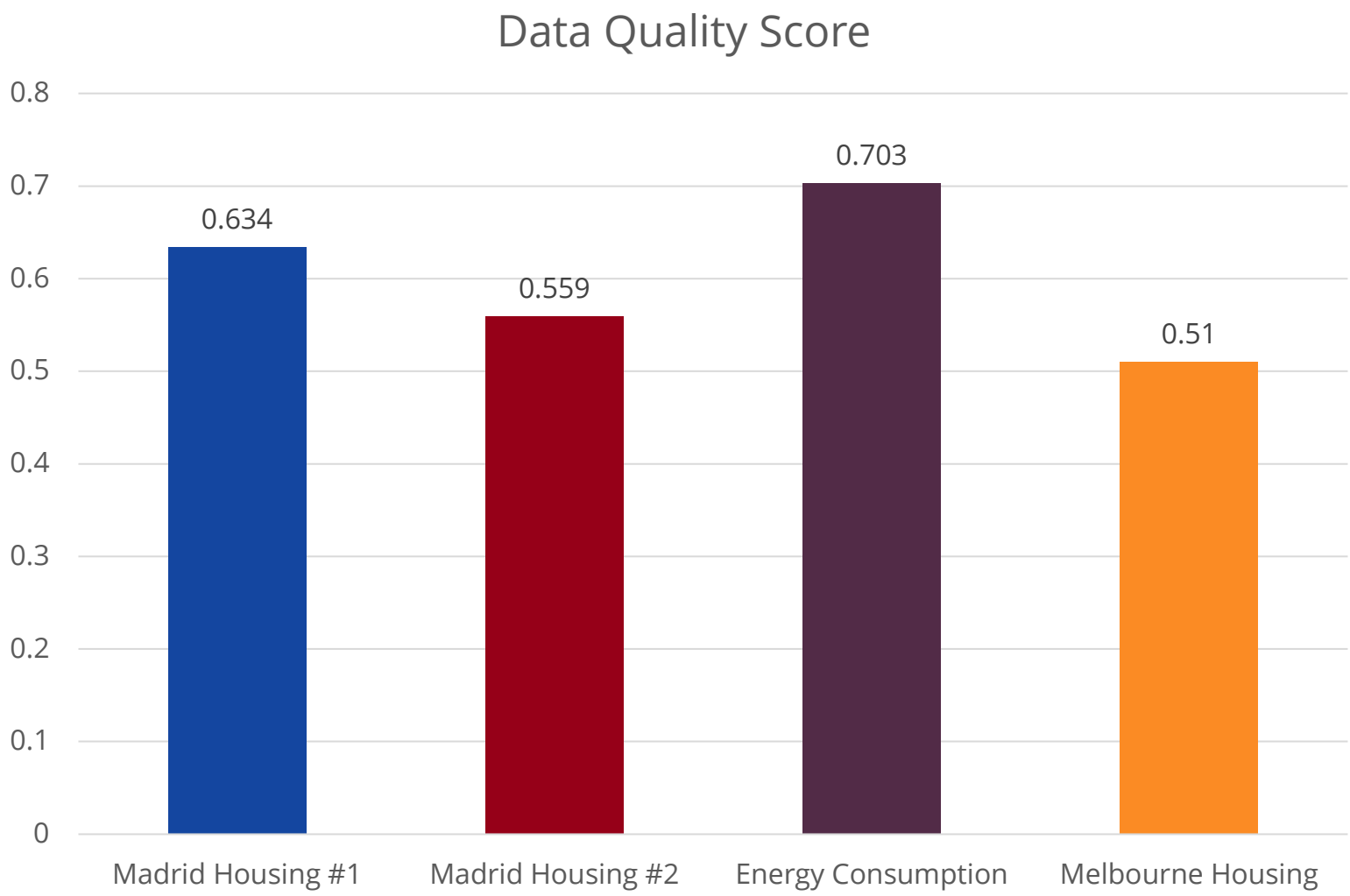
Általános jelentés

- Great Expectation generálja
  - HTML
- Oszlopszintű statisztikák

