# Programming for economists project - excessive alcohol use in The Netherlands

Derek van der Linden (2860655), Tom Haandrikman (2866711), Anouk Knaack (2772444), Sara Hoogenb

24-6-2025

## Set-up your environment

```
require(tidyverse)
library(readr)

knitr::opts_chunk$set(
  echo = TRUE,       # Toon code (of FALSE om code te verbergen)
  message = FALSE,   # Verberg messages zoals '## Rows: ...'
  warning = FALSE    # Verberg warnings
)
```

## Excessive alcohol use in The Netherlands

Derek van der Linden (2860655), Tom Haandrikman (2866711), Anouk Knaack (2772444), Sara Hoogenboom (2825231), Levi van der Kolk (2857053), Isabel Nagel (2812985)

CPR02 - Chantal Schouwenaar,

## Part 1 - Identify a Social Problem

1.1 Describe the Social Problem

Our project is about the use of excessive alcohol in Rotterdam and a smaller town close to it called Krimpenerwaard. We are going to combine this with the distance to facilities to buy alcohol, such as cafes, nightclubs, coffee shops and other party center. We are going to compare the two places with different years: 2016, 2020 and 2022. Also, we made a graph showing the overall unhealthiness combining different variables and comparing them with the different

Excessive drinking is a big societal problem and causes different issues, as well as health consequences, legal and social consequences. The excessive abuse of alcohol can affect your breathing, heart rate and gag reflex and also potentially lead to coma and death. It can also contribute to criminal behavior and many other social consequences, such as: financial problems due to excessive spending on alcohol, impacts on work performance and the loss of friends and family (SA Health, n.d.).

Of course, there has been much research into excessive alcohol use, but no research that aligns with that of ours; comparing a big city to a smaller town combined with the distance to facilities. In Rotterdam there is a higher accessibility of facilities than in Krimpenerwaard. Our research could provide us with insights into whether this difference influences people's alcohol use. Also, the different years could show insights into if the covid-19 pandemic has had any influence on the use of excessive alcohol.

# Part 2 - Data Sourcing

## 2.1 Load in the data

```
Alcohol_data_2016<- read_csv("data/Data/Gezondheid_per_wijk_en_buurt__2022_18062025_140754.csv")

Alcohol_data_2020<- read_csv("data/Data/Gezondheid_per_wijk_en_buurt__2022_18062025_140356.csv")

Alcohol_data_2022<- read_csv("data/Data/Gezondheid_per_wijk_en_buurt__2022_18062025_140404.csv")

Nabijheid_voorzieningen2016 <- read_csv("nabijheid_voorzieningen2016.csv")

Nabijheid_voorzieningen2020 <- read_csv("nabijheid_voorzieningen2020.csv")

Nabijheid_voorzieningen2022 <- read_csv("nabijheid_voorzieningen2022.csv")
```

## 2.2 Provide a short summary of the dataset(s)

```
#head(dataset)
```

After identifying the social problem and loading the datasets, we will examine these datasets and provide a summary of how they relate to our problem and what information we can actually obtain from them.

We used three different points in time in combination with two variables in the datasets, "alcohol data" and "nabijheid voorzieningen". Our datasets will provide us with a clear comparison, and by comparing them, we may even be able to identify some relationships between the variables.

Now that the intentions of the use of our datasets are clear, what do our datasets tell us?

The Alcohol__data__2016, Alcohol__data__2020, and Alcohol__data__2022 datasets contain health-related data at a "wijk/ buurt" or neighbourhood/ district level. The datasets contain information about a lot of different data points, such as alcohol use by age groups and health indicators that are possibly linked to alcohol.

The nabijheid__voorzieningen2016, nabijheid__voorzieningen2020, and nabijheid__voorzieningen2022 datasets are used as a second variable, and these datasets contain information about access to local services such as bars, cafés, educational facility proximity, and supermarkets. These datasets are similar to the alcohol datasets, since both of them measure at the neighbourhood or district level.

Both datasets are open data, "alcohol__data" is sourced from the Buurtatlas and the "nabijheid__voorzieningen" is sourced from CBS.

```
inline_code = TRUE
```

## 2.3 Describe the type of variables included

When the datasets are talking about an "overmatige drinker", it's talking about excessive drinkers, so men who drink >= 21 glasses per week, or women who drink >= 14 glasses per week.

Both datasets contain the date as the year and have identifier codes that correspond with the neighborhood or district to which the rest of the data in that row is linked. Both also contain categorical variables like names of the neighborhood or district, while only alcohol_data also contains data about the age and gender of people used to create the data_percentage_overmatig_drinken is a prominent variable in the alcohol_data dataset, which gives us a percentage of excessive drinkers.

