VIRGINIA COMMONWEALTH UNIVERSITY

Statistical analysis and modelling (SCMA 632)

A1a: Preliminary preparation and analysis of data- Descriptive statistics

AAKASH KATHIRVEL

V01110153

Date of Submission: 16-06-2024

CONTENTS

INTRODUCTION:

The aim of this research is to explore the state of Meghalaya using data from the National Sample Survey Office (NSSO). The focus is to pinpoint the districts with the highest and lowest consumption rates within the state. To accomplish this, we will thoroughly clean and process the dataset to extract relevant information for analysis. The dataset contains consumptionrelated information for both rural and urban sectors, as well as district-level variations. This data has been imported into R, a powerful and flexible statistical programming language, wellsuited for handling and analyzing extensive datasets.

Our goals for this research include detecting and correcting missing values, managing outliers, standardizing district and sector names, summarizing consumption data by region and district, and examining the significance of differences in mean consumption. The findings from this research will provide critical insights for policymakers and stakeholders, supporting targeted initiatives and promoting balanced development throughout Meghalaya.

OBJECTIVES:

1) To 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) To 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 among the different sectors.

RESULTS AND INTERPRETATION:

1) CHECK FOR MISSING VALUES:

```
Missing Values in Subset:
> print(colSums(is.na(megData)))
     state_1     District       Region       Sector State_Region Meals_At_Home     ricepds_v   Wheatpds_q    chicken_q      pulsep_q
           0            0            0            0            0            8            0            0            0            0
  wheatos_q No_of_Meals_per_day
           0                  1
```

INTERPRETATION: From the selected variables, after sorting the data for the state of Meghalaya, the column 'No_of_Meals_per_day' has 1 missing values and 'Meals_At_Home' has 8 missing values.

```
MEG.isnull().sum().sort_values(ascending = False)
✓ 0.0s

Meals_At_Home        8
state_1              0
District             0
Sector               0
Region               0
State_Region         0
ricetotal_q          0
wheattotal_q         0
moong_q              0
Milktotal_q          0
chicken_q            0
bread_q              0
foodtotal_q          0
Beveragestotal_v     0
dtype: int64
```

Removing missing values ensures accurate and unbiased estimates by preventing misleading conclusions. Many statistical tests, including z-tests, assume complete datasets; missing values can violate these assumptions and lead to incorrect results. Additionally, most statistical functions cannot handle missing values, which can cause errors or unreliable outputs during analysis.

2) HANDLING MISSING VALUES

```
Missing Values After Imputation:
> print(colSums(is.na(megData)))
     state_1     District       Region       Sector State_Region Meals_At_Home     ricepds_v   Wheatpds_q    chicken_q      pulsep_q
           0            0            0            0            0            0            0            0            0            0
  wheatos_q No_of_Meals_per_day
           0                  0
```

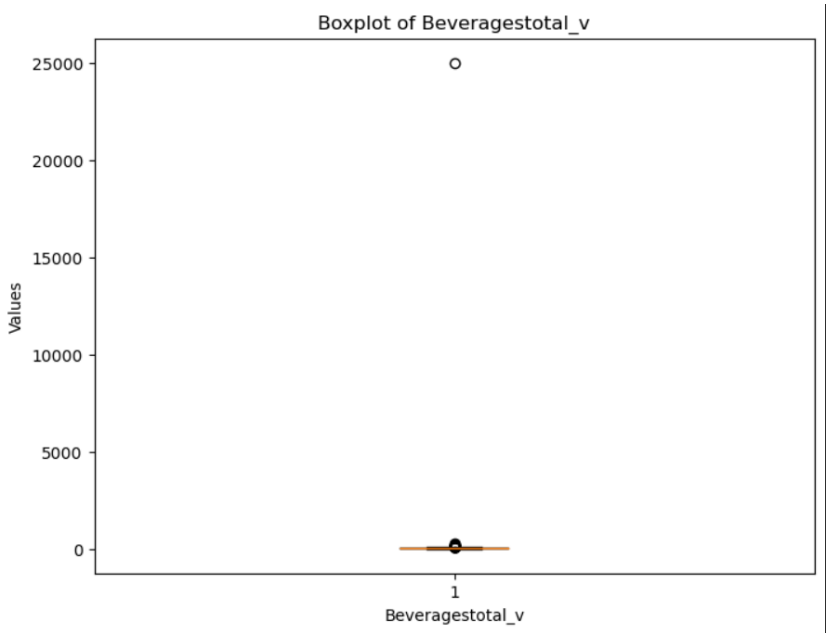
```
#CHECKING FOR MISSING VALUES AFTER IMPUTATION
new_var = MEG_clean.isnull().sum().sort_values(ascending = False)
new_var
```

