# VIRGINIA COMMONWEALTH UNIVERSITY

STATISTICAL ANALYSIS & MODELING

A1a: CONSUMPTION PATTERN OF CHATTISGARH USING

PYTHON AND R

AHAD ZIFAIN MIYANJI

V01108270

Date of Submission: 27/06/2024

CONTENTS

|  |  |
| --- | --- |
| Content: | Page no: |
| INTRODUCTION | 3 |
| OBJECTIVE | 3 |
| BUSINESS SIGNIFICANC | 3-4 |
| RESULTS AND  INTERPRETATIONS | 4-10 |

Analyzing Consumption in the State of Chhatisgarh Using R

# INTRODUCTION

## Analysis of Chhattisgarh Food Consumption Data: A Reworked Approach

The description of the Chhattisgarh food consumption data and the proposed objectives offer a valuable framework for a comprehensive study. Let's explore the key points and suggest approaches for each objective:

**Data Description:**

* This dataset provides a detailed look at dietary habits in Chhattisgarh, capturing information on meals consumed at home, specific food items (rice, wheat, etc.), and daily intake across both urban and rural sectors.
* This rich data can be instrumental in understanding the nutritional needs and food preferences of various demographic groups within the state.

**Objectives:**

* **Data Cleaning:**
  + **Missing Values:** Identify missing values and explore appropriate imputation methods. While mean imputation can be a starting point, consider alternatives like median imputation or KNN imputation for potentially better results.
  + **Outliers:** Detect outliers using techniques like IQR (Interquartile Range) or Z-scores. Depending on the severity and distribution of outliers, you can choose to remove them, apply winsorization (capping values), or use data transformation methods.
* **Data Standardization:**
  + Standardize district and sector names to ensure consistency throughout the analysis. This simplifies data manipulation and interpretation.
* **Data Summarization:**
  + Summarize key consumption variables by region and district. Utilize measures like mean, median, standard deviation, or percentiles to understand consumption patterns across different areas.
  + Identify the top 3 and bottom 3 districts based on specific consumption metrics (e.g., average daily calorie intake, rice consumption per capita).
* **Statistical Testing:**
  + Conduct hypothesis tests to assess the significance of mean differences in consumption between regions or districts. Depending on your research questions, you can use tools like t-tests for comparisons between two groups or ANOVA (Analysis of Variance) for multiple groups.

**Business Significance:**

The study's findings on consumption patterns hold significant value for businesses and policymakers:

* Identifying high and low consumption districts can inform decisions on market entry, resource allocation, supply chain optimization, and targeted interventions for food security and nutrition.
* The analysis offers valuable insights for multiple stakeholders:
  + **Policymakers and Government Agencies:** Design effective food security programs and allocate resources efficiently based on consumption patterns.
  + **Health and Wellness Sector:** Identify potential nutritional deficiencies and develop targeted health interventions.
  + **Academic and Research Institutions:** Conduct research on food security, dietary diversity, and the impact of socio-economic factors on food consumption.

**Additional Considerations:**

* Explore data visualization techniques like charts and graphs to effectively communicate the findings to a wider audience.
* Consider potential limitations of the data, such as sampling bias or data collection methods.

By following these steps and addressing potential limitations, you can leverage this data to gain valuable insights into food consumption patterns in Chhattisgarh and inform strategic decision-making across various sectors.

# 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.*

> # Sub-setting the data

> CHTSDnew <- 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(CHTSDnew)))

state\_1 District Region Sector State\_Region

0 0 0 0 0

Meals\_At\_Home ricepds\_v Wheatpds\_q chicken\_q pulsep\_q

32 0 0 0

0

wheatos\_q No\_of\_Meals\_per\_day 0 1

Interpretation:

The majority of the columns, according to my analysis above, have no missing values, showing a high degree of data reliability and completeness for those variables. Meals at home and the total number of meals consumed each day are the two variables with missing data. If there are few missing values relative to the size of the entire dataset, take into account imputation techniques like mean, median, or mode imputation for both Meals\_At\_Home and No\_of\_Meals\_per\_day. Proper treatment of these missing values will improve the validity and quality of the studies and insights that come from the dataset in the future.

