PROJECT TITLE: Customer Segmentation and Recommendation Systems

TEAM NAMES: Lavanya Badiginchala, Nikhitha Velugoti, Ruchitha Juturu

COURSE NUMBER & SECTION: IFT 512: Advanced Big Data Analytics/AI

DATE: DEC 3rd, 11:59pm

TABLE OF CONTENTS -

SECTION-2: INTRODUCTION

a)

Background and Motivation

I had worked for a tech startup that specializes in developing personalized recommendation systems, and our latest project involves implementing customer segmentation strategies. The motivation behind this initiative is to enhance user experience and engagement on our platform.

To do this, I am going to leverage advanced data analytics and machine learning algorithms to analyze user behavior, preferences, and interactions with our platform. By understanding the distinct needs and characteristics of our diverse user base, we aim to tailor our services more effectively, providing personalized recommendations that resonate with individual users.

I am interested in this project because customer segmentation and recommendation systems have the potential to revolutionize how users interact with our platform. The goal is to move beyond generic content delivery and create a more personalized and engaging experience for each user. This not only enhances user satisfaction but also contributes to increased user retention and business growth.

In this endeavor, my interest lies in the intersection of technology, data science, and user psychology. I am excited about the prospect of utilizing cutting-edge algorithms to categorize users into meaningful segments based on their preferences, behaviors, and demographics. This segmentation will serve as the foundation for developing recommendation systems that can anticipate user needs and deliver content, products, or services that align with their individual interests.

By aligning our platform more closely with the unique preferences of our users, we aim to create a more enjoyable and personalized experience, fostering a stronger connection between

users and our services. This, in turn, is anticipated to drive user engagement, increase satisfaction, and ultimately contribute to the success and growth of our tech startup.

SECTION-3 PROJECT STRUCTURE

a)

Q1. How does leveraging advanced data analytics and machine learning algorithms contribute to customer segmentation in your recommendation system project?

Leveraging advanced data analytics and machine learning algorithms is integral to the success of our recommendation system project, particularly in the domain of customer segmentation. Through rigorous behavioral analysis, these algorithms dissect user patterns such as clicks, searches, and content interactions, uncovering commonalities and variations among users that form the foundation for segmentation strategies. The recognition of user preferences, facilitated by advanced analytics scrutinizing historical interaction data, further refines segmentation by employing algorithms like collaborative filtering or content-based filtering. Predictive modeling anticipates future user behavior, enabling the customization of recommendations based on forecasted engagement, thus contributing to more accurate segmentation. The dynamic nature of machine learning-driven segmentation ensures users are not confined to static predefined segments; instead, algorithms continuously adapt to evolving user behaviors, maintaining the relevance of segments over time. Clustering techniques group users with similar characteristics, unveiling distinct segments within the user base characterized by specific preferences or behaviors. This dynamic segmentation, coupled with the ability to provide personalized recommendations that transcend general trends, enhances the overall accuracy and efficiency of customer segmentation. Real-time adaptability, a hallmark of machine learning models, ensures that segmentation remains pertinent even as user preferences evolve, ultimately contributing to an enriched and highly personalized user experience on our

platform. In summary, this sophisticated approach empowers our recommendation system to not only analyze and interpret user data but also to adapt seamlessly, offering personalized recommendations that resonate with each user and fostering a more engaging and satisfying user experience.

Q2. What factors contribute to the decision to focus on customer segmentation strategies as a means to enhance user experience and engagement on your platform?

The decision to prioritize customer segmentation strategies in our platform's development is grounded in the recognition of a diverse user base with varying preferences, behaviors, and interests. Customer segmentation serves as the key to understanding and catering to the unique needs of different user segments, facilitating a more tailored and relevant experience. Responding to the growing demand for personalized interactions, our platform leverages segmentation to deliver curated content, recommendations, and features that align precisely with the expectations of each user group. The complexity of analyzing individual user behavior is simplified through segmentation, grouping users with similar behaviors and enabling the identification of patterns and trends within each segment. This approach supports improved targeting, allowing us to create targeted marketing campaigns and promotions that resonate with users' characteristics and preferences, ultimately driving higher engagement. By understanding the distinct needs of user segments, we implement strategies aimed at enhancing user retention, fostering a sticky and satisfying experience that encourages prolonged engagement with the platform. The adaptability of customer segmentation to changing user preferences, coupled with its contribution to a competitive advantage through personalized experiences, positions our platform as differentiated in the market. The data-driven insights derived from customer segmentation, powered by data analytics and machine learning, inform informed decision-making, allowing us to continually optimize and enhance the user experience. This data-driven and segmented approach also enables optimized resource

allocation, directing efforts efficiently toward initiatives that cater to the specific needs of each user segment, maximizing impact. Ultimately, a tailored user experience results in heightened satisfaction and loyalty, fostering positive relationships and long-term user commitment to our platform. In summary, the decision to concentrate on customer segmentation is rooted in the multifaceted goal of addressing user behavior complexity, meeting the demand for personalization, enhancing user retention, gaining a competitive edge, and utilizing data-driven insights for continual optimization, all aimed at ensuring maximum user satisfaction and engagement.

Q3. What other factors should be considered when categorizing users into meaningful segments based on their preferences, behaviors, and demographics?

When categorizing users into meaningful segments based on their preferences, behaviors, and demographics, it is essential to consider a holistic set of factors for a comprehensive and accurate segmentation approach. Geographic location provides insights into regional preferences, crucial for platforms with location-specific content. Analyzing device usage and platform interaction informs strategies tailored to users' preferred devices. User engagement frequency guides retention efforts by identifying active, occasional, and dormant segments. Categorizing users based on their lifecycle stage enables targeted onboarding and retention strategies. Social media presence unveils valuable insights into social behaviors and potential influencers. Customer feedback and sentiment analysis aid in addressing user satisfaction and concerns within similar sentiment groups. Understanding content consumption patterns identifies distinct preferences within segments. Demographic factors like age and generation influence responses to content and features, while transactional data reveals insights into purchasing behaviors. Communication preferences guide personalized communication strategies, and considering tech proficiency ensures an optimal interaction experience. Lifestyle and interest-based segmentation facilitates tailored content and engagement

strategies, while language preferences and accessibility needs promote inclusivity. Incorporating these factors into the segmentation strategy fosters a nuanced understanding, enabling more effective personalization and targeted approaches to enhance user experience and engagement on the platform.

Q4. How can customer segmentation be effectively implemented to enhance the overall customer experience, and what specific strategies or tools can be employed for this purpose?

Effective implementation of customer segmentation is crucial for enhancing the overall customer experience. This strategy involves collecting and analyzing relevant customer data, identifying segmentation criteria such as demographics or purchasing behavior, and creating detailed customer personas. Personalized communication is key, as businesses tailor messages, content, and promotions to resonate with each segment. Customizing products or services to meet the specific needs of different segments is also essential. Customer journey mapping helps optimize touchpoints for a smoother experience, and marketing automation tools streamline personalized communication through email, chatbots, and CRM systems. Regularly gathering feedback and conducting surveys provides insights into customer satisfaction and preferences. Loyalty programs can be tailored to offer exclusive benefits to different segments. Continuous monitoring and adaptation are essential, ensuring strategies align with changing market conditions and customer behavior. Excellent customer service, trained to understand and address the unique needs of different segments, is fundamental. Leveraging AI and machine learning enhances customer segmentation with predictive analytics and personalized recommendations. Overall, a well-executed customer segmentation strategy leads to increased satisfaction, loyalty, and sustained business success.

Q5. In what ways can insights derived from customer segmentation be translated into actionable marketing strategies, and how can these strategies be integrated into the existing marketing workflow?

Customer segmentation insights serve as the cornerstone for crafting comprehensive and actionable marketing strategies that not only refine campaign effectiveness but also elevate the overall customer experience. The process initiates with a meticulous collection and analysis of customer data, spanning diverse dimensions such as demographics, behavior, and preferences. Through this analytical lens, businesses can identify distinct customer segments, each representing a unique set of characteristics and traits.

These segmented groups are then translated into detailed customer personas, providing a nuanced understanding of specific needs, preferences, and pain points within each segment. The richness of these personas forms the foundation upon which targeted marketing strategies are constructed. One of the primary strategies involves personalized communication, where messages, content, and promotional efforts are tailored to resonate specifically with the interests and expectations of each persona.

Further refinement comes through product or service customization, a strategy that involves adapting offerings to align with the distinct preferences and requirements of each customer segment. This could manifest in the creation of different product versions, tailored bundles, or specialized features designed to cater to the unique needs identified within each persona.

