PRESENTATION:

HOUSING PRICE PREDICTION IN PARIS

Michał Tarnawa, Wojciech Grzywocz, Michał Żarłok, Jakub Wizner

DATASET DESCRIPTION

WE ARE USING DATASET FROM KAGGLE, WHICH CONTAINS INFOMATIONS ABOUT 10,000 FLATS AND APARTMENTS IN PARIS.

NONE OF THE VALUES WERE NULL.

THIS DATASET IS DIVIDED INTO 16 COLUMNS, WHICH CAN BE SPLIT INTO FOUR CATEGORIES.

NUMERICAL

- squareMeters
- numberOfRooms
- floors
- numPrevOwners
- made
- basement
- attic
- garage
- hasGuestRoom

CATHEGORICAL

- hasYard
- hasPool
- isNewBuilt
- hasStormProtector
- hasStorageRoom

OUTPUT

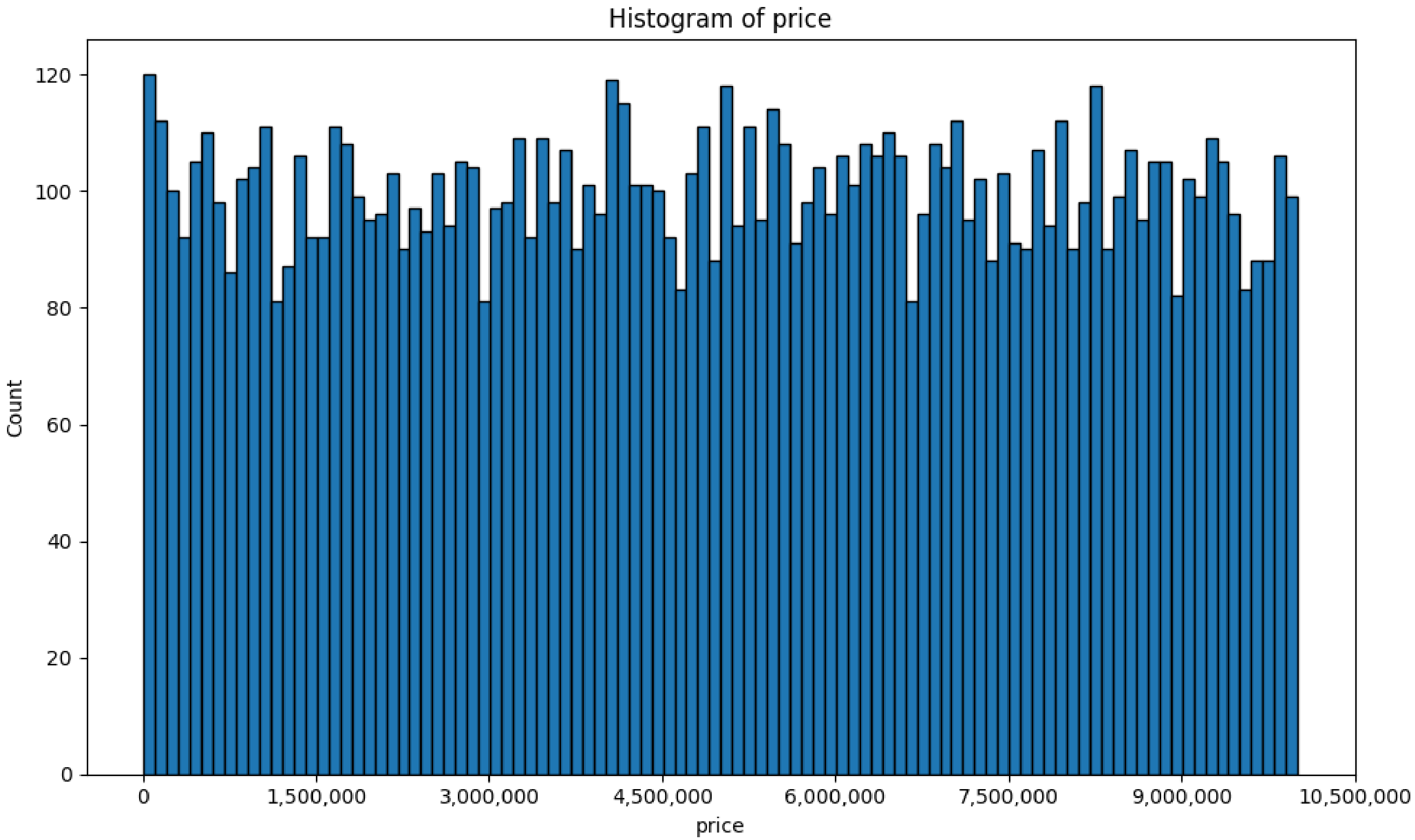
price

UNSUED

- cityPartRange
- cityCode

PRICE(€)

- min = 10,313.50
- max = 10,006,770.00
- avg = 4,993,448.00
- med = 5,016,180.00
- std = 2,877,424.00



CONCLUSIONS:

A big variety between minimum and maximum value and std at 3,000,000 € indicates that our data is diverse.

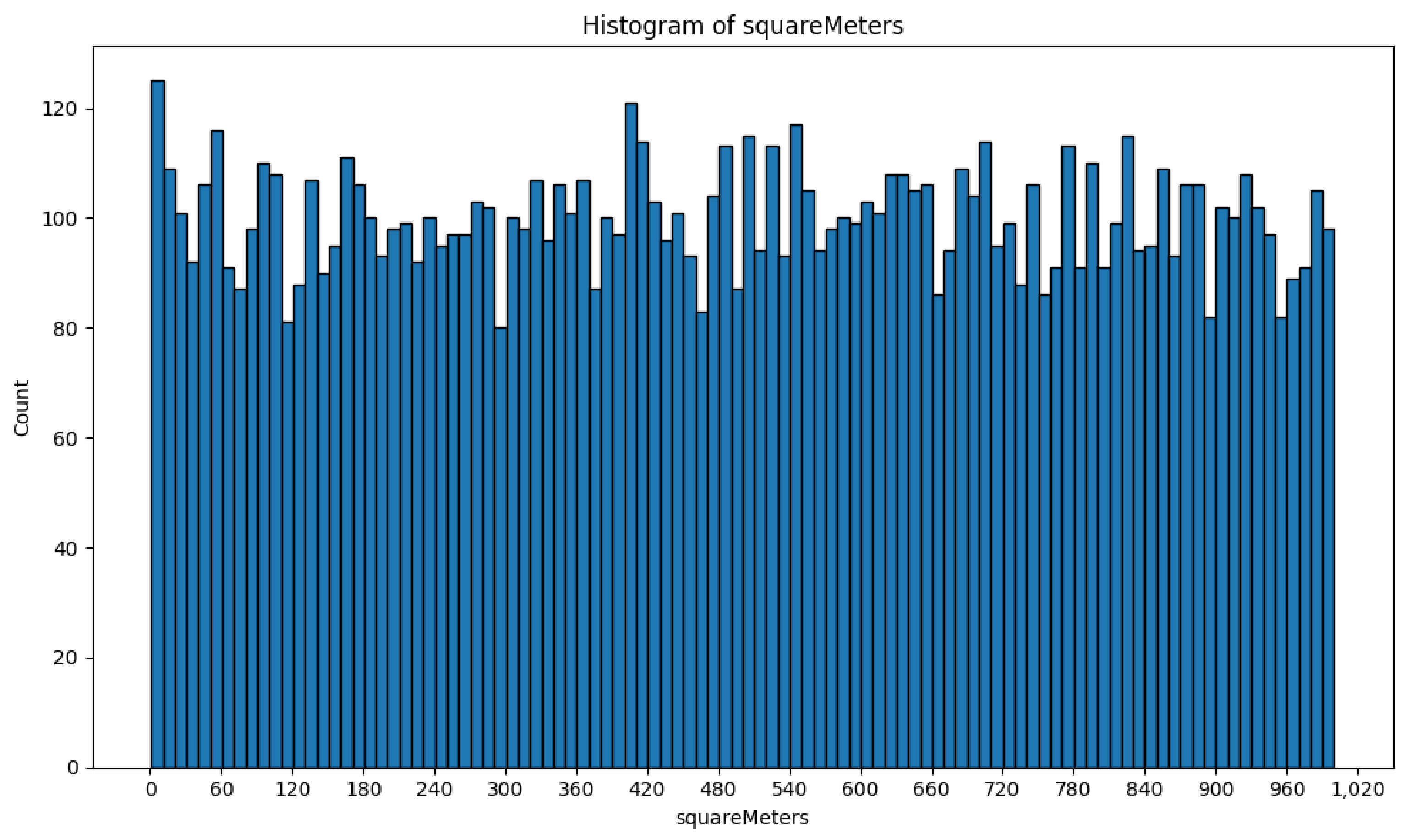
Average value shows that more expensive apartments are dominating in this dataset, which may cause a problem in estimating the price of cheaper housing.

PICTURE(1)

As can be seen on the histogram.
Price values are distributed almost equally on all values. This may suggest that most of the data came from wealthier districts of Paris.

SQUAREMETERS (M^2)

- min = 0.89
- max = 999.99
- avg = 498.70
- med = 501.06
- std = 287.74



PICTURE(2)

As can be seen on the histogram, m^2 of housing is distributed equally, with slight distribution at the beginning and end of the range.

The similarity to the price histogram suggests that these two pieces of information are connected.

CONCLUSIONS:

The smallest housing is only 0.89 m^2, which may be an error; however, it might be a microapartment with additional services.

Maximum value at 1000 m^2, standard deviation at 287, as well as average value at 500, confirms our thesis about collecting values from wealthier districts.

