SLV Assignment1

Daniel Anadria Kyuri Park Ernst-Paul Swens Emilia Loescher

September 28, 2022

Introduction

The Dataset

We explore a dataset from kaggle, which contains information about the housing market in Italy. The data were scraped from one of the most relevant housing sales websites in Italy during the month of *August 2022*. The data consist of more than 223,000 sales posts spread over 7,023 (89% of the total 7,904) Italian municipalities (*comuni*). Some of the entries were removed in dataset construction due to translation limitations (e.g., extended text-based description, specific url of the post).

In order to plot the statistics of interest to maps of Italy, we use the regional and provincial shape files which we obtained from [the Italian National Institute of Statistics] (https://www.istat.it/it/archivio/222527). These files contain the regional and provincial coding and geographical shape information which can be used cluster the comuni in our data (`location`) into (107) provinces and (20) regions.

For each sale, the dataset contains the following variables:

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 Euros
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)

Variable	Description
has_elevator	Elevator present: yes (1), no (0)
is_furnished	Furniture present: yes (1), no (0)

Exploratory Questions

We focus on four exploratory questions concerning housing prices in Italy.

Firstly, we explore if there are a geographical trends in the mean and median 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 mean regional variances of housing prices between the different provinces and regions.

Thirdly, we explore if there is a correlation between the missingness of housing price and other variables.

Fourthly, we identify the most important predictors of 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) # for wrangling data
library(skimr) # for skimming data
library(sf) # for spatial analysis
library(sp) # for spatial analysis
library(ggplot2) # for plotting
library(fuzzyjoin) # for joining on not-exact matches

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

# import shape files
munic_2022 <- st_read("data/italy_shape_2022_files/Com01012022_g")[c("COD_REG","COD_PROV", "COMUNE")]  # m
```

```
Reading layer `Com01012022_g_WGS84' from data source
  `C:\Users\emsul\OneDrive\Documents\GitHub\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
```

```
prov_2022 <- st_read("data/italy_shape_2022_files/ProvCM01012022_g") # province level
```

```
Reading layer `ProvCM01012022_g_WGS84' from data source
  `C:\Users\emsul\OneDrive\Documents\GitHub\SLV_assignment1\data\italy_shape_2022_files\ProvCM01012022_g'
  using driver `ESRI Shapefile'
Simple feature collection with 107 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
```

```
reg_2022 <- st_read("data/italy_shape_2022_files/Reg01012022_g") # region level
```

```
Reading layer `Reg01012022_g_WGS84' from data source
  `C:\Users\emsul\OneDrive\Documents\GitHub\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 the `skimr` package. This summary of the raw data set can be found in the Appendix . The original data consist 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 dataset 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, we observe that types 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).

We create a new variable `age` which contains the age of the property in 2022 by subtracting the `year_of_construction` from 2022. In the original dataset, there are some unreasonable years of construction (e.g. 2209). Thus, we set `age` to NA for any `year_of_construction` further in the future from 2026. Some property may be sold before construction is completed, but we deem it unlikely for properties whose `year_of_construction` is more than 4 years removed from the present year.

```
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 of our data.

```
# 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
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 factor variables, we see that there are no missing values. The ratio 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_bathrooms`). We will discuss how we want to deal with this in section .

Discuss price descriptives, many outliers, deal with them include other plots, density, boxplot, etc. make the plots color-blind friendly name chunks, use cache = T on the final version # Exploratory Analysis

Question 1: Geographical Differences in the Mean and Median Housing Price

Each sale in our dataset is assigned to one of 7023 municipalities. In order to create plots which visualize the differences in average housing prices across Italy, we assign each municipality (*comune*) to its corresponding province (*provincia*) and region (*regione*). We use the data from the *Italian National Institute of Statistics* (*ISTAT*) to append the province and region information to every observed municipality in our dataset.

