VIRGINIA COMMONWEALTH UNIVERSITY

STATISTICAL ANALYSIS & MODELING

A5: VISUALIZATION - PERCEPTUAL MAPPING FOR BUSINESS

ADARSH BHARATHWAJ
V01107513

Date of Submission: 13/07/2024

CONTENTS

VISUALIZATION - PERCEPTUAL MAPPING FOR BUSINESS USING PYTHON

INTRODUCTION

The National Sample Survey Office (NSSO) conducts various surveys to collect socio-economic data vital for policy formulation, planning, and decision-making. The NSSO 68th round data, encapsulated in the dataset "NSSO68.csv," provides a comprehensive overview of consumption patterns across different districts in India. This dataset is a critical resource for understanding the regional disparities in consumption, which can reflect broader socio-economic conditions, including income levels, poverty rates, and lifestyle differences.

The data encompasses a wide array of consumption metrics, capturing the expenditure on essential and non-essential goods and services. By analyzing this data, researchers, policymakers, and businesses can gain insights into consumer behavior and preferences at the district level. Such granular data is instrumental in identifying target markets, assessing the effectiveness of social welfare programs, and designing localized strategies for economic development.

OBJECTIVES

a) Plot a histogram (to show the distribution of total consumption across different districts) and a barplot (To visualize consumption per district with district names) of the data to indicate the consumption district-wise for the state assigned to you.

b) Depict the consumption on the state map, showing consumption in each district for the state of Karnataka.

BUSINESS SIGNIFICANCE

Understanding consumption patterns is crucial for businesses aiming to penetrate or expand in the diverse Indian market. The NSSO68 dataset provides valuable insights into district-wise consumption, enabling companies to tailor their products and marketing strategies to meet the specific needs and preferences of different regions.

- **Market Penetration and Expansion:** Businesses can use the NSSO68 dataset to identify high-demand districts and optimize distribution channels, allowing for more effective market penetration and expansion strategies.
- **Product and Marketing Customization:** Companies can tailor their products and

marketing campaigns to meet the specific needs and preferences of different regions, enhancing customer satisfaction and market share.

- **Policy Design and Resource Allocation:** Policymakers can leverage the data to address regional imbalances, design targeted interventions, and improve resource allocation to boost economic activity in underprivileged areas.
- **Impact Monitoring and Evaluation:** The dataset enables government agencies to monitor the impact of existing policies and programs, facilitating data-driven decision-making to foster inclusive growth and sustainable development.
- **Investment and Business Planning:** Investors and entrepreneurs can use consumption data to identify potential investment opportunities and design business plans that align with regional consumption trends and demands.

RESULTS AND INTERPRETATION

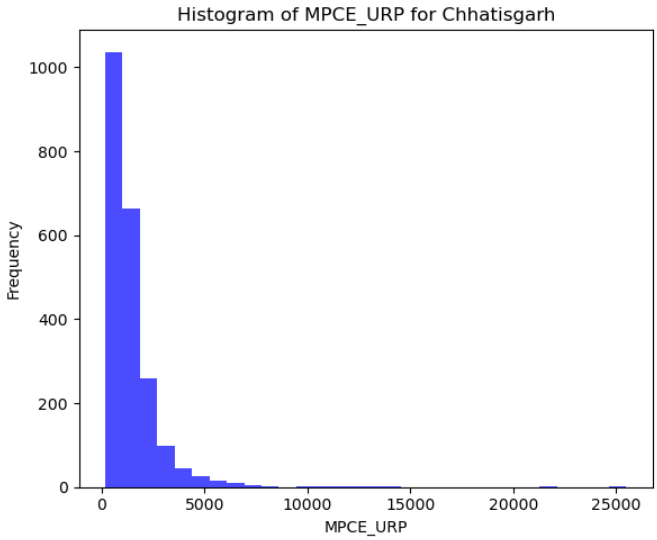
a) Plot a histogram (to show the distribution of total consumption across different districts) and a barplot (To visualize consumption per district with district names) of the data to indicate the consumption district-wise for the state assigned to you.

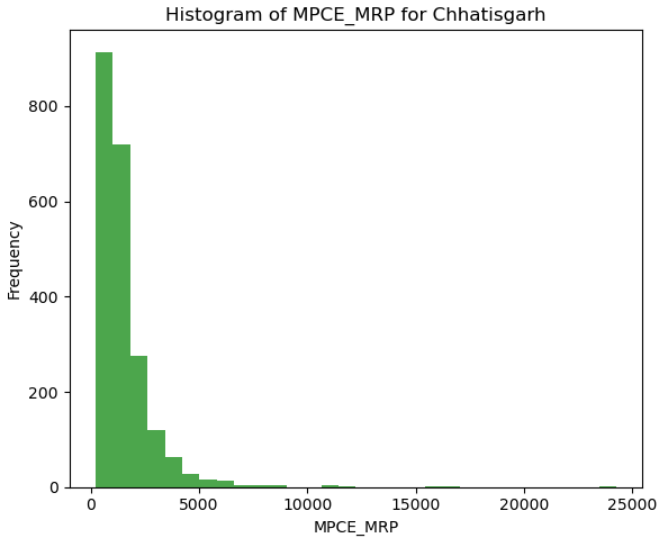
1. Code:

