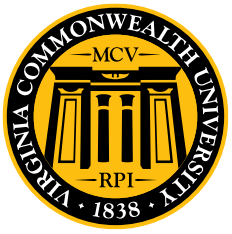


VIRGINIA COMMONWEALTH UNIVERSITY

STATISTICAL ANALYSIS AND MODELLING (SCMA 632)

A1a: CONSUMPTION PATTERN OF GUJARAT USING PYTHON AND R

GAURI VINOD NAIR

V01110160

Date of Submission: 16-06-2024

TABLE OF CONTENTS

INTRODUCTION

The focus of this study is on the state of Gujarat, utilizing data from the National Sample Survey Office (NSSO), to identify the top and bottom three consuming districts within the state. To achieve this, we meticulously manipulate and clean the dataset to extract the necessary information for analysis. The dataset comprises consumption-related data, encompassing both rural and urban sectors, along with district-wise variations. This dataset has been imported into R, a powerful and versatile statistical programming language, well-suited for handling and analysing large datasets.

Our objectives in this study include identifying and addressing missing values, handling outliers, standardizing district and sector names, summarizing consumption data both regionally and district-wise, and testing the significance of differences in mean consumption. The insights derived from this study will provide valuable information for policymakers and stakeholders, facilitating targeted interventions and promoting equitable development across Gujarat.

OBJECTIVE

1. Identify and Address Missing Values
   Check if there are any missing values in the data.
   Identify the missing values.
   Replace missing values with the mean of the respective variable.

2. Detect and Amend Outliers
   Check for outliers in the dataset.
   Describe the outcome of the outlier detection test.
   Make suitable amendments to handle the identified outliers.

3. Standardize Naming Conventions
   Rename the districts and sectors to maintain consistency, specifying sectors as rural and urban.

4. Summarize Consumption Data
   Summarize critical variables in the dataset both region-wise and district-wise.
   Identify and indicate the top and bottom three districts in terms of consumption.

5. Analyse Mean Differences
   Test whether the differences in mean consumption values are statistically significant.

BUSINESS SIGNIFICANCE

The focus of this study on Gujarat'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 Gujarat's economic growth.

RESULTS AND INTERPRETATION

1. Identifying and addressing missing values

Identifying missing values

```
> # Finding missing values
> missing_info <- colSums(is.na(df))
> cat("Missing Values Information:\n")
Missing Values Information:
> print(missing_info)
                          slno                            grp                   Round_Centre                      FSU_number 
                            0                              0                              0                              0 
                         Round                 Schedule_Number                         Sample                         Sector 
                            0                              0                              0                              0 
                         state                   State_Region                       District                  Stratum_Number 
                            0                              0                              0                              0 
                    Sub_Stratum                  Schedule_type                      Sub_Round                     Sub_Sample 
                            0                              0                              0                              0 
                FOD_Sub_Region          Hamlet_Group_Sub_Block                              t                X_Stage_Stratum 
                            0                              0                              0                              0 
                        HHS_No                          Level                         Filler                          hhdsz 
                            0                              0                              0                              0 
                      NIC_2008                       NCO_2004                        HH_type                       Religion 
                          154                            155                              0                              0 
                   Social_Group            Whether_owns_any_land             Type_of_land_owned                    Land_Owned 
                            0                              0                            460                            470 
                Land_Leased_in                Otherwise_possessed               Land_Leased_out           Land_Total_possessed 
                         2911                           3336                           3326                              0 
    During_July_June_Cultivated          During_July_June_Irrigated                            NSS                            NSC 
                         2416                           2714                              0                              0 
                            MLT                        land_tt                    Cooking_code                   Lighting_code 
                            0                              0                              0                              0 
             Dwelling_unit_code             Regular_salary_earner                Perform_Ceremony Meals_seved_to_non_hhld_members 
                            0                              0                              0                            415 
```

