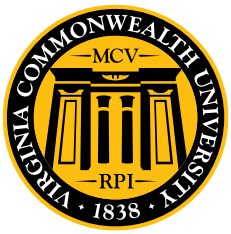


VIRGINIA COMMONWEALTH UNIVERSITY

Statistical analysis and modelling (SCMA 632)

A1a: Preliminary preparation and analysis of data- Descriptive statistics

PUSHPAK DEVAKI

V01108254

Date of Submission: 16-06-2024

CONTENTS

INTRODUCTION

The data procured from NSSO is analysed, in order to determine the top and bottom three consuming districts in Haryana. To get the data we need for analysis, we will clean and alter the dataset. The dataset contains district-specific variances together with consumption-related data for both the urban and rural sectors. The dataset has been loaded into R/Python, a potent statistical programming language renowned for its effectiveness in handling and analysing big datasets.

Our goals include summarizing consumption statistics by area and district, controlling outliers, detecting and fixing missing values, and determining the significance of mean variances. We also want to standardize district and sector names. Policymakers and other stakeholders will benefit greatly from the study's findings, which will enable focused initiatives and encourage equal development throughout 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, describe your test's outcome, and make suitable amendments.

c) Rename the districts and sectors, viz., rural and urban.

d) Summarize the critical variables in the data set region-wise and district-wise and indicate the top and bottom three districts of consumption.

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

RESULTS & 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. (From R)

Code:

```
> # Check for missing values in the subset
> cat ("Missing Values in Subset:\n")
Missing Values in Subset:
> print(colSums(is.na(hrnew)))
```

Result:

```
      state_1      District        Region        Sector
            0             0             0             0
 State_Region Meals_At_Home      ricepds_v    Wheatpds_q
            0            14             0             0
    chicken_q      pulsep_q     wheatos_q No_of_Meals_per_day
            0             0             0             0
```

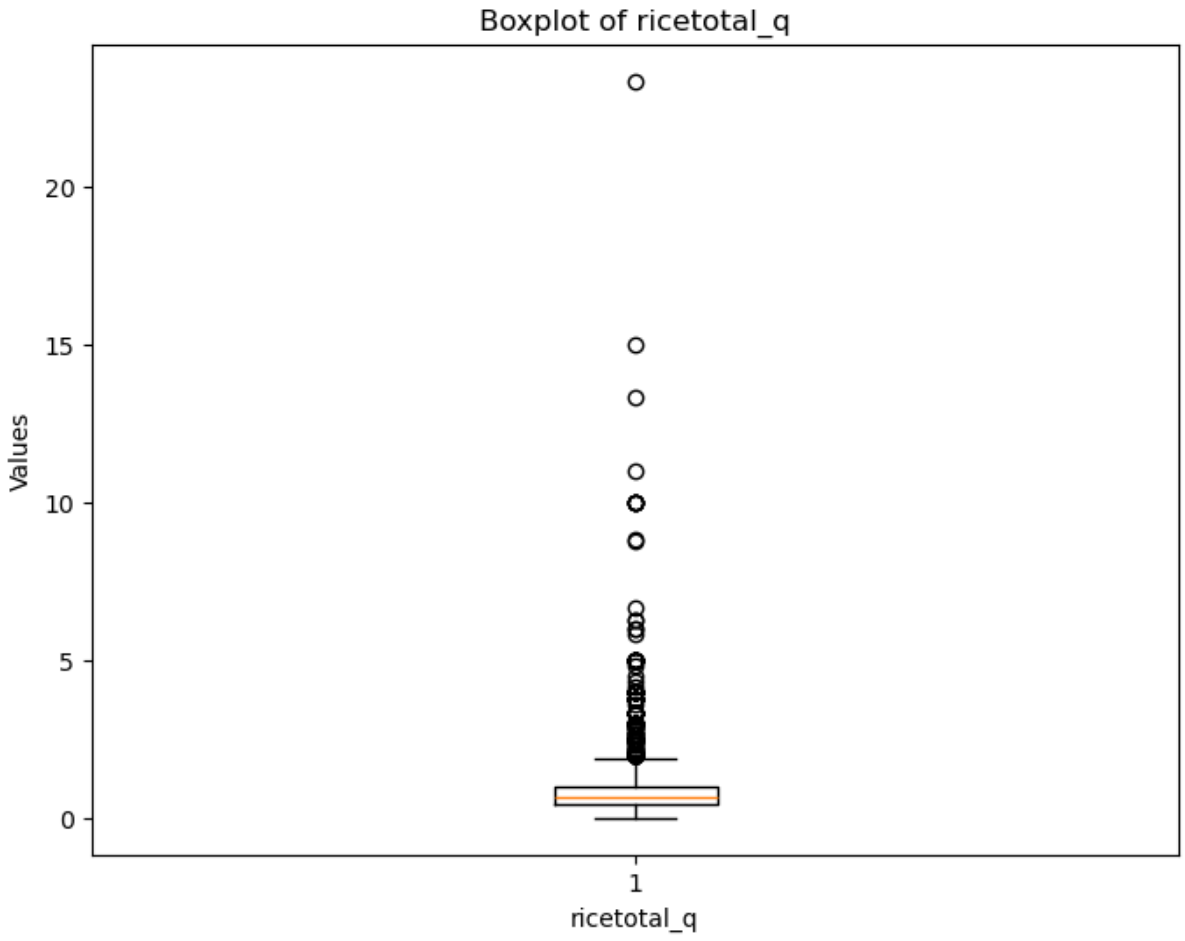
Interpretation:

