VIRGINIA COMMONWEALTH UNIVERSITY

STATISTICAL ANALYSIS & MODELING

# A1a: CONSUMPTION PATTERN OF ANDHRA PRADESH USING PYTHON AND R

CHANDHINI KM
V01107497

Date of Submission: 16/06/2024

CONTENTS

# Analyzing Consumption in the State of Mizoram Using R

## INTRODUCTION

This study centers on Mizoram, utilizing NSSO data to identify the top and bottom three districts in terms of consumption. We will manipulate and clean the dataset to extract the necessary information for analysis. The dataset, which includes consumption details across rural and urban sectors as well as district-wise variations, has been imported into R—a versatile statistical programming language well-suited for managing and analyzing large datasets.

Our objectives are to identify missing values, address outliers, standardize district and sector names, summarize consumption data by region and district, and test the significance of differences in means. The insights from this study can guide policymakers and stakeholders, enabling targeted interventions and promoting equitable development throughout the state.

## OBJECTIVES

a) Examine the dataset for any missing values, identify them, and if found, replace them with the mean of the respective variable.
b) Detect outliers in the data, describe the findings, and make appropriate adjustments.
c) Standardize the names of the districts and sectors, specifically rural and urban.
d) Provide a summary of key variables in the dataset by region and district, and identify the top three and bottom three districts in terms of consumption.
e) Conduct a test to determine whether the differences in the mean values are statistically significant.

## BUSINESS SIGNIFICANCE

The focus of this study on Mizoram's consumption patterns from NSSO data holds significant implications for businesses and policymakers. By identifying the top and bottom three consuming districts, the study provides valuable insights for market entry, resource allocation, supply chain optimization, and targeted interventions. Through data cleaning, outlier detection, and significance testing, the findings facilitate informed decision-making, fostering equitable development and promoting Mizoram's economic growth.

# RESULTS AND INTERPRETATION

**a) Identifying if there are any missing values in the data, and if there are replace them with the mean of the variable.**

