Mental Health Inequality Across the Globe: How Economic and Social Disparities Shape Psychological Well-Being

Alona Sychevska, 2769408

June 5, 2025

```r
library(psych)
library(GGally)
library(readr)
library(dplyr)
library(ggplot2)
library(tidyr)
library(scales)
library(WDI)
library(stringr)
library(rnaturalearth)
library(rnaturalearthdata)
library(sf)
library(readxl)
library(naniar)
library(plotly)
library(reshape2)
library(RColorBrewer)
library(e1071)
library(ggrepel)
```

Mental Health Inequality Across the Globe: How Economic and Social Disparities Shape Psychological Well-Being

Alona Sychevska

Tutorial lecturer's name: Chantal Schouwenaar, Jack Fitzgerald

1 Problem Motivation & Literature

Why It's a Recognized Social Problem

Mental health issues such as depression and anxiety are on the rise globally—not only in high-income countries like the Netherlands or the UK, but also across low- and middle-income countries (LMICs). According to the WHO, mental disorders are among the leading causes of disability worldwide, and their burden is projected to grow (World Health Organization, 2023).

A growing body of literature suggests that **economic inequality**, rather than just poverty or absolute income, is a key social determinant of mental health. Studies have found that countries with higher levels of income inequality tend to have higher rates of psychological distress, even after controlling for GDP per capita and other macroeconomic indicators (Wilkinson & Pickett, 2009).

A systematic review of 26 studies—primarily from high-income countries—found that greater income inequality is associated with a 19% increased risk of depression (Patel et al., 2018). While most of this literature focuses on wealthy countries, studies in LMICs have also confirmed that poverty and inequality significantly predict common mental disorders (Lund et al., 2010).

More recent cross-continental studies have demonstrated that income inequality not only exacerbates anxiety and depression in high-income countries, but also in emerging and developing economies, reinforcing the need for a more globally inclusive analysis (Patel et al., 2022).

Theoretical Framing & Mechanisms

Several mechanisms help explain how economic inequality may drive mental health problems:

- **Relative deprivation**: People evaluate their well-being not in absolute terms but in comparison to others. This social comparison can lead to chronic stress and lower self-worth, especially in unequal societies (Smith et al., 2012).
- **Social capital erosion**: Inequality weakens social cohesion and trust, contributing to social exclusion and loneliness, which are known risk factors for mental illness (Kawachi & Berkman, 2000).
- **Bidirectional poverty–mental health link**: Mental illness can both result from and reinforce poverty, creating a feedback loop that is particularly severe in LMICs (Lund et al., 2010).

Gap in the Literature

Most cross-national studies on mental health and inequality focus on wealthy countries or single-region datasets. Far fewer include **LMICs**, despite the fact that these countries now bear a disproportionately large share of the global mental health burden.

Moreover, few studies consider a **broad set of economic, environmental, and social variables together**, such as:

- Inequality (Gini, wealth concentration)
- Housing cost burden
- Urbanization rate
- Government spending
- Air pollution (PM2.5)

This project combines all of these, providing a richer understanding of **how multiple inequality-related factors relate to mental health outcomes** across different economic contexts. It uses cross-national data covering income groups from the World Bank and mental health data from the Global Burden of Disease project.

---

2 Data Sourcing & Description

Data Sources and Credibility

This project integrates two key datasets: the Global Burden of Disease (GBD) Study 2021 and a multi-source datasets compiled from the World Bank (WDI), WHO, and other internationally reputable institutions.

The GBD Study is produced by the Institute for Health Metrics and Evaluation (IHME) at the University of Washington. It is considered the gold standard for global health statistics and is widely cited by academic institutions and policymakers worldwide (IHME, 2022).

The merged compilation of datasets drawing from sources like the World Bank's World Development Indicators (WDI), WHO's Air Quality Database, and World Population Review. These are recognized institutions known for robust methodologies and transparency. WDI, in particular, is frequently used in cross-country empirical studies due to its consistency and broad temporal coverage.

Metadata Overview

Dataset 1: Global Burden of Disease Study (GBD 2021)

Citation: Global Burden of Disease Collaborative Network. GBD 2021 Results. Seattle, WA: IHME, 2022. Available from https://vizhub.healthdata.org/gbd-results/

Metadata:

Variables:

- location (country)
- cause (e.g. anxiety, depression)
- measure_name (e.g. DALYs, Deaths, Prevalence)
- val_mental (absolute burden), disorder_rate (percent)
- year, population, iso3c

Structure: Long-format panel data; each row = country-cause-year combo

Timeframe: Includes historical time series from 2012.

Unit of observation: Country-year-cause.

Multi-source datasets: World Bank & World Population Review–Compiled Dataset (Merged Socioeconomic Indicators)

Sources:

- World Bank. (n.d.). *World Development Indicators (WDI)* – GDP per capita, urbanization, population. World Bank WDI
- World Health Organization. (2022). *WHO Air Quality Database 2022.* WHO Air Quality
- World Bank. (n.d.). *Poverty and Inequality Platform: Gini index.* Gini Index – PIP
- World Bank. (n.d.). *Income share held by highest 10% (SI.DST.10TH.10).* Top 10% Income Share
- Transparency International. (2020). *Corruption Perceptions Index (CPI) 2020.* CPI 2020

- World Bank. (n.d.). *House price to income ratio (IMF Global Housing Watch).* House Price to Income – World Bank

- OECD. (n.d.). *OECD house price statistics.* OECD House Prices

- International Monetary Fund. (n.d.). *Government expenditure, percent of GDP.* Government Expenditure – IMF

- Qery. (n.d.).
- Unemployment in OECD countries. OECD Unemployment – Qery

Metadata:

- **Structure:** Cross-sectional dataset using the latest available year per country (from 2012).

- **Units:** Mixed units including percentages, index scores, and GDP in USD.

Variables:

- **Inequality:**
  - `gini_index` – Gini coefficient of income inequality
  - `wealth_share_10` – Share of income held by the top 10%
- **Economic:**
  - `gdp_per_capita` – GDP per capita (USD)
  - `income_grp` – World Bank income group classification
  - `employment_rate` – Percent of working-age population employed
  - `gov_spending_to_GDP_percent` – Government spending as % of GDP
- **Environmental:**
  - `PM2.5` – Annual mean exposure to fine particulate matter (μg/m^3)
- **Social:**
  - `urban_pct` – Urban population (% of total)
  - `housing_cost_toincome` – Ratio of housing costs to income
  - `mortgage_to_income` – Ratio of mortgage payments to income
  - `corruption_index` – Transparency International's CPI score

Complementarity of the Two Datasets

GBD offers high-quality health burden metrics (Prevalence), crucial for understanding the impact of mental health disorders, while the merged dataset offers explanatory variables on inequality, economic performance, environmental exposure, and housing. These datasets are complementary in structure and purpose: the GBD gives outcome measures, while the merged dataset gives potential predictors. Their integration allows for robust modeling of social determinants of mental health at the national level.

While I initially collected more than two datasets, the final dataset has complementary strengths and ability to support the causality-oriented focus of this project: exploring how inequality and structural factors relate to mental health burdens across countries.

Relevance to the Topic

The project's focus is on the causal relationship between inequality and mental health in low-, middle-, and high-income countries. The selected datasets are ideal because:

The GBD dataset provides mental health burden metrics (e.g., Prevalence of depression or anxiety), which are essential for measuring the societal impact of mental disorders across countries.

The merged dataset allows for correlational and regression analyses of how inequality indicators (e.g., Gini index, top 10% wealth share, housing burden) , wealth, economic performance, and environmental factors relate to mental health prevalence.

Both datasets cover a wide range of countries, including those outside the high-income bracket, which aligns with the study's comparative angle.

Limitations of the Data

Despite their strengths, datasets have limitations:

GBD data, though standardized, aggregate country-level estimates, potentially masking subnational disparities and cultural variation in diagnosis/reporting.

The multi-sourced data suffers from missing data for certain indicators in low-income countries, which could bias regression results or reduce sample size.

Differences in data collection years (some indicators are 2021, others 2022 or 2023) may introduce temporal misalignment.

Mental health prevalence estimates from sources like World Population Review may lack the methodological rigor of epidemiological surveys.

Self-reported or perception-based indicators (e.g., corruption, employment satisfaction) could carry subjective bias.

2.1 Load in the data

```
GBD <- read_csv("../data/Global-Burden-of-Disease-Study.csv")
#GBD <- read_csv("../data/Global-Burden-of-Disease-Study/IHME-GBD_2021_DATA-3c361732-1.csv")

gni_by_country_year <- read_excel("../data/gni_by_country_year.xlsx")
economic_inequality_gini_index <- read_csv("../data/economic-inequality-gini-index/economic-inequality-g
corruption_index <- read_excel("../data/corruption_index.xlsx")
housing_cost_over_income <- read_csv("../data/house_price_to_income.csv")
gov_spending_toGDP <-read_excel("../data/gov_spending_toGDP.xlsx")
population <- read_csv("../data/population.csv")
air_polution <- read_csv("../data/air_polution_who.csv")
uneml_rate <- read_csv("../data/unemployment_rate.csv")
percent_wealth_10 <- read_csv("../data/10percent_wealth/Income share held by highest 10%.csv")


world <- ne_countries(scale = "medium", returnclass = "sf")

# WDI data
urban_data <- WDI(
  country = "all",
```

```r
  indicator = "SP.URB.TOTL.IN.ZS",
  start = 2012,
  end = 2025
) %>%
  rename(urban_pct = SP.URB.TOTL.IN.ZS)

# Pull GDP per capita data
  gdp_data <- WDI(
  country = "all",
  indicator = "NY.GDP.PCAP.CD",  # GDP per capita (current US$)
  start = 2012,
  end = 2025
)
```

2.2 Checking the structure

```r
# checking structure of dataset gini
dim(economic_inequality_gini_index)
```

```
## [1] 2285    5
```

```r
names(economic_inequality_gini_index)
```

```
## [1] "Entity"
## [2] "Code"
## [3] "Year"
## [4] "Gini coefficient (2017 prices) - Income or consumption consolidated"
## [5] "1039568-annotations"
```

```r
head(economic_inequality_gini_index)
```

```
## # A tibble: 6 x 5
##   Entity  Code   Year Gini coefficient (2017 prices) - I~1 `1039568-annotations`
##   <chr>   <chr> <dbl>                                <dbl> <lgl>
## 1 Albania ALB    1996                                0.270 NA
## 2 Albania ALB    2002                                0.317 NA
## 3 Albania ALB    2005                                0.306 NA
## 4 Albania ALB    2008                                0.300 NA
## 5 Albania ALB    2012                                0.290 NA
## 6 Albania ALB    2014                                0.346 NA
## # i abbreviated name:
## #   1: `Gini coefficient (2017 prices) - Income or consumption consolidated`
```

```r
# dropping cols from urban data set
head(urban_data)
```

```
##                         country iso2c iso3c year urban_pct
## 1 Africa Eastern and Southern    ZH   AFE 2024        NA
## 2 Africa Eastern and Southern    ZH   AFE 2023  38.42490
```

```
## 3 Africa Eastern and Southern    ZH   AFE 2022  37.90901
## 4 Africa Eastern and Southern    ZH   AFE 2021  37.39363
## 5 Africa Eastern and Southern    ZH   AFE 2020  36.88403
## 6 Africa Eastern and Southern    ZH   AFE 2019  36.38427
```

```r
colSums(is.na(urban_data))
```

```
##   country     iso2c     iso3c      year urban_pct
##         0         0         0         0       302
```

```r
head(corruption_index)
```

```
## # A tibble: 6 x 32
##   Country     `CPI score 2020` `Rank 2020` `Sources 2020` `Standard error 2020`
##   <chr>                  <dbl>       <dbl>          <dbl>                 <dbl>
## 1 Denmark                   88           1              8                  1.78
## 2 New Zealand               88           1              8                  1.48
## 3 Finland                   85           3              8                  1.75
## 4 Singapore                 85           3              9                  1.20
## 5 Sweden                    85           3              8                  1.30
## 6 Switzerland               85           3              7                  1.10
## # i 27 more variables: `CPI score 2019` <dbl>, `Rank 2019` <dbl>,
## #   `Sources 2019` <dbl>, `Standard error 2019` <dbl>, `CPI score 2018` <dbl>,
## #   `Rank 2018` <dbl>, `Sources 2018` <dbl>, `Standard error 2018` <dbl>,
## #   `CPI score 2017` <dbl>, `Rank 2017` <dbl>, `Sources 2017` <dbl>,
## #   `Standard error 2017` <dbl>, `CPI score 2016` <dbl>, `Sources 2016` <dbl>,
## #   `Standard error 2016` <dbl>, `CPI score 2015` <dbl>, `Sources 2015` <dbl>,
## #   `Standard error 2015` <dbl>, `CPI score 2014` <dbl>, ...
```

```r
colnames(corruption_index)
```

```
##  [1] "Country"             "CPI score 2020"      "Rank 2020"
##  [4] "Sources 2020"        "Standard error 2020" "CPI score 2019"
##  [7] "Rank 2019"           "Sources 2019"        "Standard error 2019"
## [10] "CPI score 2018"      "Rank 2018"           "Sources 2018"
## [13] "Standard error 2018" "CPI score 2017"      "Rank 2017"
## [16] "Sources 2017"        "Standard error 2017" "CPI score 2016"
## [19] "Sources 2016"        "Standard error 2016" "CPI score 2015"
## [22] "Sources 2015"        "Standard error 2015" "CPI score 2014"
## [25] "Sources 2014"        "Standard error 2014" "CPI Score 2013"
## [28] "Sources 2013"        "Standard error 2013" "CPI Score 2012"
## [31] "Sources 2012"        "Standard error 2012"
```

```r
head(gdp_data)
```

```
##                       country iso2c iso3c year NY.GDP.PCAP.CD
## 1 Africa Eastern and Southern    ZH   AFE 2024             NA
## 2 Africa Eastern and Southern    ZH   AFE 2023       1659.515
## 3 Africa Eastern and Southern    ZH   AFE 2022       1628.025
## 4 Africa Eastern and Southern    ZH   AFE 2021       1522.590
## 5 Africa Eastern and Southern    ZH   AFE 2020       1344.081
## 6 Africa Eastern and Southern    ZH   AFE 2019       1493.780
```

```r
colSums(is.na(gdp_data))
```

```
##        country          iso2c          iso3c           year NY.GDP.PCAP.CD
##              0              0              0              0            371
```

```r
dim(GBD)
```

```
## [1] 41472    16
```

```r
head(GBD)
```

```
## # A tibble: 6 x 16
##   measure_id measure_name location_id location_name sex_id sex_name age_id
##        <dbl> <chr>              <dbl> <chr>          <dbl> <chr>     <dbl>
## 1          1 Deaths                85 Israel             3 Both         22
## 2          1 Deaths                85 Israel             3 Both         22
## 3          1 Deaths                85 Israel             3 Both         22
## 4          1 Deaths                36 Kazakhstan         3 Both         22
## 5          1 Deaths                36 Kazakhstan         3 Both         22
## 6          1 Deaths                36 Kazakhstan         3 Both         22
## # i 9 more variables: age_name <chr>, cause_id <dbl>, cause_name <chr>,
## #   metric_id <dbl>, metric_name <chr>, year <dbl>, val <dbl>, upper <dbl>,
## #   lower <dbl>
```

```r
head(housing_cost_over_income)
```

```
## # A tibble: 6 x 26
##   STRUCTURE STRUCTURE_ID   STRUCTURE_NAME ACTION REF_AREA `Reference area` FREQ
##   <chr>     <chr>          <chr>          <chr>  <chr>    <chr>            <chr>
## 1 DATAFLOW  OECD.ECO.MPD:~ Analytical ho~ I      OECD     OECD             A
## 2 DATAFLOW  OECD.ECO.MPD:~ Analytical ho~ I      OECD     OECD             A
## 3 DATAFLOW  OECD.ECO.MPD:~ Analytical ho~ I      OECD     OECD             A
## 4 DATAFLOW  OECD.ECO.MPD:~ Analytical ho~ I      OECD     OECD             A
## 5 DATAFLOW  OECD.ECO.MPD:~ Analytical ho~ I      OECD     OECD             A
## 6 DATAFLOW  OECD.ECO.MPD:~ Analytical ho~ I      OECD     OECD             A
## # i 19 more variables: `Frequency of observation` <chr>, MEASURE <chr>,
## #   Measure <chr>, UNIT_MEASURE <chr>, `Unit of measure` <chr>,
## #   TIME_PERIOD <dbl>, `Time period` <lgl>, OBS_VALUE <dbl>,
## #   `Observation value` <lgl>, OBS_STATUS <chr>, `Observation status` <chr>,
## #   UNIT_MULT <dbl>, `Unit multiplier` <chr>, ADJUSTMENT <chr>,
## #   Adjustment <chr>, DECIMALS <dbl>, Decimals <chr>, BASE_PER <dbl>,
## #   `Base period` <lgl>
```

```r
colnames(housing_cost_over_income)
```

```
##  [1] "STRUCTURE"                "STRUCTURE_ID"
##  [3] "STRUCTURE_NAME"           "ACTION"
##  [5] "REF_AREA"                 "Reference area"
##  [7] "FREQ"                     "Frequency of observation"
##  [9] "MEASURE"                  "Measure"
```

```
## [11] "UNIT_MEASURE"             "Unit of measure"
## [13] "TIME_PERIOD"              "Time period"
## [15] "OBS_VALUE"                "Observation value"
## [17] "OBS_STATUS"               "Observation status"
## [19] "UNIT_MULT"                "Unit multiplier"
## [21] "ADJUSTMENT"               "Adjustment"
## [23] "DECIMALS"                 "Decimals"
## [25] "BASE_PER"                 "Base period"
```

```r
colSums(is.na(housing_cost_over_income))
```

```
##                STRUCTURE              STRUCTURE_ID            STRUCTURE_NAME
##                        0                         0                         0
##                   ACTION                  REF_AREA            Reference area
##                        0                         0                         0
##                     FREQ Frequency of observation                   MEASURE
##                        0                         0                         0
##                  Measure              UNIT_MEASURE           Unit of measure
##                        0                         0                         0
##              TIME_PERIOD               Time period                 OBS_VALUE
##                        0                       448                         0
##        Observation value                OBS_STATUS        Observation status
##                      448                         0                         0
##                UNIT_MULT           Unit multiplier                ADJUSTMENT
##                        0                         0                         0
##               Adjustment                  DECIMALS                  Decimals
##                        0                         0                         0
##                 BASE_PER               Base period
##                        0                       448
```

```r
colSums(is.na(air_polution))
```

```
##                              WHO Region                                     ISO3
##                                       1                                        0
##                        WHO Country Name                        City or Locality
##                                       0                                        0
##                        Measurement Year                            PM2.5 ( g/m3)
##                                       0                                    17143
##                            PM10 ( g/m3)                             NO2 ( g/m3)
##                                   11082                                     9991
##               PM25 temporal coverage (%)               PM10 temporal coverage (%)
##                                   24916                                    26810
##                NO2 temporal coverage (%)                                Reference
##                                   12301                                        5
## Number and type of monitoring stations                  Version of the database
##                                   23433                                        0
##                                  Status
##                                   32191
```

```r
head(air_polution)
```

```
## # A tibble: 6 x 15
```

```
##   `WHO Region`    ISO3  `WHO Country Name` `City or Locality` `Measurement Year`
##   <chr>           <chr> <chr>              <chr>                           <dbl>
## 1 Eastern Medite~ AFG   Afghanistan        Kabul                            2019
## 2 European Region ALB   Albania            Durres                           2015
## 3 European Region ALB   Albania            Durres                           2016
## 4 European Region ALB   Albania            Elbasan                          2015
## 5 European Region ALB   Albania            Elbasan                          2016
## 6 European Region ALB   Albania            Elbasan                          2017
## # i 10 more variables: `PM2.5 ( g/m3)` <dbl>, `PM10 ( g/m3)` <dbl>,
## #   `NO2 ( g/m3)` <dbl>, `PM25 temporal coverage (%)` <dbl>,
## #   `PM10 temporal coverage (%)` <dbl>, `NO2 temporal coverage (%)` <dbl>,
## #   Reference <chr>, `Number and type of monitoring stations` <chr>,
## #   `Version of the database` <dbl>, Status <lgl>
```

```r
head(gni_by_country_year)
```

```
## # A tibble: 6 x 38
##   Country  `1987` `1988` `1989` `1990` `1991` `1992` `1993` `1994` `1995` `1996`
##   <chr>    <chr>  <chr>  <chr>  <chr>  <chr>  <chr>  <chr>  <chr>  <chr>  <chr>
## 1 Afghani~ L      L      L      L      L      L      L      L      L      L
## 2 Albania  ..     ..     ..     LM     LM     LM     L      L      L      LM
## 3 Algeria  UM     UM     LM     LM     LM     LM     LM     LM     LM     LM
## 4 America~ H      H      H      UM     UM     UM     UM     UM     UM     UM
## 5 Andorra  ..     ..     ..     H      H      H      H      H      H      H
## 6 Angola   ..     LM     LM     LM     LM     LM     LM     LM     L      L
## # i 27 more variables: `1997` <chr>, `1998` <chr>, `1999` <chr>, `2000` <chr>,
## #   `2001` <chr>, `2002` <chr>, `2003` <chr>, `2004` <chr>, `2005` <chr>,
## #   `2006` <chr>, `2007` <chr>, `2008` <chr>, `2009` <chr>, `2010` <chr>,
## #   `2011` <chr>, `2012` <chr>, `2013` <chr>, `2014` <chr>, `2015` <chr>,
## #   ...31 <chr>, ...32 <chr>, ...33 <chr>, ...34 <chr>, `2020` <chr>,
## #   `2021` <chr>, `2022` <chr>, `2023` <chr>
```

```r
head(percent_wealth_10)
```

```
## # A tibble: 6 x 69
##   `Country Name`  `Country Code` `Indicator Name` `Indicator Code` `1960` `1961`
##   <chr>           <chr>          <chr>            <chr>            <lgl>  <lgl>
## 1 Aruba           ABW            Income share he~ SI.DST.10TH.10   NA     NA
## 2 Africa Eastern~ AFE            Income share he~ SI.DST.10TH.10   NA     NA
## 3 Afghanistan     AFG            Income share he~ SI.DST.10TH.10   NA     NA
## 4 Africa Western~ AFW            Income share he~ SI.DST.10TH.10   NA     NA
## 5 Angola          AGO            Income share he~ SI.DST.10TH.10   NA     NA
## 6 Albania         ALB            Income share he~ SI.DST.10TH.10   NA     NA
## # i 63 more variables: `1962` <lgl>, `1963` <dbl>, `1964` <dbl>, `1965` <dbl>,
## #   `1966` <dbl>, `1967` <dbl>, `1968` <dbl>, `1969` <dbl>, `1970` <dbl>,
## #   `1971` <dbl>, `1972` <dbl>, `1973` <dbl>, `1974` <dbl>, `1975` <dbl>,
## #   `1976` <dbl>, `1977` <dbl>, `1978` <dbl>, `1979` <dbl>, `1980` <dbl>,
## #   `1981` <dbl>, `1982` <dbl>, `1983` <dbl>, `1984` <dbl>, `1985` <dbl>,
## #   `1986` <dbl>, `1987` <dbl>, `1988` <dbl>, `1989` <dbl>, `1990` <dbl>,
## #   `1991` <dbl>, `1992` <dbl>, `1993` <dbl>, `1994` <dbl>, `1995` <dbl>, ...
```

