
Prairie View A&M University, TX, 77446

Presentation on:

# Comparing the performance of K-Means and Hierarchical clustering analysis on Online Retail Dataset

By Uboho Victor T.

ID: P21443445

Course: Data Mining

Instructor: Dr. Lin Li

# Outline

- Introduction
- Problem Description
- Data Preprocessing
- Clustering Models
- Experimental Results
- Conclusion

# Introduction

## › What is Clustering

- Clustering is the grouping of a particular set of objects based on their characteristics.
- This methodology partitions the data implementing a specific join algorithm, most suitable for the desired information analysis

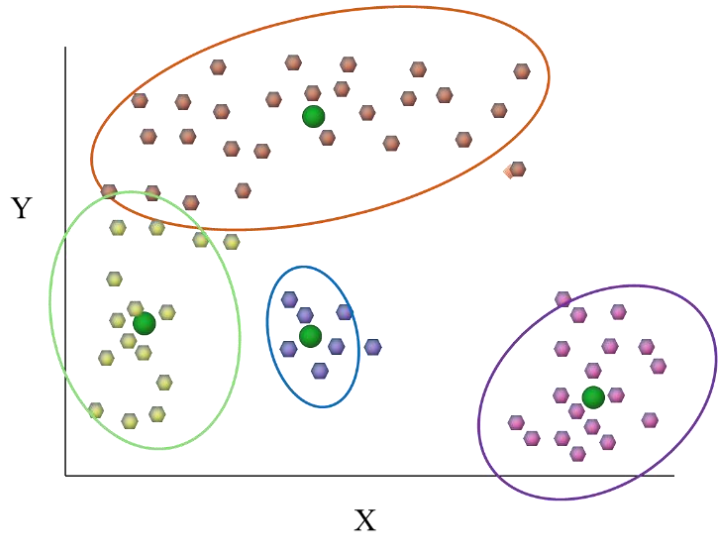



Fig 1: Clustering diagram

**Some applications of Clustering**

- Image processing
- Building co-expressed genes
- Partitioning market segments
- Lexical Ambiguity in NLP

# Problem Description

› The aim of the project is to carry out model evaluation and performance measurement on two clustering algorithms (K-Means and Hierarchical)

› Why is this important?

- ❖ There always lies the need to find the best model that represents our data and how well the chosen model will work in the future
- ❖ Performance of different cluster algorithms vary with respect to the given dataset and the associated features
- ❖ Critical in the field of machine learning

# Data Preprocessing

## › Dataset Information

### Online Retail Dataset

| | | | | | |
|---|---|---|---|---|---|
| **Data Set Characteristics:** | Multivariate, Sequential, Time-Series | **Number of Instances:** | 541909 | **Area:** | **Business** |
| **Attribute Characteristics:** | Integer, Real | Number of Attributes: | 8 | Date Donated | 2015-11-06 |
| **Associated Tasks:** | Classification, Clustering | Missing Values? | N/A | Number of Web Hits: | 272968 |

## › Source

Dr Daqing Chen, Director: Public Analytics group. chend '@' lsbu.ac.uk, School of Engineering, London South Bank University, London SE1 0AA, UK.