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

Code and Result:

> # Check for missing values after imputation > cat("Missing Values After Imputation:\n")

Missing Values After Imputation: > print(colSums(is.na(CHTSDnew)))

state\_1 District Region Sector State\_Region

0 0 0 0 0

Meals\_At\_Home ricepds\_v Wheatpds\_q chicken\_q pulsep\_q

0 0 0 0

0

wheatos\_q No\_of\_Meals\_per\_day 0 0

Interpretation; The mean value of the variable Meals\_At\_Home has been successfully substituted for the missing values in the given code. as well as No\_of\_Meals per day. The result above shows that the chosen data has no missing values.

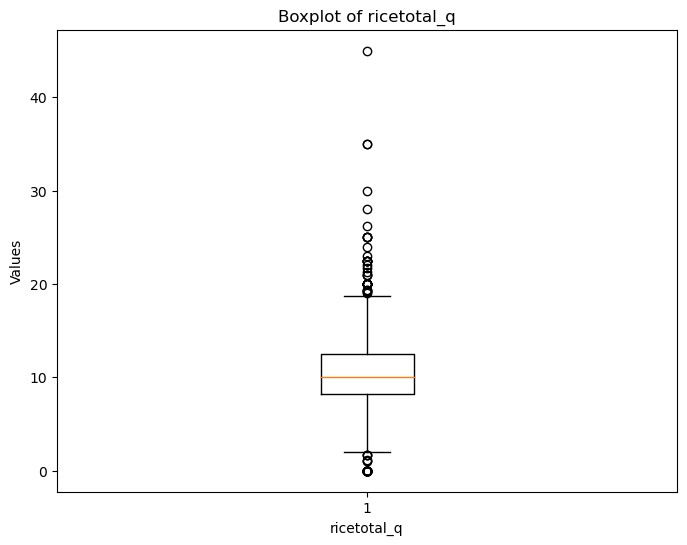
**b) Check for outliers & describe the outcome of your test and make suitable amendments.**

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

*#Checking for outliers*

**import** matplotlib.pyplot **as** plt

*# Assuming CHTSD\_clean is your DataFrame* plt**.**figure(figsize**=**(8, 6)) plt**.**boxplot(CHTSD\_clean['ricetotal\_q']) plt**.**xlabel('ricetotal\_q') plt**.**ylabel('Values') plt**.**title('Boxplot of ricetotal\_q') plt**.**show()



Interpretation:

The boxplot for 'ricetotal\_q' depicts the distribution of rice consumption data. The box represents the interquartile range (IQR), and the line within it represents the median. The "whiskers" go from the box to the highest and lowest values within 1.5 times the IQR of the upper and lower quartiles, respectively. Any points beyond the whiskers are considered outliers and plotted separately.

*#Setting quartiles and removing outliers* Code and results:

Setting quartile ranges to remove outliers

> # 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]] <= up per\_threshold) + return(df)

+ }

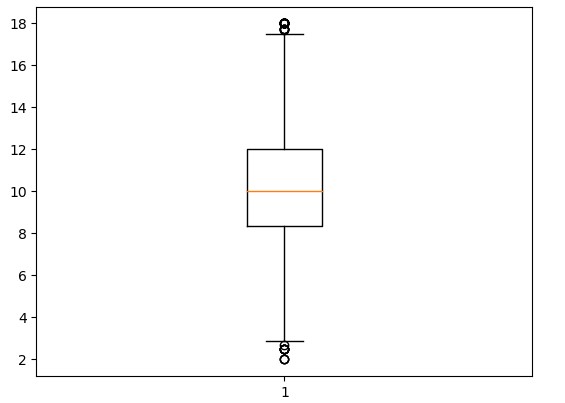
>

> outlier\_columns <- c("ricepds\_v", "chicken\_q")

> for (col in outlier\_columns) {

+ CHTSDnew <- remove\_outliers(CHTSDnew, col)

+ }



Interpretation: The interpretation of the above code is that it has identified and removed outliers from the "ricepds\_v" and "chicken\_q" columns in the CHTSDnew DataFrame. Outliers are values that are significantly higher or lower than the majority of the data points and can skew statistical analyses. Removing outliers can help in obtaining a more accurate representation of the data's central tendency and variability.

