VIRGINIA COMMONWEALTH UNIVERSITY

Statistical analysis and modelling (SCMA 632)

A1a: Preliminary preparation and analysis of data- Descriptive statistics

KATHRYN SHAJU

V01111327

Date of Submission: 16-06-2024

CONTENTS

INTRODUCTION

Consumption is an important indicator of the economic wellbeing of the country or state. In this study we aim to understand the top 3 most consuming as well as bottom 3 consuming districts in the state of Punjab. For this we have acquired our relevant data mainly related to consumption from the NSSO dataset. We have then used the R programming language that will help us to not only analyse the data but also help us to clean and manipulate the data to achieve the desired data on the state of Punjab specifically.

Our main goals include identifying missing values and replacing them, similarly to recognize the outliers in the dataset and eliminate them, renaming and summarizing the top 3 as well as bottom 3 consuming districts as well as to conduct test for significance of the means.

The research will allow policymakers and researchers to better understand the economic wellbeing of a state like Punjab, through its consumption patterns to better formulate future policies and schemes for the state.

OBJECTIVES

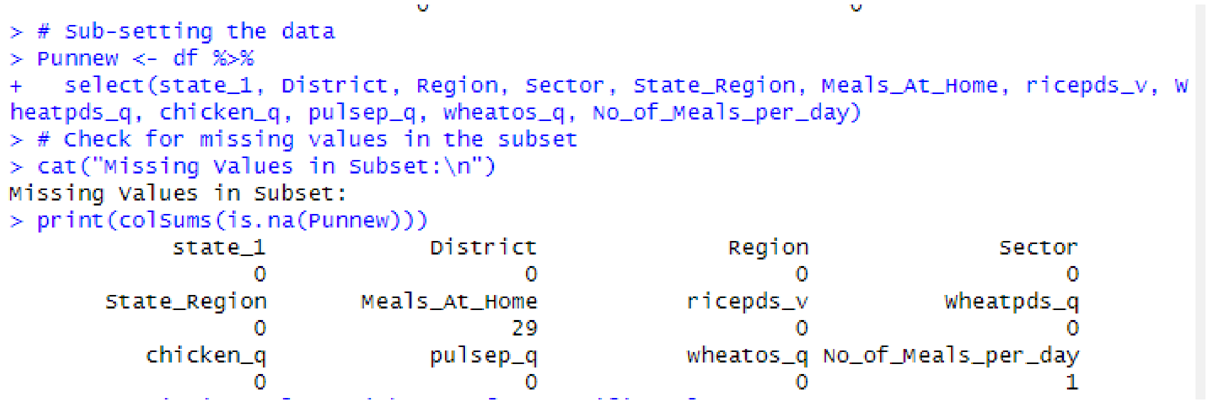
1. To check for missing values, identifying them and replacing them with the mean of the variable.
2. To check for outliers and eliminate them.
3. Summarizing critical variables in the data set district wise and identifying top 3 and bottom 3 districts for consumption.
4. Renaming the districts as well as the sector viz. rural and urban.
5. Test whether the differences between the mean are significant or not

BUSINESS SIGNIFICANCE

Punjab is one of the top states in India for production of mainly agricultural products. This makes Punjab's economic affairs of great interest to not only academicians but also business owners as well. The consumption patterns of Punjab, can firstly help researchers and academicians to understand the lifestyle, and sociological as well as cultural layout of the land. Most importantly the consumption pattern of Punjab is significant for market researchers and business owners in understanding the market trends in Punjab. Evaluating the top 3 districts as well as the bottom 3 districts allows business owners in mapping out the best areas in the markets for not only resource allocation but also efficient distribution, supply chain growth. Through data cleaning and testing the information provided only becomes more accurate and trustworthy, leading to competent decision making, nurturing growth in the state of Punjab.

