# VIRGINIA COMMONWEALTH UNIVERSITY

STATISTICAL ANALYSIS & MODELING

A1a: CONSUMPTION PATTERN OF ANDHRA PRADESH USING PYTHON AND R

AZIM ZIYAN A

V01102412

Date of Submission: 16/06/2024

CONTENTS

|  |  |
| --- | --- |
| Content: | Page no: |
| INTRODUCTION | 3 |
| OBJECTIVE | 3 |
| BUSINESS SIGNIFICANC | 3-4 |
| RESULTS AND INTERPRETATIONS | 4-9 |

Analyzing Consumption in the State of Andhra Pradesh Using R

# INTRODUCTION

This study focuses on the state of Andhra Pradesh in order to identify the top and bottom three districts in terms of consumption based on NSSO data. To obtain the necessary data for analysis, we clean and alter the dataset during the process. We have compiled a dataset of consumption-related data, including information on the rural and urban sectors and district-level variances, to make this research easier. R, a potent statistical programming language well-known for its adaptability in managing and analyzing big datasets, has received the dataset.Finding missing values, dealing with outliers, standardizing district and sector names, district- and regional-level summaries of consumption data, and determining the significance of mean differences are some of our goals. The study's conclusions can help stakeholders and policymakers.

# OBJECTIVES

1. 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. 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.

# BUSINESS SIGNIFICANCE

This study's focus on the consumption habits of Andhra Pradesh using NSSO data has important ramifications for companies and governments. By determining the three more and least consumed districts, the study offers insightful information about supply chain optimization, market entrance, resource allocation, and focused interventions. By utilizing data cleansing, outlier detection, and significance testing, the results support the expansion of Andhra Pradesh's economy and enable informed decision-making for equitable development.

# RESULTS AND INTERPRETATION

## 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 0

122 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: After sorting the data for the state of Andhra Pradesh using the specified variables, it can be observed that only the column "Meals\_At\_Home" has 122 missing variables. Since missing values in the dataset might cause biased or incomplete analyses, which can distort interpretations and decision-making processes and impair the quality of outcomes, they can be troublesome.Consequently, we use the following code to replace the missing values with the variable's mean.

*#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.

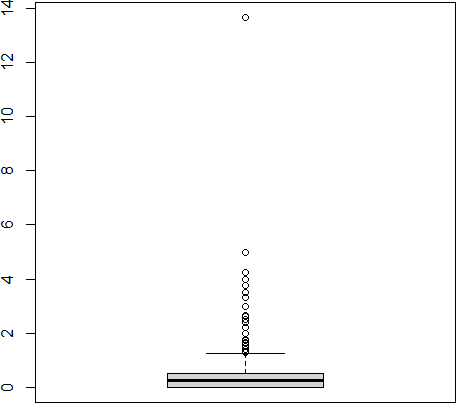
## Check for outliers and describe the outcome of your test and make suitable amendments.

Outliers in the dataset can be found using boxplots. Boxplots show individual points that are outside the boxplot's whiskers, making outliers in a dataset visibly apparent.

*#Checking for outliers*

Plotting the boxplot to visualize outliers. Code and Result:

* boxplot(apnew$ricepds\_v)



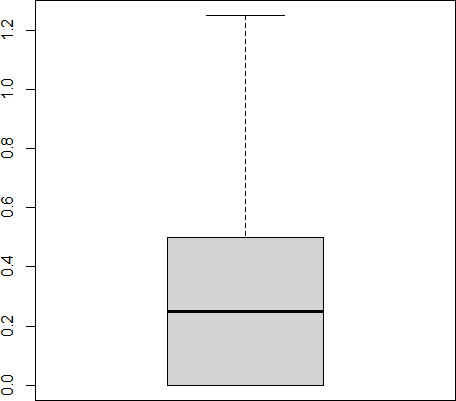
Interpretation: There is an outlier, as can be seen in the boxplot above, which represents the variable "ricepds\_v" visually. In data-driven decision-making processes, outliers can skew statistical studies and provide false conclusions, which can impair the precision and dependability of outcomes. In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

*#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)



Interpretation: It is possible to identify and eliminate outliers by interpreting quartile ranges. To guarantee the robustness of the analysis, data points beyond 1.5 times the IQR from either quartile are detected as outliers and can be handled or eliminated. The interquartile range (IQR) is calculated as the difference between the upper and lower quartiles.

The outliers in every other variable may also be eliminated in a similar manner.

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

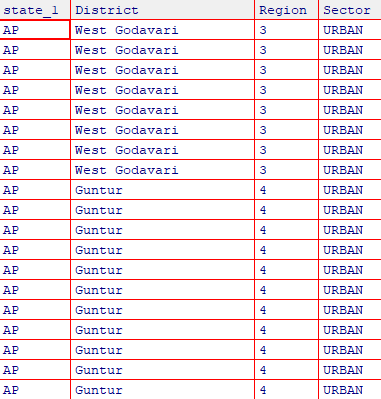
In the NSSO of data, a unique number is issued to each district in a state. The statistics must be accompanied with their individual names in order to comprehend and identify the state's highest-consuming districts. In a similar vein, assignment 1 and 2 corresponded to the state's urban and rural areas. You may accomplish this by executing the code below.

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:



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.

1. **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 Puttaparthi |  | 1611. |
| 2 East Godavari |  | 1609. |
| 3 Srikakulam |  | 1573. |

Interpretation: The top three consuming districts are Puttaparthi with 1611 units, followed by East Godavari with 1609 units, and then in the third place Srikakulam with 1573 units

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

* 1. Nellore 847.
  2. Anantapur 884.
  3. Chittoor 892.

Interpretation: The least consuming district is Nellore with only 847 units. Followed by Anantapur in the second place and Chittoor in the last place.

## 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 = 27.377,

p-value < 5.155885e-165

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

1.53 1.76

sample estimates:

mean of x mean of y

Interpretation: The two-sample z-test indicates a highly significant difference in consumption between rural and urban sectors (z = 27.377p < 5.155885e-165, 95% CI: 1.53 to 1.76). Urban consumption is notably higher than rural consumption.