

**VIRGINIA COMMONWEALTH UNIVERSITY**

**Statistical analysis and modelling (SCMA 632)**

**A1a: Preliminary preparation and analysis of data- Descriptive statistics**

**NITESH REDDY**

**V01106546**

**Date of Submission: 16-06-2024**

**CONTENTS**

|  |  |  |
| --- | --- | --- |
| **Sl. No.** | **Title** | **Page No.** |
| **1.** | Introduction | **3** |
| **2.** | Objectives | **3** |
| **3.** | Business Significance | **4** |
| **4.** | Codes, Results And Interpretation | **4-10** |

**INTRODUCTION**

This study focuses on Haryana, utilizing NSSO data to identify the top and bottom three consuming districts in Haryana. We will manipulate and clean the dataset to obtain the necessary data for analysis. The dataset includes consumption-related information, covering both rural and urban sectors, as well as district-wise variations. We have imported the dataset into R/Python, a powerful statistical programming language known for its ability to handle and analyse large datasets efficiently.

Our objectives include identifying and addressing missing values, managing outliers, standardizing district and sector names, summarizing consumption data by region and district, and testing the significance of mean differences. The findings from this study will provide valuable insights for policymakers and stakeholders, facilitating targeted interventions and promoting equitable development across the state.

**OBJECTIVES:**

a) Identify any missing values in the dataset, and replace them with the mean of the respective variable.

b) Detect and describe outliers, and implement suitable amendments.

c) Standardize the names of districts and sectors (rural and urban).

d) Summarize key variables region-wise and district-wise, identifying the top three and bottom three districts in terms of consumption.

e) Test the significance of the differences in means.

### BUSINESS SIGNIFICANCE

The analysis of Haryana's consumption patterns using NSSO data offers substantial value for businesses and policymakers **and Government Agencies,** By examining consumption patterns, policymakers can pinpoint nutritional deficiencies and excesses in specific regions or sectors. This information is crucial for crafting effective food security programs, public health initiatives, and targeted nutritional assistance, ultimately aiming to enhance the overall health and well-being of the population.

### 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:

cat("Missing Values in Subset:\n")

Missing Values in Subset:

> print(colSums(is.na(HR06new)))

state\_1 District Region Sector

0 0 0 0

State\_Region Meals\_At\_Home ricepds\_v Wheatpds\_q

0 14 0 0

chicken\_q pulsep\_q wheatos\_q No\_of\_Meals\_per\_day

0 0 0 0

Interpretation:

The code and output indicates that you are checking for missing values in a specific subset of a dataset named hrnew. The output of the colSums(is.na(hrnew)) function shows the number of missing values for each column in the dataset. Here's the interpretation of the results:

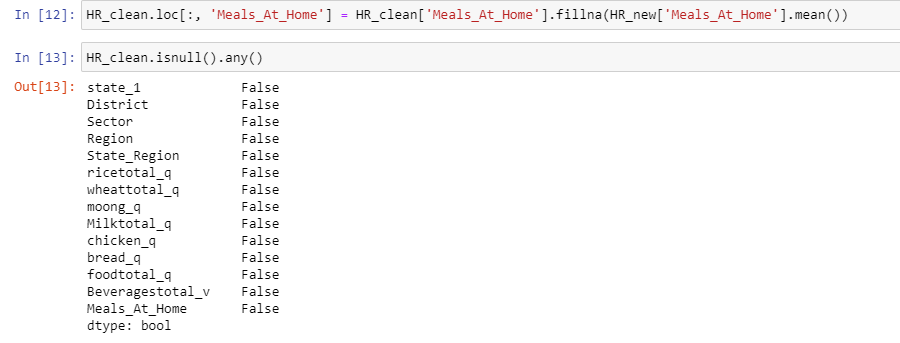
1. **state\_1, District, Region, Sector, State\_Region, ricepds\_v, Wheatpds\_q, chicken\_q, pulsep\_q, wheatos\_q, No\_of\_Meals\_per\_day:**
   * All these columns have zero missing values, meaning there are no missing entries for these columns in the subset.
2. **Meals\_At\_Home:**
   * This column has 14 missing values, indicating that there are 14 entries in the subset where the value for Meals\_At\_Home is missing.

