Senior Design Project

House Price Prediction Using Ensemble Learning Techniques

Under the guidance of:-
Dr. Manomita Chakraborty

By:-
Raghunath Singh (20BCI7012)
Anuj (20BCE7055)

AGENDA

PROBLEM DEFINITION

ABOUT DATASET

METHODOLOGY

SUMMARY

REFERENCE

THANK YOU NOTE

PROBLEM DEFINITION

To design an Machine Learning Algorithm for the accurate prediction of House Price using Random Forest, Support Vector Regression, Boosting algorithm and Ensemble models.

ABOUT DATASET

The dataset has been taken from Kaggle.

It has following features:-
ID: Unique identifier for each house listing.
Date: Date of the house listing.
Number of bedrooms: The count of bedrooms in the house.
Number of bathrooms: The count of bathrooms in the house.
Living area: Total living area of the house in square feet.
Lot area: Total area of the lot in square feet.
Number of floors: Number of floors in the house.
Waterfront present: Indicates whether the house has a waterfront view (binary: 0 for no, 1 for yes).
Number of schools nearby: Number of schools located near the house.
Distance from the airport: Distance of the house from the nearest airport in miles.
Price: Price of the house.

ABOUT DATASET

Number of views: Number of views the house has received.
Condition of the house: Condition rating of the house.
Grade of the house: Grade rating of the house.
Area of the house (excluding basement): Total area of the house excluding the basement in square feet.
Area of the basement: Area of the basement in square feet.
Built Year: Year the house was originally built.
Renovation Year: Year of the last renovation, if any.
Postal Code: Postal code of the house location.
Latitude: Latitude coordinate of the house location.
Longitude: Longitude coordinate of the house location.
Living area after renovation: Total living area of the house after renovation in square feet.
Lot area after renovation: Total area of the lot after renovation in square feet.

Kaggle Link:- https://www.kaggle.com/datasets/mohamedafsal007/house-price-dataset-of-india

Methodology

Step 1: Data Exploration:

- Loaded the dataset and checked for missing values.
- Explored data types and basic statistics.
- We are not having any null value in our dataset.

In [4]:
```
pd.isnull(df).head()
```

Out[4]:

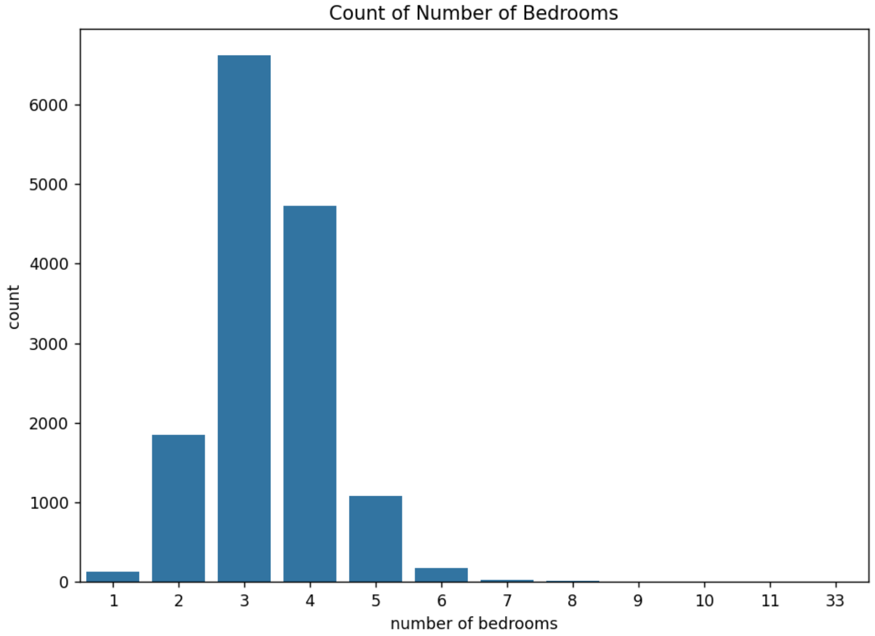
number of bathrooms	living area	lot area	number of floors	waterfront present	number of views	condition of the house	...	Built Year	Renovation Year	Postal Code	Lattitude	Longitude	living_area_renov	lot_area_renov	Number of schools nearby
False	False	False	False	False	False	False	...	False	False	False	False	False	False	False	False
False	False	False	False	False	False	False	...	False	False	False	False	False	False	False	False
False	False	False	False	False	False	False	...	False	False	False	False	False	False	False	False
False	False	False	False	False	False	False	...	False	False	False	False	False	False	False	False
False	False	False	False	False	False	False	...	False	False	False	False	False	False	False	False

