Unemployment and Level of Education

Melle Kooge 2855016 Lex Veerman 2821630 Michael Minneboo 2858823
Mateo van Gerven 2826991 Merel Vonk 2862870 Kirill Müller 2831153
Athiraiyan Visvalingam 2801095

24 June 2025

```
source("controle_data.R")
```

```
##
## 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

## Linking to GEOS 3.13.1, GDAL 3.11.0, PROJ 9.6.0; sf_use_s2() is TRUE

## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v forcats   1.0.0     v tibble    3.2.1
## v lubridate 1.9.4     v tidyr     1.3.1
## 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 error
```

Part 1

Identify a Social Problem The social problem we have chosen is that low-educated people are more often unemployed than high-educated people. In this report we will only focus on statistics in the Netherlands.

1.1 Motivation

The inequality between lower and higher educated people is a continuing challenge in the labor market. According to CBS (Centraal Bureau voor de Statistiek, a surveyor of economic data from mainly the Netherlands), lower educated people are more frequently unemployed than higher educated peers. This will cause a broadening gap between lower and higher educated employees, impacting both their opportunities in the labor market and their financial stability. This research is aimed to store our findings in R to give a general overview of the social problem.

Part 2 - Data Sourcing

2.1 Load in the Data

2.2 Summary of the data

```
head(dataset_1)
```

```
##                   Geslacht       Leeftijd Perioden             Regio.s
## 1 Totaal mannen en vrouwen 15 tot 75 jaar     2015           Nederland
## 2 Totaal mannen en vrouwen 15 tot 75 jaar     2015 Noord-Nederland (LD)
## 3 Totaal mannen en vrouwen 15 tot 75 jaar     2015  Oost-Nederland (LD)
## 4 Totaal mannen en vrouwen 15 tot 75 jaar     2015  West-Nederland (LD)
## 5 Totaal mannen en vrouwen 15 tot 75 jaar     2015  Zuid-Nederland (LD)
## 6 Totaal mannen en vrouwen 15 tot 75 jaar     2015       Groningen (PV)
##   Bevolking..aantal. Onderwijsniveau.5.categorien.11.Basisonderwijs....
## 1           12849400                                                10.2
## 2            1299400                                                 9.6
## 3            2688700                                                10.0
## 4            6100800                                                10.3
## 5            2760500                                                10.4
## 6             452500                                                 9.2
##   Onderwijsniveau.5.categorien.12.Vmbo..havo...vwo.onderbouw..mbo1....
## 1                                                                 20.6
## 2                                                                 20.8
## 3                                                                 21.6
## 4                                                                 19.5
## 5                                                                 22.1
## 6                                                                 18.5
##   Onderwijsniveau.5.categorien.21.Havo..vwo..mbo2.4....
## 1                                                  41.4
## 2                                                  45.4
## 3                                                  43.1
## 4                                                  39.7
## 5                                                  41.7
## 6                                                  45.0
##   Onderwijsniveau.5.categorien.31.Hbo...wo.bachelor....
## 1                                                  17.8
## 2                                                  17.2
## 3                                                  17.5
## 4                                                  18.2
## 5                                                  17.6
## 6                                                  18.2
##   Onderwijsniveau.5.categorien.32.Hbo...wo.master..doctor....
## 1                                                        10.0
## 2                                                         7.0
## 3                                                         7.9
## 4                                                        12.4
## 5                                                         8.2
## 6                                                         9.2
```

