# Dissertation

Emmanouil Mertzianis

10/1/2021

## Contents

# 1 Introduction

## 1.1 The real estate market and the importance of predictive and descriptive models

## 1.2 The purpose of the paper

## 1.3 The structure of the paper

~~ TODO ~~

# 2 Literature

~~ TODO ~~

# 3 Exploratory Data Analysis

Before we start searching for the best regression model through formal data analysis and model fitting, it is important to explore our data through numerical and graphical summaries. This will allow for a better understanding of the patterns in and the structure of our data and it will enable us to make educated decisions during model fitting. For this purpose, we start by exploring each variable individually and, then, we focus on the relationships between the variables with emphasis on the ones related to the sale price, which is the variable of interest.

## 3.1 Exploring variables individually

Table 1: Summary statistics for the categorical variables in the initial data set.

| Variable | # missing | Unique lvls | Counts |
|---|---|---|---|
| bath | 0 | 5 | 1: 46, 2: 222, 3: 198, 4: 33, 63: 1 |
| parking | 0 | 4 | Covered: 105, No Parking: 73, Not Provided: 126, Open: 196 |

Table 2: Summary statistics for the numerical variables in the initial data set.

| Variable | # missing | Mean | SD | Min | 25% | Median | 75% | Max |
|---|---|---|---|---|---|---|---|---|
| elevation | 0 | 30.274 | 5.198555e+00 | 9 | 27.00 | 30.0 | 34.00 | 47 |
| dist_am1 | 0 | 8258.486 | 2.590404e+03 | 604 | 6439.75 | 8219.0 | 10011.25 | 20662 |
| dist_am2 | 0 | 11036.594 | 2.592219e+03 | 4402 | 9229.25 | 11015.0 | 12848.50 | 20945 |
| dist_am3 | 0 | 13092.760 | 2.629431e+03 | 4922 | 11215.75 | 13188.0 | 14775.75 | 23294 |
| sqft | 0 | 1816.096 | 5.721306e+02 | 932 | 1588.50 | 1770.5 | 2003.00 | 12730 |
| precip | 0 | 793.160 | 2.724887e+02 | -110 | 610.00 | 790.0 | 980.00 | 1530 |
| price | 0 | 510508.840 | 5.556979e+05 | 124333 | 380271.00 | 481042.0 | 593750.25 | 12500000 |

As a first step, we are interested in the summary statistics of the individual numerical and categorical variables in our data. The tables 1 and 2 contain useful statistics about the variables, prior to making any

alterations to the original data set. We note that there are no missing values for any of our variables in the data.

Regarding the categorical variables, we observe that there are five and four unique levels for the categorical variables *"bath"* and *"parking"*, respectively. Table 1 shows that most of the sale entries refer to houses with two baths or an "open" type parking. However, the most important observation to note here is a single entry with 63 bathrooms, which is exceedingly higher than all the rest observations in our data that are limited to just 4 bathrooms at maximum. Such an observation is likely to be an outlier and the exploratory analysis to follow further underpins this assumption.

Table 2 shows statistics about the numerical variables. It becomes apparent that the numerical variables are measured in different numerical scales with differences in the magnitude of their values. In terms of magnitude and standard deviation in ascending order:

- *"elevation"* presents the smallest values that do not exceed the value of 47 and the smallest standard deviation.
- *"precip"* and *"sqft"* come second and third, respectively, with the latter having almost double the standard deviation of the former.
- The three variables representing the distance from three chosen amenities (i.e. *dist_am1*, *dist_am2* and *dist_am3*) exhibit almost the same standard deviation. However, it seems that the $75^{th}$ percentile of *"dist_am1"* is, relatively close to the $25^{th}$ percentile of *"dist_am2"*, while the $25^{th}$ percentile of *"dist_am3"* is a bit higher than the median of *"dist_am2"*. This could indicate that, on average, the distance of houses from *"Amenity 1"* could be significantly smaller than that from *"Amenity 2"* and equally for the distances of houses from the *"Amenity 2"* and *"Amenity 3"*.
- The numerical scale and the standard deviation of *"price"* are the largest among all numerical variables. Also, it is interesting to point out that there exists a high outlier in *"price"*, even relative to its large magnitude, as it is 21.5755574 times the standard deviation greater than the mean value. The boxplot in figure 1 further illustrates the extreme outlier in *"price"*.

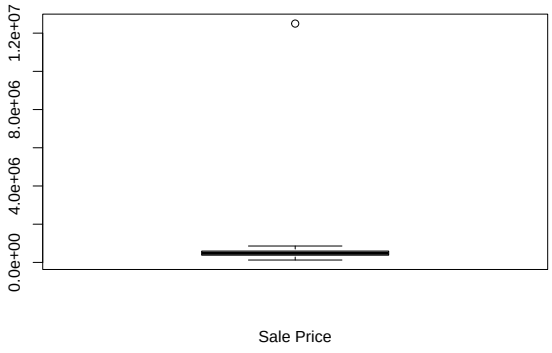


Figure 1: Boxplot of sale price.