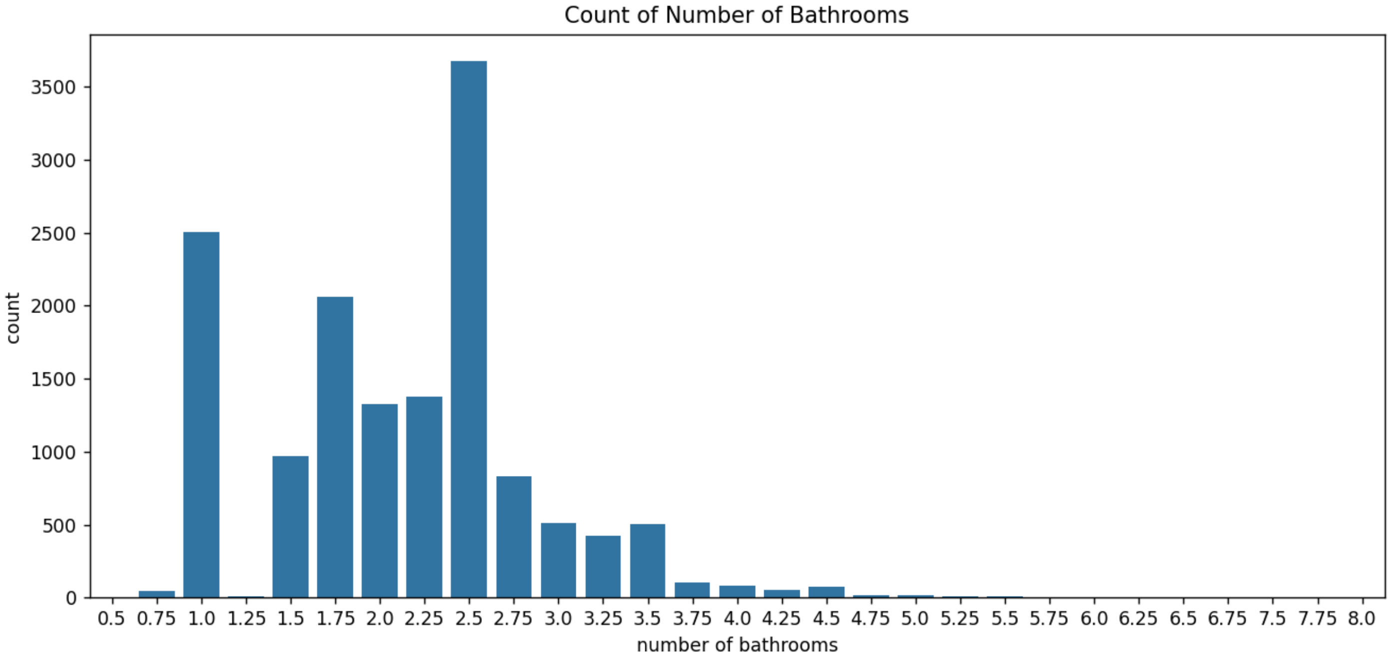
In [5]:
```
df.isna().sum()
```

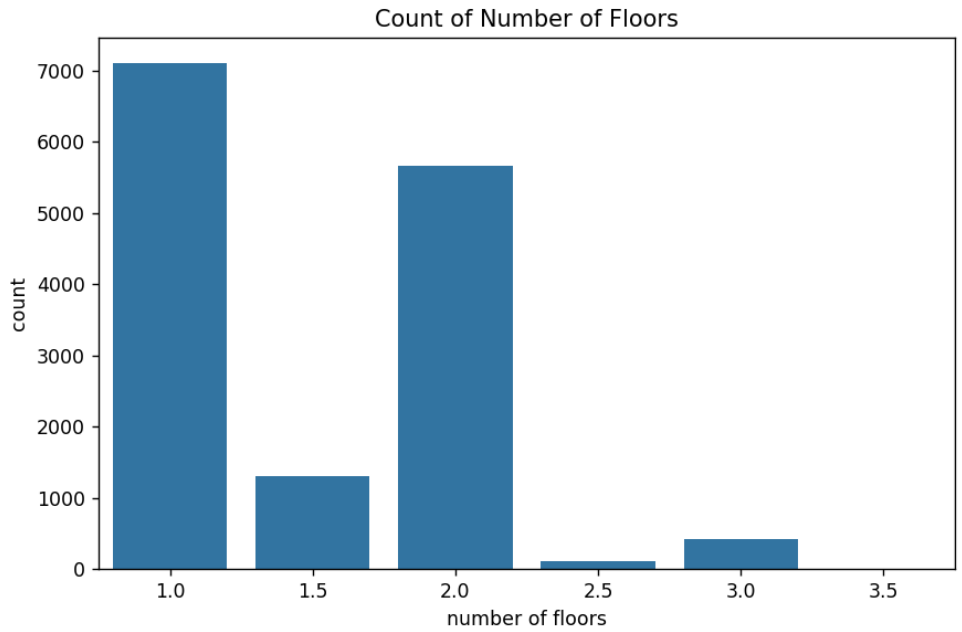
Out[5]:
```
id                                        0
Date                                      0
number of bedrooms                        0
number of bathrooms                       0
living area                               0
lot area                                  0
number of floors                          0
waterfront present                        0
number of views                           0
condition of the house                    0
grade of the house                        0
Area of the house(excluding basement)     0
Area of the basement                      0
Built Year                                0
Renovation Year                           0
Postal Code                               0
```

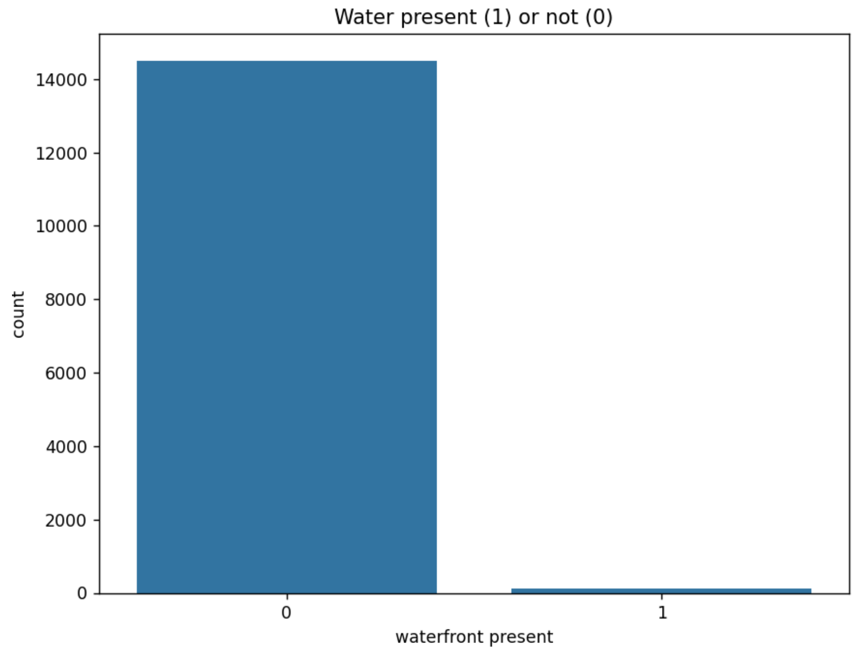
Step 2: Visualization:

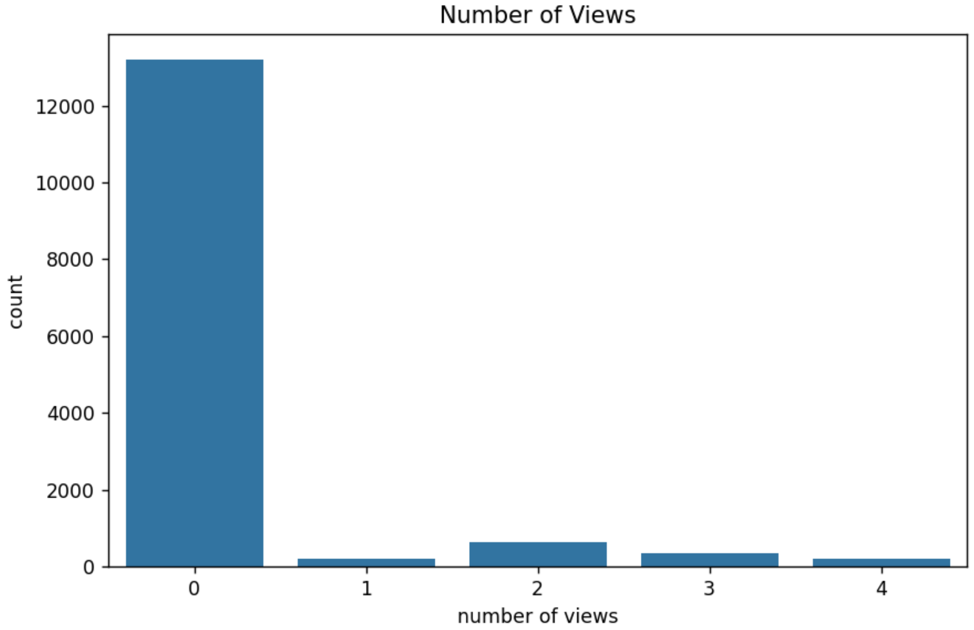
- Plotted count distributions for various features.
- Visualized geographical data and price distributions.
- Examined correlations and relationships between variables.

