Housing Purchase and Sale Market in the City of Madrid - Scraper and Data Analysis

Context and Motivations

This project stems from the desire to develop knowledge in data analysis, both in statistical treatment and in the tools and technologies of the sector.

The housing market is a topic where narratives abound but evidence-based analyses are scarce. Data doesn't tell stories by itself, but it represents a more solid anchor than any anecdotal contribution to understanding reality.

The problem of access to housing is increasingly present in the political agenda, especially after the post-COVID inflationary period. This context makes the real estate market a relevant object of study.

Project objective: Analyze the purchase-sale housing market in Madrid through a reproducible scheme of Extraction, Transformation and Analysis that allows for future periodic studies.

For extraction, I performed ethical scraping of advertisement platforms. The initial objective was to capture the complete August 2024 supply (11,416 homes), but since this wasn't possible, I implemented random sampling using publication date as the ordering criterion.

Project Phases

1. **Extraction**: Development of ethical scraper.
2. **Transformation**: Cleaning, duplicate removal, formatting and creation of derived variables.
3. **Analysis**: Visualization and pattern extraction in three notebooks:
   - `1. EDA.ipynb`: Exploratory analysis
   - `2. viviendas.ipynb`: Characterization of the real estate stock
   - `3. mercado.ipynb`: Price and correlation analysis

Documentation:

Below is a summary of the project's key documentation and its location:

1. **Folder** `0. Knowledge (Cursor context)`: Contains scraper documentation. The most relevant files are:
   - `project_description.md`: General description, objectives and project structure.
   - `navegacion_y_selectores.md`: Web navigation strategies according to detected context.
   - `proxies.md`: Proxy service configuration.
   - `reportes.md`: Reporting system for navigation testing and validation.
   - `stealth_config.md`: Anti-detection tool configuration.
   - `global_rules.md`: Cursor behavior rules.
   - `nuevos_desarrollos.md`: Change log and version control.
   - `resultados_scraper.xlsx`: Scraper output and input for visualizations. Includes:
     - `Distrib muestra`: Sample and main groupings.
     - `Variables`: Variable index and their location (`1. EDA.ipynb`, `2. viviendas.ipynb` or `3. mercado.ipynb`).
2. **Folder** `1. EDA`:
   - `1. EDA.ipynb`: Exploratory Data Analysis (EDA).

3. **Folder** `Visualización y Análisis`**:**
    - `2. viviendas.ipynb`: Real estate stock analysis.
    - `3. mercado.ipynb`: Price and correlation analysis.
4. **Development documents**: Virtual environment (`venv`), `requirements.txt`, navigation cache and other operational files.

Data Extraction

After deciding to extract data from public real estate offer platforms, I focused on ethical scraping, training myself in advanced techniques through tutorials and documentation. The objective was to achieve undetectable and sustained automated navigation, capable of simulating human behavior to extract information without being blocked.

Architecture decisions:

The first architectural decision was to opt for navigation with headless browsers versus direct HTTP requests via HTTP client. The reason was the architecture of the target websites (user-published advertisement sites), which required sequential navigation to access information about each property. This choice also allowed for JavaScript rendering and dynamic server response management.

The second decision was to use a residential rotating proxy provider to support the large amount of navigation required, enabling request distribution and avoiding IP blocks.

Scraper development:

After making these decisions, development involved putting the acquired knowledge into practice through the incorporation of numerous elements: CSS selectors and parsers for data extraction, navigation identity management (session, browser fingerprint, cookies, cache), rate limit control, latency and retries, strategies to overcome JavaScript and CAPTCHA challenges, and simulated human navigation patterns (think time, scroll time, jitter, click sequences).

With the technology stack defined—Python with automated navigation libraries and anti-detection configuration—and the key elements to implement, I documented requirements and functionalities in files that served as a technical repository and as context for AI-assisted development (Cursor IDE + Claude). From this base, I developed the code modularly, validating each phase before moving forward.

The result is a configurable scraper that allows extraction from multiple real estate platforms using CSS selectors, with tabulated and structured output.

Data Cleaning and Structuring

The data we obtain directly from the web is optimized to be viewed, not to be analyzed. This phase is what converts that original, unstructured text into the key variables and metrics we need for comparison.

Applied Transformation Steps

Ensure quality (Cleaning):

1. We remove duplicate entries and manage missing or null data.
    - We normalize formats so that all dates, currencies and units are consistent.
2. Decode information (Parsing):
    - We convert text elements (prices, surfaces) to real numerical values.
    - We identify whether key attributes exist in binary form (yes/no): Does it have an elevator? Does it have a garage?
    - We standardize geographical locations to group by zone.
3. Create key metrics (Derived Variables):
    - We calculate Price per m², our main comparability metric.

- We define price ranges by segments (percentiles) and surface.
- We establish geographical hierarchy (Main Zone → Subzone).

4. Review and Consistency (Validation):
   - We detect and handle outliers that could bias the analysis.
   - We verify that variables are consistent with each other and within valid ranges.

Tools and Workflow

The process began with a quick exploration in Excel, generating a base file (`resultados_scraper.xlsx`). This file is the starting point for our Python/Pandas script (`1. EDA.ipynb`), where transformations are completed and all analytical variables are generated.

Challenges of this Phase

1. Disorganized Information: Original ads are very heterogeneous; often, information is incomplete, poorly labeled or written in different ways.
2. Segment Definition: Establishing the correct thresholds to classify prices and surfaces required several iterations.

The final result of this rigorous cleaning is a dataset ready for analysis of 610 homes with 41 structured variables.

Visualization and Analysis

The analysis presented is descriptive and does not aim to test or reject hypotheses. The notebooks contain visualizations and additional data that allow for deeper exploration beyond what is summarized here.

This phase constitutes the main output of the project, structured in three Jupyter Notebooks:

1. Exploratory Data Analysis (EDA)

The EDA constitutes the systematic reconnaissance phase of the dataset. Its objective is to understand the structure, quality and analytical potential of the data before formulating specific analyses.

Content:

- Univariate (distributions, outliers, missing values) and bivariate analysis (correlations between variables)
- Creation of derived variables: `precio_el_metro_cuadrado` , price and surface ranges, equipment aggregations
- Definition of segments by percentiles and geographical groupings (zone, subzone)
- Establishment of thresholds for outlier filtering

The result is the preparation of the dataset for the specific analyses of the following phases.

Relevant EDA findings:

The distribution analysis detects outliers in `metros_cuadrados` that, although distant from the interquartile range (`IQR`), correspond to plausible homes: they concentrate in `zones` that house large properties (mainly `chalets` in `north` and `west`) and the rest of their variables present values consistent with real high-end homes. I decided to keep them because their removal would bias the analysis toward "typical" homes, losing relevant information about the luxury segment. For greater rigor, I report statistics with and without outliers when the difference is substantial.

The EDA also establishes the data subsets used in the rest of the analysis (defined in the `# Subconjuntos (ALL)` cell of the notebook).

Analysis approach:

I adopt an exploratory approach without prior hypotheses, letting patterns emerge from the data. Conclusions are confined to main trends; notebooks contain additional visualizations for those who want greater granularity. When I formulate explanatory hypotheses, I explicitly identify them as such.

The distinction between `2. viviendas.ipynb` and `3. mercado.ipynb` responds to two complementary perspectives detailed below.

2. Properties

The sample (610 homes from a universe of 11,416) was obtained through ordering by publication date.

This ordering criterion, although not random in the strict sense, doesn't present an obvious theoretical relationship with the variables of interest (price, surface, amenities), which reduces the risk of systematic bias. However, factors such as supply seasonality or seller profile could introduce undetected biases. This limitation should be kept in mind when interpreting results.

Results should be interpreted with this caution, especially in segments with few observations.

At the `zone` and `subzone` level, samples exceed or approach 50 observations, allowing robust analyses. At the district level, some cases have less than 20 observations, which prevents reliable analyses at that granularity.

This phase characterizes the available supply focusing on the physical properties of homes. `precio_en_euros` (`pee`) is used as a segmentation variable, but `precio_el_metro_cuadrado` (`pmc`) is reserved for market analysis.

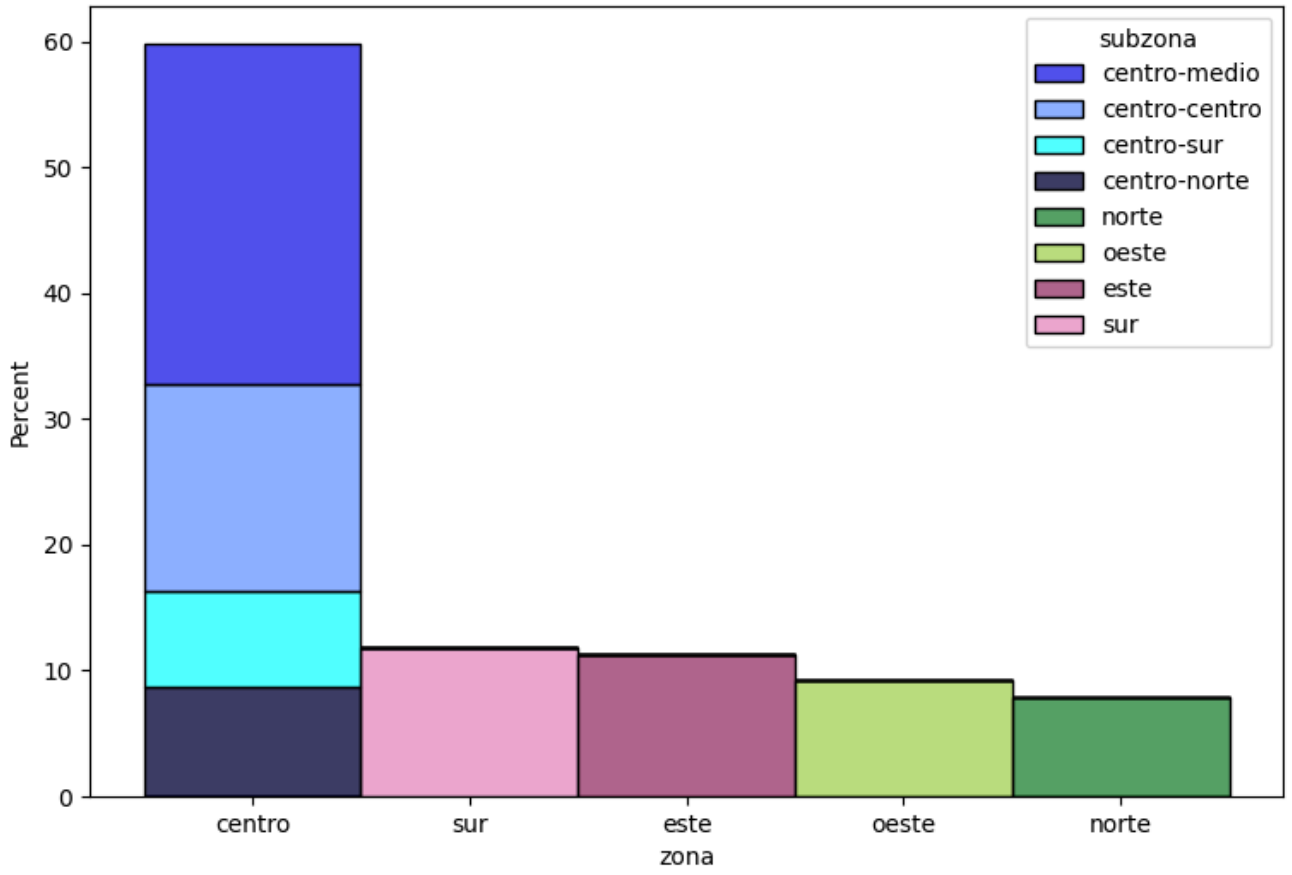
Justification for this separation:

- `pee` (total price) serves a dual function: besides indicating absolute value, it segments typologies (high-end vs. economical homes) and defines profiles when combined with physical characteristics.
- `pmc` (price per m²) is a normalized metric that eliminates the size effect and allows comparing market value regardless of surface. This distinction is analytically relevant: as will be seen, `chalets` present high `pee` but low `pmc`, revealing that their high total price responds to their large surface, not to a higher valuation per square meter. Therefore, `pmc` is reserved for market analysis in `3. mercado.ipynb`.

Objective: Dimension the supply by typology, identify characteristic patterns by zone and establish natural groups of homes.

LOCATION

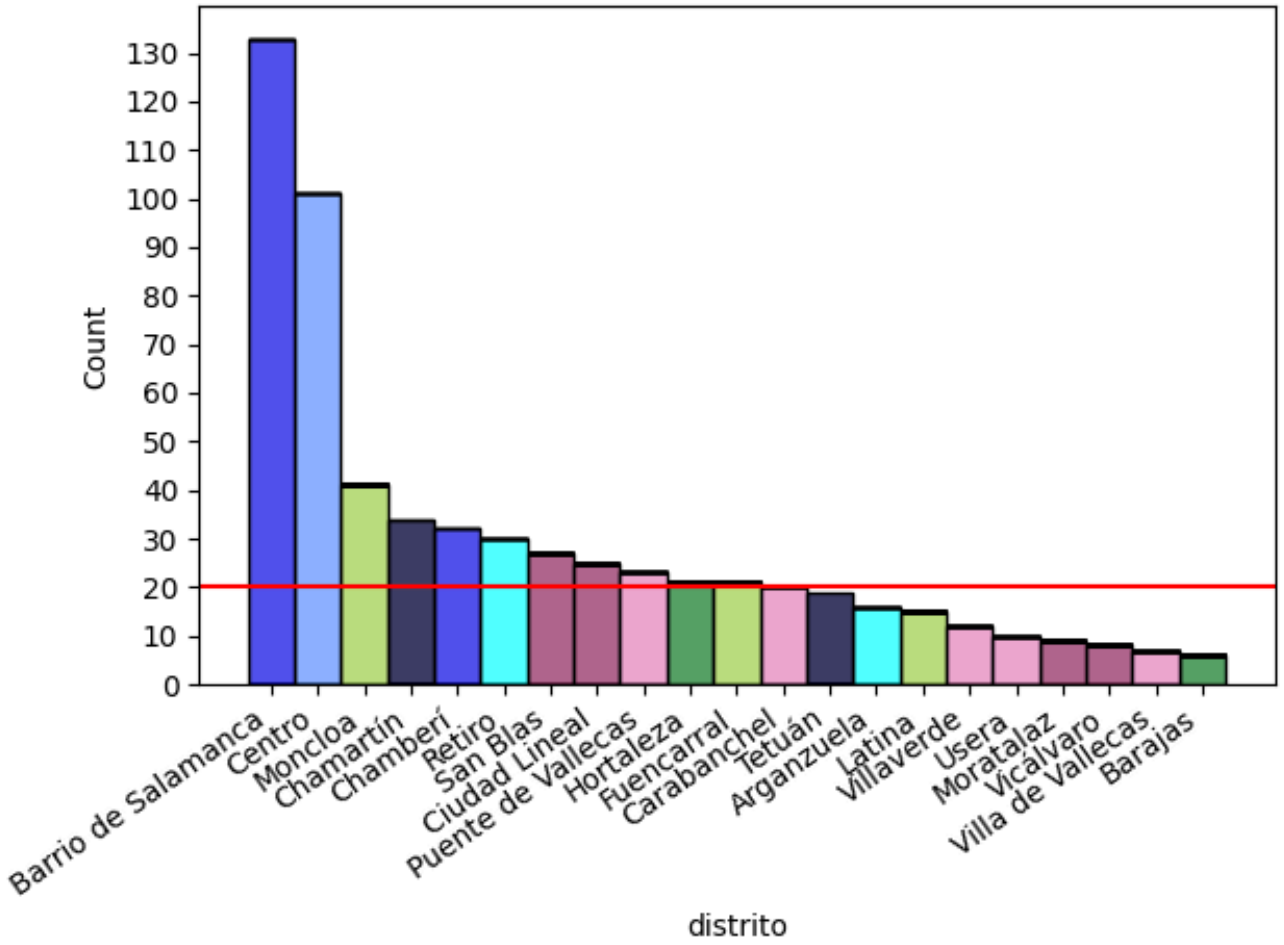
The following graph shows the geographical distribution of the sample by `zone` and `subzone`. The concentration in the `center` is notable: it represents more than 60% of observations.



The districts `Salamanca` (133 homes) and `Centro` (101) dominate, followed by `Moncloa` (41) and `Chamartín` (34). Peripheral zones (`south` , `east` , `west` , `north`) show lower and balanced representations among themselves.

This distribution may reflect greater supply rotation in the center—more active markets generate more new ads—rather than a sampling bias, although this hypothesis cannot be confirmed with available data.

The following graph breaks down the sample by district. Several `periphery` districts fall below 20 observations, a threshold I consider minimum for reliable statistical analyses. This limitation conditions the level of geographical granularity of the analysis.



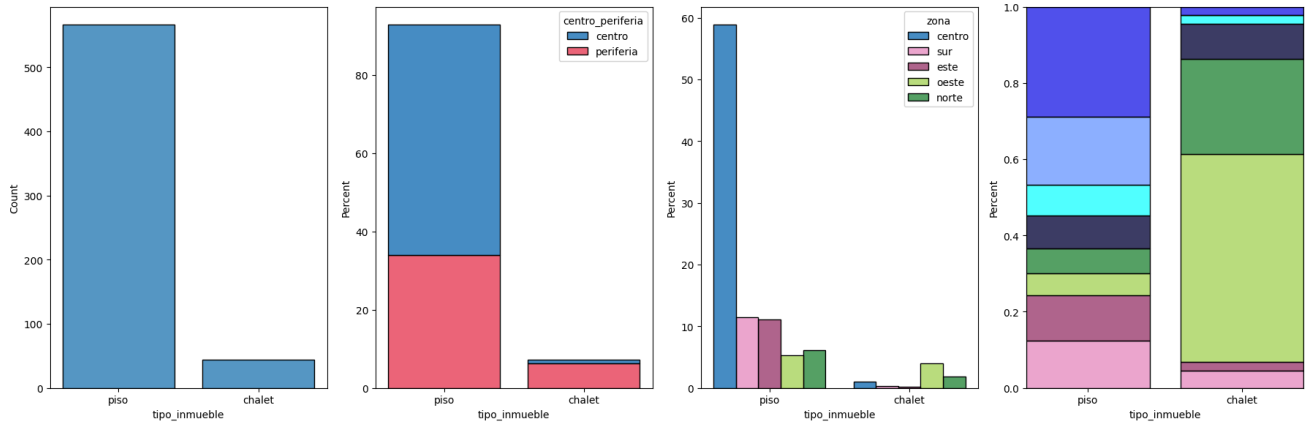
CHARACTERISTICS

Property type

The following composite graph analyzes the distribution of `pisos` (apartments) and `chalets` by `zone`, condition (`new` / `used`) and `center` / `periphery` location.

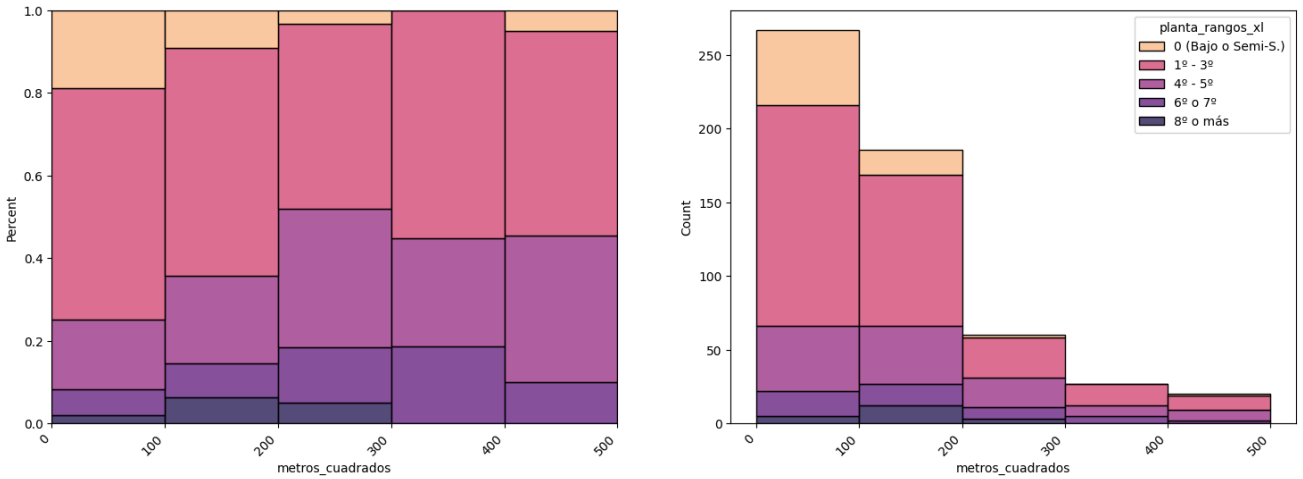
The sample is dominated by `pisos` (566 observations, 92.8%) versus `chalets` (44 observations, 7.2%). This proportion limits the robustness of specific conclusions about `chalets`, which should be interpreted with caution given the reduced sample size.

`Pisos` have presence in all zones with greater concentration in the `center`. `Chalets` are mainly distributed in peripheral zones (`north`, `west`, `south`) with marginal presence in the `center`, a pattern consistent with land availability for single-family housing. Both types show predominance of used homes over new ones, more pronounced in `pisos`.



Floor Number

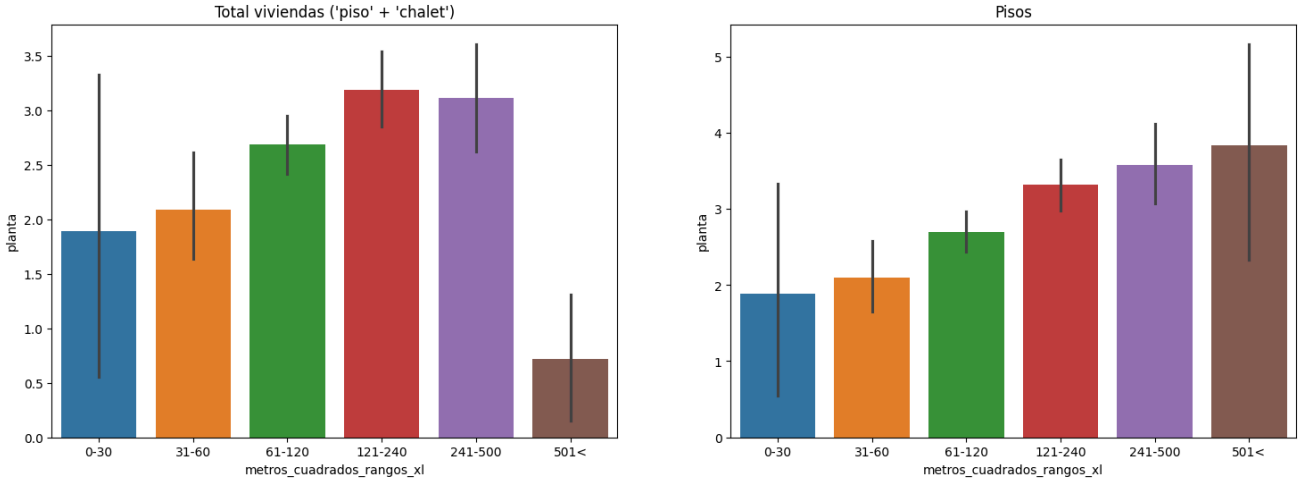
The following composite graph crosses the `floor` number with four variables: location (`center` / `periphery`), surface range, orientation (`exterior` / `interior`) and condition (`new` / `used`).



The distribution concentrates on low intermediate `floors` : `1st` - `3rd` (307 homes, ~50%) and `4th` - `5th` (120 homes). They are followed by `ground floor` / `semi-basement` (71), floors `6th` - `7th` (48), `chalets` (44) and floors `8th` + (20).

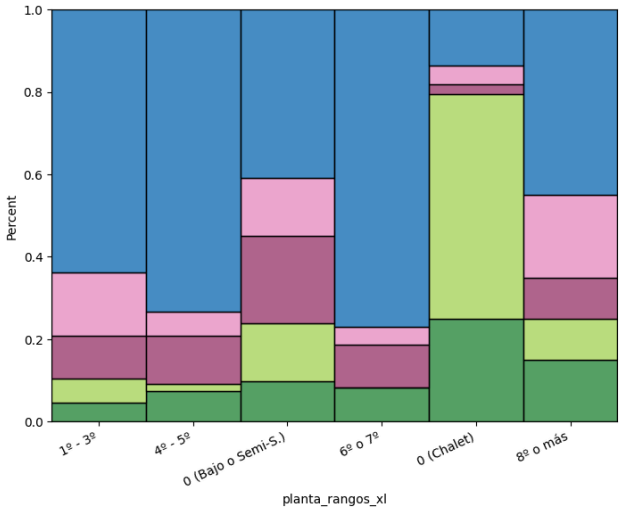
The `center` shows greater diversity of `floors` with strong presence of intermediate and high ones, reflection of its more heterogeneous building stock. `Periphery` zones present a greater proportion of ground floors and `chalets` . New homes have greater relative presence on high floors, while used ones are more evenly distributed—hypothesis: recent `new construction` in Madrid has concentrated in higher buildings.

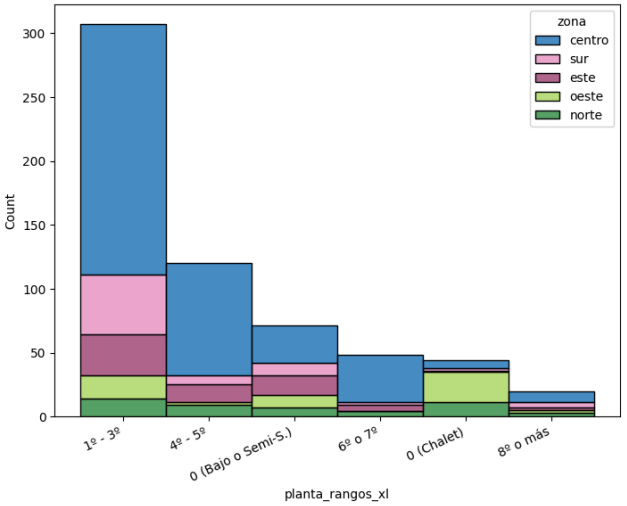
The following graph explores the relationship between `floor` and `surface` , comparing the complete dataset with the subset of `pisos` (excluding `chalets` , which distort the analysis as they typically have 1-2 floors regardless of their `surface`).



There is a positive correlation between `floor` and `square meters` , more visible when analyzing exclusively `pisos` . This correlation may reflect historical construction patterns: old buildings in the center usually have larger noble floors (1st-2nd), while modern buildings with high floors tend to offer larger homes as a premium product. However, this is a hypothesis that would require building age data to be confirmed.

The following composite graph compares the distribution of `floors` by `zone` in relative (proportions, left) and absolute (counts, right) terms.

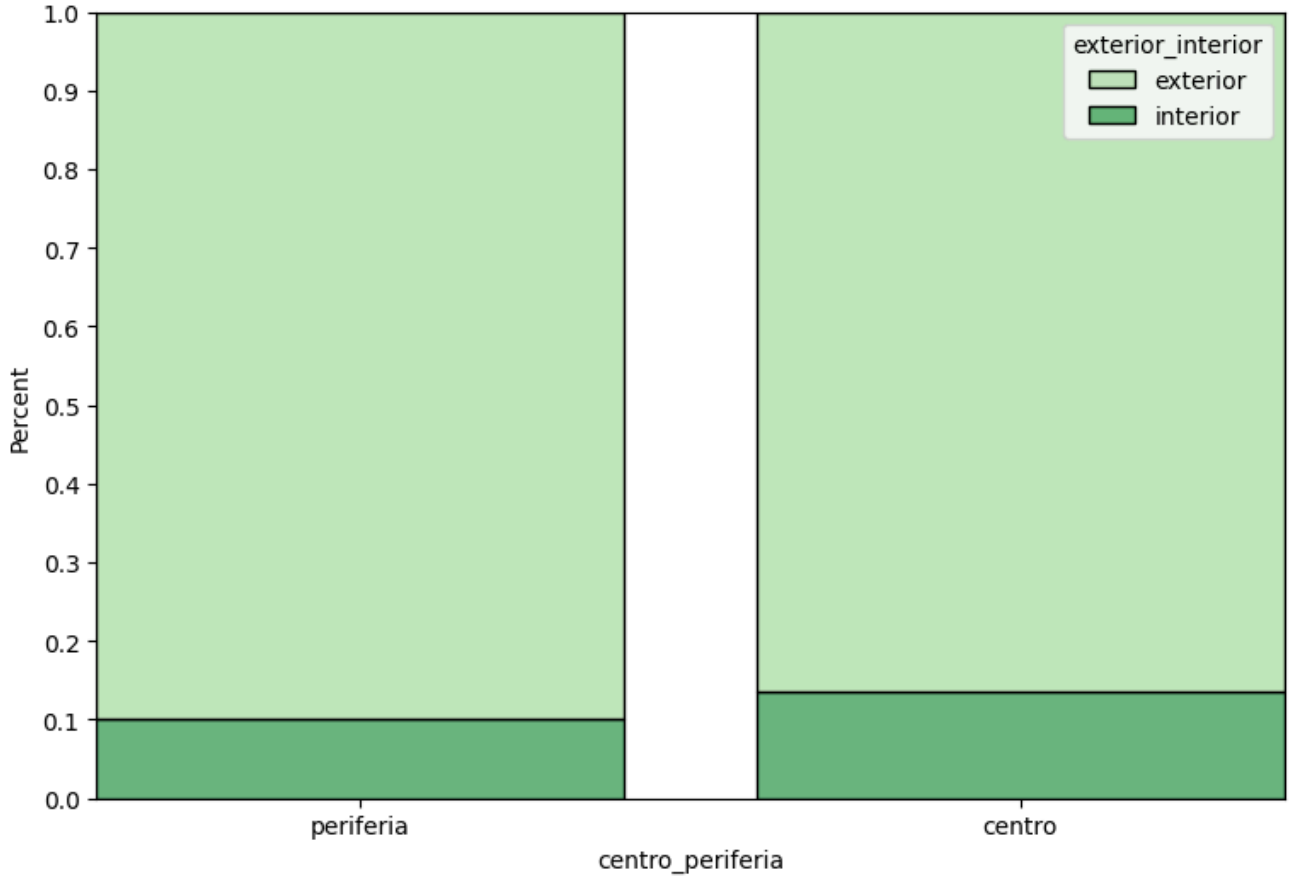




Floors 1st-3rd dominate in all `zones` (~50%), but high floors (`6th` or higher) are concentrated almost exclusively in the `center`. The count graph confirms that the `center` accumulates the greatest diversity of floors, while peripheral zones present a greater proportion of ground floors and `chalets`, especially `north` and `west`. This pattern is consistent with the predominant building typologies in each `zone`: medium-high historic buildings in the `center` versus low-density developments in the `periphery`.