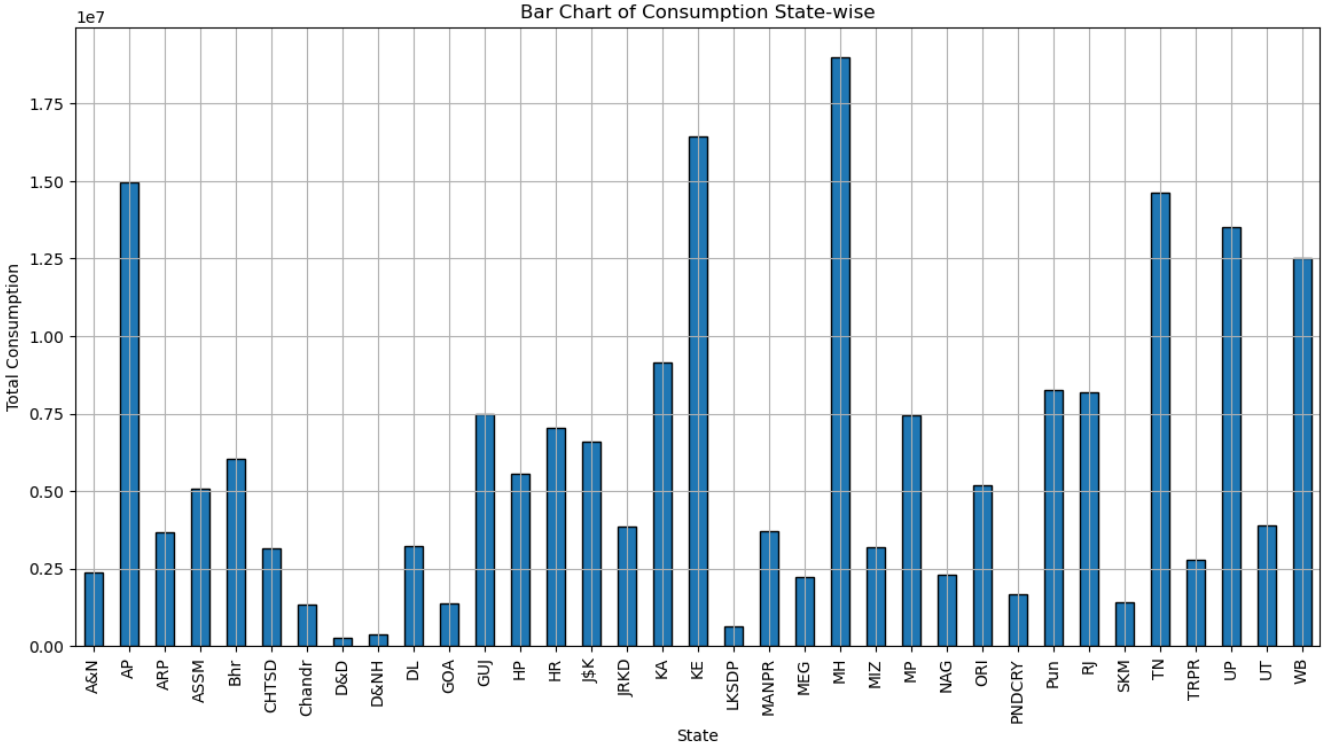
```
import pandas as pd
import matplotlib.pyplot as plt

# Load the data
file_path = 'NSSO68.csv'
data = pd.read_csv(file_path,low_memory=False)

# Inspect the first few rows of the dataset to understand its structure
print(data.head())


# Subset the data with specified columns and filter based on state_1
subset_data = data[['state', 'District', 'MPCE_URP', 'MPCE_MRP', 'state_1']]
filtered_data = subset_data[subset_data['state_1'] == 'CHTSD']


import matplotlib.pyplot as plt

# Plot histograms for MPCE_URP and MPCE_MRP
plt.figure(figsize=(12, 5))

plt.subplot(1, 2, 1)
plt.hist(filtered_data['MPCE_URP'], bins=30, color='blue', alpha=0.7)
plt.title('Histogram of MPCE_URP for Chhatisgarh')
plt.xlabel('MPCE_URP')
plt.ylabel('Frequency')

plt.subplot(1, 2, 2)
plt.hist(filtered_data['MPCE_MRP'], bins=30, color='green', alpha=0.7)
plt.title('Histogram of MPCE_MRP for Chhatisgarh')
plt.xlabel('MPCE_MRP')
plt.ylabel('Frequency')

plt.tight_layout()
plt.show()


# Assuming the relevant columns are named 'state_1' for state names and 'MPCE_URP'
for consumption
state_column = 'state_1'  # Replace with the actual column name for states
consumption_column = 'MPCE_URP'  # Replace with the actual column name for
consumption

# Group the data by state and sum the consumption
state_consumption = data.groupby(state_column)[consumption_column].sum()

# Plot the bar chart
```

```
plt.figure(figsize=(14, 7))
state_consumption.plot(kind='bar', edgecolor='black')
plt.title('Bar Chart of Consumption State-wise')
plt.xlabel('State')
plt.ylabel('Total Consumption')
plt.xticks(rotation=90)
plt.grid(True)
plt.show()
```

2. Result:







3. Interpretation:

The histograms depict the Monthly Per Capita Expenditure (MPCE) for Chhattisgarh, based on Uniform Recall Period (URP) and Mixed Recall Period (MRP):

Histogram of MPCE_URP:

- The distribution is highly skewed to the right, with most households having a low MPCE.
- The majority of households have an MPCE below 5000, with a significant concentration around 1000 to 2000.
- There are very few households with MPCE values exceeding 10,000, indicating a smaller affluent population.

