VIRGINIA COMMONWEALTH UNIVERSITY

Statistical analysis and modelling (SCMA 632)

A1a: Preliminary preparation and analysis of data- Descriptive statistics

JANY BALASIVAN

V01108262

Date of Submission: 16-06-2024

CONTENTS

Analyzing Consumption in the State of Madhya Pradesh Using R

Introduction

The focus of this study is on the state of ODISHA, from the NSSO data, to find the top and bottom three consuming districts of ODISHA. In the process, we manipulate and clean the dataset to get the required data to analyze. 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 analyzing 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 ODISHA's 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 ODISHA's 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(ODnew)))
            state_1            District              Region              Sector
                  0                   0                   0                   0
       State_Region       Meals_At_Home           ricepds_v          Wheatpds_q
                  0                  46                   0                   0
          chicken_q           pulsep_q          wheatos_q No_of_Meals_per_day
                  0                   0                   0                   2
```

Interpretation: In the subset of the Data dataset, there are a few missing values that need attention. Specifically, the Meals_At_Homecolumn has 46 missing values, while the No_of_Meals_per_day column has 2 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. Addressing these missing values is essential for ensuring the completeness and accuracy of subsequent analyses. Appropriate strategies such as data imputation or removal of rows with missing values can be employed to handle these gaps in the data.

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

```
> # 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)
+ }
> ODnew$Meals_At_Home <- impute_with_mean(ODnew$Meals_At_Home)
> ODnew$No_of_Meals_per_day <- impute_with_mean(ODnew$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(ODnew)))
            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: After imputing the missing values, the subset of the dataset (ODnew) now shows no missing values across all selected columns. This includes state_1, District, Region, Sector, State_Region, Meals_At_Home, ricepds_v, Wheatp ds_q, chicken_q, pulsep_q, wheatos_q, and No_of_Meals_per_day. The successful imputation

ensures that the dataset is complete, which is crucial for accurate and reliable analysis. This means that any subsequent analysis can be conducted without concerns about missing data skewing the results.

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) {
+   ODnew <- remove_outliers(ODnew, col)
+ }
```

c) <u>Rename the districts as well as the sector, viz. rural and urban.</u>

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("15" = "anugul", "11" = "jagatsinghapur", "3" = "sambalpur")

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

>

> ODnew$District <- as.character(ODnew$District)

> ODnew$Sector <- as.character(ODnew$Sector)

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

> ODnew$Sector <- ifelse(ODnew$Sector %in% names(sector_mapping), sector_mapping[ODnew$Sector], ODnew$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 and display top and bottom consuming districts and regions
> summarize_consumption <- function(group_col) {
+   summary <- ODnew %>%
+     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))
```

```
Top 3 Consuming Districts:
> print(head(district_summary, 3))
# A tibble: 3 × 2
  District total
     <int> <dbl>
1       26 1533.
2       29 1401.
3       12 1246.
```

Interpretation: The output displays the top three consuming districts based on their total consumption values. The district with the highest consumption is District 26 with a total consumption of 1533 units. Following District 26, District 29 has a total consumption of 1401 units, and District 12 has a total consumption of 1246 units. This ranking indicates that District 26 is the highest consumer among the districts analyzed, followed by Districts 29 and 12.

```
cat("Bottom 3 Consuming Districts:\n")
print(tail(district_summary, 3))
```

```
> cat("Bottom 3 Consuming Districts:\n")
Bottom 3 Consuming Districts:
> print(tail(district_summary, 3))
# A tibble: 3 × 2
  District total
     <int> <dbl>
1       11  644.
2        2  521.
3        3  511.
```

Interpretation: The output lists the bottom three consuming districts based on their total consumption values. District 11 has a total consumption of 644 units, making it the highest among the bottom three. District 2 follows with a total consumption of 521 units, and District 3 has the lowest consumption with a total of 511 units. This ranking highlights Districts 11, 2, and 3 as the least consuming districts in the dataset, with District 3 having the minimal consumption.

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)
```

```
> # Test for differences in mean consumption between urban and rural
> rural <- ODnew %>%
+   filter(Sector == "RURAL") %>%
+   select(total_consumption)
>
> urban <- ODnew %>%
+   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 = 2.3
4, 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_urba
n}\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"))
+ }
P value is < 0.05 i.e. 0, Therefore we reject the null hypothesis.There is a difference between mean cons
umptions of urban and rural.The mean consumption in Rural areas is 8.01106868940729 and in Urban areas it
s 5.61696346197833
>
```

CODES

```
# Set the working directory and verify it
setwd('/Users/janybalashiva/Downloads')
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 OD
df <- Data %>%
  filter(state == "21")

# 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
ODnew <- 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(ODnew)))
```

```
# 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)
}
ODnew$Meals_At_Home <- impute_with_mean(ODnew$Meals_At_Home)
ODnew$No_of_Meals_per_day <- impute_with_mean(ODnew$No_of_Meals_per_day)

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

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

```
# Summarize consumption
ODnew$total_consumption <- rowSums(ODnew[, 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 <- ODnew %>%
    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("15" = "anugul", "11" = "jagatsinghapur", "3" = "sambalpur")
sector_mapping <- c("2" = "URBAN", "1" = "RURAL")

ODnew$District <- as.character(ODnew$District)
ODnew$Sector <- as.character(ODnew$Sector)
```

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

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



# Test for differences in mean consumption between urban and rural

rural <- ODnew %>%

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

  select(total_consumption)


urban <- ODnew %>%

  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"))

}
```