VIRGINIA COMMONWEALTH UNIVERSITY

Statistical Analysis & Modelling

A1a - Data Cleaning using NSSO - Consumption Data Set

State: **Punjab**

Using R

Submitted by

MONIKA SARADHA

#V01068784

Date of Submission: 06/04/2023

Table of Contents

1. Introduction

The NSSO-Consumption dataset is a product of the National Sample Survey Organization (NSSO), which was established by the Government of India in 1969. Recognizing the challenges of collecting information from every individual in a large country like India, the NSSO employs scientific sampling methods to collect socio-economic data. The surveys are conducted in rounds, with each round spanning a period of six months to one year. There are two types of samples: the "Central Samples," conducted by the Government of India, and the "State Samples," conducted by the respective states.

1.1. About the Data

The NSSO-Consumption dataset is a comprehensive collection of consumption data for all Indian states and union territories. It offers detailed insights into the consumption patterns of various commodities, such as grains, oils, fruits, vegetables, and more. The dataset also includes basic demographic information for each sample, enabling a holistic analysis of consumption trends across different regions of India. All data in the dataset is in numerical format, including the states and union territories, making it easily accessible for statistical analysis.

1.2. Objective

The primary goal of the NSSO-Consumption dataset is to provide useful data for policymaking, planning, and research. Policymakers can develop targeted interventions to promote economic growth, social welfare, and sustainable development by studying consumption patterns. This dataset can be used by researchers to better understand the factors that influence consumption behavior, identify regional variations, and investigate the impact of demographic variables on consumption habits. The dataset's goal is to facilitate evidence-based decision-making and contribute to a better understanding of India's consumption dynamics.

For the dataset in the state of Punjab:

- Check the dataset for missing values for the assigned variables and replace them with the mean of the variable.

- Identify and describe any outliers in the dataset, and make any necessary changes.
- Rename districts and sectors to provide more descriptive and clear labels for variables.
- Summarize critical variables by region and district, emphasizing the top three and bottom three districts in terms of consumption levels.
- To determine if there are significant differences, test the significance of mean differences in consumption variables between regions or districts.

Based on the results, provide insights and analysis to inform decision-making and policy formulation regarding consumption patterns.

1.3. Business Significance

This extensive collection of primary data, auxiliary information, and socioeconomic indicators enriches the NSSO-Consumption dataset, allowing researchers, policymakers, and analysts to investigate various dimensions of consumption patterns and their underlying factors in India.

Understanding consumer behavior and consumption patterns is critical for companies operating in a variety of industries in order to conduct effective market research, product development, and marketing strategies. Businesses can gain a comprehensive understanding of the demand for various products and services across regions by leveraging the insights from this dataset, identifying potential market opportunities, and tailoring their offerings to meet consumer preferences. Businesses can also use the dataset to examine the impact of socioeconomic factors on consumption, identify target demographics, and optimize resource allocation for maximum profitability.

The NSSO primarily conducts four types of surveys: household surveys, enterprise surveys, village facilities, and land and livestock holdings. Provided state of Punjab comprises the following 4 division: Survey Design and Research (SDR), Field Operation Division (FOD), Data Process, and Economic Analysis.

2. Results

2.1. R- output and Interpretation

A **subset** was constructed using certain vital variables specific to the data set of the state Punjab.

- Sector: Refers to the sector of the economy or the type of area, such as rural or urban.
- State_Region: Represents the region or state within the dataset.
- District: Refers to the specific district within a state or region.
- Sex: Represents the gender of the individual.
- Age: Indicates the age of the individual.
- No_of_Meals_per_day: Represents the number of meals consumed per day by the individual.
- wheattotal_q: Refers to the quantity of wheat consumed.
- cerealtot_q: Represents the quantity of cereals consumed.
- moong_q: Indicates the quantity of moong (lentils) consumed.
- pulsestot_q: Represents the total quantity of pulses consumed.
- milk_q: Indicates the quantity of milk consumed.
- onion_q: Represents the quantity of onions consumed.
- potato_q: Indicates the quantity of potatoes consumed.

Structure of the dataset

```
> # View the structure
> str(subset_punjabds)
tibble [3,118 × 13] (S3: tbl_df/tbl/data.frame)
 $ Sector             : num [1:3118] 2 2 2 2 2 2 2 2 2 ...
 $ State_Region       : num [1:3118] 32 32 32 32 32 32 32 32 32 32 ...
 $ District           : num [1:3118] 9 9 9 9 9 9 9 13 13 ...
 $ Sex                : num [1:3118] 1 1 1 1 1 2 1 1 1 ...
 $ Age                : num [1:3118] 75 60 33 42 50 60 28 39 40 35 ...
 $ No_of_Meals_per_day: num [1:3118] 3 3 3 3 3 3 3 3 3 ...
 $ wheattotal_q       : num [1:3118] 8 10 5 3.75 10 7 5 7 10 10 ...
 $ cerealtot_q        : num [1:3118] 8.84 10.8 6.25 4.75 10.67 ...
 $ moong_q            : num [1:3118] 0.2 0.2 0.25 0.25 0.167 ...
 $ pulsestot_q        : num [1:3118] 1.1 1 1.25 1.25 2.17 ...
 $ milk_q             : num [1:3118] 18.7 18.7 11.7 11.7 31.2 ...
 $ onion_q            : num [1:3118] 0.8 1 1 1.25 1.33 ...
 $ potato_q           : num [1:3118] 1.4 1 1 1.25 2 ...
>
```

Inference:

The `str` function in R provides a concise summary of the structure of a dataset.

