|  |
| --- |
| **Info 212: Data Science Programming** |
| **Smart Urban Mobility Solutions for Public Transportation** |
| ***Team Project Report*** |



**Group Number:** T17

**Group Members:**

Qinru Yang 320220940961 Rongshen He 20220940271

Qirui Zhang [320220941080](mailto:320220941080@lzu.edu.cn) Wei Dou [320220940190](mailto:320220940190@lzu.edu.cn)

**Abstract**

This project is based on American Census Demographic Data. Through exploratory data analysis and modeling analysis, the General Research Question is concluded as follows:

What are the key factors influencing employment rate in America?

During the exploratory data analysis phase, we refined the questions and asked and answered the four questions:

After analyzing the correlation coefficients, bar plots, line plots, box plots, and descriptive data, we find that (see the following sections for details):

American inland cities need more entrepreneurship supporting policies than coastal cities. The salary gap between different industries in the US is very large. We found that professional occupations have much higher salaries than other occupations, while construction occupations have poor salaries and prospects. There is a difference in the number of different industries, but this difference is not related to the unemployment rate and poverty rate, so we speculate that the employment structure is balanced, which is in line with supply and demand. There are significant differences in employment rates between different races, whites have higher employment rates than other races, and among minorities, we suspect there may be intrspecific competition.

After exploratory data analysis, we made assumptions on the data and built models for further analysis and prediction, and mainly solved 3 prediction problems:

**Catalog**

[1. Data Collection 2](#_Toc169627732)

[2. Data Preparation 3](#_Toc169627733)

[3. Exploratory data analysis 3](#_Toc169627734)

[3.1 Observe and understand the data 3](#_Toc169627735)

[3.2 Analyze questions 4](#_Toc169627736)

[3.2.1 Analyze questions 1 4](#_Toc169627737)

[3.2.2 Analyze question 2 4](#_Toc169627738)

[3.2.3 Analyze question 3 5](#_Toc169627739)

[3.2.4 Analyze question 4 7](#_Toc169627740)

[3.3 Other findings 8](#_Toc169627741)

[3.4 Summary of EDA 8](#_Toc169627742)

[4. Hypothesis 8](#_Toc169627743)

[5. Modeling, prediction, and Evaluation 8](#_Toc169627744)

[5.1 Modeling and analyze question 4 8](#_Toc169627745)

[5. 2 Modeling and analyze question 5 10](#_Toc169627746)

# 1. Data Collection

The dataset we are going to use is called `**American Census Demographic Data**`, which were taken from the DP03 and DP05 tables of the 2015 American Community 5-year estimates.

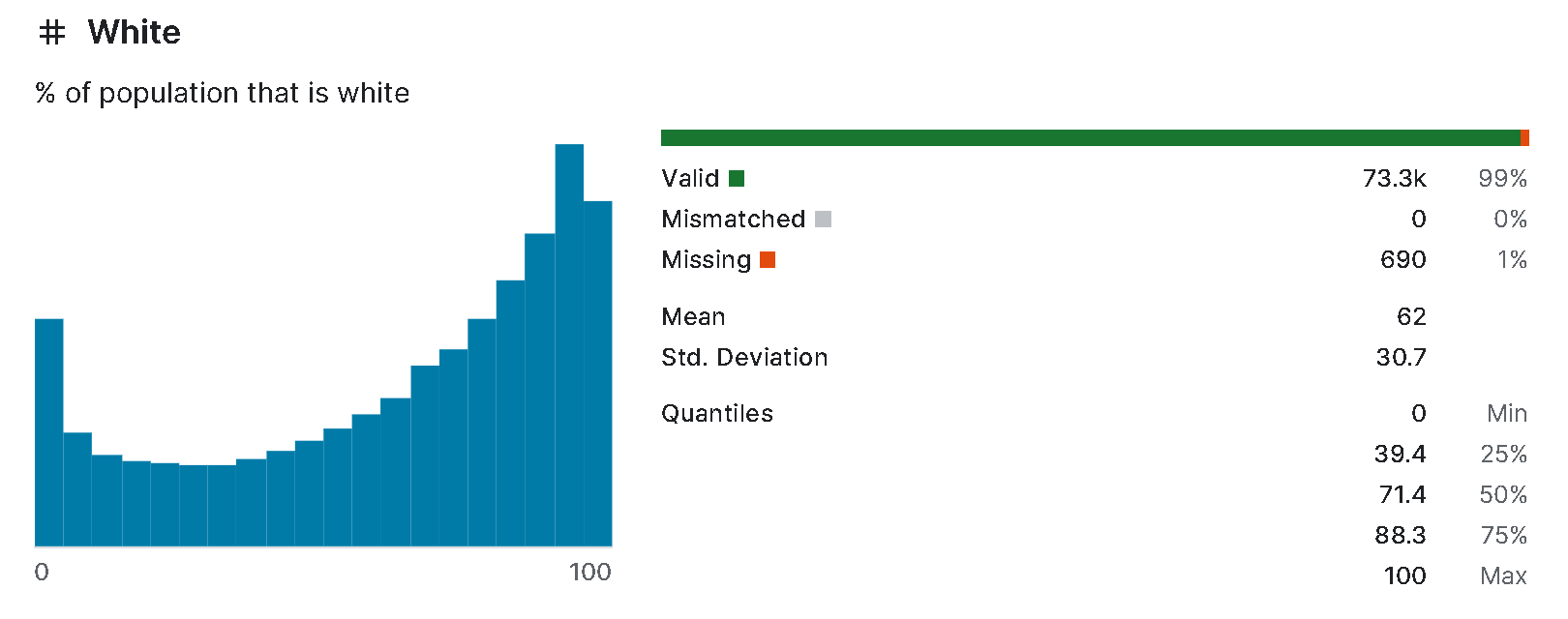
Comprehensive coverage: The dataset includes data for each census tract, county, and county equivalent in the United States. This comprehensive geographic coverage ensures that our analysis captures regional variations and trends across the entire country. Additionally, the dataset encompasses a wide range of demographic variables, providing a detailed snapshot of the population's characteristics. This breadth is essential for examining the multifaceted factors that influence employment rates.

Validity and reliability: The data comes from the American Community Survey (ACS), conducted by the U.S. Census Bureau. The ACS is a reputable and reliable source, known for its rigorous data collection and processing methodologies. Additionally, the use of five-year estimates enhances the reliability of the data by smoothing out annual fluctuations and providing a more stable and accurate representation of demographic trends.

Analytical depth: The specific tables we are using (DP03 and DP05) contain detailed information on economic characteristics (e.g., employment status, income levels) and demographic characteristics (e.g., age, race, gender), crucial for in-depth exploratory data analysis and building predictive models. The richness of the dataset enables multifactorial analysis, allowing us to consider various demographic and economic factors simultaneously. This is essential for identifying key drivers of employment rates and understanding their interrelationships.

