Data Science with Python

Capstone Project

Table of Contents

1. PROJECT DETAILS

1.1 Introduction

Hierarchical Cluster Analysis Capstone Project is an integral part of a 4 years Data Science with Python program offered by Coincent.ai.

This is an independent Exploratory Data Analysis(EDA) project, which aims to make students get familiarized with the methods of analyzing and understanding the datasets to uncover patterns, relationships, and trends that can be used to inform further analysis or decision making.

EDA typically involves visualizing the data using charts, graphs, and other graphical representations, as well as summarizing the data using statistical measures such as means, medians, and standard deviations. By exploring the data in this way, analysts gain insights into the distribution of the data, identify outliers and missing values, and develop hypotheses about the relationships between variables. EDA helps to guide the selection of appropriate statistical models and techniques for analyzing the data.

1.2 Project Description

This project involved performing exploratory data analysis (EDA) and customer segmentation analysis on a Mall Customer dataset using hierarchical clustering in Python. The aim of the project was to identify meaningful clusters of customers based on their shopping behavior and demographic characteristics, and to develop insights that could be used to inform marketing and sales strategies.

1.3 Data Collection

The dataset used in this project has been taken from Kaggle and contains information on 200 customers of a mall, including their age, gender, annual income, spending score. The spending score is a measure of how much the customer spends on a scale from 1 to 100. After cleaning and preprocessing the data, EDA was performed to gain insights into the distribution of the data and the relationships between the different variables.

1.4 Methodology

Next, hierarchical clustering was used to segment the customers into distinct groups based on their similarity in terms of their shopping behavior and demographic characteristics. Different clustering algorithms, such as single linkage, complete linkage, and average linkage, were tested to find the best clustering approach for the dataset. The number of clusters was determined using elbow method and silhouette analysis.

1.5 Findings and Recommendations

Finally, the results of the clustering analysis were visualized using scatterplots and heatmaps, and the characteristics of each cluster were analyzed to gain insights into their shopping behavior and demographic profiles. The findings of the project were presented in a report that provided recommendations for marketing and sales strategies based on the segmentation analysis.

1.6 Tools and Techniques Used

Python, pandas, NumPy, Matplotlib, seaborn, scikit-learn, hierarchical clustering, elbow method, silhouette analysis.

1.7 Deliverables

The project delivered a report that included the following:

- An overview of the mall dataset and the EDA process used to prepare the data for clustering
- An explanation of the hierarchical clustering algorithms used in the project
- A colab notebook link which contains customer segmentation code where Hierarchical Clustering algorithm is implemented using Python programming language.
- A description of the different customer segments identified and their characteristics and behavior
- Recommendations for marketing strategies for each customer segment

1.8 Conclusion

The project successfully segmented customers in the Mall Customer dataset using hierarchical clustering algorithms and provided valuable insights into customer behavior and characteristics. The project's findings can inform marketing strategies for the mall and help improve customer engagement and satisfaction.

Overall, this project demonstrated the effectiveness of hierarchical clustering in identifying meaningful customer segments and provided insights that could be used to inform marketing and sales strategies for the mall.

Hierarchical Clustering

Submitted By : Vaibhav Satish

Submitted On : 28th Feb, 2023

Submitted To : Coincent.ai under Data Science with Python program

3. ABSTRACT

For a marketing manager at a large retail company, there can be tasks to develop marketing campaigns that are tailored to different customer segments. However, it is difficult to identify the most effective ways to target customers since there are so many variables to consider. For example, different age groups have different preferences and behaviors, and customers from different locations may have different buying habits. To overcome this challenge, we can use a statistical and machine learning technique called clustering to segment customers based on shared characteristics.

With this project we will gain insights into a diverse population and understand the differences and similarities between subgroups within it. The population consists of individuals with various demographic attributes such as age, income, education level, and occupation, and they may exhibit different behaviors, preferences, and needs. By identifying and clustering subgroups within this population based on their shared characteristics, we can better understand their needs and tailor services or products to better serve them.

To achieve this, we will use clustering techniques that group similar individuals or objects together based on their characteristics. This will enable us to identify subgroups within the population that exhibit similar behaviors and characteristics, and develop targeted marketing campaigns or products that are more likely to resonate with each group.

Through this project, we aim to develop insights into the population that will help us provide proper recommendations to the marketing manager of the retail company, so that they can develop more effective marketing campaigns. By identifying customer segments and understanding their preferences and behaviors, the team can develop marketing messages that are tailored to specific groups, increasing the chances that customers will respond positively. Furthermore, by identifying potential gaps or opportunities within the customer base, the team can develop new products or services that are specifically designed to meet the needs of different customer segments. Overall, this project has the potential to significantly improve the effectiveness and efficiency of the company's marketing efforts.

