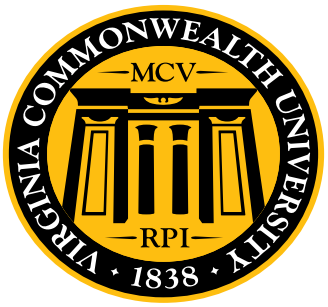


VIRGINIA COMMONWEALTH UNIVERSITY

Statistical analysis and modelling (SCMA 632)

A5.a : Visualisation – Perceptual Mapping for Business

JYOTHIS KANIYAMPARAMBIL THANKACHAN

V01110144

Date of Submission: 15-07-2024

CONTENTS

Introduction

Histogram and Barplot to indicate the consumption district-wise for Manipur

The state of Manipur is the main focus of this study. Data from the National Sample Survey Office (NSSO) are used to look into spending trends at the district level. With the help of a histogram, we want to show how total usage is spread out across different areas. In addition, it used a barplot to show a full picture of usage by area. The NSSO68 collection has a lot of information about spending in both rural and urban areas. Taking care of missing numbers, finding and getting rid of outliers, and standardising district and sector names are all parts of the research. By putting together, the purchasing data at the regional and local levels, we hope to give you important information about how people in Manipur buy things. These graphics will help policymakers and other interested parties understand how consumption trends vary from one area to the next. In this way, they will be able to make certain changes and support fair and equal growth across the state.

Objectives:

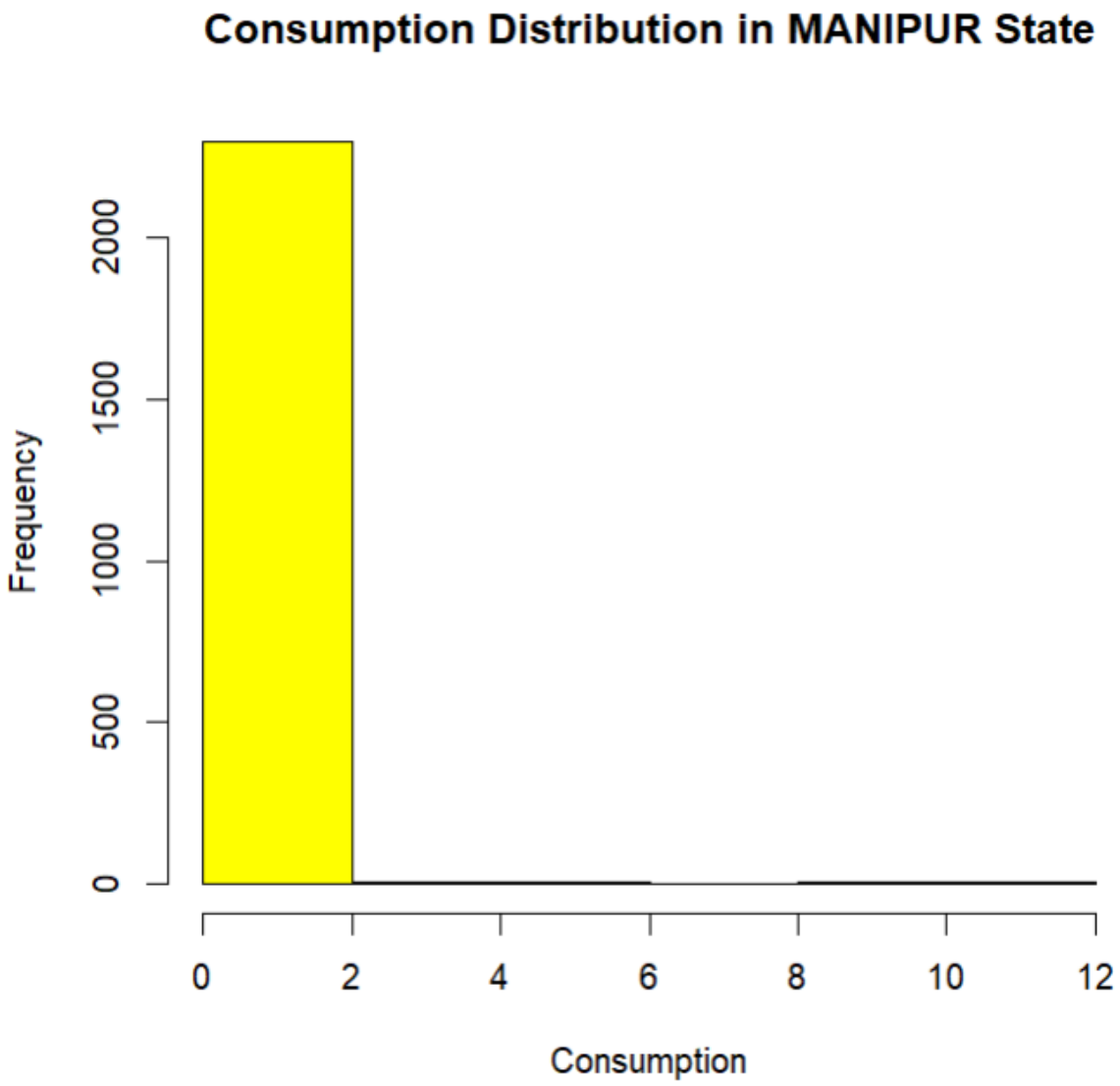
- Visualize the distribution of consumption.
- Conduct a detailed analysis of consumption by district.
- Identify patterns in consumption.

