VIRGINIA COMMONWEALTH UNIVERSITY

Statistical analysis and modelling (SCMA 632)

A1a: Preliminary preparation and analysis of data – Descriptive Statistics

VIJAYATHITHYAN B B

V01107268

Date of Submission: 16-06-2024

CONTENTS

Analysing Consumption in the state of Bihar using R

INTRODUCTION

The focus of the study is consumption patterns across Bihar. We leverage data from the National Sample Survey Organisation (NSSO) to identify the districts with the highest and lowest consumption levels. To achieve this, we will meticulously clean and analyze the dataset, focusing on each district's rural and urban sectors.

R, a powerful statistical software renowned for its data manipulation capabilities, is our primary data exploration and analysis tool.

The insights gleaned from this study are not just academic, but hold significant value for policymakers and stakeholders. By identifying areas with high and low consumption, we can guide targeted interventions that promote a more equitable distribution of resources and foster balanced development across the entire state of Bihar. This has the potential to drive significant change and improve the lives of the people in Bihar.

OBJECTIVES

Our investigation will encompass several key objectives:

1. Identifying and addressing any missing values within the dataset.
2. Detecting and handling outliers that might skew the results.
3. Standardizing district and sector names for consistency.
4. Summarizing consumption data at both the regional and district levels.
5. Employing statistical tests to assess the significance of differences in average consumption between districts.

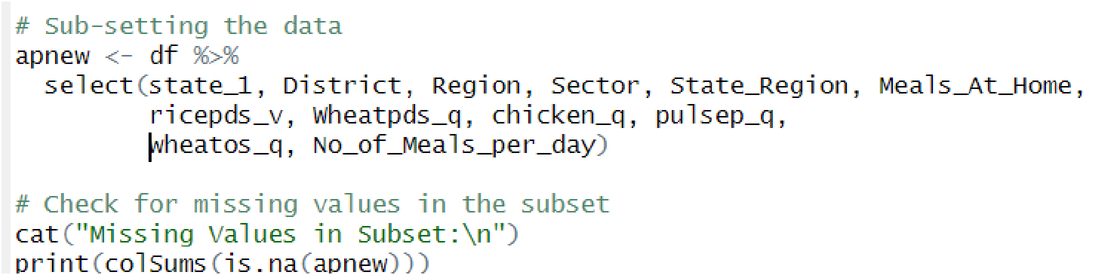
BUSINESS SIGNIFICANCE

Unveiling consumption patterns in Bihar through NSSO data offers a treasure trove of information for businesses and policymakers alike. Identifying the districts with the highest and lowest consumption levels sheds light on ideal locations for market entry, efficient resource allocation, and optimized supply chains. Additionally, the study employs data cleaning, outlier detection, and rigorous testing to ensure reliable findings. This robust approach empowers informed decision-making, ultimately driving balanced development and propelling Bihar's economic prosperity.

RESULTS AND INTERPRETATION

A. Identification of missing values in the data, and replacing them with the average value of the variables:

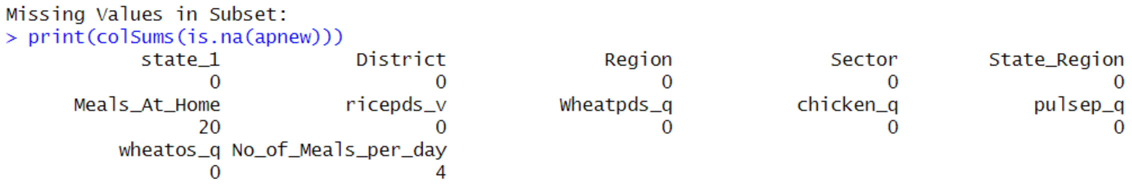
Code used:



```
# Sub-setting the data
apnew <- 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")
print(colSums(is.na(apnew)))
```

Output:

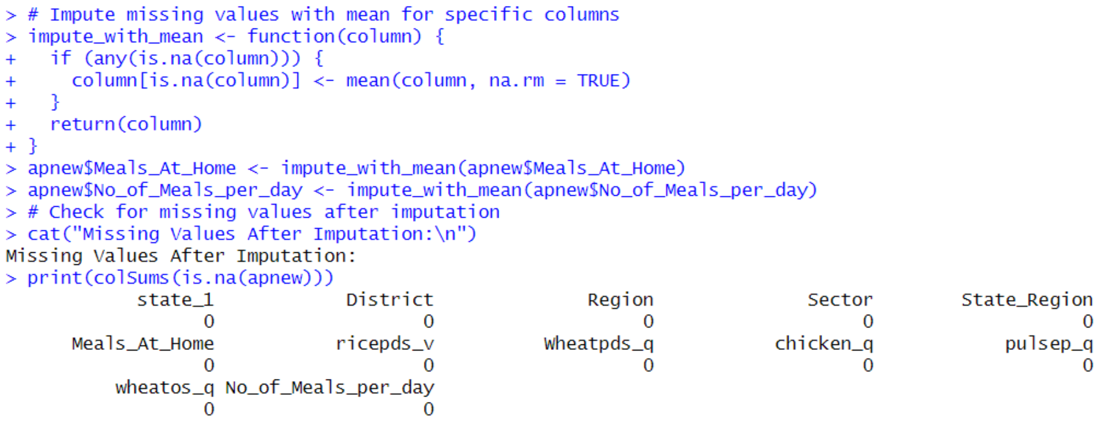


```
Missing Values in Subset:
> print(colSums(is.na(apnew)))
             state_1            District              Region              Sector        State_Region 
                   0                   0                   0                   0                   0 
       Meals_At_Home           ricepds_v          Wheatpds_q           chicken_q            pulsep_q 
                  20                   0                   0                   0                   0 
           wheatos_q No_of_Meals_per_day 
                   0                   4 
```

Interpretation: From the selected variables, after sorting the data for the state of Bihar, it is seen that only the column 'Meals_At_Home has 20 missing values and 'No_of_Meals_per_day has 4 missing values. Missing values in the dataset can be problematic; they lead to incomplete or biased analyses, hindering the accuracy of results and potentially skewing interpretations and decision-making processes. Therefore, we replace the missing values with the variable's mean using the following code.

# Imputing the missing values in variables 'Meals_At_Home and 'No_of_Meals_per_day with their respective mean

Code and Results:



```
> # Impute missing values with mean for specific columns
> impute_with_mean <- function(column) {
+   if (any(is.na(column))) {
+     column[is.na(column)] <- mean(column, na.rm = TRUE)
+   }
+   return(column)
+ }
> apnew$Meals_At_Home <- impute_with_mean(apnew$Meals_At_Home)
> apnew$No_of_Meals_per_day <- impute_with_mean(apnew$No_of_Meals_per_day)
> # Check for missing values after imputation
> cat("Missing Values After Imputation:\n")
Missing Values After Imputation:
> print(colSums(is.na(apnew)))
             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 
```

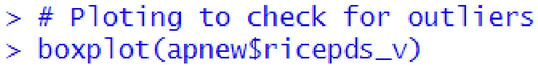
Interpretation: The above code has replaced the missing values in variables variables 'Meals_At_Home and 'No_of_Meals_per_day with their respective mean.

