

Dimensionality Reduction

Matthew E. Heino

Data Mining II

Introduction

In this assessment, the competency that will be demonstrated is the ability to reduce the number of features to a minimum by identifying the significant features of the dataset. This assessment will make use of the concept of principal component analysis (PCA). This concept was first introduced in a previous course and assessment. This assessment will build upon the concepts that were introduced in the previous course.

Background

This assessment will make use of the same dataset. This dataset is composed of 50 columns and 10,000 rows. The dataset is the **medical_clean.csv**. This is a cleaned dataset and should only need very remedial measures to bring the data into a usable manner. The data must be in a state that would be acceptable for principal component analysis (PCA). Please note that not all the columns of the CSV will be used for the assessment. Only columns that are pertinent to the business question will be read in from the CSV and these columns will be cleaned to be ready for principal component analysis. All the other columns will be ignored and these columns will not be read or handled in any way.

Part I: Research Question

In this section, there will be a discussion about the proposed business research question and how it can be answered using the data that is found in the CSV file. The goal of the analysis will be discussed and summarized.

A1. The Research Question: A question that could be answered using the data that is found in the university-provided dataset, is the following "What are factors that can contribute to the readmission to the hospital?" This can be answered by utilizing the features that are available in the provided dataset.

This will align with Scenario 1 as described in the requirements of the assessment. This research question can aid in determining if any of the features in the dataset play an integral role in the patient being readmitted. Continued re-admittance of a patient to the hospital takes a toll on the resources of the hospital and does not reflect well on the reputation of the hospital. This added expense is not welcomed by the patient. If analysis can define the principal components that lead to a patient being readmitted then these components can be addressed in one form or another.

A2. The Goal of the Analysis: The goal of the analysis is to determine which features are most significant when it comes to determining whether or not a patient will be readmitted. If there is a determination of features can these features be addressed in some way? Thereby reducing the economic burden on the hospital and the patient. Reducing this burden would be beneficial to the hospital and the patient.

Part II: Method Justification

In this section, there will be an explanation of how PCA analyzes the dataset. There will be a summarization of one assumption that must be present for PCA.

B1. Explanation of PCA: Datasets are often comprised of many different features. These features compose the columns of the dataset. When we look to answer a question we may not need all the features that are included in the dataset. We may include them all when we build

a model or conduct other types of research. This inclusion of all the dataset's features may make analysis cumbersome and the processing time may become quite long. There is a way to reduce the dimensionality of the dataset without losing any information. This technique is called principal component analysis. This is what will be covered in this section.

What is principal component analysis or PCA? PCA is a method that seeks to reduce the number of features in a dataset to only features that are considered significant. Significant features are those that allow the capture of the most variance. You want to capture as much information from the dataset as possible. There are trade-offs. One of which is when you start omitting features there is the possibility of sacrificing the accuracy of the model or the analysis. With a reduced feature set, there are a few advantages one of which it is easier to perform certain types of calculation or modeling. Fewer features lead to a less complex model. This means that fewer resources both in terms of the amount of data required but also the computing complexity to arrive at the answer of the analysis.

There are a few steps that need to be carried out to undertake PCA. These steps will need to be carried out. These steps are the following (Jaadi, 2023):

1. There is a need to standardize the data so that each of the features is on the same "scale."

This is required because there is a need to ensure that all the prospective components contribute equally to the analysis. If this is not done then features with a large variance will play a more integral part in the analysis. This effect is not desired. We want to look at the features on a level playing field. This is why we need to make sure that we scale these larger variances.

2. In the second step, we need to look at how the features vary regarding one another. This is where we need to create a covariance matrix These values are not to be confused with

the idea of correlation. The values are on a different scale and do not necessarily exhibit values on a scale that falls on a range from -1 and 1 like the correlation matrix will.

3. The third step is to compute the eigenvector and/or eigenvalues. These eigenvector values are composed of a magnitude and a direction relative to a plane. This line is referred to as a line of "best fit" (Adewumi, 2021). After these values are calculated it is now possible to order them. This ordering will make it possible to find the principal components since after the ordering the most significant components will appear at the top of the list.
4. The fourth step involves the creation of a feature matrix. This matrix is a multi-dimensional array that contains the eigenvectors. This matrix is the features that are going to be kept for analysis.
5. Use the feature vector that is composed of the eigenvectors of the covariance matrix to re-orientate from the original axes to the axes that are represented by the principal components.

After these steps are completed you will have reduced the features and the only ones that will be present are significant. In essence, there will be a creation of a new set of components that contain most of the information or only the essential information contained in the features. The components are now an amalgam or a linear combination of the features.

