***CUSTOMER SEGMENTATION***

***USING***

***K-means CLUSTERING ALGORITHM (KNN)***

***IN MACHINE LEARNING.***

##### **A PROJECT REPORT**

***Submitted by***

***Manish N P***

#### 

#### 

**Index**

|  |  |  |
| --- | --- | --- |
| **S.no** | **Title** | **Page no** |
| **1.** | **Introduction** | 6 |
| 1.1 | Customer Segmentation | 6 |
| 1.2 | The Business Problem | 6 |
| 1.3 | Data | 7 |
| 1.4 | Scope of Analysis | 9 |
| **2.** | **Segmentation Analysis** | 10 |
| 2.1 | Brief Introduction | 10 |
| 2.2 | RFM | 12 |
| 2.3 | Product Recommendation | 15 |
| **3.** | **Modelling** | 17 |
| 3.1 | Data Processing | 18 |
| 3.2 | K means Clustering | 19 |
| 3.3 | Customer behaviour prediction | 19 |
| 3.4 | Product Recommendation | 19 |
| **4.** | **Challenges of Performing analysis** | 20 |
| 4.1 | Clustering algorithm | 21 |
| 4.2 | Clustering(Other ML methods) | 21 |
| 4.3 | Similarity Measure | 21 |
| **5.** | **K means Clustering** | 24 |

|  |  |  |
| --- | --- | --- |
| **S.no** | **TITLE** | **Page no** |
| 5.1 | Centroid based | 24 |
| 5.2 | Data Preparation | 35 |
| 5.3 | Feature Engineering | 36 |
| **6.** | **Marketers** | 38 |
| **7.** | **RFM Analysis** | 39 |
| 7.1 | RFM Example | 40 |
| 7.2 | RFM Score | 42 |
| 7.3 | RFM Segmentation | 44 |
| **8.** | **Proposed Model -1** | 46 |
| **9.** | **Proposed Model -2** | 48 |
| **10.** | **Application of RFM Model** | 50 |
| 10.1 | Definition and Scoring Scheme RFM model | 50 |
| 10.2 | Comparing RFM with other variables on other models | 52 |
| 10.3 | Criteria of Clustering | 52 |
| 10.4 | Performing Analysis and Results | 54 |
| **11.** | **Brief Overview of Code** | 55 |
| 11.1 | Clustering Results | 55 |
| 11.2 | Managerial Implication of Results | 58 |
| **12.** | **Work,Conclusion and Reference** | 60 |

**Abstract**

As the legal cannabis industry emerges from its nascent stages, there is increasing motivation for retailers to look for data or strategies that can help them segment or describe their customers in a succinct, but informative manner. While many cannabis operators view the state-mandated traceability as a necessary burden, it provides a goldmine for internal customer analysis. Traditionally, segmentation analysis focuses on demographic or RFM (recency- frequency-monetary) segmentation. Yet, neither of these methods has the capacity to provide insight into a customer’s purchasing behavior. With the help of 4Front Ventures, a battle-tested multinational cannabis operator, this report focuses on segmenting customers using cannabis-specific data (such as flower and concentrate consumption) and machine learning methods (K-Means and Agglomerative Hierarchical Clustering) to generate newfound ways to explore a dispensary’s consumer base. The findings are that there are roughly five or six clusters of customers with each cluster having unique purchasing traits that define them. Although the results are meaningful, this report could benefit with exploring more clustering algorithms, comparing results across dispensaries within the same state, or investigating segmentations in other state markets

**Chapter -1 Introduction**

## Customer Segmentation

An important marketing strategy that is widely used by businesses is customer segmentation. As previously stated, the point of customer segmentation is to split the user-base into smaller groups that can be targeted with specialized content and offers.

The produced customer groups are drawn from user behaviour data which gives the business a deeper understanding of the types of users that exist in the system. The benefit of customer segmentation is twofold. Firstly, a better knowledge about the types of users in a system can lead to better business and marketing strategies. Secondly, a user is likely to use an application more often if he/she always receives relevant content. Another essential point is that if a customer is pleased, he/she is more likely to recommend the application to other people which helps in the expansion of a company.

**The Business Problem**

Any company in retail, no matter the industry, ends up collecting, creating, and manipulating data over the course of their lifespan. These data are produced and recorded in a variety of contexts, most notably in the form of shipments, tickets, employee logs, and digital interactions. Each of these in- stances of data describes a small piece of how the company operates, for better or for worse. The more access to data that one has, the better the picture that the data can delineate. With a clear picture made from data, details previously unseen begin to emerge that spur new insights and innovations.The sheer size and complicated nature of data in the real world make the above task much easier said than done, though. The rise of performance metrics and interactive dashboards have ushered in a new era of looking at data. Many times, the data included in dashboards are at the superficial level: How much did store X make during December?, What are our top 5 products?, What is our monthly COGS (Cost of Goods Sold)?

While dashboards supply data that often have important significance in supply chain management and operations, they are limited in the sense that they omit data and insights that require higher levels of data mining and analysis.

Companies that utilize proper data science and data mining practices allow themselves to dig further into their own operating strategies, which in turn allows them to optimize their commercial practices. As a result, there are increasing motivations for investigating phenomena and data that cannot be simply answered: Why is product B purchased more on the first Saturday of every month compared to other weekends?, If a customer bought product B, will they like product C?, What are the defining traits of our customers? Can we predict what customers will want to buy?

is the latter half of the last question that will be the broad focus of this paper.

In particular, this paper discusses the results of a customer segmentation analysis project done in conjunction with Front Ventures. Front Ventures (hereby referred to as Front) is a consulting and management firm in the legal cannabis industry that operates various cultivation, production, and retail sites across the country. Not just the products they like to purchase, but when they like to purchase the While some of these questions are more straightforward than others, it is clear that they all require data munging, analysis, and presentation that involve skills and techniques beyond what is required of a traditional analyst. By integrating machine learning practices and conventional business un- derstandings, the paths to answering these questions became more intertwined with that of a similar question.

## Data

The available data is the most vital part of any clustering algorithm. The most important aspects are the quality and amount of the available data. In order to run some sort of similarity function to cluster items or users in a system, the data needs to be arranged into feature vectors with a set of feature values. To achieve the best results, a large amount of data is needed and more importantly the absence of data points needs to be minimal. The amount of data in most cases is not a problem nowadays since companies store all kinds of user and item data in large databases.

Another important aspect in customer segmentation is to understand the available data. In a system where items are rated using some sort of scale, e.g. a rating from zero to five, it is fairly easy to interpret a user’s preferences

**Acquisition of Data**

Finding readied, usable data for analysis in a business context is a rarity. As such, it is imperative to collect as much data as possible, but also in a format that meets a wide variety of financial, ethical, and computational considerations. But before discussing these, it is first important to describe the ways in which the relevant retail data are stored and utilized across the company.

Without delving into confidential details, the broad idea is that the vast majority of retail data are stored in various SQL databases. Because of em- phasis on seed-to-sale traceability, various state regulations, and lack of com- petition in the software market, most businesses are required to integrate their entire business up to one point of sale (POS) system that is consistent across the company. If the company is vertically integrated, the POS extends to their cultivation and production software. Some software providers, such as Bio- Track, Greenbits, Viridian, have flourished in the industry by providing fully integrated software known as seed-to-sale systems. In the backend, servers store their data in SQL databases built to comply with state regulations and standards. On the front end, they deliver relevant data or insight via inter- active dashboards, reporting modules, or simple visuals to retail managers or analysts.

As a direct consequence of 4Front’s successful expansion into new and de- veloping markets, they have incurred unforeseen challenges with data handling and storage. Even though the tactics and strategies that 4Front uses to sell and market their products are, for the most part, consistent across state mar- kets, their data storage and data accessibility is contingent upon their markets and access to third-party software. Certain software, while allowing for nice.

reports and key visuals, do not have any built-in backend functionality for retailers to access the raw data. Luckily, one of the software used in a couple of 4Front’s operating markets allows for a backend SQL editor that allows for direct queries, though there is very limited documentation on the database structure provided by the software company. Nonetheless, it is possible for customer data to be collected for customers who exist in the state markets with the appropriate seed-to-sale software.

However, just having access to the data/knowing where it is is a small step in the overall data gathering process. Roughly speaking, it is possible to classify the various data acquisition processes into three distinct categories.

First, it was necessary to establish any ethical considerations or constraints to the usage of data. When first-time customers enter a dispensary, they are presented with a form that asks for verifiable demographic information such as their name, age, and address. In addition, they are also asked if they con- sent to the company using their data for analysis and marketing purposes. Each customer’s answer to the previous question is one-hot encoded into the database: 0 for “no'' and 1 for “yes”. The customers included in this anal- ysis, thus, are only the customers who answered “yes” to the question and have a 1 for the value for the appropriate feature. Furthermore, to protect the anonymity of each of the customers, it is also necessary to prune away all sensitive information from each customer. In other words, the only demo- graphic/sensitive information of each customer that the analysis will use is the age of the customer. The sex, address, name, and other sensitive or personal information is detached from the customer during analysis. Each customer is uniquely identified with an ID that allows for consistent analysis, but the IDs are generated internally, which means that the customer has no knowledge of their ID. Essentially, while there is a way for the program to keep track of a particular customer’s purchases, it is not possible for the program to include customers who do not consent to using their data for this purpose, or for the program to tie the purchases to a particular name or address.

Second, collecting the data in an efficient manner heavily relies on a strong understanding of the structure of the database. Without revealing too many details, there were four important datatables in the database that contained relevant information.

The *customers* table includes the customer id, number of visits, total amount spent, whether or not they consent to us using their data, and age of each customer.

These data are needed for identifying unique customers and also providing the beginnings of some of the data used in clustering.

The *tickets* table contained all information regarding tickets , such as the ticket ID, time of transaction, total amount spent, the customer ID involved in the ticket, and which employee completed the ticket. The time of transaction, total amount spent, and associated customer ID are relevant for this particular analysis

The *sales* table hosts data related to each individual sale (i.e each indi- vidual product sold). This consists of a sales ID, the ticket ID that the sale is associated to, the price of the sale (price of the item), and the product ID associated with the sale.

This table contains many IDs and other data that intersect with other tables that are important for this analysis. From this table, it is possible to gather almost all the relevant data for each ticket/customer.

The *products* table includes the necessary information about each prod- uct the store has, such as its ID, when it was added to the system, and which product category [5](#_48pi1tg) it belongs to. This table is mostly used for debugging purposes and for providing some context that makes it easier to identify and classify products.

Lastly, there were certain computational considerations to take into account when collecting data as well.

Though the database is set up to handle missing values already, there were several columns in several tables that had malformed or missing values that required additional attention.

Incorrect self- reported dates, voided tickets, and tickets with $0 in sales needed to be pruned from the dataset. In addition, any relevant field with a missing or negative value needed to be pruned or corrected from the dataset. Though the num- ber of affected instances is small, it was crucial to handle these malformed instances because they prevented smooth analysis later on.

**Scope of Analysis**

In general, the methods used to gather the data for this project can easily be extended into other relevant contexts/analyses. While there is clear value in using the same data to investigate purchasing patterns or to build an item- based collaborative filtering recommender system, neither of these is the focus for this paper. The scope of the paper is limited to the following four inter- twined goals:

To cluster customers based on common purchasing behaviors for future operations/marketing projects

To incorporate best mathematical, visual, programming, and business practices into a thoughtful analysis that is understood across a variety of contexts and disciplines

To investigate how similar data and algorithms could be used in future data mining projects

To create an understanding and inspiration of how data science can be used to solve real-world problems

Before delving into the details of the project and its implications, the next chapter discusses what customer segmentation analysis actually is and the reasons for its importance.

**Chapter-2 Segmentation Analysis**

**Brief Introduction**

For a retailer, understanding the components of their consumer base is key to maximizing their potential in a market; the retailer that attracts the most customers will acquire the most market share, ceteris paribus. In fact, the high costs of gaining a new customer or getting back an old customer force retailers to seriously consider how to allocate resources to optimize not just volume of customers, but the retention of them as well. Additionally, it is a common understanding in the retail industry that the Pareto Principle—more likely than not—applies to the company: 80% of profits come from 20% of the customer's . One crucial reason why this principle holds is because retail businesses thrive on repeat purchases[9](#_3mzq4wv). As a consequence, a net change of one customer can significantly impact a business’ profit in the long run. Therefore, it is generally in the best interest of the retailer to devote efforts to retaining customers by understanding them on as deep of a level as necessary.

However, examining the intricate, rich relationships between a retailer and their consumer base involves understanding how different components of the base behave. Namely, how different segments of customers act similarly or differently from other segments . One method of approaching customer un- derstanding is through the lens of customer segmentation. In short, customer segmentation analysis is the process of grouping customers in such a way that customers within one particular group are similar to each other but different from customers in other groups. In general, there are two paths of segmenta- tion: a priori and post hoc. A priori analysis involves creating the segments beforehand and then, after examining data, placing each customer within the segments[11](#_haapch). Rather than having the customer data dictate the types of seg- ments formed, certain outside knowledge or structure would dictate the pre- ferred segmentations. As such, the key unit of analysis here are the created segments, not necessarily the customers themselves.

On the other hand, post hoc analysis leverages the data to form the seg- ments, rather than the other way around. In a sense, post hoc analysis is a direct consequence of advancements in data collection and reliability whereas a prior analysis rose to prominence several years before such beneficial advace- ments. Regardless of the context, advancing technology has opened doors for post hoc analysis to succeed as a segmentation method in the retail industry. So, modern retailers and data scientists tend to perform customer segmenta- tion using techniques residing under the post hoc umbrella, which will be the focus of the remainder of the paper.

While the goal of customer segmentation analysis has been consistent among retailers for many years, approaches in the past relied on much weaker analytical techniques than available today. It is nonsensical to blame com- panies in the past who failed to utilize their data properly; the technology and data infrastructure simply were not ubiquitous or cheap enough to al- low for companies to collect massive amounts of data as they do today. Yet, many companies still found rudimentary methods to attempt to understand their customers, the most traditional involving purely demographic analysis.

Standard implementation of RFM model is cheap and simple: once each of the components are defined in a way that makes them easy to collect, it is a relatively menial task for a retailer to visual- ize the results, which makes interpretation easy as well. Usually, the results of an RFM analysis would include three plots—one for each combination of two variables (e.g. Recency and Frequency)—with the inferred segments and their defining characteristics. RFM analysis became a staple of modern marketing for its simplicity and its cheap cost to implement as well to communicate ef- ficiently[16](#_2fk6b3p). In a way, the visualization aspect alone gave utility to the RFM model, allowing managers to effectively glean insights from the analysis.

Yet, as the retail industry evolved in parallel with the technology boom, it became dramatically easier for retailers to collect data at a larger scale, which also meant it became easier to mine at a larger scale. In the case of the cannabis industry, the mandate that each operator must have a secure and sound traceability system allows operators— who know how to access their data— virtually unlimited potential in performing higher level analysis. While RFM modelling is based on only three features, modern customer segmentation can involve several hundred or even several thousand features. As a result, the segments of the analysis become much finer, much richer to allow retailers to understand their customers at levels simply unattainable from RFM or demographic analysis [17](#_upglbi).

One of the more popular ways retailers have been able to acquire such specific data regarding their customers is through a loyalty program [18](#_3ep43zb). In a loyalty program, the customer benefits by receiving certain discounts, but the natural by-product[19](#_1tuee74) of the loyalty card is the data that the retailer can mine to better serve their customers and boost profits [20](#_4du1wux). By using this data, retailers can create specific marketing campaigns, target certain customer seg- ments with uniquely tailored discounts, or even invite old customers back into the store. This data allows for retailers to conduct ultra-specific marketing strategies that have transformed the way retailers compete in the age of Big Data.

In order to perform customer segmentation analysis at a high level, retailers have begun to incorporate aspects of machine learning into the analysis of their customers. More specifically, retailers are utilizing unsupervised machine learning tools such as clustering and dimensionality reduction to approaches.analysis in ways that cannot be matched without machine learning. Instead of focusing on only a few features or customers at a time, it is possible to write programs and implement algorithms that can take into account several more features or several more instances than traditional spreadsheets can hold or process. Because of this massive potential, retailers across all industries are attempting to leverage clustering algorithms such as K-Means or hierarchical clustering to more accurately and quickly segment their customers. The faster and better retailers are able to cluster their customers, the quicker they can market to them and thus acquire market share

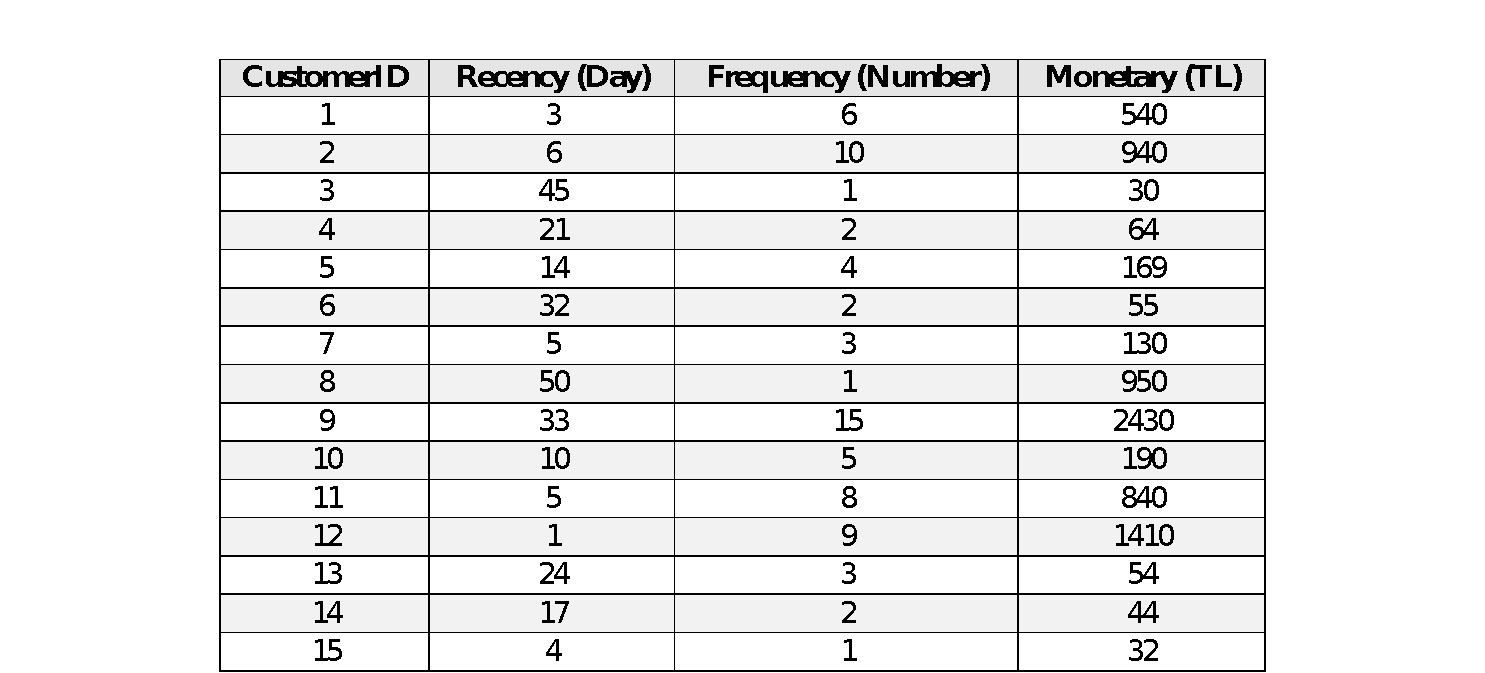
.