In this data set the different level of education has been measured per Region (provinces) per year in the Netherlands. It is a percentage based on the total inhabitants per region.

```
head(dataset_2)
```

```
##                       Quarter Nederland Groningen Friesland Drenthe Overijssel
## 1 het eerste kwartaal van 2015       8.5      10.5       8.8     9.3        8.1
## 2 het tweede kwartaal van 2015       7.9       9.6       8.4     7.8        7.8
## 3  het derde kwartaal van 2015       7.5       8.8       8.1     7.2        7.4
## 4 het vierde kwartaal van 2015       7.6       9.2       8.1     7.8        7.3
## 5 het eerste kwartaal van 2016       7.8       8.9       8.7     8.0        7.6
## 6 het tweede kwartaal van 2016       7.2       8.5       7.4     7.1        7.2
##   Flevoland Gelderland Utrecht Noord.Holland Zuid.Holland Zeeland Noord.Brabant
## 1       9.4        7.9     7.9           8.2          9.4     6.9           7.9
## 2       8.9        7.6     7.1           7.4          8.9     6.7           7.6
## 3       8.0        7.0     7.1           7.2          8.4     6.2           7.2
## 4       8.1        7.0     7.1           7.4          8.4     6.5           7.3
## 5       9.1        7.6     6.8           7.3          8.6     6.4           7.2
## 6       7.7        6.9     6.4           7.0          8.2     6.0           6.8
##   Limburg Year
## 1     8.2 2015
## 2     7.7 2015
## 3     7.1 2015
## 4     7.4 2015
## 5     7.8 2016
## 6     6.8 2016
```

In this dataset the unemployment has been measured per Region (provinces) per year in the Netherlands. It is a percentage based on the working population. Currently the data is measured per quarter of a year. We need to have an average from each year to have it corresponds with our other dataset. The years that have been measured also differs with the other dataset, so we only use the years that both datasets inlcude.

2.3 Describe the variables included

```
head(dataset_1)
```

```
##                   Geslacht       Leeftijd Perioden              Regio.s
## 1 Totaal mannen en vrouwen 15 tot 75 jaar     2015            Nederland
## 2 Totaal mannen en vrouwen 15 tot 75 jaar     2015 Noord-Nederland (LD)
## 3 Totaal mannen en vrouwen 15 tot 75 jaar     2015  Oost-Nederland (LD)
## 4 Totaal mannen en vrouwen 15 tot 75 jaar     2015  West-Nederland (LD)
## 5 Totaal mannen en vrouwen 15 tot 75 jaar     2015  Zuid-Nederland (LD)
## 6 Totaal mannen en vrouwen 15 tot 75 jaar     2015       Groningen (PV)
##   Bevolking..aantal. Onderwijsniveau.5.categorien.11.Basisonderwijs....
## 1           12849400                                               10.2
## 2            1299400                                                9.6
## 3            2688700                                               10.0
## 4            6100800                                               10.3
## 5            2760500                                               10.4
## 6             452500                                                9.2
##   Onderwijsniveau.5.categorien.12.Vmbo..havo...vwo.onderbouw..mbo1....
## 1                                                                 20.6
## 2                                                                 20.8
```

```
## 3                                                                    21.6
## 4                                                                    19.5
## 5                                                                    22.1
## 6                                                                    18.5
##   Onderwijsniveau.5.categorien.21.Havo..vwo..mbo2.4....
## 1                                                  41.4
## 2                                                  45.4
## 3                                                  43.1
## 4                                                  39.7
## 5                                                  41.7
## 6                                                  45.0
##   Onderwijsniveau.5.categorien.31.Hbo...wo.bachelor....
## 1                                                  17.8
## 2                                                  17.2
## 3                                                  17.5
## 4                                                  18.2
## 5                                                  17.6
## 6                                                  18.2
##   Onderwijsniveau.5.categorien.32.Hbo...wo.master..doctor....
## 1                                                        10.0
## 2                                                         7.0
## 3                                                         7.9
## 4                                                        12.4
## 5                                                         8.2
## 6                                                         9.2
```

