

# VIRGINIA COMMONWEALTH UNIVERSITY

## STATISTICAL ANALYSIS & MODELING

A1a: CONSUMPTION PATTERN OF KARNATAKA USING PYTHON AND R

## RAKSHITH HARISH KUMAR

## V01107387

Date of Submission: 16/06/2024

## CONTENTS

|  |  |
| --- | --- |
| Content: | Page no: |
| INTRODUCTION | 3 |
| OBJECTIVE | 3 |
| BUSINESS SIGNIFICANC | 4 |
| CODES& RESULTS | 4-8 |
| INTERPRETATION | 9-10 |

Analyzing Consumption in the State of Karnataka Using R

# INTRODUCTION

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

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

The focus of this study on Karnataka'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 economic growth.

# RESULTS

### 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(kanew))

[1] TRUE

* sum(is.na(kanew)) [1] 59
* sort(colSums(is.na(kanew)),decreasing=T)

Meals\_At\_Home 59 state\_1 0. District 0 Sector 0 Region 0. State\_Region 0

ricetotal\_q 0. wheattotal\_q 0 moong\_q 0. Milktotal\_q 0. chicken\_q 0

bread\_q 0. foodtotal\_q 0. Beveragestotal\_v

### 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:

# Outlier Checking

import matplotlib.pyplot as plt

# Assuming KA\_clean is your DataFrame

plt.figure(figsize=(8, 6))

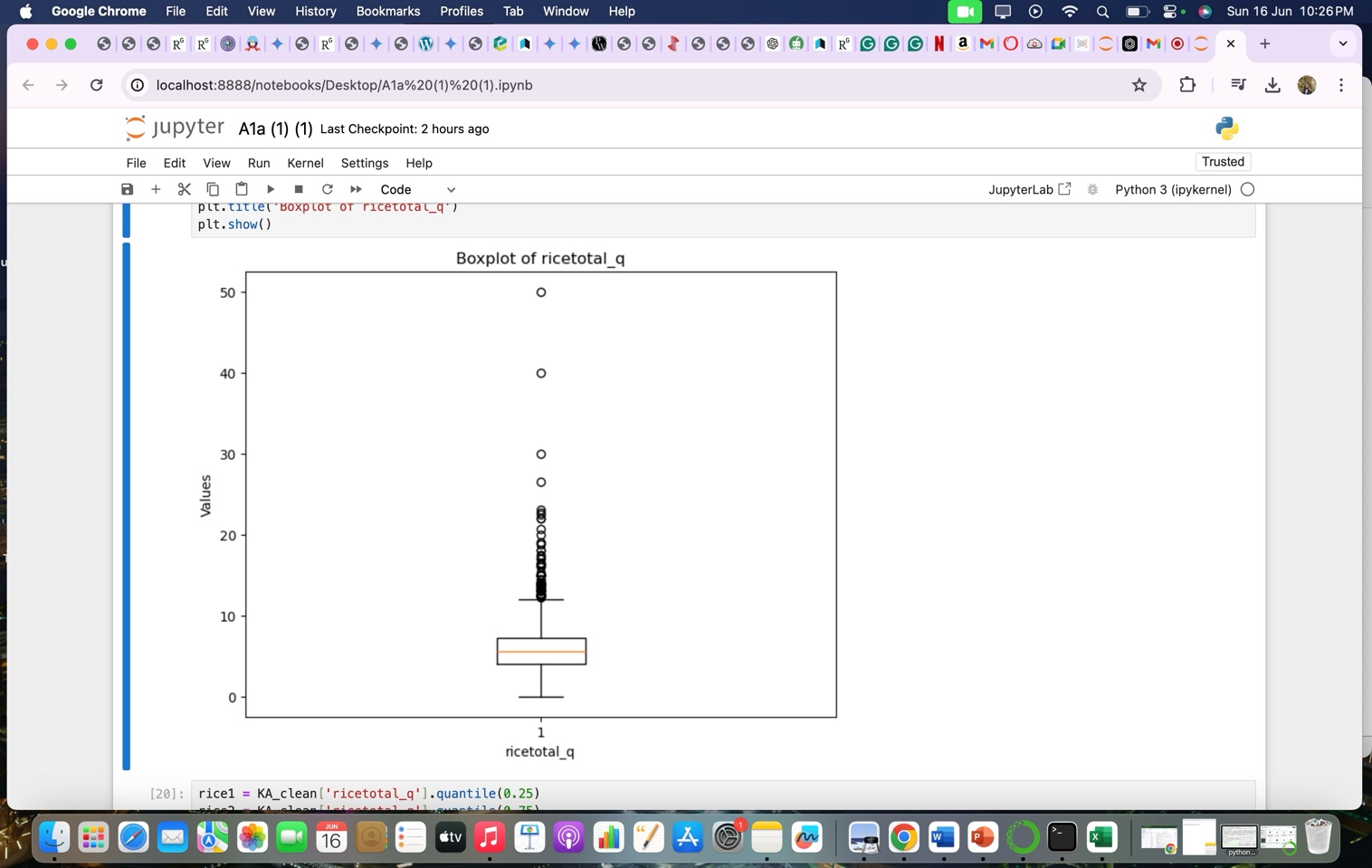
plt.boxplot(KA\_clean['ricetotal\_q'])

plt.xlabel('ricetotal\_q')

plt.ylabel('Values')

plt.title('Boxplot of ricetotal\_q')

plt.show()