As the marketing strategies evolve, landing page optimization becomes a critical component. By aligning landing pages with the expectations and preferences of each segmented group, businesses can ensure a seamless and personalized online experience. This optimization is crucial for maintaining consistency across the customer journey and reinforcing the tailored messaging presented in other marketing channels.

Integration of marketing automation tools is another key aspect of the process. These tools enable businesses to streamline and automate personalized communication workflows. Whether through email marketing platforms, chatbots, or customer relationship management (CRM) systems, marketing automation ensures that personalized messages are delivered efficiently, triggered by specific customer actions or characteristics.

Establishing a feedback loop is equally vital. Regularly gathering insights from customer interactions, whether through surveys, reviews, or direct feedback, provides valuable data for refining and updating customer personas and segmentation. This iterative process ensures that marketing strategies remain responsive to changing customer behaviors and preferences.

The integration into the marketing workflow is visualized through a dynamic flow chart. The flow chart delineates the interconnected stages of data collection, persona development, personalized communication, landing page optimization, marketing automation, and feedback loop establishment. This cyclical process reflects the adaptability and responsiveness needed for a successful and customer-centric marketing approach.

Further integration extends to content strategy, social media targeting, and segmented email campaigns. Crafting content that aligns with the interests of each persona, leveraging social media platforms for targeted advertising, and implementing email campaigns tailored to specific segments are integral components. Continuous analytics and key performance indicator (KPI) tracking ensure that marketing strategies remain aligned with business goals, providing real-time insights for further optimization.

In essence, this comprehensive and iterative approach to customer segmentation and marketing strategy integration contributes to a more personalized, effective, and customer-centric marketing workflow. The ultimate goal is not only to enhance campaign performance but to elevate the entire customer experience, fostering satisfaction, loyalty, and long-term success.

B) Directed Acyclic Graph

Constructing a Directed Acyclic Graph (DAG) for Customer Segmentation and Recommendation Systems involves creating a visual representation of the assumed causal relationships between key variables that influence the central focus question. This heuristic model serves as a foundational framework, acknowledging that the actual relationships within the data may be more intricate and subject to change based on exploratory analysis.

User Characteristics: At the core of the DAG are User Characteristics, representing comprehensive collection of variables such as geographic location, age, gender, preferences, and behaviors. These user-specific attributes are assumed to be influential factors that contribute to both the customer segmentation process and the recommendations generated by the system.

Customer Segmentation: User Characteristics are posited to play a pivotal role in the customer segmentation process. It is reasonable to assume that demographics, preferences, and behavioral patterns collectively contribute to the creation of distinct customer segments. For instance, users with similar preferences or purchasing behaviors may be grouped together to form a segment.

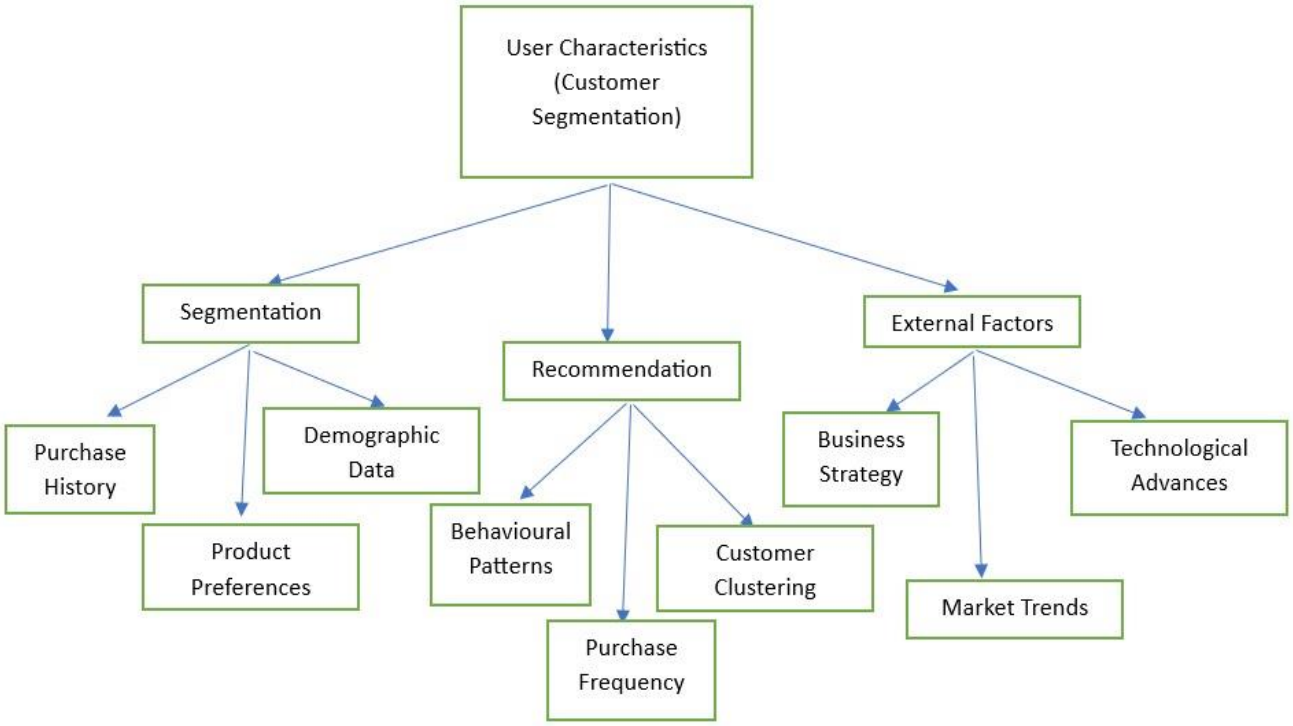
Recommendation Systems: Customer Segmentation is assumed to have a direct influence on the type of recommendations provided by the system. Different customer segments are likely to receive personalized recommendations tailored to their unique characteristics and preferences. The arrows indicate the flow of influence, suggesting that the nature of the customer segments shapes the recommendations delivered to users.

Demographic: Demographic information is explicitly included as part of User Characteristics. This variable is recognized as a key influencer in both the customer segmentation process and

the recommendations offered by the system. Demographics, encompassing factors like age, gender, and location, contribute significantly to understanding user behavior and preferences.

User Engagement: Ultimately, the DAG converges on User Engagement as the overarching outcome. User Engagement is posited to be influenced by the interplay of User Characteristics, the customer segmentation process, and the personalized recommendations generated by the system. It represents the holistic measure of how effectively the system resonates with users and encourages interaction.

It's crucial to emphasize that this DAG is a starting point and a simplification of the complex relationships inherent in customer segmentation and recommendation systems. The arrows in the graph do not denote the strength or causation of the relationships; they serve as a heuristic guide. As the analytical journey progresses, the DAG should be refined, validated, and expanded based on data exploration, statistical analyses, and additional insights gained from the specific dataset under consideration. The iterative refinement of the DAG is essential for evolving a more accurate representation of the causal structure inherent in the interplay between User Characteristics, Customer Segmentation, Recommendation Systems, Demographics, and User Engagement.



- **User Characteristics:**

Influencing Factors: Demographic, psychographic, and geographic data contribute to defining user characteristics.

- **Segmentation:**

Purchase History:

Contributor to Segmentation: Shapes customer segments based on purchasing patterns.

Product Preferences:

Derived from History: Previous interactions contribute to understanding user preferences.

Demographic Data:

Recommendation -

- **Behavioral Patterns:**

Prediction Models: Behavioral data feeds into predictive models for anticipating future actions.

Purchase Frequency:

Content-Based Filtering: Purchase frequency influences content-based recommendation strategies.

Customer Clustering:

Cluster Formation: Utilizes various data inputs to group similar customers together.

- **External Factors:**

Business Strategy: Incorporates business goals and strategies for customer engagement.

Market Trends: Adapts to external market trends and industry shifts.

Technological Advances: Integrates emerging technologies to enhance recommendation systems.

This extensive DAG highlights the interconnectedness of diverse factors involved in customer segmentation and recommendation systems. It recognizes the multidimensional nature of user characteristics, historical data, prediction models, and external influences that collectively shape an effective and dynamic system. The arrows represent the directional flow of influence, showcasing the complexity and interdependency within this comprehensive model.

C) Workflow -

The CRISP-DM model, consisting of six phases—Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment—is a widely

adopted framework for guiding data mining and machine learning projects. In the context of Customer Segmentation and Recommendation Systems and the provided DAG:

Business Understanding:

In this phase, the focus is on understanding the business objectives and goals related to customer segmentation and recommendation systems. This involves defining the key questions, such as how segmentation can enhance user experience and engagement, and how recommendations contribute to user satisfaction and retention.