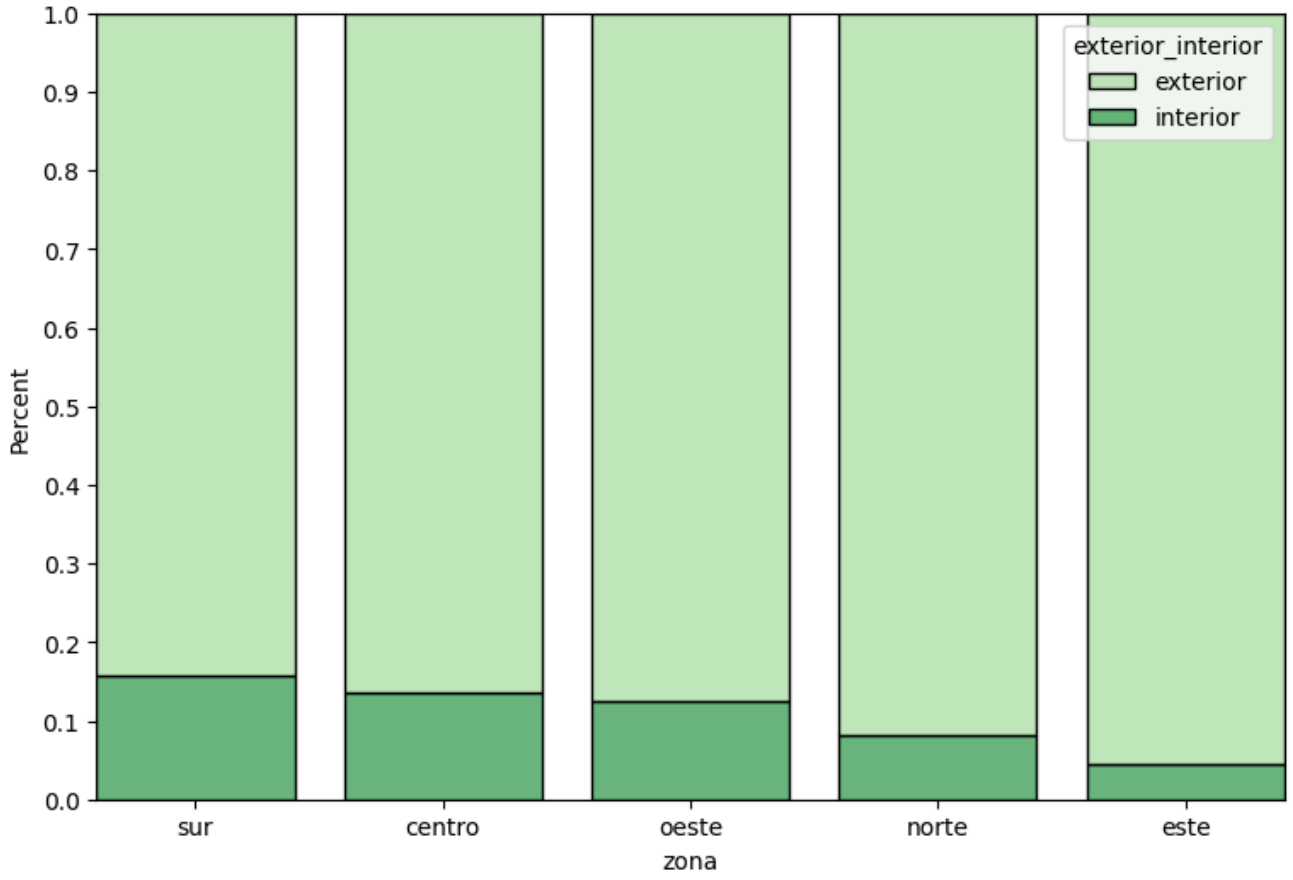
Exterior / Interior

The following graph shows the proportion of `exterior` and `interior` homes according to location (`center` / `periphery`).

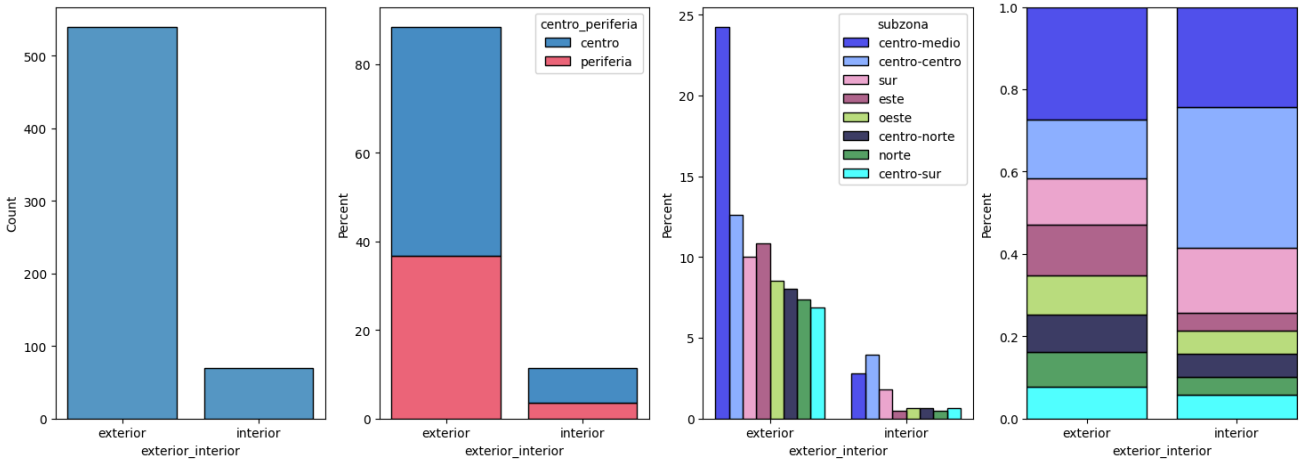


`Exterior` homes dominate the sample (540 obs., 88.5%) versus `interior` (70 obs., 11.5%). The proportion of exteriors is slightly lower in the `center` (~86%) than in the `periphery` (~91%). This difference could reflect the higher historical building density of the `center`, with more interior courtyards and homes without street facades, although it's not possible to confirm this with available data.

The following graph breaks down the `exterior` / `interior` proportion by `zone` and `subzone`, and analyzes the geographical distribution of `interior` homes.



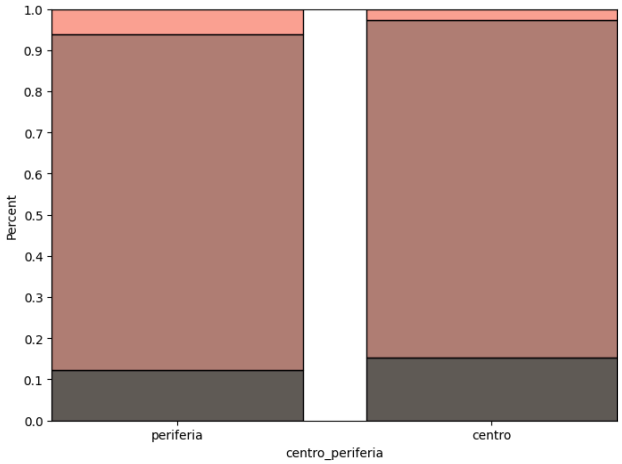
The `east` and `north` zones reach ~95% of `exterior` homes; `center` and `west` present the highest proportions of `interior` (~13-15%). Among `interior` homes, `center-center` concentrates the highest proportion (31.2%), while `north` represents only 6.6%. This pattern probably reflects two combined factors: the higher density of old buildings with interior courtyards in the historic `center`, and the greater sample weight of this subzone.

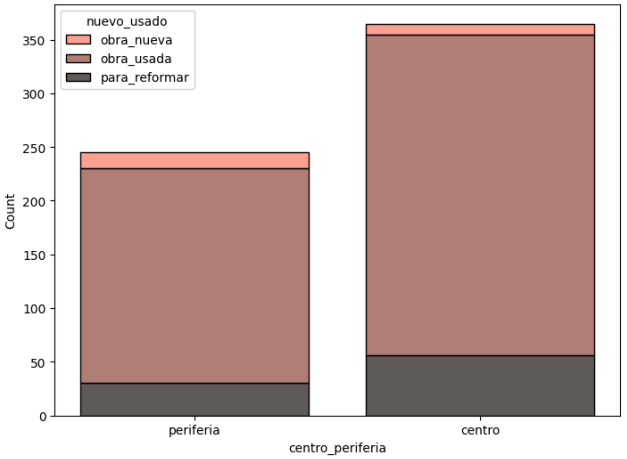


The previous composite graph crosses the `exterior` / `interior` condition with multiple variables. It stands out that exterior homes show greater diversity of `surfaces`, while `interior` ones concentrate in intermediate ranges (`61 m²` - `240 m²`). The absence of large `interior` homes (> `300 m²`) is consistent with the physical limitations of interior courtyards in urban buildings.

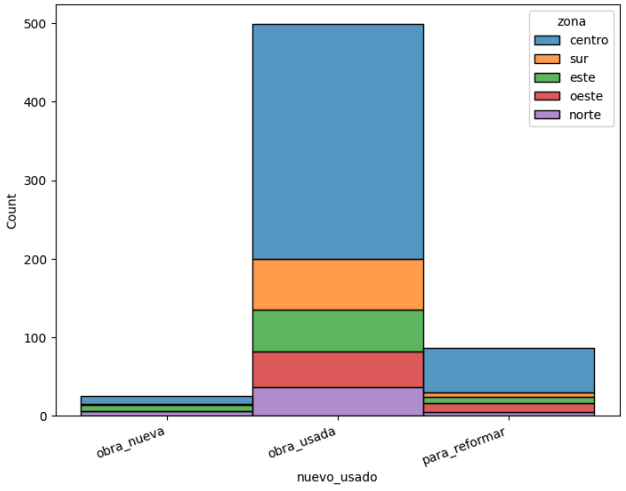
Property Condition

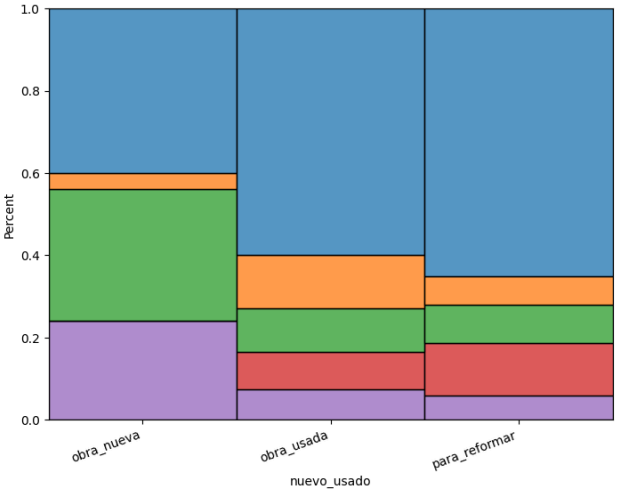
The following composite graph analyzes the distribution of conservation status (`new_construction`, `used_construction`, `to_reform`) by location and zone.





The sample shows predominance of `used_construction` (499 obs., 81.8%), followed by homes `to_reform` (86 obs., 14.1%) and `new_construction` (25 obs., 4.1%). The scarce representation of `new_construction` limits conclusions about this segment.





Both zones (`center` and `periphery`) show similar patterns with predominance of `used_construction` (~82%). The `center` presents the greatest relative diversity among categories. Homes `to_reform` concentrate in intermediate-high surface ranges (`121 m²` - `500 m²`), while `new_construction` shows more dispersed distribution.

SPACE

The following table shows descriptive statistics of `metros_cuadrados` , with and without outlier filtering.

```
df[['metros_cuadrados']].describe()
```

	metros_cuadrados
count	610.000000
mean	178.991803
std	224.723165
min	20.000000
25%	70.000000
50%	109.500000
75%	198.750000
max	3015.000000

The average `surface` is `179 m²` with a median of `109.5 m²` (difference of `69.5 m²`) including outliers. After filtering to 3× `IQR`, the average drops to `133.5 m²` and the median to `103 m²` (difference reduced to `30.5 m²`). This divergence between mean and median confirms an asymmetric distribution with right tail: outliers significantly skew the mean but barely affect the median, which justifies prioritizing the median as a central tendency statistic in this analysis.

```
df_sin_outliers_metros_cuadrados[['metros_cuadrados']].describe()
```

	metros_cuadrados
count	569.000000
mean	133.488576
std	91.215638
min	20.000000
25%	69.000000
50%	103.000000
75%	165.000000
max	450.000000

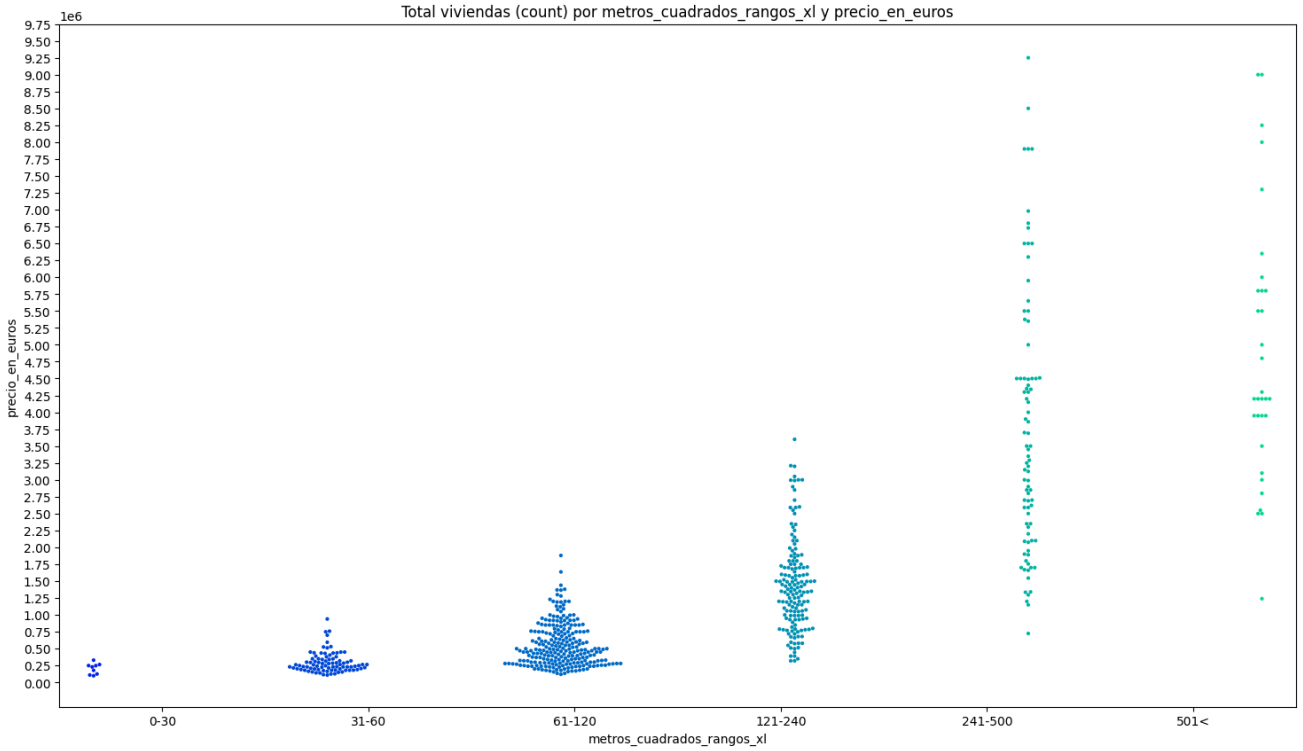
Surface and total price

Variables: `precio_en_euros` + `metros_cuadrados`

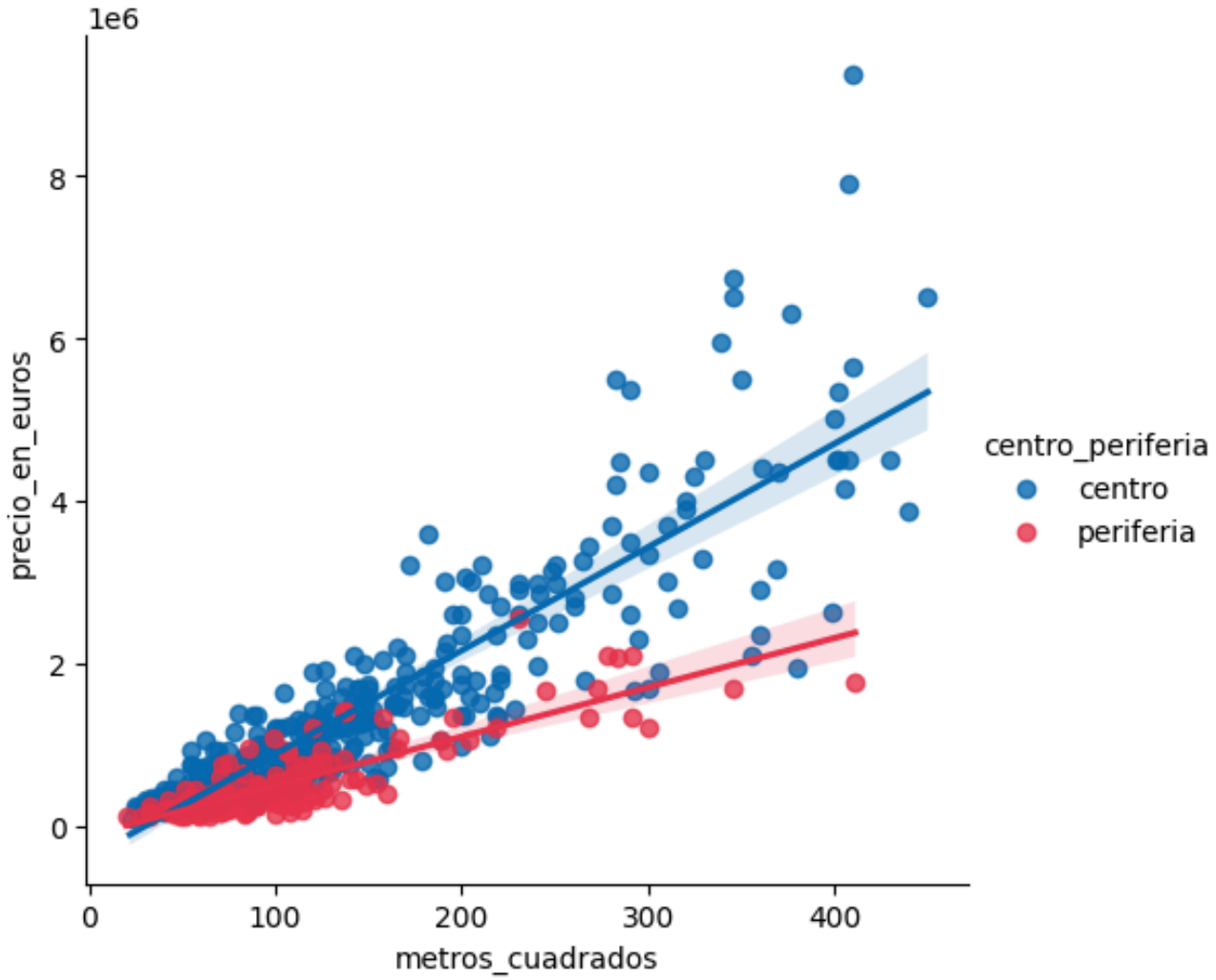
The following scatter plot represents the relationship between `surface` and `total price`, differentiating by `square meter` ranges.

The relationship between `surface` and `total price` is linearly positive, with marked differences between `center` and `periphery`. From ~ `150 m²` onwards, dispersion increases notably, especially in the `center` where high-value homes appear (>4M€).

This heteroscedasticity pattern—increasing variability with size—has a plausible explanation: small homes constitute a relatively homogeneous product, while in large ones more differentiating factors intervene (construction qualities, premium location, luxury amenities) that widen the range of possible prices.



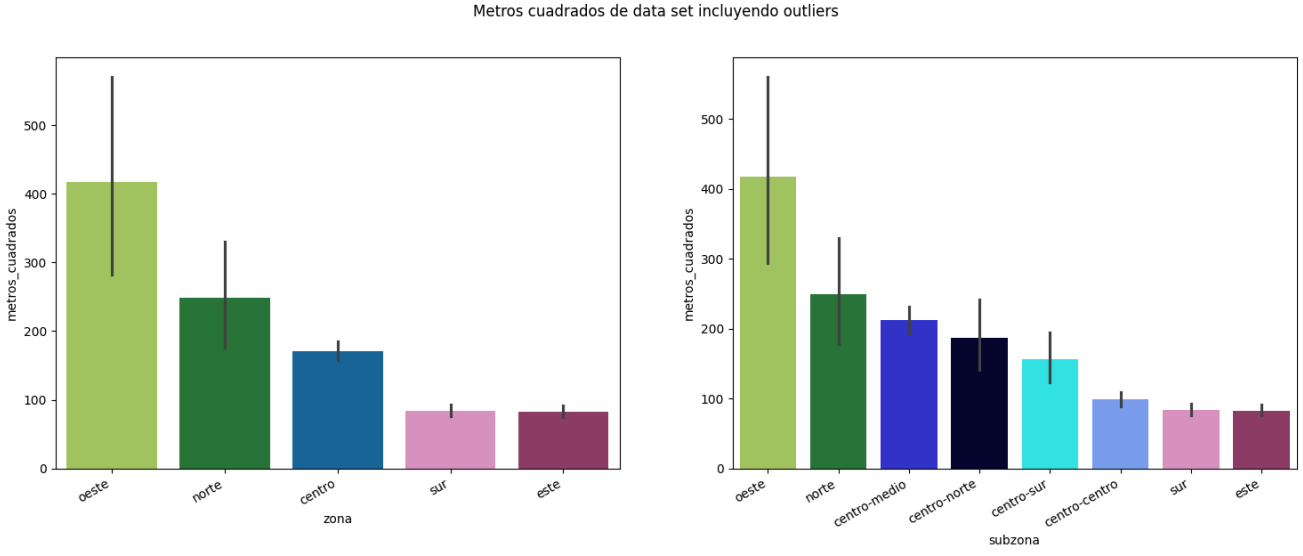
The following scatter plot with regression lines compares the `price` - `surface` relationship between `center` and `periphery`.



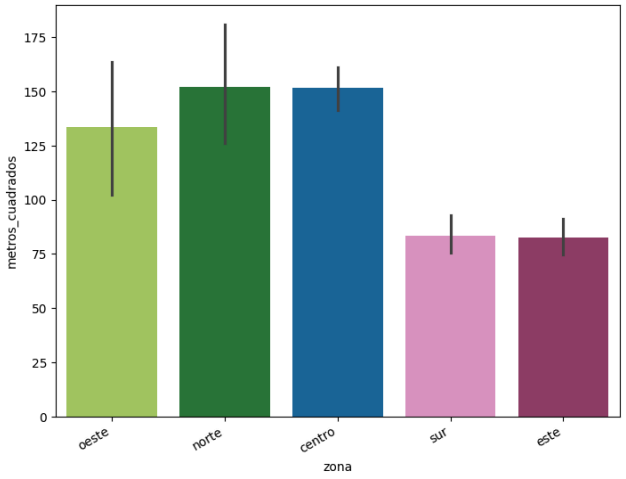
Trend lines reveal different slopes: the `center` presents greater price increase per additional m², indicating that each added square meter "is worth more" in central locations. The `periphery` shows more compact grouping in low-medium ranges. Most homes concentrate between `61 m²` - `240 m²` with prices between `200K€` - `1M€`.

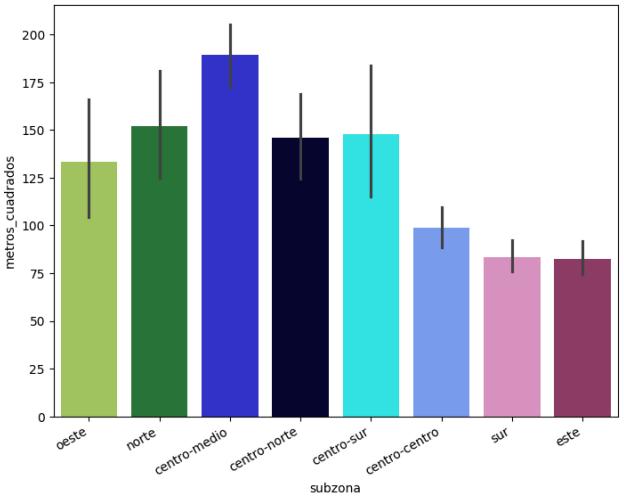
Upper extreme values are not data errors but significantly skew averages: they correspond to homes with a very specific profile (large `chalets`), concentrated in the `west` and `north` zones. The `west` zone illustrates this effect: its average goes from `417 m²` (with outliers) to `133 m²` (without outliers), a 68% reduction that makes it fall from first to third place among `zones`. This example justifies my decision to report both metrics and prioritize medians when distribution is asymmetric.

The following graphs compare average surface by `zone` and `subzone`, with and without outliers:



Including outliers: `west` leads with 417 m² average, followed by `north` (249 m²) and `center` (170 m²). However, standard deviations are extreme (`west` : 533 m², `north` : 282 m²), an unmistakable sign of very dispersed distributions where the average is not very representative.



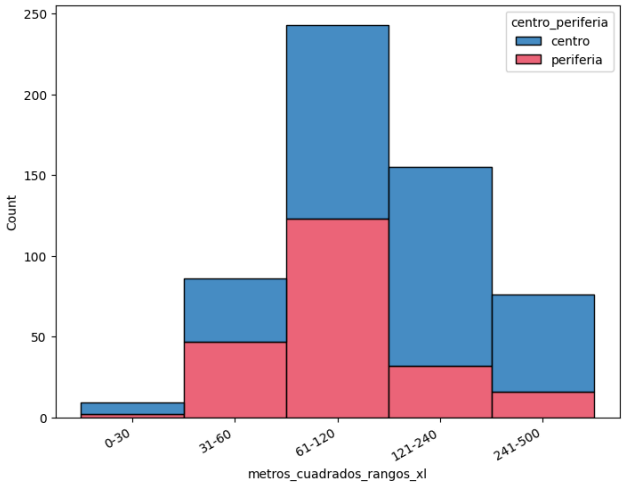


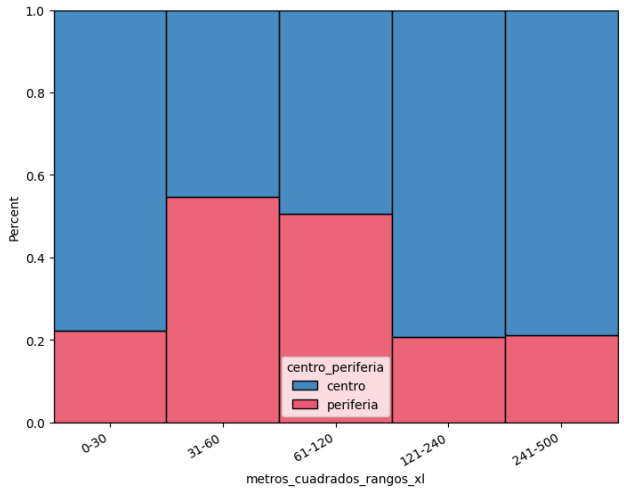
Excluding outliers: the ranking changes substantially. `Center` and `north` lead (~152 m²), `west` falls to third place (134 m²). `South` and `east` maintain the lowest averages (~83 m²). That medians are systematically lower than means in all zones confirms asymmetric distributions with right tail, even after filtering.

Surface according to zone

Variables: `centro_periferia` + `zona y subzona` + `metros_cuadrados`

The following composite graph shows the distribution of surfaces by `zone` , both in absolute values (counts) and relative (proportions).

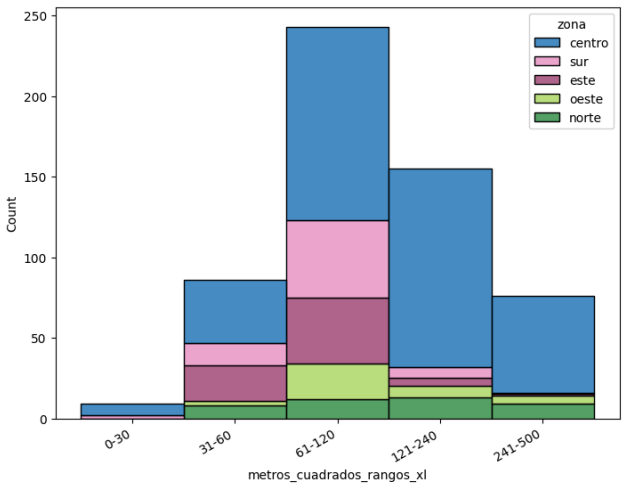


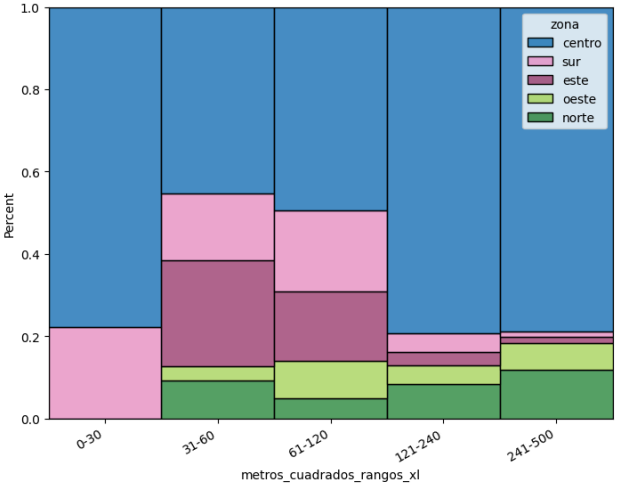


Surface distribution differs between `center` and `periphery` . The `center` presents greater relative concentration in small-medium ranges (`31 m²` - `120 m²`), while the `periphery` has greater relative weight in large surfaces (`>241 m²`). This pattern is consistent with land availability: large homes require plots that are scarce in the consolidated urban `center` .