Business Significance: Use a graph and a bar plot to show how much each area in Manipur uses. The barplot and histogram jobs, which show how people in different parts of Manipur spend their money, have big business effects. They give important information about how to distribute resources, which lets people plan ahead for food supplies and build infrastructure that fits the needs of each district's residents. Businesses can use this information to tailor their marketing strategies and products to the tastes of customers in different areas. Understanding differences in consumption also helps improve supply chain management, making sure that goods are supplied efficiently to meet a range of demand levels. Policymakers can benefit from focusing their efforts on creating programmes that aim to improve public health and encourage fair usage. These visuals also help with competition analysis, which lets businesses find market gaps and chances for strategic entry and placing. Looking at trends of consumption helps figure out the effects on society and the economy and push for actions that promote better eating and living choices in Manipur.

Results and Interpretation using R

Histogram

```
# histogram to show the distribution of total consumption across different districts
hist(MANPRnew$total_consumption, breaks = 6, col = 'yellow', border = 'black',
     xlab = "Consumption", ylab = "Frequency", main = "Consumption Distribution in MANIPUR State")
```



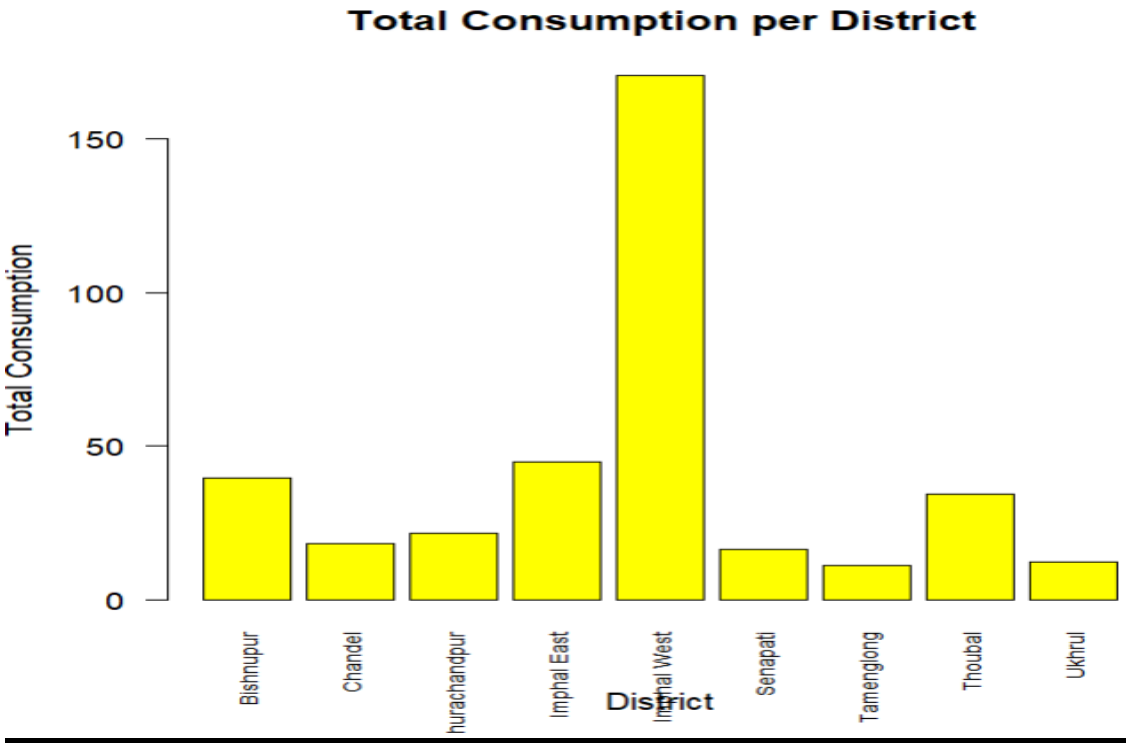
Interpretation:

The histogram illustrates the consumption distribution in Manipur State, revealing that the majority of consumption values fall between 0 and 2 units, with a frequency exceeding 2000. This indicates a significant concentration of low consumption levels within the state. The

graph also shows a rapid decline in frequency as consumption values increase, with very few instances of consumption levels beyond 2 units. This suggests that most households or entities in Manipur have minimal consumption, reflecting potential constraints in economic activity, resource availability, or purchasing power. The data underscores the need for policies aimed at boosting economic growth and resource access to improve overall consumption levels in the state.

Barplot

```
# barplot to visualize consumption per district with district names
??barplot
barplot(MANPR_consumption$total_consumption,
        names.arg = MANPR_consumption$District,
        las = 2, # Makes the district names vertical
        col = 'yellow',
        border = 'black',
        xlab = "District",
        ylab = "Total Consumption",
        main = "Total Consumption per District",
        cex.names = 0.7)
```



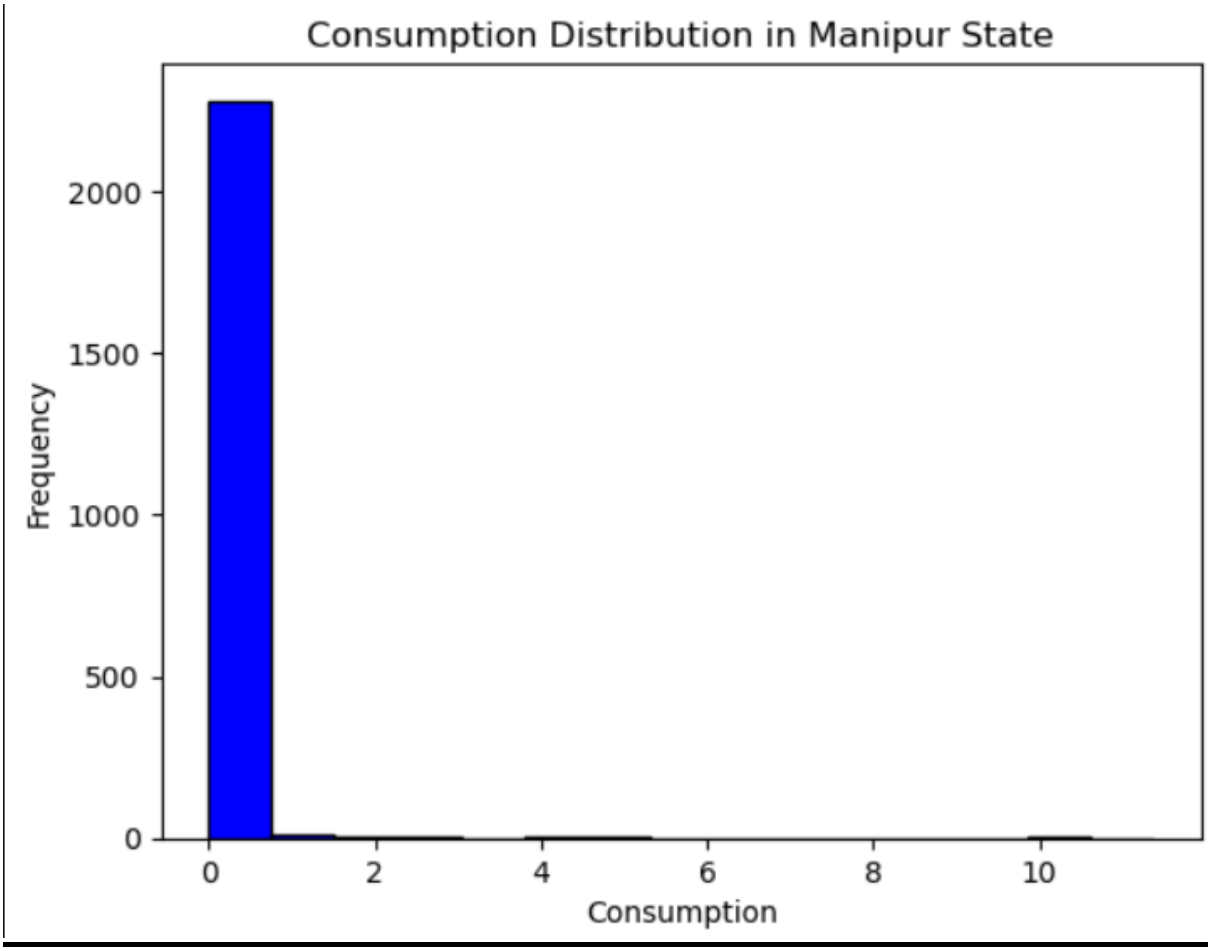
Interpretation:

The bar plot displays the total consumption per district. Among the districts, Imphal West stands out significantly with the highest total consumption, reaching over 150 units. Other districts such as Bishnupur, Imphal East, and Thoubal also show notable consumption levels but are considerably lower than Imphal West. The districts of Chandel, Churachandpur, Senapati, Tamenglong, and Ukhrul have relatively low consumption, with values much closer to each other and significantly below the highest values. This indicates a substantial disparity in total consumption among the districts, with Imphal West having an exceptionally high consumption compared to the rest.

Results and Interpretation using Python

Histogram

```
# Histogram to show the distribution of total consumption across different districts
plt.hist(MANPRnew['total_consumption'], bins= 15, color='blue', edgecolor='black')
plt.xlabel('Consumption')
plt.ylabel('Frequency')
plt.title('Consumption Distribution in Manipur State')
plt.show()
```



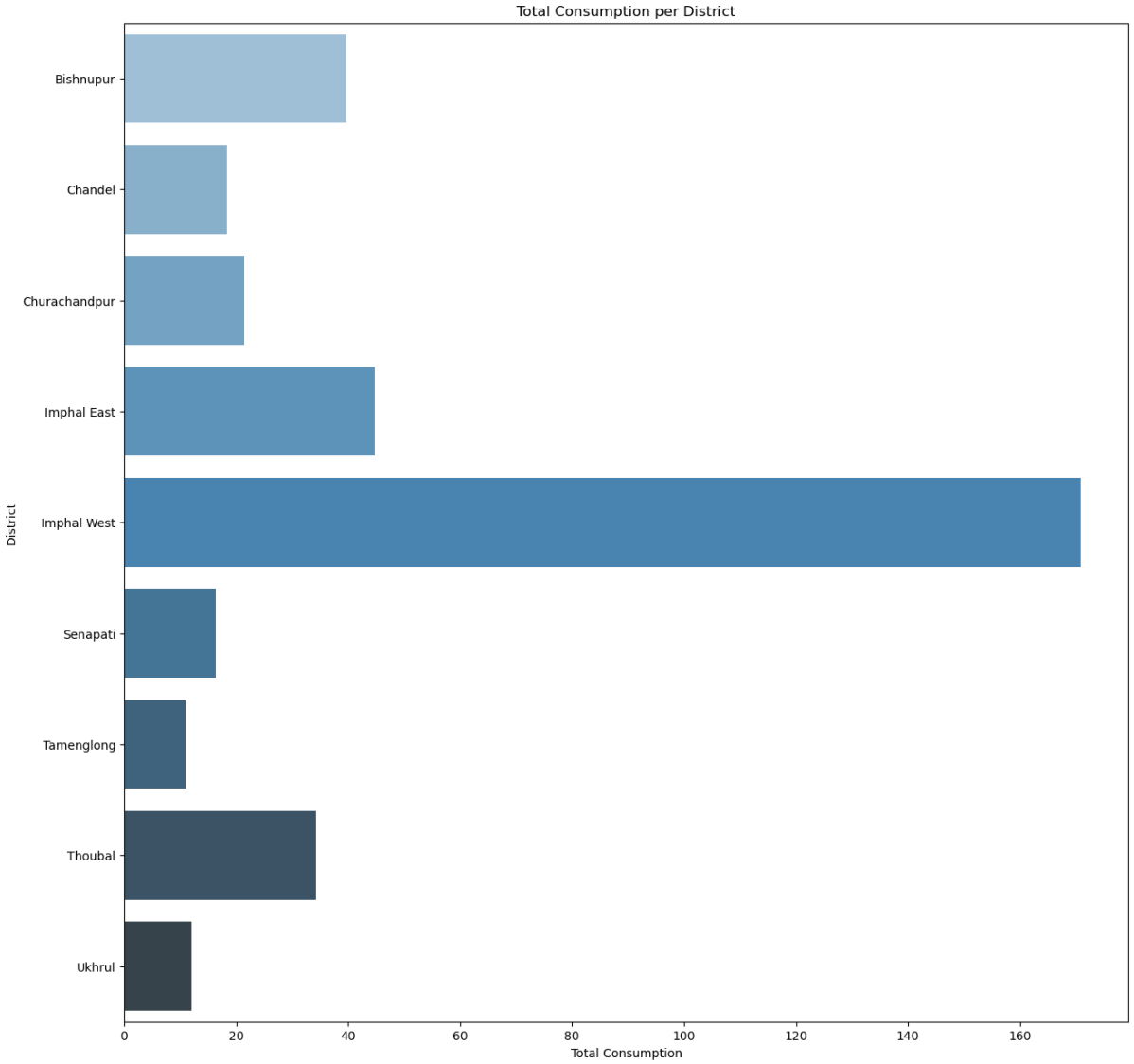
Interpretation:

The histogram displays the consumption distribution in Manipur State. It shows a highly skewed distribution with the majority of consumption values concentrated at the lower end of the scale. The frequency of consumption values is exceedingly high near zero, indicating that most observations have very low consumption levels. There are very few instances of higher

consumption, as evidenced by the rapid decline in frequency as consumption increases. This suggests that the overall consumption in Manipur State is predominantly low, with a small number of outliers having higher consumption values.

Barplot

```
# Barplot to visualize consumption per district with district names
plt.figure(figsize=(15, 15))
sns.barplot(x='total_consumption', y='District', data=MANPR_consumption, palette='Blues_d')
plt.xlabel('Total Consumption')
plt.ylabel('District')
plt.title('Total Consumption per District')
plt.show()
```



Interpretation:

The bar plot illustrates the total consumption per district in Manipur State. Imphal West leads significantly in total consumption, reaching around 160 units, which is far higher than any other district. Following Imphal West, Bishnupur, Imphal East, and Thoubal show moderate levels of consumption, each ranging between 40 to 60 units. Other districts like Chandel, Churachandpur, Senapati, Tamenglong, and Ukhrul exhibit much lower total consumption, all below 30 units. This distribution highlights a significant disparity, with Imphal West's consumption vastly exceeding that of the other districts, indicating a potentially uneven distribution of resources or varying levels of demand across the districts.

Recommendations

To reduce differences in consumption, it is suggested that low-consumption families be helped by making it easier for them to get to resources and facilities. Promoting energy economy and environmentally friendly habits can help cut down on overuse in places with a lot of it. These attempts will be guided by an analysis of the causes that lead to different amounts of consumption. We can make the state's consumption patterns more balanced and long-lasting by improving infrastructure and resource access in areas with low consumption and supporting efficiency in areas with high consumption.

