****

**VIRGINIA COMMONWEALTH UNIVERSITY**

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

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

**SAMPREETH S SHETTY**

**V01107634**

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

**CONTENTS**

|  |  |  |
| --- | --- | --- |
| **Sl. No.** | **Title** | **Page No.** |
| **1.** | Introduction | **1** |
| **2.** | Results | **3** |
| **3.** | Interpretations | **4** |
| **4.** | Recommendations | **5** |
| **5.** | Codes | **-** |
| **6.** | References | **5** |

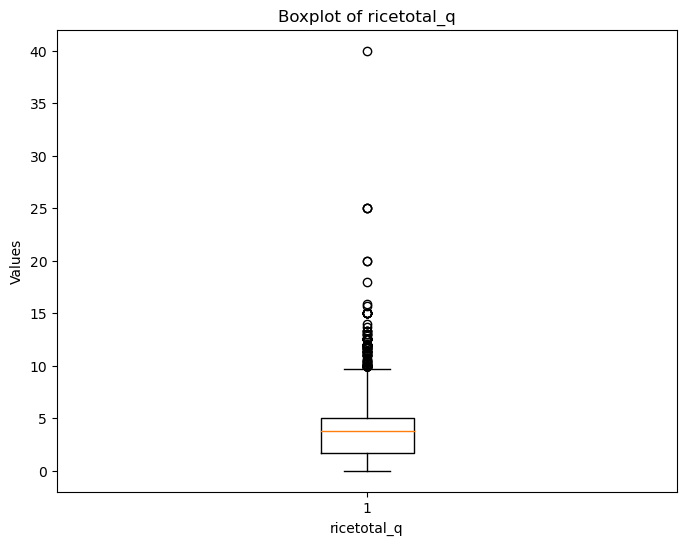
**Introduction**

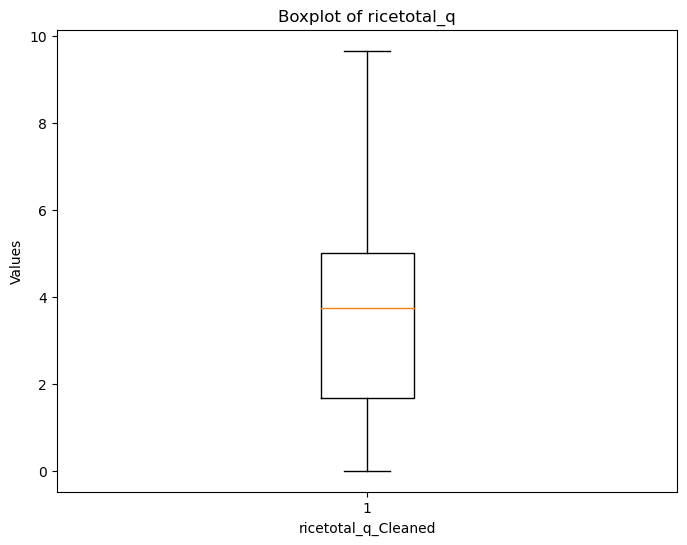
The NSSO (National Sample Survey Office) collected the dataset used in this project. The National Sample Survey Office (NSSO), headed by a Director General, is India’s premier data collection agency responsible for conducting large-scale sample surveys across diverse fields on an all-India basis. The NSSO is involved in significant surveys such as Socio-economic Surveys, Annual Surveys of Industries, and Agricultural Surveys. It takes full responsibility for the socio-economic survey from its design to the release of survey reports and collects field data for other surveys. The data collected through nationwide household surveys cover various socio-economic subjects, including housing conditions and basic amenities like drinking water, bathrooms, sewerage, lavatories, lighting, and more.

The particular data set in hand deals with the consumption of products, mainly food products, by the households of India. The households can be categorized based on the state and whether they belong to the rural or the urban sector.

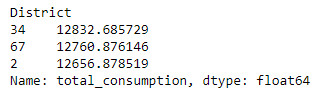
The state that was assigned was Uttar Pradesh. The data was cleaned, and only the data of Uttar Pradesh was taken for analysis. The null values after viewing were imputed with the mean values of the variable, the variables with a considerably more significant null value were dropped, and the only variable whose null values were imputed was ‘Meals\_AT\_Home.’

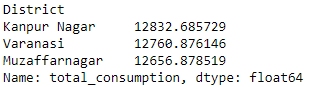
A boxplot was plotted for the variable ‘ricetotal\_q’ to check for any outliers, and upon finding them, the outliers were removed, and the boxplot was plotted again. Both the boxplots are below.





The data was further manipulated by converting the Sector values in codes to ‘Urban’ and ‘Rural.’ a new variable was added to the dataset called ‘total\_consumption,’ which was the sum of all the rows across the dataset, which indicated the total consumption of the products across the districts of Uttar Pradesh. The new variable's maximum, minimum, and standard deviations were then computed as per the regions and districts. The top three districts with the highest total consumption were found out and decoded to their actual district names in string values. The before and after of the same is below.





Finally, rural and urban entries were grouped separately from the cleaned data to perform a z-test.

**Results**

The hypothesis of this particular exercise is as follows

H\_0: There is no significant difference between the mean consumption of the Rural and Urban areas of Uttar Pradesh.

H\_1: There is a significant difference between the mean consumption of the Rural and Urban areas of Uttar Pradesh.

The output from Python for the test is



**Interpretation**

The z-score is 15.349380604793554, which is relatively high and indicates that the data point is many standard deviations from the mean. The p-value is extremely low (3.5758420317152064e-53), suggesting the result is highly statistically significant. Thus, we **reject the null hypothesis** and conclude that there is a substantial difference between the mean consumption of the Rural and urban parts of Uttar Pradesh.

**Recommendations**

The test results are apparent indicators that the consumption patterns of goods in rural households are significantly different from that of the urban household, this being about only the state of Uttar Pradesh because the test for said hypothesis was done for this state alone. The feasible recommendation that can be given to the NSSO is that they can further look into why there is a disparity between these two sectors in this State. However, it can also be assumed that this is a mere observation from the data set and might bear no significance.

**References**

(National Sample Survey Office(NSSO), 2024)