In terms of relative proportion, the `center` represents approximately 20-25% of homes in small-medium ranges (31-120 m²), but this proportion decreases in large ranges where the `periphery` dominates.

Regarding patterns by `zone` , the `center` shows more balanced distribution between small and medium ranges, while the `periphery` presents greater weight of large surfaces (>241 m²), reflecting the greater presence of `chalets` and single-family homes.





Descriptive statistics by `zone` reveal significant differences in average surface. `North` and `center` lead with averages of ~152 m² and ~151 m² respectively, followed by `west` (~134 m²). `South` and `east` present the lowest averages (~83 m²).

In terms of variability, all `zones` show high dispersion (standard deviations >36 m²), with `center`, `west` and `north` being the most variable (~97 m²). Medians are systematically lower than means, indicating asymmetric distributions.

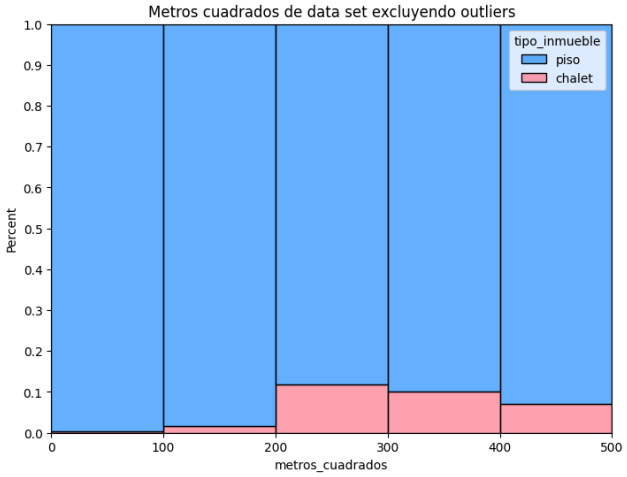
By subzones, `center-medium` shows the highest values, followed by `center-north`; `center` subzones present greater heterogeneity of surfaces, while in the `periphery` distribution is more homogeneous but with lower average surface.

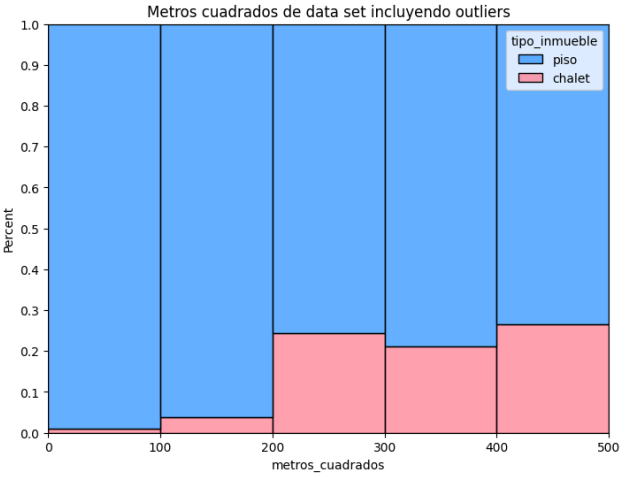
In terms of concentration, small homes (<100 m²) predominate in peripheral zones, while large ones (>300 m²) have greater relative presence in `center`, `north` and `west`.

Surface and property type

Variables: `tipo_inmueble` + `metros_cuadrados`

The following composite graph compares surface distribution between `pisos` and `chalets`, with and without outlier filtering.





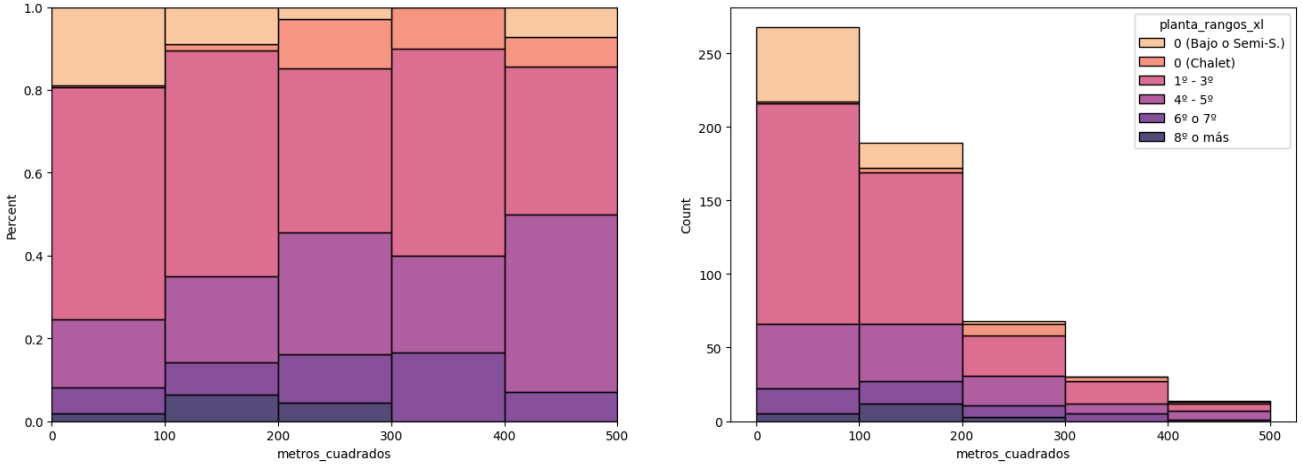
`Pisos` dominate all `surface` ranges up to `240 m²`, representing >95% of homes in small-medium ranges (`31 m²` - `200 m²`). `Chalets` concentrate on larger surfaces (`>200 m²`).

Outlier filtering has differential impact by type: `pisos` go from 566 to 553 observations (-2.3%), while `chalets` fall from 44 to 16 (-63.6%). This data is revealing: `chalets` contain a very high proportion of extreme values, confirming that they are the main source of surface outliers in the dataset. Distribution graphs confirm that `pisos` present concentrated distribution between `50 m²` - `200 m²`, while `chalets` show greater dispersion and systematically higher values.

Surface and floor number

Variables: `planta, planta_rangos(_xl)` + `metros_cuadrados`

The following composite graph crosses `surface` with `floor` number, showing both proportions and absolute values.

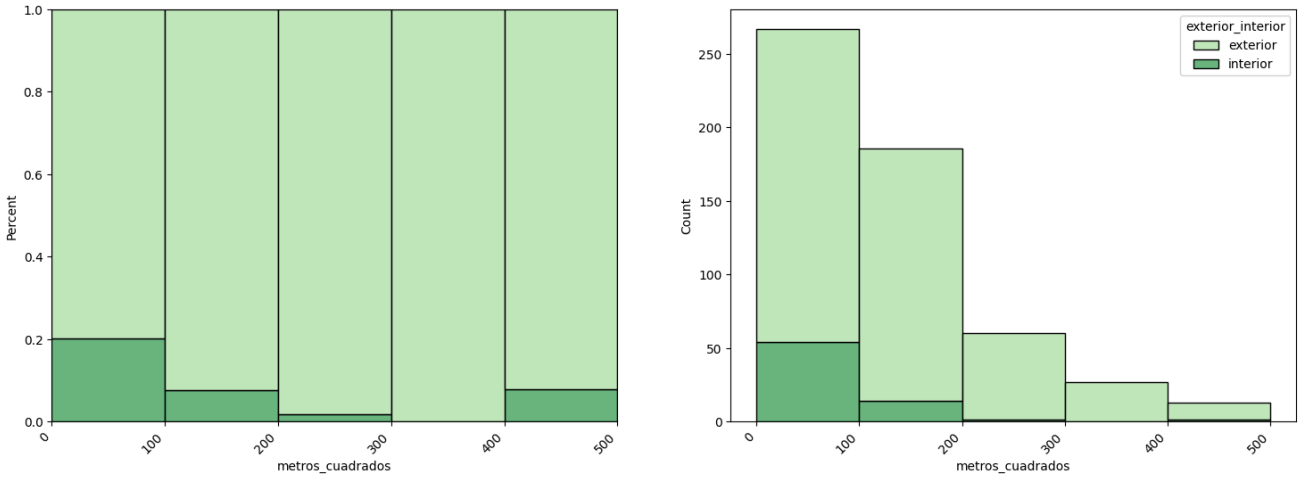


`Surface` varies according to height in a non-linear way. `Ground` / `semi-basement` floors show bimodal distribution (small and large ranges), a pattern that could reflect two distinct typologies: premises converted to housing and ground floors of stately buildings. `Chalets` dominate upper ranges (>200 m²). Intermediate floors (`1st` - `7th`) concentrate on `61 m²` - `240 m²`, with slight shift toward smaller surfaces on higher floors. Floors `8th` + present balanced distribution, although with limited sample (<20 obs.) that advises interpretive caution.

Surface and Exterior/Interior

Variables: `metros_cuadrados` + `exterior_interior`

The following composite graph analyzes the relationship between `surface` and orientation (`exterior` / `interior`), showing both proportions and absolute distribution.

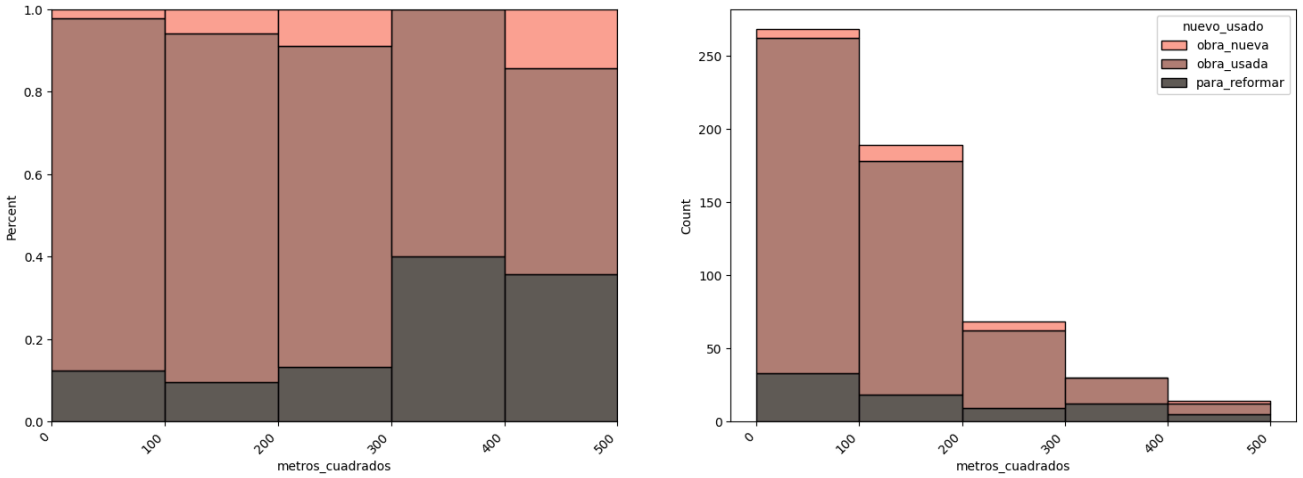


Exterior homes dominate in all surface ranges. Interior ones represent ~20% of small homes (< `100 m²`), a proportion that decreases progressively until being practically non-existent above `300 m²`. This inverse relationship between `surface` and proportion of interiors makes physical sense: interior homes depend on courtyards whose dimensions limit the maximum achievable size. The histogram confirms that most homes concentrate between `50 m²` - `200 m²`.

Surface and Property Condition

Variables: `nuevo_usado` + `metros_cuadrados`

The following composite graph relates surface with conservation status (`new_construction` , `used` , `to_reform`).

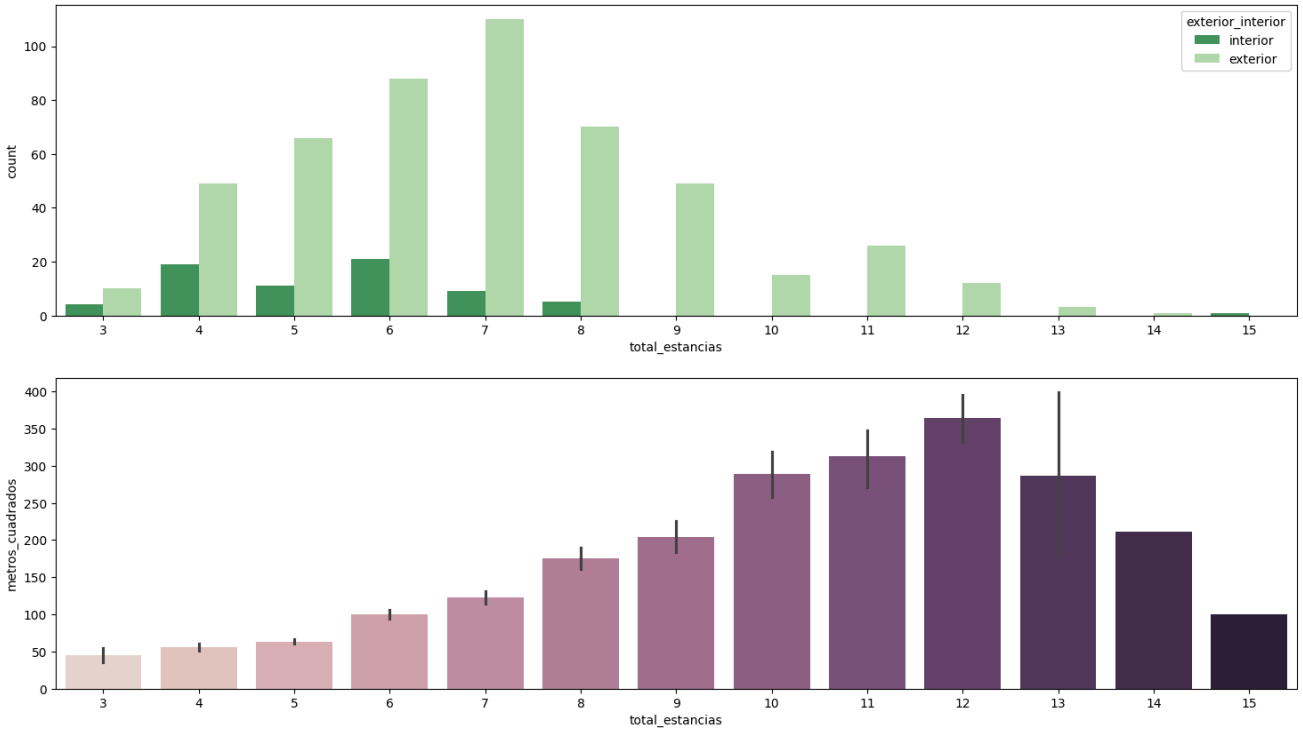


`New_construction` dominates in all ranges with uniform distribution between `50 m²` - `300 m²`. Homes `to_reform` concentrate in medium-high ranges (`121 m²` - `300 m²`) with low presence at extremes—hypothesis: large and old homes are more frequent candidates for comprehensive renovation. `New_construction`, with lower absolute representation, appears dispersed in all ranges without clear concentration, although the limited sample (25 obs.) prevents extracting robust conclusions.

Surface and total rooms

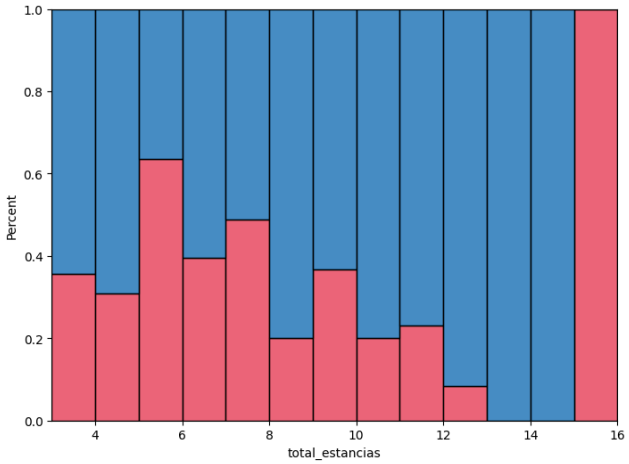
Variables: `total_estancias` + `metros_cuadrados`

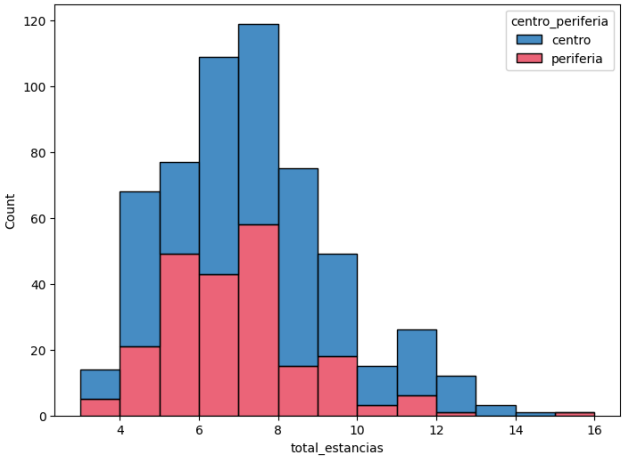
The following composite graph analyzes the relationship between number of `total_rooms` and `surface`, differentiating by orientation (`exterior` / `interior`) and location (`center` / `periphery`).



The relationship between `total_rooms` and `surface` is positive, as expected, with mode at 6-7 rooms and average surface growing from ~50 m² (3 rooms) to ~350 m² (12 rooms). Dispersion increases from 8 rooms onwards.

An interesting finding: for the same number of `rooms`, `center` homes usually have greater surface, a difference especially notable in medium ranges (6-8 rooms). This suggests larger rooms in the `center`, hypothesis consistent with the typology of historic buildings with high ceilings and generous layouts. The `center` dominates proportionally in homes with few rooms (3-5, ~60-70%); this proportion balances in medium ranges and reverses in high ranges where the `periphery` is majority, reflecting the presence of `chalets` with many rooms.

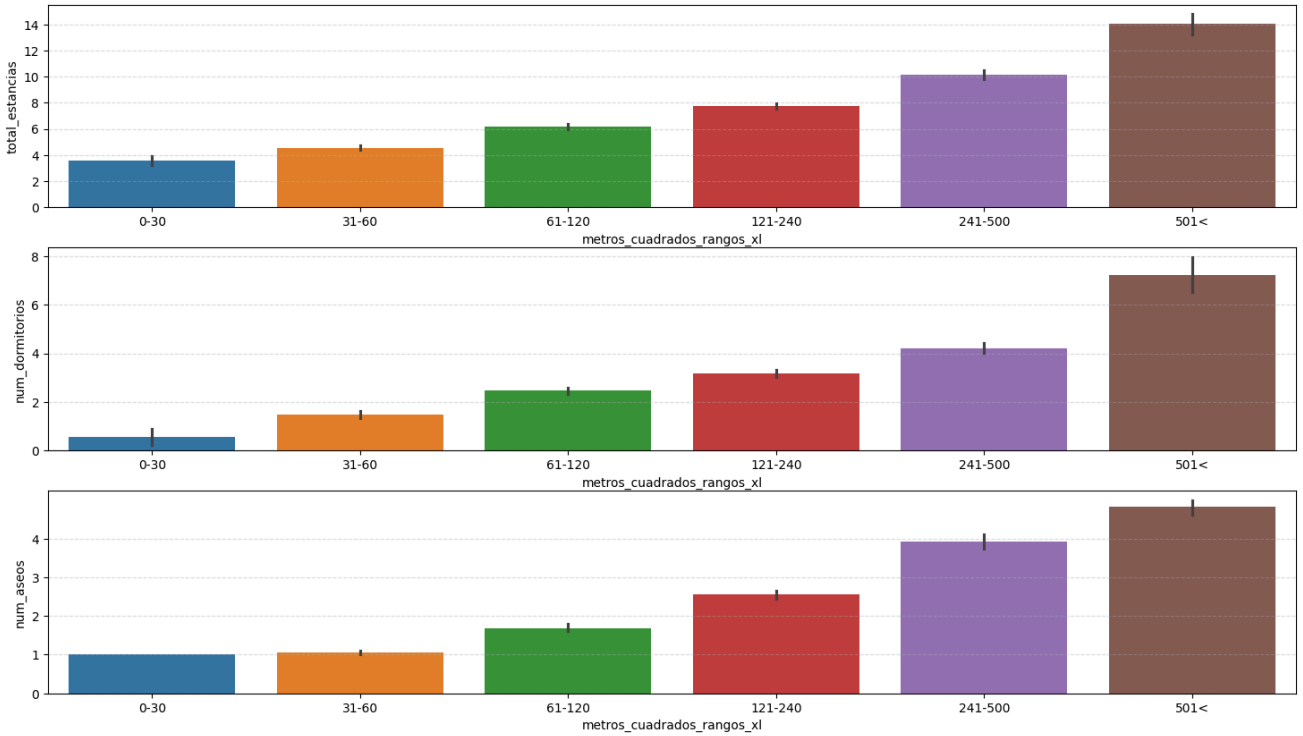




Surface, bedrooms and bathrooms

Variables: `num_dormitorios` **&** `num_aseos` **+** `metros_cuadrados`

The following composite graph relates `surface` with number of `bedrooms`, number of `bathrooms` and total `rooms`, segmented by surface ranges.



Regarding `bedrooms`, the mode is 3 (~120 obs.), followed by 2 (~110 obs.). 1-bedroom homes are scarce (~15 obs.) and 4+ represent a decreasing right tail. Average surface progresses approximately linearly: ~ `50 m²` (1 bedr.) → ~ `80 m²` (2) → ~ `110 m²` (3) → ~ `150 m²` (4) → > `200 m²` (5+), with increasing dispersion from 4 bedrooms.

Regarding `bathrooms`, 1-2 predominate (~70% of total). Surface follows similar pattern: ~ `60 m²` (1 bath) → ~ `100 m²` (2) → ~ `150 m²` (3) → > `200 m²` (4+). The most frequent combinations are 2-3 bedrooms with 1-2 bathrooms, typical configuration of urban family housing.

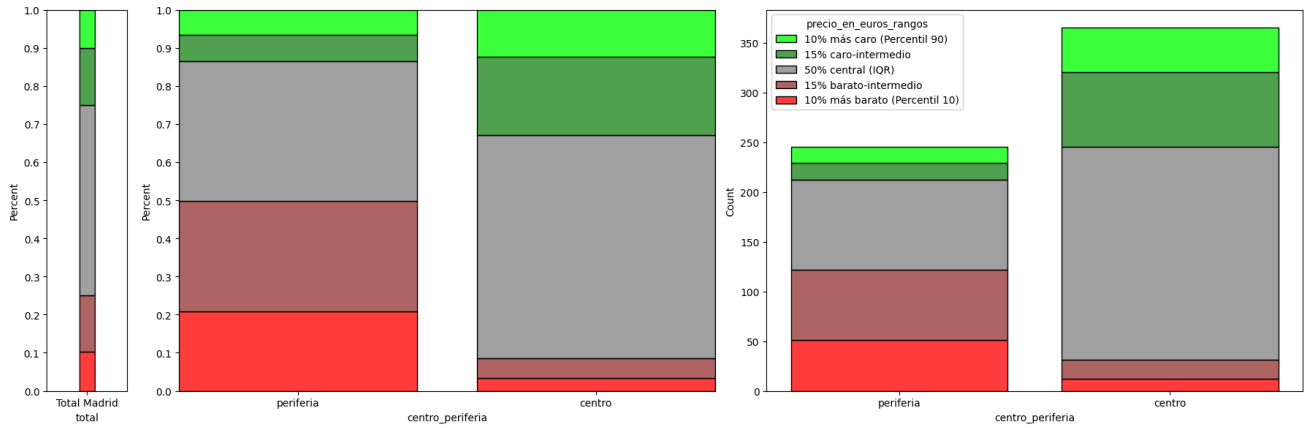
TOTAL PRICE and PEE

Total price (`precio_en_euros` or `pee`) is a composite variable that implicitly captures the combined effect of `surface`, `location`, `amenities` and `qualities`. Unlike price per square meter (`pmc`, which is analyzed in section 3. Market), `pee` is

useful for segmenting product typologies: a 2M€ home and another of 200K€ belong to different market segments regardless of their individual characteristics, and probably target very different buyer profiles.

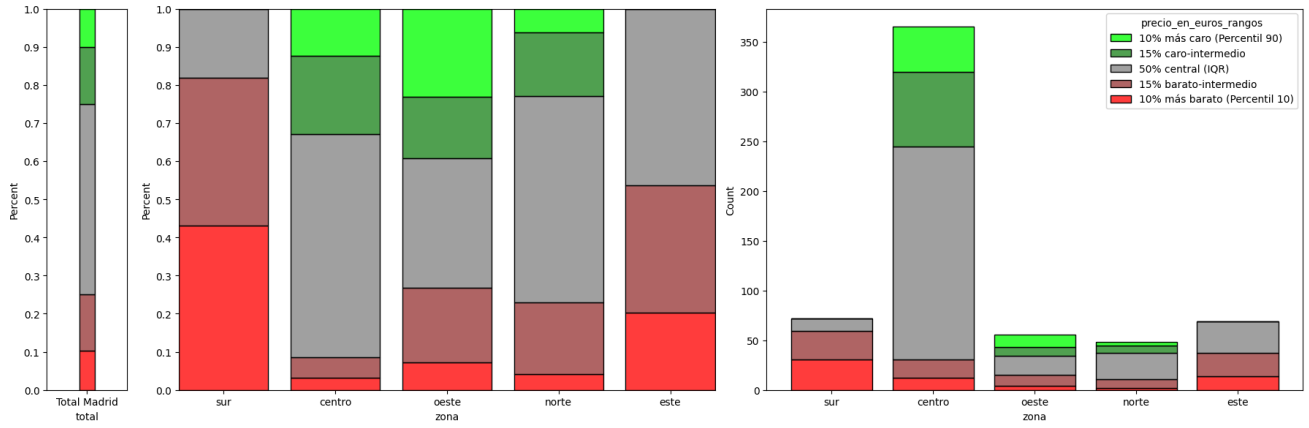
LOCATION and PEE

The following composite graph shows the distribution of `pee` percentiles by location (`center` / `periphery`), both in proportions and absolute values.



Spatial price segregation is clear. The `center` (365 obs., 59.8%) dominates high percentiles: it concentrates 45 of the 61 homes in `p90` (74%), well above its sample weight. The `periphery` (245 obs., 40.2%) presents the inverse pattern, with 51 of the 63 homes in `p10` (81%). In other words: the periphery doubles its representation in cheap homes (from 25% population to 50%) and underrepresents expensive ones (from 25% to ~15%). The center shows the inverse pattern although less pronounced, favored by its greater sample weight.

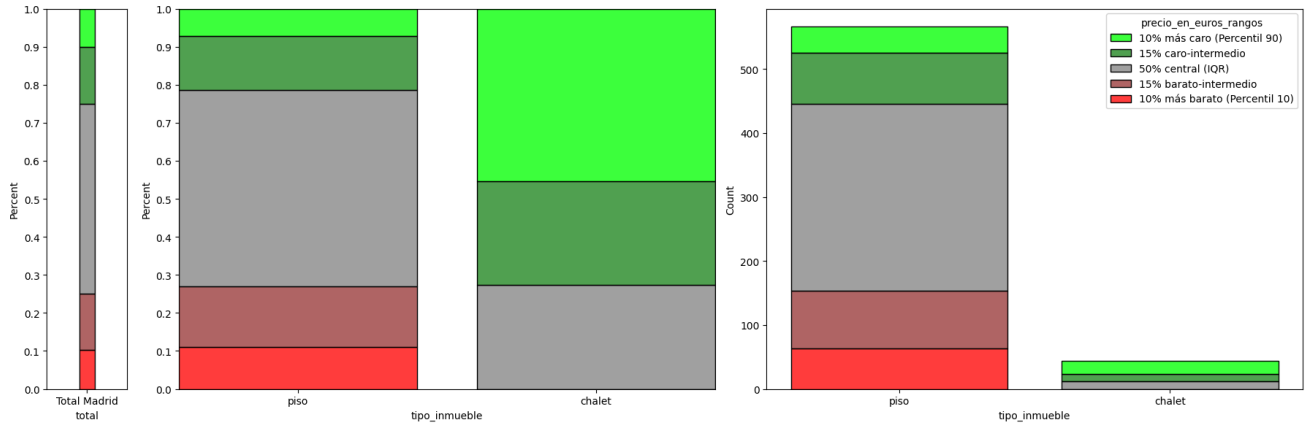
The following graph breaks down distribution by specific `zones` .



`South` and `east` concentrate greater weight of low `pee` percentiles. `West` and `center` concentrate ~40% of homes in high percentiles (vs 25% population). `North` behaves similarly to the general population, without clear bias toward any extreme. Peripheral zones— `south` (72 obs.), `east` (69), `west` (56) and `north` (48)—present more homogeneous price distributions among themselves, with greater concentration in low-medium ranges. The `center` shows the greatest variability and reaches maximum market values, reflecting its urban heterogeneity (from modest homes in popular neighborhoods to ultra-luxury product on the Castellana-Salamanca axis).

