Six Steps for Data Clustering

Executive Summary

Data in the business world is growing exponentially. According to IDC report "The Digitization of the World: From Edge to Core", 79 zeta bytes of data was generated in 2021. Most data are generated by operational activities, such as customer transactions, supply chain management, and product development. This is also reflected in the prominence of services to manage and analyze data. This is not at all surprising. Business professionals experience this every day. There is a steady accumulation of data pertaining to shipments, invoices, customers, inventory, people etc.

The challenge is specifically gaining insight from this large amount of data. There are tools such as Tableau or PowerBI that are suitable for data visualization and creation of appealing charts. However, these tools are helpful when one knows what to look for in the data.

A particularly appealing technique that could speed up the understanding of a large dataset is clustering. This technique refers to the segmentation of a dataset into groups based on similarity. The similarity of datapoints is assessed across all the available attributes. The ability to utilize all available attributes by the clustering algorithms is particularly interesting and beneficial.

The resulting segments of input data are also referred to as clusters. Each cluster can be considered as a particular type of input data represented in the dataset and can therefore enable deeper understanding. For instance, clustering can be used to identify:

1. Key customer types from a dataset of customer transactions [e.g. 5 customer groups from 500 customers]
2. Main cylinder profiles from part master data for cylinder commodity [e.g. 10 main cylinder profiles from 600 SKUs]
3. Major types of shipment lanes from historic shipment data [50 main lanes from 10,000 monthly shipments]

In the above examples, the main benefit is that it is easier to review 5 customer groups [vs. 500 individual customers], 10 cylinder profiles [vs. 600 SKUs] or 50 main shipment lanes [vs. 10,000 shipments] and thereby, gain a deeper understanding of the data.

The purpose of this article is to describe the clustering technique through a case study.

Case Study

Introduction

This case study covers a scenario overlapping between engineering and procurement. The dataset contains synthetic information about hydraulic cylinders. Using the dataset, the case study demonstrates the approach for clustering. The sequence of steps utilized in this case study follow the CRISP-DM methodology.

Step 1: Business Understanding

The dataset utilized in this case study contains synthetic information about hydraulic cylinders. Each line item in the data represents combination of a particular SKU [Stock Keeping Unit] and using location. From a procurement perspective, the objective is to gain a deep understanding of the types of products. This involves exploring the commonalities as well as the differences among the SKUs. Such an

understanding will enable procurement specialists to identify and evaluate the key supplier capabilities for the clusters. In this case study, clustering algorithms will be utilized to segment the SKUs and thereby, understand the similarities among SKUs within each cluster and differences among clusters.

Step 2: Data Understanding

The dataset contains 648 rows and 28 columns [features]. Among the features, 21 are numeric and 5 are alpha-numeric. The numeric features [excluding Attr3_Costamount_norm] contain technical specifications of the SKUs. The dataset has been generalized and therefore, the attributes do not reflect real values.

Each SKU in the dataset belongs to a particular region and business unit. Figure 1 shows the stratification of SKUs by region. Region_1 has the highest number of SKUs (at 48% of total) followed by Region_2 (at 31% of total).

