Big Data Analytics

BYGB-7990-001

Yilu Zhou, Ph.D

**e-Commerce user profile &**

**product recommendation system**

**Group 4**

**Annuo Qian, Jing Zhu, Haojia Li, Ting Ding**

**1. Executive Summary**

In this project, we want to deal with e-Commerce user behavior data to understand customers better and bring more profits to business. We did three important analyses to e-Commerce users and products: Data visualization, User attribute generation and user-product rating and recommendation system.

We collected data about online users’ action from Kaggle. Our dataset contains behavior data for 5 months (Oct 2019 – Feb 2020) from a medium cosmetics online store, and we created 6 new columns/variables that are calculated from the original data, which are: event hour, session count, purchases per session, purchase times per session, total purchase, cart per view to do the analysis.

In terms of data visualization we used python to do descriptive analysis about our data and presented the data attributes with four charts. According to the charts we found that the sessions has an obviously decrease in Dec. 2019 due to the Christmas festival and so on. We also noticed that purchase only occupies 6.2% in all the events, indicating that most users browsing the website do not purchase anything in the end. Also, the variation of events follows the same trend with sessions. However, the amount of view is only 13.3% of that of view, which indicates the view-to-purchase rate is very low. On the other hand, most events concentrate in the price range which is less than 10 dollars, thus the low-price product is our main market.

Then we used python and SPSS to generate user attribute, including user profile and user clustering. There were four clusters in our model. The largest cluster occupied 98.2% of our data, while the smallest cluster held less than 0.1% of the dataset. Cluster-1, occupying the most significant percentage of users, was a cluster representing “light users.” They loved browsing instead of taking actions. Only eight purchasing behaviors would happen in 100 sessions. Each user in cluster-1 contributed an average of 2.01 dollars to the website’s sales. However, considering the vast amounts of users, their purchasing power is enormous and indispensable. Cluster-4 users visited nearly surfed our shopping website every week. Every time they visited, they would like to purchase more than five products and contribute 29 dollars on average. The total sales they made were huge. They were rich, openhanded, and consumeristic. Users in cluster-2 also spent a lot of money. However, they seldom visited our shopping website. They had a concrete target. Once they found the target product, they would buy lots of them. They would not be the best targets of our recommendation system, so we may consider give up making recommendations to them to reduce costs. The most typical characteristic of users in cluster-3 views a lot while purchase few.

Finally, we created a user-product rating algorithm and used ALS algorithm to get top ten products for each user. Then we provided corresponding recommendations according to their preference. For cluster-1 users, we may make the shopping website more appealing and provide more promotions and discount to them at their most concentrate event hour. For cluster-4 users, we decided to put more advertisements about their preferred products and related categories at their concentrate event hour and enhance the advertisement frequency, since they are our important targets. According to cluster-2 users’ requests and preference, we can recommend products to them directly like sending email to satisfy their demands instead of investing a lot in the advertisement. For cluster-3 users, we can just recommend their preferred products rather than related categories to reduce costs.

**2. Business Problem**

**Background Introduction**

Nowadays online shops generate tons of user session data on a daily basis. An HTTP session is a sequence of network request-response transactions. Marketers can use this data to understand how customers are using the website, where they come from, and what users are doing before and after an important action like a purchase. How to deal with those user behavior data to understand your customers better is a big issue for e-commerce.

**Problem Statement**

Our dataset is eCommerce events history in Cosmetics Shop, it contains more than 20 million event records with various users and products. Through observing the events and sessions data, our group find that some interesting business stories such as the purchase events is only a very low proportion in all the events and most events concentrate in the price range of a fewer dollars.

Due to these issues, our group want to know:

* How a session or an event contributes to the website’s business. For example, the website could make how much income per session in average, or what is the specific composition of different event types in one session?
* The behaviors of users. Which price range is the user most interested in? When does he or she visit the website most frequently in a day? How many times of cart or purchase action the user has when he or she click through the website?
* Which product is a specific user’s favorite? The different event types of a same user for products indicate the assessment from the user and may lead to the user’s preferences.

**Project Expectation**

To answer the questions above, we decide to

* Use data visualization to take a more convenient and direct look at this dataset, and descriptive analytics could help us make a comprehensive understanding of events, sessions and even the whole activities of the website.
* Generate user attributes and profile to collect useful information of users, and the information could provide us with a database to build a machine learning model and divide users into several different clusters according to their behaviors and preferences.
* Quantify the assessment of products by a user for our group to analyze and design a recommendation system based on the quantitative results. This system could recommend the appropriate products to target users.

In one word, we will understand the session records, the product sales and user information better to optimize the operation strategies of the eCommerce website.

**3. Dataset Description**

**Dataset Summary**

Our dataset contains behavior data for 5 months (Oct 2019 – Feb 2020) from a medium cosmetics online store. The dataset has 20,692,840 rows with 9 features and the size of file is 2.0GB.

Each row in this dataset represents an event. All events are related to products and users. Each event is like many-to-many relation between products and users.

**Features Introduction**

**event\_time** Time when event happened at (in UTC)

**event\_type** Events can be:

* view - a user viewed a product
* cart - a user added a product to shopping cart
* removefromcart - a user removed a product from shopping cart
* purchase - a user purchased a product

**product\_id**  ID of a product

**category\_id** Product's category ID

**category\_code** Product's category taxonomy (code name) if it was possible to make it.

Usually present for meaningful categories and skipped for different

kinds of accessories.

