Houshold shortage

Joris heeremans(2787294), Tom Nethe(2784796), Mohamed Hannun(2824390), Jelle Menninga(2864828),

2025-06-24

Set-up your environment

Houshold Shortage

Joris Heeremans, Tom Nethe, Mohamed Hannun, Jelle Menninga, Daan Notenboom, Suleyman Yigitsoy, Soufyan Attokoki

Tutorial group number: 5

Tutorial lecturer's name: Jack Fitzgerald

Part 1 - Identify a Social Problem

1.1 Describe the Social Problem

The Netherlands is dealing with a severe housing shortage, estimated at around 400,000 homes. (CBS, 2025) This imbalance between supply and demand, was created by population growth (including immigration), an increase in single-person households, and a lack of new construction (Langen, January 2025). This imbalance has led to skyrocketing prices and fierce competition for available houses. The house prices have risen significantly, making it difficult for many, especially solo buyers and those with low to middle incomes, to find suitable and affordable housing. The government needs to come up with new ideas, otherwise this problem will only grow bigger.

Part 2 - Data Sourcing

2.1 Load in the data

```
Voorraad_woningen <- read.csv("Vooraad_woningen_Google - Voorraad_woningen.csv")
KW1_voorraad <- Voorraad_woningen[grepl("1e kwartaal", Voorraad_woningen$Perioden), ]

Huishoudens <- read.csv("Aantal_huishoudens - Blad1 (2).csv")

Migratie <- read.csv("Migratie - Blad1 (2).csv")

Bevolking <- read.csv("Bevolking - Blad1.csv")
```

2.2 Provide a short summary of the dataset(s)

```
head(KW1_voorraad)
```

```
##           Regio.s         Perioden  aantal
## 2       Nederland 2019 1e kwartaal 7830489
## 6       Nederland 2020 1e kwartaal 7909246
## 10      Nederland 2021 1e kwartaal 7987921
## 12 Groningen (PV) 2019 1e kwartaal  279996
## 16 Groningen (PV) 2020 1e kwartaal  283748
## 20 Groningen (PV) 2021 1e kwartaal  285538
```

```
head(Huishoudens)
```

```
##          Regio.s Perioden  aantal
## 1      Nederland     2018 7857914
## 2      Nederland     2019 7924691
## 3      Nederland     2020 7997800
## 4      Nederland     2021 8043443
## 5 Groningen (PV)     2018  292255
## 6 Groningen (PV)     2019  293740
```

Both data sets include information about the stock houses and the number of households in the Netherlands per province and period.

Part 3 - Quantifying

3.1 Data cleaning

We already filtered the data through the CBS filter.

```
# Add a year column by extracting the first 4 characters of the period column
Voorraad_woningen$Year <- substr(Voorraad_woningen$Perioden, 1, 4)
Huishoudens$Year <- substr(Huishoudens$Periode, 1, 4)


# Set which years we want to use
years_to_keep <- c("2019", "2020", "2021")

#Set which provinces
provincies <- c("Groningen (PV)", "Fryslân (PV)", "Drenthe (PV)", "Overijssel (PV)", "Flevoland (PV)",
                "Gelderland (PV)", "Utrecht (PV)", "Noord-Holland (PV)", "Zuid-Holland (PV)",
                "Zeeland (PV)", "Noord-Brabant (PV)", "Limburg (PV)")

# Select only data for Netherlands or provinces, and only the needed years, only first quarter
housing_stock_selected <- subset(Voorraad_woningen,
                                 Regio.s %in% provincies &
                                 Year %in% years_to_keep &
                                 grepl("1e kwartaal", Perioden))
```

```r
households_selected <- subset(Huishoudens,
                              Regio.s %in% provincies &
                              Year %in% years_to_keep)

# --- 1. Reshape migration data from wide to long ---
migratie_long <- Migratie %>%
  filter(Onderwerp %in% c("Immigratie", "Emigratie")) %>%
  select(-X) %>%  # Verwijder de kolom die problemen veroorzaakt
  pivot_longer(
    cols = starts_with("X"),
    names_to = "Year",
    names_prefix = "X",
    values_to = "Aantal"
  ) %>%
  select(Regio.s, Onderwerp, Year, Aantal) %>%
  pivot_wider(
    names_from = Onderwerp,
    values_from = Aantal
  ) %>%
  mutate(
    Immigratie = as.numeric(Immigratie),
    Emigratie = as.numeric(Emigratie)
  )
# --- 2. Reshape population data from wide to long ---
bevolking_long <- Bevolking %>%
  filter(Onderwerp == "Bevolking") %>%
  select(-X) %>%  # Verwijder de kolom die de fout veroorzaakt
  pivot_longer(
    cols = starts_with("X"),
    names_to = "Year",
    names_prefix = "X",
    values_to = "Bevolking"
  ) %>%
  select(Regio.s, Year, Bevolking) %>%
  mutate(Bevolking = as.numeric(Bevolking))

# --- 3. Merge and calculate net migration as % of population ---
migratie_bevolking <- migratie_long %>%
  left_join(bevolking_long, by = c("Regio.s", "Year")) %>%
  mutate(
    Netto_Migratie = Immigratie - Emigratie,
    Netto_Migratie_Perc = 100 * Netto_Migratie / Bevolking
  )

head(migratie_bevolking)
```

