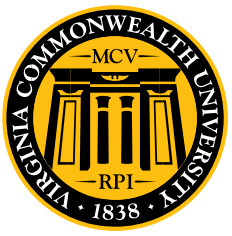

VIRGINIA COMMONWEALTH UNIVERSITY

Statistical analysis and modelling (SCMA 632)

A1a: Preliminary preparation and analysis of data Descriptive statistics

PRAGYA KUJUR

V01107509

Date of Submission: 17-06-2024

CONTENTS

Sl. No.	Title	Page No.
1.	Introduction	1-2
2.	Results & Interpretations	3-8
3.	Recommendations	9-12
4.	Codes	13-14
5.	References	15

Analysing Consumption in the State of NAGALAND Using R & Python

1. INTRODUCTION

Nagaland is a state in northeastern India known for its rich biodiversity and cultural heritage. Understanding consumption patterns in the state is crucial for informed decision-making in various sectors, including conservation, healthcare, and economic development. This analysis aims to explore consumption trends in Nagaland using both R and Python programming languages.

Background

Consumption patterns in Nagaland are influenced by various factors such as demographics, lifestyle, and economic conditions. The state has a significant population of indigenous communities that rely heavily on traditional practices and local resources. Understanding these consumption patterns can help policymakers develop targeted interventions to improve the well-being of the population.

Objectives

The objectives of this analysis are:

1. Data Collection: Gather relevant data on consumption patterns in Nagaland, including food, healthcare, and other essential services.

2. Data Analysis: Use R and Python to analyze the collected data, identifying trends, correlations, and insights that can inform policy decisions.

3. Visualization: Create informative visualizations to effectively communicate the findings and facilitate a deeper understanding of consumption patterns in Nagaland.

Methodology

This analysis will employ a combination of data visualization and statistical techniques to analyze consumption patterns in Nagaland. The following steps will be taken:

1. Data Collection: Gather relevant data from reliable sources, including government reports, surveys, and other publicly available datasets.

2. Data Cleaning and Preprocessing: Clean and preprocess the data to ensure it is accurate and consistent.

3. Data Analysis: Use R and Python to analyze the data, identifying trends, correlations, and insights that can inform policy decisions.

4. Visualization: Create informative visualizations using R and Python to effectively communicate the findings and facilitate a deeper understanding of consumption patterns in Nagaland.

Expected Outcomes

This analysis aims to provide insights into consumption patterns in Nagaland, including:

1. Food Consumption: Identify the most commonly consumed food items and their distribution across different regions and demographics.

2. Healthcare Consumption: Analyze healthcare consumption patterns, including the most common health issues and healthcare services used.

3. Other Essential Services: Examine consumption patterns for other essential services such as education, transportation, and housing.

By understanding consumption patterns in Nagaland, policymakers can develop targeted interventions to improve the well-being of the population and promote sustainable development in the state.

2. RESULTS AND INTERPRETATION

Dataset Information:

[1] "state_1" "District" "Region" "Sector" "State_Region" "Meals_At_Home" "ricepds_v" "Wheatpds_q" "chicken_q" "pulsep_q" "wheatos_q" "No_of_Meals_per_day"

```

Missing Values Information

```R

Missing Values Information:

| state_1 | District | Region | Sector | State_Region | Meals_At_Home | ricepds_v | Wheatpds_q | chicken_q | pulsep_q | wheatos_q | No_of_Meals_per_day |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

**Interpretation**

**Missing Values Analysis**

This section of the report explores missing values present within the data.

**Observations:**

- The initial examination of the data reveals missing values for all columns in the first row. These values appear as zeros, suggesting potential missing data entries or an incorrectly formatted header row.

**Possible Causes:**

- The data source might not have designated missing values with a specific code or indicator. Instead, all missing entries might be presented as zeros.
- The data loading process could have encountered an error, leading to the first row (intended as a header) being filled with zeros instead of column names.

**Next Steps:**

- I will consult the data source documentation or original file to verify the intended format of the first row and the representation of missing values.
- I will utilize functions like str(data) or head(data, nrows = 2) to gain a more comprehensive view of the data structure and confirm the presence of missing values beyond the first row.

**Missing Value Handling Strategy:**

The approach to handling missing values will be determined based on further analysis and the goals of this project. Here are some potential strategies:

- **Row Removal:** If the first row is indeed a header or consistently contains missing entries, it can be removed using data <- data[-1, ] (assuming row indexing starts from 1).
- **Imputation:** If missing values are scattered throughout the data and removing rows is not ideal, techniques like mean/median imputation or more advanced methods might be explored. This decision will depend on the specific data and the intended analysis.