With the creation of the eigenvalues and the eigenvectors, these values will capture how much of the variation of the dataset has been captured with the principal components. The principal components also overcome the problem of multicollinearity. This is because the principal components are considered orthogonal to each other. So the components are not linearly dependent on each other (Duvva, 2021).

The outcome of the PCA analysis is we are left with features that are significant and we maintain as much information about the original dataset. Some of the features that were initially included will no longer be relevant to the analysis as they do not offer any information that warrants them being included going further with future analysis.

B2. Summary of One PCA Assumption: An assumption of PCA is that of orthogonality. As stated in the previous section, the principal components are not linearly dependent on each other. Each principal component must not be dependent on other components (Jain, 2022). If the principal components did not exhibit orthogonality they cannot be considered unique.

Part III Data Preparation

This section will cover the steps that were required to get the supplied dataset into a state where the data can be used for PCA analysis. There will be an identification of the variables that were used in the PCA analysis. There will be a discussion on the standardization of the variables that were used in PCA.

C1. Identification of the Variables: To perform the appropriate PCA analysis there is a need to gather the right variables to answer the question. The table below lists all the variables that will be used to complete the PCA analysis.

Variable Name		Description of the Variable
1	Lat	The latitude of the patient
2	Lng	Longitude of the patient

Dimensionality Reduction

3	Population	Population of the town
4	Children	Number of children in the hospital
5	Age	Age of the patient at the time of the admission.
6	Income	Annual income of the patient.
7	VitD_levels	Measured vitamin D levels
8	Doc_visits	The number of times a doctor visited the patient while in the hospital.
9	Full_meals_eaten	Number of full meals eaten while the patient was in the hospital.
10	VitD_supp	The number of times a vitamin D supplement was administered in the hospital.
11	Initial_days	Number of days the patient stayed in the hospital during the first admission.
12	TotalCharge	The total charge incurred by the patient while they were in the hospital.
13	Additional_charges	Additional charges incurred by the patient.

The rationale for Using These Variables:

The bullet points below offer some of the rationale of why these particular features were singled out for determining readmission.

- Children – the number of children can add additional stress upon their return home.
- Age – older patients will most likely have additional health concerns. These concerns may have not been covered in the hospital or captured in the answers that were given to questions like high blood pressure.
- Income – Lower-income patients may not be able to afford the appropriate care and other items that contribute to better health.
- VitD_levels – the amount of this supplement might have a bearing on whether a patient is readmitted to the hospital. Is there a correlation between the levels and readmission?
- Doc_visits – if the doctor did not visit the patient it could lead to the patient being readmitted.
- Full_meals_eaten – the patient did not receive enough meals before being discharged from the hospital.
- VitD_supp – if the patient did not receive enough supplements to keep healthy and during recovery.
- Initial_days – longer the hospital stay could mean the patient was in for a serious condition. This could indicate the condition was not adequately solved and this led to readmission.

These are the eight continuous variables out of the fifty that would most likely lead to readmission to the hospital. The other variables did not seem like they would contribute to readmission to the hospital. This included a few continuous variables like the latitude, longitude,

and the population of the town or city. If this analysis does not seem to be productive other variables could be examined at a later date.

C2. Standardization of the Selected Data: Since the variables that were described in the previous section vary a great amount in terms of their scale. There are a few ways to scale these variables. One way is to use the following formula:

- $(\text{Values of the data frame} - \text{mean of the data frame}) / \text{data frame's standard deviation}$

This formula was discussed in the DataCamp video "Dimensionality Reduction in Python" by J. Boeye a Datacamp video and WGU course resource. The method that will be used in this assessment will make use of a method that is provided by the **Sklearn** library. The method is **StandardScaler**. This method will scale the values by removing the mean and then will scale each of the features to unit variance (Loukas, 2020).

The code used to accomplish this can be found in section C2 of the accompanying Jupyter Notebook. Additional information may be found in this section.

This standardized dataset can be found in the following file:

- **Heino_D212_Task2_Standardized_Data.csv**

Part IV Analysis

In this section, there will be a determination of the matrix of all the components that are relevant and significant features that can be used. In one section there will be a discussion on how the **elbow rule** or the **Kaiser rule** was used to identify the total number of components. There will be an identification of the total variance that was captured by each of the principal components identified in a previous section. There will be a summary of the results of the PCA analysis.

