VIRGINIA COMMONWEALTH UNIVERSITY

Statistical analysis and modelling (SCMA 632)

A1a: Preliminary preparation and analysis of data, Descriptive statistics

SAI PREETHI ANANYA NAIDU

V01151224

Date of Submission: 12-06-2025

CONTENTS

Sl. No.	Title	Page No.
1.	Introduction	1-2
2.	Results & Interpretations	3-7
3.	Recommendations	8-9
4.	Codes	10-15

ANALYSING CONSUMPTION IN THE STATE OF NAGALAND USING R AND PYTHON

INTRODUCTION

The northeastern Indian state of Nagaland is well known for its rich cultural legacy and diverse ecosystem. Developing evidence-based policies in areas like socioeconomic development, healthcare planning, and environmental conservation requires a thorough grasp of the state's consumption patterns. The purpose of this study is to use analytical tools and the R and Python programming languages to examine and interpret consumption patterns in Nagaland.

The available dataset offers a thorough summary of food consumption trends in the Indian state of Nagaland. It records key elements of dietary habits in both urban and rural areas, such as the number of meals eaten at home, the consumption of particular foods like rice, pulses, meat, and cereals, and the average number of meals eaten daily. The information provides insightful information about food preferences and nutritional intake in different Nagaland districts and regions.

Determining inequalities in food access and developing evidence-based plans to enhance the state's nutrition and general well-being depend on an understanding of these trends. Additionally, this analysis makes it easier to compare districts and demographic groups in a meaningful way, which helps to make interventions more focused and effective.

OBJECTIVES

The following objectives are the focus of this study:

a) Check if there are any missing values in the data, identify them, and if there are, replace them with the mean of the variable
b) Check for outliers, describe your test's outcome, and make suitable amendments.
c) Rename the districts and sectors, viz., rural and urban.
d) Summarize the critical variables in the data set region-wise and district-wise and indicate the top and bottom three districts of consumption.
e) Test whether the differences in the means are significant or not.

BUSINESS SIGNIFICANCE

Analyzing Nagaland's food consumption statistics is important for business and policy decisions. This study enables different stakeholders to make well-informed decisions about resource distribution, nutritional program design, and market expansion by identifying consumption trends and regional disparities.

- The results can be used by government agencies and policymakers to promote equitable development by implementing customized food security and nutrition programs, particularly in underprivileged areas.
- Health and Wellness Sector: Nutritionists and other medical professionals can assess dietary imbalances and employ culturally relevant approaches to address public health concerns based on actual consumption data.
- Academic and Research Institutions: Researchers can further fields like food economics, sociology, and public health by examining the relationships between dietary habits and socioeconomic conditions.

Reliable insights that support focused planning and sustainable development throughout Nagaland are ensured by the structured analysis using data cleaning, descriptive statistics, and hypothesis testing.

Results & Interpretations

(a) Handling Missing Data

```
## Missing Values Before Imputation:
```

```
print(missing_info)
```

```
##             state_1            District              Region              Sector
##                   0                   0                   0                   0
##        State_Region       Meals_At_Home         ricetotal_v        wheattotal_v
##                   0                   1                   0                   0
##         Milktotal_v         pulsestot_v       nonvegtotal_v          fruitstt_v
##                   0                   0                   0                   0
## No_of_Meals_per_day
##                   0
```

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

missing_info <- colSums(is.na(state_subset))
cat("Missing Values After Imputation:\n")
```

```
## Missing Values After Imputation:
```

```
print(missing_info)
```

```
##             state_1            District              Region              Sector
##                   0                   0                   0                   0
##        State_Region       Meals_At_Home         ricetotal_v        wheattotal_v
##                   0                   0                   0                   0
##         Milktotal_v         pulsestot_v       nonvegtotal_v          fruitstt_v
##                   0                   0                   0                   0
## No_of_Meals_per_day
##                   0
```

The variable Meals_At_Home had a missing entry(one), according to an initial analysis of the dataset, which showed that missing values were minimal. The mean of each corresponding column was used to impute missing values in numeric columns to solve this problem. This technique, which assumes that data are missing at random (MAR), offers a straightforward but efficient way to preserve dataset completeness while reducing the possibility of bias. This method of imputing missing values avoids the information loss that would come from row-wise deletion while maintaining the total number of observations.

(b) Outlier Detection and Treatment

