AVENGERS

Data Science Project
on Azure ML

Nattharika Sae Tang 6031748621
Natakorn Ammy Kam 6031772621
Phongsun Worrawattanapreecha 6031795021
Pachara Pattarabodee 6031796721

AGENDA

Our journey today

- Types of Machine Learning
- Introduction
- Azure ML
- K-Means Clustering Model
- Market Basket Analysis
- Summary

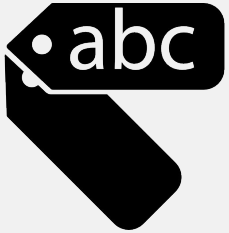
1

Types of Machine Learning

Types of Machine Learning

As with any method, there are different ways to train machine learning algorithms, each with their own advantages and disadvantages.

In Machine Learning, there are 2 types of data...



Labeled data
has both the input and output parameters

Unlabeled data
has one or none of the parameters

Types of Machine Learning

As with any method, there are different ways to train machine learning algorithms, each with their own advantages and disadvantages.

There are 3 main machine learning algorithm.

1

Supervised Learning
Task Driven

2

Unsupervised Learning
Data Driven

3

Reinforcement Learning
Learn from Mistakes

Types of Machine Learning

As with any method, there are different ways to train machine learning algorithms, each with their own advantages and disadvantages.

Supervised Learning

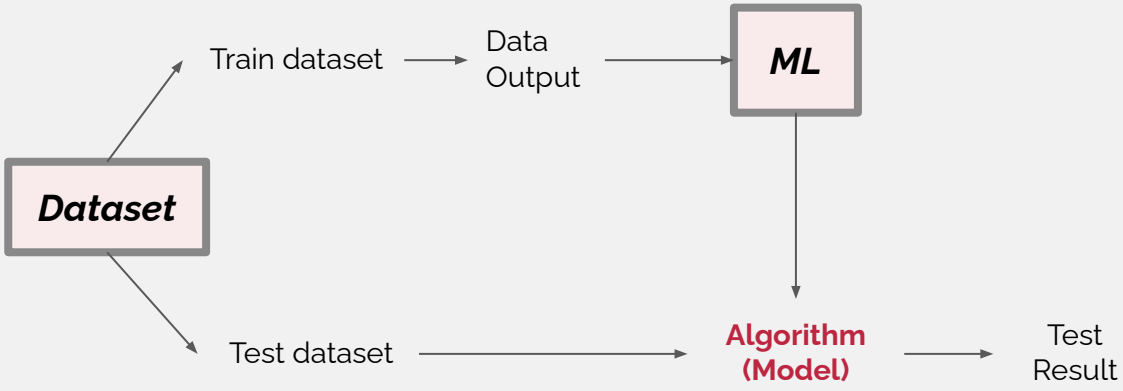
Data: Labeled Data

Algorithm:
- Train Data
- Test Data

Output: Relationship of data

"Teach the data of what the input is"

Goal To find relationship of data and output to create model for further usage



Types of Machine Learning

As with any method, there are different ways to train machine learning algorithms, each with their own advantages and disadvantages.

Supervised Learning

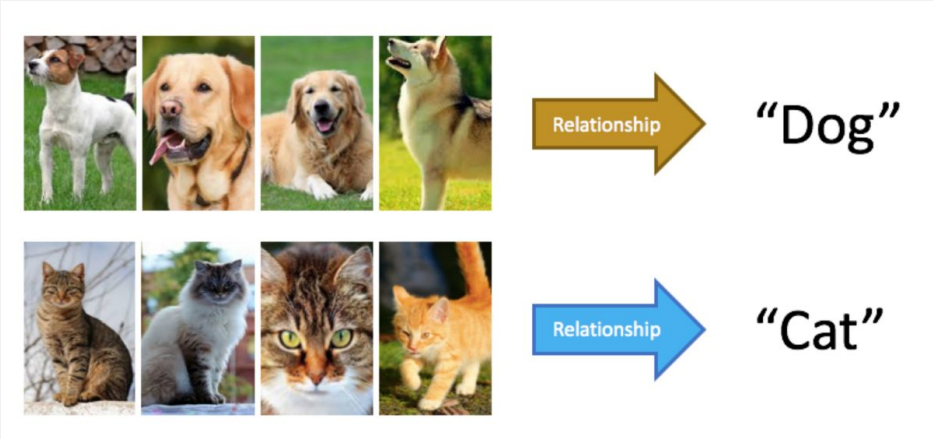
Data: Labeled Data

Algorithm:
- Train Data
- Test Data

Output: Relationship of data

"Teach the data of what the input is"

Ex: Image Classification



Types of Machine Learning

As with any method, there are different ways to train machine learning algorithms, each with their own advantages and disadvantages.

Supervised Learning

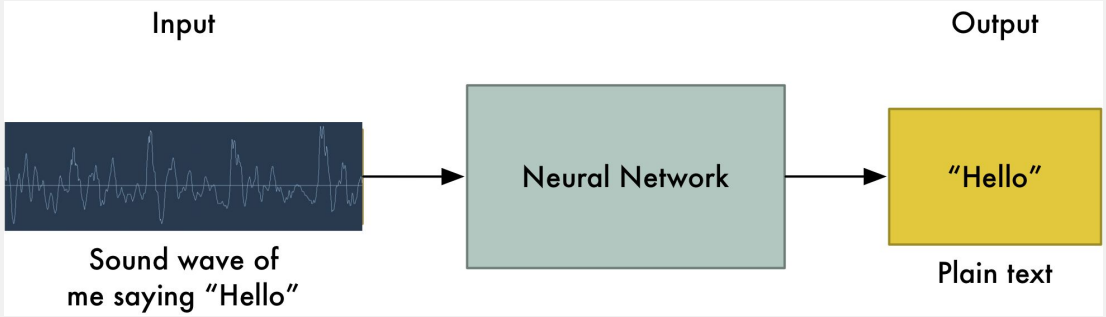
Data: Labeled Data

Algorithm:
- Train Data
- Test Data

Output: Relationship of data

"Teach the data of what the input is"

Ex: Speech Recognition



Types of Machine Learning

As with any method, there are different ways to train machine learning algorithms, each with their own advantages and disadvantages.

Unsupervised Learning

Data: Unlabeled Data

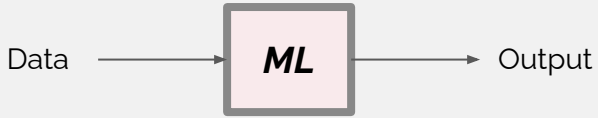
Algorithm: Find relationship of data without labeling

Output: Relationship of data

"Find the descriptive meaning of data"

Goal

1. **Dimensionality**: reduction of information dimension to reduce the complexity
2. **Clustering**: separating the information in different characteristics



No need to input the output of data like Supervised Learning

Types of Machine Learning

As with any method, there are different ways to train machine learning algorithms, each with their own advantages and disadvantages.

Unsupervised Learning

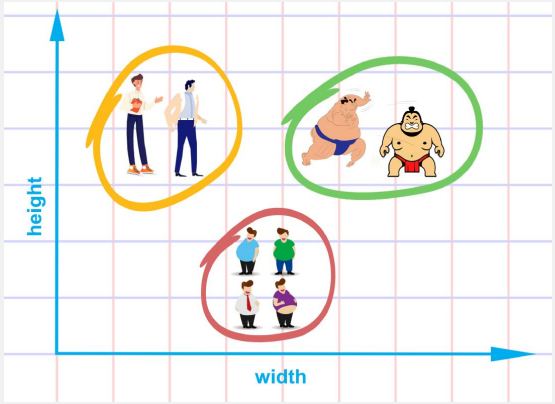
Data: Unlabeled Data

Algorithm: Find relationship of data without labeling

Output: Relationship of data

"Find the descriptive meaning of data"

Ex: Clustering



Types of Machine Learning

As with any method, there are different ways to train machine learning algorithms, each with their own advantages and disadvantages.

Unsupervised Learning

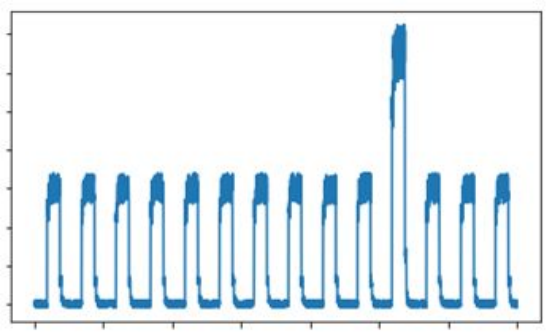
Data: Unlabeled Data

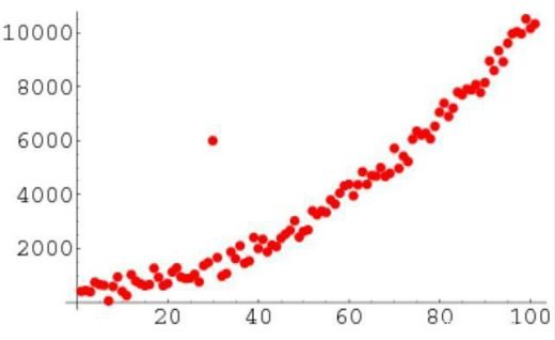
Algorithm: Find relationship of data without labeling

Output: Relationship of data

"Find the descriptive meaning of data"

Ex: Anomaly Detection





Types of Machine Learning

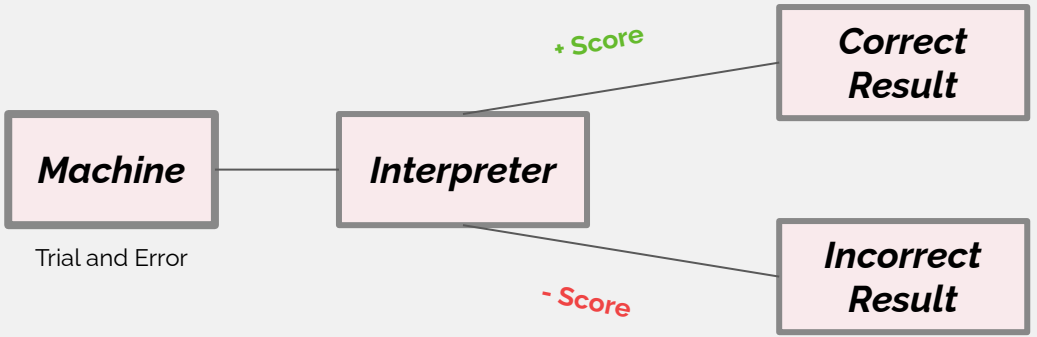
As with any method, there are different ways to train machine learning algorithms, each with their own advantages and disadvantages.

Reinforcement Learning

Algorithm: Find the best algorithm by using an interpreter and reward system with trial-and-error method.

Output: The best solution

Goal Teach Machine to do something for the best result



Types of Machine Learning

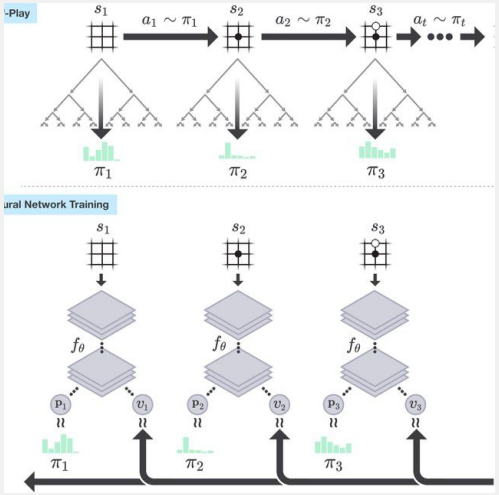
As with any method, there are different ways to train machine learning algorithms, each with their own advantages and disadvantages.

Reinforcement Learning

Algorithm: Find the best algorithm by using an interpreter and reward system with trial-and-error method.

Output: The best solution

Ex: AlphaGo



Types of Machine Learning

As with any method, there are different ways to train machine learning algorithms, each with their own advantages and disadvantages.

Supervised Learning

Data: Labeled Data
Algorithm:
- Train Data
- Test Data

Output: Relationship of data

"Teach the data of what the input is"

Unsupervised Learning

Data: Unlabeled Data

Algorithm: Find relationship of data without labeling

Output: Relationship of data

"Find the descriptive meaning of data"

Reinforcement Learning

Algorithm: Find the best algorithm by using an interpreter and reward system with trial-and-error method.

Output: The best solution

Types of Machine Learning

As with any method, there are different ways to train machine learning algorithms, each with their own advantages and disadvantages.

