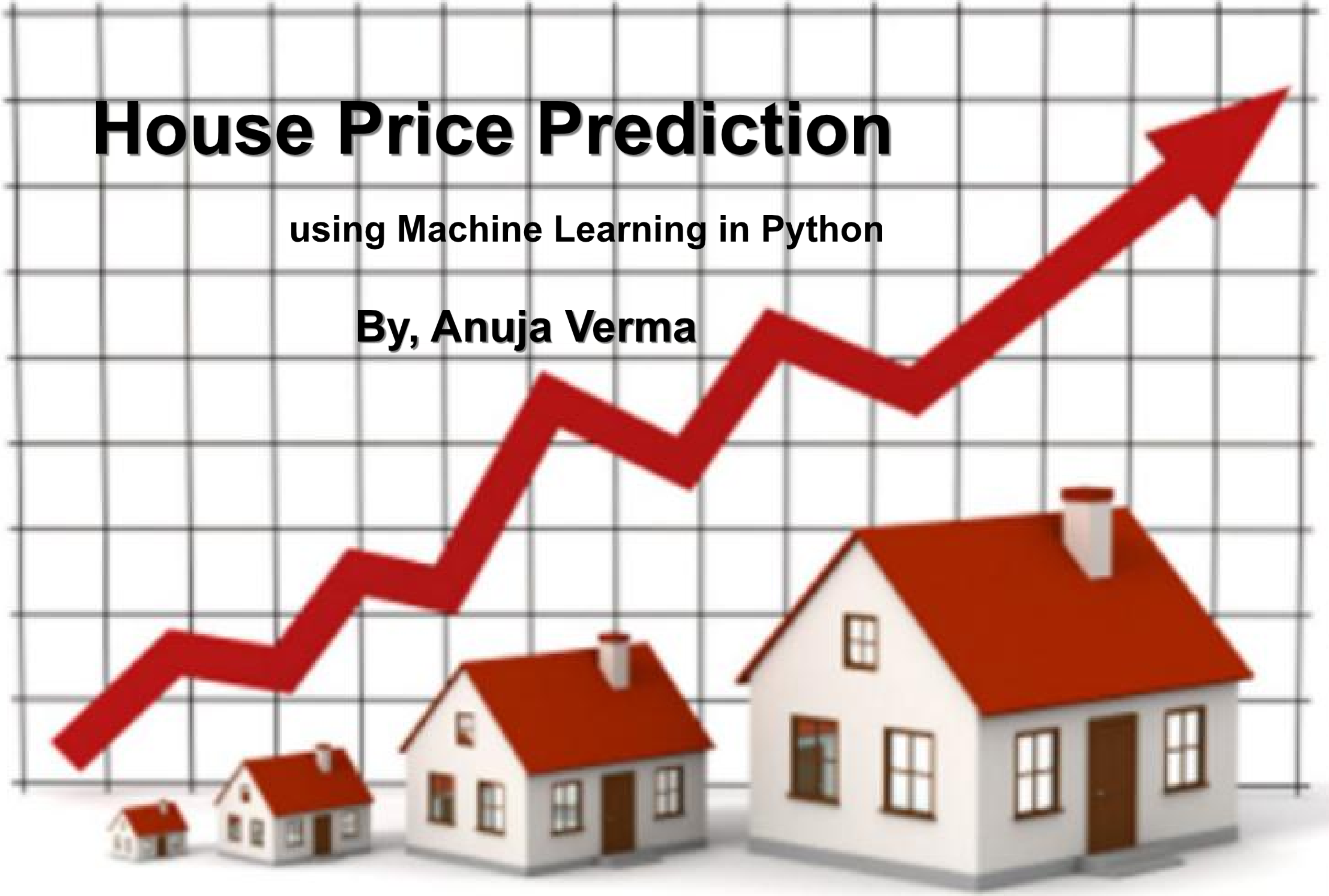
House Price Prediction

using Machine Learning in Python

By, Anuja Verma



Abstract

House Price Index (HPI) is commonly used to estimate the changes in housing price. Since housing price is strongly correlated to other factors such as location, area, population, it requires other information apart from HPI to predict individual housing price. There has been a considerably large number of papers adopting traditional machine learning approaches to predict housing prices accurately, but they rarely concern about the performance of individual models and neglect the less popular yet complex models. As a result, to explore various impacts of features on prediction methods, this paper will apply both traditional and advanced machine learning approaches to investigate the difference among several advanced models. This paper will also comprehensively validate multiple techniques in model implementation on regression and provide an optimistic result for housing price prediction.

