Customer Income-Level Data Analysis

Ziling Gong

February 2026

1 Introduction

The objective of this project is to develop a predictive model that classifies individuals into income groups ($< 50K$ vs. $> 50K$) using demographic, education, and employment-related attributes from the Census Bureau dataset. To be specific, I utilized a XGBoost model to conduct binary classification, achieving a test accuracy of 0.88 and recall rate of 0.87. Furthermore, I build a customer segmentation unsupervised model clustering 5 groups for marketing purposes. 5 customer segments are Traditional Households, Growth-oriented Households, Next-Gen Saver, Aggressive Investors, and Affluent Retirees. From a business perspective, these models can support targeted marketing campaigns and product personalization.

2 Exploratory Data Analysis

The Census Bureau dataset consists of 199,523 rows and 42 columns. The dataset contains a mix of numerical variables such as age, $wageperhour$, $capitalgains/losses$, etc, and categorical variables such as $education$, $occupation$, $maritalstatus$, etc. The target, $label$, is a binary variable that refers to the income level of the individual. Initial exploration showed a strong class imbalance, with 93.7% of $< 50K$ group and 6.21% of $> 50K$ group. It implies that a stratified train-test split should be implemented, and the classification model should consider balanced error control (ie. recall rate, f1 score). Another variable $weight$ refers to sampling weight by census sampling method which will also be considered during model training, as I want to weight samples by their population.

I observed log-scaled frequency distribution of some economic indicators. From Fig. 1, wage_per_hour, capital_gains and dividends from stock are heavily right-skewed. It suggests income separation is driven by a small subset of financially active individuals. In particular, the large number of individuals with high capital gains stands out in the bar chart on the right. But the log-scaled capital_loss performs a bell-shaped distribution.

Fig. 2 shows that the income level is relatively balanced between gender across different races. Fig. 3 shows that the mean age of 46 for $> 50K$ group is higher than the one of 46 for $< 50K$ group.

Fig. 4 compares the distribution of income level across major industries. The 'Not in Universe or Children' category was excluded due to its high volume and minor information. The data demonstrates that top 3 major industries for the $< 50K$ group are retail trade, education, and durable goods manufacturing. In contrast, the for $> 50K$ group is most prevalent in durable goods manufacturing, finance/real estate industry, and 'other professional industries'. Notably, $> 50K$ group represents approximately double the proportion of the $< 50K$ group in 'mining', 'communications', 'other professional services', and 'utilities and sanity services'.