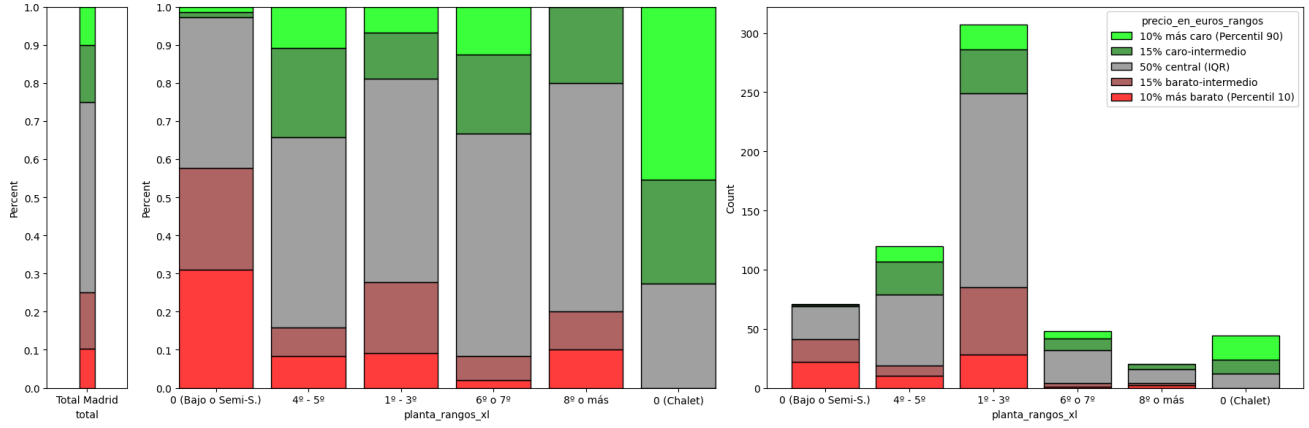
CHARACTERISTICS and PEE

The following composite graph analyzes price percentile distribution according to property type (`piso` / `chalet`).



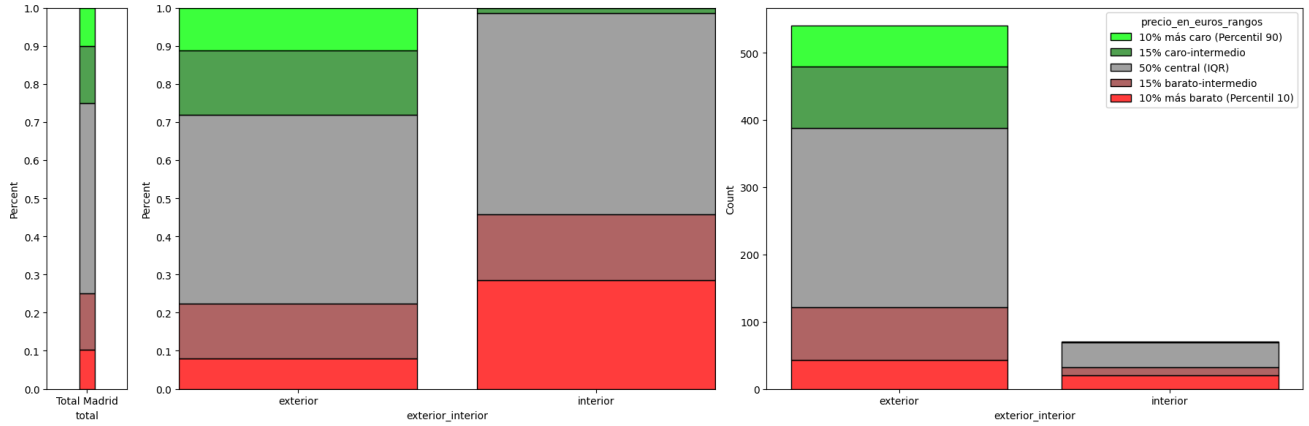
`Chalets` (44 obs.) show marked bias toward high prices: ~70% are situated in upper percentiles versus 25% population. This result is expected given that `chalets` combine greater surface with locations in certain-level residential areas. `Pisos` (566 obs., 92.8%) align almost perfectly with total distribution, covering all price ranges.

The following composite graph crosses `pee` percentiles with `floor` number.



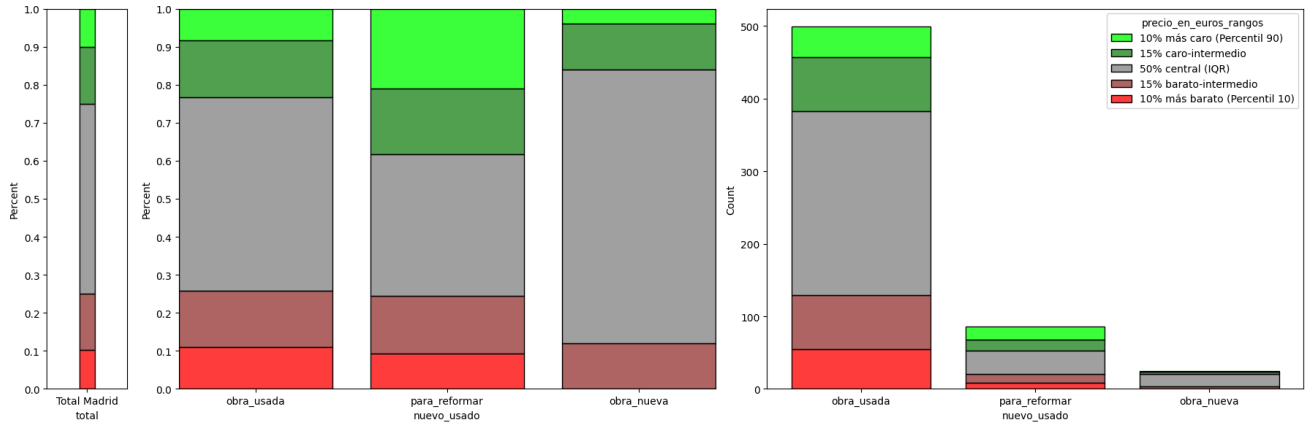
Floors `1st` - `3rd` (307 obs.) present balanced distribution between percentiles. Floors `4th` - `5th` (120 obs.) skew toward medium-high prices. `Ground floors` / `semi-basements` (71 obs.)

The following graph analyzes `pee` distribution according to orientation (`exterior` / `interior`).



`Exterior` homes (540 obs., 88.5%) dominate in all `pee` ranges. `Interior` homes (70 obs., 11.5%) have greater representation in low-medium percentiles, suggesting a price penalty for this characteristic that I will quantify in the Market section (`pmc`).

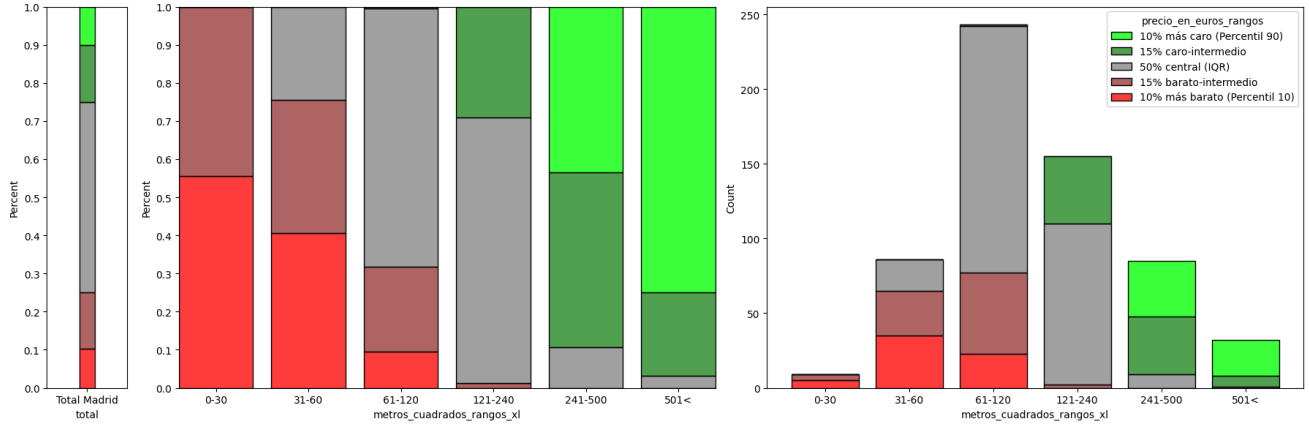
The following graph shows `pee` distribution according to conservation `condition` .



`Used_construction` (499 obs., 81.8%) covers all percentiles, reflecting its heterogeneity. Homes `to_reform` (86 obs., 14.1%) concentrate in medium-low ranges, which makes economic sense: the offer price should discount the necessary renovation cost. `New_construction` (25 obs., 4.1%) skews toward high percentiles, although the limited sample prevents robust conclusions.

SPACE and PEE

The following composite graph analyzes `pee` percentile distribution according to `surface` ranges.



The relationship between `surface` and `pee` is positive and progressive, as expected. Small ranges (0- `60 m²`) concentrate on low percentiles; range `61 m²` - `120 m²` (243 obs., 39.8%) skews toward medium-low percentiles; ranges `121 m²` - `240 m²` and `241 m²` - `500 m²` shift toward high percentiles; homes > `501 m²` concentrate almost exclusively in the highest percentile.

A relevant finding: medium ranges (`61 m²` - `240 m²`, ~65% of sample) present heterogeneous price distribution, indicating that other factors—location, characteristics, qualities—significantly modulate price beyond surface. This justifies the multivariate analysis that follows.

AMENITIES and PEE

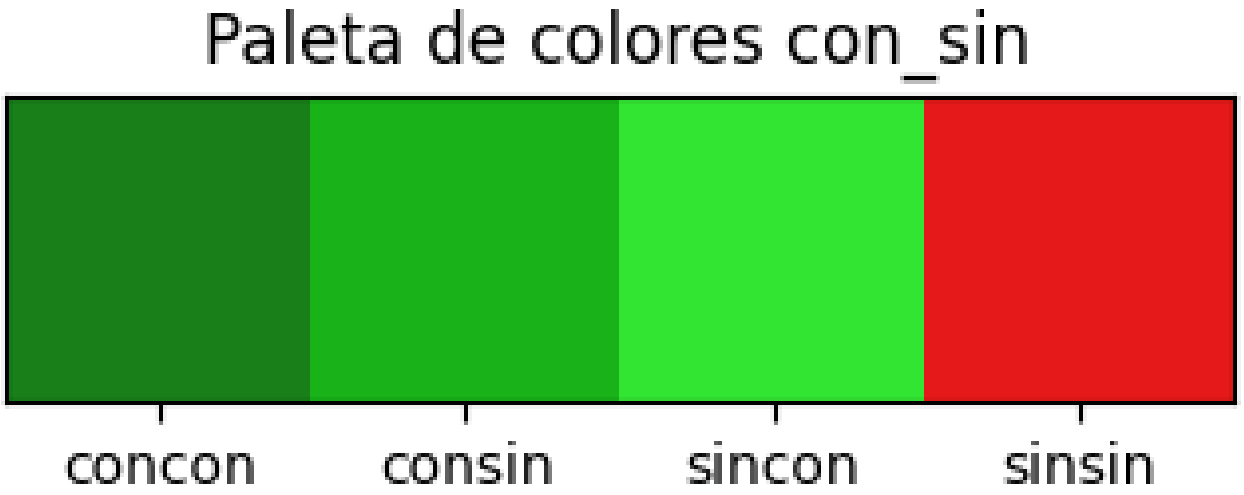
This section analyzes the relationship between `amenities` and `total price`. Amenities are coded as binary variables (has/doesn't have) and are analyzed through two complementary metrics: count (prevalence of each configuration in three price groups: `p10`, `IQR`, `p90`) and average `pee` of each group. This dual perspective allows distinguishing between amenities frequent in expensive homes versus amenities that raise average price.

For analysis, I group amenities into five categories.

Variable grouping:

- `elevator`
- `equip_services`: garage + storage room

- `equip_leisure_spaces`: garden + pool
- `equip_views`: balcony + terrace
- `equip_climate`: air conditioning + heating

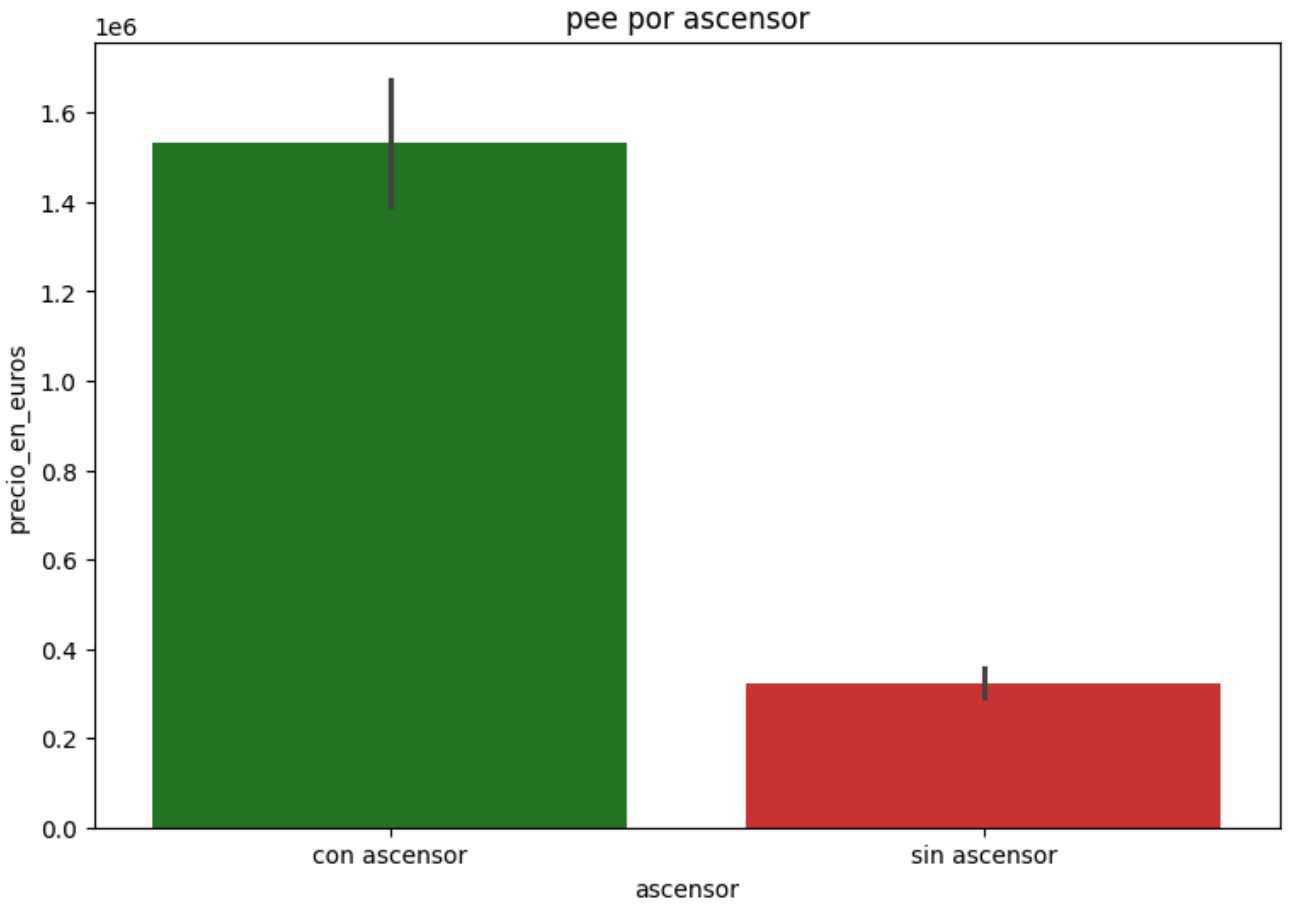


Grouped amenities are coded in four combinations: `with/with`, `with/without`, `without/with`, `without/without`. The graph's color palette follows a gradient from green (complete amenity) to red (without amenity). In some subgroups samples are small, which limits analysis robustness and advises interpreting observed differences with caution.

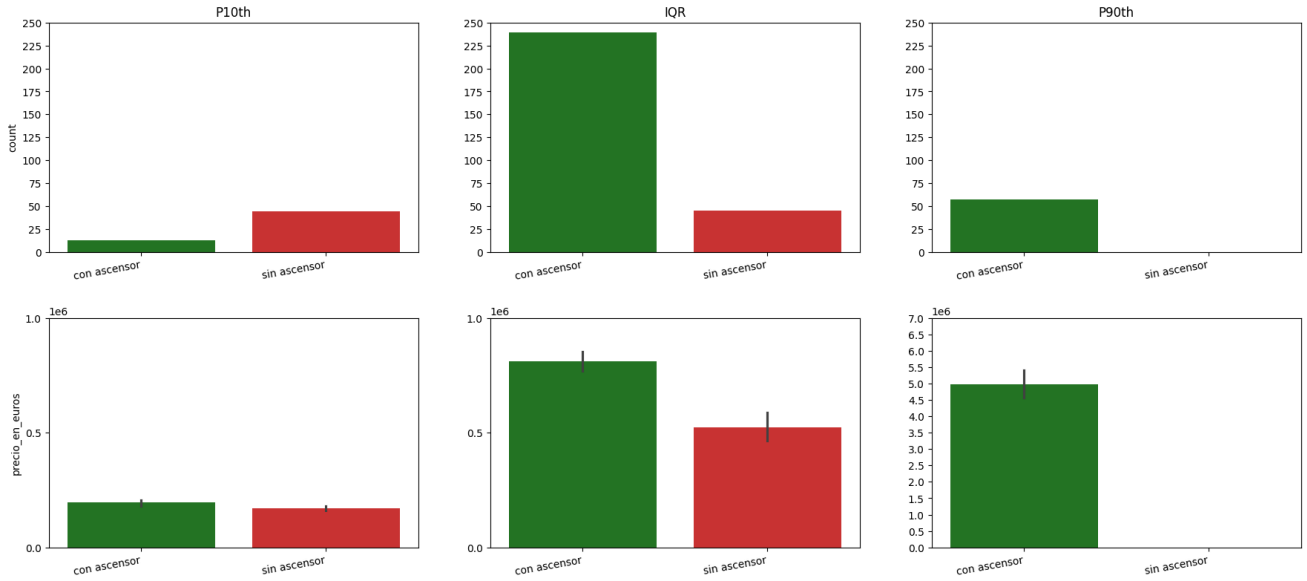
PEE / Elevator Amenity

Variables: `ascensor`

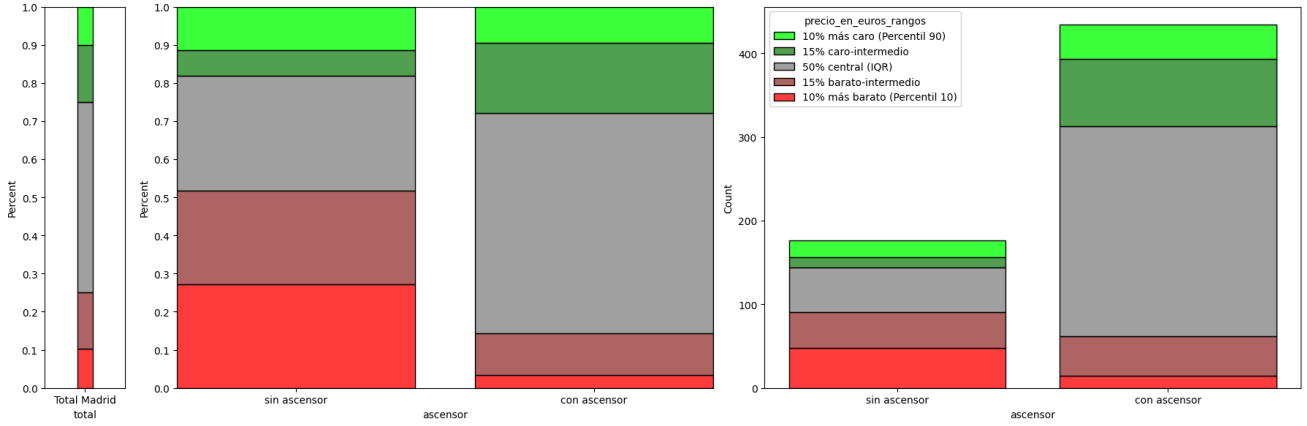
The following composite graph shows `pee` and distribution by percentiles according to `elevator` availability.



Homes `with elevator` present average `pee` of ~1.5M€ versus ~325K€ `without elevator`, a difference of almost 5x. This pronounced gap reflects, most probably, not only the intrinsic value of the amenity but its correlation with other characteristics: buildings with elevators tend to be more modern, be in better locations and have higher construction qualities. By percentiles, homes `without elevator` concentrate on p10 (~50 obs. vs ~15 with elevator), while p90 is composed almost exclusively of homes `with elevator`.



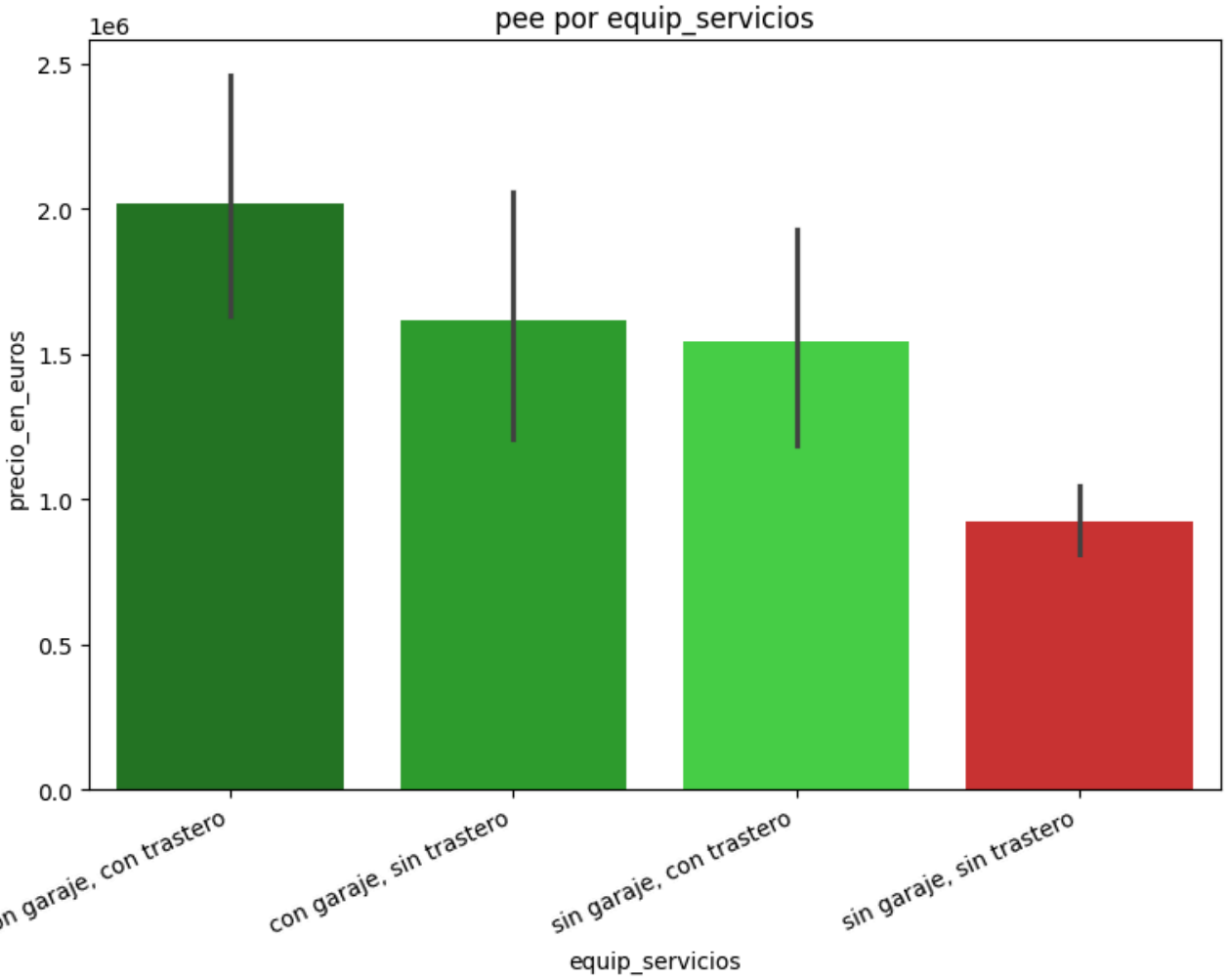
The proportion histogram confirms the pattern: homes `without elevator` concentrate most of their observations in low price percentiles (`p10` and cheap-intermediate), while homes `with elevator` present distribution shifted toward high percentiles, dominating practically alone `p90`. The elevator emerges as one of the variables with greatest price discriminating power.



PEE / Services Amenity

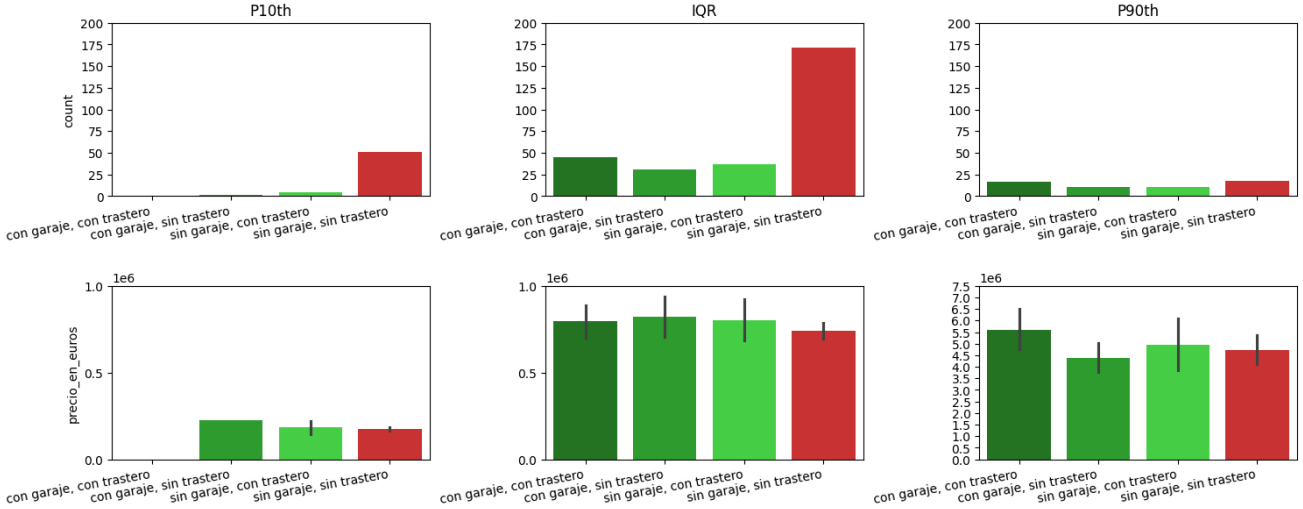
Variables: `equip_servicios`

The following composite graph shows `pee` and distribution by percentiles according to `garage` and `storage room` availability.

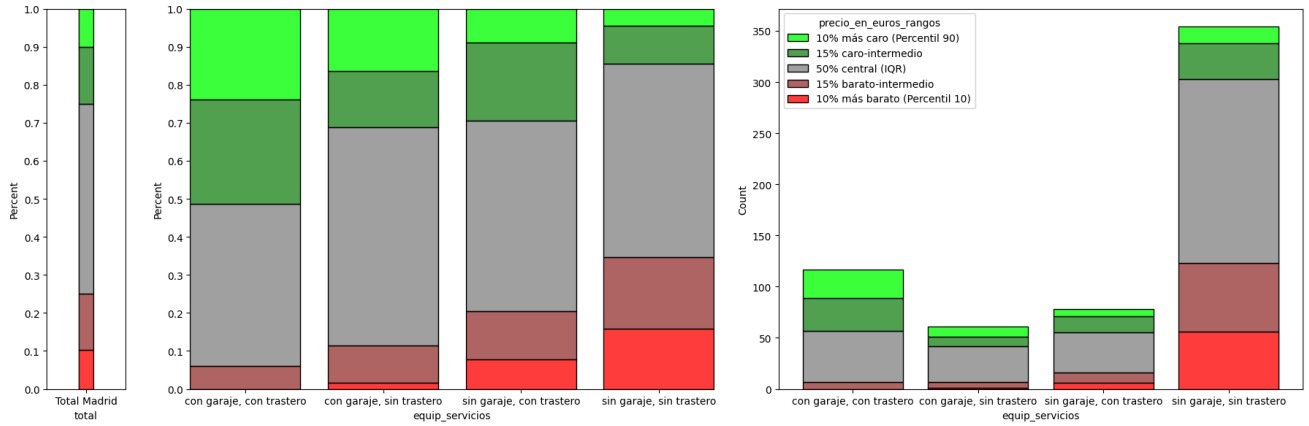


`Pee` decreases as services are reduced: ~2M€ in `with/with` (`garage` and `storage room`) down to ~1M€ in `without/without`, with high dispersions especially in categories with more amenities. The four categories have representation in all sections, with `without/without` being the most numerous.

Important methodological note: ads don't always specify whether `garage` and/or `storage room` are included in price (`pee`) or are offered as optional extras. This heterogeneity in coding can attenuate real differences between categories, as homes classified as "without garage" could have it available as an option not reflected in published price. Results of this variable should be interpreted with this limitation present.



