VIRGINIA COMMONWEALTH UNIVERSITY

Statistical analysis and modelling (SCMA 632)

A5.a : Visualisation – Perceptual Mapping for Business

SAYA SANTHOSH

V01101901

Date of Submission: 15-07-2024

CONTENTS

Introduction - Histogram and Barplot to indicate the consumption district-wise for West Bengal

This study primarily examines the state of West Bengal, utilizing data from the National Sample Survey Office (NSSO) to investigate consumption patterns at the district level. Our objective is to graphically represent the distribution of total consumption in different districts using a histogram. Additionally, it present a comprehensive picture of consumption by district using a barplot. The NSSO68 dataset contains extensive consumption-related data for both rural and urban sectors. The analysis entails managing missing values, detecting and eliminating outliers, and normalizing district and sector names. Our objective is to offer significant insights into consumption patterns within West Bengal by summarizing the consumption statistics at both regional and district levels. These visualizations will aid policymakers and stakeholders in comprehending the variations in consumption patterns among different districts. This will enable them to implement specific interventions and promote fair and balanced development throughout the state.

Objectives :

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

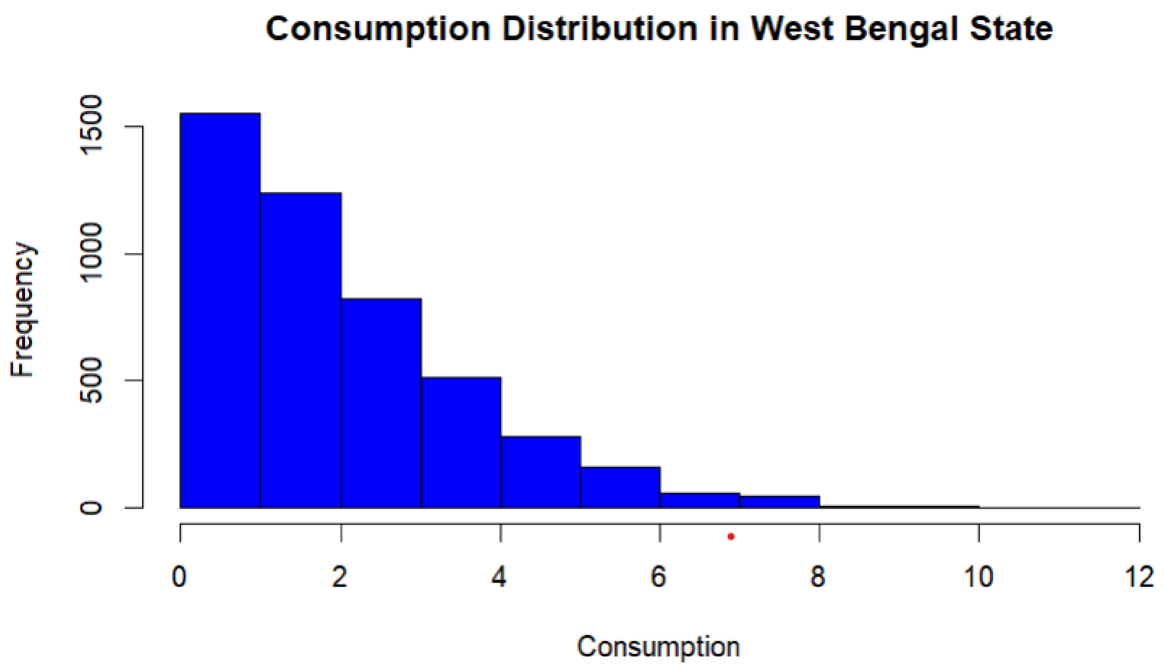
Business Significance :

Use a histogram and barplot to display the consumption of West Bengal's districts. The barplot and histogram tasks, which display consumption patterns across districts in West

Bengal, have substantial business consequences. They offer crucial observations for the distribution of resources, allowing for effective strategizing of food provisions and the construction of infrastructure tailored to the consumption requirements of each district. Companies can utilize this data to customize marketing strategies and product offerings, efficiently focusing on consumer preferences in various districts. Furthermore, comprehending discrepancies in consumption aids in enhancing supply chain management, guaranteeing that products are delivered effectively to fulfill diverse levels of demand. Policymakers can gain advantages by developing focused initiatives that aim to promote equitable consumption and enhance public health outcomes. These visualizations also aid in competitive analysis, enabling organizations to identify market opportunities and gaps for strategic market entry and positioning. Conducting an examination of consumption patterns helps in evaluating the socioeconomic impact and promoting measures that encourage healthier eating habits and lifestyle choices in West Bengal.