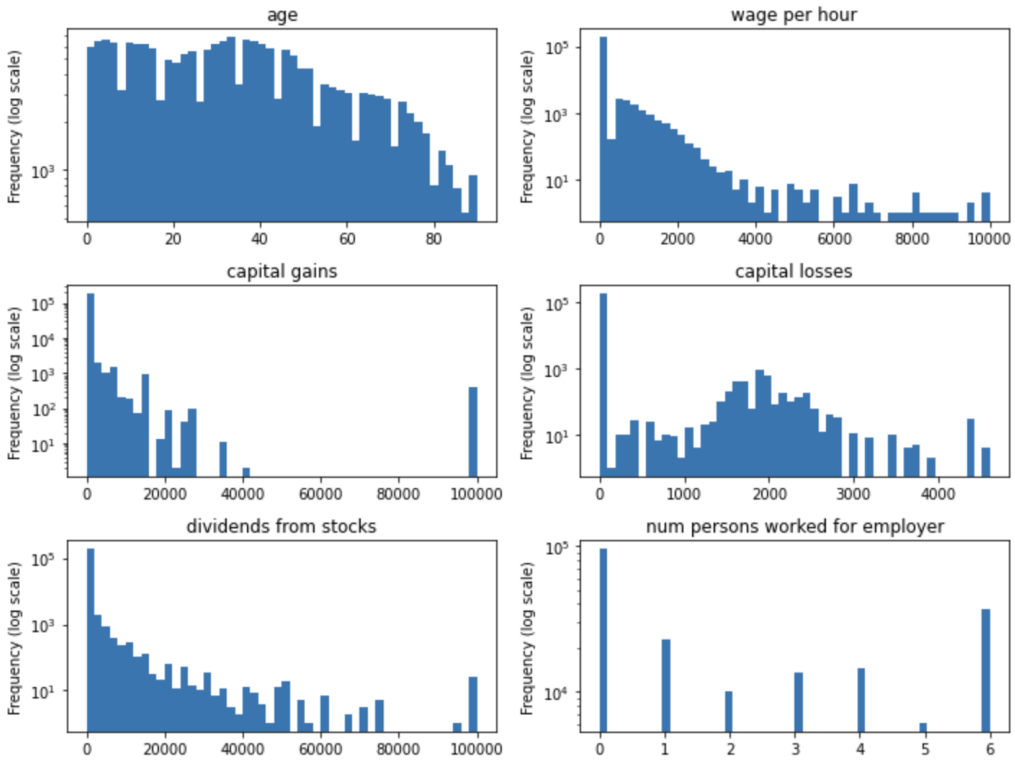

Figure 1: Distribution of Numerical Features

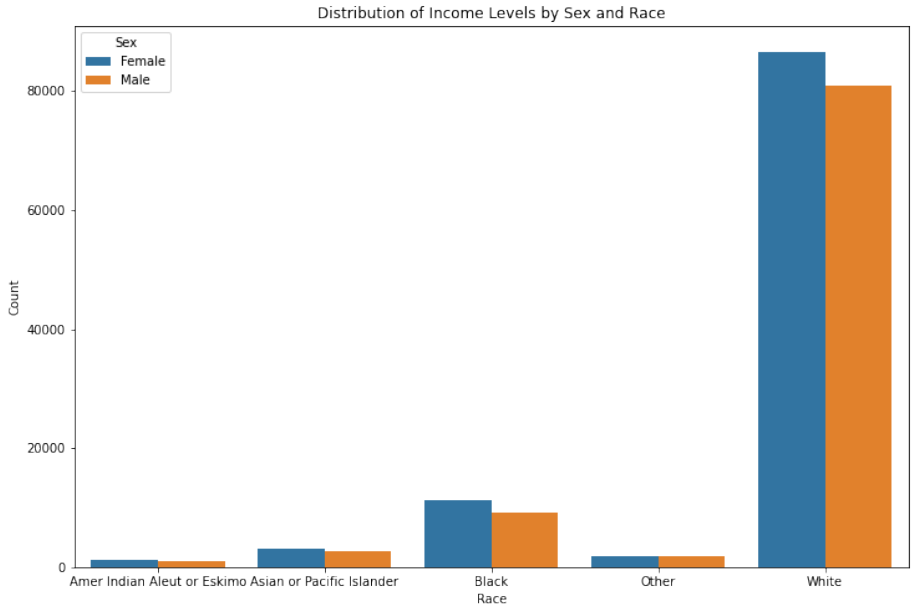

Figure 2: Income Level distribution

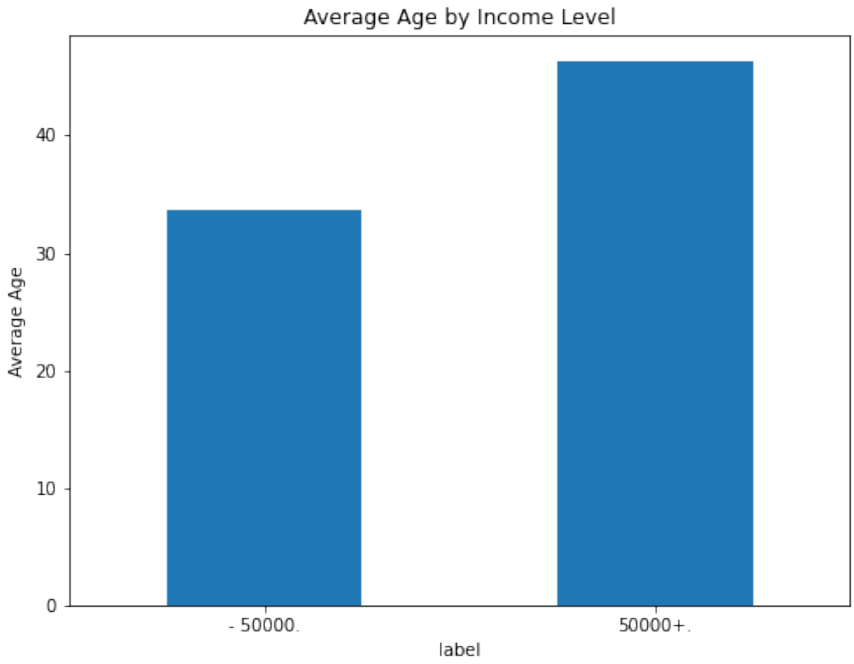

Figure 3: Average Age by income level

3 Data Preprocess

The dataset doesn't showcase large missingness, except more than 800 entries missing in hispanic_origin. I simply dropped the missing data as we had a large dataset. Features were divided into 7 numerical columns and 33 categorical columns. The target is $label$, which is recoded into '0' and '1' where 1 represents $> 50K$ group. The sampling weight variable is not included in features as used during training to reflect population-level importance. Since categorical variables such as 'country of birth father', 'country of birth mother', 'country of birth self', 'citizenship' contain high cardinality, I dropped these three features. Plus, 'state of previous residence', 'detailed industry recode', 'detailed occupation recode', have high cardinality and have duplicated information in other features such as 'major industry code' and 'region of previous residence', I also dropped former columns.

I implemented a preprocessing pipeline to **standardize** numerical features and **one-hot encode** selected categorical variables. Standardization was important for model stability, as it accounts for the high variance between large-scale values in capital-related columns and smaller-scale values such as hourly wages. One-hot encoding expanding to **328** features creating sparse and high-dimensional data.

For the classification model, I used stratified train-valid-test splitting, with 60% training, 20% validation and 20% testing. Stratification preserves income distribution across splits, ensuring generalized evaluation under imbalance. In contrast, the segmentation model uses full data set to do clustering, as it is unsupervised learning.

4 Income-Level Classification

Due to the significant class imbalance, accuracy is an insufficient metric to measure the model performance. I prioritized a high recall rate to capture the maximum number of potential high-income customers. In this

