VIRGINIA COMMONWEALTH UNIVERSITY

Statistical analysis and modelling (SCMA 632)

A1a: Consumption Pattern of Gujarat using R and Python

Cherussery Jayaprakash Lakshmy

V01110023

Date of Submission: 16-06-2024

CONTENTS

Sl. No.	Title	Page No.
1.	Introduction	1
2.	Objective	2
3.	Business Significance	3
4.	Results	4-6
5.	Interpretations	7
6.	Codes	8
7.	Recommendations	9

Introduction

The focus of this study is on the state of Gujarat, utilizing data from the National Sample Survey Office (NSSO) to identify the top and bottom three consuming districts within the state. To achieve this, we carefully manipulate and clean the dataset to extract the necessary information for analysis. The dataset includes consumption-related data, covering both rural and urban sectors, along with district-wise variations. This dataset has been imported into R, a powerful and versatile statistical programming language well-suited for handling and analyzing large datasets.

The objectives of this study include identifying and addressing missing values, handling outliers, standardizing district and sector names, summarizing consumption data both regionally and district-wise, and testing the significance of differences in mean consumption. The insights derived from this study will provide valuable information for policymakers and stakeholders, facilitating targeted interventions and promoting equitable development across Gujarat.

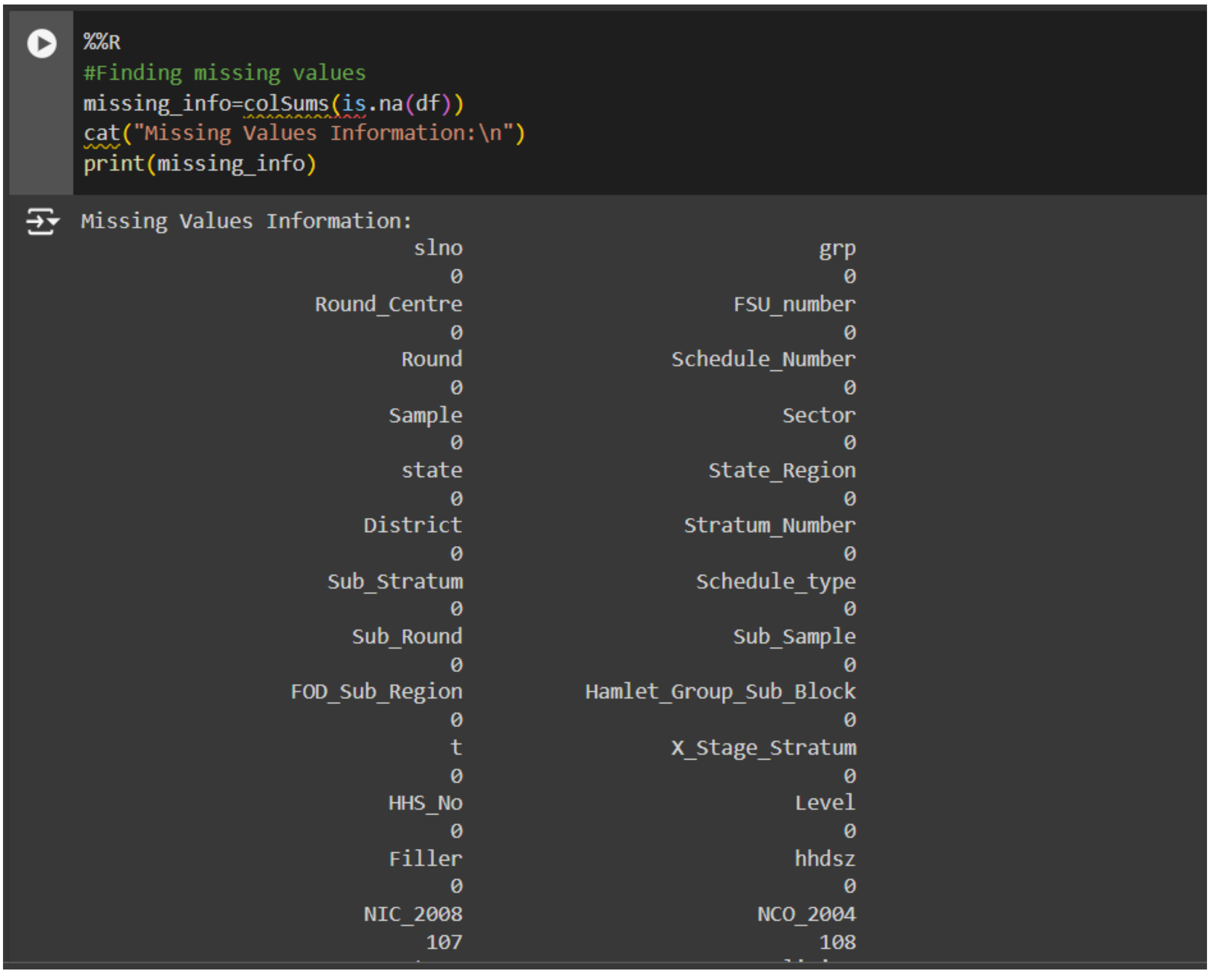
Objective

1. Identify and Address Missing Values