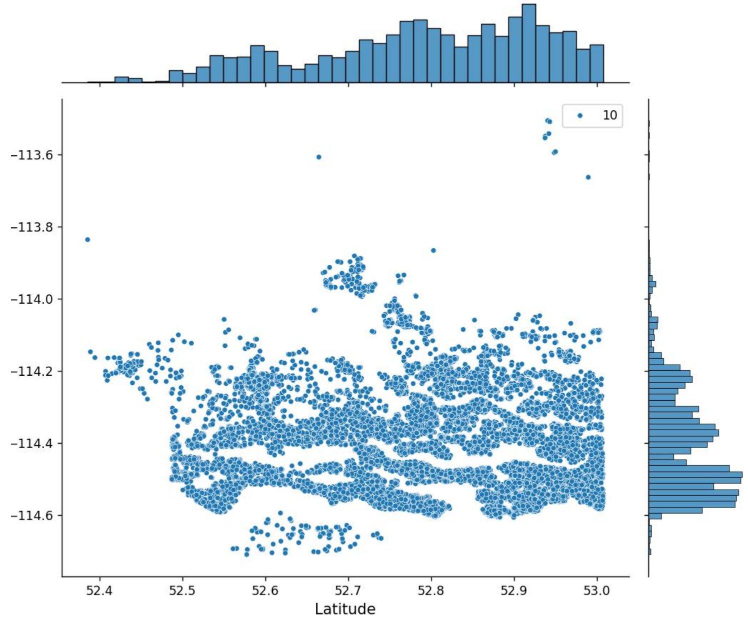


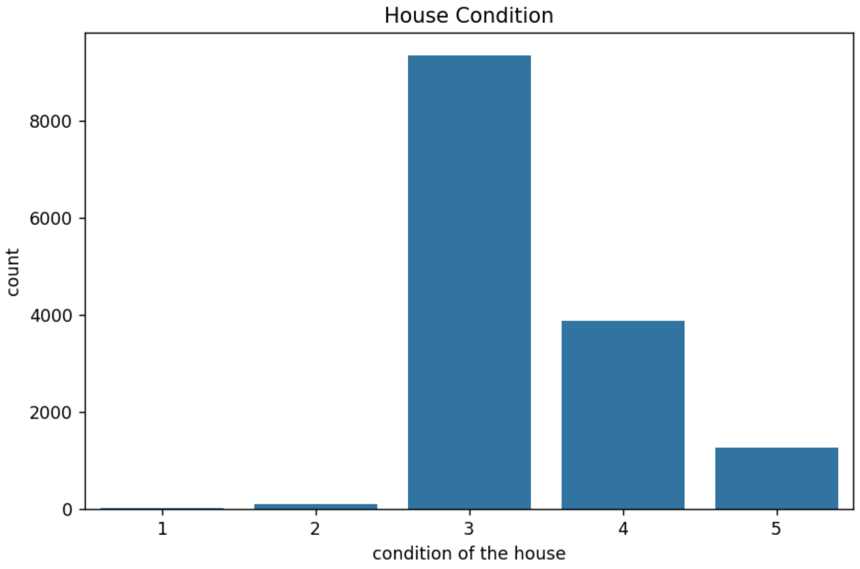


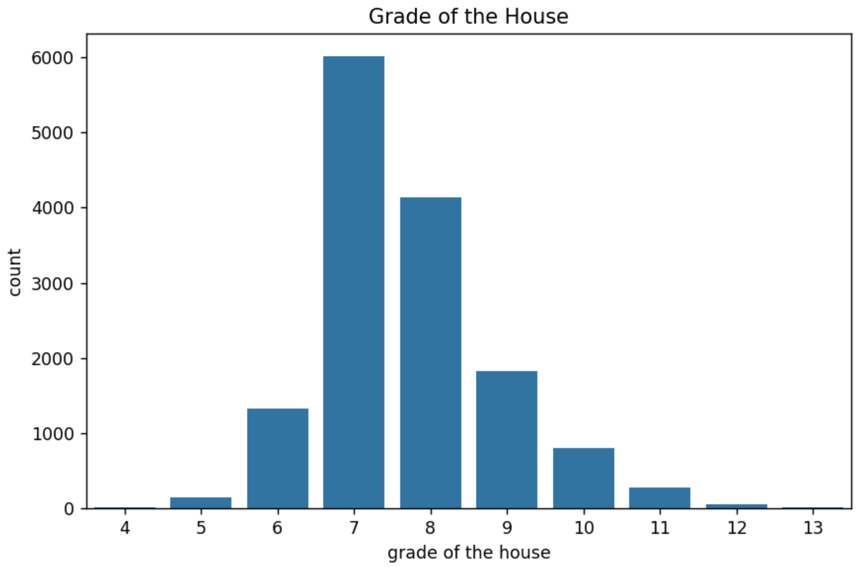

Count of Number of Floors
count
0
1000
2000
3000
4000
5000
6000
7000
1.0
1.5
2.0
2.5
3.0
3.5
number of floors

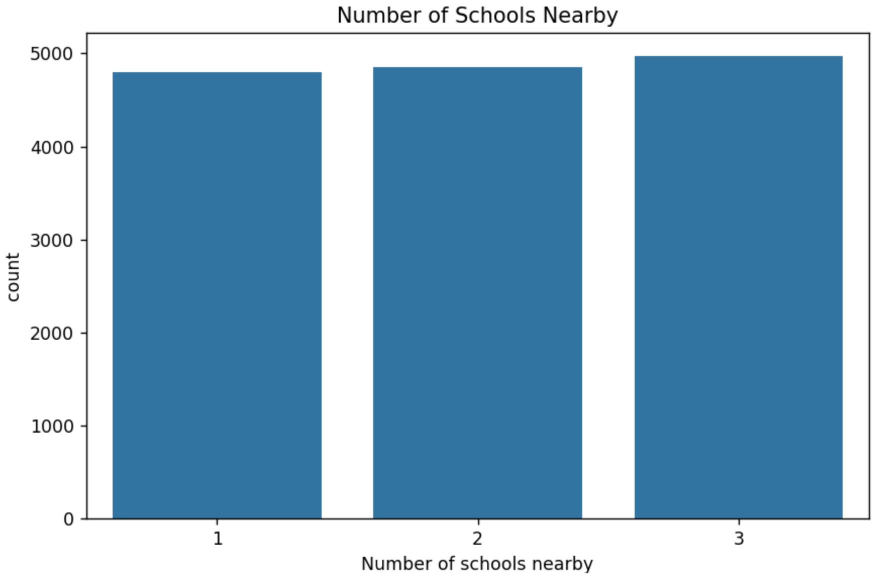

Water present (1) or not (0)
count
0
2000
4000
6000
8000
10000
12000
14000
0
1
waterfront present

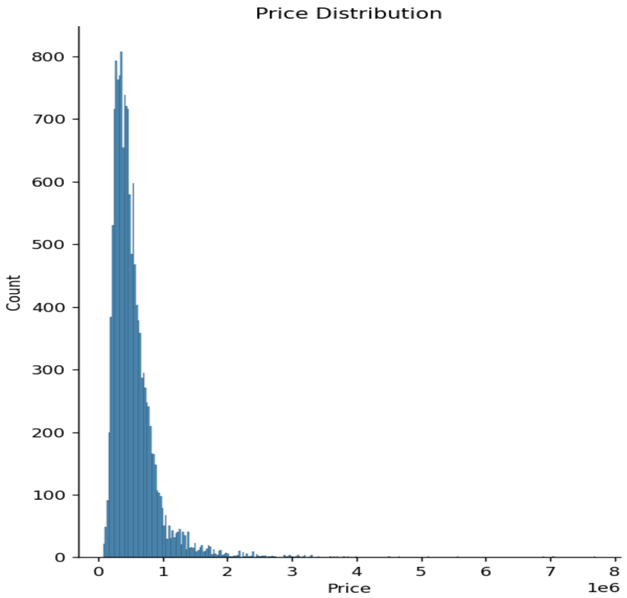

Number of Views
count
0
2000
4000
6000
8000
10000
12000
0
1
2
3
4
number of views

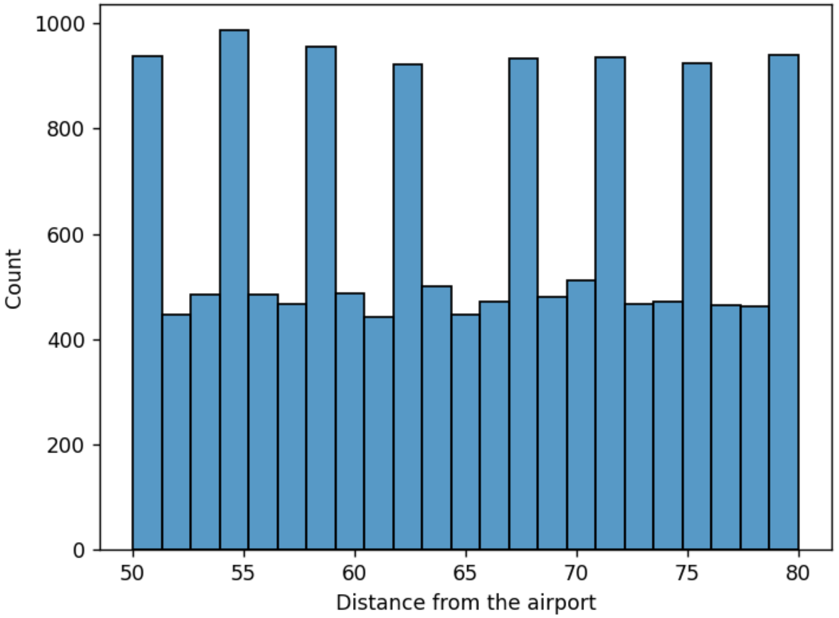

10
−113.6
−113.8
−114.0
−114.2
−114.4
−114.6
52.4
52.5
52.6
52.7
52.8
52.9
53.0
Latitude

