Identifying Cannabis Dispensary Purchase Patterns with Market Basket Analysis

By

Paul Kitko

A Capstone Project Paper Submitted in Partial Fulfillment of the

Requirements for the Degree of

Master of Science

In

Data Science

University of Wisconsin – Oshkosh

Oshkosh, Wisconsin

August 2021

ABSTRACT

Identifying Cannabis Dispensary Purchase Patterns with Market Basket Analysis

Paul Kitko
Capstone Project for Data Science Master's Degree
University of Wisconsin-Oshkosh
Oshkosh, WI
August 2021

Market Basket Analysis (MBA) sometimes referred to as Association Rule Mining, Affinity Analysis or Frequent Itemset Mining, was developed as a method to evaluate "if/then" associations that arise between elements in a dataset (Agrawal & Srikant, 2000). Historically MBA rule sets have been applied to retail grocery stores' Point of Sale data to develop likely product associations that can then be used to anticipate and recommend combinations of future purchases. These recommendations or "cross-sells" have been found to be useful in improving retail sales volume. The newly legalized recreational cannabis market offers an opportunity to apply MBA to an unexplored retail industry.

This project used MBA on a retail cannabis dataset representing multiple dispensaries across the state of Washington. The project's purpose was to verify if MBA was feasible in uncovering useful product association rules from a cannabis sales dataset to use in cross-selling recommendations.

the results of the study show that it is possible to derive meaning MBA rule sets from cannabis retail data but that some limitations were uncovered that offer three future opportunities for research. First that similar product with highly differentiated names may need to be re-categorized into more generalized and meaningful products. Second, that it is possible that product churn may introduce signal noise into the MBA process resulting in a higher number of less useful rule sets. Third, that cannabis customers tend to purhase within product families which is an atypical finding in MBA and should be futher explored.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1 - INTRODUCTION

Market Basket Analysis (MBA) sometimes referred to as Association Rule Mining, Affinity Analysis or Frequent Itemset Mining, was developed as a method to evaluate "if/then" associations that arise between elements in a dataset (Agrawal & Srikant, 2000). For example, in American grocery store transactions, if customers buy a product like peanut butter, then MBA can show how likely those customers are to buy the other complimentary products to make a sandwich, like bread and jelly. The associations that emerge from MBA can be grouped together to form "rules" that are useful in understanding consumer purchasing patterns. Historically MBA association rules add evidence-based insights that inform retail marketing strategies. Applied MBA has led to improved cross-selling opportunities, creation of attractive product bundling promotions (Ruchika & Warikoo, 2017), generated customer friendly retail floorplans (Joe, et al., 2019) and has even been used to develop improved item pairings for restaurant menus (Ting, et al., 2010). Uncovering associations with MBA has proven so versatile that it has also been applied to diverse fields like genomics, aviation, energy, and medical diagnosis, but a review of literature shows that most of MBA's contributions are made in the retail space.

An area of unprecedented retail growth where advanced analytical techniques like MBA have seen very little use is in the US recreational cannabis market. According to analysis by FlowHub (2020), a cannabis data and software platform company, the US cannabis market is expected to be worth between $80 and $100 billion by 2030. To date 17 states have legalized the use of recreational cannabis driving growth and clearing the way for establishing retail outlets, called dispensaries. Each state also recognizes cannabis as a controlled substance, and as such, it must be managed and regulated across the entire supply chain. This concept is

commonly known as “seed to sale” (STS) tracking (Kees et al., 2019). As the name implies, STS tracking starts with sourcing seeds or plant clones. From there, planting, cultivation and, harvesting activities are tracked by producers buy tagging “batches” of product commonly referred to as flower or bud, with barcoded Radio Frequncy ID tags. While harvesting, batch samples are transported via certified delivery companies to cannabis testing labs for contaminant compliance. The same delivery companies also ship tested batches to processors who turn the raw cannabis flower into derived products like edibles, inhalers, and oils. These derived products are also tracked with barcoded RF ID tags. All products, whether in raw form direct from the producer or derived products from a processor, are shipped to dispensaries for final sale to the customer.

Each state’s cannabis market requires a STS platform that collects and tracks each transaction for compliance and tax purposes. These platforms are developed by third party vendors and generate an immense amount of data at each stage of the STS value chain. The final link in the chain is the dispensary, and when equipped with point-of-sale data tracking systems integrated with the STS platform, opens the door to many of the same data-driven marketing opportunities that conventional retailers use like MBA.

Inspiration

The inspiration for this project comes from two sources. The first is the opportunity to apply proven data science concepts to a new data source from a relatively untapped market, thereby increasing the chances of making a concrete contribution. The second is to assist a new software services startup company called Cannlytics based in Olympia, Washington. Cannlytics is owned by Keegan Skeate and currently specializes in developing productivity improvement

software for cannabis testing labs. Conducting retail sales analysis in the form of MBA explores a part of the STS value chain that Cannlytics can offer as a new service to potential customers.

Objectives and Purpose

The project's overall objective is to show if MBA can uncover retail purchasing patterns (association rules) in a large STS cannabis data set and if those patterns can be leveraged by dispensaries to make cross-sell marketing decisions. The project's results will serve as either a potential service that the client can offer to retail dispensaries or as content for a whitepaper that the client can use as promotional research material.

This project will explore association rules generated from specific product names generated from the sales of three representative dispensaries and results will be contrasted to make recommendations on how to use the data. Sub-objectives to achieving the main goal include:

1. Explain, in accessible terms appropriate for the client, what MBA is, how it works and how it is useful in retails sales
2. Procure a representative STS cannabis data set containing raw retail sales transactions appropriate for conducting MBA. This project will be using a dataset supplied by the Washington State Liquor and Cannabis Board.
3. Due to the size of the Washington State STS database, this project will leverage a cloud computing environment for data pre-processing, querying and MBA model development.
4. Use exploratory data analysis to understand the data sets and identify a group of representative dispensaries that make suitable candidates for a comparative MBA analysis.

5. Develop key visual results that Cannlytics can leverage as a possible sales tool for offering MBA services to dispensaries.

Limitations

Due the substantial size of STS datasets, considerations must be made to effectively handle the data. In particular effectively uploading, pre-processing, and querying the raw data must be taken into account. I selected Google Cloud's Big Query service for these activities due to its ease of use when compared to alternatives like Amazon Web Services and Microsoft Azure. Big Query leverages standard SQL for data manipulation. Developing the MBA model will also be accomplished in Google Cloud using their AI Platform. Derived modeling datasets are anticipated to be quite large (sparse matrices) but the AI Platform environment sould have ample computing resources to process it. All AI Platform MBA models will be developed in the R programming language within a "notebooks" environment.

Organization of Paper

The paper's final structure will be as follows:

I) Introduction
    a) Overview of Market Basket Analysis and Its Applicability to Cannabis Retail
    b) Inspiration
    c) Objectives and purpose
    d) Limitations

II) Review of Current Literature and Research
    a) Origins and development of MBA
    b) Understanding MBA association rules and measuring their strength

c) Use cases for MBA

d) Recent developments in MBA research

III) Methods

a) Data acquisition and understanding seed-to sale data schema

b) Transformation of data formats, perform data quality checks

c) Exploratory data analysis to identify candidate retail dispensaries

d) MBA data preprocessing and modeling

IV) Results

a) Fine-tuning MBA association rules

b) Applicability of results for retail dispensaries

V) Conclusion

a) Summary and takeaways

b) Recommendations for further development

CHAPTER 2 - LITERATURE REVIEW

The introduction of legalized recreational cannabis in 17 states has dramatically increased the number of retail dispensaries across the country. Like any traditional retail outlet, dispensaries are faced with typical marketing decisions to increase product sales. Typical retail practices are being used but to date there is very little evidence to suggest that sophisticated analytical technique like MBA have been employed in the cannabis industry. This literature review confirms that dispensaries have made little progress using advanced analysis to support traditional retail decisions at the consumer level. More specially, this review highlights the potential for using MBA to develop cross-selling strategies in the emerging cannabis market.

The Link Between Market Basket Analysis and Cross-selling

Many traditional retail sales strategies have been refined using MBA to increase customer sales. Cross-selling, up-selling, store layout optimization, product shelf placement, customer affiliation marketing and product bundling have all been used based on MBA-generated association rules (Kamakura, 2008; Derdenger & Kumar, 2013; Ting et al., 2010; Bermúdez et al., 2016; Berg et al., 2018). Cross-selling appears to be the most popular retails sales strategy in much of the MBA literature and is intended to increase customer "share of wallet" while providing value-added goods and services. According to Kamakura (2008) there are five advantages to cross-selling:

1. It takes much less time to cross-sell a new product to an existing customer that trying to recruit a new customer to the same product.
2. Cross-selling response rate are between 2 to 5 times more effective than cold-calls.

3. There is evidence that suggests cross-selling deepens the customer to seller relationship improving prospects for increased “share of wallet.”
4. Deeper customer relationships lead to improved customer retention.
5. The customer to seller relationship deepens as the seller learns more about the customers wants and needs resulting in a competitive advantage.

The five advantages of traditional cross-selling create an obvious business incentive, but it wasn’t until retail technology developments in the early 1990’s that cross-selling could be scaled and adopted across the retail industry.

Market Basket Analysis and Cross-selling Adoption in Retail

The adoption of MBA for cross-selling can be attributed to the proliferation of PoS (Point of Sale) systems. Traditionally salespeople would broker cross-sell opportunities, but PoS systems have turned human-mediated sales into streamlined transactions that collect and store vast amounts of consumer data (Kamakura, 2008). Advancements in data storage drove down the cost of storing PoS data and improved processing capability which made analyzing very large data sets feasible (Avcilar & Yakut, 2014). By early 1990’s the convergence of affordable storage and improved processing power with ubiquitous PoS systems opened the door to a variety of data analysis techniques including MBA. Combining MBA with cross-selling became an obvious approach once PoS technology was in place.

The concept behind cross-selling is that a retailer can offer customers alternate products and services based what they have already purchased. Said another way, cross-selling poses the question, “If you like this, then you may like that.” The cross-selling idea aligns nicely with the “if/then” MBA association rules, mentioned previously, that arise between elements in a dataset

(Agrawal & Srikant, 2000). The conceptual overlap between MBA and cross-selling has led to a variety of applications to the retail market.

Market Basket Analysis and Cross-selling Use Cases

Foundational work from Agrawal, Imielinski and Swami introduced MBA in 1993 and it quickly became one of the most used algorithms in retail. MBA refers to the act of multiple shoppers loading various products into their respective shopping baskets resulting in interesting aggregate combinations. These combinations can then be analyzed to potentially improve product offerings (Avcilar & Yakut, 2014). A survey of the literature reveals a diverse set of MBA use cases.

In one interesting example, MBA was used to develop an "ideal" menu assortment of entrees and side dishes for a restaurant (Ting, Pan & Chou, 2010). The authors examined over 3,700 dining transactions for the associations generated between 24 entrees and 49 side dishes. The combinations uncovered with the highest frequency and strength of association were then used to make dining suggestions to undecided customers. The cross-sell suggestions generated in the study were welcomed by customers two out of three times.

In a sperate study, MBA was applied to improving grocery store layouts, commonly called block designs (Ozgormus & Smith, 2020). The underlying logic goes like this: MBA was used to identify those product combinations of most interest to the customers. The combinations were optimized in conjunction with the stores' physical layout and those product combinations that could realize the most cross-sell revenue. From there a set of block designs were created from which the retailer could choose the most appropriate one.

MBA not only offers a data driven approach to block designs, but optimized floorplans can also be used to improve cross-sell promotional strategies in the store. Bermúdez, Apolinario & Abad (2016) first analyzed 24 product families for association rule strength. The resulting associations were then positioned in the store in such a way as to maximize the travel distance customers take to collect complementary products. The strategy of maximizing travel distance means customers are exposed to additional offerings during their shopping visit which raises the likelihood of unplanned purchases. Once the block design was implemented promotional advertisements were placed adjacent to complimentary products incentivizing trips though the store to gather the additional cross-sell goods.

These examples show MBA can play a key role in data driven retail strategies like cross-selling, but this versatile data mining technique has been applied to many domains other than retail. Genetics research by Anandhavalli et al. (2010) developed association rules that determine the relationships between a single gene expression and thousands of other genes expression patterns. In another interesting use case, Bhagwandin et al, (2017) generated association rules with MBA to show which home appliances are used in conjunctions with others to determine energy saving patterns. These patterns, when used by smart home management systems as a rule set, were able to automatically turn on and off co-used appliances to increase energy efficiency. A recent study by Distefano and Leonardi (2020) used association rule mining to uncover which conditions were most associated with aircraft runway excursions (accidents) at airports. In this study the equivalent of the retail transactions that we see in traditional MBA are the taxi routes aircraft takes before or after takeoff. Various contributing factors like type of aircraft, aircraft systems fault and human error were then used to generate patterns associated with the runway excursion incidents.