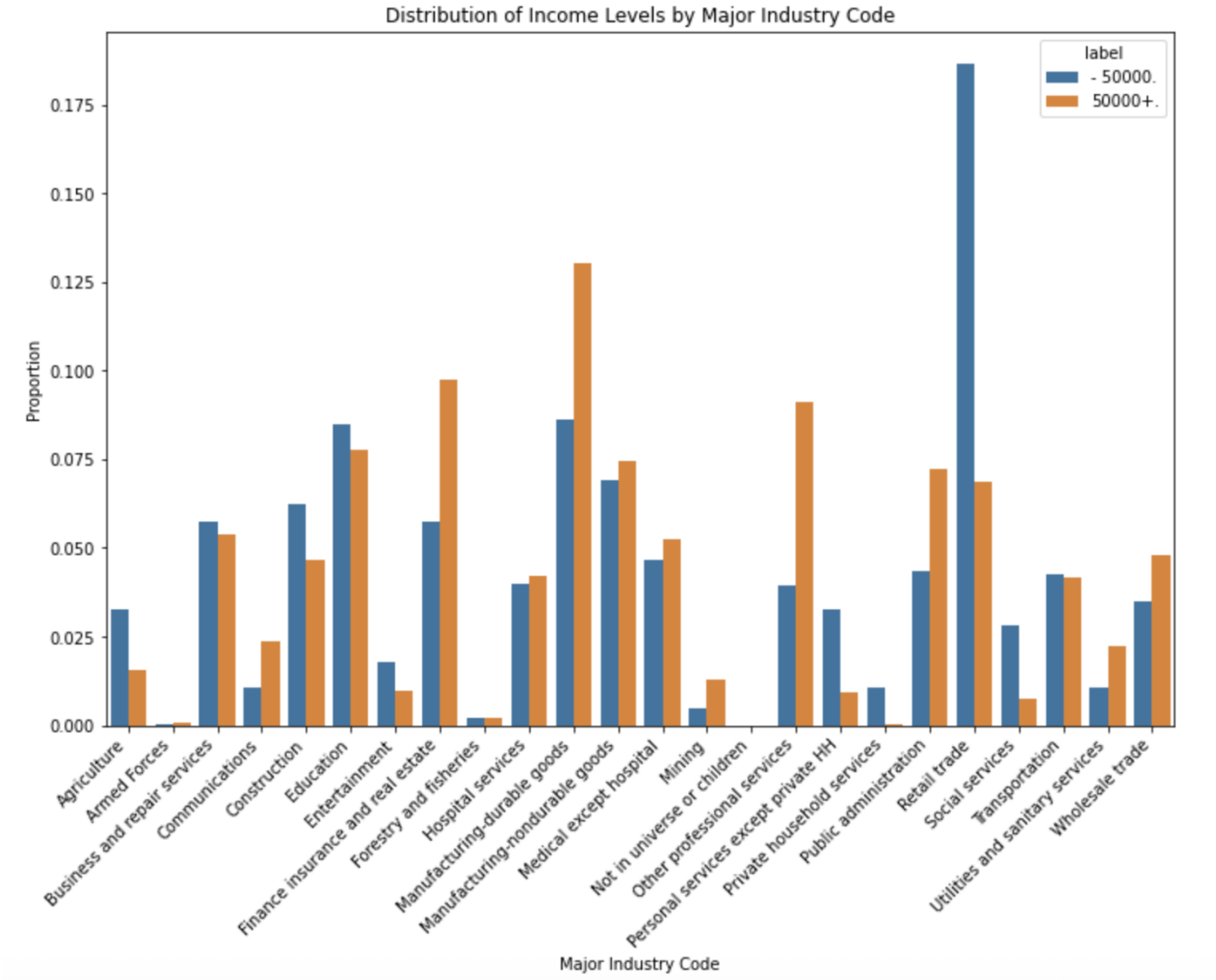

Figure 4: Income Level across Major Industry

marketing context, recall is preferred over precision because our objective is to identify as many qualified leads as possible for specialized investment outreach.

4.0.1 Baseline (Logistic Regression)

I chose Logistic Regression with **L1** regularization for sparse learning and reduced dimension. Plus, I used **SAGA solver** which supports sparse high-dimensional data, had **class-weight balancing** and included **sample weights** during the training. It results in 86% validation accuracy, 0.89 recall rate and 0.42 f1 score.

As baseline uses 0.5 threshold in default, I further tuned threshold using validation set, and the optimal threshold for high recall rate is 0.61. Using the optimal threshold, I obtained 89% validation accuracy, 0.84 recall rate and 0.49 f1 score.

4.0.2 XGBoost

As income prediction likely involves nonlinear relationship and feature interactions, I also trained XGBoost to capture nonlinear relationships. Because I want to ensure precision and recall balance, not overly predict false positives. XGBoost is trained with:

- `n_estimators = 500`
- `max_depth = 5`
- `scale_pos_weight` to address imbalance
- `eval_metric = aucpr`

XGBoost achieves 88% validation accuracy, 0.88 recall and 0.47 f1 score.

Comparing PR-AUC between Logistic Regression (0.603) and XGBoost (0.66), XGBoost achieves better precision and recall balance, so I use XGBoost to classify income level finally. The test accuracy is 0.88, recall is 0.87 and f1 score is 0.47.

Model	Accuracy	Recall	F1
Logistic Regression (default threshold)	0.86	0.89	0.42
Logistic Regression (tuned threshold = 0.61)	0.89	0.84	0.49
XGBoost (Validation)	0.88	0.88	0.47
XGBoost (Test)	**0.88**	**0.87**	**0.47**

4.0.3 Discussion

Since **XGBoost** captures nonlinear interaction between features, we are able to model rare but high-impact patterns, such as finding small groups with high income probability. It is good for our final classification outcome. However, logistic regression is interpretable so that we are able to explain which feature impacts income differs the most. If we want to analyze indicators, **Logistic Regression** remains valuable for interpretability and feature attribution. Since both models already achieved good accuracy and recall rate above 85%, we can even tune more for better specialized prediction in future.

In summary, XGBoost generates prediction scores and labels. Marketing team can target top customers for high-value investment, but also differentiate outreach strategies for lower-income segments.

5 Customer Segmentation

The objective is to better understand the customer base for targeted marketing campaigns or allocate promotion budgets more efficiently. I'm using **K-Means** unsupervised learning to cluster customers. Because K-Means scales well to large dataset and produce interpretable centroid-based segments.

Using the previously cleaned dataset of demographic and financial attributes, I applied **Truncated SVD** to reduce dimensionality before clustering. Evaluating the silhouette score on a random subset revealed that a **10**-dimensional latent space is optimal. This indicates that a lower-dimensional projection captures the core behavioral variances while filtering out the noise inherent in high-cardinality features. Additional dimensions likely encode 'micro-categories' that do not contribute meaningfully to segment separation. The usage of silhouette score is due to its ability to consider both inter-cluster and intra-cluster distance.

I then used the same method to find **k=5** is the optimal cluster size. K-means produces 5 clusters with distribution in Fig. 5 and Fig. 6 shows the reduced 2D visualization.

