Contents

House Prices 1
Abstract: 1
Key words: 1
Introduction: 1
Literature Review: 2
Methology: 2
Discussion of Findings: 2
Conclusion: 2
References / Bibliography: 2
Appendices: 3

House Prices

CUNY SPS DATA 621

GROUP 2: William Aiken, Donald Butler, Michael Ippolito, Bharani Nittala, and Leticia Salazar

Abstract:

Use 250 words or less to summarize your problem, methodology, and major outcomes.

[insert text here]

Key words:

Select a few key words (up to five) related to your work.

[insert text here]

Introduction:

Describe the background and motivation of your problem.

[insert text here]

Literature Review:

Discuss how other researchers have addressed similar problems, what their achievements are, and what the advantage and drawbacks of each reviewed approach are. Explain how your investigation is similar or different to the state-of-the- art. Please cite the relevant papers where appropriate.

[insert text here]

Methology:

Discuss the key aspects of your problem, data set and regression model(s). Given that you are working on real-world data, explain at a high-level your exploratory data analysis, how you prepared the data for regression modeling, your process for building regression models, and your model selection.

[insert text here]

Discussion of Findings:

Describe the specifics of what you did (data exploration, data preparation, model building, model selection, model evaluation, etc.), and what you found out (statistical analyses, interpretation and discussion of the results, etc.). Conclude your findings, limitations, and suggest areas for future work.

[insert text here]

Recommendations: [insert text here]

Limitations: [insert text here]

Conclusion:

[insert text here]

References / Bibliography:

Be sure to cite all references used in the report (APA format).

- H, M. Y. (2022, January 12). Housing prices dataset. Kaggle. Retrieved November 28, 2022, from https://www.kaggle.com/datasets/yasserh/housing-prices-dataset
- [insert text here]
- [insert text here]

Appendices:

- Supplemental tables and/or figures.
- R statistical programming code.

[insert graphs, tables, etc here]

START OF OUR CODING HERE! WILL NOT SHOW FOR PDF VERSION

Load Libraries:

These are the libraries used to explore, prepare, analyze and build our models

```
library(tidyverse)
library(dplyr)
library(corrplot)
```

Load Data:

We have included the original data sets in our GitHub account and read from this location. Since our data set doesn't come with a training and evaluation data sets we will be splitting our data using the 70% - 30% split. Below we are showing the training data set:

```
##      price  area bedrooms bathrooms stories mainroad guestroom basement
## 1 13300000  7420        4         2       3      yes        no       no
## 2 12250000  8960        4         4       4      yes        no       no
## 3 12250000  9960        3         2       2      yes        no      yes
## 5 11410000  7420        4         1       2      yes       yes      yes
## 8 10150000 16200        5         3       2      yes        no       no
## 9  9870000  8100        4         1       2      yes       yes      yes
##   hotwaterheating airconditioning parking prefarea furnishingstatus
## 1              no             yes       2      yes        furnished
## 2              no             yes       3       no        furnished
## 3              no              no       2      yes   semi-furnished
## 5              no             yes       2       no        furnished
## 8              no              no       0       no      unfurnished
## 9              no             yes       2      yes        furnished
```

Data Exploration:

Based on this our training data includes 386 records and 13 variables whereas the evaluation data includes 159 records and 13 variables.

Training:

```
## 'data.frame':    386 obs. of  13 variables:
##  $ price           : int  13300000 12250000 12250000 11410000 10150000 9870000 9800000 9800000 96810
##  $ area            : int  7420 8960 9960 7420 16200 8100 5750 13200 6000 6550 ...
##  $ bedrooms        : int  4 4 3 4 5 4 3 3 4 4 ...
##  $ bathrooms       : int  2 4 2 1 3 1 2 1 3 2 ...
##  $ stories         : int  3 4 2 2 2 2 4 2 2 2 ...
##  $ mainroad        : chr  "yes" "yes" "yes" "yes" ...
##  $ guestroom       : chr  "no" "no" "no" "yes" ...
##  $ basement        : chr  "no" "no" "yes" "yes" ...
##  $ hotwaterheating : chr  "no" "no" "no" "no" ...
##  $ airconditioning : chr  "yes" "yes" "no" "yes" ...
##  $ parking         : int  2 3 2 2 0 2 1 2 2 1 ...
##  $ prefarea        : chr  "yes" "no" "yes" "no" ...
##  $ furnishingstatus: chr  "furnished" "furnished" "semi-furnished" "furnished" ...
```

