VIRGINIA COMMONWEALTH UNIVERSITY

Statistical analysis and modelling (SCMA 632)

A5: VISUALIZATION – PERCEPTUAL MAPPING FOR BUSINESS

SARATH SABU

V01109792

Date of Submission: 15-07-2024

CONTENTS

Introduction

The National Sample Survey Office (NSSO) is a premier organization under the Ministry of Statistics and Programme Implementation, Government of India. It is responsible for conducting large-scale surveys across India to collect data on various socio-economic parameters. These surveys provide critical insights into the living conditions, consumption patterns, and economic activities of households across the country.

The 68th round of NSSO, conducted during 2011-2012, focused on household consumption expenditure. This survey collected data on the consumption of various items, both food and non-food, across rural and urban areas. The data from this survey helps in understanding the expenditure patterns and the standard of living of households in different regions of India.

In this analysis, we will focus on the state of Madhya Pradesh and Karnataka using data from the 68th round of NSSO.

Objectives

1. **Analyze Consumption Patterns in Madhya Pradesh:**
   - **Histogram:** Create a histogram to show the distribution of total consumption across different districts of Madhya Pradesh. This will help visualize how consumption varies across the state.
   - **Bar Plot:** Generate a bar plot to visualize the total consumption per district in Madhya Pradesh, with district names clearly labeled. This will provide a detailed view of how each district compares in terms of consumption.
2. **Visualize Data on Karnataka State Map:**
   - Select a variable from the dataset and plot it on the Karnataka state map. This spatial visualization will help in understanding the geographical distribution of the selected variable across different districts of Karnataka.

Business Significance

1. **Policy Making:**
   - **Targeted Interventions:** By understanding the consumption patterns across districts, policymakers can design targeted interventions to improve living standards. Districts with lower consumption may require more focused attention and resources.
   - **Resource Allocation:** Effective allocation of resources is crucial for balanced regional development. Identifying regions with higher or lower consumption helps in prioritizing areas for development projects and welfare schemes.
2. **Market Research:**
   - **Identifying Potential Markets:** Businesses can use consumption data to identify potential markets for their products and services. Districts with higher consumption may represent lucrative markets for launching new products or expanding business operations.
   - **Consumer Behavior Analysis:** Understanding consumption patterns helps businesses tailor their marketing strategies to meet the needs and preferences of consumers in different regions.
3. **Investment Decisions:**
   - **Infrastructure Development:** Investors can use this data to make informed decisions about investing in infrastructure projects. Areas with higher consumption might indicate better economic conditions and a higher potential for returns on investment.
   - **Retail and Services:** Retailers and service providers can use this data to plan the expansion of their outlets and services in regions with higher consumption levels.
4. **Social and Economic Research:**
   - **Understanding Regional Disparities:** Researchers can use this data to study regional disparities in consumption and living standards. This helps in identifying the factors contributing to these disparities and in formulating strategies to address them.
   - **Socio-Economic Indicators:** The data serves as an important socio-economic indicator for assessing the overall development and well-being of different regions.