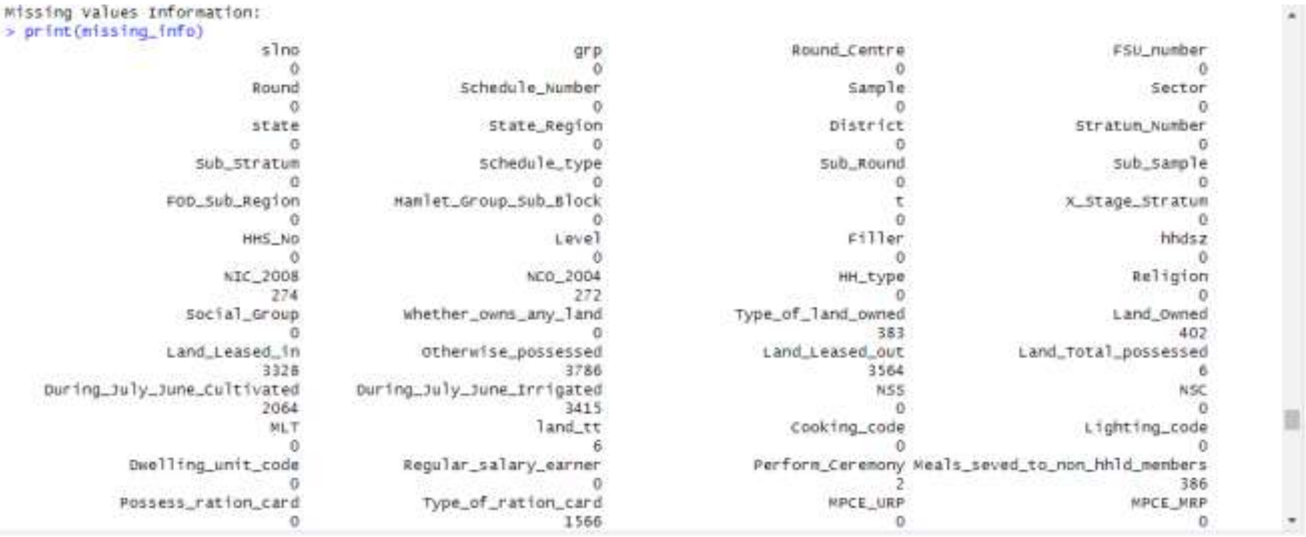
*#Finding the missing values*

**Code :**

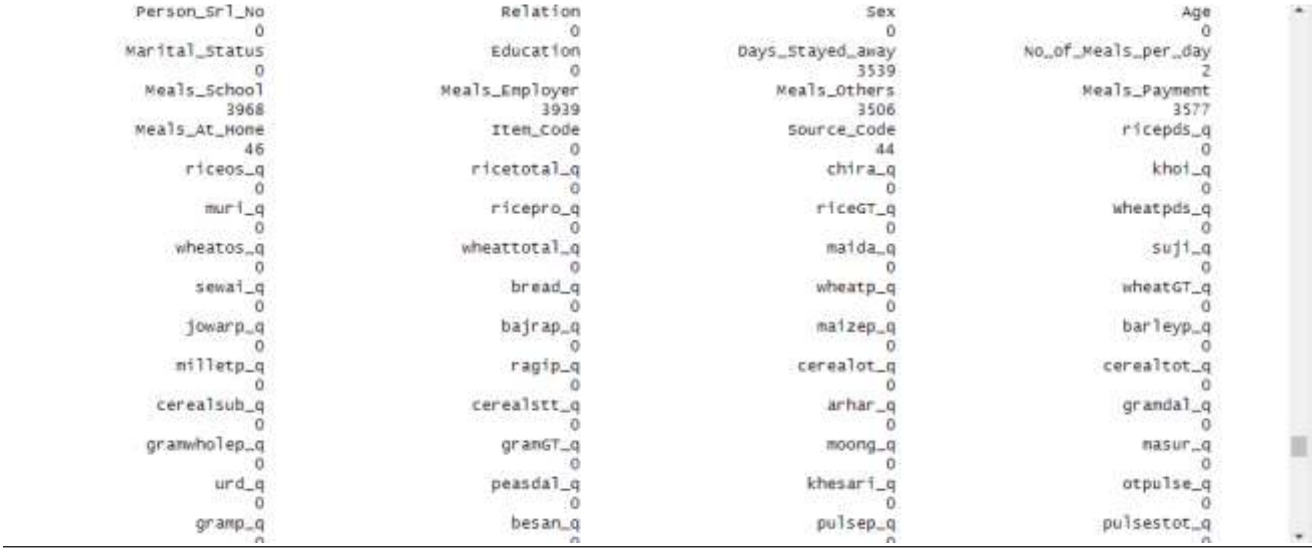
```
# Finding missing values
missing_info <- colSums(is.na(df))
cat("Missing Values Information:\n")
print(missing_info)

# Subsetting the data
apnew <- 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)
```

**Result:**



```
Missing Values Information:
> print(missing_info)
                           slno                             grp                    Round_Centre                      FSU_number
                              0                               0                               0                               0
                          Round                 Schedule_Number                          Sample                          Sector
                              0                               0                               0                               0
                          state                    State_Region                        District                  Stratum_Number
                              0                               0                               0                               0
                    Sub_Stratum                   Schedule_type                       Sub_Round                      Sub_Sample
                              0                               0                               0                               0
                 FOD_Sub_Region          Hamlet_Group_Sub_Block                               t                 X_Stage_Stratum
                              0                               0                               0                               0
                         HHS_No                           Level                          Filler                           hhdsz
                              0                               0                               0                               0
                       NIC_2008                        NCO_2004                         HH_type                        Religion
                            274                             272                               0                               0
                   Social_Group           Whether_owns_any_land              Type_of_land_owned                      Land_Owned
                              0                               0                             383                             402
                 Land_Leased_in             Otherwise_possessed                 Land_Leased_out            Land_Total_possessed
                           3328                            3786                            3564                               6
    During_July_June_Cultivated      During_July_June_Irrigated                             NSS                             NSC
                           2064                            3415                               0                               0
                            MLT                         land_tt                    Cooking_code                   Lighting_code
                              0                               6                               0                               0
             Dwelling_unit_code           Regular_salary_earner                Perform_Ceremony Meals_seved_to_non_hhld_members
                              0                               0                               2                             386
            Possess_ration_card             Type_of_ration_card                        MPCE_URP                        MPCE_MRP
                              0                            1566                               0                               0
```



```
    Person_Srl_No              Relation                   Sex                   Age
                0                     0                     0                     0
   Marital_Status             Education      Days_Stayed_away   No_of_Meals_per_day
                0                     0                  3539                     2
     Meals_School        Meals_Employer          Meals_Others         Meals_Payment
             3968                  3939                  3506                  3577
    Meals_At_Home             Item_Code           Source_Code            ricepds_q
               46                     0                    44                     0
         riceos_q           ricetotal_q               chira_q                khoi_q
                0                     0                     0                     0
           muri_q             ricepro_q              riceGT_q           wheatpds_q
                0                     0                     0                     0
        wheatos_q          wheattotal_q               maida_q                suji_q
                0                     0                     0                     0
          sewai_q               bread_q              wheatp_q             wheatGT_q
                0                     0                     0                     0
         jowarp_q              bajrap_q              maizep_q             barleyp_q
                0                     0                     0                     0
        milletp_q               ragip_q            cerealot_q           cerealtot_q
                0                     0                     0                     0
      cerealsub_q           cerealstt_q               arhar_q             grandal_q
                0                     0                     0                     0
     gramwholep_q              gramGT_q               moong_q               masur_q
                0                     0                     0                     0
            urd_q             peasdal_q             khesari_q             otpulse_q
                0                     0                     0                     0
          gramp_q               besan_q              pulsep_q           pulsestot_q
```

