VIRGINIA COMMONWEALTH UNIVERSITY

Statistical analysis and modelling (SCMA 632)

A5: Visualization - Perceptual Mapping for Business

DEEPTHI ANNA ALEX
V01101949

Date of Submission: 15-07-2024

CONTENTS

I

Introduction:

In this report, we analyze the consumption patterns across different districts of Arunachal Pradesh . The data used for this analysis is sourced from Assignment A1, which provides insights into the total consumption in various districts. The objective is to visualize and understand how consumption varies across districts within the state. Through these visualizations and analyses, we gain valuable insights into the consumption patterns and geographical distributions of {'any variable of your choice'} across Arunachal Pradesh. These insights are crucial for understanding regional dynamics and informing targeted policy interventions aimed at addressing disparities and optimizing resource allocation.

OBJECTIVES:

A) Plot a **histogram** (to show the distribution of total consumption across different districts) and a **bar-plot** (To visualize consumption per district with district names) of the data in **Assignment A1** to indicate the consumption district-wise for the state assigned to you.
B) Plot {'any variable of your choice'} on the **Karnataka** (or the state assigned to you) state map using NSSO68.csv data

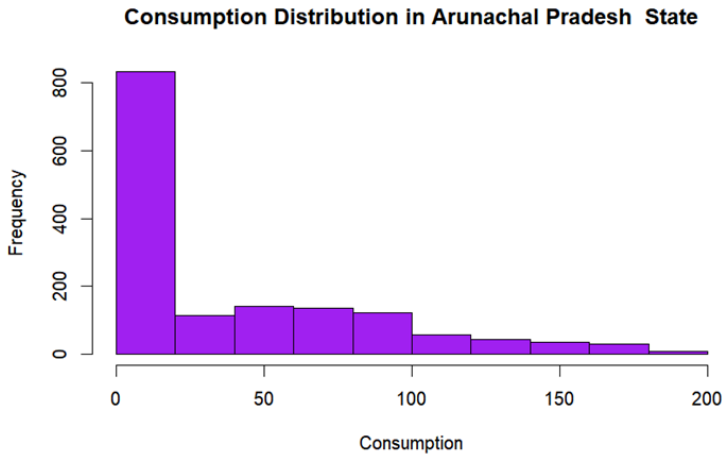
PART-A

Results and Interpretation

	state_1	District	Region	Sector	State_Region	Meals_At_Home	ricepds_v	Wheatpds_q	chicken_q	pulsep
1	ARP	changlang	1	URBAN	121	85	0.00000	0.0000000	0.0000000	0.0
3	ARP	changlang	1	URBAN	121	90	64.00000	0.0000000	0.0000000	0.0
4	ARP	changlang	1	URBAN	121	89	54.00000	0.0000000	0.4000000	0.0
5	ARP	changlang	1	URBAN	121	84	45.00000	0.0000000	0.1666667	0.0
6	ARP	changlang	1	URBAN	121	87	67.50000	0.0000000	0.2500000	0.0
7	ARP	changlang	1	URBAN	121	4	0.00000	0.0000000	0.0000000	0.0
8	ARP	changlang	1	URBAN	121	90	56.25000	0.0000000	0.0000000	0.0
9	ARP	west siang	1	URBAN	121	60	0.00000	0.0000000	0.0000000	0.0
10	ARP	west siang	1	URBAN	121	60	0.00000	0.0000000	0.5000000	0.0
11	ARP	west siang	1	URBAN	121	90	0.00000	0.0000000	0.2500000	0.0
12	ARP	west siang	1	URBAN	121	60	0.00000	0.0000000	0.0000000	0.0
13	ARP	west siang	1	URBAN	121	90	0.00000	0.0000000	0.0000000	0.0
14	ARP	west siang	1	URBAN	121	90	0.00000	0.0000000	0.0000000	0.0
15	ARP	west siang	1	URBAN	121	60	0.00000	0.0000000	0.0000000	0.0
16	ARP	west siang	1	URBAN	121	60	33.75000	0.0000000	0.2500000	0.0

District	total
<chr>	<dbl>
papum pare	8168.
east siang	6291.
tirap	6178.
west kameng	5760.
tawang	5729.
kurungkumey	3707.
lohit	3679.
west siang	3501.
changlang	3351.
east kameng	2809.
lower dibang	1987.
upper subansiri	1916.

