VIRGINIA COMMONWEALTH UNIVERSITY

STATISTICAL ANALYSIS & MODELING

A1a: CONSUMPTION PATTERN OF KERALA USING PYTHON AND R

Kirthan Shaker Iyangar
V01072700

Date of Submission: 16/06/2024

CONTENTS

Analyzing Consumption in the State of Kerala Using R

INTRODUCTION

The focus of this study is on the state of Kerala, from the NSSO data, to find the top and bottom three consuming districts of Kerala. 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.

f) * Use the dataset [data "NSSO68.csv"]

BUSINESS SIGNIFICANCE

The focus of this study on Kerala'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 Kerala's economic growth.

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.

Code and Result:

```
any(is.na(kenew))
[1] TRUE
> sum(is.na(kenew))
[1] 104
> sort(colSums(is.na(kenew)),decreasing=T)
   Meals_At_Home            state_1            District              Region
Sector
             104                  0                   0                   0
0
    State_Region          ricepds_v          Wheatpds_q           chicken_q
pulsep_q
               0                  0                   0                   0
0
       wheatos_q No_of_Meals_per_day
               0                   0
```

Interpretation: From the selected variables, after sorting the data for the state of Kerala, it is seen that only the column ‘Meals_At_Home has 104 missing variables. 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 we replace the missing values with the mean of the variable using following code.

#Subesetting the Data

Code and Result:

Before Imputation

```
> kenew <- 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")
Missing Values in Subset:
> print(colSums(is.na(kenew)))
      state_1       District        Region        Sector     State_Region     Meals_At_Home
```

```
ricepds_v
          0          0          0          0          0        104          0
   Wheatpds_q     chicken_q     pulsep_q     wheatos_q No_of_Meals_per_day
          0          0          0          0          0
```

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

Code and Result: (After Imputation of missing Value)

```
# 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)
+ }
> kenew$Meals_At_Home <- impute_with_mean(kenew$Meals_At_Home)
> # Check for missing values after imputation
> cat("Missing Values After Imputation:\n")
Missing Values After Imputation:
> print(colSums(is.na(kenew)))
            state_1            District              Region              Sector 
State_Region       Meals_At_Home           ricepds_v 
                  0                   0                   0                   0 
0                   0                   0 
         Wheatpds_q           chicken_q            pulsep_q           wheatos_q 
No_of_Meals_per_day 
                  0                   0                   0                   0 
0
```

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 for No_of_meals_per_day.

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

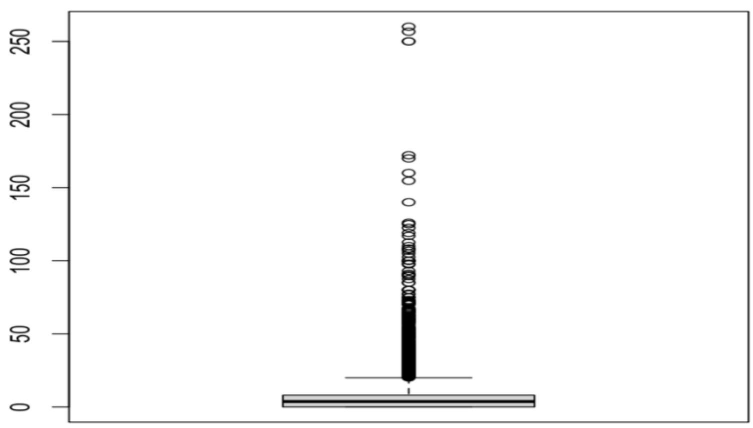
Boxplots can be used to find outliers in the dataset. Boxplots visually reveal outliers in a dataset by displaying individual points located beyond the whiskers of the boxplot.

#Checking for outliers

Plotting the boxplot to visualize outliers.

