Structured Data Assignment

Data Description:

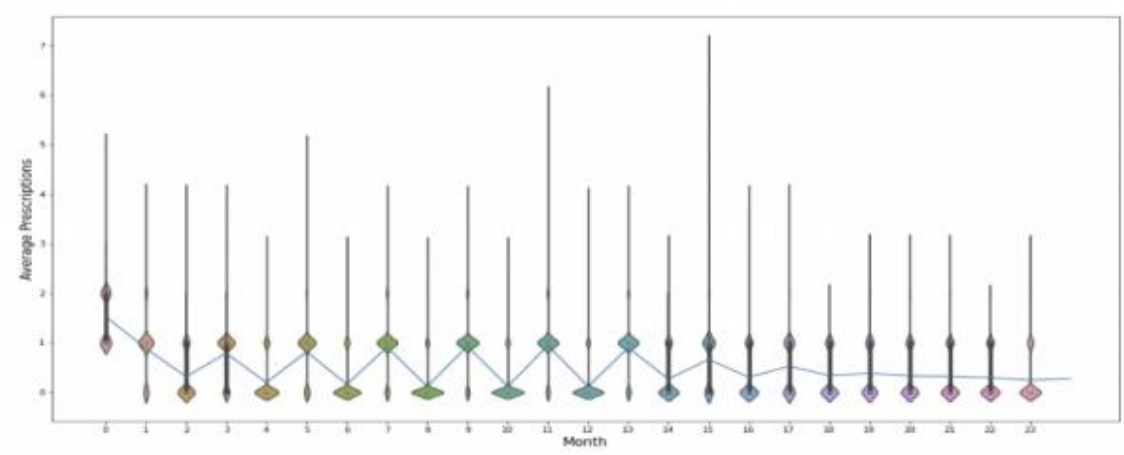
- **Train.parquet** - Dataset to be used for training
- **Test.parquet** - Dataset to be used for testing

The data consists of three primary columns:

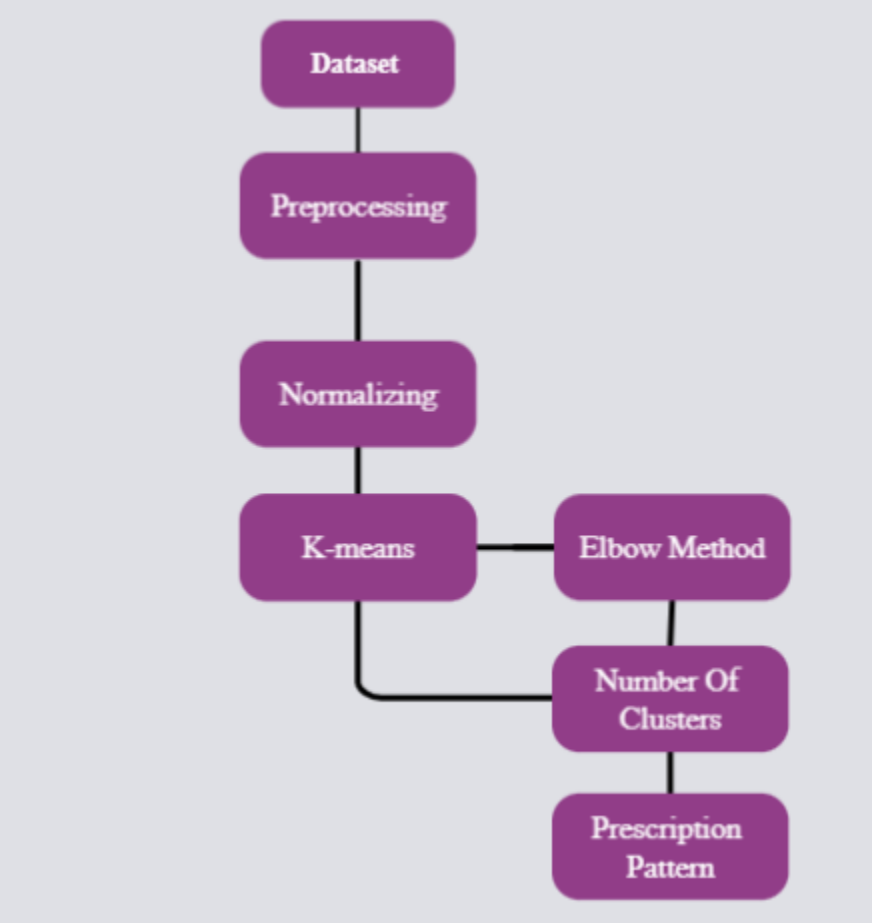
- **Patient-Uid:** A unique alphanumeric identifier assigned to each patient who are taking the medicine.
- **Date:** The specific date when the patient experienced the event.
- **Incident:** This column provides a description of the event that took place on the given date for the patient.