Additionally, we are going to use the census demographic data of 2017 to validate the generalization ability of our model, enabling our work to be evaluated fairly.

# 2. Data Preparation

First, we look at the original dataset and see that some columns have missing values (the original dataset on kaggle has been visualized by its author). The good news is that the percentage of true values is extremely small, so we discard the missing values.

To process our data, we dropped the columns "Income", "IncomeErr", "IncomePerCapErr" firstly, decause we wanted to aggregated data by state, while the "Income" indicator showed the homehold income, "IncomeErr" showed homehold income err and the "IncomePerCapErr" showed per capita income err. We didn't know the specific number of homeholds in America, and we didn't know the specific income of everyone contained in this dataset. Hence, we cannot use the data in "Income", "IncomeErr", "IncomePerCapErr" columns.

The data in 'Hispanic', 'White', 'Black', 'Native', 'Asian', 'Pacific', 'IncomePerCap', 'Poverty', 'ChildPoverty', 'Professional', 'Service', 'Office', 'Construction', 'Production', 'Drive', 'Carpool', 'Transit', 'Walk', 'OtherTransp', 'WorkAtHome' columns are percentage number. To aggregate data by state, we need to multiply the data in these indicators by the total number of people, group them by state, and add them together to calculate the exact number of people with this indicator.

While data in "PrivateWork", 'PublicWork', 'SelfEmployed', 'FamilyWork', 'Unemployment' columns showed persentage of employed people in different industries, so we multiplied them with the total number of employed people.

For other indicators like 'IncomePerCap', 'Unployment', 'MeanEmploymentRate', 'Poverty', we

# 3. Exploratory data analysis

Our Exploratory Data Analysis will be guided by the questions posed in the proposal, adopting a **question-driven** approach. We will obtain statistical and measurement data through basic mathematical calculations. Additionally, we will visualize the data to make it more observable and to gain insights into the relationships between the data. The questions are as follows:

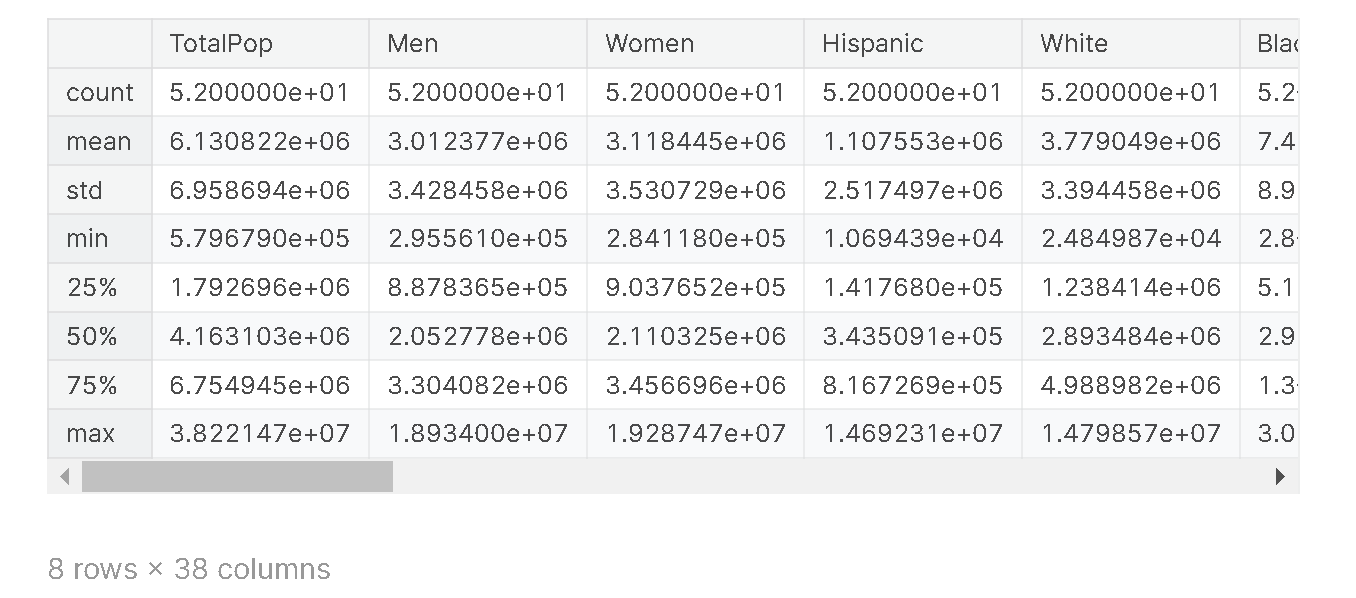
1) Which areas should be covered with additional entrepreneurship supporting policies?

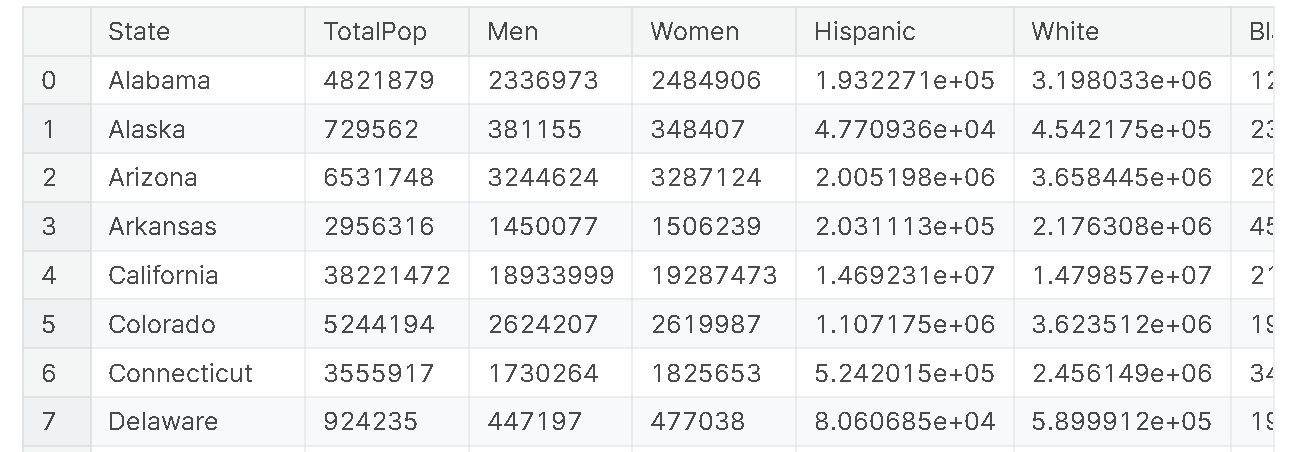
2) What are the salary levels in different industries? Whether there is a significant pay gap?

3) Whether the number of practitioners in certain industries is too large or too small, resulting in an unbalanced employment structure?

