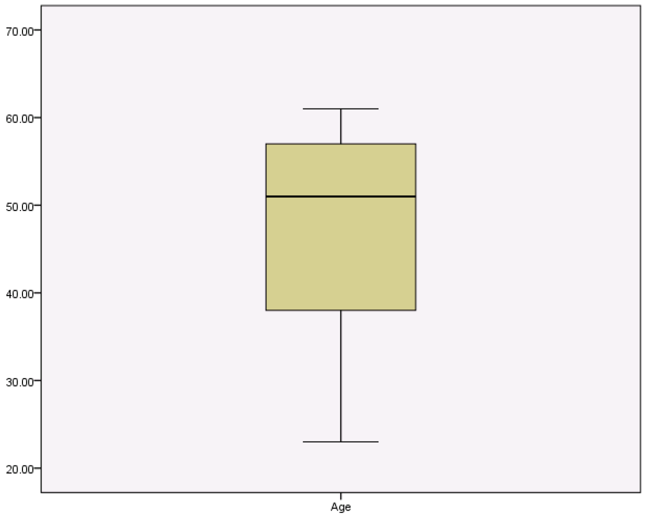
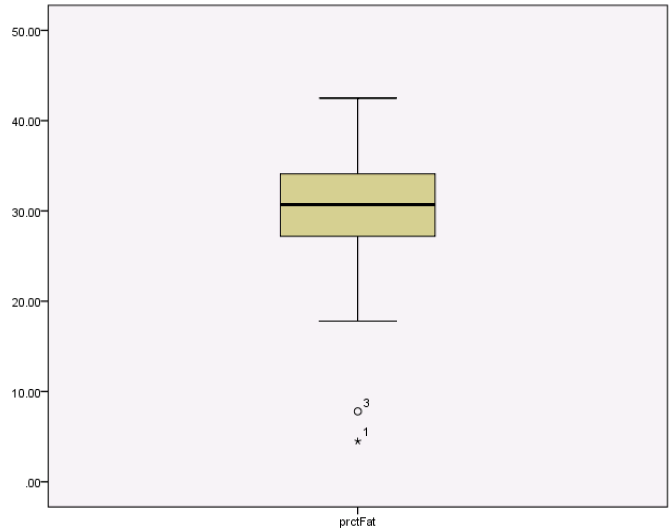
Assignment 2- IS 467
Sarah Cummings

Problem 1: Age and Percent Fat Data—
a) Box plots for Age and %fat





Looking at the boxplot, we see that the age data is skewed left. The median is a little above 50 years old, with minimum of 23 and maximum of 61. There are no outliers in the age data.
For the percent fat data, we have a fairly symmetric boxplot with mimim of 7.8% and max of 42.5%. The median is around 30%. There are also two influential points on the lower end of the data, 7.8% and 9.5%.

b) Z-score normalization of data:

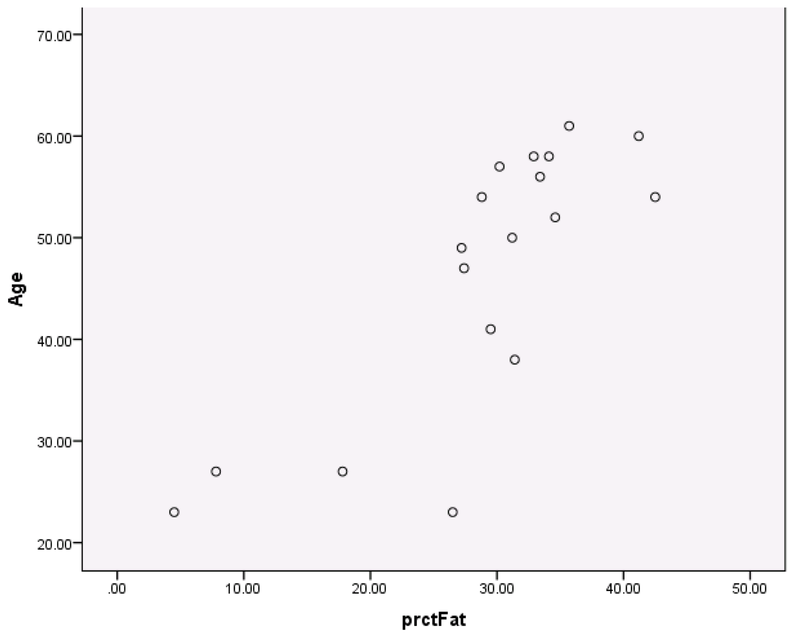
ZAge	ZprctFat
-1.76469	-2.44523
-1.76469	-.22280
-1.46289	-2.11186
-1.46289	-1.10167
-.63294	.27219
-.40659	.08025
.04611	-.13189
.19701	-.15209
.27246	.25199
.42336	.59545
.57426	1.39350
.57426	.00954
.72516	.47423
.80061	.15097
.87606	.54494
.87606	.42372
1.02696	1.26218
1.10241	.70657