Aim

The aim of house price prediction is to estimate the potential selling price of a property based on various factors. This is crucial for buyers, sellers, and investors in making informed decisions about the real estate market. House price prediction helps buyers find affordable homes, sellers set competitive prices, and investors identify profitable opportunities.

- These are the Parameters on which we will evaluate ourselves -

- Create an effective price prediction model
- Validate the model's prediction accuracy
- Identify the important home price attributes which feed the model's predictive power.

Implementation

Step Involved :

- Importing the required packages in our python environment.
- Importing data - Set

```
[8]: import pandas as pd
     import numpy as np
     import seaborn as sns
     import matplotlib.pyplot as plt
```

```
[9]: ### 1.Loading the Data ###
```

```
[10]: Ds1 = pd.read_csv(r"C:\Users\Anuja Verma\Downloads\housing_data.csv")
```

Data Selection

Data selection is defined as the process of determining the appropriate data type and source, as well as suitable instruments to collect data.Data selection precedes the actual practice of data collection.

```
### 1.Loading the Data ###
```

```
Ds1 = pd.read_csv(r"C:\Users\Anuja Verma\Downloads\housing_data.csv")
```

```
Ds1.head()
```

ıtage	LotArea	Street	Alley	LotShape	LandContour	Utilities	...	PoolArea	PoolQC	Fence	MiscFeature	MiscVal	MoSold	YrSold	SaleType	SaleCondition	SalePrice
65	8450	Pave	NaN	Reg	Lvl	AllPub	...	0	No	No	No	0	Feb	2008	WD	Normal	208500
80	9600	Pave	NaN	Reg	Lvl	AllPub	...	0	No	No	No	0	May	2007	WD	Normal	181500
68	11250	Pave	NaN	IR1	Lvl	AllPub	...	0	No	No	No	0	Sep	2008	WD	Normal	223500
60	9550	Pave	NaN	IR1	Lvl	AllPub	...	0	No	No	No	0	Feb	2006	WD	Abnorml	140000
84	14260	Pave	NaN	IR1	Lvl	AllPub	...	0	No	No	No	0	Dec	2008	WD	Normal	250000

Data visualization

Data visualization is the graphical representation of information and data. By using visual elements like charts,graphs, and maps , data visualization tools provide an accessible way to see and understand trends, outliers, and patterns in data. In the world of big Data, data visualization tools and technologies are essential to analyse massive amounts of information and make data-driven decisions.

