DAWN OR DOOM 2018
EMERGING TECHNOLOGY – RISKS AND REWARDS

Balanced K-Means Algorithm with Equitable Distribution of Power Ratings

Ananthapadmanabhan Sivasankaran, Muthuraja Palaniappan, Matthew A. Lanham
Purdue University Krannert School of Management
asivasa@purdue.edu; palaniam@purdue.edu; lanhamm@purdue.edu

Abstract

- K-Means algorithm divides data into clusters that are similar
- Similarity is based on Euclidean distance
- No limits on minimum and maximum number of observations per cluster
- Cluster-to-cluster similarity ignored at the cost of within-cluster similarity
- We implemented a heuristic k-means algorithm similar to (Shunzhi Zhu, et al., 2010) to transform the size-constrained clustering problem into a Linear Programming (LP) approach
- Formed clusters within specified upper and lower bound constraints, and ensured similarity (rather than difference) among clusters based on a specified feature
- Business case used for demonstration purpose is DIII Men’s Wrestling conference realignment problem, where schools should be geographically located as close as possible to one another (reduce travel time and costs), have a similar number of schools within a conference, and be similar on average power-rating.

Introduction

- K-Means is efficient in terms of clustering based on Euclidean distances
- But, clusters imbalanced in terms of number of observations might be infeasible to use it in practical scenarios.