The variables we have in this data set are level of education, this the different levels of education people have in the Netherlands. Secondly the region, these are the twelve provinces in the Netherlands. Lastly there is the years, these are the years from 2013 to 2023.

```
head(dataset_2)
```

```
##                       Quarter Nederland Groningen Friesland Drenthe Overijssel
## 1 het eerste kwartaal van 2015       8.5      10.5       8.8     9.3        8.1
## 2 het tweede kwartaal van 2015       7.9       9.6       8.4     7.8        7.8
## 3  het derde kwartaal van 2015       7.5       8.8       8.1     7.2        7.4
## 4 het vierde kwartaal van 2015       7.6       9.2       8.1     7.8        7.3
## 5 het eerste kwartaal van 2016       7.8       8.9       8.7     8.0        7.6
## 6 het tweede kwartaal van 2016       7.2       8.5       7.4     7.1        7.2
##   Flevoland Gelderland Utrecht Noord.Holland Zuid.Holland Zeeland Noord.Brabant
## 1       9.4        7.9     7.9           8.2          9.4     6.9           7.9
## 2       8.9        7.6     7.1           7.4          8.9     6.7           7.6
## 3       8.0        7.0     7.1           7.2          8.4     6.2           7.2
## 4       8.1        7.0     7.1           7.4          8.4     6.5           7.3
## 5       9.1        7.6     6.8           7.3          8.6     6.4           7.2
## 6       7.7        6.9     6.4           7.0          8.2     6.0           6.8
##   Limburg Year
## 1     8.2 2015
## 2     7.7 2015
## 3     7.1 2015
## 4     7.4 2015
## 5     7.8 2016
## 6     6.8 2016
```

The variables we have in this data set are: - level of unemployment: the percentage of people that are unemployed from the working population - region: the twelve provinces in the Netherlands - quarter per year: the years go through the first quarter of 2015 to the first quarter of 2025

#Part 3 Quantifying

3.1 Data cleaning

```
#Load the dataset
Unemployment <- read.csv2("Werkloosheidpercentage.csv")

#Rename the first column and extract the year from the 'Quarter' column
colnames(Unemployment)[1] <- "Quarter"
Unemployment$Year <- str_extract(Unemployment$Quarter, "\\d{4}")

#Replace commas with dots in the numeric columns and convert to numeric
#(exclude 'Quarter' and 'Jaar' columns)
numeric_cols <- setdiff(colnames(Unemployment), c("Quarter", "Year"))
Unemployment[numeric_cols] <- lapply(Unemployment[numeric_cols], function(x) {
  as.numeric(str_replace_all(x, ",", "."))
})

#Group by year and calculate the mean per column
Unemployment_yearly <- Unemployment %>%
  group_by(Year) %>%
  summarise(across(all_of(numeric_cols), ~round(mean(.x, na.rm = TRUE), 2)))
```

The file is created where the total unemployment per year with columns is clear. Firstly the first column was renamed to Quarter. Then the years from the data set were extracted, so they could be grouped by year. Previously the data was grouped per quarter per year, so to get a yearly average the mean from each year per region was calculated. To calculate the mean, the necessary numbers were converted to numeric.

```
Unemployment_yearly <- Unemployment_yearly[-c(9, 10, 11), ]
```

The rows for the years 2023, 2024 and 2025 were removed, as these years were not used to make the necessary computations.

```
#Load in the education level by year datasets
Educationlevel2015=read.csv2("2015_onderwijsniveau.csv", fileEncoding = "UTF-8")
Educationlevel2016=read.csv2("2016_onderwijsniveau.csv", fileEncoding = "UTF-8")
Educationlevel2017=read.csv2("2017_onderwijsniveau.csv", fileEncoding = "UTF-8")
Educationlevel2018=read.csv2("2018_onderwijsniveau.csv", fileEncoding = "UTF-8")
Educationlevel2019=read.csv2("2019_onderwijsniveau.csv", fileEncoding = "UTF-8")
Educationlevel2020=read.csv2("2020_onderwijsniveau.csv", fileEncoding = "UTF-8")
Educationlevel2021=read.csv2("2021_onderwijsniveau.csv", fileEncoding = "UTF-8")
Educationlevel2022=read.csv2("2022_onderwijsniveau.csv", fileEncoding = "UTF-8")
Educationlevel2023=read.csv2("2023_onderwijsniveau.csv", fileEncoding = "UTF-8")
Educationlevel_by_year=rbind(Educationlevel2015, Educationlevel2016, Educationlevel2017,
                             Educationlevel2018, Educationlevel2019, Educationlevel2020,
                             Educationlevel2021,Educationlevel2022,Educationlevel2023)
```