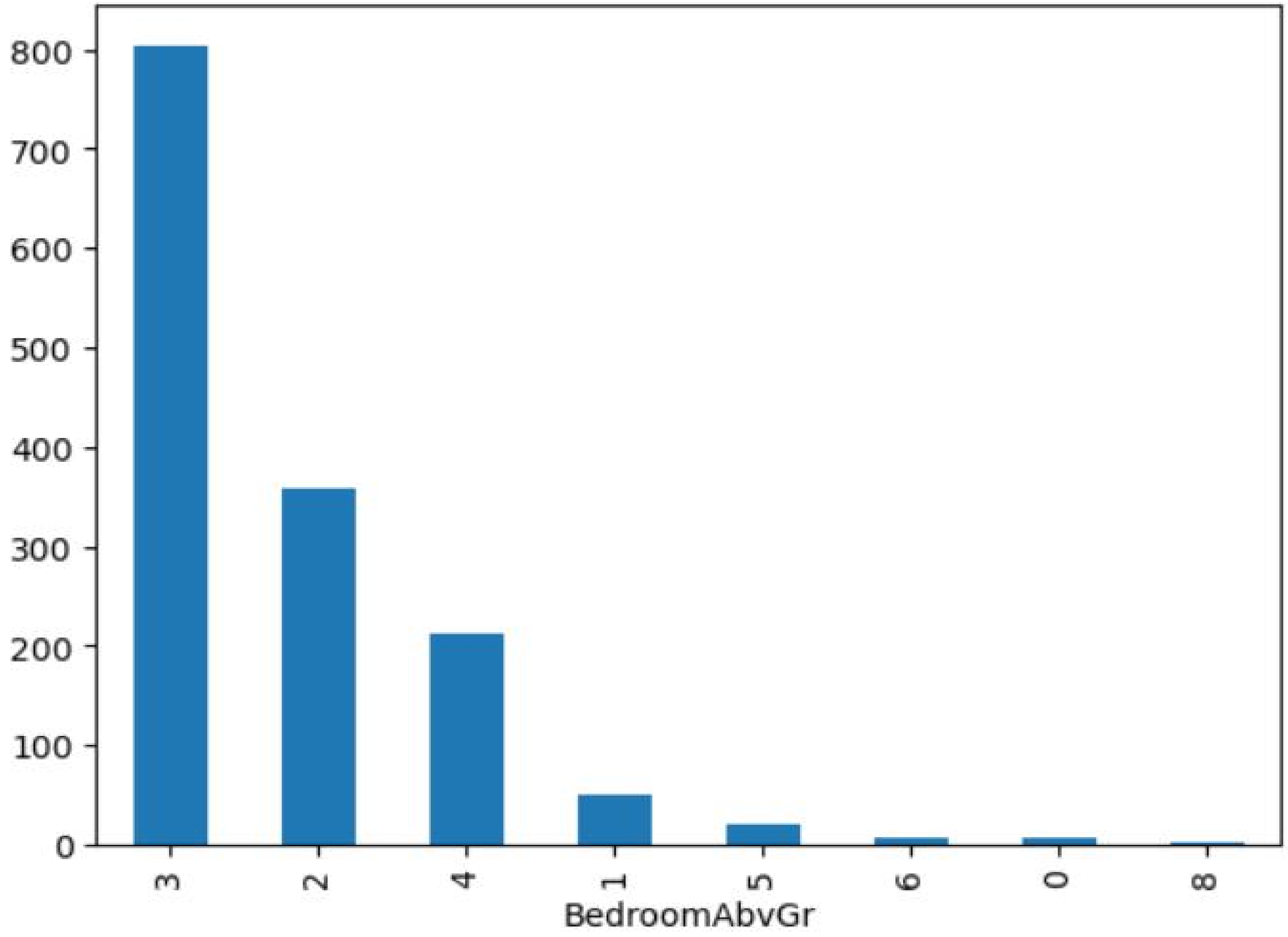
Exploratory Data Analysis

Refers to the deep analysis of data so as to discover different patterns and spot anomalies. Before making inferences from data it is essential to examine all your variables. we can infer from above describe function that the dataset has a house where the house and would be interesting to know more about it as we progress.

Maximum square feet is 16200 where as the minimum is 1650.we can see that the data is distributed.

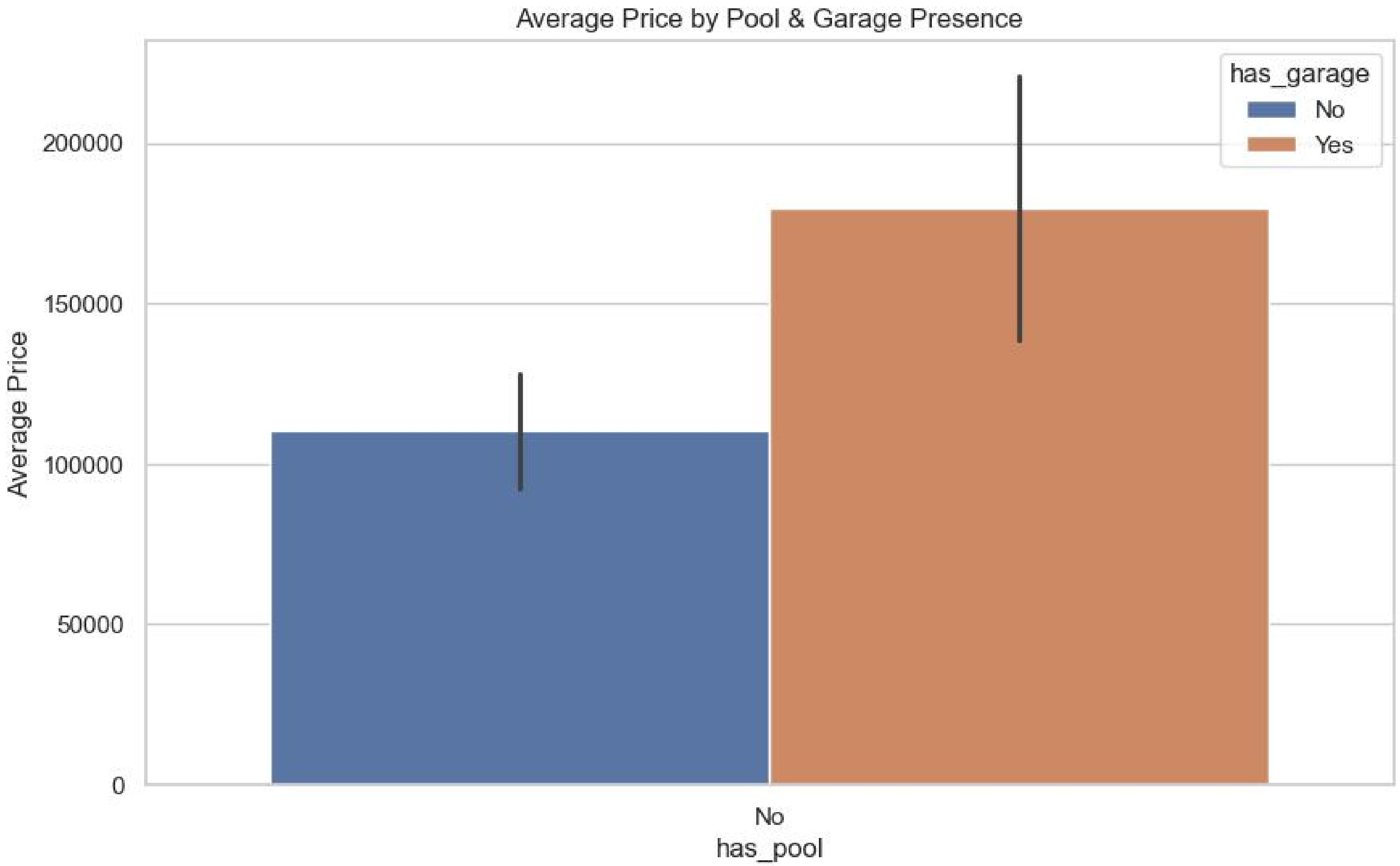
```
Ds1['BedroomAbvGr'].value_counts().plot( kind = 'bar')
```

```
<Axes: xlabel='BedroomAbvGr'>
```