2.3 Data cleaning

```
economic_inequality_gini_index <- economic_inequality_gini_index %>%
  filter(Year >= 2012) %>%
  rename("Gini_index" = "Gini coefficient (2017 prices) - Income or consumption consolidated") %>%
  select(-Code, -`1039568-annotations`)

# Creating gini buckets
gini_cats <- economic_inequality_gini_index %>%
  mutate(gini_bucket = case_when(
    Gini_index < 0.25               ~ "Low",
    Gini_index >= 0.25 & Gini_index < 0.35 ~ "Moderate",
    Gini_index >= 0.35              ~ "High",
    TRUE                           ~ NA_character_
  ))

colSums(is.na(gini_cats))
```

```
##      Entity        Year  Gini_index gini_bucket
##           0           0           0           0
```

```
# dropping cols from urban data set
urban_data <- urban_data %>%
  select(-iso2c)
```

```
# selecting cols needed

colnames(corruption_index)
```

```
##  [1] "Country"              "CPI score 2020"      "Rank 2020"
##  [4] "Sources 2020"         "Standard error 2020" "CPI score 2019"
##  [7] "Rank 2019"            "Sources 2019"        "Standard error 2019"
## [10] "CPI score 2018"       "Rank 2018"           "Sources 2018"
## [13] "Standard error 2018" "CPI score 2017"      "Rank 2017"
## [16] "Sources 2017"         "Standard error 2017" "CPI score 2016"
## [19] "Sources 2016"         "Standard error 2016" "CPI score 2015"
## [22] "Sources 2015"         "Standard error 2015" "CPI score 2014"
## [25] "Sources 2014"         "Standard error 2014" "CPI Score 2013"
## [28] "Sources 2013"         "Standard error 2013" "CPI Score 2012"
## [31] "Sources 2012"         "Standard error 2012"
```

```
corruption_index <- corruption_index %>%
  select(Country, "CPI Score 2012", "CPI Score 2013",`CPI score 2014`, `CPI score 2015`, `CPI score 201

corruption_index <- corruption_index %>%
  pivot_longer(
    cols = starts_with("CPI score"),
    names_to = "Year",
    values_to = "CPI_score"
  ) %>%
  mutate(
```

```r
    Year = gsub("CPI score ", "", Year),          # Remove text to keep only the year
    Year = as.integer(Year)                       # Convert to integer if needed
  )
```

```
## Warning: There was 1 warning in `mutate()`.
## i In argument: `Year = as.integer(Year)`.
## Caused by warning:
## ! NAs introduced by coercion
```

```r
uneml_rate <- uneml_rate %>%
  pivot_longer(
    cols = matches("^\\d{4}$"),
    names_to = "year",
    values_to = "unemployment_rate"
  ) %>%
  mutate(year = as.integer(year)) %>%
  rename(Country = `Country Name`) %>%
  select(Country, year, unemployment_rate) %>%
  filter(year >=2012)
```

```r
gov_spending_toGDP <- gov_spending_toGDP %>%
  select(Country, starts_with("201")) %>%  # Select year columns, assuming they are named like "2014",
  mutate(across(where(is.character), ~na_if(., "no data")))

gov_spending_toGDP <- gov_spending_toGDP %>%
  pivot_longer(
    cols = matches("^\\d{4}$"),
    names_to = "year",
    values_to = "gov_spending_toGDP"
  ) %>%
  mutate(year = as.integer(year)) %>%
  filter(year >= 2012)
```

```r
# Select only the columns needed, no duplicates
gdp_data <- gdp_data %>%
  select(country, year, NY.GDP.PCAP.CD) %>%
  rename(gdp_per_capita = NY.GDP.PCAP.CD) %>%
  filter(year >= 2012)
```

```r
# renaming col names for housing cost over income data
housing_cost_over_income <- housing_cost_over_income %>%
    select(`Reference area`, TIME_PERIOD, OBS_VALUE) %>%
    rename(year = TIME_PERIOD, housing_CosttoIncome = OBS_VALUE) %>%
  filter(year >= 2012)
```

```r
air_polution <- air_polution %>%
  rename("PM2.5" = "PM2.5 ( g/m3)", "Country" = "WHO Country Name") %>%
  filter(!is.na(`Measurement Year`), !is.na(PM2.5)) %>%  # Remove rows without a year
  group_by(Country, `Measurement Year`) %>%               # Group by country and year
  summarise(across(where(is.numeric), ~ mean(.x, na.rm = TRUE)), .groups = "drop") %>%
  rename("year" = `Measurement Year`) %>%
```

```r
  select(Country, PM2.5, year) %>%
  filter(year>=2012)
```

```r
gni_by_country_year <- gni_by_country_year %>%
  rename("2016" = "...31", "2017" = "...32", "2018" = "...33", "2019" = "...34") %>%
  pivot_longer(
    cols = matches("^\\d{4}$"),
    names_to = "year",
    values_to = "income_group"
  ) %>%
  mutate(year = as.integer(year)) %>%
  select(Country, year, income_group)


unique(gni_by_country_year$income_group)
```

```
## [1] "L"   ".."  "LM"  "UM"  "H"   NA    "LM*"
```

```r
gni_by_country_year <- gni_by_country_year %>%
  mutate(income_group = ifelse(income_group == "..", NA, income_group)) %>%
  filter(year >= 2012)
```

```r
unique(population$Year)
```

```
##   [1] -10000  -9000  -8000  -7000  -6000  -5000  -4000  -3000  -2000  -1000
##  [11]      0    100    200    300    400    500    600    700    800    900
##  [21]   1000   1100   1200   1300   1400   1500   1600   1700   1710   1720
##  [31]   1730   1740   1750   1760   1770   1780   1790   1800   1801   1802
##  [41]   1803   1804   1805   1806   1807   1808   1809   1810   1811   1812
##  [51]   1813   1814   1815   1816   1817   1818   1819   1820   1821   1822
##  [61]   1823   1824   1825   1826   1827   1828   1829   1830   1831   1832
##  [71]   1833   1834   1835   1836   1837   1838   1839   1840   1841   1842
##  [81]   1843   1844   1845   1846   1847   1848   1849   1850   1851   1852
##  [91]   1853   1854   1855   1856   1857   1858   1859   1860   1861   1862
## [101]   1863   1864   1865   1866   1867   1868   1869   1870   1871   1872
## [111]   1873   1874   1875   1876   1877   1878   1879   1880   1881   1882
## [121]   1883   1884   1885   1886   1887   1888   1889   1890   1891   1892
## [131]   1893   1894   1895   1896   1897   1898   1899   1900   1901   1902
## [141]   1903   1904   1905   1906   1907   1908   1909   1910   1911   1912
## [151]   1913   1914   1915   1916   1917   1918   1919   1920   1921   1922
## [161]   1923   1924   1925   1926   1927   1928   1929   1930   1931   1932
## [171]   1933   1934   1935   1936   1937   1938   1939   1940   1941   1942
## [181]   1943   1944   1945   1946   1947   1948   1949   1950   1951   1952
## [191]   1953   1954   1955   1956   1957   1958   1959   1960   1961   1962
## [201]   1963   1964   1965   1966   1967   1968   1969   1970   1971   1972
## [211]   1973   1974   1975   1976   1977   1978   1979   1980   1981   1982
## [221]   1983   1984   1985   1986   1987   1988   1989   1990   1991   1992
## [231]   1993   1994   1995   1996   1997   1998   1999   2000   2001   2002
## [241]   2003   2004   2005   2006   2007   2008   2009   2010   2011   2012
## [251]   2013   2014   2015   2016   2017   2018   2019   2020   2021   2022
## [261]   2023   1555   1640   1785   1788
```

```r
population <- population %>%
  rename(name = Entity, pop_est = "Population (historical)", pop_year = Year) %>%
  select(name, pop_est, pop_year) %>%
  filter(pop_year >= 2012)
```

```r
#transform percent wealth held by Top 10% merge
wealth10_long <- percent_wealth_10 %>%
  pivot_longer(
    cols = matches("^\\d{4}$"),  # Select only year columns
    names_to = "year",
    values_to = "wealth_share_10"
  ) %>%
  mutate(
    year = as.integer(year)
  ) %>%
  filter(!is.na(wealth_share_10)) %>%
  group_by(`Country Name`) %>%
  rename(
    country = `Country Name`
  ) %>%
  select(country, wealth_share_10, year) %>%
  filter(year >= 2012)

head(wealth10_long)
```

```
## # A tibble: 6 x 3
## # Groups:   country [2]
##   country wealth_share_10  year
##   <chr>             <dbl> <int>
## 1 Angola             39.6  2018
## 2 Albania            22.9  2012
## 3 Albania            25.5  2014
## 4 Albania            24.8  2015
## 5 Albania            25    2016
## 6 Albania            24.6  2017
```

```r
#renaming col names
GBD <- GBD %>%
  rename(Entity = location_name, cause = cause_name, Year = year)
```

```r
GBD <- GBD %>%
  inner_join(population,
          by = c("Entity" = "name", "Year" = "pop_year"))

# Calculate rate per 100,000 and percent
GBD <- GBD %>%
  filter(metric_name == "Number", measure_name == "Prevalence") %>%
  mutate(
    val_m = round((val / pop_est) * 100000, 2),
    val_rate = format(val_m, big.mark = ",", scientific = FALSE),
    val_percent = (val / pop_est) * 100
  ) %>%
```

```r
  select(Entity, cause, Year, val_rate, val_percent, pop_est) %>%
  #pivot_wider(names_from = cause,
  #             values_from = val_rate) %>%
  filter(Year >= 2012)
```

2.4 Datasets merging

```r
# Mental + gini_cats data
merged_data <- GBD %>%
  inner_join(gini_cats, by = c("Entity" = "Entity", "Year" = "Year"))

merged_data <- merged_data %>%
  inner_join(gni_by_country_year,by = c("Entity" = "Country", "Year" = "year"))

merged_data <- merged_data %>%
  inner_join(gdp_data, by = c("Entity" = "country", "Year" = "year"))

# merge unemployment rate merge
merged_data <- merged_data %>%
  inner_join(uneml_rate, by = c("Entity" = "Country", "Year" = "year"))

# Merge wealth share data
merged_data <- merged_data %>%
  inner_join(wealth10_long, by = c("Entity" = "country", "Year" = "year"))

merged_data <- merged_data %>%
  inner_join(urban_data, by = c("Entity" = "country", "Year" = "year"))

###

# merge corruption index
merged_data <- merged_data %>%
  left_join(corruption_index, by = c("Entity" = "Country", "Year" = "Year"))

# Merge gov_spending to GDP %
merged_data <- merged_data %>%
  left_join(gov_spending_toGDP, by = c("Entity" = "Country", "Year" = "year"))

# merge housing_cost_over_income
merged_data <- merged_data %>%
  left_join(housing_cost_over_income, by = c("Entity" = "Reference area", "Year" = "year"))

#Merge airpolution
merged_data <- merged_data %>%
  left_join(air_polution, by = c("Entity" = "Country", "Year" = "year"))
```

```r
# Post-Merge Checks
summary(merged_data$urban_pct)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   32.78   66.31   77.45   75.31   85.60   98.12
```

```r
summary(merged_data$corruption_score)
```

```
## Warning: Unknown or uninitialised column: `corruption_score`.

## Length  Class   Mode
##      0   NULL   NULL
```

```r
summary(merged_data$wealth_share_10)
```

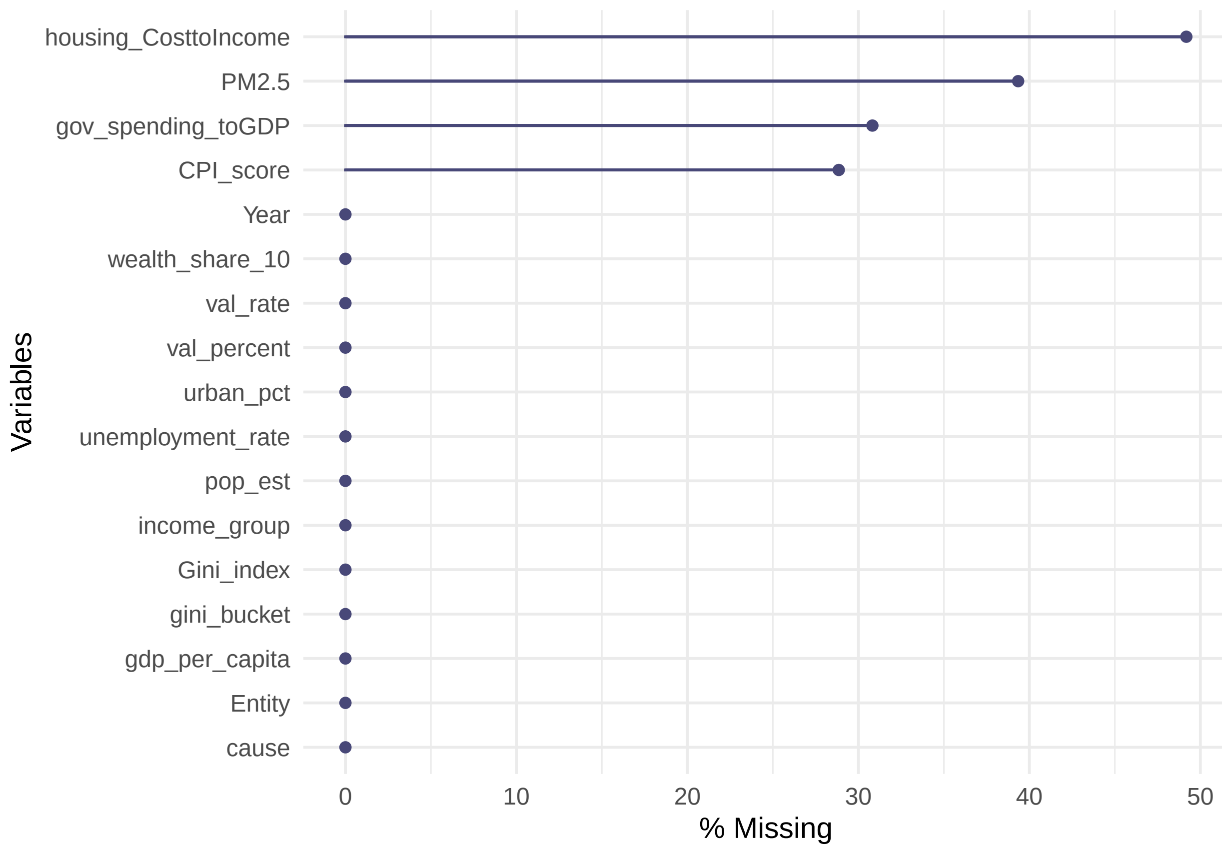
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    20.6    23.0    25.3    26.6    27.8    43.7
```

