D212: Data Mining II

Performance Assessment

Task 2

Logan Rosemeyer

Western Governor's University

D212: Data Mining II

Dr. Keiona Middleton

February 28, 2023

Contents

A1: Proposal of Question

One question I will answer is: How are population, children, age, income, vitamin D levels, doctor visits, full meals eaten, vitamin D supplements, initial days, total charge, and additional charges related, and can we identify the principal components of our patients using these variables?

A2: Defined Goal

The goal of the data is to reduce the dimensions in the data set. By reducing the dimensionality of the variables population, children, age, income, vitamin D levels, doctor visits, full meals eaten, vitamin D supplements, initial days, total charge, and additional charges, the organization can identify the principal components of their patients, which will allow for better business and strategic decision-making.

B1: Explanation of PCA

PCA will help us analyze the data set by reducing the dimensionality, which will increase interpretability by creating new uncorrelated components (Jolliffe and Cadima, 2016). This is accomplished by linearly transforming the data into a new coordinate system where the variation in the data can be described with fewer dimensions that the initial data. I would expect for the PCA to give us 2-3 principal components that we can use to make better business decisions.

B2: PCA Assumption

One assumption of PCA is that it assumes that the variables have a linear relationship. If the relationship is non-linear, PCA will not be able to find a relationship between the variables. (2022)

B3: Packages or Libraries List

I used the dplyr package to create a new dataframe with only the variables I needed to perform PCA. I used the factoextra package for the prcomp function needed to perform PCA.

C1: Continuous Dataset Variables

The continuous variables I am using to answer my question are population, children, age, income, vitamin D levels, doctor visits, full meals eaten, vitamin D supplements, initial days, total charge, and additional charges.

C2: Standardization of Dataset Variables

Variables have been standardized using the scale function in R. A copy of the cleaned data set is included in the submission.

D1: Principal Components

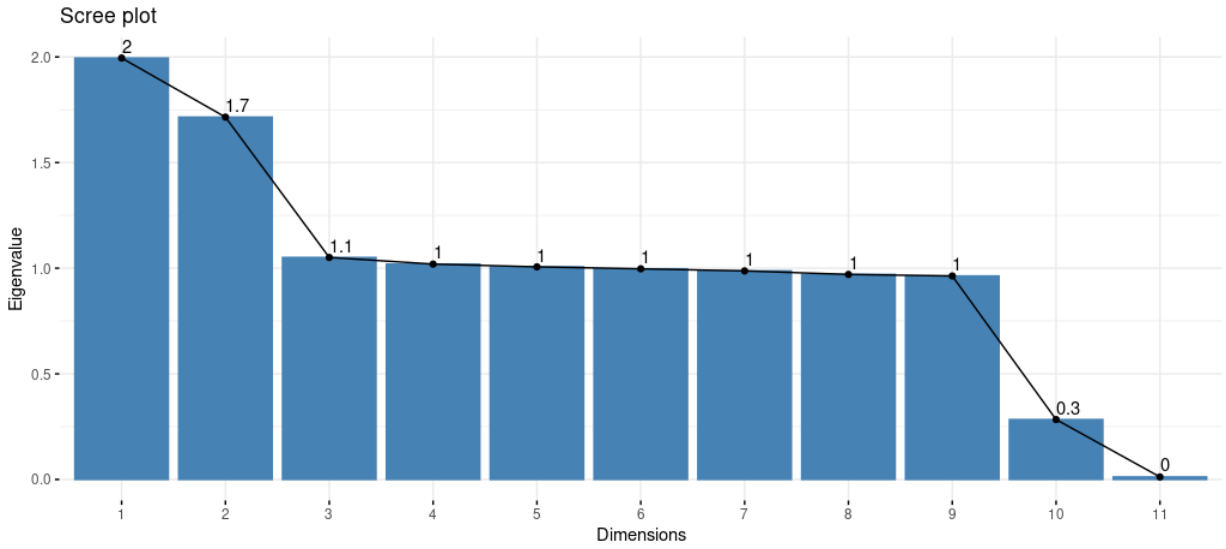
	PC1	PC2	PC3	PC4	PC5
Population	0.024451174	-0.02835355	0.42185533	-0.370208759	0.235895710
Children	0.034519574	0.01741021	-0.09673081	-0.348999558	-0.518806655
Age	0.084872008	0.70076170	0.02352361	0.017158878	-0.006688610
Income	-0.020183888	-0.01904644	0.32525229	-0.196995274	-0.658377615
VitD_levels	-0.002039300	0.02043746	-0.35991297	-0.572299909	0.328572558
Doc_visits	-0.006888796	0.01544715	0.21314372	-0.593921973	0.106641745
Full_meals_eaten	-0.020702590	0.03214605	-0.60361328	-0.129857514	-0.030405883
vitD_supp	0.025366294	0.01452231	0.40701545	0.069521191	0.348283329
Initial_days	0.700221817	-0.09023546	-0.01841090	0.005489042	-0.006565308
TotalCharge	0.701357789	-0.07964337	-0.01960962	0.002230420	-0.004801098
Additional_charges	0.085222600	0.70076898	0.02524196	0.006438039	-0.007918954

