­

**Data Mining II**

Jason Willis

College of Information Technology,

Western Governors University

Dr. Kesselly Kamara

June 30th, 2022

Table of Contents for Each Rubric

Part I: Research Question

[Describe Purpose, Summarize Research Question and Define Objectives:](#RubricA) 3

***Part II: Technique Justification***

Summarize Clustering Assumptions and List Python Libraries w/Justifications7

***Part III: Data Preparation***

[Describe](#RubricC4) Data Preprocessing, Summary Statistics, Visualizations and Code 8

Part IV: Analysis

Split Data Into Training/Test Data Sets w/Provided Files 15

Describe Analysis Technique w/Screenshots15

Provide Code Used for Clustering Analysis from Previous Step15

Part V: Data Summary and Implications

Compare Accuracy of Clustering Technique 16

Discuss Results and Implications of the Clustering Analysis 16

Discuss ONE Limitation of the Data Analysis 16

Recommend a Course of Action on Hypothesis and Implications 16

Part VI: Demonstration

[Video](#RubricF) 17

[Sources for Third-Party Code](#RubricG) 17

[Sources](#RubricH) 18

# Hospital Readmission Problem

For our chain of hospitals to lower readmission concerns, we need to identify patients who have increased risk of rehospitalization within a month of their release. According to Schuller (2020), non-obese adults were 21% less likely to be readmitted than obese adults. A readmission study by Gert, et. al. (2002) showed a correlation between longer initial hospital stays and readmission. Within the provided dataset, I’m leveraging these studies to help create my hypothetical question and shape my approach in finding potential patient groups with a statistically significant chance for readmission outcomes.

After viewing the provided medical\_clean.csv data set and accompanying data dictionary, there seems to be some patient groupings which are aligned with the research mentioned above. For instance, the following patient data fields: Initial patient admin days, Total Charges, and Initial Says (inpatient) both caught my attention and were underscored by the research mentioned above. While my initial feelings towards these variables might make them feel related, are they?

## A1 – Proposal of Question

From information about previous patients who were readmitted, can we predict which patients are likely to be readmitted in the future?

**A2 – Defined Goal**

The goal of our analysis is to logically investigate the provided data set and, with evidence, support or reject the hypothesis. Some data will need to be converted from categorical to numerical data types prior to processing. Our objective is to see how, if at all, any patient’s data correlate with potential for readmission.

**B1 – Explanation of Clustering Technique**

According to Larose (2019) “A *cluster* is a collection of records that are similar to one another and dissimilar to records in other clusters.” Unlike KNN, where the method tries to classify a record, clustering attempts to segment data into uniform subdivisions of the whole data set.

**B2 – Summary of Technique Assumption:**

From the DataCamp lesson “Cluster Analysis in Python” the following K-Means assumptions to consider (Daityari, 2022):

* Bias of analysis towards uniformly sized clusters

**B3 – Packages or Libraries List**

The following Python libraries were used followed by their corresponding reason for use:

* Pandas – Used to import dataset and data analysis tasks.
* Numpy – Used for describing the data set and computing distances in KNN.
* Matplotlib – Used for viewing the testing and actual data as a scatter plot.
* Seaborn – Used for creating a heatmap when looking for null values in the original dataset and ggplot style graph matrix to help visualize univariate data.
* Scipy – Used for statistical techniques
* Sklearn – Used for preprocessing, PCA, model splitting and K-Means tasks.

**C1** – **Data Preprocessing**

A preprocessing goal achieved for this model was standardization, as seen below in Figure 1 using sklearn. Standardization helps by formatting variables on similar scales (subtracting the mean and dividing the standard deviation) as to reduce variables from being a dominant influence to the model.