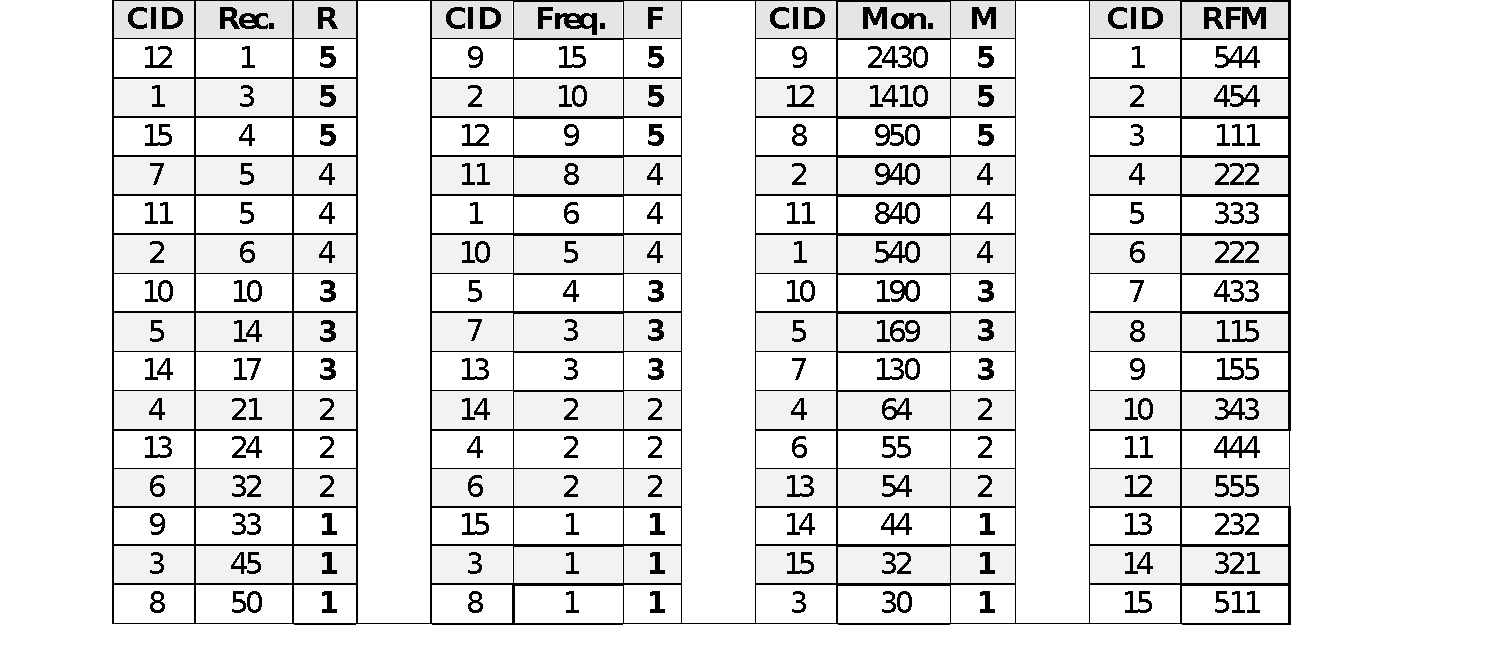
**RFM**

RFM stands for Recency, Frequency and Monetary value. RFM analysis is a marketing technique used for analyzing customer behavior such as how recently a customer has purchased (recency), how often the customer purchases (frequency), and how much the customer spends (monetary). It is a useful method to improve customer segmentation by dividing customers into various groups for future personalization services and to identify customers who are more likely to respond to promotions. In recent years, data mining applications based on RFM concepts have also been proposed for different areas such as for the computer security (Kim et al., 2010), for automobile industry (Chan, 2008) and for the electronics industry (Chiu et al., 2009). Research cases of data mining with RFM variables include different data mining techniques such as neural network and decision tree (Olson et al., 2009), rough set theory (Cheng & Chen, 2009), self organizing map (Li et al., 2008), CHAID (McCarty and Hastak, 2007), genetic algorithm (Chan, 2008) and sequential pattern mining (Chen et al., 2009; Liu et al., 2009). Integration of RFM analysis and data mining techniques provides useful information for current and new customers. Clustering based on RFM attributes provides more behavioral knowledge of customers’ actual marketing levels than other cluster analyses. Classification rules discovered from customer demographic variables and RFM variables provide useful knowledge for managers to predict future customer behavior such as how recently the customer will probably purchase, how often the customer will purchase, and what will be the value of his/her purchases.

Current RFM values of the customer, potential future customer behavior and products frequently purchased together. To the best of our knowledge, this chapter is the first in applying the RFM criterion in three data mining tasks, applied one after another, using customer demographic data, customer transaction data, and product properties. Experiments, which were carried out using the datasets collected by a sports store in Turkey through its e-commerce website, empirically demonstrate the benefits of using our model in direct marketing.

**RFM analysis**

The concept of RFM was introduced by Bult and Wansbeek (1995) and has proven very effective (Blattberg et al., 2008) when applied to marketing databases. RFM analysis depends on Recency (R), Frequency (F), and Monetary (M) measures which are three important purchase-related variables that influence the future purchase possibilities of the customers. Recency refers to the interval between the time that the latest consuming behavior happens, and present. Many direct marketers believe that most-recent purchasers are more likely to purchase again than less-recent purchasers. Frequency is the number of transactions that a customer has made within a certain period. This measure is used based on the assumption that customers with more purchases are more likely to buy products than customers with fewer purchases. Monetary refers to the cumulative total of money spent by a particular customer. In order to demonstrate RFM analysis, an example dataset (customer transaction data) is given in Table 1. Table 2 shows the steps of RFM analysis, which involves scaling customers based on each RFM factor separately. The segmentation starts with recency, then frequency, and finally monetary value. It begins with sorting customers based on recency, i.e. period since last purchase, in order of lowest to highest (most recent purchasers at the top). The customers are then split into quintiles (five equal groups), and given the top 20% a recency score of 5, the next 20% a score of 4 and so on. Customers are then sorted and scored for frequency – from the most to least frequent, coding the top 20% as 5, and the less frequent quintiles as 4, 3, 2, and 1. This process is then undertaken for monetary as well. Finally, all customers are ranked by concatenating R, F, and M values. This example shows that RFM analysis can be useful even if the database is small of only 15 transactions whereas it would be more powerful when the database grows. RFM analysis assigns value-scores to each customer on the basis of her past behavior. Using the quintile system explained above, at the most, 125 different scores (5x5x5) can be assigned. These cells differ in size from one another. A customer’s score can range from 555 being the highest, to 111 being the lowest. The best customers are in quintile 5 for each factor (555) that have purchased most recently, most frequently and have spent the most money 



RFM provides a simple framework for quantifying customer behavior. For example, it is possible to infer from Table 2 that a customer with id 9, which has RFM score 155, has made a high number of purchases with high monetary values but not for a long time. Something might have gone wrong with this customer

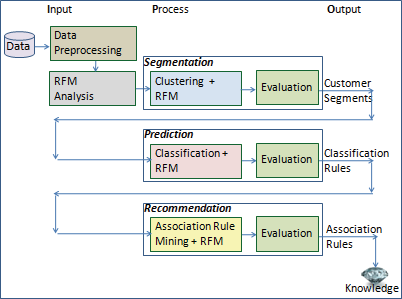
## Classification using RFM

Recently, integration of classification techniques and RFM was studied by Olson et al. (2009) to analyze customers’ response possibilities to a specific product promotion. They compared three data mining techniques: logistic regression, decision trees and neural networks, and discussed the relative tradeoffs among these data mining algorithms in the context of customer segmentation.

Furthermore, in order to evaluate the accuracy rate of the generated classification rules, they compared their approach with different three methods: Decision Tree, Artificial Neural Networks and Naive Bayes. According to the empirical results, their procedure outperforms the other methods listed in terms of accuracy rate.

## Association rule mining using RFM

In data mining, association rules are descriptive patterns of the form XY, where X is termed the left-hand-side, and is the conditional part of an association rule; meanwhile, Y is called the right-hand-side, and is the consequent part. Association rule mining (ARM) is a task for discovering the hidden, interesting association rules between items in the database, having support ≥ minsup threshold. The support of an association rule indicates how frequently that rule occurs in the data. Higher support corresponds to a stronger correlation between the items in the database.



**RFM Analysis**

In this step, RFM analysis is applied by defining the scaling of R–F–M attributes. This process is divided into four parts introduced in the following:

\*Sort the data of three R–F–M attributes by descending or ascending order.

\*Partition the three R–F–M attributes respectively into 5 equal parts and each part is equal to 20% of all. The five parts are assigned 5, 4, 3, 2 and 1 score that refer to the customer contributions. The ‘5’ refers to the most customer contribution, while ‘1’ refers to the least contribution to revenue.

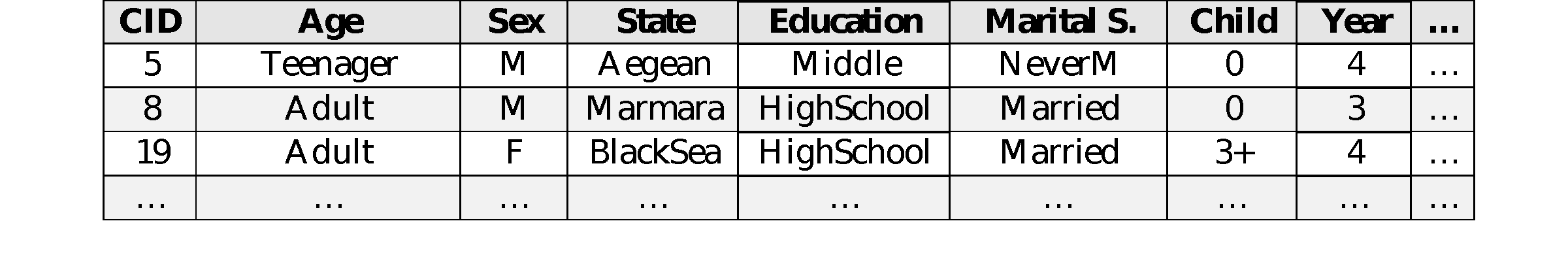
**Product recommendation**

The core concept of this work is to extract recommendation rules from each customer group by considering classification rules and using FP-Growth Algorithm (Han et al., 2000). So, the purpose of this step is to identify the associations between customer segments, customer profiles and product items purchased together. By applying such an algorithm, it is possible to recommend products with associated rankings, which results in better customer satisfaction and cross selling.

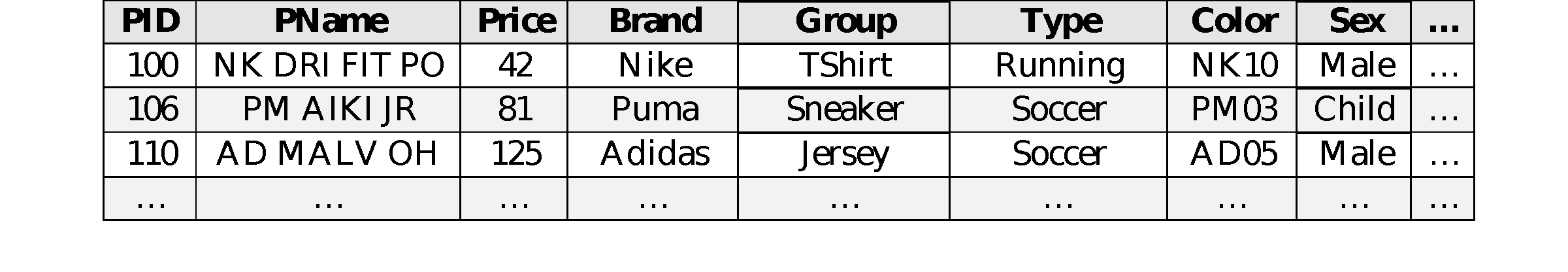
The detail process of this stage is expressed into two sub-steps.

*ARM*: FP-Growth (Frequent Pattern Growth) is one of the Association Rule Mining (ARM) algorithms. Among the other ARM algorithms such as Apriori, Eclat, Mafia, it extracts the rules very fast from data by constructing a prefix tree and traversing this tree to generate rules. The algorithm scans the database two times only. Because of these reasons, FP-Growth algorithm is preferred in this study.

## Customers



**Products**



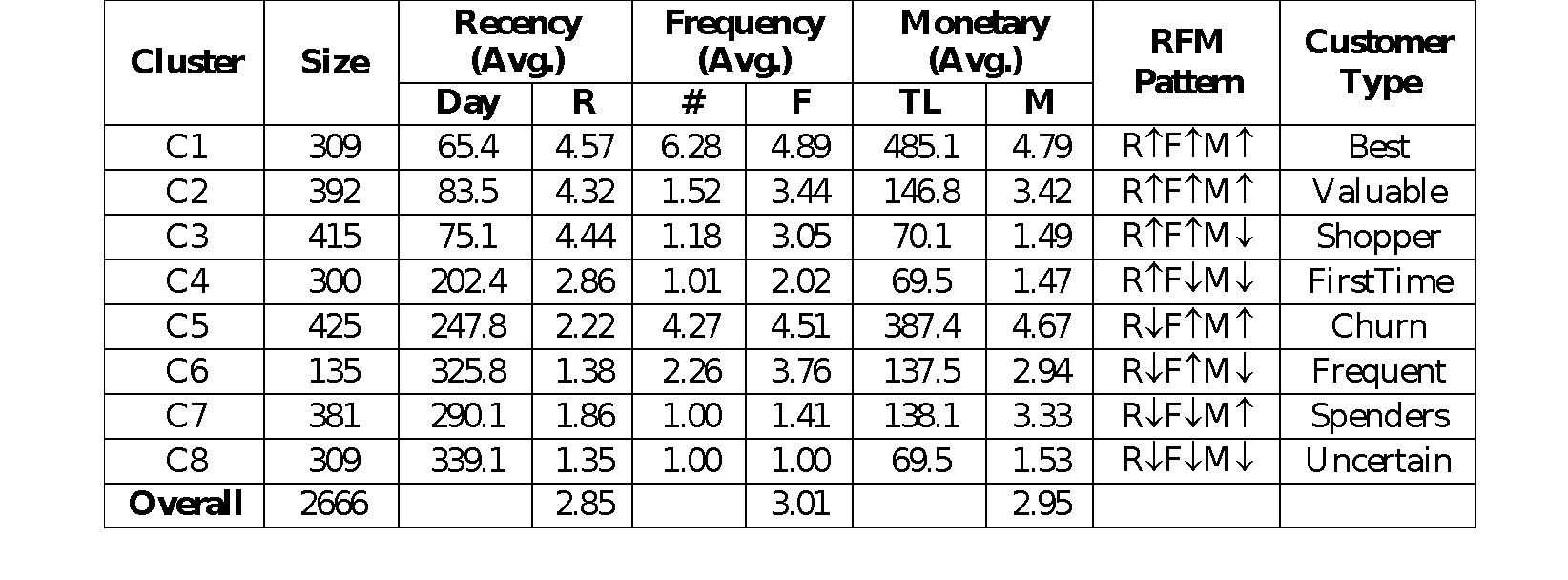
## Orders

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **TID** | **PID** | **CID** | **Date** | **Quantity** | **Discount** | **Total** | **Type** | **…** |
| T1 | 106 | 19 | 2008.12.2 | 1 | 0 | 81 | SS | … |
| T2 | 100 | 8 | 2008.12.2 | 1 | 0 | 42 | YS | … |
| T3 | 110 | 5 | 2008.12.3 | 1 | 0 | 125 | SS | … |
| … | … | … | … | … | … | … | … | … |

## Chapter -3 Modelling

All customers were ranked by considering their recency, frequency and monetary values and they were represented by R-F-M codes. Table 4 shows example R-F-M values of some customers after RFM analysis. For example, it is possible to infer from the first row in Table 4 that a customer with id 5 has R-F-M values 4-3-4 respectively. This customer has made a high number of purchases with high monetary values, not long ago.

Figure 2 shows the distribution of the number of customers with respect to their RFM values. The distribution of RFM values varies within the limits of 0 - 4.6%. At the most, the customers have the RFM value 555 (125 customers), followed by RFM value 113 (108 customers), and next, 107 customers have the RFM value 321. Some RFM values such as 121, 125, 231, 311 etc. were not assigned to any customer



The clusters that have RFM values with at least two upper arrow () can be selected as target ones, all customers who belong to these clusters become candidates for conducting suitable marketing strategies, which attract the most attention.

After customer segmentation, standard deviation and SSE metrics were used to evaluate clustering results. All clusters had a lower standard deviation and SSE values. The result confirmed that these eight clusters were significantly distinguished by recency, frequency, and monetary. Standard deviation values range from 0.67 being the highest, to 0.33 being the lowest. In the experiments, K-Means++ algorithm was run 10 times

Standard Deviation

final result.

0.8

0.7

0.6

0.5

0.4

0.3

0.2

0.1

0.0

C1

C2

C3

C4

Clusters

C5

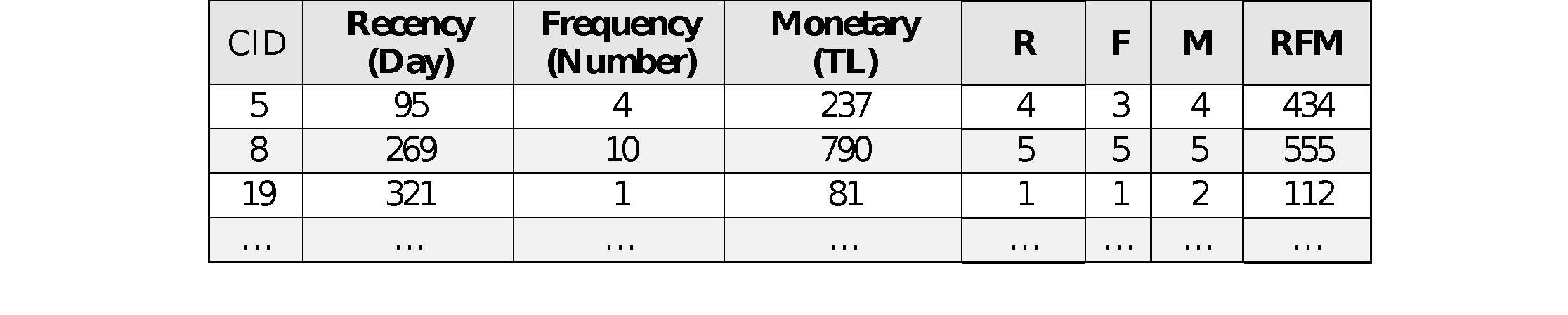
C6

After customers were classified by demographic variables, the recommendation list was generated by feature attributes determined using a classification rule inducer. Parameters were set up to identify association rules that had at least 40% confidence and 2% support imposed on the FP-Growth association rule algorithm. Figure 5 shows a part of association rules, found in the case study

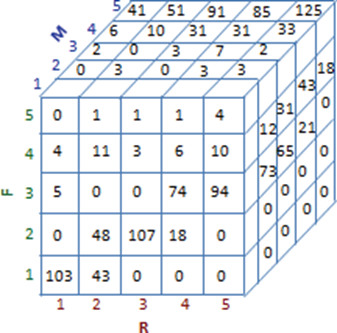
## Data preprocessing

Dataset used in this case study was provided by a sports store in Turkey and collected through its e-commerce website within two years. The complete dataset included 1584 different product demands in 54 sub-groups and 6149 purchase orders of 2666 individual customers. The purchase orders included many columns such as transaction id, product id, customer id, ordering date, quantity, ordering amount (price), sales type, discount and whether or not promotion was involved. While the customer table included demographic variables such as age, gender, marital status, education level and geographic region; product table included attributes such as barcode, brand, color, category, sub- category, usage type and season.

Data preprocessing step handles outliers, fills missing values and makes dimensionality reduction, transformation, concept hierarchy generation, normalization and discretization.



Example R-F-M values of some customers after RFM analysis

****

RFM distribution: 125 possible RFM values and the number of customers

**K-Means clustering**

K-Means clustering was employed to group customers with similar RFM values. Customers were segmented into eight target markets in terms of the period since the last transaction (recency), purchase frequency and total purchase expenditure (monetary). The *k* parameter was set to 8, since eight (2x2x2) possible combinations of inputs (RFM) can be obtained by assigning  or , according to the average to R,F,M values of a cluster being less than or greater than the overall average. If the average R (F, M) value of a cluster exceeded the overall average R (F, M), then an upward arrow  was included, otherwise and downward arrow  was included. For example, RFM represents that the average recency value of a customer segment is greater than overall average, while frequency and monetary average values are smaller than overall averages. These eight customer groups include best customers (most valuable), valuable customers, shoppers, first-time customers, churn customers, frequent customers, spenders, and uncertain customers (least valuable).

**Customer behavior prediction**

A customer segment is not as enough to identify, and then to predict customer’s behavior. Many direct marketers believe that the RFM variables of customers are generally associated with customer profiling. For example, customers with profiles *age = teenager* and *gender = female* and *state = Aegean* can generally have *R**F**M* pattern, while customers with profiles *age = senior* and *gender = male* and *state = EasternAnatolia* can generally have *R**F**M* pattern. For this reason, in this step, classification rules were discovered using demographic variables (age, gender, education level etc.) and RFM values of customer segments

## Product recommendation

In the proposed approach, after generating classification rules, association rule mining was applied to extract recommendation rules, namely, frequent purchase patterns from each group of customers. The extracted frequent purchase patterns represent the common purchasing behavior of customers with similar RFM values and with similar demographic variables. For example, not all women aged 45-54 have the same tendency to purchase a product; so we should also consider their RFM values, customer segments and the other products frequently purchased together with that product.