Top consuming districts from top to bottom



To interpret the histogram in relation to the consumption data across different districts in Arunachal Pradesh, here's a detailed analysis:

Title and Axes:

- **Title:** "Consumption Distribution in Arunachal Pradesh State"
- **X-Axis (Horizontal Axis):** This axis represents the consumption values, which range from 0 to 200 units.
- **Y-Axis (Vertical Axis):** This axis shows the frequency, indicating how many districts fall within each consumption range.

Observations:

1. **High Frequency of Low Consumption**

   The first bar (0-20 units) has the highest frequency, with about 800 instances. This suggests that the majority of the districts have very low consumption values.

2. **Moderate Consumption**

The subsequent bars (20-40, 40-60 units) show a frequency of approximately 200-300. These bars represent districts with moderate consumption values.

3. **Decreasing Frequency**

As consumption increases (60-80 units and beyond), the frequency continues to decrease, indicating fewer districts have higher consumption values.

4. **Minimal High Consumption**

There are very few instances of high consumption (above 100 units), with frequencies nearing zero.

District Interpretation:

Based on the histogram and the districts listed, we can infer the following:

1. **Majority with Low Consumption:**

Districts like Anjaw, Changlang, Dibang Valley, and Longding likely fall into the low consumption range (0-20 units), contributing to the high frequency in this range.

2. **Moderate Consumption:**

Districts such as East Kameng, Kurung Kumey, Lohit, and Lower Dibang might fall into the moderate consumption ranges (20-60 units).

3. **Higher Consumption:**

Districts like Papum Pare, East Siang, West Kameng, and West Siang, which showed higher total consumption in the previous bar plot, might contribute to the higher consumption ranges (60 units and above) seen in the histogram.

4. **Outliers:**

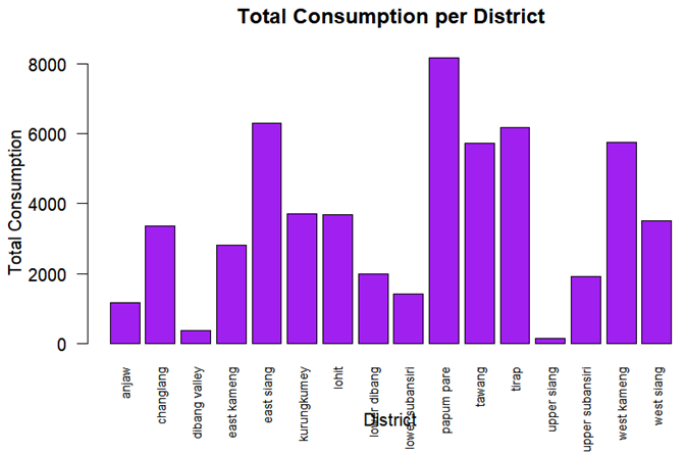
Districts such as Papum Pare, which had the highest consumption in the bar plot, might be represented by the rightmost bars in the histogram (although the histogram doesn't extend beyond 200 units, Papum Pare likely represents an extreme value).

Interpretation

This histogram provides a comprehensive view of how consumption is distributed across the districts in Arunachal Pradesh. It indicates that most districts have low to moderate consumption levels, with very few districts exhibiting high consumption. This distribution is skewed to the right, suggesting a few districts have significantly higher consumption levels

compared to the rest. Understanding this distribution helps in identifying which districts might need more resources or focused interventions to manage their consumption effectively.

