Software Used: Rstudio & MS Word (for documentation)

All R codes:

```
setwd("D:/Playground_all_/Rprogramming/MospiCPI")
df=read.csv("CPIndex_Jan13-To-Jan25_F&B.csv",header = T,skip = 1)
View(df)

#Data Cleaning
colSums(is.na(df))

df=df[,-c(4,5,11)]

summary(df)
str(df)

df_ap=df[df$State == "Arunachal Pradesh", ];View(df_ap)
#It is clear that AP has only NA values in urban , i.e we can fill it as 0 value.
#and the combined col only consists of rural cpi for AP so we impute combined col using
# rural cpi.

# Replace NA values in Urban column with 0 for Arunachal Pradesh
df$Urban[df$State == "Arunachal Pradesh" & is.na(df$Urban)] <- 0

# Replace Combined column values with Rural values for Arunachal Pradesh
df$Combined[df$State == "Arunachal Pradesh"] <- df$Rural[df$State == "Arunachal Pradesh"]
```

```
### Visualization
library(ggplot2)
library(dplyr)

attach(df)

# Extract ALL India CPI data
all_india_cpi <- df %>%
  filter(State == "ALL India") %>%
  select(Year, Month, All_India_Combined = Combined)

# Convert Month-Year into a proper Date format
all_india_cpi$Date <- as.Date(paste0(all_india_cpi$Year, "-", all_india_cpi$Month, "-01"), format = "%Y-%B-%d")

# Line plot with correct ordering
ggplot(all_india_cpi, aes(x = Date, y = All_India_Combined)) +
  geom_line(color = "blue", linewidth = 1) +
  labs(title = "Monthly Combined CPI for All India",
       x = "Year-Month", y = "All India Combined CPI") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))


df_compare <- df %>%
  filter(State != "ALL India") %>%  # Exclude "ALL India" from main data
```

```
  left_join(all_india_cpi, by = c("Year", "Month"))



df_compare <- df_compare %>%
  mutate(
    CPI_Diff = Combined - All_India_Combined,
    CPI_Percent_Diff = ((Combined - All_India_Combined) / All_India_Combined) * 100
  )

# Line plot comparing Combined CPI with All India CPI for a few sample states
ggplot(df_compare %>% filter(State %in% c("Bihar", "Delhi", "Gujarat", "West Bengal")),
       aes(x = Date)) +
  geom_line(aes(y = Combined, color = State)) +
  geom_line(aes(y = All_India_Combined), color = "black", linetype = "dashed") +
  labs(title = "State vs All India Combined CPI",
       x = "Date", y = "Combined CPI") +
  theme_minimal()

### finding top 5 performing states over combined cpi

top_5_states <- df_compare %>%
  group_by(State) %>%
  summarize(Avg_Difference = mean(CPI_Diff, na.rm = TRUE)) %>%
  arrange(desc(Avg_Difference)) %>%
  slice_head(n = 5)
```

```
# Find top 5 overperforming states for each year
top_5_states_per_year <- df_compare %>%
  group_by(Year, State) %>%
  summarize(Avg_Difference = mean(CPI_Diff, na.rm = TRUE), .groups = "drop") %>%
  arrange(Year, desc(Avg_Difference)) %>%
  group_by(Year) %>%
  slice_head(n = 5)  # Pick top 5 for each year



ggplot(top_5_states_per_year, aes(x = factor(Year, levels = sort(unique(Year))), y =
Avg_Difference, fill = factor(State))) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(title = "Top 5 States Overperforming All India CPI (Year-wise)",
       x = "Year", y = "Avg CPI Difference",
       fill = "State") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))



##############################################################################
#########

#Rural vs Urban CPI Comparison
