term paper - statistics

Hannah Behrens

Abstract

Due to Germany's history a disequilibrium between its western and eastern states was a common topic for which a solidarity tax was payed a long time. Are there still any financial differences of the human's life in the west compared to the east? In order to answer this question or to get an idea, a housing data set will be analysed concerning the geography and time.

Keywords: Statistics, Regression, Forecasting

Meschede

2nd February 2022

Wir geben Impulse

Declaration of Authorship

I hereby declare that the thesis submitted is my own unaided work. All direct or indirect sources used are acknowledged as references. This paper was not previously presented to another examination board and has not been published.

I am aware that the thesis in digital form can be examined for the use of unauthorized aid and in order to determine whether the thesis as a whole or parts incorporated in it may be deemed as plagiarism. For the comparison of my work with existing sources I agree that it shall be entered in a database where it shall also remain after examination, to enable comparison with future theses submitted.

Meschede, 2nd February 2022.

Hannah Behrens
MatNr: xx
Email: behrens.hannah@fh-swf.de
Corresponding Author

This Declaration must be signed by all listed Authors to be valid.

Checklist

I hereby declare . . .

- ☐ List item 1 goes here.
- ☐ List item 2 goes here.
    - ☐ Sublist item 1 goes here.
    - ☐ Sublist item 2 goes here.
- ☐ List item 3 goes here
- ☐ List item 4 goes here.

All list elements must be checked.

Contents

1 Introduction

1.1 Context

Germany's history: separation of Germany in West and East (DDR 1949-1990)

missing financial support/investigations in the eastern states concerning e.g. its infrastructure/housing

→ as a cause: different living conditions

Since 1990: adaptation of *new states* (East):

- Brandenburg
- Sachsen
- Sachsen-Anhalt
- Thüringen
- Mecklenburg-Vorpommern

→ reunion: Federal Republic of Germany

After Germany's reunion, a solidarity tax has been payed in order to balance the drawbacks in East-Germany.

1.2 Research Question

Hypothesis:

Despite the solidarity tax there is still a significant difference concerning the housing prices in the west vs. east.

Or in other words, the research question is: Does a significant difference concerning the housing prices between the east and west of Germany still exist which can be possibly seen as a remnant of Germany's separation?

Therefore, the temporal development of houseprices are of interest.

- assumption: no increase in housing prices with increasing years, i.e. no inflation, will be considered

- increases in value due to e.g. reduction of living space are not considered

1.3 Paper Structure

In order to answer the research question and to examine the mentioned hypotheses respectively, the different states have to be marked: This means, which state belongs to the former eastern states and which is mapped to the west?

Therefore, some variables such as **id**, **price**, **year_construction** and **state** will be picked off. Furthermore, the German housing data set has to be modified by introducing a further (categorical) variable which is called **location** and which gives information about west/east-affiliation.

In order to examine if there is a significant difference between the housing prices in west and east germany, a t-test will be performed.

2 Data

```r
# Load data
x <- read.csv("https://bchwtz.github.io/bchwtz-stat/data/2021_fhswf_stat_housing.csv")


# Structure of selected variables of the housing-data set
head(data.frame(x$id, x$price, x$year_construction, x$state))
```

```
##   x.id x.price x.year_construction              x.state
## 1    0  498000                2005 Baden-WÃ¼rttemberg
## 2    1  495000                1994 Baden-WÃ¼rttemberg
## 3    2  749000                2013 Baden-WÃ¼rttemberg
## 4    3  259000                1900 Baden-WÃ¼rttemberg
## 5    4  469000                1968 Baden-WÃ¼rttemberg
## 6    5 1400000                1969 Baden-WÃ¼rttemberg
```

The file consists of 10552 observations and 25 variables.

