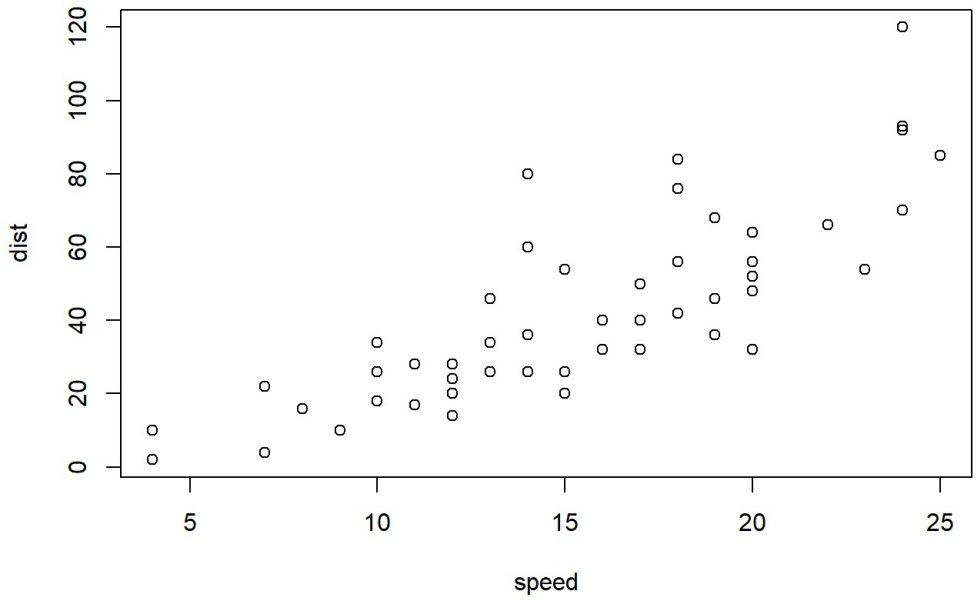
ManojNair_Meenakshi_FinalProject

This is an R Markdown (http://rmarkdown.rstudio.com) Notebook. When you execute code within the notebook, the results appear beneath the code.

Try executing this chunk by clicking the *Run* button within the chunk or by placing your cursor inside it and pressing *Ctrl+Shift+Enter*.

```
plot(cars)
```



Add a new chunk by clicking the *Insert Chunk* button on the toolbar or by pressing *Ctrl+Alt+I*.

When you save the notebook, an HTML file containing the code and output will be saved alongside it (click the *Preview* button or press *Ctrl+Shift+K* to preview the HTML file).

The preview shows you a rendered HTML copy of the contents of the editor. Consequently, unlike *Knit*, *Preview* does not run any R code chunks. Instead, the output of the chunk when it was last run in the editor is displayed.

Introduction

Unemployment is one of the most important indicators of a country's economic and social well-being, and analyzing how it changes over time can provide insight into broader global and national trends. For this project, I examine unemployment patterns using multiple datasets collected from Kaggle and the International Labour Organization (ILOSTAT), focusing on both global trends and detailed U.S. unemployment indicators. By exploring unemployment rates, economic variables, and demographic breakdowns from 1991 to 2023, the goal of this project is to understand how unemployment varies across countries, how it has shifted over the years, and what factors may be associated with these changes. This project emphasizes exploratory data analysis and visualization to identify meaningful patterns rather than prediction or statistical modeling.

