Data Acquisition – D206

Jeffrey Williams

Western Governors University

Student #: 001173968

Date: 01/20/2023

Mentor: Jared Knepp

**Table of Contents**

**Part A.** Question or Decision 3

**Part B.** Required Variables3

**Part C1.** Plan to Find Anomalies4

**Part C2.** Justification of Approach**5-6**

**Part C3.** Justification of Tools6

**Part C4.** Provide Code7-15

**Part D1.** Cleaning Findings**7-19**

**Part D2.** Justification of Mitigation Methods19-20

**Part D3.** Summary of The Outcomes 20

**Part D4.** Mitigation Code**7-15**

**Part D5.** Clean Data.21

**Part D6.** Limitations.21

**Part D7**. Impact of Limitations.**22**

**Part E1.** Principal Components.22

**Part E2.** Criteria.22-23

**Part E3.** Benefits.**23**

**Part F.** Video23

**Part** **G& H.** Sources for Third-Party Code & Paper Sources24

**Part A: Question or Decision**

The research question focuses on discovering a relationship between various health determinants and medical complication risks.

**Part B. Required Variables**

The dependent variable is patient’s complication risk. The independent variables fields are gender, age, risk of high blood pressure, anxiety, obesity, reflux esophagitis, number of children, and allergic rhinitis. The medical raw dataset analyzed by viewing certain characteristics such as number of rows, number of columns, data type of each column, etc. The below images highlights the variable names, data type, and examples of each of the variables found within the providing dataset. Utilizing the total number of rows and columns, can determine the various relationships between data.

Table, calendar

Description automatically generatedTable

Description automatically generated

**Part C1: Plan to Find Anomalies.**

* Using R Studio import the CSV file into the platform - *RStudio: (2023, January, 19)*



* Check if each column is a consistent data type- *r-lang.com: (2012, February, 16)*



* Validate the data types within the columns are the same - *r-lang.com: (2012, February, 16)*



* Delete Columns not relevant - *r-lang.com: (2012, February, 16)*



* Use Central Tendency to identify and replace missing values (Mean, Median, or Mode) (*Larcose, 2019, p.29-43*)

A screenshot of a computer

Description automatically generated with medium confidence

* Create function to change categorical columns to numeric and to apply- *r-lang.com: (2012, February, 16)*



* Use boxplot function to detect outliners for numerical variables boxplot- *datamentor.io: (2018, March, 13)*

Text

Description automatically generated

* Use Principal Component Analysis (PCA) for chosen dependent and independent variables and standardize the input data to eliminate mean and variance. (*Larcose, 2019, p.29-43*). *Principal component analysis for dimensionality reduction (2019, May, 24)-* 
* Extract and view the eigen vlaues and variances- *www.geeksforgeeks.org: (2020, June, 03)*



**Part C2. Justification of Approach**

My approach to cleaning the medical raw data was by using the seven aspects of data quality listed below: (*Precisely.com, 2022, November,14*)

|  |  |
| --- | --- |
| Accuracy | is the data relating to the column and not misleading |
| Completeness | The data set can provide a full view into the information needed to answer the research question |
| Consistency | Does data stored in one place match relevant data stored elsewhere? |
| Relevancy | Data pertains logically to research question |
| Validity | The data follows a set constraint and parameters of measure is appropriate |
| Uniqueness | Does the data have duplicates or redundant information |

The First challenge is to change the raw data that does not accurately represent the dataset. The format also needs to be accurate by added adjustments, misleading values will be changed, proper index field will be created, and finally change categorial data into numeric.

I will be using R software to clean the data to implement functions that provide solutions by, manipulating the data and creating visual resources to reflect the changes made. R also has packages I can upload to clean the data and identify outliers.

Listed below are the packages installed, and libraries added to R Studio and used to clean the raw dataset: (*kdnugget (2016, November, 16*)

Text, letter

Description automatically generated

**Part C3. Justification of tools**

The programming language R cleans and analyze large datasets. R Studio is an easy to use tool that cleans and analyzes datasets efficiently using the R programming language. This platform also has a plethora of packages designed to read and clean data. (RStudio).