   Check if there are any missing values in the data.
   Identify the missing values.
   Replace missing values with the mean of the respective variable.
2. Detect and Amend Outliers Check for outliers in the dataset.
   Describe the outcome of the outlier detection test.
   Make suitable amendments to handle the identified outliers.
3. Standardize Naming Conventions
   Rename the districts and sectors to maintain consistency, specifying sectors as rural and urban.
4. Summarize Consumption Data
   Summarize critical variables in the dataset both region-wise and district-wise.
   Identify and indicate the top and bottom three districts in terms of consumption.
5. Analyse Mean Differences
   Test whether the differences in mean consumption values are statistically significant.

Business Significance

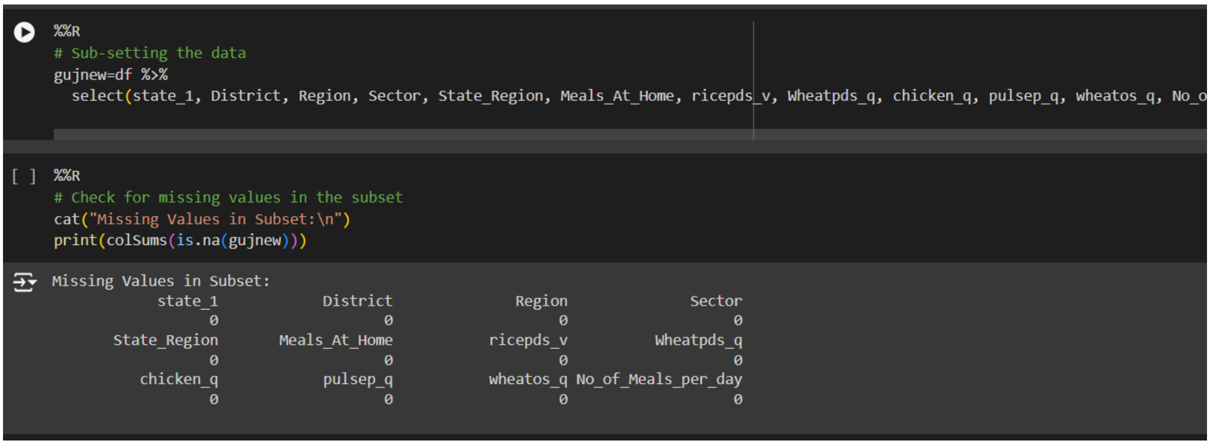
The focus of this study on Gujarat's consumption patterns using NSSO data has significant implications for businesses and policymakers. By identifying the top and bottom three consuming districts, the study provides valuable insights for market entry, resource allocation, supply chain optimization, and targeted interventions. Through data cleaning, outlier detection, and significance testing, the findings facilitate informed decision-making, fostering equitable development and promoting economic growth in Gujarat.

Results and Interpretations



```
%%R
#Finding missing values
missing_info=colSums(is.na(df))
cat("Missing Values Information:\n")
print(missing_info)
```

