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)

2025-06-19

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!

```
dataset <- midwest
```

midwest is an example dataset included in the tidyverse package

2.2 Provide a short summary of the dataset(s)

```r
head(dataset)
```

```
## # A tibble: 6 x 28
##     PID county   state  area poptotal popdensity popwhite popblack popamerindian
##   <int> <chr>    <chr> <dbl>    <int>      <dbl>    <int>    <int>         <int>
## 1   561 ADAMS    IL    0.052    66090      1271.    63917     1702            98
## 2   562 ALEXAND~ IL    0.014    10626       759      7054     3496            19
## 3   563 BOND     IL    0.022    14991       681.    14477      429            35
## 4   564 BOONE    IL    0.017    30806      1812.    29344      127            46
## 5   565 BROWN    IL    0.018     5836       324.     5264      547            14
## 6   566 BUREAU   IL    0.05     35688       714.    35157       50            65
## # i 19 more variables: popasian <int>, popother <int>, percwhite <dbl>,
## #   percblack <dbl>, percamerindan <dbl>, percasian <dbl>, percother <dbl>,
## #   popadults <int>, perchsd <dbl>, percollege <dbl>, percprof <dbl>,
## #   poppovertyknown <int>, percpovertyknown <dbl>, percbelowpoverty <dbl>,
## #   percchildbelowpovert <dbl>, percadultpoverty <dbl>,
## #   percelderlypoverty <dbl>, inmetro <int>, category <chr>
```

In this case we see 28 variables, but we miss some information on what units they are in. We also don't know anything about the year/moment in which this data has been captured.

These are things that are usually included in the metadata of the dataset. For your project, you need to provide us with the information from your metadata that we need to understand your dataset of choice.

2.3 Describe the type of variables included

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.

```r
mean(dataset$percollege)
```

```
## [1] 18.27274
```

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: this is a variable for the house to income ratio we used for the Netherlands, Zeeland and Flevoland.

```
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: This variable is the change in house prices between the years 2023 and 2019 in each province.

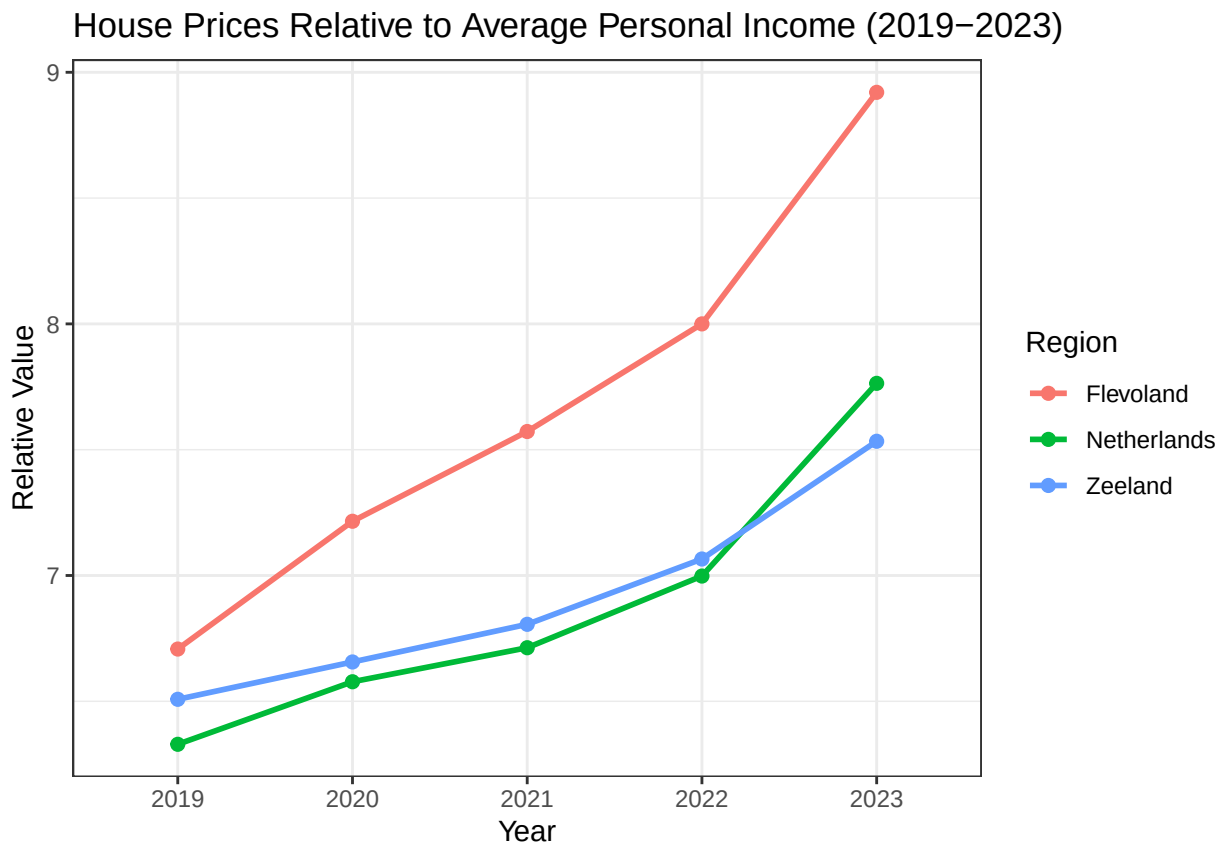
```
dif <- c() # give variable dif the value of a vector