From the above dataset, we can see there is complete data for most columns, including geographical and sectoral information, as well as quantities of rice, wheat, chicken, pulses, wheat output, and number of meals per day. However, the "Meals_At_Home" column has 14 missing values, indicating incomplete records for this variable. This missing data needs addressing, either through imputation or exclusion, before further analysis. Overall, the dataset is robust, allowing for comprehensive analysis of food consumption patterns across different regions and sectors.

B) Check for outliers, describe your test's outcome, and make suitable amendments.

(Code from Python)

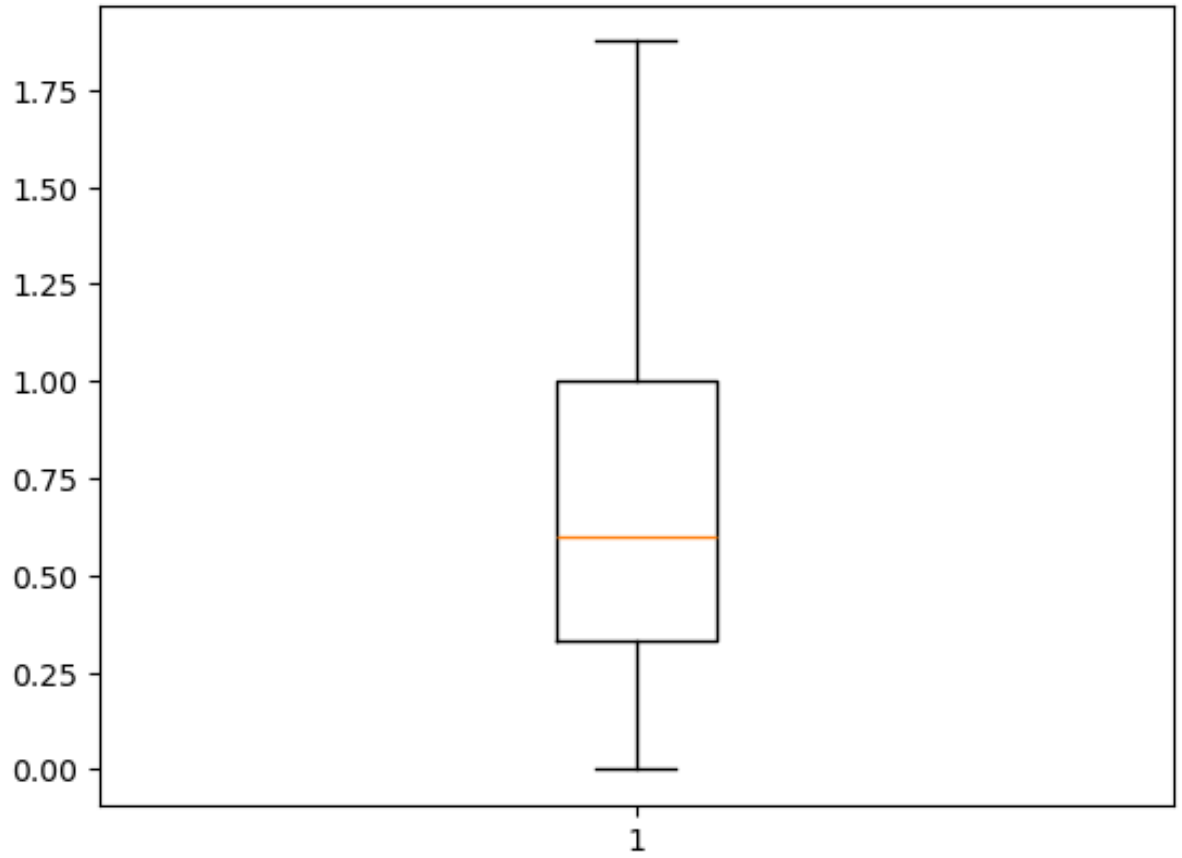
```
# Outlier Checking
import matplotlib.pyplot as plt
# Assuming HR_clean is your DataFrame
plt.figure(figsize=(8, 6))
plt.boxplot(HR_clean['ricetotal_q'])
plt.xlabel('ricetotal_q')
plt.ylabel('Values')
plt.title('Boxplot of ricetotal_q')
plt.show()
```



```
HR_clean=HR_new[(HR_new['ricetotal_q']<=up_limit)&(HR_new['ricetotal_q']>=low_limit
)]

plt.boxplot(HR_clean['ricetotal_q'])
```

```
{'whiskers': [<matplotlib.lines.Line2D at 0x133acd7d0>,  <matplotlib.lines.Line2D at 0x133
ace410>],  'caps': [<matplotlib.lines.Line2D at 0x133acf090>,  <matplotlib.lines.Line2D at 0
x133acfbd0>],  'boxes': [<matplotlib.lines.Line2D at 0x133accb90>],  'medians': [<matplotlib
.lines.Line2D at 0x133ad46d0>], 'fliers': [<matplotlib.lines.Line2D at 0x133ad5250>],
 'means': []}
```



