

**VIETNAM NATIONAL UNIVERSITY  
HANOI INTERNATIONAL SCHOOL**

-----



**Subject: SEMINAR**

**Report Final Examination**

Class	:	INS3075
Lecture	:	Oanh Tran Thi
Topic	:	Amazon Customer Analysis
Group number	:	3
Members	:	Do Thi Minh Phuong – 21070074 Nguyen Phuong Thao – 20070981 Vu Thi Thao – 21070528

*Hanoi, 21th May 2024*

# TABLE OF CONTENT

<b>MEMBER'S CONTRIBUTION .....</b>	<b>3</b>
<b>I. INTRODUCTION .....</b>	<b>4</b>
<b>II. CUSTOMER CHURN ANALYSIS – AMAZON PERSONALIZED .....</b>	<b>4</b>
1. Definition .....	4
2. Amazon Prime's Churn rate.....	4
3. Consequence .....	7
4. Reason.....	9
5. Use models to predict customer churn.....	10
6. Conclusion .....	14
<b>III. CUSTOMER SEGMENTATION ANALYSIS .....</b>	<b>14</b>
1. Definition .....	14
2. About the dataset.....	16
3. About the research problem .....	17
4. Techniques/Methods/Models and Tools use to solve problem.....	18
5. Case study .....	26
6. Benefits and Recommendation .....	27
<b>IV. CUSTOMER RECOMMENDATION SYSTEM.....</b>	<b>28</b>
1. Definition .....	28
2. About the research problem .....	29
3. Types of recommendation.....	29
4. Implementation .....	33
5. Demo .....	40
6. Conclusion .....	41
<b>V. CONCLUSION .....</b>	<b>41</b>
<b>REFERENCES .....</b>	<b>42</b>

## LIST OF FIGURES

Figure 1. Amazon Prime retention rates in the United States _____	5
Figure 2. Online shopping cart abandonment rates worldwide _____	5
Figure 3. Share of retail eCommerce sales _____	8
Figure 4. Annual net sales revenue of Amazon _____	8
Figure 5. Age Distribution of Amazon Prime Users _____	19
Figure 6. Gender Distribution of Amazon Prime Users _____	19
Figure 7. Subscription Plan Distribution of Amazon Prime _____	20
Figure 8. Renewal Status Distribution of Amazon Prime _____	20
Figure 9. Distribution of Amazon Prime User Engagement Levels _____	21
Figure 10. Scatter plot of Feedback/Ratings vs Customer Support Interactions by Engagement Level _____	21
Figure 11. Favourite Content Genres Distribution Among Amazon Prime Users _____	22
Figure 12. Device Usage Patterns Distribution Among Amazon Prime Users _____	23
Figure 13. RFM Segmentation of Amazon Prime Users _____	24
Figure 14. Amazon Prime Revenue from 2014 to 2023 _____	26
Figure 15. Content-based systems _____	30
Figure 16. User-Based Collaborative Filtering _____	31
Figure 17. Item-Based Collaborative Filtering _____	32
Figure 18. Rating Distribution _____	34
Figure 19. Books ordered for support & Genres more than 3K entires _____	34
Figure 20. Number of books published per year _____	35
Figure 21. Top 10 most viewed books _____	36
Figure 22. Best suggestion for customer _____	37
Figure 23. Selected books by user _____	38
Figure 24. Correlations between users _____	38
Figure 25. User with highest correlation _____	38
Figure 26. Recommend books for user _____	39
Figure 27. Best suggestion for item _____	39
Figure 28. Top 5 recommended books for item _____	40
Figure 29. Demo Amazon Book Recommendation System _____	40
Figure 30. Demo Amazon Book Recommendation System _____	41

## MEMBER'S CONTRIBUTION

No.	Name	Contribution
1	Vu Thi Thao 21070528	Customer Churn Analysis – Amazon Personalized
2	Nguyen Phuong Thao 20070981	Customer Segmentation Analysis
3	Do Thi Minh Phuong 21070074	Customer Recommendation System

## **I. INTRODUCTION**

E-commerce is the largest growing thing on the internet using all the possible directions that are involved technically. Machine learning, business intelligence, and artificial intelligence-based solutions are a few of the best solutions developed to generate leads in e-commerce. The world's business hubs just turned after the COVID-19 breakthrough into online smart places. People overall recommend online marketplaces more over the regular shopping markets.

Developing such online marketplaces can build the economy, can overcome the current fear of the pandemic, and develop more reliable technical smart markets to generate business leads. Developing such solutions, which compete in the latest tech-based problems, especially highly involving artificial intelligence-based solutions. Recommending items based on the search history of the person browsing, is one example, telling the store owner what products are generating revenue and what products are still just filling up storage space. Adding more features like responding to the voice of the customer is a new artificial intelligence-based solution . Generating bills online and sending the items or products directly to the delivery address without the involvement of the human being is what Amazon is achieving.

## **II. CUSTOMER CHURN ANALYSIS – AMAZON PERSONALIZED**

### **1. Definition**

Customer churn is the percentage of customers who stop using a service or product over a given period of time. It is an important metric to measure customer satisfaction, loyalty, and retention. For E-commerce businesses like Amazon, customer churn can have a significant impact on revenue and growth.

### **2. Amazon Prime's Churn rate**

According to recent research from Daniela Coppola on Statista Web, Amazon Prime's Churn rate in the United States allways under 5% and average of the world is about 7%.

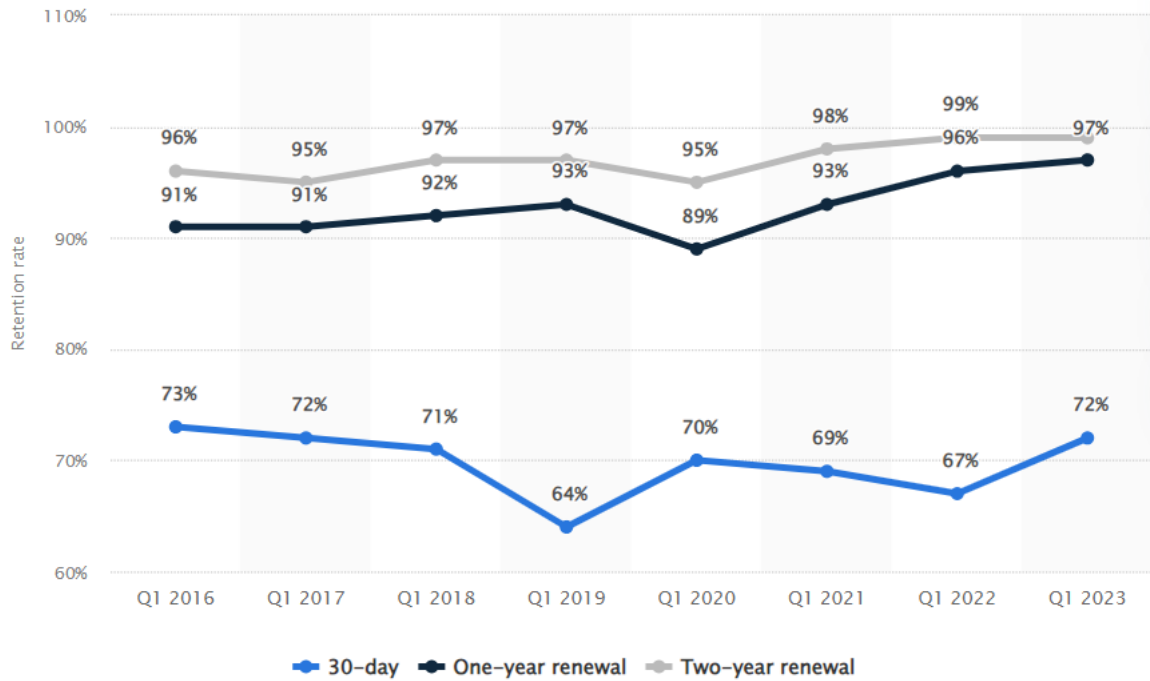


Figure 1. Amazon Prime retention rates in the United States

This means that customers using Amazon's services are less likely to stop using them, which is significant feat. Comparing with average of E-commerce of the World's rate is lower about 15%.

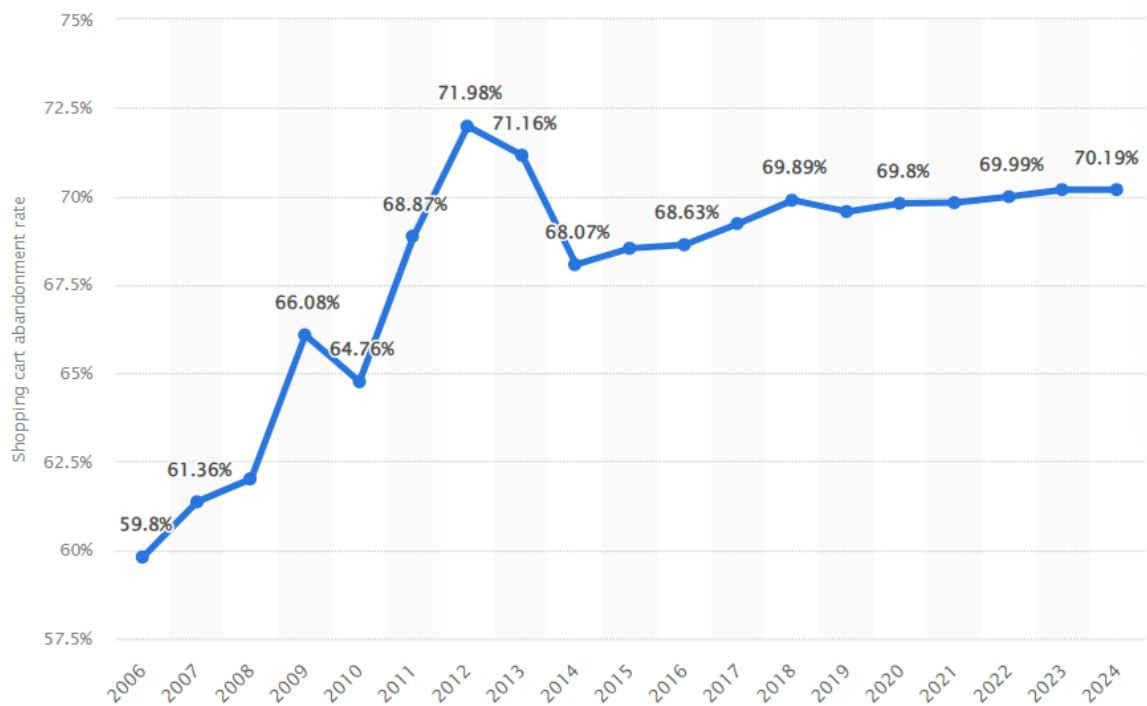


Figure 2. Online shopping cart abandonment rates worldwide

## Amazon Maintains Low Churn & High Customer Retention:

Customer-centricity as a mission. Amazon's mission is seen through its innovative use of technology to create seamless, intuitive experiences for its users. From personalized recommendations to Alexa's voice-activated shopping, Amazon uses data analytics and machine learning to anticipate customer needs and preferences. Such strategies showcase Amazon's desire to place the customer right at the center of its business model.

