

COLUMBIA UNIVERSITY

IEOR 4111: OPERATIONS CONSULTING

DEPARTMENT OF INDUSTRIAL ENGINEERING AND OPERATIONS RESEARCH

---

## Patient Prioritization Model Report

---

*Authors*

IMPACT CONSULTING GROUP:

Salman ALAHMADI

Paula ANGEL

Roupya BEHERA

Akshat BHUSHAN

Shin Ler LOW

Jordan TANUDJAJA

Hua YAO

Mavis YE

*Client*

NOVO NORDISK



# Contents

<b>Novo Nordisk</b>	<b>1</b>
<b>1 Introduction</b>	<b>2</b>
1.1 Abstract . . . . .	2
1.2 Background . . . . .	2
1.3 Methodology . . . . .	2
<b>2 Patient Relational Database</b>	<b>4</b>
2.1 Objective . . . . .	4
2.2 Data Sources . . . . .	4
2.3 SQL Workbench Tables . . . . .	4
2.3.1 Patient Table . . . . .	4
2.3.2 Prescriber and Plantrak Tables . . . . .	5
2.4 Socioeconomic Flat Table . . . . .	5
2.4.1 Census Tract, County and ZIP Codes . . . . .	6
2.4.2 Conversion to ZIP Codes . . . . .	6
2.4.3 Final Socioeconomic Table . . . . .	7
2.5 Overall structure of database . . . . .	7
<b>3 Data Preprocessing</b>	<b>8</b>
3.1 Objective . . . . .	8
3.2 Methodology . . . . .	8
3.2.1 Creation of a single dataset . . . . .	8
3.2.2 Patient filtration and Feature selection . . . . .	8
3.2.3 Converting categorical variables to numerical . . . . .	8
<b>4 Patient Level Clustering</b>	<b>10</b>
4.1 Objective . . . . .	10
4.2 Scaling and Dimensionality Reduction . . . . .	10
4.2.1 Standard Scaler . . . . .	10
4.2.2 Principal Component Analysis . . . . .	10
4.3 Clustering Algorithm . . . . .	11
4.3.1 Testing different models . . . . .	11
4.3.2 Implementing the GMM algorithm . . . . .	12
4.4 Data visualization . . . . .	12
4.5 Patient Cluster Characteristics . . . . .	13
4.5.1 Cluster 1 - High Priority Patients . . . . .	13
4.5.2 Cluster 2 - Medium High Priority Patients . . . . .	13
4.5.3 Cluster 3 - Medium Low Priority Patients . . . . .	14
4.5.4 Cluster 4 - Low Priority Patients . . . . .	14
4.5.5 Cluster 5 - Separate Group . . . . .	14
<b>5 ZIP Code Level Clustering</b>	<b>15</b>
5.1 Objective . . . . .	15
5.2 ZIP Code Flat Table . . . . .	15
5.3 Clustering Algorithm . . . . .	15
5.4 Data Visualization . . . . .	17
5.5 ZIP Code Group Characteristics . . . . .	18
5.5.1 Group A: High Priority ZIP Codes . . . . .	18

5.5.2	Group B: Medium High Priority ZIP Codes . . . . .	19
5.5.3	Group C: Medium Priority ZIP Codes . . . . .	19
5.5.4	Group D: Medium Low Priority ZIP Codes . . . . .	19
5.5.5	Group E: Low Priority ZIP Codes . . . . .	19
5.5.6	Group F: Separate Group of ZIP Codes . . . . .	19
<b>6</b>	<b>Overall Patient Classification and Ranking</b>	<b>20</b>
6.1	Objective . . . . .	20
6.2	Data Visualization . . . . .	20
6.3	Ranking Process . . . . .	21
6.4	Categories Characteristics . . . . .	22
6.4.1	NNI AOM Enthusiasts . . . . .	22
6.4.2	NNI AOM Convertibles . . . . .	22
6.4.3	NNI AOM Potentials . . . . .	22
6.4.4	NNI AOM Rejects . . . . .	22
<b>7</b>	<b>Recommended Strategies</b>	<b>23</b>
7.1	Objective . . . . .	23
7.2	Marketing Campaigns . . . . .	23
7.3	Investment Efforts . . . . .	24
7.4	Future portfolio: Sema-O . . . . .	24
<b>8</b>	<b>Conclusion</b>	<b>26</b>
<b>A</b>	<b>Appendix: Features of patient flat table</b>	<b>i</b>
A.1	General Information . . . . .	i
A.2	Diagnosis (Dx) information . . . . .	i
A.3	Procedure (Px) information . . . . .	iii
A.4	Prescription (Rx) information . . . . .	iv
A.5	Combined and Foreign Features . . . . .	viii
<b>B</b>	<b>Appendix: Features of Prescriber flat table</b>	<b>x</b>
<b>C</b>	<b>Appendix: Features of Plantrak flat table</b>	<b>xi</b>
<b>D</b>	<b>Appendix: Features of Socioeconomic Flat Table</b>	<b>xii</b>
D.1	Census Tract Level . . . . .	xii
D.2	County Level . . . . .	xiv
<b>E</b>	<b>Appendix: K-Means Algorithm</b>	<b>xv</b>
<b>F</b>	<b>Appendix: K-Prototypes Algorithm</b>	<b>xvi</b>
<b>G</b>	<b>Appendix: Gaussian Mixture Model</b>	<b>xvii</b>
<b>H</b>	<b>Appendix: Patient Level Visualizations</b>	<b>xviii</b>
<b>I</b>	<b>Appendix: ZIP Code Level Visualizations</b>	<b>xx</b>
<b>J</b>	<b>Appendix: Combined Patient and ZIP code level Visualizations</b>	<b>xxii</b>

# NOVO NORDISK

# Introduction

## 1.1 Abstract

Novo Nordisk's (NN) overarching goal is to use their limited resources to prioritize and target patients most likely to take their FDA-approved drug Saxenda®. The objectives of the project is to generate a predictive model to score and rank Persons with Obesity (PwOs) most receptive to taking Saxenda®. The final deliverable is to communicate our model and insights, as well as generate recommendations to the Obesity Leadership Team (OLT) to inform them of potential patient targeting strategies, consumer media investments and employer and health system prioritization.

## 1.2 Background

NN is a leading global healthcare company that focuses on obesity, diabetes and haemophilia. More than one-third of US adults live with obesity, a serious chronic disease that increases the risk of having or acquiring other medical conditions, including depression, high blood pressure, etc. NN is developing anti-obesity medicines (AOMs) that can help PwOs lose excess weight. It currently has one FDA-approved medication, Saxenda®, and has more AOMs in its pipeline.

Despite the large proportion of PwOs in the US, NN has limited promotional resources that must be directed to selected segments of patients and prescribers that are expected to be the most appropriate and promotionally responsive. NN has considerable structured and unstructured commercial data, but lacks the data science knowledge required to extract relevant promotional response insights from that data.

Therefore, our project seeks to apply machine learning and predictive modeling techniques to help our client process their data and derive insights that can guide their marketing decisions.

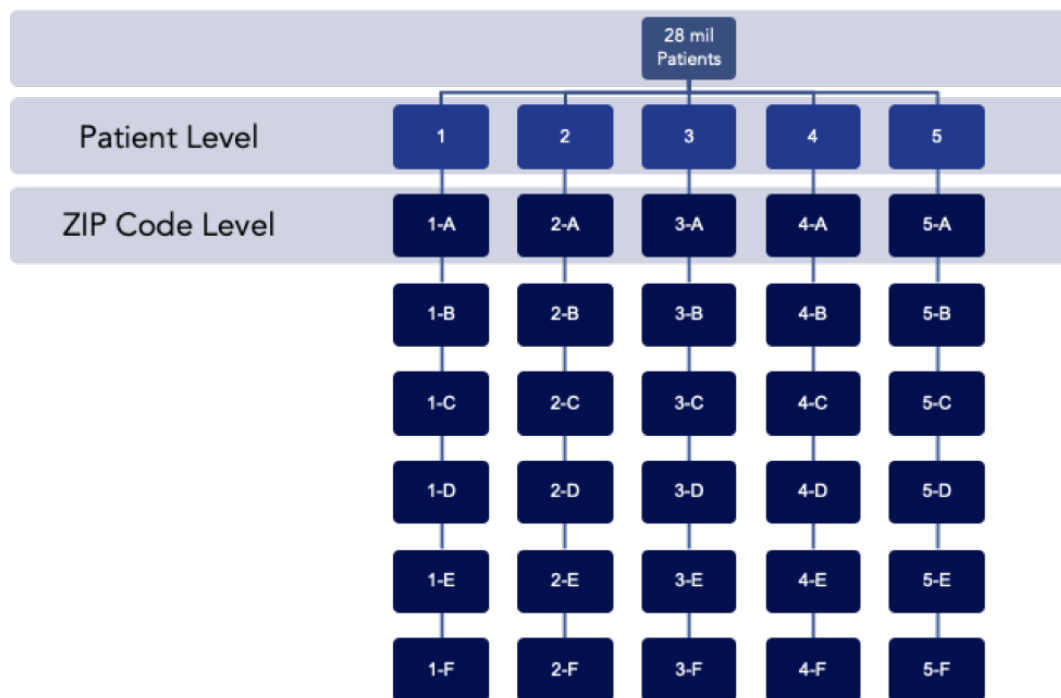
## 1.3 Methodology

Our approach to this problem is divided into 5 major steps. First, we created a patient-level relational database in SQL workbench by aggregating all the claim-level dataset from the Diagnosis (Dx), Procedure (Px), and Prescription (Rx) database and combining it with a socioeconomic dataset, created in Microsoft Excel using Policy Map datafiles from 2018. This relational database will act as the source of data for our machine learning model.

Our next step is to build the predictive model by using an unsupervised machine learning algorithm to cluster similar patients together into distinct groups, visualizing their characteristics and ranking them based on our insights and analysis of each group.

Next, we compiled a ZIP code level dataset from our patient level relational database, and ran a separate clustering algorithm to group similar ZIP codes together and conducted a similar visualization analysis to rank the different ZIP code groups.

Subsequently, we combined both of the clustering models to classify each patient into one of 30 small buckets (1-A to 5-F), as shown in Figure 1. We analyzed each of these buckets to rank the top 15 according to how receptive each group of patients are to taking Saxenda®, and designated similar buckets to higher-level categories.



**Figure 1:** *Patient Classification*

Finally, we recommended a few investment strategies and marketing campaigns to target certain patient categories over others, and by using NN's patient investment data, we calculated an estimate of the revenue earnings to justify our proposed strategies.

# Patient Relational Database

## 2.1 Objective

The purpose of creating this database is to build a data source where we can extract relevant information for our model creation and visual insights. The database consists of 4 flat tables - Patient, Prescriber, Plantrak and Socioeconomic.

## 2.2 Data Sources

We had access to the Longitudinal Access and Adjudication Data (LAAD) at the medical claim level in SQL Workbench as well as Policy Map data files in Microsoft Excel containing socioeconomic information at the Census Tract and County level.

The LAAD database is divided into 3 datasets, Diagnosis (Dx), Procedure (Px), and Prescription (Rx). The Dx dataset has information on diagnosis of obesity, overweight, and weight-related conditions, the Px dataset has details on weight consultations and bariatric surgeries, and finally the Rx dataset shows a patient's prescription fill rates and out of pocket cost. All of these information are at the medical claim level, in which the unique column identifiers are claim IDs, as shown in Figure 2.

Result 1 Messages					
claim_id	patient_id	service_date	diagnosis_code	rendering_provider_id	refe
126566431530	10357	2015-11-11	M25561		0
123012161551	10357	2015-09-19	71946		7531060
10420089730200331027	10357	2017-01-26	M25511		7858172
126342719480	10357	2015-11-20	M25561		1019765
126723616475	10357	2015-11-24	M25561		1019765
10420089732201553635	10357	2017-01-26	M25511		7858172
125169379252	10357	2015-10-07	M25561		7987272
124515750339	10357	2015-09-23	71946		7987272
120030243324	10357	2015-04-06	71946		7984779
10490172223904464575	10357	2019-08-01	O99213		7576467
10490170973506549001	10357	2019-08-19	O99213		7576467
126115198177	10357	2015-11-10	M25561		1019765
129953168836	10357	2015-10-28	M25561		1019765
125738036573	10357	2015-10-28	M25561		0
126865050845	10357	2015-12-09	M25561		0
125775789222	10357	2015-10-26	M25561		1019765
125633420970	10357	2015-10-21	M25561		1019765
10420162171504100383	10357	2019-05-31	M546		22182561
125633420971	10357	2015-10-23	M25561		1019765
10420163091702987295	10357	2019-06-07	M546		22182561
126127817219	10357	2015-11-11	M25561		1019765
126212993084	10357	2015-11-18	M25561		1019765
10420163091702987295	10357	2019-06-11	M546		22182561
125972807672	10357	2015-11-06	M25561		1019765
10420162171502153132	10357	2019-06-03	M546		22182561
125858178065	10357	2015-10-14	M25561		12023544

Figure 2: Claim Level Dataset

Socioeconomic data in the Policy Map datafiles included the percentage of age groups, race and ethnicity, education level, household spending and income at the census tract level and the unemployment rate at the county level.