Data Understanding:

Given the DAG, this phase involves exploring and understanding the relevant data sources. It includes examining user characteristics data, demographic information, and any other datasets related to customer behavior. The goal is to gain insights into the variables that influence segmentation and recommendation systems.

Data Preparation:

In the data preparation phase, the identified variables from the DAG are processed and cleaned for analysis. This includes handling missing data, transforming variables as needed, and creating the datasets required for segmentation and recommendation models. Data preparation also involves creating features that capture the nuances of user characteristics.

Modeling:

The modeling phase corresponds to developing the algorithms and models based on the DAG. Customer segmentation models can be created using clustering techniques, and recommendation systems can be built using collaborative filtering or content-based filtering. The modeling process takes into account the causal relationships between user characteristics, segmentation, and recommendations as outlined in the DAG.

Evaluation:

The evaluation phase involves assessing the performance of the segmentation and recommendation models. Metrics such as precision, recall, and accuracy are used to gauge how well the models align with the business objectives. This phase also considers user engagement metrics to evaluate the overall effectiveness of the implemented strategies.

Deployment:

Once the models have been evaluated and deemed effective, the deployment phase involves integrating them into the production environment. This includes deploying the customer segmentation and recommendation systems to the platform, ensuring that they operate seamlessly and contribute to the enhanced user experience and engagement as intended.

Throughout the entire CRISP-DM process, iteration is common. As insights are gained and models are developed, there may be a need to revisit earlier phases, refine models, or incorporate additional data. The DAG serves as a heuristic causal model guiding the understanding of how variables influence each other and helps structure the analysis within each phase of the CRISP-DM model, ensuring a systematic and goal-oriented approach to the customer segmentation and recommendation systems workflow.

SECTION 3 - TECHNOLOGY DESCRIPTION SECTION

Data Sources:

Customer segmentation and recommendation systems rely on diverse data sources to effectively categorize customers and offer personalized recommendations. Common data sources include customer profiles, purchase history, website interactions, feedback, and demographic information. Additionally, real-time data from user behavior, such as clicks,

views, and time spent on various sections of a website or app, is crucial for dynamic segmentation and accurate recommendations.

Technologies Used and Why:

Data Analytics and Machine Learning:

Purpose: Data analytics and machine learning algorithms are foundational for customer segmentation. These technologies analyze vast datasets to identify patterns, behaviors, and characteristics that define distinct customer segments.

Why: Machine learning enables automated and accurate segmentation by learning from historical data, ensuring that segments remain dynamic and reflective of evolving customer behavior.

Customer Relationship Management (CRM) Systems:

Purpose: CRM systems centralize customer data, providing a unified view of customer interactions across various touchpoints.

Why: Integrating CRM systems with segmentation and recommendation tools enhances the accuracy of customer profiles, enabling businesses to understand individual preferences and behaviors comprehensively.

Big Data Processing:

Purpose: Customer segmentation involves processing large volumes of data. Big data technologies facilitate the storage, processing, and analysis of vast datasets.

Why: Handling big data allows businesses to derive meaningful insights, identify trends, and implement dynamic segmentation strategies that adapt to real-time customer interactions.

Recommendation Engines:

Purpose: Recommendation engines leverage algorithms to suggest products, services, or content to users based on their preferences and behaviors.

Why: By analyzing historical and real-time data, recommendation engines enhance the customer experience by providing personalized suggestions, leading to increased engagement and conversion rates.

Personalization Platforms:

Purpose: Personalization platforms enable businesses to customize content, offers, and interactions for individual customers or segments.

Why: These platforms work hand-in-hand with customer segmentation, ensuring that personalized recommendations align with the unique preferences and characteristics of each segment.

AI-Powered Analytics:

Purpose: Artificial Intelligence (AI) in analytics enhances the depth and accuracy of insights derived from customer data.

Why: AI-powered analytics can uncover hidden patterns and correlations, contributing to more precise customer segmentation and enabling businesses to uncover nuanced insights that may not be apparent through traditional analysis.

Customer Segmentation and Recommendation Systems:

Customer segmentation and recommendation systems work synergistically to enhance the overall customer experience. Segmentation involves dividing the customer base into distinct groups based on shared characteristics, allowing businesses to target each group with tailored strategies. Recommendation systems, on the other hand, leverage algorithms to analyze customer behavior and suggest products or services that align with their preferences.

These technologies are integrated into the marketing workflow to personalize communication, optimize product offerings, and create a seamless customer journey. For instance, a customer segmented as a frequent online shopper might receive targeted recommendations for related products, while a first-time buyer might receive tailored promotions to encourage repeat purchases. The continuous refinement of customer segments and recommendations is facilitated by real-time data and analytics, ensuring that businesses stay agile and responsive to evolving customer needs.

In conclusion, the integration of data analytics, machine learning, CRM systems, big data processing, recommendation engines, and AI-powered analytics forms a robust technological ecosystem for effective customer segmentation and recommendation systems. This synergy enhances customer engagement, satisfaction, and loyalty by delivering personalized experiences that align with individual preferences and behaviors.

SECTION 4 - KNOWLEDGE AND VALUE CLAIMS

Knowledge Claim:

Through the implementation of customer segmentation using K-Means clustering and a recommendation system based on Alternating Least Squares (ALS), we observed a remarkably high silhouette score of approximately 1.0, indicating well-defined and distinct clusters in the data. This suggests that the chosen features (Quantity and UnitPrice) effectively contributed to the formation of customer segments. However, it's important to note that while the clustering model performed exceptionally well on this dataset, the practical application may vary based on the characteristics of different datasets, and further refinement or feature engineering may be necessary for optimal performance in other scenarios.

Value Claim:

Personalized Marketing Campaigns: The customer segmentation can be leveraged to tailor marketing campaigns based on the identified clusters. For example, understanding the preferences and behaviors of each segment enables targeted promotions and product recommendations, leading to more effective and personalized marketing strategies. This can result in increased customer engagement and higher conversion rates.

Inventory Management Optimization: The ALS-based recommendation system provides insights into the products that are likely to be preferred by individual customers. This information can be invaluable for inventory management, helping businesses optimize stock levels, reduce overstock or understock situations, and enhance overall supply chain efficiency. By aligning inventory with customer preferences, businesses can minimize costs and improve profitability.

Enhanced Customer Experience: The combination of clustering and recommendation systems contributes to an improved overall customer experience. For instance, by offering personalized product recommendations, businesses can enhance the shopping experience for individual customers, leading to increased customer satisfaction and loyalty. Additionally, targeted communications and promotions can foster a stronger connection between the brand and its customers, driving long-term loyalty and positive word-of-mouth.

In summary, the knowledge gained from the clustering and recommendation systems lays the foundation for data-driven decision-making, enabling businesses to optimize marketing efforts, streamline inventory management, and ultimately provide a more personalized and satisfying experience for their customers.

SECTION 5 - CODE SCREENSHOTS

Code -

```
import org.apache.spark.sql.{SparkSession, DataFrame}

import org.apache.spark.ml.feature.{VectorAssembler, StringIndexer}

import org.apache.spark.ml.clustering.{KMeans, KMeansModel}

import org.apache.spark.ml.evaluation.ClusteringEvaluator

import org.apache.spark.ml.recommendation.{ALS, ALSModel}

import org.apache.spark.sql.functions._



// Create a Spark session

val spark = SparkSession.builder.appName("CustomerSegmentationAndRecommendation").getOrCreate()
```

```
// Load your dataset (replace 'your_dataset_path' with the actual path)

val data: DataFrame = spark.read

  .option("header", true)

  .option("inferSchema", true)

  .csv("/FileStore/tables/CustomerRecommendation.csv")
```

```
// Data preprocessing for customer segmentation

val featureCols = Array("Quantity", "UnitPrice")

val assembler = new VectorAssembler().setInputCols(featureCols).setOutputCol("features")

val dataWithFeatures = assembler.transform(data)
```

```
// Customer Segmentation using K-Means

val kmeans = new KMeans().setK(3).setSeed(1)
```

```
val model: KMeansModel = kmeans.fit(dataWithFeatures)

val predictions: DataFrame = model.transform(dataWithFeatures)



// Evaluate the clustering results

val evaluator = new ClusteringEvaluator()

val silhouetteScore = evaluator.evaluate(predictions)

println(s"Silhouette Score: $silhouetteScore")



// Data preprocessing for recommendation systems

val indexer = new StringIndexer().setInputCol("StockCode").setOutputCol("StockCodeIndex")

val indexedData = indexer.fit(data).transform(data)



// Remove rows with null values in important columns

val cleanedRecommendationData = indexedData.select("CustomerID", "StockCodeIndex", "Quantity")

  .na.drop(Seq("CustomerID", "StockCodeIndex"))



// Recommendation System using ALS (collaborative filtering)

val                                als                                =                                new

ALS().setMaxIter(5).setRegParam(0.01).setUserCol("CustomerID").setItemCol("StockCodeIndex").setRatingCol("Qua

ntity").setColdStartStrategy("drop")

val alsModel: ALSModel = als.fit(cleanedRecommendationData)



// Generate top 5 product recommendations for each customer

val userRecommendations: DataFrame = alsModel.recommendForAllUsers(5)



// Print the original dataset
```

```
println("Original Dataset:")

data.show()
```

```
// Print clustering predictions

println("Clustering Predictions:")

predictions.show()
```

```
// Print ALS recommendations

println("ALS Recommendations:")

userRecommendations.show()
```