Preparation

At the beginning of the assignment, we loaded the *ISTAT* shape files. These files are useful for two reasons. First, they contain the list of all Italian municipalities, their respective provinces and regions. Therefore, we can use this data to append our original dataset with additional location indicators. Second, they contain the shapes of Italy divided into provinces and regions. This is particularly useful for creating map plots using `ggplot2`.

For completeness of our dataset, we append the province and region information. We use fuzzy matching for inexact matches as we found that there were some minor inconsistencies in how the municipalities were named in our dataset as opposed to their names in the ISTAT shape files. The result of the following chunk of code is that all the municipalities are assigned their regions and provinces.

```
housing_red <- stringdist_left_join(housing_red, munic_2022,by = c("location" = "COMUNE"), distance_col =
  group_by(location) %>% slice_min(distance) %>%
  select(-geometry,-distance) %>%
  left_join(., as.data.frame(reg_2022[,c("COD_REG","DEN_REG")])) %>%
  select(-geometry, -COMUNE) %>%
  left_join(., as.data.frame(prov_2022[,c("DEN_UTS", "COD_PROV")], by = "COD_PROV")) %>%
  select(-geometry, -COD_REG, -COD_PROV) %>%
  rename(., "region" = "DEN_REG", "province" = "DEN_UTS") %>%
  relocate(c(region, province), .after=location)
```

To answer the first exploratory question, we aggregate our data on two levels: 1) regional and 2) provincial level by computing two aggregate statistics: 1) the mean housing price and 2) the variance in housing price on the two respective levels. This yields two datasets, one per aggregation level. To each, we attach geometric information needed for plotting and convert it to an `sf` object which is a requirement for plotting maps.

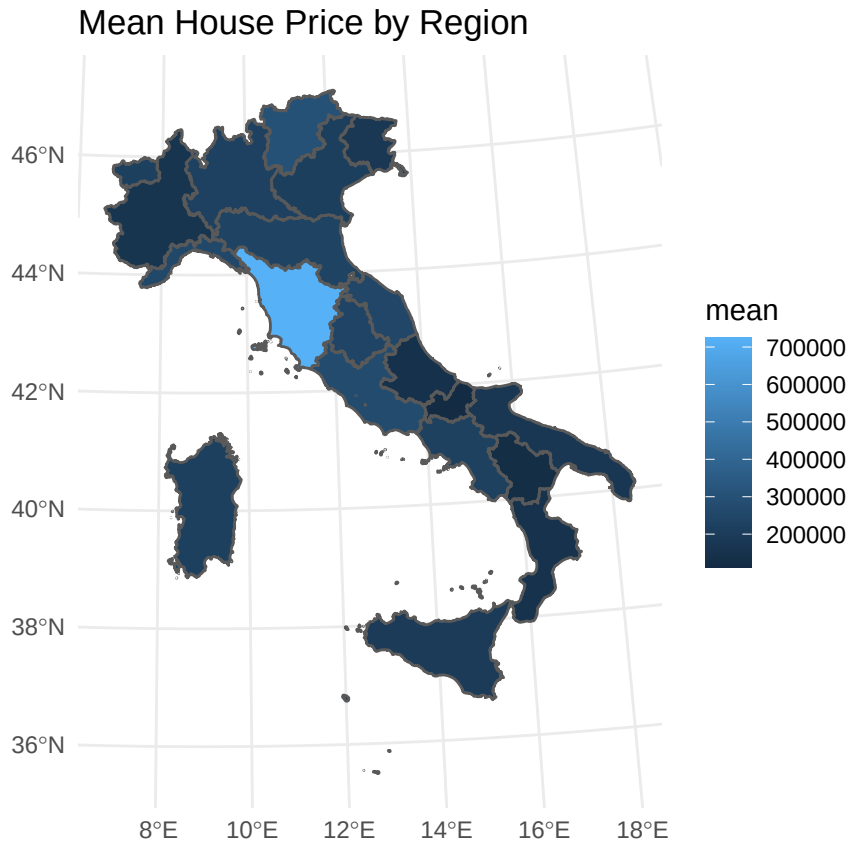
```
price_by_reg <- housing_red %>% group_by(region) %>%
  summarize(mean = mean(price, na.rm=T), median = median(price, na.rm=T), variance = var(price, na.rm=T), r
```

```
price_by_prov <- housing_red %>% group_by(province) %>%
  summarize(mean = mean(price, na.rm=T), median = median(price, na.rm=T), variance = var(price, na.rm=T), r
```