Supervised Learning

Data: Labeled Data
Algorithm:
- Train Data
- Test Data

Output: Relationship of data

"Teach the data of what the input is"

Unsupervised Learning

Data: Unlabeled Data

Algorithm: Find relationship of data without labeling

Output: Relationship of data

"Find the descriptive meaning of data"

Reinforcement Learning

Algorithm: Find the best algorithm by using an interpreter and reward system with trial-and-error method.

Output: The best solution

2

Introduction

RETAIL STORE

Improve retail business and find insight in many aspects from machine learning

Curiosity

?

Curiosity

?

?

? ? Curiosity ? ? ?

?
?
?
?
Curiosity
?
?
?
?

Curiosity

What's the purchase behavior of customers?
Can we group them in order to analyse type of customers?

Curiosity

What's the purchase behavior of customers? Can we group them in order to analyse type of customers?

Does any group of customer have outstanding characteristic?

Curiosity

What's the purchase behavior of customers? Can we group them in order to analyse type of customers?

Does any group of customer have outstanding characteristic?

Which product is likely to have higher sales when they bundle with product A?

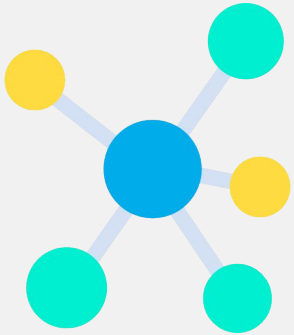
Curiosity

What's the purchase behavior of customers? Can we group them in order to analyse type of customers?

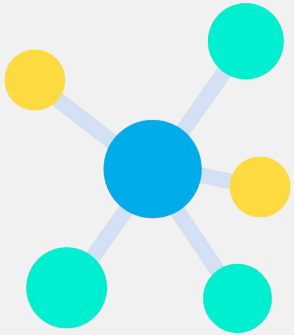
Does any group of customer have outstanding characteristic?

Which product is likely to have higher sales when they bundle with product A?

How shall the staff arrange the product on the shelf? What kind of product shall be place near each other to increase purchase rate?

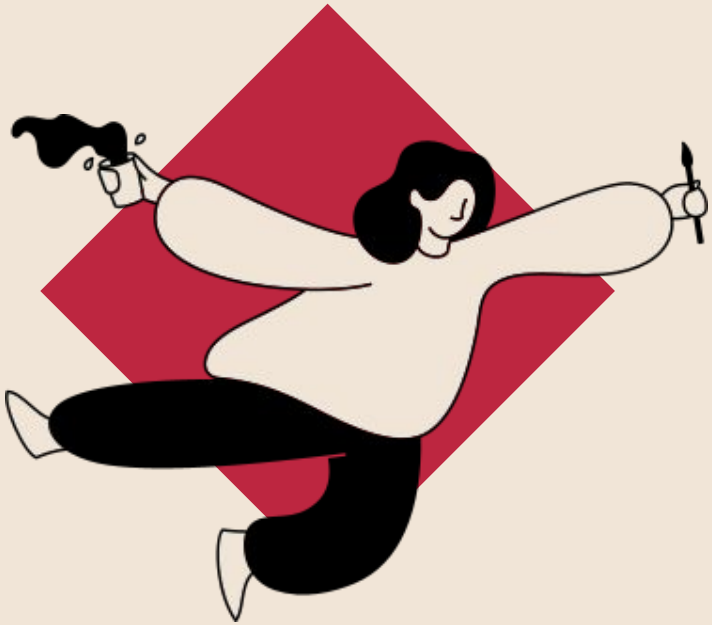


K-Means Clustering



K-Means Clustering

Market Basket Analysis



3

Azure ML

What is Azure Machine Learning studio?

Machine Learning as a Service (MLaaS)

AZURE ML SERVICE

Reasons

AZURE ML SERVICE

Reasons

- Machine Learning as a Service (MLaaS)
- Easy & Flexible building interface

AZURE ML SERVICE

Reasons

- Machine Learning as a Service (MLaaS)
- Easy & Flexible building interface
- Wide range of supported algorithms

AZURE ML SERVICE

Reasons

- Machine Learning as a Service (MLaaS)
- Easy & Flexible building interface
- Wide range of supported algorithms
- Easy implementation of web services

AZURE ML SERVICE

Reasons

- Machine Learning as a Service (MLaaS)
- Easy & Flexible building interface
- Wide range of supported algorithms
- Easy implementation of web services
- Great documentation for Machine Learning Solutions

IMPLEMENTATION

Create Project

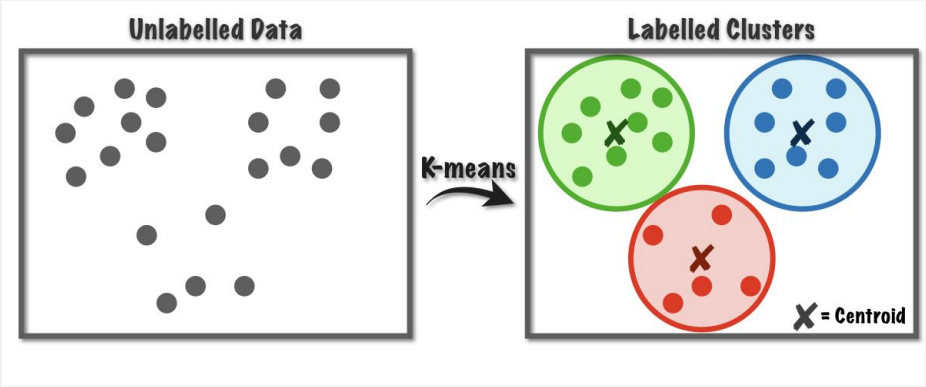
4

K-Means Clustering

K-Means Clustering

Types: Unsupervised Learning

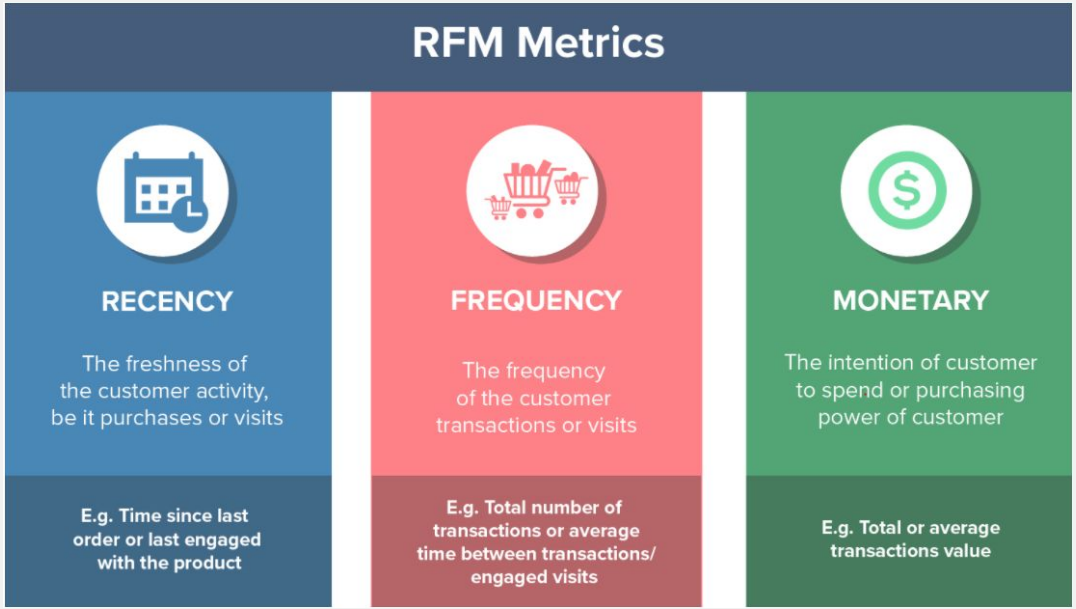
This analysis is used to perform customer **segmentation** based on their profile and behavior. This technique divides the population of data points into a number of groups such that data points in the same group are **similar** to other points in the **same group** and dissimilar to the data points in other groups.



What

RFM Analysis

This analysis is an effective segmentation methods to enable markets to analyze customer behavior. RFM stands for Recency, Frequency, and Monetary value, each corresponding to some key customer trait.



Why

RFM analysis helps answer questions such as:

Who are the best customers?

Why

RFM analysis helps answer questions such as:

Who are the best customers?

Who has the potential to become valuable customers?

Why

RFM analysis helps answer questions such as:

Who are the best customers?

Who has the potential to become valuable customers?

Which of the customers can be retained?

Why

RFM analysis helps answer questions such as:

Who are the best customers?

Who has the potential to become valuable customers?

Which of the customers can be retained?

Which of the customers could contribute to your churn rate?

Why

RFM analysis helps answer questions such as:

Who are the best customers?

Who has the potential to become valuable customers?

Which of the customers can be retained?

Which of the customers could contribute to your churn rate?

Which of your customers are most likely to respond to engagement campaigns?

Why

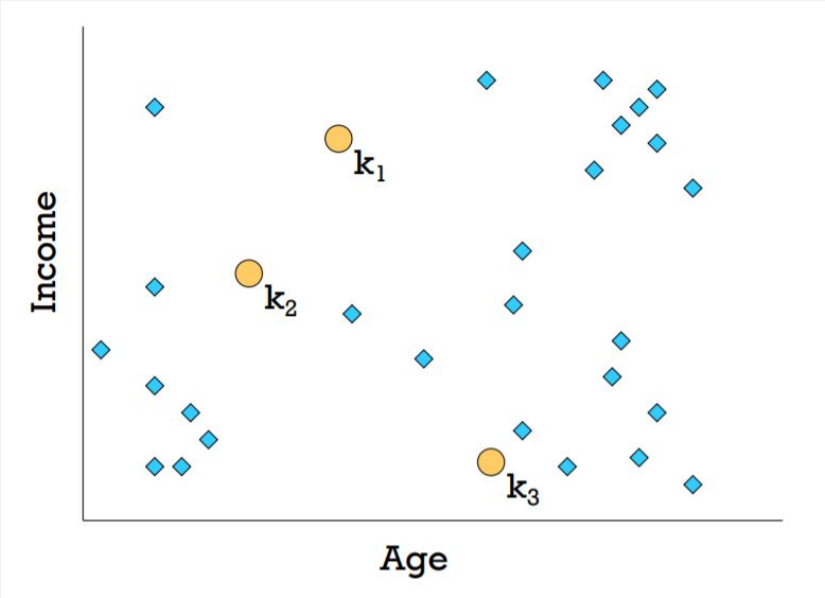
Algorithm

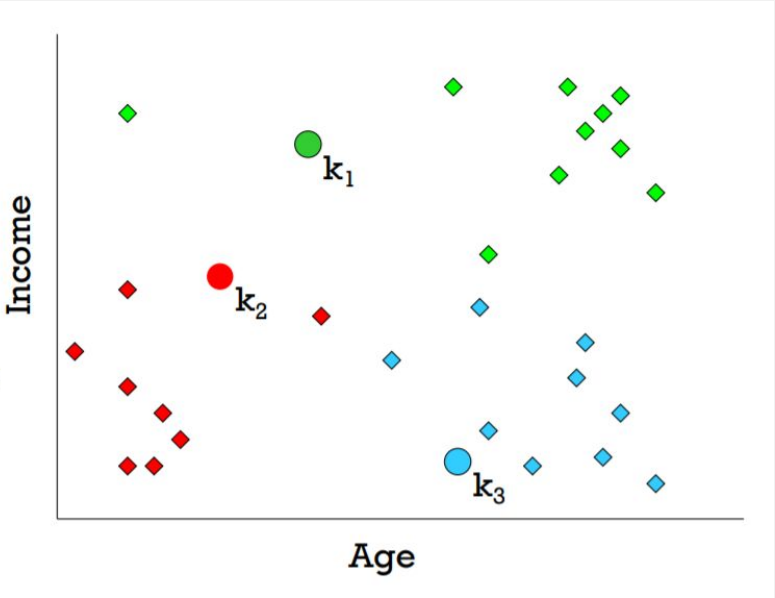
K-Means Clustering

Algorithm

K-Means Clustering

1 Randomly pick centroids and assign each point to the closest centroid





Algorithm

K-Means Clustering

2 Move each centroid to the mean of each cluster

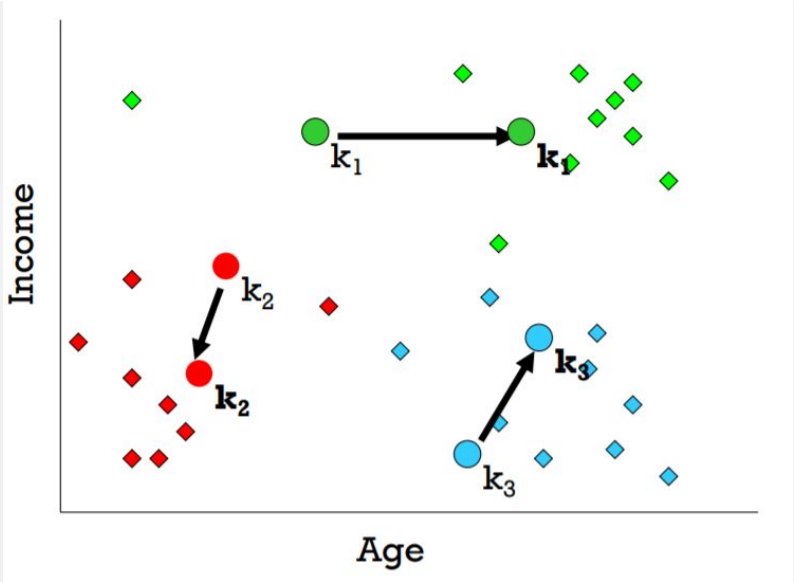
Algorithm

K-Means Clustering

2 Move each centroid to the mean of each cluster

The centroid of a finite set of k points $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k$ in $\mathbb{R}^n$ is

$$\mathbf{C} = \frac{\mathbf{x}_1 + \mathbf{x}_2 + \cdots + \mathbf{x}_k}{k}.$$ [2]



