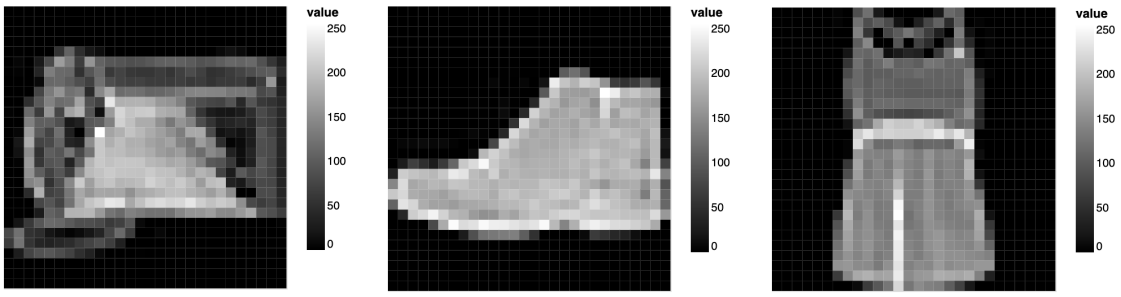
Model Analysis

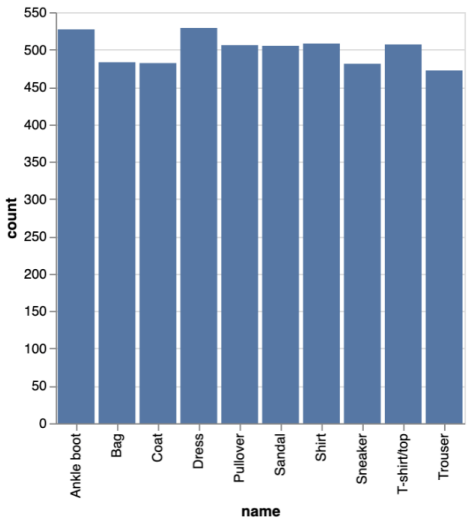
Dataset:

- **What dataset did you choose?**
  The dataset I chose was Fashion MNIST from Kaggle. It's structured very similarly to MNIST, in that it has 60k images that are 28x28 pixels in grayscale and there are 10 different class labels. The labels are: 0: T-shirt/top, 1: Trouser, 2: Pullover, 3: Dress, 4: Coat, 5: Sandal, 6: Shirt, 7: Sneaker, 8: Bag, 9: Ankle boot. For the purposes of this assignment, I made the dataset smaller, choosing to only sample 5,000 images. Here are some examples:

