Keiland Pullen Assignment 2
DSC 441 – Fall 2021

Problem 1.)

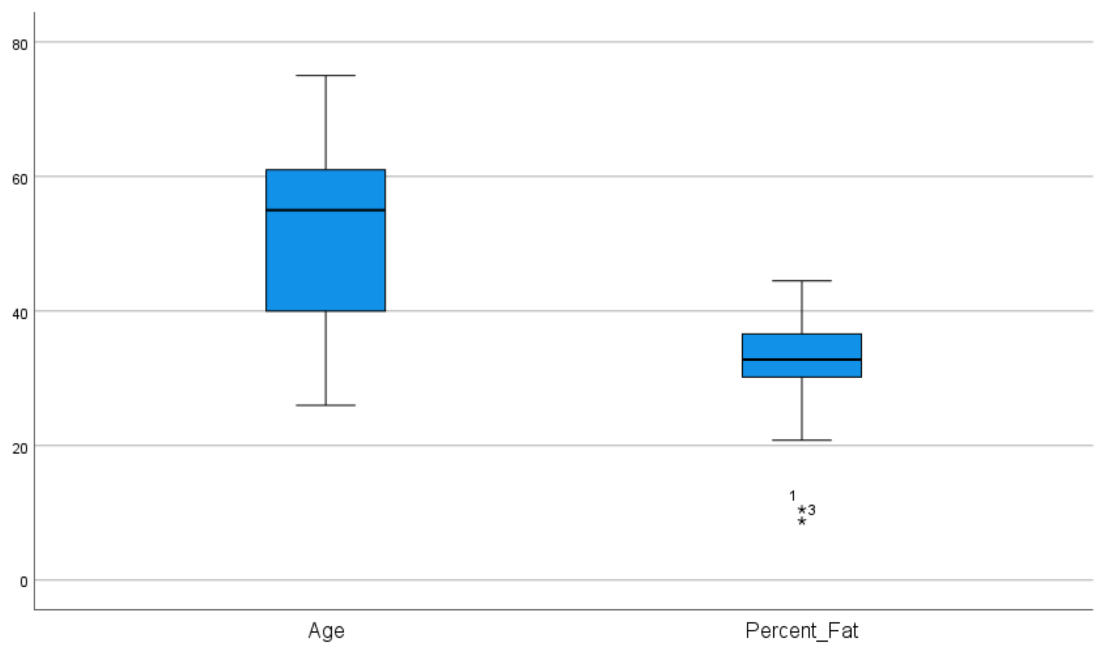
a.)

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
Age	18	100.0%	0	0.0%	18	100.0%
Percent_Fat	18	100.0%	0	0.0%	18	100.0%

Statistics

		Age	Percent_Fat
N	Valid	18	18
	Missing	0	0
Mean		50.1111	31.0611
Median		55.0000	32.8000