This chapter presents incorporating RFM analysis into data mining techniques to provide market intelligence. It aims to bring attention to data miners and marketers to the importance and advantages of using RFM analysis in data mining. In order to evaluate the proposed model and empirically demonstrate the benefits of using this model in direct marketing, a case study was carried out using the datasets collected within two years period by a sports store in Turkey through its e-commerce website. According to experimental study results, the proposed approach provides better product recommendations than simple recommendations, by considering several parameters together: customer’s segment, the current RFM values of the customer, potential future customer behavior and products frequently purchased together.

**Chapter-4 Challenges of Performing Analysis**

The benefits of customer segmentation analysis are clear. By having a stronger understanding of their consumer base, retailers can properly allocate resources to collect and mine relevant information to boost profits. However, getting to the point of performing high-level customer segmentation analysis is more difficult than originally thought for many retailers. Many retailers may have the rights to the necessary data to perform the analysis, but do not have either the ability to access it in a user-friendly manner or have an employee that has the skills to work with it. The lack of proper personnel or equipment to handle the necessary volume of data is perhaps the biggest hindrance to smaller firms being able to perform such analysis. The popularity of open- source programming software such as R or Python has certainly helped make this type of analysis more accessible, but it still would require retailers having someone on their team who can code in either of those languages. Additionally, some retailers are simply unaware of either the extent of their data collection or are not yet inspired to dig into it. Nevertheless, retailers that have not fully adopted customer segmentation analysis are likely not doing so simply because they cannot afford to spend the time, money, or labor to perform the analysis. Therefore, it is an aim of this paper to show that this rich analysis can be performed cheaply and efficiently.

However, there is a far subtler but still consequential reason why retailers do not implement customer segmentation analysis: it is too complicated to un- derstand. When compared to traditional demographic segmentation or RFM analysis, high-level customer segmentation analysis requires far more precise knowledge of machine learning and the mathematics that describe how the algorithms work. In addition, traditional marketing analysts are not equipped with the math or programming skills necessary to successfully implement customer segmentation analysis with machine learning methods [21](#_2szc72q); similarly, pro- grammers and data analysts are not well-suited to handle marketing tasks. This poses another conundrum as it involves transforming a typical marketing assignment—segmenting customers based on purchasing behaviors— into a purely programming one, which means the marketing team does not have the skills to code it up themselves but the programming team does not have the marketing skills to interpret the results. Hence, there is a necessity for a hybrid role that involves knowledge of the business, programming, and marketing. In modern workspaces, this role is dubbed the data scientist or information specialist.

In sum, customer segmentation analysis is the process of trying to under- stand a consumer base by splitting it up into segments. While traditional analysts found some success with demographic or RFM analysis, these models simply do not have the technological capabilities to provide rich insight into more specific details regarding the customers. On the other hand, customer segmentation analysis that is combined with machine learning methods has the ability to transform the way a retailer thinks about their data. As such, retailers are trying to find cheap, easy ways to implement and communicate how clustering can be used to segment their customers.

Now that there has been plenty of introduction into customer segmentation analysis, it is time to take a look under the hood of some clustering algorithms before finally engaging in discussion of the analysis

## Clustering Algorithms

Clustering algorithms are used to assign users into groups so that users belonging to the same group are more similar than users in another group. The goal of this division is to find meaningful underlying patterns within the data space. User similarity is determined by a distance measure. This section will introduce the most common similarity measures and clustering algorithms.

**Clustering other ML Methods**

While many applications of machine learning, such as regression and classifica- tion, focus on predicting the outcome or value of an instance, these applications do not attempt to understand similarities between instances, just the relation- ship between instances and their respective outputs. Thus, when it comes to searching for algorithms or methods that look for similarities between fea- tures of instances, the focus must turn from supervised machine learning to unsupervised machine learning.

Determining whether an algorithm is a part of supervised and unsupervised machine learning is contingent upon whether the instances used to train the model in the training data contain their target value. In all cases of supervised machine learning training, instances are paired with a target value, which could be a scalar or a vector depending on the context. In contrast, unsupervised machine learning deals with data that is not paired with a target value. To clearly spell out these differences — and also certain similarities — it may be best to examine them through an example.

For instance, consider a retail store owner who has a store that has been open for over a year and they are interested in examining their data to help boost understanding of their customers while also predicting how much they will spend next visit. To predict their next ticket, the owner takes their pre- vious purchases and comes up with a way to guess, based on the previous tickets, the value of the next purchase. Since this example involves prediction and the outcomes of previous data and its outcomes (the tickets themselves), this is an example of supervised machine learning. To be more specific, since the owner is likely trying to predict a dollar amount the customer will spend, this type of algorithm is called regression. On the other hand, to boost the understanding of their customers, the owner decides to look at some collected customer data and see if there are broader patterns or similarities between the customers. Since there is no clear outcome or target value associated with the data or the process, this is a type of unsupervised machine learning. More precisely, this exemplifies clustering. In technical terms, clustering is an unsupervised machine learning technique that groups instances into clusters based on the similarities between instances.

**Similarity Measures**

The success of a clustering algorithm rests upon the ability to choose the proper similarity measure before engaging in clustering. Choosing the best similarity measure, however, depends on an acute awareness of what similarity is and how it can be defined mathematically.

First and foremost, similarity in data science is a function of distance; the closer together two instances’ values are, the more similar they will be. In certain contexts, defining distance between two instances is more obvious than others. If a data scientist were to cluster instances based solely on one numerical feature, then the clustering algorithm would take into account the differences between the instances and group them based on that. If the data scientist were to consider two features, the distance between features is a little more complicated. Instead of just the difference between the instances, we have the Euclidean Distance[22](#_279ka65) between instances x and y:

*distance*(*x, y*) =sqrt[(*x*1 − *y*1)2 + (*x*2 − *y*2)2] (2D Distance Formula)

This formula comes from the Pythagorean Theorem and the fact that in Euclidean Geometry, the shortest distance between any two points is a straight line. The distance of the straight line in this case is calculated using the above formula. But, this is a good time to pause and evaluate certain understandings and motivations of what is going on here.

In order to find a way to compare how similar two instances are, it was first necessary to define the relationship between similarity and distance. In most contexts, the natural relationship to establish is an inverse relationship, which is what is used in this paper. The next necessity is to define the distance between two instances. With numerical data, such as the data in this project or in the previous examples, the natural distance measure to use is the standard Euclidean Distance Formula. The main reason why this is the natural measure for distance is that the data we are interested in clustering is numeric in nature. In other contexts such as Natural Language Processing, notions of similarity begin to diverge from the simple numerical notation presented here. But, the essential point is that finding the similarity between two instances involves at least two forethoughts; what is the relationship between distance and similarity, and how is distance defined in this context?

As alluded previously, this project makes use of Euclidean Distance as a way to define the distance between two instances. While the two examples above talk about distance on a one or two-dimensional level, the data in this project involves much more than two features, and so the intuition that guided the lower dimensional thinking needs to be expanded into higher dimensions. It turns out, when expanded into n features, the distance formula gets a more general look:

*n*

*distance*(*x, y*) = (*xi* − *yi*)2

*i*=0

(n-D Distance Formula)

Many textbooks or academic papers may choose to refer to the distance formula in the context of clustering without the square root sign.

*n*

*distance*(*x, y*) = (*xi* − *yi*)2

*i*=0

This is a shorthand way of saving computational power in the actual clus- tering algorithm, but it ultimately is tied to the fact that when it becomes important to minimize the distance, finding the minimum of a squared dis- tance or the square root of some squared distance yields the same minimum[23](#_meukdy). Now that there has been a discussion on similarity, it is appropriate to begin to explore two different types of clustering algorithms. Both were used in the context of this project, which makes it crucial to compare the two algorithms in this paper because they reveal deeper insight into how different

types of clustering can be used in differing contexts.

### Chapter-5 K-means Clustering

K-means clustering is widely used in the field of cluster analysis and customer segmentation. K-means is an algorithm designed to group a set of items into *K* subgroups or clusters. The algorithm is dependent on a manually set value for *K*. The *K* centroids are initialized to random observations in the dataset. K-means is then tasked with iteratively moving these centroids to minimize the cluster variance using two steps:

* + - * for each centroid *c* identify the subset of items that are closer to c than any other centroid using some similarity measure.
      * calculate a new centroid each cluster after every iteration which is equal to the mean vector of all the vectors in the cluster.

This two-step process is repeated until convergence is reached.

The standard implementation of K-means uses Euclidean distance mea- sure described in the section above to find the subset of items that corre- sponds to each cluster. This is done by calculating mean squared error, which in this case is equivalent with the Euclidean distance, of each item’s feature vector with the *K* centroid and choosing the closest result. However, other distance measures can be used instead of Euclidean distance. Aggarwal et al. claim that for high dimensional data, the choice of distance measure used in clustering is vital for its success

**Centroid-based: K-Means**

Although there are numerous types of clustering that each deserve their own mentions and explanations, this paper will focus on two types of clustering algorithms: centroid-based and hierarchical-based. Though, before beginning an ample discussion of centroid-based clustering, it is first necessary to under- stand what a centroid is and how they fit into clustering.

In the context of centroid-based clustering, a centroid is the center of a cluster of data. Although there are numerous ways to define the center of a cluster, the center in a k-means cluster is the arithmetic mean of each feature in the space in which the data exist. In other words, the centroid is the mean of the features of the instances that are assigned to that cluster[24](#_36ei31r). However, it might not be immediately clear why centroids are necessary in the first place or how to initially define them. After all, there has not been any discussion on how exactly an algorithm would go about grouping data into clusters, let alone how centroids fit into that process.

To begin this discussion, it is appropriate to also begin with an example. Imagine that there is some 2D data that, when plotted, looks like the following:

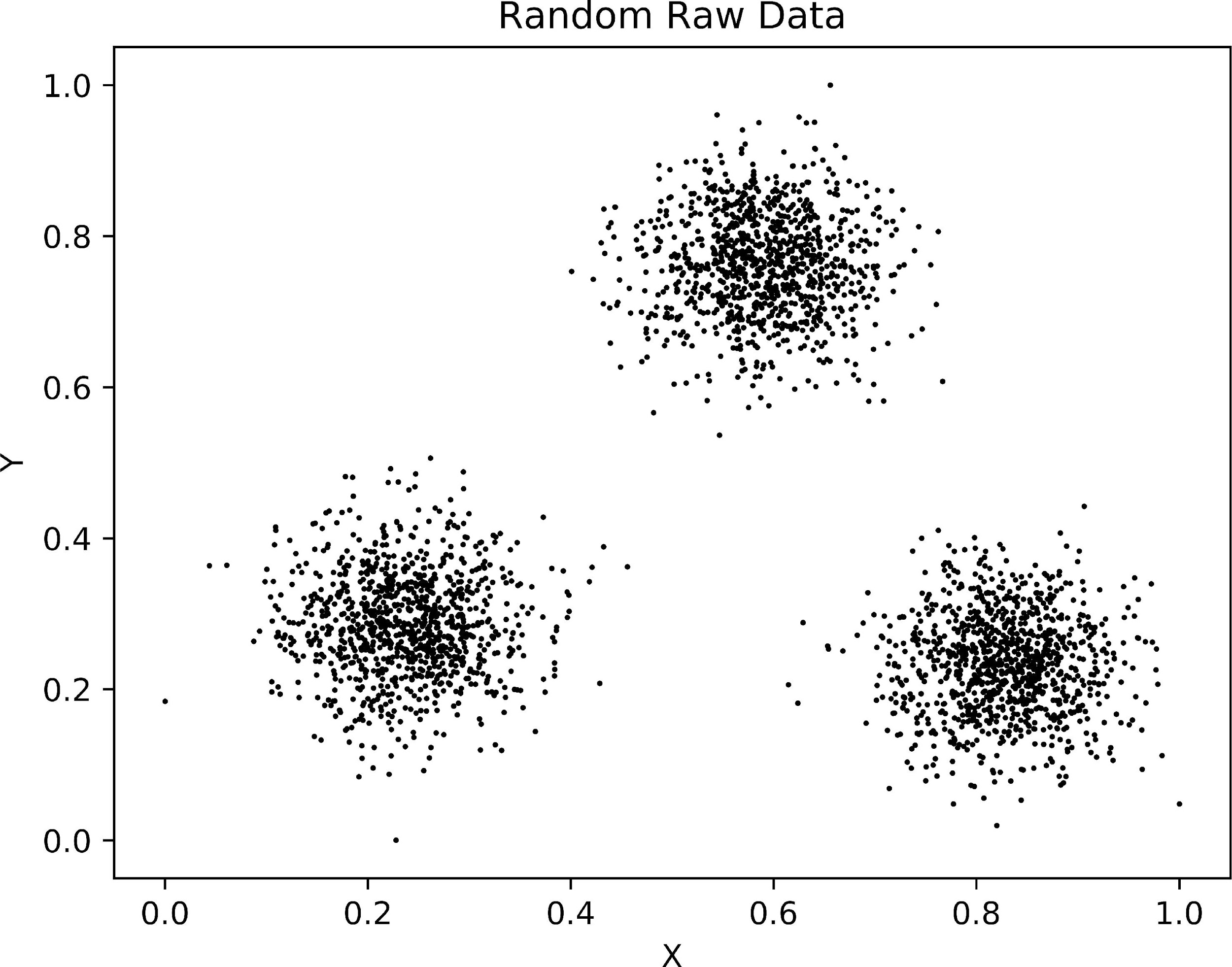


Figure 1: Random Raw Data

Immediately after looking at this figure, there are three things that are apparent. First, each data point exists in two dimensions: x and y. Although naming the axes x and y is convenient, it does not provide much insight into what the axes represent. So, if x and y are too simple, perhaps think of them as age and income or time between visits and average ticket. Regardless of what the axes’ names are, the crucial point is that there are two dimensions. Second, the data is scaled between 0 and 1 in both dimensions. Given this paper has yet to discuss the importance of scaled data in clustering, it might not make much sense why this is an important thing to note. Without revealing too much detail, the gist is that scaling keeps features that have large ranges from overwhelming data with smaller ranges. Furthermore√, the 0-to-1

scale means that the maximum distance between any two points is *M* , where

M is the number of dimensions (in this case it is 2). To go along with this, the 0-to-1 scale also makes sure that none of the numbers, when squared, are bigger than 1; this is an often understated point in discussions of K-Means. When there are several hundred—or even thousand—numbers being summed and squared during the distance calculation, it is important to have smaller numbers because they will take up less memory in the long run. Thus, despite the 0-to-1 scale at first appearing meaningless, the scaling makes it easier to compute distance.Lastly, to a human eye, there appear to be at least three distinct clusters. To some, this might be trivial to point out: by looking at the figure, it feels natural and also simple to place each data point into one of three clusters. In essence, this natural feeling is a reflection of the idea that humans are excellent at finding patterns/commonalitiesbetween instances under two conditions: when there are not that many instances and when there are not that many features. In the figure, there are two features and although there are several thousand instances, plotting them all at once makes it easier to see the differences between each datum. Because of the low number of features and the ability to see all the data clearly, the human mind has little difficulty dividing up the data into clusters. However, teaching a computer to perform the same task is slightly more difficult. For all the incredible tasks that a CPU can perform, it cannot visualize the data and divide it into nice groups like a human can. Therefore, it is reasonable to wonder how a computer would go about the task of clustering.

In centroid-based clustering, the most popular algorithm is k-means. There are two important parts to the name: k and means. Here, k refers to the number of centroids (clusters) the algorithm will generate and “means” refers to what the centroids are: arithmetic means of the data [27](#_2afmg28). Roughly the k- means algorithm can be broken up into four sections, each with their own important attributes.

To start, while it is clear what a centroid is, it is unclear how it fits into the algorithm at the beginning. First, if a centroid is supposed to be the arithmetic mean of the points that belong to it, how is it possible to use them initially? In other words, how does one know where to put the centroids? In short, the smartest and most common decision to make is to randomly place the centroids throughout the dataset. Here, it might help to think of centroids as points in the same space that the data belong. In the raw data shown in figure [1,](#_2xcytpi) centroids would appear as a random point between 0 and 1 for both its x and y component. In relation to this, it is also reasonable to wonder how many centroids to randomly place throughout the dataset. For the sake of simplicity, let’s initialize three centroids and place them randomly throughout the dataset.

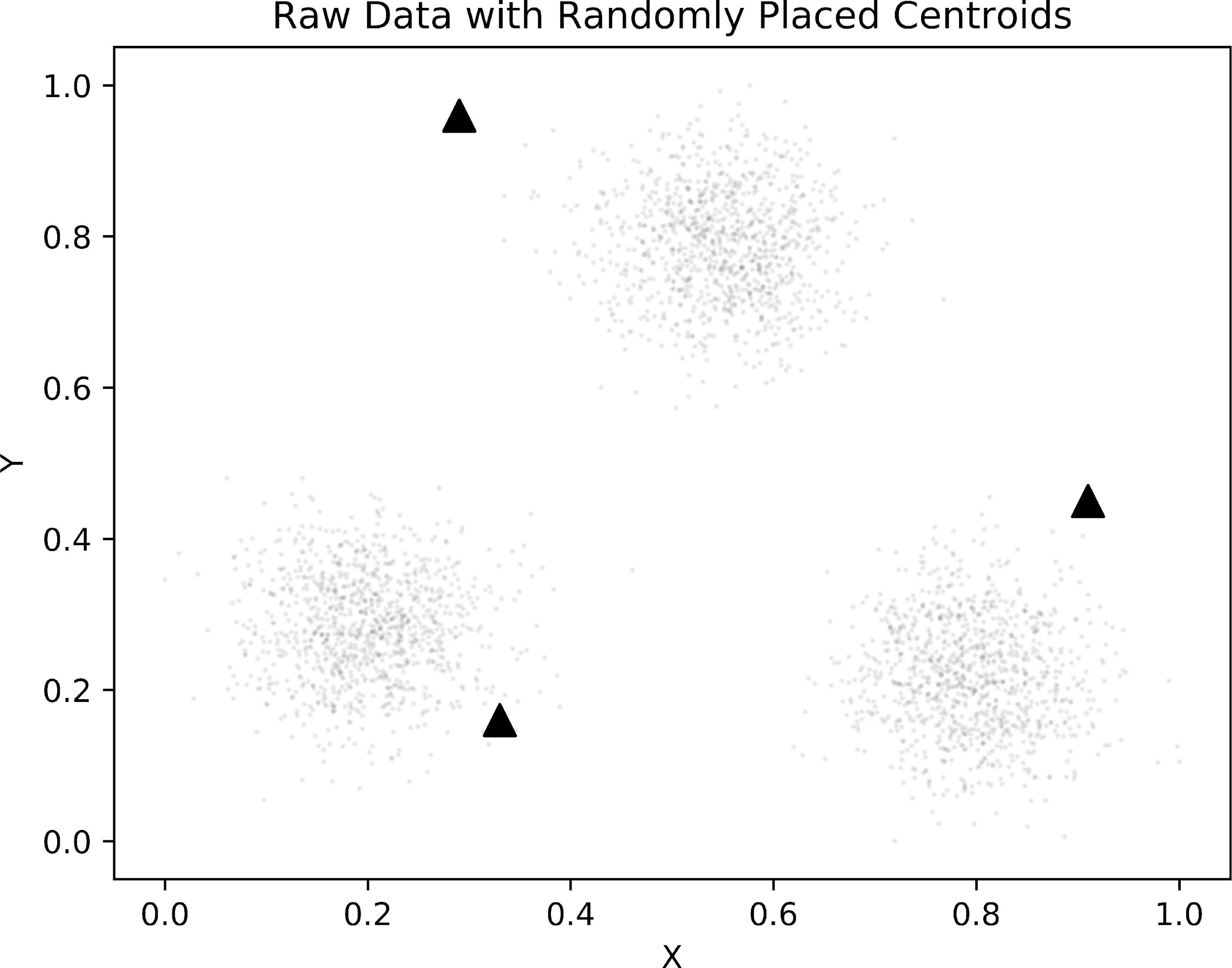


Figure 2: Raw Data with Centroids

Now that the centroids exist in space, it is nearly time to begin cluster- ing. Before beginning the main chunk of the k-means algorithm, it is necessary to assign each point to one —and only one— of the centroids. To assign a point to a centroid, one must first find the distance from each point to each centroid. The data point, thus, will be assigned to the centroid that is closest to it or, equivalently, the one to which it is the most similar. Figure [3](#_2bn6wsx) shows an example of taking one point and computing the distance (shown in red) between itself and each of the three centroids. From the figure, it becomes easy to see that the randomly selected point should be assigned to the left-most centroid, since it is the closest centroid to the point. This process of assigning points to the closest centroid repeats for all remaining points in the dataset.

