VIRGINIA COMMONWEALTH UNIVERSITY

Statistical analysis and modelling (SCMA 632)

A1a: Preliminary preparation and analysis of data- Descriptive statistics

JYOTHIS KANIYAMPARAMBIL THANKACHAN
V01110144

Date of Submission: 16-06-2024

CONTENTS

Introduction

This study focuses on the state of Manipur using NSSO data to identify the top and bottom three districts in terms of consumption. To obtain the necessary data for analysis, I clean up and alter the dataset during this procedure. In order to make this study easier, I have assembled a dataset. include statistics on consumption, district-specific variances, and data on the rural and urban sectors. R, a potent statistical programming language well-known for its adaptability in managing and analysing big datasets, has been used to import the dataset.

Identifying missing values, dealing with outliers, standardising district and sector names, district- and regional-level summaries of consumption data, and determining the significance of mean differences are some of our goals. Policymakers and other stakeholders can benefit from the study's results, which can support targeted actions and fair development throughout the state.

OBJECTIVES

- Check if there are any missing values in the data, identify them, and if there are, replace them with the mean of the variable
- Check for outliers, describe your test’s outcome, and make suitable amendments.
- Rename the districts and sectors, viz., rural and urban.
- Summarize the critical variables in the data set region-wise and district-wise and indicate the top and bottom three districts of consumption.
- Test whether the differences in the means are significant or not.

BUSINESS SIGNIFICANCE

The investigation's emphasis on Manipur consumption habits using NSSO data has important ramifications for decision-makers in industry and government. The study offers important insights for market entrance, resource allocation, supply chain optimisation, and focused interventions by identifying the top and bottom three consuming districts. The results support educated decision-making, equitable development, and Manipur's economic progress through data cleansing, outlier detection, and significance testing.

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.

Yes, there are some missing value in the data.

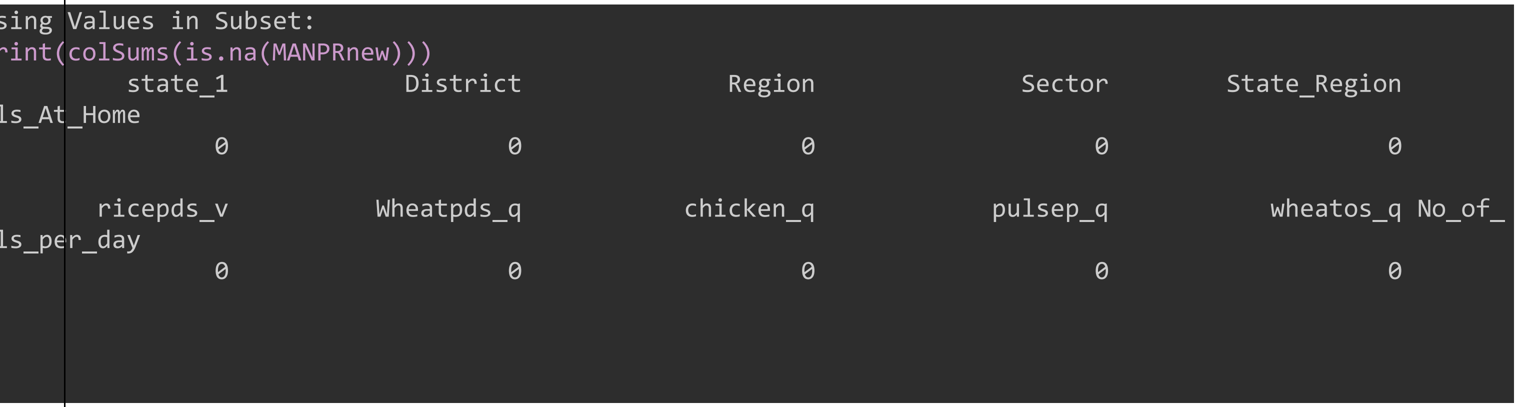
Code:

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

# Sub-setting the data for varibles of interes
```

MANPRnew <- 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(MANPRnew))

Result:



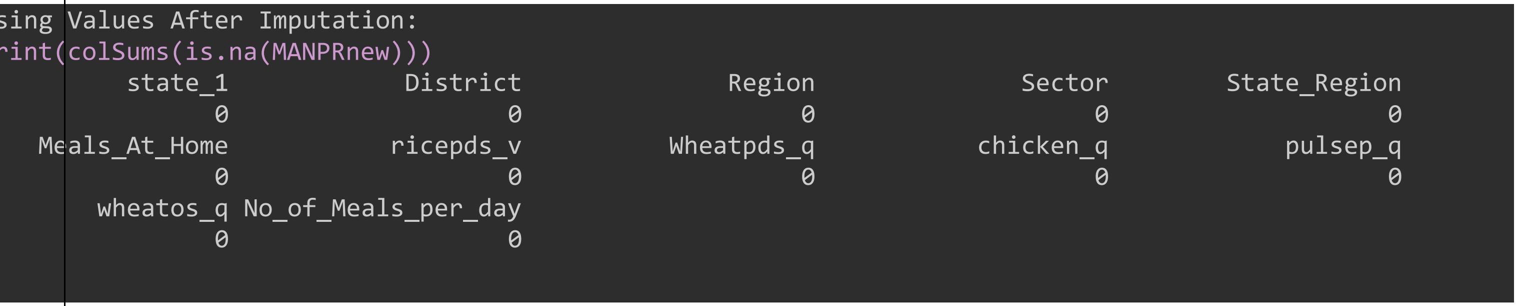
```
sing Values in Subset:
rint(colSums(is.na(MANPRnew)))
        state_1            District              Region              Sector        State_Region 
ls_At_Home 
              0                   0                   0                   0                   0 

      ricepds_v          Wheatpds_q           chicken_q            pulsep_q           wheatos_q No_of_
ls_per_day 
              0                   0                   0                   0                   0 
```

Founded meals at home there are 7 missing values

Replace the missing value with the mean of the variable.
Code:

```
# 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)
}
MANPRnew$Meals_At_Home <- impute_with_mean(MANPRnew$Meals_At_Home)

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

Result:



```
sing Values After Imputation:
rint(colSums(is.na(MANPRnew)))
        state_1            District              Region              Sector        State_Region 
              0                   0                   0                   0                   0 
  Meals_At_Home           ricepds_v          Wheatpds_q           chicken_q            pulsep_q 
              0                   0                   0                   0                   0 
      wheatos_q No_of_Meals_per_day 
              0                   0 
```

Interpretation: Based on the variables chosen, the data for the state of Manipur is sorted, and the only column with seven missing values is "Meals_At_Home." Since missing values in the dataset might cause biased or incomplete analyses, which can distort interpretations and decision-making processes and impair the accuracy of outcomes, they can be troublesome. Consequently, we use the following code to replace the missing values with the variable's mean.

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

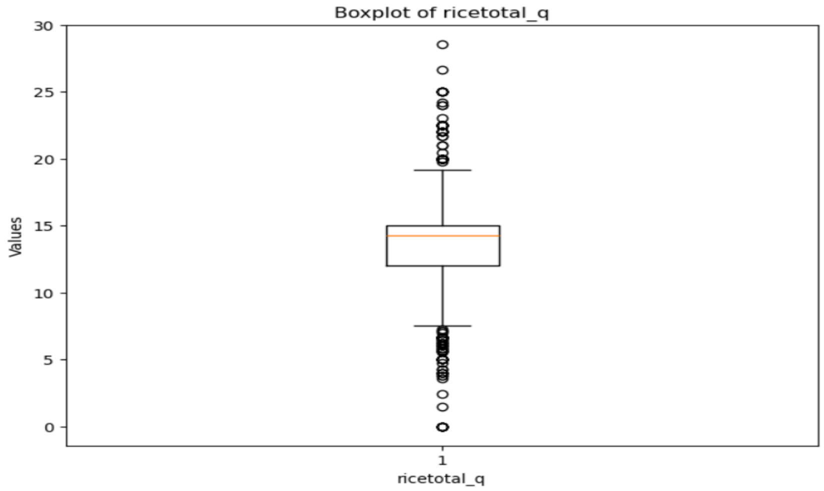
Code:

```
# 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) {
  MANPRnew <- remove_outliers(MANPRnew, col)
}
```