The datasets for "alcohol_data" were created from a health monitor in the Netherlands, utilizing data from multiple data collection sources. Buurtatlas used sample-sized groups to conclude about the data created, having between 350.000 and 550.000 respondents.

# Part 3 - Quantifying

## 3.1 Data cleaning

```r
library(dplyr)
library(readr)

setwd("~/Documents/GitHub/School_Opdracht")
nabijheid_voorzieningen2016 <- read_csv("nabijheid_voorzieningen2016.csv")
nabijheid_voorzieningen2020 <- read_csv("nabijheid_voorzieningen2020.csv")
nabijheid_voorzieningen2022 <- read_csv("nabijheid_voorzieningen2022.csv")
merged_data <- nabijheid_voorzieningen2016 %>%
  full_join(nabijheid_voorzieningen2020, by = "Regioaanduiding.Codering..code.", suffix = c("_2016", "_2
  full_join(nabijheid_voorzieningen2022, by = "Regioaanduiding.Codering..code.")
nabijheid_voorzieningen2016$...1 <- NULL
voorzien <- rbind(nabijheid_voorzieningen2016, nabijheid_voorzieningen2020,nabijheid_voorzieningen2022)
write_csv(voorzien,"data/Data/voorzieningen.csv")
merged_data <- voorzien %>%
  merge(AlcoholBuurten, by = "Regioaanduiding.Codering..code.")

setwd("~/Documents/GitHub/School_Opdracht")

AlcoholBuurten <- read_csv("data/Data/AlcoholBuurten.csv")
nabijheid_voorzieningen2016$Perioden <- 2016
nabijheid_voorzieningen2020$Perioden <- 2020
nabijheid_voorzieningen2022$Perioden <- 2022

nabijheid_voorzieningen2016$...1 <- NULL

voorzien <- rbind(nabijheid_voorzieningen2016, nabijheid_voorzieningen2020,nabijheid_voorzieningen2022)
write_csv(voorzien,"data/Data/voorzieningen.csv")
merged_data <- voorzien %>%
  merge(AlcoholBuurten, by = c("Regioaanduiding.Codering..code.", "Perioden"))
write_csv(merged_data,"data/Data/merged_data.csv")
```

## 3.2 Generate necessary variables

**Variable 1:**
First we created the new variable Distance_cafes, we created this one to classify whether the distance to facilities as cafes, nightclubs, coffee shops and other party centers are located closer or further than 1 kilometer averagely for people from Rotterdam and Krimpenerwaard. We introduced this categorization

("Cafes closer than 1 km" vs. "Cafes further than 1km") to make it easier to analyze whether proximity to facilities is associated with patterns of excessive alcohol use. By simplifying this distance measure, we can more clearly examine potential relationships between access to drinking venues and alcohol consumption behavior.

```
  mutate(Distance_cafes = if_else(
    Horeca.Cafés.en.dergelijke.Afstand.tot.café.e.d...km.> 1,
    "Cafes further than 1 km",
    "Cafes closer than 1 km"
  ))
```

**Variable 2:**
The second variable Unhealthiness score represents the average level of several key health risk factors within a given population. We calculated this as the mean of the following six indicators, all looking at both Rotterdam & Krimpenerwaard:

- Percentage of smokers
- Percentage of heavy drinkers
- Percentage of excessive drinkers
- Percentage reporting one or more chronic physical conditions
- Percentage at high risk of anxiety or depression
- Percentage experiencing serious or very serious loneliness

This data was also retrieved from Rijksinstituut voor Volksgezondheid en Milieu (RIVM).

