VIRGINIA COMMONWEALTH UNIVERSITY

Statistical analysis and modelling (SCMA 632)

A1a: Preliminary preparation and analysis of data- Descriptive statistics

M K NITHEESH

V01107616

Date of Submission: 16-06-2024

CONTENTS

Analyzing Consumption in the State of West Bengal Using R

Introduction

The focus of this study is on the state of West Bengal, from the NSSO data, to find the top and bottom three consuming districts of West Bengal. In the process, we manipulate and clean the dataset to get the required data to analyse. To facilitate this analysis, we have gathered a dataset containing consumption-related information, including data on rural and urban sectors, as well as district-wise variations. The dataset has been imported into R, a powerful statistical programming language renowned for its versatility in handling and analysing large datasets.

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

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.

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

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

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.

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

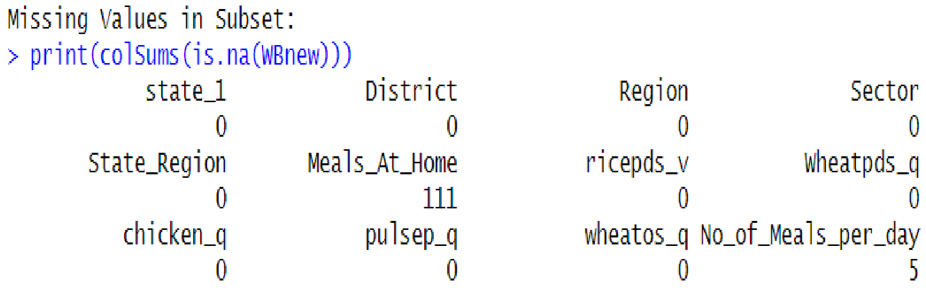
BUSINESS SIGNIFICANCE

The focus of this study on West Bengals 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 West Bengals economic growth.

A) RESULTS AND INTERPRETATION

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.

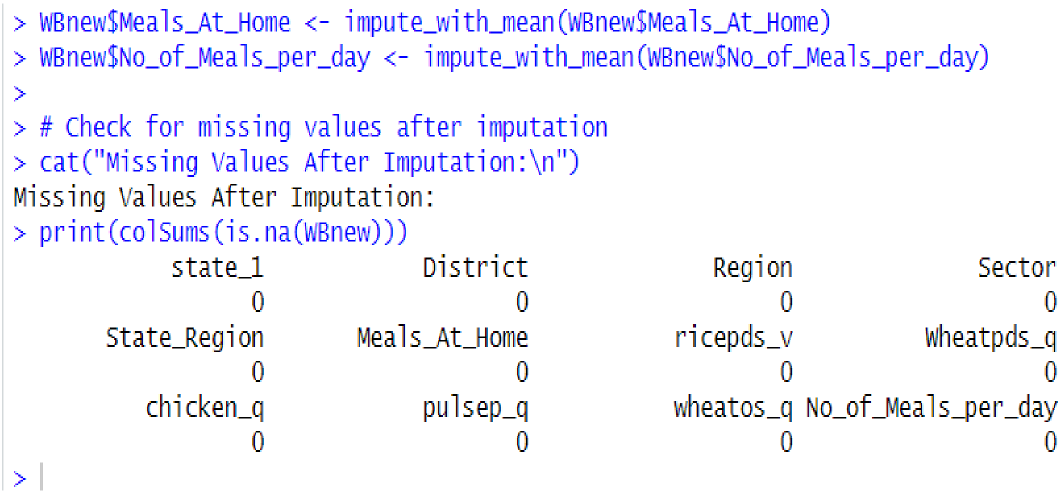
#Identifying the missing values.



```
Missing Values in Subset:
> print(colSums(is.na(WBnew)))
          state_1          District            Region            Sector 
                0                 0                 0                 0 
     State_Region      Meals_At_Home         ricepds_v        Wheatpds_q 
                0               111                 0                 0 
        chicken_q          pulsep_q         wheatos_q No_of_Meals_per_day 
                0                 0                 0                 5 
```

Interpretation: In the subset of the WBnew dataset, there are several missing values that need to be addressed. The Meals_At_Home column has 111 missing values, while the No_of_Meals_per_day column has 5 missing values. All other selected columns, including state_1, District, Region, Sector, State_Region, ricepds_v, Wheatpds_q, chicken_q, pulsep_q, and wheatos_q, have no missing values. It is crucial to manage these missing values appropriately to maintain the integrity and reliability of any analysis performed on this dataset. Potential strategies include data imputation or the exclusion of rows with missing data.

