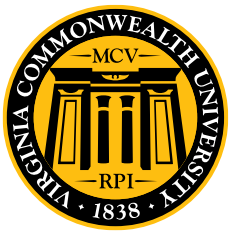


VIRGINIA COMMONWEALTH UNIVERSITY

Statistical analysis and modelling (SCMA 632)

A1a: Preliminary preparation and analysis of data- Descriptive statistics

SATHWIK NAG CHANNAGIRI VENKATESH

V01107764

Date of Submission: 16-06-2024

CONTENTS

INTRODUCTION

This analysis aims to process and scrutinize the NSSO68 dataset for **Goa**, focusing on consumption-related information across its districts. Our primary objectives are to identify missing values and outliers, standardize district and sector names, summarize consumption data regionally and district-wise, and perform statistical tests to assess the significance of mean differences. By methodically cleaning and analyzing the dataset imported into R, a versatile statistical programming language, we aim to reveal the top and bottom three consuming districts in Goa. This report details the steps to achieve these goals, presenting the findings and recommendations that can guide policymakers and stakeholders in fostering targeted interventions and promoting equitable development throughout the state.

Objectives:

- Check if there are any missing values in the data, identify them, and if there are, replace them with the mean of the variable
- Check for outliers, describe your test's outcome, and make suitable amendments.
- Rename the districts and sectors, viz., rural and urban.
- Summarize the critical variables in the data set region-wise and district-wise and indicate the top and bottom three consumption districts.
- Test whether the differences in the means are significant or not.

Business Significance:

The focus of this study on Goa's consumption patterns from NSSO data holds considerable implications for businesses and policymakers. By analyzing the consumption data across Goa's two districts, the study provides crucial insights into regional consumption behaviours, which are valuable for market entry strategies, resource allocation, and supply chain optimization.

This analysis aids in fostering economic growth and ensuring balanced resource distribution across the state, ultimately enhancing Goa's overall economic landscape.

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.

Result:

All missing values in the dataset were identified and replaced with the mean of their respective variables to maintain data integrity and consistency.

```
Missing Values in Subset:
> print(colSums(is.na(ganew)))
           state_1           District             Region             Sector       State_Region
                 0                  0                  0                  0                  0
     Meals_At_Home          ricepds_v         Wheatpds_q          chicken_q          pulsep_q
                 2                  0                  0                  0                  0
         wheatos_q        fishprawn_q No_of_Meals_per_day
                 0                  0                  0
```

```
Missing Values After Imputation:
> print(colSums(is.na(ganew)))
           state_1           District             Region             Sector       State_Region
                 0                  0                  0                  0                  0
     Meals_At_Home          ricepds_v         Wheatpds_q          chicken_q          pulsep_q
                 0                  0                  0                  0                  0
         wheatos_q        fishprawn_q No_of_Meals_per_day
                 0                  0                  0
```

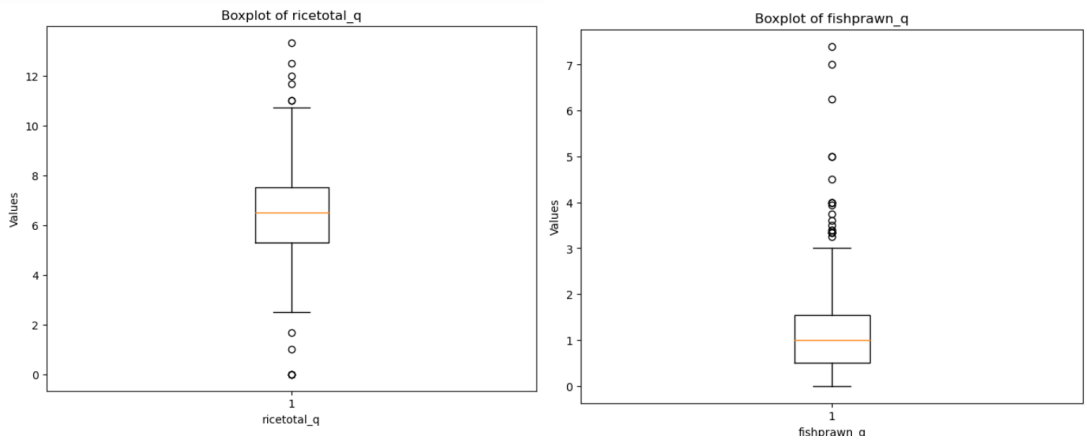
Interpretation:

The above dataset shows complete data for most columns, including geographical and sectoral information and quantities of rice, wheat, chicken, pulses, wheat output, fish prawn, and number of meals per day. However, the "Meals_At_Home" column has 2 missing values, indicating incomplete records for this variable. This missing data needs addressing, either through imputation or exclusion, before further analysis. The robust dataset allows for a comprehensive analysis of food consumption patterns across different regions and sectors.