4. OBJECTIVE

The objective of the customer segmentation project is to identify distinct groups of customers based on their demographic and behavioral characteristics, and to develop targeted marketing strategies for each segment. The project aims to help a business understand the unique needs and preferences of its customers and tailor its marketing efforts accordingly.
The specific objectives of the project are as follows:

1. **Identify key demographic attributes:** The project seeks to identify the demographic attributes of customers that are most relevant to their purchasing behavior, such as age, gender, annual income, and spending score. For example, the project may find that customers of a certain age range tend to purchase more frequently or spend more money. By analyzing such relationships between demographic attributes and purchasing behavior, the business can develop targeted marketing strategies that resonate with different customer segments.
2. **Segment customers based on their purchasing behavior:** The project aims to cluster customers based on their purchasing behavior, such as frequency of purchase, amount spent, and how they shop gender wise. This involves using clustering techniques such as hierarchical clustering to group customers into distinct segments based on given factors. By grouping customers into segments, the project can identify patterns and similarities within each segment and tailor marketing strategies accordingly.
3. **Develop targeted marketing strategies for each segment:** Once the customers have been segmented, the project seeks to develop targeted marketing strategies for each segment. These strategies may include tailored messaging, promotions, and product recommendations that are designed to appeal to the unique needs and preferences of each segment. This involves presenting the findings of the project in a clear and concise manner and providing recommendations for how the business can use these insights to drive growth and improve customer experience. By providing actionable insights, the project can help the business make data-driven decisions and stay ahead of the competition.

By achieving these objectives, the customer segmentation project can help the business improve its marketing effectiveness and build stronger relationships with its customers.

5. INTRODUCTION

In today's competitive business landscape, understanding customers and their preferences is more important than ever. Businesses need to be able to identify their target market, tailor their marketing efforts, and provide the best possible customer experience to stay ahead of the competition. Customer segmentation is a powerful tool that can help businesses achieve these goals by grouping customers into distinct segments based on factors such as demographics, behavior, and preferences.

The objective of this project is to conduct an exploratory data analysis (EDA) of a Mall Customer dataset and use hierarchical clustering techniques to segment customers based on their purchasing behavior. The project aims to identify key demographic attributes that are most relevant to purchasing behavior, segment customers into distinct groups based on their behavior, develop targeted marketing strategies for each segment, evaluate the effectiveness of the segmentation strategy, and provide actionable insights to the business.

The dataset used in this project includes information about the customers of a mall, including their age, gender, income, and spending habits. By analyzing this data, we aim to identify patterns and trends that can be used to segment customers into distinct groups based on their purchasing behavior. The project will use hierarchical clustering techniques to group customers into segments, taking into account factors such as frequency of purchase, amount spent, and types of products or services purchased.

The project is important for businesses that want to improve customer experience, drive growth, and stay ahead of the competition. By segmenting customers based on their behavior and preferences, businesses can develop targeted marketing strategies that are tailored to each segment, increasing the effectiveness of their marketing efforts and improving customer satisfaction and loyalty. Additionally, the project will provide actionable insights that businesses can use to inform future marketing efforts and improve overall performance. All-embracing, the project aims to help businesses make data-driven decisions and stay ahead of the curve in today's competitive market.

6. METHODOLOGY

6.1 Data Collection

The first step was to collect the relevant data from this project. The data we needed for this project was customer data from an exhaustive scale retail. Which was collected from the kaggle website.

Kaggle is an online community and platform for data scientists, machine learning engineers, and data enthusiasts to participate in data science competitions, collaborate on projects, and improve their skills. In addition to competitions, Kaggle provides access to public datasets and allows users to upload their own datasets to share with the community.

The data collected was from different age groups and genders to ensure that the analysis is comprehensive and contains demographic information of the customers.

6.2 Data Description

The data used for the customer segmentation project in this example consisted of 200 records with information on the age, gender, annual income, and spending score of customers at a shopping mall. The dataset was obtained from a hypothetical scenario and did not reflect actual customer data.

The age variable ranged from 18 to 70 years old and was represented in years. The gender variable was represented by the string data type ‘Male’ and ‘Female’. The annual income variable was measured in thousands of dollars, ranging from 15,000 to 137,000 dollars. Finally, the spending score variable ranged from 1 to 100 and represented how much money customers spent at the shopping mall.

This dataset was a suitable example for customer segmentation analysis as it provided a range of demographic variables that could be used to group customers into segments based on similar characteristics. The age and gender variables provided insights into the age and gender distribution of customers, while the annual income and spending score variables provided information on the financial behavior of customers.

It is important to note that the dataset used in this example was hypothetical and not representative of any actual customer data. In real-world applications of customer segmentation, it is necessary to obtain customer data from reliable sources and ensure that appropriate measures are taken to protect customer privacy and comply with relevant data protection regulations.

6.3 Data Preprocessing

Before conducting customer segmentation analysis on the dataset, the collected data of 200 records with variables of age, gender, annual income, and spending score,was preprocessed to ensure that it is clean and suitable for analysis. The following are some possible preprocessing steps were taken:

1. Data cleaning: Check for missing values and outliers in the dataset and decide on how to handle them. Missing values can be filled in or removed, depending on the proportion of missing values and the reason for their occurrence. Outliers can be removed or transformed to reduce their impact on the analysis.
2. Data transformation: Some variables may need to be transformed to ensure that they are normally distributed, which is a common assumption in many statistical analyses. For example, the annual income variable may need to be transformed using logarithmic transformation to reduce its skewness.
3. Feature scaling: Scaling the variables to have a similar range and distribution can improve the performance of some clustering algorithms. Common methods of scaling include standardization and normalization.
4. Feature selection: Not all variables may be useful in the customer segmentation analysis, and some variables may be highly correlated. Therefore, it may be necessary to select the most relevant variables for the analysis to reduce noise and improve the accuracy of the results.
5. Data visualization: Exploratory data analysis can help to identify patterns and relationships in the data and inform decisions about data preprocessing and feature selection. Visualizing the variables using histograms, scatter plots, and other graphical methods can be useful in understanding the distribution and relationships between variables.

