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 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, almost equal to the 25^{th} percentile of *"dist_am2"*, while the 75^{th} percentile of *"dist_am2"* is roughly the same as the 25^{th} percentile of *"dist_am3"*. This could indicate that, on average, the distance of houses in the data set 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"*.

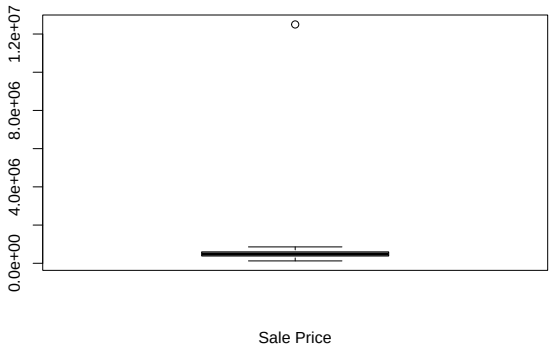

Figure 1: Boxplot of sale price.

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.**

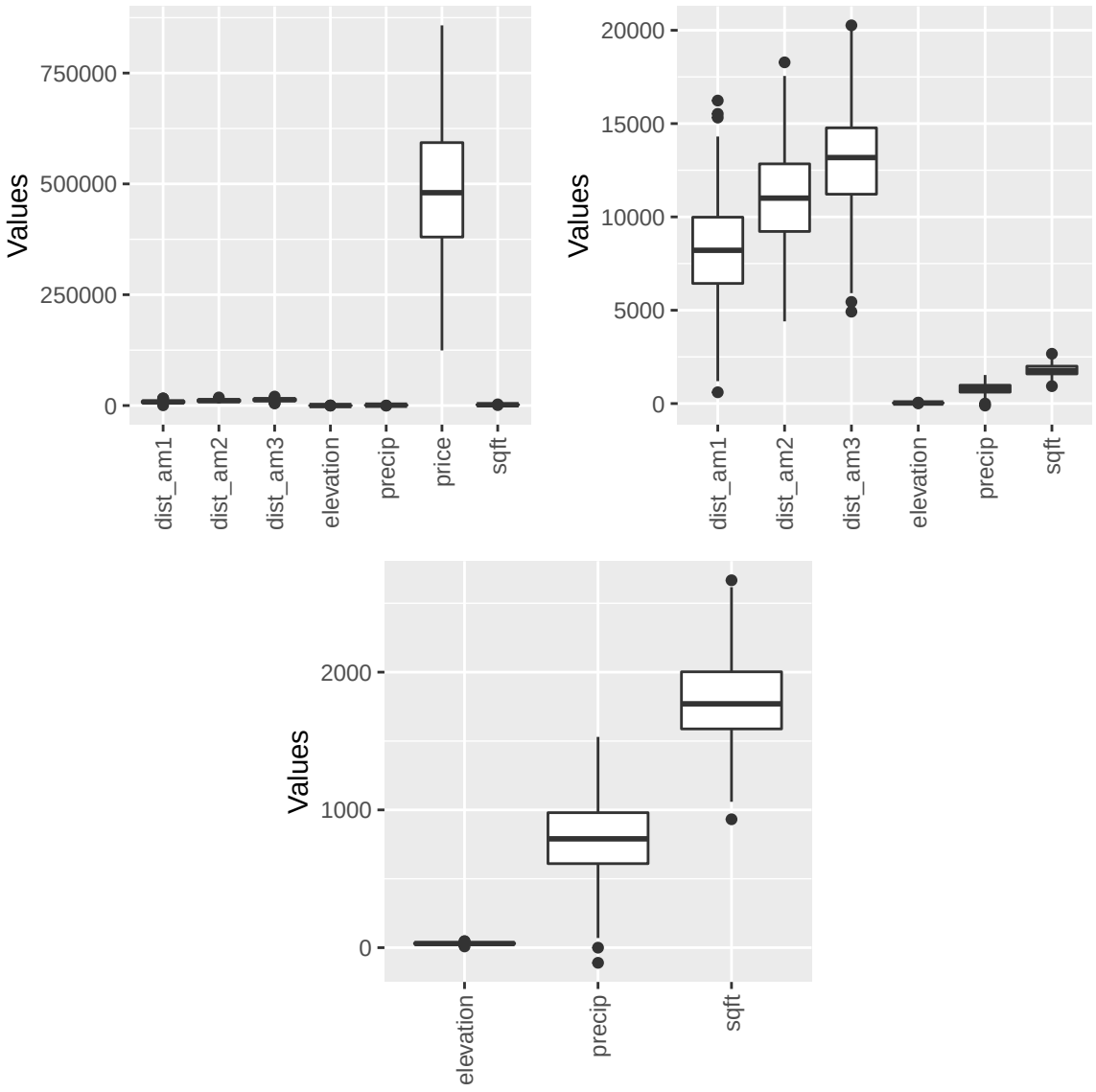

Figure 2: Boxplots on all numerical variables without the extreme outlier.

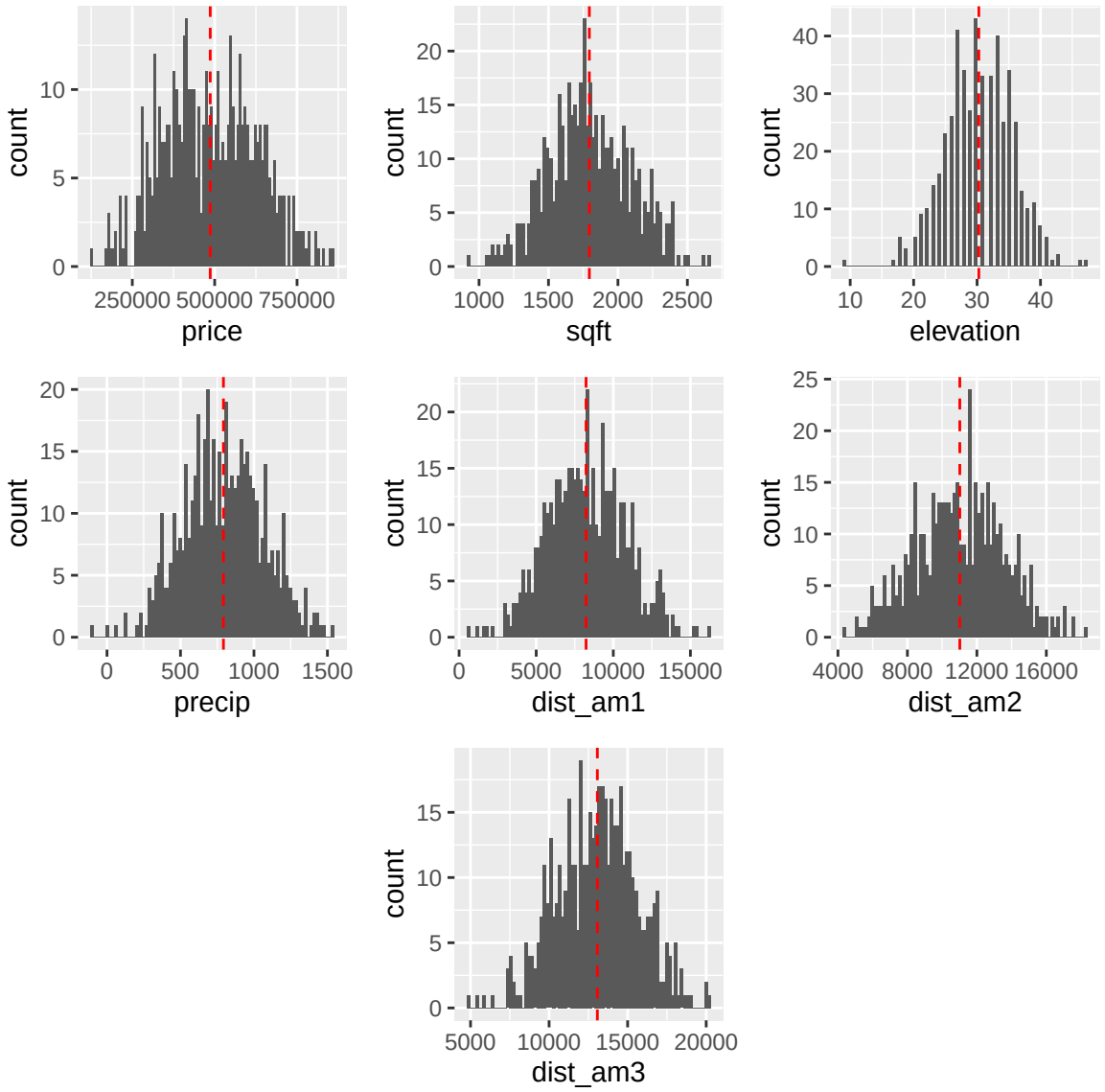

Figure 3: Histograms on all numerical variables without the extreme outlier. The red dashed line represents the mean value of the variable's values.

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 each one of the rest of the numerical 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, while considering the different levels of the categorical variables.

3.2.1 Numerical

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.

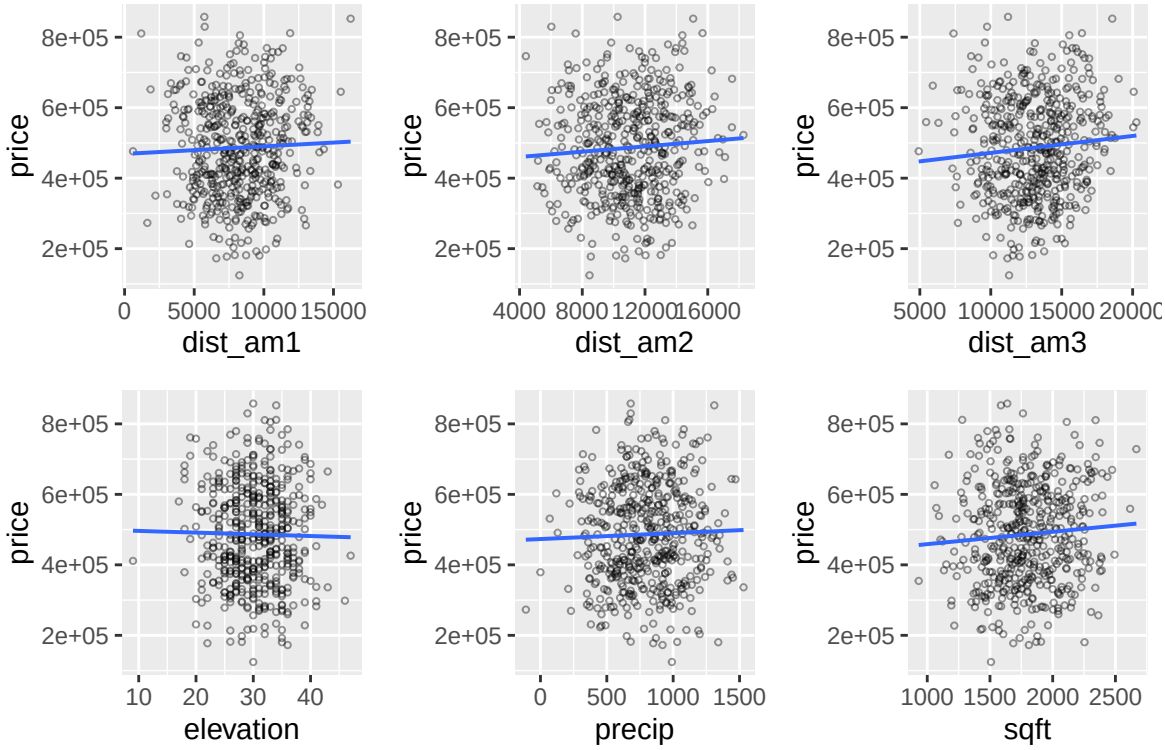

Figure 4: Scatterplots of 'price' against all the rest numerical variables. The simple linear regression line is superimposed on the plots.

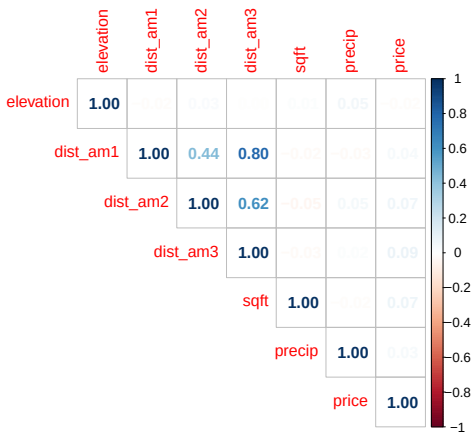

Figure 5: Correlation between all numerical variables.

3.2.2 Categorical

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"*.

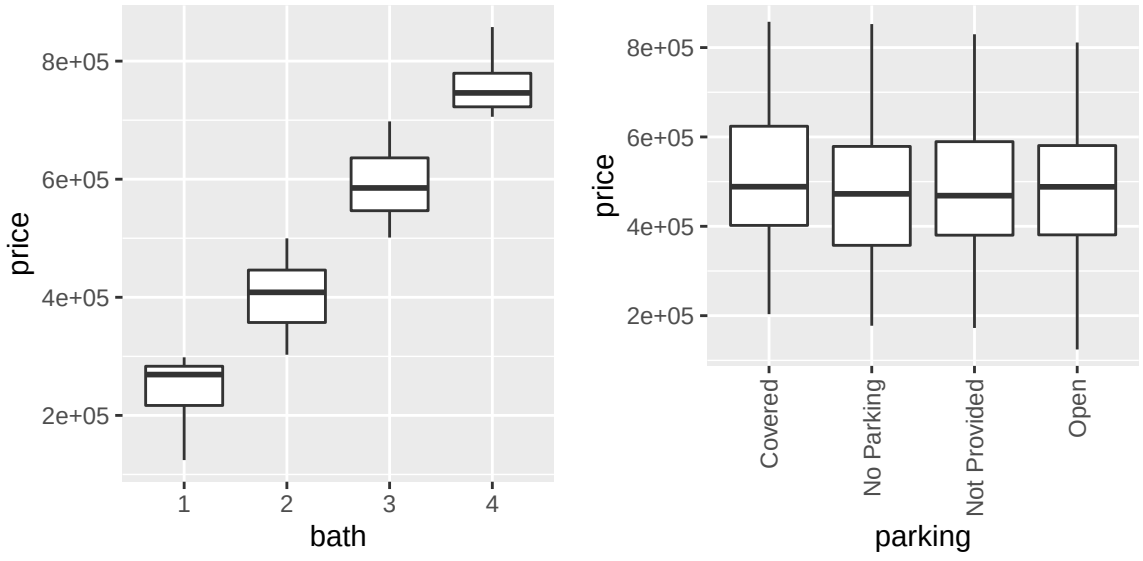

Figure 6: Boxplots of 'price' by 'bath' and by 'parking'.

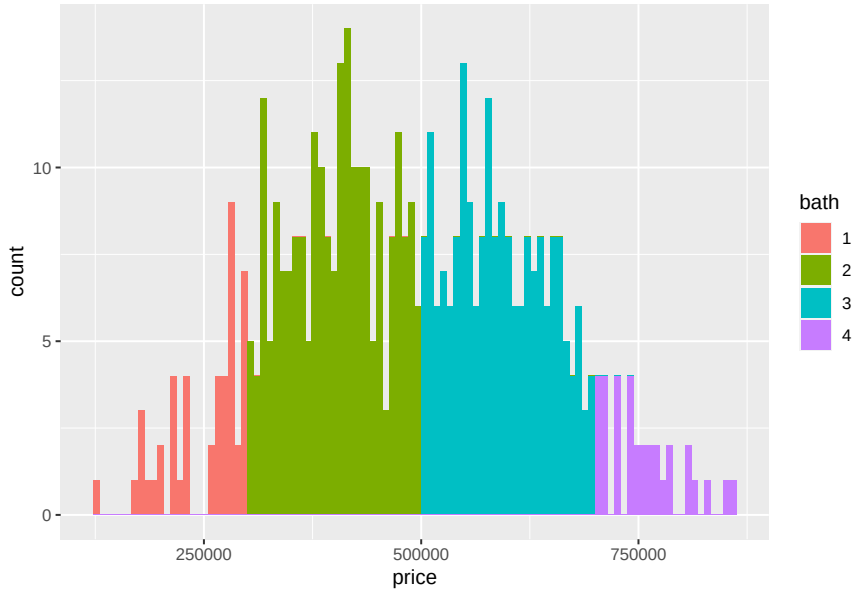

Figure 7: Histogram of 'price' coloured by 'bath'.

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. On all plots, we observe four distinct layers of data corresponding to each *"bath"* level. As we have already explained, these layers represent the full partitioning of the sample distribution of *"price"* into four non-overlapping chunks when considering the variable *"bath"*. It is apparent that there are observations across the whole range of values on all numerical explanatory variables on each different *"bath"* level and that the nature of the relationship between those and *"price"* stays the same regardless of the level of *"bath"*. This indicates that there is no interaction between *"bath"* and any of the numerical variables and that the additive model is probably appropriate in this case. The same can be proven about the categorical variable *"parking"* as well.

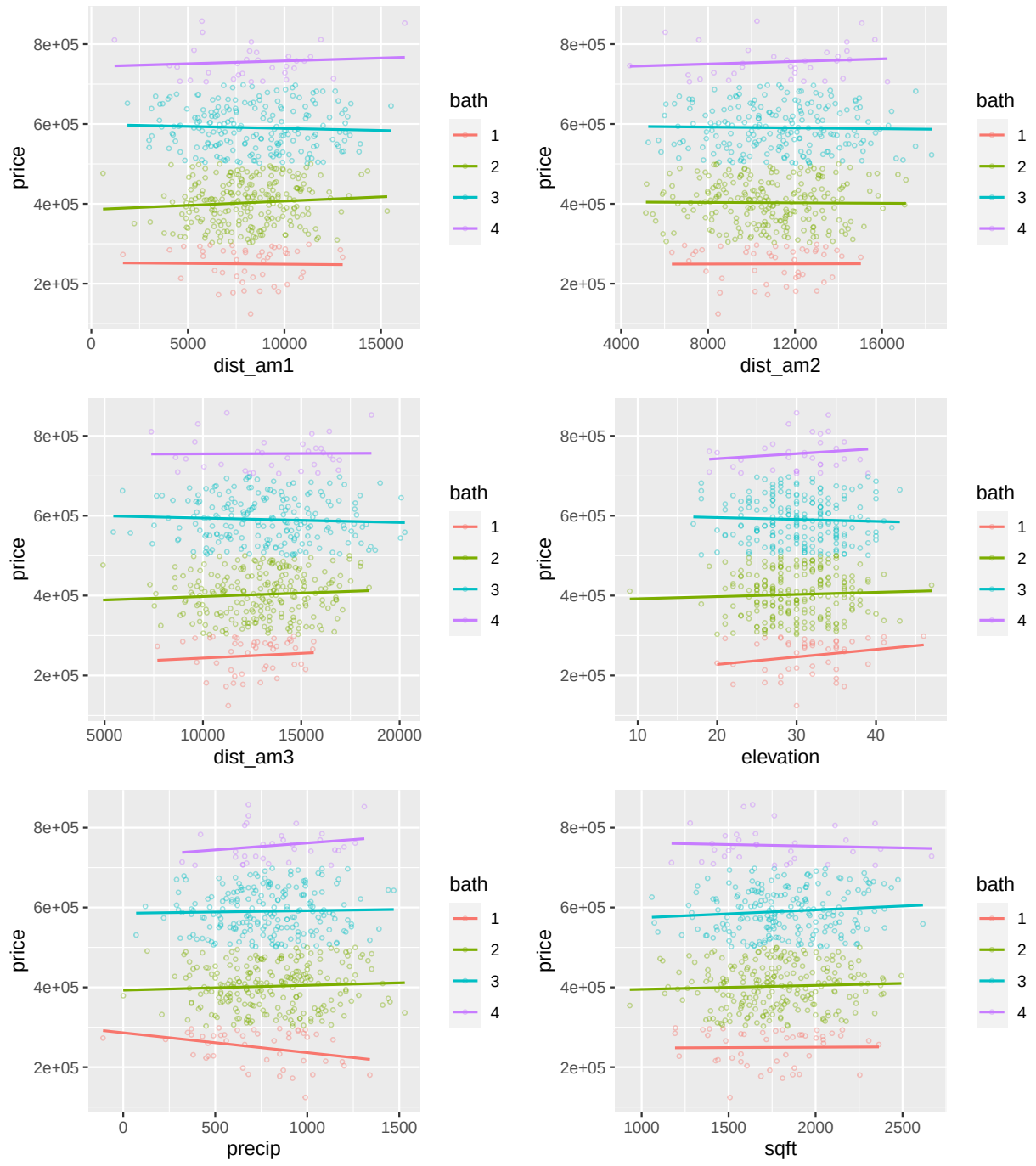

Figure 8: Scatterplots of 'price' against all the rest numerical variables and a simple linear regression line superimposed, coloured by 'bath'.

4 Model Fitting: Selecting the best possible regression model

~~ TODO ~~

5 Conclusions

~~ TODO ~~

6 Further Work

~~ TODO ~~

7 References