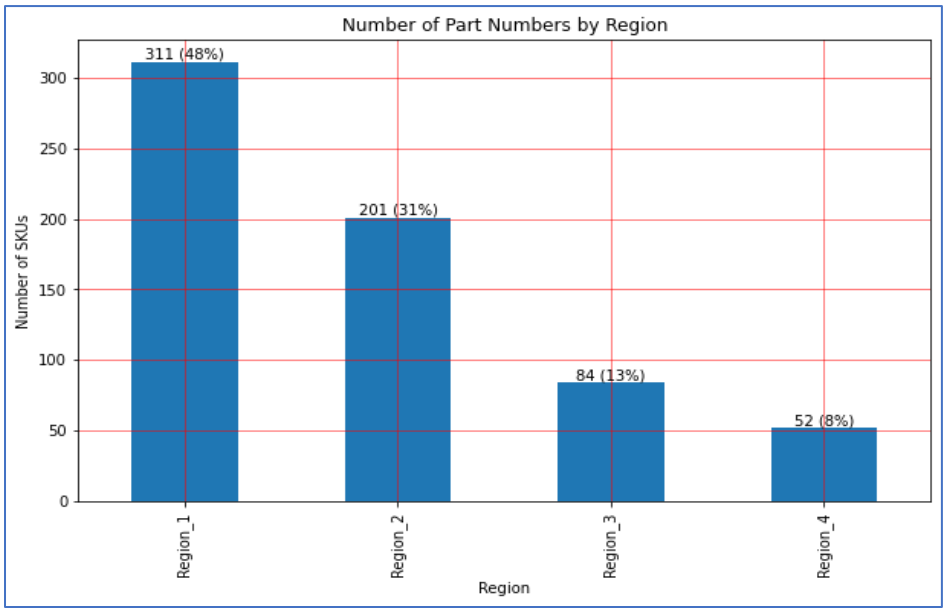

Figure 1: Number of SKUs by Region

Figure 2 shows the stratification of SKUs by business unit. “AG” business unit (p_bu = “AG”) has the highest number of SKUs (44%) followed by “AS” at 30%. Based on the significance of the business units, this case study will focus on “AG” and “CE” business units.

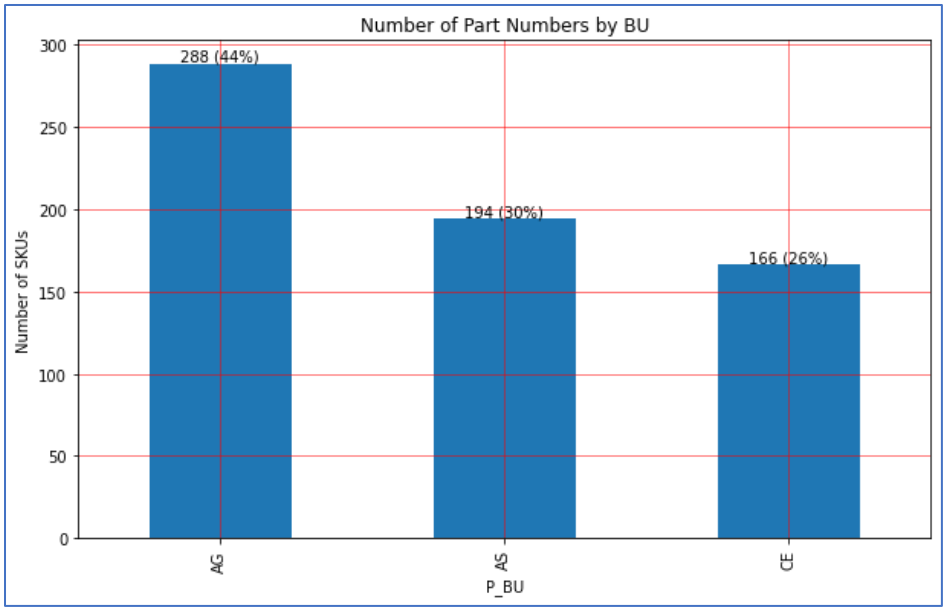

Figure 2: Number of SKUs by Business Unit

Figure 3 shows the average cost by region. Average SKU_cost for Region_1 is higher than the other regions. Region_2 and Region_4 have similar average costs.

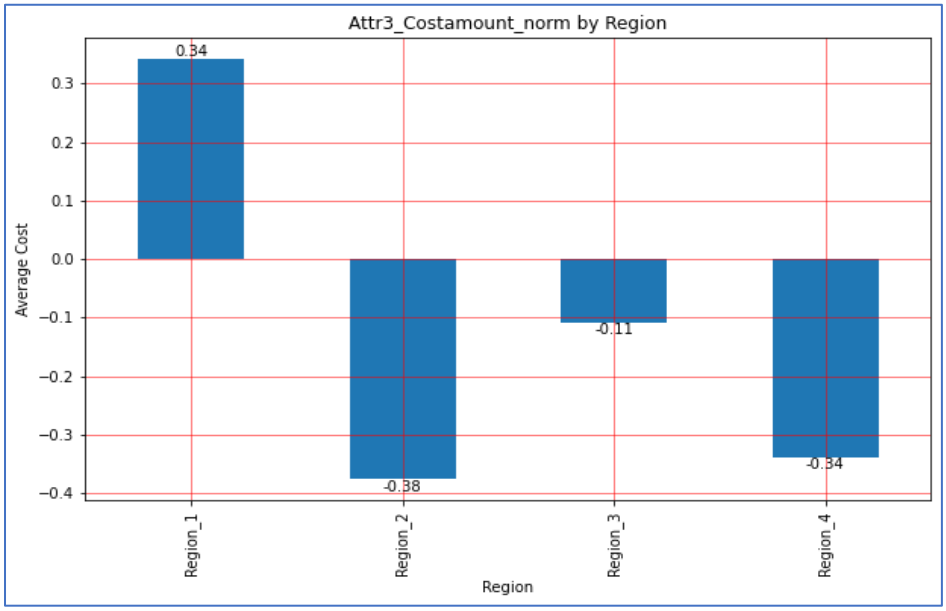

Figure 3: Average SKU Cost by Region

Correlation

One of the first steps is to understand the features whose values have an influence on the cost of the SKU. Covariance among the attributes and specifically versus cost, can be used to identify the features that are correlated with cost.