✓ 0.0s

```
state_1             0
District            0
Sector              0
Region              0
State_Region        0
ricetotal_q         0
wheattotal_q        0
moong_q             0
Milktotal_q         0
chicken_q           0
bread_q             0
foodtotal_q         0
Beveragestotal_v    0
Meals_At_Home       0
dtype: int64
```

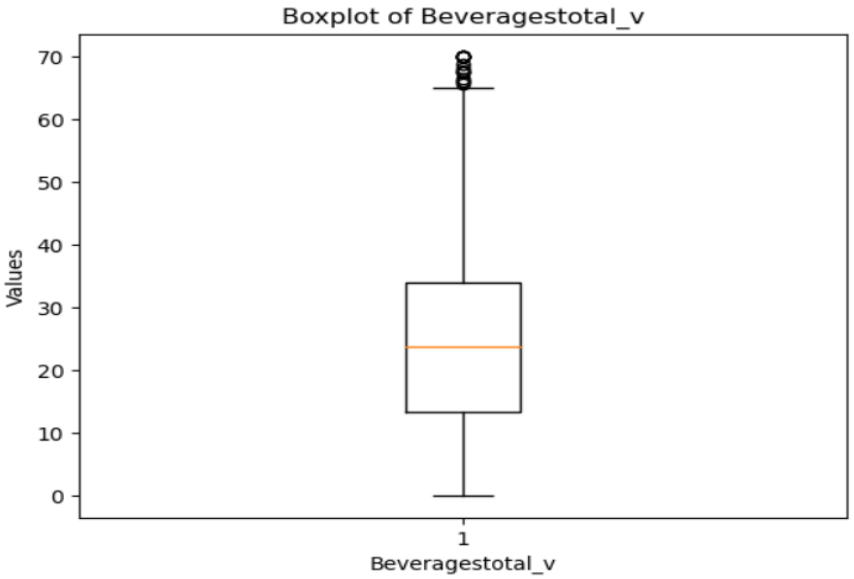
INTERPRETATION: The above code has successfully replaced the missing values with the mean value of the variable. As can be seen from the result above, there are no missing values in the selected data.

3) CHECK FOR OUTLIERS



INTERPRETATION: From the boxplot above, which is a visual representation of the variable 'Beveragestotal_v' shows that there is outliers. Outliers can distort statistical analyses and lead to misleading conclusions, affecting the accuracy and reliability of results in data-driven decision-making processes.

4) SETTING QUARTILES AND REMOVING OUTLIERS:



INTERPRETATION: 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 and can be excluded or treated to ensure the robustness of the analysis. In the similar way the outliers in all other variables can be removed

5) RENAMING AND DISPLAYING TOP 3 & LAST 3 DISTRICTS OF CONSUMPTION:

```
Top 3 Consuming Districts:
> print(head(district_summary, 3))
# A tibble: 3 x 2
  District         total
  <chr>            <dbl>
1 East Khasi Hills 6648.
2 Jaintia Hills    4545.
3 West Khasi Hills 4007.
```

INTERPRETATION: The top three consuming districts are East Khasi Hills with 6648 units, followed by Jaintia Hills with 4545 units, and West Khasi Hills with 4007 units. Similarly the bottom three districts can be found by sorting the total consumption.

```
Bottom 3 Consuming District
> print(tail(district_summa
# A tibble: 3 x 2
  District         total
  <chr>            <dbl>
1 Ri Bhoi          2381.
2 East Garo Hills  2181.
3 South Garo Hills 1573.
```

INTERPRETATION: The least consuming district is Ri Bhoi with 2381 units, East Garo Hills with 2181 units and South Garo Hills with 1573 units.

6) REGION CONSUMPTION SUMMARY:

7)

```
Region Consumption Summary:
> print(region_summary)
# A tibble: 2 x 2
  Sector  total
  <chr>   <dbl>
1 RURAL  20492.
2 URBAN   4628.
```

INTERPRETATION:

Rural sector consists of 20492 units and Urban sector consists of 4628 units.

