

Template

Studentnames and studentnumbers here

2025-06-11

Set-up your environment

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ----- tidyverse 2.0.0 --
## v dplyr      1.1.4      v readr      2.1.5
## v forcats    1.0.0      v stringr   1.5.1
## v ggplot2    3.5.2      v tibble    3.3.0
## 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 errors
```

Title Page

Include your names

Include the tutorial group number

Include your tutorial lecturer's name

Part 1 - Identify a Social Problem

Use APA referencing throughout your document. Here's a link to some explanation.

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)

```
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>, percamerindian <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.

```
inline_code = TRUE
```

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.

```
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

Variable 2

3.3 Visualize temporal variation

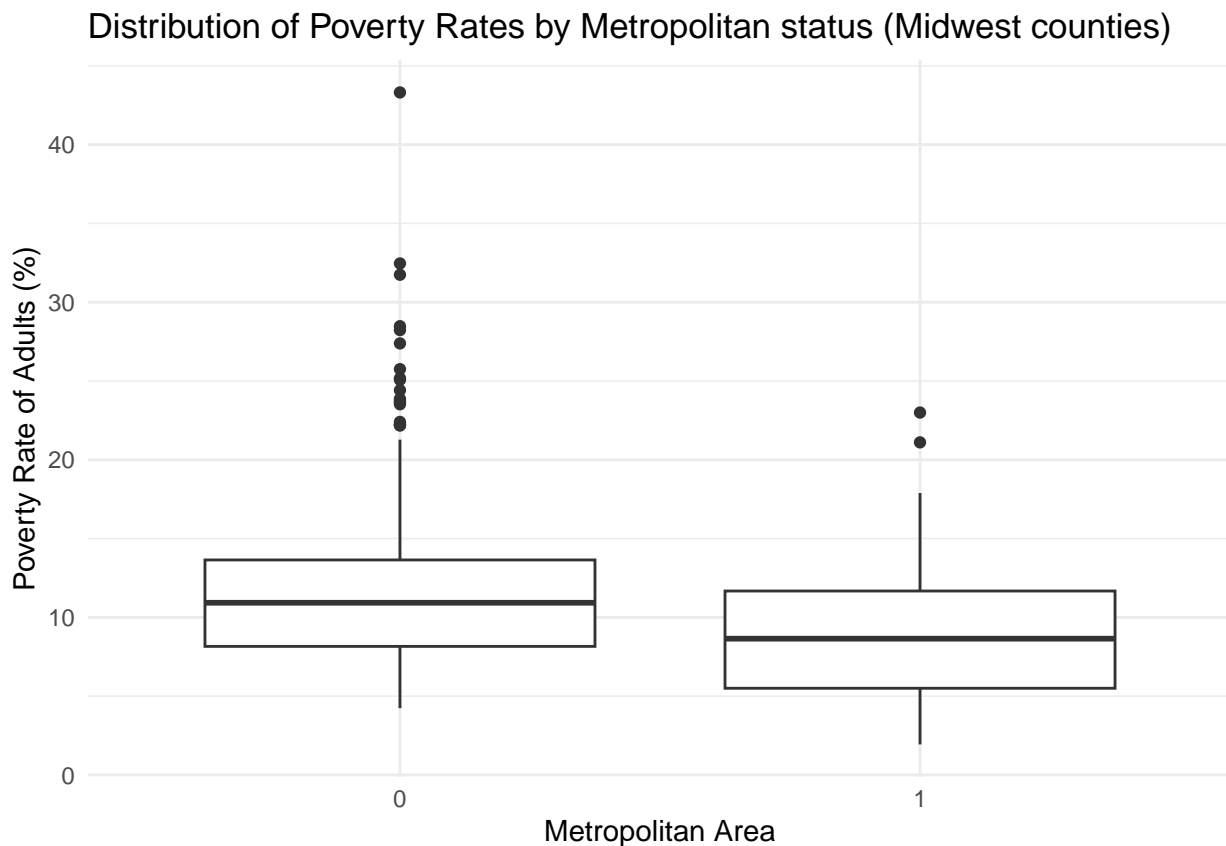
3.4 Visualize spatial variation

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?

```
dataset$inmetro <- dataset$inmetro %>% as.factor()  
# Boxplot of poverty rate by state using the 'midwest' dataset  
ggplot(dataset, aes(x = inmetro, y = percadultpoverty)) +  
  geom_boxplot() +  
  labs(  
    title = "Distribution of Poverty Rates by Metropolitan status (Midwest counties)",  
    x = "Metropolitan Area",  
    y = "Poverty Rate of Adults (%)"  
  ) +  
  theme_minimal() +  
  theme(  
    legend.position = "right"  
  )
```



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: ...

5.2 Reference list

Use APA referencing throughout your document.