Results and Interpretation using R

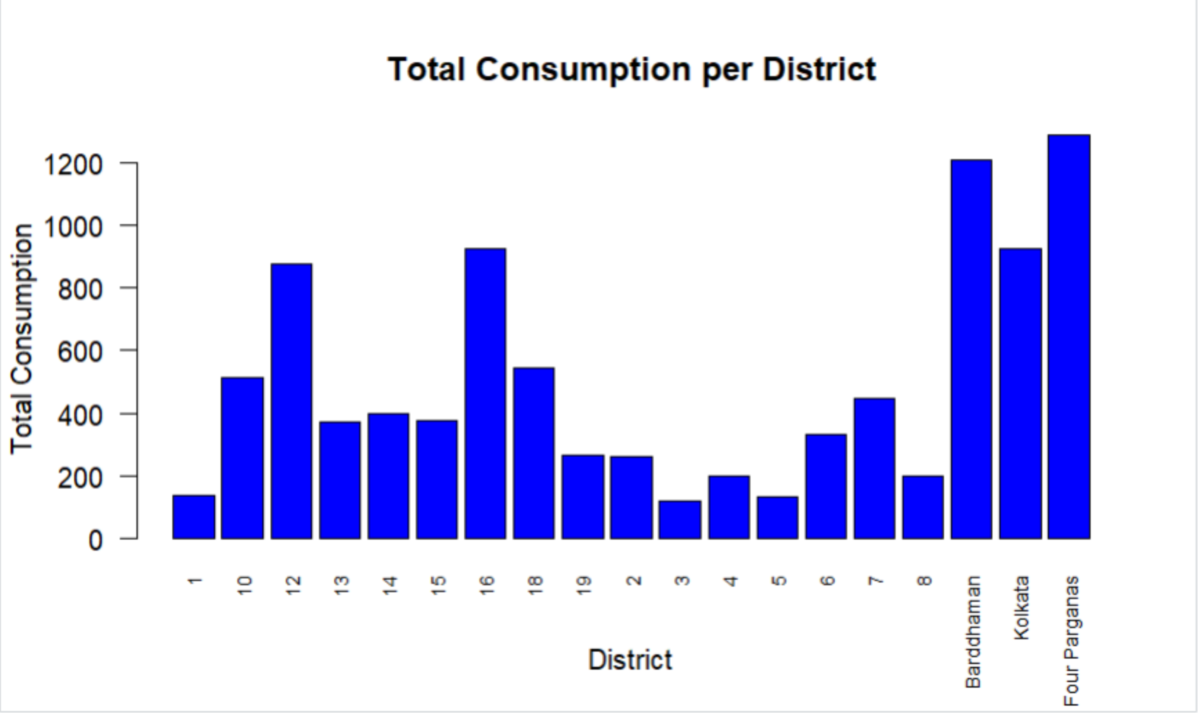
```
> # histogram to show the distribution of total consumption across differe
nt districts
> hist(wbnew$total_consumption, breaks = 15, col = 'blue', border = 'black
',
+      xlab = "Consumption", ylab = "Frequency", main = "Consumption Distr
ibution in West Bengal State")
```



Interpretation:

The histogram titled "Consumption Distribution in West Bengal State" illustrates the frequency distribution of consumption values within the state. The x-axis represents the consumption levels, ranging from 0 to 12, while the y-axis shows the frequency of each consumption level. The data displays a right-skewed distribution, with the majority of observations clustered around lower consumption values. The highest frequency is observed for consumption levels between 0 and 2. As consumption increases, the frequency significantly drops, indicating fewer instances of higher consumption values. There is also an outlier visible around the consumption level of 8. This distribution suggests that most individuals or households in West Bengal have low to moderate consumption, with a sharp decline in the number of high-consumption cases.

