Clustering Techniques – D212

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

Dr. Kesselly Kamara

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 clustering, can we identify a simple way to group readmitted patients based on key numerical identifying variables?**
2. Data Analysis Goals and Objective
   1. **I will be using the *k-*means clustering method on patient data to group those that were readmitted to the medical center. My analysis will only look at patients with a ReAdmis value of “Yes.” I am confident that clustering will be possible, given the given variables.**

## **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.
   1. **Expected outcomes for this analysis are: the *k-*means clustering technique will identify actionable groups that can be utilized by hospital stakeholders, analyzed in more depth, or passed as a feature in a different algorithm for further examination.**
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.
   2. **Upon further evaluation, I added sklearn, specifically KMeans from the .cluster library, and KneeLocator, from the kneed library to complete my evaluation. I did so to better build visualizations.**

## **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 remove non-relevant and categorical data. In addition, my research question specifically asks about readmitted patients, so I used the .drop function to remove all data that did not align.**

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, 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. I also ran a quick .dtypes to make sure everything was as expected.

A screenshot of a computer

Description automatically generated with medium confidence

* 1. **I created plots, using .pairplot to look for potential data combinations/concerns. I noted that several of the variables had outliers but was not concerned with them.**

Graphical user interface

Description automatically generated

Diagram

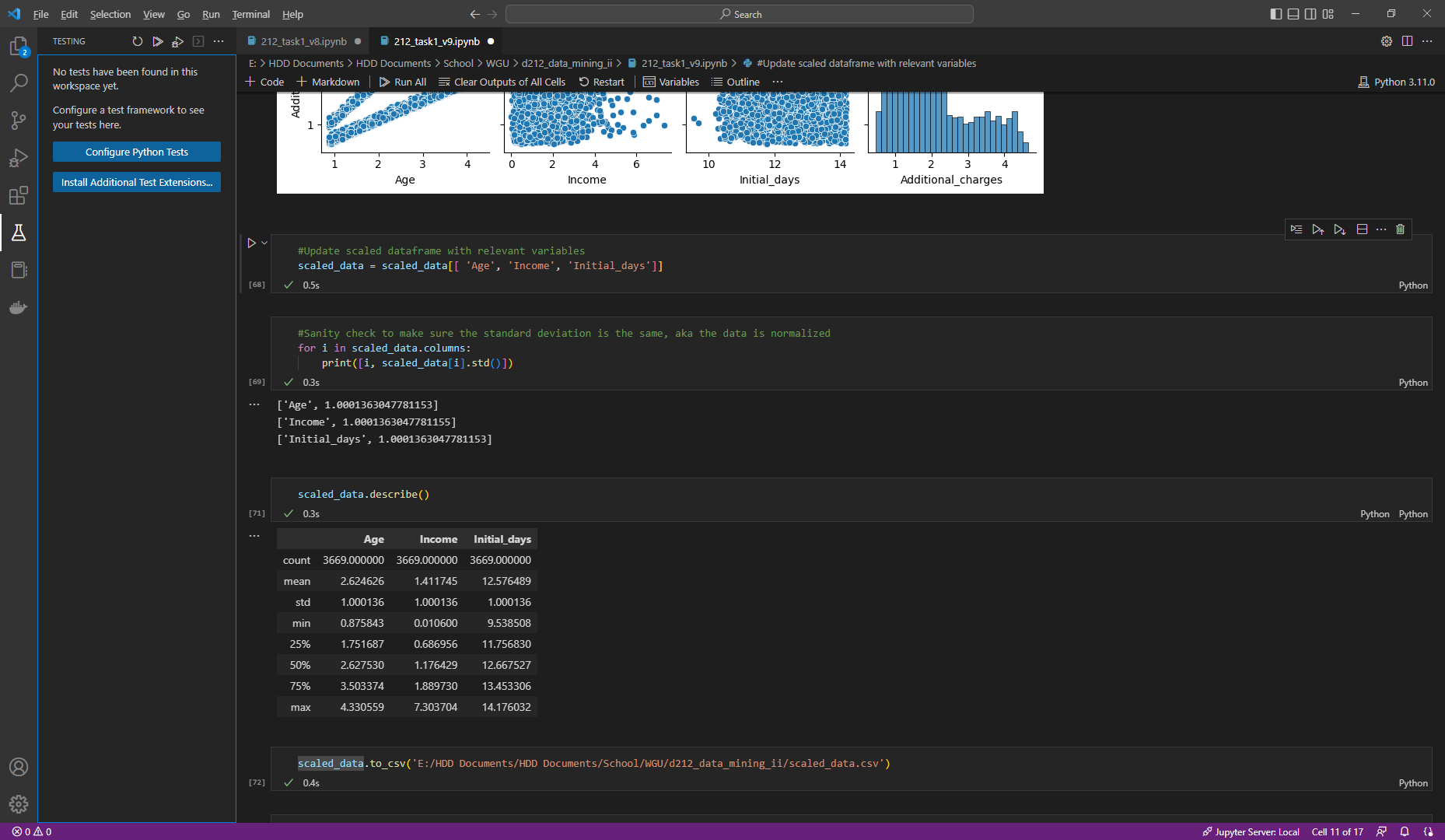
Description automatically generated

* 1. **I dropped the Additional\_charges column due to the correlation between Age—despite knowing that correlation is not a considerable factor in k-means. I did so for clarity in my visuals. This was previously marked incorrect; however, I hope explaining the why behind my choice clarifies. I didn’t think I had to remove this variable, I did it because leaving it made the final scatterplot confusing and harder to read.**

A screenshot of a computer

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 **Initial\_days has a considerably higher average than the other variables.**



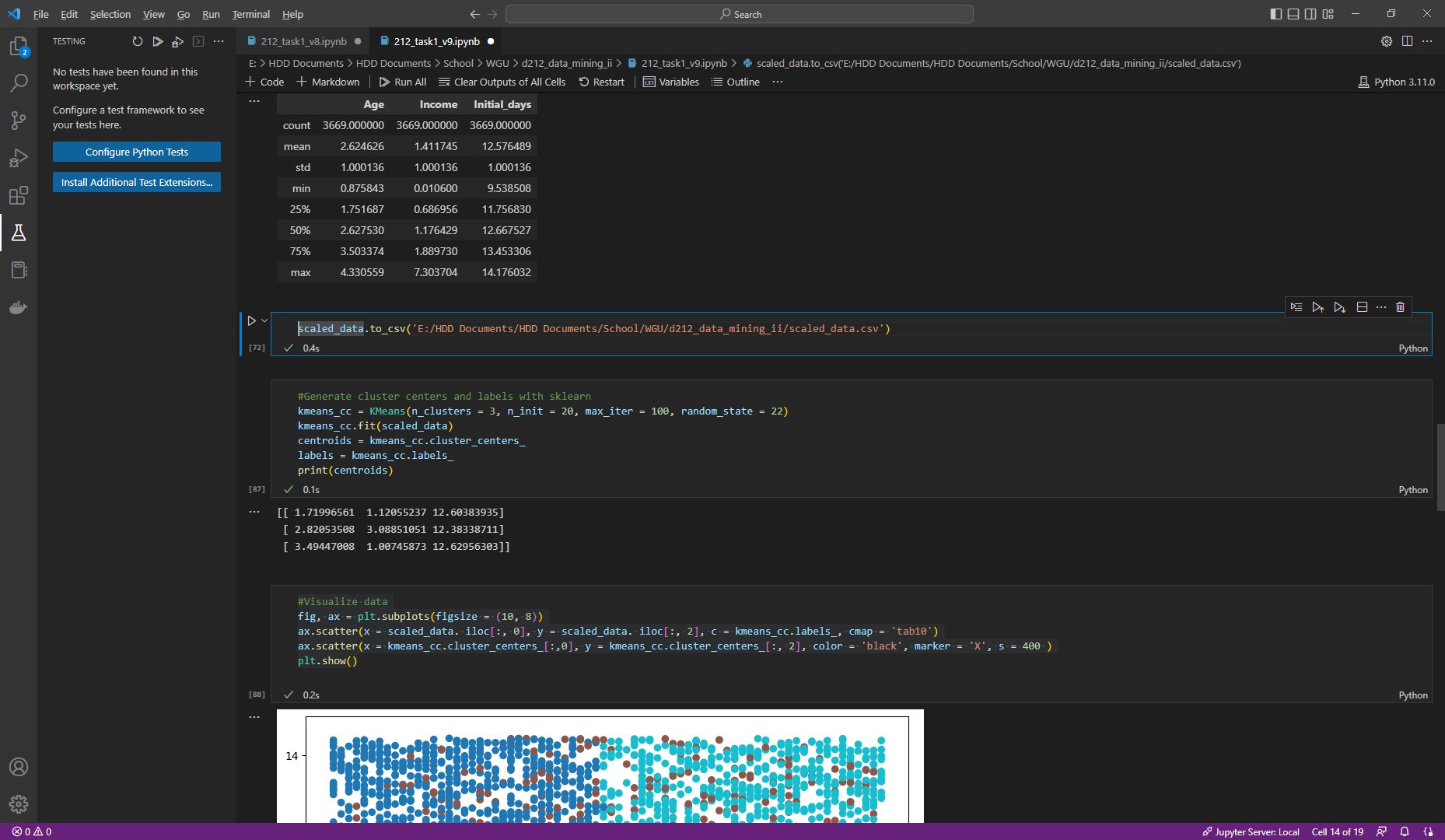
1. Cleaned dataset can be found in my submission.