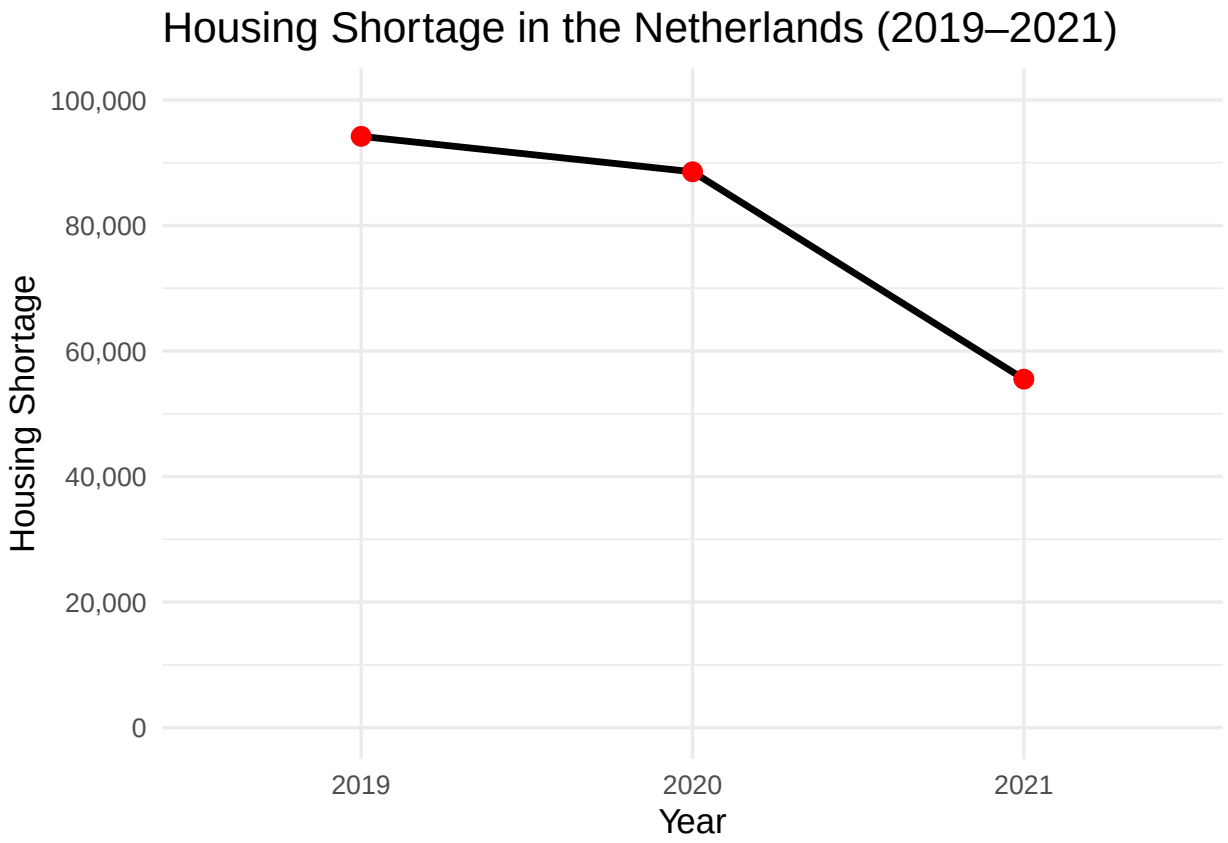
```
## # A tibble: 6 x 7
##   Regio.s        Year  Immigratie Emigratie Bevolking Netto_Migratie
##   <chr>          <chr>      <dbl>     <dbl>     <dbl>          <dbl>
## 1 Groningen (PV) 2019       16664      6239    583990          10425
## 2 Groningen (PV) 2020       14043      6406    585866           7637
## 3 Groningen (PV) 2021       22575      5665    586937          16910
## 4 Fryslân (PV)   2019        4497      2676    647672           1821
```

```
## 5 Fryslân (PV)   2020        3513      2532    649957            981
## 6 Fryslân (PV)   2021        3349      2520    651435            829
## # i 1 more variable: Netto_Migratie_Perc <dbl>
```

