Part I: Cleaning Data

Data Cleaning Goals

- Eliminating missing values and errors
- Handling duplicate entries
- Converting data types

Cleaning Methods

- Removing or replacing missing values
- Correcting incorrect or inconsistent data
- Normalizing data

Tools

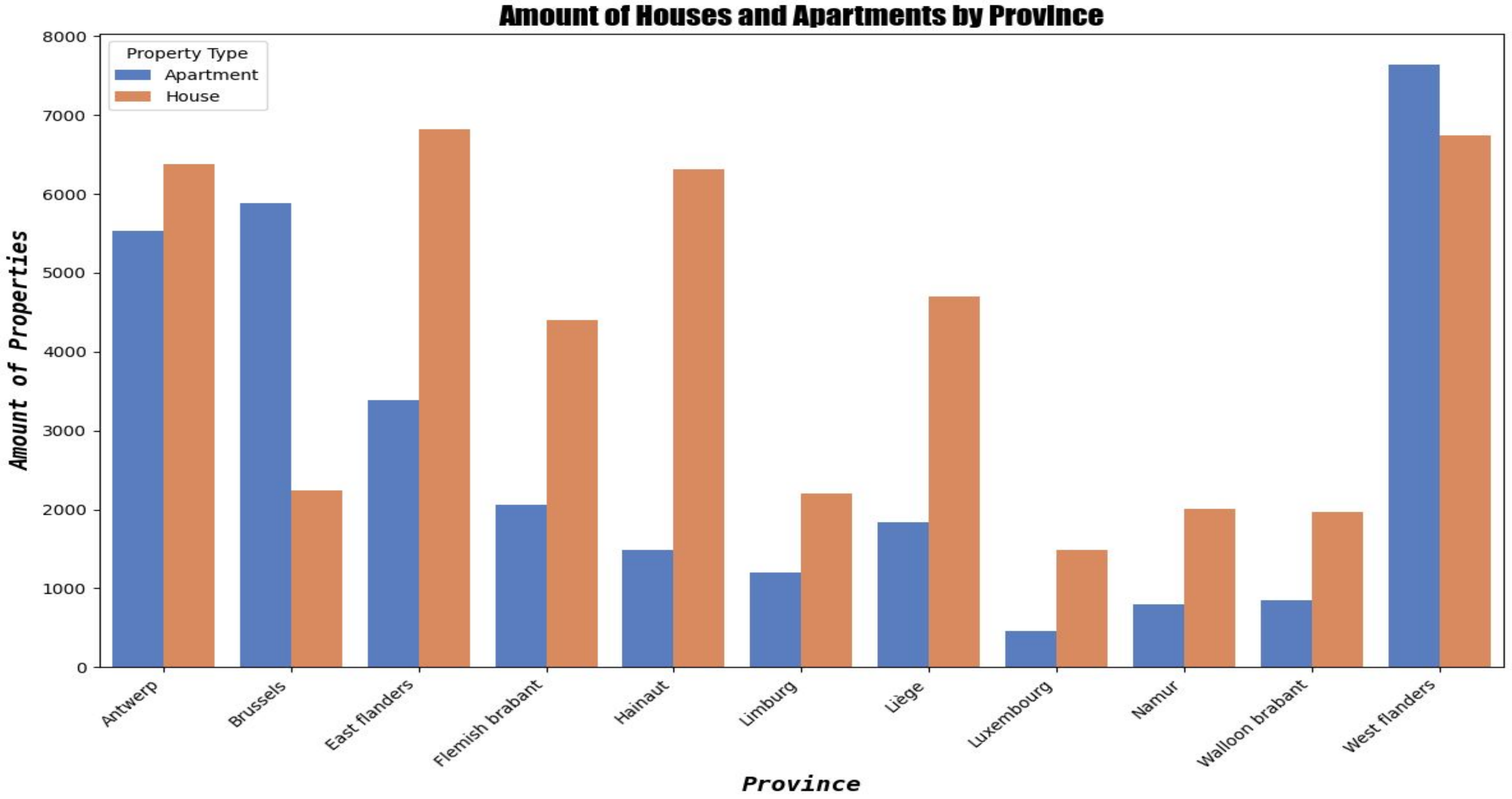