```r
### Dropping unnecessary cols
head(merged_data)
```

```
## # A tibble: 6 x 18
##   Entity cause          Year val_rate val_percent pop_est Gini_index gini_bucket
##   <chr>  <chr>         <dbl> <chr>          <dbl>   <dbl>      <dbl> <chr>
## 1 China  Schizophrenia  2012 "   366~       0.367  1.37e9      0.422 High
## 2 China  Schizophrenia  2013 "   366~       0.366  1.38e9      0.397 High
## 3 China  Schizophrenia  2014 "   365~       0.365  1.39e9      0.392 High
## 4 China  Schizophrenia  2015 "   364~       0.365  1.40e9      0.386 High
## 5 China  Schizophrenia  2016 "   365~       0.365  1.40e9      0.385 High
## 6 China  Schizophrenia  2017 "   365~       0.366  1.41e9      0.391 High
## # i 10 more variables: income_group <chr>, gdp_per_capita <dbl>,
## #   unemployment_rate <dbl>, wealth_share_10 <dbl>, iso3c <chr>,
## #   urban_pct <dbl>, CPI_score <dbl>, gov_spending_toGDP <dbl>,
## #   housing_CosttoIncome <dbl>, PM2.5 <dbl>
```

```r
merged_data <- merged_data %>%
  select(-iso3c)
```

```r
# Visualize missiness
gg_miss_var(merged_data, show_pct = TRUE)
```




```r
# short summaries of the final dataset
head(merged_data)
```

```
## # A tibble: 6 x 17
##   Entity cause          Year val_rate val_percent pop_est Gini_index gini_bucket
##   <chr>  <chr>         <dbl> <chr>          <dbl>   <dbl>      <dbl> <chr>
## 1 China  Schizophrenia  2012 "    366~        0.367  1.37e9      0.422 High
## 2 China  Schizophrenia  2013 "    366~        0.366  1.38e9      0.397 High
## 3 China  Schizophrenia  2014 "    365~        0.365  1.39e9      0.392 High
## 4 China  Schizophrenia  2015 "    364~        0.365  1.40e9      0.386 High
## 5 China  Schizophrenia  2016 "    365~        0.365  1.40e9      0.385 High
## 6 China  Schizophrenia  2017 "    365~        0.366  1.41e9      0.391 High
## # i 9 more variables: income_group <chr>, gdp_per_capita <dbl>,
## #   unemployment_rate <dbl>, wealth_share_10 <dbl>, urban_pct <dbl>,
## #   CPI_score <dbl>, gov_spending_toGDP <dbl>, housing_CosttoIncome <dbl>,
## #   PM2.5 <dbl>
```

2.5 Describe the type of variables included

```r
describe_variables <- function(df) {
  desc <- lapply(names(df), function(var) {
    col <- df[[var]]
    var_class <- class(col)
```

```r
    example_vals <- if (is.numeric(col)) {
      sprintf("mean = %.2f, sd = %.2f", mean(col, na.rm = TRUE), sd(col, na.rm = TRUE))
    } else if (is.factor(col) || is.character(col)) {
      vals <- unique(na.omit(col))
      paste("levels:", paste(head(vals, 5), collapse = ", "), if (length(vals) > 5) "...", collapse = "
    } else if (is.logical(col)) {
      "logical (TRUE/FALSE)"
    } else {
      paste("class:", var_class)
    }

    type_label <- if (is.numeric(col)) {
      "continuous"
    } else if (is.factor(col) || is.character(col)) {
      "categorical"
    } else if (is.logical(col)) {
      "logical"
    } else {
      "other"
    }

    paste0("- ", var, ": ", type_label, " (", var_class, "), ", example_vals)
  })

  cat(paste(unlist(desc), collapse = "\n"))
}
```

```r
describe_variables(merged_data)
```

```
## - Entity: categorical (character), levels: China, Mongolia, Belarus, Kazakhstan, Armenia ...
## - cause: categorical (character), levels: Schizophrenia, Attention-deficit/hyperactivity disorder, Id
## - Year: continuous (numeric), mean = 2016.37, sd = 2.82
## - val_rate: categorical (character), levels:    366.55,    366.01,    365.29,    364.93,    365.07 .
## - val_percent: continuous (numeric), mean = 2.78, sd = 4.25
## - pop_est: continuous (numeric), mean = 104628265.19, sd = 315184396.73
## - Gini_index: continuous (numeric), mean = 0.34, sd = 0.07
## - gini_bucket: categorical (character), levels: High, Moderate, Low
## - income_group: categorical (character), levels: UM, LM, H
## - gdp_per_capita: continuous (numeric), mean = 30698.76, sd = 24331.04
## - unemployment_rate: continuous (numeric), mean = 8.29, sd = 4.76
## - wealth_share_10: continuous (numeric), mean = 26.60, sd = 5.16
## - urban_pct: continuous (numeric), mean = 75.31, sd = 13.04
## - CPI_score: continuous (numeric), mean = 59.12, sd = 19.38
## - gov_spending_toGDP: continuous (numeric), mean = 42.56, sd = 10.15
## - housing_CosttoIncome: continuous (numeric), mean = 104.57, sd = 8.51
## - PM2.5: continuous (numeric), mean = 17.16, sd = 14.36
```

```r
write.csv(merged_data, "../data/merged_data.csv")
```

3 Quantifying

3.1 Final data cleaning

```r
unique(merged_data$Year)
```

```
##  [1] 2012 2013 2014 2015 2016 2017 2018 2019 2020 2021
```

```r
colSums(is.na(merged_data))
```

```
##                Entity                 cause                  Year
##                     0                     0                     0
##              val_rate           val_percent               pop_est
##                     0                     0                     0
##            Gini_index           gini_bucket          income_group
##                     0                     0                     0
##        gdp_per_capita     unemployment_rate       wealth_share_10
##                     0                     0                     0
##             urban_pct             CPI_score    gov_spending_toGDP
##                     0                   968                  1034
## housing_CosttoIncome                 PM2.5
##                  1650                  1320
```

```r
dep_anx <- merged_data[
  (merged_data$cause %in% c("Anxiety disorders", "Depressive disorders")),
  ]

mental_disorders <- merged_data[
  (merged_data$cause %in% c("Mental disorders")),
  ]
```

Generate necessary variables

3.2 Visualizations

```r
filtered_2019 <- merged_data %>%
  filter(Year == 2019 & cause == "Depressive disorders")

ggplot(filtered_2019, aes(x = Gini_index, y = val_percent,
                          color = Entity, size = gdp_per_capita)) +
  geom_point(alpha = 0.8) +
  geom_text_repel(aes(label = Entity), size = 3, max.overlaps = 40) +
```

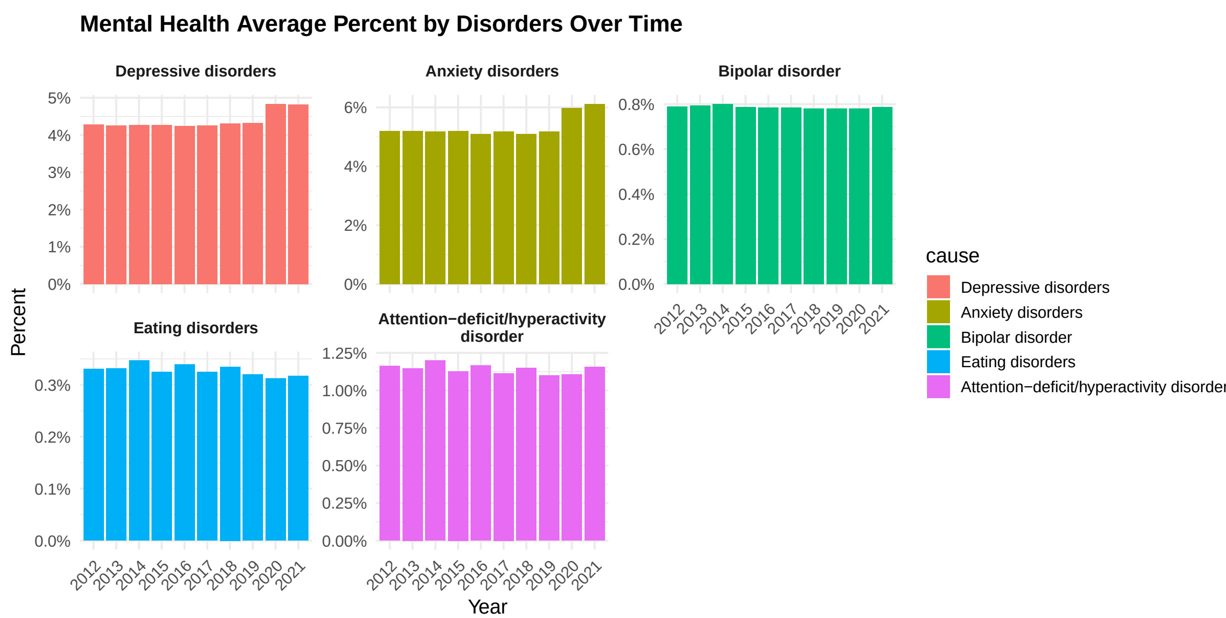
```r
  scale_size_continuous(range = c(5, 15)) +
  labs(
    title = "Depression Rates vs. Income Inequality (Gini) in 2019",
    x = "Gini Index (Income Inequality)",
    y = "Depression Rate (%)",
    color = "Country",
    size = "GDP per Capita"
  ) +
  theme_minimal()
```

3.2.1 Temporal Variations

```r
tendencies <- merged_data %>%
  filter(cause %in% c("Anxiety disorders", "Bipolar disorder",
                      "Attention-deficit/hyperactivity disorder",
                      "Eating disorders", "Depressive disorders")) %>%
  mutate(cause = factor(cause, levels = c("Depressive disorders",
                                          "Anxiety disorders",
                                          "Bipolar disorder",
                                          "Eating disorders",
                                          "Attention-deficit/hyperactivity disorder"))) %>%
  group_by(cause, Year) %>%
  summarise(avg_val_percent = mean(val_percent, na.rm = TRUE)) %>%
  select(cause, Year, avg_val_percent)
```

```
## `summarise()` has grouped output by 'cause'. You can override using the
## `.groups` argument.
```

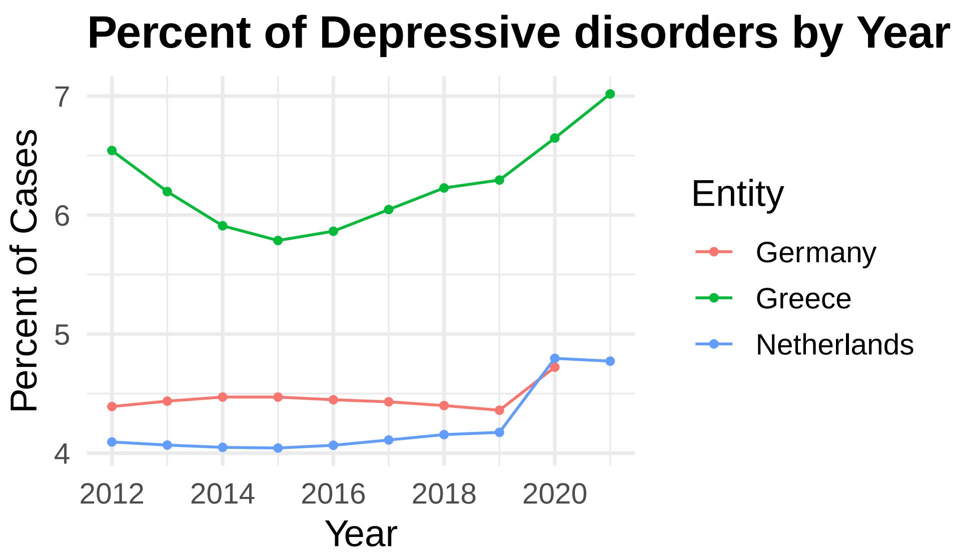
```r
ggplot(tendencies,
       aes(x = factor(Year), y = avg_val_percent, fill = cause)) +
  geom_col() +
  facet_wrap(~ cause, scales = "free_y", labeller = label_wrap_gen(25)) +
  scale_y_continuous(labels = scales::percent_format(scale = 1)) +
  labs(
    title = "Mental Health Average Percent by Disorders Over Time",
    x = "Year",
    y = "Percent"
  ) +
  theme_minimal(base_size = 14) +
  theme(
    axis.text.x = element_text(angle = 45, hjust = 1),
    plot.title = element_text(face = "bold", size = 16),
    strip.text = element_text(face = "bold")
  )
```




```r
prevalence_depr <- dep_anx %>%
  filter(cause == "Depressive disorders", Entity %in% c("Netherlands", "Germany", "Sweeden", "Greece"))
  group_by(Entity, Year)


ggplot(prevalence_depr, aes(x = Year, y = val_percent, color = Entity)) +
  geom_line(size = 0.5) +
  geom_point(size = 1) +
  scale_y_continuous(labels = scales::label_comma()) +
  labs(title = "Percent of Depressive disorders by Year",
       x = "Year", y = "Percent of Cases", color = "Entity") +
  theme_minimal(base_size = 14) +
  theme(plot.title = element_text(face = "bold"))
```

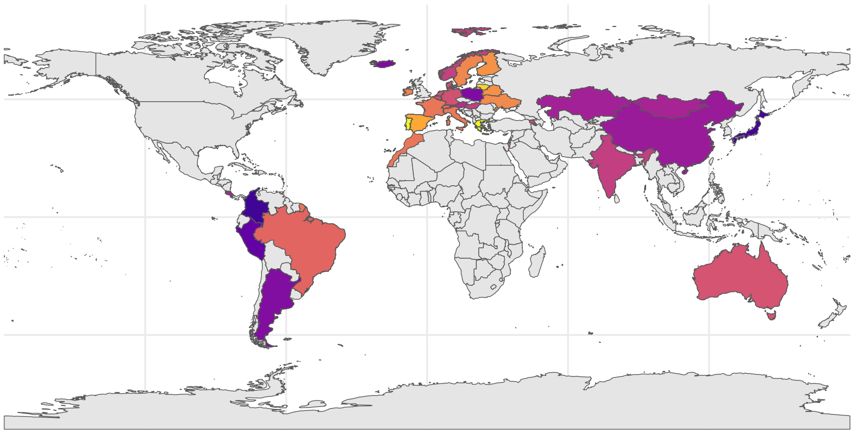
```
## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
```

```
## 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.
```




```r
# Filter to depressive disorders only, select relevant columns for joining
dep_map_data <- merged_data %>%
  filter(cause == "Depressive disorders") %>%
  select(Entity, val_percent)

# Join with world data (assuming 'name' matches 'Entity')
world_dep <- left_join(world, dep_map_data, by = c("name" = "Entity"))

