Looking for CSV files in: C:\Users\Marcin\Documents\Uczelnia PG\Uczenie maszynowe\Projekt\v5\Data

Found 11 CSV files:

- apartments\_pl\_2023\_08.csv

- apartments\_pl\_2023\_09.csv

- apartments\_pl\_2023\_10.csv

- apartments\_pl\_2023\_11.csv

- apartments\_pl\_2023\_12.csv

- apartments\_pl\_2024\_01.csv

- apartments\_pl\_2024\_02.csv

- apartments\_pl\_2024\_03.csv

- apartments\_pl\_2024\_04.csv

- apartments\_pl\_2024\_05.csv

- apartments\_pl\_2024\_06.csv

Combined dataframe shape: (195568, 28)

--- Basic information about the data ---

id city type squareMeters rooms floor floorCount buildYear latitude ... ownership buildingMaterial condition hasParkingSpace hasBalcony hasElevator hasSecurity hasStorageRoom price

0 f8524536d4b09a0c8ccc0197ec9d7bde szczecin blockOfFlats 63.00 3.0 4.0 10.0 1980.0 53.378933 ... condominium concreteSlab NaN yes yes yes no yes 415000

1 accbe77d4b360fea9735f138a50608dd szczecin blockOfFlats 36.00 2.0 8.0 10.0 NaN 53.442692 ... cooperative concreteSlab NaN no yes yes no yes 395995

2 8373aa373dbc3fe7ca3b7434166b8766 szczecin tenement 73.02 3.0 2.0 3.0 NaN 53.452222 ... condominium brick NaN no no no no no 565000

3 0a68cd14c44ec5140143ece75d739535 szczecin tenement 87.60 3.0 2.0 3.0 NaN 53.435100 ... condominium brick NaN yes yes no no yes 640000

4 f66320e153c2441edc0fe293b54c8aeb szczecin blockOfFlats 66.00 3.0 1.0 3.0 NaN 53.410278 ... condominium NaN NaN no no no no no 759000

[5 rows x 28 columns]

--- Column information ---

<class 'pandas.core.frame.DataFrame'>

RangeIndex: 195568 entries, 0 to 195567

Data columns (total 28 columns):

# Column Non-Null Count Dtype

--- ------ -------------- -----

0 id 195568 non-null object

1 city 195568 non-null object

2 type 153307 non-null object

3 squareMeters 195568 non-null float64

4 rooms 195568 non-null float64

5 floor 160974 non-null float64

6 floorCount 193185 non-null float64

7 buildYear 163352 non-null float64

8 latitude 195568 non-null float64

9 longitude 195568 non-null float64

10 centreDistance 195568 non-null float64

11 poiCount 195568 non-null float64

12 schoolDistance 195400 non-null float64

13 clinicDistance 194840 non-null float64

14 postOfficeDistance 195320 non-null float64

15 kindergartenDistance 195361 non-null float64

16 restaurantDistance 195089 non-null float64

17 collegeDistance 190132 non-null float64

18 pharmacyDistance 195291 non-null float64

19 ownership 195568 non-null object

20 buildingMaterial 118186 non-null object

21 condition 49261 non-null object

22 hasParkingSpace 195568 non-null object

23 hasBalcony 195568 non-null object

24 hasElevator 185866 non-null object

25 hasSecurity 195568 non-null object

26 hasStorageRoom 195568 non-null object

27 price 195568 non-null int64

dtypes: float64(16), int64(1), object(11)

memory usage: 41.8+ MB

None

--- Descriptive statistics ---

squareMeters rooms floor floorCount buildYear latitude longitude ... clinicDistance postOfficeDistance kindergartenDistance restaurantDistance collegeDistance pharmacyDistance price

count 195568.000000 195568.000000 160974.000000 193185.000000 163352.000000 195568.000000 195568.000000 ... 194840.000000 195320.000000 195361.000000 195089.000000 190132.00000 195291.000000 1.955680e+05

mean 58.697667 2.679222 3.332414 5.309113 1985.976346 52.026288 19.465989 ... 0.970287 0.516340 0.367560 0.345257 1.44327 0.358114 7.841833e+05

std 21.407206 0.915024 2.531684 3.312234 33.812810 1.335275 1.783264 ... 0.888884 0.498013 0.444673 0.463510 1.10457 0.457679 4.097092e+05

min 25.000000 1.000000 1.000000 1.000000 1850.000000 49.978999 14.447100 ... 0.001000 0.001000 0.001000 0.001000 0.00400 0.001000 1.500000e+05