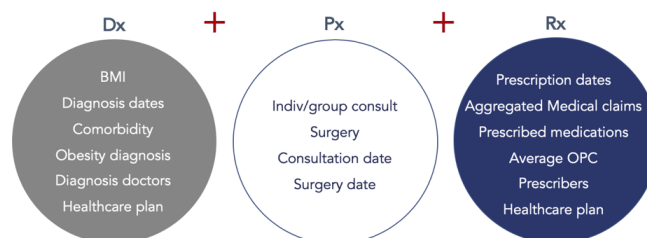
## 2.3 SQL Workbench Tables

### 2.3.1 Patient Table

Using these LAAD datasets, we transformed the data from a medical claim level into a patient level data, aggregating the features such that the patient ID is the unique column identifier. The description of significant features of the table can be found in Appendix A.

Patients in the Dx and Rx dataset are the most important patients for our machine learning model and hence we decided to combine all patients in the Dx and Rx dataset together to create the Patient flat table. Patients in the Px dataset were also included as long as they satisfy the

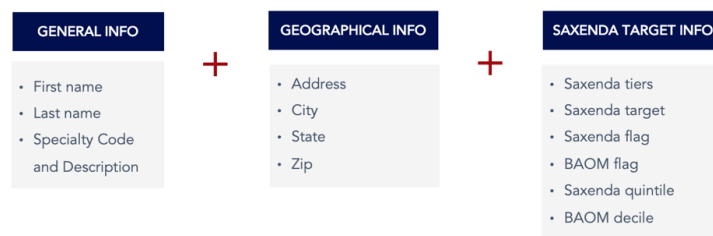
condition of also being at least in either the Dx or Rx dataset. Due to the fact that some patients in the Dx dataset are not in the Rx dataset and vice versa, many of these patients will have missing information for features that were not in their specified dataset. These missing information are filled with 0 values, and we have a specific flag feature that denotes where this patient resides in the database (Dx or Rx or both). The table contains information of more than 28.3 million patients and 105 features. Features of the patient flat table are shown in Figure 3.



**Figure 3:** *Features of Patient Flat table*

### 2.3.2 Prescriber and Plantrak Tables

Using the same methodology to create the patient table, the medical claim level in the LAAD dataset was transformed such that the Prescriber ID is the unique column identifier for the Prescriber table and the Plantrak ID for the Plantrak table. Features of the Prescriber Table and the Plantrak are shown in Figure 4 and 5 respectively. A more detailed description of these features can be found in Appendix B and Appendix C respectively.



**Figure 4:** *Features of Prescriber Flat Table*



**Figure 5:** *Features of Plantrak Flat Table*

## 2.4 Socioeconomic Flat Table

The purpose of the socioeconomic flat table is to allow us to identify any existing trends or characteristics of the different groups of patients according to where they live, and determine



whether these factors play a role in deciding whether a certain patient will take Saxenda®. At the census tract level, our data was extracted from Policy Map [1]. At the county level, we have the annual unemployment rate in the USA in 2018, sourced from the U.S. Bureau of Labor Statistics (BLS). A more detailed description of these features can be found in Appendix D.

### 2.4.1 Census Tract, County and ZIP Codes

Understanding geographic relationships is key to understanding how to properly use the data in our analysis. The level of data granularity that we are focusing on is a ZIP code level because it is the common link between the socioeconomic table and the patient and prescriber tables. Therefore, we need to establish the relationship between census tracts, counties and ZIP codes.

#### Census Tracts vs ZIP Codes

Census tracts represent the smallest territorial entity for which population data are available in the United States. Census tracts are therefore more specific geographic values, and there are approximately 50% more census tracts than ZIP codes. The relationship between census tracts and ZIP codes is hence as such: a ZIP code may contain several census tracts. Although the boundaries of ZIP codes and census tracts may not be a perfect match, the overlap of boundaries does not create a major problem in terms of relative social and economic characteristics.

#### Counties vs ZIP Codes

A county is an administrative or political subdivision of a state that consists of a geographic region with specific boundaries. Counties are hence broader geographical areas, which contain census tracts and hence ZIP codes. Since ZIP codes refer to delivery routes by the U.S. Postal Service, these routes can cross county (or even state) lines, but such occurrences are few and far between, and we do not consider it an issue.

### 2.4.2 Conversion to ZIP Codes

Noting that the information of patients in the patient flat table is on a ZIP code level, this arises the need for the socioeconomic data to be transformed to the same level of data granularity.

#### Census Tracts to ZIP Codes

We made use of a Census Tract - ZIP crosswalk dataset obtained from HUD's Office of Policy Development and Research. This dataset lists the corresponding ZIP codes for every census tract.

As mentioned, a ZIP code can consist of multiple census tracts, and in such instances, we aggregated all the data for that particular ZIP code and took the average. There were also instances where a census tract can relate to multiple ZIP codes, and our treatment of the data is the same. Figure 6 shows an illustration of this process.

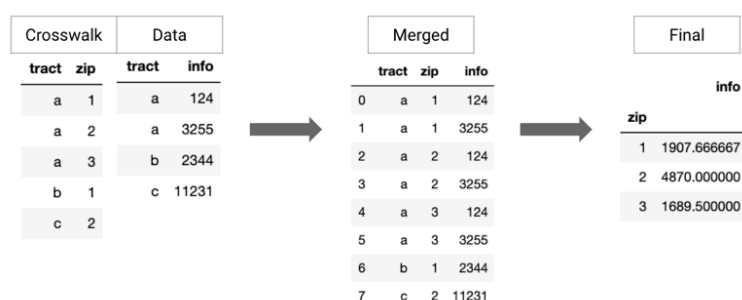


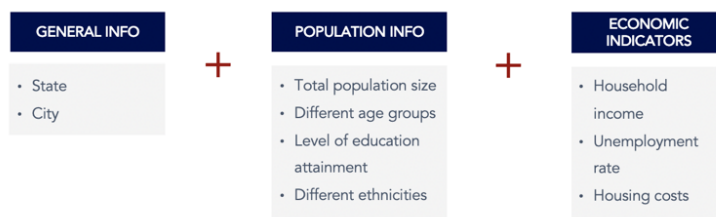
Figure 6: Illustration of mechanics of conversion

### Counties to ZIP Codes

We made use of a ZIP - County crosswalk dataset obtained from HUD's Office of Policy Development and Research. This dataset lists the corresponding counties for every ZIP code. In this case, all of the ZIP codes belonging to the same county had the same unemployment rate assigned to them.

### 2.4.3 Final Socioeconomic Table

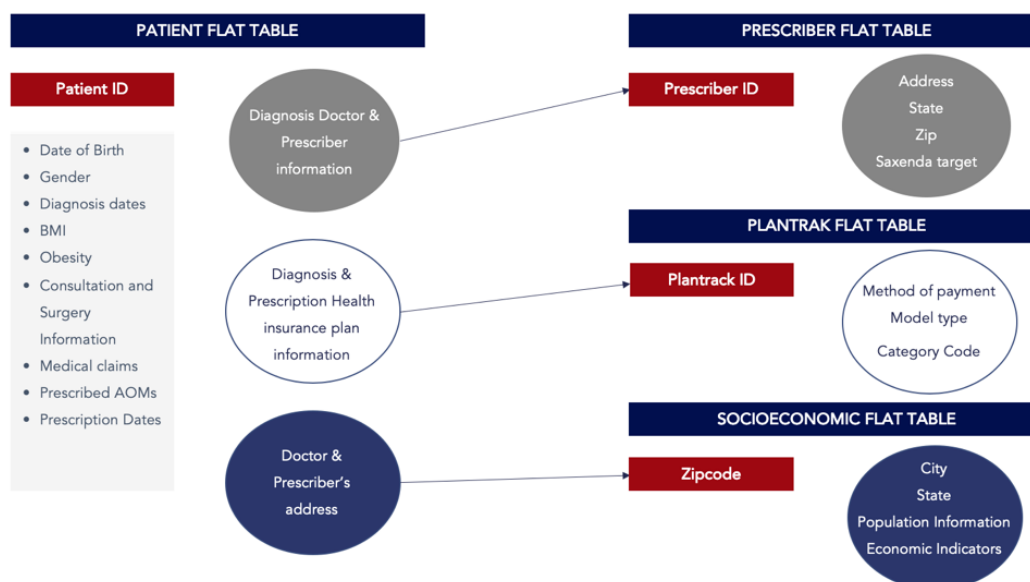
Figure 7 shows the final features of the socioeconomic flat table.



**Figure 7:** *Features of Socioeconomic Table*

### 2.5 Overall structure of database

These 4 tables are related to each other through the patient flat table. The patient flat table contains each patient's doctor/prescriber ID, which is the foreign key that links to the prescriber table, as well as the patient's insurance information denoted by the patient's plantrak ID, the foreign key that links to the plantrak table. Finally, the doctor/prescriber's address that the patient goes to has a ZIP code which is the foreign key that links to the socioeconomic table. As a result, any information in all 4 tables are readily accessible through the patient table. A visual representation of the database is shown in Figure 8.



**Figure 8:** *Patient Relational Database Schematic*

# Data Preprocessing

## 3.1 Objective

Since the size of the patient database is extremely large, not all the patients and features will be included in the clustering algorithm. Furthermore, the dataset has many missing values, categorical variables as well as high correlation between certain variables. Therefore, we had to clean and drop certain features to ensure a successful clustering.

## 3.2 Methodology

The data preprocessing methodology involves 3 significant steps:

1. Creation of a dataset in which the machine learning algorithm will run on by merging the patient table with certain features from the other 3 tables
2. Patient filtration and feature selection
3. Conversion of categorical variables to numerical variables

### 3.2.1 Creation of a single dataset

We merged the patient flat table with certain features in the prescriber, plantrak and socioeconomic table from the patient relational database to create a combined dataset that measures 28.3 million patients by 132 features. The features that were included from the Prescriber table were the prescriber's geographical information, Saxenda® tier and target information. Features extracted from the Plantrak table were the method of payment, and model type. All features from the socioeconomic table were kept.

### 3.2.2 Patient filtration and Feature selection

Some of these patients in our database have erroneous values, ranging from an age of over 2000, to missing prescriber ID values. These types of patient represent about 3.5% of our patient database, and because it is a very small portion of our dataset, we decided to drop these patients, reducing the total patient count to approximately 27.9 million. Additionally, out of the 132 features, some of them were highly correlated, while others were thought to not value-add to the clustering process. Some of these variables include addresses of the prescribers, number of days between diagnosis and prescription, total out of pocket cost, first and latest doctor and prescriber information. As a result, we decided to drop these features. The final number of features kept is 59.

### 3.2.3 Converting categorical variables to numerical

Since machine learning algorithms are only able to take in numerical variables, any other variables have to be converted from their original form to a numerical form that would convey the same meaning as when it is in its categorical format. The way we decided to process this is by using one hot encoding and label encoding techniques. These encoding techniques encode such variables to certain numbers so that the algorithm recognizes these variables as numbers that signify categorical values. One such example is converting a feature that has Yes, No values into 1, 0 values, with 1 representing 'Yes', and 0 representing 'No', as shown in Figure 9. An important feature that was label encoded is the Saxenda® Tier features, which has values Tier 1, Tier 2, Tier 3, No Tier, and Non-Tiers. In this special case, the values are arranged in an

order where Tier 1 is ranked highest and Non-Tiers are ranked lowest. Hence, the label encoding converts these values into 10, 8, 6, 2, 0 to signify the same ranking hierarchy in a numerical format.



**Figure 9:** *Label encoding illustration*

The final cleaned dataset consists of all numerical variables, and consisted of 28.3 million patients with 59 features. This patient level dataset will be our primary dataset in both our patient level and ZIP code level clustering analysis.

# Patient Level Clustering


## 4.1 Objective

The patient level clustering method aims to divide the entire patient level dataset into smaller datasets, and group similar patients together into a certain number of clusters, so that we can extract the properties of each group of patients and compare their differences across the clusters. Before running the clustering algorithm, it is important to further inspect and clean the data by scaling the numerical variables and reduce the correlation between features.

## 4.2 Scaling and Dimensionality Reduction

### 4.2.1 Standard Scaler

The purpose of scaling the data is to ensure that the clustering algorithm does not put a larger bias to a certain feature over others simply because the magnitude of that feature is significantly larger than that of other features. For example, a feature named ‘age’ will have values in the range of 40 to 50. However, another numerical feature such as ‘number of claims’ could potentially exceed 200. The machine learning algorithm is not capable of distinguishing between these 2 features and will cluster the patients based on the much larger numerical value, because the smaller value seems insignificant compared to the larger one. To prevent this kind of biasness from occurring, scaling the dataset is important in order to ensure the magnitude of all features are transformed in a way that allows the algorithm to recognize all features as equally important attributes as best as possible, as shown in Figure 10. The Standard Scaler method from the sklearn library in Python was used to transform the continuous variables in the dataset to have a mean value of 0 and a standard deviation of 1.



