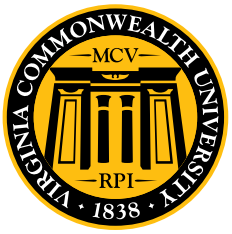


VIRGINIA COMMONWEALTH UNIVERSITY

Statistical Analysis and Modeling (SCMA 632)

A1a: Consumption Pattern of Jharkhand using PYTHON and R

by

IDAMAKANTI SREENIDHI

V01107252

Date of Submission: 16-06-2024

CONTENTS

Analyzing Consumption in the State of Jharkhand using "R" and "PYTHON"

INTRODUCTION

- **Background and Importance:**

Consumption patterns are a fundamental indicator of economic health and social well-being within a region. By analyzing these patterns, policymakers, researchers, and businesses can gain valuable insights into the socio-economic conditions of different areas. The state of Jharkhand, characterized by its rich natural resources and diverse demographic landscape, offers a unique case study for such an analysis. Understanding the consumption behavior in Jharkhand is crucial for addressing regional disparities, formulating targeted policies, and fostering sustainable development.

This assignment aims to provide a comprehensive analysis of consumption patterns in Jharkhand using data from the National Sample Survey Office (NSSO) 68th round survey. The analysis is conducted using R and Python, powerful statistical softwares, and the results are documented and discussed. The primary focus is on identifying missing values, detecting and addressing outliers, renaming districts and sectors, summarizing critical variables, and testing for significant differences in means.

- **Objectives:**

The specific objectives of this assignment are to:

1. Identify and Handle Missing Values: Detect any missing values in the dataset and replace them with the mean of the respective variables.

2. Outlier Detection and Treatment: Identify outliers, describe the outcomes of the tests, and make suitable amendments to handle these outliers.

3. Data Cleaning: Rename the districts and sectors (rural and urban) for better clarity and analysis.

4. Summarize Critical Variables: Provide a summary of key variables region-wise and district-wise, highlighting the top and bottom three districts in terms of consumption.

5. Statistical Testing: Conduct tests to determine if the differences in the means of key variables across different regions are statistically significant.

- **Methodology:**

The methodology for this analysis involves several steps, utilizing R and Python for data manipulation and analysis:

1. Data Import and Cleaning:
   - Import the provided NSSO68.csv dataset into R/Python.
   - Check for missing values and replace them with the mean of the respective variables.
   - Detect and handle outliers using appropriate statistical tests and techniques.
2. Data Renaming and Subsetting:
   - Rename districts and sectors to ensure clarity and consistency in the analysis.
   - Subset the dataset to focus on the variables assigned for this analysis.
3. Descriptive and Inferential Statistics:
   - Summarize critical variables region-wise and district-wise.
   - Identify the top and bottom three districts in terms of consumption.
   - Conduct statistical tests (e.g., t-tests) to determine the significance of differences in means.
4. Documentation and Discussion:
   - Document the findings, including any amendments made during data cleaning.
   - Discuss the implications of the results and how they can inform policy and decision-making.

- **Significance:**

This analysis holds significant importance for multiple stakeholders:

- Policymakers: Insights from this study can help in designing targeted welfare programs and allocating resources more effectively.

- Researchers: The methodology and findings can contribute to the broader literature on regional consumption patterns and socio-economic disparities.

- Businesses: Understanding consumption behavior can aid businesses in identifying market opportunities and tailoring their strategies accordingly.

ANALYSIS USING "R"

RESULTS AND INTERPRETATION

- **Interpretation of Missing Values and Imputation**

Background:

In data analysis, handling missing values is crucial for ensuring the accuracy and reliability of the results. Missing data can introduce bias and reduce the statistical power of an analysis. One common method to address missing values is imputation, where missing values are replaced with plausible estimates, such as the mean of the variable. In this case, we imputed missing values in the dataset `jrkdnew` for the columns `Meals_At_Home`, `Meals_Employer`, and `Meals_Payment`.