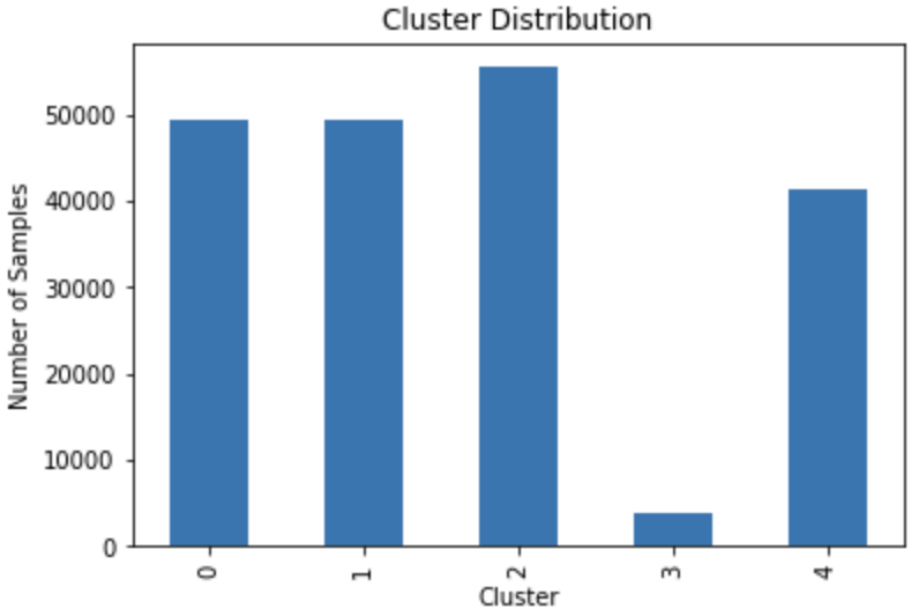

Figure 5: Cluster Distribution

5.0.1 Segments Profile

Mean values of numerical features by cluster is displayed in Fig. 7, highlighting the financial and demographic centroids of each group. 5 groups are defined as following:

Cluster 0: **Traditional Household** (stable income)

Cluster 1: **Growth-Oriented Households** (actively growing wealth through market instruments)

Clusters 0 and 1 share nearly identical demographic profiles in terms of gender and education. Their average age is 38 in both, similar gender distribution, and mostly married with a spouse at present. However, Cluster 1 is distinguished by higher average capital gains and dividends, suggesting a more active investment portfolio despite similar backgrounds.