## › Preprocessing stages

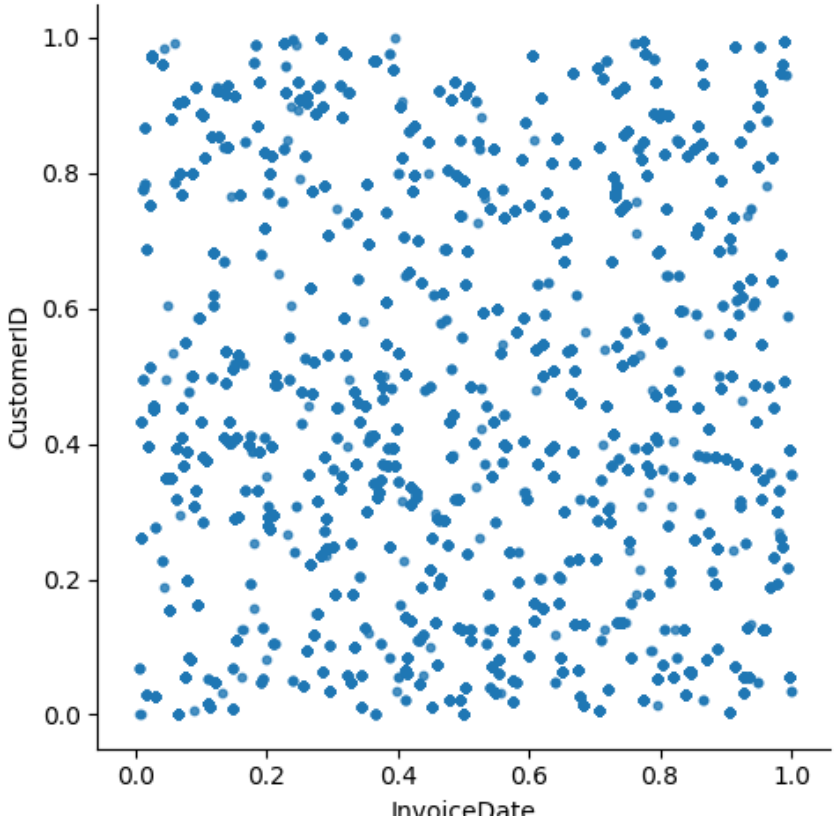
- › Importing libraries and dataset
- › Splitting the dataset
- › Converting categorical features to numeric
- › Normalize the data.

# Data Preprocessing

- Preprocessing stages
- Importing libraries and dataset
  - Some of the libraries used for the data preprocessing includes NumPy, Pandas, Sklearn, Seaborn, Time, etc
- Splitting the dataset
  - Due to the limited computational power, I used 4% of the dataset (20,000 samples) for the clustering.
- Converting categorical features to numeric
  - Alphabetic characters were removed from the “InvoiceNo” column as to convert the column to Numeric.
  - Pandas was also used to convert the categorical features such as ‘Country’ and ‘InvoiceDate’ to numeric features.
- Normalize the data.
  - Firstly, we used the *stats* feature of the *scipy* module to generate the *zscore* of columns with numeric features for the clustering.
  - For the purpose of visualizing the data after clustering, for this, we used the preprocessing feature of *sklearn* module

# Data Preprocessing

› Scatterplot of our data



# CLUSTERING MODELS

## K-MEANS

How it works:

- Place K points (centroids) into the space represented by the objects that are being clustered.
- Assign each object to the group that has the closest centroid.
- Recalculate the positions of the K centroids.
- Repeat Steps 2 and 3 until the centroids no longer move.

This produces a separation of the objects into groups from which the metric to be minimized can be calculated.

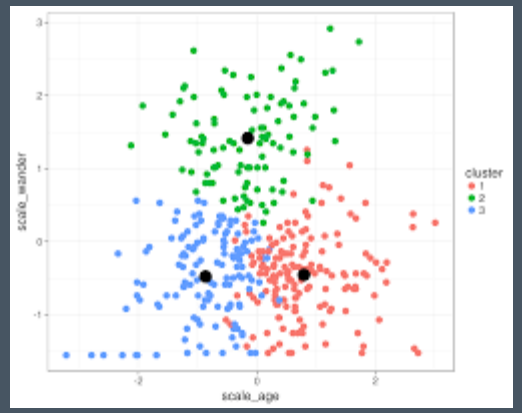



Fig 2: K-Means Clustering

## HIERARCHICAL

How it works:

› Start by assigning each item to a cluster, so that if you have N items, you now have N clusters, each containing just one item

› Find the closest (most similar) pair of clusters and merge them into a single cluster, so that now you have one cluster less.

› Compute distances (similarities) between the new cluster and each of the old clusters.