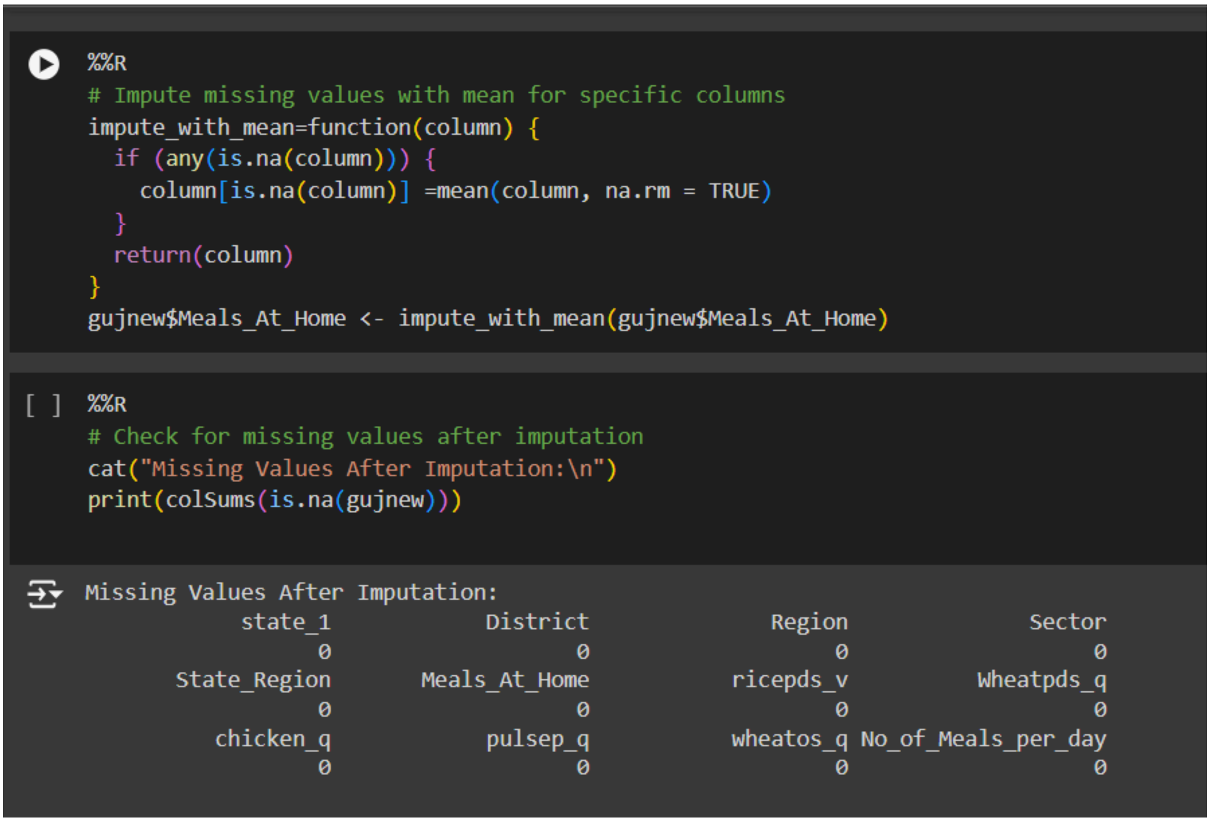
```
Missing Values Information:
                  slno                    grp
                     0                      0
          Round_Centre             FSU_number
                     0                      0
                 Round        Schedule_Number
                     0                      0
                Sample                 Sector
                     0                      0
                 state           State_Region
                     0                      0
              District         Stratum_Number
                     0                      0
           Sub_Stratum          Schedule_type
                     0                      0
             Sub_Round             Sub_Sample
                     0                      0
        FOD_Sub_Region Hamlet_Group_Sub_Block
                     0                      0
                     t        X_Stage_Stratum
                     0                      0
                HHS_No                  Level
                     0                      0
                Filler                  hhdsz
                     0                      0
              NIC_2008               NCO_2004
                   107                    108
```



```
%%R
# Sub-setting the data
gujnew=df %>%
  select(state_1, District, Region, Sector, State_Region, Meals_At_Home, ricepds_v, Wheatpds_q, chicken_q, pulsep_q, wheatos_q, No_o
```

```
[ ]  %%R
# Check for missing values in the subset
cat("Missing Values in Subset:\n")
print(colSums(is.na(gujnew)))
```

