Alex Teboul
DSC 441
Assignment 1
Due Date: Saturday, *October 6th, 2018, by midnight*

Problem 1 (10 points): This problem is an example of data preprocessing needed in a data mining process.

Suppose that a hospital tested the age and body fat data for 18 randomly selected adults with the following results:

Age	26	26	29	29	40	45	50	55	60
%fat	10.5	30.5	8.8	20.8	32.4	26.9	30.4	30.2	33.2
Age	55	45	60	55	61	62	63	75	66
%fat	36.6	44.5	30.8	35.4	33.2	36.1	37.9	43.2	37.7

a. (2 points) Draw the box-plots for age and %fat. Interpret the distribution of the data.

Table 1: Five Number Summary

		Age	Percent_Fat
N	Valid	18	18
	Missing	0	0
1. Median		55.000	32.800
2. Minimum		26.000	8.800
3. Maximum		75.000	44.500
Quartiles	4. Q1	37.250	29.375
	5. Q3	61.250	36.875

- To interpret the distribution of the data, boxplots for Age and %Fat were created. The two boxplots show a fairly normal distribution. The %Fat boxplot also identified the first and third data points as outliers. Table 1 describes the data distribution shown in box plots with the Five Number Summary.

- The z-value calculation and Q-Q plots were also analyzed to confirm the interpretation that the data is normally distributed.

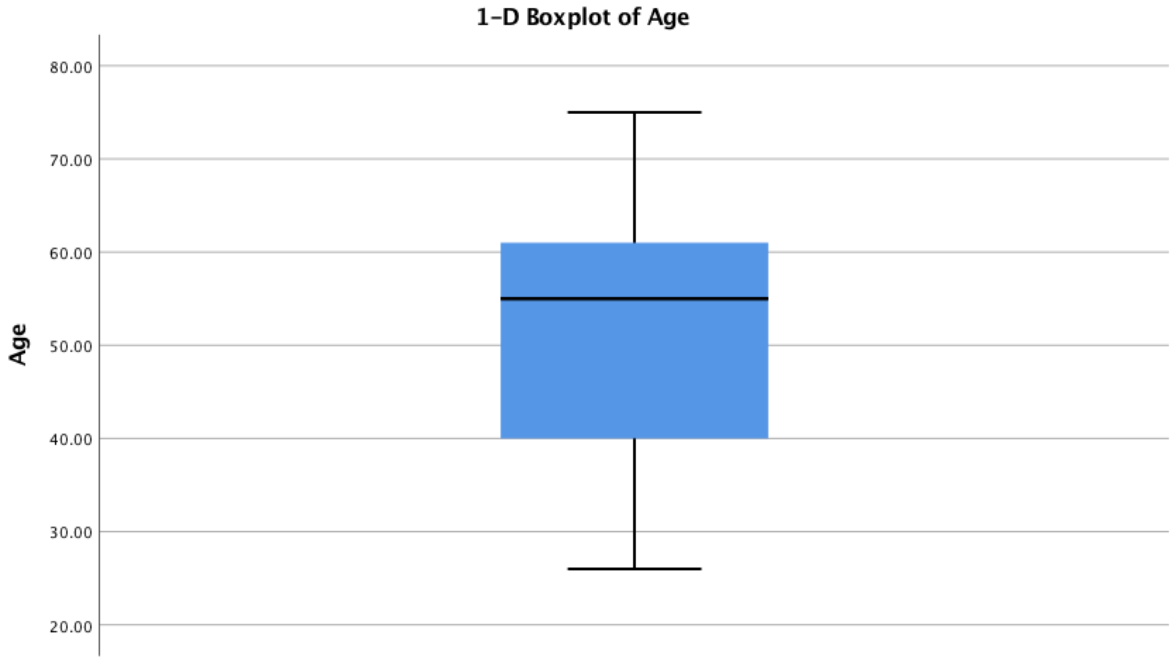

Figure 1: This is the boxplot for age. 5 Number summary is found in the Table 1.