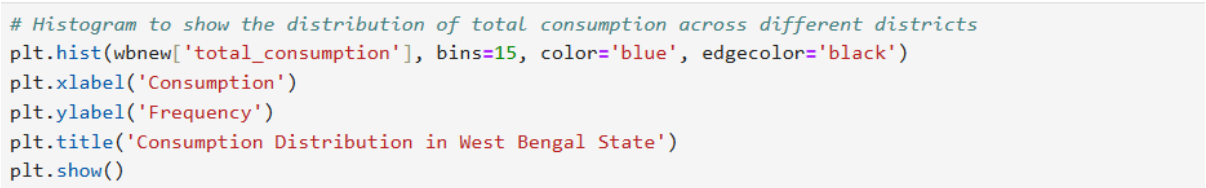
```
> # barplot to visualize consumption per district with district names
> ??barplot
> barplot(wb_consumption$total_consumption,
+         names.arg = wb_consumption$District,
+         las = 2, # Makes the district names vertical
+         col = 'blue',
+         border = 'black',
+         xlab =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



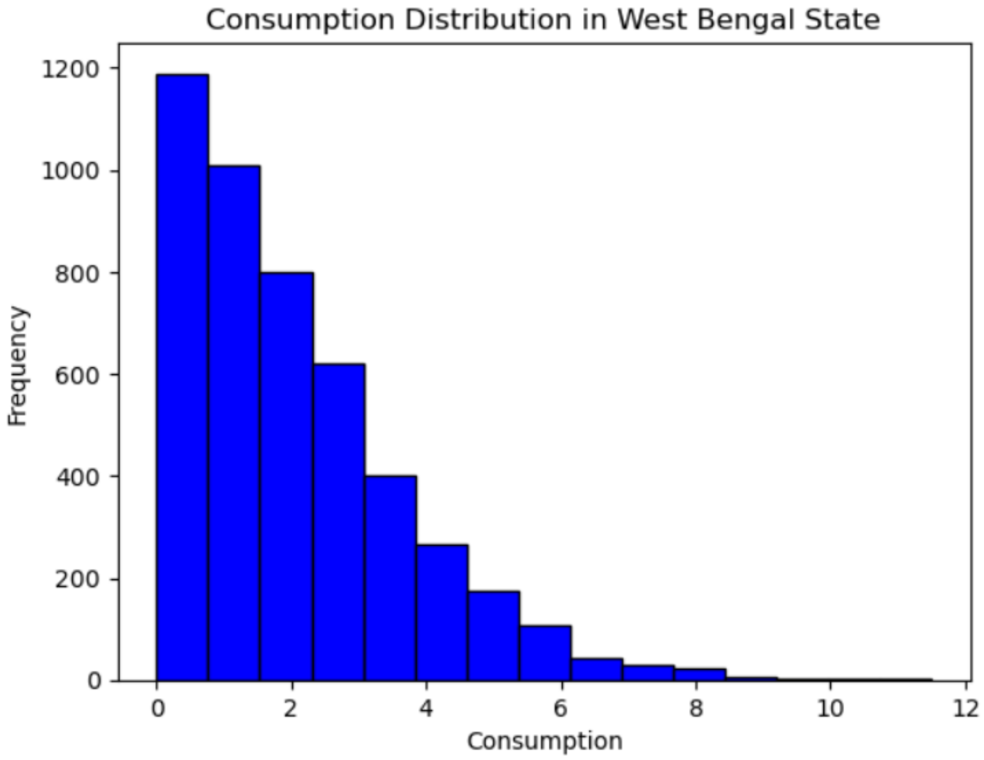
Interpretation:

The bar chart titled "Total Consumption per District" shows the total consumption values for various districts in West Bengal. Each bar represents a district, with the height of the bar indicating the total consumption for that district. The x-axis lists the districts, with some labeled numerically and others by name, such as Barddhaman, Kolkata, and Four Parganas. The y-axis represents the total consumption, ranging up to 1200 units. Notable peaks in total consumption are observed in districts 12, 16, Barddhaman, and Four Parganas, with these districts having the highest consumption values. Conversely, several districts exhibit significantly lower consumption, indicating a varied distribution of consumption across the state. This chart highlights the disparities in consumption levels among different districts, with a few districts contributing disproportionately to the total consumption.

Results and Interpretation using Python

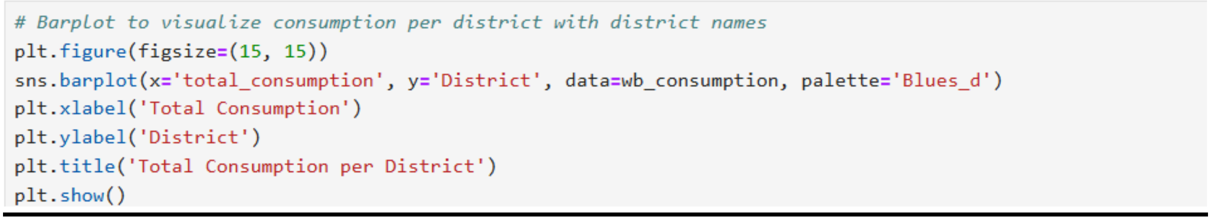