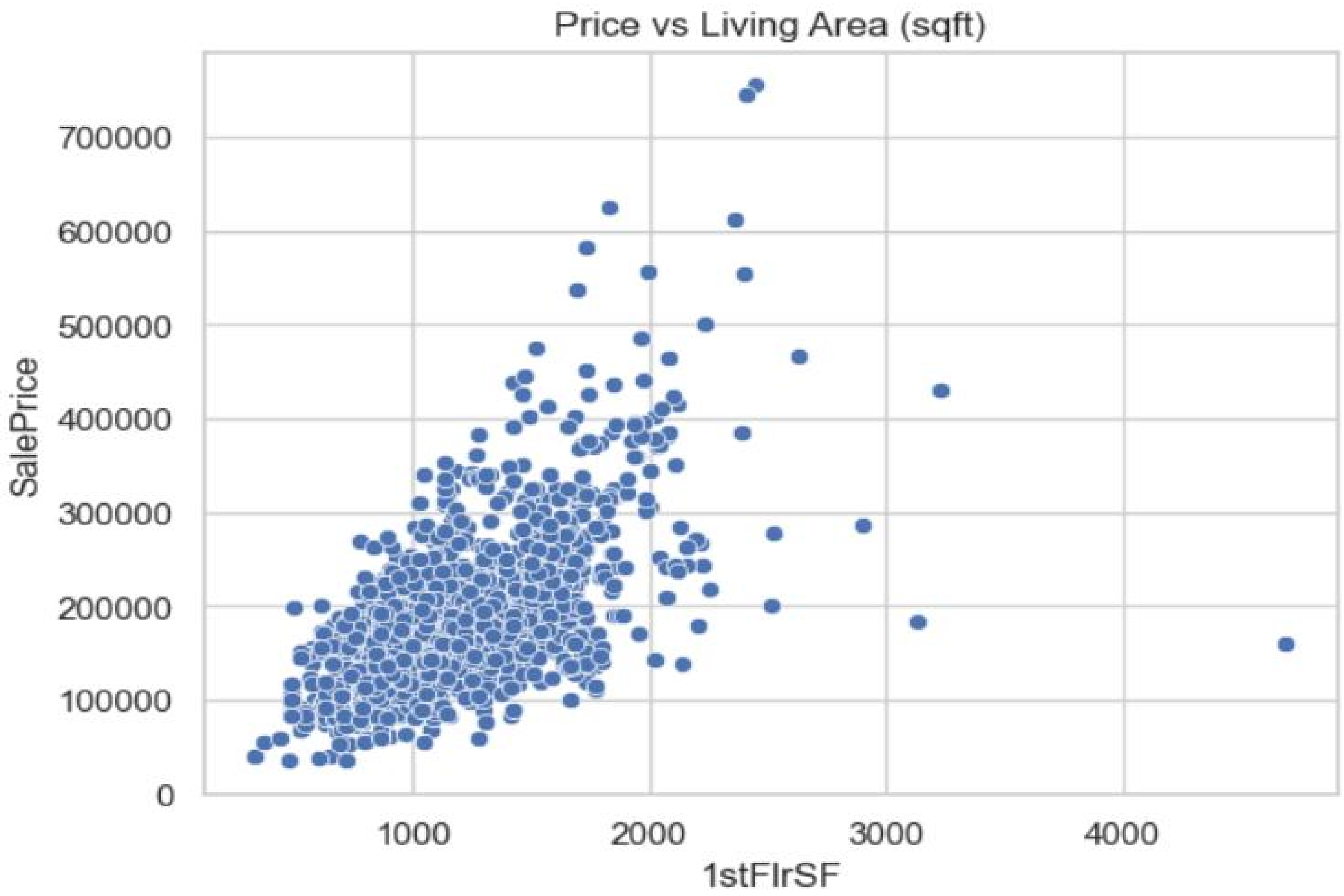
As we can see from the visualization 3 bedroom house are most commonly sold followed by 2 bedroom. So how is it useful? For a builder having this data , He can make a new building with more 2 and 3 bedroom's to attract more buyers.

```python
plt.figure(figsize=(10,6))
sns.barplot(data=avg_prices, x='has_pool', y='SalePrice', hue='has_garage')
plt.title('Average Price by Pool & Garage Presence')
plt.ylabel('Average Price')
plt.show()
```