Plot

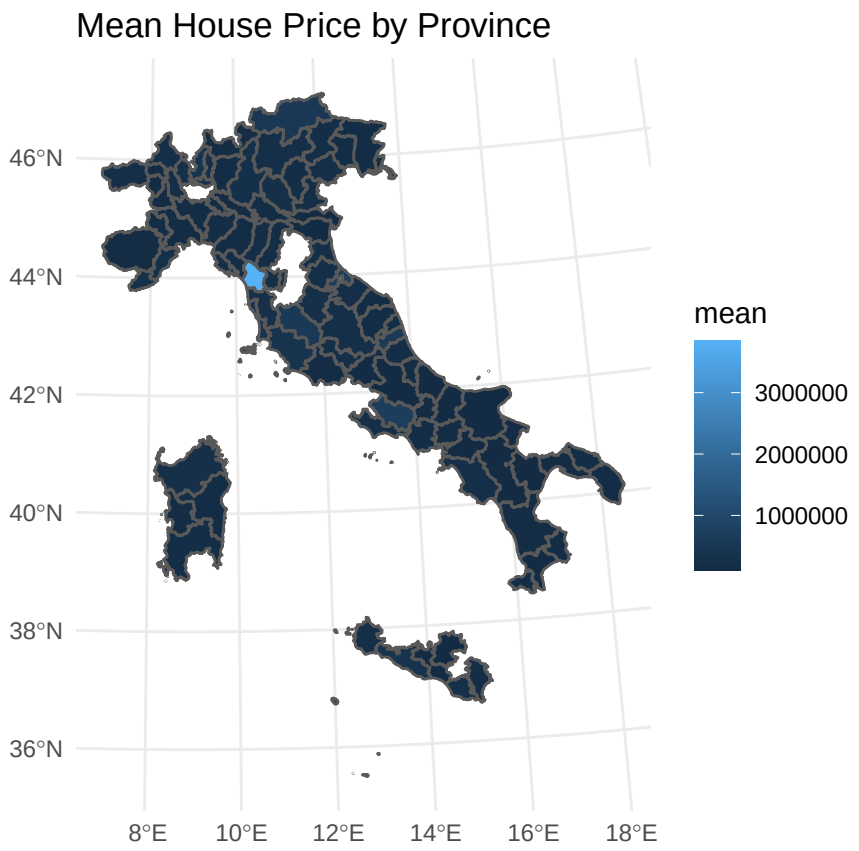
We inspect the mean housing price by region.

```
ggplot(price_by_reg) +
  geom_sf(aes(fill = mean))+
  ggtitle("Mean House Price by Region") +
  theme_minimal()
```