```r
# 8. Remove outliers from specific columns ####
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('Meals_At_Home', 'ricetotal_v', 'wheattotal_v', 'Milktotal_v',
                     'pulsestot_v','nonvegtotal_v', 'fruitstt_v', 'No_of_Meals_per_day')

for (col in outlier_columns) {
  state_subset <- remove_outliers(state_subset, col)
}
```

The Interquartile Range (IQR) approach was used to identify outliers. The observations that fell outside of the range specified by Q1 - 1.5×IQR and Q3 + 1.5×IQR were deemed outliers and were subsequently eliminated for each chosen numeric variable (e.g., ricetotal_v, milktotal_v, meals_at_home, etc.). This method successfully reduces the impact of extreme or spurious values on further statistical analysis and is robust to non-normal distributions. More stable estimates of central tendency and variation are ensured by eliminating these outliers, especially when comparing means across groups.

(c) Renaming Districts and Sectors

```r
# 11. Rename district and sector codes (based on NSS codes) ####
district_mapping <- c(
  "1" = "Mon", "2" = "Tuensang", "3" = "Mokokchung", "4" = "Zunheboto",
  "5" = "Wokha", "6" = "Dimapur", "7" = "Kohima", "8" = "Phek",
  "9" = "Kiphire", "10" = "Longleng", "11" = "Peren"
)

sector_mapping <- c("2" = "URBAN", "1" = "RURAL")

state_subset$District <- as.character(state_subset$District)
state_subset$Sector <- as.character(state_subset$Sector)

state_subset$District <- ifelse(state_subset$District %in% names(district_mapping),
                                district_mapping[state_subset$District],
                                state_subset$District)

state_subset$Sector <- ifelse(state_subset$Sector %in% names(sector_mapping),
                              sector_mapping[state_subset$Sector],
                              state_subset$Sector)

# Update summaries again after mapping
district_summary <- summarize_consumption("District")
region_summary <- summarize_consumption("Region")
sector_summary <- summarize_consumption("Sector")
```

District and sector codes were mapped to their corresponding names to enhance the dataset's interpretability. For example, sector codes "1" and "2" were renamed as "RURAL" and "URBAN," respectively, and district codes like "1" were mapped to "Mon." Stakeholders can readily comprehend which geographic regions are being mentioned in the summary statistics and comparative tests thanks to this transformation, which makes it possible for analyses and visualizations to be clearer and more meaningful.

(d) Summary of Consumption Patterns by Region and District

```
print(head(district_summary, 4))
```

```
## # A tibble: 4 × 2
##   District    total
##   <chr>       <dbl>
## 1 Mokokchung 45477.
## 2 Kohima     38150.
## 3 Mon        31954.
## 4 Tuensang   30952.
```

```
cat("Updated Bottom Consuming Districts:\n")
```

```
## Updated Bottom Consuming Districts:
```

```
print(tail(district_summary, 4))
```

```
## # A tibble: 4 × 2
##   District  total
##   <chr>     <dbl>
## 1 Longleng 26529.
## 2 Kiphire  23318.
## 3 Wokha    19026.
## 4 Peren    18259.
```

By adding up important food categories like rice, wheat, milk, pulses, non-vegetarian foods, and fruits, a new variable called total_consumption was created. This variable functioned as a composite indicator of food consumption in households. District, regional, and sector-level summaries were produced.

District-level Results:

Districts with the highest total consumption:

- Mokokchung
- The Kohima
- Mon

Districts with the lowest total consumption:

- Peren (18,259 units)
- Wokha (19,026 pieces)
- 23,318 units of Kiphire

The differences in household consumption throughout Nagaland are reflected in these findings. While districts like Peren and Wokha might encounter obstacles because of geography, infrastructure, or financial limitations, districts like Mokokchung and Kohima might gain from improved market access and infrastructure.

Region-level Summary: Because Nagaland is only included in one NSS region, it is difficult to perform intra-regional comparisons.

Sector-level Summary: Compared to urban areas, rural areas had substantially higher aggregate consumption, which may be explained by larger household sizes or a higher reliance on food produced locally.

(e) Testing Hypotheses for Mean Differences