Histogram of MPCE_MRP:

- Similar to the MPCE_URP, the MPCE_MRP distribution is also right-skewed.
- Most households have an MPCE below 5000, with a significant concentration around 1000 to 2000.
- There are very few households with MPCE values exceeding 10,000.

Business Significance:

a. **Market Segmentation:** The data indicates that the majority of the population has a low MPCE. Businesses should focus on products and services that cater to the low-income segment, offering affordable and essential goods.

b. **Targeted Marketing:** Marketing strategies should highlight value for money and affordability, emphasizing the utility and necessity of products.

c. **Product Development:** Developing smaller pack sizes and budget-friendly options can be more attractive to this demographic, ensuring affordability without compromising quality.

For the bar chart:

Andhra Pradesh (AP), Maharashtra (MH), and Tamil Nadu (TN) have the highest total consumption. These states represent significant market opportunities due to their higher consumption levels. States like Gujarat (GUJ), Karnataka (KA), and West Bengal (WB) also show substantial consumption, indicating robust markets but slightly lower than the top states. States like Nagaland (NAG), Sikkim (SKM), and Goa (GOA) have the lowest total consumption, suggesting smaller markets or lower demand.

- **Market Expansion:** Businesses should prioritize expanding operations and marketing efforts in states with high consumption like Andhra Pradesh, Maharashtra, and Tamil Nadu. These states are likely to offer the highest returns on investment.
- **Resource Allocation:** Allocate resources such as marketing budgets, sales teams, and distribution networks strategically to the high-consumption states to maximize market penetration and sales.

b) Depict the consumption on the state map, showing consumption in each district for the state of Karnataka. [NSSO68.csv]

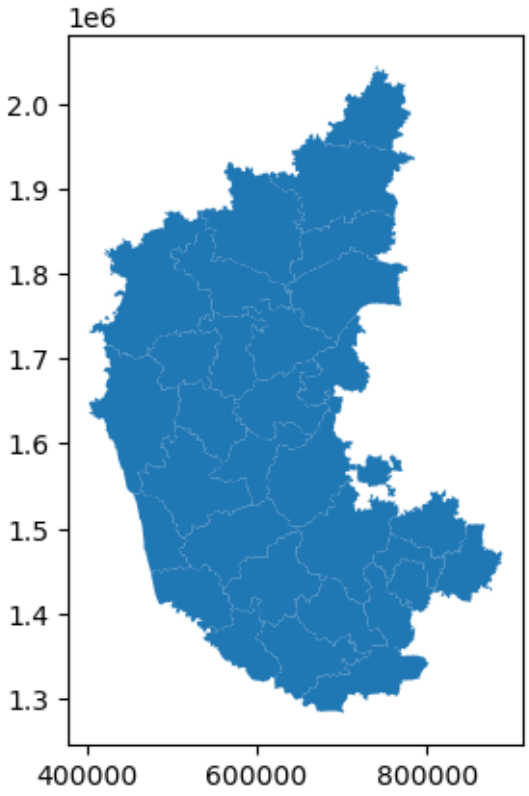
1. Code:

```
#Import GeoPandas library
import geopandas as gpd

# Load the shapefile
gdf_districts = gpd.read_file('District.shp')

gdf_districts.head()

#plot the map
gdf_districts.plot()

import pandas as pd

# Load the NSSO68.csv file
nsso_data = pd.read_csv('NSSO68.csv',low_memory=False)

# Subset the data with specified columns and filter based on state_1
subset_data = nsso_data[['state', 'District', 'MPCE_URP', 'MPCE_MRP', 'state_1']]
filtered_data = subset_data[subset_data['state_1'] == 'KA']

# Load the district-codes.xlsx file
district_codes = pd.read_excel('district-codes.xlsx')

# Filter district codes for Karnataka
karnataka_districts = district_codes[district_codes['state name'] == 'Karnataka']

# Create a mapping from district codes to district names
district_mapping = dict(zip(karnataka_districts['dc'],
karnataka_districts['district name']))

# Replace district codes in the filtered NSSO data with district names
filtered_data['District'] = filtered_data['District'].map(district_mapping)

# Sum all values district-wise
district_wise_sum = filtered_data.groupby('District').sum().reset_index()

# Exclude 'state_1' from the sum result
district_wise_sum = district_wise_sum[['District', 'MPCE_URP', 'MPCE_MRP']]
```