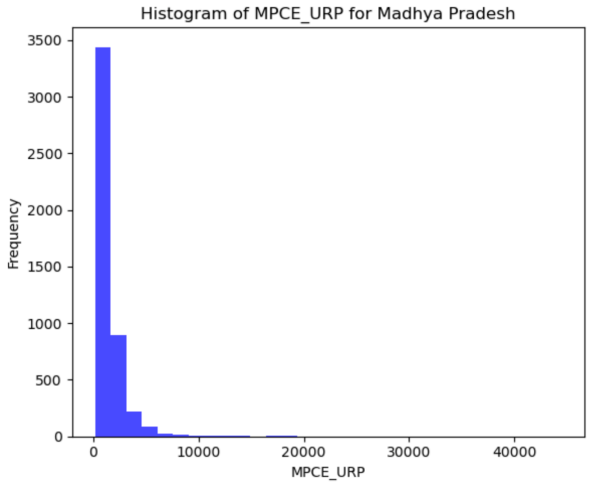
RESULTS AND INTERPRETATIONS

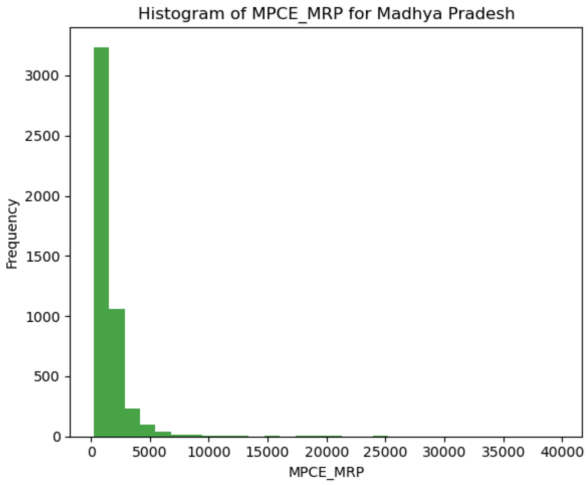
PYTHON

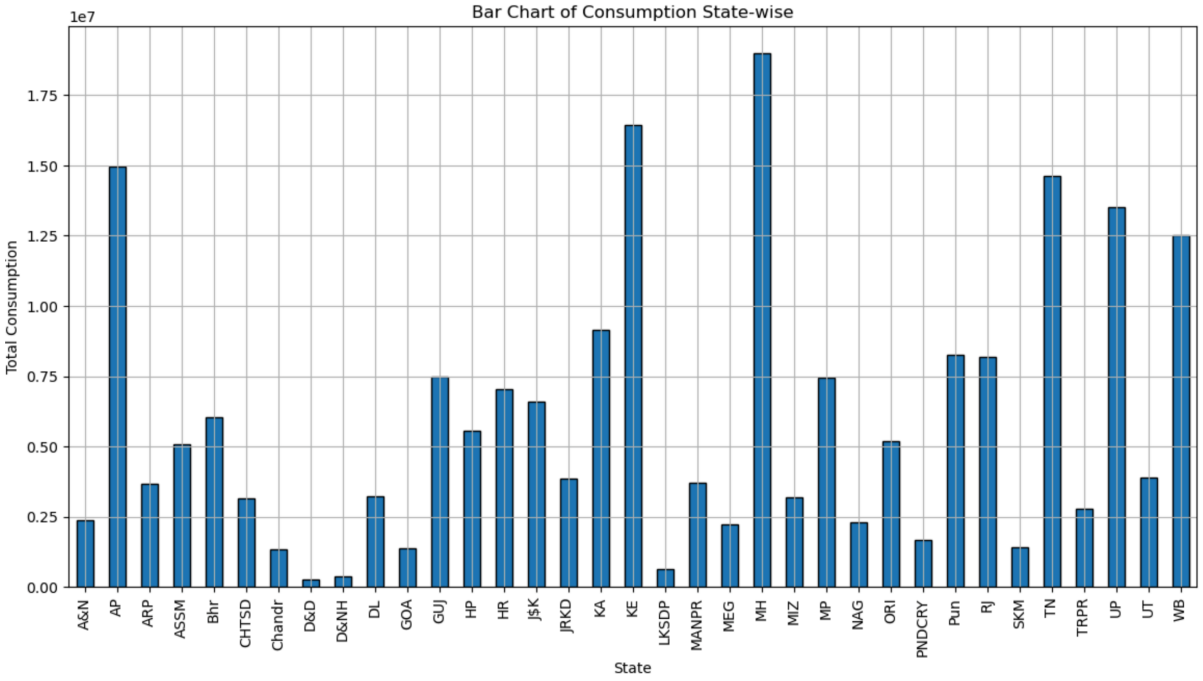
PART A- CODES

```
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 Madhya Pradesh')
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 Madhya Pradesh')
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_1')['MPCE_URP'].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()
```

RESULTS







INTERPRETATION

1. **Histogram of MPCE_URP for Madhya Pradesh (Left Histogram)**
   - **MPCE_URP**: Monthly Per Capita Expenditure based on Uniform Reference Period (URP).
   - **Distribution**: The histogram shows the distribution of MPCE_URP across different districts in Madhya Pradesh. The x-axis represents the MPCE_URP values, and the y-axis represents the frequency of these values.

- **Observation**: The majority of the districts have MPCE_URP values concentrated at the lower end of the spectrum. Very few districts have higher MPCE_URP values, indicating a right-skewed distribution. Most of the consumption values are clustered between 0 and 10,000, with a sharp decline in frequency as the values increase.