The table with education levels were sorted by year. So, the tables were downloaded by year and then bounded.

```r
#Renamed all the column names to an easier accessible name or to an English word
Educationlevel_by_year <- Educationlevel_by_year %>%
  rename(
    Basisonderwijs = Onderwijsniveau.5.categorien.11.Basisonderwijs....,
    VMBO_HAVO_VWO_Onderbouw_MBO_1 =
      Onderwijsniveau.5.categorien.12.Vmbo..havo...vwo.onderbouw..mbo1....,
    HAVO_VWO_MBO_2_4 = Onderwijsniveau.5.categorien.21.Havo..vwo..mbo2.4....,
    HBO_WO_Bachelor = Onderwijsniveau.5.categorien.31.Hbo...wo.bachelor....,
    HBO_WO_Master_Doctor = Onderwijsniveau.5.categorien.32.Hbo...wo.master..doctor....,
    Gender = Geslacht,
    Age = Leeftijd,
    Year = Perioden,
    Region = Regio.s,
    Population = Bevolking..aantal.
  )
```

The column names in the dataset for the education level by year were simplified and changed to English, to make improve the readability of the dataset.

```r
Low_and_High_Education <- Low_and_High_Education[, -c(6, 7, 8, 9, 10)]
```

Some columns were also deleted to improve the readability. This piece of code was not the only instance when this was used.

```r
Low_and_High_Education_Provinces_2022 <- Low_and_High_Education_Provinces_2022 %>%
  mutate(Region = gsub(" \\(PV\\)", "", Region))
```

Within the province, the letters (PV) were written behind the name of the province, so these were removed.

```r
Low_and_High_Education_provinces$Region <- ifelse(Low_and_High_Education_provinces$Region
                                                  == "Fryslan", "Friesland",
                                                  Low_and_High_Education_provinces$Region)
```

The province Friesland, was as Fryslan in one dataset, so this name was changed into Friesland.

3.2 Generate necessary variables

```r
Low_and_High_Education <- Educationlevel_by_year %>%
  mutate(
    Low_Education_rate =
      Basisonderwijs +
      VMBO_HAVO_VWO_Onderbouw_MBO_1 +
      HAVO_VWO_MBO_2_4,
    High_Education_rate =
      HBO_WO_Bachelor +
      HBO_WO_Master_Doctor
  )
```