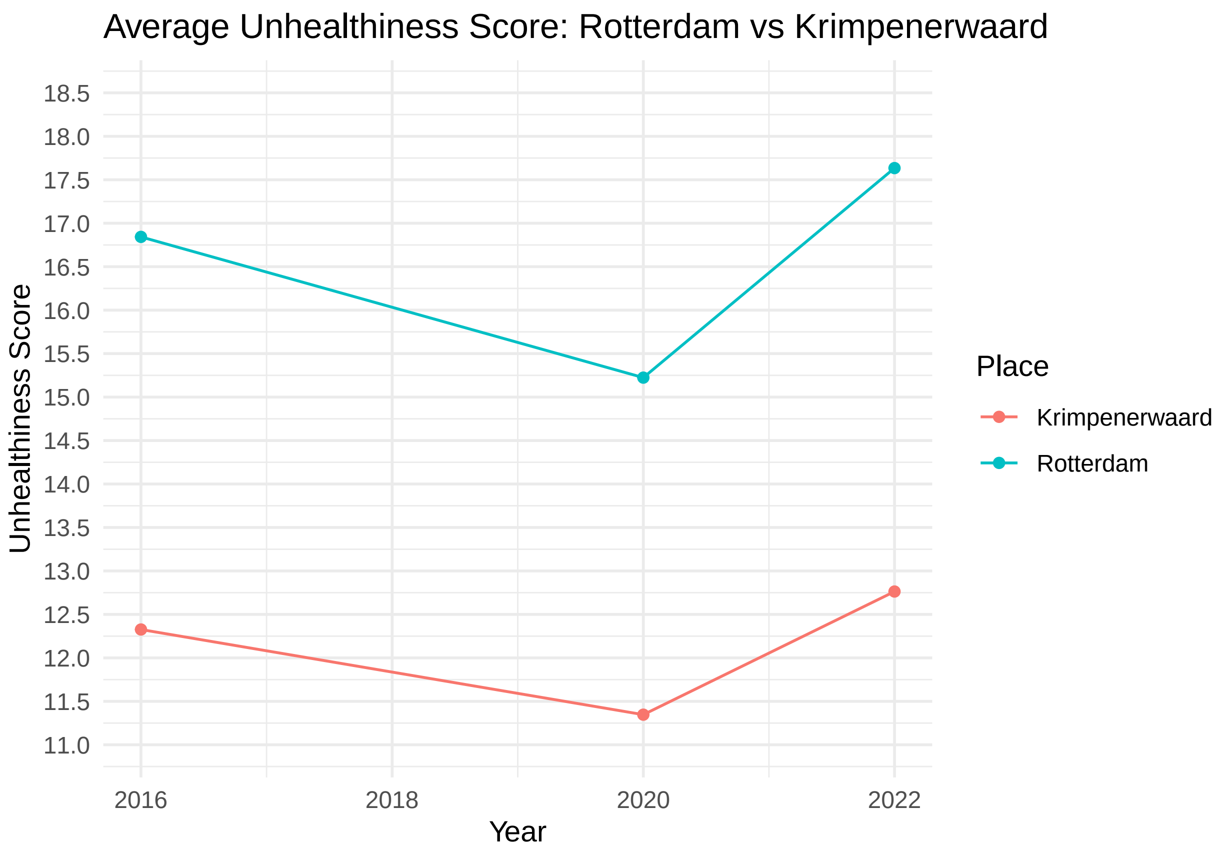
By combining these indicators into a single score, we get a general measure of overall unhealthiness, making it way easier to compare regions. It also makes it easier to see how the health changed through the years in both places.

```
library(dplyr)

# Select columns to include in the score
cols_to_convert <- c(
  "Roker....",
  "Alcoholgebruik.Zware.drinker....",
  "Alcoholgebruik.Overmatige.drinker....",
  "Lichamelijke.gezondheid.Eén.of.meer.langdurige.aandoeningen....",
  "Hoog.risico.op.angst.of.depressie....",
  "Eenzaamheid.Ernstig.zeer.ernstig.eenzaam...."
)

# Replace commas with periods and convert character values to numeric
merged_data[cols_to_convert] <- lapply(merged_data[cols_to_convert], function(x) {
  x <- gsub(",", ".", x)
  as.numeric(x)
})

# Calculate the unhealthiness_score as the average of the selected columns
merged_data <- merged_data %>%
  mutate(
    ongezondheid_score = rowMeans(select(., all_of(cols_to_convert)), na.rm = TRUE)
  )
```

## 3.3 Visualize temporal variation

This line graph titled "Average Unhealthiness Score: Rotterdam vs Krimpenerwaard" uses our second new variable we've created. It compares the average unhealthiness scores of two places Rotterdam and Krimpeneraard, over three years: 2016, 2020 and 2022. This shows us very easily if the type of place, more urban like Rotterdam, or more rural and small-town like Krimpenerwaard, has any influence on the healthiness of the people that live there. In the discussions we will go further in the observations from this graph.

```r
# Gemiddelde ongezondheid_score per jaar berekenen
library(dplyr)
library(ggplot2)

# Calculate average unhealthiness score per year and plot
# Make sure the column with city name is correctly named

merged_data %>%
  filter(Regioaanduiding.Gemeentenaam..naam..x %in% c("Rotterdam", "Krimpenerwaard")) %>%
  group_by(Perioden, Regioaanduiding.Gemeentenaam..naam..x) %>%
  summarise(avg_unhealthiness = mean(ongezondheid_score, na.rm = TRUE), .groups = "drop") %>%
  ggplot(aes(x = Perioden, y = avg_unhealthiness, color = Regioaanduiding.Gemeentenaam..naam..x)) +
  geom_line() +
  scale_y_continuous(breaks = seq(11,19, by = 0.5), limits =c(11,18.5)) +
  geom_point() +
  labs(
    title = "Average Unhealthiness Score: Rotterdam vs Krimpenerwaard",
    x = "Year",
    y = "Unhealthiness Score",
    color = "Place"
  ) +
  theme_minimal()
```



