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.

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 almost equal to the 25^{th} percentile of *"dist_am2"* and the 75^{th} percentile of *"dist_am2"* is roughly the same as the 25^{th} percentile of *"dist_am3"*. This indicates 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 standard deviation of *"price"* are the largest among all numerical variables.

Additionally, it is interesting to point out that: 1. no missing values 2. high outlier for price even for its magnitude (times the standard deviation from the mean) 3. all numerical variables have a mean that's close to the median.

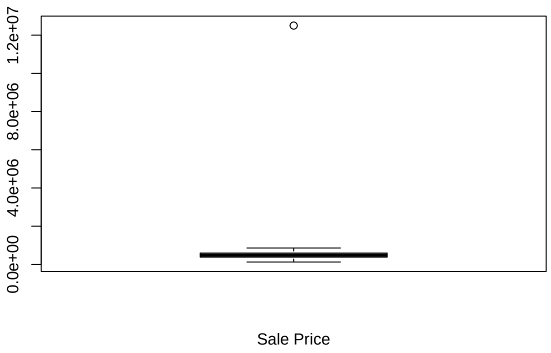

Figure 1: Boxplot of sale price.

Table 3: The extreme observation in the data.

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

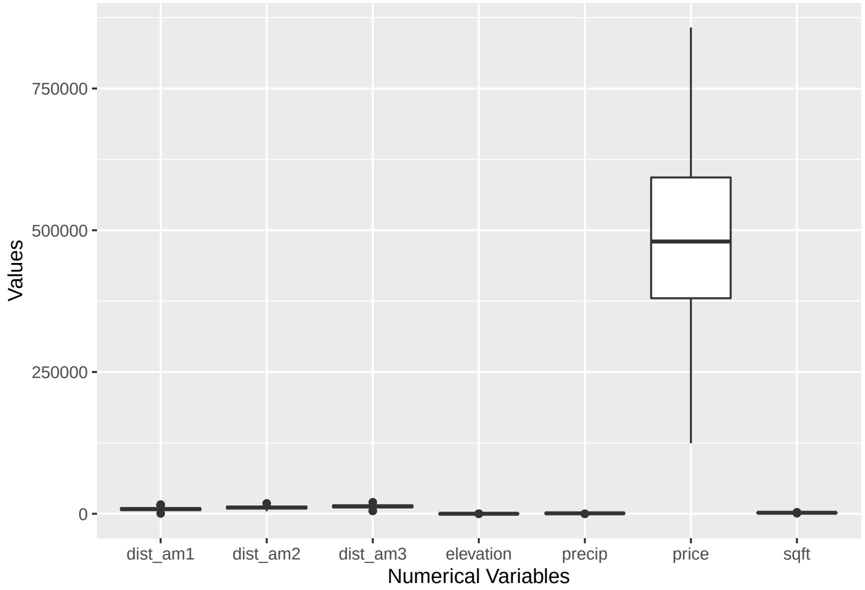

Figure 2: Boxplots on all numerical variables in the initial data set.

~~ TODO ~~

4 Model Fitting: Selecting the best possible regression model

~~ TODO ~~

5 Conclusions

~~ TODO ~~

6 Further Work

~~ TODO ~~

7 References