3,118 observations (rows) and 13 variables (columns).

The variables have different data types:

- Sector, State_Region, District, Sex: These variables are represented as **numeric values**. They likely serve as identifiers or **categorical indicators**.

- Age, No_of_Meals_per_day, wheattotal_q, cerealtot_q, moong_q, pulsestot_q, milk_q, onion_q, potato_q: These variables are **numeric**, representing **continuous values**. They likely capture quantities, ages, and number of meals.

To view the first few rows:

```
> # View the first few rows
> head(subset_punjabds)
# A tibble: 6 × 13
  Sector State_Region Distr…¹   Sex   Age No_of…² wheat…³ cerea…⁴ moong_q
   <dbl>        <dbl>   <dbl> <dbl> <dbl>   <dbl>   <dbl>   <dbl>   <dbl>
1      2           32       9     1    75       3     8       8.84   0.2
2      2           32       9     1    60       3    10      10.8    0.2
3      2           32       9     1    33       3     5       6.25   0.25
4      2           32       9     1    42       3     3.75    4.75   0.25
5      2           32       9     1    50       3    10      10.7    0.167
6      2           32       9     2    60       3     7       7.52   0.1
# … with 4 more variables: pulsestot_q <dbl>, milk_q <dbl>,
#   onion_q <dbl>, potato_q <dbl>, and abbreviated variable names
#   ¹District, ²No_of_Meals_per_day, ³wheattotal_q, ⁴cerealtot_q
# i Use `colnames()` to see all variable names
```

Inference:

The output can be used to make an initial inference of the kind of variables in the dataset and their values. Possible missing values, data entry errors and formatting issues could be observed here.

To view last few rows:

```
> # View the last few rows
> tail(subset_punjabds)
# A tibble: 6 × 13
  Sector State_Region Distr…¹   Sex   Age No_of…² wheat…³ cerea…⁴ moong_q
   <dbl>        <dbl>   <dbl> <dbl> <dbl>   <dbl>   <dbl>   <dbl>   <dbl>
1      1           31       1     1    50       3    6.25    8.75   0.125
2      1           31       1     2    30       3    3.33    6.67   0.0833
3      1           31       1     1    36       3    7.5     8.6    0.25
4      1           31       1     1    50       3    6.67   10      0.0833
5      1           31       1     2    22       3    5       6.25   0.125
6      1           31       1     2    30       3    6       6.7    0.25
# … with 4 more variables: pulsestot_q <dbl>, milk_q <dbl>,
#   onion_q <dbl>, potato_q <dbl>, and abbreviated variable names
#   ¹District, ²No_of_Meals_per_day, ³wheattotal_q, ⁴cerealtot_q
# i Use `colnames()` to see all variable names
>
```

Inference:

Just to ensure the dataset is complete and avoid any discrepancies in the dataset the last few rows are observed. Both these tests for head and tail are carried out to ensure consistency in data.

To view the summary of the dataset:

```
> #View the Summary
> summary(subset_punjabds)
     Sector       State_Region      District          Sex
 Min.   :1.000   Min.   :31.00   Min.   : 1.000   Min.   :1.000
 1st Qu.:1.000   1st Qu.:31.00   1st Qu.: 4.000   1st Qu.:1.000
 Median :2.000   Median :32.00   Median : 9.000   Median :1.000
 Mean   :1.502   Mean   :31.53   Mean   : 9.274   Mean   :1.118
 3rd Qu.:2.000   3rd Qu.:32.00   3rd Qu.:14.000   3rd Qu.:1.000
 Max.   :2.000   Max.   :32.00   Max.   :20.000   Max.   :2.000
      Age        No_of_Meals_per_day  wheattotal_q     cerealtot_q
 Min.   : 8.00   Min.   :2.00        Min.   : 0.000   Min.   : 0.000
 1st Qu.:38.00   1st Qu.:3.00        1st Qu.: 6.250   1st Qu.: 7.650
 Median :46.00   Median :3.00        Median : 7.500   Median : 8.986
 Mean   :47.78   Mean   :2.89        Mean   : 7.743   Mean   : 9.057
 3rd Qu.:58.00   3rd Qu.:3.00        3rd Qu.: 9.000   3rd Qu.:10.250
 Max.   :95.00   Max.   :3.00        Max.   :41.667   Max.   :50.500
    moong_q         pulsestot_q         milk_q          onion_q
 Min.   :0.00000   Min.   : 0.0000   Min.   :  0.00   Min.   :0.000
 1st Qu.:0.08333   1st Qu.: 0.6667   1st Qu.:  7.80   1st Qu.:0.750
 Median :0.14286   Median : 0.9000   Median : 10.40   Median :1.000
 Mean   :0.16796   Mean   : 0.9755   Mean   : 12.82   Mean   :1.183
 3rd Qu.:0.25000   3rd Qu.: 1.1667   3rd Qu.: 15.60   3rd Qu.:1.500
 Max.   :1.50000   Max.   :14.6667   Max.   :124.80   Max.   :8.333
    potato_q
 Min.   : 0.000
 1st Qu.: 1.000
 Median : 1.333
 Mean   : 1.456
 3rd Qu.: 1.750
 Max.   :16.667
>
```

