VIRGINIA COMMONWEALTH UNIVERSITY

STATISTICAL ANALYSIS & MODELING

A1a: CONSUMPTION PATTERN OF CHATTISGARH USING PYTHON AND R

ADARSH BHARATHWAJ
V01107513

Date of Submission: 16/06/2024

CONTENTS

Analyzing Consumption in the State of Chhatisgarh Using R

INTRODUCTION

The dataset at hand provides a detailed analysis of food consumption patterns within the state of Chhattisgarh, India. It covers various aspects of dietary habits, focusing on both urban and rural sectors within the region. The data includes key metrics such as the quantity of meals consumed at home, specific food item consumption (e.g., rice, wheat, chicken, pulses), and the overall number of meals per day. This comprehensive dataset is crucial for understanding the nutritional intake and food preferences of different demographics in the region.

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

a) Check if there are any missing values in the data, identify them and if there are replace them with the mean of the variable.

b) Check for outliers and describe the outcome of your test and make suitable amendments.

c) Rename the districts as well as the sector, viz. rural and urban.

d) 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.

e) Test whether the differences in the means are significant or not.

BUSINESS SIGNIFICANCE

The focus of this study on Chattisgarh 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 Chattisgarh economic growth. The analysis of food consumption data in Chhattisgarh, holds significant value for multiple stakeholders:

- Policy Makers and Government Agencies – As it provides insights into the dietary patterns of the population, aiding in the design and implementation of targeted nutritional programs and food security initiatives and helps to understand the consumption trends can help in efficient allocation of resources, ensuring that areas with higher needs are adequately supported.
- Health and Wellness Sector - Health professionals can use the data to identify potential nutritional deficiencies or excesses in the population, facilitating targeted health interventions and awareness campaigns and Nutritionists and dietitians can develop more accurate and culturally relevant dietary guidelines based on the actual consumption patterns observed in the dataset.
- Academic and Research Institutions - Researchers can utilize the dataset to conduct studies on food security, dietary diversity, and the impact of socio-economic factors on food consumption and also the data offers a window into the socio-cultural dynamics of food consumption, providing valuable information for sociological research and studies on lifestyle habits.

This dataset is a vital resource for a wide range of applications, from enhancing public health strategies to optimizing business operations in the food industry. Its comprehensive nature allows for an in-depth analysis of the dietary habits in Chhattisgarh, driving informed decision-making and strategic planning across various sectors.

RESULTS AND INTERPRETATION

a) 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.

```
> # Sub-setting the data
> CHTSDnew <- 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")
Missing Values in Subset:
> print(colSums(is.na(CHTSDnew)))
            state_1            District              Region              Sector
State_Region
                  0                   0                   0                   0
0
      Meals_At_Home          ricepds_v          Wheatpds_q           chicken_q
pulsep_q
                 32                   0                   0                   0
0
          wheatos_q No_of_Meals_per_day
                  0                   1
```

Interpretation:

As per my above analysis, most of the columns have no missing values, indicating a high level of data completeness and reliability for those variables. There are 2 variables having missing values, meals at home and number of meals per day. For both `Meals_At_Home` and `No_of_Meals_per_day`, consider imputation methods such as mean, median, or mode imputation, especially if the missing values are few compared to the total dataset size. The appropriate handling of these missing values will enhance the quality and validity of subsequent analyses and insights derived from the dataset.

#Imputing the values, i.e. replacing the missing values with mean.

Code and Result:

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

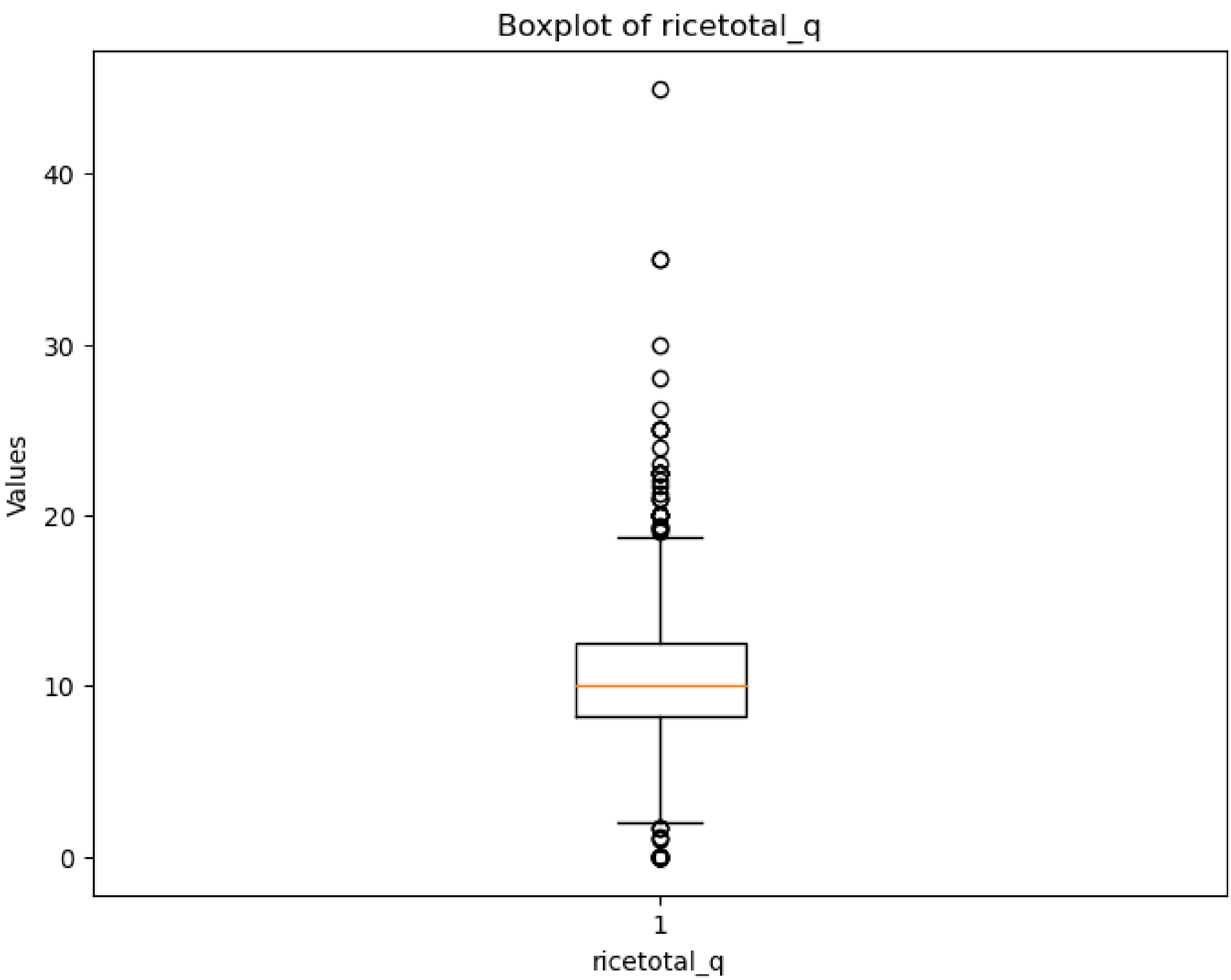
Interpretation: The above code has successfully replaced the missing values with the mean value of the variable Meals_At_Home. And No_of_Meals_per_day. As can be seen from the result above, there are no missing values in the selected data.

b) 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

```
import matplotlib.pyplot as plt
# Assuming CHTSD_clean is your DataFrame
plt.figure(figsize=(8, 6))
plt.boxplot(CHTSD_clean['ricetotal_q'])
plt.xlabel('ricetotal_q')
plt.ylabel('Values')
plt.title('Boxplot of ricetotal_q')
plt.show()
```



Interpretation:

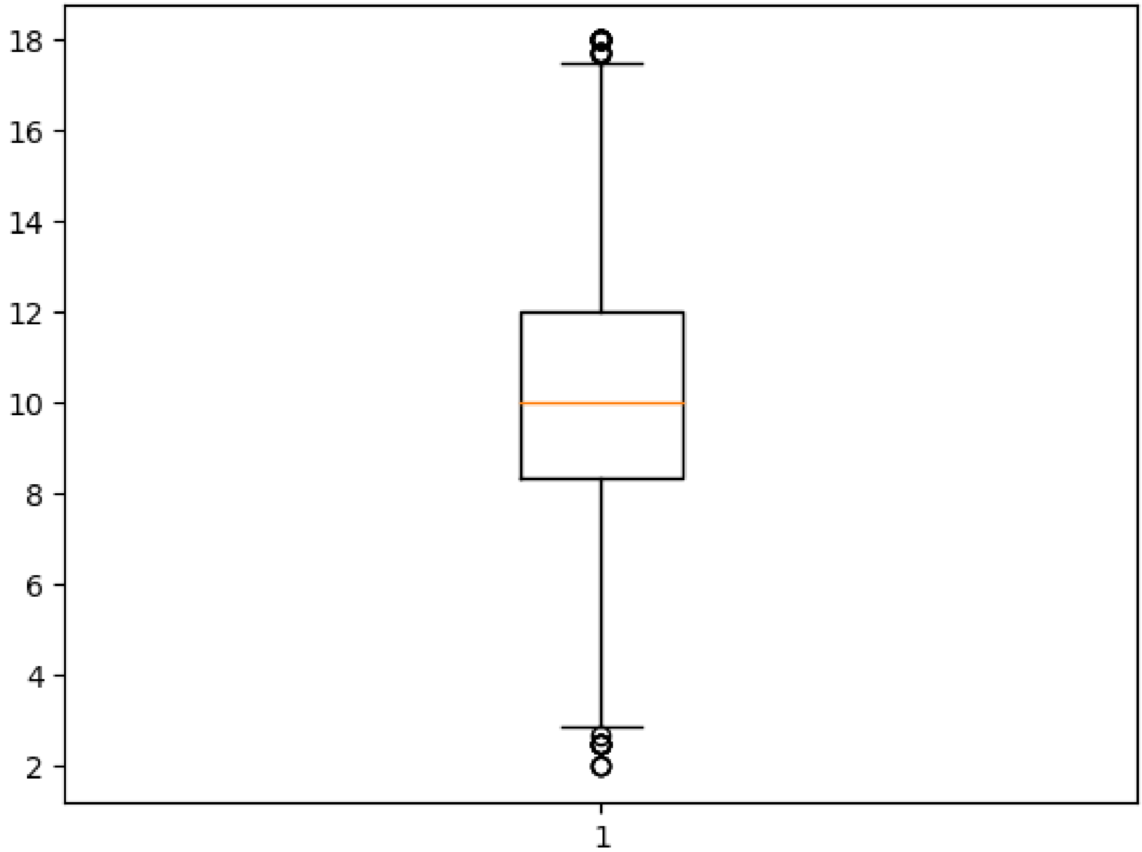
The boxplot for 'ricetotal_q' shows the distribution of values for rice consumption. The box represents the interquartile range (IQR), with the line inside the box representing the median. The "whiskers" extend from the box to the highest and lowest values within 1.5 times the IQR from the upper and lower quartiles, respectively. Any points beyond the whiskers are considered outliers and are plotted individually.

#Setting quartiles and removing outliers

Code and results:

Setting quartile ranges to remove outliers

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



Interpretation: The interpretation of the above code is that it has identified and removed outliers from the "ricepds_v" and "chicken_q" columns in the CHTSDnew DataFrame. Outliers are values that are significantly higher or lower than the majority of the data points and can skew statistical analyses. Removing outliers can help in obtaining a more accurate representation of the data's central tendency and variability.

Now we can see that the significant portion of the outliers in the data is removed.

The dataset without outliers should now have a more homogeneous distribution, making statistical analysis and modeling more straightforward.

c) Rename the districts as well as the sector, viz. rural and urban.