ggplot() +
  geom_sf(data = world, fill = "grey80", color = "white") +
  geom_sf(data = world_dep, aes(fill = val_percent)) +
  scale_fill_viridis_c(option = "plasma", na.value = "grey90") +
  theme_minimal() +
  labs(title = "Depression Rates by Country", fill = "Rate (%)")
```

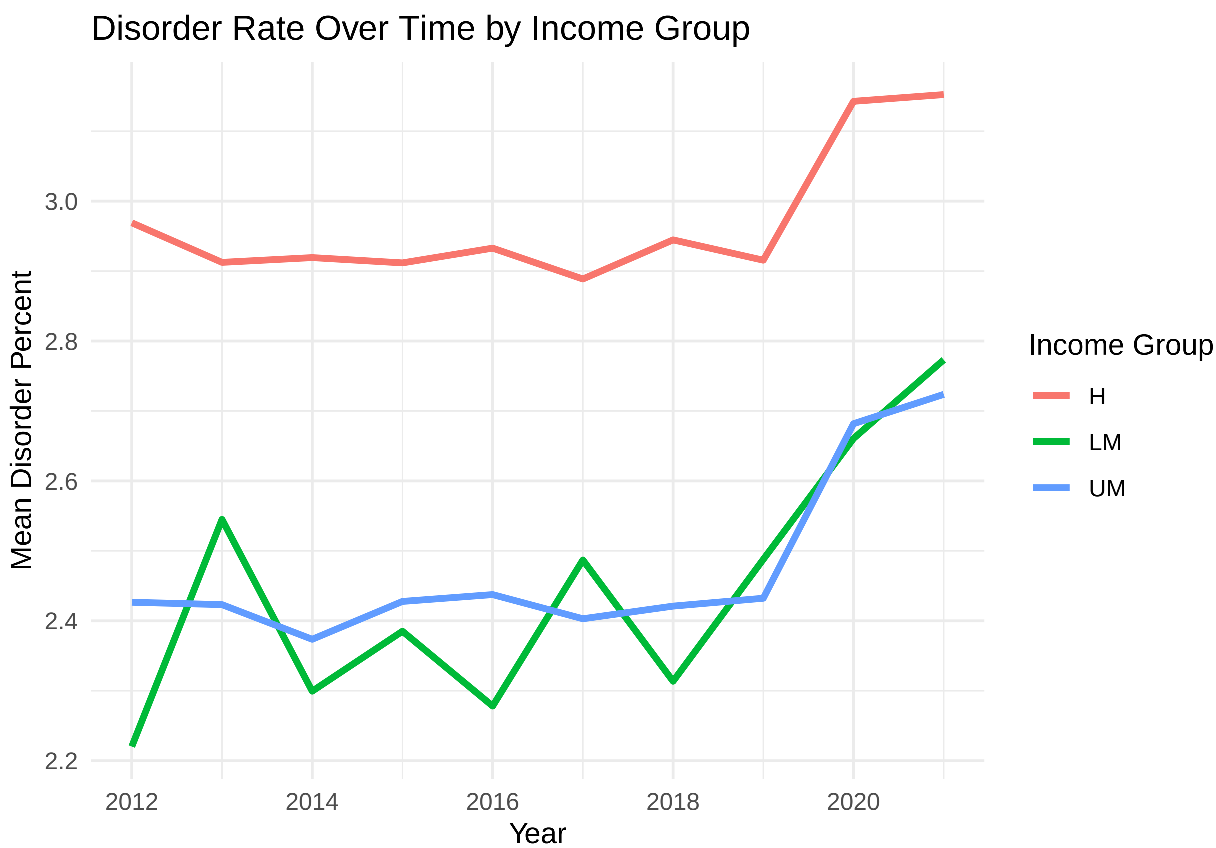
Depression Rates by Country



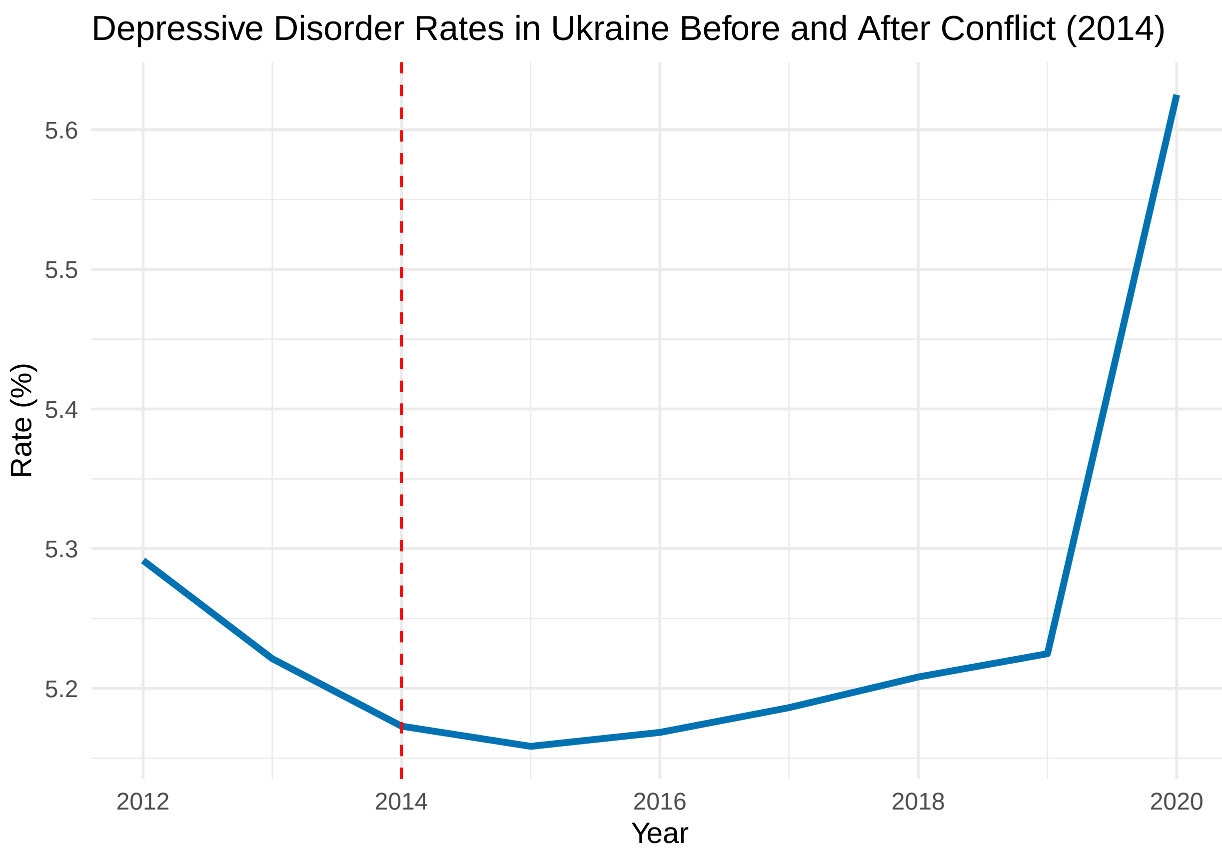
3.2.2 Spatial Analysis (Maps)

```r
# Analyze the relationship between two variables.
# Trend in disorder over time by income group
merged_data %>%
  filter(!is.na(income_group)) %>%
  group_by(Year, income_group) %>%
  summarise(mean_disorder = mean(val_percent, na.rm = T)) %>%
  ggplot(aes(x = Year, y = mean_disorder, color = income_group)) +
  geom_line(size = 1.2) +
  labs(
    title = "Disorder Rate Over Time by Income Group",
    x = "Year",
    y = "Mean Disorder Percent",
    color = "Income Group"
  ) +
  theme_minimal()
```

```
## `summarise()` has grouped output by 'Year'. You can override using the
## `.groups` argument.
```




```r
merged_data %>%
  filter(Entity == "Ukraine",
         cause == "Depressive disorders") %>%
  ggplot(aes(x = Year, y = val_percent)) +
  geom_line(color = "#0072B2", size = 1.2) +
  geom_vline(xintercept = 2014, linetype = "dashed", color = "red") +
  labs(
    title = "Depressive Disorder Rates in Ukraine Before and After Conflict (2014)",
    x = "Year",
    y = "Rate (%)"
  ) +
  theme_minimal()
```



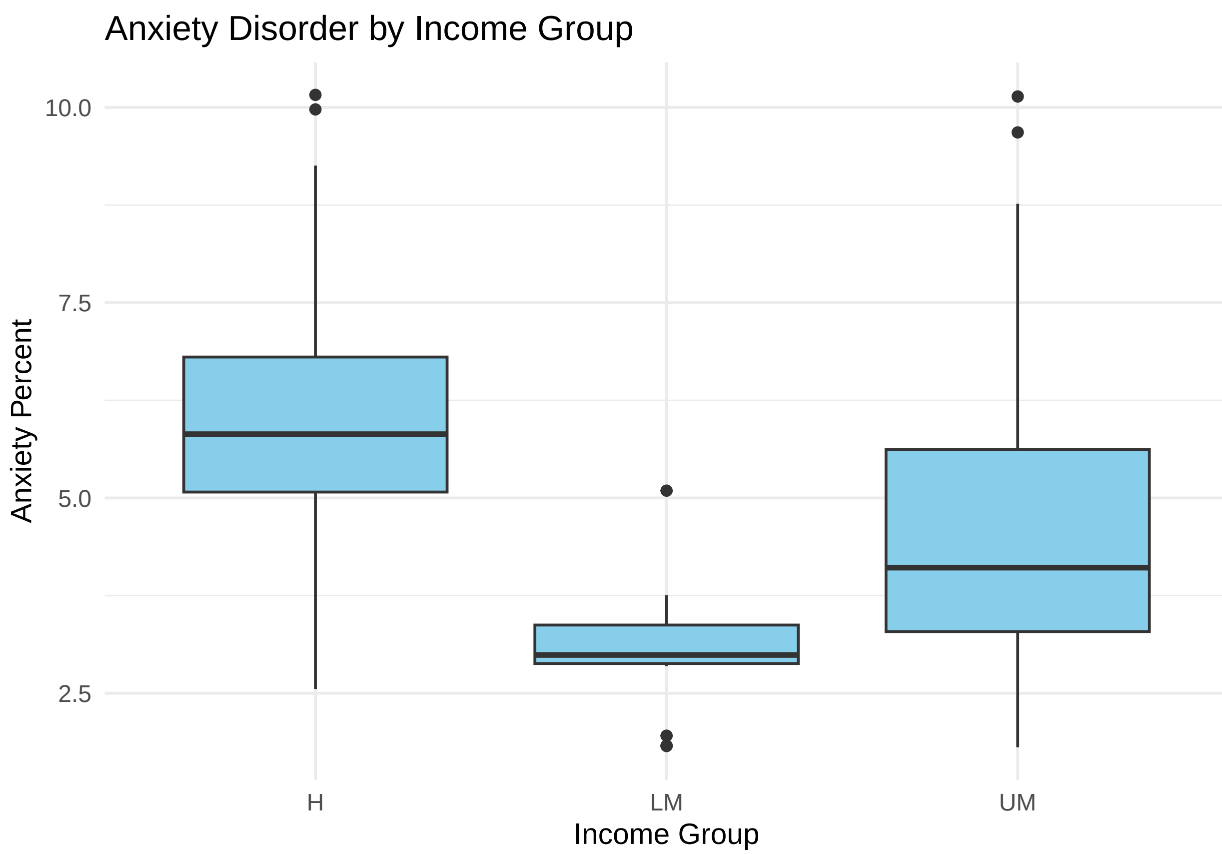
3.2.3 Variation & Subgroup Analysis

```
# Rename causes for cleaner legend
mental_Sweden <- dep_anx %>%
  filter(Entity == "Sweden") %>%
  mutate(cause = recode(cause,
                        "Depressive Disorders" = "Depression",
                        "Anxiety Disorders" = "Anxiety"))


# Aggregate by year and cause, averaging disorder_rate across sex and age groups
mental_summary <- mental_Sweden %>%
  group_by(Year, cause)

# Plot
p <- ggplot(mental_summary, aes(x = factor(Year), y = val_rate, fill = cause,
                               text = paste0("Year: ", Year, "<br>",
                                             "Cause: ", cause, "<br>"
                                             #,"Avg Rate: ", comma(round(val_rate, 2)
                                             ))) +
  geom_col(position = position_dodge(width = 0.7), width = 0.6) +
  scale_fill_manual(values = c( "Depressive disorders" = "#2E8B57", "Anxiety disorders" = "#6A5ACD")) +
  #scale_y_continuous(labels = comma) +
  labs(
```

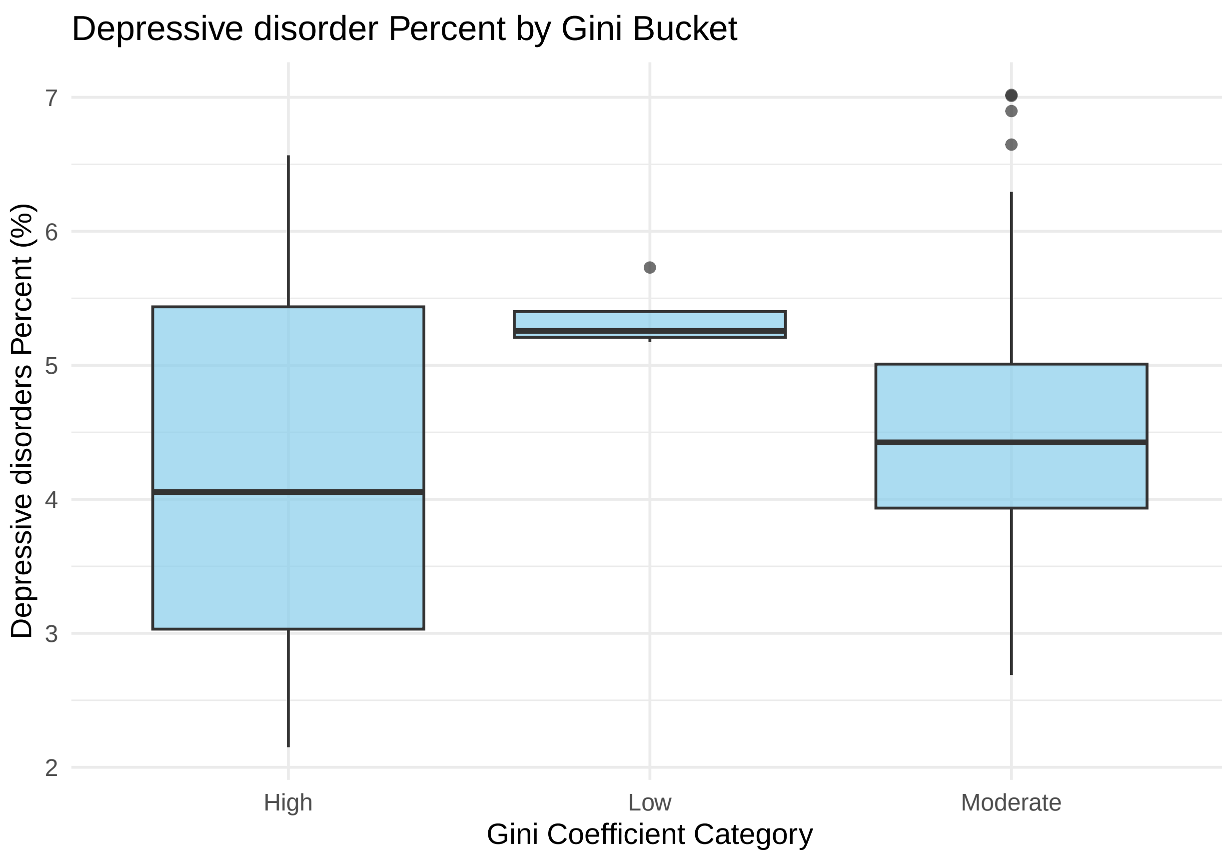
```r
    title = "Disorder Rates in Sweden Over Time (Aggregated by Cause)",
    x = "Year", y = "Disorder Rate", fill = "Cause"
  ) +
  theme_minimal(base_size = 14) +
  theme(
    panel.grid.major.y = element_blank(),
    panel.grid.minor.y = element_blank(),
    panel.grid.major.x = element_blank(),
    panel.grid.minor.x = element_blank()
  )

# Interactive plot with better tooltips
ggplotly(p, tooltip = "text")
```

3.3 Event analysis

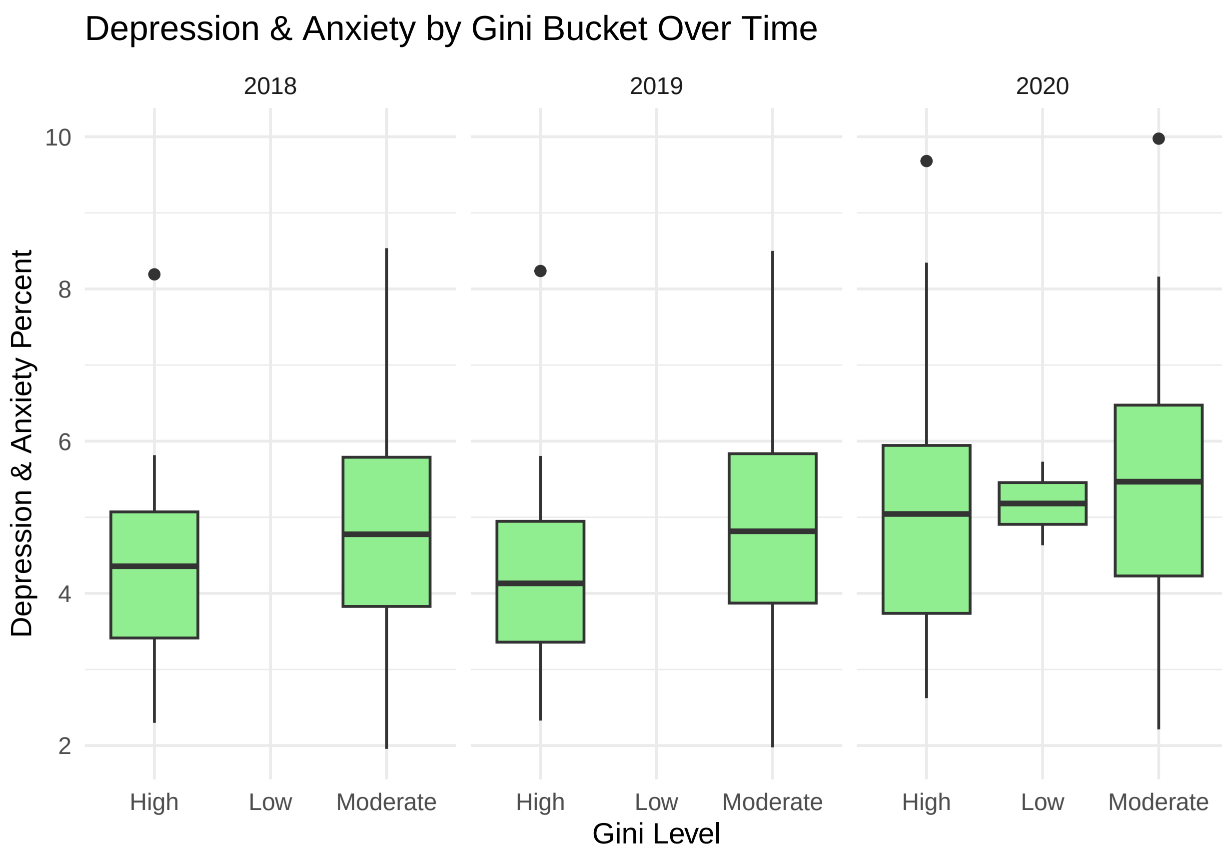
```r
# Boxplot: Anxiety disorders by income group
anxiety <- dep_anx %>%
  filter(cause == "Anxiety disorders")

anxiety %>%
  filter(!is.na(income_group)) %>%
ggplot(aes(x = income_group, y = val_percent)) +
  geom_boxplot(fill = "skyblue") +
  labs(
    title = "Anxiety Disorder by Income Group",
    x = "Income Group",
    y = "Anxiety Percent"
  ) +
  theme_minimal()
```




```r
# Shows the distribution (median, spread, outliers) of disorder rates for each gini group.
depr <- dep_anx %>%
  filter(cause == "Depressive disorders") %>%
  mutate(logged_depr = log(val_percent + 1))

# Basic boxplot
depr %>%
  filter(!is.na(gini_bucket)) %>%
ggplot(aes(x = gini_bucket, y = val_percent)) +
  geom_boxplot(fill = "skyblue", alpha = 0.7) +
  labs(title = "Depressive disorder Percent by Gini Bucket",
       x = "Gini Coefficient Category",
       y = "Depressive disorders Percent (%)") +
  theme_minimal()
```




```r
dep_anx %>%
  filter(Year %in% c(2018, 2019, 2020)) %>%  # Limit to a few years for clarity
  ggplot(aes(x = gini_bucket, y = val_percent)) +
  geom_boxplot(fill = "lightgreen") +
  facet_wrap(~Year) +
  labs(title = "Depression & Anxiety by Gini Bucket Over Time",
       x = "Gini Level",
       y = "Depression & Anxiety Percent") +
  theme_minimal()
```