2. **Histogram of MPCE_MRP for Madhya Pradesh (Right Histogram)**
   - **MPCE_MRP**: Monthly Per Capita Expenditure based on Mixed Reference Period (MRP).
   - **Distribution**: The histogram shows the distribution of MPCE_MRP across different districts in Madhya Pradesh. The x-axis represents the MPCE_MRP values, and the y-axis represents the frequency of these values.
   - **Observation**: Similar to the MPCE_URP, the MPCE_MRP values are also concentrated at the lower end, with most districts having values between 0 and 5,000. The distribution is also right-skewed, with a sharp decline in frequency as the values increase beyond 5,000.

Interpretation of Bar Chart

- **Bar Chart of Total Consumption State-wise**
  - **Total Consumption**: The y-axis represents the total consumption values for different states, while the x-axis represents the state names abbreviated.
  - **Observation**: The bar chart shows the total consumption across various states in India. Some states, like Maharashtra (MH), Uttar Pradesh (UP), and West Bengal (WB), show significantly higher total consumption compared to others. These states have tall bars indicating higher total consumption values. In contrast, smaller states or states with lower total consumption like Arunachal Pradesh (ARP) and Sikkim (SKM) have shorter bars.
  - **Comparison**: The chart allows for a comparative analysis of total consumption across states, highlighting significant regional differences in consumption patterns.

Insights and Business Implications

1. **Histograms (Madhya Pradesh)**

- The right-skewed distribution in both MPCE_URP and MPCE_MRP indicates that while a few districts have high consumption, the majority have lower consumption levels.
- **Policy Implications**: This suggests a need for targeted economic and social interventions in districts with lower consumption to bridge the disparity.
- **Market Research**: Businesses could focus on districts with higher MPCE values for potential market expansion, while also considering strategies to tap into the lower consumption districts.

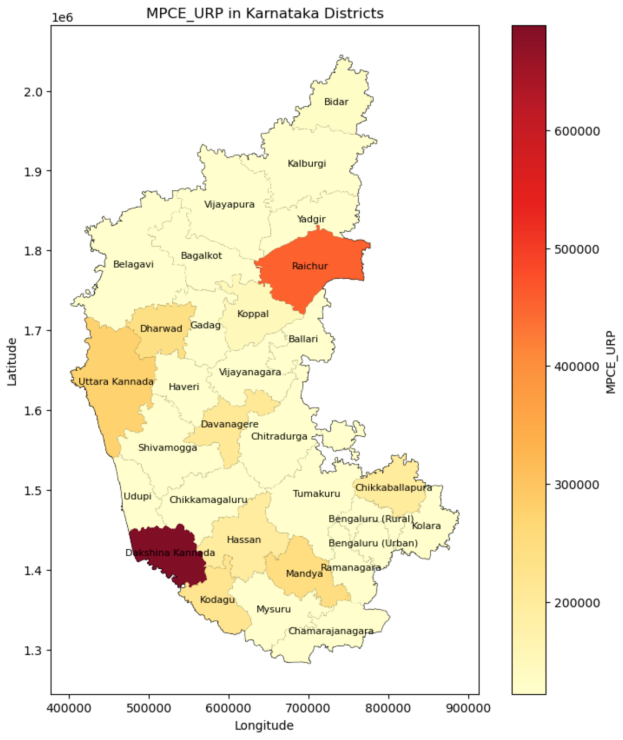
2. **Bar Chart (State-wise)**
   - The significant differences in total consumption across states indicate varied economic activity and living standards.
   - **Investment Decisions**: States with higher total consumption may present better opportunities for investments in infrastructure, retail, and services.
   - **Resource Allocation**: Policymakers can use this data to allocate resources more effectively, ensuring that states with lower consumption receive adequate support to boost economic activities.

PART B-CODES

```
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
```

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

# Define color scale
cmap = plt.colormaps['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, ax=ax)
cbar.set_label('MPCE_URP')

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

RESULTS



	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
22	Mysore	518860.65	546168.69
23	Raichur	454758.00	225172.47
24	Ramanagar	180952.08	189096.87
25	Shimoga	200683.07	212535.34
26	Tumkur	375582.21	374431.64
27	Udupi *	291457.61	308433.24
28	Uttara Kannada	276601.83	273993.15

INTERPRETATION

Shapefiles are essential for mapping and spatial analysis due to their wide compatibility with various GIS software. They consist of three mandatory files: the main file (.shp) containing geometric data, the index file (.shx) for indexing, and a dBASE table (.dbf) that holds attribute data in a tabular format. Additional optional files can also be included to provide more comprehensive information.