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

Result:

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.



Interpretation:

The "ricetotal_q" and “fishprawn_ q" graphs’ boxplots displayed above show the outliers. The distribution is displayed in the above boxplots, where several points are above the upper whisker, designating values more than 20 as outliers. 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. 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. These plots present high-value outliers, indicating the need for cautious handling in additional investigation.

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

Result:

Data Editor

File Edit Help

	row.names	state_1	District	Region	Sector	State_Region	Meals_At_Home	ricepds_v	Wheatpds_q	chicken_q	pulsep_q	wheatos_q	fishprawn_q	No_of_Meals_per_day	total_consumption	var16
1	2	GOA	South Goa	1	URBAN	301	60	0	0	1	0	1	1.5	2	3.5	
2	5	GOA	South Goa	1	URBAN	301	60	22.5	0.5	0	0	1.25	0.5	2	24.75	
3	6	GOA	South Goa	1	URBAN	301	60	0	0	0.5	0	0.75	0.75	2	2	
4	9	GOA	North Goa	1	URBAN	301	60	0	0	0.25	0	2.5	1.25	2	4	
5	10	GOA	North Goa	1	URBAN	301	60	0	0	0	0	3	1.2	2	4.2	
6	11	GOA	North Goa	1	URBAN	301	60	30	0	0	0	1.666667	1.666667	2	33.33333	
7	12	GOA	North Goa	1	URBAN	301	60	22.5	0	0.25	0	1.25	0.75	2	24.75	
8	14	GOA	North Goa	1	URBAN	301	60	0	0	0.6666667	0	1.666667	2.666667	2	5	
9	15	GOA	North Goa	1	URBAN	301	60	18	0.4	0.6	0	0.8	0.6	2	20.4	
10	16	GOA	North Goa	1	URBAN	301	60	18	0.4	0	0	1	0.2	2	19.6	
11	18	GOA	North Goa	1	URBAN	301	60	15	1	0.3333333	0	0	2.5	2	18.83333	
12	22	GOA	North Goa	1	URBAN	301	60	0	0	0.5	0	0.5	1.375	2	2.375	
13	24	GOA	North Goa	1	URBAN	301	60	0	0	0.2	0	0.9	0.68	2	1.78	
14	27	GOA	South Goa	1	URBAN	301	60	30	0	1	0	1.666667	1.333333	2	34	
15	31	GOA	South Goa	1	URBAN	301	60	18	0	0	0	0.4	0	2	18.4	
16	36	GOA	North Goa	1	URBAN	301	60	33.5	0.5	0	0	0.75	0.875	2	35.625	
17	37	GOA	North Goa	1	URBAN	301	60	0	0.75	0.125	0	0.5	0.95	2	2.325	
18	40	GOA	North Goa	1	URBAN	301	60	0	0	1	0	0	1.5	2	2.5	
19	43	GOA	South Goa	1	URBAN	301	60	0	0	0.3	0	1.4	1.78	2	3.48	
20	44	GOA	South Goa	1	URBAN	301	60	18	0	0.4	0	1	1.8	2	21.2	
21	46	GOA	South Goa	1	URBAN	301	60	0	0	0.2	0	2.4	2.44	2	5.04	
22	47	GOA	South Goa	1	URBAN	301	60	15	0.5	0.3333333	0	0.5	1.416667	2	17.75	
23	48	GOA	South Goa	1	URBAN	301	60	18	0	1	0	1	2.1	2	22.1	
24	49	GOA	North Goa	1	URBAN	301	60	0	0	0	0	0	0	2	0	
25	51	GOA	North Goa	1	URBAN	301	60	0	0	0	0	0	0	2	0	
26	53	GOA	North Goa	1	URBAN	301	60	0	0	0	0	1.666667	0.5	2	2.166667	
27	55	GOA	North Goa	1	URBAN	301	60	0	0	0	0	1	0.2	2	1.2	
28	56	GOA	North Goa	1	URBAN	301	60	0	0	0	0	0	0	2	0	
29	57	GOA	South Goa	1	URBAN	301	60	0	0	0	0	1.25	1.15	2	2.4	
30	58	GOA	South Goa	1	URBAN	301	60	0	0	0.5	0	2	1	2	3.5	
31	59	GOA	South Goa	1	URBAN	301	60	0	0	0.25	0	1.25	1.125	2	2.625	
32	60	GOA	South Goa	1	URBAN	301	60	0	0	0.25	0	2	0.75	2	3	
33	63	GOA	South Goa	1	URBAN	301	60	0	0	0.4285714	0	2.857143	1.071429	2	4.357143	
34	64	GOA	South Goa	1	URBAN	301	60	0	0	0	0	3.75	0	2	3.75	
35	66	GOA	South Goa	1	URBAN	301	60	0	0	0	0	1.5	0	2	1.5	
36	67	GOA	South Goa	1	URBAN	301	60	30	1	0	0	0.3333333	0.6	2	31.93333	
37	68	GOA	South Goa	1	URBAN	301	60	22.5	0.75	0.125	0	0.75	1	2	25.125	

