VIRGINIA COMMONWEALTH UNIVERSITY

STATISTICAL ANALYSIS & MODELING

A5: VISUALIZATION – PERCEPTUAL MAPPING FOR BUSINESS USING PYTHON AND R

MOHITH KUMAR
V01106540

Date of Submission: 15/07/2024

CONTENTS

Content:	Page no:
INTRODUCTION	3
OBJECTIVE	3
BUSINESS SIGNIFICANC	3-4
RESULTS AND INTERPRETATIONS	4-18

VISUALIZATION – PERCEPTUAL MAPPING FOR BUSINESS USING PYTHON AND R

INTRODUCTION

In analyzing the dataset "NSSO68.csv," the focus is on using statistical models to uncover meaningful insights and relationships within the data. This report explores three key methodologies: logistic regression, decision tree analysis, and Tobit regression. Logistic regression is employed to predict binary outcomes by modeling the probability of a categorical response variable based on one or more predictor variables. Decision trees offer a non-linear approach by recursively partitioning data into subsets, making them interpretable and suitable for complex decision-making processes. Tobit regression addresses scenarios involving censored data, where the dependent variable is constrained by upper or lower limits, providing robust estimation methods essential in fields like economics and healthcare. By comparing these methodologies, this report aims to demonstrate their applicability and effectiveness in extracting valuable insights from the "NSSO68.csv" dataset, thereby contributing to informed decision-making in various research and practical domains.

OBJECTIVES

- Plot a **histogram** (to show the distribution of total consumption across different districts) and a **barplot** (To visualize consumption per district with district names) to indicate the consumption district-wise for the Chattisgarh state.
- Plot for any variable on the **Chattisgarh** state map using NSSO68.csv data

BUSINESS SIGNIFICANCE

Visualization and perceptual mapping are powerful tools for businesses to gain insights and make informed decisions. Visualization simplifies complex data into easy-to-understand graphical formats, helping identify trends, patterns, and outliers. It enhances communication with stakeholders and supports data-driven decision-making while enabling real-time performance monitoring. Perceptual mapping, on the other hand, allows businesses to understand their market positioning relative to competitors by visualizing consumer perceptions. It helps identify market gaps, informs competitive analysis, guides product development, and shapes effective marketing strategies. Together, these tools provide a comprehensive understanding of market dynamics and consumer preferences, leading to better strategic planning and execution.

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) to indicate the consumption district-wise for the Chattisgarh state.

R Code:

```
# Summarize consumption
chtsdnew$total_consumption <- rowSums(chtsdnew[, 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 <- chtsdnew %>%
    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("1" = "Koriya","2" = "Surguja","3" = "Jashpur","4" = "Raigarh","5" =
"Korba","6" = "Janjgir - Champa","7" = "Bilaspur","8" = "Kawardha","9" = "Rajnandgaon","10" =
"Durg","11" = "Raipur","12" = "Mahasamund","13" = "Dhamtari","14" = "Kanker","15" =
"Bastar","16" = "Dantewada","17" = "Narayanpur","18" = "Bijapur")
sector_mapping <- c("2" = "URBAN", "1" = "RURAL")

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

View(chtsdnew)

hist(chtsdnew$total_consumption, breaks = 10, col = 'blue', border = 'black',
```

```
    xlab = "Consumption", ylab = "Frequency", main = "Consumption Distribution in Chattisgarh State")

CHTSD_consumption <- aggregate(total_consumption ~ District, data = chtsdnew, sum)
View(CHTSD_consumption)
??barplot
barplot(CHTSD_consumption$total_consumption,
        names.arg = CHTSD_consumption$District,
        las = 2, # Makes the district names vertical
        col = 'blue',
        border = 'black',
        xlab = "District",
        ylab = "Total Consumption",
        main = "Total Consumption per District",
        cex.names = 0.6) # Adjust the size of district names if needed
```

Python Code:

```
 # Summarize consumption
chtsdnew['total_consumption'] = chtsdnew[['ricepds_v', 'Wheatpds_q', 'chicken_q', 'pulsep_q', 'wheatos_q']].sum(axis=1)

# Summarize and display top consuming districts and regions
def summarize_consumption(df, group_col):
    summary = df.groupby(group_col)['total_consumption'].sum().reset_index().sort_values(by='total_consumption', ascending=False)
    return summary

district_summary = summarize_consumption(chtsdnew, 'District')
region_summary = summarize_consumption(chtsdnew, 'Region')

print("Top Consuming Districts:\n")
print(district_summary.head(4))
print("Region Consumption Summary:\n")
print(region_summary)
```

