VIRGINIA COMMONWEALTH UNIVERSITY

STATISTICAL ANALYSIS & MODELING

A1a: CONSUMPTION PATTERN OF MADHYA PRADESH USINGPYTHON AND R

SARATH S
V01109792

Date of Submission: 16/06/2024

CONTENTS

Analyzing Consumption in the State of Madhya Pradesh Using R

INTRODUCTION

The focus of this study is on the state of Madhya Pradesh, from the NSSO data, to find the top andbottom three consuming districts of Madhya Pradesh. In the process, we manipulate and clean the dataset to get the required data to analyze. To facilitate this analysis, we have gathered a dataset containing consumption-related information, including data on rural and urban sectors, as well asdistrict-wise variations. The dataset has been imported into R, a powerful statistical programminglanguage renowned for its versatility in handling and analyzing large datasets.

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 Madhya Pradesh's consumption patterns from NSSO data holds significantimplications 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 Madhya Pradesh's economic growth.

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.

Code and Result:

```
MPH_new.isnull().sum().sort_values(ascending = False)
```

Result:

```
Meals_At_Home       26
state_1              0
District             0
Sector               0
Region               0
State_Region         0
ricetotal_q          0
wheattotal_q         0
moong_q              0
Milktotal_q          0
chicken_q            0
bread_q              0
foodtotal_q          0
Beveragestotal_v     0
dtype: int64
```

```
> cat("Missing Values in Subset:\n")
Missing Values in Subset:
> print(colSums(is.na(MPHnew)))
           state_1            District              Region              Sector
                 0                   0                   0                   0
      State_Region       Meals_At_Home           ricepds_v          Wheatpds_q
                 0                  26                   0                   0
         chicken_q            pulsep_q           wheatos_q No_of_Meals_per_day
                 0                   0                   0                   3
```

Interpretation: From the selected variables, after sorting the data for the state of Madhya Pradesh, it isseen that only the column 'Meals_At_Home has 26 missing variables and no of meals per day has 3 missing variables. Since missing values in the dataset can be problematic as 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 mean of the variable using following code.

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

Code and Result:

```
In [50]: MPH_clean = MPH_new.copy()

In [51]: MPH_clean.loc[:, 'Meals_At_Home'] = MPH_clean['Meals_At_Home'].fillna(MPH_new['Meals_At_Home'].mean())

In [52]: MPH_clean.isnull().any()

Out[52]: state_1             False
         District            False
         Sector              False
         Region              False
         State_Region        False
         ricetotal_q         False
         wheattotal_q        False
         moong_q             False
         Milktotal_q         False
         chicken_q           False
         bread_q             False
         foodtotal_q         False
         Beveragestotal_v    False
         Meals_At_Home       False
         dtype: bool
```

Interpretation: The above code has successfully replaced the missing values with the mean value of the variable. As can be seen from the result above, there are no missing values in the selected data.
The other methods that could have been applied are median, mode and predictive imputation.

