CUSTOMER SEGMENTATION ANALYSIS OF MALL CUSTOMERS DATA: A MACHINE LEARNING APPROACH

Gaganajit Singh

A report submitted to The Exposys Data Labs
In fulfillment of the Internship Programme in Data Science

ABSTRACT

As the development of mall services progresses, retailers are increasingly motivated to seek data and strategies that can effectively segment or describe their customers in a concise yet informative manner. While many mall operators consider state-mandated traceability as necessary burden, it actually presents a valuable opportunity for internal customer analysis. Traditionally, segmentation analysis has primarily focused on demographic segmentation. However, these methods do not have the capability to provide insights into a customer's purchasing behavior. By employing machine learning algorithms such as K-Means and Agglomerative Hierarchial Clustering, new avenues for exploring a dispensary's consumer base have emerged . The results revealed the presence of approximately five or six customer clusters each characterized by unique purchasing traits. While these findings are important, further exploration of additional clustering algorithms comparing results across dispensaries within the same state in other state markets could enhance the value of this report.

Contents

List of Tables

1 Introduction

1.1 The Business Problem

Any company in retail, no matter the industry, ends up collecting, creating, and manipulating [1] 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 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 particular month? What are our top 5 products?* While dashboards supply data that often have important signifance in supply chain management and operations, they are limited in the sense that they omit data and insights that require higher level 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?* It is the latter half of the last question that will be the broad focus of this paper.

1.2 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.

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 analysis, 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 anonymitity 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[3].

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, an 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.

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 number of affected instances is small, it was crucial to handle these malformed instances because they prevented smooth analysis later on.

After taking the above processes and considerations into account, it was possible to collect the relevant data in a single query using the software's SQL editor. The data was then outputted into a CSV file (with around 250,000 rows) for easy viewing, importing, and analysis.

1.3 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 reccommender system, neither of these is the focus for this paper. The scope of the paper is limited to the following four inter- twined goals:

1. To cluster customers based on common purchasing behaviors for future operations/marketing projects
2. To incorporate best mathematical, visual, programming, and business practices into a thoughtful analysis that is understood across a variety of contexts and disciplines
3. To investigate how similar data and algorithms could be used in future data mining projects
4. 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.

2 Customer Segmentation Analysis

2.1 Brief Introduction

Customer Segmentation is a popular application of unsupervised learning. Using clustering, identify segments of customers to target the potential user base. They divide customers into groups according to common characteristics like gender, age, interests, and spending habits so they can market to each group effectively. 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.

2.2 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 peforming 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 skillset 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, retatilers 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.

3 Clustering Using Machine Learning 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 previous 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 understandings 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 tech- nique that groups instances into clusters based on the similarities between instances.

3.1 Centroid-based: K-Means

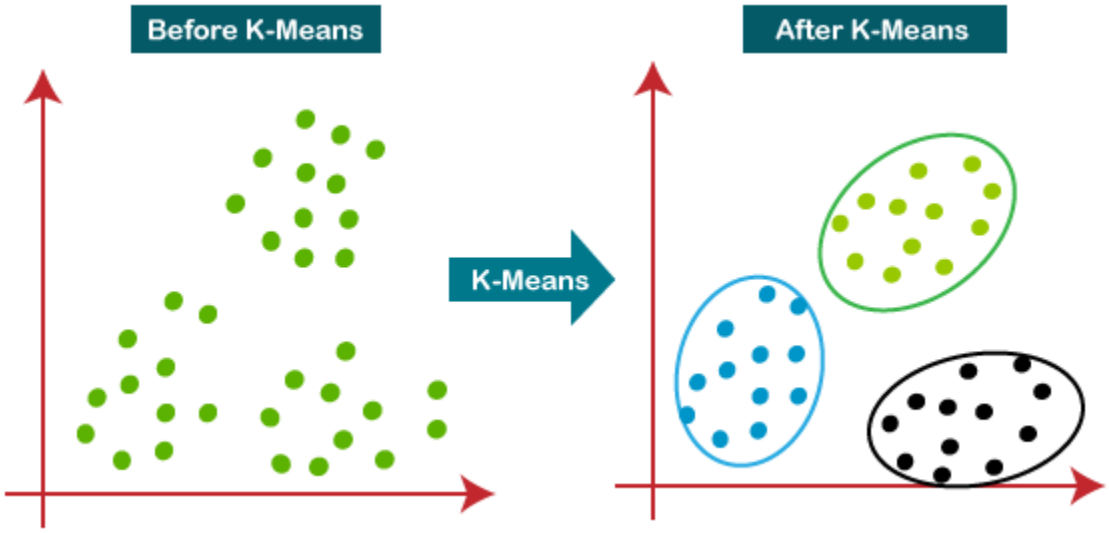
From the universe of unsupervised learning algorithms, K-means is probably the most recognized one. This algorithm has a clear objective: partition the data space in such a way so that data points within the same cluster are as similar as possible (intra-class similarity), while data points from different clusters are as dissimilar as possible