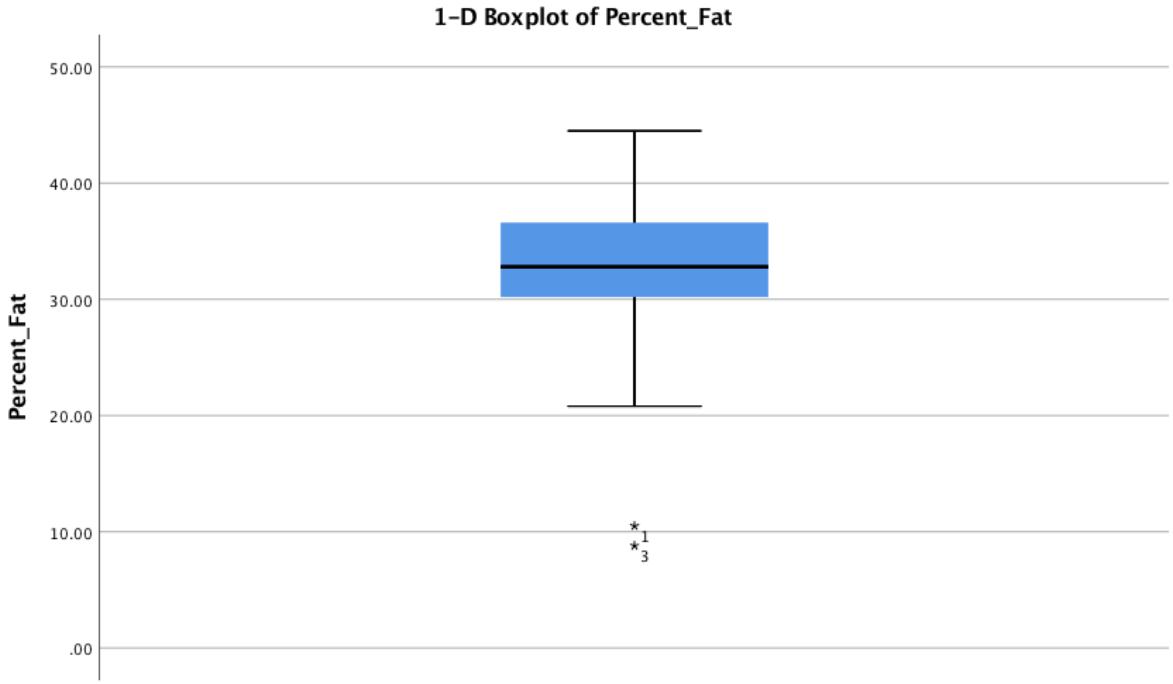

Figure 2: This is the boxplot for %Fat. 5 Number summary is found in the Table 1.

- Want a z-value between -1.96 and +1.96 to confirm normal distribution. This z-value is calculated by dividing the skewness and kurtosis by their standard error.

Table 2: Skewness and Kurtosis

	Statistic	Std. Error	z-value	Between -1.96 and +1.96?
Age Skewness	-.427	.536	-0.797	Yes
Age Kurtosis	-.865	1.038	-0.833	Yes
%Fat Skewness	-1.193	.536	-2.226	No – Skewed due to outliers
%Fat Kurtosis	1.409	1.038	1.357	Yes

- The Q-Q Plots below confirm that the data is fairly normally distributed as well, as the points fall close to the line.

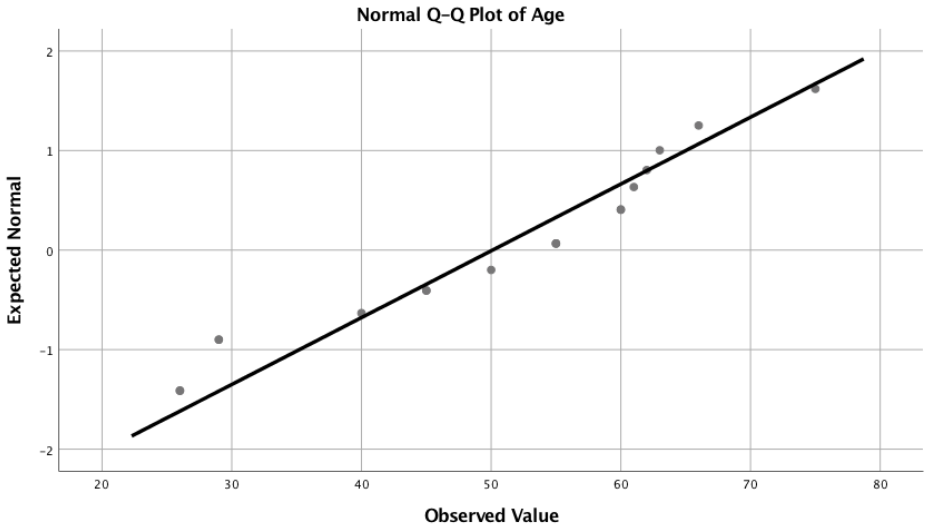

Figure 3: This plot confirms that the data is fairly normally distributed as the points line up well along the trendline.

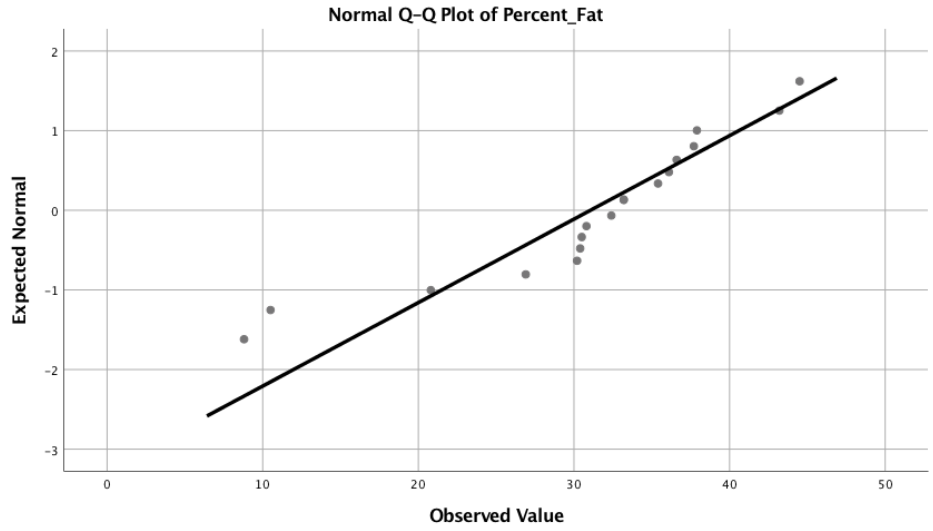

Figure 4: This plot confirms that the data is fairly normally distributed as the points line up along the trendline.

b. (2 points) Normalize the two attributes based on z-score normalization.

Z_Age	-1.61828	-1.6182	-1.41693	-1.41693	-0.67863	-0.34305	-0.00746	0.32813	0.6637
Z_%fa	-2.15530	-0.5882	-2.33351	-1.07561	0.14035	-0.43619	-0.6930	-0.09027	0.2242
Z_Age	0.32813	-0.3430	0.66372	0.32813	0.73084	0.79795	0.86507	1.67048	1.0664
Z_%fa	0.58061	1.40872	-0.02737	0.45482	0.22421	0.52820	0.71688	1.27245	0.6959

Table 3: z-score normalization

	N	Mean	Std. Deviation
Age	18	50.1111	14.89923
Zscore(Age)	18	.0000000	1.00000000
Percent_Fat	18	31.0611	9.53977
Zscore(Percent_Fat)	18	.0000000	1.00000000

- The z-score normalization performed on the two attributes worked, as indicated by the mean of zero and standard deviation of 1 shown in Table 3. I did not use the mean absolute deviation of A for the z-score normalization, just the standard deviation. The two boxplots below show the z-score normalized values.