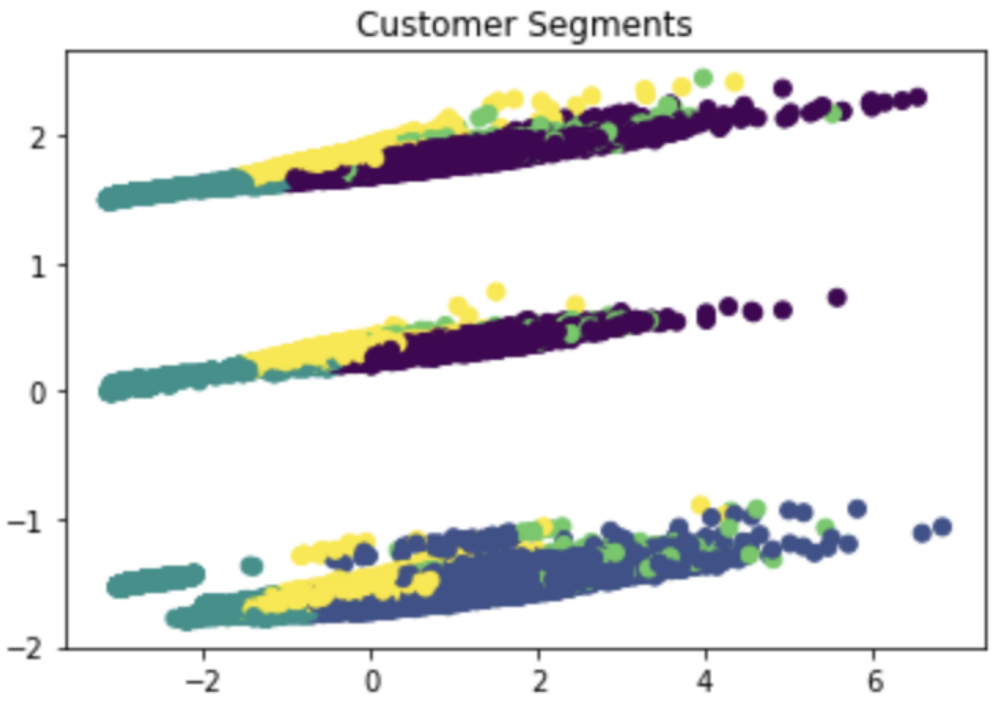

Figure 6: Customer Segment Dimension-Reduced View

clusters	age	wage per hour	capital gains	capital losses	dividends from stocks	num persons worked for employer	weeks worked in year
0	38.357379	107.085968	744.817838	0.319772	176.433524	3.666457	44.586120
1	38.387110	109.870073	838.624598	0.384785	203.298417	3.823987	44.925199
2	8.379821	0.510192	0.102356	0.000000	0.897013	0.049464	0.224736
3	43.978175	79.422298	0.000000	1943.011307	727.906127	3.469103	41.672890
4	59.298533	0.184171	204.592291	0.503401	430.053840	0.101843	0.718135

Figure 7: Mean Value of Features by Cluster

Cluster 2: **Next-Gen Savers**
This segment consists of children with a mean age of 8, primarily enrolled in school and showing no current capital gains or dividend income. While not immediate revenue drivers, this group represents a high-growth segment with significant long-term lifetime value. Marketing efforts should target their parents, who are likely to prioritize educational savings and long-term custodial investments.

Cluster 3: **Aggressive Investors**
With a mean age of 43, this professional segment is characterized by a high-activity investment profile. From Fig. 7, while they maintain significant dividend-yielding assets (averaging 717), they also report substantial capital losses exceeding 1000, that is much higher loss than cluster 0 and 1. This suggests a group of active traders or investors currently undergoing portfolio rebalancing or tax-loss harvesting within their professional-tier income bracket.

Cluster 4: **Affluent Retirees**.
This segment is characterized by an older demographic (mean age of 59) with significant passive income, averaging 430 in stock dividends. From 7 and 8, most members of this group are no longer in the

workforce. Notably, this cluster has the highest concentration of high-income individuals, with a 65:1 ratio of the $> 50K$ group relative to the $< 50K$ group, shown in 9.

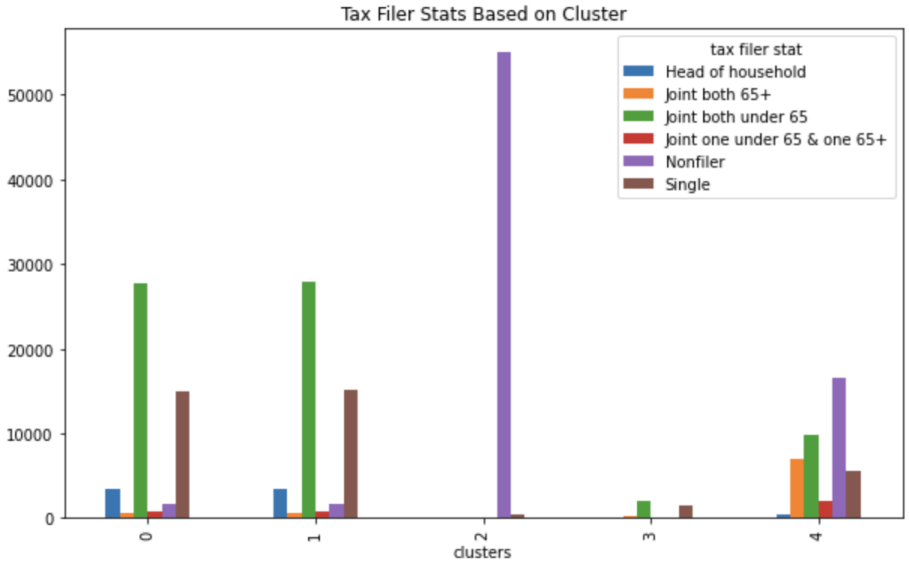

Figure 8: Tax Filer stats on Cluster

clusters	label	
0	-50000.	44427
	50000+.	4952
1	-50000.	43811
	50000+.	5640
2	-50000.	55433
	50000+.	1
3	-50000.	2645
	50000+.	1158
4	-50000.	40825
	50000+.	631

Figure 9: Count of Income Level by Cluster