4) How labor force participation rates differ among different groups?

## 3.1 Observe and understand the data

First, we examined the columns of the dataset `grouped\_df\_by\_state` generated from preprocessing. Columns typically represent the meanings of various features, so observing the column names helps us understand what each column represents. We observed that it has 35 columns. Additionally, we performed descriptive statistics on the columns of `grouped\_df\_by\_state`. The mean and median from these descriptive statistics help us understand the overall characteristics of the data and the magnitude of the data in the corresponding columns.

Besides descriptive statistics, we also calculated the correlation coefficients between features. Through the correlation coefficient matrix, we can gain early insights into which features have a high correlation with each other. 

## 3.2 Analyze questions

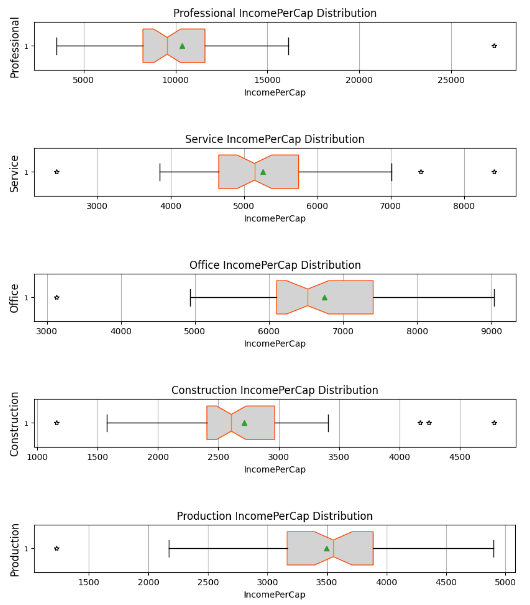
### 3.2.1 Analyze questions 1

We visualize the number of people performing different types of work in each state on Map. We find:

* The private employment rate varies significantly across the United States. This is mainly characterized by high private employment rates in coastal cities (indicated in yellow on the map), such as California, while inland cities have lower private employment rates.
* Other employment rates, such as self-employment, public sector employment, and household employment, are generally low across the country.
* We speculate that coastal cities often engage more in international trade, thus offering more job opportunities. Additionally, coastal cities are usually national technology innovation hubs, resulting in more private companies and higher private employment rates. In contrast, other forms of employment do not have high regional requirements, so they are more evenly distributed nationwide.

### 3.2.2 Analyze question 2

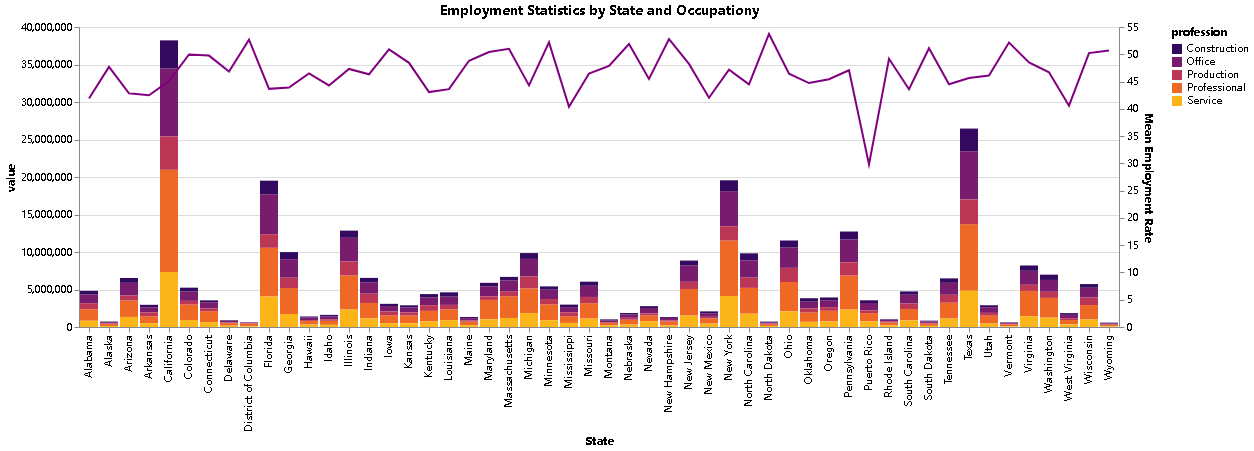
We have plotted box plots of IncomePerCap for each industry. These five sets of box plots reflect the distribution differences in salaries across different professions. Additionally, we have created a table of descriptive statistics for IncomePerCap across industries. From the graphs and tables, we observe:

* The salary level for the Professional industry is significantly higher than that of other industries (its first quartile value is higher than the third quartile values of other industries). Moreover, the variance in this industry is the highest, indicating a considerable salary fluctuation within the industry.
* The Construction industry has the lowest median and mean salaries, indicating poor salary levels in this sector. Coupled with the previous line chart, the employment population in this industry is almost the lowest in all states. Therefore, we can conclude that the population with low wages is relatively small. Additionally, the variance in this industry is the smallest, suggesting that it is challenging to achieve higher salaries if employed in this industry, making its employment prospects not optimistic.
* Comparing the Office and Service industries, we find that the standard deviations of both industries are similar, indicating comparable stability. However, the Office industry has higher mean and quartile values than the Service industry. This suggests that the Office industry can maintain stability while also offering higher salaries compared to the Service industry. Therefore, individuals in the Service industry might consider transitioning to the Office industry.

### 3.2.3 Analyze question 3

We visualize the employed population and each industry employed rate accross different state.

* We found that the employment rate in Professional is the highest among all states, while the employment rate in Construction is the lowest. It is also evident from the chart that the employment rate in Professional is significantly higher than in Construction.
* The employment rates of Production, Service and Office increase sequentially. The chart shows that the bars for Office are generally much higher than those for Production. However, the difference between the bars for Office and those for Service is not particularly significant.
* The above analysis suggests that there may be substantial differences in employment rates between different industries.



But do these differences indicate an imbalance in the employment structure?

* We believe that employment imbalance is an abnormal phenomenon. It can lead to quantifiable impacts such as the overall employment rate, poverty rate, and so on.
* Therefore, we attempted to analyze the relationship between the current employment structure and the overall employment rate and poverty rate to determine whether the current employment structure is reasonable.

Thus, we created a heatmap of the correlation coefficients between various industries and the overall employment rate and poverty rate. Through this chart, we found:

* The correlation coefficients between each industry and the overall employment rate and poverty rate are nearly zero, indicating that the current industry structure does not significantly impact the national employment situation.
* Additionally, we found that the correlation between various industries is relatively high and positively correlated. A reasonable explanation is that as the overall employment rate increases, the population flows into various industries in a supply-demand balanced manner.

From this, we conclude that although there may be substantial differences in employment rates between different industries, this significant difference is in line with the balance of supply and demand, meaning each industry has employment numbers that match its needs.