### 3.4 Visualize spatial variation

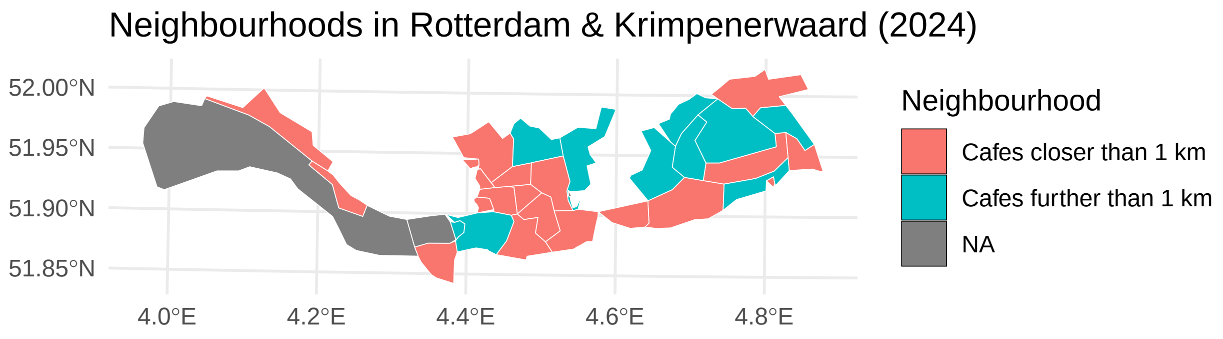
This map “Neighborhoods in Rotterdam & Krimpenerwaard (2024) is based on their proximity to facilities to buy alcohol, such as cafes, nightclubs, coffee shops and other party centers. It uses the first new variable we created: Distance_cafes.

- The Red (Cafes closer than 1 km) parts show us the part where the neighborhoods have a cafe or similar venue within 1 kilometer.
- The blue/teal (Cafes further than 1 km) parts show where the neighborhoods are that are more than 1 kilometer away from the nearest cafe or similar venue.
- Then we have the grey (NA) parts that have no data available for these neighborhoods.