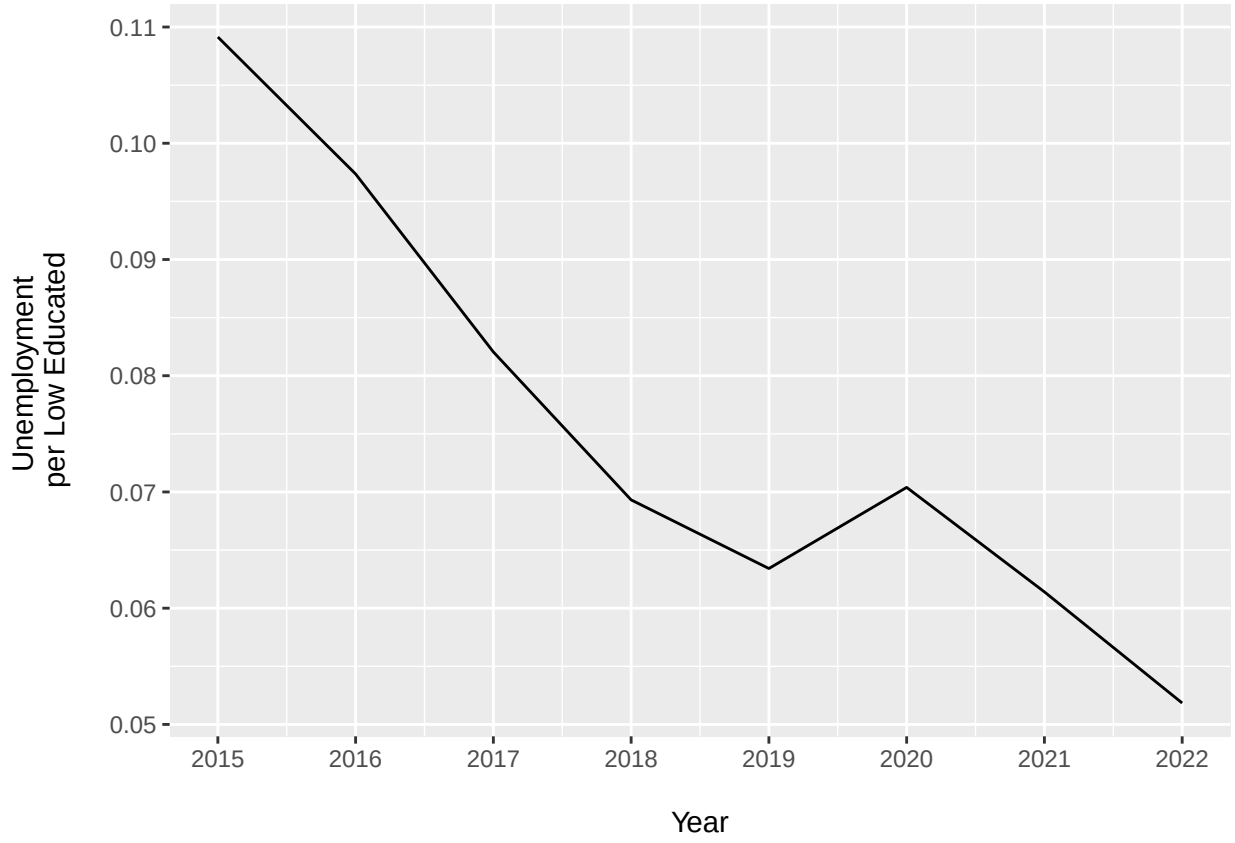
The used datasets divided the education level in 5 different parts. To make the computations easier two variables were created, dividing these 5 parts into low education and high education. This was done by adding the percentages in the different parts. Now there were only two variables to use in the computations.

```r
#Added the Unemployment_by_Low_Education column
Education_Unemployment_yearly_transposed_provinces <-
  Education_Unemployment_yearly_transposed_provinces %>%
  mutate(Unemployment_by_Low_Education = Unemployment / Low_Education_rate)
```

The third variable was created by dividing the unemployment by low education rate. This was done to realise a variable to find a connection between the unemployment and the low education rate in each province. By using unemployment and the low education rate, intuitively there should be a high unemployment by low education rate in provinces with a low education rate, and it should be low in provinces with a high education rate.

##3.3 Visualize temporal variation

```r
ggplot(Education_Employment, aes(x = Year, y = Unemployment_by_Low_Education)) +
  geom_line() +
  labs(x = "\nYear", y = "Unemployment \nper Low Educated\n")+
  scale_x_continuous(breaks = seq(2015, 2022, by = 1), lim = c(2015, 2022))
```