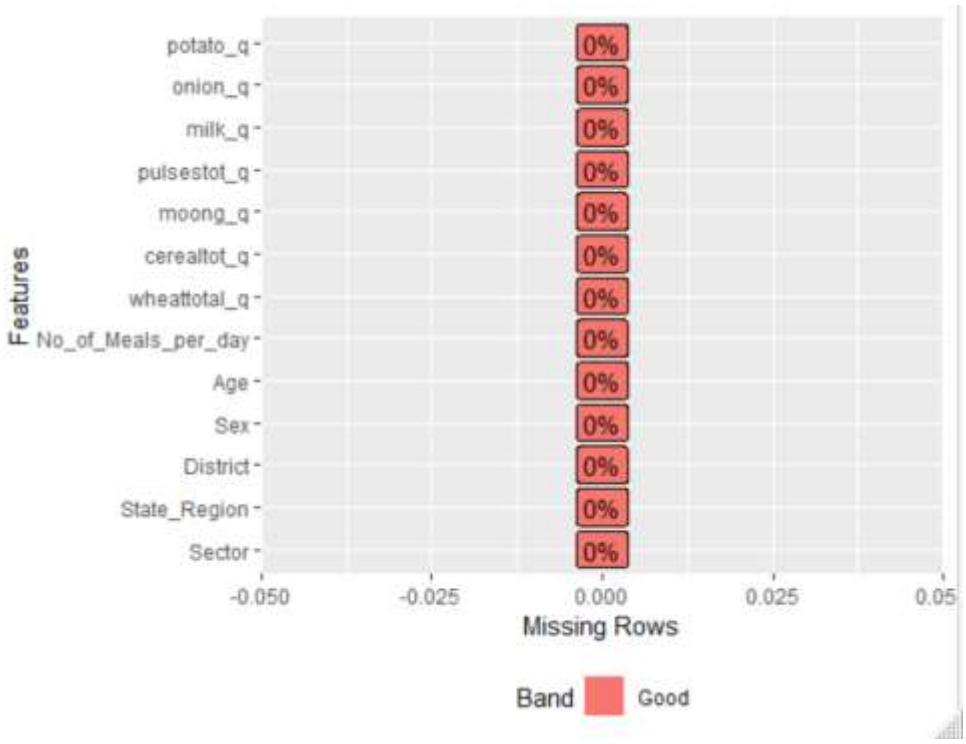
Inference:

- Provides an overview of the subset created.
- "Sector" variable indicates that there are two sectors present in the dataset.
- "State_Region" variable shows that the state/region codes range from 31 to 32.
- "District" variable ranges from 1 to 20, indicating different districts within Punjab.
- "Sex" variable indicates the gender of the individuals, with values 1 and 2 representing male and female, respectively.
- "Age" variable ranges from 8 to 95, representing the age of the individuals.

- "No_of_Meals_per_day" variable shows that the majority of individuals consume three meals per day.
- The remaining variables (wheattotal_q, cerealtot_q, moong_q, pulsestot_q, milk_q, onion_q, potato_q) represent the quantities of respective food items consumed by the individuals.
- Summary provides information about the minimum, maximum, median, mean, and quartiles for each variable.

2.2. Missing Value Analysis

Missing Plot:



Sum of missing values:

```
> sum(is.na(subset_punjabds))
[1] 0
```

Inference:

The obtained missing plot of the chosen subset of Punjab showed there are no missing values, which means all values in the data are available for analysis.

```
> na_count <- sum(is.na(subset_punjabds))
> print(na_count)
[1] 0
>
```

Since there are 0 missing values and NA values in the subset chosen, we could proceed with the current data for further analysis.