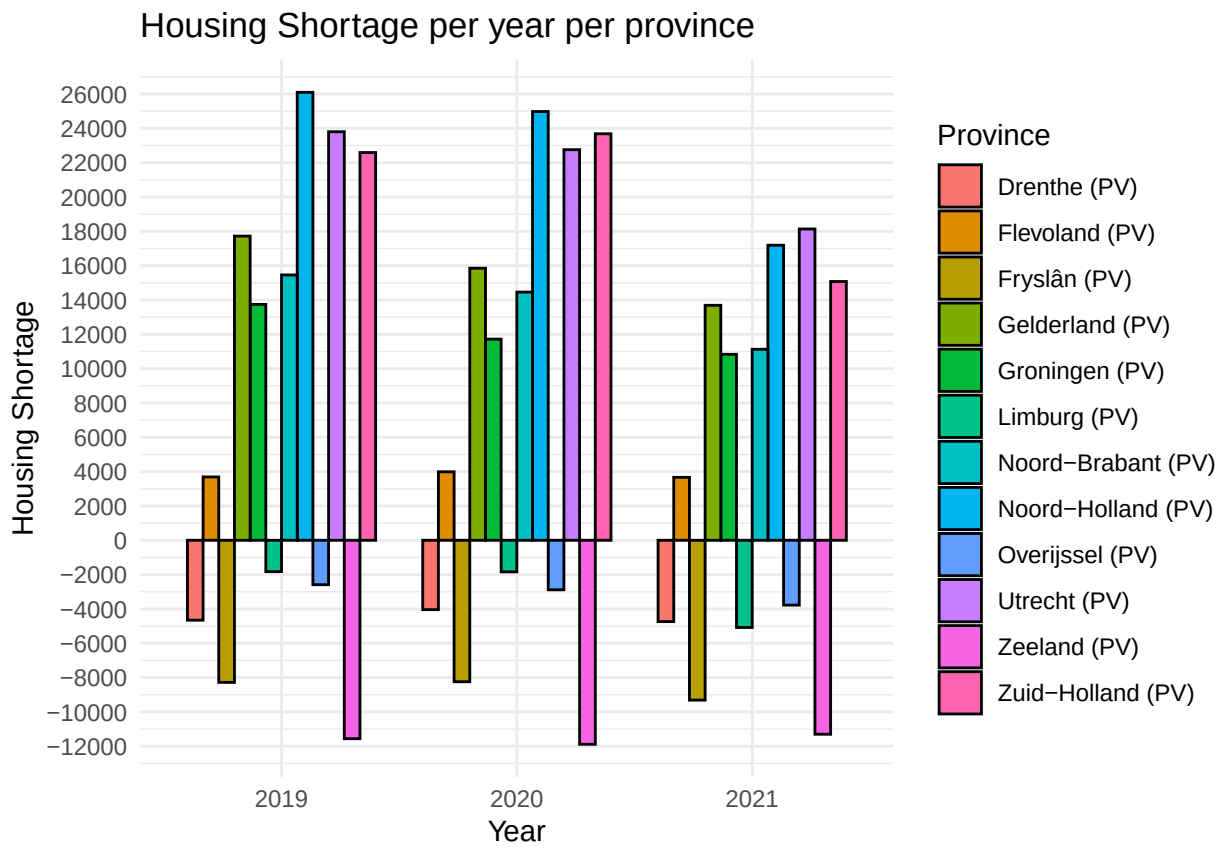
```r
## 3.2 Generate necessary variables

# Merge datasets
data_combined <- housing_stock_selected %>%
  select(Regio.s, Year, Voorraad = aantal) %>%  # Replace value_column if needed
  left_join(
    households_selected %>%
      select(Regio.s, Year, Huishoudens = aantal), # Replace value_column if needed
    by = c("Regio.s", "Year")
  )

# Create shortage variable
data_combined <- data_combined %>%
  mutate(Tekort = Huishoudens - Voorraad)
```

```r
# SUM the shortage across all provinces for each year
nederland_shortage <- data_combined %>%
  group_by(Year) %>%
  summarize(Total_Tekort = sum(Tekort, na.rm = TRUE))

# Zorg dat Year een factor is
nederland_shortage$Year <- as.factor(nederland_shortage$Year)

# Maak de lijnplot met vaste y-as (0 tot 100.000)
ggplot(nederland_shortage, aes(x = Year, y = Total_Tekort, group = 1)) +
  geom_line(color = "black", size = 1.2) +
  geom_point(size = 3, color = "red") +
  scale_y_continuous(
    limits = c(0, 100000),
    breaks = seq(0, 100000, by = 20000),
    labels = scales::comma
  ) +
  labs(
    title = "Housing Shortage in the Netherlands (2019-2021)",
    x = "Year",
    y = "Housing Shortage"
  ) +
  theme_minimal(base_size = 13)
```

```
## Warning: Using ‘size‘ aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use ‘linewidth‘ instead.
## This warning is displayed once every 8 hours.
## Call ‘lifecycle::last_lifecycle_warnings()‘ to see where this warning was
## generated.
```