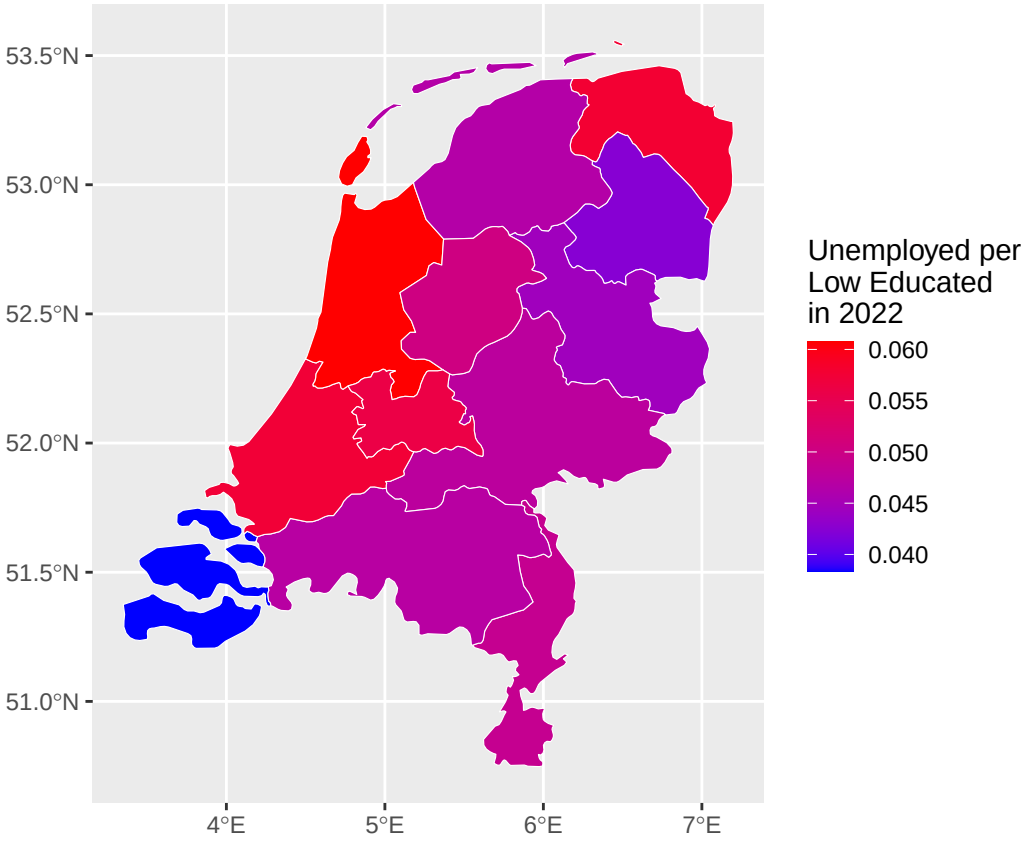


This line graph shows the national trend in unemployment rates per low-educated individuals in the Netherlands from 2015 to 2022. It is relevant for emphasizing the overall long-term progress, while showing that low-educated people still face employment instability.

##3.4 Visualize spatial variation

```r
ggplot(map_data, aes(fill = Unemployment_by_Low_Education)) +
  geom_sf(color = "white", size = 0.2) +
```

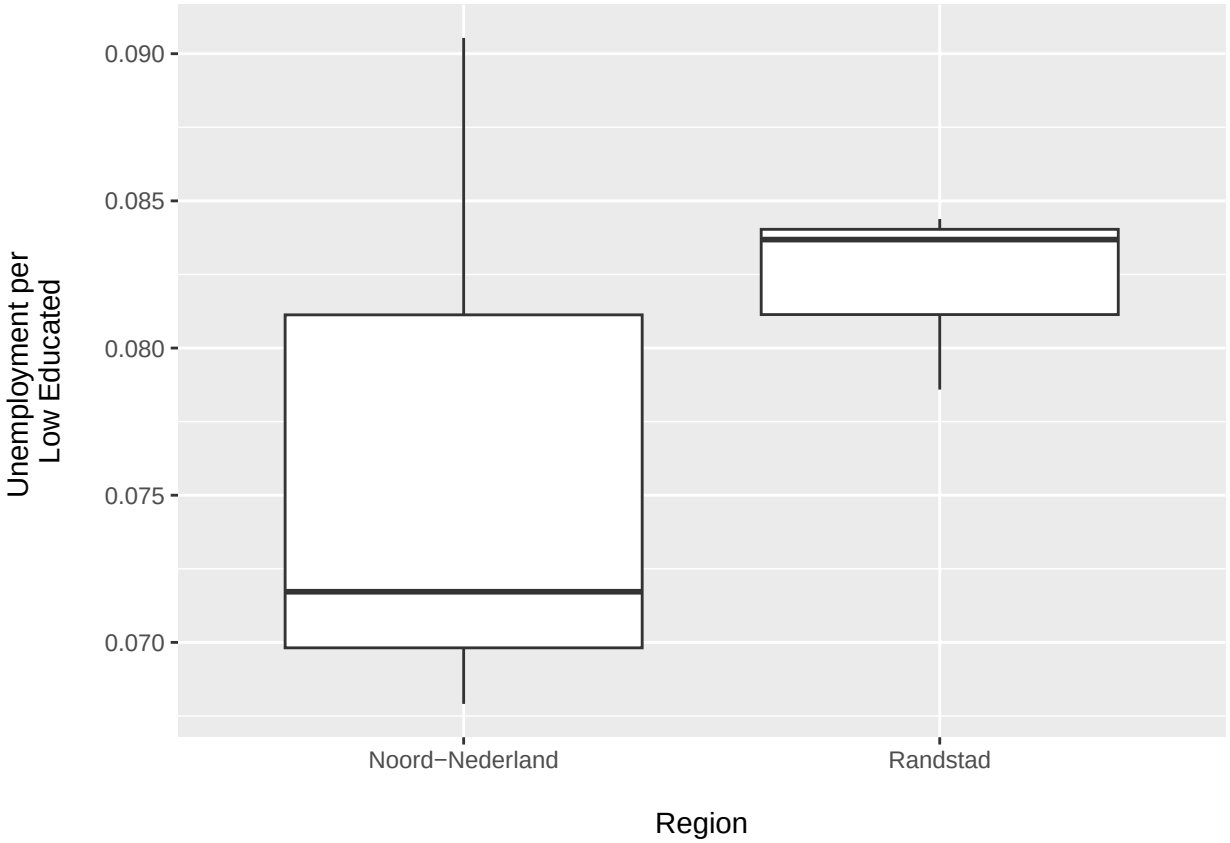
```
  scale_fill_gradient(low = "blue", high = "red",
                      name = "Unemployed per\nLow Educated\nin 2022")
```