- Pandas library for data manipulation

Demo

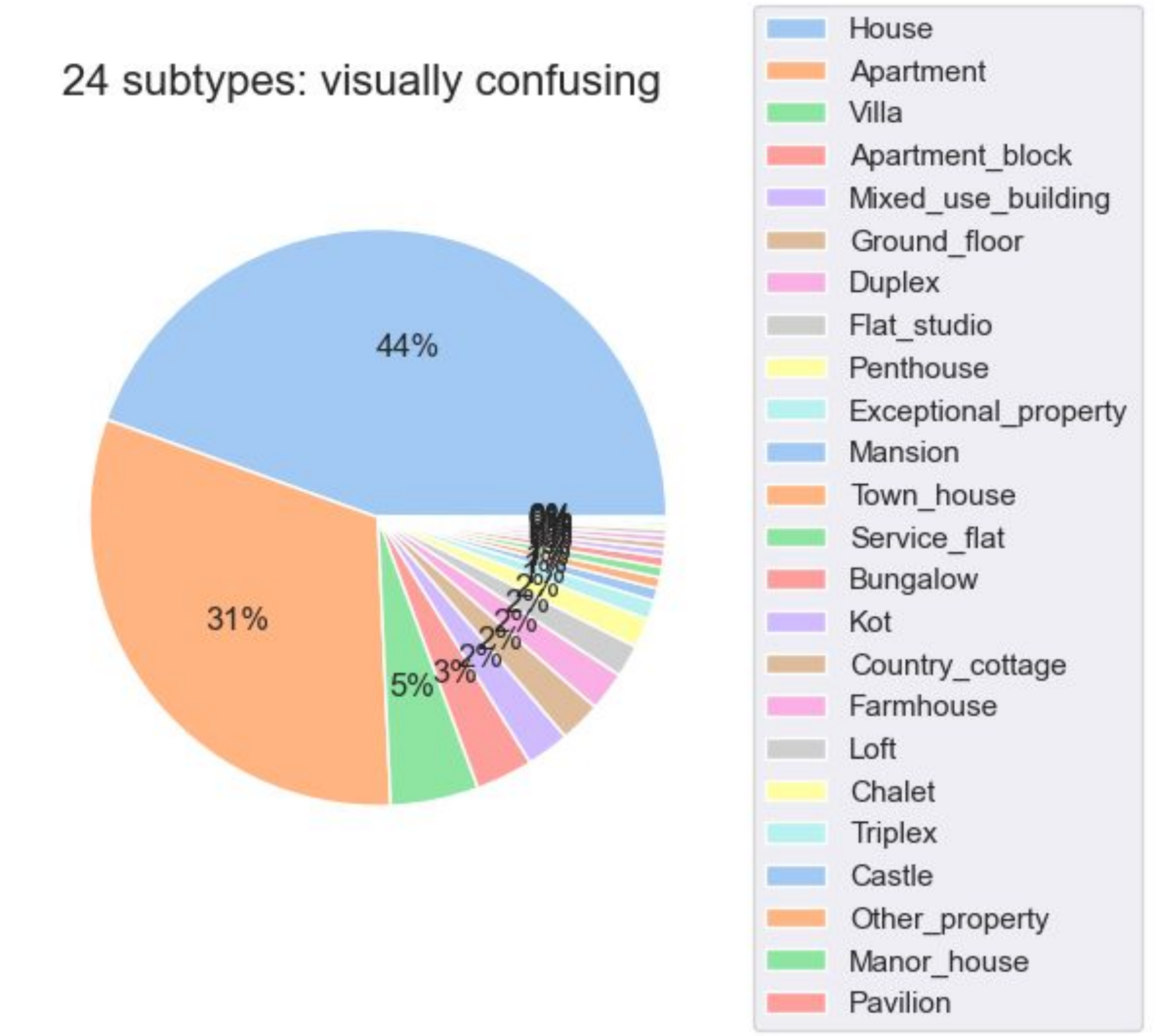
Part II: Visualisation

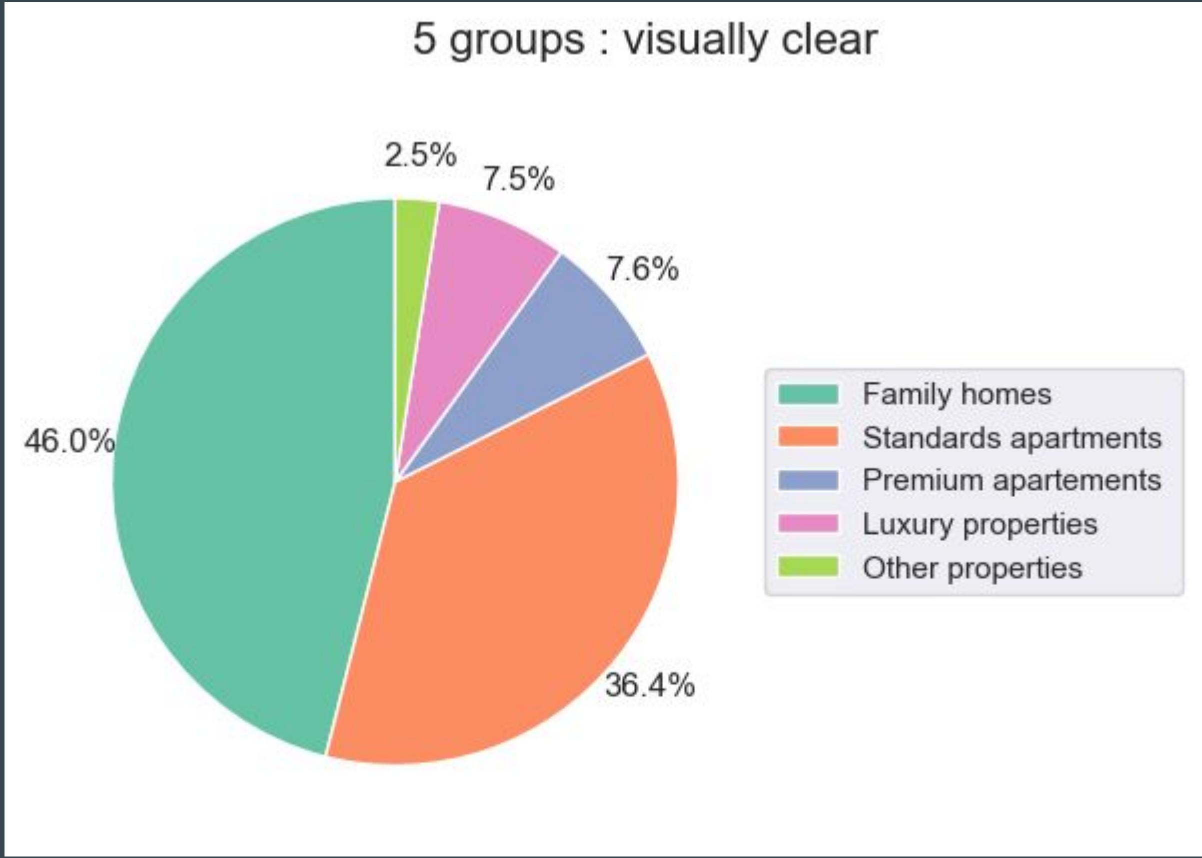
Distribution of Property Types

Distribution of Property Types: counts of different property types (Type_of_property and Subtype_of_property)

