# Lab 6

Please submit your answers in this word document, including detailed discussion for each question with supportive screenshots of your analysis in SAS EM.

## Dimension Reduction with PCA

The Pollution.xlsx data set includes regional climate, pollution, and population demographic statistics from 1960 in the United States. Below are the descriptions for these variables.

|  |  |
| --- | --- |
| **Variable** | **Description** |
| PREC | Average annual precipitation in inches |
| JANT | Average January temperature in degrees F |
| JULT | Average July temperature in degrees F |
| OVR65 | Percent population aged 65 or older |
| POPN | Average household size |
| EDUC | Median school years completed by those over 22 |
| HOUS | Percent housing units, which are in good repair and with all facilities |
| DENS | Population per square mile in urbanized areas, 1960 |
| NONW | Percent non-white population in urbanized areas, 1960 |
| WWDRK | Percent employed in white collar occupations |
| POOR | Percent of families with income < $3000 |
| HC | Relative hydrocarbon pollution potential |
| NOX | Relative nitric oxides pollution potential |
| SO2 | Relative sulfur dioxide pollution potential |
| HUMID | Annual average % relative humidity at 1:00 pm |
| MORT | Total age-adjusted mortality rate per 100,000 |

1. How many components are required for us to explain at least 80% of variance. Provide a plot or a table from the output to support your claim.

We see that explained variance at least 80% of variance **at PC 6.**

A screenshot of a computer

Description automatically generated

A graph on a computer screen

Description automatically generated

1. Locate the loadings for the 2nd component and show PC2 in the form of linear combination of predictors.

A screenshot of a graph

Description automatically generated

1. Look at the pairwise scatter plots of the selected principal components, are any of the components highly correlated to each other? Is this expected? Please remember to include your first and last name in the footer of the graph.

There is **no component with highly correlation** to each other.

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

## Clustering

You will be using the German Credit Benchmark Data DMAGECR to understand the patterns in the customer clusters based on their attributes. The variable setting is as below. Note: there is no target variable in this analysis.

Graphical user interface, table

Description automatically generated

1. Is there any missing data in the data set? Provide appropriate screenshot to support your answer.

Using StatsExplore node, I found that there are **no missing values** in any columns in the dataset.

A screenshot of a computer

Description automatically generated

1. Prepare a clustering model, set the Selection Criterion Final Maximum to be 10 and Clustering Method to be Average. How many clusters are created? Provide a screenshot of the Segment Size Pie chart with your First and Last Name as footnote to support your answer.

A screenshot of a pie chart with Crust in the background

Description automatically generated

1. Provide a screenshot of the Input Means Plot with Segment 1 highlighted with your First and Last Name in the footnote. And then use the Input Means Plot to describe customers in Segment 1. Note: If the plot doesn’t show the correct axes labels or legend, use Data Options under Edit to edit the plot.

Graphical user interface, text, application

Description automatically generated Graphical user interface, application

Description automatically generated

We see that 4 out of 7 variables are used to cluster in segment 1 are having normalized mean 1. While amount and duration are nearer to zero normalized mean.

A screenshot of a computer

Description automatically generated

1. What are the top 3 most important variables for the decision tree considering the segment as the target variable? Provide the appropriate output to support your answer.

We see the top 3 most important variables are,

* + - 1. **Amount**
      2. **Duration**
      3. **age**

A screenshot of a computer

Description automatically generated

1. Prepare another clustering model, set the Selection Criterion Final Maximum to be 10 and Clustering Method to be Ward. How does the results differ in cluster numbers and top 3 most important variables? Provide screen shots to support your answer.

There are 4 clusters in the previous model and there are only 3 clusters used to cluster same dataset using this model.

We see the top 3 most important variables are,

* + - 1. **Duration**
      2. **Amount**
      3. **age**

A screenshot of a pie chart

Description automatically generated

A screenshot of a computer

Description automatically generated