B. Checking for outliers and the outcome of the test with suitable amendments.

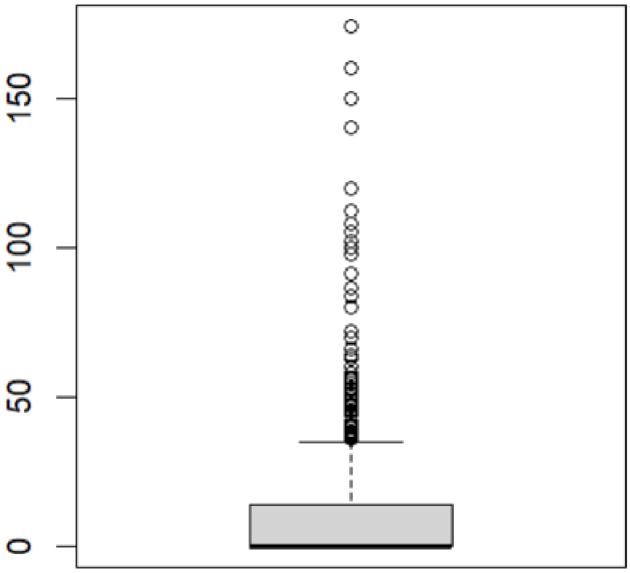
Boxplot is used to find outliers in a dataset. It is a visual representation that revels outliers in a dataset by plotting individual data points located beyond the whiskers of the boxplot.

# Checking for outliers

Code and Result:



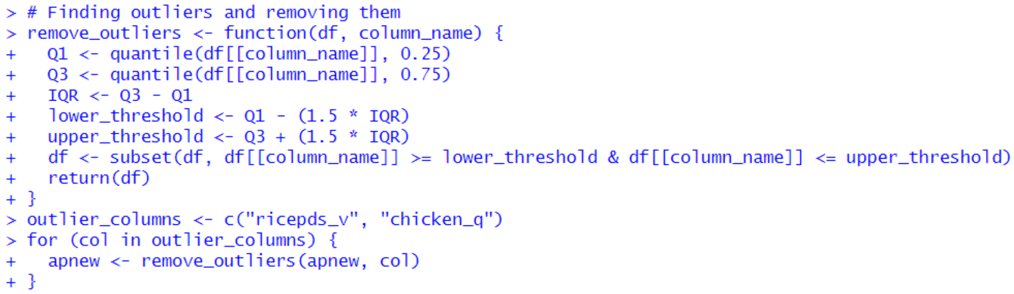
```
> # Ploting to check for outliers
> boxplot(apnew$ricepds_v)
```



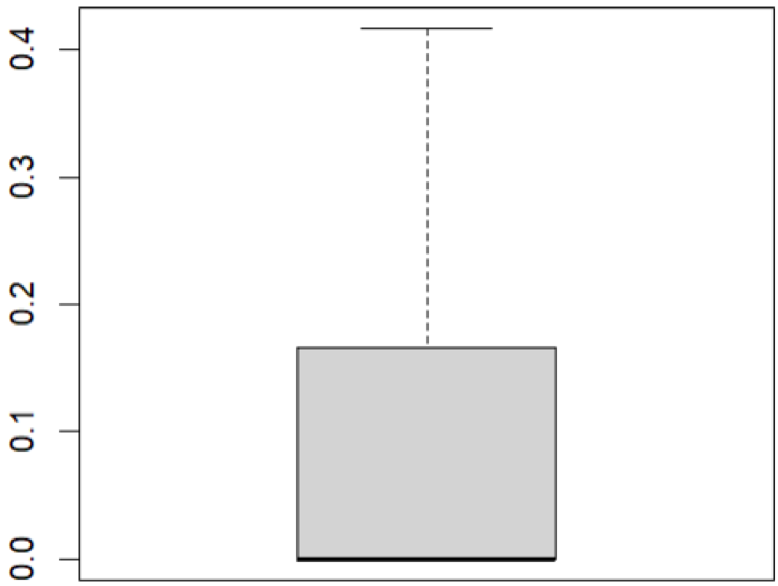
Interpretation: This boxplot reveals a substantial spread in consumption levels across Bihar's districts. The median consumption is around 50, with the middle half of districts between 0 and 150. However, whiskers reaching 0 and 200 indicate a range extending beyond that, with outliers potentially skewing the distribution towards higher consumption in some districts.