Code and Result:

```
> boxplot(apnew$ricepds_v)
```



Interpretation: From the boxplot above, which is a visual representation of the variable 'ricepds_v' shows that there is an outlier. Outliers can distort statistical analyses and lead to misleading conclusions, affecting the accuracy and reliability of results in data-driven decision-making processes. Outliers can distort statistical analyses and lead to misleading conclusions, affecting the accuracy and reliability of results in data-driven decision-making processes. The outliers can be removed using the following code.

#Setting quartiles and removing outliers

Code and results:

Setting quartile ranges to remove outliers

#Remove Outliers in the dtatset

```
> outlier_columns <- c("ricepds_v", "chicken_q","Wheatpds_p","ricepds_v", "Wheatpds_q",
"chicken_q", "pulsep_q",
"wheatos_q","No_of_Meals_per_day","total_consumption","District","Region","Sector")
> for (col in outlier_columns) {
+  kenew <- remove_outliers(apnew, col)
+ }
> # Summarize consumption
> apnew$total_consumption <- rowSums(apnew[, c("ricepds_v", "Wheatpds_q", "chicken_q",
"pulsep_q",
"wheatos_q","No_of_Meals_per_day","total_consumption","District","Region","Sector")], na.rm =
TRUE)
> # Summarize and display top and bottom consuming districts and regions
> summarize_consumption <- function(group_col) {
+  summary <- apnew %>%
+    group_by(across(all_of(group_col))) %>%
+    summarise(total = sum(total_consumption)) %>%
+    arrange(desc(total))
+  return(summary)
```

To change

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

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

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.

Code and Result:

```
# Rename districts and sectors , get codes from appendix of NSSO 68th ROund Data
district_mapping
<- c("14" = "Thiruvananthapuram", "04" = "Kozhikode", "2" = "Kannur")
sector_mapping <- c("2" = "URBAN", "1" = "RURAL")

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

fix(kenew)))
```

```
Result :
```

Result:

R Data Editor

Copy Paste Quit

	row.names	state_1	District	Region	Sector	State_Region	Meals_At_Home	ricepds_v	Wheatpds_q
1	1	KE	5	1	URBAN	321	83.48381	4	0
2	2	KE	5	1	URBAN	321	89	4	0.5
3	3	KE	5	1	URBAN	321	90	0	0
4	4	KE	5	1	URBAN	321	77	0	0
5	5	KE	5	1	URBAN	321	88	0	0
6	6	KE	5	1	URBAN	321	80	0	0
7	7	KE	5	1	URBAN	321	90	2.333333	0
8	8	KE	5	1	URBAN	321	80	0	0
9	9	KE	4	1	URBAN	321	90	9.6	0.4
10	10	KE	4	1	URBAN	321	90	4.666667	0
11	11	KE	4	1	URBAN	321	90	2.4	0
12	12	KE	4	1	URBAN	321	90	3.5	2.5
13	13	KE	4	1	URBAN	321	90	1.5	0.125
14	14	KE	4	1	URBAN	321	90	0	0
15	15	KE	4	1	URBAN	321	89	1.5	0
16	16	KE	4	1	URBAN	321	90	1	1.25
17	17	KE	4	1	URBAN	321	90	0	0
18	18	KE	4	1	URBAN	321	85	0	0
19	19	KE	4	1	URBAN	321	90	12.4	0
20	20	KE	4	1	URBAN	321	83.48381	0	0
21	21	KE	4	1	URBAN	321	83.48381	0	0
22	22	KE	4	1	URBAN	321	90	3	0
23	23	KE	4	1	URBAN	321	89	11.2	0
24	24	KE	4	1	URBAN	321	90	9.333333	0
25	25	KE	5	1	URBAN	321	83	4	0

Interpretation: The result as show above has successfully assigned the district names to the given number. Also the sectors 1 and 2 have been replaced as urban and rural sectors respectively.

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.

