Question 3:

- **Correct predictions:** 338/354
- **Accuracy:** 0.955

Question 4:

- Swapped training set to 20% and test set to 80%
- Correct Predictions: 1293/1415
- Accuracy: 0.914
- This is pretty much what I expected, due to a lower training set the accuracy of the predictive model is lower. This makes me question at what point does increasing the training set fail to improve accuracy, and also what quantity of tests is needed to know our accuracy number is statistically valid.

Question 5:

- To me, it makes sense why the misclassified digits were messed up by the model. I could not tell what most of them were going for based on the heatmap.

Question 6:

<u>How the data was collected:</u>

- The MNIST dataset was collected by having people write digits on paper, which were then scanned and normalized to 8x8 pixel images.
    - Issues arise during the scan process, which can introduce blur or distortions. Additionally, normalizing the scans to 8x8 pixel images causes a significant loss of detail, making it hard to distinguish between similar-looking numbers (eg. 1 vs 7)

<u>Who the data was collected from:</u>

- The data was primarily collected from American Census Bureau employees and high school students.
    - This is a narrow sample group, which can cause significant bias as the dataset may not represent how people from different countries, cultures, age groups, or educational backgrounds write digits.

<u>What the data represents:</u>

- The data represents isolated written digits in a controlled format (8x8 normalized grayscale pixels).
    - This data doesn't capture real-world complexity. For example, the surrounding text is not included, and it doesn't handle connected digits. It also doesn't account for varying writing instruments (pen vs pencil vs marker).

Question 8:

- For my initial testing value of K, I chose K = 3 mainly because it is a small odd number. I kept it small because this is just a test value, and I wanted it to be odd so that we don’t have any edge cases where a digit is closest to an equal amount of two different numbers (like 1 vote 4 and 1 vote 9).

- Accuracy was 351/354 for 3, which I believe is as good as it will get. Keeping it small ensures that we only compare it to the nearest neighbors, so the data isn't compromised by having insufficient close digits. 3 is also big enough that we can still get some good comparisons in.

Question 9:

- For the selected seeds we went with, the best_k was the same. We went with the two given numbers and our selected number; 12345, and so K = 1 was the best for all. However, with different seeds we would sometimes get K = 8 or K = 5, so while 1 might be the most common, it is still possible for other seeds to have different Ks. Thus, we chose K = 1 as the best value because it was the most common of the seeds we tried. comparison of predicted vs actual labels:

Question 10:

I only included the first 20 lines of output for the sake of time/space.

Comparison of predicted vs actual labels:

- 0 6 6
- 1 7 7
- 2 1 1
- 3 3 3
- 4 6 6
- 5 5 5
- 6 6 6
- 7 1 1
- 8 7 7
- 9 0 0
- 10 3 3
- 11 7 7
- 12 9 9
- 13 4 4
- 14 8 8
- 15 0 0
- 16 5 5
- 17 6 6
- 18 2 2
- 19 5 5

- Correct: 349 out of 353
- Accuracy: 0.989