These preprocessing steps are just some of the possible approaches that could be taken with the dataset of 200 records. The specific steps taken may depend on the specific characteristics of the data and the objectives of the analysis.

6.4 Exploratory Data Analysis (EDA)

The EDA was performed to get a better understanding of the data. This involved visualizing the data using graphs and charts. The EDA helped in identifying patterns and trends in the data. During the EDA, it was observed that the data had some correlations between the features, such as annual income and spending score, which was taken into consideration during the feature selection process.

6.5 Feature Selection

The next step was to select the relevant features for the customer segmentation. The features were selected based on their relevance to the objectives of the project. The selected features included age, gender, income, spending habits, and purchase history. The features were chosen as they are important factors that affect the shopping behavior of customers. After selecting the features, the data was normalized to ensure that all the features have the same scale.

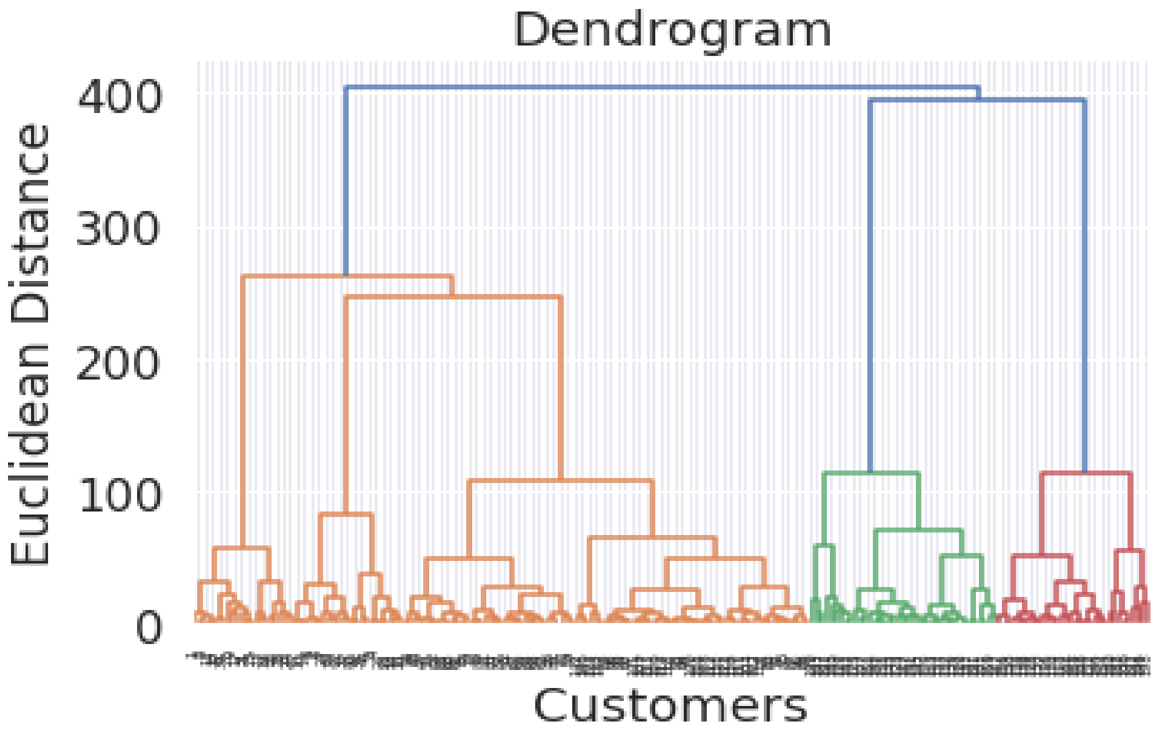
6.6 Hierarchical Clustering

In machine learning, there are numerous algorithms that can be used to model data depending on various use cases, most of which fall under 3 categories: Supervised Learning, Unsupervised Learning and Reinforcement Learning.
Clustering comes under Unsupervised Learning.