Shapefile Attributes

The displayed shapefile data includes several key attributes for different districts:

- **District Identifiers**: KGISDistri and KGISDist_1
- **Administrative Code**: LGD_Distri
- **Unique Identifier**: BhuCodeDis
- **Metadata**: Creation and editing dates and users
- **Geometrical Properties**: Area (SHAPE_STAr) and perimeter (SHAPE_STLe)

- **Geometric Shapes**: Representing district boundaries

Map Interpretation

The map above illustrates the Monthly Per Capita Expenditure (MPCE) based on the Uniform Reference Period (URP) in various districts of Karnataka. This visualization provides a clear view of consumption patterns across different regions within the state, highlighting economic disparities.

Observations:

- **High MPCE**: Dakshina Kannada and Raichur districts show the highest levels of MPCE, indicated by dark red and orange shades, respectively.
- **Low MPCE**: Districts such as Belagavi and Ballari exhibit lower levels of MPCE, represented by lighter yellow shades.

Purpose of the Map

- **Identifying Economic Disparities**: The map helps in pinpointing economic disparities within Karnataka.
- **Policy Design**: It is valuable for policymakers to design targeted interventions to uplift economically disadvantaged regions.
- **Resource Allocation**: Understanding these patterns aids in efficient resource allocation and planning for regional development.

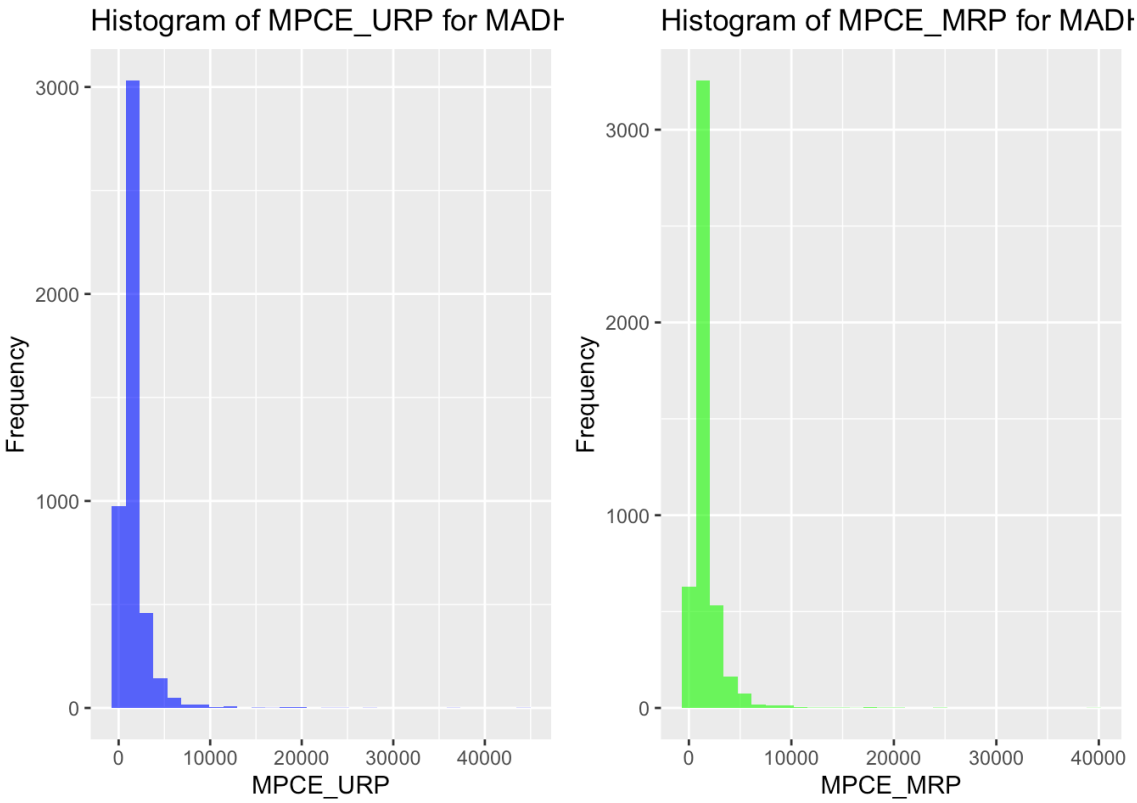
This spatial representation of expenditure patterns is crucial for recognizing areas of economic disparity and can significantly impact policy formulation and regional development strategies.