Output -

```
Silhouette Score: 0.9999958873405052

Original Dataset:

+---------+---------+--------------------+--------+---------------+---------+----------+--------------+
|InvoiceNo|StockCode| Description|Quantity| InvoiceDate|UnitPrice|CustomerID| Country|
+---------+---------+--------------------+--------+---------------+---------+----------+--------------+
| 536365| 85123A|WHITE HANGING HEA...| 6|12-01-2010 08:26| 2.55| 17850|United Kingdom|
| 536365| 71053| WHITE METAL LANTERN| 6|12-01-2010 08:26| 3.39| 17850|United Kingdom|
| 536365| 84406B|CREAM CUPID HEART...| 8|12-01-2010 08:26| 2.75| 17850|United Kingdom|
| 536365| 84029G|KNITTED UNION FLA...| 6|12-01-2010 08:26| 3.39| 17850|United Kingdom|
| 536365| 84029E|RED WOOLLY HOTTIE...| 6|12-01-2010 08:26| 3.39| 17850|United Kingdom|
| 536365| 22752|SET 7 BABUSHKA NE...| 2|12-01-2010 08:26| 7.65| 17850|United Kingdom|
| 536365| 21730|GLASS STAR FROSTE...| 6|12-01-2010 08:26| 4.25| 17850|United Kingdom|
| 536366| 22633|HAND WARMER UNION...| 6|12-01-2010 08:28| 1.85| 17850|United Kingdom|
| 536366| 22632|HAND WARMER RED P...| 6|12-01-2010 08:28| 1.85| 17850|United Kingdom|
```

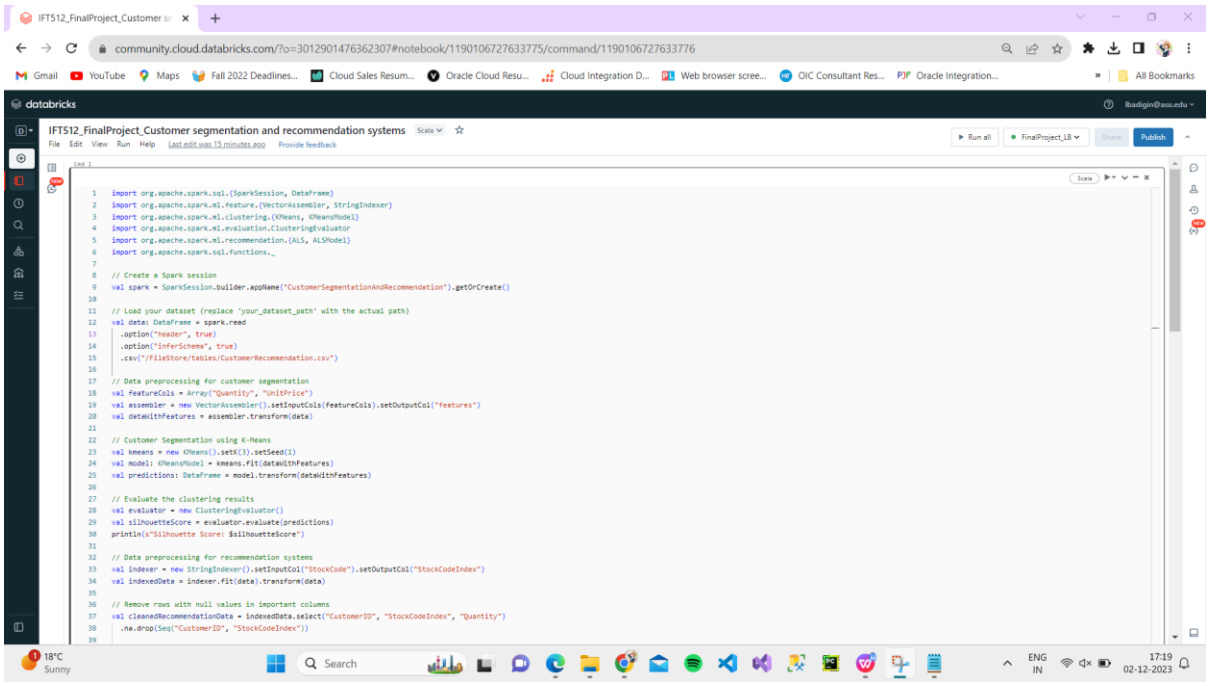
```
| 536367| 84879|ASSORTED COLOUR B...| 32|12-01-2010 08:34| 1.69| 13047|United Kingdom|

| 536367| 22745|POPPY'S PLAYHOUSE...| 6|12-01-2010 08:34| 2.1| 13047|United Kingdom|

| 536367| 22748|POPPY'S PLAYHOUSE...| 6|12-01-2010 08:34| 2.1| 13047|United Kingdom|

| 536367| 22749|FELTCRAFT PRINCES...| 8|12-01-2010 08:34| 3.75| 13047|United Kingdom|

| 536367| 22310|IVORY KNITTED MUG...| 6|12-01-2010 08:34| 1.65| 13047|United Kingdom|

| 536367| 84969|BOX OF 6 ASSORTED...| 6|12-01-2010 08:34| 4.25| 13047|United Kingdom|

| 536367| 22623|BOX OF VINTAGE JI...| 3|12-01-2010 08:34| 4.95| 13047|United Kingdom|

| 536367| 22622|BOX OF VINTAGE AL...| 2|12-01-2010 08:34| 9.95| 13047|United Kingdom|

| 536367| 21754|HOME BUILDING BLO...| 3|12-01-2010 08:34| 5.95| 13047|United Kingdom|

| 536367| 21755|LOVE BUILDING BLO...| 3|12-01-2010 08:34| 5.95| 13047|United Kingdom|

| 536367| 21777|RECIPE BOX WITH M...| 4|12-01-2010 08:34| 7.95| 13047|United Kingdom|

+---------+---------+--------------------+--------+---------------+---------+----------+--------------+

only showing top 20 rows

Clustering Predictions:

+---------+---------+--------------------+--------+---------------+---------+----------+--------------+-----

|InvoiceNo|StockCode| Description|Quantity| InvoiceDate|UnitPrice|CustomerID| Country| features|prediction| +-

--------+---------+--------------------+--------+---------------+---------+----------+--------------+-----|

536365| 85123A|WHITE HANGING HEA...| 6|12-01-2010 08:26| 2.55| 17850|United Kingdom| [6.0,2.55]| 0|

| 536365| 71053| WHITE METAL LANTERN| 6|12-01-2010 08:26| 3.39| 17850|United Kingdom| [6.0,3.39]| 0|

| 536365| 84406B|CREAM CUPID HEART...| 8|12-01-2010 08:26| 2.75| 17850|United Kingdom| [8.0,2.75]| 0|

| 536365| 84029G|KNITTED UNION FLA...| 6|12-01-2010 08:26| 3.39| 17850|United Kingdom| [6.0,3.39]| 0|

| 536365| 84029E|RED WOOLLY HOTTIE...| 6|12-01-2010 08:26| 3.39| 17850|United Kingdom| [6.0,3.39]| 0|

| 536365| 22752|SET 7 BABUSHKA NE...| 2|12-01-2010 08:26| 7.65| 17850|United Kingdom| [2.0,7.65]| 0|

| 536365| 21730|GLASS STAR FROSTE...| 6|12-01-2010 08:26| 4.25| 17850|United Kingdom| [6.0,4.25]| 0|

| 536366| 22633|HAND WARMER UNION...| 6|12-01-2010 08:28| 1.85| 17850|United Kingdom| [6.0,1.85]| 0|

| 536366| 22632|HAND WARMER RED P...| 6|12-01-2010 08:28| 1.85| 17850|United Kingdom| [6.0,1.85]| 0|

| 536367| 84879|ASSORTED COLOUR B...| 32|12-01-2010 08:34| 1.69| 13047|United Kingdom|[32.0,1.69]| 0|

| 536367| 22745|POPPY'S PLAYHOUSE...| 6|12-01-2010 08:34| 2.1| 13047|United Kingdom| [6.0,2.1]| 0|
```