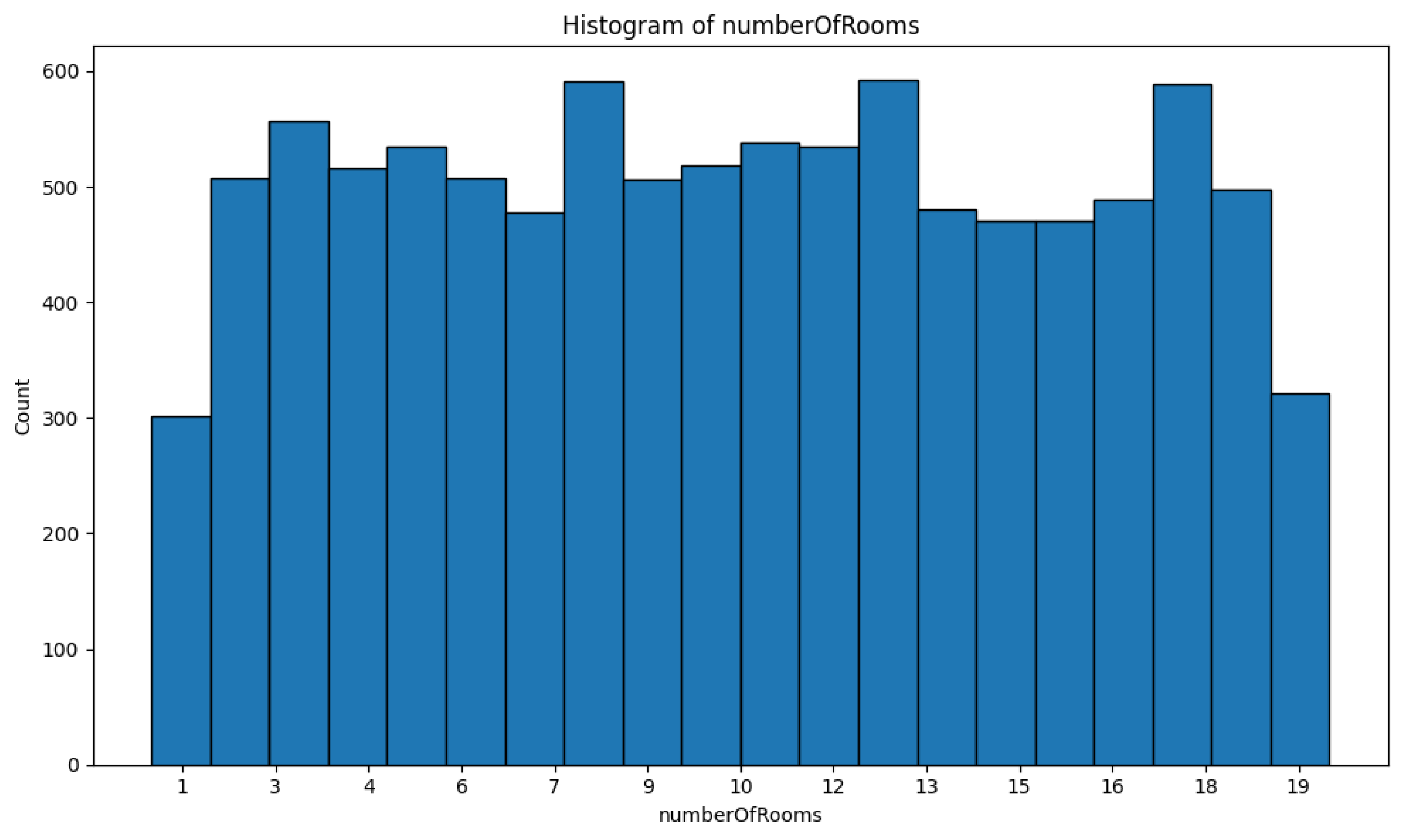
NUMBEROFROOMS

- min = 1.00
- max = 20.00
- avg = 10.47
- med = 10.00
- std = 5.55

CONCLUSIONS:

Minimum value of 1 suggests that there are microapartments in our dataset.

Average and median at 10 mean that there might be a lot of older buildings that were renovated (previous old townhouses), which may have more smaller rooms.



PICTURE(3)

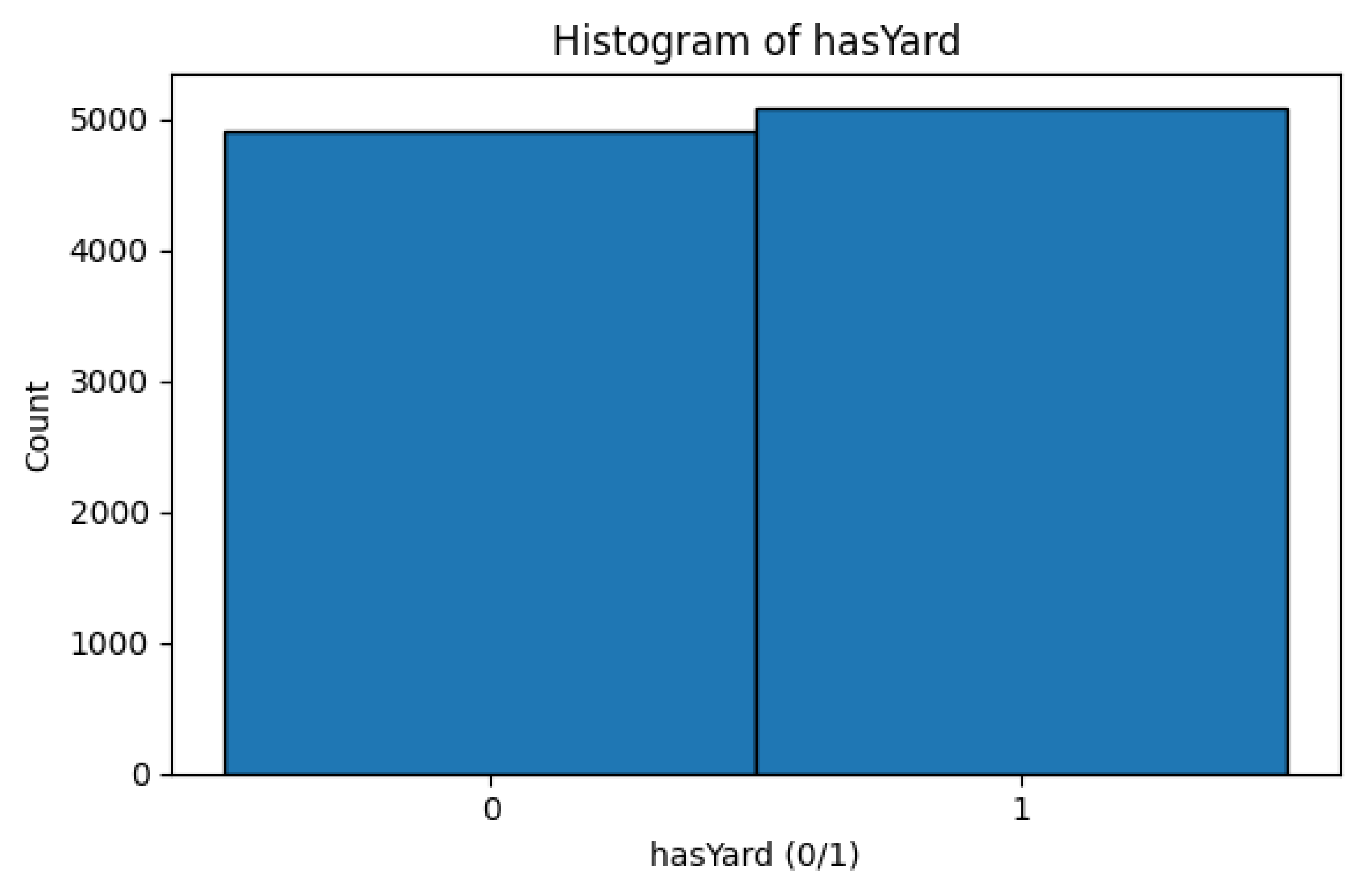
As can be seen, the amount of 2- to 18-room housing is distributed almost equally (500), with 8, 13, and 18 deviating from the rest, reaching almost 600 units. Fortunately for us, the amount of microapartments and penthouses (20) is low and equals 300

HASYARD

- Values are true or false(1 or 0).

CONCLUSIONS:

Half of the housing has a yard; the second does not.
We do not know the area of it as well as if it is shared or not.
That might cause some problems with more accurate predictions.



PICTURE(3)

Most of the housing has a yard. It is distributed equally at 5000 counts each.

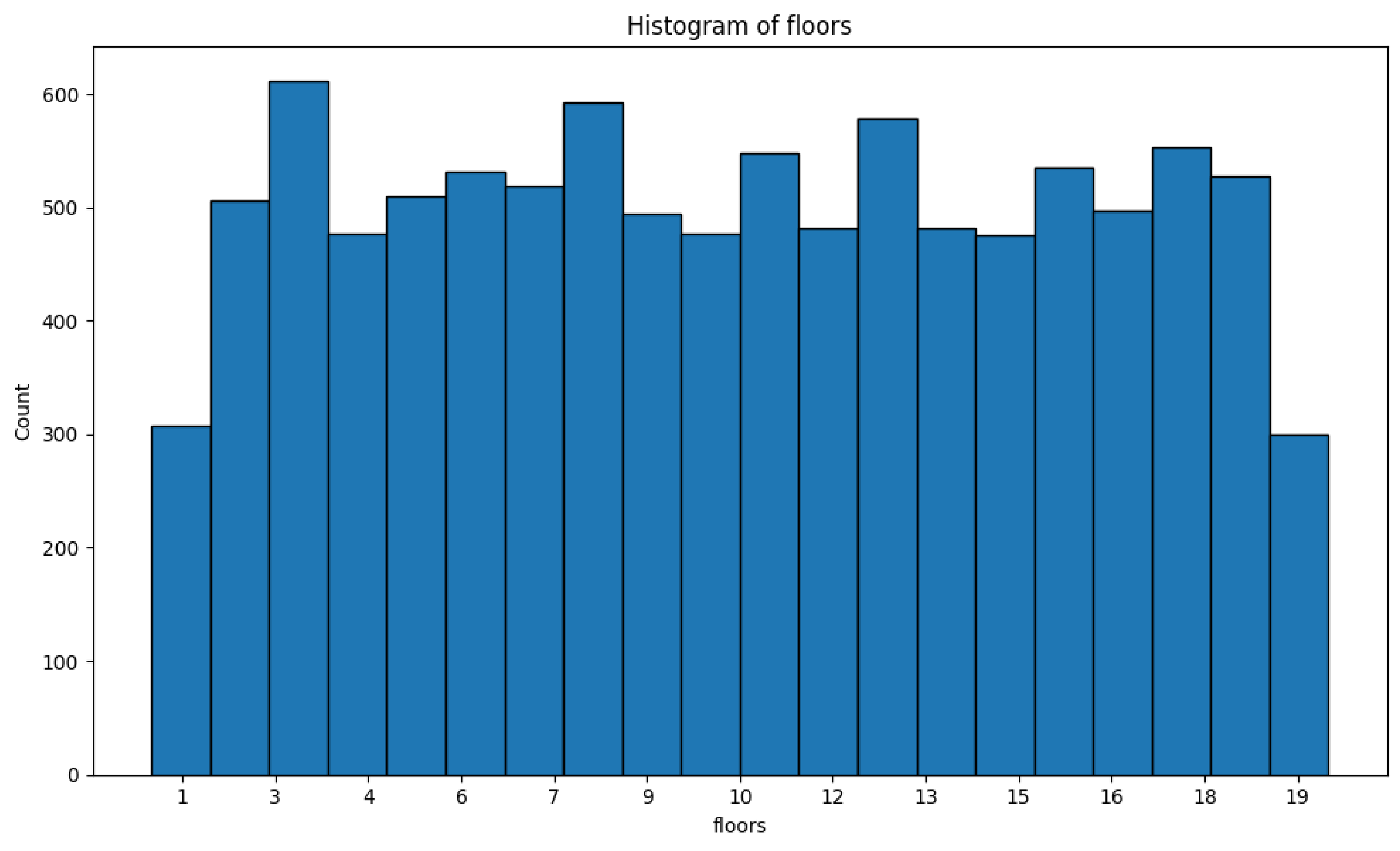
FLOORS

- min = 1.00
- max = 20.00
- avg = 10.46
- med = 10.00
- std = 5.57

CONCLUSIONS:

As can be seen, average housing is a 10-floor building.

There are bigger skyscrapers with 20 floors, which is normal in newer districts of Paris.
There are a few detached houses.