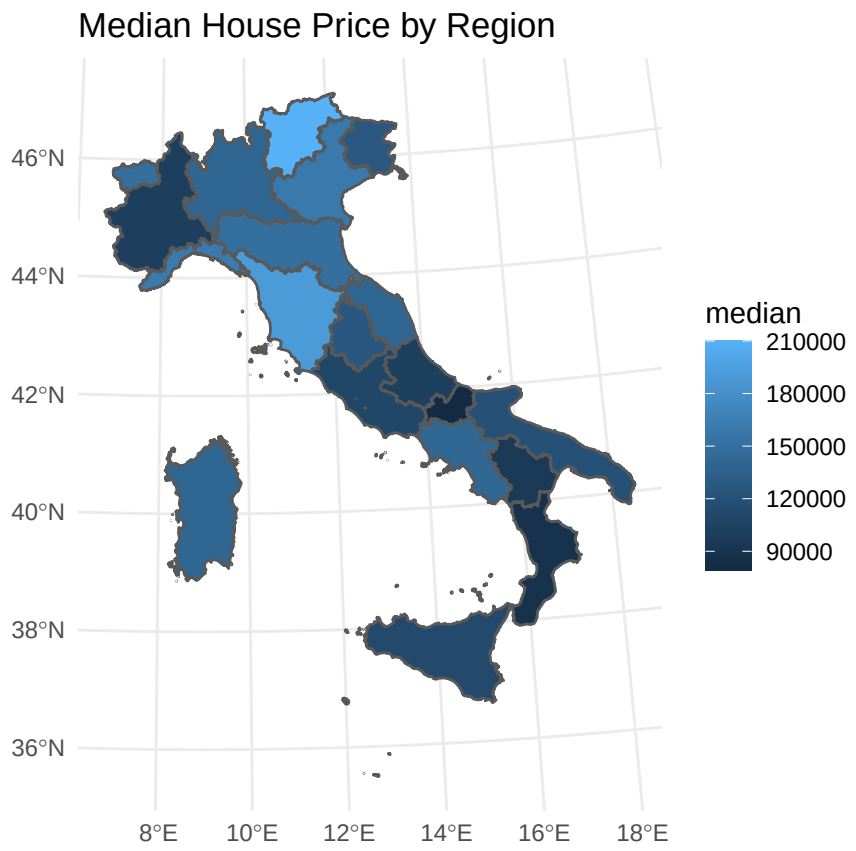
We inspect the average housing price by province.

```
ggplot(price_by_prov) +
  geom_sf(aes(fill = mean))+
  ggtitle("Mean House Price by Province") +
  theme_minimal()
```



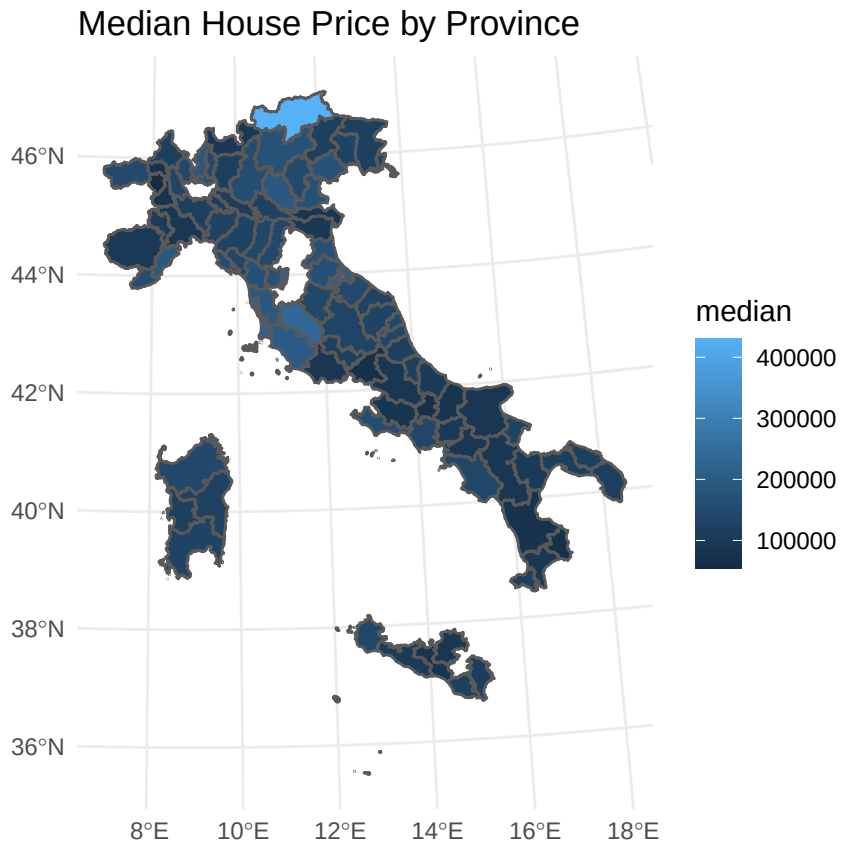
We inspect the median housing price by region.

```
ggplot(price_by_reg) +
  geom_sf(aes(fill = median))+
  ggtitle("Median House Price by Region") +
  theme_minimal()
```



