**IBM NAAN MUDHALVAN**

**PHASE 5 - DOCUMENTATION**

**DATA LOADING AND PREPROCESSING FOR MARGINAL WORKERS DATASET**

|  |  |
| --- | --- |
| **DOMAIN :** | **DATA ANALYTICS**  **WITH COGNOS** |
| **PROJECT TITLE :** | **TN MARGINAL WORKERS ASSESSMENT**  **A SOCIOECONOMIC ANALYSIS** |
| **TEAM MEMBERS :** | **BHARATH KUMAR B-420421104012**  **SANJAY KUMAR A-420421104068**  **NOOR MOHAMED K-420421104048**  **NITHISH KANNAN G-420421104302**  **MOHAMMED JAVITH M-420421104042** |

**BASIC IDEA ON THE PROJECT**

**Background:**

Tamil Nadu, a southern state in India, has a significant population of marginalized workers who are often engaged in low-paying, informal sector jobs. Understanding the socio-economic conditions, employment patterns, and challenges faced by these marginal workers is crucial for informed policy-making and targeted interventions to improve their well-being.

**Problem Statement:**

The problem at hand is to conduct a comprehensive analysis of marginal workers in Tamil Nadu, with a focus on gathering and interpreting relevant data to gain insights into their socio-economic status, employment patterns, and associated challenges. This analysis aims to inform policymakers, government agencies, and non-governmental organizations in developing strategies and policies that can uplift the living conditions and opportunities for these marginalized individuals.

**Key Objectives:**

1. **Data Collection:** Collect and compile relevant data on marginal workers in Tamil Nadu, including demographic information, employment details, income levels, and geographical distribution.
2. **Employment Patterns:** Analyze the types of employment these workers are engaged in, including sectors such as agriculture, construction, domestic work, and informal labor markets.
3. **Income and Livelihoods:** Examine the income levels and sources of income for marginal workers, identifying disparities among various sub-groups based on gender, age, and location.
4. **Living Conditions:** Investigate the living conditions of marginal workers, including access to housing, sanitation, and basic amenities.
5. **Education and Skill Levels:** Assess the educational background and skill levels of these workers, exploring opportunities for skill development and education.
6. **Gender Disparities:** Analyze gender-specific challenges faced by female marginal workers, including wage gaps, workplace safety, and access to healthcare.
7. **Migration Trends:** Study migration patterns among marginal workers, both within Tamil Nadu and to other states, and understand the reasons for migration.
8. **Social Welfare Programs:** Evaluate the effectiveness of existing government programs and schemes targeted at marginal workers and identify areas for improvement.
9. **Recommendations:** Based on the analysis, provide actionable recommendations for policymakers and relevant stakeholders to improve the socio-economic conditions and opportunities for marginal workers in Tamil Nadu.

**Data Sources:**

1. Census data
2. Labor force surveys
3. Government reports and publications
4. NGO and research organization reports
5. Field surveys and interviews with marginal workers
6. Academic studies and scholarly articles

**Methodology:**

The analysis will involve a combination of quantitative and qualitative research methods, including data collection, surveys, interviews, and statistical analysis. Data visualization techniques will be employed to present findings effectively.

**Deliverables:**

1. Comprehensive report summarizing the analysis and findings.
2. Data sets and statistical analysis for reference.
3. Visualizations (charts, graphs, maps) to illustrate key trends.
4. Policy recommendations for improving the conditions of marginal workers in Tamil Nadu.

**DESIGN THINKING**

Applying design thinking principles to the analysis of marginal workers in Tamil Nadu can help ensure that the research is not only comprehensive but also focused on addressing the real needs and challenges of this vulnerable population. Here's a design thinking framework for the analysis:

**1. Empathize:**

* **User Research:** Start by deeply understanding the lives and experiences of marginal workers. Conduct in-depth interviews, surveys, and ethnographic studies to gather personal stories and insights.
* **Personas:** Create personas that represent different segments of the marginal worker population, considering factors like age, gender, location, and occupation.
* **Journey Mapping:** Map out the typical journey of a marginal worker, from seeking employment to living conditions, to understand pain points and opportunities for improvement.

**2. Define:**

* **Problem Statement:** Based on the empathy phase, distill the key challenges and needs of marginal workers into a clear and concise problem statement. For example, "How might we improve the access to skill development for female marginal workers in rural Tamil Nadu?"
* **Stakeholder Mapping:** Identify all stakeholders involved, including government agencies, NGOs, and community organizations, and understand their roles and interests.