Evaluation:

```
## 'data.frame':    159 obs. of  13 variables:
##  $ price           : int  12215000 10850000 10150000 9240000 9100000 8960000 8855000 8750000 8400000
##  $ area            : int  7500 7500 8580 7800 6600 8500 6420 4320 7950 6840 ...
##  $ bedrooms        : int  4 3 4 3 4 3 3 3 5 5 ...
##  $ bathrooms       : int  2 3 3 2 2 2 2 1 2 1 ...
##  $ stories         : int  2 1 4 2 2 4 2 2 2 2 ...
##  $ mainroad        : chr  "yes" "yes" "yes" "yes" ...
##  $ guestroom       : chr  "no" "no" "no" "no" ...
##  $ basement        : chr  "yes" "yes" "no" "no" ...
##  $ hotwaterheating : chr  "no" "no" "no" "no" ...
##  $ airconditioning : chr  "yes" "yes" "yes" "no" ...
##  $ parking         : int  3 2 2 0 1 2 1 2 2 1 ...
##  $ prefarea        : chr  "yes" "yes" "yes" "yes" ...
##  $ furnishingstatus: chr  "furnished" "semi-furnished" "semi-furnished" "semi-furnished" ...
```

Using the `summary()` function lets start exploring the training and evaluation data.

Training:

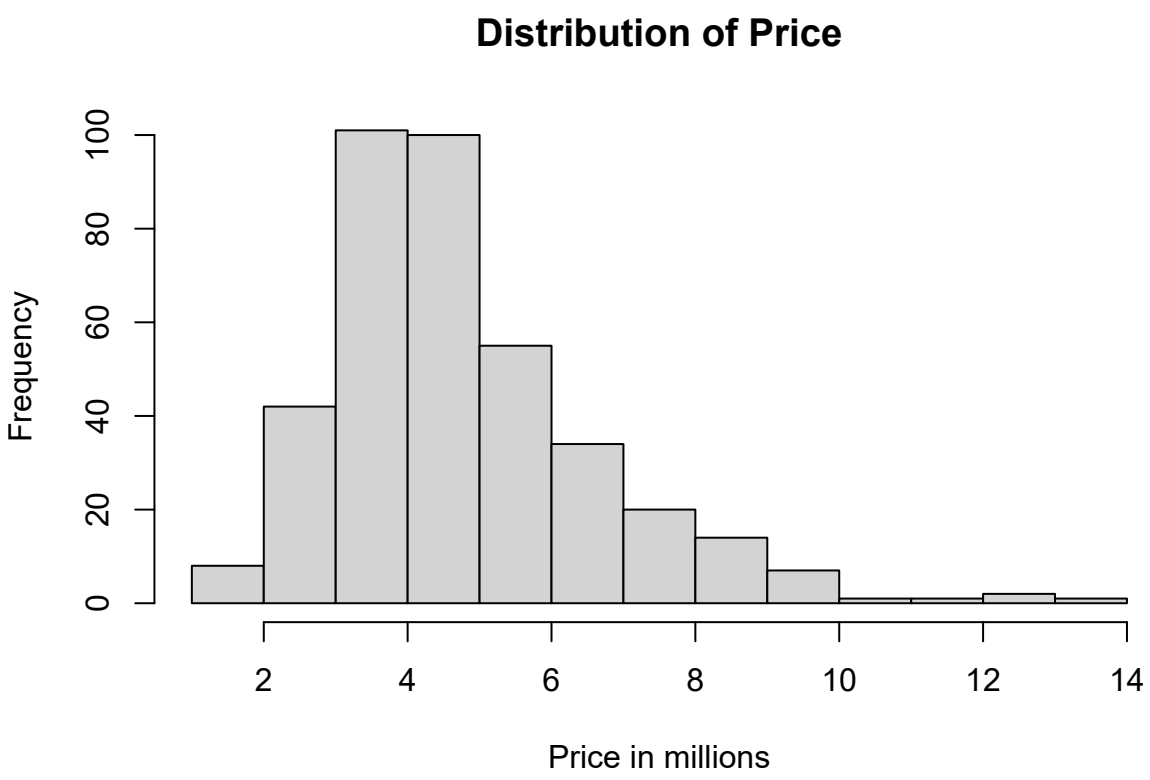
```
##      price               area          bedrooms       bathrooms
##  Min.   : 1750000   Min.   : 1650   Min.   :1.000   Min.   :1.00
##  1st Qu.: 3473750   1st Qu.: 3588   1st Qu.:2.000   1st Qu.:1.00
##  Median : 4340000   Median : 4600   Median :3.000   Median :1.00
##  Mean   : 4763635   Mean   : 5178   Mean   :2.953   Mean   :1.28
##  3rd Qu.: 5740000   3rd Qu.: 6360   3rd Qu.:3.000   3rd Qu.:2.00
##  Max.   :13300000   Max.   :16200   Max.   :6.000   Max.   :4.00
##     stories        mainroad          guestroom           basement
##  Min.   :1.000   Length:386         Length:386         Length:386
##  1st Qu.:1.000   Class :character   Class :character   Class :character
##  Median :2.000   Mode  :character   Mode  :character   Mode  :character
##  Mean   :1.793
##  3rd Qu.:2.000
##  Max.   :4.000
##  hotwaterheating    airconditioning       parking        prefarea
```

```
##  Length:386         Length:386         Min.   :0.000   Length:386
##  Class :character   Class :character   1st Qu.:0.000   Class :character
##  Mode  :character   Mode  :character   Median :0.000   Mode  :character
##                                        Mean   :0.715
##                                        3rd Qu.:1.000
##                                        Max.   :3.000
##  furnishingstatus
##  Length:386
##  Class :character
##  Mode  :character
##
##
##
```

