VIRGINIA COMMONWEALTH UNIVERSITY

Statistical analysis and modelling (SCMA 632)

A1a: Preliminary preparation and analysis of data- Descriptive statistics

ANJELINA SAJAN

V01107258

Date of Submission: 16-06-2024

CONTENTS

INTRODUCTION

This study focuses on Maharashtra, utilizing NSSO data to identify districts with the highest and lowest consumption levels. The dataset provides comprehensive insights into household consumption patterns across diverse districts of Maharashtra, encompassing staple food items such as rice, wheat, poultry, pulses, and other essential commodities. It categorizes data by geographical regions, sectors, and meal frequencies, serving as a critical resource for understanding dietary behaviors and nutritional intake among households in the region. This understanding is pivotal for formulating targeted interventions and policy initiatives.

To facilitate the analysis, the dataset was processed and refined to extract relevant information. Our dataset includes consumption-related data from rural and urban sectors, as well as district-specific variations. This dataset has been imported into R, a powerful statistical programming language well-regarded for its capabilities in managing and analyzing large datasets effectively.

OBJECTIVES

a) Detect any missing values in the dataset and replace them with the mean of the respective variable to ensure data completeness and accuracy.
b) Conduct statistical tests to identify outliers in the consumption data and implement appropriate measures to mitigate their impact, ensuring a more robust analysis.
c) Rename districts and sectors (rural and urban) for consistency and clarity, facilitating a more straightforward analysis and interpretation of the data.
d) Analyze and summarize key consumption variables across different regions and districts, highlighting variations and identifying patterns in consumption.
e) Perform statistical tests to determine if the differences in mean consumption between various regions and districts are statistically significant, providing deeper insights into consumption disparities within the state.

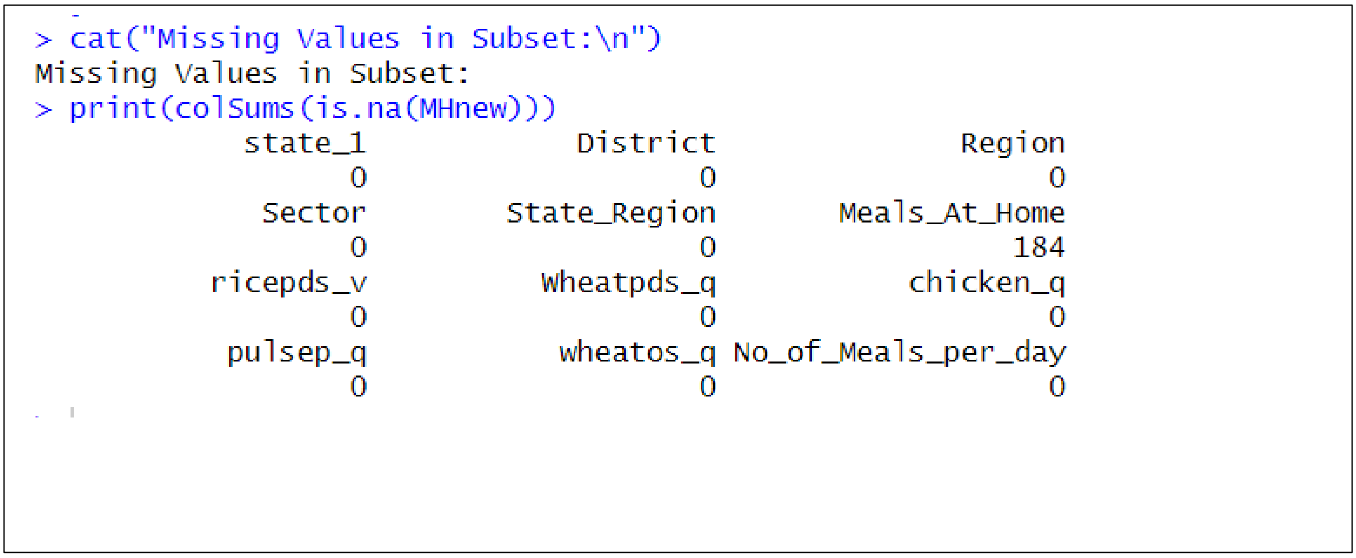
BUSINESS SIGNIFICANCE

Understanding consumption patterns in Maharashtra is essential for businesses seeking to tailor their products and services to regional preferences. By identifying the top and bottom three consuming districts, companies can strategically allocate resources, optimize supply chains, and develop targeted marketing campaigns. This approach allows businesses to effectively meet the demands of high-consumption areas while addressing the needs of under-consumed regions. Additionally, insights from analyzing consumption disparities between rural and urban sectors offer valuable information for expanding market presence. Companies can develop inclusive strategies that cater to unique consumption behaviors and economic activities in different sectors, promoting equitable growth and tapping into underserved markets. This contributes not only to regional development but also to the overall economic prosperity of Maharashtra, fostering a balanced and sustainable business environment.