This map helps us to identify areas where people are more likely to have easy access to drinking venues, which could contribute to higher alcohol consumption.

```r
setwd("~/Documents/GitHub/School_Opdracht")
if (!require("remotes")) install.packages("remotes")
if (!require("sf")) install.packages("sf")
if (!require("tidyverse")) install.packages("tidyverse")
if (!require("cbsodataR")) install.packages("cbsodataR")

library(tidyverse)
library(sf)
library(cbsodataR)
library(ggplot2)
setwd("~/Documents/GitHub/School_Opdracht")
merged_data <- read_csv("data/Data/merged_data.csv")
maps <- cbs_get_maps()
prov_map_yr <- max(maps$year[maps$region == "provincie"])
prov_sf <- cbs_get_sf(region = "provincie", year = prov_map_yr)
zh_sf <- prov_sf %>% filter(statnaam == "Zuid-Holland")

#wijken en buurten er over heen
prov_map_yr <- max(cbs_get_maps()$year[cbs_get_maps()$region == "provincie"])
zh_sf <- cbs_get_sf("provincie", prov_map_yr) %>%
  filter(statnaam == "Zuid-Holland")

wijk_map_yr <- max(cbs_get_maps()$year[cbs_get_maps()$region == "wijk"])
wijken_sf <- cbs_get_sf("wijk", wijk_map_yr)


merged_data <- merged_data %>%
  mutate(Distance_cafes = if_else(
    Horeca.Cafés.en.dergelijke.Afstand.tot.café.e.d...km.> 1,
    "Cafes further than 1 km",
    "Cafes closer than 1 km"
  ))

#write.csv(merged_data, "data/Data/merged_data.csv")
#multipolygon koppelen aan merged_goeie (die wijkcodes)
wijken_sf<-sf::st_as_sf(wijken_sf)
merged_2016 <- merged_data %>%
  filter(Perioden == 2016)

