VIRGINIA COMMONWEALTH UNIVERSITY

Statistical Analysis and Modeling (SCMA 632)

A5: Visualization – Perceptual Mapping for Business

by

MITHILESH GURUSAMY SIVARAJ

V01107530

Date of Submission: 15-07-2024

CONTENTS

Introduction

Surveys are carried out by the National Sample Survey Office (NSSO) in India to collect extensive socio-economic data, including trends of household spending. At the district level, the NSSO68 dataset sheds light on these trends. The purpose of this analysis is to use the NSSO68 data to depict the consumption trends in the state of Jharkhand. We are able to comprehend the distribution of total consumption among the various districts by generating a bar plot and a histogram. Furthermore, putting the total consumption on the state map of Jharkhand provides a geographical perspective on the patterns in consumption.

Business Insights

- **Consumption Distribution:** Analyzing how consumption is spread across different districts helps to pinpoint areas with varying levels of consumption. This is vital for businesses and policymakers to understand regional differences and effectively target their interventions or marketing strategies.

- **Resource Allocation:** Insight into consumption patterns aids in more efficient allocation of resources. Districts with lower consumption levels may require additional economic activities or welfare programs to enhance their consumption rates.

- **Market Potential:** Businesses can recognize districts with higher consumption, signaling a greater market potential for their products and services.

- **Policy Making:** Policymakers can utilize these insights to design and implement policies that address regional inequalities and encourage balanced economic development.

- **Balanced Growth:** Ensuring balanced growth across districts is essential for the state's overall development. Identifying underperforming districts enables the creation of interventions aimed at reducing regional disparities and fostering inclusive growth.

- **Geospatial Distribution:** Mapping the total consumption provides a clear visual representation of consumption patterns across districts, aiding in the identification of regional disparities within the state.

Visualizations

1. **Histogram of Total Consumption**

The histogram will display the distribution of total consumption across various districts in Jharkhand. This will highlight whether consumption is concentrated around specific values and provide an understanding of the overall range of consumption levels within the state.

- **Purpose:** To analyze the distribution of total consumption among districts.
- **Insight:** Detect districts with particularly high or low consumption and gain an understanding of the general spread of consumption levels.

2. **Bar Plot of Consumption per District**

The bar plot will depict the total consumption for each district, with district names along the x-axis. This visualization will facilitate comparison of consumption levels between districts, revealing which districts have the highest and lowest consumption.

- **Purpose:** To compare total consumption levels across different districts.
- **Insight:** Identify the districts with the highest and lowest consumption, enabling businesses and policymakers to direct their efforts where needed.

3. **Map of Total Consumption in Jharkhand**

Mapping the total consumption onto the Jharkhand state map offers a geographical view of consumption patterns. This visualization will help in recognizing regional clusters of high or low consumption and understanding the spatial distribution of consumption levels.

- **Purpose:** To visualize the geographical distribution of total consumption across Jharkhand.
- **Insight:** Pinpoint regional clusters of high or low consumption, aiding targeted interventions and resource distribution.

Conclusion

Visualizing the NSSO68 data provides significant insights into the consumption patterns across different districts in Jharkhand. The histogram gives an overview of the distribution of total consumption, while the bar plot highlights consumption per district. Mapping the consumption data onto the state map offers a spatial perspective, which can guide more informed business decisions and policymaking. These insights are essential for businesses to identify market potential and for policymakers to address regional disparities and promote balanced economic growth.

ANALYSIS USING “R”:

Setting Working Directory and Loading Libraries

1. **Setting the Working Directory:** The working directory is set to 'D:\CHRIST\Boot camp\DATA', ensuring all file operations (reading and writing) occur in this location.
2. **Loading Required Libraries:** The `install_and_load` function checks if a package is installed, installs it if necessary, and then loads it. This guarantees that all essential libraries are ready for use.

Reading and Filtering Data

3. **Reading the CSV File:** The CSV file "NSSO68.csv" is loaded into a data frame named `data`.
4. **Filtering for JRKD:** The data frame is filtered to include only the rows where `state_1` is "JRKD", representing Jharkhand state.