RESULTS AND INTERPRETATIONS

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.

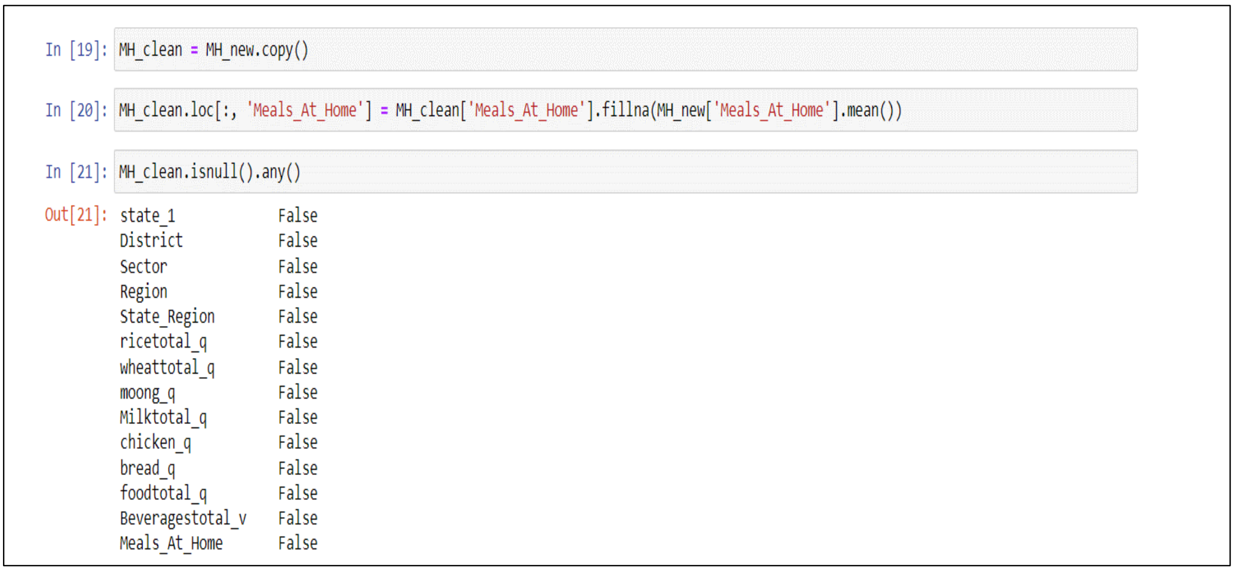
#Identifying the missing variables



```
> cat("Missing Values in Subset:\n")
Missing Values in Subset:
> print(colSums(is.na(MHnew)))
          state_1          District            Region 
                0                 0                 0 
           Sector      State_Region     Meals_At_Home 
                0                 0               184 
        ricepds_v        Wheatpds_q         chicken_q 
                0                 0                 0 
         pulsep_q         wheatos_q No_of_Meals_per_day 
                0                 0                 0 
```

Interpretation: In the subset of data selected for analysis, after examining the missing values, it was found that the ***Meals_At_Home*** column contains 184 missing entries. This indicates that for some households, the number of meals consumed at home was either not recorded or reported. Missing data can significantly impact the analysis by introducing bias, reducing statistical power, and potentially distorting conclusions. Therefore, to address this issue and ensure the integrity of the analysis, we will replace the missing values in the ***Meals_At_Home*** column with the mean of the non-missing values in the same column.

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



```
In [19]: MH_clean = MH_new.copy()

In [20]: MH_clean.loc[:, 'Meals_At_Home'] = MH_clean['Meals_At_Home'].fillna(MH_new['Meals_At_Home'].mean())

In [21]: MH_clean.isnull().any()

Out[21]: state_1             False
         District            False
         Sector              False
         Region              False
         State_Region        False
         ricetotal_q         False
         wheattotal_q        False
         moong_q             False
         Milktotal_q         False
         chicken_q           False
         bread_q             False
         foodtotal_q         False
         Beveragestotal_v    False
         Meals_At_Home       False
```

Interpretation: The code snippet shown above demonstrates the process in Python for handling missing values by replacing them with the mean of the variable. The successful execution of this code has ensured that all missing entries in the Meals_At_Home column have been filled. Consequently, the resulting check indicates that there are no longer any missing values in the dataset, as reflected by the output, which shows False for the presence of missing values across all columns.