	PC6	PC7	PC8	PC9
Population	-0.349304908	0.479600344	-0.222539551	0.475576302
Children	-0.626415103	-0.279714581	0.352458633	0.057534073
Age	0.003293583	0.024529276	-0.008834421	-0.014554347
Income	0.322788071	-0.081858092	-0.554280623	-0.055189636
VitD_levels	-0.133582884	-0.092902481	-0.414384153	-0.486165172
Doc_visits	0.545545433	-0.072843268	0.535508626	0.021684777
Full_meals_eaten	0.228782705	-0.172001750	-0.184048574	0.707447655
vitD_supp	-0.106249605	-0.799984712	-0.161924229	0.172815164
Initial_days	0.034372722	0.005663564	-0.009233506	-0.004639713
TotalCharge	0.033306520	0.005327961	-0.009057134	-0.002110143
Additional_charges	-0.001301828	0.029718898	-0.014130359	0.007696578

	PC10	PC11
Population	-0.0143888892	0.0009303120
Children	-0.0036278384	0.0009221462
Age	-0.7065734425	-0.0262453810
Income	-0.0022592555	-0.0012955594
VitD_levels	0.0023256664	0.0014915518
Doc_visits	-0.0010128499	0.0010986718
Full_meals_eaten	-0.0108030429	0.0016289252
vitD_supp	-0.0004330654	0.0005986480
Initial_days	-0.0314681168	0.7062748671
TotalCharge	0.0314432208	-0.7064904473
Additional_charges	0.7059932141	0.0367205953

D2: Identification of Total Number of Components

Using the Kaiser Rule, I determined that I would use nine principal components. I decided on nine because they have an eigenvalue of at least 1, which means that the component has as much, or more, information than a single variable.



D3: Total Variance of Components

The first principal component explains 18.14% of the variance, while the second principal component explains 15.59% of the variance. The third explains 9.55% of the variance. The fourth explains 9.267% of the variance. The fifth explains 9.149% of the variance. The sixth explains 9.064% of the variance. The seventh explains 8.975% of the variance. The eighth explains 8.823% of the variance. The ninth explains 8.757% of the variance.

```
Importance of components:
                          PC1     PC2     PC3     PC4     PC5     PC6     PC7
Standard deviation     1.4124 1.3097 1.02510 1.00964 1.00320 0.99851 0.99359
Proportion of Variance 0.1814 0.1559 0.09553 0.09267 0.09149 0.09064 0.08975
Cumulative Proportion  0.1814 0.3373 0.43282 0.52549 0.61699 0.70762 0.79737
                           PC8     PC9    PC10    PC11
Standard deviation     0.98514 0.98145 0.53240 0.10823
Proportion of Variance 0.08823 0.08757 0.02577 0.00106
Cumulative Proportion  0.88560 0.97317 0.99894 1.00000
```

