***MNIST dataset:*** *Exploratory Data Analysis of Handwritten digits classification.* by import-pandas-as-pd

This blog post intends to explore and understand MNIST dataset by conducting a comprehensive exploratory data analysis (EDA). The dataset is a collection of greyscale pixel images- each representing a handwritten numerical digit from 1-9. Each image is represented as a pixel, and each pixel is represented by an integer between 0 and 255 indicating the brightness of the pixel - with the label of the image i.e the numerical digit it illustrates also being presented.

***Understanding the MNIST Dataset :***

We shall be using two CSV files for the MNIST data set; ***mnist\_test***and ***mnist\_train*** to perform our EDA.

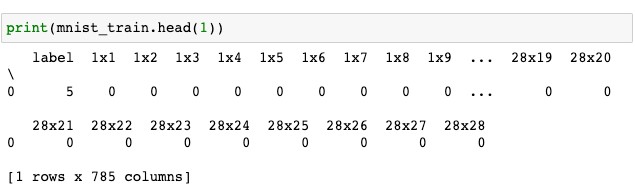
***mnist\_train*** Shape: (60000, 785) ***mnist\_test*** Shape: (10000, 785)

***mnist\_train*** Rows: 60000 ***mnist\_test*** Rows: 10000

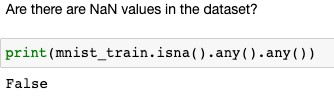
***mnist\_train*** Columns: 785 ***mnist\_test*** Columns: 785

To begin with, we load the data into the appropriate variables and display its shape to get a sense of the dimensions of the data. We find that the dataset consists of 70,000 images of handwritten digits, with 60,000 images in the training set and 10,000 images in the test set. The exact number of pixels per image is 784 represented as an array of integers which suggests a 28 X 28-pixel image size.

Here's a general shape of the data represented by visualizing the head of the dataset **mnist\_train :**



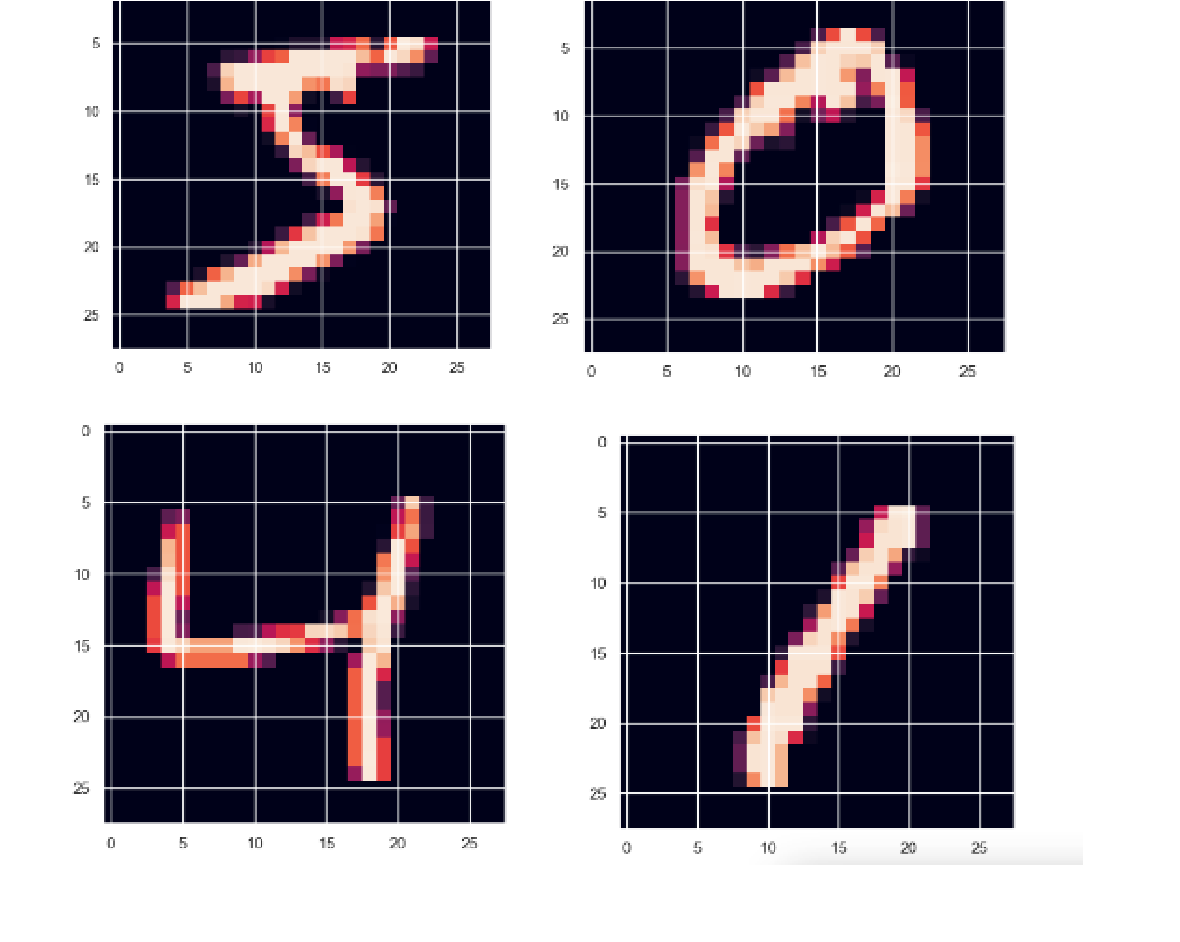
We ensure our data doesn't contain any **NA** Values :