**Interpretation:** The missing values summary indicates that most columns have no missing data, ensuring data completeness. Key columns like NIC_2008, NCO_2004, Type_of_land_owned, and Land_Owned have moderate missing values, which may require imputation for accuracy. Columns such as Land_Leased_in, Otherwise_possessed, Land_Leased_out, and Meals_School have substantial missing values, potentially impacting analysis and requiring careful handling.

- Most columns have zero missing values, indicating that the data in these columns is complete.

- **Columns with Few Missing Values**:

  - Perform_Ceremony: 2 missing values
  - Meals_seved_to_non_hhld_members: 386 missing values

- **Columns with Moderate Missing Values**:

  - NIC_2008: 274 missing values
  - NCO_2004: 272 missing values
  - Type_of_land_owned: 383 missing values
  - Land_Owned: 402 missing values

- **Columns with High Missing Values**:

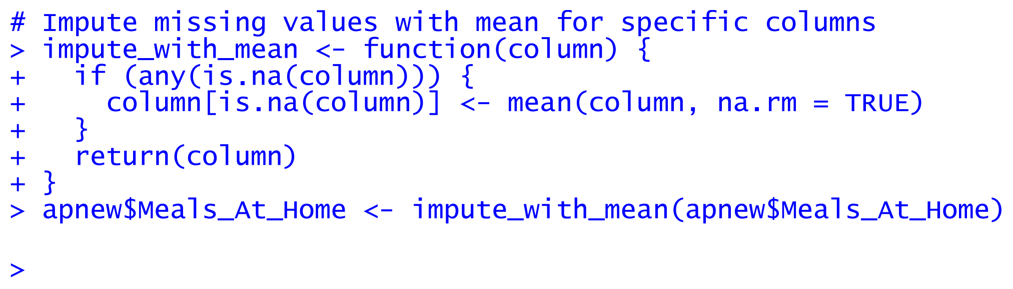
  - Land_Leased_in: 3328 missing values
  - Otherwise_possessed: 3786 missing values
  - Land_Leased_out: 3564 missing values
  - During_July_June_Cultivated: 2064 missing values
  - During_July_June_Irrigated: 3415 missing values
  - Type_of_ration_card: 1566 missing values
  - Days_Stayed_away: 3539 missing values
  - Meals_School: 3968 missing values
  - Meals_Employer: 3939 missing values
  - Meals_Others: 3506 missing values
  - Meals_Payment: 3577 missing values
  - babyfood_q: 4026 missing values

*#Imputing the values, i.e. replacing the missing values with mean.*

**Code :**

```
# 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)
}
[1] FALSE
```

**Result:**



```
# 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)
+ }
> apnew$Meals_At_Home <- impute_with_mean(apnew$Meals_At_Home)

>
```