**brand** Downcased string of brand name. Can be missed**.**

**price** Float price of a product. Present.

**user\_id** Permanent user ID

**user\_session** Temporary user's session ID. Same for each user's session. Is changed

every time user come back to online store from a long pause.

**Dataset Discussion**

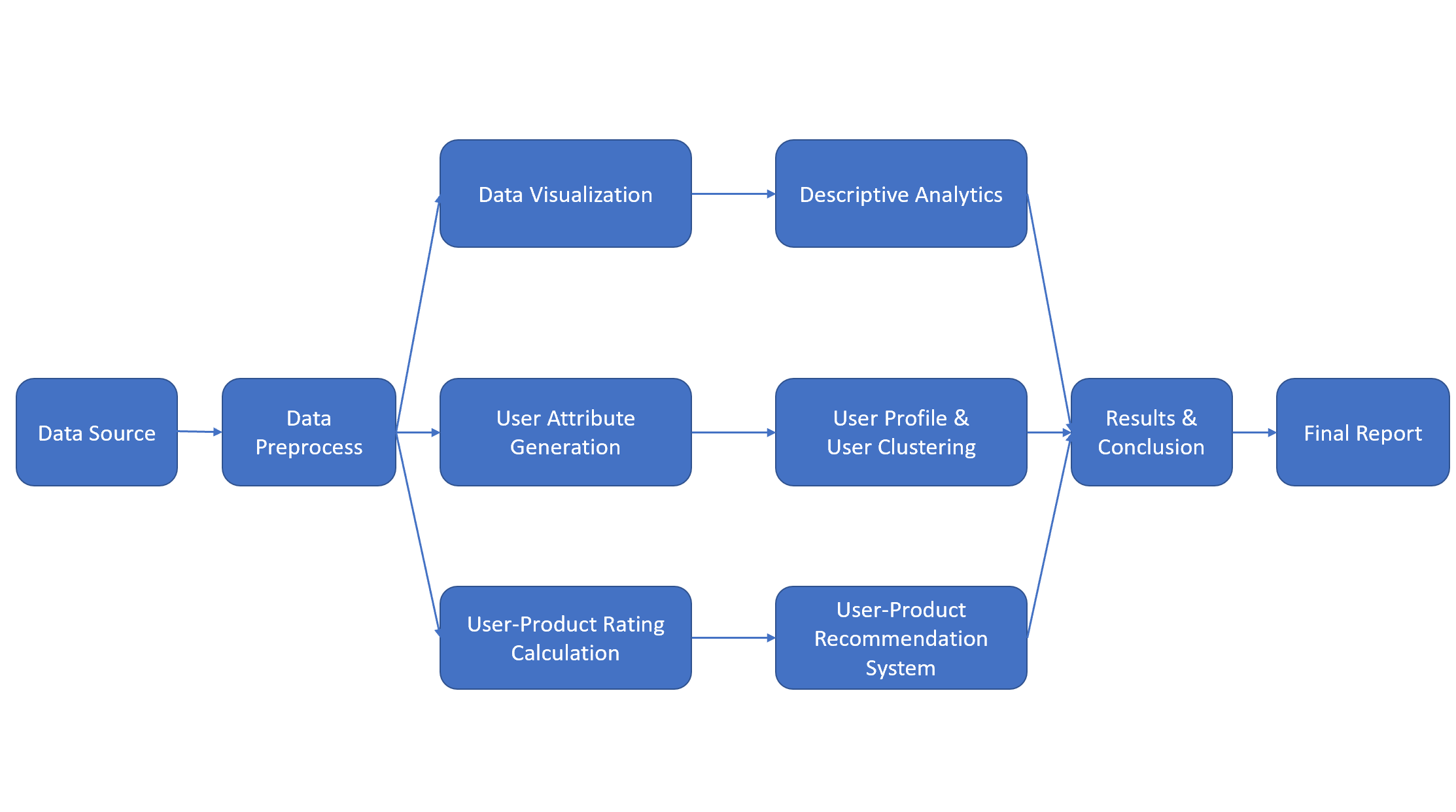
We collect information and for the product, there are 54571 products with 525 categories and 274 brands, many brands information are empty, in this dataset, so each category has nearly 104 products in average. On the other hand, the price varies from -79.37 to 327.78 and the negative value is because promotions

For the events and sessions, as we mentioned before, there are 20,692,840 events with 4,535,942 sessions, which means each session has 4.5 events in average. In those events, we find they include 9,657,821 views, 5,768,333 carts, 3,979,679 removefromcarts and only 1,287,007 purchases. The purchase proportion is less than one tenth!

For the user information, we could find 1,639,358 user ids in the dataset, which means each user has nearly average 2.7 sessions in these five months. And if we consider the events information calculated above, we could notice that one user only has less than 0.7 purchase as expectation. We will pay continuous attention to this issue and try to solve it later.