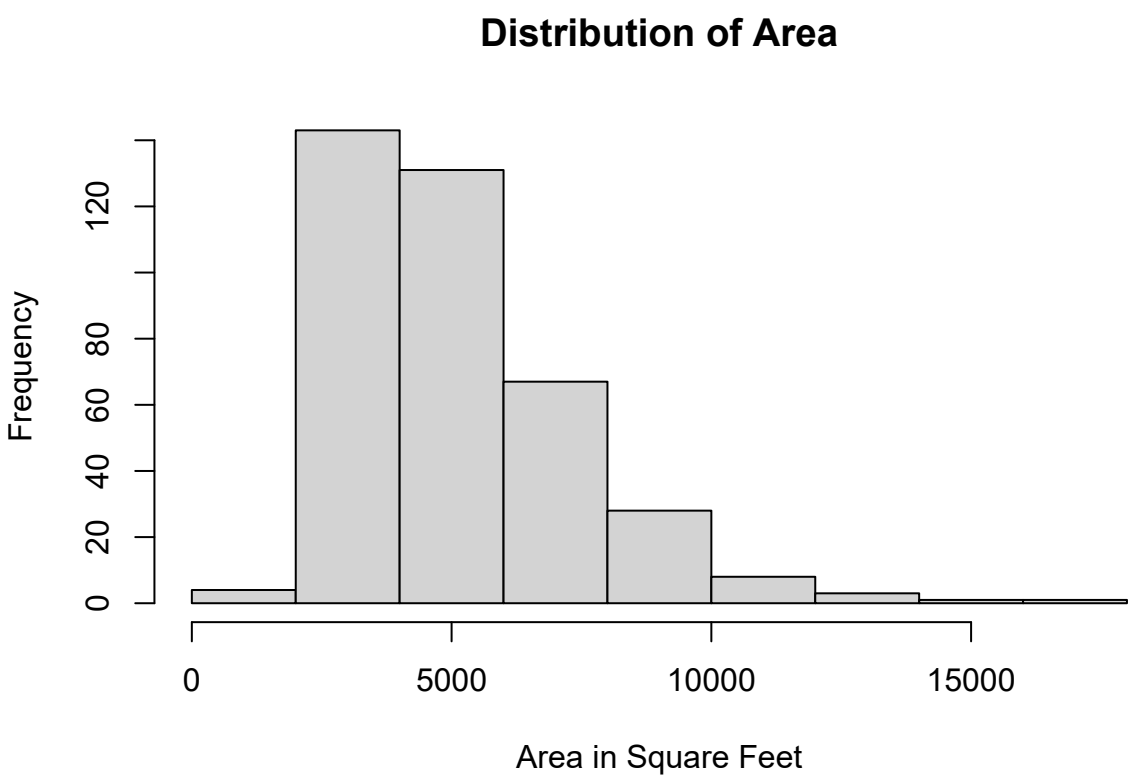
Evaluation:

```
##      price               area          bedrooms       bathrooms
##  Min.   : 1767150   Min.   : 1836   Min.   :1.000   Min.   :1.000
##  1st Qu.: 3430000   1st Qu.: 3600   1st Qu.:3.000   1st Qu.:1.000
##  Median : 4270000   Median : 4500   Median :3.000   Median :1.000
##  Mean   : 4774240   Mean   : 5083   Mean   :2.994   Mean   :1.302
##  3rd Qu.: 5771500   3rd Qu.: 6450   3rd Qu.:3.000   3rd Qu.:2.000
##  Max.   :12215000   Max.   :12944   Max.   :5.000   Max.   :3.000
##     stories        mainroad          guestroom          basement
##  Min.   :1.000   Length:159         Length:159         Length:159
##  1st Qu.:1.000   Class :character   Class :character   Class :character
##  Median :2.000   Mode  :character   Mode  :character   Mode  :character
##  Mean   :1.836
##  3rd Qu.:2.000
##  Max.   :4.000
##  hotwaterheating    airconditioning       parking         prefarea
##  Length:159         Length:159         Min.   :0.0000   Length:159
##  Class :character   Class :character   1st Qu.:0.0000   Class :character
##  Mode  :character   Mode  :character   Median :0.0000   Mode  :character
##                                        Mean   :0.6415
##                                        3rd Qu.:1.0000
##                                        Max.   :3.0000
##  furnishingstatus
##  Length:159
##  Class :character
##  Mode  :character
##
##
##
```