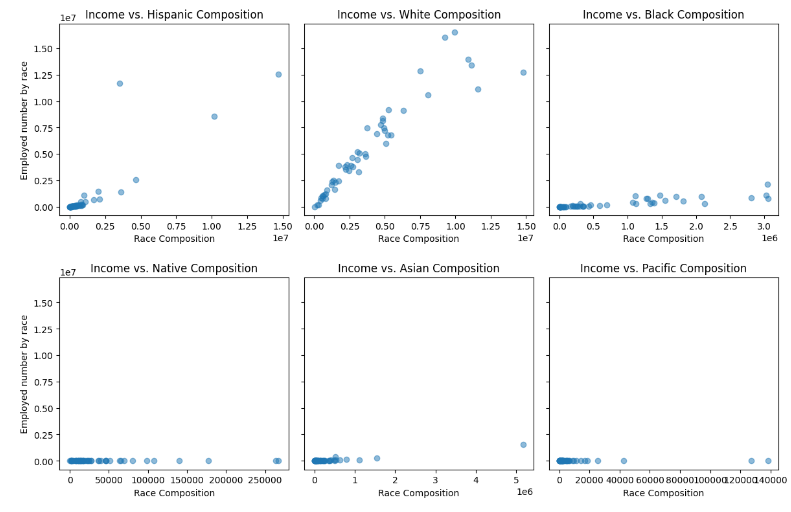
### 3.2.4 Analyze question 4

We separate people by race and plot scatter plots of race versus employed population corresponding to that race, where:

Employed\_population\_corresponding\_to\_a\_race=Total\_population\_of \_that\_race/Total\_employed\_population ∗ Total\_population\_of\_that\_race

We find:

* As the population increases, only White and Black populations show an upward trend in employed numbers. The trend is more pronounced for White population, suggesting better employment rates compared to the Black population. Additionally, the Hispanic population shows a very slight upward trend in employed numbers.
* For the second row representing races, we observe nearly horizontal trends, indicating minimal variation in employed numbers with changes in population size. This suggests poor employment opportunities for these three races, potentially accompanied by significant competition within each race.



## 3.3 Other findings

## 3.4 Summary of EDA

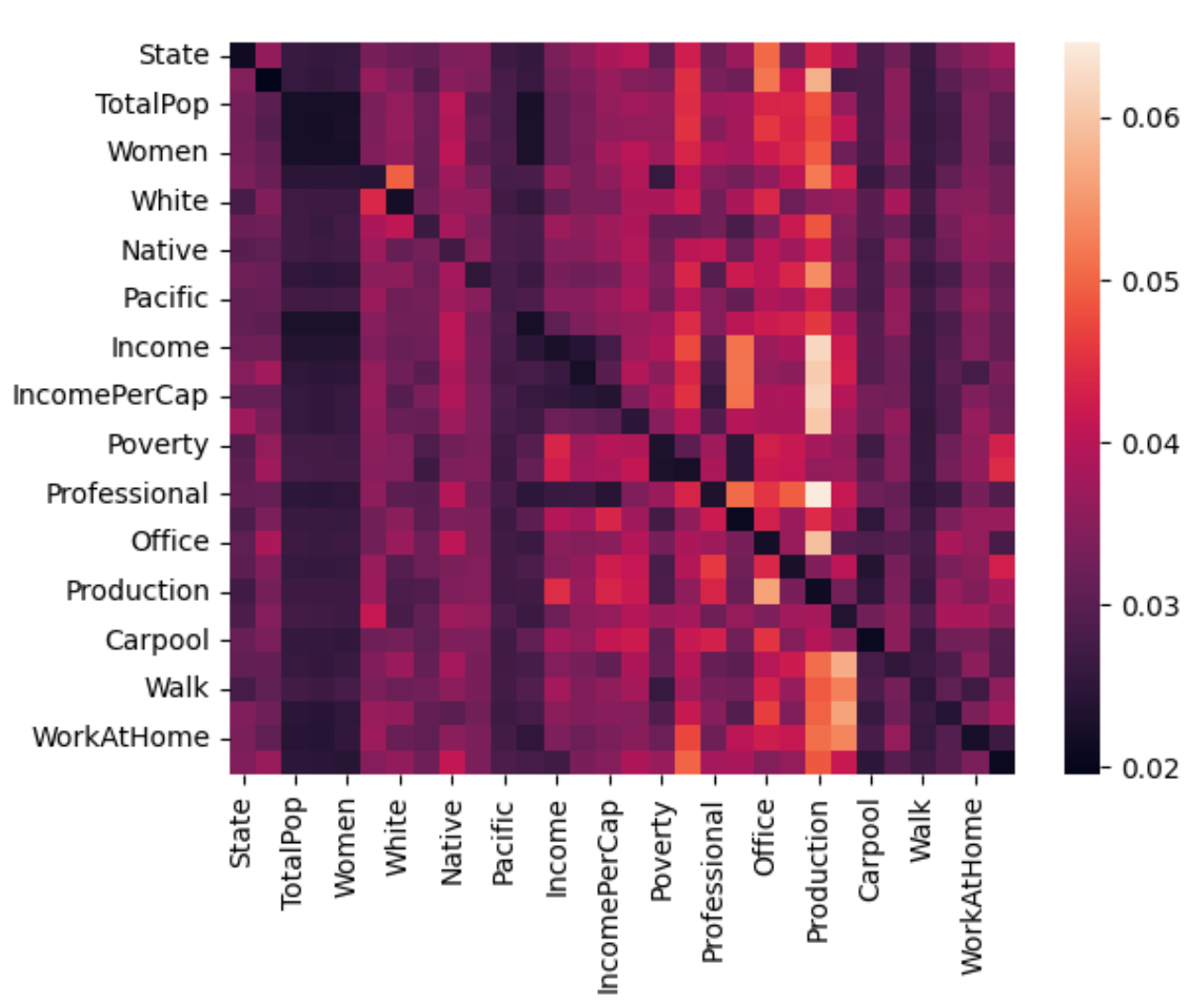
# 4. Modeling, prediction, and Evaluation