```
# Histogram to show the distribution of total consumption across different districts
plt.hist(wbnew['total_consumption'], bins=15, color='blue', edgecolor='black')
plt.xlabel('Consumption')
plt.ylabel('Frequency')
plt.title('Consumption Distribution in West Bengal State')
plt.show()
```

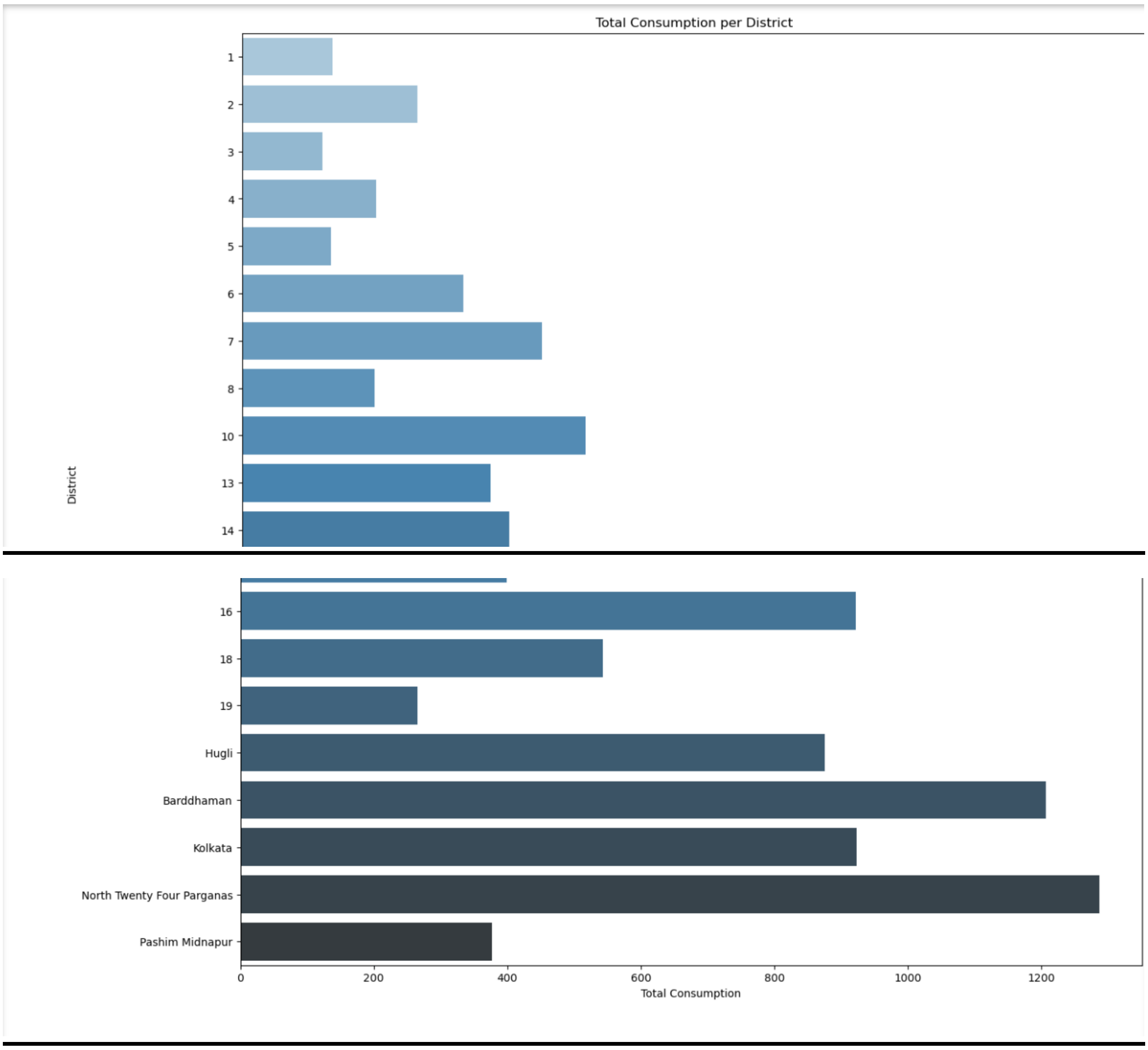


Interpretation:

The histogram titled "Consumption Distribution in West Bengal State" presents the frequency of different consumption levels within the state. The x-axis indicates consumption values ranging from 0 to 12, while the y-axis shows the frequency of occurrences for each consumption level. The distribution is highly right-skewed, with the highest frequency observed at the lowest consumption levels (0 to 2). As consumption increases, the frequency of occurrences steadily decreases, with very few instances of consumption levels above 6. This pattern suggests that most households or individuals in West Bengal have relatively low consumption, and higher consumption levels are quite rare. The overall trend highlights a significant concentration of lower consumption values, indicating a possible economic disparity within the state.



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



Interpretation:

The bar charts depict the total consumption across various districts.The chart shows a range of total consumption values for districts with a noticeable increase from district 1 to district 14 and additional districts and named locations such as Hugli, Barddhaman, Kolkata, North Twenty Four Parganas, and Pashim Midnapur. Among these, North Twenty Four Parganas shows the highest total consumption, while Pashim Midnapur has the lowest. This suggests a significant disparity in consumption levels across the districts, with urban or more densely populated areas likely exhibiting higher consumption rates.

Recommendations

It is recommended to address consumption disparities by supporting low-consumption households through improved access to resources and infrastructure. In high-consumption areas, promoting energy efficiency and sustainable practices can help reduce excessive use. Analyzing the factors behind varying consumption levels will inform these efforts. By enhancing infrastructure and resource availability in low-consumption districts and encouraging efficiency in high-consumption areas, we can achieve more balanced and sustainable consumption patterns across the state.

R Codes