ruralcpi=df %>%
  mutate(CPI_Diff = Rural - Urban) %>%
  group_by(State) %>%
```

```
  summarize(Avg_CPI_Diff = mean(CPI_Diff, na.rm = TRUE)) %>%
  arrange(desc(Avg_CPI_Diff))

# If Avg_CPI_Diff > 0, Rural CPI is generally higher than Urban CPI.
#
# If Avg_CPI_Diff < 0, Urban CPI is higher.
#
# Helps identify if rural areas face higher inflation than urban areas.

#Top 5 States Where Rural CPI > Urban CPI

top_rural_cpi_states <- df %>%
  filter(Rural > Urban) %>%
  group_by(State) %>%
  summarize(Avg_Rural_CPI = mean(Rural, na.rm = TRUE),
            Avg_Urban_CPI = mean(Urban, na.rm = TRUE),
            Diff = mean(Rural - Urban, na.rm = TRUE)) %>%
  arrange(desc(Diff)) %>%
  slice_head(n = 5)

#Shows which states have higher Rural CPI, indicating possible supply chain
#issues, lack of subsidies, or increased transportation costs in rural areas.

#Trend of Rural vs Urban CPI Over Time

library(lubridate)
```

```
df %>%
  group_by(Year, Month) %>%
  summarize(Avg_Rural_CPI = mean(Rural, na.rm = TRUE),
            Avg_Urban_CPI = mean(Urban, na.rm = TRUE),
            .groups = "drop") %>%
  mutate(Date = as.Date(paste0(Year, "-", Month, "-01"), format = "%Y-%B-%d")) %>%
  ggplot(aes(x = Date)) +
  geom_line(aes(y = Avg_Rural_CPI, color = "Rural CPI"), linewidth = 1) +
  geom_line(aes(y = Avg_Urban_CPI, color = "Urban CPI"), linewidth = 1) +
  labs(title = "Rural vs Urban CPI Trend Over Time",
       x = "Year-Month", y = "CPI Value") +
  scale_x_date(date_labels = "%b-%Y", date_breaks = "6 months") +  # Adjust x-axis format
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, vjust = 1, hjust = 1))

# If Urban CPI fluctuates more, cities experience more inflation volatility than villages.
#
# If both increase similarly, inflation affects both rural and urban regions at a similar rate.

#Identifying Outlier States
#Goal: Find states where Rural CPI or Urban CPI is unusually high/low.
library(gridExtra)

rural=ggplot(df, aes(x = State, y = Rural)) +
  geom_boxplot(fill = "blue", alpha = 0.5) +
```

```
  labs(title = "Distribution of Rural CPI Across States",
       x = "State", y = "Rural CPI") +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust = 1))

urban=ggplot(df, aes(x = State, y = Urban)) +
  geom_boxplot(fill = "red", alpha = 0.5) +
  labs(title = "Distribution of Urban CPI Across States",
       x = "State", y = "Urban CPI") +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust = 1))

grid.arrange(rural,urban)

# Boxplots help spot outliers (e.g., a state where inflation is unusually high).
#
# Helps investigate whether certain states have inflation control issues.

#Correlation Between Rural & Urban CPI
#Goal: Check how strongly correlated Rural and Urban CPI are.
cor(df$Rural, df$Urban, use = "complete.obs")

# If correlation close to 1, both CPIs move together.
# If correlation < 0.5, urban and rural inflation behave differently.

## States with the Largest CPI Volatility
#Goal: Find states where CPI fluctuates the most.
```

```
df %>%
  group_by(State) %>%
  summarize(CPI_SD = sd(Combined, na.rm = TRUE)) %>%
  ggplot(aes(x = reorder(State, CPI_SD), y = CPI_SD, fill = CPI_SD)) +
  geom_bar(stat = "identity") +
  coord_flip() +
  labs(title = "States with Highest CPI Volatility",