### a. Algorithm recommendations:

- Amazon uses data from its member, taggers, and machine learning algorithms to create personalized suggestions for what to then decide 3 types:
  - **User Activity (Events):** Includes interactions such as clicks, purchases, or views. This data type provides the strongest signal for personalization because it directly reflects user behavior and preferences. It is essential for creating accurate and relevant recommendations tailored to individual users.
  - **Item Details:** Includes attributes such as price, category, style, and genre. While optional, item details are very useful for enhancing recommendation accuracy, especially for new items that lack historical interaction data. This data helps in understanding the context and characteristics of items, allowing for better matching with user preferences.
  - **User Details:** Includes demographic information such as location, age, gender, and subscription tier. This data helps to refine and personalize recommendations further based on user demographics. It is particularly useful for segmenting users and tailoring recommendations to different user groups, enhancing the overall user experience.
- Data Ingestion Methods:

- **Bulk Import:** Use dataset import jobs to upload historical event data and comprehensive details about items and users. This method is ideal for initializing models with a rich dataset.
- **Real-Time Streaming:** Use the PutEvents, PutItems, and PutUsers APIs to stream the latest user interactions and metadata into the service. This keeps your data sets current and ensures that recommendations are based on the most recent user activities and item details.
- Once your data is integrated into Amazon Personalize, you can easily create a custom private personalization model, trained and hosted specifically for your application, enhancing the accuracy and relevance of recommendations with minimal effort.

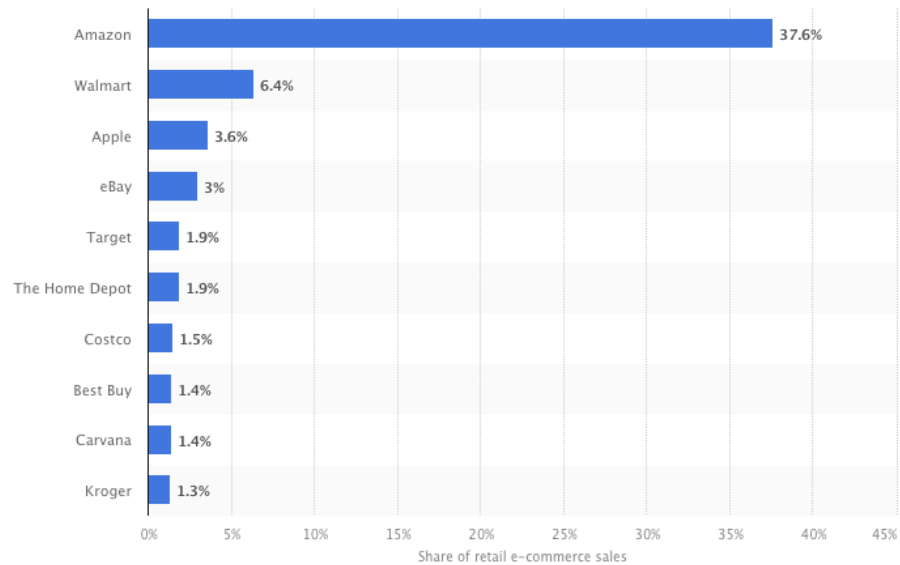
b. Email recommendations and reminders:

- Amazon sends emails to notify membership of new releases and recommendations based on their viewing history.

### 3. Consequence

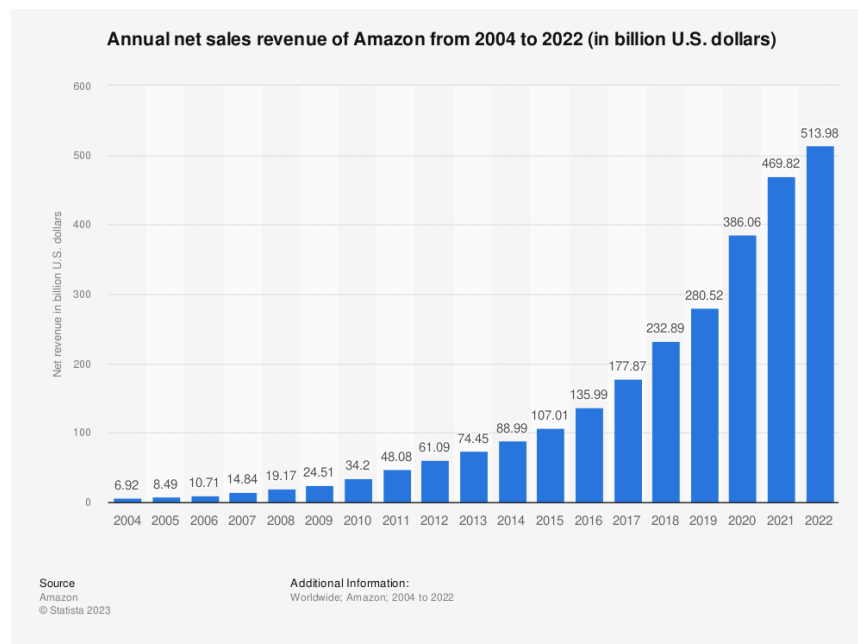
- Amazon's Share of the US Ecommerce Market is 37.8%: Amazon itself has an inventory of about 12 million items across all its categories and services. But if you go broader and look at all the items that Marketplace sellers list, that number expands to about 350 million. It's a lot of competition, yes, but it's also a lot of visibility and sales. According to Statista, Amazon was responsible for 37.6% of US eCommerce spending in 2023 — a figure which is expected to rise by another 11.7% in 2024.





*Figure 3. Share of retail eCommerce sales*

- Amazon's Net Revenue Continues to Grow: Amazon's net revenue increased by over \$40 billion year-on-year, from \$513.98 billion in 2022 to \$554.02 billion in 2023.



*Figure 4. Annual net sales revenue of Amazon*

- The US marketplace is still the highest-performing: It's not surprising that Amazon started its journey in the United States, as this is where Jeff Bezos originally founded the company. Since then, Amazon has expanded globally, encompassing 21 marketplaces. Following the U.S., the highest-

earning Amazon marketplaces are in Canada, the UK, Mexico, and Germany, with France also being a close contender.

- **Prime Members Spend a Lot:** There are over 200 million Amazon Prime members around the world (more Prime members than non), and they typically spend over \$1,000 a year. This might be an extra incentive to fulfil with Amazon, or at least make your items Prime-eligible. For those without a Prime membership, they tend to spend a little less freely than their Prime counterparts. About three-quarters of non-Prime shoppers spend between \$100 to \$500 a year on Amazon.
- **Third-party sellers make up the majority of total Amazon sellers:** Around the world, close to 2 million small and medium-sized third-party businesses engage in selling on Amazon. As of 2023, approximately 70% of these sellers operate as independent third-party (3P) sellers, utilizing Amazon's Seller Central platform.
- **Amazon Sales Per Second, Minute and Hour:** Each second, Amazon records \$4,722. Each minute, those sales amount to \$283,000. And in an hour, that averages more than \$17 million.

#### **4. Reason**

Building a churn rate prediction model has a direct impact on personalized services:

- **Targeted Interventions:**  
*Personalized Offers:* By identifying customers at risk of churning, Amazon can deliver personalized offers, discounts, or recommendations to these individuals. This targeted approach makes the intervention more effective because it aligns with the specific preferences and behaviors of each customer.  
*Customized Communication:* Amazon can tailor communications to at-risk customers based on their past interactions and preferences, making the outreach more relevant and engaging.
- **Enhanced Customer Experience:**  
*Proactive Service Improvements:* By understanding the reasons behind potential churn, Amazon can personalize the user experience to address specific issues, thereby improving customer satisfaction and loyalty.

*Personalized Support:* Amazon can offer tailored support and solutions to at-risk customers, ensuring they receive the help they need to stay engaged with the service.

- Predictive Personalization:

*Anticipating Needs:* A churn prediction model can help Amazon anticipate the needs and preferences of at-risk customers, allowing for more proactive and personalized product recommendations.

*Behavioral Insights:* The insights gained from churn prediction can feed back into personalization algorithms, enhancing their accuracy and relevance by understanding what drives customer retention and satisfaction.

- Resource Optimization:

*Efficient Allocation:* By predicting churn, Amazon can allocate resources more efficiently, focusing on personalizing experiences for those most likely to leave. This ensures that efforts and investments in personalization have the highest possible impact.

*Strategic Focus:* Personalized retention strategies can be developed for different segments of at-risk customers, ensuring that interventions are both effective and cost-efficient.

- Customer Lifetime Value (CLV):

*Maximizing CLV:* Personalized strategies based on churn predictions help maximize customer lifetime value by reducing churn and encouraging continued engagement with Amazon services.

*Long-term Relationships:* Building long-term relationships through personalized interactions increases the overall value of each customer, as they are more likely to remain loyal and continue using Amazon's services.

- Data-Driven Insights:

*Feedback Loop:* The data from churn prediction models provides valuable insights that can enhance personalization algorithms. For example, if certain products or services are linked to higher retention rates, these can be emphasized in personalized recommendations.

*Continuous Improvement:* Insights from churn predictions help Amazon continuously refine its personalization strategies, ensuring they remain effective as customer preferences and behaviors evolve.

## **5. Use models to predict customer churn**

Models use machine learning algorithms to analyze historical data and identify patterns that can help predict future behavior. Some common techniques used to predict customer churn:

#### **A. Random forest:**

Random forest works by constructing multiple decision trees and combining their predictions to make a final prediction.

In the context of predicting customer churn for Amazon, a random forest model would take into account various input variables that may be related to the likelihood of churn. These could include factors such as the customer's researching history, purchase and interactions with the service, as well as demographic information and other relevant data.

The random forest algorithm works by constructing multiple decision trees using a modified version of the standard decision tree algorithm. Each tree is constructed using a bootstrap sample of the training data, meaning that some samples may be used multiple times while others may not be used at all. Additionally, at each split in the tree, only a random subset of the input variables is considered when choosing the best split.

Once all trees have been constructed, they can be used to make predictions for new customers by averaging the predictions of all trees in the forest. The final prediction is then determined by taking a majority vote among the predictions of all trees.

Random forests have several advantages for predicting customer churn. They are relatively fast to train and can handle large datasets. They are also less prone to overfitting than individual decision trees, due to the use of bootstrap sampling and random variable selection. However, random forests can be more difficult to interpret than individual decision trees, as the final prediction is determined by a combination of multiple trees.

#### **B. Decision tree:**

This is a type of machine learning algorithm that can be used for classification and regression problems, including predicting customer churn for Amazon. Decision trees work by recursively splitting the data into subsets based on the values of the input variables, in order to create a tree-like structure where each leaf node represents a prediction.

