Report

Approach : Basically , we have taken the dataset and converted it into a pandas dataframe and then first we removed rows having 0 as prices(As they are anomalies).Then we split the Seat Fare Type 1 (4 Categories) as Category11 , Category12,Category13,Category14 and Seat Fare Type 2 as Category21,Category22,Category23,Category24. The idea behind the split is to check pattern in individual categories and summing then up to produce final result. Later, we plotted we made the dataframe compact by keeping only unique bus ids and their corresponding price change and time as list . Then we remove nan value from the price list of each buses and , later sorted the prices with respect to time.

Further we plotted time series to analyse data. We found that Category21, Category22, Category23 have very less data.

So then we compared the time series of Categories11 and by using Dynamic Time Wrapping we calculated distance . Basically we used dtw comparation and compared Category11 of a bus with every other buses and similarly for Category12 and so on.If we unfortunately get a infinity as dtw distance we changed it with maximum of the dtw distance+1000 in the list so as to get rid of infinity.

Then we found minimum of dtw distance and stored which bus have this value for which , will resemble most closely and normalised that dtw distance using a normalisation formula of (M-S(x,y))/M where M is the maximum dtw distance in the list. Later we passed distance to sigmoid function to produce confidence score .Finally we stored the data in a dataframe and saved it in a csv file.