An important clarification is realized when the data contains no NaN or missing values: the data is clean, and we can pursue exploratory data analysis without worries.

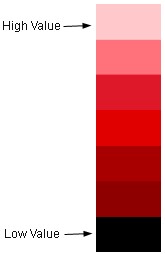
Following are the images of the first 4 instances:





**Investigating pixel data:**

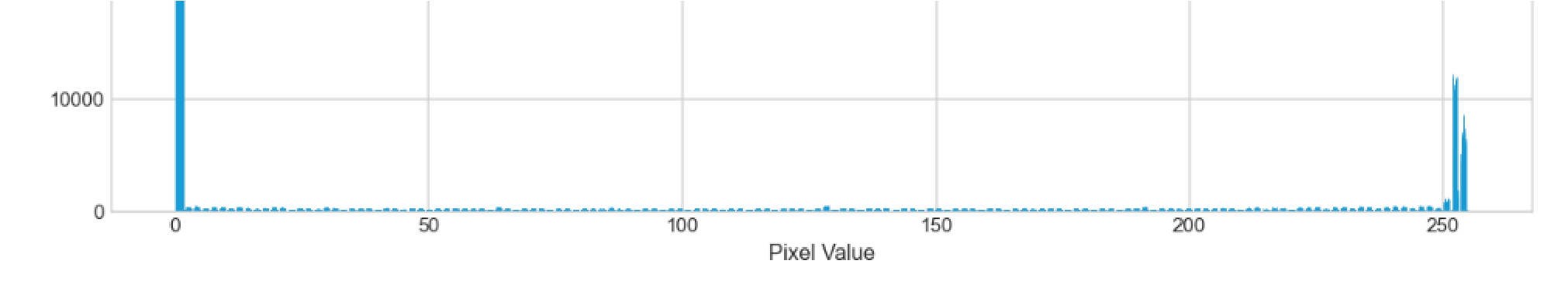
Let's first understand the concept of Intensity here. From the above pictures, we can say that the regions with white density have a greater intensity than those regions that don't have white/red density. Since the redness can be visualized, we imagine say completely white has a pixel value of 250 and completely black regions are a 0 on the intensity scale



(Pixel values ranging between 0 - 250)

To begin better understanding of MNIST dataset, we attempted to gauge the distribution of pixel values in the data, we calculated the frequency of each unique pixel value across all images. We make use of the unique pixel values and the number of occurrences of each value in the dataset to plot the following:

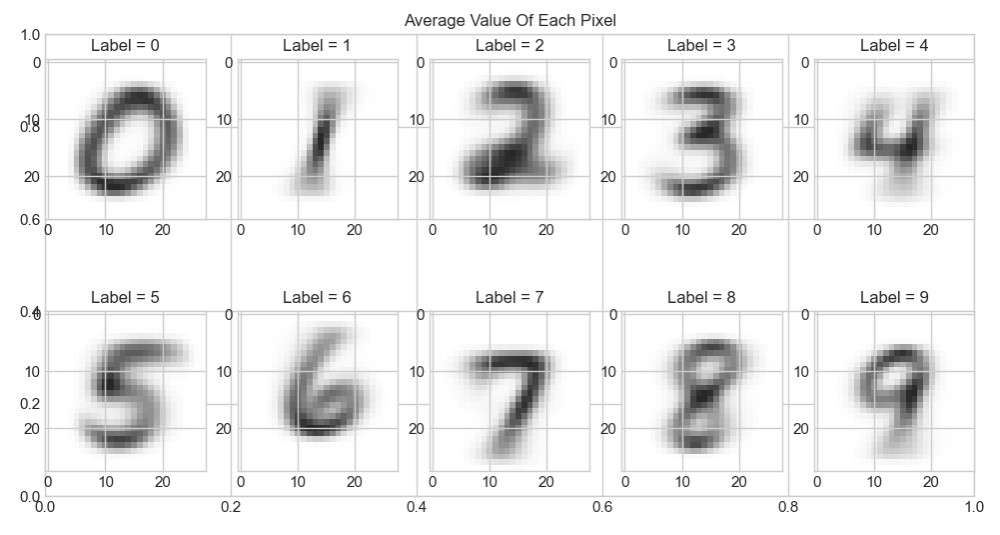




The results show a majority of the pixels in the dataset are entirely white, while a smaller number of pixels in the dataset are entirely dark. We may have observed a lot more diversity if we had been dealing with more natural images such as those with any realistic subject in them i.e, landscapes. For the current dataset, however, this suggests that we could presumably normalize the data by replacing each pixel with a more standardized range and not face much loss of data while retaining greater efficiency in storing the data itself.

This still begs the question of individual variability within each digit label. Do all instances of each label look like each other, if not what is the example of the most prevalent example of a digit?

To answer this, we can determine the mean value for each dimension within each label individually and display the results.

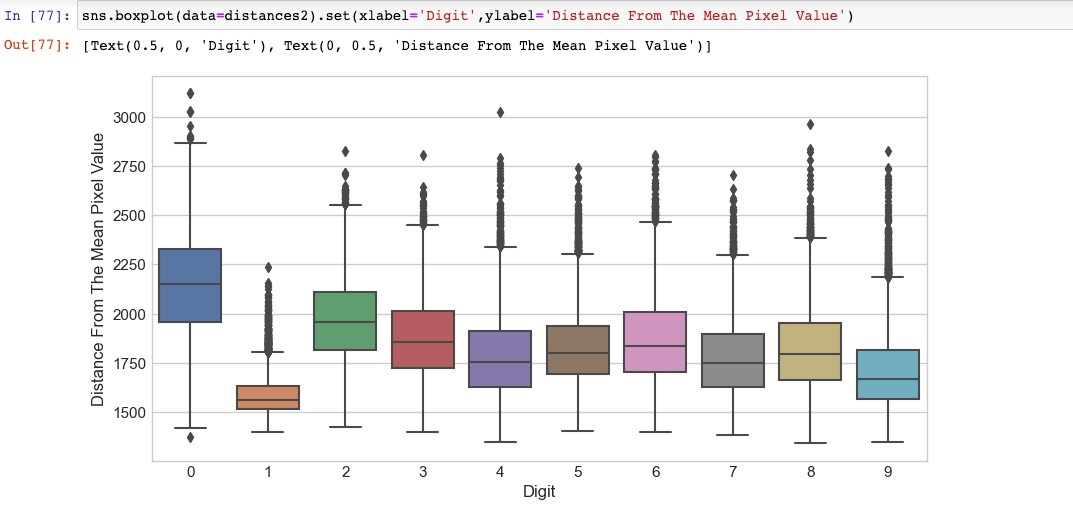


So far, this visualization doesn't show anything typically unusual, since all of the above digits are very typical versions of what we would expect. However, what about variability within each label itself? Are some labels much more di cult to classify compared to others?