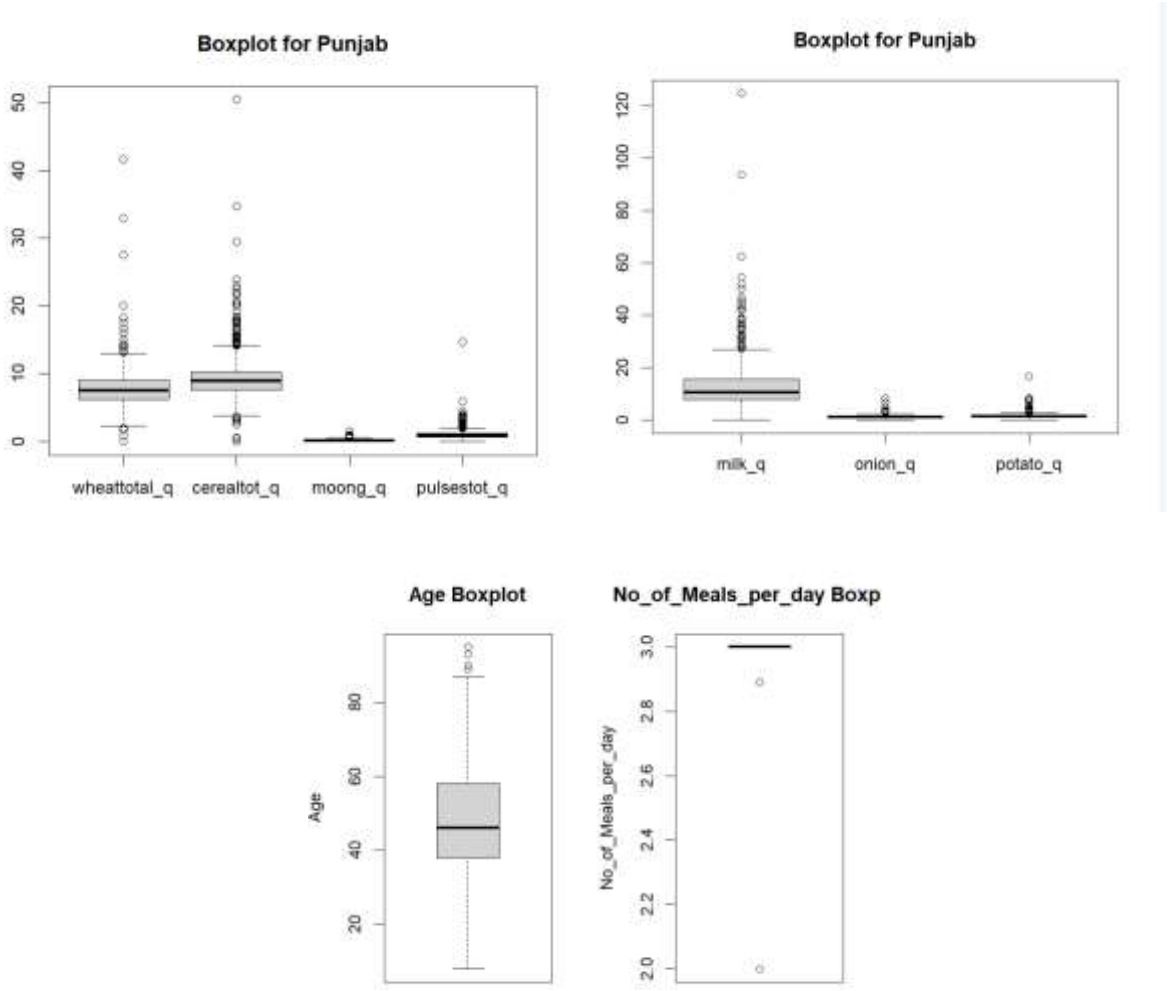
Imputation of missing values:

```
> if (any(is.na(subset_punjabds))) {
+   # Impute missing values with the mean of the respective column
+   punjab_imputed <- subset_punjabds
+   for (col in colnames(subset_punjabds)) {
+     if (any(is.na(subset_punjabds[[col]]))) {
+       col_mean <- mean(subset_punjabds[[col]], na.rm = TRUE)
+       punjab_imputed[[col]][is.na(subset_punjabds[[col]])] <- col_mean
+     }
+   }
+
+  # Print the imputed dataset
+   print(punjab_imputed)
+ } else {
+   print("No missing values found.")
+ }
[1] "No missing values found."
>
```

Inference:

Other methods to handle missing values would be to remove them or imputation by means of mean, median and mode.

2.3. Outliers Identification and Amendments



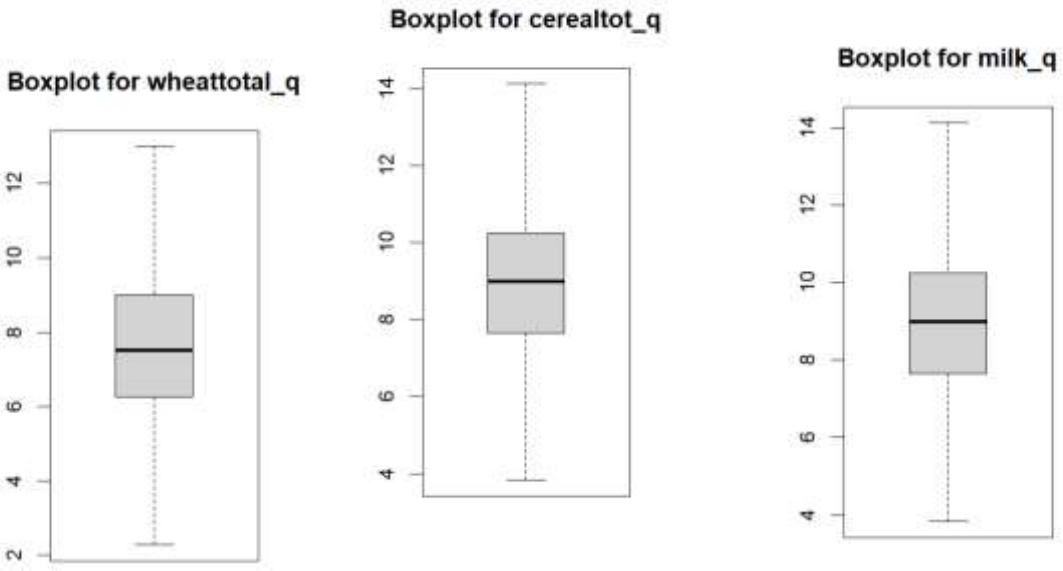
Inference:

The categorical variables can be ignored in terms of analyzing the outliers. If required these can be converted to numeric in order to analyze, since in this particular instance they do not hold any significant value, we choose to ignore them.

The variables such as wheattotal_q, cerealtot_q and milk_q could be observed to have an ample number of outliers which is to worked on before proceeding with analyzing this subset.