	District	total_consumption
1	anjaw	1174.0994
2	changlang	3351.4529
3	dibang valley	368.7500
4	east kameng	2809.2576
5	east siang	6291.1357
6	kurungkumey	3706.5894
7	lohit	3678.8722
8	lower dibang	1987.2769
9	lower subansiri	1417.0187
10	papum pare	8167.7724
11	tawang	5729.0270
12	tirap	6177.7667
13	upper siang	139.8988
14	upper subansiri	1915.9606
15	west kameng	5759.7734



This bar plot displays the "Total Consumption per District" for various districts, likely in a specific region. Here's a detailed explanation:

Title and Axes:

Title: The title of the graph is "Total Consumption per District".

X-Axis (Horizontal Axis): This axis represents the different districts. The districts listed are:

- anjaw
- changlang
- dibang valley
- east kameng
- east siang
- kurung kumey
- lohit
- longding
- lower dibang
- lower subansiri
- papum pare
- siang
- tawang
- tirap
- upper siang
- upper subansiri
- west kameng
- west siang

Y-**Axis (Vertical Axis):** This axis represents the "Total Consumption". The units of measurement are not specified but are likely in a quantitative unit such as liters, kilograms, or another metric relevant to consumption.

Observations:

1. **Papum Pare:** This district has the highest total consumption, with a value approaching 8000 units.
2. **East Siang:** This district also has a high total consumption, above 6000 units.
3. **West Siang and West Kameng:** Both districts have significant consumption levels, around 5000 units.
4. **Tawang and Lower Subansiri:** These districts have moderate consumption levels, around 4000 units.
5. **East Kameng, Kurung Kumey, Lohit, Lower Dibang:** These districts have lower consumption levels, ranging between approximately 3000 to 4000 units.
6. **Anjaw, Changlang, Upper Siang, Tirap:** These districts have comparatively low total consumption, with values between approximately 1000 to 2000 units.
7. **Dibang Valley and Upper Subansiri:** These districts have the lowest total consumption, with Dibang Valley showing minimal consumption.

Insights:

- The bar plot visually emphasizes that Papum Pare is a standout district in terms of total consumption.
- There is significant variation in consumption across different districts.
- A few districts exhibit minimal consumption, which might indicate lower population, lesser resource availability, or other socio-economic factors.

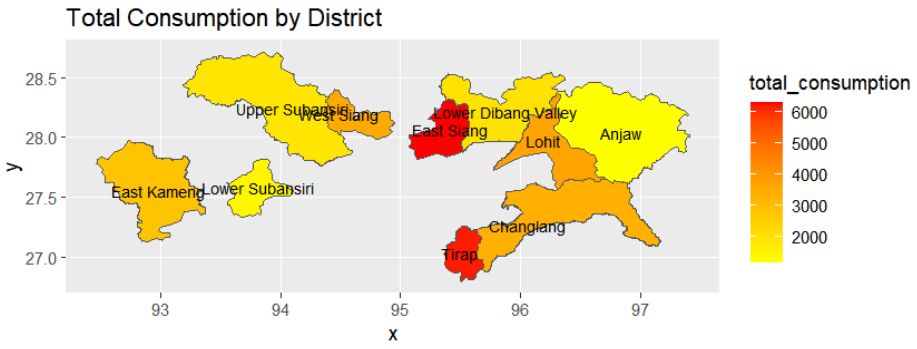
- The consumption trends could be influenced by factors such as population density, economic activities, resource availability, and infrastructure in the respective districts.

Interpretation

This bar plot provides a clear comparison of total consumption across various districts, highlighting significant disparities and pinpointing the districts with the highest and lowest consumption. This information can be valuable for resource allocation, planning, and policy-making in the respective region.

PART-B

Results and Interpretation



The provided map displays "Total Consumption by District" in a region, likely Arunachal Pradesh, based on the district names. Here's a detailed interpretation of the map:

Title and Axes:

- **Title:** "Total Consumption by District."
- **X-Axis (Horizontal Axis):** Represents longitude, ranging from approximately 92° to 97°.
- **Y-Axis (Vertical Axis):** Represents latitude, ranging from approximately 26.5° to 28.5°.