25% 44.000000 2.000000 2.000000 3.000000 1967.000000 51.108796 18.523270 ... 0.356000 0.239000 0.156000 0.114000 0.57700 0.142000 5.200000e+05

50% 54.600000 3.000000 3.000000 4.000000 1994.000000 52.194596 19.899434 ... 0.676000 0.393000 0.262000 0.229000 1.12000 0.239000 6.990000e+05

75% 68.550000 3.000000 4.000000 6.000000 2016.000000 52.409006 20.989907 ... 1.237000 0.623000 0.416000 0.409000 2.05500 0.406000 9.300000e+05

max 150.000000 6.000000 29.000000 29.000000 2024.000000 54.606460 23.208873 ... 4.999000 4.970000 4.961000 4.985000 5.00000 4.992000 3.250000e+06

[8 rows x 17 columns]

--- Number of unique values ---

id 92967

longitude 49315

latitude 47484

price 8092

squareMeters 7201

collegeDistance 4825

clinicDistance 4282

postOfficeDistance 2713

schoolDistance 2537

pharmacyDistance 2446

restaurantDistance 2437

kindergartenDistance 2342

centreDistance 1489

poiCount 196

buildYear 166

floorCount 29

floor 27

city 15

rooms 6

type 3

ownership 3

buildingMaterial 2

hasParkingSpace 2

condition 2

hasBalcony 2

hasElevator 2

hasSecurity 2

hasStorageRoom 2

dtype: int64

--- Analysis of missing data ---

Missing count Missing percent

condition 146307 74.811319

buildingMaterial 77382 39.567823

type 42261 21.609363

floor 34594 17.688988

buildYear 32216 16.473043

hasElevator 9702 4.960934

collegeDistance 5436 2.779596

floorCount 2383 1.218502

clinicDistance 728 0.372249

restaurantDistance 479 0.244928

pharmacyDistance 277 0.141639

postOfficeDistance 248 0.126810

kindergartenDistance 207 0.105846

schoolDistance 168 0.085904

--- Filtering out top N% most expensive apartments per city ---

Shape after removing top 0%: 194692

Removed columns: ['id', 'buildingMaterial', 'condition', 'type', 'ownership', 'latitude', 'longitude']

--- Filling missing data ---

--- NaN check after groupwise KNNImputer ---

buildYear 0

squareMeters 0

rooms 0

poiCount 0

dtype: int64

Data processing completed.

Total missing values in dataframe:

squareMeters 0

rooms 0

floor 0

floorCount 0

buildYear 0

centreDistance 0

poiCount 0

schoolDistance 0

clinicDistance 0

postOfficeDistance 0

kindergartenDistance 0

restaurantDistance 0

collegeDistance 0

pharmacyDistance 0

price 0

city\_bialystok 0

city\_bydgoszcz 0

city\_czestochowa 0

city\_gdansk 0

city\_gdynia 0

city\_katowice 0

city\_krakow 0

city\_lodz 0

city\_lublin 0

city\_poznan 0

city\_radom 0

city\_rzeszow 0

city\_szczecin 0

city\_warszawa 0

city\_wroclaw 0

hasParkingSpace\_yes 0

hasBalcony\_yes 0

hasElevator\_yes 0

hasSecurity\_yes 0

hasStorageRoom\_yes 0

dtype: int64

--- Filtering out top N% most expensive apartments per city (LOO pipeline) ---

Data processing for XGB/LOO completed.

