Apply Data Science to Improving Addiction Treatment

Zhe Du

August 1st, 2019

Contents

Motivation

Every day, more than 130 people in the United States die after overdosing on opioids[1]. Addiction to and misuse of opioids, such as heroin, prescription pain relievers, and synthetic opioids such as fentanyl, has reached alarming levels. This issue has destroyed countless families, and has placed a heavy burden on the overall economy, including increased costs of healthcare, addiction treatment, criminal justice involvement, and lost productivity, etc.

Therefore, it has become ever more critical for healthcare providers to extract insights and better understand the addiction population for providing effective treatment.

Data science is one of the ways to help achieve this.

As our organization continuously strives to improve treatment quality, I've recently worked on a data science project to apply machine learning techniques to understand key common attributes among patients, and discover any unique clusters. Continuity of care is crucial for treating substance use disorder. Our goal is to discover key causes for patients to quit recovery programs, and identify areas for improvement, and design more effective intervention.

The analysis presented in this report is only one piece of a working-in-progress effort towards the goal. Upon approval by our organization, I'd like to share some of the key findings from this project, hoping to inspire meaningful discussions about how data science and technologies can transform addiction treatment.

Summary

In this report, I will go through step by step on how I used unsupervised machine learning to extract clinical insights and conduct patients segmentation analysis from survey results.

The survey used in this project is called Brief Addiction Monitor (BAM). The complete survey with scoring & clinical guidelines are publicly available, which can be found following this link.

This report follows the BSPF project framework developed by Matt Dancho, an enhanced CRISP-DM Methodology designed specifically for solving business problems using data science.

Here is the outline of this report:

1. Problem Understanding
2. Data Understanding
3. Data Preparation
4. Modeling
5. Validation
6. Impact
7. What's Next

Disclaimer

The findings extracted from this analysis is entired based on data sets from within our organization. Although we'd like to think insights drawn from this analysis could represent a larger scope (given the sample size), you might find otherwise. To reproduce the analysis, the dataset used in this report can be found here: [TBD]. **Please note, to completely ensure patients' privacy, the data has been de-identified to stay HIPAA compliant. Column "display_id" contains an id arbitrarily assigned to each patient for analysis purposes only.**

[1] National Institute on Drug Abuse - https://www.drugabuse.gov/drugs-abuse/opioids/opioid-overdose-crisis

1. Problem Understanding

There are various reasons why patients might choose to stop going to clinics for treatment. Some patients go back for readmission, and some don't. Although it's not rare to see patients leave the program, we want to have a much better understanding of the causes. We'd also like to examine if there are any common attributes shared among patients who left our program, versus patients who didn't. Then we can identify ways to reduce the patients churn rate.

2. Data Understanding

Exploratory Data Analysis (EDA) is an observational step that allows us to understand the characteristics of the data we are working with. This step helps us evaluate key drivers, develop KPIs, identify problems and opportunities, and ultimately guide us to choose an appropriate approach for a project.

2.1 Dataset Overview

The original survey contains a total of 17 questions (see link above). In addition to the 17 questions, a couple of demographic questions are also added, such as:

- Question 18: Living Situation
- Question 19: Employment Status

These two questions, along with "Question 17 : Patient Satisfaction", will be held out initially, but they will be used for cluster analysis at a later stage. Only the scored questions will be used in this analysis, and therefore, "Question 7 A - G" are not considered.

Note, each patient completes this survey regularly. The dataset contains all responses from all surveys each patient takes. Therefore, I calculated the most frequent answers from each patient for each question, so each patient only has one row of the most frequent responses. Here is a glimpse of the dataset: (values represent response encodings)

```
## Observations: 2,731
## Variables: 15
## $ Question_16 <dbl> 3, 3, 4, 4, 4, 4, 4, 4, 2, 3, 4, 1, 3, 4, 1, 4, 4,...
## $ Question_14 <dbl> 4, 4, 4, 4, 4, 4, 0, 4, 0, 4, 4, 4, 0, 4, 4, 4, 0,...
## $ Question_15 <dbl> 0, 1, 0, 0, 0, 0, 0, 0, 1, 0, 0, 1, 0, 0, 0, 0, 0,...
## $ Question_12 <dbl> 3, 4, 4, 4, 0, 3, 0, 4, 1, 2, 1, 2, 4, 3, 1, 4, 4,...
## $ Question_11 <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 2, 1, 1, 0, 0, 0, 0, 3,...
## $ Question_10 <dbl> 0, 2, 0, 0, 2, 0, 0, 0, 1, 0, 1, 1, 0, 0, 1, 2, 0,...
## $ Question_8  <dbl> 2, 0, 0, 0, 4, 0, 3, 0, 0, 1, 0, 2, 0, 0, 0, 0, 1,...
## $ Question_9  <dbl> 4, 4, 3, 4, 4, 4, 1, 4, 4, 0, 1, 2, 4, 4, 4, 3, 1,...
## $ Question_5  <dbl> 0, 0, 0, 0, 0, NA, 3, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
## $ Question_3  <dbl> 1, 0, 0, 0, 4, 0, 2, 0, 0, 2, 0, 1, 1, 0, 0, 2, 3,...
## $ Question_2  <dbl> 0, 1, 0, 0, 0, 0, 2, 0, 0, 3, 0, 0, 1, 4, 0, 4, 4,...
## $ Question_1  <dbl> 2, 2, 2, 0, 2, 2, 3, 2, 2, 2, 2, 2, 2, 4, 0, 2, 3,...
## $ Question_6  <dbl> 0, 0, 0, 0, 4, 0, 4, 0, 0, 2, 1, 1, 0, 0, 0, 0, 3,...
## $ Question_13 <dbl> 0, 2, 0, 0, 0, 4, 0, 2, 1, 0, 0, 0, 0, 0, 0, 0, 0,...
## $ Question_4  <dbl> 0, 0, 0, 0, 1, 0, 4, 0, 0, 0, 1, 0, 0, 0, 0, 0, 3,...
```

2.2 Feature Variables Exploration

Let's take a look at how the response data for each question is distributed, based on question categories:

Based on the scoring guideline provided (see link above), questions are grouped into three categories: "Protective Factors", "Risk Factors", and "Drug Use". The facet chart above is colored and ordered to reflect each category accordingly.