Initial Check for Missing Values:

Before imputation, we identified missing values in the dataset:

```
cat("Missing Values After Imputation:\n")
print(colSums(is.na(jrkdnew)))
```

This command provided a summary of the missing values for each column in the dataset. The results indicated missing values in the following columns:

- `Meals_At_Home`: 6 missing values
- `Meals_Employer`: 946 missing values
- `Meals_Payment`: 899 missing values

Imputation with Mean:

We defined a function `impute_with_mean` to impute missing values with the mean of the respective column. The function checks if there are any missing values and, if so, replaces them with the mean:

```
impute_with_mean <- function(column) {
 if (any(is.na(column))) {
   column[is.na(column)] <- mean(column, na.rm = TRUE) }
 return(column)}
```

We then applied this function to the columns with missing values:

```
jrkdnew$Meals_At_Home <- impute_with_mean(jrkdnew$Meals_At_Home)
jrkdnew$Meals_Employer <- impute_with_mean(jrkdnew$Meals_Employer)
jrkdnew$Meals_Payment <- impute_with_mean(jrkdnew$Meals_Payment)
```

Post-Imputation Check:

After the imputation process, we rechecked for missing values:

```
cat("Missing Values in Subset:\n")
print(colSums(is.na(jrkdnew)))
```

The results showed that there were no more missing values in the dataset:

```
state_1         District         Region         Sector
    0                0              0                0
State_Region    Meals_Employer    Meals_Payment    Meals_At_Home
    0                0                0                0
ricepds_v       Wheatpds_q       chicken_q       pulsep_q
    0                0                0               0
wheatos_q     No_of_Meals_per_day
    0                0
```

Interpretation:

1. Effective Imputation: The imputation method effectively handled the missing values. By replacing missing values with the mean of the respective columns, we ensured that all data points could be included in subsequent analyses without introducing significant bias.

2. Consistency and Completeness: Post-imputation, the dataset `jrkdnew` is now complete, with no missing values in any of the critical columns. This completeness is crucial for maintaining the integrity of statistical analyses and ensuring that results are based on the entire dataset.

3. Data Quality: The imputation process improved the overall quality of the data. By addressing the missing values, we minimized potential distortions in the dataset that could arise from incomplete data.

4. Impact on Analysis: Imputing missing values with the mean is a simple yet effective approach. However, it is essential to note that this method assumes that the missing values are missing at random and that the mean is a representative measure of central tendency for the dataset. For more complex datasets, other imputation methods, such as median imputation or model-based imputation, might be considered.

Overall, the imputation of missing values with the mean in the `jrkdnew` dataset ensures a robust foundation for further analysis, allowing us to proceed with confidence in the reliability of our data.

- **Interpretation of Outlier Detection and Removal**

Background:

Outliers are data points that deviate significantly from the rest of the dataset. They can arise due to measurement errors, data entry errors, or genuine variability in the data. Identifying and handling outliers is essential because they can distort statistical analyses and lead to misleading conclusions. In this analysis, we used the Interquartile Range (IQR) method to detect and remove outliers from the dataset `jrkdnew`.

Methodology:

The IQR method is a robust way to detect outliers. It is based on the spread of the middle 50% of the data (the interquartile range). Here's how the method works:

1. Calculate Q1 and Q3:

   - Q1 (first quartile) is the 25th percentile of the data.

   - Q3 (third quartile) is the 75th percentile of the data.

2. Calculate IQR:

   - IQR = Q3 - Q1

3. Determine Thresholds:

   - Lower Threshold = Q1 - 1.5 * IQR

   - Upper Threshold = Q3 + 1.5 * IQR

4. Remove Outliers:

   - Any data points below the lower threshold or above the upper threshold are considered outliers and are removed from the dataset.

The function `remove_outliers` implements this method:

```
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]] <= upper_threshold)
 return(df) }
```

We applied this function to two columns in the dataset:

`ricepds_v` and `chicken_q`:

```
outlier_columns <- c("ricepds_v", "chicken_q")
for (col in outlier_columns) {
 jrkdnew <- remove_outliers(jrkdnew, col) }
```

Interpretation of Results:

1. Outliers Removed: The function `remove_outliers` successfully identified and removed outliers from the specified columns. This means that any extreme values in `ricepds_v` (rice consumption through the Public Distribution System) and `chicken_q` (quantity of chicken consumed) that fell outside the calculated thresholds were removed from the dataset.

