VIRGINIA COMMONWEALTH UNIVERSITY

Statistical analysis and modelling (SCMA 632)

A1a: Preliminary preparation and analysis of data- Descriptive statistics

FERAH SHAN SHANAVAS RABIYA

V01101398

Date of Submission: 16-06-2024

CONTENTS

Introduction

This report focuses on exploring consumption patterns across the districts of Maharashtra, distinguishing between rural and urban sectors. The data encapsulates crucial variables such as district-wise consumption levels, categorized into rural and urban sectors. This analysis delves into identifying missing values, addressing outliers, renaming sectors and districts for clarity, and conducting statistical tests to discern significant differences in consumption between regions. The dataset has been imported into R, a powerful statistical programming language renowned for its versatility in handling and analyzing large datasets.

The 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. Through these methodologies, I aim to provide a comprehensive understanding of consumption dynamics within Maharashtra, highlighting key insights that can inform policy and development strategies tailored to the state's diverse communities.

Objectives

- Examine for any missing values in the data, identify them, and replace them with the mean of the variable.
- Check for outliers, explain your test's results, and make the necessary adjustments.
- Rename the sectors—rural and urban—as well as the districts.
- List the top three and lowest three districts of consumption, as well as a summary of the important factors in the data set by region and district.
- Determine whether or not the mean differences are significant.

Business Significance

The focus of this study on Maharashtra's purchasing patterns using NSSO data has significant implications for decision-makers in business and government. The study offers important insights for market entry, resource allocation, supply chain optimization, and focused interventions by identifying the top and bottom three consuming districts. By means of data cleaning, outlier detection, and significance testing, the results aid in making well-informed decisions that promote fair development and propel Maharashtra's economic expansion.

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.**

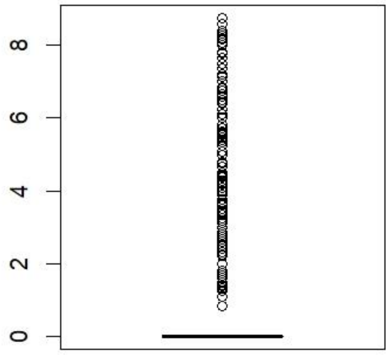
```
#Identifying the missing values
> any(is.na(mh_df))
[1] TRUE

> sum(is.na(mh_df))
[1] 93676

# Sub-setting the data
> mhnew <- mh_df %>%
Select (state_1, District, Region, Sector, State_Region, Meals_At_Hom
e, ricepds_v, Wheatpds_q, chicken_q, pulsep_q, wheatos_q, No_of_Meals
_per_day)
> print(colSums(is.na(mhnew)))
state_1            District
                0                 0
           Region            Sector
                0                 0
     State_Region      Meals_At_Home
                0               184
        ricepds_v        Wheatpds_q
                0                 0
        chicken_q          pulsep_q
                0                 0
        wheatos_q No_of_Meals_per_day
                0                 0

# 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)
  }
  > mhnew$Meals_At_Home <- impute_with_mean(mhnew$Meals_At_Home)
```

Interpretation:

After sorting the data for the state of Maharashtra based on the variables that were chosen, it is evident that only the column "Meals_At_Home" has 184 missing variables. Since missing values in the dataset might cause biased or incomplete analyses, which can skew interpretations and decision-making processes and impair the quality of outcomes, they can be troublesome. Consequently, we use the following code to replace the missing values with the variable's mean. The above code has successfully replaced the missing values with the mean value of the variable. After this, there will be no missing values in the selected data.

- **Check for outliers, describe your test’s outcome, and make suitable amendments.**

```
#Checking for Outliers
  > boxplot(mhnew$ricepds_v)
```



Interpretation:

There is an outlier, as can be seen in the boxplot above, which represents the variable "ricepds_v" visually. Outliers can skew statistical analysis and provide false conclusions, which impairs the dependability and accuracy of findings in systems that use data to make decisions. In data-driven decision-making processes, outliers can skew statistical studies and produce false conclusions, which can impair the precision and dependability of results. The code below can be used to eliminate the outliers.

```
#Setting quartile ranges for removing outliers
> remove_outliers <- function(mh_df, column_name) {
   Q1 <- quantile(mh_df[[column_name]], 0.25)
   Q3 <- quantile(mh_df[[column_name]], 0.75)
   IQR <- Q3 - Q1
   lower_threshold <- Q1 - (1.5 * IQR)
   upper_threshold <- Q3 + (1.5 * IQR)
   mh_df <- subset(mh_df, mh_df[[column_name]] >= lower_threshold & m
h_df[[column_name]] <= upper_threshold)
   return(mh_df)
 }
> outlier_columns <- c("ricepds_v", "chicken_q")
> for (col in outlier_columns) {
   mhnew <- remove_outliers(mhnew, col)
 }
```