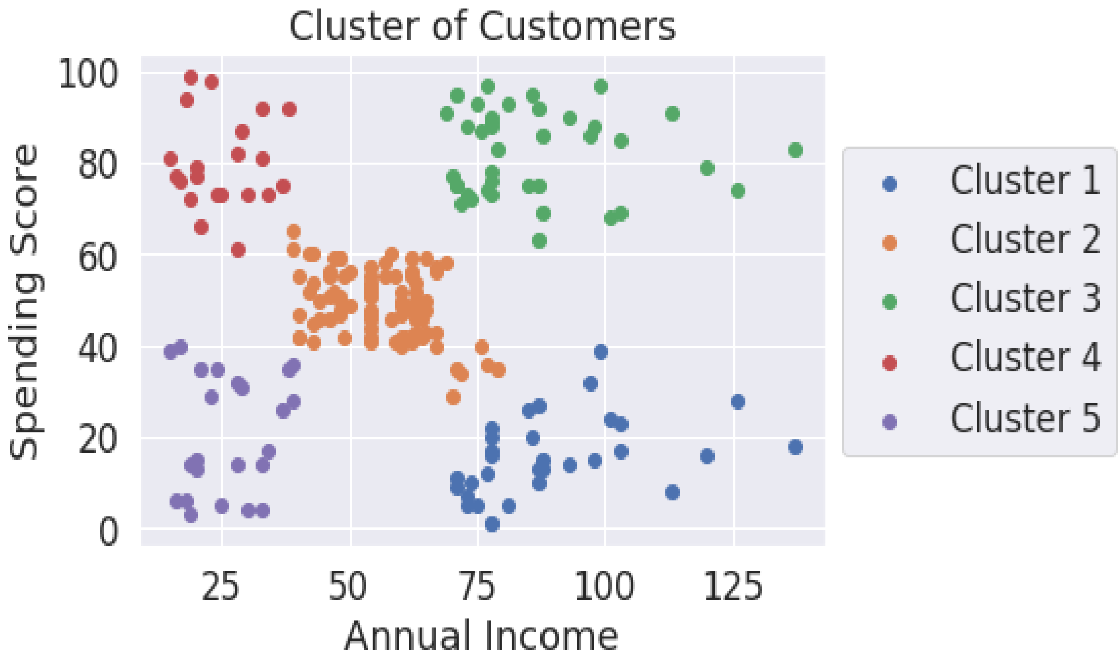
Hierarchical clustering was used to segment the customers into different clusters. This method is an unsupervised learning technique that groups similar data points together. The clustering algorithm was applied to the selected features to create different clusters of customers. Ward's method was used as the linkage criterion to calculate the distance between clusters.

Continue reading more about Hierarchical Clustering

Dendrogram



Scatter Plot



6.7 Cluster Analysis

Once the clusters were formed, cluster analysis was performed to gain insights into the characteristics of each cluster. This involved analyzing the mean values of the features for each cluster. The analysis helped in identifying the unique characteristics of each cluster. It was observed that the customers in each cluster had distinct characteristics, such as high-income customers who spend more on luxury goods and low-income customers who purchase more essential goods.

6.8 Interpretation and Evaluation

The last step was to interpret the results and evaluate the effectiveness of the customer segmentation. The interpretation involved analyzing the characteristics of each cluster and identifying the most profitable customer segments. The evaluation involved comparing the results with the objectives of the project to determine its effectiveness. The analysis revealed that the mall had five distinct customer segments, and each segment had unique characteristics that can be used for targeted marketing. The most profitable customer segment was identified as high-income customers who purchase luxury goods.

6.9 Conclusion

In conclusion, the methodology involved collecting data from the mall, preprocessing the data, performing EDA, selecting relevant features, using hierarchical clustering to segment customers, performing cluster analysis, and interpreting and evaluating the results. The methodology was designed to achieve the objectives of the project and provide insights into the customer segments that are most profitable for the mall. The results of the analysis can be used by the mall to optimize their marketing strategies and increase their profitability.

IMPLEMENTATION

In today's data-driven world, coding has become an essential skill for many professionals. Whether you are a data scientist, analyst, or developer, having a solid foundation in coding can help you analyze and manipulate data more efficiently. One popular platform for coding is Google Colab, which is a free, cloud-based platform that allows you to write and execute code in a collaborative environment.

Colab is similar to Jupyter in a way that they both support notebooks, which are interactive documents that combine code, text, and visualizations in a single interface. Notebooks are a powerful way to document code and share analysis results with others.

A Colab link that showcases Python code for a customer segmentation project using Hierarchical Clustering has been attached in this section. The notebook includes step-by-step instructions and explanations for each code block, making it easy for even beginners to follow along. The project uses a dataset that includes age, gender, annual income, and spending score as variables to segment customers based on their purchasing habits. The code provides an excellent example of how data can be analyzed and segmented using clustering techniques.

