Raquel Ocasio

D206 – Data Cleaning

September 17, 2023

Western Governors University

Part A

Given the available data, can I determine how many readmitted patients were employed full-time and had at least one child in the household?

Part B

Variable Name	*Data Type*	*Description*	*Example*
CaseOrder	Numerical	Indicates the order of the record in the dataset	1
Customer_id	Categorical	A unique patient identifier	C412403
Interaction	Categorical	A unique identifier related to procedures, transaction, and admissions of the patient	8cd49b13-f45a-4b47-a2bd-173ffa932c2f
UID	Categorical	A unique identifier related to procedures, transaction, and admissions of the patient	3a83ddb66e2ae73798bdf1d705dc0932
City	Categorical	City of residence for the patient	Eva
State	Categorical	State of residence for the patient	AL
County	Categorical	County of residence for the patient	Morgan
Zip	Categorical	Zipcode of residence for the patient	35621
Lat	Numerical	Latitude portion of GPS coordinates for the patient's residence	34.35
Lng	Numerical	Longitude portion of GPS coordinates for	-86.73

Variable Name	*Data Type*	*Description*	*Example*
		the patient's residence	
Population	Numerical	The population within a one mile radius of the patient	2951
Area	Categorical	The type of living area, related to population density	Suburban
Timezone	Categorical	The timezone of the patient's residence	America/Chicago
Job	Categorical	The job of the patient or primary insurance holder	Psychologist, sport and exercise
Children	Numerical	How many children are in the patient's household	1
Age	Numerical	Patient's age	53
Education	Categorical	Highest degree earned by the patient	Some College, Less than 1 Year
Employment	Categorical	Patient's employment status	Full Time
Income	Numerical	Patient's annual income	86575.93
Marital	Categorical	Patient's marital status	Divorced
Gender	Categorical	Patient's choice to identify as male, female, or nonbinary	Male
ReAdmis	Categorical	Whether the patient was readmitted within a month of release or not	No

Variable Name	*Data Type*	*Description*	*Example*
VitD_levels	Numerical	Measurement in ng/mL of the patient's vitamin D levels	17.8
Doc_visits	Numerical	During initial hospitalization, number of times the primary physician visited the patient	6
Full_meals_eaten	Numerical	While hospitalized, the number of full meals consumed by the patient	0
VitD_supp	Numerical	Number of times the patient was given vitamin D supplements	0
Soft_drink	Categorical	Indicates if the patient drinks three or more sodas a day.	NA
Initial_admin	Categorical	How the patient was admitted to the hospital	Emergency Admission
HighBlood	Categorical	Indicates if the patient has high blood pressure	Yes
Stroke	Categorical	Indicates if the patient has had a stroke	No
Complication_risk	Categorical	Assessment by a primary physician of the patient's complication risk	Medium
Overweight	Categorical	Indicates if the patient	No

Variable Name	*Data Type*	*Description*	*Example*
		is considered over-weight	
Arthritis	Categorical	Indicates if the patient has arthritis	Yes
Diabetes	Categorical	Indicates if the patient has diabetes	Yes
Hyperlipidemia	Categorical	Indicates in the patient has hyperlipidemia	No
BackPain	Categorical	Indicates if the patient has chronic back pain	Yes
Anxiety	Categorical	Indicates if the patient has anxiety disorder	Yes
Allergic_rhinitis	Categorical	Indicates if the patient has allergic rhinitis	Yes
Reflux_esophagitis	Categorical	Indicates if the patient has reflux esophagitis	No
Asthma	Categorical	Indicates if the patient has asthma	Yes
Services	Categorical	Indicates the type of service the patient received while hospitalized	Blood Work
Initial_days	Numerical	Indicated how many days the patient stayed in the hospital during the initial visit	10.59
TotalCharge	Numerical	Indicates the daily amount charged to the patient	3191.05
Additional_charges	Numerical	Indicates the average	17939.4

Variable Name	*Data Type*	*Description*	*Example*
		amount charged for miscellaneous procedures, treatments, etc to the patient	
Item1	Categorical	Survey response regarding timely admission	3
Item2	Categorical	Survey response regarding timely treatment	3
Item3	Categorical	Survey response regarding timely visits	2
Item4	Categorical	Survey response regarding reliability	2
Item5	Categorical	Survey response regarding options	4
Item6	Categorical	Survey response regarding hours of treatment	3
Item7	Categorical	Survey response regarding courteous staff	3
Item8	Categorical	Survey response regarding evidence of active listening from doctor	4

Part C1

The quality of the data will be assessed by examining the data profile, then detecting and treating duplicates, missing values, outliers, as well as re-expressing any variables with inconsistent presentation by using the Python programming language.