D1. Matrix of the Principal Components: The Principal Component matrix was composed of the eight variables that were discussed in section C of this paper. These variables were first standardized using the method that was discussed in section C2. To create the Principal Component matrix, Python provides a library function that allows the creation of the matrix. The function is the **PCA()** function that is found in **sklearn.decomposition** library. This will take a few arguments. The first is the **n_components**. This is the number of features that will be used to create the principal component. The next is the **randoms_state** to make sure the code is reproducible at a later date.

There was a matrix created as well as a heat map to show the results The results are shown below.

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11	I
Lat	-0.018834	0.000913	-0.715570	-0.036559	0.128188	-0.018260	-0.039974	-0.005117	-0.067661	-0.039423	0.679459	0.00
Lng	-0.011011	0.009716	0.274895	-0.474659	-0.554592	-0.289613	0.229759	0.320779	0.056053	0.033702	0.384029	-0.00
Population	0.028719	-0.029027	0.626046	0.295638	0.250669	0.142253	-0.174676	-0.135732	-0.083567	0.038751	0.615001	0.01
Children	0.034537	0.017244	-0.034510	0.344621	0.158969	0.231131	0.427505	0.717166	-0.131085	0.292473	-0.006222	0.00
Age	0.084650	0.700793	0.011244	-0.020860	0.010691	0.011755	0.006632	-0.017856	-0.013308	-0.020631	-0.001154	0.70
Income	-0.019701	-0.019176	0.075776	-0.067301	0.412381	-0.149024	0.651545	-0.162893	0.461862	-0.359436	0.056064	0.00
VitD_levels	-0.001995	0.020340	-0.020176	0.526197	-0.213021	-0.366372	-0.208667	0.305325	0.061710	-0.634109	-0.003265	-0.00
Doc_visits	-0.006991	0.015446	0.017291	0.096735	0.282211	-0.820104	0.040698	-0.076493	-0.285582	0.381544	-0.056573	0.00
Full_meals_eaten	-0.020712	0.031960	-0.103248	0.454738	-0.385982	-0.050904	0.062235	-0.238447	0.590939	0.462602	0.073982	0.01
vitD_supp	0.025381	0.014511	0.029741	-0.262904	0.377611	-0.097049	-0.508283	0.424062	0.565530	0.137073	-0.018434	0.00
Initial_days	0.699994	-0.089859	-0.022902	-0.007101	-0.018751	-0.017957	0.013322	-0.023812	0.008988	-0.007002	0.000171	0.03
TotalCharge	0.701146	-0.079267	-0.020888	-0.003830	-0.019601	-0.019199	0.012132	-0.022769	0.009702	-0.005149	0.000729	-0.00
Additional_charges	0.085029	0.700745	0.013730	-0.004630	0.019713	0.016979	0.006236	-0.025023	-0.006886	-0.010633	0.020560	-0.70

Image 1. The PCA matrix of all components.