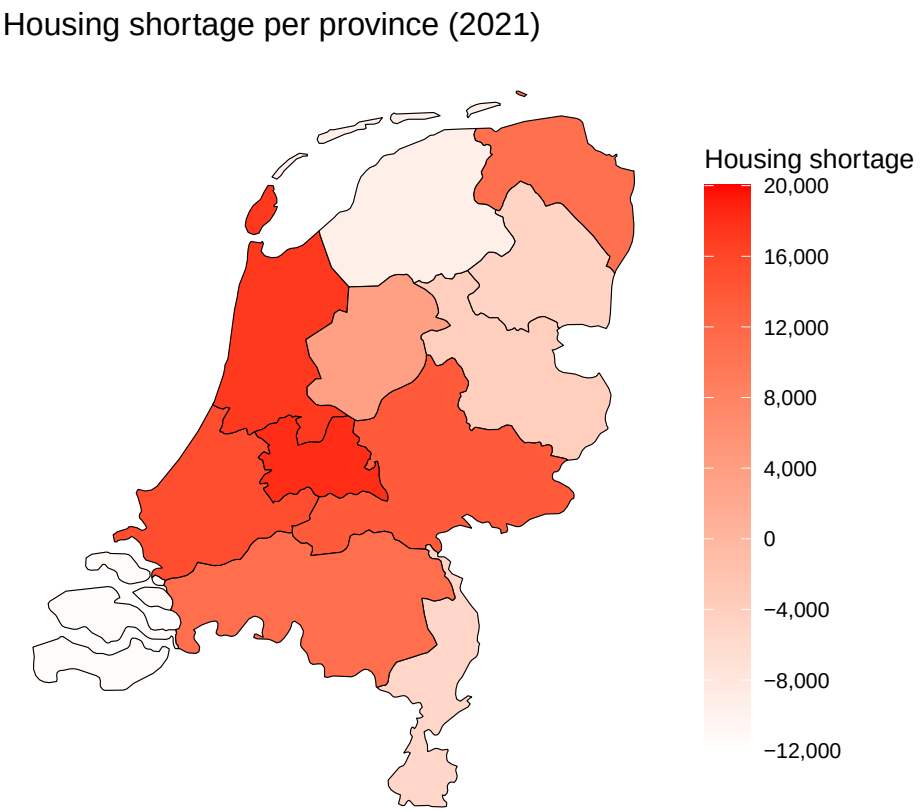

```
## 3.3 Visualize temporal variation


# Zorg dat je factor-niveau voor jaar logisch wordt geordend (voor de x-as)
data_combined$Year <- as.factor(data_combined$Year)

# zwarte rand, smallere staven optioneel
ggplot(data_combined, aes(x = factor(Year), y = Tekort, fill = Regio.s)) +
  geom_col(position = "dodge", color = "black", width = 0.8) +
  scale_y_continuous(
    breaks = seq(
      floor(min(data_combined$Tekort, na.rm = TRUE) / 2000) * 2000,
      ceiling(max(data_combined$Tekort, na.rm = TRUE) / 2000) * 2000,
      by = 2000)
  ) +
  labs(
    title = "Housing Shortage per year per province",
    x = "Year",
    y = "Housing Shortage",
    fill = "Province"
  ) +
  theme_minimal()
```