In the context of predicting customer churn for Amazon, a decision tree model would take into account various input variables that may be related to the likelihood of churn. These could include factors such as the customer's researching history, purchase and interactions with the service, as well as demographic information and other relevant data.

The decision tree algorithm works by selecting the best input variable to split the data on at each step, based on a measure of impurity such as Gini impurity or information gain.

Gini impurity is a measure of the impurity of a set of samples. It is calculated as the probability that two randomly chosen samples from the set will have different class labels. A Gini impurity of 0 indicates that all samples in the set have the same class label, while a Gini impurity of 1 indicates that the samples are evenly distributed among all class labels. When building a decision tree, the algorithm will try to choose splits that minimize the Gini impurity of the resulting subsets.

Information gain is another measure that can be used to evaluate splits in decision tree algorithms. It is based on the concept of entropy from information theory.

In the context of decision tree algorithms, entropy is used to evaluate the quality of a split by measuring the reduction in uncertainty that results from splitting the data on a particular variable.

Entropy is calculated as the negative sum of the probabilities of each class label, multiplied by the logarithm of the probabilities. For a binary classification problem with two class labels, the entropy can be calculated as follows:

$$\text{Entropy} = -p_1 * \log_2(p_1) - p_2 * \log_2(p_2)$$

Where  $p_1$  and  $p_2$  are the probabilities of the two class labels.

The entropy is 0 when all samples in the set have the same class label, indicating that there is no uncertainty. The entropy is maximized when the samples are evenly distributed among all class labels, indicating maximum uncertainty.

When building a decision tree, the algorithm will try to choose splits that maximize the information gain, meaning that they result in the greatest reduction in uncertainty.

The data is then split into two subsets based on the chosen variable, and the process is repeated recursively on each subset until a stopping criterion is met, such as reaching a maximum tree depth or a minimum number of samples per leaf.

Once the tree is constructed, it can be used to make predictions for new customers by traversing the tree from the root to a leaf node based on the values of the input variables. The prediction at the leaf node is then returned as the final prediction.

Decision trees have several advantages for predicting customer churn. They are easy to interpret and can handle both numerical and categorical data. They are also relatively fast to train and can handle large datasets. However, decision trees can be prone to overfitting, especially when the tree is very deep. To overcome this issue, techniques such as pruning or ensemble methods like random forests can be used.

### **C. Logistic regression:**

Logistic regression is a type of generalized linear model that is commonly used for binary classification problems, such as predicting whether a customer will churn or not. The model estimates the probability of the binary outcome as a function of the input variables, which can include both numerical and categorical data.

In the context of predicting customer churn for Amazon, a logistic regression model would take into account various input variables that may be related to the likelihood of churn. These could include factors such as the customer's researching history, purchase and interactions with the service, as well as demographic information and other relevant data.

The logistic regression model works by estimating the probability of churn using a mathematical function called the logistic function. This function takes in a linear combination of the input variables, weighted by a set of coefficients, and outputs a value between 0 and 1 representing the estimated probability of churn.

The logistic function has an S-shaped curve, with the output approaching 0 as the input becomes very negative, and the output approaching 1 as the input becomes very positive. The function is defined as follows:

$$f(x) = 1 / (1 + \exp(-x))$$

where  $x$  is the input value and  $\exp$  is the exponential function.

In the context of logistic regression, the input to the logistic function is a linear combination of the input variables, weighted by a set of coefficients. The output of the function represents the estimated probability of the binary outcome (e.g., churn or not churn).

The coefficients in the model are determined during the training process, where the model is fit to historical data using a technique called maximum likelihood estimation (MLE).

The basic idea behind MLE is to find the set of parameters that makes the observed data most likely under the assumed model. In other words, MLE tries to find the parameter values that maximize the likelihood function, which measures how likely the observed data is given the model and the parameter values.

Once trained, the logistic regression model can be used to predict the likelihood of churn for new customers based on their input variables. The predicted probabilities can then be used to identify customers who are at high risk of churning and take appropriate actions to retain them.

It is important to note that while logistic regression is a powerful technique for predicting binary outcomes, it is not always the best choice for every problem. Other techniques, such as decision trees and random forests, may also be effective for predicting customer churn and should be considered when building a predictive model

## **6. Conclusion**

By integrating churn prediction with personalized services, Amazon can create a more engaging and satisfying customer experience, ultimately reducing churn and enhancing loyalty. This integration helps ensure that the efforts to retain customers are not just broad and generic, but finely tuned to the needs and preferences of each individual.

## **III. CUSTOMER SEGMENTATION ANALYSIS**

### **1. Definition**

- **What is a customer segment?**

A customer segment is a group of consumers who share similar characteristics and needs. By identifying and understanding your different customer segments, businesses can tailor their products, services, and marketing efforts to better meet the specific needs of each segment. This can lead to more effective marketing, increased customer loyalty, and better overall profitability.

- **What is customer segmentation?**

Customer segmentation is the practice of dividing a customer base into groups of individuals that are similar in specific ways relevant to marketing, such as age, gender, interests, spending habits, and so forth. This enables companies to tailor their marketing efforts to specific groups, improving the effectiveness of their marketing strategies and enhancing customer engagement.

In business-to-business (B2B) contexts, customer segmentation often involves criteria such as:

- Industry: Different industries have different needs and challenges, affecting their buying decisions.
- Number of employees: The size of a company can influence its purchasing power and the type of products it requires.
- Products previously purchased: Understanding what a business has bought in the past can help predict future needs and cross-selling opportunities.
- Location: Geographic factors can influence product requirements, logistics, and marketing strategies.

In business-to-consumer (B2C) marketing, segmentation is frequently based on:

- Demographic Segmentation: Based on demographic information such as age, gender, income, education, and family size.
- Geographic Segmentation: Divides the market based on location like country, region, or city. This can be crucial for businesses that need to adjust their products or marketing to fit local cultures or needs.
- Psychographic Segmentation: Involves segmenting customers based on their lifestyles, interests, attitudes, values, and personalities.
- Behavioural Segmentation: Focuses on dividing customers based on their behaviours and patterns such as purchase history, brand loyalty, user status, and spending habits.



Customer segmentation helps businesses to more effectively target their products and services, craft more personalised communications, and ultimately increase efficiency in marketing efforts.

## **2. About the dataset**

<https://www.kaggle.com/datasets/arnavsmayan/amazon-prime-userbase-dataset>

This dataset is designed to provide a holistic view of our customer base, enabling us to effectively categorize customers into distinct segments based on their behaviors, preferences, and interactions with our services.

The dataset includes key customer attributes such as:

- User ID: Unique identifier for each customer.
- Name: Full name for personalization purposes.
- Email Address: Used for communication and as a unique contact identifier.
- Username: Identifier on our platform.
- Date of Birth: Helps in age-based segmentation.
- Gender: Used for demographic segmentation.
- Location: Geographic data for regional targeting.
- Membership Start Date and Membership End Date: Tracks the duration of engagement and customer lifecycle.
- Subscription Plan: Details of the service plan to understand preferences.
- Payment Information: Payment methods used, important for financial analysis.
- Renewal Status: Indicates subscription renewals or cancellations, related to customer retention.
- Usage Frequency: Frequency of service usage, indicates engagement levels.
- Purchase History: Record of transactions for analyzing purchasing behaviors.
- Favorite Genres: Preferences in content or products, used for personalized recommendations.
- Devices Used: Types of devices used to access services, informs tech support and development.
- Engagement Metrics: Includes activity levels, time spent, and interaction data.
- Feedback/Ratings: Direct customer feedback on services and products.
- Customer Support Interactions: Engagement records with customer support, used for service quality assessment.

### **3. About the research problem**

- **The research problem?**

Customer segmentation analysis for Amazon revolves around effectively categorizing Amazon's diverse and vast customer base into distinct segments to optimize marketing strategies, improve customer service, and enhance overall business performance. This involves addressing the challenges of analyzing large volumes of complex customer data to uncover meaningful patterns and characteristics. The core question is how to accurately identify and characterize these segments in a way that reflects differences in customer behaviors, preferences, and needs. Furthermore, the analysis must consider the dynamic nature of customer data, evolving market trends, and the diverse range of products and services offered by Amazon. The ultimate goal is to leverage this segmentation to provide more personalized customer experiences, improve product recommendations, tailor marketing communications, and drive strategic business decisions, all while maintaining data privacy and ethical standards.

- **How to solve the problem?**

**Step 1: Data Preprocessing**

Clean data to address missing values, remove duplicates, and correct errors. Normalize or standardize the data if necessary. Finally, feature engineering to create new metrics or combine existing variables to better capture customer behaviors and characteristics.

**Step 2: Exploratory Data Analysis (EDA)**

Conduct an exploratory analysis to understand the characteristics and patterns in the data. This may include analyzing purchase frequency, average spending, product preferences, and customer demographics.

**Step 3: RFM Segmentation**

Calculate RFM metrics for each customer.

Segment Customers by using RFM Scores: Customers can be divided into segments such as "High", "Medium", and "Low" based on RFM scores.

**Step 4: Applying K-Means Clustering**

Use the RFM metrics as input features for K-Means. Use methods like the elbow method, silhouette analysis, or other heuristic techniques to decide the optimal

number of clusters. Apply the K-Means algorithm to the RFM data. This will group customers into clusters based on their RFM characteristics.

#### **Step 5: Build a Predictive Model**

Predictive models can be built to predict user behavior. For example, a predictive model can estimate the probability that a user will like a particular genre movie based on their past behavior and the behavior of similar users.

#### **Step 6: Analyze and Interpret Clusters**

Analyze the characteristics of each cluster. Identify which clusters are most valuable or at risk.

#### **Step 7: Strategy Development and Implementation**

Based on the clustering insights, develop strategies tailored to each segment. Deploy these strategies in marketing plans and customer interaction initiatives.

#### **Step 8: Monitor and Refine**

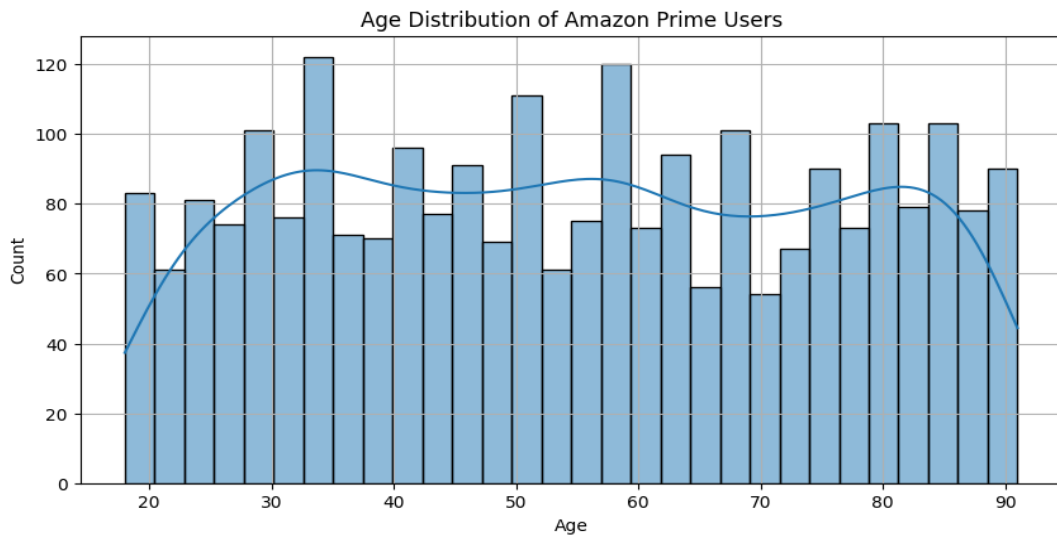
Monitor the performance of strategies implemented for different segments. Regularly revisit and refine the segmentation as new data becomes available or as business contexts change.