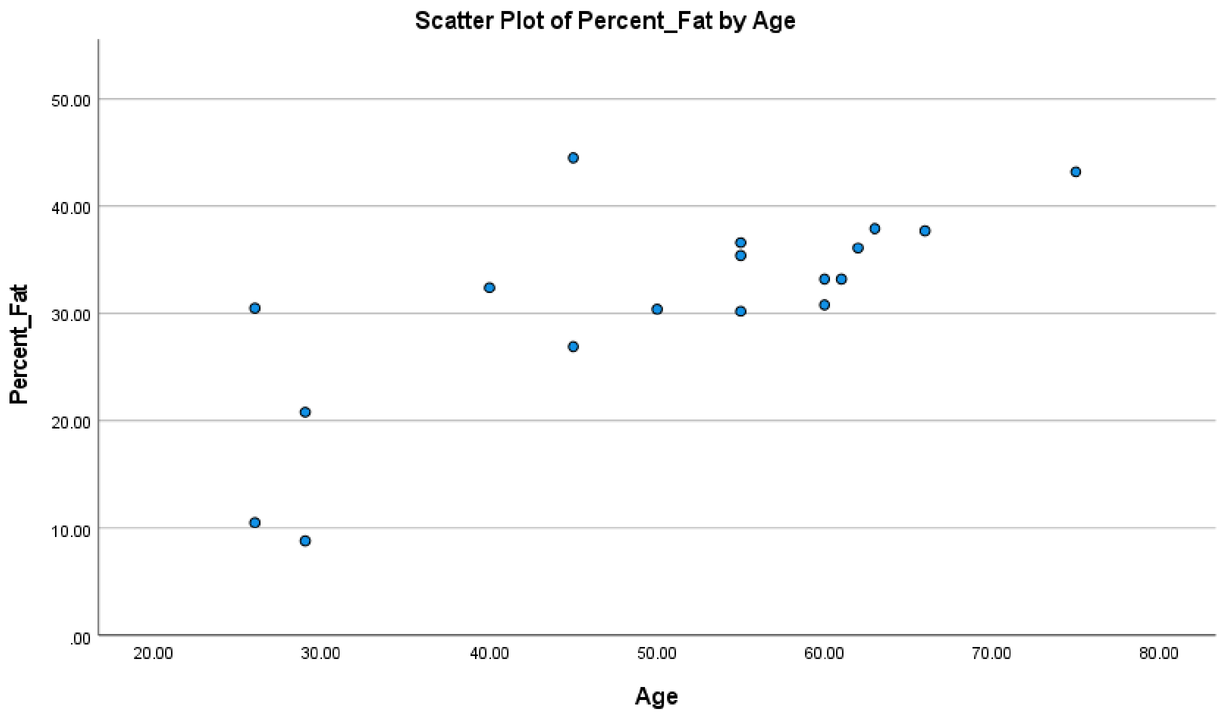


The box-plot for Age displays the median being closer to the 3rd Quartile. This tells us that the majority of the data are below the median.

The box-plot for Percent_Fat displays 2 outliers that should be examined. The two outliers are data entry 1 and data entry 3. The values for those 2 outliers are 10.5% and 8.8%.

Keiland Pullen Assignment 2
DSC 441 – Fall 2021

b.)



There appears to be a positive relationship between Age and Percent_Fat. We see that in general, as Age increases Percent_Fat also increases. There are at least 2 distinct outliers.

c.)

Correlations

		Age	Percent_Fat
Age	Pearson Correlation	1	.735**
	Sig. (2-tailed)		.001
	N	18	18
Percent_Fat	Pearson Correlation	.735**	1
	Sig. (2-tailed)	.001	
	N	18	18

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

This Correlation Matrix tells us that they are positively correlated.

Keiland Pullen Assignment 2
DSC 441 – Fall 2021

Correlations

		Age	Percent_Fat
Age	Pearson Correlation	1	.735**
	Sig. (2-tailed)		.001
	Sum of Squares and Cross-products	3773.778	1776.578
	Covariance	221.987	104.505
	N	18	18
Percent_Fat	Pearson Correlation	.735**	1
	Sig. (2-tailed)	.001	
	Sum of Squares and Cross-products	1776.578	1547.123
	Covariance	104.505	91.007
	N	18	18

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

Covariance is the Correlation of X and Y times the Standard Deviation of each variable. This calculation is hard to understand at times, hence the Correlation Matrix may be easier to understand, as the Correlation range is from -1 to 1.

Problem 2.)

12 sales price records:
[8, 13, 14, 15, 17, 37, 55, 60, 77, 95, 208, 218]

Bin 1 [8, 13, 14]
Bin 2 [15, 17, 37]
Bin 3 [55, 60, 77]
Bin 4 [95, 208, 218]

Smooth by Boundary:

Bin 1 [8, 14, 14]
Bin 2 [15, 15, 37]
Bin 3 [55, 55, 77]
Bin 4 [95, 218, 218]

Keiland Pullen
DSC 441 – Fall 2021

Assignment 2

Problem 3.)

a.)

In this case, an "eye-test" for the classification would suggest that Unemployed/Employed data plot would be the easiest to use because it answers a simple Yes or No question about employment. However, we know that we can calculate the entropy of both cases and see which is closer to zero. A set of only one class is extremely predictable meaning it would have low entropy. A set of mixed classes is unpredictable, meaning it would have high entropy. We should select the classification that has the lowest entropy.

b.)
i.)