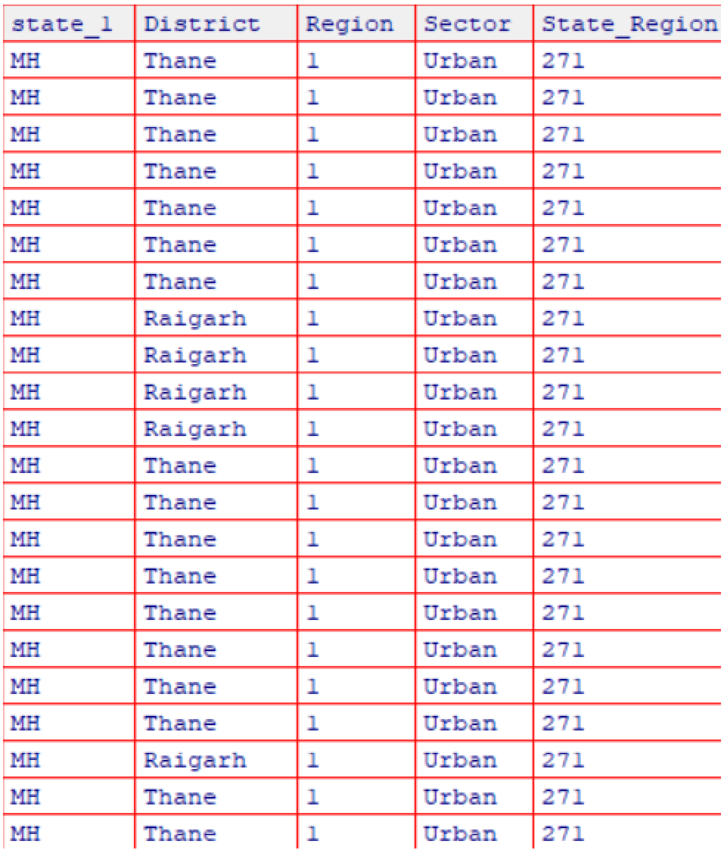
Interpretation:

It is possible to identify and eliminate outliers by interpreting quartile ranges. Data points that are more than 1.5 times the interquartile range (IQR) from either quartile are considered outliers and can be removed or handled in order to maintain the analysis's robustness. The IQR is computed as the difference between the upper and lower quartiles.

− Rename the districts and sectors, viz., rural and urban.

In the NSSO of data, a unique number is assigned to each district in a state. The statistics must be accompanied by their individual names in order to comprehend and identify the state's highest-consuming districts. Likewise, the state's urban and rural areas have been placed to assignments 1 and 2, respectively. To accomplish this, execute the subsequent code.

```
# Renaming districts and sectors
> district_mapping <- c( "1" = "Nandurbar", "2" = "Dhule", "3" = "J
algaon", "4" = "Buldana", "5" = "Akola", "6" = "Washim", "7" = "Amr
avati", "8" = "Wardha", "9" = "Nagpur", "10" = "Bhandara", "11" = "
Gondiya", "12" = "Gadchiroli", "13" = "Chandrapur", "14" = "Yavatma
l", "15" = "Nanded", "16" = "Hingoli", "17" = "Parbhani", "18" = "J
alna", "19" = "Aurangabad", "20" = "Nashik", "21" = "Thane", "22" =
"Mumbai", "24" = "Raigarh", "25" = "Pune", "26" = "Ahmadnagar", "27
" = "Bid", "28" = "Latur", "29" = "Osmanabad", "30" = "Solapur", "3
1" = "Satara", "32" = "Ratnagiri", "33" = "Sindhudurg", "34" = "Kol
hapur", "35" = "Sangli" )
> sector_mapping <- c("1" = "Rural", "2" = "Urban")
```



state_1	District	Region	Sector	State_Region
MH	Thane	1	Urban	271
MH	Thane	1	Urban	271
MH	Thane	1	Urban	271
MH	Thane	1	Urban	271
MH	Thane	1	Urban	271
MH	Thane	1	Urban	271
MH	Thane	1	Urban	271
MH	Raigarh	1	Urban	271
MH	Raigarh	1	Urban	271
MH	Raigarh	1	Urban	271
MH	Raigarh	1	Urban	271
MH	Thane	1	Urban	271
MH	Thane	1	Urban	271
MH	Thane	1	Urban	271
MH	Thane	1	Urban	271
MH	Thane	1	Urban	271
MH	Thane	1	Urban	271
MH	Thane	1	Urban	271
MH	Thane	1	Urban	271
MH	Raigarh	1	Urban	271
MH	Thane	1	Urban	271
MH	Thane	1	Urban	271

Interpretation:

The district names have been successfully assigned to the specified number in the result as it is seen above. Also, the sectors 1 and 2 have been replaced as urban and rural sectors respectively.