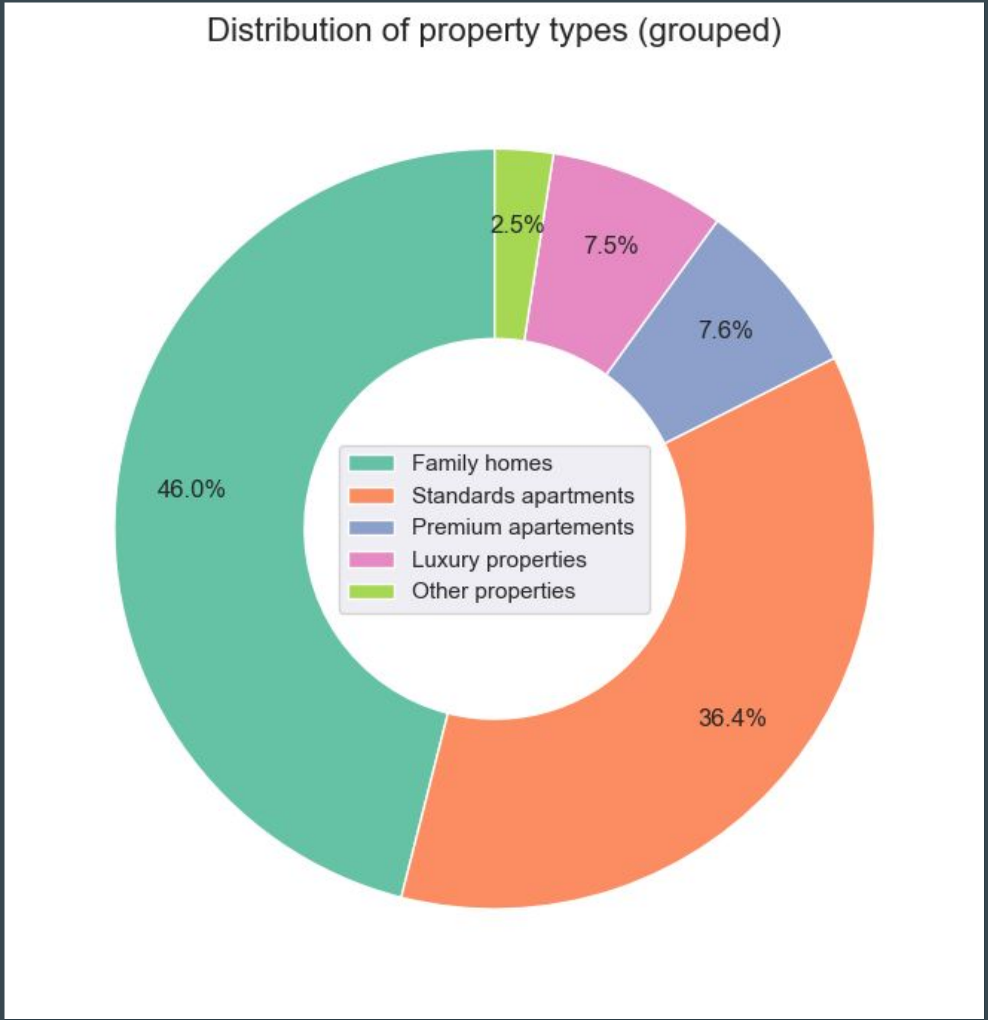


Amount of Houses and Apartments by Province
Property Type
Apartment
House
Amount of Properties
0
1000
2000
3000
4000
5000
6000
7000
8000
Antwerp
Brussels
East flanders
Flemish brabant
Hainaut
Limburg
Liège
Luxembourg
Namur
Walloon brabant
West flanders
Province