Amendment of outliers using Quantiles:

```
> for (wheattotal_q in names(subset_punjabds)) {
+
+   lower_quantile <- quantile(subset_punjabds[[wheattotal_q]], 0.25)
+   upper_quantile <- quantile(subset_punjabds[[wheattotal_q]], 0.75)
+
+
+   iqr <- upper_quantile - lower_quantile
+
+
+   lower_bound <- lower_quantile - 1.5 * iqr
+   upper_bound <- upper_quantile + 1.5 * iqr
+
+
+   subset_punjabds[[wheattotal_q]][subset_punjabds[[wheattotal_q]] < lower_bound]
<- lower_quantile
+   subset_punjabds[[wheattotal_q]][subset_punjabds[[wheattotal_q]] > upper_bound]
<- upper_quantile
+ }
```



Inference:

Using the above mentioned code the outliers have been replaced with the upper or lower quantile values. Post which the box plot result of these variables showed the absence of outliers as they have been replaced.

2.4. Renaming

Renaming the Districts and the Sector (Rural & Urban):

```
# A tibble: 6 × 13
  Sector State_Reg…¹ Distr…²   Sex   Age No_of…³ wheat…⁴ cerea…⁵ moong_q pulse…⁶
  <chr>        <dbl> <chr>    <dbl> <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>
1 Urban           32 Moga         1    75       3    8       8.84    0.2     1.1
2 Urban           32 Moga         1    60       3   10      10.8     0.2     1
3 Urban           32 Moga         1    33       3    5       6.25    0.25    1.25
4 Urban           32 Moga         1    42       3    3.75    4.75    0.25    1.25
5 Urban           32 Moga         1    50       3   10      10.7     0.167   1.17
6 Urban           32 Moga         1    60       3    7       7.52    0.1     1
# … with 3 more variables: milk_q <dbl>, onion_q <dbl>, potato_q <dbl>, and
#   abbreviated variable names ¹State_Region, ²District, ³No_of_Meals_per_day,
#   ⁴wheattotal_q, ⁵cerealtot_q, ⁶pulsestot_q
# ℹ Use `colnames()` to see all variable names
```

Count of the data collected from Urban & Rural and different districts:

```
> # Count of urban and rural sectors
> sector_count <- table(subset_punjabds$Sector)
> sector_count

Rural Urban
 1552  1566
> # Count of different districts
> district_count <- table(subset_punjabds$District)
> district_count

          Amritsar            Barnala          Bathinda           Faridkot
               288                 96               192                224
   Fatehgarh Sahib           Firozpur         Gurdaspur         Hoshiarpur
               184                 64               224                 96
         Jalandhar         Kapurthala          Ludhiana              Mansa
               128                 96               224                 87
              Moga Mohali (SAS Nagar)           Muktsar          Pathankot
               383                 96               160                 64
           Patiala           Rupnagar           Sangrur         Tarn Taran
               256                 64                96                 96
>
```

Inference:

The dataset includes both rural and urban areas. According to the count, there are slightly more urban sectors (1566) than rural sectors (1552). Moga has the highest number of districts with 383, followed by Amritsar with 288. Barnala, Firozpur, Hoshiarpur, Kapurthala, Pathankot, Rupnagar, Sangrur, and Tarn Taran have relatively lower counts ranging from 64 to 96. These data aid in analyzing the distribution between districts and sectors.

2.5. Summary of Critical Variables region wise and district wise

Critical Variables Chosen: wheattotal_q, cerealtot_q, moong_q, pulsestot_q, milk_q, onion_q, and potato_q

Region-wise Summary:

```
> print(region_summary)
# A tibble: 2 × 8
  State_Region mean_wheattotal_q mean_…¹ mean_…² mean_…³ mean_…⁴ mean_…⁵ mean_…⁶
         <db l>             <db l>  <db l>  <db l>  <db l>  <db l>  <db l>  <db l>
1           31              7.26    8.77   0.139   0.941    11.5    1.17    1.41
2           32              8.10    9.18   0.177   0.893    11.6    1.10    1.32
# … with abbreviated variable names ¹mean_cerealtot_q, ²mean_moong_q,
#   ³mean_pulsestot_q, ⁴mean_milk_q, ⁵mean_onion_q, ⁶mean_potato_q
>
```

Inference:

According to the summary, state region 31 has a slightly lower mean value for wheat consumption than state region 32. State region 32, on the other hand, has a higher mean value for cereal consumption, indicating a potentially higher cereal consumption in that region. State region 32 has slightly higher mean values for moong dal, pulses, milk, onion, and potato than state region 31.