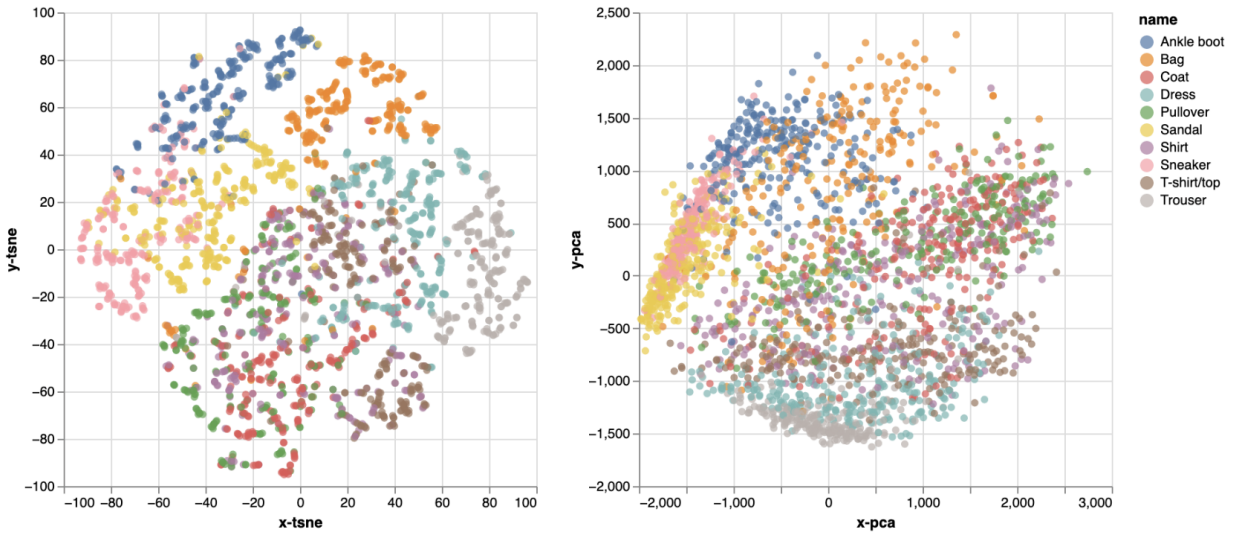


Purse | Ankle Boot | Dress

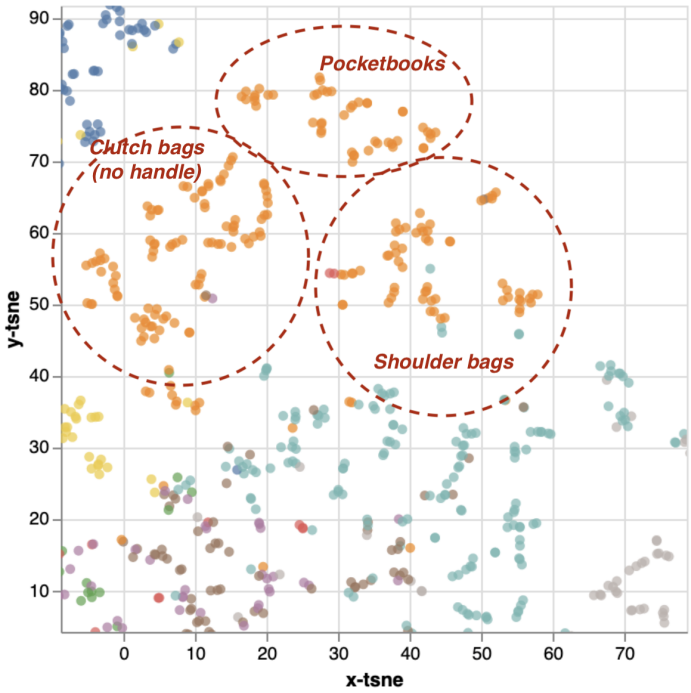
- **What is the class distribution of this dataset?**
  Classes are approximately evenly balanced.



- **Do any features have interesting distributions or interesting relationships with other features?**
  Some of the classes that overlap seem like they could be in the same class; for example, I'm not completely sure of the definition of "pullover", and it appears to have a lot of overlap with T-shirt/top and Coat. Is it a sweater? If so, it makes sense that the model has trouble differentiating, because the silhouettes of sweaters are very similar to long-sleeve shirts or jackets. Here are two embeddings of features: t-SNE and PCA, both showing overlaps in the long-sleeve categories (top, pullover, coat):

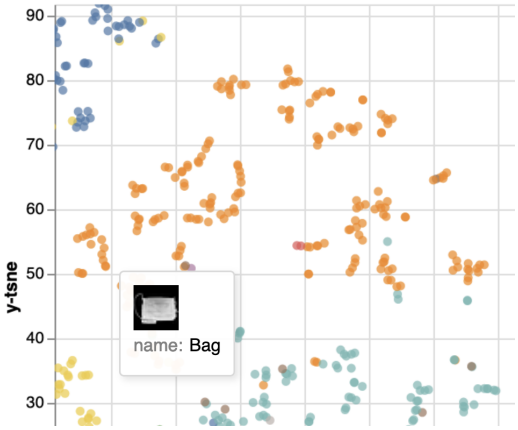


- **Did you do any feature engineering or feature selection? If so, what informed those decisions?**
  If I were to develop a model like this, and making that distinction was important to me, I might want to collect additional data, e.g. about the weight or fabric blend, to help separate those classes. Unfortunately this data is not included. Conversely, I might actually suggest a second or even third class of bag (clutch versus shoulder bag or purse) because it appears that the model has made a distinction in its understanding of this class.