### **4. Techniques/Methods/Models and Tools use to solve problem**

To facilitate the analysis and derive insights on user trends, targeted marketing, and improvements in user experience, we analyse and insights into user behaviour, preferences, and interactions with the Amazon Prime platform.

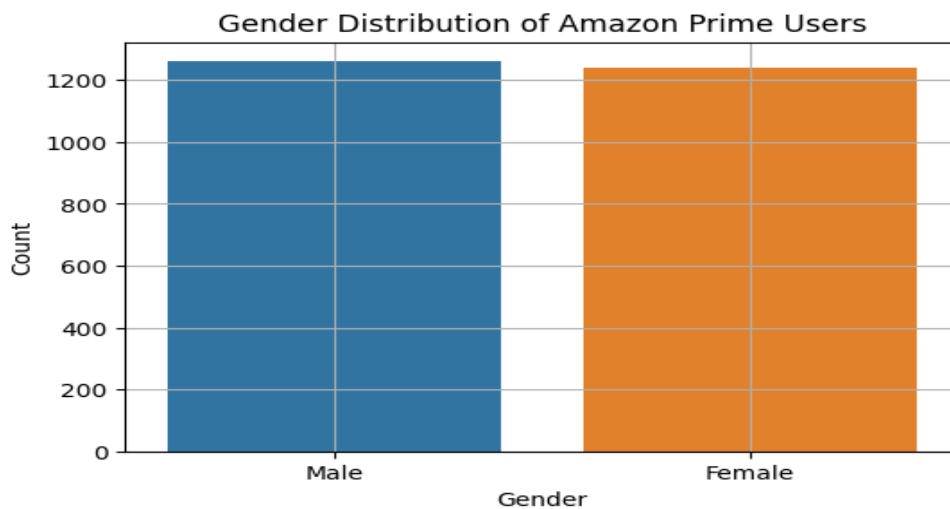
#### **4.1 Methods use to solve problem**

##### **a. Demographics Segmentation: Age and gender distribution.**



*Figure 5. Age Distribution of Amazon Prime Users*

The age distribution of Amazon Prime users shows a fairly wide spread, suggesting that the platform appeals to a diverse age range. We can see peaks in specific age groups, which could be further investigated to understand which age groups are the most active or valuable for targeted marketing.



*Figure 6. Gender Distribution of Amazon Prime Users*

The gender representation among Amazon Prime users appears relatively balanced. This suggests that the content and marketing can continue to be designed to appeal to a gender-diverse audience, or potentially, campaigns could be customized to target underrepresented genders more effectively if any business goals align with such initiatives.

## b. Behavioural Segmentation:

- Subscription and Renewal Analysis: Patterns in subscription types and renewal status.

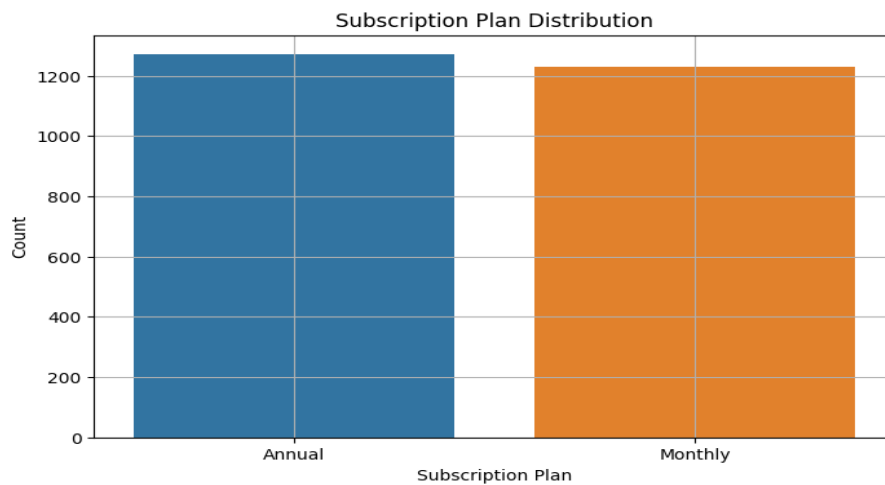


Figure 7. Subscription Plan Distribution of Amazon Prime

The data reveals a variety of subscription plans, with 'Monthly' plans being the most common among users. This preference for monthly subscriptions might indicate a need for flexibility among users or a lower commitment barrier, which can be leveraged in marketing strategies to promote these plans further or to introduce trial periods to encourage transitions to annual plans.

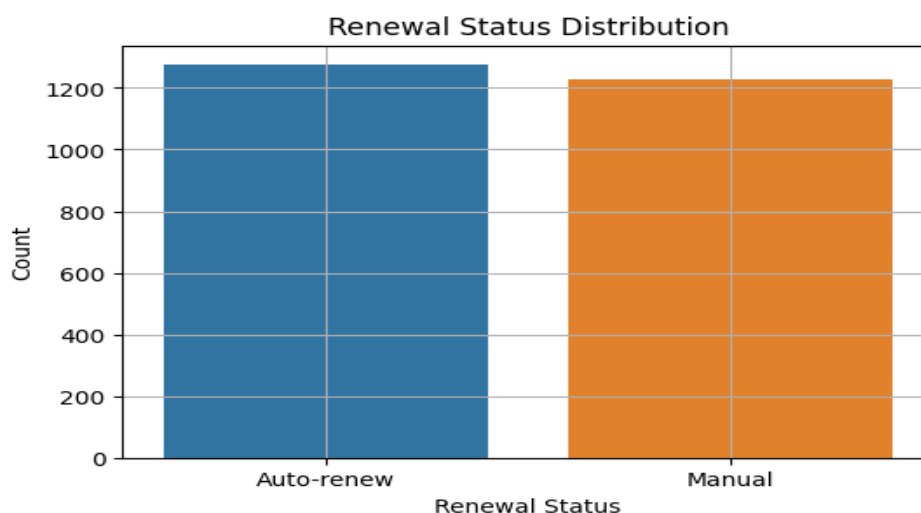
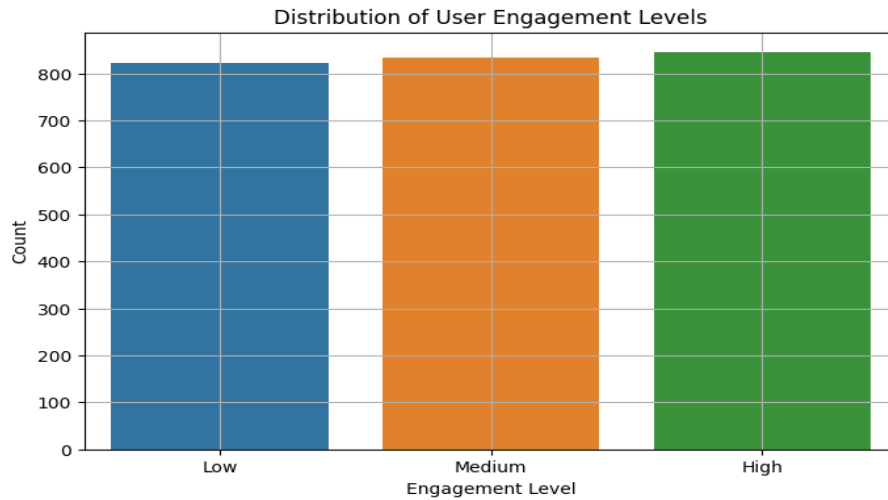


Figure 8. Renewal Status Distribution of Amazon Prime

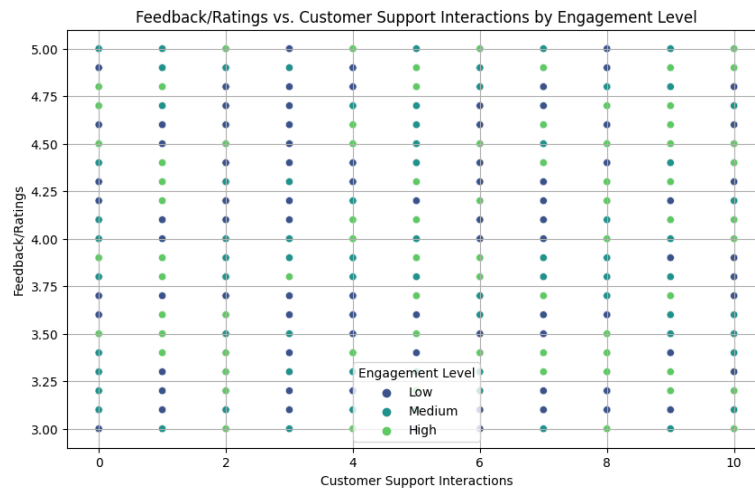
Most users are on an 'Auto-renew' plan, which indicates a high level of user retention and less friction at the renewal stage. This is a positive sign for stable recurring revenue. Marketing efforts might focus on converting users from 'Manual' to 'Auto-renew' by promoting the convenience and possibly incentives for switching to auto-renew.

- **Engagement Analysis:** Analyse user engagement and its correlation with feedback ratings and support interactions.



*Figure 9. Distribution of Amazon Prime User Engagement Levels*

The engagement levels among Amazon Prime users are fairly distributed, with 'High' engagement being the most common. We could be directed towards increasing engagement by enhancing content recommendations or introducing new features to engage users more deeply.



*Figure 10. Scatter plot of Feedback/Ratings vs Customer Support Interactions by Engagement Level*

The scatter plot shows a spread across different levels of customer support interactions and their corresponding feedback ratings. Interestingly, users with fewer support interactions tend to have higher ratings, suggesting that a smoother user experience correlates with better satisfaction. Users with higher engagement levels generally provide a range of feedback, possibly indicating more critical engagement with the platform's offerings. This visualization indicates potential

areas to improve the customer support process or to proactively address issues before they require support intervention, potentially enhancing user ratings and satisfaction.

c. Psychographic Segmentation: Explore favourite genres and purchasing behaviour.