Algorithm

K-Means Clustering

3 Reassign points closest to a different new centroid

Algorithm

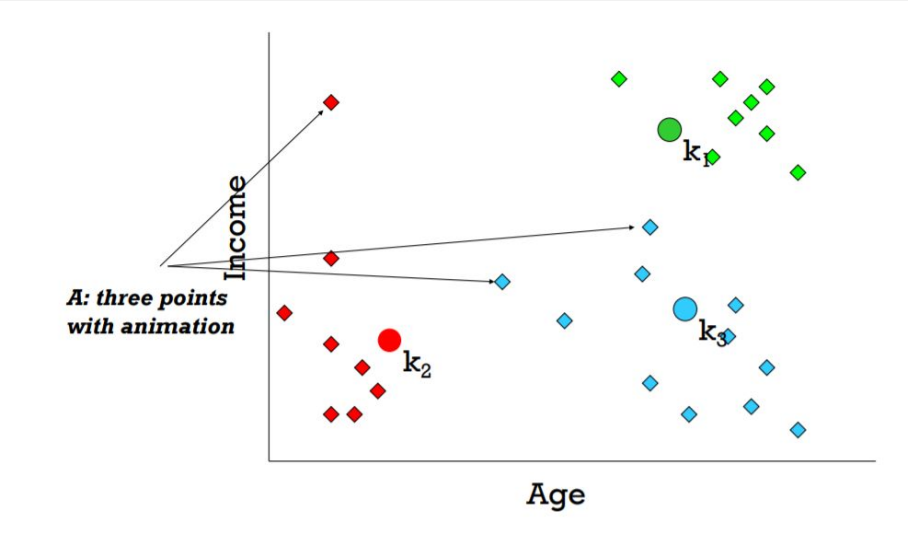
K-Means Clustering

3 Reassign points closest to a different new centroid

The distance to the closest centroid can be calculated by using the distance formula between a pair of samples p and q in an n-dimensional feature space

The distance to the closest centroid

$$\sqrt{\sum_{i=1}^{n}(q_i - p_i)^2}.$$



Algorithm

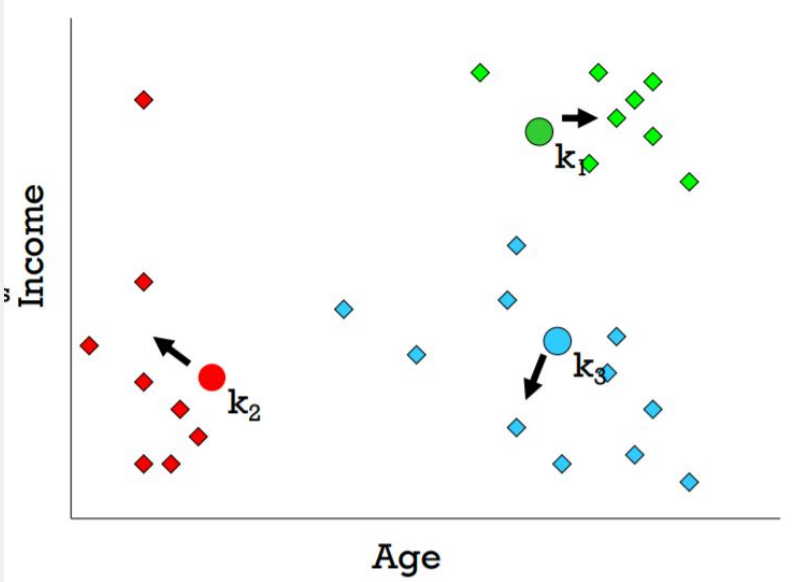
K-Means Clustering

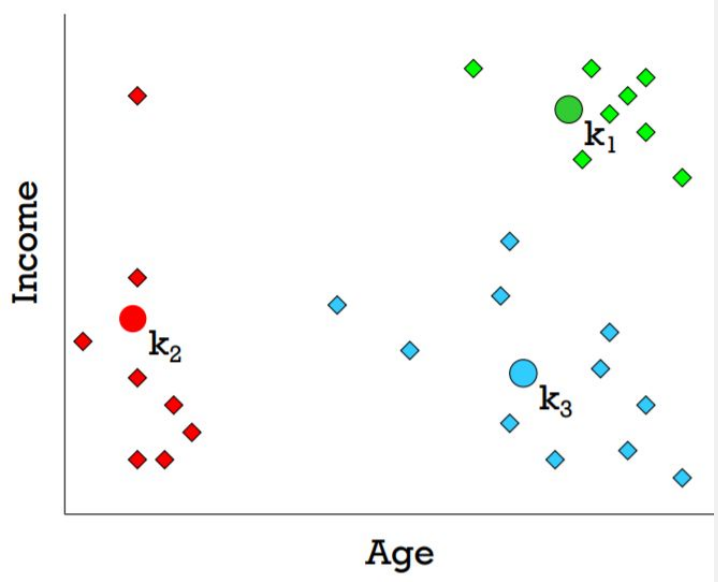
4 Repeat steps 2 & 3 until the centroid converges to a certain value

Algorithm

K-Means Clustering

4 Repeat steps 2 & 3 until the centroid converges to a certain value





Dataset

Transactions.csv

Rate	Qty	Store_type	Tax	Teenhome	total_amt	tran_date	Year	Quarter	Month	Day	transaction_id
210	5	Flagship store	110.25	0	1160.25	12/12/2013	2013	Qtr 4	12	12	16197868036
-210	-5	Flagship store	110.25	0	-1160.25	16/12/2013	2013	Qtr 4	12	16	16197868036
321	3	TeleShop	101.115	0	1064.115	14/8/2012	2012	Qtr 3	8	14	12644501524
187	5	TeleShop	98.175	0	1033.175	13/1/2014	2014	Qtr 1	1	13	87243835584
806	1	Flagship store	84.63	0	890.63	26/4/2012	2012	Qtr 2	4	26	63314547725
650	4	e-Shop	273	0	2873	7/9/2012	2012	Qtr 3	9	7	19516063887
868	5	e-Shop	455.7	0	4795.7	30/1/2013	2013	Qtr 1	1	30	56902862040
312	3	Flagship store	98.28	0	1034.28	6/5/2013	2013	Qtr 2	5	6	64633435931
1175	5	MBR	616.875	0	6491.875	14/10/2012	2012	Qtr 4	10	14	56844530655

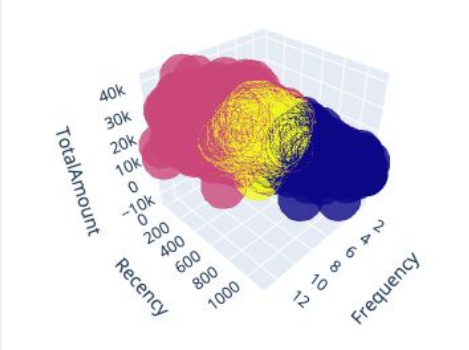
This dataset contains about 23000 rows of data.

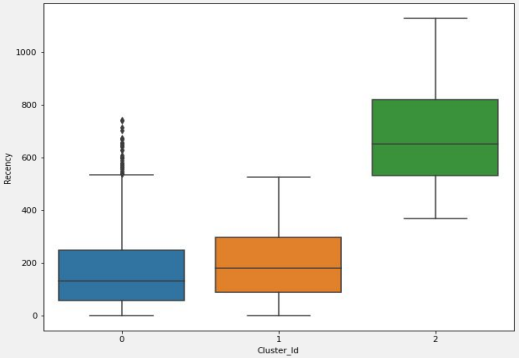
Each row refers to 1 transaction

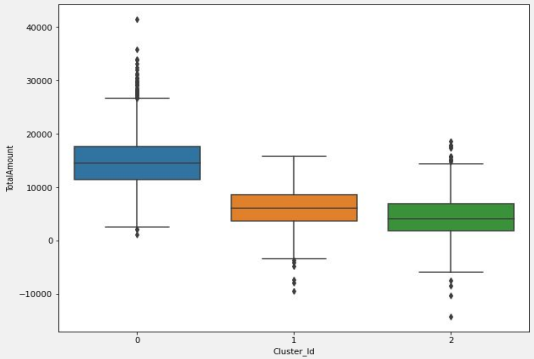
IMPLEMENTATION

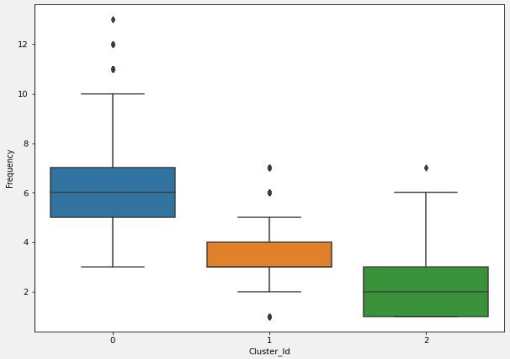
K-Means Clustering With Code

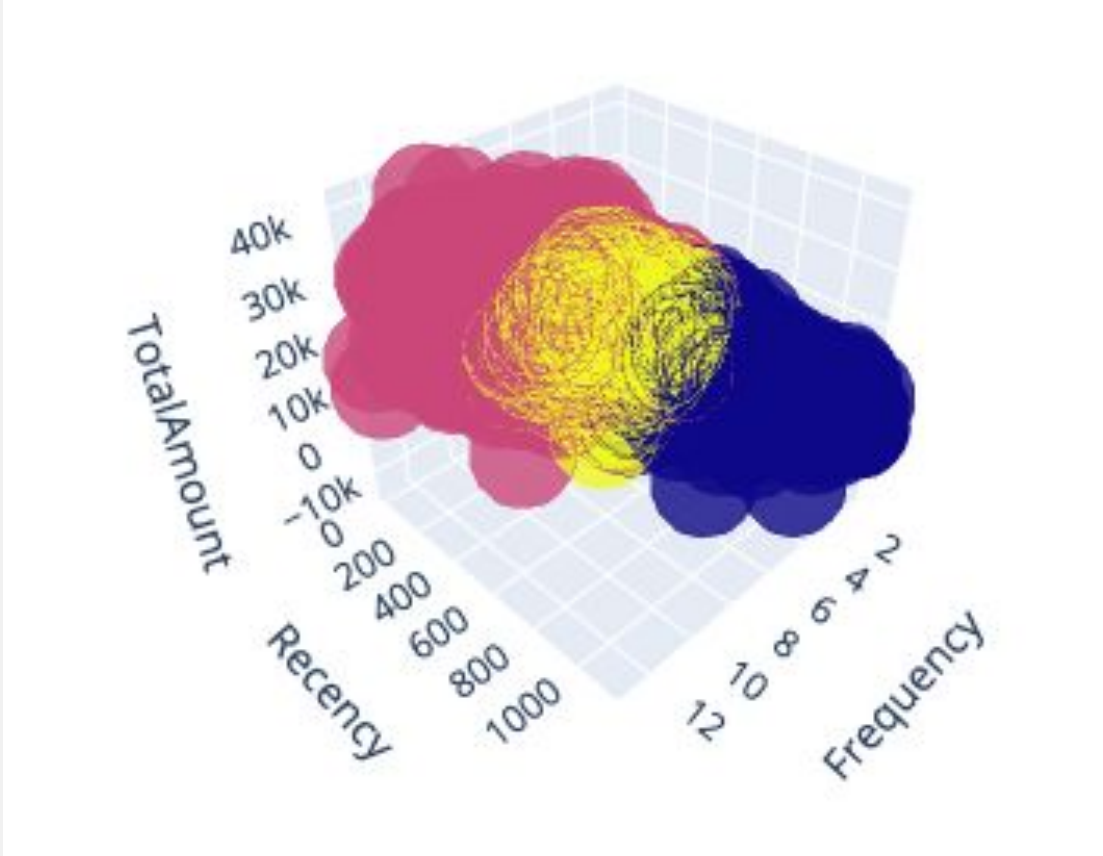
FINDINGS


TotalAmount
40k
30k
20k
10k
0
−10k
Recency
0
200
400
600
800
1000
Frequency
2
4
6
8
10
12