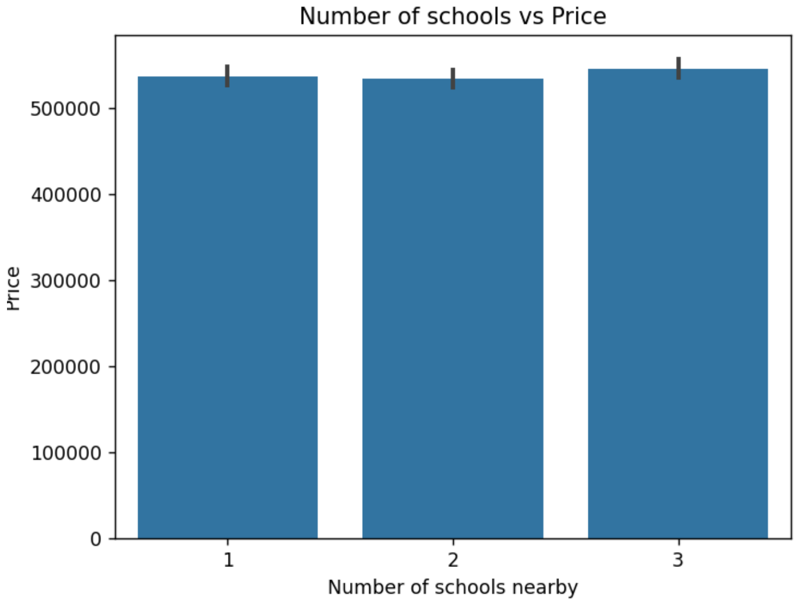

House Condition
count
0
2000
4000
6000
8000
1
2
3
4
5
condition of the house

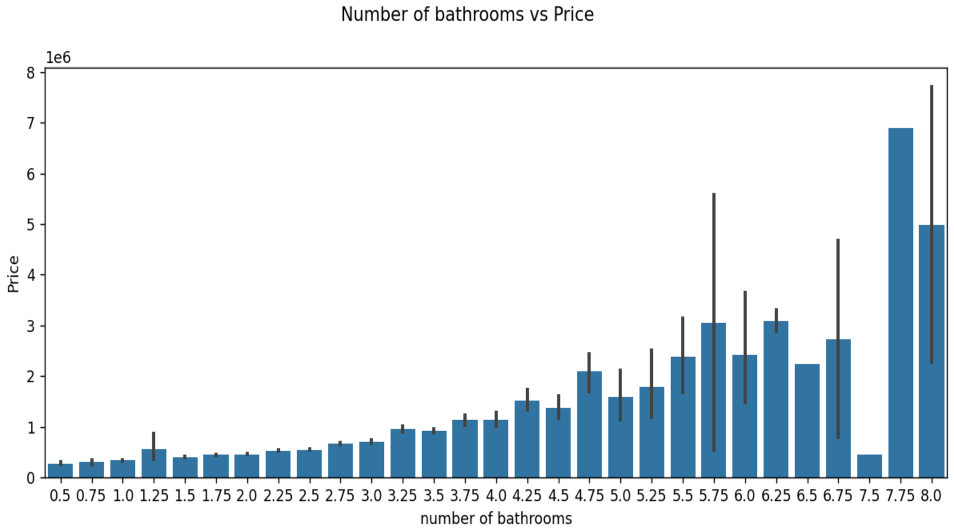

Grade of the House
count
0
1000
2000
3000
4000
5000
6000
4
5
6
7
8
9
10
11
12
13
grade of the house

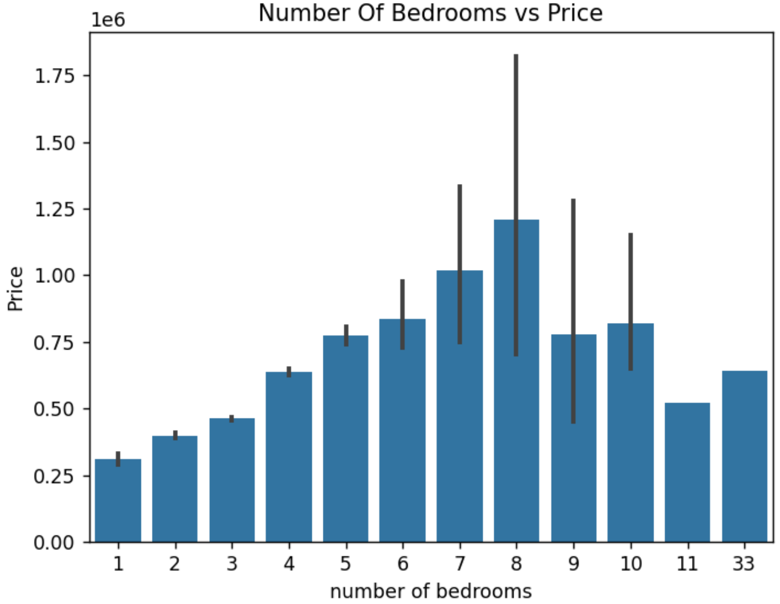

Number of Schools Nearby
count
0
1000
2000
3000
4000
5000
1
2
3
Number of schools nearby

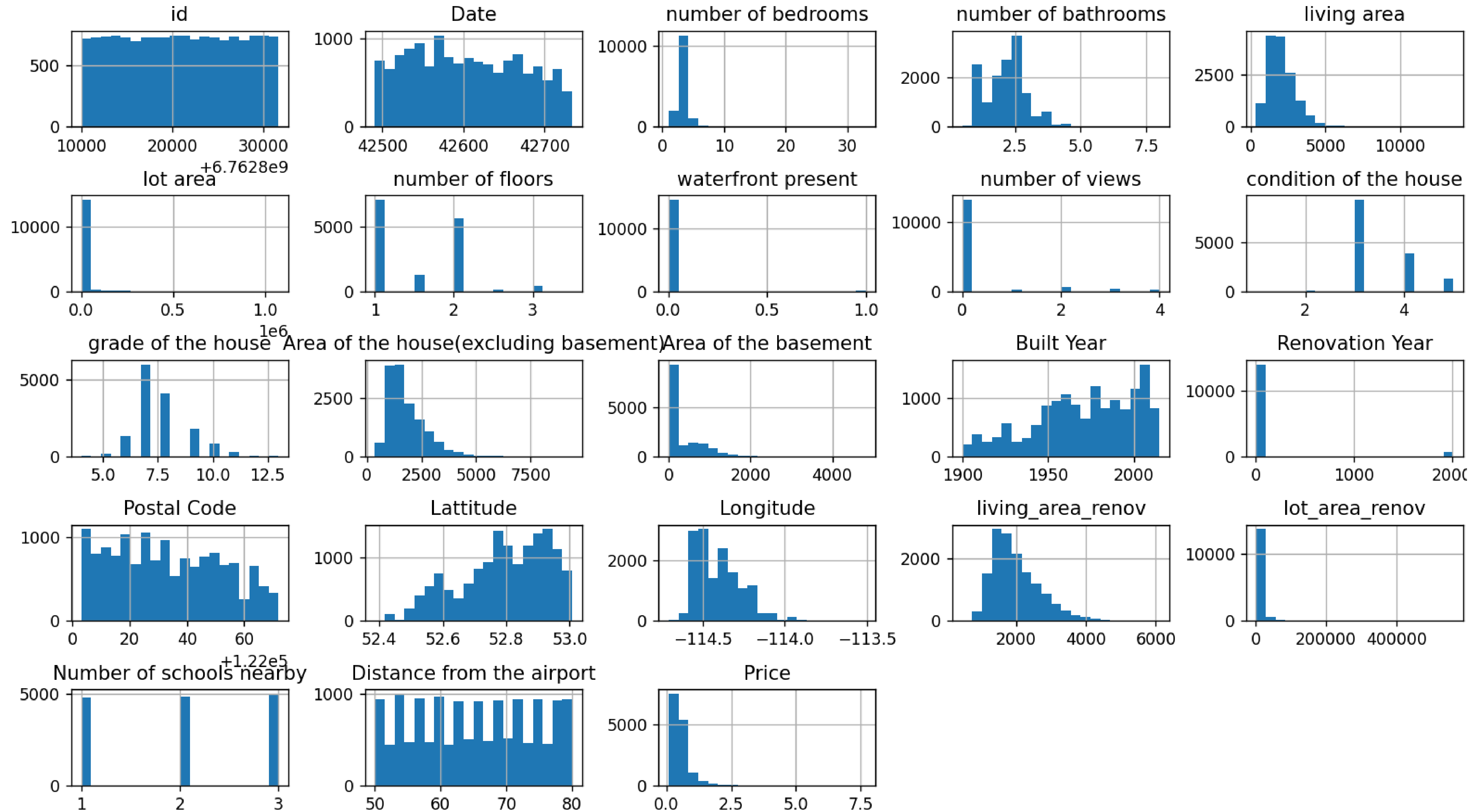

Price Distribution
Count
0
100
200
300
400
500
600
700
800
0
1
2
3
4
5
6
7
8
Price
1e6