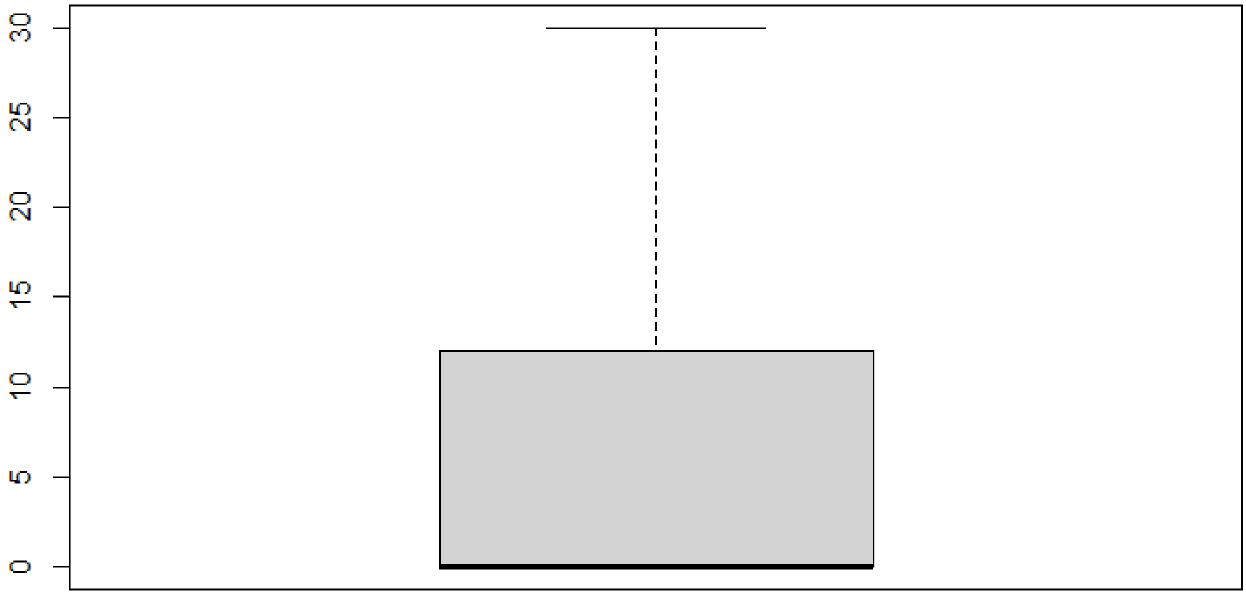
**<u>Interpretation</u>:** The above code has replaced the missing values with the mean value of the variable..

## b) Check for outliers and describe the outcome of your test and make suitable amendments.

Boxplots can be used to find outliers in the dataset. Boxplots visually reveal outliers in a dataset by displaying individual points located beyond the whiskers of the boxplot.

*#Checking for outliers using box plot*

```
> boxplot(apnew$ricepds_v)
```



**Interpretation:** From the boxplot above, which is a visual representation of the variable ‘ricepds_v’ shows that there is an outlier. Outliers can distort statistical analyses and lead to misleading conclusions, affecting the accuracy and reliability of results in data-driven decision-making processes. Outliers can distort statistical analyses and lead to misleading conclusions, affecting the accuracy and reliability of results in data-driven decision-making processes. The outliers can be removed using the following code.

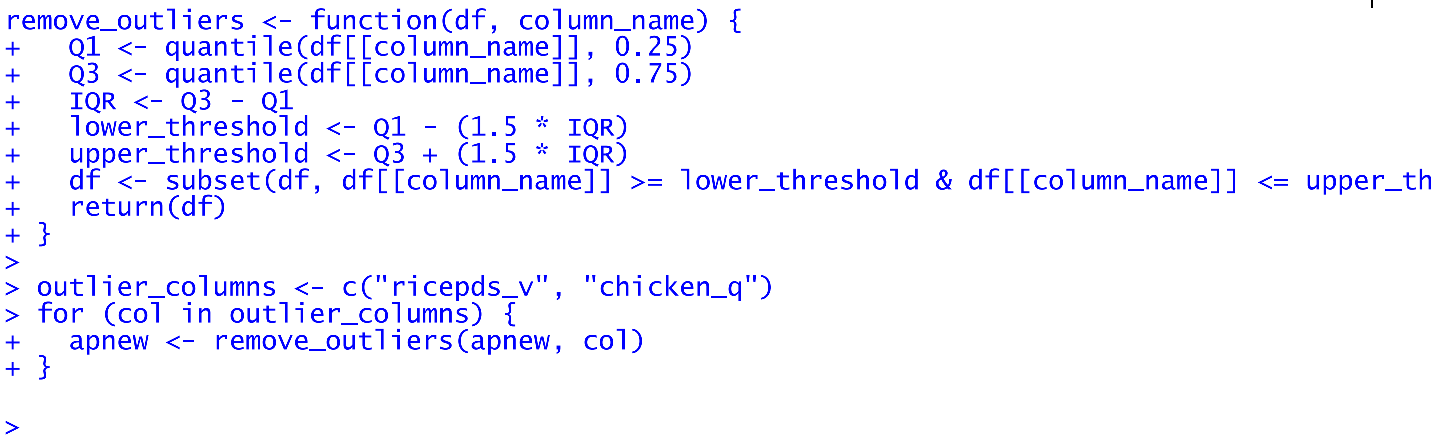
*#Setting quartiles and removing outliers*

