# VIRGINIA COMMONWEALTH UNIVERSITY

STATISTICAL ANALYSIS & MODELING

A1a: CONSUMPTION PATTERN OF HIMACHAL USING PYTHON AND R

ROHAN ARPIT DUNGDUNG

V01106712

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 Himachal Pradesh Using R

# INTRODUCTION

The focus of this study is on the state of Himachal Pradesh, from the NSSO data, to find the top and bottom three consuming districts of Himachal Pradesh. In the process, we manipulate and clean the dataset to get the required data to analyze. For this analysis, we have gathered a dataset containing consumption-related information, including data on rural and urban sectors.

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

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.
2. Check for outliers and describe the outcome of your test and make suitable amendments.
3. Rename the districts as well as the sector, viz. rural and urban.
4. 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.
5. Test whether the differences in the means are significant or not.

# BUSINESS SIGNIFICANCE

The focus of this study on Himachal Pradesh'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 Himachal Pradesh's economic growth.

# INTERPRETATION

MISSING VALUES

# Finding missing values

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

cat("Missing Values Information:\n")

print(missing\_info)

# 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)

}

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

hpnew$Days\_Stayed\_away <- impute\_with\_mean(hpnew$Days\_Stayed\_away)

hpnew$Meals\_School <- impute\_with\_mean(hpnew$Meals\_School)

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

hpnew$Meals\_Others <- impute\_with\_mean(hpnew$Meals\_Others)

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

hpnew$Source\_Code <- impute\_with\_mean(hpnew$Source\_Code)

hpnew$soyabean\_q <- impute\_with\_mean(hpnew$soyabean\_q)

#### Interpretation:

The dataset was checked for missing values, and they were identified across various columns. These missing values were then replaced with the mean of the respective variable. Imputing missing values with the mean is a common technique to maintain the dataset's integrity while ensuring that analyses can proceed without interruptions due to missing data.

Outliers:

# 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)

}

# Remove outliers in the dataset

outlier\_columns <- c("ricepds\_v", "chicken\_q","Meals\_At\_Home","Wheatpds\_q","pulsep\_q","wheatos\_q","No\_of\_Meals\_per\_day")

for (col in outlier\_columns) {

hpnew <- remove\_outliers(hpnew, col)

}

#### Interpretation:

Outliers were identified and removed from the dataset. Outliers can significantly skew the results of statistical analyses, leading to inaccurate conclusions. By removing these extreme values, the analysis becomes more robust and reflective of the central tendency of the data.

Rename Districts and Sectors:

# Rename districts and sectors

district\_mapping <- c("2" = "Kangra", "5" = "Mandi", "11" = "Shimla", "6" = "Hamirpur")

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

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

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

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

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

#### Interpretation:

The districts and sectors were renamed for clarity and ease of interpretation. This step makes the dataset more readable and user-friendly, helping to better understand the geographic and demographic distribution of the data.

Summarize Critical Variables:

# Summarize consumption

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

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)

#### Interpretation:

The data was summarized to identify the top and bottom districts and regions based on total food consumption. This summary helps pinpoint areas with the highest and lowest food consumption, providing valuable insights for targeted policy interventions or resource allocation.

Test for Differences in Means:

# Test for differences in mean consumption between urban and rural

rural <- hpnew %>%

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

select(total\_consumption)

urban <- hpnew %>%

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)

summary(z\_test\_result)

z\_test\_result$statistic

z\_test\_result$p.value

# 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"))

}

#### Interpretation:

A z-test was performed to test whether there are significant differences between the mean consumptions of urban and rural areas. The result indicates whether the mean consumptions are statistically significantly different. If the p-value is less than 0.05, we reject the null hypothesis, indicating a significant difference. If it is greater than or equal to 0.05, we fail to reject the null hypothesis, indicating no significant difference.

# CODES

# Set the working directory and verify it

setwd('C:\\Users\\HP\\Downloads')

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("NSSO68.csv")

# Filtering for HP

df <- data %>%

filter(state\_1 == "HP")

#checking the filter

unique(df$state\_1)

# 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 for variable of interest

hpnew <- 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)

unique(apnew$Meals\_At\_Home)

#chcek for missing values in the subset

cat("Missing values in subset:\n")

print(colSums(is.na(hpnew)))

# 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)

}

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

hpnew$Days\_Stayed\_away <- impute\_with\_mean(apnew$Days\_Stayed\_away)

hpnew$Meals\_School <- impute\_with\_mean(apnew$Meals\_School)

hpnew$Meals\_Employer <- impute\_with\_mean(apnew$Meals\_Employer)

hpnew$Meals\_Others <- impute\_with\_mean(apnew$Meals\_Others)

hpnew$Meals\_Payment <- impute\_with\_mean(apnew$Meals\_Payment)

hpnew$Source\_Code <- impute\_with\_mean(apnew$Source\_Code)

hpnew$soyabean\_q <- impute\_with\_mean(apnew$soyabean\_q)

# 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)

}

names(hpnew)

#Remove outliers in the dataset

outlier\_columns <- c("ricepds\_v", "chicken\_q","Meals\_At\_Home","Wheatpds\_q","pulsep\_q","wheatos\_q","No\_of\_Meals\_per\_day")

for (col in outlier\_columns) {

hpnew <- remove\_outliers(apnew, col)

}

# 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("2" = "Kangra", "5" = "Mandi", "11" = "Shimla", "6" = "Hamirpur")

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)

fix(apnew)

# 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)

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)

summary(z\_test\_result)

z\_test\_result$statistic

z\_test\_result$p.value

# 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"))

}