BIG DATA LAB 3

Timothy Cassel and Iñigo Exposito

June 2025

Pipelines

We show the basic outline of the piepline from raw data sources, to landing zone, to formatted zone to exploitaiton zone in Figure 1.

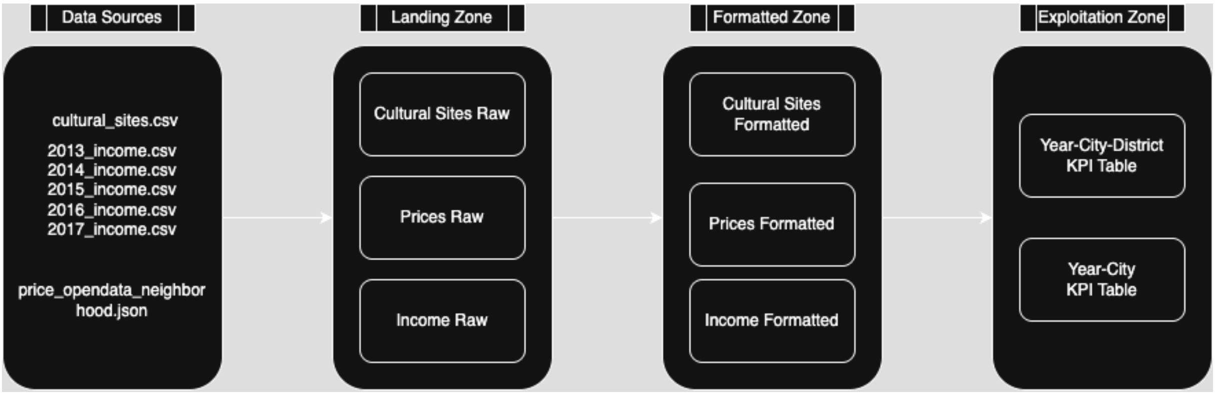

Figure 1: Pipeline

Selected datasets and analysis

To conduct our analysis, we selected three key datasets—income, cultural sites, and price open data—that collectively provide a multidimensional view of urban inequality by capturing economic conditions, cultural infrastructure, and housing market dynamics. Income data reflects economic capacity and distribution; cultural sites represent access to nearby monuments and historical places; and housing prices reveal market pressures and potential gentrification.

Across all data pipelines, we applied rigorous validation steps including schema standardization, consistent data type casting, handling of missing values, and duplicate removal. We assumed that normalized neighborhood names correspond to the same geographic boundaries across datasets, and that missing data were either random or minimally biased. We confirmed that both income and price datasets span 2013 to 2017, ensuring a consistent timeframe for comparison. The cultural sites dataset differs, as some monuments predate this period. Exploring whether the creation of cultural sites within 2013–2017 influences housing prices could provide valuable insights.

- Income Data. The income dataset was composed of five yearly CSV files ranging from 2013 to 2017. Each file was read and augmented with a year column, then all years were unioned into a single DataFrame. Column names were standardized, data types were cast appropriately, and missing values in the RFD index were imputed using the mean. Rows with placeholder labels such as *No consta* were filtered out, and duplicate records were removed. The final formatted dataset was saved as a Parquet file without partitioning.
- Cultural Sites Data. The cultural sites dataset was processed by selecting only relevant attributes such as the name of the site, district, neighborhood, geographic coordinates, and the creation date. Column names were standardized, and the dataset was filtered to ensure consistency in neighborhood names (e.g., “el Poble-sec” was normalized to “el Poble Sec”). This processing step

was essential for integrating this dataset with the income and price data. The cleaned dataset was saved as a Parquet file without partitioning.

- Price Open Data. The housing price dataset was initially provided as a nested JSON structure. We flattened this structure by exploding the info array, generating one row per year and neighborhood. Relevant variables were extracted, including total price, price per square meter, and their used counterparts. Inconsistencies in neighborhood name formatting were corrected, and duplicates were removed. The resulting dataset was validated and saved in Parquet format, partitioned by year. Missing values in `used_amount` column, were filled with zeros. This approach reflects the assumption that no transactions occurred during those years, justifying the use of zero as a placeholder.

Neighborhood Consistency. To ensure seamless integration of the datasets, we compared neighborhood names across all three sources. Differences—primarily arising from inconsistent formatting or naming conventions—were addressed through normalization. The cultural sites dataset contains 63 neighborhoods, while the others include 73. This discrepancy is expected, as some neighborhoods lack cultural sites. Such differences will be accounted for in the subsequent analysis.

In light of the preceding analysis, we selected the following KPIs.

- Cultural Sites per District. This KPI captures the number of cultural facilities available in each district. It serves as a proxy for access to cultural infrastructure and public amenities, which are important components of urban quality of life. Uneven distribution of cultural sites may indicate disparities in investment, public services, or accessibility, and can be related to broader patterns of social inequality.

- Housing Affordability Trend. Calculated as the ratio between the RFD Index (Relative Family Disposable Income) and the average housing price per square meter, this KPI reflects how affordable housing is for residents of a given area. A lower value suggests that housing costs are consuming a larger portion of income, which may indicate housing stress or exclusionary market pressures. Monitoring this trend over time is essential for identifying vulnerable neighborhoods.

- City Price Evolution. This metric measures the yearly increase in average housing prices across the city. It provides a macroeconomic view of the housing market's dynamics and helps assess whether the city is experiencing inflationary pressures, gentrification, or speculative investment trends.

- Income Inequality Progression. This KPI evaluates the income disparity across neighborhoods by computing the difference between the maximum and minimum RFD Index values for each year. Tracking its progression over time helps determine whether economic inequality is growing or shrinking.

- City Price Index. Representing the average housing price across the entire city, this index provides a benchmark for comparing local neighborhood prices. It enables the identification of areas that are significantly above or below the city average, supporting analyses of affordability, segregation, and investment hotspots.

- Neighborhood Population Growth Rate. This metric measures the rate at which each neighborhood's population is changing annually. Rapid growth may imply increasing demand and pressure on housing and infrastructure, while decline may indicate economic stagnation or reduced livability.

Given the large number of neighborhoods in the dataset, we often aggregated the data at the district level to simplify interpretation and facilitate more meaningful analysis.

At this stage of the project, we have chosen Descriptive Analysis and Dashboarding as our primary analytical approaches. In the Descriptive Analysis, we compute basic statistics and key socio-economic indicators for a variety of key performance indicators. These include the total number of sites citywide, the average number of sites per district, and the average annual growth rate of the affordability ratio. We also analyze the average price per square meter and its changes over time, the average annual growth in the Real Financial Demand (RFD), and the percentage growth per district. Additionally, we evaluate differences compared to previous years in income-related indicators, such as the inequality gap and the RFD index.

We report the citywide average housing price, which currently stands at €2,732/m^2. Further analysis includes the price index by district and year, the average price per square meter in each district, and a comparison of district prices relative to the city average. We also categorize district price levels by year and assess population growth by computing average growth rates for each district. All of these indicators are presented in a dashboard format to enable clear visualization and facilitate data-driven insights and decision-making. Plots are interpretated in the code section.