We inspect the median housing price by province.

```
ggplot(price_by_prov) +
  geom_sf(aes(fill = median))+
  ggtitle("Median House Price by Province") +
  theme_minimal()
```



Conclusion

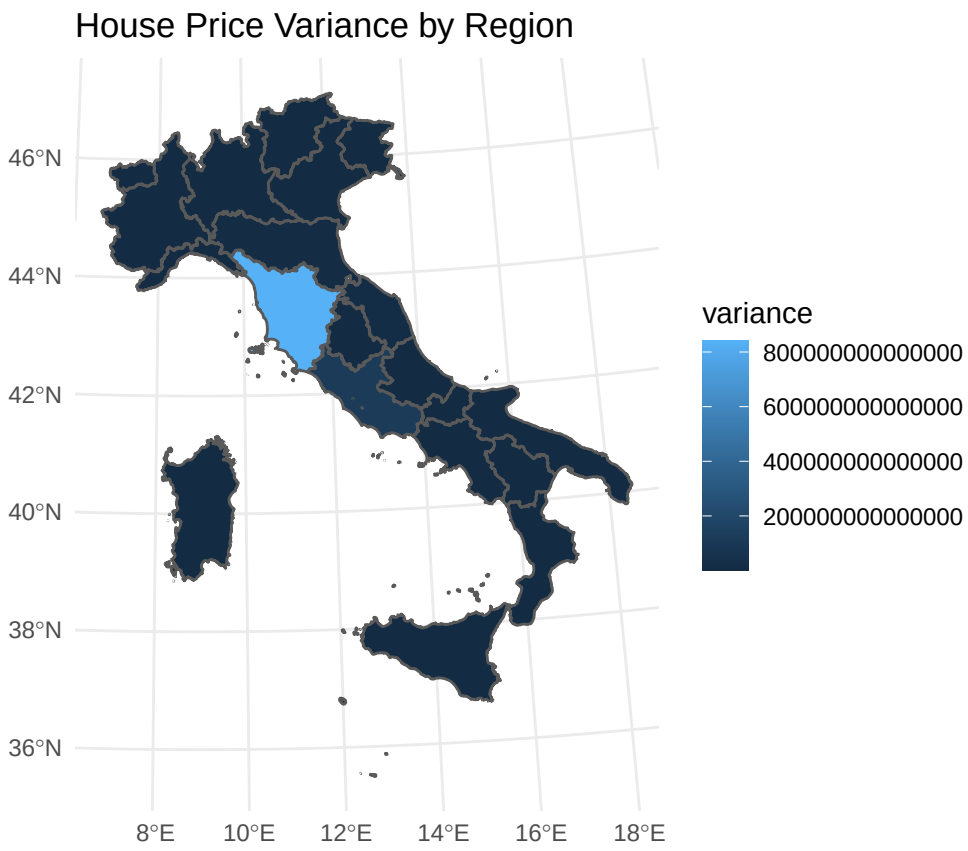
Question 2: Geographical Differences in Variance and Median Absolute Deviation of Housing Prices