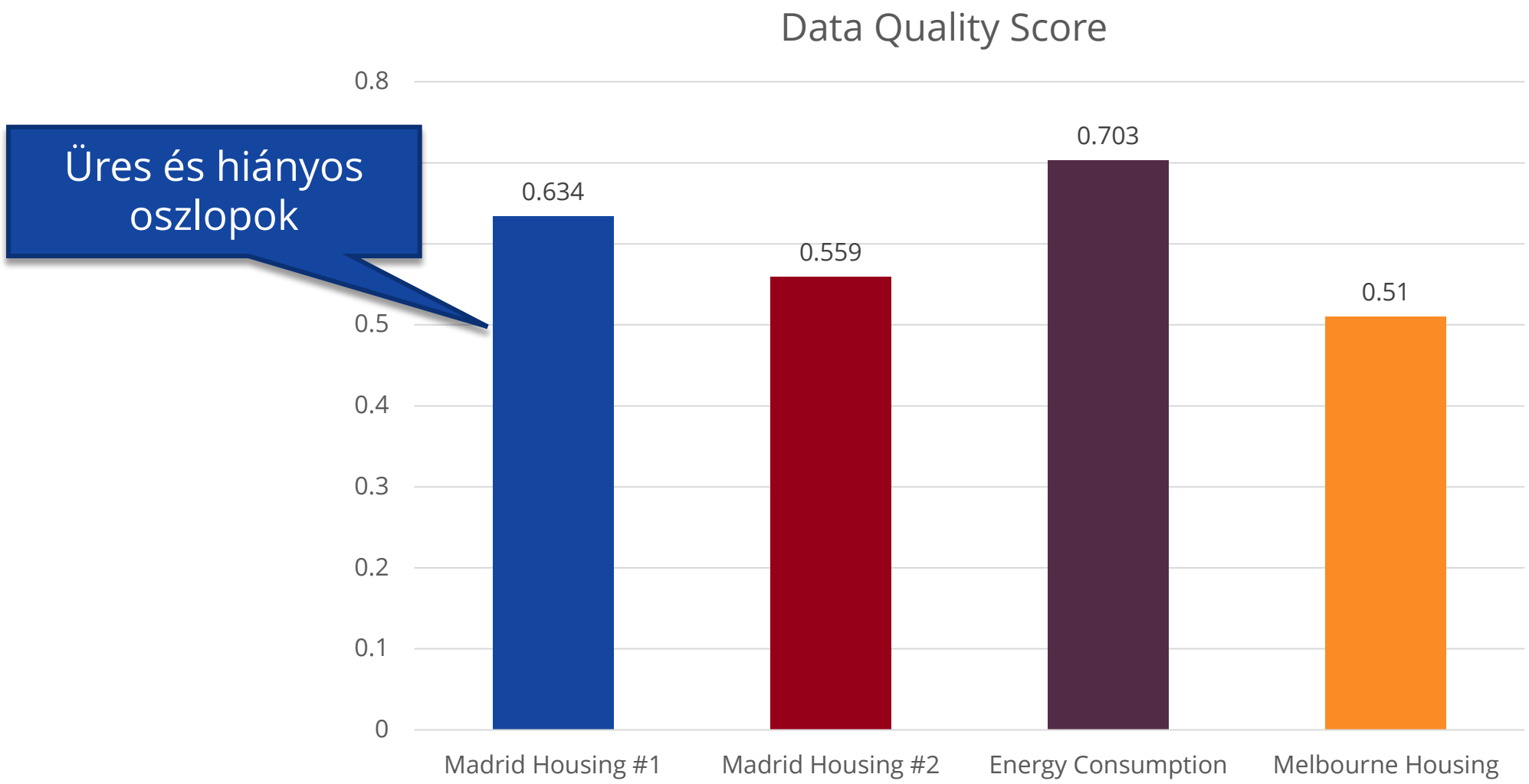
Hibagyűjtő funkció

- Minden sor, ahol legalább 1 elvárás nem teljesül
  - Eredeti adatokkal
  - Elvárásokkal
- csv formátum