Results

```
Top Consuming Districts:

    District  total_consumption
10        11        1530.338289
9         10        1503.413102
1          2        1367.023922
6          7         967.668297
Region Consumption Summary:

   Region  total_consumption
1       2        8800.393114
2       3        2369.273470
0       1        1817.004637
```

```
Top Consuming Districts:
> print(head(district_summary, 4))
# A tibble: 4 × 2
  District total
     <int> <dbl>
1       11 1530.
2       10 1503.
3        2 1367.
4        7  968.
> cat("Region Consumption Summary:\n")
Region Consumption Summary:
> print(region_summary)
# A tibble: 3 × 2
  Region total
   <int> <dbl>
1      2 8800.
2      3 2369.
3      1 1817.
```

Interpretation:

The above figures show the summary of the districts and region based on total consumption. The top consuming districts are displayed here. District with the most consumption is 11 with 1530 units. The same is with region. With region 2 being the highest of 8800 units.

b) Plot for any variable on the **Chattisgarh** state map using NSSO68.csv data

R Code:

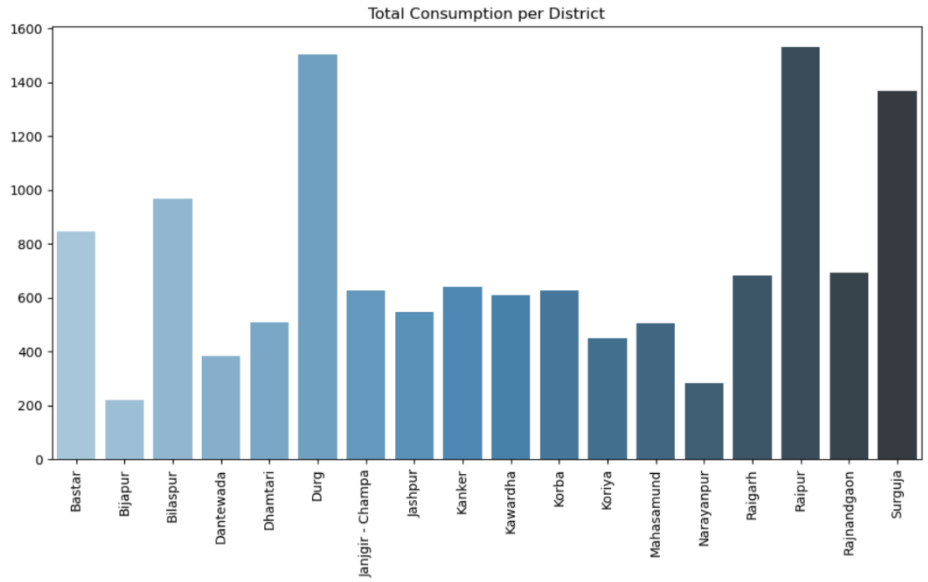
```
data_map <- st_read("E:\\R\\Assignment 5\\CHHATTISGARH_DISTRICTS.geojson")
View(data_map)

data_map <- data_map %>%
  rename(District = dtname)
colnames(data_map)
data_map_data <- merge(CHTSD_consumption,data_map,by = "District")
View(data_map_data)
ggplot(data_map_data) +
  geom_sf(aes(fill =total_consumption, geometry = geometry)) +
  scale_fill_gradient(low = "yellow", high = "red") +
  ggtitle("Total Consumption_by_District")

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:

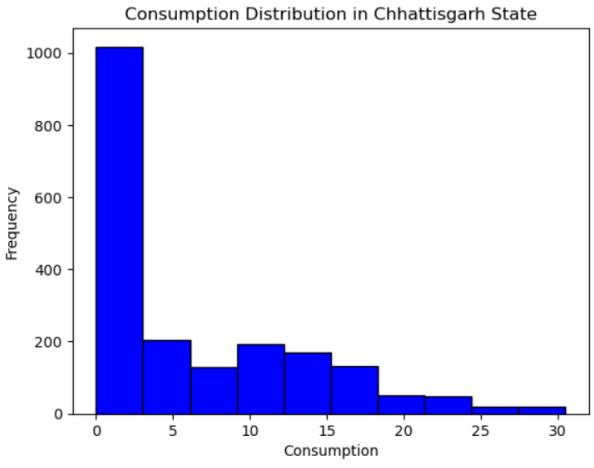
```
# Plot using GeoPandas and Matplotlib
fig, ax = plt.subplots(1, 1, figsize=(15, 10))
data_map_data.plot(column='total_consumption', ax=ax, legend=True, cmap='YlOrRd')
plt.title("Total Consumption by District")
plt.show()