3.4 Visualize spatial variation

```
# Haal shapefile met Nederlandse provincies binnen
nl_prov <- ne_states(country = "Netherlands", returnclass = "sf")

# Selecteer alleen de 12 echte provincies van Nederland
provincies_nederland <- c(
  "Drenthe", "Flevoland", "Friesland", "Gelderland", "Groningen",
  "Limburg", "Noord-Brabant", "Noord-Holland", "Overijssel",
  "Utrecht", "Zeeland", "Zuid-Holland"
)
nl_prov_alleen_prov <- nl_prov %>% filter(name %in% provincies_nederland)

# Maak een data-frame voor 2021 met provincie-kolom zonder (PV)
data_2021 <- data_combined %>%
  filter(Year == "2021") %>%
  mutate(provincie_kort = gsub(" \\(PV\\)", "", Regio.s),
         provincie_kort = ifelse(provincie_kort == "Fryslân", "Friesland", provincie_kort),
         )

# Join: voeg de tekorten toe aan de kaart
kaart_met_tekort <- nl_prov_alleen_prov %>%
  left_join(data_2021, by = c("name" = "provincie_kort"))
```

```r
# Plot de kaart met kleurverloop blauw (laag) tot rood (hoog)
ggplot(kaart_met_tekort) +
  geom_sf(aes(fill = Tekort), color = "black", size = 0.5) +
  scale_fill_gradient(
    low = "white",
    high = "red",
    name = "Housing shortage",
    limits = c(-12000, 20000),
    breaks = seq(-12000, 20000, by = 4000),
    labels = scales::comma
  ) +
  labs(
    title = "Housing shortage per province (2021)",
  ) +
  theme_minimal(base_size = 10) +
  theme(
    legend.key.height = unit(1.5, "cm"),
    axis.text = element_blank(),
    axis.ticks = element_blank(),
    panel.grid = element_blank()
  )
```

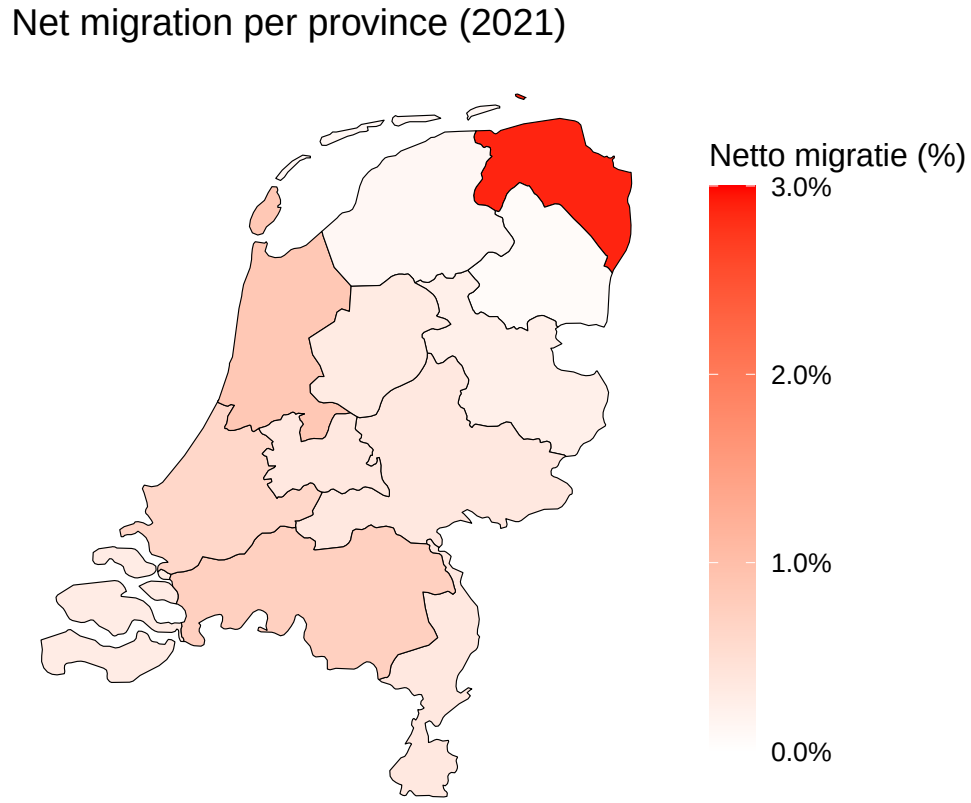


Here you provide a description of why the plot above is relevant to your specific social problem.