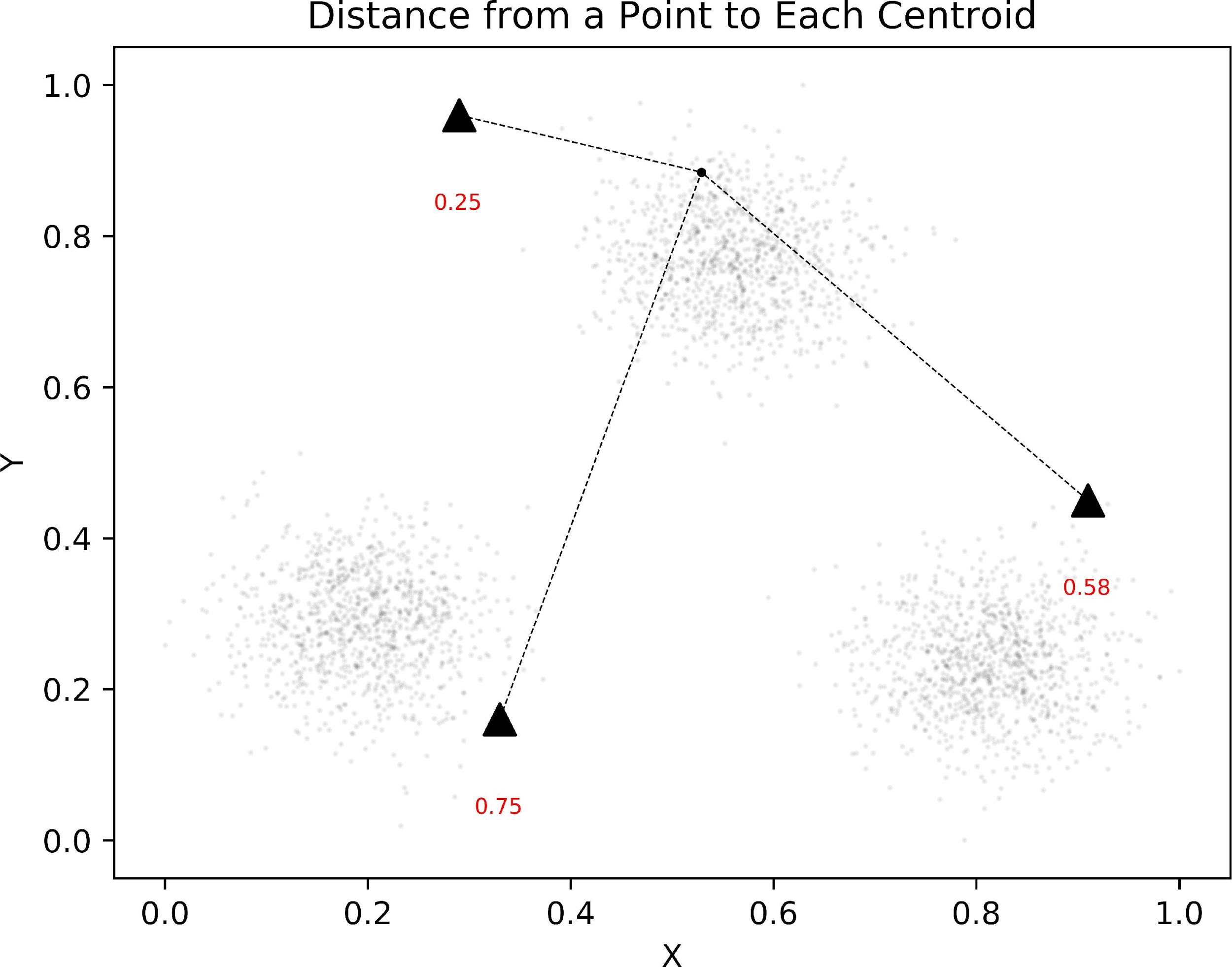


Figure 3: Distance from a Random Point to Each Centroid

Once each point is assigned to a centroid, it is time to update the position of each centroid. In k-means, recall that the centroid is the arthmetic mean of the data that belong to that centroid. So, in two dimensions, this idea can mathematically expressed as:

*centroidi*

*n*

= ( *X*

Σ1 ∗

*n*

*j*=1

(*j,*1)

*n*

*, X*

Σ

*j*=1

(*j,*2)

) (2D Centroid Update)

Here, each instance resides in *X* and instance *Xj* is assigned to centroid

*i*. Any instance not assigned to centroid *i* does not affect the reassignment of the centroid. Furthermore, the terms *X*(*j,*1) and *X*(*j,*2) indicate the value of the

first and second features of instance *Xj* respectively. Lastly, the 1

*n*

is the way

this calculation becomes an average, since *n* represents the number of instances belonging to centroid *i*.

However, it is also generally useful to understand how similar concepts can be applied outside of two dimensions. When expanded into higher spaces, the update formula changes to (in k dimensions):

*centroidi*

*n*

= ( *X*

*n*

*j*=1

(*j,*1)

*n*

*, X*

*j*=1

(*j,*2)

*n*

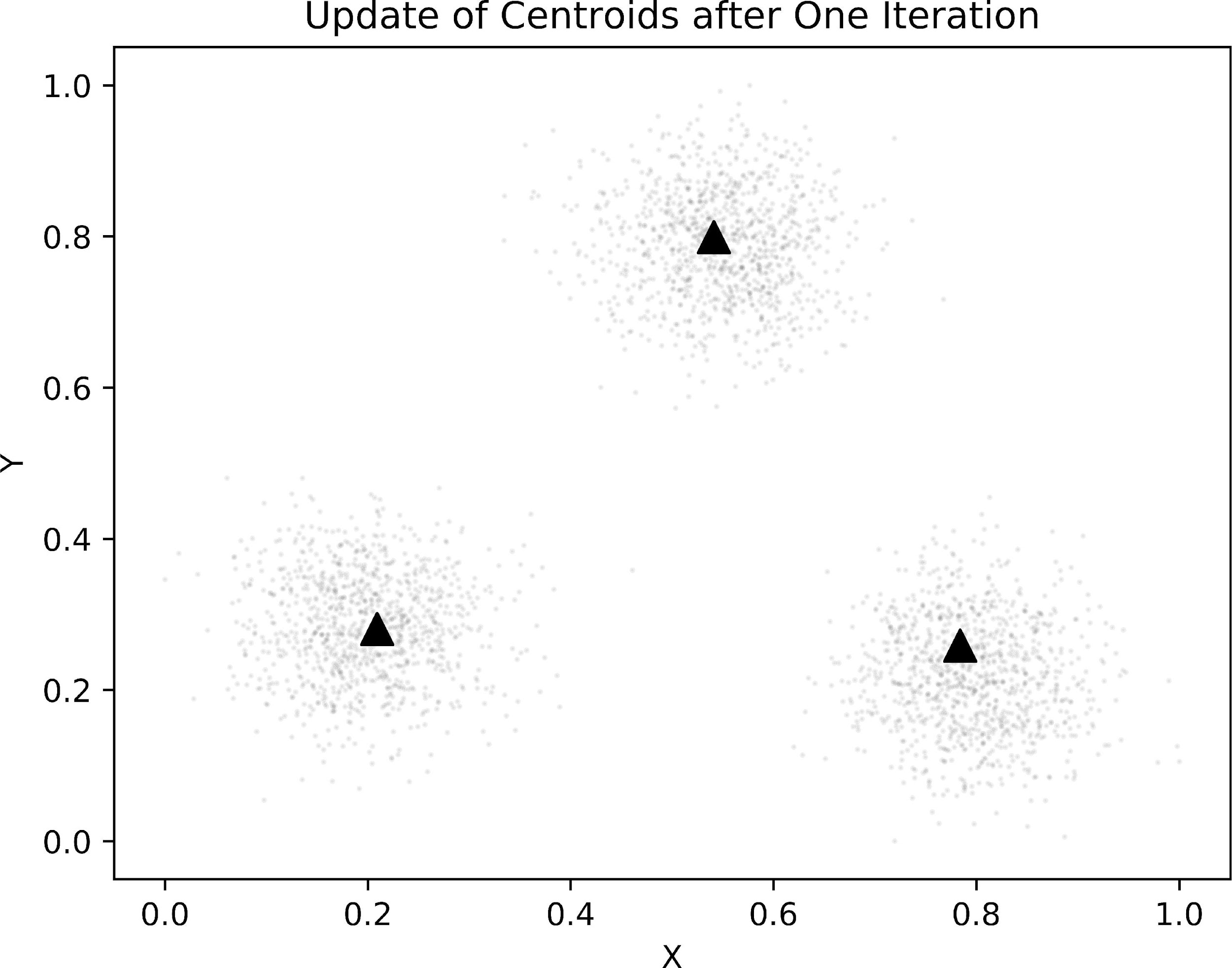
*, . . . , X*

*j*=1

(*j,k*)

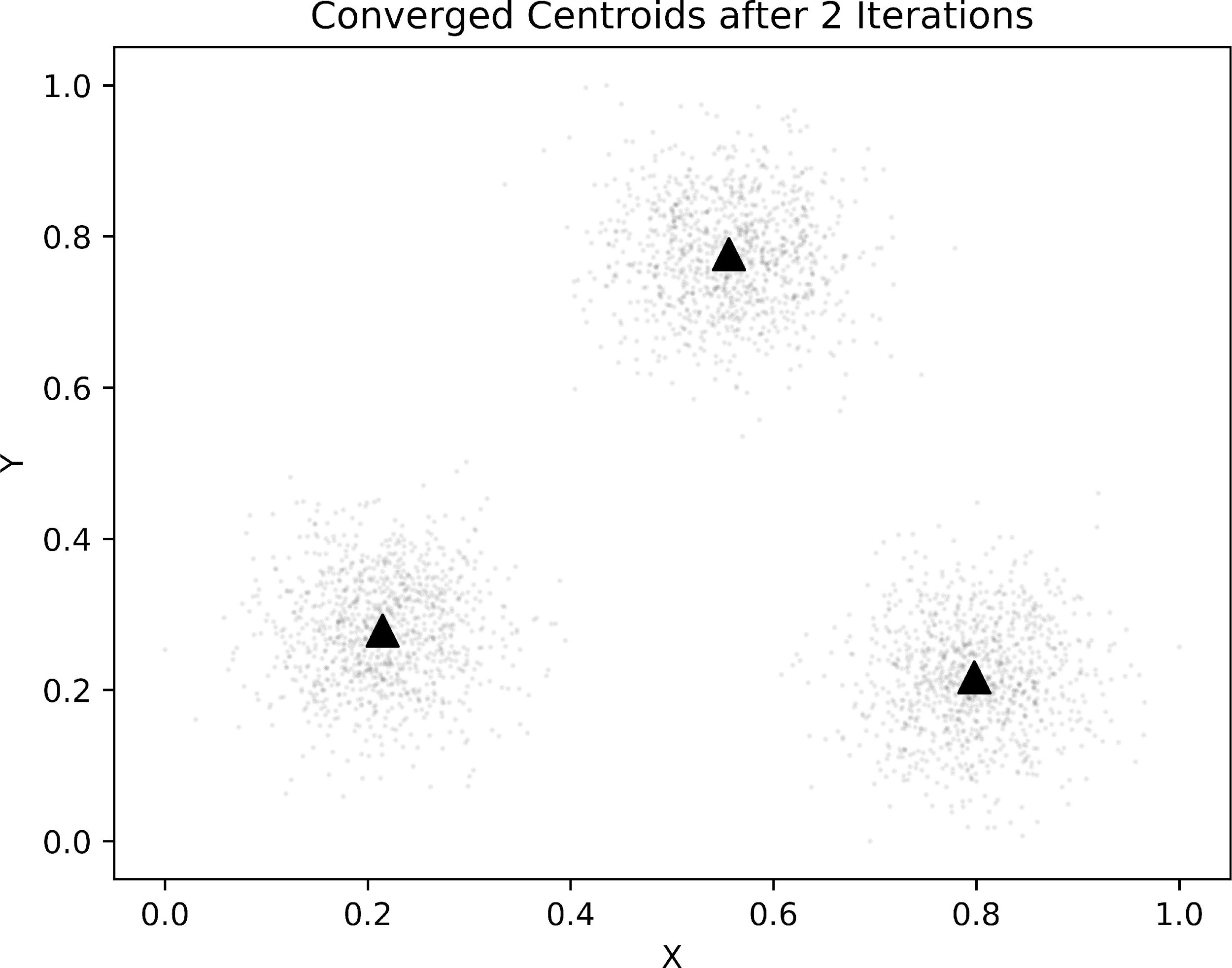
) (k-D Centroid Update)

Figure [4](#_qsh70q) graphically shows the placement of the new centroids after up- dating their positions. When comparing this figure to figure [2,](#_1ci93xb) it is evident that each of the centroids moved toward the direction of the points closest to them.



Update of Centroids after One Iteration

The update of centroids, now, is not too difficult to explain or implement, but when should the algorithm stop updating centroids? In practice, the k-means algorithm stops when either the centroids remain unchanged from the previous iteration or, equivalently, the labelling of each point to a centroid remains unchanged from the previous iteration; this is called convergence. Since the data in this example are particularly well-suited for clustering, it should not be a surprise that the first iteration of k-means yields centroids that are very much close to the ideal centers of each of the clusters. Similarly, the algorithm here, as shown in figure [5](#_3as4poj) actually converges only after two iterations, a very fast convergence.



Converged k-means Algorithm

In this particularly simple example, the centroids converged rather quickly mainly due to the fact that the centroids were already close to nicely shaped clusters. But, consider a case where the centroids are not nicely placed, per- haps closer to each other or closer toward the middle of the data. In this case, the centroids do not converge anywhere near as quickly, and they may converge in different spots than the simple example. Because of this, it is common for data scientists to run the K-Means clustering with several different random initial assignments and take the one that has the smallest inertia, which is the sum of the square distance between each point and the centroid. Inertia will be discussed more specifically in section 5 in the context of choosing the optimal k—or the number of centroids— to cluster with.

Before continuing, it is worth the time to quickly summarize k-means and centroid-based clustering. The k-means algorithm works by taking k centroids and randomly placing them across the dataset, ideally so that they are evenly spaced out. Each datum then gets assigned to the centroid it is closest to, which is the centroid with the smallest Euclidean Distance between it and the datum.

Once all the data have been assigned, the centroids update by becoming the mean of all the data assigned to it. When the centroids stop moving or the assignments stop changing, the algorithm stops. To get close to the optimal solution for a particular k, it is recommended to rerun the algorithm with different initial centroid assignments.

In sum, centroid-based clustering is one of the most common ways to cluster data with machine learning methods, but it is not flawless. For example, the number of centroids, k, has to be chosen beforehand, which makes it more difficult to find the optimal k to cluster with. Furthermore, k-means operates under the assumption that the data will have nice “centers”[34](#_45jfvxd) to their segments that will allow the centroids to converge, which is often not the case with real- world data .

This also implies that k-means is sensitive to outlier data that can cause centroids to converge far from the optimal spot. So, for as powerful and influential as k-means is, there are clear reasons to explore other strategies that do not suffer from the same weaknesses. Thus, it is necessary to dive into an alternate form of clustering known as hierarchical clustering.

Hierarchical-based: Agglomerative

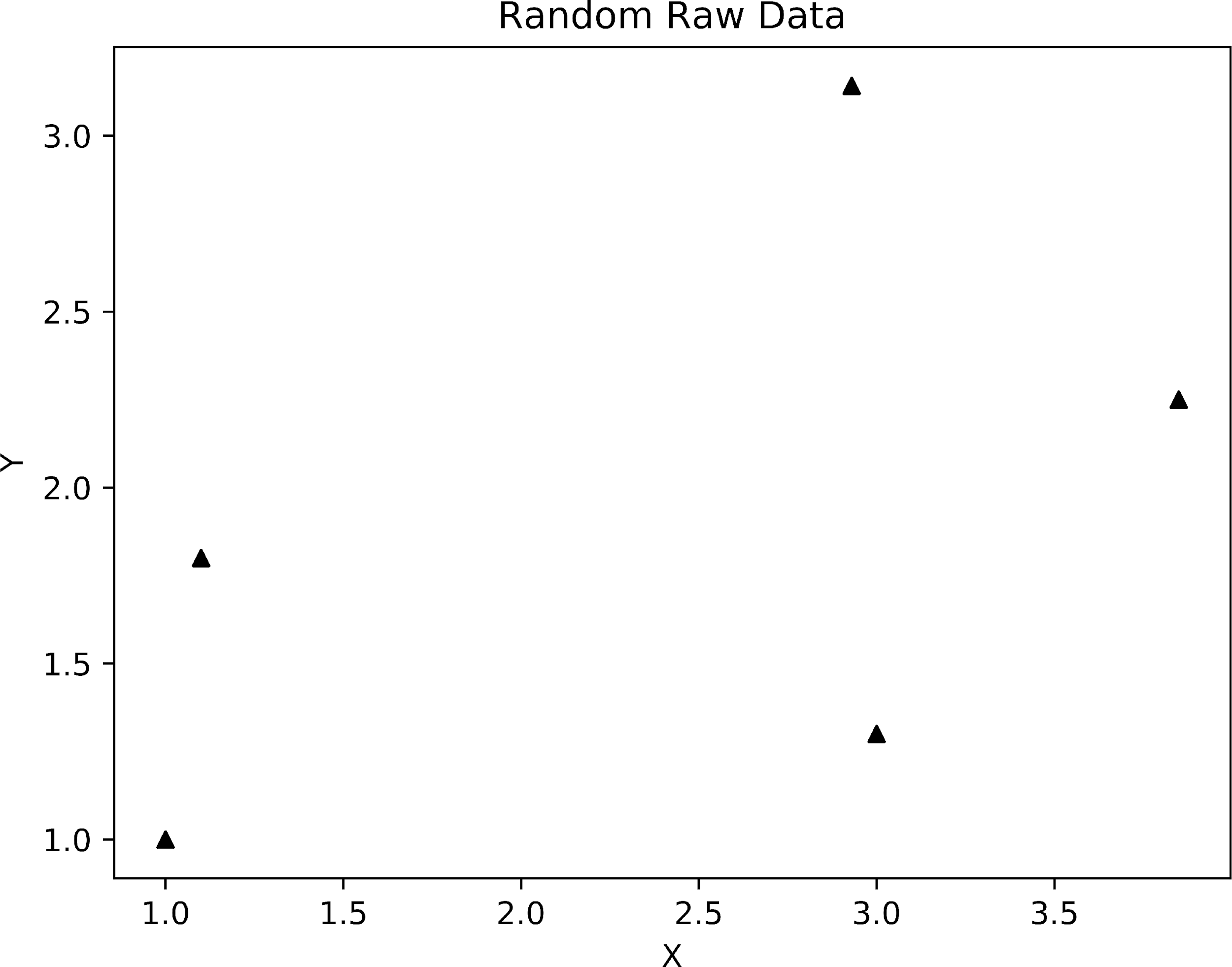
Centroid-based clustering is extremely popular but often improperly applied to real-world datasets. In addition to being susceptible to outliers, it is also frus- trating for analysts to determine the proper number of centroids, k, to specify beforehand. To circumvent these shortcomings in practice, it is common to explore a different kind of clustering algorithm.

While centroid-based clustering is intuitive and easy to implement, hier- archical clustering is comparable in its implementation but does not suffer from the drawbacks that centroid-based clustering does. In short, hierarchical clustering is a type of clustering based on either a top-bottom or a bottom- top approach.

More specifically, a bottom-up approach—where each datum starts as one cluster until they all merge into one giant cluster— is known as agglomerative clustering whereas the converse is known as divisive clus- tering.

Because agglomerative clustering is more intuitive to explain than divisive, this paper will make use of it.

figure : Random Raw Data for Agglomerative Clustering

In any clustering project, it is important to develop some intuitions about the structure of the data prior to engaging in clustering. When looking at the data, the two points in the bottom left corner should stand out due to their relative closeness. In a similar fashion to centroid-based clustering, it is important to explicitly define what “close” means in this context. When each point is its own cluster, it is common to use the Euclidean Distance formula if the features are all numeric such as here. So, it is natural that the first step of the agglomerative clustering algorithm is to find the two closest points: in this case, it is the two leftmost points. In order for the computer to find the two closest points, it is necessary to compute the distance from each point to every other point. This is most easily accomplished by creating a distance matrix where position (*i, j*) represents the distance between points *i* and *j*. The distance matrix is important because it holds the key to the order of grouping for the data. Lastly, while the order of grouping for the rightmost points might not be immediately clear, it should be simple to see that the first grouping will occur with the leftmost points and move on from there. Figure [7,](#_147n2zr) below, visualizes a sample distance calculation for one random point.