Table 3 focuses on the aforementioned extreme outlier in *"price"*. The table reveals that the outlier (that is, the $348^{th}$ observation) contains extreme values in other variables as well. Specifically, that same observation is the one related to the 63 bathrooms, which we have already noted as a possible extreme value, and through further exploration we can show that its value of 12730 square feet is also extremely high. Those findings suggest that observation 348 could have probably come from a different population compared to the rest of the observations in the data. In any case, we lack enough data between this extreme observation and the

rest ones, to the point that any model fitting with this outlier included would result in speculating after some range of values and would probably lead to a heavily influenced model. Therefore, we conclude that **we have enough evidence to support our decision on removing observation 348 before we move on any further**.

Table 3: The extreme observation in the variable *'price'*.

| | elevation | dist_am1 | dist_am2 | dist_am3 | bath | sqft | parking | precip | price |
|---|---|---|---|---|---|---|---|---|---|
| 348 | 31 | 20662 | 20945 | 23294 | 63 | 12730 | Covered | 1130 | 12500000 |

Table 4: Summary statistics for the numerical variables after removing the extreme outlier.

| Variable | # missing | Mean | SD | Min | 25% | Median | 75% | Max |
|---|---|---|---|---|---|---|---|---|
| elevation | 0 | 30.27255 | 5.20367 | 9 | 27.0 | 30 | 34.0 | 47 |
| dist_am1 | 0 | 8233.62926 | 2532.61067 | 604 | 6434.5 | 8210 | 9984.5 | 16233 |
| dist_am2 | 0 | 11016.73747 | 2556.47368 | 4402 | 9219.5 | 11006 | 12842.0 | 18281 |
| dist_am3 | 0 | 13072.31663 | 2591.98862 | 4922 | 11215.5 | 13179 | 14771.0 | 20263 |
| sqft | 0 | 1794.22445 | 297.20037 | 932 | 1588.0 | 1770 | 2002.5 | 2667 |
| precip | 0 | 792.48497 | 272.34339 | -110 | 610.0 | 790 | 980.0 | 1530 |
| price | 0 | 486481.80361 | 142096.23500 | 124333 | 380125.0 | 480167 | 593167.0 | 857667 |

After removing the outlier, our conclusions about the variables *"elevation"*, *"dist_am1"*, *"dist_am2"*, *"dist_am3"* and *"precip"* are similar to the ones we derived earlier. However, we observe a significant drop in the maximum value of *"sqft"* along with a significant decrease in its standard deviation, which has now become relatively close to that of *"precip"*. Also, the maximum value and the standard deviation of *"price"* incurred a large drop.

The boxplots in figure 2 present graphically the already discussed differences in the magnitude and the variation between the numerical variables, by gradually removing variables from plot to plot. Interestingly, the boxplots suggest that the sample distributions of all numerical variables are fairly symmetrical as we observe the median to lie almost at the middle of the IQR box and roughly equal tails at the top and the bottom. This observation is backed by the computed numerical statistics, where the median is reported to be quite close to the mean value for every numerical variable. **A closer view using histograms in figure 3 reveals that the sample distribution of all numerical variables in the data resembles that of a sample coming from a Normal Distribution.**