**3. Ideate:**

* **Brainstorming:** Organize brainstorming sessions involving a multidisciplinary team, including researchers, designers, policymakers, and representatives from marginalized communities. Generate a wide range of innovative ideas to address the defined problem.
* **Co-Creation Workshops:** Involve marginal workers themselves in ideation sessions to ensure their voices are heard and to gather additional insights.

**4. Prototype:**

* **Concept Development:** Select the most promising ideas and develop rough prototypes or mockups of potential solutions. These could include new policies, programs, or interventions.
* **Pilot Programs:** Implement small-scale pilot programs to test the feasibility and effectiveness of selected solutions. This could involve collaborating with local organizations.

**5. Test:**

* **Feedback Gathering:** Collect feedback from the target audience and stakeholders involved in the pilot programs. Use surveys, interviews, and observations to assess the impact and gather insights for improvement.
* **Iteration:** Based on the feedback received, iterate on the prototypes and pilot programs as needed. This may involve refining the solutions or trying alternative approaches.

**6. Implement:**

* **Scaling Up:** Once a solution has been successfully tested and refined, work on scaling it up to reach a larger portion of the marginal worker population in Tamil Nadu.
* **Collaboration:** Collaborate with government agencies, NGOs, and community organizations to implement the solutions effectively and sustainably.

**7. Evaluate:**

* **Impact Assessment:** Continuously monitor and evaluate the impact of the implemented solutions on the well-being and conditions of marginal workers.
* **Feedback Loop:** Maintain an ongoing feedback loop with the target audience to ensure that the solutions remain relevant and effective over time.

**8. Communicate:**

* **Sharing Insights:** Share the findings and lessons learned from the design thinking process with stakeholders, policymakers, and the wider public through reports, presentations, and media to promote awareness and support.

Applying design thinking to the analysis of marginal workers in Tamil Nadu ensures that the research and solutions are not only data-driven but also deeply rooted in the needs and perspectives of the people being studied. It encourages a holistic approach that considers both the quantitative data and the human experiences of marginal workers, ultimately leading to more effective and empathetic policies and programs.

**CLUSTERING ANALYSIS**

The model chosen for the analysis is Clustering

Clustering analysis is a method used in unsupervised machine learning to group similar data points together. Here are the general steps for performing clustering analysis:

* **Data Collection**: First, you need to obtain the data you want to cluster. It is collected from <https://tn.data.gov.in/resource/marginal-workers-classified-age-industrial-category-and-sex-scheduled-caste-2011-tamil>.
* **Data Preprocessing**: This step involves cleaning and preparing the data for analysis. You may need to handle missing values, normalize the data, and convert categorical variables into numerical formats if necessary.
* **Feature Selection**: Choose the relevant features (attributes) that you want to use for clustering. Features should be selected based on the problem you're trying to solve
* **Normalization/Standardization**: Depending on the algorithm chosen, you may need to normalize or standardize the data to ensure that features with different scales do not bias the clustering
* **Choosing a Clustering Algorithm**: There are various clustering algorithms available, such as K-Means, Hierarchical Clustering, DBSCAN, etc. You need to select an appropriate algorithm for your dataset and problem
* **Cluster Analysis**: Apply the chosen clustering algorithm to the preprocessed data. This will group data points into clusters based on their similarity.
* **Summarizing the Model:** Summarize the model by specifying the Silhouette Score, Cluster Assignments, Cluster centers, No of clusters used, Cluster inertia etc…
  + **Cluster Inertia:**

Cluster inertia, also known as within-cluster sum of squares (WCSS), is a metric used to evaluate the quality of clusters in K-Means clustering or similar centroid-based clustering algorithms. It is a measure of how internally coherent the clusters are. The objective of K-Means clustering is to minimize this value

* + **Calculation:**

Cluster Inertia (WCSS) is calculated as the sum of the squared distances between each data point in a cluster and its centroid. The formula for calculating the WCSS for a given cluster "C" is:

WCSS(C) = Σ(dist(p, centroid(C))^2 for all p in C

where ,

* "dist(p, centroid(C))" is the Euclidean distance (or any other distance metric) between a data point "p" and the centroid of cluster "C."
* Σ denotes the summation over all data points within the cluster "C."
  + **Interpretation:**

A lower WCSS indicates that the data points within each cluster are closer to their respective cluster centroids, meaning that the clusters are more internally cohesive. Lower WCSS values are generally desirable as they suggest better-defined and more compact clusters

**INSIGHTS GAINED:**

**From Silhouette Score:**

The Silhouette Score ranges from -1 to 1, where a high value indicates that the object is well matched to its own cluster and poorly matched to neighboring clusters. A score around 0 means that the object is on or very close to the decision boundary between two neighboring clusters and could potentially be assigned to either.