(inter-class similarity). In K-means, each cluster is represented by its center (called a "centroid"), which corresponds to the arithmetic mean of data points assigned to the cluster. A **centroid** is a data point that represents the center of the cluster (the mean), and it might not necessarily be a member of the dataset. This way, the algorithm works through an iterative process until each data point is closer to its own cluster's centroid than to other clusters' centroids, minimizing intra-cluster distance at each step. But how?

K-means searches for a predetermined number of clusters within an unlabelled dataset by using an iterative method to produce a final clustering based on the number of clusters defined by the user (represented by the variable K). For example, by setting "k" equal to 2, your dataset will be grouped in 2 clusters, while if you set "k" equal to 4 you will group the data in 4 clusters.

K-means triggers its process with arbitrarily chosen data points as proposed centroids of the groups and iteratively recalculates new centroids in order to converge to a final clustering of the data points. Specifically, the process works as follows:

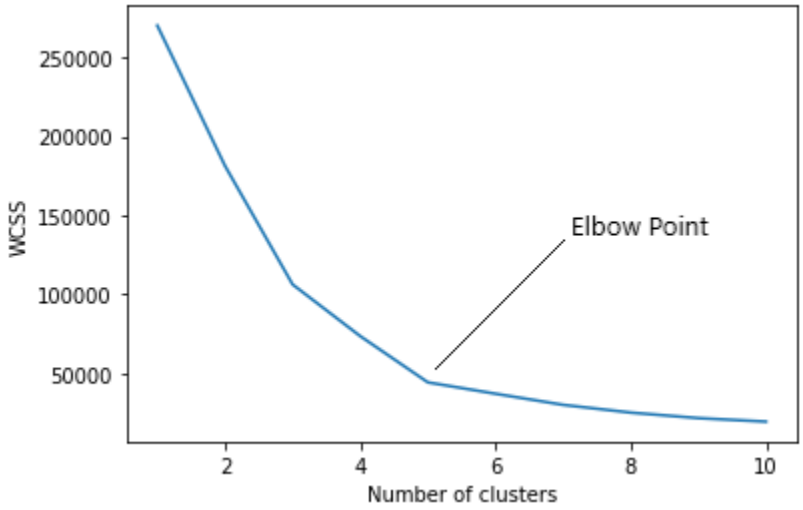
1. The algorithm randomly chooses a centroid for each cluster. For example, if we choose a "k" of 3, the algorithm randomly picks 3 centroids.
2. K-means assigns every data point in the dataset to the nearest centroid, meaning that a data point is considered to be in a particular cluster if it is closer to that cluster's centroid than any other centroid.
3. For every cluster, the algorithm recomputes the centroid by taking the average of all points in the cluster, reducing the total intra-cluster variance in relation to the previous step. Since the centroids change, the algorithm re-assigns the points to the closest centroid.



Finding the value of K

How do you choose the right value of “k”? When you define “k” you are telling the algorithm how many centroids you want, but how do you know how many clusters to produce?