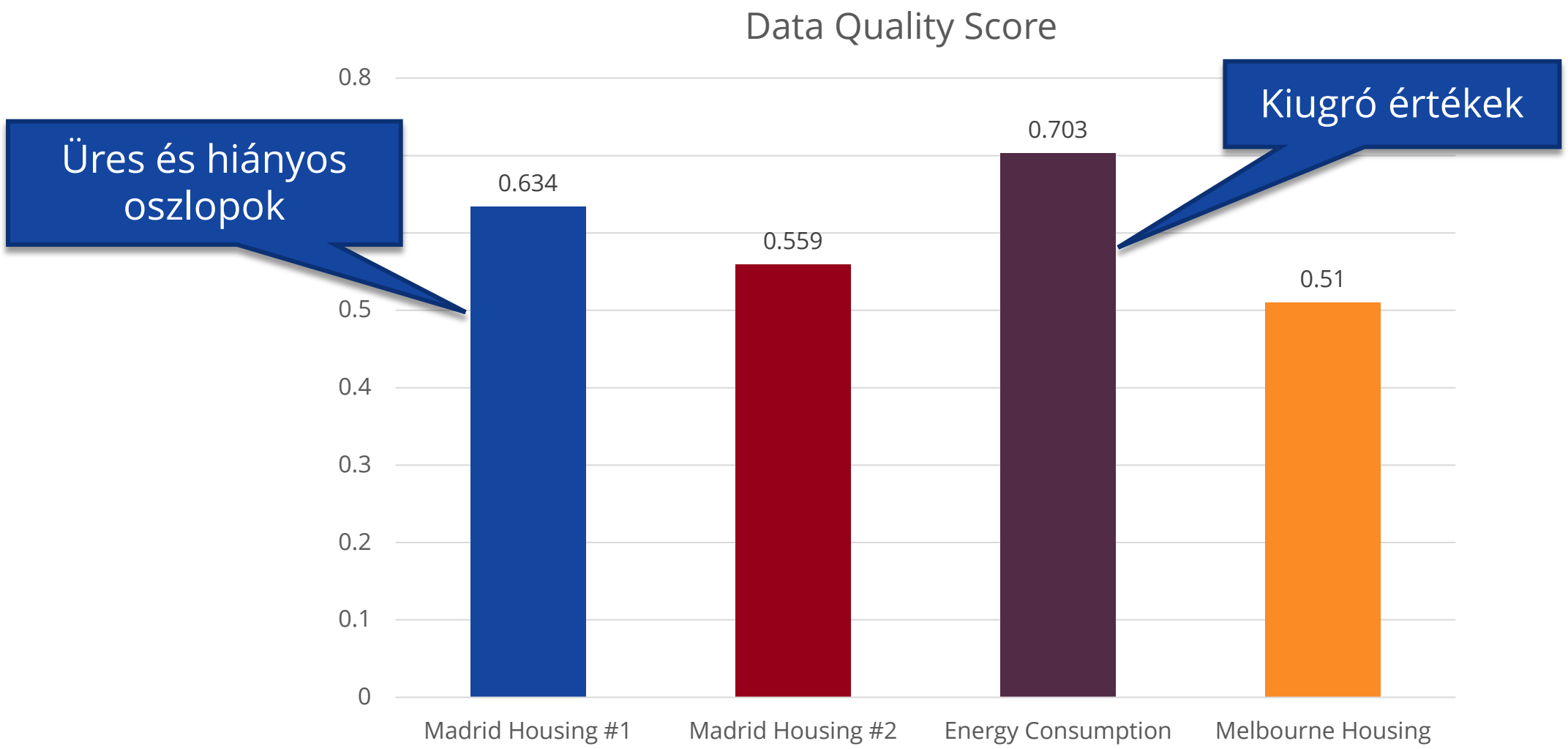
Adatforrások értékelése



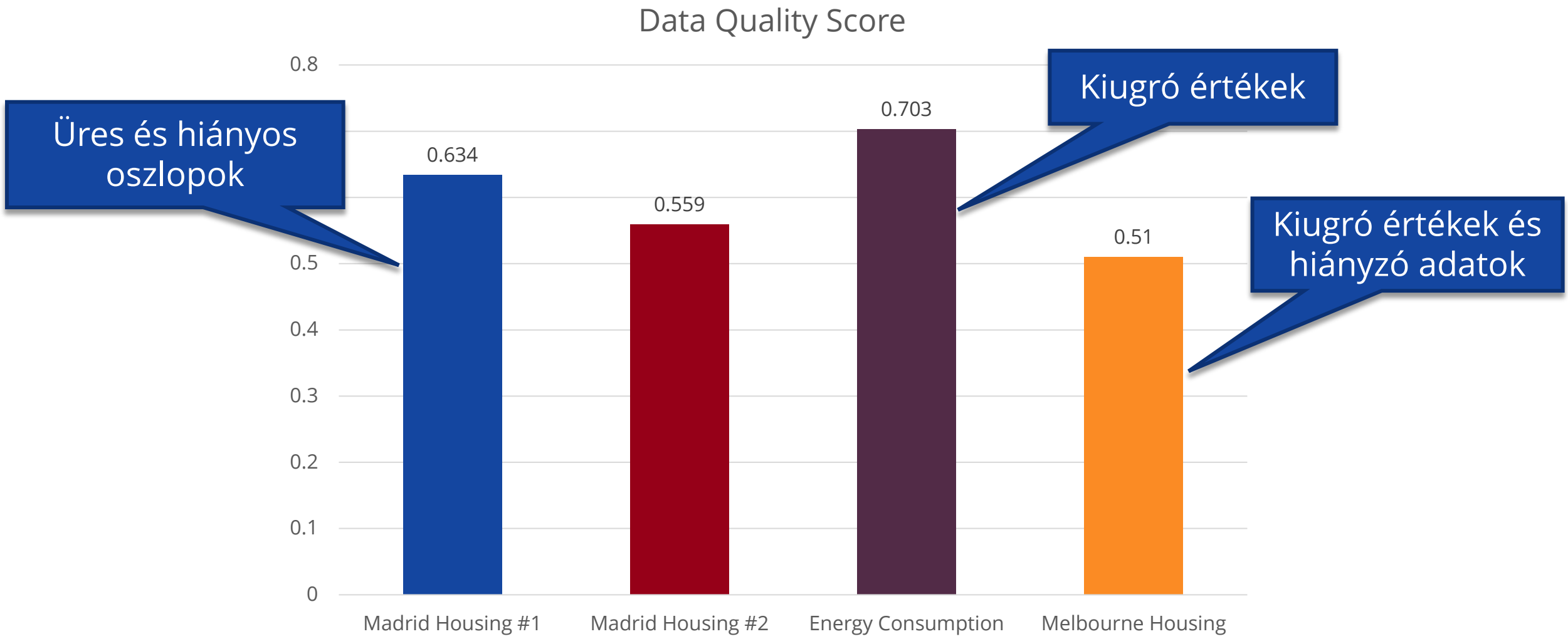
Adatforrások értékelése



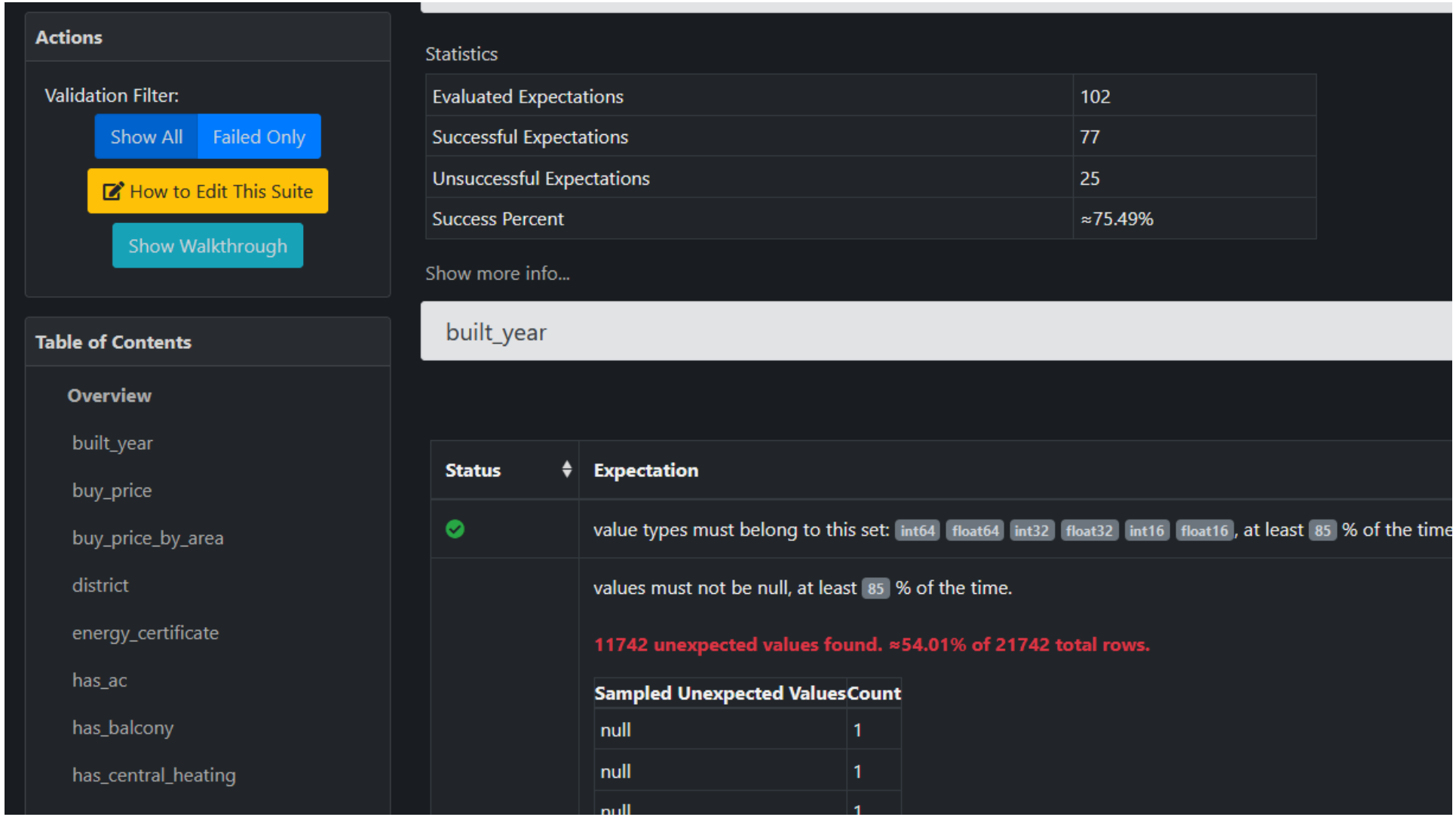
Adatforrások értékelése



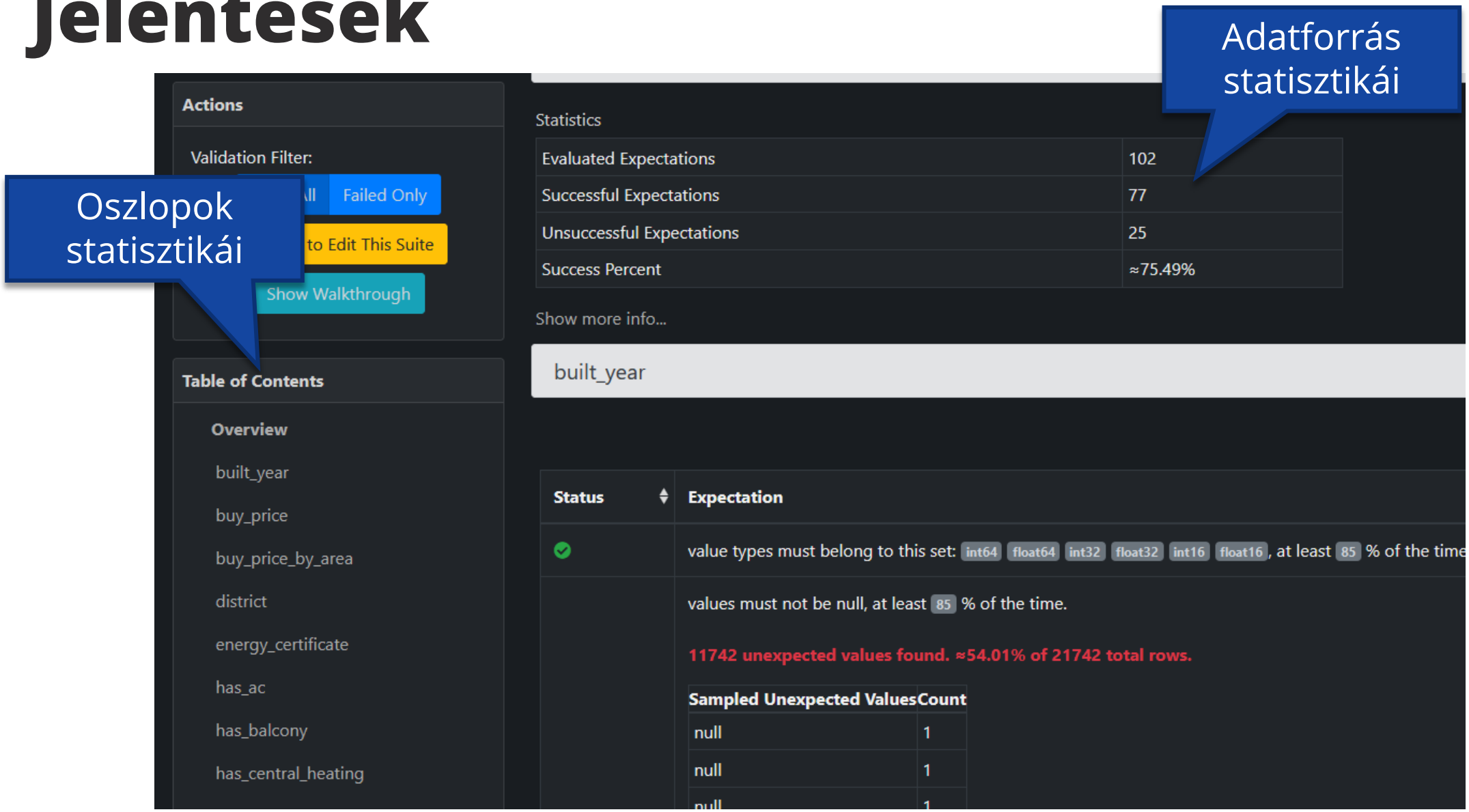
Adatforrások értékelése



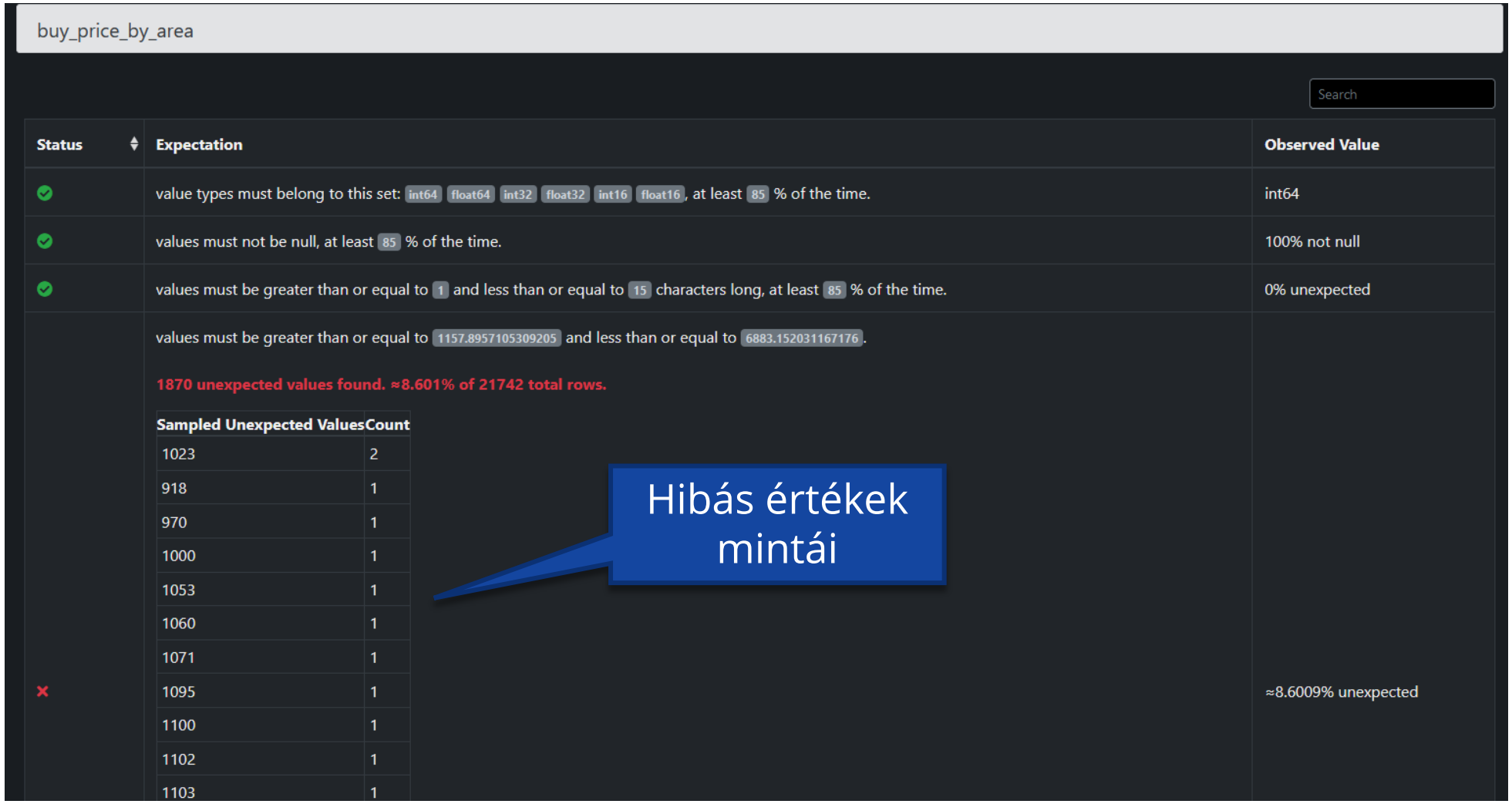
Jelentések


Actions
Validation Filter:
Show All
Failed Only
How to Edit This Suite
Show Walkthrough
Table of Contents
Overview
built_year
buy_price
buy_price_by_area
district
energy_certificate
has_ac
has_balcony
has_central_heating
Statistics
Evaluated Expectations 102
