Part I

A. Question

Can principal component analysis be used to identify preexisting conditions which correlate with the patient's length of stay?

B. Description of Variables

In the medical data CaseOrder, Zip, Population, Doc_visits, Full_meals_eaten, VitD_Supp, Item1, Item2, Item3, Item4, Item5, Item6, Item7, and Item8 are all non-null, integer data type. These fields are numeric (example: Zip is 78229 or 80134, Doc_visits counts the number of visits, 1, 2 so on and so forth). All of the proceeding fields are entered as whole numbers, so integer is the appropriate way to store this data.

Of this set, CaseOrder, Zip, and Item1 through Item8 are qualitative, categorical data. There is a limited selection for what values can be placed in all of these columns, which makes the data categorical and the value of these fields is assigned based on the list of available inputs, not the numeric value (qualitative). To expand on this, the CaseOrder is dependent on the number of rows. Each value for CaseOrder is based on the values around it, not the numerical value. We could theoretically start the case order at 10 or 100 and the meaning of the field would remain unchanged. For a quantitative variable this does not hold true. If all values in the Doc_visits field were increased by 100 there would be a substantial change in the value of this field. Item1-8 are scored 1 through 8. Zip codes are groupings based on area. We have assigned these fields value beyond the numeric value. This supports that these fields are qualitative values. For CaseOrder and Item1 through Item8, the values have an order, for example a rating of 1 is worse than 8, so these fields are ordinal. Zip codes codes are nominal, while some values are grouped by geographic region, these groupings do not convey a direction.

On the other hand, Population, Doc_visits, VitD_Supp, and Full_meals_eaten are discrete, quantitative numerical values. These values can only be a whole number which makes the values discrete and not continuous. For example, we would not have half a doctor visit or half a person for that matter. These fields are quantitative because the number conveys value to the field. A value of 1 for the Doc_visit field conveys that there was 1 doctor visit, 1000 in the Population field indicates 1000 people. The same is true for VitD_Supp and Full_meals_eaten.

The Customer_id, Interaction, UID, City, State, County, Area, Timezone, Job, Education, Employment, Martial, Gender, ReAdmis, Soft_drink, Initial_admin, HighBlood, Stroke, Complication_risk, Arthritis, Diabetes, Hyperlipidemia, BackPain, Allergic_rhinitis, Reflux_esophagitis, Asthma, Services, columns are non-null, categorical, object data types. These are categorical data types because there is a limit to the allowable inputs we have to choose from and the input conveys meaning beyond the value of the input. Customer_id is a unique mix of numbers and letters used to represent each individual, since each individual is limited to one Customer_id there is a limited amount of values to chose from, which fits the definition for categorical data. Similar is true for Interaction and UID. While these values are unique to each interaction, each assigned value is used to separate the interaction from another and there allowable values to fill this field are limited to the number of interactions. There is a limited number of Cities, States, Countries, areas, and time zones for us to input values from, so these also fall under the definition of categorical variable. For Education we have a selection of options to represent the education level of the individual such as "Associate's Degree". While there are many inputs in the Job field, this field is still limited to the possible career fields in the world (vast as these options may be). There are a limited amount of marital statuses to input to the Marital field. Gender is limited to "Male", "Female" and "Prefer not to answer". Intial_admin is assigned based on how the patient was admitted to the hospital. Complication_risk

values are limited to high, medium and low. ReAdmis, Soft_drink, HighBlood, Stroke, Arthritis, Diabetes, Hyperlipidemia, BackPain, Anxiety, Allergic_rhinitis, Reflux_esophagitis, and Asthma, are limited to Yes/No. Finally the Services field is limited to the primary service that the patient used while admitted.