**Atypical Instances in data:**

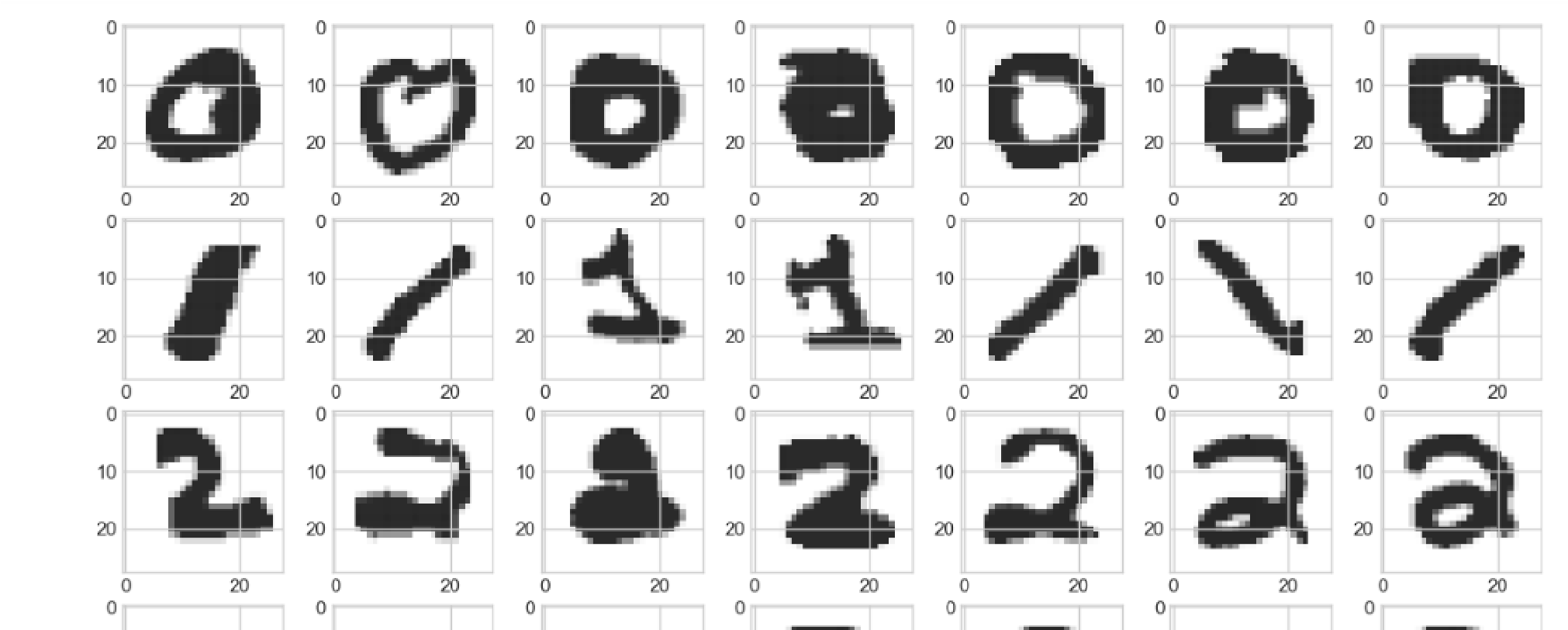
Based on intuition we already have suspicions on how some digits might be easier to classify compared to others. For example, a ‘1’ may be relatively easy to distinguish compared to the di erence between say a ‘3’ and an ‘8’.