Successful Expectations 77
Unsuccessful Expectations 25
Success Percent ≈75.49%
Show more info...
built_year
Status
Expectation
value types must belong to this set: int64 float64 int32 float32 int16 float16, at least 85 % of the time
values must not be null, at least 85 % of the time.
11742 unexpected values found. ≈54.01% of 21742 total rows.
Sampled Unexpected Values Count
null 1
null 1

Jelentések


Adatforrás statisztikái
Oszlopok statisztikái
Actions
Validation Filter:
Failed Only
to Edit This Suite
Show Walkthrough
Table of Contents
Overview
built_year
buy_price
buy_price_by_area
district
energy_certificate
has_ac
has_balcony
has_central_heating
Statistics
Evaluated Expectations 102
Successful Expectations 77
Unsuccessful Expectations 25
Success Percent ≈75.49%
Show more info...
built_year
Status
Expectation
value types must belong to this set: int64 float64 int32 float32 int16 float16, at least 85 % of the time
values must not be null, at least 85 % of the time.
11742 unexpected values found. ≈54.01% of 21742 total rows.
Sampled Unexpected Values Count
null 1
null 1

Jelentések



buy_price_by_area

Search

Status	Expectation	Observed Value
✓	value types must belong to this set: int64 float64 int32 float32 int16 float16, at least 85 % of the time.	int64
✓	values must not be null, at least 85 % of the time.	100% not null
✓	values must be greater than or equal to 1 and less than or equal to 15 characters long, at least 85 % of the time.	0% unexpected
✗	values must be greater than or equal to 1157.8957105309205 and less than or equal to 6883.152031167176. **1870 unexpected values found. ≈8.601% of 21742 total rows.**	≈8.6009% unexpected

Sampled Unexpected Values	Count
1023	2
918	1
970	1
1000	1
1053	1
1060	1
1071	1
1095	1
1100	1
1102	1
1103	1