	Test 1 = T	Test 1 = F
+	4	0
-	3	3
	7	3

	Test 2 = T	Test 2 = F
+	4	0
-	0	6
	4	6

	Test 3 = T	Test 3 = F
+	3	1
-	1	5
	4	6

Test 1:

E_{True} = - [$\frac{4}{7}\log_2\frac{4}{7} + \frac{3}{7}\log_2\frac{3}{7}$] = 0.29658

E_{False} = - [$\frac{0}{3}\log_2\frac{0}{3} + \frac{3}{3}\log_2\frac{3}{3}$] = 0

$E_{Test\ 1}$ = ($\frac{7}{10} * 0.29658 + \frac{3}{10} * 0$) = 0.2076

Test 2:

E_{True} = - [$\frac{4}{4}\log_2\frac{4}{4} + \frac{0}{4}\log_2\frac{0}{4}$] = 0

E_{False} = - [$\frac{0}{6}\log_2\frac{0}{6} + \frac{6}{6}\log_2\frac{6}{6}$] = 0

$E_{Test\ 2}$ = ($\frac{4}{10} * 0 + \frac{6}{10} * 0$) = 0

Test 3:

E_{True} = - [$\frac{3}{4}\log_2 3/4 + \frac{1}{4}\log_2\frac{1}{4}$] = 0.2442

E_{False} = - [$\frac{1}{6}\log_2\frac{1}{6} + \frac{5}{6}\log_2\frac{5}{6}$] = 0.1956

$E_{Test\ 3}$ = ($\frac{4}{10} * 0.2442 + \frac{6}{10} * 0.1956$) = .21504

I would say that Test 1 would be used first as it has the lowest Entropy.

ii.)

Test 1 – Gini Index
True : $1 - (\frac{4}{7})^2 - (\frac{3}{7})^2 = 0.4897959184$
False: $1 - (\frac{0}{3})^2 - (\frac{3}{3})^2 = 0$

Gini = 7/10 * 0.4897959184 + 3/10 * 0 = 0.342

Test 3 – Gini Index
True : $1 - (\frac{3}{4})^2 - (\frac{1}{4})^2 = 0.375$
False: $1 - (\frac{1}{6})^2 - (\frac{5}{6})^2 = 0.2777$

Gini = 4/10 * 0.375 + 6/10 * 0.277 = 0.3162

In this case, Test 3 would be preferred.

4.)

a.)

The dataset contains approximately 1420 cases or instances.

There are 18 total variables. Remove the index variable and there are 17 variables.

The class distribution would be dinner, party, sleep, and workout.

The image below is the Correlation matrix for this data set. The matrix indicates that there is significant correlation between a number of the variables.

