# Segmentation of US Demographics Data

Target Internship Project Report

Vishwajit Prakash Hegde
Manager:Yasaswi Chodavarapu
Mentor:Vivek Payasi

July 31, 2020

---

## Contents

---

# 1 Project Overview

Each local area has a set of unique population demographic attributes. The population demographics of a given area has a great influence on the products which are being bought in that area. Knowing the demographic attributes which are unique for a given area is important for the retail stores as it helps them in displaying the products which are usually bought by people of that demographics and educating their availability to them. It also helps in identifying the population segments which are potentially under served due to the absence of targeted marketing or the unavailability of some products in the store. **This project aims at grouping the localities based on their demographic attributes and finding the attributes which are unique to them.**

# 2 Data Selection

US demographics data is available on the US Census Bureau website. The population count estimates is available at national, state, county, zip code (5 digit code) and block level. The population counts at **zip code** level is chosen. The reason is that each Target store serves a few zip codes (10-15 zip codes). Also, one cannot generalize the demographics at the county level and it is not practical to define demographics at the block level (9 digit code) as the population is very small. There are a number of choices available for selecting the demographic attributes on the website. Both 1-year and 5-year population estimates are available. Since the latest 1-year estimates is not available, **5-year estimates** is chosen. The demographic attributes data is available at different levels of detail.
The following options are available to choose from:

- **Detailed Tables** : contain the most detailed cross-tabulations, many of which are published down to block groups. The data are population counts. There are over 20,000 variables in this data set.
- **Subject Tables** : provide an overview of the estimates available in a particular topic. The data are presented as population counts and percentages. There are over 18,000 variables in this data set.
- **Data Profiles** : contain broad social, economic, housing, and demographic information. The data are presented as population counts and percentages. There are over 1,000 variables in this data set.

**Data Profiles** is chosen for the analysis as the number of variables is comparatively small and hence easy to explore and interpret the data.

# 3 Data Overview

Data is available for a total of **33120** zip codes. Each zip code is a five-digit number. There are a total of **673** attributes in the Data Profiles data set. For most of the attributes both the **population estimates** and the **percentage estimates** are available. Some of the attributes are not population counts. For example, average household size, unemployment rate etc.
The attributes are broadly categorized into following groups:

- Social Characteristics (**DP02**)
- Social Characteristics in Puerto Rico (**DP02PR**)
- Economic Characteristics (**DP03**)
- Housing Characteristics (**DP04**)
- Demographic and Housing Estimates (**DP05**)

Each group data can be separately downloaded using its ID (**DPXX**).
Each attribute has an ID. It starts with the table ID (**DPXX**) followed by an underscore and a four digit number. If it is population count estimate, then the suffix is **E**. If the attribute is percentage estimate, then the suffix is **PE**. An example is **DP02_0001PE**. Here is the Link to a spreadsheet containing all the attributes and their descriptions.

# 4 Data Preprocessing

**Percentage Estimates (DPXX_XXXXPE)** are used for all the analysis as comparing population count estimates between different zip codes does not make sense.

## 4.1 Dropping the attributes corresponding to Puerto Rico

There is a separate set of social characteristics attributes (**DP02PR**) for zip codes corresponding to **Puerto Rico**. These columns have values for only Puerto Rican zip codes and rest of the cells are empty. DP02 attributes have values in non Puerto Rican zip codes only. Hence DP02PR attribute values are copied to DP02 attribute columns as they are mutually exclusive.

## 4.2 Dropping the zip codes with no population

It is found that for **578** zip codes, the population count estimate is zero. Hence, rows corresponding to these zip codes are dropped. The total number of remaining rows is **32542**.

## 4.3 Filling the missing values

For some attributes like average household size, average household income, **DPXX_XXXXPE** columns have missing values as there are no equivalent percentage values. Such columns are replaced by **DPXX_XXXXE** columns. In some columns, few of the rows have missing values. They are replaced by median values of the respective columns.

## 4.4 Manual attribute selection

Not all the attributes in Data Profiles are useful for our application. Many attributes are repetitive. For example, there are separate columns for average household income and median household income. Some attributes are irrelevant for demographic analysis. For example, number of grandparents living with own grandchildren. Some of the attributes are too specific and their populations are very small. For example, population percentage of American Indian and Alaska Native belonging to Navajo tribe. The housing characteristics (DP04) attributes are also unnecessary for the analysis. Hence all such attributes are removed in order to get better interpretable results.

## 4.5 Dropping the correlated attributes

Some of the attributes have high correlation among each other. Keeping the highly correlated variables is equivalent to representing the same data twice and that leads to higher weight given to the variable compared to the non-correlated ones. This can produce distorted results. Hence only one among the highly correlated attributes is retained and rest are dropped. **Figure 1** shows correlation heatmap for attributes belonging to the category Education Attainment. Correlation between **DP02_0060PE** and **DP02_0066PE** is **-0.81**. The respective attributes are percentage of population with 9th to 12th grade education (without high school diploma) and percentage of population who are high school graduate or higher. They are clearly inversely correlated. Hence keeping both the attributes is not ideal.