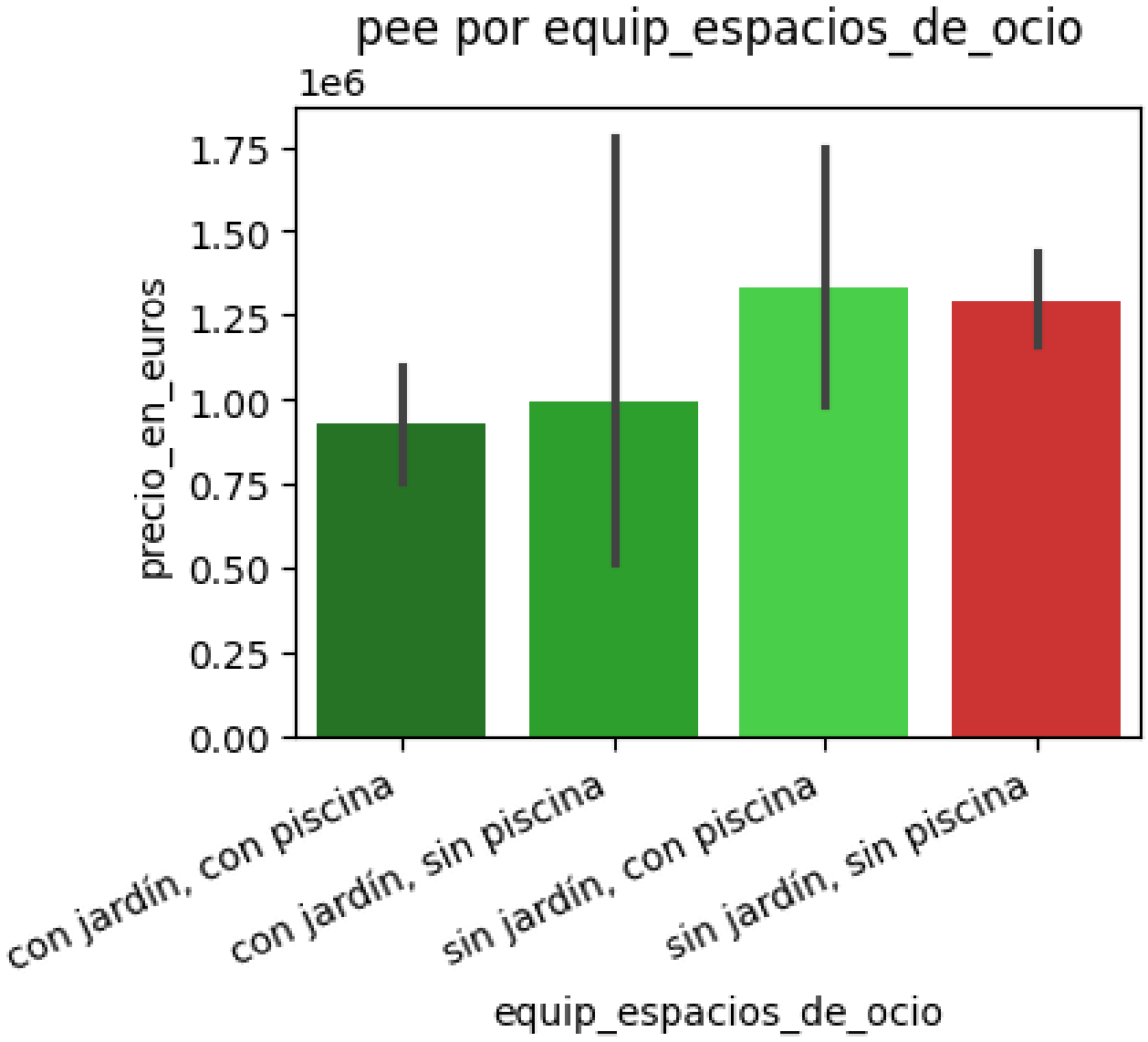
The proportion histogram shows gradual evolution: `with/with` presents greater relative weight in high percentiles, a pattern that gradually attenuates until `without/without`, where concentration in low percentiles increases. Differences are less pronounced than in `elevator`, consistent with lower `pee` dispersion between categories and with the mentioned methodological limitation.



PEE / Leisure Amenity

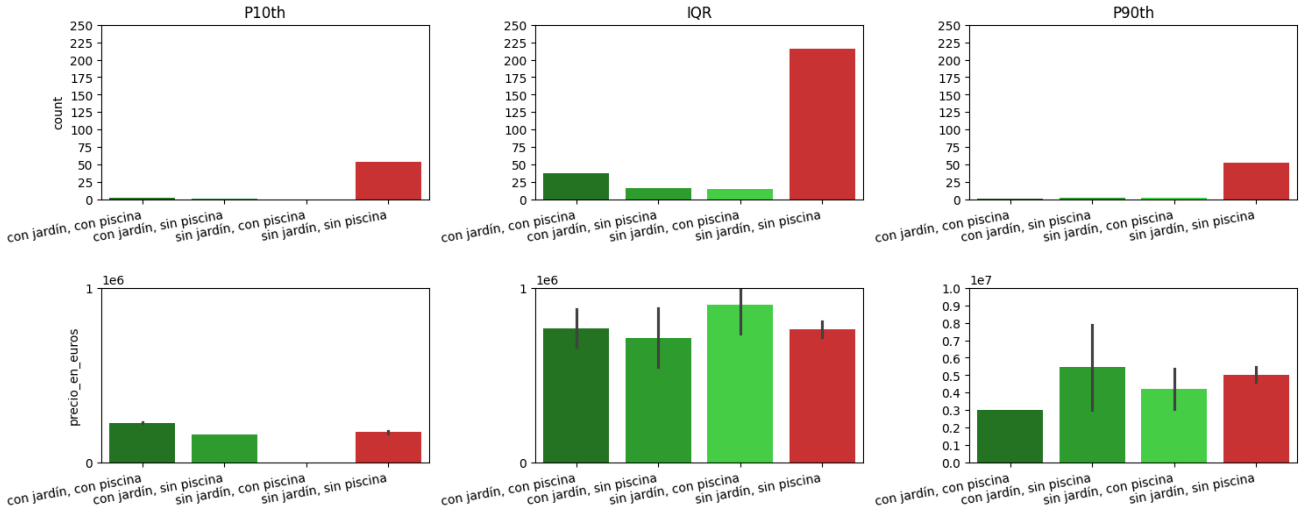
Variables: `equip_espacios_de_ocio`

The following composite graph shows average price and distribution by percentiles according to garden and pool availability.

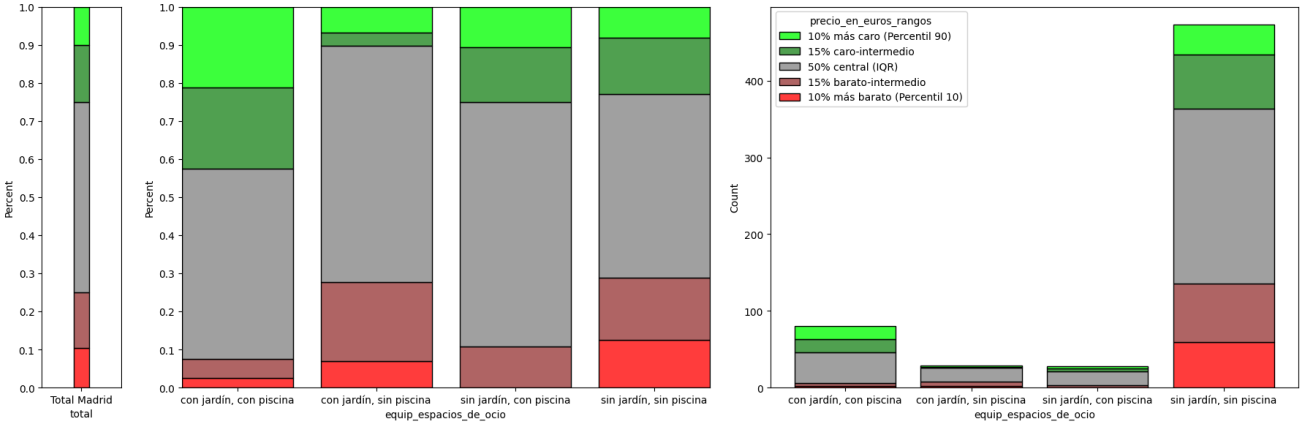


The pattern is counterintuitive at first sight: `pee` increases as amenities decrease, going from ~0.9M€ in `with/with` to ~1.25M€ in `without/without`. However, this result clearly illustrates the effect of a confounding variable: `garden` and `pool` are more frequent in peripheral zones (`chalets`, `developments`) where price per square meter is systematically lower.

The negative correlation `amenity` - `price` doesn't imply that these elements subtract value from the home; it reflects their unequal geographical distribution. This pattern underlines a fundamental limitation of bivariate analysis: the observed association between two variables can be mediated or confounded by a third. Dispersion is high in all categories; `without/without` dominates in count and in all percentile sections.



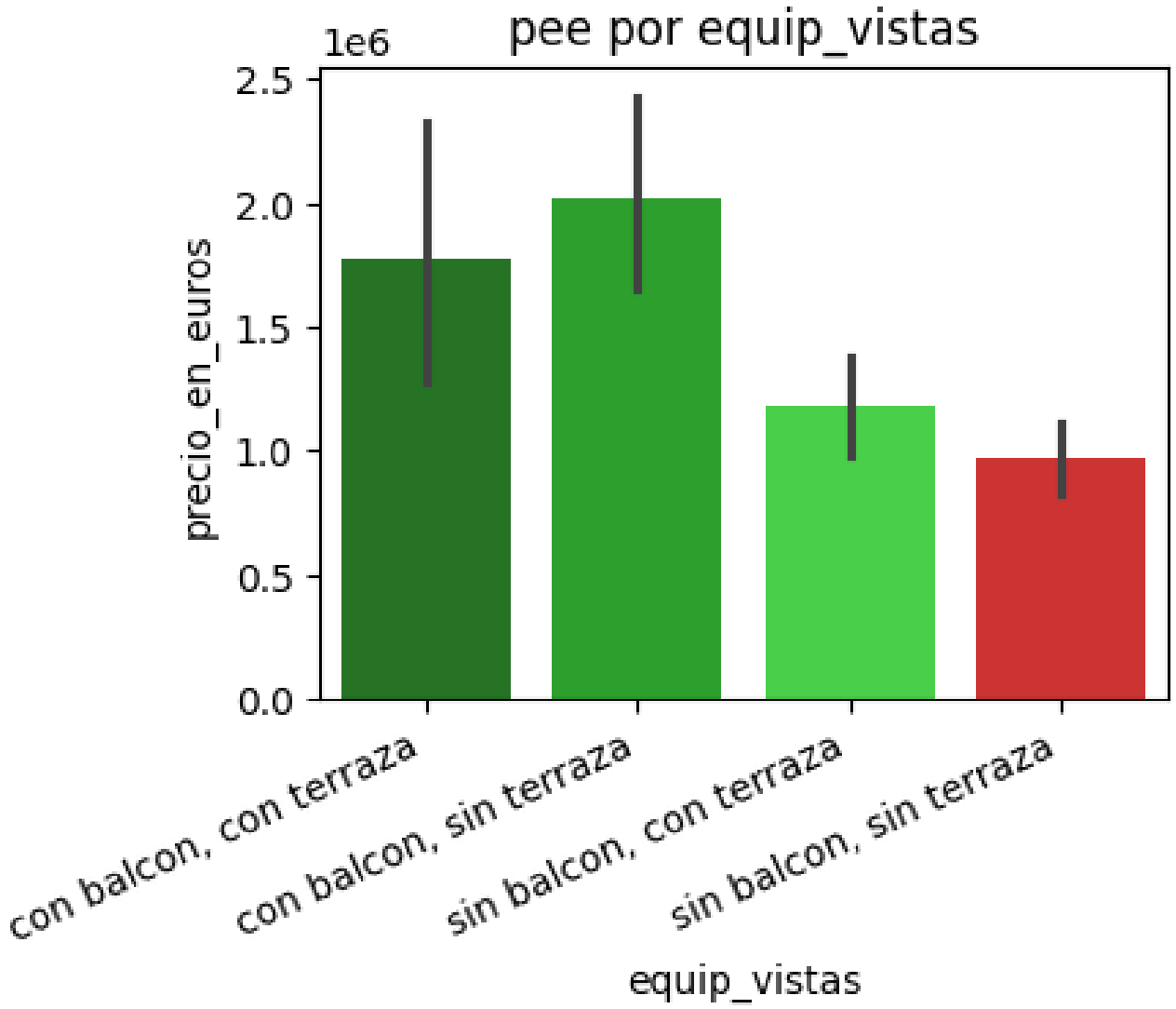
The proportion histogram reflects the inverse pattern: `without/with` (`with pool without garden`) shows greater concentration in high `pmc` percentiles, while `with/without` (`with garden without pool`) presents greater weight in low percentiles. `Without/without`, despite dominating in absolute count, shows distribution close to population. This pattern is consistent with the hypothesis that garden and pool correlate with peripheral location of lower `pmc`.



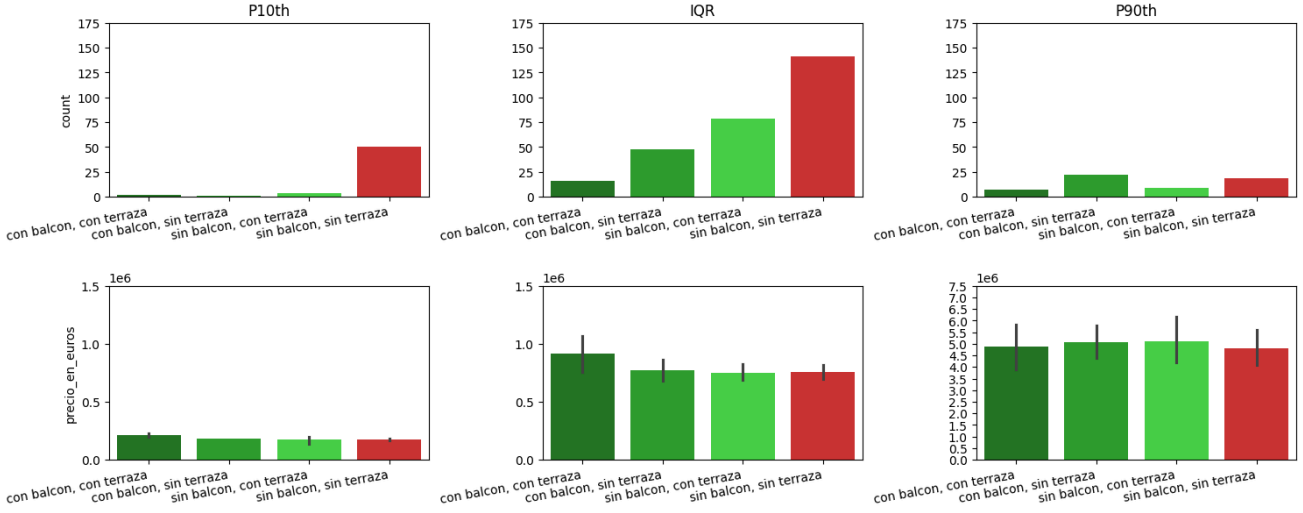
PEE / Views Amenity

Variables: `equip_vistas`

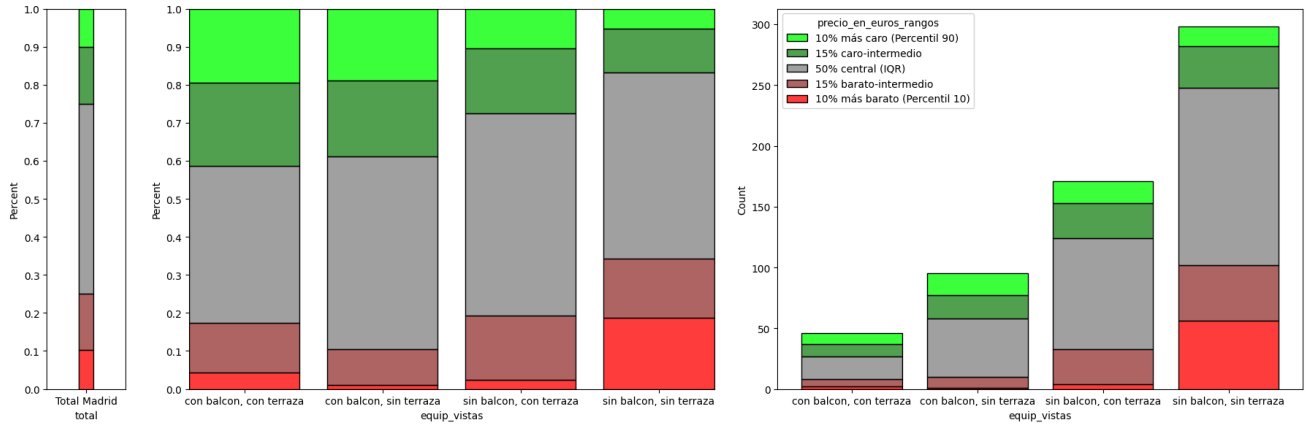
The following composite graph shows `pee` and distribution by percentiles according to `balcony` and `terrace` availability.



`Pee` is higher in categories with balcony: `with/without` (balcony without terrace, ~2.0M€) and `with/with` (~1.75M€), both with high dispersion. Categories without balcony present lower values: `without/with` (terrace without balcony, ~1.2M€) and `without/without` (~1.0M€). Balcony seems to have greater association with high price than terrace, possibly because balcony is more frequent in central apartments of historic buildings, while terrace (without balcony) usually associates with penthouses or peripheral homes. The `without/without` category dominates in absolute count, especially in IQR.



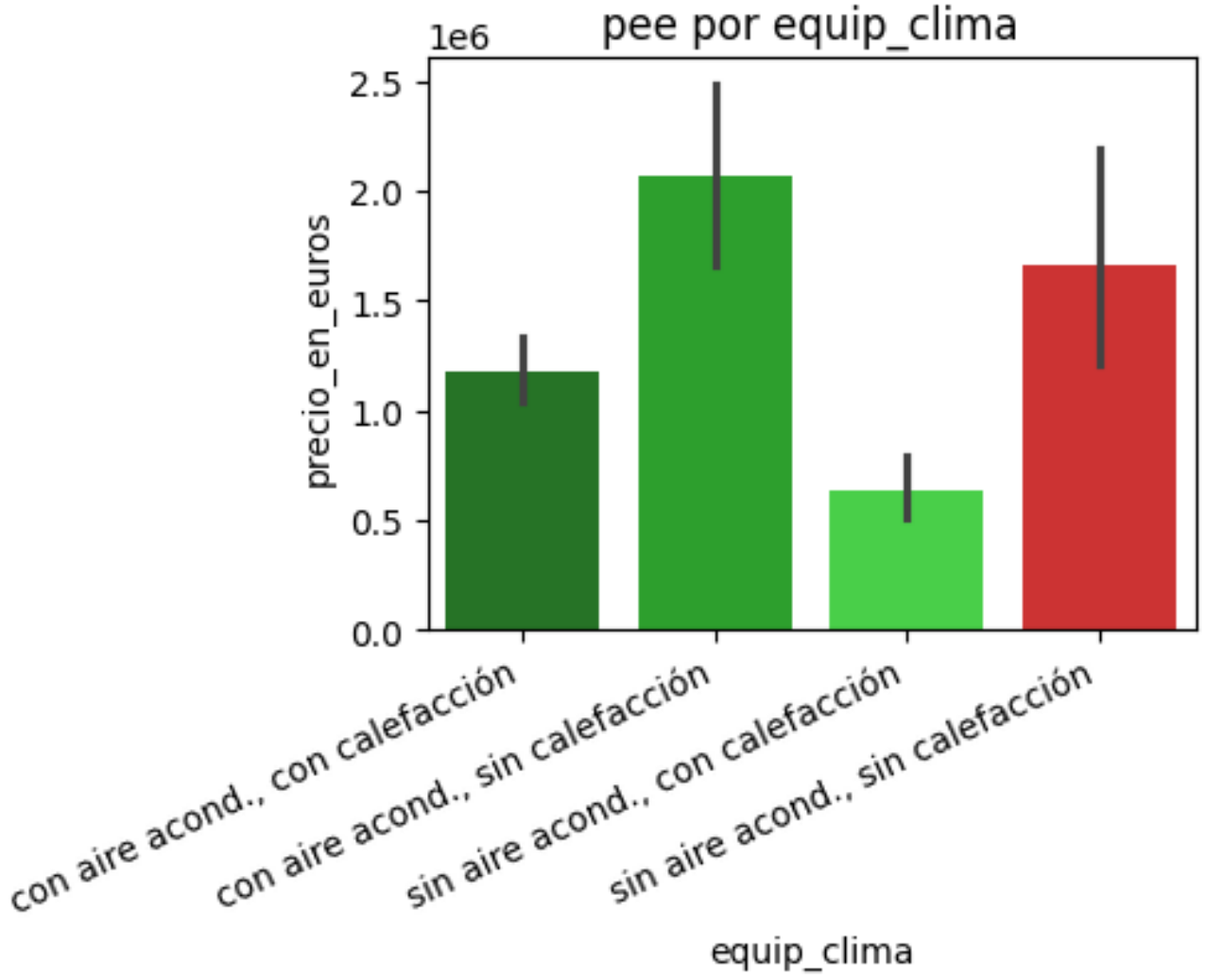
The proportion histogram confirms that categories `with balcony` (`with/with` and `with/without`) concentrate greater weight in high `pmc` percentiles, especially `with/without` which dominates `p90`. Categories `without balcony` (`without/with` and `without/without`) present distributions shifted toward low-intermediate percentiles.



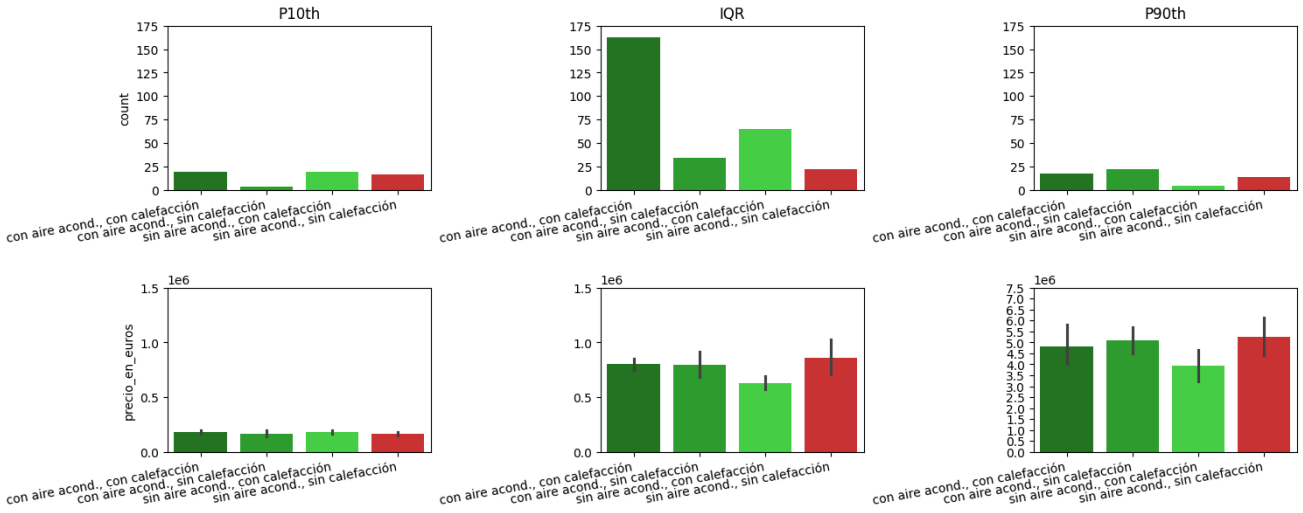
PEE / Climate Amenity

Variables: `equip_clima`

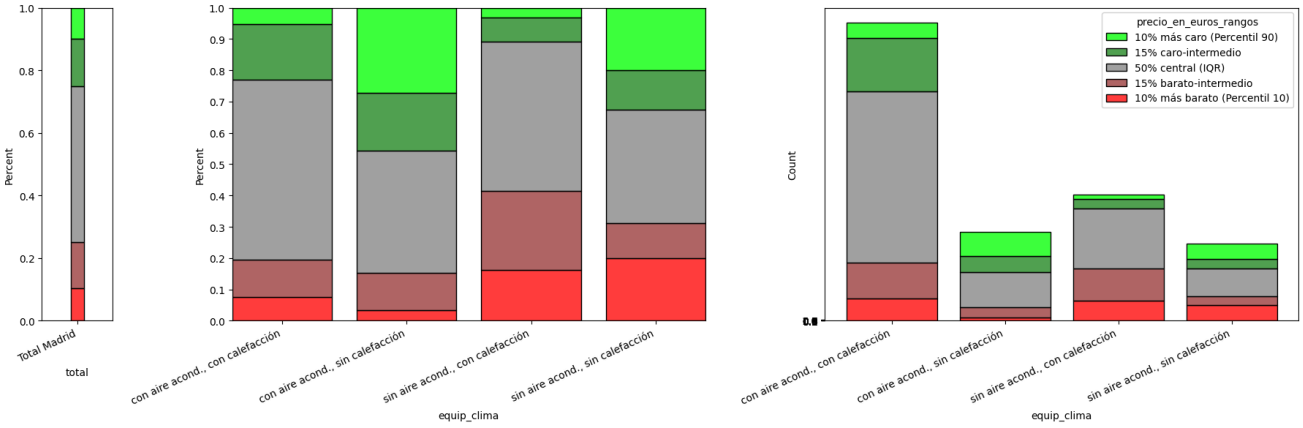
The following composite graph shows `pee` and distribution by percentiles according to `air conditioning` and `heating` availability.



The pattern is not monotonic: `with/without` (`air` without `heating`) presents the highest `pee` (~2.0M€), followed by `without/without` (~1.6M€, high dispersion), while `without/with` (`heating` without `air`) has the lowest (~0.65M€). This result suggests that `air conditioning` has greater association with high price than `heating`. The plausible interpretation is that `air conditioning` functions as an indicator of property quality/modernity, while `heating` is more universal and doesn't discriminate between segments. The `without/without` category is the most numerous.



The proportion histogram shows clear segregation: `with/without` (with `air`, without `heating`) concentrates the highest proportion of homes in `p90`, while `without/with` (with `heating`, without `air`) dominates in `p10`. The `without/without` category presents more balanced distribution between percentiles. This pattern reinforces that `air conditioning` is a better predictor of high price than `heating`, a finding I will also verify in `pmc` analysis (section 3).



3. Market (PMC)

This section analyzes `pmc` and its relationship with the characteristics described in the previous phase. Unlike total price (`pee`), `pmc` is a normalized metric that allows comparing market value regardless of property `size`.

Central question: What variables are most associated with `pmc` differences in the Madrid market?

The approach combines univariate analysis (`pmc` / `characteristic` correlation), segmentation (home profiles by `pmc` range) and geographical analysis (`pmc` distribution by `zone` and `subzone`).

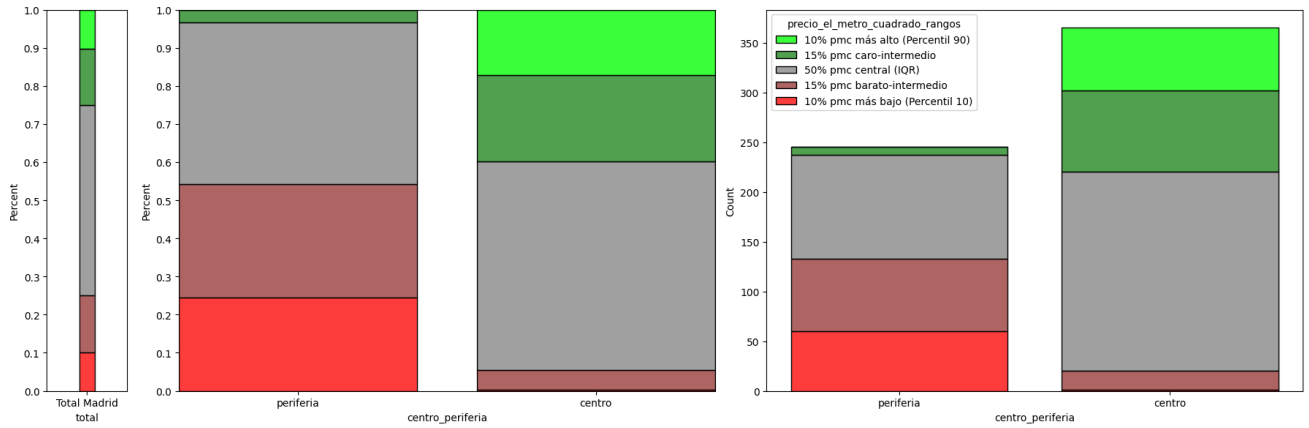
The comparison between `pee` (previous section) and `pmc` (this section) allows obtaining the complete picture: `pee` indicates absolute value and product typology; `pmc` indicates normalized market valuation. As we will see, some variables that discriminate strongly in `pee` have attenuated or even inverse effect on `pmc`, which reveals valuable information about market structure.

LOCATION

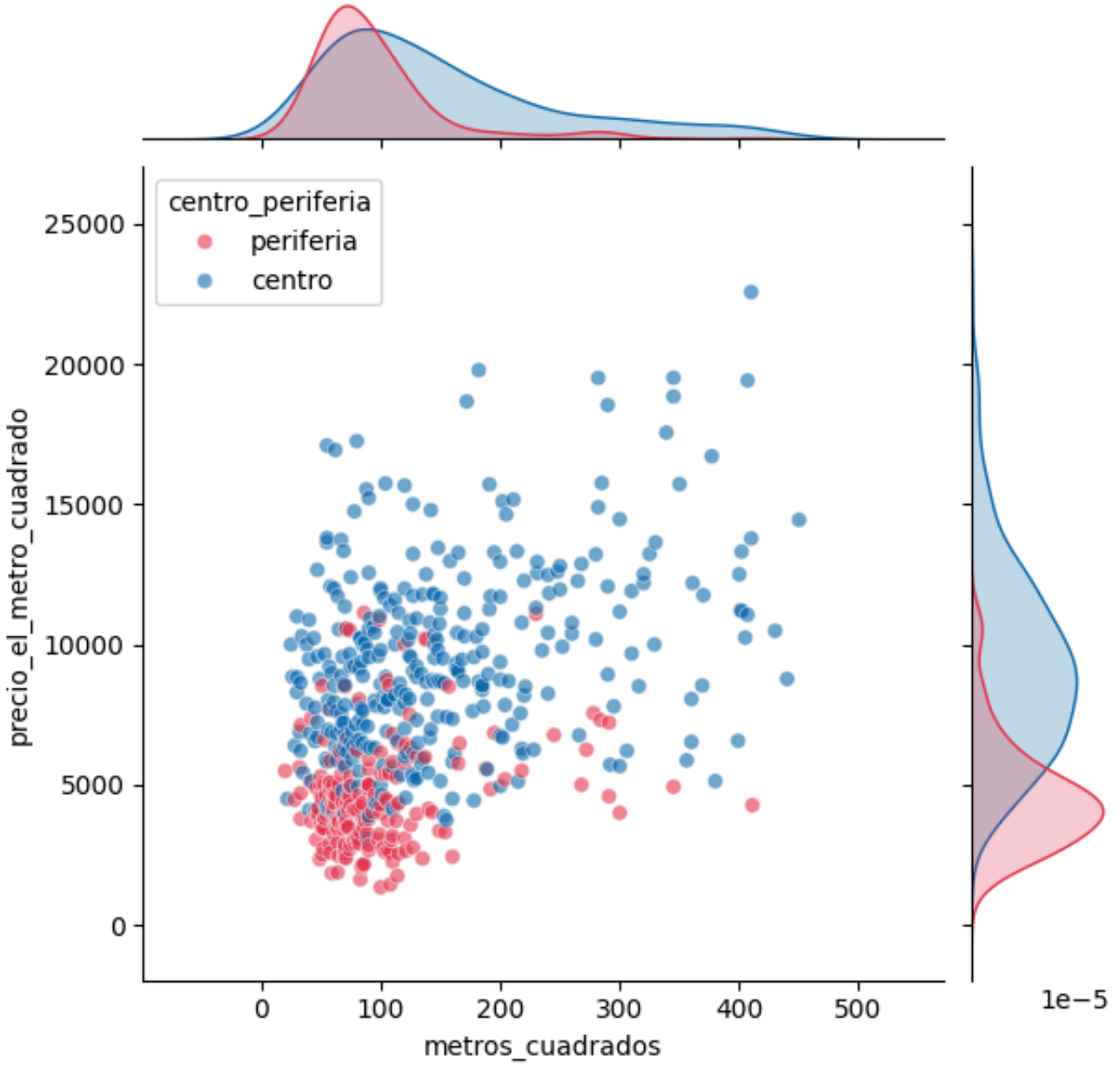
PMC / Center or Periphery

Variables: `pmc` + `centro_periferia`

The following composite graph shows `pmc` `percentile` distribution by location (`center` / `periphery`), along with a scatter diagram relating surface and `pmc`.