Problem Statement 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 analyze 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. Visualize 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).



Process Flow:



Implementation:

Initial analysis of data:

- Loading the dataset

	Patient-Uid	Date	Incident
0	a0db1e73-1c7c-11ec-ae39-16262ee38c7f	2019-03-09	PRIMARY_DIAGNOSIS
1	a0dc93f2-1c7c-11ec-9cd2-16262ee38c7f	2015-05-16	PRIMARY_DIAGNOSIS
3	a0dc94c6-1c7c-11ec-a3a0-16262ee38c7f	2018-01-30	SYMPTOM_TYPE_0
4	a0dc950b-1c7c-11ec-b6ec-16262ee38c7f	2015-04-22	DRUG_TYPE_0
8	a0dc9543-1c7c-11ec-bb63-16262ee38c7f	2016-06-18	DRUG_TYPE_1

- Calculating the total number of distinct patients and the total number of distinct incident

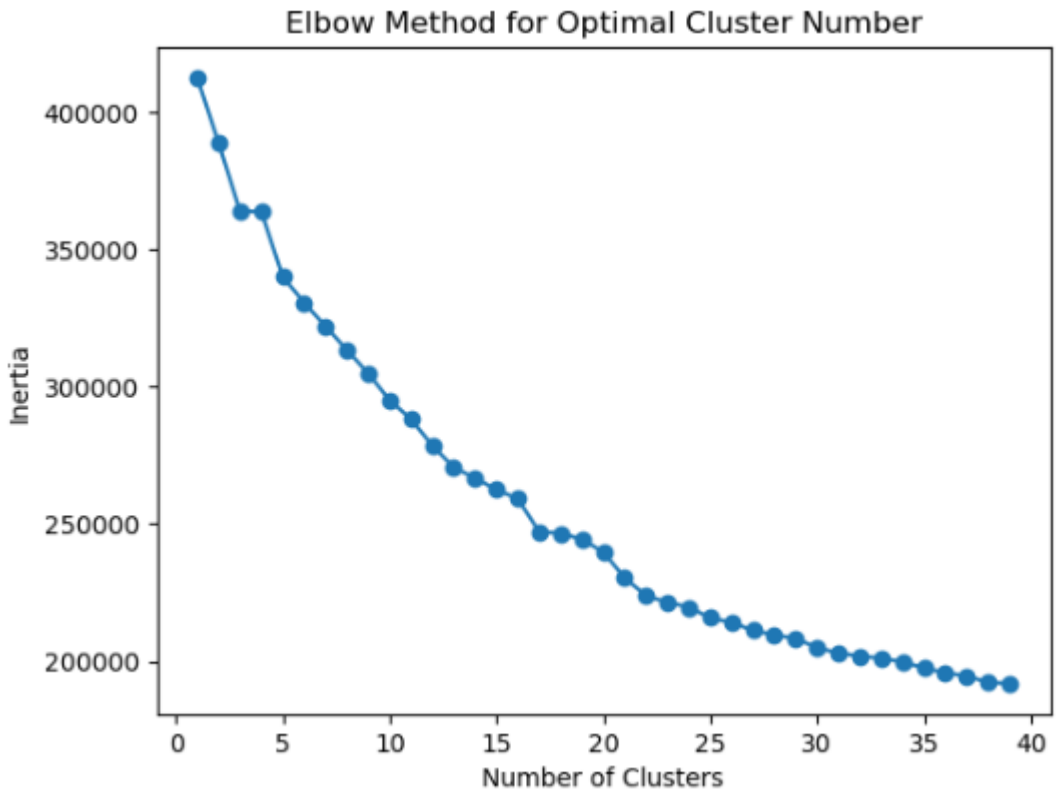
Pivot Table:

Initially, we'll add a column to the data frame to represent months. This month column will be utilized during the clustering process and when constructing a pivot table.