In the silhouette score calculated (0.9719930172313797), it is very close to 1, which is a high score. This indicates that the data points in your clusters are well matched to their own clusters and significantly separated from neighboring clusters, suggesting that the clustering result is excellent.

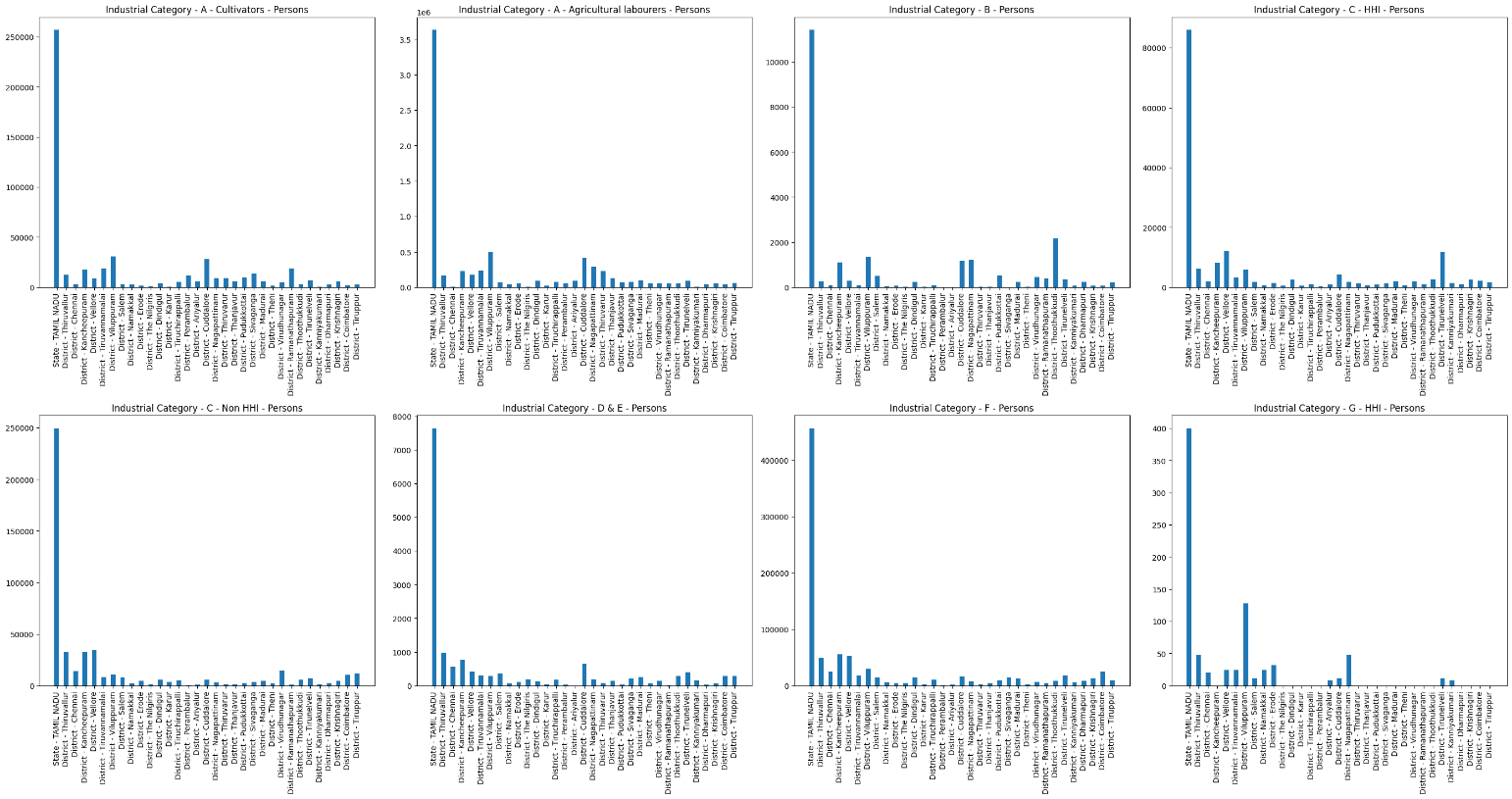
In summary, a silhouette score of approximately 0.9719930172313797 is a very good score, indicating that the clusters are well-defined and the data points within each cluster are similar to each other while being distinct from data points in other clusters.