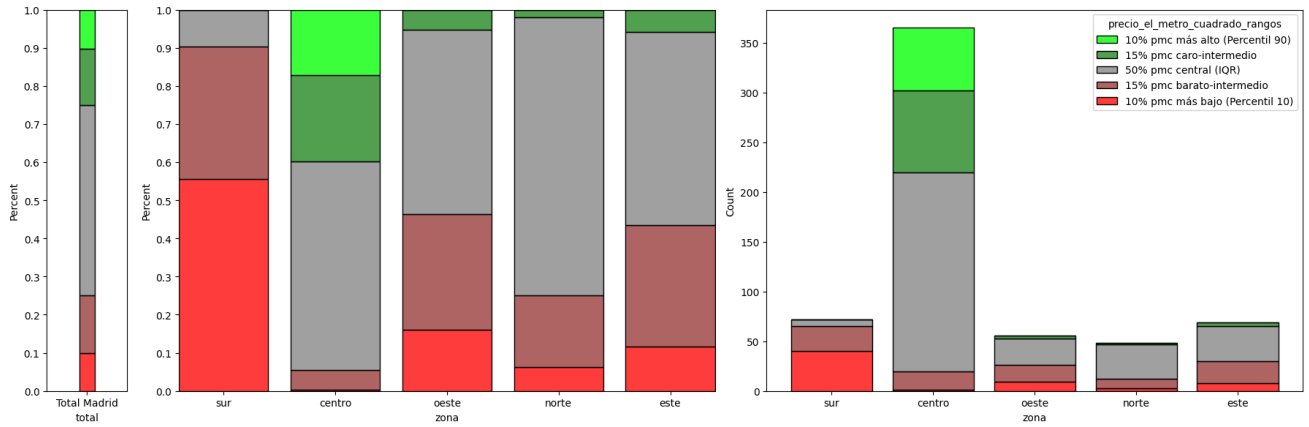


The `center` / `periphery` division is notably accentuated compared to `pee`. A resounding fact: the totality of homes in p90 of `pmc` is in the `center` (literally 0 in `periphery`); the inverse pattern occurs in low percentiles. This confirms that location is the main determinant of price per square meter (`pmc`) in Madrid.

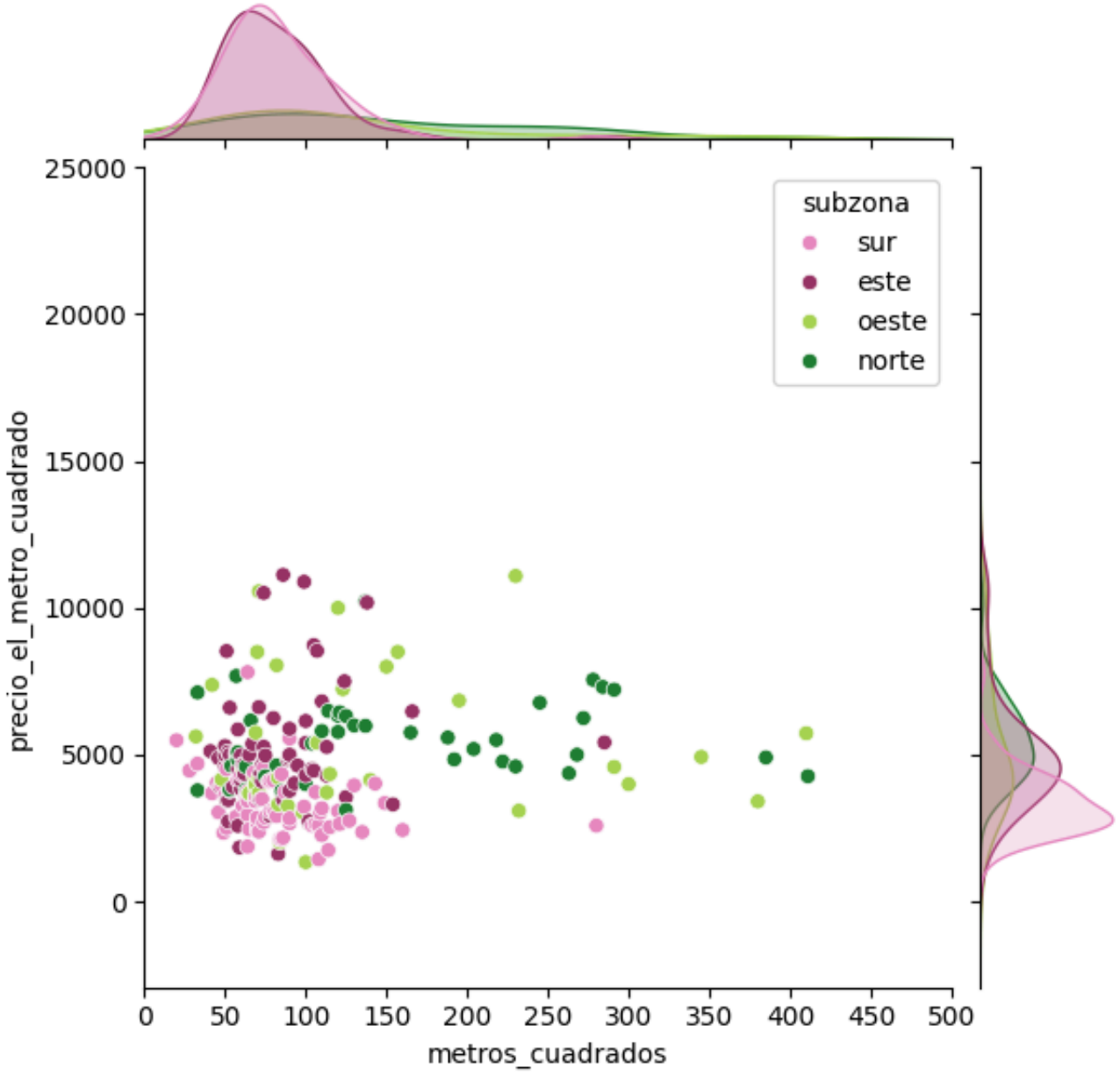


PMC / Zone

The following composite graph breaks down pmc distribution by `zones`, including scatter diagrams (`surface` vs `pmc`) for `periphery` and `center`, and comparative boxplots.

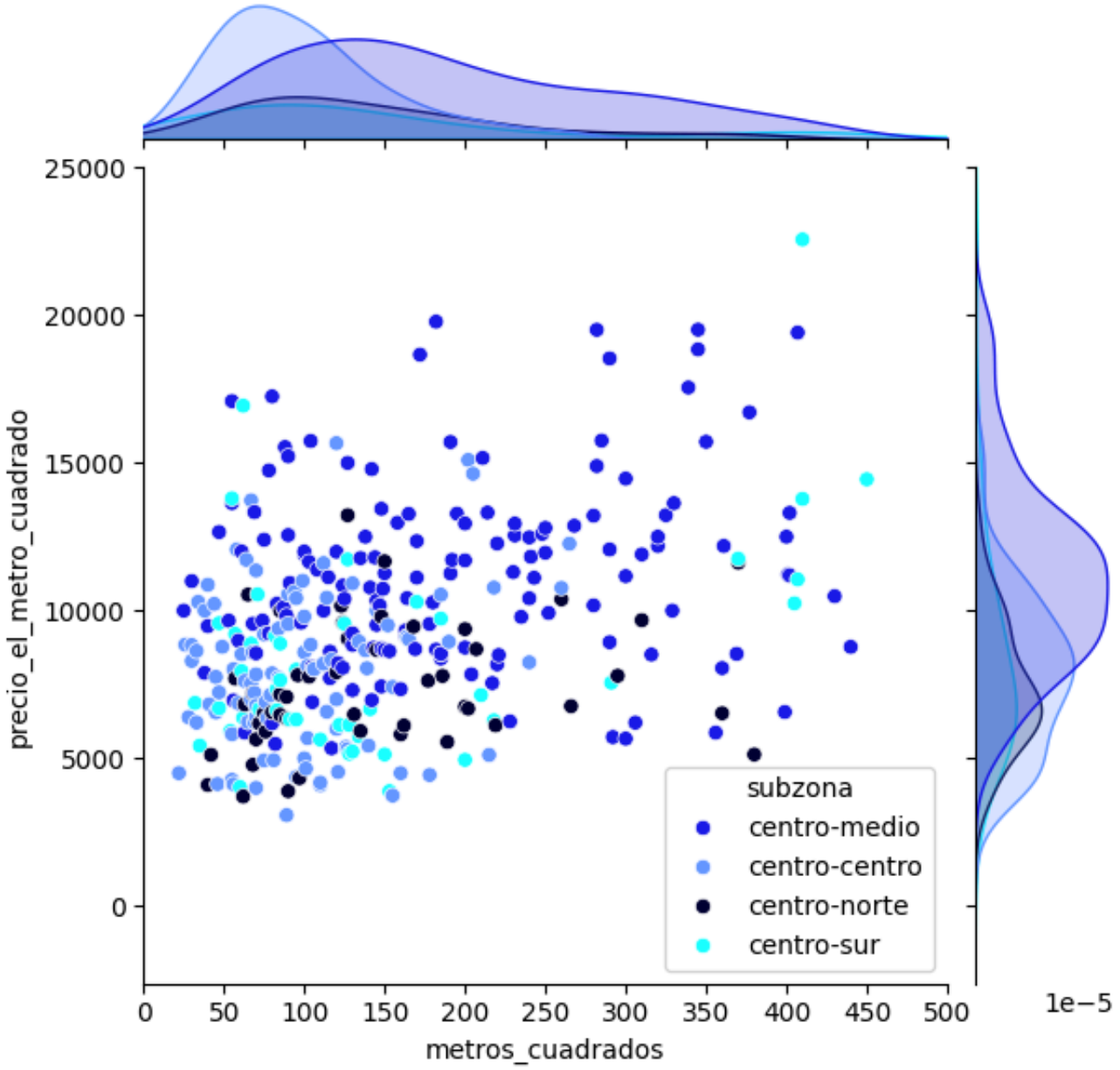


When disaggregating by `zones`, the pattern becomes more precise. The increase of `p90` and decrease of `p10` concentrates exclusively in the `center`. `West` and `north`, which showed presence in high `pee` percentiles, lose all their `p90` homes in `pmc` —their high `pee` responded to surface, not to premium valuation per square meter. `South` slightly increases its presence in `p90`, while east reduces `p90` and increases its intermediate-high band.



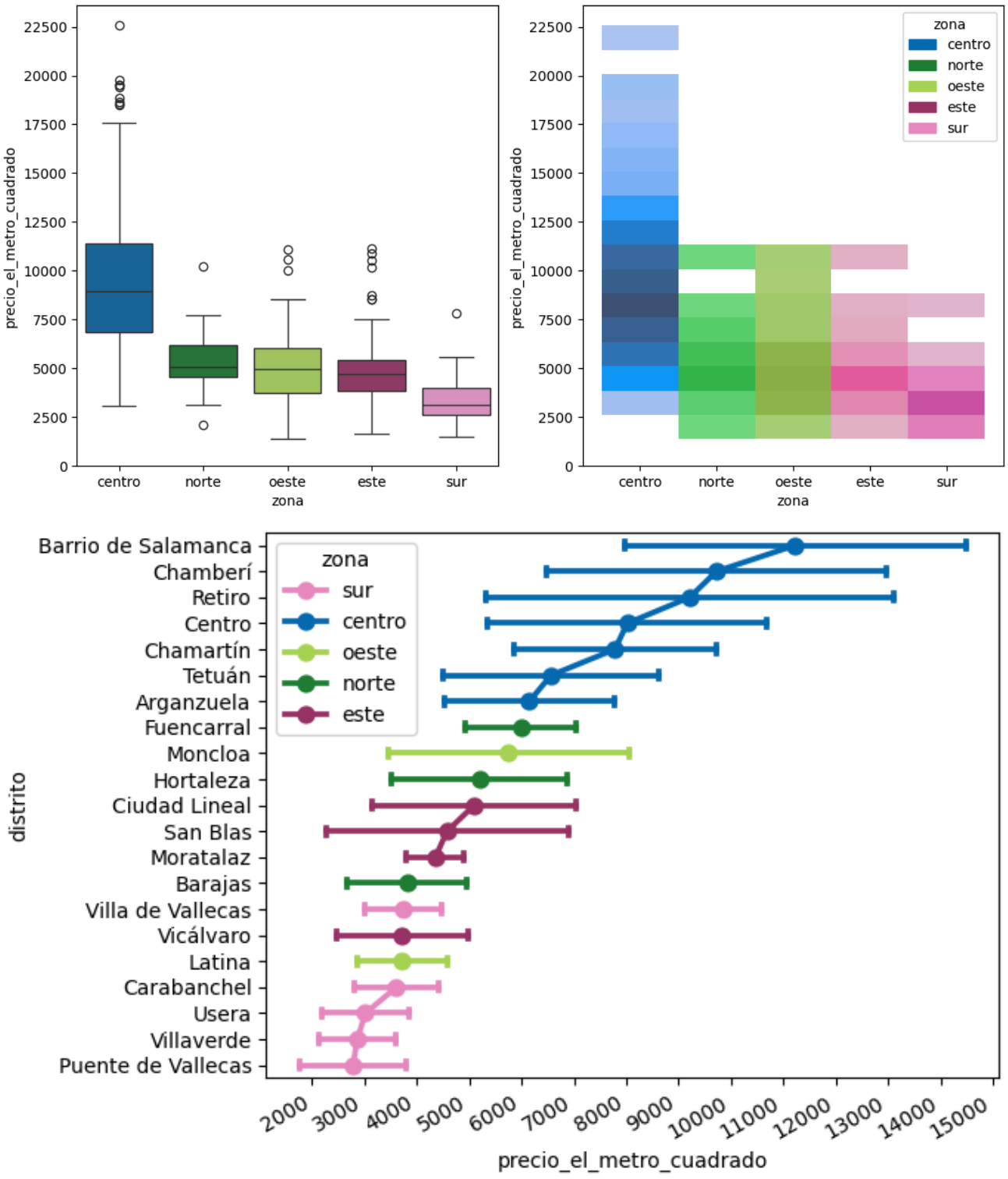
The previous graph shows `pmc` / `surface` dispersion for peripheral zones. Concentration in low-medium `pmc` ranges (`2,000€/m²` - `8,000€/m²`) is evident, with scarce presence above `10,000€/m²` regardless of `surface` .

The following graph shows the same analysis for center subzones.

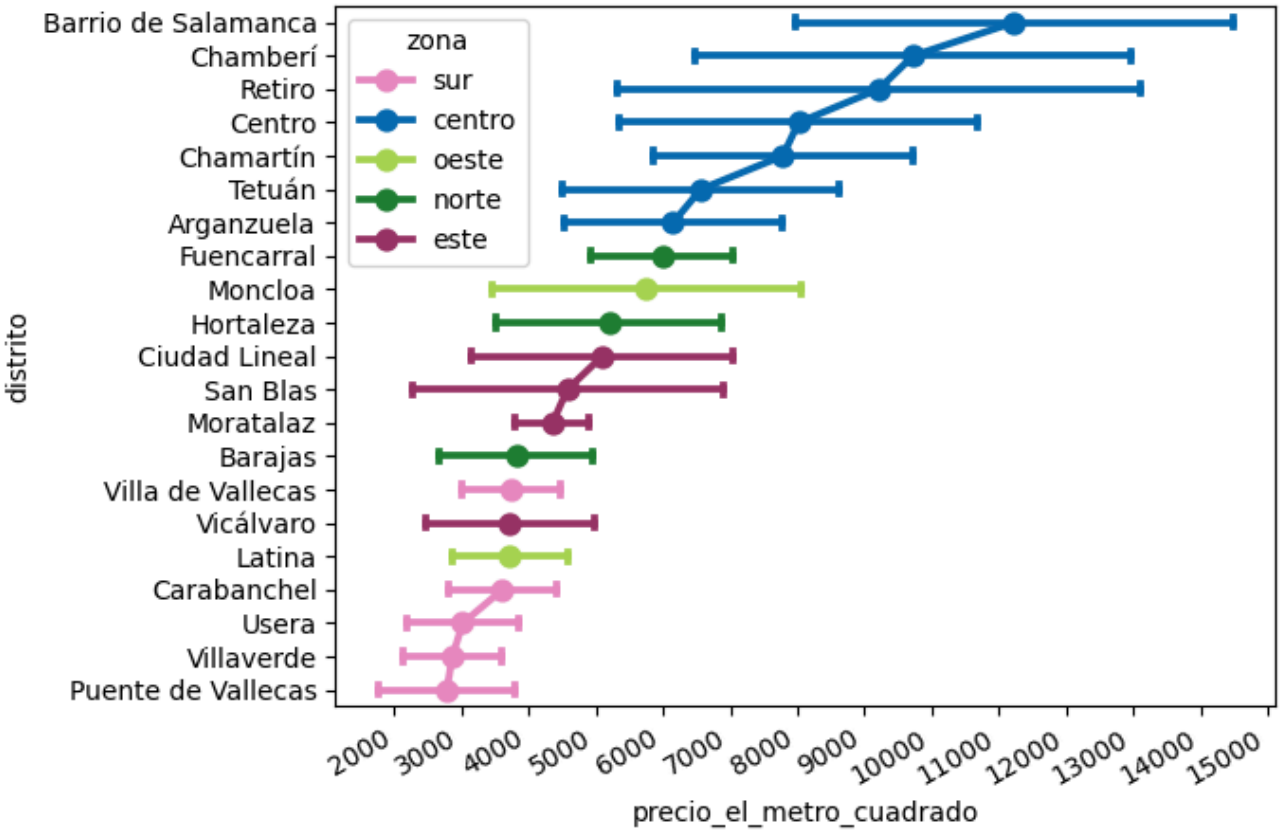


The contrast with `periphery` is marked: the `center` shows greater vertical dispersion (wider `pmc` range for similar surfaces) and reaches values above `20,000€/m²` . `Center-medium` presents the highest values and greatest variability.

The following boxplots compare `pmc` distribution by zone, complemented with a density heatmap.



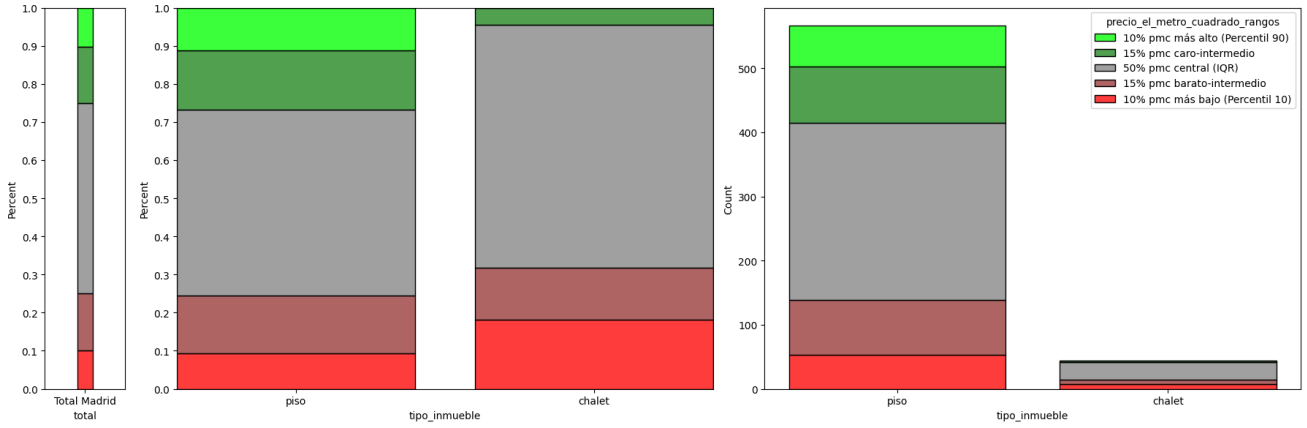
The following graph shows average `pmc` by `district`, with confidence intervals and coding by `zone`. It allows identifying the `price` hierarchy at `district` level, although some present reduced samples that limit estimation robustness.



CHARACTERISTICS

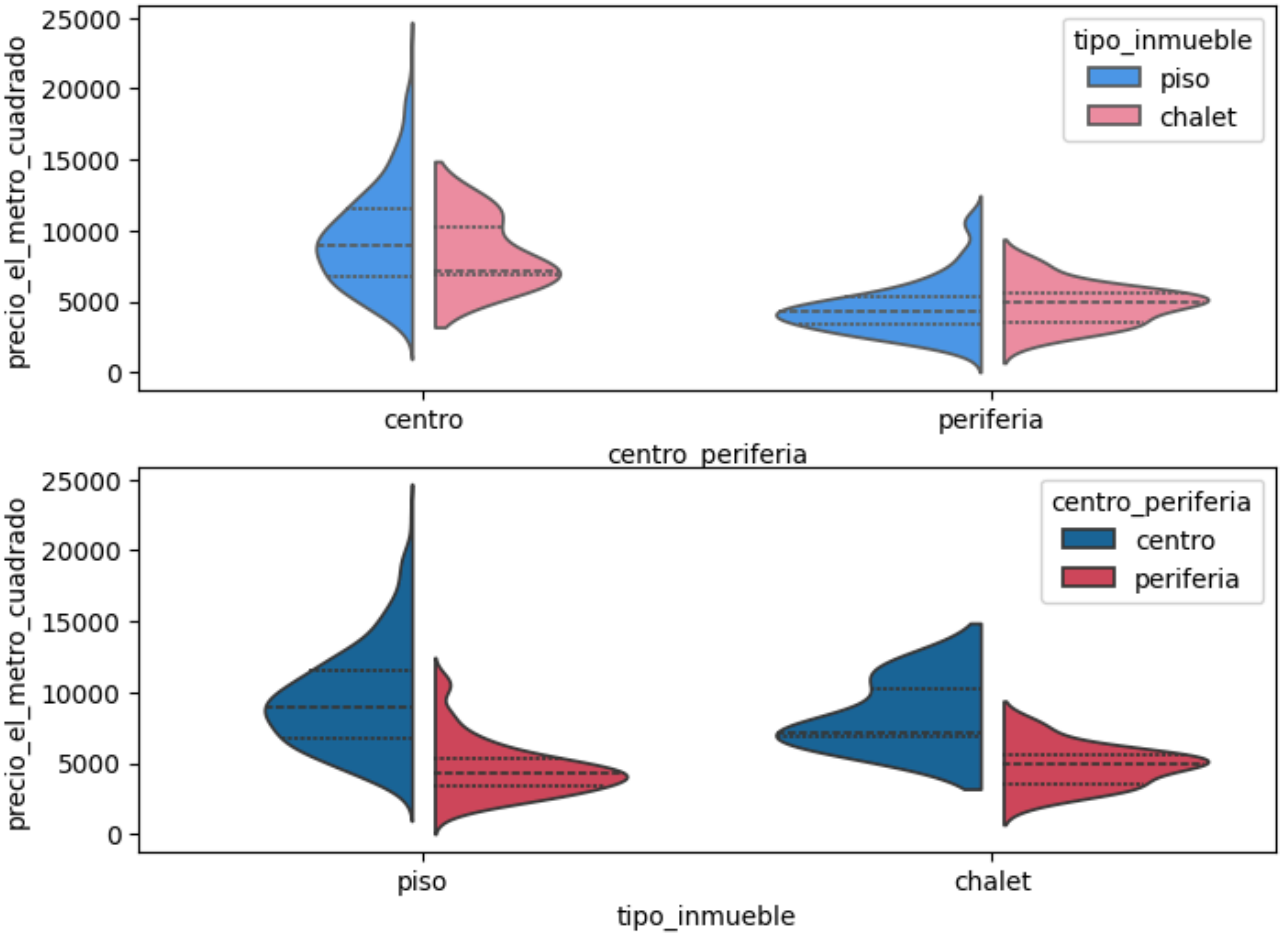
PMC / Property type

The following composite graph analyzes `pmc` distribution according to property type (`piso` / `chalet`), including density histograms by location.



`Chalets` show inverse behavior in `pmc` compared to `pee`: although they concentrated 70% of their observations in high `pee` percentiles, in `pmc` they have overrepresentation of low values and literally 0 homes in `p90`.

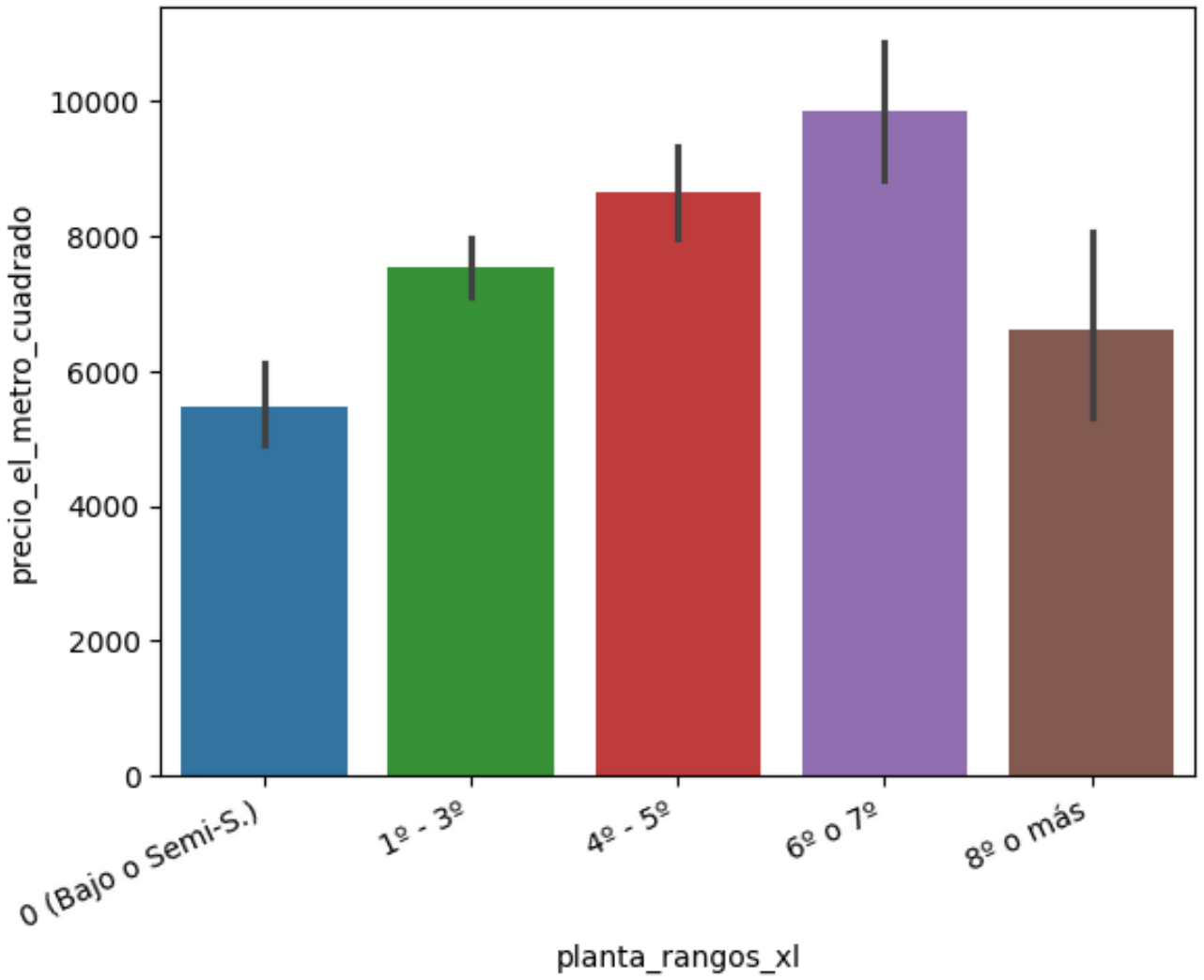
This contrast is revealing: it confirms that the high total price of `chalets` responds to their large surface, not to premium valuation per square meter. In fact, the average `pmc` of `chalets` is lower than `pisos`, a direct consequence of their location in peripheral zones where land is cheaper. `Pisos` align almost perfectly with total distribution in both metrics, which is expected given their dominant sample weight (92.8%).



