CS 1675 Homework #1
Ava Chong

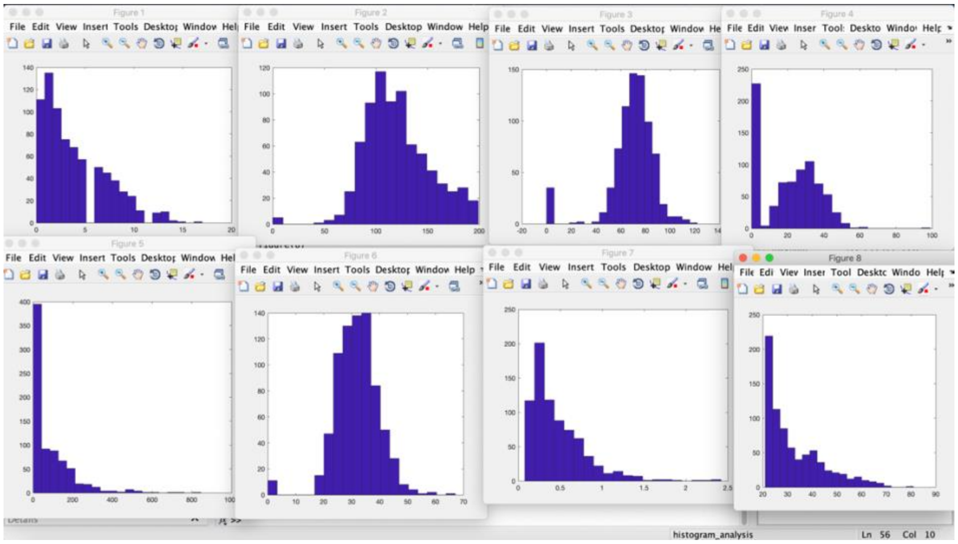
Problem 2: Exploratory Data Analysis

a. Max values for each attribute = [17, 199, 122, 99, 846, 67.1, 2.42, 81]
   Min values for each attribute = [0, 0, 0, 0, 0, 0, 0.0780, 21]
   Range for each attribute = [17, 199, 122, 99, 67.1, 2.34, 60]

b. Means for each attribute = [3.84, 120.89, 69.10, 20.53, 79.79, 31.99, 0.47, 33.24]
   Variances for each attribute = [11.35, 1022.2, 374.64, 254.46, 13281.18, 62.16, 0.1098, 138.30]

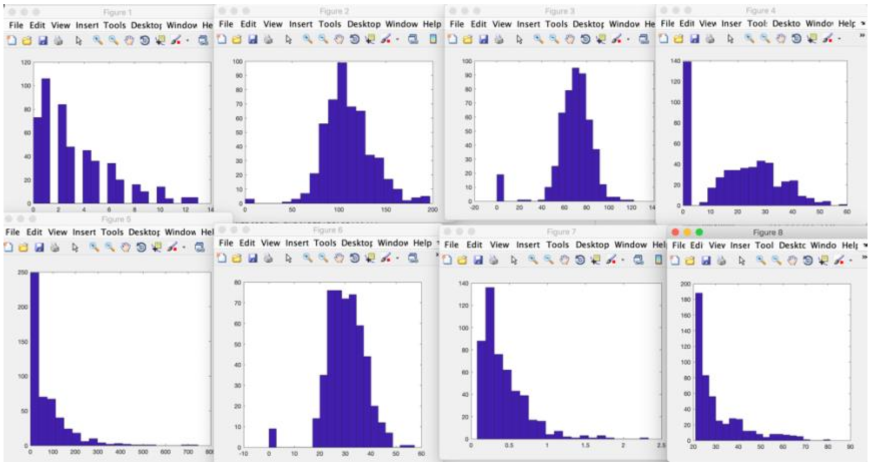
c.
   a. For data with 9th attribute = 0
      i. Mean values for each attribute = [3.298, 109.98, 68.184, 19.664, 68.792, 30.304, 0.429734, 31.19]
      ii. Standard deviation for each attribute = [3.017, 26.14, 18.063, 14.88, 98.86, 7.689, 0.299, 11.66]
   b. For data with 9th attribute = 1
      i. Mean values for each attribute = [4.865, 141.257, 70.824, 22.1641, 100.335, 35.1425, 0.550, 37.067]
      ii. Standard deviation for each attribute = [3.74, 31.93, 21.49, 17.679, 138.689, 7.262, 0.372, 10.9682]
   c. Analyze data
      i. Attribute 2 has vastly different means, 109.98 and 141.25, between the two subsets. The variances between attribute 2 in each subset are 26.14 and 31.93 respectively. The consistent variances between the subsets can lead me to believe that Subset 1 will generally have higher values for attribute 2 making attribute 2 a discriminatory attribute.