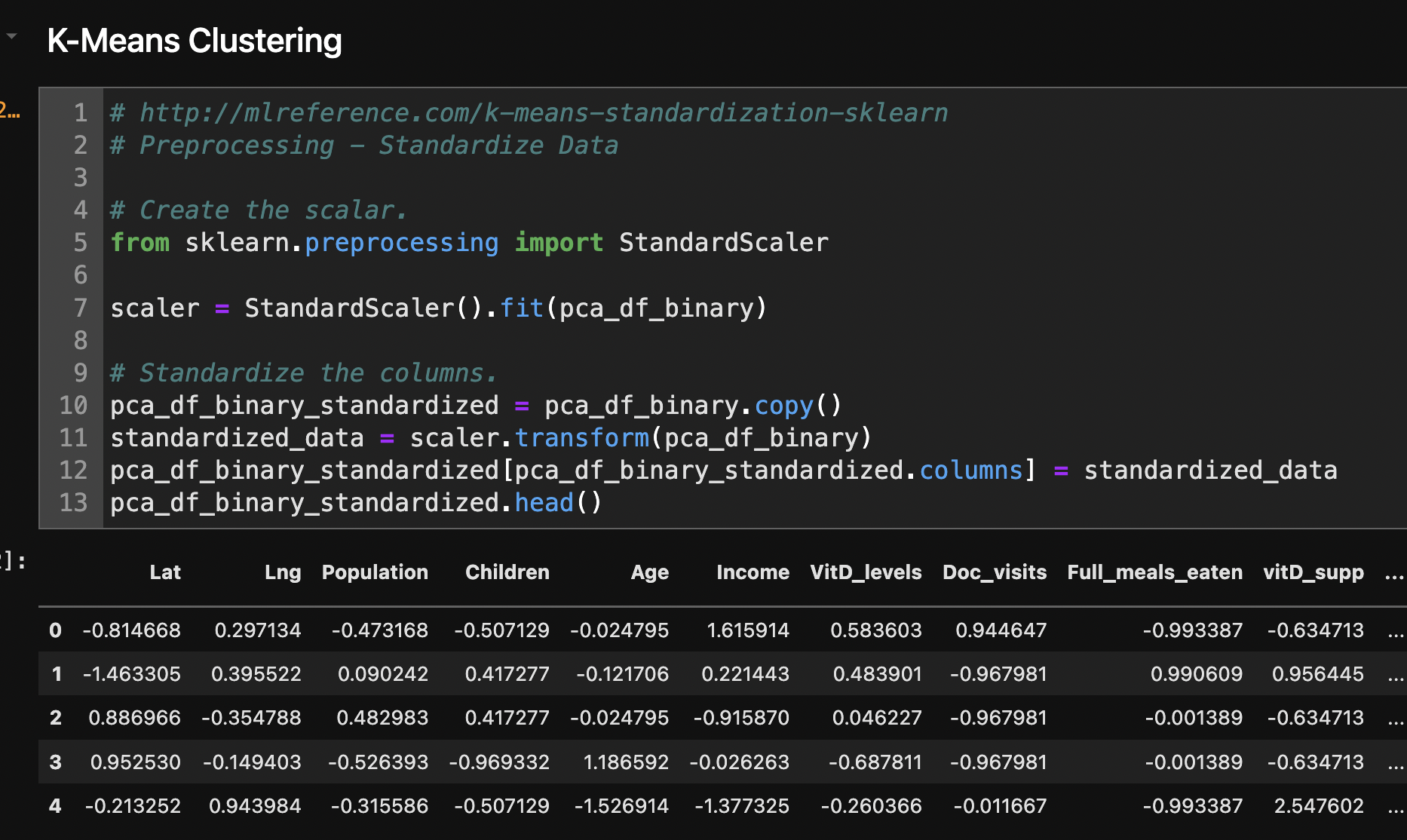


Figure - Preprocessing Goal: Scaling

**C2 – Data Set Variables**

Once preprocessing was completed, to include creating dummy variables (Figure 3), the data as shown in Figure 4, had all numerical data types:



Figure - Data Types

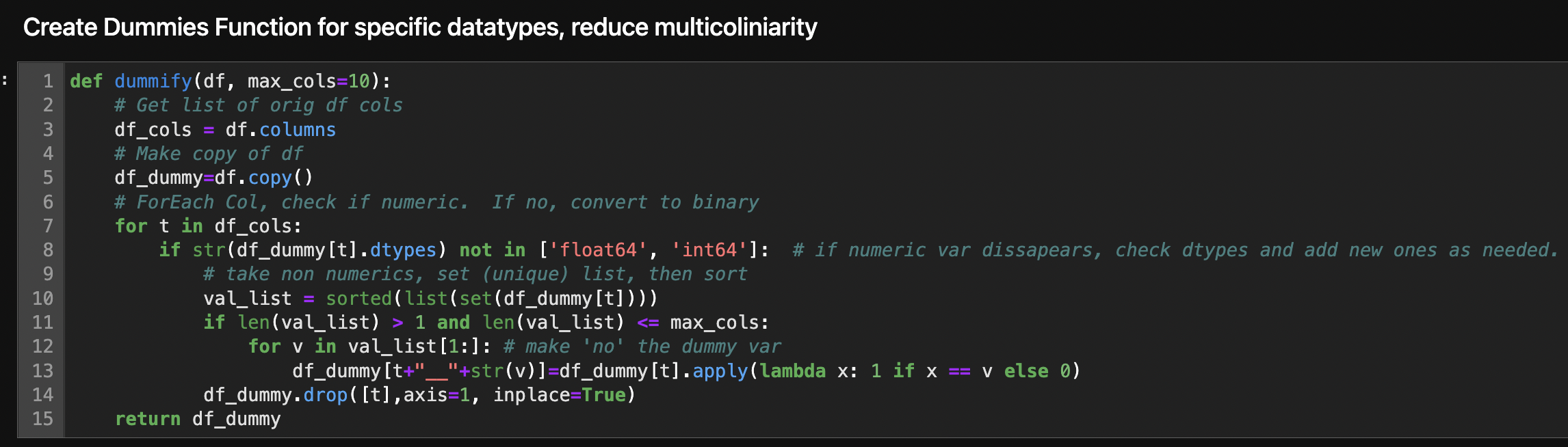


Figure - Create Dummy Data

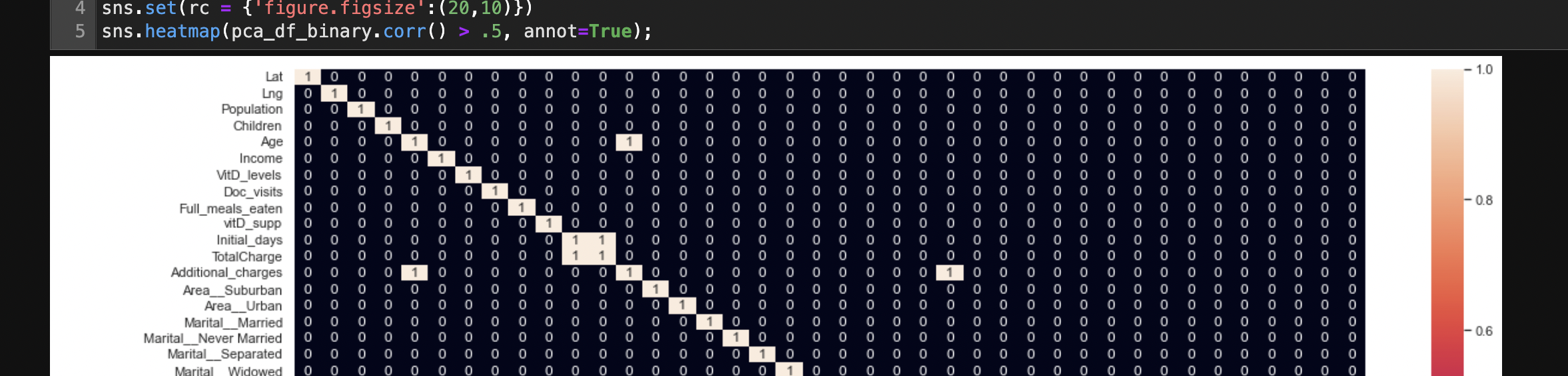
****

Figure - Correlation Matrix Showing Results Above 0.5

**C3 Steps for Analysis**

Initially, the dataset was loaded using pd.read\_csv(‘medical\_clean.csv’) and a data frame was created. Some exploratory data analysis was performed to familiarize myself to the data, look for missing values and view data statistics using df.describe(). Since the target and many predictor variable were initially a categorical data series, I converted this column to integers (0,1) using dummy data. A correlation matrix identified data greater than 0.5 with our target “ReAdmis\_Yes” for classification. Using PCA and experimenting with various *n* components, 3 was used with a shape of 10,000 rows by 3 components and 40 columns.

Next, the pca\_df\_binary data was standardized (Figure 1) and fit to the K-Means clustering technique.

|  |
| --- |
| Text  Description automatically generated |
|  |

Figure - K-Means Clustering – and Data Standardization

Chart, line chart

Description automatically generated

Figure - Scree Diagram

First, the algorithm iterates through a number of cluster means by finding the closest squared distance of each point to a user-defined K(s). Each row is assigned to the nearest cluster mean and the algorithm continues to converge around the K-mean (cluster group) until data stops adjusting. A Scree diagram helped figure out an optimal number of clusters (Figure 6).