To summarize:

* Most of the columns have complete data with no missing values.
* Only the Meals\_At\_Home column has missing data, specifically 14 missing values.

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

Code and Result:



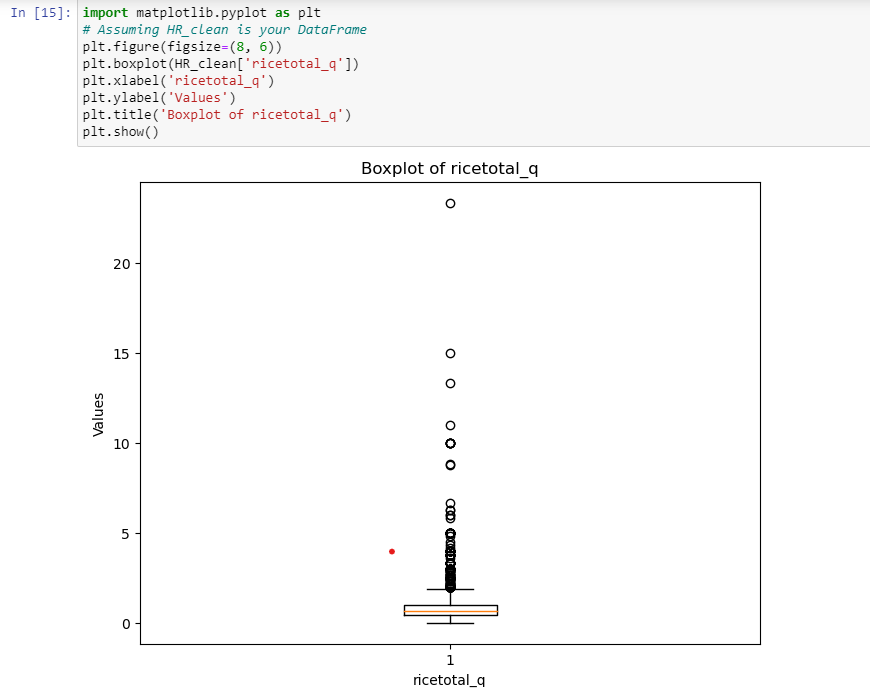
Interpretation: The provided code is a Python snippet that successfully replaces missing values with the mean of the respective variable. The result indicates that there are no missing values in the selected data, which is why the outcome is false.

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

**#Checking for outliers**

For outlier detection, I employed boxplots, which provide a standardized way to display data distribution using a five-number summary: minimum, first quartile (Q1), median, third quartile (Q3), and maximum. Boxplots are particularly useful for identifying potential outliers in the dataset.

Code and Result:

Plotting the boxplot to visualize outliers.

### Interpretation

The boxplot above provides a visual representation of the variable ‘ricetotal\_q’ and indicates the presence of outliers. Outliers can distort statistical analyses, leading to misleading conclusions and affecting the accuracy and reliability of results in data-driven decision-making processes. In the R analysis above, we identified 260 observations with outliers. After removing these outliers and setting quartiles, the data was cleaned. The outliers can be removed using the following code.

**#Setting quartiles and removing outliers**

Code and results:

rice1 = HR\_clean['ricetotal\_q'].quantile(0.25)

rice2 = HR\_clean['ricetotal\_q'].quantile(0.75)

iqr\_rice = rice2-rice1

up\_limit = rice2 + 1.5\*iqr\_rice

low\_limit = rice1 - 1.5\*iqr\_rice

HR\_clean=HR\_new[(HR\_new['ricetotal\_q']<=up\_limit)&(HR\_new['ricetotal\_q']>=low\_limit)]