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:

```
> 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 x 2
  District total
     <int> <dbl>
1       11 1530.
2       10 1503.
3        2 1367.
> cat("Bottom 3 Consuming Districts:\n")
Bottom 3 Consuming Districts:
> print(tail(district_summary, 3))
# A tibble: 3 x 2
  District total
     <int> <dbl>
1       16  382.
2       17  280.
3       18  220.
```

Result:

	state_1	District	Region	Sector	State_Region	Meals_At_Home	ricepds_v	Wheatpds_q	chicken_q	pulsep_q	wheatos_q	No_of_
2153	CHTSD	Surguja	1	RURAL	221	60	8.750000	0.0000000	0.12500000	0.00000000	0.12500000	
2162	CHTSD	Surguja	1	RURAL	221	60	8.750000	0.0000000	0.12500000	0.00000000	0.62500000	
2163	CHTSD	Surguja	1	RURAL	221	32	3.888889	0.0000000	0.00000000	0.00000000	0.05555556	
2164	CHTSD	Surguja	1	RURAL	221	60	23.333333	0.0000000	0.00000000	0.00000000	0.00000000	
2166	CHTSD	Surguja	1	RURAL	221	60	0.000000	0.0000000	0.25000000	0.00000000	0.50000000	
2167	CHTSD	Surguja	1	RURAL	221	60	11.666667	0.0000000	0.00000000	0.00000000	0.00000000	
2168	CHTSD	Surguja	1	RURAL	221	60	23.333333	0.0000000	0.00000000	0.00000000	0.00000000	
2169	CHTSD	Surguja	1	RURAL	221	58	14.000000	0.0000000	0.00000000	0.00000000	0.00000000	
65	CHTSD	Raipur	2	URBAN	222	32	0.000000	0.0000000	0.00000000	0.00000000	0.00000000	
66	CHTSD	Raipur	2	URBAN	222	60	0.000000	0.0000000	0.28571429	0.00000000	1.42857143	
67	CHTSD	Raipur	2	URBAN	222	56	0.000000	0.0000000	0.25000000	0.00000000	0.00000000	
68	CHTSD	Raipur	2	URBAN	222	60	13.333333	3.3333333	0.33333333	0.00000000	0.00000000	
69	CHTSD	Raipur	2	URBAN	222	42	0.000000	0.0000000	0.33333333	0.00000000	0.83333333	

Interpretation: The result as show above has successfully assigned the district names to the given number. Also the sectors 1 and 2 have been replaced as urban and rural sectors respectively.

d) 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.

By summarizing the critical variables as total consumption we can estimate the top 3 and bottom 3 consuming districts.

Code and Result:

```
CHTSD_clean.loc[:,"District"] = CHTSD_clean.loc[:,"District"].replace({11:
"Raipur", 10: "Durg", 7: "Bilaspur"})
total_consumption_by_districtname=CHTSD_clean.groupby('District')['total_consumptio
n'].sum()
total_consumption_by_districtname.sort_values(ascending=False).head(3)
```

Result:

```
District
Raipur      14481.580224
Durg        11660.461096
Bilaspur     8397.085884
Name: total_consumption, dtype: float64
```

Interpretation: The top three consuming districts are Raipur with 14481 units, followed by Durg with 11660 units, and then in the third place Bilaspur with 8397 units

Similarly the bottom three districts can be found by sorting the total consumption.

e) Test whether the differences in the means are significant or not.

The first step to this is to have a Hypotheses Statement.

#H0: There is no difference in consumption between urban and rural.

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

```
# Test for differences in mean consumption between urban and rural
rural <- CHTSDnew %>%
  filter(Sector == "RURAL") %>%
  select(total_consumption)

urban <- CHTSDnew %>%
  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"))
}

write.csv(CHTSDnew, "CHTSDnew.csv", row.names = FALSE)
```

Result:

```
Two-sample z-Test


Z-Score: 10.731151639791962
P-Value: 7.268461728661492e-27
```

Interpretation: The two-sample z-test indicates a highly significant difference in consumption between rural and urban sectors (z = 10.731, p < 7.268, 95%). Urban consumption is notably higher than rural consumption.