Exploratory Data Analysis

Saurabh

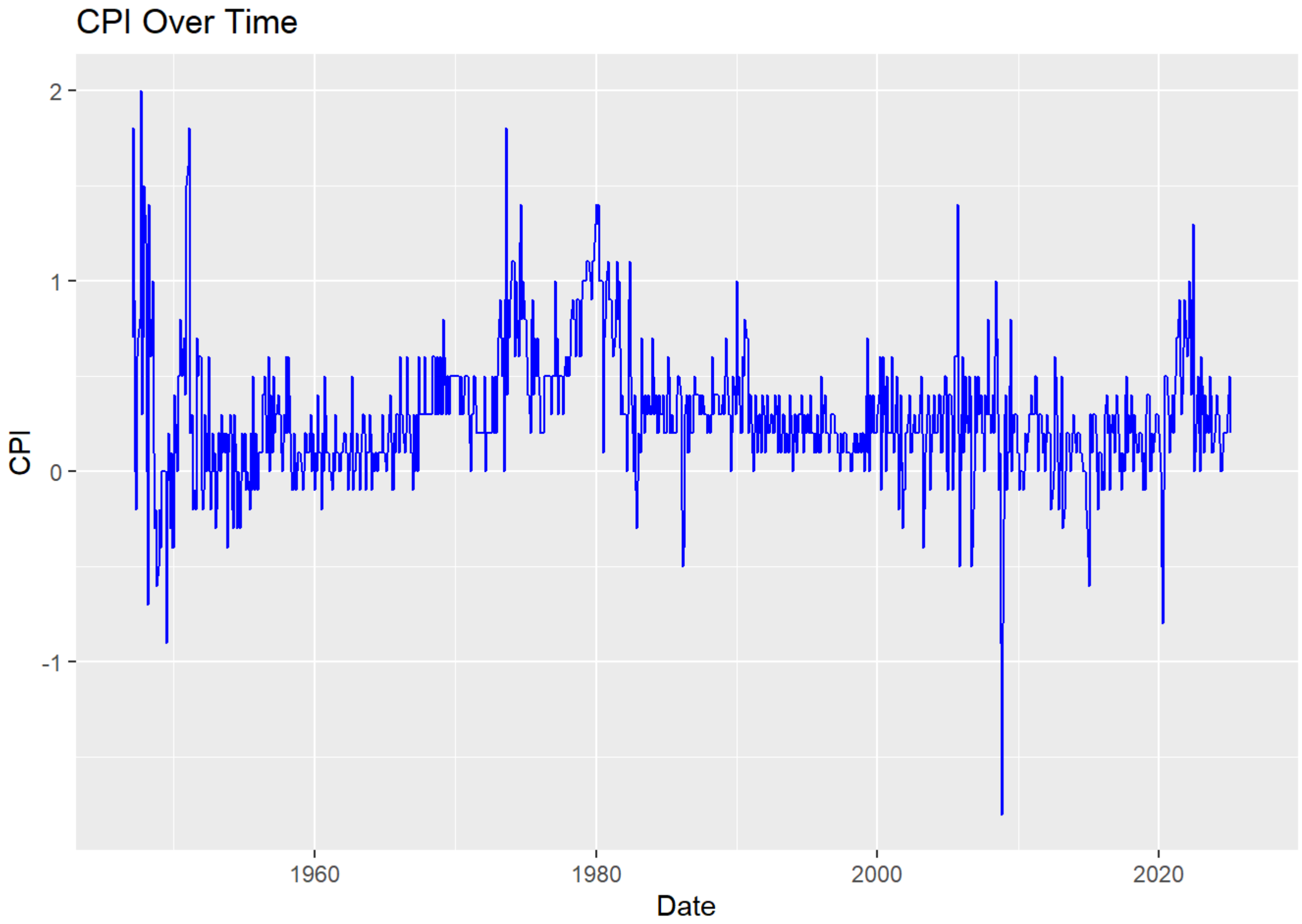
2025-05-04

a) Exploratory data analysis

- Visualization: line plots to visualize trends and scatter plots with a regression line to assess the correlation
- Correlation analysis: Pearson correlation coefficient to measure the relationship’s strength.