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: This variable is the calculation of the growth of new businesses in each province

```
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
}
```

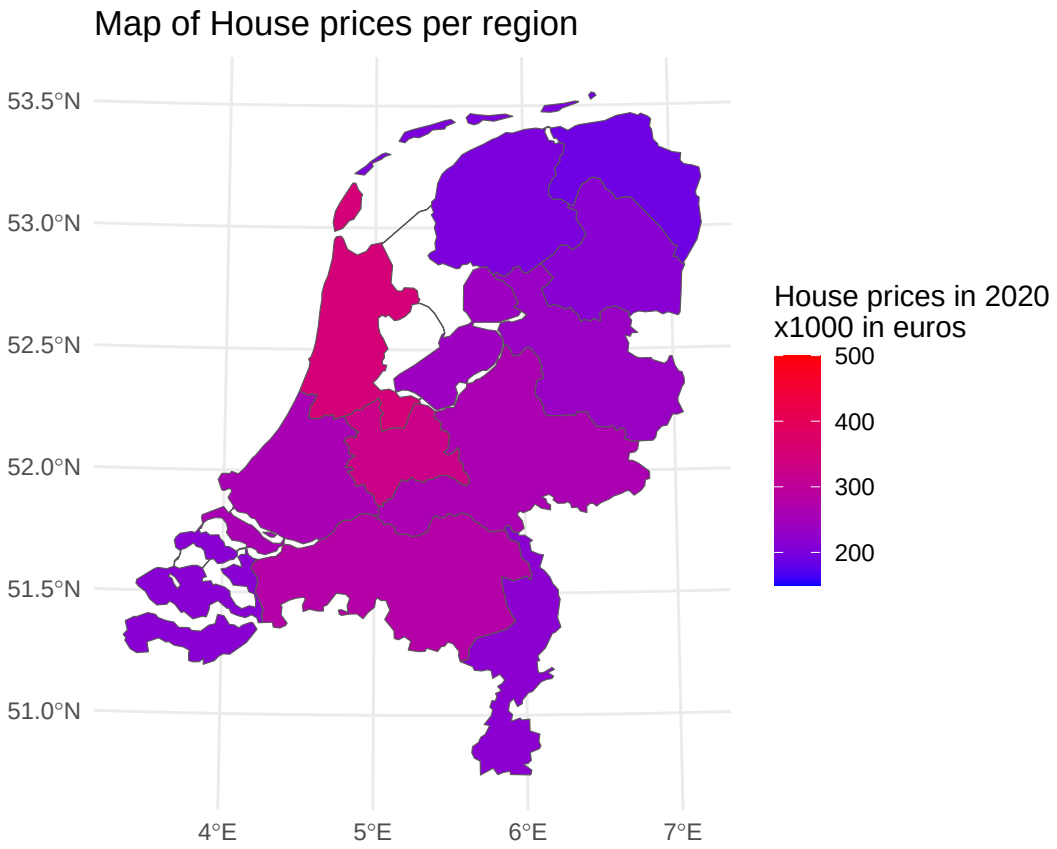
3.3 Visualize temporal variation

```
graph_data <- data.frame(
  Year = colnames(data)[1:5],  # Extract years
  Netherlands = as.numeric(data[5, 1:5]),  # Row 5: Netherlands
  Flevoland = as.numeric(data[6, 1:5]),  # Row 6: Flevoland
  Zeeland = as.numeric(data[7, 1:5])   # Row 7: Zeeland
)

# Convert data to long format
graph_data_long <- pivot_longer(graph_data, cols = -Year, names_to = "Region", values_to = "Value")

# Create the multi-line plot
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()
```