Price: It is important to recognize that this dataset contains homes with prices above 1 million. It is not clear that this is a US dataset, which would indicate that this is for luxury homes and/or high value markets.



Area: The area variable appears to be square footage of the home. We would traditionally expect that increases in area would lead to increases in price.



Bedrooms: While we expect increases in the number of bedrooms to increase the price, we also realize that at some point there are diminishing returns that an additional bedroom doesn't have as much of an impact. For example, increasing from one to two bedrooms should have significant increase in price, while increasing from four to five, perhaps not so much.

```
##   bedrooms   n
## 1        1   1
## 2        2 102
## 3        3 207
## 4        4  68
## 5        5   6
## 6        6   2
```

Based on the distribution of the number of Bedrooms, it may be best to categorize these with dummy variables; 2, 3, and 4+.

Bathrooms: Similar to the number of bedrooms, we would expect that an increase in bathroom count would lead to increases in price. Although similarly, having more than four bathrooms is likely going to lead to smaller increases.

```
##   bathrooms   n
## 1         1 288
## 2         2  89
## 3         3   8
## 4         4   1
```

Based on the distribution of the number of bathrooms, it may be best to categorize these with dummy variables; 2, and 3+.

Stories: Similar to the number of bedrooms and bathrooms, it would seem to make sense to classify homes with 3 or more floors together by introducing dummy variables; 2, and 3+.

```
##   stories   n
## 1       1 169
## 2       2 161
## 3       3  23
## 4       4  33
```

Parking: We are assuming that the parking variable represents the size of a garage. Similar to other variable the increase in price from no garage to a one car garage would be significant, while additional cars would add some lesser value. It would initially seem to make sense to introduce dummy variables; 1, and 2+.

```
##   parking   n
## 1       0 203
## 2       1  97
## 3       2  79
## 4       3   7
```

Furnishing Status: The furnishing status variable is taking on three values; unfurnished, semi-furnished, and furnished. Since we would consider unfurnished as the default state, we will use dummy variables; semi-furnished and furnished.

```
##   furnishingstatus   n
## 1        furnished 103
## 2   semi-furnished 160
## 3      unfurnished 123
```

Main Road: The main road variable is yes/no based on the street of the home. We will replace this with a dummy variable.

```
##   mainroad   n
## 1       no  50
## 2      yes 336
```