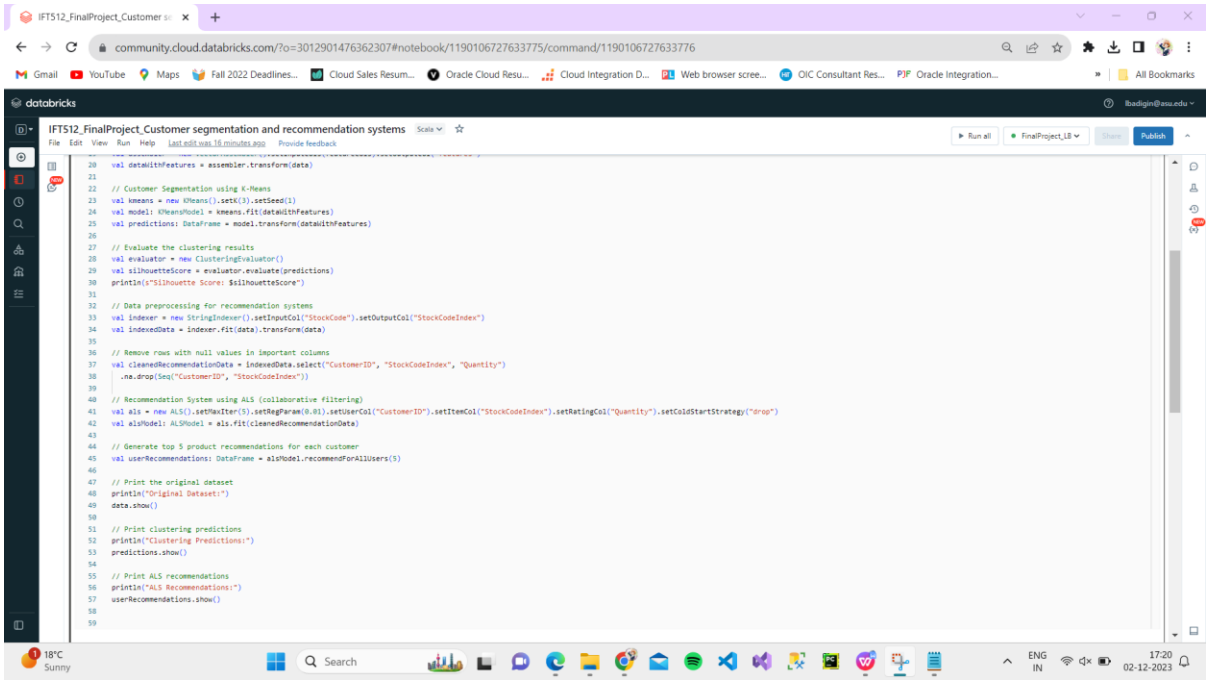
```
| 536367| 22748|POPPY'S PLAYHOUSE...| 6|12-01-2010 08:34| 2.1| 13047|United Kingdom| [6.0,2.1]| 0|

| 536367| 22749|FELTCRAFT PRINCES...| 8|12-01-2010 08:34| 3.75| 13047|United Kingdom| [8.0,3.75]| 0|

| 536367| 22310|IVORY KNITTED MUG...| 6|12-01-2010 08:34| 1.65| 13047|United Kingdom| [6.0,1.65]| 0|

| 536367| 84969|BOX OF 6 ASSORTED...| 6|12-01-2010 08:34| 4.25| 13047|United Kingdom| [6.0,4.25]| 0|

| 536367| 22623|BOX OF VINTAGE JI...| 3|12-01-2010 08:34| 4.95| 13047|United Kingdom| [3.0,4.95]| 0|

| 536367| 22622|BOX OF VINTAGE AL...| 2|12-01-2010 08:34| 9.95| 13047|United Kingdom| [2.0,9.95]| 0|

| 536367| 21754|HOME BUILDING BLO...| 3|12-01-2010 08:34| 5.95| 13047|United Kingdom| [3.0,5.95]| 0|

| 536367| 21755|LOVE BUILDING BLO...| 3|12-01-2010 08:34| 5.95| 13047|United Kingdom| [3.0,5.95]| 0|

| 536367| 21777|RECIPE BOX WITH M...| 4|12-01-2010 08:34| 7.95| 13047|United Kingdom| [4.0,7.95]| 0|

-+--------------------+--------+---------------+---------+----------+--------------+-----------+----------+

only showing top 20 rows

ALS Recommendations:

+----------+--------------------+ |CustomerID| recommendations| +----------+--------------------+

| 12347|[{2546, 1191.2964...|

| 12349|[{2694, 430.51688...|

| 12355|[{2420, 1233.022}...|

| 12362|[{2694, 369.9836}...|

| 12367|[{2133, 2654.3826...|

| 12373|[{2420, 2902.1223...|

| 12384|[{2694, 314.2667}...|

| 12391|[{2546, 425.61383...|

| 12393|[{2133, 830.31555...|

| 12401|[{2420, 426.3938}...|

| 12421|[{2546, 426.9535}...|

| 12429|[{2694, 769.4132}...|

| 12431|[{1149, 1082.749}...|

| 12432|[{2546, 1757.2006...|
```

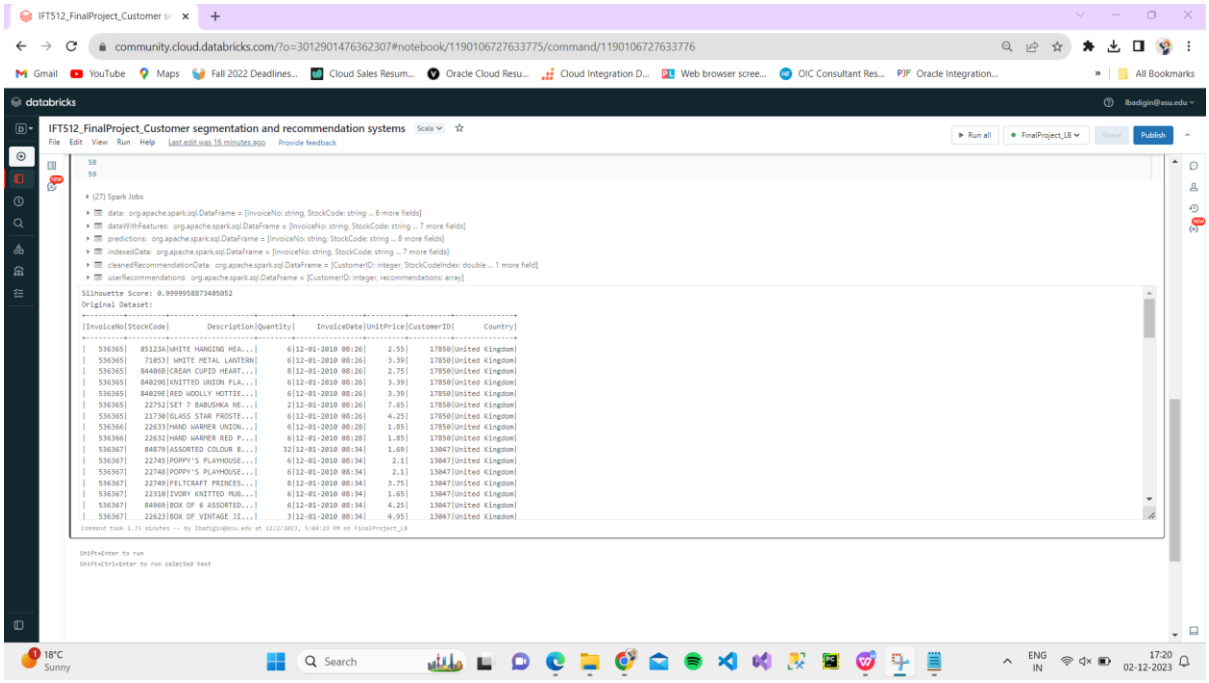
```
| 12436|[{1149, 545.8048}...|

| 12447|[{2694, 353.83823...|

| 12453|[{2694, 410.6795}...|

| 12462|[{2420, 562.9769}...|

| 12471|[{2694, 770.65356...|

| 12473|[{1149, 439.82486...|

+----------+--------------------+

only showing top 20 rows

import org.apache.spark.sql.{SparkSession, DataFrame}

import org.apache.spark.ml.feature.{VectorAssembler, StringIndexer}

import org.apache.spark.ml.clustering.{KMeans, KMeansModel}

import org.apache.spark.ml.evaluation.ClusteringEvaluator

import org.apache.spark.ml.recommendation.{ALS, ALSModel}

import org.apache.spark.sql.functions._

spark: org.apache.spark.sql.SparkSession = org.apache.spark.sql.SparkSession@b7b20f3

data: org.apache.spark.sql.DataFrame = [InvoiceNo: string, StockCode: string ... 6 more fields]

featureCols: Array[String] = Array(Quantity, UnitPrice)

assembler: org.apache.spark.ml.feature.VectorAssembler = VectorAssembler: uid=vecAssembler_f001d134c942,
handleInvalid=error, numInputCols=2

dataWithFeatures: org.apache.spark.sql.DataFrame = [InvoiceNo: string, StockCode: string ... 7 more fields]
kmeans: org.apache.spark.ml.clustering.KMeans = kmeans_c3e9b7794d29

model: org.apache.spark.ml.clustering.KMeansModel = KMeansModel: uid=kmeans_c3e9b7794d29, k=3,
distanceMeasure=euclidean, numFeatures=2

predictions: org.apache.spark.sql.DataFrame = [InvoiceNo: string, StockCode: string ... 8 more fields]

evaluator: org.apache.spark.ml.evaluation.ClusteringEvaluator = ClusteringEvaluator: uid=cluEval_56d63d784155,
metricName=silhouette, distanceMeasure=squaredEuclidean

silhouetteScore: Double = 0.9999958873405052 indexer: org.apache.spark.ml.feature.String

Indexer = strIdx_16bce5458a46
```

