****

**VIRGINIA COMMONWEALTH UNIVERSITY**

**Statistical analysis and modelling (SCMA 632)**

**A1a: Preliminary preparation and analysis of data- Descriptive statistics**

**SHREYA MISHRA**

**V01108272**

**Date of Submission: 16-06-2024**

**CONTENTS**

|  |  |  |
| --- | --- | --- |
| **Sl. No.** | **Title** | **Page No.** |
| **1.** | Introduction | 1 |
| **2.** | Objectives | **1** |
| **3.** | Business Significance | **1** |
| **4.** | Results and analysis | **2-5** |
| **5.** | Codes | **6-10** |
| **6.** | References |  |

**Analysing Consumption in the State of Tripura Using R**

**Introduction**

The focus of this study is on the state of Tripura, utilizing data from the NSSO to identify the top and bottom three consuming districts. In the process, we manipulate and clean the dataset to obtain the required data for analysis. The dataset includes consumption-related information, covering both rural and urban sectors, as well as district-wise variations. This data has been imported into R, a powerful statistical programming language renowned for its versatility in handling and analyzing large datasets.

Our objectives include identifying missing values, addressing outliers, standardizing district and sector names, summarizing consumption data regionally and district-wise, and testing the significance of mean differences in consumption between rural and urban areas. The findings from this study can inform policymakers and stakeholders, fostering targeted interventions and promoting equitable development across the state of Tripura.

**OBJECTIVES**

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 and describe the outcome of your test and make suitable amendments.

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

d) Summarize the critical variables in the data set region wise and district wise and indicate the top three districts and the bottom three districts of consumption.

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

**BUSINESS SIGNIFICANCE**

The study of Tripura's consumption patterns using NSSO data provides vital insights for businesses and policymakers. Identifying the top and bottom consuming districts helps in:

Market Entry: Target high-consumption districts for launching new products.

Resource Allocation: Optimize distribution and stock levels based on consumption patterns.

Supply Chain Optimization: Enhance logistics efficiency by focusing on high-demand areas.

Targeted Interventions: Policymakers can design strategies to boost consumption in low-performing districts.

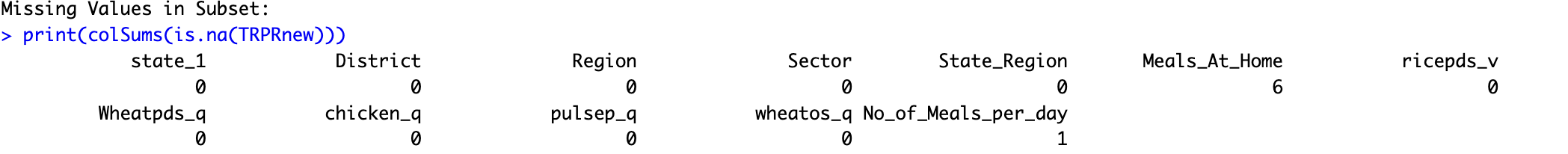
Economic Growth: Attract investments, create jobs, and stimulate economic activities in high-consumption areas.

These findings support informed decision-making, fostering equitable development and promoting economic growth in Tripura

**A)RESULTS AND INTERPRETATION**

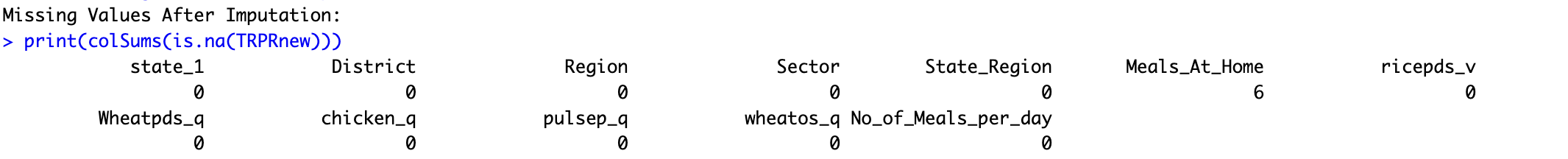
1. Check if there are any missing values in the data, identify them and if there are replace them with the mean of the variable.

#Identifying the missing values.