d. Histograms for each attribute

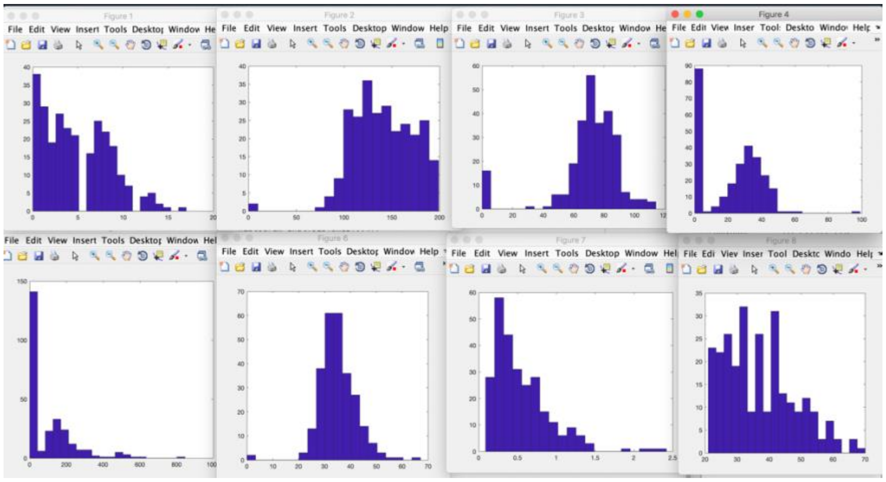


a. Figure 2, displaying attribute 2 which is plasma glucose concentration, is the closest to normally distributed. It has the majority of the data centered around the middle with an even distribution tapering off. Figures 3 and 6 also follow normally distributed patterns but with more apparent outliers and a narrower spread.

e. Discriminating between two classes (Subset 0 vs Subset 1)



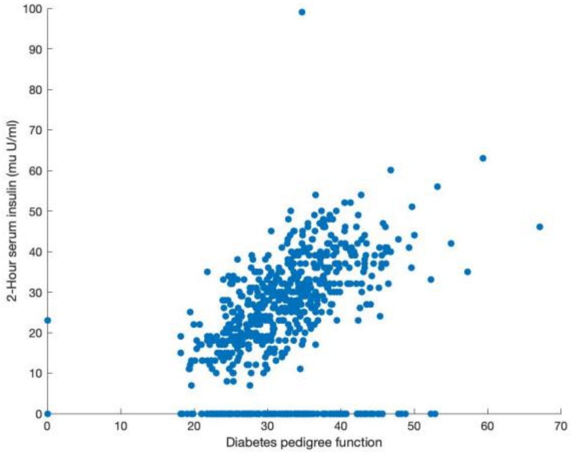
Histograms for Subset 0

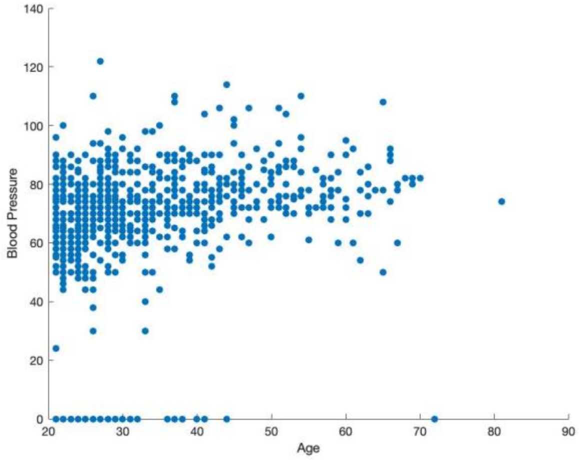


Histograms for Subset 1

Looking at the histograms for all 8 of the attributes in Subset 0 and Subset 1, we can see that the histogram with the biggest difference is Figure 8. In Subset 0, attribute 8 is heavily skewed to the left while in Subset 1 attribute 8 is less skewed and more evenly distributed. Therefore, attribute 8 which is Age in years, can be considered the discriminating attribute.

f. Using the *scatter_plot* function, we can graph any two attributes against each other. If the attributes are independent and random, we expect to see minimal correlation and clustering. I found that the Diabetes Pedigree Function vs 2-Hour serum insulin (mu U/ml) scatter plot had a general positive correlation between the attributes. This is interesting because it shows that diabetes pedigree function is related to use of the insulin. I also found within the Age vs Blood pressure scatter plot, age played little to no role in change in blood pressure and that blood pressure was not affected by age. I found this interesting because it shows that other factors have an effect to your blood pressure and age has little to no effect.





Problem 3: Data Preprocessing

a.

[0, 0, 0, 1, 0, 0, 0, 0;
0, 0, 0, 0, 0, 0, 0, 1;
0, 0, 0, 0, 1, 0, 0, 0;
0, 0, 0, 1, 0, 0, 0, 0;
0, 0, 0, 0, 0, 0, 1, 0;
0, 1, 0, 0, 0, 0, 0, 0;
0, 1, 0, 0, 0, 0, 0, 0]

b. Normalize
   a. First five normalized values of attribute 3 = [0.149, -0.160, -0.263, -0.160, -1.50]

c. Discretize Attribute

   a. First five discretized values of attribute 3 = [6, 6, 6, 6, 5]

Problem 4: Splitting data into training and testing sets

a. In this problem, we needed to split the data by percentage (ptrain) while making sure the data was random. First, I used randperm(length(data)) to randomly create a vector that held a random ordering of each row from the data. I then multiplied the length of the data by the ptrain percentage and rounded that number to find how many rows I would need to allot for the training set. I used a for loop to iterate through the random row vector the number of rows needed for the training set. Each iteration, I add the data at the random row to the training set. Following that for loop is another for loop but for the rest of the random row vector. The ending random row vector points to the rows in data that need to be stored in the testing set. I tested my code with different ptrain values and found that it splits the data accordingly.

Problem 5: Matrix operations practice problems

a. A’ = [1, 3; 2, 4; 5, 6]
b. B^{-1} = [-1, -5.5, 1.25; 0, -0.5, 0.25; -.667, 4.33, -1]
c. B + C = [15, 7, 14; 3, -1, 7; 3, 6, 10]
d. B – C = [-1, -5, 4; 1, 5, -1; 5, 10, 2]
e. A*B = [3, 45, 45; 53, 59, 75]
f. B*C = [48, 21, 75; 15, 0, 30; 34, -12, 76]
g. B*A = Cannot compute because matrix dimensions do not align.