identifying missing values of the subset of the dataset

```
> # Sub-setting the data
> gujnew <- 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)
> # Check for missing values in the subset
> cat("Missing Values in Subset:\n")
Missing Values in Subset:
> print(colSums(is.na(gujnew)))
     state_1     District       Region       Sector  State_Region Meals_At_Home    ricepds_v 
           0            0            0            0            0            1            0 
  Wheatpds_q    chicken_q    pulsep_q   wheatos_q No_of_Meals_per_day 
           0            0            0            0            0 
> 
```

imputing the missing values, i.e, replacing it with the mean

```
> # 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)
+ }
> gujnew$Meals_At_Home <- impute_with_mean(gujnew$Meals_At_Home)
> # Check for missing values after imputation
> cat("Missing Values After Imputation:\n")
Missing Values After Imputation:
> print(colSums(is.na(gujnew)))
     state_1     District       Region       Sector  State_Region Meals_At_Home    ricepds_v 
           0            0            0            0            0            0            0 
  Wheatpds_q    chicken_q    pulsep_q   wheatos_q No_of_Meals_per_day 
           0            0            0            0            0 
```

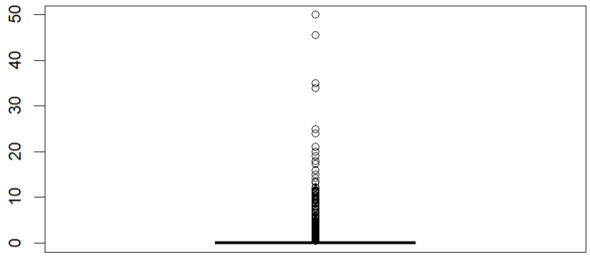
Interpretation

A subset of the dataset was created with the attributes state_1, District, Region, Sector, State_Region, Meals_At_Home, ricepds_v, Wheatpds_q, chicken_q, pulsep_q, wheatos_q, and No_of_Meals_per_day. During the data cleaning process, we identified that Meals_At_Home had one missing value. Missing values in the dataset can be problematic as they lead to incomplete or biased analyses, hindering the accuracy of results and potentially skewing interpretations and decision-making processes. Therefore, to ensure data integrity, we replaced the missing value with the mean of the variable.

2. Detecting and amending outliers
   Here , we use boxplot to detect if there are any outliers present.

#Checking for outliers in ricepds_v

```
#checking for outliers
#box plot
boxplot(gujnew$ricepds_v)
```



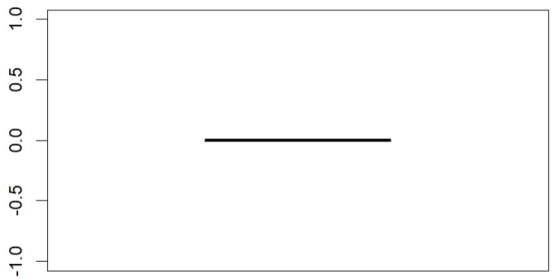
Finding outliers using quartiles and removing them

```
# 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]]
  return(df)
}

outlier_columns <- c("ricepds_v")
for (col in outlier_columns) {
  gujnew <- remove_outliers(gujnew, col)
}
```

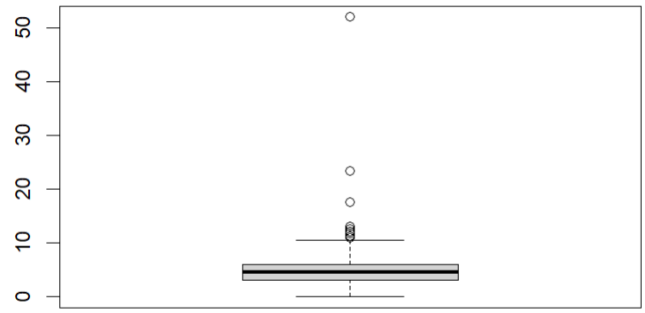
box plot after removing outliers

```
#after removing outliers
boxplot(gujnew$ricepds_v)
```