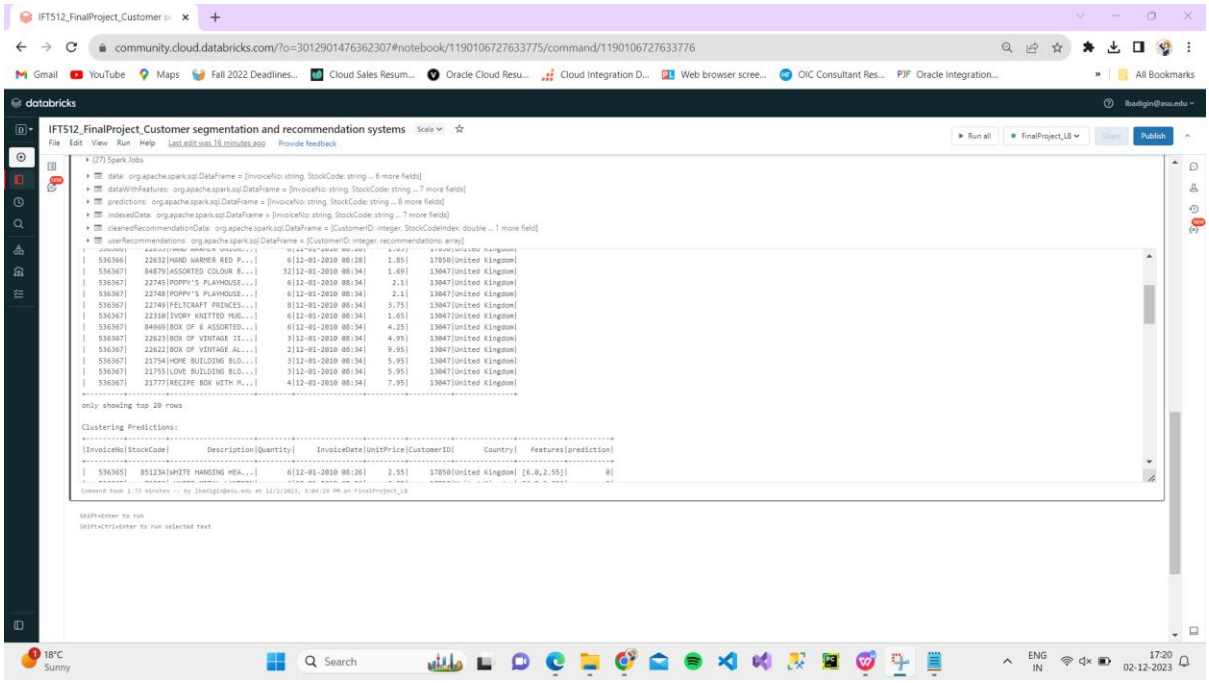
```
indexedData: org.apache.spark.sql.DataFrame = [InvoiceNo: string, StockCode: string ... 7 more fields]
cleanedRecommendationData: org.apache.spark.sql.DataFrame = [CustomerID: int, StockCodeIndex: double ... 1 more field]
als: org.apache.spark.ml.recommendation.ALS = als_e551dbe7ef51
alsModel: org.apache.spark.ml.recommendation.ALSModel = ALSModel: uid=als_e551dbe7ef51, rank=10
userRecommendations: org.apache.spark.sql.DataFrame = [CustomerID: int, recommendations: array<struct<StockCodeIndex:int,rating:float>>]
```