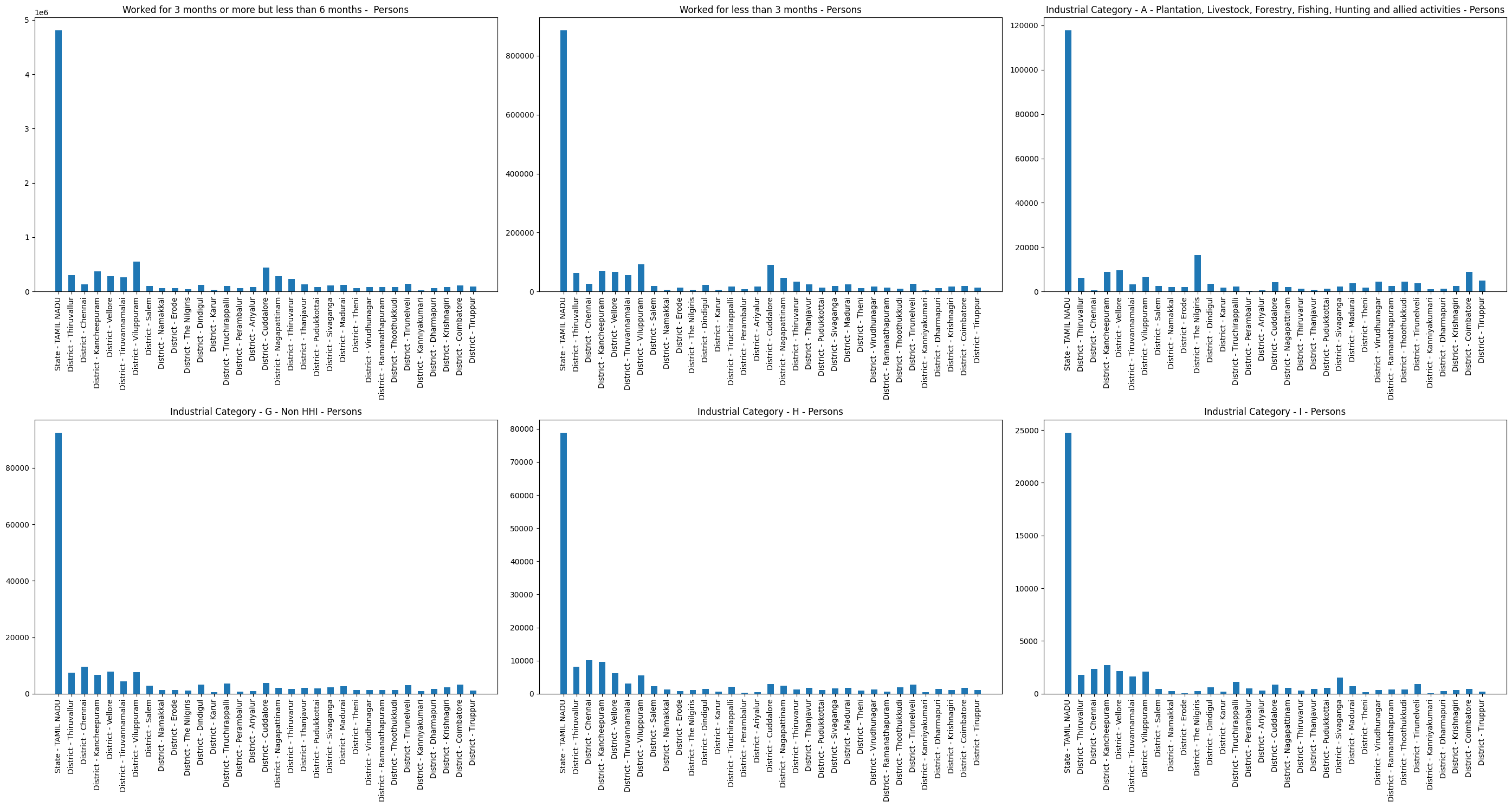
**From Cluster Inertia:**

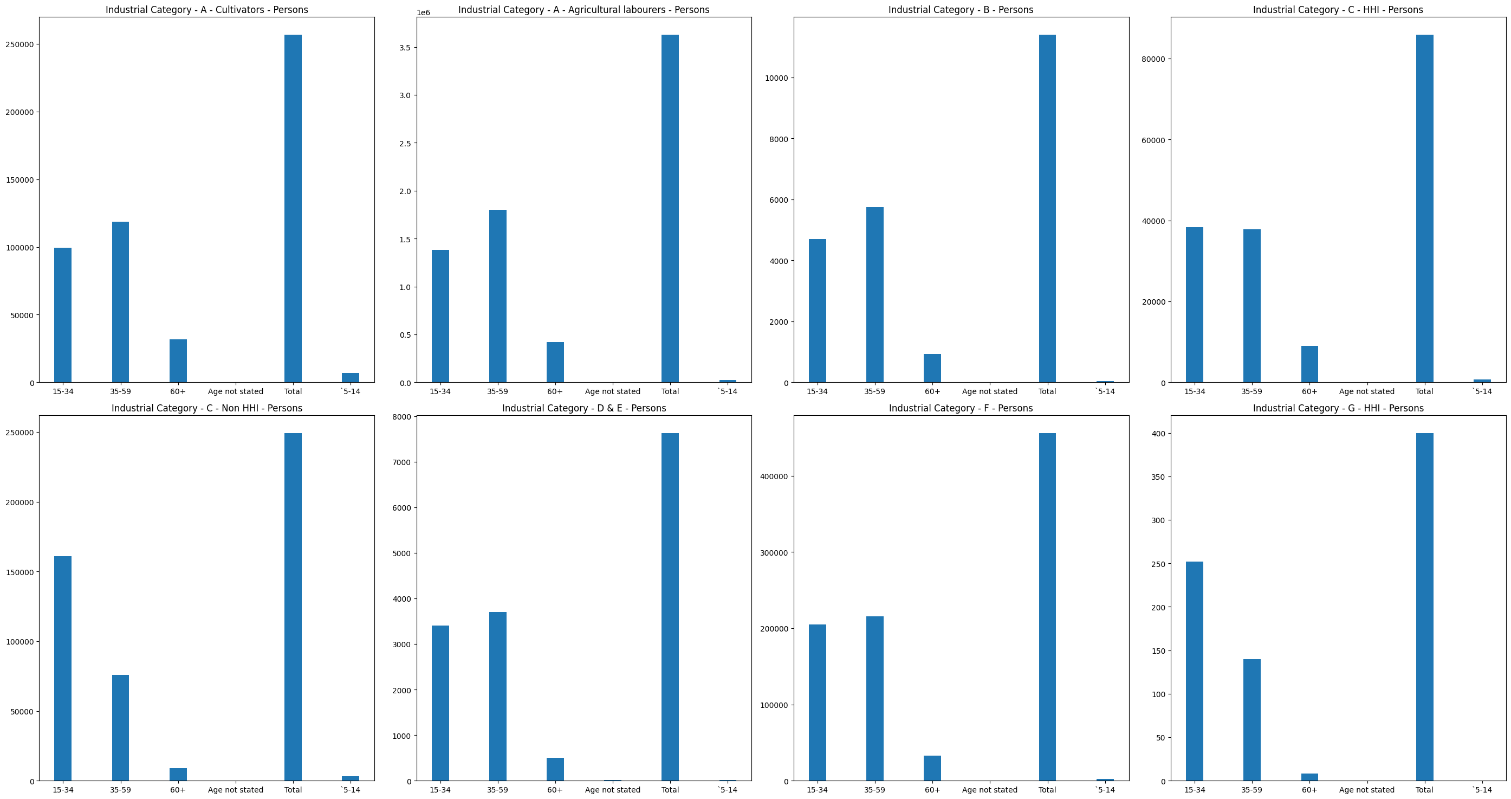
A WCSS value in the billions is relatively large. This suggests that the data points within the clusters are not very tightly packed around their respective centroids. In other words, the clusters may not be very internally cohesive, and the data points within each cluster are relatively spread out.