**4. System Design**

**Workflow Chart**



*Figure 4.1 Workflow charts*

From Figure 4.1, our group will:

1. Collect feature information and analyze our dataset, preprocess data, and refine the dataset according to our analytics results to make preparations for our whole project.
2. Visualize our data to explore the relationships among the different features. We want to use descriptive analytics to introduce several types of charts, help us understand the sessions and events in this e-commerce website, find the business stories and problems, and try to design some solutions.
3. Generate user attributes which are defined by our group from the original dataset to create unique user profile. Through this user profile, we could understand the users’ behaviors and preferences, apply K-means machine learning algorithm to it and split the users to different clusters. By analyzing these clusters, we may know different demands of users and make specific strategies.
4. To meet the need that we should know which user would like which product, we decide to create a User-Product recommendation system. We design a formula which is involved by the different events for a specific product by a unique user to calculate the rating given by the user to the product. Our recommendation system is based on the rating results and will tell us a given user would like which product.
5. Finally, we will have charts describing the dataset and find business issues, user profile about user behaviors and preferences, clustering model with different groups of users, a table including all the ratings a user give to a product and a recommendation system based on the ratings. We could conclude and complete our final report.

**5. Data Preprocessing**

**Feature engineering**

As we mentioned in the data description, we have row-based session data of user behaviors like view product, add a product to cart, remove from cart or purchase. In the feature engineering part, we want to generate more meaningful features to know our user habits better.

Therefore, we decide to generate a dataset that shows user behavior features. Each row in this data set is a unique customer. We add 6 columns/variables that are calculated from the original data, which are: event hour, session count, purchases per session, purchase times per session, total purchase, cart per view. We use python and pandas package in this step.

1. **event hour**: This information is extracted from the original attribute ‘event time’. We convert the string type to datetime in pandas and extract the hour that event happens because we only care about the time period when the customer is used to shop most. Then we grouped the data by user id and event hour to count the most frequent hour that each customer shops online. At last, we order the dataset in descending order by count, group by user id and remained the hour that has the highest count for each user.
2. **Session count**: A user can have many session records and a session may contain multiple events such as view or add to cart. So to calculate how many sessions a user has, we first group by user id and session number and then aggregate unique sessions that a user has. This process takes a lot of computing power and CPU memory.
3. **total purchase**: This attribute is calculated to show the purchasing power of each customer, that is in the five-month-period data we have, the total purchasing price of each user. To calculate, we first use logical operation to extract rows of records whose event type is ‘purchase’. Then we group by user id to sum up the total purchase each customer did.
4. **purchases per session**: This is another attribute that shows the user's purchasing power. This is how much did a user average pay per session. This column is calculated by simply dividing the total purchase of a user by the session count.
5. **purchase times per session**: This is an attribute that represents a user’s shopping habits. It measures on average how many purchase behaviors a user did in a session that may include behaviors like view and cart. This is an important matric that a marketer may look at. This is calculated by selecting event type as ‘purchase’, grouping by user id and count purchase times.
6. **cart per view**: This is another attribute that shows a user’s shopping habits. It represents how many cart behaviors a customer would do in terms of one view. This is calculated by using logical operation to extract rows of records whose event type is ‘cart’ and records are ‘view’ separately and divide them in the end.

**Handling NaN and missing value**

The calculated fields ‘total purchase’ and ‘purchase times’ have many NaN values in the dataset due to some customers not having any purchase behavior, we replace those fields with zeros.

**6. Algorithm Development**

**User Profile & Clustering**

We put together session information and generate user profiles in this step because we want to get to know our clients in multiple dimensions. A user profile commonly includes details of customers, such as their: location, purchasing habits, affordable prices, preferable brands, etc. We preprocessed our dataset and added the following variables to represent user habits: *event hour, session count, purchases per session, purchase times per session, total purchase, cart per view*.

We extract *event hour* information from event time because we want to know when our customers usually go to our website most. We calculate the *session count* of each customer wondering how often a customer visits our online shop and if he returns very frequently. We aggregate the *total purchase* of each customer to see how much a user spent on our products. The *purchases per session* also indicates the purchasing power of a customer, we want to target customers that purchase in different price ranges with different advertising strategies. The *purchase times per sessio*n represent the purchase habit of a customer, some customers may view and add and remove a product several times before he purchases while others merely view a lot. *Cart per view* also represents a customer’s online shopping habit.

We get user habit information in terms of shopping time and purchasing habit, we can then cluster them into different types and decide customized advertisement for them.

After data engineering, newly generated features were then put into the clustering process. The input variables included session count, total purchase, purchases per session, purchase times per session, and cart per view. In the SPSS Modeler, these variables set as input in the “type” node, and then they were put into the K-mean clustering node to build a model. The stopping rule here was iterations times reach 20. Compared with models with other K values, the model whose K equaled to four performed the best.

**User-Product Rating**

There are four kinds of events in the dataset: view, cart, remove from cart and purchase. There is no rating in the dataset. However, to apply the ALS recommendation system, users, products, and ratings are necessary. Therefore, we decided to generate ratings through these variables.

Initially, we simply gave each event a score according to its contribution to increasing sales. We gave “view” 1 point, “cart” 2 points, “remove from cart” -2 points, and “purchase” 5 points. We then grouped data by user ID and product ID and summed up the scores of each user against each product as the rating. However, a lot of negative ratings showed up, which was out of our expectations. Because our design of -2 points for “remove from cart” was to counteract the 2 points that “cart” added, and we did not expect negative scores, let alone a lot of negative numbers. Our dataset covered the actions that users made from October 2019 to February 2020. Negative ratings may relate to the users who added products to the cart before October and dropped them after October.