# 5 Exploratory Data Analysis

## 5.1 Comparing attribute mean values

The plot shown in **Figure 2** compares the mean values of all the attributes used in the analysis. It can be seen from the figure that some attributes have a very large mean and some very small. Attributes like percentage of white population, percentage of people working in private sector, percentage of households

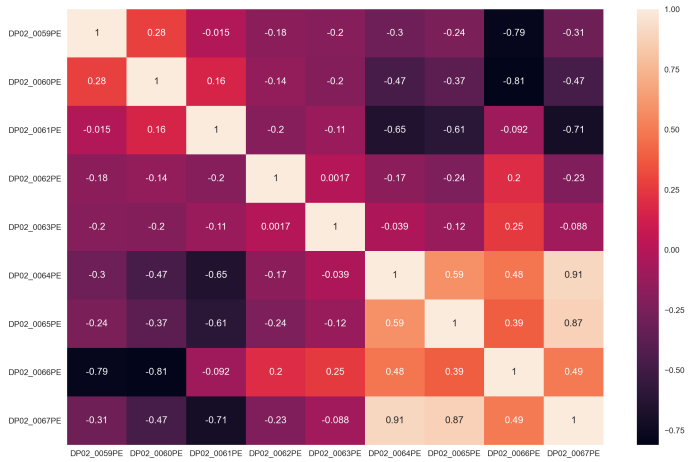



Figure 1: *Correlation Heatmap; Education Attainment*

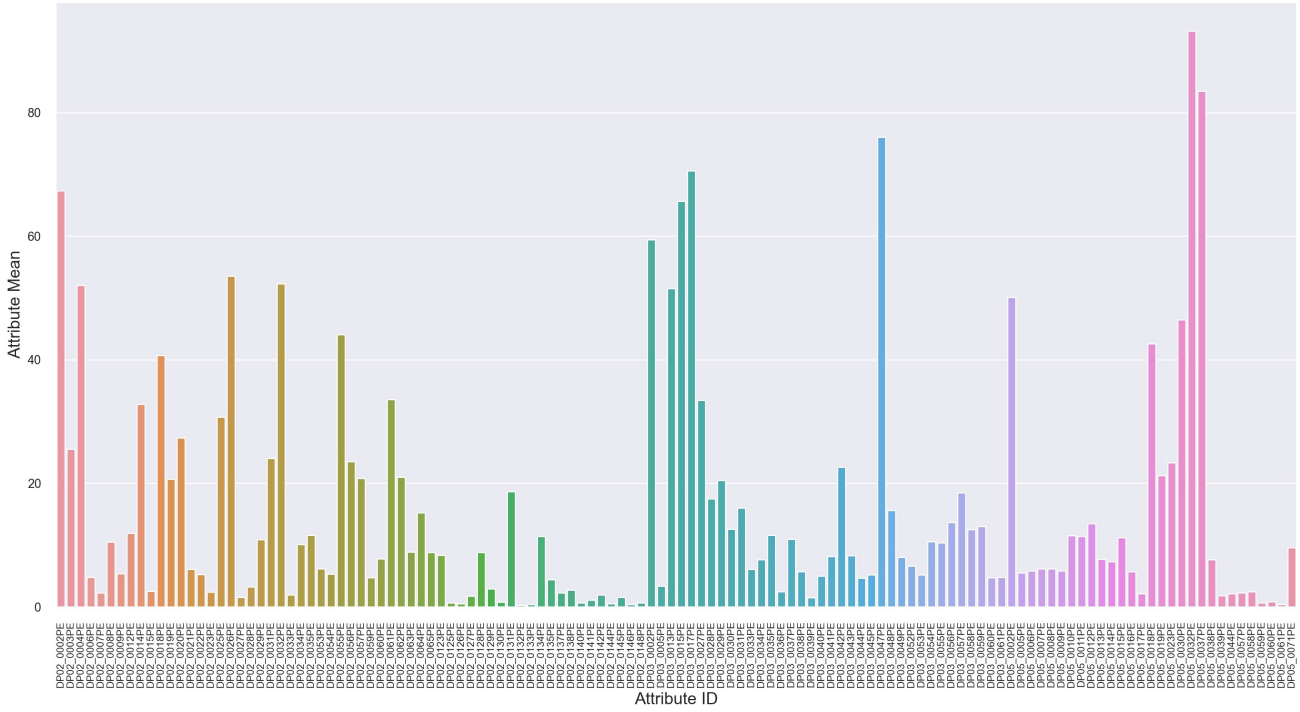
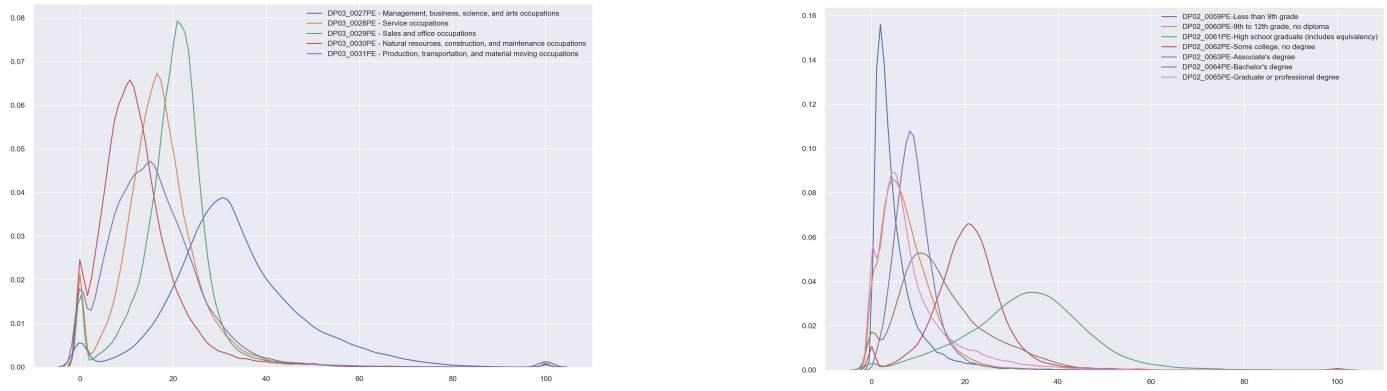