Using sklearn’s train-test-split method the standardized data set was run to create training and testing models (Figure 7).

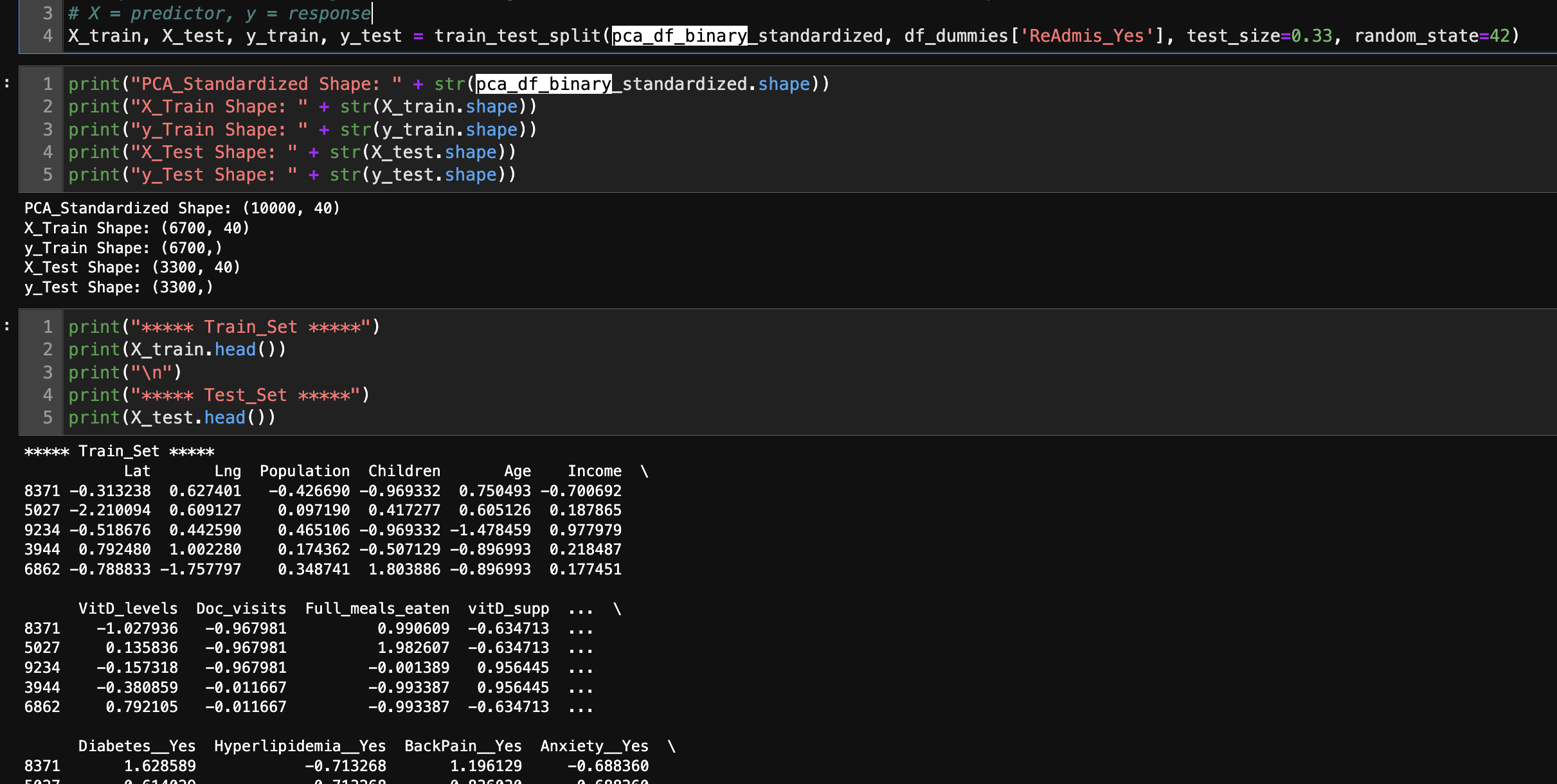


Figure - Create Training and Testing Models

**C4 – Cleaned Data Set**

The cleaned data set was saved to “cleaned\_pca\_df\_binary\_standardized\_plus.csv”.

