VIRGINIA COMMONWEALTH UNIVERSITY

STATISTICAL ANALYSIS & MODELING

A1a: CONSUMPTION PATTERN OF ANDHRA PRADESH USING PYTHON AND R

AMBALA GOWTHAM REDDY
V01072700

Date of Submission: 04/06/2023

CONTENTS

Analyzing Consumption in the State of Andhra Pradesh Using R

INTRODUCTION

The focus of this study is on the state of Andhra Pradesh, from the NSSO data, to find the top and bottom three consuming districts of Andhra Pradesh. 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 Andhra Pradesh'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 Andhra Pradesh'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(apnew))
[1] TRUE
> sum(is.na(apnew))
[1] 122
> sort(colSums(is.na(apnew)),decreasing=T)
      Meals_At_Home              state_1              District              Region
Sector
                122                    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 Andhra Pradesh, it is seen that only the column ‘Meals_At_Home has 122 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.

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

Code and Result:

```
> apnew=apnew%>%
+  mutate(across(all_of(c("Meals_At_Home")), ~ifelse(is.na(.), mean(., na.rm =
TRUE), .)))
> any(is.na(apnew))
[1] FALSE
```

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.

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

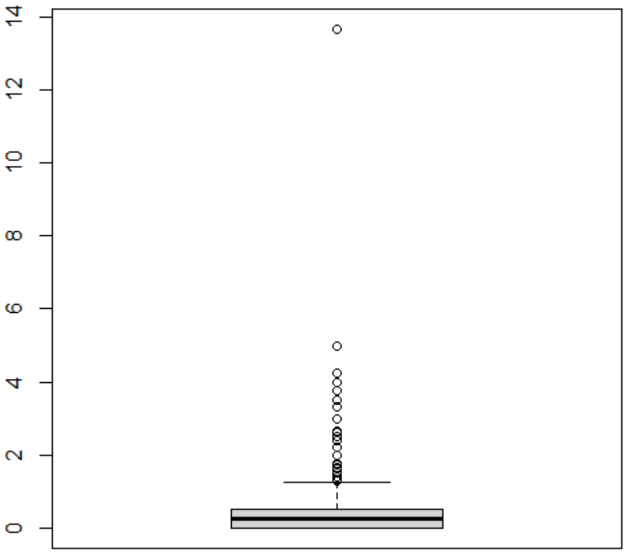
Boxplots can be used to find outliers in the dataset. Boxplots visually reveal outliers in a dataset by displaying individual points located beyond the whiskers of the boxplot.

#Checking for outliers

Plotting the boxplot to visualize outliers.

Code and Result:

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

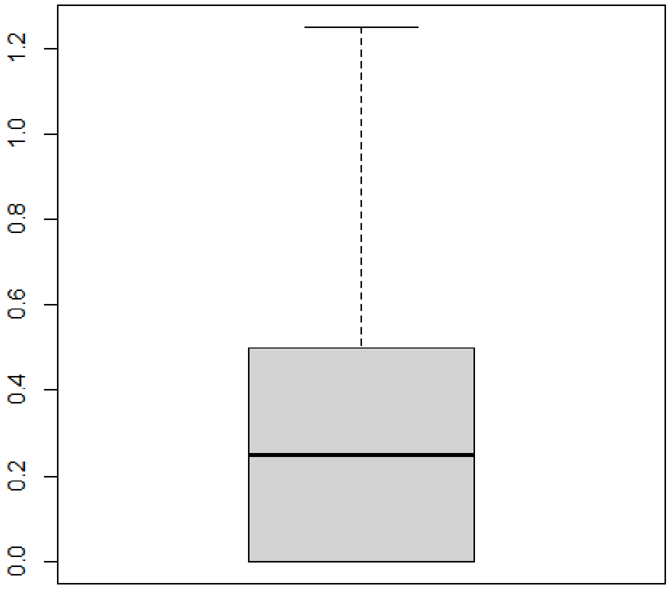


Interpretation: From the boxplot above, which is a visual representation of the variable 'ricepds_v' shows that there is an outlier. Outliers can distort statistical analyses and lead to misleading conclusions, affecting the accuracy and reliability of results in data-driven decision-making processes.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

```
> # Calculate quartiles and IQR
> Q1 <- quantile(apnew$ricepds_v, 0.25)
> Q3 <- quantile(apnew$ricepds_v, 0.75)
> IQR <- Q3 - Q1
> # Define outlier thresholds
> lower_threshold <- Q1 - (1.5 * IQR)
> upper_threshold <- Q3 + (1.5 * IQR)
> boxplot(apnew$ricepds_v)
```