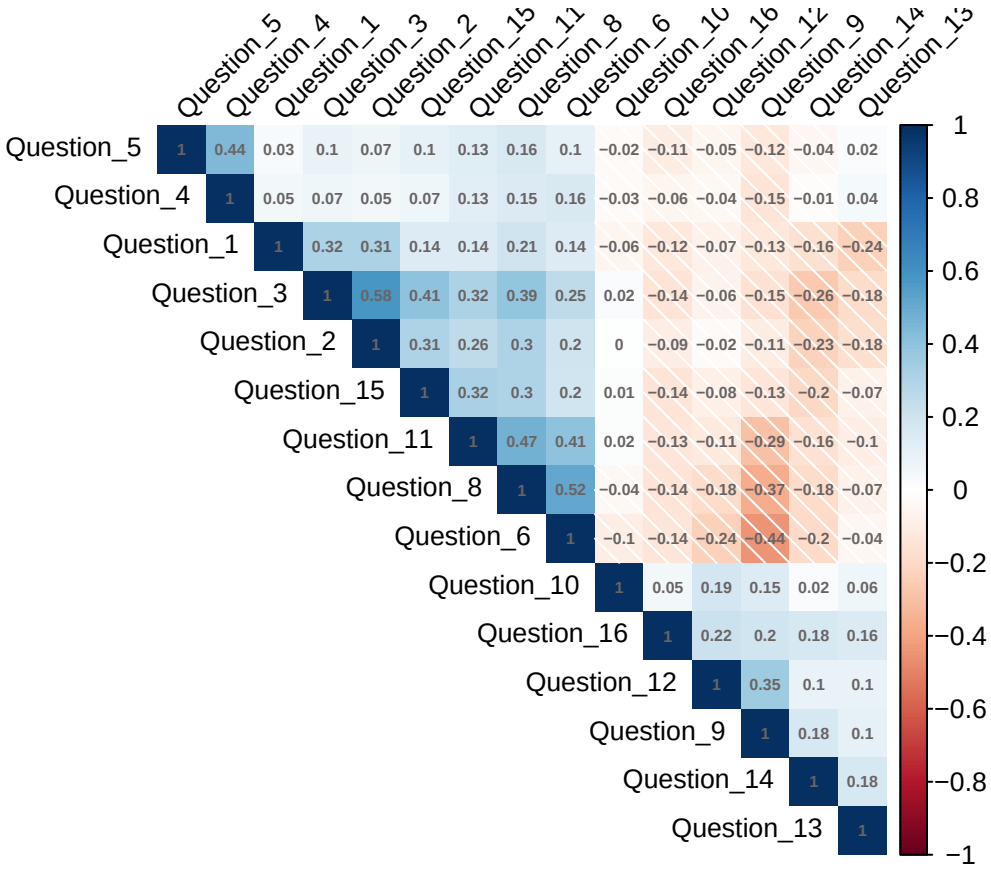
The response distributions above might also suggest possible correlations among certain variables. A correlation plot will help us further investigate the relationships. But first, it's always a good practice to format datasets in a "human-readable" format for data exploration, and a "machine-readable" format for modeling.

Currently (as shown above), the response data is not quite "human-readable", as it's filled with encodings. To make it more readable, let's convert score encodings into actual definition description, and glimpse the output:

```
## Observations: 2,731
## Variables: 15
## $ Question_1  <fct> Good, Good, Good, Excellent, Good, Good, Fair, Goo...
## $ Question_10 <fct> 0, 4-8, 0, 0, 4-8, 0, 0, 0, 1-3, 0, 1-3, 1-3, 0, 0...
## $ Question_11 <fct> 0, 0, 0, 0, 0, 0, 0, 0, 0, 4-8, 1-3, 1-3, 0, 0, 0,...
## $ Question_12 <fct> Considerably, Extremely, Extremely, Extremely, Not...
## $ Question_13 <fct> 0, 4-8, 0, 0, 0, 16-30, 0, 4-8, 1-3, 0, 0, 0, 0, 0...
## $ Question_14 <fct> Yes, Yes, Yes, Yes, Yes, Yes, No, Yes, No, Yes, Ye...
## $ Question_15 <fct> Not at all, Slightly, Not at all, Not at all, Not ...
## $ Question_16 <fct> 9-15, 9-15, 16-30, 16-30, 16-30, 16-30, 16-30, 16-...
```

```
## $ Question_2  <fct> 0, 1-3, 0, 0, 0, 0, 4-8, 0, 0, 9-15, 0, 0, 1-3, 16...
## $ Question_3  <fct> 1-3, 0, 0, 0, 16-30, 0, 4-8, 0, 0, 4-8, 0, 1-3, 1-...
## $ Question_4  <fct> 0, 0, 0, 0, 1-3, 0, 16-30, 0, 0, 0, 1-3, 0, 0, 0, ...
## $ Question_5  <fct> 0, 0, 0, 0, 0, NA, 9-15, 0, 0, 0, 0, 0, 0, 0, 0, 0...
## $ Question_6  <fct> 0, 0, 0, 0, 16-30, 0, 16-30, 0, 0, 4-8, 1-3, 1-3, ...
## $ Question_8  <fct> Moderately, Not at all, Not at all, Not at all, Ex...
## $ Question_9  <fct> Extremely, Extremely, Considerably, Extremely, Ext...
```

Now the response data is in a much more readable format. I can proceed with the correlation plot for all feature variables. The responses are ordinal categorical variables, and they are not normally distributed. Therefore, instead of using the popular Pearson correlation, the Spearman correlation will be used to examine the ranked ordinal association. Using hierarchical clustering order for the correlation matrix, we can observe correlation patterns of the feature variables much more clearly:



The plot above clearly shows that some questions are much more correlated than others, either positively or negatively. Understanding the correlations of our feature variables facilitates feature selections and model performance.

