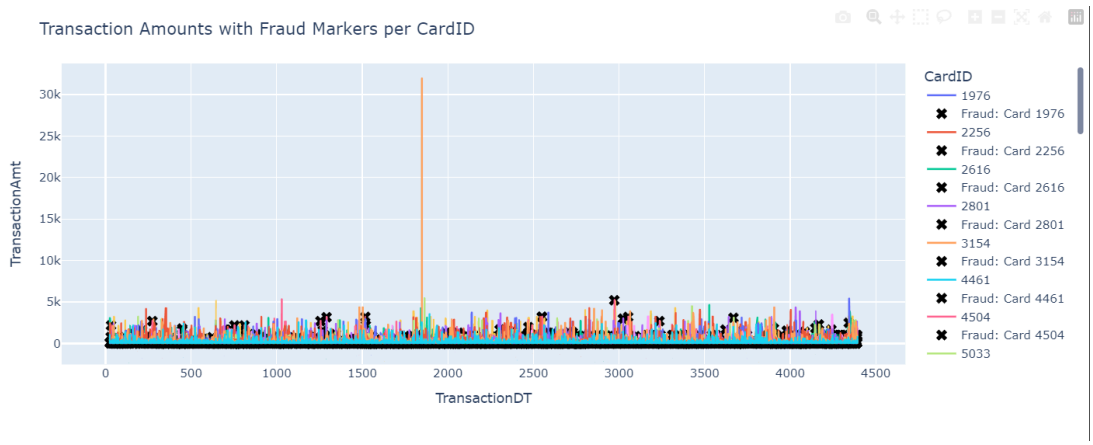
Working with Time Series Patterns (TransactionDT and CardID)

1. Extract only Relevant Columns (CardID, TransactionDT, TransactionAmt and Engineering Encoded Columns (14), isFraud)

2. Define Fraud and non Fraud CardID
    - Fraud: 1735 (total flagged)
    - Non fraud: 11808 (total)
    - In which:
        + The more transaction per card, the higher likelihood of Fraud that card becomes -> Card with exceptionally high number of transactions is more likely to be fraud

3. Retrieve Fraud-flagged cardID from fraud data and map back the transaction logs of them



=> Approaches:
- Similarity Analysis between Card transaction log to define whether Fraud cards share the same patterns
- Feature Based Analysis: Extract meaningful feature from the Date and Amount columns: mean, std, skewness, kurtosis,..... to map relevant information (catch22)

I. Similarity Analysis (Using Clustering)

1. Vary Labels
    - Get Flagged CardId and get number of Fraud transactions - > Fraud Ratio
    - Looking at the Fraud Ratio histogram, choose 0.05 as the threshold for Fraud Level (1 for low fraud, 2 for high fraud)
    - Create a test dataset to store CardID and their relative Fraud value (0 - 1 - 2) (1491 cardID: 971 (0), 520(1))
    - Test Data Final distribution: (Ensure balance between Fraud and non-fraud cards)
        + 0: 971
        + 1: ~ 300
        + 2: ~ 200

2. Define Approaches
    - By TransactionAmt
    - By Engineering Encoded features
3. Data Preparation
    - Keep the TransactionAmt column, try Weighted-sum and Autoencoder to scale down 14 E-E features to 1 (to maximize the information gain instead of using classic dimensionality reduction)

    - Choose the prune threshold high enough for equally observing fraud/ non-fraud data and low enough to ensure time series: Cut at 35 transactions per card (fraud/ non fraud ratio = 1.03) and Map all card to a fixed size of 50 transactions each to reduce computational cost

4. Result
    - The TransactionAmt shows a balanced, mostly flat DTW centroid for most of the cards, the 95% CI in the highest - distributed clusters also show a steady and constant value over time without any highlighted peaks. As the centroid gets more chaotic and spiky, the number of actual related cards decreases. This shows that the actual rate of Fraudity doesn't really depend on the transaction time series.
    - In the next part, the time series will be aggregated to find meaningful patterns
    - The same test is also conducted with the Engineering Encoded features, varying in 2 ways of preprocessed (weighted sum by Explain variance and Autoencoder)
        + The autoencoder method shows a highly superior result both from the Scatterplot and the clusters, which returns more meaningful information
        + The Weighted-sum performs poorly for having too many noises in the distribution and fails to differentiate the data points.

II. Feature Based Analysis

1. Idea
    - Extract Tabular data of Time series: getting the relative information from each CardID and its respective Time/ Value
    - Try to fit Classification ML models
    - Result:
        + If fit well (accuracy high): Create a predictor which ingests a time series and automatically extract

features. Predict the Probability of that Input series to be flagged as Fraud
    + If not fit well (accuracy low): Conclude that the fraudity is not closely related to the time/value relationship

2. Prepare
    - Get the dataset of flagged CardID, map them back to the original data to get the Fraud/ non-Fraud data
    - Apply the Extract function and returns the Fraud/ non-fraud data, concat them vertically after assigning the right label
    - Final data (7267 cards: 6334 (0), 933 (1))

3. Result
    - Applying a bunch of Tuned Classification models returned a rather unexpectable result where the maximum accuracy is 0.86 which is really high.
    - Choosing the Gradient boosting, create an application to input custom time series for 1 card and return the probability of its being flagged by the system. Also the model can be used for calculating the variable importance.