Figure 2: *Attribute Means*

with family have very high means while percentage of people older than 85 years, percentage of people belonging to White and Asian race (mixed race) are very low. Hence it is important to scale all the attributes to same range so that all the attributes are used fairly in forming the clusters.

## 5.2 Density plots

Since it is not practical to visualise the density plots of all the attributes at once, attributes belonging to a certain category like income, race or occupation are visualised at once. The plots in **Figure 3(a)** and in **Figure 3(b)** depict the density plots of all the attributes corresponding to categories occupation and education respectively.




(a) *Occupation* (b) *Education Attainment*

Figure 3: *Density Plots*

## 5.3 Principal Component Analysis

As the number of variables is too large, a direct application of any distance based clustering algorithm would not yield good results as the euclidean distance loses meaning in higher dimensions (curse of dimensionality). Hence the first attempt at dimensionality reduction is done using Principal Component Analysis. The plot in **Figure 4** shows the **explained variance ratio** of each principal component. Larger the explained variance ratio for the first few principal components, the initial components would

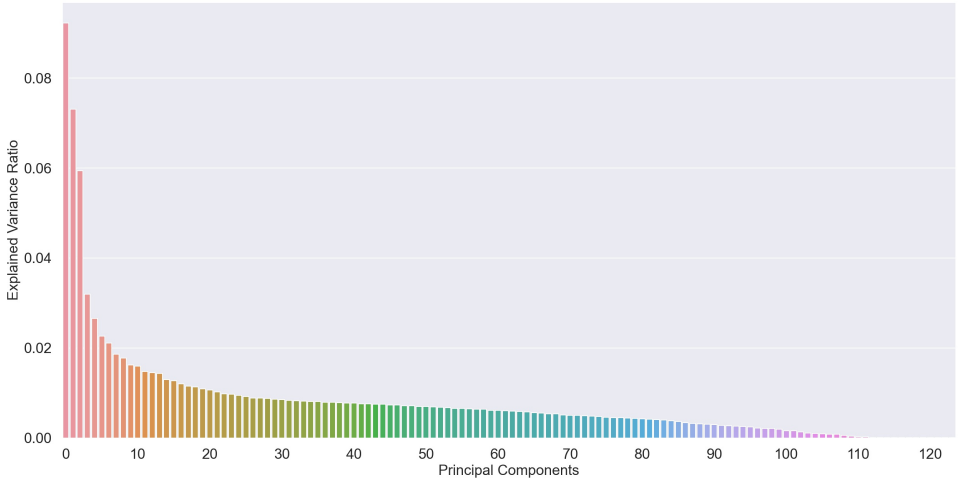



Figure 4: *PCA; Explained Variance Ratio*

have captured the variance in the data and hence better the results. The sum of initial few principal

components explained variance ratio should be as close to one as possible. This is not the case with our data. Hence dimensionality reduction using PCA is not a viable option.

# 6 Attempts at Clustering The Data

The following clustering methods are attempted and the issues faced (results in some cases) are noted.

## 6.1 K-Means

Since the euclidean distance loses its meaning in high dimensions, direct application of K-Means clustering to high dimensional data does not produce a good set of clusters. **Figure 5** shows K-Means inertia plot for number of clusters between 100 and 500. The plot does not contain an "elbow".