2.3 Key Takeaways From EDA

- All responses are categorical data (either ordinal or discretized).
- All responses have the same scoring scale. Each question has five levels, except for "Questions 14", which has two levels.
- Some responses are highly skewed and have almost zero variance, such as "Question 5".
- Certain variables are much more correlated than others, further investigation is needed for the optimal feature selection.
- There are missing values to be handled during data preparation step.
- Feature variables are categorical, which requires an appropriate modeling approach.

3. Data Preparation

Now that I have a good grasp of the data I'm working with, the next stage is to prepare our data for modeling. Why data preparation? Remember the dataset was converted into a "human-readable" format? I now need to convert it into a "machine-readable" format as well, so that I can apply machine learning techniques for modeling.

There are several common key steps for data preprocessing:

1. Imputation

The first step is to identify any missing values in the dataset. The Skimr package in R provides us a very convenient way for doing this:

```
## # A tibble: 15 x 10
##    variable    missing complete  mean    p0  p100   p25   p50   p75    sd
##    <chr>         <dbl>    <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
##  1 Question_5      232     2499 0.182     0     4     0     0     0 0.634
##  2 Question_1        0     2731 2.21      0     4     2     2     3 0.917
##  3 Question_10       0     2731 0.684     0     4     0     0     1 1.13
##  4 Question_11       0     2731 1.07      0     4     0     0     2 1.47
##  5 Question_12       0     2731 2.21      0     4     1     2     4 1.48
##  6 Question_13       0     2731 1.33      0     4     0     0     4 1.76
##  7 Question_14       0     2731 2.03      0     4     0     4     4 2.00
##  8 Question_15       0     2731 0.534     0     4     0     0     1 0.937
##  9 Question_16       0     2731 2.89      0     4     2     4     4 1.48
## 10 Question_2        0     2731 1.51      0     4     0     1     3 1.60
## 11 Question_3        0     2731 1.36      0     4     0     1     3 1.50
## 12 Question_4        0     2731 0.395     0     4     0     0     0 0.935
## 13 Question_6        0     2731 1.56      0     4     0     1     3 1.53
## 14 Question_8        0     2731 1.19      0     4     0     1     2 1.27
## 15 Question_9        0     2731 2.30      0     4     1     2     4 1.38
```

From the output, We can clearly see that "Question 5" contains 232 missing values. There are many ways to deal with missing data. Here is an article for a good overview of them.

Before deciding which approach to use, let's see if "Question_5" is actually needed as a feature variable. In the correlation plot above, we see that "Question_4" and "Question_5" are highly correlated, which suggests one might provide the same information as the other. Furthermore, looking at these two questions, they essentially ask the same thing — alcohol usage. Therefore, the modeling power should still be maintained without "Question_5" as a variable.

2. Remove Zero Variance Features

Since no variables have zero variance, this step will be skipped.

3. Individual Transformations for Skewness

The dataset does contain skewed variables, but since they are all categorical variables, feature transformation won't be performed here.

4. Normalization (center, scale, range, etc)

Because I will create a dummy variable for each categorical variable, this step will be skipped as well. It doesn't make sense to normalize dummy variables.

5. Create Dummy Variables

To create dummy variables or any other preprocessing steps, we can use the Recipes package from Tidymodels. Recipes is a powerful package that does a lot of heavy lifting for preprocessing data.

As you'd imagine, the output will be fairly wide, but we can easily glimpse it:

```
## Observations: 2,731
## Variables: 53
## $ Question_1_Very.Good     <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
## $ Question_1_Good          <dbl> 1, 1, 1, 0, 1, 1, 0, 1, 1, 1, 1, 1, 1...
## $ Question_1_Fair          <dbl> 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0...
## $ Question_1_Poor          <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
## $ Question_10_X1.3         <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 1, 1, 0...
## $ Question_10_X4.8         <dbl> 0, 1, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0...
## $ Question_10_X9.15        <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
## $ Question_10_X16.30       <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
## $ Question_11_X1.3         <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 0...
## $ Question_11_X4.8         <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0...
## $ Question_11_X9.15        <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
## $ Question_11_X16.30       <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
## $ Question_12_Slightly     <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 1, 0, 0...
## $ Question_12_Moderately   <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 1, 0...
## $ Question_12_Considerably <dbl> 1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0...
## $ Question_12_Extremely    <dbl> 0, 1, 1, 1, 0, 0, 0, 1, 0, 0, 0, 0, 1...
## $ Question_13_X1.3         <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0...
## $ Question_13_X4.8         <dbl> 0, 1, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0...
## $ Question_13_X9.15        <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
## $ Question_13_X16.30       <dbl> 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0...
## $ Question_14_Yes          <dbl> 1, 1, 1, 1, 1, 1, 0, 1, 0, 1, 1, 1, 0...
## $ Question_15_Slightly     <dbl> 0, 1, 0, 0, 0, 0, 0, 0, 1, 0, 0, 1, 0...
## $ Question_15_Moderately   <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
## $ Question_15_Considerably <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
## $ Question_15_Extremely    <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
## $ Question_16_X1.3         <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0...
## $ Question_16_X4.8         <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0...
## $ Question_16_X9.15        <dbl> 1, 1, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 1...
## $ Question_16_X16.30       <dbl> 0, 0, 1, 1, 1, 1, 1, 1, 0, 0, 1, 0, 0...
## $ Question_2_X1.3          <dbl> 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1...
## $ Question_2_X4.8          <dbl> 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0...
## $ Question_2_X9.15         <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0...
## $ Question_2_X16.30        <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
## $ Question_3_X1.3          <dbl> 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1...
## $ Question_3_X4.8          <dbl> 0, 0, 0, 0, 0, 0, 1, 0, 0, 1, 0, 0, 0...
## $ Question_3_X9.15         <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
## $ Question_3_X16.30        <dbl> 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0...
## $ Question_4_X1.3          <dbl> 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 1, 0, 0...
## $ Question_4_X4.8          <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
## $ Question_4_X9.15         <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
```

```
## $ Question_4_X16.30        <dbl> 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0...
## $ Question_6_X1.3          <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 0...
## $ Question_6_X4.8          <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0...
## $ Question_6_X9.15         <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
## $ Question_6_X16.30        <dbl> 0, 0, 0, 0, 1, 0, 1, 0, 0, 0, 0, 0, 0...
## $ Question_8_Slightly      <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0...
## $ Question_8_Moderately    <dbl> 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0...
## $ Question_8_Considerably  <dbl> 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0...
## $ Question_8_Extremely     <dbl> 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0...
## $ Question_9_Slightly      <dbl> 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0...
## $ Question_9_Moderately    <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0...
## $ Question_9_Considerably  <dbl> 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
## $ Question_9_Extremely     <dbl> 1, 1, 0, 1, 1, 1, 0, 1, 1, 0, 0, 0, 1...
```

6. Multivariate Transformation (e.g. PCA, spatial sign, etc)

Feature reduction can be very useful in some cases. For example, when dealing with a large number of features, we can use PCA to construct a much smaller number of components that capture the most variance within the dataset. Then we can use these components to build predictive models, such as a classification model using LDA.