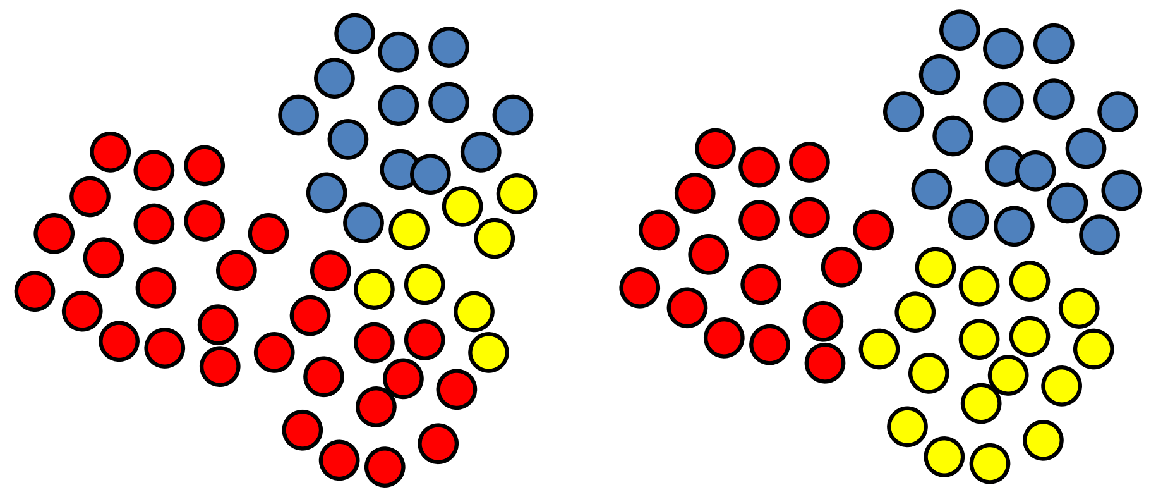


Figure 1. Imbalanced clustering Balanced clustering

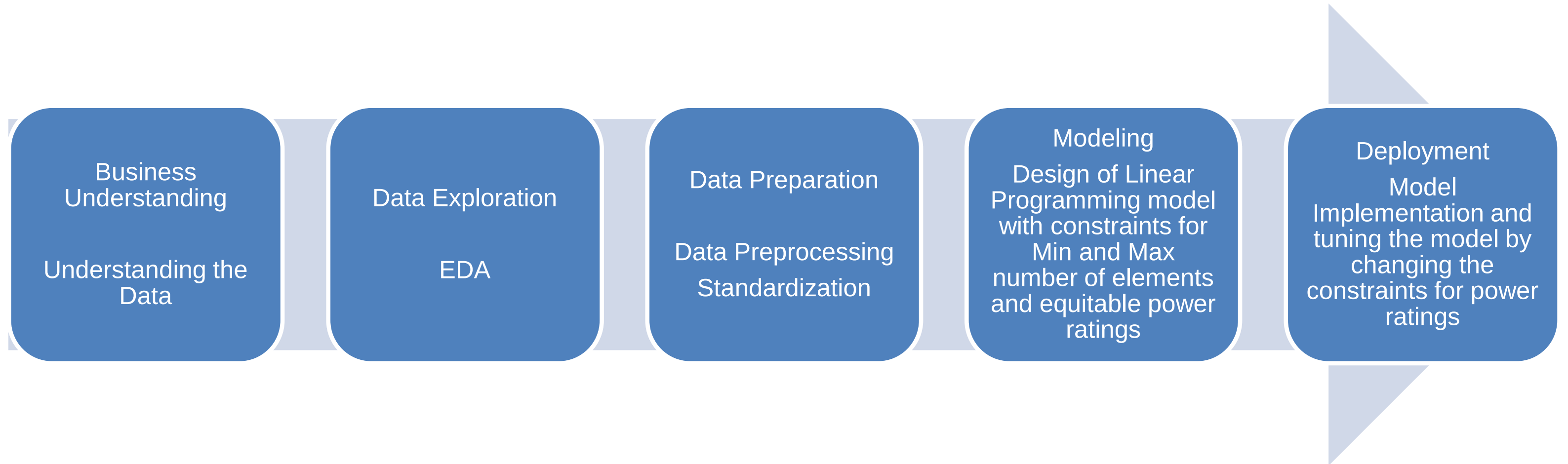
- Applications of balanced clustering with equitable power ratings:
- Supply Chain modeling where distribution centers supply products to geographically clustered stores with different demands.
- Roll out of marketing campaigns in a geography to customers of different earning potential

Literature Review

Our method is unique in that it accounts for additional external information (power ratings). Also, we used an interpretable Linear Programming approach that is low on math complexity.

Balanced k-Means methods	Easily implemented	Low math complexity	Cluster size controlling	Robustness to initialization	Scalable
Multicenter clustering (Liang, et al., 2012)	●		●	●	
MinMax K-Means (Tzortzis et al., 2014)	●		●	●	●
Min-Cut Clustering (Chang, Nie, et al., 2014)	●	●			●
Weight point sets (Borgwardt, Brieden, et al., 2016)			●	●	
Background knowledge (Wagstaff et al., 2001)	●	●	●		
Undersampled (Kumar, Rao, et al., 2014)	●	●			●
FSCL (C. T. Althoff, A. Ulges, A. Dengel, 2000)	●		●	●	
Balanced K-Means with Hungarian algorithm (Mikko I. Malinen et al., 2014)			●	●	●
Heuristic with Linear Programming (Shunzhi Zhu, et al., 2010)	●	●	●	●	
Size-regularized inter-cluster similarity (Chen, et al. 2005)			●	●	
Balanced K-Means with equitable distribution of power ratings (this approach)	●	●	●	●	

Methodology



Business Understanding

Understanding the Data

Data Exploration

EDA

Data Preparation

Data Preprocessing

Standardization

Modeling

Design of Linear Programming model with constraints for Min and Max number of elements and equitable power ratings

Deployment

Model Implementation and tuning the model by changing the constraints for power ratings

- We implemented a Linear Programming (LP) approach to optimize constraints of maximum and minimum number of elements per cluster as well as equitable distribution of power ratings in each cluster.

Optimization Parameters:

$$Minimize: \sum_{i=1}^{n}\sum_{j=1}^{k} d_{ij} * b_{ij}$$

$$Subject\ to:$$

$$\sum b_i \geq Minimum\ size\ of\ the\ cluster$$

$$\sum b_i \leq Maximum\ size\ of\ the\ cluster$$

$$\Sigma p_i b_i \geq (\mu - delta * \sigma) *\ size\ of\ cluster$$
$$\Sigma p_i b_i \leq (\mu + delta * \sigma) *\ size\ of\ cluster$$

