****

**VIRGINIA COMMONWEALTH UNIVERSITY**

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

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

**Manoranjan Mohankumar**

**V01107254**

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

**CONTENTS**

|  |  |  |
| --- | --- | --- |
| **Sl. No.** | **Title** | **Page No.** |
| **1.** | Introduction | **1** |
| **2.** | Results | **2** |
| **3.** | Interpretations | **4** |
| **4.** | Recommendations | **6** |
| **5.** | Codes | **8** |
| **6.** | References | **13** |

**Introduction**

This report presents a comprehensive analysis of household consumption data from the National Sample Survey Office (NSSO) 68th round for Uttar Pradesh (UP). The NSSO conducts surveys to collect data on various socio-economic parameters, and this particular survey focuses on household consumption patterns across different regions and sectors in UP.

**The primary objectives of this report are to:**

* Summarize Consumption Patterns: Provide an overview of household consumption across different districts and regions within UP.
* Identify Top-Consuming Districts and Regions: Highlight the areas with the highest consumption levels to understand geographic disparities.
* Analyse Urban vs. Rural Consumption: Examine the differences in consumption patterns between urban and rural sectors.
* Data Cleaning and Preparation: Address missing values, outliers, and other data quality issues to ensure accurate analysis.
* Statistical Testing: Conduct statistical tests to determine significant differences in consumption patterns.

The data cleaning process involves filtering the dataset to include only records from UP, handling missing values, and removing outliers. Subsequently, the cleaned data is used to identify top-consuming districts and regions, and a z-test is conducted to compare the mean consumption between urban and rural sectors.

This analysis provides valuable insights into the consumption behaviour of households in UP, which can inform policy-making and resource allocation to address regional disparities and improve living standards.

**Results**

**Dataset Information After Filtering for UP**After filtering for the state of Uttar Pradesh, the dataset contains 9015 rows and 384 columns.

**Missing Values Information**The dataset has missing values in several columns. Below are the columns with notable missing values:

* NIC\_2008: 546 missing values
* NCO\_2004: 535 missing values
* Land\_Owned: 615 missing values
* Land\_Leased\_in: 7704 missing values
* Otherwise\_possessed: 8829 missing values
* Land\_Leased\_out: 8526 missing values
* During\_July\_June\_Cultivated: 4379 missing values
* During\_July\_June\_Irrigated: 4464 missing values

**Subset Data Before Imputation**The subset data includes the following columns:

* 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, ricetotal\_q, wheattotal\_q, sugartotal\_q.

**Data After Imputation**Missing values in relevant columns were imputed with the mean of the respective columns. The imputed columns include:

* Meals\_At\_Home, ricepds\_v, Wheatpds\_q, chicken\_q, pulsep\_q, wheatos\_q, No\_of\_Meals\_per\_day, ricetotal\_q, wheattotal\_q, sugartotal\_q.

Data After Outlier Removal  
Outliers in the columns ricetotal\_q, wheattotal\_q, and sugartotal\_q were removed based on the IQR method.

**Top Consuming Districts**The top consuming districts in UP are:

|  |  |
| --- | --- |
| **District** | **Total Consumption** |
| Moradabad | 2215 |
| Gorakhpur | 2192 |
| Allahabad | 2171 |
| Kanpur | 2041 |

Region Consumption Summary

|  |  |
| --- | --- |
| **Region** | **Total Consumption** |
| 3 | 40180 |
| 5 | 21884 |
| 2 | 14313 |
| 1 | 12527 |
| 4 | 6145 |

**Renaming Districts**  
The districts were renamed as follows:

* District 4: Moradabad
* District 67: Gorakhpur
* District 58: Allahabad
* District 45: Kanpur

**Test for Differences in Mean Consumption Between Urban and Rural**

A z-test was conducted to compare the mean consumption between urban and rural sectors.

Result: P value is < 0.05. Therefore, we reject the null hypothesis. There is a difference between urban and rural mean consumption.

**Interpretations**

The analysis of the NSSO 68th round household consumption data for Uttar Pradesh reveals several key insights into the consumption patterns across different regions and sectors within the state.

**Consumption Variation**  
The data indicates a significant variation in total consumption across different districts and regions in Uttar Pradesh. The top-consuming districts identified are Moradabad (formerly District 4), Gorakhpur (formerly District 67), Allahabad (formerly District 58), and Kanpur (formerly District 45). Moradabad shows the highest total consumption, followed closely by Gorakhpur, Allahabad, and Kanpur. This variation in consumption can be attributed to several factors, including population density, economic activities, and the availability of resources in these regions.

Region-wise analysis shows that Region 3 has the highest total consumption, indicating that this region has a higher overall demand for goods and services than other regions. This could be due to higher income levels, better infrastructure, and greater access to markets and services.

**Urban vs. Rural Consumption**  
A significant finding from the data is the difference in consumption patterns between urban and rural sectors. Urban areas exhibit higher mean consumption levels compared to rural areas. This disparity can be attributed to several factors:

* Income Levels: Urban households generally have higher income levels than rural households, allowing for higher spending on goods and services.
* Access to Markets: Urban areas have better access to markets and a wider variety of goods and services, contributing to higher consumption.
* Lifestyle Differences: Urban lifestyles often demand higher spending on amenities and services, such as transportation, education, healthcare, and entertainment.
* Economic Activities: Urban areas are typically hubs of economic activities, providing more employment opportunities and higher wages, leading to increased consumption.