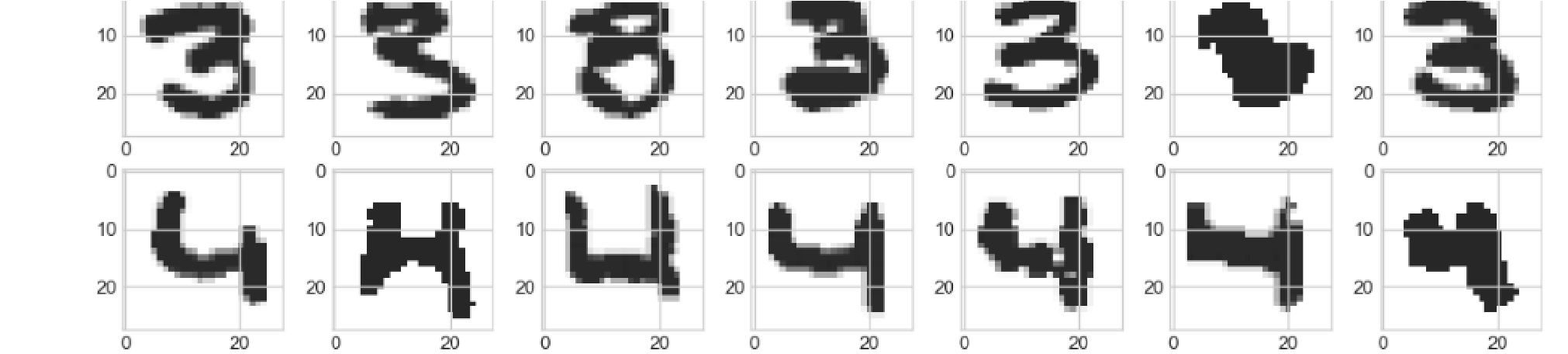
To go about visualizing this difference in the perceived instance we calculated the Euclidean distance (square root of the sum of squares) of each image to its label’s previously calculated typical instance. By doing this we were able to envision, on average, which digits have more of variability in perception and thus classification.

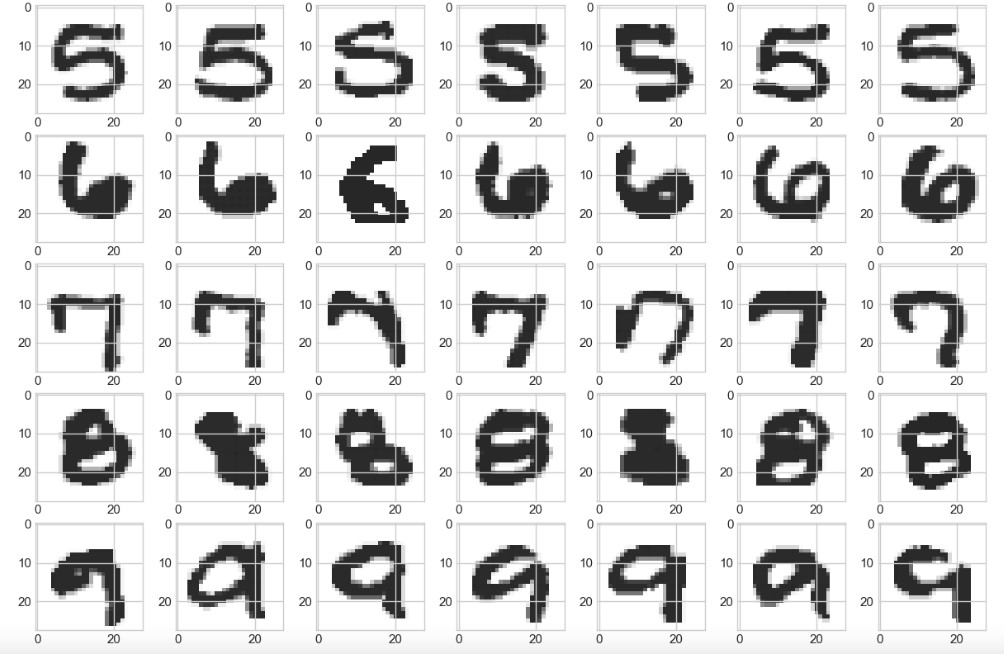


The above boxplot shows that, in fact there isn’t a lot of variation in how people depict hand-drawn digits. As expected, the distances between 1s and their typical instances appear to be the least, with digits from 3 to 9 exhibiting somewhat similar variabilities. The most variability, however, is shown to be in instances of 0s and 2s. Another thing to note, however, is that every digit has an outlier instance with an uncommonly long distance from its typical instance. What do these instances look like?

To visualize, we can create a visualization for 7 instances of each label that had the least resemblance to their most typical instance.