3.5 Visualize sub-population variation

What is the poverty rate by state?

```
# 2. Filter to the 12 main provinces
provincies_nederland <- c(
  "Drenthe", "Flevoland", "Friesland", "Gelderland", "Groningen",
  "Limburg", "Noord-Brabant", "Noord-Holland", "Overijssel",
  "Utrecht", "Zeeland", "Zuid-Holland"
)
nl_prov_12 <- nl_prov %>% filter(name %in% provincies_nederland)

# 3. Prepare CBS migration data for 2021
data_migratie <- migratie_bevolking %>%
  filter(Year == "2021") %>%
  mutate(
    provincie_kort = gsub(" \\(PV\\)", "", Regio.s),
    provincie_kort = ifelse(provincie_kort == "Fryslân", "Friesland", provincie_kort),
    Netto_Migratie_Perc = Netto_Migratie / Bevolking * 100
  )

# 4. Join shapefile with migration data
kaart_cbs_migratie <- nl_prov_12 %>%
  left_join(data_migratie, by = c("name" = "provincie_kort"))

ggplot(kaart_cbs_migratie) +
  geom_sf(aes(fill = Netto_Migratie_Perc), color = "black", size = 0.4) +
  scale_fill_gradient(
    low = "white", high = "red",
    name = "Netto migratie (%)",
    limits = c(0, 3),
    breaks = seq(0, 3, by = 1),
    labels = scales::percent_format(accuracy = 0.1, scale = 1)
  ) +
  labs(
    title = "Net migration per province (2021)",
    fill = "Migration %"
  ) +
  theme_minimal(base_size = 12) +
  theme(
    legend.key.height = unit(1.5, "cm"),
    axis.text = element_blank(),
    axis.ticks = element_blank(),
    panel.grid = element_blank()
  )
```



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

3.6 Event analysis

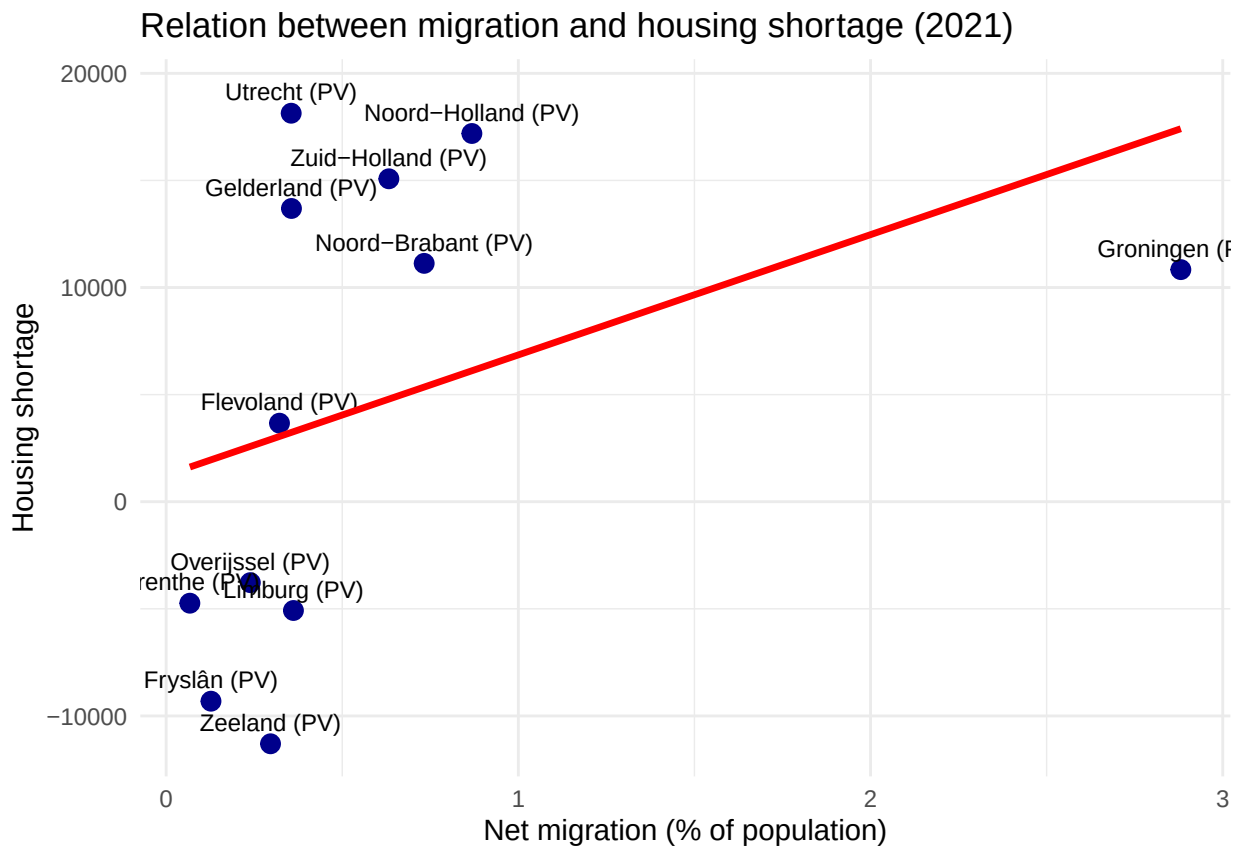
Analyze the relationship between two variables.

```
# Clean Regio.s before the join
event_analysis_data <- data_combined %>%
  filter(Year == "2021") %>%
  mutate(provincie_kort = gsub(" \\(PV\\)", "", Regio.s),
         provincie_kort = ifelse(provincie_kort == "Fryslân", "Friesland", provincie_kort)) %>%
  left_join(
    migratie_bevolking %>%
      filter(Year == "2021") %>%
      mutate(
        provincie_kort = gsub(" \\(PV\\)", "", Regio.s),
        provincie_kort = ifelse(provincie_kort == "Fryslân", "Friesland", provincie_kort),
        Netto_Migratie_Perc = Netto_Migratie / Bevolking * 100
      ) %>%
      select(provincie_kort, Netto_Migratie_Perc),
    by = "provincie_kort"
  )

# Clean for plotting
event_analysis_data_clean <- event_analysis_data %>%
```

```r
  filter(!is.na(Netto_Migratie_Perc), !is.na(Tekort))

# Plot with regression line and province labels
ggplot(event_analysis_data_clean, aes(x = Netto_Migratie_Perc, y = Tekort, label = Regio.s)) +
  geom_point(color = "darkblue", size = 3) +
  geom_text(nudge_y = 1000, size = 3) +
  geom_smooth(method = "lm", se = FALSE, color = "red", linewidth = 1.2) +
  labs(
    title = "Relation between migration and housing shortage (2021)",
    x = "Net migration (% of population)",
    y = "Housing shortage"
  ) +
  theme_minimal(base_size = 11)
```

```
## `geom_smooth()` using formula = 'y ~ x'
```

```
## Warning: The following aesthetics were dropped during statistical transformation: label.
## i This can happen when ggplot fails to infer the correct grouping structure in
##   the data.
## i Did you forget to specify a `group` aesthetic or to convert a numerical
##   variable into a factor?
```