Current age	Median HH income
42	36976.23
36	68221.00
50	104771.13
47	107304.59

➡➡➡

Current age	Median HH income
-0.6111	-1.3773
-1.1667	-0.2384
0.1296	1.0940
-0.1482	1.1864

Figure 10: *Scaling data process illustration*

### 4.2.2 Principal Component Analysis

Principal Component Analysis (PCA) is a technique used to reduce the dimensionality of the dataset while maintaining as much useful information about the data as possible [2]. Machine learning algorithms become less accurate if the data is too large, and if the features are highly correlated. PCA was used to reduce the dimensionality of the dataset by aggregating correlated variables together and combining them to form a single independent variable, and in the process, reduces the number of features. To ensure that most of the information held in the dataset is kept, we set a 95% explained variance threshold that the PCA algorithm should maintain. The result is a reduction from 59 features to 34 features that are relatively independent from each other.

## 4.3 Clustering Algorithm

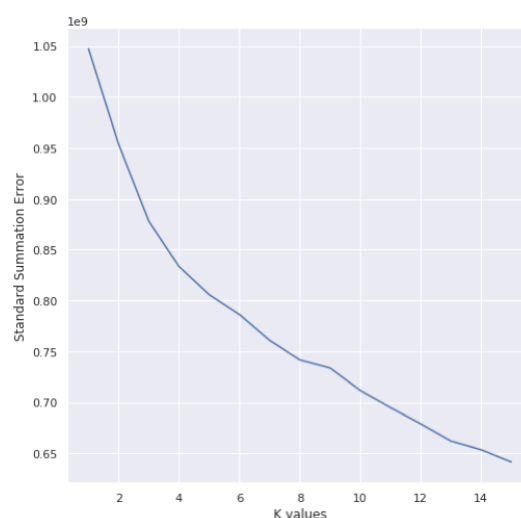
### 4.3.1 Testing different models

There are numerous unsupervised machine learning methods that cluster data into specific groups. After a thorough research on the conditions of implementing each method, we decided to test from 3 algorithms that would represent our dataset the best out of all the available algorithms. These are K-Means, K-Prototype and Gaussian Mixture Model.

#### K-Means

K-Means clustering is the most well-known clustering algorithm and is the easiest to implement. It works by selecting a number of groups to use and randomly initializing their respective center points. A more detailed description of the algorithm can be found in Appendix E.

In order to test whether K-Means would be a good clustering method for our dataset, we plotted an elbow plot (shown in Figure 11) to find the best point that optimizes the bias-variance tradeoff.



**Figure 11:** *Patient Clustering Elbow Plot*

The elbow plot is not a good representation of clustering of our dataset because there is no obvious point in which the bias-variance tradeoff is optimal. This shows that K-Means is not a suitable clustering algorithm for our dataset.

#### K-Prototypes

K-Prototypes is a modified form of K-Means, meant to handle both numerical values and categorical values. Similar to K-Means, in this algorithm, k ‘prototypes’ are chosen at the start of the algorithm and the points are clustered by measuring the distance of the points from these prototypes. A more detailed explanation of the algorithm can be found in Appendix F.

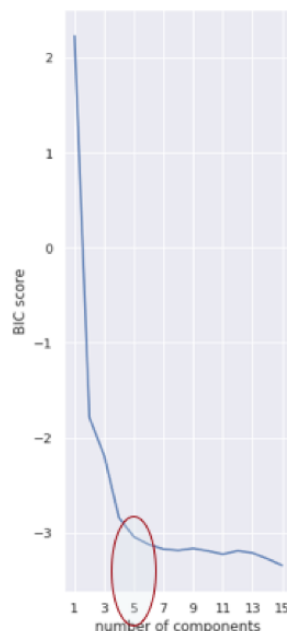
We tested this algorithm on a 1.3 million patient subset of the data, but unfortunately, it was not able to produce results even after 5 hours of run time. Since the run time of the algorithm increases exponentially with the number of data points, we decided not to implement this model because it is extremely inefficient.

#### Gaussian Mixture Models

Gaussian Mixture Model (GMM) is a soft clustering method that measures the probability of a datapoint belonging to a specific cluster and chooses the cluster with the highest probability. A more detailed description of the GMM can be found in Appendix G. We think that it is a

more appropriate clustering method because the scaling method we used distributes the data in a gaussian fashion.

Similar to the K-Means elbow plot, we plotted the GMM equivalent, the Bayesian Information Criterion (BIC) score to determine the point that optimizes the bias-variance trade-off. Figure 12 shows the plot of BIC score against the number of clusters for our dataset.



**Figure 12:** *BIC score plot*

From the plot, there is a distinct point in which the reduction of the BIC score becomes insignificant with an increase of 1 cluster. Through this plot, we decided to implement the GMM algorithm using  $n = 5$  as our optimal number of clusters.

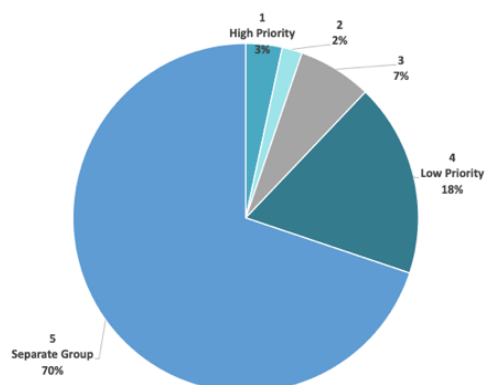
#### 4.3.2 Implementing the GMM algorithm

Before implementing the GMM on the dataset, we first split our patient-level dataset into training and testing data using a 80-20 split. The split was randomized to ensure minimal bias as possible. With the training dataset, we fitted the GMM algorithm with 5 as our chosen number of clusters. The model is saved in a serialized format using the joblib module specific to Python so that it can be loaded anytime in the future for new datasets. The model is then used to predict each data point in the testing dataset, and cluster those points according to how it was trained with the training dataset. The clusters are initially numbered from 0 to 4, but for conventional naming purposes, we decided to number them from 1 to 5. We conducted preliminary visualization tests on these clusters of patients, and decided to rank each cluster based on how receptive the patients are to taking Saxenda®. Hence, we re-arranged the numbering of the clusters such that Cluster 1 represents highest priority patients, and Cluster 4 represents the lowest priority, and Cluster 5 represents a separate group of patients. The patient-cluster dictionary is then serialized using the joblib module for future references.

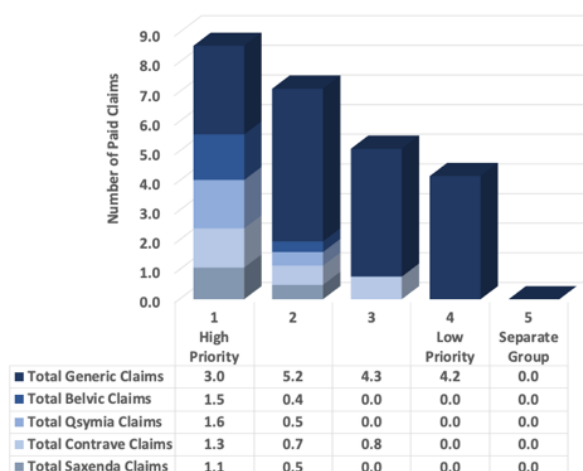
#### 4.4 Data visualization

To identify the characteristics of patients in each cluster, we decided to group the patients by their cluster number, and calculate the average value of each feature in the dataset. We visualized certain features and compared them across different clusters to generate insights

about each group. Figure 13 shows the number of patients in each cluster and Figure 14 shows the distribution of paid claims across the various AOMs in each group.



**Figure 13:** *Patient Distribution*



**Figure 14:** *Distribution of paid claims across AOMs*

Based on the figures above, cluster 5 represents most of our patient population but it has 0 paid claims in any medication. This is because cluster 5 patients only have diagnosis information in the Dx dataset and no prescription information in the Rx dataset. Some of the prescription information may not have been added to the original database by IQVIA, hence resulting in 0 paid claims, but this does not reflect the actual reality of the patient's past medication history. According to NN's market share statistic, we assume that at least 6% of these patients have some combination of paid claims and have been targeted by NN to take Saxenda® before. Additional visualization analysis can be seen in Appendix H.

## 4.5 Patient Cluster Characteristics

Based on our visualizations and comparisons, we derived the following characteristics of each cluster.

### 4.5.1 Cluster 1 - High Priority Patients

- Largest number of paid claims and stay-times across all AOMs and the most Saxenda® prescriptions
- Highest percentage of Tier 1 and Tier 2 Saxenda® prescribers, and lowest percentage of Tier 3 Saxenda® prescribers
- Highest median household income

### 4.5.2 Cluster 2 - Medium High Priority Patients

- Second largest Saxenda® paid prescriptions, paid claims and stay-times across branded AOMs
- Second highest percentage of Tier 1 and Tier 2 Saxenda® prescribers
- Second highest percentage of patients diagnosed with obesity and comorbidities
- Only group of patients with consultations, surgeries and screenings
- Second highest median household income



#### **4.5.3 Cluster 3 - Medium Low Priority Patients**

- Prescribed mostly generic AOMs, and very few branded AOMs
- Only cluster with rejected Saxenda® prescriptions
- Third highest percentage of Tier 1 and Tier 2 Saxenda® prescribers
- Third highest median household income

#### **4.5.4 Cluster 4 - Low Priority Patients**

- Only prescribed Generic AOMs
- Highest percentage of Tier 3 Saxenda® prescribers, and second lowest Tier 1 and Tier 2 Saxenda® prescribers
- Lowest percentage of patients diagnosed with obesity and comorbidities
- Highest Pharmacy Benefit Manager (PBM) Commercial Insurance Model
- Lowest median household income

#### **4.5.5 Cluster 5 - Separate Group**

- Patients who exist only in the Diagnosis dataset (Dx) with no prescription information
- Highest percentage of Non-Targeted Saxenda® prescribers
- Highest percentage of patients diagnosed with obesity and comorbidities
- Highest percentage of male patients

# ZIP Code Level Clustering

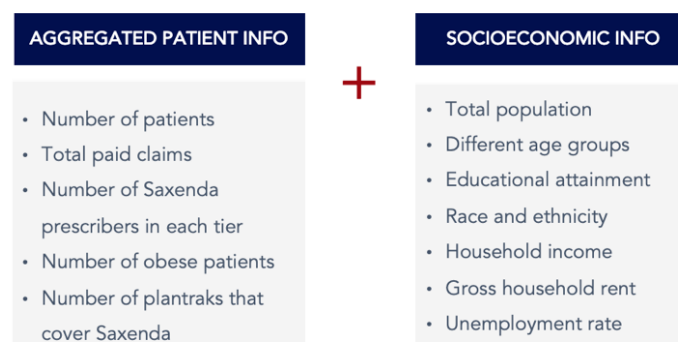
## 5.1 Objective

The ZIP code clustering analysis is a complementary analysis to the patient level clustering. The ZIP code clustering aims to divide patients again into distinct groups which contain similar ZIP codes based on certain features such as socioeconomic data, number of paid claims across AOMs, etc. A more detailed analysis could be achieved when we combine both ZIP code and patient level clustering results.

## 5.2 ZIP Code Flat Table

Before executing the clustering algorithm, a ZIP code flat table was created from the patient level dataset that was already preprocessed and cleaned. We aggregated the patient level features of the patient dataset in a way such that the unique column identifier becomes the ZIP code as associated by each patient. The socioeconomic information was left untouched because they are already at the ZIP code level.

Figure 15 shows some of the features in the ZIP code flat table. The dataset measures 19.5K ZIP codes by 30 features.



**Figure 15:** *Features of ZIP code flat table*

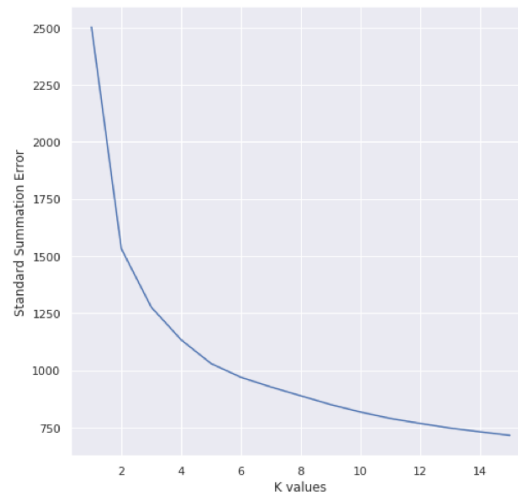
Using this ZIP code level dataset, we created a normalized version by dividing the aggregated patient info with the population of each ZIP code, in order to reduce the biases of having different population values for each ZIP code. There were some ZIP codes in which the population was 0 as well as ZIP codes in which the number of unique patients exceeded the population. This amounts to 1055 ZIP codes, and these values were taken out of the dataset and put in a separate group, Group F. The cleaned normalized dataset, which measured 18.4K ZIP codes by 30 features was then scaled using Min-Max Scaler before running it through our clustering algorithm. Min-Max Scaler transformed the data such that the new values were between 0 and 1 and we used this scaling method because Standard Scaler did not work well for our dataset, and Min-Max Scaler was considered the next best option.