Guest Room: The guest room variable is yes/no based on the home having a guest room. It is unclear from the dataset source if this is in addition to the number of bedrooms, but we would expect houses with a guest room to have a higher price. We will replace this with a dummy variable.

```
##   guestroom   n
## 1        no 312
## 2       yes  74
```

Basement: The basement variable is yes/no based on the home having a basement. It is unclear if having a basement or not would lead to an increase in home price, but we will replace this with a dummy variable for analysis.

```
##   basement   n
## 1       no 249
## 2      yes 137
```

Hot Water Heating: Based on the distribution, we assume that the hot water heating variable represents if the house has in-floor heating, rather than forced air. Based on this assumption, we assume that having this feature would lead to higher house price. The variable will be replaced with a dummy variable for analysis.

```
##   hotwaterheating   n
## 1              no 366
## 2             yes  20
```

Air Conditioning: The air conditioning variable indicates if the house has central air conditioning. We would expect homes with air conditioning would have a higher price than those without. The variable will be replaced with a dummy variable.

```
##   airconditioning   n
## 1              no 264
## 2             yes 122
```

Preferential Area: The dataset source doesn't specify exactly what this variable represents. We are assuming that this is a yes/no value if the house is in a preferred neighborhood. We would expect houses with a yes to be higher price than those not.

```
##   prefarea   n
## 1       no 298
## 2      yes  88
```

Data Preparation:

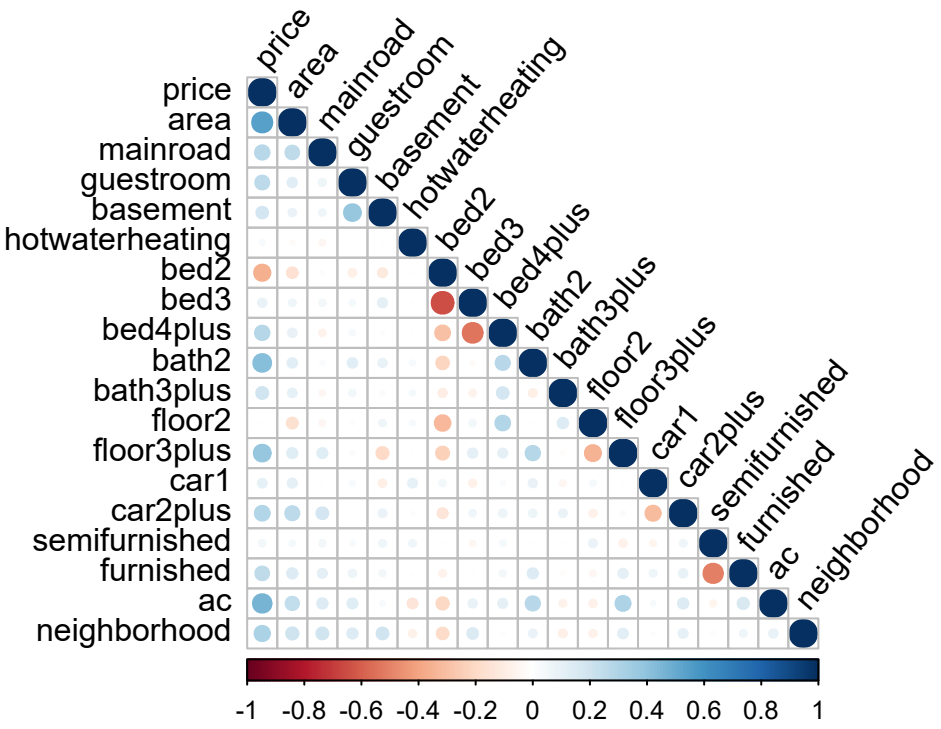
Based on our exploration, we do not have any blank values in our dataset.

Clean Function: We will introduce a clean function to replace our categorical variables with the dummy values. This will also ensure that our test and train datasets are processed in the same way.

Coorelation Plot: After cleaning the dataset looking at a correlation plot will give us confirmation about our initial examination for the variables.

```
cor_res <- cor(dftrain_clean)

corrplot(cor_res,
         type = "lower",
         order = "original",
         tl.col = "black",
         tl.srt = 50,
         tl.cex = 1)
```



The correlation plot generally confirms our initial expectations for the data.

Model Building:

Model Selection: