Midterm Project

Analysis of Global Health Indicators

KOMAL BHOSLE

RStudio Link

https://posit.cloud/spaces/603138/content/9862907

Library Calls

```
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   3.5.1     v tibble    3.2.1
v lubridate 1.9.3     v tidyr     1.3.1
v purrr     1.0.2
-- 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 beco
```

```
library(readxl)
library(readr)
```

Datasets

The first data set gives us the measures of Child Immunization for measles from age 12 - 23 months. It indicates the number of children in different developing countries of age between 12 to 23 months who received at least one dose of vaccination against measles. This immunization

measure plays a role towards making society healthier by preventing disease severity, hence saving lives of children. (Kabacoff,n.d.) To note on data wrangling steps- The data set was imported from the World bank group data set and converted to a csv file by using the function read.csv. Then from the csv file data the data was tidied to get details precise to three required columns, which is the Country, Year and the immunization rate.

```
# First Data set

Measles_immunization <-
  read.csv(
  file = "API_SH.IMM.MEAS_DS2_en_csv_v2_2015.csv",
  skip = 3
  )

Measles_immunization_clean <-
  Measles_immunization |>
  select(
    country = Country.Name,
    `X1960`:`X2023`
  ) |>
   pivot_longer(
    cols      = "X1960":"X2023",
    names_to  = "year",
    values_to = "immunization_rate"
  ) |>
  mutate(
    year = as.integer(sub("X","",year))
  ) |>
  drop_na(immunization_rate)

str(Measles_immunization_clean)
```

```
tibble [9,810 x 3] (S3: tbl_df/tbl/data.frame)
 $ country          : chr [1:9810] "Africa Eastern and Southern" "Africa Eastern and Southern
 $ year             : int [1:9810] 1980 1981 1982 1983 1984 1985 1986 1987 1988 1989 ...
 $ immunization_rate: num [1:9810] 9.13 17.03 19.72 31.8 35.52 ...
```

The second data set is about the Male employment rate, which gives the details about percentage of contribution of male employment in different countries. This indicates how employment can have an impact on the economy of different countries. (Kabacoff,n.d.) Note on data wrangling- The data set was imported from the World bank group data and converted it into an excel format by using the function read_excel. The data sheet was used and tidied

by selecting relevant columns of the Country, Year and employment rate. The missing values were handled, and a clean wrangled data set was created.

```
# Second Dataset

read_excel("API_SL.FAM.WORK.MA.ZS_DS2_en_excel_v2_916.xls")
```

```
New names:
* `` -> `...3`
* `` -> `...4`
* `` -> `...5`
* `` -> `...6`
* `` -> `...7`
* `` -> `...8`
* `` -> `...9`
* `` -> `...10`
* `` -> `...11`
* `` -> `...12`
* `` -> `...13`
* `` -> `...14`
* `` -> `...15`
* `` -> `...16`
* `` -> `...17`
* `` -> `...18`
* `` -> `...19`
* `` -> `...20`
* `` -> `...21`
* `` -> `...22`
* `` -> `...23`
* `` -> `...24`
* `` -> `...25`
* `` -> `...26`
* `` -> `...27`
* `` -> `...28`
* `` -> `...29`
* `` -> `...30`
* `` -> `...31`
* `` -> `...32`
* `` -> `...33`
* `` -> `...34`
* `` -> `...35`
* `` -> `...36`
```

```
* `` -> `...37`
* `` -> `...38`
* `` -> `...39`
* `` -> `...40`
* `` -> `...41`
* `` -> `...42`
* `` -> `...43`
* `` -> `...44`
* `` -> `...45`
* `` -> `...46`
* `` -> `...47`
* `` -> `...48`
* `` -> `...49`
* `` -> `...50`
* `` -> `...51`
* `` -> `...52`
* `` -> `...53`
* `` -> `...54`
* `` -> `...55`
* `` -> `...56`
* `` -> `...57`
* `` -> `...58`
* `` -> `...59`
* `` -> `...60`
* `` -> `...61`
* `` -> `...62`
* `` -> `...63`
* `` -> `...64`
* `` -> `...65`
* `` -> `...66`
* `` -> `...67`
* `` -> `...68`

# A tibble: 269 x 68
   `Data Source`      World Development In~1 ...3  ...4  ...5  ...6  ...7  ...8
   <chr>              <chr>                  <chr> <chr> <chr> <chr> <chr> <chr>
 1 Last Updated Date  45685                  <NA>  <NA>  <NA>  <NA>  <NA>  <NA>
 2 <NA>               <NA>                   <NA>  <NA>  <NA>  <NA>  <NA>  <NA>
 3 Country Name       Country Code           Indi~ Indi~ 1960  1961  1962  1963
 4 Aruba              ABW                    Cont~ SL.F~ <NA>  <NA>  <NA>  <NA>
 5 Africa Eastern an~ AFE                    Cont~ SL.F~ <NA>  <NA>  <NA>  <NA>
 6 Afghanistan        AFG                    Cont~ SL.F~ <NA>  <NA>  <NA>  <NA>
```

```
 7 Africa Western an~ AFW                   Cont~ SL.F~ <NA>  <NA>  <NA>  <NA>
 8 Angola             AGO                   Cont~ SL.F~ <NA>  <NA>  <NA>  <NA>
 9 Albania            ALB                   Cont~ SL.F~ <NA>  <NA>  <NA>  <NA>
10 Andorra            AND                   Cont~ SL.F~ <NA>  <NA>  <NA>  <NA>
# i 259 more rows
# i abbreviated name: 1: `World Development Indicators`
# i 60 more variables: ...9 <chr>, ...10 <chr>, ...11 <chr>, ...12 <chr>,
#   ...13 <chr>, ...14 <chr>, ...15 <chr>, ...16 <chr>, ...17 <chr>,
#   ...18 <chr>, ...19 <chr>, ...20 <chr>, ...21 <chr>, ...22 <chr>,
#   ...23 <chr>, ...24 <chr>, ...25 <chr>, ...26 <chr>, ...27 <chr>,
#   ...28 <chr>, ...29 <chr>, ...30 <chr>, ...31 <chr>, ...32 <chr>, ...
```

```r
Male_employment<- read_excel(
  path  = "API_SL.FAM.WORK.MA.ZS_DS2_en_excel_v2_916.xls",
  sheet = "Data",
  skip  = 3
)

Male_employment_contribution<-
  Male_employment |>
  pivot_longer(
    cols      = "1960":"2023",
    names_to  = "year",
    values_to = "employment_rate"
  ) |>
  rename(
    "country" = "Country Name"
  ) |>
  select(
    country,
    year,
    employment_rate
  ) |>
  mutate(
    year = as.integer(year)
  ) |>
  drop_na(employment_rate)
```

The joined data set called tidied_join was created by joining the above two data sets which have the world developmental indicators such as Measles immunization rate among children and Male employment rate. Both of which are collectively joined with Country and Year as common factors. (Kabacoff,2025) Note on data wrangling- The wrangled tidied final data sets of Measles immunization and Male employment were combined using the function inner_join.

```r
# Joined Dataset

tidied_join<-
  inner_join(
    x  = Measles_immunization_clean,
    y  = Male_employment_contribution,
    by = c("country", "year")
  )
```

Univariate Analyses

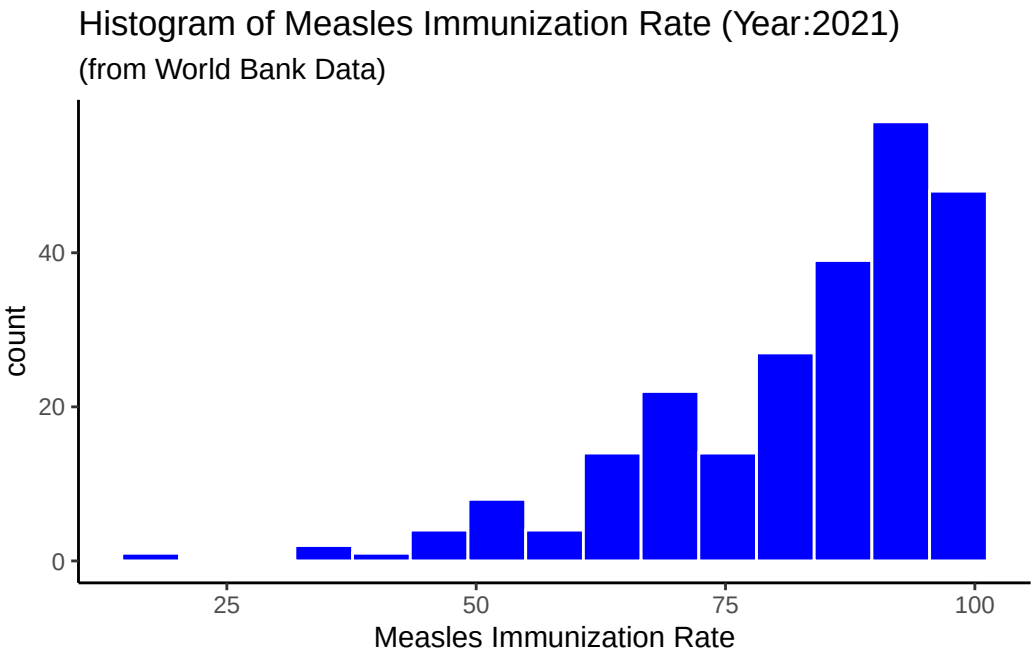
To visualize a data set as a histogram, the first tidied data set was used which consists of three columns of Country, Year and Immunization rate. Out of which just the data set of specific year 2021 was used to get a visual representation of immunization rate. Then to generate a histogram, the function ggplot was used to generate histogram and geom_histogram with labs to make it readable. The x axis being immunization rate with the frequency rate on the y axis. (Kabacoff,2025) The histogram shows an increasing rate with increasing count which indicates that there are more countries with higher immunization rate. This shows the positive health measures are being taken by majority of the countries.

```r
# Histogram

tidied_measles_same_year <-
  Measles_immunization_clean |>
  filter(
    year == 2021
  )

tidied_measles_same_year |>
  ggplot(
    mapping = aes(
      x     = immunization_rate
    )
  ) +
  geom_histogram(
    bins  = floor(sqrt(length(tidied_measles_same_year$immunization_rate))),
    fill  = "blue",
    color = "white"
  ) +
  labs(
    x        = "Measles Immunization Rate",
```

```
    title    = "Histogram of Measles Immunization Rate (Year:2021)",
    subtitle = "(from World Bank Data)"
  ) +
  theme_classic()
```



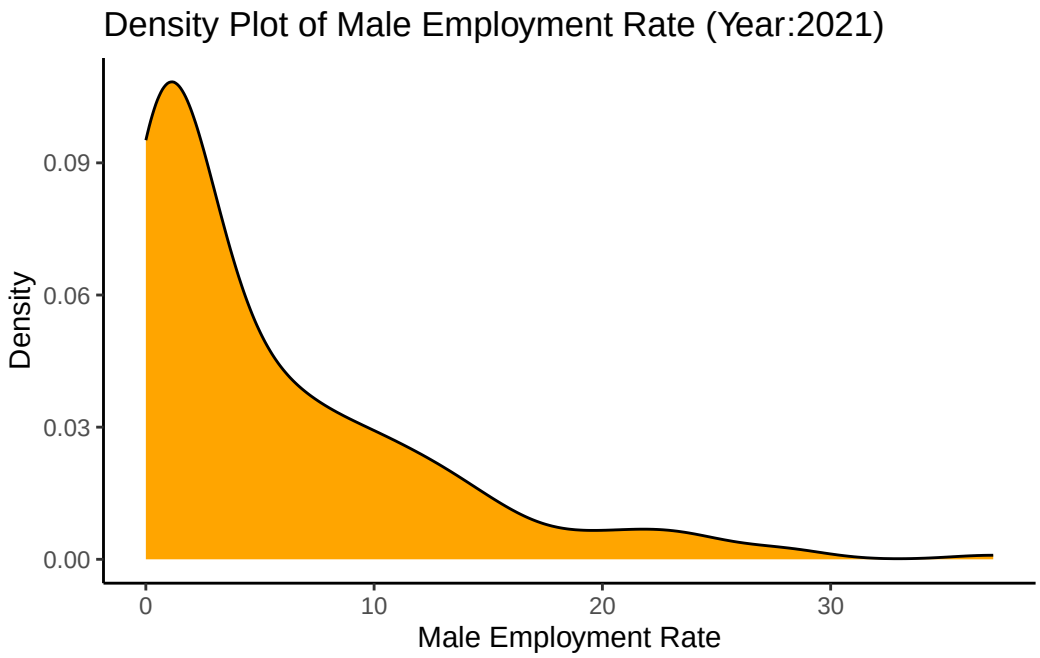
To visualize a data set as a density plot, the second tidied data set was used which consists of three columns of Country, Year and Employment rate. Out of which just the data set of specific year 2021 was used to get a visual representation of Employment rate. Then to generate a density plot, the function ggplot was used to generate density plot and geom_density with labs to make it readable. The x axis being male employment rate with density on y axis. (Kabacoff,2025) The density plot shows left skew distribution, which shows that a few number of countries have high male employment rate with majority of countries having moderate to low employment rates.

```
# Density Plot

tidied_employment_same_year <-
  Male_employment_contribution |>
  filter(
    year == 2021
  )

tidied_employment_same_year |>
```

```r
ggplot(
  mapping = aes(
    x     = employment_rate
  )
) +
geom_density(
  fill = "orange",
) +
labs(
  x     = "Male Employment Rate",
  y     = "Density",
  title = "Density Plot of Male Employment Rate (Year:2021)"
) +
theme_classic()
```



Bivariate Analyses

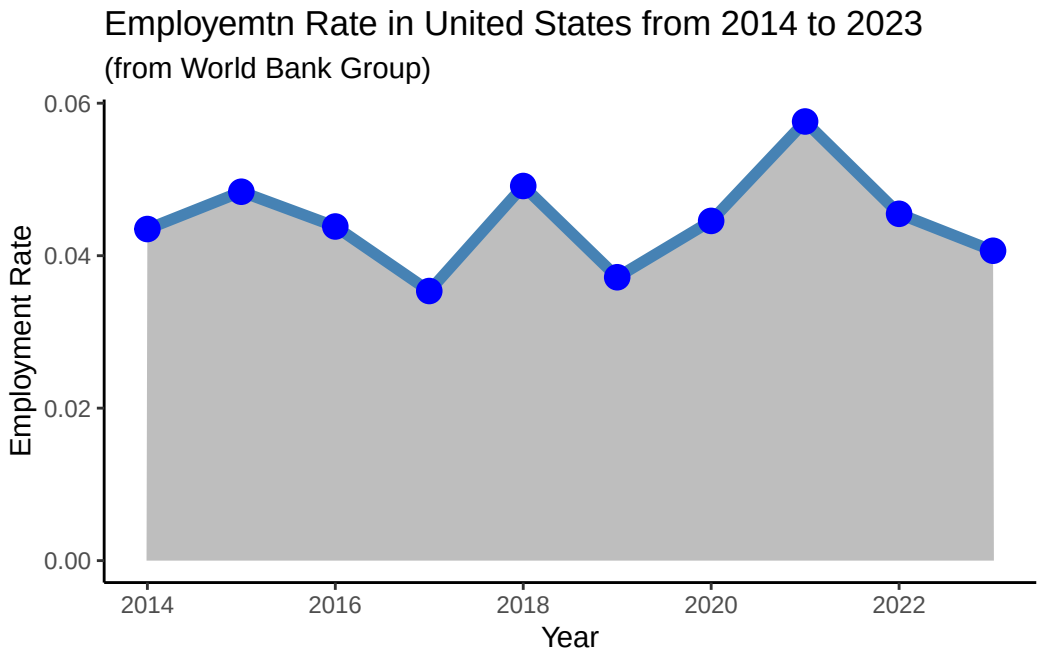
To visualize a data set in bivariate analysis, here is the Male employment rate with specification to "United States" for the ten year period from 2014 to 2023. This analysis was done by using the function ggplot to create a map, where geom_area is used to highlight the area spread of the analysis, geom_line gives a clear graph by joining the dots which we get from the function geom_point. The labs have Year on x axis and employment rate on Y axis. (Kabacoff,2025) To

understand the bivariate analysis in the line plot- there was a decline in the employment rate during 2019 with a gradual increase till 2021. To my knowledge this sharp shift could have been during the pandemic which has a major impact on employment and all major industries.

```
# Line Plot

United_States_data <-
  Male_employment_contribution |>
  filter(
    country == "United States",
    year >=2014 & year <=2023
  )

United_States_data |>
  ggplot(
    mapping = aes(
      x     = year,
      y     = employment_rate
    )
  ) +
  geom_area(
    fill = "grey"
  ) +
  geom_line(
    color     = "steelblue",
    linewidth = 2
  ) +
  geom_point(
    color = "blue",
    size  = 4
  ) +
  labs(
    x        = "Year",
    y        = "Employment Rate",
    title    = "Employemtn Rate in United States from 2014 to 2023",
    subtitle = "(from World Bank Group)"
  ) +
  theme_classic()
```



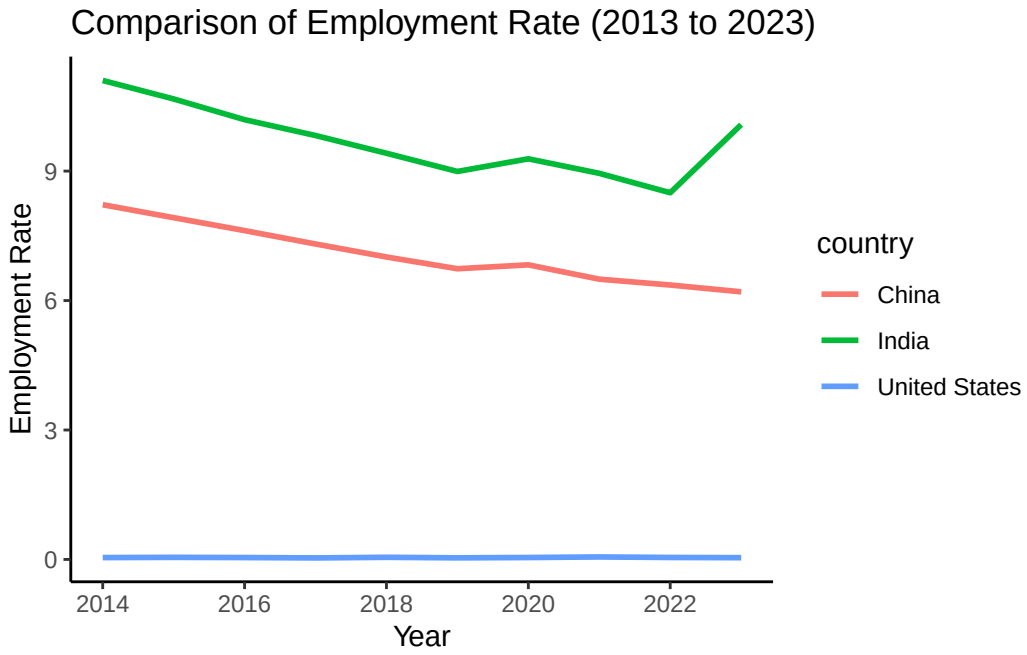
The multivariate plot between three different countries, United States, India and China over the 10-year period from 2014 to 2023 was analyzed to see the Male employment rate and how they vary among these countries during this time period. (Nahhas,2024) The function ggplot creates visualization plot with geom_line acting as an indicator of employment rate uniquely for different countries. Labs on X axis as Year and Employment rate on Y axis. (Kabacoff,2025) To analyze the line plot- The employment rate has been stable in the United States since a few years, while there was a sharp decline in China and increase in employment rate of India after 2022.

```
### Multivariate Plot -- Line Plot

Countries_Employment_rate <-
  tidied_join |>
    filter(country %in%
      c("United States", "India", "China"),
    year >=2014 & year <=2023
  )

Countries_Employment_rate |>
  ggplot(
    mapping = aes(
      x     = year,
      y     = employment_rate,
      color = country
```

```
  )
)+
geom_line(
  linewidth = 1
)+
labs(
  x     = "Year",
  y     = "Employment Rate",
  title = "Comparison of Employment Rate (2013 to 2023)"
)+
theme_classic()
```



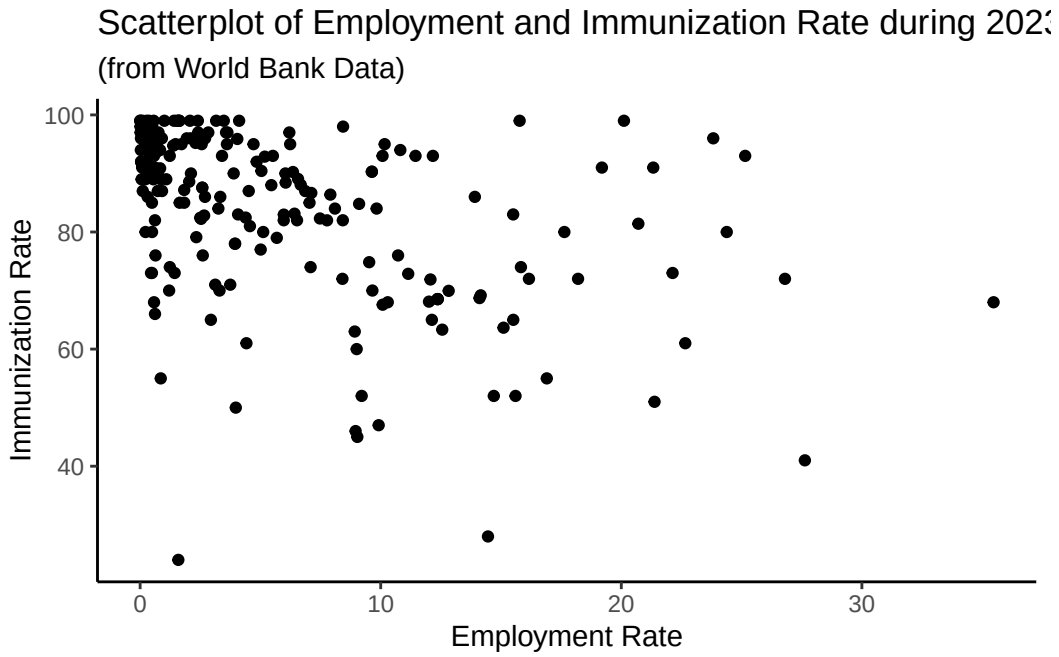
The scatter plot visualizes the relationship between Employment rate and Immunization rate for the year 2023. The function ggplot is used for data visualization, where labs have Employment rate on X axis and Immunization rate on Y axis. (Kabacoff,2025) To analyze the scatter plot- There is an indistinct spread where it looks like when there was an increase in employment rate, the immunization rate decreased. This could be due to various factors such as employers providing health coverage or decrease in employment made the government or public health sector provide immunization for unemployed people.

```
# Scatterplot

tidied_join_year_2023 <-
```

```
  tidied_join |>
  filter (
    year == 2023)

tidied_join_year_2023 |>
  ggplot(
    mapping = aes(
      x     = employment_rate,
      y     = immunization_rate,
    )
  ) +
  geom_point() +
  labs (
    x        = "Employment Rate",
    y        = "Immunization Rate",
    title    = "Scatterplot of Employment and Immunization Rate during 2023",
    subtitle = "(from World Bank Data)"
  ) +
  theme_classic()
```