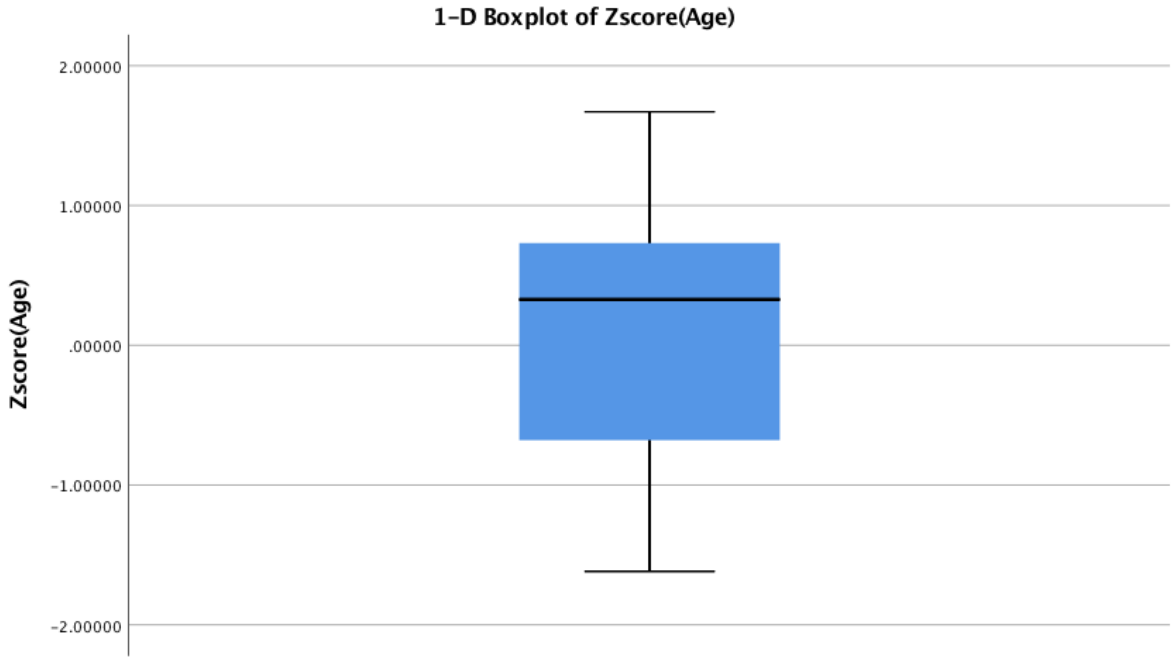

Figure 5: This boxplot shows the z-score normalized Age values.

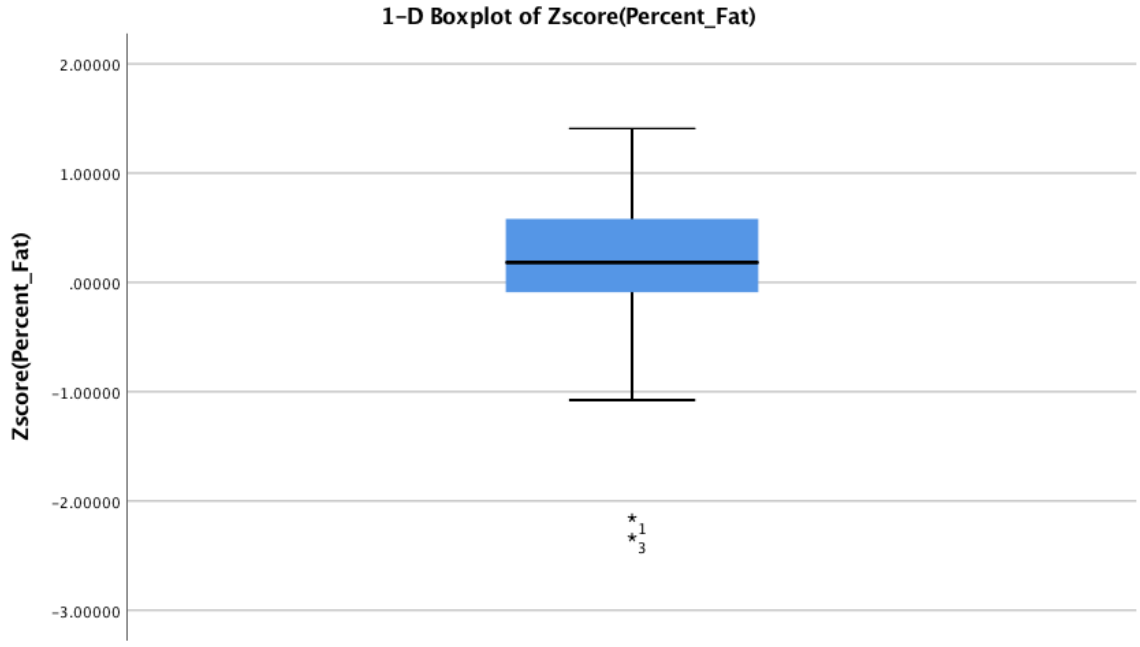

Figure 6: This boxplot shows the z-score normalized %Fat values.

c. (2 points) Regardless of the original ranges of the variables, normalization techniques transform the data into new ranges that allow to compare and use variables on the same scales. What are the values ranges of the following normalization methods? Explain your answer.