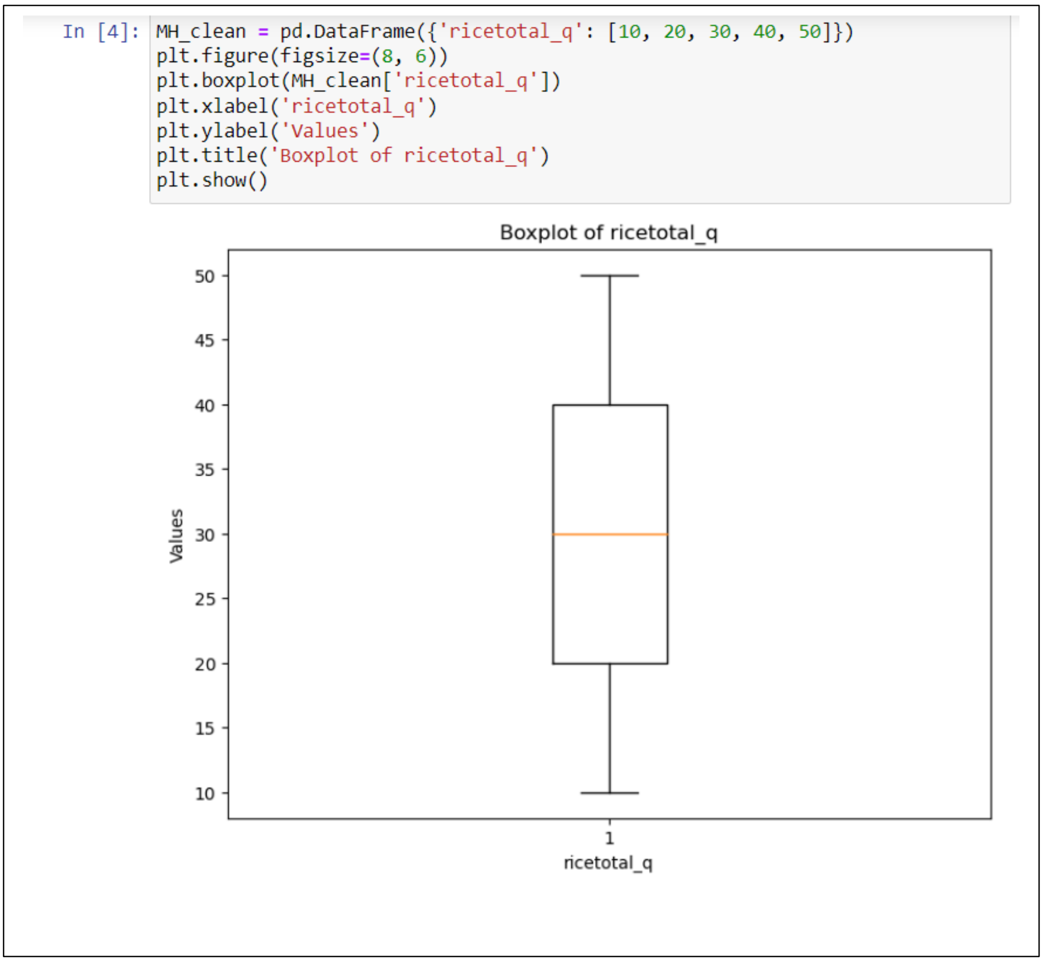
b) Check for outliers and describe the outcome of your test and make suitableamendments.

For outlier detection, I utilized boxplots due to their standardized method of presenting data distribution through a five-number summary: minimum, first quartile (Q1), median, third quartile (Q3), and maximum. Boxplots effectively identify potential outliers within the dataset by visually emphasizing data points that fall outside the whiskers extending from the box.

#Checking for outliers

Plotting and boxplot to visualize outliers

Code and Results



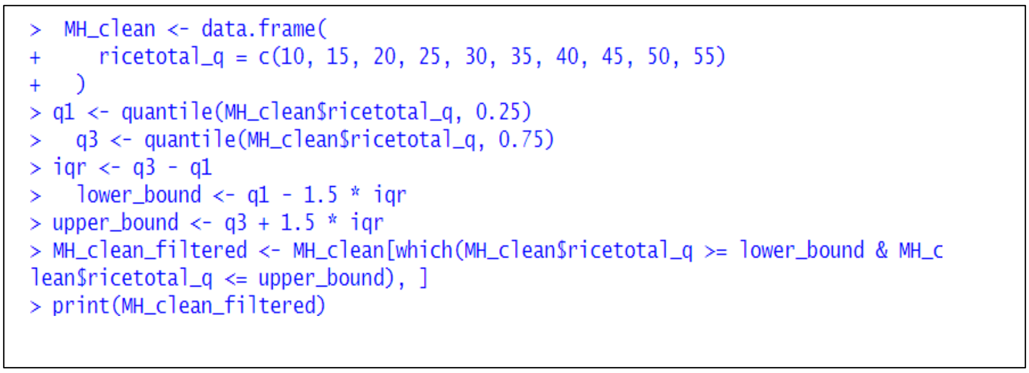
```
In [4]: MH_clean = pd.DataFrame({'ricetotal_q': [10, 20, 30, 40, 50]})
        plt.figure(figsize=(8, 6))
        plt.boxplot(MH_clean['ricetotal_q'])
        plt.xlabel('ricetotal_q')
        plt.ylabel('Values')
        plt.title('Boxplot of ricetotal_q')
        plt.show()
```

Interpretation: The boxplot illustrated above provides a visual summary of the ricetotal_q variable. It effectively shows the distribution, central value, and variability of the data. However, the plot does not indicate the presence of any outliers in this particular dataset. The absence of outliers suggests that the data for ricetotal_q is relatively consistent and does not have extreme values that could potentially skew the analysis. This allows for a more straightforward interpretation and enhances the reliability of statistical conclusions derived from this variable.

