Name: Syed Muhammad Askari

Roll No :232386

Summited to: Sir Safdar Ali



**Introduction:**

Geographic data often involves analyzing regional differences in demographic variables such as income, education, employment, and population distributions. However, collecting data for large populations can be resource-intensive and time-consuming. **Cluster sampling**, a cost-effective statistical technique, enables analysts to work with representative subsets while maintaining accuracy.

This summary focuses on the application of cluster sampling to a dataset of gender-based employment demographics in Pakistan for 2023. The dataset explores key indicators across provinces, divisions, and districts. Using clustering methods, we aim to reduce the data collection effort while effectively identifying regional patterns in demographic variables.

**Why use Cluster Sampling?**

**Cluster sampling is particularly effective for this analysis because of the following reasons:**

1. **Geographical Grouping:**

Provinces, divisions, and districts represent natural clusters in the dataset. Each cluster contains demographic and income-related data that can be studied as a group. This makes it easier to identify regional differences.

1. **Cost and Time Efficiency:**

Instead of analyzing every individual or district, cluster sampling allows us to focus on representative clusters. This reduces the amount of data to process while still capturing the overall trends within each province.

1. **Diversity Within Clusters:**

Provinces and districts are diverse, with significant variations in income, gender distribution, and employment levels. By sampling clusters, we can better understand these variations and identify patterns that are unique to specific regions.

1. **Better Representation:**

Clusters provide a more comprehensive view of regional disparities compared to random sampling. For example, clustering ensures that every province is included, avoiding underrepresentation of smaller or less populated provinces like Balochistan.

1. **Data Comparisons:**

Using clusters, we can compare provinces effectively to see how indicators such as employment and income vary, making it easier to draw actionable conclusions for policy development.

**2. Objectives and Dataset Description**

**Objective:**

• To analyze geographic patterns in demographics, focusing on regional differences in labor force, employment, and population distribution.

• To demonstrate how clustering can reduce data collection while preserving accuracy.

**Dataset Overview:**

The dataset includes the following columns:

• **Province, Division, District:** Geographic hierarchy of the data.