Displaying Dataset Information

5. **Dataset Information:** The code prints the column names, the first few rows, and the dimensions of the filtered data frame to provide an understanding of its structure and content.
6. **Missing Values Information:** The code calculates and prints the number of missing values in each column of the data frame.

Data Subsetting and Missing Values Imputation

7. **Subsetting the Data:** A new data frame, `jrkdnew`, is created by selecting specific columns relevant for the analysis.
8. **Missing Values in Subset:** The code checks and prints the number of missing values in the subsetted data frame.
9. **Imputing Missing Values:** The columns `Meals_At_Home`, `Meals_Employer`, and `Meals_Payment` are imputed with the mean of their respective columns if they have any missing values.

Outlier Detection and Removal

10. **Removing Outliers:** The code defines a function to remove outliers using the Interquartile Range (IQR) method and applies this function to the `ricepds_v` and `chicken_q` columns.

Summarizing Consumption

11. **Calculating Total Consumption:** A new column, `total_consumption`, is created by summing specified food consumption columns.
12. **Summarizing by District and Region:** The code groups the data by District and Region, summing the total consumption for each group. It prints the top and bottom consuming districts along with the region consumption summary.

Renaming Districts and Sectors

13. **Renaming Districts and Sectors:** The code maps numerical codes to their respective district and sector names based on a predefined mapping.

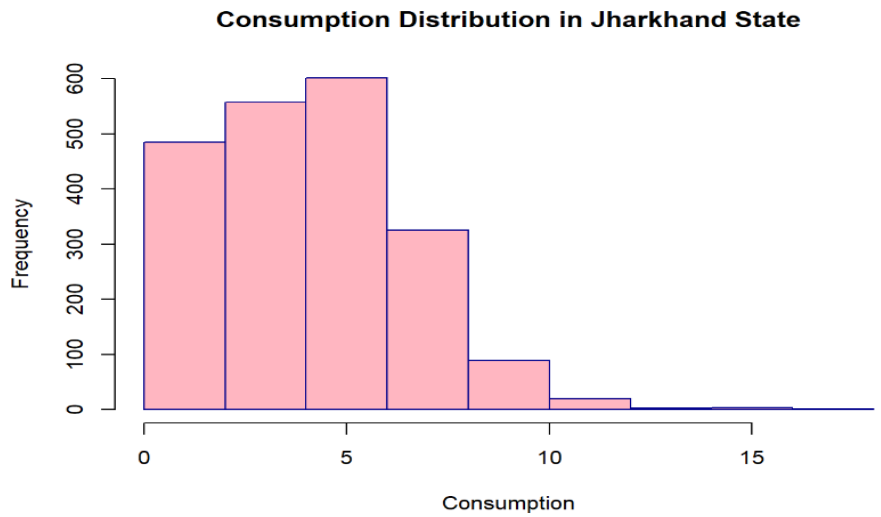
Visualizations

14. **Histogram for Consumption Distribution:** A histogram is plotted to display the distribution of total consumption across different districts, visualizing the frequency of districts within various consumption ranges.
15. **Bar Plot for Total Consumption per District:** A bar plot is created to visualize the total consumption per district, allowing easy comparison of consumption levels across districts.

Map Visualization

16. **Reading GeoJSON File:** The GeoJSON file "JHARKHAND_DISTRICTS.geojson" is loaded into a data frame named `data_map`.
17. **Merging Consumption Data with Map Data:** The consumption data is merged with the map data based on district names.
18. **Plotting Total Consumption on the Map:** A map of Jharkhand is plotted, showing total consumption by district with a gradient color scale from yellow (low consumption) to red (high consumption). District names are displayed on the map.