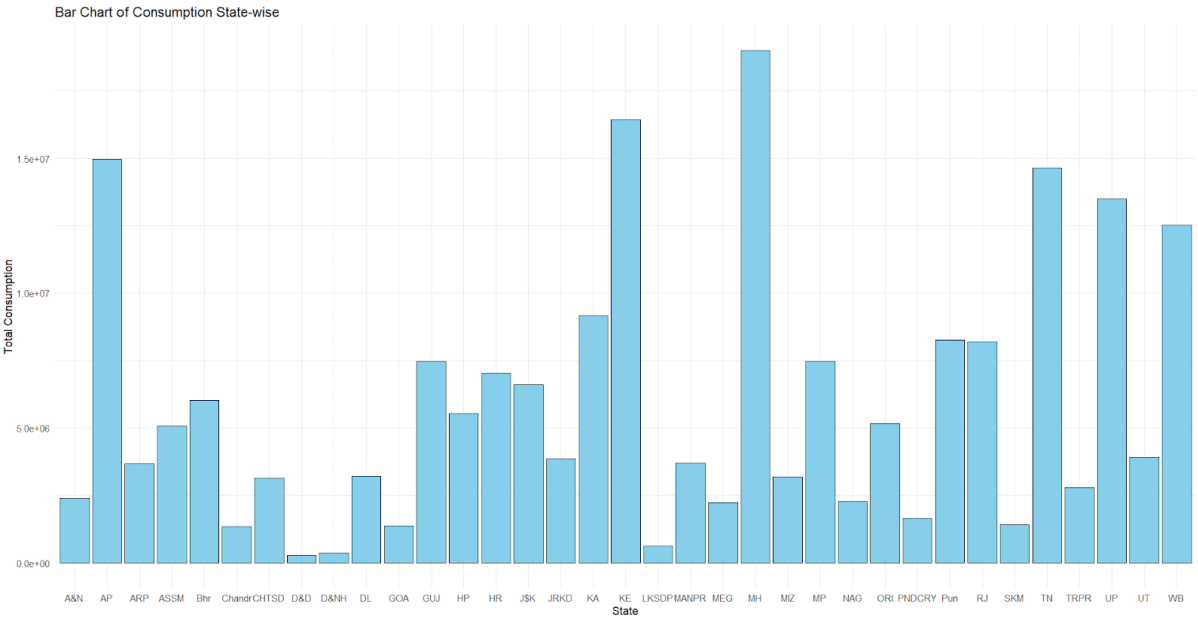
R-PROGRAMMING

PART A – CODES

```
# 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 == 'MP')

# 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 MADHYAPRADESH') +
  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 MADHYAPRADESH') +
  xlab('MPCE_MRP') +
  ylab('Frequency')

# 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()
```

RESULTS





Interpretation of Histograms

1. **Histogram of MPCE_URP for Madhya Pradesh (Left Histogram)**
   - **MPCE_URP**: Monthly Per Capita Expenditure based on Uniform Reference Period (URP).

- **Distribution**: The histogram shows the distribution of MPCE_URP across different districts in Madhya Pradesh. The x-axis represents the MPCE_URP values, and the y-axis represents the frequency of these values.
- **Observation**: Most districts have MPCE_URP values concentrated at the lower end of the spectrum. The majority of the consumption values are clustered between 0 and 10,000, with a sharp decline in frequency as the values increase. This indicates that few districts have high MPCE_URP values, resulting in a right-skewed distribution.