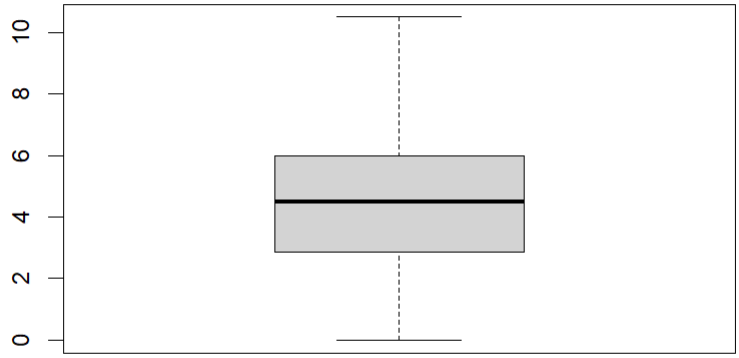


Similarly, the box plot before and after removing outliers for wheatos_q

```
#checking for outliers
#box plot
boxplot(gujnew$wheatos_q)
```




```
#after removing outliers
boxplot(gujnew$wheatos_q)
```



Interpretation

From the boxplot above, which is a visual representation of the variable 'ricepds_v' shows that there is an outlier. Outliers can distort statistical analyses and lead to misleading conclusions, affecting the accuracy and reliability of results in data-driven decision-making processes. The outliers were removed by using the method of inter quartile range by calculating it as the difference between the upper and lower quartiles, data points beyond 1.5 times the IQR from either quartile are identified as outliers and can be excluded or treated to ensure the robustness of the analysis.

3. Standardize naming conventions

```
# Rename districts and sectors , get codes from appendix of NSSO 68th ROund Data
district_mapping <- c("1" = "Ahmedabad", "2"="Amreli","3"="Anand","4"="Aravalli","5"="Banaskantha",
                      "6"="Bharuch","7"="Bhavnagar","8"="Botad","9"="Chota Udaipur","10"="Dahid","11"="Dang",
                      "12"="Dwarka","13"="Gandhinagar","14"="Gor Somnath","15"="Jamnagar","16"="Junagadh",
                      "17"="Kutch","18"="Kheda","19"="Mahisagar","20"="Mehsana","21"="Morbi","22"="Narmada")
sector_mapping <- c("2" = "URBAN", "1" = "RURAL")

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