merged_2016<- as.data.frame(merged_2016)

#merged_2016$geometry <- NULL
wijken_sf <- wijken_sf %>%
  rename("Regioaanduiding.Codering..code." = "statcode")

merged_2016 <- merged_2016 %>% merge(wijken_sf,
                                     by = "Regioaanduiding.Codering..code.")
library(dplyr)
library(stringr)

# 1. Most-recent vintages
maps      <- cbs_get_maps()
prov_year <- max(maps$year[maps$region == "provincie"])
wijk_year <- max(maps$year[maps$region == "wijk"])

# 2. Read layers (will return data.frames if Tibble Hunter is on)
zh_sf_raw     <- cbs_get_sf("provincie", prov_year) |>
  filter(statnaam == "Zuid-Holland")
wijken_sf_raw <- cbs_get_sf("wijk", wijk_year)

# 3. Keep neighbourhoods in Rotterdam & Krimpenerwaard
```

```
gem_codes <- c("0599", "1931")                      # GM-codes → digits
wijken_subset_raw <- wijken_sf_raw |>
  mutate(gem_code = str_extract(statcode, "\\d{4}")) |>
  filter(gem_code %in% gem_codes)
# 4. Convert back to sf **here**
zh_sf        <- st_as_sf(zh_sf_raw)
wijken_subset <- st_as_sf(wijken_subset_raw)

# 5. Plot
merged_2016 <- st_as_sf(merged_2016)

ggplot() +
  geom_sf(data = merged_2016,
          aes(fill = Distance_cafes), colour = "white", linewidth = .15) +
  theme_minimal() +
  labs(title = sprintf("Neighbourhoods in Rotterdam & Krimpenerwaard (%s)", wijk_year),
       fill  = "Neighbourhood") +
  guides(fill = guide_legend(override.aes = list(colour = "black")))
```



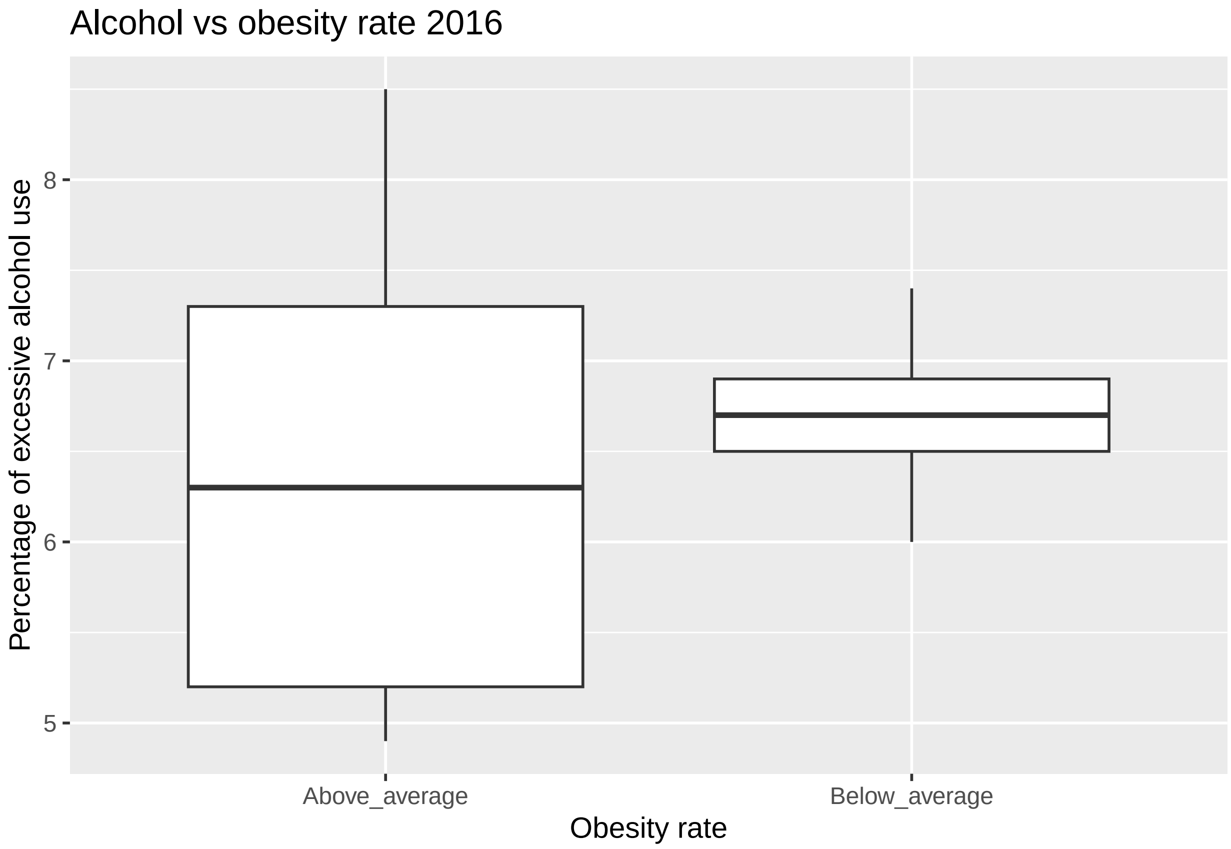
### 3.5 Visualize sub-population variation

