SLV Assignment1

Daniel Anadria Kyuri Park Ernst-Paul Swens Emilia Loescher

September 23, 2022

Introduction

About dataset

We will use 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*. The data consist of more than 223000 sales posts and some of the data were removed due to translation limitations (e.g., extended text-based description, specific url of the post). Additionally, we use Italy regional data from [Italian National Institute of Statistics] (https://www.istat.it/it/archivio/222527). We will extract the regional coding and geographical shape information so that we can cluster the comunes in our data (`location`) into regions and thereby we could represent values of interest on the regional map of Italy. ***… describe better the regional data!***

Table 1: Description of variables

Variable	Description
id	
timestamp	timestamp of the post
location	community name
title	title of the post
price	sale price of house
n_rooms	
floor	
mq	
n_bathrooms	
year_of_construction	
availability	
energy_class	
status	
heating	
has_garage	
has_terrace	
has_garden	
has_balcony	
has_fireplace	
has_alarm	
has_air_conditioning	
has_pool	
has_parking	
has_elevator	
is_furnished	

Exploratory question

1. What are the important factors (highly-correlated) for housing price in Italy?
    - Sub-question: Are there any differences in the set of important factors across different regions?

Maybe better: What are revelant predictors for housing price in Italy? Sub-question: How good are the predictions using a national model across different regions?

2. Is there a geological trend in housing price in Italy? (e.g., Is housing more/less expensive in northern Italy compared to Southern Italy?) **Only if we can figure out the location of the regions**

Preparation

Load packages & Import data

```
# load packages
library(tidyverse) # for wrangling data
library(skimr) # for skimming data
library(sf) # for spatial analysis
library(sp) # for spatial analysis

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

# additional data on Italy regional information
# comune level shape file
com2022 <- st_read("data/italy_shape_2022_files/Com01012022_g")[c("COD_REG","COD_CM", "COMUNE")]
```

```
Reading layer `Com01012022_g_WGS84' from data source
  `/Users/Kyuri1/Desktop/slv_assign1/slv_assignment1/data/italy_shape_2022_files/Com01012022_g'
  using driver `ESRI Shapefile'
Simple feature collection with 7904 features and 12 fields
Geometry type: MULTIPOLYGON
Dimension:     XY
Bounding box:  xmin: 313279.3 ymin: 3933846 xmax: 1312016 ymax: 5220292
Projected CRS: WGS 84 / UTM zone 32N
```

```
#  region level shape file
reg2022 <- st_read("data/italy_shape_2022_files/Reg01012022_g")
```

```
Reading layer `Reg01012022_g_WGS84' from data source
  `/Users/Kyuri1/Desktop/slv_assign1/slv_assignment1/data/italy_shape_2022_files/Reg01012022_g'
  using driver `ESRI Shapefile'
Simple feature collection with 20 features and 5 fields
Geometry type: MULTIPOLYGON
Dimension:     XY
Bounding box:  xmin: 313279.3 ymin: 3933846 xmax: 1312016 ymax: 5220292
Projected CRS: WGS 84 / UTM zone 32N
```

Preliminary analysis

We skim through our data using `skimr` package.

- We see our data consists of 223,409 rows and 25 columns.
- Given our question, we conclude that `timestamp` (timestamp of the post) and `title` (title of the post) are irrelevant and hence exclude them from the further analysis.
- In addition, we remove two columns that have only one unique value (i.e., `status` and `availibility`), as one cannot differentiate variables with respect to constant.
- 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).
- ***Remove id ? is it redundant?***
- ***What to do with the location??... there are 7023 unique locations. How do we want to deal with this?***
- Lastly, we observe that several variables 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 the following section.

```
# 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),
                     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)
```

Table 2: Data summary

Name	housing
Number of rows	223409
Number of columns	25
Column type frequency:	
character	6
numeric	19
Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate	empty	n_unique
location	0	1	0	7023
title	0	1	0	199305
availability	0	1	0	1
energy_class	679	1	0	12
status	0	1	0	1
heating	0	1	0	2

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	median	min	max	n_unique
id	0	1.00	111705.00	64492.77	111705	1	223409	223409
timestamp	0	1.00	1661135705.37	12645.42	1661135577	1661114079	1661158618	42238
price	39116	0.82	239938.98	7562062.01	135000	1	2147483647	2852
n_rooms	60323	0.73	3.50	0.99	3	2	5	4
floor	72365	0.68	1.82	1.13	2	1	52	22
mq	4034	0.98	158.63	128.68	117	1	999	976
n_bathrooms	14397	0.94	1.59	0.67	1	1	3	3
year_of_construction	10	1.00	1965.13	76.75	1980	1000	2209	389
has_garage	0	1.00	0.18	0.38	0	0	1	2
has_terrace	0	1.00	0.11	0.32	0	0	1	2
has_garden	0	1.00	0.17	0.37	0	0	1	2
has_balcony	0	1.00	0.10	0.30	0	0	1	2
has_fireplace	0	1.00	0.05	0.23	0	0	1	2
has_alarm	0	1.00	0.01	0.10	0	0	1	2
has_air_conditioning	0	1.00	0.30	0.46	0	0	1	2
has_pool	0	1.00	0.02	0.15	0	0	1	2
has_parking	0	1.00	0.02	0.12	0	0	1	2
has_elevator	0	1.00	0.06	0.23	0	0	1	2
is_furnished	0	1.00	0.08	0.27	0	0	1	2

```r
housing <- housing %>%
  # select variables that have more than one unique value
  select(where(~n_distinct(.) > 1)) %>%
  # remove timestamp and title
  select(-c(timestamp, title, id)) %>%
  # fix the data type
  mutate(across(c(starts_with("has"), is_furnished, heating, energy_class), factor))

# check the data again
my_skim(housing)
```