R Codes

```
# Set the working directory and verify it
setwd('D:\\Assignments_SCMA632')
getwd()

# Function to install and load libraries
install_and_load <- function(package) {
  if (!require(package, character.only = TRUE)) {
    install.packages(package, dependencies = TRUE)
    library(package, character.only = TRUE)
  }
}

# Load required libraries
libraries <- c("dplyr", "readr", "readxl", "tidyr", "ggplot2", "BSDA","glue","sf")
lapply(libraries, install_and_load)

# Reading the file into R
data <- read.csv("NSSO68.csv")



# a)Plotting a histogram and a Barplot of the data to indicate the consumption district-wise for the West Bengal

# Filtering for MANPR
df <- data %>%
  filter(state_1 == "MANPR")

# Display dataset info
cat("Dataset Information:\n")
print(names(df))
print(head(df))
print(dim(df))
```

```
# Sub-setting the data
MANPRnew <- df %>%
  select(state_1, District, Region, Sector, State_Region, Meals_At_Home, ricepds_v, W
heatpds_q, chicken_q, pulsep_q, wheatos_q, No_of_Meals_per_day)

# Check for missing values in the subset
cat("Missing Values in Subset:\n")
print(colSums(is.na(MANPRnew)))

# Impute missing values with mean for specific columns
impute_with_mean <- function(column) {
  if (any(is.na(column))) {
    column[is.na(column)] <- mean(column, na.rm = TRUE)
  }
  return(column)
}
MANPRnew$Meals_At_Home <- impute_with_mean(MANPRnew$Meals_At_Home
)

# Check for missing values after imputation
cat("Missing Values After Imputation:\n")
print(colSums(is.na(MANPRnew)))

# Finding outliers and removing them
remove_outliers <- function(df, column_name) {
  Q1 <- quantile(df[[column_name]], 0.25)
  Q3 <- quantile(df[[column_name]], 0.75)
  IQR <- Q3 - Q1
  lower_threshold <- Q1 - (1.5 * IQR)
  upper_threshold <- Q3 + (1.5 * IQR)
  df <- subset(df, df[[column_name]] >= lower_threshold & df[[column_name]] <= up
per_threshold)
  return(df)
```

```
}
outlier_columns <- c("ricepds_v", "chicken_q")
for (col in outlier_columns) {
  MANPRnew <- remove_outliers(MANPRnew, col)
}

# Summarize consumption
MANPRnew$total_consumption <- rowSums(MANPRnew[, c("ricepds_v", "Wheatpd
s_q", "chicken_q", "pulsep_q", "wheatos_q")], na.rm = TRUE)

# Summarize and display top and bottom consuming districts and regions
summarize_consumption <- function(group_col) {
  summary <- MANPRnew %>%
    group_by(across(all_of(group_col))) %>%
    summarise(total = sum(total_consumption)) %>%
    arrange(desc(total))
  return(summary)
}
district_summary <- summarize_consumption("District")
region_summary <- summarize_consumption("Region")

cat("Top 3 Consuming Districts:\n")
print(head(district_summary, 3))
cat("Bottom 3 Consuming Districts:\n")
print(tail(district_summary, 3))

cat("Region Consumption Summary:\n")
print(region_summary)

# Rename districts and sectors , get codes from appendix of NSSO 68th ROund Data
district_mapping <- c("6" = "Imphal West", "7" = "Imphal East", "4" = "Bishnupur", "
1"= "Senapati", "2"= "Tamenglong", "3"= "Churachandpur","5"= "Thoubal","8"="Uk
hrul", "9"="Chandel")
sector_mapping <- c("2" = "URBAN", "1" = "RURAL")
```

```
MANPRnew$District <- as.character(MANPRnew$District)
MANPRnew$Sector <- as.character(MANPRnew$Sector)
MANPRnew$District <- ifelse(MANPRnew$District %in% names(district_mapping),
district_mapping[MANPRnew$District], MANPRnew$District)
MANPRnew$Sector <- ifelse(MANPRnew$Sector %in% names(sector_mapping), se
ctor_mapping[MANPRnew$Sector], MANPRnew$Sector)
View(MANPRnew)

# MANPR_consumption stores the aggregate of the consumption district wise
MANPR_consumption <- aggregate(total_consumption ~ District, data = MANPRnew
, sum)
View(MANPR_consumption)

# histogram to show the distribution of total consumption across different districts
hist(MANPRnew$total_consumption, breaks = 6, col = 'yellow', border = 'black',
    xlab = "Consumption", ylab = "Frequency", main = "Consumption Distribution in
MANIPUR State")

# barplot to visualize consumption per district with district names
??barplot
barplot(MANPR_consumption$total_consumption,
      names.arg = MANPR_consumption$District,
      las = 2, # Makes the district names vertical
      col = 'yellow',
      border = 'black',
      xlab = "District",
      ylab = "Total Consumption",
      main = "Total Consumption per District",
      cex.names = 0.7)
```