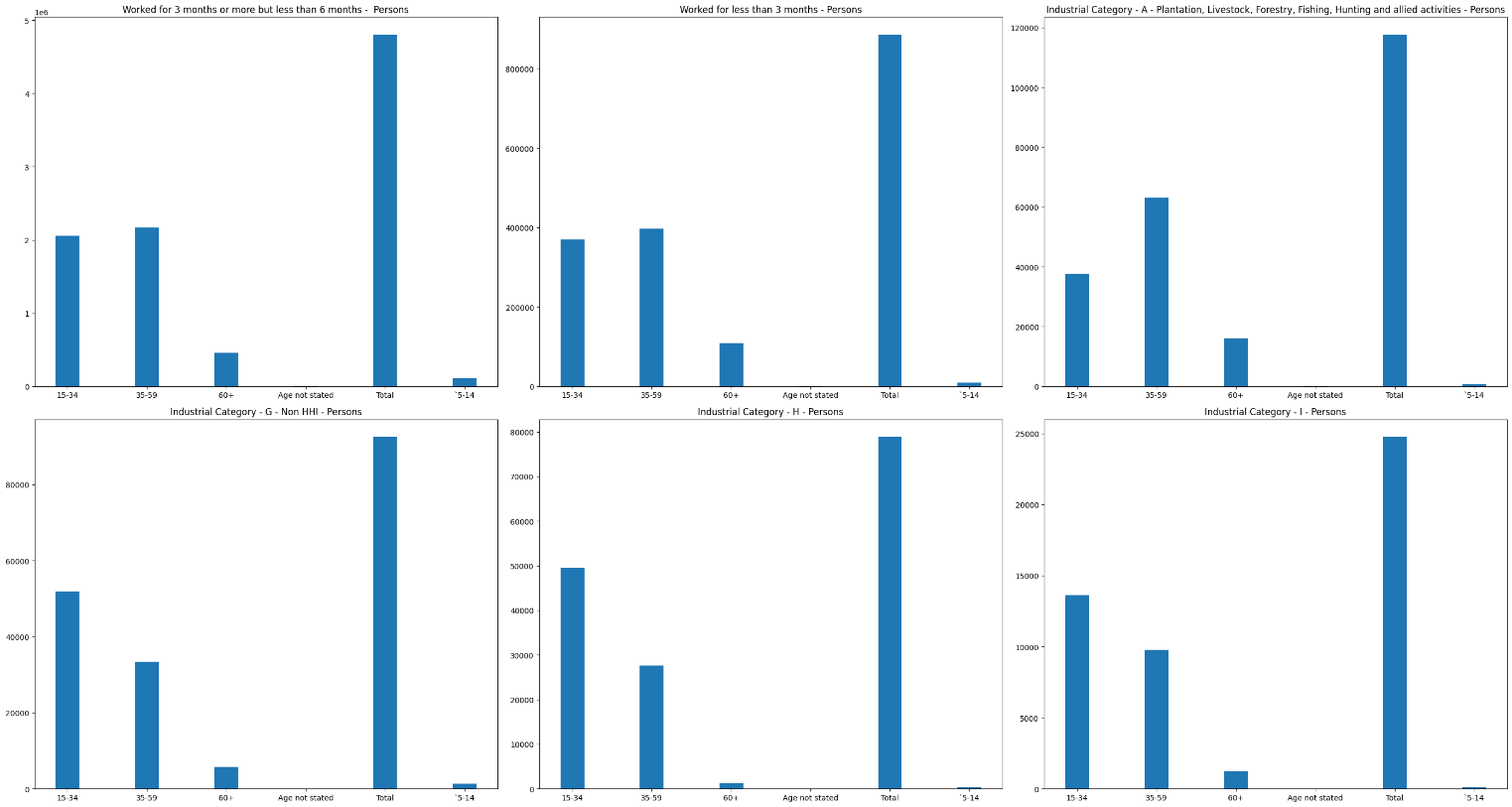
Without knowing the context and the specific number of clusters used to obtain this WCSS value, it's challenging to provide a definitive interpretation. It is required to compare this value to others, ideally on a graph, to determine whether it's relatively high or low in the context of your specific clustering problem.