PMC / Floor number

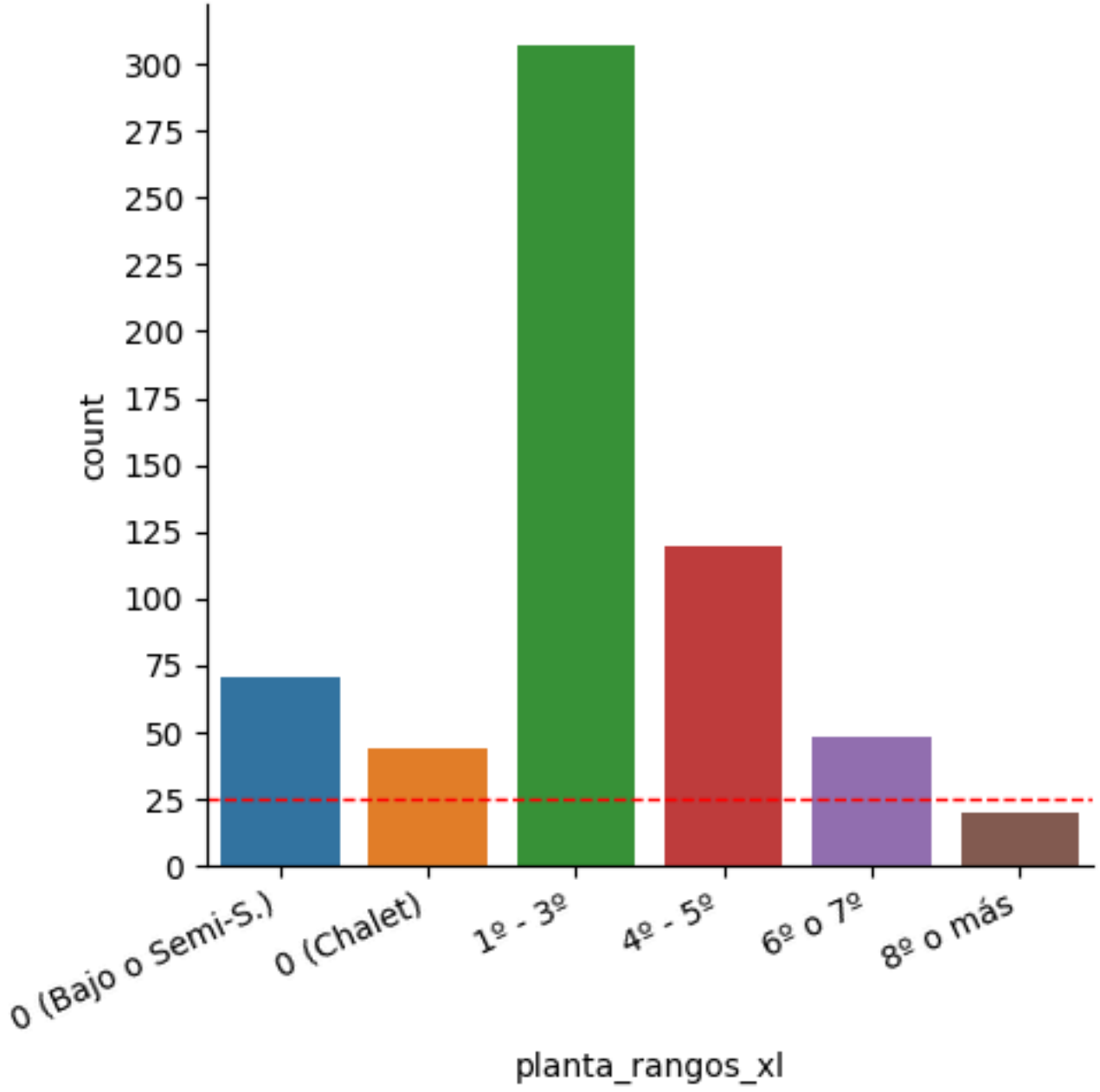
Variables: `planta`, `planta_rangos(_xl)` + `pmc`

The following composite graph shows average `pmc` by floor range and percentile distribution.

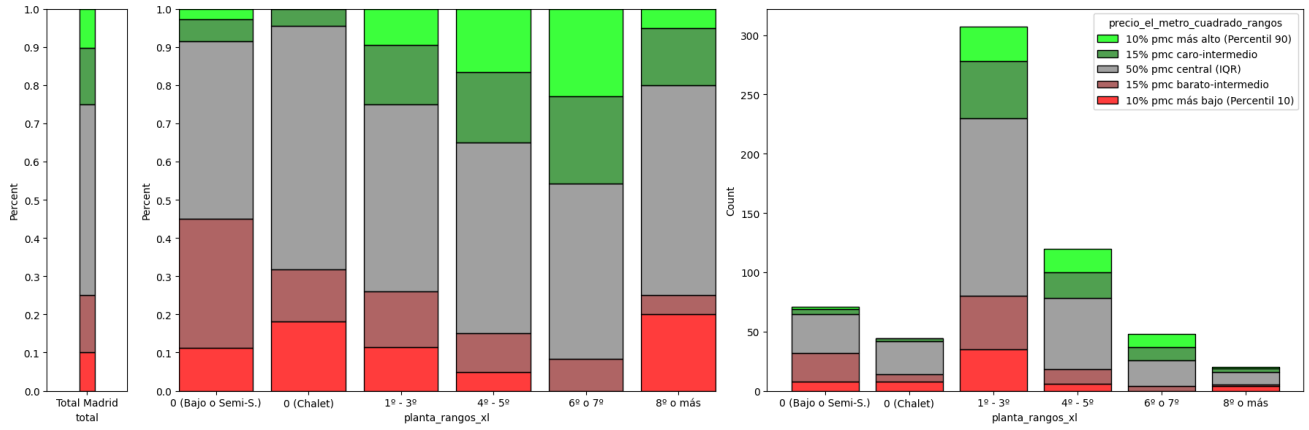


`Pmc` shows positive relationship with height, although not linear. `Ground floors` / `semi-basements` present the lowest `pmc` (~ `4,800€/m²`), followed by `chalets` (~ `6,600€/m²`). Intermediate floors oscillate between `7,500€/m²` - `8,700€/m²`, with peak on floors 6th-7th (~ `10,000€/m²`).

Floors `8th` + descend to ~ `6,600€/m²`, a result that should be interpreted with caution due to reduced sample (<20 obs.). This descent could be statistical artifact, or reflect that very high floors in Madrid don't always correspond to premium product—there are social housing buildings in height, for example, that would reduce the segment average.



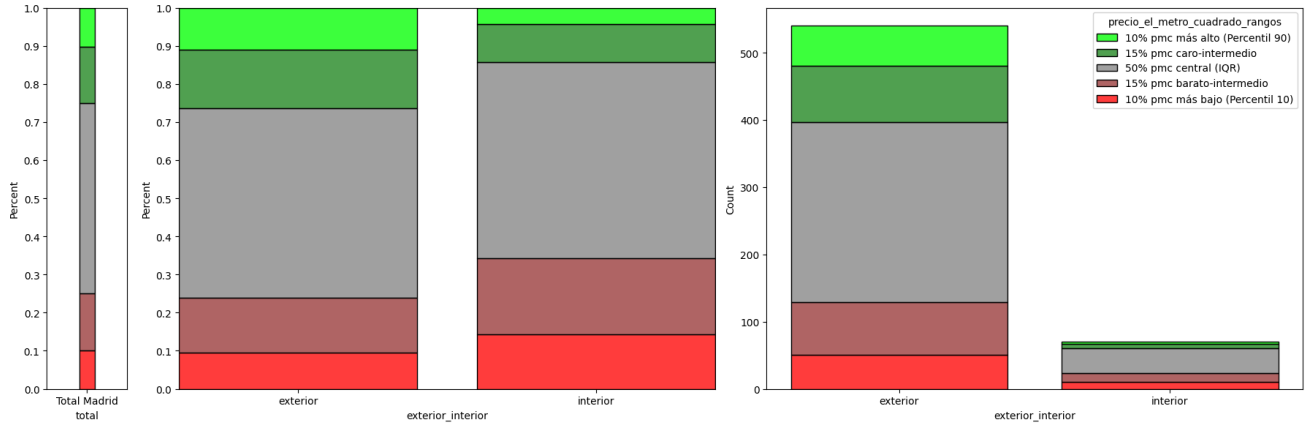
The counts graph confirms that floors `1st–3rd` dominate the sample (~307 obs.), while `ground floors`, `chalets` and `8th` + have limited representation (<75 obs. each), which advises prudence in conclusions about these segments.



PMC / Exterior/Interior

Variables: `exterior_interior` + `pmc`

The following composite graph analyzes `pmc` distribution according to orientation (`exterior` / `interior`).



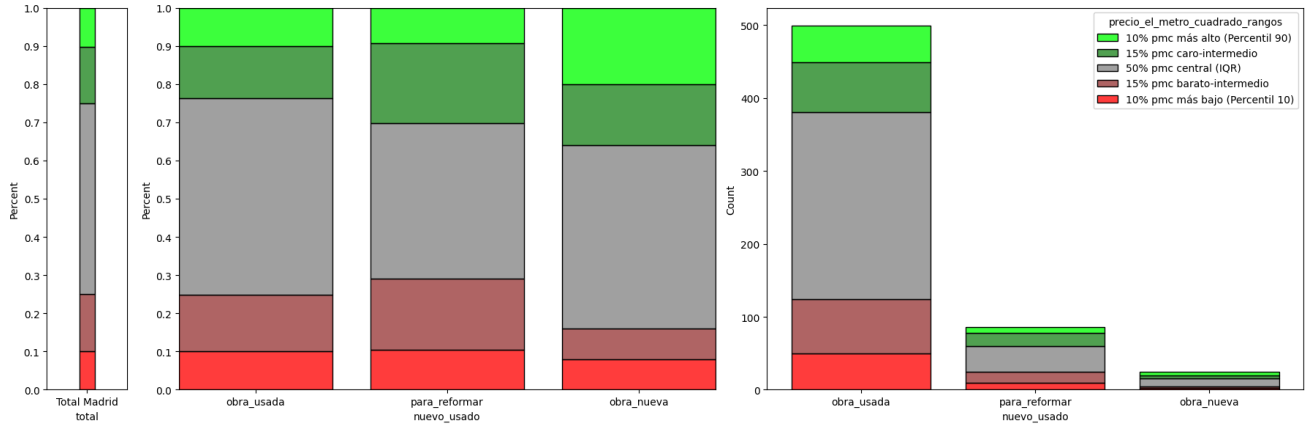
An interesting finding: `interior` apartments slightly increase their presence in high `pmc` percentiles, approaching parity with `exterior`.

This suggests that the `interior` / `exterior` condition explains differences in `pee` (interiors are cheaper in absolute terms) but is more neutral in `pmc`. The interpretation: `interior` homes cost less mainly because they tend to be smaller, not because the market significantly penalizes them per square meter. This is a clear example of how joint analysis of `pee` and `pmc` allows decomposing factors that determine price.

PMC / Property Condition

Variables: `nuevo_usado` + `pmc`

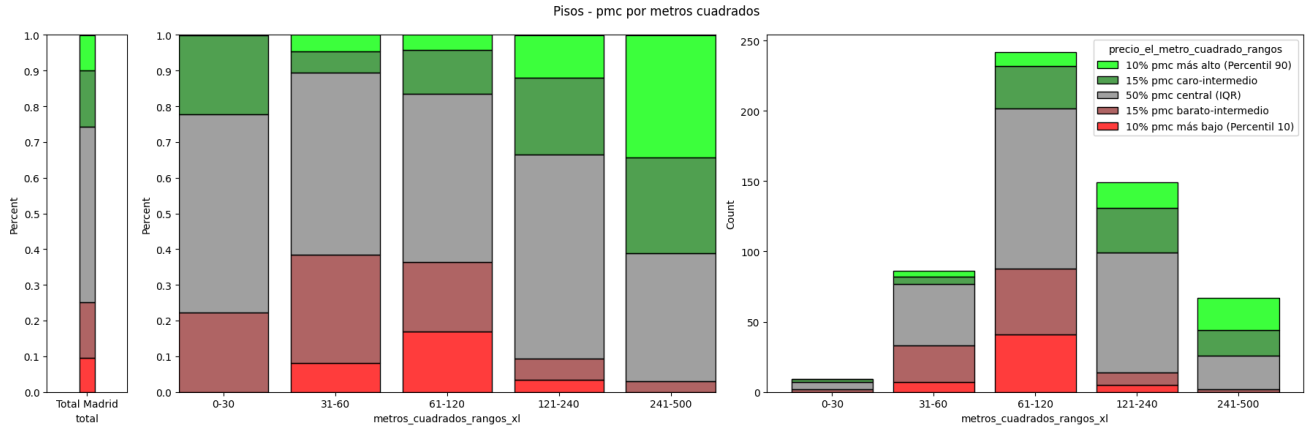
The following composite graph shows `pmc` distribution according to conservation status.



`Used_construction` aligns with total distribution in both `pmc` and `pee`, reflecting its heterogeneity and dominant sample weight. Homes `to_reform` align in `pmc` but had greater weight of high percentiles in `pee` —this indicates they are large homes (high `pee`) but with per square meter valuation similar to general market. `New_construction` presents high variability between both metrics, although the reduced sample (25 obs.) prevents robust conclusions.

SPACE

The `pee` / `metros_cuadrados_rangos_xl` cross showed the expected: total price grows with surface.



The `pee` /surface cross showed the expected: total price grows with surface. The `pmc` /surface cross reveals a less obvious but equally relevant pattern: high `pmc` percentile representation grows with size, reaching 60% in range `241–500 m²`. Small ranges (`0–60 m²`) have greater concentration of low `pmc` .

This pattern suggests that large homes not only cost more in absolute terms (`pee`), but have greater value per square meter (`pmc`).

The most plausible explanation is not that surface itself generates a price premium, but that large homes tend to be located in premium zones and incorporate superior qualities. That is, the surface- `pmc` correlation is mediated by location and quality, it's not a direct relationship.

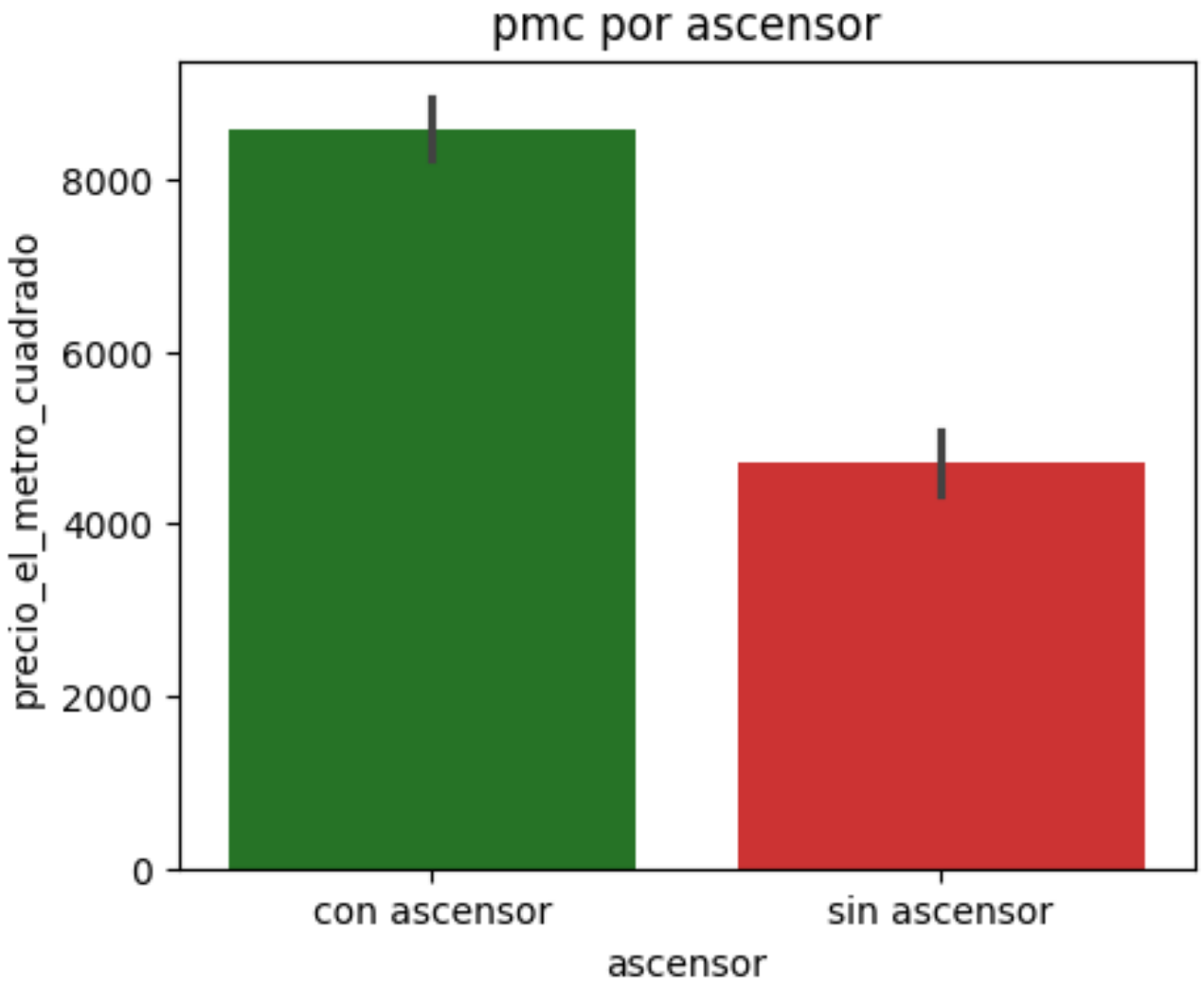
AMENITIES

This section replicates the amenities analysis performed for `pee` , now with `pmc` as dependent variable. The comparison allows distinguishing which amenities affect total price (`pee`) due to their correlation with `size` , and which have independent effect on valuation per square meter (`pmc`). Some relationships are clear; others are not conclusive but are documented for completeness.

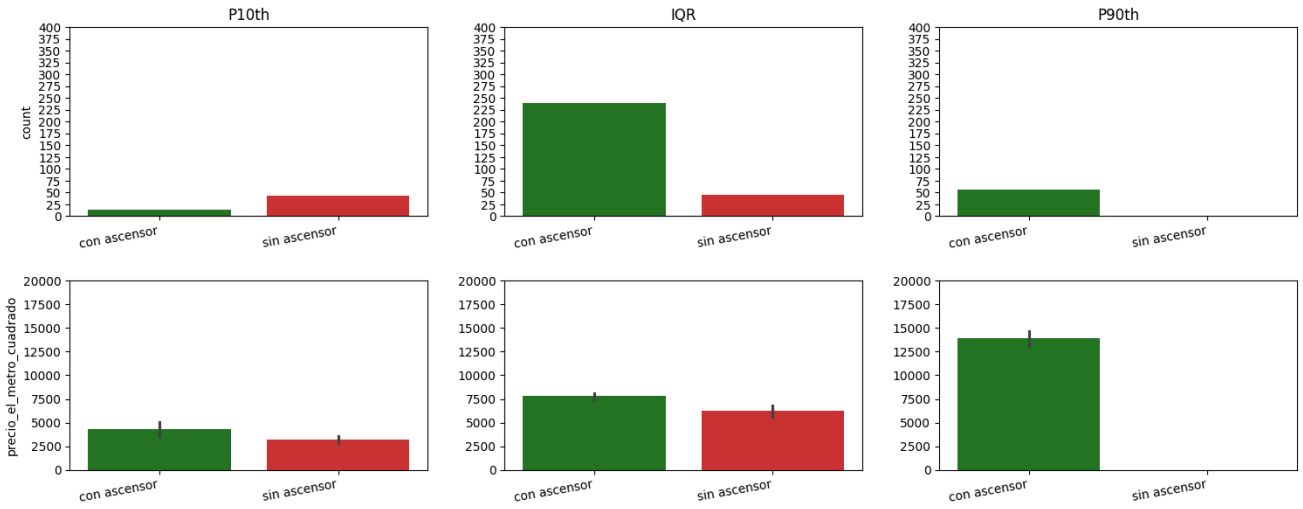
PMC / elevator

Variables: `ascensor` + `pmc`

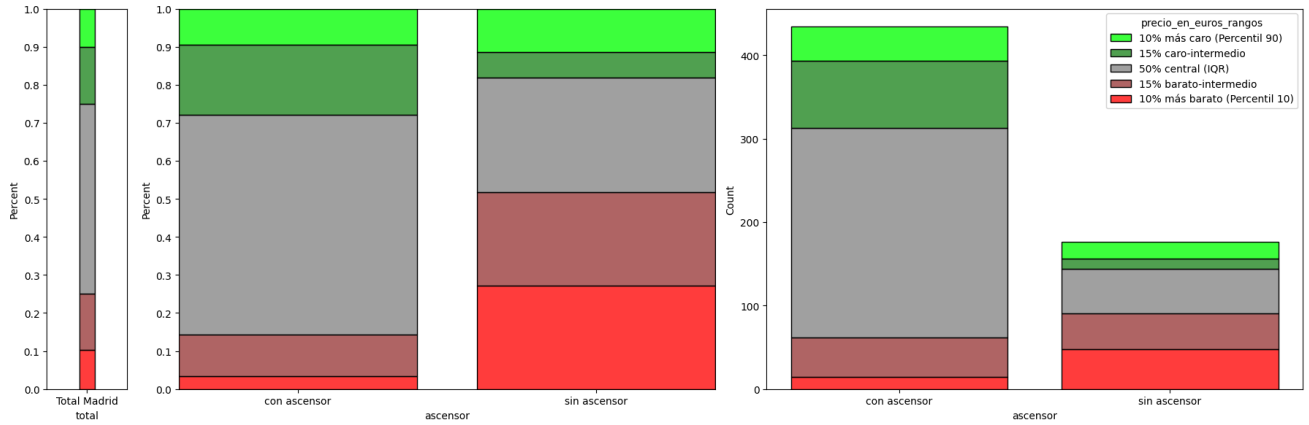
The following composite graph shows average `pmc` and distribution by percentiles according to `elevator` availability.



Homes `with elevator` present `pmc` of ~8,600€/m² vs ~4,700€/m² `without elevator`, a difference of ~83% that is maintained after normalizing by surface. This result is key: the `elevator` not only correlates with larger homes (which would explain differences in `pee`), but with greater valuation per square meter. It's the amenity variable with greatest discriminating power in both metrics, confirming its role as proxy for building quality and location.



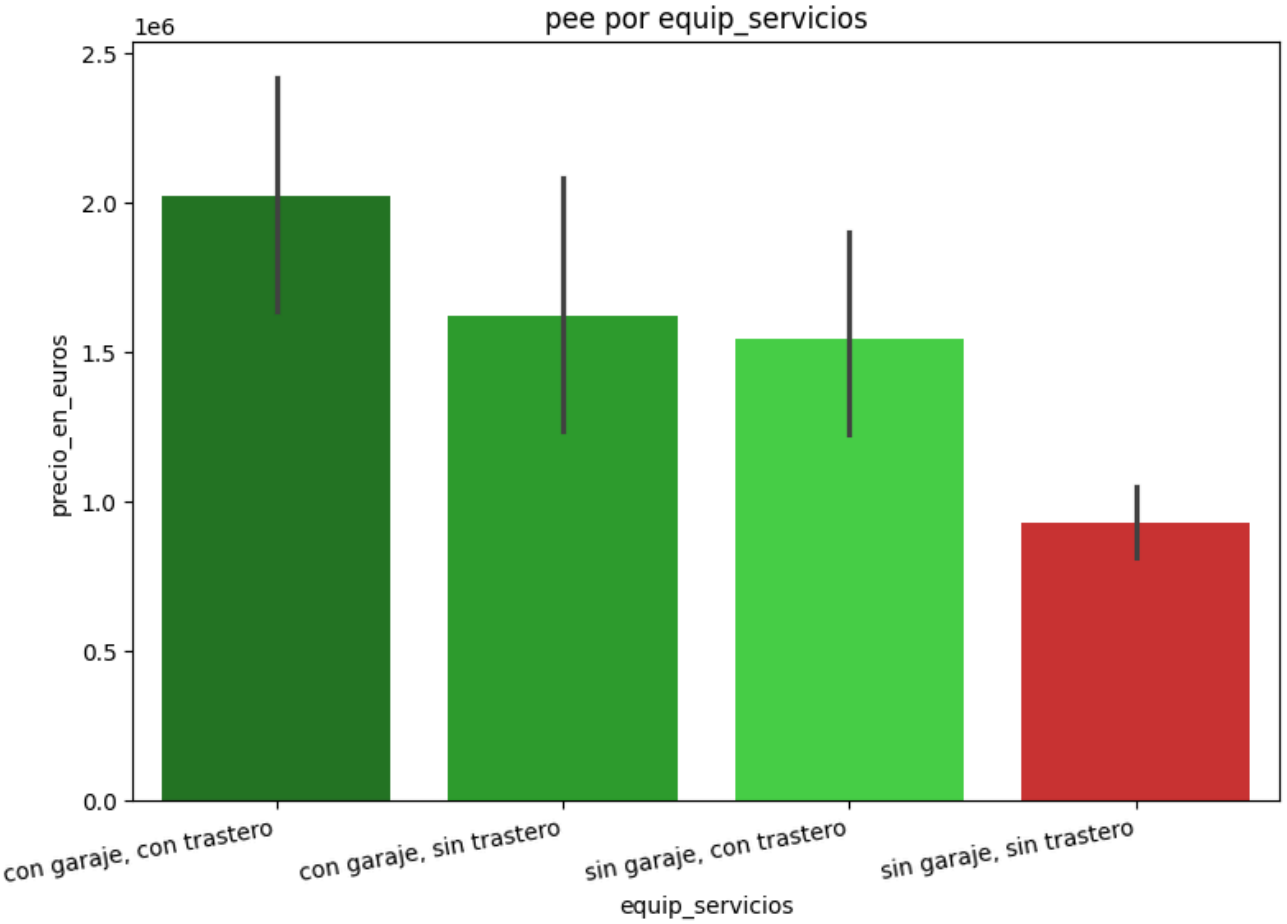
The proportion histogram confirms the pattern: homes `without elevator` concentrate most of their observations in low `pmc` percentiles (`p10` and `cheap-intermediate`), while homes `with elevator` present distribution shifted toward high percentiles, dominating practically alone `p90`. The pattern replicates almost exactly what was observed in `pee`, confirming effect robustness.



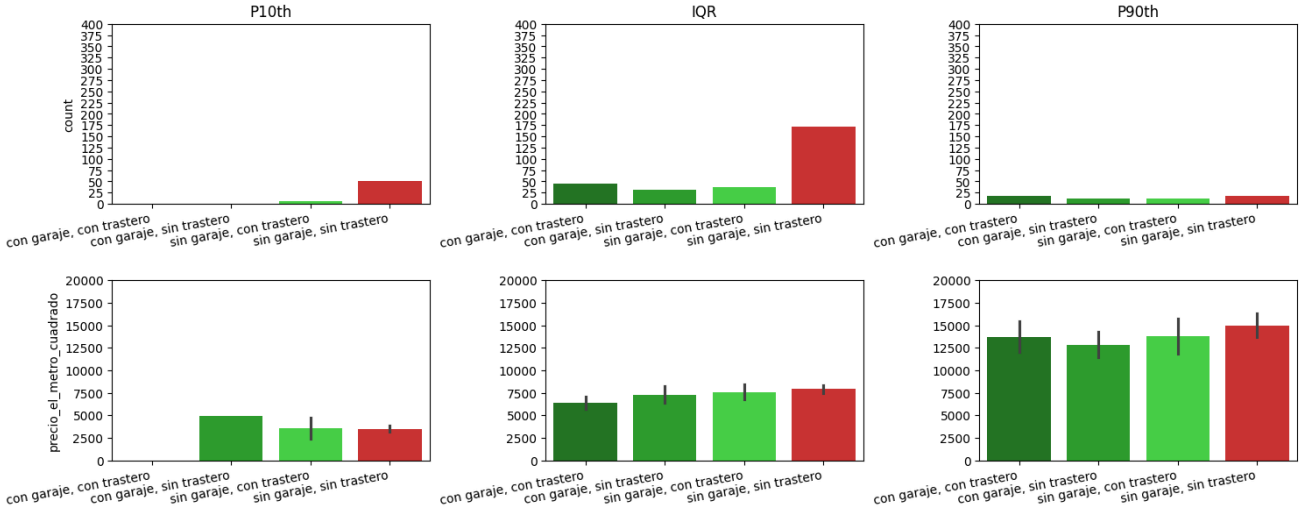
PMC / services amenity

Variables: `equip_servicios` + `pmc`

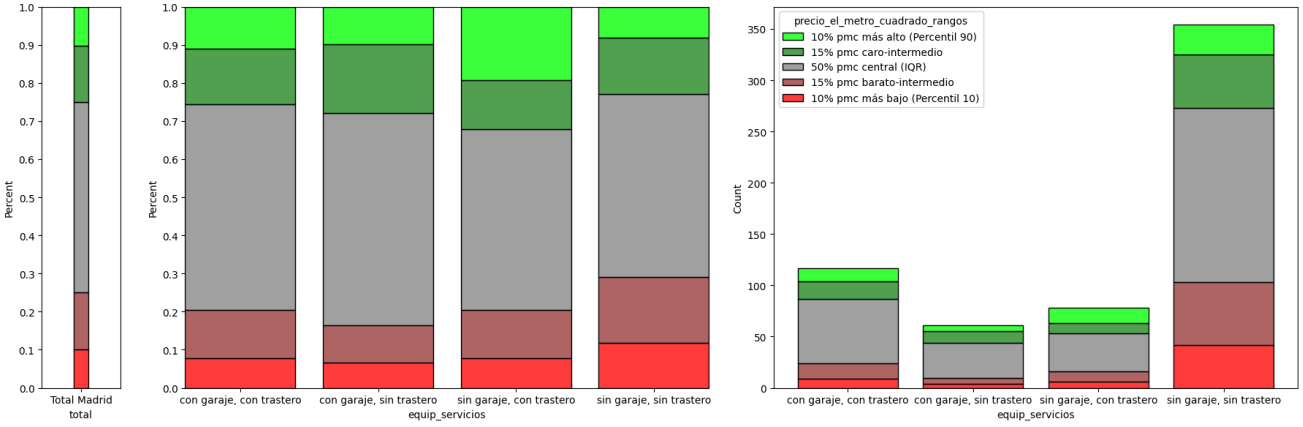
The following composite graph shows average `pmc` and distribution by percentiles according to `garage` and `storage room` availability.



The four categories show similar `pmc` (~8,000-8,300€/m²), with `without/without` slightly lower (~7,300€/m²). Discriminating power in `pmc` is lower than in `pee`, suggesting that `garage` and `storage room` correlate with larger homes (affecting `pee`) but don't add significant premium per square meter. The methodological limitation about optional inclusion of these elements in price, mentioned in the `pee` section, equally applies here.