```
> gujnew$District
  [1] "Bhavnagar"     "Bhavnagar"     "Bhavnagar"     "Bhavnagar"
  [5] "Bhavnagar"     "Bhavnagar"     "Bhavnagar"     "Bhavnagar"
  [9] "Bhavnagar"     "Bhavnagar"     "Bhavnagar"     "Bhavnagar"
 [13] "Bhavnagar"     "Bhavnagar"     "Bhavnagar"     "Bhavnagar"
 [17] "Bhavnagar"     "Bhavnagar"     "Bhavnagar"     "Bhavnagar"
 [21] "Bhavnagar"     "Bhavnagar"     "Bhavnagar"     "Bhavnagar"
 [25] "Bhavnagar"     "Gor Somnath"   "Gor Somnath"   "Gor Somnath"
 [29] "Gor Somnath"   "Gor Somnath"   "Gor Somnath"   "Gor Somnath"
 [33] "Bhavnagar"     "Bhavnagar"     "Bhavnagar"     "Bhavnagar"
 [37] "Bhavnagar"     "Bhavnagar"     "Bhavnagar"     "Bhavnagar"
 [41] "Bhavnagar"     "Bhavnagar"     "Bhavnagar"     "Gor Somnath"
 [45] "Gor Somnath"   "Gor Somnath"   "Bhavnagar"     "Bhavnagar"
 [49] "Bhavnagar"     "Bhavnagar"     "Bhavnagar"     "Gor Somnath"
 [53] "Gor Somnath"   "Gor Somnath"   "Gor Somnath"   "Gor Somnath"
 [57] "Bhavnagar"     "Bhavnagar"     "Bhavnagar"     "Bhavnagar"
 [61] "Bhavnagar"     "Bhavnagar"     "Bhavnagar"     "Bhavnagar"
 [65] "Bhavnagar"     "Bhavnagar"     "Bhavnagar"     "Bhavnagar"
 [69] "Bhavnagar"     "Bhavnagar"     "Bhavnagar"     "Bhavnagar"
 [73] "Bhavnagar"     "Bhavnagar"     "Bhavnagar"     "Bhavnagar"
 [77] "Bhavnagar"     "Bhavnagar"     "Bhavnagar"     "Bhavnagar"
 [81] "Bharuch"       "Bharuch"       "Bharuch"       "Bharuch"
 [85] "Bharuch"       "Bharuch"       "Bharuch"       "Ahmedabad"
 [89] "Ahmedabad"     "Ahmedabad"     "Ahmedabad"     "Ahmedabad"
 [93] "Bhavnagar"     "Bhavnagar"     "Bhavnagar"     "Bhavnagar"
 [97] "Bhavnagar"     "Bhavnagar"     "Bhavnagar"     "Bhavnagar"
[101] "Bhavnagar"     "Bhavnagar"     "Bhavnagar"     "Bhavnagar"
[105] "Bhavnagar"     "Bhavnagar"     "Bhavnagar"     "Bhavnagar"
```

```
> gujnew$Sector
  [1] "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN"
 [16] "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN"
 [31] "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN"
 [46] "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN"
 [61] "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN"
 [76] "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN"
 [91] "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN"
[106] "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN"
[121] "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN"
[136] "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN"
[151] "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN"
[166] "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN"
[181] "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN"
[196] "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN"
[211] "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN"
[226] "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN"
[241] "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN"
[256] "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN"
[271] "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN"
[286] "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN"
[301] "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN"
[316] "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN"
[331] "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN"
[346] "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN" "URBAN"
```

```
    state_1      District Sector
1       GUJ     Bhavnagar  URBAN
3       GUJ     Bhavnagar  URBAN
4       GUJ     Bhavnagar  URBAN
5       GUJ     Bhavnagar  URBAN
6       GUJ     Bhavnagar  URBAN
7       GUJ     Bhavnagar  URBAN
8       GUJ     Bhavnagar  URBAN
10      GUJ     Bhavnagar  URBAN
11      GUJ     Bhavnagar  URBAN
12      GUJ     Bhavnagar  URBAN
13      GUJ     Bhavnagar  URBAN
14      GUJ     Bhavnagar  URBAN
15      GUJ     Bhavnagar  URBAN
16      GUJ     Bhavnagar  URBAN
17      GUJ     Bhavnagar  URBAN
19      GUJ     Bhavnagar  URBAN
21      GUJ     Bhavnagar  URBAN
22      GUJ     Bhavnagar  URBAN
24      GUJ     Bhavnagar  URBAN
25      GUJ     Bhavnagar  URBAN
26      GUJ     Bhavnagar  URBAN
29      GUJ     Bhavnagar  URBAN
30      GUJ     Bhavnagar  URBAN
31      GUJ     Bhavnagar  URBAN
32      GUJ     Bhavnagar  URBAN
33      GUJ   Gor Somnath  URBAN
34      GUJ   Gor Somnath  URBAN
```

Interpretation

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 assignment 1 and 2 respectively.
The result as show above has successfully assigned the district names to the given number. Also the sectors 1 and 2 have been replaced as urban and rural sectors respectively.

4. Summarize consumption data

```
# Summarize consumption
gujnew$total_consumption <- rowSums(gujnew[, 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 <- gujnew %>%
    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)
```