The Lat, Lng, Children, Age, Income, VitD_levels, Overweight, Initial_days, TotalCharge, Additional_charges, columns are non-null, float data types. Other than the Overweight column (which is boolean) these columns are numerical data. The Lat, Lng, Income, VitD_levels, Initial_days, TotalCharge, and Additional_charges values are continuous variables as they can technically be infinitesimal. Since these values can have a decimal point, storing them as float is appropriate. Age is only displayed as discrete whole numbers so it would be more appropriate to assign the data type of float. Children is also discrete, as one cannot have half a child, so we can identify that this field to be changed to the integer data type. For the Overweight column, if we need to work with this field we could turn this into a true boolean value by changing the 0.0 to False and 1.0 to True and the data type to boolean or change the 0.0 to No and 0.1 to Yes and the data type to float. If needed for the analysis, these fields will be examined later in the cleaning section.

Part II

C.1. Plan

To start this project, the data will be examined for duplicated data. If present, duplicated data will be removed. The data will then be assessed for misrepresented data by looking at the unique values in each column. Any misrepresented data, such as misspellings, unique capitalization will be corrected. Then the data will be assessed for missing, or null data. Based on the data these values may be removed or replaced with an appropriate stand in (mean, median, mode). Following this, the data will be examined for outliers and these outliers will be addressed. These values may be recalculated, removed or replaced, depending on the nature of the outlier.

C.2. Justification

To begin to assess the data, all data will be assessed for duplication. All the data needs to be included to assess for true duplicated information. It is important to assess for duplication because duplication could impact the level of correlation between variables and will result in over representation of the duplicated data.

Next, the categorical data will be reviewed for uniqueness. This is appropriate because categorical data needs to be assigned to matching values, like Yes or No. Alterations in capitalization and spelling will impact whether a value is assigned to a matching category or not. If this is not addressed it may result in having to analyze too many categories, obscuring the relationship to other variables.

All the data needs to be assessed for missing inputs because too much missing data may impact correlations and future calculations. Missing data can cause too much of the data to be excluded for in calculations. Just like the above issues, this could lead to variables with false levels of correlation.

Outlying data points will be the last point of assessment before proceeding to PCA analysis. Outliers need to be removed so that calculations are not skewed. Age, VitD_levels and Initial_days are quantitative data and as such are appropriate to look at for outliers.

C.3. Programming Language Justification

I chose to write my code in python because I wanted more practice with python. While going through the lessons I felt that python was more challenging for me to learn. For this reason, I wanted to challenge myself so that I would be able to become more comfortable with python. I also chose to do

this project in python because python can be used for so many different areas. By being well versed in python I feel that I will be more competitive in my field and I will be able to apply this to many future projects. For packages I used pandas, numpy and os as standard packages. I also used missingno to visualize the missing data and matplotlib.pyplot to create histograms. The PCA feature from sklearn.decomposition and seaborn will be used to create the PCA analysis.

C.4. Annotated Code

See D206_Wood_Annotated_Code.exe

Part III

D.1. Findings

To begin, the data was screened for duplication and the data appears to have no duplication so no mediation is needed for this. Review of unique data entries does show some data that could be cleaned, such as the Timezone data which shows multiple names for the same time zone. An example of this is America/Denver and America/Boise which should be combined into Mountain Time. Job is another field that could have unique entries reduce, values such as 'Clinical psychologist' and 'Psychologist, counseling' could be combined into Psychologist.

There is data missing in a number of columns. These columns are Children, Age, Income, Soft_drink, Overweight, Anxiety and Initial_days. While cleaning it was also discovered that the Initial_days field has a split that correlates to the patient's readmission status. It was also noted that the VitD_levels data had a significant split which will be addressed in the below, Justify Mitigation section. Finally, there were some outliers noted as well in the Income and VitD_levels data.

D.2. Justify Mitigation

As stated above, no duplicated data was found so no mitigation was needed at this step. The Employment and Job column could be cleaned to reduce the number of unique entries. Since neither of these fields are continuous, numeric values they would not be appropriate for PCA. Since our question is interested in whether or not PCA is an appropriate tool to examine the length of admission we do not need to address these unique values. So, I will forgo cleaning this.