Criticisms and Considerations

For all its support and broad application MBA does have some detractors. Zhang et al. (2021) argue that although MBA uncovers associations rules it falls short in attributing any linkage to consumer demographics. Zhang argues that rules may have weak associations in aggregate but when attributed to consumer subpopulations the strength of association can be substantially increased. Essentially Zhang sees this as a lost opportunity to better inform retail decision making. An additional criticism from Vindevogel et al. (2005) goes a step further suggesting that complimentary elements of associations rules (the products) can be substituted thereby making the association rules useless to retailers. Instead Vindevogel suggested promoting sales based on cross-price elasticates.

Cross-selling does have its limitations and may not be appropriate in all circumstances. For cross-selling to work it must be seen by the customer as a beneficial service and not just a sales tactic. For instance, not meeting the customers original needs first before introducing cross-selling suggestions or “over-touching” can result in turning off the customer (Kamakura, 2008). Additionally, if a cross-selling strategy is employed on the sales floor, then salespersons should exercise the skills for making recommendations to customers based on MBA analysis and results. Salespeople must also be effectively incentivized to deliver appropriate cross-selling suggestions to customers. A balance must be met to not overly incentivize salespersons leading to disruptive competition at the expense of the customers’ experience (Kamakura, 2007). Although these criticisms and limitations may have merit in their specific use cases this project defers to the broad and proven application of MBA to cross-selling.

How does Market Basket Analysis Work?

At its core, MBA is an exploratory data mining technique used to derive frequent and highly associated “in/then” patterns from large retail datasets. Foundational work from Agrawal, Imielinski and Swami introduced MBA in 1993 and it became one of the most popular algorithms in retail. MBA refers to the act of multiple shoppers loading various products into their respective shopping baskets resulting in interesting aggregate combinations. These combinations can then be analyzed to potentially improve product offerings (Avcilar & Yakut, 2014).

As previously mentioned, MBA and the association rules it generates are derived from PoS transactional datasets, but two challenges arise when generating association rules from large data sets. The first is that the number of rule combinations grow exponentially with the size of the database. It’s easy to imagine the huge number of rule combinations that can be generated from a grocery store filled with thousands of products. The second challenge is selecting interesting subset of rules form a potentially large pool (Hipp, et al. 2000). The first challenge is addressed with a concept called the apriori principal which “prunes” the size of association rule sets and the second challenge is managed by applying thresholds called support, confidence, and lift that restrict the size and applicability of the association rules. Let’s look at the first challenge in more detail.

The apriori principal asserts that if an item set (association rule) is infrequent then all supersets that contain the same item set must also be infrequent. For example, in a grocery store, if cookie sales are infrequent then cookie and milk sales should also be infrequent. The concept is quite intuitive and regulates association rule frequencies. The name apriori originates from idea that association rules use “prior knowledge” about subsets to deduce what can be contained in a superset. The apriori principal was incorporated into the “apriori algorithm” developed by Agrawal, et al. (1993) in their landmark paper, “Mining Association

Rules between Sets of Items in Large Databases." The paper established the foundational concepts for MBA. As a testament to this paper's importance to the data mining community, as of this writing Google Scholar showed over 23,000 citations.

As mentioned previously, there is second method to limiting a potentially large number of MBA association rules. The second method is based on the understanding that association rules generated by the apriori algorithm come in the form of two "if/then" parts. The first part is called the antecedent and the second part is called the consequent. The rule format takes the form of {antecedent} ⇒{consequent} which is read as "antecedent implies consequent." Here are some examples:

{peanut butter} ⇒{jelly}

{hotdogs} ⇒ {buns}

{hotdogs, buns} ⇒ {ketchup}

{hotdogs, buns} ⇒ {ketchup, mustard}

Notice the hotdog examples. Both the antecedent and consequent can have more than one item. Intuitively it's easy to understand how quickly the number of combinations can grow from large retail data sets. Due to the power of the apriori algorithm it's possible to limit the number of association rules generated by incorporating thresholds called support, confidence, and lift.

Support is a measure of how frequent an association rule is in proportion to the total number rules generated from a dataset. Support can be thought of as measure of a rule's "popularity." For example, if the rule {hotdogs, buns} ⇒ { ketchup} occurs 10 times in set of 100 possible association rules the support value is 10/100 or 0.1.

$$Support = \frac{\{\text{hotdogs, buns}\} \Rightarrow \{\text{ketchup}\}}{\text{Total}}$$

What's convenient when using the apriori algorithm is that you can set a support threshold to discriminate infrequently occurring rule sets. In the example above if we set support to .2 then the {hotdogs, buns} ⇒ {ketchup} rule would not be included in the result set of association rules helping surface more frequently occurring rules. Support can never have a value greater than 1 due to the apriori principal.

Another threshold used for association rules is called confidence. Confidence is the implied strength between the antecedent and the consequent and can be thought of as how likely the consequent would occur given the antecedent was purchased.

$$Confidence = \frac{\{\text{hotdogs, buns}\} \Rightarrow \{\text{ketchup}\}}{\{\text{hotdogs, buns}\}}$$

Like support, confidence can be set to a minimum value thus discriminating lower value associations that are not deemed as important for making retail decisions. Confidence values are also limited to value less than or equal to 1, but some caution should be considered when using it. In the above formula the antecedent's popularity is taken into account but the consequent's popularity is not. If the consequent is also popular then there is a possibility that confidence could be artificially inflated. To address the possibility of artificial inflation the measure of lift is used. Lift works the same way confidence does except it controls for how popular the consequent is.

$$Lift = \left(\frac{\left(\frac{\{\text{hotdogs, buns}\} \Rightarrow \{\text{ketchup}\}}{\{\text{hotdogs, buns}\}} \right)}{\left(\frac{\{\text{ketchup}\}}{\text{Total}} \right)} \right)$$

Lift values greater than 1 show strength of association with larger values indicating more strength. For this project lift will be the primary measure used to evaluate MBA associations rules in the Washington State retail cannabis dataset.

An Opportunity to Apply Market Basket Analysis to Cannabis Dispensaries

To date the literature shows very little research in cannabis retail sales analysis other than high-level parallels to the cannabis market drawn from the tobacco and alcohol industries. (Berg, et al. 2019). Results from Berg's Marijuana Retail Surveillance Tool survey confirmed that dispensaries are employing several strategies to improve product availability, price, and promotion. Price discounts, social media promotions, loyalty programs and indoor signage for novel products were being used to influence customer buying decisions but Burg made no mention of more sophisticated marketing techniques being used. Additional research for this project identified a single vendor called Headset which claims to offer "basket analysis" for its clients but a deeper investigation into the offering revealed that the firm offers summary statistics of items purchased at an aggregate market basket level and not true MBA association rules that are applicable to cross-selling strategies. The lack of evidence showing data driven approaches to cannabis retail analysis creates an opportunity for applying more sophisticated techniques like exploring how MBA can be used for cross-selling opportunities.

CHAPTER 3 - METHODOLOGY

The increasing legalization of recreational cannabis has necessitated the need to develop and implement STS tracking systems. Since cannabis is considered a controlled substance STS tacking systems provide the traceability that states need to monitor the entire cannabis supply chain including retail sales. Resulting large retail data sets offer an opportunity to uncover product cross-selling strategies that round out dispensary product offerings potentially boosting revenue. STS tracking systems encompass retail sales transactions in the same way standalone PoS systems do, therefore this project attempts to uncover cross-selling opportunities in the form of association rules with MBA. The literature bears out this approach as it shows a tight coupling between product-cross-selling, PoS systems and MBA (Agrawal & Srikant, 2000; Kamakura, 2008; Avcilar & Yakut, 2014; Gupta & Mamtora, 2014). The objective of this project was to validate that MBA associations rules can be used to develop product cross-selling lists for cannabis dispensaries. Three primary steps were used to accomplish this as seen in figure 1.

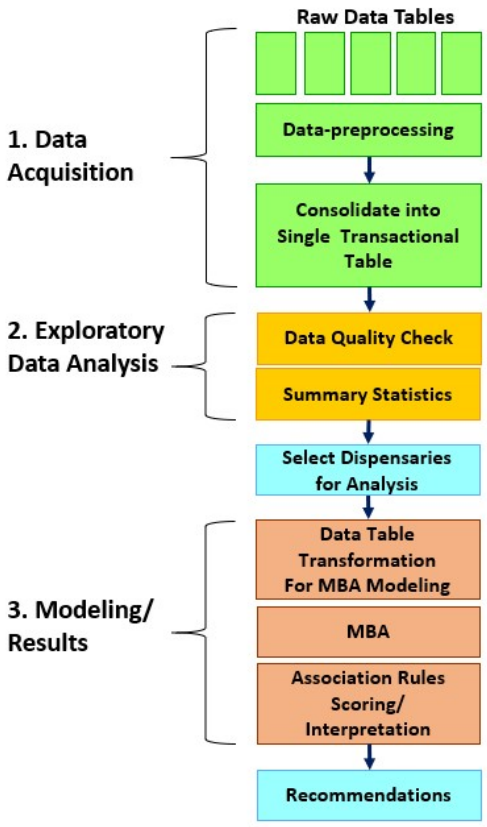


Figure 1. Analytical approach used in this project

The first step, data acquisition, delt mainly with pre-processing non-standard STS data files into an usable format appropriate for importing into a cloud environment. In the second step, exploratory data analysis, the data was checked for overall quality and completeness. A series of summary statistics were developed to identify three candidates' dispensaries appropriate for further MBA analysis. In the third step, MBA was run against the three candidate dispensary data sets. Assocation rules were fine-tuned and then model results were interpreted in terms of cross-selling opportunities. Although the process seems sequential there were some iteration loops which Iare highlighted later where appropriate. The following sections explain the three steps in further detail.

Data Acquisition and Preprocessing

This project used an STS data set provided by the Washington State Liquor and Cannabis Board. Washington State uses a proprietary STS application from Leaf Data Systems (Leaf Data Systems, 2021) which has collected data since 2018. The entire dataset consisted of 22 files ranging from 192 bytes to 5.5 GB in size. The combined dataset was over 29.8 GB. The size of the dataset made it too large to be explored and analyzed in its entirety on a conventional laptop, therefore I decided to use Google Cloud, a suite of cloud computing services, to carry out data storage, exploratory data analysis and modeling. I chose the R statistical programming language in a Jupyter Notebooks development environment for my exploratory data analysis and data modeling since it was also available as a service in the Google Cloud suite.

The Google Cloud suite required a specific data encoding format for data imports called Unicode Transformation Format 8 (UTF-8) which is considered a standard for encoding tabular text data. The STS dataset posed a challenge in that it was encoded using UTF-16, a less common alternative to UTF-8, and a format incompatible with Google Cloud services. Additionally, the dataset was in tab delimited file format. This is not typically an issue for data imports into Google Cloud, but the decision was made to convert the dataset into the more common comma delimited file type as preventative measure to future issues.

Carrying out the data encoding and file type translation presented a “chicken and egg” dilemma. Converting data encoding and file types is typically carried out by opening conventionally sized local files in a tool like Microsoft Excel and then saving them in the required format. This approach was not possible due to the size of most of the dataset files. Instead Microsoft PowerShell was used which is a Windows administrative scripting utility available on most Microsoft laptops. Using PowerShell allowed me to convert the files from their original

format to UTF-8 comma delimited files without opening them. The large files took up to 6 hours to process. PowerShell scripting examples are available in Appendix A. Once properly converted all files were uploaded into the Google Cloud Storage service.

Along with the STS data files, Washington State Liquor and Cannabis Board also supplied a data design document defining the various tables, fields, primary/secondary keys (table joins) used by the LEAF application. Based on the document, the following data schema was composed to complete the view of dispensary sales transactions (see Figure 2). A complete singular transactional view of the data was prerequisite for MBA processing.