To fix it, we designed a more comprehensive program to calculate the rating. We gave a score to each event as above. We added a column called “count” to count the event times of every kind of event for each user against products. And we also upgraded our logical design. If the “remove from cart” and “purchase” times equaled to or less than the “cart” times, there was still enough products in the cart, so that we could sum up the scores directly. If the “remove from cart” and “purchase” times were greater than the “cart” times, we needed to make another logical judgment. Under this situation, if the “remove” times were greater than “cart” times, considering the “cart” events before October, all the “remove from cart” should be offset. Hence, we only summed up the scores of “purchase” and “view.” If the “remove” times were less than “cart” times, but the “remove” and “purchase” times were greater than the “cart” times, we need to make up the “cart” times before October. So we summed up all the scores and added 2 points for the difference between the sum of times of “remove” & ”purchase,” and “cart” times. We implemented this program as our final rating calculating method.

**User-Product recommendation system**

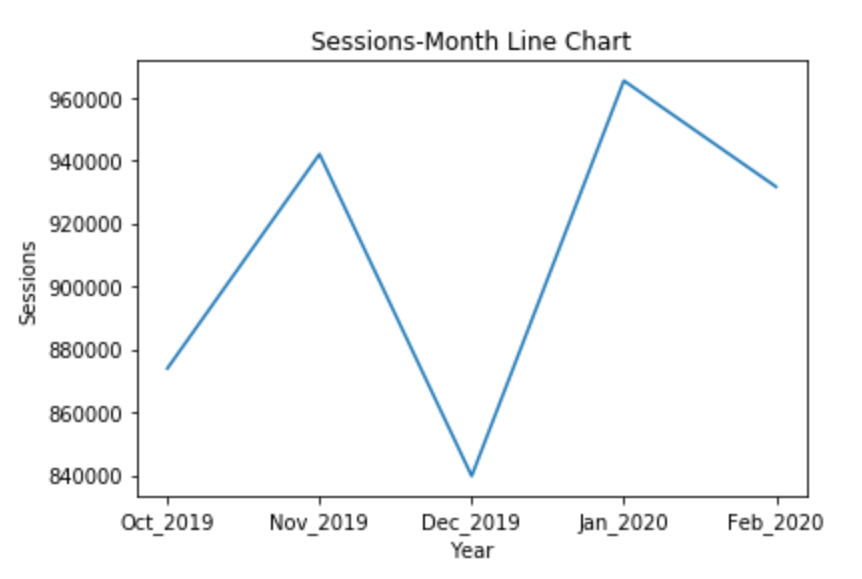
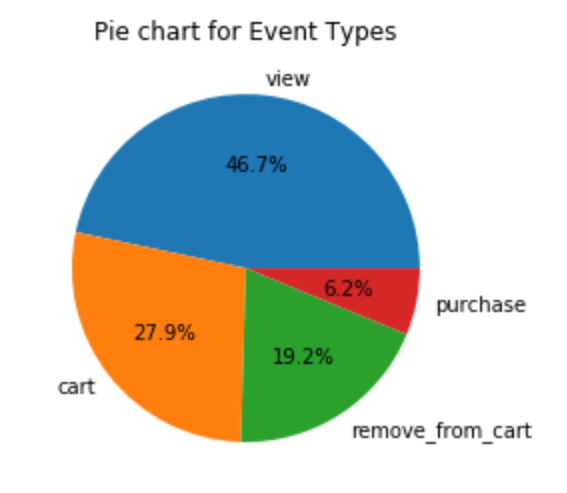
Given its popularity or rating there are three main techniques used for providing recommendations online-Collaborative filtering, Content-based and Hybrid technique. Here we will be making use of Alternating least square matrix factorization method, a collaborative filtering algorithm. The idea is basically to take a large (or potentially huge) matrix and factor it into some smaller representation of the original matrix through alternating least squares. ALS comes inbuilt in Spark.

We are going to build a recommendation system in python using spark and Jupyter Notebook. Our recommendation system is based on the rating results, and we do not need all the columns present in the dataframe. Only assign which is user ID, product ID and overall (rating given by users to each product) is required. We will import important modules we need to create the recommendation system, including Spark Context, SQLContext, RegressionEvaluator, ALS, SparkSession, etc. We will hold out 80% for training and leave 20% for testing. We built the recommendation model using ALS on the training data and we set cold start strategy to 'drop' to ensure we do not get NaN evaluation metrics. The ALS model was evaluated by computing the RMSE on the test data, the standard deviation of the residuals (prediction errors), which are a measure of how far from the regression line data points are and it tells you how concentrated the data is around the line of best fit. Finally, we can get top 10 products recommendations for each user and export the result to the csv file.