c) Value ranges for he following normalization methods, with explanation:
i.Min-max normalization: Min-max normalization allows us to transform the data into any range we see fit. Suppose we wanted to compare the age and percent fat values on a 1-10 scale, then the min-max formula would allow us to do just that.

ii.Z score standardization: As seen in the z score standardization of these variables (left), ZAge ranges from [-1.76, 1.102] and ZperctFat ranges from [-2.44, 1.39]. These values are found by subtracting the mean from each value, and then dividing by the standard deviation for that variable.

iii. Decimal Scaling: The value range is [0,1] for any variable that is transformed with decimal scaling. This is because decimal scaling is designed to produce values between 0 and 1 every time, since the transformed value of v,
v'= v/(10^j) Where *j* is the smallest integer such that
Max(l v' l)<1

d) Scatterplot of the two variables, and interpretation of their relationship:



As seen in the scatterplot, age and percent fat have a positive relationship. Subjects who are older tend to have a higher percentage of fat.

e) Correlation coefficient between the variables, and covariance matrix:

Correlations

		Age	prctFat
Age	Pearson Correlation	1	.805**
	Sig. (2-tailed)		.000
	N	18	18
prctFat	Pearson Correlation	.805**	1
	Sig. (2-tailed)	.000	
	N	18	18

**. Correlation is significant at the 0.01 level (2-tailed).

Inter-Item Covariance Matrix

	Age	prctFat
Age	175.663	105.604
prctFat	105.604	97.992

As seen in the SPSS correlation output, our variables have a correlation coefficient of 0.805, which is statistically significant and confirms a positive relationship between age and percent fat.

Also, note the covariance matrix above (right).

Problem 2: Data preprocessing/ Binning of sales price data—

Data: 5, 10, 11, 13, 15, 35, 50, 55, 72, 92, 204, 215

a) equal-depth binning with 3 values per bin:
Bin 1: 5, 10, 11
Bin 2: 13, 15, 35
Bin 3: 50, 55, 72
Bin 4: 92, 204, 215

Smoothing by bin means:
Bin 1: 8.66, 8.66, 8.66
Bin 2: 63, 63, 63
Bin 3: 59, 59, 59
Bin 4:170.33, 170.33, 170.33

Smoothing by bin boundaries:
Bin 1: 5, 5, 11
Bin 2: 13, 13, 35
Bin 3: 50, 50, 72
Bin 4: 92, 92, 215

Interpretation: Using the equi-depth binning, we have very different ranges for each bin. With equi-depth binning for this data, I think its best to smooth by bin boundaries so we get a better picture of the data.

b) equal-width binning with 3 bins:

Width= 215-5/ 3 —> Width= 70

Bin 1: from 5 to 75, contains 5, 10, 11, 13, 15, 35, 50, 55, 72
Bin 2: from 75 to 145, contains 92
Bin 3 from 145 to 215, contains 204, 215

Smoothing by bin means:
Bin 1: 29.55, 29.55, 29.55, 29.55, 29.55, 29.55, 29.55, 29.55, 29.55
Bin 2: 92
Bin 3: 209.5, 209.5