Color Scale:

- The color scale on the right side indicates "total_consumption" values.

- **Yellow:** Represents lower total consumption (around 2000 units).
- **Orange:** Represents medium total consumption (around 4000 units).
- **Red:** Represents higher total consumption (around 6000 units).

Districts and Consumption:

➢ High Consumption (Red Areas):

- **East Siang:** Shown in red, indicating the highest consumption, around 6000 units.
- **Tirap:** Also shown in red, indicating high consumption.

➢ Medium Consumption (Orange Areas):

- **Changlang:** Shown in orange, indicating medium consumption, around 4000 units.
- **Lower Dibang Valley:** Shown in orange, indicating medium consumption.
- **Lohit:** Shown in light orange, indicating slightly lower medium consumption.

➢ Low Consumption (Yellow Areas):

- **Anjaw:** Shown in yellow, indicating lower consumption, around 2000 units.
- **Upper Subansiri, Upper Siang, East Kameng, Lower Subansiri:** All shown in yellow, indicating low consumption.

Geographical Context:

- The map provides a spatial distribution of consumption across different districts in the region. Each district is color-coded to represent its total consumption.

Insights:

1. **High Consumption Concentration:**

   High consumption districts (East Siang and Tirap) are clearly distinguishable with red shading.

2. **Regional Variations:**

   There is a notable variation in consumption across different districts, with some having significantly higher consumption than others.

3. **Planning and Resource Allocation:**

This map can be a valuable tool for policymakers and planners to allocate resources efficiently. Areas with high consumption might need more resources or interventions to manage demand.

4. **Targeted Interventions:**

   Districts with medium to high consumption might benefit from targeted interventions to ensure sustainable consumption patterns.

Interpretation

This thematic map effectively visualizes the total consumption across different districts in a region. By using a color scale from yellow to red, it highlights districts with low, medium, and high consumption. This visual representation aids in understanding the spatial distribution of consumption and helps in making informed decisions regarding resource management and policy planning.