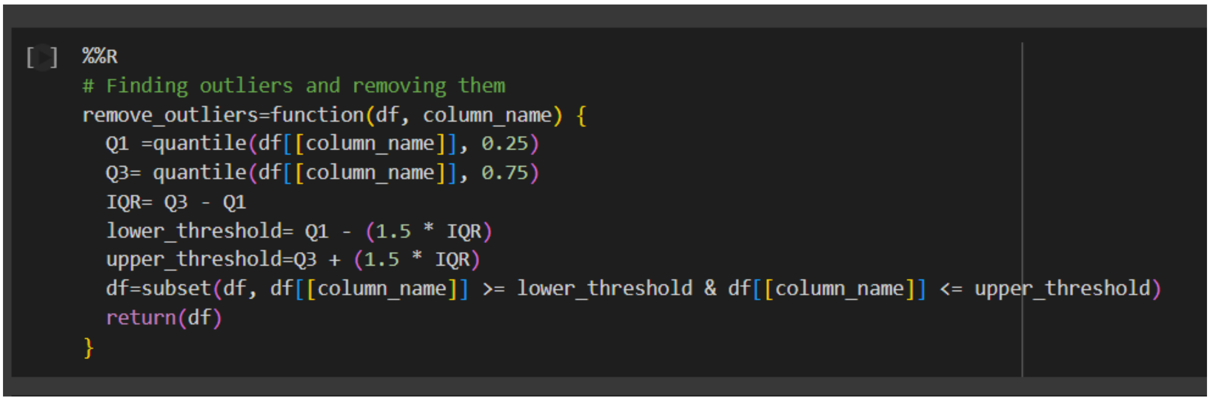
```
Missing Values in Subset:
          state_1          District            Region            Sector
                0                 0                 0                 0
     State_Region     Meals_At_Home         ricepds_v        Wheatpds_q
                0                 0                 0                 0
        chicken_q          pulsep_q         wheatos_q No_of_Meals_per_day
                0                 0                 0                 0
```



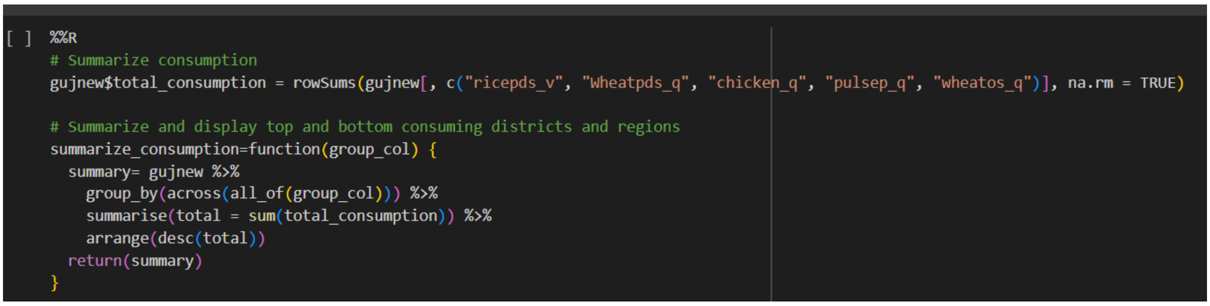
```
%%R
# Impute missing values with mean for specific columns
impute_with_mean=function(column) {
  if (any(is.na(column))) {
    column[is.na(column)] =mean(column, na.rm = TRUE)
  }
  return(column)
}
gujnew$Meals_At_Home <- impute_with_mean(gujnew$Meals_At_Home)
```

```
%%R
# Check for missing values after imputation
cat("Missing Values After Imputation:\n")
print(colSums(is.na(gujnew)))
```

```
Missing Values After Imputation:
          state_1          District            Region            Sector 
                0                 0                 0                 0 
     State_Region     Meals_At_Home        ricepds_v        Wheatpds_q 
                0                 0                 0                 0 
        chicken_q          pulsep_q         wheatos_q No_of_Meals_per_day 
                0                 0                 0                 0 
```

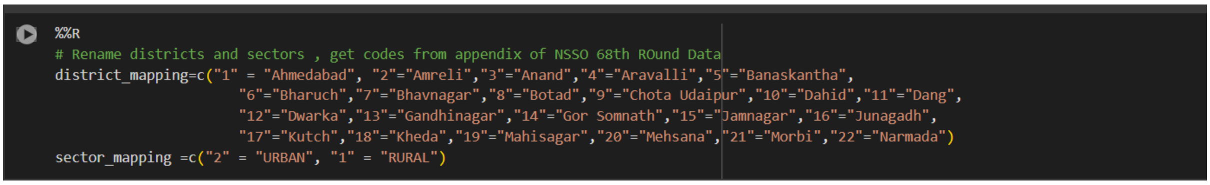


```
%%R
# Finding outliers and removing them
remove_outliers=function(df, column_name) {
  Q1 =quantile(df[[column_name]], 0.25)
  Q3= quantile(df[[column_name]], 0.75)
  IQR= Q3 - Q1
  lower_threshold= Q1 - (1.5 * IQR)
  upper_threshold=Q3 + (1.5 * IQR)
  df=subset(df, df[[column_name]] >= lower_threshold & df[[column_name]] <= upper_threshold)
  return(df)
}
```

