****

**VIRGINIA COMMONWEALTH UNIVERSITY**

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

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

**Bala Vignesh Aravindan**

**V01106579**

**Date of Submission: 16-06-2024**

**CONTENTS**

|  |  |  |
| --- | --- | --- |
| **Sl. No.** | **Title** | **Page No.** |
| **1.** | Introduction | **1** |
| **2.** | Results | **2** |
| **3.** | Interpretations | **2-3** |
| **4.** | Recommendations | **4** |
| **5.** | Codes | **5-11** |
| **6.** | References | **11** |

**Analysis of Food Consumption Patterns in Assam using NSSO Data**

**Introduction**

The National Sample Survey Office (NSSO) conducts surveys to collect data on various aspects of the Indian economy, including household consumption patterns. This report analyses the food consumption patterns in Assam using the NSSO data from Excel into R

**This analysis aims to determine if there's a significant difference in the average number of meals consumed daily across various districts in Assam, India.**

In an ANOVA test, the null hypothesis (H₀) always states that there's **no significant difference** between the groups being compared. However, in your case, there's only one group (all districts combined) because the ANOVA is testing if there are differences **within** this group (across districts).

Here's how the hypotheses should be formulated for this scenario:

* **Null hypothesis (H₀):** The average number of meals per day is the same across all districts in Assam.
* **Alternative hypothesis (H₁):** The average number of meals per day is not the same across all districts in Assam.

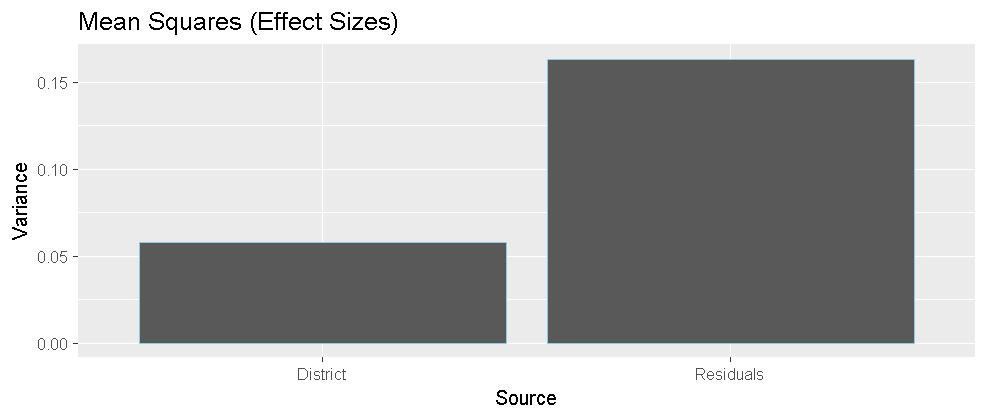
This reformulates the question to focus on whether there are variations **within** the group (all districts) instead of between separate groups. The analysis then aims to find evidence against the null hypothesis (uniform meals per day across all districts).