INTERPRETATIONS

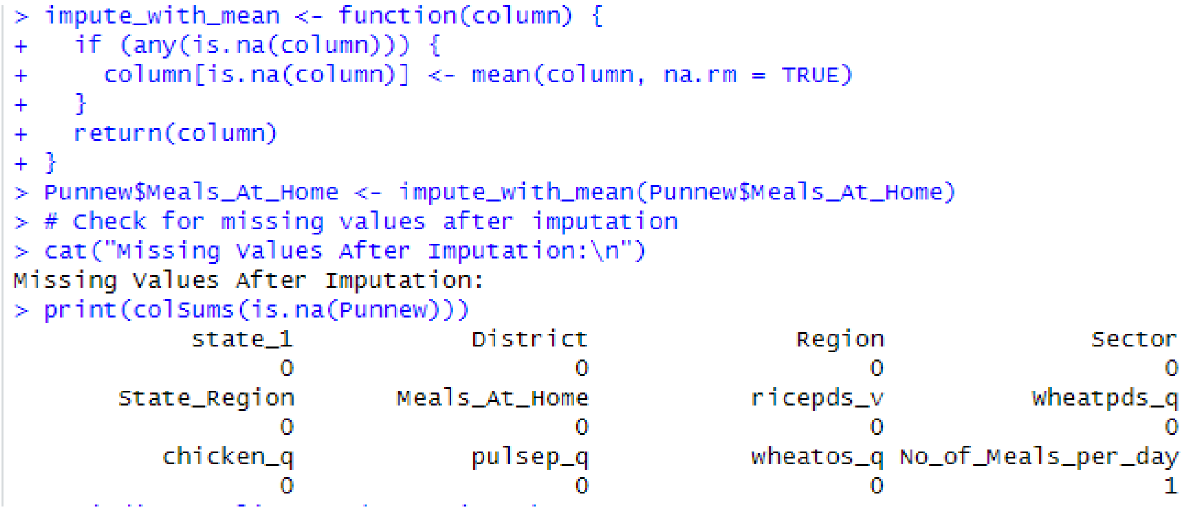
1. ***To check for missing values, identifying them and replacing them with the mean of the variable.***



```
> # Sub-setting the data
> Punnew <- df %>%
+   select(state_1, District, Region, Sector, State_Region, Meals_At_Home, ricepds_v, W
heatpds_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(Punnew)))
            state_1            District              Region              Sector 
                  0                   0                   0                   0 
       State_Region       Meals_At_Home           ricepds_v          Wheatpds_q 
                  0                  29                   0                   0 
          chicken_q            pulsep_q           wheatos_q No_of_Meals_per_day 
                  0                   0                   0                   1 
```

Interpretation: to check for missing value we first have to subset the data into the variables of our choice. For the purpose of this study we have chosen the following variables; 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.

Once we have created the subset of variables, we now run the code that will identify the missing values present within the subset. Here for our study purposes we have identified the variable Meals_At_Home which showcased 29 missing values. This causes the data to be inaccurate and can cause faulty results, hence we replace the missing value with the mean of the variables.