Since this dataset doesn't contain a significantly large number of features, and we are primarily interested in clustering the data, we can skip this step.

4. Modeling

Now that the data is in a "machine-readable" format, it's time to conduct unsupervised machine learning on the dataset to identify if any patients share a common set of characteristics or trends. Clustering, or patient segmentation analysis, is a powerful way to extract those insights.

4.1 Choose The Appropriate Algorithm

One of the most well-known clustering algorithms is K-Means, which uses a distance measurement to calculate the similarity among observations. Because of this, it performs very well with numeric features. However, as previously seen, this dataset contains only categorical features, and to measure a euclidean distance between "Slightly" and "Moderately" doesn't make much sense.

Therefore, instead of using K-Means to calculate distance similarity among categorical observations, I will use K-Modes, a clustering algorithm developed by Zhexue Huang, that essentially records which answer to each question got the most votes, and then uses these modes for clustering.

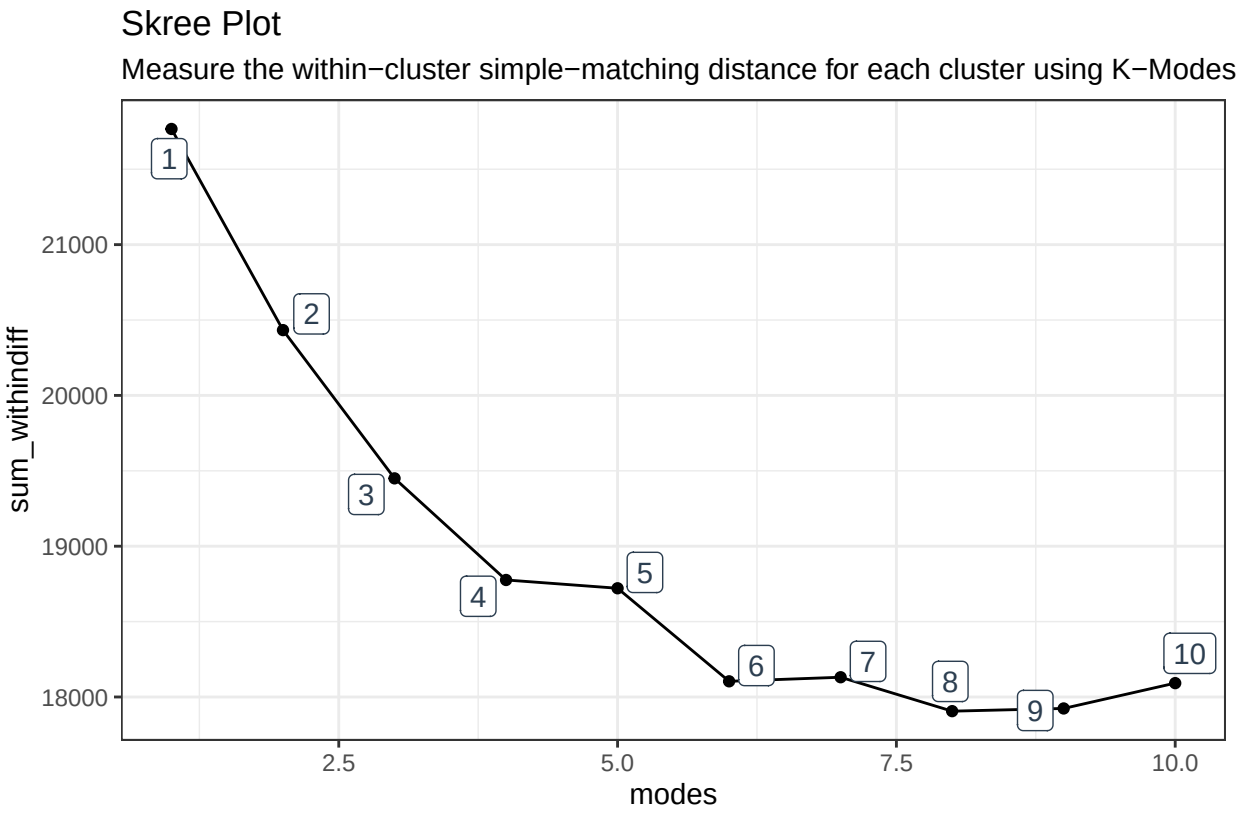
I find this algorithm works well with surveys composed of all categorical variables. K-means and K-modes are similar concepts, as they are all centroid based, but they are different under the hood. Having researched the topic in-depth, I found this article does a good job explaining the differences.

4.2 K-Modes Clustering

Like K-Means, we still need to choose the number of clusters (K) in advance, which can be one disadvantage of such algorithms in some cases.