<u>Interpretation</u>: 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:

```
> apnew$District <- recode(apnew$District, `1` = "Nellore", `2` = "Anantapur",
`3` = "Kurnool", `4` = "Krishna", `5` = "West Godavari", `6` = "East Godavari",
`7` = "Tirupati", `8` = "Eluru", `9` = "Vizag", `10` = "Guntur", `11` =
"Prakasam", `12` = "Chittoor", `13` = "Kadapa", `14` = "Srikakulam", `15` =
"Vizianagaram", `16` = "NTR District", `17` = "Alluri Sitaramaraju dt", `18` =
"Warangal", `19` = "Bapatla", `20` = "Vijayawada", `21` = "Rajamundry", `22` =
"Machilipatanam", `23` = "Puttaparthi")

> apnew$Sector <- ifelse(apnew$Sector == 2, "URBAN",
+                 ifelse(apnew$Sector == 1, "RURAL",apnew$Sector))
```

Result:

state_1	District	Region	Sector
AP	West Godavari	3	URBAN
AP	West Godavari	3	URBAN
AP	West Godavari	3	URBAN
AP	West Godavari	3	URBAN
AP	West Godavari	3	URBAN
AP	West Godavari	3	URBAN
AP	West Godavari	3	URBAN
AP	West Godavari	3	URBAN
AP	Guntur	4	URBAN
AP	Guntur	4	URBAN
AP	Guntur	4	URBAN
AP	Guntur	4	URBAN
AP	Guntur	4	URBAN
AP	Guntur	4	URBAN
AP	Guntur	4	URBAN
AP	Guntur	4	URBAN
AP	Guntur	4	URBAN
AP	Guntur	4	URBAN
AP	Guntur	4	URBAN

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 consuming districts.

Code and Result:

```
> apnew$total_consumption=
apnew$ricepds_v+apnew$Wheatpds_q+apnew$chicken_q+apnew$pulsep_q+apnew$wheatos_q
> apnew%>%
+   group_by(District)%>%
+   summarise(total=sum(total_consumption))%>%
+   arrange(-total,District)
```

Result:

```
 1 Srikakulam             1986.
 2 Machilipatanam         1977.
 3 Vizag                  1929.
 4 NTR District           1913.
 5 Puttaparthi            1759.
 6 East Godavari          1709.
 7 Kadapa                 1692.
 8 Alluri Sitaramaraju dt 1612.
 9 West Godavari          1549.
10 Krishna                1487.
```

Interpretation: The top three consuming districts are Srikakulam with 1986 units, followed by Machilipatanam with 1977 units, and then in the third place Viag with 1929 units

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

Result:

```
 1 Nellore    878.
 2 Anantapur  901.
 3 Guntur     981.
 4 Chittoor  1152.
 5 Bapatla   1190.
 6 Eluru     1215.
 7 Kurnool   1280.
 8 Prakasam  1336.
 9 Warangal  1342.
10 Tirupati  1435.
```

Interpretation: The least consuming district is Nellore with only 878 units. Followed by Anantapur in the second place and Guntur in the last place.

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.

```
> rural=apnew%>%
+  select(Sector,total_consumption)%>%
+  filter(Sector=="RURAL")
> fix(rural)
> urban=apnew%>%
+  select(Sector,total_consumption)%>%
+  filter(Sector=="URBAN")
> fix(urban)
> cons_rural=rural$total_consumption
> cons_urban=urban$total_consumption

> z.test(cons_rural,
+        cons_urban,
+        alternative="two.sided",
+        mu=0,
+        sigma.x = 2.56,sigma.y=2.34,
+        conf.level = 0.95)
```

Result:

```
Two-sample z-Test

data:  cons_rural and cons_urban
z = 29.202, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 1.614254 1.846533
sample estimates:
mean of x mean of y
```

<u>Interpretation:</u> The two-sample z-test indicates a highly significant difference in consumption between rural and urban sectors ($z = 29.202$, $p < 2.2e-16$, 95% CI: 1.614 to 1.847). Urban consumption is notably higher than rural consumption.

