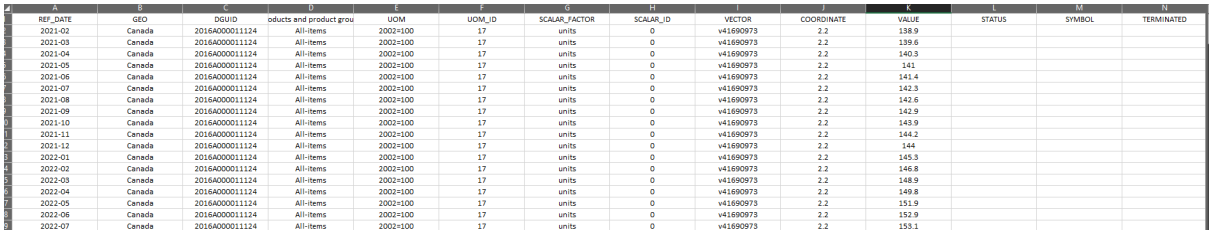
Big Data 2
Group 5 – Case Study 2

Q1: Cleaning the dataset:

Code:

```
library(dplyr)
library(readr)
library(lubridate)

# Step 1: Load the dataset
df <- read_csv("C:/Users/yukti/Downloads/48Months.csv")

# Step 2: Select relevant columns
df_2 <- df %>%
  select(REF_DATE, GEO, `Products and product groups`, VALUE) %>%
  rename(Date = REF_DATE, Location = GEO, Category = `Products and product groups`, CPI_Value =
VALUE)

# Check structure of the data
str(df_2)

# Get summary statistics before data clean up
summary(df_2)

#removing null values
df_2 <- na.omit(df_2)

#checking for duplicates
df_2 <- df_2 %>% distinct()

#filtering Date column

#df_2$Date <- as.Date(df_2$Date, format = "%Y-%m")

#df_2$Date <- ym(df_2$Date)


#summary after data clean up
summary(df_2)

# View first few rows
head(df_2)
```

Sample of Table data after tidying:



A	B	C	D	E	F	G	H	I	J	K	L	M	N
REF_DATE	GEO	DGUID	oducts and product grou	UOM	UOM_ID	SCALAR_FACTOR	SCALAR_ID	VECTOR	COORDINATE	VALUE	STATUS	SYMBOL	TERMINATED
2021-02	Canada	2016A000011124	All-items	2002=100	17	units	0	v41690973	2.2	138.9			
2021-03	Canada	2016A000011124	All-items	2002=100	17	units	0	v41690973	2.2	139.6			
2021-04	Canada	2016A000011124	All-items	2002=100	17	units	0	v41690973	2.2	140.3			
2021-05	Canada	2016A000011124	All-items	2002=100	17	units	0	v41690973	2.2	141			
2021-06	Canada	2016A000011124	All-items	2002=100	17	units	0	v41690973	2.2	141.4			
2021-07	Canada	2016A000011124	All-items	2002=100	17	units	0	v41690973	2.2	142.3			
2021-08	Canada	2016A000011124	All-items	2002=100	17	units	0	v41690973	2.2	142.6			
2021-09	Canada	2016A000011124	All-items	2002=100	17	units	0	v41690973	2.2	142.9			
2021-10	Canada	2016A000011124	All-items	2002=100	17	units	0	v41690973	2.2	143.9			
2021-11	Canada	2016A000011124	All-items	2002=100	17	units	0	v41690973	2.2	144.2			
2021-12	Canada	2016A000011124	All-items	2002=100	17	units	0	v41690973	2.2	144			
2022-01	Canada	2016A000011124	All-items	2002=100	17	units	0	v41690973	2.2	145.3			
2022-02	Canada	2016A000011124	All-items	2002=100	17	units	0	v41690973	2.2	146.8			
2022-03	Canada	2016A000011124	All-items	2002=100	17	units	0	v41690973	2.2	148.9			
2022-04	Canada	2016A000011124	All-items	2002=100	17	units	0	v41690973	2.2	149.8			
2022-05	Canada	2016A000011124	All-items	2002=100	17	units	0	v41690973	2.2	151.9			
2022-06	Canada	2016A000011124	All-items	2002=100	17	units	0	v41690973	2.2	152.9			
2022-07	Canada	2016A000011124	All-items	2002=100	17	units	0	v41690973	2.2	153.1			