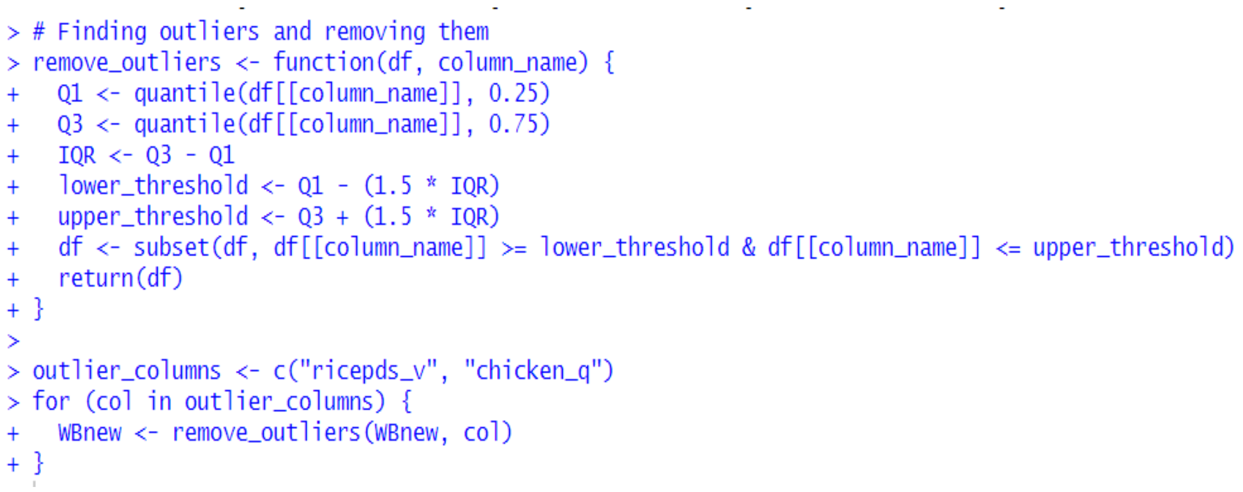
#Imputing the values, i.e. replacing the missing values with mean.



```
> WBnew$Meals_At_Home <- impute_with_mean(WBnew$Meals_At_Home)
> WBnew$No_of_Meals_per_day <- impute_with_mean(WBnew$No_of_Meals_per_day)
> 
> # Check for missing values after imputation
> cat("Missing Values After Imputation:\n")
Missing Values After Imputation:
> print(colSums(is.na(WBnew)))
          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 above code 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.

B) Check for outliers and describe the outcome of your test and make suitable amendments.



```
> # 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) {
+   WBnew <- remove_outliers(WBnew, col)
+ }
```

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 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. This is done by running the following code.

```
> # Rename districts and sectors, get codes from appendix of NSSO 68th ROund Data
> district_mapping <- c ("11" = "North Twenty-Four Parganas", "9" = "Barddhaman", "17" = "Kolkata")
> sector_mapping <- c ("2" = "URBAN", "1" = "RURAL")
>
> WBnew$District <- as.character(WBnew$District)
> WBnew$Sector <- as.character(WBnew$Sector)
> WBnew$District <- ifelse(WBnew$District %in% names(district_mapping), district_mapping[WBnew$District], WBnew$District)
> WBnew$Sector <- if else (WBnew$Sector %in% names(sector_mapping), sector_mapping [WBnew$Sector], WBnew$Sector)
```

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

```
> # Summarize consumption

> WBnew$total_consumption <- rowSums(WBnew[, 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 <- WBnew %>%

+     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")


Top 3 Consuming Districts:

> print(head(district_summary, 3))

# A tibble: 3 × 2
  District                    total
  <chr>                       <dbl>
1 North Twenty-Four Parganas 1287.
2 Barddhaman                 1206.
3 Kolkata                     924.
```

Interpretation: The top three consuming districts are North Twenty-Four Parganas with 1287 units, followed by Barddhaman with 1206 units, and then in the third place Kolkata with 924 units

```
> cat("Bottom 3 Consuming Districts:\n")
Bottom 3 Consuming Districts:

> print(tail(district_summary, 3))
```

```
# A tibble: 3 × 2
  District total
  <chr>    <dbl>
1 1         136.
2 5         133.
3 3         120.
```

Interpretation: The bottom three consuming districts are 1 with 136 units, followed by 5 with 133 units, and then in the third place 3 with 120 units

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

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

#H0: There is no difference in consumption between urban and rural.

#H1: There is difference in consumption between urban and rural.

mean_rural <- mean(rural$total_consumption)

mean_urban <- mean(urban$total_consumption)

z_test_result <- z.test(rural, urban, alternative = "two.sided", mu = 0, sigma.x = 2.56, sigma.y = 2.34, conf.level = 0.95)

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 1.6404780907106 and in Urban areas its 2.47491594640355

CODES

```
# Set the working directory and verify it
setwd('C:/Users/nithe/OneDrive/Desktop')
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","glue")
lapply(libraries, install_and_load)

# Reading the file into R
NIT <- read.csv("NSSO68.csv")

# Filtering for TN
df <- NIT %>%
  filter(state == "19")

# 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 IPL1
WBnew <- 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(WBnew)))

# 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)
}
```

```
WBnew$Meals_At_Home <-
impute_with_mean(WBnew$Meals_At_Home)
WBnew$No_of_Meals_per_day <-
impute_with_mean(WBnew$No_of_Meals_per_day)


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


# 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) {
  WBnew <- remove_outliers(WBnew, col)
}


# Summarize consumption
WBnew$total_consumption <- rowSums(WBnew[, 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 <- WBnew %>%
    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("11" = "North Twenty Four Parganas", "9" = "Barddhaman", "17" = "Kolkata")
sector_mapping <- c("2" = "URBAN", "1" = "RURAL")

WBnew$District <- as.character(WBnew$District)
```

```
WBnew$Sector <- as.character(WBnew$Sector)

WBnew$District <- ifelse(WBnew$District %in%
names(district_mapping), district_mapping[WBnew$District],
WBnew$District)

WBnew$Sector <- ifelse(WBnew$Sector %in% names(sector_mapping),
sector_mapping[WBnew$Sector], WBnew$Sector)



# Test for differences in mean consumption between urban and rural

rural <- WBnew %>%

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

  select(total_consumption)


urban <- WBnew %>%

  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"))
}
```