i. Min-max normalization

- In min-max normalization, the range is [new_min, new_max]. The original minimum value is becomes the new minimum and the original maximum becomes the new maximum. The formula to accomplish this is shown below:

$$v' = \frac{v - min_A}{max_A - min_A}(new\_max_A - new\_min_A) + new\_min_A$$

- In the formula, v' is the new value that the original value of v gets mapped to. The subscript '$_A$' refers to the original variable. The min_A and max_A are the variable's original minimum and maximum values respectively. The new_max_A and new_min_A are the new minimum and maximum values respectively. This is a useful technique for getting all your values onto a range of your choosing. A new range of [0, 1] or [-1, 1] is commonly chosen.

ii. Z-score normalization

- For z-score normalization, the possible range is theoretically [-∞, +∞]. It is unlikely to approach + or - infinity, but given the formula for calculating z-score below, it is clear that this range is possible. That said, 99.997% of the data should fall within the range [-4, +4].
  Each additional ±1 is 1 standard deviation after all, so you're unlikely to get a large range.

$$v' = \frac{v - mean_A}{standard\_deviation_A}$$

- The benefit of this normalization technique is that you end up with a new range of values that has a mean of zero and standard deviation of 1. Values get centered around zero.

iii. Normalization by decimal scaling.

- For normalization by decimal scaling the new range technically becomes [$min_A/10^j$, $max_A/10^j$], where j is the smallest integer such that $\max(|v'_i|) < 1$. What this really means is that the new range is going to fall within [-1, 1]. Unlike z-score normalization these values are not centered around 0. You could have all positive numbers and end up with normalized values between 0 and 1, all negative numbers and end up with normalized values between -1 and 0, or some mix of positive and negative numbers in your dataset and end up with normalized values between -1 and 1.

$$v'_i = \frac{v_i}{10^j}, \text{ where } j \text{ is the smallest integer such that } max(|v'_i|) < 1$$

- Comparing all three normalization techniques, it's important to note that min-max normalization does a better job of preserving the relationship among the original data values. Different scenarios can call for different normalization techniques, and all have pros and cons.

d. (2 points) Draw a scatter-plot based on the two variables and interpret the relationship between the two variables.

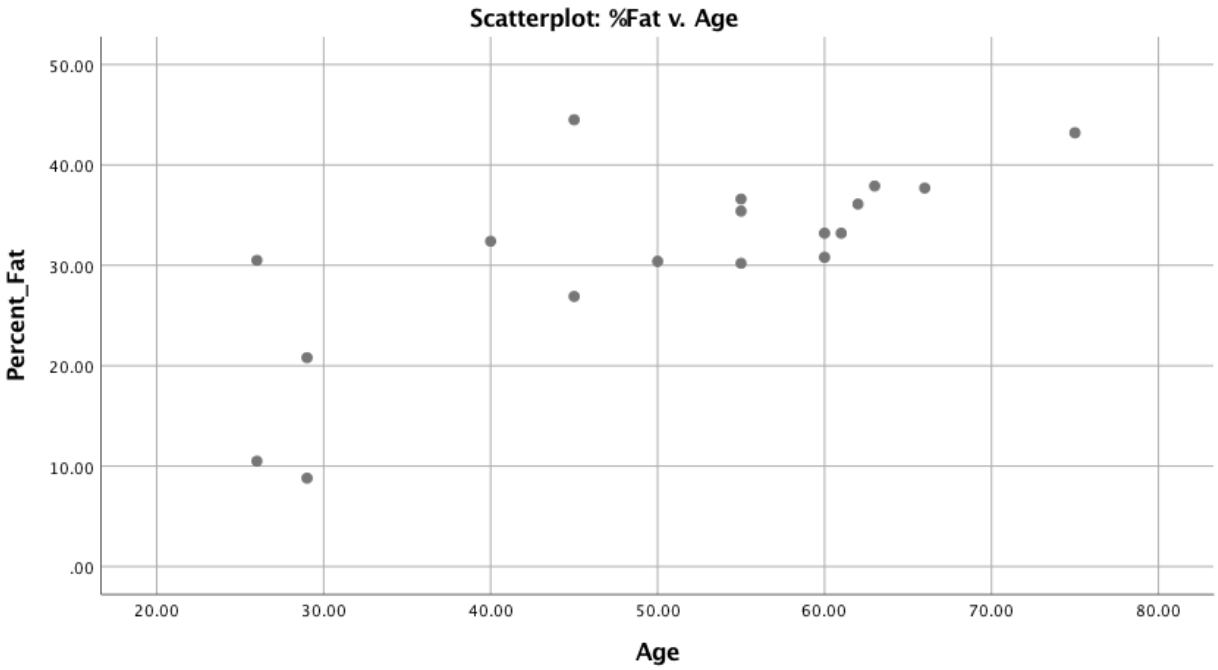

Figure 7: This figure shows the original 18 patient records in a scatterplot.

Figure 8: This figure shows the scatterplot with z-score normalized values.

- **Interpretation:** It appears that there is a positive linear relationship present in the data. Specifically, as Age increases the patients appear to also experience an increase in their

%fat. This could lead one to conclude that in general, older people have higher body fat percentages, although more data points are needed.

e. (2 points) Calculate the correlation matrix.

Table 4: Correlation Matrix

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

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

- The correlation 'r' is given by the formula:

$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[\, n\Sigma x^2 - (\Sigma x)^2 \,][\, n\Sigma y^2 - (\Sigma y)^2 \,]}}$$

Are these two attributes positively or negatively correlated?