Principal Component Analysis (PCA) can classify datasets into smaller variable groups. Principal Component Analysis, or PCA, is a dimensionality-reduction method used to reduce the number of input variables in a data set by changing a large set of variables into a smaller one that still contains most of the information acquired from the larger dataset. **Eigen()** function in R Language is used to calculate eigenvalues and eigenvectors of a matrix and needs to be viewed in order to perform the PCA.( *Principal component analysis for dimensionality reduction (2019, May, 24)*). *(www.geeksforgeeks.org: (2020, June, 03)).*

***\*\*\*(Description of each library chosen is included below (via “#”” note”)\*\*\****

Text

Description automatically generated

Graphical user interface, text, application, chat or text message

Description automatically generated

**Part C4. Provide the Code; Part D4. Mitigation Code**

Text, letter

Description automatically generated

Application

Description automatically generated with low confidence



Text

Description automatically generated

A picture containing text

Description automatically generated

Graphical user interface, text, application, chat or text message

Description automatically generated



Graphical user interface, text, application

Description automatically generated

Text

Description automatically generated

Graphical user interface, text, application

Description automatically generated

 







A screenshot of a computer

Description automatically generated with medium confidence

Chart, box and whisker chart

Description automatically generated

A screenshot of a computer

Description automatically generated with medium confidence

Diagram

Description automatically generated

Graphical user interface, text, application, email

Description automatically generated

Graphical user interface, text, application

Description automatically generated

Graphical user interface, text

Description automatically generated

Text

Description automatically generated

Graphical user interface

Description automatically generated with medium confidence



Text, letter

Description automatically generated

Table

Description automatically generated





Chart

Description automatically generated



****

**Part D1. Cleaning Findings**

While cleaning the raw dataset I found and addressed the following anomalies:

* No duplications were found.
* The below values had missing variables or a value of zero, I replaced all the values using Use Central Tendency to identify and replace missing values (Mean, Median, or Mode) *(Larcose, 2019, p.29-43*).
  + ‘Age’
  + ‘Children’
  + ‘Household\_Income’
  + ‘Population’
  + ‘Soft Drink’
  + ‘Anxiety’,
  + ‘Initial\_days’
  + ‘Overweight’
* The survey data was in normal range
* Discovered outliers by utilizing numerical data creating a function to change categorical columns to numeric and to apply- *r-lang.com: (2012, February, 16)*



* Use boxplot function to detect outliners for numerical variables boxplot- *datamentor.io: (2018, March, 13)*

Text

Description automatically generated

Chart, box and whisker chart

Description automatically generated

Diagram

Description automatically generated

* Use Principal Component Analysis (PCA) for chosen dependent and independent variables and standardize the input data to eliminate mean and variance. (*Larcose, 2019, p.29-43*). *Principal component analysis for dimensionality reduction (2019, May, 24)-* 

Text, letter

Description automatically generated

Table

Description automatically generated

* Extract and view the eigen vlaues and variances- *www.geeksforgeeks.org: (2020, June, 03)*

Graphical user interface, text, application, email

Description automatically generated

Graphical user interface, text, application

Description automatically generated

Graphical user interface, text

Description automatically generated

Chart, line chart

Description automatically generated

**Part D2. Justification of Mitigation Methods**

* Replaced categorical variables by changing the null values with mode. Any numerical columns containing an “NA” values replaced using mean. Implemented median to maintain data distribution. (*Larcose, 2019, p.29-43*).
* After identifying all the outliers, I decided to keep them. If I removed the outliers, approximately 90% of the information would be unavailable and I would not have been able to utilize box plots to analyze the dataset.
* By converting categorical values into numeric, I was able to utilize numeric data for boxplots, PCA, and eigenvalues.
* Both “Full meals” and “Vit D” contained outliners I was unable to verify them without discussing with the dataset originator.

**Part D3. Summary of the Outcomes**

* Focusing on duplication within the dataset, I was able to solidify future data integrity and consistency.
* Utilizing the “**colSums(is.na(medical)).** **Is.na”** function it was able to show all the columns with “NA” listed in the dataset. Using the median to impute NAs in numeric columns and the mode to impute NAs in categorical columns.NA values in the children, age, income, overweight, anxiety columns were imputed using the median while NA values in the soft drink column were imputed using the mode. This prevents inconsistency when performing future statistical analysis. (*www.geeksforgeeks.org: (2020, June, 03)).*
* Created a data frame with the outliers imputed with median and one frame left the outliners alone to analyze eigenvalues for the entire dataset
* By taking the categorical columns and making them factored, I was able to utilize the boxplot tool to have accurate information that identify outliers. Conduct PCA to classify the dataset, and lastly to visualize the eigenvalues and variance presented in the data.
* Utilizing the previously discussed R coding to remove all NA’s presented, removing irrelevant information, and factoring the categorical columns, I was able to conduct PCA and utilize the boxplot tool accurately displaying the dataset in its entirety.