       x = "State", y = "Standard Deviation of CPI") +
  theme_minimal()

#High variability states means Price instability.
#Consistently low-variance states means Stable inflation trends.

##Rural vs Urban CPI Difference by State
#Goal: See which states have the biggest gap between Rural and Urban CPI.

df %>%
  group_by(State) %>%
  summarize(Rural_CPI = mean(Rural, na.rm = TRUE),
            Urban_CPI = mean(Urban, na.rm = TRUE),
            CPI_Diff = mean(Urban - Rural, na.rm = TRUE)) %>%
  ggplot(aes(x = reorder(State, CPI_Diff), y = CPI_Diff, fill = CPI_Diff > 0)) +
  geom_bar(stat = "identity") +
  coord_flip() +
  labs(title = "Rural vs Urban CPI Difference by State",
       x = "State", y = "Urban CPI - Rural CPI") +
```

```
  scale_fill_manual(values = c("red", "blue"), labels = c("Rural Higher", "Urban Higher")) +
  theme_minimal()

# Urban CPI higher than Rural means Costlier urban lifestyle.
# Rural CPI higher than Urban means Potential reverse urbanization trends.

##CPI Heatmap for All States
#Goal: Show CPI changes over years in heatmap style.

df %>%
  group_by(State, Year) %>%
  summarize(Avg_CPI = mean(Combined, na.rm = TRUE)) %>%
  ggplot(aes(x = Year, y = State, fill = Avg_CPI)) +
  geom_tile() +
  scale_fill_gradient(low = "skyblue", high = "darkblue") +
  labs(title = "CPI Heatmap Across States and Years",
       x = "Year", y = "State") +
  theme_minimal()

##CPI Inflation Rate Over Time
#Goal: Show % change in CPI over time.

df_compare %>%
  group_by(Year) %>%
  summarize(Avg_CPI = mean(Combined, na.rm = TRUE)) %>%
  mutate(Inflation_Rate = (Avg_CPI - lag(Avg_CPI)) / lag(Avg_CPI) * 100) %>%
```

```
  ggplot(aes(x = as.factor(Year), y = Inflation_Rate)) +
  geom_line(group = 1, color = "red", linewidth = 1.5) +
  labs(title = "Year-over-Year CPI Inflation Rate",
       x = "Year", y = "Inflation Rate (%)") +
  theme_minimal()

################################################################################
########

#Some more questions to get deeper understanding of the problem
##1. What is the overall CPI inflation trend over the past 5 years?

df_trend <- df %>%
  filter(State == "ALL India", Year >= max(Year) - 4) %>%
  group_by(Year) %>%
  summarize(Avg_CPI = mean(Combined, na.rm = TRUE))

ggplot(df_trend, aes(x = Year, y = Avg_CPI)) +
  geom_line(color = "blue", linewidth = 1) +
  geom_point(size = 2, color = "red") +
  labs(title = "CPI Inflation Trend (Last 5 Years)", x = "Year", y = "Average CPI") +
  theme_minimal()

#2. How did inflation fluctuate during the COVID-19 pandemic?

df_covid <- df %>%
```

```
  filter(Year %in% c(2019, 2020, 2021)) %>%
  group_by(Year) %>%
  summarize(Avg_CPI = mean(Combined, na.rm = TRUE))

ggplot(df_covid, aes(x = Year, y = Avg_CPI)) +
  geom_bar(stat = "identity", fill = "orange") +
  labs(title = "CPI During COVID-19 (2019-2021)", x = "Year", y = "Average CPI") +
  theme_minimal()

##3. How does inflation fluctuate during festive seasons?

festive_months <- c("October", "November", "December")

df_festive <- df %>%
  filter(Month %in% festive_months) %>%
  group_by(Year, Month) %>%
  summarize(Avg_CPI = mean(Combined, na.rm = TRUE))

ggplot(df_festive, aes(x = as.factor(Year), y = Avg_CPI, fill = Month)) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(title = "CPI During Festive Seasons", x = "Year", y = "CPI") +