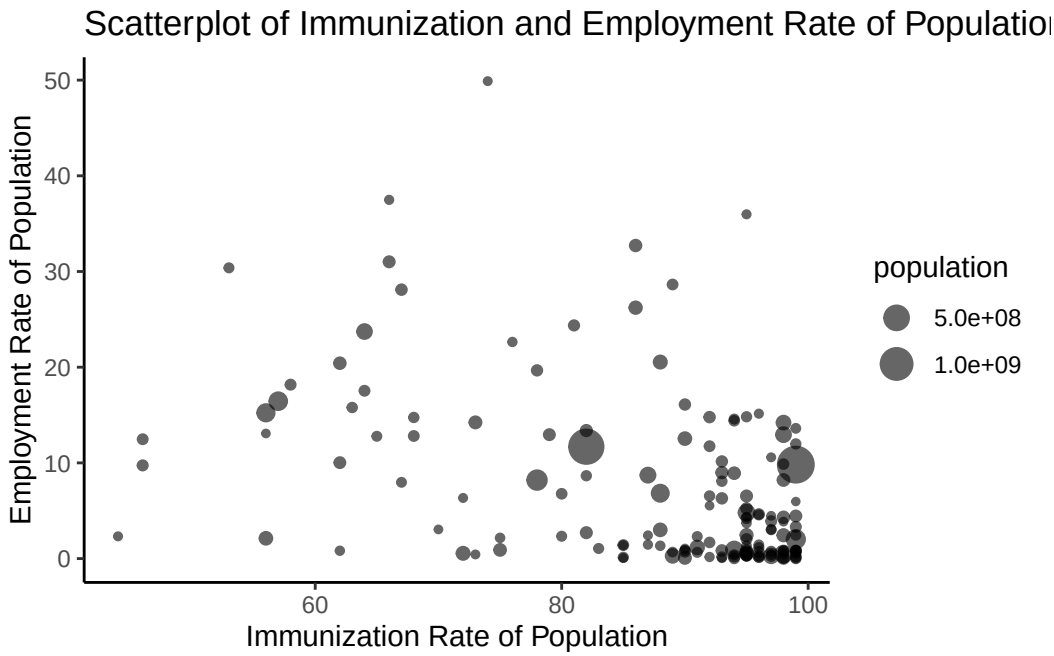
The multivariate scatter plot between the population data set and the joined data set was created to analyze the relation between Immunization rate and Employment rate. The inner_join function was used to combine the Country and year columns of both the data sets.

The rename function was used to name the newly created immunization and employment column. The data to be visualized is from the year 2010. Function ggplot maps the scatter plot, geom_point of alpha 0.6 was used. Labs Immunization rate on X axis and Employment rate on Y axis were plotted. (Kabacoff,2025) To analyze the scatter plot- With increasing employment rate, there was a decrease in immunization rate, with maximum immunization rate was seen when employment rate was the least.

```
### Multivariate Plot -- Scatterplot

tidied_population_join <-
  tidied_join |>
  inner_join(
    population,
    by = c("country", "year")
  ) |>
  rename(
    immunized_population_rate = immunization_rate,
    employed_population_rate = employment_rate
  ) |>
  filter(
    year == 2010
  )


tidied_population_join |>
  ggplot(
    mapping = aes(
      x     = immunized_population_rate,
      y     = employed_population_rate,
      size  = population
    )
  ) +
  geom_point(
    alpha = .6
  ) +
  labs (
    x     = "Immunization Rate of Population",
    y     = "Employment Rate of Population",
    title = "Scatterplot of Immunization and Employment Rate of Population during 2010",
    size  = "population",
    color = "country"
  ) +
  theme_classic()
```



References

Kabacoff, R. (n.d.). Chapter 2 Data Preparation | Modern Data Visualization with R. In rkabacoff.github.io. https://rkabacoff.github.io/datavis/DataPrep.html#importing

Kabacoff, R. (2025). Chapter 4 Univariate Graphs | Modern Data Visualization with R. Github.io. https://rkabacoff.github.io/datavis/Univariate.html#quantitative

Kabacoff, R. (2025). Chapter 5 Bivariate Graphs | Modern Data Visualization with R. Github.io. https://rkabacoff.github.io/datavis/Bivariate.html#quantitative-vs.-quantitative

Nahhas, R. W. (2024, June 25). Chapter 5 Descriptive statistics | An Introduction to R for Research. Bookdown.org. https://bookdown.org/rwnahhas/IntroToR/descriptives.html