Figure 4 shows the correlation values among the features. Attr4 and Attr5 show the highest correlation with cost. This is intuitive since Attr4_RdDi_norm and Attr5_BrDi_norm describe the geometry of the SKU. Geometry is one of the main drivers of the raw material needed to manufacture and therefore, impact the cost.

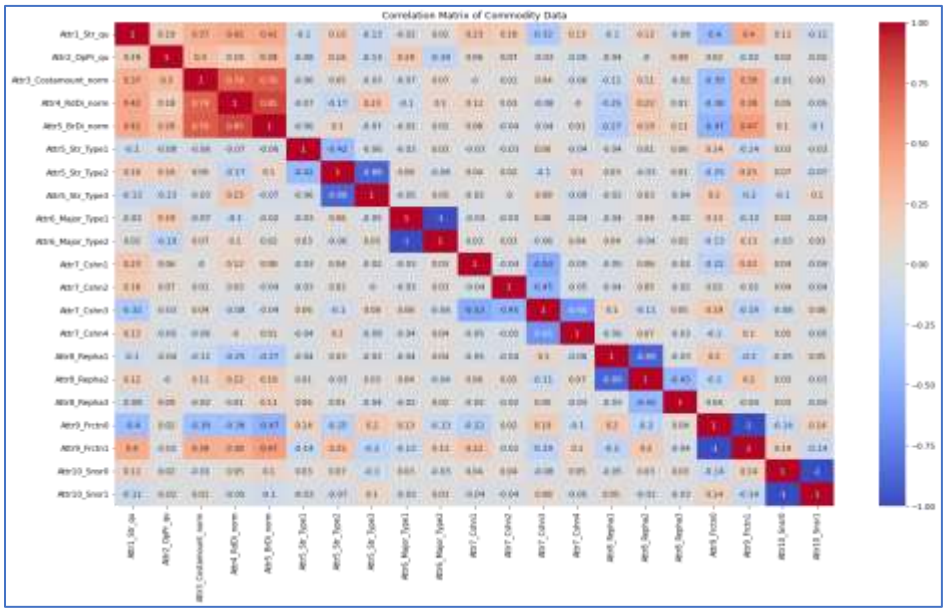

Figure 4: Correlation heatmap for all numeric features

Figure 5 shows the correlation heatmap for a subset of features with correlation value (versus Attr3_Costamount_norm) greater than 0.3.

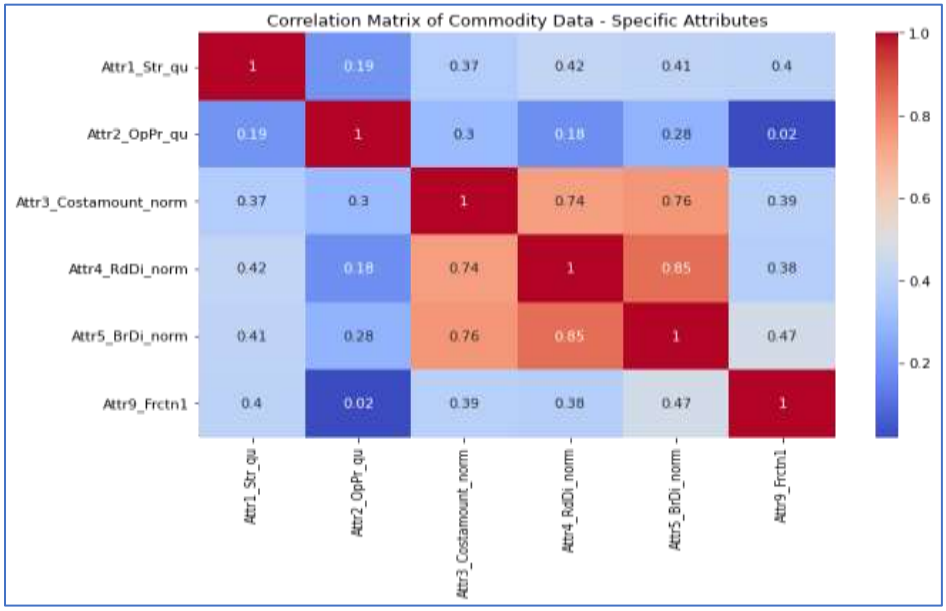

Figure 5: Correlation heatmap for feature subset

Attr1_Str_qu and Attr9_Frct1 also show meaningful correlation with cost. The influence of Attr1_Str_qu on the cost is understandable as it partially describes the geometry. Attr9_Frctn1 refers to a specialized welding requirement and therefore, the impact on cost is as expected.