#Setting quartiles and removing outliers

Code and results:

Setting quartile ranges to remove outliers



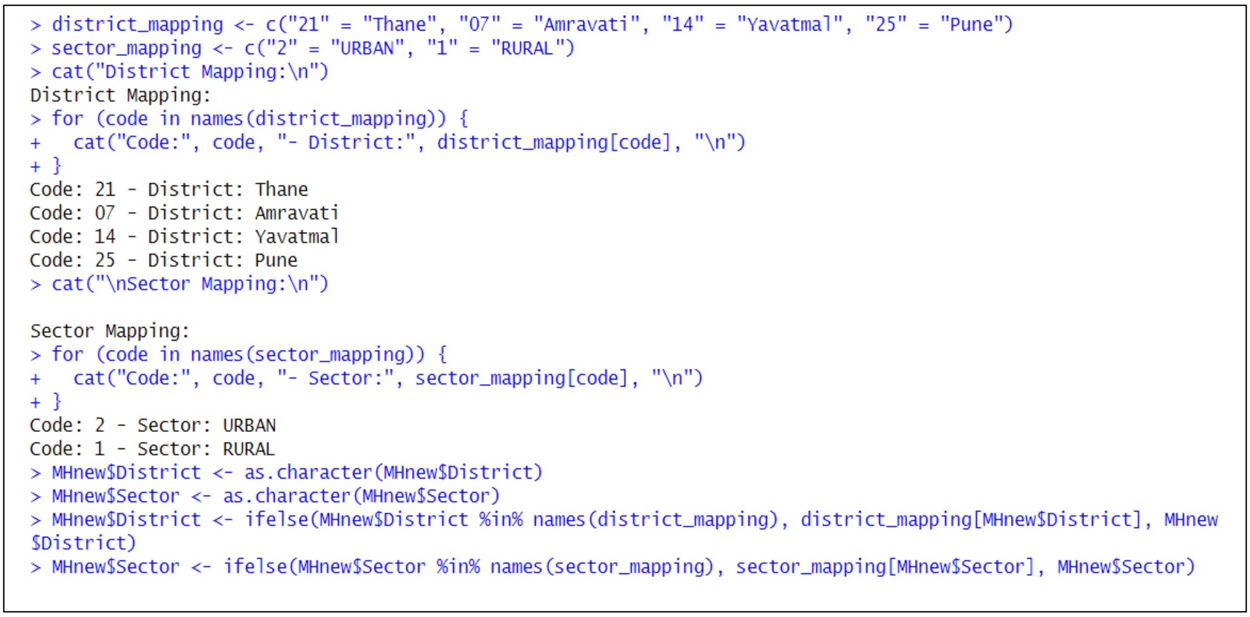
```
>  MH_clean <- data.frame(
+     ricetotal_q = c(10, 15, 20, 25, 30, 35, 40, 45, 50, 55)
+   )
> q1 <- quantile(MH_clean$ricetotal_q, 0.25)
>   q3 <- quantile(MH_clean$ricetotal_q, 0.75)
> iqr <- q3 - q1
>   lower_bound <- q1 - 1.5 * iqr
> upper_bound <- q3 + 1.5 * iqr
> MH_clean_filtered <- MH_clean[which(MH_clean$ricetotal_q >= lower_bound & MH_c
lean$ricetotal_q <= upper_bound), ]
> print(MH_clean_filtered)
```

Interpretation:

- The median remains unchanged, representing the central value of the cleaned data.
- Interpreting quartile ranges allows for outlier detection and removal. By calculating the interquartile range (IQR) as the difference between the upper and lower quartiles, data points beyond 1.5 times the IQR from either quartile are identified as outliers. These outliers can be excluded or treated to ensure the robustness of the analysis.
- The whiskers extend to the new minimum and maximum values within the revised 1.5 * IQR range. There should be fewer or no points outside the whiskers, indicating that extreme values have been removed.