This choropleth map shows the regional distribution of unemployment rates per low-educated individuals in 2022 across Dutch provinces. The color gradient from blue to red indicates lower to higher unemployment by low-educated. This plot is key in showing that regional differences play a major role in unemployment outcomes. Despite having the highest share of low-educated residents, Zeeland has the lowest unemployment, contradicting the expectation that a higher rate of low-educated people automatically leads to higher unemployment.

3.5 Visualize sub-population variation

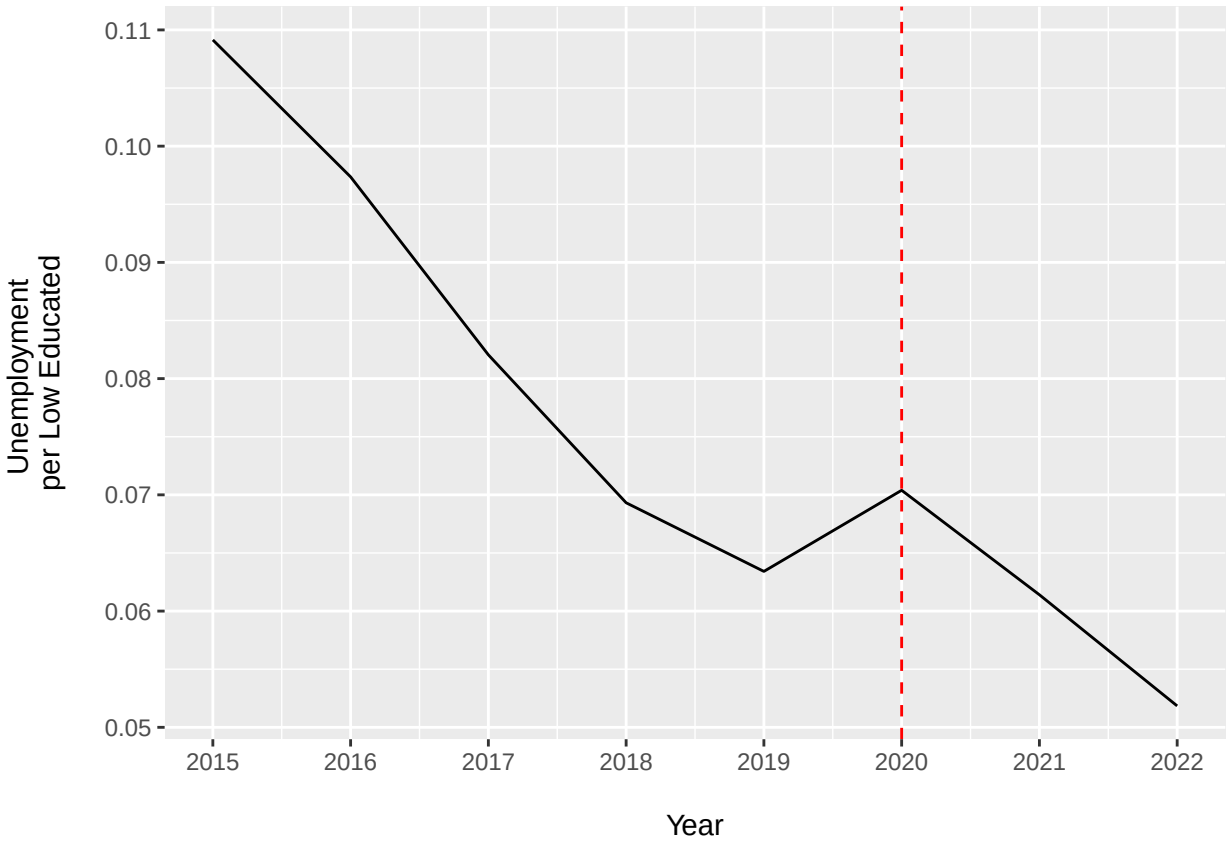
```
ggplot(data = mean_by_provinces, mapping = aes(x = Region_Group,
                                               y = Mean_Unemployment_by_Low_Education)) +
  geom_boxplot() +
  labs(x = "\nRegion", y = "Unemployment per \nLow Educated\n")
```