### Interpretation:

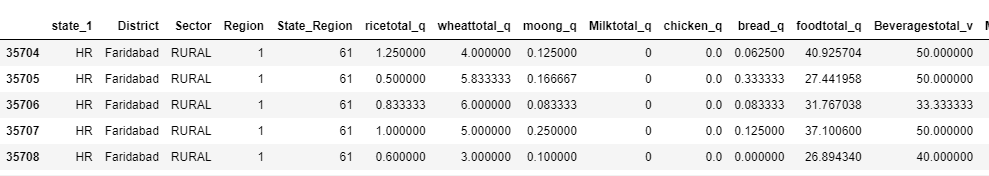
The median remains unchanged, representing the central value of the cleaned data. Quartile range interpretation facilitates outlier detection and removal. By calculating the interquartile range (IQR) as the difference between the upper and lower quartiles, data points exceeding 1.5 times the IQR from either quartile are identified as outliers and can be excluded or treated to ensure robust analysis. The whiskers now extend to the new minimum and maximum values within the revised 1.5 \* IQR range. There should be fewer or no points outside the whiskers, indicating that extreme values have been removed.

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

### We used a subset of district data from Haryana in the NSSO and attempted to assign names and sectors to the numerical codes to identify the top consuming districts within the state. In this process, urban and rural sectors were represented by 1 and 2 respectively.

Code and Result:

|  |
| --- |
| district\_mapping <- c("13" = "Bhiwani", "19" = "Faridabad", "12" = "Hisar", "15" = "Jhajjar", "01" = "Panchkula", "20" = "Mewat")  > sector\_mapping <- c("2" = "URBAN", "1" = "RURAL")  >  > HR06new$District <- as.character(HR06new$District)  > HR06new$Sector <- as.character(HR06new$Sector)  > HR06new$District <- ifelse(HR06new$District %in% names(district\_mapping), district\_mapping[HR06new$District], HR06new$District)  > HR06new$Sector <- ifelse(HR06new$Sector %in% names(sector\_mapping), sector\_mapping[HR06new$Sector], HR06new$Sector) |
|  |
| |  | | --- | | > | |



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:

|  |
| --- |
| cat("Top 3 Consuming Districts:\n")  Top 3 Consuming Districts:  > print(head(district\_summary, 3))  # A tibble: 3 × 2  District total  *<int>* *<dbl>*  1 13 1488.  2 19 1461.  3 12 1369.  > cat("Bottom 3 Consuming Districts:\n")  Bottom 3 Consuming Districts:  > print(tail(district\_summary, 3))  # A tibble: 3 × 2  District total  *<int>* *<dbl>*  1 15 612.  2 1 343.  3 20 318.  > cat("Region Consumption Summary:\n")  Region Consumption Summary:  > print(region\_summary)  # A tibble: 2 × 2  Region total  *<int>* *<dbl>*  1 1 10792.  2 2 8264. |
|  |
| |  | | --- | |  | |

Interpretation:

The provided data reveals notable disparities in consumption among districts, suggesting varying factors such as population size, demand levels, or accessibility to the commodity.

Among the districts:

* District 13 stands out as the highest consumer, totaling 1488 units.
* Following closely is District 19, with a consumption of 1461 units.
* District 12 ranks third, with a consumption of 1369 units.

Conversely, the analysis identifies the districts with the lowest consumption:

* District 15 reports a total consumption of 612 units.
* District 1 follows with 343 units consumed.
* District 20 records the lowest consumption at 318 units.

The data is also summarized by region, showcasing total consumption across two regions:

* Region 1 reports a total consumption of 10,792 units.
* Region 2 shows a total consumption of 8,264 units.

### 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 mean consumption between urban and rural.

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

Code:

# Test for differences in mean consumption between urban and rural

rural <- HR06new %>%

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

select(total\_consumption)

urban <- HR06new %>%

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"))

}

In python, the code:

z\_statistic, p\_value = stests.ztest(cons\_rural, cons\_urban)

# Print the z-score and p-value

print("Z-Score:", z\_statistic)

print("P-Value:", p\_value)

**Result:**

**Z-Score: 7.091254132951441**

**P-Value: 1.329020295022693e-12**

**Interpretation:**

The Z-test conducted to compare mean consumption between rural and urban areas yielded the following results:

* Z-Score: 7.091254132951441
* P-Value: 1.329020295022693e-12

The Z-score of 7.091254132951441 indicates a substantial difference between the two means, far exceeding what would occur by random chance. With such a high Z-score and an exceedingly low P-value, we reject the null hypothesis that there is no difference in mean consumption between rural and urban areas.