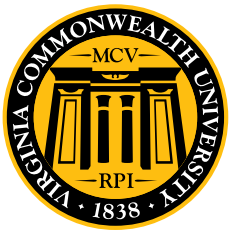


VIRGINIA COMMONWEALTH UNIVERSITY

Statistical analysis and modelling (SCMA 632)

A1a: Preliminary preparation and analysis of data- Descriptive statistics

RAKSHITHA VIGNESH SARGURUNATHAN

V01109007

Date of Submission: 16-06-2024

CONTENTS

INTRODUCTION:

The study focuses on the state Bihar, using NSSO68 data to identify the top and bottom three consuming districts within the state. The dataset includes consumption-related information, with distinctions between rural and urban sectors, as well as district-wise variations. Utilizing Python and R for analysis, it begin by importing the dataset into Excel for a structured format. The approach involves identifying and treating missing values and outliers, renaming districts and sectors for consistency, and summarizing key variables regionally and district-wise. Beyond data cleaning, it also aim to derive insights that guide policymakers and stakeholders. By testing the significance of mean differences, it provides a comprehensive overview of consumption patterns in Bihar. These insights support targeted interventions and equitable development, contributing to informed decision-making and resource allocation.

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(BiharData)))
            state_1            District              Region              Sector
                  0                   0                   0                   0
       State_Region       Meals_At_Home           ricepds_v          Wheatpds_q
                  0                  20                   0                   0
          chicken_q           pulsep_q          wheatos_q No_of_Meals_per_day
                  0                   0                   0                   4
```

```
Bhr.isnull().sum().sort_values(ascending = False)

Meals_At_Home       20
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
```

INTERPRETATION: From the selected variables, after sorting the data for the state of Bihar, the column 'No_of_Meals_per_day' has 4 missing values and 'Meals_At_Home' has 20 missing values. Since 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, missing values has to be handled.

2) HANDLING MISSING VALUES

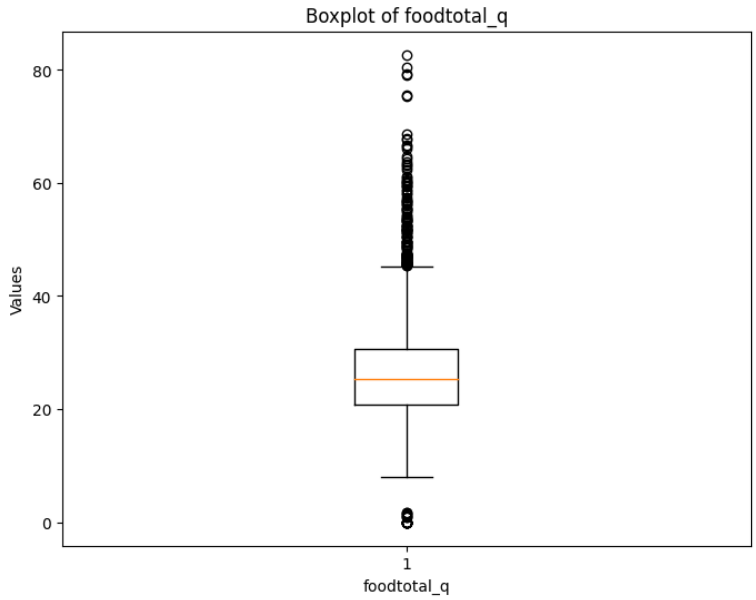
```
Missing Values After Imputation:
> print(colSums(is.na(BiharData)))
            state_1            District              Region              Sector
                  0                   0                   0                   0
       State_Region       Meals_At_Home           ricepds_v         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

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

```
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
```