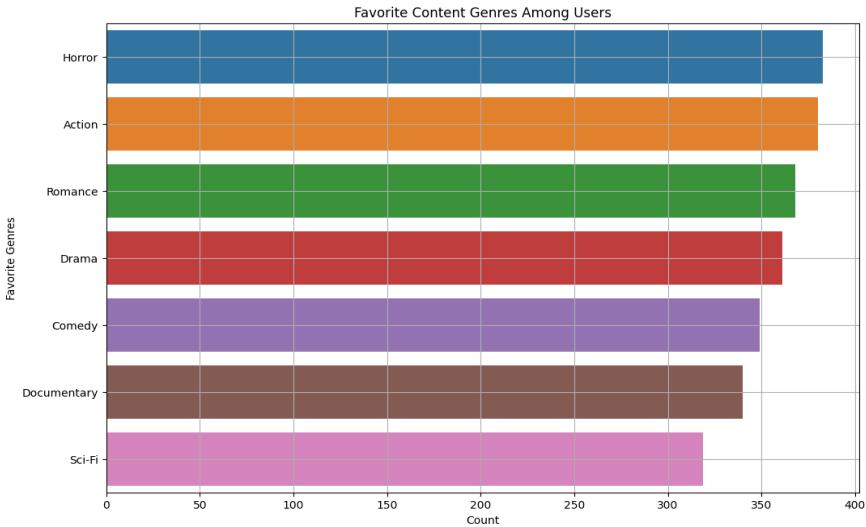
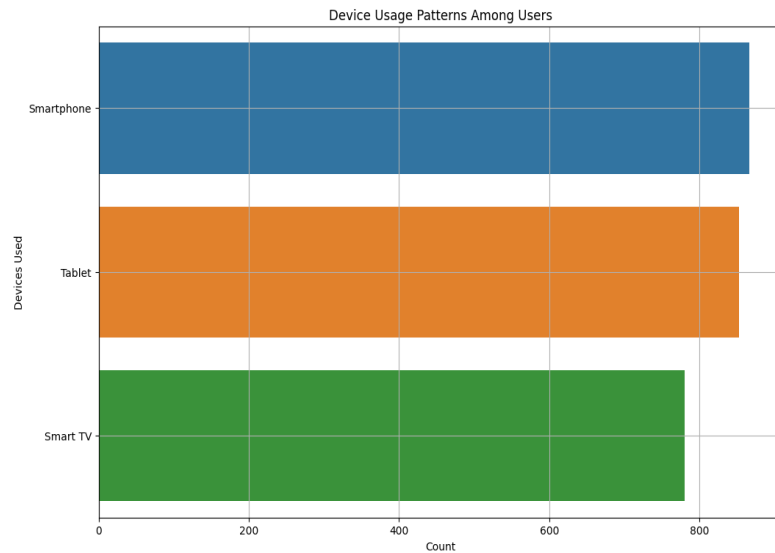


Figure 11. Favourite Content Genres Distribution Among Amazon Prime Users

The most popular genres among Amazon Prime users include Horror, Action, Romance, and Drama. This highlights diverse interests, indicating the platform's broad appeal. Marketing strategies could focus on promoting popular genres more aggressively and possibly curating specialized content or events around these genres to drive deeper engagement.



*Figure 12. Device Usage Patterns Distribution Among Amazon Prime Users*

Smartphones dominate as the most preferred device for accessing Amazon Prime, followed by Tablet. This suggests that users value the convenience of mobile access than a larger screen experience for consuming content. Enhancements and optimizations for these devices in the user interface and streaming quality can be prioritized to improve user experience.

#### **4.2 Technique use to solve problem (RFM Segmentation)**

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 behaviour because frequency and monetary value affects a customer's lifetime value, and recency effects retention, a measure of engagement.

- Recency (R): The number of days since the last purchase.
- Frequency (F): The total number of purchases.
- Monetary (M): The total money spent.



(	RFM Segment	Number of Users	
0	High	1262	
1	Medium	1236	
2	Low	0,	
	User ID	RFM Score	RFM Segment
0	1	6	High
1	2	5	Medium
2	3	5	Medium
3	4	5	Medium
4	5	7	High)

*Figure 13. RFM Segmentation of Amazon Prime Users*

- **High RFM Segment:** This segment consists of 1271 users who have the highest scores in recency, frequency, and monetary value, indicating they are the most valuable and engaged customers. Marketing Strategy: Concentrate on loyalty programs and exclusive offers to retain these high-value members. Introduce referral incentives to harness their potential as brand ambassadors.
- **Medium RFM Segment:** Comprising 1220 users, this group shows moderate engagement and spending on the platform. Marketing Strategy: Implement campaigns to boost engagement, such as personalized content recommendations and membership upgrades, to enhance their frequency and monetary scores.
- **Low RFM Segment:** Currently, there are no users in this dataset that fall into the low RFM category based on the current scoring system. This may suggest that the scoring thresholds or the RFM model itself require adjustments, informed by deeper business insights or a re-evaluation of the RFM factors.

### 4.3 Models use to solve problem

K-means clustering is a popular unsupervised machine learning algorithm used to group data into a predefined number of clusters, where each data point belongs to the cluster with the nearest mean. The algorithm aims to partition the data into `K` distinct, non-overlapping clusters based on distance to the centroid of the clusters. Using this technique, we can categorise customers into k-numbers of groups based on their recent activity, frequency, and monetary value.

- **Formulate and Calculations:**
  - Euclidean Distance (commonly used to measure distance between data point and centroid):

$$D(x, y) = \sqrt{\sum_{i=1}^n (x_i - y_i)^2}$$

Here,  $D(x, y)$  is the distance between two points  $x$  and  $y$  in  $n$ -dimensional space,  $x_i$  and  $y_i$  are the  $i$ -th coordinates of  $x$  and  $y$ , respectively.

- Centroid Calculation:

The centroid  $C_j$  of a cluster  $j$  is calculated as:

$$C_j = \frac{1}{s_j} \sum_{x_i \in S_j} x_i$$

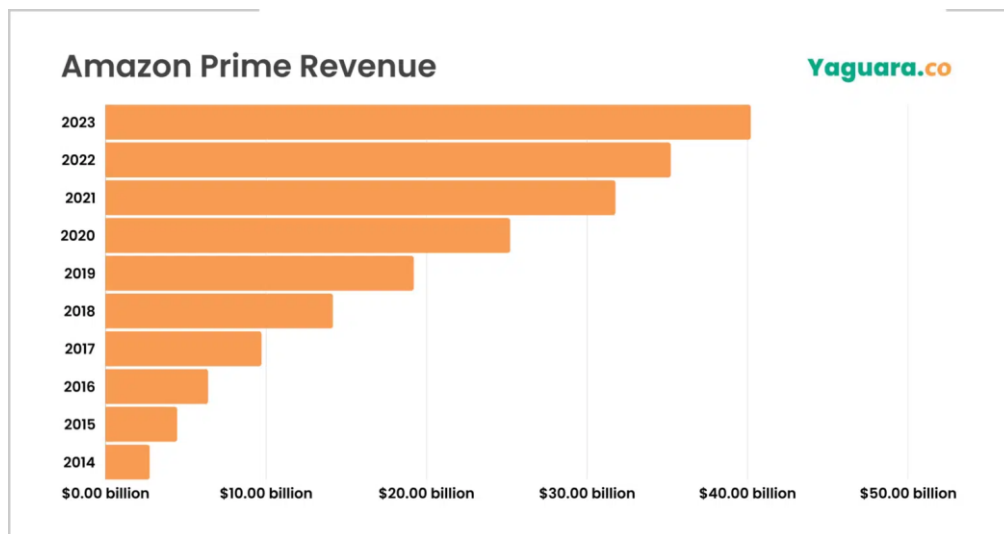
- **How K-means Clustering Works?**

- Initialization: Start by selecting  $K$  initial centroids (the number of clusters you want). The selection can be random or based on a specific criterion.
- Assignment Step: Assign each data point to the nearest cluster by calculating its distance to each centroid. Common distance metrics used are Euclidean distance, Manhattan distance, etc., with Euclidean being the most used.
- Update Step: After all points have been assigned to clusters, recalculate the centroids of these clusters. The new centroid of each cluster is typically the mean of all points assigned to that cluster.
- Iteration: Repeat the assignment and update steps until the centroids no longer move (the clusters are stable) or until a specified number of iterations is reached. This means the algorithm has converged on a solution.

#### 4.4 Tools use to solve problem

- Data Manipulation and Analysis:
  - Pandas: Essential for data manipulation in Python.
  - NumPy: Supports large, multi-dimensional arrays and matrices in Python.
- Machine Learning Libraries: Scikit-learn: Offers a variety of machine learning algorithms for classification, regression, clustering, and dimensionality reduction.
- Visualization Tools: Matplotlib and Seaborn: Python libraries for creating static, animated, and interactive visualizations.
- Development Environments and Notebooks: Jupyter Notebook: Allows integration of live code, visualizations, and text.

## 5. Case study



*Figure 14. Amazon Prime Revenue from 2014 to 2023*

Amazon, founded in 1994 by Jeff Bezos, started as a small online bookstore and has grown into the world's largest online retailer, with annual revenue surpassing \$554.02 billion by 2023. . A key factor in Amazon's impressive growth has been its innovative use of customer segmentation to deliver personalized recommendations. By effectively leveraging data such as purchase history, browsing behavior, and demographic information, Amazon has set a new standard in personalized marketing. A key part of Amazon's strategy is sophisticated machine learning algorithms that analyze customer data to make appropriate product recommendations. This approach started with books and quickly expanded to a range of different product categories. This personalization not only enhances the customer's shopping experience but also significantly increases customer loyalty and repeat purchases.

Additionally, Amazon Prime, Amazon's flagship subscription service, generated \$40.2 billion in revenue in 2023. This is an increase of about \$5 billion from the previous year when the service generated \$35.22 billion. billion USD. Amazon Prime not only brings in significant revenue but also greatly contributes to maintaining and expanding Amazon's loyal customer base.

Amazon's ability to deliver these personalized experiences stems from detailed segmentation of customer data. The company's algorithms identify and respond to

individual preferences, needs, and shopping habits, helping Amazon deliver highly relevant product recommendations. This strategy not only solidified Amazon's dominance in e-commerce but also drove significant revenue growth, with personalized recommendations accounting for approximately 35% of the company's total revenue ( according to McKinsey).

This success story illustrates the profound impact of effective customer segmentation and personalized marketing in e-commerce, showing how data-driven strategies can transform customer engagement and business efficiency.

## **6. Benefits and Recommendation**

### **a. Benefits**

Customer segmentation brings many benefits. It can help you create personalized or customized marketing campaigns, optimize resources, minimize risk, grow brand loyalty, and identify niche markets. Explore the some core benefits of marketing and audience segmentation.

