Template

Mark Barkalov (2782988) Joël Dijkstra (2854215) Shabana Esmati (2868438)
Jasper Hoogendoorn (2856970) Giovanni Koek (2857652)
Robin Kolmus (2862402) Elijah Kruize (2868400) Tim van der Laan (2854528)

2025-06-23

Set-up your environment

```
library(ggplot2)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(tidyr)
library(readr)
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v forcats   1.0.0     v stringr   1.5.1
## v lubridate 1.9.4     v tibble    3.2.1
## v purrr     1.0.4

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(cbsodataR)
library(sf)
```

```
## Linking to GEOS 3.13.1, GDAL 3.10.2, PROJ 9.5.1; sf_use_s2() is TRUE
```

Title Page

Include your names:

Mark Barkalov (2782988)

Joël Dijkstra (2854215)

Shabana Esmati (2868438)

Jasper Hoogendoorn (2856970)

Giovanni Koek (2857652)

Robin Kolmus (2862402)

Elijah Kruize (2868400)

Tim van der Laan (2854528)

Include the tutorial group number:

Group 4

Include your tutorial lecturer’s name:

Miss. C. Schouwenaar

Part 1 - Identify a Social Problem

Use APA referencing throughout your document.

1.1 Describe the Social Problem

Include the following:

- Why is this relevant?
- ...

Part 2 - Data Sourcing

2.1 Load in the data

Preferably from a URL, but if not, make sure to download the data and store it in a shared location that you can load the data in from. Do not store the data in a folder you include in the Github repository!

```r
data2 <- read.csv("wozdata.csv")
data <- read.csv("inkdata.csv")
zeefle <- read.csv2("zeefle.csv")
provincie_2020 <- cbs_get_sf("provincie", 2020)
df_2023 <- read_csv2("bedrijvenNL2023.csv", show_col_types=FALSE)
```

```
## i Using "’,’" as decimal and "’.’" as grouping mark. Use ‘read_delim()‘ for more control.
```

```
df_2019 <- read_csv2("bedrijvenNL2019.csv", show_col_types=FALSE)
```

```
## i Using "','" as decimal and "'.'" as grouping mark. Use `read_delim()` for more control.
```

2.2 Provide a short summary of the dataset(s)

Data2: Data2 is a dataset from CBS that mentions the average house prices of houses over the years 2019-2023. This dataset gives us data for the Netherlands, per city and per province. The dataset gives average house prices of houses you can buy and the value of houses that people rent out. We used this dataset to find out what the changes in house prices were over the years.

Data and zeefle: Data is a dataset from CBS that mentions the average income of groups with different hour contracts over the years 2019-2023 in the Netherlands. It also gives the total average of personal income which we used. We also have the dataset zeefle which is the same as dataset Data but it mentions the average income of Zeeland and Flevoland which we used for the temporal variation.

provincie_2020: Dataset provincie_2020 is a dataset from the package cbsodataR that we used to make a map for the spatial variation. It provides the coordinates of every province.

df_2023 and df_2019: These datasets are from the same dataset from CBS and they give the same information. They provide us with data about the amount of businesses per province but one in the year 2023 and the other in 2019. We used this for our subgroup analysis.

2.3 Describe the type of variables included

The datasets makes use of multiple variables, including: WOZ Value in the Netherlands in each province from 2019 - 2023, Average income of working population in the Netherlands, Zeeland and Flevoland from 2019-2023 and Number of businesses per province in 2019 and 2023

These variables were all from the CBS so these are all administrative sources.

Think of things like:

- Do the variables contain health information or SES information?
- Have they been measured by interviewing individuals or is the data coming from administrative sources?

For the sake of this example, I will continue with the assignment…

Part 3 - Quantifying

3.1 Data cleaning

Say we want to include only larger distances (above 2) in our dataset, we can filter for this.

Please use a separate 'R block' of code for each type of cleaning. So, e.g. one for missing values, a new one for removing unnecessary variables etc.

