VIRGINIA COMMONWEALTH UNIVERSITY

STATISTICAL ANALYSIS & MODELING

A1a: CONSUMPTION PATTERN OF HARYANA USING PYTHON AND R

RIDDHI RUNGTA
V01107488

Date of Submission: 16/06/2024

CONTENTS

Analyzing Consumption in the State of Haryana Using R

INTRODUCTION

The focus of this study is on the state of Haryana, from the NSSO data, to find the top and bottom three consuming districts of Haryana. This dataset provides comprehensive information on household consumption patterns in many districts of Haryana, India. The data includes details on the consumption of various food items such as rice, wheat, chicken, pulses, and other essential commodities, categorized by regions, sectors, and meal frequency. The dataset serves as a critical source for understanding the dietary habits and nutritional intake of households in this region, which is pivotal for formulating targeted interventions and policies.
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 as district-wide variations. The dataset has been imported into R, a powerful statistical programming language 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, social welfare organizations 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 Haryana’s consumption patterns from NSSO data holds significant implications for businesses and policymakers.

1) Policy Makers and Government Agencies can identify nutritional deficiencies and excesses in specific regions or sectors by examining the consumption patterns. This insight aids in designing effective food security programs, public health interventions, and targeted nutritional assistance to improve the overall health of the population.
2) Agricultural and Food Supply chain managers understands the demand for different food commodities enabling better planning and distribution within the supply chain. This helps in reducing wastage, ensuring timely delivery, and maintaining the balance between supply and demand, ultimately leading to more efficient food distribution networks.
3) NGOs focused on nutrition and food security can use this data to tailor their programs to the actual needs of the community. They can identify areas with higher nutritional gaps and direct their resources more effectively.
4) Companies in the food and beverage sector can leverage this information to identify market opportunities, develop new products, and create targeted marketing strategies that cater to the specific dietary preferences and needs of different consumer segments.

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:

```
In [10]: HR_new.isnull().sum().sort_values(ascending = False)

Out[10]: Meals_At_Home       14
         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(HR06new)))
         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 selected variables that is the subset, after sorting the data for the state of Haryana, in both python and R, we can see that only column 'Meals_At_Home has 14 missing variables. The presence of missing values in this column suggests that there might be some households for which this information was not recorded or reported. 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 [12]: HR_clean.loc[:, 'Meals_At_Home'] = HR_clean['Meals_At_Home'].fillna(HR_new['Meals_At_Home'].mean())

In [13]: HR_clean.isnull().any()

Out[13]: 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 is a snippet from python programming. It 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 and that's why the outcome is false.

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

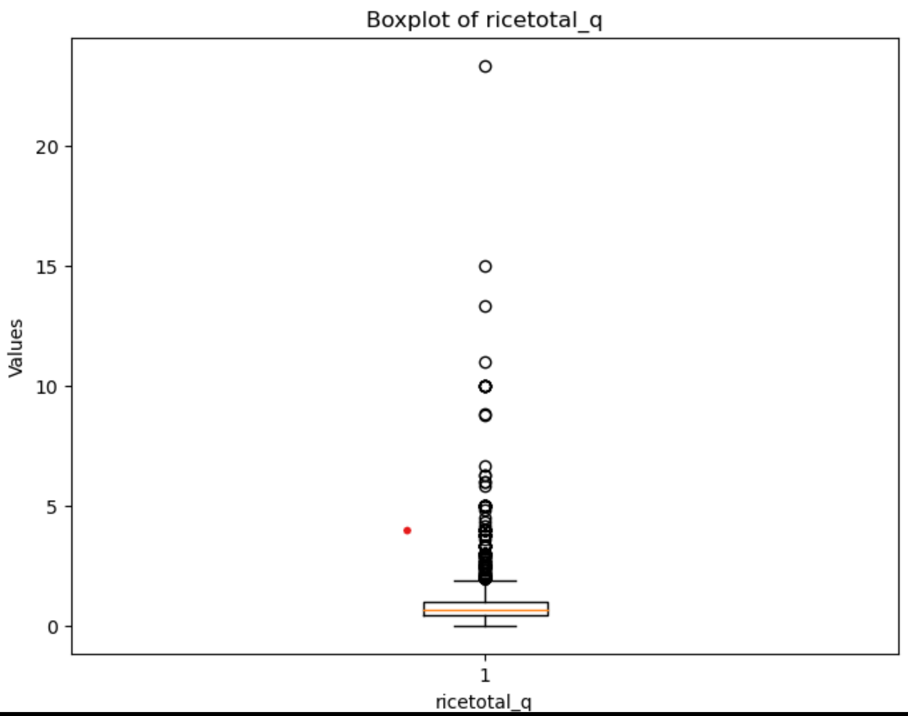
For outlier detection, I have used boxplots as its a standardized way of displaying the distribution of data based on a five-number summary: minimum, first quartile (Q1), median, third quartile (Q3), and maximum. It also highlights potential outliers in the dataset.

#Checking for outliers

Plotting the boxplot to visualize outliers.

Code and Result:

```
In [15]: 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()
```



<u>Interpretation</u>: 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 theaccuracy and reliability of results in data-driven decision-making processes.

In the above analysis using R, we could see that there were 260 observations having outlier when we removed outlier and set quartiles. The outliers can be removed using the following code.

#Setting quartiles and removing outliers

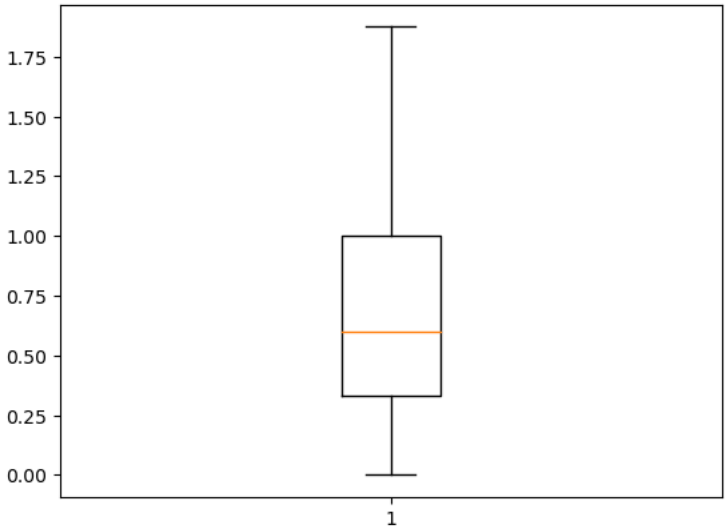
Code and results:

Setting quartile ranges to remove outliers

```
rice1 = HR_clean['ricetotal_q'].quantile(0.25)
rice2 = HR_clean['ricetotal_q'].quantile(0.75)
iqr_rice = rice2-rice1
up_limit = rice2 + 1.5*iqr_rice
low_limit = rice1 - 1.5*iqr_rice

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

Post this, when we tried making a box plot, we got a figure that had no outliers such as given below.



Interpretation: The following observations were made: -

1) The median remains the same, representing the central value of the cleaned data.
2) 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.
3) The whiskers extend to the new minimum and maximum values within the revised 1.5 * IQR range. There should be fewer or no points outside the whiskers, indicating that extreme values have been removed.

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 assigned 1 and 2 respectively. This is done by runningthe following code.

In the below code, we took a subset of Haryana district and tried to map the names and sectors instead of the code numbers.

Code and Result:

```
district_mapping <- c("13" = "Bhiwani", "19" = "Faridabad", "12" = "Hisar", "15" = "Jhajjar", "01" = "Panchkula", "20" = "Mewa
t")
> sector_mapping <- c("2" = "URBAN", "1" = "RURAL")
```

```
>
> HR06new$District <- as.character(HR06new$District)
> HR06new$Sector <- as.character(HR06new$Sector)
> HR06new$District <- ifelse(HR06new$District %in% names(district_mapping), district_mapping[HR06new$District], HR06new$District)
> HR06new$Sector <- ifelse(HR06new$Sector %in% names(sector_mapping), sector_mapping[HR06new$Sector], HR06new$Sector)

>
```

Result:

	state_1	District	Region	Sector	State_Region	Meals_At_Home	ricepds_v	Wheatpds_q	chicke
	All	All	All	All	All	All	All	All	All
6	HR	Mewat	1	RURAL	61	60	0	0.0000000	
7	HR	Mewat	1	RURAL	61	56	0	0.0000000	
8	HR	Mewat	1	RURAL	61	60	0	5.8333333	
9	HR	Mewat	1	RURAL	61	60	0	8.7500000	
0	HR	Mewat	1	RURAL	61	60	0	0.0000000	
1	HR	Mewat	1	RURAL	61	90	0	0.0000000	
2	HR	Mewat	1	RURAL	61	90	0	5.0000000	
3	HR	Mewat	1	RURAL	61	60	0	0.0000000	
6	HR	Jhajjar	1	URBAN	61	90	0	0.0000000	
7	HR	Jhajjar	1	URBAN	61	90	0	0.0000000	
0	HR	Jhajjar	1	URBAN	61	90	0	0.0000000	
1	HR	Jhajjar	1	URBAN	61	90	0	0.0000000	
2	HR	Jhajjar	1	URBAN	61	90	0	0.0000000	
0	HR	Jhajjar	1	URBAN	61	84	0	0.0000000	
1	HR	Jhajjar	1	URBAN	61	90	0	0.0000000	
2	HR	Jhajjar	1	URBAN	61	90	0	0.0000000	
4	HR	Jhajjar	1	URBAN	61	60	0	5.0000000	
5	HR	Jhajjar	1	URBAN	61	90	0	0.0000000	

	state_1	District	Sector	Region	State_Region	ricetotal_q	wheattotal_q	moong_q	Milktotal_q	chicken_q	bread_q	foodtotal_q	Beveragestotal_v
35704	HR	Faridabad	RURAL	1	61	1.250000	4.000000	0.125000	0	0.0	0.062500	40.925704	50.000000
35705	HR	Faridabad	RURAL	1	61	0.500000	5.833333	0.166667	0	0.0	0.333333	27.441958	50.000000
35706	HR	Faridabad	RURAL	1	61	0.833333	6.000000	0.083333	0	0.0	0.083333	31.767038	33.333333
35707	HR	Faridabad	RURAL	1	61	1.000000	5.000000	0.250000	0	0.0	0.125000	37.100600	50.000000
35708	HR	Faridabad	RURAL	1	61	0.600000	3.000000	0.100000	0	0.0	0.000000	26.894340	40.000000

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.
As we can see in the district and the sector column in the above table.

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:

```
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.
```

```
In [29]: total_consumption_by_districtcode.sort_values(ascending=False).head(3)
Out[29]: District
         19    16829.789693
         9     14855.129103
         11    10702.037831
         Name: total_consumption, dtype: float64

In [35]: HR_clean.loc[:,"District"] = HR_clean.loc[:,"District"].replace({13:"Bhiwani", 9:"Jind", 11:"Sirsa"})

In [36]: total_consumption_by_districtname=HR_clean.groupby('District')['total_consumption'].sum()

In [37]: total_consumption_by_districtname.sort_values(ascending=False).head(3)
Out[37]: District
         Faridabad    16829.789693
         Jind         14855.129103
         Sirsa        10702.037831
         Name: total_consumption, dtype: float64
```

Interpretation:

The above first figure shows that there are some districts that show significantly higher consumption compared to others, indicating potentially larger populations, higher demand, or greater accessibility to the commodity.
District 13: The highest consuming district with a total consumption of 1488 units.
District 19: The second highest consuming district with a total consumption of 1461 units.
District 12: The third highest consuming district with a total consumption of 1369 units.

The analysis also highlights the bottom three districts with the lowest total consumption:
District 15: This district has a total consumption of 612 units.
District 1: This district has a total consumption of 343 units.
District 20: The district with the lowest consumption, totaling 318 units.

Region Consumption Summary - The data is aggregated by region, summarizing total consumption across two regions:
Region 1: This region has a total consumption of 10,792 units.
Region 2: This region has a total consumption of 8,264 units.

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 mean consumption between urban and rural.

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

```
# Test for differences in mean consumption between urban and rural
rural <- HR06new %>%
  filter(Sector == "RURAL") %>%
  select(total_consumption)

urban <- HR06new %>%
  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:

```
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
```

In python, the code was following:-

```
z_statistic, p_value = stests.ztest(cons_rural, cons_urban)
# Print the z-score and p-value
print("Z-Score:", z_statistic)
print("P-Value:", p_value)
```

Result:

Z-Score: 7.091254132951441
P-Value: 1.329020295022693e-12

Interpretation:

The Z-test was conducted to compare the mean consumption between rural and urban areas. The results are as follows:

- **Z-Score**: 7.091254132951441
- **P-Value**: 1.329020295022693e-12

A Z-score of 7.091254132951441 is very high, indicating that the difference between the two means is significantly larger than what would be expected by random chance. Given the high Z-score and the extremely low P-value, we can reject the null hypothesis that there is no difference in mean consumption between rural and urban areas.

There is a significant difference in consumption patterns between rural and urban areas. This insight can inform targeted policy interventions, marketing strategies, and resource allocation to address the specific needs and behaviors of rural and urban populations differently.