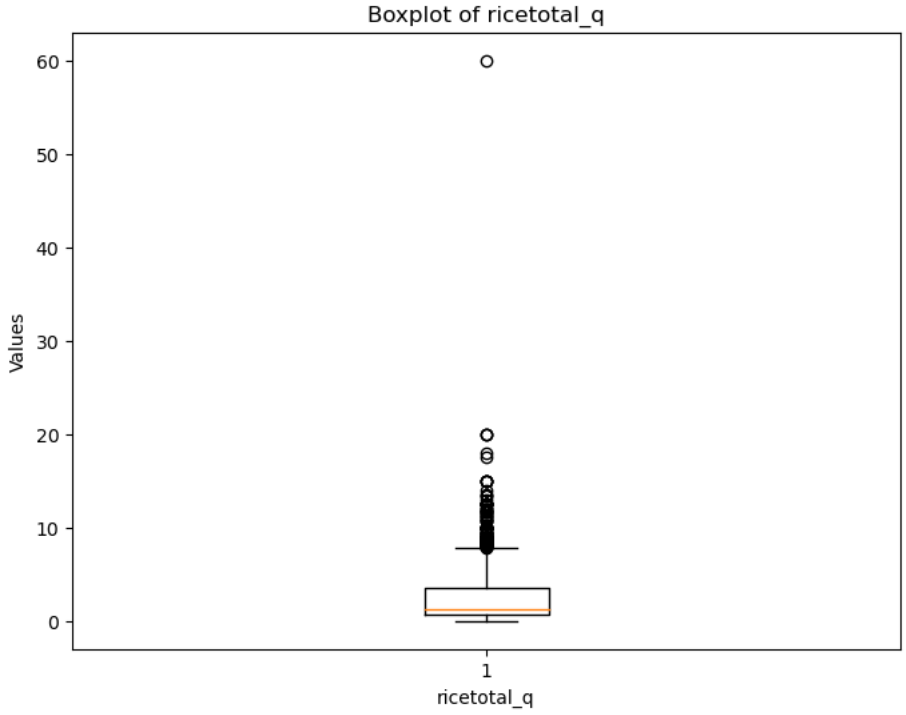
b) Check for outliers and describe the outcome of your test and make suitable amendments.