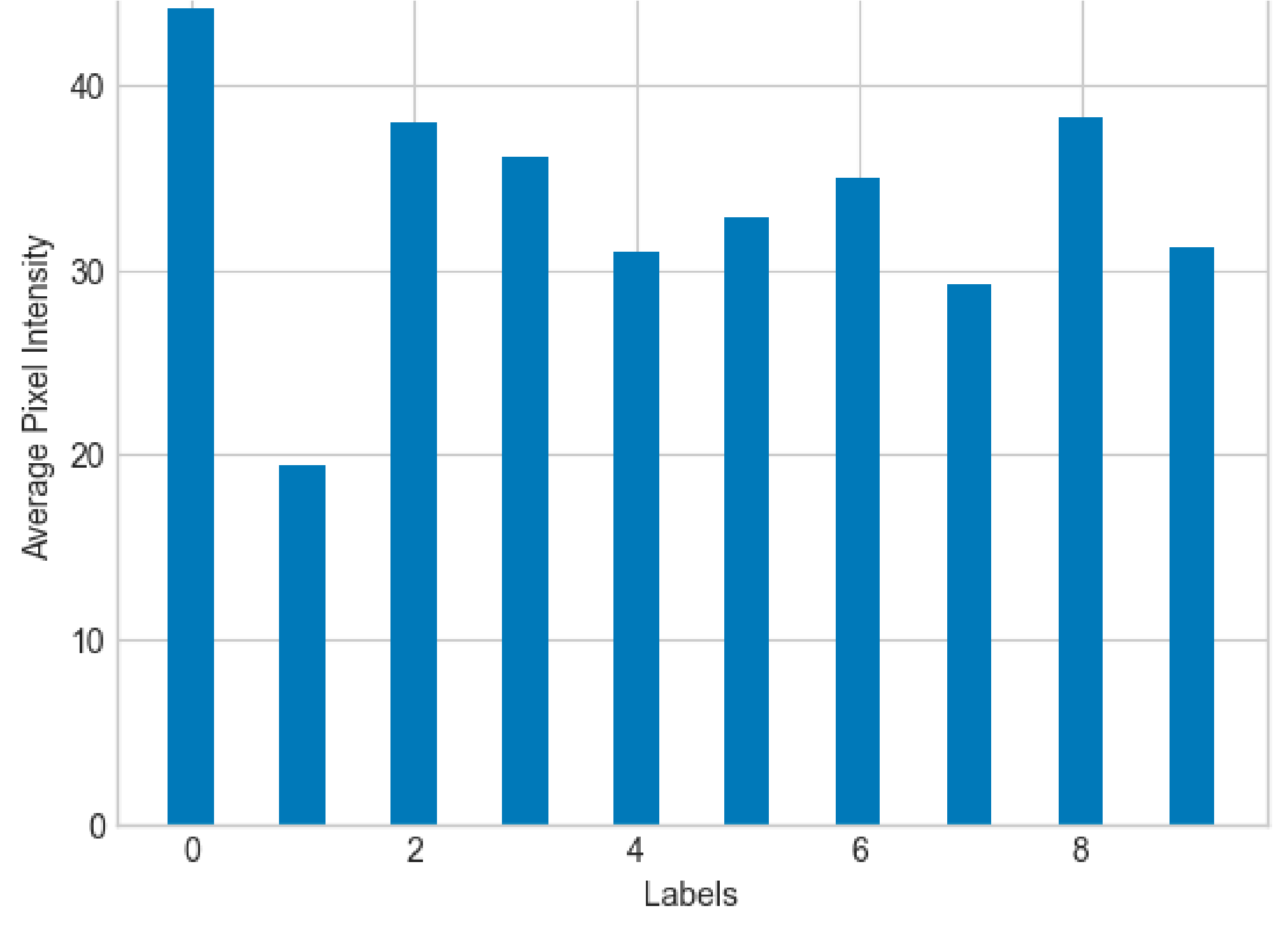


This is a helpful method to comprehend the Achilles heel of the data. For instance, while most 1s had the same appearance, they might also have a flat line with a flag on top or be drawn diagonally. A human may draw a 7 with a bar in the center instead of the way shown above and so on.

For classification situations in particular, this helps achieve a realistic sense of the accuracy it may be able to generate. Here even a human may struggle to achieve a 100% accuracy rate.