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 and generate the labels. I switched to the sklearn version of the method because it was better documented. I selected 22 as my random state for reproducing.**



* 1. **After I had my centroids and labels, I was able to plot the data.**

A screenshot of a computer

Description automatically generated with medium confidence

* 1. **I used KneeLocator and the adjacent packages in order to generate a visualization of for the optimum point.**

Graphical user interface

Description automatically generated

* 1. **Based on the Elbow graph, I regenerated the previous assessment with the optimum value for n\_clusters. It appears that there are 3 distinct groupings and a single, overlapping group.**

Graphical user interface, qr code

Description automatically generated

Chart, scatter chart

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. ***K-*means requires only a few steps. I normalized the data, identified the centroids and labels, and repeated the variance calculation 20 times. I’ve reviewed the steps I took and am confident that the work is correct—the k-means algorithm returned an expected model.**
2. Discuss the results and implications of the analysis:
   1. **My final analysis showed three discernable groups, as well as one overlapping group, which is to be expected, since “most real-world medical datasets have inherently overlapping information, which could be best explained by overlapping clustering methods that allow one sample belong to more than one cluster.”** (Khanmohammadi, Adibeig, & Shanehbandy, 2017) ***K-*means is one such clustering method.**
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.
4. Recommendations:
   1. **The model showed clear patient groups. This is important as clustering algorithms can find patterns across patients that are difficult for medical practitioners to find. (Alashwal, Halaby, , Crouse, Abdalla, & Moustafa, 2019)**
   2. **Now that clusters have been identified, hospital centers should run additional analysis in order to create campaigns that will improve the readmission rates of each impacted group.**
   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 to target segmented 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=74442cb3-a349-46da-81d6-af6a016a0325)

## **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/

Kumar, V. (2022, February 26). Tutorial for K Means Clustering in Python Sklearn. MLK - Machine Learning Knowledge. https://machinelearningknowledge.ai/tutorial-for-k-means-clustering-in-python-sklearn/

*matplotlib.axes.Axes.axvline — Matplotlib 3.6.2 documentation*. (n.d.). https://matplotlib.org/stable/api/\_as\_gen/matplotlib.axes.Axes.axvline.html

Kleine, D. (2022, January 6). Detecting knee- / elbow points in a graph - Towards Data Science. Medium. https://towardsdatascience.com/detecting-knee-elbow-points-in-a-graph-d13fc517a63c

*Choosing Colormaps in Matplotlib — Matplotlib 3.6.2 documentation*. (n.d.). https://matplotlib.org/stable/tutorials/colors/colormaps.html

*seaborn.boxplot — seaborn 0.12.1 documentation*. (n.d.). https://seaborn.pydata.org/generated/seaborn.boxplot.html

## **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

Alashwal, H., El Halaby, M., Crouse, J. J., Abdalla, A., & Moustafa, A. A. (2019). The Application of Unsupervised Clustering Methods to Alzheimer's Disease. Frontiers in computational neuroscience, 13, 31. https://doi.org/10.3389/fncom.2019.00031

*Example of K-Means Clustering in Python*. (n.d.). https://datatofish.com/k-means-clustering-python/

*scipy.cluster.vq.vq — SciPy v1.9.3 Manual*. (n.d.). https://docs.scipy.org/doc/scipy/reference/generated/scipy.cluster.vq.vq.html