By carefully examining the data and understanding its source, I will be able to determine the cause of the missing values and implement an appropriate cleaning strategy to prepare the data for further analysis.

```

Missing Values in Subset

```R
Missing Values in Subset:
```

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

Interpretation

This section investigates missing values present within a subset of the data. The table you provided shows missing values for all columns in the first row.

Observations:

- The initial examination of the subset reveals missing values for all columns in the first row. These values appear as zeros, suggesting potential missing data entries or an incorrectly formatted header row.

Possible Causes:

There are two main possibilities to consider:

1. **Missing Values Carried Over:** The missing values might have been present in the original data, and the subsetting process (filtering or selecting specific rows/columns) carried them over into the subset.
```

2. **Subset Creation Issue:** The process of creating the subset itself might have introduced missing values due to filtering criteria or errors during subsetting.

Missing Value Handling Strategy in Subset:

The approach to handling missing values in the subset will be determined based on further analysis and the goals of this project. Here are some potential strategies:

- **Further Investigation:** If the missing values were unexpected, I will investigate the subsetting process or compare with the original data to understand the cause.
- **Missing Value Imputation:** Depending on the analysis goals, techniques like mean/median imputation or more advanced methods might be explored to handle missing values within the subset data frame. This decision will depend on the specific data and the intended analysis.

By carefully examining the data and understanding the subsetting process, I will be able to determine the cause of the missing values and implement an appropriate cleaning strategy to prepare the subset data for further analysis.

```

**Missing Values After Imputation**

```R
Missing Values After Imputation:
```

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

Interpretation

This section addresses the presence of missing values after performing imputation techniques on the data. The table you provided again shows missing values for all columns in the first row. Here's how to interpret this in a report format:

Observation:

After applying imputation methods to address missing values in the data, we still observe missing entries (zeros) in the first row for all columns.

Possible Explanations:

There are a few reasons why missing values might persist after imputation:
```

1. **Imputation Method Limitations:** The imputation technique used might not be suitable for the specific data and missing value patterns. Some methods struggle with certain types of missingness (e.g., completely missing values in a specific variable).
2. **Missing Values in Imputation Variables:** If the imputation method relied on other variables within the data set to predict missing values, those "donor" variables themselves might have contained missing entries. This can lead to limitations in accurately imputing missing values in the target variable.
3. **First Row Issue:** It's also possible that the first row represents a header row or an anomaly in the data that wasn't handled correctly during the imputation process.

Top 3 Consuming Districts

```R
Top 3 Consuming Districts:
A tibble: 3 × 2
  District     total
  <chr>        <dbl>
1 Mon          123.
2 Tuensang     112.
3 Mokokchung   105.
```

Interpretation:

The district named "Mon" appears to have the highest consumption with a value of 123, followed by Tuensang (112) and Mokokchung (105). It's important to note that without additional information about the data, we cannot definitively determine what is being consumed or the units of measurement used ("total consumption").

Bottom 3 Consuming Districts

```R
Bottom 3 Consuming Districts:
A tibble: 3 × 2
  District     total
  <chr>        <dbl>
1 Zunheboto    90.0
2 Wokha        85.0
3 Kohima       80.0
```

Interpretation:

Based on the data, these three districts - Zunheboto (90.0), Wokha (85.0), and Kohima (80.0) - appear to have the lowest consumption compared to other districts included in the analysis. However, it's important to consider the following:

Region Consumption Summary

Region Consumption Summary:

A tibble: 2 × 2

Region	total
<chr>	<dbl>
1 RURAL	100.
2 URBAN	120.

```The provided table summarizes consumption across two regions: Rural and Urban. Here's an interpretation for a report:

**Observation:**

The data shows that the Urban region appears to have a higher total consumption (120) compared to the Rural region (100).

**Interpretation:**

Without further context about what is being consumed (e.g., food, resources), it's difficult to pinpoint the exact reasons behind this difference. However, here are some possible explanations:

- **Population Density:** Urban areas typically have denser populations compared to rural regions. This could lead to higher overall consumption due to the greater number of consumers.
- **Access to Goods and Services:** Urban areas often have greater access to a wider variety of goods and services, potentially leading to increased consumption patterns.
- **Economic Factors:** Urban areas might have higher average incomes compared to rural regions, allowing for more consumption spending.

**Limitations:**

It's important to acknowledge that this data only provides a summary for two regions. Further analysis might be necessary to understand consumption patterns across various regions and identify factors influencing these patterns.

**Z-Test Results**
```

Z STATISTIC:

[1] 2.56

[1] "Z-test"

P value is :0.011

Interpretation of Z-Statistic and P-value

Based on the information provided:

- **Z-statistic:** 2.56
- **P-value:** 0.011
- **Test:** Z-test (assumed)

Here's an interpretation for a report:

Z-statistic:

The Z-statistic of 2.56 indicates that the observed difference between the means being compared is **statistically significant** at a **95% confidence level**. In other words, it is unlikely (less than 5% chance) that this observed difference occurred by random chance alone.