Count
1000
800
600
400
200
0
50 55 60 65 70 75 80
Distance from the airport

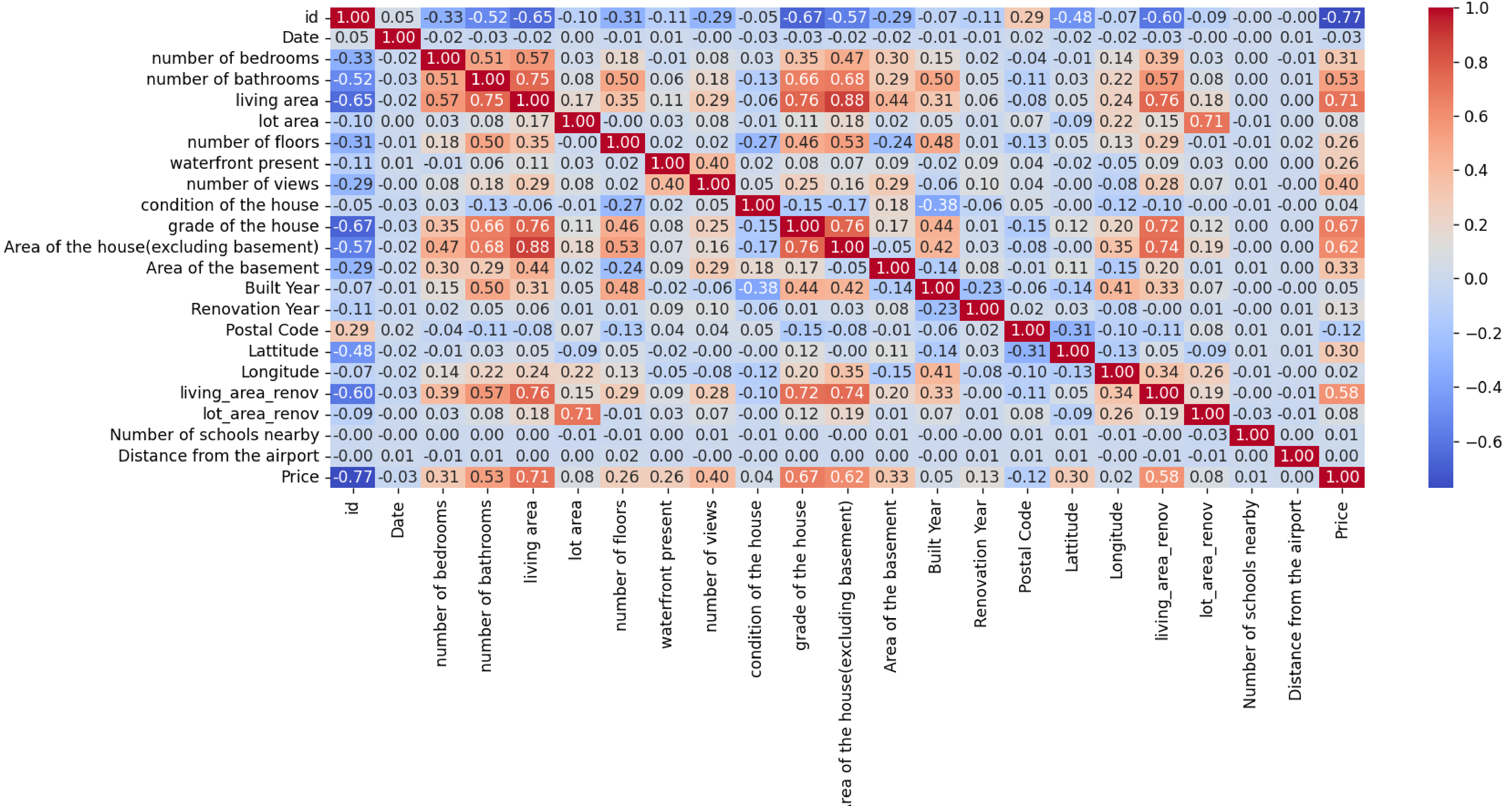

Number of schools vs Price
Price
500000
400000
300000
200000
100000
0
1 2 3
Number of schools nearby


Number of bathrooms vs Price
1e6
Price
8 7 6 5 4 3 2 1 0
0.5 0.75 1.0 1.25 1.5 1.75 2.0 2.25 2.5 2.75 3.0 3.25 3.5 3.75 4.0 4.25 4.5 4.75 5.0 5.25 5.5 5.75 6.0 6.25 6.5 6.75 7.5 7.75 8.0
number of bathrooms

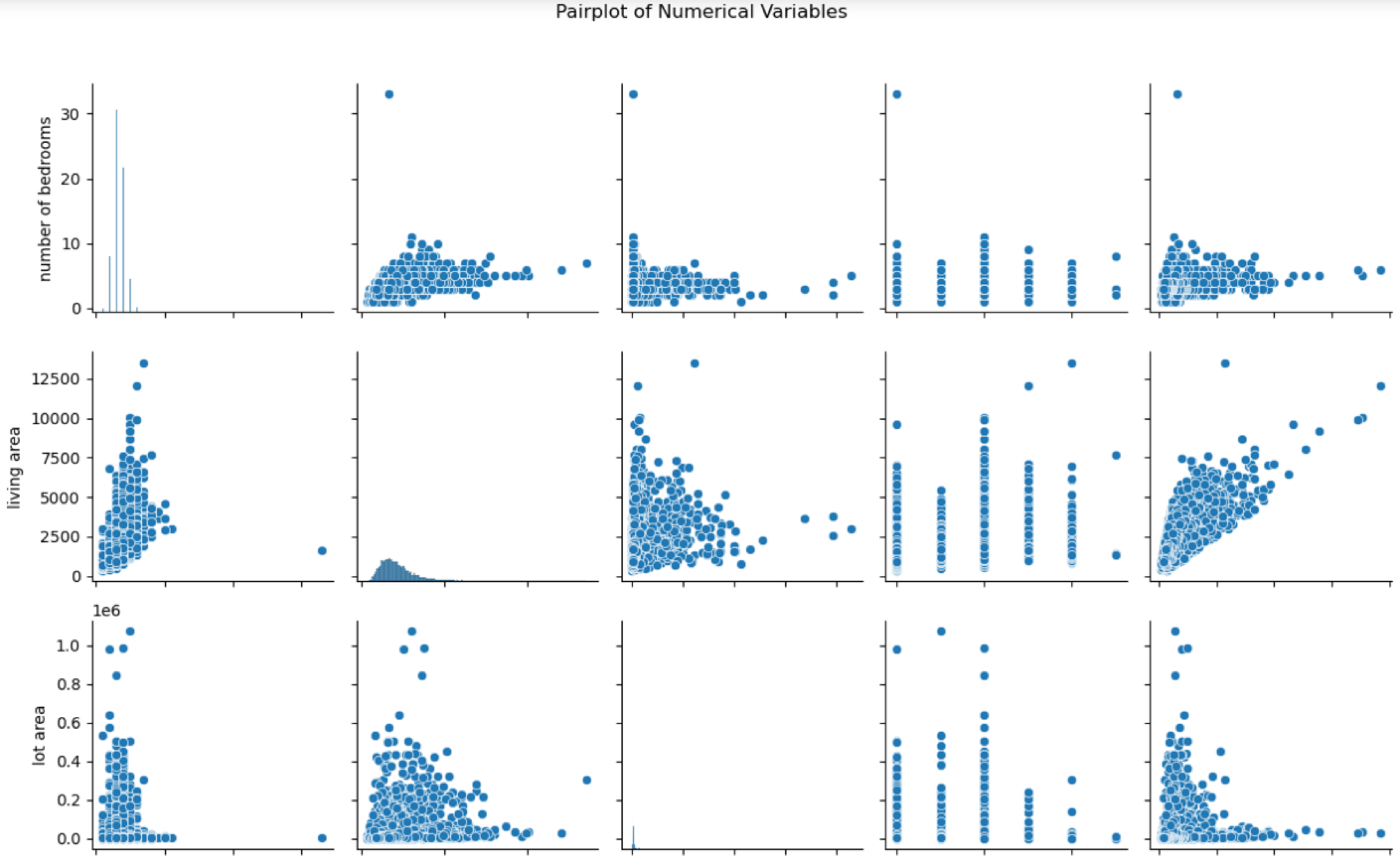

Number Of Bedrooms vs Price
1e6
Price
1.75 1.50 1.25 1.00 0.75 0.50 0.25 0.00
1 2 3 4 5 6 7 8 9 10 11 33
number of bedrooms