D4: Total Variance Captured by Components

My principal components explain 97.317% of the variance.


```
Importance of components:
                          PC1     PC2     PC3     PC4     PC5     PC6     PC7
Standard deviation     1.4124 1.3097 1.02510 1.00964 1.00320 0.99851 0.99359
Proportion of Variance 0.1814 0.1559 0.09553 0.09267 0.09149 0.09064 0.08975
Cumulative Proportion  0.1814 0.3373 0.43282 0.52549 0.61699 0.70762 0.79737
                           PC8     PC9    PC10    PC11
Standard deviation     0.98514 0.98145 0.53240 0.10823
Proportion of Variance 0.08823 0.08757 0.02577 0.00106
Cumulative Proportion  0.88560 0.97317 0.99894 1.00000
```

D5: Summary of Data Analysis

My analysis found that we can use three principal components to better understand the data. The first principal component found that there is a high correlation between initial days and total charge. The second principal component found a high correlation between age and additional charges. The third principal component found a positive correlation with population and vitamin D supplements, and a negative correlation with full meals eaten. The fourth found a negative correlation with doctor visits and vitamin D levels. The fifth found a negative correlation with children and income. The sixth found a positive correlation with doctor visits and a negative correlation with children. The seventh found a negative correlation with vitamin D supplements and a positive correlation with population. The eighth found a positive correlation with doctor visits and a negative correlation with income. The ninth found a positive correlation with full meals eaten and population, and a negative correlation with vitamin D levels.

	PC1	PC2	PC3	PC4	PC5
Population	0.024451174	-0.02835355	0.42185533	-0.370208759	0.235895710
Children	0.034519574	0.01741021	-0.09673081	-0.348999558	-0.518806655
Age	0.084872008	0.70076170	0.02352361	0.017158878	-0.006688610
Income	-0.020183888	-0.01904644	0.32525229	-0.196995274	-0.658377615
VitD_levels	-0.002039300	0.02043746	-0.35991297	-0.572299909	0.328572558
Doc_visits	-0.006888796	0.01544715	0.21314372	-0.593921973	0.106641745
Full_meals_eaten	-0.020702590	0.03214605	-0.60361328	-0.129857514	-0.030405883
vitD_supp	0.025366294	0.01452231	0.40701545	0.069521191	0.348283329
Initial_days	0.700221817	-0.09023546	-0.01841090	0.005489042	-0.006565308
TotalCharge	0.701357789	-0.07964337	-0.01960962	0.002230420	-0.004801098
Additional_charges	0.085222600	0.70076898	0.02524196	0.006438039	-0.007918954

	PC6	PC7	PC8	PC9
Population	-0.349304908	0.479600344	-0.222539551	0.475576302
Children	-0.626415103	-0.279714581	0.352458633	0.057534073
Age	0.003293583	0.024529276	-0.008834421	-0.014554347
Income	0.322788071	-0.081858092	-0.554280623	-0.055189636
VitD_levels	-0.133582884	-0.092902481	-0.414384153	-0.486165172
Doc_visits	0.545545433	-0.072843268	0.535508626	0.021684777
Full_meals_eaten	0.228782705	-0.172001750	-0.184048574	0.707447655
vitD_supp	-0.106249605	-0.799984712	-0.161924229	0.172815164
Initial_days	0.034372722	0.005663564	-0.009233506	-0.004639713
TotalCharge	0.033306520	0.005327961	-0.009057134	-0.002110143
Additional_charges	-0.001301828	0.029718898	-0.014130359	0.007696578

E: Sources for Third-Party Code

Scale: Scaling and Centering of Matrix-like Objects. RDocumentation. Retrieved March 7, 2023. https://www.rdocumentation.org/packages/base/versions/3.6.2/topics/scale

F: Sources

Jolliffe Ian T. and Cadima Jorge.*Principal component analysis: a review and recent developments*. Philosophical Transactions R. Soc. A. April 13, 2016. Retrieved March 6, 2023. https://doi.org/10.1098/rsta.2015.0202

A Guide to Principal Component Analysis (PCA) for Machine Learning. Keboola. April 2, 2022. Retrieved March 7, 2023. https://www.keboola.com/blog/pca-machine-learning#:~:text=PCA%20assumes%20a%20linear%20relationship,methods%20such%20as%20log%20transforms.