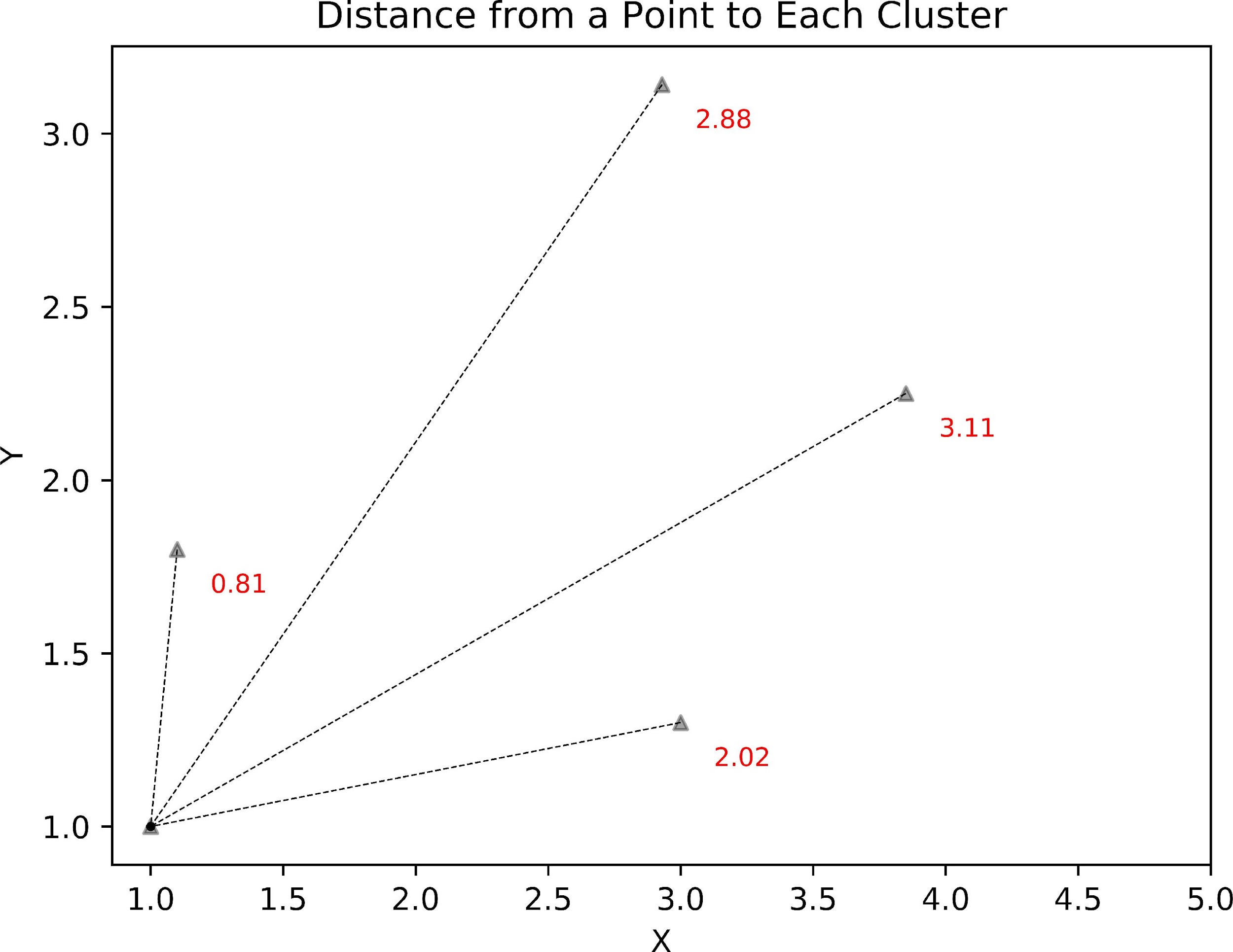


Figure 7: Sample Distance Calculation from One Point to Each Cluster

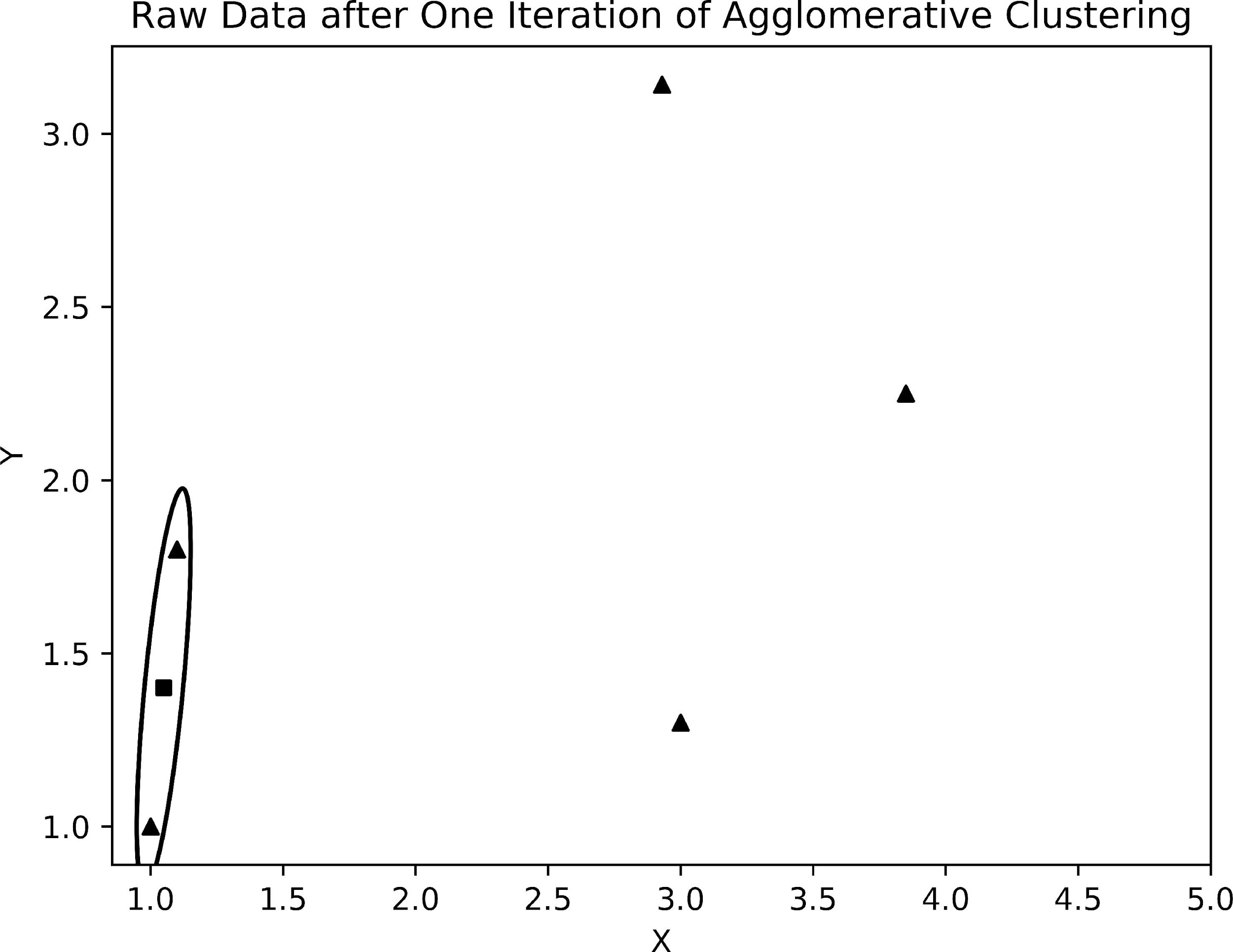
Once the shortest distance is found between two clusters, the two clusters merge into one cluster. In this example, there were five initial clusters but after one iteration, there are four remaining clusters. In general, the pattern of the number of clusters starts at *n*, goes to *n* 1, *n* 2, . . . , 2, then finally

With one merged cluster, it now becomes a little trickier to define how this merged cluster should act in the specified distance formula. In essence, the three most common methods of defining this interaction are to:

Take the shortest possible distance between one of the points in the merged cluster and the desired point. This is the defining characteristic of the single linkage method.

Take the largest possible distance between one of the points in the merged cluster and the desired point. Conversely, this is often referred to as the complete linkage method.

Take the distance from the desired point to the center of the merged cluster.



Example of a Data Update in Agglomerative Clustering

In future iterations of the algorithm, each point will compare itself to the centroid of the merged cluster. As the algorithm progresses, the grouping of clusters and subclusters naturally forms a tree-like structure. This structure, known as a dendrogram, displays crucial information regarding the way the algorithm clustered the data. For this particular example, the results are summarized below in figure [9](#_23ckvvd)

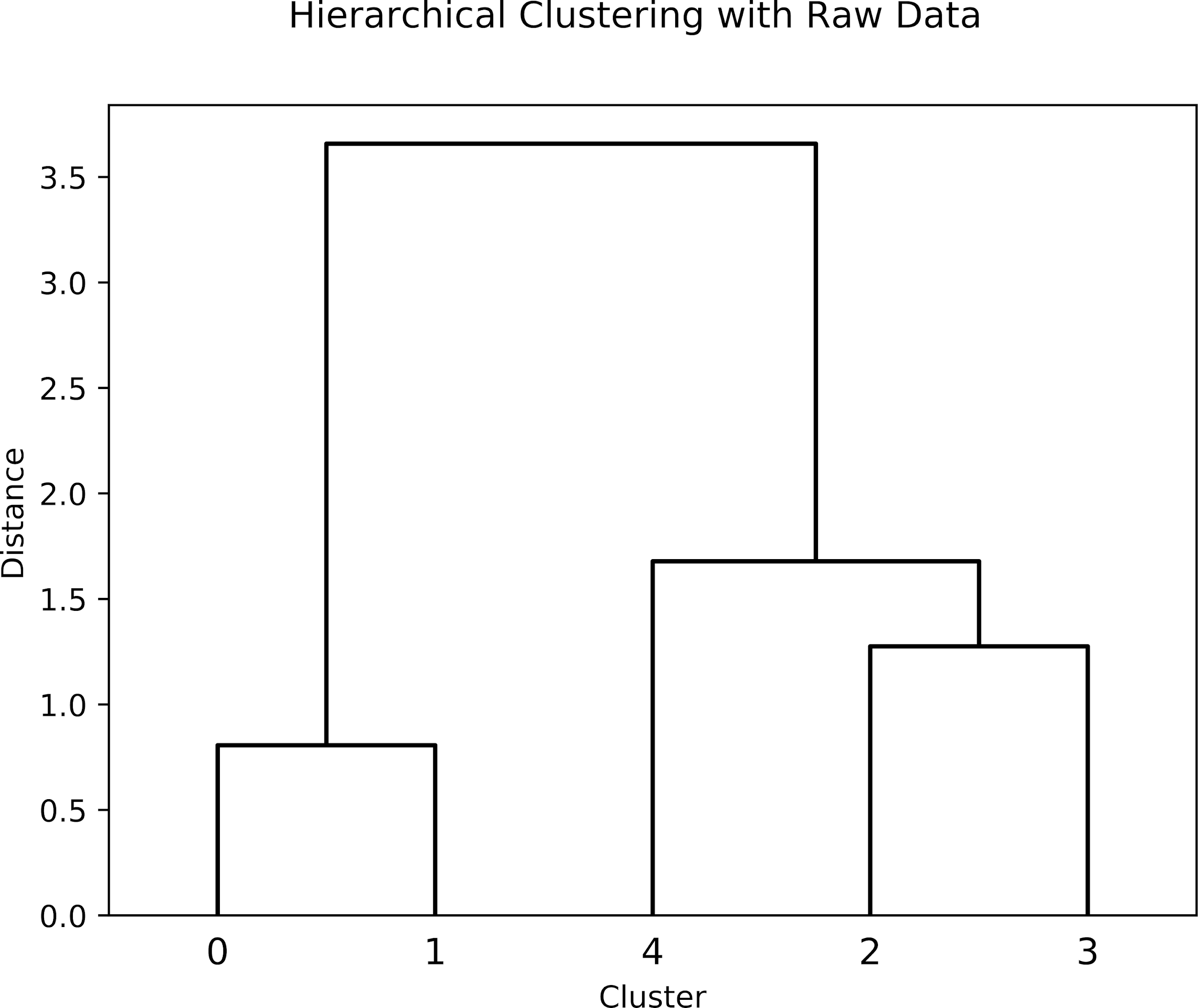


Figure 9: Dendrogram of Clustered Random Raw Data

When analyzing a dendrogram, it is important to examine the locations of the connections and how their distance from the closest connection, which appear as the flat lines on the figure. For example, points 0 and 1 — the leftmost and two closest points — join first and thus have the connection closest to the bottom. The shorter the distance between the connections, the more closely related the two clusters are. Once all the clusters finish agglomerating, each of the connections and their relationship to one another is beautifully clear.

To recap, hierarchical clustering is a type of clustering that involves estab- lishing a hierarchy in terms of how similar two clusters are. In agglomerative clustering, each datum starts as its own individual cluster and proceeds to join with the most similar cluster. With numeric data and working with Euclidean Space, the Euclidean Distance Formula is the most common dis- tance/dissimilarity metric. However, since the number of clusters updates each iteration of the algorithm, it is important to keep track of the distance between each cluster to every other cluster via a distance matrix. Once there is only one cluster remaining, the algorithm converges and the results are commonly visualized with a dendrogram.Analysis shows that these two topics are important for retailers to understand. While deserving of their own research and recognition, these topics are far better understood when they work in conjunction with one another. Clustering can transform a retailer’s customer segmentation analysis into a formidable research tool by providing richer analysis and utilizing more data. In fact, the harmony between customer segmentation and clustering is precisely the relationship that overarches the motivations and implications of writing this paper. Without further ado, it is finally time to begin discussing the setup and results of the clustering experiment performed for Front.

**Data Preparation**

The digressions of clustering and customer segmentation analysis were im- portant, but it is now time to think back to the previously stated business problem and the associated data. Although several variables from each data table were listed, not all the variables could be used in the analysis as is. Certain variables, such as the ID columns, provide necessary information to corroborate data and keep accurate calculations between instances, but are not necessarily features that merit analysis. On a similar note, features, such as the time of a specific transaction and the value of its ticket, contain essential information for mining, but need to be transformed into a more usable for- mat. In particular, these transaction-based variables need to be converted into customer-based variables. However, variables such as the product category are on an item-based level, which require a separate transformation of their own. Nonetheless, the salient point is that it is necessary to consider the raw data, examine its format and original features, and transform them into a workable format for the task at hand.

**Feature Engineering**

The process of creating or extracting features from raw data is commonly re- ferred to as feature engineering. Often, it is the first and most important step of data preprocessing because it establishes the features that the model will consider when clustering. Essentially, feature engineering involves inspecting and manipulating the raw data to somehow extract features that are worth- while for analysis. Because the concept of a “worthwhile” feature is subjective, the data scientist must place the task’s mission and constraints at the forefront of their decision-making process with regard to engineering features. In this project specifically, one of the main goals is to obtain a better understanding of 4Front’s customers based on their purchasing patterns. So, the features that will appear on a customer-based level, describe purchasing patterns, and extract the most information from the raw data will be optimal features for the project. After poring through the datatables, there were 11 unique features that were engineered that are summarized as follows:

Visits: the total number of visits the customer has made to the store since inception

Total Spent: the total amount spent across all visits

Average Ticket: total spent / visits

Average Time between Visits: the average number of days between visits. By default, it is -1 if the customer has visited less than twice

Flower: the proportion of purchased items that are categorized as flower

Vape: the proportion of purchased items that are categorized as vape, which includes live resin cartridges as well as the standard distillate car- tridges

Concentrate: the proportion of purchased items that are categorized as concentrate. Included in this is kief, shatter, wax, batter, and other dabbable forms of cannabis

Preroll: the proportion of purchased items that are categorized as a pre roll. This includes individually packaged pre rolls as well as those sold in packs of two or more.

Edible: the proportion of purchased items that are categorized as edibles. Chocolates, drinks, teas, gummies, and mints are examples of what is considered an edible.

Topical/Other: the proportion of purchased items that are categorized as either topicals or anything not in the above categories.

visits, the cannabis-related data drives the segmentation to focus on cannabis- specific purchasing behavior. This is crucial to recognize because there is not significant research of cannabis retail data. Thus, there is motivation to perform analysis in a way that is directly applicable to cannabis data.

Lastly, computing or finding the necessary data for analysis was more dif- ficult than hypothesized, particularly finding the cannabis-related purchase behavior data. In the 4Front database, each product has a specific product category ranging from 1-20. However, there is insufficient documentation on what each of the product categories refer to. As a result, it was necessary to take a sample from each product category and manually classify them into the broader categories (Flower, Edible, etc.) used in the clustering. Moreover, if this analysis is to be repeated across stores, the product categorization would have to be completely redone. This presents a problem for cannabis retail data scientists that will only likely be addressed if similar struggles are publicly shared.

**Chapter-6 MARKETERS**

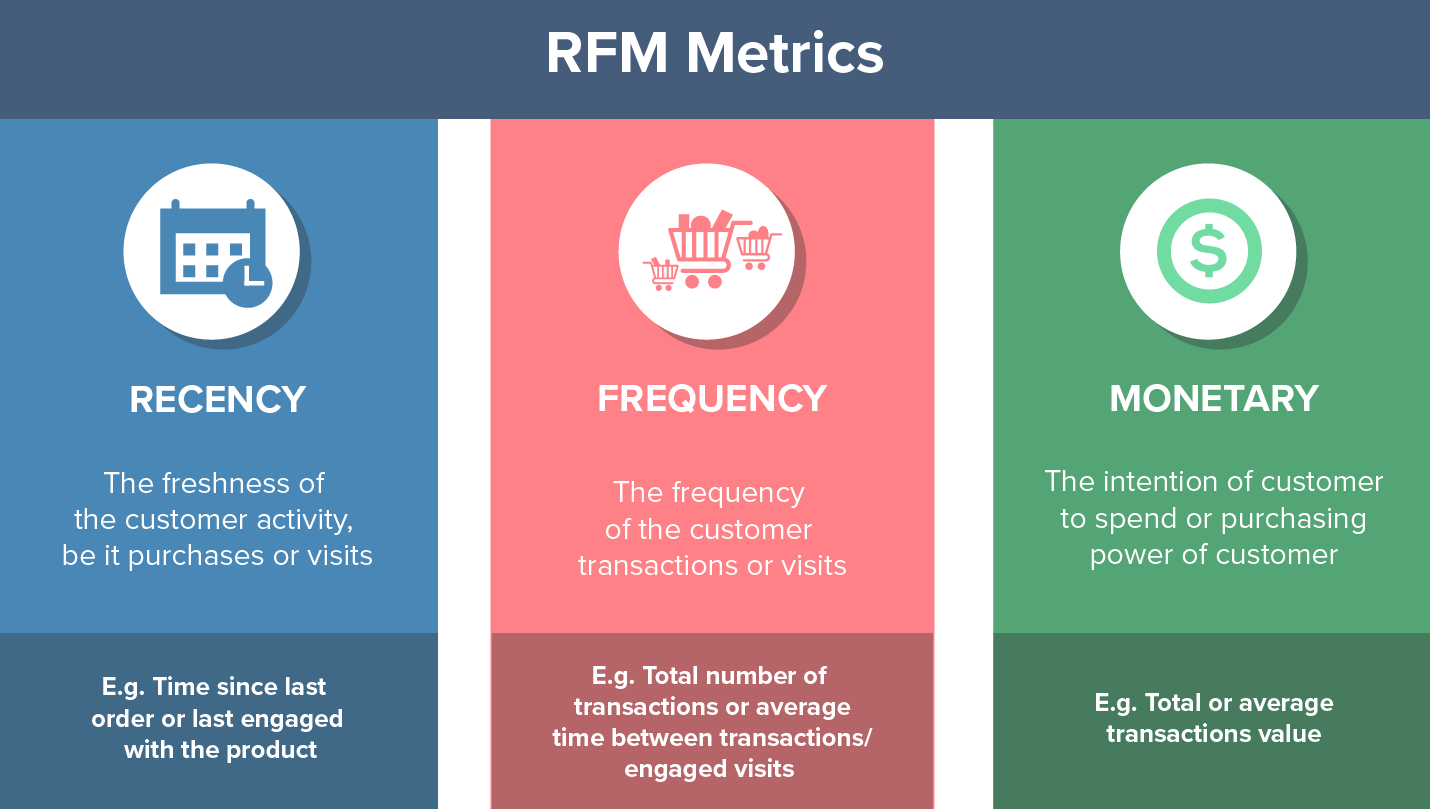
Smart marketers understand the importance of “know thy customer.” Instead of simply focusing on generating more clicks, marketers must follow the paradigm shift from increased CTRs (Click-Through Rates) to retention, loyalty, and building customer relationships.