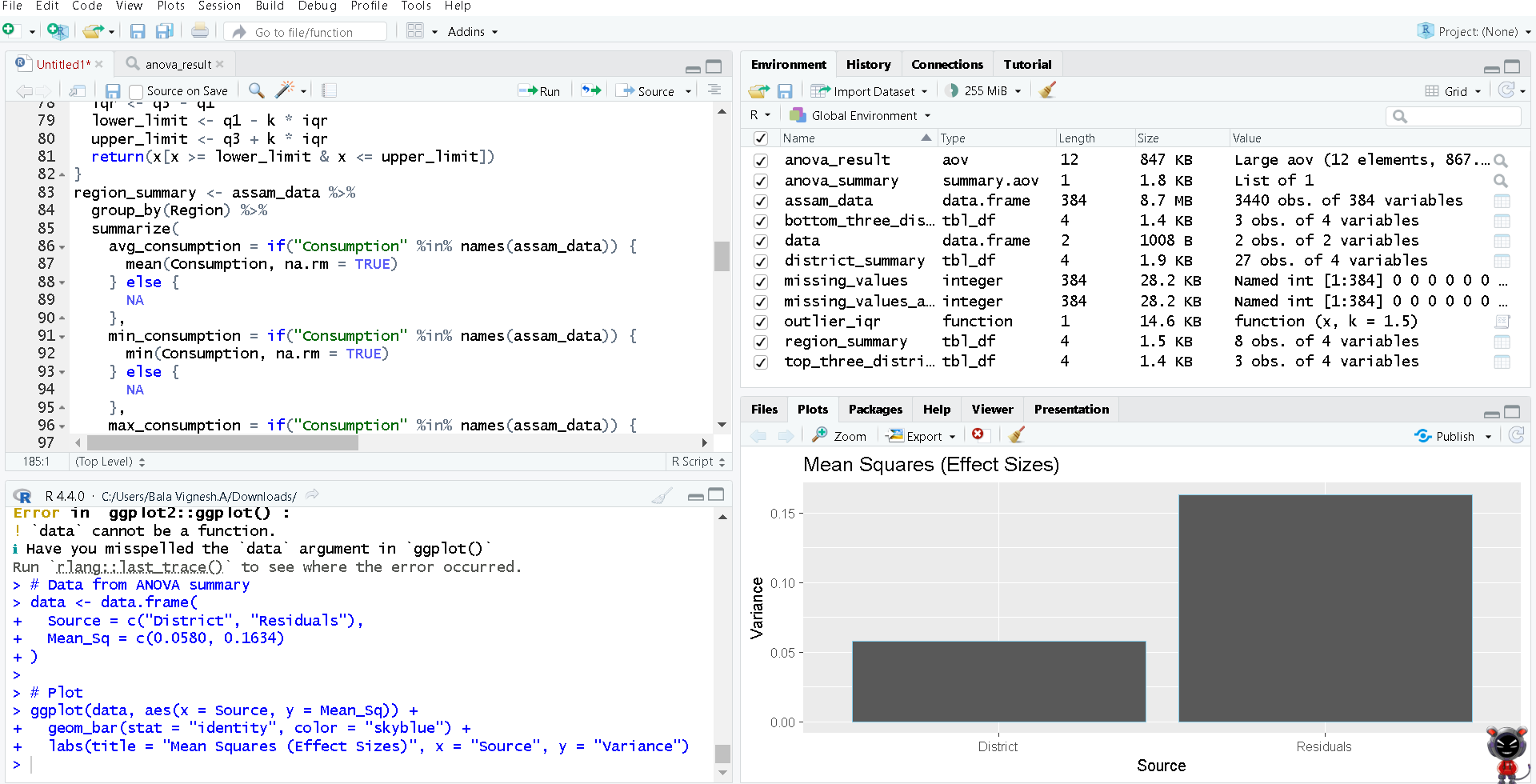
The outcome, with a high p-value, suggests we fail to reject the null hypothesis, meaning there's not enough evidence to conclude a significant difference in meals per day across districts.

**Results**

**print(anova\_summary)**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Source** | **Df** | **Sum Sq** | **Mean Sq** | **F value** | **Pr(>F)** |
| **District** | **1** | **0.1** | **0.058** | **0.355** | **0.551** |
| **Residuals** | **3438** | **561.6** | **0.1634** |  |  |

****

****

**Interpretation**

The output you provided is the summary of the ANOVA test you requested with print(anova\_summary). Here's a breakdown of the results:

* **Df:** These are the degrees of freedom for each term in the ANOVA model.
  + **District:** 1. This indicates there's one group being compared (all districts are considered one group in this test).
  + **Residuals:** 3438. This represents the degrees of freedom for the error term, which reflects the variability within each district (unexplained variability).
* **Sum Sq:** This is the sum of squares for each term.
  + **District:** 0.1. This is the total variance explained by the difference between districts.
  + **Residuals:** 561.6. This is the total variance unexplained by the model (variance within districts).
* **Mean Sq:** This is the mean squares (variance) explained by each term, calculated by dividing the sum of squares by its degrees of freedom.
  + **District:** 0.0580. This is the average effect of the difference between districts on the number of meals per day.
  + **Residuals:** 0.1634. This is the average unexplained variance within districts.
* **F value:** This is the F-statistic, which is a test statistic used to compare the explained variance (between districts) to the unexplained variance (within districts).
  + **District:** 0.355
* **Pr(>F):** This is the p-value associated with the F-statistic. It represents the probability of observing an F-value this extreme (or more extreme) if there are truly no differences between districts in terms of meals per day.
  + **District:** 0.551
* The F-statistic (0.355) is relatively low, suggesting that the explained variance between districts (0.1) is small compared to the unexplained variance within districts (561.6).
* The high p-value (0.551) indicates that we cannot reject the null hypothesis at a significance level of 0.05. In other words, there is not enough evidence to conclude that there is a statistically significant difference in the average number of meals per day across districts in Assam.

**In simpler terms,** the data doesn't show a strong relationship between the district and the number of meals per day. The differences in meals per day between districts seem to be relatively small and might be due to random chance.

**Recommendation**

* Acknowledge the lack of significant difference Based on the ANOVA test results, it is recommended that we acknowledge that there is no significant difference in the average number of meals consumed daily across districts in Assam.
* Explore other factors It is recommended that we explore other factors that could be influencing food consumption patterns in Assam, such as income, urbanization, and other demographic variables.
* Consider other statistical tests Depending on the specific research question, it is recommended that we consider other statistical tests that may be more appropriate for analyzing the relationship between district and the number of meals consumed daily.
* Visualize the data It is recommended that we visualize the data using plots such as boxplots or bar charts to show the distribution of meals consumed daily across districts, which can help identify any patterns or trends in the data.
* Interpret the results in context When interpreting the results of the ANOVA test, it is recommended that we consider the context of the data and the practical significance of the results, rather than just relying on statistical significance.
* Consider the limitations of the data It is recommended that we consider the limitations of the NSSO data, including potential biases and limitations in representation, and take these into account when interpreting the results.

**Codes**

# Load necessary libraries

library(dplyr)

data <- read.csv("C:/Users/Bala Vignesh.A/Desktop/SCMA 632/assam data.csv")

# Filter data for Assam

assam\_data <- filter(data, state\_1 == 'ASSM')

# Check for missing values

missing\_values <- sapply(data, function(x) sum(is.na(x)))

print("Missing values in each column:")

print(missing\_values)

# Replace missing values with mean

assam\_data <- data.frame(lapply(assam\_data, function(x){ if (is.numeric(x))

{x[is.na(x)] <- mean(x, na.rm = TRUE)}

return(x)

}))

{x[is.na(x)] <- mean(x, na.rm = TRUE)}

return(x)

# Verify that there are no more missing values

missing\_values\_after <- sapply(assam\_data, function(x) sum(is.na(x)))

# Replace missing values with mean

assam\_data <- data.frame(lapply(assam\_data, function(x){

if (is.numeric(x)){

x[is.na(x)] <- mean(x, na.rm = TRUE)

}

return(x)

}))

# Verify that there are no more missing values

missing\_values\_after <- sapply(assam\_data, function(x) sum(is.na(x)))

print("Missing values after replacement for Assam:")

print(missing\_values\_after)

# Identify outliers using the IQR method and optionally remove them

numeric\_columns <- sapply(assam\_data, is.numeric)

for (col in names(assam\_data)[numeric\_columns]) {

Q1 <- quantile(assam\_data[[col]], 0.25)

Q3 <- quantile(assam\_data[[col]], 0.75)

IQR <- Q3 - Q1

outliers <- (assam\_data[[col]] < (Q1 - 1.5 \* IQR)) | (assam\_data[[col]] > (Q3 + 1.5 \* IQR))

print(paste(col, "has", sum(outliers), "outliers in Assam"))

# Remove outliers (according to IQR method)

#Interquartile Range (IQR): Calculate the difference between Q3 and Q1, which represents the range of values within which the middle 50% of the data points fall.