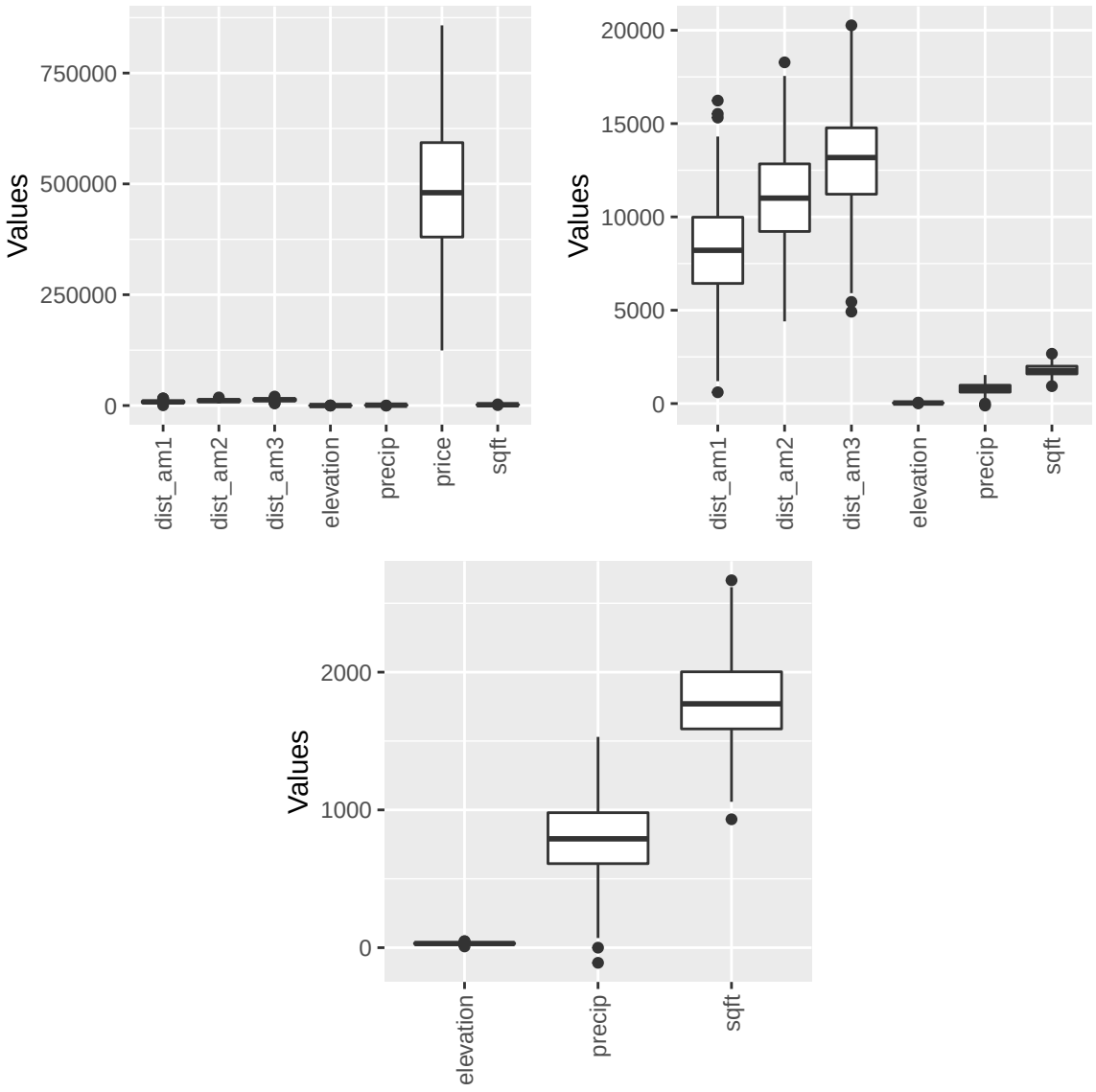


Figure 2: Boxplots on all numerical variables without the extreme outlier.

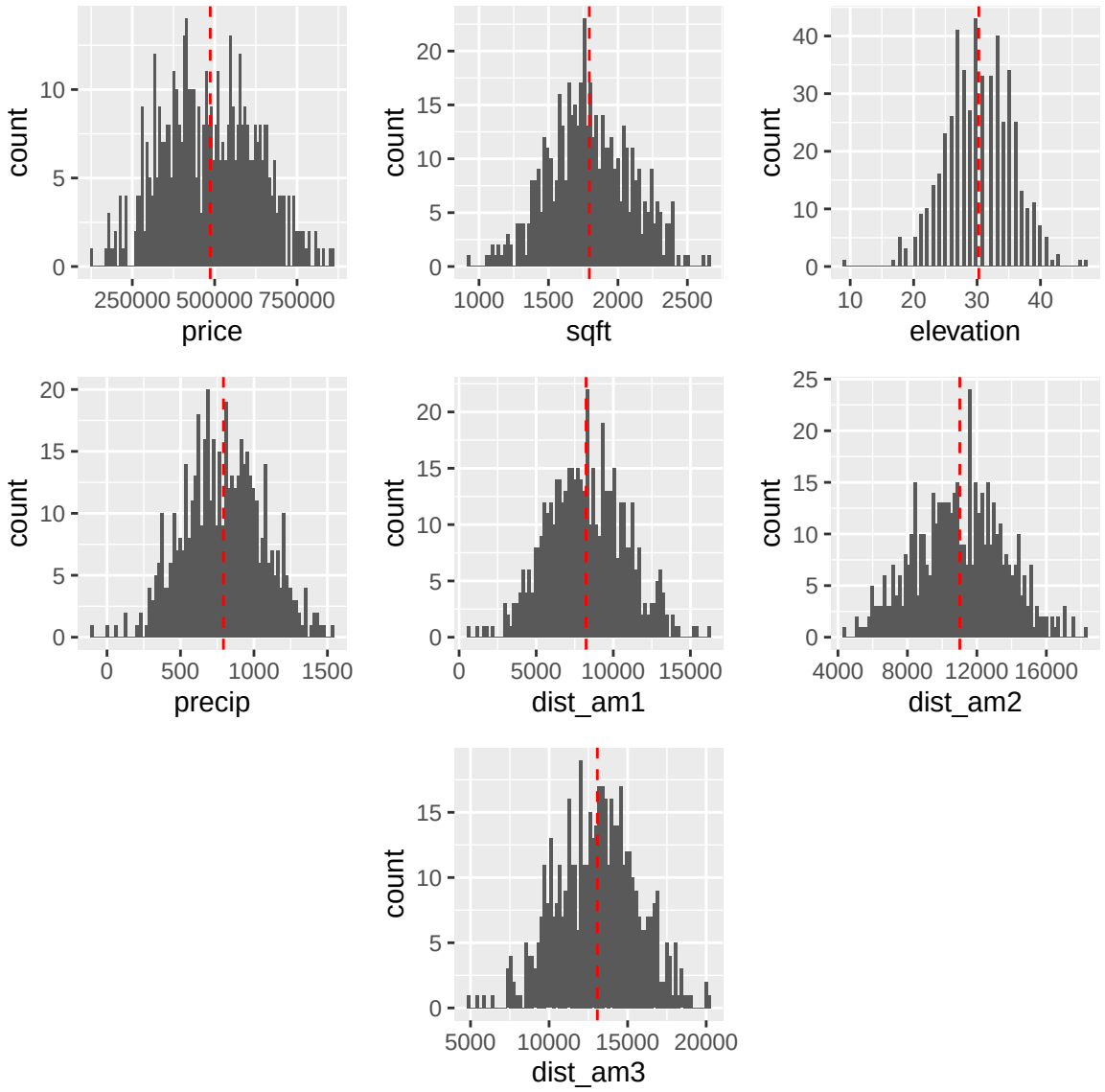


Figure 3: Histograms on all numerical variables without the extreme outlier. The red dashed line represents the mean value of the variable's values.