- These two attributes are positively correlated. The r value was calculated at 0.735, which is a strong positive correlation. A perfectly positive correlation is indicated by an r value of 1, a perfectly negative correlation is indicated by an r value of -1, and no correlation is shown by an r value of 0. The diagonal axis of the correlation matrix has 1's because it is showing the correlation between variables and themselves.

Calculate the covariance matrix.

Table 5: Covariance Matrix

		Age	Percent_Fat
Age	Pearson Correlation	1	.735**
	Sig. (2-tailed)		.001
	Covariance	**221.987**	**104.505**
	N	18	18
Percent_Fat	Pearson Correlation	.735**	1
	Sig. (2-tailed)	.001	
	Covariance	**104.505**	**91.007**
	N	18	18

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

- The covariance matrix is calculated using the formula below.

$$COV(X,Y) = \frac{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)}{n-1}$$

How is the correlation matrix different from the covariance matrix?

- Both correlation and covariance can both be used to determine the relationship and measure dependency between two variables. In general correlation describes how a change in one variable can lead to a change in another variable, while covariance describes how two variables can change together. It's a nuanced difference between the two but the calculation method is clear enough.
- The correlation can be thought of as the standardized or normalized covariance. Where correlation values have a range between -1 and 1, covariance values are not bound this way. The diagonal in the covariance matrix represents the variance in that particular variable, and off diagonal values are the covariances. The diagonal in a correlation matrix is always 1 as the same variable is perfectly correlated with itself. So in the example above, Age has a variance of 221.987 and %Fat has a variance of 91.007. Age and %Fat have a covariance of 104.505, and a correlation of 0.735.

Problem 2 (5 points): This problem is an example of data preprocessing needed in a data mining process.

Suppose a group of 12 sales price records has been sorted as follows:

8, 13, 14, 15, 17, 37, 55, 60, 77, 95, 208, 218

Partition them into bins by each of the following method, smooth the data and interpret the results:

a. (2.5 points) equal-depth partitioning with 4 values per bin

Bins	Sales Price Records	Range
Bin_1	8, 13, 14, 15	[0, 17)
Bin_2	17, 37, 55, 60	[17, 77)
Bin_3	77, 95, 208, 218	[77, +)

- **Interpretation:** The sales price records have been equally partitioned into 3 bins, each bin containing 4 values. The ranges for each been vary greatly and more data points are needed in order to determine if these bins will work for future records.

b. (2.5 points) equal-width partitioning with 4 bins

W = (B - A)/N

W = (218 – 8)/4 = 52.5

Bins	Sales Price Records	Range
Bin_1	8, 13, 15, 17, 37, 55, 60	[8, 60.5)
Bin_2	77, 95	[60.5, 113)
Bin_3		[113, 165.5)
Bin_4	208, 218	[165.5, 218]

- **Interpretation:** The 208 and 218 are outliers that lead to unequal distribution of the sales price records amongst the 4 bins. With equal-width partitioning, skewness is not handled well, a fact that is evident in this example. Bin_1 ends up with 7 values, Bin_2 with 2, Bin_3 with none, and Bin_4 with 2. Without the two high values, the bins would have more even frequency within bins, as seen below:

Bins	Sales Price Records	Range
Bin_1	8, 13, 15, 17	[8, 29.75)
Bin_2	37	[29.75, 51.5)

Bin_3	55, 60	[51.5, 73.25)
Bin_4	77, 95	[73.25, 95]

Problem 3 (10 points):

a) (2 points) Figure 1 illustrates the plots for some data with respect to two variables: balance and employment status.

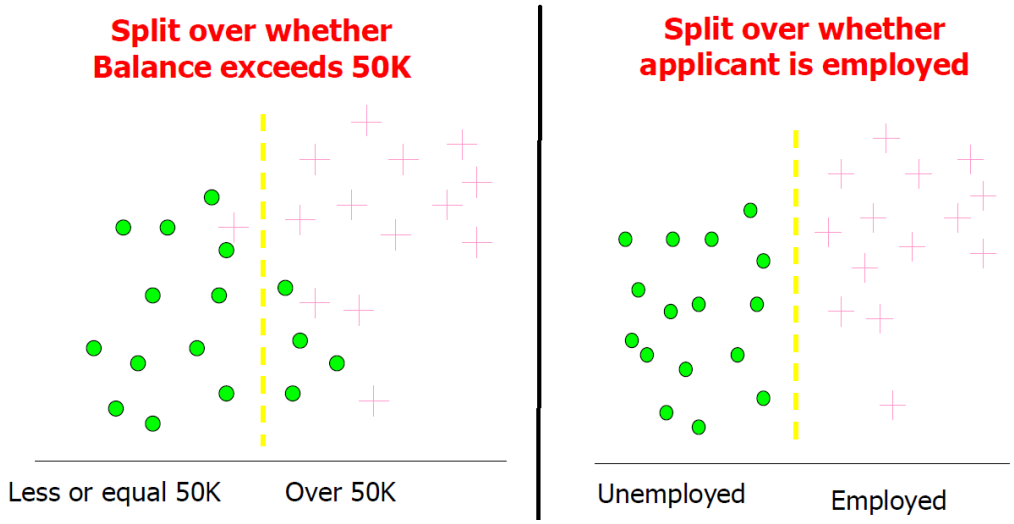

Figure 1: Data Plots for Problem 3.a.

If you have to select one of these two variables to classify the data into two classes (circle class and plus class), which one would you select?

- Employment status

Is there any approach/criterion that you can use to support your selection? Explain your answer.

- Better split of the data. One simple approach to support this selection is to tally the number of misclassified cases given each approach. The classifier with the least number of misclassified points is the best.

- Assuming that the balance classifier classifies points over 50K as pluses and under 50K as circles: Splitting over whether the balance exceeds 50K misclassifies 4 circles as pluses and 1 plus as a circle. Therefore the total number of misclassified points is (4c + 1p) = 5 out of a total of (16c + 15p) = 31.