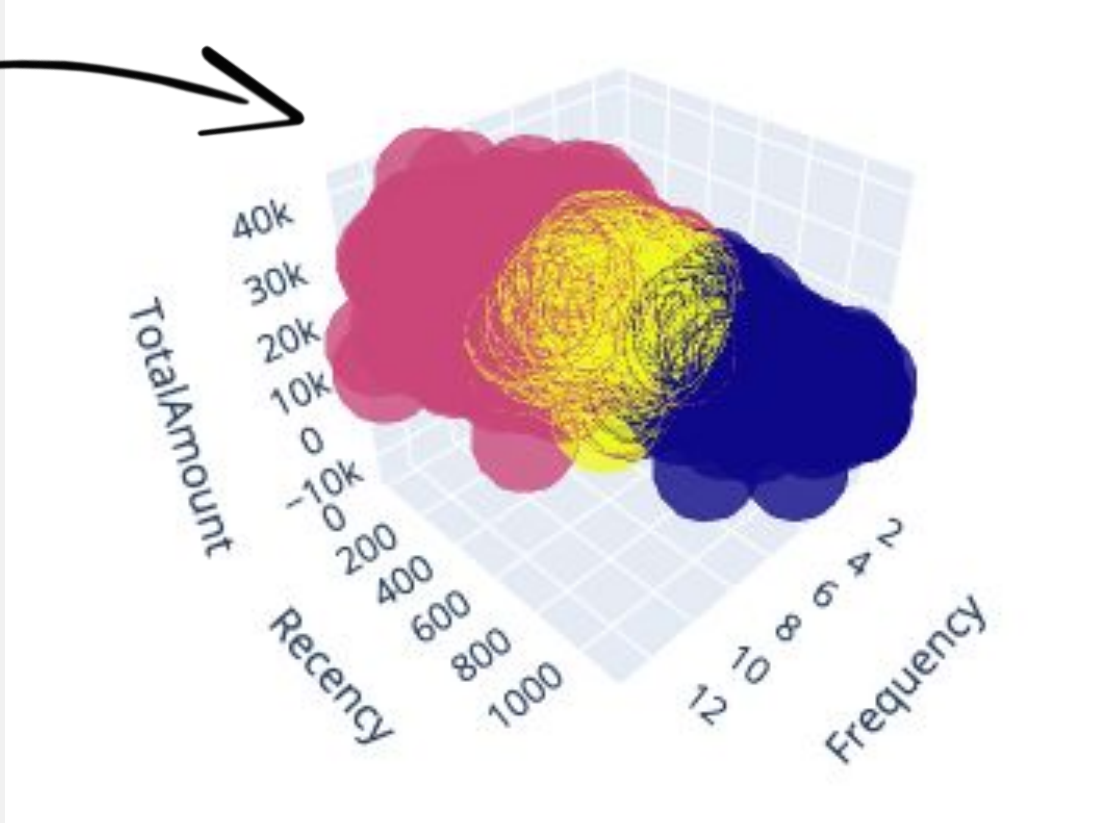

Recency
0
200
400
600
800
1000
Cluster_Id
0
1
2


TotalAmount
40000
30000
20000
10000
0
−10000
Cluster_Id
0
1
2

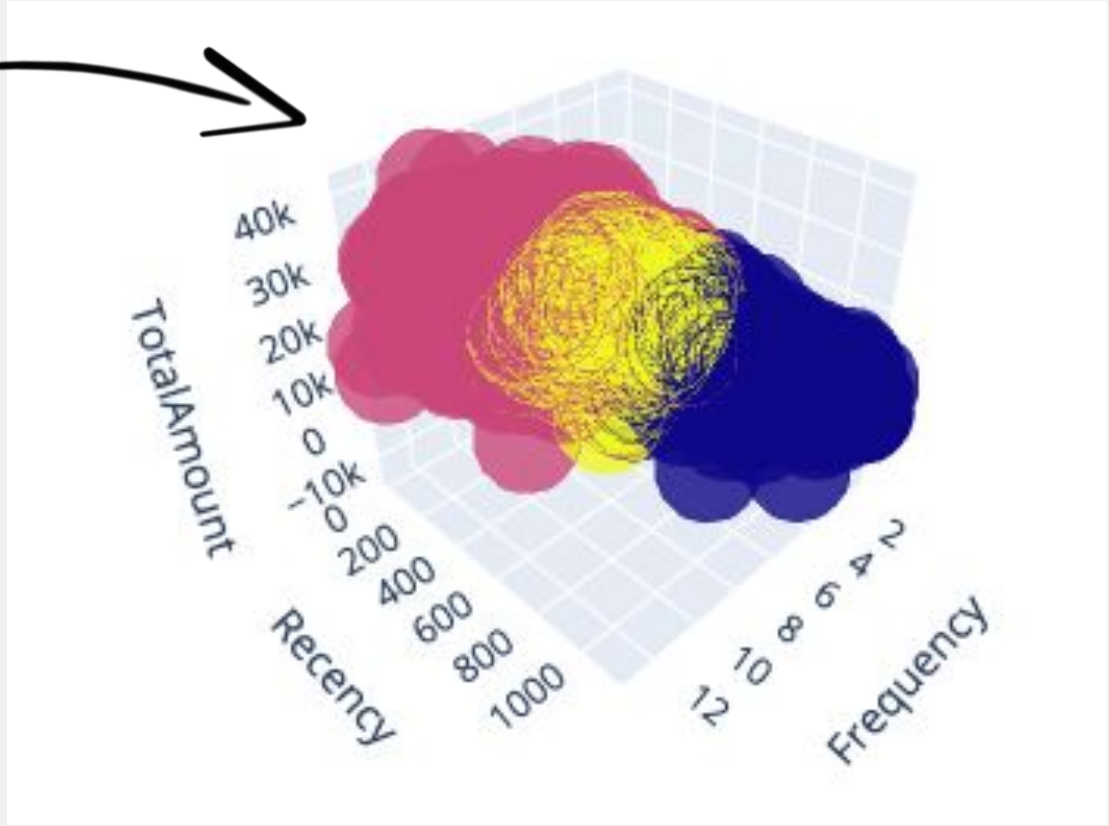

Frequency
2
4
6
8
10
12
Cluster_Id
0
1
2

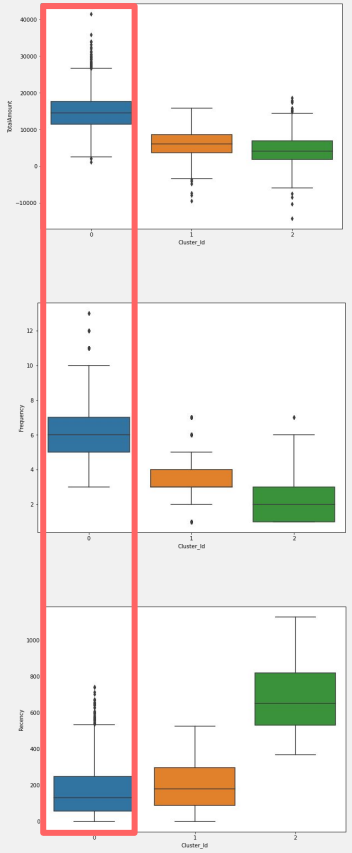

TotalAmount
40k
30k
20k
10k
0
-10k
Recency
0
200
400
600
800
1000
Frequency
12
10
8
6
4
2

Cluster 0


TotalAmount
40k
30k
20k
10k
0
-10k
Recency
0
200
400
600
800
1000
Frequency
12
10
8
6
4
2

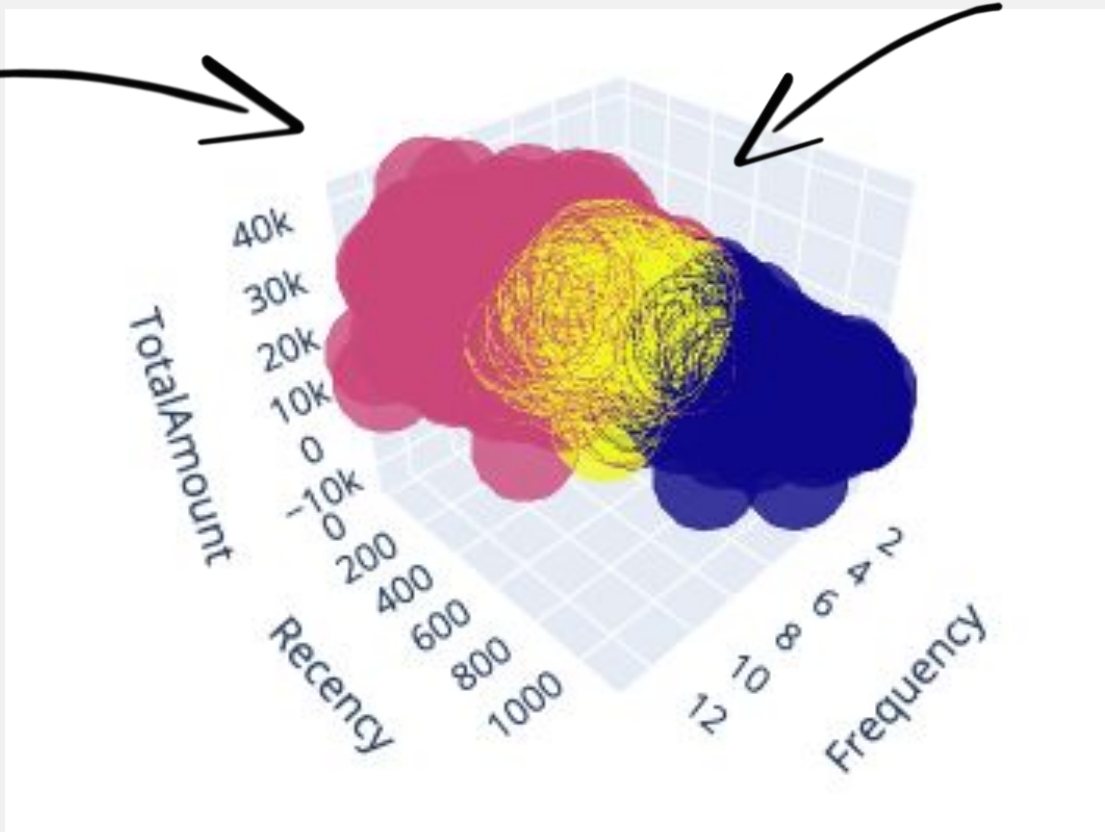
Cluster 0


TotalAmount
40k
30k
20k
10k
0
-10k
Recency
0
200
400
600
800
1000
Frequency
12
10
8
6
4
2