3.2 Generate necessary variables

Variable 1: House prices to average personal income ratio. We used this to see how much yearly income is needed to fully purchase the average house in each province and the country as a whole. We did this for the Netherlands and for the provinces Zeeland and Flevoland. For the provinces we used the average personal income in their respective province and for the Netherlands we used the average personal income of the Netherlands as a whole.

```
new_row <- data[5, ] / data[3,]  # Element-wise division of average Netherlands house prices divided by
```

```
new_row_fle <- data[11, ] / data[19,]  # Element-wise division of Flevoland's houseprices divided by th
new_row_zee <- data[16, ] / data[20,]  # Element-wise division of Zeeland's houseprices divided by the
```

Variable 2(dif): This variable is the house prices of 2023 divided by the house prices in 2019. This way we can see which city had the most increase during the years. We could find the city with the most increase and the city with the least amount of increase and find out if both cities increased more than the income. We can also use this in the subgroup analysis to find if an increase in businesses has a direct effect on the house prices.

```
dif <- c() # give variable dif the value of a vector

for(i in 1:nrow(data2))
{
  dif[i] = data2$"2023"[i] / data2$"2019"[i] # divides the house prices in 2023 by the houseprices in 2
}
```

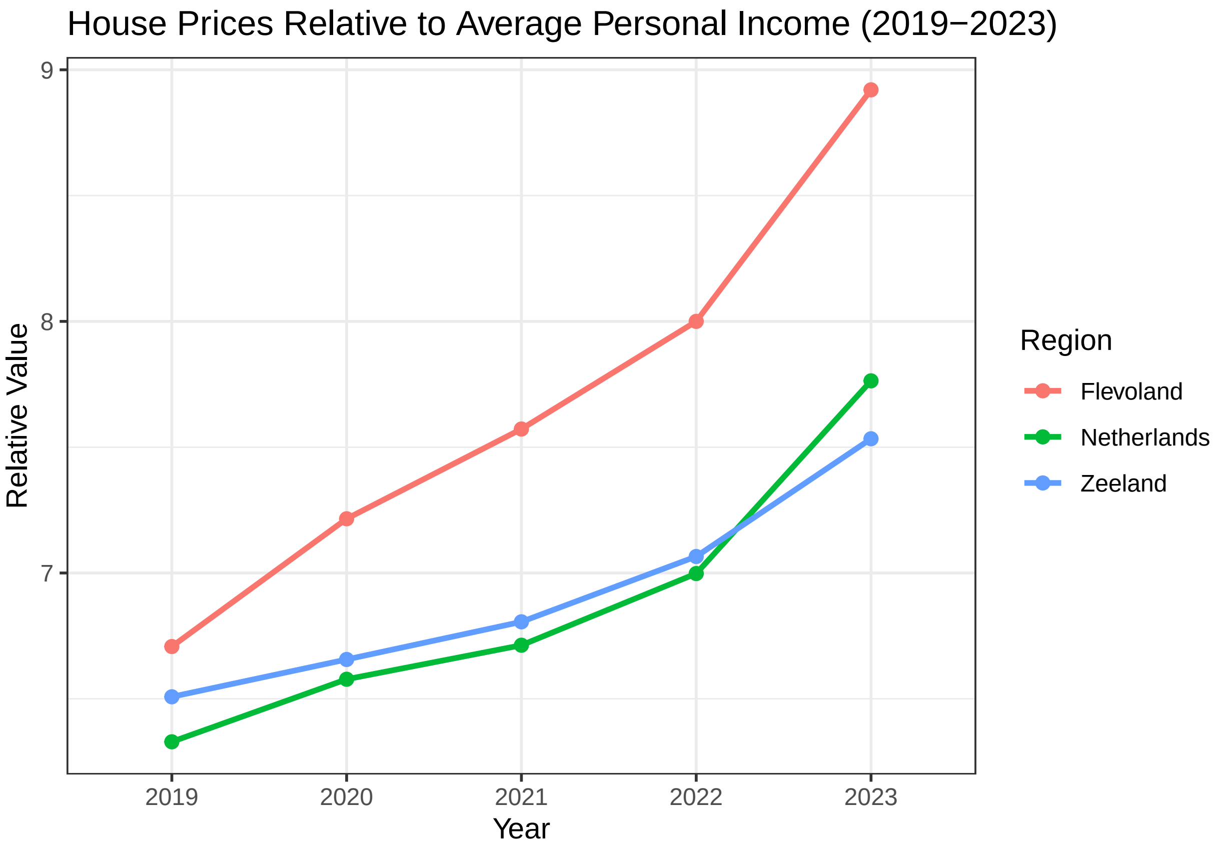
Variable 3(busi): This variable is the amount of businesses in each province in 2023 divided by the businesses in 2019. This way we can compare the change in house prices with the change in businesses, to see if an increase in businesses has a direct impact on the house prices in the Netherlands.

```
busi <- c() # give variable busi the value of a vector

for(i in 1:nrow(df_2019))
{
  busi[i] = df_2019$"Vestigingen (Aantal).x"[i] / df_2019$"Vestigingen (Aantal).y"[i] # Element divisio
}
```