## 4.1 Modeling and analyze question 5

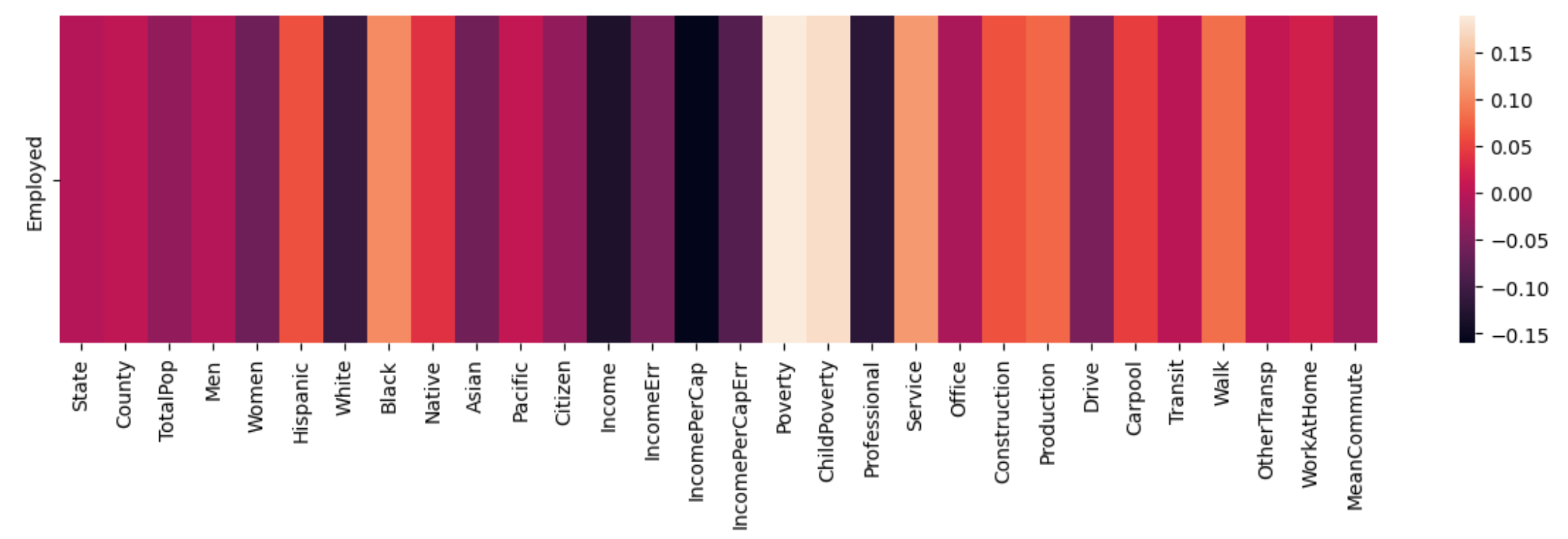
We established a simple deep learning model with several MLP layers that includes an Attention structure. We expect the Attention scores to reflect the model's perceived correlations between input features, akin to a correlation coefficient matrix within the model. When the model's accuracy is sufficiently high, these attention scores can replace the correlation coefficient matrix to more accurately reflect the importance between input features.

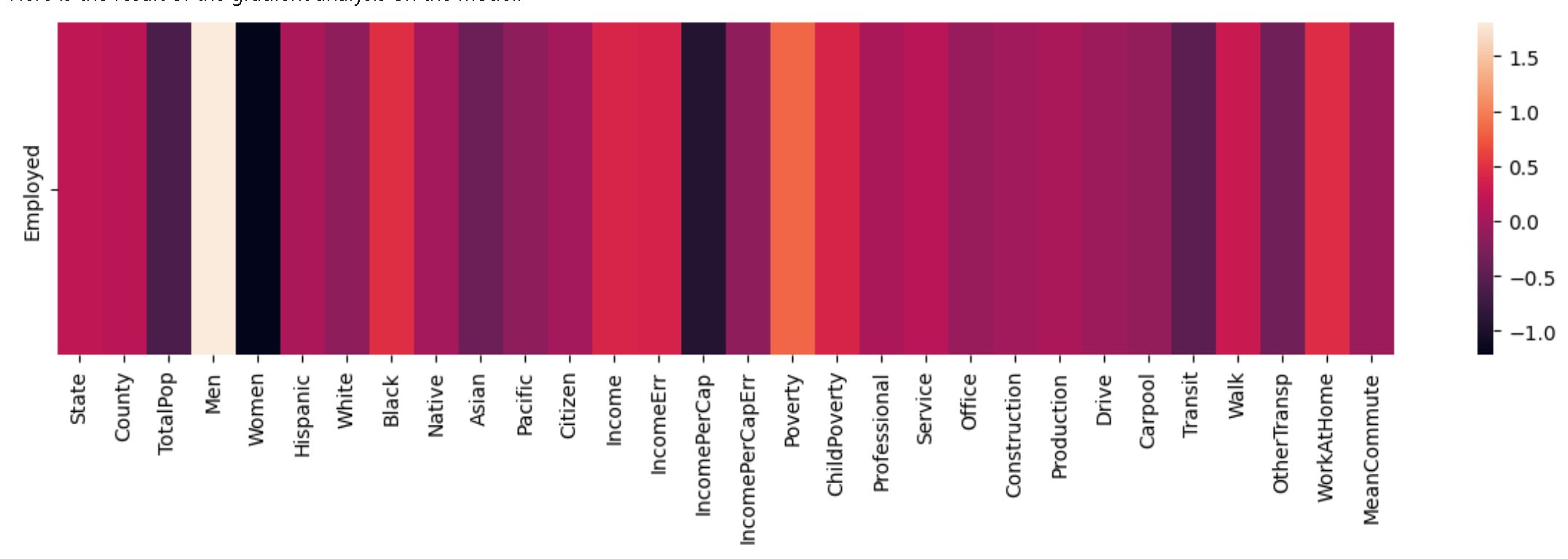
Before feeding data into the model for training, we standardized the data. This step eliminates the influence of different data scales on the model and enhances the model's stability.

Finally, we achieved an error (mse) of 0.1610 on the test set. Here is the heatmap of the model's Attention scores:

For instance, regarding income, the model identifies a strong positive correlation with production and a negative correlation with TotalPop and gender.

We attempted to use model gradients to show which input features are deemed important by the model. By backpropagating the output gradients through the model, we can compute the gradients of the input features. The larger the absolute gradient of an input feature, the more important the model considers that input to be.

