Dataset 1: Chocolates

ID	Cocoa (%)	Shape	Tree nuts?	Price ($/oz)	Review (stars)	Seller	Found online	Found in-store	# in bag	Color
1	30	Round	0 (No)	0.25	3	Hershey	1	0	40	Milk
2	15	Square	0	0.5	3	H & D	0	1	15	White
3	20	Square	0	0.44	2	Neuhaus	1	0	20	White
4	55.5	Square	0	0.8	4	Mars	1	0	12	Milk
5	60	Round	1 (Yes)	1.3	4	Ferrero	0	1	6	Dark
6	5.5	Round	0	1.1	3	Lindt	0	1	2	White
7	10	Square	1	0.75	5	H & D	0	1	8	White
8	70	Square	0	0.9	1	Hershey	0	1	32	Milk
9	85	Square	0	1.8	2	Mars	1	0	14	Dark
10	60	Round	1	2.4	5	Lindt	0	1	1	Dark
11	14.5	Square	0	1.5	4	Lindt	1	0	6	Dark
12	77	Round	0	1.4	3	H & D	0	1	12	Dark

Q: Given Dataset 1, plot the samples using

Price along the x-axis and

Review (stars) along the y-axis.

Draw a hard margin boundary to separate the positive class (Color = Dark) from the negative class (Color != Dark).

Draw the boundary and the margins.

If a hard margin boundary does not exist, draw a soft margin boundary that prioritizes accuracy over margin width.

Q: Use the polynomial kernel with (r = 1, d = 4) to calculate distance between the following samples, using only features **Cocoa (%)**, **Price ($/oz)**, **Review (stars)**, and **# in bag**.

Sample 1 and Sample 2

Sample 3 and Sample 4

Sample 5 and Sample 6

Q: Use the RBF kernel with (γ = 0.1) to calculate similarity between the following samples, using only features **Cocoa (%)**, **Price ($/oz)**, **Review (stars)**, and **# in bag**.

Sample 7 and Sample 8

Sample 9 and Sample 10

Sample 11 and Sample 12

Q: Use Manhattan distance as your distance metric, k-nearest neighbors with k=5, and the features **Cocoa (%)** and **# in bag**, for classification.
Predict **Color** of a sample with 45% Cocoa and 10 pieces in the bag

Q: Use Manhattan distance as your distance metric, k-nearest neighbors with k=3, and the features **Cocoa (%)** and **# in bag**, for regression.
Predict **Price ($/oz)** of a sample with 45% Cocoa and 10 pieces in the bag

Q: Use the feature **Review (stars)** to build the best Decision tree with a single split.
Predict **Color** of a sample with a 2-star review

Q: Use the feature **Review (stars)** to build the best Decision tree with a single split.
Predict **Price ($/oz)** of a sample with a 4-star review

Q: What would be the next step if I plan on making a Random Forest model?
What about an Adaboost model?

Q: Train a Naïve Bayes Classifier to predict **Shape** from **Seller** and **Color**
Predict the **Shape** of a sample with Color = White and Seller = H & D

Q: What differentiates PCA from LDA? In what situation can you use PCA, but cannot use LDA?

Q: Cluster the data using only **Cocoa (%)**, **Price ($/oz)**, **Review (stars)**, **# in bag** with the Euclidean distance metric and Agglomerative Hierarchical clustering. Stop when the single-linkage and complete-linkage would result in a different merge

Q: Cluster the data using only **Cocoa (%)**, **Price ($/oz)**, **Review (stars)**, **# in bag** with the Euclidean distance metric and k-means where k=2

Q: Name an alternative clustering method and what advantage it would have over k-means and agglomerative hierarchical clustering. This could be the same advantage, or a different advantage for each.

Dataset 2: Grocery store transactions

ID	Itemset				Support
1	Bread				80%
2	Milk				35%
3	Leafy vegetables				25%
4	Root vegetables				25%
5	Bread	Milk			30%
6	Bread	Root vegetables			20%
7	Root vegetables	Milk			15%
8	Bread	Leafy vegetables			5%
9	Leafy vegetables	Milk			5%
10	Root vegetables	Leafy vegetables			5%
11	Root vegetables	Bread		Milk	10%
12	Leafy vegetables	Bread		Milk	3%
13	Root vegetables	Leafy vegetables		Bread	2%
14	Root vegetables	Leafy vegetables		Milk	1%
15	Leafy vegetables	Root vegetables	Bread	Milk	1%

Q: Find all informative rules from this dataset where itemsets are frequent if Support >= 10%, and rules are strong if Confidence >= 55%