Now we can see that the significant portion of the outliers in the data is removed.

The dataset without outliers should now have a more homogeneous distribution, making statistical analysis and modeling more straightforward.

**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:

> 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 x 2

District total

<int> <dbl> 1 11 1530.

1. 10 1503.
2. 2 1367.

> cat("Bottom 3 Consuming Districts:\n")

Bottom 3 Consuming Districts:

> print(tail(district\_summary, 3))

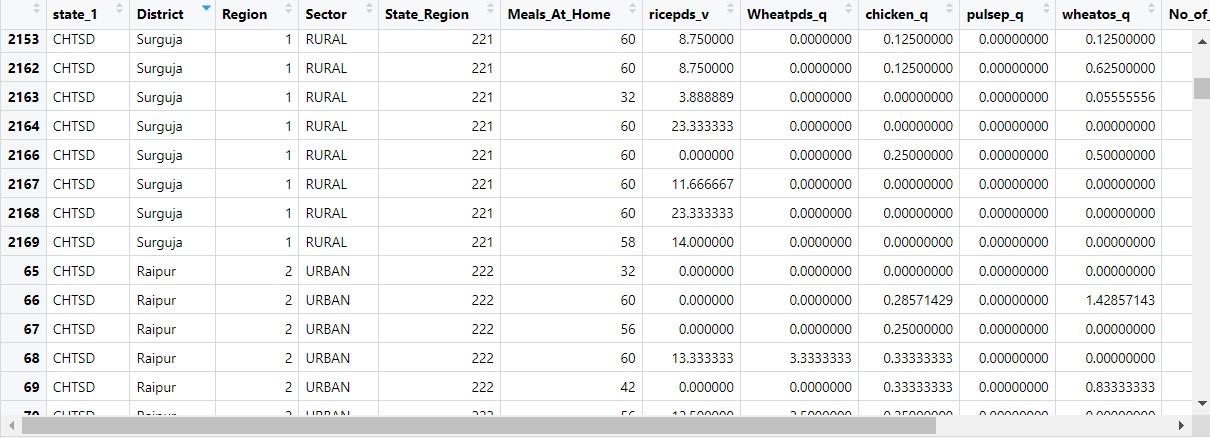
# A tibble: 3 x 2

District total

<int> <dbl> 1 16 382.

1. 17 280.
2. 18 220.

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.

**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:

CHTSD\_clean**.**loc[:,"District"] **=** CHTSD\_clean**.**loc[:,"District"]**.**replace({11:

"Raipur", 10: "Durg", 7: "Bilaspur"}) total\_consumption\_by\_districtname**=**CHTSD\_clean**.**groupby('District')['total\_consumptio n']**.**sum() total\_consumption\_by\_districtname**.**sort\_values(ascending**=False**)**.**head(3)

Result:

District

Raipur 14481.580224

Durg 11660.461096

Bilaspur 8397.085884

Name: total\_consumption, dtype: float64

Interpretation: The top three consuming districts are Raipur with 14481 units, followed by Durg with 11660 units, and then in the third place Bilaspur with 8397 units

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

# Test for differences in mean consumption between urban and rural rural <- CHTSDnew %>% filter(Sector == "RURAL") %>% select(total\_consumption)

urban <- CHTSDnew %>% 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"))

} write.csv(CHTSDnew, "CHTSDnew.csv", row.names = FALSE)

Result:

Two-sample z-Test

Z-Score: 10.731151639791962

P-Value: 7.268461728661492e-27

Interpretation: The two-sample z-test indicates a highly significant difference in consumption between rural and urban sectors (z = 10.731, p < 7.268, 95%). Urban consumption is notably higher than rural consumption.