**7. Results and Evaluation**

**Data Visualization**

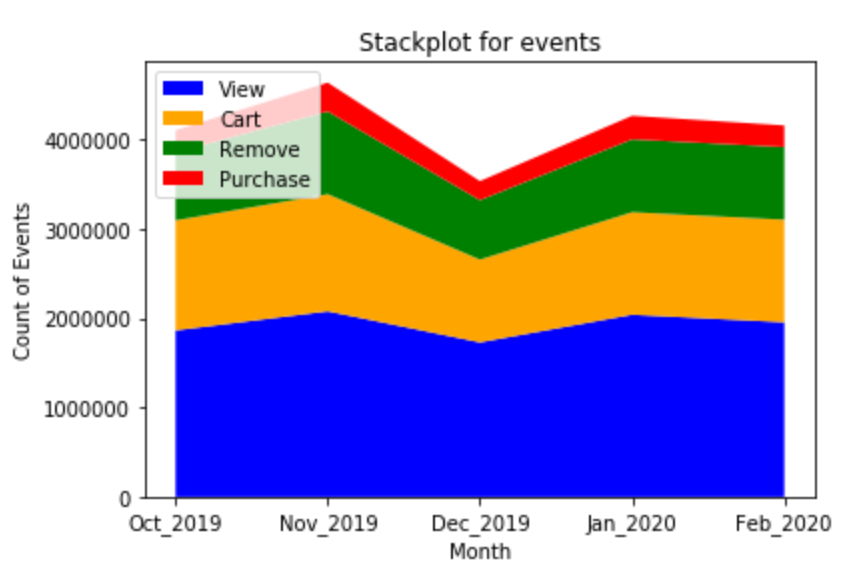
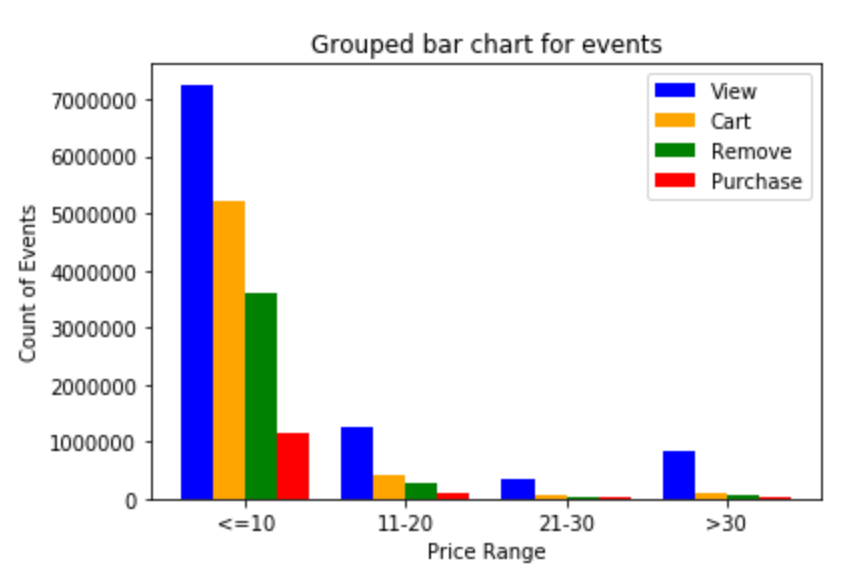
We count sessions among different months and generate proportions of event types in session.

*Figure 7.1 Figure 7.2*

From the two figures above, we could find that the sessions have an obviously decrease in Dec. 2019, we assume that people may prefer shopping offline when the Christmas is coming and affect their online shopping activities. We also notice that purchase only occupies 6.2% in all the events, it indicates most users browsing the website do not purchase anything in the end. This is not good news.

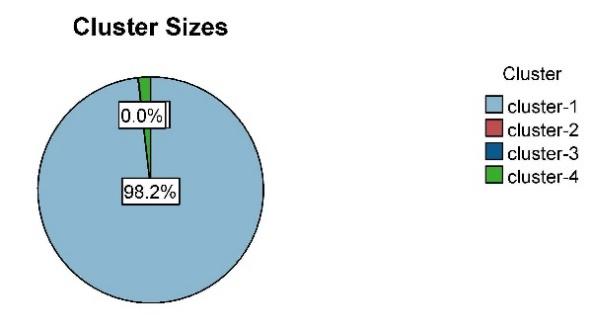
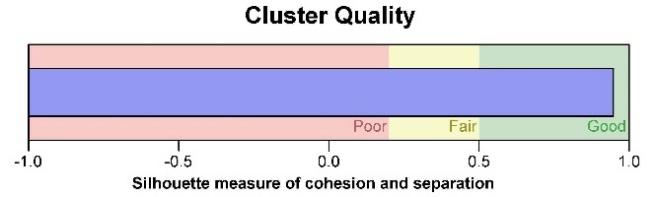
To make further understanding, we collect the events among the five months and observe the distribution of event types in different price ranges. The charts show below:

*Figure 7.3 Figure 7.4*

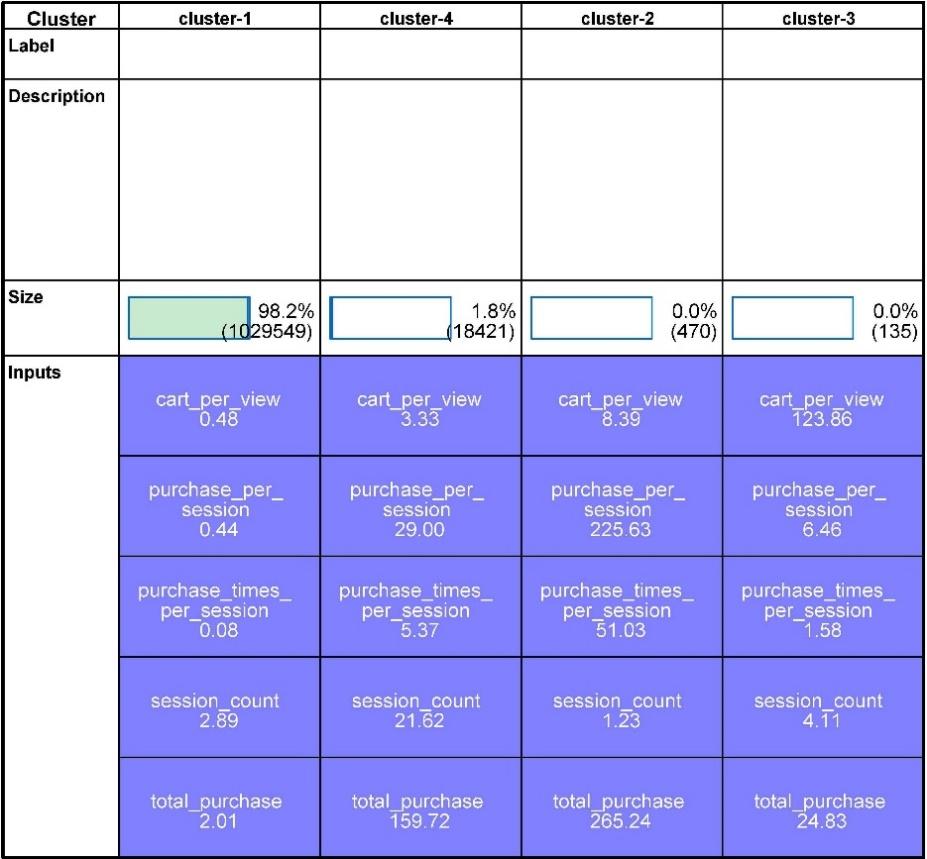
From the two figures above, we find the variation of events follows the same trend with sessions, it fits our common sense. However, the amount of purchase is only 13.3% of that of view, which indicates the view-to-purchase rate is very low. On the other hand, most events concentrate in the price range which is less than 10 dollars, thus the low price product is our main market according to this grouped bar chart.

**User Clustering**



*Figure 7.5: Cluster quality and cluster size*

As mentioned above, after comparing models with different K values, K equal to four was the most rational one considering the cluster quality and cluster’s features. Figure 7.5 is the cluster quality and cluster sizes of our K-mean model. Figure 7.5 shows that the average silhouette was 0.9, which was good. There were four clusters in our model as we set K equal to four. The largest cluster had 1029549 records, occupied 98.2% of our data, while the smallest cluster had 135 records, held less than 0.1% of the dataset. The ratio of sizes that the largest cluster to the smallest cluster was 7626.29.



*Figure 7.6: Cluster description*

Figure 7.6 is a table depicting the features in each cluster. It is worth noticing that the cluster’s size ordered the table. So, its order was cluster-1, 4, 2, and 3 from left to right. The background color represented the predictor importance of those inputs. Figure 7.6 shows that all their colors are deep purple, which means the predictor importance of five variables were all 1. The number under the input’s name was the average number of that variable in that cluster.

Cluster-1, occupying the most significant percentage of users, was a cluster representing “light users.” Figure 7.6 shows that cluster-1 had the lowest mean score in each variable except for “session\_count.” A session is a series of events users made in a log-in operation. According to the “session\_count” cell, cluster-1’s users visited our shopping website less than three times on average in five months. They loved browsing instead of taking actions, less than half of the products they looked would be put into shopping cart. Only eight purchasing behaviors would happen in 100 sessions, as shown in the “purchase\_times\_per\_session” cell. Each user in cluster-1 contributed an average of 2.01 dollars to the website’s sales. However, considering the vast amounts of users, their purchasing power is enormous and indispensable. The small improvements in their consumption habits would be a considerable change to the shopping website. To arouse cluster-1 users’ interest in viewing, adding to cart, and even buying is the priority of our recommendation system.