**D1 – Output and Intermediate Calculations**

Data is standardized to compare variables with a similar scale (Figure 5) and fit to the model, using predicted and actual data from the split (Figure 7). As seen in Figure 8 the predicted and actual cluster data is almost identical.

|  |  |
| --- | --- |
| **Predicted** | **Actual** |
|  |  |

Figure 8 - Plotting Predicted and Actual Data

**D2 – Code Execution**

Code is located in the “D212 - DataMiningII\_PA\_jw.ipynb” document.

**E1 – Accuracy of Clustering Technique**

As seen in Figure 8 the predicted and actual cluster data is almost identical. The cluster centroids are well placed and there is a clear separation between clusters.

**E2 – Results and Implications**

The model is highly accurate. Looking at Figure 5, you can see that each plotted point is a patient from the data set. As the patient’s initial stay is longer and their total charge is higher, the chance that they are readmitted is closer to “Yes”.

Graphical user interface, text

Description automatically generated

Figure 9 – Model Findings

**E3 – Limitation**

One limitation is finding the best number of *k* clusters. Using a scree diagram helps (Figure 6) though this can feel a bit loose and won’t always provide the best number of clusters to input into your model. Meaning, sometimes, you will need to play with the number of *+/- k provided.*

**E4 – Course of Action**

Our model is highly predictive of patient readmissions rates and follows previously analyzed performance assessment outcomes. My recommendation would be to focus on researching and identifying patient’s threshold criteria’s for initial days and total charges as these predictor variables helped predict readmission with a high accuracy rate. By focusing on these predictors, we should be able to predict patients that will have a higher likelihood of readmission within 30 days.

**F – Panopto Demonstration**

Panopto video Will be uploaded once report is finalized.

<https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=6906d4d4-ad26-4449-8cd6-aeba01419d20>

**G – Sources for Third-Party Code**

* Help using Markdown: <https://www.markdownguide.org/basic-syntax/>
* Help to see ALL columns: <https://stackoverflow.com/questions/24524104/pandas-describe-is-not-returning-summary-of-all-columns>
* Help to create a better histogram design: <https://mode.com/example-gallery/python_histogram/>
* Matplotlib Help: <https://matplotlib.org/2.1.2/api/_as_gen/matplotlib.pyplot.plot.html>
* Multiple ways to conduct ANOVA: <https://www.marsja.se/four-ways-to-conduct-one-way-anovas-using-python/>
* Numpy Help: <https://numpy.org/doc/stable/>
* Pandas Help: <https://pandas.pydata.org/docs/user_guide/index.html#user-guide>
* Python Help: <https://docs.python.org/3.9/library/index.html>
* Scipy.stats Help: <https://docs.scipy.org/doc/scipy/reference/tutorial/stats.html>

## References

Boeye, J., et. al., Dimensionality Reductionin Python, (2022). Found At: <https://app.datacamp.com/learn/courses/dimensionality-reduction-in-python>

Gert P Westert, Ronald J Lagoe, Ilmo Keskimäki, Alastair Leyland, Mark Murphy,

An international study of hospital readmissions and related utilization in Europe and the USA, Health Policy, Volume 61, Issue 3, 2002, Pages 269-278, ISSN 0168-8510, <https://doi.org/10.1016/S0168-8510(01)00236-6>. (<https://www.sciencedirect.com/science/article/pii/S0168851001002366>)

Larose, D., C., & Larose, D., T. (2019). Data Science Using Python and R. Wiley. <https://www.wiley.com/en-us/Data+Science+Using+Python+and+R-p-9781119526810>

Paul, S. (2018). K-Means Clustering in Python with Scikit-Learn. *DataCamp*.com. Retrieved From: <https://www.datacamp.com/tutorial/k-means-clustering-python>

Schuller K. A. (2020). Is obesity a risk factor for readmission after acute myocardial infarction?  *Journal of healthcare quality research*, *35*(1), 4–11. <https://doi.org/10.1016/j.jhqr.2019.09.002>

Urbonas K. (2022). Customer Segmentation in Python. Data Preprocessing for K-Means Clustering. *DataCamp.* <https://campus.datacamp.com/courses/customer-segmentation-in-python/data-pre-processing-for-clustering?ex=1>