## 3.2 Exploring relationships

The primary objective of this paper is to create the most accurate predictive model about *"price"* using the rest of the available variables in the data. Therefore, it is of interest to explore the relationships of *"price"* against the other variables. For this purpose, we first explore the relationships of *"price"* against all other numerical variables and, then, we focus on the categorical variables. Finally, we analyse the relationships of *"price"* against the numerical variables on different levels of the categorical variables.

### 3.2.1 Numerical explanatory variables

Figure 4 depicts the relationships of *"price"* against all the rest numerical variables in the form of scatterplots with a simple linear regression line superimposed. From the scatterplots, we observe a random scattering of the data points across all values with no obvious patterns suggesting that there is little association between *"price"* and any one of the other numerical variables. Additionally, the small slope of the superimposed regression lines along with the little correlation revealed in table 5 show that there is no linear relationship between any of the numerical variables and *"price"*. However, we observe a high correlation between *"dist_am3"*/*"dist_am1"* and a moderate correlation between *"dist_am1"*/*"dist_am2"* and *"dist_am2"*/*"dist_am3"* indicating that there is possibly multicollinearity between the distance variables.

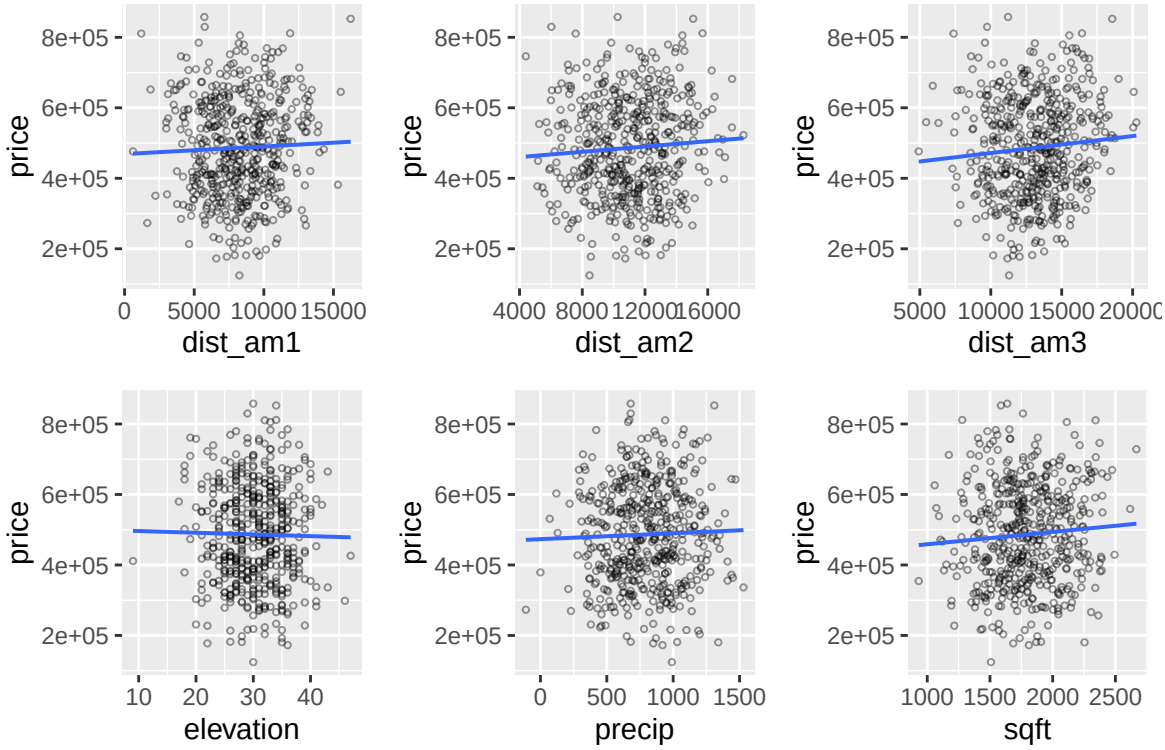


Figure 4: Scatterplots of 'price' against all the rest numerical variables. The simple linear regression line is superimposed on the plots.

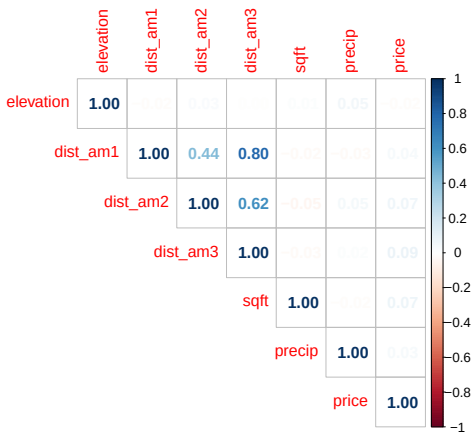


Figure 5: Correlation between all numerical variables.