4.2.1 Identify The Best K

There is no magic formula for choosing the optimal K, because often times the desired number of clusters is really dictated by the project goal. Although some technical methods, (such as the "Elbow"method, Silhouette method, Gap Statistic, etc.) can guide us towards picking a good K, we should also rely on our domain knowledge and consider what we are trying to achive and communicate.
The goal of this project is to identify any common patterns or characteristics of patients who left the program vs. those who didn't. Therefore, without looking at any outputs from technical methods, an ideal number of clusters would be 2. But let's first output a scree plot and see what the "Elbow" method shows us:



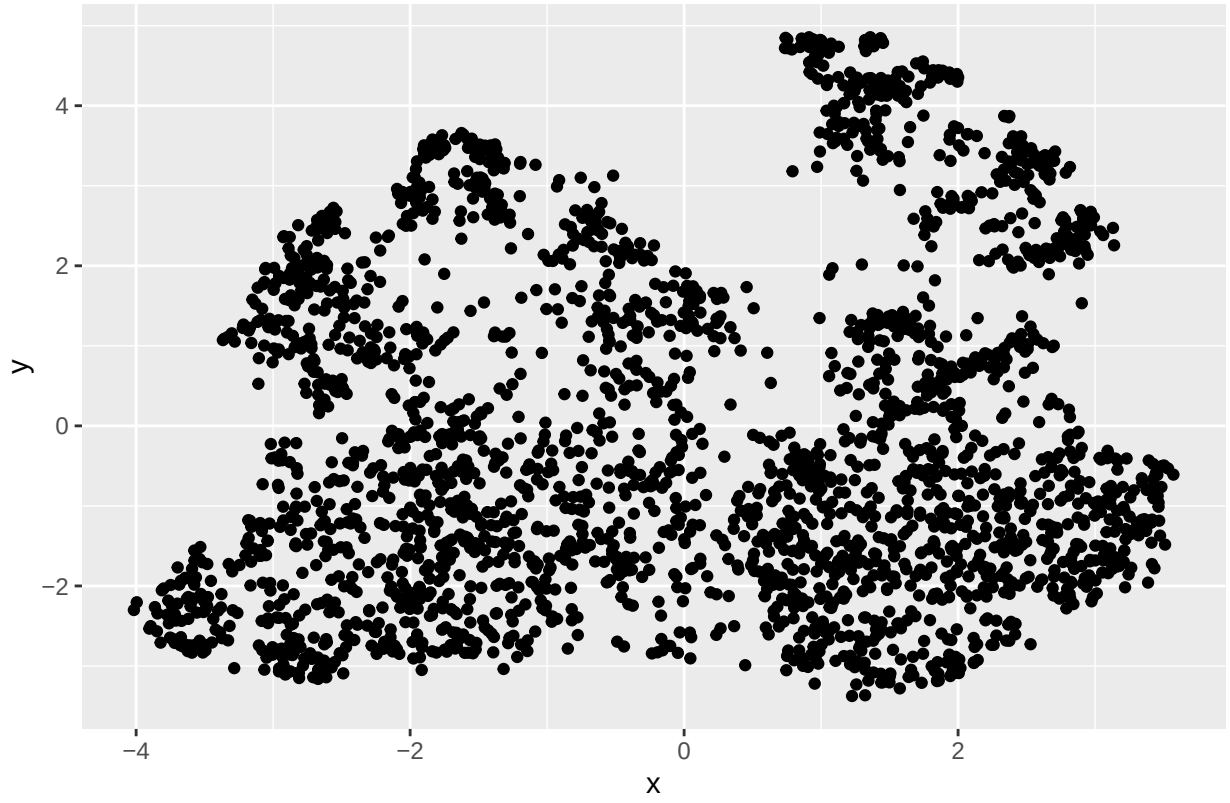
Conclusion: Based on the Scree Plot and our project goal, 3 clusters are selected to segment the patients population.

Using the "Elbow" method above, choosing either k = 3, or k = 5 is ideal. Since the dataset is not large (2611 rows), when K increases, clusters might get too thin, which results in a drop in performance, as shown above. Again, whether to choose a K of 3 or 5 is totally up to the goal of the project. Here I will select K = 3, because it gives me good modeling performance while keeping the sizes of clusters meaningful.

4.2.2 Build and Visualize Clusters

Having determined the optimal number of clusters for this analysis, I can now run the K-Modes algorithm with K = 3, and visualize the clusters. To perform K-Modes clustering, I will use the klaR package. Visualization of clusters can be performed in 2 steps:

- **Step 1:** Since the dataset used for modeling contains more than 50 feature variables (dimensions), dimension reduction is needed to plot the results on a 2-D plane. Here I choose to use the UMAP algorithm for dimension reduction, as it's fast and also supports non-linear dimension reduction. Here is the output:



Patient Segmentation Using BAM Survey: 2D Projection

- **Step 2:** Once all patients are represented on a 2-D plane, they can then be assigned to a cluster based on the K-Modes output. Each cluster will be represented in a different color. Here is the visualization output:

Conclusion: 3 Customer Segments identified using 2 algorithms

Although there is some noise, overall the clusters are fairly present. To understand if these clusters contain any useful insights, I'll need to further validate the clusters by adding back in the external known labels — the previously held out data (churn-rate, satisfaction, living situation, and employment status).

5. Validation

The purpose of the validation process is to see if there are clear differences of patient churn rates among the clusters, and if so, evaluate what factors (labels) might be the cause(s).

5.1 Patient Churn Rate

The table below suggests that cluster 1 and cluster 2 have similar churn rates, 15%, and 16% respectively. However, cluster 3 has a churn rate of 22%, which is 47% higher than that of cluster 1. The significance in difference warrants further analysis of the clusters.