The boxplot is relevant to our specific social problem since it gives an easy and clear overview of the percentage of excessive alcohol use in combination with above or below average obesity rates in Krimpenerwaard and Rotterdam. Our plot is mainly important for clearly stating whether there is a strong correlation between heavy drinkers and obesity rates within an average neighborhood in 2016.

```r
library(tidyverse)

mean(as.numeric(merged_2016$Onder.en.overgewicht.Mate.van.overgewicht.Ernstig.overgewicht....), na.rm =
```

```
## [1] NA
```

```r
#subgroup
merged_2016 = mutate(merged_2016, Obesity_rate = ifelse(Onder..en.overgewicht.Mate.van.overgewicht.Erns
                                                   "Above_average", "Below_average"))

write.csv(merged_2016, "Data/data/merged_2016.csv")
merged_2016$Alcoholgebruik.Overmatige.drinker.... <- merged_2016$Alcoholgebruik.Overmatige.drinker....

#plotting
ggplot(merged_2016, aes(x=Obesity_rate, y=Alcoholgebruik.Overmatige.drinker....))+
  geom_boxplot() +
  labs(x="Obesity rate", y="Percentage of excessive alcohol use", title = "Alcohol vs obesity rate 2016
```

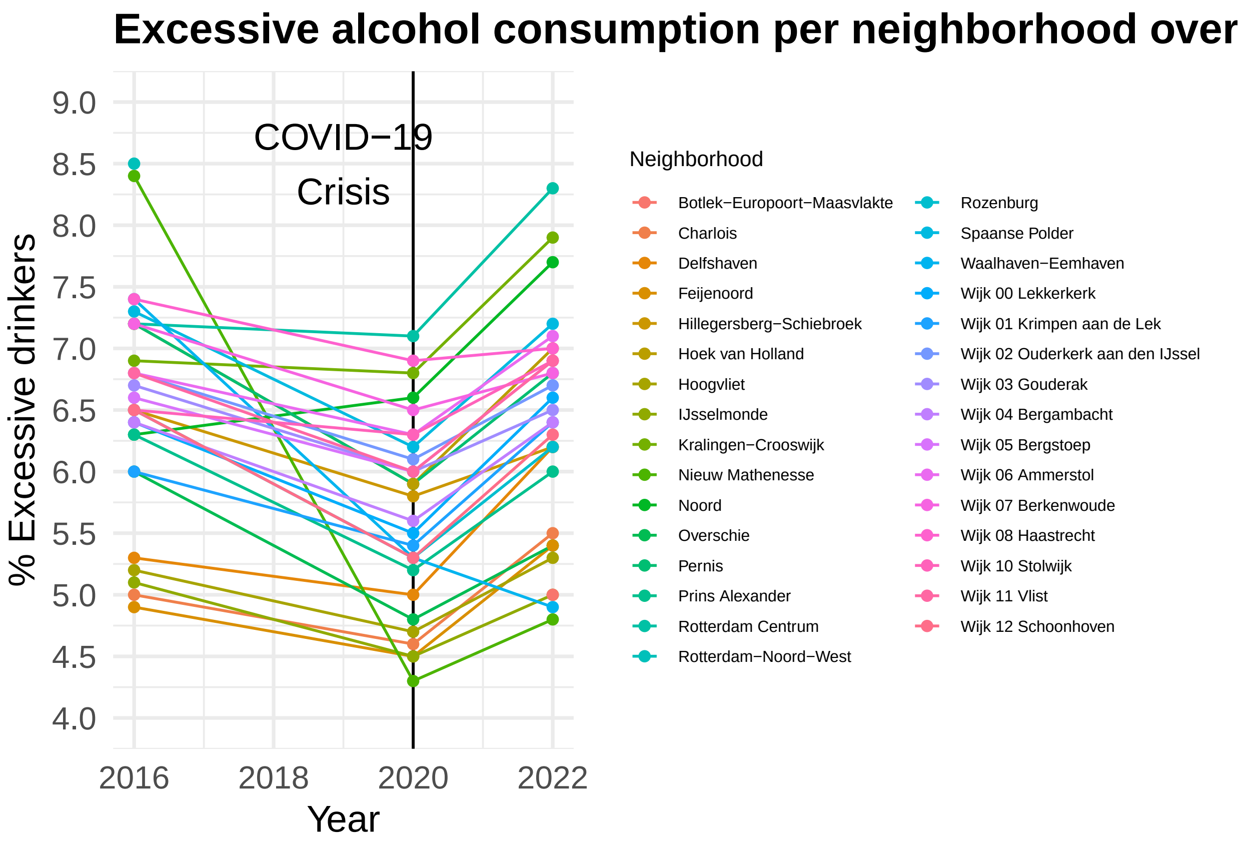


### 3.6 Event analysis

The graph gives us the relationship between the years 2016, 2018, 2020, 2022, and the percentage of excessive alcohol consumption per neighbourhood over this period of time. Each colored line represents a different neighbourhood, and the combination of all these variables gives a somewhat consistent outcome for each neighbourhood. The vertical line is representative of the “COVID-19 Crisis”, since this crisis would be the cause of a noticeable decline in excessive drinking from 2016 to 2020. The graph also shows that after COVID-19, the amount of excessive drinkers returned to their original state, and for some neighbourhoods, even ended up higher than before COVID-19.

This graph would be relevant to our research on alcohol abuse and access to local services, since most of these local services where you would be able to drink alcohol were closed during COVID-19, so this shows the strong correlation that these variables ultimately have, regardless of the distance between residents and their local services such as bars and café's.

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

wijken_data <- merged_data %>%
  filter(Regioaanduiding.Soort.regio..omschrijving..x == "Wijk")
wijken_data <- wijken_data %>%
  filter(Alcoholgebruik.Overmatige.drinker.... != ".")


wijken_data$Alcoholgebruik.Overmatige.drinker.... <-
  as.numeric(gsub(",", ".", wijken_data$Alcoholgebruik.Overmatige.drinker....))


  library(ggplot2)