24 subtypes: visually confusing
44%
31%
5%
3%
House
Apartment
Villa
Apartment_block
Mixed_use_building
Ground_floor
Duplex
Flat_studio
Penthouse
Exceptional_property
Mansion
Town_house
Service_flat
Bungalow
Kot
Country_cottage
Farmhouse
Loft
Chalet
Triplex
Castle
Other_property
Manor_house
Pavilion

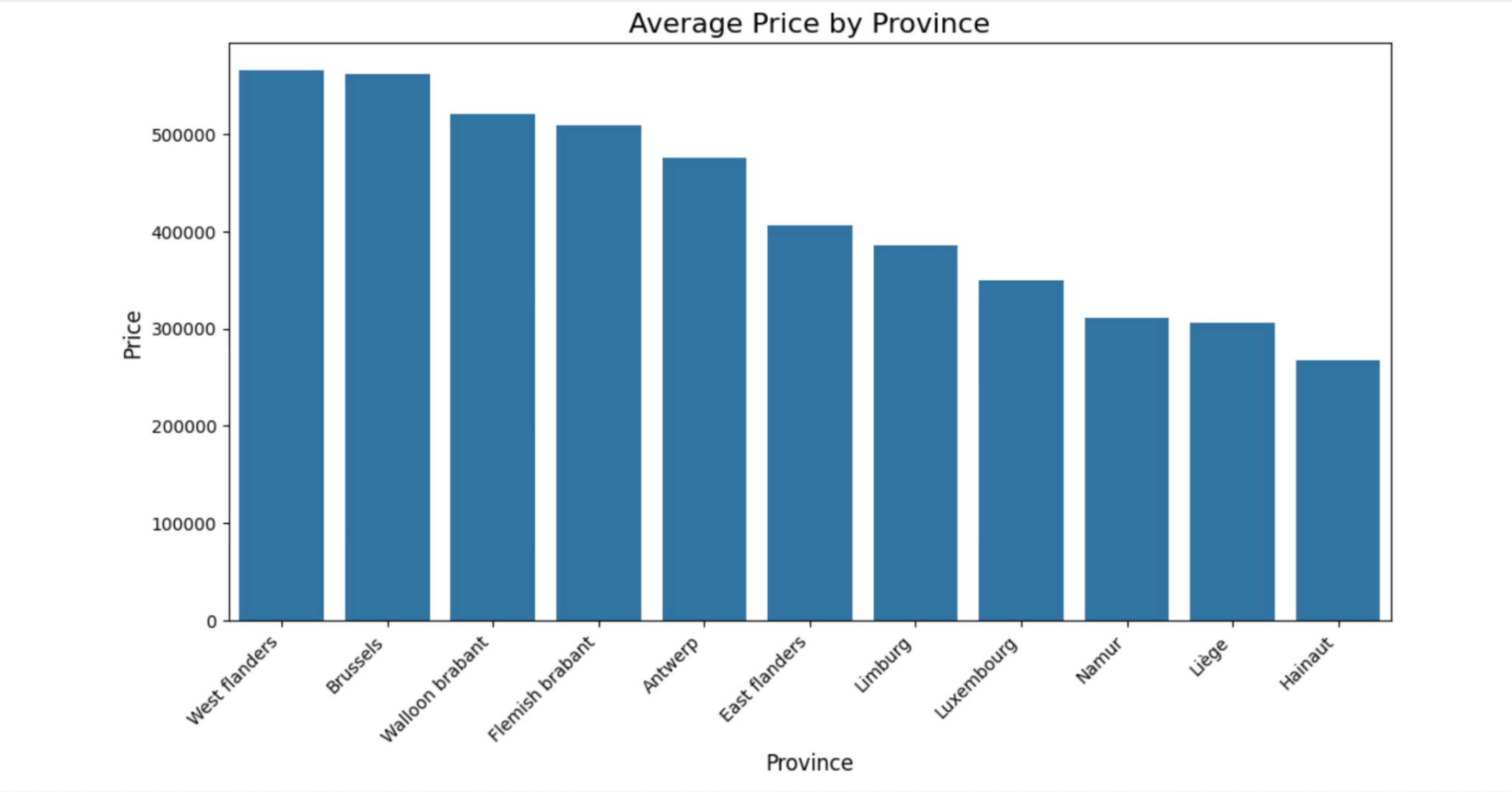


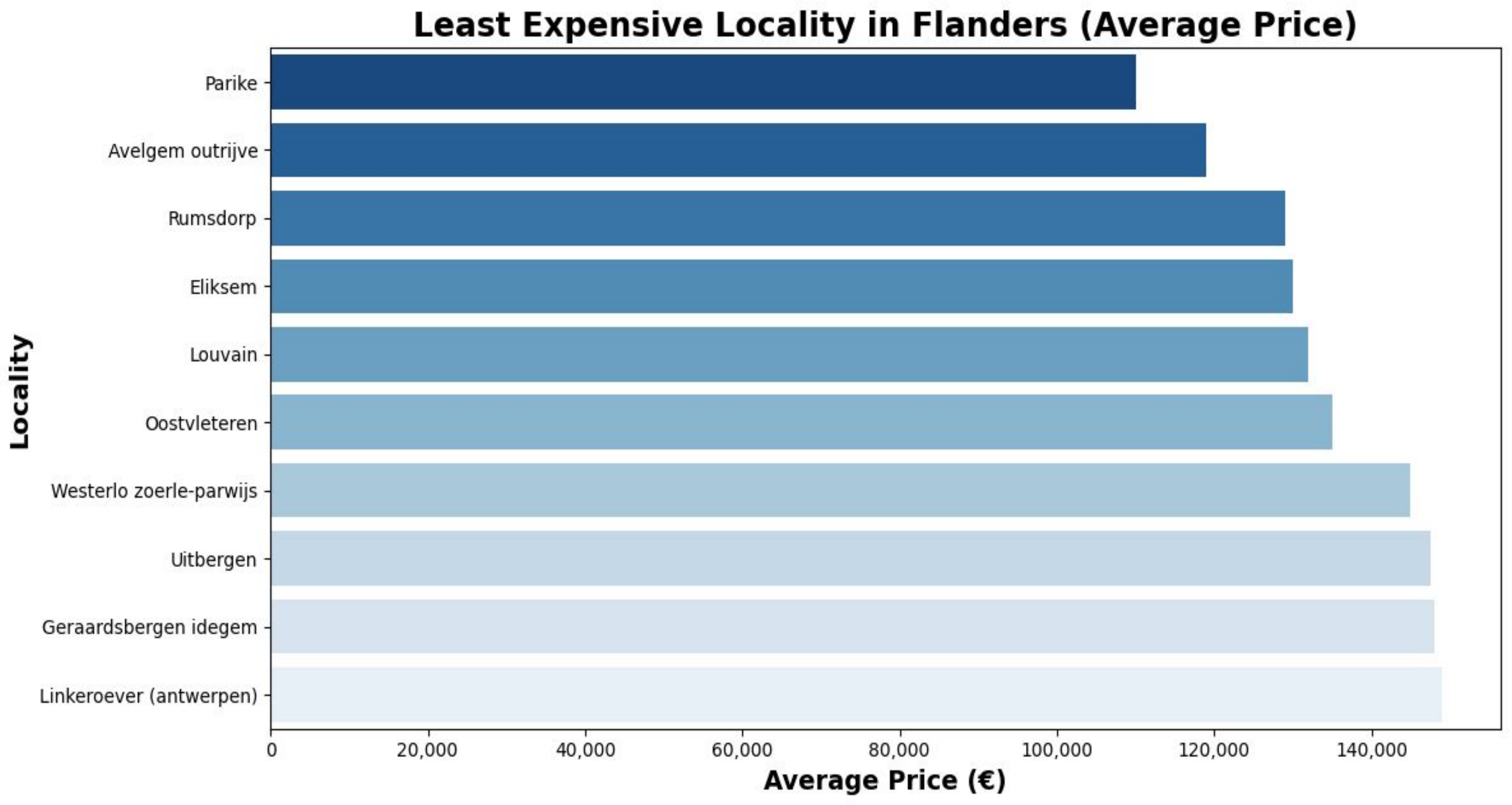
What criteria should we use ?


Distribution of property types (grouped)
2.5%
7.5%
7.6%
46.0%
36.4%
Family homes
Standards apartments
Premium apartements
Luxury properties
Other properties

Regional Insights / Price Analysis

- **Price vs. Region** : property prices across different regions or provinces.
- **Features by Region** : how features like Number_of_rooms, Net_habitable_surface, or Garden_surface vary by region.
- **Price by Region/Province** : compare prices across different regions or provinces.


Average Price by Province
Price
500000
400000
300000
200000
100000
0
West flanders
Brussels
Walloon brabant
Flemish brabant
Antwerp
East flanders
Limburg
Luxembourg
Namur
Liège
Hainaut
Province



Least Expensive Locality in Flanders (Average Price)
Locality
Parike
Avelgem outrijve
Rumsdorp
Eliksem
Louvain
Oostvleteren
Westerlo zoerle-parwijs
Uitbergen
Geraardsbergen idegem
Linkeroever (antwerpen)
0
20,000
40,000
60,000
80,000
100,000
120,000
140,000
Average Price (€)

