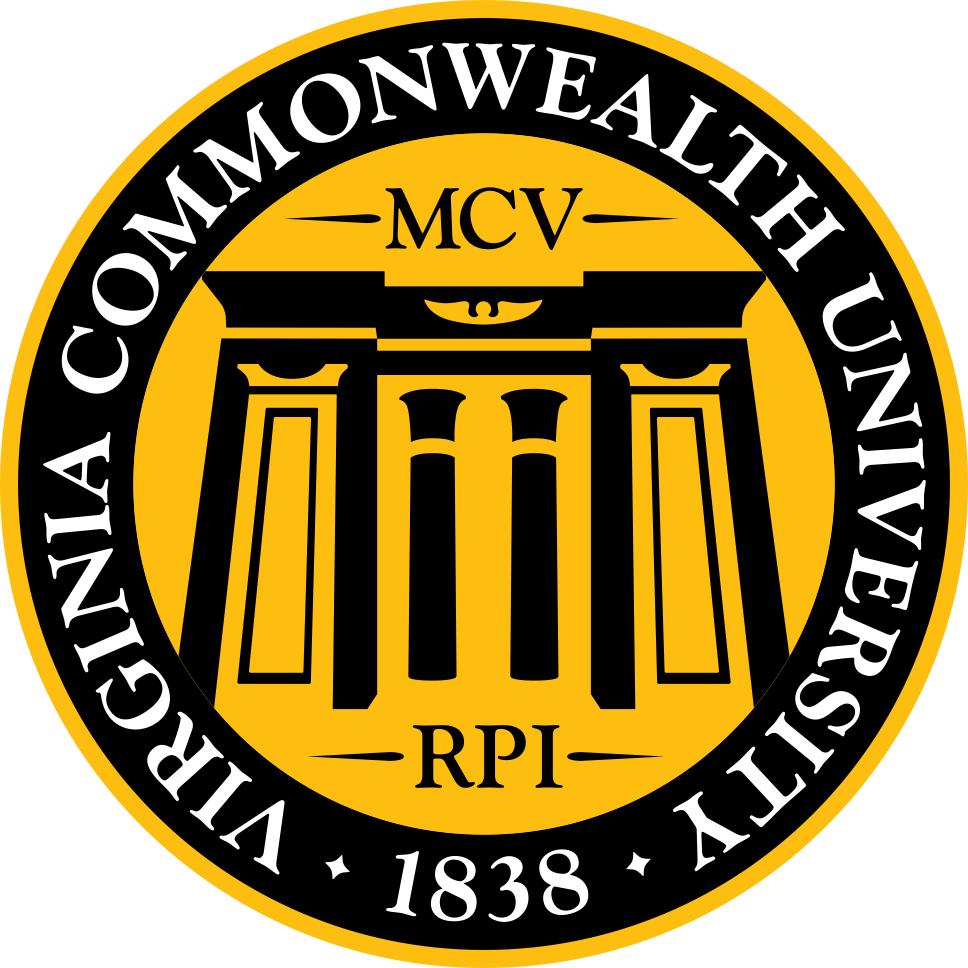
****

**VIRGINIA COMMONWEALTH UNIVERSITY**

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

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

**Nandita Reddy Mantri**

**V01108257**

**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 | **5-10** |
| **6.** | References |  |

**Analyzing Consumption in the State of SIKKIM Using R**

**Introduction**

The focus of this study is on the state of Sikkim from the NSSO data, to find the top and bottom three consuming districts of Sikkim. In the process, we manipulate and clean the dataset to get the required data to analyze. To facilitate this analysis, we have gathered a dataset containing consumption-related information, including data on rural and urban sectors, as well as district-wise variations. The dataset 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. The findings from this study can inform policymakers and stakeholders, fostering targeted interventions and promoting equitable development across the state.

**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 focus of this study on Sikkim'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 Sikkim's economic growth.

**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.

**A close-up of a computer code

Description automatically generated**

**Interpretation**: From the selected variables, after sorting the data for the state of Sikkim, it is seen that there are no missing values.

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

**A computer code with blue text

Description automatically generated**

**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 rural and urban 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("4" = "East", "3" = "South", "2" = "West")**

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

**skmnew$District <- as.character(skmnew$District)**

**skmnew$Sector <- as.character(skmnew$Sector)**

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

**skmnew$Sector <- ifelse(skmnew$Sector %in% names(sector\_mapping), sector\_mapping[skmnew$Sector], skmnew$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 and display top and bottom consuming districts and regions**

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

**+ summary <- skmnew %>%**

**+ .... [TRUNCATED]**

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

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

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

**Top 3 Consuming Districts:**

**> print(head(district\_summary, 3))**

**A black screen with white text

Description automatically generated**

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

**print(tail(district\_summary, 3))**

A black screen with white text

Description automatically generated

**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.

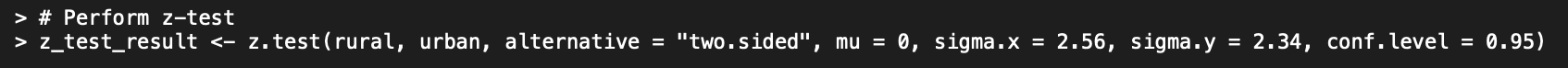
#H1: There is difference in consumption between urban and rural.

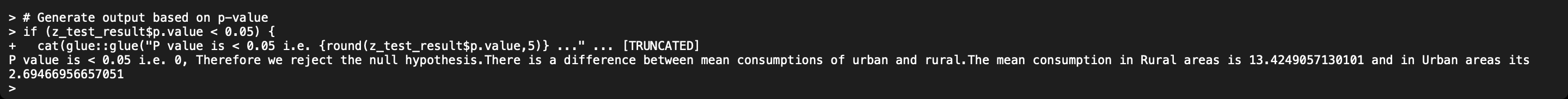
mean\_rural <- mean(rural$total\_consumption)

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





****

****

**CODES**

# Set the working directory and verify it

setwd('/Users/nanditareddy/Desktop/R BOOTCAMP/ASSIGNMENT A1')

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")

lapply(libraries, install\_and\_load)

# Reading the file into R

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

# Filtering for SKM

df <- data %>%

filter(state\_1 == "SKM")

# 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)

# Sub-setting the data

skmnew <- df %>%

select(state\_1, District, Region, Sector, State\_Region, Meals\_Employer, Meals\_Payment, 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(skmnew)))

# 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)

}

skmnew$Meals\_At\_Home <- impute\_with\_mean(skmnew$Meals\_At\_Home)

skmnew$Meals\_Employer <- impute\_with\_mean(skmnew$Meals\_Employer)

skmnew$Meals\_Payment <- impute\_with\_mean(skmnew$Meals\_Payment)

# Check for missing values after imputation

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

print(colSums(is.na(skmnew)))

# 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) {

skmnew <- remove\_outliers(skmnew, col)

}

view (remove\_outliers)

# Summarize consumption

skmnew$total\_consumption <- rowSums(skmnew[, 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 <- skmnew %>%

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("4" = "East", "3" = "South", "2" = "West")

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

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

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

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

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

# Test for differences in mean consumption between urban and rural

rural <- skmnew %>%

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

select(total\_consumption)

urban <- skmnew %>%

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

select(total\_consumption)

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

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

# Perform z-test

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

# Generate output based on p-value

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"))

}

# Summarize and display top and bottom consuming districts and regions

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

summary <- skmnew %>%

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")