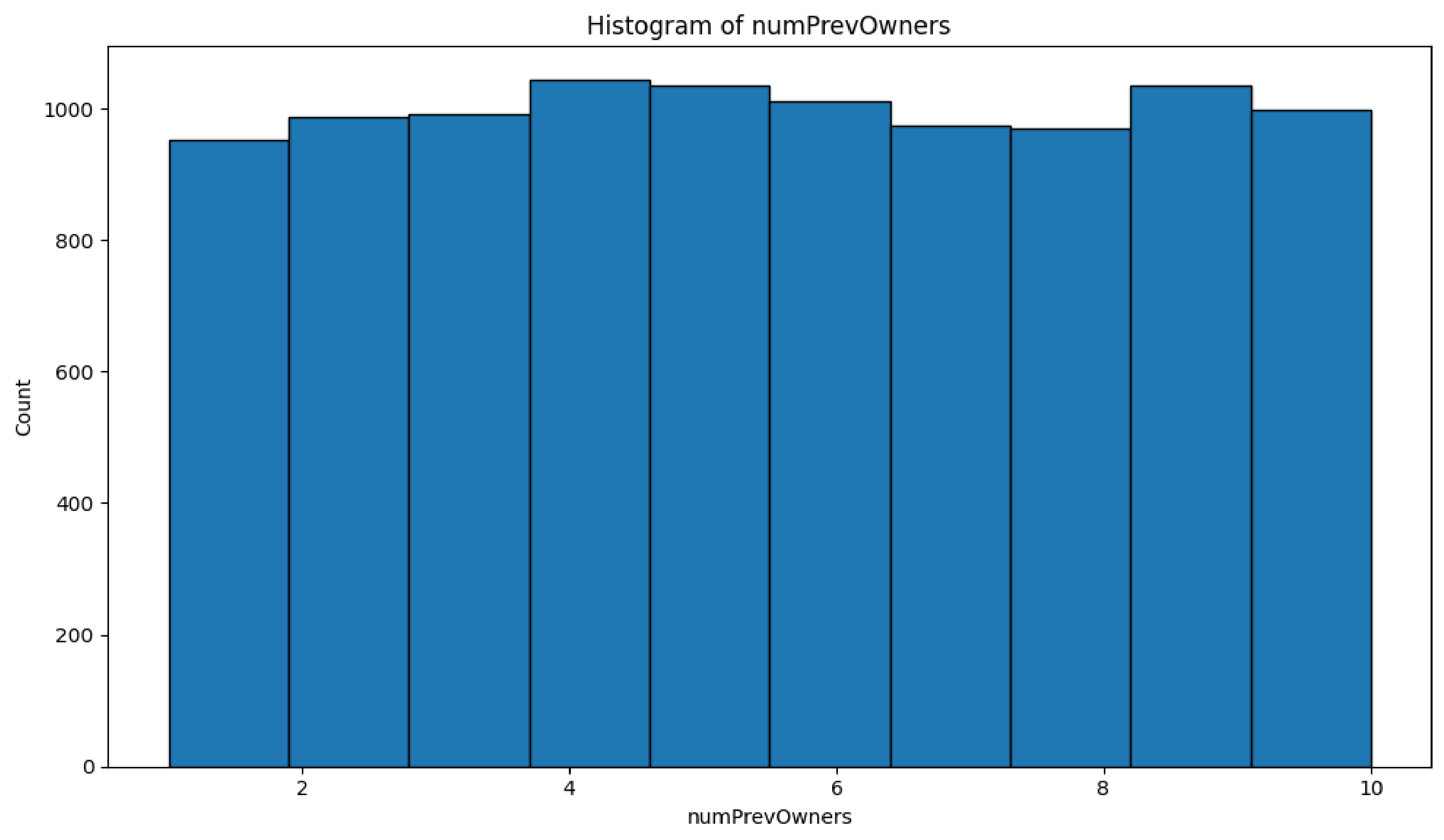


PICTURE(4)

As can be seen, there are 300 one-floor housing units.
The amount of all numbers of floors is around 500, with only 3-, 9-, 10-, and 12-floor housing rising to 600.
There is also about 300 20-floor buildings.

PREVIOUS OWNERS

- min = 1.00
- max = 10.00
- avg = 5.52
- med = 5.00
- std = 2.86



CONCLUSIONS:

Minimum number of previous owners at 1 suggests that housing was brand new (there was 1 person before selling – the builder).

Average value at 5 in the middle between the maximum value and the minimum one suggests that data was selected equally.

PICTURE(5)

As can be seen above, the number of previous owners is distributed equally at around 900 counts at each value.

This suggests that data was selected carefully.

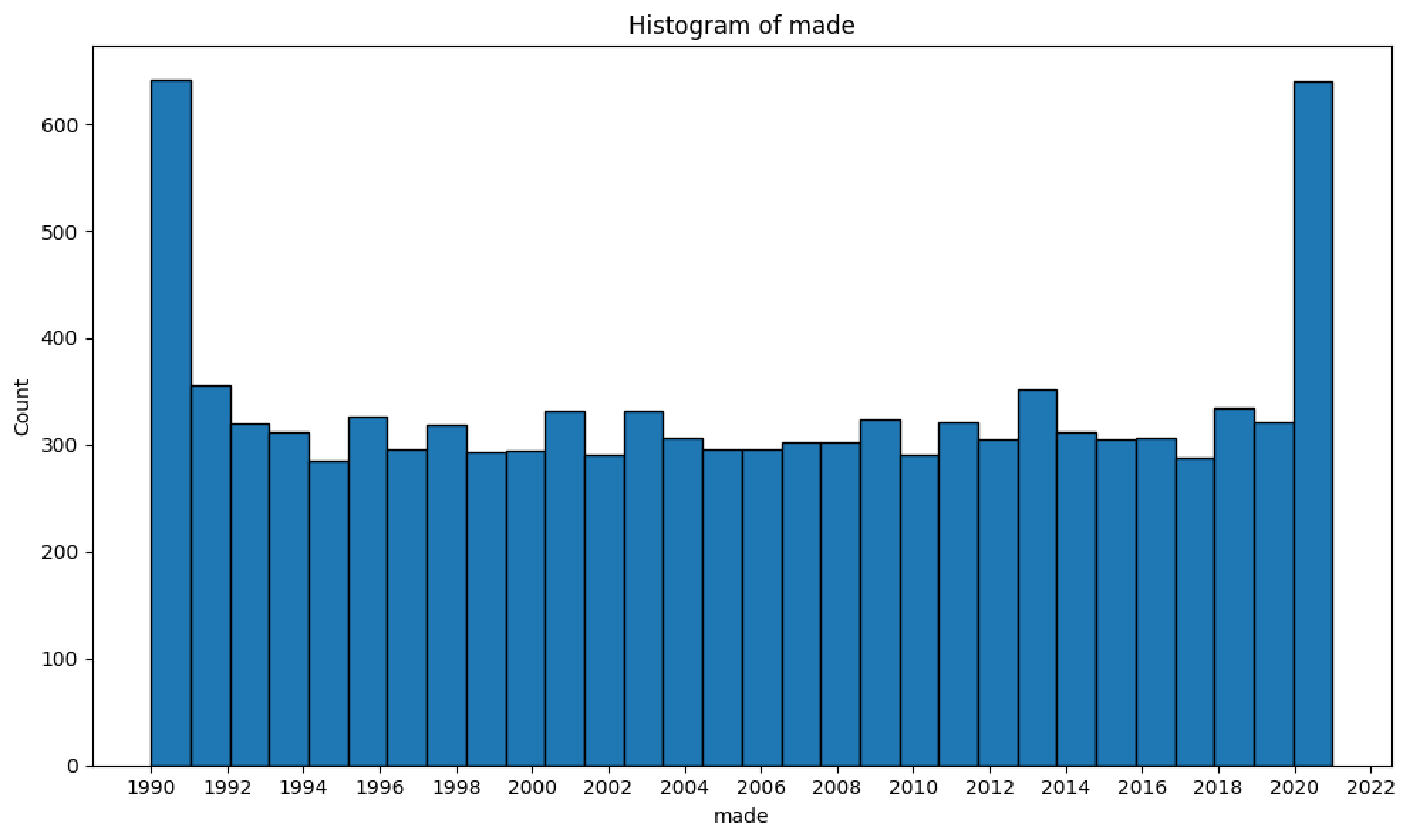
MADE(YEAR)

- min = 1,990.00
- max = 2,021.00
- avg = 2,005.49
- med = 2,005.50
- std = 9.31

CONCLUSIONS:

The oldest building in the dataset was built in 1990, whereas the newest in 2021. That means that our prediction cannot be 100% accurate.

It can be said that average housing was built in 2005 and the deviation between years is almost 9 years.



PICTURE(5)

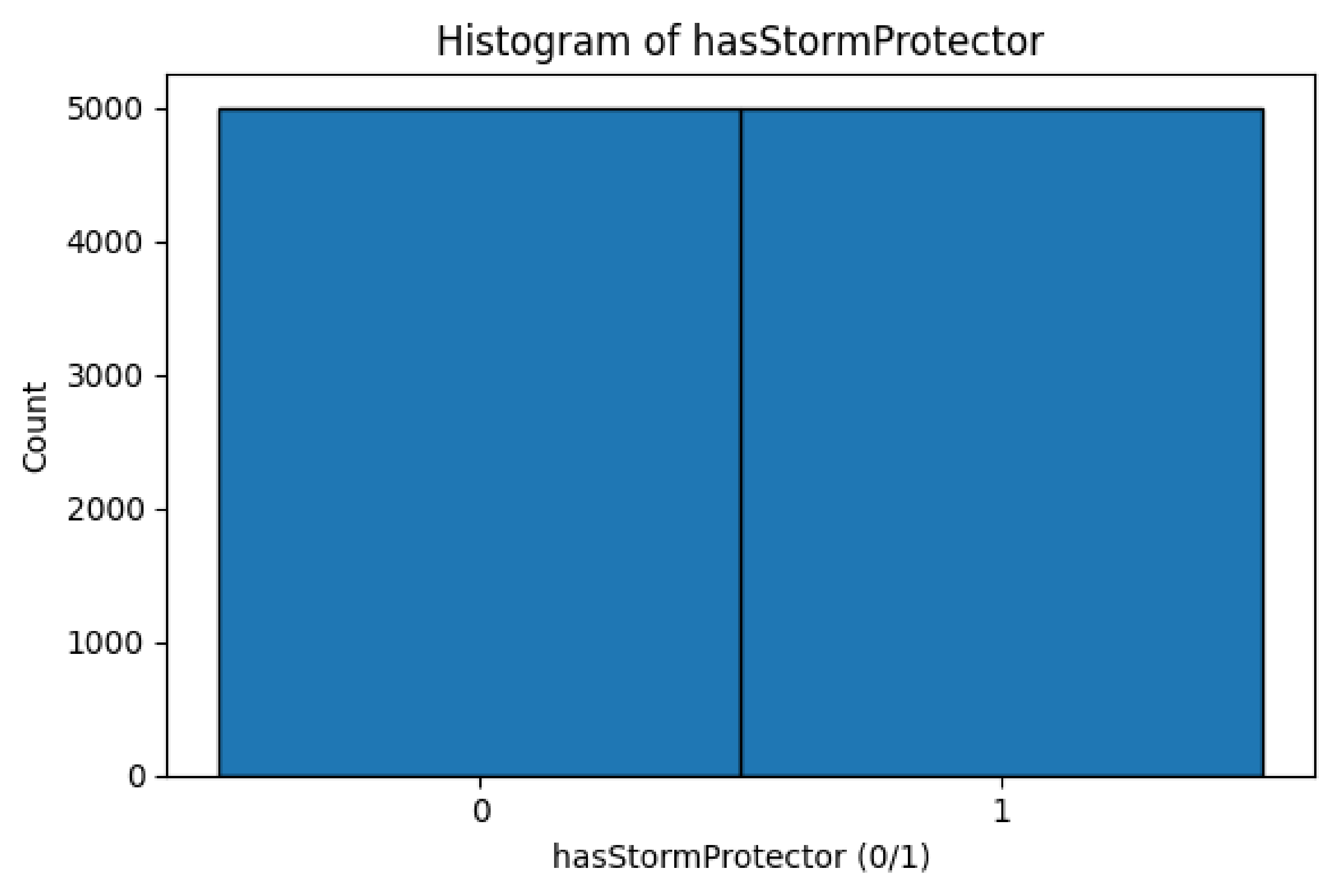
As can be seen, all housing built between 1992 and 2020 is represented equally at around 375 each. The only deviation from the norm is the year 1990 and 2021 with over 600 buildings built in these years. We can observe a pattern between it and other graphs.