Dimensionality Reduction

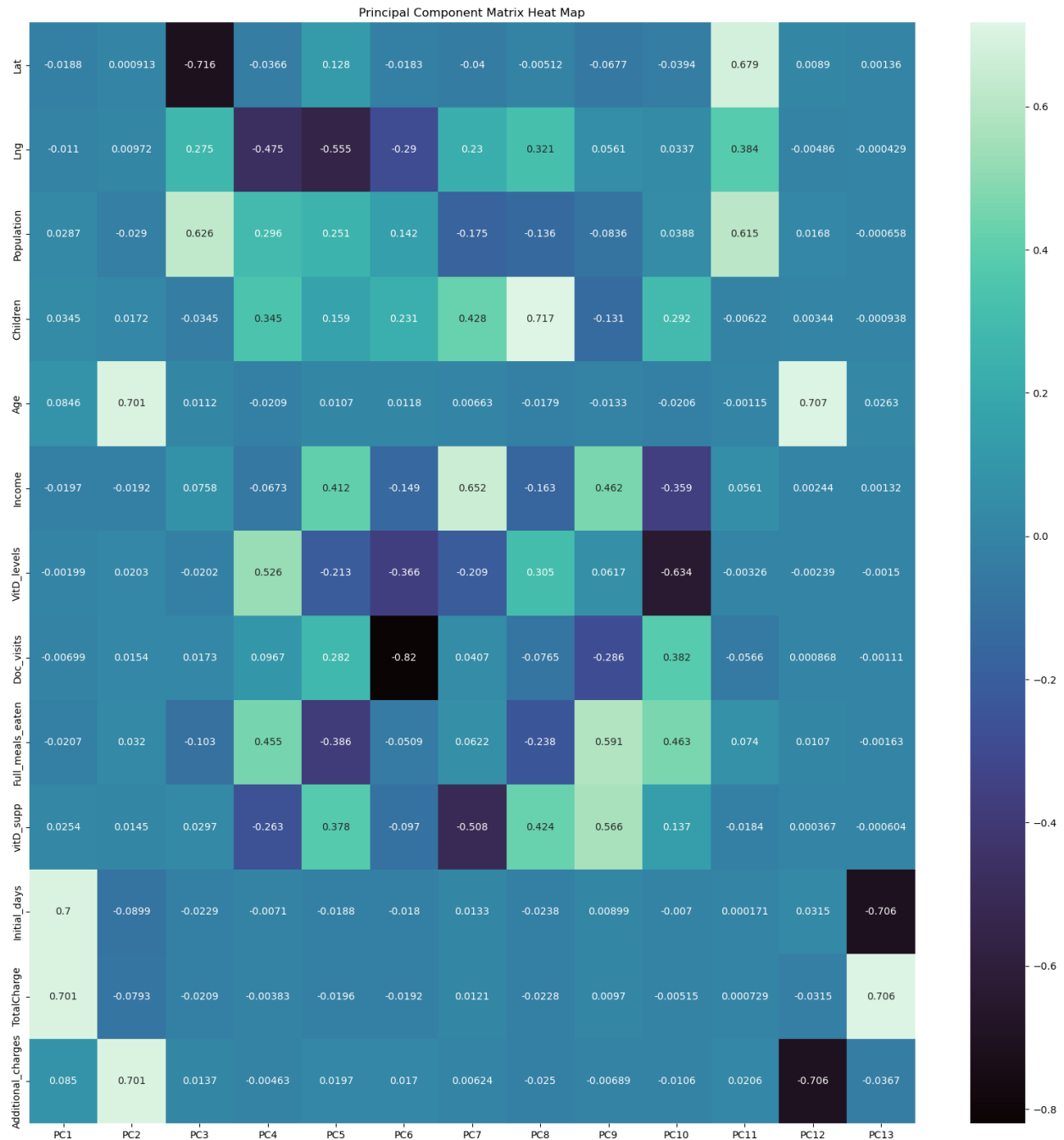


Image 2. The heat map PCA matrix of all components.

The code that was used to create the matrix can be found in the following section of the accompanying Jupyter Notebook – **Section D1**.

D2. Identification of the Components: To help identify the total number of principal components it was helpful to plot these components as a graph. Using a Scree plot aided in

Dimensionality Reduction

visualizing the number of principal components relative to the percentage of the variance that was explained with each inclusion of an additional plot. There is an included table that has the values that were used to create the required Scree plot.

	Captured Variance by PC	Eigenvalues by PC
PC1	15.35	1.995695
PC2	13.19	1.715488
PC3	9.44	1.227082
PC4	8.00	1.040267
PC5	7.98	1.037347
PC6	7.74	1.005982
PC7	7.73	1.004430
PC8	7.61	0.989870
PC9	7.51	0.975979
PC10	7.44	0.967452
PC11	5.74	0.746599
PC12	2.18	0.283395
PC13	0.09	0.011713

Image 3. The values used to create the Scree plot.