### 3.2.2 Categorical explanatory variables

Figure 6 depicts the sample distribution of *"price"* for each level of *"bath"* or *"parking"*. Regarding *"parking"*, we observe quite a significant overlap between the boxplots with small differences between them suggesting that there is little to no difference in the sale price for houses in different *"parking"* categories. In contrast, we see that there is no overlap between the boxplots of *"price"* in different *"bath"* categories with the sale price actually increasing as the number of bathrooms increases. As it can be seen more clearly in figure 7, the seemingly normally-distributed sample of *"price"* is completely partitioned based on *"bath"* into 4 non-overlapping chunks. These findings indicate strongly that the categorical variable *"bath"* is a significant predictor with a positive relationship with *"price"*.

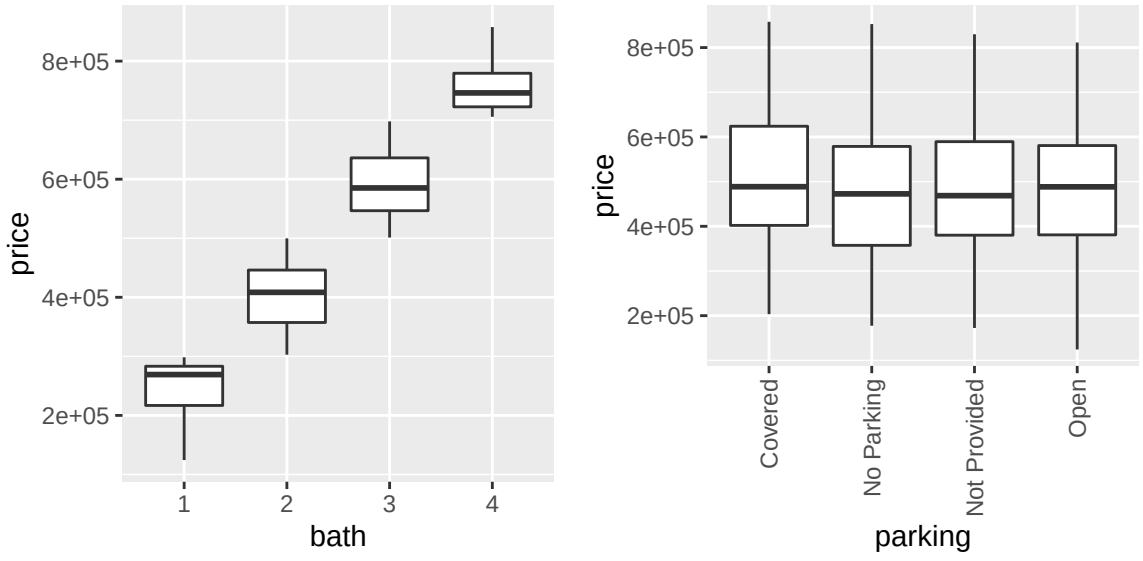


Figure 6: Boxplots of 'price' by 'bath' and by 'parking'.

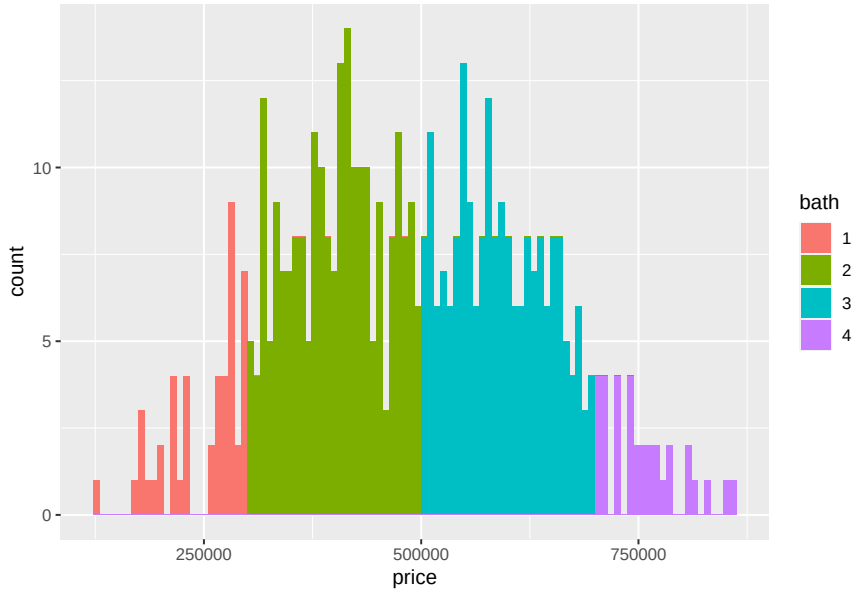


Figure 7: Histogram of 'price' coloured by 'bath'.