HAS STORM PROTECTOR

- Values are true or false(1 or 0).

CONCLUSIONS:

Data is distributed equally, so we will be able to see the impact of having storm protection in our dataset.
Probably for each unit of housing there is a similar (not equal) number with or without storm protector.



PICTURE(5)

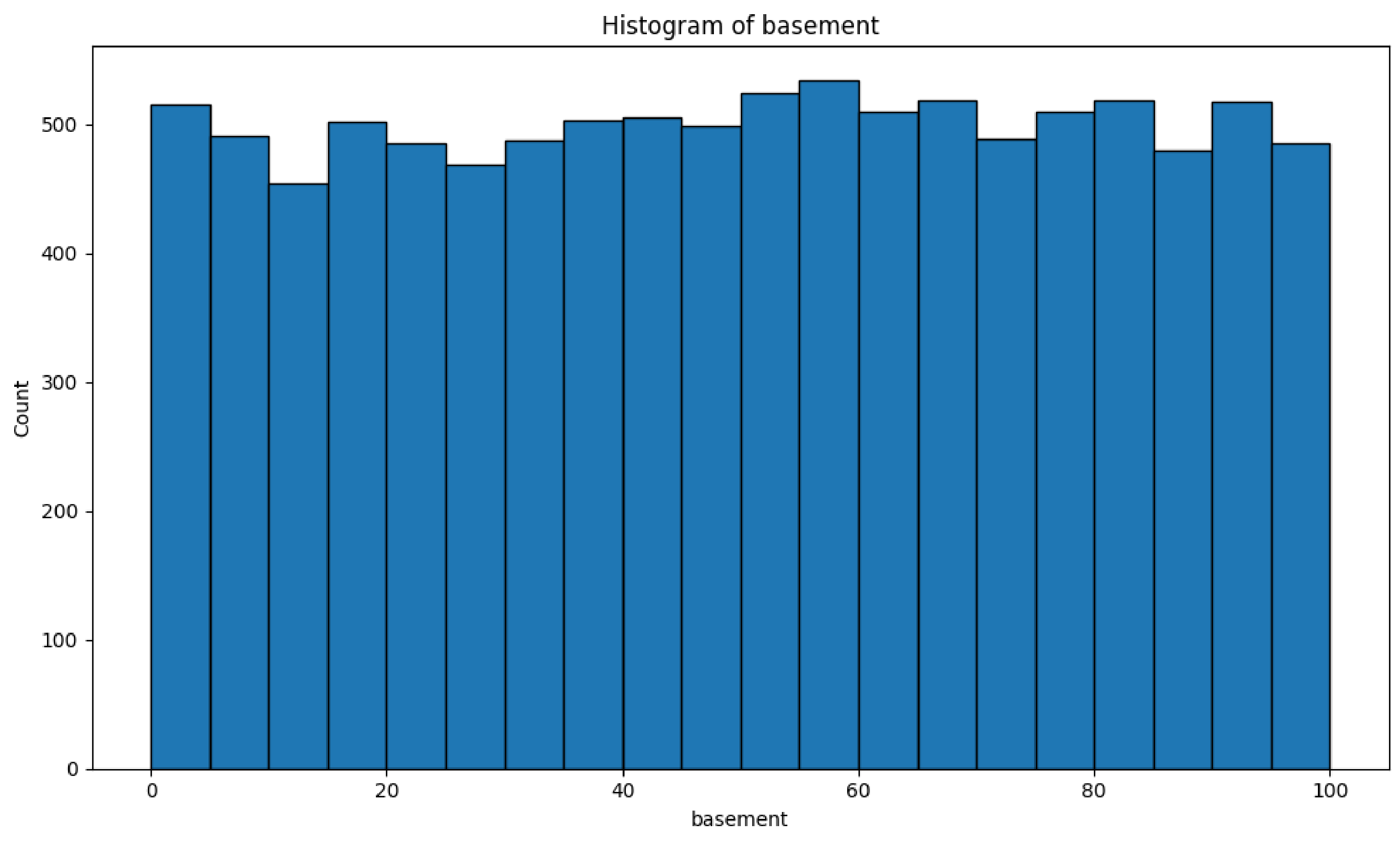
Most of the housing has a yard. It is distributed equally at 5000 counts each.

BASEMENT

- min = 0.00
- max = 100.00
- avg = 50.33
- med = 50.93
- std = 28.77

CONCLUSIONS:

A value at 0 means that the flat does not have it.
A maximum value at 100 belongs to a more expensive building.
Average value at 50 m^2 and maximum at 100 might mean that this is the shared area of the whole basement between all residents.

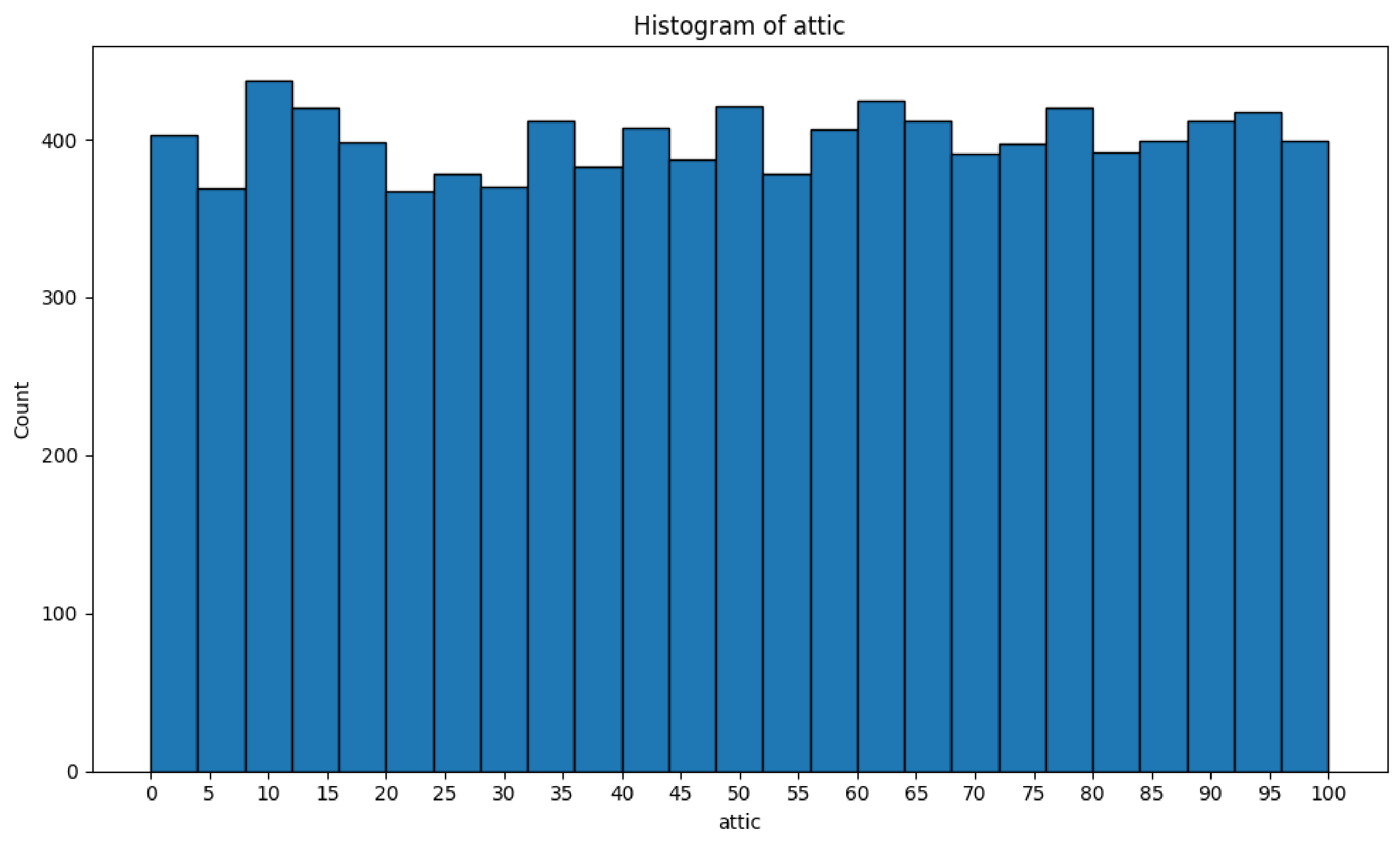


PICTURE(5)

As we can see, almost all m^2 of basements are distributed equally inside the dataset, with counts at around 500 for all categories.

ATTIC(M²)

- min = 1.00
- max = 100.00
- avg = 50.28
- med = 50.45
- std = 28.94



CONCLUSIONS:

Minimum value at 1 might mean that everyone has some place at shared attic.

After research, we came to the conclusion that the maximum value at 100 or average at 50 suggests that this is a shared area between all residents. Only more expensive apartments have private attics.

PICTURE(5)

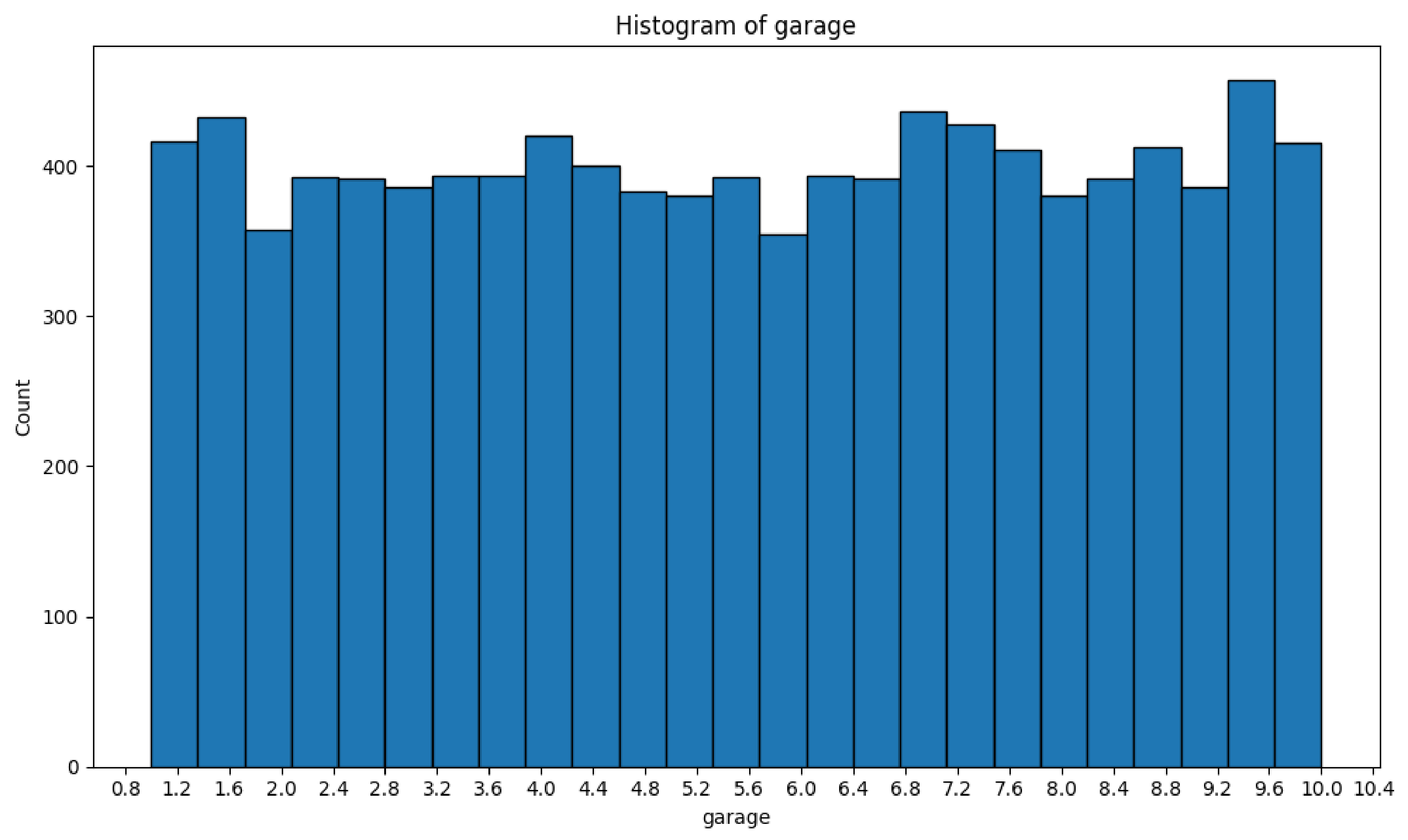
As we can see, almost all m^2 of attics are distributed equally inside the dataset with counts at around 400 for all categories.

