SLV Assignment1

Daniel Anadria Kyuri Park Ernst-Paul Swens Emilia Loescher

September 26, 2022

Introduction

The data set

We will explore the data from kaggle, which contains information about housing market in Italy. The data are scraped from one of the most relevant housing sales websites in Italy during the month of *August 2022*. and some of the data were removed due to translation limitations (e.g., extended text-based description, specific url of the post).

For each sale the following variables are available:

Table 1: Description of variables

Variable	Description
id	ID of the sale
timestamp	Timestamp consisting of 10 digits
location	Location on municipality level
title	Short description of property
price	Price in Euro
n_rooms	Number of rooms
floor	Floor
mq	Size in square meters
n_bathrooms	Number of bathrooms
year_of_construction	Year of construction
availability	Availability of property
energy_class	Energy class ranging from a+ to g
status	Status of the property
heating	Type of heating
has_garage	Garage present: yes (1), no (0)
has_terrace	Terrace present: yes (1), no (0)
has_garden	Garden present: yes (1), no (0)
has_balcony	Balcony present: yes (1), no (0)
has_fireplace	Fireplace present: yes (1), no (0)
has_alarm	Alarm present: yes (1), no (0)
has_air_conditioning	Air Conditioning present: yes (1), no (0)
has_pool	Pool present: yes (1), no (0)
has_parking	Parking present: yes (1), no (0)
has_elevator	Elevator present: yes (1), no (0)
is_furnished	Furniture present: yes (1), no (0)

Exploratory question

We choose to focus on three four exploratory questions focusing on housing prices in Italy.

Firstly, we explore if there is a geological trend in average housing prices in Italy. (E.g., Is housing more/less expensive in Northern Italy compared to Southern Italy?)

Secondly, we examine if there are differences in the amount of variance between the different regions.

Thirdly, it is of interest to us if there is a correlation between the missingness of housing price and the other variables.

Fourthly, we identify the variables that are most relevant when predicting housing prices in Italy.

Preparation

In order to start our exploratory analysis, we first load relevant packages and import the full data set. ## Load packages & Import data

```
# load packages
library(tidyverse)
library(skimr)

# import data
housing <- read.csv("data/housing_data_italy_august2022.csv", na.strings=c("","NA"), header = TRUE)
```

Preliminary analysis

We skim through our data using the `skimr` package.
The data consists of 223,409 rows (sales) and 25 columns (variables). Given our questions, we conclude that `id` (ID of the sale) `timestamp` (timestamp of the sale) and `title` (description of the property) are irrelevant and, hence, we exclude them from the data set for further analysis. In addition, we remove two columns that have only one unique value (`status`: "other" and `availibility`: "not free/other"), as these variables do not provide information specific to certain sales. Furthermore, in this first step, we see the data type of some variables are wrongly specified. We convert them to a correct type (e.g., `heating`: character –> factor, `has_xxx`: numeric –> factor, `is_furnished`: numeric –> factor).
Also, we create a new variable `age` which gives the age of the property in 2022 by subtracting the year of construction from 2022. There are some unreasonable years (e.g. 2209). Thus, we set `age` to NA for an age smaller than -4. The reason that we allow for negative age at all is that some property might be sold before construction is completed. Hence, 2026 as year of construction is reasonable.

```
housing_red <- housing %>%
  #For model building exclude the following variables:
  # select variables that have more than one unique value
  select(where(~n_distinct(.) > 1)) %>%
  # remove timestamp and title
  select(-c(id, timestamp, title)) %>%
  # fix the data type
  mutate(across(c(starts_with("has"), is_furnished, heating), factor)) %>%
  #Setting "year_of_construction" > 2026 to NA
  mutate(year_of_construction = replace(year_of_construction, year_of_construction > 2026, NA)) %>%
  #Transforming "year_of_construction" in age (2022-year of construction)
  mutate(age = 2022 - as.numeric(year_of_construction)) %>%
```

```
  # Removing "year_of_construction"
  select(-year_of_construction)
```

After the modification of the data, we take a look at the summary statistics for the data set to get a better overview over the data set.

```
# round up by 2 decimal places + disable scientific notation
options(digits = 2, scipen = 999)
# specify skimming function
my_skim <- skim_with(character = sfl(whitespace=NULL, min = NULL, max = NULL),
                     factor = sfl(ordered = NULL, top_counts = NULL, "ratio (autonomous or 1)" = ~(sum(. ==

                     numeric = sfl(p0 = NULL, p25=NULL, p50=NULL, p75=NULL,
                                   p100=NULL, hist=NULL, median = ~median(., na.rm=T),
                                   min = ~min(., na.rm=T), max = ~max(., na.rm=T), n_unique=n_unique))
# skim the data
my_skim(housing_red)
```

Table 2: Data summary

Name	housing_red
Number of rows	223409
Number of columns	20
Column type frequency:	
character	2
factor	12
numeric	6
Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate	empty	n_unique
location	0	1	0	7023
energy_class	679	1	0	12

Variable type: factor

skim_variable	n_missing	complete_rate	n_unique	ratio (autonomous or 1)
heating	0	1	2	0.90
has_garage	0	1	2	0.18
has_terrace	0	1	2	0.11
has_garden	0	1	2	0.17
has_balcony	0	1	2	0.10
has_fireplace	0	1	2	0.05
has_alarm	0	1	2	0.01
has_air_conditioning	0	1	2	0.30
has_pool	0	1	2	0.02
has_parking	0	1	2	0.02
has_elevator	0	1	2	0.06

skim_variable	n_missing	complete_rate	n_unique	ratio (autonomous or 1)
is_furnished	0	1	2	0.08

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	median	min	max	n_unique
price	39116	0.82	239939.0	7562062.01	135000	1	2147483647	2852
n_rooms	60323	0.73	3.5	0.99	3	2	5	4
floor	72365	0.68	1.8	1.13	2	1	52	22
mq	4034	0.98	158.6	128.68	117	1	999	976
n_bathrooms	14397	0.94	1.6	0.67	1	1	3	3
age	26	1.00	56.9	76.74	42	-3	1022	381

From the tables, we can see that there are 12 different energy classes and 7023 different locations. When taking a closer look at the location variable, one can see that they are given on a municipality level.

Regarding the variables of type factor, we see that there are no missing values. The ration of properties with an alarm is the lowest with 1% and the highest for air conditioning (30%). 90% of buildings have autonomous heating.

For the numeric variables, we observe that several have lots of missing values (e.g., `price`, `n_rooms`, `floor`, `mq`, `n_batohrooms`). We will discuss how we want to deal with this in section .

Exploratory analysis

Question 1: Geographical differences in average housing prices

We have described above, that each sales is assigned to one of 7023 locations on a municipality level. In order to create a plot to visualize the differences in average housing prices across Italy, we group them together into 20 regions. For this process, we used - ***Daniel, could you briefly describe the procedure here, please? :)***

Preparation

Plot

Conclusion

Question 2: Geographical differences in variation of housing prices

Preparation

Plot

Conclusion

Question 3: Missingness and Imputation

- ***Plot the missingness to see if there is any pattern for NA vs non-NA, correlation, imputation***

Preparation

Plots

Conclusion

Question 4: Relevant predictors for housing price

Preparation

Analysis

Conclusion

Overall Conclusion