2. **Histogram of MPCE_MRP for Madhya Pradesh (Right Histogram)**
   - **MPCE_MRP**: Monthly Per Capita Expenditure based on Mixed Reference Period (MRP).
   - **Distribution**: The histogram shows the distribution of MPCE_MRP across different districts in Madhya Pradesh. The x-axis represents the MPCE_MRP values, and the y-axis represents the frequency of these values.
   - **Observation**: Similar to MPCE_URP, the MPCE_MRP values are also concentrated at the lower end, with most districts having values between 0 and 5,000. The distribution is right-skewed, with a significant decline in frequency as the values increase beyond 5,000.

Interpretation of Bar Chart

- **Bar Chart of Total Consumption State-wise**
  - **Total Consumption**: The y-axis represents the total consumption values for different states, while the x-axis represents the abbreviated state names.
  - **Observation**: The bar chart shows the total consumption across various states in India. States such as Andhra Pradesh (AP), Karnataka (KA), Maharashtra (MH), Tamil Nadu (TN), and Uttar Pradesh (UP) exhibit significantly higher total consumption compared to other states, indicated by taller bars. In contrast, states like Arunachal Pradesh (ARP) and Mizoram (MZ) show lower total consumption, represented by shorter bars.
  - **Comparison**: This chart highlights substantial regional differences in consumption patterns, with some states showing higher overall consumption levels.

Insights and Business Implications

1. **Histograms (Madhya Pradesh)**

- **Right-Skewed Distribution**: Both MPCE_URP and MPCE_MRP histograms show right-skewed distributions, indicating that while a few districts have high consumption levels, the majority have lower consumption levels.
- **Policy Implications**: This suggests a need for targeted economic and social interventions in districts with lower consumption to address economic disparities.
- **Market Research**: Businesses could focus on districts with higher MPCE values for potential market expansion while considering strategies to penetrate districts with lower consumption.

2. **Bar Chart (State-wise)**
   - **Regional Consumption Differences**: The significant differences in total consumption across states indicate varied economic activities and living standards.
   - **Investment Decisions**: States with higher total consumption, such as Maharashtra and Uttar Pradesh, may present better opportunities for investments in infrastructure, retail, and services.
   - **Resource Allocation**: Policymakers can use this data to allocate resources more effectively, ensuring that states with lower consumption receive adequate support to boost economic activities.

These visualizations provide valuable insights into regional consumption patterns, aiding in making informed decisions for policy-making, business expansion, and resource allocation.

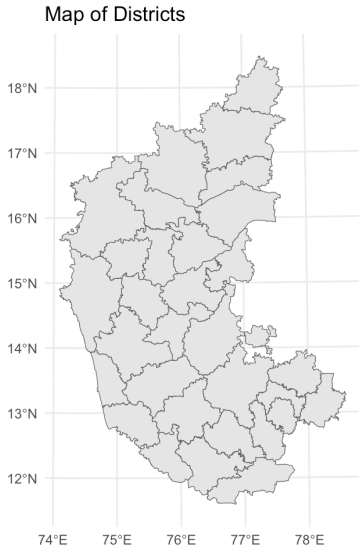
PART B- CODES

```
# Plot the map
ggplot() +
  geom_sf(data = gdf_districts) +
  labs(title = "Map of Districts") +
  theme_minimal()
# Subset the data with specified columns and filter based on state_1 (assum
ing 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
```