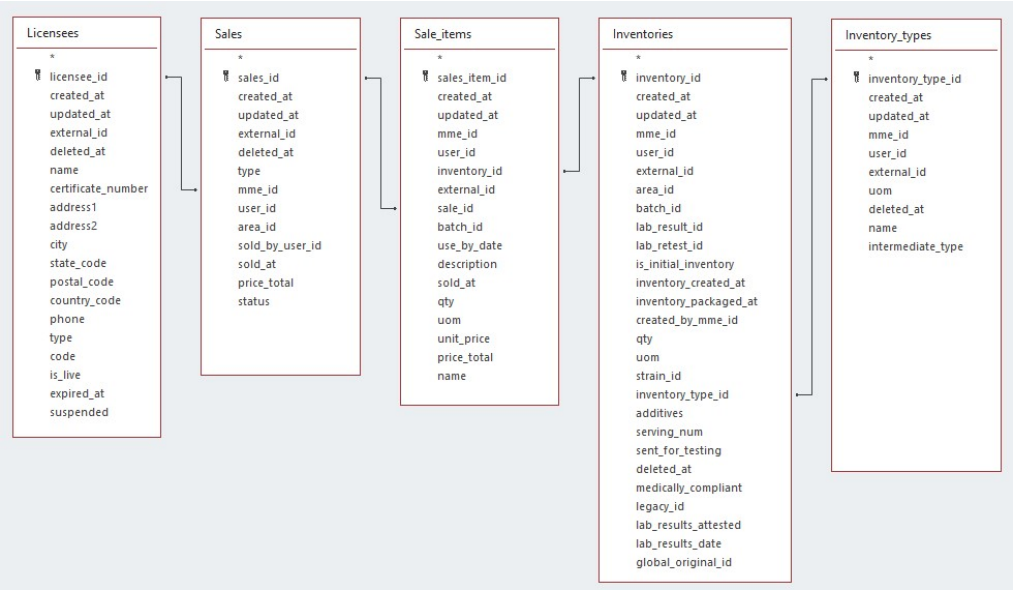


Figure 2. A view of the five Seed to Sale retail data used in this analysis. Tables were joined by primary and secondary keys.

To give further context to the data here are short descriptions for each table in figure 2 and how they are related:

- Licensees – This table is a listing of licensed facilities: producers, processors, testing labs, dispensaries, and transporters). Each licensee holds an official registration

number issued by Washington State Liquor and Cannabis Board. This project only references licensed dispensaries. The Licensees table is linked to the Sales table by a unique licensee ID

- Sales – The sales table contains all wholesale and retail transactions with one unique sales identifier assigned to each transaction. This project only references retail sales. The Sales table is linked to Sales_Items table via a unique Sales ID.
- Sale_Items – Each sale can contain a collection of one or more sale items. Each sale item has a product name and a unique identifier assigned. All products are assigned to a unique inventory "lot" so each sale item can be traced back to the original producer. The Sales_Items table is linked to Inventories table via a unique Sales_Item ID.
- Inventories – Inventories are "lots" of products originating from growers and valued added processors. The Inventories table is linked to Inventory_Types table via a unique Inventories ID.
- Inventory_Types – There are many inventory types ranging from originating products like "Harvest Materials" which are typically various types of cannabis flowers/buds to "End Products" like edibles, usable marijuana and concentrates for inhalation. This project only refers to end products sold through dispensaries. This table also has a unique identifier for each inventory type.

Data Quality and Exploratory Data Analysis

Data quality of STS source tables were verified once they were loaded into the Google Cloud Storage service (Merino et al., 2016). The Skimr reporting package for R (Using Skimr, 2021) was chosen to check table quality due to its convenience, coding simplicity and its

comprehensive reports. A Skimr report was generated for each table which examined four variable types: character, logical, numeric and date. Reports were based on a 5% simple random sample due to the number or records in each table. Skimr evaluated fields for missing values, completion rate, minimum value, maximum value, empty values, and unique values. Each report showed that the data quality was high with all required fields showing 100% completion rates and all date fields falling within anticipated ranges (see Appendix B for an example of Skimr report results).

Exploratory Data Analysis (EDA) queries were developed in Google BigQuery to gain a better business understanding of the five STS source tables' content. More specifically, exploratory queries were developed that identified which dispensaries made good candidates for MBA. To achieve this objective a query was written to join all five STS tables together into a single view of cannabis dispensary sales. Data was limited to the most recent calendar quarter available - 10/1/2020 to 12/31/2020. Since MBA's purpose is to identify useful product combinations in high volume transactional data when the product base is large (Agrawal et al. 1993) the main criteria for identifying the candidates were:

1. Dispensaries that had the highest transactional volume
2. Dispensaries that had the greatest product mix
3. Dispensaries that had the greatest earnings
4. Dispensaries that represented three different socio-economic categories based on median household income (low, medium, high)

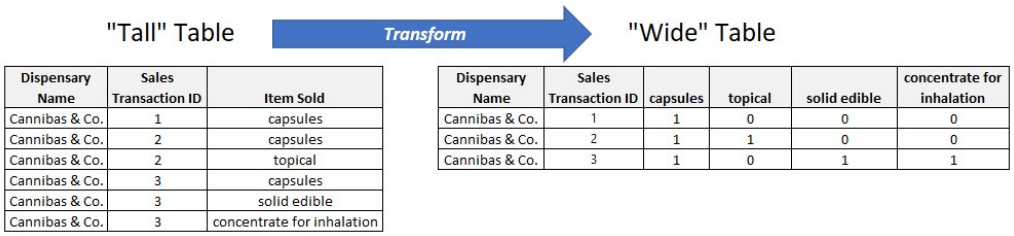
The first three items above were used to generate an ordered list of candidates with the fourth point being used to select the three final top scoring dispensaries from their respective household median income categories. Washington state median household income data (by zip

code) data was gathered from US Census Bureau website (Median Income in the Past 12 Months (In 2019 Inflation-adjusted Dollars), 2020). The median income amounts were matched to each dispensaries' zip code. Once the candidate dispensaries were identified and matched to respective median household zip codes it was possible to move forward with the MBA process.

Dispensary Market Basket Analysis

This project's objective was to show if MBA can uncover purchasing patterns (association rules) in a large STS cannabis data set and to determine if those patterns were potentially useful for developing retail cross-selling strategies. The previous section outlined the process for consolidating the STS data into a usable set of sales transactions, but one additional data processing step was required as a prerequisite to using MBA that altered the structure of the transactional data from a "tall" table into a "wide" table.

Figure 3 illustrates the concept of transforming a tall table into a wide table. Transactional tables are sometimes called tall due the fact that retail PoS systems collect large numbers of sales records resulting in a long list of transactions. This project used the arules MBA algorithm in R (Comprehensive R Archive Network (CRAN), 2021) that required transaction data to be pivoted into a wide table where each product is given its own column as in Figure 3.



Dispensary Name	Sales Transaction ID	Item Sold
Cannibas & Co.	1	capsules
Cannibas & Co.	2	capsules
Cannibas & Co.	2	topical
Cannibas & Co.	3	capsules
Cannibas & Co.	3	solid edible
Cannibas & Co.	3	concentrate for inhalation

Dispensary Name	Sales Transaction ID	capsules	topical	solid edible	concentrate for inhalation
Cannibas & Co.	1	1	0	0	0
Cannibas & Co.	2	1	1	0	0
Cannibas & Co.	3	1	0	1	1

Figure 3. Transforming a tall transactional table into a wide table.

Every transaction shows a binary value for each product column depending on whether that product is present in the transaction. For example, in Figure 3 the tall table shows four possible products against transactions 1 through 3. Each one of the products is assigned its own column in the wide table with a 1 or a 0 being assigned if the product is present in a transaction. Transforming a long table in short table was done in R using the "reshape" function. A code example can be found in Appendix C. Once the data was pivoted into the correct form it was ready for MBA.

Running the arules MBA algorithm is relatively straightforward using only one line of R code but it requires two numeric input values: support and confidence. Selecting these two values was a subjective and iterative process based on the list of association rules generated. Recall from Chapter 2 that support is a measure of association rule "popularity" in relation to all the association rules generated form a data set. Also recall that confidence is the "strength of association" going from the left side of the rule (antecedent) to the right side of an association rule (consequent). Support was used as a filter to narrow rule set to the most significant items and confidence was adjusted to tease out those rules that had the highest level of association.

By varying support and confidence values for each of the three dispensaries, several different visualizations were generated that expressed the relationships of the association rules. Greater context and intuition of the cannabis product rule space was gained. A graph matrix was used which shows how the various elements in the list of association rules are related. A scatter plot was used to show the relationship between support, confidence, and lift. Recall from chapter 2 that lift works the same way confidence does except it controls for how popular the consequent is. Lift values greater than 1 show strength of association with larger values indicating more strength. After taking support and lift into consideration as the main measures a

final list of association rules for all three dispensaries was developed. The final list was used as a source for cross-selling recommendations.

CHAPTER 4- RESULTS & DISCUSSION

The objective of this project was to analyze cannabis retails sales data to validate if product cross-selling suggestions can be generated through MBA association rules. The approach used to answer this question consisted of three steps: data acquisition, exploratory data analysis and modeling results. The specific activities around data acquisition was covered previously. Here we will outline which candidate dispensaries were selected for MBA analysis and the associations rules generated by MBA.

Identifying Candidate Dispensaries for Market Basket Analysis

Before MBA modeling could be used to develop cross-selling association rules it was important to identify a list of representative dispensaries for analysis. As mentioned in chapter 3, four specific criteria were used to identify a list of candidate dispensaries for MBA. Those criteria were dispensaries with 1) the highest transactional volume, 2) the greatest product mix, 3) the highest earnings and 4) representation from low, medium, and high median regions. The first three criteria identified the most commercially active dispensaries with a diverse product set. The fourth criteria, median household income, was then applied to the top dispensaries to segment the market into three income demographics. The four criteria were applied to 2020 dispensary sales data.

Figure 4 contains four histograms and boxplots combinations labeled A though D representing each of the previously mentioned four criteria. Each chart's histogram shows the relationship between the number dispensaries that fall within a criterion's range.

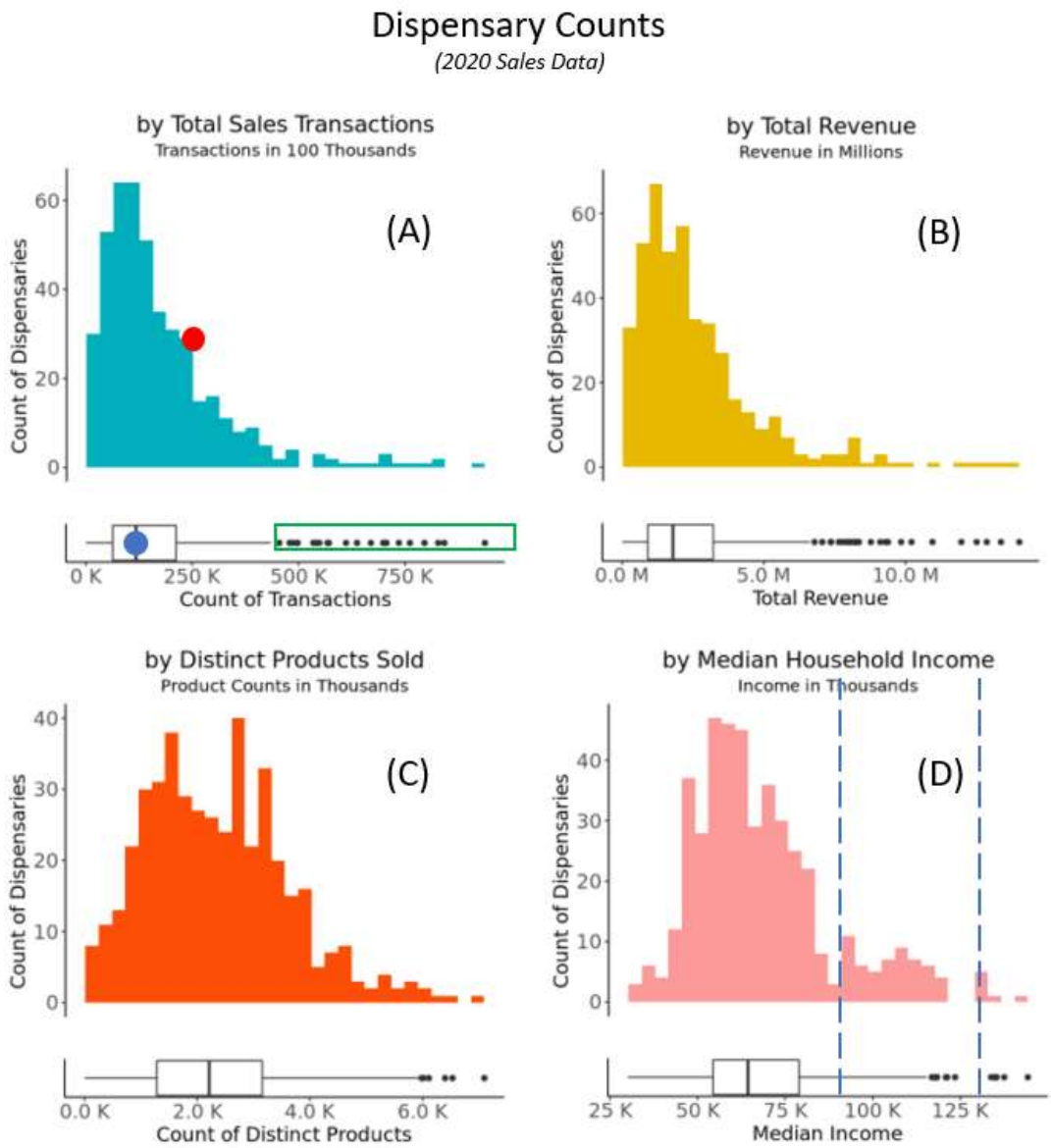


Figure 4. The four charts depict the criteria that was used to select representative dispensaries for MBA.