squareMeters 0

rooms 0

floor 0

floorCount 0

buildYear 0

centreDistance 0

poiCount 0

schoolDistance 0

clinicDistance 0

postOfficeDistance 0

kindergartenDistance 0

restaurantDistance 0

collegeDistance 0

pharmacyDistance 0

price 0

city\_loo 0

hasParkingSpace\_yes 0

hasBalcony\_yes 0

hasElevator\_yes 0

hasSecurity\_yes 0

hasStorageRoom\_yes 0

dtype: int64

Set shapes - training: (116815, 34), validation: (38938, 34), test: (38939, 34)

Set shapes - training: (116815, 20), validation: (38938, 20), test: (38939, 20)

--- Optimizing XGBoost hyperparameters ---

[I 2025-05-17 00:50:15,963] A new study created in memory with name: no-name-4187309b-7975-4343-8b96-c8182288bb1c

[I 2025-05-17 00:50:16,232] Trial 0 finished with value: 60807.55742504381 and parameters: {'n\_estimators': 85, 'max\_depth': 6, 'learning\_rate': 0.0891008440465982, 'subsample': 0.6714448912752855, 'colsample\_bytree': 0.7976369338823369, 'min\_child\_weight': 9, 'gamma': 0.7970240381633398}. Best is trial 0 with value: 60807.55742504381.

[I 2025-05-17 00:50:16,560] Trial 1 finished with value: 131952.2551228284 and parameters: {'n\_estimators': 105, 'max\_depth': 6, 'learning\_rate': 0.022026398376484343, 'subsample': 0.677922242446083, 'colsample\_bytree': 0.5826348768461345, 'min\_child\_weight': 8, 'gamma': 0.06475021121978353}. Best is trial 0 with value: 60807.55742504381.

[I 2025-05-17 00:50:18,484] Trial 2 finished with value: 28210.19049917955 and parameters: {'n\_estimators': 248, 'max\_depth': 10, 'learning\_rate': 0.09650330257626354, 'subsample': 0.8461813940655651, 'colsample\_bytree': 0.5715324111617597, 'min\_child\_weight': 3, 'gamma': 0.4084637210860691}. Best is trial 2 with value: 28210.19049917955.

[I 2025-05-17 00:50:20,710] Trial 3 finished with value: 20645.727887386292 and parameters: {'n\_estimators': 440, 'max\_depth': 9, 'learning\_rate': 0.07928899842472259, 'subsample': 0.9144159323338059, 'colsample\_bytree': 0.9381428553553359, 'min\_child\_weight': 3, 'gamma': 0.04889599814548762}. Best is trial 3 with value: 20645.727887386292.

[I 2025-05-17 00:50:21,265] Trial 4 finished with value: 31586.499394519804 and parameters: {'n\_estimators': 79, 'max\_depth': 9, 'learning\_rate': 0.21917046126232856, 'subsample': 0.9770990215902673, 'colsample\_bytree': 0.6250030637866117, 'min\_child\_weight': 5, 'gamma': 0.7705281353690161}. Best is trial 3 with value: 20645.727887386292.

[I 2025-05-17 00:50:22,046] Trial 5 finished with value: 36436.64089896323 and parameters: {'n\_estimators': 461, 'max\_depth': 4, 'learning\_rate': 0.2630837862902919, 'subsample': 0.9484407667032675, 'colsample\_bytree': 0.9446029062931063, 'min\_child\_weight': 5, 'gamma': 0.19620038563596343}. Best is trial 3 with value: 20645.727887386292.

[I 2025-05-17 00:50:23,080] Trial 6 finished with value: 31024.1495612692 and parameters: {'n\_estimators': 443, 'max\_depth': 6, 'learning\_rate': 0.2349748363182505, 'subsample': 0.8595717634812805, 'colsample\_bytree': 0.5627124499402121, 'min\_child\_weight': 4, 'gamma': 0.6076550474261102}. Best is trial 3 with value: 20645.727887386292.

[I 2025-05-17 00:50:24,386] Trial 7 finished with value: 37346.7505413791 and parameters: {'n\_estimators': 425, 'max\_depth': 7, 'learning\_rate': 0.04372396871155885, 'subsample': 0.8452100349761655, 'colsample\_bytree': 0.5900067106626353, 'min\_child\_weight': 1, 'gamma': 0.081411905913784}. Best is trial 3 with value: 20645.727887386292.