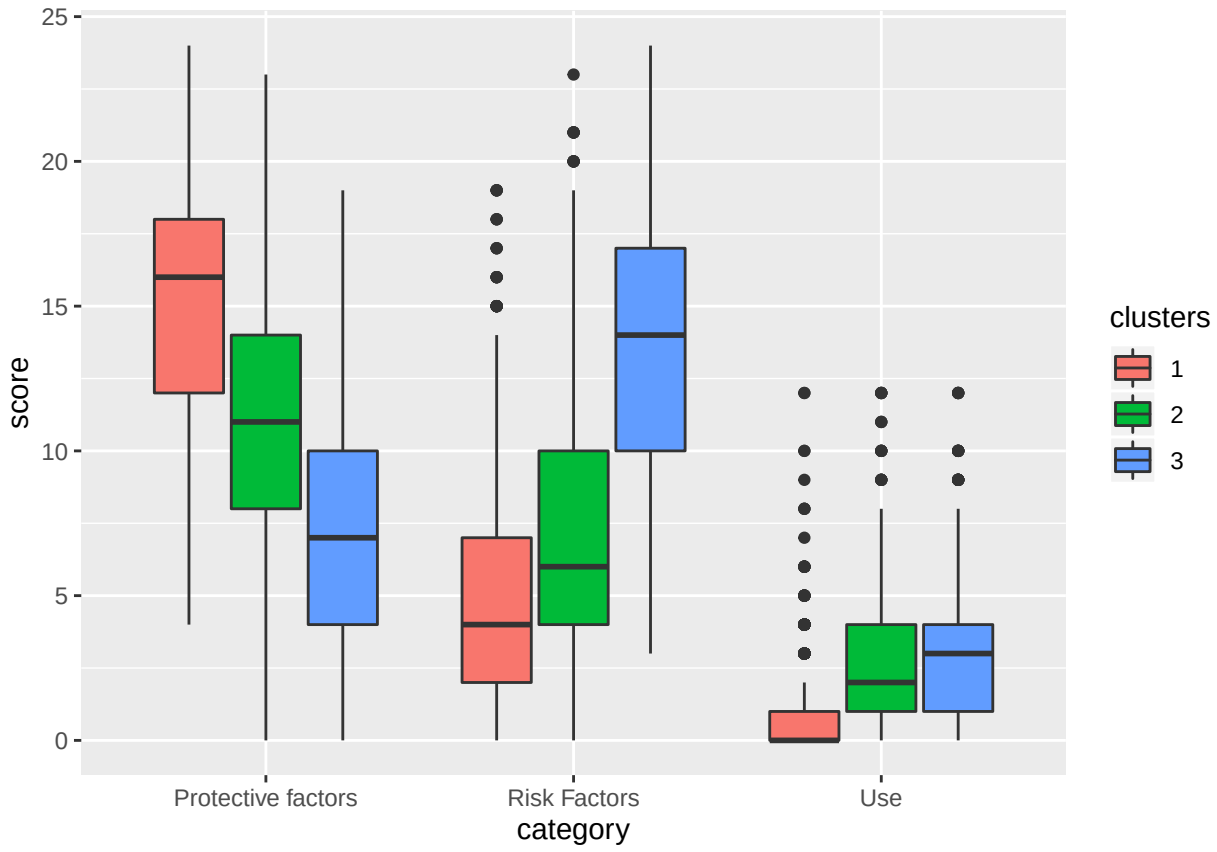
```
## # A tibble: 3 x 2
##   clusters churn_rate
##   <fct>         <dbl>
## 1 1              0.15
## 2 2              0.16
## 3 3              0.22
```

To continue, I will examine if the three clusters have any distinct patterns in terms of survey responses, treatment satisfaction, living situation, and employment status. Again, the latter three labels were held out. This analysis will also be very helpful for determining which features are significant for future modeling.

Since cluster 1 and cluster 2 have similar patients churn rates, using both clusters is helpful, as we can cross-validate to see if both clusters also share similar patterns, especially compared with cluster 3.

5.2 Survey Responses

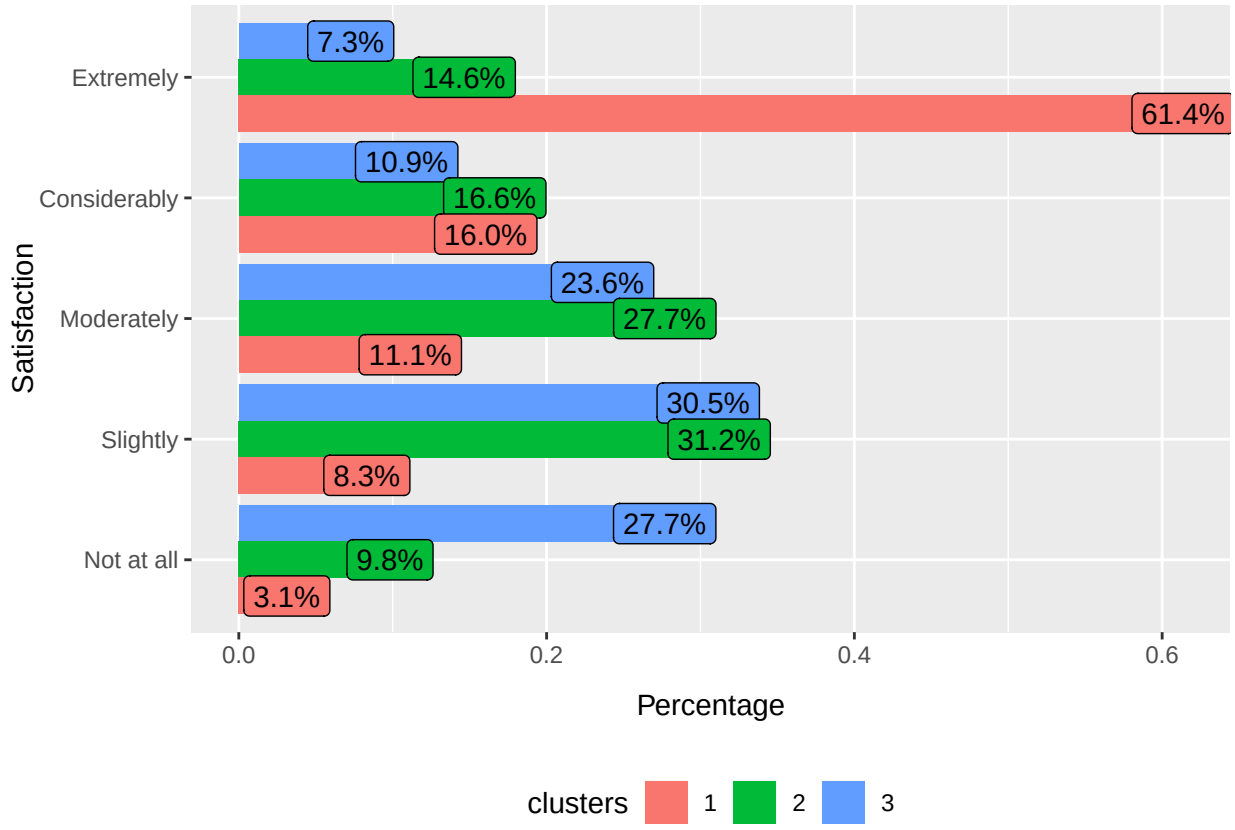
Based on the official survey scoring guideline, responses are grouped into three categories: Protective, Risk, and (drug) Use. Here is the visualization of the three categories among the clusters:



Key Differences:

- Based on the official scoring guideline, patients of both cluster 1 (churn rate - 15%) and cluster 2 (churn rate - 16%) clearly have much higher "Protective Factor", and lower "Risk Factor", and lower "Drug Use", even with outliers included.
- The scoring gap indicates that the clustering results are significant, and it's meaningful to carry out further analysis and examine how each cluster responded to the custom added questions (treatment satisfaction, living situation, and employment status).