For example, interpreting the red dot in Figure 4(A) shows about 29 dispensaries had 250 thousand transactions for 2020. Additionally, the blue dot in the boxplot shows a mean transaction value of 125 thousand. The dispensaries of interest are the outliers to the right (green box). In this case these dispensaries offer the greatest transactional volume. Charts B

and C work the same way but for total revenue and distinct products sold. Chart D shows how the count of dispensaries broke down by median household income. The two dashed lines indicate naturally occurring separation in income levels. I selected 0-85K, 85-125K and anything above 125K as my income buckets and labeled them low, medium, and high respectively. Once I had identified all four criteria, I combined them into a single representation so I could identify my top cannabis dispensaries from each income level (Figure 5).

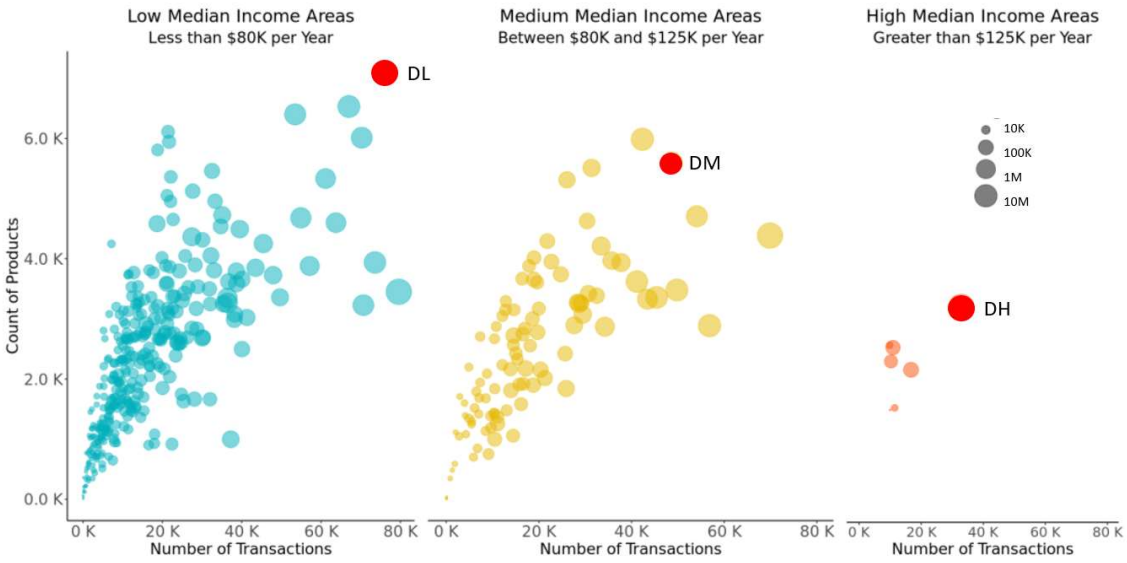


Figure 5. Comparison by Levels of Median Houshold Income.

Each of the three charts above represent a level of median household income (low, medium, high). The remaining three dispensary selection criteria are represented in the charts like this:

1. Highest transactional volume = X-axis,
2. Greatest product mix = Y-axis,
3. Greatest earnings = Size of bubble

The red dots in the upper right side of each chart identify the selected dispensary based on the highest criteria values. There were several dispensaries in the median income chart that could have been selected. The dispensary that made the best tradeoff in criteria values was selected. Table 1 lists the three dispensaries selected for the final MBA analysis.

Label	Median Income Level	Dispensary Name	ZIP Code	Median Income by ZIP	Total Sales	Total Transactions	Product Count
DL	LOW	ZIPS CANNABIS	98408	$ 59,207.00	$ 12,510,607	761858	7095
DM	MEDIUM	PRC	98223	$ 88,295.00	$ 7,357,044	486777	5598
DH	HIGH	GREEN THEORY FACTORIA	98006	$ 144,247.00	$ 5,704,415	329164	3212

Table 1. Dispensary totals sales, transactions and product counts.

Market Basket Analysis Findings

MBA generates lists of product combinations from retail sales data. The resulting association rules are used promote cross-selling opportunities in complimentary products (Ruchika & Warikoo, 2017). As mentioned in Chapter 2, this analysis used two primary measures to generate rules sets: support and confidence. Support can be thought of as a rule's popularity when compared with all other rules sets. Support values are measured in percentages and it is not uncommon for the values to be in the low single percentages when there are many products represented in a data set. Confidence was the second measure used to generate rule sets which can be thought of as the strength of the relationship between a rules first element and second element. Confidence values are also measured in percentages but are not typically as low as support percentages.

Various combinations of support and confidence values were tried during the analysis of the three dispensary data sets. If the support and confidence values were set too low, then there was a chance of generating large "noisy" sets with low value rules. Balancing the two thresholds to generate a useful and reasonably sized rule set was the objective. Figure 6 shows a

support vs. confidence scatterplot which incorporated a support value of 0.004 and a confidence value of 0.4 into the MBA algorithm.

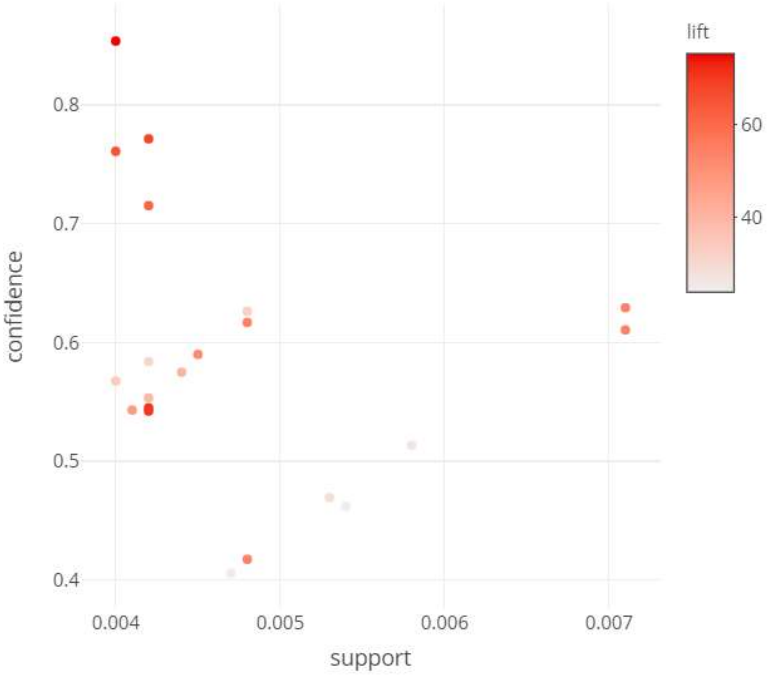


Figure 6. Confidence vs. Support scatter plot.

Notice that the plot area is not inundated with data points. This means that the topmost data points are represented in terms of support, confidence, and lift. If any of the thresholds were lowered more datapoints would be present in the chart.

The additional legend on the right side of Figure 6 represents lift. Lift is like confidence in that it indicates the likelihood of the combination, but the strength of association is bidirectional between the two elements whereas confidence is unidirectional. Lift was used as the key metric to identify “close couplings” for potential cross-selling opportunities. The stronger the bond between the rule elements the greater the lift value. Based on a support of 0.004 and a confidence of 0.4 the MBA algorithm produced 21 rule sets from the DL data as

represented by the red dots in Figure 6. The top ten resulting rules are listed in Table 2 and follow an "if-then" format. For instance, the first DL rule:

{Wax - Chernobyl - 01.0 g, Wax - Pink #1 - 01.0 g} => {Wax - Starfighter - 01.0 g}

Reads like this:

"If Wax - Chernobyl - 01.0 g *and* Wax - Pink #1 - 01.0 g *are purchased together, then* Wax - Starfighter - 01.0 g *is also likely to be purchased"*

Additionally, the DM and DH dispensary top 10 rule sets can be seen in tables 3 and 4 respectively. All tables include the support and confidence values used to generate the rule sets.

Support = 0.004, Confidence = 0.40, Total Rules Generated = 21

Top 10 DL Association Rules	Support	Confidence	Lift
{Wax - Chernobyl - 01.0 g, Wax - Pink #1 - 01.0 g} => {Wax - Starfighter - 01.0 g}	0.0040	0.8537	75.3223
{Wax - Lemon OG - 01.0 g } => {Wax - Pink Panther - 01.0 g}	0.0042	0.5423	70.3375
{Wax - Pink Panther - 01.0 g} => {Wax - Lemon OG - 01.0 g }	0.0042	0.5450	70.3375
{Wax - God's Gift - 01.0 g,Wax - Chernobyl - 01.0 g} => {Wax - Starfighter - 01.0 g}	0.0042	0.7714	68.0668
{Wax - Pink #1 - 01.0 g,Wax - Starfighter - 01.0 g} => {Wax - Chernobyl - 01.0 g}	0.0040	0.7609	65.1410
{Wax - God's Gift - 01.0 g,Wax - Starfighter - 01.0 g } => {Wax - Chernobyl - 01.0 g}	0.0042	0.7152	61.2338
{Wax - Lemon OG - 01.0 g} => {Wax - Blueberry Muffin - 1.0 g}	0.0048	0.6169	53.8835
{Wax - Blueberry Muffin - 1.0 g} => {Wax - Lemon OG - 01.0 g}	0.0048	0.4175	53.8835
{Wax - Starfighter - 01.0 g } => {Wax - Chernobyl - 01.0 g}	0.0071	0.6293	53.8727
{Wax - Chernobyl - 01.0 g} => {Wax - Starfighter - 01.0 g}	0.0071	0.6106	53.8727

Table 2. Top 10 Association Rules for low median income dispensaries.

Support = 0.001, Confidence = 0.01, Total Rules Generated = 34

Top 10 DM Association Rules	Support	Confidence	Lift
{#Blunt Sativa} => {#Blunt Hybrid}	0.0012	0.4091	151.0130
{#Blunt Hybrid} => {#Blunt Sativa}	0.0012	0.4286	151.0130
{#Blunt Indica} => {#Blunt Hybrid}	0.0011	0.4359	160.9084
{#Blunt Hybrid} => {#Blunt Indica}	0.0011	0.4048	160.9084
{Wax - Chocolate Trip - 01.0 g } => {Wax - Timewreck - 01.0g}	0.0011	0.2464	52.3264
{Wax - Timewreck - 01.0g} => {Wax - Chocolate Trip - 01.0 g }	0.0011	0.2329	52.3264
{Wax - Chocolate Trip - 01.0 g]} => {Wax - Ninja Nectar - 01.0 g }	0.0011	0.2464	32.3714
{Wax - Ninja Nectar - 01.0 g} => {Wax - Chocolate Trip - 01.0 g }	0.0011	0.1441	32.3714
{Wax - American Pie - 01.0g } => {Wax - Lodi Dodi - 01.0 g }	0.0013	0.2667	29.9594
{Wax - Lodi Dodi - 01.0 g} => {Wax - American Pie - 01.0g }	0.0013	0.1449	29.9594

Table 3. Top 10 Association Rules for medium median income dispensaries.

Support = 0.001, Confidence = 0.01, Total Rules Generated = 48

Top 10 DH Associaiton Rules	Support	Confidence	Lift
{Legend of Nigeria Prerolls 1g} => {Jack Herer Prerolls 1g}	0.0011	0.2407	46.3627
{Jack Herer Prerolls 1g} => {Legend of Nigeria Prerolls 1g}	0.0011	0.2167	46.3627
{Panda Fruit Drops- Green Apple 100mg } => {Panda Fruit Drops- Peach Mango - 100mg}	0.0010	0.2791	44.1733
{Panda Fruit Drops- Peach Mango - 100mg } => {Panda Fruit Drops- Green Apple 100mg }	0.0010	0.1644	44.1733
{Panda Fruit Drops- Strawberry Kiwi 100mg} => {Panda Fruit Drops- Peach Mango - 100mg}	0.0012	0.2500	39.5719
{Panda Fruit Drops- Peach Mango - 100mg } => {Panda Fruit Drops- Strawberry Kiwi 100mg}	0.0012	0.1918	39.5719
{Blueberry Belts - 100mg} => {Apple Rings - 100mg }	0.0015	0.2787	38.7981
{Apple Rings - 100mg } => {Blueberry Belts - 100mg}	0.0015	0.2048	38.7981
{Lodi Dodi .5G Preroll 2 pk} => {Narnia .5G Preroll 2 pk}	0.0016	0.2727	37.0749
{Narnia .5G Preroll 2 pk} => {Lodi Dodi .5G Preroll 2 pk}	0.0016	0.2118	37.0749

Table 4. Top 10 Association Rules for low median income dispensaries.