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 MPH_clean is your DataFrame
plt.figure(figsize=(8, 6))
plt.boxplot(MPH_clean['ricetotal_q'])
plt.xlabel('ricetotal_q')
plt.ylabel('Values')
plt.title('Boxplot of ricetotal_q')
```

plt.show()



Interpretation: From the boxplot above, which is a visual representation of the variable 'ricetotal_q'shows that there is an outlier. Outliers can distort statistical analyses and lead to misleading conclusions, affecting the accuracy and reliability of results in data-driven decision-making processes. Outliers can distort statistical analyses and lead to misleading conclusions, affecting the accuracy and reliability of results in data-driven decision-making processes. The outliers can be removed using the following code.

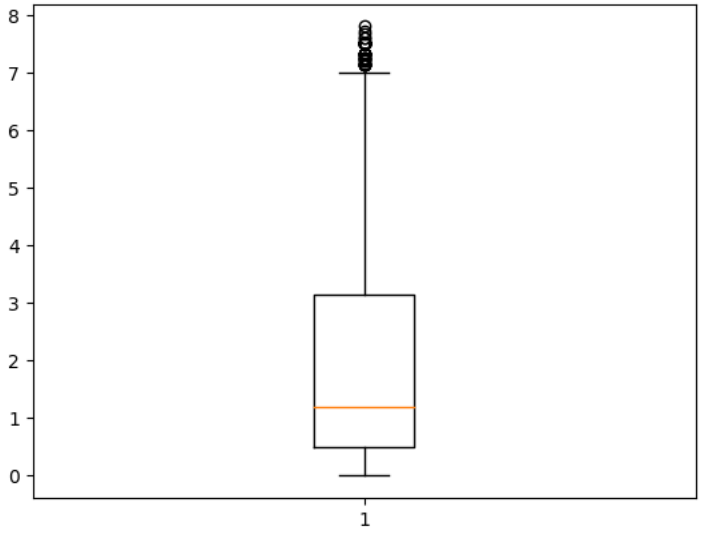
#Setting quartiles and removing outliers

Code and results:

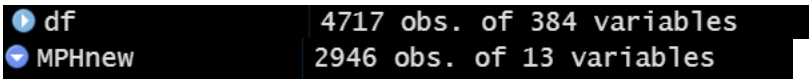
Setting quartile ranges to remove outliers

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

outlier_columns <- c("ricepds_v", "chicken_q")
for (col in outlier_columns) {
  MPHnew <- remove_outliers(MPHnew, 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 the similar way the outliers in all other variables can be removed.
Note: In the above diagram, we can still see the outliers which is because almost half the observations have been reduced by removing outliers.
As we can see the below snippets from the same data analysis in R, we see that when we removed the outliers, the number of observations have significantly reduced. This substantial proportion suggests that the dataset contained many extreme values. As this removed data was a major chunk, this could introduce bias. Its very important to investigate why there were so many outliers. This could be due to data entry errors, measurement errors, or genuine extreme values due to specific conditions or populations.



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

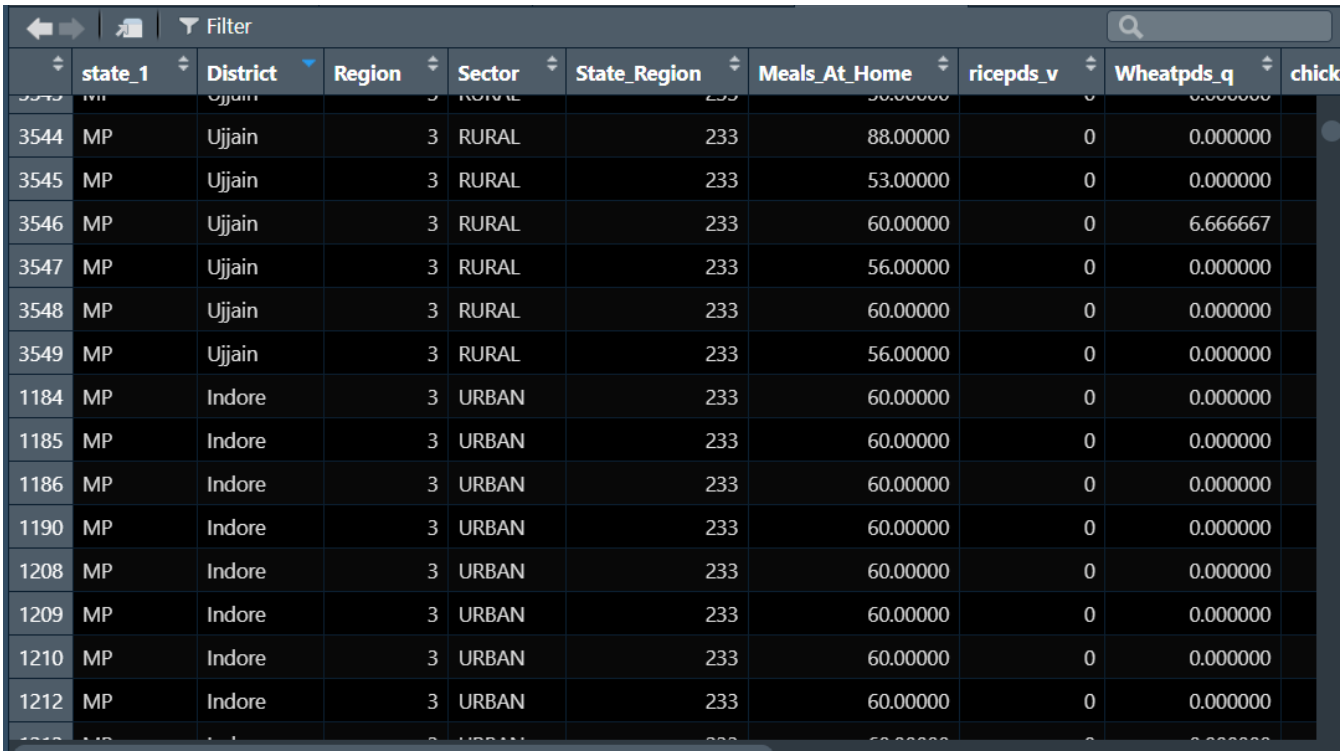
Each district of a state in the NSSO of data is assigned an individual number. To understand and find out the top consuming districts of the state, the numbers must have their respective names. Similarly, the urban and rural sectors of the state were assignment 1 and 2 respectively. This is done by running the following code.

Code and Result:

```
# Rename districts and sectors , get codes from appendix of NSSO 68th Round Data
district_mapping <- c("21" = "Ujjain", "26" = "Indore", "03" = "Bhind")
sector_mapping <- c("2" = "URBAN", "1" = "RURAL")