One popular approach is testing different numbers of clusters and measuring the resulting Sum of Squared Errors (SSE), choosing the “k” value at which an increase will cause a very small decrease in the error sum, while a decrease will sharply increase the error sum. This point that defines the optimal number of clusters is known as the “elbow point”.



4 Preparing the Data

The digressions of clustering and customer segmentation analysis were important, 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[42]. 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.

```
In [5]: data
```

	CustomerID	Gender	Age	Annual Income (k$)	Spending Score (1-100)
0	1	Male	19	15	39
1	2	Male	21	15	81
2	3	Female	20	16	6
3	4	Female	23	16	77
4	5	Female	31	17	40
...	...	...	...	...	...
195	196	Female	35	120	79
196	197	Female	45	126	28
197	198	Male	32	126	74
198	199	Male	32	137	18
199	200	Male	30	137	83

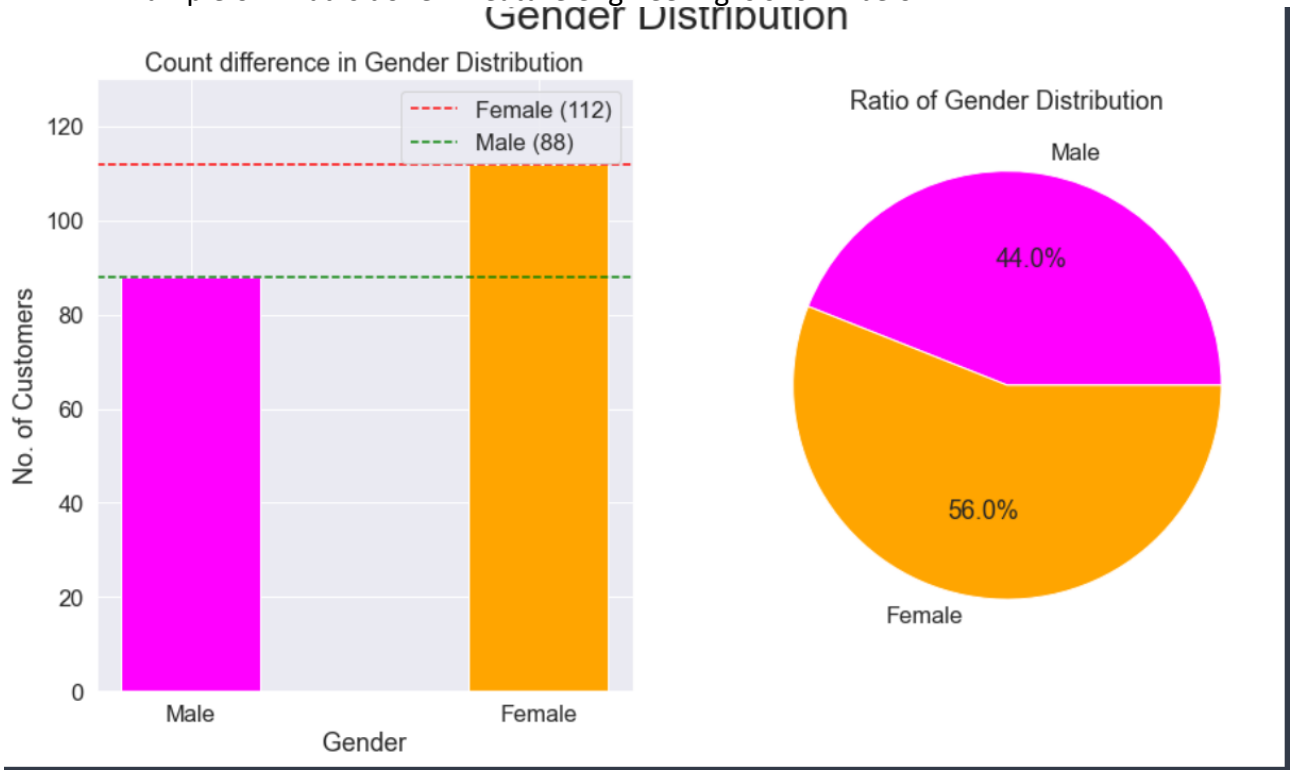
200 rows × 5 columns

4.1 Feature Engineering

The process of creating or extracting features from raw data is commonly referred 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.