```
# 12. Test for mean difference between Urban and Rural consumption ####
rural <- state_subset %>% filter(Sector == "RURAL") %>% select(total_consumption)
urban <- state_subset %>% filter(Sector == "URBAN") %>% select(total_consumption)

z_test_result <- z.test(rural, urban, alternative = "two.sided",
                        mu = 0, sigma.x = 2.1, sigma.y = 2.3, conf.level = 0.95)

if (z_test_result$p.value < 0.05) {
  cat("P value is <", 0.05, "→ Reject H0: Urban and Rural means differ significantly.\n")
} else {
  cat("P value is >=", 0.05, "→ Fail to reject H0: No significant difference in Urban vs Rural means.\n")
}
```

```
## P value is < 0.05 → Reject H0: Urban and Rural means differ significantly.
```

```
# 13. Test for mean difference between Top and Bottom Districts ####
top_district <- head(district_summary$District, 1)
bottom_district <- tail(district_summary$District, 1)

top_data <- state_subset %>% filter(District == top_district) %>% select(total_consumption)
bottom_data <- state_subset %>% filter(District == bottom_district) %>% select(total_consumption)

z_test_result2 <- z.test(top_data, bottom_data, alternative = "two.sided",
                         mu = 0, sigma.x = 2.1, sigma.y = 2.3, conf.level = 0.95)

if (z_test_result2$p.value < 0.05) {
  cat("P value is <", 0.05, "→ Reject H0: Top and Bottom district means differ significantly.\n")
} else {
  cat("P value is >=", 0.05, "→ Fail to reject H0: No significant difference between top and bottom district consumptio
n.\n")
}
```

```
## P value is < 0.05 → Reject H0: Top and Bottom district means differ significantly.
```

To determine whether the variations in mean consumption between the chosen groups were statistically significant, two-sample z-tests were used.

Comparing Rural and Urban Areas:

- Z-Score: roughly -4.62
- P-Value: less than 0.001
- Conclusion: There may be structural or economic differences between the two sectors,

as evidenced by the statistically significant difference in mean total consumption between rural and urban households.

Comparison of Top and Bottom Districts (Mokokchung vs. Peren):

- Z-Score: roughly 5.13
- P-Value: less than 0.001
- Conclusion: Intra-state disparity is confirmed by the notable difference in mean consumption between the districts with the highest consumption (Mokokchung) and the lowest consumption (Peren).

Overall Conclusion

Analysing Nagaland's household consumption data provides important new information about the state's nutritional situation. Following meticulous handling of outliers and missing data, it was found that: The amount of food consumed by households overall varies significantly between districts. Generally speaking, rural households consume more than their urban counterparts. There are statistically significant differences in mean consumption between the districts with the highest and lowest consumption as well as between sectors. Public policy may be affected by these findings, especially when it comes to focusing interventions and directing funds to low-use districts like Peren, Wokha, and Kiphire. The approach used here can be used as a model for comparable analyses with NSSO datasets in other states.

RECOMMENDATIONS

The following suggestions are put forth to address the observed differences in household food consumption in light of the empirical results of the NSSO68 data analysis for Nagaland:

1. Focused Nutritional Assistance in Districts with Low Consumption

- The districts with the lowest levels of overall food consumption were Peren, Wokha, and Kiphire. Because of their remote location, inadequate infrastructure, or financial difficulties, these regions might not have easy access to basic food supplies.
- Take action by putting in place state-sponsored or centrally supported food distribution programs that are tailored to these districts (for example, via local cooperatives or the Public Distribution System).
- Justification: In these underprivileged areas, increasing access to staple foods can lower nutritional disparities and enhance health outcomes.

2. Enhancing Rural Food Supply Chains

- Even though overall consumption is higher in rural areas, issues with food quality and variety may still exist.
- Take action to guarantee a consistent and varied supply of food items by investing in rural infrastructure, such as roads, storage facilities, and market connections.
- Justification: In remote rural areas, improved supply chains will increase food availability, affordability, and freshness.

3. Development of District-Specific Policies

- Considerable differences in consumption levels raise the possibility that a "one-size-fits-all" approach may not work.
- Take action: Create district-level action plans for food security that take into account opportunities and challenges specific to the area.
- Justification: While Peren and Wokha might need simple access interventions, Mokokchung and Kohima might gain more from behavior-based programs (such as dietary diversification).

4. Encourage dietary diversity and nutrition education

- A balanced diet that includes dairy, fruits, vegetables, and proteins is crucial, even though overall consumption is significant.
- Take action: Run neighborhood-based nutrition education programs, particularly in areas with high carbohydrate dependence or low consumption.
- Justification: Public health outcomes can be greatly enhanced by teaching households about inexpensive and nutrient-dense food combinations.

5. Increase Opportunities for Livelihood to Increase Purchasing Power

- Rather than just availability, lower consumption in some districts may be linked to economic disadvantages.
- Take action by implementing or growing revenue-generating initiatives (such as agricultural extension services, microfinance, and rural employment programs).
- Justification: Households can eat more food of higher quality and quantity when their economic capacity improves.

6. Combine Health and Food Information for Holistic Planning

- Important information on food consumption is provided by this analysis. Planning an intervention can be made more comprehensive by combining this with data on malnutrition and health.
- Action: Encourage data integration across departments (Health, Rural Development, Agriculture) to build a composite index of food and nutritional security.
- Justification: Holistic insights lead to more effective and better-targeted public health and nutrition policies.

RESULTS & INTERPRETATIONS

R code:

```
# 0. Set Working Directory
setwd("C:\\Users\\user\\Desktop\\VCU\\BOOT CAMP\\SCMA-632-C51 - STATISTICL ANALYSIS &
MODELING\\VCU_christ")
getwd()


# 1. Installing and Importing Necessary libraries ####

# 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", "tidyr", "ggplot2", "BSDA") # Vector of required packages
lapply(libraries, install_and_load)

# 2. Reading the dataset into R ####

data <- read.csv("NSSO68.csv")

# 3. Filtering data for Nagaland (State Code 13) ####
state_name <- "NAG"
state_data <- data %>%
  filter(state_1 == state_name)  # Filter Nagaland
write.csv(state_data, 'Nagaland_filtered_data.csv', row.names = FALSE)


unique(data$state_1)
unique(state_data$state_1)


# 4. Display dataset information ####
cat("Dataset Information:\n")
print(names(state_data))
print(head(state_data))
print(dim(state_data))
sum(is.na(state_data))

# 5. Check for missing values ####
missing_info <- colSums(is.na(state_data))
cat("Missing Values Information:\n")
print(missing_info)

# 6. Select relevant columns for analysis ####
state_subset <- state_data %>%
  select(state_1, District, Region, Sector, State_Region,
         Meals_At_Home, ricetotal_v, wheattotal_v, Milktotal_v,
         pulsestot_v, nonvegtotal_v, fruitstt_v, No_of_Meals_per_day)
```

```
# 7. Impute missing values with mean ####
impute_with_mean <- function(column) {
  if (any(is.na(column))) {
    column[is.na(column)] <- mean(column, na.rm = TRUE)
  }
  return(column)
}

missing_info <- colSums(is.na(state_subset))
cat("Missing Values Before Imputation:\n")
print(missing_info)

state_subset$Meals_At_Home <- impute_with_mean(state_subset$Meals_At_Home)

missing_info <- colSums(is.na(state_subset))
cat("Missing Values After Imputation:\n")
print(missing_info)

# 8. Remove outliers from specific columns ####
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('Meals_At_Home', 'ricetotal_v', 'wheattotal_v', 'Milktotal_v',
                     'pulsestot_v','nonvegtotal_v', 'fruitstt_v', 'No_of_Meals_per_day')

for (col in outlier_columns) {
  state_subset <- remove_outliers(state_subset, col)
}

# 9. Create total consumption variable ####
state_subset$total_consumption <- rowSums(state_subset[, c('ricetotal_v', 'wheattotal_v', 'Milktotal_v',
                                                           'pulsestot_v','nonvegtotal_v', 'fruitstt_v')], na.rm = TRUE)

# 10. Summarize consumption by district and region ####
summarize_consumption <- function(group_col) {
  summary <- state_subset %>%
    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")
sector_summary <- summarize_consumption("Sector")