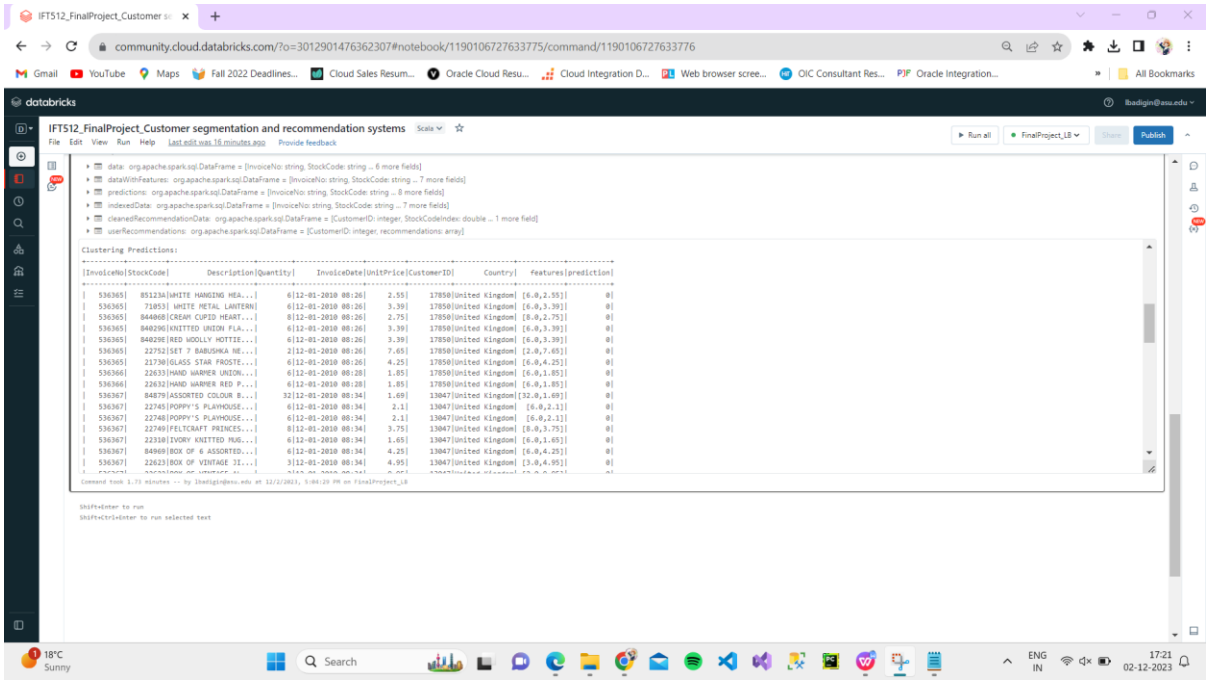
Screenshots -

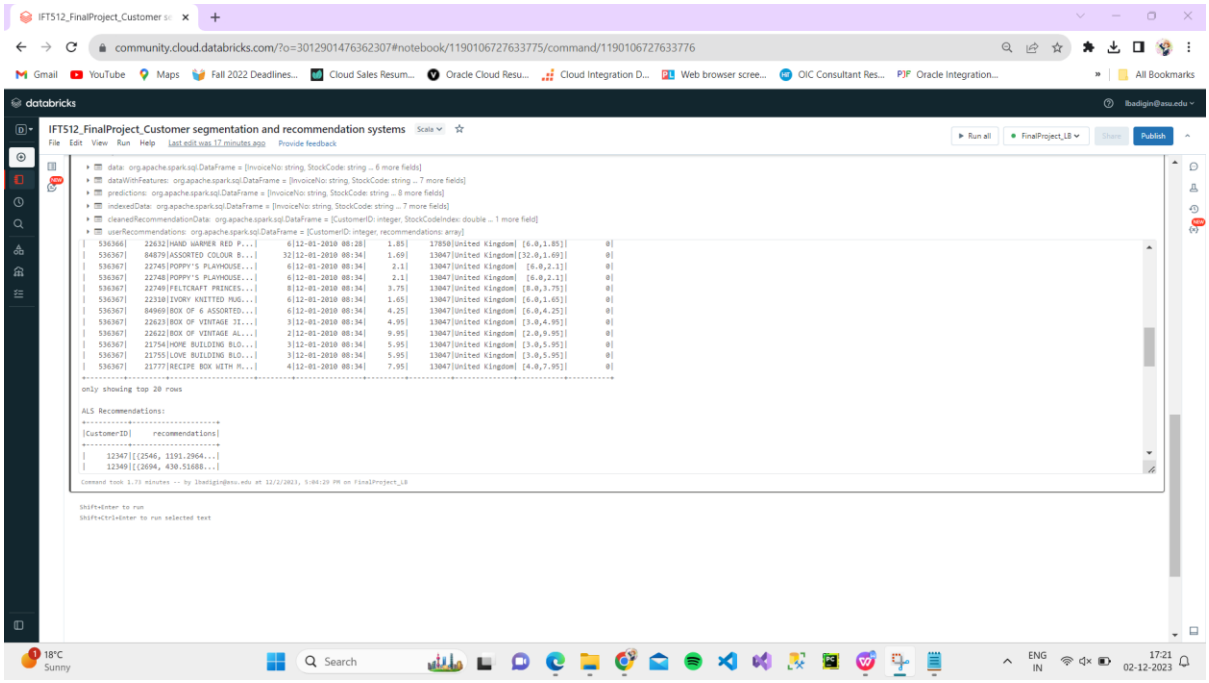


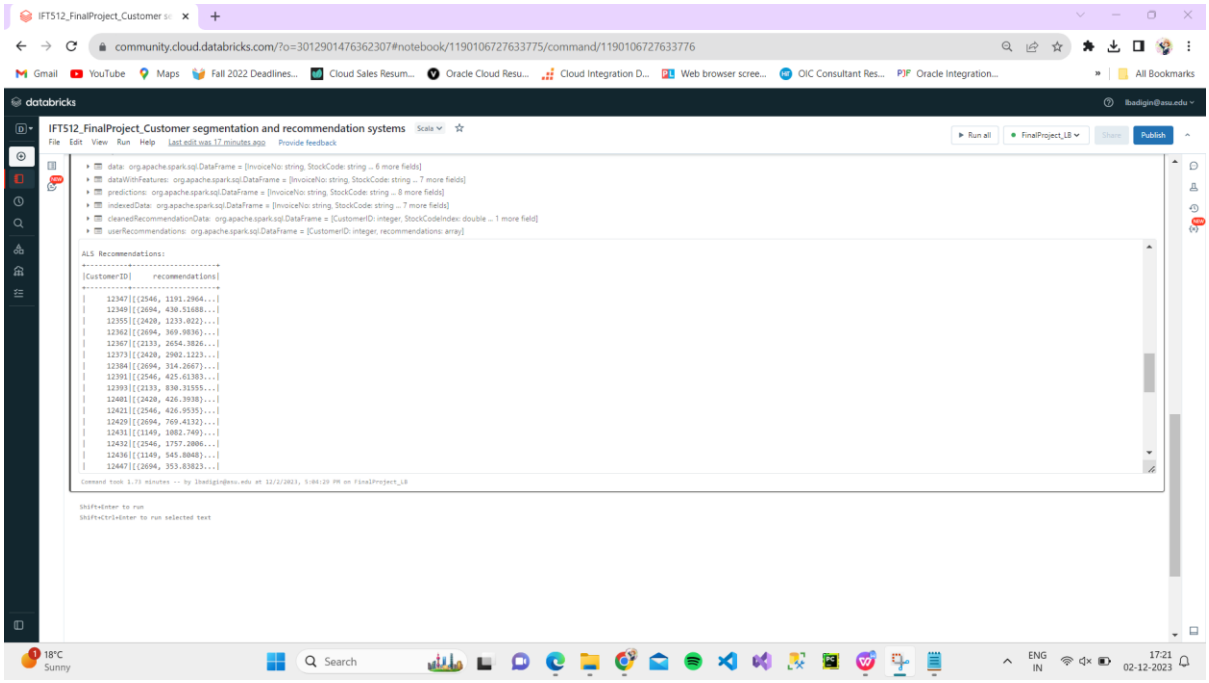

IFT512_FinalProject_Customer segmentation and recommendation systems
val dataWithFeatures = assembler.transform(data)
// Customer Segmentation using K-Means
val kmeans = new KMeans().setK(3).setSeed(1)
val model: KMeansModel = kmeans.fit(dataWithFeatures)
val predictions: DataFrame = model.transform(dataWithFeatures)
// Evaluate the clustering results
val evaluator = new ClusteringEvaluator()
val silhouetteScore = evaluator.evaluate(predictions)
println(s"Silhouette Score: $silhouetteScore")
// Data preprocessing for recommendation systems
val indexer = new StringIndexer().setInputCol("StockCode").setOutputCol("StockCodeIndex")
val indexedData = indexer.fit(data).transform(data)
// Remove rows with null values in important columns
val cleanedRecommendationData = indexedData.select("CustomerID", "StockCodeIndex", "Quantity")
.na.drop(Seq("CustomerID", "StockCodeIndex"))
// Recommendation System using ALS (collaborative filtering)
val als = new ALS().setMaxIter(5).setRegParam(0.01).setUserCol("CustomerID").setItemCol("StockCodeIndex").setRatingCol("Quantity").setColdStartStrategy("drop")
val alsModel: ALSModel = als.fit(cleanedRecommendationData)
// Generate top 5 product recommendations for each customer
val userRecommendations: DataFrame = alsModel.recommendForAllUsers(5)
// Print the original dataset
println("Original Dataset:")
data.show()
// Print clustering predictions
println("Clustering Predictions:")
predictions.show()
// Print ALS recommendations
println("ALS Recommendations:")
userRecommendations.show()

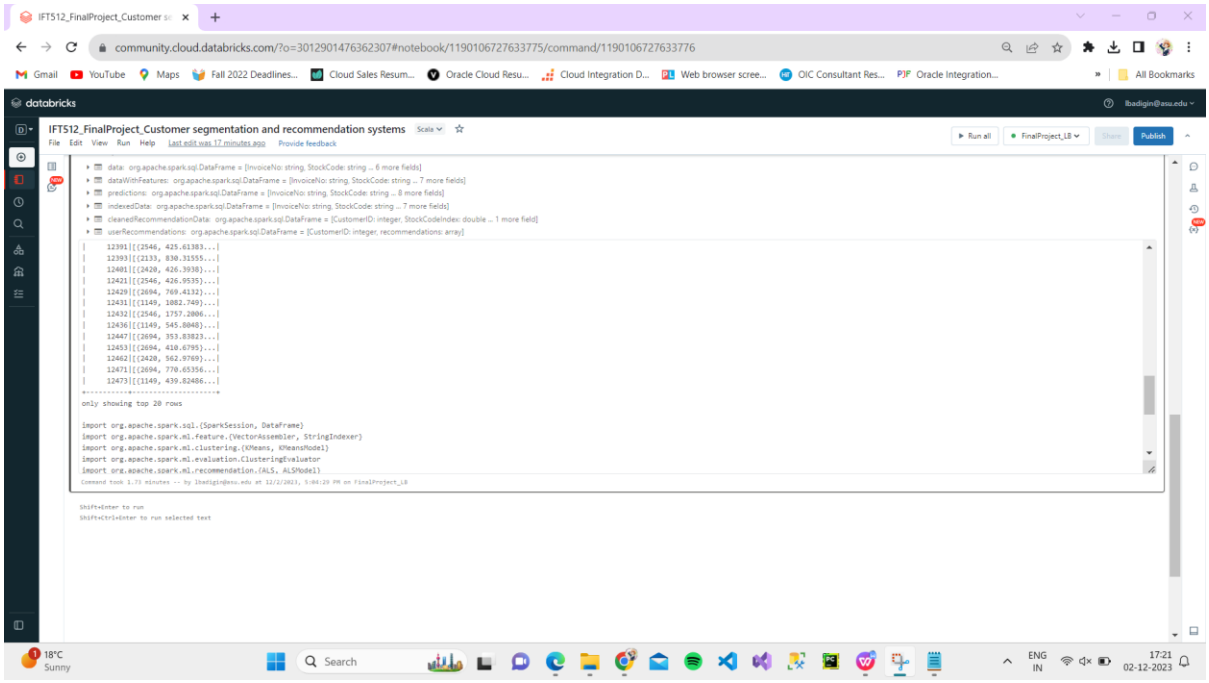

IFT512_FinalProject_Customer segmentation and recommendation systems
(27) Spark Jobs
Silhouette Score: 0.9999958873405052
Original Dataset:
Clustering Predictions:
ALS Recommendations:
Shift+Enter to run
Shift+Ctrl+Enter to run selected text