After assessing the data for missing data, it becomes clear that there is a significant amount of missing data in the Age, Income (socio-economic status is a predictor of disease, so it can be considered a pre-existing condition), Soft_drink, Overweight, Anxiety and Initial_days columns. As this will impact the results, I will address this missing data. To begin addressing this missing data, I created histograms of all the columns with missing data. The histogram for Age appeared to be bimodal so the median was used to fill missing data. The Income histogram was skewed so the median was used to input missing data. The Soft-Drink, Overweight, and Anxiety data are categorical data, so I applied the mode to the missing values. Initial_days was found to be bimodal upon examination of the histogram. I suspected that readmission status may factor into this split. I changed ReAdmis to float data type and then created a scatter plot of ReAdmis and Initial_days. It was clear that patients who were readmitted had longer hospital stays. As the question is specifically interested in the length of stay, I calculated the median initial days for patients who were not readmitted and then calculated the median stay for patients who were readmitted. I then filled these values in based off readmission.

Finally I reviewed the boxplots for Age, Income, VitD_levels and Initial_days. Income had outliers which were removed. VitD_levels demonstrated that the data was split. Upon review it appears that the data is reported in ng/ml and nmol/l as the low levels would be extremely low for nmol/L and the high levels would be very high for ng/ml. The values below 30 are recalculated to be expressed as

nmol/L. The resulting boxplot had outliers which were removed. Age and Initial_days had no outliers so no changes were made.

D.3. Outcome of Each Step

The initial step of reviewing for duplicated data resulted in no findings so no changes were made to the data. The next step of reviewing the data for unique entries did have findings. These findings were in the Job and Time_zone but these fields will not impact the data needed for this question so no changes were made.

To review the missing data I created a count of the missing values in each column and then created a matrix of the missing data to verify that there was not a pattern to the missing fields. At this point it appears that the data is missing at random. Missing data was found in the Age, Income, Soft_drink, Overweight, Anxiety and Initial_days columns. These fields were chosen for cleaning because they are considered preexisting conditions. I created a histogram for each Age, Income and Initial_days to examine the data shape. Measures of central tendency were assigned to the missing data based on the shape and type of data. As mentioned above, Age was bimodal. As the data is numerical I chose the median rather than the mode to input for the missing values. Income was skewed so the median was assigned to the missing values. The remaining fields contain categorical data, so the mode was used to input missing data. Initial_days was bimodal. I was suspicious of this because I would expect a skewed dataset with a majority of the stays being short. To examine this inconsistency, I converted the ReAdmis data into float type and exchanged ‘Yes’ for 1 and “no’ for 2. This showed there was a clear split in the data with those identified as Readmitted having far greater Intial_days. Upon referring back to the initial histogram, it is clear that both the high and low grouping are skewed. I calculated the median for the Initial_days of patients who were readmitted and for those patients who were not readmitted. I then assigned the median values to the ReAdmis field, 10.360185 for 0 (the value which I had previously exchanged ‘No’ for) and 64.276773 for 1 (the ‘Yes’ value) I then swapped these median values for the missing values in the Initial_days field. This step results in no missing data in the above columns.

Finally I created a boxplot of Age, Income and VitD_levels. Age had no outliers, so no values were changed. The boxplot for Income showed outliers. So I used the describe function to find the upper and lower quartiles, I used these to calculated the interquartile range, which I multiplied by 1.5 and added to the third quartile value to find the upper whisker value of 106011.89875. To remove the outliers I dropped all values greater than this. Then I created a new boxplot, with the remaining data. This boxplot showed some newly created outliers. I could have removed these, but decided not to to preserve the amount of data available. VitD_levels demonstrated a large gap between an upper and lower group. Upon researching, I found that this gap was likely due to two units of measurement (Moyad). Using the formula found in Moyad’s article, I multiplied the values below 30 by 2.5 to convert all values to nmol/L. I then created a boxplot of the new values for VitD_levels and used the same method as outlined above to find the upper whisker. To find the lower whisker, subtracted the interquartile range from the first quartile value. I then dropped all values above and below these values to remove the outliers. I then created a new boxplot which showed some outliers in the new data set. I again left these to preserve the amount of data.