Codes

R code

```
#install.packages(dplyr)

# 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)

  }

}

# Load required libraries

libraries <- c("dplyr", "readr", "readxl", "tidyr", "ggplot2", "BSDA")

lapply(libraries, install_and_load)
```

```
# Reading the file into R
data <- read.csv("C:\\Users\\HP\\Downloads\\NSSO68 (2).csv")



# Filtering for ARP
df <- data %>%
  filter(state_1 == "ARP")


# 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)


# Subsetting the data
arpnew <- 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)
```

```
# 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)
}
arpnew$Meals_At_Home <- impute_with_mean(arpnew$Meals_At_Home)


# Finding outliers and removing them
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) {
  arpnew <- remove_outliers(arpnew, col)
}
```

```
# Summarize consumption

arpnew$total_consumption <- rowSums(arpnew[, c("ricepds_v", "Wheatpds_q",
"chicken_q", "pulsep_q", "wheatos_q")], na.rm = TRUE)


# Summarize and display top consuming districts and regions

summarize_consumption <- function(group_col) {

  summary <- arpnew %>%

    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")


cat("Top Consuming Districts:\n")

print(head(district_summary, 4))

cat("Region Consumption Summary:\n")

print(region_summary)


# Rename districts and sectors

district_mapping <- c ("4" = "papum pare","8" = "east siang","13" = "tirap","2" =
"west kameng","1" = "tawang","15" = "kurungkumey","11" = "lohit","7" = "west
siang","12" = "changlang","3" = "east kameng","9"="upper siang","6"="upper
```

```
subansiri","5"="lower subansiri","16"="lower dibang","14"="anjaw","10"="dibang valley")



sector_mapping <- c("2" = "URBAN", "1" = "RURAL")


arpnew$District <- as.character(arpnew$District)

arpnew$Sector <- as.character(arpnew$Sector)

arpnew$District <- ifelse(arpnew$District %in% names(district_mapping), district_mapping[arpnew$District], arpnew$District)

arpnew$Sector <- ifelse(arpnew$Sector %in% names(sector_mapping), sector_mapping[arpnew$Sector], arpnew$Sector)

View(arpnew)


hist(arpnew$total_consumption, breaks = 10, col = 'purple', border = 'black',

    xlab = "Consumption", ylab = "Frequency", main = "Consumption Distribution in Arunachal Pradesh  State")


ARP_consumption <- aggregate(total_consumption ~ District, data = arpnew, sum)

View(ARP_consumption)

??barplot

barplot(ARP_consumption$total_consumption,

        names.arg = ARP_consumption$District,

        las = 2, # Makes the district names vertical

        col = 'purple',

        border = 'black',
```

```
      xlab = "District",

      ylab = "Total Consumption",

      main = "Total Consumption per District",

      cex.names = 0.7) # Adjust the size of district names if needed

# b) Plot {'any variable of your choice'} on the Arunachal Pradesh state map using NSSO68.csv data

# Filtering for Arunachal Pradesh

df_arp <- data %>%

  filter(state_1 == "ARP")



# Sub-setting the data

arp_new <- df_arp %>%

  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(arp_new)))



# Impute missing values with mean for specific columns

arp_new$Meals_At_Home <- impute_with_mean(arp_new$Meals_At_Home)



# Check for missing values after imputation

cat("Missing Values After Imputation:\n")

print(colSums(is.na(arp_new)))
```

```
# Finding outliers and removing them

outlier_columns <- c("ricepds_v", "chicken_q")

for (col in outlier_columns) {

  arp_new <- remove_outliers(arp_new, col)

}


# Summarize consumption

arp_new$total_consumption <- rowSums(arp_new[, c("ricepds_v", "Wheatpds_q",
"chicken_q", "pulsep_q", "wheatos_q")], na.rm = TRUE)


district_summary <- summarize_consumption("District")

cat("District Consumption Summary:\n")

print(district_summary)


# mapping districts so that meging of the tables will be easier

district_mapping <- c(

  "1"="tawang",

  "2"="west kameng",

  "3"="East Kameng",

  "4"="Papum Pare *",

  "5"="Lower Subansiri",

  "6"="Upper Subansiri",

  "7"="West Siang",

  "8"="East Siang",
```

```
  "9"="Upper Siang *",
  "10"="Dibang Valley",
  "11"="Lohit",
  "12"="Changlang",
  "13"="Tirap",
  "14"="Anjaw",
  "15"="Kurungkumey",
  "16"="Lower Dibang Valley"
)

arp_new$District <- as.character(arp_new$District)
arp_new$District <- district_mapping[arp_new$District]
#arp_new$District <- ifelse(arp_new$District %in% names(district_mapping),
district_mapping[arp_new$District], arp_new$District)
View(arp_new)

# arp_consumption stores aggregate of total consumption district wise
arp_consumption <- aggregate(total_consumption ~ District, data = arp_new, sum)
View(arp_consumption)

#Plotting total consumption on the Arunachal Pradesh state

Sys.setenv("SHkaE_RESTORE_SHX" = "YES")
```

```
data_map <- st_read("C:\\Users\\HP\\OneDrive\\Desktop\\ARUNACHAL
PRADESH_DISTRICTS.geojson")

View(data_map)


data_map <- data_map %>%

  rename(District = dtname)


# merging arp_consumption and data_map tables

data_map_data <- merge(arp_consumption,data_map,by = "District")

View(data_map_data)


# Plot without labeling district names

ggplot(data_map_data) +

  geom_sf(aes(fill =total_consumption, geometry = geometry)) +

  scale_fill_gradient(low = "yellow", high = "red") +

  ggtitle("Total Consumption by District")


# Plot with labelled district names

ggplot(data_map_data) +

  geom_sf(aes(fill = total_consumption, geometry = geometry)) +

  scale_fill_gradient(low = "yellow", high = "red") +

  ggtitle("Total Consumption by District") +

  geom_sf_text(aes(label = District, geometry = geometry), size = 3, color = "black")
```