## 5.3 Clustering Algorithm

Similar to the patient level clustering, we tested both the K-Means and the GMM algorithm on the dataset to see which would be a better representation of our model. K-prototype was not tested because the dataset had no categorical variables.

### **K-Means**

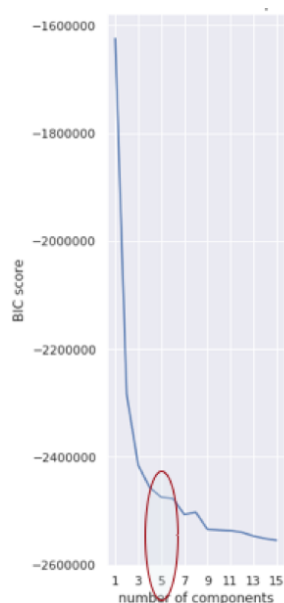
Figure 16 shows an elbow plot graph using a range of initial clusters from 1 to 15, and looking at the graph, it is shown that like the patient level clustering method, there is no distinguishable point which the bias-variance tradeoff is minimized. Hence, this shows that the K-Means is not a reliable algorithm for our ZIP code clustering analysis.



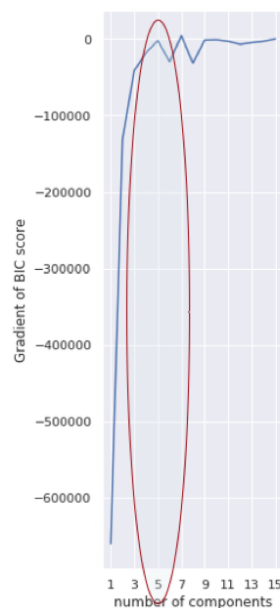
**Figure 16:** *Zip code elbow plot*

### **Gaussian Mixture Model**

Figure 17 shows a BIC score plot against the number of components. This graph is a much better representation than the elbow plot. However, it is still a bit difficult to figure out which point minimizes the bias-variance trade-off. We have narrowed it down to a choice between 5 or 6 distinct clusters. To drill down on our analysis, we plotted the slope of the BIC score against the number of components to visualize the optimal number of clusters in which the bias-variance trade off is minimized. Figure 18 shows this plot, and it is generally known that the point at which the gradient is closest to 0 is the optimal point. The x-coordinate of the point shows that 5 is the optimal number of components we should use for our clustering analysis.



**Figure 17:** *BIC score plot*

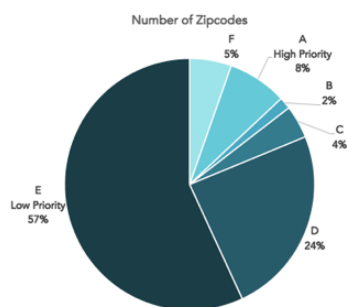


**Figure 18:** *Gradient of BIC score plot*

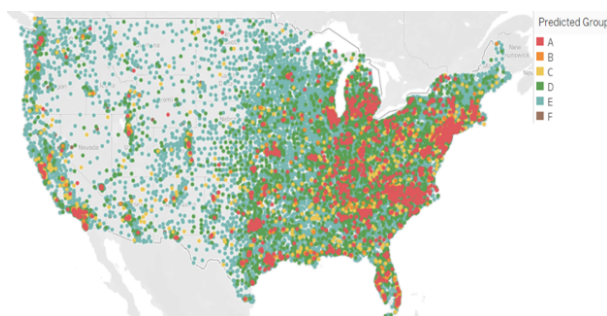
With this justification, we ran our GMM clustering with  $n = 5$  clusters on the entire cleaned ZIP code level dataset. This time, we did not split the data into a training-testing split, because there are only 18.4K ZIP codes, which is significantly smaller than the patient level dataset. Hence, we decided to use the entire set in our algorithm in order to train the model with as many data points as possible. Like the patient clustering algorithm, the groups that the model outputted were numbered 0 to 4. To differentiate these groups from the patient level, we decided to name them groups 'A' to 'E' instead. Using a similar methodology to the patient level clusters, we did a preliminary visualization analysis on these ZIP code groups, and decided to rearrange the names of the groups such that group 'A' represents the ZIP codes with the highest priority for NN to target to market Saxenda®, and 'E' represents the lowest priority, with 'F' being a separate group containing the 1055 ZIP codes that were not included in our algorithm but still very relevant in our analysis. The ZIP code - group dictionary is serialized using the joblib module for future references.

## 5.4 Data Visualization

To identify the characteristics of ZIP codes in each group, similar to the patient level clustering, we calculated the average value of each feature in the normalized dataset and visualized the data. Furthermore, at the ZIP code level, we conducted an additional heatmap analysis to show the distribution of ZIP codes at a geographical platform using Tableau. Figure 19 shows the number of ZIP codes in each group and Figure 20 shows the geographical distribution across the entire map of the United States.

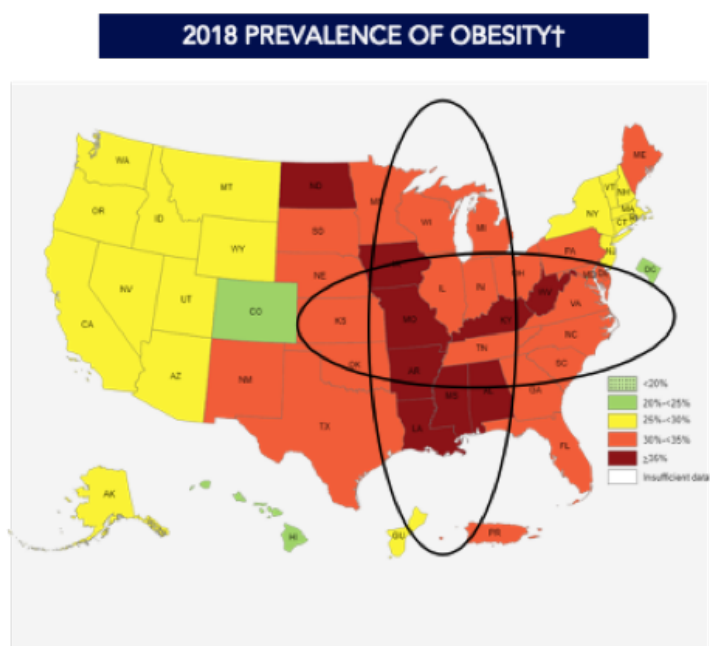


**Figure 19:**  
*ZIP code  
distribution*



**Figure 20:** *Geographical distribution*

From Figure 19, most of the ZIP codes in the dataset are categorized as low priority. From Figure 20, it looks like the concentration of high priority ZIP codes is more prevalent in the eastern part of the United States than the western parts. This matches the 2018 distribution of people with obesity in the United States, shown in Figure 21.



**Figure 21:** *2018 Obesity prevalence [3]*

Additional visualization analysis can be found in Appendix I.

## 5.5 ZIP Code Group Characteristics

Based on our visualizations and comparisons, we derived the following characteristics of each group.

### 5.5.1 Group A: High Priority ZIP Codes

- Highest ratio of paid Saxenda® claims to other AOM claims
- Highest percentage of Tier 1 and Tier 2 Saxenda® prescribers
- Highest percentage of people with high education and median household income

### **5.5.2 Group B: Medium High Priority ZIP Codes**

- Second highest number of paid Saxenda® claims
- Second highest percentage of Tier 1 and Tier 2 Saxenda® prescribers
- Second highest percentage of patients diagnosed with obesity and highest percentage for comorbidities
- Second highest median household income

### **5.5.3 Group C: Medium Priority ZIP Codes**

- Third highest overall number of paid claims with very little Saxenda® prescriptions
- Only Tier 3 Saxenda® prescribers
- Highest percentage of patients diagnosed with comorbidities and second highest percentage for obesity
- Highest unemployment rate

### **5.5.4 Group D: Medium Low Priority ZIP Codes**

- Second lowest overall number of paid claims
- Second lowest percentage of patients diagnosed with obesity and comorbidities
- Third lowest unemployment rate
- Lowest percentage of people with high education

### **5.5.5 Group E: Low Priority ZIP Codes**

- Lowest overall number of paid claims
- Lowest percentage of all types of Saxenda® prescribers
- Lowest percentage of patients diagnosed with obesity and comorbidities
- Second lowest median household income

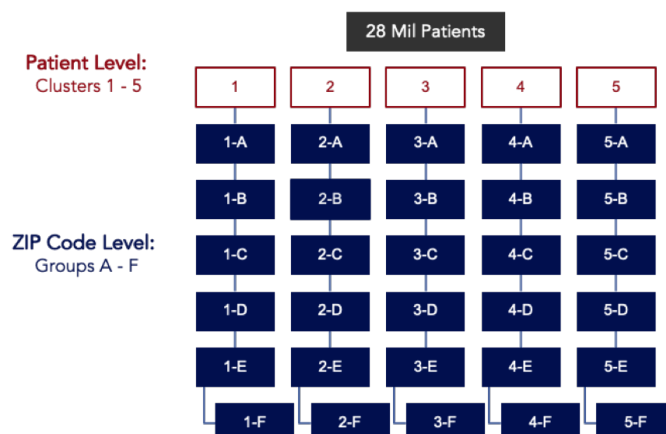
### **5.5.6 Group F: Separate Group of ZIP Codes**

- Number of patients in each ZIP Code exceed the population of that ZIP Code
- Second lowest unemployment rate
- Second highest percentage of people with high education
- Lowest median household income

# Overall Patient Classification and Ranking

## 6.1 Objective

After conducting both independent patient level and ZIP code level clustering analysis, we decided to combine both results to create a hierarchy that classifies the entire 27.9 million patients into 30 sub-clusters, labelled (1-A to 5-F), as shown in Figure 22.



**Figure 22:** *Detailed Patient Classification*

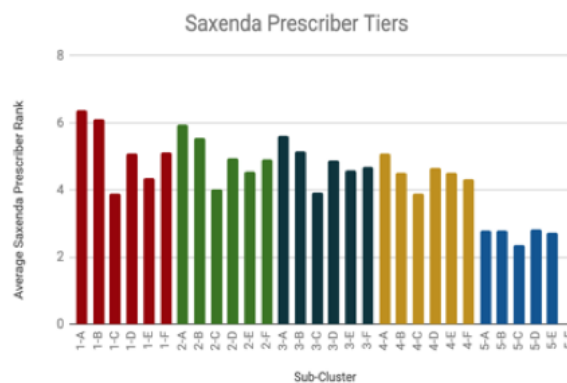
The purpose of this hierarchy is to do a deep-dive analysis on each of these sub-clusters of patients, which is a much more detailed analysis compared to the patient level clustering or the ZIP code level clustering by itself. The insights generated from the analysis would be used to rank the top 15 sub-clusters and further group similar ranks together to higher-level categories that would be easier to remember for the Obesity Leadership Team.

## 6.2 Data Visualization

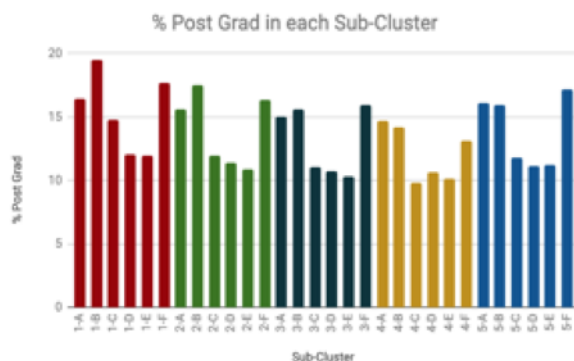
By assigning each of the 27.9 million patients to these sub-clusters, we did a group by analysis on these sub-clusters and calculated the average values of the features in that dataset. Figures 23 to 26 are some examples of the visualizations done. More visualizations can be found in Appendix J.



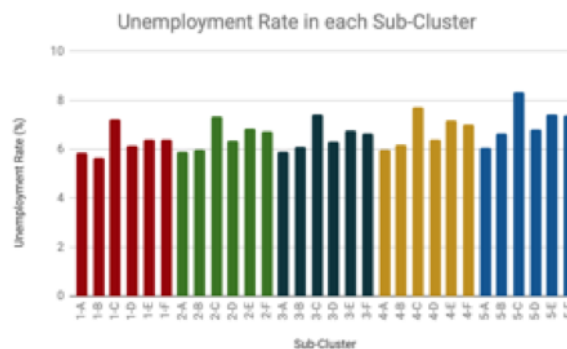
**Figure 23:** *Average Paid Saxenda® Claims*



**Figure 24:** *Average rank of Prescribers*



**Figure 25:** % of Post Graduates



**Figure 26:** Unemployment Rate

Based on these visualizations, we decided to target certain patient clusters and ZIP code groups over others. Patient clusters 1 and 2 are our high priority patients because they are the only groups that have a past history of taking Saxenda®. However, we also plan to target cluster 5 because they represent the largest pool of untapped patients in our database. Cluster 3 is another low priority target because they have been prescribed Saxenda®, but have no record of paid claims. Additionally, ZIP codes groups A and B are also our high priority targets because they show a lot more Saxenda® claims and have very strong socioeconomic status. Groups D and F are our other targets because they are our second highest ZIP code groups with more paid Saxenda® claims than groups C and E, and also have above average socioeconomic status.