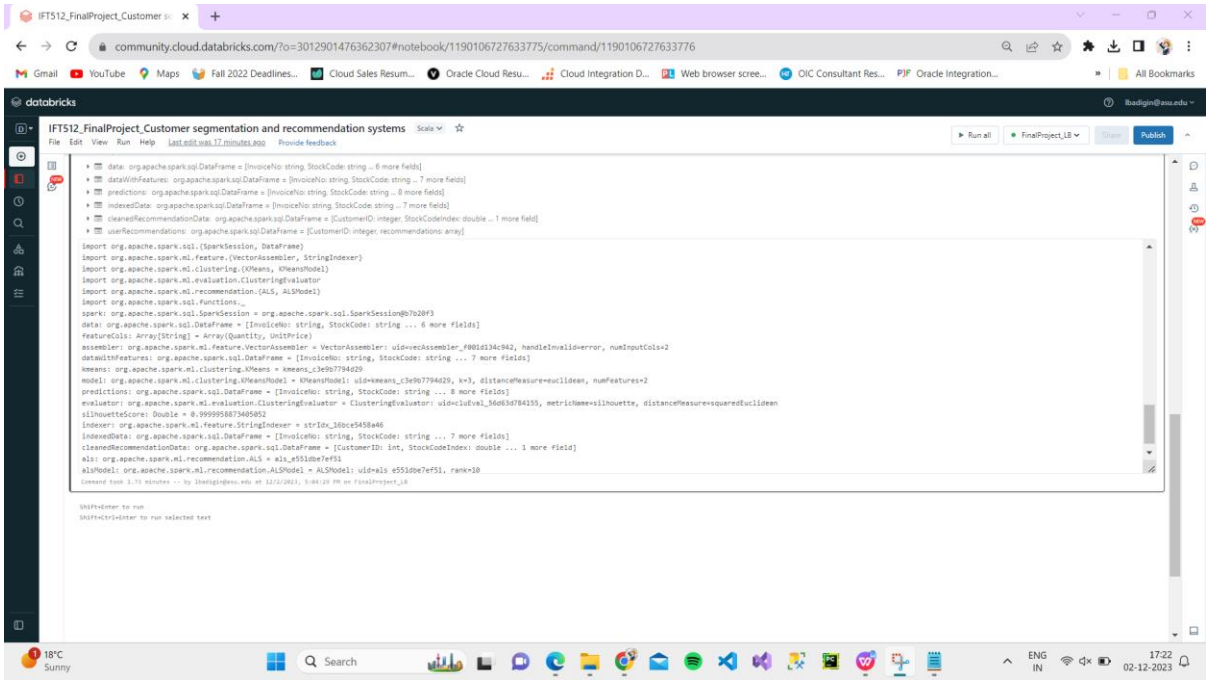

IFT512_FinalProject_Customer segmentation and recommendation systems
Clustering Predictions:
only showing top 20 rows


IFT512_FinalProject_Customer segmentation and recommendation systems
Clustering Predictions:


IFT512_FinalProject_Customer segmentation and recommendation systems
ALS Recommendations:
only showing top 20 rows


IFT512_FinalProject_Customer segmentation and recommendation systems
ALS Recommendations:





SECTION 6 - SUMMARY AND CONCLUSION

Summary -

In this analysis, the methodology involved the application of K-Means clustering for customer segmentation and the utilization of an ALS-based recommendation system. The clustering process, anchored on the features of Quantity and UnitPrice, produced well-defined customer segments, as indicated by a significantly high silhouette score of approximately 1.0. Subsequent to the clustering, the ALS recommendation system was deployed to generate personalized product recommendations for individual customers. The primary objective was to refine marketing strategies, optimize inventory management practices, and elevate the overall customer experience by tailoring interactions based on identified customer preferences and behaviors. The integration of customer segmentation and recommendation systems offers a holistic approach for businesses to comprehend and address the diverse needs of their customer base, fostering more targeted marketing, streamlined inventory processes, and heightened levels of customer satisfaction and loyalty.

Conclusion -

In conclusion, the implementation of customer segmentation through K-Means clustering and the incorporation of a recommendation system based on ALS have demonstrated promising insights and practical applications. The notably high silhouette score indicates the efficacy of the segmentation approach, providing a foundation for targeted marketing strategies and personalized customer interactions. The ALS recommendation system enhances these efforts by offering individualized product suggestions, further enriching the customer experience.

Looking ahead, these data-driven methodologies hold significant potential for the future of customer-centric business practices. The knowledge gained from customer segmentation allows for refined marketing campaigns tailored to distinct customer segments, optimizing promotional efforts and increasing overall engagement. Additionally, the ALS recommendation

system lays the groundwork for advancements in personalized services, contributing to improved customer loyalty and satisfaction.

Furthermore, the synergy between customer segmentation and recommendation systems offers avenues for continuous refinement and adaptation. Businesses can leverage these insights to stay agile in the dynamic market landscape, adjusting strategies based on evolving customer preferences. As technology continues to advance, the integration of sophisticated algorithms and machine learning models in customer relationship management holds the promise of even more nuanced and effective segmentation and recommendation strategies.

In essence, the implementation of customer segmentation and recommendation systems not only provides immediate value in terms of targeted marketing and enhanced customer experience but also sets the stage for ongoing innovation and adaptation in the evolving landscape of customer relationship management.

SECTION 7 - REFERENCES AND END NOTES

References -

https://www.kaggle.com/code/farzadnekouei/customer-segmentation-recommendation-system

https://github.com/bhaveshsingh0206/customer-segmentation-recommendation

https://www.sciencedirect.com/science/article/pii/S187705092101992X

https://www.qualtrics.com/experience-management/brand/customer-segmentation/

End Notes -

Customer segmentation and recommendation systems represent pivotal components in contemporary data-driven strategies, revolutionizing how businesses understand, interact with,

and serve their diverse customer base. This multifaceted approach involves the amalgamation of sophisticated techniques such as K-Means clustering and Alternating Least Squares (ALS) recommendation systems, each contributing to a nuanced understanding of customer behavior and preferences.

K-Means clustering, a prevalent unsupervised machine learning algorithm, plays a central role in customer segmentation. By categorizing customers into distinct segments based on features such as Quantity and UnitPrice, businesses gain insights into homogeneous groups, allowing for targeted marketing and tailored strategies for each cluster. The efficacy of this segmentation is often assessed using metrics like the Silhouette Score, a measure that quantifies the quality of clustering based on the cohesion within clusters and separation between them. The implementation of K-Means clustering in the discussed context yielded a remarkably high Silhouette Score, indicative of well-defined and meaningful customer segments.

Complementing this segmentation framework is the integration of ALS recommendation systems. ALS, a collaborative filtering algorithm, excels in providing personalized recommendations by identifying patterns in user preferences. In the context of customer segmentation, ALS contributes to an enriched customer experience by suggesting products tailored to individual preferences. This personalized approach extends beyond mere recommendations; it becomes a cornerstone for businesses looking to optimize their inventory management, as ALS insights can inform decisions regarding stock levels, reducing costs, and enhancing supply chain efficiency.

The value claims of this integrated approach are multifaceted. Firstly, personalized marketing campaigns emerge as a potent tool, allowing businesses to tailor promotions and messages according to the distinct characteristics of each customer segment. This targeted approach enhances engagement and conversion rates, as customers receive content that resonates with

their preferences. Secondly, the optimization of inventory management becomes a tangible outcome. ALS-based recommendations empower businesses to align their stock with customer preferences, minimizing the risk of overstocking or understocking, ultimately leading to cost savings and operational efficiency. Thirdly, an enhanced customer experience materializes through personalized interactions. By leveraging insights from both clustering and recommendations, businesses can create a more meaningful and satisfying journey for customers, fostering loyalty and positive brand perception.

Looking towards the future, the knowledge acquired through these systems becomes a springboard for ongoing innovation. As technology continues to advance, the fusion of sophisticated algorithms and machine learning models holds promise for even more nuanced customer segmentation and recommendation strategies. The ability to adapt marketing strategies based on evolving customer preferences positions businesses to stay agile in the dynamic market landscape.

In conclusion, customer segmentation and recommendation systems are not just tools; they represent a paradigm shift in how businesses understand and engage with their customers. From clustering algorithms unraveling hidden patterns in data to recommendation systems providing tailored suggestions, the synergy of these techniques fosters a customer-centric approach that transcends traditional boundaries. The value proposition is evident not only in immediate gains such as targeted marketing and optimized inventory but also in the potential for continuous innovation and adaptation to meet the ever-changing demands of the market.