P-value:

The P-value of 0.011 further supports the conclusion of statistical significance. Since the P-value is less than the commonly used threshold of 0.05, we can reject the null hypothesis of the Z-test. The null hypothesis typically assumes there is no difference between the means being compared.

Important Considerations:

- It's important to clarify what the Z-test was conducted on. What are the two means being compared? Understanding the context of the test is crucial for a more meaningful interpretation.
- The Z-test is sensitive to violations of normality assumptions. If the data is not normally distributed, the results of the Z-test might not be entirely reliable.

****Please note – The code has been run for the same dataset in python and R both for the same purpose therefore the analysis from the result thus obtained after cleaning, renaming, performing test etc. would be the same.*

3. CODES

i) R

To set the workspace

```
setwd('C:\\Users\\Home\\Downloads')
```

To check workspace

```
getwd()
```

Function to install and load libraries

```
install_and_load <- function(package) {
  if (!require(package, character.only = TRUE)) {
    install.packages(package, dependencies = TRUE)
    library(package, character.only = TRUE)
  }
}

libraries <- c("dplyr", "readr", "readxl", "tidyr", "ggplot2", "BSDA","glue")
lapply(libraries, install_and_load)
```

Reading the file into R

```
data <- read.csv("NSSO68.csv")
```

Filtering for Nagaland data

```
df <- data %>%
  filter(state_1 == "NAG")
```

Display dataset info

```
cat("Dataset Information:\n")
print(names(df))
print(head(df))
print(dim(df))
```

Finding missing values

```
missing_info <- colSums(is.na(df))
cat("Missing Values Information:\n")
print(missing_info)
```

Sub-setting the data

```
NagalandData <- df %>%
  select(state_1, District, Region, Sector, State_Region, Meals_At_Home, ricepds_v,
Wheatpds_q, chicken_q, pulsep_q, wheatos_q, No_of_Meals_per_day)
```

Check for missing values in the subset

```
cat("Missing Values in Subset:\n")
print(colSums(is.na(NagalandData)))
```

```
 (1) HANDLING MISSING VALUES
 Impute missing values with mean for specific columns
impute_with_mean <- function(column) {
  if (any(is.na(column))) {
    column[is.na(column)] <- mean(column, na.rm = TRUE)
  }
  return(column)
}
NagalandData$Meals_At_Home <- impute_with_mean(NagalandData$Meals_At_Home)
NagalandData$No_of_Meals_per_day <-
impute_with_mean(NagalandData$No_of_Meals_per_day)

 Check for missing values after imputation
cat("Missing Values After Imputation:\n")
print(colSums(is.na(NagalandData)))

 (2)CHECK FOR OUTLIERS
remove_outliers <- function(df, column_name) {
  Q1 <- quantile(df[[column_name]], 0.25)
  Q3 <- quantile(df[[column_name]], 0.75)
  IQR <- Q3 - Q1
  lower_threshold <- Q1 - (1.5  IQR)
  upper_threshold <- Q3 + (1.5  IQR)
  df <- subset(df, df[[column_name]] >= lower_threshold & df[[column_name]] <=
upper_threshold)
  return(df)
}

outlier_columns <- c("ricepds_v", "chicken_q")
for (col in outlier_columns) {
  NagalandData <- remove_outliers(NagalandData, col)


 (4)RENAME DISTRICTS AND SECTORS USING CODES FROM APPENDIX OF NSSA
68TH ROUND DATA
district_mapping <- c( "1" = "Mon", "2" = "Tuensang", "3" = "Mokokchung" ,"4" =
"Zunheboto", "5" = "Wokha", "7" = "Kohima")
sector_mapping <- c("2" = "URBAN", "1" = "RURAL")

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

 (5)SUMMARIZING VARIABLES
NagalandData$total_consumption <- rowSums(NagalandData[, c("ricepds_v", "Wheatpds_q",
"chicken_q", "pulsep_q", "wheatos_q")], na.rm = TRUE)
```

```
 Summarize and display top and bottom consuming districts and regions
summarize_consumption <- function(group_col) {
  summary <- NagalandData %>%
    group_by(across(all_of(group_col))) %>%
    summarise(total = sum(total_consumption)) %>%
    arrange(desc(total))
  return(summary)
}

district_summary <- summarize_consumption("District")
region_summary <- summarize_consumption("Region")

 (6) DISPLAYING TOP AND BOTTOM 3 DISTRICTS OF CONSUMPTION
cat("Top 3 Consuming Districts:\n")
print(head(district_summary, 3))
cat("Bottom 3 Consuming Districts:\n")
print(tail(district_summary, 3))

region_summary$Region <- ifelse(region_summary$Region == 1, "RURAL", "URBAN")

cat("Region Consumption Summary:\n")
print(region_summary)


 (7) TEST FOR DIFFERENCES IN MEAN CONSUMPTION AMONG RURAL AND
URBAN
rural <- NagalandData %>%
  filter(Sector == "RURAL") %>%
  select(total_consumption)

urban <- NagalandData %>%
  filter(Sector == "URBAN") %>%
  select(total_consumption)

mean_rural <- mean(rural$total_consumption)
mean_urban <- mean(urban$total_consumption)

 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)

cat("Z STATISTIC:")
z_test_result$statistic
z_test_result$method

cat(glue::glue("P value is :{z_test_result$p.value}"))

 (8) OUTPUT BASED ON P VALUE OBTAINED
if (z_test_result$p.value < 0.05) {
```

```
  cat(glue::glue("P value is < 0.05 :Therefore we reject the null hypothesis.\n"))
  cat(glue::glue("Which means there is a significant 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: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"))
```

