MATH 4310 **Name:**

Homework 10

Problem 1. We consider a data set extracted from the 2017-2018 US National Health and Nutrition Examination Survey consisting of 230 participants aged between 20-25 years. For each participant, data were collected about body measures to estimate the prevalence of overweight and obesity. Data is stored in bodydata.csv which you need to download from my webpage.

a) Draw a scatterplot of the data in pairs using the R command pair as in the notes.
b) Compute the correlation of the data matrix. Note: some data row have missing numbers. To apply the R command correctly, you will need to use this version of the command: cor(bodydata,use = "complete.obs"). You will have to do the same below.
c) Plot the correlation matrix using numbers as well as circles to display the size of correlation coefficients.
d) Apply hierarchical clustering as in the lectures on the correlation plot using 2 clusters.
e) Compute the p-values on the correlation matrix.
f) Analyze the results: which variables are strongly correlated (correlation coefficient > 0.7) to each other? Which variables are not statistically correlated (use alpha = 0.05)?

Problem 2. Load the Iris dataset in R

```
library(datasets)
data(iris)
```

It is a data frame with 150 samples (rows) and 5 variables (columns) named Sepal.Length, Sepal.Width, Petal.Length, Petal.Width, and Species.

a) Generate a statistical summary of the data
b) Apply PCA and LDA analysis to the data and generate a plot to represent the features with respect to 2 dimensions. Make sure to label the axes appropriately and display the different species using different colors or symbols.
c) Concisely discuss the performance of the two methods.