```r
colnames(merged_data)
```

```
##  [1] "Entity"                "cause"                 "Year"
##  [4] "val_rate"              "val_percent"           "pop_est"
##  [7] "Gini_index"            "gini_bucket"           "income_group"
## [10] "gdp_per_capita"        "unemployment_rate"     "wealth_share_10"
## [13] "urban_pct"             "CPI_score"             "gov_spending_toGDP"
## [16] "housing_CosttoIncome" "PM2.5"
```

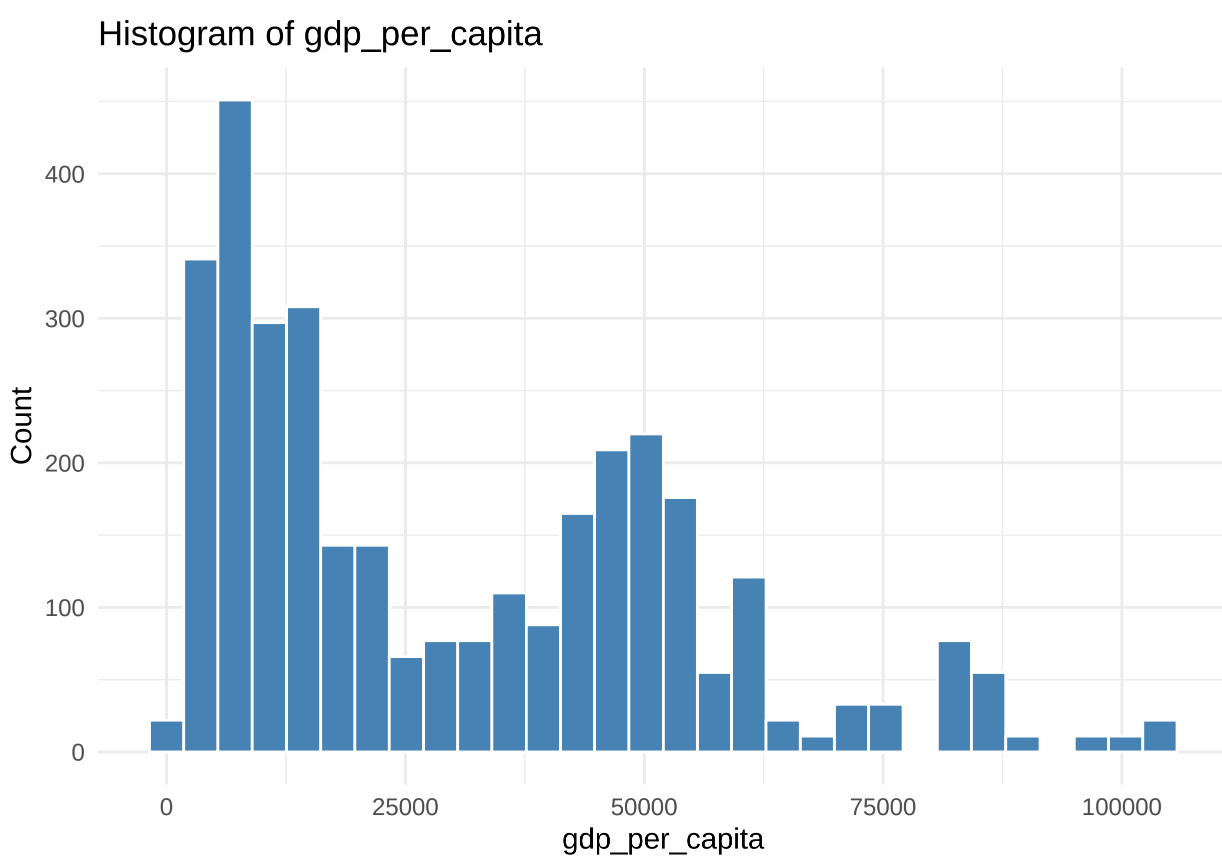
```r
# Pick numeric variables from dataset
vars_to_check <- c("gdp_per_capita", "Gini_index", "urban_pct", "pop_est", "housing_CosttoIncome",
                   "wealth_share_10",
                   "CPI_score", "PM2.5", "unemployment_rate", "gov_spending_toGDP"
                   )
```

```r
# Loop through and print skewness + plot histogram
for (var in vars_to_check) {
  cat("\n\n==========", var, "==========\n")

  # Skewness (higher than |1| = very skewed)
  skew_val <- skewness(merged_data[[var]], na.rm = TRUE)
  cat("Skewness:", round(skew_val, 2), "\n")

  # Histogram
  print(
```

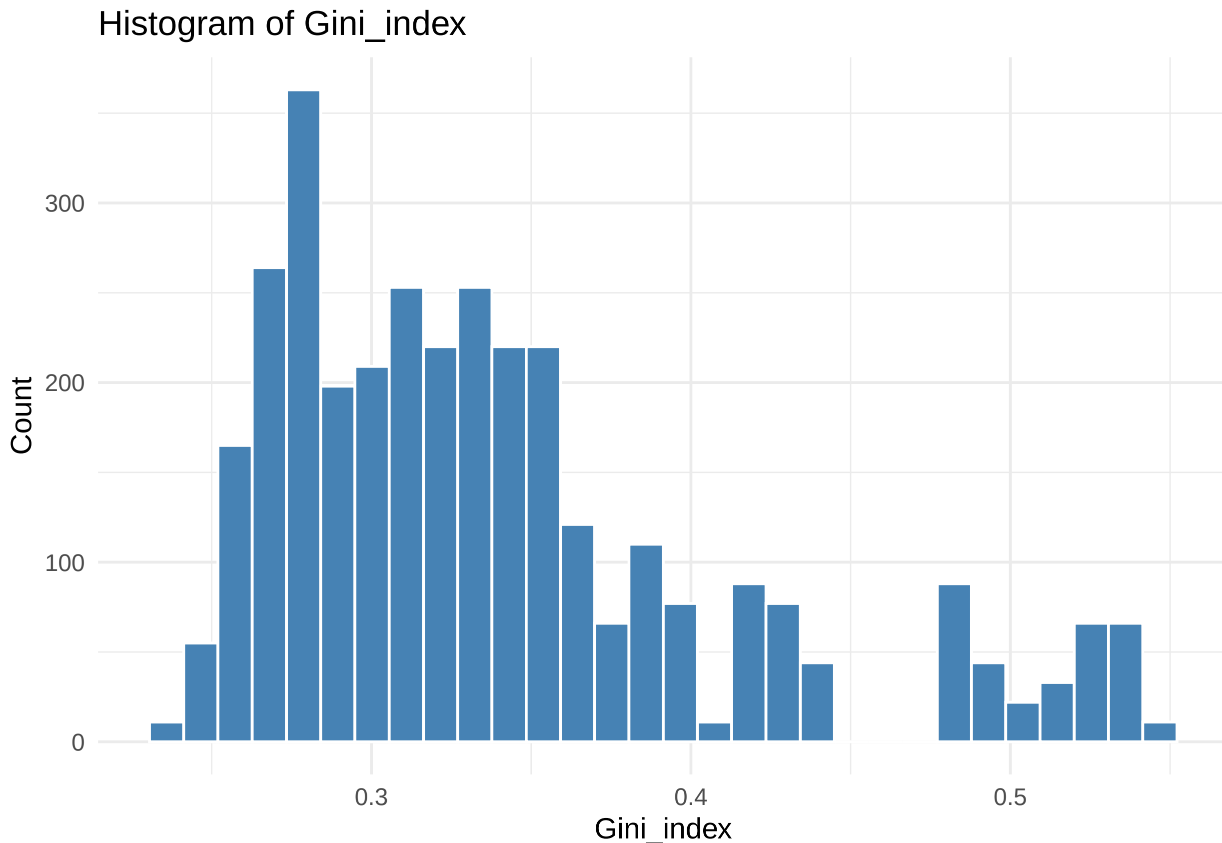
```r
    ggplot(merged_data, aes_string(x = var)) +
      geom_histogram(bins = 30, fill = "steelblue", color = "white") +
      labs(title = paste("Histogram of", var), x = var, y = "Count") +
      theme_minimal()
  )
}
```

```
##
##
## ========== gdp_per_capita ==========
## Skewness: 0.77
```

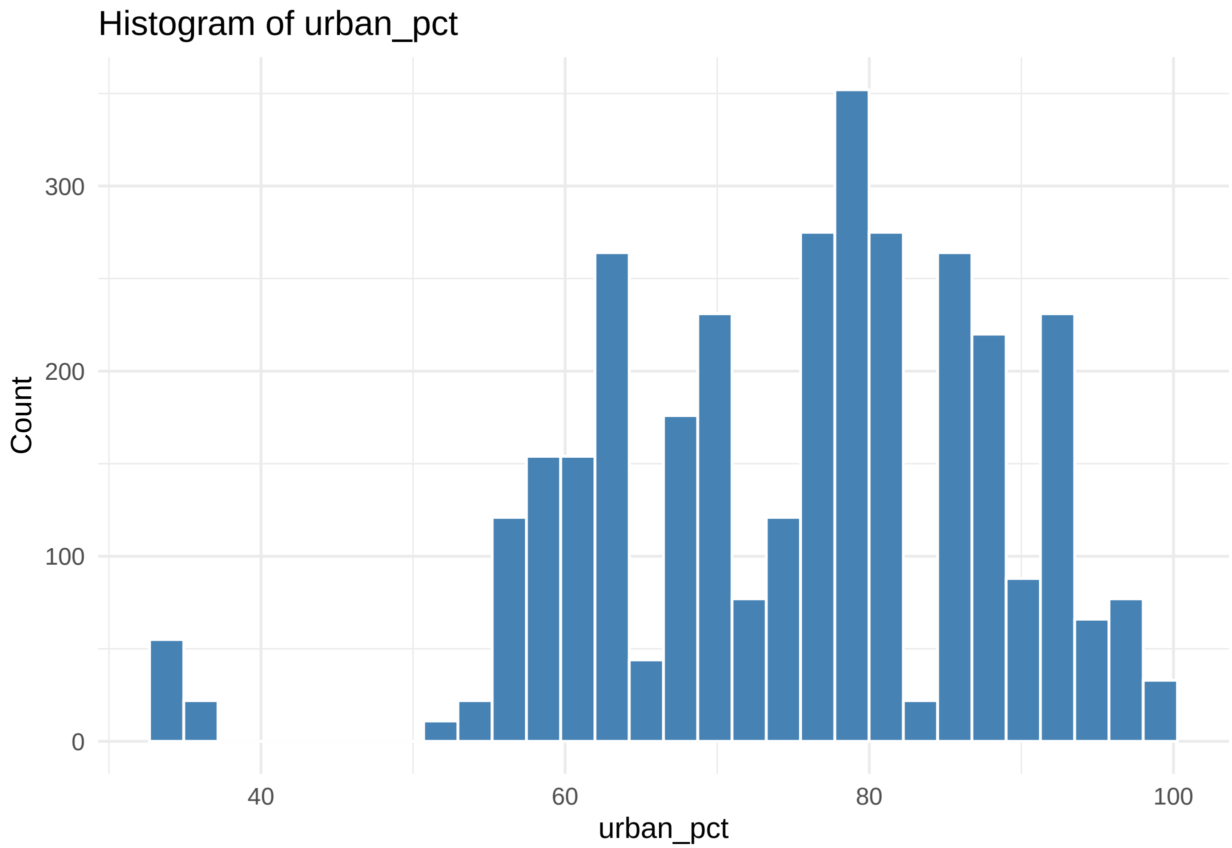
```
## Warning: `aes_string()` was deprecated in ggplot2 3.0.0.
## i Please use tidy evaluation idioms with `aes()`.
## i See also `vignette("ggplot2-in-packages")` for more information.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```



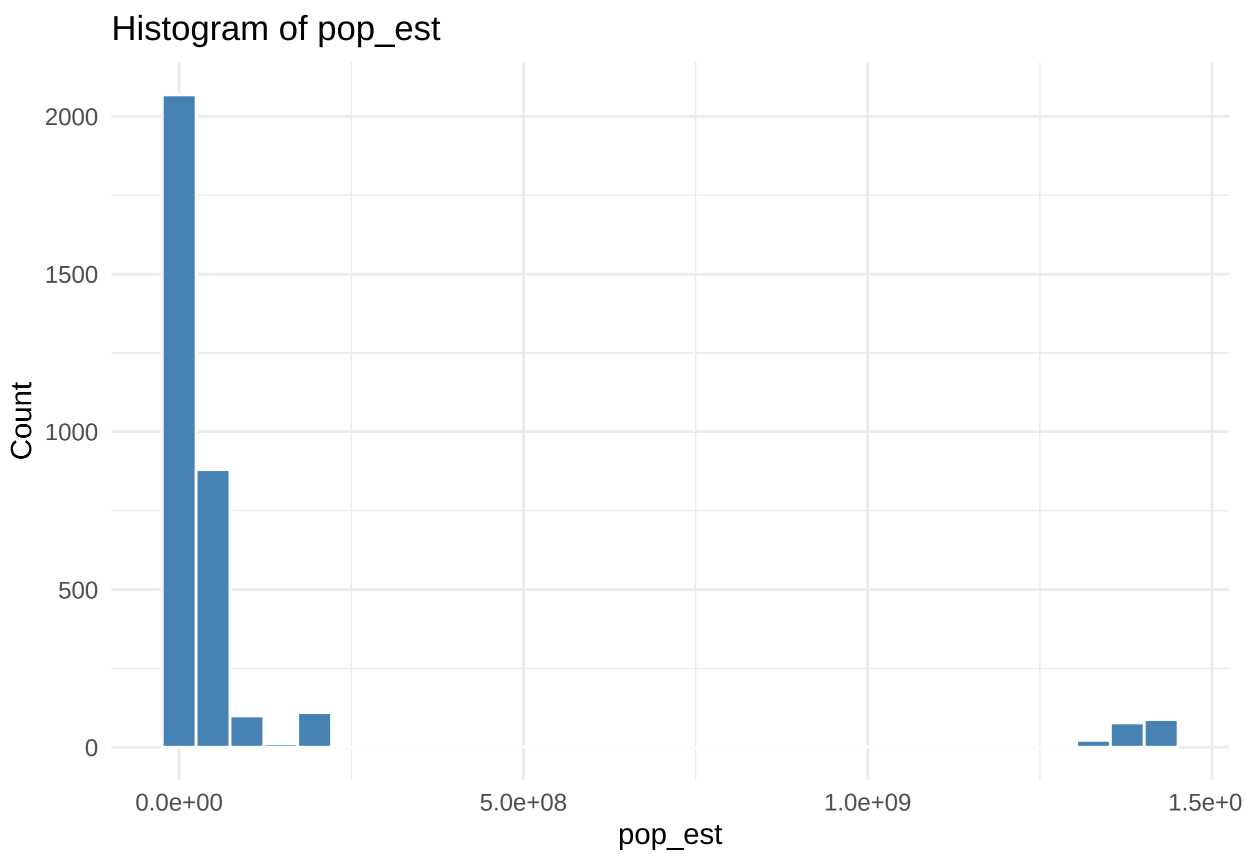

```
##
##
## ========== Gini_index ==========
## Skewness: 1.18
```



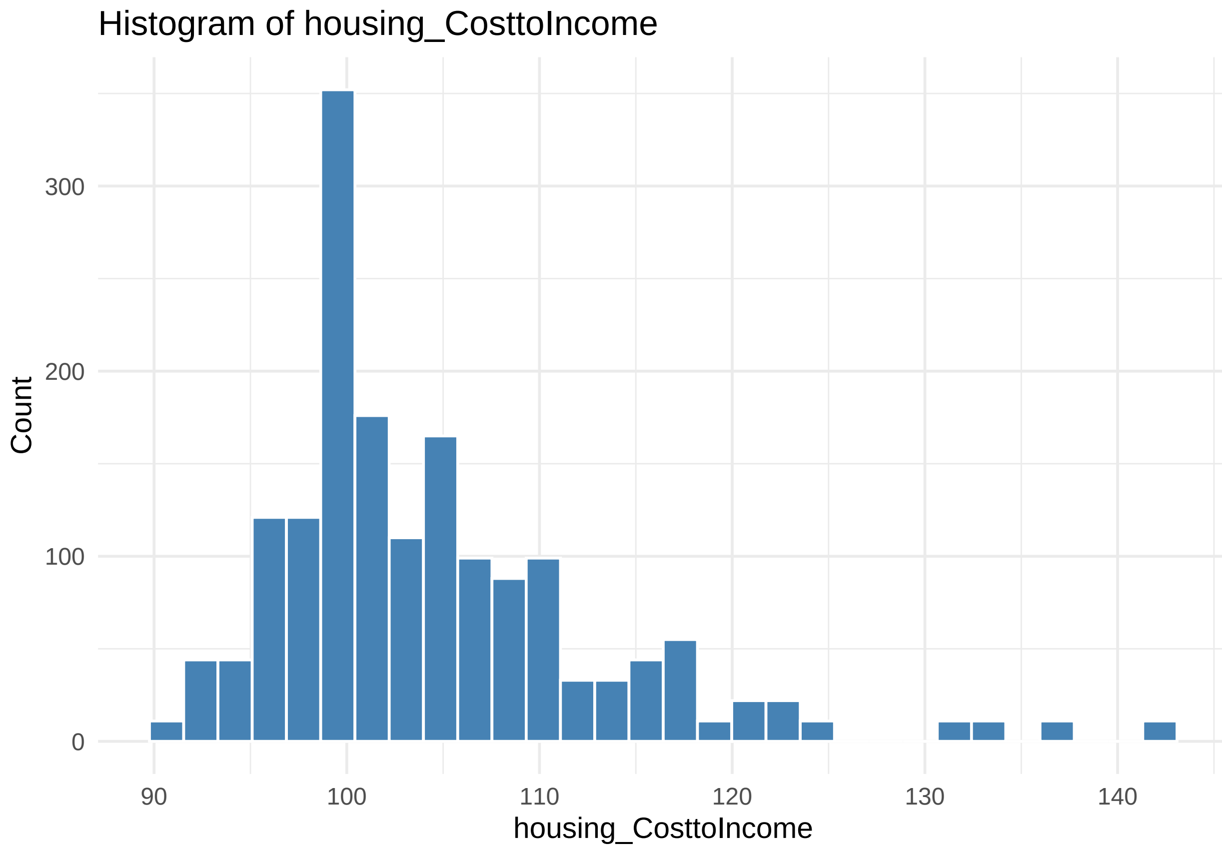

```
##
##
## ========== urban_pct ==========
## Skewness: -0.61
```