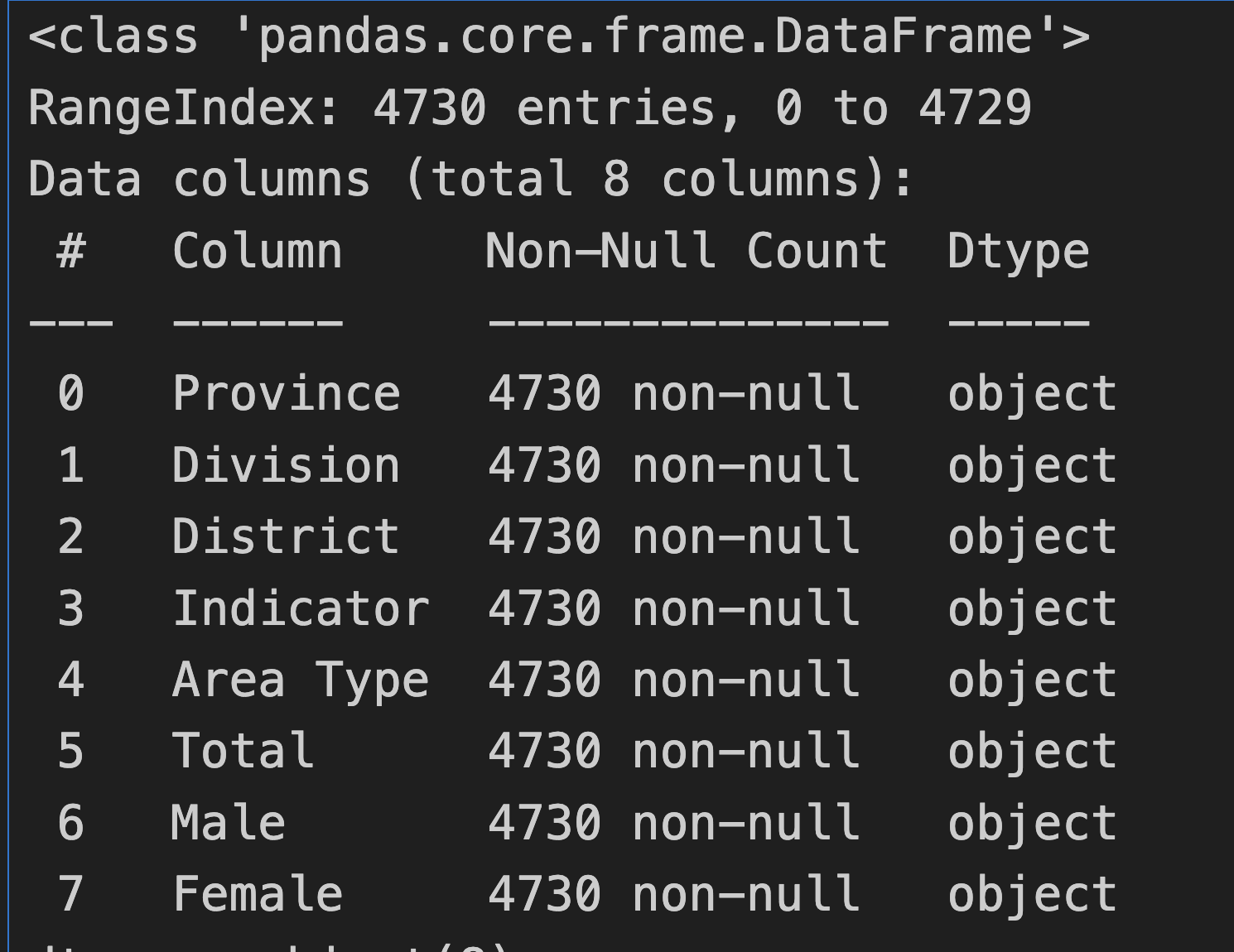
• **Indicator:** Key demographic categories (e.g., total population, labor force, employed individuals).

• **Area Type:** Rural or urban classification.

• **Demographic Variables:** Total, Male, and Female populations.

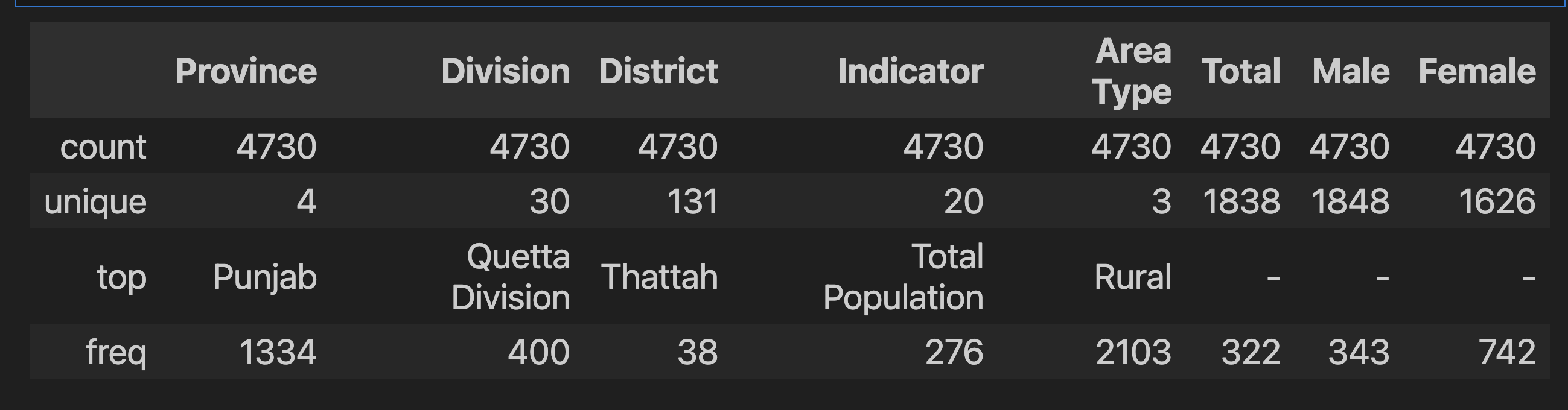
The dataset covers 4,730 entries across multiple provinces in Pakistan, representing both urban and rural areas.