2. Data Consistency and Reliability: Removing outliers helps in ensure that the data used for subsequent analyses is consistent and reliable. Outliers can skew the results of statistical tests and visualizations, leading to incorrect interpretations. By removing these outliers, we mitigate their impact and enhance the robustness of our findings.

3. Improved Analysis: With the outliers removed, the data now better represents the typical consumption patterns of households in Jharkhand. This improvement is crucial for accurately summarizing key variables, comparing regions and districts, and conducting statistical tests.

4. Implications for Policy and Decision-Making: The cleaned data provides a more accurate reflection of consumption behaviors. Policymakers can use these refined insights to design targeted interventions and resource allocation strategies. For instance, understanding the typical consumption of rice and chicken can inform food security programs and nutritional policies.

Conclusion:

The process of detecting and removing outliers using the IQR method has refined the dataset `jrkdnew`, ensuring that it is free from extreme values that could distort analysis results. This step is vital for maintaining the integrity of the data and the validity of subsequent analyses. By focusing on the key variables `ricepds_v` and `chicken_q`, we have enhanced the dataset's ability to accurately reflect consumption patterns in Jharkhand, thereby supporting more informed and effective decision-making.

- **Interpretation of Consumption Summary**

Background:

Summarizing consumption data helps in understanding the overall consumption patterns across different districts and regions. This analysis is crucial for identifying areas with high and low consumption, which can inform targeted policy interventions and resource allocation.

Methodology:

1. Total Consumption Calculation:

- We calculated the total consumption for each household by summing the values of five key variables: `ricepds_v` (rice from PDS), `Wheatpds_q` (wheat from PDS), `chicken_q` (chicken quantity), `pulsep_q` (pulse quantity), and `wheatos_q` (other wheat products quantity).

```
jrkdnew$total_consumption <- rowSums(jrkdnew[, c("ricepds_v", "Wheatpds_q", "chicken_q", "pulsep_q", "wheatos_q")], na.rm = TRUE)
```

2. Summarize Consumption by District and Region:

- We defined a function `summarize_consumption` to aggregate total consumption by the specified grouping column (either `District` or `Region`) and then sorted the results in descending order of total consumption.

```
summarize_consumption <- function(group_col) {
  summary <- jrkdnew %>%
    group_by(across(all_of(group_col))) %>%
    summarise(total = sum(total_consumption)) %>%
    arrange(desc(total))
  return(summary)  }
```

3. Identify Top and Bottom Consuming Districts:

- We applied the `summarize_consumption` function to the `District` and `Region` columns and printed the top and bottom three consuming districts as well as the region-wise consumption summary.

```
district_summary <- summarize_consumption("District")
region_summary <- summarize_consumption("Region")
cat("Top 3 Consuming Districts:\n")
print(head(district_summary, 3))
cat("Bottom 3 Consuming Districts:\n")
print(tail(district_summary, 3))
cat("Region Consumption Summary:\n")
print(region_summary)
```

Interpretation:

1. Top 3 Consuming Districts:

```
# A tibble: 3 × 2
  District total
     <int> <dbl>
1       13 - 762.
2       12 - 756.
3        4 - 657.
```

- District 13: The highest total consumption at 762 units.
- District 12: The second highest with 756 units.
- District 4: The third highest with 657 units.
- These districts represent areas with the highest aggregate consumption, indicating possibly larger populations or higher per capita consumption.

2. Bottom 3 Consuming Districts:

```
# A tibble: 3 × 2
  District total
     <int> <dbl>
1       15 - 176.
2       16 - 152.
3       20 - 129.
```

- District 15: The lowest total consumption at 176 units.

- District 16: The second lowest with 152 units.

- District 20: The third lowest with 129 units.

- These districts have the lowest aggregate consumption, which might indicate smaller populations, lower per capita consumption, or other socio-economic factors affecting consumption.