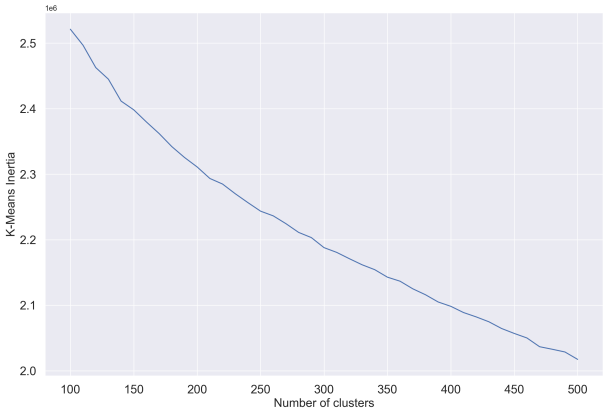



Figure 5: *K-means Inertia Plot*

## 6.2 Hierarchical Clustering

This algorithm also uses distance metrics like Euclidean, Manhattan etc. for finding the clusters, hence the problems similar to K-Means are encountered. Also, finding the appropriate number of clusters is difficult and time consuming.

## 6.3 Density Based Clustering (DBSCAN)

Using density based clustering results in many points being classified as noise, a large number of points in a single cluster and a very few in others. This is not ideal for our purpose.

## 6.4 Self Organising Maps

Self Organising Map is used to project data into two dimensions. It contains a rectangular grid of nodes, each with a weight vector whose dimension is same as the dimension of the data. The data points are assigned to the nodes based on the distance between the node's weight vector and the data point. The weight vectors are learned while assigning the data points to the nodes. Clustering is done using the nodes' weight vectors.
Deciding the number of nodes is very important for getting good results. Different grid sizes ranging from [30, 30] to [100, 100] are tried. Based on topographic error and quantization error, a grid size of **[60, 60]** is chosen. For demographic segmentation, we need the number of clusters to be in the order of 100.

Using "elbow" method, the number of clusters is decided as **200**. **Figure 6** shows the resulting clusters in two dimensions. **Silhouette score** is used to see whether a point belonging to a cluster is closer to

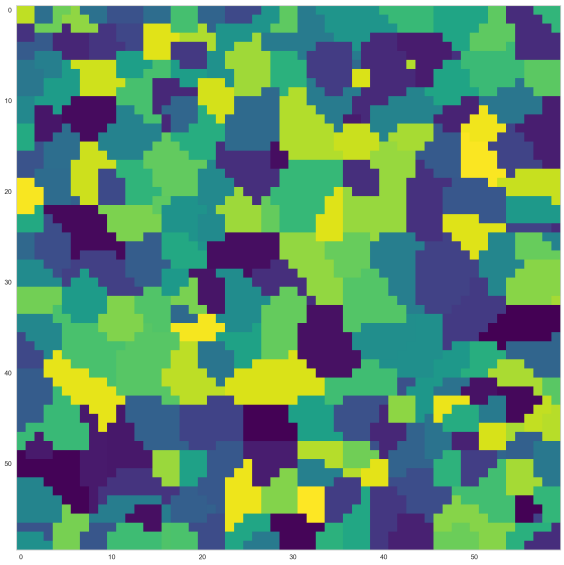



Figure 6: *SOM grid*

its cluster centroid or to the centroid of the cluster nearest to its own. It is used to validate consistency within clusters of data. It lies between -1 and 1, 1 being the best score. Silhouette score for the above set of clusters is found to be **-0.06**. This means that the clusters lack consistency.

### 6.5 t-SNE

The next attempt at dimensionality reduction is done using **tSNE**. This technique converts similarities between data points to joint probabilities and tries to minimize the Kullback-Leibler divergence between the joint probabilities of the low-dimensional embedding and the high-dimensional data. First of all tSNE does not work well if the input dimension is very large(greater than 50). Hence the dimension is reduced using **PCA** so that **95 percent of the variance** is retained. **Perplexity** is a very important hyper parameter for tSNE algorithm. Different values of perplexity can produce **tSNE** plots shaped differently altogether. The empirical criteria proposed in "Automatic Selection of t-SNE Perplexity" by Cao and Wang is used for selecting perplexity. The criteria is to minimize S where:

$$S = 2KL(P||Q) + \log(n)\frac{Perp}{n} \tag{1}$$

where $Perp$ is the perplexity value, $n$ is the number of samples and $KL(P||Q)$ is the KL divergence. The resulting value of perplexity is 1000. **Figure 7** shows the resulting tSNE plot. The plot does not provide any information about the existence of clusters.

### 6.6 UMAP

**UMAP** is a very recent dimensionality reduction technique. **UMAP** constructs a high dimensional graph representation of the data, then optimizes a low-dimensional graph to be as structurally similar