Preparation

Plot

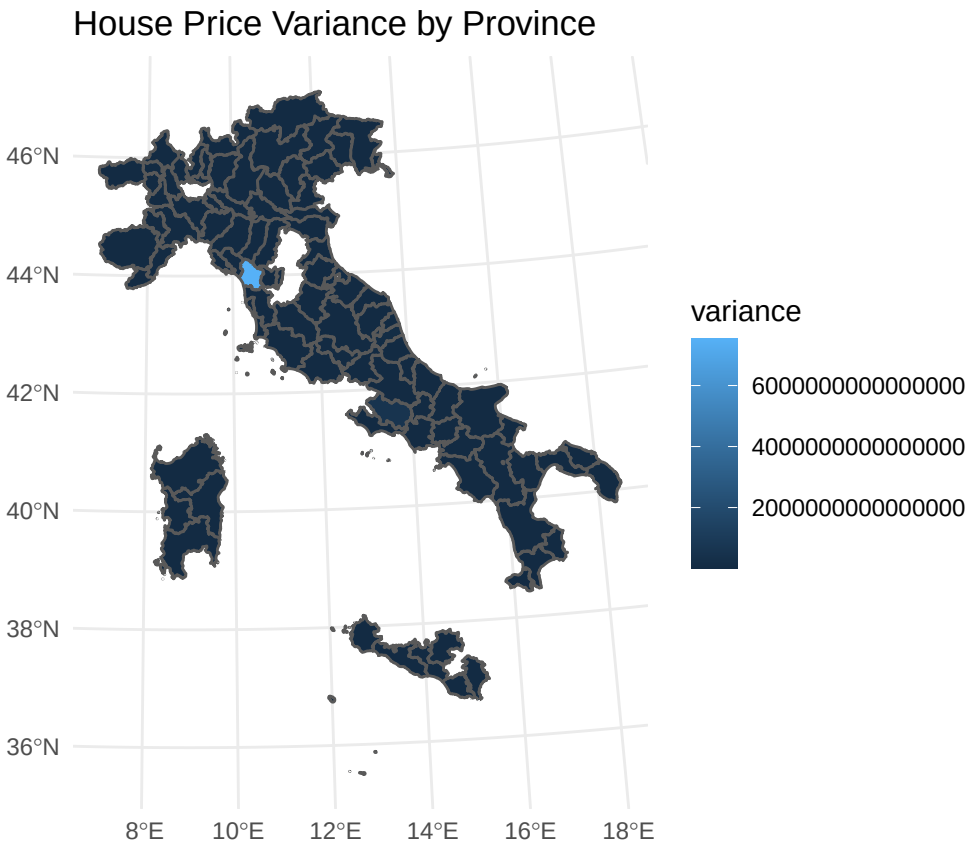
We inspect the housing price variance by region.

```
ggplot(price_by_reg) +
  geom_sf(aes(fill = variance))+
  ggtitle("House Price Variance by Region") +
  theme_minimal()
```



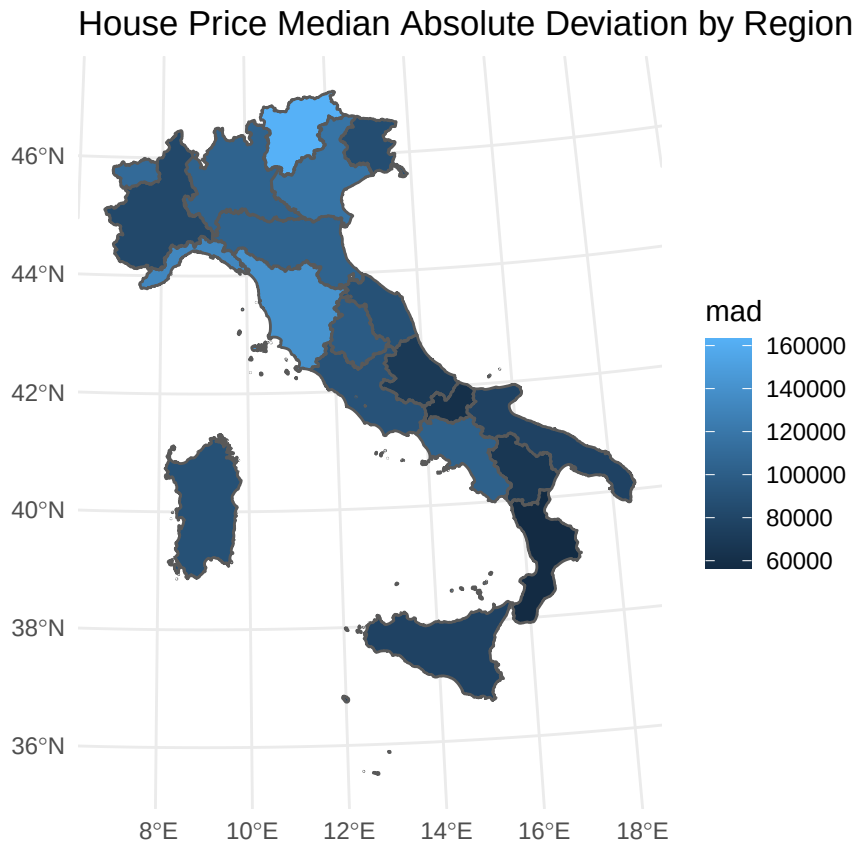
We inspect the housing price variance by province.

```
ggplot(price_by_prov) +
  geom_sf(aes(fill = variance))+
  ggtitle("House Price Variance by Province") +
  theme_minimal()
```