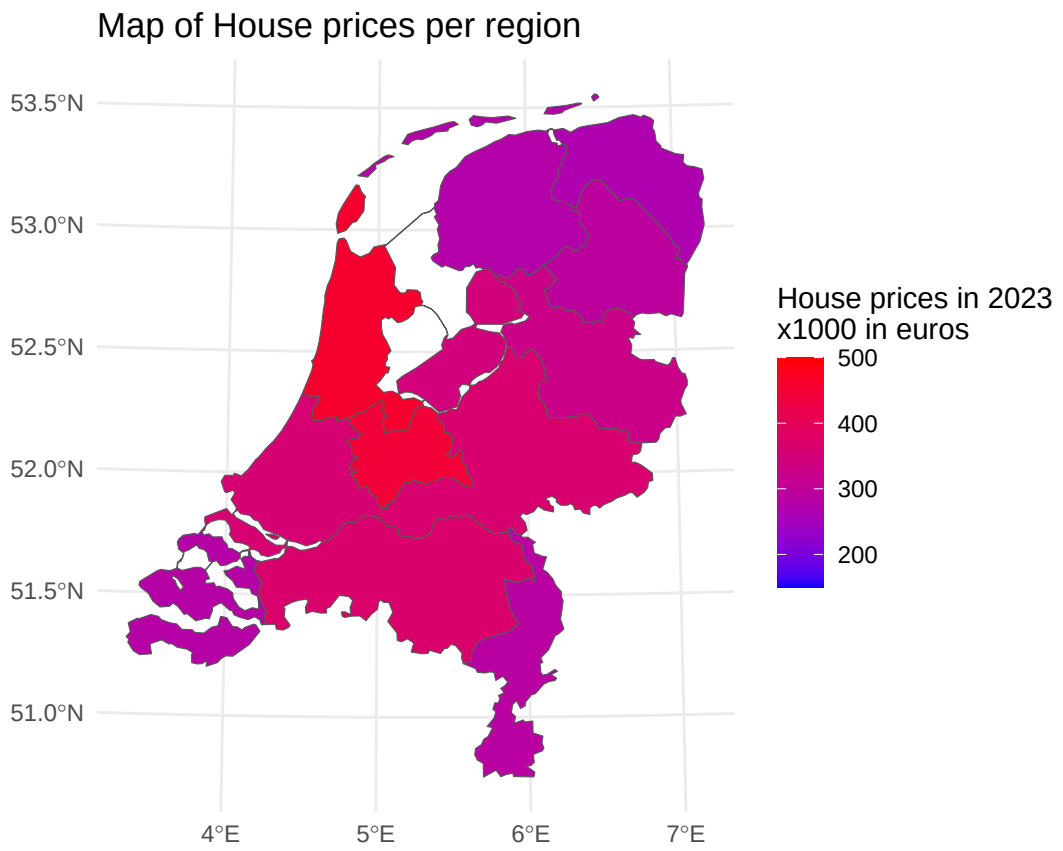


3.4 Visualize spatial variation

```
# 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")
```




```r
# 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")
```