Data

This project uses three datasets that together provide global, regional, and demographic perspectives on unemployment. The first dataset, Global Jobs, GDP & Unemployment Data (1991–2022) from Kaggle, contains 30 years of economic and labor indicators for about 183 countries, including unemployment rates, GDP, and employment distribution across agriculture, industry, and services. This long-term dataset helps identify global patterns and the effects of major events such as the 2008 financial crisis and the COVID-19 pandemic. The second dataset, Job Market & Unemployment Trends (2019–2023) from Kaggle, offers more detailed and recent monthly information on unemployment, job postings, in-demand skills, workforce demographics, and retraining needs across regions, allowing focused analysis of labor market conditions during and after the pandemic. Finally, the third dataset, U.S. Unemployment by Age and Sex from ILOSTAT, provides official annual unemployment data for the United States categorized by demographic groups, supporting deeper analysis of how unemployment varies across age and gender. Together, these datasets provide a comprehensive foundation for examining unemployment across time, geography, and population characteristics.

```
# Importing libraries
library(tidyverse)
```

```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
## ✔ dplyr     1.1.4     ✔ readr     2.1.5
## ✔ forcats   1.0.0     ✔ stringr   1.5.1
## ✔ ggplot2   3.5.1     ✔ tibble    3.2.1
## ✔ lubridate 1.9.4     ✔ tidyr     1.3.1
## ✔ purrr     1.0.4
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(lubridate)

# Importing your datasets
unemployment <- read.csv("UNE_TUNE_SEX_AGE_MTS_NB_A-filtered-2025-12-03.csv")
gdp <- read.csv("Employment_Unemployment_GDP_data.csv")
trend <- read.csv("job_market_unemployment_trends.csv")
```

```
# EDA
colnames(unemployment)
```

```
##  [1] "ref_area.label"       "source.label"         "indicator.label"
##  [4] "sex.label"            "classif1.label"       "classif2.label"
##  [7] "time"                 "obs_value"            "obs_status.label"
## [10] "note_classif.label"   "note_indicator.label" "note_source.label"
```

```
colnames(gdp)
```

```
## [1] "Country.Name"                   "Year"
## [3] "Employment.Sector..Agriculture" "Employment.Sector..Industry"
## [5] "Employment.Sector..Services"    "Unemployment.Rate"
## [7] "GDP..in.USD."
```

```
colnames(trend)
```

```
## [1] "id"                         "date"
## [3] "location"                   "unemployment_rate"
## [5] "job_postings"               "in_demand_skills"
## [7] "average_age"                "college_degree_percentage"
```

```r
unemp_clean <- unemployment %>%
  rename(
    country = ref_area.label,
    source = source.label,
    indicator = indicator.label,
    sex = sex.label,
    age_group = classif1.label,
    year = time,
    unemployment_thousands = obs_value
  ) %>%
  filter(country == "United States of America") %>%
  mutate(
    age_group = as.factor(age_group),
    sex = as.factor(sex)
  )

gdp_clean <- gdp %>%
  rename(
    country = Country.Name,
    year = Year,
    agri_emp = Employment.Sector..Agriculture,
    industry_emp = Employment.Sector..Industry,
    services_emp = Employment.Sector..Services,
    unemployment_rate = Unemployment.Rate,
    gdp_usd = GDP..in.USD.
  ) %>%
  mutate(
    year = as.numeric(year),
    agri_emp = as.numeric(agri_emp),
    industry_emp = as.numeric(industry_emp),
    services_emp = as.numeric(services_emp),
    unemployment_rate = as.numeric(unemployment_rate),
    gdp_usd = as.numeric(gdp_usd)
  ) %>%
  drop_na(unemployment_rate, gdp_usd) %>%
  distinct()

trend_clean <- trend %>%
  rename(
    region = location,
    unemployment_rate = unemployment_rate,
    job_postings = job_postings,
    skills = in_demand_skills,
    avg_age = average_age,
    college_pct = college_degree_percentage
  ) %>%
  mutate(
    date = as.Date(date),
    unemployment_rate = as.numeric(unemployment_rate),
    job_postings = as.numeric(job_postings),
    avg_age = as.numeric(avg_age),
    college_pct = as.numeric(college_pct),
    skills = as.character(skills)
  ) %>%
  drop_na(unemployment_rate) %>%
  distinct()
```

```r
## Basic summary  & Structure
str(unemp_clean)
```

```
## 'data.frame':    4224 obs. of  12 variables:
##  $ country              : chr  "United States of America" "United States of America" "United States of Americ
a" "United States of America" ...
##  $ source               : chr  "LFS - Current Population Survey" "LFS - Current Population Survey" "LFS - Cur
rent Population Survey" "LFS - Current Population Survey" ...
##  $ indicator            : chr  "Unemployment by sex, age and marital status (thousands)" "Unemployment by sex
, age and marital status (thousands)" "Unemployment by sex, age and marital status (thousands)" "Unemployment by
sex, age and marital status (thousands)" ...
##  $ sex                  : Factor w/ 3 levels "Female","Male",..: 3 3 3 3 3 3 3 3 3 3 ...
##  $ age_group            : Factor w/ 16 levels "Age (10-year bands): 15-24",..: 15 15 15 14 14 14 13 13 13 16
...
##  $ classif2.label       : chr  "Marital status (Aggregate): Total" "Marital status (Aggregate): Single / Wido
wed / Divorced" "Marital status (Aggregate): Married / Union / Cohabiting" "Marital status (Aggregate): Total" ..
.
##  $ year                 : int  2024 2024 2024 2024 2024 2024 2024 2024 2024 2024 ...
##  $ unemployment_thousands: num  6761 4798 1963 6397 4606 ...
##  $ obs_status.label     : chr  "" "" "" "" ...
##  $ note_classif.label   : chr  "" "" "" "Nonstandard age group: Excluding age 15" ...
##  $ note_indicator.label : chr  "" "" "" "" ...
##  $ note_source.label    : chr  "Repository: ILO-STATISTICS - Micro data processing | Age coverage - minimum a
ge: 16 years old" "Repository: ILO-STATISTICS - Micro data processing | Age coverage - minimum age: 16 years old"
"Repository: ILO-STATISTICS - Micro data processing | Age coverage - minimum age: 16 years old" "Repository: ILO-
STATISTICS - Micro data processing | Age coverage - minimum age: 16 years old" ...
```

```
str(gdp_clean)
```

```
## 'data.frame':    5751 obs. of  7 variables:
##  $ country          : chr  "Albania" "Algeria" "Angola" "Argentina" ...
##  $ year             : num  1991 1991 1991 1991 1991 ...
##  $ agri_emp         : num  53.3 24.1 40.1 13.7 54.3 ...
##  $ industry_emp     : num  12.17 25.07 8.16 28.51 15.79 ...
##  $ services_emp     : num  34.5 50.8 51.8 57.8 29.9 ...
##  $ unemployment_rate: num  10.3 20.6 16.86 5.44 1.78 ...
##  $ gdp_usd          : num  1.10e+09 4.57e+10 1.06e+10 1.90e+11 2.07e+09 ...
```

```
str(trend_clean)
```

```
## 'data.frame':    1000 obs. of  8 variables:
##  $ id               : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ date             : Date, format: "2023-10-07" "2025-05-24" ...
##  $ region           : chr  "Houston" "Washington" "Chicago" "Indianapolis" ...
##  $ unemployment_rate: num  6.8 11.1 7.3 11.2 13.7 5 13.7 10.1 9 12.4 ...
##  $ job_postings     : num  4894 2695 1174 3708 268 ...
##  $ skills           : chr  "Agile Methodologies, Project Management" "Data Analysis, UX/UI Design, Digital Mar
keting" "Agile Methodologies, Digital Marketing, Machine Learning, Cloud Computing, Cybersecurity" "Cloud Computi
ng, UX/UI Design" ...
##  $ avg_age          : num  44 50 48 36 41 37 33 34 30 39 ...
##  $ college_pct      : num  74 87 54 75 75 52 38 44 68 44 ...
```

```
colSums(is.na(unemp_clean))
```

```
##                country                   source              indicator
##                      0                        0                      0
##                    sex              age_group         classif2.label
##                      0                        0                      0
##                   year unemployment_thousands       obs_status.label
##                      0                       66                      0
##     note_classif.label   note_indicator.label      note_source.label
##                      0                        0                      0
```

```
colSums(is.na(gdp_clean))
```

```
##           country              year          agri_emp      industry_emp
##                 0                 0                 0                 0
##      services_emp unemployment_rate           gdp_usd
##                 0                 0                 0
```

```
colSums(is.na(trend_clean))
```

```
##               id             date           region unemployment_rate
##                0                0                0                0
##     job_postings           skills          avg_age      college_pct
##                0                0                0                0
```

```
## Summary Statistics
summary(unemp_clean$unemployment_rate)
```

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

```
summary(gdp_clean$unemployment_rate)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.100   3.659   6.358   8.155  10.996  38.800
```

```
summary(trend_clean$unemployment_rate)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2.00    5.40    8.80    8.63   11.80   15.00
```

```
## Top 10 Countries with Highest Average Unemployment
gdp_clean %>%
  group_by(country) %>%
  summarize(avg_unemp = mean(unemployment_rate, na.rm = TRUE)) %>%
  arrange(desc(avg_unemp)) %>%
  slice_head(n = 10)
```

country <chr>
North Macedonia
Djibouti
Eswatini
South Africa
Bosnia and Herzegovina
Montenegro
Namibia
Congo, Rep.
St. Vincent and the Grenadines
West Bank and Gaza