#IQR = Q3 - Q1

#Outlier detection: Any data point that falls outside the range of 1.5 times the IQR below Q1 or above Q3 is considered an outlier.

assam\_data <- assam\_data[!(assam\_data[[col]] < (Q1 - 1.5 \* IQR) | assam\_data[[col]] > (Q3 + 1.5 \* IQR)), ]

}

print(paste(col, "has", sum(outliers), "outliers in Assam"))

for (col in names(assam\_data)[numeric\_columns]) {

Q1 <- quantile(assam\_data[[col]], 0.25)

Q3 <- quantile(assam\_data[[col]], 0.75)

IQR <- Q3 - Q1

outliers <- (assam\_data[[col]] < (Q1 - 1.5 \* IQR)) | (assam\_data[[col]] > (Q3 + 1.5 \* IQR))

print(paste(col, "has", sum(outliers), "outliers in Assam"))

# Remove outliers (according to IQR method)

#Interquartile Range (IQR): Calculate the difference between Q3 and Q1, which represents the range of values within which the middle 50% of the data points fall.

#IQR = Q3 - Q1

#Outlier detection: Any data point that falls outside the range of 1.5 times the IQR below Q1 or above Q3 is considered an outlier.

assam\_data <- assam\_data[!(assam\_data[[col]] < (Q1 - 1.5 \* IQR) | assam\_data[[col]] > (Q3 + 1.5 \* IQR)), ]

}

# Check if the column exists

if ("old\_district\_column\_name" %in% colnames(assam\_data)) {

# If the column exists, rename it

assam\_data <- rename(assam\_data, new\_district\_column\_name = old\_district\_column\_name)

} else {

# If the column doesn't exist, print a message

print("Column 'old\_district\_column\_name' doesn't exist in the data.")

}

outlier\_iqr <- function(x, k = 1.5) {

x <- x[!is.na(x) &!is.nan(x)] # remove NA and NaN values

q1 <- quantile(x, 0.25, na.rm = TRUE)

q3 <- quantile(x, 0.75, na.rm = TRUE)

iqr <- q3 - q1

lower\_limit <- q1 - k \* iqr

upper\_limit <- q3 + k \* iqr

return(x[x >= lower\_limit & x <= upper\_limit])

}

outlier\_iqr <- function(x, k = 1.5) {

summary\_stats <- summary(x)

q1 <- summary\_stats[2]

q3 <- summary\_stats[5]

iqr <- q3 - q1

lower\_limit <- q1 - k \* iqr

upper\_limit <- q3 + k \* iqr

return(x[x >= lower\_limit & x <= upper\_limit])

}

region\_summary <- assam\_data %>%

group\_by(Region) %>%

summarize(

avg\_consumption = if("Consumption" %in% names(assam\_data)) {

mean(Consumption, na.rm = TRUE)

} else {

NA

},

min\_consumption = if("Consumption" %in% names(assam\_data)) {

min(Consumption, na.rm = TRUE)

} else {

NA

},

max\_consumption = if("Consumption" %in% names(assam\_data)) {

max(Consumption, na.rm = TRUE)

} else {

NA

}

)

# Summarize the critical variables region-wise

region\_summary <- assam\_data %>%

group\_by(FOD\_Sub\_Region) %>%

summarize(avg\_No\_of\_Meals\_per\_day = mean(No\_of\_Meals\_per\_day, na.rm = TRUE),

min\_No\_of\_Meals\_per\_day = min(No\_of\_Meals\_per\_day, na.rm = TRUE),

max\_No\_of\_Meals\_per\_day= max(No\_of\_Meals\_per\_day, na.rm = TRUE))