TotalAmount
Frequency
Recency
Cluster_Id

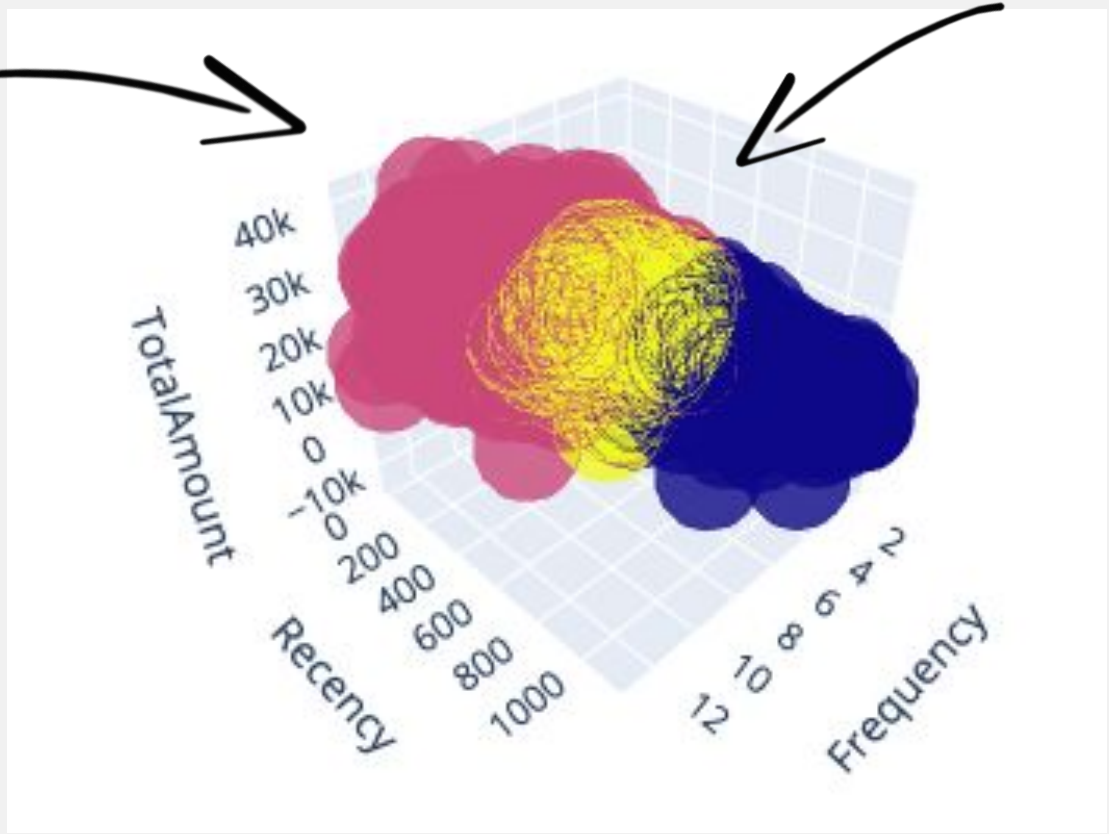
Cluster 1

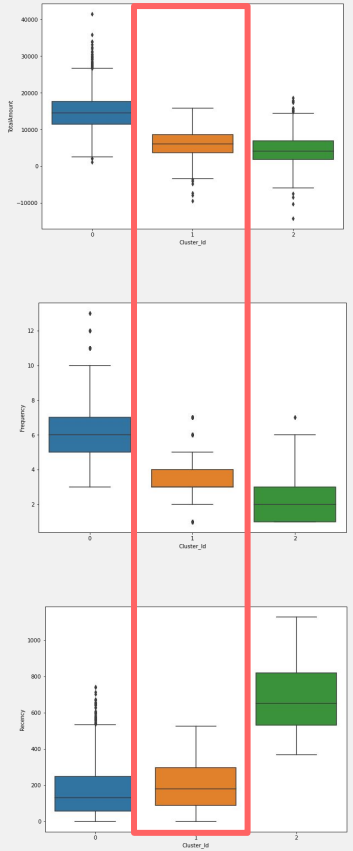
Cluster 0



Cluster 1

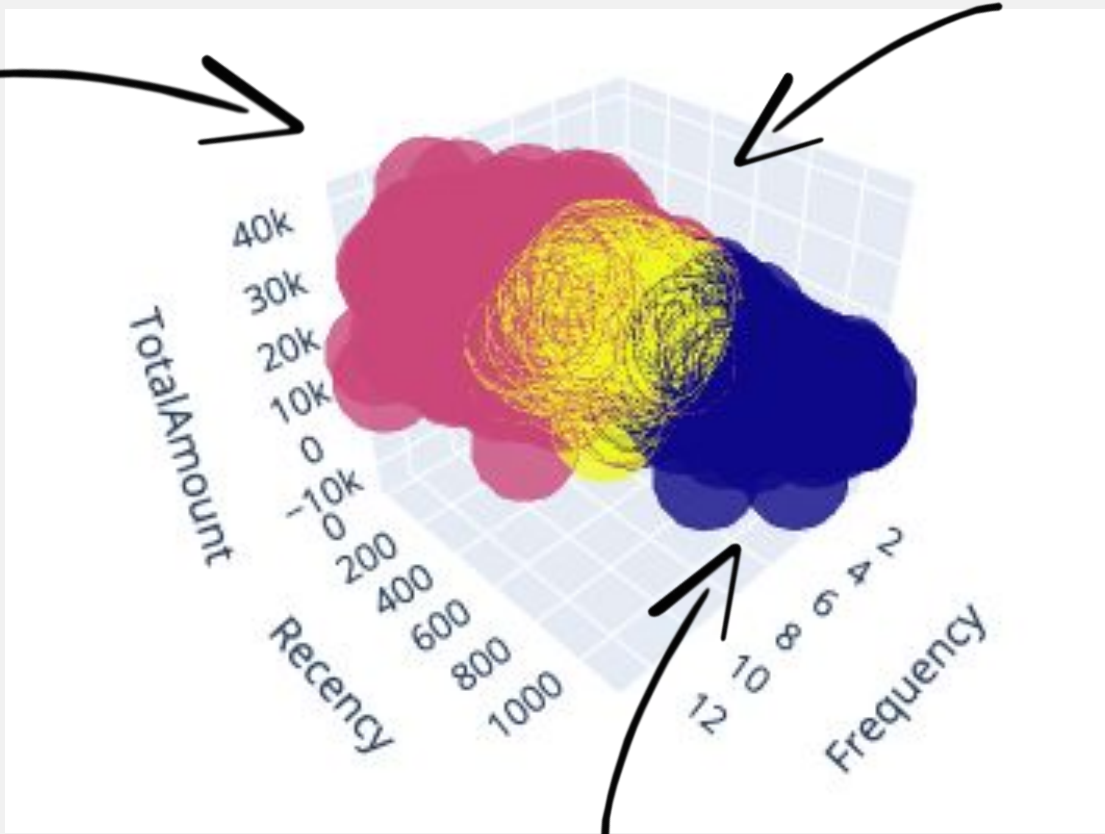
Cluster 0





Cluster 1

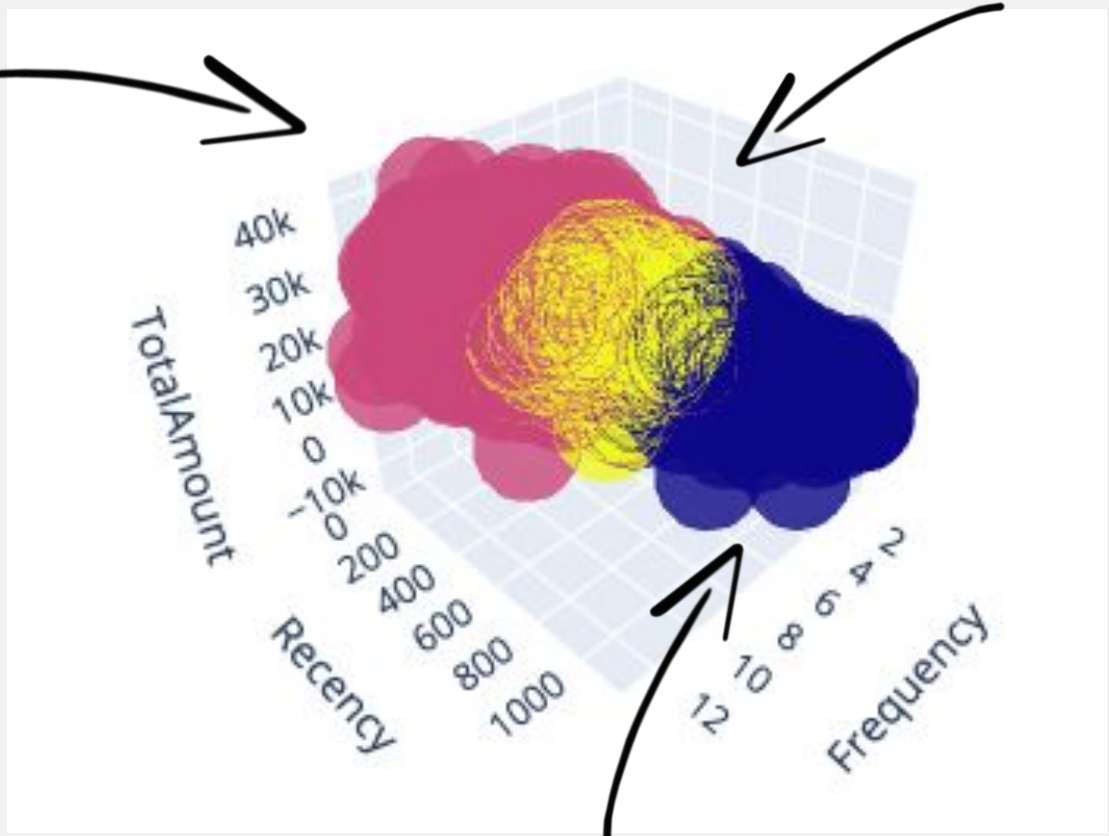
Cluster 0

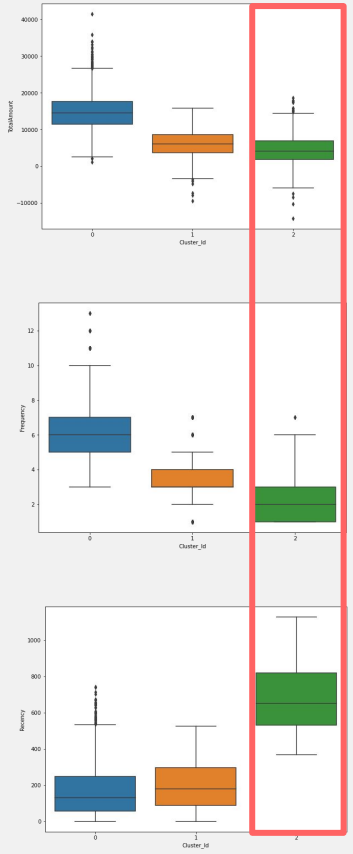


Cluster 2

Cluster 1

Cluster 0





Cluster 2

FINDING

Cluster 0 (Most Important)

- Highest Transaction Amount
- Most Frequent Buyers
- Most Recent Buyers

Cluster 1 (Moderate Importance)

- Moderate Transaction Amount
- Moderate Frequent Buyers
- Moderate Recent Buyers

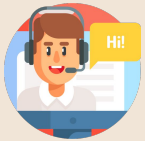
Cluster 2 (Least Importance)

- Lowest Transaction Amount
- Least Frequent Buyers
- Not Recent buyers

FINDING

Cluster 0 (Most Important)

- Highest Transaction Amount
- Most Frequent Buyers
- Most Recent Buyers



Cluster 1 (Moderate Importance)

- Moderate Transaction Amount
- Moderate Frequent Buyers
- Moderate Recent Buyers

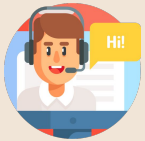
Cluster 2 (Least Importance)

- Lowest Transaction Amount
- Least Frequent Buyers
- Not Recent buyers

FINDING

Cluster 0 (Most Important)

- Highest Transaction Amount
- Most Frequent Buyers
- Most Recent Buyers



Cluster 1 (Moderate Importance)

- Moderate Transaction Amount
- Moderate Frequent Buyers
- Moderate Recent Buyers

Cluster 2 (Least Importance)

- Lowest Transaction Amount
- Least Frequent Buyers
- Not Recent buyers

FINDING

Cluster 0 (Most Important)

- Highest Transaction Amount
- Most Frequent Buyers
- Most Recent Buyers

Cluster 1 (Moderate Importance)

- Moderate Transaction Amount
- Moderate Frequent Buyers
- Moderate Recent Buyers

Cluster 2 (Least Importance)

- Lowest Transaction Amount
- Least Frequent Buyers
- Not Recent buyers

FINDING

Cluster 0 (Most Important)

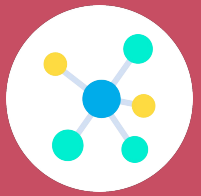
- Highest Transaction Amount
- Most Frequent Buyers
- Most Recent Buyers

Cluster 1 (Moderate Importance)

- Moderate Transaction Amount
- Moderate Frequent Buyers
- Moderate Recent Buyers

Cluster 2 (Least Importance)

- Lowest Transaction Amount
- Least Frequent Buyers
- Not Recent buyers



IMPLEMENTATION

K-Means Clustering Without Code

5

Market Basket Analysis
or
Association Analysis

Market Basket Analysis

or Association Analysion

Types: Unsupervised Learning

This analysis is used to find **association** in information. For example, we would like to know that selecting A first will affect the decision on selecting B or not.

There are 2 types of algorithm to apply in this analysis.