By summarizing the critical variables as total consumption we can estimate the top 3 and bottom 3 consuing districts

Code and Result:

```
# Summarize and display top and bottom consuming districts and regions
> summarize_consumption <- function(group_col) {
+   summary <- kenew %>%
+     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")
Top 3 Consuming Districts:
> print(head(district_summary, 3))
# A tibble: 3 × 2
  District total
     <int> <dbl>
1       14 2328.
2        4 2124.
3        2 1896.
> cat("Bottom 3 Consuming Districts:\n")
Bottom 3 Consuming Districts:
> print(tail(district_summary, 3))
# A tibble: 3 × 2
  District total
     <int> <dbl>
1       12  782.
2        9  735.
3        3  640.
>
> cat("Region Consumption Summary:\n")
Region Consumption Summary:
> print(region_summary)
# A tibble: 2 × 2
  Region  total
   <int>  <dbl>
1      2 11497.
```

```
2       1  9142.
```

Result:

```
1 Thiruvananthapuram      2328.
2 Kozhikode               2124.
3 Kannur                  1896.
```

Interpretation: The top three consuming districts are Thiruvananthapuram with 2328 units, followed by Kozhikode with 2124 units, and then in the third place Kannur with 1896 units

Similarly the bottom three districts can be found by sorting the total consumption.

Result:

```
1 Pathanamthitta       782.
2 Idukki              735.
3 Wayanand            640.
```

Interpretation: The least consuming district is Wayanad with only 640 units. Followed by Idukki in the second place and Pathanamthitta in the last place.

Total Consumption in Urban – 11497 Units

Total Consumption in Rural – 9142 Units

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.

```
# 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"))
+ }
```

Result:

```
Generated Output based on P-value

P value is < 0.05 i.e. 0, Therefore we reject the null hypothesis.There is a
difference between mean consumptions of urban and rural.The mean consumption in
Rural areas is 5.84450156641106 and in Urban areas its 4.76583963179114

>
```

Interpretation : P-value < 0.05:

A p-value less than 0.05 typically indicates that the results are statistically significant.

This means that there is strong evidence against the null hypothesis, leading to its rejection.

Rejecting the Null Hypothesis:

The null hypothesis generally states that there is no difference between the groups being compared. By rejecting the null hypothesis, we conclude that there is a significant difference between the mean consumptions in urban and rural areas.

Difference in Mean Consumptions:

The mean consumption in rural areas is approximately 5.84.

The mean consumption in urban areas is approximately 4.77.

This suggests that rural areas have higher mean consumption compared to urban areas.

Conclusion:

Statistical Significance: The test has shown that the difference in mean consumptions between urban and rural areas is statistically significant.
Practical Significance: Urban areas consume more on average than rural areas based on the provided mean values

Understandings: Urban Areas have a higher consumption rate because of the total units consumed compared to the Rural areas. This could be either because of the number of people or Households being more in Urban Areas.

CODES

```
# Set the working directory and verify it

setwd('/Users/kirthanshaker/Desktop/SCMA 631 Data Files ')

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
```

```r
data <- read.csv('/Users/kirthanshaker/Desktop/SCMA 631 Data
Files /NSSO68.csv')


# Filtering for AP

df <- data %>%

  filter(state_1 == "KE")


# 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

kenew <- 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(kenew)))


# 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)

}

kenew$Meals_At_Home <-
impute_with_mean(kenew$Meals_At_Home)


# Check for missing values after imputation

cat("Missing Values After Imputation:\n")
```

```
print(colSums(is.na(kenew)))



# 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) {

  kenew <- remove_outliers(kenew, col)

}



# Summarize consumption
```

```
kenew$total_consumption <- rowSums(kenew[, 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 <- kenew %>%

    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("14" = "Thiruvananthapuram", "04" =
"Kozhikode", "2" = "Kannur")

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



kenew$District <- as.character(kenew$District)

kenew$Sector <- as.character(kenew$Sector)

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

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



fix(kenew)

# Test for differences in mean consumption between urban and rural
```

```
rural <- kenew %>%
  filter(Sector == "RURAL") %>%
  select(total_consumption)


urban <- kenew %>%
  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"))

}
```