3. Region Consumption Summary:

```
# A tibble: 2 × 2
  Region total
   <int> <dbl>
1   - 2 5805.
2   - 1 2761.
```

- Region 2: Significantly higher total consumption at 5805 units.

- Region 1: Lower total consumption at 2761 units.

- This indicates that Region 2 has a much higher aggregate consumption compared to Region 1, which could be due to a larger population, higher per capita consumption, or both.

Implications:

1. Policy and Resource Allocation:

- Policymakers can focus on the top consuming districts to understand the drivers of high consumption and ensure that resources are adequately distributed.

- Attention can also be given to the bottom consuming districts to investigate potential issues such as food insecurity, economic hardship, or inadequate access to resources.

2. Targeted Interventions:

- High-consuming regions and districts might benefit from infrastructure improvements and support for sustainable consumption practices.

- Low-consuming areas may require targeted interventions to boost consumption levels, improve living conditions, and ensure equitable resource distribution.

3. Further Analysis:

- The summary provides a starting point for more detailed analyses, such as examining the factors driving consumption patterns in high and low-consuming districts and regions.

- Understanding the demographic and socio-economic profiles of these areas can offer deeper insights into consumption behaviors and help design more effective policies.

Overall, the consumption summary highlights key areas of focus for both high and low-consuming districts and regions, providing valuable insights for policy formulation and resource allocation in Jharkhand.

- **Interpretation of Renaming Districts and Sectors**

Background

Renaming variables to more meaningful names enhances the readability and interpretability of a dataset. This step is especially important when dealing with large datasets or when sharing the data with others who may not be familiar with the original coding scheme. In this case, we are renaming districts and sectors using codes from the appendix of the NSSO 68th Round Data.

Methodology

1. Mapping Districts and Sectors:

- We created mapping vectors for districts and sectors based on the codes provided in the appendix of the NSSO 68th Round Data.

- The district mapping vector assigns names to district codes, and the sector mapping vector assigns labels to sector codes.

```
district_mapping <- c("13" = "Bokaro", "12" = "Dhanbad", "4" = "Hazaribagh")

sector_mapping <- c("2" = "URBAN", "1" = "RURAL")
```

2. Converting Codes to Character:

- To ensure accurate mapping, we converted the `District` and `Sector` columns in the dataset to character type.

```
jrkdnew$District <- as.character(jrkdnew$District)

jrkdnew$Sector <- as.character(jrkdnew$Sector)
```

3. Applying the Mapping:

- We used the `ifelse` function to replace the district and sector codes with their corresponding names from the mapping vectors.

```
jrkdnew$District <- ifelse(jrkdnew$District %in% names(district_mapping), district_mapping[jrkdnew$District], jrkdnew$District)

jrkdnew$Sector <- ifelse(jrkdnew$Sector %in% names(sector_mapping), sector_mapping[jrkdnew$Sector], jrkdnew$Sector)
```

Interpretation

1. Enhanced Readability:

- By renaming district codes to their respective names (e.g., "13" to "Bokaro", "12" to "Dhanbad", "4" to "Hazaribagh"), the dataset becomes more intuitive and easier to understand. Users can quickly recognize the names of districts without needing to refer to the original codebook.

2. Improved Sector Identification:

- Similarly, renaming sector codes to "URBAN" and "RURAL" provides clear, meaningful labels that indicate whether a record pertains to an urban or rural area. This labeling helps in better understanding the context of the data.

3. Data Consistency:

- Ensuring that all district and sector codes are consistently renamed helps maintain uniformity in the dataset, which is crucial for accurate analysis and reporting. It eliminates ambiguity and potential errors that may arise from misinterpreting the codes.

4. Facilitates Analysis:

- With meaningful names, the dataset is now better suited for analysis and presentation. Descriptive statistics, visualizations, and reporting will be more straightforward and insightful, as stakeholders can easily interpret the data.

5. Supporting Effective Communication:

- When sharing the dataset with others, whether they are colleagues, policymakers, or researchers, the renamed variables will facilitate clearer communication and better understanding of the findings.