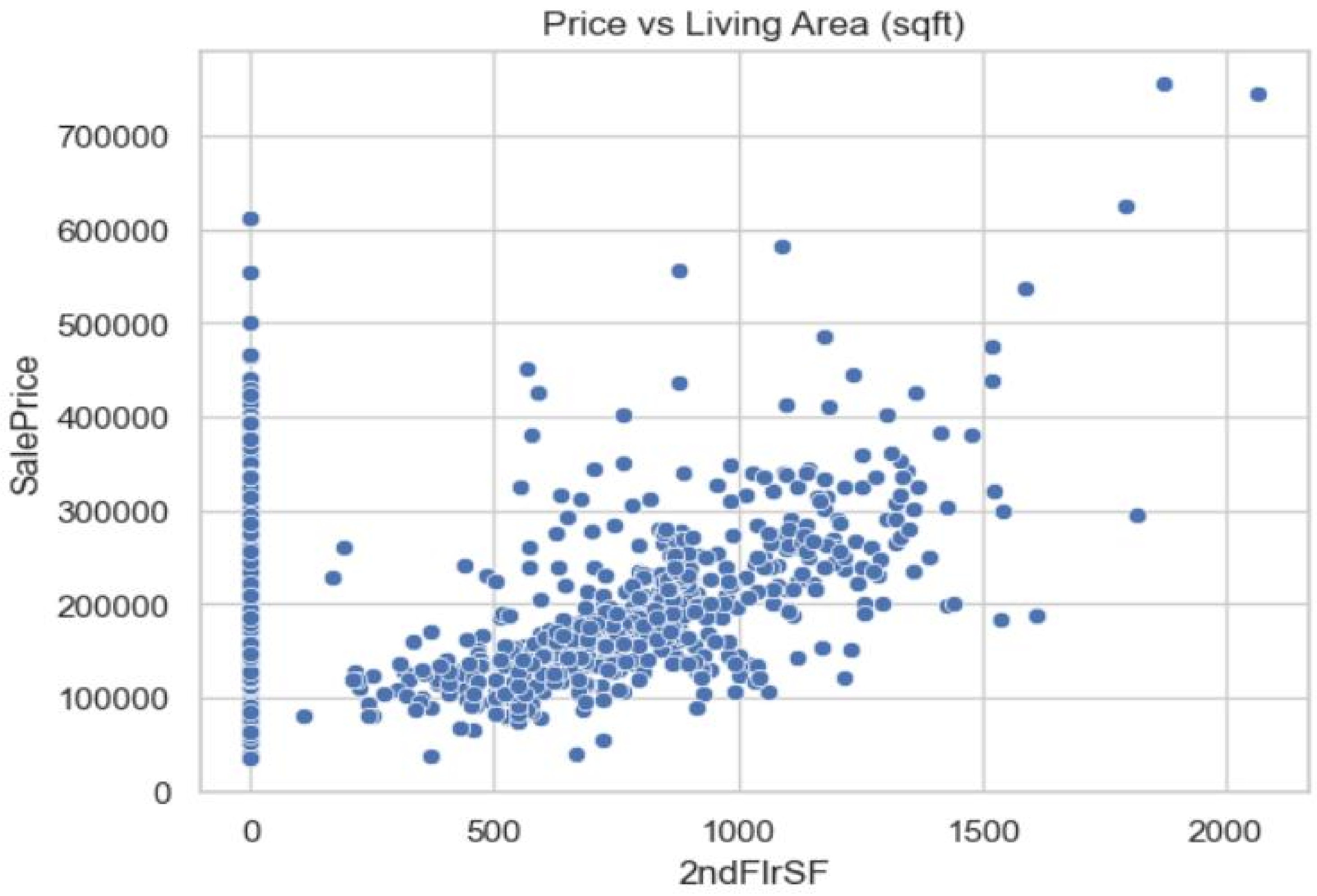
Form above we can see that house which are Pool & garage will get Average price

```
# Scatter plot for sqft_living vs price
sns.scatterplot(x='1stFlrSF' , y='SalePrice', data=Ds1)
plt.title('Price vs Living Area (sqft)')
plt.show()
```