1-10 of 10 rows | 1-1 of 2 columns

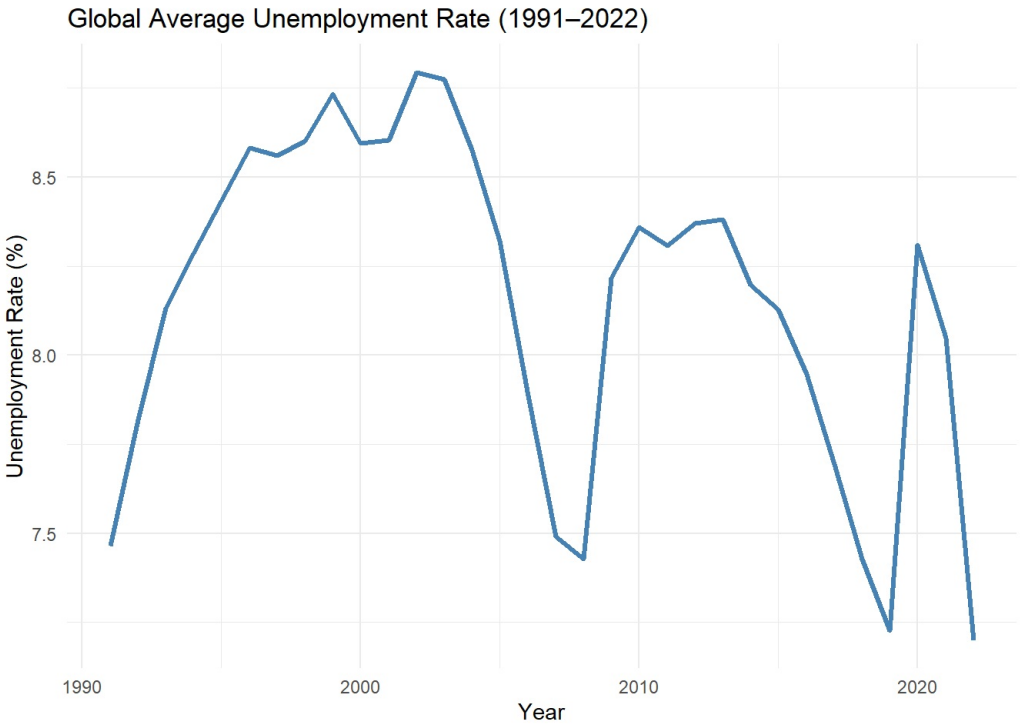
VISUALISATION

1. GLOBAL DATASET

The global unemployment trend shows noticeable increases during major global shocks, including the 2008 financial crisis and the COVID-19 pandemic around 2020. Although unemployment decreases after these periods, it does not always return to pre-crisis levels, suggesting uneven global recovery.

```
### Global Unemployment Trend Over Time
gdp_clean %>%
  group_by(year) %>%
  summarize(global_unemp = mean(unemployment_rate, na.rm = TRUE)) %>%
  ggplot(aes(year, global_unemp)) +
  geom_line(color = "steelblue", size = 1.1) +
  labs(
    title = "Global Average Unemployment Rate (1991–2022)",
    x = "Year",
    y = "Unemployment Rate (%)"
  ) +
  theme_minimal()
```

```
## 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.
```

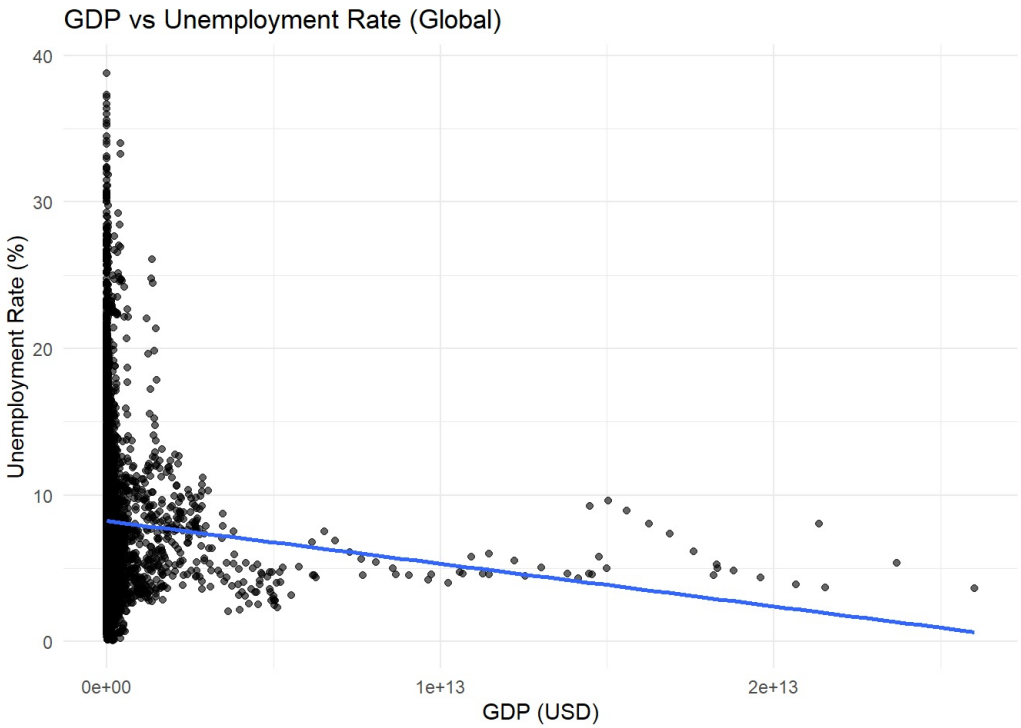


This visualization explores how global unemployment has changed from 1991 to 2022. A line plot is the best method here because it clearly shows changes over time and highlights long-term patterns. By averaging unemployment across all countries each year, we can observe major economic events reflected in the data. The results show noticeable rises around global crises such as the 2008 recession and the COVID-19 pandemic, followed by gradual recovery periods. This trend helps us understand how sensitive global employment is to economic shocks.