Month	2017-02	2017-03	2017-04	2017-05	2017-06	2017-07	2017-08	2017-09	2017-10	2017-11	...	2
Patient-Uid												
a0e9c384-1c7c-11ec-81a0-16262ee38c7f	0	0	0	0	0	0	0	0	0	0	...	
a0e9c3b3-1c7c-11ec-ae8e-16262ee38c7f	0	0	0	0	0	0	0	0	0	0	...	
a0e9c3e3-1c7c-11ec-a8b9-16262ee38c7f	0	0	0	0	0	0	0	0	0	0	...	
a0e9c414-1c7c-11ec-889a-16262ee38c7f	0	0	0	0	0	0	0	0	0	0	...	
a0e9c443-1c7c-11ec-9eb0-16262ee38c7f	0	0	0	0	0	0	0	0	0	0	...	

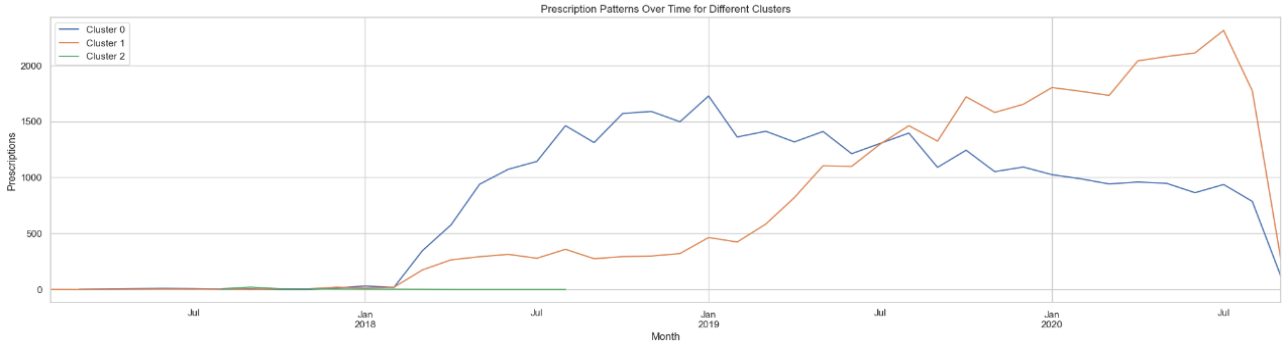
Clustering:

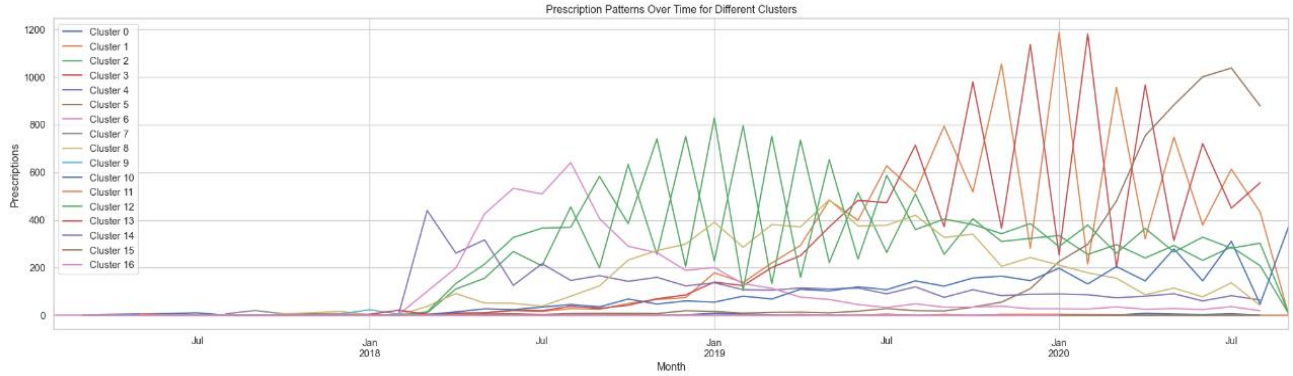
- Normalize the data.
- Applying Kmeans-to determine optimal number of cluster
- Elbow method to plot the optimal number of cluster.



Prescription pattern with respect to time:

The presence of distinct patterns among various patient groups, likely attributable to different medical conditions or factors, indicates that the provided dataset contains valuable information.





Results:

The existence of unique patterns within different patient groups, which can be attributed to various medical conditions or factors, suggests that the dataset offers valuable insights.