GARAGE(M²)

- min = 0.01
- max = 10.00
- avg = 5.53
- med = 5.54
- std = 2.62

CONCLUSIONS:

Minimum value at 0.01 is probably an error, as well as below average 5.5 m^2. We will create two datasets, one with excluded garage size, to show the impact of this parameter.



PICTURE(5)

After seeing the distribution of this data, we can see that there might be an error in the dataset.

Or these are shares of common space.

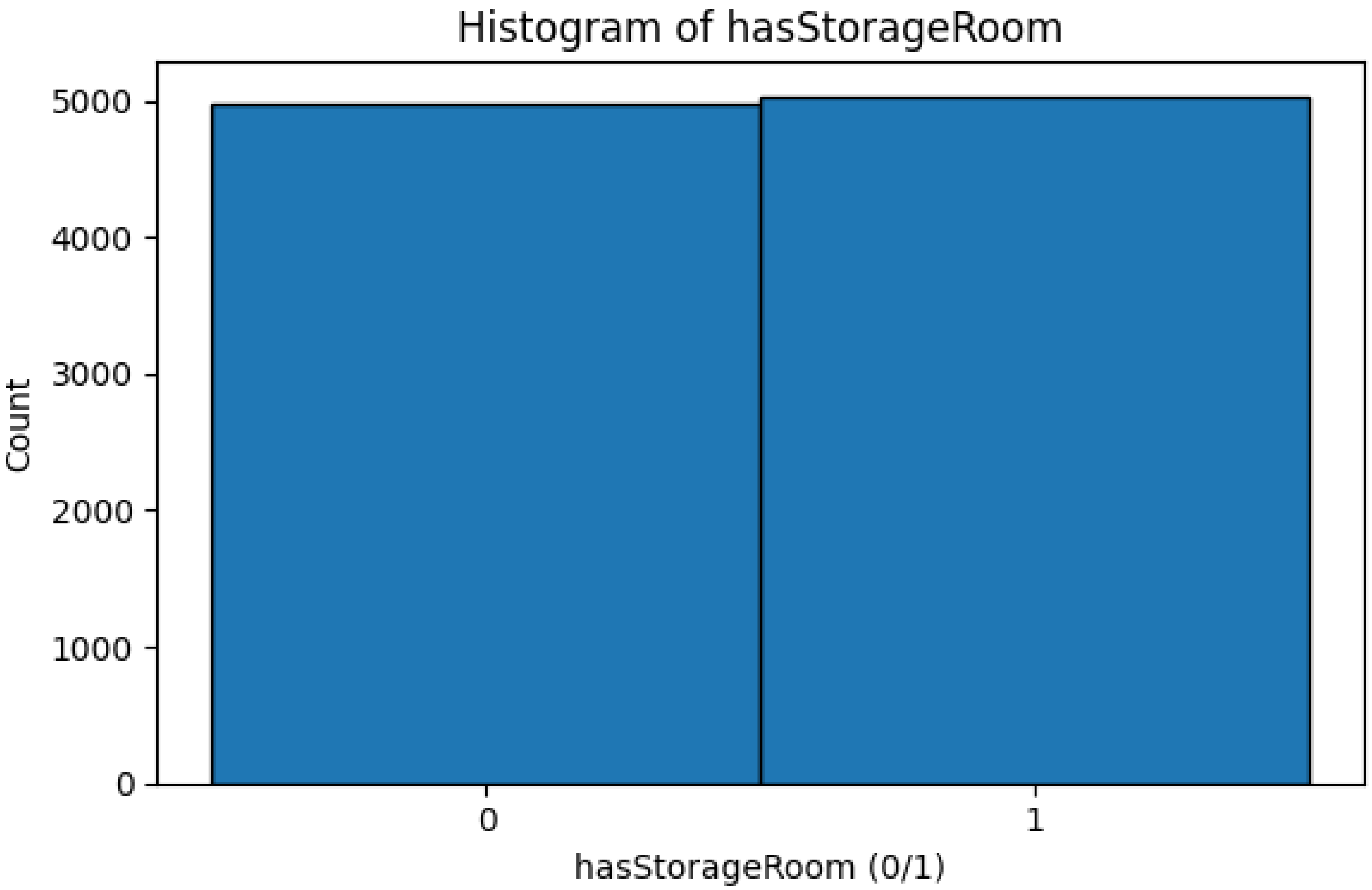
HASSTORAGE ROOM

- Values are true or false(1 or 0).

CONCLUSIONS:

Data is distributed equally, so we will be able to see the impact of having a storage room in our dataset.

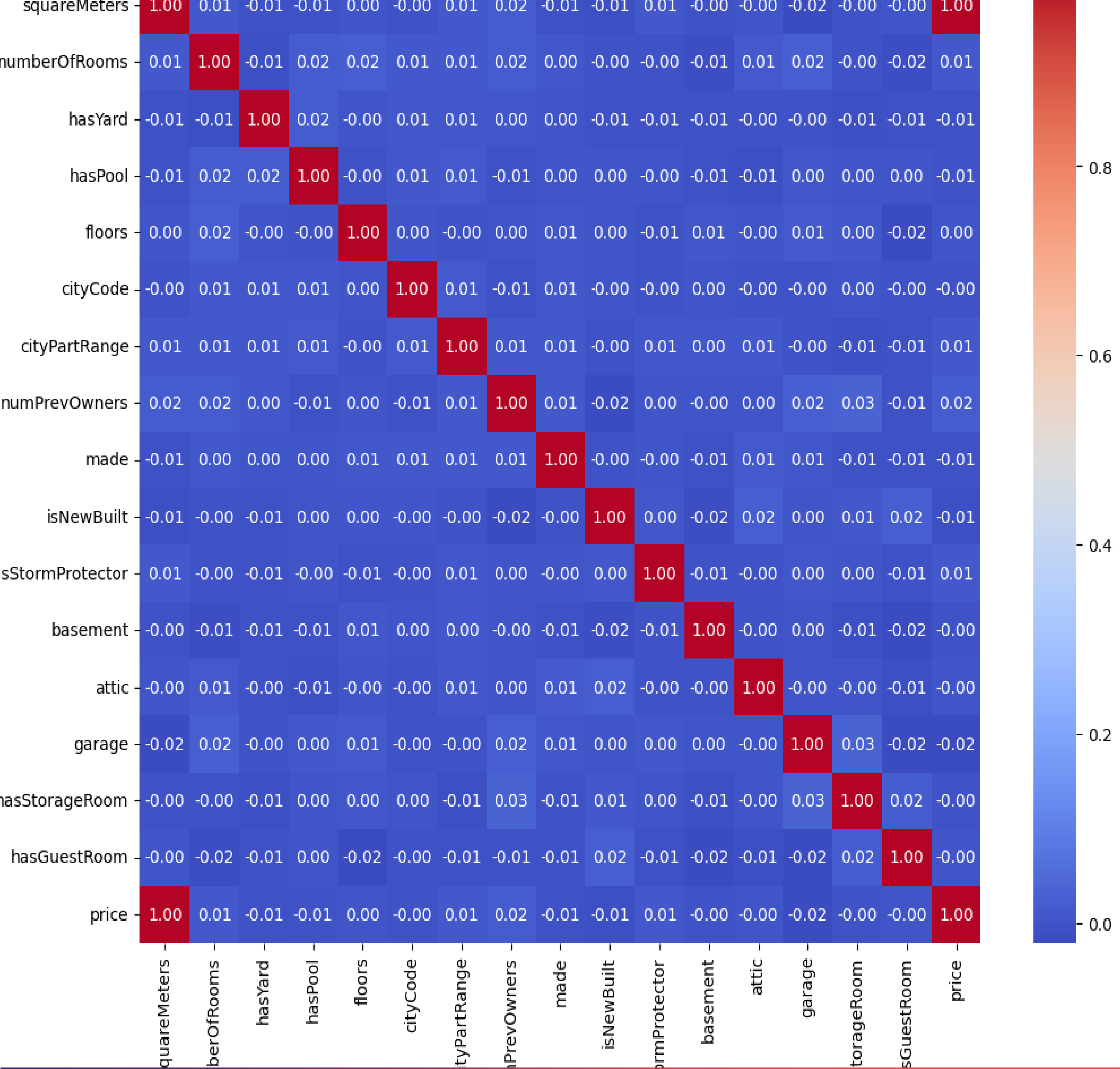
In some cases, a storage room might not be included in housing because it might be located in the basement and counted as its area.



PICTURE(5)

After seeing the distribution of this data, we can see that there might be an error in the dataset.

Or these are shares of common space



	quareMeters	berOfRooms	hasYard	hasPool	floors	cityCode	tyPartRange	PrevOwners	made	isNewBuilt	rmProtector	basement	attic	garage	torageRoom	sGuestRoom	price
squareMeters	1.00	0.01	-0.01	-0.01	0.00	-0.00	0.01	0.02	-0.01	-0.01	0.01	-0.00	-0.00	-0.02	-0.00	-0.00	1.00
numberOfRooms	0.01	1.00	-0.01	0.02	0.02	0.01	0.01	0.02	0.00	-0.00	-0.00	-0.01	0.01	0.02	-0.00	-0.02	0.01
hasYard	-0.01	-0.01	1.00	0.02	-0.00	0.01	0.01	0.00	0.00	-0.01	-0.01	-0.01	-0.00	-0.00	-0.01	-0.01	-0.01
hasPool	-0.01	0.02	0.02	1.00	-0.00	0.01	0.01	-0.01	0.00	0.00	-0.00	-0.01	-0.01	0.00	0.00	0.00	-0.01
floors	0.00	0.02	-0.00	-0.00	1.00	0.00	-0.00	0.00	0.01	0.00	-0.01	0.01	-0.00	0.01	0.00	-0.02	0.00
cityCode	-0.00	0.01	0.01	0.01	0.00	1.00	0.01	-0.01	0.01	-0.00	-0.00	0.00	-0.00	-0.00	0.00	-0.00	-0.00
cityPartRange	0.01	0.01	0.01	0.01	-0.00	0.01	1.00	0.01	0.01	-0.00	0.01	0.00	0.01	-0.00	-0.01	-0.01	0.01
numPrevOwners	0.02	0.02	0.00	-0.01	0.00	-0.01	0.01	1.00	0.01	-0.02	0.00	-0.00	0.00	0.02	0.03	-0.01	0.02
made	-0.01	0.00	0.00	0.00	0.01	0.01	0.01	0.01	1.00	-0.00	-0.00	-0.01	0.01	0.01	-0.01	-0.01	-0.01
isNewBuilt	-0.01	-0.00	-0.01	0.00	0.00	-0.00	-0.00	-0.02	-0.00	1.00	0.00	-0.02	0.02	0.00	0.01	0.02	-0.01
sStormProtector	0.01	-0.00	-0.01	-0.00	-0.01	-0.00	0.01	0.00	-0.00	0.00	1.00	-0.01	-0.00	0.00	0.00	-0.01	0.01
basement	-0.00	-0.01	-0.01	-0.01	0.01	0.00	0.00	-0.00	-0.01	-0.02	-0.01	1.00	-0.00	0.00	-0.01	-0.02	-0.00
attic	-0.00	0.01	-0.00	-0.01	-0.00	-0.00	0.01	0.00	0.01	0.02	-0.00	-0.00	1.00	-0.00	-0.00	-0.01	-0.00
garage	-0.02	0.02	-0.00	0.00	0.01	-0.00	-0.00	0.02	0.01	0.00	0.00	0.00	-0.00	1.00	0.03	-0.02	-0.02
hasStorageRoom	-0.00	-0.00	-0.01	0.00	0.00	0.00	-0.01	0.03	-0.01	0.01	0.00	-0.01	-0.00	0.03	1.00	0.02	-0.00
hasGuestRoom	-0.00	-0.02	-0.01	0.00	-0.02	-0.00	-0.01	-0.01	-0.01	0.02	-0.01	-0.02	-0.01	-0.02	0.02	1.00	-0.00
price	1.00	0.01	-0.01	-0.01	0.00	-0.00	0.01	0.02	-0.01	-0.01	0.01	-0.00	-0.00	-0.02	-0.00	-0.00	1.00