### 3.2.3 Interactions

Finally, it is important to investigate any interactions between categorical and numerical explanatory variables. Figure 8 shows the relationship of each numerical variable with *"price"*, on each *"bath"* level, in the form of scatterplots with simple regression lines superimposed. On all plots, we observe four distinct layers of data corresponding to each *"bath"* level. As we have already explained, these layers correspond to the full partitioning of the sample distribution of *"price"* into four non-overlapping chunks when considering the variable *"bath"*. It is apparent that, regardless of the *"bath"* category they belong to, the observations in the data extend fairly across the whole range of values on all numerical explanatory variables.

Excluding *"precip"*, the nature of the relationship between the rest of the explanatory variables and *"price"* is roughly the same on each level of *"bath"* and the relationships do not seem to improve for any of the variables when we take *"bath"* into consideration. In the case of the relationship between *"precip"* and *"price"*, we observe a slightly positive trend on all levels except for the houses with one bathroom, where the relationship becomes negative. In order to explore that interaction even further, we isolate the level "1" of *"bath"* by introducing a new categorical variable called *"bathBinary1"* which takes the value "More than 1" in the case of observations with 2 or more bathrooms and the value "1" in the case of just one bathroom.

When we assess the same relationships under each category of *"parking"*, it can be shown that there is an almost complete overlap between the layers of data of the different *"parking"* levels, with a random scattering of the data points. This suggests that the single regression line model should be suitable for each relationship with *"price"*, with no difference in the slopes or the intercept terms among the categories (i.e. no interaction with *"parking"*). However, it appears that in the relationship of *"price"* and *"dist_am1"* (left plot in figure 9) there is a difference in the slope of the simple linear regression lines between the *"parking"* categories, except for the slopes in the categories "Covered" and "Not Provided" where they are roughly equal. Therefore, to study the possible interaction of *"dist_am1"* and *"parking"*, we introduce a new categorical variable called *"parkingCNoP"*, which merges the categories "Covered" and "Not Provided" into a single category (right plot in figure 9).

Finally, we note that other transformations we applied on the numerical explanatory variables and on *"price"* did not seem to improve the relationships between them.