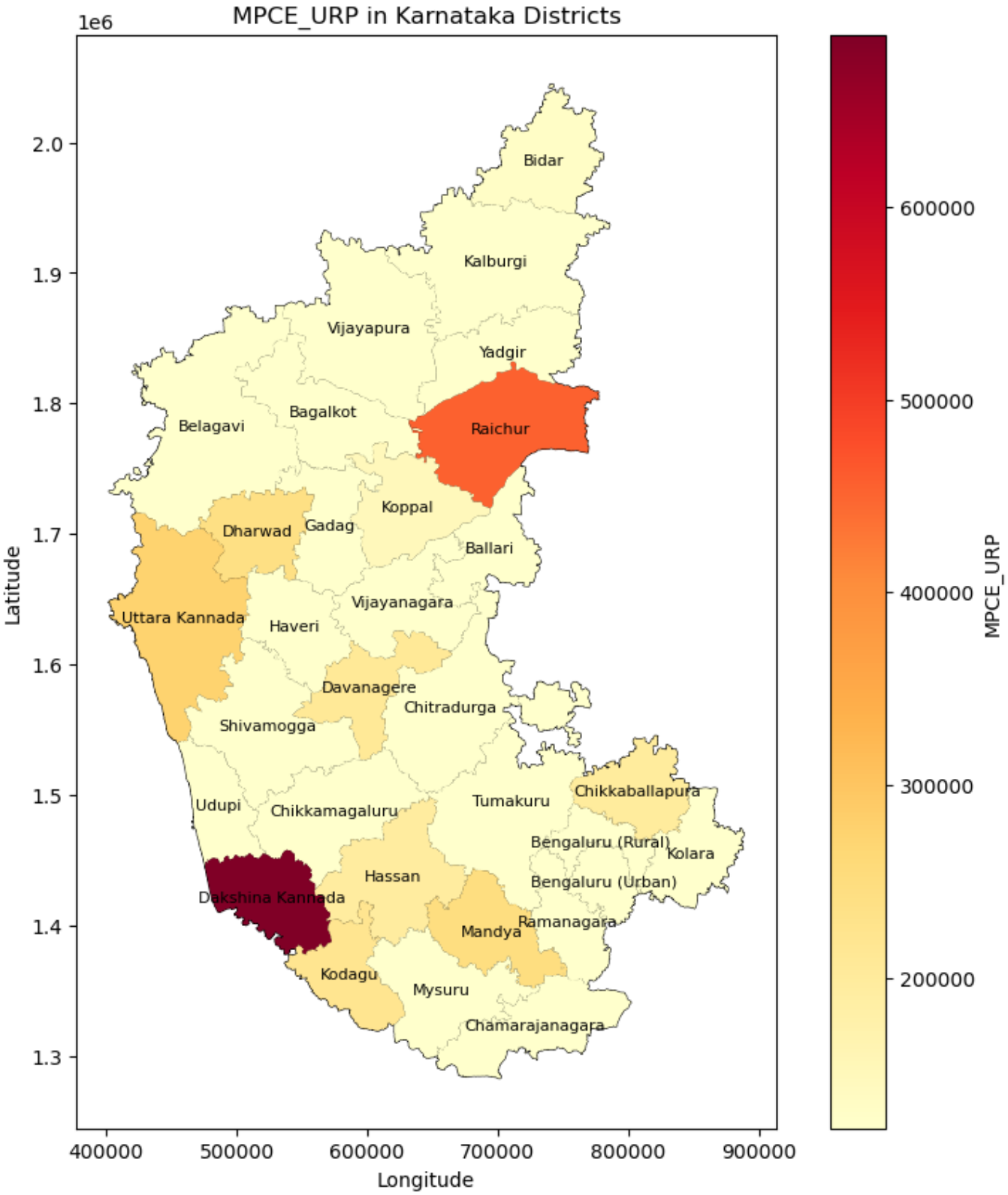
```
# Display the resulting DataFrame
district_wise_sum

df = district_wise_sum

# Merge shapefile with varibale related to Districts
#KGISDist_1 is our Distrcit Name in gdf_districts dataframe
#District is our Distrcit Name in df dataframe
gdf_merged = gdf_districts.merge(df, left_on='KGISDist_1', right_on='District',
how='left')

import matplotlib.pyplot as plt
from matplotlib import colors
import numpy as np

# Define color scale
cmap = plt.cm.get_cmap('YlOrRd')  # Red to green colormap (reversed)
cmap.set_bad('white')  # Set NaN values to white
normalize = colors.Normalize(vmin=gdf_merged['MPCE_URP'].min(),
vmax=gdf_merged['MPCE_URP'].max())

# Plot the map
fig, ax = plt.subplots(figsize=(10, 10))
gdf_districts.plot(ax=ax, facecolor='none', edgecolor='black', linewidth=0.8)  #
Plot the district outlines

# Fill districts with color based on population values
infes_values = gdf_merged['MPCE_URP'].fillna(np.nanmin(gdf_merged['MPCE_URP']) - 1)
# Replace NaN values with a value lower than min
gdf_districts.plot(ax=ax, column=infes_values, cmap=cmap, linewidth=0,
legend=False)

# Add district labels
for x, y, label in zip(gdf_merged.geometry.centroid.x,
gdf_merged.geometry.centroid.y, gdf_merged['KGISDist_1']):
    ax.text(x, y, label, fontsize=8, ha='center', va='center')

# Set plot title and axis labels
ax.set_title('MPCE_URP in Karnataka Districts')
ax.set_xlabel('Longitude')
ax.set_ylabel('Latitude')

# Create and add colorbar
sm = plt.cm.ScalarMappable(cmap=cmap, norm=normalize)
sm.set_array([])
cbar = fig.colorbar(sm)
cbar.set_label('MPCE_URP')

# Show the plot
plt.show()
```

2. Result:

	KGISDistri	LGD_Distri	KGISDist_1	BhuCodeDis	created_us	created_da	last_edite	last_edi_1	SHAPE_STAr	SHAPE_STLe	geometry
0	01	527	Belagavi	01	None	0000/00/00	SURESHBV1	2022-11-24	1.339772e+10	1.141488e+06	MULTIPOLYGON (((537523.31 1865366.861, 537555....
1	02	524	Bagalkot	02	None	0000/00/00	SURESHBV1	2022-09-08	6.561826e+09	6.682456e+05	POLYGON ((581917.898 1811433.959, 581946.875 1...
2	03	530	Vijayapura	03	None	0000/00/00	SURESHBV1	2022-11-24	1.050271e+10	7.032618e+05	POLYGON ((537523.31 1865366.861, 537516.168 18...
3	04	538	Kalburgi	04	None	0000/00/00	SURESHBV1	2022-11-09	1.097395e+10	9.181459e+05	MULTIPOLYGON (((680992.661 1951255.947, 681227...
4	05	529	Bidar	05	None	0000/00/00	SURESHBV1	2022-11-16	5.454415e+09	5.733925e+05	MULTIPOLYGON (((766506.612 1970249.909, 766481...