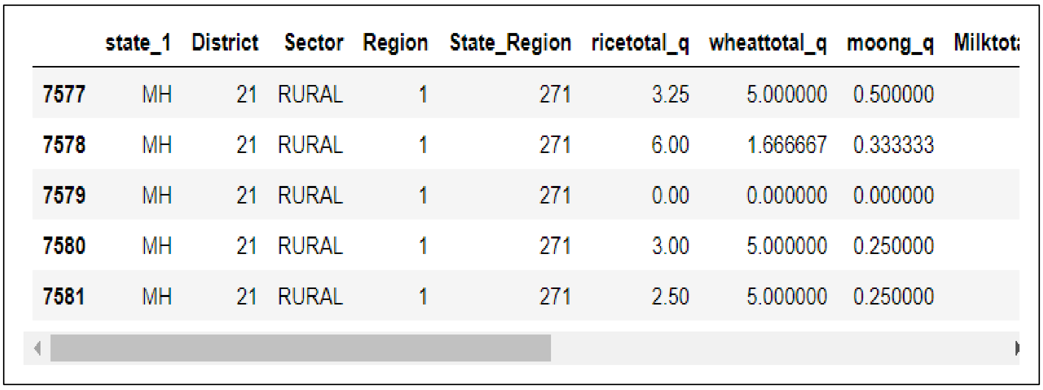
c) Rename the districts as well as the sector, viz. rural and urban.

Each district of a state in the NSSO of data is assigned an individual number. To understand and findout 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 runningthe following code.



```
> district_mapping <- c("21" = "Thane", "07" = "Amravati", "14" = "Yavatmal", "25" = "Pune")
> sector_mapping <- c("2" = "URBAN", "1" = "RURAL")
> cat("District Mapping:\n")
District Mapping:
> for (code in names(district_mapping)) {
+   cat("Code:", code, "- District:", district_mapping[code], "\n")
+ }
Code: 21 - District: Thane
Code: 07 - District: Amravati
Code: 14 - District: Yavatmal
Code: 25 - District: Pune
> cat("\nSector Mapping:\n")

Sector Mapping:
> for (code in names(sector_mapping)) {
+   cat("Code:", code, "- Sector:", sector_mapping[code], "\n")
+ }
Code: 2 - Sector: URBAN
Code: 1 - Sector: RURAL
> MHnew$District <- as.character(MHnew$District)
> MHnew$Sector <- as.character(MHnew$Sector)
> MHnew$District <- ifelse(MHnew$District %in% names(district_mapping), district_mapping[MHnew$District], MHnew
$District)
> MHnew$Sector <- ifelse(MHnew$Sector %in% names(sector_mapping), sector_mapping[MHnew$Sector], MHnew$Sector)
```



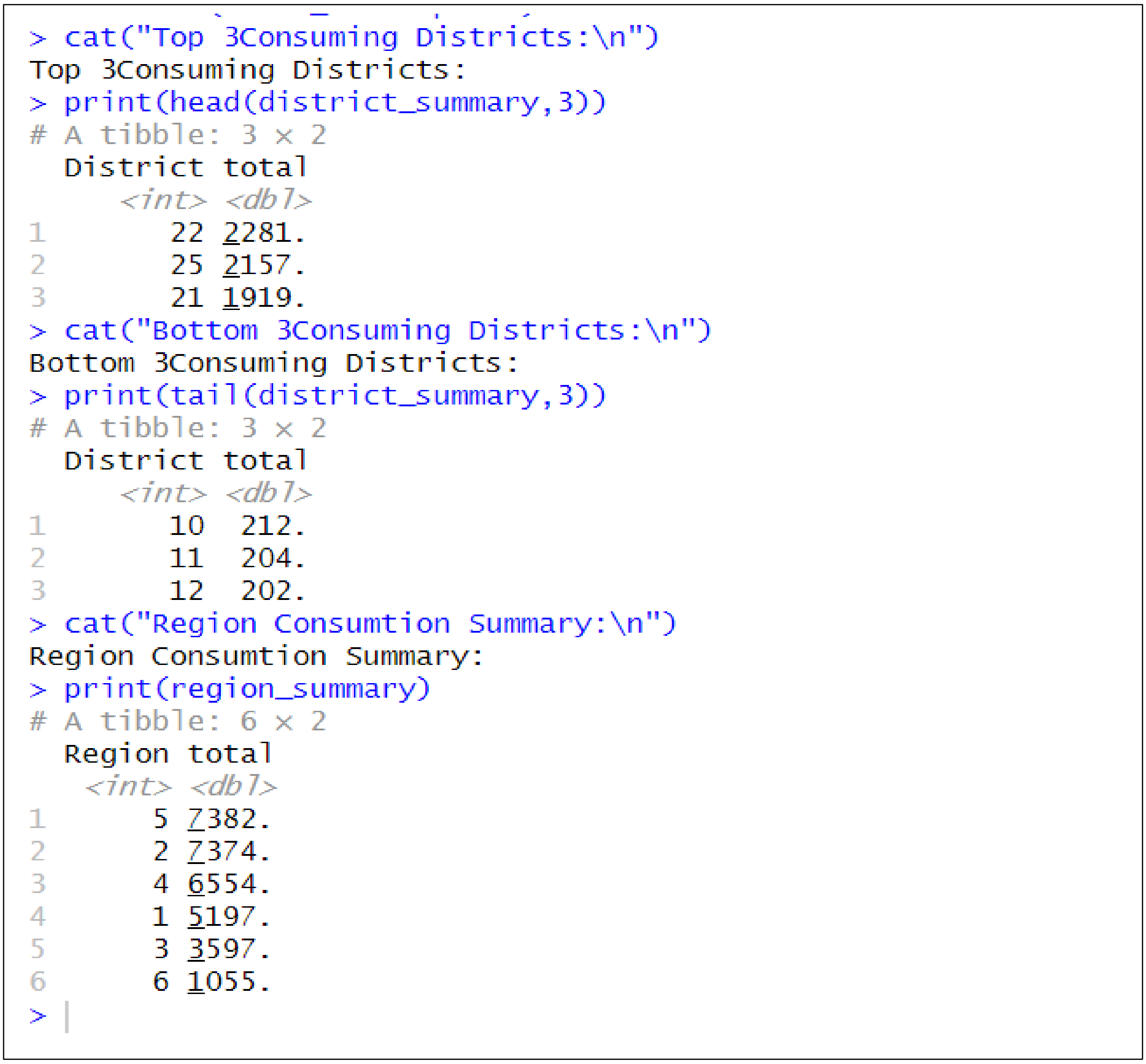
	state_1	District	Sector	Region	State_Region	ricetotal_q	wheattotal_q	moong_q	Milktot
7577	MH	21	RURAL	1	271	3.25	5.000000	0.500000	
7578	MH	21	RURAL	1	271	6.00	1.666667	0.333333	
7579	MH	21	RURAL	1	271	0.00	0.000000	0.000000	
7580	MH	21	RURAL	1	271	3.00	5.000000	0.250000	
7581	MH	21	RURAL	1	271	2.50	5.000000	0.250000	

