****

**VIRGINIA COMMONWEALTH UNIVERSITY**

**Statistical analysis and modelling (SCMA 632)**

**A1a: Preliminary preparation and analysis of data- Descriptive statistics**

**M K NITHEESH**

**V01107616**

**Date of Submission: 23-06-2024**

**CONTENTS**

|  |  |  |
| --- | --- | --- |
| **Sl. No.** | **Title** | **Page No.** |
| **1.** | Introduction | 1 |
| **2.** | Objectives | **1** |
| **3.** | Business Significance | **1** |
| **4.** | Results and analysis | **2-7** |
| **5.** | Codes | **8-13** |
| **6.** | References |  |

**Analyzing Consumption in the State of West Bengal Using R**

**Introduction**

The focus of this study is on the state of West Bengal, from the NSSO data, to find the top and bottom three consuming districts of West Bengal. In the process, we manipulate and clean the dataset to get the required data to analyse. To facilitate this analysis, we have gathered a dataset containing consumption-related information, including data on rural and urban sectors, as well as district-wise variations. The dataset has been imported into R, a powerful statistical programming language renowned for its versatility in handling and analysing 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, 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.

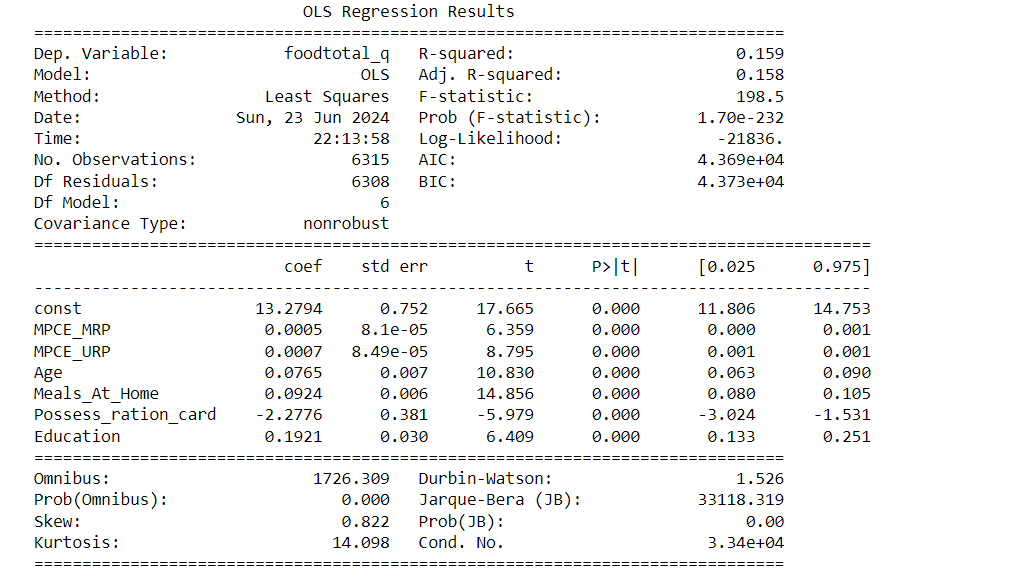
d) Summarize the critical variables in the data set region wise and district wise and indicate the top three districts and the bottom three districts of consumption.

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

**BUSINESS SIGNIFICANCE**

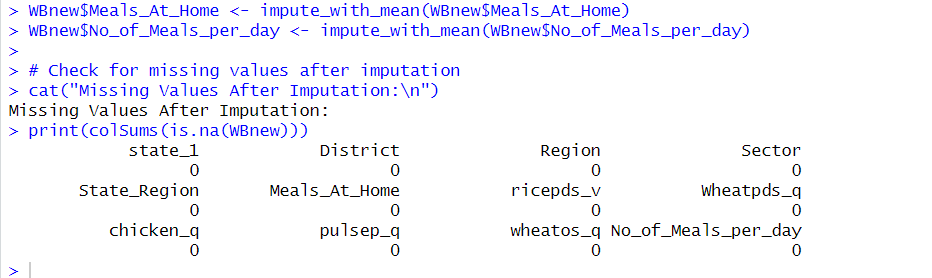
The focus of this study on West Bengals consumption patterns from NSSO data holds significant implications for businesses and policymakers. By identifying the top and bottom three consuming districts, the study provides valuable insights for market entry, resource allocation, supply chain optimization, and targeted interventions. Through data cleaning, outlier detection, and significance testing, the findings facilitate informed decision-making, fostering equitable development and promoting West Bengals economic growth.