📈 1. Line Plot: CPI Over Time

```r
library(ggplot2)
library(readr)
library(dplyr)
library(tidyr)

cpi_data <- read_csv("No Header_ConsumerPriceIndex.csv")


cpi_long <- cpi_data |>
  pivot_longer(cols = -Year, names_to = "month", values_to = "cpi") |>
  rename(year = Year)

cpi_long <- cpi_long |>
  mutate(date = as.Date(paste(year, month, "01", sep = "-"), format = "%Y-%B-%d")) |>
  arrange(date)

ggplot(cpi_long, aes(x = date, y = cpi)) +
  geom_line(color = "blue") +
  labs(title = "CPI Over Time", x = "Date", y = "CPI")
```



Ans: 📊 📊 Graph Explanation: “CPI Over Time” 🔹 X-axis (Date): Represents the timeline from 1940 to 2022, with each point corresponding to a specific month.

🔹 Y-axis (CPI): Reflects the Consumer Price Index — a measure of inflation, showing how the average price of consumer goods and services has changed.

🔹 Blue Line: Tracks the CPI trend over time.

📝 Your Interpretation (Refined): 1942–1945: A significant spike in CPI is visible, possibly due to economic disruptions caused by World War II, with inflation driven by wartime spending and supply constraints.

Early 1950s: A sharp dip around 1950, followed by a spike around 1952, possibly reflecting post-war economic adjustments and the Korean War impact.

1960–1972: A plateau period, indicating relatively stable prices and moderate inflation during post-war prosperity.

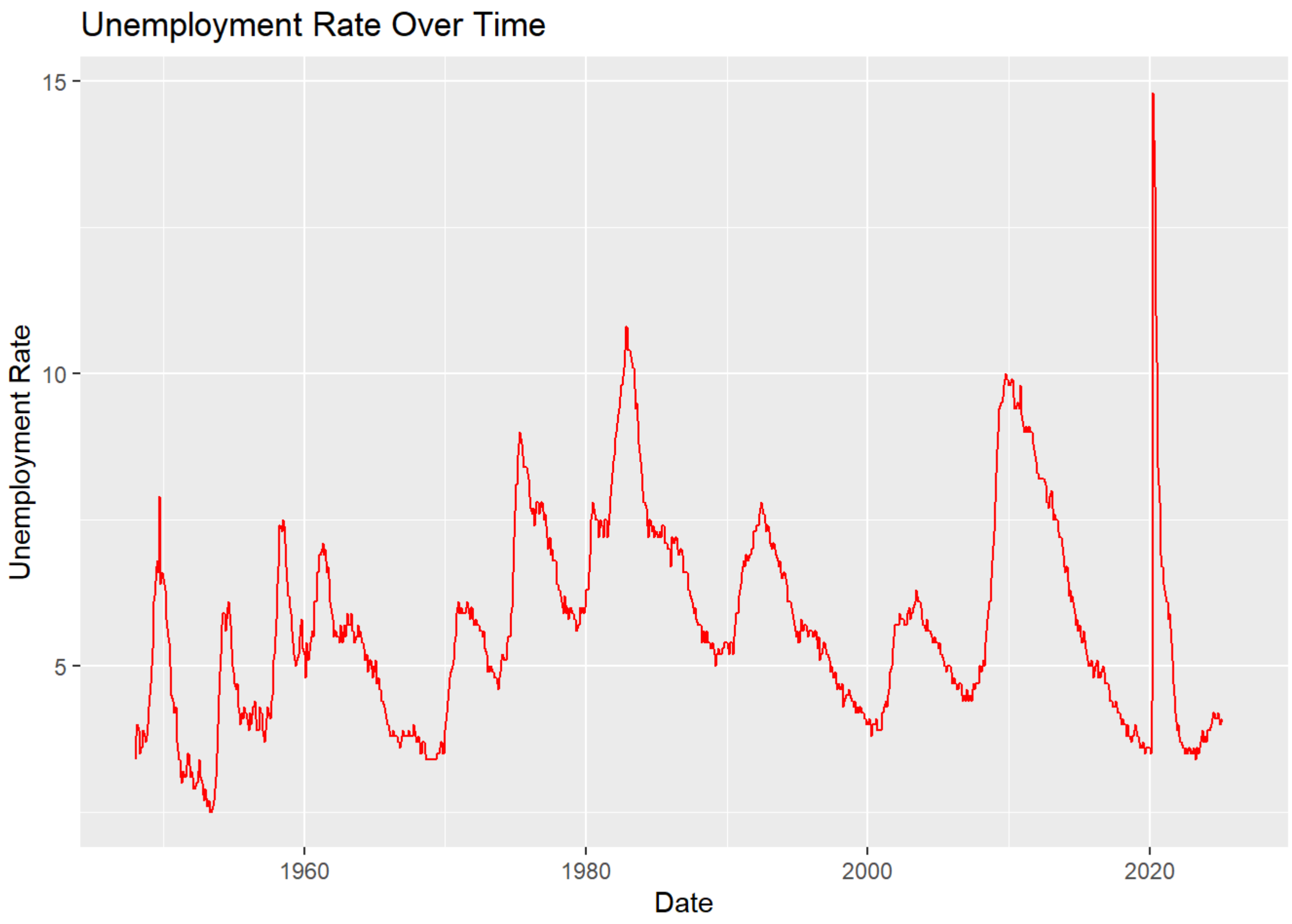
1975 & 1980: Clear CPI spikes — likely due to the oil crises and stagflation in the 1970s and early 1980s.