Table 5: Data summary

Name	housing
Number of rows	223409
Number of columns	20
Column type frequency:	
character	1
factor	13
numeric	6
Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate	empty	n_unique
location	0	1	0	7023

Variable type: factor

skim_variable	n_missing	complete_rate	ordered	n_unique	top_counts
energy_class	679	1	FALSE	12	g: 116713, f: 25680, e: 17320, a: 16033
heating	0	1	FALSE	2	aut: 200364, oth: 23045
has_garage	0	1	FALSE	2	0: 183044, 1: 40365
has_terrace	0	1	FALSE	2	0: 198444, 1: 24965
has_garden	0	1	FALSE	2	0: 186523, 1: 36886
has_balcony	0	1	FALSE	2	0: 200657, 1: 22752
has_fireplace	0	1	FALSE	2	0: 211298, 1: 12111
has_alarm	0	1	FALSE	2	0: 221370, 1: 2039
has_air_conditioning	0	1	FALSE	2	0: 156732, 1: 66677
has_pool	0	1	FALSE	2	0: 218571, 1: 4838
has_parking	0	1	FALSE	2	0: 219938, 1: 3471
has_elevator	0	1	FALSE	2	0: 210650, 1: 12759
is_furnished	0	1	FALSE	2	0: 206091, 1: 17318

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
year_of_construction	10	1.00	1965.1	76.75	1980	1000	2209	389

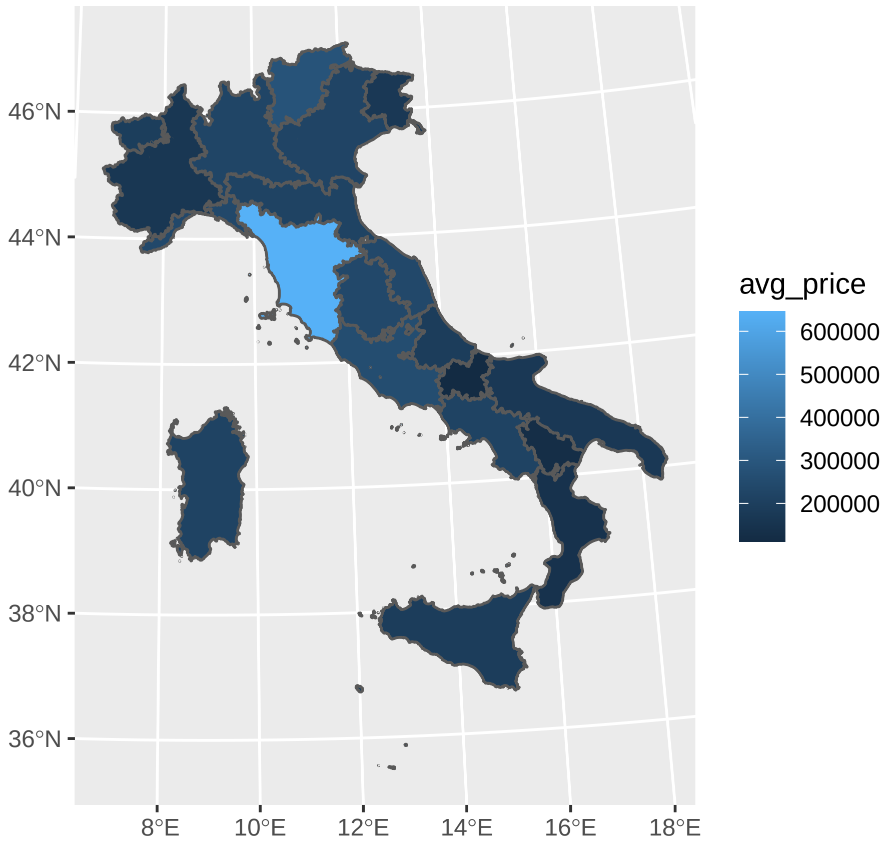
Add regional-level geographical information

```r
# add regional info to the comune level data
dat_with_geo <- st_join(com2022, reg2022, by="COD_REG") %>%
  # we don't need comune level geometry info
  select(-c(COD_REG.x)) %>%
  # lower the case of COMUNEs to match them with the comunes in our data
  mutate(location = tolower(COMUNE)) %>%
  # rename the COD_REG column
  rename(COD_REG = COD_REG.y) %>%
  # merge with the housing data
  left_join(housing, by = "location") %>%
  # relocate geometry column
  relocate(geometry, .after=Shape_Area)
```

Try out plotting the map for the average price per region

```r
# get average price per regions
avg_price_region <- dat_with_geo %>%
  group_by(DEN_REG) %>%
  summarize(avg_price = mean(price, na.rm=T))

# not working in one-go
ggplot(avg_price_region) +
  geom_sf(aes(fill = avg_price))
```



Imputation? Or plot the missingness to see if there is any pattern for NA vs non-NA

Summary statistics

Descriptive plots

Exploratory plots

Conclusion

Reducing/Modifying the data set

We need to modify the data to be able to work with it. For the visualization we have to create an "age" variable and exclude (?) housing units with a negative age (< -...). Furthermore, For the second step, identifying the predictors, we need to exclude variables of class "character" and factors with more than 2 (3?) levels.

```
#
# 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(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)
```

Summary statistics

Visualization

Exploring predictors for price

Backward selection

```
#housing_red %>%  step(lm(price ~., .), direction = backward)
```

Forward selection

```
#housing_red %>%  select(-energy_class) %>%  step(lm(price ~., .), direction = forward)
```