› Repeat steps 2 and 3 until all items are clustered into a single cluster of size N. (*)


Fig 2: Hierarchical Clustering

# Experimental Results

## K-MEANS

› Elapsed time: 0.09s

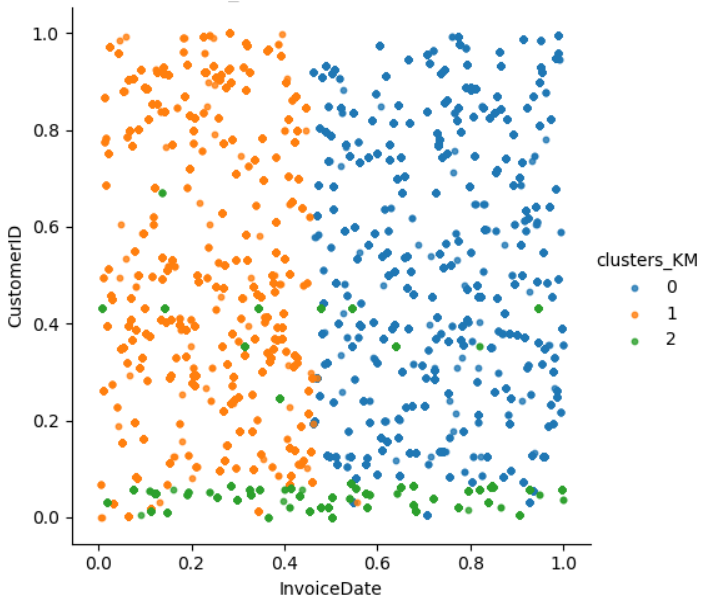



Fig. 3: KMeans Clustering with 3 clusters

## HIERARCHICAL

› Elapsed time: 14.31s

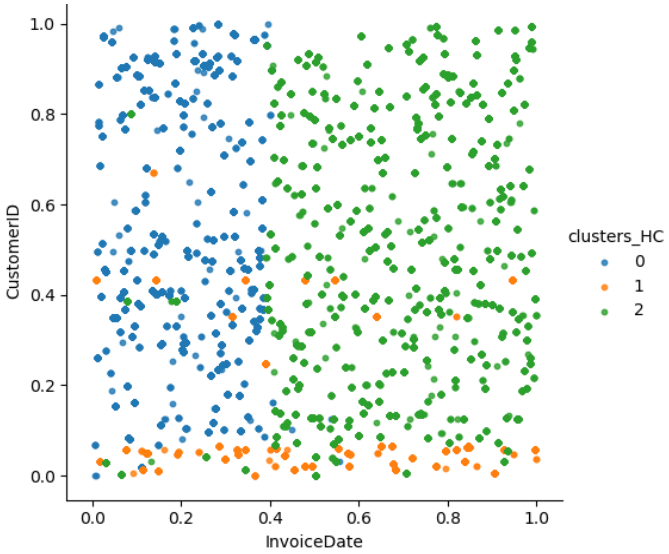



Fig. 4: Hierarchical Clustering with 3 clusters

# Experimental Results

Using different features to plot

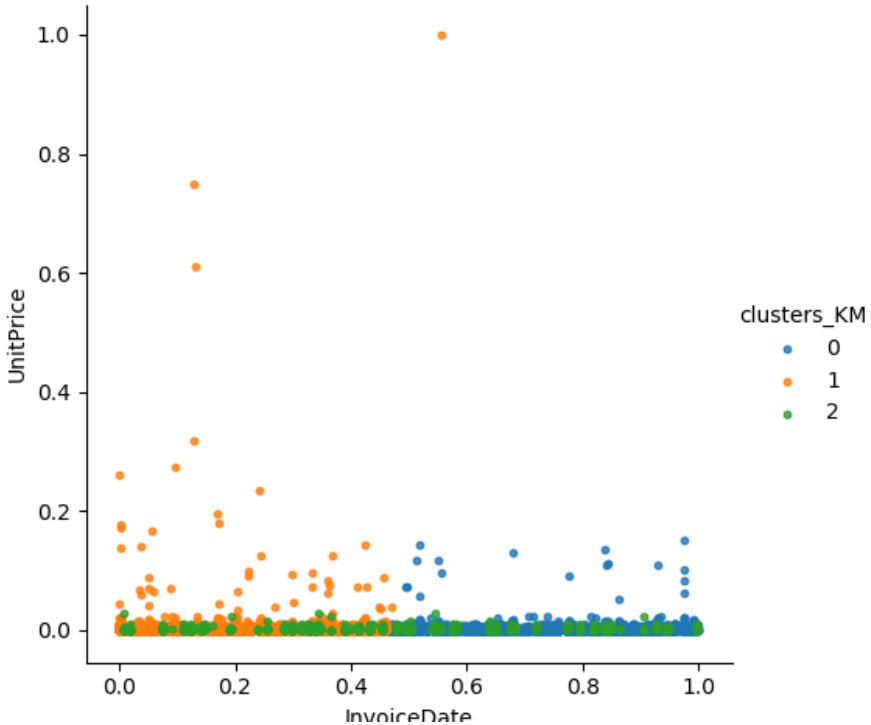



Fig. 3: KMeans Clustering with 3 clusters

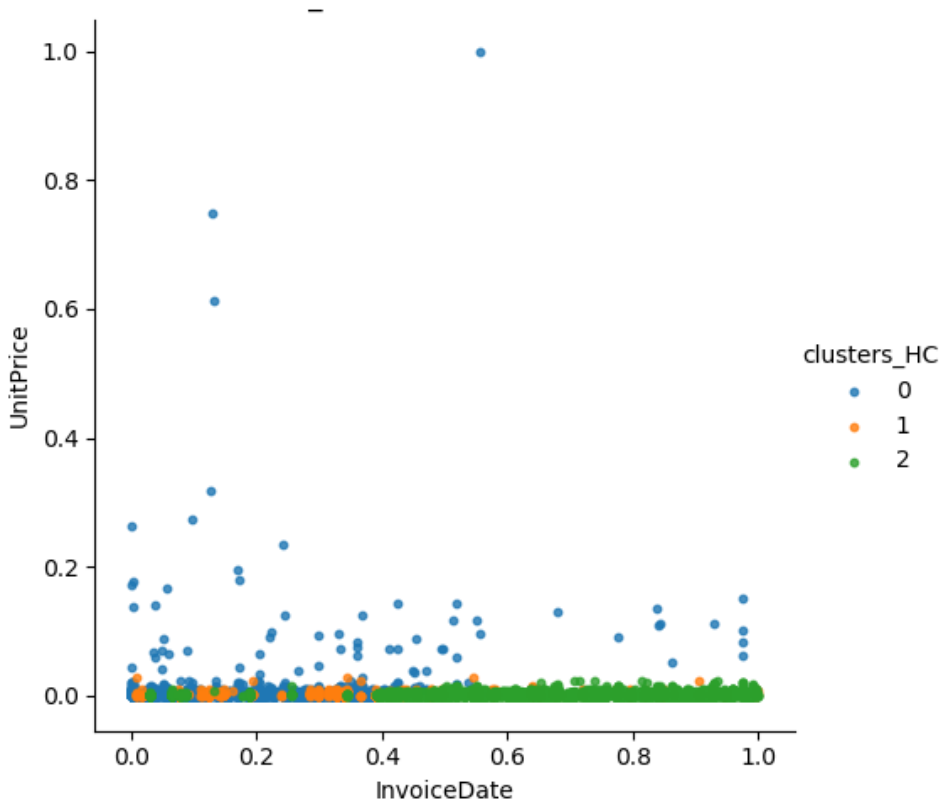



Fig. 4: Hierarchical Clustering with 3 clusters

# Conclusion

The results show that for the given dataset, the two clustering techniques produces similar work. It was also discovered that for visualization purposes, it is important that when plotting the scatter plot graph, the features used in both axes is important in order to better visualize and evaluate the performance of the clustering models.

For future work, I will research and implement different metrics for evaluating clusters such as cluster cohesiveness, completeness and cardinality

# References


Daqing Chen, Sai Liang Sain, and Kun Guo, Data mining for the online retail industry: *A case study of RFM model-based customer segmentation using data mining, Journal of Database Marketing and Customer Strategy Management*, Vol. 19, No. 3, pp. 197â€“208, 2012.

www.bigdatamadesimple.com *What is Clustering in Data Mining*, 2015

J. B. MacQueen (1967) *"Some Methods for classification and Analysis of Multivariate Observations*, *Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability"*, Berkeley, University of California Press, 1:281-297

S. C. Johnson, 1967. *"Hierarchical Clustering Schemes"* Psychometrika, 2:241-254

# Thank you for listening

Q & A