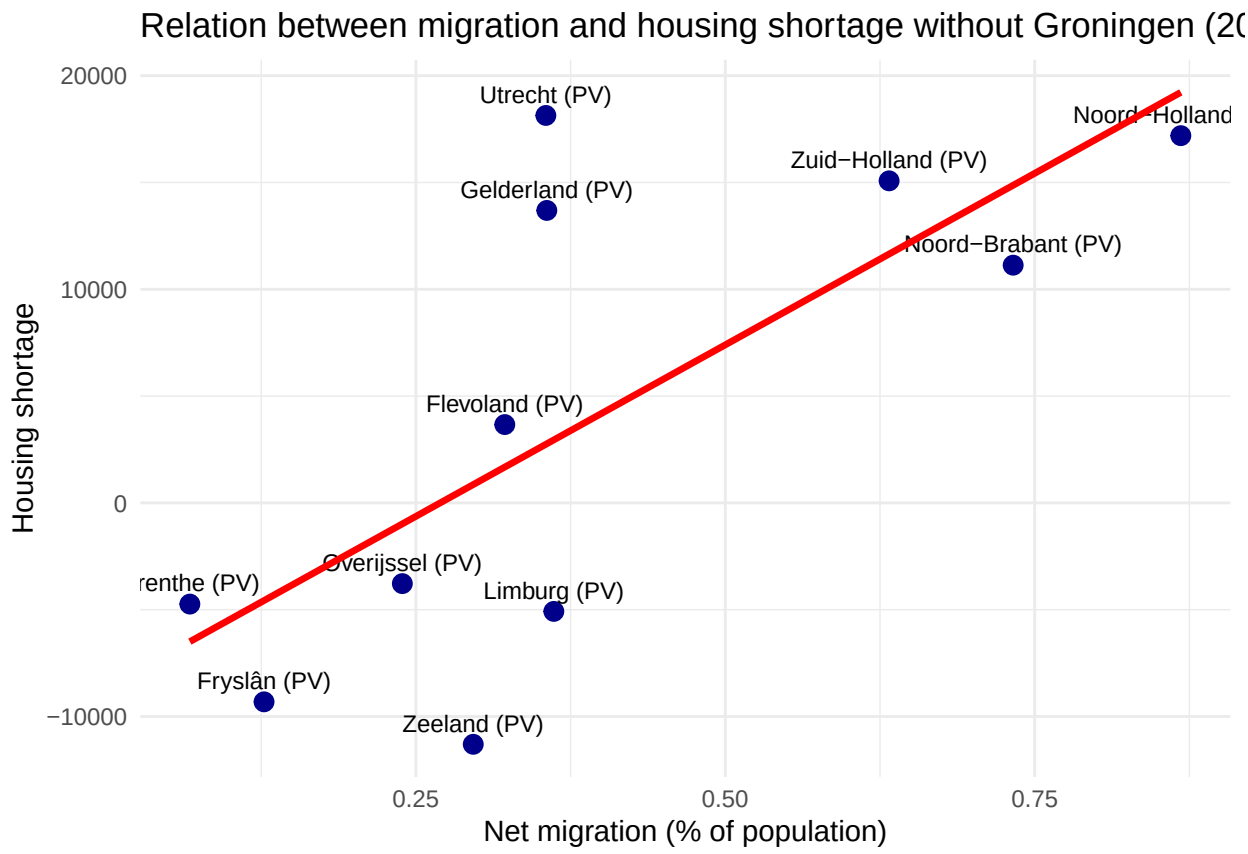



```r
# Maak dataset en sluit Groningen uit
event_analysis_data <- data_combined %>%
```

```r
  filter(Year == "2021") %>%
  mutate(
    provincie_kort = gsub(" \\(PV\\)", "", Regio.s),
    provincie_kort = ifelse(provincie_kort == "Fryslân", "Friesland", provincie_kort)
  ) %>%
  filter(provincie_kort != "Groningen") %>%  #Groningen weghalen
  left_join(
    migratie_bevolking %>%
      filter(Year == "2021") %>%
      mutate(
        provincie_kort = gsub(" \\(PV\\)", "", Regio.s),
        provincie_kort = ifelse(provincie_kort == "Fryslân", "Friesland", provincie_kort),
        Netto_Migratie_Perc = Netto_Migratie / Bevolking * 100
      ) %>%
      select(provincie_kort, Netto_Migratie_Perc),
    by = "provincie_kort"
  )

# Filter op geldige rijen
event_analysis_data_clean <- event_analysis_data %>%
  filter(!is.na(Netto_Migratie_Perc), !is.na(Tekort))

# Maak scatterplot met labels
ggplot(event_analysis_data_clean, aes(x = Netto_Migratie_Perc, y = Tekort, label = Regio.s)) +
  geom_point(color = "darkblue", size = 3) +
  geom_text(nudge_y = 1000, size = 3) +
  geom_smooth(method = "lm", se = FALSE, color = "red", linewidth = 1.2) +
  labs(
    title = "Relation between migration and housing shortage without Groningen (2021)",
    x = "Net migration (% of population)",
    y = "Housing shortage"
  ) +
  theme_minimal(base_size = 11)
```

```
## `geom_smooth()` using formula = 'y ~ x'
```

```
## Warning: The following aesthetics were dropped during statistical transformation: label.
## i This can happen when ggplot fails to infer the correct grouping structure in
##   the data.
## i Did you forget to specify a `group` aesthetic or to convert a numerical
##   variable into a factor?
```