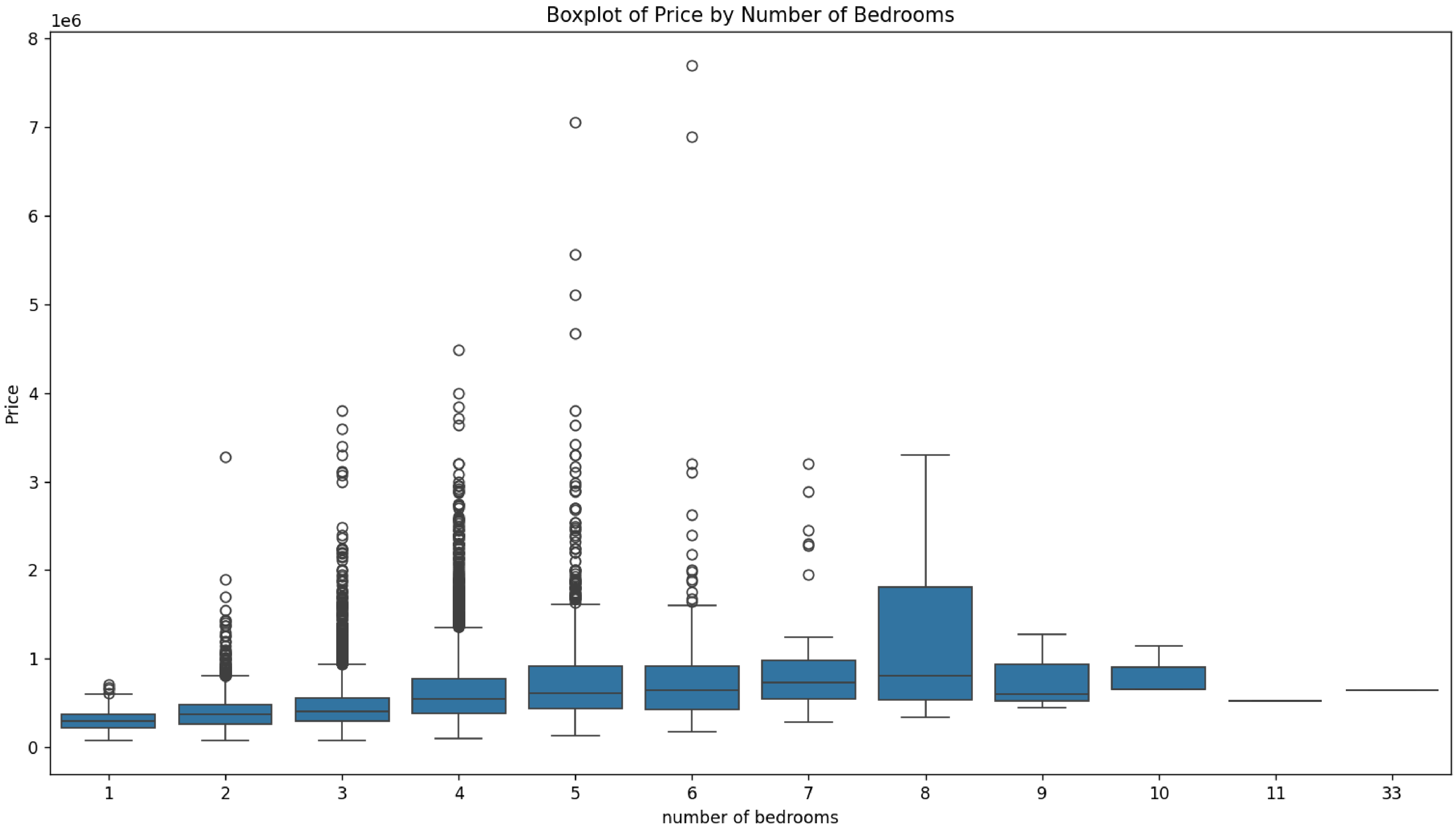


id
500
0
10000
20000
30000
+6.7628e9
Date
1000
0
42500
42600
42700
number of bedrooms
10000
0
10
20
30
number of bathrooms
2000
0
2.5
5.0
7.5
living area
2500
0
5000
10000
lot area
10000
0
0.0
0.5
1.0
1e6
number of floors
5000
0
1
2
3
waterfront present
10000
0
0.0
0.5
1.0
number of views
10000
0
2
4
condition of the house
5000
0
2
4
grade of the house
5000
0
5.0
7.5
10.0
12.5
Area of the house(excluding basement)
2500
0
2500
5000
7500
Area of the basement
5000
0
2000
4000
Built Year
1000
0
1900
1950
2000
Renovation Year
10000
0
1000
2000
Postal Code
1000
0
20
40
60
+1.22e5
Lattitude
1000
0
52.4
52.6
52.8
53.0
Longitude
2000
0
−114.5
−114.0
−113.5
living_area_renov
2000
0
2000
4000
6000
lot_area_renov
10000
0
200000
400000
Number of schools nearby
5000
0
1
2
3
Distance from the airport
1000
0
50
60
70
80
Price
5000
0
0.0
2.5
5.0
7.5