Correlation by Region

The SKUs in the dataset belong to 4 different regions. The list of features with highest correlation versus cost varies by region:

1. Region 1, Region 3, Region 4: Attr5_BrDi_norm [related to geometry] shows the highest correlation [range: 0.71 – 0.91] with cost
2. Region 2: Attr1_Str_qu [related to geometry] shows the highest correlation [0.56] with cost

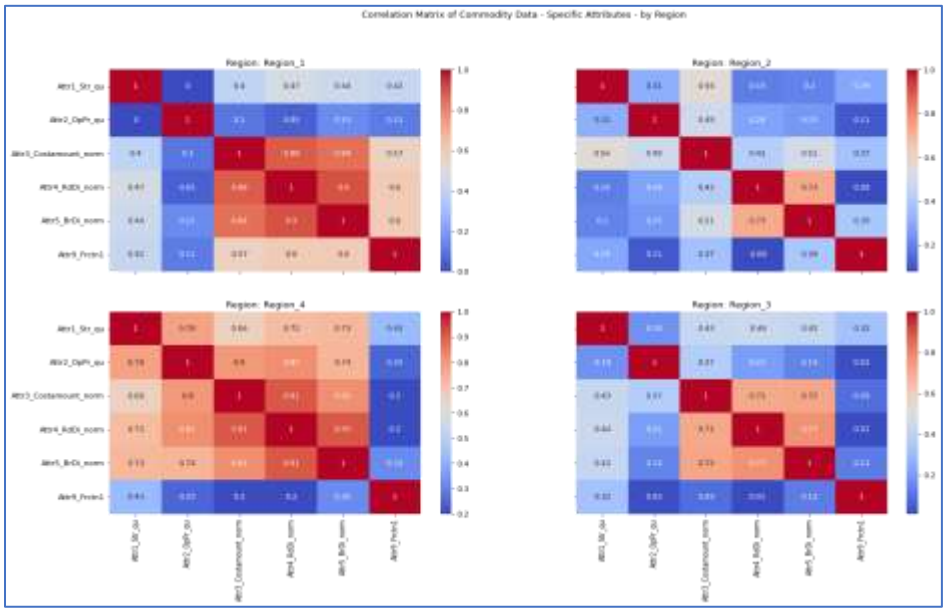

Figure 6: Correlation heatmap by region

Correlation by Business Unit

The SKUs in the dataset belong to 2 main business units [AG & CE]. Figure 7 shows the correlation values for each business unit.