# Summarize the critical variables district-wise

district\_summary <- assam\_data %>%

group\_by(District) %>%

summarize(avg\_No\_of\_Meals\_per\_day = mean(No\_of\_Meals\_per\_day, na.rm = TRUE),

min\_No\_of\_Meals\_per\_day = min(No\_of\_Meals\_per\_day, na.rm = TRUE),

max\_No\_of\_Meals\_per\_day = max(No\_of\_Meals\_per\_day, na.rm = TRUE))

assam\_data <- assam\_data %>%

mutate(District = case\_when(

District == "18" ~ "ASSM Urban",

TRUE ~ as.character(District) # or TRUE ~ District, depending on your needs

))

# Identify the top and bottom three districts of consumption

top\_three\_districts <- district\_summary %>%

arrange(desc(avg\_No\_of\_Meals\_per\_day)) %>%

head(3)

bottom\_three\_districts <- district\_summary %>%

arrange(avg\_No\_of\_Meals\_per\_day) %>%

head(3)

# Filter data for Assam

assam\_data <- filter(data, state\_1 == 'ASSM')

summary(anova\_result)

# Test whether the differences in the means are significant or not

anova\_result <- aov(No\_of\_Meals\_per\_day ~ District, data = assam\_data)

summary(anova\_result)

print(anova\_summary)

# Summarize the critical variables district-wise

district\_summary <- assam\_data %>%

group\_by(District) %>%

summarize(avg\_No\_of\_Meals\_per\_day = mean(No\_of\_Meals\_per\_day, na.rm = TRUE),

min\_No\_of\_Meals\_per\_day = min(No\_of\_Meals\_per\_day, na.rm = TRUE),

max\_No\_of\_Meals\_per\_day = max(No\_of\_Meals\_per\_day, na.rm = TRUE))

# Summarize the critical variables district-wise

district\_summary <- assam\_data %>%

group\_by(District) %>%

summarize(avg\_No\_of\_Meals\_per\_day = mean(No\_of\_Meals\_per\_day, na.rm = TRUE),

min\_No\_of\_Meals\_per\_day = min(No\_of\_Meals\_per\_day, na.rm = TRUE),

max\_No\_of\_Meals\_per\_day = max(No\_of\_Meals\_per\_day, na.rm = TRUE))

# Summarize the critical variables district-wise

district\_summary <- assam\_data %>%

group\_by(District) %>%

summarize(avg\_No\_of\_Meals\_per\_day = mean(No\_of\_Meals\_per\_day, na.rm = TRUE),

min\_No\_of\_Meals\_per\_day = min(No\_of\_Meals\_per\_day, na.rm = TRUE),

max\_No\_of\_Meals\_per\_day = max(No\_of\_Meals\_per\_day, na.rm = TRUE))

# Identify the top and bottom three districts of consumption

top\_three\_districts <- district\_summary %>%

arrange(desc(avg\_No\_of\_Meals\_per\_day)) %>%

head(3)

bottom\_three\_districts <- district\_summary %>%

arrange(avg\_No\_of\_Meals\_per\_day) %>%

head(3)

# Test whether the differences in the means are significant or not

anova\_result <- aov(No\_of\_Meals\_per\_day ~ District, data = assam\_data)

anova\_summary <- summary(anova\_result)

cat("ANOVA results:\n")

print(anova\_summary)

View(anova\_result)

View(anova\_result)

View(outlier\_iqr)

# Load the ggplot2 package

library(ggplot2)

# Libraries

library(ggplot2)

# Data from ANOVA summary

data <- data.frame(

Source = c("District", "Residuals"),

Mean\_Sq = c(0.0580, 0.1634)

)

# Plot

ggplot(data, aes(x = Source, y = Mean\_Sq)) +

geom\_bar(stat = "identity", color = "skyblue") +

labs(title = "Mean Squares (Effect Sizes)", x = "Source", y = "Variance")

q()

**References:**

 National Sample Survey Office (NSSO) website: <https://mospi.gov.in/current-surveys-0>

 RStudio website (for learning R programming): <https://posit.cloud/>