[I 2025-05-17 00:50:24,725] Trial 8 finished with value: 37549.67184943165 and parameters: {'n\_estimators': 96, 'max\_depth': 7, 'learning\_rate': 0.1876866814768199, 'subsample': 0.7387524437428674, 'colsample\_bytree': 0.8073150563567748, 'min\_child\_weight': 8, 'gamma': 0.2523057958155599}. Best is trial 3 with value: 20645.727887386292.

[I 2025-05-17 00:50:26,629] Trial 9 finished with value: 25731.438824908335 and parameters: {'n\_estimators': 329, 'max\_depth': 9, 'learning\_rate': 0.029078746501474906, 'subsample': 0.6610244108868253, 'colsample\_bytree': 0.8906030981133523, 'min\_child\_weight': 3, 'gamma': 0.009171960530676149}. Best is trial 3 with value: 20645.727887386292.

[I 2025-05-17 00:50:27,211] Trial 10 finished with value: 72549.98828394116 and parameters: {'n\_estimators': 338, 'max\_depth': 3, 'learning\_rate': 0.1404631515865609, 'subsample': 0.5065409162274742, 'colsample\_bytree': 0.9944530371073079, 'min\_child\_weight': 1, 'gamma': 0.9953235438240186}. Best is trial 3 with value: 20645.727887386292.

[I 2025-05-17 00:50:28,949] Trial 11 finished with value: 22310.110353828375 and parameters: {'n\_estimators': 341, 'max\_depth': 9, 'learning\_rate': 0.07723549473520679, 'subsample': 0.5645222925332943, 'colsample\_bytree': 0.8744719717712621, 'min\_child\_weight': 3, 'gamma': 0.008387733281618681}. Best is trial 3 with value: 20645.727887386292.

[I 2025-05-17 00:50:30,783] Trial 12 finished with value: 21495.591361951407 and parameters: {'n\_estimators': 372, 'max\_depth': 9, 'learning\_rate': 0.10533726995138934, 'subsample': 0.5880986650513522, 'colsample\_bytree': 0.875965398663957, 'min\_child\_weight': 3, 'gamma': 0.31931158551056066}. Best is trial 3 with value: 20645.727887386292.

[I 2025-05-17 00:50:32,550] Trial 13 finished with value: 25694.50057891766 and parameters: {'n\_estimators': 494, 'max\_depth': 8, 'learning\_rate': 0.13156478672300126, 'subsample': 0.5862271152170566, 'colsample\_bytree': 0.708819134546373, 'min\_child\_weight': 6, 'gamma': 0.3723406398744745}. Best is trial 3 with value: 20645.727887386292.

[I 2025-05-17 00:50:34,285] Trial 14 finished with value: 20837.48545290434 and parameters: {'n\_estimators': 234, 'max\_depth': 10, 'learning\_rate': 0.16709789978437348, 'subsample': 0.7823683822380447, 'colsample\_bytree': 0.8712068679805197, 'min\_child\_weight': 2, 'gamma': 0.24832223809284515}. Best is trial 3 with value: 20645.727887386292.

[I 2025-05-17 00:50:35,881] Trial 15 finished with value: 20547.368103968936 and parameters: {'n\_estimators': 217, 'max\_depth': 10, 'learning\_rate': 0.1712828175226464, 'subsample': 0.9003084029153451, 'colsample\_bytree': 0.9936798925250518, 'min\_child\_weight': 1, 'gamma': 0.15869961536405214}. Best is trial 15 with value: 20547.368103968936.

[I 2025-05-17 00:50:37,286] Trial 16 finished with value: 21600.09777755647 and parameters: {'n\_estimators': 179, 'max\_depth': 10, 'learning\_rate': 0.2973723619014626, 'subsample': 0.917772238887907, 'colsample\_bytree': 0.9980099709152453, 'min\_child\_weight': 1, 'gamma': 0.5111675041728856}. Best is trial 15 with value: 20547.368103968936.