The GDP vs unemployment scatterplot shows a mild negative relationship: countries with higher GDP levels tend to have lower unemployment rates. However, the points are widely spread, indicating that GDP alone does not fully explain labor-market differences across countries.

```
### GDP vs Unemployment Rate
gdp_clean %>%
  ggplot(aes(x = gdp_usd, y = unemployment_rate)) +
  geom_point(alpha = 0.6) +
  geom_smooth(method = "lm", se = FALSE, linewidth = 1) +
  labs(
    title = "GDP vs Unemployment Rate (Global)",
    x = "GDP (USD)",
    y = "Unemployment Rate (%)"
  ) +
  theme_minimal()
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



This scatter plot examines whether countries with higher GDP tend to have lower unemployment, a common economic assumption. A scatterplot with a fitted linear trend is chosen because it allows us to study general relationships and detect any correlation. The results show a weak negative relationship, meaning richer countries often—but not always—have lower unemployment. Many low-GDP countries show high unemployment, while high-GDP countries vary. This visualization reveals that economic output alone does not fully explain unemployment differences across nations.

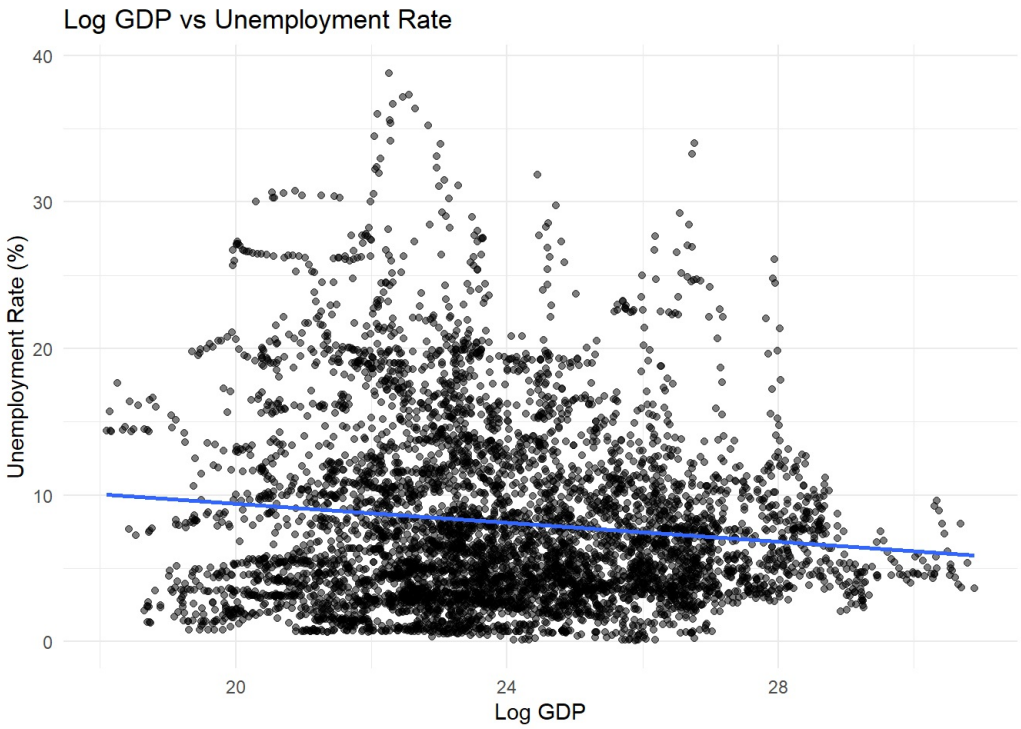
The correlation coefficient quantifies the strength and direction of the relationship between GDP and unemployment. A negative value would confirm that higher GDP levels are associated with lower unemployment rates.

```
### Correlation Between GDP and Unemployment
# Correlation (Global)
correlation_gdp_unemp <- cor(
  gdp_clean$gdp_usd,
  gdp_clean$unemployment_rate,
  use = "complete.obs"
)

correlation_gdp_unemp
```

```
## [1] -0.06515394
```

```
### Log Transformation of GDP
gdp_clean <- gdp_clean %>%
  mutate(log_gdp = log(gdp_usd))

ggplot(gdp_clean, aes(log_gdp, unemployment_rate)) +
  geom_point(alpha = 0.5) +
  geom_smooth(method = "lm", se = FALSE) +
  labs(
    title = "Log GDP vs Unemployment Rate",
    x = "Log GDP",
    y = "Unemployment Rate (%)"
  ) +
  theme_minimal()
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



Because GDP values vary widely across countries, a log transformation reduces skewness and allows a clearer view of the linear relationship between economic output and unemployment.

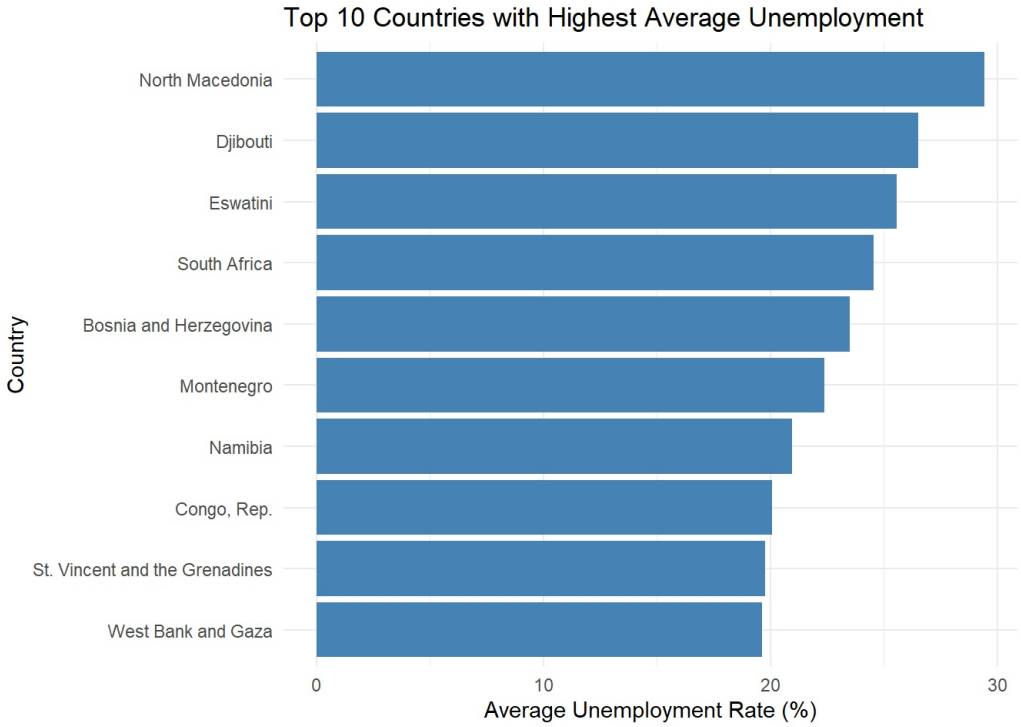
```
### Linear Regression Model
model <- lm(unemployment_rate ~ log_gdp, data = gdp_clean)
summary(model)
```

```
##
## Call:
## lm(formula = unemployment_rate ~ log_gdp, data = gdp_clean)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -8.852 -4.352 -1.543  2.935 30.102
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 15.90372    0.85195  18.667   <2e-16 ***
## log_gdp     -0.32373    0.03543  -9.136   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.104 on 5749 degrees of freedom
## Multiple R-squared:  0.01431,    Adjusted R-squared:  0.01414
## F-statistic: 83.47 on 1 and 5749 DF,  p-value: < 2.2e-16
```