Example of what is done in feature engineering is shown below:



4.2 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 200 unique customers, clustered in the dataset.

- 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.

4.3 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:

$$feature_i = \frac{feature_i - min(feature)}{max(feature) - min(feature)} \qquad \text{(MinMax Scale Formula)}$$

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. 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.

5 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 theworkflow.

5.1 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 cre- ate visualizations of data

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 [50]. 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 reformatted via the reformatting procedure laid out in earlier section.

The reformatted data are now ready to be clustered. Consequently, the next step is the clustering of the data with K-Means.

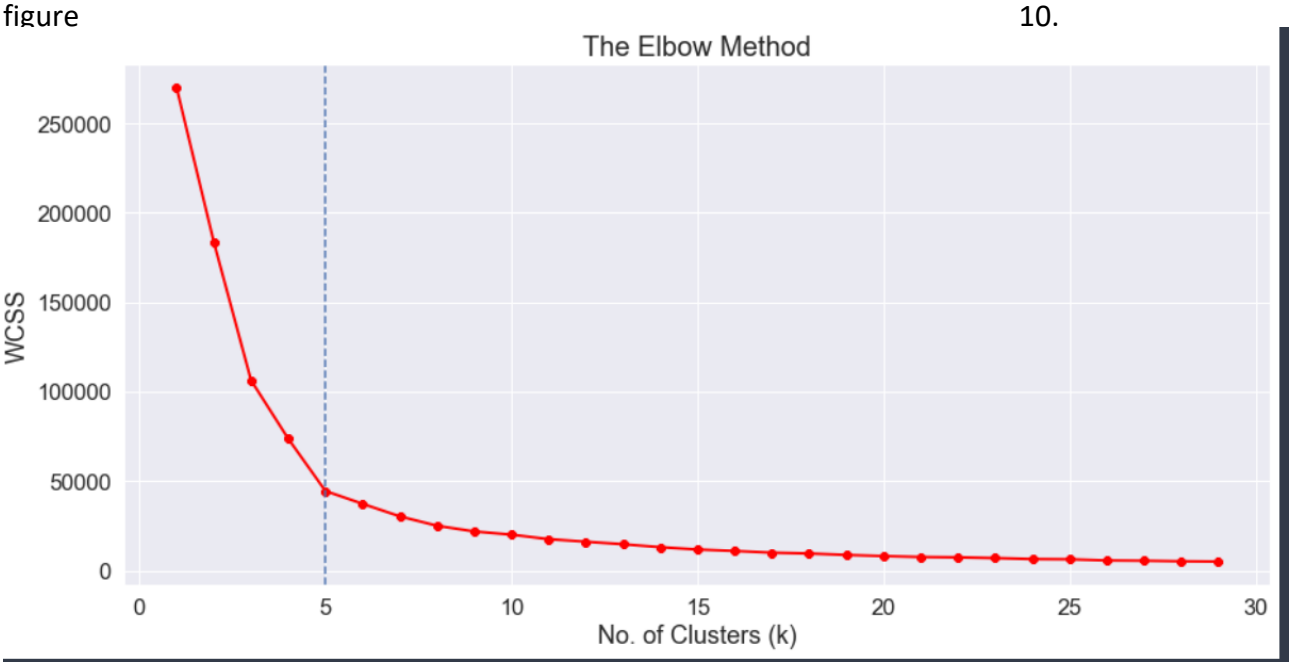
The results of the clustering are saved into a variety of lo- cations based on the