### 6.3 Ranking Process

With this analysis, we decided to rank the top 15 sub-clusters in the order of how receptive they are to taking Saxenda®, and classify close ranking groups into higher level categories: NNI AOM Enthusiasts, Convertibles, Potentials, and Rejects. The rest of the sub-clusters ranked 16 to 30 are unranked and classified into a single category labelled NNI AOM Hopeless. An illustration of this ranking process is shown in Figure 27.

RANK	SUB-CLUSTER	CATEGORY
1	1-A	NNI AOM Enthusiasts
2	1-F	
3	1-B	
4	1-D	
5	2-A	NNI AOM Convertibles
6	2-F	
7	2-B	
8	2-D	
9	5-A	NNI AOM Potentials
10	5-D	
11	5-B	
12	3-A	NNI AOM Rejects
13	3-B	
14	3-D	
15	3-F	

**Figure 27:** Patient Ranking Model



## 6.4 Categories Characteristics

Using our visualizations and analysis, we identified several distinct characteristics of each NNI AOM category.

### 6.4.1 NNI AOM Enthusiasts

- Highest paid Saxenda® to Generic AOMs claims ratio
- Highest number of High-tier Saxenda® prescribes and stay-times
- Strong socioeconomic status in the form of low unemployment rate, high household income, and high percentage of college graduates and post-doctorates
- Strong availability of health insurance that covers Saxenda® but not other branded AOMs

### 6.4.2 NNI AOM Convertibles

- Lower ratio of Saxenda® to Generic AOMs claims than Saxenda® Enthusiasts
- Second highest number of High-tier Saxenda® prescribers and stay-times
- Strong socioeconomic status, but prefers Generic and other branded AOMs
- High out of pocket costs for Saxenda® shows health insurance largely covers other branded AOMs more than Saxenda®
- Higher percentage of obese and comorbidity patients than Saxenda® Enthusiasts

### 6.4.3 NNI AOM Potentials

- Largest patient population that is untapped
- Very high percentage of patients diagnosed with obesity and comorbidity
- Socioeconomic status that is comparable to Saxenda® Enthusiasts group shows patients have the purchasing power to take Saxenda®
- Modest amounts of High-tier Saxenda® prescribers show that prescribers are willing to prescribe Saxenda®

### 6.4.4 NNI AOM Rejects

- Has paid claims only for Generic AOMs and Contrave
- Patients are being prescribed Saxenda® but all the claims are rejected
- Above average socioeconomic status, slightly lower than Saxenda® Enthusiasts
- Above average amounts of High-tier Saxenda® prescribers

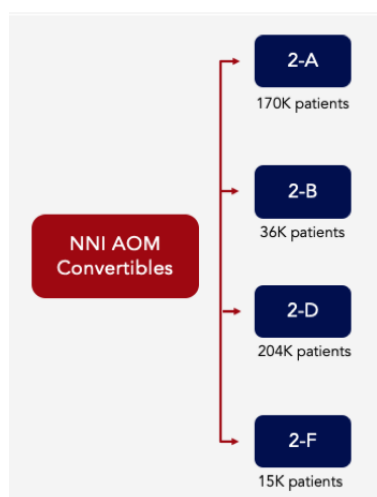
# Recommended Strategies

## 7.1 Objective

The final step of our project is to come up with some recommended strategies for NN to implement in order to use their limited promotional resources to efficiently target specific groups of patients that are most likely to take Saxenda®. Through our analysis of each Saxenda® category, we recommended NN to focus on 2 groups: NNI AOM Convertibles and NNI AOM Potentials because we believed the company can reap the most benefits by focusing their efforts primarily on these two groups of patients. For each category, we came up with different strategies that catered to the characteristics of each group.

## 7.2 Marketing Campaigns

The marketing campaigns should focus on depth, and target NNI AOM Convertibles. Figure 28 shows the distribution of the NNI AOM Convertibles category.



**Figure 28:** *NNI AOM Convertibles Distribution*

NNI AOM Convertibles represent only 15% of the entire 27.9 million patient database, but they are currently on the edge between being loyal Saxenda® patients and being loyal other branded AOM patients. Because they show a history of taking Saxenda®, they represent high priority patients that can be converted to loyal customers.

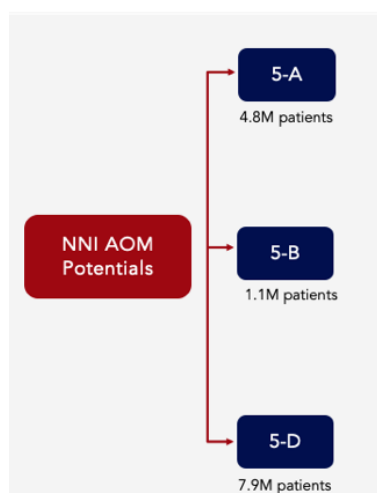
The issue that NN is facing with regards to these patients is that they face direct competition with Contrave because the average number of paid claims for Saxenda® and Contrave are relatively equal, but Saxenda® claims outnumber Qsymia and Belviq claims. Despite this, the average out of pocket cost for Saxenda® is significantly higher than the other branded AOMs. This is a stark contrast to NNI AOM Enthusiasts whose out of pocket cost for Saxenda® is so much lower than that of other branded AOMs, and this out of pocket cost is comparable to the cost for Generic claims. Since both Enthusiasts, and Convertibles have strong socioeconomic status, it goes to show that Convertibles are unwilling to pay for Saxenda® even though they have the purchasing power to do so.

One reason why we think this is the case for Convertibles is because there is a lack of health insurances or lack of knowledge of health insurances that cover Saxenda®. These patients have health insurances that cover other branded AOMs, hence the lower out of pocket costs. Therefore, we recommend that the marketing campaigns for Convertibles should first focus

on highlighting the differences between Saxenda® and Contrave, and promote Saxenda®'s qualities. Secondly, the marketing campaigns should also promote the availability of health insurances that cover Saxenda® in an attempt to lower the average out of pocket cost for these patients. These campaigns will focus more on deepening the relationship of patients with NN, and thus will have a more qualitative than quantitative nature.

### 7.3 Investment Efforts

The investment efforts should focus on breadth, and target NNI AOM Potentials. Figure 29 shows the distribution of the NNI AOM Potentials category.



**Figure 29:** *Saxenda® Potentials Distribution*

NNI AOM Potentials represent about 50% of the entire 27.9 million patient dataset, and they currently have no information on any kind of prescriptions. This could be due to missing information from IQVIA's side, so we will assume that NN has targeted 6% of these patients, according to Saxenda®'s most recent market share.

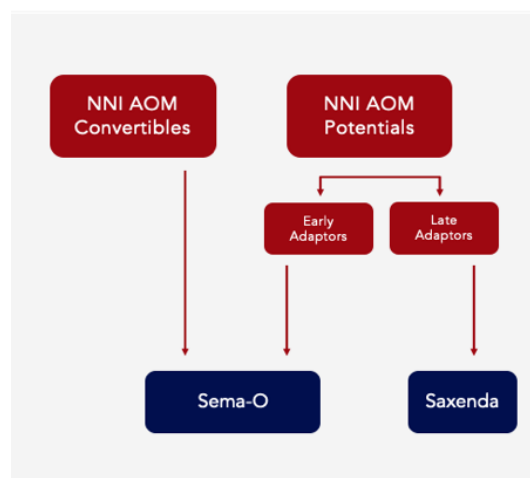
These patients have very high cases of obesity and weight-related comorbidities, which are diseases that Saxenda® can target effectively and efficiently. However, these patients do not seem to have much knowledge of Saxenda® and the drastic effects of obesity and weight-related comorbidities. As a result, the number of paid claims for any type of AOM is extremely low. Despite this, the number of high-tier Saxenda® prescribers are relatively higher than expected, which goes to show that there are prescribers in the area who are willing to prescribe Saxenda® to their patients.

Because of the large patient pool, NN will benefit by focusing their efforts on breadth rather than depth in order to capture as many patients as possible to be part of their customer base. NN can also promote Obesity educational programs to spread the understanding of the severity of obesity among these patients to attract them to taking Saxenda®. Since these patients are not yet taking any other branded AOMs, this is a prime opportunity for NN to target these patients with Saxenda® before they try out other branded AOMs. Under the assumptions of Saxenda®'s lifetime value of \$6,300, and a captured patient's stay-time of at least 6 months, if NN manages to capture just 20% (an increase from 6%) of these patients, the potential revenue increase will be at least \$12 Billion.

### 7.4 Future portfolio: Sema-O

Sema-O is a future product that NN is still developing in its pipeline. It is a GLP-1 medication that treats patients diagnosed with type 2 diabetes and helps them achieve substantial weight

loss and a lower risk of hypoglycemia. The product is still undergoing phase 3 trials and about to enter phase 4. Based on our analysis of the NNI AOM categories, we believe that Sema-O, like Saxenda®, can be targeted to Convertibles and Potentials, as shown in Figure 30.



**Figure 30:** *Sema-O Targeting Strategy*

Patients in the Convertibles category have a lower Saxenda® stay-time than Enthusiasts, so it is possible that these patients either do not like the product or they believe it is not effective. Therefore, instead of losing these patients completely to other branded AOMs like Contrave, they can be re-targeted with Sema-O when it becomes FDA-approved and enters markets. This is because some of these patients might be willing to try a new product.

Furthermore, in the Potentials category, a future enhancement of our classification and ranking model is to further divide this category into early and late adapters. Early adapters are patients who are eager and willing to try something new, and hence are potential targets for Sema-O. Late adapters can still be targeted with Saxenda® because they are patients who prefer traditional and more reliable medication that has seen previous success in the past, and are less willing to try something new until significant results have been released.

## Conclusion

Using two completely independent datasets (LAAD and Policy Map Data Files), we came up with a methodology to classify the entire 28 million patients from Q2 2015 to Q4 2019 in the NN database. In this report, we highlighted the major steps of data processing, execution of the clustering algorithm in patient and ZIP code level, and the insights we extracted from the data visualizations. Using this methodology, we came up with a ranking model that prioritizes certain patients over others, and we recommended NN to target two specific patient categories with the aforementioned strategies because we believe that by focusing their limited promotional resources on NNI AOM Convertibles and NNI AOM Potentials, NN would yield the best results in terms of attracting a larger patient base and maximizing profits for its current and future products in its pipeline.

## References

- [1] “Polycymap,” Apr. 2020. Available from: <https://www.policymap.com/> [Accessed 10 May 2020].
- [2] Z. Jaadi, “A step by step explanation of principal component analysis,” Sept. 2019. Available from: <https://builtin.com/data-science/step-step-explanation-principal-component-analysis> [Accessed 10 May 2020].
- [3] “Adult obesity prevalence maps,” Oct. 2019. Available from: <https://www.cdc.gov/obesity/data/prevalence-maps.html> [Accessed 10 May 2020].
- [4] G. Seif, “The 5 clustering algorithms data scientists need to know,” Sept. 2019. Available from: <https://towardsdatascience.com/the-5-clustering-algorithms-data-scientists-need-to-know-a36d136ef68> [Accessed 10 May 2020].
- [5] “K-means elbow method.” Available from: <https://pythonprogramminglanguage.com/kmeans-elbow-method/> [Accessed 10 May 2020].
- [6] Z. Huang, “Clustering large data sets with mixed numeric and categorical values,” 1997.