ggplot(wijken_data, aes(
    x = Perioden,
    y = Alcoholgebruik.Overmatige.drinker....,
    group = Wijken.en.buurten.x,
    color = Wijken.en.buurten.x
  )) +
    geom_line() +
  geom_vline(xintercept = 2020) +
  annotate("text", x =2019, y=8.5,size = 5, label="COVID-19\nCrisis") +
    geom_point(size = 1.5) +
    scale_y_continuous(breaks = seq(4, 9, 0.5), limits = c(4, 9)) +
    labs(
      title = "Excessive alcohol consumption per neighborhood over time",
      x = "Year",
      y = "% Excessive drinkers",
      color = "Neighborhood"
    ) +
    theme_minimal(base_size = 14) +
    theme(
      legend.text = element_text(size = 6),
      legend.title = element_text(size = 8),
      legend.key.size = unit(0.4, "cm"),
      plot.title = element_text(size = 16, face = "bold"),
      axis.text = element_text(size = 12)
    )
```



# Part 4 - Discussion

## 4.1 Discuss your findings

We will now discuss the findings per graph we made, since they are all so different from each other. Then, we will combine these findings to find a conclusion corresponding to all graphs.

**3.3 visualize temporal variation**
The key observation from this graph is that Rotterdam consistently has a higher healthiness score than Krimpenerwaard in all years. Also, it is very clear that from both places the unhealthiness scores dropped in 2020. This could indicate a connection with the covid-19 crisis and the isolation that people were in, which made them live healthier. Both places are showing a U-shaped trend, meaning that unhealthiness decreased from 2016 to 2020 and then increased again by 2022. However, the gap between the both of them remains consistent, with Rotterdam showing higher levels of the unhealthiness score across all those time points. This could suggest that factors related to urban environments (as in Rotterdam) are associated with higher health risk indicators as in the unhealthiness score such as: smoking, heavy drinking, chronic illness, mental health issues and loneliness.

**3.4 visualize spatial variation**
In Rotterdam, which is on the left side of the map, you can see that most of the neighborhoods are colored red, meaning cafés are generally very accessible. That fits with what we'd expect in a large, urban city like Rotterdam. Comparing this to Krimpenerwaard, on the right side of the map, there are more blue areas. This means there is more rural character and thus a greater distance to cafés. This difference between areas supports further analysis on how the distance to cafés might relate to excessive alcohol use, as previously discussed.

**3.5 Visualize sub-population variation**
Since each box in the boxplot of part 3.5 represents the distribution of excessive alcohol usage within the above-average and below-average obesity rates, we can conclude from this graph that the above-average group shows higher variation in excessive alcohol usage. The group with below-average obesity rates has a tighter distribution of excessive alcohol usage, but has more outliers in both directions of the boxplot. This tells us that areas with higher obesity rates may also tend to have more variation in alcohol behaviour, and since both alcohol usage and obesity rates are major public health concerns, this graph would be especially useful for our problem.

**3.6 Event analysis**
The graph of 3.6 shows heavy drinking dropped during the 2020 COVID-19 pandemic in every neighborhood, perhaps due to lockdowns and limited nightlife availability, such as access to local services such as bars and café'e. However, by 2022, consumption levels had returned in most neighborhoods, especially cities. This reflects a return to pre-pandemic behaviors and the continuation of regional differences in alcohol consumption.

**Combined conclusions**

1. Urban areas show higher excessive alcohol risk - Rotterdam consistently shows a higher unhealthiness score and a closer distance to facilities than Krimpenerwaard. These factors suggest that urban areas may have higher levels of excessive alcohol consumption. Living in a city might make it easier or more tempting to drink more often or in larger amounts, simply because there are more cafés, bars, and social opportunities around.

2. COVID-19 temporarily reduced excessive alcohol use - In both urban and rural areas, there was a noticeable drop in heavy drinking during 202. This could likely be due to the pandemic where lockdowns and limited access to nightlife and cafés were going on. However, by 2022, alcohol use went up again and showed a return to a more pre-pandemic behavior. This could indicate that changes were temporary, not structural

3. Health disparities highlight at-risk-populations - The sub-population analysis reveals that areas with higher obesity rates also have more variation in excessive alcohol use. This may suggest that excessive drinking could occur with other health issues.

4. Persistent regional gaps in alcohol use - Even though drinking patterns changed over time, the difference between Rotterdam and Krimpenerwaard stayed about the same. This suggests that local factors like culture, the distance to cafés etc, and whether an area is more urban or rural, play a role in shaping how much people drink.

# Part 5 - Reproducibility

## 5.1 Github repository link

Provide the link to your PUBLIC repository here: https://github.com/Tombo14/School_Opdracht

## 5.2 Reference list

**Rijksinstituut voor Volksgezondheid en Milieu (RIVM). (n.d.).** *Buurtatlas – Overmatige drinkers.* RIVM / Buurtatlas. Retrieved June 10, 2025, from https://buurtatlas.vzinfo.nl/#overmatige_drinkers

**Rijksinstituut voor Volksgezondheid en Milieu (RIVM). (2023).** *Gezondheid per wijk en buurt; indeling 2016.* RIVM StatLine. Retrieved June 11, 2025, from https://statline.rivm.nl/#/RIVM/nl/dataset/50120NED/table?dl=C2E97 thuas.com+5data.overheid.nl+5statline.rivm.nl+5

**Rijksinstituut voor Volksgezondheid en Milieu (RIVM). (2023).** *Gezondheid per wijk en buurt; indeling 2020.* RIVM StatLine. Retrieved June 11, 2025, from https://statline.rivm.nl/#/RIVM/nl/dataset/50120NED/table?dl=C2E95

**Rijksinstituut voor Volksgezondheid en Milieu (RIVM). (2023).** *Gezondheid per wijk en buurt; indeling 2022.* RIVM StatLine. Retrieved June 11, 2025, from https://statline.rivm.nl/#/RIVM/nl/dataset/50120NED/table?dl=C2E96

**SA Health. (n.d.)**. *Health, safety, legal and social consequences of drinking too much.* Retrieved June 10, 2025, from https://www.sahealth.sa.gov.au/wps/wcm/connect/public+content/sa+health+internet/conditions/alcohol/health+safety+legal+and+social+consequences+of+drinking+too+much

**Statistics Netherlands (CBS). (n.d.).** *Nabijheid voorzieningen; afstand locatie, regionale cijfers.* Statistics Netherlands. Retrieved Retrieved June 10, 2025, from https://opendata.cbs.nl/statline/#/CBS/nl/dataset/80305ned/table?dl=C2E93