VIRGINIA COMMONWEALTH UNIVERSITY

STATISTICAL ANALYSIS & MODELING

A1a: CONSUMPTION PATTERN OF MIZORAM USING

PYTHON AND R

CHANDHINI KM

V01107497

Date of Submission: 16/06/2024

# CONTENTS

|  |  |
| --- | --- |
| Content: | Page no: |
| INTRODUCTION | 3 |
| OBJECTIVE | 3 |
| BUSINESS SIGNIFICANC | 3-4 |
| RESULTS AND  INTERPRETATIONS | 4-9 |
| CODES | 10-14 |

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")

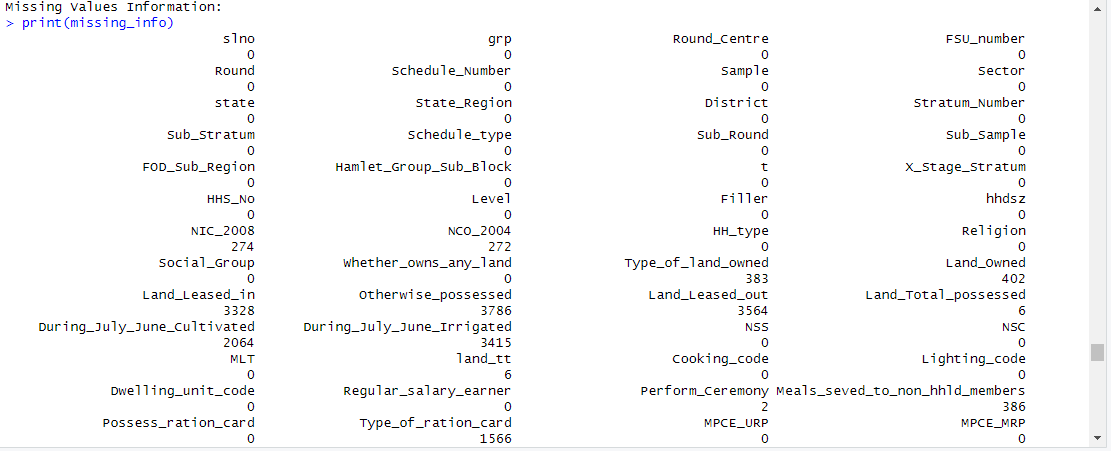
print(missing\_info)

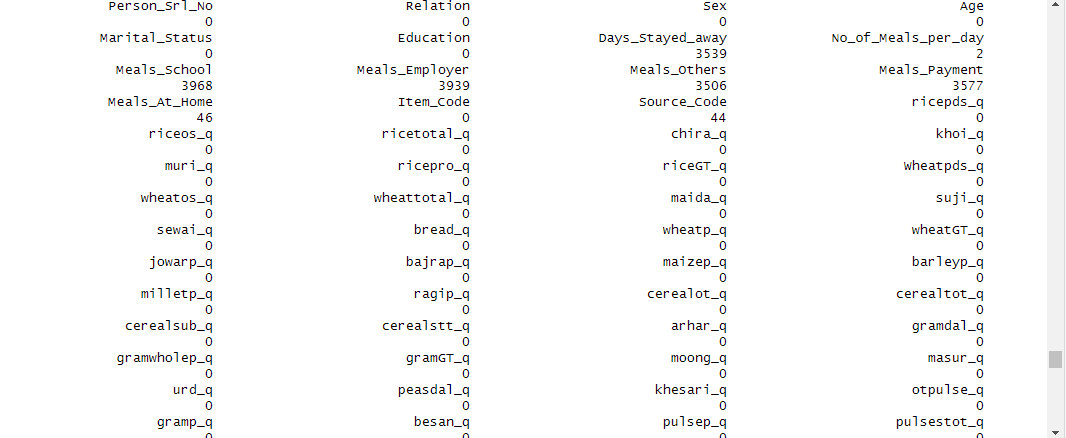
# Subsetting the data

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:**





**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) |
|  |
| |  | | --- | | > | |

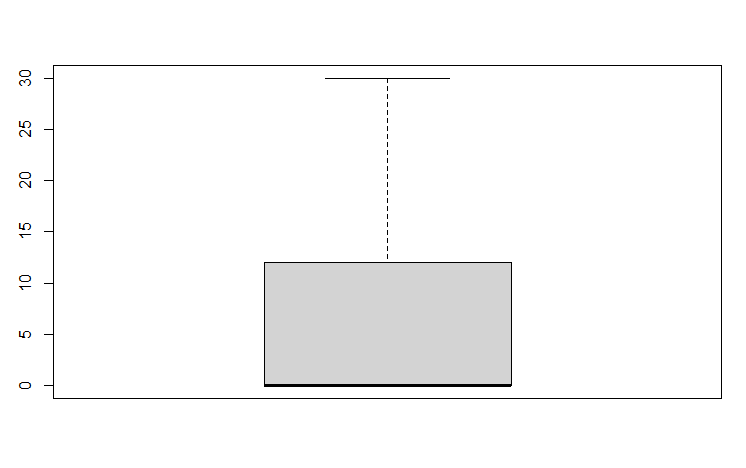
**Interpretation:** The above code has replaced the missing values with the mean value of the variable..

**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")

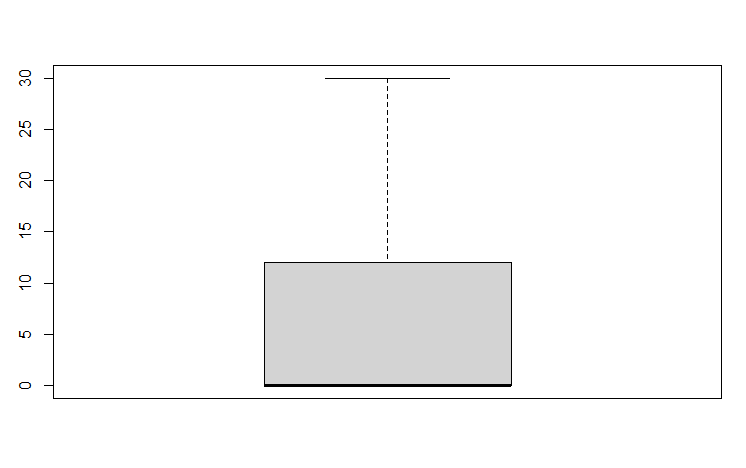
for (col in outlier\_columns) {

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\_threshold)  + return(df)  + }  >  > outlier\_columns <- c("ricepds\_v", "chicken\_q")  > for (col in outlier\_columns) {  + apnew <- remove\_outliers(apnew, col)  + } |
|  |
| |  | | --- | | > | |



**Interpretation:** 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 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)

**Interpretation:** 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 and 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")

}