Dimensionality Reduction

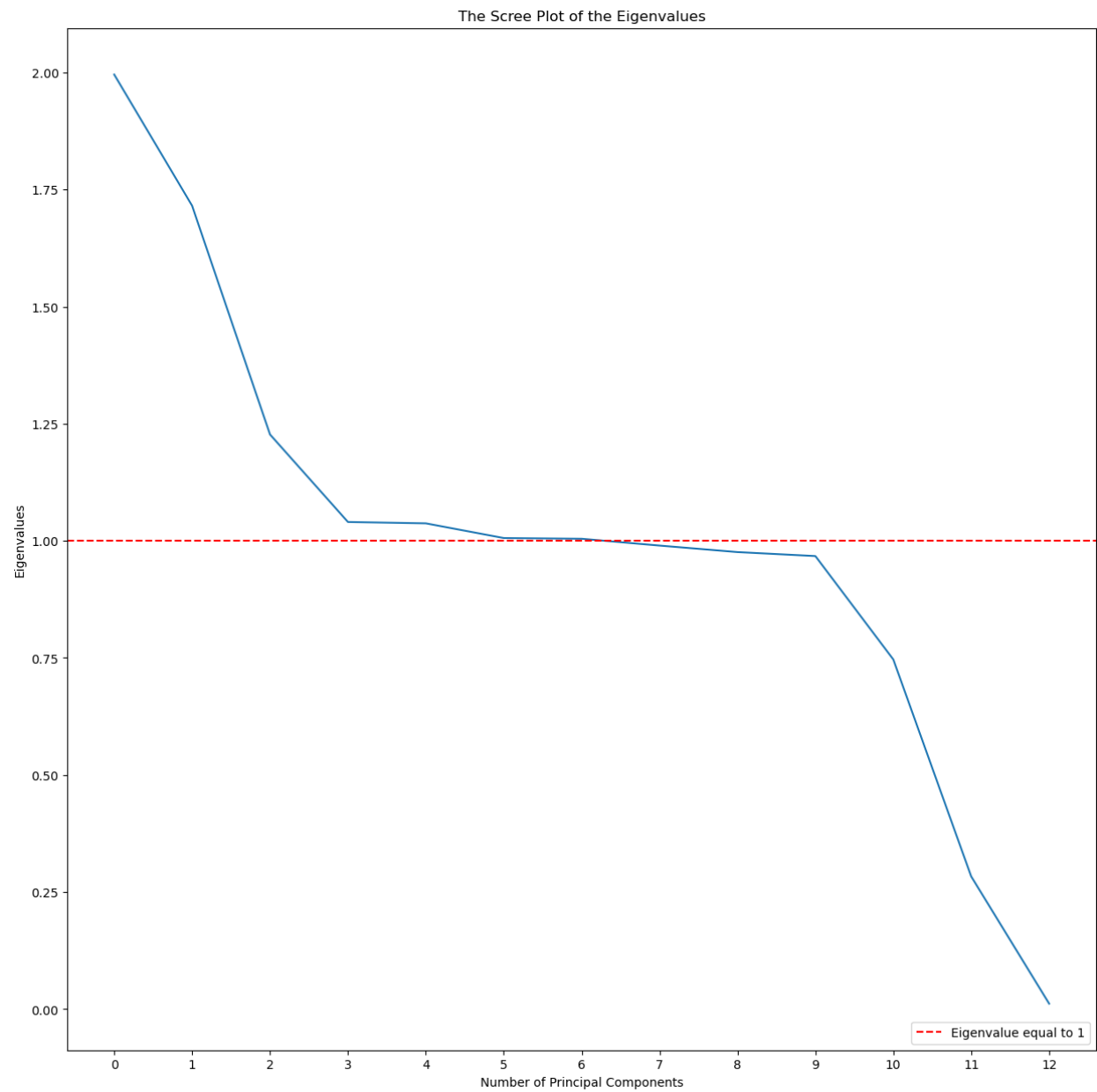


Image 4. The Scree plot.