The z-test comparing the mean consumption between urban and rural sectors confirms a statistically significant difference in consumption patterns. The p-value obtained from the z-test is less than 0.05, rejecting the null hypothesis and confirming that the mean consumption in urban areas is significantly different from that in rural areas.

Implications

These findings have important implications for policymaking and resource allocation. The significant variation in consumption across districts and regions suggests the need for targeted interventions to address regional disparities. Policy measures could include improving infrastructure, enhancing market access, and promoting economic activities in lower-consuming areas to boost consumption levels.

The difference in consumption patterns between urban and rural areas highlights the need for tailored strategies to address rural households' unique challenges. Improving income levels through job creation, enhancing market access, and providing better infrastructure and services can help bridge the consumption gap between urban and rural areas.

Overall, this analysis provides valuable insights into the consumption behaviour of households in Uttar Pradesh, which can inform the development of effective policies and programs to enhance living standards and promote balanced regional development.

**Recommendations**

**Policy Interventions**

1. Addressing Urban-Rural Disparities:

* Implement policies to improve income levels in rural areas through job creation, vocational training, and support for small and medium enterprises (SMEs).
* Enhance access to markets for rural households by improving transportation infrastructure and developing rural markets.
* Promote rural development programs to improve education, healthcare, and other essential services.

1. Boosting Economic Activities in Lower-Consuming Areas:

* Develop economic zones or industrial parks in lower-consuming regions to attract investment and create employment opportunities.
* Support agricultural development and agribusiness in rural areas to enhance productivity and income levels.
* Implement financial inclusion initiatives to provide rural households with easier access to credit and banking services.

**Resource Allocation**

1. Optimizing Resource Distribution:

* Allocate resources based on detailed consumption data to ensure that regions with lower consumption receive adequate support.
* Implement food security programs targeting low-consumption areas to meet basic needs.
* Provide subsidies or support for essential goods and services in regions with lower consumption to improve living standards.

1. Infrastructure Development:

* Invest in infrastructure projects such as roads, schools, hospitals, and sanitation facilities in lower-consuming regions to improve overall quality of life.
* Enhance connectivity between rural and urban areas to facilitate better access to markets and services.

**Monitoring and Evaluation**

1. Establishing Monitoring Frameworks:

* Develop robust monitoring frameworks to track the effectiveness of policy interventions and resource allocation strategies.
* Use real-time data collection methods to assess consumption patterns and make data-driven decisions continuously.

1. Engaging Stakeholders:

* Involve local communities, government agencies, non-governmental organizations (NGOs), and private sector partners in designing and implementing interventions.
* Ensure transparency and accountability in executing policies and programs to gain public trust and support.

By implementing these recommendations, policymakers can address regional and sectoral disparities, improve resource allocation, and promote balanced economic development across Uttar Pradesh.

**Codes**

# Step 1: Set the working directory and verify it

setwd('E:\\VCU Bridge course\\R-WD')

print(getwd())

# Step 2: 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)

}

}

# Step 3: Load required libraries

libraries <- c("dplyr", "readr", "readxl", "tidyr", "ggplot2", "BSDA")

lapply(libraries, install\_and\_load)

# Step 4: Reading the file into R

data <- read.csv("E:/VCU Bridge course/R-WD/Assignments/A1a/NSSO68.csv")

# Step 5: Filtering for UP

df <- data %>%

filter(state\_1 == "UP")

# Display dataset info

cat("Dataset Information After Filtering for UP:\n")

print(names(df))

print(head(df))

print(dim(df))

# Step 7: Finding missing values

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

cat("Missing Values Information After Filtering for UP:\n")

print(missing\_info)

# Step 8: Subsetting the data

upnew <- 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, ricetotal\_q, wheattotal\_q, sugartotal\_q)

# Check the subset data

cat("Subset Data Before Imputation:\n")

print(head(upnew))

# Step 9: 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)

}

# Apply the imputation function to all relevant columns

upnew <- upnew %>%

mutate(across(c(Meals\_At\_Home, ricepds\_v, Wheatpds\_q, chicken\_q, pulsep\_q, wheatos\_q, No\_of\_Meals\_per\_day, ricetotal\_q, wheattotal\_q, sugartotal\_q), impute\_with\_mean))

# Check the data after imputation

cat("Data After Imputation:\n")

print(head(upnew))

# Step 10: 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("ricetotal\_q", "wheattotal\_q", "sugartotal\_q")

for (col in outlier\_columns) {

upnew <- remove\_outliers(upnew, col)

}

# Check the data after outlier removal

cat("Data After Outlier Removal:\n")

print(head(upnew))

# Step 11: Summarize consumption

upnew$total\_consumption <- rowSums(upnew[, c("ricetotal\_q", "wheattotal\_q", "sugartotal\_q")], na.rm = TRUE)

# Check the data after adding total\_consumption

cat("Data After Adding Total Consumption:\n")

print(head(upnew))

# Step 12: Summarize and display top consuming districts and regions

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

summary <- upnew %>%

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

summarise(total = sum(total\_consumption, na.rm = TRUE)) %>%

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)

# Step 13: Rename districts and sectors

district\_mapping <- c("04" = "Moradabad", "67" = "Gorakhpur", "58" = "Allahabad", "45" = "Kanpur")

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

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

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

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

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

# Step 14: Test for differences in mean consumption between urban and rural

rural <- upnew %>%

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

select(total\_consumption) %>%

unlist()

urban <- upnew %>%

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

select(total\_consumption) %>%

unlist()

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")

}

**References**

* NSSO 68th round survey data
* ChatGPT