The final three tables confirm that MBA can generate product rules sets from STS cannabis dispensary data. There rules uncovered here are the basis for developing cross-selling opportunities for complimentary cannabis products.

Discussion

Analysis of the three dispensary MBA rule sets has confirmed MBA can be applied to retail cannabis data and has also uncovered some interesting observations. The DL association rules in Table 2 show that wax products (wax products are infused with cannabis' main psychoactive chemical, THC, and are typically smoked or vaped) were the only type of product represented in the top ten rules. Further analysis of all 21 rules show the same pattern of wax-only association rules suggesting that this dispensary may focus on wax product sales. The combinations of wax products in Table 2 creates a potential opportunity for DL to cross-promote those combinations either through product placement, advertising, or salesperson recommendations.

Table 3's list of product combinations offers a different view of top ten product combinations. Like DL, DM shows rules that include wax products but in addition we see blunt combinations. A

blunt is the common term for a hollowed-out cigar filled with cannabis. Interestingly blunts are only bought with other blunts and the same holds for the waxes indicating that these products may not be complimentary from the consumers perspective. Analysis shows 24 of the 34 DM rules show a consistent pattern of wax-to-wax or blunt-to-blunt combinations.

Table 4's DH rule set represents the highest mean income demographic of all three groups and shows a departure from the products listed in the DL and DM rule sets. Here we see no wax products at all and instead we see prerolls (a pre-rolled cannabis cigarette) and edibles (referred to as drops, rings, and belts in the rule sets). Again, these two classes of products are bought independently of each other. Analysis of the remaining 48 rules also list extracts for inhalation in the rule set, but again, these products appear to be bought in combination with the same product types. Although we see various of cross-selling opportunities in all three rules sets this initial analysis appears to show a "barrier" to mixing product categories which is not typically present in MBA. Traditional retail MBA, especially in grocery store sales, show a mix of products, as we would expect. Dispensaries seem to show that consumers are connoisseurs of specific products with little propensity to mix and match outside their area of preference. Would this self-selecting behavior limit the applicability to cross-selling between main product categories? Conducting an analysis of a larger sample of dispensaries would need to be carried out to confirm if this product cross-selling barrier truly exists.

Limitations and Recommendations

As previously explained, MBA rules for cross-selling opportunities are generated using support and confidence thresholds. Although these values are relative, low thresholds indicate that there are many low value rule combinations. Ideally rules with higher support, confidence values have more utility. In this analysis all three sets of rules had small support and confidence

values. This can be result of a few factors. High product churn can introduce and take away product combinations injecting low value noisy combinations into market before they have a chance to mature and rise to the top. Complex product naming schemes, like we see in the cannabis data set, can also introduce a high level of granularity whereas assigning products to representative categories can strengthen the probability of generating stronger rule sets. Two possible next steps emerge from these observations. First, conducting a product churn analysis would uncover to what extent cannabis products are transient in the marketplace. Churn rules could be developed to filter the cannabis data set to make it more representative of strong product signals. Secondly, developing a cannabis product taxonomy would generate MBA rule sets into more meaningful product buckets. This would also help confirm if the product cross-selling barrier mentioned in the previous section truly exists. Additionally, a more meaningful product taxonomy would help confirm if segmentation by income impacts product purchasing combinations.

CHAPTER 5 - CONCLUSION

We have shown that market basket analysis (MBA) can be applied to a cannabis seed to sale (STS) dataset for the purpose of generating association rules. Furthermore, we have given examples how the resulting association rules can be used by three different cannabis dispensaries to generate cross-sell opportunities. This project has also uncovered observations that suggest there are further steps available to improve the effectiveness of MBA in this space. Based on the results of this study it is feasible for cannabis dispensaries or analytical service providers to conduct MBA analysis and potentially drive higher retail revenue in the cannabis market.

Several key opportunities for follow on work emerged during this project. First, when conducting MBA on retail cannabis data is the size of the dataset itself. This project used data supplied by the Washington State Liquor and Cannabis Board. Some of the files provided were over 15GB in size which necessitated the use of cloud services to ingest and preprocess the data before any meaningful analysis could be conducted. Those intent on working with large cannabis data sets in the future should consider encouraging state authorities to release their data through cloud services for easier public access.

Another opportunity for future work regards how retail cannabis products are labeled before a meaningful MBA is attempted. MBA data relies on products that are labeled in such a way as to be reasonably descriptive. Multiple uniquely names products within a category may lower the quality of MBA results by injecting large numbers of meaningless combinations which was a limitation in this dataset. In a grocery store setting, labeling cleaning products as launder

detergent and stain removers may yield better MBA rules sets than labeling all the various laundry detergents and stain removers with the brand name. Results from this project showed that overly granular labeling can potentially washout strong association rules necessary to identify cross-selling opportunities. Future work should consider the appropriate level of product labeling before the MBA analysis is conducted. This could take the form of developing a product taxonomy of labels for various cannabis product categories. The resulting taxononmy could then be applied to classifying each product accordingly creating stronger association rules signals.

Additionally, effective MBA analysis requires that a stable list of products is offered over a reasonable period. Quickly introducing new products while rapidly taking away existing ones introduces product churn which prohibits strong association rules from maturing. It may be quite reasonable to expect new retail markets, like recreational cannabis, to initially see product churn. If that is the case, then alternatives to MBA should be considered to develop cross-selling opportunities. The limitations potentially imposed by project churn was not confirmed in this study but follow on work should consider developing churn models to detect and mitigate its presence in retail cannabis data sets.

Finally, this project identified that customers do not typically mix and match product purchases outside of a main product family. If this affinity can be further confirmed then It would be a great help to others conducting MBA in this space in ushowing that cross-selling opportunities may be more effectively executed if they are limited to specific product groupings.

REFERENCES

Agrawal, R., Imieliński, T., & Swami, A. (1993). Mining association rules between sets of items in large databases. *ACM SIGMOD Record*, *22*(2), 207–216. https://doi.org/10.1145/170036.170072

Agrawal, Rakesh & Srikant, Ramakrishnan. (2000). Fast Algorithms for Mining Association Rules. Proc. 20th Int. Conf. Very Large Data Bases VLDB. 1215.

Anandhavalli, M., Ghose, M. K., & Gauthaman, K. (2010). Association Rule Mining in Genomics. *International Journal of Computer Theory and Engineering*, 269–273. https://doi.org/10.7763/ijcte.2010.v2.151

Avcilar, M.Y., & Yakut, E. (2014). Association Rules in Data Mining: An Application on a Clothing and Accessory Specialty Store. *Canadian Social Science, 10*, 75-83.

Berg, C. J., Henriksen, L., Cavazos-Rehg, P. A., Haardoerfer, R., & Freisthler, B. (2018). The emerging marijuana retail environment: Key lessons learned from tobacco and alcohol retail research. *Addictive Behaviors*, *81*, 26–31. https://doi.org/10.1016/j.addbeh.2018.01.040

Bermudez, Jonathan & Apolinario, Kevin & Abad, Andres. (2016). Layout Optimization and Promotional Strategies Design in a Retail Store based on a Market Basket Analysis. 10.18687/LACCEI2016.1.1.307.

Comprehensive R Archive Network (CRAN). (2021, May 17). CRAN - Package arules. Arules: Mining Association Rules and Frequent Itemsets. https://cran.r-project.org/web/packages/arules/index.html

Derdenger, T., & Kumar, V. (2013). The Dynamic Effects of Bundling as a Product Strategy. Marketing Science, 32(6), 827–859. https://doi.org/10.1287/mksc.2013.0810

Gupta, S., & Mamtora, R. (2014). A Survey on Association Rule Mining in Market Basket Analysis. *International Journal of Information and Computation Technology*, *4*(4), 409–414.

Hipp, J., Güntzer, U., & Nakhaeizadeh, G. (2000). Algorithms for association rule mining — a general survey and comparison. ACM SIGKDD *Explorations Newsletter*, *2*(1), 58–64. https://doi.org/10.1145/360402.360421

How the essential industry performed last year. (2021). Flowhub. https://flowhub.com/cannabis-industry-statistics

Joe, T., Sreejith, R., Sekar, K., (2019) Optimization of Store Layout using Market Basket Analysis. International Journal of Recent Technology and Engineering, 8(2), 6459–6463. https://doi.org/10.35940/ijrte.b2207.078219

Kamakura, W. A. (2008). Cross-Selling. *Journal of Relationship Marketing*, *6*(3–4), 41–58. https://doi.org/10.1300/j366v06n03_03

Kees, J., Fitzgerald, P., Dorsey, J. D., & Hill, R. P. (2019). Evidence-Based Cannabis Policy: A Framework to Guide Marketing and Public Policy Research. *Journal of Public Policy & Marketing*, *39*(1), 76–92. https://doi.org/10.1177/0743915619848422

Leaf Data Systems. (2021). LEAF Data Systems. https://leafdatasystems.com/

Malati, N., . R., & Warikoo, B. (2017). Market Basket Analysis and Product Affinity in Retail. Effulgence-A Management Journal, 15(1), 25. https://doi.org/10.33601/effulgence.rdias/v15/i1/2017/25-38

MEDIAN INCOME IN THE PAST 12 MONTHS (IN 2019 INFLATION-ADJUSTED DOLLARS). (2020). United States Census Bureau.

https://data.census.gov/cedsci/table?q=ACSST1Y2019.S1903&tid=ACSST1Y2019.S1903&hidePreview=true

Merino, J., Caballero, I., Rivas, B., Serrano, M., & Piattini, M. (2016). A Data Quality in Use model for Big Data. Future Generation Computer Systems, 63, 123–130. https://doi.org/10.1016/j.future.2015.11.024

Ozgormus, E., & Smith, A. E. (2020). A data-driven approach to grocery store block layout. Computers & Industrial Engineering, 139, 105562. https://doi.org/10.1016/j.cie.2018.12.009

Using Skimr. (2021, March 4). Using Skimr. https://cran.r-project.org/web/packages/skimr/vignettes/skimr.html

Ting, P. H., Pan, S., & Chou, S. S. (2010b). Finding Ideal Menu Items Assortments: An Empirical Application of Market Basket Analysis. Cornell Hospitality Quarterly, 51(4), 492–501. https://doi.org/10.1177/1938965510378254

Ugih Rizqi, Zakka. (2019). Implementation of Association Rule-Market Basket Analysis in Determining Product Bundling Strategy: Case Study of Retail Businesses in Indonesia.

Vindevogel, Bernd & Van den Poel, Dirk & Wets, Geert. (2004). Why promotion strategies based on market basket analysis do not work. Expert Systems with Applications. 28. 583-590. 10.1016/j.eswa.2004.12.019.

Yazgana, Pinar & Kusakci, Ali. (2016). A Literature Survey on Association Rule Mining Algorithms. Southeast Europe Journal of Soft Computing. 5. 10.21533/scjournal.v5i1.102.

Zhang, L., Priestley, J., DeMaio, J., Ni, S., & Tian, X. (2021). Measuring Customer Similarity and Identifying Cross-Selling Products by Community Detection. Big Data, 9(2), 132–143. https://doi.org/10.1089/big.2020.0044

APPENDIX A

PowerShell Scipting Examples

The following PowerShell script was used to convert very large UTF-16 tab delimited files into UTF-8 comma delimited files suitable for use in Google Cloud's BigQuery.