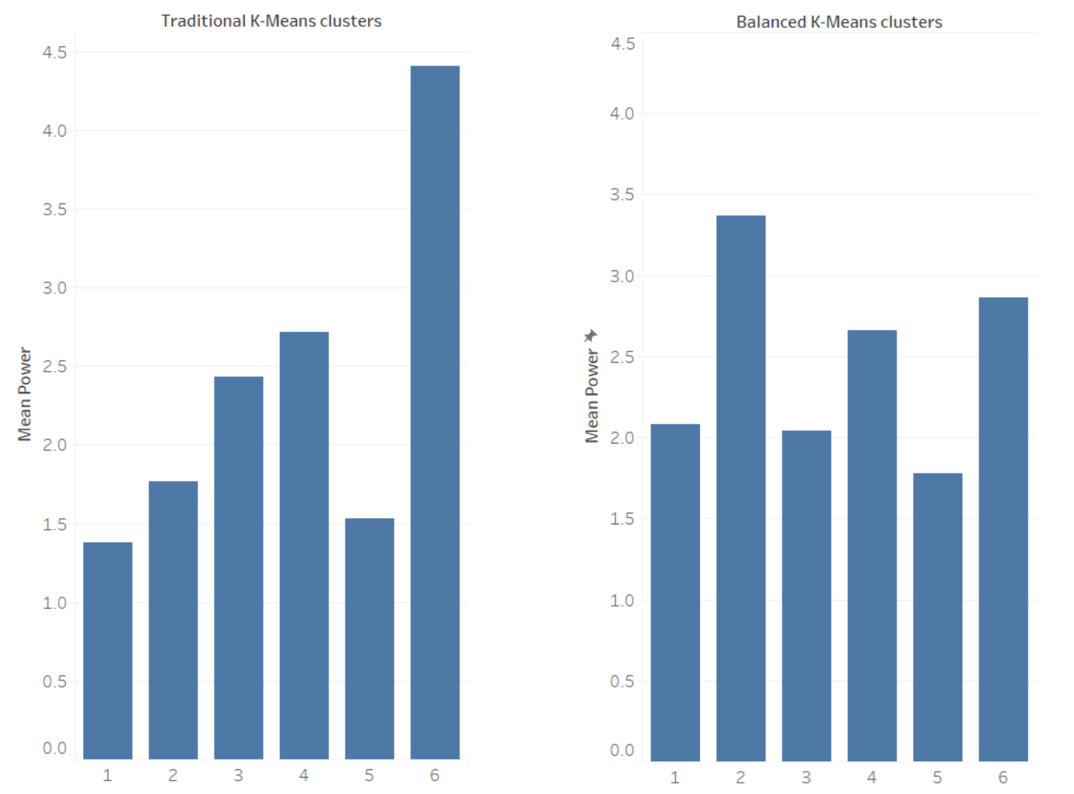
$$\sum b_j = 1$$

$$b_{ij} = \{0,1\}$$

where **μ** and **σ** are the average and standard deviation values of the power rating (target variable) of the population and d is the distance matrix between points and center of each cluster. ***b*** is the matrix of optimal cluster allocation that we want to find. ***p*** is the power rating (target variable) that we would like to maintain near the mean and delta is an user defined value for tolerance in p.

- If minimum is set too high or maximum set too low, the model will not be able to converge.
- Low delta makes the model to force-fit elements in such a way that power ratings are closer to mean of all the power ratings. This leads to the points being widely dispersed in terms of distance.
- With increasing delta, the model fits the elements in a more natural way.

Figure 2. The comparison of mean power ratings between traditional K-Means and Balanced K-Means shows the disparity clearly



Results

- Clusters produced by our Balanced K-Means algorithm satisfy the problem requirements of equitable power ratings.
- These are much better than the traditional K-Means algorithm.
- Implementation in Division III Men’s wrestling will increase the competitiveness of the sport and increase fan following at the national level.
- Also, the fairness of the sport will be restored as the strong teams need not face each other at the regional level and miss out on reaching national level.

