DAPT 622 (Statistics II)
Spring 2019
Assignment #1

Due: By 11:59pm on Sunday, January 27

As a team, you are to perform the tasks listed in Parts I and II below. Organize your analyses and conclusions in a typed document (Word, Excel, PowerPoint, LaTex formats are all fine). Be sure to include all relevant software output in your submission.

This assignment is to be submitted electronically **by 11:59pm on Sunday, January 27** via Blackboard. All team member names should be included somewhere in the document.

Part I: Exploratory Data Analysis

Consider the data in the file CandyBars.jmp or CandyBars.txt. A study was conducted to investigate the nutrition content of a variety of candy brands and products. The data set includes the following variables (most of which should be self-explanatory):

Brand – Candy brand
Name – Specific name of candy product (e.g., Snickers, Kit Kat)
Oz/pkg – Weight in ounces per package
Calories
Total fat g
Saturated fat g
Cholesterol g
Sodium mg
Carbohydrate g
Dietary fiber g
Sugars g
Protein g
Vitamin A %RDI
Vitamin C %RDI
Calcium %RDI
Iron %RDI

Perform the following exploratory data analysis tasks:

1. Consider all of the numeric variables (i.e. all of the variables except Brand and Name): Determine the **variance/covariance matrix** *and* the **correlation matrix** of these variables. Interpret briefly what you see.

2. Construct a **scatterplot matrix** of the variables specified in #1. Comment on what you observe.

3. Construct a **color map on correlations** for the variables specified in #1. Comment on what you observe.

4. Use a probability plot to assess the multivariate normality of the variables specified in #1. Interpret.

Part II: Principal Component Analysis

Reconsider the candy bar data. Perform a Principal Component Analysis on this data by carrying out the following steps:

a) i) Find and display the eigenvalues of the correlation matrix. Use these along with a Scree plot (and/or other means) to determine the number of principal components that are to be retained. How much of the variation in the data is explained by your chosen number of principal components? Remember that your goal is dimension reduction…so please be sensible in your choice of number of components.

ii) Provide the loadings matrix. What do you learn from this matrix? Using this matrix, ***and as best as possible***, interpret the first two principal components (I will expect some interpretation).

iii) Construct a biplot of the loadings and scores for the first two principal components. For this plot, **color the observations based on Brand**. Are there any natural groupings of the observations? Are there any unusual observations?

b) Show how the results change if PCA were performed on the covariance matrix instead of the correlation matrix. What do you learn from this?