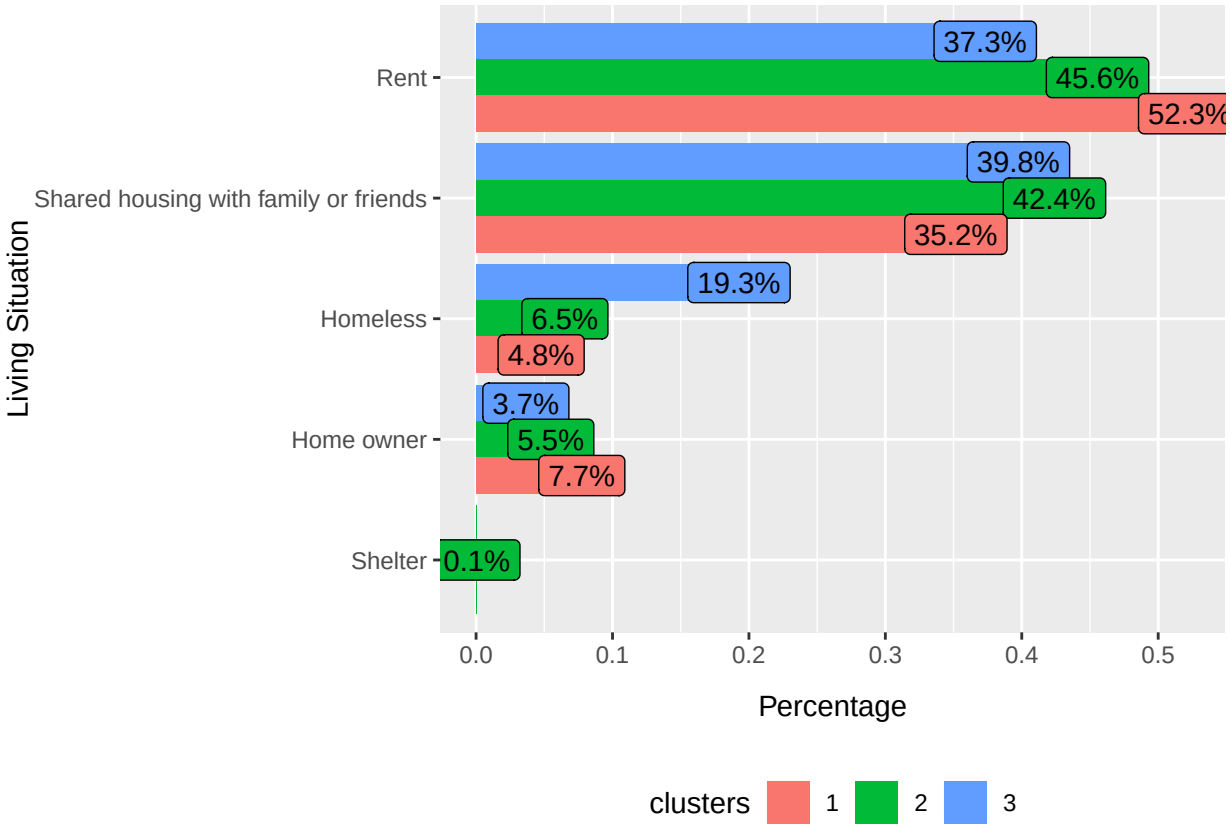
5.3 Treatment Satisfaction



Key Differences:

- Patients of both cluster 1 and cluster 2 are clearly much more satisfied overall with their treatment than patients of cluster 3, especially among those who responded "Extremely".
- On the other hand, cluster 3 has three to four times more patients who reported "not at all" than those of cluster 1 or cluster 2.
- In conclusion, throughout all levels of satisfaction ratings, cluster 3 overall provides much more negative feedback than cluster 1 or cluster 2 does. Therefore, satisfaction rating can be a useful feature in determining whether a patient is more likely to leave the program.