- Assuming that the employment classifier classifies points as circles if unemployed and pluses if employed: Splitting this way misclassifies zero points. The accuracy of this classifier is 100%.

- ErrorRate(BalanceClassifier) = 5/31 ; ErrorRate(EmploymentClassifier) = 0

- This is a simple example and can be classified well with this simple approach. But 100% accuracy isn't always the best choice, as it can be a sign of overfitting. The approach of tallying

misclassified points and choosing the classifier that misclassifies the least number of points won't always work perfectly.

b) (8 points) For the data in Figure 2 with three variables (X, Y, and Z) and two classes (I and II): which variable you would choose to classify the data?

- The variable Y should be used to classify the data. In this example, when Y = 1, C = I and when Y = 0, C = II.

Show all the steps of your calculations and interpret your answer.

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

Figure 2: Data for Problem 3.b

- Determining which variable can be used to best classify the data is a feature selection problem. In this case the variables are discrete (1 or 0), so relevance analysis can be applied to determine that Y is the best variable to classify the data out of X, Y, and Z.
- When doing relevance analysis, we can calculate the total information of the data and the entropy for a particular attribute. The best attribute is the one with the lowest entropy and therefore highest information gain (information gain = total information – entropy).

Calculate entropy and gain for each attribute and the one with the lowest entropy and gain closest to 1 is the best attribute.

- # of classes = 2 (CI and CII)
- # of attributes = 3 (X, Y, and Z)
- # of cases = 4
- # of splitting conditions = 2 (1 or 0)
- Steps: 1. Test on X, 2. Test on Y, 3. Test on Z, 4. Choose the best attribute for classification.
- 1. Test on X
    - X=1 → I I II → (3/4) of the total and (2/3 I) to (1/3 II)

- X=0 → II → (1/4) of the total and (1/1 II)
  - $E_{x=1} = -(2/3)\log_2(2/3) – (1/3)\log_2(1/3) = 0.9183$
  - $E_{x=0} = -(1/1)\log_2(1/1) = 0$
- Gain(X) = 1 – (3/4)(0.9183) – (1/4)(0) = 0.3113

- 2. Test on Y
  - Y=1 → I I → (2/4) of the total and (2/2 I)
  - Y=0 → II II → (2/4) of the total and (2/2 II)
    - $E_{Y=1} = -(2/2)\log_2(2/2) = 0$
    - $E_{Y=0} = -(2/2)\log_2(2/2) = 0$
  - Gain(Y) = 1 – (2/4)(0) – (2/4)(0) = 1

- 3. Test on Z
  - Z=1 → I I II → (3/4) of the total and (2/3 I) to (1/3 II)
  - Z=0 → II → (1/4) of the total and (1/1 II)
    - $E_{Z=1} = -(2/3)\log_2(2/3) – (1/3)\log_2(1/3) = 0.9183$
    - $E_{Z=0} = -(1/1)\log_2(1/1) = 0$
  - Gain(Z) = 1 – (3/4)(0.9183) – (1/4)(0) = 0.3113

- 4. Choose the best attribute:
  - Y is the best attribute. It has Information Gain = 1 and Entropy 0 for both splitting criteria. This means that in a decision tree, splitting for Y=1 and Y=0 results in to homogenous nodes, containing I and II classes respectively.
  - X and Z are equal in terms of classification effectiveness. Both have Information gains of 0.3113.

Problem 4 (10 points): Download the Spotify Dataset along with the description from D2L.

a) (5 points) Describe the data in terms of number of attributes, number of cases, class distribution. Is there any correlation between features? Explain your answer.

- Number of attributes: 18
    - Id, name, uri, artist(s), acounticness, danceability, duration_ms, energy, instrumentalness, key, liveness, loudness, mode, speechiness, tempo, time signature, valence, mood(s)
- Number of cases: 1420
- Class distribution: class distributions for mood(s), mode, time_signature, and key are evaluated. I didn't evaluate the class artists for example because there are many artists, and insights were minimal from this.
- The class mood(s) contains 4 classes (dinner, sleep, party, and workout). A song can fall under multiple classes this dataset. Songs belonging to the mood class 'dinner' appear most often at 32.9%. Interestingly, a 'dinner, workout' song is present in the dataset. I find this to be an odd mood classification, and it should be investigated further to determine if it is a misclassified song. I present the summarized table below:

Table 6: Class distribution 'mood(s)'

	Frequency	Percent	Valid Percent	Cumulative Percent

Valid	dinner	467	32.9	32.9	32.9
	dinner, party	3	.2	.2	33.1
	dinner, workout	1	.1	.1	33.2
	party	225	15.8	15.8	49.0
	party, workout	52	3.7	3.7	52.7
	sleep	362	25.5	25.5	78.2
	workout	310	21.8	21.8	100.0
	Total	1420	100.0	100.0	

- Mode can also be considered a class as songs are either in major or minor. Most songs, 58.8%, are in major. The distribution for the class 'mode' is summarized in the table below.

Table 7: Class distribution 'mode'

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	585	41.2	41.2	41.2
	1	835	58.8	58.8	100.0
	Total	1420	100.0	100.0	

- The class 'time_signature' has integer values between 1 and 5 and 87.3% of songs had a time signature of 4. The rest is summarized below in Table 8.

Table 8: Class distribution 'time_signature'

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	21	1.5	1.5	1.5
	3	126	8.9	8.9	10.4

	4	1239	87.3	87.3	97.6
	5	34	2.4	2.4	100.0
	Total	1420	100.0	100.0	

- The class ‘key’ has integer values between 0 and 11, and fairly even percentages across keys. The majority of songs were in the first key (1) at 12.7% and the key (3) had the lowest frequency at 4.1%.