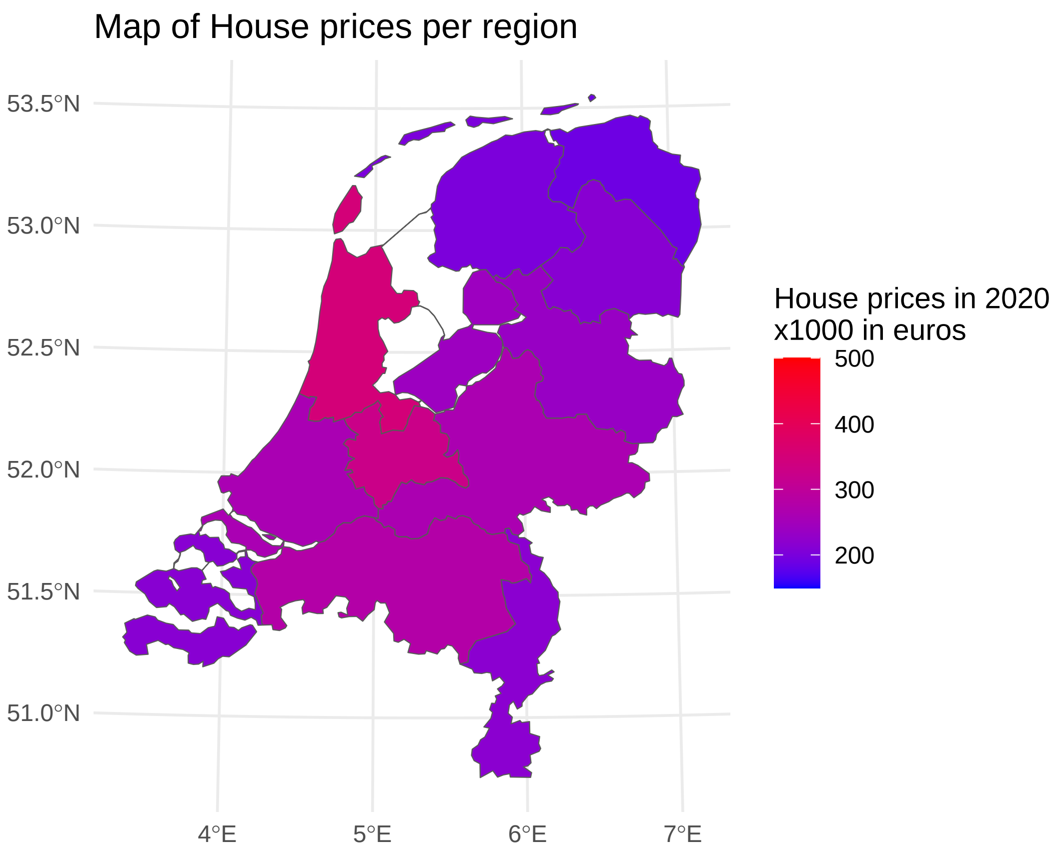
3.3 Visualize temporal variation

```
ggplot(graph_data_long, aes(x = Year, y = Value, color = Region, group = Region)) +
  geom_line(linewidth = 1) +
  geom_point(size = 2) +
  labs(
    title = "House Prices Relative to Average Personal Income (2019-2023)",
    x = "Year",
    y = "Relative Value"
  ) +
  theme_bw()
```