CORRELATION MATRIX

The price is most dependent on the number of square meters. Other things are not much correlated with it.

A similar situation can be observed in other countries where demand is so high that other things don't matter.

Even having a pool does not affect the price much.

Number of floor and previous owners does not affect it either

SUMMARY

DISTRIBUTION

All data have a similar distribution - equal counts of values rather than a messy one

BINARY DATA

We can assume that our categorical data is split evenly, ensuring that everything is correct.

CORRELATION MATRIX

The correlation matrix appears incorrect. Something must have gone wrong - it’s likely a bug in the Matplotlib library.

ERRORS IN DATASET

There are some errors in the dataset.
For instance, a flat listed as being below 1 m^2,
and a garage listed as being below 0.1 m^2.

NOVELTY OF THE APPROACH AND IMPORTANCE OF THE PROBLEM

NOVELTY OF OUR APPROACH

- Hybrid architecture: **Deep Learning model + AI Agent (Gemini)**

- Real-time predictions with **intelligent reasoning**

- Integration with conversation-based interface, so predictive model becomes interactive, accessible, and user-friendly

- Fully automated pipeline: data loading, preprocessing, model, agent reasoning, prediction

Two-level neural network setup:

- Simple baseline (Adaline-like model)
- More advanced **Multi-layer model with separate processing for numerical & boolean features**

WHY IS THIS PROBLEM IMPORTANT?

Importance of housing price prediction

- Housing prices = one of the most important economic indicators
- High demand + limited supply in Paris = strong price volatility
- **App useful for:**
  - Buyers & renters
  - Urban planners
  - Real estate agencies
  - Government policy & tax planning
- Manual evaluation of real estate value = slow, expensive, and often inaccurate

WHY OUR SOLUTION IS USEFUL AND BETTER THAN TYPICAL METHODS

Benefits of our approach

- Improved usability - NLP
- Flexible inputs
- Speed
- Better adaptability
- Lower cost, no specialists needed
- Scalability
- More accurate than simple ML models:
  - Neural network models capture non-linear relationships
  - Multi-layer model analyzes numerical and categorical data separately -> more precized infrastructure

COMPARISON WITH OTHER APPROACHES

TRADITIONAL APPROACHES

- Manual appraisal
- Linear regression – to simple, ignores non-linear relations
- Standard neural networks – no interactions with user
- Decision trees / Random Forest – inflexible approach and without user influence

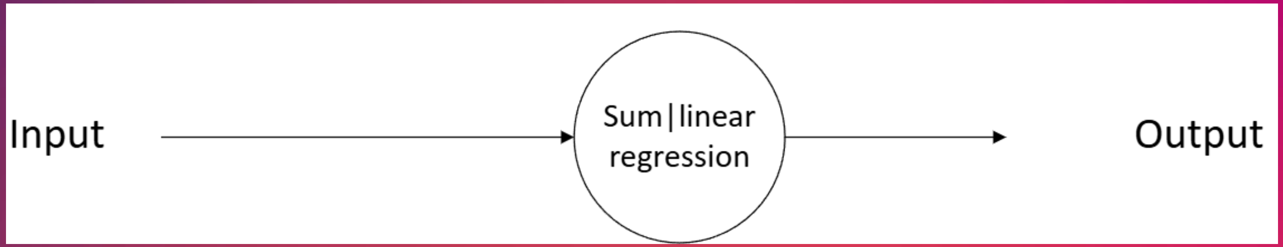
OUR SOLUTION

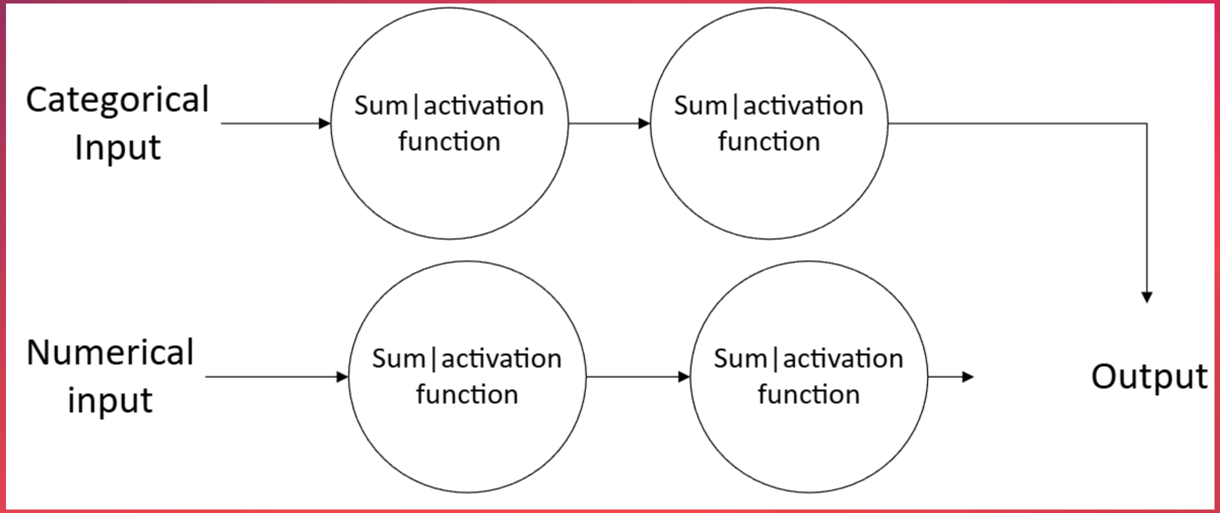
- Deep Learning model + AI agent reasoning
- Predictive + interactive system
- LLM tuned for specific housing data
- Flexible model adaptation
- Dialog graphical user interface, for frendly-usage

SYSTEM DESIGN AND DESIGN RATIONALE

Neural network- System design

Project splits into two main parts: neural network and agents. Neural network is responsible for giving right housing prices predictions. To achieve that we are using simple but effective one-layer linear regression. That decision is based on data and expected results which are numbers hard to classify. Another approach was to split categorical and numeric data by using residual network. For activation function we are also trying different options like tanh, sigmoid or relu.





Agentic system – System design

For agentic system we are using Google Gemini AI. Project consists of one agent. House price prediction agent is responsible for collecting house parameters from user and based on that it is going to predict price for that type of house. If some parameters are missing agent should give three answers: the lowest possible price, medium quality house and the most expensive house price matching given parameters. All communication with agent is going to be on user interface.

EXPLANATION OF USED TECHNIQUES USED AND DEVELOPED

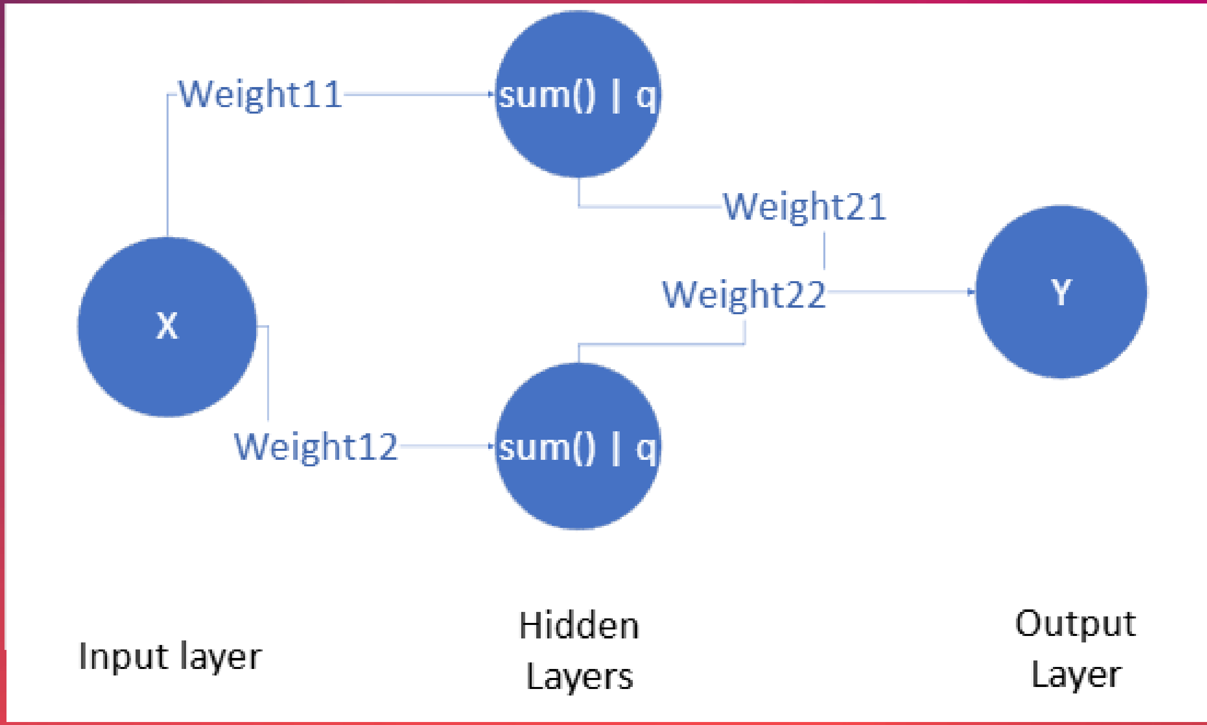
Neural network

LINEAR REGRESSION

ACTIVATION FUNCTIONS: RELU, SIGMOID, TANH

Neural network – Explanation of techniques

Neural network is structure based on human brain. There are three types of layer: input, hidden layers, output. Every neuron consist of sum function which sum received data from previous layer multiplied by wages, activation function which process summary from neuron input. After that output is send to next layer. Every connection has weight and they are calculated during process called learning. There are three types of learning: supervised, unsupervised and reinforced. In project we are using supervised learning. Using back propagation all weights are changed to make output closer to desired outcome.



