EPFL

Milestone 1: Visualizing Real Estate Data

VisualEstate Group : Élise Boyer, Charles Girardot, Gauthier Nielly

March 21, 2025

1 Problematic

The real estate market is dynamic and ever-growing, offering stability, appreciation, passive income, tax benefits, and inflation protection. However, navigating property markets—especially in unfamiliar cities—can be complex. Investors and buyers need clear insights to make informed decisions, while travelers seek neighborhoods that match their preferences in terms of design, location, space, and budget.

Our project provides data-driven real estate analysis across Berlin, London, and Madrid, three of Europe's largest and most dynamic cities, chosen for their diverse housing markets and distinctive urban characteristics. Each of these cities offers a unique real estate landscape: Berlin, with its evolving property market and rent control policies; London, known for its historic housing stock and premium real estate; and Madrid, where vibrant neighborhoods and affordable housing drive market dynamics. Together, these cities represent different facets of the European real estate ecosystem, making them ideal for comparative analysis.

By analyzing real estate data from these cities, we aim to offer valuable insights into market trends, neighborhood-level price variations, and key factors influencing property values. Through visual analytics, we uncover patterns in real estate pricing, highlight high-cost areas, and identify the most influential drivers of property values. This approach enhances decision-making for both long-term investors and short-term visitors, making property selection more transparent, efficient, and accessible.

Key questions to explore:

- How do property prices vary across different neighborhoods?
- What impact do amenities and property features have on pricing?
- Where are the luxury real estate hotspots in each city?

- What are the key drivers of property pricing, and do they vary between countries?

By leveraging market data and advanced visualizations, we empower buyers, investors, and travelers to navigate real estate markets with confidence.

2 Dataset

We will use 4 datasets coming from Kaggle. One for each city and one with House Price Index in different European countries

- **Berlin's dataset** can be found here. The dataset is well-structured and includes both numerical and categorical features, such as price, area, energy source, heating type, number of rooms, and zip code. The dataset provides a diverse range of attributes, with a sufficient number of entries for visualization and analysis (around 5000 unique values). Before proceeding with visualization, preprocessing steps like one-hot encoding for categorical variables are recommended to facilitate correlation analysis and ensure meaningful insights as well as some data cleaning for outliers. Depending on the need of the website some features could be added.

- **London's dataset** can be found here. The dataset is clean and well-structured, with both numerical and categorical features properly categorized and clearly labeled, without any typos.There are no missing values in any of the features, and the dataset provides a diverse range of attributes along with a sufficient amount of data for visualization and analysis, comprising 17 features and 1,000 data entries. Before proceeding with visualization, some preprocessing steps may be necessary, such as one-hot encoding for categorical variables to facilitate correlation analysis and ensure meaningful insights.

- **Madrid's dataset** can be found here. The dataset is well-structured and contains over 15 000 property listings with a diverse set of numerical and categorical features. Most attributes are complete, with only limited missing values in *house type 2*, which were imputed using the mode. The dataset offers good coverage for analysis, but preliminary exploration revealed potential outliers in price and surface area that require further attention. Preprocessing steps prefix cleaning in *neighborhood* variable and outlier removing are recommended to ensure reliable analysis.

- **HPI dataset** can be found here The House Price Index (HPI) is a key measure of inflation in the residential property market across Europe. It reflects price changes for various types of dwellings purchased by households, including flats, detached houses, and terraced houses, while excluding self-built properties. The HPI accounts for both the structure and land components of the properties, providing a broad perspective on housing price trends.

The dataset includes HPI values from 2005 to 2021 for European Union Member States (with some exceptions), the United Kingdom, Iceland, Norway, Switzerland, and Turkey.

While this dataset offers valuable insights into long-term housing market trends, we did not have the opportunity to conduct an in-depth analysis at this stage of the project. However, it serves as an important complement to the city-level real estate data by providing a macro-level view of price evolution across Europe.

3 Exploratory data analysis

Our exploratory data analysis for each city is available in our Git repository, under `EDA/exploratory_analysis_city-name`.

3.1 Berlin EDA

Based on the deeper analysis of the dataset, we observe that it captures a broad range of properties across Berlin, from modest apartments to high-end houses, with prices spanning from under 34 000€ to over 16M€. The average property price sits over 565 000€. The dataset includes listings from various Berlin neighborhoods, comprising around 5,000 unique entries. As highlighted by the heatmap, features such as surface area and location show stronger relationships with price, making them key variables for further modeling and analysis. ven if energy source, heating type, and construction year are not strongly correlated with price, they remain useful for visualizations aimed at identifying whether certain neighborhoods have specific energy types, heating systems, or more modern buildings.

3.2 London EDA

Based on the deepper analysis of the dataset, we have that The dataset covers a diverse range of properties, from basic to more luxurious properties with prices ranging from £400,000 to £5 million and an average price of £1.85 million. It includes data from 10 neighborhoods in North and Central London, encompassing 770 unique addresses.

A variety of features help assess property quality and desirability, including: 4 types of heating systems, 4 interior styles, 5 classification types for property views, 4 primary construction materials.

As highlighted by the correlation matrix, certain features exhibit stronger relationships with price, making them key variables for further analysis.

3.3 Madrid EDA

Based on a deeper analysis of the dataset, we observe a wide variety of residential properties in Madrid, ranging from compact apartments to large houses. Prices span from below 100 000€ up to over 11M€, with some high-end outliers identified beyond the 99th percentile. The dataset includes hundreds of properties with diverse characteristics across multiple neighborhoods.

Key categorical features include property type, secondary classification, neighborhood, and availability of amenities such as elevators and garages. Continuous features like surface area (m^2) and price exhibit meaningful variation, with outlier filtering applied to enhance data quality. One-hot encoding was used for categorical variables to enable correlation analysis. The correlation matrix highlights that property size and number of rooms are among the most predictive variables with respect to price, indicating their relevance for subsequent modeling tasks.

4 Related work

Our main goal is to share meaningful real estate insights through accessible and effective visualizations. A key tool will be maps, including heat maps and choropleth maps, to display city-specific characteristics such as price variations and neighborhood trends. Additional visualizations, such as dot maps for pinpointing property locations, word clouds to highlight key features by city, and bar charts or scatter plots for comparative analysis, will provide deeper insights. Most real estate websites lack effective visualizations and often focus on a single type of data. What sets our work apart is the integration of diverse features and insights, all presented on a single page through varied visualization techniques, providing a comprehensive and intuitive user experience without the need to search across multiple platforms.

To refine our approach and enhance our visualizations, we draw inspiration from various sources and leverage specialized visualization tools. Below are some key references:

- The Pudding – Climate Zones Visualization: Inspiration for user experience and animations, including smooth scrolling, zoom effects, and intuitive data transitions, as well as effective map representations.
- New York Times – NYC Neighborhood Map: Provides inspiration for layout, color scheme, and overall design aesthetics, to ensure a engaging visual experience.
- Airbnb – Property Listings Visualization: A primary reference for its simple yet powerful map-based price visualization, offering a highly intuitive way to explore real estate data.

Additionally, we will visualization techniques and tools from:

- Data Viz Catalogue – Visualization Techniques
- Datawrapper – Chart Gallery

These references guide our design choices, ensuring a rich and intuitive visual representation of real estate data.