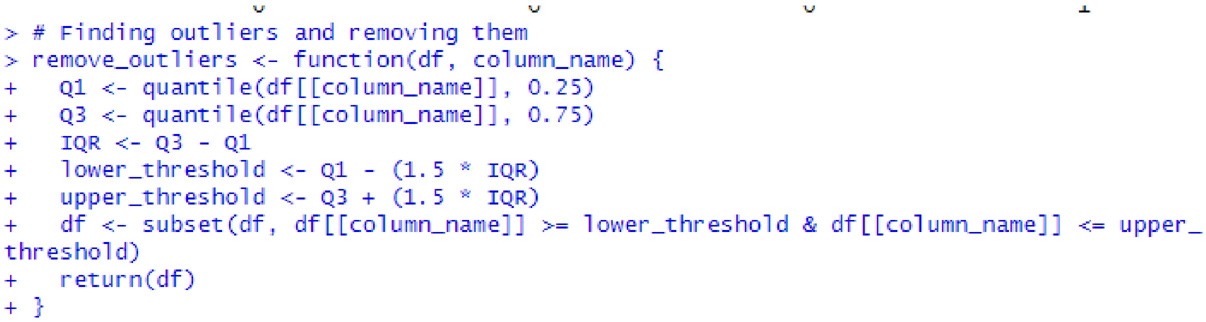


```
> impute_with_mean <- function(column) {
+   if (any(is.na(column))) {
+     column[is.na(column)] <- mean(column, na.rm = TRUE)
+   }
+   return(column)
+ }
> Punnew$Meals_At_Home <- impute_with_mean(Punnew$Meals_At_Home)
> # Check for missing values after imputation
> cat("Missing Values After Imputation:\n")
Missing Values After Imputation:
> print(colSums(is.na(Punnew)))
            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                   1 
```

To achieve this we again run the code, for replacing the missing value with the mean of the variable, which is the imputed mean and now when we check again we find them. As seen above we have successfully managed to make the data smooth and accurate again, allowing for a better result.