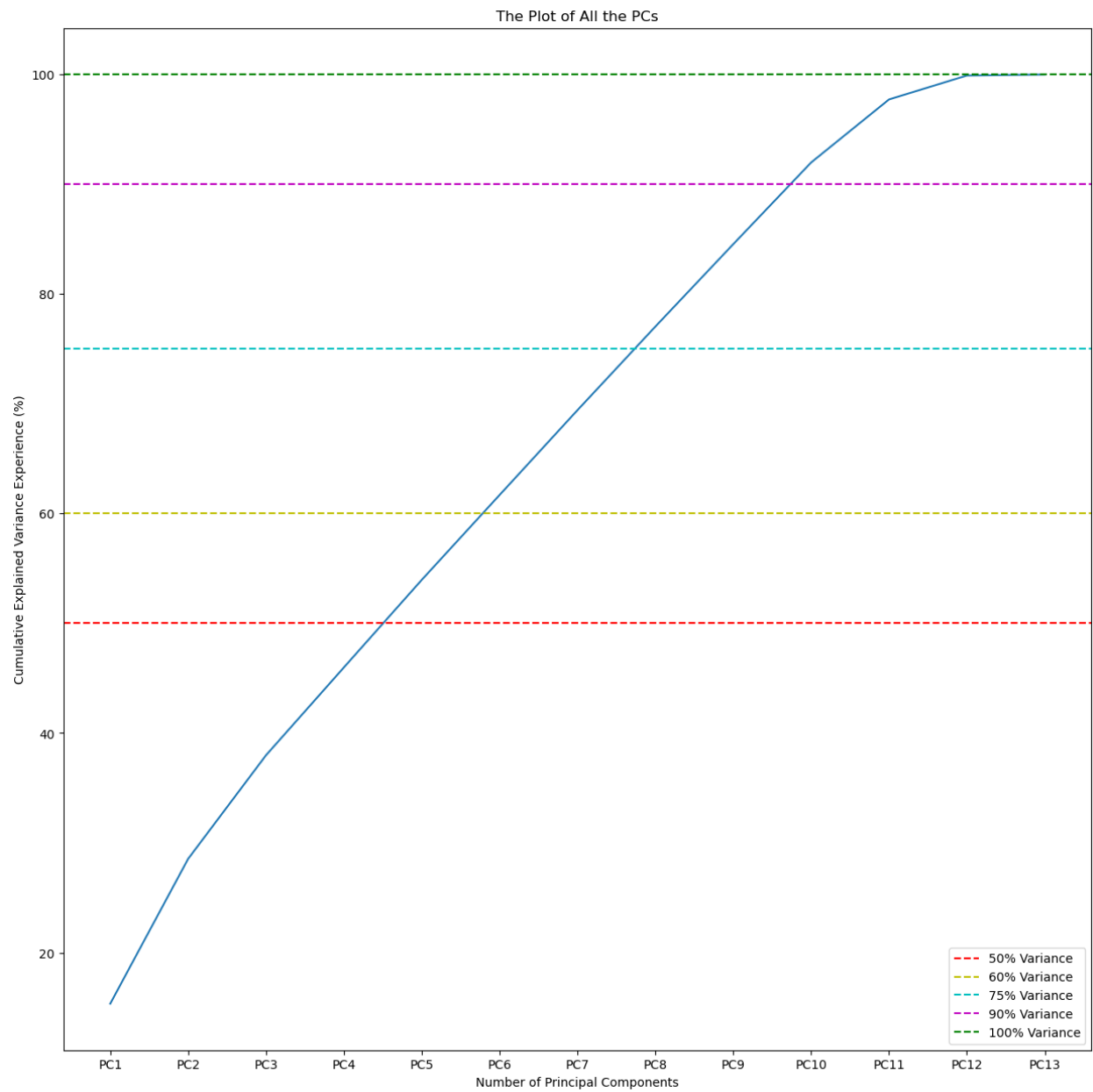
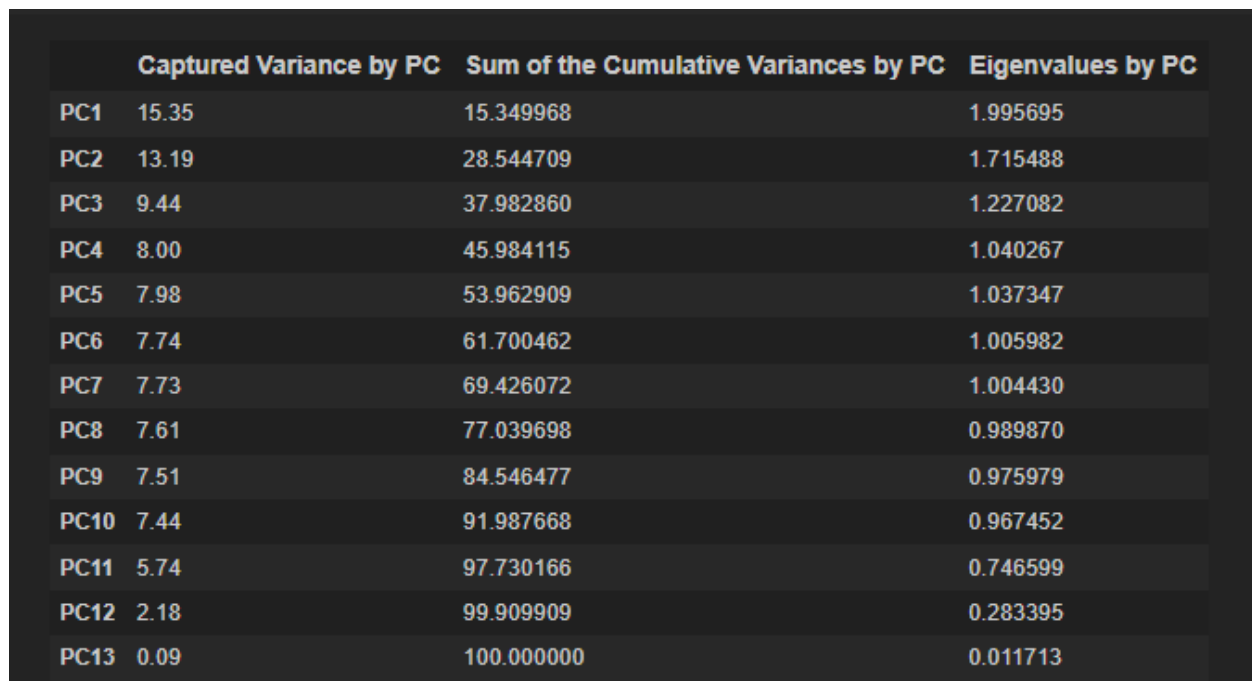


Image 5. Cumulative Plot of all the PC variances.