```
# Set the working directory and verify it
setwd('C:\\Users\\sayas\\OneDrive\\New folder\\python projects')
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 WB
df <- data %>%
  filter(state_1 == "WB")

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

# Sub-setting the data
wbnew <- df %>%
  select(state_1, District, Region, Sector, State_Region, Meals_At_Home, ricepds_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(wbnew)))

# 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)
}
```

```
wbnew$Meals_At_Home <- impute_with_mean(wbnew$Meals_At_Home)

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

# 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) {
  wbnew <- remove_outliers(wbnew, col)
}

# Summarize consumption
wbnew$total_consumption <- rowSums(wbnew[, 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 <- wbnew %>%
    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("11" = "North Twenty Four Parganas", "9" = "Barddhaman", "17" = "Kolkata")
sector_mapping <- c("2" = "URBAN", "1" = "RURAL")

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

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

# histogram to show the distribution of total consumption across different districts
hist(wbnew$total_consumption, breaks = 15, col = 'blue', border = 'black',
     xlab = "Consumption", ylab = "Frequency", main = "Consumption Distribution in West Bengal State")

# barplot to visualize consumption per district with district names
```

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

Python Codes

```
# Import necessary libraries
!pip install geopandas pandas numpy matplotlib seaborn
import os
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import geopandas as gpd
# Set the working directory and verify it
os.chdir('C:\\Users\\sayas\\OneDrive\\New folder\\python projects')
print(os.getcwd())

# Reading the file into Python
data = pd.read_csv("NSSO68.csv")
# a) Plotting a histogram and a barplot of the data to indicate the consumption district-wise for West Bengal

# Filtering for WB
```

```
df = data[data['state_1'] == "WB"]

# Display dataset info
print("Dataset Information:")
print(df.columns)
print(df.head())
print(df.shape)
# Sub-setting the data
wbnew = df[['state_1', 'District', 'Region', 'Sector', 'State_Region', 'Meals_At_Home', '
ricepds_v', 'Wheatpds_q', 'chicken_q', 'pulsep_q', 'wheatos_q', 'No_of_Meals_per_day'
]]

# Check for missing values in the subset
print("Missing Values in Subset:")
print(wbnew.isnull().sum())

# Impute missing values with mean for specific columns
wbnew['Meals_At_Home'].fillna(wbnew['Meals_At_Home'].mean(), inplace=True)

# Check for missing values after imputation
print("Missing Values After Imputation:")
print(wbnew.isnull().sum())
# Function to remove outliers
def remove_outliers(df, column_name):
    Q1 = df[column_name].quantile(0.25)
    Q3 = df[column_name].quantile(0.75)
    IQR = Q3 - Q1
    lower_threshold = Q1 - (1.5 * IQR)
    upper_threshold = Q3 + (1.5 * IQR)
    df = df[(df[column_name] >= lower_threshold) & (df[column_name] <= upper_thr
eshold)]
    return df
outlier_columns = ['ricepds_v', 'chicken_q']
for col in outlier_columns:
    wbnew = remove_outliers(wbnew, col)
# Summarize consumption
wbnew['total_consumption'] = wbnew[['ricepds_v', 'Wheatpds_q', 'chicken_q', 'pulsep_q', 'wheatos_q']].sum(axis=1)

# Summarize and display top and bottom consuming districts and regions
district_summary = wbnew.groupby('District')['total_consumption'].sum().reset_index().sort_values(by='total_consumption', ascending=False)
region_summary = wbnew.groupby('Region')['total_consumption'].sum().reset_index().sort_values(by='total_consumption', ascending=False)

print("Top 3 Consuming Districts:")
print(district_summary.head(3))
print("Bottom 3 Consuming Districts:")
print(district_summary.tail(3))

print("Region Consumption Summary:")
print(region_summary)
# Rename districts and sectors
district_mapping = {"11": "North Twenty Four Parganas", "9": "Barddhaman", "17": "Kolkata"}
sector_mapping = {"2": "URBAN", "1": "RURAL"}

wbnew['District'] = wbnew['District'].astype(str).map(district_mapping).fillna(wbnew['District'])
wbnew['Sector'] = wbnew['Sector'].astype(str).map(sector_mapping).fillna(wbnew['Sector'])
print(wbnew)
# wb_consumption stores the aggregate of the consumption district-wise
wb_consumption = wbnew.groupby('District')['total_consumption'].sum().reset_index()
print(wb_consumption)
# Histogram to show the distribution of total consumption across different districts
plt.hist(wbnew['total_consumption'], bins=15, color='blue', edgecolor='black')
plt.xlabel('Consumption')
plt.ylabel('Frequency')
plt.title('Consumption Distribution in West Bengal State')
plt.show()
# Barplot to visualize consumption per district with district names
plt.figure(figsize=(15, 15))
sns.barplot(x='total_consumption', y='District', data=wb_consumption, palette='Blues_d')
plt.xlabel('Total Consumption')
plt.ylabel('District')
plt.title('Total Consumption per District')
plt.show()
```

References

1. www.github.com