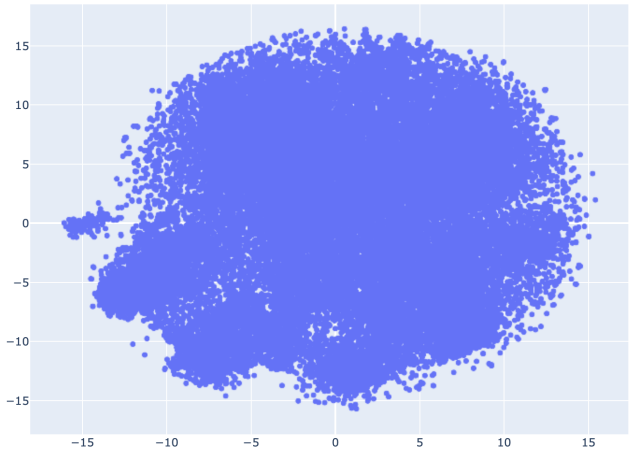



Figure 7: *tSNE plot*

as possible. The two important parameters to be fed to the algorithm are **n_neighbors** and **min_dist**, which are effectively used to control the balance between local and global structure in the final projection. **n_neighbors** is the number of nearest neighbors used to construct the initial high-dimensional graph. Low **n_neighbors** values will push **UMAP** to focus more on local structure by constraining the number of neighboring points considered when analyzing the data in high dimensions, while high values will push **UMAP** towards representing the big-picture structure while losing fine detail. For our purpose, low values of this parameter are preferred as local structure is more important for producing a large number of clusters. **min_dist** controls how tightly **UMAP** clumps points together, with low values leading to more tightly packed embeddings. Larger values of **min_dist** will make **UMAP** pack points together more loosely, focusing instead on the preservation of the broad topological structure. **min_dist** values are chosen in the range [0, 1]. Since we are more concerned with local structure, smaller values of this parameter are preferred. **n_neighbors** is chosen as **5** and **min_dist** is chosen as **0.001**. The plot in **Figure 8** shows the **UMAP embedding** for our dataset. For our purpose, we need the number of

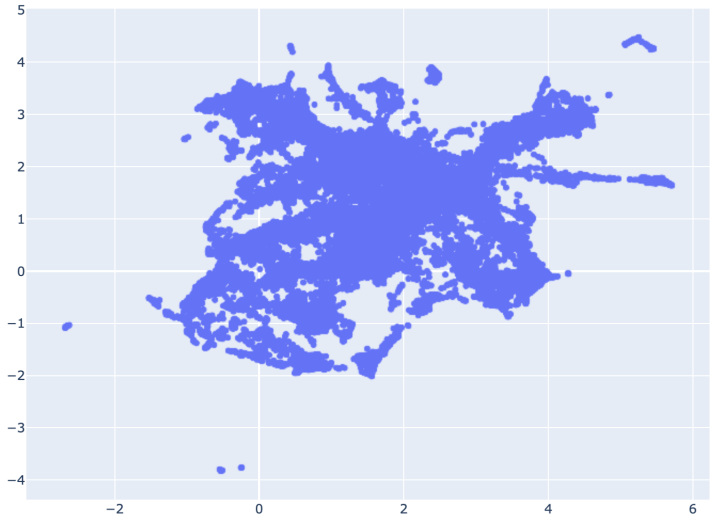



Figure 8: *UMAP plot*

clusters to be in the order of 100. The above plot consists of a few small cluster of points and a big clump of points in the middle. This is not very suited for our purpose.

# 7 Proposed Clustering Method

Direct application of any of the above methods does not produce accurate and interpretable results with the demographics data. A two stage clustering method is proposed for getting more interpretable results.

## 7.1 Segregating the attributes into different categories

The attributes can be segregated into different categories like age, gender, income, race, etc. An example of an attribute in the category race is the percentage of population belonging to the race Asian. The following categories are identified and the attributes are manually segregated:

1. Households
2. Relationship
3. Marital Status
4. School Enrollment
5. Education Attainment
6. Ancestry
7. Employment Status
8. Occupation
9. Industry
10. Class of Worker
11. Income
12. Sex and Age
13. Race

There are 124 columns (attributes) altogether.

## 7.2 Performing clustering separately in all the categories

On an average, each category contains 10 attributes. Hence even the direct application of K-Means algorithm can produce decent results. On an average, 9 clusters are formed in each category. The details of stage 1 clustering are presented in section 8.

## 7.3 Using stage 1 clustering results for stage 2 clustering

The results of the stage 1 clustering can be represented by a new data set which contains category names as columns and zip codes as rows. The values in the data set represent the cluster label to which a given zip code belongs to in a given category. One cannot use Euclidean distance as a metric for clustering this data because it is categorical. The distance between two data points is proportional to the number of categories (columns) in which the cluster labels are different. **Hamming distance** uses this principle in calculating the distance between any two data points. The number of clusters to be formed can be decided based on the silhouette score. The details of stage 2 clustering are presented in section 9.

# 8 Stage 1 Clustering

K-Means clustering algorithm is directly applied to the data belonging to each of the category. "Elbow" method is used to find the optimum number of clusters.