**Code :**

```
# 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)
}
```

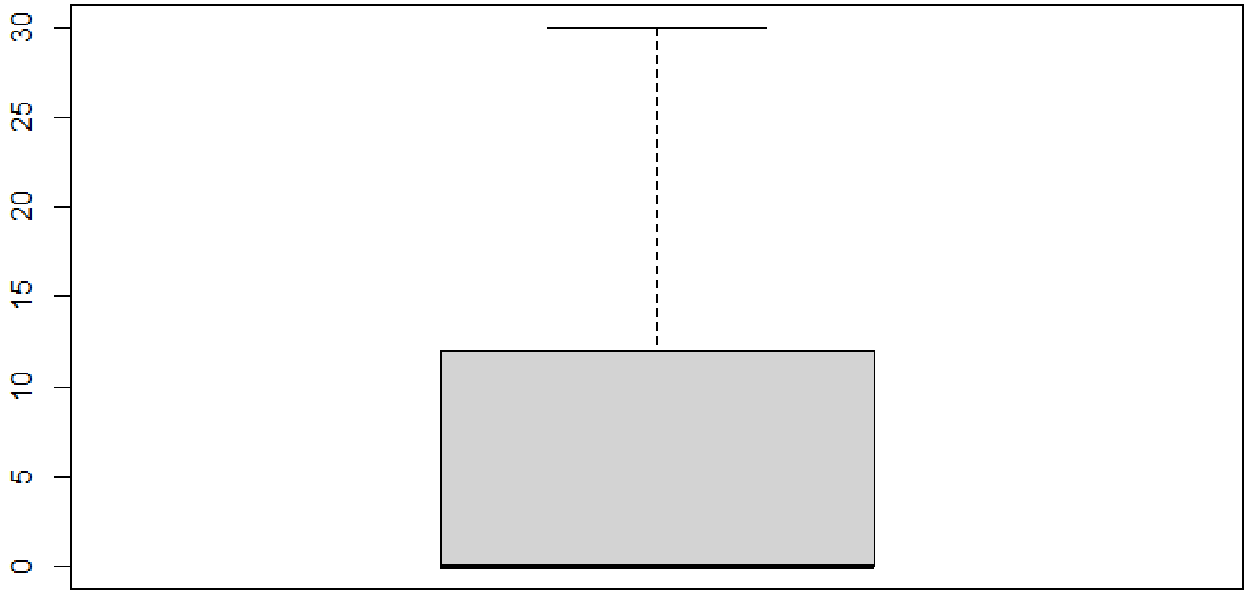
```
outlier_columns <- c("ricepds_v", "chicken_q")
for (col in outlier_columns) {
  apnew <- remove_outliers(apnew, col)
}
```

**Result:**



```
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_th
+   return(df)
+ }
>
> outlier_columns <- c("ricepds_v", "chicken_q")
> for (col in outlier_columns) {
+   apnew <- remove_outliers(apnew, col)
+ }

>
```



**<u>Interpretation</u>:** Interpreting quartile ranges allows for the identification and removal of outliers. The interquartile range (IQR), calculated as the difference between the upper and lower quartiles, helps in pinpointing data points that lie beyond 1.5 times the IQR from either quartile. These outliers can then be excluded or adjusted to improve the robustness of the analysis. This technique can also be used to remove outliers from other variables.

### c) Rename the districts as well as the sector, viz. rural and urban.

Each district in the NSSO data for a state is given a unique number. To determine the top consuming districts, these numbers need to be mapped to their respective names. Likewise, the urban and rural sectors of the state are assigned the numbers 1 and 2, respectively. This can be achieved by running the following code.