# Appendix: Features of patient flat table

## A.1 General Information

- patient\_birth\_year
  - Patient's year of birth
- patient\_gender
  - Patient's gender, has values Male, Female, Undefined

## A.2 Diagnosis (Dx) information

- age\_during\_first\_diagnosis
  - Patient's earliest age when he/she goes to a diagnosis doctor from the time period Q2 2015 to Q4 2019
- age\_during\_latest\_diagnosis
  - Patient's age when he/she last visited a diagnosis doctor from the time period Q2 2015 to Q4 2019
- bmi\_latest
  - Patient's body mass index when he/she last visited a diagnosis doctor from the time period Q2 2015 to Q4 2019
- common\_wt\_cm\_dx\_yn
  - Whether the patient is diagnosed any common weight-related comorbidities, has values Yes, No
- overweight\_dx\_yn
  - Whether the patient is diagnosed overweight, has values Yes, No
- any\_wt\_cm\_dx\_yn
  - Whether the patient is diagnosed with any type (common or uncommon) weight-related comorbidities, has values Yes, No
- obesity\_dx\_yn
  - Whether the patient is diagnosed with obesity, has values Yes, No
- baom\_label\_adult\_yn
  - Whether the patient is labelled as an adult, has values Yes, No
- baom\_label\_adolescent\_yn
  - Whether the patient is labelled as an adolescent, has values Yes, No
- overweight\_and\_wt\_cm\_dx\_yn

- Whether the patient is diagnosed overweight and has any weight-related comorbidities, has values Yes, No
- obesity\_or\_ow\_and\_cm\_yn
  - Whether the patient is diagnosed either overweight/obese and has any weight-related comorbidities, has values Yes, No
- dx\_most\_freq\_prescriber\_id
  - Identification number of the most frequent diagnosis doctor the patient goes to
- dx\_most\_freq\_state
  - The US State that the most frequent diagnosis doctor the patient goes to lives in
- dx\_most\_freq\_zip
  - The zip code of the most frequent diagnosis doctor the patient goes to
- dx\_most\_freq\_plantrak\_id
  - The insurance identification number of the most frequent (Dx) healthcare insurance plan the patient has signed for
- first\_diagnosis\_date
  - The earliest date the patient visits the diagnosis doctor from the time period of Q2 2015 to Q2 2019
- dx\_first\_prescriber\_id
  - Identification number of the first diagnosis doctor the patient goes to from the time period of Q2 2015 to Q2 2019
- dx\_first\_state
  - The US State that the first diagnosis doctor the patient goes to lives in from the time period of Q2 2015 to Q2 2019
- dx\_first\_zip
  - The zip code of the first diagnosis doctor the patient goes to from the time period of Q2 2015 to Q2 2019
- dx\_first\_plantrak\_id
  - The insurance identification number of the first (Dx) healthcare insurance plan the patient has signed for from the time period of Q2 2015 to Q2 2019
- latest\_diagnosis\_date
  - The latest date the patient visits the diagnosis doctor from the time period of Q2 2015 to Q2 2019
- dx\_latest\_prescriber\_id
  - Identification number of the latest diagnosis doctor the patient goes to from the time period of Q2 2015 to Q2 2019



- dx\_latest\_state
  - The US State that the latest diagnosis doctor the patient goes to lives in from the time period of Q2 2015 to Q2 2019
- dx\_latest\_zip
  - The zip code of the latest diagnosis doctor the patient goes to from the time period of Q2 2015 to Q2 2019
- dx\_latest\_plantrak\_id
  - The insurance identification number of the latest (Dx) healthcare insurance plan the patient has signed for from the time period of Q2 2015 to Q2 2019

### A.3 Procedure (Px) information

- group\_consult\_yn
  - Whether the patient undergoes any group consultations, has values Yes, No
- count\_group\_consult
  - Number of times the patient undergoes group consultation
- individual\_consult\_yn
  - Whether the patient undergoes any individual consultation, has values Yes, No
- count\_individual\_consult
  - Number of times the patient undergoes individual consultation
- screening\_yn
  - Whether the patient undergoes any screening, has values Yes, No
- count\_screening
  - Number of times the patient undergoes screening
- surgery\_yn
  - Whether the patient undergoes any bariatric surgery, has values Yes, No
- count\_surgery
  - Number of times the patient undergoes bariatric surgery
- first\_consult\_service\_date
  - Earliest date that the patient undergoes consultation from the time period of Q2 2015 to Q2 2019
- last\_consult\_service\_date
  - Latest date that the patient undergoes consultation from the time period of Q2 2015 to Q2 2019
- first\_surgery\_service\_date

- Earliest date that the patient undergoes surgery from the time period of Q2 2015 to Q2 2019
- Last\_surgery\_service\_date
  - Latest date that the patient undergoes surgery from the time period of Q2 2015 to Q2 2019

#### A.4 Prescription (Rx) information

- Total\_rx\_claims
  - Total number of medical claims the patient has from the time period of Q2 2015 to Q2 2019
- Total\_pd\_claims
  - Total number of paid medical claims the patient has time period of Q2 2015 to Q2 2019
- total\_pd\_saxenda\_claims
  - Total number of paid Saxenda® claims the patient has time period of Q2 2015 to Q2 2019
- total\_pd\_contrave\_claims
  - Total number of paid Contrave claims the patient has time period of Q2 2015 to Q2 2019
- total\_pd\_ksymia\_claims
  - Total number of paid Ksymia claims the patient has time period of Q2 2015 to Q2 2019
- Total\_pd\_belviq\_claims
  - Total number of paid Belviq claims the patient has time period of Q2 2015 to Q2 2019
- Total\_pd\_generic\_claims
  - Total number of paid Generic claims the patient has time period of Q2 2015 to Q2 2019
- Stdaln\_pd\_nonlifecycle\_claims
  - Number of Standalone paid non-lifecycle medical claims
- stdaln\_pd\_lifecycle\_claims
  - Number of Standalone paid lifecycle medical claims
- final\_pd\_claims
  - Number of paid lifecycle (Final status) medical claims
- stdaln\_rj\_nonlifecycle\_claims
  - Number of Standalone rejected non-lifecycle medical claims

- stdaln\_rj\_lifecycle\_claims
  - Number of Standalone rejected lifecycle medical claims
- stdaln\_rv\_nonlifecycle\_claims
  - Number of Standalone reversed non-lifecycle medical claims
- stdaln\_rv\_lifecycle\_claims
  - Number of Standalone reversed lifecycle medical claims
- initial\_rv\_claims
  - Number of reversed lifecycle (Initial status) medical claims
- Initial\_rj\_claims
  - Number of rejected lifecycle (Initial status) medical claims
- final\_rj\_claims
  - Number of rejected lifecycle (Final status) medical claims
- final\_rv\_claims
  - Number of reversed lifecycle (Final status) medical claims
- prescribed\_saxenda\_yn
  - Whether the patient has been prescribed Saxenda® from the time period of Q2 2015 to Q4 2019 regardless of claim type (Paid, rejected, reversed)
- prescribed\_other\_branded\_aoms\_yn
  - Whether the patient has been prescribed other branded AOMs from the time period of Q2 2015 to Q4 2019 regardless of claim type (Paid, rejected, reversed)
- prescribed\_generic\_aoms\_yn
  - Whether the patient has been prescribed Generic AOMs from the time period of Q2 2015 to Q4 2019 regardless of claim type (Paid, rejected, reversed)
- total\_opc\_saxenda
  - Total out of pocket cost the patient paid for all his/her Saxenda claims
- avg\_opc\_saxenda
  - Average out of pocket cost the patient paid for a single Saxenda claim, normalized over a 30 day supply
- total\_opc\_other\_branded\_aoms
  - Total out of pocket cost the patient paid for all his/her other branded AOMs claims
- Avg\_opc\_other\_branded\_aoms
  - Average out of pocket cost the patient paid for a single other branded AOM claim
- total\_opc\_generic\_aoms

- Total out of pocket cost the patient paid for all his/her oGeneric AOMs claims
- avg\_opc\_generic\_aoms
  - Average out of pocket cost the patient paid for a single Generic AOM claim
- rx\_most\_freq\_prescriber\_id
  - Identification number of the most frequent prescriber the patient goes to
- rx\_most\_freq\_state
  - The US State that the most frequent prescriber the patient goes to lives in
- rx\_most\_freq\_zip
  - The zip code of the most frequent prescriber the patient goes to
- rx\_most\_freq\_plantrak\_id
  - The insurance identification number of the most frequent (Rx) healthcare insurance plan the patient has signed for
- First\_prescription\_date
  - The earliest date the patient visits the prescriber from the time period of Q2 2015 to Q2 2019
- First\_paid\_prescription\_date
  - The earliest date the patient visits the prescriber and gets a paid claim from the time period of Q2 2015 to Q2 2019
- rx\_first\_prescriber\_id
  - Identification number of the first prescriber the patient goes to from the time period of Q2 2015 to Q2 2019
- rx\_first\_prescriber\_state
  - The US State that the first prescriber the patient goes to lives in from the time period of Q2 2015 to Q2 2019
- rx\_first\_prescriber\_zip
  - The US State that the first prescriber the patient goes to lives in from the time period of Q2 2015 to Q2 2019
- rx\_first\_plantrak\_id
  - The insurance identification number of the first (Rx) healthcare insurance plan the patient has signed for from the time period of Q2 2015 to Q4 2019
- first\_brand\_prescribed\_saxenda\_yn
  - Whether the first medication prescribed to the patient is Saxenda from the time period of Q2 2015 to Q4 2019, has values Yes, No
- first\_brand\_prescribed\_other\_branded\_aoms\_yn

- Whether the first medication prescribed to the patient is other branded AOMs from the time period of Q2 2015 to Q4 2019, has values Yes, No
- first\_brand\_prescribed\_generic\_aoms\_yn
  - Whether the first medication prescribed to the patient is Generic AOMs from the time period of Q2 2015 to Q4 2019, has values Yes, No
- latest\_prescription\_date
  - The latest date the patient visits the prescriber from the time period of Q2 2015 to Q2 2019
- Latest\_paid\_prescription\_date
  - The latest date the patient visits the prescriber and gets a paid claim from the time period of Q2 2015 to Q2 2019
- rx\_latest\_prescriber\_id
  - Identification number of the latest prescriber the patient goes to from the time period of Q2 2015 to Q2 2019
- rx\_latest\_prescriber\_state
  - The US State that the latest prescriber the patient goes to lives in from the time period of Q2 2015 to Q2 2019
- rx\_latest\_prescriber\_zip
  - The US State that the latest prescriber the patient goes to lives in from the time period of Q2 2015 to Q2 2019
- rx\_latest\_plantrak\_id
  - The insurance identification number of the latest (Rx) healthcare insurance plan the patient has signed for from the time period of Q2 2015 to Q4 2019
- latest\_brand\_prescribed\_saxenda\_yn
  - Whether the latest medication prescribed to the patient is Saxenda from the time period of Q2 2015 to Q4 2019, has values Yes, No
- latest\_brand\_prescribed\_other\_branded\_aoms\_yn
  - Whether the latest medication prescribed to the patient is other branded AOMs from the time period of Q2 2015 to Q4 2019, has values Yes, No
- Latest\_brand\_prescribed\_generic\_aoms\_yn
  - Whether the latest medication prescribed to the patient is Generic AOMs from the time period of Q2 2015 to Q4 2019, has values Yes, No

## A.5 Combined and Foreign Features

- joined\_prescriber\_id
  - From the values of dx\_most\_freq\_prescriber\_id, dx\_first\_prescriber\_id, dx\_latest\_prescriber\_id, rx\_most\_freq\_prescriber\_id, rx\_first\_prescriber\_id, rx\_latest\_prescriber\_id, choose the prescriber id that appears the most, or if it does not exist, or if the number is tied, choose from rx\_most\_freq\_prescriber\_id, if that does not exist, then choose dx\_most\_freq\_prescriber\_id
- nni\_saxenda\_gsb
  - Prescriber's Tier in terms of prescribing Saxenda® to a patient, Tier 1 has the highest chance of prescribing Saxenda® to a patient, No Tier has the lowest chance
- nni\_saxenda\_target
  - Whether the prescriber is a Saxenda® target, meaning is he/she worth marketing Saxenda® to so that he/she can prescribe Saxenda® to a patient, has values Yes, No
- zip
  - The zip code the joined\_prescriber lives in
- state
  - The US state the joined\_prescriber lives in
- joined\_plantrak\_id
  - From the values of dx\_most\_freq\_plantrak\_id, dx\_first\_plantrak\_id, dx\_latest\_plantrak\_id, rx\_most\_freq\_plantrak\_id, rx\_first\_plantrak\_id, rx\_latest\_plantrak\_id, choose the plantrak id that appears the most, or if it does not exist, or if the number is tied, choose from rx\_most\_freq\_plantrak\_id, if that does not exist, then choose dx\_most\_freq\_plantrak\_id
- method\_of\_payment
  - Patient's method of payment of paying for a medical claim
- model\_type
  - Patient's health insurance model type
- days\_between\_first\_diag\_latest\_diag
  - Number of days between the patient's first diagnosis and the latest diagnosis from the time period of Q2 2015 to Q4 2019
- days\_between\_first\_consult\_latest\_consult
  - Number of days between the patient's first consultation and the latest consultation from the time period of Q2 2015 to Q4 2019
- days\_between\_first\_surgery\_latest\_surgery
  - Number of days between the patient's first surgery and the latest surgery from the time period of Q2 2015 to Q4 2019
- days\_between\_first\_prescr\_latest\_prescr

- Number of days between the patient's first prescription and the latest prescription from the time period of Q2 2015 to Q4 2019
- days\_between\_first\_pd\_prescr\_latest\_pd\_prescr
  - Number of days between the patient's first paid prescription and the latest paid prescription from the time period of Q2 2015 to Q4 2019
- days\_between\_first\_consult\_latest\_surgery
  - Number of days between the patient's first consultation and the latest surgery from the time period of Q2 2015 to Q4 2019
- days\_between\_first\_diag\_latest\_prescr
  - Number of days between the patient's first diagnosis and the latest prescription from the time period of Q2 2015 to Q4 2019
- days\_between\_first\_diag\_first\_prescr
  - Number of days between the patient's first diagnosis and the first prescription from the time period of Q2 2015 to Q4 2019
- days\_between\_latest\_diag\_latest\_prescr
  - Number of days between the patient's latest diagnosis and the latest prescription from the time period of Q2 2015 to Q4 2019
- days\_between\_latest\_diag\_first\_prescr
  - Number of days between the latest diagnosis and the first prescription from the time period of Q2 2015 to Q4 2019