Zoomed in t-SNE projection showing highlighted clusters of bags (in an earlier t-SNE not screenshotted, these groups were actually in completely opposite locations from one another in embedding space)

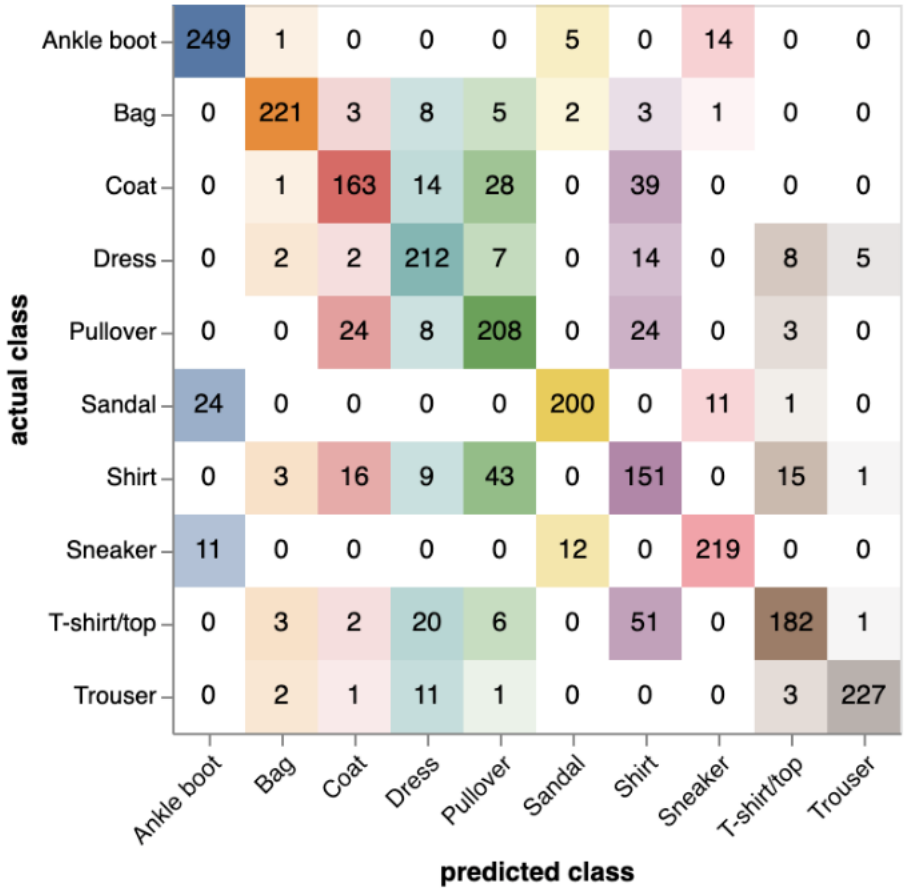
Hovering over a shoulder bag



Hovering over a clutch bag

Error analysis:

- **What is the overall performance of the model?**
  Using only 5,000 total instances (2,500 for testing, the other half for training), the model still gets 81.2% accuracy.

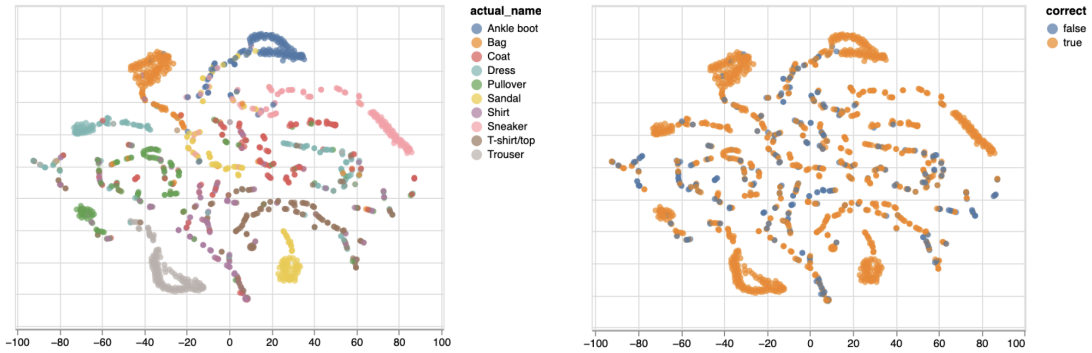


- **What kind of errors does the model make? What are the types of instances the model is confused about? What is the potential reason for the confusion?**
  There are certain classes that are most confusing to the model: in particular, there is major confusion across coats, pullovers, and shirts. Ironically, these are the classes that anecdotally are most difficult to differentiate in humans, e.g. children. It makes sense that a black and white dataset would especially confuse these classes; color, texture, and material are the biggest factors in differentiating between these. Notably, there is also some confusion between sandals, sneakers, and ankle boots.