Code used in Python:

```
import matplotlib.pyplot as plt
# Assuming MANPR_clean is your DataFrame
plt.figure(figsize=(8, 6))
plt.boxplot(MANPR_clean['ricetotal_q'])
plt.xlabel('ricetotal_q')
plt.ylabel('Values')
plt.title('Boxplot of ricetotal_q')
plt.show(
```

Result :



.

Interpretation: There is an outlier, as can be seen in the boxplot above, which represents the variable "ricetotal_q" visually. In data-driven decision-making processes, outliers can skew statistical studies and provide false conclusions, which can impair the precision and dependability of outcomes. In data-driven decision-making processes, outliers can skew statistical studies and provide false conclusions, which can impair the precision and dependability of outcomes. The following can be used to eliminate the outliers.

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

In the NSSO of data, a unique number is issued to each district in a state. The statistics must be accompanied with their individual names in order to comprehend and identify the state's highest-consuming districts. Likewise, the state's urban and rural areas were assigned to assignments 1 and 2, respectively. Running is how this is accomplished the code that follows.

Code:

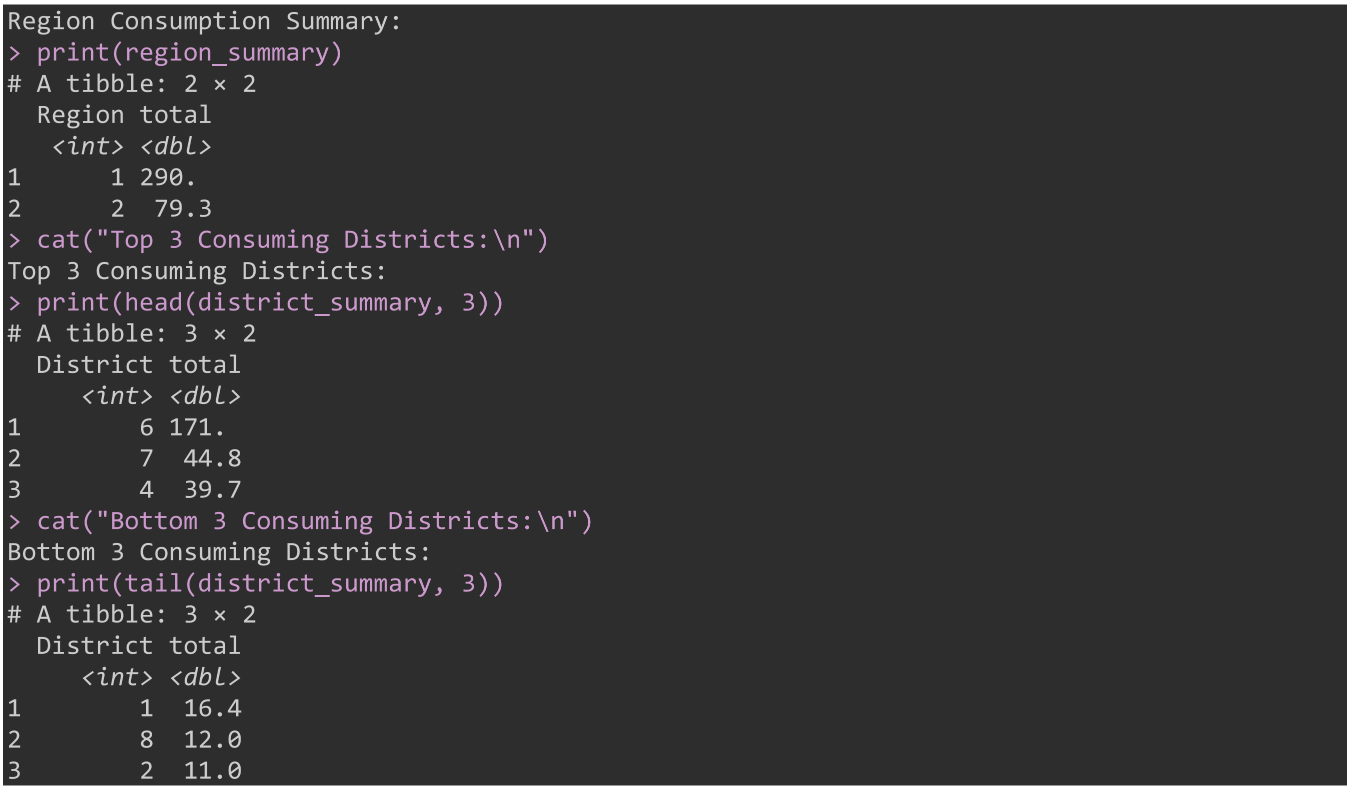
code:

```
# Rename districts and sectors , get codes from appendix of NSSO 68th ROund Data
district_mapping <- c("06" = "Imphal West", "07" = "Imphal East", "04" = "Bishnupur")
sector_mapping <- c("2" = "URBAN", "1" = "RURAL")
MANPRnew$District <- as.character(MANPRnew$District)
MANPRnew$Sector <- as.character(MANPRnew$Sector)
MANPRnew$District <- ifelse(MANPRnew$District %in% names(district_mapping),
district_mapping[MANPRnew$District], MANPRnew$District)
MANPRnew$Sector <- ifelse(MANPRnew$Sector %in% names(sector_mapping),
sector_mapping[MANPRnew$Sector], MANPRnew$Sector)

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

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

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

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 and display top and bottom consuming districts and regions
summarize_consumption <- function(group_col) {
summary <- MANPRnew %>%
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)
```