Python Codes

```r
# Set the working directory and verify it
setwd('D:\\Assignments_SCMA632')
getwd()
# Function to install and load libraries
install_and_load <- function(package) {
  if (!require(package, character.only = TRUE)) {
    install.packages(package, dependencies = TRUE)
    library(package, character.only = TRUE)
  }
}

# Load required libraries
libraries <- c("dplyr", "readr", "readxl", "tidyr", "ggplot2", "BSDA","glue
","sf")
lapply(libraries, install_and_load)
# Reading the file into R
data <- read.csv("NSSO68.csv")



# a)Plotting a histogram and a Barplot of the data to indicate the consumpt
ion district-wise for the West Bengal


# Filtering for MANPR
df <- data %>%
  filter(state_1 == "MANPR")


# Display dataset info
cat("Dataset Information:\n")
print(names(df))
print(head(df))
print(dim(df))
# Sub-setting the data
MANPRnew <- df %>%
  select(state_1, District, Region, Sector, State_Region, Meals_At_Home, ri
cepds_v, Wheatpds_q, chicken_q, pulsep_q, wheatos_q, No_of_Meals_per_day)


# Check for missing values in the subset
cat("Missing Values in Subset:\n")
```

```
print(colSums(is.na(MANPRnew)))
# Impute missing values with mean for specific columns
impute_with_mean <- function(column) {
  if (any(is.na(column))) {
    column[is.na(column)] <- mean(column, na.rm = TRUE)
  }
  return(column)
}
MANPRnew$Meals_At_Home <- impute_with_mean(MANPRnew$Meals_At_Home)

# Check for missing values after imputation
cat("Missing Values After Imputation:\n")
print(colSums(is.na(MANPRnew)))
# Finding outliers and removing them
remove_outliers <- function(df, column_name) {
  Q1 <- quantile(df[[column_name]], 0.25)
  Q3 <- quantile(df[[column_name]], 0.75)
  IQR <- Q3 - Q1
  lower_threshold <- Q1 - (1.5 * IQR)
  upper_threshold <- Q3 + (1.5 * IQR)
  df <- subset(df, df[[column_name]] >= lower_threshold & df[[column_name]]
<= upper_threshold)
  return(df)
}
outlier_columns <- c("ricepds_v", "chicken_q")
for (col in outlier_columns) {
  MANPRnew <- remove_outliers(MANPRnew, col)
}

# Summarize consumption
MANPRnew$total_consumption <- rowSums(MANPRnew[, c("ricepds_v", "Wheatpds_q
", "chicken_q", "pulsep_q", "wheatos_q")], na.rm = TRUE)

# Summarize and display top and bottom consuming districts and regions
summarize_consumption <- function(group_col) {
  summary <- MANPRnew %>%
    group_by(across(all_of(group_col))) %>%
    summarise(total = sum(total_consumption)) %>%
```

```
    arrange(desc(total))
  return(summary)
}
district_summary <- summarize_consumption("District")
region_summary <- summarize_consumption("Region")


cat("Top 3 Consuming Districts:\n")
## Top 3 Consuming Districts:
print(head(district_summary, 3))
cat("Bottom 3 Consuming Districts:\n")
## Bottom 3 Consuming Districts:
print(tail(district_summary, 3))
cat("Region Consumption Summary:\n")
## Region Consumption Summary:
print(region_summary)
# Rename districts and sectors , get codes from appendix of NSSO 68th ROund Data
district_mapping <- c("6" = "Imphal West", "7" = "Imphal East", "4" = "Bishnupur", "1"= "Senapati", "2"= "Tamenglong", "3"= "Churachandpur","5"= "Thoubal","8"="Ukhrul", "9"="Chandel")
sector_mapping <- c("2" = "URBAN", "1" = "RURAL")


MANPRnew$District <- as.character(MANPRnew$District)
MANPRnew$Sector <- as.character(MANPRnew$Sector)
MANPRnew$District <- ifelse(MANPRnew$District %in% names(district_mapping), district_mapping[MANPRnew$District], MANPRnew$District)
MANPRnew$Sector <- ifelse(MANPRnew$Sector %in% names(sector_mapping), sector_mapping[MANPRnew$Sector], MANPRnew$Sector)
View(MANPRnew)


# MANPR_consumption stores the aggregate of the consumption district wise
MANPR_consumption <- aggregate(total_consumption ~ District, data = MANPRnew, sum)
View(MANPR_consumption)


# histogram to show the distribution of total consumption across different districts
hist(MANPRnew$total_consumption, breaks = 6, col = 'yellow', border = 'black',
     xlab = "Consumption", ylab = "Frequency", main = "Consumption Distribution in MANIPUR State")
```

```
# barplot to visualize consumption per district with district names
??barplot
barplot(MANPR_consumption$total_consumption,
        names.arg = MANPR_consumption$District,
        las = 2, # Makes the district names vertical
        col = 'yellow',
        border = 'black',
        xlab = "District",
        ylab = "Total Consumption",
        main = "Total Consumption per District",
        cex.names = 0.7)
```

References

www.github.com