		acousticness	danceability	duration_ms	energy	instrumentalness	key	liveness	loudness	mode	speechiness	tempo	time_signature	valence
acousticness	Pearson Correlation	1	-.526**	.056*	-.816**	.566**	-.042	-.217**	-.724**	.077**	-.319**	-.220**	-.254**	-.365**
	Sig. (2-tailed)		<.001	.035	.000	<.001	.113	<.001	<.001	.004	<.001	<.001	<.001	<.001
	N	1420	1420	1420	1420	1420	1420	1420	1420	1420	1420	1420	1420	1420
danceability	Pearson Correlation	-.526**	1	-.302**	.436**	-.569**	.031	-.105**	.652**	-.067*	.208**	.146**	.296**	.627**
	Sig. (2-tailed)	<.001		<.001	<.001	<.001	.240	<.001	<.001	.011	<.001	<.001	<.001	<.001
	N	1420	1420	1420	1420	1420	1420	1420	1420	1420	1420	1420	1420	1420
duration_ms	Pearson Correlation	.056*	-.302**	1	.046	.155**	-.071**	.180**	-.203**	.042	-.014	-.119**	-.076**	-.216**
	Sig. (2-tailed)	.035	<.001		.080	<.001	.008	<.001	<.001	.116	.598	<.001	.004	<.001
	N	1420	1420	1420	1420	1420	1420	1420	1420	1420	1420	1420	1420	1420
energy	Pearson Correlation	-.816**	.436**	.046	1	-.538**	.045	.332**	.777**	-.055*	.282**	.211**	.238**	.400**
	Sig. (2-tailed)	.000	<.001	.080		<.001	.091	<.001	<.001	.038	<.001	<.001	<.001	<.001
	N	1420	1420	1420	1420	1420	1420	1420	1420	1420	1420	1420	1420	1420
instrumentalness	Pearson Correlation	.566**	-.569**	.155**	-.538**	1	-.014	-.062*	-.726**	-.026	-.263**	-.173**	-.261**	-.505**
	Sig. (2-tailed)	<.001	<.001	<.001	<.001		.599	.020	<.001	.324	<.001	<.001	<.001	<.001
	N	1420	1420	1420	1420	1420	1420	1420	1420	1420	1420	1420	1420	1420
key	Pearson Correlation	-.042	.031	-.071**	.045	-.014	1	.033	.021	-.178**	.088**	-.044	.021	.083**
	Sig. (2-tailed)	.113	.240	.008	.091	.599		.209	.429	<.001	<.001	.099	.437	.002
	N	1420	1420	1420	1420	1420	1420	1420	1420	1420	1420	1420	1420	1420
liveness	Pearson Correlation	-.217**	-.105**	.180**	.332**	-.062*	.033	1	.111**	-.018	.128**	.014	.023	-.067*
	Sig. (2-tailed)	<.001	<.001	<.001	<.001	.020	.209		<.001	.500	<.001	.609	.391	.012
	N	1420	1420	1420	1420	1420	1420	1420	1420	1420	1420	1420	1420	1420
loudness	Pearson Correlation	-.724**	.652**	-.203**	.777**	-.726**	.021	.111**	1	-.034	.252**	.262**	.299**	.488**
	Sig. (2-tailed)	<.001	<.001	<.001	<.001	<.001	.429	<.001		.201	<.001	<.001	<.001	<.001
	N	1420	1420	1420	1420	1420	1420	1420	1420	1420	1420	1420	1420	1420
mode	Pearson Correlation	.077**	-.067*	.042	-.055*	-.026	-.178**	-.018	-.034	1	-.081**	-.015	-.008	-.064*
	Sig. (2-tailed)	.004	.011	.116	.038	.324	<.001	.500	.201		.002	.572	.750	.015
	N	1420	1420	1420	1420	1420	1420	1420	1420	1420	1420	1420	1420	1420
speechiness	Pearson Correlation	-.319**	.208**	-.014	.282**	-.263**	.088**	.128**	.252**	-.081**	1	.145**	.122**	.150**
	Sig. (2-tailed)	<.001	<.001	.598	<.001	<.001	<.001	<.001	<.001	.002		<.001	<.001	<.001
	N	1420	1420	1420	1420	1420	1420	1420	1420	1420	1420	1420	1420	1420
tempo	Pearson Correlation	-.220**	.146**	-.119**	.211**	-.173**	-.044	.014	.262**	-.015	.145**	1	.054*	.094**
	Sig. (2-tailed)	<.001	<.001	<.001	<.001	<.001	.099	.609	<.001	.572	<.001		.042	<.001
	N	1420	1420	1420	1420	1420	1420	1420	1420	1420	1420	1420	1420	1420
time_signature	Pearson Correlation	-.254**	.296**	-.076**	.238**	-.261**	.021	.023	.299**	-.008	.122**	.054*	1	.180**
	Sig. (2-tailed)	<.001	<.001	.004	<.001	<.001	.437	.391	<.001	.750	<.001	.042		<.001
	N	1420	1420	1420	1420	1420	1420	1420	1420	1420	1420	1420	1420	1420
valence	Pearson Correlation	-.365**	.627**	-.216**	.400**	-.505**	.083**	-.067*	.488**	-.064*	.150**	.094**	.180**	1
	Sig. (2-tailed)	<.001	<.001	<.001	<.001	<.001	.002	.012	<.001	.015	<.001	<.001	<.001	
	N	1420	1420	1420	1420	1420	1420	1420	1420	1420	1420	1420	1420	1420

b.)

The following table displays the range of each of the numerical variables. The range is calculated by subtracting the maximum value from the minimum value in the respective data column. Yes, the data should be normalized. In this case, I would suggest Z-score normalization as it will assist with detecting outliers, but at the expense of each variable having the same scale. Min-Max is possible, as all features will have the same scale, but it isn't good for outliers. I would suggest using both.

Statistics

		danceability	duration_ms	energy	key	liveness	loudness	mode	speechiness	tempo	time_signature
N	Valid	1420	1420	1420	1420	1420	1420	1420	1420	1420	1420
	Missing	0	0	0	0	0	0	0	0	0	0
Range		.9085	4445704	.99846	11	.9563	41.058	1	.4971	161.174	4