The plot above shows the results of the plot. We can ascertain that using 10 principal components we cover about 91% of the variance within the dataset. This will reduce the number of components by two.

The code to accomplish this section can be found in Jupyter Notebook – **Section D2**
Create the Elbow plot.

D3. Identification of the Variance: The table displays the variances of the principal components. You will see the breakdown of the captured variance by principal component, a running sum of the variances as well as a listing of the eigenvalues.



	Captured Variance by PC	Sum of the Cumulative Variances by PC	Eigenvalues by PC
PC1	15.35	15.349968	1.995695
PC2	13.19	28.544709	1.715488
PC3	9.44	37.982860	1.227082
PC4	8.00	45.984115	1.040267
PC5	7.98	53.962909	1.037347
PC6	7.74	61.700462	1.005982
PC7	7.73	69.426072	1.004430
PC8	7.61	77.039698	0.989870
PC9	7.51	84.546477	0.975979
PC10	7.44	91.987668	0.967452
PC11	5.74	97.730166	0.746599
PC12	2.18	99.909909	0.283395
PC13	0.09	100.000000	0.011713

Image 5. Principal component and cumulative variances by percentage.

The code to accomplish this section can be found in Jupyter Notebook – **Section D2**
Create the Elbow plot.

D4. Identification of the Total Variance: The graph below displays the total variance of the principal components. This graph depicts the individual variance captured across the bottom of the graph as a bar chart. The cumulative variance is depicted as a step graph. Each

Dimensionality Reduction

graph has the relevant information included. The bar graph has the percentage that this particular component contributed to capturing the variance. The steps have the total variance as they are added to the previous principal component.