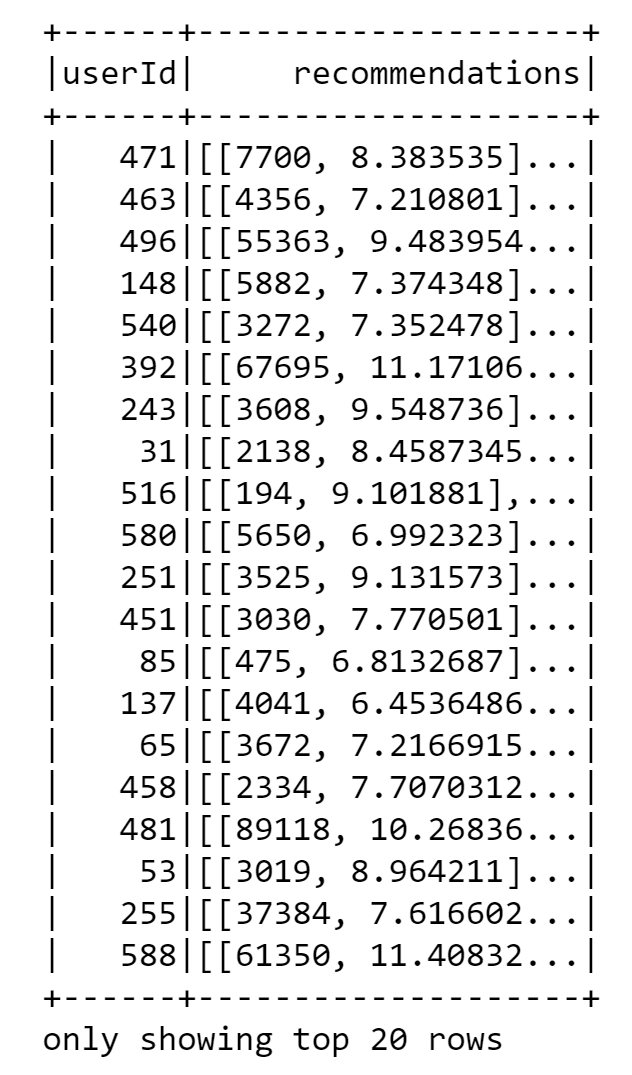
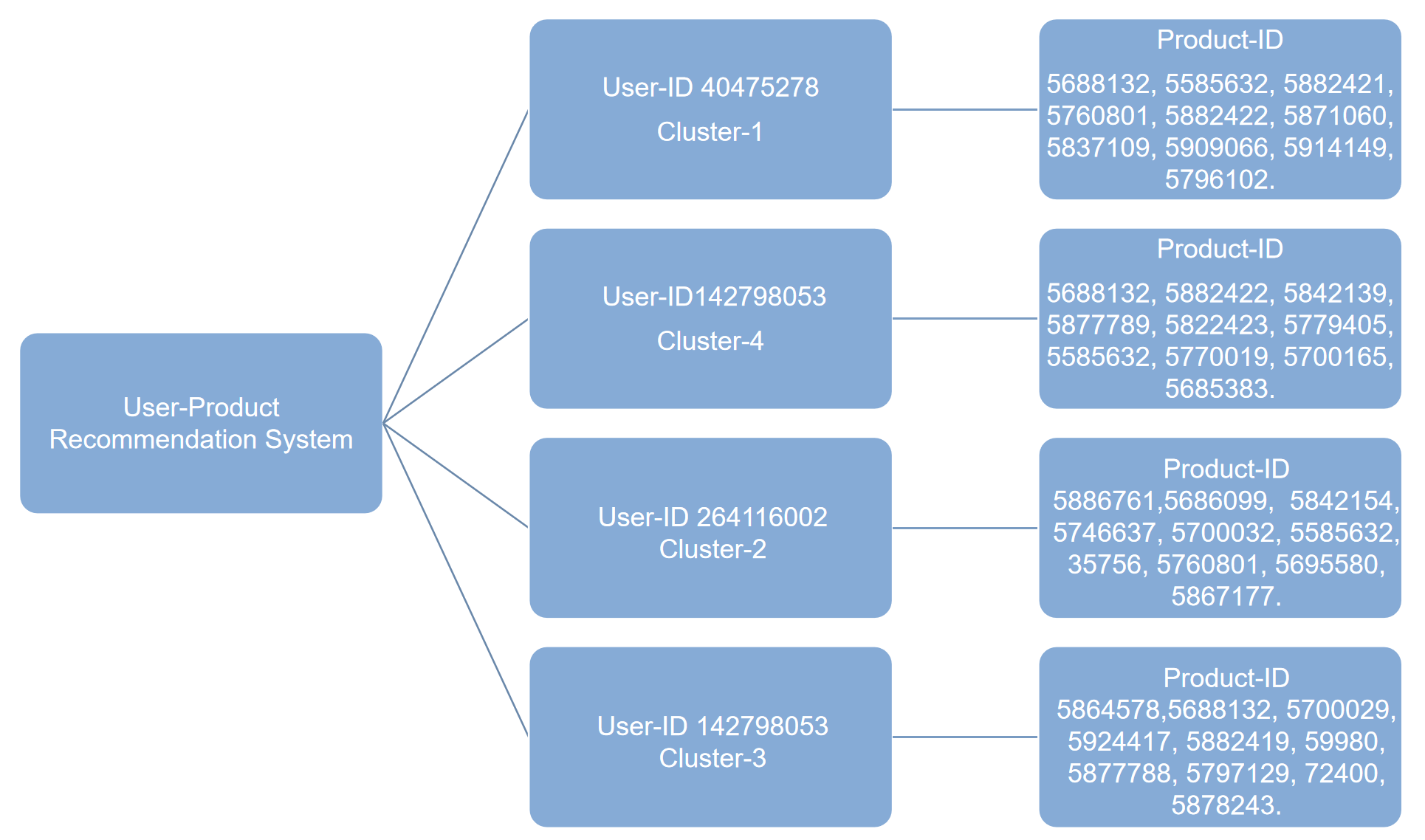
Cluster-4 on the second column was a cluster for active customers. They visited our shopping website 21.62 times on average in the past five months. In other words, they surfed our shopping website every week. Every time they visited, they would like to purchase more than five products and contribute 29 dollars on average. The total sales they made were even greater than that of cluster-1. They were rich, openhanded, and consumeristic. Investing money on advertising them would be the most effective and rewarding. Keep pushing new and grand product to cluster-2’s users would be a proper strategy.