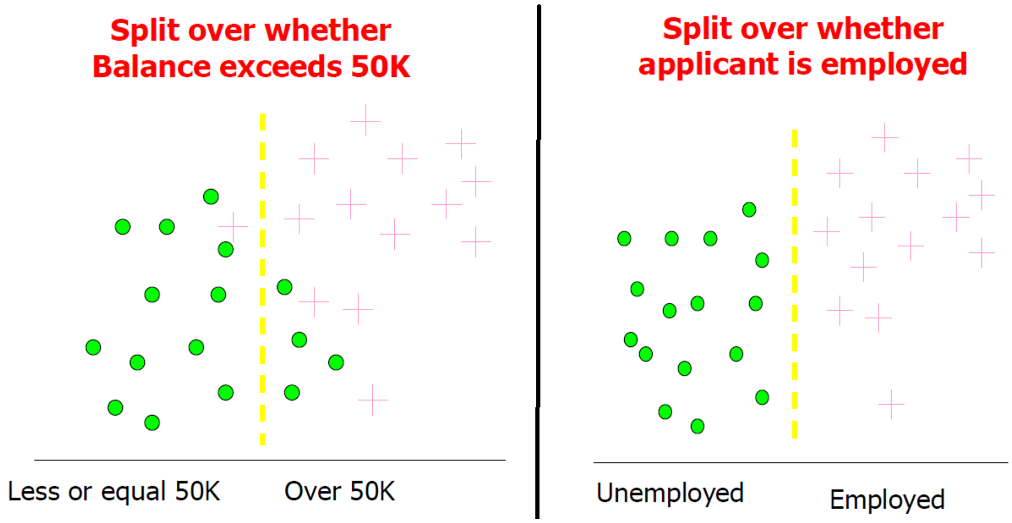
Smoothing by bin boundaries :
Bin 1: 5, 5, 5, 5, 5, 5, 5, 5, 72
Bin 2: 92
Bin 3: 204, 215

Interpretation: Using equi-width binning, we have most of our data in the first of three bins. This data is skewed and has what appears to be a couple upper-end outliers, thus equi-width binning is not best for representing the data.

Problem 3: Classification—

a) Which variable should we use to classify?

Looking at the dat plots, it makes most since to divide the data into classes by employment status. Using employment status, we have a clean division between the classes with no overlap. The split for whether or not balance excesses 50K has some overlap and entropy.



b) Three variables, two classes; which variable should we use to classify?

Right away, according to the chart, it appears as though variable Y is most consistent in deterring the classes for the data. Observations with Y=1 belong to class one consistently, and observations with Y=0 belong to class two consistently.

X	Y	Z	C
1	1	1	I
1	1	0	I
0	0	1	II
1	0	0	II

Calculations:

Information(D)= $-2/4 \log2(2/4)-2/4 \log2(2/4) = -1/2(-1)- 1/2(-1)=$ **1**

$\text{Info}_X(D)= 3/4(-2/3*\log_2(2/3) -1/3*\log_2(1/3)) + 1/4(-1/1*\log_2(1/1))$

$= 3/4(0.3899 + 0.52824) + 1/4(-1*(0))$

$= 0.6886$

$\text{Gain}(X)= \text{Info}(D) - \text{Info}_X(D)=$ **0.3114**

$\text{Info}_Y(D)=2/4(-2/2*\log_2(2/2)) +2/4(-2/2\log_2(2/2))= 0$

$\text{Gain}(Y)= \text{Info}(D)- \text{Info}_Y(D) =$ **1**

$\text{Info}_Z(D)= 2/4(-1/2*\log_2(1/2) - 1/2*\log_2(1/2)) + 2/4(-1/2*\log_2(1/2) - 1/2*\log_2(1/2))$

$=1/2(1/2 + 1/2) + 1/2(1/2 + 1/2) = 1$

$\text{Gain}(Z)= \text{Info}(D)- \text{Info}_Z(D) =$ **0**

As expected, Gain(Y) is greater than that of X or Z, thus confirming that Y is best in classifying our variables.