### 8.1 Stage 1 Clustering Results

| Category | Number of Attributes | Number of Clusters | Silhouette Score |
|---|---|---|---|
| Households | 10 | 8 | 0.168 |
| Relationship | 6 | 8 | 0.198 |
| Marital Status | 10 | 10 | 0.145 |
| School Enrollment | 5 | 7 | 0.224 |
| Education Attainment | 7 | 8 | 0.193 |
| Ancestry | 21 | 17 | 0.117 |
| Employment Status | 5 | 8 | 0.176 |
| Occupation | 5 | 8 | 0.204 |
| Industry | 13 | 12 | 0.110 |
| Class of Worker | 3 | 7 | 0.347 |
| Income | 10 | 11 | 0.161 |
| Sex and Age | 19 | 9 | 0.128 |
| Race | 10 | 8 | 0.386 |

### 8.2 Cluster validation using XGBoost Classifier

XGBoost is a decision-tree-based ensemble Machine Learning algorithm that uses a gradient boosting framework. It is widely used in classification problems because of its ability to learn non linear patterns. Apart from silhouette score there is no other reliable way to check the accuracy of clustering. Hence using cluster labels as target labels, XGBoost classifier is trained on 80% of the data and the accuracy score on the test set is noted. The following table displays accuracy scores for all the categories.

| Category | Accuracy Score |
|---|---|
| Households | 0.967 |
| Relationship | 0.973 |
| Marital Status | 0.964 |
| School Enrollment | 0.991 |
| Education Attainment | 0.976 |
| Ancestry | 0.949 |
| Employment Status | 0.965 |
| Occupation | 0.981 |
| Industry | 0.97 |
| Class of Worker | 0.992 |
| Income | 0.966 |
| Sex and Age | 0.964 |
| Race | 0.984 |

### 8.3 Feature importance using XGBoost Classifier

XGBoost classifier can be used to find the attributes based on which a cluster is formed. Since it is decision-tree based classifier, it outputs a feature importance score for each attribute which basically tells us to what extent the given attribute is used in the classification. In order to find important attributes for each cluster, XGBoost is trained to perform binary classification separately for each cluster in each category. **Figure 9** shows Feature Importance for 4 clusters belonging to the category Race.


Figure 9: *XGBoost Feature Importance; Race*

# 9 Stage 2 Clustering

**Figure 10** shows the resulting data set from stage 1 clustering. Each value indicates the cluster to which a zip code belongs to in a category. Similarity between two zip codes is proportional to the number of categories in which they belong to the same cluster.

| | Households | Relationship | Marital Status | School Enrollment | Education Attainment | Ancestry | Employment Status | Occupation | Industry | Class of Worker | Income | Sex and Age | Race |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **zcta** | | | | | | | | | | | | | |
| **43964** | 6 | 0 | 6 | 1 | 1 | 1 | 3 | 0 | 7 | 0 | 9 | 8 | 0 |
| **28216** | 5 | 2 | 4 | 0 | 7 | 12 | 7 | 2 | 9 | 2 | 5 | 3 | 1 |
| **28277** | 7 | 1 | 0 | 1 | 5 | 1 | 7 | 1 | 9 | 0 | 1 | 3 | 6 |
| **28278** | 7 | 2 | 0 | 0 | 7 | 0 | 0 | 2 | 9 | 0 | 10 | 3 | 6 |
| **28303** | 0 | 0 | 4 | 0 | 7 | 1 | 3 | 2 | 7 | 1 | 9 | 8 | 1 |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| **98279** | 4 | 5 | 0 | 0 | 7 | 4 | 2 | 6 | 1 | 6 | 5 | 2 | 0 |
| **98280** | 4 | 5 | 0 | 2 | 5 | 4 | 1 | 1 | 9 | 6 | 5 | 4 | 0 |
| **98311** | 6 | 0 | 0 | 0 | 1 | 1 | 7 | 2 | 7 | 1 | 10 | 8 | 6 |
| **98326** | 0 | 6 | 4 | 2 | 0 | 1 | 3 | 0 | 4 | 1 | 9 | 2 | 6 |
| **98332** | 6 | 1 | 0 | 0 | 7 | 1 | 3 | 1 | 7 | 2 | 1 | 8 | 6 |

Figure 10: *Resulting dataframe from stage 1*

### 9.1 Dimensionality reduction using UMAP

Since the data is categorical, direct application of K-Means is not possible as K-Means algorithm uses euclidean distance. UMAP lets one choose a distance metric from a long list of metrics. Hamming distance is chosen as the distance metric. n_neighbors is set as 5 and min_dist is set as 0.001. **Figure 11** shows the UMAP embedding in two dimensions.