# Setting quartiles and removing outliers

Code and Result:



```
> # 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]] <= upper_threshold)
+   return(df)
+ }
> outlier_columns <- c("ricepds_v", "chicken_q")
> for (col in outlier_columns) {
+   apnew <- remove_outliers(apnew, col)
+ }
```

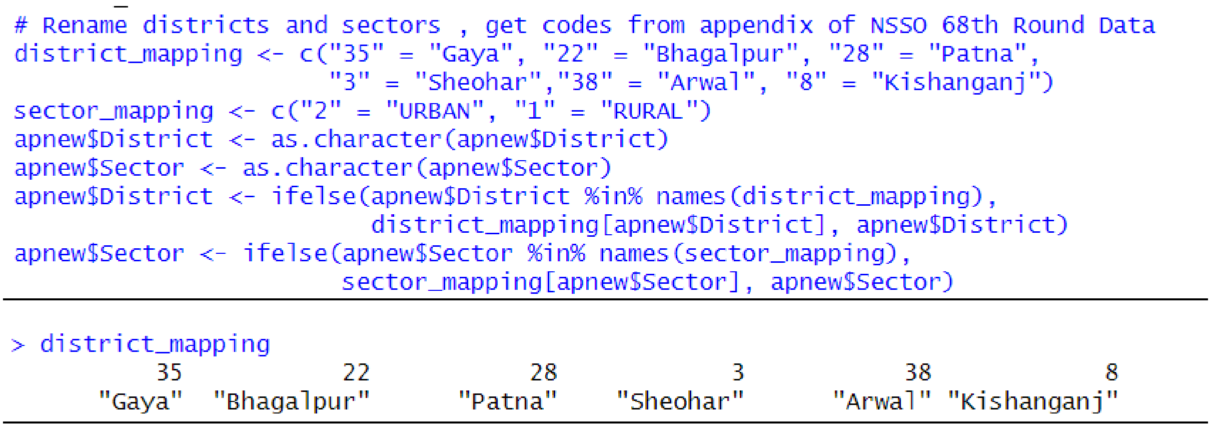


Interpretation: Interpreting quartile ranges allows for outlier detection and removal. By calculating the interquartile range (IQR) as the difference between the upper and lower quartiles, data points beyond 1.5 times the IQR from either quartile are identified as outliers and can be excluded or treated to ensure the robustness of the analysis. In a similar way the outliers in all other variables can be removed.

C. Renaming the District and Sector, viz. rural and urban

The NSSO data assigns unique numerical identifiers to districts within a state. However, to identify the state's top consumers, we need to know the actual names corresponding to these numbers. This is a crucial step in our data analysis. Likewise, urban and rural sectors are simply coded as 1 and 2, respectively. To address this, we'll utilize the following code snippet, which is instrumental in achieving our goal.

Code and Results:



```
# Rename districts and sectors , get codes from appendix of NSSO 68th Round Data
district_mapping <- c("35" = "Gaya", "22" = "Bhagalpur", "28" = "Patna",
                      "3" = "Sheohar","38" = "Arwal", "8" = "Kishanganj")
sector_mapping <- c("2" = "URBAN", "1" = "RURAL")
apnew$District <- as.character(apnew$District)
apnew$Sector <- as.character(apnew$Sector)
apnew$District <- ifelse(apnew$District %in% names(district_mapping),
                         district_mapping[apnew$District], apnew$District)
apnew$Sector <- ifelse(apnew$Sector %in% names(sector_mapping),
                       sector_mapping[apnew$Sector], apnew$Sector)
```