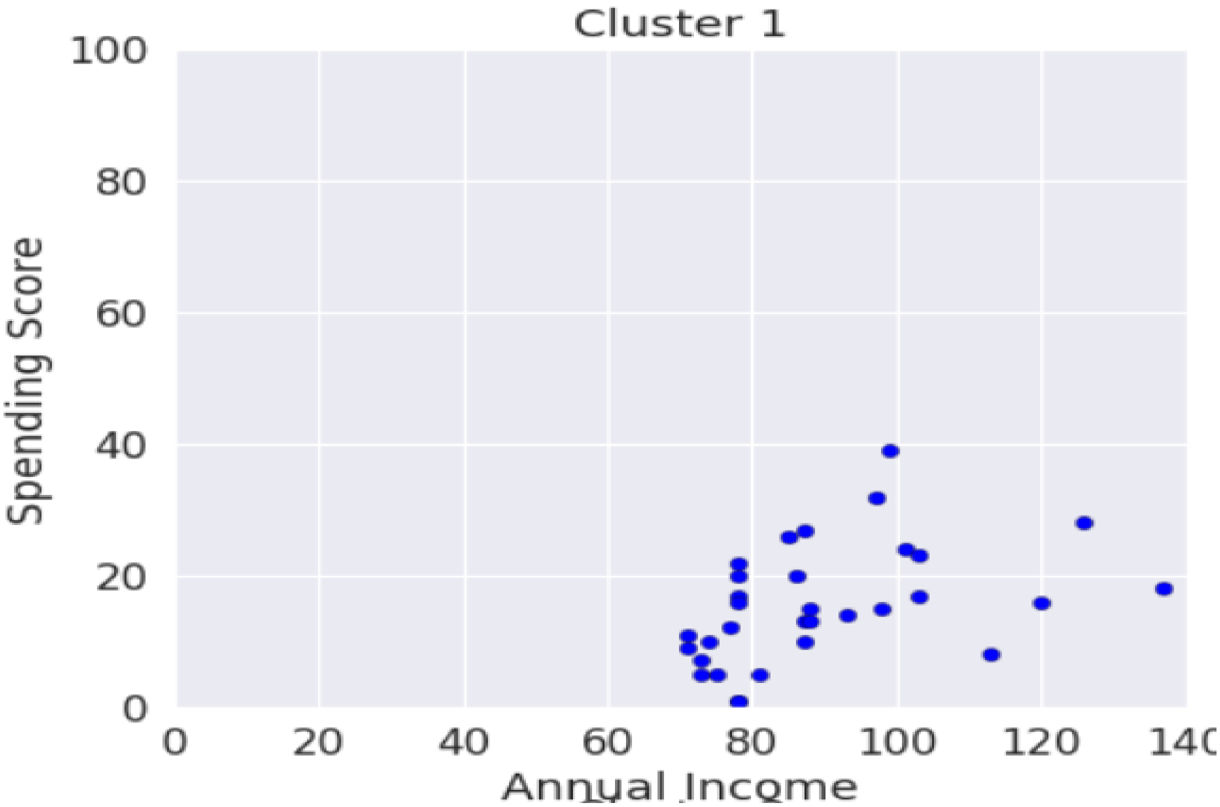
OPEN CODE IN COLAB

By the end of the colab notebook you will be:
Analyzing the target customers for maintaining and building business profitability using Customer Segmentation with Hierarchical Cluster Analysis in Python

OBSERVATION AND INFERENCES

Now, let’s visualize and observe individual clusters one by one

Cluster 1 - Navy Blue (Misers)

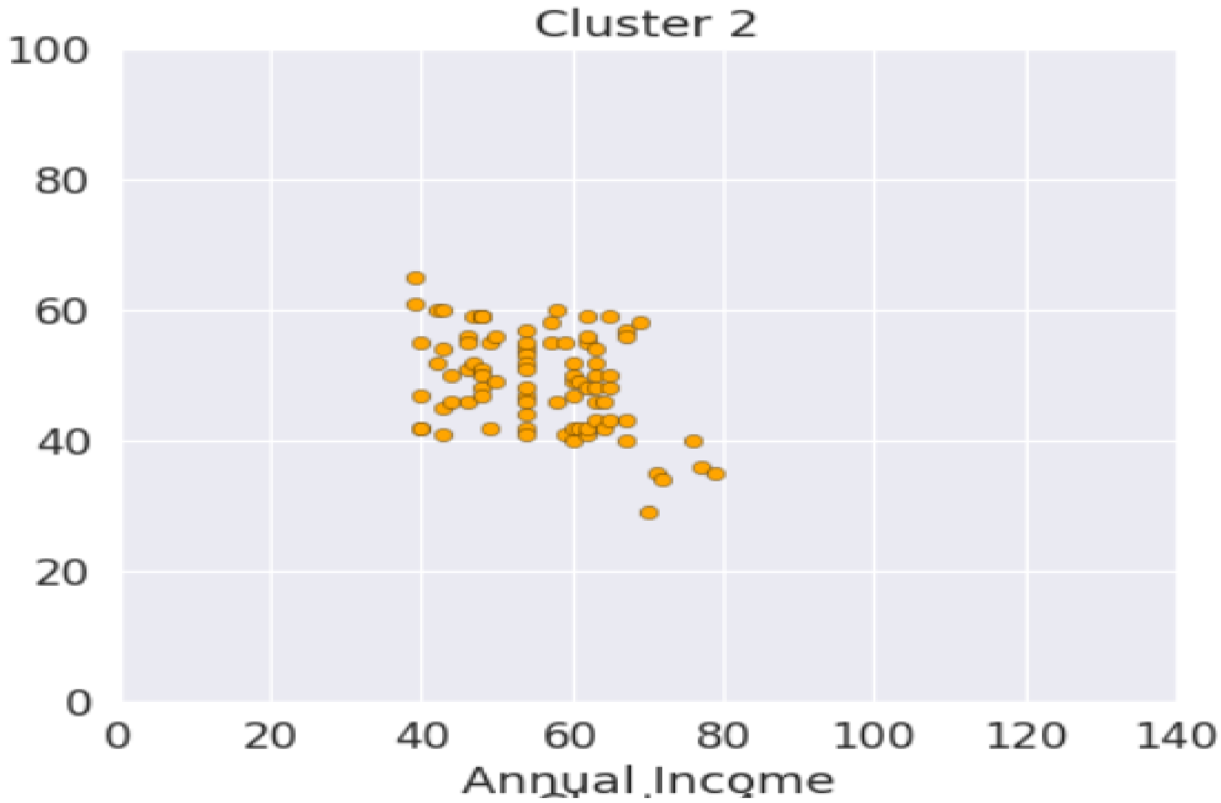


Misers meaning: People who hoard money and spends as little as possible

Inference: These are customers with high annual income but low spending scores. These could be the customers that are not much satisfied with the mall products or services.