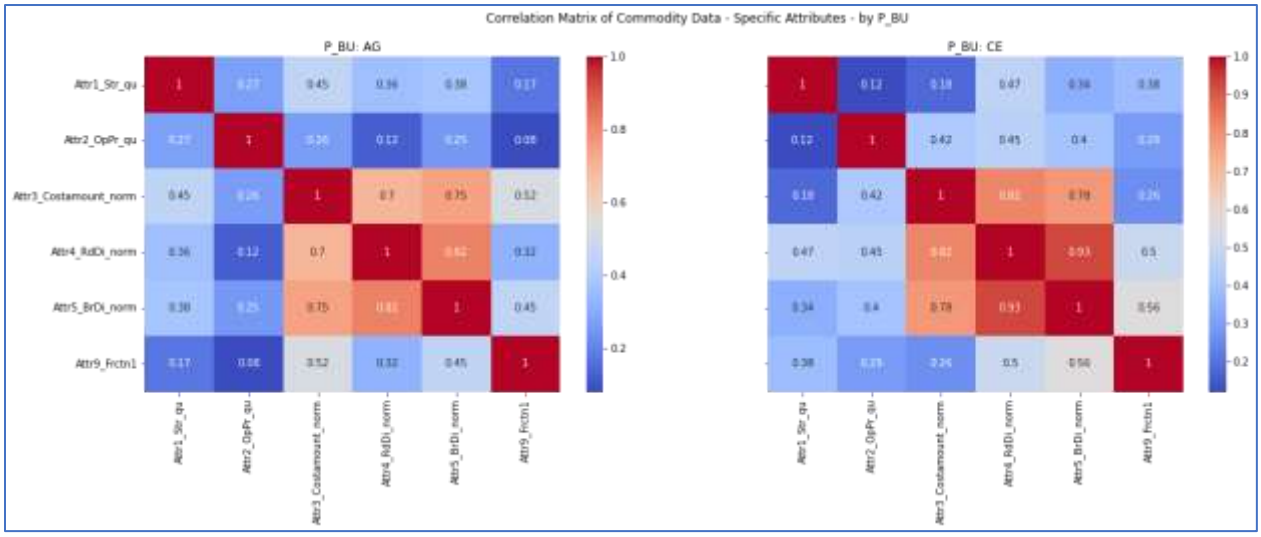

Figure 7: Correlation by Business Unit

For the "AG" business unit, Attr4_RdDi_norm and Attr5_BrDi_norm show the highest correlation with cost. In addition, Attr9_Frctn1 and Attr1_Str_qu also show meaningful correlations of 0.52 and 0.45 respectively.

For the "CE" business unit, Attr4_RdDi_norm and Attr5_BrDi_norm also show the highest correlation with cost. However, the correlation values for Attr4_RdDi_norm and Attr5_BrDi_norm for the CE business unit is greater than the AG business unit. The variance in Attr9_Frct1 within the "CE" business unit is minimal (majority of the SKUs have this special welding requirement) and therefore, lack of correlation with Atr3_Cost_norm is as expected.

Distribution of Specific Attributes

In this section, the distribution of the attributes that exhibited meaningful correlation with Attr3_Cost_norm has been reviewed. The two attributes whose distribution (within each business unit) will be studied are Attr1_Str_qu and Attr2_OpPr_qu

Distribution of Attr1_Str_qu by Business Unit

Figure 8 shows the distribution of Attr1_Str_qu by Business Unit. From figure 8, it can be observed that the variance for Attr1_Str_qu for "AG" is significantly greater than "CE". This is also reflected in the variance coefficient [AG = 0.69, CE = 0.38]. Therefore, it can be concluded that Attr1_Str_qu has a higher influence over Attr3_Cost_norm for the "AG" business unit.

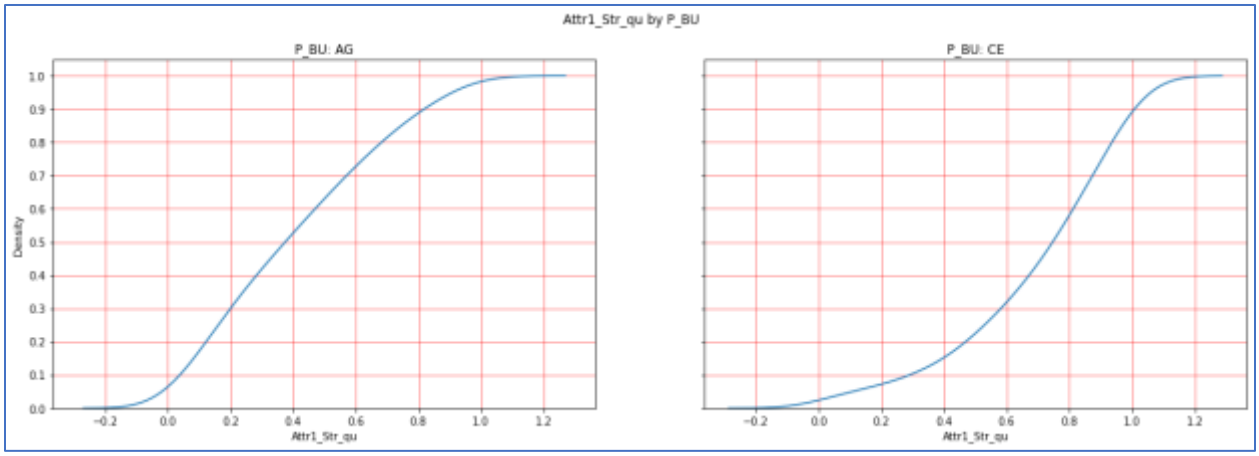

Figure 8: Distribution of Attr1_Str_qu by Business Unit

Distribution of Attr2_OpPr_qu by Business Unit

Figure 9 shows the distribution of Attr2_OpPr_qu by business unit. From figure 9, it can be observed that the variance for Attr2_OpPr_qu is greater for "AG" business unit compared to "CE" business unit. This is also reflected in the variance coefficients. The variance coefficient for "AG" is 41% greater than "CE" (AG = 0.65, CE = 0.46).