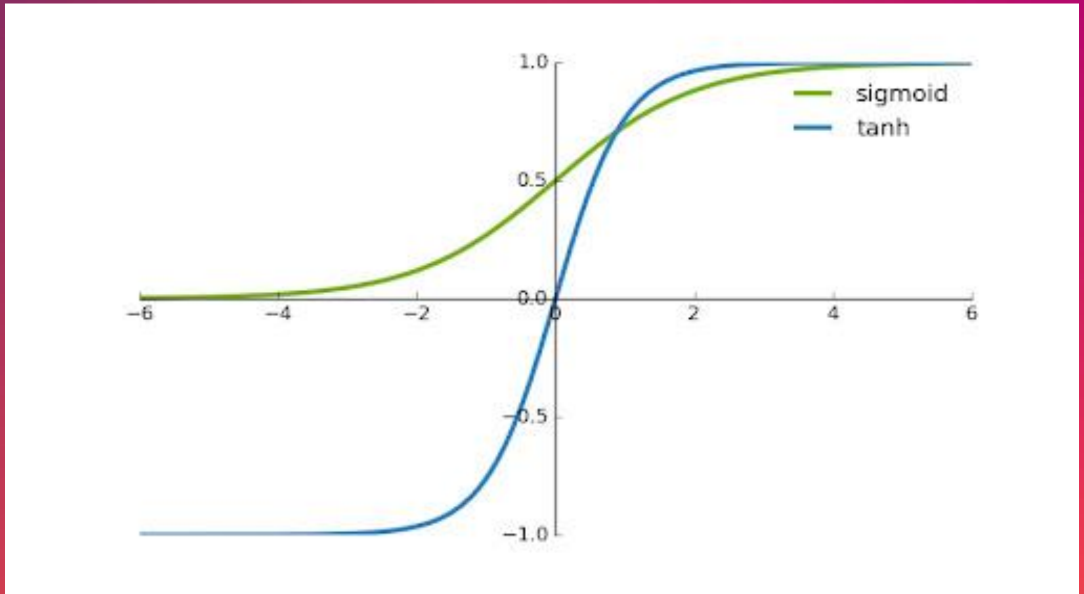
Activation functions – Explanation of techniques

Simple linear regression – using one variable we are trying to get all data as close to function as possible. Function formula: ax

Sigmoid – gives values from 0 to 1 or –1 to 1. Function formula: $\frac{1}{1+e^{-x}}$

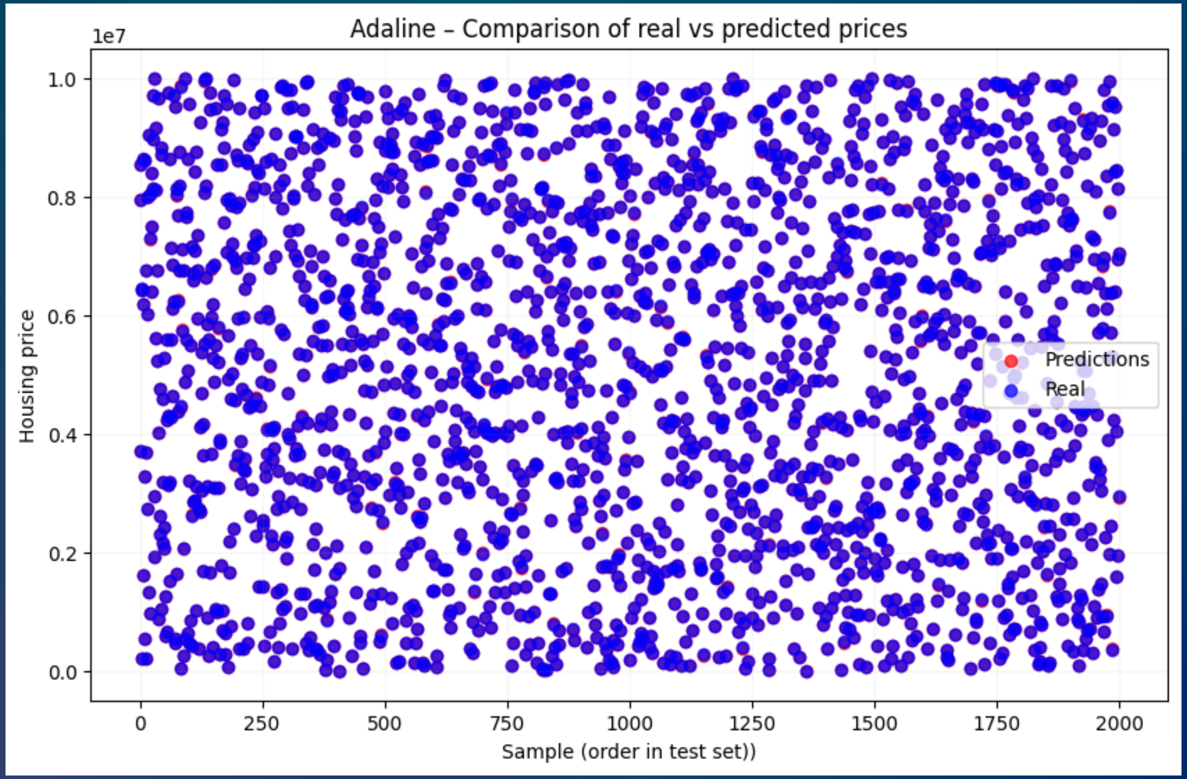
ReLu – gives values from 0 to infinity. $\begin{cases} 0, x < 0 \\ x, x \geq 0 \end{cases}$

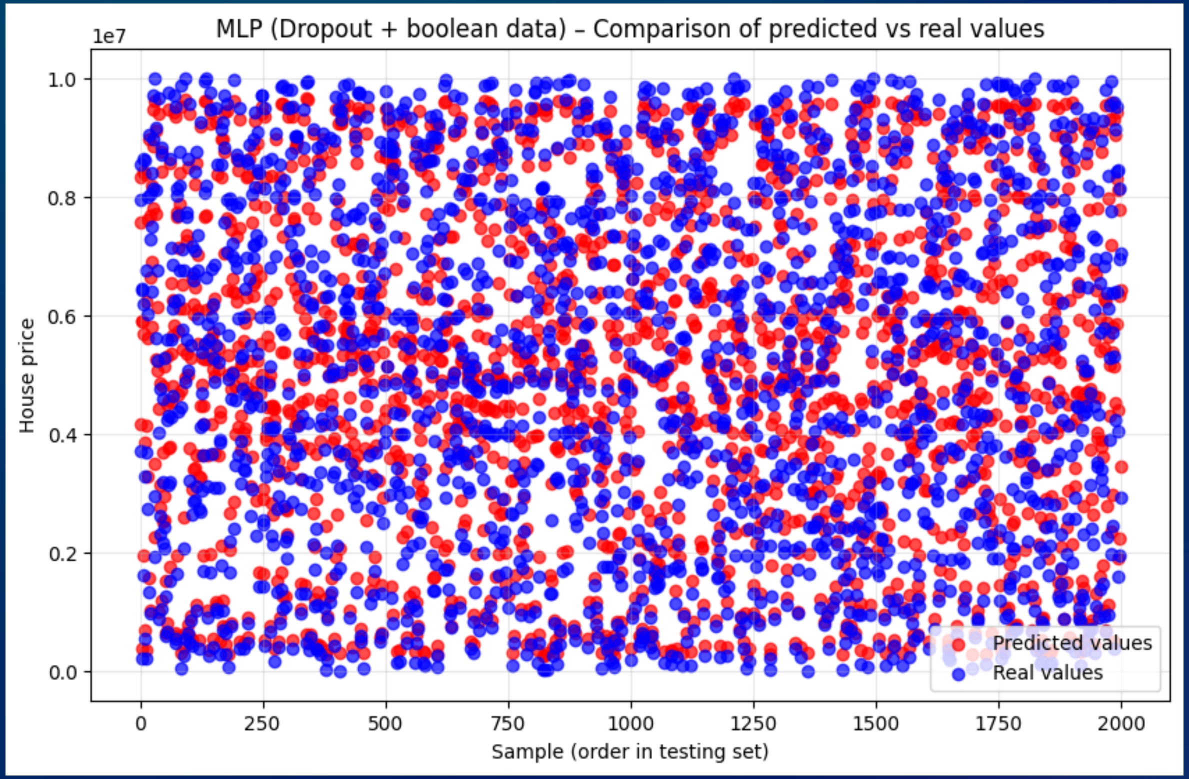
Tanh- Gives values from –1 to 1. Function formula: $\frac{e^{2x} - 1}{e^{2x} + 1}$