Before any treatments were applied, the profile of the data was examined by using the print and .info() functions.

The print and duplicated().value_counts functions were used to detect any duplicate values.

To detect missing values, the print and isnull().sum() functions were used.

To detect categorical variables with inconsistent presentation for re-expression, the print and .unique() functions were used.

The functions used for the detection of outliers were stats.zscore(), plt.hist(), plt.title(), plt.xlabel(), plt.ylabel(), and plt.show() functions.

Part C2

The print and .info() functions were used to profile the data. These functions produced a table that included the total number of rows, column names, number of non-null values, and the data types. This information was used to determine the detection and treatment methods for each variable.

The print and duplicated().value_counts functions were used to create a sum count of any duplicate records in the dataset.

The print and isnull().sum() functions were used to detect missing values because they produce a count of missing values for each variable.

The print and .unique() functions were combined to display a list of unique values in a categorical variable. The results were compared to information from the medical data dictionary to determine which variables needed to be re-expressed.

The detection of outliers was a two-step process. The first step was to calculate the z-score using the stats.zscore() function. Then a histogram was created using the z-score and the plt.hist() function. A title, x-axis label, and y-axis label were added to the histogram using the plt.title(), plt.xlabel(), and plt.ylabel() functions, respectively. Finally the histogram was displayed using the plt.show() function.

Part C3

The Python programming language was used to clean the data. Python was selected because of its ease of use, simple syntax, as well as the variety of packages available for data cleaning and analysis.

The Pandas library was imported to be able to access the .info() function for profiling the data.

The pylot package from the Matplotlib library was imported to be able to access the functions needed for creating histogram plots.

The numpy package was imported to be able to perform calculations needed for principal component analysis.

The sklearn.decomposition package was imported to access the PCA library for principal component analysis.

Part C4

See code attached.

Part D1

The data set does not have any duplicate values.

The following variables had missing values. The number of missing values is listed after the variable name.

Variable Name	***Number of missing values***
Children	2,588
Age	2,414
Income	2,464
Soft_drink	2,467
Overweight	982
Anxiety	984
Initial_days	1,056

Two categorical variables did not present values consistently. The Overweight and Anxiety variables presented the data with numerical values instead of Yes/No values.

The following variables had outlier values. The number of outliers, and the value range for the outliers is listed after the variable name below.

Variable Name	***Number of outliers***	***Value range of outliers***
Population	218	54,453 to 122,814
Children	303	8.0 to 10.0
Income	180	114,215.99 to 207,249.13
VitD_levels	500	40.8416712 to 53.01912416
Doc_visits	8	1 to 9
Full_meats_eaten	33	5 to 7

Variable Name	*Number of outliers*	*Value range of outliers*
VitD_supp	70	3 to 5
TotalCharge	276	16,053.46288 to 21,524.22421

Part D2

The dataset did not have duplicate values so no duplicate value treatment was necessary.

Variables with missing values were treated with Univariate Imputation. A histogram of the variable was created for each variable. Based on the distribution observed in the histogram, the appropriate imputation was performed. The imputation was verified by confirming that the variable no longer has missing values. Then a new histogram was created to confirm that the distribution of the variable after imputation was in relative alignment to the distribution of the variable before imputation. (Kumar, 2022)

For the categorical variables, the values were re-expressed before applying the treatment above. These variables were expressed to create numerical values that would be used in the above calculations.

For variables with outlier values, the values were retained. The values were retained because the outlier values were found to be expected and acceptable.

Part D3

The treatment for missing values reduced the number of missing values for numerical and categorical variables to zero. It also created additional columns to show the numerical values for categorical variables that were re-expressed.

The information produced by the detection of outlier values caused those values to be retained.

Part D4

See attached code.

Part D5

See attached file named “medical_raw_data_cleaned.csv”

Part D6

For the treatment of missing values, a limitation of Univariate Imputation is that it could distort the data and/or distribution of the data.