HISTOGRAM:



1. **Top 3 Consuming Districts**:
    - **District 13**: Total Consumption = 761.55
    - **District 12**: Total Consumption = 756.30
    - **District 4**: Total Consumption = 656.74
2. **Bottom 3 Consuming Districts**:
    - **District 15**: Total Consumption = 175.56
    - **District 16**: Total Consumption = 152.38
    - **District 20**: Total Consumption = 128.89
3. **Region Consumption Summary**:
    - **Region 2**: Total Consumption = 5805.04
    - **Region 1**: Total Consumption = 2760.81

Analysis of the Histogram: Consumption Distribution in Jharkhand State

The histogram displays the consumption distribution in Jharkhand State.

1. **X-axis (Consumption)**:
    - The range of consumption values appears to be from 0 to around 15.
    - The consumption values are divided into bins, each representing a range of consumption values.
2. **Y-axis (Frequency)**:
    - The frequency ranges from 0 to 600.
    - This represents the number of occurrences (or frequency) of consumption values within each bin.
3. **Distribution Shape**:
    - The distribution is right-skewed, meaning that most of the consumption values are concentrated on the left side (lower values) of the distribution.
    - The frequency peaks in the bins representing lower consumption values and gradually decreases as consumption increases.
4. **Frequency Peaks**:
    - The highest frequency is in the range of 5-6, indicating that this is the most common consumption range.
    - There is a noticeable drop in frequency as the consumption values increase beyond this range.
5. **Tail**:
    - The tail on the right side of the distribution extends to higher consumption values but with much lower frequency.

Overall, the histogram suggests that in Jharkhand State, most of the consumption values are relatively low, with fewer instances of higher consumption. This is a common pattern in many distributions where a few instances of higher values extend the range but are not common.

Integrated Analysis

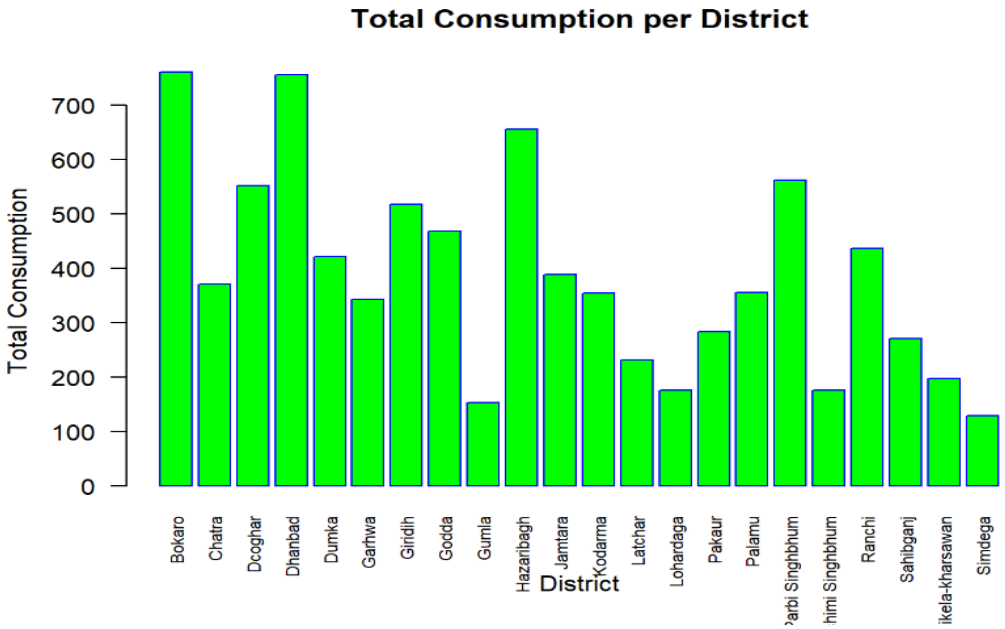
1. **Top Consuming Districts**:
    - The top three consuming districts (13, 12, and 4) are significantly higher in total consumption, suggesting that these districts have either a higher population, greater economic activity, or better access to resources.
2. **Bottom Consuming Districts**:
    - The bottom three consuming districts (15, 16, and 20) show much lower total consumption, indicating potential areas for development and resource allocation to improve living standards and economic activity.
3. **Regional Disparities**:
    - The regional consumption summary highlights a stark contrast between Region 1 and Region 2. Region 2 consumes more than double compared to Region 1, pointing towards significant regional disparities that may be influenced by factors such as population density, industrial activity, or economic policies.