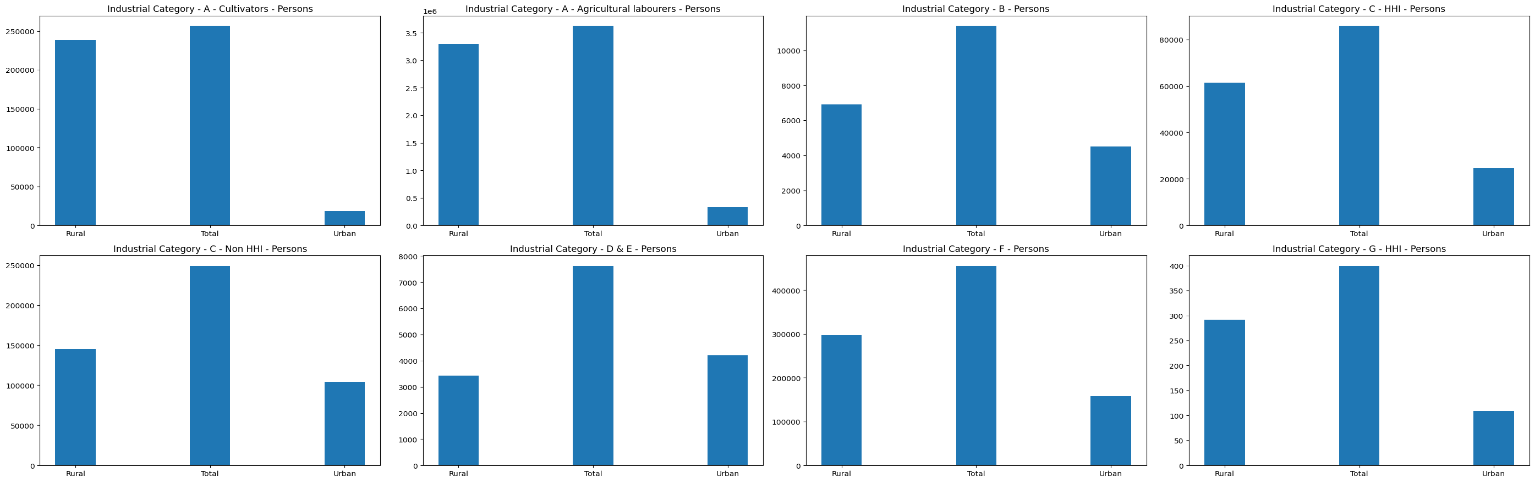
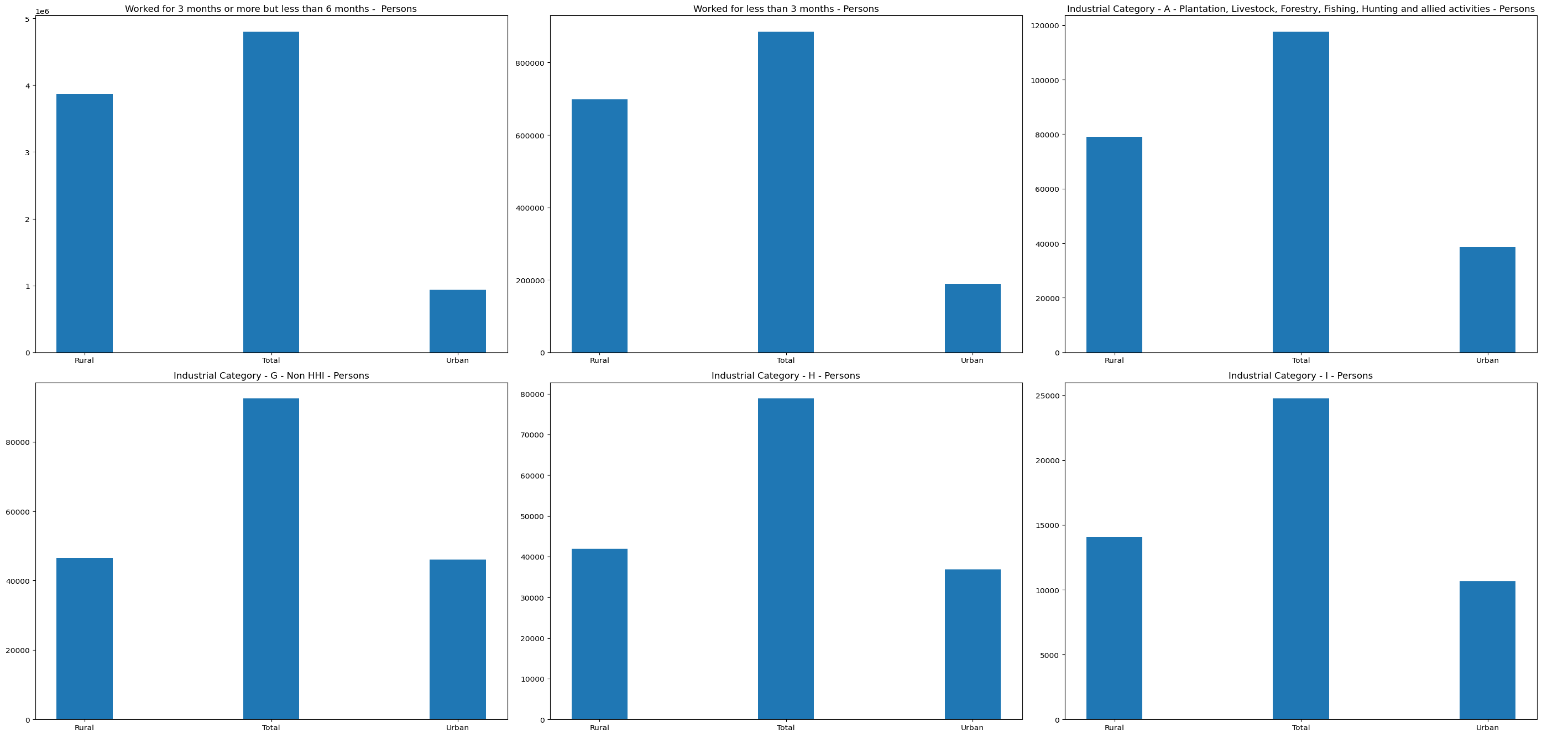
**SAMPLE VISUALIZATIONS AND OUTPUTS**