Similar to cluster-4, users in cluster-2 also spent a lot of money. However, they seldom visited our shopping website. And they seldom browsing around, which made the “cart\_per\_view” very low. They had a concrete target. Once they found the target product, they would buy lots of them. As a result, cluster-2’s customers’ average of “pruchase\_per\_session” and “total\_purchase” were very close, and the “purchase\_times\_per\_session” was very high. They would not be the best targets of our recommendation system, for which they seldom viewed and goal-oriented.

The most typical characteristic of users in cluster-3 on the last column was the extremely high average number of “cart\_per\_view”. Even though, only few of these products were purchased at the end, referring “purchase\_times\_per\_session”. How to convert the carted products to the the sale would be a task. In consideration of the number of users in this cluster and the cost of developing a method to solve this task, building a recommendation system generally and let go the minority would be a better choice.

**User-Product recommendation System**

By using ALS algorithm with jupyter and spark, we get the result of top 10 products for each user. Below is the sample result.

*Figure 7.7: top 10 products for each user Figure 7.8: User-Product recommendation system*

A client whose user ID is 40475278 belongs to cluster1 “light users”, and his top 10 product id are 5688132, 5585632, 5882421, 5760801, 5882422, 5871060, 5837109, 5909066, 5914149, 5796102. According to the cluster 1 user attribute, cluster 1 users loved browsing instead of taking actions, less than half of the products they looked would be put into shopping cart, and they only contributes an average of 2.01 dollars to the website’s sales. However, he is the potential customer considering the vast amounts of this kind of users. Their purchasing power is enormous and indispensable. Those users are the priority of our recommendation system. For them, we may make the shopping website more appealing and provide more promotions and discount to them at their most concentrate event hour.

A user whose user ID is 142798053 comes from cluster-4, and she is a representative of active customers, who visit our shopping website 21.62 times on average in the past five months, which means at least once a week. She can bring many profits to business, since users in this group usually purchase more than five products and contribute 29 dollars on average. We guess that this user is rich and consumeristic, and she is our VIP customer, so we are supposed to invest money on advertising her. For example, putting more advertisements about her preferred products, which are 5688132, 5882422, 5842139, 5877789, 5822423, 5779405, 5585632, 5770019, 5700165, 5685383 as well as related categories at her usual event hour and enhance the advertisement frequency.

Since users in cluster-2 are goal-oriented and they have a concrete target. They are not important targets of our recommendation system. User 264116002 is a representative of this group, and his top 10 products id are 5886761, 5686099, 5842154, 5746637, 5700032, 5585632, 35756, 5760801, 5695580, 5867177. According to his request and preference, we can recommend those products to them directly like sending email to satisfy their demands instead of investing a lot in the advertisement to reduce the cost and bring more profits.

A customer whose id is 142798053 belongs to cluster-3, a group that has extremely high average number of “cart\_per\_view” while end up with few purchases. In consideration of the number of users in this cluster is relatively small and the cost of developing a method to solve this task. We can just recommend their preferred products which are 5864578, 5688132, 5700029, 5924417, 5882419, 59980, 5877788, 5797129, 72400, 5878243 rather than related categories to reduce costs.

**8. Conclusion & Lessons Learned**

**Conclusion**

To conclude our session data, the total session amount dropped in December 2019 and peaked at 2020. Among all event types only 6.2% are purchase event, most events are view and cart operations. The event amounts fluctuate with the session amount overtime. The most purchased products price range is 10 dollars and below. The average purchase times per session is 0.21.

The user behaviors have apparent patterns. Regarding event time, there are three peaks of online users: 5, 11, and 18 o’clock, these relate to different groups of people. Regarding purchasing habits, there are 4 clusters of customers. The first cluster are users who loved browsing instead of taking actions, they visited our shopping website less than three times in five months. They each contributed a little to the website’s sales, but the number of users is large in this cluster, so their purchasing power is enormous and indispensable. The second cluster are active customers. They visited our shopping website very often, purchased a lot. Keep pushing new and grand products to this cluster’s users would be a proper strategy. The third cluster customers seldom visited our shopping website but hey purchased a lot. The fourth cluster of users are highly active in terms of putting products into cart, but they seldom purchase.

Regarding different types of users we would recommend different products and using different advertising strategy. For this purpose, we developed a rating generation formula based on the session data that we have. The recommendation system we then built based on ALS algorithm recommend the top 10 most suitable products for each costumer. We would see more clearly if we have more information about products that the recommended products correspond to the user cluster features.

**Lesson learned**

In this project we dig deep into the insight of the collaborative filtering recommendation algorithm. Different from the conventional user-item figure rating, the data we have doesn’t contain explicit rating information. We explore the user-item relation through developing our own formula of rate based on their shopping behavior online. The logic behind is that rating shows customers' attitude toward each product, and the view, cart and purchase behavior shows that too. That’s how we combine our own thoughts into existing algorithm in this project.

**9. References**

<https://www.kaggle.com/mkechinov/ecommerce-events-history-in-cosmetics-shop>