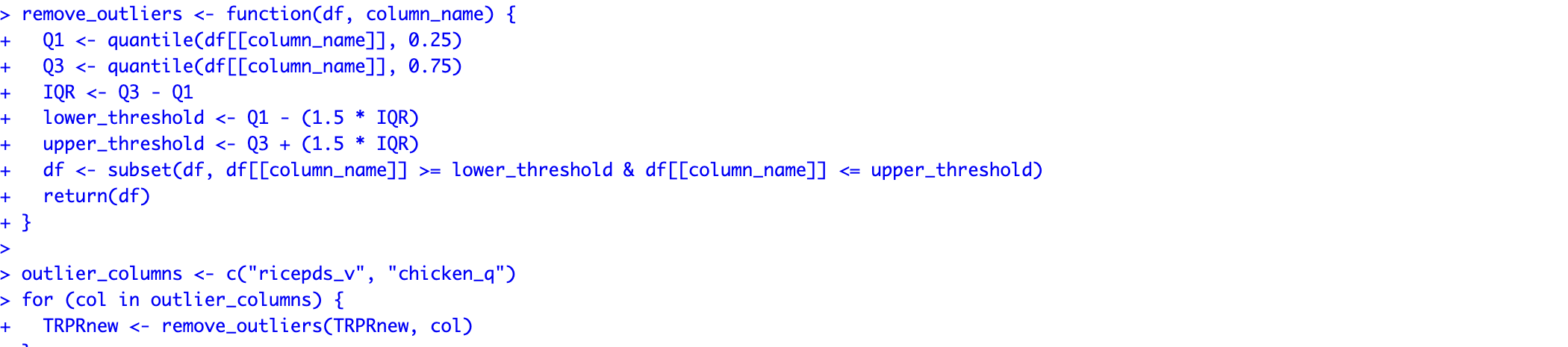
**Interpretation**: From the selected variables, after sorting the data for the state of Tripura, it is seen that only the column No. Of meals per day has 01 missing variable. Since missing values in the dataset can be problematic as they lead to incomplete or biased analyses, hindering the accuracy of results and potentially skewing interpretations and decision-making processes. Therefore we replace the missing values with the mean of the variable using following code.

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



Interpretation: The above code has successfully replaced the missing values with the mean value of the variable. As can be seen from the result above, there are no missing values in the selected data.

**B) Check for outliers and describe the outcome of your test and make suitable amendments.**



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

Each district of a state in the NSSO of data is assigned an individual number. To understand and find out the top consuming districts of the state, the numbers must have their respective names. Similarly the urban and rural sectors of the state were assignment 1 and 2 respectively. This is done by running the following code.

**# Rename districts and sectors , get codes from appendix of NSSO 68th ROund Data**

district\_mapping <- c("01" = "West Tripura", "02" = "South Tripura", "03" = "Dhalai", "04" = "North Tripura")

TRPRnew$District <- as.character(TRPRnew$District)

TRPRnew$Sector <- as.character(TRPRnew$Sector)

TRPRnew$District <- ifelse(TRPRnew$District %in% names(district\_mapping), district\_mapping[TRPRnew$District], TRPRnew$District)

TRPRnew$Sector <- ifelse(TRPRnew$Sector %in% names(sector\_mapping), sector\_mapping[TRPRnew$Sector], TRPRnew$Sector

**d) Summarize the critical variables in the data set region wise and district wise and indicate the top three districts and the bottom three districts of consumption**

summarize\_consumption <- function(group\_col) {

summary <- TRPRnew %>%

group\_by(across(all\_of(group\_col))) %>%

summarise(total = sum(total\_consumption)) %>%

arrange(desc(total))

return(summary)

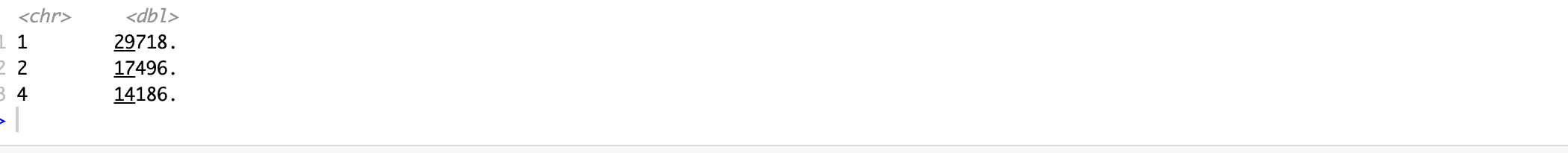
}

district\_summary <- summarize\_consumption("District")

region\_summary <- summarize\_consumption("Region")

cat("Top 3 Consuming Districts:\n")

print(head(district\_summary, 3))