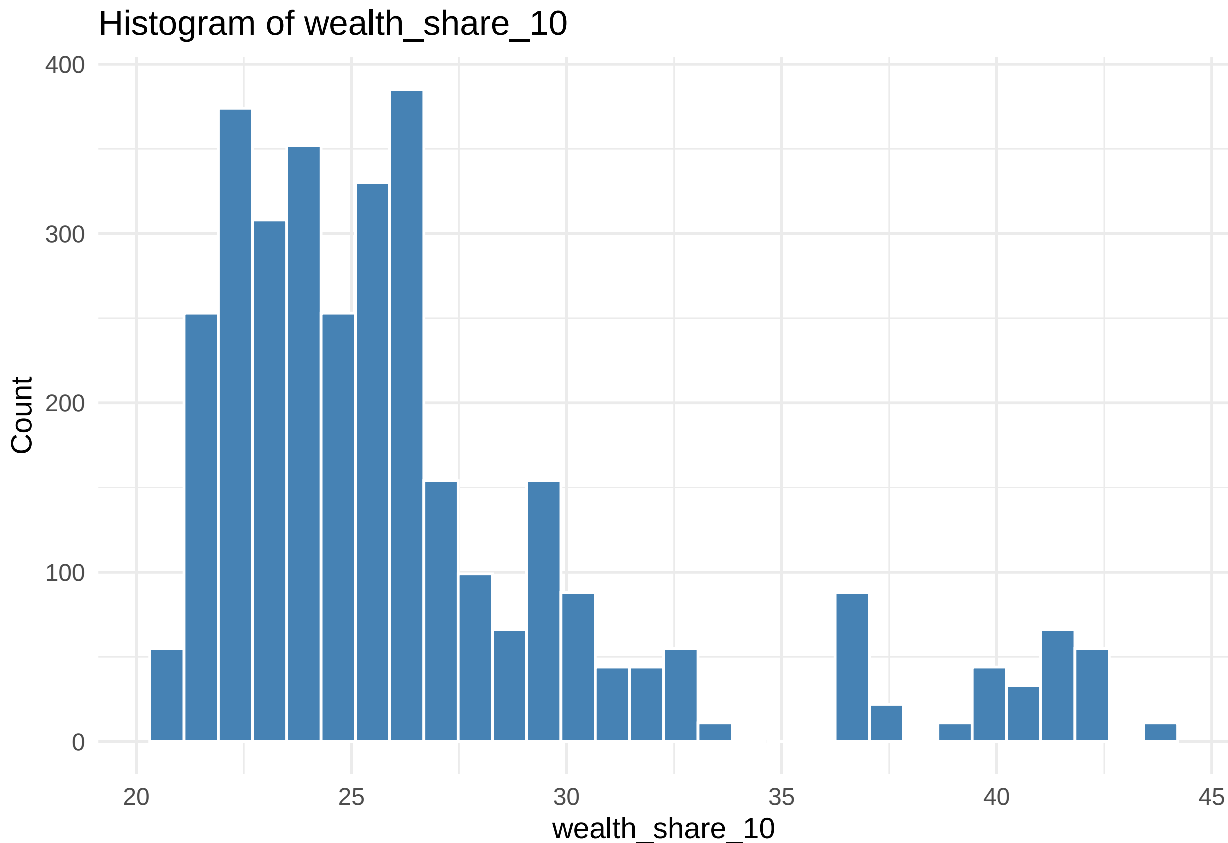

```
##
##
## ========== pop_est ==========
## Skewness: 3.78
```



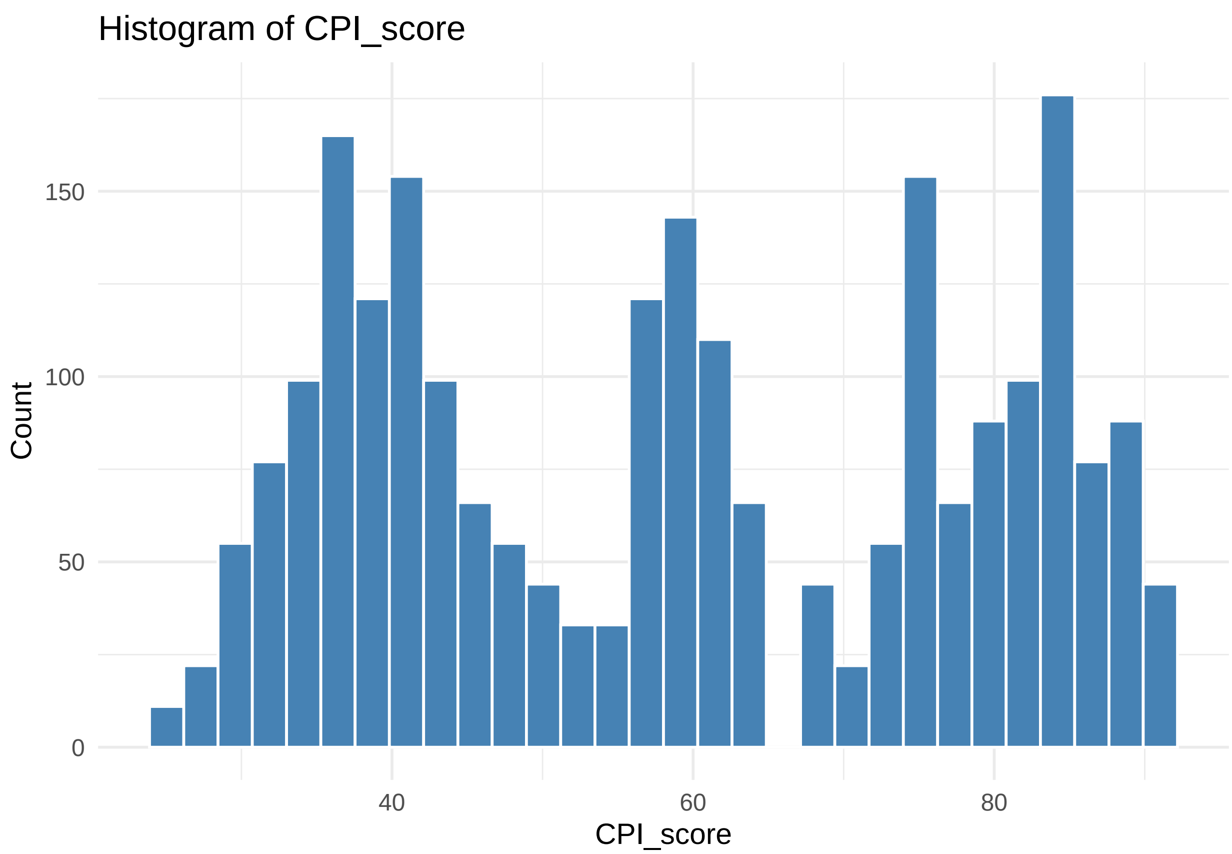

```
##
##
## ========== housing_CosttoIncome ==========
## Skewness: 1.57

## Warning: Removed 1650 rows containing non-finite outside the scale range
## (`stat_bin()`).
```



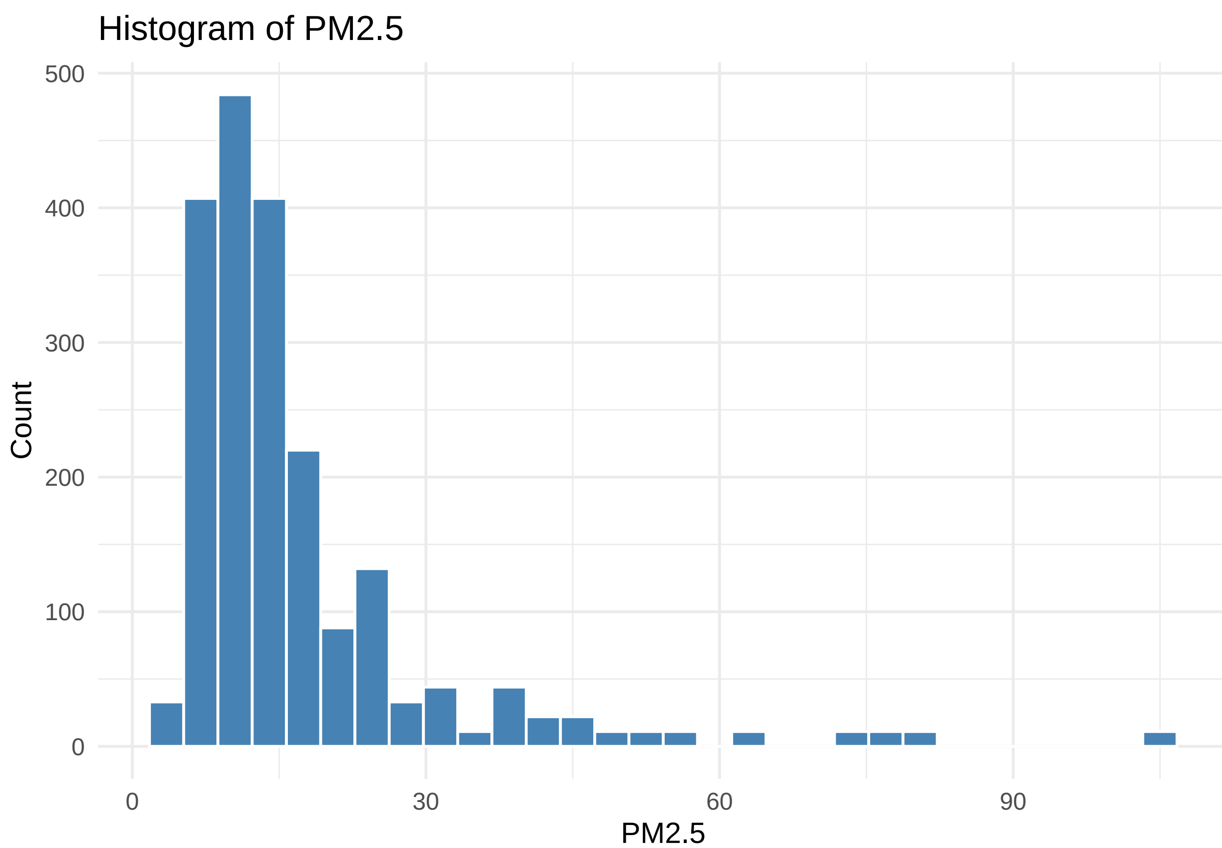

```
##
##
## ========== wealth_share_10 ==========
## Skewness: 1.64
```



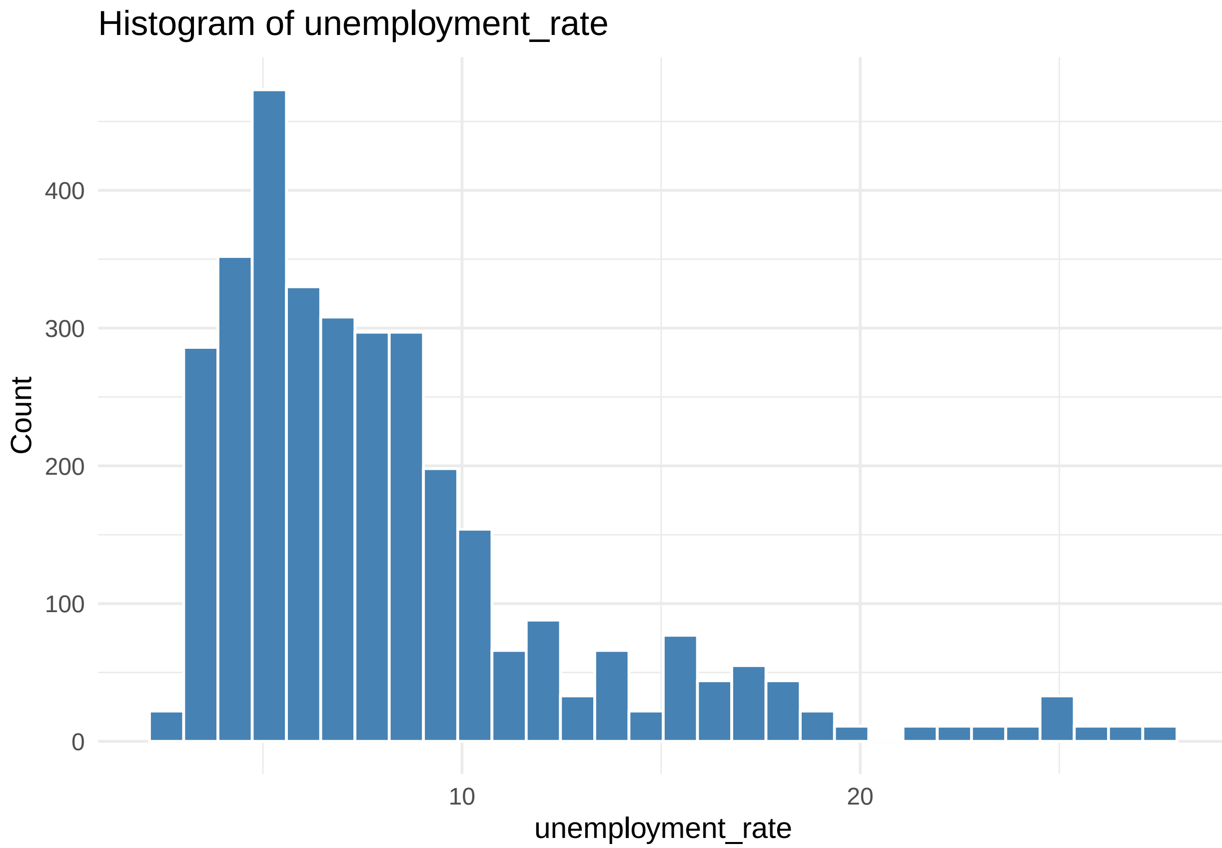

```
##
##
## ========== CPI_score ==========
## Skewness: 0.07

## Warning: Removed 968 rows containing non-finite outside the scale range
## (`stat_bin()`).
```




```
##
##
## ========== PM2.5 ==========
## Skewness: 3

## Warning: Removed 1320 rows containing non-finite outside the scale range
## (`stat_bin()`).
```

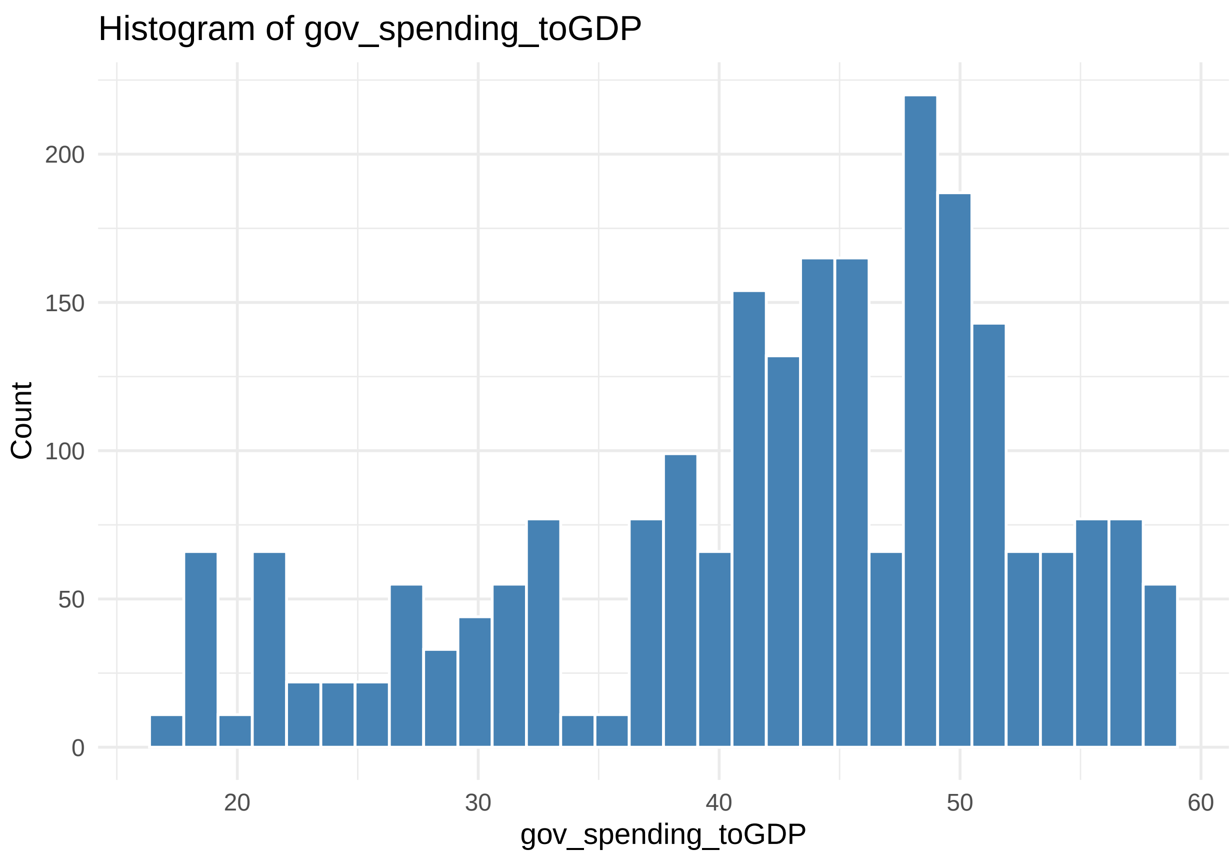



```
##
##
## ========== unemployment_rate ==========
## Skewness: 1.74
```




```
##
##
## ========== gov_spending_toGDP ==========
## Skewness: -0.75
```

```
## Warning: Removed 1034 rows containing non-finite outside the scale range
## (`stat_bin()`).
```




```
disorders_wide <- merged_data %>%
  filter(cause %in% c("Depressive disorders", "Anxiety disorders", "Schizophrenia", "Bipolar disorder",
  select(Entity, Year, cause, val_percent, gdp_per_capita, wealth_share_10, unemployment_rate, pop_est,
  pivot_wider(names_from = cause, values_from = val_percent)


merged_data_log <- disorders_wide %>%
  mutate(
    log_eating = log(`Eating disorders` + 1),

    log_gdp_per_capita = log(gdp_per_capita + 1),
    log_wealth_share_10 = log(wealth_share_10 + 1),
    log_unemployment_rate = log(unemployment_rate + 1),
    log_population = log(pop_est + 1),
    log_PM2.5 = log(PM2.5 + 1)
  )
```