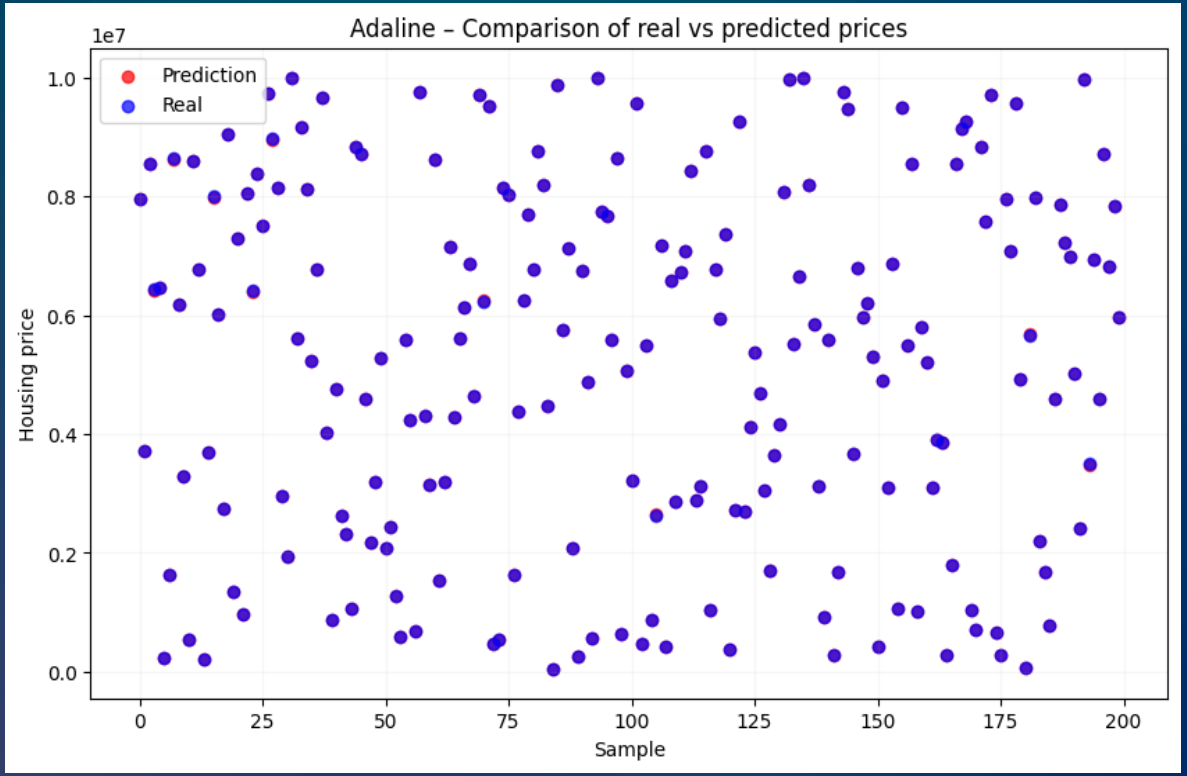
PRELIMINARY PERFORMANCE ANALYSIS

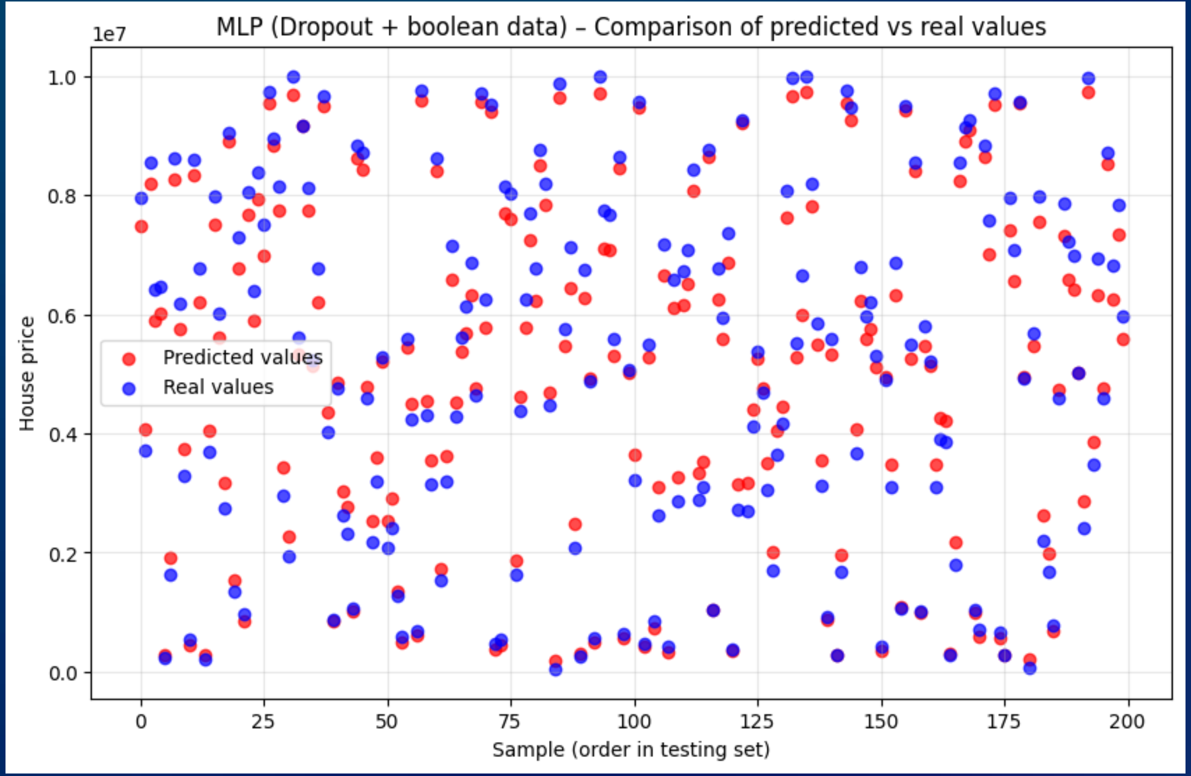
Predicted vs Real values for 2000 samples





Predicted vs Real values for 200 samples

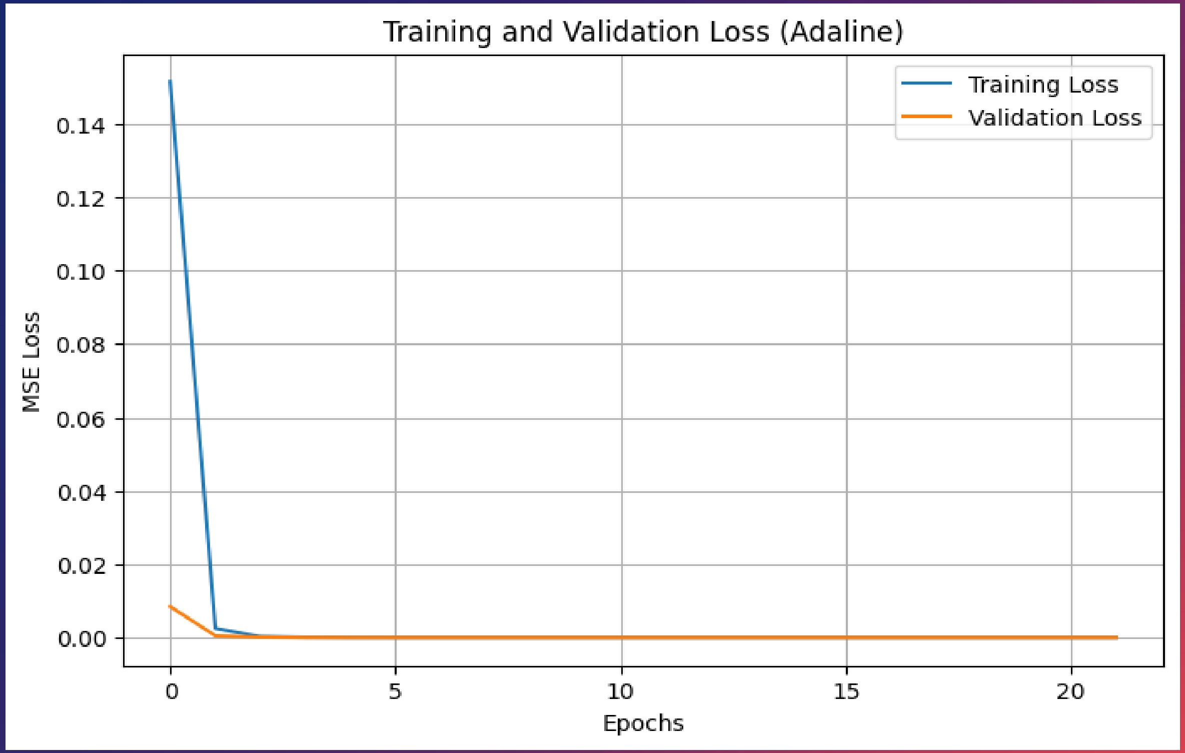


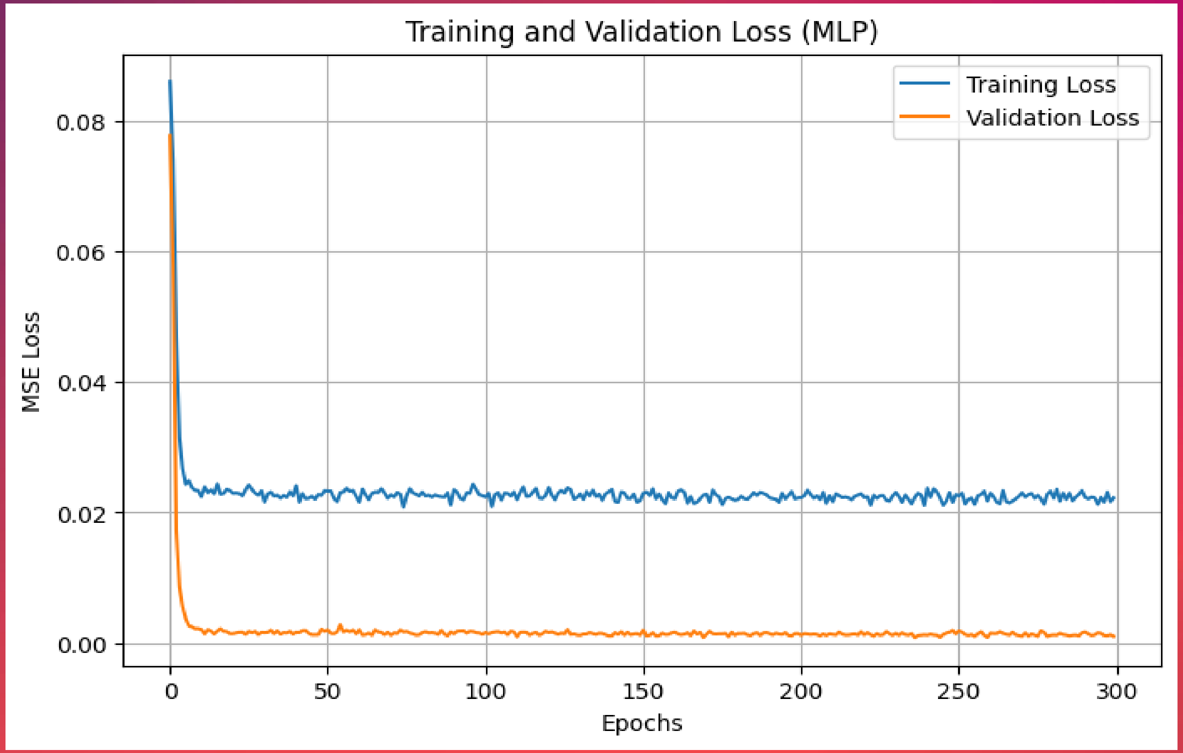


Predicted vs Real values

	🟢 **MLP (Multi-Layer Perceptron)**	🔴 **Adaline (Adaptive Linear Neuron)**
Prediction Quality	**Strong Regressor.** Predictions (Red dots) closely track the Real values (Blue dots) across the entire price range.	**Poor Regressor.** Predictions (Red dots) are severely **restricted** and concentrated, failing to capture the wide variance of prices.
Tracking	**Successful.** The model learned the complexity of the data, resulting in minimal and random errors.	**Failed.** The model largely predicts a value close to the mean, indicating **Underfitting** due to its simplicity.
Takeaway	**Excellent Generalization.** Provides high-quality, reliable price estimates.	**Inadequate.** Not suitable for this complex, non-linear dataset.

Loss functions





Loss functions

	🟢 **MLP (Multi-Layer Perceptron)**	🔴 **Adaline (Adaptive Linear Neuron)**
Convergence Speed	**Steady and controlled.** Too approximately 50 epochs for the initial sharp drop, with refinement continuing up to 300 epochs.	**Instantaneous.** Dropped to near-zero loss within the first **2 epochs**.
Final Validation Loss	**Low and Stable.** Demonstrates excellent generalization to unseen data.	**Deceptively Low.** A near-zero loss value is **misleading** given the poor predictive power shown in the scatter plots.
Takeaway	**Robust Learning.** The higher complexity (more parameters) requires more epochs but achieves a genuinely accurate solution.	**Shallow Learning.** The model quickly found a simple, sub-optimal linear solution that minimizes MSE *on paper* but fails *in practice*.

Error comparison

	🟢 **MLP (Multi-Layer Perceptron)**	🔴 **Adaline (Adaptive Linear Neuron)**
Visual Performance	**Strong Regressor.** Predictions closely track the wide range of real house prices.	**Failed Regressor.** Predictions are heavily restricted and do not track the real values.
R^2 Score (Key Metric)	**0.81 (Strong)**	**Extremely Low** (Implied by failure to track variance).
Reported MAPE	12.10%	0.10%
Technical Interpretation	High R^2 and realistic 12.10% MAPE confirm **robust and accurate predictions**.	**MISLEADING *MAPE*.** Low value is an **artifact of normalization**; the model is simply predicting the mean of the scaled data (approx 0.0).
Final Conclusion	**Selected Model.** Performs best across all metrics and visualizations, demonstrating high predictive power.	**Rejected Model.** Lacks the complexity to solve this non-linear problem, resulting in **severe underfitting**.

THANK YOU FOR YOUR ATTENTION

SOURCES

- Paris panorama
  - https://www.klook.com/destination/c107-paris/
- Dataset:
  - https://www.kaggle.com/datasets/mssmartypants/paris-housing-price-prediction
- Gramar correction:
  - Gramarly
- Opening picture:
  - https://skillfloor.com/blog/data-analysis-industry-interpretation