- **Personalized digital marketing:** Segmentation helps marketers create more personalized digital advertising by providing insights and identifying audience characteristics, helping you target online marketing efforts to targeted customers. age groups, locations, purchasing habits, specific interests and more.
- **Optimize resources:** Marketers can optimize resources through segmentation by identifying and understanding distinct customer segments with diverse needs, preferences, and behaviors. Tailoring marketing strategies to meet the specific requirements of each segment helps allocate resources more efficiently and effectively, thereby maximizing the impact of marketing efforts.
- **Growth in the number of loyal customers of the brand:** When brands understand their customers' specific needs and preferences through segmentation, they can deliver targeted messaging, relevant offers, and customized products or services tailored to their needs each segment. This personalized approach can help grow a brand's customer loyalty and make customers feel valued and understood, thereby promoting a deeper emotional connection with the brand.
- **Identify niche markets:** Segmentation can help marketers identify niche markets. Advertisers can use audience insights to identify new or underserved markets and explore approaches to better serve their campaigns

and messages existing markets. These opportunities can be leveraged to strategically expand market reach and boost brand trust.

- **Improved Customer Service:** Amazon can provide better service to customers by understanding the needs and wants of each segment. This not only increases customer loyalty but also encourages them to shop more.
- **Analyze and Predict Customer Behavior:** Amazon can predict future shopping trends and customer behavior, helping them prepare and react quickly to changes in the market.

#### **b. Recommendation**

- **Engagement Strategy:** Develop targeted strategies to enhance customer engagement, as higher engagement levels are strongly linked to better customer satisfaction. This could involve personalized recommendations, gamification elements, or incentives for active participation.
- **User Experience Optimization:** Continuously optimize the user experience to maintain high satisfaction levels among frequent and regular users. Conduct usability testing, gather feedback, and make iterative improvements to ensure a seamless and enjoyable experience.
- **Customer Support Quality:** While the frequency of support interactions does not directly impact satisfaction, ensure that customer support services are of high quality. Prioritize efficient resolution of issues, friendly interactions, and effective communication to maintain customer trust and confidence.
- **Targeted Promotions:** Leverage the identified usage frequency segments (Regular, Frequent, Occasional) to develop targeted marketing campaigns and personalized offers. Tailor promotions and incentives to encourage occasional users to become more frequent and retain the loyal customer base.
- **Feedback Analysis:** Implement a robust feedback analysis system to identify areas for improvement and address any recurring issues or pain points reported by customers. Regularly monitor feedback trends and take proactive measures to address concerns and enhance customer satisfaction.

By implementing these recommendations, the company can effectively leverage the insights gained from the data analysis, fostering higher customer engagement, satisfaction, and retention, ultimately driving business growth and success.

## **IV. CUSTOMER RECOMMENDATION SYSTEM**

### **1. Definition**

A recommendation engine is a tool that uses machine learning to filter specific items from a larger dataset. There are different types of filtering methods, but in general, the recommendation engine uses existing user insights - what a user

has previously interacted with, what products were purchased, which movies were watched, and so on, to match other items similar to these, and then makes an item recommendation. The Amazon recommendation system works in a similar way.

A recommendation system helps an organization to create loyal customers and build trust by them desired products and services for which they came on your site. They recommend the products which are currently trending or highly rated and they can also recommend the products which bring maximum profit to the company.

## 2. About the research problem

E-commerce companies like Amazon , flipkart uses different recommendation systems to provide suggestions to the customers. Amazon uses currently item-item collaborative filtering, which scales to massive datasets and produces high quality recommendation system in the real time. This system is a kind of a information filtering system which seeks to predict the "rating" or preferences which user is interested in. In 2021, Amazon's net revenue from e-commerce sales was US\$470 billion, and about 35 percent of all sales on Amazon happen via recommendations.

This clearly elucidates the power of recommendations. In this case study, we look at how Amazon is using recommendations across its store and even off it, the technology behind the recommendation engine, and how you can implement similar strategies to increase engagement and conversions on your e-commerce store.

## 3. Types of recommendation

There are 3 main types of recommendation system:

- **Popularity-based systems (simple)** recommend the most popular items to a user. They tend to be not very accurate since they don't give any personalized suggestions based on the behavior and taste of the user.

Rating-Weight formula:

$$W = Rv + Cmv + m$$

$v$  = number of votes for the item

$m$  = minimum votes required to be listed (threshold)

$R$  = average rating for the item

$C$  = the mean vote across the whole dataset

- **Content-based systems** use user-profiles and item attributes to recommend items that match better the taste of the user. In this case, a matrix including user preferences and one including specific characteristics of each book are multiplied and suggestions are given based on the resulting weighted matrix.

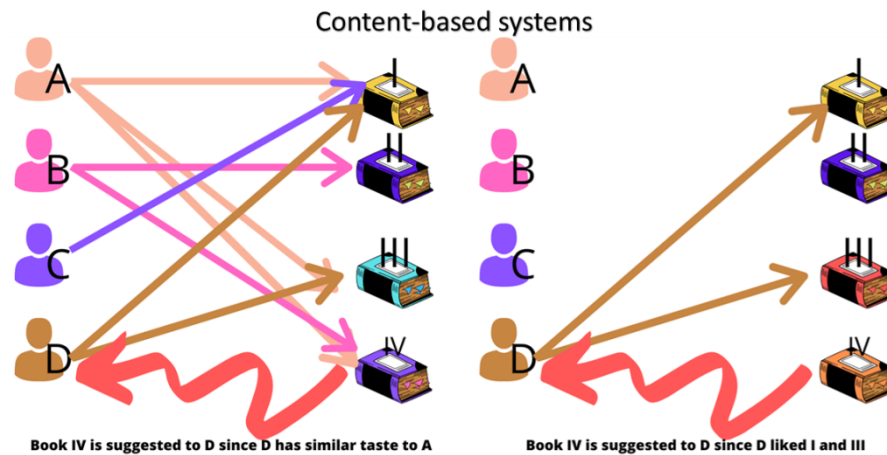


Figure 15. Content-based systems

Advantage:

- No need for data on users - no cold start and sparsity
- Able to recommend users with unique taste
- Able to recommend new and unpopular items

Limitation:

- Data should be in structured format
- Unable to use quality evaluation from other users

- **Collaborative filtering** systems collect and evaluate users' behavioral information in the form of feedback, ratings, preferences. Based on this information, they try to use similarities among users and items for predicting missing ratings and make suitable recommendations:

**User-based** systems use similarities among several users to suggest to the target user something that he/she didn't evaluate. Example: One needs to know if the target will like "Animal Farm" by George Orwell. This approach finds that a certain number N of users have rated other books similar to the target. To evaluate the rating that the target would give to the book, a weighted average of the n most similar users is taken. Similarity metrics are used to create the weighted matrix. If the predicted rating is high, then the book is suggested to the target.









	User Based			
				Animal Farm
A 	5	2	3	4
B 	3	3	5	5
C 	3	3	4	
D 	4		4	1
E 	5	2	3	?

Figure 16. User-Based Collaborative Filtering

Pearson correlation:

$$sim(u, v) = \frac{\sum_{i \in com(u, v)} (r_{u,i} - \bar{r}_u)(r_{v,i} - \bar{r}_v)}{\sqrt{\sum_{i \in com(u, v)} (r_{u,i} - \bar{r}_u)^2} \sqrt{\sum_{i \in com(u, v)} (r_{v,i} - \bar{r}_v)^2}}$$

- Let  $r_{u,i}$  be the rating of the  $i$ th item under user  $u$ ,  $\bar{r}_u$  be the average rating of user  $u$ , and  $com(u, v)$  be the set of items rated by both user  $u$  and user  $v$
- The similarity between user  $u$  and user  $v$  is then given by Pearson's correlation coefficient

Prediction function:

$$r_{up} = \bar{r}_u + \frac{\sum_{i \in users} sim(u, i) * r_{ip}}{\sum_{i \in users} |sim(u, i)|}$$

Advantage: No knowledge about item features needed

Limitation: New user item cold start problem: items with few rating cannot easily be recommended => Sparsity problem and popularity bias

**Item-based** systems use similarities among several items to suggest to the target user something that he/she didn't evaluate. - Example: One needs to know if the target will like "Animal Farm" by George Orwell. This approach finds that a certain number N of books have been rated similarly to "Animal Farm" by other users. To evaluate the rating that the target would give to the book, a weighted average of the n most similar books is taken. Similarity metrics are used to create the weighted matrix. If the predicted rating is high, then the book is suggested to the target



**Item Based**









				 Animal Farm
Genre A	<b>5</b>	<b>2</b>	<b>3</b>	
Genre B	<b>3</b>		<b>1</b>	<b>2</b>
Genre C	<b>3</b>		<b>1</b>	<b>5</b>
Genre D		<b>2</b>	<b>4</b>	<b>1</b>
Genre E	<b>5</b>	<b>2</b>	<b>3</b>	<b>4</b>
	 ECB	 DEB	 DAE	 ACD

Figure 17. Item-Based Collaborative Filtering

Cosine Similarity Metric:

$$\text{CosSim}(x, y) = \frac{\sum_i ((x_i - \bar{l})(y_i - \bar{l}))}{\sqrt{\sum_i (x_i - \bar{l})^2} \sqrt{\sum_i (y_i - \bar{l})^2}}$$

- The cosine similarity can calculate the smallest distance between product dimensions, regardless of the magnitude of the products themselves
- To obtain effective similarity, the result of the cosine similarity calculation should be as close to 1 as possible. If the result is 1, the two items are considered to have perfect similarities.

Prediction Function:

$$p_{u,i} = \bar{r}_u + \frac{\sum_{v \in nn(u)} \text{sim}(u, v) \cdot (r_{v,i} - \bar{r}_v)}{\sum_{v \in nn(u)} |\text{sim}(u, v)|}$$

Advantages:

- No knowledge about item features needed
- Better scalability, because correlations between limited number of items instead of very large number of users
- Reduced sparsity problem

Problems:

- New user, item cold start problem

Most widely-used recommendation approach:

- *k*-Nearest Neighbor: Similar to the spectral clustering based approach from the homework. Create a similarity matrix for pairs of users. Use *k*-NN to find the *k* closest users to a target user. Use the ratings of the *k* nearest neighbors to make predictions
- Matrix factorization: Decomposing large matrix of user-item ratings into smaller matrices of user and item features. Assumes that each user and item can be represented by a vector of latent features. Find the optimal feature vectors that minimize the difference between the estimated and actual ratings.

The final goal of this project would be to have a recommendation system of three different types:

- A popularity-based system
- A content-based system based on similarities among genres and authors to suggest to Bill a book similar to the ones he read but that he didn't evaluate.
- A collaborative filtering system

#### 4. Implementation

a. About the data

Link: <https://www.kaggle.com/datasets/saurav9786/amazon-product-reviews>

There are 3 files in our dataset which is extracted from several books selling websites:

- Books – first are about books which contain all the information related to books like an author, title, publication year, etc.
- Users – The second file contains registered user's information like user id, location.
- ratings – Ratings contain information like which user has given how much rating to which book.