[I 2025-05-17 00:50:37,977] Trial 17 finished with value: 28780.644607096623 and parameters: {'n\_estimators': 174, 'max\_depth': 8, 'learning\_rate': 0.18814739774489325, 'subsample': 0.8988185842167997, 'colsample\_bytree': 0.6935986512719589, 'min\_child\_weight': 6, 'gamma': 0.15393766803210462}. Best is trial 15 with value: 20547.368103968936.

[I 2025-05-17 00:50:39,292] Trial 18 finished with value: 25737.738206765567 and parameters: {'n\_estimators': 292, 'max\_depth': 8, 'learning\_rate': 0.06039546970341722, 'subsample': 0.9990240794484722, 'colsample\_bytree': 0.9389674432028944, 'min\_child\_weight': 2, 'gamma': 0.1661220501084665}. Best is trial 15 with value: 20547.368103968936.

[I 2025-05-17 00:50:39,750] Trial 19 finished with value: 52404.7411595554 and parameters: {'n\_estimators': 187, 'max\_depth': 5, 'learning\_rate': 0.1243322621958009, 'subsample': 0.791586774359937, 'colsample\_bytree': 0.7915933370053172, 'min\_child\_weight': 4, 'gamma': 0.4572256037076733}. Best is trial 15 with value: 20547.368103968936.

[I 2025-05-17 00:50:42,510] Trial 20 finished with value: 20015.639485162596 and parameters: {'n\_estimators': 400, 'max\_depth': 10, 'learning\_rate': 0.1634000152748696, 'subsample': 0.909221237398669, 'colsample\_bytree': 0.9470297491314039, 'min\_child\_weight': 2, 'gamma': 0.5459991219012698}. Best is trial 20 with value: 20015.639485162596.

[I 2025-05-17 00:50:45,316] Trial 21 finished with value: 20210.70924040025 and parameters: {'n\_estimators': 396, 'max\_depth': 10, 'learning\_rate': 0.16691252767131357, 'subsample': 0.9066870199167639, 'colsample\_bytree': 0.9427724658084637, 'min\_child\_weight': 2, 'gamma': 0.5703827331319076}. Best is trial 20 with value: 20015.639485162596.

[I 2025-05-17 00:50:48,312] Trial 22 finished with value: 30720.140624678137 and parameters: {'n\_estimators': 394, 'max\_depth': 10, 'learning\_rate': 0.16429016735259142, 'subsample': 0.8756355029799775, 'colsample\_bytree': 0.5036284009793041, 'min\_child\_weight': 2, 'gamma': 0.6009083821686425}. Best is trial 20 with value: 20015.639485162596.

[I 2025-05-17 00:50:50,479] Trial 23 finished with value: 21673.25577757066 and parameters: {'n\_estimators': 279, 'max\_depth': 10, 'learning\_rate': 0.20532112554717005, 'subsample': 0.9446809858776344, 'colsample\_bytree': 0.9581019023139483, 'min\_child\_weight': 1, 'gamma': 0.5817256803439155}. Best is trial 20 with value: 20015.639485162596.

[I 2025-05-17 00:50:52,077] Trial 24 finished with value: 21823.638560056846 and parameters: {'n\_estimators': 383, 'max\_depth': 8, 'learning\_rate': 0.15952224027785392, 'subsample': 0.8003341200868981, 'colsample\_bytree': 0.9068892759360587, 'min\_child\_weight': 2, 'gamma': 0.7056661591852678}. Best is trial 20 with value: 20015.639485162596.

[I 2025-05-17 00:50:53,186] Trial 25 finished with value: 27487.573919864226 and parameters: {'n\_estimators': 141, 'max\_depth': 10, 'learning\_rate': 0.2409267624581057, 'subsample': 0.8242401281387738, 'colsample\_bytree': 0.847778342298086, 'min\_child\_weight': 4, 'gamma': 0.5466144579303496}. Best is trial 20 with value: 20015.639485162596.

[I 2025-05-17 00:50:54,408] Trial 26 finished with value: 21181.100254708206 and parameters: {'n\_estimators': 225, 'max\_depth': 9, 'learning\_rate': 0.1860854545502316, 'subsample': 0.753539939388124, 'colsample\_bytree': 0.9588708043911657, 'min\_child\_weight': 2, 'gamma': 0.6803189891604184}. Best is trial 20 with value: 20015.639485162596.