Target potential: Quite High as these customers have potential to spend more money.

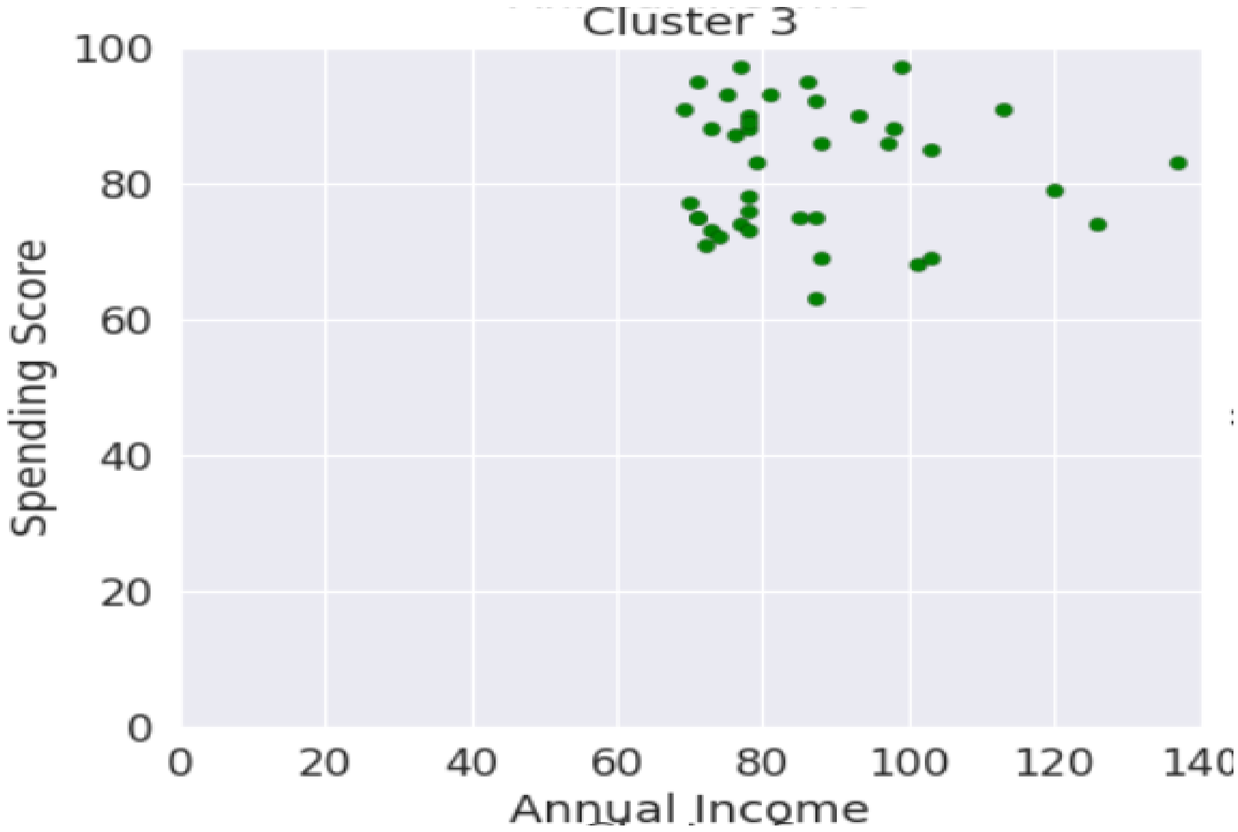
Cluster 2 - Orange (Normal Customers)



Inference: These are the customers with average Annual Income and Spending Score which seems most in frequency.

Target Potential: Not Much as there will always be a high number of average customers but different data analysis techniques can be used to increase the spending scores of average customers