## Appendix: Features of Prescriber flat table

- first\_name
  - Diagnosis Doctor/Prescriber's first name
- last\_name
  - Diagnosis Doctor/Prescriber's last name
- address
  - Diagnosis Doctor/Prescriber's home address
- city
  - City the Diagnosis Doctor/prescriber lives in
- state
  - US state the Diagnosis Doctor/prescriber lives in
- zip
  - Zip code the Diagnosis Doctor/prescriber lives in
- specialty\_code
  - The Diagnosis Doctor/prescriber's specialty initials
- specialty\_desc
  - The Diagnosis Doctor/prescriber's specialty
- nni\_saxenda\_gsb
  - The Diagnosis Doctor/prescriber's Saxenda® prescribing Tier
- nni\_saxenda\_segment
  - The Diagnosis Doctor/prescriber's Saxenda® segment
- nni\_saxenda\_target
  - Whether the Diagnosis Doctor/prescriber is a Saxenda® target, has values Yes, No
- nni\_c3\_saxenda\_flag
  - Whether the Diagnosis Doctor/prescriber has prescribed Saxenda® before, has values Yes, No
- nni\_c3\_baom\_flag
  - Whether the Diagnosis Doctor/prescriber's has prescribed other branded AOMs before
- nni\_saxenda\_quintile
  - The Diagnosis Doctor/prescriber's Saxenda® quintile
- nni\_baom\_decile
  - The Diagnosis Doctor/prescriber's branded AOM decile



## Appendix: Features of Plantrak flat table

- payer\_name
  - The health insurance's type, whether its government based, private or cash
- plan\_name
  - The health insurance's specific type, more specific to region
- method\_of\_payment
  - Patient's method of payment using this health insurance type
- model\_type
  - Health insurance's model type
- mc\_category\_cd
  - Health insurance's category code
- mc\_category\_cd\_name
  - Health insurance's category code name

# Appendix: Features of Socioeconomic Flat Table

## D.1 Census Tract Level

### Age

- % age 25-34
  - Estimated percent of the population age 25 to 34 years, between 2014-2018
- % age 35-44
  - Estimated percent of the population age 35 to 44 years, between 2014-2018
- % age 45-54
  - Estimated percent of the population age 45 to 54 years, between 2014-2018
- % age 55-64
  - Estimated percent of the population age 55 to 64 years, between 2014-2018
- % age 65 or older
  - Estimated percent of the population age 65 years and older, between 2014-2018.
- Median age
  - Estimated median age of the population between 2013-2017

### Education

- % less than HS diploma
  - Estimated percent of population 25 years and older with less than a high school diploma or equivalent between 2013-2017
- % HS Diploma
  - Estimated percent of population 25 years and older with a high school diploma or equivalent and no college between 2013-2017
- % College
  - Estimated percent of the population 25 years and older whose educational attainment is some college, or an Associate's degree, between 2013-2017
- % Bachelor's
  - Estimated percent of population 25 years and older with a Bachelor's degree, between 2014-2018
- % Post Grad
  - Estimated percent of population 25 years and older with a graduate or professional degree between 2013-2017

## **Household and Spending**

- Average household size
  - Estimated average household size, between 2014-2018. A household includes all the people who occupy a housing unit as their usual place of residence
- Median gross rent
  - Estimated median gross rent for rental units with cash rent, between 2014-2018. Gross rent is the contract rent plus the estimated average monthly cost of utilities and fuels if these are paid by the renter). Gross rent is intended to eliminate differentials that result from varying practices with respect to the inclusion of utilities and fuels as part of the rental payment.
- Median value of an owner-occupied home
  - Estimated median value of an owner-occupied housing unit, between 2014-2018. The value is based on survey respondents' estimates of how much their properties and lots would sell for if they were for sale.
- Median household income
  - Estimated median household income in the past 12 months, as reported between 2014-2018. A household includes all the people who occupy a housing unit as their usual place of residence
- Median renter cost burden
  - Median gross rent as a percentage of household income, between 2014-2018
- Median owner cost burden
  - Estimated median selected monthly owner costs as a percentage of household income, for all owner-occupied housing units (with and without a mortgage), between 2014-2018

**Ethnicity** Data was obtained from the Census' American Community Survey 2013-2017 estimates

- Predominant ethnic group
  - Predominant racial or ethnic group, by percentage of the population in the group
- % White
  - Estimated percent of the population that is White, by single classification of Census race, between 2013-2017
- % Hispanic
  - Estimated percent of the population that is Hispanic or Latino between 2013-2017
- % Black
  - Estimated percent of the population that is Black or African American, by single classification of Census race, between 2013-2017
- % Asian
  - Estimated percent of the population that is Asian, by single classification of Census race, between 2013-2017

## D.2 County Level

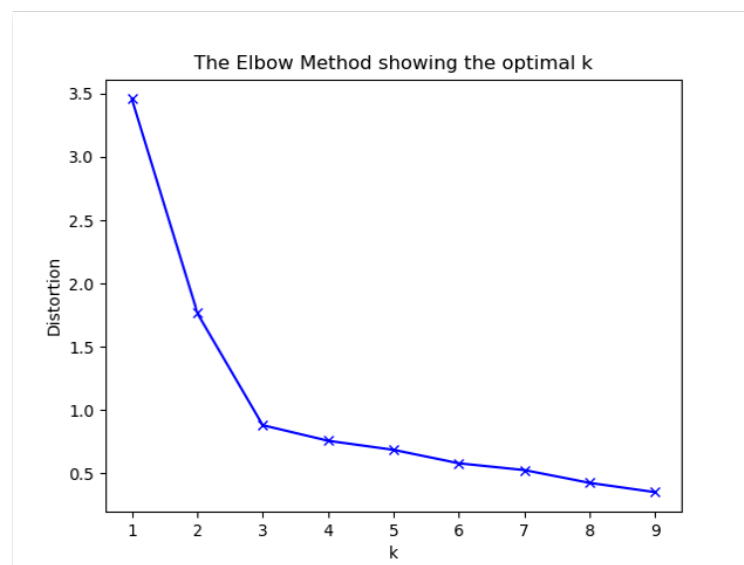
**Unemployment Rate** Annual unemployment rate in 2018. The unemployment rate represents the number of unemployed people as a percent of people in the labor force.

## Appendix: K-Means Algorithm

In the algorithm, each data point is classified by computing the distance between that point and each group center, and then classifying the point to be in the group whose center is closest to it. Based on these classified points, the algorithm recomputes the group center by taking the mean of all the vectors in the group, and these steps are repeated until the centers do not change much between iterations [4].

The disadvantage of K-Means is that it is considered a hard clustering method and initializes a data point to a specific cluster and not any other clusters. However, large datasets may not be as binary and clear-cut, it is possible for a datapoint to exist in different clusters with different probabilities of belonging to that cluster. K-Means also fails in cases where the clusters are not circular.

A strong indicator of whether K-Means is a good clustering method for our dataset is by plotting an elbow plot, which looks at how the standard error is reduced over a range of clusters. When there is a distinct point (the elbow) in the plot that shows the reduction in standard error is extremely insignificant with an increase of 1 cluster, that point is considered the best point that optimizes the bias-variance trade-off, and the x-coordinate shows the optimal number of clusters to fit the data. The figure below shows an example of a good elbow plot with  $k = 3$  as the optimal number of clusters.



**Figure 31:** *Illustration of a good Elbow Plot [5]*

## Appendix: K-Prototypes Algorithm

Traditionally machine learning models could only handle either totally numerical data or totally categorical data. For these two extreme ends of the spectrum there are various set of clustering algorithms. If we have a mix of categorical data and numerical data and the categorical data is hard to encode, or the cardinality of the data is high, there is no robust clustering algorithm.

K-Prototypes was created as a way to cluster this kind of data [6]. In this algorithm each prototype has a value for the numerical columns as well as the categorical values and the distance of each point is measured as a dissimilarity number form these particular values. Due to this, the distance calculation becomes tedious and time consuming and the complexity of the algorithm scales up very fast as the number of data point increase.

A typical flow of this algorithm is given below and it is very similar to K-Means:

1. Read parameter
2. Initial prototypes
3. Initial allocation
4. Reallocation
5. Satisfy loop ending condition
6. Program output

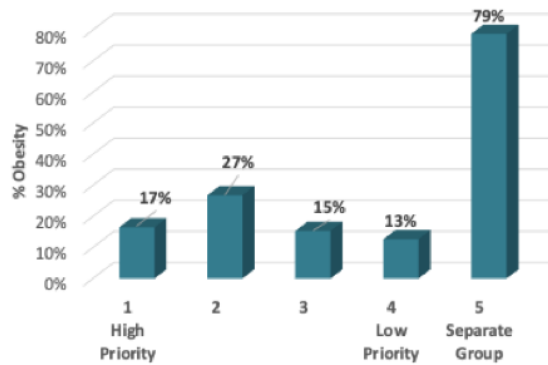
Like K-Means, this algorithm is a hard clustering algorithm which sees the data in black and white. Since, this algorithm is an extension of how K-Means works, this algorithm also suffers from the same shortcomings as the K-Means algorithm.

## Appendix: Gaussian Mixture Model

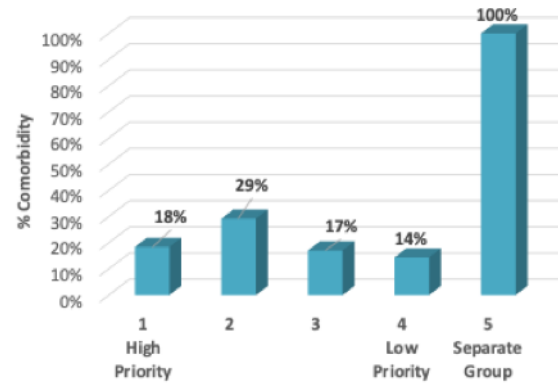
GMM assumes the data points are Gaussian distributed, which is a less restrictive assumption than saying they are circular by using the mean. GMM uses the mean and standard deviation to describe the shape of the cluster. As a result, clusters can vary in shape and size more significantly than K-Means, which proves the higher flexibility that this algorithm provides more than others. GMM follows the same methodology as the K-Means algorithm but instead of just calculating the optimized mean, it also computes the optimized standard deviation using a built-in algorithm called Expectation-Maximization (EM) [4].

Similar to the K-Means elbow plot, a strong indicator of whether a GMM is a good clustering algorithm for the dataset is to plot the Bayesian Information Criterion (BIC) over a range of clusters. The BIC score estimates how accurate the model fits the data based on the initial number of numbers selected. The optimal number of clusters is selected by looking at the point that optimizes the bias-variance trade-off (the elbow).

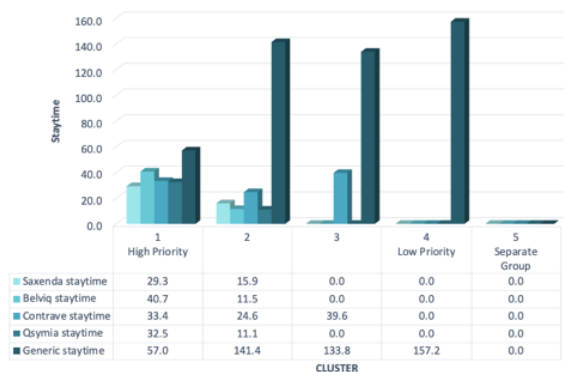
## Appendix: Patient Level Visualizations



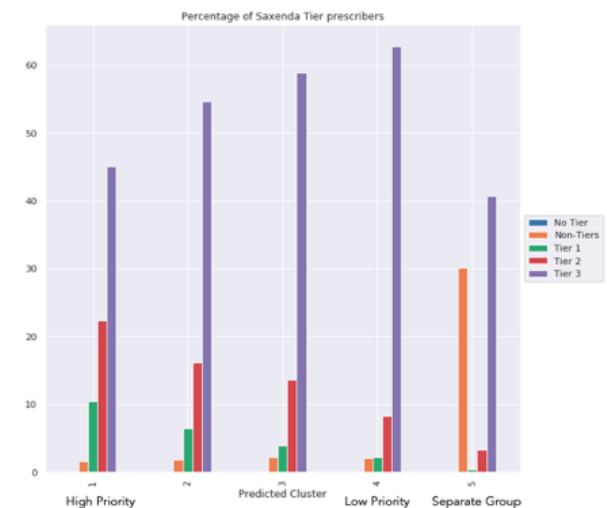
**Figure 32:** % of patients diagnosed with obesity