[I 2025-05-17 00:50:57,030] Trial 27 finished with value: 22941.756515140685 and parameters: {'n\_estimators': 306, 'max\_depth': 10, 'learning\_rate': 0.11417833393088926, 'subsample': 0.9747938745205422, 'colsample\_bytree': 0.8327481940806426, 'min\_child\_weight': 1, 'gamma': 0.8867100333568475}. Best is trial 20 with value: 20015.639485162596.

[I 2025-05-17 00:50:58,479] Trial 28 finished with value: 23705.052963450642 and parameters: {'n\_estimators': 486, 'max\_depth': 7, 'learning\_rate': 0.15311597781775266, 'subsample': 0.8883259705690028, 'colsample\_bytree': 0.9136674699169622, 'min\_child\_weight': 10, 'gamma': 0.4309465580037712}. Best is trial 20 with value: 20015.639485162596.

[I 2025-05-17 00:50:59,334] Trial 29 finished with value: 34026.6323928772 and parameters: {'n\_estimators': 408, 'max\_depth': 5, 'learning\_rate': 0.20823607885213377, 'subsample': 0.9451217910117066, 'colsample\_bytree': 0.7547557888877745, 'min\_child\_weight': 7, 'gamma': 0.6901893768594286}. Best is trial 20 with value: 20015.639485162596.

[I 2025-05-17 00:51:00,758] Trial 30 finished with value: 22592.364019730205 and parameters: {'n\_estimators': 362, 'max\_depth': 8, 'learning\_rate': 0.13985512759303956, 'subsample': 0.7091554644833935, 'colsample\_bytree': 0.9779980142915073, 'min\_child\_weight': 4, 'gamma': 0.34906547283374934}. Best is trial 20 with value: 20015.639485162596.

[I 2025-05-17 00:51:03,226] Trial 31 finished with value: 20588.565370127177 and parameters: {'n\_estimators': 458, 'max\_depth': 9, 'learning\_rate': 0.18399368477959363, 'subsample': 0.9122697967899961, 'colsample\_bytree': 0.9134215940903513, 'min\_child\_weight': 2, 'gamma': 0.09512312809832067}. Best is trial 20 with value: 20015.639485162596.

[I 2025-05-17 00:51:03,647] Trial 32 finished with value: 27294.725644343816 and parameters: {'n\_estimators': 51, 'max\_depth': 9, 'learning\_rate': 0.17704570482271537, 'subsample': 0.9179068247763696, 'colsample\_bytree': 0.9185416998601084, 'min\_child\_weight': 2, 'gamma': 0.48096127451715903}. Best is trial 20 with value: 20015.639485162596.

[I 2025-05-17 00:51:06,974] Trial 33 finished with value: 20197.041367487465 and parameters: {'n\_estimators': 469, 'max\_depth': 10, 'learning\_rate': 0.20364782044751772, 'subsample': 0.8238234021757729, 'colsample\_bytree': 0.9662248312514075, 'min\_child\_weight': 1, 'gamma': 0.10422330767362287}. Best is trial 20 with value: 20015.639485162596.

[I 2025-05-17 00:51:08,952] Trial 34 finished with value: 20495.563617524647 and parameters: {'n\_estimators': 255, 'max\_depth': 10, 'learning\_rate': 0.20518960280008566, 'subsample': 0.8355140864771828, 'colsample\_bytree': 0.9755253508889165, 'min\_child\_weight': 1, 'gamma': 0.28208079544103687}. Best is trial 20 with value: 20015.639485162596.

[I 2025-05-17 00:51:11,152] Trial 35 finished with value: 20630.12554494034 and parameters: {'n\_estimators': 257, 'max\_depth': 10, 'learning\_rate': 0.23452959809812088, 'subsample': 0.8255423321234816, 'colsample\_bytree': 0.9600276348527748, 'min\_child\_weight': 3, 'gamma': 0.6470828187533131}. Best is trial 20 with value: 20015.639485162596.