Result:



```
Region Consumption Summary:
> print(region_summary)
# A tibble: 2 × 2
  Region total
   <int> <dbl>
1      1 290. 
2      2  79.3
> cat("Top 3 Consuming Districts:\n")
Top 3 Consuming Districts:
> print(head(district_summary, 3))
# A tibble: 3 × 2
  District total
     <int> <dbl>
1        6 171. 
2        7  44.8
3        4  39.7
> cat("Bottom 3 Consuming Districts:\n")
Bottom 3 Consuming Districts:
> print(tail(district_summary, 3))
# A tibble: 3 × 2
  District total
     <int> <dbl>
1        1  16.4
2        8  12.0
3        2  11.0
```

We may estimate the top three and bottom three consuming districts by adding together all of the important variables to get total consumption.

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

Code :

```
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"))
}
```

Result :

```
P value is >= 0.05 i.e. 0.45698, Therefore we fail to reject the null hypothesis.There is no significant difference between mean consumptions of urban and rural.The mean consumption in Rural area is 0.123707407475369 and in Urban area its 0.199510810648995
```

Interpretation:

Based on the analysis, the p-value of 0.45698 indicates that there is no statistically significant difference between the mean consumptions of urban and rural areas, as it exceeds the common significance threshold of 0.05. Therefore, we fail to reject the null hypothesis, concluding that the observed difference in mean consumptions—0.1237 in rural areas and 0.1995 in urban areas—is not significant. This suggests that, given the current data, the mean consumption levels in urban and rural areas are similar.

Whole code of R

```
# Set the working directory and verify it
setwd('D:\\Assignments_SCMA632')
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("NSSO68.csv")

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

# 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
MANPRnew <- 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(MANPRnew)))
```

```
# 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)
}
apnew$Meals_At_Home <- impute_with_mean(apnew$Meals_At_Home)

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

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

# Summarize consumption
MANPRnew$total_consumption <- rowSums(MANPRnew[, 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 <- MANPRnew %>%
    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(("06" = "Imphal West", "07" = "Imphal East", "04" = "Bishnupur")
sector_mapping <- c("2" = "URBAN", "1" = "RURAL")

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


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

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

Reference :

Data camp Website: https://www.datacamp.com/tutorial/r-studio-tutorial.
Open AI