cat("Top Consuming Districts:\n")
print(head(district_summary, 4))
cat("Bottom Consuming Districts:\n")
print(tail(district_summary, 4))
```

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

# 11. Rename district and sector codes (based on NSS codes) ####
district_mapping <- c(
  "1" = "Mon", "2" = "Tuensang", "3" = "Mokokchung", "4" = "Zunheboto",
  "5" = "Wokha", "6" = "Dimapur", "7" = "Kohima", "8" = "Phek",
  "9" = "Kiphire", "10" = "Longleng", "11" = "Peren"
)

sector_mapping <- c("2" = "URBAN", "1" = "RURAL")

state_subset$District <- as.character(state_subset$District)
state_subset$Sector <- as.character(state_subset$Sector)

state_subset$District <- ifelse(state_subset$District %in% names(district_mapping),
                                district_mapping[state_subset$District],
                                state_subset$District)

state_subset$Sector <- ifelse(state_subset$Sector %in% names(sector_mapping),
                              sector_mapping[state_subset$Sector],
                              state_subset$Sector)

# Update summaries again after mapping
district_summary <- summarize_consumption("District")
region_summary <- summarize_consumption("Region")
sector_summary <- summarize_consumption("Sector")


cat("Updated Top Consuming Districts:\n")
print(head(district_summary, 4))
cat("Updated Bottom Consuming Districts:\n")
print(tail(district_summary, 4))
cat("Updated Region Consumption Summary:\n")
print(region_summary)
cat("Updated Sector Consumption Summary:\n")
print(sector_summary)

# 12. Test for mean difference between Urban and Rural consumption ####
rural <- state_subset %>% filter(Sector == "RURAL") %>% select(total_consumption)
urban <- state_subset %>% filter(Sector == "URBAN") %>% select(total_consumption)

z_test_result <- z.test(rural, urban, alternative = "two.sided",
                        mu = 0, sigma.x = 2.1, sigma.y = 2.3, conf.level = 0.95)

if (z_test_result$p.value < 0.05) {
  cat("P value is <", 0.05, "→ Reject H0: Urban and Rural means differ significantly.\n")
} else {
  cat("P value is >=", 0.05, "→ Fail to reject H0: No significant difference in Urban vs Rural means.\n")
}

# 13. Test for mean difference between Top and Bottom Districts ####
top_district <- head(district_summary$District, 1)
bottom_district <- tail(district_summary$District, 1)

top_data <- state_subset %>% filter(District == top_district) %>% select(total_consumption)
bottom_data <- state_subset %>% filter(District == bottom_district) %>% select(total_consumption)
```

```
z_test_result2 <- z.test(top_data, bottom_data, alternative = "two.sided",
                         mu = 0, sigma.x = 2.1, sigma.y = 2.3, conf.level = 0.95)

if (z_test_result2$p.value < 0.05) {
  cat("P value is <", 0.05, "→ Reject H0: Top and Bottom district means differ significantly.\n")
} else {
  cat("P value is >=", 0.05, "→ Fail to reject H0: No significant difference between top and bottom district
consumption.\n")
}
```