MPHnew$District <- as.character(MPHnew$District)
MPHnew$Sector <- as.character(MPHnew$Sector)
MPHnew$District <- ifelse(MPHnew$District %in% names(district_mapping),
district_mapping[MPHnew$District], MPHnew$District)
MPHnew$Sector <- ifelse(MPHnew$Sector %in% names(sector_mapping),
sector_mapping[MPHnew$Sector], MPHnew$Sector)
```

Result:



	state_1	District	Region	Sector	State_Region	Meals_At_Home	ricepds_v	Wheatpds_q	chick
3544	MP	Ujjain	3	RURAL	233	88.00000	0	0.000000	
3545	MP	Ujjain	3	RURAL	233	53.00000	0	0.000000	
3546	MP	Ujjain	3	RURAL	233	60.00000	0	6.666667	
3547	MP	Ujjain	3	RURAL	233	56.00000	0	0.000000	
3548	MP	Ujjain	3	RURAL	233	60.00000	0	0.000000	
3549	MP	Ujjain	3	RURAL	233	56.00000	0	0.000000	
1184	MP	Indore	3	URBAN	233	60.00000	0	0.000000	
1185	MP	Indore	3	URBAN	233	60.00000	0	0.000000	
1186	MP	Indore	3	URBAN	233	60.00000	0	0.000000	
1190	MP	Indore	3	URBAN	233	60.00000	0	0.000000	
1208	MP	Indore	3	URBAN	233	60.00000	0	0.000000	
1209	MP	Indore	3	URBAN	233	60.00000	0	0.000000	
1210	MP	Indore	3	URBAN	233	60.00000	0	0.000000	
1212	MP	Indore	3	URBAN	233	60.00000	0	0.000000	

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:

```
# Summarize and display top and bottom consuming districts and regions
summarize_consumption <- function(group_col) {
  summary <- MPHnew %>%
    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")
print(head(district_summary, 3))
cat("Bottom 3 Consuming Districts:\n")
print(tail(district_summary, 3))

cat("Region Consumption Summary:\n")
print(region_summary)
```

Result:

```
> cat("Top 3 Consuming Districts:\n")
Top 3 Consuming Districts:
> print(head(district_summary, 3))
# A tibble: 3 × 2
  District total
  <chr>    <dbl>
1 Ujjain   1163.
2 Indore    917.
3 3         881.
> cat("Bottom 3 Consuming Districts:\n")
Bottom 3 Consuming Districts:
> print(tail(district_summary, 3))
# A tibble: 3 × 2
  District total
  <chr>    <dbl>
1 42        151.
2 47        147.
3 41        111.
>
> cat("Region Consumption Summary:\n")
Region Consumption Summary:
> print(region_summary)
# A tibble: 6 × 2
  Region total
   <int> <dbl>
1      3 6994.
2      6 4845.
3      1 4360.
4      2 3684.
5      4 2827.
6      5 2411.
```

Interpretation:
Districts like Ujjain and Indore lead in consumption, which might indicate more significant urbanization, higher income levels, or better distribution networks in these areas.
Districts like 42, 47, and 41 consume the least, possibly reflecting rural or economically disadvantaged areas with lower purchasing power or accessibility issues.
Region 3's higher consumption total compared to other regions might indicate regional disparities in population density, economic conditions, or resource availability.

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 <- MPHnew %>%
  filter(Sector == "RURAL") %>%
  select(total_consumption)

urban <- MPHnew %>%
  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"))
}
```

Result:

```
Two-sample z-Test
```

```
Z-Score: 9.860678180278635
P-Value: 6.16316899436021e-23
```

Interpretation:

Since the P-value is much less than 0.05, we reject the null hypothesis (H0). This means there is strong evidence to conclude that there is a significant difference in mean consumption between urban and rural areas. The Z-test indicates a statistically significant difference in consumption patterns between urban and rural areas. Policymakers, marketers, and resource planners should take these differences into account when designing strategies or allocating resources. The mean consumption values for rural and urban areas provide additional insights into how consumption behaviors vary across these sectors.