**Figure 33:** % of patients diagnosed with comorbidity



**Figure 34:** Stay-times across AOMs



**Figure 35:** % of the different Saxenda®-Tiered prescribers



**Figure 36:** Number of graduates at different education levels



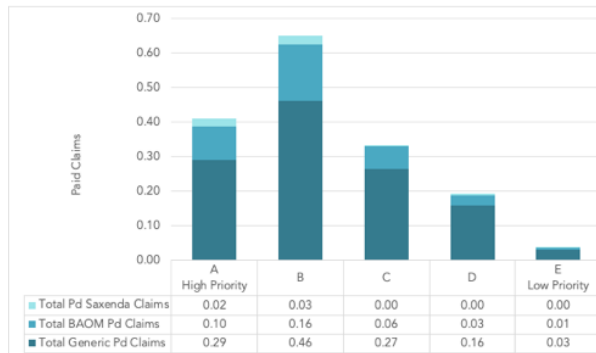


**Figure 37:** *Median Gross rent*

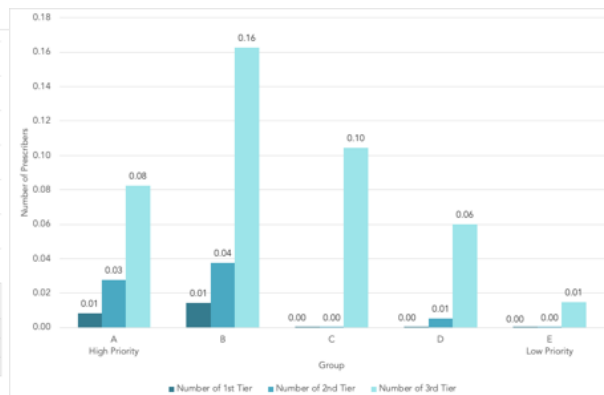


**Figure 38:** *Median household income*

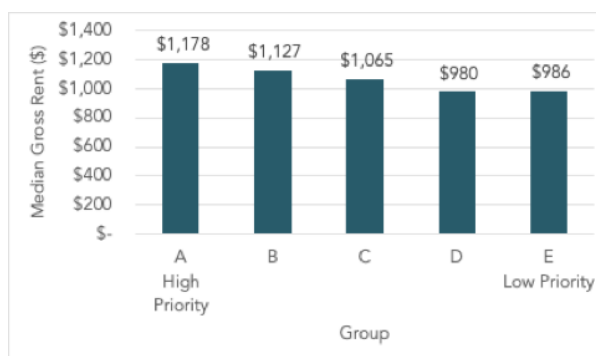
## Appendix: ZIP Code Level Visualizations



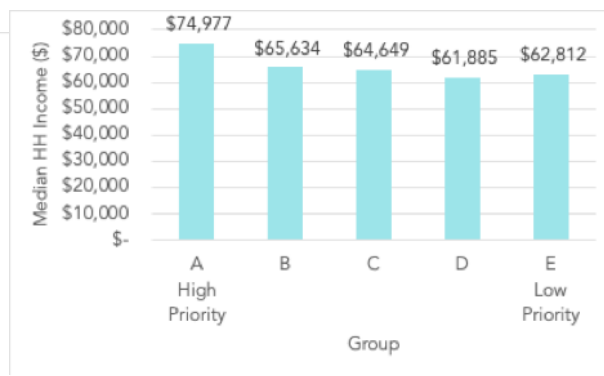
**Figure 39:** *Normalized total paid claims across AOMs*



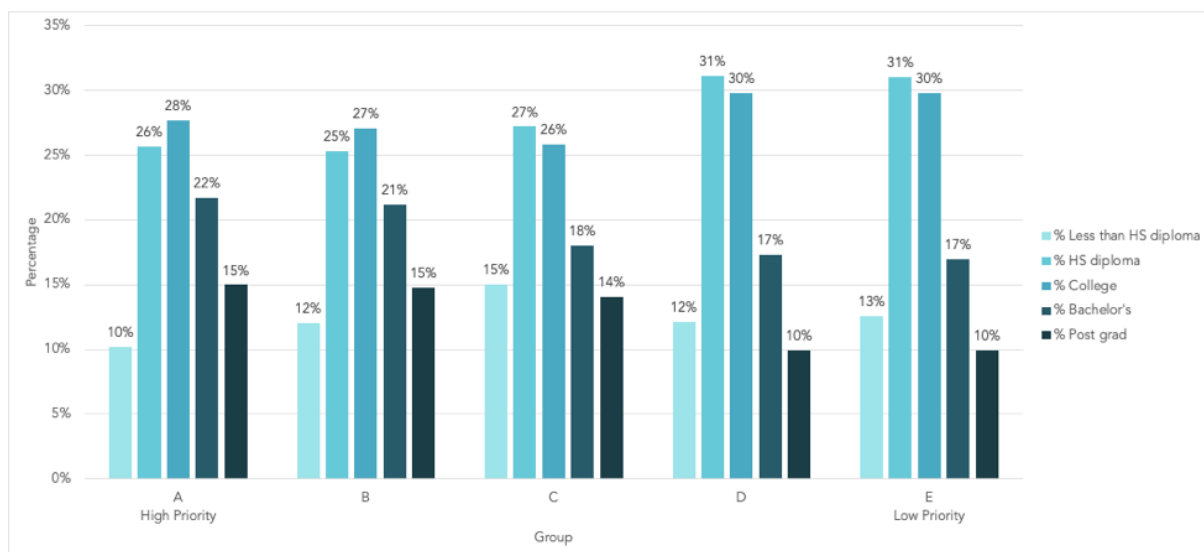
**Figure 40:** *Normalized number of Saxenda®-Tiered prescribers*



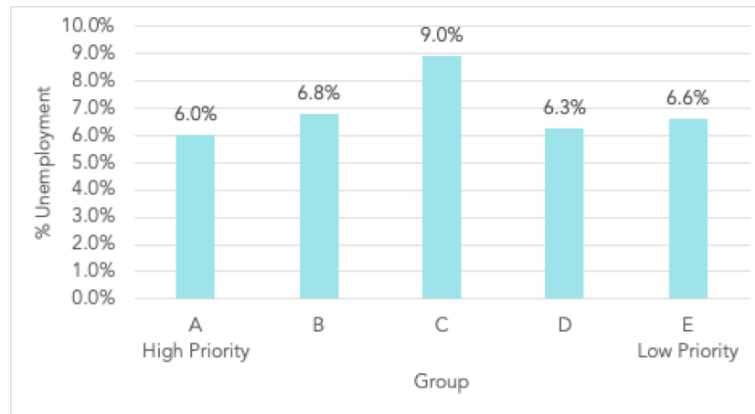
**Figure 41:** *Median Gross Rent*



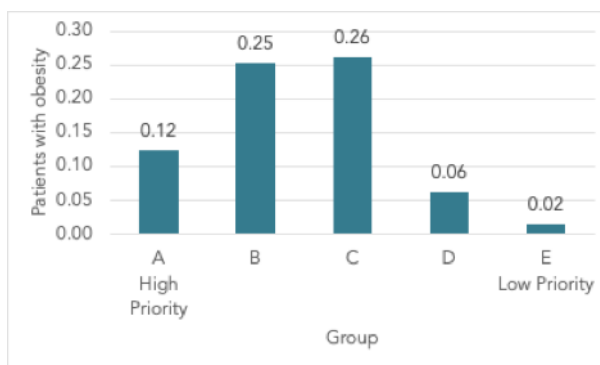
**Figure 42:** *Median Household Income*



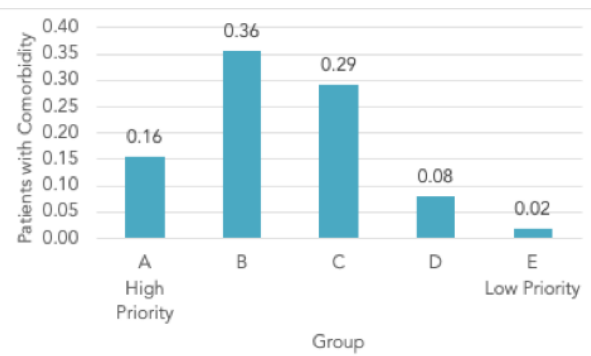
**Figure 43:** *Percentage of graduates at each education level*



**Figure 44:** *Unemployment Rate*



**Figure 45:** *% of patients diagnosed with obesity*



**Figure 46:** *% of patients diagnosed with comorbidities*

## Appendix: Combined Patient and ZIP code level Visualizations

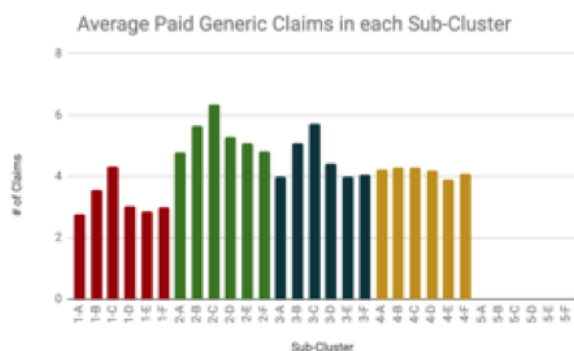


Figure 47: Median Gross Rent

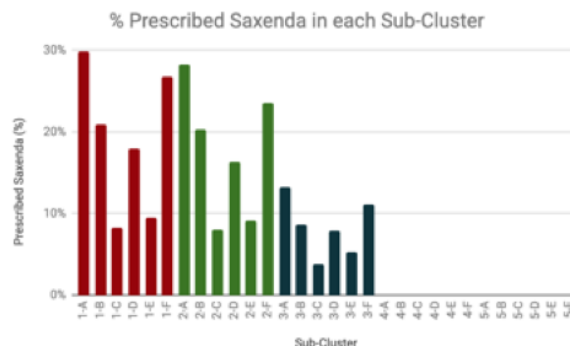


Figure 48: Median Household Income

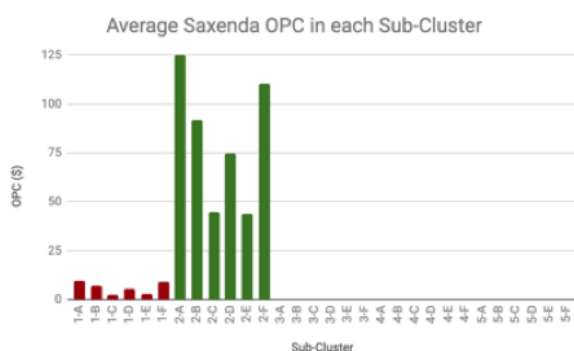


Figure 49: Median Gross Rent

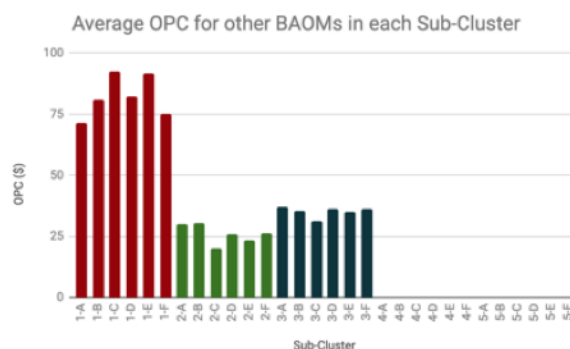


Figure 50: Median Household Income



Figure 51: Median Gross Rent

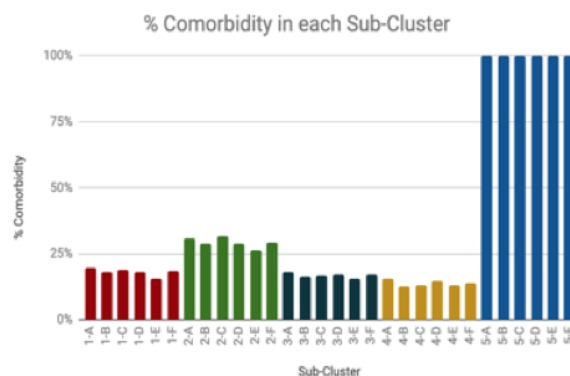
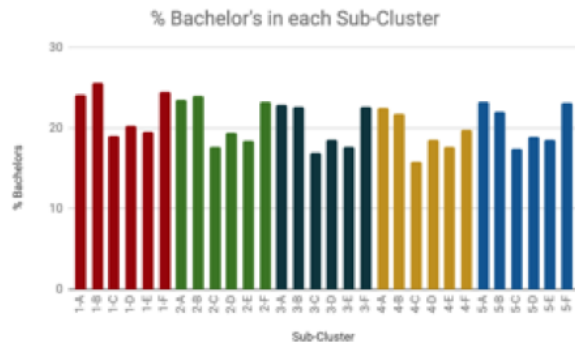
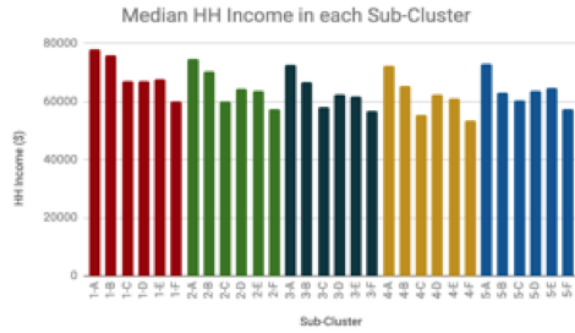


Figure 52: Median Household Income



**Figure 53:** *Median Gross Rent*



**Figure 54:** *Median Household Income*