**Problem One**

The code used for this problem can be found in understanding\_data.py. It can be run using the following command:

*python understanding\_data.py*

**A.** I examined 50 images of the digit 8. I chose this digit because it is a relatively intricate number with two stacked loops, and can be misinterpreted as the digit 3, if written incompletely, or the digit 0 if written too thickly. With this intuition, I examined the set of images, and identified the following pictures to be potential difficult cases.



The above three figures demonstrate a key challenge with classifying the digit 8. When written messily, often times either one or two of the loops is not fully connected. Because the fundamental characteristic of the letter 8 is two closed loops, the digit becomes a suddenly very complex image of curling lines. One could easily misinterpret an unclosed digit 8 with a messy digit 3, which likely has a complementary issue of accidentally closing loops.



The above three figures demonstrate another key challenge when classifying the digit 8. When written extremely thickly, the loops become hard to identify. In turn, the classifier is likely to struggle when discerning a thick blob, and may assume the bottom loop is a thick line (far left), or perhaps a messy 0. With that, it is also challenging when handling finely written 8s (far right). If one loop is larger and significantly clearer, the classifier may only capture it and the thin stroke would make it increasingly difficult for the classifier to even recognize the smaller, messier loop, leading to misclassifications.

**B.**

The total number of images for digit 0: 5923

The total number of images for digit 1: 6742

The total number of images for digit 2: 5958

The total number of images for digit 3: 6131

The total number of images for digit 4: 5842

The total number of images for digit 5: 5421

The total number of images for digit 6: 5918

The total number of images for digit 7: 6265

The total number of images for digit 8: 5851

The total number of images for digit 9: 5949

The total number of images: 60000

I made the training set such that it consists of ¾ of each digit’s images, enabling the other ¼ of each digit’s images to be put in the testing set. While within each digit image set the partition is relatively random, I wanted to ensure that each digit was properly represented in both the training and testing set. As such, the classifier should be trained for all digits within the testing set, in order to appropriately assess its ability in classifying unseen data within the same realm. Moreover, I chose this distribution such that the model is exposed to a substantial number of images for each digit during training. As shown above, there is a large variation in handwriting for each given digit. As such, it is important the classifier sees a breadth of data during classification, so it can further refine its interpretation of each digit and understand the disparities.

**Problem Two**