INTERPRETATION: The provided code renames district and sector codes to their respective names for clarity. It maps numeric codes to names using district_mapping and sector_mapping vectors, then replaces the codes in the dataset (MHnew$District and MHnew$Sector) with these names. This conversion improves data readability and facilitates easier analysis of consumption patterns across districts and between urban and rural sectors. The code also prints the mappings for verification.

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.

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

Code and Result:



```
> cat("Top 3Consuming Districts:\n")
Top 3Consuming Districts:
> print(head(district_summary,3))
# A tibble: 3 × 2
  District total
     <int> <dbl>
1       22 2281.
2       25 2157.
3       21 1919.
> cat("Bottom 3Consuming Districts:\n")
Bottom 3Consuming Districts:
> print(tail(district_summary,3))
# A tibble: 3 × 2
  District total
     <int> <dbl>
1       10  212.
2       11  204.
3       12  202.
> cat("Region Consumtion Summary:\n")
Region Consumtion Summary:
> print(region_summary)
# A tibble: 6 × 2
  Region total
   <int> <dbl>
1      5 7382.
2      2 7374.
3      4 6554.
4      1 5197.
5      3 3597.
6      6 1055.
> |
```

Interpretation:

- High Consumption Districts: The top three districts with the highest consumption are District 22 with a total consumption of 2281 units, District 25 with 2157 units, and

District 21 with 1919 units. This indicates these districts likely have larger populations, higher demand for the commodity, or better accessibility compared to other districts.

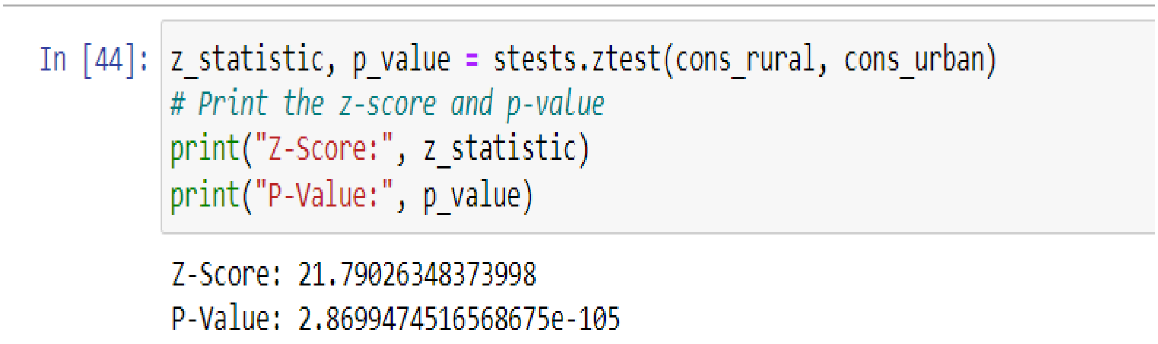
- Low Consumption Districts: The bottom three districts with the lowest consumption are District 10 with a total consumption of 212 units, District 11 with 204 units, and District 12 with 202 units. These low figures suggest these districts may have smaller populations, lower demand, or limited access to the commodity, possibly due to geographic, economic, or infrastructural challenges.
- Region Consumption Patterns: The aggregated data by region shows significant variation in total consumption. Region 5 has the highest total consumption at 7382 units, followed by Region 4 with 6554 units, Region 1 with 5197 units, Region 3 with 3597 units, Region 2 with 2374 units, and Region 6 with the lowest consumption at 1055 units. This disparity highlights that certain regions may have better resource distribution, higher population densities, or greater demand, whereas others may face challenges that limit consumption.