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

https://github.com/ProgrammingforEcon-Team-5-Groep-1/Programming-Group5-Team1

5.2 Reference list

CBS Statline. (May 23, 2025). Voorraad woningen (en niet woningen), https://opendata.cbs.nl/statline/#/CBS/nl/dataset/81955NED/table?fromstatweb (Last used on June 2, 2025) CBS Statline. (June 3, 2025). Huishoudens, https://opendata.cbs.nl/statline/#/CBS/nl/dataset/71486ned/table?fromstatweb (Last used on June 2, 2025) CBS Statline. (May 28, 2025). Bevolkingsontwikkeling, https://opendata.cbs.nl/statline/#/CBS/nl/dataset/37230ned/table?ts=1750160957830 (Last used on June 17,2025) CBS Statline. (May 28, 2025). Bevolkingsontwikkeling, https://opendata.cbs.nl/statline/#/CBS/nl/dataset/37230ned/table?ts=1750159602778 (Last used on June 17,2025) Centraal Bureau voor de Statistiek. (2025). Wonen. Centraal Bureau Voor de Statistiek. https://www.cbs.nl/nl-nl/visualisaties/monitor-brede-welvaart-en-de-sustainable-development-goals/hier-en-nu/

wonen#:~:text=Door%20ABF%20Research%20is%20voor,een%20kleine%20aanvullende%20aardgaslevering%20plaatsvindt.
Mike Langen, ABN AMRO Bank. (16 January, 2025). Housing market - building according to need. https://www.abnamro.com/research/en/our-research/housing-market-building-according-to-need

Use APA referencing throughout your document.