**Correlation Matrix: (write after the data cleaning and outlier detection, as 2.3):**

After processing the data, calculate the correlation between the indicators to generate a correlation matrix. The heat map generated from the correlation matrix shows that red indicates high correlation and blue indicates low correlation. From the heat map, we can see that the correlation between the indicators is not strong, which means these indicators can be used for further analysis.

图表

描述已自动生成

**Clustering （Could Precede Conclusion Part as 7.）**

This article utilizes the HDBSCAN clustering algorithm, which is a density-based clustering method particularly suited for high-dimensional and large datasets. One of its main advantages is that it does not require pre-specifying the number of clusters to be generated; instead, it determines the number of clusters based on the density distribution of the data, making it highly effective for datasets with complex structures.

Focusing on data from 2023, the study involves 46,707 data points. It includes cross-validation primarily involving systematic adjustments and tests on two parameters: the minimum cluster size (min\_cluster\_size) and the minimum number of samples (min\_samples). Specific settings are as follows: min\_cluster\_size varied from 5 to 20, in increments of 5, i.e., [5, 10, 15, 20]; min\_samples ranged from 3 to 9, in increments of 3, i.e., [3, 6, 9]. A total of 12 different parameter combination experiments were conducted to find the optimal clustering setup. After each clustering, the silhouette score was calculated to assess the quality of the clusters, indicating higher similarity within clusters and lower similarity between different clusters, thus ensuring good distinction in the clustering results. The analysis found the optimal settings to be: min\_cluster\_size of 15, min\_samples of 9, silhouette score of 0.78639, and 670 clusters.

Subsequently, the top six clusters were selected for further analysis based on the number of points in each cluster, with pie charts illustrating the proportion of occupations. The occupations include: OCC\_2310: Elementary and middle school teachers, OCC\_3255: Registered nurses, OCC\_4700: First-Line supervisors of retail sales workers, OCC\_4720: Cashiers, OCC\_9130: Driver/sales workers and truck drivers, OCC\_4760: Retail salespersons.

From the graphs, we observe the following:

Cluster 1: First-Line supervisors of retail sales workers and Retail salespersons together account for about 50%.

Cluster 2: Driver/sales workers and truck drivers, and Elementary and middle school teachers together account for about 50%.

Cluster 3: Driver/sales workers and truck drivers and First-Line supervisors of retail sales workers together account for about 50%.

Cluster 4: Registered nurses and Elementary and middle school teachers together account for about 50%.

Cluster 5: Cashiers and Driver/sales workers and truck drivers together account for about 50%.

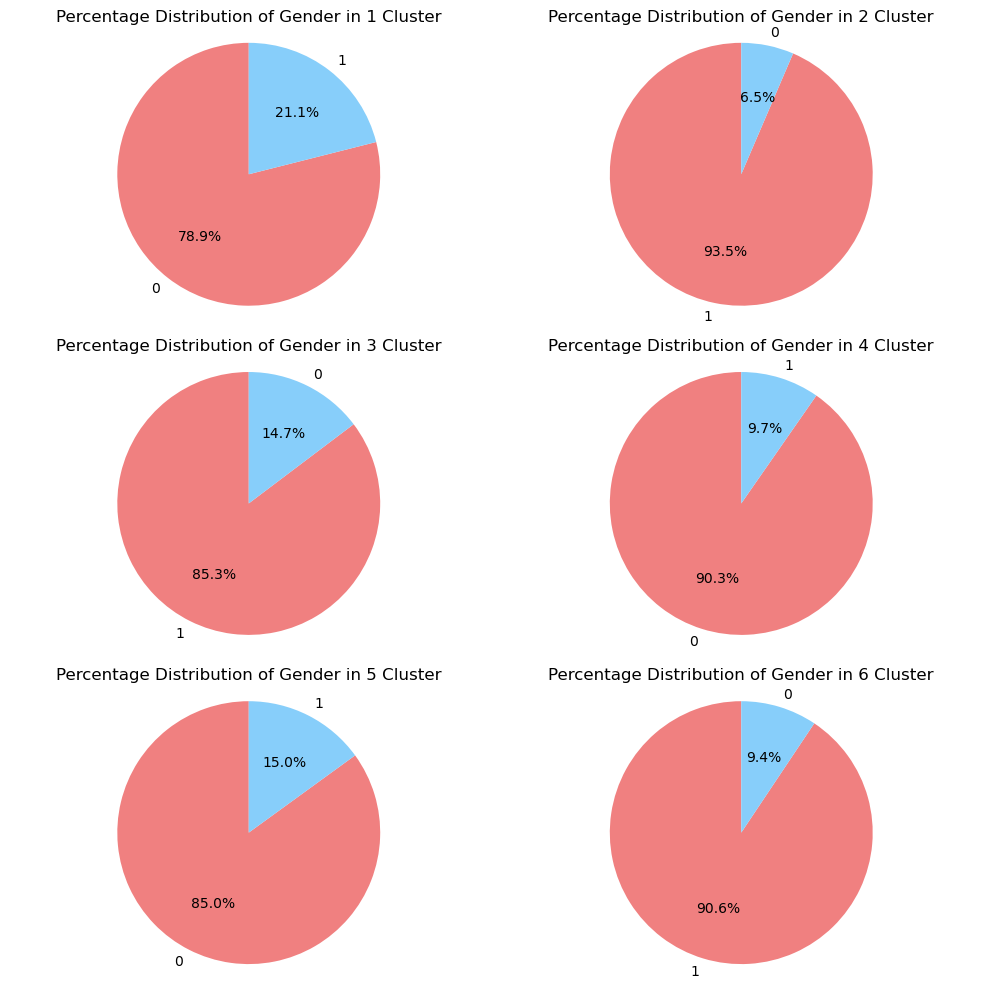
Cluster 6: Driver/sales workers and truck drivers and First-Line supervisors of retail sales workers together account for about 50%.图表, 饼图

描述已自动生成

Next, further analysis is conducted on the positions that have the highest proportions in these six clusters, including the distribution of male and female ratios. In 2023, the position of First-Line supervisors of retail sales workers has a higher proportion of females. In Driver/sales workers and truck drivers, males have a higher proportion. In Registered nurses and Cashiers, females have a higher proportion.

Subsequently, we analyzed the average income of females and males in these positions. In Cluster 1: First-Line supervisors of retail sales workers, the average income for females is $30,002.07, and for males, it is $30,000.75. In Cluster 2: Driver/sales workers and truck drivers, the average income for females is $50,000.00, and for males, it is $50,007.00. In Cluster 3: Driver/sales workers and truck drivers, the average income for females is $40,006.40, and for males, it is $40,003.62. In Cluster 4: Registered nurses, the average income for females is $20,000.56, and for males, it is $20,003.167. In Cluster 5: Cashiers, the average income for females is $20,000.56, and for males, it is $20,003.17. In Cluster 6: Driver/sales workers and truck drivers, the average income for females is $35,006.00, and for males, it is $35,001.1.

Combining the data results, in 2023, in the fields of First-Line supervisors of retail sales workers, Driver/sales workers and truck drivers, Registered nurses, and Cashiers, there is little difference in income between men and women.



**Add to the Conclusion Part as a Last Paragraph**

According to the analysis results, in 2023, in the fields of First-Line supervisors of retail sales workers, Driver/sales workers and truck drivers, Registered nurses, and Cashiers, there is little difference in income between men and women.