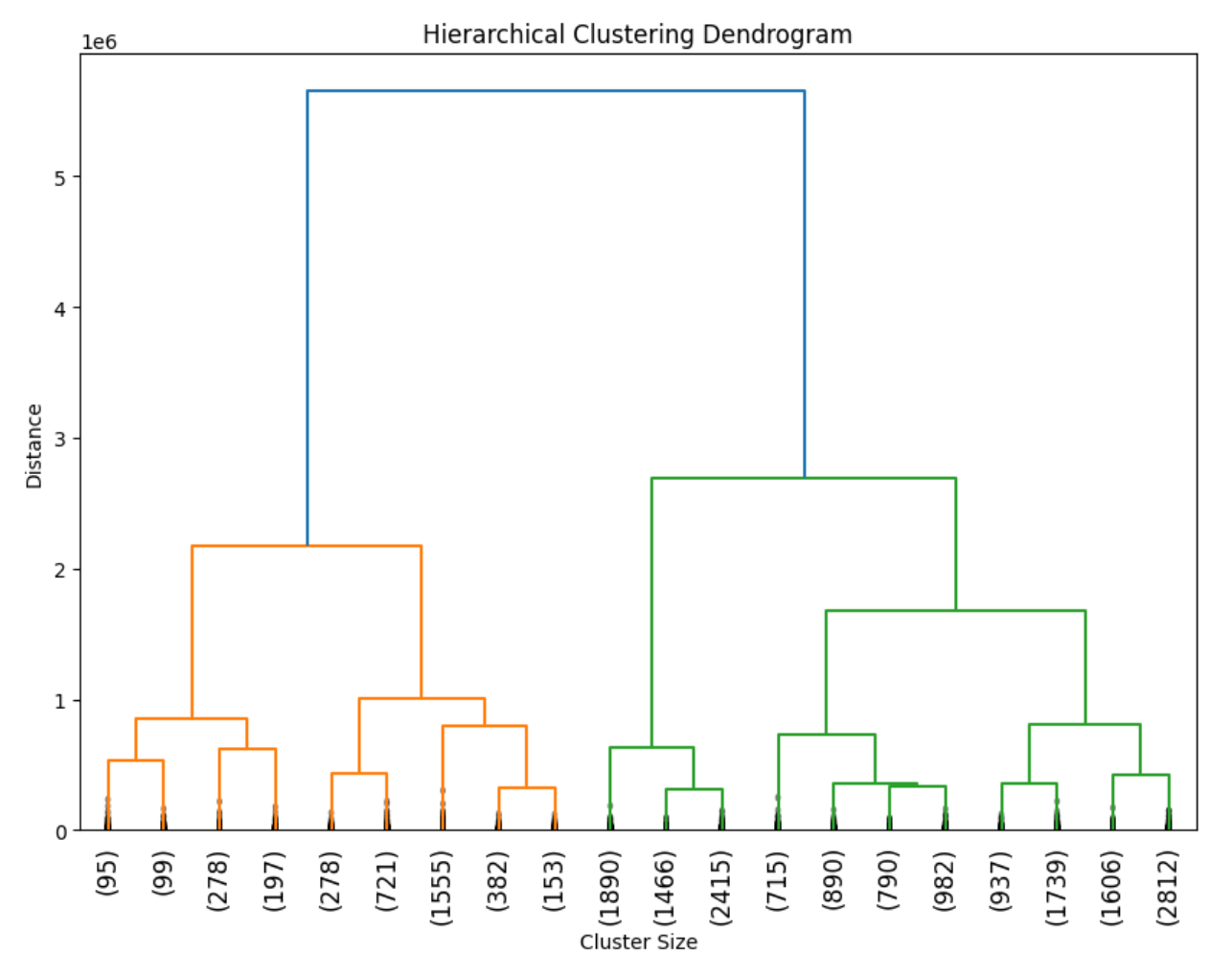
Here is the correlation matrix of input features to the output feature:

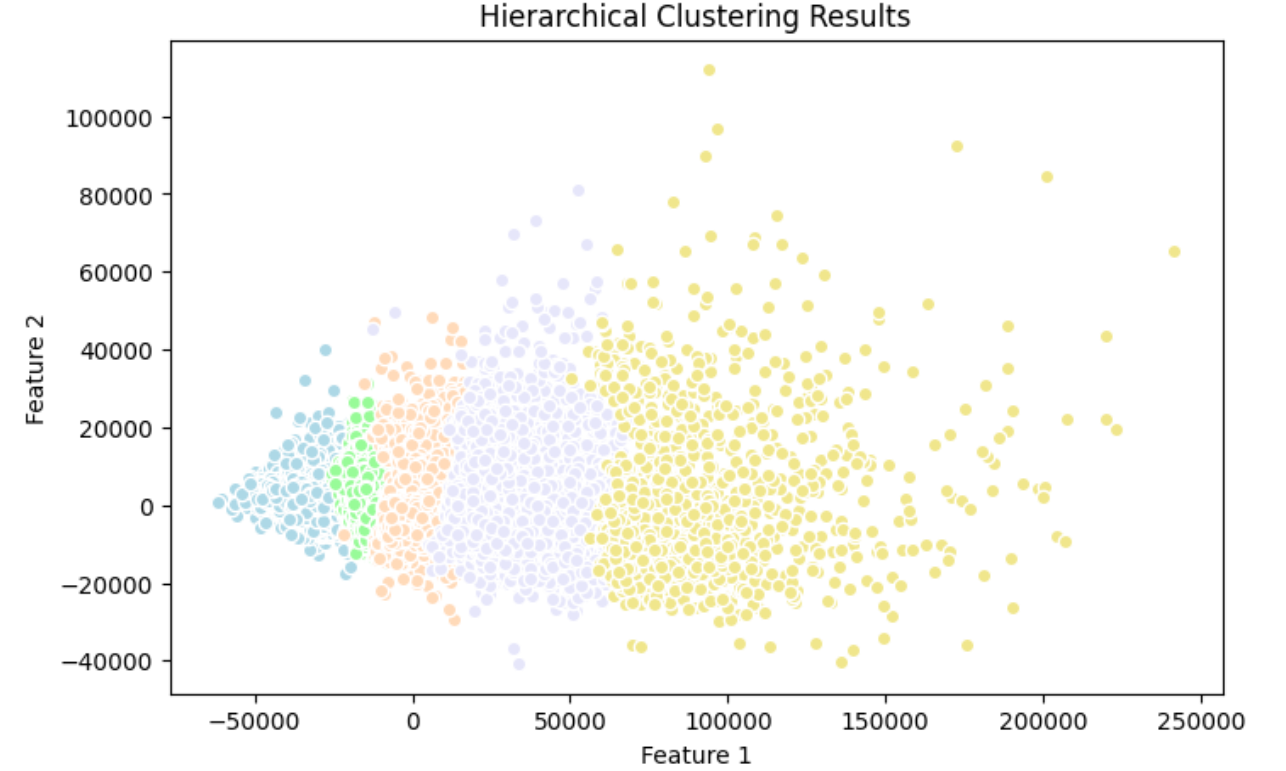
Here is the result of the gradient analysis on the model:

We find that features highly correlated according to the correlation matrix are indeed related to the output. The gradient analysis results more precisely reflect the importance of input features to the output. Notably, the model indicates that the input features with strong correlations to the employment rate are:

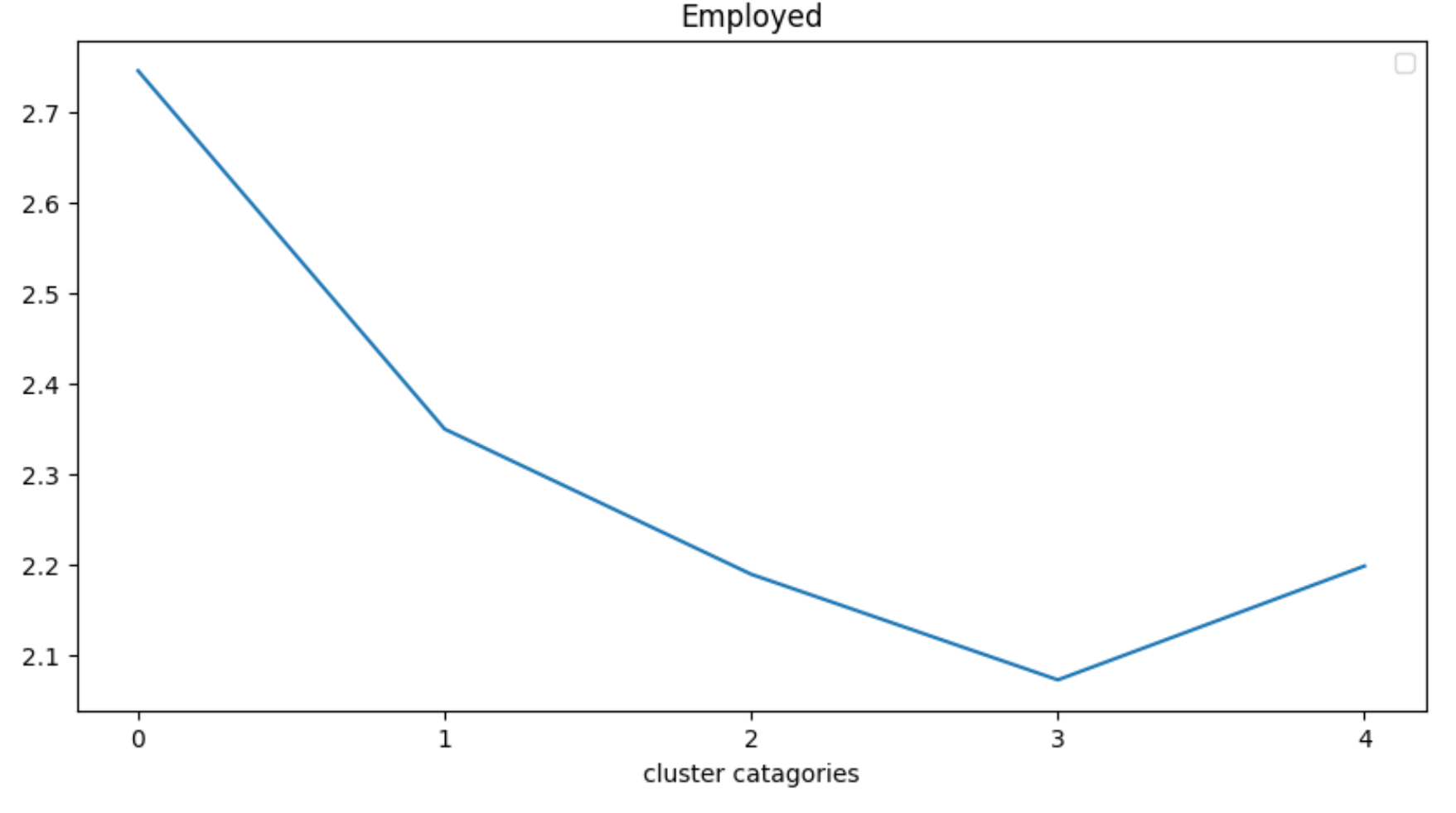
* Positively correlated: Men, Poverty.
* Negatively correlated: Women, IncomePerCap.

## 4. 2 Modeling and analyze question 6

We used a hierarchical clustering model to cluster the input features. After training, the model's dendrogram is as follows:

We visualized the clustering results with a scatter plot combined with PCA, where each color represents a different cluster:

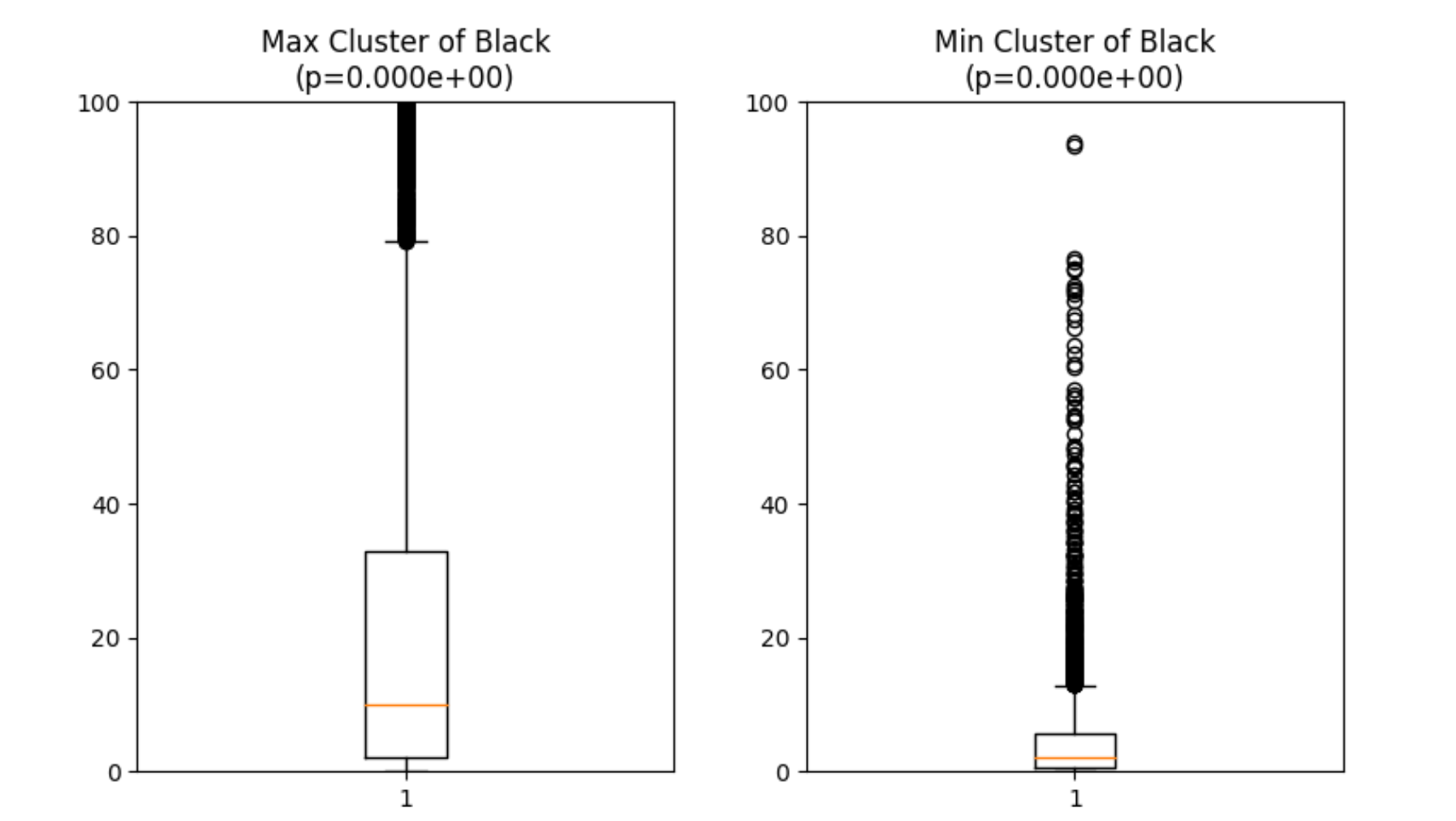
The results show a clear clustering phenomenon, indicating that the model can easily classify the data into multiple categories. This implies that data within the same cluster tends to have similar characteristics.