Table 9: Class distribution ‘key’

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	154	10.8	10.8	10.8
	1	181	12.7	12.7	23.6
	2	126	8.9	8.9	32.5
	3	58	4.1	4.1	36.5
	4	104	7.3	7.3	43.9
	5	142	10.0	10.0	53.9
	6	86	6.1	6.1	59.9
	7	146	10.3	10.3	70.2
	8	104	7.3	7.3	77.5
	9	113	8.0	8.0	85.5
	10	89	6.3	6.3	91.8
	11	117	8.2	8.2	100.0
	Total	1420	100.0	100.0	

- The 5 Number summary for each numeric variable is presented below to further describe the distribution of the dataset.

		acousticness	danceability	duration_ms	energy	instrumentalness	key
N	**Valid**	1420	1420	1420	1420	1420	1420
	Missing	0	0	0	0	0	0
Median		0.265	0.588	226800	0.5965	0.002075	5
Minimum		0.0000228	0.0585	54333	0.00154	0	0
Maximum		0.996	0.967	4500037	1	0.996	11
Percentiles	**25**	0.031375	0.44725	196043.75	0.31725	1.0225E-06	2
	75	0.815	0.691	274670.25	0.801	0.7985	8

		liveness	loudness	mode	speechiness	tempo	time_signature	valence
N	**Valid**	1420	1420	1420	1420	1420	1420	1420
	Missing	0	0	0	0	0	0	0
Median		0.121	-7.811	1	0.04875	118.0165	4	0.3645
Minimum		0.0227	-41.808	0	0.0229	52.799	1	0.00001
Maximum		0.979	-0.75	1	0.52	213.973	5	0.974
Percentiles	**25**	0.096625	-14.41	0	0.037	95.04975	4	0.18625
	75	0.23175	-5.0335	1	0.081475	131.5045	4	0.588

- Correlation between features: Yes. To summarize the findings:
    - The following pairs did not meet the significance test value of 0.05: Duration with energy, key with acousticness/danceability/energy/instrumental, liveness with key, mode with duration/instrumental/liveness/loudness, speechiness with duration, tempo with key/liveness/mode, time signature with key/liveness/mode.
    - All other pairs were correlated with significance at the 0.05 level and many at the 0.01 level.
    - Energy and acousticness were the most negatively correlated with an r = -0.816.
    - Loudness and energy were the most positively correlated with an r value of 0.777.
- The correlation matrix is displayed below. Because of the large number of attributes, the correlation matrix displayed in SPSS is not ideal for presenting the correlations. I display it below but it has been split to fit on the page. A heatmap would be a better visualization method for this particular problem, but I was unable to create one in SPSS.

Correlations

		index	acousticness	danceability	duration_ms	energy	instrumental	key	liveness	l
index	Pearson Corr	1	-.245**	-0.008	.065*	.223**	.106**	0.001	.172**	
	Sig. (2-tailed)		0	0.762	0.014	0	0	0.972	0	
	N	1420	1420	1420	1420	1420	1420	1420	1420	
acousticness	Pearson Corr	-.245**	1	-.526**	.056*	-.816**	.566**	-0.042	-.217**	.
	Sig. (2-tailed)	0		0	0.035	0	0	0.113	0	
	N	1420	1420	1420	1420	1420	1420	1420	1420	
danceability	Pearson Corr	-0.008	-.526**	1	-.302**	.436**	-.569**	0.031	-.105**	.
	Sig. (2-tailed)	0.762	0		0	0	0	0.24	0	
	N	1420	1420	1420	1420	1420	1420	1420	1420	
duration_ms	Pearson Corr	.065*	.056*	-.302**	1	0.046	.155**	-.071**	.180**	.
	Sig. (2-tailed)	0.014	0.035	0		0.08	0	0.008	0	
	N	1420	1420	1420	1420	1420	1420	1420	1420	
energy	Pearson Corr	.223**	-.816**	.436**	0.046	1	-.538**	0.045	.332**	.
	Sig. (2-tailed)	0	0	0	0.08		0	0.091	0	
	N	1420	1420	1420	1420	1420	1420	1420	1420	
instrumental	Pearson Corr	.106**	.566**	-.569**	.155**	-.538**	1	-0.014	-.062*	.
	Sig. (2-tailed)	0	0	0	0	0		0.599	0.02	
	N	1420	1420	1420	1420	1420	1420	1420	1420	
key	Pearson Corr	0.001	-0.042	0.031	-.071**	0.045	-0.014	1	0.033	
	Sig. (2-tailed)	0.972	0.113	0.24	0.008	0.091	0.599		0.209	
	N	1420	1420	1420	1420	1420	1420	1420	1420	
liveness	Pearson Corr	.172**	-.217**	-.105**	.180**	.332**	-.062*	0.033	1	.
	Sig. (2-tailed)	0	0	0	0	0	0.02	0.209		
	N	1420	1420	1420	1420	1420	1420	1420	1420	
loudness	Pearson Corr	0.052	-.724**	.652**	-.203**	.777**	-.726**	0.021	.111**	
	Sig. (2-tailed)	0.051	0	0	0	0	0	0.429	0	
	N	1420	1420	1420	1420	1420	1420	1420	1420	
mode	Pearson Corr	-.103**	.077**	-.067*	0.042	-.055*	-0.026	-.178**	-0.018	
	Sig. (2-tailed)	0	0.004	0.011	0.116	0.038	0.324	0	0.5	
	N	1420	1420	1420	1420	1420	1420	1420	1420	
speechiness	Pearson Corr	.341**	-.319**	.208**	-0.014	.282**	-.263**	.088**	.128**	.
	Sig. (2-tailed)	0	0	0	0.598	0	0	0.001	0	
	N	1420	1420	1420	1420	1420	1420	1420	1420	
tempo	Pearson Corr	.074**	-.220**	.146**	-.119**	.211**	-.173**	-0.044	0.014	.
	Sig. (2-tailed)	0.005	0	0	0	0	0	0.099	0.609	
	N	1420	1420	1420	1420	1420	1420	1420	1420	
time_signatu	Pearson Corr	0.017	-.254**	.296**	-.076**	.238**	-.261**	0.021	0.023	.
	Sig. (2-tailed)	0.524	0	0	0.004	0	0	0.437	0.391	
	N	1420	1420	1420	1420	1420	1420	1420	1420	
valence	Pearson Corr	-.162**	-.365**	.627**	-.216**	.400**	-.505**	.083**	-.067*	.
	Sig. (2-tailed)	0	0	0	0	0	0	0.002	0.012	
	N	1420	1420	1420	1420	1420	1420	1420	1420	