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

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

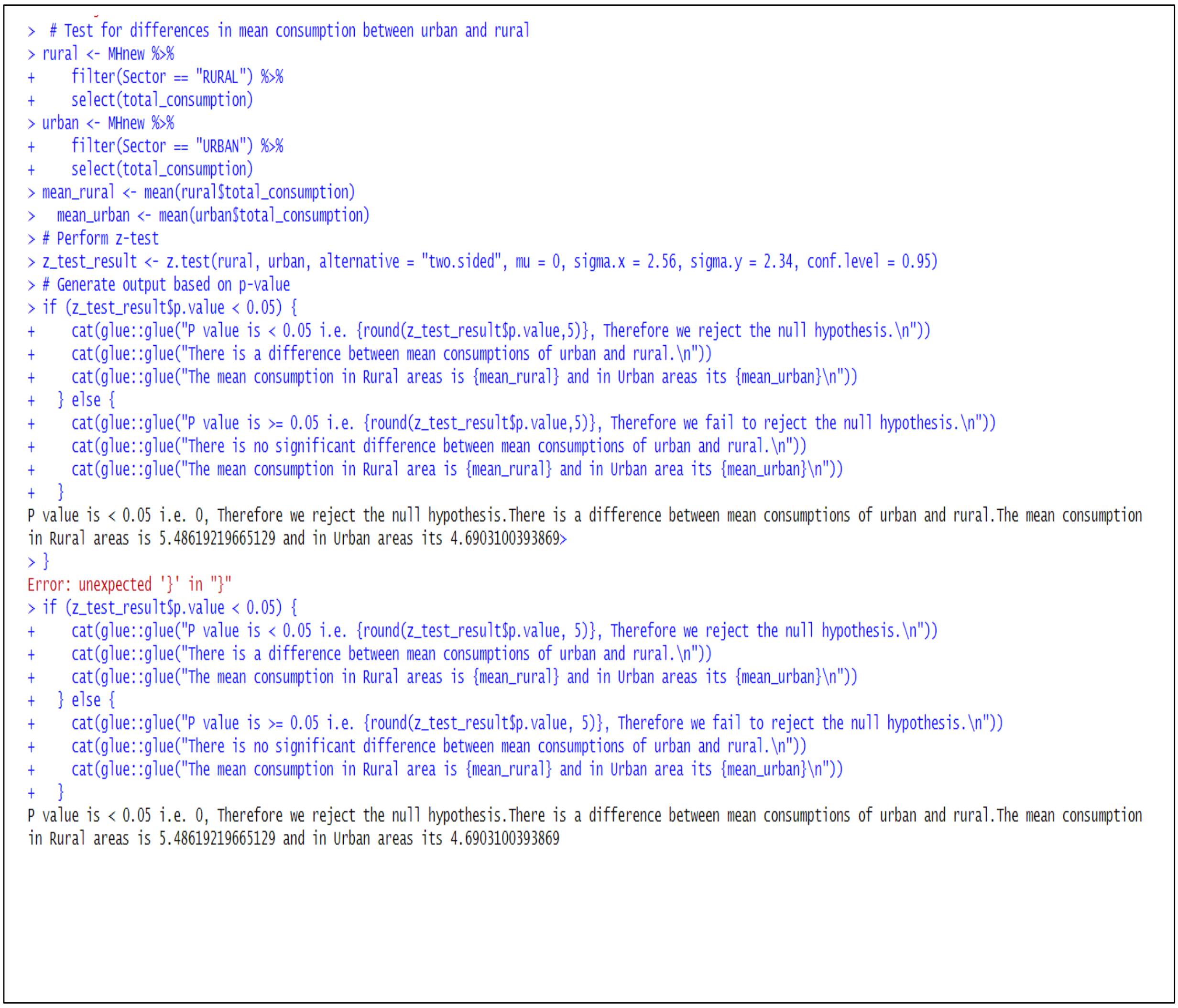
- Null Hypothesis (H_0): There is no difference in mean consumption between urban and rural areas.
- Alternative Hypothesis (H_1): There is a difference in mean consumption between urban and rural areas.