3 Methodology

3.1 Data manipulation

```r
# remove the row in which no state is listed:
x2<-x[-which(x$state==""),]


## Data manipulation:
x2$state<-gsub("Ã¼", "ue", x2$state)

east_states<-c("Brandenburg", "Sachsen", "Sachsen-Anhalt",
               "Thueringen", "Mecklenburg-Vorpommern")
```

```r
# introducing the categorical variable "location":
x2$location<-0
x2$location[which(x2$state %in% east_states)]<-"east"
x2$location[which(!x2$state %in% east_states)]<-"west"


x3<-x2[which(x2$year_construction>=1949),]


# house prices of 0 | 1 make no sense
# --> remove the corresponding observations
x3<-x3[which(x3$price>1),]


x3_extract<-data.frame(x3$id, x3$price, x3$year_construction,
                       x3$state, x3$location)
colnames(x3_extract)<-c("id", "price", "yc", "state", "location")


# separate into west and east
x3_west<-x3_extract[which(x3_extract$location=="west"),]
x3_east<-x3_extract[which(x3_extract$location=="east"),]

## annual housing prices
# in the west
x3_west_mean<-x3_west %>%
  group_by(yc) %>%
  summarise(m = mean(price))


# in the east
x3_east_mean<-x3_east %>%
  group_by(yc) %>%
  summarise(m = mean(price))


comb<-x3_east_mean %>%
  left_join(x3_west_mean, by="yc", suffix=c("_east", "_west"))
```

Table 1: *Mean and standard deviation of annual housing prices after applying left join.*

	prices_east	prices_west
mean_x	381167.5	639777.2
sd_x	209075.0	145705.5

In total, 6991 observations exist. Most of them are part of **West-Germany** (6135). Only 856 observations belong to **East-Germany**.

3.2 Summary of prices

```
## $east_summary
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   12000  220000  359000  444125  508183 4900000
##
## $west_summary
##     Min.  1st Qu.   Median     Mean  3rd Qu.      Max.
##     7900   330000   482000   629743   725000 13000000
```

```
# annual housing prices
comb2<-comb[,2:3]
colnames(comb2)<-c("mean_annual_east", "mean_annual_west")
lapply(comb2, summary)
```

```
## $mean_annual_east
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   89000  226167  342731  381168  527667 1031457
##
## $mean_annual_west
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  348879  550749  599584  639777  727734 1183153
```
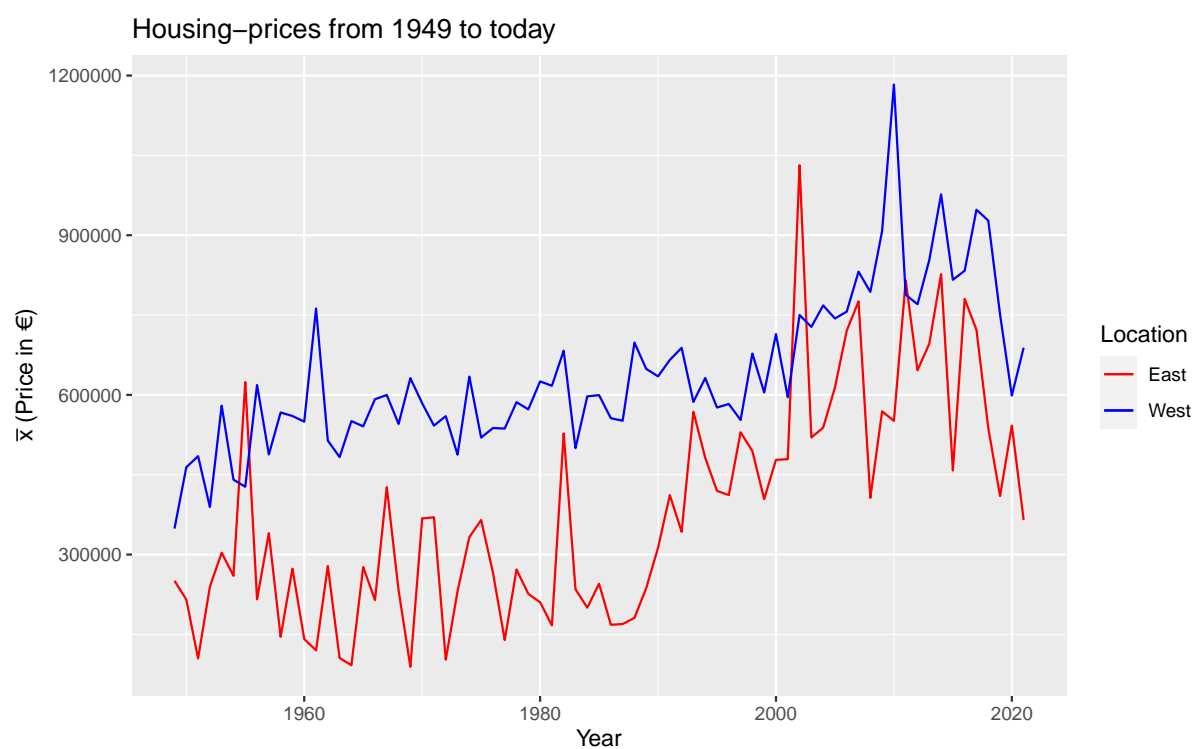
3.3 Plot the data

$\rightarrow$ obvious differences can be seen from Figures 1 and 2 as well as from Table 1.