	District	MPCE_URP	MPCE_MRP
0	Bagalkot *	174637.12	188630.85
1	Bangalore	1971811.41	2097996.99
2	Bangalore Rural	147689.44	167381.05
3	Belgaum	359775.85	386126.91
4	Bellary	268798.33	307397.14
5	Bidar	132880.05	126224.45
6	Bijapur	216815.91	226354.48
7	Chamarajanagar *	223580.93	203602.53
8	Chikkaballapura	197961.96	191656.92
9	Chikmagalur	119842.93	137185.02
10	Chitradurga	121807.68	129759.68
11	Dakshina Kannada	689269.48	673994.71
12	Davanagere	210405.49	217903.53
13	Dharwad	240198.89	267251.33
14	Gadag *	112849.90	117375.49
15	Gulbarga	308941.08	309402.83
16	Hassan	195295.98	218821.13
17	Haveri *	151676.51	146201.05
18	Kodagu	219822.06	233656.29
19	Kolar	401164.07	412801.30
20	Koppal	149191.48	154493.29
21	Mandya	248067.76	257508.17


MPCE_URP in Karnataka Districts
1e6
Latitude
Longitude
MPCE_URP
2.0
1.9
1.8
1.7
1.6
1.5
1.4
1.3
400000
500000
600000
700000
800000
900000
600000
500000
400000
300000
200000
Bidar
Kalburgi
Vijayapura
Yadgir
Bagalkot
Belagavi
Raichur
Koppal
Dharwad
Gadag
Ballari
Vijayanagara
Uttara Kannada
Haveri
Davanagere
Chitradurga
Shivamogga
Udupi
Chikkamagaluru
Tumakuru
Chikkaballapura
Bengaluru (Rural)
Kolara
Hassan
Bengaluru (Urban)
Dakshina Kannada
Ramanagara
Mandya
Kodagu
Mysuru
Chamarajanagara

3. Interpretation:

Shapefiles are commonly used for mapping and spatial analysis due to their compatibility with various GIS software. They consist of at least three mandatory files: a main file (.shp) containing geometry data, an index file (.shx) for indexing, and a dBASE table (.dbf) that stores attribute data in tabular form. Additional optional files can also be included to provide more information.

The displayed shapefile data includes several key attributes for different districts: a district identifier (KGISDistri and KGISDist_1), an administrative code (LGD_Distri), a unique identifier (BhuCodeDis), and metadata about creation and editing dates and users. Additionally, it contains geometrical properties like area (SHAPE_STAr) and perimeter (SHAPE_STLe), and the geometric shapes (geometry) representing the district boundaries.

The map above depicts the Monthly Per Capita Expenditure (MPCE) based on Uniform Reference Period (URP) in various districts of Karnataka. This representation provides a clear visualization of the consumption patterns across different regions within the state. This helps visualize the economic disparity and consumption patterns across the state.

In the map, we can observe that Dakshina Kannada and Raichur districts exhibit the highest levels of MPCE, as denoted by the dark red and orange shades, respectively. Conversely, districts such as Belagavi and Ballari show lower levels of MPCE, represented by the lighter yellow shades. The spatial distribution of expenditure patterns can help in identifying areas of economic disparity and can be crucial for targeted policy interventions.

Purpose:

- The map helps in identifying economic disparities within the state.
- It is useful for policymakers to design targeted interventions to uplift economically backward regions.
- Understanding these patterns can aid in resource allocation and planning for regional development.