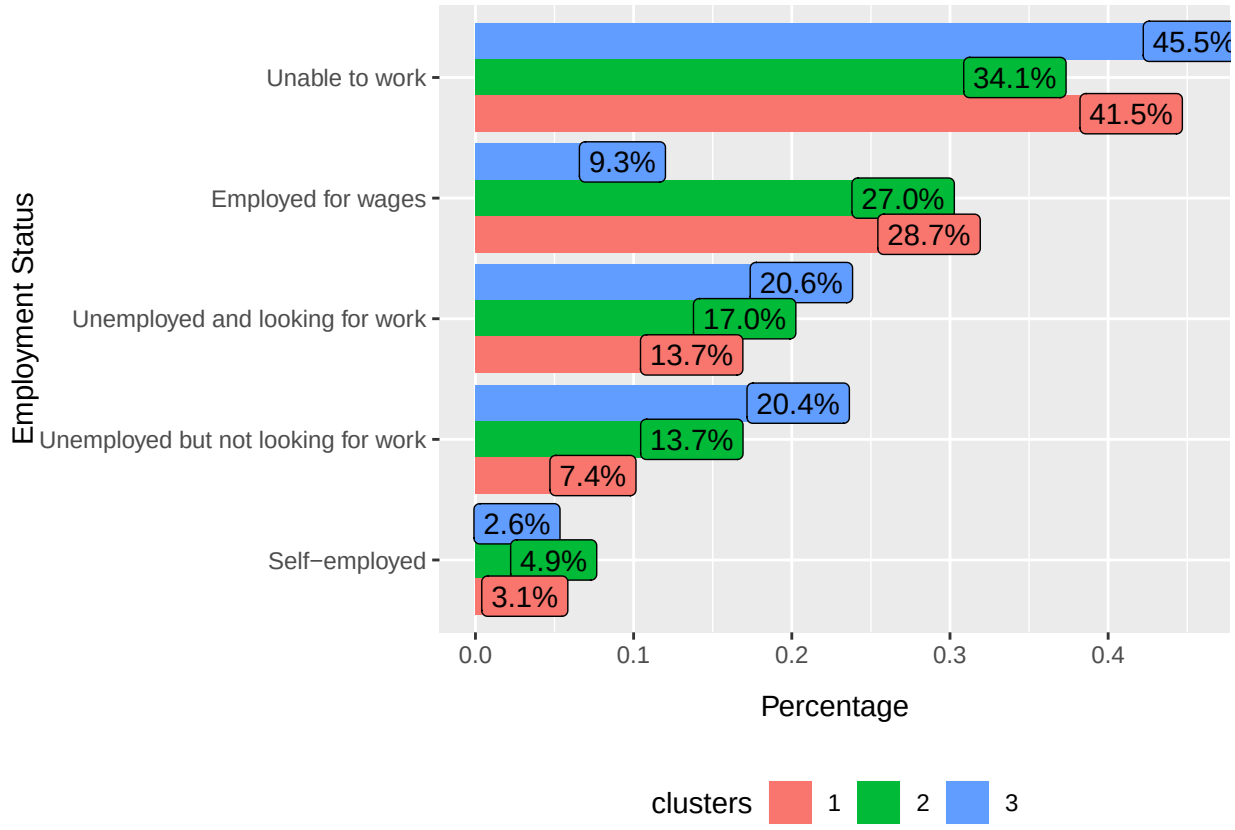
5.4 Living Situation



Key Differences:

- Cluster 3 has a much higher homeless percentage, which could indicate that a lack of stable places to live drives these patients to move from place to place, resulting in higher percentage of patients leaving the program.
- Cluster 1 has the highest percentage of patients renting and owning a place to live. This could indicate that cluster 1 has more income or overall more stable means of earning income.
- In conclusion, "Rent", "Homeless", "Home Owner" contribute most to the variance of patients' living situation. This is useful to know for future modeling.

5.5 Employment Status



Key Differences:

- It's clear that cluster 3 has the highest percentage of patients who are unable to work, and the lowest percentage of patients who are employed for wages. This could indicate that income stability plays a key role in whether a patient would leave the program or not.
- Interestingly, cluster 1 has a significantly lower percentage of patients who are unemployed and looking for work, which needs to be further looked into. Maybe this is related to the level of income or types of work they perform.
- In conclusion, "unable to work", "Employed for Wages", and "Looking for work" contribute most to the variance of patients' employment status. This is useful to know for future modeling.

6. Impact

- Increase Treatment Success

Clustering patients into distinct groups based on their habits, behaviors, living environment, treatment satisfaction, etc., can aid our clinicians to better understand issues that affect patients and improve treatment outcomes and effectiveness, such as personalized treatment, lower relapse rates, and higher life quality.

- Reduce Wastes and Costs

This patients segmentation analysis can equip our healthcare stakeholders to make much more informed decisions. It will also help our senior management team make more intelligent and focused use of resources, which otherwise could be largely wasted due to a lack of clear strategy or direction. For example, with this analysis, our teams will be able to identify which patients need stable places to live, so that we can launch dedicated programs to help them, and at the same time, keep track of treatment results to measure any improvement.

- Improve Retension and Admission Rate

One of the most important ways to facilitate patients to overcome drug addiction is to help them receive professional treatment. Our business teams can use the analysis findings to organize focused efforts for admitting, retaining, and engaging patients. For example, based on clusters characteristics, patients can be introduced to specific programs, rather than to all programs, so that relevancy, patients' interests, and treatment experience are enhanced.

7. What's next?

With attitudinal and behavioral data rather than only demographics information, this analysis provided our clinicians with a deeper understanding of common characteristics and behavior patterns shared by certain groups of patients. However, this is only the beginning of diving deeper and discovering more impactful insights. Here are a few more ideas to try:

- In this analysis, after conducting K-Mode clustering, I've only compared clusters using "Employment Status", "Living Situation", and "Satisfaction". However, there are many other important features we can use to further understand the characteristics of each feature, such as the amount of counseling, treatment consistency, treatment progress, etc. After extracting the most relevant features, we can then build visual applications, for instance, an interactive dashboard, to guide our counselors to help each patient where he/she needs the most.
- I can further extend this analysis by including more relevant features into the clustering algorithm, making this model more robust and accurate to this specific patient population that we serve. I can then test model performance via internal, external, or relative cluster validation. I can also try other clustering methods, for instance, the ROCK algorithm, and compare findings and performance.
- This unsupervised learning analysis helped me identify which features are likely to contribute most to patients' churn rate. Building on top of it, I can try to include these features to develop a binary classification model for predicting the likelihood of a patient quitting the rehabilitation program, and identifying the causes at an individual level. By doing so, not only can we discover areas for improvement, but equally importantly, we can prevent patients from giving up treatment and reduce the probability of relapses.
- Of course, we can always modify the existing surveys to exclude questions that don't contribute to clustering patients, such as questions that are highly correlated and give us essentially the same information. Or replace these questions with questions that focus on other aspects.

Finishing Thoughts

As a data scientist, working on this project has been a rich and fulfilling experience for me, especially when building, developing and owning the entire project from beginning to end: accumulate domain knowledge, understand business needs, create research ideas, define project scope, implement data science techniques, synthesize and communicate findings, and suggest a plan of actions.

Data science is a vast field that combines mathematics, statistics, programming, problem-solving, business understanding, creativity, and much more. This is why data science is so powerful and fascinating, as it can be applied in any domain to make a big positive impact. At Turning Point Clinic, we embrace data-driven applications and data-informed treatment. By leveraging data and technologies, we firmly believe we will continue to provide personalized and outcome-driven professional care. To learn more about our efforts and support our cause, please go to turningpointclinic.org.