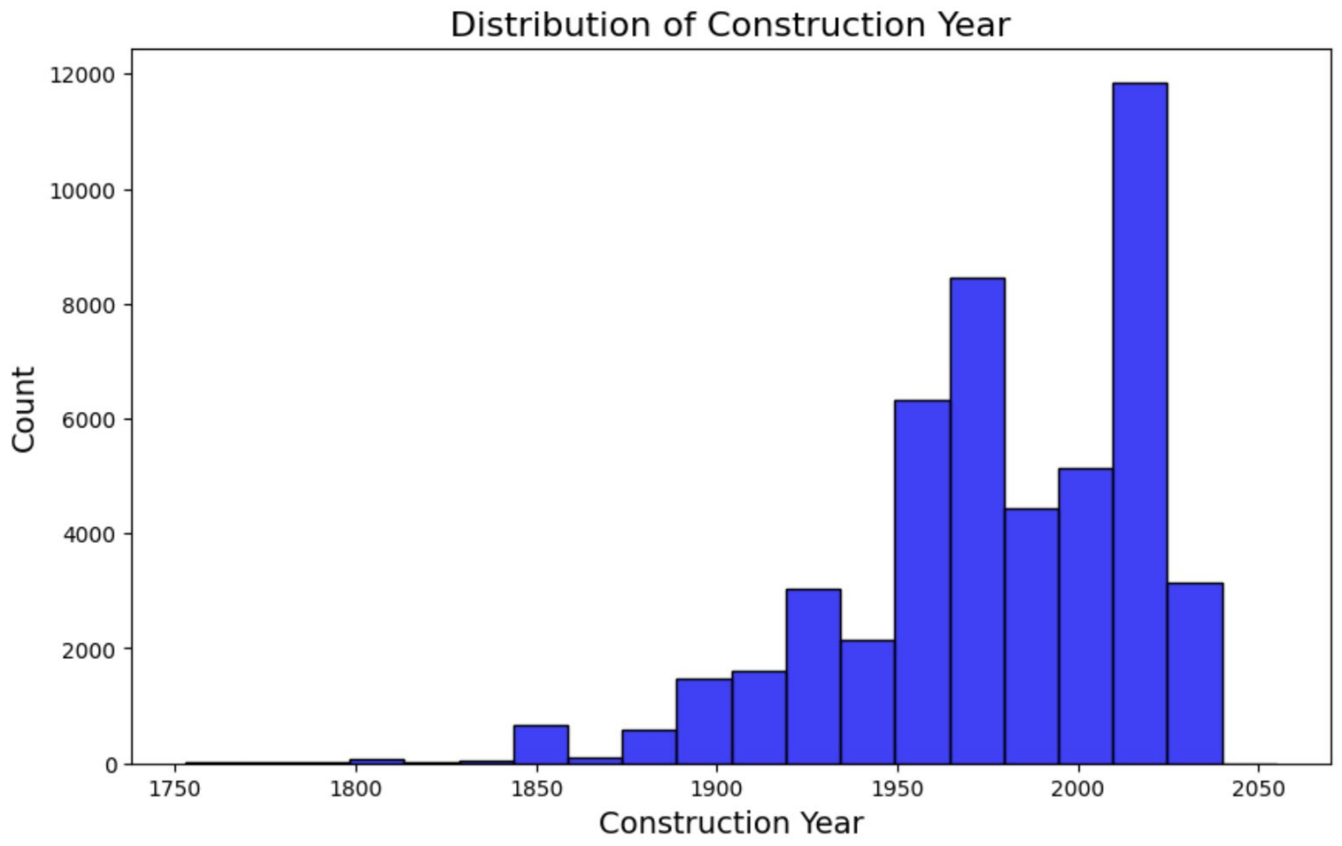

Most Expensive Locality in Wallonia (Price Per M2)
Montignies s/s
Taviers
Ophain
Rendeux (beffe)
Lasne-chapelle-saint-lambert
Tintigny lahage
Louvain-la-neuve
Eghezée (waret-la-chaussée)
Bouge (namur)
Arlon (arlon)
Court-st-etienne
Genappe (bousval)
Limerlé
Ophain-seigneur-bois-isaac
Ottignies-louvain-la-neuve céroux-mousty
Virginal-samme (ittre)
Raeren (hauset)
Liège (4000)
Braine-le-château (braine-le-château)
Ottignies-louvain-la-neuve
Locality
0
5,000
10,000
15,000
20,000
25,000
30,000
35,000
Price Per M2 (€)