Format:

> Import-Csv **< *import file name* >** -Delimiter "`t" | export-csv **< *import file name* >** -NoTypeInformation -Delimiter "," -Encoding UTF8

Example:

> Import-Csv Inventories_0.csv -Delimiter "`t" | export-csv Inventories_0_utf8.csv -NoTypeInformation -Delimiter "," -Encoding UTF8

APPENDIX B

Data Quality Check with SkimR

The following R output comes form the SkimR package. SkimR reviews each field in a table for missing values and value ranges.

```
── Data Summary ────────────────────────
                           Values
Name                       Piped data
Number of rows             2886
Number of columns          19
_________________________
Column type frequency:
  character                12
  logical                  2
  numeric                  1
  POSIXct                  4
_________________________
Group variables            None

── Variable type: character ──────────────────────────────────────────────────
   skim_variable n_missing complete_rate   min   max empty n_unique whitespace
 1 global_id             0         1         9    11     0     2886          0
 2 external_id          12         0.996     1    16     0     1744          0
 3 name                  0         1         1    45     0     2237          0
 4 address1              3         0.999     8    44     0     2710          0
 5 address2           2357         0.183     2    22     0      202          0
 6 city                  3         0.999     3    17     0      309          0
 7 state_code            1         1.00      2     2     0        2          0
 8 postal_code           3         0.999     5    10     0     1882          0
 9 country_code         17         0.994     2     3     0        2          0
10 phone                 4         0.999     7    14     0     2157          0
11 type                  1         1.00      3    21     0        9          0
12 code                  0         1         2    11     0     2882          0

── Variable type: logical ────────────────────────────────────────────────────
  skim_variable n_missing complete_rate  mean count
1 is_live               0             1 0.672 TRU: 1939, FAL: 947
2 suspended             0             1 0.352 FAL: 1871, TRU: 1015

── Variable type: numeric ────────────────────────────────────────────────────
  skim_variable      n_missing complete_rate       mean        sd    p0
1 certificate_number         7         0.998 606557599. 37613967.     1
         p25       p50        p75      p100 hist
1 603348680. 603358352 603560870. 987654321 ▁▁▁▁▇▁

── Variable type: POSIXct ────────────────────────────────────────────────────
  skim_variable n_missing complete_rate min                 max
1 created_at            0       1       1900-01-01 00:00:00 2021-01-05 18:06:47
2 updated_at            0       1       1900-01-01 00:00:00 2021-01-05 18:07:43
3 deleted_at         2879       0.00243 2018-02-04 11:45:47 2020-01-22 02:43:46
4 expired_at           36       0.988   2014-09-29 17:00:00 2022-08-30 17:00:00
  median              n_unique
1 1900-01-01 00:00:00      666
2 2021-01-05 18:07:00      328
3 2018-02-21 07:47:19        7
4 2021-03-30 17:00:00       81
```

APPENDIX C

Exploratory Data Analysis R Code

In [1]:
```
# Install Packages/Load Libraries
install.packages("ggdist")
install.packages("tidyquant")
install.packages("tidyverse")
install.packages("bigrquery")
install.packages("scales")
install.packages('gridExtra')

library(ggdist)
library(tidyquant)
library(tidyverse)
library(bigrquery) # used for querying BigQuery
library(scales)
library(gridExtra)
```

In [2]:
```
# References
# https://cloud.google.com/architecture/data-science-with-r-on-gcp-eda
```

In [3]:
```
# Authenticate bigrquery using out-of-band authentication - its a type of two-factor identification that Google Cloud uses
bq_auth(use_oob = True)

# Set a variable to the name of the Google Cloud project I created - the ID can be used by APIs
PROJECT_ID <- "woven-sensor-314415"

# Set a variable to the name of the Cloud Storage bucket I created...this is the Google Cloud data repository
BUCKET_NAME <- "washington_data"
```

In [20]:
```
# Assign a query to a variable
sql_query <- "
    SELECT
*
 FROM woven-sensor-314415.Washington_State_Data.2020_summary_stats_for_all_dispensaries
    ORDER BY total_sales DESC
"
```

In [21]:
```
# Run a query against one of the tables in my Cloud Storage bucket

dispensary_data <- bq_table_download(
    bq_project_query(
        PROJECT_ID,
        query=sql_query
    )
)
```

```
Auto-refreshing stale OAuth token.
```

In [22]:

```
dispensary_data$median_income_by_zip = as.numeric(as.character(dispensary_data$median_income_by_zip))

dispensary_data <- dispensary_data %>%
  mutate(median_income_flag = case_when(median_income_by_zip <= 80000 ~ 'low',
          median_income_by_zip > 80000 & median_income_by_zip <= 125000 ~ 'medium',
          median_income_by_zip > 125000 ~ 'high'
          ))


head(dispensary_data)
```

```
Warning message in eval(expr, envir, enclos):
"NAs introduced by coercion"
```

A tibble: 6 × 12

mme_id	name	certificate_number	address1	city	state	zip_code	median_income_by_zip	total_sales	total_transactions	p
<chr>	<chr>	<int>	<chr>	<chr>	<chr>	<chr>	<dbl>	<dbl>	<int>	
WAWA1.MMJI	CRAFT CANNABIS	604338188	1510 N WENATCHEE AVE	WENATCHEE	WA	98801	58670	14015087	796106	
WAWA1.MM4A	MAIN STREET MARIJUANA	603357161	2314 MAIN ST	VANCOUVER	WA	98660	47823	13369120	938088	
WAWA1.MM3F	GREENSIDE	603348991	23407 PACIFIC HWY S	DES MOINES	WA	98198	66906	12868059	828456	
WAWA1.MM4V	ZIPS CANNABIS	604273684	317 S 72ND ST	TACOMA	WA	98408	59207	12510607	761858	
WAWA1.MM13M	MAIN STREET MARIJUANA EAST	603571920	16219 SE 12TH ST STE 104	VANCOUVER	WA	98683	70099	11970832	842792	
WAWA1.MM18B	NIRVANA CANNABIS COMPANY	603359191	4950 ARENA RD	RICHLAND	WA	99352	81410	10950094	698883	

In [23]:

```
summary(dispensary_data)
```

```
   mme_id             name           certificate_number    address1
Length:444         Length:444         Min.   :600314288   Length:444
Class :character   Class :character   1st Qu.:603352391   Class :character
Mode  :character   Mode  :character   Median :603401426   Mode  :character
                                      Mean   :603555018
                                      3rd Qu.:603587950
                                      Max.   :604641709

    city              state              zip_code         median_income_by_zip
Length:444         Length:444         Length:444         Min.   : 30083
Class :character   Class :character   Class :character   1st Qu.: 54436
Mode  :character   Mode  :character   Mode  :character   Median : 64400
                                                         Mean   : 68965
                                                         3rd Qu.: 78966
                                                         Max.   :144247
                                                         NA's   :1
 total_sales       total_transactions product_count  median_income_flag
Min.   :      90   Min.   :    11    Min.   :   2   Length:444
1st Qu.:  901243   1st Qu.: 63151    1st Qu.:1284   Class :character
Median : 1790907   Median :117852    Median :2213   Mode  :character
Mean   : 2436722   Mean   :162152    Mean   :2310
3rd Qu.: 3211780   3rd Qu.:212306    3rd Qu.:3155
Max.   :14015087   Max.   :938088    Max.   :7095
```

In [24]:

```
#dispensary_data %>% filter(median_income_flag == 'high')
dispensary_data %>% filter(median_income_by_zip > 85000)

write_csv(dispensary_data, "dispensary_data.csv")
```

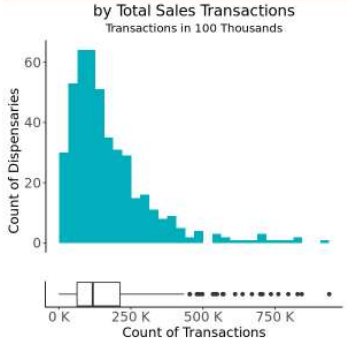
A tibble: 70 × 12

mme_id	name	certificate_number	address1	city	state	zip_code	median_income_by_zip	total_sales	total_transactio
<chr>	<chr>	<int>	<chr>	<chr>	<chr>	<chr>	<dbl>	<dbl>	<in
WAWA1.MM1AI	NXNW RETAIL LLC \/ CANNABIS AND GLASS	603351098	25101 E APPLEWAY AVE	LIBERTY LAKE	WA	99019	85131	8357453	5683
WAWA1.MMB8	MR. BILLS OF BUCKLEY	603405237	29393 STATE ROUTE 410 E STE D	BUCKLEY	WA	98321	96183	7994432	4979
WAWA1.MMFX	HIGHER LEAF MARIJUANA BOUTIQUE	603414213	12525 WILLOWS RD NE STE 10	KIRKLAND	WA	98034	101674	7930156	4542

In [25]:

```
options(repr.plot.height=5, repr.plot.width=5)
sales1 <- ggplot(dispensary_data, aes(x=total_transactions))+
    geom_histogram( fill = "#00AFBB") +
    theme_classic() +
    ggtitle(label = "by Total Sales Transactions",
            subtitle = "Transactions in 100 Thousands") +
    ylab("Count of Dispensaries       ") + xlab("") +


    theme(axis.text.x=element_blank(),
        axis.ticks.x=element_blank(),
        plot.title = element_text(size=16, hjust = 0.5),
        plot.subtitle = element_text(size=12, hjust = 0.5),
        axis.title = element_text(size=14),
        axis.line.x = element_blank(), axis.text.y = element_text(size=14))

sales2 <- ggplot(dispensary_data,aes(y=total_transactions))+
    geom_boxplot(height = .05) +
    ylab("Count of Transactions") + xlab(" ") +
    theme_classic() +
    coord_flip() +
    scale_y_continuous(labels = unit_format(unit = "K", scale = 1e-3)) +
    theme(axis.text.y=element_blank(), axis.ticks.y=element_blank(), axis.title.y = element_text(size=35), axis.title.x = el

grid.arrange(sales1, sales2, ncol = 1, heights = c(1, .25))
```

```
Warning message:
“Ignoring unknown parameters: height”
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



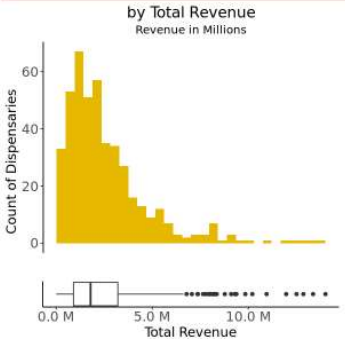
In [27]:

```
options(repr.plot.height=5, repr.plot.width=5)

prod1 <- ggplot(dispensary_data, aes(x=product_count))+
    geom_histogram(fill="#FC4E07") +
theme_classic() +
    ggtitle(label = "by Distinct Products Sold",
            subtitle = "Product Counts in Thousands") +
    ylab("Count of Dispensaries       ") + xlab("") +


    theme(axis.text.x=element_blank(),
        axis.ticks.x=element_blank(),
        plot.title = element_text(size=16, hjust = 0.5),
        plot.subtitle = element_text(size=12, hjust = 0.5),
        axis.title = element_text(size=14),
        axis.line.x = element_blank(), axis.text.y = element_text(size=14))

prod2 <- ggplot(dispensary_data,aes(y=product_count))+
    geom_boxplot() +
    ylab("Count of Distinct Products") + xlab(" ") +
    theme_classic() +
    coord_flip() +
    scale_y_continuous(labels = unit_format(unit = "K", scale = 1e-3)) +
    theme(axis.text.y=element_blank(), axis.ticks.y=element_blank(), axis.title.y = element_text(size=35), axis.title.x = el

grid.arrange(prod1, prod2, ncol = 1, heights = c(1, .25))
```

In [26]:

```
options(repr.plot.height=5, repr.plot.width=5)

rev1 <- ggplot(dispensary_data, aes(x=total_sales))+
    geom_histogram(fill = "#E7B800") +
theme_classic() +
    ggtitle(label = "by Total Revenue",
            subtitle = "Revenue in Millions") +
    ylab("Count of Dispensaries       ") + xlab("") +


    theme(axis.text.x=element_blank(),
        axis.ticks.x=element_blank(),
        plot.title = element_text(size=16, hjust = 0.5),
        plot.subtitle = element_text(size=12, hjust = 0.5),
        axis.title = element_text(size=14),
        axis.line.x = element_blank(), axis.text.y = element_text(size=14))

rev2 <- ggplot(dispensary_data,aes(y=total_sales))+
    geom_boxplot(show.legend = TRUE) +
    ylab("Total Revenue") + xlab(" ") +
    theme_classic() +
    coord_flip() +
    scale_y_continuous(labels = unit_format(unit = "M", scale = 1e-6)) +
    theme(axis.text.y=element_blank(), axis.ticks.y=element_blank(), axis.title.y = element_text(size=35), axis.title.x = el

grid.arrange(rev1, rev2, ncol = 1, heights = c(1, .25))
```

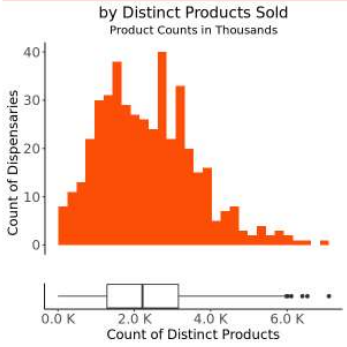
```
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