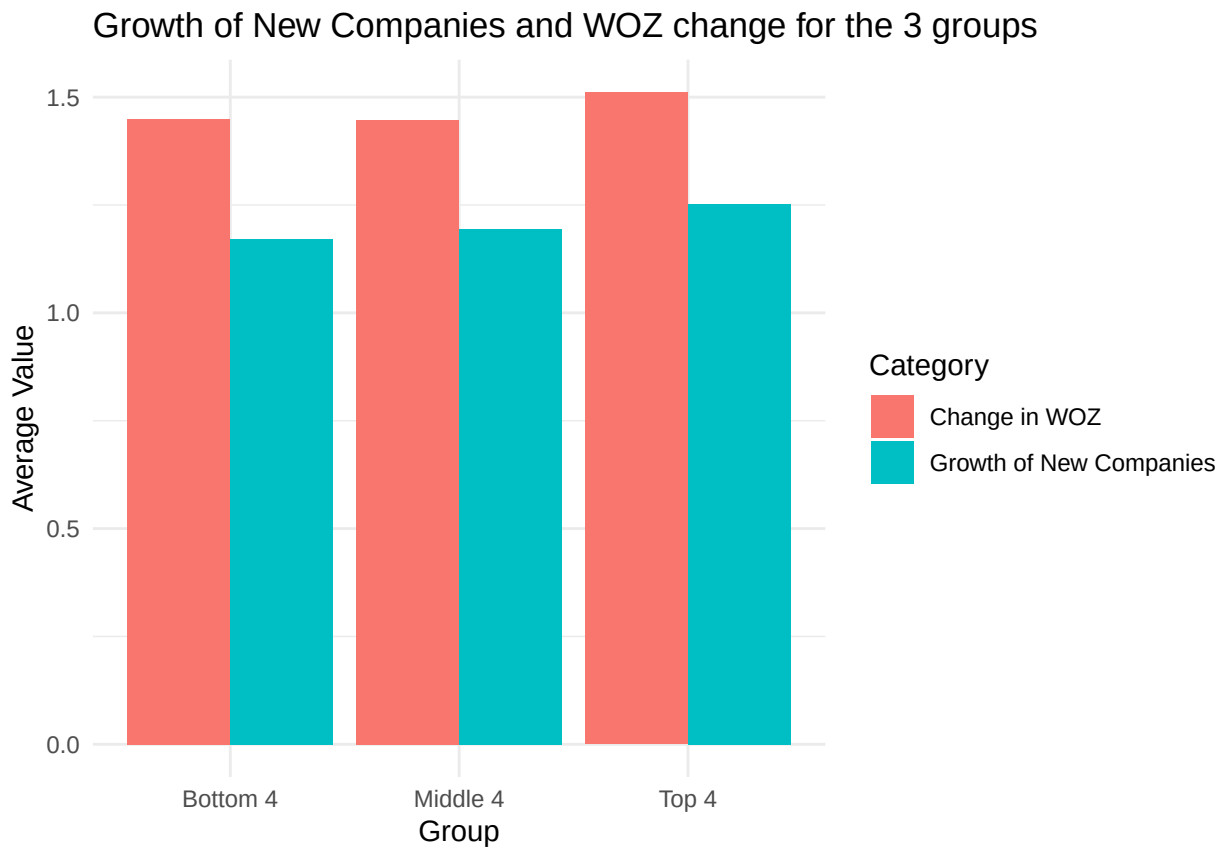


Here you provide a description of why the plot above is relevant to your specific social problem.

3.5 Visualize sub-population variation

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

```
data <- data.frame(
  Group = c("Bottom 4", "Middle 4", "Top 4"),
  Category = rep(c("Growth of New Companies", "Change in WOZ"), each = 3),
  Value = c(bottom, middle, top, bottom_woz, middle_woz, top_woz)
)

# Create the double bar plot
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.

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.