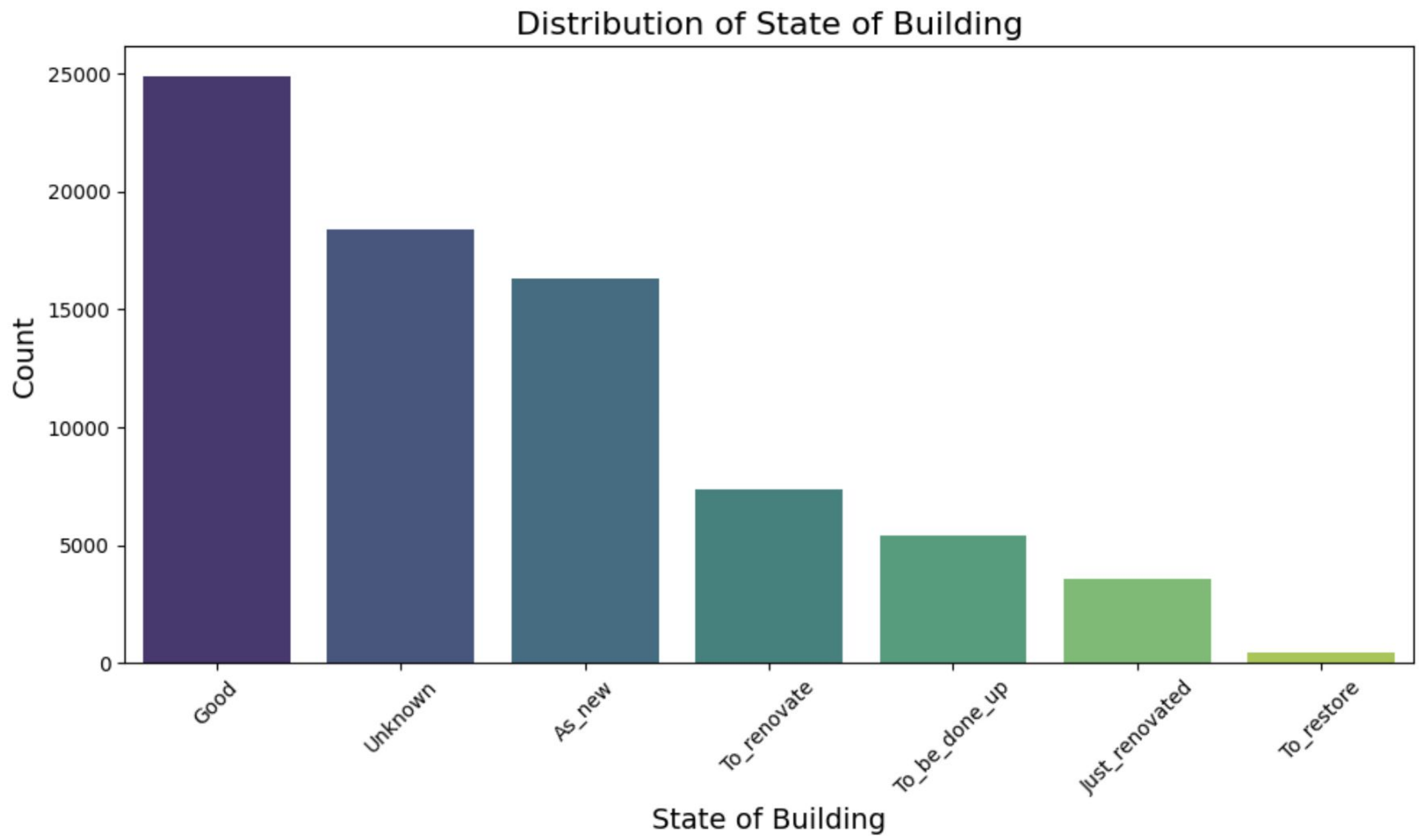

Most Expensive Locality in Brussels (Median Price)
Woluwe-saint-lambert - sint-lambrechts-woluwe
Woluwe-saint-pierre - sint-pieters-woluwe
Haren (brussel)
Bruxelles ville
St-pieters-woluwe
Etterbeek (etterbeek)
Elsene
Woluwe-saint-pierre
Oudergem
Watermael-boitsfort - métro demey
Locality
0
200,000
400,000
600,000
800,000
1,000,000
1,200,000
1,400,000
Median Price (€)

Construction and Building State

- **Construction Year** : distribution of Construction_Year to understand the age of properties.

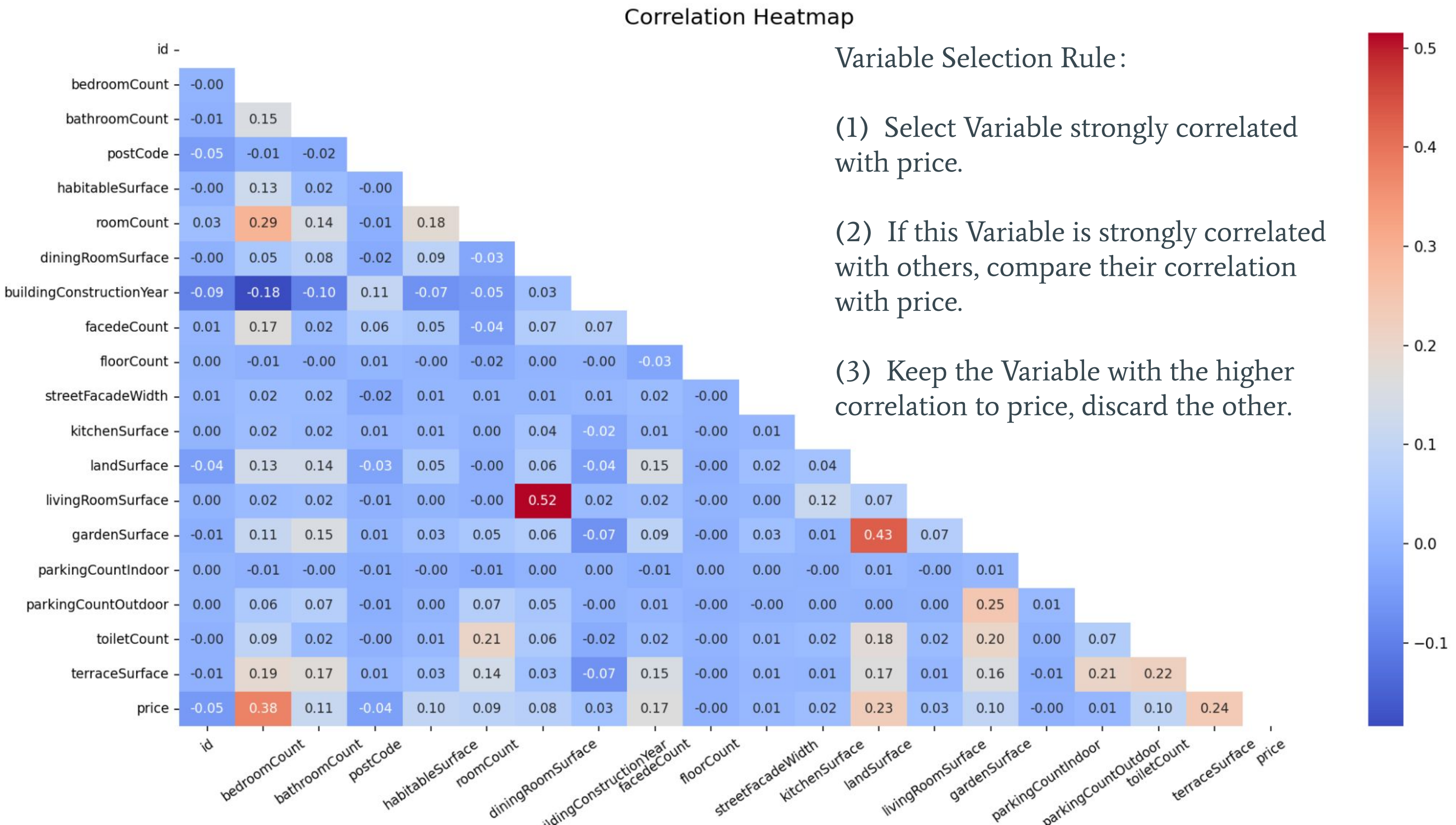
- **State of Building** : distribution of State_of_building (e.g., new, good, to renovate)