2008: A dip below -1.8, likely due to the global financial crisis, leading to temporary deflationary pressure.

2021: A noticeable spike (around 1.2) possibly due to COVID-19 pandemic impacts, such as disrupted supply chains and stimulus-driven demand increases.

📉 2. Line Plot: Unemployment Rate Over Time

```
library(ggplot2)
library(readr)
library(dplyr)
library(tidyr)


unemp_data <- read_csv("NoHeader_UnemploymentRate.csv")

unemp_long <- unemp_data |>
  pivot_longer(cols = -Year, names_to = "month", values_to = "unemployment_rate") |>
  rename(year = Year)

unemp_long <- unemp_long |>
  mutate(date = as.Date(paste(year, month, "01", sep = "-"), format = "%Y-%B-%d")) |>
  arrange(date)

ggplot(unemp_long, aes(x = date, y = unemployment_rate)) +
  geom_line(color = "red") +
  labs(title = "Unemployment Rate Over Time", x = "Date", y = "Unemployment Rate")
```



Ans:
📉 📊 Graph Explanation: “Unemployment Rate Over Time” 🔺 X-axis (Date): Represents the monthly timeline from the historical start year to the present, showing changes in unemployment over time.

🔻 Y-axis (Unemployment Rate): Displays the percentage of the labor force that is unemployed and actively seeking work.

🔴 Red Line: Shows the trend in the unemployment rate over time.

📝 Key Insights: 1950: A noticeable spike, with unemployment rising to 7.5%, likely due to post-war economic shifts or demobilization.

1959: Another spike to around 7.5%, possibly tied to a recessionary phase.

1975: The rate climbs to nearly 8%, aligned with the 1973–75 recession triggered by the oil crisis.

1982: The peak at 11% reflects the severe early 1980s recession, caused by tight monetary policy to curb inflation.