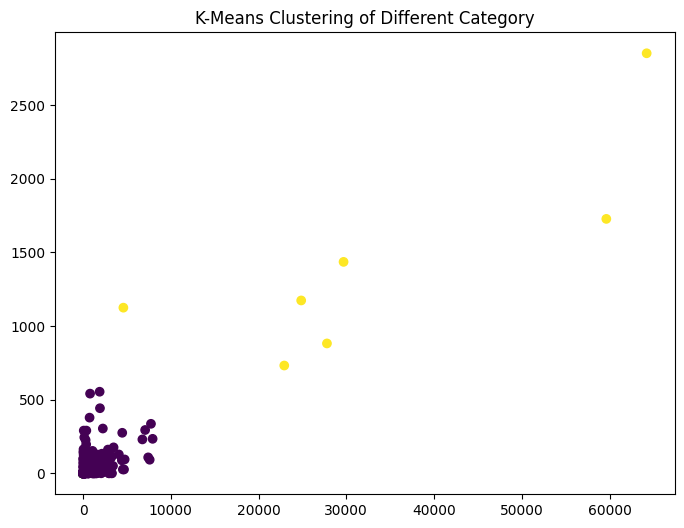


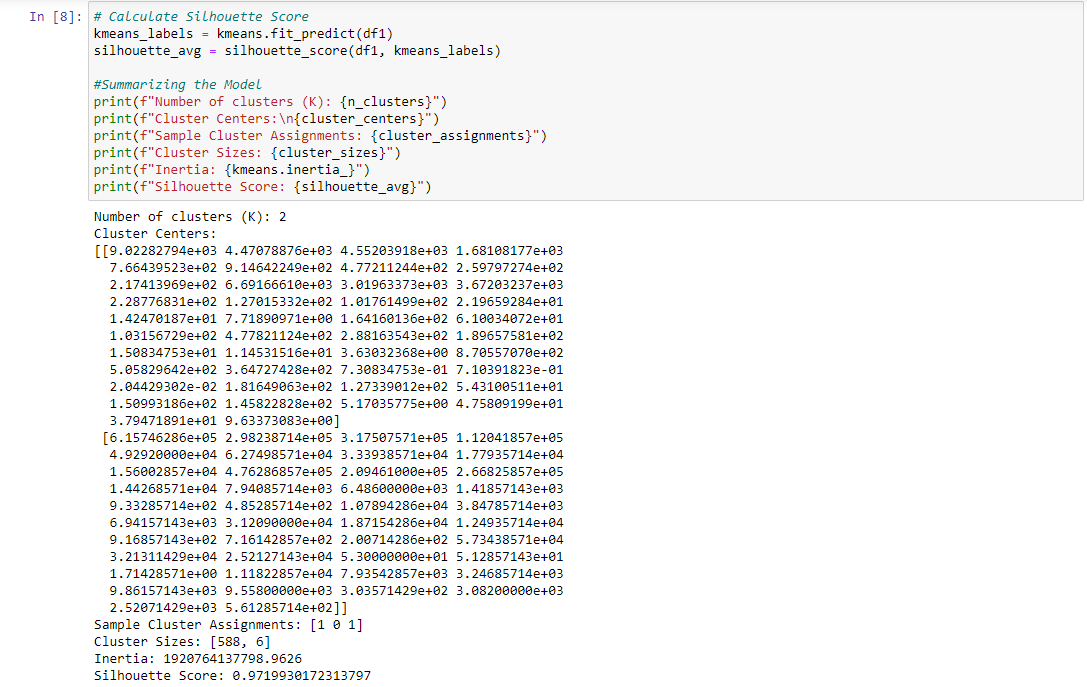










****