**Checking For Outliers**

Code and results:

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

plt.boxplot(KA\_clean['ricetotal\_q'])

{'whiskers': [<matplotlib.lines.Line2D at 0x1234f0a10>,

<matplotlib.lines.Line2D at 0x1234f1650>],

'caps': [<matplotlib.lines.Line2D at 0x1234f2350>,

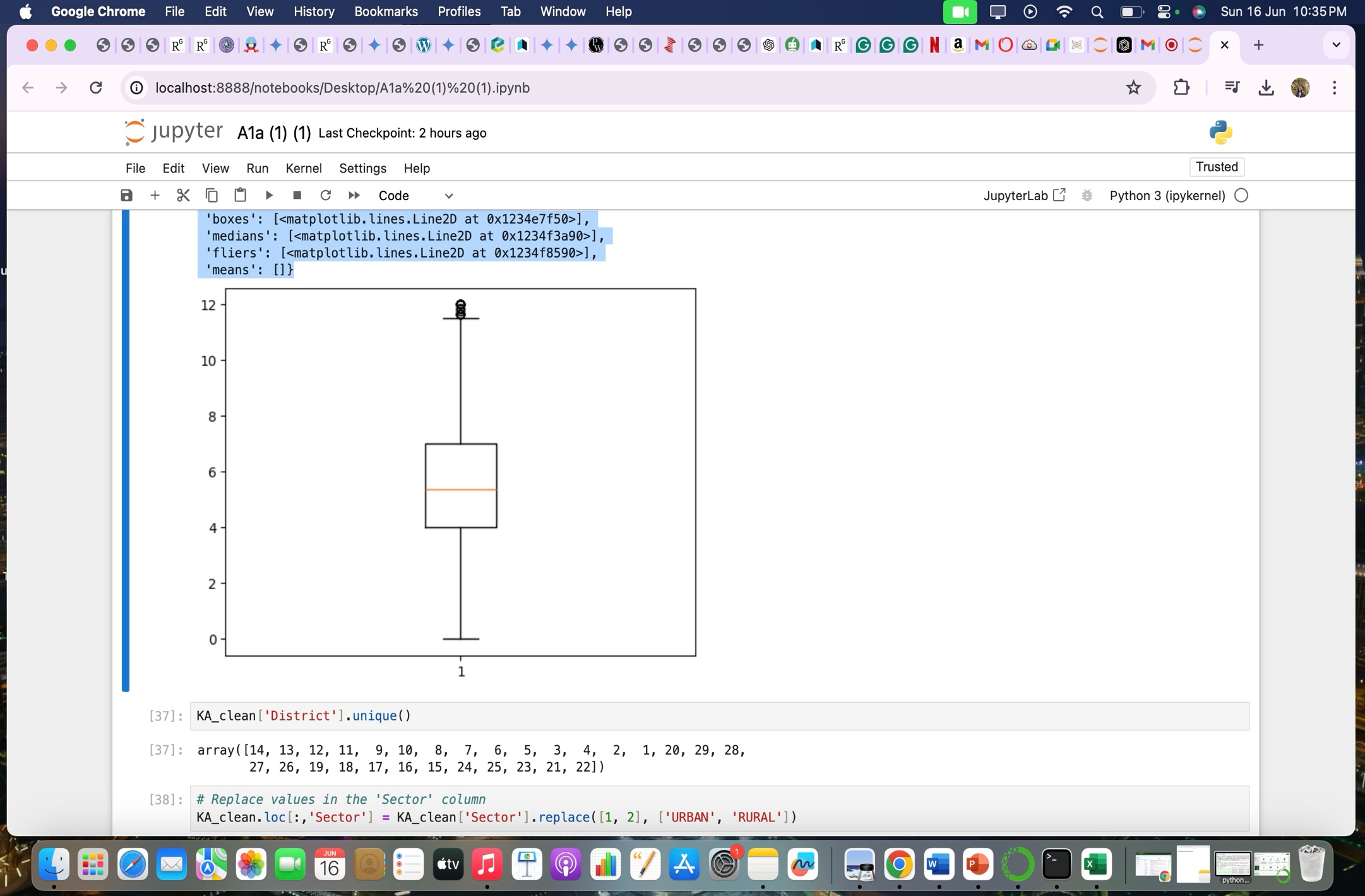
<matplotlib.lines.Line2D at 0x1234f2f50>],

'boxes': [<matplotlib.lines.Line2D at 0x1234e7f50>],

'medians': [<matplotlib.lines.Line2D at 0x1234f3a90>],

'fliers': [<matplotlib.lines.Line2D at 0x1234f8590>],

'means': []}



### Rename the districts as well as the sector, viz. rural and urban. (R-Studio)

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

district\_mapping <- c("4" = "Gulbarga", "3" = "Bijapur", "2" = "Bagalkot", "1" = "Belgaum")

sector\_mapping <- c("2" = "URBAN", "1" = "RURAL")