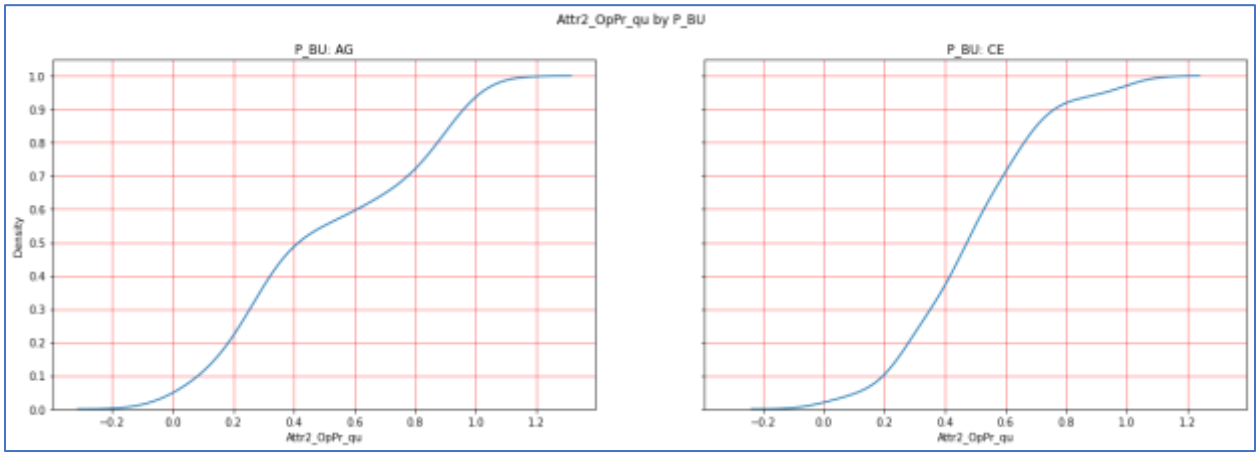

Figure 2: Distribution of Attr2_OpPr_qu by Business Unit

Step 3: Data Preparation

In this step, subset of attributes based on variance are selected. The list below shows the 7 main attributes to be included in the model building step.

```
['Attr1_Str_qu', 'Attr2_OpPr_qu', 'Attr3_Costamount_norm',
       'Attr4_RdDi_norm', 'Attr5_BrDi_norm', 'Attr9_Frctn0', 'Attr9_Frctn1']
```

Step 4: Model Building

The objective of the model is to create clusters based on the input data. As shown in figure 10, there are