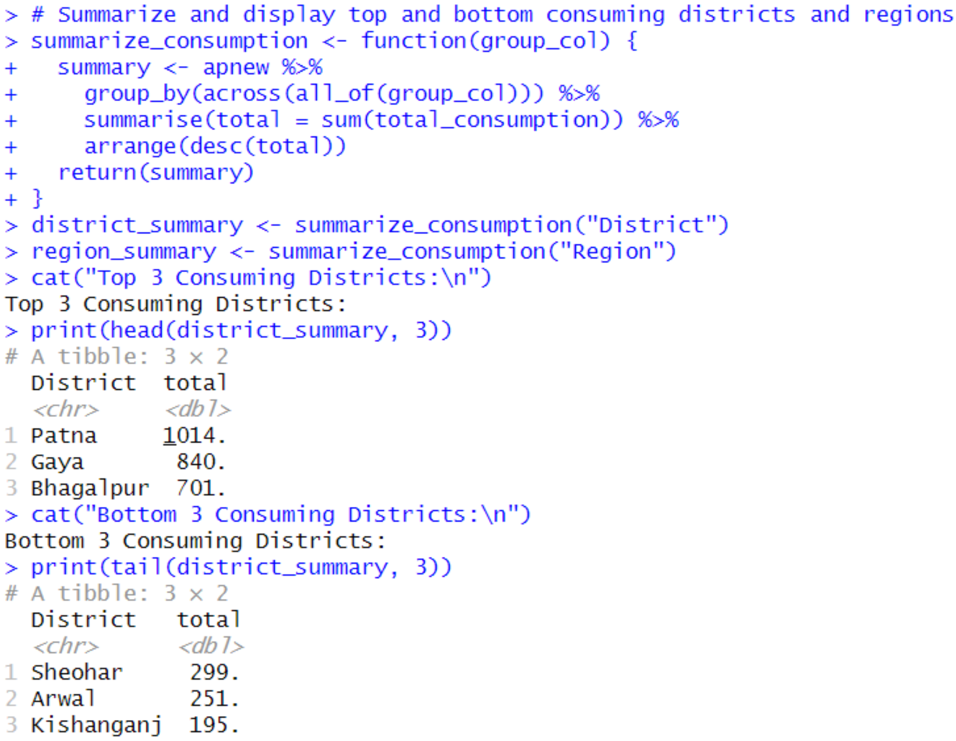
```
> district_mapping
          35           22           28            3           38            8 
      "Gaya"  "Bhagalpur"      "Patna"    "Sheohar"      "Arwal" "Kishanganj" 
```

<u>Interpretation:</u> Here only the districts Gaya, Bhagalpur, Patna, Sheohar, Arwal and Kishanganj are renamed. These are the top three districts in consumption.

D. Summarizing the critical variables in the dataset region-wise and district-wise, indicating the top three and bottom three districts in consumption

Summarizing the top three and bottom three districts in consumption.

Code and Results:



```
> # Summarize and display top and bottom consuming districts and regions
> summarize_consumption <- function(group_col) {
+   summary <- apnew %>%
+     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 3 Consuming Districts:\n")
Top 3 Consuming Districts:
> print(head(district_summary, 3))
# A tibble: 3 × 2
  District  total
  <chr>     <dbl>
1 Patna     1014.
2 Gaya       840.
3 Bhagalpur  701.
> cat("Bottom 3 Consuming Districts:\n")
Bottom 3 Consuming Districts:
> print(tail(district_summary, 3))
# A tibble: 3 × 2
  District    total
  <chr>       <dbl>
1 Sheohar      299.
2 Arwal        251.
3 Kishanganj   195.
```

<u>Interpretation:</u> Patna with 1014 units, Gaya with 840 units and Bhagalpur with 701 units are the top three consuming districts. Similarly, Sheohar with 299 units, Arwal with 251 units and Kishanganj with 195 units are the bottom three consuming units.

E. Testing the significance of difference in the means.

Forming the hypothesis:

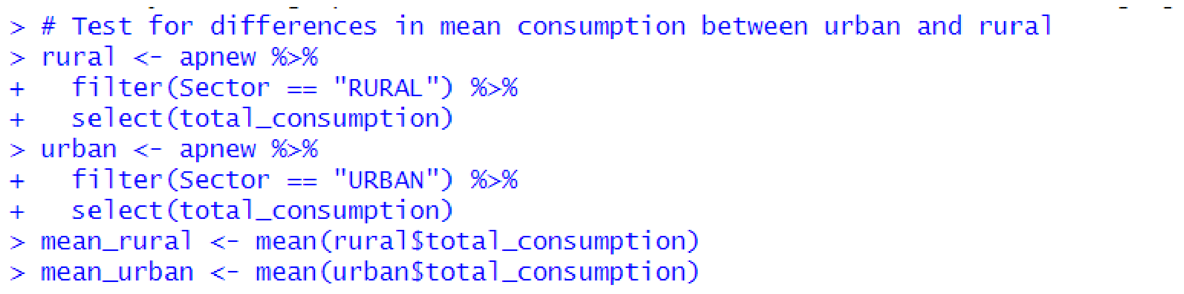
Ho: There is no significant difference in consumption between rural and urban.