1. **Apriori**
2. **FP-Growth**

What

Market Basket Analysis

or Association Analysion

Types: Unsupervised Learning

This analysis is used to find **association** in information. For example, we would like to know that selecting A first will affect the decision on selecting B or not.

There are 2 types of algorithm to apply in this analysis.

1. **Apriori**
2. **FP-Growth**

What

Why

What can we recommend customer?

If the user select item A, we could suggest them item B which have association with item A.

Why

What can we recommend customer?

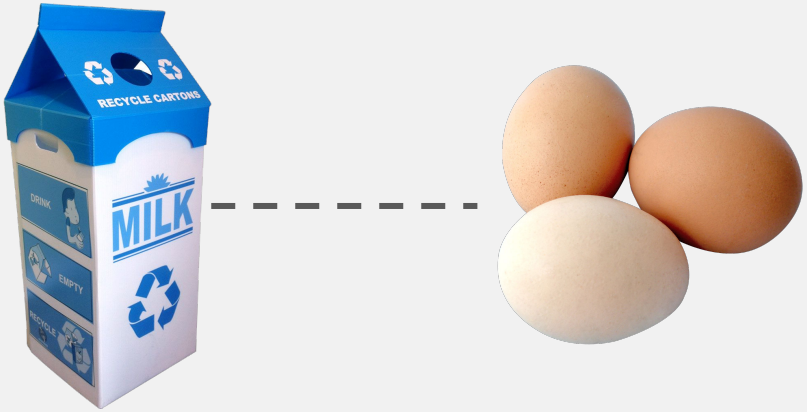
If the user select item A, we could suggest them item B which have association with item A.

Milk

Why

What can we recommend customer?

If the user select item A, we could suggest them item B which have association with item A.



Milk **Egg**

Why

What can we recommend customer?

If the user select item A, we could suggest them item B which have association with item A.

What should be combined to increase sale?

We could combine product that has association into promotion item.

Why

What can we recommend customer?

If the user select item A, we could suggest them item B which have association with item A.

What should be combined to increase sale?

We could combine product that has association into promotion item.

Milk and Egg

Why

What can we recommend customer?

If the user select item A, we could suggest them item B which have association with item A.

What should be combined to increase sale?

We could combine product that has association into promotion item.

Milk and Egg

Why

What can we recommend customer?

If the user select item A, we could suggest them item B which have association with item A.

What should be combined to increase sale?

We could combine product that has association into promotion item.

How can we adjust the shelves?

We can re-adjust how we arrange the item on shelves. Providing item with association together increases the chance to sell products.

Why

Algorithm

Market Basket Analysis

or Association Analysion

Associated Term

1. **Support:** the relative frequency that the relationship show up
   High - implies the useful relationship
   Low - implies the hidden relationship

2. **Confidence:** the measure of the reliability of the rule.
   For example, a confidence level of 0.5 means that we are 50% sure that the association products will be purchased.

3. **Lift:** the ratio of the two data were independent or not
   Lift > 1: Data is more interesting as it is dependent for each other
   Lift = 1: Data is independent to each other making it not interesting data

Algorithm

Market Basket Analysis

or Association Analysion

Apriori Algorithm

is a commonly-applied technique in computational statistics that identifies itemsets that occur with a support greater than a pre-defined value (frequency) and calculates the confidence of all possible rules based on those itemset.

There are 2 major steps in implementing this algorithm.

1. **Find frequency in item set**
   Find the data pattern that occurs together many times, which means that it passes the level of support that we set.

2. **Create association rule**

Algorithm

Market Basket Analysis

or Association Analysion

Apriori Algorithm

is a commonly-applied technique in computational statistics that identifies itemsets that occur with a support greater than a pre-defined value (frequency) and calculates the confidence of all possible rules based on those itemset.

There are 2 major steps in implementing this algorithm.

1. **Find frequency in item set**
   Find the data pattern that occurs together many times, which means that it passes the level of support that we set.

2. **Create association rule**

Let's look into more detail...

Algorithm

Market Basket Analysis

or Association Analysion

1 Clean data

We want to "**One Hot Encoded**" (encode data into category with binary value) the data; therefore, we separate data into rows.

InvoiceNo	Item
1	bread,pasta,eggs
2	cereals,pasta,cider
3	bread,pasta,cereals,cider
4	cereals,cider

Algorithm

Market Basket Analysis

or Association Analysion

1 Clean data

We want to "**One Hot Encoded**" (encode data into category with binary value) the data; therefore, we separate data into rows.

InvoiceNo	Item
1	bread,pasta,eggs
2	cereals,pasta,cider
3	bread,pasta,cereals,cider
4	cereals,cider

InvoiceNo	Bread	Pasta	Eggs	Cereals	Cider
1	1	1	1	0	0
2	0	1	0	1	1
3	1	1	0	1	1
4	0	0	0	1	1

Algorithm

Market Basket Analysis

or Association Analysion

1 Clean data

We want to "**One Hot Encoded**" (encode data into category with binary value) the data; therefore, we separate data into rows.

InvoiceNo	Item
1	**bread,pasta,eggs**
2	cereals,pasta,cider
3	bread,pasta,cereals,cider
4	cereals,cider

InvoiceNo	Bread	Pasta	Eggs	Cereals	Cider
1	**1**	**1**	**1**	0	0
2	0	1	0	1	1
3	1	1	0	1	1
4	0	0	0	1	1

Algorithm

Market Basket Analysis

or Association Analysion

2 Finding support value

Algorithm

Market Basket Analysis

or Association Analysion

2 Finding support value

$$\text{Support (A)} = \frac{\text{Frequent of A}}{\text{N}}$$

We shall set minimum **support of 50%**; therefore, anything that is below this level will be discarded.

Algorithm

Market Basket Analysis

or Association Analysion

2 Finding support value

$$\text{Support (A)} = \frac{\text{Frequent of A}}{\text{N}}$$

We shall set minimum **support of 50%**; therefore, anything that is below this level will be discarded.

InvoiceNo	Bread	Pasta	Eggs	Cereals	Cider
1	1	1	1	0	0
2	0	1	0	1	1
3	1	1	0	1	1
4	0	0	0	1	1
Support	2/4	3/4	1/4	3/4	3/4

Algorithm

Market Basket Analysis

or Association Analysion

2 Finding support value

$$\text{Support (A)} = \frac{\text{Frequent of A}}{\text{N}}$$

We shall set minimum **support of 50%**; therefore, anything that is below this level will be discarded.

InvoiceNo	Bread	Pasta	Eggs	Cereals	Cider
1	1	1	1	0	0
2	0	1	0	1	1
3	1	1	0	1	1
4	0	0	0	1	1
Support	2/4	3/4	1/4	3/4	3/4

Algorithm

Market Basket Analysis

or Association Analysion

2 Combine Items

Algorithm

Market Basket Analysis

or Association Analysion

2 Combine Items

$$\text{Support}(A,B) = \frac{\text{Frequent of } A,B}{N}$$

We shall combine item by 2 and calculate the support level again.

Algorithm

Market Basket Analysis

or Association Analysion

2 Combine Items

$$\text{Support (A,B)} = \frac{\text{Frequent of A,B}}{\text{N}}$$

We shall combine item by 2 and calculate the support level again.

InvoiceNo	Bread, Pasta	Bread, Cereals	Bread, Cider	Pasta, Cereals	Pasta, Cider	Cereals, Cider
1	1	0	0	0	0	0
2	0	0	0	1	1	1
3	1	1	1	1	1	1
4	0	0	0	0	0	1
Support	2/4	1/4	1/4	2/4	2/4	3/4

Algorithm

Market Basket Analysis

or Association Analysion

2 Combine Items

$$\text{Support (A,B)} = \frac{\text{Frequent of A,B}}{\text{N}}$$

We shall combine item by 2 and calculate the support level again.

InvoiceNo	Bread, Pasta	Bread, Cereals	Bread, Cider	Pasta, Cereals	Pasta, Cider	Cereals, Cider
1	1	0	0	0	0	0
2	0	0	0	1	1	1
3	1	1	1	1	1	1
4	0	0	0	0	0	1
Support	2/4	1/4	1/4	2/4	2/4	3/4

Algorithm

Market Basket Analysis

or Association Analysion

3 More combination

Algorithm

Market Basket Analysis

or Association Analysion

3 More combination

$$\text{Support (A,B,C)} = \frac{\text{Frequent of A,B,C}}{\text{N}}$$

We can further the combination on 1 rule: **the first item is the same.**

Algorithm

Market Basket Analysis

or Association Analysion

3 More combination

$$\text{Support (A,B,C)} = \frac{\text{Frequent of A,B,C}}{\text{N}}$$

We can further the combination on 1 rule: **the first item is the same.**

InvoiceNo	Bread, Pasta	**Pasta, Cereals**	**Pasta, Cider**	Cereals, Cider
1	1	0	0	0
2	0	1	1	1
3	1	1	1	1
4	0	0	0	1
Support	2/4	2/4	2/4	3/4

Algorithm

Market Basket Analysis

or Association Analysion

3 More combination

$$\text{Support (A,B,C)} = \frac{\text{Frequent of A,B,C}}{\text{N}}$$

We can further the combination on 1 rule: **the first item is the same.**

InvoiceNo	Pasta, Cereals, Cider
1	0
2	1
3	1
4	0
Support	2/4

Algorithm

Market Basket Analysis

or Association Analysion

3 More combination

$$\text{Support (A,B,C)} = \frac{\text{Frequent of A,B,C}}{\text{N}}$$

We can further the combination on 1 rule: **the first item is the same.**

InvoiceNo	Pasta, Cereals, Cider
1	0
2	1
3	1
4	0
Support	2/4

No more combination available -- Algorithm stops

Algorithm

Market Basket Analysis

or Association Analysion

4 Get frequent item set

Algorithm

Market Basket Analysis

or Association Analysion

4 Get frequent item set

This is the frequent item set that passes the **support level of 50%** which means that this combination occurs about 50% of the transaction.

InvoiceNo	Support	Size
Bread	2/4	1
Pasta	3/4	1
Cereals	3/4	1
Cider	3/4	1
Bread, Pasta	2/4	2
Pasta, Cereals	2/4	2
Pasta, Cider	2/4	2
Cereals, Cider	3/4	2
Pasta, Cereals, Cider	2/4	3

Algorithm

Market Basket Analysis

or Association Analysion

5 Find Confidence

Algorithm

Market Basket Analysis

or Association Analysion

5 Find Confidence

$$\text{Confidence}(A,B) = \frac{\text{Support of } A,B}{\text{Support } A}$$

Then, we will see the **confidence** level in each combination.

Algorithm

Market Basket Analysis

or Association Analysion

5 Find Confidence

$$\text{Confidence (A,B)} = \frac{\text{Support of A,B}}{\text{Support A}}$$

Then, we will see the **confidence** level in each combination.

InvoiceNo	Support	Size	Confidence
Bread, Pasta	2/4	2	1
Pasta, Cereals	2/4	2	2/3
Pasta, Cider	2/4	2	2/3
Cereals, Cider	3/4	2	1
Pasta, Cereals, Cider	2/4	3	1

Algorithm

Market Basket Analysis

or Association Analysion

6 Find Lift

Algorithm

Market Basket Analysis

or Association Analysion

6 Find Lift

$$\text{Lift}(A,B) = \frac{\text{Support of A,B}}{\text{Support A x Support B}}$$

Lastly, the **lift** level of each combination is calculated.

Algorithm

Market Basket Analysis

or Association Analysion

6 Find Lift

$$\text{Lift}(A,B) = \frac{\text{Support of A,B}}{\text{Support A x Support B}}$$

Lastly, the **lift** level of each combination is calculated.