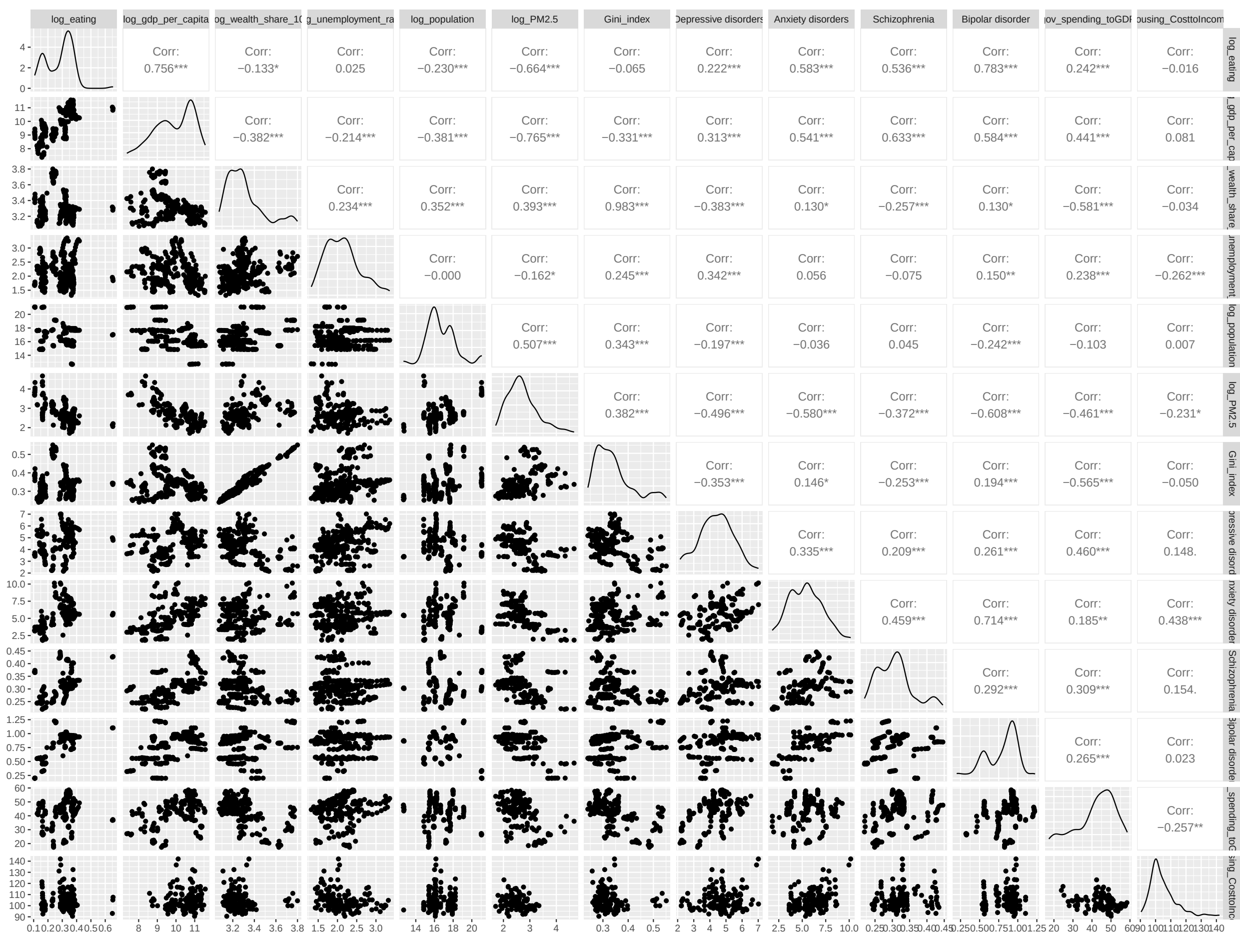
```
merged_data_log %>%
  select(log_eating, log_gdp_per_capita, log_wealth_share_10,  log_unemployment_rate, log_population, l
         Gini_index,
         `Depressive disorders`, `Anxiety disorders`, Schizophrenia, `Bipolar disorder`,
         gov_spending_toGDP, housing_CosttoIncome) %>%
  cor(use = "complete.obs") %>%
  round(2)
```

```
##                          log_eating log_gdp_per_capita log_wealth_share_10
```

```
## log_eating              1.00            0.61              -0.18
## log_gdp_per_capita      0.61            1.00              -0.64
## log_wealth_share_10    -0.18           -0.64               1.00
## log_unemployment_rate   0.11           -0.38               0.30
## log_population         -0.12           -0.44               0.40
## log_PM2.5              -0.55           -0.60               0.40
## Gini_index             -0.15           -0.67               0.98
## Depressive disorders    0.46            0.36              -0.33
## Anxiety disorders       0.34            0.49              -0.08
## Schizophrenia           0.52            0.63              -0.35
## Bipolar disorder        0.74            0.46              -0.08
## gov_spending_toGDP      0.04            0.11              -0.51
## housing_CosttoIncome   -0.04            0.06               0.02
##                       log_unemployment_rate log_population log_PM2.5 Gini_index
## log_eating                             0.11          -0.12     -0.55      -0.15
## log_gdp_per_capita                    -0.38          -0.44     -0.60      -0.67
## log_wealth_share_10                    0.30           0.40      0.40       0.98
## log_unemployment_rate                  1.00           0.23      0.08       0.37
## log_population                         0.23           1.00      0.45       0.45
## log_PM2.5                              0.08           0.45      1.00       0.44
## Gini_index                             0.37           0.45      0.44       1.00
## Depressive disorders                   0.45          -0.14     -0.50      -0.27
## Anxiety disorders                     -0.04          -0.11     -0.38      -0.13
## Schizophrenia                         -0.11          -0.33     -0.40      -0.41
## Bipolar disorder                       0.24          -0.06     -0.38      -0.06
## gov_spending_toGDP                     0.15          -0.01     -0.14      -0.50
## housing_CosttoIncome                  -0.34          -0.02     -0.22      -0.01
##                       Depressive disorders Anxiety disorders Schizophrenia
## log_eating                            0.46              0.34          0.52
## log_gdp_per_capita                    0.36              0.49          0.63
## log_wealth_share_10                  -0.33             -0.08         -0.35
## log_unemployment_rate                 0.45             -0.04         -0.11
## log_population                       -0.14             -0.11         -0.33
## log_PM2.5                            -0.50             -0.38         -0.40
## Gini_index                           -0.27             -0.13         -0.41
## Depressive disorders                  1.00              0.45          0.30
## Anxiety disorders                     0.45              1.00          0.43
## Schizophrenia                         0.30              0.43          1.00
## Bipolar disorder                      0.58              0.42          0.38
## gov_spending_toGDP                    0.20             -0.11         -0.08
## housing_CosttoIncome                  0.05              0.28          0.09
##                       Bipolar disorder gov_spending_toGDP housing_CosttoIncome
## log_eating                        0.74               0.04                -0.04
## log_gdp_per_capita                0.46               0.11                 0.06
## log_wealth_share_10              -0.08              -0.51                 0.02
## log_unemployment_rate             0.24               0.15                -0.34
## log_population                   -0.06              -0.01                -0.02
## log_PM2.5                        -0.38              -0.14                -0.22
## Gini_index                       -0.06              -0.50                -0.01
## Depressive disorders              0.58               0.20                 0.05
## Anxiety disorders                 0.42              -0.11                 0.28
## Schizophrenia                     0.38              -0.08                 0.09
## Bipolar disorder                  1.00               0.27                -0.09
## gov_spending_toGDP                0.27               1.00                -0.26
```

```
## housing_CosttoIncome                    -0.09                -0.26                  1.00
```

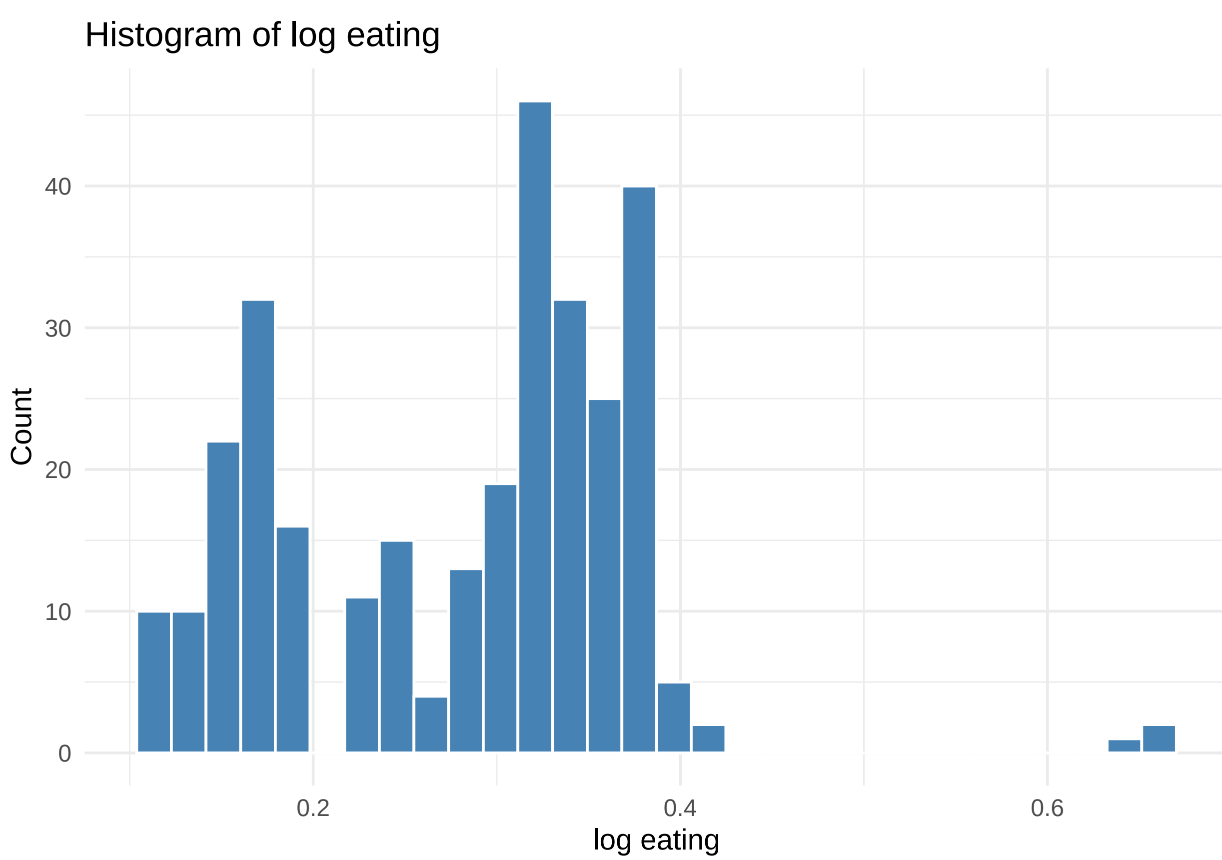
```r
p <- merged_data_log %>%
  select(log_eating,
    log_gdp_per_capita, log_wealth_share_10,  log_unemployment_rate, log_population, log_PM2.5,
         Gini_index,
    `Depressive disorders`, `Anxiety disorders`, Schizophrenia, `Bipolar disorder`,
    gov_spending_toGDP, housing_CosttoIncome) %>%
  ggpairs()

# Save bigger image
ggsave("correlation_plot.png", plot = p, width = 16, height = 12, dpi = 300)

# Show the plot in the knitted HTML
p
```



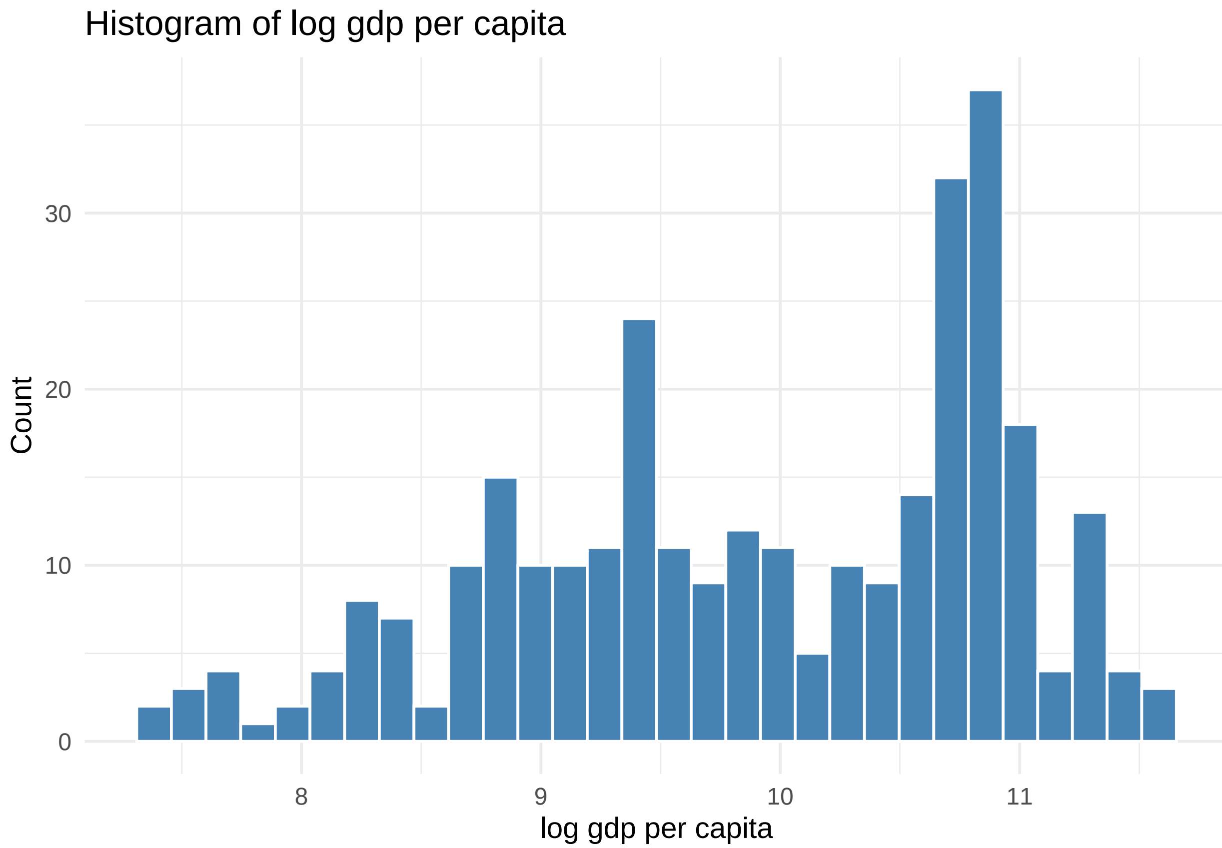

```r
logs <- c("log_eating",
  "log_gdp_per_capita", "log_wealth_share_10", "log_unemployment_rate", "log_population", "log_PM2.5")

for (var in logs) {
  cat("\n\n==========", var, "==========\n")

  # Skewness (higher than |1| = very skewed)
```

```
skew_val <- skewness(merged_data_log[[var]], na.rm = TRUE)
cat("Skewness:", round(skew_val, 2), "\n")

# Histogram
print(
  ggplot(merged_data_log, aes_string(x = var)) +
    geom_histogram(bins = 30, fill = "steelblue", color = "white") +
    labs(
        title = paste("Histogram of", gsub("_", " ", var)),
        x = gsub("_", " ", var),
        y = "Count"
    ) +
    theme_minimal())
}
```

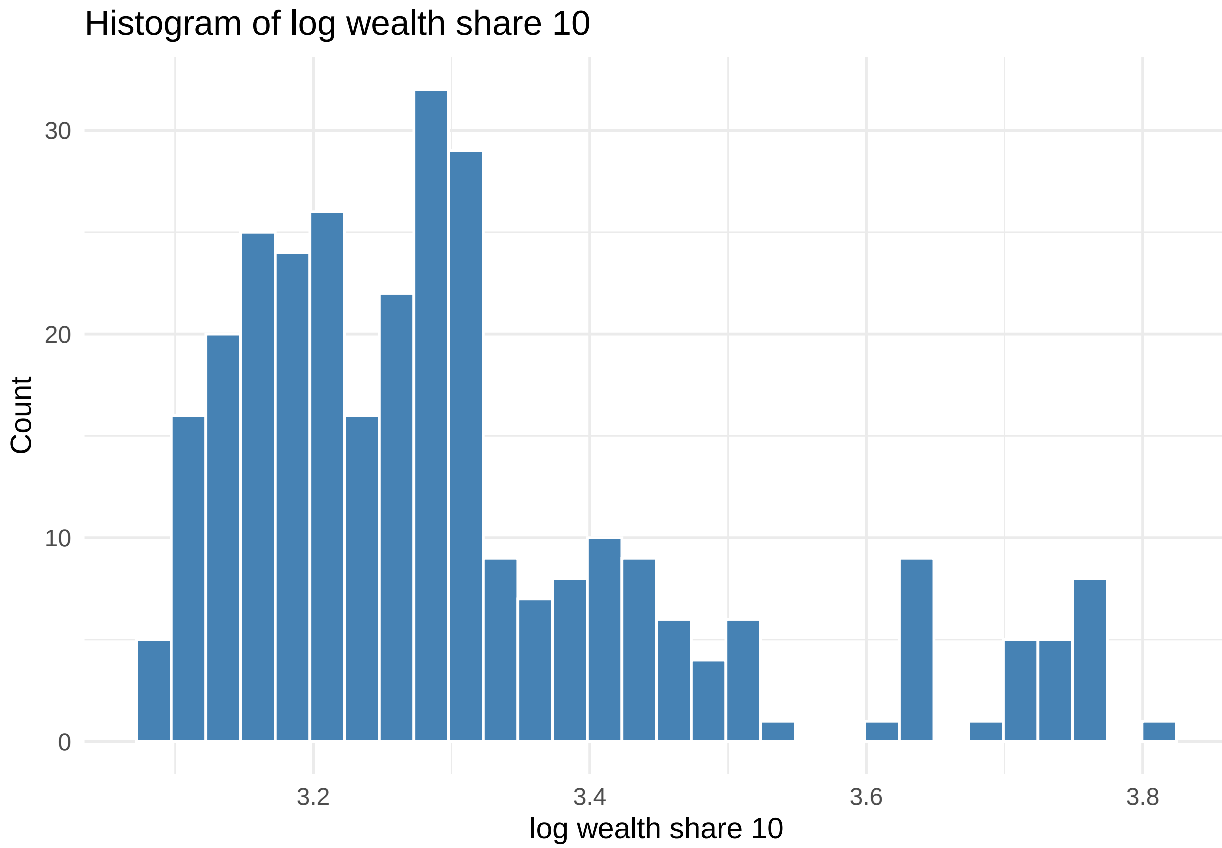
```
##
##
## ========== log_eating ==========
## Skewness: 0.14
```



