Problem 3

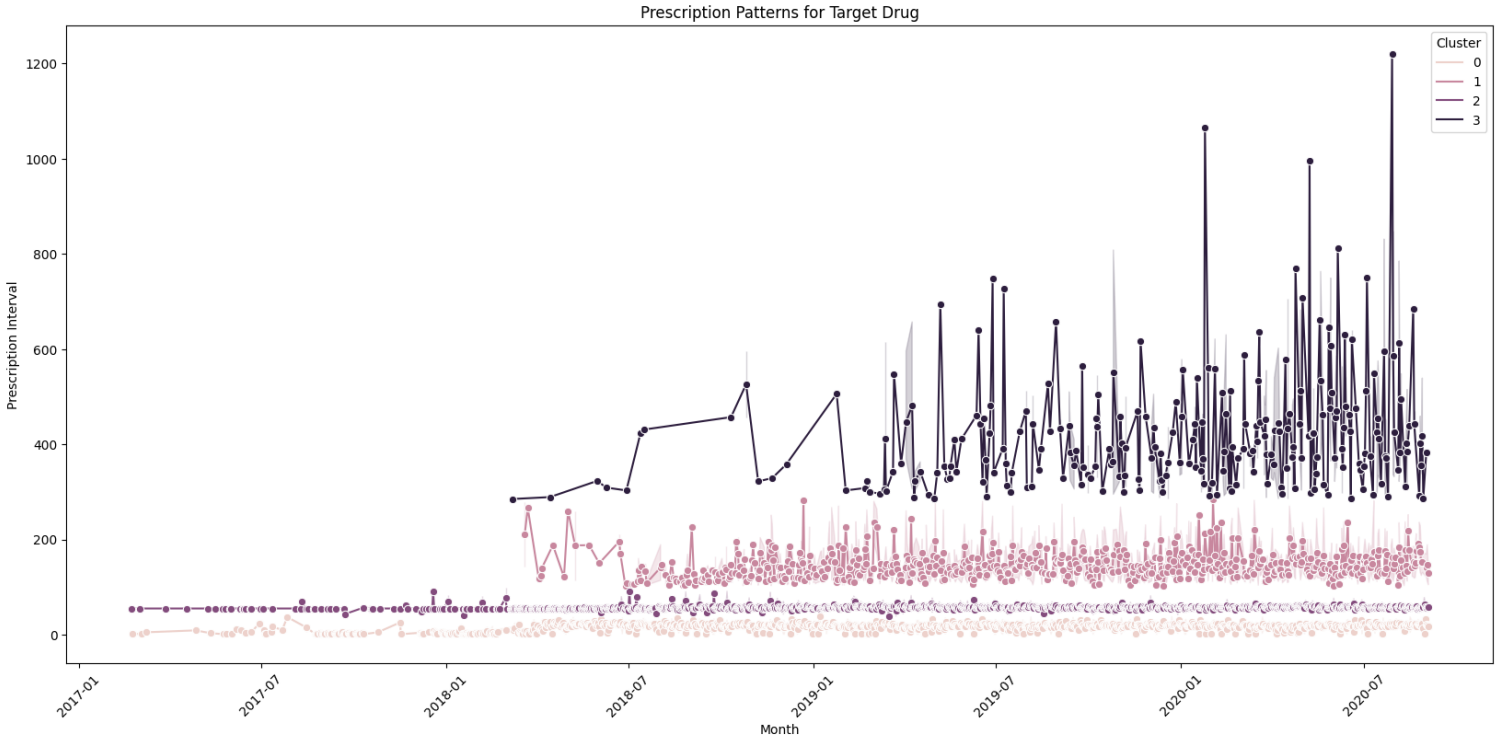
A drug is generally administered to a patient in certain patterns or in regular intervals of time. For example Chemotherapy which is drug treatment in case of Cancer is generally given to patients in an interval 3-4 weeks, i.e. every 3-4 weeks patients are administered with the drug.

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).

Code Explanation:

1. The 'Date' column in the target drug data is converted to the datetime format using the pd.to_datetime() function.
2. Prescription intervals are calculated for each patient by taking the difference in days between consecutive dates. Missing values in the intervals are filled with the median interval using the fillna() method.
3. The 'PrescriptionInterval' column is extracted from the data and stored in the variable 'X'. This column represents the prescription intervals and will be used for clustering.
4. The code then determines the optimal number of clusters using the elbow method. It initializes an empty list called 'wcss' (within-cluster sum of squares) and iterates over a range of cluster numbers from 1 to 10. For each iteration, K-means clustering is performed on the 'X' data, and the inertia (WCSS) is calculated and appended to 'wcss'.
5. The elbow curve is plotted to visualize the change in WCSS for different cluster numbers. This helps in choosing the appropriate number of clusters based on the point where the curve levels off.
6. The chosen number of clusters(here 4) is assigned to the variable 'num_clusters'.
7. K-means clustering is performed again, this time with the chosen number of clusters.
8. The data is grouped by cluster, and the average prescription interval and prescription count are calculated for each cluster.
9. The prescription patterns are visualized using a line plot. The 'Date' is plotted on the x-axis, 'PrescriptionInterval' on the y-axis, and each cluster is represented by a different color.
10. The plot is labeled appropriately with axes labels and a title. A legend is added to indicate the clusters.

Result


Prescription Patterns for Target Drug
Cluster
0
1
2
3
Prescription Interval
0
200
400
600
800
1000
1200
2017-01
2017-07
2018-01
2018-07
2019-01
2019-07
2020-01
2020-07
Month