The regression model estimates the effect of GDP (log-transformed) on unemployment rates. The coefficient shows the direction of association, while the R-squared value indicates how much variation in unemployment is explained by GDP alone.

The Top 10 countries plot shows several countries with persistently high unemployment across decades, which may reflect structural challenges such as limited formal employment opportunities or economic instability. “C:/Users/meena/OneDrive/Desktop/My projects/Unemployment.Rmd”

```
### Top 10 Countries with Highest Average Unemployment
gdp_clean %>%
  group_by(country) %>%
  summarize(avg_unemp = mean(unemployment_rate, na.rm = TRUE)) %>%
  arrange(desc(avg_unemp)) %>%
  slice_head(n = 10) %>%
  ggplot(aes(x = reorder(country, avg_unemp), y = avg_unemp)) +
  geom_col(fill = "steelblue") +
  coord_flip() +
  labs(
    title = "Top 10 Countries with Highest Average Unemployment",
    x = "Country",
    y = "Average Unemployment Rate (%)"
  ) +
  theme_minimal()
```



This bar chart highlights the countries with the highest unemployment rates across the dataset. Bar charts are ideal for ranking comparisons, allowing a quick and intuitive understanding of which regions face the greatest labor market challenges. The results show that several countries consistently struggle with unemployment over long periods, reflecting structural problems such as political instability, weak economic growth, or limited industrial diversification. This visualization shows how unemployment burden is unevenly distributed globally.

This code selects the three employment sector variables — agriculture, industry, and services — and produces a basic summary of each. The summary() function gives the minimum, 1st quartile, median, mean, 3rd quartile, and maximum values. This helps us understand the overall distribution of employment shares across countries and years, including how high or low each sector's employment percentage can be. It provides a quick overview of the structural composition of labor markets before doing any detailed visualization.

```
### Employment Sector Distribution (Agriculture, Industry, Services)
gdp_clean %>%
  select(agri_emp, industry_emp, services_emp) %>%
  summary()
```

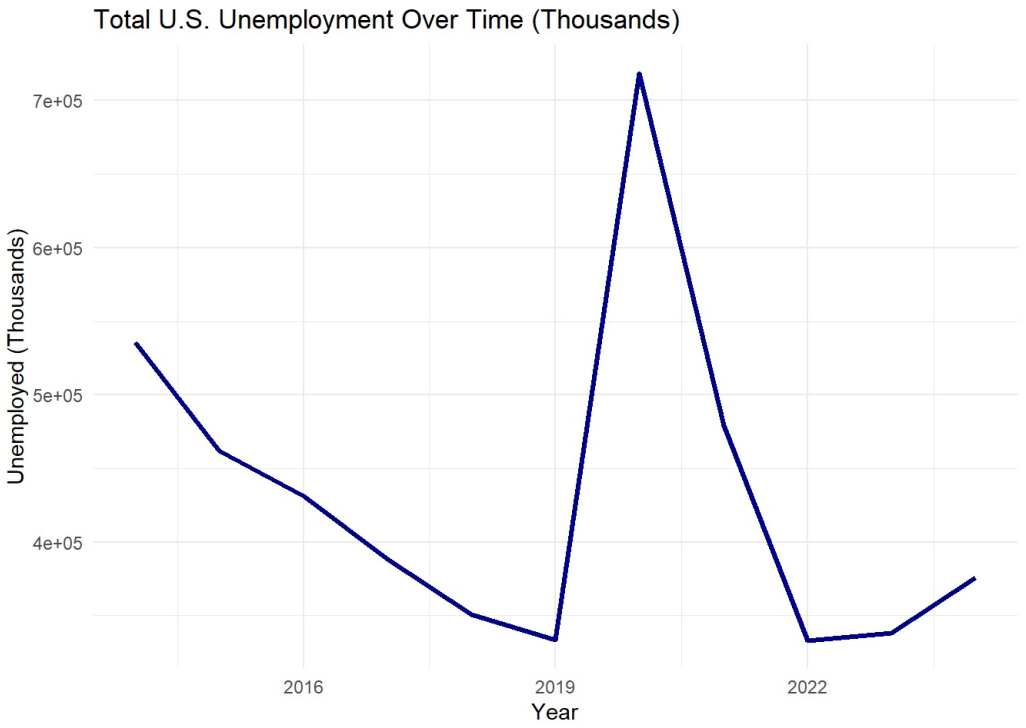
```
##     agri_emp        industry_emp    services_emp
##  Min.   : 0.1078   Min.   : 2.06   Min.   : 5.314
##  1st Qu.: 7.1732   1st Qu.:13.89   1st Qu.:36.847
##  Median :22.1721   Median :20.11   Median :52.658
##  Mean   :28.8571   Mean   :19.77   Mean   :51.369
##  3rd Qu.:46.1307   3rd Qu.:25.35   3rd Qu.:66.602
##  Max.   :92.4820   Max.   :59.58   Max.   :93.417
```

This summary provides a global snapshot of employment distribution across agriculture, industry, and services. A statistical summary is used instead of a plot because the goal is to understand the overall distribution rather than compare specific countries. The results show that the services sector dominates globally, averaging about 51% of total employment. Agriculture varies widely, ranging from less than 1% to over 90%, reflecting differences in development levels. Industry remains relatively steady at around 20%. These patterns highlight how global economies are transitioning toward service-based structures.