Housing–prices from 1949 to today
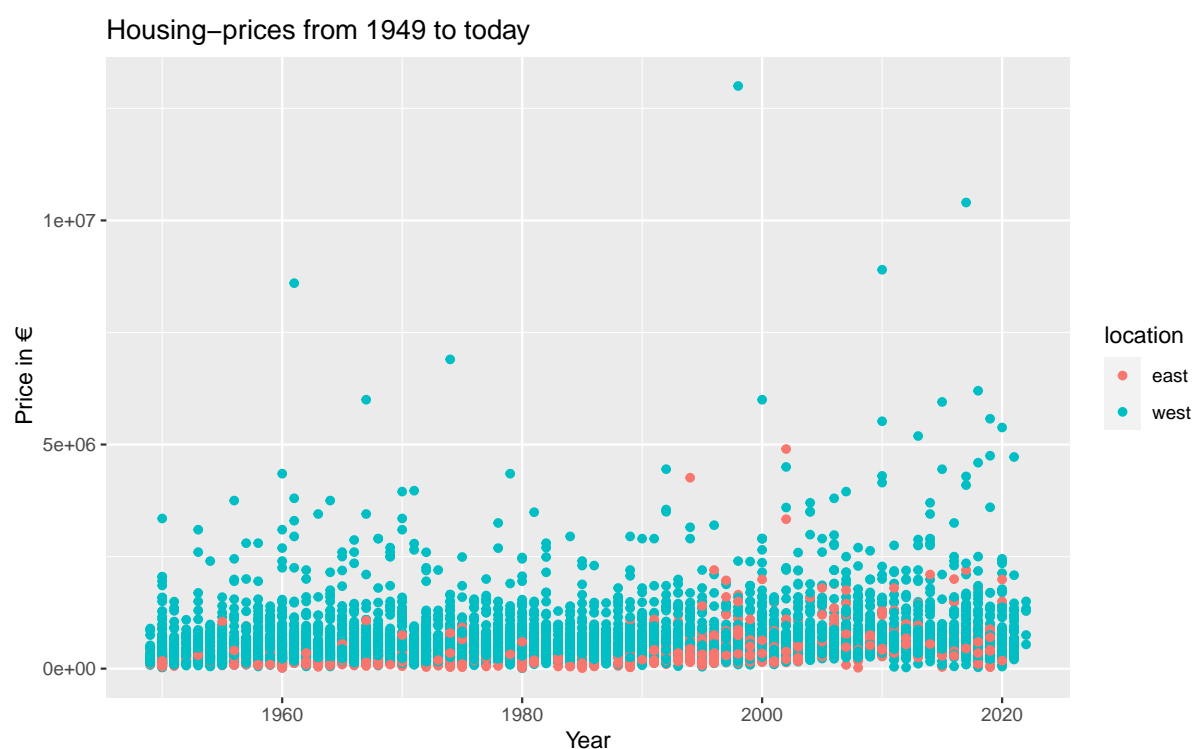
Figure 1: *Raw data of housing prices in the west and east of Germany.*

Housing–prices from 1949 to today

Figure 2: *Annual housing prices.*

4 Analysis

t-test

H_0: no difference concerning the housing prices in West- and East-Germany

H_A: the housing prices in West-Germany are *really/much higher* compared to those in East-Germany or the other way around: the housing-prices in East-Germany are *much lower* than those in West-Germany, so there is a *real/significant* difference

$\alpha = 0.05$ (5 %) $\rightarrow \frac{\alpha}{2} = 0.025$

```
# Two-sided t-test based on the annual housing prices: 1949-today
tt<-t.test(x=comb$m_east, y=comb$m_west, alternative = "two.sided")
tt
```

```
##
##  Welch Two Sample t-test
##
## data:  comb$m_east and comb$m_west
## t = -8.6705, df = 128.59, p-value = 1.597e-14
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -317624.0 -199595.3
## sample estimates:
## mean of x mean of y
##  381167.5  639777.2
```

```
comb_1990<-comb[which(comb$yc>=1990),]
```

```
# Two-sided t-test based on the annual housing prices: 1990-today
tt_1990<-t.test(x=comb_1990$m_east, y=comb_1990$m_west,
                alternative = "two.sided")
tt_1990
```

```
##
##  Welch Two Sample t-test
##
## data:  comb_1990$m_east and comb_1990$m_west
```

```
## t = -4.9389, df = 60.296, p-value = 6.554e-06
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   -266279.2 -112774.7
## sample estimates:
## mean of x mean of y
##   558161.4  747688.3
```