4. **Histogram Insights**:
    - The histogram complements the summarized data by visually demonstrating the distribution of consumption values. The right-skewed distribution aligns with the presence of a few high-consuming districts and regions, while most districts have moderate to low consumption values.
5. **Policy Implications**:
    - Policymakers can leverage this integrated analysis to design targeted interventions for the bottom-consuming districts, aiming to boost their consumption levels and reduce disparities.
    - The significant consumption in top districts and Region 2 can be further analyzed to replicate their success factors in other areas.

Recommendations

1. **Focused Development Programs**:
    - Implement development programs in the bottom-consuming districts to improve infrastructure, access to resources, and economic activities.
2. **Resource Allocation**:
    - Allocate more resources to Region 1 to bridge the consumption gap between the two regions, promoting balanced regional development.
3. **Further Investigation**:
    - Conduct detailed studies on the high-consuming districts and Region 2 to identify best practices and success factors that can be applied to other districts and regions.
4. **Monitoring and Evaluation**:
    - Establish a monitoring and evaluation framework to track consumption patterns over time and assess the impact of interventions.

BARPLOT:

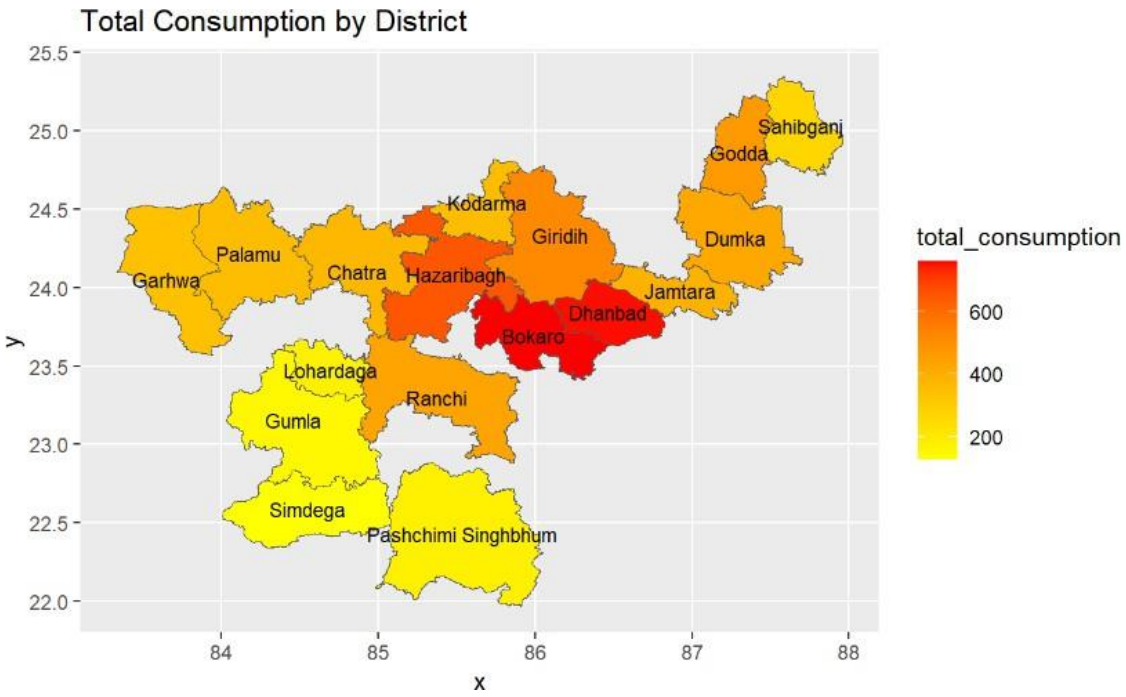


1. **Districts with Highest Consumption**:
    - Bokaro and Hazaribagh have the highest total consumption, both exceeding 700 units.
    - Deoghar and Giridih also show relatively high consumption, around 550 units.
2. **Districts with Moderate Consumption**:
    - Several districts like Chatra, Dhanbad, Garhwa, Dumka, Jamtara, Kodarma, and Palamu show moderate consumption levels, ranging between 400 to 500 units.
    - Pakur, Pashchimi Singhbhum, and Ranchi have consumption levels between 400 to 500 units as well.
3. **Districts with Lower Consumption**:
    - Districts such as Godda, Gumla, Lohardaga, Latehar, and Simdega exhibit lower consumption levels, with Gumla and Simdega showing the lowest values around 100 to 200 units.
    - Lohardaga and Latehar also have low consumption, falling between 200 to 300 units.
4. **General Trends**:
    - The consumption levels vary significantly across districts, with some showing very high values and others showing much lower values.
    - There isn't a clear regional pattern visible from the chart alone, indicating that consumption levels might be influenced by district-specific factors rather than geographic location.

Recommendations:

- **Further Analysis**:
    - It might be useful to look into the factors contributing to the high consumption in districts like Bokaro and Hazaribagh. This could include population size, industrial activity, or economic status.
    - Similarly, understanding the reasons for low consumption in districts like Gumla and Simdega could provide insights for policy-making or resource allocation.

MAP PLOTTING:



1. **High Consumption Districts**:
    - **Bokaro and Dhanbad**: These districts are shaded in dark red, indicating the highest total consumption levels, above 600 units.
    - **Hazaribagh**: This district is also prominently high in consumption, although slightly less than Bokaro and Dhanbad.

2. **Moderate Consumption Districts**:
    - **Giridih, Kodarma, Jamtara, Chatra, Godda, and Sahibganj**: These districts are shaded in orange, indicating moderate consumption levels, ranging between 400 to 600 units.
    - **Palamu and Ranchi**: These districts fall into the moderate consumption category as well.

3. **Low Consumption Districts**:
    - **Simdega, Gumla, Pashchimi Singhbhum**: These districts are shaded in yellow, indicating the lowest consumption levels, below 200 units.
    - **Lohardaga and Garhwa**: These districts are also in the low consumption range.

Geographical Insights:

- **Central and Eastern Regions**: The central region, including Bokaro and Hazaribagh, shows significantly higher consumption levels. This might indicate higher population density, industrial activities, or better infrastructure.
- **Southern and Western Regions**: The southern regions, such as Simdega and Gumla, show much lower consumption, which could be due to lower population density, less industrialization, or other socio-economic factors.
- **North-Eastern Regions**: Moderate consumption is observed in the north-eastern parts, like Sahibganj and Godda.

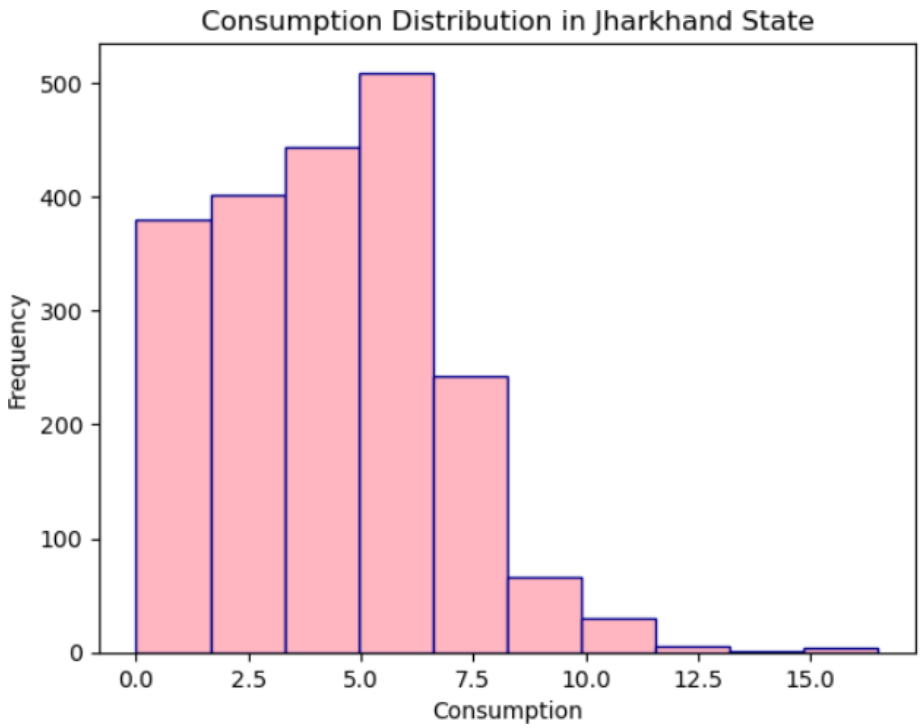
Recommendations:

1. **Investigate High Consumption Areas**:
    - Understanding the factors contributing to high consumption in Bokaro, Dhanbad, and Hazaribagh could provide insights into effective resource utilization and economic activities driving the demand.
2. **Focus on Low Consumption Areas**:
    - Analyzing the reasons behind low consumption in districts like Simdega and Gumla might help in identifying developmental needs and potential for infrastructure improvement.
3. **Policy Implications**:
    - Tailoring policies to address the specific needs of high and low-consumption areas could optimize resource distribution and economic development efforts.
    - Enhancing infrastructure and industrial opportunities in low-consumption areas might help in balancing the consumption levels across the districts.
4. **Further Research**:
    - Combining this data with other socio-economic indicators like population density, income levels, and industrial presence could provide a more comprehensive understanding of consumption patterns.

ANALYSIS USING "PYTHON":

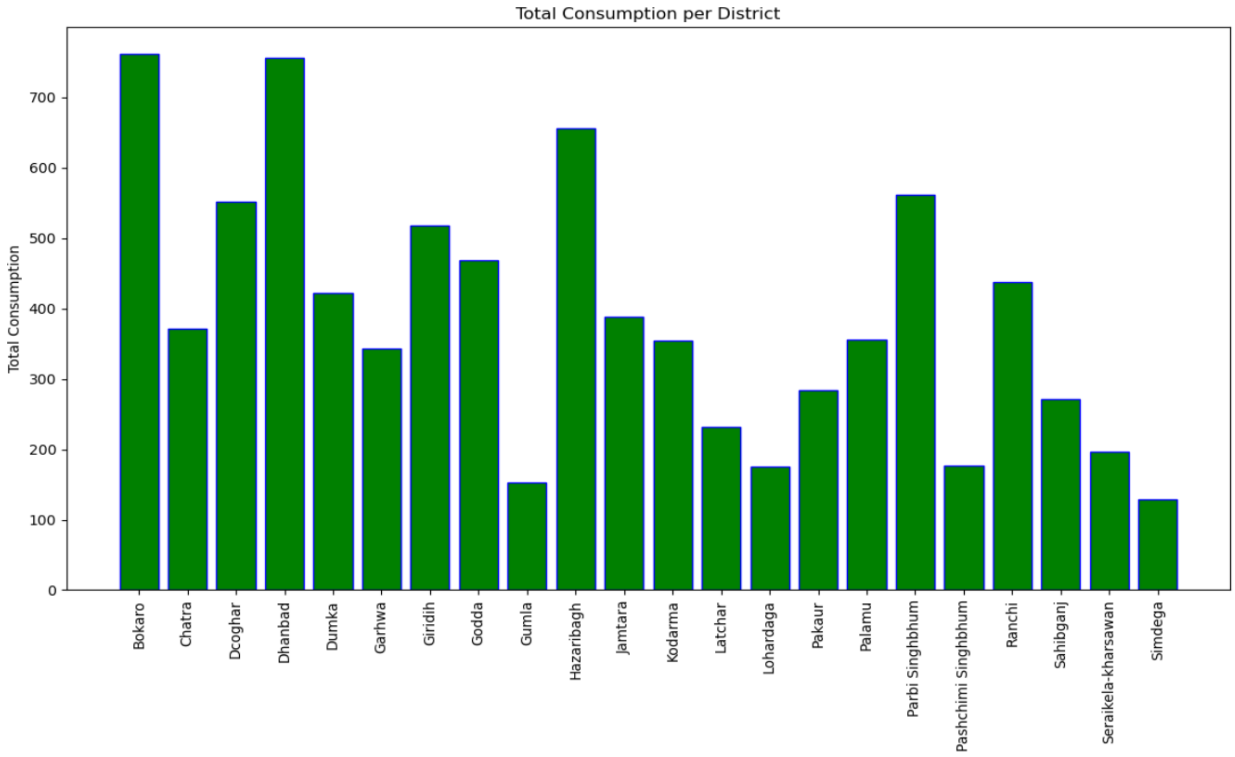
1. **Setting Working Directory and Loading Libraries**:
   - The working directory is set to ensure all file operations are performed in the correct folder.
   - Required libraries are installed and loaded to ensure all necessary functionalities are available for data manipulation, visualization, and geospatial operations.
2. **Reading and Filtering Data**:
   - The data is read from a CSV file and filtered to include only rows where `state_1` is "JRKD", representing the state of Jharkhand.
3. **Displaying Dataset Information**:
   - The structure of the dataset is displayed, including column names, the first few rows, and the dimensions. This provides an overview of the data.
4. **Missing Values**:
   - Missing values are identified and displayed. This is crucial for understanding the completeness of the dataset.
   - Missing values in certain columns are imputed with the mean of the respective columns to ensure there are no gaps in the data.