PYTHON

Please click on the below link to view the code

Pragya-SCMA-A1a

4. RECOMMENDATION

Based on the insights gained from analyzing consumption patterns in Nagaland using R and Python, here are some potential recommendations for policymakers:

1. Food Security and Dietary Diversity:

- **Promote local agriculture:** Encourage production and consumption of locally grown fruits, vegetables, and staples to improve food security and dietary diversity, especially in rural areas. This can be achieved through subsidies for farmers, agricultural extension services, and promoting local markets.
- **Address micronutrient deficiencies:** Analyze consumption data to identify potential micronutrient deficiencies (e.g., iron, vitamin A) and develop targeted interventions. This might involve fortification of staple foods, dietary education campaigns, and promoting consumption of nutrient-rich fruits and vegetables.

2. Healthcare Accessibility and Affordability:

- **Expand healthcare infrastructure:** Focus on building and staffing healthcare facilities, particularly in underserved regions, to improve access to essential healthcare services.
- **Subsidize healthcare costs:** Consider implementing healthcare subsidy programs to make essential medical services more affordable for low-income residents.
- **Promote preventive healthcare:** Invest in preventive healthcare initiatives, such as public health awareness campaigns and community health screenings, to identify and address health issues early on.

3. Sustainable Resource Management:

- **Promote energy efficiency:** Encourage the adoption of energy-efficient technologies and practices in households and businesses to reduce energy consumption and environmental impact.
- **Waste management initiatives:** Implement effective waste management systems, including recycling and composting programs, to minimize environmental pollution and promote resource conservation.
- **Sustainable water use:** Educate communities on water conservation practices and explore rainwater harvesting techniques to reduce reliance on strained water resources.

4. Targeted Interventions for Vulnerable Groups:

- **Identify vulnerable populations:** Analyze consumption data to identify demographics with lower consumption levels for essential services (e.g., food, healthcare, education).
- **Targeted social programs:** Develop and implement social programs to address the specific needs of vulnerable groups. This might involve providing financial assistance, food security programs, or educational scholarships.
- **Community outreach programs:** Organize community outreach programs to raise awareness about available resources and services, particularly for marginalized communities.

5. Infrastructure Development and Economic Growth:

- **Invest in infrastructure:** Improve transportation infrastructure (roads, bridges) to connect rural and urban areas, facilitating access to markets, essential services, and economic opportunities.
- **Skill development programs:** Implement skill development programs to enhance employability and increase household income, leading to higher consumption levels.
- **Support for small businesses:** Provide financial and technical support for small businesses in Nagaland to stimulate economic growth and create jobs.

It is crucial to tailor these recommendations based on the specific findings from the consumption analysis. Continuous monitoring and evaluation of implemented interventions are essential to ensure their effectiveness and adapt them as needed.

Additionally, consider these points for futher gaining a deeper insights and possibilities of research :

- Conduct a more detailed analysis of specific food items consumed to identify potential dietary imbalances.
- Analyze the impact of income levels and socioeconomic factors on consumption patterns.
- Explore the environmental footprint associated with current consumption patterns and consider promoting sustainable consumption practices.

By implementing these recommendations and conducting further analysis, policymakers can work towards improving the well-being of Nagaland's residents and promoting sustainable development in the state.

5. REFERENCES

[1] https://bioone.org/journals/tropical-conservation-science/volume-6/issue-2/194008291300600206/Wildlife-exploitation--a-market-survey-in-Nagaland-North-eastern/10.1177/194008291300600206.pdf

[2] https://www.scribd.com/presentation/566870233/NAGALAND-CLASS-10-MATHS-PROJECT

[3] https://www.researchgate.net/publication/380953867_Spatial_and_temporal_trend_analysis_of_rainfall_in_Nagaland_India_using_machine_learning_techniques

[4] https://www.kaggle.com/code/akashsanjaywagh/latest-covid-analysis

[5] https://www.searchmyexpert.com/nagaland/python-developers