1. The training data is used to train the neural network, the testing data (remaining data) is used to see how well the network performs on other images/data that is new/not familiar. The purpose is to teach the computer to pick up patterns/equations, and be able to use them independently. The images variables are pixel arrays while the labels are numbers which represent the clothing.
2. Relu helps ensure that there are no negative numbers. So it sets any number that is less than 0, to 0. Softmax is usually used in the last layer, and sets the largest value to 1, and the rest to 0. This helps find the most likely answer since less data needs to be sorted through. There are 10 neurons in the last layer because there were 10 pieces of clothing, they are the output. Each neuron gives the probability that its correlated article is in the class.
3. The loss function finds the amount of error and works to minimize it. The optimizer then looks at the results from the loss function, and attempts to make another estimate with less error. The loss function will then analyze it, and the cycle continues. The loss function is categorical because the items which are being analyzed are also categorical. You cannot really measure them continuously.
4.
    a. What is the shape of the images training set: (60,000,28,28)
    b. What is the length of the labels training set? 6000
    c. What is the shape of the images test set? (1000,28,28)
    d. Estimate a probability model and apply it to the test set in order to produce the array of probabilities that a randomly selected image is each of the possible numeric outcomes:

```
array([1.5539440e-09, 8.2805976e-12, 4.7423885e-07, 4.3320365e-06,
       1.9139490e-12, 2.6702085e-10, 1.1747801e-14, 9.9999344e-01,
       1.0673884e-06, 7.2961376e-07], dtype=float32)
```

    e. Use np.argmax() with your predictions object to return the numeral with the highest probability from the test labels dataset: 7

f.