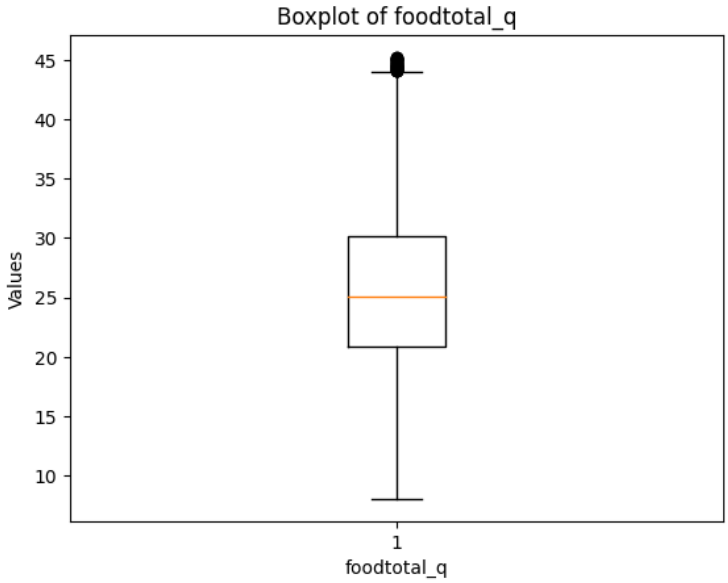
INTERPRETATION: The above code has successfully replaced the missing values with the mean value of the variable. As can be seen from the result above, there are no missing values in the selected data.

3) CHECK FOR OUTLIERS



INTERPRETATION: From the boxplot above, which is a visual representation of the variable 'foodtotal_q' 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 × 2
  District    total
  <chr>       <dbl>
1 Muzaffarpur 2068.
2 Madhubani   2013.
3 Patna       1899.
```

INTERPRETATION: The top three consuming districts are Muzaffarpur with 2068 units, followed by Madhubani with 2013 units, and Patna with 1899 units. Similarly the bottom three districts can be found by sorting the total consumption.

```
Bottom 3 Consuming Districts:
> print(tail(district_summary, 3))
# A tibble: 3 × 2
  District   total
  <chr>      <dbl>
1 Nawada      762.
2 Arwal       754.
3 Aurangabad  731.
```

INTERPRETATION: The least consuming district is Aurangabad with 731 units, Arwal with 754 units and Nawada with 762 units.

6) REGION CONSUMPTION SUMMARY:

```
Region Consumption Summary:
> print(region_summary)
# A tibble: 2 × 2
  Region  total
  <chr>   <dbl>
1 RURAL  30447.
2 URBAN  18244.
```

INTERPRETATION:

Rural sector consists of 30447 units and Urban sector consists of 18244 units.

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

```
> cat("Z STATISTIC:")
Z STATISTIC:
> z_test_result$statistic
       z
33.21954
> z_test_result$method
[1] "Two-sample z-Test"
> cat(glue::glue("P value is :{z_test_result$p.value}"))
P value is :5.62269352709633e-242

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 sector.The mean consumption in Rural areas is 12.3443836749735 and in Urban areas its 9.58032322468237
>
```

```
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)
```

```
Z-Score: 8.33414798698796
P-Value: 7.80582434945592e-17
```

```
# 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.")
```

```
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:

```
#To set the workspace
setwd('E:\\VCU\\SCMA\\DATA')
#To check workspace
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)
  }}

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 Bihar data
df <- data %>%
  filter(state_1 == "Bhr")

# 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
BiharData <- 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(BiharData)))

# (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)
}
```

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

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

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

# (4)RENAME DISTRICTS AND SECTORS USING CODES FROM APPENDIX OF NSSA 68TH ROUND DATA
district_mapping <- c( "14" = "Muzaffarpur", "5" = "Madhubani", "28" = "Patna" ,"36" = "Nawada", "38" = "Arwal", "34" = "Aurangabad")
sector_mapping <- c("2" = "URBAN", "1" = "RURAL")

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

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

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

# (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))

region_summary$Region <- ifelse(region_summary$Region == 1, "RURAL", "URBAN")

cat("Region Consumption Summary:\n")
print(region_summary)


# (7) TEST FOR DIFFERENCES IN MEAN CONSUMPTION AMONG RURAL AND URBAN
rural <- BiharData %>%
```

```
  filter(Sector == "RURAL") %>%
  select(total_consumption)

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

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

# 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)

cat("Z STATISTIC:")
z_test_result$statistic
z_test_result$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 sector.\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 identifies consumption patterns in Bihar by analyzing NSSO data. Using Python and R for data cleaning, summarization, and statistical testing, we provide insights to guide policymakers in promoting equitable development and targeted interventions across districts, ultimately supporting informed decision-making and resource allocation. This study indicates a highly significant difference in consumption between rural and urban sectors in Bihar. In Bihar, Rural consumption is notably higher than in Urban Bihar.