**Part D5. Clean Data**

Table

Description automatically generated





**Part D6. Limitations**

* The missing values is the first limitation when working this raw dataset. I had two options, I could of filled in those null values randomly, but that would complete negate the overall findings, therefore since I could not obtain the missing information my only option was to delete those rows and columns.
* Exact values were not accurate nor preserved with the boxplot tool results, if I were to combine boxplots with other visuals tools along with being provided information of variables, I was uncertain of, would have led to a more thorough analysis of the data.
* Not having a background or medical knowledge greatly limited me when analyzing the dataset, if I had the knowledge base, I would have had better rational for outliers and possibly could of discovered information not bias to my lack of knowledge or skewing data that could of contributed to my overall analysis. Not having the ability to contact the originator of the dataset greatly limited me determining what working through the mitigation of data.

**Part D7. Impact of Limitations**

* Not having an understanding nor knowledgeable medical background of this data, I was unable to clean the data accurately. If I was able to discuss all the anomalies found during the cleaning process with someone who originated it, I could have fixed a lot of the information to have a more thorough accurate final dataset.
* Due to my ignorance with dataset given, models and the clean data produced are limited.
* Not having a source to verify that the mitigations performed are correct, is a direct impact of limitations.
* A plethora of variable are not clearly defines.

**Part E1. Principal Components**

Table

Description automatically generatedA picture containing text

Description automatically generated

**Part E2. Criteria**

* In order to utilize PCA I needed to change all the values to numerical, I was able to standardize the values afterwards.
* Taking the large dataset and being able to condense it to a much smaller dataset while trying to preserve as much information as possible.
  + Step 1 : Isolated numerical data
  + Step 2 : Standardize data
  + Step 3: created boxplot
  + Step 4: Identify eigenvalues
  + Step 5: Explore Feature Vector (*Principal component analysis for dimensionality reduction (2019, May, 24)).*

* Following lesson 6 the data was normalized. Identified components using Kaiser’s rule that have eigenvalues of at least 1. Added additional Gradient Styling to the PCA. *(Psychological Methods © 2016 American Psychological Association 2017, Vol. 22, No. 3, 450 – 466).*
* To answer my original question, I chose my components based off my research question focusing on discovering a relationship between various health determinants and medical complication risks. My analysis indicates if patients list multiples of the PC’s identified, they have higher medical complication risks. To summarize, the more PCs listed in part E1 have a direct relationship with medical complication risks. Some components push the threshold higher (PC1-HighBlood), but when stacked with multiple components, is causes significant change in the overall risk.

**Part E3. Benefits**

The resulting eigenvalues analyze the numerical values created from the PCA’s identified in part E1. Principle Component 1 had a majority of variance in the dataset. Given this indication, if the organization that produced this data focuses on the first principle component variables, they can articulate further in depth analysis. Principle Component 1 variables listed below. (rdocumentation.org). *(Principal component analysis for dimensionality reduction (2019, May, 24))*

Table

Description automatically generated

**Part F. Video**

Panopto video link:

<https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=e4cd1306-342e-493c-ada4-af930164ef16>

**G-Sources for Third-Party Code & H-Sources**

*RStudio: (2023, January, 19)*

*https://en.wikipedia.org/wiki/RStudio*

*r-lang.com: (2012, February, 16)*

*https://r-lang.com/view-function-in-r https://r-lang.com/view-function-in-r*

*https://r-lang.com/category/r-charts/*

*Principal component analysis for dimensionality reduction (2019, May, 24) https://towardsdatascience.com/principal-component-analysis-for-dimensionality-reduction-115a3d157bad*

*www.geeksforgeeks.org: (2020, June, 03)*

*https://www.geeksforgeeks.org/calculate-the-sum-of-matrix-or-array-columns-in-r-programming-colsums-function/*

*colSums: Form Row and Column Sums and Means (2019, December, 31)*

*https://www.rdocumentation.org/packages/base/versions/3.6.2/topics/colSums*

*kdnugget: (2016, November, 16)*

*https://www.kdnuggets.com/2019/03/top-r-packages-data-cleaning.html*

*Precisely.com: (2022, November, 14)*

*https://www.precisely.com/blog/data-quality/data-quality-dimensions-measure#:~:text=How%20can%20you%20assess%20your,timeliness%2C%20validity%2C%20and%20uniqueness.*

*rdocumentation.org: (2022, October, 19)*

*https://www.rdocumentation.org/packages/FactoMineR/versions/2.4/topics/PCA*

*datamentor.io: (2018, March, 13)*

*https://www.datamentor.io/r-programming/box-plot/#:~:text=In%20R%2C%20boxplot%20(and%20whisker,numeric%20vectors%20as%20its%20components.*

*{bibliography}*

*Chantal D. Larose, & Daniel T. Larose. (2019). Data Science Using Python and R. Wiley*

*{bibliography}*

*Psychological Methods © 2016 American Psychological Association 2017, Vol. 22, No. 3, 450 – 466*