**6.2 Project 2: Project Check-In/Milestone 2**

1. *Any surprises from your domain from the data?*
   1. Most of the surprises I found in this domain will be explained below in the next section about the dataset itself (which is related to the domain of the problem).
2. *The dataset is what you thought it was?*
   1. While the dataset contains information needed to create a model to detect credit card fraud, it puzzled me at first. I understand that any credit card transaction information will be anonymized to remove any indication who it belongs to. What surprised me is that this dataset has almost 30 columns of anonymized data! This makes it a little more challenging in determining which parameters are more important for the model. I expected to see some more benign parameters like the retails name, whether a pin was inputted or not, and other kinds of information that is not indicative of who the user is.
3. *Have you had to adjust your approach or research questions?*
   1. As of now no adjustment have had to be done yet. In anticipation for any problems, or the ability to expand the project, I do have an alternative plan. Since the model be using a supervised model of some kind, I would like to try my hand at creating one using an unsupervised method such as a neural network. While I was already planning on dedicating my next project to neural networks, I feel like this could be a good application. If I were to use both methods I could compare and contrast the benefits and negatives of each in a production setting.
4. *Is your method working?*
   1. So far, my methods are working pretty well! Had a little bit of a rough start (as described in the next section) but everything else seems to be running smoothly now. After I completed EDA I moved onto my first attempt at modeling with little to no issues. My first model resulted in a ROC value around 0.78! This is a very promising start. As I am researching Spark more and more I want to try and implement some validation and model hyperparameter tuning functions to try and achieve a better model. At first glance they seem very complicated and I’m unsure as to how much it will improve the model. With that being said, I am up for a challenge to try and further my knowledge!
5. *What challenges are you having?*
   1. As with my last project where I wanted to use Spark, my main challenges are running Spark on my computer. When I was preforming EDA on the dataset I had no trouble loading the file, but when I tried to create histograms in Spark I encountered lots of error about the worker and driver having mismatched versions. For purposes of EDA, I ended up using Pandas the load and explore the data. While I wanted to use Spark for this process, I decided it wasn’t worth the time to debug and mess around to make it work. Spark will still be used when creating the model.