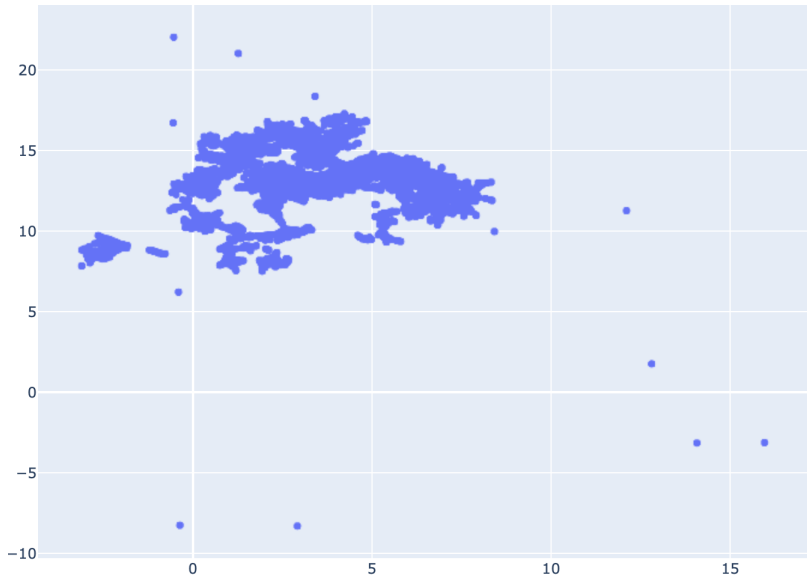



Figure 11: *UMAP plot*

### 9.2 Clustering

UMAP reduced data is fed to K-Means clustering algorithm. Number of clusters ranging from 100 to 500 are tried and the one with the highest silhouette score is chosen. The resulting number of clusters is **300**. Silhouette score when the number of clusters is 300 is **0.506**.

### 9.3 Finding important categories for each cluster

Consider a cluster out of 300. If all the zip codes belonging to this cluster also belong to the same cluster in the category income, one can conclude that income is an important category in defining that cluster. Hence the important categories for a cluster (stage 2) have the same value for most of the zip codes belonging to the cluster (stage 2). Hence importance of a category for a cluster can be quantified as the ratio of maximum number of zip codes with the same value (cluster label corresponding to stage 1) to the total number of zip codes in the cluster. **Figure 12** shows the importance of different categories for clusters 0 to 3.

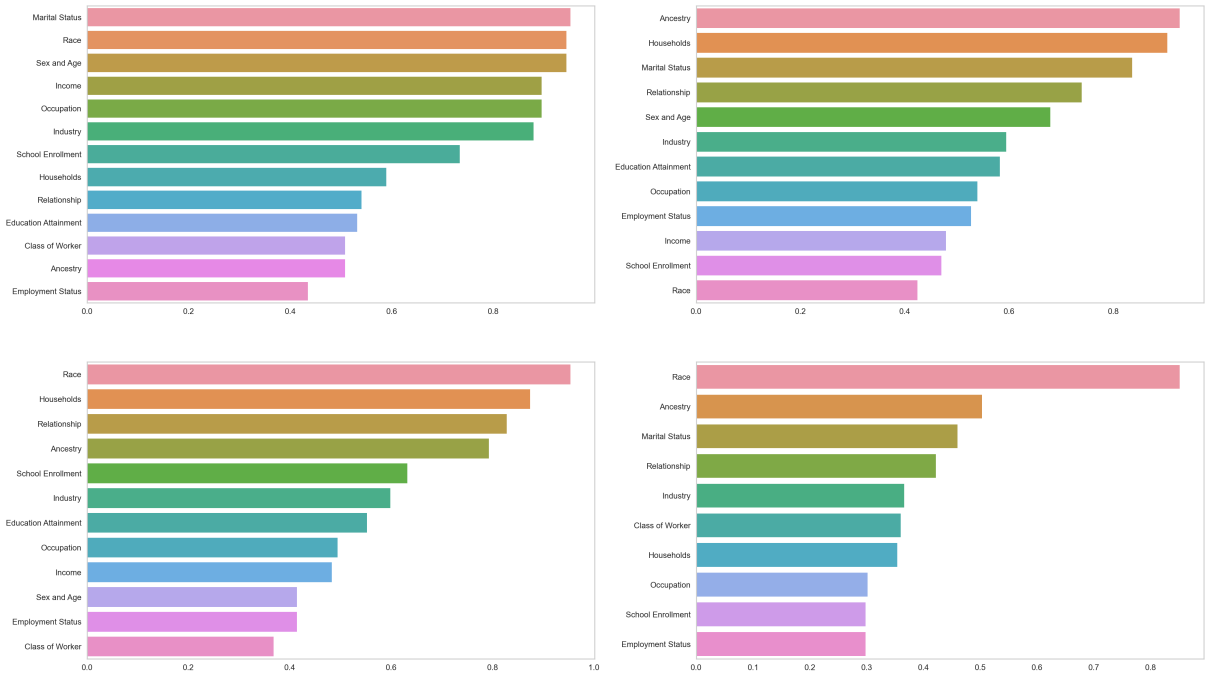



Figure 12: *Category Importance*

## 10 Results

The aim of the project is to find the attributes which define a given zip code's demographic. Steps used for finding the attributes are mentioned below:

- Find the cluster (stage 2) to which the given zip code belongs to.
- Select the categories for which the zip code's cluster label is same as the cluster label of maximum number of zip codes and the ratio (as described in section 9.3) is greater than 0.5. If more than three categories satisfy this condition, select the first three with the highest ratios.
- For each of the categories, zip code's cluster label is known. The feature importance score for all the attributes is found as described in section 8.3. Select the first two with the highest feature importance score.
- It is important to make sure that the zip code is not an outlier in these two attributes. The attributes cannot be considered as important if the cluster mean is far from the actual values of the zip code for these attributes. If the zip code's value is more than 2 sigma apart from the cluster mean for the selected attribute, then the attribute is discarded and the next attribute is selected based on the feature importance score.