# Annotate districts with their names
fig, ax = plt.subplots(1, 1, figsize=(15, 10))
```

```
data_map_data.plot(column='total_consumption', ax=ax, legend=True, cmap='YlOrRd')
data_map_data.apply(lambda x: ax.annotate(text=x['District'], xy=x.geometry.centroid.coords[0], ha='center'), axis=1)
plt.title("Total Consumption by District")
plt.show()
```

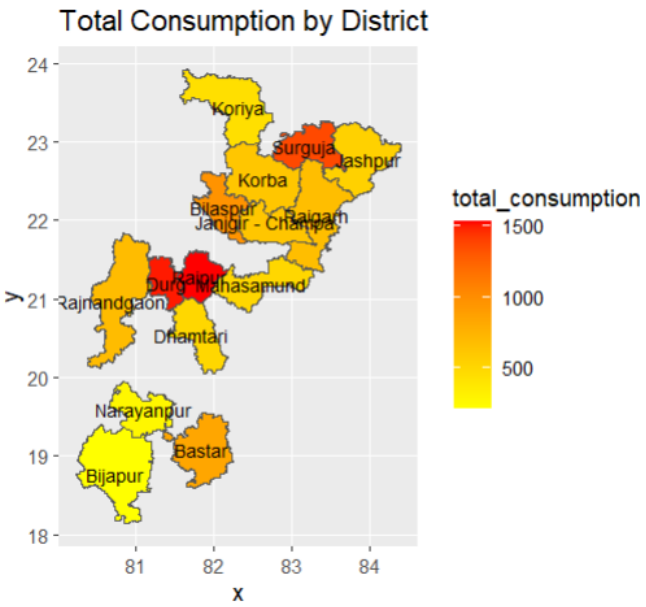
Result:



Bar Plot



Histogram



Interpretation:

- Histogram of Consumption Distribution in Chhattisgarh State

The histogram illustrates the distribution of consumption across Chhattisgarh State. The x-axis represents consumption levels, while the y-axis represents the frequency of each consumption level. The histogram shows that the majority of the population has low levels of consumption, with a steep decline in frequency as consumption levels increase. Most of the consumption values are concentrated in the lower ranges, with very few individuals exhibiting higher consumption levels. This indicates a skewed distribution where a small portion of the population consumes significantly more than the rest.

- Bar Chart of Total Consumption per District

The bar chart compares the total consumption across different districts in Chhattisgarh. Each bar represents a district, and its height corresponds to the total consumption in that district. The chart reveals significant variability in consumption levels across districts. Durg stands out with the highest total consumption, followed by Raipur and Surguja. On the other end, districts like Bijapur and Narayanpur show much lower total consumption. This variation highlights the economic disparities and consumption patterns within the state, suggesting that some districts have higher resource availability or better economic conditions than others.

- Choropleth Map of Total Consumption by District

The choropleth map visually represents the total consumption per district within Chhattisgarh. Districts

are color-coded based on their total consumption, with a gradient from yellow (lower consumption) to red (higher consumption). The map vividly shows that districts like Durg, Raipur, and Surguja have higher consumption levels, indicated by the red shading. In contrast, districts like Bijapur and Narayanpur are shaded in yellow, denoting lower consumption levels. This geographic representation provides a spatial understanding of consumption distribution, revealing regional disparities and potentially guiding resource allocation and policy-making to address these inequalities.

These visualizations provide a comprehensive understanding of consumption patterns in Chhattisgarh. The histogram highlights a skewed distribution with most people having low consumption levels. The bar chart and choropleth map emphasize significant regional disparities, with certain districts exhibiting much higher consumption levels. These insights can inform government and organizational strategies to address resource distribution, economic development, and targeted interventions in districts with lower consumption.

Conclusion

These visualizations provide a comprehensive understanding of consumption patterns in Chhattisgarh. The histogram highlights a skewed distribution with most people having low consumption levels. The bar chart and choropleth map emphasize significant regional disparities, with certain districts exhibiting much higher consumption levels. These insights can inform government and organizational strategies to address resource distribution, economic development, and targeted interventions in districts with lower consumption.