District-wise Summary:

```
> print(district_summary)
# A tibble: 20 × 8
   District           mean_whe…¹ mean_…² mean_…³ mean_…⁴ mean_…⁵ mean_…⁶ mean_…⁷
   <chr>                   <db l>  <db l>  <db l>  <db l>  <db l>  <db l>  <db l>
 1 Amritsar                 7.56    9.07   0.125    1.02    11.0    1.17    1.64
 2 Barnala                  8.37    9.00   0.196   0.795    13.2    1.14    1.17
 3 Bathinda                 6.93    8.54   0.138   0.964    12.3    1.31    1.42
 4 Faridkot                 7.27    8.70   0.171   0.971    11.4    1.03    1.34
 5 Fatehgarh Sahib          7.38    8.50   0.190   0.936    12.2    1.09    1.30
 6 Firozpur                 9.47   10.5    0.202   0.900    11.8    1.01    1.14
 7 Gurdaspur                9.48   10.5    0.186   0.872    11.6    1.12    1.27
 8 Hoshiarpur               7.52    8.74   0.160   0.765    12.2    1.17    1.40
 9 Jalandhar                7.07    8.44   0.127   0.749    11.9    1.05    1.30
10 Kapurthala              10.1    10.6    0.152   0.771    11.2    1.10    1.38
11 Ludhiana                 6.75    8.91   0.116    1.03    11.4    1.42    1.43
12 Mansa                    7.42    8.94   0.159    1.02    11.4   0.904    1.06
13 Moga                     7.17    8.55   0.174    1.00    10.8    1.13    1.53
14 Mohali (SAS Nagar)       7.88    8.93   0.169    1.01    11.3   0.950    1.14
15 Muktsar                  8.28    8.87   0.169   0.667    13.4    1.28    1.22
16 Pathankot                7.55    8.85   0.157   0.896    11.0   0.961    1.26
17 Patiala                  7.14    8.47   0.156   0.845    11.0    1.10    1.35
18 Rupnagar                 7.84    8.73   0.167   0.878    12.0   0.706    1.01
19 Sangrur                  8.97    9.98   0.183   0.843    10.6    1.19    1.26
20 Tarn Taran               7.81    8.90   0.135   0.990    12.5    1.05    1.56
# … with abbreviated variable names ¹mean_wheattotal_q, ²mean_cerealtot_q,
#   ³mean_moong_q, ⁴mean_pulsestot_q, ⁵mean_milk_q, ⁶mean_onion_q,
#   ⁷mean_potato_q
```

Inference:

From the summary we can see differences in consumption patterns across different districts. Districts such as Kapurthala, Gurdaspur, and Muktsar, for example, have relatively higher mean values for wheat consumption, indicating potentially higher wheat consumption in these areas. Wheat consumption is mean values are relatively lower in districts such as Firozpur, Rupnagar, and Bathinda.

Top three districts and the bottom three districts of consumption:

```
> cat("Top Three Districts (Overall Consumption):\n")
Top Three Districts (Overall Consumption):
> print(top_three_districts)
# A tibble: 3 × 8
  District  mean_wheattotal_q mean_cer…¹ mean_…² mean_…³ mean_…⁴ mean_…⁵ mean_…⁶
  <chr>                 <dbl>      <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>
1 Patiala                7.14       8.47   0.156   0.845    11.0    1.10    1.35
2 Moga                   7.17       8.55   0.174   1.00     10.8    1.13    1.53
3 Jalandhar              7.07       8.44   0.127   0.749    11.9    1.05    1.30
# … with abbreviated variable names ¹mean_cerealtot_q, ²mean_moong_q,
#   ³mean_pulsestot_q, ⁴mean_milk_q, ⁵mean_onion_q, ⁶mean_potato_q
> cat("Bottom Three Districts (Overall Consumption):\n")
Bottom Three Districts (Overall Consumption):
> print(bottom_three_districts)
# A tibble: 3 × 8
  District   mean_wheattotal_q mean_ce…¹ mean_…² mean_…³ mean_…⁴ mean_…⁵ mean_…⁶
  <chr>                  <dbl>     <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>
1 Gurdaspur               9.48      10.5   0.186   0.872    11.6    1.12    1.27
2 Firozpur                9.47      10.5   0.202   0.900    11.8    1.01    1.14
3 Kapurthala             10.1       10.6   0.152   0.771    11.2    1.10    1.38
# … with abbreviated variable names ¹mean_cerealtot_q, ²mean_moong_q,
#   ³mean_pulsestot_q, ⁴mean_milk_q, ⁵mean_onion_q, ⁶mean_potato_q
>
```

Inference:

Patiala, Moga, and Jalandhar are the top three districts in terms of overall consumption. These districts have relatively higher mean values, indicating that their residents consume more of these food items on average. **Gurdaspur, Firozpur, and Kapurthala are the bottom three districts** with the lowest overall consumption. These districts have significantly lower mean values. This implies that residents of these districts consume fewer of these food items on average.

Factors that are affecting this disparity are income, education, food accessibility, cultural dietary preferences, government policies and health awareness.

2.6. Hypothesis Testing

Null Hypothesis (H0): There is no significant difference in the means of rural consumption and urban consumption.

Alternate Hypothesis (Ha): There is a significant difference between the means of rural consumption and urban consumption.

```
> print(result)

        Two-sample z-Test

data:  z_rural and z_urban
z = 5.3024, p-value = 1.143e-07
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.2257141 0.4904224
sample estimates:
mean of x mean of y
 4.727947  4.369879

>
```

Inference:

The test produced a highly significant test statistic (z-value) of 5.3024. The p-value (1.143e-07) is extremely low, providing strong evidence that the true difference in means between rural and urban consumption is not zero.

P value < alpha (0.05) Reject Null Hypothesis (H0) and accept Alternate Hypothesis (Ha).

The 95% confidence interval for the mean difference also supports the conclusion that there is a significant difference.

Therefore, we can conclude that the consumption levels in rural and urban areas differ significantly. The rural consumption rate is significantly higher.