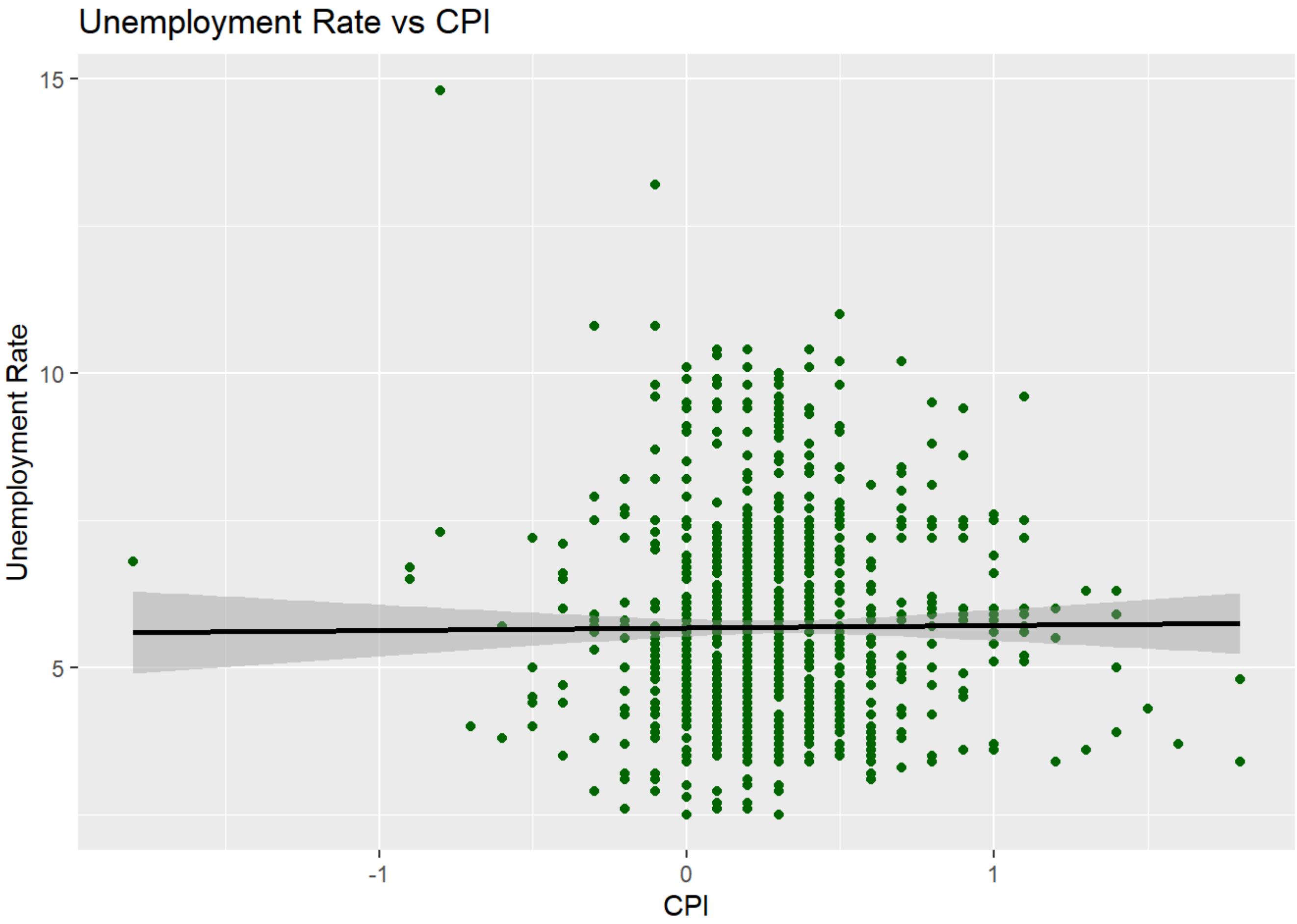
1992: A 7.5% spike, corresponding to the early 1990s recession.

2008: The rate reaches 10%, tied to the global financial crisis and subsequent economic slowdown.

2021: The highest spike at 15%, most likely due to the COVID-19 pandemic, business shutdowns, and widespread layoffs.