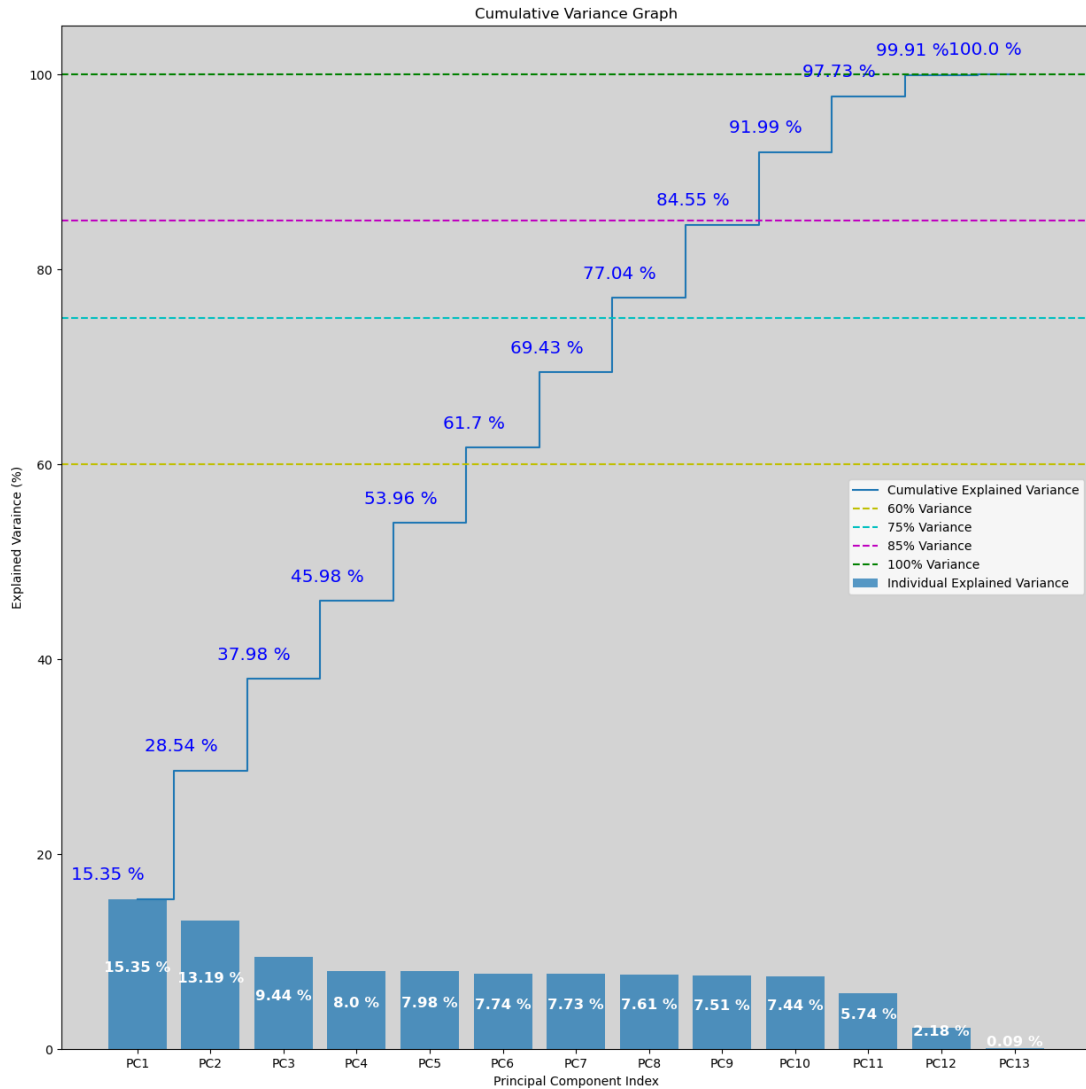


Image 6. Principal component and cumulative variances.

D5. Summary of the Data Analysis: Reviewing the composition of the matrix we can see that some components are more likely to contribute to the likelihood of being readmitted. It

is not surprising that features like age and the number of days spent in the hospital played a major factor in the readmission of the patient to the hospital.

Features like latitude and longitude did not seem to play a major role in the patient being admitted. This does not seem surprising since these values are not a good indicator of admittance.

It can be debated if features like the ones that record charges can be safely used to determine if a patient will be readmitted to the hospital. The granularity of these charges is not apparent with this feature. The meaning here is that we do not know what the charges were that the patient incurred. We do not know if the charges come from a single lengthy hospital stay or are the charges from a series of procedures that could add a hefty amount to the charges of the initial bill. Without this information, we cannot be sure what effect this feature will have on the ability to predict if the patient will be readmitted.

Analyzing the PCA analysis, we can capture the majority of the variance, about 91%, using only ten principal components. If we are willing to accept a lower capture percentage we can use only nine principal components. This would reduce the size of the features that are needed to model and answer the question proposed at the beginning of the paper. Using fewer features would mean that we will require less resources and less data. This would also lead to less computational complexity and reduced runtime in answering the question if we choose to run a model with larger datasets than the one that was provided.

References

Part V Attachments

E. Web Resources

Web sources that were used to create the source code and other references that were not part of the WGU or DataCamp resources. Please note that some of the code was used from some of my other assessments and will not have any citations.

Kumar, A. (2023, November 24). *PCA Explained Variance Concepts with Python*

Example. Analytics Yogi. Retrieved January 6, 2024, from <https://vitalflux.com/pca-explained-variance-concept-python-example/>

Loukas, S., Ph.D. (2020, May 26). *How and why to Standardize your data: A python tutorial* |

Towards Data Science. Medium. Retrieved January 4, 2024, from

<https://towardsdatascience.com/how-and-why-to-standardize-your-data-996926c2c832>

F. In-text Citations

Any citations that are used to create or referenced in this Word document. This may include references to code segments that were discussed in the paper.

Adewumi, J. (2021, December 9). Understanding the role of eigenvectors and eigenvalues in

PCA dimensionality reduction. *Medium.* Retrieved January 4, 2024, from

<https://medium.com/@dareyadewumi650/understanding-the-role-of-eigenvectors-and-eigenvalues-in-pca-dimensionality-reduction-10186dad0c5c>

Boeye, J. (n.d.). *Dimensionality Reduction in Python*. DataCamp. Retrieved October 29, 2023, from <https://app.datacamp.com/learn/courses/dimensionality-reduction-in-python>

Duvva, P. (2021, December 29). The power of Eigenvectors and Eigenvalues in dimensionality reduction techniques such as PCA. *Medium*. Retrieved January 4, 2024, from <https://medium.com/wicds/the-power-of-eigenvectors-and-eigenvalues-in-dimension-reduction-techniques-such-as-pca-8540322124ea>

Jaadi, Z. (2023, March 29). *A Step-by-Step Explanation of Principal Component Analysis (PCA)*. Built In. Retrieved January 4, 2024, from <https://builtin.com/data-science/step-step-explanation-principal-component-analysis>

Jain, S. (2022, January 6). Limitations, Assumptions Watch-Outs of principal component analysis. *Medium*. Retrieved January 4, 2024, from <https://codatalicious.medium.com/limitations-assumptions-watch-outs-of-principal-component-analysis-8483ceaa2800>