Clustering Techniques – D212

Gooden, Nina S. Gooden [ID #: 009823504]

Clustering Techniques – D212

This analysis will explore the medical data of a theoretical real-world organization. By creating this clustering evaluation, I will address the business’ concern about readmission in a data-first evaluation. I will also provide visualizations to support my assessments, in tandem with code for my models. As previously requested, I will also discuss the limitations and potential course of action this data supports.

## **A. The Research Question:**

1. Research Question
   1. Using the *k*-means method of classification, can we identify unseen groupings between patient observation variables such as Age, Income, Initial\_days, Zip, and Additional\_charges?
2. Data Analysis Goals and Objective
   1. I will be continuing my assessment of the aforementioned variables. By looking at these variables, I will evaluate whether there is an unseen grouping that will allow the medical center’s decision makers to better target and segment long-term, cost-effective campaigns.

## **B. Justification for Technique**

1. *K-*means clustering is a centroid-based technique that tries to minimize the distance of the data points from a central point—the centroid. By doing do, this method can then measure the existing data points and find the ones that have the least distance between the generated centroids and the data points, thus creating groups or associations with each centroid.
2. *K*-means assumption summary *(Nagar, 2021)*:
   1. Assumes data can and will be grouped in a spherical shape.
   2. Assumes the size and densities of the clusters will be similar.
3. Packages and libraries:
   1. I’ve opted to use Python for my analysis as this is the language, I am most comfortable with. Python offers packages and libraries that make visualizations and analysis easy and straightforward. I imported pandas, Numpy, Matplotlib, and Seaborn in this evaluation. In addition, I will also be using the scipy.cluster.vq library, with specific imports of kmeans, vq, and whiten. This package is uniquely suited for clustering and normalizing data.

## **C. Data Preperation**

1. Preprocessing goal
   1. My first data processing goal is to ensure my data is as clean as possible and ready for clustering. As such, I needed to make sure my data was numeric, as well as not obviously correlated with anything else in the DataFrame. I also needed to normalize my data before jumping into the *k-*means clustering.

A screenshot of a computer

Description automatically generated with medium confidence

1. I created a DataFrame with only the columns I would be using in my evaluation. I also ran a few quick tests to make sure the variables were acceptable.
   1. All of my variables for this assessment are continuous. Age, Income, Initial\_days, Zip, and Additional\_charges.

A screenshot of a computer

Description automatically generated with medium confidence

1. Data preparation steps:
   1. Normalizing data and storing it in a new DataFrame.

A screenshot of a computer

Description automatically generated with medium confidence

* 1. I created two separate graphs for this analysis because it was so difficult to read. I was already concerned about the viability of my content, due to the variance in the data.

|  |  |
| --- | --- |
| Graphical user interface  Description automatically generated | Graphical user interface  Description automatically generated |

* 1. Finally, I checked the standard deviations and the .describe() method to make sure my data had been sufficiently normalized. I noted that Age still has a considerably higher average than the other variables.

A screenshot of a computer

Description automatically generated

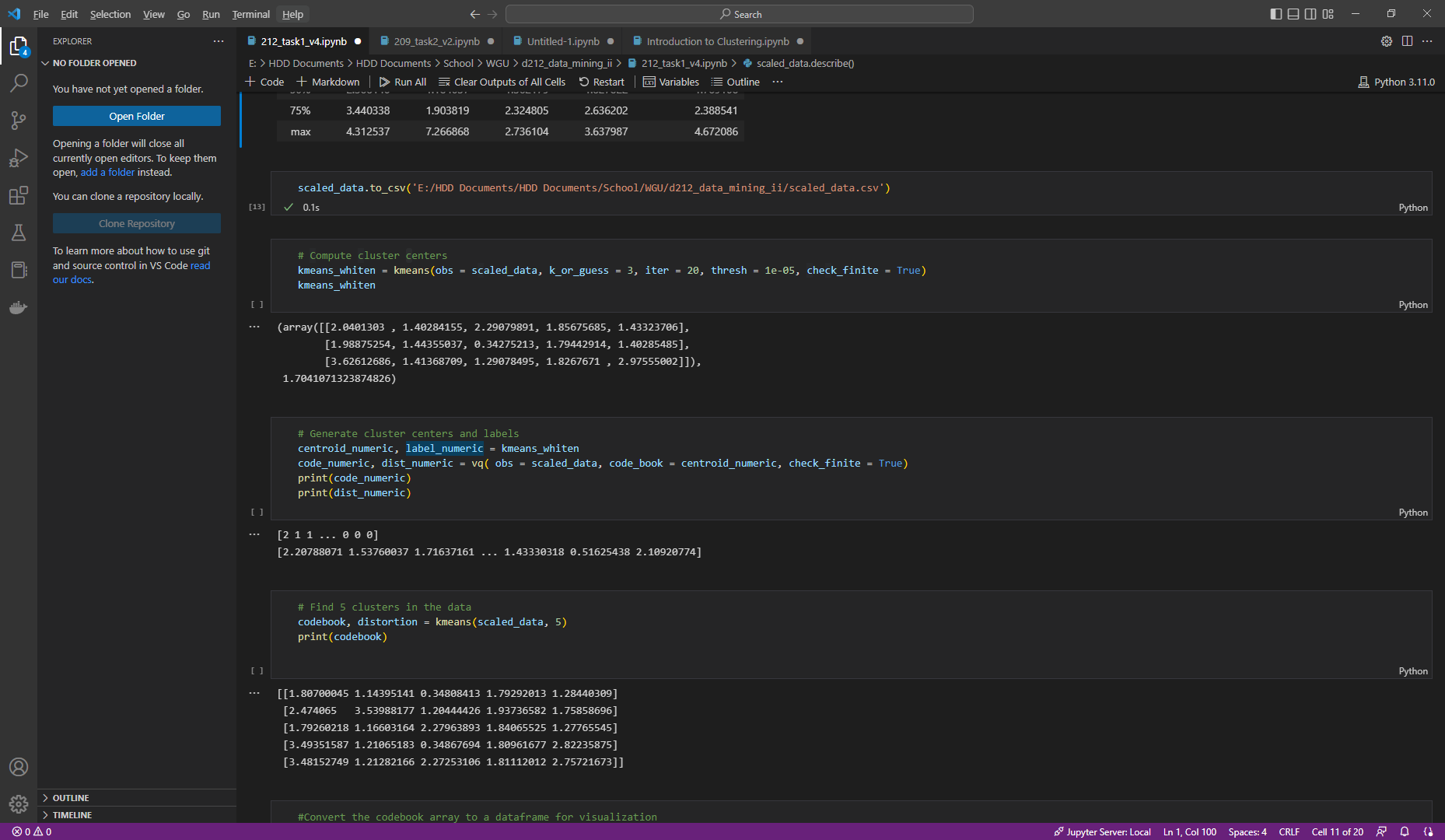
1. Cleaned dataset can be found in my submission.