Additionally, we averaged the employment rates of different clusters and observed a clear stratification of employment rates across different clusters:

This essentially proves our hypothesis that data in the same cluster is more likely to be characterized by specific features.

Then, we aim to compare clusters with the highest and lowest employment rates and identify common features within the same cluster and across different clusters.

* We used a t-test to identify the top n features that contribute the most to these clusters.
* We then compared the data distribution of these features using a box plot.

For reasons of space, only some of the box plots are shown, and readers are referred to the submitted notebook on their own. 

We found significant differences in data distribution for these top n features between clusters with **highest and lowest employment rates**. For instance:

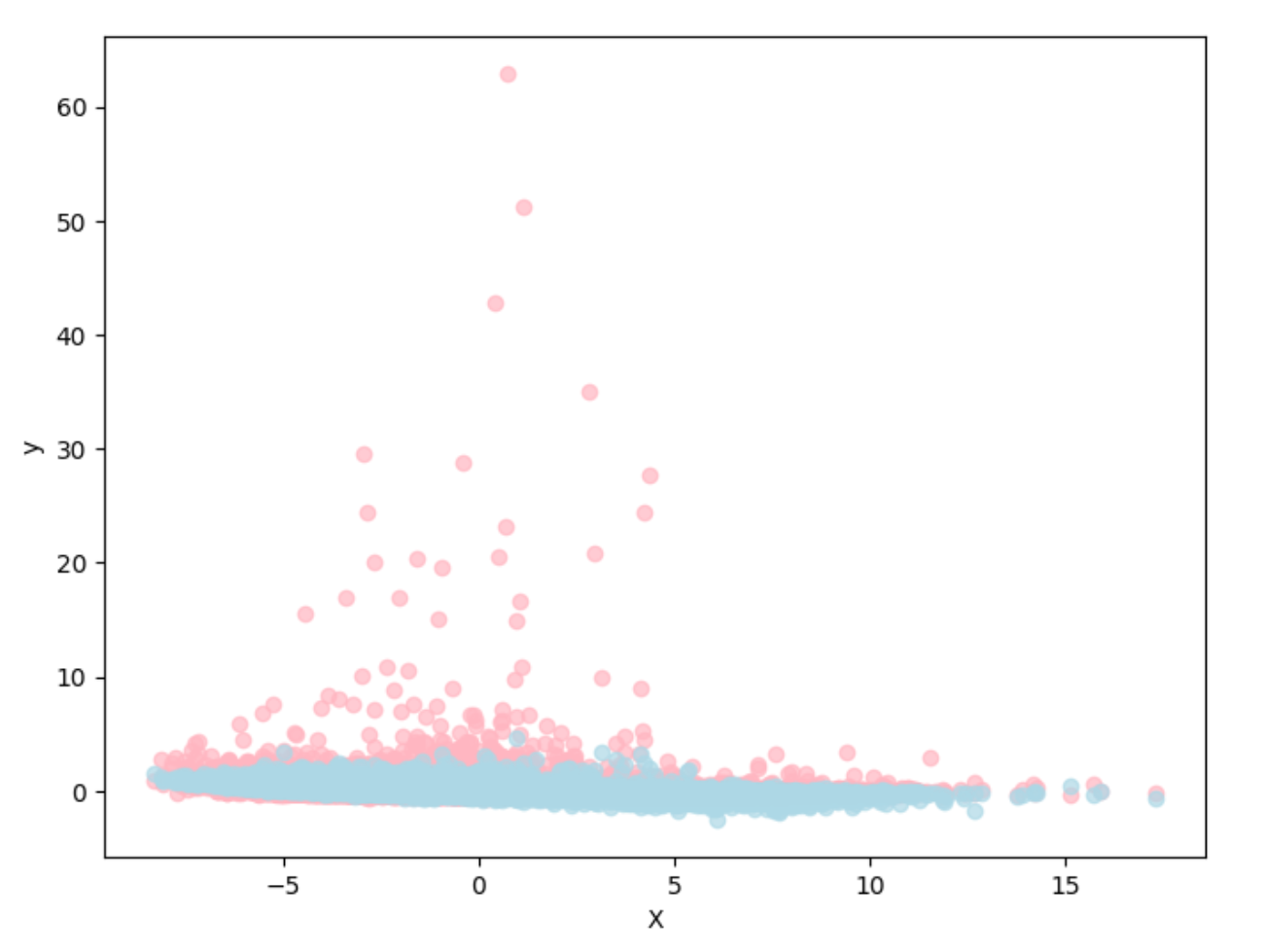
* Clusters with higher employment rates tend to have lower income, whereas clusters with lower employment rates tend to have higher income.
* Clusters with higher employment rates have a higher childhood poverty rate, whereas clusters with lower employment rates have a lower childhood poverty rate.
* Clusters with higher employment rates have a higher percentage of black individuals and a lower percentage of white individuals, while the opposite is true for clusters with lower employment rates.

This suggests that clusters with lower employment rates are more representative of the elite class, while clusters with higher employment rates are more representative of the lower class, who are more likely to engage in physical and repetitive labor and have higher poverty rates.

## 4. 3 Modeling and analyze question 7

We attempted to generalize our model to the 2017 dataset. If the model trained on the 2015 dataset generalizes well to the 2017 dataset, it suggests that our conclusions may be generalized to datasets over a longer time span.

We generalized our deep learning model trained on the 2015 dataset to the 2017 dataset, achieving an mse of 0.3248.

Here is the visualization of the generalization results:

The generalization results are barely acceptable. The model shows some degree of overfitting. Our conclusions can be partially transferred to the 2017 dataset.

We performed hierarchical clustering on the 2017 dataset and extracted its employment rate. We observed a stratification phenomenon in employment rates across different clusters, suggesting that our clustering conclusions can be somewhat applied to future datasets.

In summary, our model performed worse on the 2017 dataset than on the 2015 test set, but it is still within an acceptable range. We can partially apply our conclusions over a certain time span, but the precision of our conclusions diminishes as the time span increases.