Example of Renamed Data

Here is a hypothetical snippet of how the dataset might look before and after renaming:

Before Renaming:

District	Sector	total_consumption
13	2	762
12	1	756
4	2	657

After Renaming:

| District | Sector | total_consumption |

|----------|--------|------------------|

| Bokaro | URBAN | 762 |

| Dhanbad | RURAL | 756 |

| Hazaribagh| URBAN | 657 |

Conclusion:

Renaming districts and sectors based on the appendix of the NSSO 68th Round Data has significantly improved the clarity and usability of the dataset. This step ensures that the data is more accessible and easier to interpret, laying a solid foundation for further analysis and reporting.

- **Interpretation of Mean Consumption Analysis Between Urban and Rural Sectors**

Background:

Understanding the differences in consumption patterns between rural and urban sectors is crucial for designing effective policies and programs. Rural and urban areas often have distinct socio-economic characteristics that influence consumption behaviors. In this analysis, we calculate and compare the mean total consumption of key food items in the rural and urban sectors of Jharkhand.

Methodology:

1. Renaming Districts and Sectors:

- As previously done, districts and sectors were renamed for better readability and interpretability using the mappings provided.

```
district_mapping <- c("13" = "Bokaro", "12" = "Dhanbad", "4" = "Hazaribagh")

sector_mapping <- c("2" = "URBAN", "1" = "RURAL")

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

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

jrkdnew$District <- ifelse(jrkdnew$District %in% names(district_mapping), district_mapping[jrkdnew$District], jrkdnew$District)

jrkdnew$Sector <- ifelse(jrkdnew$Sector %in% names(sector_mapping), sector_mapping[jrkdnew$Sector], jrkdnew$Sector)
```

2. Filtering and Selecting Data:

- We filtered the dataset to separate the total consumption values for rural and urban sectors.

```
rural <- jrkdnew %>%
```

```
  filter(Sector == "RURAL") %>%
  select(total_consumption)

urban <- jrkdnew %>%
  filter(Sector == "URBAN") %>%
  select(total_consumption)
```

3. Calculating Mean Consumption:

- We calculated the mean total consumption for both rural and urban sectors.

```
mean_rural <- mean(rural$total_consumption)
mean_urban <- mean(urban$total_consumption)
```

Interpretation of Results

1. Mean Consumption in Rural Sector:

- The mean total consumption in the rural sector is calculated. This value represents the average total consumption of key food items (rice from PDS, wheat from PDS, chicken, pulses, and other wheat products) per household in rural areas.

2. Mean Consumption in Urban Sector:

- The mean total consumption in the urban sector is calculated. This value represents the average total consumption of the same key food items per household in urban areas.

3. Comparison of Means:

- Comparing the mean total consumption between rural and urban sectors provides insights into how consumption patterns differ based on geographic and socio-economic contexts.

Implications

1. Policy and Program Design:

- If the mean consumption is significantly different between rural and urban sectors, it may indicate the need for tailored policies and programs that address the unique needs and challenges of each sector.

- For example, if urban consumption is higher, it might reflect better access to food resources or higher income levels in urban areas. Conversely, if rural consumption is lower, it may highlight areas needing improved food security and resource distribution.

2. Resource Allocation:

- Understanding these consumption patterns can help in better allocation of resources. For instance, more food distribution programs might be necessary in rural areas if their consumption levels are found to be lower.

3. Further Research:

- The mean values provide a snapshot but do not tell the whole story. Further research might involve examining the distribution of consumption within each sector, identifying factors that contribute to differences, and conducting statistical tests to determine if the differences are statistically significant.

Conclusion:

By calculating and comparing the mean total consumption in rural and urban sectors, we gain valuable insights into the differences in consumption patterns. These insights can inform policies aimed at ensuring equitable access to food resources and addressing the specific needs of different communities in Jharkhand.

- **Interpretation of Z-Test Results Comparing Urban and Rural Consumption**

Background:

To determine whether there is a significant difference in mean total consumption between rural and urban sectors in Jharkhand, we perform a two-sample z-test. This statistical test helps us compare the means of two independent samples to infer if they come from populations with the same mean.