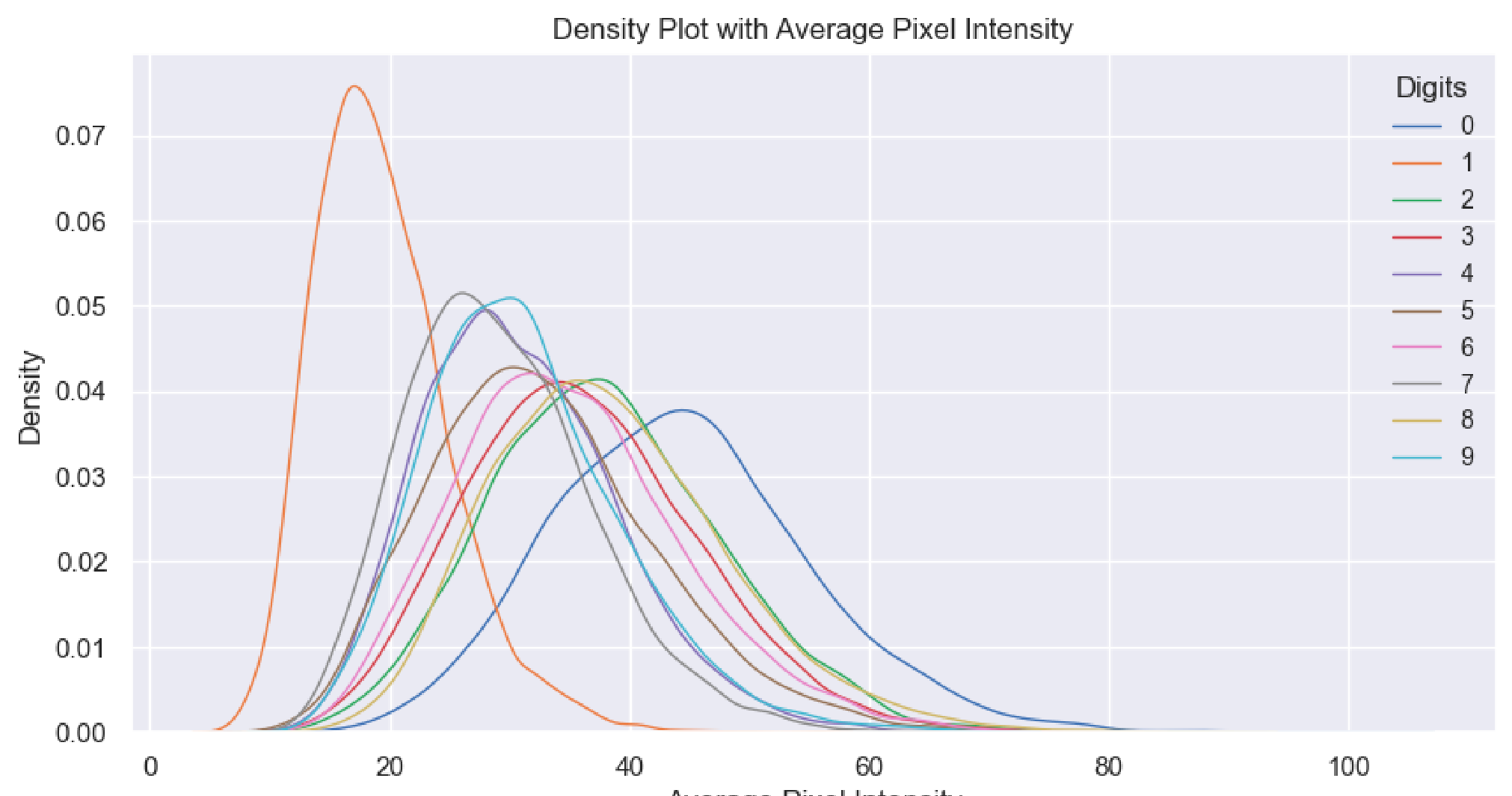
Moving on, we can gain further insights on the average intensity of different digits in an image, which refers to the average value of a pixel. Again intuition leads us to believe the digit "1" has a lower average intensity compared to an "8". To achieve this, we create a bar plot that showcases the average pixel intensity for each label.