Jelentések

id	title	subtitle	raw_address	street_name	street_number	expectation_failed	house_type_id_length_between	has_central_heating_not_null	has_individual_heating_not_null
21216	Piso en ve	Los Ãngele	Calle de la del M	Calle de la del M	65 -69	TRUE	FALSE	TRUE	TRUE
19700	Piso en ve	MoscardÃ³	Calle de Antonio	Calle de Antonio LÃ³pez		TRUE	FALSE	TRUE	TRUE
19474	Piso en ve	MoscardÃ³	Calle de Antonio	Calle de Antonic	115	TRUE	FALSE	FALSE	FALSE
19235	Piso en ve	TetuÃ¡n, Madrid				TRUE	FALSE	TRUE	TRUE
18986	Piso en ve	TetuÃ¡n, Madrid				TRUE	FALSE	FALSE	FALSE
18926	Piso en ve	TetuÃ¡n, Madrid				TRUE	FALSE	TRUE	TRUE
18910	Estudio en	Cuatro Car	Calle Orden	Calle Orden		TRUE	TRUE	FALSE	FALSE
18771	Estudio en	Cuatro Car	Calle Orden, 6	Calle Orden	6	TRUE	TRUE	TRUE	TRUE
18730	Piso en ve	Bellas Vist	Calle de Bravo M	Calle de Bravo Murillo		TRUE	FALSE	TRUE	TRUE
18676	Piso en ve	Cuzco-Cas	Calle de la Infan	Calle de la Infanta Mercedes, Mad		TRUE	FALSE	TRUE	TRUE
18512	DÃºplex en	Cuzco-Cas	Calle Pensamier	Calle Pensamie	27	TRUE	FALSE	TRUE	TRUE
18426	Estudio en	TetuÃ¡n, Madrid				TRUE	TRUE	TRUE	TRUE
18366	Piso en ve	Cuatro Car	Calle del Aviado	Calle del Aviadc	32	TRUE	FALSE	TRUE	TRUE
18306	Piso en ve	TetuÃ¡n, Madrid				TRUE	FALSE	FALSE	FALSE
18260	Piso en ve	Cuzco-Cas	Calle Rosario Pir	Calle Rosario Pi	18	TRUE	FALSE	TRUE	TRUE
18251	Piso en ve	Cuzco-Cas	Calle de Rosario	Calle de Rosaric	18	TRUE	FALSE	TRUE	TRUE
18250	Piso en ve	Cuzco-Cas	Calle de Rosario	Calle de Rosaric	18	TRUE	FALSE	TRUE	TRUE
18131	Piso en ve	Bellas Vist	Wad Ras, 10	Wad Ras	10	TRUE	FALSE	TRUE	TRUE
18076	Piso en ve	Bellas Vist	Calle de Wad-Ra	Calle de Wad-R	10	TRUE	FALSE	TRUE	TRUE
18075	Piso en ve	Bellas Vist	Calle de Wad-Ra	Calle de Wad-R	10	TRUE	FALSE	TRUE	TRUE
18036	Estudio en	TetuÃ¡n, Madrid				TRUE	TRUE	FALSE	FALSE
18030	Piso en ve	Cuatro Car	Calle del Aviado	Calle del Aviador Zorita		TRUE	FALSE	TRUE	TRUE
17929	Piso en ve	Cuzco-Cas	Calle de Bravo M	Calle de Bravo Murillo		TRUE	FALSE	FALSE	FALSE
17906	Piso en ve	Retiro, Madrid				TRUE	FALSE	FALSE	FALSE
17837	Piso en ve	Adelfas, Ma	Avenida de la Ci	Avenida de la Ciudad de Barcelona		TRUE	FALSE	TRUE	TRUE
17797	Piso en ve	PacÃfico	Calle del Doctor	Calle del Doctor Esquerdo		TRUE	FALSE	TRUE	TRUE

Jelentések

id	title	subtitle	raw_address	street_name	street_number	expectation_failed	house_type_id_length_between	has_central_heating_not_null	has_individual_heating_not_null
21216	Piso en ve	Los Ãngele	Calle de la del M	Calle de la del M	65 -69	TRUE	FALSE	TRUE	TRUE
19700	Piso en ve	MoscardÃ³	Calle de Antonio	Calle de Antonio LÃ³pez		TRUE	FALSE	TRUE	TRUE
19474	Piso en ve	MoscardÃ³	Calle de Antonio	Calle de Antonic	115	TRUE	FALSE	FALSE	FALSE
19235	Piso en ve	TetuÃ¡n, Madrid				TRUE	FALSE	TRUE	TRUE
18986	Piso en ve	TetuÃ¡n, Madrid				TRUE	FALSE	FALSE	FALSE
18926	Piso en ve	TetuÃ¡n, Madrid				TRUE	FALSE	TRUE	TRUE
18910	Estudio en	Cuatro Car	Calle Orden	Calle Orden		TRUE	TRUE	FALSE	FALSE
18771	Estudio en	Cuatro Car	Calle Orden, 6	Calle Orden	6	TRUE	TRUE	TRUE	TRUE
18730	Piso en ve	Bellas Vist	Calle de Bravo M	Calle de Bravo Murillo		TRUE	FALSE	TRUE	TRUE
18676	Piso en ve	Cuzco-Cas	Calle de la Infan	Calle de la Infanta Mercedes, Mad		TRUE	FALSE	TRUE	TRUE
18512	DÃºplex en	Cuzco-Cas	Calle Pensamier	Calle Pensamie	27	TRUE	FALSE	TRUE	TRUE
18426	Estudio en	TetuÃ¡n, Madrid				TRUE	TRUE	TRUE	TRUE
18366	Piso en ve	Cuatro Car	Calle del Aviado	Calle del Aviadc	32	TRUE	FALSE	TRUE	TRUE
18306	Piso en ve	TetuÃ¡n, Madrid				TRUE	FALSE	FALSE	FALSE
18260	Piso en ve	Cuzco-Cas	Calle Rosario Pir	Calle Rosario Pi	18	TRUE	FALSE	TRUE	TRUE
18251	Piso en ve	Cuzco-Cas	Calle de Rosario	Calle de Rosaric	18	TRUE	FALSE	TRUE	TRUE
18250	Piso en ve	Cuzco-Cas	Calle de Rosario	Calle de Rosaric	18	TRUE	FALSE	TRUE	TRUE
18131	Piso en ve	Bellas Vist	Wad Ras, 10	Wad Ras	10	TRUE	FALSE	TRUE	TRUE
18076	Piso en ve	Bellas Vist	Calle de Wad-Ra	Calle de Wad-R	10	TRUE	FALSE	TRUE	TRUE
18075	Piso en ve	Bellas Vist	Calle de Wad-Ra	Calle de Wad-R	10	TRUE	FALSE	TRUE	TRUE
18036	Estudio en	TetuÃ¡n, Madrid				TRUE	TRUE	FALSE	FALSE
18030	Piso en ve	Cuatro Car	Calle del Aviado	Calle del Aviador Zorita		TRUE	FALSE	TRUE	TRUE
17929	Piso en ve	Cuzco-Cas	Calle de Bravo M	Calle de Bravo Murillo		TRUE	FALSE	FALSE	FALSE
17906	Piso en ve	Retiro, Madrid				TRUE	FALSE	FALSE	FALSE
17837	Piso en ve	Adelfas, Ma	Avenida de la Ci	Avenida de la Ciudad de Barcelona		TRUE	FALSE	TRUE	TRUE
17797	Piso en ve	PacÃ-fico	Calle del Doctor	Calle del Doctor Esquerdo		TRUE	FALSE	TRUE	TRUE

