Features

1. address: String. Bitcoin address.
2. year: Integer. Year.
3. day: Integer. Day of the year. 1 is the first day, 365 is the last day.
4. length: Integer. Quantifies mixing rounds on Bitcoin, where transactions receive and distribute similar amounts of coins in multiple rounds with newly created addresses to hide the coin origin
5. weight: Float. Quantifies the merge behavior (i.e., the transaction has more input addresses than output addresses), where coins in multiple addresses are each passed through a succession of merging transactions and accumulated in a final address.
6. count: Integer.
7. looped: Integer. Intended to count how many transactions
    a. split their coins
    b. move these coins in the network by using different paths, and finally
    c. merge them in a single address.
8. neighbors: Integer.
9. income: Integer. Satoshi amount (1 bitcoin = 100 million satoshis)
10. label: Category String. Name of the ransomware family (e.g., CryptXXX, CryptoLocker, etc) or white (i.e., not known to be ransomware).

Prompt:

1. Look through the dataset and preprocess it in a way that makes sense to you.
2. Extract trends and patterns from the data using:
    a. Data visualization: Create an infographic with no less than three charts (that can help you better understand the data).
    b. Hypothesis/experimental Testing: Explore the data and come up with a hypothesis.
3. Determine the top three ransom labels that have the most ransom transactions.
4. Define a machine learning model most appropriate for classifying heist incidents into ransomware families. You will be graded on this step.
5. Then, define a model to predict:
    a. If a future transaction is ransom or not, and if it is,
    b. The ransomware family it belongs to.

    Fit it on the train set and predict values on the test set.
6. Prepare a report containing your results from the analysis. It should contain the following: Intro, data cleaning/pre-processing, visualizations (at least 3), analysis, proposal, conclusion.