As expected the digit "1" has the least intensity, however unexpectedly the digit "0" has the highest. This recently discovered feature seems to hold promise for making predictions, particularly in distinguishing whether a given digit represents a "1" or not.

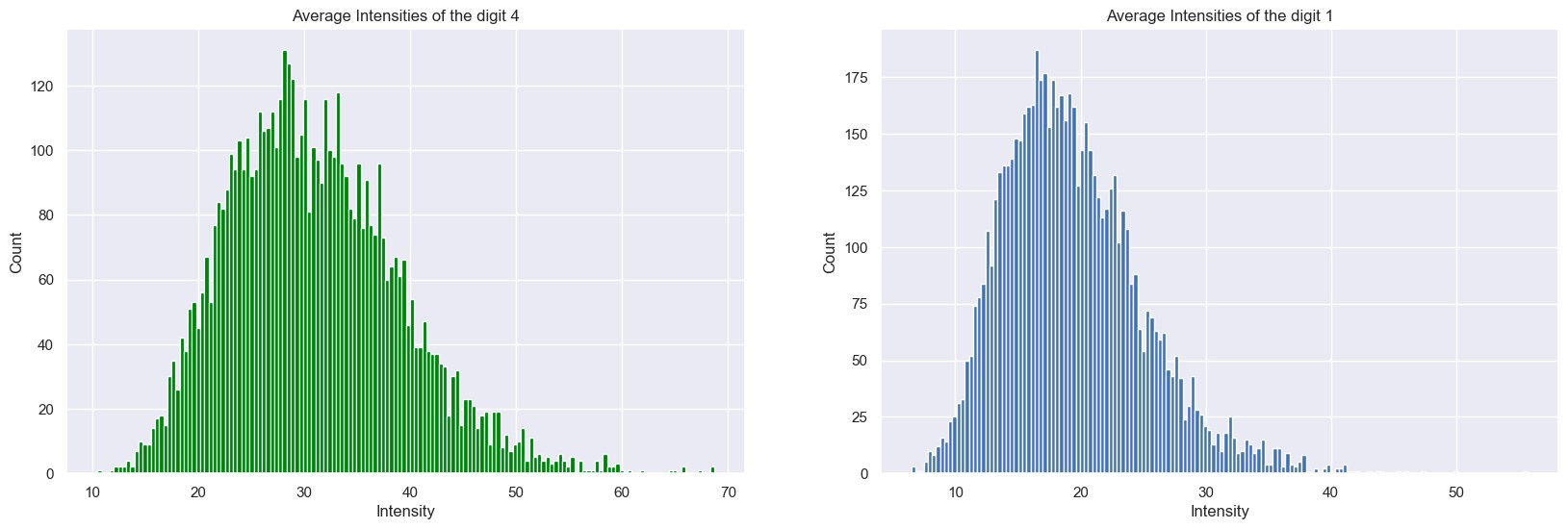
However, as hinted earlier the primary issue lies in the fact that individuals write their digits di erently. To actualize this, we employ a density plot to illustrate the average pixel intensities. This can be achieved by examining the distribution of the average intensity based on the label.





The density diagrams shown above provide us with information about several aspects. Firstly, they reveal that some intensity distributions exhibit higher variance compared to others. Additionally, the overall pattern of intensity distributions appears to follow a roughly normal distribution. Furthermore, the digit "1" appears to be the most consistently written across various scenarios. It is observed that the two digits that exhibit the most variation in writing styles are 4 and 7. For example, some individuals may write their 7s with a line across them . Additionally, there exist multiple acceptable ways to write the digit 4. Unfortunately, the intensity function does not o er much practical value beyond this.

To compensate, we can attempt to view the anticipated differences in variance between most and least various digits - "1" and "4" - through visual means and to achieve this, we utilize histograms.



As the above graphs indicate, the intensity distribution for the digit 4 is not as "normal" as that for 1. Specifically, the distribution for 4 appears to be almost bimodal, which indicates that people may write their fours in two distinct ways. This finding implies that the variability in how individuals write these digits is considerably higher than that of 1. Such results could potentially have implications for classification systems.

It may also be worthwhile to investigate whether different populations or cultures exhibit varying levels of variability in their handwriting and whether this variability impacts their ability to recognize handwritten digits effectively.