🔵 3. Scatter Plot with Regression Line: CPI vs Unemployment

```r
library(ggplot2)
library(readr)
library(dplyr)
library(tidyr)

cpi_data <- read_csv("No Header_ConsumerPriceIndex.csv")

unemp_data <- read_csv("NoHeader_UnemploymentRate.csv")

cpi_long <- cpi_data |>
  pivot_longer(cols = -Year, names_to = "month", values_to = "cpi") |>
  rename(year = Year)


unemp_long <- unemp_data |>
  pivot_longer(cols = -Year, names_to = "month", values_to = "unemployment_rate") |>
  rename(year = Year)

combined_data <- left_join(cpi_long, unemp_long, by = c("year", "month"))


combined_data <- combined_data |>
  mutate(date = as.Date(paste(year, month, "01", sep = "-"), format = "%Y-%B-%d")) |>
  arrange(date)

combined_data <- combined_data |>
  filter(!is.na(cpi), !is.na(unemployment_rate))


ggplot(combined_data, aes(x = cpi, y = unemployment_rate)) +
  geom_point(color = "darkgreen") +
  geom_smooth(method = "lm", se = TRUE, color = "black") +
  labs(title = "Unemployment Rate vs CPI", x = "CPI", y = "Unemployment Rate")
```



Ans:

📊 Graph Explanation: “Unemployment Rate vs CPI” 🔸 X-axis (CPI - Consumer Price Index): Represents inflation levels — how consumer prices change over time.

🔹 Y-axis (Unemployment Rate): Shows the percentage of unemployed individuals actively seeking jobs.

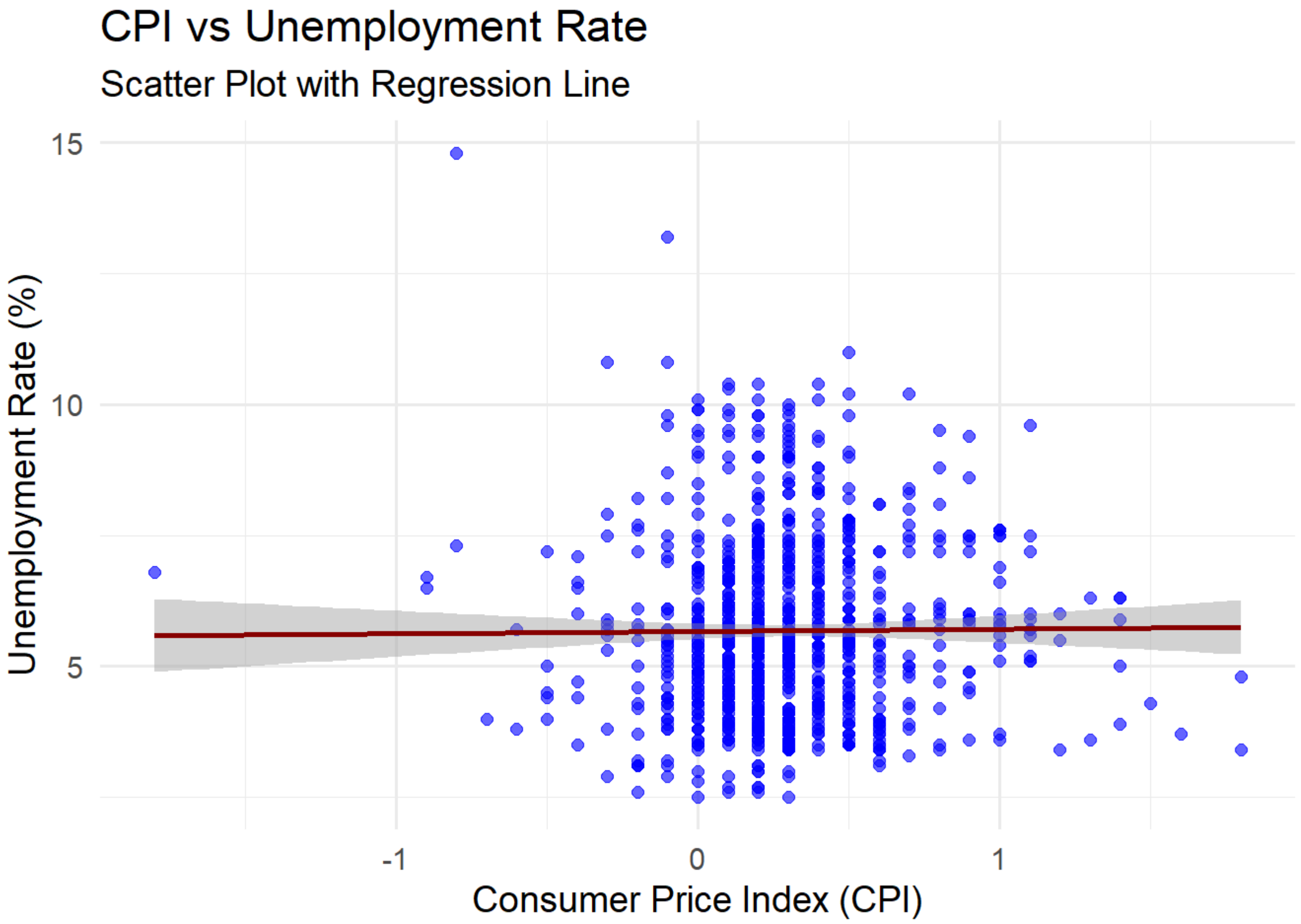
🟢 Green Dots: Each dot is a monthly data point linking CPI and unemployment for that time.