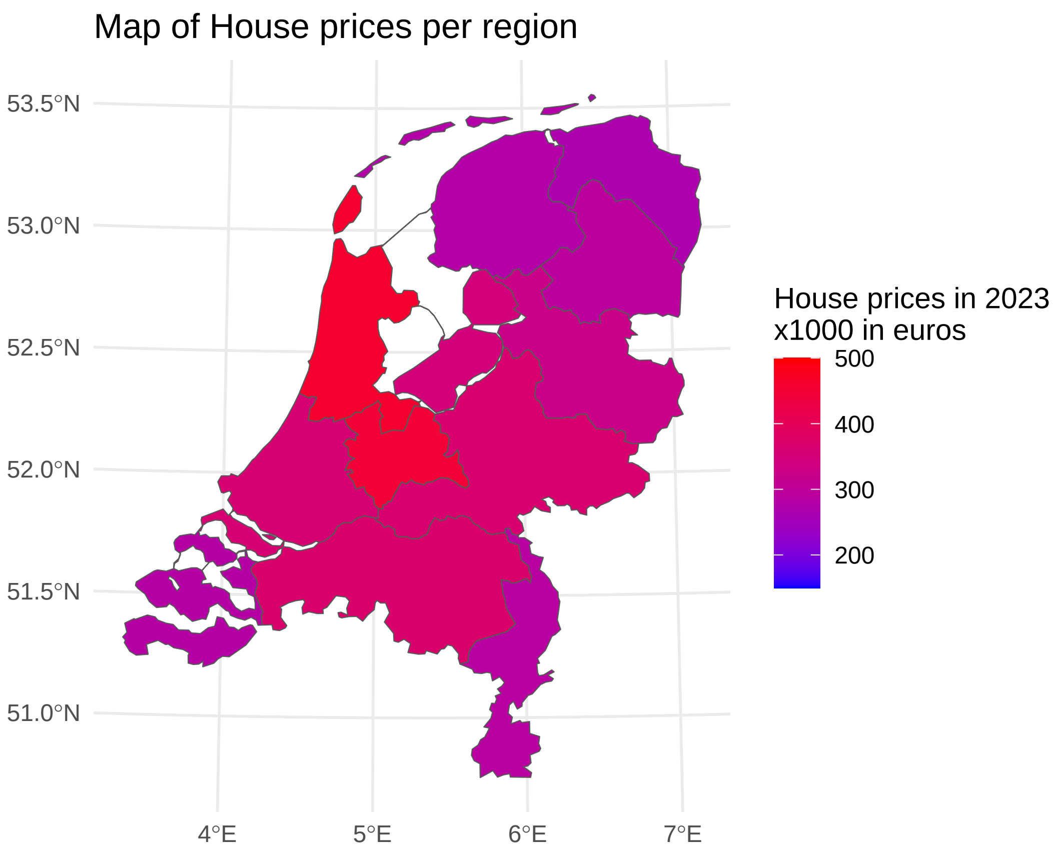


3.4 Visualize spatial variation

```r
# Create the map using ggplot2
ggplot(data2) +
  geom_sf(aes(fill = `2020`)) +  # Mapping values from the 2020 column
  scale_fill_gradient(low = "blue", high = "red", limits = c(150, 500)) +
  theme_minimal() +
  labs(title = "Map of House prices per region", fill = "House prices in 2020\nx1000 in euros")
```



```
# Create the map using ggplot2
ggplot(data2) +
  geom_sf(aes(fill = `2023`)) +  # Mapping values from the 2020 column
  scale_fill_gradient(low = "blue", high = "red", limits = c(150, 500)) +
  theme_minimal() +
  labs(title = "Map of House prices per region", fill = "House prices in 2023\nx1000 in euros")
```