Instead of analyzing the entire customer base as a whole, it’s better to segment them into homogeneous groups, understand the traits of each group, and engage them with relevant campaigns rather than segmenting on just customer age or geography.

One of the most popular, easy-to-use, and effective segmentation methods to enable marketers to analyze customer behavior is RFM analysis.

**Chapter-7 RFM Analysis**

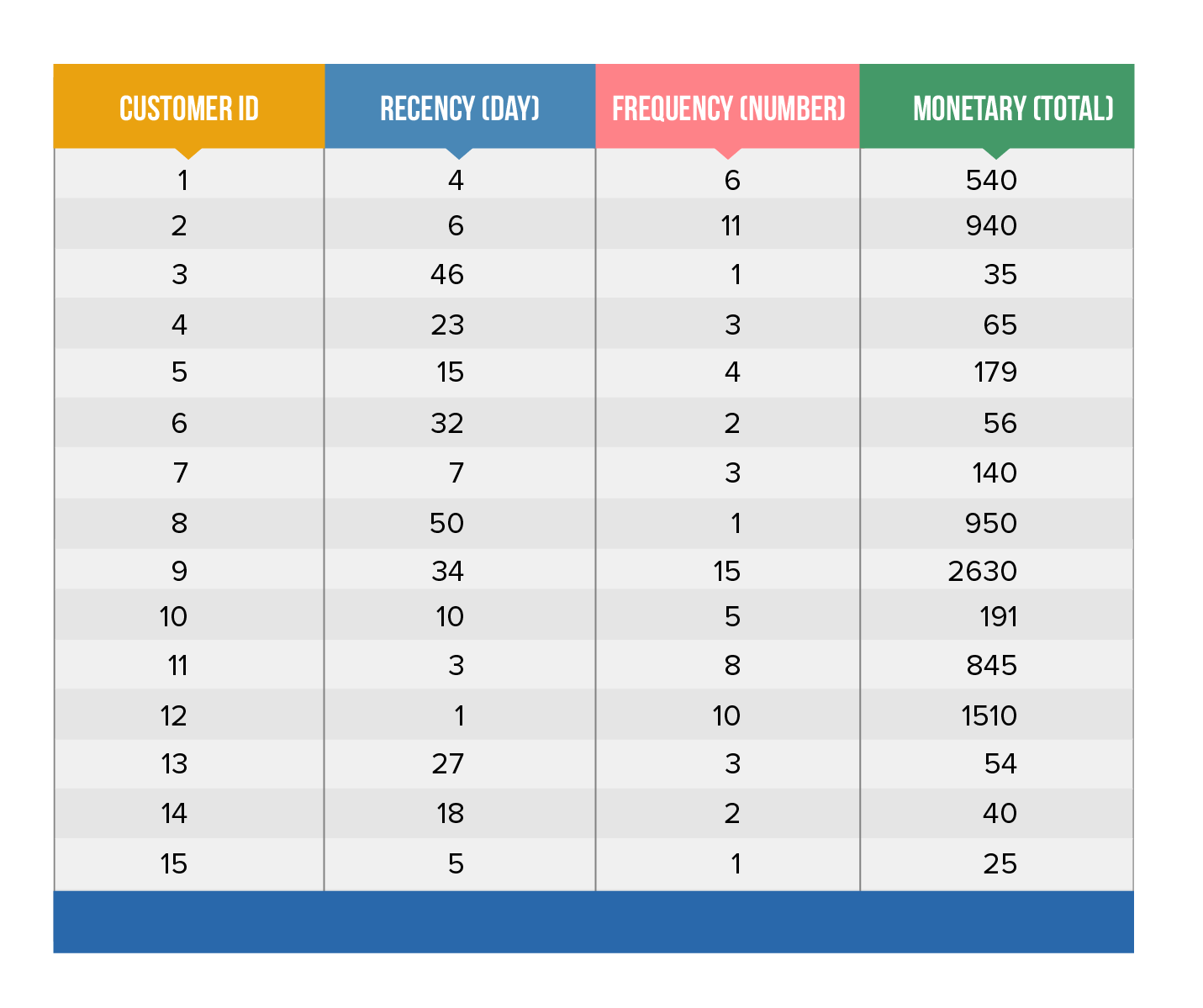
RFM stands for Recency, Frequency, and Monetary value, each corresponding to some key customer trait. These RFM metrics are important indicators of a customer’s behavior because frequency and monetary value affect Customer lifetime value and recency effects retention, a measure of engagement.



Businesses that lack the monetary aspect, like viewership, readership, or surfing-oriented products, could use Engagement parameters instead of Monetary factors. This results in using RFE – a variation of RFM. Furthermore, this Engagement parameter could be defined as a composite value based on metrics such as bounce rate, visit duration, number of pages visited, time spent per page, etc.

RFM factors illustrate these facts:

* the more recent the purchase, the more responsive the customer is to promotions
* the more frequently the customer buys, the more engaged and satisfied they are
* monetary value differentiates heavy spenders from low-value purchasers



##### *Table 1: Example Customer transactions dataset*

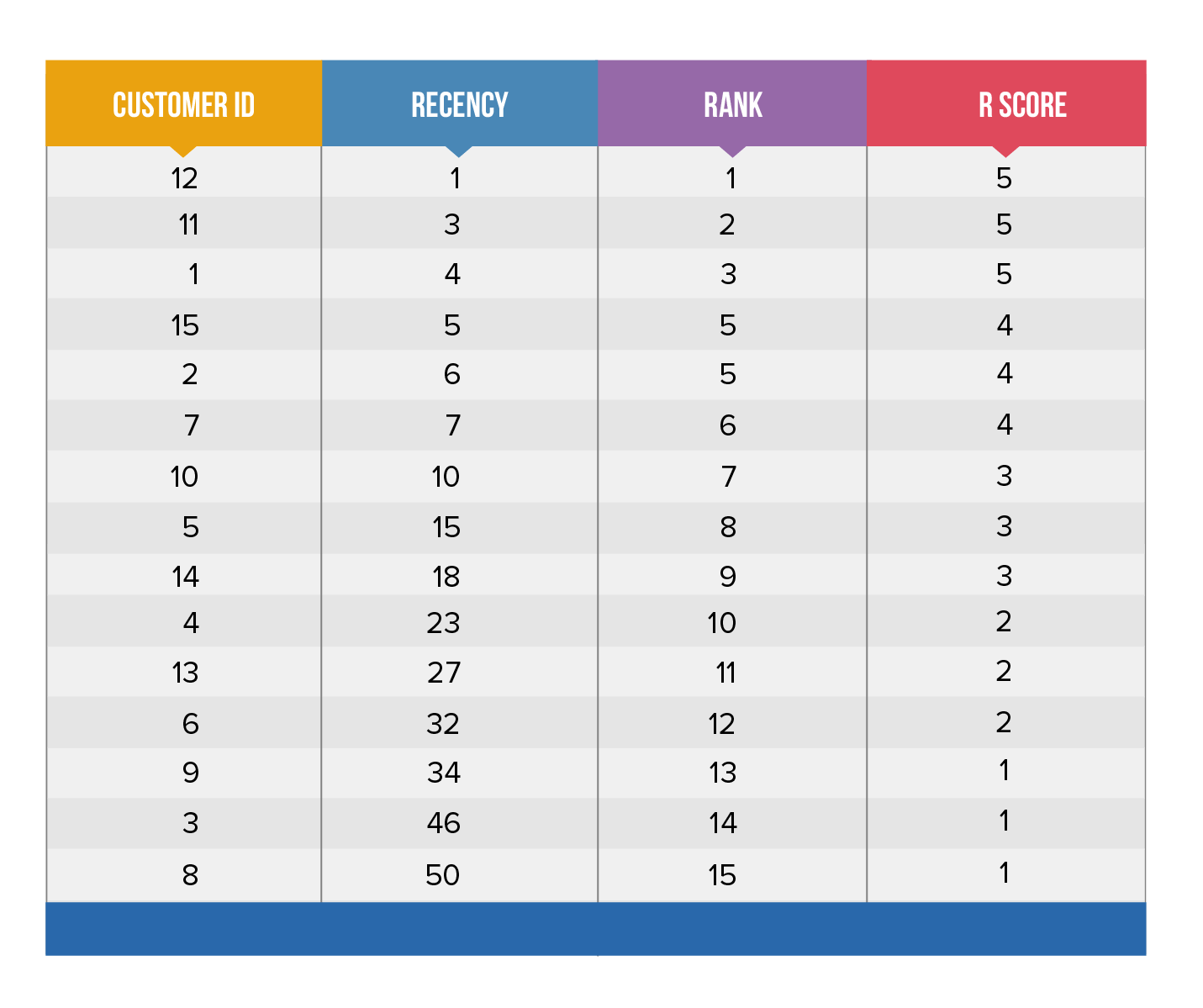
Table 1 contains recency, frequency, and monetary values for 15 customers based on their transactions.

## RFM Analysis Example

To conduct RFM analysis for this example, let’s see how we can score these customers by ranking them based on each RFM attribute separately.

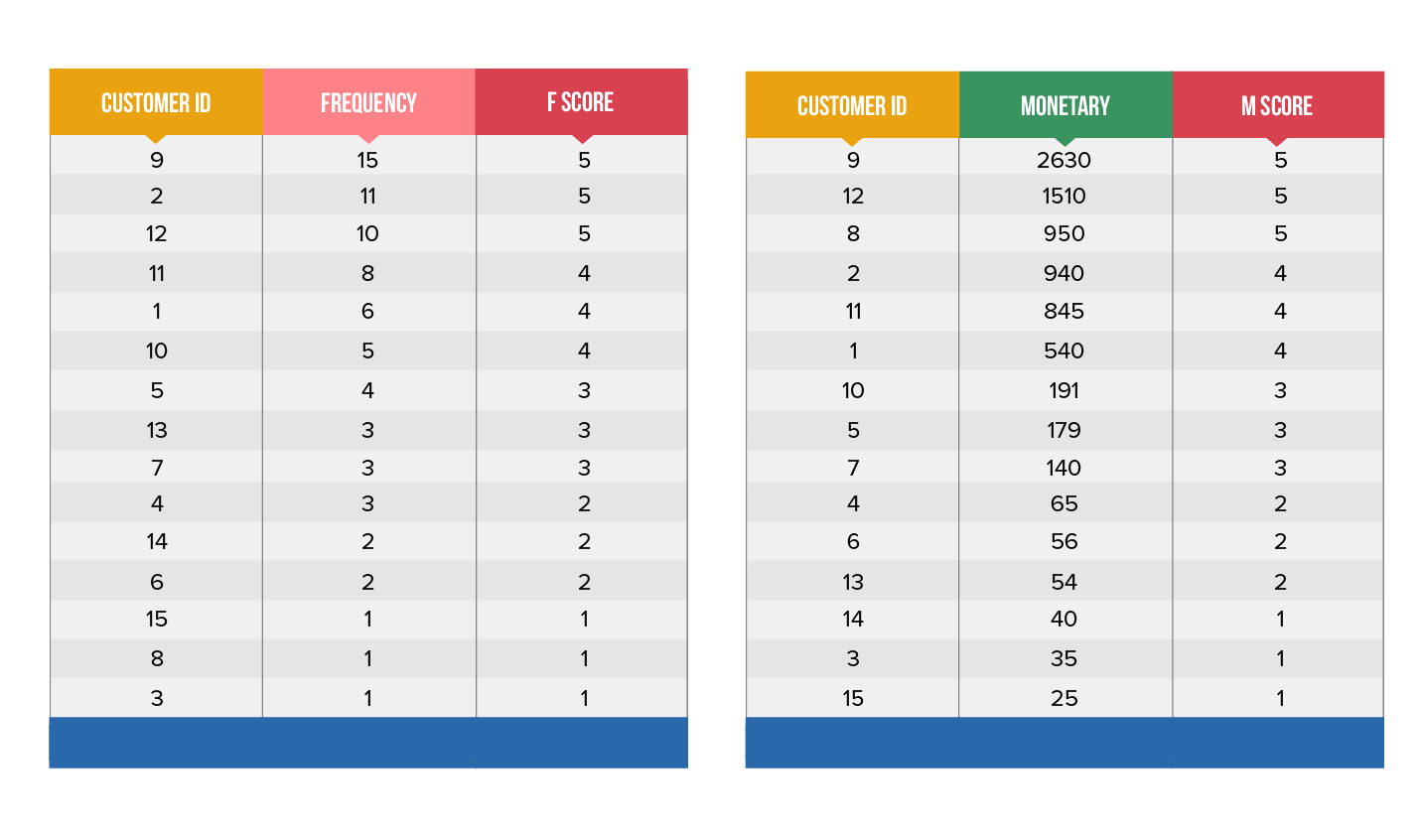
Assume that we rank these customers from 1-5 using RFM values.

Let’s begin with ranking customers on recency first, as shown in the below table:



As seen in the above table, we have sorted customers by recency, with the most recent purchasers at the top. Since customers are assigned scores from 1-5, the top 20% of customers (customer 12, 11, 1) receive a recency score of 5, the next 20% (the next 3 customers 15, 2, 7) a score of 4, and so on.

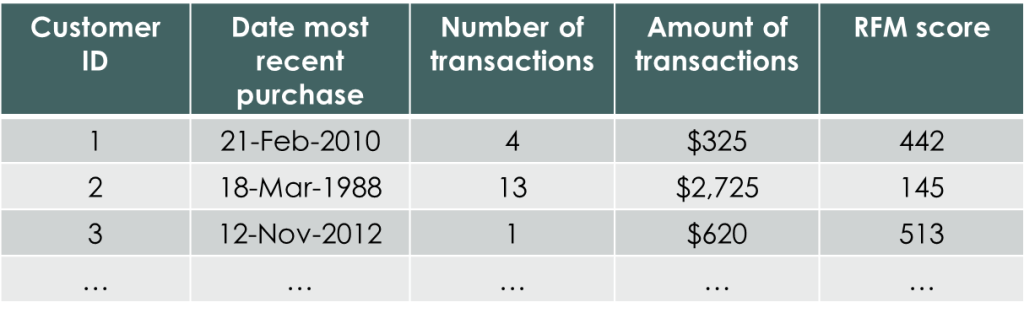
Similarly, we can then sort customers by frequency from most to least frequent, assigning the top 20% a frequency score of 5, etc. For the monetary factor, the top 20% of customers (big spenders) will be assigned a score of 5 and the lowest 20% a score of 1. These F and M scores are summarized below:

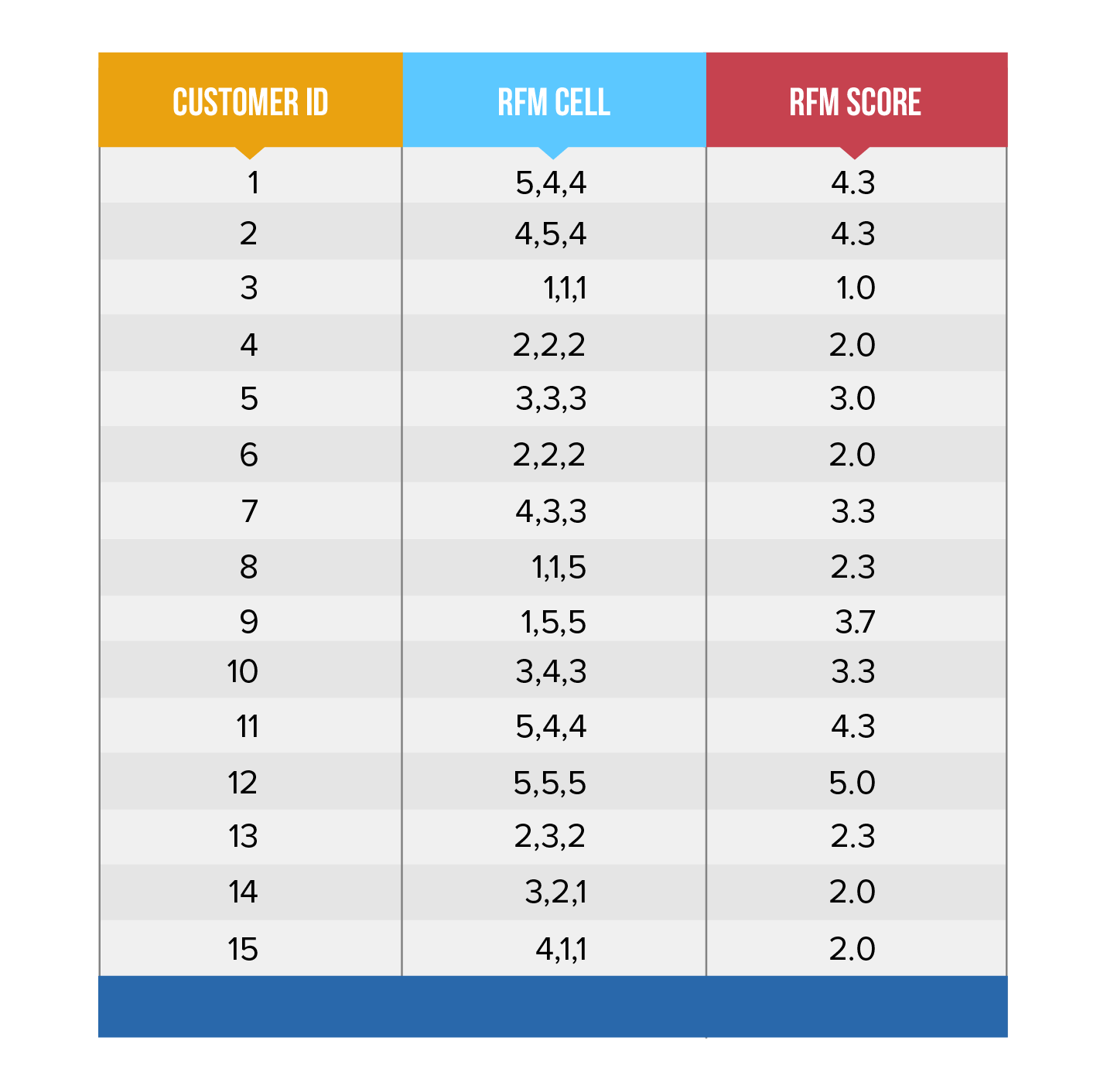


## RFM Score

Finally, we can rank these customers by combining their individual R, F, and M rankings to arrive at an aggregated RFM score. This RFM score, displayed in the table below, is simply the average of the individual R, F, and M scores, obtained by giving equal weights to each RFM attribute.

The RFM analysis assigns a 3 digit RFM score (from 111 thru 555) to each customer. A 555 indicates that a customer has purchased a product or service most recently, most frequently, and at the highest monetary value.





## Recency, Frequency, and Monetary Analysis

The next question that arises is: Is it fair to average out the individual R, F, and M scores for each customer and assign them to the RFM segment, as per their purchase or engagement behavior?

Depending on the nature of your businesses, you might increase or decrease the relative importance of each RFM variable to arrive at the final score. For example:

* In a consumer durables business, the monetary value per transaction is normally high but frequency and recency is low. For example, you can’t expect a customer to purchase a refrigerator or air conditioner on a monthly basis. In this case, a marketer could give more weight to monetary and recency aspects rather than the frequency aspect.
* In a retail business selling fashion/cosmetics, a customer who searches and purchases products every month will have a higher recency and frequency score than monetary score. Accordingly, the RFM score could be calculated by giving more weight to R and F scores than M.
* For content apps like Hotstar or Netflix, a binge watcher will have a longer session length than a mainstream consumer watching at regular intervals. For bingers, engagement and frequency could be given more importance than recency, and for mainstreamers, recency and frequency can be given higher weights than engagement to arrive at the RFE score.

This simple approach of scaling customers from 1-5 will result in, at the most, 125 different RFM scores (5x5x5), ranging from 111(lowest) to 555(highest). Each RFM cell will differ in size and vary from one another, in terms of the customer’s key habits, captured in the RFM score. Obviously, marketers can’t analyze all 125 segments individually if each RFM cell is considered a segment, and it’s difficult and overwhelming to visualize this imaginary 3D cube!