**Code and Result:**

```
district_mapping <- c("26" = "Chittoor", "29" = "Rangareddi", "12" = "East Godavari", "7" =
"Visakhapatnam")
sector_mapping <- c("2" = "URBAN", "1" = "RURAL")

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

**Result:**

```
# Rename districts and sectors
> district_mapping <- c("26" = "Chittoor", "29" = "Rangareddi", "12" = "East Goda
vari", "7" = "Visakhapatnam")
> sector_mapping <- c("2" = "URBAN", "1" = "RURAL")
>
> apnew$District <- as.character(apnew$District)
> apnew$Sector <- as.character(apnew$Sector)
> apnew$District <- ifelse(apnew$District %in% names(district_mapping), district_
mapping[apnew$District], apnew$District)
> apnew$Sector <- ifelse(apnew$Sector %in% names(sector_mapping), sector_mapping[
apnew$Sector], apnew$Sector)
```

**<u>Interpretation</u>:** After running this code, the "District" and "Sector" columns in the `apnew` data frame will be updated. Any numeric codes originally present will be replaced with their corresponding human-readable names from the dictionaries. This improves the readability and clarity of your data for further analysis.

### d) Summarizing the critical variables region wise and district wise and we are indicate the top three districts and the bottom three districts of consumption.

By aggregating the critical variables into total consumption, we can identify the top 3 and bottom 3 consuming districts.

**Code :**

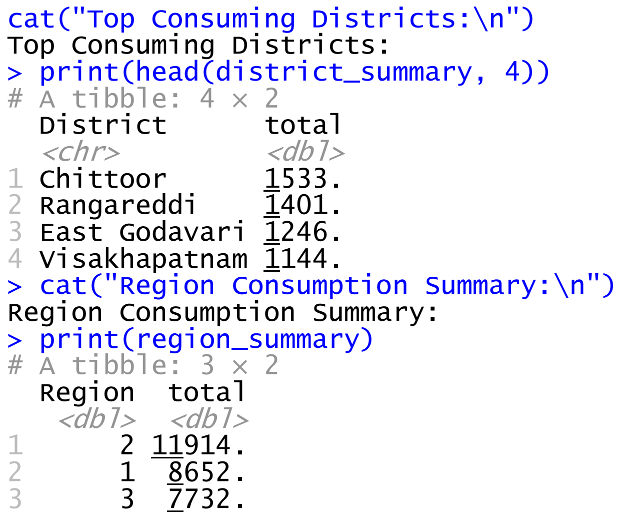
```
 # Summarize consumption
apnew$total_consumption <- rowSums(apnew[, c("ricepds_v", "Wheatpds_q", "chicken_q",
"pulsep_q", "wheatos_q")], na.rm = TRUE)

# Summarize and display top consuming districts and regions
summarize_consumption <- function(group_col) {
  summary <- apnew %>%
    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 Consuming Districts:\n")
print(head(district_summary, 4))
cat("Region Consumption Summary:\n")
print(region_summary)
```

**Result:**



```
cat("Top Consuming Districts:\n")
Top Consuming Districts:
> print(head(district_summary, 4))
# A tibble: 4 × 2
  District      total
  <chr>         <dbl>
1 Chittoor      1533.
2 Rangareddi    1401.
3 East Godavari 1246.
4 Visakhapatnam 1144.
> cat("Region Consumption Summary:\n")
Region Consumption Summary:
> print(region_summary)
# A tibble: 3 × 2
  Region  total
   <dbl>  <dbl>
1      2 11914.
2      1  8652.
3      3  7732.
```

**Interpretation:**

The output shows the "Top Consuming Districts" table with the four districts having the highest total consumption. You can see "Chittoor" has the highest total consumption (1533), followed by "Rangareddi" and so on.

The "Region Consumption Summary" table displays the total consumption for each region (identified by a numeric code). Region 2 has the highest total consumption (11914), followed by region 1 and region 3.

We can identify which districts have the highest overall consumption based on the combined quantities of the five food items. It also provides a summary of the total consumption for each region.

By analyzing these results, you might investigate further to understand the reasons behind the variations in consumption patterns across districts and regions.

**# Test for differences in mean consumption between urban and rural**

**Code :**