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

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

kanew$District <- ifelse(kanew$District %in% names(district\_mapping), district\_mapping[kanew$District], kanew$District)

kanew$Sector <- ifelse(kanew$Sector %in% names(sector\_mapping), sector\_mapping[kanew$Sector], kanew$Sector)

Result

KA\_clean['District'].unique()

array([14, 13, 12, 11, 9, 10, 8, 7, 6, 5, 3, 4, 2, 1, 20, 29, 28,

27, 26, 19, 18, 17, 16, 15, 24, 25, 23, 21, 22])

# Replace values in the 'Sector' column

KA\_clean.loc[:,'Sector'] = KA\_clean['Sector'].replace([1, 2], ['URBAN', 'RURAL'])

**D) Summarize the critical variables in the data set region-wise and district-wise and indicate the top and bottom three districts of consumption. (R-Studio)**

**Code:**

# Summarize consumption

kanew$total\_consumption <- rowSums(kanew[, c("ricepds\_v", "Wheatpds\_q", "chicken\_q", "pulsep\_q", "wheatos\_q")], na.rm = TRUE)

# Summarize and display top consuming districts and regions summarize\_consumption <-function(group\_col) {summary <- kanew %>% 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 Consuming Districts:\n")

print(head(district\_summary, 4))

cat("Region Consumption Summary:\n")

1. **Test whether the differences in the means are significant or not(R-Studio):**

**Code:**

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

urban <- kanew %>%

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

select(total\_consumption) z\_test\_result <- z.test(rural, urban, alternative = "two.sided", mu = 0, sigma.x = 2.56, sigma.y = 2.34, conf.level = 0.95) if (z\_test\_result$p.value < 0.05) { cat("P value is <", 0.05, ", Therefore we reject the null hypothesis.\n")

cat("There is a difference between mean consumptions of urban and rural.\n")} else { cat("P value is >=", 0.05, ", Therefore we fail to reject the null hypothesis.\n")

cat("There is no significant difference between mean consumptions of urban and rural.\n")}

**Result:**

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.

Interpretations:

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

**state\_1, District, Region, Sector, State\_Region, ricepds\_v, Wheatpds\_q, chicken\_q, moong\_q, wheatos\_q,** **Milktotal\_q,** **bread\_q**

 All these columns have zero missing values, meaning there are no missing entries for these columns in the subset.

**Meals\_At\_Home:** This column has 14 missing values, indicating that there are 59 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 59 missing values.

1. **Boxplot Analysis**
2. **Boxplot Visualization**

* The boxplot of ricetotal\_q was created to visually identify outliers. Boxplots display the distribution of data and highlight outliers as individual points beyond the whiskers.
* The plot shows many outliers above the upper whisker, indicating higher values of ricetotal\_qthat deviate significantly from the rest of the data.

**Outlier Removal**

1. **Outlier Removal Code**

* The code provided for removing outliers sets the acceptable range for ricetotal\_q values between low\_limit and up\_limit.
* The new DataFrame KA\_clean includes only those rows where ricetotal\_q values fall within this specified range.
* The second boxplot of KA\_clean['ricetotal\_q'] confirms the removal of outliers, showing a more compact range of values without extreme points.

**Interpretation**

* **Boxplot Interpretation:**

The initial boxplot for ricetotal\_q clearly indicates the presence of several outliers. These outliers can affect statistical analyses and modelingby skewing the results.

* **Outlier Handling:**
* By filtering the dataset to include only values within the low\_limit and up\_limit range, the cleaned dataset (KA\_clean) excludes these outliers. This process ensures a more reliable analysis by focusing on the central tendency of the data.
* The subsequent boxplot of KA\_clean['ricetotal\_q'] should show no or fewer outliers, indicating that the data has been effectively cleaned.

**C.)**

* **District Renaming:**

The District column in the kanew dataset is updated by replacing numeric district codes with their respective names from district\_mapping. Any district code that doesn't exist in district\_mapping remains unchanged.

* **Sector Renaming:**

The Sector column in the kanew dataset is updated similarly, replacing numeric sector codes (1 and 2) with "URBAN" and "RURAL", respectively.

* **Unique District Values:**

After the renaming process, the unique district values in the KA\_clean dataset should reflect the mapped names if the mapping is correctly applied. However, the output array([14, 13, 12, 11, 9, 10, 8, 7, 6, 5, 3, 4, 2, 1, 20, 29, 28, 27, 26, 19, 18, 17, 16, 15, 24, 25, 23, 21, 22], indicates that the unique district values are still numeric, suggesting that the KA\_clean DataFrame might not have been updated with the renamed district names yet.

* **Final Adjustments:**

Ensure that the correct dataset is being used for the renaming operations. Verify that the district\_mapping and sector\_mapping dictionaries include all necessary mappings for the dataset.