# **VIRGINIA COMMONWEALTH UNIVERSITY**

## **STATISTICAL ANALYSIS & MODELING**

**A1a: CONSUMPTION PATTERN OF UTTARANCHAL USING**

**R**

## **MARIA MARTIN**

**V01107759**

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

## **CONTENTS**

|  |  |
| --- | --- |
| **Content:** | **Page no:** |
| INTRODUCTION | 3 |
| OBJECTIVE | 3 |
| BUSINESS SIGNIFICANCE | 3-4 |
| RESULTS AND INTERPRETATIONS | 5-12 |

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

# **INTRODUCTION**

The focus of this study is on the state of Uttaranchal, from the NSSO data, to find the top and bottom three consuming districts of Uttaranchal. This dataset provides comprehensive information on household consumption patterns in many districts of Uttaranchal, India. The data includes details on the consumption of various food items such as rice, wheat, chicken, pulses, and other essential commodities, categorized by regions, sectors, and meal frequency. The dataset serves as a critical source for understanding the dietary habits and nutritional intake of households in this region, which is pivotal for formulating targeted interventions and policies.

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-wide variations. The dataset has been imported into R, a powerful statistical programming language renowned for its versatility in handling and analyzing large datasets.

# **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 Uttaranchal’s consumption patterns from NSSO data holds significant implications for businesses and policymakers.

1. Policy Makers and Government Agencies can identify nutritional deficiencies and excesses in specific regions or sectors by examining the consumption patterns. This insight aids in designing effective food security programs, public health interventions, and targeted nutritional assistance to improve the overall health of the population.

# **RESULTS AND INTERPRETATION**

### 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.**

Code and Result:

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

Missing Values in Subset:

> print(colSums(is.na(uttnew)))

state\_1 District Region Sector

0 0 0 0

State\_Region Meals\_At\_Home ricepds\_v Wheatpds\_q

0 0 0 0

chicken\_q pulsep\_q wheatos\_q No\_of\_Meals\_per\_day

0 0 0 0

>

Interpretation: The provided R code checks for missing values in the dataframe uttnew and outputs the count of missing values for each column. The code first prints a label "Missing Values in Subset:" to introduce the results. It then calculates and prints the number of missing values in each column using the colSums(is.na(uttnew)) function. The results show that all listed 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) have zero missing values. This indicates that the dataset is complete with respect to these columns, meaning there are no entries with missing data, and the dataset is ready for further analysis without the need for handling missing values in these columns.

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

Code and 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)

+ }

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

>

Interpretation: The above code is a snippet from python programming. It 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 and that’s why the outcome is false.

### 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 assigned 1 and 2 respectively. This is done by running the following code.

In the below code, we took a subset of Uttaranchal district and tried to map the names and sectors instead of the code numbers.

Code and Result:

|  |
| --- |
| > # Rename districts and sectors , get codes from appendix of NSSO 68th ROund Data  > district\_mapping <- c("05" = "Dehradun", "13" = "Hardwar", "06" = "Garhwal")  > sector\_mapping <- c("2" = "URBAN", "1" = "RURAL")  >  > uttnew$District <- as.character(uttnew$District)  > uttnew$Sector <- as.character(uttnew$Sector)  > uttnew$District <- ifelse(uttnew$District %in% names(district\_mapping), district\_mapping[uttnew$District], uttnew$District)  > uttnew$Sector <- ifelse(uttnew$Sector %in% names(sector\_mapping), sector\_mapping[uttnew$Sector], uttnew$Sector)  > |
|  |
| |  | | --- | |  | |

Interpretation: The provided R code transforms the District and Sector columns in the dataframe uttnew by replacing numeric codes with their corresponding descriptive names. This is done using predefined mappings stored in the vectors district\_mapping and sector\_mapping. The district\_mapping vector translates the district codes "05", "13", and "06" to "Dehradun", "Hardwar", and "Garhwal", respectively. Similarly, the sector\_mapping vector converts the sector codes "2" and "1" to "URBAN" and "RURAL".

To achieve this, the code first converts the District and Sector columns to character type to ensure proper handling of the mappings. Then, it uses the ifelse function to check if each entry in these columns matches any of the codes in the mapping vectors. If a match is found, the corresponding descriptive name from the mapping vector replaces the code. If no match is found, the original code is retained. This process enhances the readability and interpretability of the dataset by substituting obscure numeric codes with clear and meaningful labels, thus facilitating better understanding and analysis of the data.

1. **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**. By summarizing the critical variables as total consumption we can estimate the top 3 and bottom 3 consuming districts.

Code and Result:

|  |
| --- |
| 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("05" = "Dehradun", "13" = "Hardwar", "06" = "Garhwal")  sector\_mapping <- c("2" = "URBAN", "1" = "RURAL")  uttnew$District <- as.character(uttnew$District)  uttnew$Sector <- as.character(uttnew$Sector)  uttnew$District <- ifelse(uttnew$District %in% names(district\_mapping), district\_mapping[uttnew$District], uttnew$District)  uttnew$Sector <- ifelse(uttnew$Sector %in% names(sector\_mapping), sector\_mapping[uttnew$Sector], uttnew$Sector) |
|  |
| |  | | --- | |  | |

Interpretation:

The provided R code performs several important operations on the uttnew dataframe, enhancing its readability and providing insights into consumption patterns. Initially, the code prints summaries of consumption data. It identifies and displays the top three and bottom three consuming districts from the district\_summary dataframe. Additionally, it provides a comprehensive summary of regional consumption through the region\_summary dataframe.

Following the summary outputs, the code aims to improve the interpretability of the dataset by renaming district and sector codes to their descriptive names, based on mappings derived from the NSSO 68th Round Data appendix. It defines two mapping vectors: district\_mapping, which maps district codes like "05", "13", and "06" to "Dehradun", "Hardwar", and "Garhwal", respectively, and sector\_mapping, which translates sector codes "2" and "1" to "URBAN" and "RURAL".

To apply these mappings, the code first converts the District and Sector columns in the uttnew dataframe to character type. It then uses the ifelse function to replace the codes in these columns with their corresponding names from the mapping vectors. If a code does not have a corresponding name in the mapping vectors, it retains the original code. This transformation ensures that the dataset is more understandable, replacing obscure numeric codes with meaningful labels. As a result, the dataset becomes easier to interpret and analyze, facilitating better insights and decision-making.

Top of Form

Bottom of Form

### Test whether the differences in the means are significant or not.

### # Test for differences in mean consumption between urban and rural

### rural <- uttnew %>%

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

### select(total\_consumption)

### urban <- uttnew %>%

### 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"))

### }

Interpretation:

The analysis conducted aimed to investigate potential differences in mean consumption between urban and rural areas. The data were segregated into rural and urban sectors, focusing on total consumption values. The mean consumption in rural areas was calculated to be [mean\_rural], while in urban areas, it was [mean\_urban].

A z-test was then performed to statistically compare these means, with a significance level of 0.05. The null hypothesis stated no difference in mean consumption between urban and rural areas. The z-test resulted in a p-value of [round(z\_test\_result$p.value,5)], which is [less than/greater than] 0.05. Based on this result, we [reject/fail to reject] the null hypothesis.

Consequently, we infer that there is [a significant difference/no significant difference] between mean consumptions in urban and rural areas. If the null hypothesis is rejected, it suggests that the mean consumption values are significantly different. Conversely, if we fail to reject the null hypothesis, it implies that there is no significant distinction between the mean consumptions of urban and rural areas.

In conclusion, the analysis provides evidence regarding the comparative mean consumptions, highlighting whether urban and rural areas exhibit meaningful differences in this context.