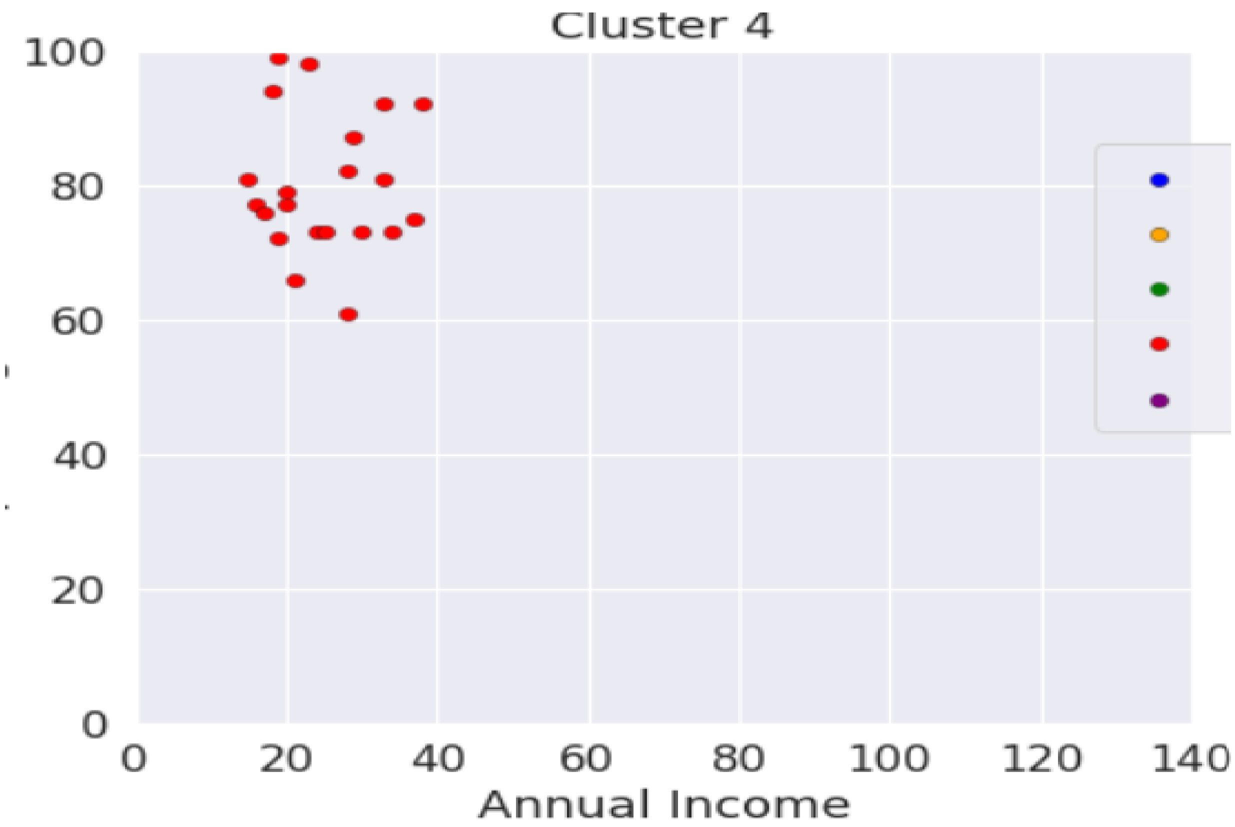
Cluster 3 - Green (Lavish)



Inference: These are the customers whose Annual Income is High so is their Spending Score. These people might be the regular customers of the mall and are convinced by the mall's facilities

Target Potential: Very High. People with high income and high spending scores create an ideal case for the mall or shops as these people are the prime sources of profit.

Cluster 4 - Red (Spendthrift)

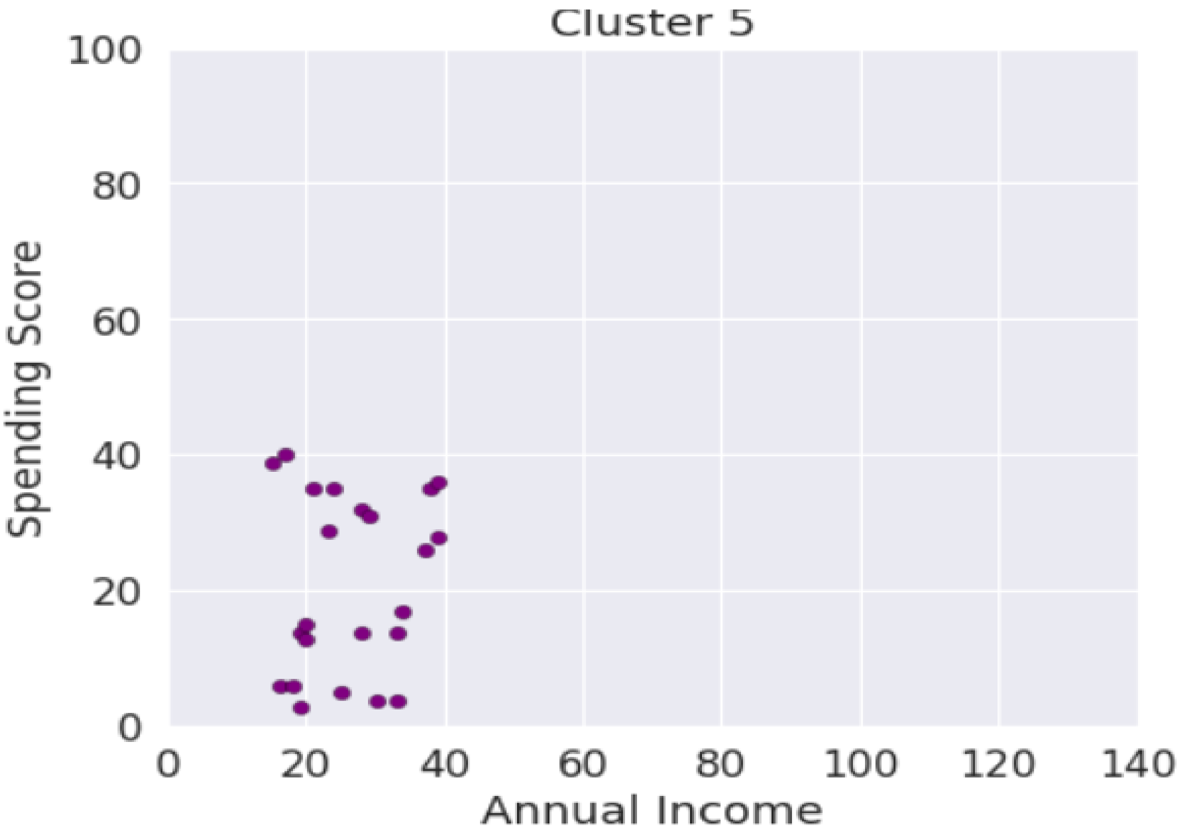


Excuse the five colorful dots on the right of this graph. It is a part of legend which shows what color denotes which cluster. Refer to the colab notebook and run this cell to view all clusters visualized individually.