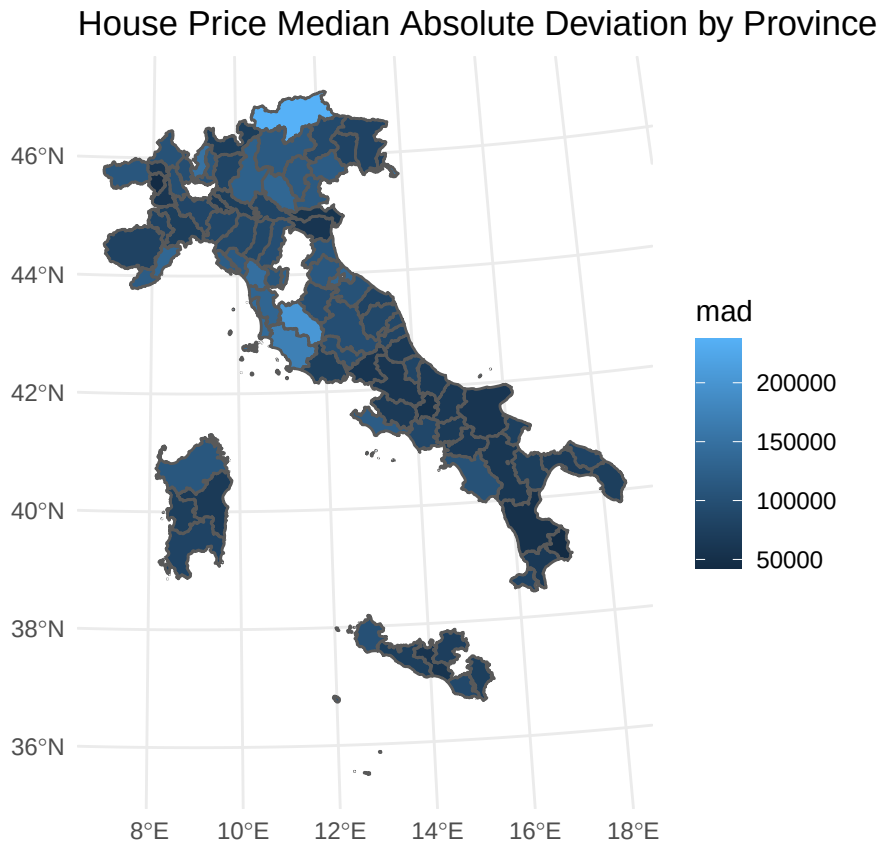
We inspect the median absolute deviation of housing price by region.

```
ggplot(price_by_reg) +
  geom_sf(aes(fill = mad))+
  ggtitle("House Price Median Absolute Deviation by Region") +
  theme_minimal()
```



We inspect the median absolute deviation of housing price by province.

```
ggplot(price_by_prov) +
  geom_sf(aes(fill = mad))+
  ggtitle("House Price Median Absolute Deviation by Province") +
  theme_minimal()
```



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

Appendix

```
my_skim(housing)
```

Table 6: 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

skim_variable	n_missing	complete_rate	mean	sd	median	min	max	n_unique
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