**Part I: Research Question**

A.  Describe the purpose of this data mining report by doing the following:

1.  Can the information in the numeric variables of our dataset be reduced to a smaller number of features while capturing a majority of the information provided by these variables?

2.  The goal of this analysis is to use a smaller number of principal components to represent the initial features, achieving feature reduction. We hope to capture at least 90% of the variance in as few principal components as possible.

**Part II: Method Justification**

B.  Explain the reasons for using PCA by doing the following:

1.  PCA analyzes data by creating new features which are linear combinations of the original features. To achieve feature reduction, you can select a lower number of principal components than initial features while still capturing a high amount of the variance in the data. Principal components are organized such that the first principal component captures the most variance while each subsequent principal component captures less than the last. We expect a smaller number of principal components than initial features will capture a high proportion of the variance in the data set and allow for feature reduction.

2.  PCA assumes that data has been standardized. This assumption is due to PCA being sensitive to the value of the variance of each variable, so a variable with a greater range will overwhelm a variable with a smaller range unless they are scaled to a similar range.

**Part III: Data Preparation**

C.  Perform data preparation for the chosen dataset by doing the following:

1. PCA is designed to work with numeric data, not categorical. The numeric predictor variables we use for this analysis are Lat (Continuous), Lng (Continuous), Population (Discrete), Children (Discrete), Age (Discrete), VitD\_levels (Continuous), Doc\_visits (Discrete), Full\_meals\_eaten (Discrete), vitD\_supp (Discrete), Initial\_days (Discrete), TotalCharge (Continuous), and Additional\_charges (Continuous).

2.  The data is standardized using sklearn’s StandardScaler method and the resulting, scaled data is provided as the attachment OFM3\_scaled.csv.

**Part IV: Analysis**

D.  Perform PCA by doing the following:

1. Input:

loadings = pd.DataFrame(pca.components\_.T, columns=labels, index=data\_sel.columns)

loadings

Output:

Table

Description automatically generated

1. There are 12 total Principal Components, as shown in the scree plot below.

Input:

sum\_var = []

sum\_var.append(per\_var[0])

for i in range(1, len(per\_var)):

sum\_var.append(per\_var[i] + sum\_var[i-1])

plt.figure(figsize=(10,10))

sns.lineplot(x=range(1, len(per\_var) + 1), y=per\_var, color='r')

sns.lineplot(x=range(1, len(per\_var) + 1), y=sum\_var, color='g')

plt.xticks(range(1, len(per\_var) + 1))

plt.ylabel('% Explained Variance')

plt.xlabel('Principal Component')

plt.title('Scree Plot')

plt.grid(b=True)

plt.legend()

Output:

Chart, line chart

Description automatically generated

We see two distinct changes in the variance per principal component, in red, one elbow and one knee. The elbow occurs at PC4 and the knee occurs at PC9. While conventional wisdom tells us to select only the principal components before the elbow, the cumulative variance line, in green, shows that four principal components captures only approximately 50% of the variance while nine principal components allows the capture of approximately 90% of the variance. This information implies we should keep the first nine principal components.

1. Input:

for i in range(len(labels)):

print(labels[i] + ': ' + str(per\_var[i]))

Output:

A picture containing table

Description automatically generated

1. Input:

for i in range(len(labels)):

print(labels[i] + ': ' + str(np.round(sum\_var[i],decimals=2)))

Output:

Text

Description automatically generated

1. To achieve our desired variance captured of 90%, we must retain the first nine principal components. This allows us to reduce the twelve initial variables to nine features which are all linear combinations of the initial twelve features.

**Part V: Attachments**

E.  No web sources were used to supply code for this analysis and all code was written by the student.

F.  No works cited.