CODES

```
setwd('C:\\Users\\SERVICE POINT\\Desktop\\SCMA\\A1A')
getwd()
library(dplyr)
library(readr)
library(readxl)
library(tidyr)
install.packages("ggplot2")
library(ggplot2)

#READING THE FILE INTO R
data=read.csv("4. NSSO68 data set.csv")

#FILTERING FOR AP
df=data%>%
  filter(state_1=="AP")
names(df)
head(df)
dim(df)

#FINDING MISSING VALUES
is.na(df)
any(is.na(df))
sum(is.na(df))

sort(colSums(is.na(df)),decreasing=T)

# SUBSETIING
apnew = 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)
fix(apnew)
```

```
any(is.na(apnew))
sum(is.na(apnew))
head(apnew)
sort(colSums(is.na(apnew)),decreasing=T)


#IMPUTING THE VALUES i.e REPLACING MISSING VALUES WITH MEAN
apnew=apnew%>%
  mutate(across(all_of(c("Meals_At_Home")), ~ifelse(is.na(.), mean(., na.rm = TRUE), .)))
any(is.na(apnew))
fix(apnew)

# FINDING OUTLIERS AND MAKING AMENDMENTS
boxplot(apnew$ricepds_v)
boxplot(apnew$Wheatpds_q)
boxplot(apnew$chicken_q)
boxplot(apnew$pulsep_q)
boxplot(apnew$No_of_Meals_per_day)

# Calculate quartiles and IQR
Q1 <- quantile(apnew$ricepds_v, 0.25)
Q3 <- quantile(apnew$ricepds_v, 0.75)
IQR <- Q3 - Q1

# Define outlier thresholds
lower_threshold <- Q1 - (1.5 * IQR)
upper_threshold <- Q3 + (1.5 * IQR)

apnew = subset(apnew,apnew$ricepds_v>=lower_threshold & apnew$ricepds_v<=upper_threshold)
fix(apnew)
boxplot(apnew$ricepds_v)

Q1 <- quantile(apnew$chicken_q, 0.25)
Q3 <- quantile(apnew$chicken_q, 0.75)
```

```
IQR <- Q3 - Q1

# Define outlier thresholds
lower_threshold <- Q1 - (1.5 * IQR)
upper_threshold <- Q3 + (1.5 * IQR)

apnew = subset(apnew,apnew$chicken_q>=lower_threshold &
apnew$chicken_q<=upper_threshold)
fix(apnew)
boxplot(apnew$chicken_q)

#Renaming the districts as well as the sector, viz. rural and urban.

apnew$District <- ifelse(apnew$District == 5, "East Godavari",
                  ifelse(apnew$District == 10, "West Godavari",
                  ifelse(apnew$District == 6, "Nellore",
                  ifelse(apnew$District == 3, "Anantapur", apnew$Dist))))
fix(apnew)



apnew$Sector <- ifelse(apnew$Sector == 2, "URBAN",
                ifelse(apnew$Sector == 1, "RURAL",apnew$Sector))
fix(apnew)

# 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.
# 1. Districts
apnew$total_consumption=
apnew$ricepds_v+apnew$Wheatpds_q+apnew$chicken_q+apnew$pulsep_q+apnew$wheatos_q
apnew%>%
  group_by(District)%>%
  summarise(total=sum(total_consumption))%>%
  arrange(total,District)
#TOP 3 Consuming distircts are Anantapur, (3), District 23, Nellore(6)
```

```
# 2. Region
apnew%>%
  group_by(Region)%>%
  summarise(total=sum(total_consumption))%>%
  arrange(-total,Region)
# Region 3,1 and 5 are the top 3 consuming regions.

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

#H0: There is no difference in consumption between urban and rural.

#H1: There is difference in consumption between urban and rural.

rural=apnew%>%
  select(Sector,total_consumption)%>%
  filter(Sector=="RURAL")
fix(rural)

urban=apnew%>%
  select(Sector,total_consumption)%>%
  filter(Sector=="URBAN")
fix(urban)

cons_rural=rural$total_consumption
cons_urban=urban$total_consumption


length(cons_rural)
length(cons_urban)

install.packages("BSDA")
library(BSDA)
```

```
z.test(cons_rural,
       cons_urban,
       alternative="two.sided",
       mu=0,
       sigma.x = 2.56,sigma.y=2.34,
       conf.level = 0.95)

# P value is <0.05, Therefore we reject the null hypothesis.
#There is difference between mean consumptions of urban and rural.
```