Exploring Housing Prices with Key Property Features

1 Introduction

In the real estate market, several factors influence the price of a property, such as the number of bedrooms, the size of the house, its distance from the city center, and proximity to public transportation. This analysis is based on a housing dataset, focusing on variables like `Bedroom Count`, `Net Square Meters`, `Center Distance`, `Metro Distance`, `Floor`, `Age`, and `Price`. We will explore how each of these factors contributes to determining housing prices by cleaning the data, visualizing relationships, and performing statistical tests.

2 Key Variables

1. **Bedroom Count**: Represents the number of bedrooms in the housing unit, indicating its size and capacity.
2. **Net Square Meters (Net Sqm)**: Refers to the total usable interior space within the housing unit, quantifying the property size.
3. **Center Distance**: Measures the distance from the housing unit to the city center, which affects convenience and price.
4. **Metro Distance**: Indicates the distance to the nearest metro or subway station, influencing the appeal of the property for commuters.
5. **Floor**: Specifies the floor or level of the housing unit within the building.
6. **Age**: Represents the number of years since the property was constructed or renovated.
7. **Price**: The price of the property, which we aim to analyze in relation to the other features.

3 Data Cleaning

Before analysis, the dataset requires cleaning to address missing or inconsistent values. We imputed missing values using the median for numeric features,

ensuring that the dataset was ready for analysis. Additionally, we ensured all variables were correctly formatted, with `Net Sqm` and `Price` in numerical formats.

4 Data Visualization

To explore the relationships between different variables and house prices, we created the following visualizations:

- **Scatter plot** of `Price` vs. `Net Square Meters` to examine how property size affects price.
- **Scatter plot** of `Price` vs. `Center Distance` to see the impact of proximity to the city center.
- **Box plot** to visualize price variations across different `Bedroom Count` categories.
- **Histogram** to show the distribution of property `Age`.

5 Data Summarization

Descriptive statistics provide an overview of the dataset, including the mean and median of house prices, distribution of bedroom counts, and the variability in the `Net Sqm` of housing units. These statistics offer insights into the central tendencies and spread of key variables.

6 Statistical Tests

To quantify relationships between variables, we performed the following statistical tests:

- **Pearson correlation** between `Net Sqm` and `Price`, which revealed a positive relationship indicating that larger properties tend to have higher prices.
- **T-test** to compare prices between newer properties (`Age < 10 years`) and older properties, showing that newer properties tend to command higher prices.