**A) RESULTS AND INTERPRETATION**

****

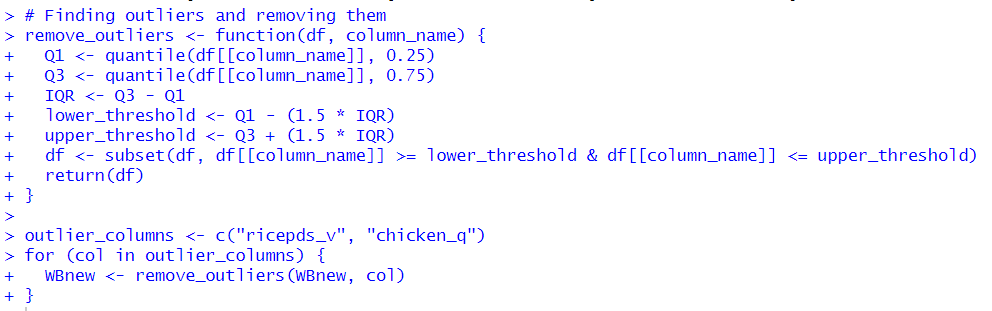
**Interpretation**: he models shows a moderate relationship (15.9% explained) between various factors and food expenditure. People who are older, eat more at home, and are more educated tend to spend more on food. Those with ration cards spend less. The model has limitations, so interpret cautiously.

**#Imputing the values, i.e. replacing the missing values with mean.**

****

Interpretation: The above code 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.

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

****

****

**#NSSO data regression**

**library(dplyr)**

**setwd("C:\\Users\\nithe\\Downloads")**

**# Loading the dataset**

**data = read.csv("NSSO68.csv")**

**unique(data$state\_1)**

**# Subset data to state assigned**

**subset\_data <- data %>%**

**filter(state\_1 == 'WB') %>%**

**select(foodtotal\_q, MPCE\_MRP, MPCE\_URP,Age,Meals\_At\_Home,Possess\_ration\_card,Education, No\_of\_Meals\_per\_day)**

**print(subset\_data)**

**sum(is.na(subset\_data$MPCE\_MRP))**

**sum(is.na(subset\_data$MPCE\_URP))**

**sum(is.na(subset\_data$Age))**

**sum(is.na(subset\_data$Possess\_ration\_card))**

**sum(is.na(data$Education))**

**impute\_with\_mean = function(data, columns) {**

**data %>%**

**mutate(across(all\_of(columns), ~ ifelse(is.na(.), mean(., na.rm = TRUE), .)))**

**}**

**# Columns to impute**

**columns\_to\_impute = c("Education")**

**# Impute missing values with mean**

**data = impute\_with\_mean(data, columns\_to\_impute)**

**sum(is.na(data$Education))**

**# Fit the regression model**

**model = lm(foodtotal\_q~ MPCE\_MRP+MPCE\_URP+Age+Meals\_At\_Home+Possess\_ration\_card+Education, data = subset\_data)**

**# Print the regression results**

**print(summary(model))**

**library(car)**

**# Checking for multicollinearity using Variance Inflation Factor (VIF)**

**vif(model)**

**# Extracting the coefficients from the model**

**coefficients <- coef(model)**

**# Construct the equation**

**equation <- paste0("y = ", round(coefficients[1], 2))**

**for (i in 2:length(coefficients)) {**

**equation <- paste0(equation, " + ", round(coefficients[i], 6), "\*x", i-1)**

**}**

**# Print the equation**

**print(equation)**