  theme_minimal()


#4. How much has inflation reduced the purchasing power of ₹100?
```

```
df_pp=df_compare %>%
  group_by(Year) %>%
  summarize(Purchasing_Power = 100 / mean(Combined, na.rm = TRUE))

#5. How much has inflation increased food expenses?
df_inf_f=df %>%
  group_by(Year) %>%
  summarize(Avg_CPI = mean(Combined, na.rm = TRUE)) %>%
  mutate(Cost_Index = Avg_CPI / first(Avg_CPI) * 100)



################################################################################
#######
################### Analysis ###################

#1. Chi-Square Test: Inflation Trends (Categorical Analysis)

#null: Inflation trends (increasing or decreasing) are independent of state.
#alt: Inflation trends are dependent on state.

df$Inflation_Trend <- ifelse(df$Combined > lag(df$Combined), "Increase", "Decrease")
chisq.test(table(df$State, df$Inflation_Trend))

# Pearson's Chi-squared test
#
# data:  table(df$State, df$Inflation_Trend)
```

```
# X-squared = 1563.2, df = 36, p-value < 2.2e-16

## interpret: p-value < 0.05, rejecting null, Inflation trends depend on the state.

#2. Time Series Analysis
## A. CPI Trend Forecasting (ARIMA)
library(forecast)
cpi_ts <- ts(df[df$State == "ALL India", "Combined"], start = c(2013,1), frequency = 12)
fit <- auto.arima(cpi_ts)
forecasted_cpi <- forecast(fit, h = 12)
plot(forecasted_cpi)

# 3. CPI Stability Analysis
# A. Standard Deviation of CPI Over Years
# Objective: Identify which states have the most volatile CPI.

df_volatility <- df %>%
  group_by(State) %>%
  summarize(Std_Dev = sd(Combined, na.rm = TRUE)) %>%
  arrange(desc(Std_Dev))
#High Volatility States (Manipur, Lakshadweep, Andaman, Telangana, West Bengal)
#States with high standard deviation experience unstable inflation.

# B. CPI Stationarity Test (ADF Test)
#Test for Stationarity (Augmented Dickey-Fuller Test)
```

```
library(tseries)
adf.test(cpi_ts)

# Augmented Dickey-Fuller Test
#
# data:  cpi_ts
# Dickey-Fuller = -1.6814, Lag order = 5, p-value = 0.7091
# alternative hypothesis: stationary

##p-value > 0.05, CPI follows a trend and needs differencing.

# 4. Clustering Analysis (Identifying Similar States)
# A. K-Means Clustering on CPI Trends
# Objective: Group states with similar inflation trends.

df_cluster <- df_compare %>%
  group_by(State) %>%
  summarize(Mean_CPI = mean(Combined, na.rm = TRUE))

set.seed(123)
kmeans_result <- kmeans(df_cluster$Mean_CPI, centers = 3)
df_cluster$Cluster <- kmeans_result$cluster
ggplot(df_cluster, aes(x = State, y = Mean_CPI, color = as.factor(Cluster))) +
  geom_point(size = 4) + coord_flip() +
  labs(title = "Clustering States Based on CPI Trends")
```

cluster 1 (High inflation): Andaman & Nicobar, Kerala, Manipur, Puducherry

cluster 2 (Medium inflation): Andhra Pradesh, Arunachal Pradesh, Daman and Diu, Goa, Jammu and Kashmir,

Karnataka, Lakshadweep, Mizoram, Nagaland, Odisha, Sikkim, Tamil Nadu,

Telangana, Tripura, West Bengal.

cluster 3 (Low inflation): Assam, Bihar, Chandigarh, Chhattisgarh, Dadra and Nagar Haveli, Delhi,

Gujarat, Haryana, Himachal Pradesh, Jharkhand, Madhya Pradesh, Maharashtra,

Meghalaya, Punjab, Rajasthan, Uttar Pradesh, Uttarakhand.

This groups states into high, medium, and low inflation zones.