This boxplot compares the distribution of unemployment rates per low-educated individuals in two regions of the Netherlands: the Randstad and the Northern Netherlands This plot is especially relevant to our research question because it directly visualizes regional differences in unemployment per low-educated individuals.

3.6 Event analysis

```
ggplot(Education_Employment, aes(x = Year, y = Unemployment_by_Low_Education)) +
  geom_line() +
  geom_vline(xintercept = 2020,
             color = "red",
             linetype = "dashed") +
  labs(x = "\nYear", y = "Unemployment \nper Low Educated\n")+
  scale_x_continuous(breaks = seq(2015, 2022, by = 1), lim = c(2015, 2022))
```



This line graph shows the national trend in unemployment rates per the low-educated individuals in the Netherlands from 2015 to 2022. The vertical red line at 2020 marks the start of the COVID-19 pandemic. This plot is important because it illustrates that unemployment among low-educated individuals was steadily decreasing from 2015 to 2019. However, there is a slight increase in 2020, likely due to the pandemic, after which the rate continues to decline. The main reason for this sudden shift is that unemployment suddenly rose during the pandemic.

Part 4 Discussion

At the start of our research, we hypothesized that there would be a clear correlation between education level and unemployment rates within the Netherlands. Specifically, we expected that regions with a lower average education level would have higher unemployment, while regions with a higher education level would experience lower unemployment. This expectation was based on the general assumption that higher education provides individuals with better job opportunities and employability. However, after processing the datasets and visualizing the results in graphs, our findings challenged this expectation. For example, the province of Zeeland had the highest proportion of low-educated individuals, yet it also had the lowest unemployment rate of all provinces. This was a surprising outcome that did not align with our initial hypothesis. Another notable finding was the contrast between the Randstad region and the northern provinces of the Netherlands. Despite the Randstad being more urbanized and economically developed, the unemployment rate among low-educated individuals was higher in the Randstad than in the northern provinces. This was unexpected, as we initially assumed that urban areas with more diverse economies would offer more opportunities, even for those with lower levels of education.

Part 5 Reproducibility

5.1 Github repository link

https://github.com/Lex0803/project.git

5.2 Reference List

CBS Statline. (z.d.). https://opendata.cbs.nl/#/CBS/nl/dataset/85525NED/table

Centraal Bureau voor de Statistiek. (z.d.). Werkloosheid per provincie. Centraal Bureau Voor de Statistiek. https://www.cbs.nl/nl-nl/visualisaties/dashboard-arbeidsmarkt/werklozen/werkloosheid-per-provincie