D.4. Annotated Code

See D206PA.py

D.5. Cleaned Data

See Medical_data_clean.csv

D.6. Limitations

There are some limitations with this data set. The most influential limitation is to the Initial_days and VitD_levels data. Ideally I would verify that the Initial_days column was not including days the patient was out of the hospital. The data could just be date of initial intake to date of final discharge, which would include days that readmitted patients were not in the hospital. I would ideally also put the number of days that the patient was in the hospital upon readmission in a separate column. As it is now, it is impossible to say whether or not the Initial_days column includes the days the patient was readmitted for.

For the VitD_levels it would be ideal to verify that the levels were measured on two different scales. Looking at the data we can see a large split between an upper and a lower cluster of data but there could be other factors that create this effect.

Beyond these limitations with the dataset itself, there were some limitations imposed by cleaning and mitigation. During the cleaning phase, I used measures of central tendency to replace missing data. This can artificially inflate the mean, median or mode. This could cause a false correlation. I was unable to find any implications that data was missing not at random but it it were, this could cause a false findings. For example if all the Ages for newborns were missing and I input the median for these values, this would affect all future analysis of this field. Another example of this is the Initial_days field. I chose to divide this field into readmitted and not readmitted patients, there may have been a better way to divide this field. This different division may have resulted in different values for the median. Even if all data was missing at random and I chose the very best measure of central tendency for each data set, by exchanging the measure of central tendencies for missing values, the Quartile measures are altered. By adding to the entries which carry the value of the median, some data that may have not been previously excluded, may be excluded as an outlier.

D.7. Impact of Limitations

Since the question posed is looking at length of stay the limitation outlined above for the Initial_days column is significant. For our question, this is the column we are most interested in. While I moved forward with principal component analyis, in the real world it would be imperative to ascertain how the values for the Initial_days column are calculated. We would need to know if this field is including days that the patient was out of the hospital as well as if this field is the total number of days hospitalized or if it is just the days the patient was hospitalized this admission.

Verifying that the VitD-levels do contain two units is also very important. I increased the low values, however if these low values are actually just increadibly low values I may have removed valuable information that may correlate with the length of patient admission.

As mentioned above using the measures of central tendency to input missing values will increase the number of findings with the median value which may cause false levels of correlation. In the example above I outlined an example where all newborns were missing the Age. In this case it would be ideal to pull birth dates from the hospital and fill in the Age with the actual value. As this was not available to us, it cannot be done. By exchanging the median for these missing values in this scenario it would certainly shift the data set towards the higher values.

E.1. Principal Components

The following numeric variables were identified as preexisting conditions that may impact patient length of stay (Initial_days), Age, VitD_levels, and Income. Socioeconomic status is a well documented indicator of health, so it is appropriate to include in the list of preexisting conditions. Ideally, PCA should be done on continous, numeric values, but for the sake of this analysis, I converted HighBlood, Stroke, Overweight, Arthritis, Diabetes, Hyperlipidemia, BackPain, Anxiety, Allergic_rhinitis, Reflux_esophagitis and Asthma to numeric values with No set to 0 and yes set to 1.