The following example shows how the important attributes are found for a randomly selected zip code. The randomly selected zip code is **67001**. It belongs to the cluster **119**. **Figure 13** shows category importance for cluster 119. The top three categories are selected. They are **Race**, **Education Attain-**

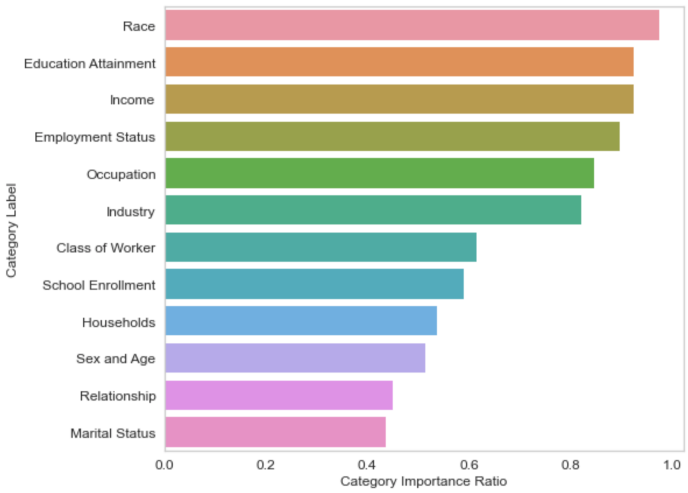



Figure 13: *Category Importance; cluster 119*

**ment** and **Income** and the zip code belongs to clusters **0**, **7** and **10** respectively in these categories. **Figure 14** shows the XGBoost Feature Importance for these clusters. The final set of important features

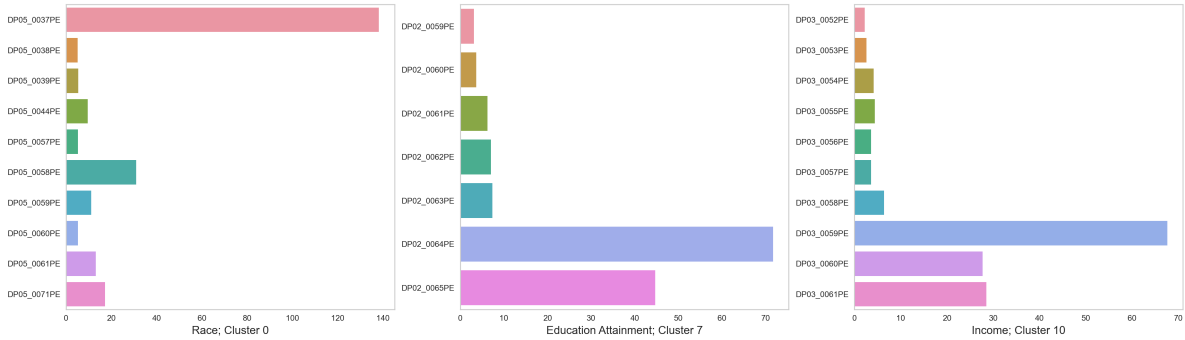



Figure 14: *XGBoost Feature Importance*

for zip code 67001 with their descriptions and categories are listed in the table below.

| Category | Attribute | Description | Value |
|---|---|---|---|
| Race | DP05_0037PE | people belonging to race white | 100 |
| Race | DP05_0058PE | people belonging to mixed race | 0 |
| Education Attainment | DP02_0064PE | people with bachelor's degree | 22.4 |
| Education Attainment | DP02_0065PE | people with graduate or professional degree | 12.4 |
| Income | DP03_0059PE | Household income in the range $100,000 to $149,999 | 22.2 |
| Income | DP03_0061PE | Household income more than $200,000 | 10.2 |

## 11 Proposed Continuation of The Project

One of the uses of knowing the attributes which uniquely define a zip code's demographics is in comparing the Target store's customer data and the population demographics data. This comparison would help in identifying the under served population segments in the area. Also, one can formulate marketing strategies to target such population segments. The following method is a possible way for identifying

under served population segments.

- For each zip code find the closest Target Store.
- Hence, one can find all the zip codes which are associated with a given Target Store.
- Estimate the percentage of customers coming from each zip code based on the population of the zip code and distance from the store. For example, if there are 5 zip codes associated with the store with populations, 10000, 5000, 12000, 13000, 20000 and if they are all at same distance from the store, estimated percentage of customers from these zip codes would be 16.67, 8.33, 20, 21.67, 33.33 respectively (A more rigorous formulation is required in order to include the distance from the store).
- Find the actual percentages using the Target customer data. This can be done given that each customer's zip code is available in the database.
- If for a given zip code, the actual percentage is less than the estimated percentage then it is possible that it is because of the demographics of the zip code. The attributes which uniquely define the zip code's demographics can be used to find the under served population segments in that case.
- It is possible that for all the zip codes belonging to a given cluster, the difference between the estimated and the actual percentage are close as they share similar demographics.