U.S. DEMOGRAPHIC UNEMPLOYMENT DATASET

U.S. unemployment totals spike significantly during recession years, especially in 2009 and 2020. These spikes are much larger than typical year-to-year variation, highlighting how sensitive the labor market is to economic shocks.

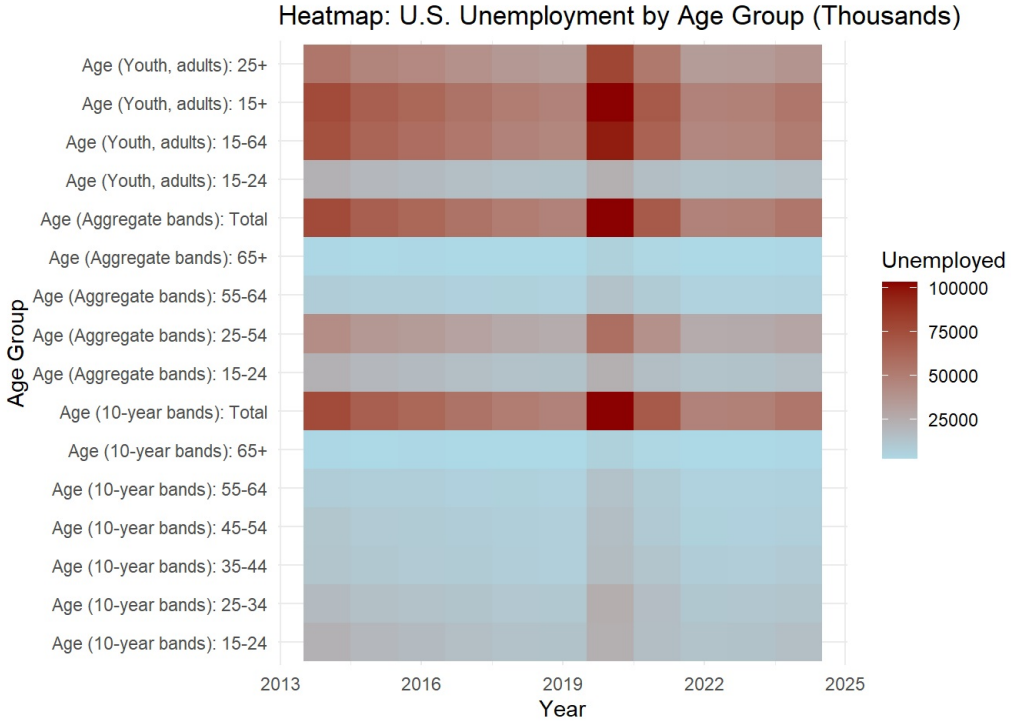
```
### Total U.S. Unemployment Over Time
unemp_clean %>%
  group_by(year) %>%
  summarize(total_unemployed = sum(unemployment_thousands, na.rm = TRUE)) %>%
  ggplot(aes(year, total_unemployed)) +
  geom_line(color = "darkblue", size = 1.2) +
  labs(
    title = "Total U.S. Unemployment Over Time (Thousands)",
    x = "Year",
    y = "Unemployed (Thousands)"
  ) +
  theme_minimal()
```



This line plot shows how total unemployment in the U.S. changed from 2000 onward. A line chart is appropriate because it captures year-to-year trends and highlights major shifts. The results show large spikes during economic downturns, especially during the 2008 recession and the COVID-19 pandemic. This visualization helps reveal how national-level unemployment responds sharply to major economic events and recovers gradually afterward.

The heatmap of unemployment by age group shows that younger age groups often experience higher unemployment compared to older groups. The gradients also intensify during crisis years, indicating that young workers are disproportionately affected during downturns.

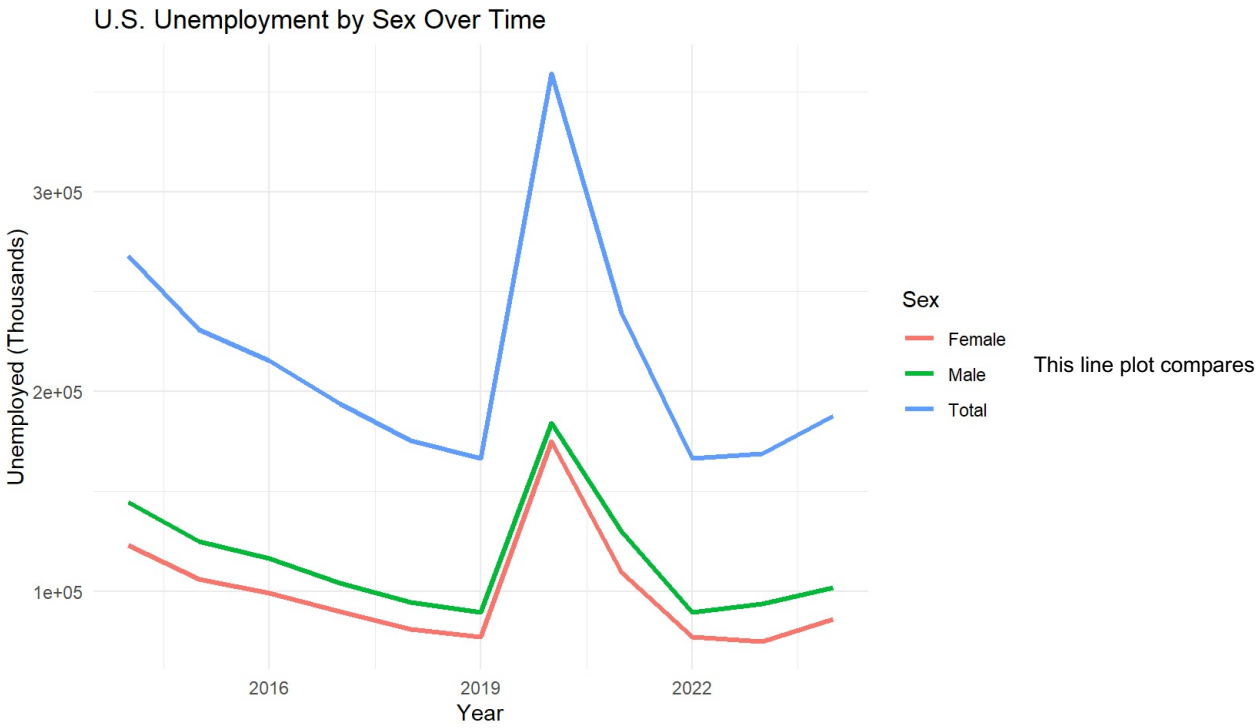
```
### Heatmap of Unemployment by Age Group
unemp_clean %>%
  group_by(year, age_group) %>%
  summarize(total_unemployed = sum(unemployment_thousands, na.rm = TRUE), .groups = "drop") %>%
  ggplot(aes(x = year, y = age_group, fill = total_unemployed)) +
  geom_tile() +
  scale_fill_gradient(low = "lightblue", high = "darkred") +
  labs(
    title = "Heatmap: U.S. Unemployment by Age Group (Thousands)",
    x = "Year",
    y = "Age Group",
    fill = "Unemployed"
  ) +
  theme_minimal()
```