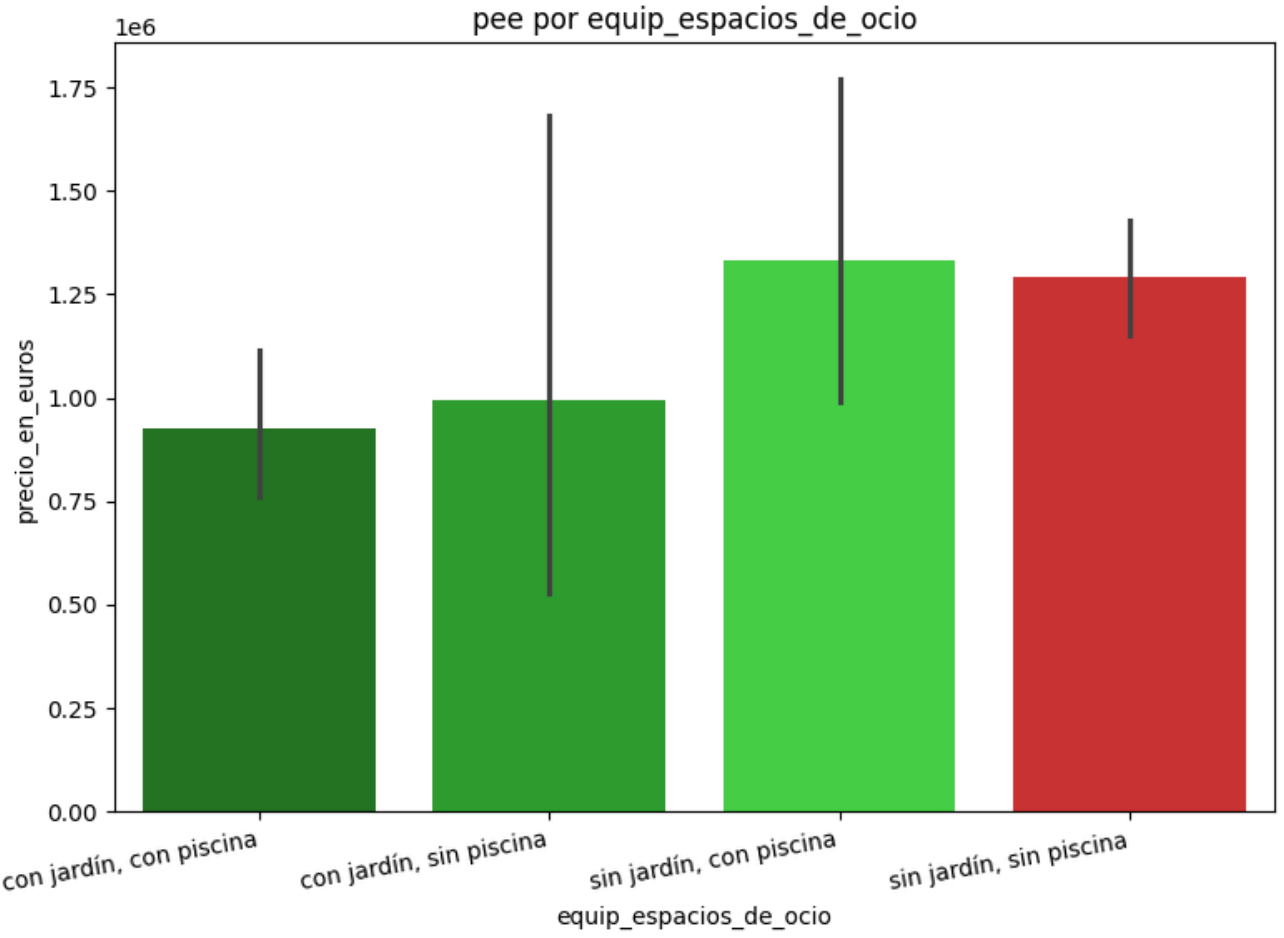
The proportion histogram shows gradual evolution: `with/with` presents greater relative weight in high `pmc` percentiles, a pattern that gradually attenuates until `without/without`, where concentration in low percentiles increases. Differences are less pronounced than in `elevator`, consistent with lower discriminating power of this variable.



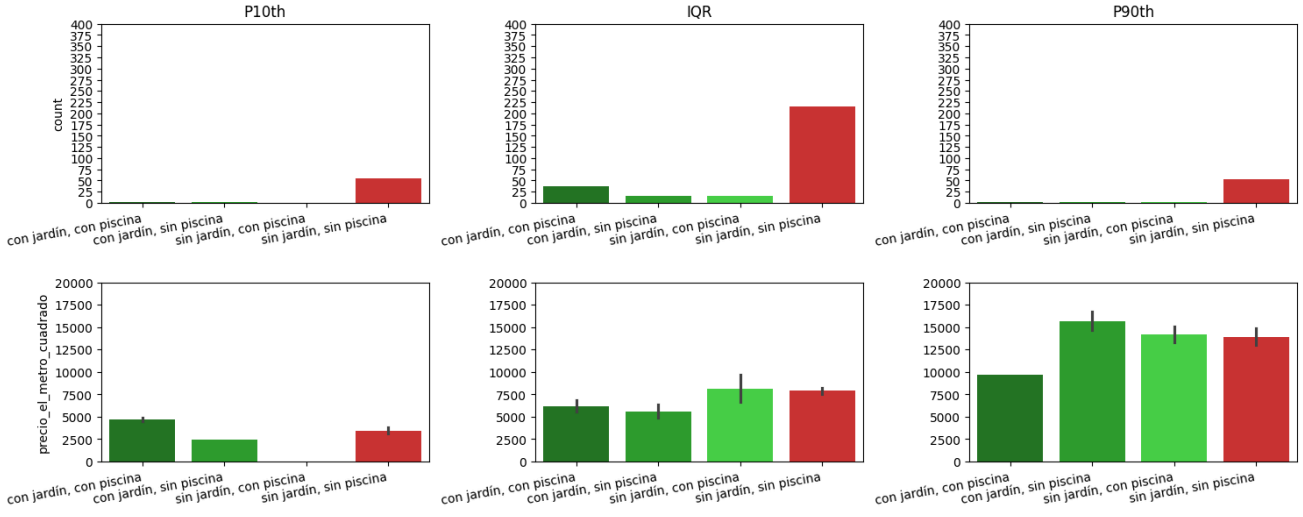
PMC / leisure spaces amenity

Variables: `equip_espacios_de_ocio` + `pmc`

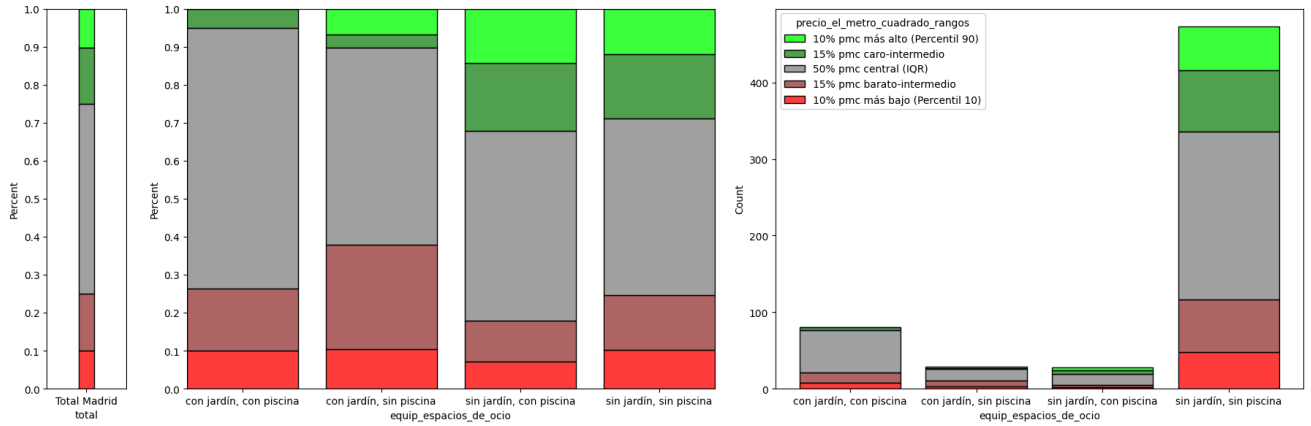
The following composite graph shows average `pmc` and distribution by percentiles according to `garden` and `pool` availability.



The pattern replicates what was observed in `pee`: `with/without` (garden without pool) has the lowest `pmc` (~ `6,000€/m²`), `without/with` (pool without garden) the highest (~ `8,400€/m²`). The `without/without` category (~ `7,900€/m²`) dominates in count. The persistence of inverse amenity/price relationship in `pmc` confirms this is a location effect, not size: `garden` and `pool` are more frequent in peripheral zones of lower `pmc`, and this effect doesn't disappear when normalizing by surface.



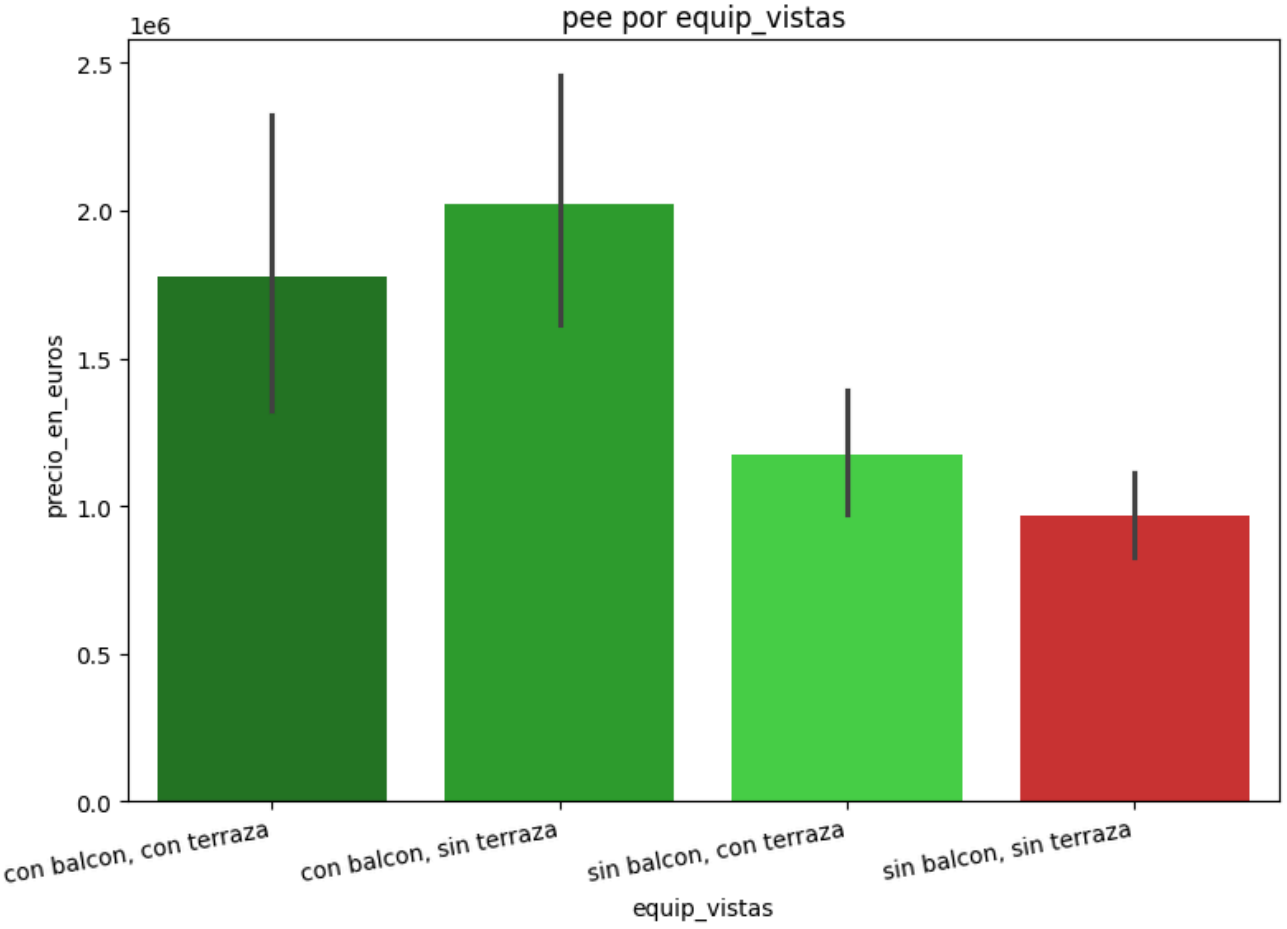
The proportion histogram reflects the pattern: `without/with` (pool without garden) shows greater concentration in high `pmc` percentiles, while `with/without` (garden without pool) presents greater weight in low percentiles. The `without/without` category, despite dominating in absolute count, shows distribution close to population. This pattern reinforces the hypothesis that `garden` is more associated with peripheral locations of low `pmc`, while community `pool` can also appear in certain-standing central buildings.



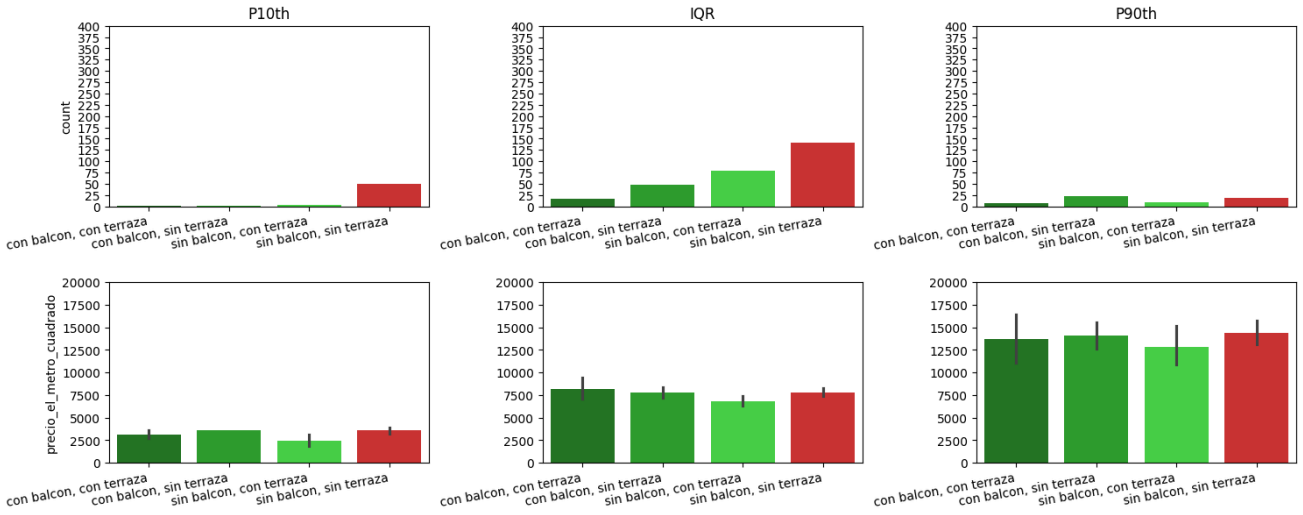
PMC / views amenity

Variables: `equip_vistas` + `pmc`

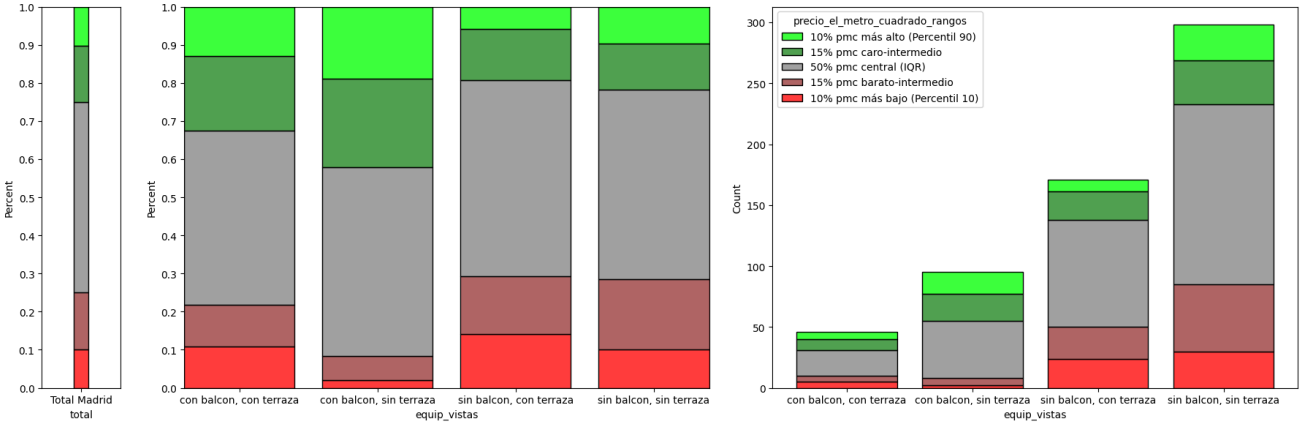
The following composite graph shows average `pmc` and distribution by percentiles according to `balcony` and `terrace` availability.



`Pmc` oscillates between `7,000€/m²` - `9,500€/m²`. The `with/without` category (balcony without terrace) presents the highest value (~ `9,500€/m²`); `without/with` (terrace without balcony) the lowest (~ `7,000€/m²`). `Balcony` shows greater association with high `pmc` than `terrace`, pattern consistent with `pee`. The explanatory hypothesis remains: `balcony` is more frequent in central apartments of historic buildings, while `terrace` (without balcony) usually associates with penthouses or `periphery` homes.



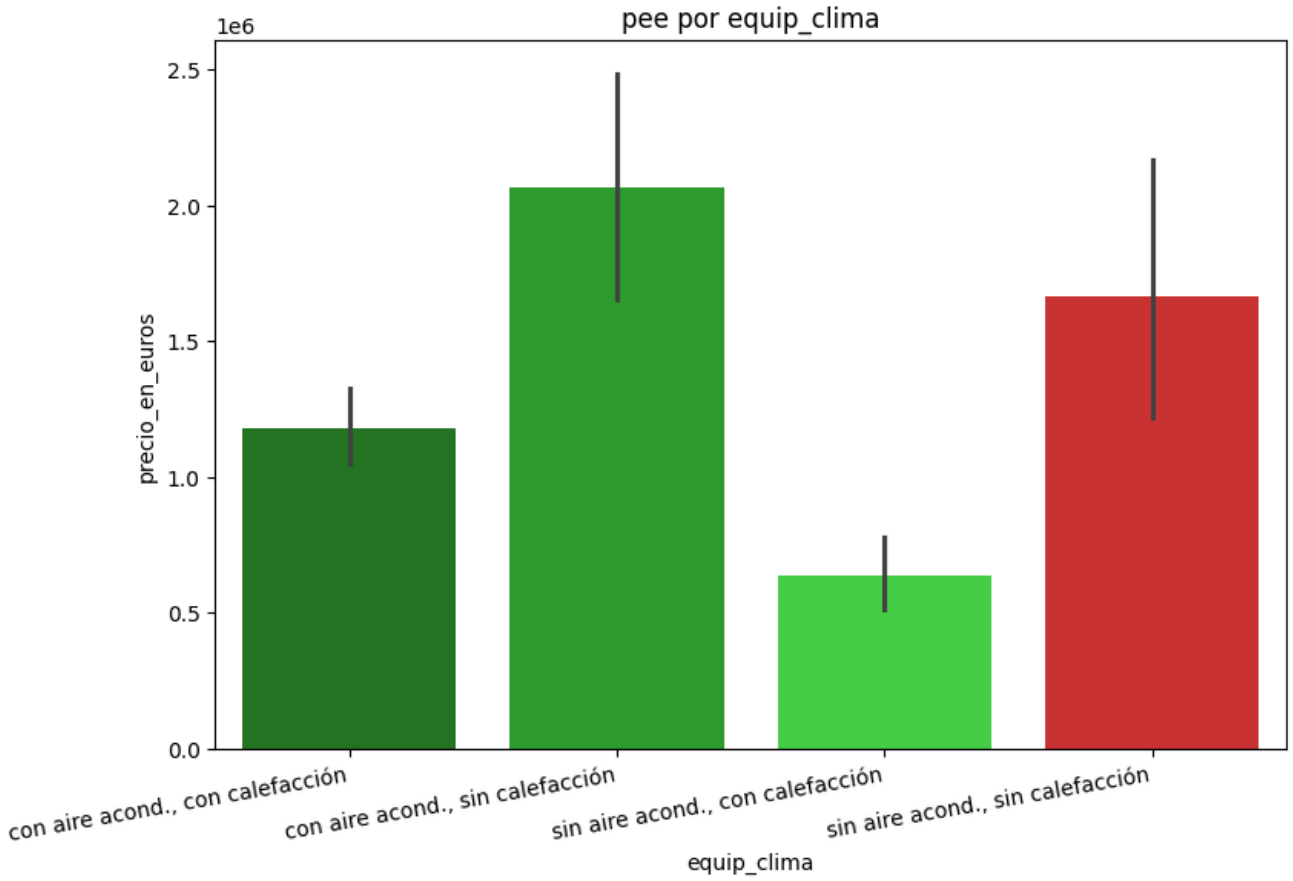
The proportion histogram confirms that categories with `balcony` (`with/with` and `with/without`) concentrate greater weight in high `pmc` percentiles, especially `with/without` (balcony without terrace) which dominates `p90`. Categories without `balcony` (`without/with` and `without/without`) present distributions shifted toward low-intermediate percentiles. The pattern is consistent between `pee` and `pmc`, indicating that the `balcony` effect is not only due to property size.



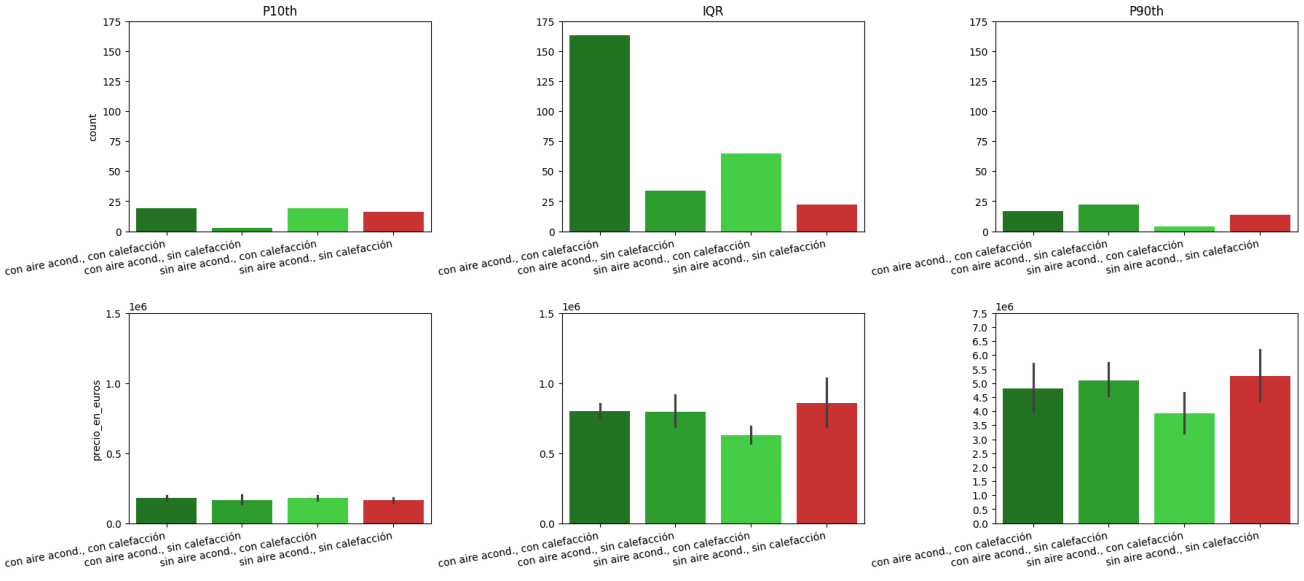
PMC / climate amenity

Variables: `equip_clima` + `pmc`

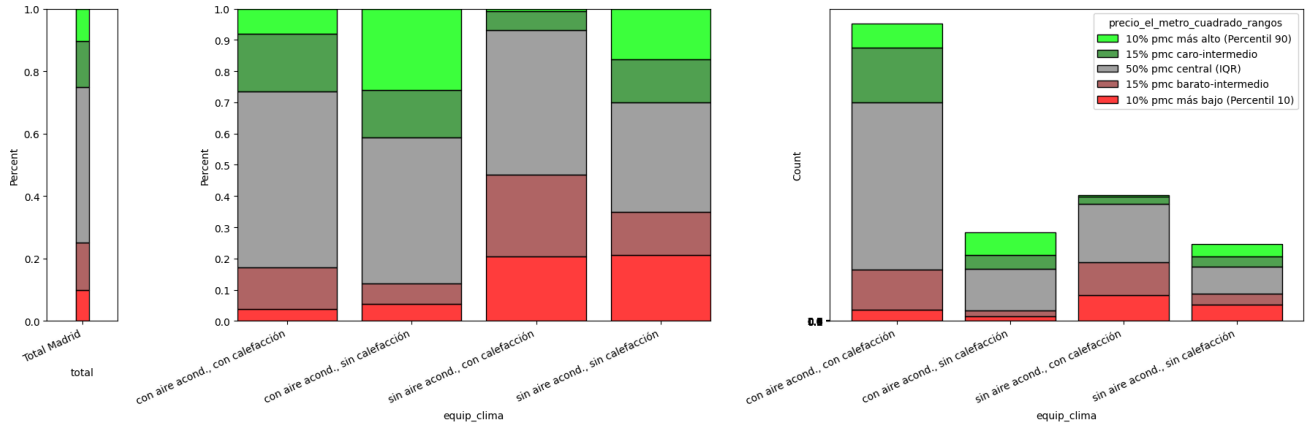
The following composite graph shows average `pmc` and distribution by percentiles according to `air conditioning` and `heating` availability.



The `with/without` category (air without heating) presents the highest `pmc` (~9,900€/m²); `without/with` (heating without air) the lowest (~5,400€/m²). `Air conditioning` shows greater association with high `pmc` than `heating`, pattern consistent with `pee`. This difference is even more pronounced in `pmc` than in `pee`, suggesting that `air conditioning` functions as an indicator of property quality/modernity that is directly reflected in per square meter valuation.



The proportion histogram shows clear segregation: `with/without` (air without heating) concentrates the highest proportion of homes in `p90`, while `without/with` (heating without air) dominates in `p10`. The `without/without` category presents more balanced distribution between percentiles. This pattern reinforces that `air conditioning` is a better predictor of high `pmc` than `heating`, consolidating as one of the amenity variables with greatest discriminating power along with `elevator`.



Conclusions

The descriptive analysis of the purchase-sale housing market in Madrid, based on a sample of 610 homes extracted in August 2024, reveals consistent patterns that I synthesize below. These conclusions should be interpreted considering the methodological limitations exposed: sampling is not strictly random, some segments have reduced representation, and data correspond to offer prices, not transaction prices.

Regarding location, the `center` concentrates the highest-priced homes in both `pee` and `pmc`, with more pronounced differences in `pmc` where 100% of `p90` is located in the `center`—a resounding fact that underlines the primacy of location as a determinant of market value. The median `pmc` in the `center` (~ `9,000€/m²`) practically doubles that of `south` and `east` (~ `4,000€/m²` - `5,000€/m²`). `North` and `west` occupy intermediate positions (~ `6,000€/m²` - `7,000€/m²`).

Regarding physical characteristics, I highlight three findings. First, `chalets` show inverse behavior depending on metric: high `pee` (absolute value) but low `pmc` (normalized price), confirming that their high total price responds to surface, not to premium valuation per square meter. Second, high floors (`6th-7th`) present the highest `pmc`, although floors `8th` + descend—a result that could reflect heterogeneity in this segment or sampling limitations. Third, `interior` homes have lower `pee` but `pmc` similar to `exterior`, indicating that their lower total price is mainly due to size, not to a penalty per square meter.

Regarding amenities, the `elevator` shows the clearest and most robust association with high price, in both `pee` and `pmc`. This variable probably functions as proxy for multiple correlated factors: building quality, age, central location. Leisure amenities (`garden` / `pool`) present inverse relationship with price, an effect that persists in `pmc` and confirms it's due to their concentration in lower-value peripheral zones, not to property size. `Air conditioning` has greater association with high price than `heating`, suggesting it functions as an indicator of property quality/modernity. `Balcony` shows greater association with high `pmc` than `terrace`, possibly due to its greater frequency in historic central buildings.

In terms of variable hierarchy, location (`zone` / `subzone`) and `elevator` emerge as the strongest and most consistent price predictors in both metrics. Surface has positive relationship with `pee` (expected) and also with `pmc` (less obvious), although this second effect is probably mediated by correlation between size and location/quality. Leisure amenities (`garden` / `pool`) illustrate the risk of bivariate analysis: their negative association with price reflects geographical distribution, not causal effect. `Heating` presents weak or even inverse association with `pmc`, being of little use as predictor. Condition (`new` / `used` / `to_reform`) and `exterior` / `interior` condition have moderate effects on `pee` but limited on `pmc`, indicating they affect price mainly through size.

Regarding limitations, the analysis is descriptive: it identifies patterns and associations that can guide hypotheses, but doesn't demonstrate causal relationships. Uncollected variables (building age, construction qualities, proximity to public transport, solar orientation) could nuance or modify some results. Some segments—peripheral districts, `new_construction`, floors `8th` +, `chalets` —have reduced samples (<50 observations) that limit conclusion robustness in those groups.

Finally, data correspond to offer prices, not transaction prices. This can introduce bias if negotiation discounts vary systematically between segments—for example, if luxury homes have greater negotiation margin than economical ones. This limitation is inherent to real estate portal data and should be kept in mind when interpreting results.

Despite these limitations, the analysis provides systematic characterization of the Madrid purchase-sale market that can serve as basis for deeper studies or as reference for future periodic analyses, the original project objective.