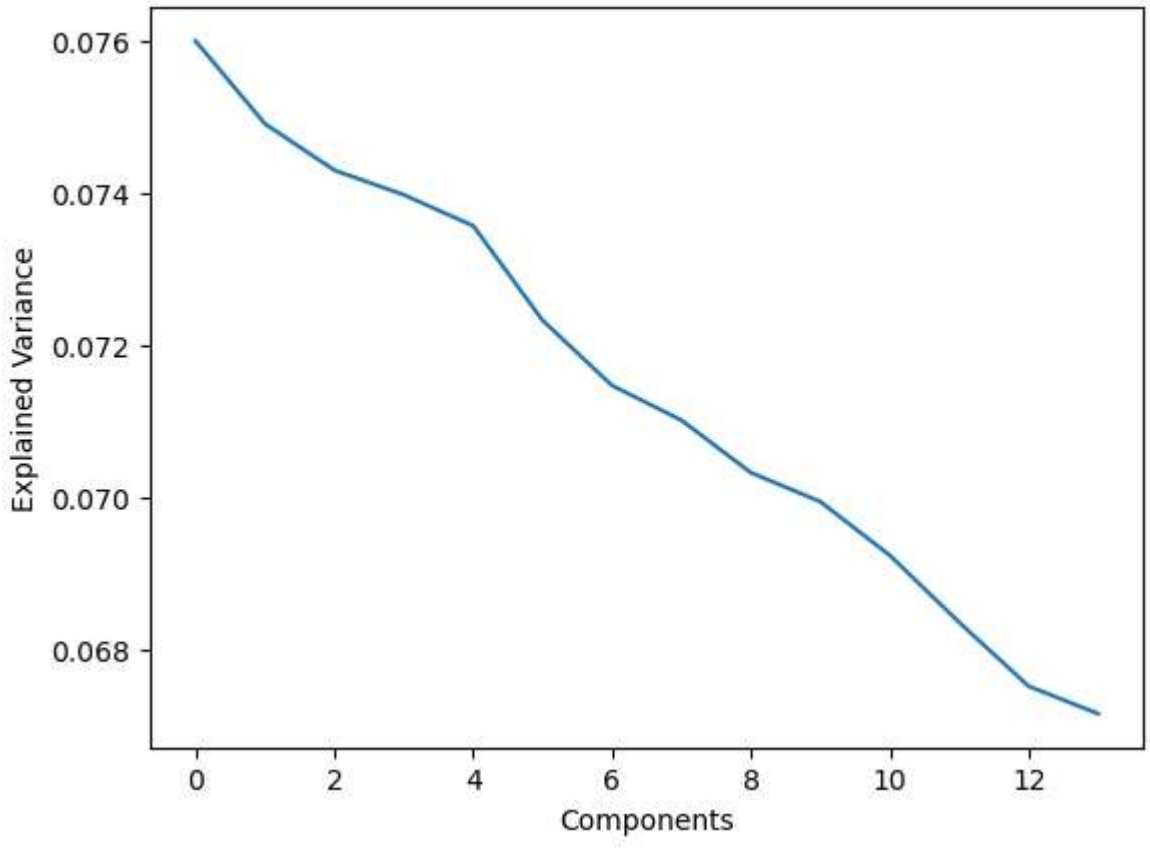
After identifying these variables of interest the data was normalized and the following PCA loadings matrix resulted:

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11	PC12	PC13	PC14
Age	-0.296499	-0.150409	0.165503	0.363858	0.377620	0.160764	-0.273762	0.321387	0.162705	0.012442	-0.418547	0.072921	0.086948	-0.410266
VitD_levels	-0.310505	-0.038272	-0.322184	-0.310577	0.196809	0.171251	-0.488367	0.131505	0.019231	-0.066232	0.456514	-0.133226	-0.379694	-0.054096
Income	0.218109	0.412669	0.225934	0.073237	-0.354173	0.264160	-0.102255	0.249607	0.252426	0.350400	-0.019402	0.230428	-0.465060	-0.038612
Initial_days	0.139771	-0.337000	0.443749	-0.095673	0.185878	-0.044555	0.047005	0.098244	-0.265386	0.559606	0.425994	-0.022278	0.106967	-0.188239
Overweight	-0.473215	0.077190	0.043871	-0.152116	-0.273367	-0.393881	0.071663	-0.007007	-0.386368	-0.034560	-0.057007	0.488035	-0.157340	-0.305502
HighBlood	-0.397387	0.124574	0.139583	-0.087339	-0.297302	0.143150	0.330134	0.525334	-0.088313	-0.051240	0.029101	-0.518523	0.145136	0.066385
Stroke	-0.101659	0.291987	0.006826	0.095113	0.437075	0.458105	0.365719	0.071024	-0.236151	-0.146162	0.223483	0.425513	0.060593	0.209328
Arthritis	-0.081564	-0.588975	0.107264	-0.043090	-0.277073	0.303544	-0.192051	0.101840	-0.175135	-0.037432	-0.208347	0.250161	-0.069326	0.523509
Diabetes	0.076494	-0.130211	-0.071416	0.649425	-0.157497	0.128901	0.068237	-0.171179	-0.419532	-0.153394	0.148445	-0.260270	-0.388839	-0.185615
Hyperlipidemia	0.322407	-0.011408	-0.036547	0.246101	-0.036896	-0.404483	-0.179705	0.596565	-0.007063	-0.340125	0.302739	0.203504	0.138843	0.113110
BackPain	-0.103422	0.171024	0.572113	-0.011351	0.339403	-0.333517	-0.094797	-0.091391	-0.052266	-0.181787	-0.118462	-0.203256	-0.389543	0.382069
Allergic_rhinitis	-0.293263	-0.315171	0.161615	0.187682	-0.093211	-0.066807	0.331522	-0.145304	0.644768	-0.139786	0.369503	0.148379	-0.123137	-0.013101
Reflux_esophagitis	0.238664	-0.014479	0.436394	-0.320110	-0.156564	0.310495	-0.117136	-0.121012	-0.027133	-0.576856	0.068962	0.007061	0.100735	-0.390107
Asthma	-0.293726	0.304213	0.193850	0.307928	-0.230644	0.055712	-0.468576	-0.294481	-0.015505	0.091493	0.260975	0.011008	0.467680	0.175024

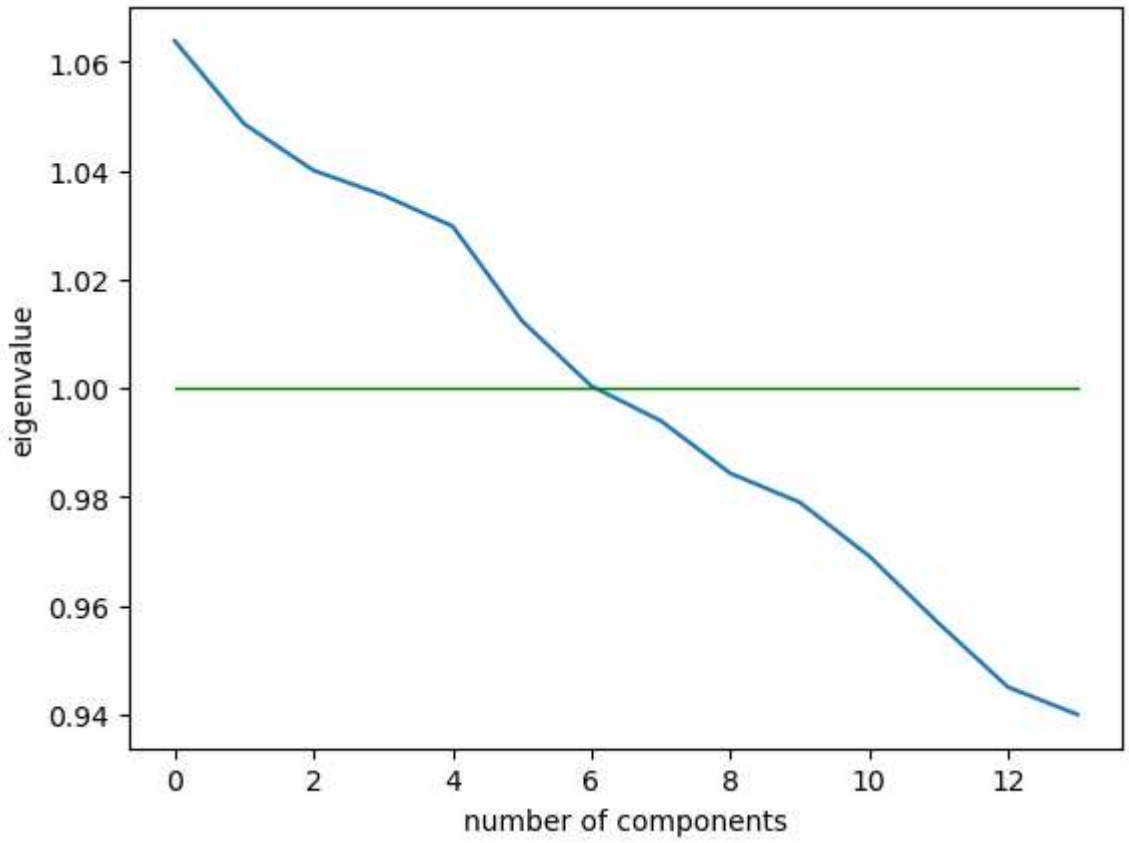
E.2. Justify Principal Component Reduction