3. Recommendation

3.1. Business Implications

- Ludhiana district in Punjab shows high wheat consumption, presenting an opportunity for businesses in the wheat product industry.

- Fazilka district in Punjab has low milk consumption, indicating a potential market gap for dairy products.
- Rural areas exhibit higher fruit consumption compared to urban areas, suggesting businesses should target rural markets for fruit products.
- Urban areas in Punjab have higher consumption of milk compared to rural areas, indicating a potential market for businesses in the beverage industry to target urban consumers.

3.2. Business Recommendations

- Targeted Marketing Strategies: To cater to specific consumer preferences, businesses can develop targeted marketing strategies based on regional consumption patterns.
- Product diversification: Businesses can broaden product offerings to meet the diverse consumption habits of different regions and districts.
- Collaboration with Local Suppliers: Form alliances with local suppliers to ensure a consistent supply of desired food items while also supporting the local economy.
- In high consuming regions businesses can focus on market expansion, offering premium products and increased customer engagement.
- In low consuming regions business can focus on price optimization, market penetration, product adaptations and increasing the awareness.

4. Reference:

- NSS & Tabulation | Department of Economic and Statistical Affairs Haryana | India. (n.d.). NSS & Tabulation | Department of Economic and Statistical Affairs Haryana | India. https://esaharyana.gov.in/nss-tabulation/#:~:text=The%20National%20Sample%20Survey%20Organization,done%20by%20E.S.O.%2C%20Planning%20Department.

5. Codes

5.1. R-studio

```
library(readxl)

punjab_ds<-read_excel("C:\\Users\\monis\\OneDrive\\Desktop\\ASSG1.xlsx")

#Subset the variables

subset_punjabds<-
punjab[,c("Sector","State_Region","District","Sex","Age","No_of_Meals_per_day","wheattotal_q","cerealtot_q","moong_q", "pulsestot_q","milk_q","onion_q","potato_q")]

# View the structure

str(subset_punjabds)

# View the first few rows

head(subset_punjabds)

# View the last few rows

tail(subset_punjabds)

#View the Summary

summary(subset_punjabds)

#-----------------------------------------------

#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.

library(DataExplorer)

plot_missing(subset_punjabds)
```

```
is.na(subset_punjabds)

sum(is.na(subset_punjabds))

# Check if there are any missing values (NA)

any_na <- any(is.na(subset_punjabds))

# Count the number of missing values (NA)

na_count <- sum(is.na(subset_punjabds))

print(na_count)
```

#Imputation of MV with Mean

```
if (any(is.na(subset_punjabds))) {

  # Impute missing values with the mean of the respective column

  punjab_imputed <- subset_punjabds

  for (col in colnames(subset_punjabds)) {

    if (any(is.na(subset_punjabds[[col]]))) {

      col_mean <- mean(subset_punjabds[[col]], na.rm = TRUE)

      punjab_imputed[[col]][is.na(subset_punjabds[[col]])] <- col_mean

    }

  }

 # Print the imputed dataset

  print(punjab_imputed)

} else {

  print("No missing values found.")

}

#-----------------------------------------------------
```

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

```
# Boxplot 1 with variables: wheattotal_q, cerealtot_q, moong_q, pulsestot_q
boxplot(subset_punjabds$wheattotal_q, subset_punjabds$cerealtot_q, subset_punjabds$moong_q, subset_punjabds$pulsestot_q,
        main = "Boxplot for Punjab",
        names = c("wheattotal_q", "cerealtot_q", "moong_q", "pulsestot_q"))
# Boxplot 2 with variables: milk_q, onion_q, potato_q
boxplot(subset_punjabds$milk_q, subset_punjabds$onion_q, subset_punjabds$potato_q,
        main = "Boxplot for Punjab",
        names = c("milk_q", "onion_q", "potato_q"))
# Create a new plotting window
par(mfrow = c(1, 2))
# Boxplot for Age
boxplot(subset_punjabds$Age,
        main = "Age Boxplot",
        ylab = "Age")

# Boxplot for No_of_Meals_per_day
boxplot(subset_punjabds$No_of_Meals_per_day,
        main = "No_of_Meals_per_day Boxplot",
        ylab = "No_of_Meals_per_day")
# Reset the plotting layout
par(mfrow = c(1, 2))
#Amendment of outliers using Quantiles
for (wheattotal_q in names(subset_punjabds)) {
```

```
  lower_quantile <- quantile(subset_punjabds[[wheattotal_q]], 0.25)
  upper_quantile <- quantile(subset_punjabds[[wheattotal_q]], 0.75)



  iqr <- upper_quantile - lower_quantile
  lower_bound <- lower_quantile - 1.5 * iqr
  upper_bound <- upper_quantile + 1.5 * iqr
  subset_punjabds[[wheattotal_q]][subset_punjabds[[wheattotal_q]] < lower_bound] <- lower_quantile
  subset_punjabds[[wheattotal_q]][subset_punjabds[[wheattotal_q]] > upper_bound] <- upper_quantile
}
boxplot(subset_punjabds$wheattotal_q,
        main = "Boxplot for wheattotal_q",
        names = "wheattotal_q")
for (cerealtot_q in names(subset_punjabds)) {
  lower_quantile <- quantile(subset_punjabds[[cerealtot_q]], 0.25)
  upper_quantile <- quantile(subset_punjabds[[cerealtot_q]], 0.75)
  iqr <- upper_quantile - lower_quantile
  lower_bound <- lower_quantile - 1.5 * iqr
  upper_bound <- upper_quantile + 1.5 * iqr
  subset_punjabds[[cerealtot_q]][subset_punjabds[[cerealtot_q]] < lower_bound] <- lower_quantile
  subset_punjabds[[cerealtot_q]][subset_punjabds[[cerealtot_q]] > upper_bound] <- upper_quantile
}
boxplot(subset_punjabds$cerealtot_q,
        main = "Boxplot for cerealtot_q",
```

```
      names = "cerealtot_q")
for (milk_q in names(subset_punjabds)) {
  lower_quantile <- quantile(subset_punjabds[[milk_q]], 0.25)
  upper_quantile <- quantile(subset_punjabds[[milk_q]], 0.75)
  iqr <- upper_quantile - lower_quantile
  lower_bound <- lower_quantile - 1.5 * iqr
  upper_bound <- upper_quantile + 1.5 * iqr
  subset_punjabds[[milk_q]][subset_punjabds[[milk_q]] < lower_bound] <- lower_quantile
  subset_punjabds[[milk_q]][subset_punjabds[[milk_q]] > upper_bound] <- upper_quantile
}
boxplot(subset_punjabds$cerealtot_q,
      main = "Boxplot for milk_q",
      names = "milk_q")
#---------------------------------------------------------------------
```