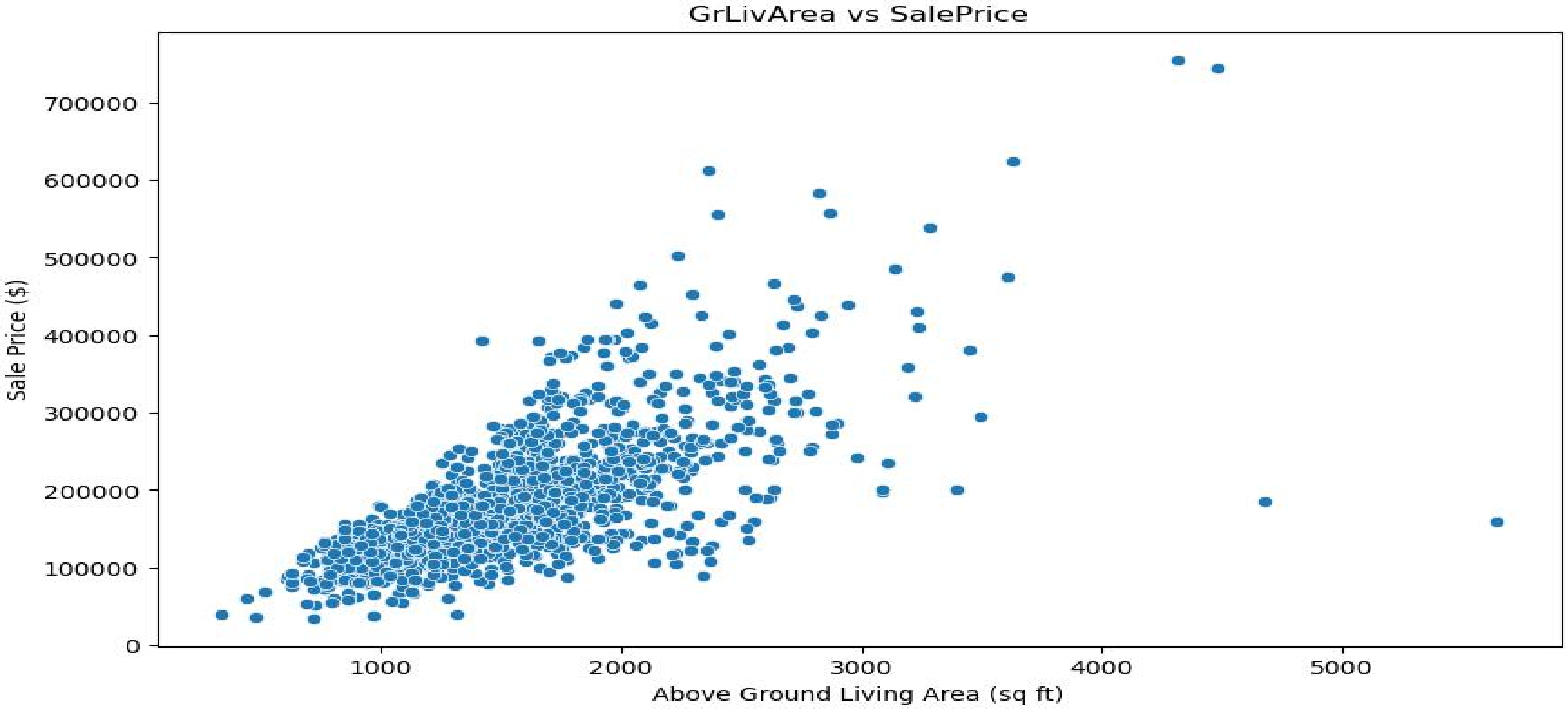
The pricing of the first floor square feet is shown in the chart.

```
# Scatter plot for sqft_living vs price
sns.scatterplot(x='2ndFlrSF' , y='SalePrice', data=Ds1)
plt.title('Price vs Living Area (sqft)')
plt.show()
```