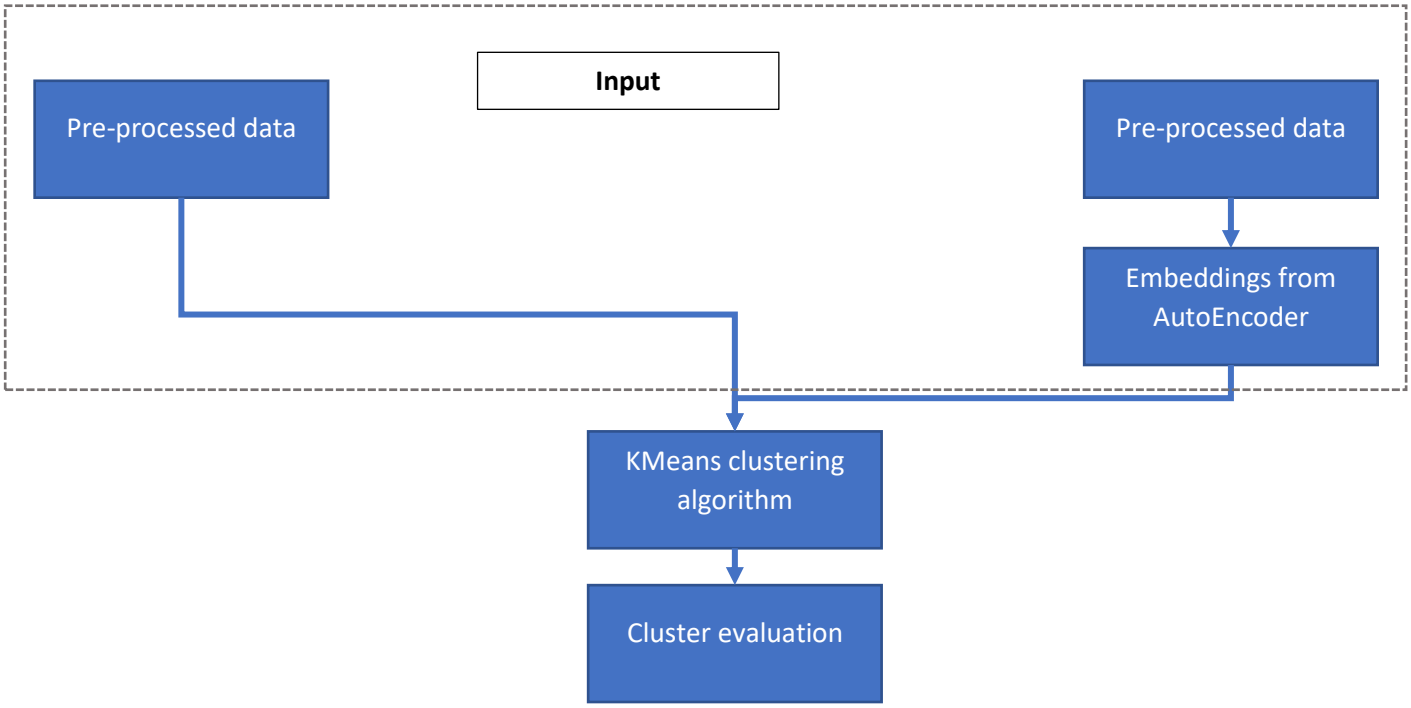

Figure 10: Clustering Model

2 different approaches have been utilized for generating the input data used for creating the clusters. The quality of the output clusters from the different scenarios have been evaluated to the select a particular approach.

Step 5: Model Evaluation

The most suitable configuration of the clustering model for this case study has been selected based on the results from 12 experiments. There are 2 main experiment variables. The first one is the source of the input data. As shown in figure 11, the input data can be the plain pre-processed input data or the embeddings generated from the AutoEncoder. There are 3 different configurations considered for the AutoEncoder. The second experiment variable is the options for the cluster count. There are 3 different values considered for the cluster count. The list of experiments considered in this case study are shown in figure 12.

Experiment number	Variable 1: Input	Variable 2: Cluster count
1	Pre-processed data	4
2	Pre-processed data	6
3	Pre-processed data	8
4	AutoEncoder (Embedding size = 4)	4
5	AutoEncoder (Embedding size = 4)	6
6	AutoEncoder (Embedding size = 4)	8
7	AutoEncoder (Embedding size = 6)	4
8	AutoEncoder (Embedding size = 6)	6
9	AutoEncoder (Embedding size = 6)	8
10	AutoEncoder (Embedding size = 8)	4
11	AutoEncoder (Embedding size = 8)	6
12	AutoEncoder (Embedding size = 8)	8

Figure 11: Experiments for clustering

As shown in figure 13, there are 3 metrics utilized to evaluate the quality of clusters from each experiment. The 3 metrics use different approaches to basically compare the separation between the clusters versus the proximity of the points within a cluster.

Metric number	Name	Description
1	Silhouette score	Best value is 1 and worst value is -1
2	Calinski Harabasz	Higher value is better
3	Davies Bouldin	Lower value is better

Figure 3: Cluster evaluation metrics

The experiment results are shown in figure 13. Experiment_1.2 shows the best overall cluster quality. Experiment_1.2 belongs to the simplest group of scenarios and involves direct clustering of the pre-processed data.

Experiment number	Name	Training Loss	Silhouette Score	Calinski Harabasz Score	Davies Bouldin Score
0	Experiment_1.1	0.0000	0.3232	557.8044	1.0660
1	Experiment_1.2	0.0000	0.3598	544.8101	0.9070
2	Experiment_1.3	0.0000	0.3098	501.3407	1.0119
3	Experiment_2.1	0.0008	0.2694	351.6826	1.2175
4	Experiment_2.2	0.0012	0.2706	415.8404	1.1357
5	Experiment_2.3	0.0009	0.2471	362.8166	1.3458
6	Experiment_3.1	0.0002	0.3428	469.4927	0.9601
7	Experiment_3.2	0.0000	0.2732	393.0455	1.2373
8	Experiment_3.3	0.0001	0.2509	378.9276	1.4002
9	Experiment_4.1	0.0001	0.3016	410.2725	0.9816
10	Experiment_4.2	0.0001	0.2571	390.1845	1.2173
11	Experiment_4.3	0.0001	0.3202	463.9392	1.0491

Figure 4: Experiment results

Among the experiments utilizing embeddings, Experiment_3.1 shows the best overall cluster quality. The hypothesis is that the clustering approaches using embeddings will yield better results with significantly higher input data.