In general, the monetary aspect of RFM is viewed as an aggregation metric for summarizing transactions or aggregate visit length. Therefore, these 125 RFM segments are reduced to 25 segments by using just R and F scores.

## RFM Table

We use recency and frequency scores to visualize RFM analysis on a 2-dimensional graph. This enables users to consume and make sense of the scores more easily. Moreover, instead of creating 25 segments, we have combined a few segments to arrive at more manageable and intuitive segments.

As illustrated in the above RFM grid, we can get the following information for each of the segments:

* brief description of the segment
* recency (last activity)
* frequency (activity count)
* average monetary value
* reachability of users across different channels

Now, let’s discuss how to interpret the RFM segments to understand the behaviors of those users, and recommend some effective marketing strategies.

## RFM Segmentation

Let’s delve into few interesting segments:

* **Champions** are your best customers, who bought most recently, most often, and are heavy spenders. Reward these customers. They can become early adopters for new products and will help promote your brand.
* **Potential Loyalists** are your recent customers with average frequency and who spent a good amount. Offer membership or loyalty programs or recommend related products to upsell them and help them become your Loyalists or Champions.
* **New Customers** are your customers who have a high overall RFM score but are not frequent shoppers. Start building relationships with these customers by providing onboarding support and special offers to increase their visits.
* **At Risk Customers** are your customers who purchased often and spent big amounts, but haven’t purchased recently. Send them personalized reactivation campaigns to reconnect, and offer renewals and helpful products to encourage another purchase.
* **Can’t Lose Them** are customers who used to visit and purchase quite often, but haven’t been visiting recently. Bring them back with relevant promotions, and run surveys to find out what went wrong and avoid losing them to a competitor.

# Chapter-8 Proposed model 1: Two step cluster results

During the analysis, the number of clusters was not fixed to evaluate the clusters. That means the number of clusters determined automatically. We have chosen a log-likelihood method to measure the distance and Schwarz's Bayesian Criterion (BIC) as clustering criteria. There are three clusters evaluated by the results of two step cluster analysis. We named clusters as Bronze\*, Gold\* and Premium\* to ease comparison with the current segments and we have used \* sign to not confuse with the current segments. First cluster named as Bronze\* consisted of 279717 customers, 40% of all population. Customers of this cluster had scores below the overall mean for all indicators. Thus, we signed this cluster as R-, F-, and M-. Second cluster called as Gold\* (R+, F-, M-) which consisted377379 customers, 40% of all population. Mean of the recency scores of customers of this cluster had better than overall mean. But their shopping frequency and total expenses are below the mean. And customers of the third cluster which named Premium\* (R+, F+, M+) had better scores for all indicators. This cluster includes 42936 customers, 6,1% of all population. The results are summarized in Table 5.

Table 5: Model 1 Results

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Indicator | Overall  Mean | Bronze\* (40,0  R- F- M- | %) | Gold\* (53,9%) R+ F- M- | Premium\* (6,1%) R+ F+ M+ |
| R | **119,46** | **193,23** | **72,31** | | **53,25** |
| F | **1,9** | **1,49** | **1,57** | | **7,49** |
| M | **336,67** | **261,18** | **282,07** | | **1308,18** |
| Cluster Size (N) | | **279717** | **377379** | | **42936** |

Comparison between current customer segmentation and model 1 is shown in Table 6.

Table 6: Comparison between model 1 and current segmentation

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Bronze\* Gold\*  Premium\* | Bronze  **279717**  **377379**  **37551** | Card Type Gold  **0**  **0**  **4469** | Premium  **0**  **0**  **916** | Total  **279717**  **377379**  **42936** |
| Total | **694647** | **4469** | **916** | **700032** |

The company defined 694647 customers that should have a bronze card by considering just their expense. But according the results of proposed model 1, 54,3% of this customers (377379 customers) should be defined as gold member and 5,4% of this customers (37551 customers) should be defined as premium member. 4469 customers defined as gold member according to current profiling. But pursuant to proposed model 1, all of these customers should have premium card. 916 customers defined as premium member according to current profiling. They also should have premium card according to proposed model 1. But in proposed model 1 there are 42936 customers right to have premium card. If the company distributes card types according to proposed model 1, 60% of the customers should change their card type. In other words there are 40% similarity between current profiling and proposed profiling.

# Chapter-9 Proposed model 2: K-means clustering analysis results

It has been aimed to conduct K-means analysis to build clusters by considering the R, F and M indicators in proposed model 2. As mentioned before, the number of cluster should be defined in k-means method. Many values of k (2 to 8) has been tested and optimal solution have been evaluated for k=4. The clusters that obtained by the k-means clustering analysis have been entitled according to their RFM scores. First cluster named as *“Regular”* consisted of 644081customers, 92% of all population. Customers of this cluster had values below the overall mean for all indicators. It seems that, the member of this clusters are likely one time buyers. Their F value is almost 1. Second cluster called as *“Loyal”* which includes 514 customers. RFM values of customers of this cluster had better than overall mean. Customers of the third cluster which named *“Star”* had elegant scores for all indicators. The company has very few customers which have such RFM scores. This cluster includes just 97 customers. The individuals of fourth cluster which is called *“Advanced”* (55340 customers) also have better RFM score comparing with all population. But their RFM values are less than Loyal customers and Advanced customers such that their values are so close to average scores. The results are summarized in Table 7.

Table 7: Model 2 Results

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Indicator | Overall Mean | Regular | Loyal | Star | Advanced |
| R | **119,46** | 120,16 | 88,5 | 54,1 | 111,7 |
| F | **1,9** | 1,12 | 2,63 | 6,03 | 2,01 |
| M | **336,67** | 327,2 | 719,2 | 2823,2 | 439,1 |
| Cluster Size (N) | | 644081 | 514 | 97 | 55340 |

|  |  |  |  |
| --- | --- | --- | --- |
| loyal | 0 | 0 | 51 |
| star | 0 | 0 | 4 |
| advanced | 50544 | 4969 | 55434 |
| Total | 694647 | 4469 | 700032 |

Comparison between current customer segmentation and model 1 is shown in Table 8.

Table 8: Comparison between model 2 and current segmentation

As could be seen in Table 8, some of the bronze customers are placed in Regular group, while some of them are placed in Gold group. All of the current gold customers placed in Advanced group by model 2. And current premium customers disperse three clusters; Loyal, Star and Advanced. Main reason behind this is proposed model segmenting the customers according to three indicators; R, F and M while current segmentation segmenting the customers just according to their expense.

**Chapter-10 THE APPLICATION OF THE RFM MODEL**

The RFM model measures when people buy, how often they buy and how much they buy. While past purchases of customers can effectively predict their future purchase

behavior, firms can identify which customer is worthy to be contacted based on his or her past purchase behavior via K-means clustering technique to develop a CLV model by

determining customers’ CLV and segmentation by taking into account the RFM measures.

RFM model has also been widely used to identify customers and analyze customer profitability. For instance, Kaymak (2001) used RFM variables as features for

characterizing the customers when examining how fuzzy clustering can be used to obtain target selection models. The methods for target selection include segmentation

methods and scoring methods. In order to improve unreliable segmentation result due to the traditional adoption of general variables such as customer demographics and lifestyle to segment a market, Tsai and Chiu (2004) introduced a novel purchased-based market segmentation methodology in accordance with product specific variables (that is, the purchased items and associated monetary expenses from transactional

customer histories). After segmentation, they used a designated RFM model to analyze the relative profitability for each customer cluster, which helps provide more

marketing opportunities and aid marketers to revise their marketing strategies (Sohrabi and Khanlari, 2007). Jonker et al. (2006) provided a decision support system to determine mailing frequency for active customers such that direct mailers with tools can define the preferred response behavior and provide advices on the mailing strategy that can motivate customers towards this preferred response behavior. The system observes the mailing pattern of customers in terms of RFM variables and provides mailing policies for multiple time periods. As such, the mailing decision process is modeled through a Markov decision chain.

**The definition and scoring scheme of RFM model**

The RFM model is the most frequently adopted segmentation technique that comprises three measures (recency, frequency and monetary), which are combined

into a three-digit RFM cell code, covering five equal quintile (20% group). Among the three RFM measures, recency is often regarded as the most important one.

However, according to prior findings, RFM values are inclined to be firm-specific and are based on the nature of the products (Lumsden et al., 2008). For example, Fader

et al. (2005) found that for lower recency, customers with higher frequency tended to have lower future purchasing potential than those with lower pre-purchasing rates.

Lumsden et al. (2008) have similar findings that there are significant differences between groups across recency and frequency.

The process to quantify customer behavior via RFM model is as follows. First, sort the database by each dimension of RFM and then divide the customer list into

five equal segments. The method is known to have an exactly equal size. Different RFM quintiles have different response rates. For recency, customers are sorted by

purchase dates. Recency is commonly defined by the number of periods since the last purchase, which measures the interval between the most recent transaction

time and the analyzing time (days or months), that is, the lower the number of days, the higher the score of recency.

A customer having a high score of recency implies that he or she is more likely to make a repeat purchase. The top 20% segment is coded as 5, while the next 20% segment

is coded as 4 and so forth. Finally, the recency for each customer in the database is denoted by a number from 5 to 1 (Hughes, 1996; Kahan, 1998; Tsai and Chiu, 2004).

For frequency, the database is sorted by purchase frequency (the number of purchases) made in a certain time period. The definition of frequency is often simplified

to consider two states, including single and repeated purchases. The top quintile is assigned a value of 5 and the others are given the values of 4, 3, 2 and 1. However,

higher frequency score indicates greater customer loyalty. A customer having a high score of frequency implies that he or she has great demand for the product and is more

likely to purchase the products repeatedly. For monetary,customers are coded by the total amount of money spent during a specified period of time. The definition of

monetary is defined by the dollar value that the customer spent in this time period or by the average dollar amount per purchase or all purchases to date. Marcus (1998)

suggested that it is better to use the average purchase amount rather than the total accumulated purchase amount so as to reduce co-linearity of frequency and

monetary. Finally, all customers are presented by 555, 554,553, …, 111, which thus creates 125 (5×5×5) RFM cells.

Moreover, the best customer segment is 555, whereas the worst customer segment is 111. Based on the assigned RFM behavior scores, customers can be grouped into

segments and their profitability can be further analyzed (Bult and Wansbeek, 1995; Bitran and Mondschein, 1996; Miglautsch, 2000; Chang et al., 2010).

In the study of Miglautsch (2000), the RFM scorin. Once a customer’s total

frequency value is lower than the mean, a score of 2 is given to this customer. The process may be repeated more than two times. For monetary, five quintiles are still created and each has equal amounts of sales. In addition to use the value of each cell to judge whether the customer is valuable, some studies suggest that the

possible combinations of RFM can be obtained by assigning  or  based on the average R (F, M) value of a cluster being less than or greater than the overall average

R (F, M) value. In this case, 8 segments are created. The composite value of RFM is obtained via multiplying normalized RFM values of each customer and the weight

of RFM variables (Liu and Shih, 2005a, 2005b; Sohrabi

and Khanlari, 2007)

Firms can get much benefit from the adoption of RFM, encompassing increased response rates, lowered order cost and greater profit. In the application of RFM model,

each customer’s name and address needs to be assigned by a unique key (that is, an account number) and order, and the sales information needs to be stored with the

unique key included in each transaction record (Hughes, 1996; Kahan, 1998). The analysis of RFM is to examine customer transaction history, including an observation of

the purchasing time, purchasing frequency and purchasing monetary amount, and thus to help identify significant and valuable customers (Miglautsch, 2000,

2002). Customers can be classified into different types of groups via RFM. Thompson (2002) applied RFM model to classify customers into uncertain customers, spenders,

frequent customers and the best customers.

**Comparing RFM with other variables on other models**

Given the weakness of RFM models, some papers have attempted to improve the predictability of RFM models via adding more additional variables to predict customer

behavior or develop new models to test whether they perform better than RFM. For instance, Buckinx and Poel (2005) built a model to predict partial defection by

behaviorally loyal clients adopting three classification techniques: Logistic regression, Automatic Relevance Determination (ARD) neural networks and random forests

in a non-contractual setting and obtained data from an FMCG retailer. They used the observed past purchase behavior variables (including RFM variables) and

additional customer variables to predict partial churn behavior. Their findings revealed that past purchase behavior variables, particularly RFM variables are the best

predictors of parietal customer defection. Also, they confirm the importance of demographic variables and some additional variables such as the length of customer relationship, which are also useful to be incorporated in the attrition models. Suh et al. (1999) proposed RFM as a method that has a low correlation coefficient when combined with neural networks or logistic regression. The findings showed that the combined response model is superior to the single models when the correlation coefficient is low. However, the low correlation coefficient cannot assure improved performance.

Fader et al. (2005) proposed a model that links RFM with CLV by using iso-value curves so as to visualize the interactions and trade-offs among the linkage. Besides, in

order to generate valuable information on customer purchasing behavior, by using data collected via a retail chain in Taiwan, Chen et al. (2009) incorporated the

concept of RFM in the marketing literature to define the RFM sequential pattern and developed a novel algorithm:

RFM-Apriori for generating all RFM sequential patterns from customers’ purchasing data. Hosseini et al. (2010) proposed a procedure according to the expanded RFM model by adding one additional parameter, period of product activity, to classify customer product loyalty under B2B concept. The findings showed that the developed methodology for CRM produces better results than other commonly used models.

Yeh et al. (2008) introduced a comprehensive methodology to select targets for direct marketing from a database by extending the RFM model to RFMTC model by adding two parameters, namely: time since first purchase and churn probability. This model can estimate the probability that one customer will purchase at the next time and the expected value of the total number of times that the customer will purchase in the future.

The findings summarized that the proposed methodology provides more predictive accuracy than the RFM model.

**Criteria for Clustering**

To expedite the clustering process, the new customer data needed to undergo minor data preprocessing. In this step, certain customers were pruned from the dataset if they did not meet certain self-imposed constraints. At the particular dispensary studied, there were 15,489 unique customers as of July 10, 2019 that had spent a total of $3,743,454. However, only 4,975[47](#_3jtnz0s) (32.12%) of the customers were able to be clustered in the dataset. These 4,975 were chosen for meeting the following criteria:

The customer checked “Yes” to allowing their data be used for internal and marketing purposes

Their birthday and other necessary data had no malformed values. The birthday deserves special mention because some birthday inputs had only two digit years or months bigger than 12, which made it hard to absolutely determine their age. Less than 100 of the several thousand customers failed this criteria

The customer must have visited at least three times. This is to ensure that there is ample data collection and that time-based features, such as average time between visits, can be meaningful.

While there is only around a third of the original dataset remaining, the data is now dense enough to allow for customer preferences to truly appear. In turn, the results of the clustering will be much richer than having to include the whole dataset, provided that there is as little Survivorship Bias[48](#_1yyy98l) as possible.

Scaling and Reformatting Data

As mentioned previously, it is often a good idea to have data scaled between 0 and 1 before engaging in clustering. This is true because it prevents the distance formula from accumulating computationally taxing sums, since each term of the sum is between 0 and 1. Scaling data between 0 and 1 is relatively straightforward:

*featurei*

= *feature* − *min*(*feature*)

*max*(*feature*) − *min*(*feature)*

*(*MinMax Scale Formula)

Where *feature* is the feature that is becoming scaled. It is important to scale features rather than instances in this context because instances are the focus of comparison, not the features; in order words, we are clustering instances, not features.

While it would have been preferred that all data would work well with a simple MinMax scaling, one drawback of MinMax scaling is that it is very susceptible to outliers. Certain features such as total spent and average time between visits vary so widely across customers that a MinMax scale would not be adequate in the sense that it would not mitigate the variability in the data. Thus, it is often common to apply a log transform (or some other transform such as a power) to the data, then MinMax scale the transformed data instead. This achieves the ultimate purpose of scaling— to get the data between 0 and 1— while circumventing the issue of outliers.

Visits - Log2 transform, then MinMax scale

Total Spent - Log10 transform, then MinMax scale

Average Ticket - Log10 transform, then MinMax scale

Average Time between Visits - Log10 transform, then MinMax scale

All other features are already within a 0-1 range

Once the data passed the criteria for clustering and was scaled/reformatted, the data was then prepared for clustering. The following section describes these results and the implications of them.

**Performing Analysis and Results**

After the data was formatted in an appropriate way, it was time to begin clustering the data. While the clustering algorithms were implemented using sklearn — a popular, open-source Python data science library — there was significant coding needed to not only get to the point of clustering, but also recording the results in a reasonable manner. Hence, it is only proper to first provide an overview of the programming needed to create the workflow.

**Chapter-11 Brief Overview of Code**

The entirety of the code written for this project was in Python. The following Python packages played pivotal roles in the execution and development of this project:

pandas, numpy, sklearn.preprocessing, os, datetime, and time were all used for data collection, handling, and manipulation

seaborn, matplotlib.pyplot, and scipy.cluster.hierarchy were used to create visualizations of data

sklearn.cluster and sklearn.decomposition were used for clustering the data or decomposing it into three dimensions for plotting

In general, the dataflow consists of five separate steps. First, the raw data collected from the database is cleaned for malformed values, voided tickets, and items that were not sold. If necessary, this data can be saved and stored for future access. Next, the remaining data is turned into customer-based data. Each unique customer is initialized with their purchase data encapsulated by the features used for clustering. Additionally, customers that do not meet the criteria for clustering are pruned from the dataset, leaving only customer data that is able to be clustered. Once the customer data are formed, the data are then scaled and source packages such as sklearn, the clustering is extremely fast, regardless of the number of clusters chosen. The results of the clustering are saved into a variety of lo- cations based on the format of the results; results that involve labelling the raw data are moved into a separate location than the data that describes the structure of the clusters. Furthermore, some data on the runtime or other meta results of clustering were collected. Altogether, the dataflow, when done in its fullest form, takes around eight minutes to complete, which is far from optimal.

**Clustering Results**

K-Means

The first clustering performed in the dataflow was K-Means. Since K-Means requires a prespecified k to cluster, K-Means was run with ks from 1-25 to ensure an ample range for sufficient clustering to occur. Tied to this, each it- eration of clustering was run with randomly initialized centroids 100 times,with the best [51](#_4iylrwe) clustering chosen from each one.While typical examples of K-Means yield nice elbow curves, this clustering does not. There is no clear, discernible elbow point on the curve in figure [10](#_46r0co2) that indicates an optimal k. Some arguments can be made that one occurs around 6 or 7, but this is not conspicuous. Perhaps the optimal k is beyond the range, but that would likely involve results that are significantly overfitted or not actionable. Since the result of the optimal k is less obvious than desired, it was decided to explore clustering with five clusters. Five was chosen because it was small enough to be actionable but large enough to provide specific breakdowns within the clustering.

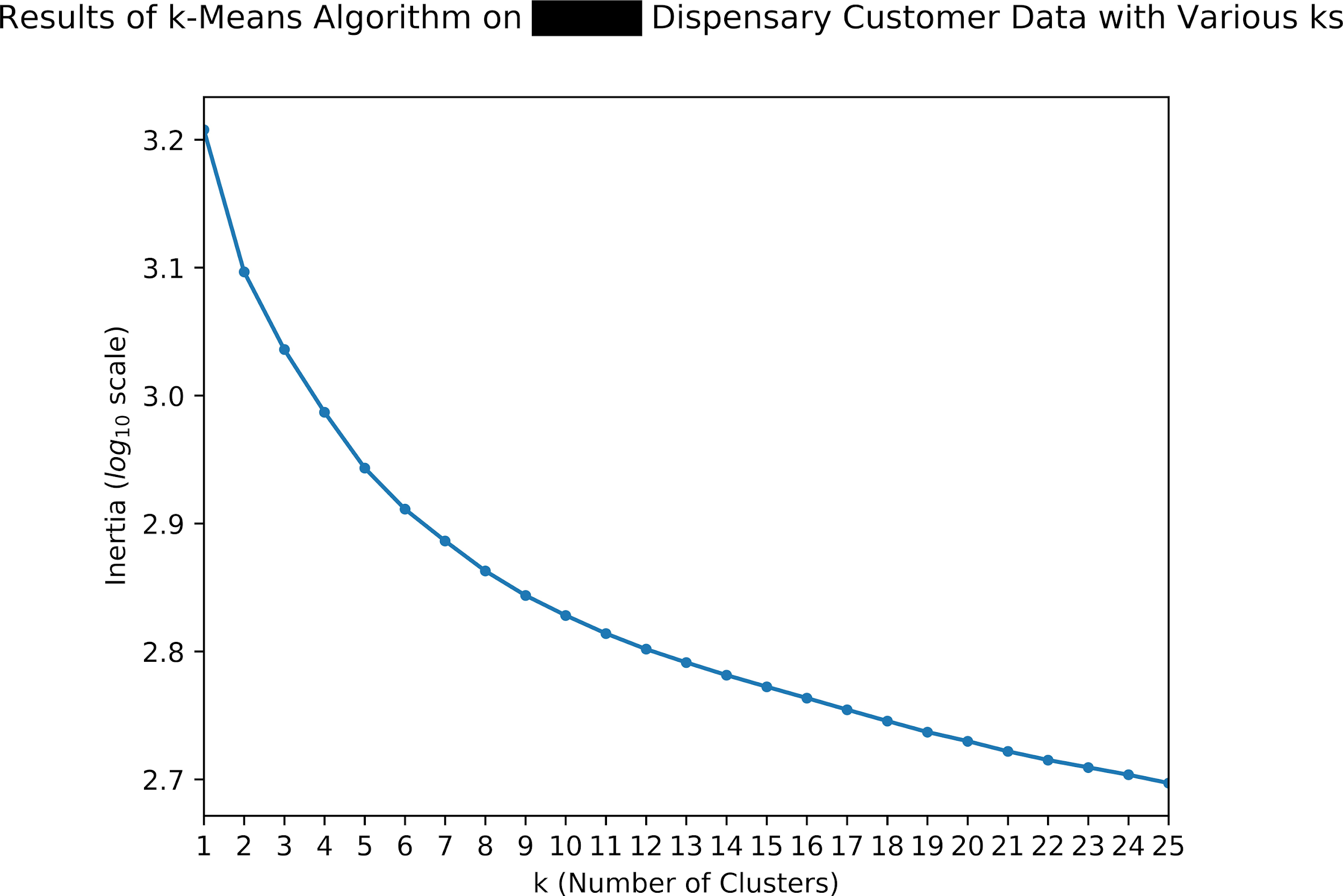


Figure 10: Inertia for Several Ks on 4Front (dispensary name redacted) Dis- pensary Data.