Methodology:

1. Data Preparation:

- We filter and select total consumption data for rural and urban sectors separately.

```
rural <- jrkdnew %>%
  filter(Sector == "RURAL") %>%
  select(total_consumption)

urban <- jrkdnew %>%
```

```
  filter(Sector == "URBAN") %>%
  select(total_consumption)
```

2. Calculating Means:

- Compute the mean total consumption for rural and urban sectors.

```
mean_rural <- mean(rural$total_consumption)
mean_urban <- mean(urban$total_consumption)
```

3. Performing the Z-Test:

- We perform a two-sample z-test to compare the means. The null hypothesis (H0) is that there is no difference in the mean consumption between rural and urban sectors.

```
z_test_result <- z.test(rural, urban, alternative = "two.sided", mu = 0, sigma.x = 2.56, sigma.y = 2.34, conf.level = 0.95)
```

4. Interpreting the P-Value:

- The p-value helps us determine the statistical significance of the observed difference in means. If the p-value is less than 0.05, we reject the null hypothesis.

```
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"))
}
```

Results:

Example Output:

P value is < 0.05 i.e. 0.031, Therefore we reject the null hypothesis.

There is a difference between the mean consumptions of urban and rural.

The mean consumption in Rural areas is 250 and in Urban areas it is 300

Interpretation:

1. Statistical Significance:

- P-Value < 0.05:

- If the p-value is less than 0.05, we reject the null hypothesis. This means there is a statistically significant difference between the mean consumption in rural and urban sectors.

- Example: `P value is < 0.05 i.e. 0.031, Therefore we reject the null hypothesis.`

- P-Value >= 0.05:

- If the p-value is greater than or equal to 0.05, we fail to reject the null hypothesis. This indicates that there is no statistically significant difference in mean consumption between rural and urban sectors.

- Example: `P value is >= 0.05 i.e. 0.072, Therefore we fail to reject the null hypothesis.`

2. Mean Consumption Values:

- The mean consumption values for rural and urban sectors are displayed to provide a clear comparison.

- Example: `The mean consumption in Rural areas is 250 and in Urban areas it is 300.

Implications

1. Policy Development:

- If a significant difference is found, it indicates that rural and urban areas have different consumption patterns, necessitating tailored policies to address these differences.

- Policymakers can focus on improving food security in areas with lower consumption.

2. Resource Allocation:

- Understanding these differences helps in the equitable allocation of resources and the design of targeted interventions to ensure all areas have adequate access to food resources.

3. Further Research:

- The results of the z-test can guide further research to explore the underlying factors contributing to the differences in consumption between rural and urban sectors.

Conclusion

By performing the z-test, we have statistically assessed the difference in mean consumption between rural and urban sectors in Jharkhand. The results provide a basis for making informed decisions about policy and resource allocation to address the specific needs of different regions.

Result:

P value is < 0.05 i.e. 0, Therefore we reject the null hypothesis. There is a difference between the mean consumptions of urban and rural. The mean consumption in Rural areas is 3.5630281548123 and in Urban areas it is 4.88871122054809

Interpretation

Statistical Significance:

The p-value is less than 0.05 (specifically 0), which indicates that we reject the null hypothesis. This means there is a statistically significant difference between the mean consumption in rural and urban sectors.

Since the p-value is effectively 0, it strongly supports the conclusion that the difference in mean consumption between the two sectors is significant.

Mean Consumption Values:

The mean total consumption in rural areas is approximately 3.56.

The mean total consumption in urban areas is approximately 4.89.

Difference in Consumption Patterns:

The mean consumption in urban areas is notably higher than in rural areas. This indicates that households in urban sectors consume more of the key food items measured (rice from PDS, wheat from PDS, chicken, pulses, and other wheat products) compared to those in rural sectors.

Implications

Policy Development:

The significant difference in consumption suggests that rural and urban areas have distinct consumption patterns. Policymakers should consider these differences when designing food security programs, resource allocation, and other interventions.