PYTHON CODE:

```
# 1. Setting the working directory
import os
os.chdir("C:\\Users\\user\\Desktop\\VCU\\BOOT CAMP\\SCMA-632-C51 - STATISTICL ANALYSIS &
MODELING\\VCU_christ")

# 2. Installing and Importing Necessary Libraries
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from statsmodels.stats import weightstats as stests

# 3. Reading the dataset
df = pd.read_csv("NSSO68.csv", encoding="Latin-1", low_memory=False)

# 4. Filtering data for Nagaland
state_data = df[df['state_1'] == "NAG"]
state_data.to_csv("C:/Users/user/Desktop/VCU/BOOT CAMP/SCMA-632-C51 - STATISTICL ANALYSIS &
MODELING/VCU_christ/nagaland_data.csv", index=False)

# 5. Display dataset information
print("Dataset Information:\n")
print("Column Names:")
print(state_data.columns.tolist())
print("\nFirst 5 Rows:")
print(state_data.head())
print("\nDimensions (rows, columns):")
print(state_data.shape)
print("\nTotal Missing Values:")
print(state_data.isna().sum().sum())

# 6. Check for missing values in each column
missing_values = state_data.isnull().sum().sort_values(ascending=False)
print("Missing Values per Column (Descending Order):\n")
print(missing_values)

# 7. Subsetting the dataset
state_subset = state_data[[
    'state_1', 'District', 'Region', 'Sector', 'State_Region',
    'Meals_At_Home', 'ricetotal_v', 'wheattotal_v', 'Milktotal_v',
    'pulsestot_v', 'nonvegtotal_v', 'fruitstt_v', 'No_of_Meals_per_day'
]]

# 8. Impute missing values with mean
print("Missing Values Before Imputation:\n")
print(state_subset.isna().sum())
```

```
state_cleaned = state_subset.fillna(state_subset.mean(numeric_only=True))

print("\n  Missing Values After Imputation:\n")
print(state_cleaned.isna().sum())

# 9. Removing outliers using IQR
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
    return df[(df[column_name] >= lower_threshold) & (df[column_name] <= upper_threshold)]

outlier_columns = [
    'Meals_At_Home', 'ricetotal_v', 'wheattotal_v', 'Milktotal_v',
    'pulsestot_v', 'nonvegtotal_v', 'fruitstt_v', 'No_of_Meals_per_day'
]

for col in outlier_columns:
    state_cleaned = remove_outliers(state_cleaned, col)

print("\n  Columns in the Cleaned Dataset:")
print(state_cleaned.columns.tolist())
# 10. Create total consumption variable
state_cleaned['total_consumption'] = state_cleaned[[
    'ricetotal_v', 'wheattotal_v', 'Milktotal_v',
    'pulsestot_v', 'nonvegtotal_v', 'fruitstt_v'
]].sum(axis=1)

# 11. Summarize consumption
def summarize_consumption(df, group_col):
    summary = df.groupby(group_col)['total_consumption'].sum().reset_index()
    summary = summary.sort_values(by='total_consumption', ascending=False)
    return summary

district_summary = summarize_consumption(state_cleaned, 'District')
region_summary = summarize_consumption(state_cleaned, 'Region')
sector_summary = summarize_consumption(state_cleaned, 'Sector')

print("\n Top 4 Consuming Districts:")
print(district_summary.head(4))
print("\n Region Consumption Summary:")
print(region_summary)
print("\n Sector Consumption Summary:")
print(sector_summary)
print("\n Bottom 4 Consuming Districts:")
print(district_summary.tail(4))

# 12. Rename district and sector codes
state_cleaned['District'] = state_cleaned['District'].astype(str)
state_cleaned['Sector'] = state_cleaned['Sector'].astype(str)

district_mapping = {
    "1": "Mon", "2": "Tuensang", "3": "Mokokchung", "4": "Zunheboto",
    "5": "Wokha", "6": "Dimapur", "7": "Kohima", "8": "Phek",
    "9": "Kiphire", "10": "Longleng", "11": "Peren"
}
sector_mapping = {"1": "RURAL", "2": "URBAN"}
```

```
state_cleaned['District'] = state_cleaned['District'].map(district_mapping).fillna(state_cleaned['District'])
state_cleaned['Sector'] = state_cleaned['Sector'].map(sector_mapping).fillna(state_cleaned['Sector'])

# Updated summaries
district_summary = summarize_consumption(state_cleaned, 'District')
region_summary = summarize_consumption(state_cleaned, 'Region')
sector_summary = summarize_consumption(state_cleaned, 'Sector')

print("\n Updated District Summary (After Mapping):")
print(district_summary.head(4))
print("\n Region Summary:")
print(region_summary)
print("\n Sector Summary:")
print(sector_summary)
# 13. Z-Test: Urban vs Rural
consumption_rural = state_cleaned[state_cleaned['Sector'] == 'RURAL']['total_consumption']
consumption_urban = state_cleaned[state_cleaned['Sector'] == 'URBAN']['total_consumption']

z_statistic, p_value = stests.ztest(consumption_rural, consumption_urban, alternative='two-sided')

print("\n Z-Test for Rural vs Urban Consumption")
print("Z-Score:", round(z_statistic, 4))
print("P-Value:", round(p_value, 4))

if p_value < 0.05:
    print("Significant difference between Rural and Urban mean consumption (Reject H₀)")
else:
    print("No significant difference between Rural and Urban mean consumption (Fail to reject H₀)")

# 14. Z-Test Between Top and Bottom Consuming Districts
top_district = district_summary.head(1).iloc[0]['District']
bottom_district = district_summary.tail(1).iloc[0]['District']

top_data = state_cleaned[state_cleaned['District'] == top_district]['total_consumption']
bottom_data = state_cleaned[state_cleaned['District'] == bottom_district]['total_consumption']

z_statistic, p_value = stests.ztest(top_data, bottom_data, alternative='two-sided')

print(f"\n Z-Test: {top_district} vs {bottom_district}")
print("Z-Score:", round(z_statistic, 4))
print("P-Value:", round(p_value, 4))

if p_value < 0.05:
    print(f'Significant difference between {top_district} and {bottom_district} mean consumption (Reject H₀)")
else:
    print(f" No significant difference between {top_district} and {bottom_district} mean consumption (Fail to
reject H₀)")
```