RESULTS AND INTERPRETATION

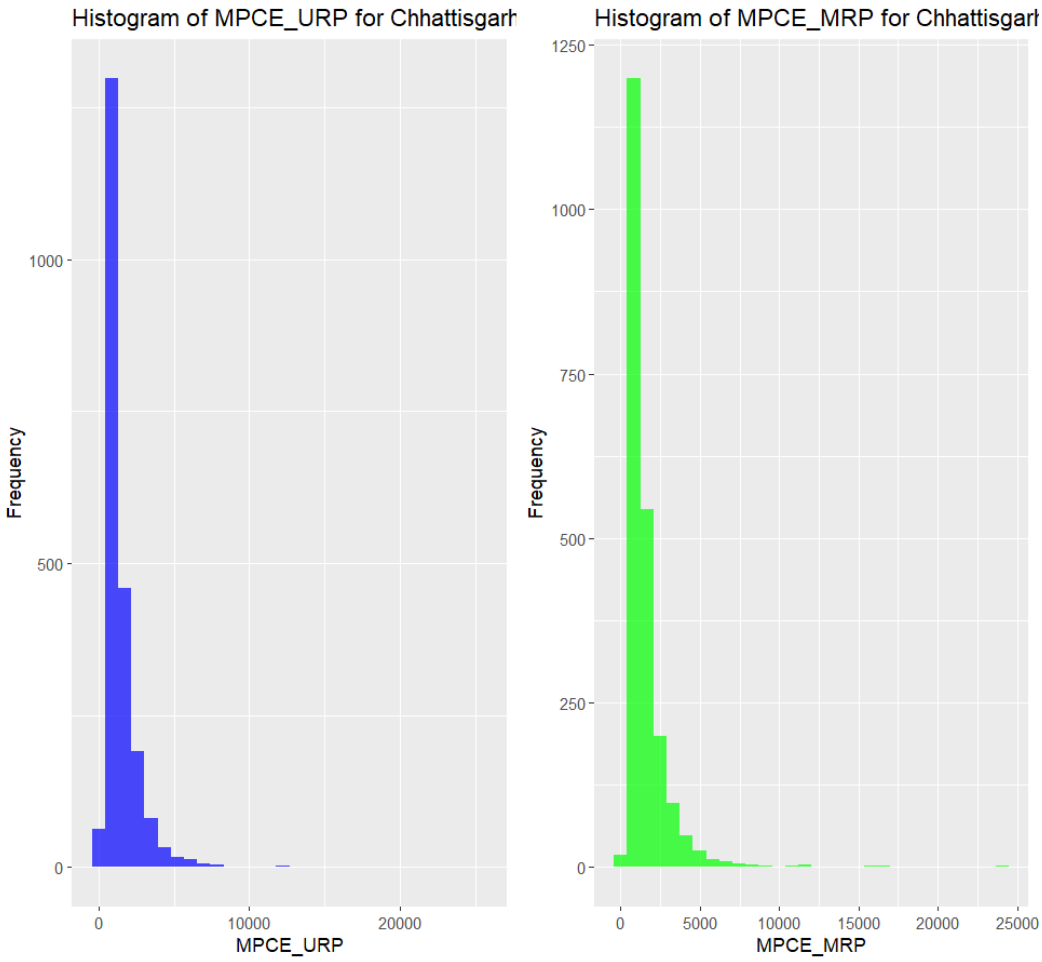
c) Plot a histogram (to show the distribution of total consumption across different districts) and a barplot (To visualize consumption per district with district names) of the data to indicate the consumption district-wise for the state assigned to you.

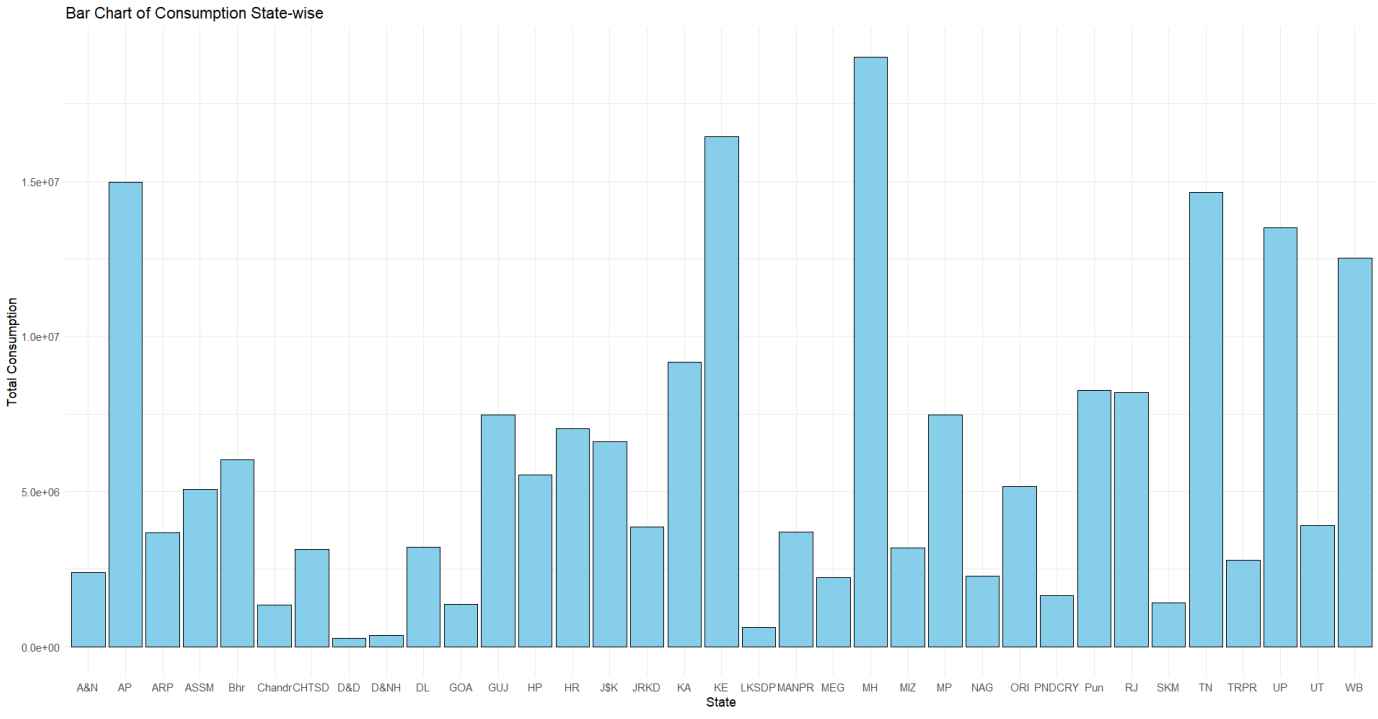
1. Code:

```
# 1. Draw a histogram of the data to indicate the consumption district-wise.

# Function to auto-install and load packages
install_and_load <- function(packages) {
  for (package in packages) {
    if (!require(package, character.only = TRUE)) {
      install.packages(package, dependencies = TRUE)
    }
    library(package, character.only = TRUE)
  }
}

# List of packages to install and load
packages <- c("readr", "ggplot2", "dplyr", "gridExtra")

# Call the function
install_and_load(packages)

# Load the data
file_path <- 'NSSO68.csv'
data <- read_csv(file_path)

# Inspect the first few rows of the dataset to understand its structure
print(head(data))

# Subset the data with specified columns and filter based on state_1
subset_data <- data %>%
  select(state, District, MPCE_URP, MPCE_MRP, state_1)
```

```
filtered_data <- subset_data %>%
  filter(state_1 == 'CHTSD')

# Plot histograms for MPCE_URP and MPCE_MRP
p1 <- ggplot(filtered_data, aes(x = MPCE_URP)) +
  geom_histogram(bins = 30, fill = 'blue', alpha = 0.7) +
  ggtitle('Histogram of MPCE_URP for Chhattisgarh') +
  xlab('MPCE_URP') +
  ylab('Frequency')

p2 <- ggplot(filtered_data, aes(x = MPCE_MRP)) +
  geom_histogram(bins = 30, fill = 'green', alpha = 0.7) +
  ggtitle('Histogram of MPCE_MRP for Chhattisgarh') +
  xlab('MPCE_MRP') +
  ylab('Frequency')

# Arrange the plots side by side
grid.arrange(p1, p2, ncol = 2)

# Group the data by state and sum the consumption
state_consumption <- data %>%
  group_by(state_1) %>%
  summarise(total_consumption = sum(MPCE_URP))

# Plot the bar chart
ggplot(state_consumption, aes(x = state_1, y = total_consumption)) +
  geom_bar(stat = "identity", color = "black", fill = "skyblue") +
  theme(axis.text.x = element_text(angle = 90, hjust = 1)) +
  labs(title = "Bar Chart of Consumption State-wise",
       x = "State",
       y = "Total Consumption") +
  theme_minimal()
```