A screenshot of a computer

Description automatically generated with medium confidence

## **D. ANALYSIS**

1. Analysis technique used:
   1. I first used the *k*-means method to partition the data into 3 groups. At this point, I thought three would be more than enough since I only had 5 variables.



* 1. After I generated the centers, I generated the rest of the cluster centers, as well as the labels that I would use for the visualization. I also opted to print the code and distribution values.

A screenshot of a computer

Description automatically generated with medium confidence

* 1. My next step was to isolate the code\_book variable so that I could convert it into a Dataframe. This was necessary to create the scatterplot visualization.
     1. This was also the point where I realized my assessment would likely not end with comfortable groups. We want the values in the code\_book to be as low as possible and several were over 2. A handful were also over 3.

A screenshot of a computer

Description automatically generated with medium confidence

* 1. I created my visualization and noted that the groupings were all over the place. The visualization confirmed my concern that I would not have discernable groups.

A screenshot of a computer

Description automatically generated with medium confidence

* 1. My last step was to use the elbow method to check my assessment. As expected, the resulting visualization is a fairly smooth curve, and it’s unclear what would have been the best k value to choose. This is because the data is not largely clustered.

A screenshot of a computer

Description automatically generated

1. All code has been provided with detailed steps and screenshots.

## **E. Data SUMMARY AND IMPLICATIONS**

1. Explaining the accuracy of your clustering technique:
   1. Ultimately, this was not a successful attempt at grouping. This makes sense, as the dataset is grouped very clearly in the instructions and further by the work we accomplished in D211. I knew going in that I had selected variables that would not likely result in grouping, as they are not even remotely related to one another, so I feel this was a satisfactory test..
2. Discuss the results and implications of the analysis:
   1. According to the elbow method graph, this was not a successful clustering attempt. I’ve reviewed the steps I took and am confident that the work is correct—the chosen variables simply did not reflect unseeing grouping.
3. Data analysis limitations:
   1. *k-*means evaluations do not provide a clear answer to the question “how many clusters should I use?” As such, there is the need to use outside sources to attempt to guesstimate this answer—ergo, the elbow method. However, this is not a perfect solution, as we’ve seen in my output..
4. Recommendations:
   1. This analysis should be rerun with changes to the explored variables in order to better understand potential groupings.
   2. This medical center should spend resources to house data in a more way that better designates actionable information for evaluation.
   3. In order to maintain the health of this analysis, the medical center should launch campaigns to gather as much additional data as possible. I would also recommend keeping track of the variation KPIs in order to segment response efforts.
   4. This analysis should be run quarterly in order to ensure the value of this assessment.

## **F. Panopto**

[Data Mining II – OFM3](https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=badaea64-888f-4d48-bcee-af6700593c4c)

## **G. Reference web sources**

*Round elements of the array to the nearest integer in Numpy*. (n.d.). https://www.tutorialspoint.com/round-elements-of-the-array-to-the-nearest-integer-in-numpy/

VanderPlas, J. (n.d.). *In Depth: k-Means Clustering | Python Data Science Handbook*. https://jakevdp.github.io/PythonDataScienceHandbook/05.11-k-means.html

*only integers, slices (`:`), ellipsis (`. . .`), numpy.newaxis (`None`) and integer or boolean arrays are valid indices*. (2016, January 22). Stack Overflow. https://stackoverflow.com/questions/34952651/only-integers-slices-ellipsis-numpy-newaxis-none-and-intege

Real Python. (2022, September 1). *K-Means Clustering in Python: A Practical Guide*. https://realpython.com/k-means-clustering-python/

## **H. Acknowledge sources**

Sharma, P. (2022, June 15). *The Most Comprehensive Guide to K-Means Clustering You’ll Ever Need*. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2019/08/comprehensive-guide-k-means-clustering/

Nagar, A. (2021, December 13). *K-means Clustering — Everything you need to know - Analytics Vidhya*. Medium. https://medium.com/analytics-vidhya/k-means-clustering-everything-you-need-to-know-175dd01766d5