```
rural <- apnew %>%

  filter(Sector == "RURAL") %>%

  select(total_consumption)



urban <- apnew %>%

  filter(Sector == "URBAN") %>%

  select(total_consumption)
```

**Result :**

```
# Test for differences in mean consumption between urban and rural
> rural <- apnew %>%
+   filter(Sector == "RURAL") %>%
+   select(total_consumption)
> 
> urban <- apnew %>%
+   filter(Sector == "URBAN") %>%
+   select(total_consumption)
```

**Explaination :**

We will have two separate data frames:

- `rural`: Contains the total consumption values for all entries classified as "RURAL" in the original `apnew` dataset.
- `urban`: Contains the total consumption values for all entries classified as "URBAN" in the original `apnew` dataset.

This separation allows us to perform further analysis on the consumption patterns of rural and urban populations.

### e) Test whether the differences in the means are significant or not.

### Code:

```
z_test_result <- z.test(rural, urban, alternative = "two.sided", mu = 0, sigma.x = 2.56, sigma.y = 2.34,
conf.level = 0.95)


if (z_test_result$p.value < 0.05) {
  cat("P value is <", 0.05, ", Therefore we reject the null hypothesis.\n")
  cat("There is a difference between mean consumptions of urban and rural.\n")
} else {
  cat("P value is >=", 0.05, ", Therefore we fail to reject the null hypothesis.\n")
  cat("There is no significant difference between mean consumptions of urban and rural.\n")
}
```

**Result:**

```
z_test_result <- z.test(rural, urban, alternative = "two.sided", mu = 0, sigma.x
= 2.56, sigma.y = 2.34, conf.level = 0.95)
>
> if (z_test_result$p.value < 0.05) {
+   cat("P value is <", 0.05, ", Therefore we reject the null hypothesis.\n")
+   cat("There is a difference between mean consumptions of urban and rural.\n")
+ } else {
+   cat("P value is >=", 0.05, ", Therefore we fail to reject the null hypothesis
.\n")
+   cat("There is no significant difference between mean consumptions of urban an
d rural.\n")
+ }
P value is < 0.05 , Therefore we reject the null hypothesis.
There is a difference between mean consumptions of urban and rural.
```

**Interpretation:** The two-sample z-test indicates a highly significant difference in consumption between rural and urban sectors . The output shows "P value is < 0.05", which means the p-value from the z-test is less than 0.05. Therefore, we reject the null hypothesis. This suggests there's a statistically significant difference between the mean consumption of urban and rural populations.

# CODES

```
# Set the working directory and verify it
setwd('C:\\Users\\Chand\\Downloads\\Assignment1')
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")
lapply(libraries, install_and_load)

# Reading the file into R
data <- read.csv("C:\\Users\\Chand\\Downloads\\Assignment1\\NSSO68.xlsx")

# Filtering for Mizoram
df <- data %>%
  filter(state_1 == "15")

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

# Finding missing values
missing_info <- colSums(is.na(df))
cat("Missing Values Information:\n")
print(missing_info)

# Subsetting the data
apnew <- 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)

# 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)
}
apnew$Meals_At_Home <- impute_with_mean(apnew$Meals_At_Home)

boxplot(apnew$ricepds_v)

# 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) {
  apnew <- remove_outliers(apnew, col)
}
boxplot(apnew$ricepds_v)

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

# Summarize and display top consuming districts and regions
summarize_consumption <- function(group_col) {
  summary <- apnew %>%
    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 Consuming Districts:\n")
print(head(district_summary, 4))
cat("Region Consumption Summary:\n")
print(region_summary)

# Rename districts and sectors
district_mapping <- c("26" = "Chittoor", "29" = "Rangareddi", "12" = "East Godavari", "7" =
"Visakhapatnam")
sector_mapping <- c("2" = "URBAN", "1" = "RURAL")

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


# Test for differences in mean consumption between urban and rural
rural <- apnew %>%
  filter(Sector == "RURAL") %>%
  select(total_consumption)

urban <- apnew %>%
  filter(Sector == "URBAN") %>%
  select(total_consumption)


z_test_result <- z.test(rural, urban, alternative = "two.sided", mu = 0, sigma.x = 2.56, sigma.y = 2.34,
conf.level = 0.95)

if (z_test_result$p.value < 0.05) {
  cat("P value is <", 0.05, ", Therefore we reject the null hypothesis.\n")
  cat("There is a difference between mean consumptions of urban and rural.\n")
} else {
  cat("P value is >=", 0.05, ", Therefore we fail to reject the null hypothesis.\n")
  cat("There is no significant difference between mean consumptions of urban and rural.\n")
}
```