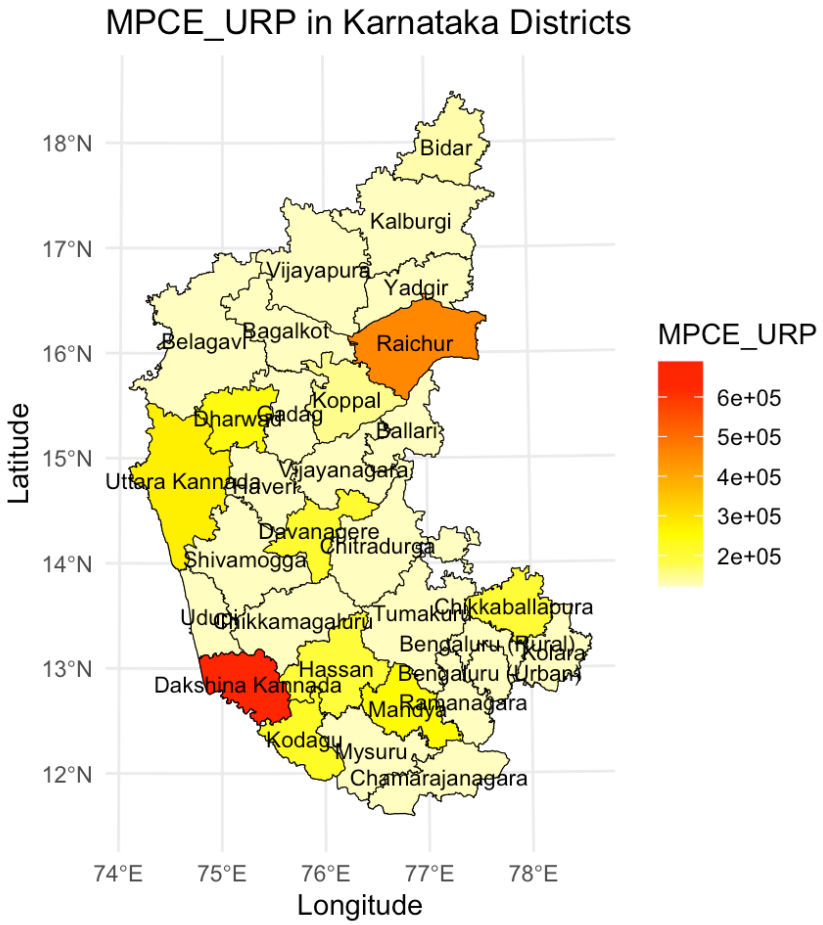
```
district_codes <- read_excel('district-codes.xlsx')
district_codes <- clean_names(district_codes)
names(district_codes)
# Filter district codes for Karnataka
karnataka_districts <- district_codes %>%
  filter(state_name == 'Karnataka')

# 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", n
ame = "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()
```

RESULTS




```
## # A tibble: 29 × 3
##    District         MPCE_URP MPCE_MRP
##    <chr>               <dbl>    <dbl>
##  1 Bagalkot *        174637.  188631.
##  2 Bangalore        1971811. 2097997.
##  3 Bangalore Rural   147689.  167381.
##  4 Belgaum           359776.  386127.
##  5 Bellary           268798.  307397.
##  6 Bidar             132880.  126224.
##  7 Bijapur           216816.  226354.
##  8 Chamarajanagar *  223581.  203603.
##  9 Chikkaballapura   197962.  191657.
## 10 Chikmagalur       119843.  137185.
## # ℹ 19 more rows
```



INTERPRETATIONS

The plot displays the Monthly Per Capita Expenditure (Urban) (MPCE_URP) in various districts of Karnataka, India. The data and the map provide a visual representation of the MPCE_URP values across these districts.

Interpretation:

1. **Top Plot - Map of Districts**:
   - A map of Karnataka divided into its districts.
   - A table (tibble) listing 29 districts with their respective MPCE_URP and MPCE_MRP values.
2. **Bottom Plot - MPCE_URP in Karnataka Districts**:
   - The map highlights the MPCE_URP values for each district using a color gradient.
   - The color gradient ranges from yellow (lower MPCE_URP values) to red (higher MPCE_URP values).

Key Observations:

1. **Highest MPCE_URP**:
   - Districts like Bangalore, which is shown in a dark red color, have the highest MPCE_URP values, indicating higher urban monthly per capita expenditure.
2. **Lowest MPCE_URP**:
   - Districts like Bidar, depicted in light yellow, have lower MPCE_URP values, indicating lower urban monthly per capita expenditure.
3. **Regional Variation**:
   - Northern districts (like Raichur) show a mix of moderate to high MPCE_URP values.
   - Southern districts have a varied distribution with some areas like Dakshina Kannada showing high MPCE_URP.

Specific Data Points:

- **Bagalkot**: MPCE_URP = 174637
- **Bangalore**: MPCE_URP = 1971811
- **Bangalore Rural**: MPCE_URP = 147689

- **Belgaum**: MPCE_URP = 359776
- **Bellary**: MPCE_URP = 268798
- **Bidar**: MPCE_URP = 132880
- **Bijapur**: MPCE_URP = 216816
- **Chamarajanagar**: MPCE_URP = 223581
- **Chikkaballapura**: MPCE_URP = 197962
- **Chikmagalur**: MPCE_URP = 119843

The data and the map together provide a comprehensive view of the economic disparities in urban per capita expenditure across different districts of Karnataka.