For the treatment of outlier values, a limitation of retaining the values is that the dataset may not contain values that are acceptable or appropriate.

Part D7

The limitation of the missing value treatment could affect analysis outcomes that depend on the distribution of the data, such as histograms.

The limitations of the outlier treatment could affect analysis outcomes that depend on the values, such as the mean or median.

Part E1

The principal component analysis produced five principal components. A screenshot of the output of the principal components is below.

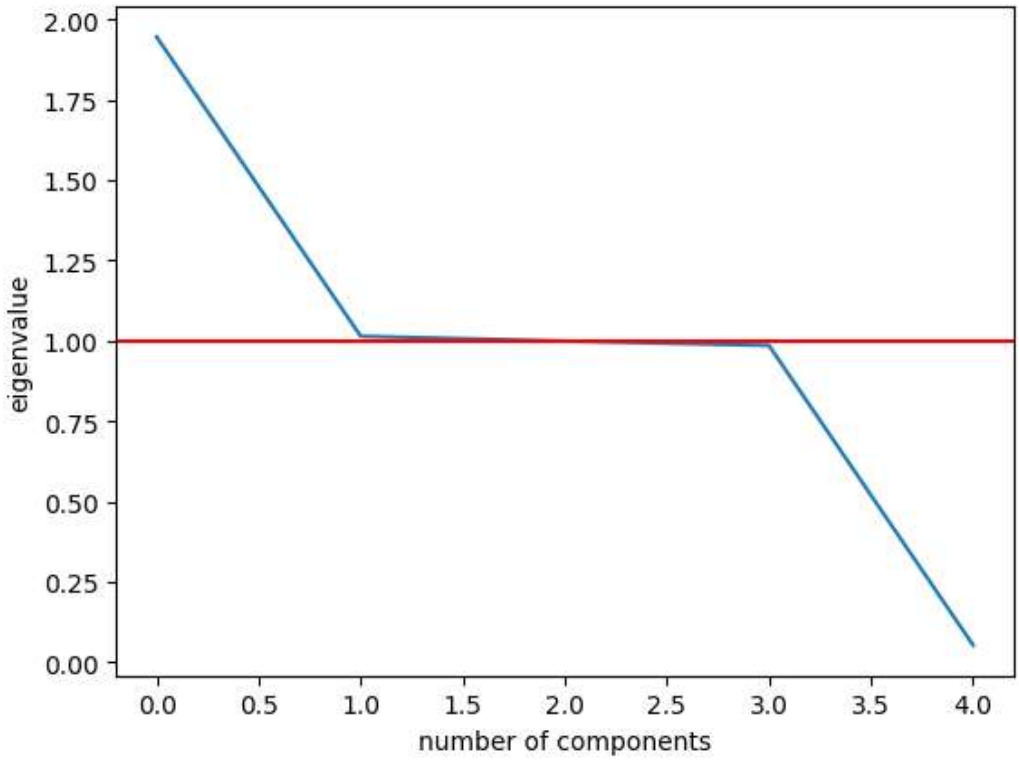
	PC1	PC2	PC3	PC4	PC5
Income	-0.006704	0.199350	0.966018	-0.164389	-0.001042
VitD_levels	0.544446	0.446406	-0.010167	0.455943	0.544346
Initial_days	0.450135	-0.570290	0.033296	-0.517132	0.451236
TotalCharge	0.707128	-0.006515	0.006546	0.006215	-0.706998
Additional_charges	0.029704	0.660082	-0.256038	-0.705427	0.015054

Part E2

The Kaiser Rule was used determine which principal components should be retained. The eigenvalue was calculated for each principal component, and the components with an eigenvalue greater than or equal to one were retained.

The eigenvalues were saved to a dataframe and inspected. The most important principal components are PC1 and PC2, with values of 1.9457323614540885 and 1.014538331865836, respectively.

A screenshot of the scree plot for the eigenvalues is below.



Part E3

The organization could benefit from PCA by using the most important principal components to understand the connection between the variables of the data.

Part F

See attached video link.

Part G

Kumar, A. (2022, September 29). *How to add a matplotlib title*. Scaler Topics. https://www.scaler.com/topics/matplotlib/matplotlib-title/

Part H

Not applicable