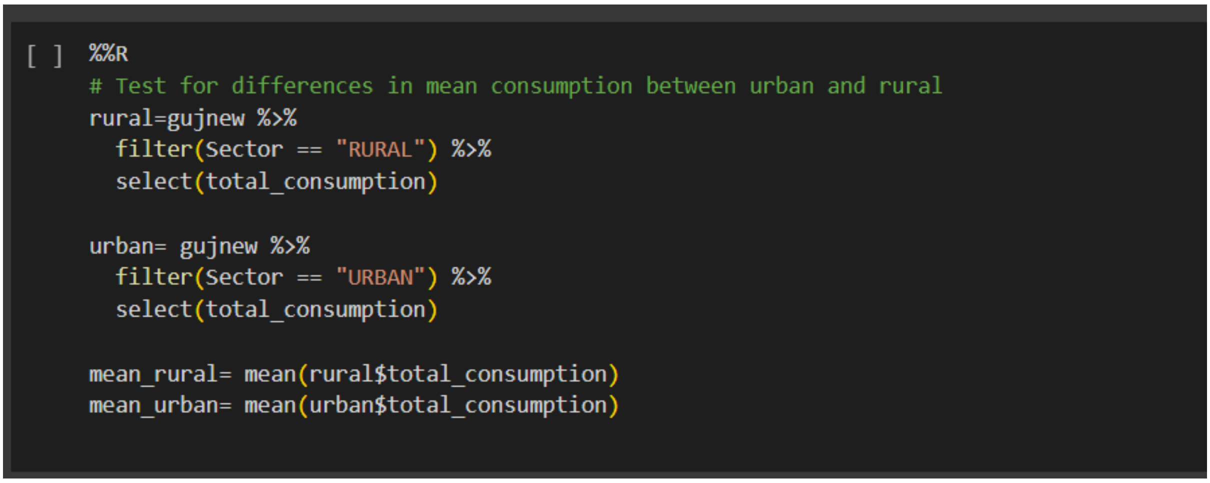


```
%%R
# Summarize consumption
gujnew$total_consumption = rowSums(gujnew[, c("ricepds_v", "Wheatpds_q", "chicken_q", "pulsep_q", "wheatos_q")], na.rm = TRUE)

# Summarize and display top and bottom consuming districts and regions
summarize_consumption=function(group_col) {
  summary= gujnew %>%
    group_by(across(all_of(group_col))) %>%
    summarise(total = sum(total_consumption)) %>%
    arrange(desc(total))
  return(summary)
}
```