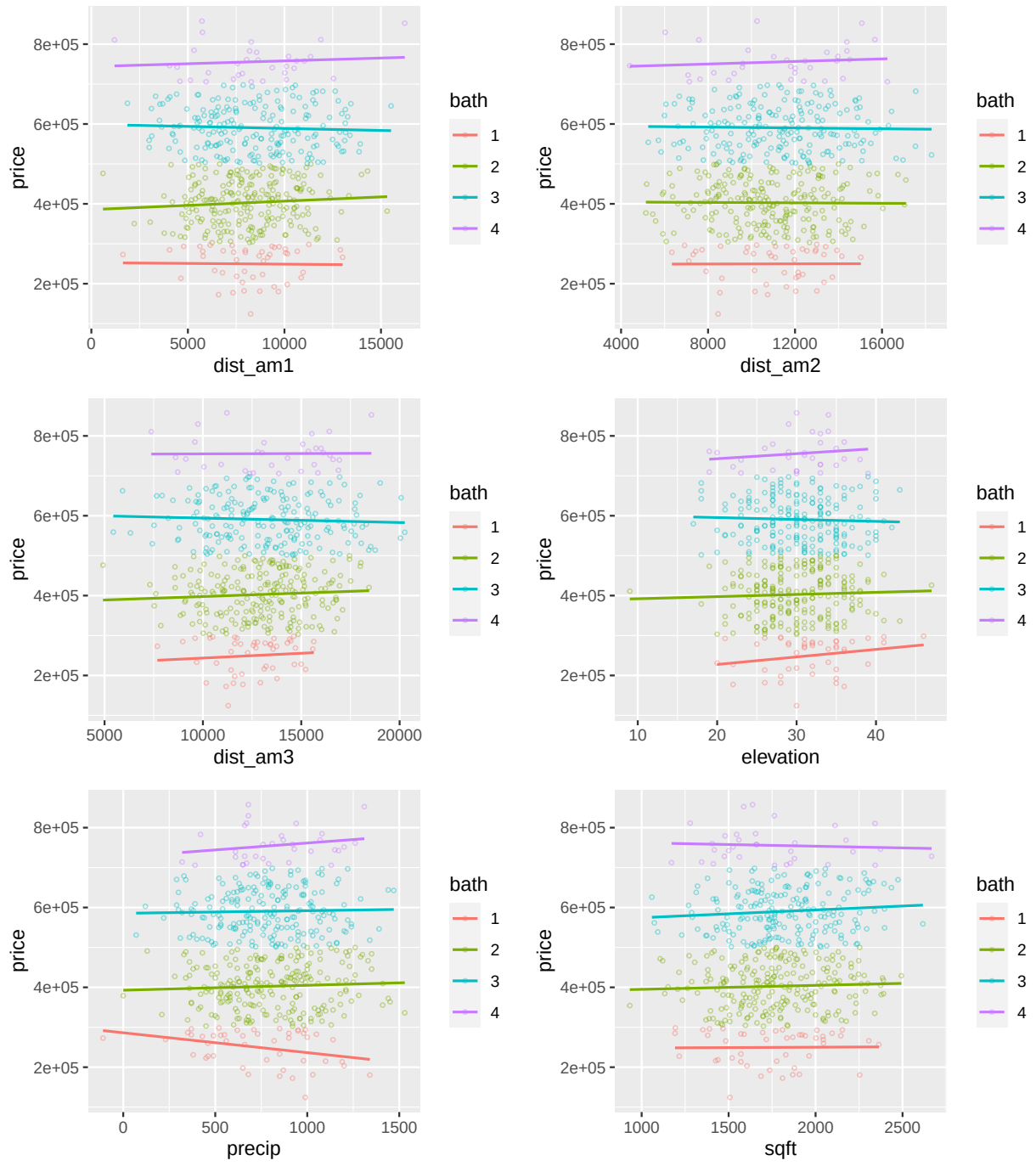


Figure 8: Scatterplots of 'price' against all the rest numerical variables and a simple linear regression line superimposed, coloured by 'bath'.

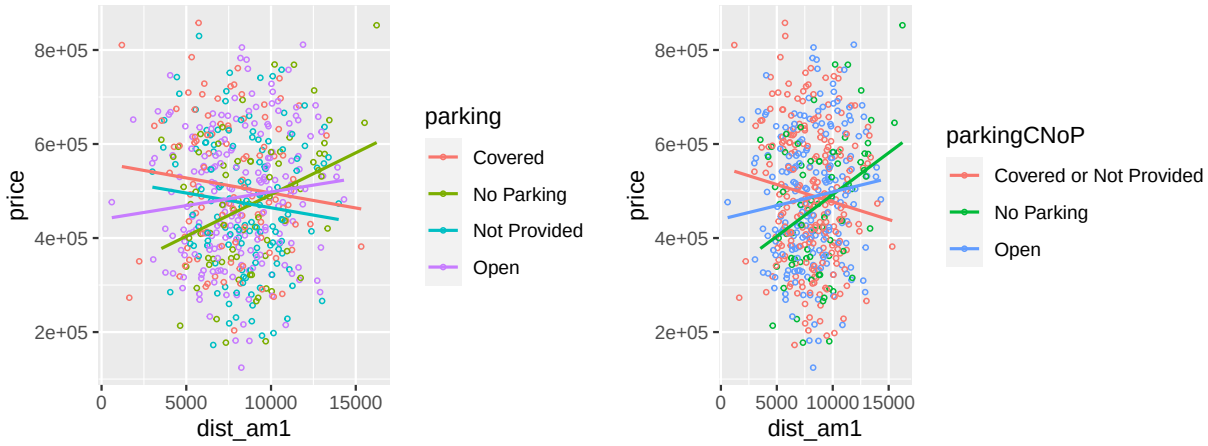


Figure 9: 'Price' against 'dist_am1' on all 'parking' levels and on the levels of 'parkingCNoP'.

# 4 Model Fitting: Selecting the best possible regression model

The **primary objective** of this paper is to search for and design the best possible *predictive* regression model to accurately predict the sale price of a house based on the rest of its attributes. The main approach we chose for this problem is **variable selection on the full model via backward elimination**.

For this purpose, we decided on assessing our models' performance and conduct variable selection based on the validation (mean squared)prediction error:

$$\hat{MSPE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

We achieve this by splitting our original data set into three subsets; training, validation and test set. These sets will be used for training(fitting) the models, assessing their prediction performance(out-of-sample prediction) and calculating the expected prediction performance on the best model, respectively. The split we chose to use is:

- 299 observations for training,
- 150 observations for assessment and
- 50 observations for reporting the final expected prediction performance.

Finally, we will consider moving beyond linear models by smoothing the best regression model we found using Ridge regression on its subset of predictors and see whether we can improve the prediction performance even further. Also, we will investigate whether variable selection via Lasso regression yields a more powerful model than backward elimination.

## 4.1 Fitting the full model

In section 3, we discussed about the different available variables in our data and we explored numerically and graphically the relationships between them and especially with *"price"*. Based on our findings, we select as our starting model the full additive model with the original variables, including interactions between *"precip"* - *"bathBinary1"* and *"dist_am1"* - *"parkingCNoP"*:

$$price_i = \beta_0 + \beta_1 \cdot \mathbb{I}bath|2_i + \beta_2 \cdot \mathbb{I}bath|3_i + \beta_3 \cdot \mathbb{I}bath|4_i +$$

$$\beta_4 \cdot \mathbb{I}parking|NoParking_i + \beta_5 \cdot \mathbb{I}parking|NotProvided_i + \beta_6 \cdot \mathbb{I}parking|Open_i +$$
$$\beta_7 \cdot sqft_i + \beta_8 \cdot elevation_i + \beta_9 \cdot dist\_am2_i + \beta_{10} \cdot dist\_am3_i +$$
$$\beta_{11} \cdot precip_i + \beta_{12} \cdot \mathbb{I}bathBinary1|Morethan1_i \cdot precip_i +$$
$$\beta_{13} \cdot dist\_am1_i + \beta_{14} \cdot \mathbb{I}parkingCNoP|NoParking_i \cdot dist\_am1_i + \beta_{15} \cdot \mathbb{I}parkingCNoP|Open_i \cdot dist\_am1_i$$
$$+ \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2), \quad i = 1, ..., 299$$

, where the indicator variables(starting with $\mathbb{I}$) are equal to 1 when the $i^{th}$ observation's corresponding categorical variable takes the relevant value(separated from the variable name with "|") and 0 otherwise. The baseline categories are **1** for *"bath"* and *"bathBinary1"*, **Covered** for *"parking"* and **Covered or Not Provided** for *"parkingCNoP"*.