Here we can interpret that, West Tripura is the most consuming district followed by the, South Tripura and North Tripura

cat("Bottom 3 Consuming Districts:\n")

print(tail(district\_summary, 3))  
  
U,{c90d5a79-1df2-44aa-9af2-b4d4563afcd0}{214},3.125,3.125

cat("Region Consumption Summary:\n")

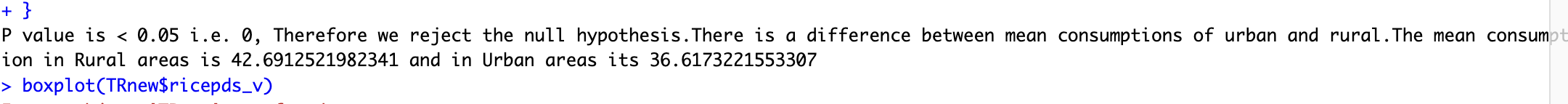
print(region\_summary)

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

The first step to this is to have a Hypotheses Statement.

#H0: There is no difference in consumption between urban and rural.

#H1: There is difference in consumption between urban and rural.



**CODES data <- read.csv('/Users/shreyamishra/Desktop/NSSO68 (3).csv')**

library(dplyr)

library(readr)

library(readxl)

library(tidyr)

library(ggplot2)

library(BSDA)

library(glue)

if (!require(dplyr)) {

install.packages("dplyr")

library(dplyr)

}

df <- data %>%

filter(state\_1 == 'TRPR')

names(df)

head(df)

dim(df)

missing\_info <- colSums(is.na(df))

cat("Missing Values Information:\n")

print(missing\_info)

TRPRnew <- 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)

cat("Missing Values in Subset:\n")

print(colSums(is.na(TRPRnew)))

impute\_with\_mean <- function(column) {

if (any(is.na(column))) {

column[is.na(column)] <- mean(column, na.rm = TRUE)

}

return(column)

}

TRPRnew$No\_of\_Meals\_per\_day <- impute\_with\_mean(TRPRnew$No\_of\_Meals\_per\_day)

cat("Missing Values After Imputation:\n")

print(colSums(is.na(TRPRnew)))

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) {

TRPRnew <- remove\_outliers(TRPRnew, col)

}

TRPRnew$total\_consumption <- rowSums(TRPRnew[, c("ricepds\_v", "Wheatpds\_q", "chicken\_q", "pulsep\_q", "wheatos\_q")], na.rm = TRUE)

summarize\_consumption <- function(group\_col) {

summary <- TRPRnew %>%

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)

district\_mapping <- c("01" = "West Tripura", "02" = "South Tripura", "03" = "Dhalai", "04" = "North Tripura")

sector\_mapping <- c("2" = "URBAN", "1" = "RURAL")

TRPRnew$District <- as.character(TRPRnew$District)

TRPRnew$Sector <- as.character(TRPRnew$Sector)

TRPRnew$District <- ifelse(TRPRnew$District %in% names(district\_mapping), district\_mapping[TRPRnew$District], TRPRnew$District)

TRPRnew$Sector <- ifelse(TRPRnew$Sector %in% names(sector\_mapping), sector\_mapping[TRPRnew$Sector], TRPRnew$Sector)

rural <- TRPRnew %>%

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

select(total\_consumption)

urban <- TRPRnew %>%

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

select(total\_consumption)

mean\_rural <- mean(rural$total\_consumption)

mean\_urban <- mean(urban$total\_consumption)

if (!require(BSDA)) {

install.packages("BSDA")

library(BSDA)

}

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(glue::glue("P value is < 0.05 i.e. {round(z\_test\_result$p.value, 5)}, Therefore we reject the null hypothesis.\n"))

cat(glue::glue("There is a difference between mean consumptions of urban and rural.\n"))

cat(glue::glue("The mean consumption in Rural areas is {mean\_rural} and in Urban areas its {mean\_urban}\n"))

} else {

cat(glue::glue("P value is >= 0.05 i.e. {round(z\_test\_result$p.value, 5)}, Therefore we fail to reject the null hypothesis.\n"))

cat(glue::glue("There is no significant difference between mean consumptions of urban and rural.\n"))

cat(glue::glue("The mean consumption in Rural area is {mean\_rural} and in Urban area its {mean\_urban}\n"))

}

boxplot(TRPRnew$ricepds\_v)

TRPRnew %>%

group\_by(District) %>%

summarise(total = sum(total\_consumption)) %>%

arrange(-total, District)