```
Top 3 Consuming Districts:
> print(head(district_summary, 3))
# A tibble: 3 × 2
  District      total
  <chr>         <dbl>
1 Bhavnagar     1466.
2 Narmada       1113.
3 Chota Udaipur  781.
> cat("Bottom 3 Consuming Districts:\n")
Bottom 3 Consuming Districts:
> print(tail(district_summary, 3))
# A tibble: 3 × 2
  District total
  <chr>    <dbl>
1 Mehsana  155.
2 Botad    135.
3 23        48.8
> cat("Region Consumption Summary:\n")
Region Consumption Summary:
> print(region_summary)
# A tibble: 5 × 2
  Region total
   <int> <dbl>
1      2 3027.
2      1 2994.
3      5 2891.
4      3  681.
5      4  215.
```

Interpretation

Here, we find out the top 3 and bottom 3 consuming districts. The district with the most total consumption is Bhavnagar with a total consumption of 1466 units followed by Narmada with 1113 units and then Chota Udaipur with 781 units.

The district with the least total consumption is Navsari with 48.8 units followed by Botad and Mehsana with 135 units and 155 units respectively.

5. Analyze mean differences

Here, we test the following hypothesis

H0: there is no significant difference between the total consumption of rural and urban areas.

H1: there is a significant difference between the total consumption of rural and urban areas.

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

urban <- gujnew %>%
  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"))
}
```

```
> z_test_result

        Two-sample z-Test

data:  rural and urban
z = -11.994, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -1.483885 -1.067022
sample estimates:
mean of x mean of y
 3.857037  5.132491
```

```
> # 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 3.85703728364783 and in Urban areas its 5.13249100113808
```

Interpretation

Here we observe that the p value (p = 2.2e-16) is less than the level of significance (α = 0.05). Therefore we reject the null hypothesis.

This implies that there is a significant difference in the total consumption between the rural and the urban areas. We also observe that the average total consumption of urban is greater than that of the rural area.

CODE

```
# 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("C:/Users/gauri/Downloads/NSSO68.csv")
# Display the first few rows of the data
head(data)

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

# 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
gujnew <- 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)

# Check for missing values in the subset
cat("Missing Values in Subset:\n")
print(colSums(is.na(gujnew)))

# 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)
}
gujnew$Meals_At_Home <- impute_with_mean(gujnew$Meals_At_Home)

# Check for missing values after imputation
cat("Missing Values After Imputation:\n")
print(colSums(is.na(gujnew)))

#checking for outliers
#box plot
boxplot(gujnew$wheatos_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)
}
outlier_columns <- c("wheatos_q")
for (col in outlier_columns) {
  gujnew <- remove_outliers(gujnew, col)
}

#after removing outliers
boxplot(gujnew$wheatos_q)

# Summarize consumption
gujnew$total_consumption <- rowSums(gujnew[, 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 <- gujnew %>%
    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("1" = "Ahmedabad",
"2"="Amreli","3"="Anand","4"="Aravalli","5"="Banaskantha",
                      "6"="Bharuch","7"="Bhavnagar","8"="Botad","9"="Chota
Udaipur","10"="Dahid","11"="Dang",
                      "12"="Dwarka","13"="Gandhinagar","14"="Gor
Somnath","15"="Jamnagar","16"="Junagadh",

"17"="Kutch","18"="Kheda","19"="Mahisagar","20"="Mehsana","21"="Morbi","22"="Nar
mada")
sector_mapping <- c("2" = "URBAN", "1" = "RURAL")
gujnew$District <- as.character(gujnew$District)
gujnew$Sector <- as.character(gujnew$Sector)
gujnew$District <- ifelse(gujnew$District %in% names(district_mapping),
district_mapping[gujnew$District],
                          gujnew$District)
gujnew$Sector <- ifelse(gujnew$Sector %in% names(sector_mapping),
sector_mapping[gujnew$Sector], gujnew$Sector)

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

}
```