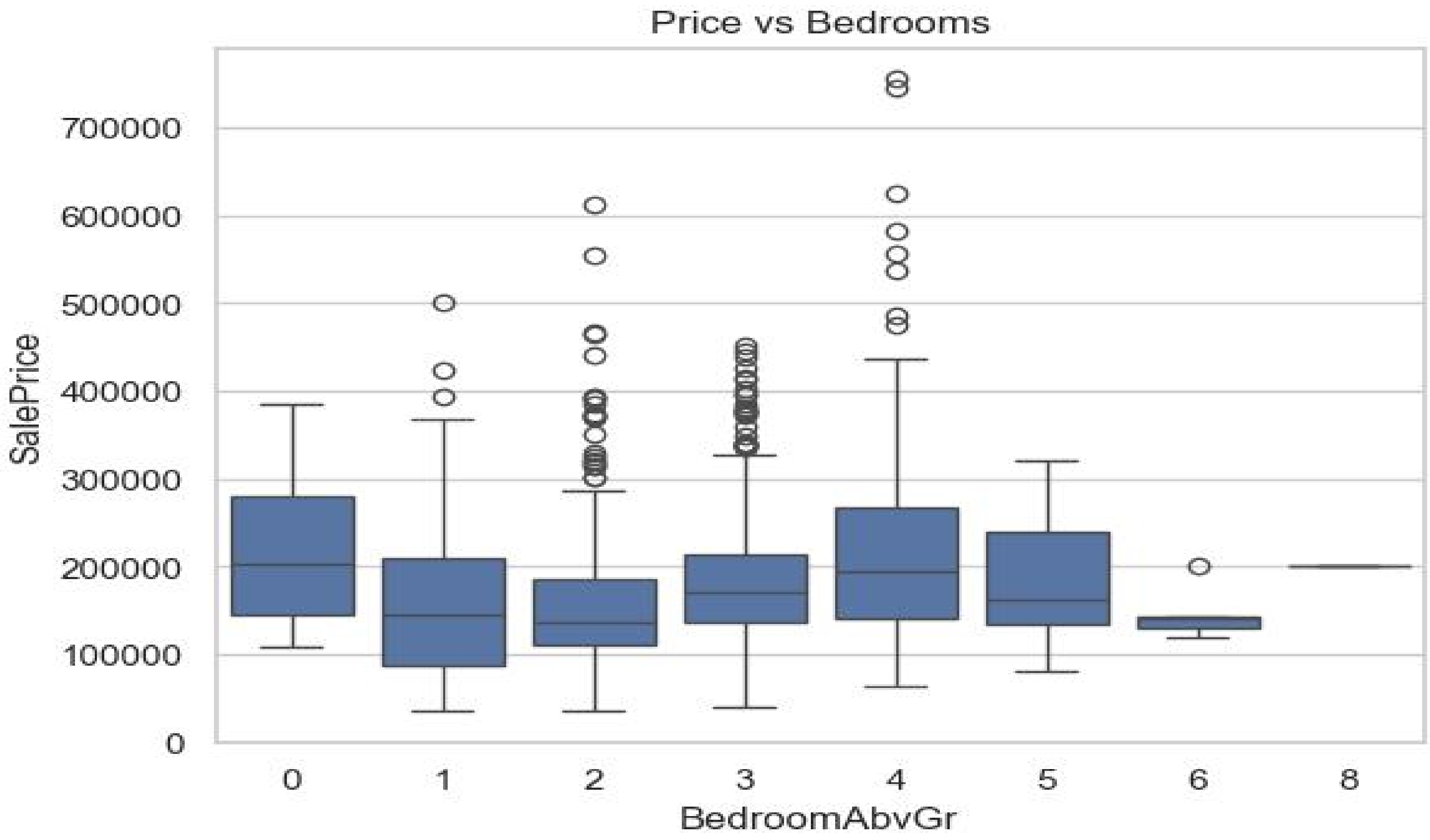


The pricing of the Seceond floor square feet is shown in the chart.