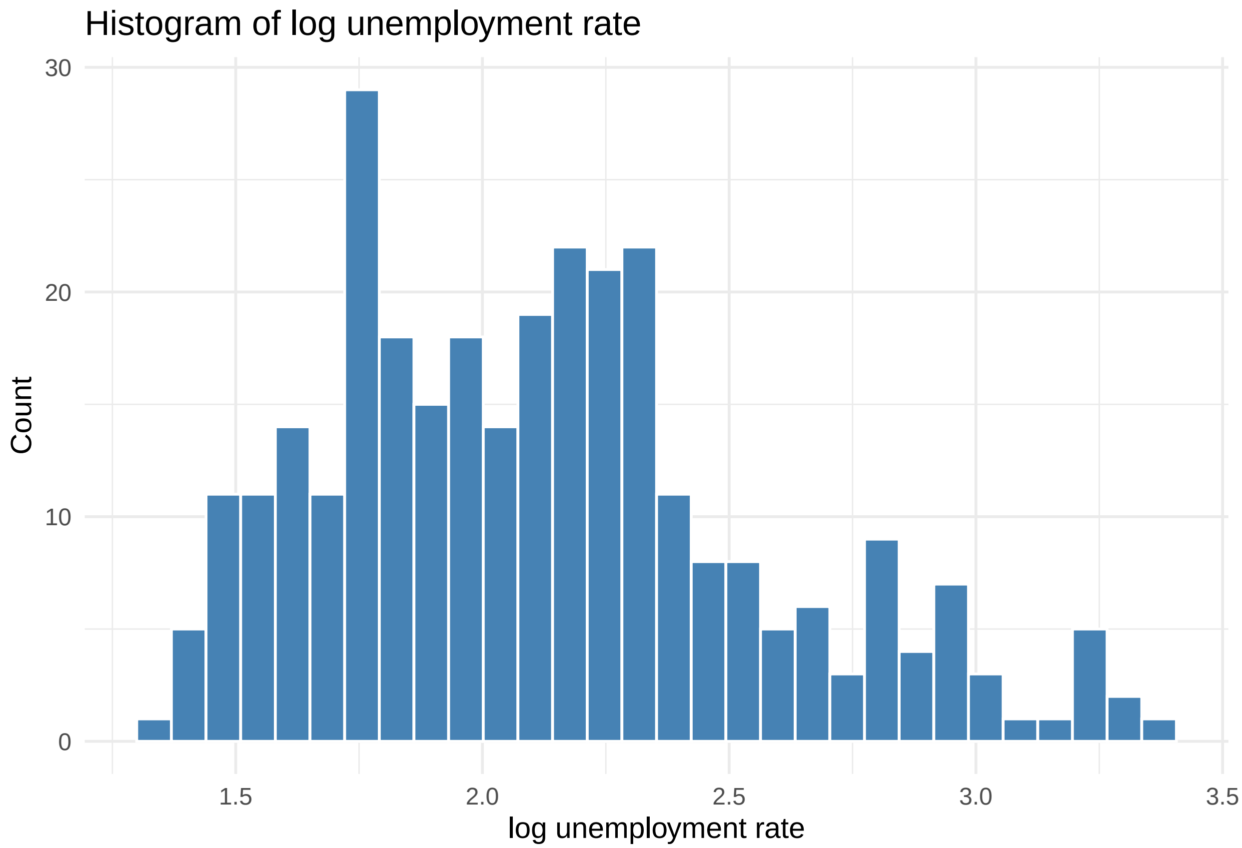

```
##
##
## ========== log_gdp_per_capita ==========
## Skewness: -0.45
```



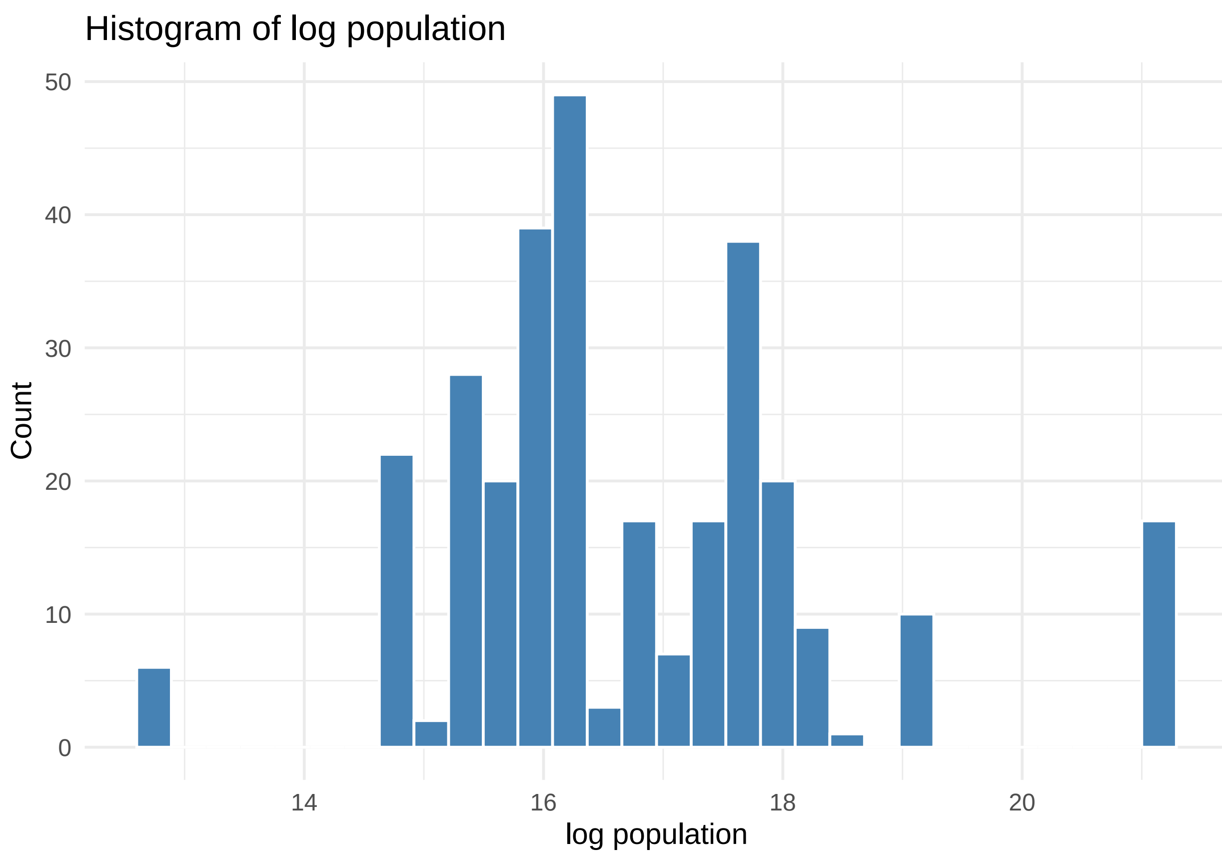

```
##
##
## ========== log_wealth_share_10 ==========
## Skewness: 1.27
```




```
##
##
## ========== log_unemployment_rate ==========
## Skewness: 0.62
```

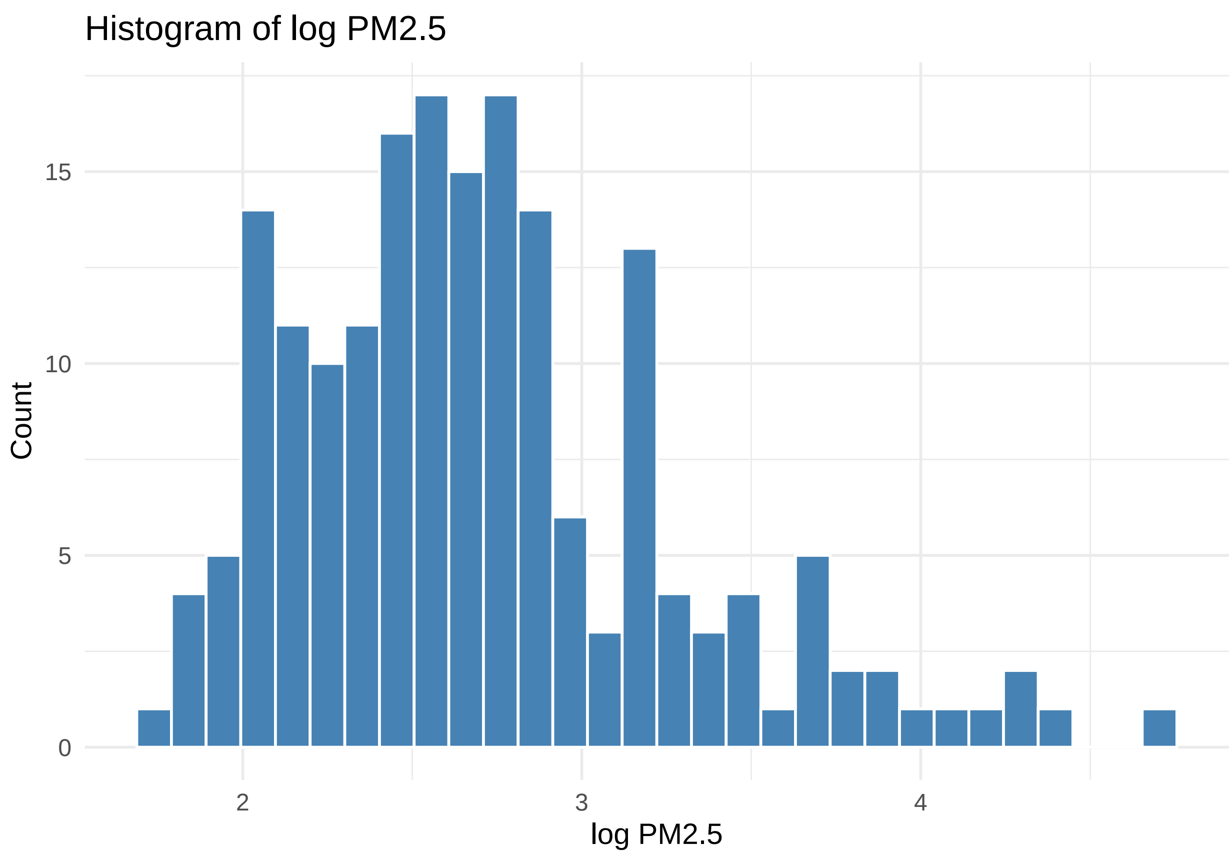



```
##
##
## ========== log_population ==========
## Skewness: 0.79
```



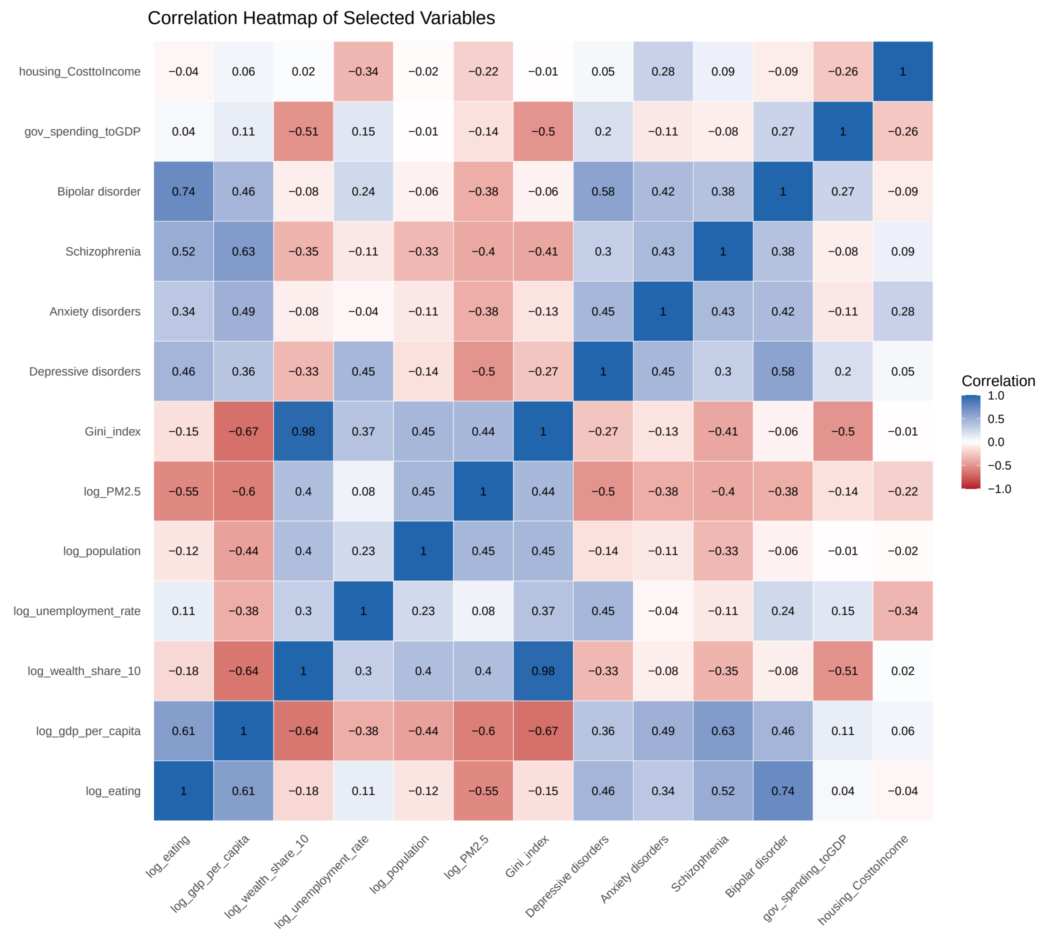

```
##
##
## ========== log_PM2.5 ==========
## Skewness: 0.89
```

```
## Warning: Removed 120 rows containing non-finite outside the scale range
## (`stat_bin()`).
```




```
# Calculate correlation matrix
cor_matrix_log <- merged_data_log %>%
  select(log_eating, log_gdp_per_capita, log_wealth_share_10,  log_unemployment_rate, log_population, lo
         Gini_index,
         `Depressive disorders`, `Anxiety disorders`, Schizophrenia, `Bipolar disorder`,
         gov_spending_toGDP, housing_CosttoIncome) %>%
  cor(use = "complete.obs") %>%
  round(2)


# Melt the correlation matrix
cor_df_log <- melt(cor_matrix_log)

# Plot as heatmap
# Create heatmap
heatmap <- ggplot(cor_df_log, aes(x = Var1, y = Var2, fill = value)) +
  geom_tile(color = "white") +
  scale_fill_gradient2(
    low = "#B2182B", high = "#2166AC", mid = "white",
    midpoint = 0, limit = c(-1, 1), space = "Lab",
    name = "Correlation"
  ) +
  theme_minimal(base_size = 14) +
  theme(
    axis.text.x = element_text(angle = 45, vjust = 1, hjust = 1),
    panel.grid = element_blank()
```

```r
  ) +
  coord_fixed() +
  geom_text(aes(label = value), color = "black", size = 4) +
  labs(
    title = "Correlation Heatmap of Selected Variables",
    x = NULL,
    y = NULL
  )


# Display
print(heatmap)
```




```r
#Save a high-resolution image
ggsave("correlation_heatmap.png", plot = heatmap, width = 12, height = 10, dpi = 300)
```

4 Discussion

Result: get a sharp, labeled heatmap with intuitive coloring:

Red = strong negative correlation

Blue = strong positive

White = neutral

4.1 Discuss your findings

5 Reproducibility

5.1 Github repository link

https://github.com/qgelena/Quantifying_a_social_problem/tree/main

5.2 Reference list

- Kawachi, I., & Berkman, L. F. (2000). Social cohesion, social capital, and health. *Social Epidemiology*, 174–190.
- Lund, C., Breen, A., Flisher, A. J., Kakuma, R., Corrigall, J., Joska, J. A., … & Patel, V. (2010). Poverty and common mental disorders in low and middle income countries: A systematic review. *Social Science & Medicine, 71*(3), 517–528. https://doi.org/10.1016/j.socscimed.2010.04.027
- Patel, V., Burns, J. K., Dhingra, M., Tarver, L., Kohrt, B. A., & Lund, C. (2018). Income inequality and depression: A systematic review and meta-analysis of the association and a scoping review of mechanisms. *World Psychiatry, 17*(1), 76–89. https://doi.org/10.1002/wps.20492
- Patel, V., Saxena, S., Lund, C., Thornicroft, G., Baingana, F., Bolton, P., … & UnÜtzer, J. (2022). The Lancet Commission on global mental health and sustainable development. *The Lancet, 392*(10157), 1553–1598. https://doi.org/10.1016/S0140-6736%2818%2931612-X
- Smith, K. E., Bambra, C., Hill, S. E., & Watt, R. G. (2012). Health inequalities and the social determinants of health: What works? *Journal of Public Health, 34*(4), 523–529. https://doi.org/10.1093/pubmed/fds052
- Wilkinson, R., & Pickett, K. (2009). *The Spirit Level: Why more equal societies almost always do better.* London: Allen Lane.
- World Health Organization. (2023). *Mental health.* Retrieved from https://www.who.int/health-topics/mental-health

---

5.3 Databases:

- Global Burden of Disease Collaborative Network. (2022). *Global Burden of Disease Study 2021 (GBD 2021) results.* Institute for Health Metrics and Evaluation (IHME). GBD Results
- Gapminder. (n.d.). *Population data documentation (GD003).* Gapminder Population
- World Bank. (n.d.). *World Development Indicators (WDI).* Retrieved June 2025, from World Bank WDI
- World Bank. (n.d.). *The world by income and region.* World by Income and Region
- World Bank. (n.d.). *Poverty and Inequality Platform: Gini index.* Gini Index – PIP
- Transparency International. (2020). *Corruption Perceptions Index (CPI) 2020.* CPI 2020

- World Bank. (n.d.). *House price to income ratio (IMF Global Housing Watch).* House Price to Income – World Bank
- OECD. (n.d.). *OECD house price statistics.* OECD House Prices
- International Monetary Fund. (n.d.). *Government expenditure, percent of GDP.* Government Expenditure – IMF
- World Bank. (n.d.). *Income share held by highest 10% (SI.DST.10TH.10).* Top 10% Income Share
- OECD. (n.d.). *Gini index – disposable income.* OECD Gini Index
- Qery. (n.d.). *Unemployment in OECD countries.* OECD Unemployment – Qery
- World Health Organization. (2022). *WHO Air Quality Database 2022.* WHO Air Quality
- World Population Review. (2023). *Depression rates by country.* Depression by Country
- World Population Review. (2023). *Anxiety rates by country.* Anxiety by Country