Upon graphing the Eigen values, using the Kaiser Criterion, we can see that six components explain the majority of the variation in this data. The Kaiser criterion allows us to include all components greater than one. I have added a line at one on the Eigen Plot to illustrate this. As we can see components zero through five are above one. The graph falls below one at component six. The first six principal components can be retained and the remaining 8 should be discarded. Of the first six principal components Initial_days has the highest contribution to PC3. BackPain and Reflux_esophagitis also contribute highly to PC3, these ffactors should be further investigated for potential correlation with length of admission.

Scree Plot



Plot of Eigen Values



E.3. Benefit

The principal component analysis suggests that we should further examine the relationship between back pain, reflux esophagitis and admission length. We can see that Initial_days has the largest contribution to the variation in principal component 3 (PC3). In this model back pain is the factor with the highest contribution to the model, followed by reflux esophagitis. This suggests that these factors may have a correlation to the length of stay (Initial_days). From this finding we may want to look at other variables that may change with back pain, such as activity level or use of narcotic pain medications.

PCA provides a benefit to the hospital because it allows us to simplify complex data, which allows us to identify areas that may need further investigation. With out PCA, we would need to gather data on every aspect of disease to identify variable that potentially correlate with length of admission. In addition, these finding allows the hospital to identify areas for potential intervention to shorten the length of admission.

Part IV

F. Pantopo Video

https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=6052a5d3-198f-49d0-a324-b1360165c2d9

G. Code Sources

"Indexing and Selecting Data — Pandas 1.5.1 Documentation." *Pandas.pydata.org*, NumFOCUS, INC, pandas.pydata.org/docs/user_guide/indexing.html.

Middleton, Keiona. "D206- Getting Started with D206 | Duplicates." *WGU.hosted.pantopo*, WGU.edu, wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=3bcc452f-fa35-43be-b69f-b05901356f95.

Middleton, Keiona. "D206- Getting Started with D206 | Missing Values." *WGU.hosted.pantopo*, WGU.edu, wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=3bcc452f-fa35-43be-b69f-b05901356f95.

Middleton, Keiona. "D206- Getting Started with D206 | Outliers." *WGU.hosted.pantopo*, WGU.edu, wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=3bcc452f-fa35-43be-b69f-b05901356f95.

Middleton, Keiona. "D206- Getting Started with D206 | Re-sxpression of Categorical Variables." *WGU.hosted.pantopo*, WGU.edu, wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=3bcc452f-fa35-43be-b69f-b05901356f95.

Middleton, Keiona. "D206- Getting Started with D206 | PCA." *WGU.hosted.pantopo*, WGU.edu, wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=3bcc452f-fa35-43be-b69f-b05901356f95.

H. Citations

Moyad, Mark. "Vitamin D: A Rapid Review - Page 8." *Medscape*, WebMD, 1 Jan. 2009, www.medscape.com/viewarticle/589256_8?form=fpf. Accessed 12 Mar. 2024.

I. Professionalism

This paper strives to be professional, thank you for taking the time to review this project.