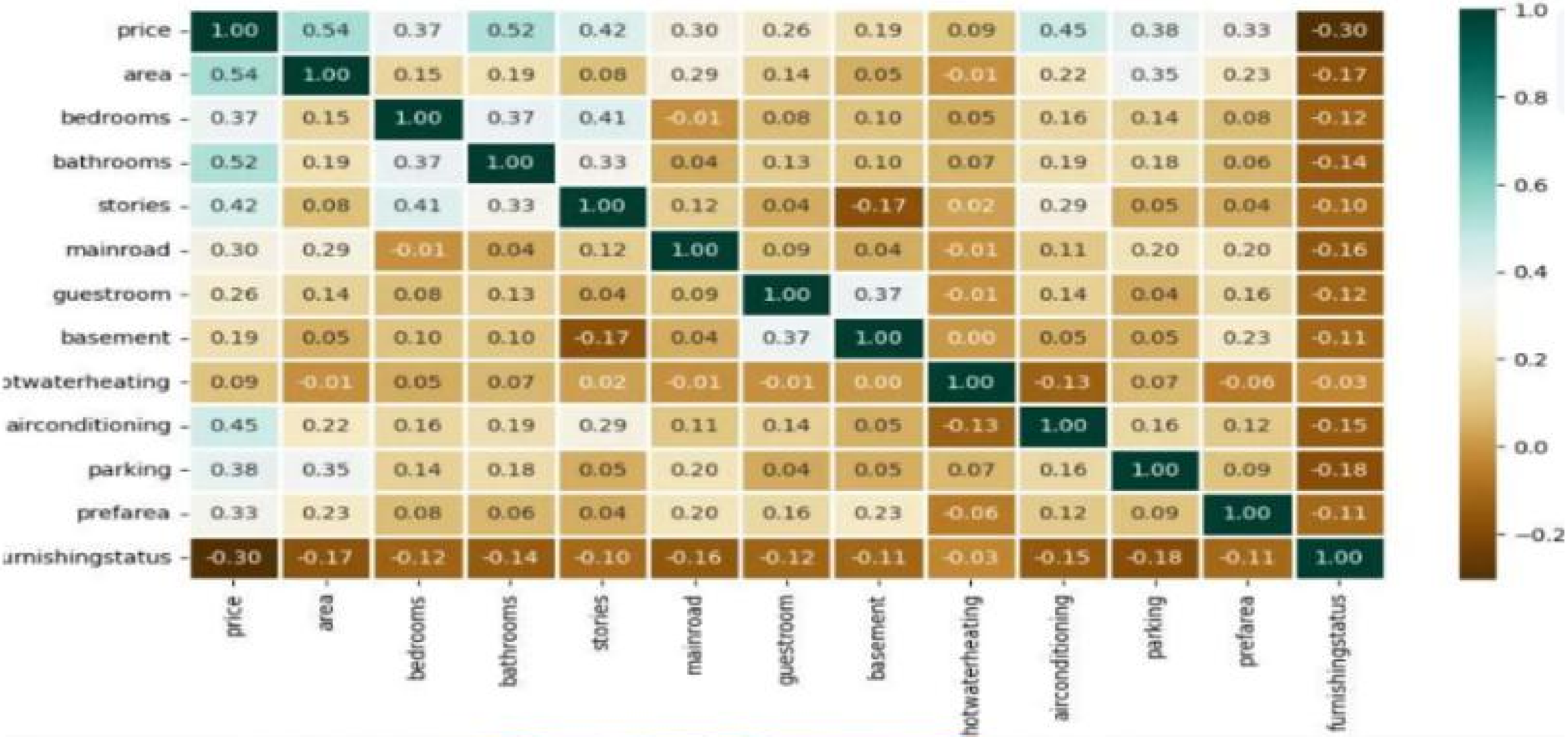
```python
# Scatter plot for GrLivArea vs SalePrice
plt.figure(figsize=(10, 6))
sns.scatterplot(data=Ds1, x='GrLivArea', y='SalePrice')
plt.title('GrLivArea vs SalePrice')
plt.xlabel('Above Ground Living Area (sq ft)')
plt.ylabel('Sale Price ($)')
plt.show()
```



```python
# Boxplot to see price variation with bedrooms
sns.boxplot(x ='BedroomAbvGr', y='SalePrice', data=Ds1)
plt.title('Price vs Bedrooms')
plt.show()
```



Correlation Heatmap



[*]:

```
# Calculate the correlation matrix
correlation_matrix = Ds1.corr()

# Set up the matplotlib fgure
plt.figure(figsize=(12, 10))

# Draw the heatmap
sns.heatmap(correlation_matrix, annot=True, fmt=".2f", cmap='coolwarm', square=True, cbar_kws={"shrink": .8})

# Title and show the plot
plt.title('Correlation Matrix of House Prices Dataset')
plt.show()

sns.heatmap(Ds1.corr(numeric_only=True), annot=True, cmap='coolwarm')
```

Feature Selection

- Feature selection is a process that chooses a subset of feature from the original feature so that the feature space is optimally reduced according to a certain criterion.

```
#Feature Selection

x = Ds1.drop(['SalePrice',],axis = 1)
y = Ds1['SalePrice']

print(x.shape,y.shape)
```

```
(1460, 80) (1460,)
```

Conclusion

- So we conclude that the system that we proposed solves most of the problem that we have with the existing system. After training and testing of datasets with all models, the linear regression performs better than gradient boost regressor. The highest accuracy score is achieved by the linear regression. So , we suggest that this regression model be used for future house price predictions. Therefore. the outcome of our project will be predicting house prices with good accuracy which can help the customer as well as developer.

Thank You