In [27]:

```
options(repr.plot.height=5, repr.plot.width=5)

prod1 <- ggplot(dispensary_data, aes(x=product_count))+
    geom_histogram(fill="#FC4E07") +
theme_classic() +
    ggtitle(label = "by Distinct Products Sold",
            subtitle = "Product Counts in Thousands") +
    ylab("Count of Dispensaries       ") + xlab("") +

                                        .

    theme(axis.text.x=element_blank(),
        axis.ticks.x=element_blank(),
        plot.title = element_text(size=16, hjust = 0.5),
        plot.subtitle = element_text(size=12, hjust = 0.5),
        axis.title = element_text(size=14),
        axis.line.x = element_blank(), axis.text.y = element_text(size=14))

prod2 <- ggplot(dispensary_data,aes(y=product_count))+
    geom_boxplot() +
    ylab("Count of Distinct Products") + xlab(" ") +
    theme_classic() +
    coord_flip() +
    scale_y_continuous(labels = unit_format(unit = "K", scale = 1e-3)) +
    theme(axis.text.y=element_blank(), axis.ticks.y=element_blank(), axis.title.y = element_text(size=35), axis.title.x = el

grid.arrange(prod1, prod2, ncol = 1, heights = c(1, .25))
```

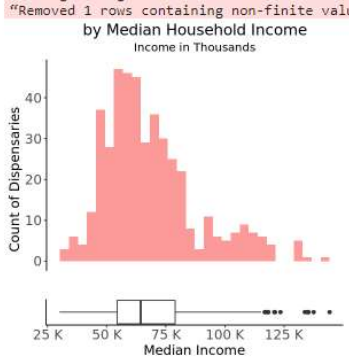
```
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



In [28]:

```
options(repr.plot.height=5, repr.plot.width=5)

med1 <- ggplot(dispensary_data, aes(x=median_income_by_zip))+
    geom_histogram(fill='#FB9A99') +
    theme_classic() +
    ggtitle(label = "by Median Household Income",
            subtitle = "Income in Thousands") +
    ylab("Count of Dispensaries       ") + xlab("") +


    theme(axis.text.x=element_blank(),
        axis.ticks.x=element_blank(),
        plot.title = element_text(size=16, hjust = 0.5),
        plot.subtitle = element_text(size=12, hjust = 0.5),
        axis.title = element_text(size=14),
        axis.line.x = element_blank(), axis.text.y = element_text(size=14))

med2 <- ggplot(dispensary_data,aes(y=median_income_by_zip))+
    geom_boxplot() +
    ylab("Median Income") + xlab(" ") +
    theme_classic() +
    coord_flip() +
    scale_y_continuous(labels = unit_format(unit = "K", scale = 1e-3)) +
    theme(axis.text.y=element_blank(), axis.ticks.y=element_blank(), axis.title.y = element_text(size=35), axis.title.x = el

grid.arrange(med1, med2, ncol = 1, heights = c(1, .25))
```

```
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.

Warning message:
"Removed 1 rows containing non-finite values (stat_bin)."
Warning message:
"Removed 1 rows containing non-finite values (stat_boxplot)."
```



In [29]:

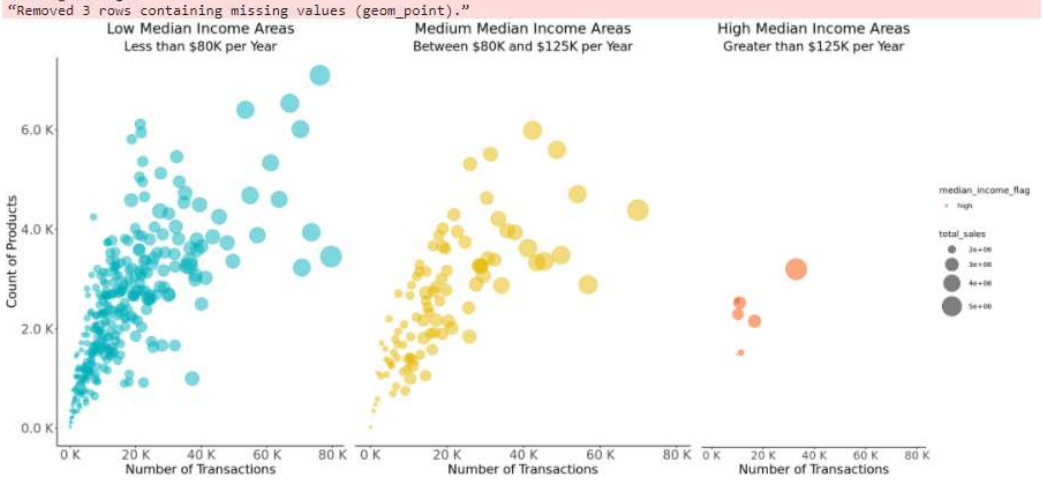
```
dispensary_data_low <- dispensary_data %>% filter(median_income_flag == 'low')
dispensary_data_med <- dispensary_data %>% filter(median_income_flag == 'medium')
dispensary_data_high <- dispensary_data %>% filter(median_income_flag == 'high')
```

In [32]:

```
options(repr.plot.height=8, repr.plot.width=18)
p1 = ggplot(dispensary_data_low, aes(x=total_transactions, y=product_count ))+
geom_point(aes(color = median_income_flag, size = total_sales), alpha = 0.5) +
    ggtitle(label = "Low Median Income Areas",
            subtitle = "Less than $80K per Year") +
    ylab("Count of Products") + xlab("Number of Transactions") +
    scale_color_manual(values = c("#00AFBB")) +
    scale_size(range = c(0.5, 12)) +
    theme_classic() +
    scale_y_continuous(labels = unit_format(unit = "K", scale = 1e-3), limits = c(0, max(dispensary_data$product_count))) +
    scale_x_continuous(labels = unit_format(unit = "K", scale = 1e-4), limits = c(0, dispensary_data$total_transactions)) +

    theme(
        plot.title = element_text(size=18, hjust = 0.5),
        plot.subtitle = element_text(size=16, hjust = 0.5),
        axis.title = element_text(size=16),
        axis.text.y = element_text(size=16),
        axis.text.x = element_text(size=16),
        legend.position = "none")

p2 = ggplot(dispensary_data_med, aes(x=total_transactions, y=product_count ))+
geom_point(aes(color = median_income_flag, size = total_sales), alpha = 0.5) +
    ggtitle(label = "Medium Median Income Areas",
            subtitle = "Between $80K and $125K per Year") +
    xlab("Number of Transactions") +
    scale_color_manual(values = c("#E7B800")) +
    scale_size(range = c(0.5, 12)) +
    theme_classic() +
    scale_y_continuous(labels = unit_format(unit = "K", scale = 1e-3), limits = c(0, max(dispensary_data$product_count))) +
    scale_x_continuous(labels = unit_format(unit = "K", scale = 1e-4), limits = c(0, dispensary_data$total_transactions)) +
    theme(
        plot.title = element_text(size=18, hjust = 0.5),
        plot.subtitle = element_text(size=16, hjust = 0.5),
        axis.title = element_text(size=16),
        axis.title.y=element_blank(),
        axis.line.y=element_blank(),
        axis.text.y=element_blank(),
        axis.ticks.y=element_blank(),
        axis.text.x = element_text(size=16),
        legend.position = "none")
```

```
p3 = ggplot(dispensary_data_high, aes(x=total_transactions, y=product_count ))+
geom_point(aes(color = median_income_flag, size = total_sales), alpha = 0.5) +
    ggtitle(label = "High Median Income Areas",
            subtitle = "Greater than $125K per Year") +
    xlab("Number of Transactions") +
    scale_color_manual(values = c("#FC4E07")) +
    scale_size(range = c(0.5, 12)) +
    theme_classic() +
    scale_y_continuous(labels = unit_format(unit = "K", scale = 1e-3), limits = c(0, max(dispensary_data$product_count))) +
    scale_x_continuous(labels = unit_format(unit = "K", scale = 1e-4), limits = c(0, dispensary_data$total_transactions)) +
    theme(
        plot.title = element_text(size=18, hjust = 0.5),
        plot.subtitle = element_text(size=16, hjust = 0.5),
        axis.title = element_text(size=16),
        axis.title.y=element_blank(),
        axis.line.y=element_blank(),
        axis.text.y=element_blank(),
        axis.ticks.y=element_blank(),
        axis.text.x = element_text(size=14),
        #Legend.position = "none"
    )

grid.arrange(p1, p2, p3, nrow = 1)
```

```
Warning message:
"Removed 3 rows containing missing values (geom_point)."
```



APPENDIX D

Market Basket Analyis R Code

Load R Packages

In [ ]:
```r
# Install Packages/Load Libraries
# install.packages("ggdist")
# install.packages("dlo")
# install.packages("RColorBrewer")
install.packages("ggplot2")

library(tidyverse)
library(readxl)
library(knitr)
library(ggplot2)
library(lubridate)
library(arules)
library(arulesViz)
library(bigrquery) # used for querying BigQuery
library(plyr)
library(plotly)
library(RColorBrewer)
```

Load Data

The following data is loaded from a view created Google Cloud BigQuery.

In [2]:
```r
# References
# https://cloud.google.com/architecture/data-science-with-r-on-gcp-eda
```

In [3]:
```r
# Authenticate bigrquery using out-of-band authentication - its a type of two-factor identification that Google Cloud uses
bq_auth(use_oob = True)

# Set a variable to the name of the Google Cloud project I created - the ID can be used by APIs
PROJECT_ID <- "woven-sensor-314415"

# Set a variable to the name of the Cloud Storage bucket I created...this is the Google Cloud data repository
BUCKET_NAME <- "washington_data"
```

In [4]:
```r
# Assign a query to a variable
sql_query <- "
    SELECT * FROM woven-sensor-314415.Washington_State_Data.2020_sales_for_top_three_dispensaries
    WHERE dispensary_id = 'WAWA1.MM4V'
"
```

In [5]:
```r
# Run a query against one of the tables in my Cloud Storage bucket

sales_df <- bq_table_download(
    bq_project_query(
        PROJECT_ID,
        query=sql_query
    )
)
```

In [6]:
```
# Show the sample of the data

head(sales_df)
```

A tibble: 6 × 10

dispensary_id	dispensary_name	transaction_id	sales_date	unit_of_measure	prod_name	prod_id	unit_price	price_total
<chr>	<chr>	<chr>	<dttm>	<chr>	<chr>	<chr>	<dbl>	<dbl>
WAWA1.MM4V	ZIPS CANNABIS	WAR427634.SA34Z5J1	2020-12-01 16:00:00	ea	Flower - MAC- 1g [MAC #10 - Usable Marijuana - 1g]	WAR427634.SI48CPAF	6.78	6.78
WAWA1.MM4V	ZIPS CANNABIS	WAR427634.SA34Z5PA	2020-12-01 16:00:00	ea	ATF 4g BB	WAR427634.SI48CPRW	16.94	16.94
WAWA1.MM4V	ZIPS CANNABIS	WAR427634.SA34Z5PA	2020-12-01 16:00:00	ea	ATF 4g BB	WAR427634.SI48CPRX	16.94	16.94
WAWA1.MM4V	ZIPS CANNABIS	WAR427634.SA34Z5PE	2020-12-01 16:00:00	ea	Usable Marijuana - Usable Marijuana 1.000g	WAR427634.SI48CPSH	5.69	5.69

In [7]:
```
# List the table fileds and their data types

summary(sales_df)
```

```
dispensary_id      dispensary_name    transaction_id
Length:53963       Length:53963       Length:53963
Class :character   Class :character   Class :character
Mode  :character   Mode  :character   Mode  :character


  sales_date                unit_of_measure     prod_name
Min.   :2020-12-01 16:00:00   Length:53963       Length:53963
1st Qu.:2020-12-09 16:00:00   Class :character   Class :character
Median :2020-12-17 16:00:00   Mode  :character   Mode  :character
Mean   :2020-12-16 13:45:59
3rd Qu.:2020-12-23 16:00:00
Max.   :2020-12-30 16:00:00
  prod_id            unit_price      price_total       prod_type
Length:53963       Min.   :  1.22   Min.   :-127.03   Length:53963
Class :character   1st Qu.:  5.81   1st Qu.:   5.76   Class :character
Mode  :character   Median :  7.62   Median :   7.62   Mode  :character
                   Mean   : 15.33   Mean   :  15.28
                   3rd Qu.: 18.97   3rd Qu.:  18.97
                   Max.   :208.67   Max.   : 208.67
```

In [8]:
```
# Create a table of transactions by product name

sql_query <- "
    SELECT transaction_id, prod_name FROM woven-sensor-314415.Washington_State_Data.2020_sales_for_top_three_dispensaries
    WHERE dispensary_id = 'WAWA1.MM4V'
"
```

In [9]:
```
# Run a query against one of the tables in my Cloud Storage bucket

sales_df <- bq_table_download(
    bq_project_query(
        PROJECT_ID,
        query=sql_query
    )
)