The dataset is well-suited for collaborative filtering, specifically user-item collaborative filtering. The 'ratings' table holds the user-book interactions, presenting an opportunity to analyze user preferences.

b. Data analysis and visualization.

The most rated books in the data set are shown in the next figure.

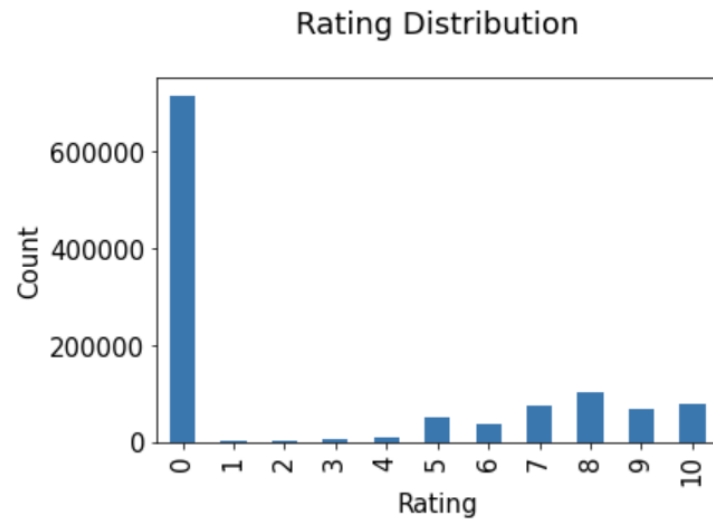


Figure 18. Rating Distribution

The last two items to be mentioned are the supports that are present in the dataset, shown in the next figure, and the most popular genres, shown in the successive figure.

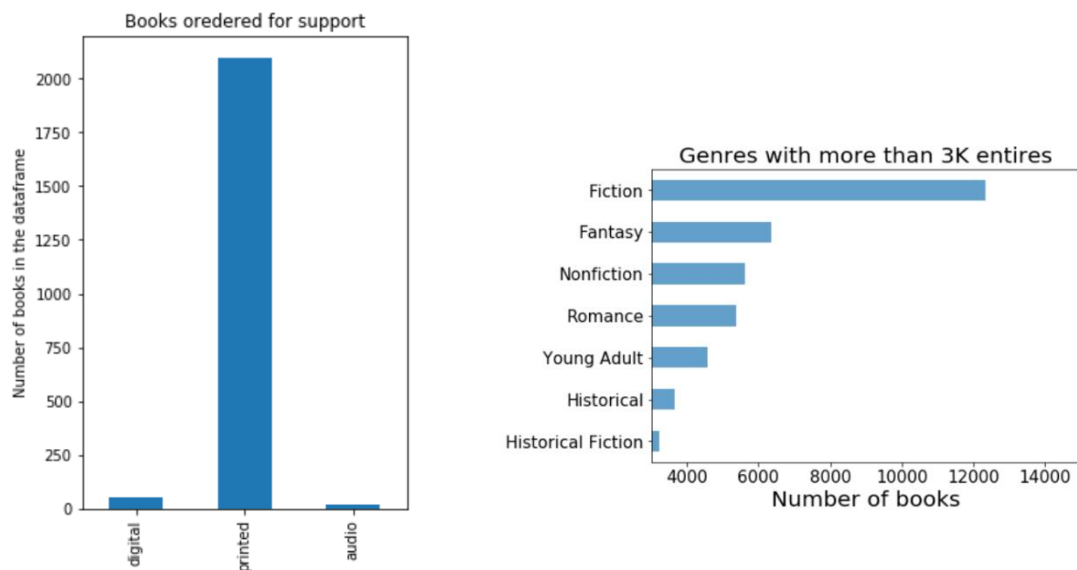


Figure 19. Books ordered for support & Genres more than 3K entires

In the dataset, there are more than 700 different genres. An interesting data to observe is that if any of the genres, despite being different is associated with the same kind of books. This is to reduce the number of different genres in case of further use of the data.

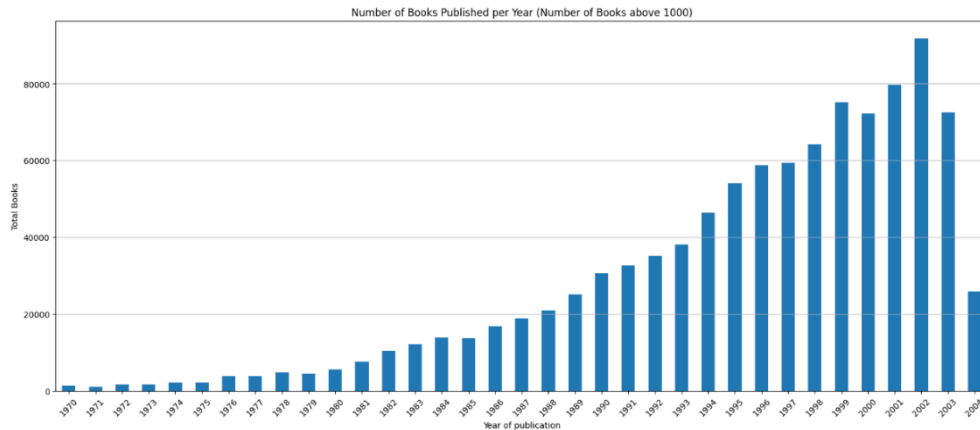


Figure 20. Number of books published per year

Based on this graph, it can be concluded that in 2002 the most people read or bought books, and the year with the least was 1971

c. The recommendation systems.

After having cleaned and pre-processed the data, one can implement the recommendation system. With the datasets available for this project, there are two recommendation systems that can be implemented. The first one is a popularity-based system and the second is a content-based system.

#### - **Popularity based recommendation system.**

Popularity based recommendation systems work with the trend or with the most popular items. In the case of this dataset, these are just the most evaluated books. It is a good recommendation system if there is no information about a user, but it can not make any use of the personalization that comes with other recommendation systems.

#### **Interpretation:**

- Book-Title is merged with Book-Author which then grouped by the merged field to get the total number of user ratings and the average ratings
- Get the Rating-Weight to get the weight of each Book-Title-Author since the rating average could be biased. Book A rated 10 with 1 number of voters weighted less than Book B rated 9.4 with 20 number of voters

#### **Observation:**

- Harry Potter books dominated the popular book list

	book	rank
	Suzanne Collins - The Hunger Games	1.0
	Stephenie Meyer - Twilight	2.0
	John Green - The Fault in Our Stars	3.0
	J.K. Rowling - Harry Potter and the Prisoner o...	4.0
	J.K. Rowling - Harry Potter and the Chamber of...	5.0
	J.K. Rowling - Harry Potter and the Order of t...	6.0
	J.K. Rowling - Harry Potter and the Half-Blood...	7.0
	J.K. Rowling - Harry Potter and the Deathly Ha...	8.0
	Anne Frank - The Diary of a Young Girl	9.0
	J.D. Salinger - The Catcher in the Rye	10.0

Figure 21. Top 10 most viewed books

- To recommend popular books we select only those books which have got more than 250 votes then rank the books according to average rating.
  - These suggestions are good to see the most popular books, but they probably do not fit customer's taste.
- ***Content-based recommendation system.***

To implement this system is enough to have a specific property of the book. In this case, I will simply use the genres as characteristics to evaluate the similarity of the books. A better system would be to use collaborative filtering but the data set does not have many different evaluations of the books but just an average.

### **Observation:**

To implement this system, it is essential to build up a matrix of the genres and calculate a cosine matrix. The books that have the most similar values are suggested as the best next book. To proceed further, one can eliminate all those books which do not have a genre, reducing the list of books to > 24 k, and Bill's list is integrated into the dataset. Among the 93 books from the original list, the system found 19 books that also have an assigned genre. In the matrix, the genres and the authors are evaluated, and then a simple mechanism to find the best rated for each book Bill's suggested is applied. In the next table, one can see the final result.Count Vectorizer.

Convert a collection of text documents to a matrix of token counts. It's a data table that is obtained after normalization of next-generation sequencing data.

## Things to do:

+ Initialize & Fit CountVectorizer into 'title' -> to create count\_matrix this is useful for cosine similarity

+ Check all words/ features in the vocabulary

+ Generate cosine similarity between Titles

Further steps, I will use Tf-Idf and Rake to see which of these give us better results. Here we consider overall document weightage of a word, useful while dealing with most frequent words:

	Bill's	Best next suggestion
0	Allie Brosh - Hyperbole and a Half: Unfortunate Situations, Flawed Coping Mechanisms, Mayhem, and Other Things That Happened	Liz Prince - Tomboy: A Graphic Memoir
1	Amor Towles - A Gentleman in Moscow	Edward Rutherfurd - Russia: the Novel of Russia
2	Carol S. Dweck - Mindset: The New Psychology of Success	John Medina - Brain Rules: 12 Principles for Surviving and Thriving at Work, Home, and School
3	Daniel Kahneman - Thinking, Fast and Slow	Dan Ariely - Predictably Irrational: The Hidden Forces That Shape Our Decisions
4	David Brooks - The Road to Character	Robert Greene - The Art of Seduction
5	Doris Kearns Goodwin - The Bully Pulpit: Theodore Roosevelt, William Howard Taft, and the Golden Age of Journalism	Doris Kearns Goodwin - No Ordinary Time: Franklin and Eleanor Roosevelt: The Home Front in World War II
6	Graeme Simsion - The Rosie Project	Graeme Simsion - The Rosie Effect
7	Hans Rosling - Factfulness: Ten Reasons We're Wrong About the World – and Why Things Are Better Than You Think	Barry Schwartz - The Paradox of Choice
8	Katherine Boo - Behind the Beautiful Forevers: Life, Death, and Hope in a Mumbai Undercity	Alice Albinia - Empires of the Indus: The Story of a River
9	Matthew Desmond - Evicted: Poverty and Profit in the American City	Michelle Alexander - The New Jim Crow: Mass Incarceration in the Age of Colorblindness
10	Matthew Walker - Why We Sleep: Unlocking the Power of Sleep and Dreams	Francis B. Colavita - Sensation, Perception, and the Aging Process
11	Peter H. Diamandis - Abundance: The Future Is Better Than You Think	Christopher Steiner - Automate This: How Algorithms Came to Rule Our World
12	Phil Knight - Shoe Dog: A Memoir by the Creator of Nike	Pat Conroy - My Losing Season: A Memoir
13	Randall Munroe - What If?: Serious Scientific Answers to Absurd Hypothetical Questions	Charles Pellegrino - The Last Train from Hiroshima: The Survivors Look Back
14	Richard Dawkins - The Magic of Reality: How We Know What's Really True	Daniel C. Dennett - Breaking the Spell: Religion as a Natural Phenomenon
15	Tara Westover - Educated: A Memoir	Barbara Eden - Jeannie Out of the Bottle
16	Thomas Piketty - Capital in the Twenty-First Century	Alan S. Blinder - After the Music Stopped: The Financial Crisis, the Response, and the Work Ahead
17	Trevor Noah - Born a Crime: Stories From a South African Childhood	Carrie Fisher - The Princess Diarist
18	Yuval Noah Harari - Homo Deus: A Brief History of Tomorrow	Jacob Bronowski - The Ascent of Man

Figure 22. Best suggestion for customer

- **Collaborative recommendation system.**

- User-based CF

## Things to do:

+ Selecting Users who have rated more than 200 books. There are only 811 users who have rated more than 200 books.