As a result, the statistics and means of each feature for the five clusters are reported in table [1.](#_2lwamvv)

Table [1](#_2lwamvv) reveals the natural segmentations of the customers. First off, the clusters are not as balanced as hoped, but are still close to within 10% of the expected (20%). The most popular clusters, clusters two and four, are the heaviest vape and flower consumers, respectively. In other words, these are the variables that most clearly delineate these clusters. While cluster two has much higher tickets and total spent, cluster four consumes far more prerolls than cluster two. Due to the low number of visits and high time between visits, cluster four are likely customers that do not shop consistently, but perhaps bulk.

Table 1: Results of K-Means Clustering with *k* = 5

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Feature/Cluster | 1 | 2 | 3 | 4 | 5 |
| Count | 588(11.82%) | 1,163(23.38%) | 755(15.18%) | 1,505(30.25%) | 964(19.38%) |
| Age | 32.7 | 35.7 | 35.9 | 36.0 | 37.1 |
| Visits | 7.45 | 6.92 | 32.56 | 5.21 | 6.23 |
| Total Spent | $453.86 | $542.37 | $1,733.12 | $270.75 | $334.57 |
| Ticket | $63.94 | $78.85 | $59.20 | $53.27 | $55.53 |
| Time between Visits | 46.59 | 45.84 | 12.49 | 50.75 | 44.97 |
| Flower | 20.57% | 19.39% | 52.76% | 70.63% | 25.46% |
| Vape | 12.63% | 56.54% | 12.09% | 5.99% | 13.07% |
| Preroll | 8.15% | 8.60% | 13.09% | 12.05% | 23.13% |
| Edible | 5.69% | 5.75% | 9.00% | 4.49% | 27.35% |
| Topical/Other | 0.28% | 0.53% | 0.52% | 0.43% | 1.53% |
| Concentrate | 48.82% | 4.57% | 6.89% | 3.18% | 4.82% |

up on cheap flower deals. On the other hand, customers within cluster two are spending more money more frequently, which makes sense given a 1-gram vape cartridge can be near $40 while one gram of flower is near $10 on average. These segments are worth highlighting not only because they are the largest in terms of size, but they are the base consumers of two of the most popular types of products at a dispensary: vapes and raw flower.

Despite consuming less vape and flower than clusters two and four, perhaps cluster three is the most intriguing cluster. Each of the other four clusters averages between 44 and 50 days between visits, but consumers in cluster three are visiting more than three times as often as the other clusters. And while their tickets are not significantly higher, their visits and thus total spent are the highest of any cluster. To make this cluster even more peculiar, this cluster also is not the highest in any of the cannabis-specific features. This suggests that they dabble in each of the product types offered at the dispensary. In a sense, this makes them the “connoisseur” cluster, which means that they are also likely the most loyal and educated customers. Their connoisseur character is also vastly similar to cluster five, which has more even balanced cannabis- specific features but is more defined by their edible consumption in addition to their low tickets.

Last but not least, concentrate consumption and age best separates cluster one from the rest of the clusters in the data. The low age and high concentrate consumption may suggest that cluster one represents the younger, newer cannabis consumers that lean towards dabbing concentrates rather than pur- chasing flowers or vapes such as in clusters two through five. In a similar sense to cluster two, cluster one has a high average ticket because the average price for gram of concentrate is much higher than flower.

Interestingly enough, this is also the smallest cluster by several hundred customers, so it may not be as influential to the total customer breakdown as the other clusters.

In sum, centroid-based K-Means yielded five clusters that were mostly dif- ferentiated by their purchasing behavior and cannabis-specific behavior rather than their demographics.

While clusters two and four are the largest clusters in terms of size, cluster three is the cluster with the highest total spend and the most visits.

Lastly, cluster one, the youngest cluster, consumes mostly concen- trates, cluster five consumes mostly edibles, clusters three and four consume mostly flowers and prerolls, and cluster two consumes mostly vapes.

**Managerial Implications of Results**

Tables [1](#_2lwamvv) and [2](#_4k668n3) describe important patterns and behaviors of the dispensary’s consumers, but there are crucial implications that accompany these results, particularly with regard to the structure of the clusters.

To begin, analysis of both algorithms indicates that the optimal number of clusters to choose is somewhere between five and six. With only five or six clusters, it is straight- forward to separate out the clusters to uncover patterns.

However, a lower number of clusters might be easier to separate, but the clusters will be far less informative and too general to make accurate predictions. This is often a problem with traditional customer segmentation analysis. Since there is motivation to keep analysis simple and not expand features unless necessary, traditional customer segmentation analysis often leads to oversimplification of the clusters and thus complicates managerial action.

Regardless, the decision to cluster with five or six segments is present throughout customer segmentation analysis research. In one sense, this implies that clustering cannabis retail data, even with cannabis-specific variables, may not be different from clustering other retail data. In turn, applying methods performed with other types of retail data to cannabis retail data is not only applicable, but perhaps even recommended as both the size and complexity of the data evolve.

Besides the difference in the number of clusters, the inherent purchasing behavior remains very much intact between both clusters. For example, clusters four and two in tables [1](#_2lwamvv) and [2](#_4k668n3) are almost identically the same size and have nearly equal values for their features; this is also the case with cluster two in table [1](#_2lwamvv) and cluster three in table [2.](#_4k668n3)

The parallelism between the two tables should not be entirely surprising, since they are clustering with the same data. Yet, the prevalence of the specific purchasing profiles in both tables reveals how easily detectable these profiles are from the given data.

When cluster profiles are consistent across numerous different clustering algorithms,

It indicates that the profiles are more than just quirks in the data: they are real signals that emerge from the noise. For managers, this translates to con- fidence that the profiles obtained from clustering algorithms are meaningful and actionable.

When the profiles obtained from customer segmentation analysis are mean- ingful, they can be converted into real marketing campaigns that target cus- tomers based on their purchase profiles, rather than traditional demographic profiles. One of the common problems with traditional demographic segmen- tation is that it is too simple to describe the convoluted purchasing nature of a retailer’s customers. When clustering data with variables that are domain- specific, there is a clear intention to uncover domain-specific patterns that are unattainable with RFM or demographic segmentation.

These uncovered pat- terns are invaluable to a retail company because they inherently communicate a retailer’s business strategy. If a retailer is able to find meaningful patterns within a clustering, the patterns can be used to form actionable customer pro- files.

These actionable profiles make it easy for retailers to trust the results of the clustering, which turns deployment into the quotidian business strat- egy. A strong trust between retailers and data scientists is integral to the success of any data-based retail project. In essence, it cannot be understated how using domain-specific data in clustering can form segmentations that are actionable and thus profitable.

While most customer segmentation analyses focused on traditional RFM analysis, it has been noted throughout this paper that there are stronger, more insightful ways to engineer relevant features for segmentation analysis.

How- ever, it is not useful to simply add features that are irrelevant for the sake of adding them. In order to come up with informative features, it is often necessary to have domain-specific knowledge to engineer proper features; the success of any machine learning project depends on the features available to it. So, deriving features is not a trivial task, but it is imperative that it is done correctly.

The combination of cannabis-specific domain knowledge and prac- tical skills of data science helped create features in this project that elegantly define each cluster. Yet, these features were not created ex nihilo: they came from taking a deeper look into the raw data that is already collected within the database

. Although many retailers and operators within the cannabis industry view the strict traceability as a burdensome necessity, it offers a goldmine of opportunity in terms of data. By encouraging deeper looks into the raw trans- actional data, retailers can develop more actionable and profound profiles of their customers and product.

**Chapter 12-Work,Conclusion and References**

**Possible Research Avenues or Expansions**

While there has been plenty of time devoted to discussions of customer segmen- tation analysis, machine learning, and the results of clustering with cannabis retail data, is time to revisit the end of the first section of the paper included a list of four goals that were important to accomplish for the paper to fully cover its scope. Goals one, two, and four were achieved in the first five sections, but goal three requires its own special attention. More specifically, there needs to be discussion into not only ways to improve the current project, but also ways to expand upon it or use its ideas or findings in other contexts.

As it stands, there are at least four visible improvements that could be made to the current project. First, perhaps most obviously, there should be more clustering algorithms used to fully understand the range of customer profiles and also the number of segments within the customer data. Although there was plenty of information gleaned from just two clusterings, different cluster algorithms can communicate additional findings or handle different sets of constraints. As mentioned previously, K-Means is the most popular clustering algorithm but it suffers from the key drawbacks of requiring the number of centroids to be established a priori in addition to requiring vari- ables to be numerical in nature. Although hierarchical clustering fixes this problem, it is not as scalable as K-Means and also assumes an inherent hier- archical structure of the data. Furthermore, neither of the algorithms provide a probability of an instance belonging to a particular group; they both simply classify the instances into clusters. An additional clustering algorithm that can provide probabilities of belonging to a cluster is called Gaussian Mixture Models. With a probability of cluster assignment, retailers can begin to look at clustering as less rigid of a process. Ultimately, this offers a more flexible approach to marketing and profiling rather than strict clustering algorithms such as the ones in this paper.

Indeed, exploring other clustering algorithms is worthwhile, but another important potential improvement is to enhance the data to make more precise clusters. While the project aimed to accomplish goals of traditional customer segmentation analysis, there was no variable to account for the recency of the customer, mostly because there was confusion over defining it. Since the dis- pensary had been open for less than two years at the time of the analysis, it proved difficult to come up with an objective way to measure the recency of a customer. The original thought was to, like the other variables, scale the number of days since last visit between 0 and 1, but this creates an incomprehensible structure: a higher value would mean less recent.

There were discussions of turning the recency into a categorical variable (have they visited within the last two months), but this would prove to be more trouble than it is worth because of the struggles—and sensitivities— that the relevant clustering algorithms have with categorical data.

So, the motivation to add a recency feature is justifiable, but the implementation of it is far more difficult than envisioned.

Another relevant implementation to the current project would be to add additional measures of cluster stability and validity.

Clustering, in its very essence, is a way to explore data; naturally, clustering projects tend to focus more on investigating the patterns that arise in the data rather than evaluat- ing rigorous loss or benefit metrics such as in supervised machine learning. As clustering research continues to expand, many data scientists have proposed a variety of measures or tools to address common concerns of clustering. Some common ways to evaluate clustering include computing the silhouette coeffi- cient, creating a proximity matrix, turning the clustering results into a decision tree and computing the entropy/purity, and calculating the inertia or SSE of the model.

While these measures do not tell the whole story, they can illu- minate the strengths or limitations of the present clustering architecture. In the end, this leads to a holistic comprehension of the data and results.

On a lesser note, the data collection process can be improved, not neces- sarily for results but for efficiency. The original code, while passable, struggled to prune the raw data in a timely fashion. After inspecting the bottlenecks of the code, it became clear that one of the issues involves updating a dataframe each iteration of cleaning rather than updating all at once; instead of updating the data frame, in its entirety, one time, the data flow now updates the entire data frame several thousand times, which is slower than it should be.

Fixing the runtime of a workable project should be the final update before publication or deployment, so

there is not urgent motivation to correct these bottlenecks as of now.

country.

Fine-tuning the analysis to one store is not optimal. Rather, the data flow should be as gen- eralizable as possible. On a practical level, this means discovering how to collect the same features from other POS systems, which may require a far more involved process than the one used in this project.

But once consistency in the dataflow is achieved, clustering customers in different markets may gen- erate advanced understanding of the economics of the area, or even the more general customer segmentation in the cannabis industry.

In short, expand- ing this analysis into each of the available markets gives 4Front a competitive advantage that can be leveraged to beat out competition.

Exploring the customer data at one point in time can provide enormous insight, but it is also possible to explore the evolutionary clustering of the customers. One method is to see how the size and optimal number of clus- ters evolve with time.

In particular, studying the evolution of the optimal number of clusters may hint at a deeper understanding of how customers nat- urally segment within a retail setting.

These results can then be compared to evolutionary cluster studies in other industries to answer a longstanding question within the cannabis industry: do cannabis customers act differently than customers in other retail industries? Simply taking a look at the results of evolutionary clustering is not only worthwhile for 4Front, it could be ground- breaking for the industry.

Lastly, clustering is not specific to customer segmentation analysis: it can be used in any segmentation project. With that said, clustering other relevant data within the retail database is also an expansion to the current project. Although the feature engineering process and dataflow would be altered, the general path of analysis would remain the same; convert the raw data into a usable format (e.g. day-based or product-based), scale and reformat the data appropriately, and implement the algorithms. Depending on the basis of the data, the results of the clustering uncover patterns not just within customers but also certain days of operation or even groups of products. As a side effect, this also promotes further discussion of machine learning in everyday retail analysis and therefore makes the concepts and practices of data science more easily understood in the business setting.

**Conclusion**

For the most part, the cannabis industry is in its nascent stages. The intense federal criminalization of cannabis for years totally hampered professional re- search into all facets of cannabis, from cultivation to retail to consumption. As a result, dispensaries are learning how to navigate not just a thicket of reg- ulations and other constraints, but also an unclear road of consumer behavior.

Conducting direct research with consumers and products is not possible, so retailers must look inward to uncover the behaviors of their customers. De- spite many retailers outside of cannabis have had tremendous success with traditional customer segmentation analysis, the supply of skilled analysts will- ing and capable to serve the cannabis industry is far smaller. To compound this, there is plenty of data in the cannabis industry— due to the enforcement of a traceability system— but few ways to access it. Although dashboards and elaborate interfaces have easened the responsibility of finding patterns or commonalities in retail data, none of them provides statistics or data that is rich enough to make advanced insights such as customer segmentations. As a result, it is necessary to bring in tools that are specifically built for situations such as this: machine learning.

With ample data, cannabis-specific domain knowledge, and a background in machine learning, developing a set of scripts to cluster the raw data was possible. After engineering relevant features and reformatting the data, it was possible to perform customer segmentation analysis with two different clus- tering algorithms: K-Means and Agglomerative. Even though the algorithms used different numbers of clusters in their clusterings, they essentially con- vey the same three pieces of information. First, flower and vape consumption were the defining characteristics of the largest clusters, which hints at the importance of these two products to a dispensary’s success. Second, both al- gorithms generated a cluster of ultra-frequent consumers, with average visits and total spent significantly higher than the rest of the clusters. Lastly, the tables also show that older consumers tend to enjoy edibles and topicals more than other consumers; on the flip side, younger consumers tend to enjoy vapes and concentrates more.

Regardless of the information provided, the results provide actionable ways for retailers to employ a marketing campaign or similar segmentation for their consumers. Despite the usefulness of the analysis as-is, there are numerous routes for improvement and growth. While there was motivation to keep the number of features low, adding a separate feature to account for the recency of the consumer would provide clearer details on whether certain purchase profiles are more common now than in the store’s past. On a similar note, finding ways to cluster a customer quicker (such as in one or two visits rather than three) could generate insights into not only the evolutionary aspect of the clustering but potentially also the leakage of customers. Finally, attempting the same analysis with numerous other clustering algorithms such as Gaussian Mixture Models or deep learning would bring about insight into the stability of cluster formation.

This project could not have been completed without the gracious wisdom of my thesis advisor, Dr. Richard Thompson, as well as the level-headedness of my boss, Joe Feltham, from 4Front Ventures. Finally, I would also like to extend a neverending gratitude towards my parents, Randall and Gina, and my partner, Rebekah: your love, motivation, and passion for my success is inspiring.

**References**

Bhatnagar, Amit; Ghose, S. (2004), ‘A latent class segmentation analysis of e-shoppers’, *Journal of Business Research* **57**, 758–767.

Chen, D., Sain, S. L. & Guo, K. (2012), ‘Data mining for the online retail in- dustry: A case study of rfm model-based customer segmentation using data mining’, *Journal of Database Marketing & Customer Strategy Management* **19**(3), 197–208.

**URL:** *https://doi.org/10.1057/dbm.2012.17*

Cooil, B., Aksoy, L. & Keiningham, T. L. (2008), ‘Approaches to customer segmentation’, *Journal of Relationship Marketing* **6**(3-4), 9–39.

Marcus, C. (1998), ‘A practical yet meaningful approach to customer segmen- tation approach to customer segmentation’, *Journal of Consumer Marketing* **15**, 494–504.

Mattson, M. P. (2014), ‘Superior pattern processing is the essence of the evolved human brain’, *Frontiers in Neuroscience* **8**, 265.

Morrison, C., Gruenewald, P. J., Freisthler, B., Ponicki, W. R. & Remer,

L. G. (2014), ‘The economic geography of medical cannabis dispensaries in california’, *International Journal of Drug Policy* **25**(3), 508 – 515.

**URL:** [*http://www.sciencedirect.com/science/article/pii/S0955395913002387*](http://www.sciencedirect.com/science/article/pii/S0955395913002387)

Rajaraman, A. & Ullman, J. D. (2011), *Mining of Massive Datasets*, Cam- bridge University Press, New York, NY, USA.

Rogers, S. & Girolami, M. (2016), *A First Course in Machine Learning, Second Edition*, Chapman & Hall/CRC.

Roux, M. (2018), ‘A comparative study of divisive and agglomerative hierar- chical clustering algorithms’, *Journal of Classification* **35**(2), 345–366.

**URL:** *https://doi.org/10.1007/s00357-018-9259-9*

Shepitsen, A., Gemmell, J., Mobasher, B. & Burke, R. (2008), Personalized recommendation in social tagging systems using hierarchical clustering, *in* ‘Proceedings of the 2008 ACM Conference on Recommender Systems’, Rec- Sys ’08, ACM, New York, NY, USA, pp. 259–266.

**URL:** [*http://doi.acm.org/10.1145/1454008.1454048*](http://doi.acm.org/10.1145/1454008.1454048)

Su, Q. & Chen, L. (2015), ‘A method for discovering clusters of e-commerce interest patterns using click-stream data’, *Electronic Commerce Research and Applications* **14**(1), 1 – 13.

**URL:** [*http://www.sciencedirect.com/science/article/pii/S1567422314000726*](http://www.sciencedirect.com/science/article/pii/S1567422314000726)

Tan, P.-N., Steinbach, M., Karpatne, A. & Kumar, V. (2018), *Introduction to Data Mining (2nd Edition)*, 2nd edn, Pearson.

Wagstaff, K., Cardie, C., Rogers, S. & Schr¨odl, S. (2001), Constrained k-means clustering with background knowledge, *in* ‘Proceedings of the Eighteenth International Conference on Machine Learning’, pp. 577–584.

Ward Jr., J. H. (1963), ‘Hierarchical grouping to optimize an objective func- tion’, *Journal of the American Statistical Association* **58**(301), 236–244.

Zhang, G. (2007), Customer segmentation based on survival character, *in* ‘2007 International Conference on Wireless Communications, Networking and Mo- bile Computing’, pp. 3391–3396.