# convert tibble to df
sales_df <- as.data.frame(sales_df)
```

In [10]:
```
head(sales_df)
```

A data.frame: 6 × 2

	transaction_id	prod_name
	<chr>	<chr>
1	WAR427634.SA34Z5J1	Flower - MAC- 1g [MAC #10 - Usable Marijuana - 1g]
2	WAR427634.SA34Z5PA	ATF 4g BB
3	WAR427634.SA34Z5PA	ATF 4g BB
4	WAR427634.SA34Z5PE	Usable Marijuana - Usable Marijuana 1.000g
5	WAR427634.SA34Z5PE	Usable Marijuana - Usable Marijuana 1.000g
6	WAR427634.SA34Z60R	(Twisted by Dabulous - Twisted Pre-Roll - 1 g)

Transform the Data into a Form Suitable for MBA

Using the aRules library to conduct MBA. The aRules library reuires that data be converted into a transaction matrix format. The first steps is to "reshape" the orginal "tall" dataframe of transations into a "wide" format

In [11]:
```
# Next convert the dataframe to a Matrix.

sales_df$const = TRUE

# Remove duplicates
dim(sales_df) #1,000 x 3
sales_df <- unique(sales_df)
dim(sales_df) #979 x 3

class(sales_df)

# Replace charater transactionID with a system generated numeric ID

#orders <- orders %>% dplyr::mutate(transactionID = row_number())
dim(sales_df)
head(sales_df)
```

53963 · 3

44323 · 3

'data.frame'

44323 · 3

A data.frame: 6 × 3

	transaction_id	prod_name	const
	<chr>	<chr>	<lgl>
1	WAR427634.SA34Z5J1	Flower - MAC- 1g [MAC #10 - Usable Marijuana - 1g]	TRUE
2	WAR427634.SA34Z5PA	ATF 4g BB	TRUE
4	WAR427634.SA34Z5PE	Usable Marijuana - Usable Marijuana 1.000g	TRUE
6	WAR427634.SA34Z60R	(Twisted by Dabulous - Twisted Pre-Roll - 1 g)	TRUE
7	WAR427634.SA34Z60S	Wax - God's Gift - 01.0 g [God's Gift - 1g - Marijuana Extract for Inhalation - 01 g]	TRUE
8	WAR427634.SA34Z60S	Usable Marijuana - Usable Marijuana 1.000g	TRUE

In [12]:
```
# Here we "pivot" the matrix from a "tall" matrix to a "wide" matrix. We are creating a column for each product.
# This ends up generating a parse matrix

sales_mat_prep <- reshape(data = sales_df,
                          idvar = "transaction_id",
                            timevar = "prod_name",
                            direction = "wide")

# Drop the transaction ID after the pivot
sales_matrix <- as.matrix(sales_mat_prep[,-1])

# The pivoted matrix is sparse so populate the missing values with "FALSE"
sales_matrix[is.na(sales_matrix)] <- FALSE

#head(low_sales_matrix)
head(sales_matrix)
```

A matrix: 6 × 1696 of type lg

	const.Flower - MAC- 1g [MAC #10 - Usable Marijuana - 1g]	const.ATF 4g BB	const.Usable Marijuana - Usable Marijuana 1.000g	const.(Twisted by Dabulous - Twisted Pre-Roll - 1 g)	const.Wax - God's Gift - 01.0 g [God's Gift - 1g - Marijuana Extract for Inhalation - 01 g]	const.Usable Marijuana - Usable Marijuana 7.000g	const.Joint Pack	const.Flower	const.WHJ FL - Grease Monkey - 28.0 Gram	const.Wax - Fruity Pebbles - 01.0g [Fruity Pebble - 1g - Marijuana Extract for Inhalation - 01 g]	...	const.FL - FMK - Unrivaled OG - 03.5 gram Jar	const.THC-Suga
1	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	...	FALSE	
2	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	...	FALSE	
4	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	...	FALSE	
6	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	...	FALSE	
7	FALSE	FALSE	TRUE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	...	FALSE	
10	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	TRUE	TRUE	FALSE	FALSE	...	FALSE	

In [13]:
```
# Remove the auto-generated word "const" form the column name.

colnames(sales_matrix) <- gsub(x=colnames(sales_matrix),
                               pattern="const\\.", replacement="")
```

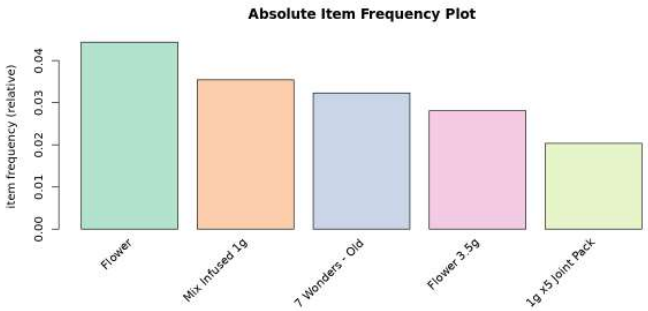
In [14]:
```
# Transform the matrix into a "transactions" table.  The Transaction table is a special object type reuired by the MBA algor

sales_transactions <- as(sales_matrix,"transactions")
#class(order_trans2)
```

In [15]:
```
summary(sales_transactions)

# Verify the number of oringal transactions by calculating them from the summary info
# Transactions as itemMatrix in sparse format with
# 15504 rows (elements/itemsets/transactions) and
# 1295 columns (items) and a density of 0.00124357

low_num_of_transactions <- 15504 * 1295 * 0.00124357
low_num_of_transactions
```

```
transactions as itemMatrix in sparse format with
 25941 rows (elements/itemsets/transactions) and
 1696 columns (items) and a density of 0.001007434

most frequent items:
          Flower   Mix Infused 1g  7 Wonders - Old      Flower 3.5g
            1151              919              838              729
1g x5 Joint Pack          (Other)
             528            40158

element (itemset/transaction) length distribution:
sizes
    1     2     3     4     5     6     7     8     9    10    11    12    13
16078  5493  2211  1085   587   261   121    60    25    11     2     2     1
   14    15    17    18
    1     1     1     1

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.000   1.000   1.000   1.709   2.000  18.000

includes extended item information - examples:
                                                labels
1 Flower - MAC- 1g  [ MAC #10 - Usable Marijuana - 1g ]
2                                          ATF 4g BB
3            Usable Marijuana - Usable Marijuana 1.000g

includes extended transaction information - examples:
  transactionID
1             1
2             2
3             4
24968.0005176
```

Modeling

In [16]:
```
# Show the Top 20 dispensary products

options(repr.plot.height=5, repr.plot.width=10)
itemFrequencyPlot(sales_transactions, topN=5, type="relative", col=brewer.pal(8,'Pastel2'), main="Absolute Item Frequency Pl
```



In [17]:
```
# Assign support and confidence thresholds.  Typically a variety of values must be tried to generate a repsentative list of
# If the values are too low then you may get a large number of low value rules.
association_rules <- apriori(sales_transactions, parameter = list(supp=0.004, conf=0.4, minlen=2))

# Write out the association rules to a csv file for review and ease of reading
rule_output <- as(association_rules, "data.frame")
write_csv(rule_output, "low_rule_output.csv")
```

```
Apriori

Parameter specification:
 confidence minval smax arem  aval originalSupport maxtime support minlen
        0.4    0.1    1 none FALSE            TRUE       5   0.004      2
 maxlen target  ext
     10  rules TRUE

Algorithmic control:
 filter tree heap memopt load sort verbose
    0.1 TRUE TRUE  FALSE TRUE    2    TRUE

Absolute minimum support count: 103

set item appearances ...[0 item(s)] done [0.00s].
set transactions ...[1696 item(s), 25941 transaction(s)] done [0.01s].
sorting and recoding items ... [87 item(s)] done [0.00s].
creating transaction tree ... done [0.00s].
checking subsets of size 1 2 3 done [0.00s].
writing ... [21 rule(s)] done [0.00s].
creating S4 object  ... done [0.00s].
```

In [18]:

```
# List the number of rules generated

summary(association_rules)
```

```
set of 21 rules

rule length distribution (lhs + rhs):sizes
 2  3
15  6

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  2.000   2.000   2.000   2.286   3.000   3.000

summary of quality measures:
    support           confidence        coverage             lift
 Min.   :0.004048   Min.   :0.4059   Min.   :0.004742   Min.   :23.78
 1st Qu.:0.004163   1st Qu.:0.5423   1st Qu.:0.007132   1st Qu.:32.23
 Median :0.004433   Median :0.5750   Median :0.007710   Median :51.53
 Mean   :0.004773   Mean   :0.5882   Mean   :0.008451   Mean   :47.78
 3rd Qu.:0.004780   3rd Qu.:0.6263   3rd Qu.:0.011333   3rd Qu.:61.23
 Max.   :0.007132   Max.   :0.8537   Max.   :0.011680   Max.   :75.32
     count
 Min.   :105.0
 1st Qu.:108.0
 Median :115.0
 Mean   :123.8
 3rd Qu.:124.0
 Max.   :185.0

mining info:
              data ntransactions support confidence
 sales_transactions         25941   0.004        0.4
```

In [19]:

```
#Review what the rules look like

inspect(association_rules[1:10])
```

```
     lhs                                                                                         rhs
support confidence    coverage     lift count
[1]  {Wax - Shark's Breath - 01.0 g}                                                            => {Wax - Blueberr
y Muffin - 1.0 g  [ Blueberry Muffin - 1g - Marijuana Extract for Inhalation - 01 g ]} 0.004124745  0.5431472 0.007594156 47.
44034   107
[2]  {Wax - Shark's Breath - 01.0 g}                                                            => {Wax - Rainbow
Runtz - 01.0 g  [ Rainbow Runtz - 1g - Marijuana Extract for Inhalation - 01 g ]}      0.004201843  0.5532995 0.007594156 39.
00310   109
[3]  {Wax - Lemon OG - 01.0 g  [ Lemon OG - 1g - Marijuana Extract for Inhalation - 01 g ]}              => {Wax - Pink Pan
ther - 01.0 g  [ Pink Panther - 1g - Marijuana Extract for Inhalation - 01 g ]}       0.004201843  0.5422886 0.007748352 70.
33754   109
[4]  {Wax - Pink Panther - 01.0 g  [ Pink Panther - 1g - Marijuana Extract for Inhalation - 01 g ]}        => {Wax - Lemon OG
- 01.0 g  [ Lemon OG - 1g - Marijuana Extract for Inhalation - 01 g ]}               0.004201843  0.5450000 0.007709803 70.3
3754   109
[5]  {Wax - Lemon OG - 01.0 g  [ Lemon OG - 1g - Marijuana Extract for Inhalation - 01 g ]}              => {Wax - Blueberr
y Muffin - 1.0 g  [ Blueberry Muffin - 1g - Marijuana Extract for Inhalation - 01 g ]} 0.004780078  0.6169154 0.007748352 53.
88351   124
[6]  {Wax - Blueberry Muffin - 1.0 g  [ Blueberry Muffin - 1g - Marijuana Extract for Inhalation - 01 g ]} => {Wax - Lemon OG
- 01.0 g  [ Lemon OG - 1g - Marijuana Extract for Inhalation - 01 g ]}               0.004780078  0.4175084 0.011449057 53.8
8351   124
[7]  {Wax - Pink Panther - 01.0 g  [ Pink Panther - 1g - Marijuana Extract for Inhalation - 01 g ]}        => {Wax - Blueberr
y Muffin - 1.0 g  [ Blueberry Muffin - 1g - Marijuana Extract for Inhalation - 01 g ]} 0.004548784  0.5900000 0.007709803 51.
53263   118
[8]  {Wax - Pink Panther - 01.0 g  [ Pink Panther - 1g - Marijuana Extract for Inhalation - 01 g ]}        => {Wax - Rainbow
Runtz - 01.0 g  [ Rainbow Runtz - 1g - Marijuana Extract for Inhalation - 01 g ]}      0.004433137  0.5750000 0.007709803 40.
53281   115
[9]  {Wax - GMO - 01.0 g  [ GMO - 1g - Marijuana Extract for Inhalation - 01 g ]}                        => {Wax - God's Gi
ft - 01.0 g  [ God's Gift - 1g - Marijuana Extract for Inhalation - 01 g ]}          0.004780078  0.6262626 0.007632705 32.
23389   124
[10] {Wax - Starfighter - 01.0 g  [ Starfighter - 1g - Marijuana Extract for Inhalation - 01 g ]}          => {Wax - Chernoby
l - 01.0 g  [ Chernobyl - 1g - Marijuana Extract for Inhalation - 01 g ]}            0.007131568  0.6292517 0.011333410 53.
87267   185
```

In [20]:

```
#Plot Rules
options(repr.plot.height=5, repr.plot.width=5)

plot(association_rules, engine = 'plotly')
```