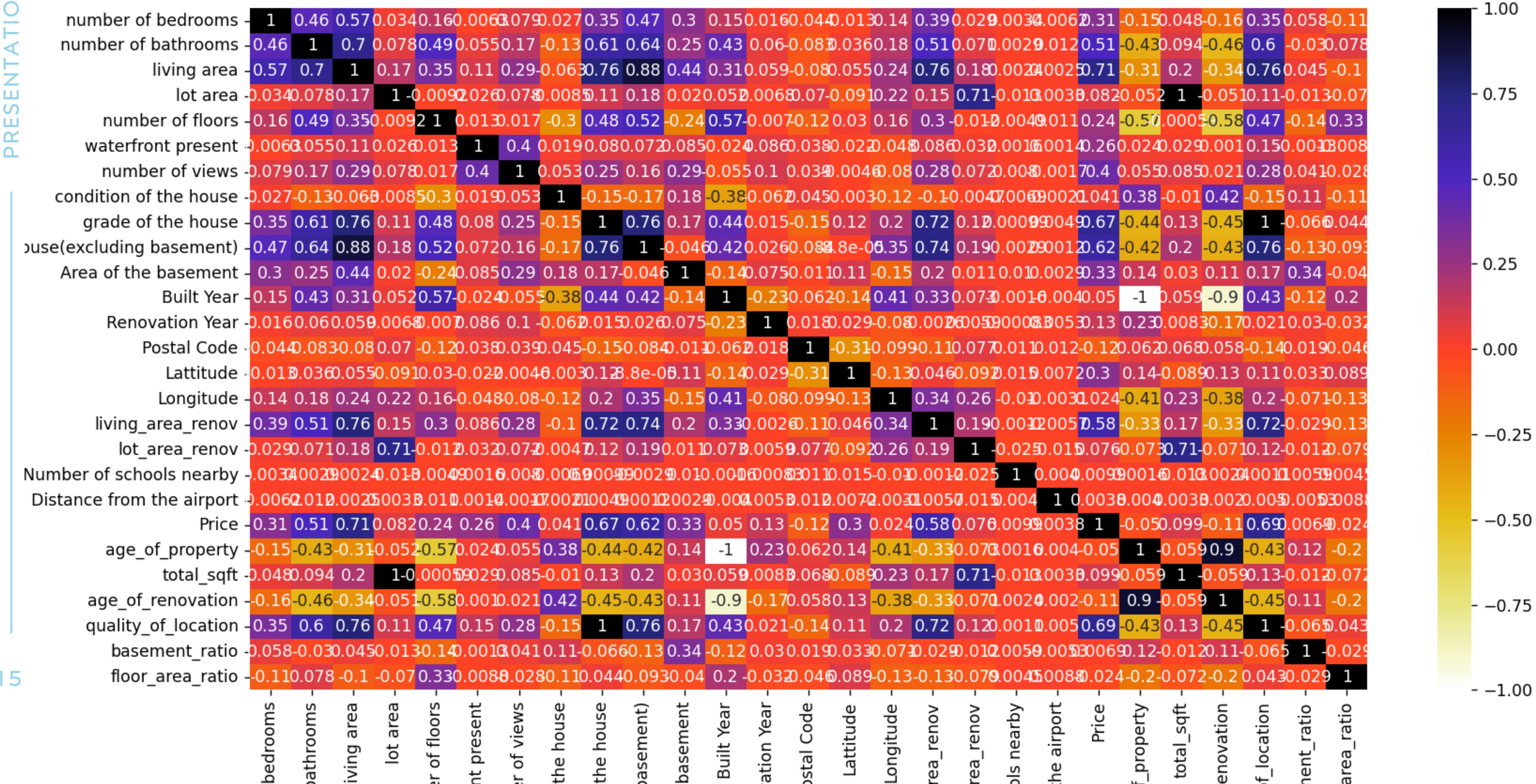
id 1.00 0.05 -0.33 -0.52 -0.65 -0.10 -0.31 -0.11 -0.29 -0.05 -0.67 -0.57 -0.29 -0.07 -0.11 0.29 -0.48 -0.07 -0.60 -0.09 -0.00 -0.00 -0.77
Date 0.05 1.00 -0.02 -0.03 -0.02 0.00 -0.01 0.01 -0.00 -0.03 -0.03 -0.02 -0.02 -0.01 -0.01 0.02 -0.02 -0.02 -0.03 -0.00 -0.00 0.01 -0.03
number of bedrooms -0.33 -0.02 1.00 0.51 0.57 0.03 0.18 -0.01 0.08 0.03 0.35 0.47 0.30 0.15 0.02 -0.04 -0.01 0.14 0.39 0.03 0.00 -0.01 0.31
number of bathrooms -0.52 -0.03 0.51 1.00 0.75 0.08 0.50 0.06 0.18 -0.13 0.66 0.68 0.29 0.50 0.05 -0.11 0.03 0.22 0.57 0.08 0.00 0.01 0.53
living area -0.65 -0.02 0.57 0.75 1.00 0.17 0.35 0.11 0.29 -0.06 0.76 0.88 0.44 0.31 0.06 -0.08 0.05 0.24 0.76 0.18 0.00 0.00 0.71
lot area -0.10 0.00 0.03 0.08 0.17 1.00 -0.00 0.03 0.08 -0.01 0.11 0.18 0.02 0.05 0.01 0.07 -0.09 0.22 0.15 0.71 -0.01 0.00 0.08
number of floors -0.31 -0.01 0.18 0.50 0.35 -0.00 1.00 0.02 0.02 -0.27 0.46 0.53 -0.24 0.48 0.01 -0.13 0.05 0.13 0.29 -0.01 -0.01 0.02 0.26
waterfront present -0.11 0.01 -0.01 0.06 0.11 0.03 0.02 1.00 0.40 0.02 0.08 0.07 0.09 -0.02 0.09 0.04 -0.02 -0.05 0.09 0.03 0.00 0.00 0.26
number of views -0.29 -0.00 0.08 0.18 0.29 0.08 0.02 0.40 1.00 0.05 0.25 0.16 0.29 -0.06 0.10 0.04 -0.00 -0.08 0.28 0.07 0.01 -0.00 0.40
condition of the house -0.05 -0.03 0.03 -0.13 -0.06 -0.01 -0.27 0.02 0.05 1.00 -0.15 -0.17 0.18 -0.38 -0.06 0.05 -0.00 -0.12 -0.10 -0.00 -0.01 -0.00 0.04
grade of the house -0.67 -0.03 0.35 0.66 0.76 0.11 0.46 0.08 0.25 -0.15 1.00 0.76 0.17 0.44 0.01 -0.15 0.12 0.20 0.72 0.12 0.00 0.00 0.67
Area of the house(excluding basement) -0.57 -0.02 0.47 0.68 0.88 0.18 0.53 0.07 0.16 -0.17 0.76 1.00 -0.05 0.42 0.03 -0.08 -0.00 0.35 0.74 0.19 -0.00 0.00 0.62
Area of the basement -0.29 -0.02 0.30 0.29 0.44 0.02 -0.24 0.09 0.29 0.18 0.17 -0.05 1.00 -0.14 0.08 -0.01 0.11 -0.15 0.20 0.01 0.01 0.00 0.33
Built Year -0.07 -0.01 0.15 0.50 0.31 0.05 0.48 -0.02 -0.06 -0.38 0.44 0.42 -0.14 1.00 -0.23 -0.06 -0.14 0.41 0.33 0.07 -0.00 -0.00 0.05
Renovation Year -0.11 -0.01 0.02 0.05 0.06 0.01 0.01 0.09 0.10 -0.06 0.01 0.03 0.08 -0.23 1.00 0.02 0.03 -0.08 -0.00 0.01 -0.00 0.01 0.13
Postal Code 0.29 0.02 -0.04 -0.11 -0.08 0.07 -0.13 0.04 0.04 0.05 -0.15 -0.08 -0.01 -0.06 0.02 1.00 -0.31 -0.10 -0.11 0.08 0.01 0.01 -0.12
Lattitude -0.48 -0.02 -0.01 0.03 0.05 -0.09 0.05 -0.02 -0.00 -0.00 0.12 -0.00 0.11 -0.14 0.03 -0.31 1.00 -0.13 0.05 -0.09 0.01 0.01 0.30
Longitude -0.07 -0.02 0.14 0.22 0.24 0.22 0.13 -0.05 -0.08 -0.12 0.20 0.35 -0.15 0.41 -0.08 -0.10 -0.13 1.00 0.34 0.26 -0.01 -0.00 0.02
living_area_renov -0.60 -0.03 0.39 0.57 0.76 0.15 0.29 0.09 0.28 -0.10 0.72 0.74 0.20 0.33 -0.00 -0.11 0.05 0.34 1.00 0.19 -0.00 -0.01 0.58
lot_area_renov -0.09 -0.00 0.03 0.08 0.18 0.71 -0.01 0.03 0.07 -0.00 0.12 0.19 0.01 0.07 0.01 0.08 -0.09 0.26 0.19 1.00 -0.03 -0.01 0.08
Number of schools nearby -0.00 -0.00 0.00 0.00 0.00 -0.01 -0.01 0.00 0.01 -0.01 0.00 -0.00 0.01 -0.00 -0.00 0.01 0.01 -0.01 -0.00 -0.03 1.00 0.00 0.01
Distance from the airport -0.00 0.01 -0.01 0.01 0.00 0.00 0.02 0.00 -0.00 -0.00 0.00 0.00 0.00 -0.00 0.01 0.01 0.01 -0.00 -0.01 -0.01 0.00 1.00 0.00
Price -0.77 -0.03 0.31 0.53 0.71 0.08 0.26 0.26 0.40 0.04 0.67 0.62 0.33 0.05 0.13 -0.12 0.30 0.02 0.58 0.08 0.01 0.00 1.00
id
Date
number of bedrooms
number of bathrooms
living area
lot area
number of floors
waterfront present
number of views
condition of the house
grade of the house
rea of the house(excluding basement)
Area of the basement
Built Year
Renovation Year
Postal Code
Lattitude
Longitude
living_area_renov
lot_area_renov
Number of schools nearby
Distance from the airport
Price
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6


Pairplot of Numerical Variables
number of bedrooms
0
10
20
30
living area
0
2500
5000
7500
10000
12500
lot area
1e6
0.0
0.2
0.4
0.6
0.8
1.0


Boxplot of Price by Number of Bedrooms
1e6
8
7
6
5
4
3
2
1
0
Price
1
2
3
4
5
6
7
8
9
10
11
33
number of bedrooms

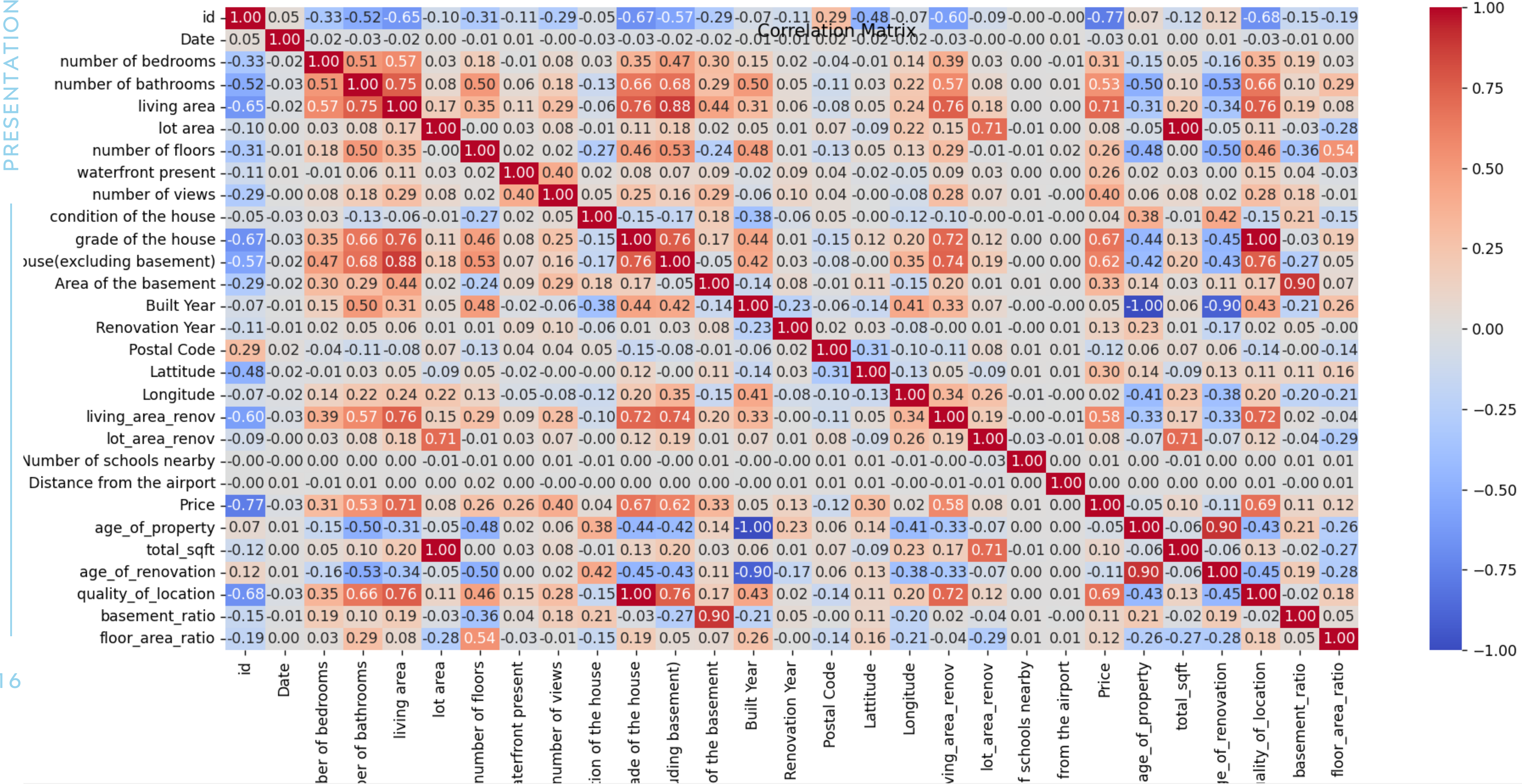
Step 3: Feature Engineering:

Created new features like property age and total square footage



Step 4: Feature Selection:

Identified and removed highly correlated features



Step 5: Modeling:

- Split data for training and testing.

```
# Split the data into train and test sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
```

- Trained RandomForest, XGBoost, and CatBoost models.

```
random_forest_model = RandomForestRegressor()
xgboost_model = XGBRegressor()

# Train individual base models
random_forest_model.fit(X_train_scaled, y_train)
xgboost_model.fit(X_train_scaled, y_train)

# Make predictions using individual base models
y_pred_rf = random_forest_model.predict(X_test_scaled)
y_pred_xgb = xgboost_model.predict(X_test_scaled)

y_pred_rf
y_pred_xgb
```

- Evaluated models using R-squared scores.

```
Random Forest Model:
R-squared: 0.6591513191729483

XGBoost Model:
R-squared: 0.6487248781534409
R-squared: 0.6720537148567195
catboost:
R-Squared:  0.6681556299940523
```

- Combined predictions to create an ensemble model.

```
Ensemble Model:
R-squared: 0.6772239409590448
Precision: 0.7687861271676301
Recall: 0.7287671232876712
```

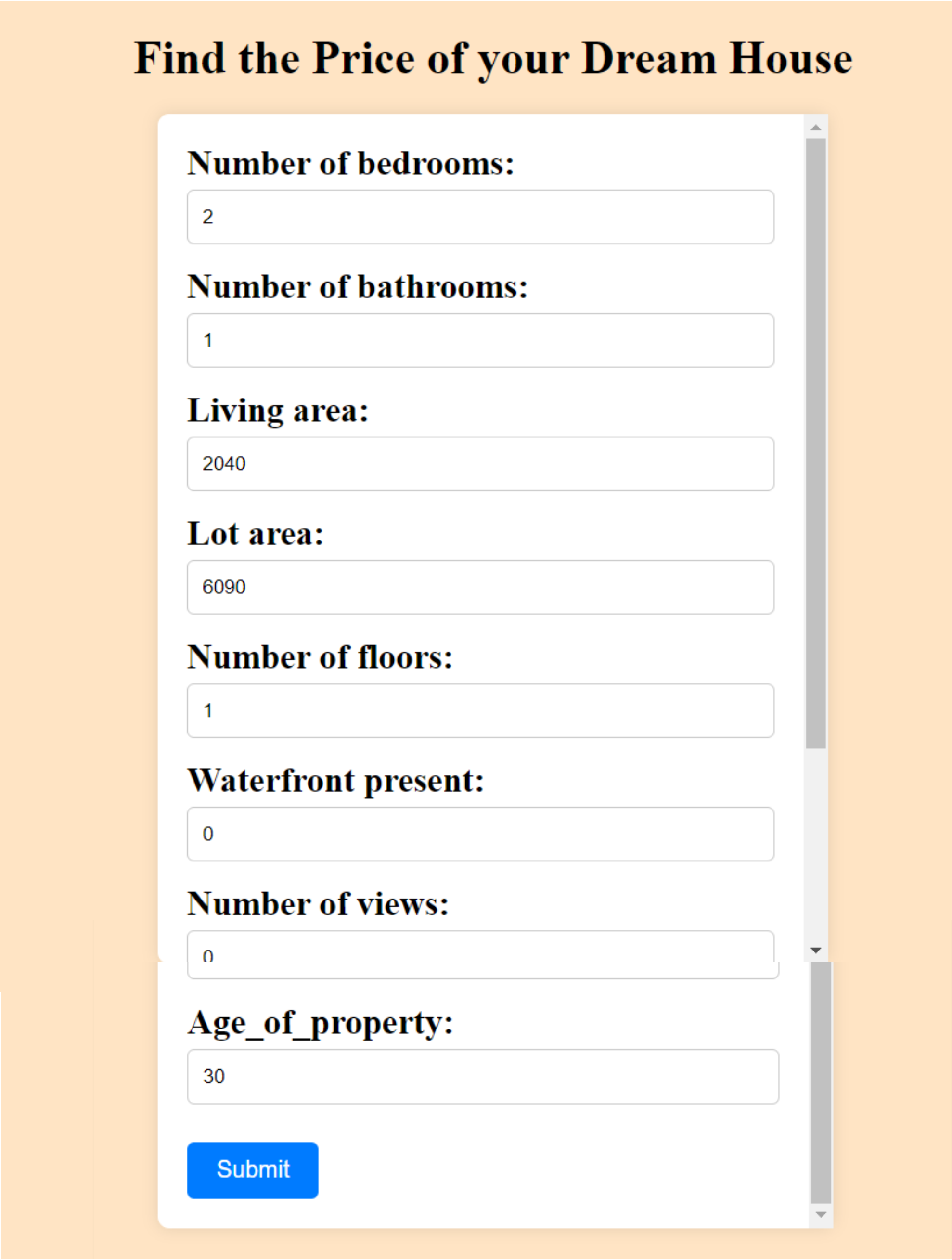
Step 6: Model Selection:

Selected the best-performing model based on R-squared scores.

```
Best Model: CatBoost
R-squared: 0.6681556299940523
```

Step 7: Frontend Integeration:

- Saved the selected model for future use.
- Integrated the Machine learning model with the website using flask and saving the executed model file as pickle file.
- Pickel file is used for reusing the compiled model again and again without compiling everytime.
- It increases the speed of website.


Find the Price of your Dream House
Number of bedrooms:
2
Number of bathrooms:
1
Living area:
2040
Lot area:
6090
Number of floors:
1
Waterfront present:
0
Number of views:
0
Age_of_property:
30
Submit


Predicted House price is: ₹587796

SUMMARY

The project involved analyzing housing data to predict prices in India. After data preprocessing and exploratory data analysis, redundant features were removed, leaving 8 informative features. Three machine learning models (Random Forest, XGBoost, CatBoost) were trained individually and combined using averaging to create an ensemble model. The best model, a combination of the three, achieved high accuracy. A Flask backend was implemented to serve predictions based on user inputs from a frontend interface.

Reference

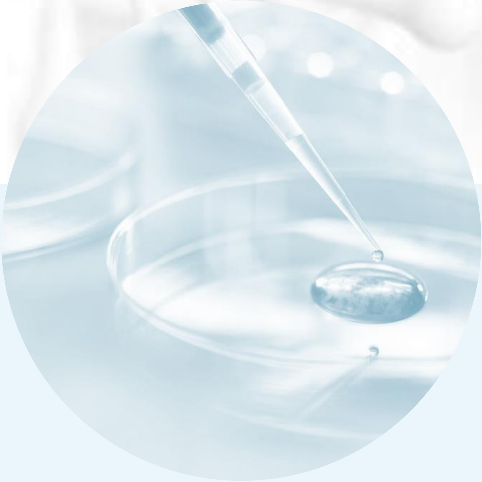
https://medium.com/shecodeafrica/predicting-house-prices-gradientboostingregressor-algorithm-ec9d381b0ebc

https://www.mdpi.com/2813-2203/3/1/3

https://gustiyaniz.medium.com/house-price-prediction-with-xgboost-using-r-markdown-d4f891d1f327

https://www.ijraset.com/research-paper/prediction-of-house-prices-using-machine-learning

https://medium.com/hackerdawn/house-prices-prediction-using-random-forest-aa8722347276



THANK YOU