df.info()



**5 point Summary Report:**

df.describe()

****

**3. Methodology**

**Cluster Sampling Process:**

Cluster sampling involves grouping a population into clusters, sampling a subset of clusters, and analyzing them to infer patterns for the whole population. In this study, KMeans clustering is employed to identify patterns among geographic regions.

1. **Data Grouping:**

• Data is grouped by provinces to allow province-wise cluster analysis.

• Numeric columns (Total, Male, and Female) are prepared for clustering by converting them to numeric types and handling missing values.

2. **Cluster Formation:**

• For each province, KMeans clustering is applied to form five clusters.

• Each cluster represents a group of districts or divisions with similar demographic profiles.

3. **Cluster Labeling:**

• Cluster labels are generated and appended to the dataset.

• These labels provide insights into demographic similarities within each cluster.

4. **Data Visualization:**

• After clustering, visualizations (scatter plots and bar charts) are created to represent the demographic distribution across clusters.

Code : Snippet

df['Male'] = pd.to\_numeric(df['Male'], errors='coerce')

df['Female'] = pd.to\_numeric(df['Female'], errors='coerce')

average\_income = df.groupby('Province')[['Male', 'Female']].mean()

average\_income.plot(kind='bar', figsize=(10, 6))

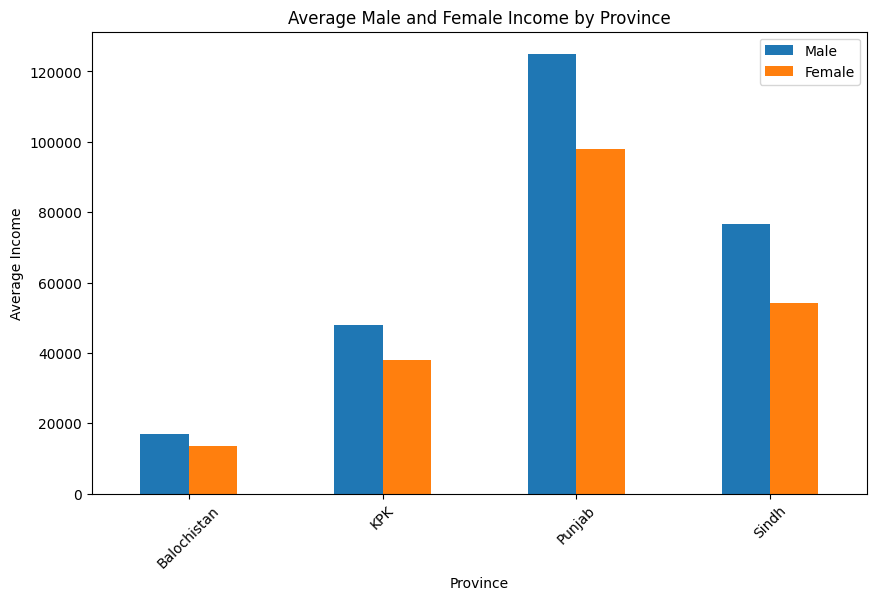
plt.title('Average Male and Female Income by Province')

plt.xlabel('Province')

plt.ylabel('Average Income')

plt.xticks(rotation=45)

plt.show()



plt.figure(figsize=(12, 6))

sns.countplot(data=df, x='Province', hue='Cluster')

plt.title('Distribution of Clusters by Province')