6) TEST WHETHER THE DIFFERENCES IN THE MEANS AMONG THE SECTORS ARE SIGNIFICANT OR NOT.

```
> cat("Z STATISTIC:\n")
Z STATISTIC:
> print(z_test_result$statistic)
       z
9.59687
> print(z_test_result$method)
[1] "Two-sample z-Test"
> cat(glue::glue("P value is :{z_test_result$p.value}"))
P value is :8.24086100214657e-22
```

P value is < 0.05 :Therefore we reject the null hypothesis.which means there is a significant difference between mean consumptions of urban and rural.The mean consumption in Rural areas is 24.929148362317 5 and in Urban areas its 12.0201144678442

```
    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)
```

✓ 0.0s

```
Z-Score: 2.006802685776415
P-Value: 0.04477067678874531
```

```
    # H1: There is a significant difference between mean consumptions of urban and rural sectors
    # Ho: There is no significant difference between mean consumptions of urban and rural sectors

    # Checking p-value against significance level(0.05)
    if p_value < 0.05:
        print("Reject Ho: There is a significant difference between mean consumptions of urban and rural sectors.")
    else:
        print("Fail to reject Ho: There is no significant difference between mean consumptions of urban and rural sectors.")
```

✓ 0.0s

```
Reject Ho: There is a significant difference between mean consumptions of urban and rural sectors.
```

INTERPRETATION: The two-sample z-test indicates a highly significant difference in consumption between rural and urban sectors. Rural consumption is notably higher than Urban consumption.

CODES:

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

# Reading the file into R
data <- read.csv("C:/Users/Aakash/Desktop/SCMA/NSSO68.csv")
# Display the first few rows of the data
head(data)

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

# 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
megData <- 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(megData)))

# (1) HANDLING MISSING VALUES
# 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)
}
megData$Meals_At_Home <- impute_with_mean(megData$Meals_At_Home)
megData$No_of_Meals_per_day <- impute_with_mean(megData$No_of_Meals_per_day)

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

# (2)CHECK FOR OUTLIERS
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) {
  megData <- remove_outliers(megData, col)
}

# (4)RENAME DISTRICTS AND SECTORS USING CODES FROM APPENDIX OF NSSA 68TH ROUND DATA
district_mapping <- c("6" = "East Khasi Hills" , "7" = "Jaintia Hills" ,"4" = "West Khasi Hills" , "1" = "West
Garo Hills" , "5" = "Ri Bhoi" , "2" = "East Garo Hills", "3" = "South Garo Hills")
```

```
sector_mapping <- c("2" = "URBAN", "1" = "RURAL")

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


# (5)SUMMARIZING VARIABLES
megData$total_consumption <- rowSums(megData[, 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 <- megData %>%
    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("Sector")

# (6) DISPLAYING TOP AND BOTTOM 3 DISTRICTS OF CONSUMPTION
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)

# (7) TEST FOR DIFFERENCES IN MEAN CONSUMPTION AMONG RURAL AND URBAN
rural <- megData %>%
  filter(Sector == "RURAL") %>%
  select(total_consumption)

urban <- megData %>%
  filter(Sector == "URBAN") %>%
  select(total_consumption)


mean_rural <- mean(rural$total_consumption)
mean_urban <- mean(urban$total_consumption)

# Z-TEST :
z_test_result <- BSDA::z.test(rural$total_consumption, urban$total_consumption,
                      alternative = "two.sided", mu = 0, sigma.x = sd_rural, sigma.y = sd_urban, conf.level =
0.95)
```

```
cat("Z STATISTIC:\n")
print(z_test_result$statistic)
print(z_test_result$method)

cat(glue::glue("P value is :{z_test_result$p.value}"))

# (8) OUTPUT BASED ON P VALUE OBTAINED
if (z_test_result$p.value < 0.05) {
  cat(glue::glue("P value is < 0.05 :Therefore we reject the null hypothesis.\n"))
  cat(glue::glue("Which means there is a significant 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: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"))
}
```

CONCLUSION:

This study investigates consumption patterns in Meghalaya using data from the NSSO. Employing Python and R for data cleaning, summarization, and statistical testing, we aim to provide insights that will aid policymakers in fostering equitable development and implementing targeted interventions across districts. This analysis is crucial for supporting informed decision-making and effective resource allocation. Our findings reveal a significant disparity in consumption between rural and urban sectors in Meghalaya. In this state, rural consumption is significantly higher than urban consumption, highlighting the need for tailored strategies to address the unique needs of these regions.