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

```
library(dplyr)
# Define district names
district_names <- c("Ludhiana", "Amritsar", "Jalandhar", "Patiala", "Bathinda", "Hoshiarpur", "Mohali (SAS Nagar)", "Pathankot", "Moga", "Sangrur", "Gurdaspur", "Kapurthala", "Firozpur", "Muktsar", "Barnala", "Fatehgarh Sahib", "Faridkot", "Mansa", "Rupnagar", "Tarn Taran")
# Replace district values with names
subset_punjabds$District <- district_names[subset_punjabds$District]
# Replace sector values with "Rural" and "Urban"
subset_punjabds$Sector <- ifelse(subset_punjabds$Sector == 1, "Rural", "Urban")
head(subset_punjabds)
```

```
# Count of urban and rural sectors
sector_count <- table(subset_punjabds$Sector)
sector_count


# Count of different districts
district_count <- table(subset_punjabds$District)
district_count
#-----------------------------------------------------
```

#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.

```
library(dplyr)
#region summary
region_summary <- subset_punjabds %>%
  group_by(State_Region) %>%
  summarize(mean_wheattotal_q = mean(wheattotal_q),
            mean_cerealtot_q = mean(cerealtot_q),
            mean_moong_q = mean(moong_q),
            mean_pulsestot_q = mean(pulsestot_q),
            mean_milk_q = mean(milk_q),
            mean_onion_q = mean(onion_q),
            mean_potato_q = mean(potato_q))
print(region_summary)
#district summary
district_summary <- subset_punjabds %>%
```

```
  group_by(District) %>%
  summarize(mean_wheattotal_q = mean(wheattotal_q),
            mean_cerealtot_q = mean(cerealtot_q),
            mean_moong_q = mean(moong_q),
            mean_pulsestot_q = mean(pulsestot_q),
            mean_milk_q = mean(milk_q),
            mean_onion_q = mean(onion_q),
            mean_potato_q = mean(potato_q))
print(district_summary)
```

#top 3 and bottom 3 districts of consumption

```
sorted_districts <- district_summary[order(rowSums(district_summary[, c("mean_wheattotal_q", "mean_cerealtot_q", "mean_moong_q", "mean_pulsestot_q", "mean_milk_q", "mean_onion_q", "mean_potato_q")])), ]
```

 # Identify the top three and bottom three districts

```
top_three_districts <- head(sorted_districts, 3)
bottom_three_districts <- tail(sorted_districts, 3)
#  top three districts
cat("Top Three Districts (Overall Consumption):\n")
print(top_three_districts)
#  bottom three districts
cat("Bottom Three Districts (Overall Consumption):\n")
print(bottom_three_districts)
#-----------------------------------------------------------------
```

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

```
library(BSDA)
```

```
# Subset the data for the rural and urban sectors

rural_consumption <- subset(subset_punjabds, Sector == "Rural")

urban_consumption <- subset(subset_punjabds, Sector == "Urban")


# Extract the variables for the z-test

z_rural <- c(rural_consumption$potato_q, rural_consumption$onion_q, rural_consumption$moong_q, rural_consumption$pulsestot_q, rural_consumption$wheattotal_q, rural_consumption$milk_q, rural_consumption$cerealtot_q)

z_urban <- c(urban_consumption$potato_q, urban_consumption$onion_q, urban_consumption$moong_q, urban_consumption$pulsestot_q, urban_consumption$wheattotal_q, urban_consumption$milk_q, urban_consumption$cerealtot_q)

# Calculate sample standard deviations

sigma.x <- sd(z_rural)

sigma.y <- sd(z_urban)

# Perform the two-sample z-test

result <- z.test(z_rural, z_urban, alternative = "two.sided", mu = 0, sigma.x = sigma.x, sigma.y = sigma.y, conf.level = 0.95)

# Print the z-test result

print(result)
```