Data Science Challenge:

Problem 1 - Classifying sensor data

For this problem, we want you to figure out how containers move through the port of Oakland. Specifically, we want to track 3 events that happen within the port:

- 1) When the container is unloaded from a ship
- 2) When the container is loaded onto a train
- 3) When the train leaves the port

For this problem, you can assume that every container will have all of these events and that they happen in order.

Let's say we've attached a couple of sensors to each container: SENSOR_A and SENSOR_B. Whenever there is activity for a specific container, we receive a sensor reading from both of its attached sensors. However, there is a lot of noise associated with these sensors. Our goal is to classify these sensor readings as one of our events or as noise:

- LOAD_TRAIN
- DEPART_TRAIN
- UNLOAD_SHIP
- NOISE

The sensor data for all of our containers is contained in two csv files: train.csv and test.csv. Each row represents a single sensor reading for a specific container.

Each sensor reading (row) is made up of the following fields:
container_id (int)
TIME: (datetime) the time that the sensor reading occurred
SENSOR_A: (int) the sensor A measurement
SENSOR_B: (int) the sensor B measurement
LABEL: (str) the label we are trying to classify

Your goal is to implement a model that predicts the value of field 'LABEL' for each sample. Please include your code, any analysis you performed, and share your thoughts on future work.

Problem 2 - Reconciling conflicting data

In this problem, we'll explore the challenges of storing and reconciling conflicting data sources. Background: Imagine a container being shipped from a factory in Asia to a distribution center in the United States. Along that container's journey, there are numerous parties involved. Typically, a container will start off in the retailer's hands, then it's passed to the trucker, then to the port, then to ocean carrier, etc.

During the journey, each party is recording event data about that container. For example, the trucker might record the events "loaded on the truck", "underwent customs check", "delivered to the port". The port might record: "received from trucker", "loaded onto vessel". The vessel might record: "loaded onto vessel", "loaded off of vessel."

In theory, at each hand off point, the events reported by the various parties should match up. For example, the trucker's "delivered to the port" event should occur at the same time as the port's "received from trucker" event. But unfortunately that is wishful thinking in this industry. Often times these events won't match up for various reasons. For instance, maybe the trucker doesn't want to admit they delivered the container late so they shift their "delivered to the port event" up by a few hours, or maybe the port is not the most technically savvy, so they tend to report their "received from trucker" event a few hours late.

These conflicts present a challenge for ClearMetal when we ingest the data for a container’s journey. Whenever we receive conflicting data for an event, we need to figure out what actually happened. This can get pretty darn complicated, because sometimes you can have 3 or more parties all reporting conflicting pieces of data for the same container event.

Problem:

We’d like you to write up a description of how you would you approach this problem, focusing on:

1. What would your first steps be for approaching this problem?
2. What statistical models might be relevant for this problem?