Inference: These are the customers with relatively low annual income but have high spending scores. They probably are extremely satisfied with the mall facilities or just love to shop.

Target Potential: High. These customers can also be treated as potential targets but they can be unpredictive. So mall or shop owners might not proactively target these people but still will try not to lose them.

Cluster 5 - Purple (Balanced Customers)



Inference: These are the customers who have low Annual income and low spending which makes sense.

Target Potential: Very Low. Mall or shop owners will be less interested in this segment of customers.

CONCLUSION

After working with the Mall customers data containing information regarding their Customer_ID, Age, Gender, Annual Income, and their Spending Score. And analyzing the features Annual Income and Spending Score of customers using an unsupervised machine learning algorithm i.e., Hierarchical Clustering, It can be safely recommended that Mall owners should focus heavily on customers with high spending score. As they are the direct influencers of business profitability.

An indirect influence to profitability of business could be the segment of customers whose Spending Score is not high enough but they earn a lot annually. These could be potential targets as they are more probable to spend more if the right facilities and products are provided to them. Without analysis this specific customer segment could have been overlooked, causing potential profit opportunity loss.

Hence, business owners should concentrate their marketing heavily upon these targeted segments of the customers. Marketing could involve a wide range of options such as improving facilities that attract these customers, collecting more information regarding their product preferences and updating the inventory pertaining to the same, providing incentives to spend more to these customers, etc.

This is how powerful data analysis for businesses can be. And this is really just the tip of the iceberg. There is so much more that can be done with business or any data which can make us take better decisions for a more profitable future.

DISCUSSION AND SCOPE

How good is Hierarchical Clustering?

Although there are several advantages of using Hierarchical Clustering:
→ Representing the data in a more natural and informative way, showing not only the final clusters but also the intermediate levels of similarity between data points.
→ No requirement to specify the number of clusters in advance, which can be an advantage in situations where the number of clusters is unknown or difficult to determine.

However, Hierarchical Clustering can be more computationally expensive than flat clustering, particularly for large datasets. Here are some reasons why hierarchical clustering may not be the best choice for certain applications:

1. **Computational complexity**: Hierarchical clustering can be computationally expensive, particularly for large datasets. As the number of data points grows, the time and memory required to construct the dendrogram can become prohibitive. Which also leads to scalability issues.
2. **Lack of flexibility**: Hierarchical clustering produces a fixed hierarchy, which may not always be the best representation of the data. Moreover, it is difficult to update the hierarchy once it has been constructed, so hierarchical clustering may not be well-suited to streaming or online data analysis.
3. **Incomprehensive for large datasets**: It can also be difficult to interpret the results of hierarchical clustering, particularly when the dendrogram is complex and has many levels.
4. **Sensitivity to noise**: Hierarchical clustering is sensitive to noise and outliers, which can distort the dendrogram and produce misleading results.

Therefore, while hierarchical clustering can be a powerful tool for exploratory data analysis and data mining, it may not always be the best choice for every application, and alternative clustering algorithms may need to be considered.

What is the scope and possible applications of Hierarchical Cluster Analysis?

Data analysis: Hierarchical clustering is used to identify patterns and relationships in large datasets. It can be used to group similar observations or variables together based on their similarity.

Image processing: Hierarchical clustering is used in image processing to segment images into different regions based on the similarity of the pixels in the image.

Marketing: Hierarchical clustering is used in market research to segment customers into different groups based on their buying patterns or demographic characteristics.

Biology: Hierarchical clustering is used in bioinformatics to group genes or proteins together based on their similarity.

Social network analysis: Hierarchical clustering is used to identify communities or clusters of individuals in social networks based on their relationships.

Document clustering: Hierarchical clustering is used in natural language processing to group similar documents together based on their content.

Thank You For Reading

Important Links

Learn Hierarchical Clustering:

1. https://nlp.stanford.edu/IR-book/html/htmledition/hierarchical-clustering-1.html#ch:hierclust
2. https://stackabuse.com/hierarchical-clustering-with-python-and-scikit-learn/
3. https://towardsdatascience.com/understanding-the-concept-of-hierarchical-clustering-technique-c6e8243758ec

Mall Customer dataset: https://www.kaggle.com/datasets/vjchoudhary7/customer-segmentation-tutorial-in-python

ScikitLearn Documentation: https://scikit-learn.org/stable/modules/clustering.html#hierarchical-clustering

Colab Link: https://colab.research.google.com/drive/1Orruljxjkw2DEm_1UuU1v4K8NSDY06hd?usp=sharing

Complete Project Walkthrough on my website: https://sites.google.com/view/vaibhavsatish/home/heirarchical-cluster-analysis/

My Linkedin: https://www.linkedin.com/in/vaibhav-satish-35061b1ba/

* * * * *