– Summarize the critical variables in the data set region-wise and district-wise and indicate the top and bottom three districts of consumption.

```
# Summarize consumption
> mhnew$total_consumption <- rowSums(mhnew[, c("ricepds_v", "wheatpds_q",
"chicken_q", "pulsep_q", "wheatos_q")], na.rm = TRUE)
> 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))
  District total
1       Mumbai  2281.
2       Pune    2157.
3       Thane   1919.

> cat("Bottom 3 Consuming Districts:\n")
Bottom 3 Consuming Districts:
> print(tail(district_summary, 3))
  District total
1       Bhandara    212.
2       Gondiya     204.
3       Gadchiroli  202.
>
> cat("Region Consumption Summary:\n")
Region Consumption Summary:
> print(region_summary)
  Region total
1      Akola     7382.
2      Dhule     7374.
3      Buldana   6554.
4      Nandurbar 5197.
5      Jalgaon   3597.
6      Washim    1055.
```

Interpretation:

The top three consuming districts are Mumbai with 2281 units, Pune with 2157 units and Tha ne with 1919 units. Similarly, the bottom three consuming districts are Bhandara with 212 uni ts, Gondiya with 204 units and Gadchiroli with 202 units.

- **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.

```
# Test for differences in mean consumption between urban a
nd 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 consumpti
ons 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 m
ean 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:

P value is < 0.05 i.e. 7e-05, Therefore we reject the null hypothesis. There is a difference between mean consumptions of urban and rural. The mean consumption in Rural areas is 4.64627249850332 and in Urban areas its 4.37012249728655.

Recommendation

Policy implications should consider unique consumption patterns in urban and rural areas, addressing disparities in economic conditions or resource accessibility. Further research should explore the causes of variations in consumption, such as infrastructure, cultural preferences, economic standing, and commodity accessibility. Resource distribution should be adjusted to address these patterns, involving marketing methods, infrastructural improvements, or distribution strategies. Long-term monitoring should be conducted to evaluate the success of changes or policies, ensuring that actions are still applicable and productive. By addressing these factors, policymakers can better address the unique needs of urban and rural populations.

Codes

```
# Set the working directory and verify it
setwd("E:\\VCU\\Summer 2024\\Statistical Analysis & Modeling")
getwd()

# 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)

# Read the file into R
data <- read.csv("NSSO68.csv")

# Filter for Maharashtra
mh_df <- data %>% filter(state_1 == "MH")

# Display dataset info
cat("Dataset Information:\n")
print(names(mh_df))
print(head(mh_df))
print(dim(mh_df))

# Finding missing values
missing_info <- colSums(is.na(mh_df))
cat("Missing Values Information:\n")
print(missing_info)
```

```
any(is.na(mh_df))
sum(is.na(mh_df))

# Sub-setting the data
mhnew <- mh_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(mhnew)))

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

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

# Find outliers and removing them
boxplot(mhnew$ricepds_v)
remove_outliers <- function(mh_df, column_name) {
  Q1 <- quantile(mh_df[[column_name]], 0.25)
  Q3 <- quantile(mh_df[[column_name]], 0.75)
  IQR <- Q3 - Q1
  lower_threshold <- Q1 - (1.5 * IQR)
  upper_threshold <- Q3 + (1.5 * IQR)
  mh_df <- subset(mh_df, mh_df[[column_name]] >= lower_threshold & mh_df[[column_name]] <= upper_threshold)
  return(mh_df)
}


outlier_columns <- c("ricepds_v", "chicken_q")
for (col in outlier_columns) {
  mhnew <- remove_outliers(mhnew, col)
}

# Renaming districts and sectors
district_mapping <- c( "1" = "Nandurbar", "2" = "Dhule", "3" = "Jalgaon", "4" = "Buldana", "5" = "Akola", "6" = "Washim", "7" = "Amravati", "8" = "Wardha", "9" = "Nagpur", "10" = "Bhandara", "11" = "Gondiya", "12" = "Gadchiroli", "13" = "Chandrapur", "14" = "Yavatmal", "15" = "Nanded", "16" = "Hingoli", "17" = "Parbhani", "18" = "Jalna", "19" = "Aurangabad", "20" = "Nashik", "21" = "Thane", "22" = "Mumbai", "24" = "Raigarh", "25" = "Pune", "26" = "Ahmadnagar", "27" = "Bid", "28" = "Latur", "29" = "Osmanabad", "30" = "Solapur", "31" = "Satara", "32" = "Ratnagiri", "33" = "Sindhudurg", "34" = "Kolhapur", "35" = "Sangli" )
sector_mapping <- c("1" = "Rural", "2" = "Urban")

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)

# Summarize consumption
mhnew$total_consumption <- rowSums(mhnew[, c("ricepds_v", "Wheatpds_q", "chicken_q", "pulsep_q", "wheatos_q")], na.rm = TRUE)

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)



# 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"))
}
```

References

1. www.github.com

2. www.geeksforgeeks.com