InvoiceNo	Support	Size	Confidence	Lift
Bread, Pasta	2/4	2	1	4/3
Pasta, Cereals	2/4	2	2/3	8/9
Pasta, Cider	2/4	2	2/3	8/9
Cereals, Cider	3/4	2	1	4/3
Pasta, Cereals, Cider	2/4	3	1	32/27

Algorithm

Market Basket Analysis

or Association Analysion

7 Summary

Algorithm

Market Basket Analysis

or Association Analysion

7 Summary

- We care about lift > 1

InvoiceNo	Support	Size	Confidence	Lift
Bread, Pasta	2/4	2	1	4/3
Pasta, Cereals	2/4	2	2/3	8/9
Pasta, Cider	2/4	2	2/3	8/9
Cereals, Cider	3/4	2	1	4/3
Pasta, Cereals, Cider	2/4	3	1	32/27

Algorithm

Market Basket Analysis

or Association Analysion

7 Summary

- We care about lift > 1

InvoiceNo	Support	Size	Confidence	Lift
Bread, Pasta	2/4	2	1	4/3
Pasta, Cereals	2/4	2	2/3	8/9
Pasta, Cider	2/4	2	2/3	8/9
Cereals, Cider	3/4	2	1	4/3
Pasta, Cereals, Cider	2/4	3	1	32/27

Algorithm

Market Basket Analysis

or Association Analysion

7 Summary

- We care about lift > 1
- ***What the store should do. . .***

InvoiceNo	Support	Size	Confidence	Lift
Bread, Pasta	2/4	2	1	4/3
Cereals, Cider	3/4	2	1	4/3
Pasta, Cereals, Cider	2/4	3	1	32/27

Algorithm

Market Basket Analysis

or Association Analysis

7 Summary

- We care about lift > 1
- ***What the store should do. . .***

InvoiceNo	Support	Size	Confidence	Lift
Bread, Pasta	2/4	2	1	4/3
Cereals, Cider	3/4	2	1	4/3
Pasta, Cereals, Cider	2/4	3	1	32/27

Bread, Pasta

Algorithm

Market Basket Analysis

or Association Analysis

7 Summary

- We care about lift > 1
- ***What the store should do. . .***

InvoiceNo	Support	Size	Confidence	Lift
Bread, Pasta	2/4	2	1	4/3
Cereals, Cider	3/4	2	1	4/3
Pasta, Cereals, Cider	2/4	3	1	32/27

Bread, Pasta

Cereals, Cider

Algorithm

Market Basket Analysis

or Association Analysis

7 Summary

- We care about lift > 1
- ***What the store should do. . .***

InvoiceNo	Support	Size	Confidence	Lift
Bread, Pasta	2/4	2	1	4/3
Cereals, Cider	3/4	2	1	4/3
Pasta, Cereals, Cider	2/4	3	1	32/27

Bread, Pasta

Cereals, Cider

Pasta, Cereals, Cider

Dataset

1

Groceries_dataset.csv

Member_number	Date	itemDescription	Quantity
1808	21/7/2015	tropical fruit	6
2552	5/1/2015	whole milk	6
2300	19/9/2015	pip fruit	8
1187	12/12/2015	other vegetables	6
3037	1/2/2015	whole milk	6
4941	14/2/2015	rolls/buns	2
4501	8/5/2015	other vegetables	6
3803	23/12/2015	pot plants	6
2762	20/3/2015	whole milk	6

Dataset

This dataset contains about 30000 rows of data.

Each row refers to 1 item that member purchases

Dataset

1

Groceries_dataset.csv

Member_number	Date	itemDescription	Quantity
1808	21/7/2015	tropical fruit	6
2552	5/1/2015	whole milk	6
2300	19/9/2015	pip fruit	8
1187	12/12/2015	other vegetables	6
3037	1/2/2015	whole milk	6
4941	14/2/2015	rolls/buns	2
4501	8/5/2015	other vegetables	6
3803	23/12/2015	pot plants	6
2762	20/3/2015	whole milk	6

This dataset contains about 30000 rows of data.

Each row refers to 1 item that member purchases

2

delimiterData.csv

InvoiceNo	Item	Quantity
53617	fresh bread	12
53617	cider	2
53617	mayonaise	12
53617	pasta	1
53617	eggs	3
53617	olive oil	4
53618	cereals	6
53618	cider	2
53618	mayonaise	72
53618	pasta	50

This dataset contains about 3000 rows of data.

Each row refers to 1 item in each invoice

IMPLEMENTATION

Market Basket Analysis

FINDING

1. Groceries_dataset.csv

FINDING

1 Groceries_dataset.csv

- **Support Level** = 10%

  The frequency of the data is set around 10% as this is real-world data and the frequency of data is not high.

- **Confidence Level** = 40%

  The combination is about 40% sure that this can be trusted.

- **Lift** > 1

  The lift level = 1 means that data is independent and not interesting in analysis, so we concern about lift > 1 only.

FINDING

1 Groceries_dataset.csv

```
rules[ (rules['lift'] >= 1) &
       (rules['confidence'] >= 0.4) ]
```

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
3	(other vegetables)	(whole milk)	0.348309	0.420604	0.161890	0.464789	1.105050	0.015390	1.082555
4	(rolls/buns)	(whole milk)	0.282727	0.420604	0.132714	0.469406	1.116029	0.013798	1.091977
6	(root vegetables)	(whole milk)	0.216628	0.420604	0.100439	0.463647	1.102336	0.009324	1.080251
8	(soda)	(whole milk)	0.235993	0.420604	0.109476	0.463895	1.102925	0.010216	1.080751
10	(tropical fruit)	(whole milk)	0.229796	0.420604	0.103796	0.451685	1.073897	0.007142	1.056685
12	(yogurt)	(whole milk)	0.230829	0.420604	0.111541	0.483221	1.148875	0.014454	1.121169

FINDING

1 **Groceries_dataset.csv**

```
rules[ (rules['lift'] >= 1) &
       (rules['confidence'] >= 0.4) ]
```

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
3	(other vegetables)	(whole milk)	0.348309	0.420604	0.161890	0.464789	1.105050	0.015390	1.082555
4	(rolls/buns)	(whole milk)	0.282727	0.420604	0.132714	0.469406	1.116029	0.013798	1.091977
6	(root vegetables)	(whole milk)	0.216628	0.420604	0.100439	0.463647	1.102336	0.009324	1.080251
8	(soda)	(whole milk)	0.235993	0.420604	0.109476	0.463895	1.102925	0.010216	1.080751
10	(tropical fruit)	(whole milk)	0.229796	0.420604	0.103796	0.451685	1.073897	0.007142	1.056685
12	(yogurt)	(whole milk)	0.230829	0.420604	0.111541	0.483221	1.148875	0.014454	1.121169

FINDING

1 Groceries_dataset.csv

```
rules[ (rules['lift'] >= 1) &
       (rules['confidence'] >= 0.4) ]
```

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
3	(other vegetables)	(whole milk)	0.348309	0.420604	0.161890	0.464789	1.105050	0.015390	1.082555
4	(rolls/buns)	(whole milk)	0.282727	0.420604	0.132714	0.469406	1.116029	0.013798	1.091977
6	(root vegetables)	(whole milk)	0.216628	0.420604	0.100439	0.463647	1.102336	0.009324	1.080251
8	(soda)	(whole milk)	0.235993	0.420604	0.109476	0.463895	1.102925	0.010216	1.080751
10	(tropical fruit)	(whole milk)	0.229796	0.420604	0.103796	0.451685	1.073897	0.007142	1.056685
12	(yogurt)	(whole milk)	0.230829	0.420604	0.111541	0.483221	1.148875	0.014454	1.121169

- **Combine** these items into package for more chance to sell
- Put promotion on **“whole milk”** as it is usually the consequent product

FINDING

2 DelimiterData.csv

FINDING

2 DelimiterData.csv

- **Support Level** = 7%

  The frequency of the data is set around 7% as this is real-world data and the frequency of data is not high.

- **Confidence Level** = 30%

  The combination is about 30% sure that this can be trusted.

- **Lift** > 1

  The lift level = 1 means that data is independent and not interesting in analysis, so we concern about lift > 1 only.

FINDING

2 DelimiterData.csv

```
[27] rules[ (rules['lift'] >= 1) &
            (rules['confidence'] >= 0.3) ]
```

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
0	(eggs)	(spaghetti)	0.213656	0.205213	0.071219	0.333333	1.624329	0.027374	1.192181
1	(spaghetti)	(eggs)	0.205213	0.213656	0.071219	0.347048	1.624329	0.027374	1.204291

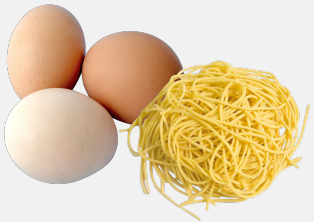
FINDING

2 DelimiterData.csv

```
[27] rules[ (rules['lift'] >= 1) &
            (rules['confidence'] >= 0.3) ]
```

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
0	(eggs)	(spaghetti)	0.213656	0.205213	0.071219	0.333333	1.624329	0.027374	1.192181
1	(spaghetti)	(eggs)	0.205213	0.213656	0.071219	0.347048	1.624329	0.027374	1.204291

- **"Eggs"** and **"Spaghetti"** should be put near each other,



6

Summary

FINAL RESULT

Avengers Project

Special Thanks to

Amazing Professor Peerapon Vateekul, Ph.D.

Super TA Supakrit Paoliwat

REFERENCES

Types of Machine Learning

- http://codeonthehill.com/machine-learning-types/
- https://www.edureka.co/blog/supervised-learning/

Azure ML Implementation

- https://docs.microsoft.com/en-us/azure/machine-learning/overview-what-is-machine-learning-studio

K-Means Clustering - Information

- https://en.wikipedia.org/wiki/Centroid
- https://sebastianraschka.com/faq/docs/euclidean-distance.html

K-Means Clustering - Data set

- https://www.kaggle.com/rodsaldanha/arketing-campaign
- https://www.kaggle.com/amark720/retail-shop-case-study-dataset?select=Transactions.csv

K-Means Clustering - Tutorials

- https://medium.com/@ruslanmv/train-a-simple-clustering-model-in-azure-d79ba64b49aa
- https://www.youtube.com/watch?v=1zTt8QVIFOI

Market Basket Analysis - Information

- https://smartbridge.com/market-basket-analysis-101/
- https://medium.com/tni-university/association-rules-a36515861b6e

Market Basket Analysis - Data set

- https://www.kaggle.com/heeraldedhia/groceries-dataset
- https://www.kaggle.com/betdev/retail-sales-transactions

Market Basket Analysis - Tutorials

- https://pbpython.com/market-basket-analysis.html

- Azure ML
- Supervised and Unsupervised and Reinforcement Learning
- K-Means Clustering Model
    - Why
    - Dataset
    - Algorithm
    - Implementation
    - Findings
- Market Basket Analysis or association analysis
    - Why
    - Dataset
    - Algorithm
    - Implementation
    - Findings
- Summary?
    - What we have done
- (Thank you Page Eiei)

Algorithm

K-Means Clustering

1 Collect Customer Profile

Since we want to perform RFM analysis on the customer data, the customer transactions dataset is used since it contains relevant information that are useful for the analysis.

2 Select Specific Fields

Recency: Amount of time since the last transaction of a particular customer
Frequency: Total number of transactions a particular customer made
Monetary: Total spendings a particular customer made

Chosen fields in the dataset to generate Recency, Frequency, and Monetary attributes includes:

1. trans_date
2. cust_id
3. total_amt

Algorithm

Monetary

Calculate the total spending of each customer

```
rfm_m = retail.groupby('cust_id')['total_amt'].sum()
rfm_m = rfm_m.reset_index()
rfm_m.head()
```

Result

	cust_id	total_amt
0	266783	3113.890
1	266784	5694.065
2	266785	21613.800
3	266788	6092.970
4	266794	27981.915

Algorithm

Frequency

Calculate the number of transactions of each customer

```
rfm_f = retail.groupby('cust_id')['transaction_id'].count()
rfm_f = rfm_f.reset_index()
rfm_f.columns = ['cust_id', 'frequency']
rfm_f.head()
```

Result

	cust_id	frequency
0	266783	5
1	266784	3
2	266785	8
3	266788	4
4	266794	12

Algorithm

Recency

Convert tran_date to DateTime Format

```
retail['tran_date'] = pd.to_datetime(retail['tran_date'], format='%d/%m/%Y')
```

Finding the last date of the recorded transactions

```
max_date = max(retail['tran_date'])
max_date
```

```
Timestamp('2014-02-28 00:00:00')
```

Compute the difference between the max_date and transaction date

```
retail['Diff'] = max_date - retail['tran_date']
```

Algorithm

Recency

Compute the minimum difference difference between the max_date and transaction date to get the recency value of each customer

```
rfm_p = retail.groupby('cust_id')['Diff'].min()
rfm_p = rfm_p.reset_index()
rfm_p.head()
```

Result

	cust_id	Diff
0	266783	373
1	266784	451
2	266785	211
3	266788	381
4	266794	16

Algorithm

Combing all the Values

```
rfm = rfm.rename(columns={'cust_id': 'CustomerID', 'total_amt': 'TotalAmount', 'frequency':'Frequency', 'recency':'Recency'})
rfm
```

	CustomerID	TotalAmount	Frequency	Recency
0	266783	3113.890	5	373
1	266784	5694.065	3	451
2	266785	21613.800	8	211
3	266788	6092.970	4	381
4	266794	27981.915	12	16
...	...	...	...	...
5513	545168	-2581.280	1	481
5514	546212	-1204.450	1	641
5515	547874	-5337.150	1	530
5516	549312	-7315.100	1	224
5517	549496	-8475.350	1	802

