Data Source (Step 5)

- **Dataset:** Real Estate Listings Berlin (DE) April 2023

- **Source**

  This is an external dataset; it is publicly available on Kaggle. The dataset was scraped from immowelt.de. At times, there are fraudulent listings on immowelt.de, but the amount is minor. Therefore, the data is mostly trustworthy.

- **Collection**

  The dataset contains real estate listings on immowelt.de for Berlin in April 2023; The craping process could introduce missing or inconsistent information, inaccuracies or outliers that could affect the results' reliability and accuracy.

- **Content**

  This data includes 1 dataset: ***real_estate_listings_clean.csv***. The columns are price, area, energy source, heating type, number of rooms, zip code, construction year and floor level.

Why I chose this dataset:

I have been living in Berlin for 8 years. At some point, I want to buy my own property instead of renting. But I have little idea of how the price varies depending on the characteristics of a property. I was happy to find this dataset, because investigating this subject matter for myself will be very fruitful and interesting.

Data Profile(Step 6-8)

Variables and Data Types:

Columns	Data Types	Data Integrity Issues	Changed/Fixed Records
energy	qualitative, time-invariant, nominal	Almost half of all rows have the value 'na'.	
heating	qualitative, time-invariant, nominal	Almost half of all rows have the value 'na'.	
price	quantitative, time-variant, discrete		
area	qualitative, time-invariant, continuous		
rooms	qualitative, time-invariant, discrete	Row 3223 and 3530 have unreasonable room and price.	Removed both rows.
fee	quantitative, time-variant, continuous	Row 346, 2098 and 2181 have fee outside of reasonable range.	Replaced values in these three rows.
zipcode	qualitative, time-invariant, discrete	Row 1910, 1095 and 1528 have zipcode outside of reasonable range.	Removed row 1910, replaced values in row 1095 and 1528.
construction_year	qualitative, time-invariant, discrete	The first 25 rows (sorted by ascendent) have the year outside of reasonable range.	Replaced these values with a random year value within the realistic range.
level	qualitative, time-invariant, discrete	Row 2557 have level outside of reasonable range.	Replaced it with the average level value within the zipcode area of 12159.
price_per_area	qualitative, time-variant, continuous	Row 2962 and 1575 have the price per area outside of reasonable range and they are duplicates. There could also be more fraudulent listings.	Removed row 2962 and 1575. Keep the other rows as is but remain aware of the potential frauds.

Summary:

	price	area	rooms	fee	zipcode	construction_year	level	price_per_area
count	4.937000e+03	4937.000000	4937.000000	4937.000000	4937.000000	4937.000000	4937.000000	4937.000000
mean	5.619536e+05	84.981057	2.797650	3.685817	11945.913713	1953.703261	2.994126	6354.012572
std	5.939564e+05	58.016705	1.530276	1.579564	1316.975885	45.052797	4.830155	2472.453375
min	3.495000e+04	13.000000	1.000000	0.000000	10115.000000	1838.000000	1.000000	997.581620
25%	2.590000e+05	54.040000	2.000000	3.570000	10589.000000	1910.000000	1.000000	4538.333333
50%	3.899000e+05	72.740000	3.000000	3.570000	12161.000000	1956.000000	1.000000	5843.852267
75%	6.690000e+05	101.000000	3.000000	3.570000	13088.000000	1994.000000	3.000000	8055.555556
max	1.590000e+07	970.000000	26.000000	48.100000	14199.000000	2026.000000	24.000000	29120.879121

4937 rows, 10 columns.

Outline any limitations and ethical considerations presented by the content of your data, its source, and/or how it was collected:

The dataset was scraped from immowelt.de. Even though it was cleaned by the owner, there are still data quality issues due to the nature of scraping and the listings themselves. There are highly likely frauds, but I could only identify them based on my own experience living in Berlin for the past 8 years. There is no way to verify, because the links to the original listings are expired. The presence of potential frauds would make the data skewed, and my fraud identification could add human errors.

Questions to explore:

1. How do housing prices vary in different districts of Berlin?
2. Berlin has different types of properties – for example, ‘Altbau’ (old/historical building), ‘Neubau’ (new building), and those that were built rapidly after the World War II, with a cheap price and bad quality. Are there any price differences in correlation to the construction year?
3. Which district(s) in Berlin tend to have larger floor area with a lower price?
4. Do the recently constructed properties have a larger area or smaller, are there any correlations?