Table 5 shows the OLS estimates of the full model's regression coefficients along with their p-value and 95% Confidence Interval(C.I.), when fitted on the training set. The reported p-values suggest that the coefficients for the categories of *"bath"* are significant, considering all the predictors we used in the model and choosing a significance level of 5%. Furthermore, the absence of overlap between the 95% C.Is of the *"bath"* categories further underlines the importance of this predictor. Also, we observe a p-value close to the 5% threshold for the difference in the coefficient of *"precip"* when considering houses with more than 1 bathrooms compared to the baseline caategory. All the rest estimates have a p-value that is greater than the chosen significance level or, equivalently, they have a 95% C.I. that includes the value of 0 as a plausible value for the coefficient in the population's regression model.

Table 5: The estimated regression coefficients along with the corresponding p-values and 95% Confidence Intervals from fitting the full additive model on the training set.

| term | estimate | p_value | lower_ci | upper_ci |
|---|---|---|---|---|
| intercept | 274279.647 | 0.000 | 190361.002 | 358198.293 |
| bath: 2 | 106553.631 | 0.000 | 53688.905 | 159418.356 |
| bath: 3 | 292193.260 | 0.000 | 239530.184 | 344856.336 |
| bath: 4 | 454783.607 | 0.000 | 398006.391 | 511560.823 |
| parking: No Parking | -47095.315 | 0.119 | -106327.118 | 12136.488 |
| parking: Not Provided | -14978.753 | 0.102 | -32957.525 | 3000.020 |
| parking: Open | -35614.978 | 0.126 | -81237.337 | 10007.381 |
| precip | -51.612 | 0.127 | -118.050 | 14.826 |
| dist_am1 | -0.645 | 0.806 | -5.800 | 4.510 |
| dist_am2 | -0.640 | 0.677 | -3.656 | 2.377 |
| dist_am3 | 0.663 | 0.766 | -3.716 | 5.042 |
| sqft | 11.229 | 0.273 | -8.907 | 31.365 |
| elevation | 435.430 | 0.445 | -684.597 | 1555.457 |
| precip:bathBinary1More than 1 | 64.755 | 0.070 | -5.417 | 134.928 |
| dist_am1:parkingCNoPNo Parking | 2.498 | 0.441 | -3.877 | 8.873 |
| dist_am1:parkingCNoPOpen | 2.157 | 0.427 | -3.184 | 7.497 |

Lastly, we can calculate the mean squared prediction error of the full model from predicting on the validation set, which is equal to $\hat{MSPE} = 3.050646 \times 10^9$. This metric will be used for performing variable selection as described in the following subsection.

## 4.2 Variable Selection

Although the full model is a good starting point to consider all predictors that could be possibly related to the response variable, it is almost certain that we have included excessive, unnecessary complexity. Usually,

such a model is very flexible, lacks stability (high variance) and captures sample specific features that are not generalisable to other data from the same population. In other words, these kinds of models excel in estimating on their training data set, but fail to demonstrate a good performance in out-of-sample prediction.

Since our goal is to construct the best predictive model, we wish to minimise the $\hat{MSPE}$ on the validation data set as much as possible and the method we chose to achieve this is **variable selection via backward elimination on the full model**. This iterative method starts by removing one variable at a time from the full model and calculating the $\hat{MSPE}$ after each removal, until it has removed all variables once. Then, the method identifies the variable that resulted in the best(lowest) $\hat{MSPE}$ value after its removal and updates the model such that it does not contain this variable any longer. The method continues by seeking the next best variable to remove on the updated model. This process is repeated until either there is only one independent variable left or no more improvement can be achieved.

In simple words, we decrease the overall complexity of our model by gradually removing unnecessary variables, while we continuously monitor the improvement of the $\hat{MSPE}$, until we reach the best possible value. It is easy to realise that this method belongs to the family of greedy algorithms as it consists of a series of consecutive, independent and non-reversible steps at which the optimal solution is chosen each time.

Table 6: Results from performing variable selection with backward elimination on the full model. The left column contains the variables that have been removed from the full model in the order of removal, while the right one shows the $\hat{MSPE}$ on the validation set after the removal of the variable on the left. The table contains the full model with its $\hat{MSPE}$ in the first row for reference.

| Removed variable | MSPE after removal |
|---|---|
| full model | 3050645988.1555 |
| parking | 2970745281.55559 |
| dist_am1*parkingCNoP - parkingCNoP | 2955058766.24976 |
| elevation | 2948877513.85971 |
| dist_am2 | 2944620560.75069 |

~~ TODO ~~

### 4.3 Regularised regression using Ridge and Lasso

~~ TODO ~~

## 5 Conclusions

~~ TODO ~~

## 6 Further Work

go beyond regression models, k-NN(but first, feature scaling) maybe. ~~ TODO ~~

## 7 References