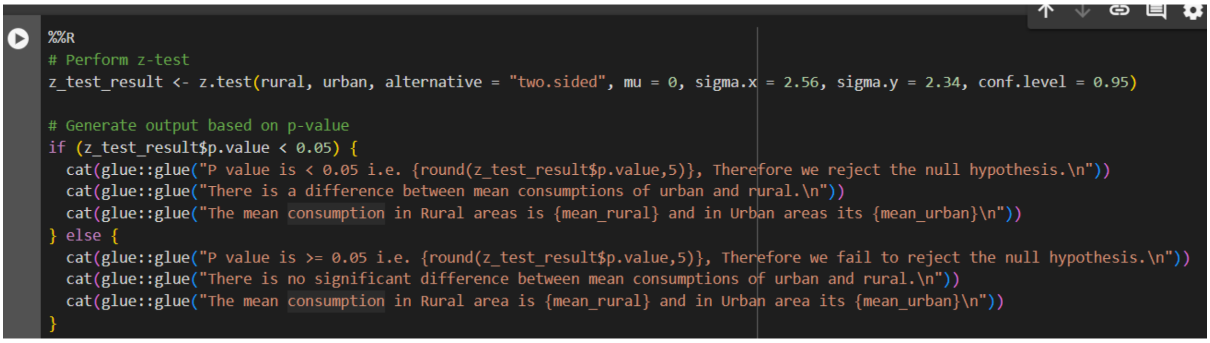
```
%%R
# Rename districts and sectors , get codes from appendix of NSSO 68th ROund Data
district_mapping=c("1" = "Ahmedabad", "2"="Amreli","3"="Anand","4"="Aravalli","5"="Banaskantha",
                     "6"="Bharuch","7"="Bhavnagar","8"="Botad","9"="Chota Udaipur","10"="Dahid","11"="Dang",
                     "12"="Dwarka","13"="Gandhinagar","14"="Gor Somnath","15"="Jamnagar","16"="Junagadh",
                     "17"="Kutch","18"="Kheda","19"="Mahisagar","20"="Mehsana","21"="Morbi","22"="Narmada")
sector_mapping =c("2" = "URBAN", "1" = "RURAL")
```



```
[ ]  %%R
     # Test for differences in mean consumption between urban and rural
     rural=gujnew %>%
       filter(Sector == "RURAL") %>%
       select(total_consumption)

     urban= gujnew %>%
       filter(Sector == "URBAN") %>%
       select(total_consumption)

     mean_rural= mean(rural$total_consumption)
     mean_urban= mean(urban$total_consumption)
```



```
%%R
# 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"))
}
```

Interpretation

- A subset of the dataset was created, including the attributes `state_1`, `District`, `Region`, `Sector`, `State_Region`, `Meals_At_Home`, `ricepds_v`, `Wheatpds_q`, `chicken_q`, `pulsep_q`, `wheatos_q`, and `No_of_Meals_per_day`. During the data cleaning process, we identified one missing value in the `Meals_At_Home` attribute. Missing values in the dataset can lead to incomplete or biased analyses, which can hinder the accuracy of results and potentially skew interpretations and decision-making processes. To ensure data integrity, we replaced the missing value with the mean of the variable.
- Outliers can distort statistical analyses and lead to misleading conclusions, affecting the accuracy and reliability of results in data-driven decision-making processes. To address this, we used the interquartile range (IQR) method to remove outliers. The IQR is calculated as the difference between the upper and lower quartiles. Data points that fall beyond 1.5 times the IQR from either quartile are identified as outliers. These outliers can then be excluded or treated to ensure the robustness of the analysis.

- Each district in the NSSO dataset is assigned a unique number. To identify the top consuming districts, it is essential to map these numbers to their respective names. Similarly, the urban and rural sectors are assigned the numbers 1 and 2, respectively.
- The result shown above demonstrates that the district numbers have been successfully mapped to their corresponding names. Additionally, the sector numbers 1 and 2 have been replaced with the labels "urban" and "rural," respectively.
- We identify the top three and bottom three consuming districts. Bhavnagar has the highest total consumption with 1,466 units, followed by Narmada with 1,113 units, and Chota Udaipur with 781 units. Conversely, Navsari has the lowest total consumption with 48.8 units, followed by Botad with 135 units, and Mehsana with 155 units.
- Here, we observe that the p-value ($p = 2.2e-16$) is less than the level of significance ($\alpha = 0.05$). Therefore, we reject the null hypothesis. This implies a significant difference in total consumption between rural and urban areas. Additionally, we observe that the average total consumption in urban areas is greater than in rural areas.

Codes

r codes a1.pdf

python codes a1.pdf

Recommendations

As a student analyzing data and drawing conclusions based on statistical tests, here are a few recommendations:

1. **Understand Hypothesis Testing**: Ensure a solid grasp of hypothesis testing concepts, including null and alternative hypotheses, p-values, significance levels (α), and interpreting test outcomes. This foundational knowledge will help you correctly interpret statistical results.
2. **Data Preparation and Cleaning**: Prioritize thorough data cleaning to handle missing values, outliers, and ensuring data integrity. This step is crucial as it directly impacts the reliability of your analyses and conclusions.
3. **Visualize Data**: Use visualizations such as boxplots, histograms, and scatter plots to explore data distributions and identify trends or anomalies visually before applying statistical tests. Visual aids can provide insights that numerical summaries might miss.
4. **Choose Appropriate Tests**: Select statistical tests based on the nature of your data (e.g., independent samples, paired samples, categorical variables) and the research question. Ensure the assumptions of the chosen test are met for robust results.