2. *To check for outliers and eliminate them.*

To check this we use the following,

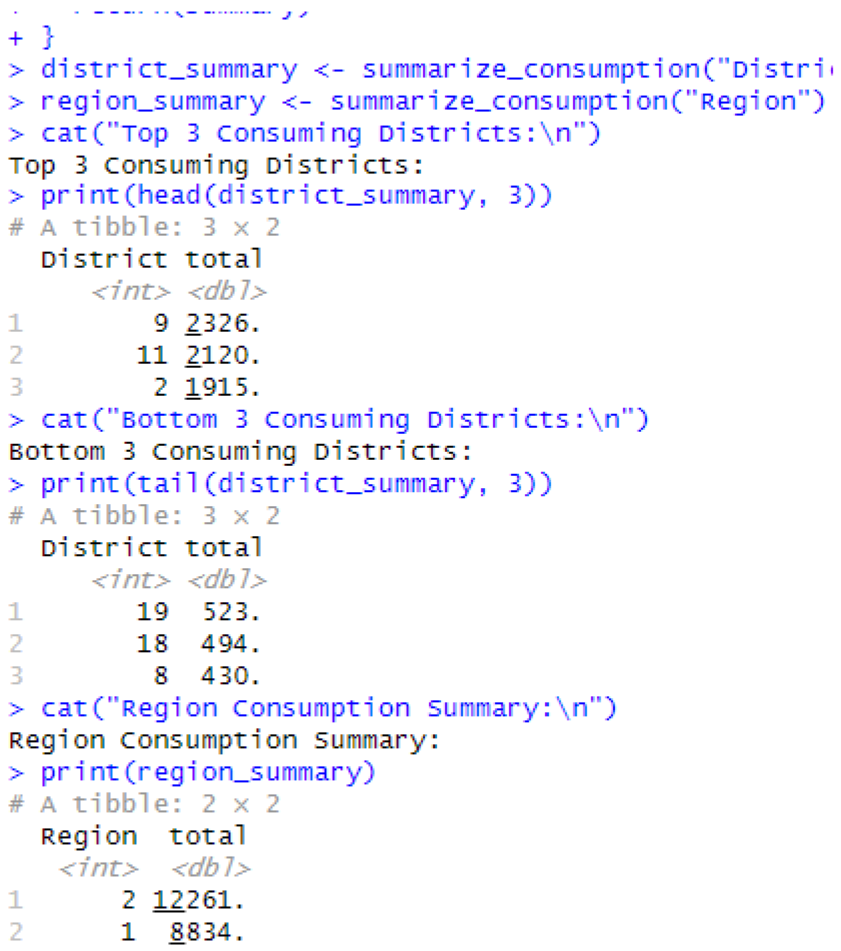


```
> # Finding outliers and removing them
> remove_outliers <- function(df, column_name) {
+   Q1 <- quantile(df[[column_name]], 0.25)
+   Q3 <- quantile(df[[column_name]], 0.75)
+   IQR <- Q3 - Q1
+   lower_threshold <- Q1 - (1.5 * IQR)
+   upper_threshold <- Q3 + (1.5 * IQR)
+   df <- subset(df, df[[column_name]]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

By interpreting quartile range we are able to detect for outliers as well as able to remove them. Calculating the range as difference between upper quartiles and lower quartiles, points beyond 1.5 times of the IQR, outliers are not only detected but also can be excluded. For this study, we checked the outliers for the variables, ricepds_v and chicken_q, using code and it results in the answer, ricepds_v has 1 outlier and chicken_q has 2.

Outliers in the dataset lead to the problem of Hetroskedasticity, which can lead to inaccurate results, because the nature of the data would no longer be normally distributed but rather skewed in nature.

3. ***Summarizing critical variables in the data set district wise and identifying top 3 and bottom 3 districts for consumption*.**



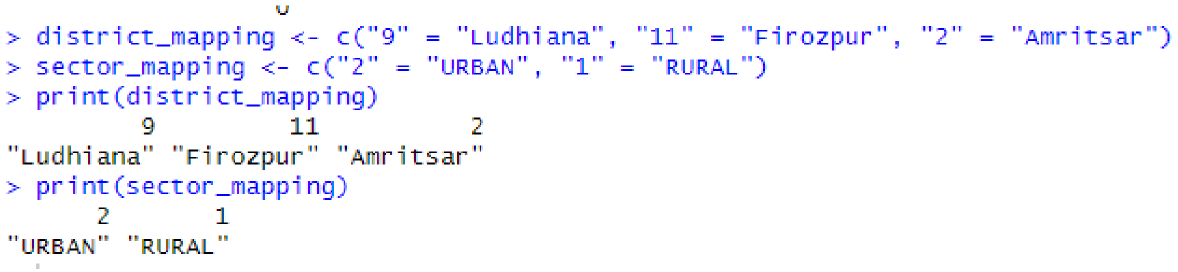
```
+ }
> district_summary <- summarize_consumption("Distri
> region_summary <- summarize_consumption("Region")
> cat("Top 3 Consuming Districts:\n")
Top 3 Consuming Districts:
> print(head(district_summary, 3))
# A tibble: 3 × 2
  District total
     <int> <dbl>
1        9 2326.
2       11 2120.
3        2 1915.
> cat("Bottom 3 Consuming Districts:\n")
Bottom 3 Consuming Districts:
> print(tail(district_summary, 3))
# A tibble: 3 × 2
  District total
     <int> <dbl>
1       19  523.
2       18  494.
3        8  430.
> cat("Region Consumption Summary:\n")
Region Consumption Summary:
> print(region_summary)
# A tibble: 2 × 2
  Region  total
   <int>  <dbl>
1      2 12261.
2      1  8834.
```

From the above code and Result we see that the top 3 districts for consumption in Punjab are given by their respected district codes ‘9’ – Ludhiana, ‘11’-Firozpur, ‘2’- Amritsar.

Similarly, the Bottom 3 districts for consumption in Punjab are given by their respected district codes ‘19’- Barnala, ‘18’- SJAS Nagar(Mohali), ‘8’- Fatehgarh Sahib.

This results allow us to identify where exactly the market entry and setup for business is feasible as well as for policymakers showcase the low and high points of the state.

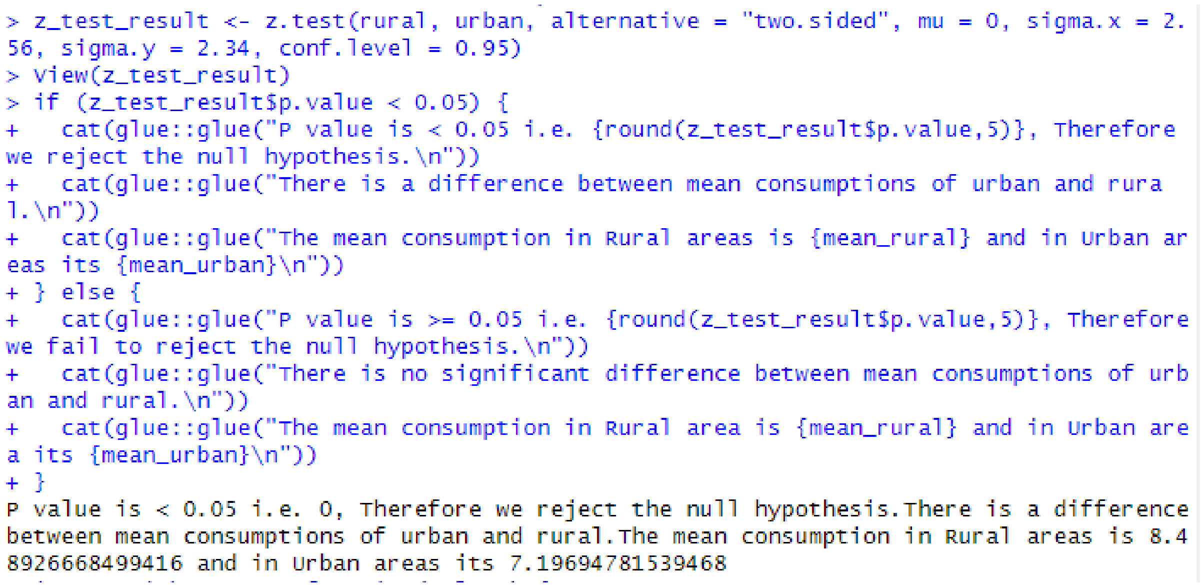
4. Renaming the districts as well as the sector viz. rural and urban.



