­

**Data Mining II**

**PA2 Dimensionality Reduction Methods**

Jason Willis

College of Information Technology,

Western Governors University

Dr. Kesselly Kamara

June 25th, 2022

Table of Contents for Each Rubric

Part I: Research Question

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

***Part II: Method Justification***

Explain PCA Analysis & Summarize Assumptions4

***Part III: Data Preparation***

Perform Data Preparation, Identify and Standardize Continuous data 6

Part IV: Perform PCA Analysis

Determine matrix, Identify All Principal Components and Variance, Summarize 8

Part V: Attachments

[Sources for Third-Party Code](#RubricG) 10

[Sources](#RubricH) 11

# 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 ascertain the minimum number of principal variables for our patients?

**A2 – Defined Goal**

The goal of our analysis is to logically investigate the provided data set and, by leveraging principal component analysis and other techniques, reduce the dimensionality of the provided dataset while keeping the model’s accuracy high.

**B1 – Explanation of PCA**

According to Larose (2019) “PCA seeks to account for the correlation structure of a set of predictor variables, using a smaller set of uncorrelated linear combinations of these variables, called components.” The approach aims to reduce the original number of predictor variables while accounting for most predictions in the original set. By reducing the predictor set, redundancies like multicollinearity are removed. Additionally, reductions in predictor values also helps to decrease overfitting.

**B2 – PCA Assumption:**

As stated in the Principal Component Analysis (PCA) explanation above, the technique assumes reduced data set *k* can provide almost indistinguishable predictive results as compared to the original complete data set *m*.

**C1** – **Continuous Dataset Variables**

The data set accurately identifies continuous (and discrete) dataset variables needed to answer the PCA question from part A1 (Figure 1, Figure 2).

Graphical user interface, text

Description automatically generated

Figure - Identifying Continuous and Discrete Variables

A picture containing text

Description automatically generated

Figure - Data Types

**C2 – Standardization of Dataset Variables**

During the preprocessing steps, unnecessary features were removed dummy variables (Figure 3), were created.

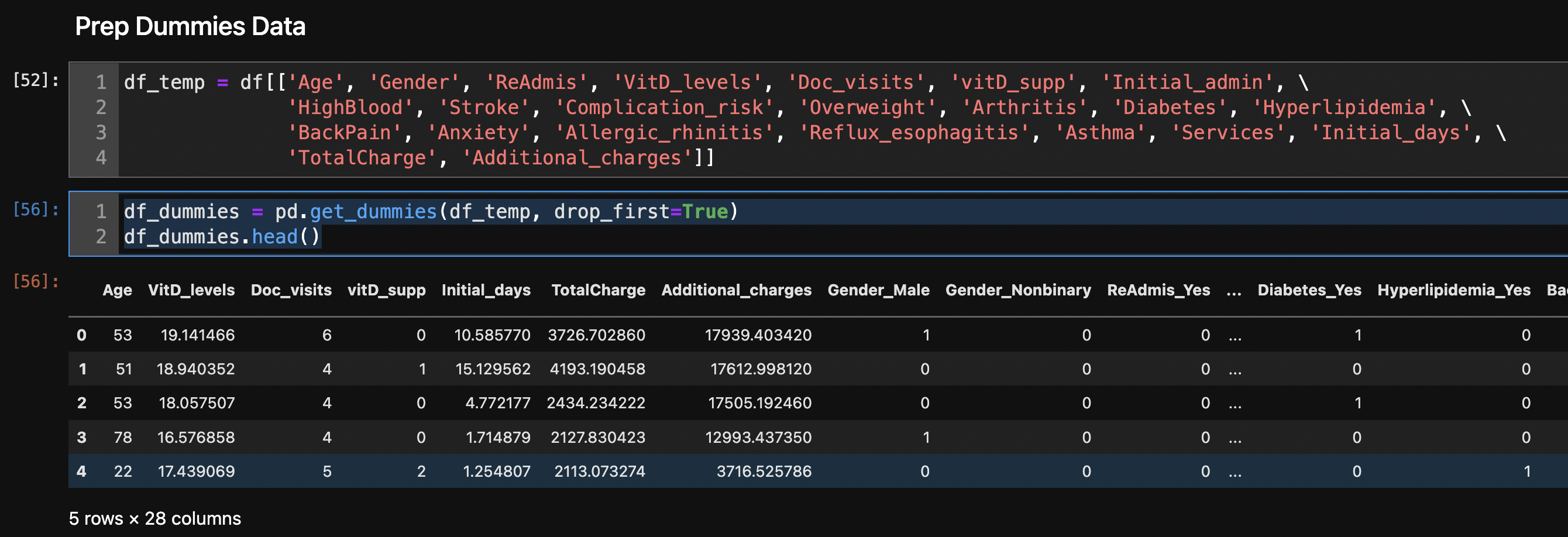


Figure - Create Dummy Data

Additionally, the dataset was standardized using sklearn.preproccessing StandardScaler() function (Figure 4) and then exported as std\_df.csv.

Text

Description automatically generated

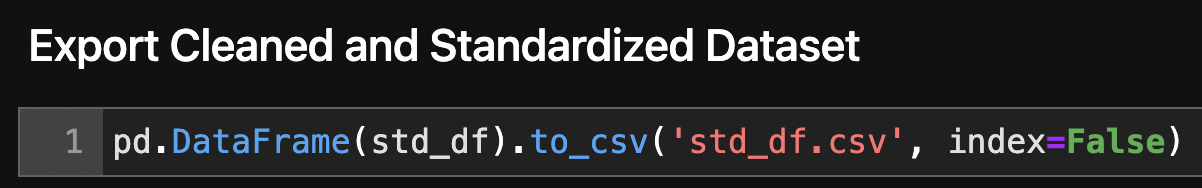


Figure - Standardize and Export Continuous Dataset

Timeline

Description automatically generated

Figure - Matrix of All Principle Components

**D1 – Principal Components**

The submission accurately determines the matrix of all of the principal components (Figure 5).

**D2 – Identification of Total Number of Components**

The submission accurately identifies the total number of principal components. (Figure 6). The ideal number of components are 3 with an accuracy of 95.5%.

|  |  |
| --- | --- |
| **n\_components = 3** | **Optimal n\_components Plot** |
| Text  Description automatically generated | **Graphical user interface  Description automatically generated with medium confidence** |

Figure - Optimal n\_components

**D3 – Total Variance of Components**

The submission accurately identifies the variance of each of the principal components identified in part D2 (Figure 6). You can see that each principal component’s variance ratio and their sum.

**D4 – Total Variance Captured by Components**

The submission accurately identifies the total variance captured by the principal components identified in part D2 using the explained\_variance\_ratio\_.sum(). (Figure 6)

**D5 – Summary of Data Analysis**

At 95.5% accuracy, this model is quite precise while reducing principal component features. I would have a high confidence in the dimension reduction process taken here to prepare the dataset. I believe we achieved our goal of figuring out the minimum number of principal variables for our patients while maintaining the model’s accuracy high.

**E – 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

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>

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>