Distribution of Construction Year
Count
Construction Year
0
2000
4000
6000
8000
10000
12000
1750
1800
1850
1900
1950
2000
2050


Distribution of State of Building
Count
25000
20000
15000
10000
5000
0
Good
Unknown
As_new
To_renovate
To_be_done_up
Just_renovated
To_restore
State of Building

Correlation Analysis

- **Correlation Heatmap** : Analyze correlations between numerical variables



Variable Selection Rule :

(1) Select Variable strongly correlated with price.

(2) If this Variable is strongly correlated with others, compare their correlation with price.

(3) Keep the Variable with the higher correlation to price, discard the other.

Setting threshold = 0.1 for (1), 0.5 for (2)

Results:
bedroomCount: 0.38
terraceSurface: 0.24
landSurface: 0.23
facedeCount: 0.17
bathroomCount: 0.11

Challenges

- **Missing Values** :

	NaN_ratio_%
hasAirConditioning	98.50%
hasSwimmingPool	97.60%
hasDressingRoom	96.60%
hasFireplace	96.00%
hasThermicPanels	95.90%
hasArmoredDoor	95.20%
gardenOrientation	92.70%
diningRoomSurface	91.00%
hasHeatPump	90.20%
hasPhotovoltaicPanels	89.50%
hasOffice	86.40%
terraceOrientation	85.30%
hasAttic	83.60%
hasDiningRoom	81.50%
streetFacadeWidth	79.70%
gardenSurface	79.10%
hasGarden	79.10%
hasVisiophone	79.10%
parkingCountOutdoor	76.30%
hasLift	75.10%
roomCount	71.30%
kitchenSurface	68.20%
parkingCountIndoor	63.60%
terraceSurface	62.60%
livingRoomSurface	62.10%
hasBasement	61.60%
floorCount	50.80%
landSurface	48.20%
kitchenType	45.10%
hasLivingRoom	43.90%
floodZoneType	43.80%
heatingType	38.30%
hasTerrace	37.90%
buildingConstructionYear	35.70%
facedeCount	30.30%
toiletCount	27.90%
buildingCondition	24.10%
epcScore	15.70%
bathroomCount	12.70%
habitableSurface	11.30%
bedroomCount	3.70%
price	0.00%

- **Category Variables** :
  - 'type'
  - 'subtype'
  - 'province'
  - 'locality'
  - 'postCode'
  - 'buildingCondition'
  - 'floodZoneType'
  - 'heatingType'
  - 'kitchenType'
  - 'gardenOrientation'
  - 'terraceOrientation'
  - 'epcScore'

Correlation Analysis - new attempt

- **New attempt:**
  Missing value:
  hasXXX (e.g. ‘hasBasement’) True/NaN → 1/0

  One-hot encoding:
  'type', 'subtype', 'province', 'heatingType', 'kitchenType', 'gardenOrientation', 'terraceOrientation', 'region'

  Target encoding:
  'postCode'

  Label encoding:
  'floodZoneType', 'buildingCondition', 'epcScore’

- **Possible solution:**
  More ‘changes’ on data: group, imputation, drop outliers...
  Use other method to select features.

Setting threshold = 0.1 for price-variable, 0.3 for variable-variable

Results:
postCode_target_encoding: 0.5
bedroomCount: 0.38
hasSwimmingPool: 0.26
subtype_Villa: 0.26
terraceSurface: 0.24
landSurface: 0.23
kitchenType_Hyper_equipped: 0.19
facedeCount: 0.17
subtype_Exceptional_property: 0.15
hasOffice: 0.17
bathroomCount: 0.11
buildingCondition: -0.11
hasArmoredDoor: 0.1
hasVisiophone: 0.1
hasFireplace: 0.1

Thank you