```
> district_mapping <- c("9" = "Ludhiana", "11" = "Firozpur", "2" = "Amritsar")
> sector_mapping <- c("2" = "URBAN", "1" = "RURAL")
> print(district_mapping)
         9         11          2
"Ludhiana" "Firozpur" "Amritsar"
> print(sector_mapping)
      2       1
"URBAN" "RURAL"
```

From the above picture, we observe how we can easily rename the top 3 districts and then also be able to sector map it into either rural or urban. Renaming the district allows information provided by the data to be more easily retained and understood.

5. Test whether the differences between the mean are significant or not



```
> z_test_result <- z.test(rural, urban, alternative = "two.sided", mu = 0, sigma.x = 2.
56, sigma.y = 2.34, conf.level = 0.95)
> View(z_test_result)
> 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 rura
l.\n"))
+   cat(glue::glue("The mean consumption in Rural areas is {mean_rural} and in Urban ar
eas 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 urb
an and rural.\n"))
+   cat(glue::glue("The mean consumption in Rural area is {mean_rural} and in Urban are
a its {mean_urban}\n"))
+ }
P value is < 0.05 i.e. 0, Therefore we reject the null hypothesis.There is a difference
between mean consumptions of urban and rural.The mean consumption in Rural areas is 8.4
8926668499416 and in Urban areas its 7.19694781539468
```

After obtaining the mean for both rural and urban area for consumption, we perform the z test, to test the significance of the difference between the mean.

Here we assume,

Null Hypothesis – There is no difference in Consumption between urban and rural areas.

Alternate Hypothesis- there is a statistically significant difference in Consumption between urban and rural areas.

Interpretation: P value is < 0.05 i.e. 0, Therefore we reject the null hypothesis. There is a difference between mean consumptions of urban and rural. The mean consumption in Rural areas is 8.48926668499416 and in Urban areas its 7.1969 4781539468