Interpretation:

The "ricetotal_q" graph's boxplots, which are displayed above, shows the outliers. The distribution is displayed in the first boxplot, where several points are above the upper whisker, designating values more than 20 as outliers. The majority of the data points are concentrated in the middle box, which represents the interquartile range (IQR), and the whiskers (1.5 * IQR). The second boxplot, which shows the core data distribution with an IQR ranging from roughly 0.25 to 1.0 and a median of about 0.5, is devoid of outliers. High-value outliers are present in these plots, indicating the necessity for cautious handling in additional investigation.

c) Rename the districts as well as the sector, viz. rural and urban. (Code from R)

```
> # Rename districts and sectors , get codes from appendix of NSSO 68th ROund Data
> district_mapping <- c("13" = "Bhiwani", "19" = "Faridabad", "12"  .... [TRUNCATED]
> sector_mapping <- c("2" = "URBAN", "1" = "RURAL")
> hrnew$District <- as.character(hrnew$District)
> hrnew$Sector <- as.character(hrnew$Sector)
> hrnew$District <- ifelse(hrnew$District %in% names(district_mapping),
district_mapping[hrnew$District], hrnew$District)
> hrnew$Sector <- ifelse(hrnew$Sector %in% names(sector_mapping),
sector_mapping[hrnew$Sector], hrnew$Sector)
```

Result:

```
HR_clean['District'].unique()
array([19, 20, 18, 17, 13, 14, 16, 15, 12, 11,  8, 10,  7,  9,  6,  5,  2, 4,  1,  3], dtype=int64)
# Replace values in the 'Sector' column
HR_clean.loc[:,'Sector'] = HR_clean['Sector'].replace([1, 2], ['URBAN', 'RURAL'])
```

Interpretation:

We first converted the 'District' and 'Sector' columns to character type. Using the mappings (`district_mapping` and `sector_mapping`). Then replaced numeric district codes with their corresponding names and sector codes with "URBAN" or "RURAL."

D) Summarize the critical variables in the data set region-wise and district-wise and indicate the top and bottom three districts of consumption. (Code from R)

```
> # Summarize consumption
> hrnew$total_consumption <- rowSums(hrnew[, c("ricepds_v", "Wheatpds_q", "chicken_q",
"pulsep_q", "wheatos_q")], na.rm = TR .... [TRUNCATED]
```

```
> # Summarize and display top and bottom consuming districts and regions
> summarize_consumption <- function(group_col) {
+   summary <- hrnew %>%
+   .... [TRUNCATED]
> 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 × 2
  District total
     <int> <dbl>
1       13 1488.
2       19 1461.
3       12 1369.
> cat("Bottom 3 Consuming Districts:\n")
Bottom 3 Consuming Districts:
> print(tail(district_summary, 3))
# A tibble: 3 × 2
  District total
     <int> <dbl>
1       15  612.
2        1  343.
3       20  318.
> cat("Region Consumption Summary:\n")
Region Consumption Summary:
> print(region_summary)
# A tibble: 2 × 2
  Region  total
   <int>  <dbl>
```

1 1 10792.

2 2 8264.

Interpretation:

From the above coded snippet our intention was to find out the top 3 and bottom 3 districts from the given data set. So here we figured out that the top 3 consuming districts are 13, 19, 12 which are Bhiwani, Faridabad and Hisar respectively. And the bottom 3 consuming districts are 15, 1, 20 which are Jhajjar, Panchkula and Mewat respectively. So that would be the interpreted data from the above commands.

e) Test whether the differences in the means are significant or not. (Code from R)

```
> # Test for differences in mean consumption between urban and rural
> rural <- hrnew %>%
+   filter(Sector == "RURAL") %>%
+   select(total_consumpti .... [TRUNCATED]

> urban <- hrnew %>%
+   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)} ..." ... [TRUNCATED]
```

Result:

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 8.73202164858282 and in Urban areas its 7.46857953285784.

Interpretation:

The statistical analysis shows that the p-value is less than 0.05, which means that the null hypothesis is rejected. This suggests that the mean consumptions in urban and rural areas varied significantly. In urban regions, the mean consumption is 7.47, while in rural areas it is 8.73. As a result, the mean consumption in rural vs urban areas is higher, and this difference is statistically significant. This conclusion suggests that urban and rural people have different resource availability or consumption habits, which calls for more research to determine the underlying causes of this discrepancy.

RECOMMENDATIONS

- As we have figured out from the above analysis about the districts consuming highest and lowest in the state, we can plan the inventory accordingly and will also be able to make the better forecast from future.
- Also, as we have categorised the urban and rural zones, it would be better for monitoring, administration and mobilisation.
- As we have seen that there is a significantly higher consumption in rural areas than the urban population, priorities can be set according to the policy.