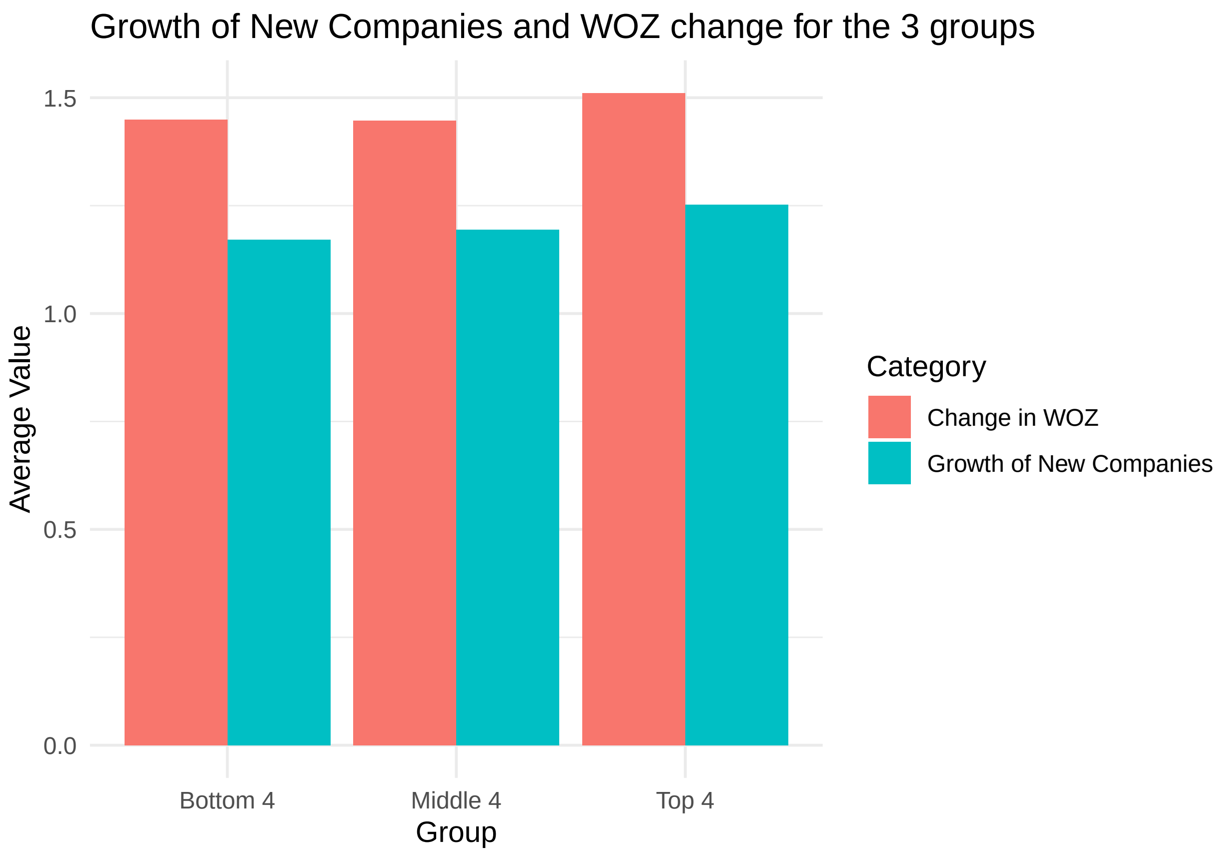


Here you provide a description of why the plot above is relevant to your specific social problem.

3.5 Visualize sub-population variation

Does the growth of new businesses have a direct impact on the change in WOZ value?

```r
ggplot(data, aes(x = Group, y = Value, fill = Category)) +
  geom_bar(stat = "identity", position = "dodge") +
  theme_minimal() +
  labs(title = "Growth of New Companies and WOZ change for the 3 groups",
       x = "Group", y = "Average Value", fill = "Category")
```