Python code

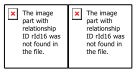
```
import pandas as pd

import geopandas as gpd

import matplotlib.pyplot as plt


# Reading the CSV file into pandas DataFrame

data = pd.read_csv("C:\\Users\\HP\\Downloads\\NSSO68 (2).csv")


# Filtering for ARP (Arunachal Pradesh)

df = data[data['state_1'] == "ARP"]


# Display dataset info

print("Dataset Information:")

print(df.columns)

print(df.head())

print(df.shape)


# Finding missing values

missing_info = df.isna().sum()

print("Missing Values Information:")

print(missing_info)


# Subsetting the data
```

```
arpnew = df[["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"]]

# Impute missing values with mean for specific columns
arpnew['Meals_At_Home'].fillna(arpnew['Meals_At_Home'].mean(), inplace=True)

# Function to remove outliers
def remove_outliers(df, column_name):
    Q1 = df[column_name].quantile(0.25)
    Q3 = df[column_name].quantile(0.75)
    IQR = Q3 - Q1
    lower_threshold = Q1 - (1.5 * IQR)
    upper_threshold = Q3 + (1.5 * IQR)
    df = df[(df[column_name] >= lower_threshold) & (df[column_name] <= upper_threshold)]
    return df

# Finding outliers and removing them
outlier_columns = ["ricepds_v", "chicken_q"]
for col in outlier_columns:
    arpnew = remove_outliers(arpnew, col)

# Summarize consumption
arpnew['total_consumption'] = arpnew[["ricepds_v", "Wheatpds_q", "chicken_q",
```

```
"pulsep_q", "wheatos_q"]].sum(axis=1)


# Summarize and display top consuming districts and regions
district_summary = arpnew.groupby("District")['total_consumption'].sum().reset_index()
district_summary = district_summary.sort_values(by='total_consumption', ascending=False)
print("Top Consuming Districts:")
print(district_summary.head())


# Renaming districts
district_mapping = {
    "1": "Tawang",
    "2": "West Kameng",
    "3": "East Kameng",
    "4": "Papum Pare",
    "5": "Lower Subansiri",
    "6": "Upper Subansiri",
    "7": "West Siang",
    "8": "East Siang",
    "9": "Upper Siang",
    "10": "Dibang Valley",
    "11": "Lohit",
    "12": "Changlang",
    "13": "Tirap",
    "14": "Anjaw",
```

```
    "15": "Kurungkumey",
    "16": "Lower Dibang Valley"
}

arpnew['District'] = arpnew['District'].map(district_mapping)

# Loading the GeoJSON file into a GeoDataFrame using geopandas
data_map = gpd.read_file("C:\\Users\\HP\\OneDrive\\Desktop\\ARUNACHAL PRADESH_DISTRICTS.geojson")

# Merging consumption data with GeoDataFrame
data_map_data = data_map.merge(district_summary, left_on='District', right_on='District')

# Plotting with labeled district names
fig, ax = plt.subplots(figsize=(12, 8))
data_map_data.plot(column='total_consumption', cmap='YlOrRd', linewidth=0.8, ax=ax, edgecolor='0.8')
ax.set_title('Total Consumption by District')
for idx, row in data_map_data.iterrows():
    ax.text(row.geometry.centroid.x, row.geometry.centroid.y, s=row['District'], ha='center', fontsize=8)
plt.show()
```


The image part with relationship ID rId16 was not found in the file.


The image part with relationship ID rId16 was not found in the file.
The image part with relationship ID rId16 was not found in the file.