16. Calculate the midrange stock price.

The midrange stock price is the central price for the entire price range and is formulated as follows:

MidRangeX = [Max(X) + Min(X)] / 2

For the problem at hand we have as follows:

MidRangeX = ($75 + $4) / 2 = ($79) / 2 = **$39.5**

17. Compute the Z-score standardized stock price for the stock worth $20.

Z-Score standardization scales an observation where the mean value is zero, the SD is 1 and most values lie between -4 and 4 (this value has no units) and is formulated as follows:

Z-Score(X) = [X_i – Mean(X)] / |SD(X)|

Given the mean of $17 (see Exercise #13) and |SD| of 18.03 (see Exercise #14), The Z-Score for the stock price of $20 is calculated as follows:

Z-Score($20) = ($20 - $17) / $18.03 = ($3) / $18.03 = **0.1664**.

Please note that this value makes sense as it is slightly greater than zero just as $20 is slightly greater than $18.03.

Solutions to Chapter 7
k-NEAREST NEIGHBOR ALGORITHM
Prepared by James Cunningham, Graduate Assistant

1. Clearly describe what is meant by classification.

Classification is a type of predictive modeling where the target variable is categorical having a predefined finite set of possible outcomes or ***classes***. The objective of the classification model is to classify each new observation determining which class each new observation belongs to.

2. What is meant by the term instance-based learning?

Instance-based learning is a type of learning where the training data is stored in memory and new observations are classified or estimated by comparing them to the training data in memory.

3. Make up a set of three records, each with two numeric predictor variables and one categorical target variable, so that the classification would not change regardless of the value of k.

The table below identifies three customers that have purchased cars at a particular dealership and the dealer wants to be able to classify new customers in order to assign the sales staff member who has the greatest expertise in the car type of interest. If we were to use KNN in order to classify a new potential buyer based on the nearest neighbors below, then our choice of k will have no effect on the classification as all three neighbors are in the same class (i.e. the SUV class).

Customer	Age	Height (Inches)	CarType
1	25	70	SUV
2	35	68	SUV
3	37	69	SUV

Table 3.1. Set of 3 records – two numeric variables and one categorical variable

4. **Refer to Exercise 3. Alter your data set so that the classification changes for different values of k.**

   If we were to alter the set of neighbors depicted in the below to classify a new potential buyer, then our choice of k will now affect the classification as these three neighbors are in different classes (i.e. the SUV and Sedan classes). When k=1, the new record will be classified as either an SUV-buyer or a Sedan-buyer. When k=3 (I am deliberately saving k=2 for last), the new record will be classified as an SUV-buyer with 66.67% confidence. Finally, when k=2, the model would need to employ additional criteria for breaking the potential tie that could arise when the two nearest neighbors were either {1,3} or {2,3}.

Customer	Age	Height (Inches)	CarType
1	25	70	SUV
2	35	72	SUV
3	37	68	Sedan

Table 4.1. Set of 3 records – altered classification

5. **Refer to Exercise 4. Find the Euclidean distance between each pair of points. Using these points, verify that Euclidean distance is a true distance metric.**

Using the records in Exercise #4, we calculate the Euclidean distance between each of the three points. The results are given the table below.

Distance Measurement	Euclidean Distance
D(1,2)	$\sqrt{(25-35)^2+(70-72)^2} = \mathbf{10.1980}$
D(1,3)	$\sqrt{(25-37)^2+(70-68)^2} = \mathbf{12.1655}$
D(2,3)	$\sqrt{(35-37)^2+(72-68)^2} = \mathbf{4.4721}$

Table 5.1. Euclidean distances

Referring to the table above, we observe that D(2,3) is the shortest distance. However, these two end points are classified differently (i.e. #2 is an SUV-buyer and #3 is a Sedan-buyer). This is due to the fact that the Euclidean distance metric is more heavily influenced by the predictors having the greatest variation. Since the range of age is **37-25 = 12** and the range of Height is only **72-68 = 4**, the Euclidean distances between records will be driven more by Age than by Height, which explains why D(2,3) is shortest.

If we were to standardize the predictors say using Min-Max normalization, the values would vary on similar scales. The results are given in the table below.

Customer	Age	Height (Inches)	CarType
1	0	0.05	SUV
2	0.83	1	SUV
3	1	0	Sedan

Table 5.2. Set of three records – Min-Max normalized

We then recalculate the Euclidean distances using these scaled values. The results are given in the table below.

Distance Measurement	Euclidean Distance
D(1,2)	$\sqrt{(0-0.83)^2+(0.50-1)^2}=\mathbf{0.9690}$
D(1,3)	$\sqrt{(0-1)^2+(0.5-0)^2}=\mathbf{1.1180}$
D(2,3)	$\sqrt{(0.83-1)^2+(1-0)^2}=\mathbf{1.0413}$

Table 5.3. Euclidean distances for normalized variables

Now that we are using similar scales for both predictors, we observe that the closest points are 1 and 2, which is what we would expect. Therefore, we conclude that Euclidean distance is a true distance metric, but care must be taken to transform the predictors so that they vary on similar scales.

6. Compare the advantages and drawbacks of unweighted versus weighted voting.

Unweighted voting is the simplest to implement, new observations are classified simply by the majority class of its k nearest neighbors. While this works well for training sets that have good separation by class, it works poorly when the training set does not have good separation by class (i.e. the classes of the training points are thoroughly mixed from a geometric perspective). For example, let's say we have classes A, B, and C and k=3. If the three neighbors happen to be {A,A,B}, then we classify the new observation as an 'A'. However, if the three neighbors happened to be {A,B,C}, then we would have a three-way tie.

In contrast, weighted voting is more difficult to implement, but it gives precedence to the neighbors that are closest to the new observation. This approach works well even for training sets that do not have good geometric separation by class. However it is not completely rigorous. In the rare case that there is k neighbors lying at a distance of zero from the new observation, the inverse distance would be undefined, and the prescribed approach is to fall back to majority classification, which can result in multi-way ties as in the non-weighted voting approach.

7. Why does the database need to be balanced?

It is important to ensure that the database is balanced when constructing KNN models so that he algorithm does not become limited to predicting the common classifications. The data must contain enough rare occurrences so that the model can predict similarly rare observations as well.