Mu-rural = Mu-urban

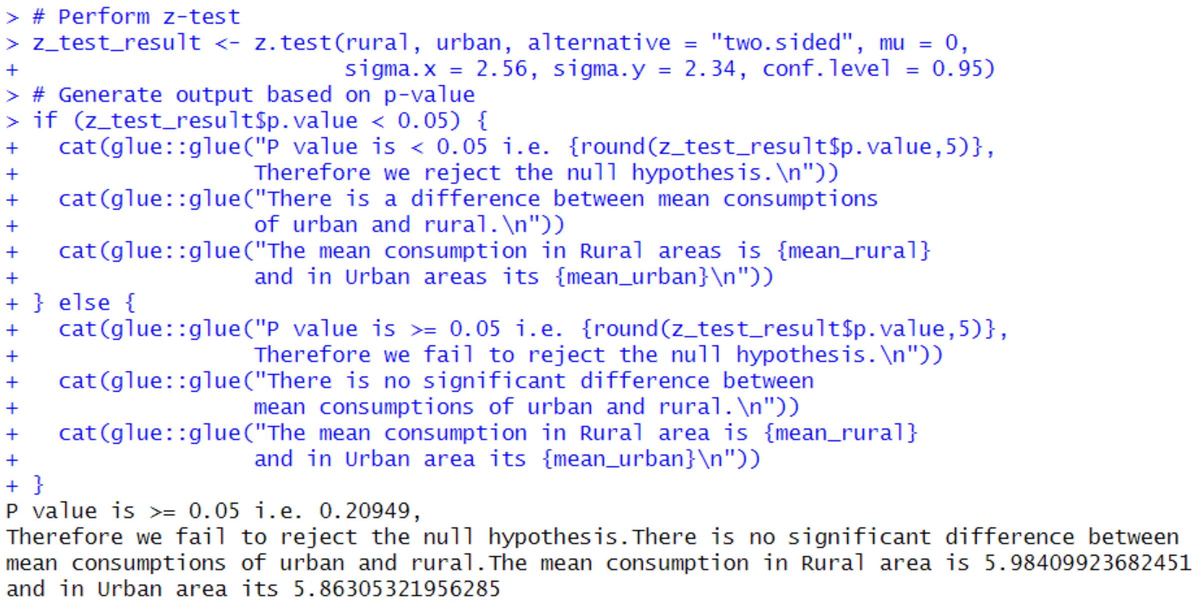
H1: There is significant difference in consumption between rural and urban.

Mu-rural ≠ Mu-urban

Code and Results:



```
> # Test for differences in mean consumption between urban and rural
> rural <- apnew %>%
+   filter(Sector == "RURAL") %>%
+   select(total_consumption)
> urban <- apnew %>%
+   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"))
+ }
P value is >= 0.05 i.e. 0.20949,
Therefore we fail to reject the null hypothesis.There is no significant difference between
mean consumptions of urban and rural.The mean consumption in Rural area is 5.98409923682451
and in Urban area its 5.86305321956285
```

Interpretation: This code snippet performs a two-sided z-test to compare the mean consumption between rural and urban areas. The analysis concludes with a **p-value of 0.20949**, **greater than the commonly used significance level of 0.05**.

We fail to reject the null hypothesis since the p-value is not less than 0.05. In simpler terms, the evidence from the data is not strong enough to claim a statistically significant difference between the mean consumption in rural and urban areas.