5. **Outlier Detection and Removal**:
   - Outliers in specified columns are identified using the IQR method and removed. This ensures the analysis is not skewed by extreme values.
6. **Summarizing Consumption**:
   - Total consumption is calculated for each row by summing specified columns.
   - The data is grouped by district and region to calculate total consumption for each group. This helps identify areas with high and low consumption.
7. **Renaming Districts and Sectors**:
   - Numerical codes for districts and sectors are mapped to their respective names for better readability and understanding.
8. **Visualizations**:
   - A histogram is plotted to show the distribution of total consumption across districts, helping to visualize the frequency distribution.
   - A barplot is created to show the total consumption per district, allowing for easy comparison of consumption levels across different districts.
   - A map is plotted to visualize the total consumption by district, highlighting geographical patterns in consumption across Jharkhand.

HISTOGRAM:



1. **Distribution Shape**:
    - The histogram shows a right-skewed distribution. Most of the consumption values are concentrated on the lower end of the scale, with fewer observations at higher consumption levels.
2. **Frequency**:
    - The highest frequency is observed around the 5.0 consumption mark, with around 500 observations.
    - There is a gradual decline in the number of observations as the consumption value increases beyond 5.0.
3. **Range and Spread**:
    - Consumption values range from 0.0 to approximately 15.0.
    - The majority of observations fall between 0.0 and 7.5, indicating that most of the consumption values are relatively low.
4. **Outliers**:
    - There are a few observations with consumption values greater than 10.0, which can be considered as outliers given their low frequency compared to the rest of the data.

BARPLOT:

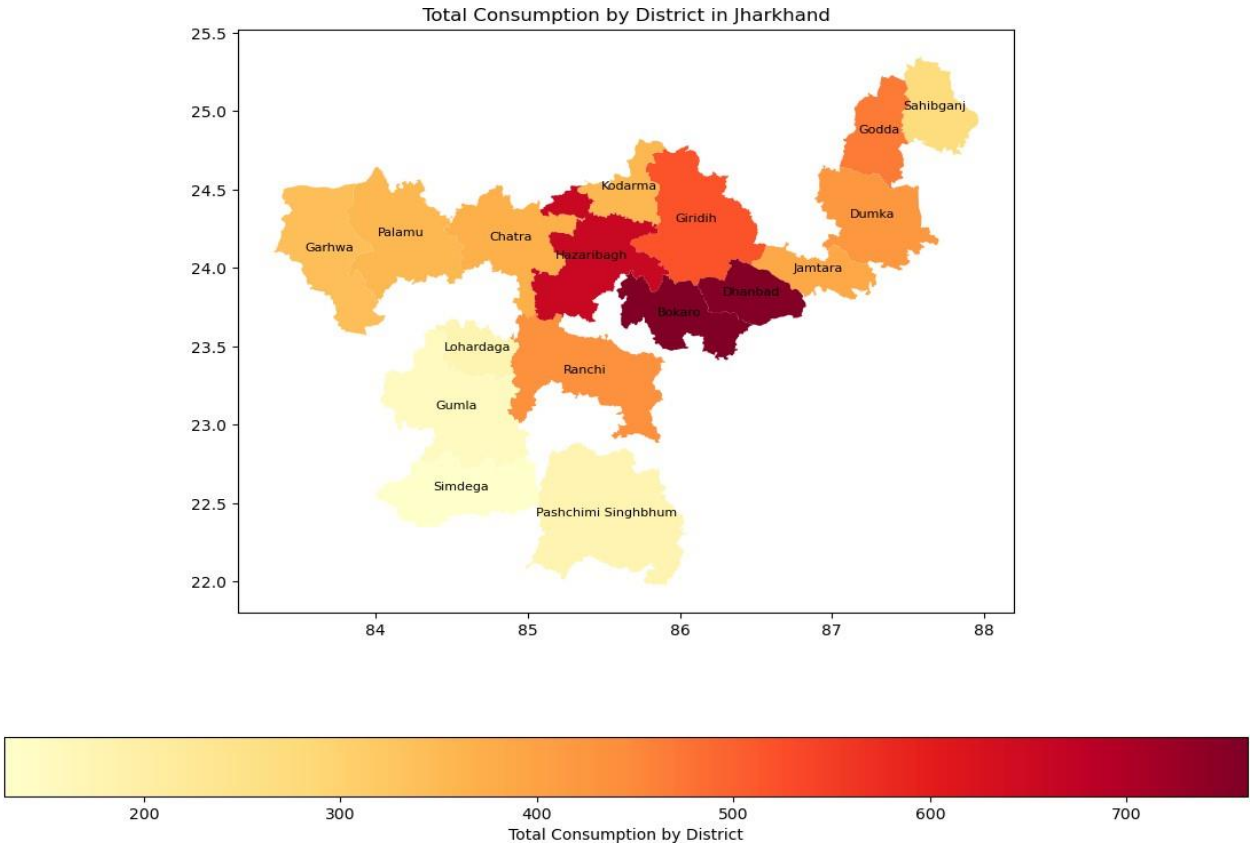


The bar chart shows total consumption per district in Jharkhand State:

1. **Highest Consumption**:
    - Bokaro and Dhanbad exceed 700 units.
    - Pashchimi Singhbhum is close to 600 units.
2. **Moderate Consumption**:
    - Districts like Chatra, Deoghar, Giridih, Hazaribagh, Jamtara, Kodarma, and Ranchi range between 400-500 units.
3. **Lowest Consumption**:
    - Gumla, Pakur, Palamu, and Simdega are below 200 units.

This indicates significant variation in consumption across the districts.

MAP PLOTTING:



This map shows total consumption by district in Jharkhand State using a color gradient:

1. **Highest Consumption**:
    - Bokaro and Dhanbad are highlighted in dark red, indicating the highest consumption (600-700 units).
2. **Moderate Consumption**:
    - Districts like Hazaribagh, Giridih, and Ranchi show moderate consumption with shades of orange (400-500 units).
3. **Lowest Consumption**:
    - Simdega, Gumla, and Pashchimi Singhbhum are in light yellow, indicating the lowest consumption (200-300 units).

This map visually emphasizes the regional disparities in consumption within Jharkhand, with the central and northeastern districts showing higher consumption levels.