Jelentések

id	title	subtitle	raw_address	street_name	street_number	expectation_failed	house_type_id_length_between	has_central_heating_not_null	has_individual_heating_not_null
21216	Piso en ve	Los Ãngele	Calle de la del M	Calle de la del M	65 -69	TRUE	FALSE	TRUE	TRUE
19700	Piso en ve	MoscardÃ³	Calle de Antonio	Calle de Antonio LÃ³pez		TRUE	FALSE	TRUE	TRUE
19474	Piso en ve	MoscardÃ³	Calle de Antonio	Calle de Antonic	115	TRUE	FALSE	FALSE	FALSE
19235	Piso en ve	TetuÃ¡n, Madrid				TRUE	FALSE	TRUE	TRUE
18986	Piso en ve	TetuÃ¡n, Madrid				TRUE	FALSE	FALSE	FALSE
18926	Piso en ve	TetuÃ¡n, Madrid				TRUE	FALSE	TRUE	TRUE
18910	Estudio en	Cuatro Car	Calle Orden	Calle Orden		TRUE	TRUE	FALSE	FALSE
18771	Estudio en	Cuatro Car	Calle Orden, 6	Calle Orden	6	TRUE	TRUE	TRUE	TRUE
18730	Piso en ve	Bellas Vist	Calle de Bravo M	Calle de Bravo Murillo		TRUE	FALSE	TRUE	TRUE
18676	Piso en ve	Cuzco-Cas	Calle de la Infan	Calle de la Infanta Mercedes, Mad		TRUE	FALSE	TRUE	TRUE
18512	DÃºplex en	Cuzco-Cas	Calle Pensamier	Calle Pensamie	27	TRUE	FALSE	TRUE	TRUE
18426	Estudio en	TetuÃ¡n, Madrid				TRUE	TRUE	TRUE	TRUE
18366	Piso en ve	Cuatro Car	Calle del Aviado	Calle del Aviadc	32	TRUE	FALSE	TRUE	TRUE
18306	Piso en ve	TetuÃ¡n, Madrid				TRUE	FALSE	FALSE	FALSE
18260	Piso en ve	Cuzco-Cas	Calle Rosario Pir	Calle Rosario Pi	18	TRUE	FALSE	TRUE	TRUE
18251	Piso en ve	Cuzco-Cas	Calle de Rosario	Calle de Rosaric	18	TRUE	FALSE	TRUE	TRUE
18250	Piso en ve	Cuzco-Cas	Calle de Rosario	Calle de Rosaric	18	TRUE	FALSE	TRUE	TRUE
18131	Piso en ve	Bellas Vist	Wad Ras, 10	Wad Ras	10	TRUE	FALSE	TRUE	TRUE
18076	Piso en ve	Bellas Vist	Calle de Wad-Ra	Calle de Wad-R	10	TRUE	FALSE	TRUE	TRUE
18075	Piso en ve	Bellas Vist	Calle de Wad-Ra	Calle de Wad-R	10	TRUE	FALSE	TRUE	TRUE
18036	Estudio en	TetuÃ¡n, Madrid				TRUE	TRUE	FALSE	FALSE
18030	Piso en ve	Cuatro Car	Calle del Aviado	Calle del Aviador Zorita		TRUE	FALSE	TRUE	TRUE
17929	Piso en ve	Cuzco-Cas	Calle de Bravo M	Calle de Bravo Murillo		TRUE	FALSE	FALSE	FALSE
17906	Piso en ve	Retiro, Madrid				TRUE	FALSE	FALSE	FALSE
17837	Piso en ve	Adelfas, Ma	Avenida de la Ci	Avenida de la Ciudad de Barcelona		TRUE	FALSE	TRUE	TRUE
17797	Piso en ve	PacÃfico	Calle del Doctor	Calle del Doctor Esquerdo		TRUE	FALSE	TRUE	TRUE

Eredmények és jövőbeli irányok

Eredmények

- Adatminőség értékelő eszköz
  - Általános
  - Minimális adatelőkészítés
  - Gyors minőség ellenőrzés
  - Jelentések

Fejlesztési irányok

- Specifikusabb algoritmusok
  - Felhasználási területenként
  - Adatforrás típusonként
- Oszlopok értékeinek együttes vizsgálata
- Eszközök használatának kiterjesztése
  - ydata quality
- AI alapú osztályozás