- **Does the model make errors near or far from the decision boundary?**
  The model makes errors close to the decision boundary; from the embeddings, we can see that several classes overlap almost entirely, so it is clear that the model is struggling to learn important features for differentiating coats, pullovers, and shirts.

- **How do the errors distribute? Are there specific subsets (areas of the decision space) where the model makes a higher number of errors?**
  Long-sleeved tops, generally, are most mis-classified together, followed by shoes. Interestingly, there are two distinct kinds of bag in the model's eyes, though they are all

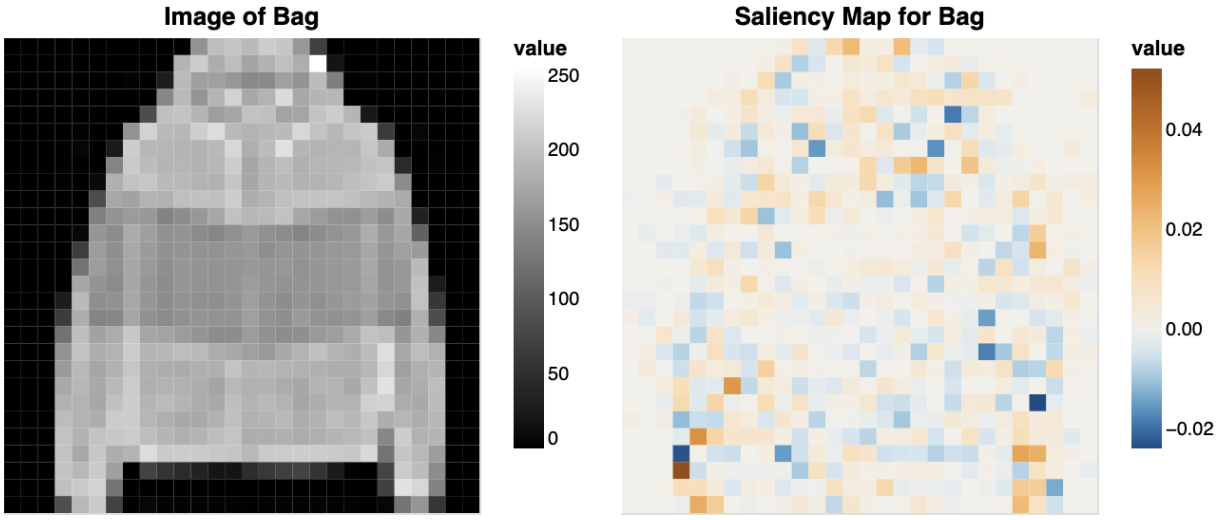
classed the same in the target variable. However, there are errors in a good number of classes, as shown by comparing these two projections of layer 4 of the MLP side-by-side:



Model logic:

- **What drives the model decisions? What are the most relevant features? What impact do those features have on the model's behavior?**
  The model is a 4-layer multi-perceptron with four layers (input, two hidden layers using the ReLU activation function, and output, which uses softmax). Based on the saliency plots, it appears that the model pays attention most to the edges of the clothing.

- **Does the logic used by the model make sense intuitively?**
  It makes sense that the model pays attention most to the edges of the clothing, since they are mostly silhouettes. Below, we can see a saliency map of a misclassified pullover; the saliency map does somewhat resemble a backpack, so it may be that the curvature of the hood caused the neural net to think this was a rounded bag.



- **Does the logic of the model differ in different data subsets?**
  It's interesting that the model divides purses into two distinct areas, despite both being the same class; there is one group of purses that have long straps/handles, and another that are more like clutch purses with a small or absent handle. The model suggests that performance could perhaps be improved by adding a new class that differentiates these two subtypes of purse. Apparently, it is using edges as a way of deciding between long-sleeve tops (coats, jackets, pullovers), but looking at lines (e.g. straps) on bags and shoes. It seems to look for straps both in its placements of different purses (despite them all being one class), and in its decisions about which shoes are sandals versus not (ankle boots, sneakers).