3.
Train = first 80%, Test = last 20%
Accuracy: 0.955

4.
Train = last 80%, Test = first 20%
Accuracy: 0.969

When I swapped the sets (trained on the last 80% and tested on the first 20%), my accuracy went up slightly to 0.969. The accuracy is pretty similar overall, which means the model performs about the same no matter which side of the data is used for training. The small difference is probably because of how the digits are ordered in the dataset

5.
When I looked at the heatmaps of the first five misclassified digits, the mistakes actually made sense.

- Pred 9 vs True 8: The 8 looks a bit open at the top, so it kind of has the same shape as a 9
- Pred 1 vs True 8 (twice): The 8s are drawn faintly, and the middle part isn't very visible, so they look more like straight lines instead of loops.
- Pred 5 vs True 9: The 9 looks like it's missing the top curve, which makes it easy for the model to confuse it with a 5.
- Pred 7 vs True 3: The 3 has a sharper top and less of a bottom curve, so it looks more like a 7 in the low-resolution image.

The mix-ups happen because of how similar some digits look when they're shrunk down to just 8×8 pixels

6.
There are a few possible issues with the dataset itself:

- How it was collected: The digits were scanned or downsampled from handwritten numbers, so fine details are lost when compressed to 8×8 pixels.
- From whom it was collected: Probably a small group of writers, so the handwriting styles aren't very diverse. That makes the model less accurate on handwriting from other people.
- What it represents: Each image is just pixel intensity values, so there's no context about how people actually write digits. The low resolution and possible duplicates can also inflate accuracy a bit.

8.

- I picked the initial k using the square-root rule to match the training set size, balancing bias and variance for k-NN. This gave a solid starting point before testing.
- With this k, the model had moderate accuracy and validated the functions. The compareLabels results served as a performance baseline for comparison.

9.

- I validated with three random seeds—8675309, 5551212, and 20251109—to see if the top k stayed consistent. The best k values were similar, though not always identical, showing some variability due to splits.
- I chose the median k for stability and highest validation accuracy, which also controlled overfitting and confirmed my functions worked as intended.