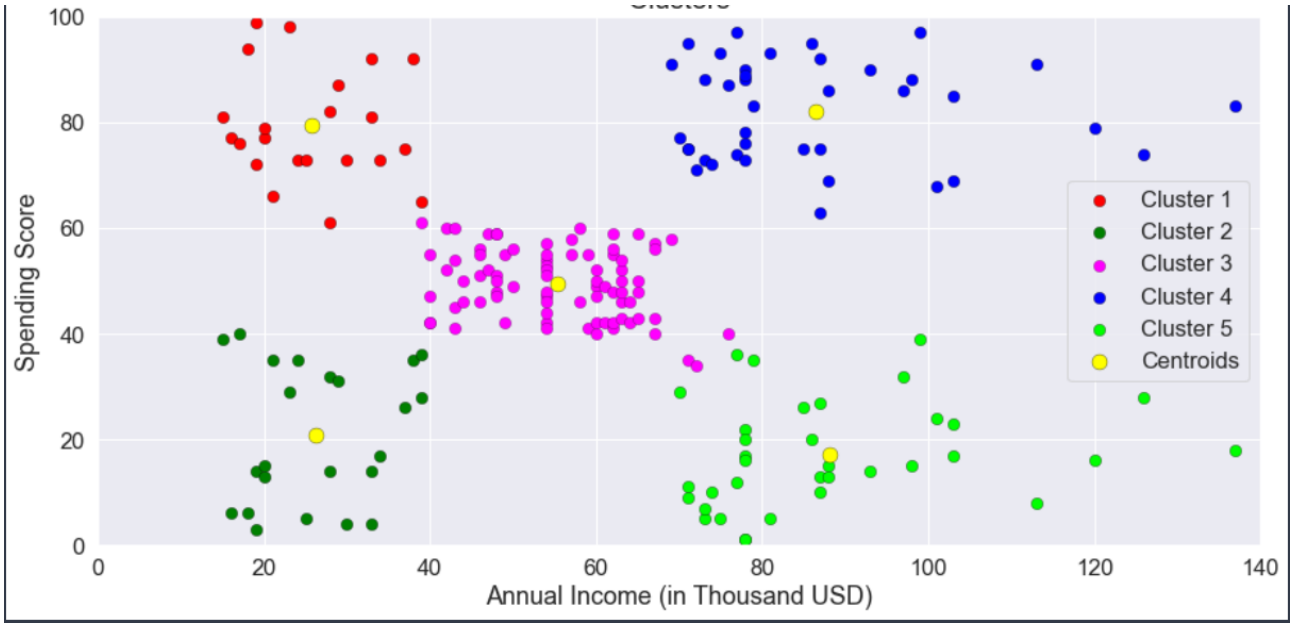
format of the results; results that involve labelling the raw data are moved into a separate location than the data that describes the structure of the clusters. Furthermore, some data on the runtime or other meta results of clustering were collected. Altogether, the dataflow, when done in its fullest form, takes around eight minutes to complete, which is far from optimal.

5.2 Clustering Results

5.2.1 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] clustering chosen from each one. The results of the clustering are summarized in figure 10.





5.3 Managerial Implications of Results

We have 5 clusters of customers or we can say 5 different types of customers.

* Customer Group 1: (Cluster Orange)

Customers in this group have annual income in the range of 80-140 but have spending score in the range of 0-140 So these customers might not be happy with our services. So we can try to add new facilities to attract these customers to increase our sales.

* Customer Group 2: (Cluster Blue)

These are pretty much Balanced customers having balanced Annual Income and Spending score so our but they have average annual income and can not be our prime customers if we are running a mall but still by giving additional features we can attract these customers.

* Customer Group 3: (Cluster Magenta)

This is group of customers earning less and also spending less which is quite obvious that they are spending wisely so they are not the targets of our mall business.

* Customer Group 4: (Cluster Red)

This is group of customers which likes to spend money on luxurious items so we will try to provide additional benefits/features for these customers because these group of people can be regular customers of our mall and our prime source of profit.

* Customer Group 5: (Cluster Green)

This group of customers spend more and earns less. It might be the reason that they are pretty much satisfied with mall services and for some reason love to buy the products from our mall and shops. So even without giving huge perks and benefits we can still attract these customers.