1.) $|$t-value$|$, i.e. 8.6705 and 4.9389 are both $> t_{n=100,0.975} = 1.984$

2.) $\frac{\alpha}{2} < 1.597 \cdot 10^{-14}$, $\frac{\alpha}{2} < 6.554 \cdot 10^{-6}$

$\rightarrow$ rejection of H_0, i.e. *significant* difference in annual housing prices

5 Discussion (manually added)

For a *detailed* analysis further factors such as the following ones have to be taken into account:

- So far, no inflation which can vary in the diverse (western/eastern) states has been considered. This can be an explanation for the high difference in the means.

- the different size of the houses, differences in any governmental grants which result in differences concerning the house prices, the *geographic* distributions of houses in villages/cities in the west/east have to be considered and further ones which may have an impact on the presented results.

- disequilibrium between the number of observations in the west vs. east of Germany $\rightarrow \frac{6135 \ obs.}{11 \ states} = 557.7273 \ \frac{obs.}{state}$ vs. $\frac{856 \ obs.}{5 \ states} = 171.2 \ \frac{obs.}{state}$

Instead of using the yearly means in order to compare the housing prices in the west and east of Germany, the median, which is a robust parameter, is possibly more suitable.

6 Conclusion

In the analysed data set, there is a significant difference in the annual housing prices between West- and East-Germany in the range from 1949 to today as well as after 1990. Consequently, the houses in West-Germany are still much more expensive than those in the east. Nevertheless, there are some more criteria which have not been considered so far.

```
# Language Options

german: true

lang: de-de
```

7 Outlook - further data sets (added manually)

7.1 What to do in the future?

- How can the large increases in housing prices in the east of Germany in 1955 as well as in 2002 be possibly explained?

- taking into account the mentioned factors

7.2 Further data sets for other examinations

- R package `housingData` → `housing` data set

- R package `fpp3`: different data sets such as `prices`, `boston_marathon`

overview of diverse R packages with data sets

7.3 Further information - literature:

- How to write texts in R Markdown

Technical Appendix

```
Sys.time()
```

```
## [1] "2022-02-02 12:24:41 CET"
```

```
sessionInfo()
```

```
## R version 4.1.2 (2021-11-01)
## Platform: x86_64-w64-mingw32/x64 (64-bit)
## Running under: Windows 10 x64 (build 19044)
##
## Matrix products: default
##
## locale:
## [1] LC_COLLATE=German_Germany.1252  LC_CTYPE=German_Germany.1252
## [3] LC_MONETARY=German_Germany.1252 LC_NUMERIC=C
## [5] LC_TIME=German_Germany.1252
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
## [1] dplyr_1.0.7   knitr_1.37    ggplot2_3.3.5
##
## loaded via a namespace (and not attached):
##  [1] latex2exp_0.9.0  highr_0.9         pillar_1.6.4     compiler_4.1.2
##  [5] tools_4.1.2      digest_0.6.29     evaluate_0.14    lifecycle_1.0.1
##  [9] tibble_3.1.6     gtable_0.3.0     pkgconfig_2.0.3  rlang_0.4.12
## [13] DBI_1.1.2        yaml_2.2.1       xfun_0.29        fastmap_1.1.0
## [17] withr_2.4.3      stringr_1.4.0    generics_0.1.1   vctrs_0.3.8
## [21] grid_4.1.2       tidyselect_1.1.1 glue_1.6.0       R6_2.5.1
## [25] fansi_1.0.2      rmarkdown_2.11   bookdown_0.24    farver_2.1.0
## [29] purrr_0.3.4      magrittr_2.0.1   scales_1.1.1     ellipsis_0.3.2
## [33] htmltools_0.5.2  assertthat_0.2.1 colorspace_2.0-2 fhswf_0.0.1
## [37] labeling_0.4.2   utf8_1.2.2       stringi_1.7.6    munsell_0.5.0
## [41] crayon_1.4.2
```