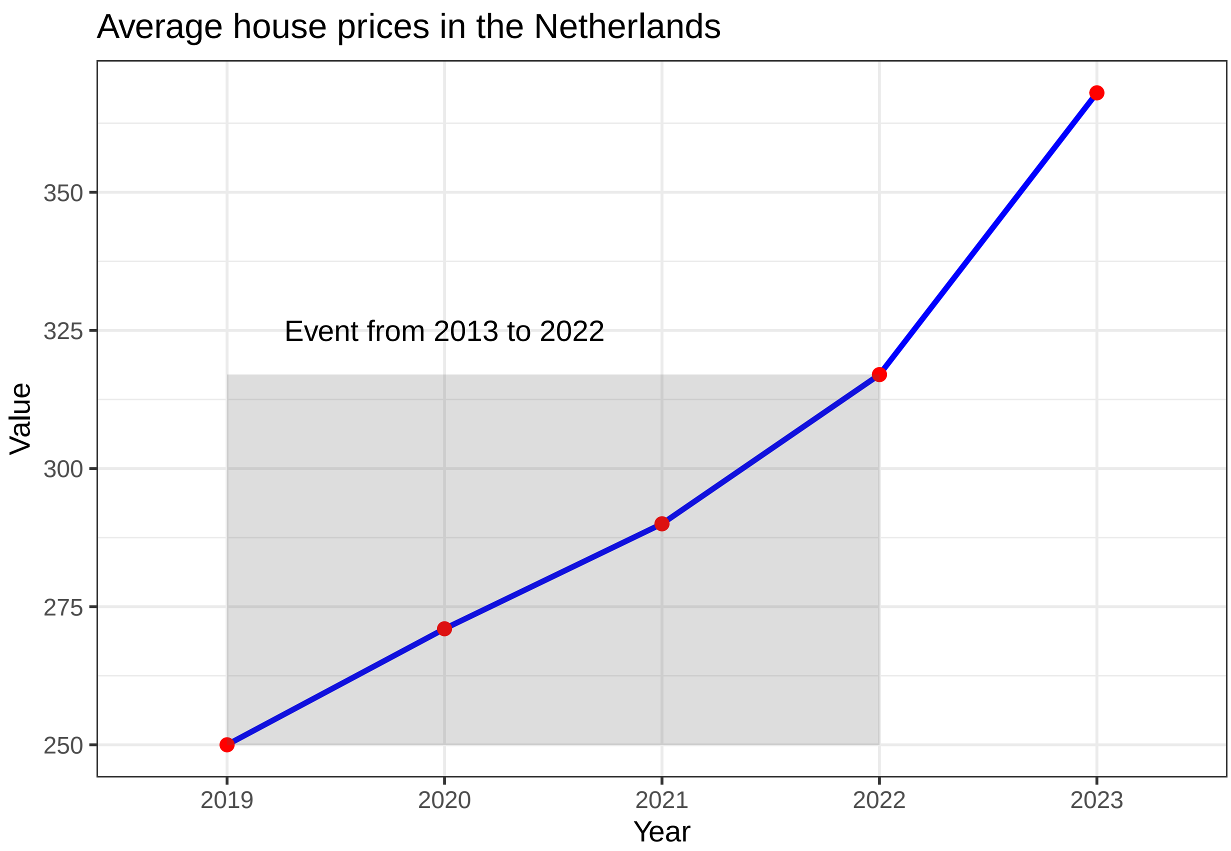


Here you provide a description of why the plot above is relevant to your specific social problem.

3.6 Event analysis

Analyze the relationship between two variables.

```r
ggplot(graph_data, aes(x = Year, y = Value, group = 1)) +  # Explicitly set `group = 1`
  geom_line(color = "blue", linewidth = 1) +
  geom_point(color = "red", size = 2) +
  annotate("text", label = "Event from 2013 to 2022", size = 4, x = 2, y = 325) +
  annotate("rect", xmin = 1, xmax = 4, ymin = 250, ymax = 317, alpha = .2) +
  labs(
    title = "Average house prices in the Netherlands",
    x = "Year",
    y = "Value"
  ) +
  theme_bw()
```



Here you provide a description of why the plot above is relevant to your specific social problem.

Part 4 - Discussion

4.1 Discuss your findings

Part 5 - Reproducibility

5.1 Github repository link

Provide the link to your PUBLIC repository here: https://github.com/jasperdaniel06/Team-Daap

5.2 Reference list

Use APA referencing throughout your document.