Step 6: Deployment

The application is deployed as a web application. As shown in figure 15, he landing page consists of an overview of the clusters and a field for the user to select a cluster number.

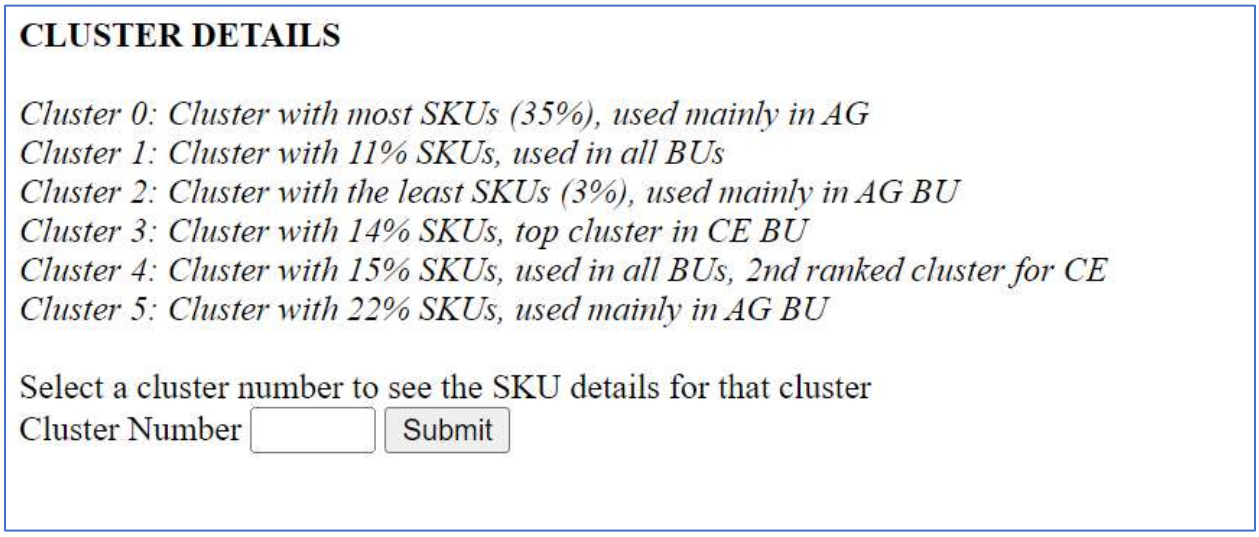

Figure 5: Landing page of web application

For the cluster selected by the user, the resulting page (show in figure 16) shows the business unit and regional details for the cluster.

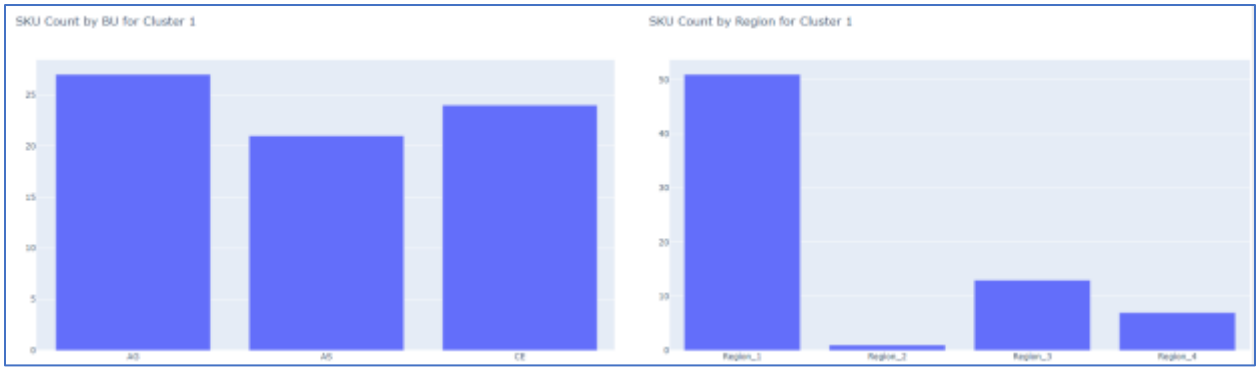

Figure 6: Cluster details (shown in web application)