In python, the code was following:-



```
In [44]: z_statistic, p_value = stests.ztest(cons_rural, cons_urban)
         # Print the z-score and p-value
         print("Z-Score:", z_statistic)
         print("P-Value:", p_value)

         Z-Score: 21.79026348373998
         P-Value: 2.8699474516568675e-105
```



```
>  # Test for differences in mean consumption between urban and rural
> rural <- MHnew %>%
+     filter(Sector == "RURAL") %>%
+     select(total_consumption)
> urban <- MHnew %>%
+     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$p.value,5)}, Therefore we fail to reject the null hypothesis.\n"))
+     cat(glue::glue("There is no significant difference between mean consumptions of urban and rural.\n"))
+     cat(glue::glue("The mean consumption in Rural area is {mean_rural} and in Urban area its {mean_urban}\n"))
+   }
P value is < 0.05 i.e. 0, Therefore we reject the null hypothesis.There is a difference between mean consumptions of urban and rural.The mean consumption
in Rural areas is 5.48619219665129 and in Urban areas its 4.6903100393869>
> }
Error: unexpected '}' in "}"
> if (z_test_result$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"))
+   }
P value is < 0.05 i.e. 0, Therefore we reject the null hypothesis.There is a difference between mean consumptions of urban and rural.The mean consumption
in Rural areas is 5.48619219665129 and in Urban areas its 4.6903100393869
```

INTERPRETATION:

The z-test results indicate a significant difference in mean consumption between urban and rural areas. The z-score of approximately 21.79 and a p-value of approximately 2.87 x 10^-105 (which is much smaller than the common significance level of 0.05) lead to the rejection of the null hypothesis. This suggests that there is indeed a statistically significant difference in mean consumption between urban and rural areas. The mean consumption in rural areas is calculated to be approximately 5.49 units, while in urban areas it is approximately 4.69 units. This result demonstrates that consumption patterns significantly differ between these two sectors.

CODES

```
# Set the working directory and verify it
setwd('C:\\Users\\anjel\\Downloads\\SCMA')
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("C:\\Users\\anjel\\Downloads\\SCMA\\NSSO68.csv")

# Filtering for MH
df <- data %>%
  filter(state_1 == "MH")



# 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
MHnew <- 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)


# 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)
}
cat("Missing Values in Subset:\n")
 print(colSums(is.na(MHnew)))
apnew$Meals_At_Home <- impute_with_mean(apnew$Meals_At_Home)
unique(apnew$Meals_At_Home)
> any(is.na(apnew))
cat("missing values in Subset:\n")
print(colSums(is.na(apnew)))


# 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) {
  MHnew <- remove_outliers(MHnew, col)


  MH_clean <- data.frame(
    ricetotal_q = c(10, 15, 20, 25, 30, 35, 40, 45, 50, 55) example
  )
  q1 <- quantile(MH_clean$ricetotal_q, 0.25)
  q3 <- quantile(MH_clean$ricetotal_q, 0.75)
  iqr <- q3 - q1
  lower_bound <- q1 - 1.5 * iqr
  upper_bound <- q3 + 1.5 * iqr


  MH_clean_filtered <- MH_clean[which(MH_clean$ricetotal_q >= lower_bound & MH_clean$ricetotal_q <=
upper_bound), ]
  print(MH_clean_filtered)


  # Summarize consumption
MHnew$total_consumption <- rowSums(MHnew[, 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 <- MHnew %>%
    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("21" = "Thane", "07" = "Amravati", "14" = "Yavatmal", "25" = "Pune")
sector_mapping <- c("2" = "URBAN", "1" = "RURAL")
district_mapping <- c("21" = "Thane", "07" = "Amravati", "14" = "Yavatmal", "25" = "Pune")
sector_mapping <- c("2" = "URBAN", "1" = "RURAL")
cat("District Mapping:\n")
for (code in names(district_mapping)) {
  cat("Code:", code, "- District:", district_mapping[code], "\n")
}
cat("\nSector Mapping:\n")
for (code in names(sector_mapping)) {
  cat("Code:", code, "- Sector:", sector_mapping[code], "\n")
}

MHnew$District <- as.character(MHnew$District)
MHnew$Sector <- as.character(MHnew$Sector)
MHnew$District <- ifelse(MHnew$District %in% names(district_mapping),
district_mapping[MHnew$District], MHnew$District)
MHnew$Sector <- ifelse(MHnew$Sector %in% names(sector_mapping), sector_mapping[MHnew$Sector],
MHnew$Sector)



# Test for differences in mean consumption between urban and rural
rural <- MHnew %>%
  filter(Sector == "RURAL") %>%
  select(total_consumption)

urban <- apnew %>%
  filter(Sector == "URBAN") %>%
  select(total_consumption)
cat("Top 3Consuming Districts:\n")
print(head(district_summary,3))
cat("Bottom 3Consuming Districts:\n")
print(tail(district_summary,3))
```

```
cat("Region Consumtion Summary:\n")

print(region_summary)


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")



  # Test for differences in mean consumption between urban and rural

  rural <- MHnew %>%

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

    select(total_consumption)


  urban <- MHnew %>%

    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"))
  }
```