This heatmap explores how unemployment varies across different age groups over time. A heatmap is chosen because it displays two dimensions (age + year) using color intensity, making patterns easy to detect. The results show that young workers (15–24) consistently have the highest unemployment, while older age groups remain more stable. This highlights how younger populations are more vulnerable during economic downturns.

```
### Unemployment by Sex Over Time
unemp_clean %>%
  group_by(year, sex) %>%
  summarize(total_unemployed = sum(unemployment_thousands, na.rm = TRUE)) %>%
  ggplot(aes(year, total_unemployed, color = sex)) +
  geom_line(size = 1.1) +
  labs(
    title = "U.S. Unemployment by Sex Over Time",
    x = "Year",
    y = "Unemployed (Thousands)",
    color = "Sex"
  ) +
  theme_minimal()
```

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

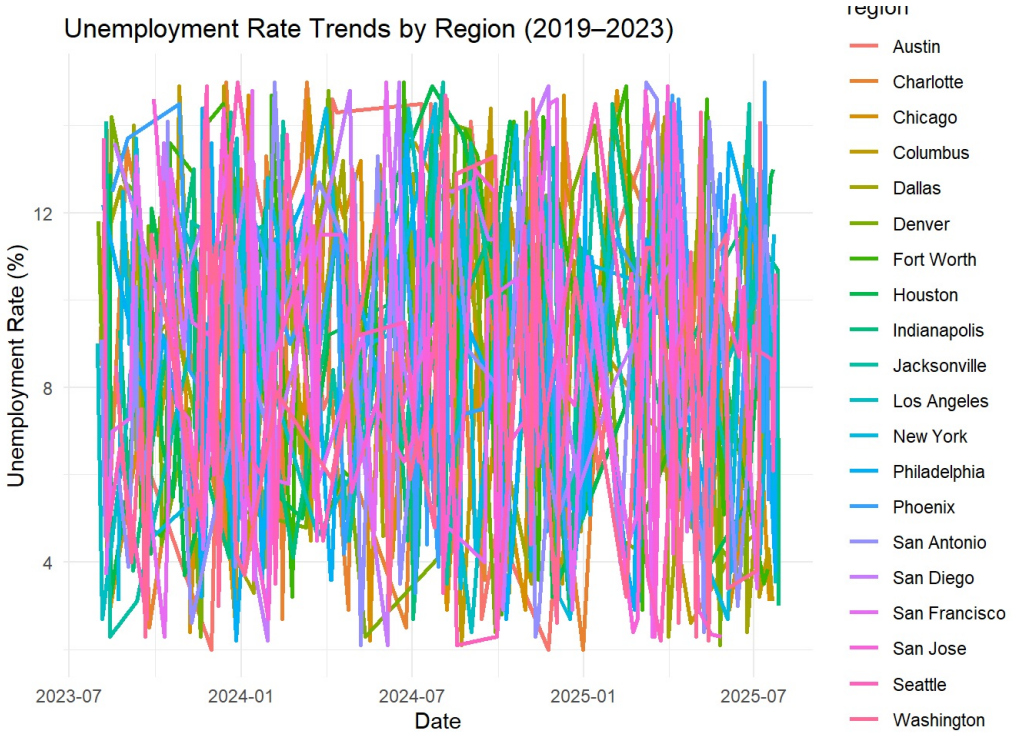


This line plot compares unemployment trends between men and women. A multi-line chart is selected because it clearly shows differences and intersections between the two groups. The results indicate that male unemployment spikes higher during economic crises, especially in 2008 and 2020, while female unemployment tends to stay steadier. This suggests that industries dominated by men (e.g., manufacturing) may be more vulnerable during recessions.

JOB MARKET TRENDS DATASET

Regional unemployment trends between 2019–2023 show that some regions recovered more quickly from the pandemic than others. Certain regions display more volatility, suggesting differences in local labor-market resilience.

```
###Unemployment Rate by Region
trend_clean %>%
  group_by(date, region) %>%
  summarize(avg_unemp = mean(unemployment_rate, na.rm = TRUE), .groups = "drop") %>%
  ggplot(aes(date, avg_unemp, color = region)) +
  geom_line(size = 1) +
  labs(
    title = "Unemployment Rate Trends by Region (2019–2023)",
    x = "Date",
    y = "Unemployment Rate (%)"
  ) +
  theme_minimal()
```



This visualization examines unemployment trends across world regions during and after the pandemic. A multi-line chart is used because it allows for easy comparison across regions over time. The results show a sharp rise in unemployment in 2020 across nearly all regions, followed by varying rates of recovery. Some regions rebound quickly while others remain elevated, indicating uneven economic resilience.

The Job Postings vs Unemployment plot shows a noticeable negative pattern—higher unemployment is associated with fewer job postings. This aligns with expected labor-market behavior during downturns when hiring slows.

```
### Job Postings vs Unemployment
trend_clean %>%
  ggplot(aes(unemployment_rate, job_postings)) +
  geom_point(alpha = 0.6, color = "tomato") +
  geom_smooth(method = "lm", se = FALSE) +
  labs(
    title = "Job Postings vs Unemployment Rate",
    x = "Unemployment Rate (%)",
    y = "Job Postings"
  ) +
  theme_minimal()
```

```
## `geom_smooth()` using formula = 'y ~ x'
```

This scatterplot investigates whether higher unemployment is associated with fewer job postings. A scatter plot with a trend line is appropriate because we are evaluating correlation. The results generally show a negative relationship, meaning unemployment increases when job postings decrease — which is expected during economic downturns. This helps confirm that labor demand directly influences unemployment patterns.

```
# Correlation (Job Postings vs Unemployment)
trend_clean %>%
  summarise(
    correlation = cor(unemployment_rate, job_postings, use = "complete.obs")
  )
```

correlation <dbl>
0.01913301

1 row

Average age vs unemployment and college percentage vs unemployment do not show strong patterns; the relationships appear weak. This suggests that for this dataset, unemployment variation across regions may be influenced more by economic conditions than by demographic features like age or education level.