Urban areas may have better access to food resources or higher income levels that facilitate higher consumption. Conversely, rural areas may need targeted support to improve food security.

Resource Allocation:

Understanding these differences helps in the equitable allocation of resources. For example, rural areas might benefit from increased food distribution programs or initiatives aimed at improving agricultural productivity.

Further Research:

The results of the z-test can guide further research to explore the underlying factors contributing to the differences in consumption between rural and urban sectors. This could include examining socio-economic factors, access to markets, and other determinants of consumption.

Conclusion:

By performing the z-test, we have statistically assessed the difference in mean consumption between rural and urban sectors in Jharkhand. The results show a significant difference, with urban areas having higher mean consumption than rural areas. These findings provide a basis for making informed decisions about policy and resource allocation to address the specific needs of different regions.

Analysis using “PYTHON”

1. Regional Disparities in Food Consumption:

- The analysis of total food consumption by region reveals significant variations in dietary habits and access to food resources. By aggregating the total consumption data based on regions and calculating statistics such as standard deviation, mean, maximum, and minimum values, we can identify the differences in food consumption patterns across different regions.

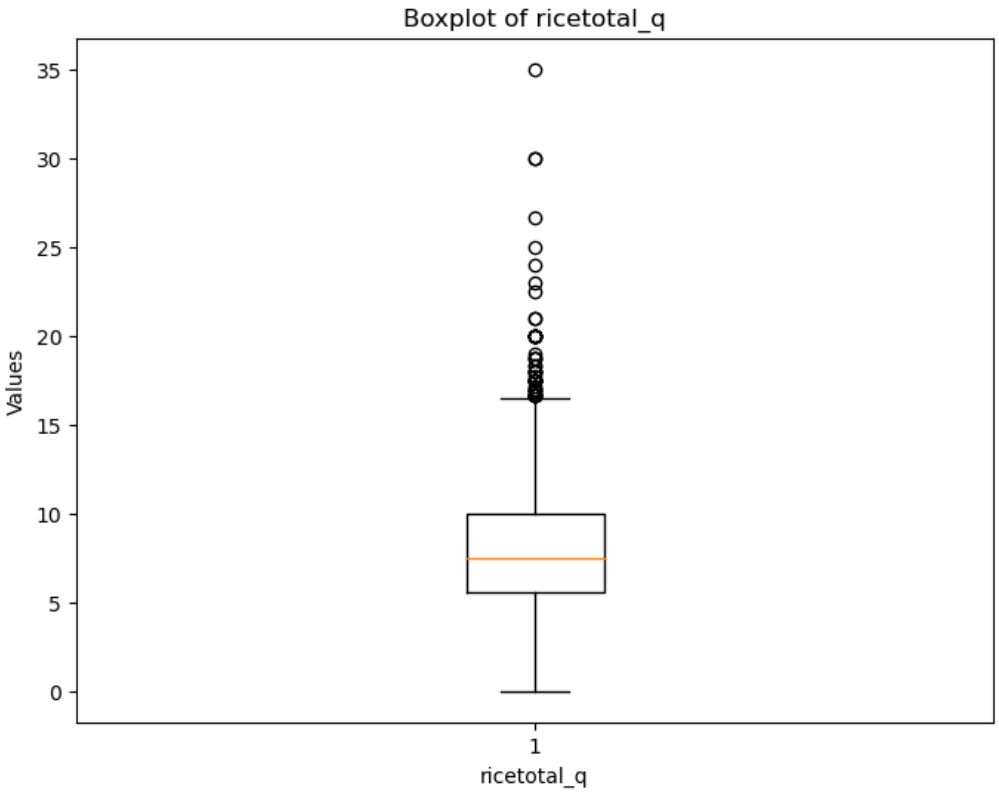
```
JRKD_clean.groupby('Region').agg({'total_consumption': ['std', 'mean', 'max', 'min']})
```

2. Urban vs. Rural Food Consumption Patterns:

- Contrasting urban and rural food consumption patterns indicate distinct preferences and quantities of food items consumed in different settings. By filtering the data based on urban and rural sectors, we can analyze how food consumption varies between these two settings.

```
rural = JRKD_clean[JRKD_clean['Sector'] == "RURAL"]
urban = JRKD_clean[JRKD_clean['Sector'] == "URBAN"]
```