+ Selecting a Random User and Determining the Movies They Watched.

```
random_user = 99252
```

```
['Anne of Green Gables (Anne of Green Gables Novels (Paperback))',
 'Message in a Bottle',
 'The Summons',
 'To Kill a Mockingbird',
 'Whispers']
```

*Figure 23. Selected books by user*

- + Accessing the Data and IDs of Other Users Who Read the Same Books.
- + Determining the Users Most Similar to the User to be Recommended.
- + Create a new dataframe named `corr_df` which will contain the correlations between users.

	user_id_1	user_id_2	corr
0	102647	7346	-1.000000
1	99252	43842	-0.997509
2	16795	112001	-0.997406
3	201017	99252	-0.994135
4	11676	112001	-0.989743
...	...	...	...
113	11676	59172	0.985136
114	55548	187145	1.000000
115	236283	30972	1.000000
116	260897	231210	1.000000
117	7346	8362	NaN

*Figure 24. Correlations between users*

- + Filter users with high correlation (above 0.50) with the selected user and create a new dataframe named `top_users`.

	User-ID	corr
4	105979	0.967868
3	187145	0.755929
2	159033	0.755929
1	231210	0.755929
0	11676	0.534861

*Figure 25. User with highest correlation*

- + Finally calculating the Weighted Average Recommendation Score and recommending a Book

```
['The Illustrated Man',  
 'This Is the Story of Archibald Frisby: Who Was As Crazy for Science As A  
ny Kid Could Be (Reading Rainbow Book)',  
 'Tearing the Silence : On Being German in America',  
 'Something to Declare',  
 'Fail Safe',  
 'In the Time of the Butterflies']
```

---

*Figure 26. Recommend books for user*

- Item-based CF

### Things to do:

- + Selecting Items which have been rated more than 200. There are only 811 items which have rated more than 200.
- + Pivoting the table on the Book-Title as index, User-ID as columns with Book-Rating as values.
- + Compute the cosine similarity between books
- + This function will suggest the top 5 books based on their similarity scores closest to the given book.

```
: recommend('Message in a Bottle')  
  
Nights in Rodanthe  
The Mulberry Tree  
A Walk to Remember  
River's End  
Nightmares & Dreamscapes
```

*Figure 27. Best suggestion for item*

+ We do not want to find a similarity between users or books. we want to do that if there is user A who has read and liked x and y books, and user B has also liked this two books and now user A has read and liked some z book which is not read by B so we have to recommend z book to user B. So this is achieved using Matrix Factorization, we will create one matrix where columns will be users and indexes will be books and value will be rating. Like we have to create a Pivot table.



+ Then now we will calculate the euclidian distance of each vector with other vectors and on the basis of that distance we will get to know which book is similar.

+ Using **cosine\_similarity Distance** to find the all recommended Books:

```
recommend('Harry Potter and the Chamber of Secrets (Book 2)')

[['Harry Potter and the Prisoner of Azkaban (Book 3)',
  'J. K. Rowling',
  'http://images.amazon.com/images/P/0439136350.01.MZZZZZZZ.jpg'],
 ['Harry Potter and the Goblet of Fire (Book 4)',
  'J. K. Rowling',
  'http://images.amazon.com/images/P/0439139597.01.MZZZZZZZ.jpg'],
 ["Harry Potter and the Sorcerer's Stone (Harry Potter (Paperback))",
  'J. K. Rowling',
  'http://images.amazon.com/images/P/059035342X.01.MZZZZZZZ.jpg'],
 ["Harry Potter and the Sorcerer's Stone (Book 1)",
  'J. K. Rowling',
  'http://images.amazon.com/images/P/0590353403.01.MZZZZZZZ.jpg']]
```

Figure 28. Top 5 recommended books for item

## 5. Demo

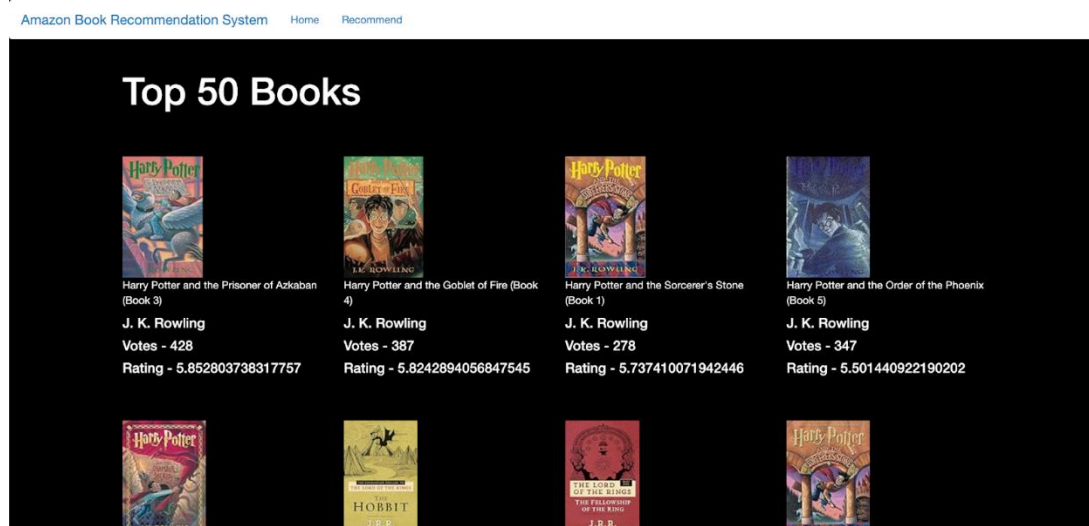


Figure 29. Demo Amazon Book Recommendation System

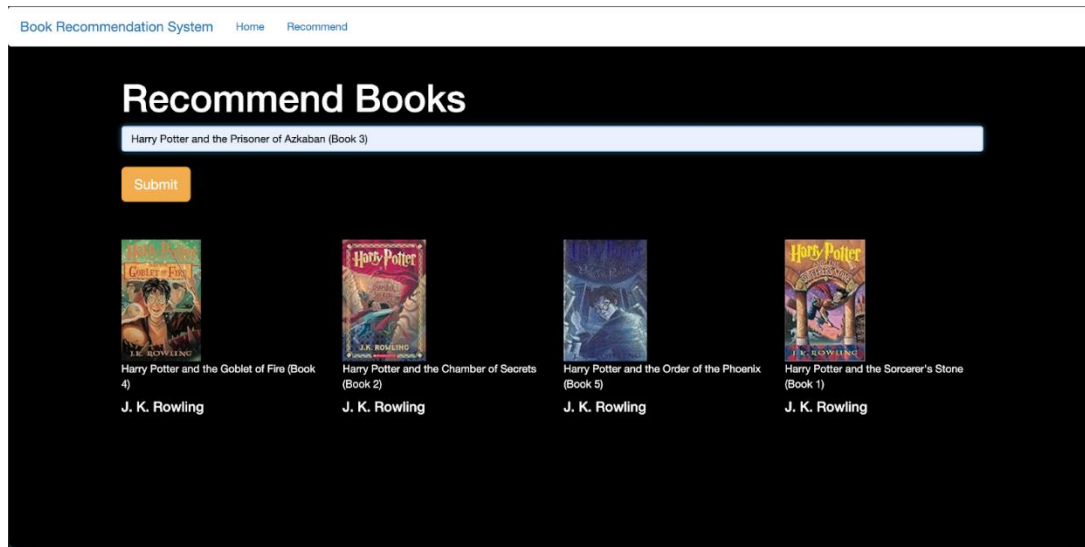


Figure 30. Demo Amazon Book Recommendation System

## 6. Conclusion

To conclude, the main issue of this work was to find a meaningful way to unify the data and clean them. The recommendation system implementation, that in theory does not look so difficult became quite complex to implement due to the type of data I started with.

Like any of this project, I am trying to work on, the main goal is to learn as much as possible. I hope you enjoyed the reading and got some good suggestions both for Bill's list and from the recommendation system.

## V. CONCLUSION

Amazon's success in applying advanced machine learning and artificial intelligence techniques has set a new standard in the e-commerce industry. By effectively implementing customer churn prediction, customer segmentation, and personalized recommendation systems, Amazon has significantly enhanced customer retention, satisfaction, and overall business performance.

The use of predictive models has allowed Amazon to anticipate customer behavior and deliver targeted interventions. Customer segmentation techniques have enabled Amazon to categorize its vast customer base into distinct segments. This strategic approach has led to more effective marketing campaigns and improved customer engagement. Moreover, Amazon's recommendation system has not only increased sales but also strengthened customer loyalty by consistently offering relevant product suggestions.

## REFERENCES

- Lalwani, P., Mishra, M.K., Chadha, J.S. and Sethi, P. (2021) Customer Churn Prediction System: A Machine Learning Approach. *Computing*, 104, 271-294.
- Gopal, P. and MohdNawi, N.B. (2021, December) A Survey on Customer Churn Prediction Using Machine Learning and Data Mining Techniques in E-Commerce. 2021 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE), Brisbane, 8-10 December 2021, 1-8.
- How Many Products Does Amazon Sell? – March 2021 (2022), *Scraphero*, Available at: How Many Products Does Amazon Sell? - March 2021
- Hernández, Blanca, Julio Jiménez, and M. José Martín. "Customer behavior in electronic commerce: The moderating effect of e-purchasing experience." *Journal of business research*, 63(9-10) (2010): 964-971.
- Adamopoulos P. and Tuzhilin A. 2014 *Proceedings of the 8th ACM Conference on Recommender systems* (ACM) On over-specialization and concentration bias of recommendations: Probabilistic neighborhood selection in collaborative filtering systems 153-160.
- Su X. and Khoshgoftaar T. M. 2009 A survey of collaborative filtering techniques *Adv. in Artif. Intell.* 4 2-4 Jan. 2009
- Gunawardana A. and Meek C. 2009 *Proceedings of the Third ACM Conference on Recommender Systems, RecSys'09* (New York, NY, USA: ACM) A unified approach to building hybrid recommender systems 117-124
- Marko Balabanovic and Yoav Shoham 1997. Fab: Contentbased, collaborative recommendation. *Communications of the ACM*, 40(3): pp. 66-72.
- Chumki Basu, Haym Hirsh, and William Cohen 1998. Recommendation as classification: using social and content- 166 based information in recommendation. In *Proceedings of the 1998 Workshop on Recommender Systems*, pages 11-15.