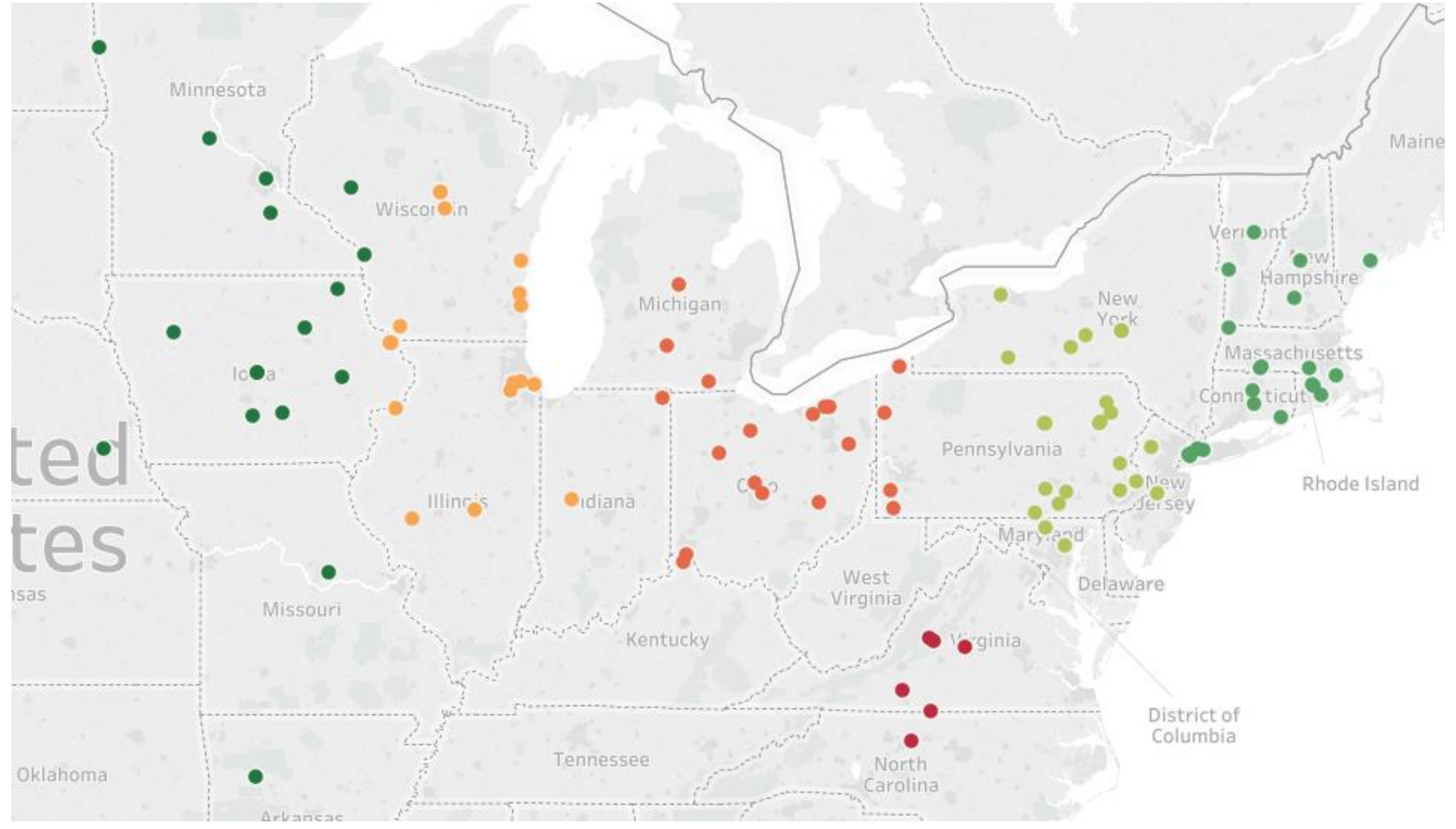


Figure 3. Clusters generated by the standard K-means algorithm

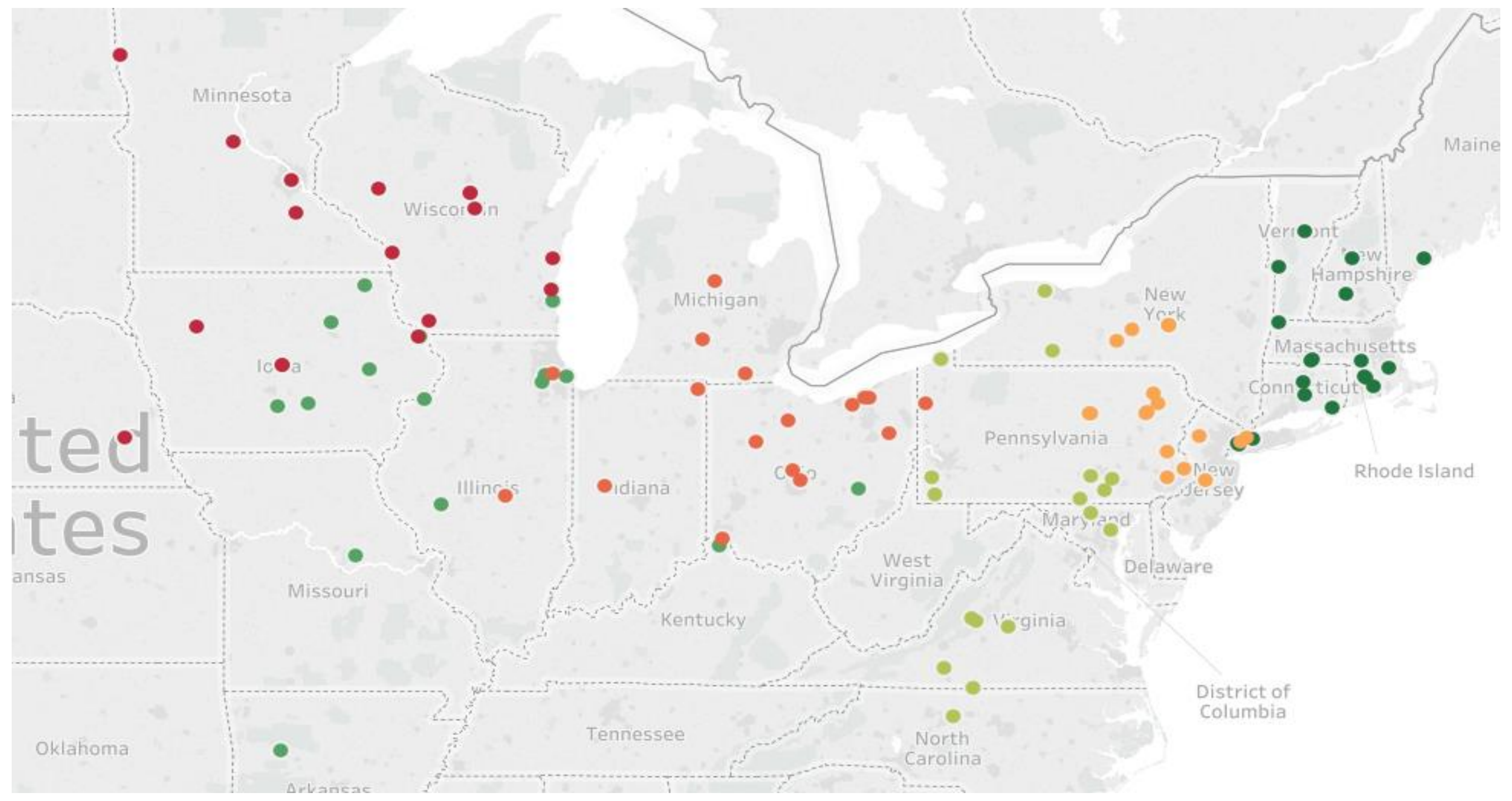


Figure 4. Clusters produced by our algorithm have observations clubbed together in terms of distance and equitably distributed power ratings eg: Wartburg College and Augsburg College are in 2 clusters.

Conclusions

- This simplified approach is beneficial to business as it is easily understood and also clusters the observations appropriately.
- Potential use cases makes it an effective tool for marketers and supply chain professionals.

Acknowledgements

The authors would like to thank Dr. Dave Schrader and Professor Matthew Lanham for providing an opportunity and guidance on this project.