Algorithm

Removing Outliers

```
# Removing (statistical) outliers for Total Amount
Q1 = rfm.TotalAmount.quantile(0.05)
Q3 = rfm.TotalAmount.quantile(0.95)
IQR = Q3 - Q1
rfm = rfm[(rfm.TotalAmount >= Q1 - 1.5*IQR) & (rfm.TotalAmount <= Q3 + 1.5*IQR)]

# Removing (statistical) outliers for Recency
Q1 = rfm.Recency.quantile(0.05)
Q3 = rfm.Recency.quantile(0.95)
IQR = Q3 - Q1
rfm = rfm[(rfm.Recency >= Q1 - 1.5*IQR) & (rfm.Recency <= Q3 + 1.5*IQR)]

# Removing (statistical) outliers for Frequency
Q1 = rfm.Frequency.quantile(0.05)
Q3 = rfm.Frequency.quantile(0.95)
IQR = Q3 - Q1
rfm = rfm[(rfm.Frequency >= Q1 - 1.5*IQR) & (rfm.Frequency <= Q3 + 1.5*IQR)]
```

Algorithm

Rescaling the Attributes

```
rfm_df = rfm[['TotalAmount', 'Frequency', 'Recency']]

# Instantiate
scaler = StandardScaler()

# fit_transform (mean = 0 , sd = 1)
rfm_df_scaled = scaler.fit_transform(rfm_df)
rfm_df_scaled.shape
```

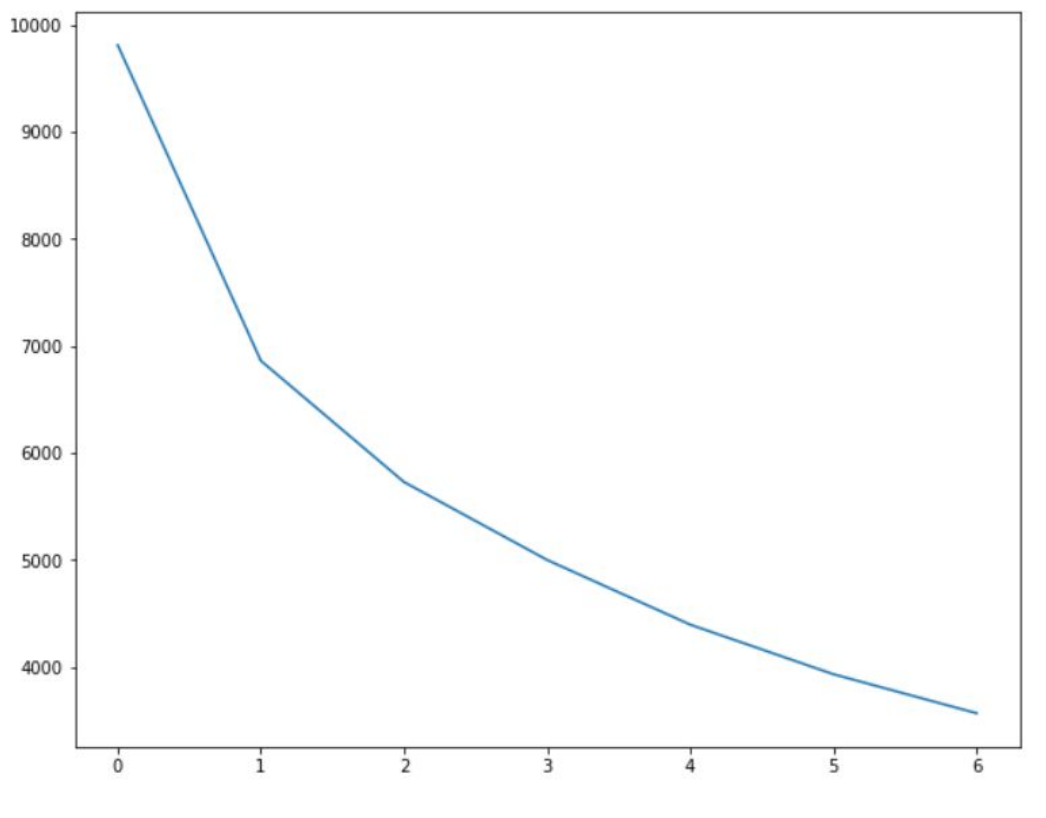
```
rfm_df_scaled = pd.DataFrame(rfm_df_scaled)
rfm_df_scaled.columns = ['TotalAmount', 'Frequency', 'Recency']
rfm_df_scaled.head()
```

	TotalAmount	Frequency	Recency
0	-0.963110	0.389941	0.369498
1	-0.526379	-0.555859	0.689210
2	2.168261	1.808642	-0.294518
3	-0.458859	-0.082959	0.402289
4	3.246155	3.700242	-1.093798

Algorithm

Finding the Optimal Number of Clusters

Elbow Curve



Algorithm

Finding the Optimal Number of Clusters

Silhouette Score

```
range_n_clusters = [2, 3, 4, 5, 6, 7, 8]

for num_clusters in range_n_clusters:

    # intialise kmeans
    kmeans = KMeans(n_clusters=num_clusters, max_iter=50)
    kmeans.fit(rfm_df_scaled)

    cluster_labels = kmeans.labels_

    # silhouette score
    silhouette_avg = silhouette_score(rfm_df_scaled, cluster_labels)
    print("For n_clusters={0}, the silhouette score is {1}".format(num_clusters, silhouette_avg))
```

```
For n_clusters=2, the silhouette score is 0.34668345289624836
For n_clusters=3, the silhouette score is 0.340611659599919
For n_clusters=4, the silhouette score is 0.28323819362116126
For n_clusters=5, the silhouette score is 0.28851115547351375
For n_clusters=6, the silhouette score is 0.29012630060006483
For n_clusters=7, the silhouette score is 0.2695550286685653
For n_clusters=8, the silhouette score is 0.26390137019471416
```

Algorithm

Chosen Model (3 Clusters)

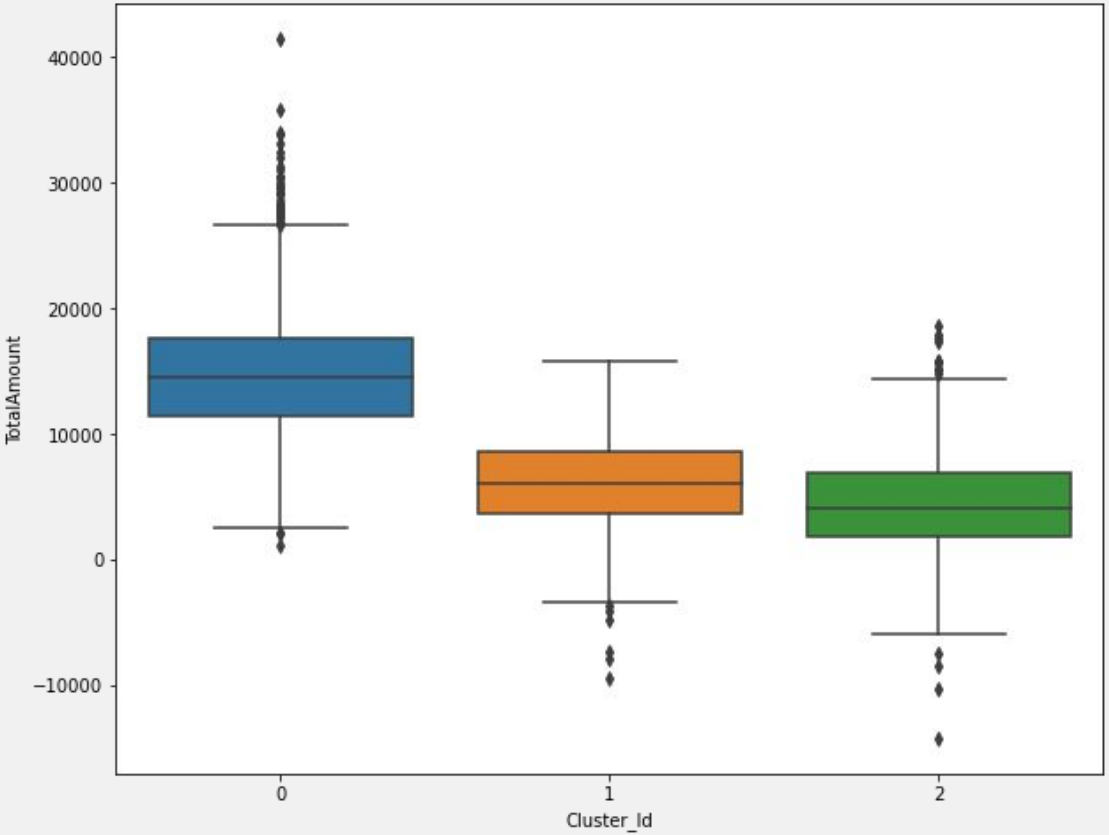
```
kmeans = KMeans(n_clusters=3, max_iter=50)
kmeans.fit(rfm_df_scaled)
```

Assigning Cluster_ID

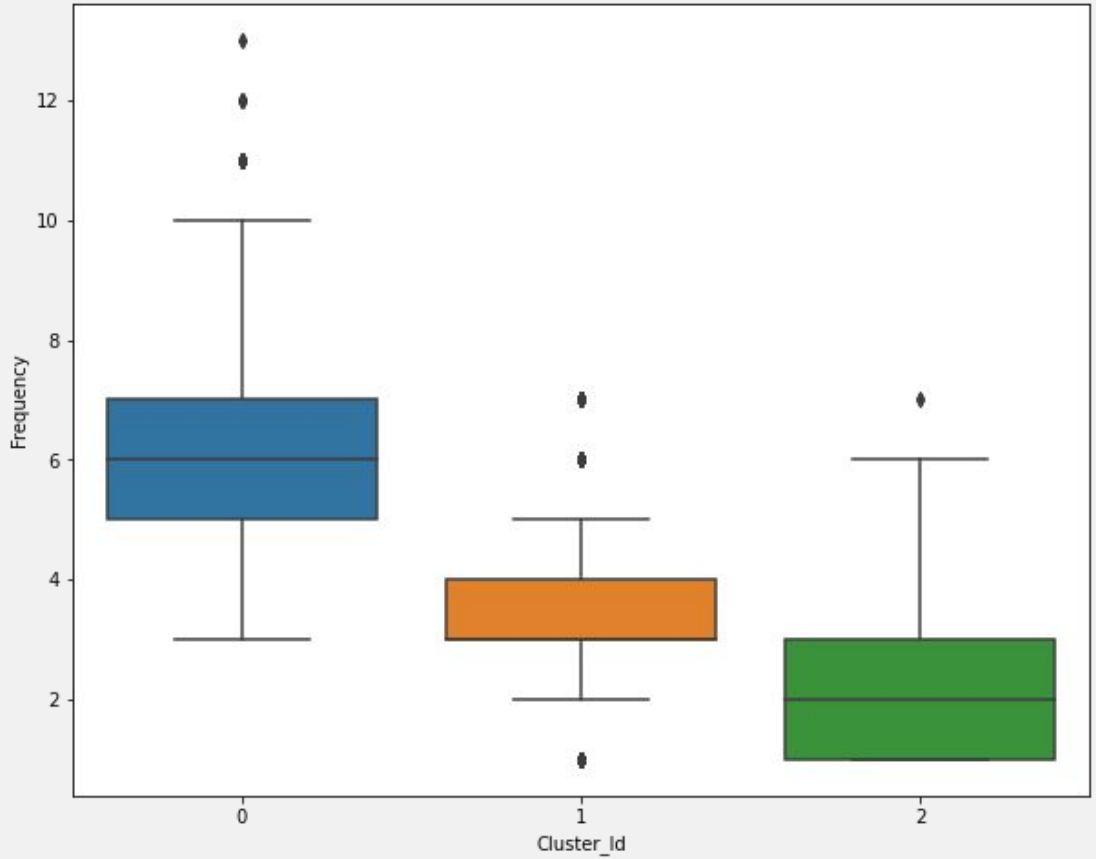
```
rfm['Cluster_Id'] = kmeans.labels_
rfm.head()
```

	CustomerID	TotalAmount	Frequency	Recency	Cluster_Id
0	266783	3113.890	5	373	1
1	266784	5694.065	3	451	2
2	266785	21613.800	8	211	0
3	266788	6092.970	4	381	1
4	266794	27981.915	12	16	0

Total Amount vs Cluster_ID



Frequency vs Cluster_ID



Recency vs Cluster_ID