Documentation supporting Problem Statement

Problem Statement

A drug is generally administered to a patient in certain patterns or in regular intervals of time. For example Chemotherapy which is drug treatment in case of Cancer is generally given to patients in an interval 3-4 weeks, i.e. every 3-4 weeks patients are administered with the drug. Similarly to Chemotherapy, “Target Drug” is also administered/prescribed in certain patterns, we want to analyse in what patterns “Target Drug“ is administered/prescribed to patients, there might be multiple patterns in which “Target Drug” is administered/prescribed, come up with an analysis which to extract the dominant patterns in the data using clustering or other unsupervised techniques. Visualise the prescription patterns with time on X-axis (month) and prescriptions on Y-axis for each of the patterns you are able to extract(Below is an example of a prescription pattern, where a prescription is made at least once in the first two months followed by one prescription for every two months).

Steps I followed to come up with a solution

1. First I read and addressed basic inconsistencies here as well and then as we have already understood the data and made inferences about data, I left that part and jumped directly into problem statement.
2. As just the positive set is of importance here, I pulled out just that out of the dataset we had and sorted them based on Uid and Date for convenience.
3. I engineered new feature called Time interval which is interval between subsequent target drug administration for all patients.
4. Used the time interval feature and found optimal number of clusters using elbow method.
5. Did clustering using k-means clustering algorithm and assigned the cluster labels to respective instances and have seen the % distribution of clusters.
6. Segregated each of the clusters and formed 4 new dataframes for ease of usage.
7. Using Grouper object and groupby method formed 4 new dataframes to create some overall visualization – Time Period against number of Prescriptions made and noted down my inferences in the notebook.
8. In order to go with problem statements demand, framed a new feature representing month using the already existing time interval column and using groupby and few other methods found average prescriptions every month.
9. Using Plotly graph objects, plotted month against average prescriptions for each clusters separately and made inferences and noted down separately and summarised things at the end.
10. Overall, there were 2 clusters performing well and 2 quite the opposite.