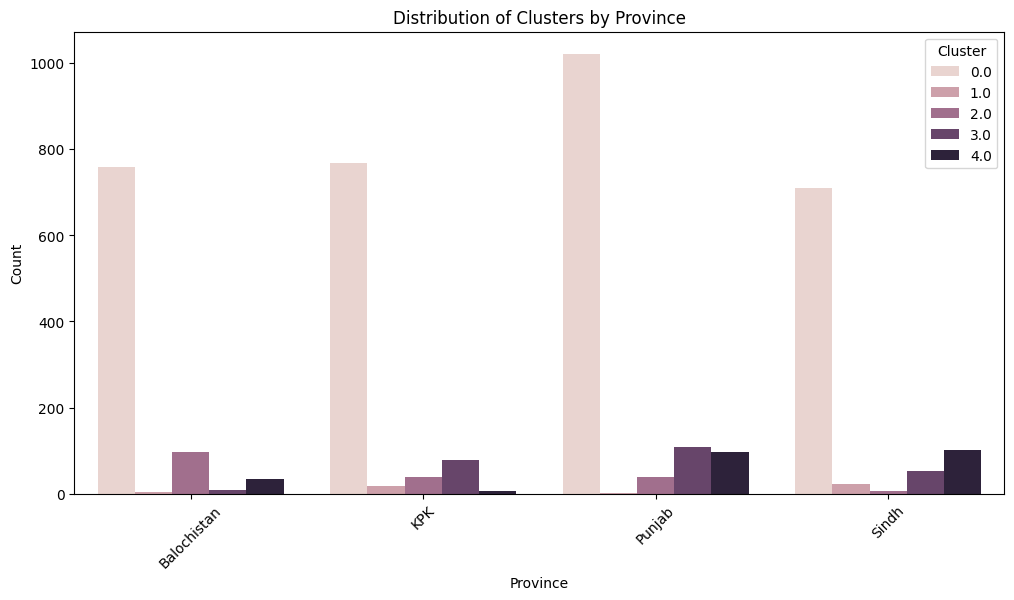
plt.xlabel('Province')

plt.ylabel('Count')

plt.xticks(rotation=45)

plt.legend(title='Cluster')

plt.show()



# Calculate the total number of employed people in each province

total\_employed = employment\_data.groupby('Province')[['Male', 'Female']].sum()

# Calculate the percentage of male and female employed in each province

total\_employed['Total'] = total\_employed['Male'] + total\_employed['Female']

total\_employed['Male\_Percentage'] = (total\_employed['Male'] / total\_employed['Total']) \* 100

total\_employed['Female\_Percentage'] = (total\_employed['Female'] / total\_employed['Total']) \* 100

# Plot the percentage of male and female employed in each province

plt.figure(figsize=(12, 6))

total\_employed[['Male\_Percentage', 'Female\_Percentage']].plot(kind='bar', stacked=True, figsize=(12, 6))

plt.title('Percentage of Male and Female Employed in Provinces')

plt.xlabel('Province')

plt.ylabel('Percentage')

plt.xticks(rotation=45)

plt.legend(title='Gender')

plt.show()

A graph with blue and orange bars

Description automatically generated

**Conclusion:**

The visualizations provide insights into the income distribution and clustering across provinces.

1. **Gender Employment Rates**: The first chart shows the employment distribution of males and females across provinces. Punjab has the highest percentage of both male and female employment, while Sindh and KPK show similar trends. Balochistan lags behind in both male and female employment percentages. Male employment dominates in all provinces, with a smaller proportion of females employed.

2. **Average Male and Female Income**: The second chart highlights the income disparity between genders in different provinces. Punjab has the highest average income for both males and females, followed by Sindh and KPK. Balochistan exhibits the lowest income levels. Male incomes are higher than female incomes across all provinces, reflecting a gender income gap.

3. **Cluster Distribution by Province**: The third chart displays the clustering of income levels and demographics in provinces. Clusters are unevenly distributed, with most provinces dominated by cluster 0, indicating similar income and demographic characteristics in those areas. Other clusters show limited representation, suggesting regional economic diversity.

These insights suggest the need for targeted policies to bridge gender and regional disparities in employment and income.