2. Result:





3. <u>Interpretation</u>:

The histograms for MPCE (Monthly Per Capita Expenditure) using URP (Uniform Reference Period) and MRP (Modified Reference Period) data for Chhattisgarh indicate a highly skewed distribution. Most households fall into the lower expenditure brackets, with a sharp decline in frequency as the expenditure increases. Both histograms reveal that a significant majority of the population in Chhattisgarh has a low monthly per capita expenditure, with only a few households reporting higher expenditures. The MPCE_URP histogram (blue) shows a slightly broader spread compared to the MPCE_MRP histogram (green), indicating that the URP method captures a wider range of expenditure values. The skewed distribution in Chhattisgarh suggests economic disparities, with most households experiencing low levels of consumption. This points to potential areas for targeted economic development and welfare programs.

The bar chart illustrating total consumption state-wise reveals substantial variation in consumption across different states. States like Andhra Pradesh (AP) and Maharashtra (MH) exhibit notably high total consumption levels, suggesting either a larger population, higher per capita consumption, or a combination of both factors. In contrast, states such as Daman and Diu (D&D) and Dadra and Nagar Haveli (D&NH) show minimal total consumption, likely due to their smaller populations. The significant differences in total consumption across states highlight the diverse economic landscapes in India. States with higher consumption may benefit from strong economic activities and better living standards, while those with lower consumption may need more attention in terms of development policies and resource allocation.

d) Depict the consumption on the state map, showing consumption in each district for the state of Karnataka. [NSSO68.csv]

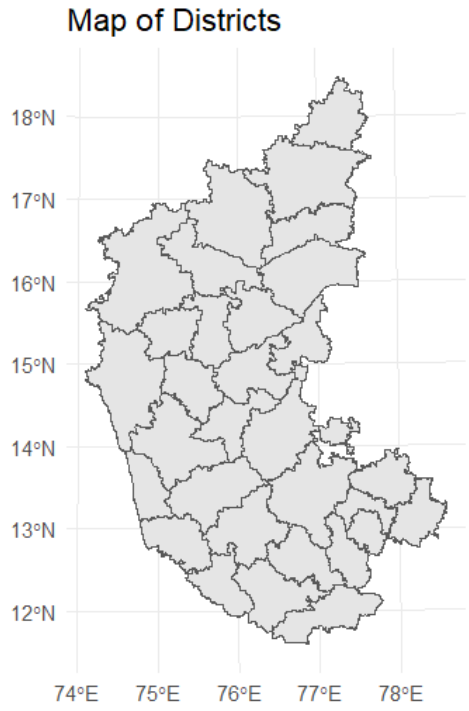
1. Code:

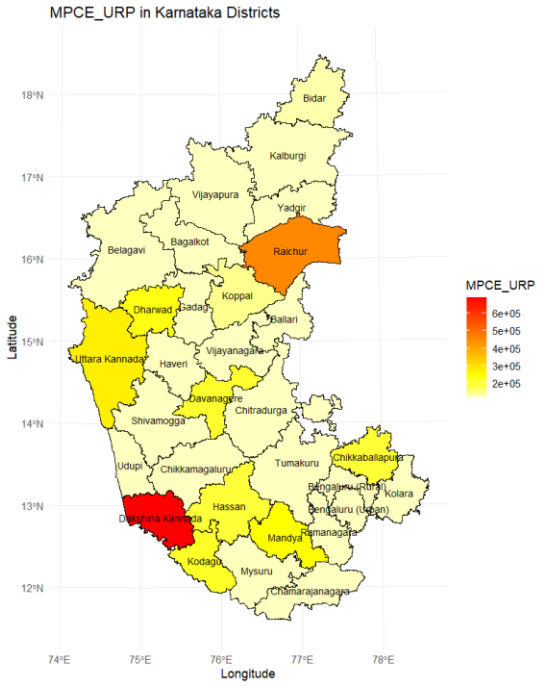
```
# 2.Depict the consumption on the state map, showing consumption in each district.

# Function to auto-install and load packages
install_and_load <- function(packages) {
  for (package in packages) {
    if (!require(package, character.only = TRUE)) {
      install.packages(package, dependencies = TRUE)
    }
    library(package, character.only = TRUE)
  }
}
```

```
# List of packages to install and load
packages <- c("sf", "ggplot2", "dplyr", "readr","readxl","janitor","tidyverse")

# Call the function
install_and_load(packages)

# Load the shapefile
gdf_districts <- st_read('District.shp')

# Display the first few rows
head(gdf_districts)

# Plot the map
ggplot() +
  geom_sf(data = gdf_districts) +
  labs(title = "Map of Districts") +
  theme_minimal()

# Load NSSO68.csv file
nsso_data <- read_csv('NSSO68.csv')

# Subset the data with specified columns and filter based on state_1 (assuming state_1 is equivalent to state in R)
subset_data <- nsso_data %>%
  select(state_1, District, MPCE_URP, MPCE_MRP) %>%
  filter(state_1 == 'KA')  # Filter for Karnataka state

# Load district-codes.xlsx file
district_codes <- read_excel('district-codes.xlsx')
district_codes <- clean_names(district_codes)
names(district_codes)


# Filter district codes for Karnataka
karnataka_districts <- district_codes %>%
  filter(state_name == 'Karnataka')

# Create a mapping from district codes to district names
district_mapping <- karnataka_districts %>%
  select(dc, `district_name`) %>%
  deframe()
```

```
# Replace district codes in the filtered NSSO data with district names
subset_data <- subset_data %>%
  mutate(District = district_mapping[District])

# Sum all values district-wise
district_wise_sum <- subset_data %>%
  group_by(District) %>%
  summarise(MPCE_URP = sum(MPCE_URP), MPCE_MRP = sum(MPCE_MRP))

# Display the resulting DataFrame
print(district_wise_sum)

df <- district_wise_sum

# Merge the shapefile with the DataFrame
gdf_merged <- gdf_districts %>%
  left_join(df, by = c('KGISDist_1' = 'District'))



# Replace NaN values with a value lower than min for visualization purposes
gdf_merged$MPCE_URP[is.na(gdf_merged$MPCE_URP)] <- min(gdf_merged$MPCE_URP, na.rm = TRUE) - 1

# Calculate the centroids
gdf_merged$centroid <- st_centroid(gdf_merged$geometry)

# Extract the coordinates of the centroids
gdf_merged$X <- st_coordinates(gdf_merged$centroid)[, "X"]
gdf_merged$Y <- st_coordinates(gdf_merged$centroid)[, "Y"]

# Plot using ggplot2
ggplot() +
  geom_sf(data = gdf_districts, color = "black", fill = "white") +
  geom_sf(data = gdf_merged, aes(fill = MPCE_URP), color = "black") +
  scale_fill_gradientn(colors = rev(heat.colors(10)), na.value = "white", name = "MPCE_URP") +
  geom_text(data = gdf_merged, aes(x = X, y = Y, label = KGISDist_1), size = 3) +
  labs(title = "MPCE_URP in Karnataka Districts",
       x = "Longitude",
       y = "Latitude") +
  theme_minimal()
```

2. Result:





3. Interpretation:

A shapefile is a widely-used digital vector storage format for storing geometric location and associated attribute information of geographic features. Developed by Esri, a shapefile can store data that represents various geographic features, such as points, lines, and polygons, making it ideal for use in Geographic Information Systems (GIS).

This dataset enables users to perform spatial analysis, create detailed maps, and make data-driven decisions based on geographic patterns and relationships. For instance, it can help in urban

planning, resource management, and environmental monitoring by visualizing and analyzing spatial distributions and relationships among different geographic features.

The map showcases the distribution of Monthly Per Capita Expenditure (MPCE) based on the Uniform Reference Period (URP) across the districts of Karnataka. The data, derived from the NSSO68 survey, has been visualized on the Karnataka state map, highlighting the variations in consumption across different districts. The color gradient from yellow to red represents increasing levels of MPCE, with yellow indicating lower expenditure and red indicating higher expenditure.

1. Color Gradient:
    - Yellow: Represents lower MPCE.
    - Red: Represents higher MPCE.
    - This gradient provides a visual cue to easily identify regions with varying levels of expenditure.
2. High MPCE Districts:
    a. Dakshina Kannada (Dark Red):
        - This district has a high MPCE, indicating a higher average consumption level.
        - Possible reasons include higher income levels, better economic activities, and more developed infrastructure.
    b. Raichur (Orange):
        - Exhibits high MPCE, similar to Dakshina Kannada.
        - Economic activities such as agriculture, power generation, and industrial presence might contribute to this.
3. Low MPCE Districts:
    a. Belagavi (Light Yellow):
        - Shows lower MPCE, indicating lower average consumption.
        - Factors might include lower income levels, less industrialization, and fewer economic opportunities.
    b. Ballari (Light Yellow):
        - Similar to Belagavi in terms of low MPCE.
        - Economic disparities and limited infrastructure development could be contributing factors.