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

* Correlation is significant at the 0.05 level (2-tailed).

Correlations

		loudness	mode	speechiness	tempo	time_signatu	valence
index	Pearson Corr	0.052	-.103**	.341**	.074**	0.017	-.162**
	Sig. (2-tailed)	0.051	0	0	0.005	0.524	0
	N	1420	1420	1420	1420	1420	1420
acousticness	Pearson Corr	-.724**	.077**	-.319**	-.220**	-.254**	-.365**
	Sig. (2-tailed)	0	0.004	0	0	0	0
	N	1420	1420	1420	1420	1420	1420
danceability	Pearson Corr	.652**	-.067*	.208**	.146**	.296**	.627**
	Sig. (2-tailed)	0	0.011	0	0	0	0
	N	1420	1420	1420	1420	1420	1420
duration_ms	Pearson Corr	-.203**	0.042	-0.014	-.119**	-.076**	-.216**
	Sig. (2-tailed)	0	0.116	0.598	0	0.004	0
	N	1420	1420	1420	1420	1420	1420
energy	Pearson Corr	.777**	-.055*	.282**	.211**	.238**	.400**
	Sig. (2-tailed)	0	0.038	0	0	0	0
	N	1420	1420	1420	1420	1420	1420
instrumental	Pearson Corr	-.726**	-0.026	-.263**	-.173**	-.261**	-.505**
	Sig. (2-tailed)	0	0.324	0	0	0	0
	N	1420	1420	1420	1420	1420	1420
key	Pearson Corr	0.021	-.178**	.088**	-0.044	0.021	.083**
	Sig. (2-tailed)	0.429	0	0.001	0.099	0.437	0.002
	N	1420	1420	1420	1420	1420	1420
liveness	Pearson Corr	.111**	-0.018	.128**	0.014	0.023	-.067*
	Sig. (2-tailed)	0	0.5	0	0.609	0.391	0.012
	N	1420	1420	1420	1420	1420	1420
loudness	Pearson Corr	1	-0.034	.252**	.262**	.299**	.488**
	Sig. (2-tailed)		0.201	0	0	0	0
	N	1420	1420	1420	1420	1420	1420
mode	Pearson Corr	-0.034	1	-.081**	-0.015	-0.008	-.064*
	Sig. (2-tailed)	0.201		0.002	0.572	0.75	0.015
	N	1420	1420	1420	1420	1420	1420
speechiness	Pearson Corr	.252**	-.081**	1	.145**	.122**	.150**
	Sig. (2-tailed)	0	0.002		0	0	0
	N	1420	1420	1420	1420	1420	1420
tempo	Pearson Corr	.262**	-0.015	.145**	1	.054*	.094**
	Sig. (2-tailed)	0	0.572	0		0.042	0
	N	1420	1420	1420	1420	1420	1420
time_signatu	Pearson Corr	.299**	-0.008	.122**	.054*	1	.180**
	Sig. (2-tailed)	0	0.75	0	0.042		0
	N	1420	1420	1420	1420	1420	1420
valence	Pearson Corr	.488**	-.064*	.150**	.094**	.180**	1
	Sig. (2-tailed)	0	0.015	0	0	0	
	N	1420	1420	1420	1420	1420	1420

** Correlation is significant

* Correlation is significant

b) (5points) Report the ranges for each numerical variable. Would you recommend to normalize the data? If yes, which approach would you apply? Justify your answer.

- Ranges:
  - Acousticness: [0.0, 1.0]
  - Danceability: [0.0, 1.0]
  - Duration_ms: [0, +inf]
  - Energy: [0.0, 1]
  - Instrumentalness: [0.0, 1]
  - Key: [0, 11] (int)
  - Liveness: [0, 1]
  - Loudness: [-60, 0]
  - Mode: [0, 1] (int)
  - Speechiness: [0, 1.0]
  - Tempo: [0, +inf]
  - Time_signature: [1, +inf] (int) * observed range [1,5]
  - Valence: [0.0, 1.0]
- Normalize the data: Yes.
  - Normalizing the data is performed to give all the attributes an equal weight. Most of these attributes are already on similar ranges, but it wouldn't hurt to use min-max normalization on a few attributes to put everything on the range of [0.0, 1].
- Specifically, I would recommend duration_ms, loudness, and tempo be normalized using min-max normalization and mapped to the new range of [0.0, 1].
- Key, mode, and time_signature don't need to be normalized, as they can be used as classes. But depending on what you want to do with the data they can also be min-max normalized.
- Also, speechiness should probably be binned in ranges [66, +inf), [0.33,0.66), and [0, 0.33) according to the spotify_data_characteristics page. These bins indicate the level of spoken words present in the song. This discretization could be useful in the 'data reduction step'.