[I 2025-05-17 00:51:14,429] Trial 36 finished with value: 21422.294928415115 and parameters: {'n\_estimators': 422, 'max\_depth': 10, 'learning\_rate': 0.25890564212040984, 'subsample': 0.8504610316363411, 'colsample\_bytree': 0.9357635523669875, 'min\_child\_weight': 1, 'gamma': 0.28931880915860253}. Best is trial 20 with value: 20015.639485162596.

[I 2025-05-17 00:51:16,967] Trial 37 finished with value: 20275.917537808247 and parameters: {'n\_estimators': 474, 'max\_depth': 9, 'learning\_rate': 0.2042442027966226, 'subsample': 0.8164347079701365, 'colsample\_bytree': 0.9704339484802268, 'min\_child\_weight': 2, 'gamma': 0.7701976023453712}. Best is trial 20 with value: 20015.639485162596.

[I 2025-05-17 00:51:19,296] Trial 38 finished with value: 24995.016463287237 and parameters: {'n\_estimators': 469, 'max\_depth': 9, 'learning\_rate': 0.25671354972854277, 'subsample': 0.8648209373097402, 'colsample\_bytree': 0.8496651162550901, 'min\_child\_weight': 3, 'gamma': 0.7752013420362234}. Best is trial 20 with value: 20015.639485162596.

[I 2025-05-17 00:51:21,447] Trial 39 finished with value: 24627.48643284572 and parameters: {'n\_estimators': 446, 'max\_depth': 9, 'learning\_rate': 0.2231277481307647, 'subsample': 0.7559684940937076, 'colsample\_bytree': 0.8124886718941835, 'min\_child\_weight': 5, 'gamma': 0.84117901611953}. Best is trial 20 with value: 20015.639485162596.

[I 2025-05-17 00:51:23,040] Trial 40 finished with value: 21825.55346377269 and parameters: {'n\_estimators': 418, 'max\_depth': 8, 'learning\_rate': 0.28140813086844074, 'subsample': 0.8024380275401701, 'colsample\_bytree': 0.9377460785662775, 'min\_child\_weight': 4, 'gamma': 0.9147185928905477}. Best is trial 20 with value: 20015.639485162596.

[I 2025-05-17 00:51:26,039] Trial 41 finished with value: 20859.470367197726 and parameters: {'n\_estimators': 477, 'max\_depth': 10, 'learning\_rate': 0.20589018663195757, 'subsample': 0.8530985413108554, 'colsample\_bytree': 0.9707673227249999, 'min\_child\_weight': 2, 'gamma': 0.5547490117761327}. Best is trial 20 with value: 20015.639485162596.

[I 2025-05-17 00:51:28,624] Trial 42 finished with value: 20185.843752491495 and parameters: {'n\_estimators': 401, 'max\_depth': 10, 'learning\_rate': 0.198661881896049, 'subsample': 0.8169277138046518, 'colsample\_bytree': 0.9816583374558233, 'min\_child\_weight': 1, 'gamma': 0.7473954403402093}. Best is trial 20 with value: 20015.639485162596.

[I 2025-05-17 00:51:30,790] Trial 43 finished with value: 20788.729254093432 and parameters: {'n\_estimators': 439, 'max\_depth': 9, 'learning\_rate': 0.1468919049889835, 'subsample': 0.7728013406802164, 'colsample\_bytree': 0.8908579904722873, 'min\_child\_weight': 1, 'gamma': 0.7321192243091873}. Best is trial 20 with value: 20015.639485162596.

[I 2025-05-17 00:51:33,260] Trial 44 finished with value: 21089.99762920802 and parameters: {'n\_estimators': 399, 'max\_depth': 10, 'learning\_rate': 0.21693358499701837, 'subsample': 0.6973701904618838, 'colsample\_bytree': 0.9334572681968581, 'min\_child\_weight': 2, 'gamma': 0.831632244861292}. Best is trial 20 with value: 20015.639485162596.