Interpretation:

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. The 'District' and 'Sector' columns were converted to character type. Using the mappings (`district_mapping` and `sector_mapping`).

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

Result:

```
Top 2 Consuming Districts:
> print(head(district_summary, 2))
# A tibble: 2 × 2
  District total
     <int> <dbl>
1        1 2258.
2        2 1964.
> cat("Bottom 2 Consuming Districts:\n")
Bottom 2 Consuming Districts:
> print(tail(district_summary, 2))
# A tibble: 2 × 2
  District total
     <int> <dbl>
1        1 2258.
2        2 1964.
```

Interpretation:

Since Goa has only two districts: North Goa and South Goa, the top consuming district is North Goa, and the bottom consuming district is South Goa.

E. Test whether the differences in the means are significant or not.

Result:

```
> z_test_result$statistic
       z
2.10968
> z_test_result$p.value
[1] 0.03488593
> z_test_result$method
[1] "Two-sample z-Test"
```

```
> summary(z_test_result)
            Length Class  Mode
statistic   1      -none- numeric
p.value     1      -none- numeric
conf.int    2      -none- numeric
estimate    2      -none- numeric
null.value  1      -none- numeric
alternative 1      -none- character
method      1      -none- character
data.name   1      -none- character
```

Interpretation:

P value is < 0.05, i.e., 0.03489; therefore, we reject the null hypothesis. There is a difference between urban and rural mean consumption. The mean consumption in Rural areas is 17.1142664996842, and in Urban areas, it's 16.4350905635443. As a result, the mean consumption in rural is higher than in urban, 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.

Summary of interpretations:

- **Data Completeness**: Most columns have complete data, but the "Meals_At_Home" column has two missing values. These needs addressing either through imputation or exclusion before further analysis.
- **Data Mapping and Transformation**: Districts and sectors were mapped to their respective names using district_mapping and sector_mapping, and columns were converted to character type for clarity.
- **Outlier Detection**: Boxplots for "ricetotal_q" and “fishprawn_q" revealed several outliers with values above 20. These outliers can distort statistical analyses and need careful handling.
- **Urban vs. Rural Consumption**: Statistical tests indicate a significant difference in mean consumption between urban and rural areas. The mean consumption in rural areas (17.11) is higher than in urban areas (16.43), suggesting differences in resource availability or consumption habits.
- **District-wise Consumption**: With only two districts in Goa, North Goa is identified as the top consuming district, and South Goa as the bottom consuming district.

Recommendations

Based on the findings, the following recommendations are proposed:

- **Targeted Policy Interventions**: Implement focused economic and social programs in South Goa to enhance consumption and overall economic conditions.
- **Sector-specific Programs**: Develop and implement programs to improve consumption in rural areas, addressing the significant difference in consumption patterns between rural and urban sectors.
- **Regular Monitoring**: Continuously monitor and update consumption data to track pattern changes, ensuring timely intervention and policy adjustments as needed.
- **Resource Allocation and Forecasting**: Use the insights on district and sector-wise consumption to optimize inventory planning and make better forecasts for future demand.
- **Zone-based Monitoring and Administration**: Utilize the categorization of urban and rural zones to improve monitoring, administration, and resource mobilization, ensuring targeted and efficient delivery of services.
- **Further Research**: Conduct additional studies to understand the underlying causes of higher rural consumption and address any disparities in resource availability and consumption habits.