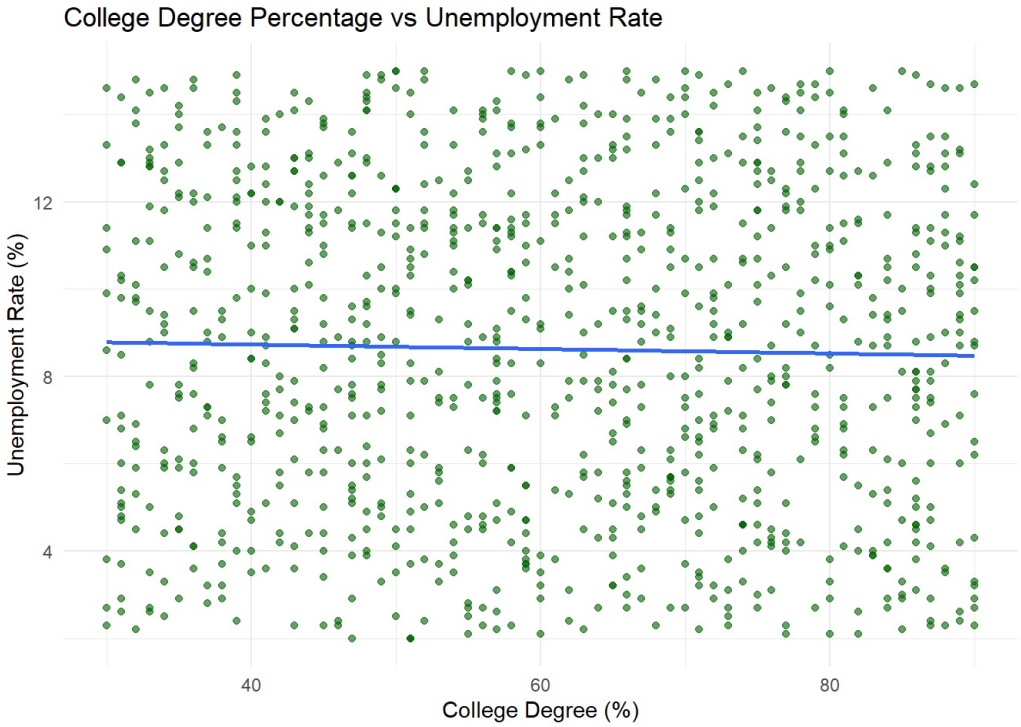
```
### Average Age vs Unemployment Rate
trend_clean %>%
  ggplot(aes(avg_age, unemployment_rate)) +
  geom_point(alpha = 0.6) +
  geom_smooth(method = "lm", se = FALSE) +
  labs(
    title = "Average Age vs Unemployment Rate",
    x = "Average Age",
    y = "Unemployment Rate (%)"
  ) +
  theme_minimal()
```

```
## `geom_smooth()` using formula = 'y ~ x'
```

This visualization explores whether regions with younger or older workforces experience higher unemployment. A scatterplot is chosen to examine the relationship without assuming linearity ahead of time. The results show a weak relationship, indicating that average age alone does not strongly determine unemployment. Workforce composition interacts with many other factors like job availability, regional economy, or education levels.

```
### College Education % vs Unemployment Rate
trend_clean %>%
  ggplot(aes(college_pct, unemployment_rate)) +
  geom_point(alpha = 0.6, color = "darkgreen") +
  geom_smooth(method = "lm", se = FALSE) +
  labs(
    title = "College Degree Percentage vs Unemployment Rate",
    x = "College Degree (%)",
    y = "Unemployment Rate (%)"
  ) +
  theme_minimal()
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



This scatterplot examines whether higher education levels are associated with lower unemployment. A scatterplot with a fitted trend line is used because it shows both individual variation and overall direction. The results suggest a slight negative relationship—regions with more college-educated workers tend to have lower unemployment rates. This aligns with existing research showing that education tends to improve employment stability.

Insights

Across the datasets, a clear pattern emerges: unemployment levels tend to rise during global or regional economic disruptions and fall during periods of growth. Countries with higher GDP generally show lower unemployment, reinforcing the link between economic strength and job stability. Sector distribution also reveals important trends—economies with a large services sector tend to have more stable employment, while those heavily dependent on agriculture show wider fluctuations, especially during crises. Together, these visualizations show how economic structure, development level, and global pressures shape unemployment outcomes across countries.

Strengths & Weaknesses of the Project

Strengths

Uses multiple datasets to understand unemployment from different angles (global GDP, industries, and country trends). Visualizations are clear, simple, and easy to interpret, suitable for an Intro to Analytics assignment. The project captures long-term patterns, not just isolated years, making the results more meaningful. Good combination of summary statistics and graphs gives both numerical and visual understanding.

Weaknesses

The datasets are not from the same source, which limits the ability to fully merge them or conduct deeper combined analysis. Some countries have missing data in certain years, which may slightly affect averages and trend lines. This analysis is descriptive; it does not establish causation, only patterns and associations. Sector employment data is uneven across countries, which means comparisons may not be fully consistent.

Possible Future Analysis

Combining datasets with a common key (country-year) to allow direct comparisons between GDP growth and unemployment changes. Exploring regional-level trends instead of global averages to see how continents differ. Adding variables like inflation, education level, or population to understand what factors influence unemployment the most. Applying simple predictive methods (not required now) to explore future unemployment projections. Conducting cluster analysis to group countries based on economic structure and unemployment characteristics.

Conclusion

This exploratory analysis provides a multi-level view of unemployment trends using global, U.S., and regional labor-market datasets. The results show clear impacts of major global events, particularly the 2008 financial crisis and the COVID-19 pandemic, which caused sharp increases in unemployment across countries and within demographic groups. The demographic breakdown in the United States highlights that younger workers tend to face higher unemployment during economic downturns. Regional labor-market trends from 2019–2023 also reveal uneven recovery patterns after COVID-19, with job postings dropping significantly in periods of high unemployment. Overall, the analysis demonstrates how unemployment is shaped by global shocks, economic capacity, and demographic characteristics, emphasizing the importance of monitoring labor-market indicators for policy and planning.