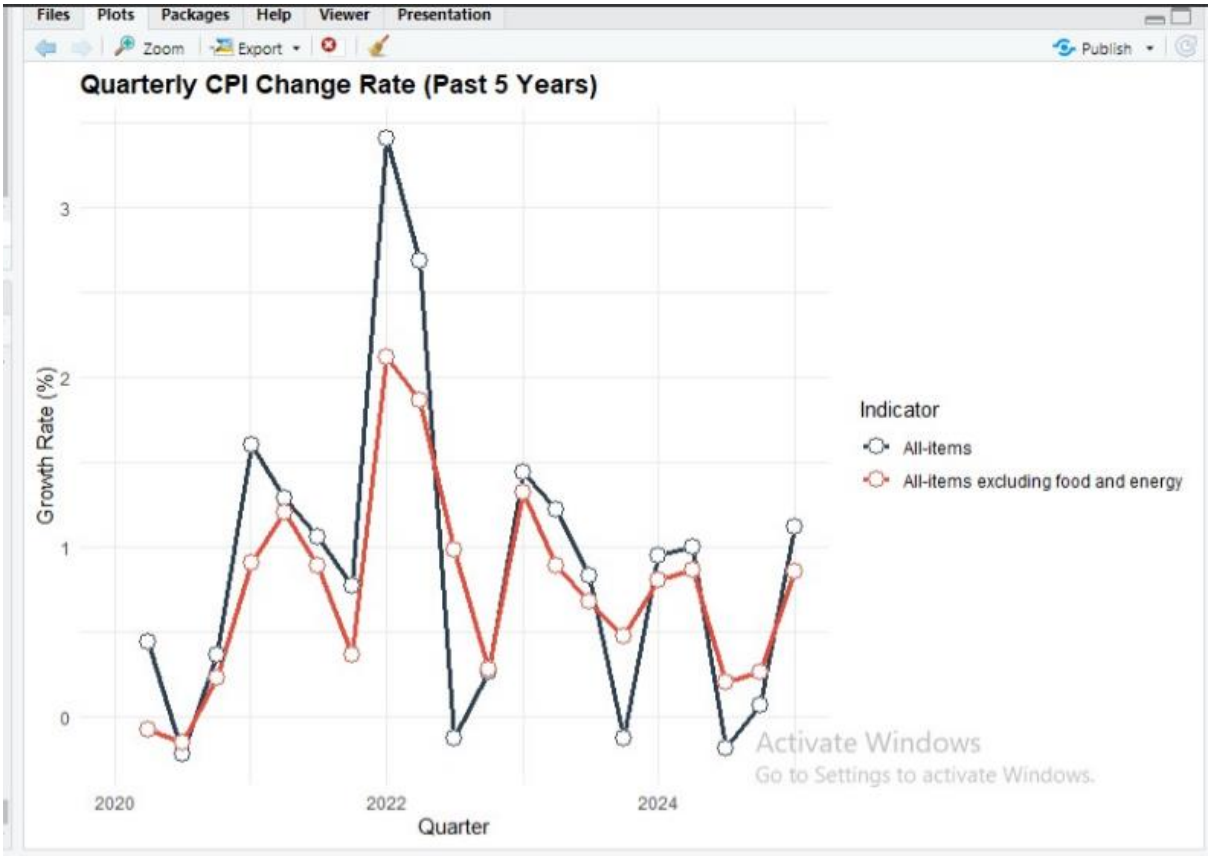
Q2: Quarterly Changes:

Code:

```
# Load libraries
install.packages("dplyr")
library(dplyr)
install.packages("lubridate")
library(lubridate)
install.packages("janitor")
library(janitor)
library(ggplot2)
# Read data
cpi_data <- read.csv("C:\\Users\\Administrator\\OneDrive - Humber College\\2025 winter\\Big Data 2 - BIA-5303-0LA\\1810000401_databaseLoadingData.csv", stringsAsFactors = FALSE) %>%
janitor::clean_names() %>%  # Clean column names
  mutate(ref_date = as.Date(paste(ref_date, "01", sep = "-"), format = "%Y-%m-%d")) %>%  #  Convert to date
  select(ref_date, product = products_and_product_groups, value)  # Keep the required fields

# Filter target product groups
q2_data <- cpi_data %>%
  filter(product %in% c("All-items", "All-items excluding food and energy")) %>%
  mutate(quarter = floor_date(ref_date, "quarter")) %>%  # Mark the quarter
  group_by(quarter, product) %>%
  summarise(value = last(value), .groups = "drop") %>%  # Get the end-of-quarter value
  arrange(product, quarter) %>%
  group_by(product) %>%
  mutate(quarter_change = (value / lag(value) - 1) * 100)  # Calculate the quarterly growth rate (%)

# Plot
ggplot(q2_data, aes(x = quarter, y = quarter_change, color = product)) +
  geom_line(size = 1.2) +
  geom_point(size = 4, shape = 21, fill = "white") +
  labs(
    title = "Quarterly CPI Change Rate (Past 5 Years)",
    x = "Quarter", y = "Growth Rate (%)", color = "Indicator"
  ) +
  scale_color_manual(values = c("#2c3e50", "#e74c3c")) +
```

```
  theme_minimal() +
  theme(plot.title = element_text(face = "bold"))

ggsave("q2_cpi_quarterly_change.png", dpi = 300, width = 10, height = 6)
```

Graph:



Conclusion:

The chart illustrates the quarterly Consumer Price Index (CPI) growth rates over the past five years, comparing the overall "All-items" CPI with the "All-items excluding food and energy" measure. The data reveals that CPI growth has been volatile, with notable spikes and dips, particularly between 2021 and 2022. This period saw the highest growth rate, surpassing 3% for the "All-items" category, reflecting the global inflationary pressures during the post-pandemic recovery phase. Excluding food and energy, the CPI still showed a significant but slightly more stable growth pattern, underscoring the high volatility of food and energy prices.

From 2023 onwards, both indicators demonstrate a trend of stabilization, with fewer extreme fluctuations and a narrower gap between the two lines. This indicates that inflation pressures may have eased, and core inflation (excluding volatile items) has been better controlled. The smaller variance also reflects improved economic conditions and potentially effective monetary policies.

Overall, while the CPI growth rate peaked dramatically during the global economic rebound, recent quarters suggest a return to more moderate and stable inflation levels. The convergence of the two

indicators toward the end of the period indicates a healthier economic balance and reduced impact from volatile commodity prices.

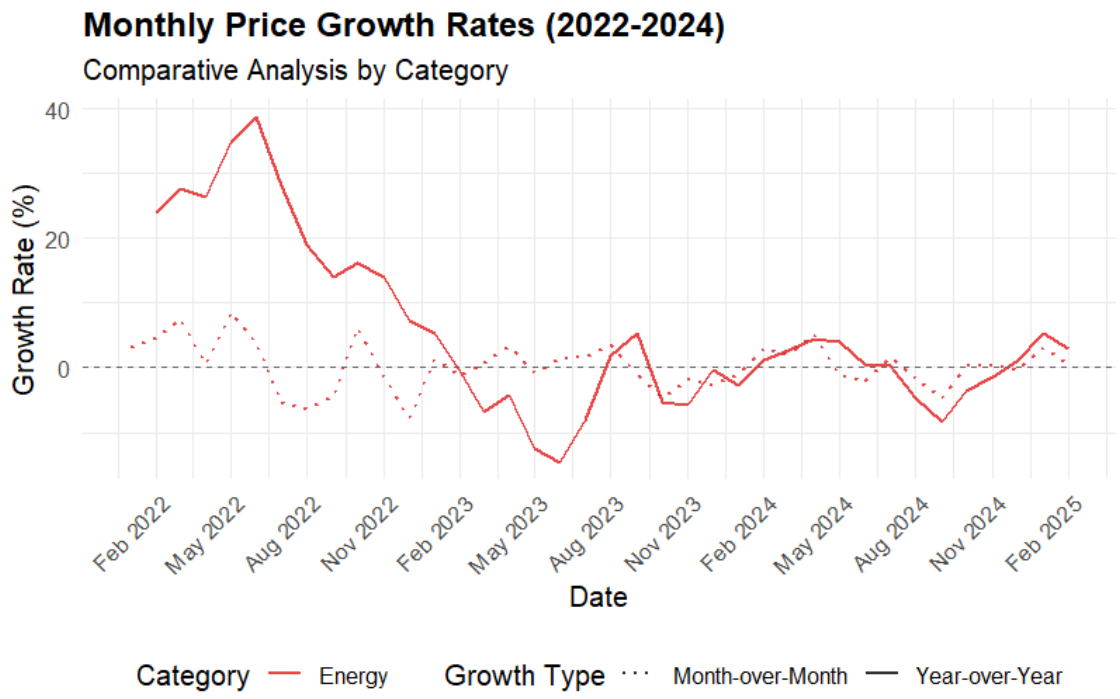
#Q3: Monthly price growth:

Code:

```
library(tidyverse)
library(lubridate)

# 1. Import and Prepare Your Data
# Replace "your_data.csv" with your actual file path
cpi_data <- read_csv("C:/Users/dhrum/Downloads/CS2.csv") %>%
  # Select and rename columns to match expected format
  select(
    REF_DATE = `REF_DATE`,            # Date column
    Category = `Products and product groups`,  # Category column
    VALUE = `VALUE`                   # Value column
  ) %>%
  # Filter for required categories (adjust names as needed)
  filter(Category %in% c("Food", "Shelter", "Energy",
                         "Food purchased from stores",
                         "Shelter - owned accommodation",
                         "Electricity")) %>%
  # Convert date and ensure proper sorting
  mutate(REF_DATE = ymd(paste0(REF_DATE, "-01"))) %>%
  arrange(Category, REF_DATE)

# 2. Calculate Growth Rates (same as before)
growth_rates <- cpi_data %>%
  group_by(Category) %>%
  mutate(
    mom_growth = (VALUE - lag(VALUE)) / lag(VALUE) * 100,
    yoy_growth = (VALUE - lag(VALUE, 12)) / lag(VALUE, 12) * 100
  ) %>%
  filter(REF_DATE >= as.Date("2022-01-01")) %>%  # Last 36 months
  pivot_longer(cols = c(mom_growth, yoy_growth),
               names_to = "Growth_Type",
               values_to = "Growth_Rate") %>%
  drop_na()

# 3. Create Combined Plot (updated with better visuals)
ggplot(growth_rates, aes(x = REF_DATE, y = Growth_Rate,
                         color = Category, linetype = Growth_Type)) +
  geom_line(linewidth = 0.8, alpha = 0.8) +
  geom_hline(yintercept = 0, color = "gray40", linetype = "dashed") +
  scale_color_brewer(palette = "Set1") +  # Colorblind-friendly palette
```

```
 scale_linetype_manual(values = c("mom_growth" = "dotted",
                                  "yoy_growth" = "solid"),
                       labels = c("Month-over-Month", "Year-over-Year")) +
 labs(title = "Monthly Price Growth Rates (2022-2024)",
      subtitle = "Comparative Analysis by Category",
      x = "Date",
      y = "Growth Rate (%)",
      color = "Category",
      linetype = "Growth Type") +
 theme_minimal(base_size = 12) +
 theme(
   legend.position = "bottom",
   legend.box = "horizontal",
   axis.text.x = element_text(angle = 45, hjust = 1),
   plot.title = element_text(face = "bold")
 ) +
 scale_x_date(date_breaks = "3 months", date_labels = "%b %Y")

# 4. Save High-Quality Output
ggsave("Q3_Monthly_Growth_Actual.png", width = 10, height = 6, dpi = 300)
```

Graph:



Conclusion:

- **Energy** exhibited extreme volatility, with YoY growth peaking at ~35% in mid-2022 before sharply declining to negative rates by late 2023. This reflects geopolitical impacts and subsequent market adjustments.

- **Shelter** maintained persistent inflation, with YoY growth consistently above 5% throughout the period, indicating sustained housing market pressures.
- **Food** prices grew moderately but accelerated in early 2023 (MoM peaks exceeding 2%), likely due to supply chain disruptions.

All categories showed stabilization by 2024, with Energy growth rates converging with other sectors. The MoM fluctuations (dashed lines) were most pronounced for Energy, while Shelter's steady growth suggests structural inflationary factors.