[I 2025-05-17 00:51:34,897] Trial 45 finished with value: 20456.373090066576 and parameters: {'n\_estimators': 348, 'max\_depth': 9, 'learning\_rate': 0.19478133897673697, 'subsample': 0.8151098669461756, 'colsample\_bytree': 0.9996033787609927, 'min\_child\_weight': 3, 'gamma': 0.6343858516926699}. Best is trial 20 with value: 20015.639485162596.

[I 2025-05-17 00:51:37,943] Trial 46 finished with value: 26638.720690003116 and parameters: {'n\_estimators': 499, 'max\_depth': 10, 'learning\_rate': 0.22711876247688517, 'subsample': 0.962392517607556, 'colsample\_bytree': 0.641119316081874, 'min\_child\_weight': 3, 'gamma': 0.7588488011608662}. Best is trial 20 with value: 20015.639485162596.

[I 2025-05-17 00:51:38,601] Trial 47 finished with value: 58363.80467378733 and parameters: {'n\_estimators': 436, 'max\_depth': 3, 'learning\_rate': 0.1959650440312709, 'subsample': 0.730256497411481, 'colsample\_bytree': 0.8909170529597682, 'min\_child\_weight': 1, 'gamma': 0.9238363368145577}. Best is trial 20 with value: 20015.639485162596.

[I 2025-05-17 00:51:40,161] Trial 48 finished with value: 20438.484483933735 and parameters: {'n\_estimators': 322, 'max\_depth': 9, 'learning\_rate': 0.24771010052682163, 'subsample': 0.9338441948893887, 'colsample\_bytree': 0.9760183576201797, 'min\_child\_weight': 2, 'gamma': 0.8220581962177548}. Best is trial 20 with value: 20015.639485162596.

[I 2025-05-17 00:51:41,665] Trial 49 finished with value: 20981.9991421218 and parameters: {'n\_estimators': 364, 'max\_depth': 9, 'learning\_rate': 0.17393692507656822, 'subsample': 0.8860551501648848, 'colsample\_bytree': 0.9570889556900096, 'min\_child\_weight': 9, 'gamma': 0.5220867850424761}. Best is trial 20 with value: 20015.639485162596.

Best parameters: {'n\_estimators': 400, 'max\_depth': 10, 'learning\_rate': 0.1634000152748696, 'subsample': 0.909221237398669, 'colsample\_bytree': 0.9470297491314039, 'min\_child\_weight': 2, 'gamma': 0.5459991219012698}

Best RMSE value: 20015.64

['squareMeters', 'rooms', 'floor', 'floorCount', 'buildYear', 'centreDistance', 'poiCount', 'schoolDistance', 'clinicDistance', 'postOfficeDistance', 'kindergartenDistance', 'restaurantDistance', 'collegeDistance', 'pharmacyDistance', 'city\_loo', 'hasParkingSpace\_yes', 'hasBalcony\_yes', 'hasElevator\_yes', 'hasSecurity\_yes', 'hasStorageRoom\_yes']

--- Evaluating the best model on the test set ---

RMSE: 20770.22

R²: 0.9973

RandomForest model saved as rf\_apartment\_price\_model.pkl

XGB model saved as best\_apartment\_price\_model.pkl

• XGB pipeline

--- Error Analysis ---

MAE: 8986.31

RMSE: 20770.22

MedianAE: 3986.06

MAPE: 1.36

Error < 50k PLN (%): 97.58

Error < 100k PLN (%): 99.34

• Random Forest pipeline

--- Error Analysis ---

MAE: 41739.48

RMSE: 79596.62

MedianAE: 18080.00

MAPE: 5.58

Error < 50k PLN (%): 75.19

Error < 100k PLN (%): 89.11

--- Model Interpretation with SHAP ---

Top 10 most important features:

1. city\_loo: 205039.5469

2. squareMeters: 119954.7266

3. buildYear: 14578.3174

4. rooms: 6863.6738

5. hasElevator\_yes: 6559.0591

6. poiCount: 4658.8394

7. centreDistance: 3834.4895

8. restaurantDistance: 2806.8057

9. clinicDistance: 2473.7742

10. floorCount: 1635.0515

(RF pipeline, one-hot) Example price prediction: 1004922.68 PLN

(XGB pipeline, LOO encoding) Example price prediction: 1006511.56 PLN