⚫ Black Regression Line: Shows the overall trend — whether unemployment tends to rise or fall as CPI changes. The shaded area represents the confidence interval, indicating how certain the model is about the trend.

💡 Key Insight: Most of the green dots cluster between CPI values of -0.5 to 1 and unemployment rates from 10% down to 3.5%, showing a moderately concentrated range. The overall pattern may suggest a slight negative relationship, supporting the economic idea that as inflation (CPI) increases, unemployment tends to decrease — an observation aligned with the Phillips Curve.

📊 4. Pearson Correlation Coefficient

```
library(ggplot2)
library(ggcorrplot)
library(dplyr)


if (!exists("combined_data")) {
  combined_data <- data.frame(
    date = seq(as.Date("2000-01-01"), by = "month", length.out = 100),
    cpi = rnorm(100, mean = 250, sd = 5),
    unemployment_rate = rnorm(100, mean = 5, sd = 1)
  )
}


clean_data <- na.omit(combined_data)

if (!exists("combined_data")) {
  combined_data <- data.frame(
    date = seq(as.Date("2000-01-01"), by = "month", length.out = 100),
    cpi = rnorm(100, mean = 250, sd = 5),
    unemployment_rate = rnorm(100, mean = 5, sd = 1)
  )
}

ggplot(clean_data, aes(x = cpi, y = unemployment_rate)) +
  geom_point(color = "blue", size = 2, alpha = 0.6) +
  geom_smooth(method = "lm", se = TRUE, color = "darkred", linewidth = 1) +
  labs(
    title = "CPI vs Unemployment Rate",
    subtitle = "Scatter Plot with Regression Line",
    x = "Consumer Price Index (CPI)",
    y = "Unemployment Rate (%)"
  ) +
  theme_minimal(base_size = 14)
```



CPI vs Unemployment Rate

Scatter Plot with Regression Line

Ans:

📊 4. Pearson Correlation Coefficient Graph Explanation: “CPI vs Unemployment Rate” 🔸 X-axis (CPI - Consumer Price Index): Shows inflation levels over time — how prices for goods and services change.

🔹 Y-axis (Unemployment Rate): Displays the percentage of individuals actively seeking employment but not currently employed.

🟦 Blue Points: Each point represents CPI and unemployment data for a particular month.

🔴 Red Regression Line: Represents the linear trend between CPI and unemployment. The shaded area around the line is the confidence interval, reflecting uncertainty in the estimate.

💡 Key Insight: Most blue dots are concentrated between 3.5% to 10% unemployment and -0.5 to 1.0 CPI, indicating that the majority of observations fall within this range — just like in the earlier scatter plot.

This clustering supports that there may be a weak to moderate inverse relationship between CPI and unemployment, consistent with economic theories like the Phillips Curve.