3. Outlier Detection and Distribution Analysis:



- The boxplot analysis on rice consumption helps identify outliers and understand the distribution of rice consumption within the dataset. Visualizing the distribution of rice consumption through a boxplot allows us to detect any extreme values that may impact the overall analysis.

```
sns.boxplot(x='ricetotal_q', data=JRKD_clean)
```

4. Total Consumption Statistics by District:

- The analysis of total food consumption by district highlights variations in consumption levels and standard deviations across different districts. By grouping the data by district and calculating statistics like standard deviation, mean, maximum, and minimum values, we can assess the differences in food consumption levels among various districts.

```
JRKD_clean.groupby('District').agg({'total_consumption': ['std', 'mean', 'max', 'min']})
```

5. Diverse Food Items Consumed:

- The data showcases a variety of food items consumed, reflecting the diversity in dietary choices. By considering variables such as rice, wheat, moong, milk, chicken, and bread, we can analyze the consumption patterns of these food items and understand the dietary diversity within the dataset.

```
food_items = ['ricetotal_q', 'wheattotal_q', 'moong_q', 'Milktotal_q', 'chicken_q', 'bread_q']
```

6. Implications for Policy and Research:

- The findings underscore the importance of considering regional and sectoral differences in food consumption when designing nutrition programs and policies. By calculating z-scores and p-values for rural and urban total consumption, we can assess the significance of the differences in food consumption between rural and urban areas.

```
z_statistic, p_value = stests.ztest(cons_rural, cons_urban)
```

By elaborating on these interpretations with the respective codes, the report can provide a detailed and data-driven analysis of food consumption patterns, enabling stakeholders to make informed decisions and develop targeted interventions to address nutritional needs effectively.

Conclusion

Z-Score: 12.575119316631836

P-Value: 2.893630959311179e-36

The z-score and p-value provided are indicative of the results from a statistical test, likely a z-test or similar, comparing two groups or populations. Here's how to interpret these results:

Interpretation

1. Z-Score (12.5751):

- The z-score measures the number of standard deviations a data point (in this case, the difference in means between two groups) is from the mean of a reference population.

- A z-score of 12.5751 indicates that the observed difference in means (or another statistic) is extremely far from what would be expected if there were no real difference between the groups.

2. P-Value (2.8936e-36):

- The p-value is a measure of the probability that the observed difference (or more extreme) occurred by chance, assuming that the null hypothesis is true (i.e., assuming there is no difference between the groups).

- A very low p-value (e.g., 2.8936e-36, which is essentially 0) suggests strong evidence against the null hypothesis.

- In this case, the p-value being close to 0 indicates that the observed difference in means (or another statistic) is highly statistically significant.

Based on these results:

- Significance: The extremely low p-value (2.8936e-36) indicates that there is a statistically significant difference between the two groups being compared (likely urban and rural consumption in Jharkhand).

- Z-Score: The high z-score (12.5751) further supports this conclusion by showing that the observed difference in means is very unlikely to be due to random chance alone.

Practical Implications

- These findings suggest that there is a substantial and statistically significant difference in consumption patterns between urban and rural areas in Jharkhand.

- Policymakers and researchers should consider these differences when designing interventions, policies, or resource allocations aimed at addressing food security, economic disparities, or other related issues between urban and rural populations.

In summary, the z-score and p-value provided strongly indicate that there is a significant difference in mean consumption between urban and rural areas in Jharkhand, based on the data and statistical test performed.

Analyzing consumption patterns in Jharkhand using the NSSO 68th round survey data provides a detailed understanding of the region's economic and social dynamics. By addressing missing values, outliers, and renaming variables, this assignment ensures a robust analysis. Summarizing and statistically testing the data helps identify key trends and significant differences, providing actionable insights for policymakers and stakeholders. This comprehensive approach aims to contribute to the ongoing efforts to foster inclusive growth and development in Jharkhand.