**Part I: Research Question**

1. Describe the purpose of this data mining report by doing the following:
2. What are the most common prescriptions and what other prescriptions are most often prescribed along with these popular prescriptions?
3. By identifying popular prescriptions and the common groupings of prescriptions, we can provide a recommendation for how to organize pharmacy storage in order to maximize efficiency in prescription filling and minimize patient wait times.

**Part II: Market Basket Justification**

1. Explain the reasons for using market basket analysis by doing the following:
2. Market basket analysis is an association analysis that is based upon probability. The analysis produces three key metrics: support, confidence, and lift (McColl, 2018). Support is the frequency of a particular outcome, whether it be a single item or a set of items which is called a rule. Confidence is the probability that a rule’s antecedent and consequent, the left- and right-hand value in the rule table respectively, will be present in a transaction. Lift is a metric that evaluates whether a rule’s antecedent increases the probability of the rule’s consequent being present in the transaction by dividing the support of the rule by the product of the antecedent’s and consequent’s support. Once these metrics are calculated for each combination of items in a transaction, the data can be filtered and sorted to show items with high lift and support, which are said to be common association rules which can be used for recommendation or organization.
3. In the unprocessed data read directly from the provided csv file, the data is organized with each transaction represented as a row. At index position zero, we see the first transaction is empty, containing no prescriptions. In contrast, the transaction at index position one depicts a patient prescribed 20 medications, filling every column of the data frame.
4. Market basket analysis assumes that the data is stored in a sparse matrix of counts with rows representing transactions and columns representing products. In this counts matrix, we store the quantity of each product purchased in each transaction. This counts matrix is then encoded such that any positive value is represented as a one and all other values are assigned zero (Jihargifari, 2020).

**Part III: Data Preparation and Analysis**

1. Prepare and perform market basket analysis by doing the following:
2. In order to transform our data into the encoded counts matrix that the apriori algorithm requires, we will create a new data frame with transactions as rows and possible prescriptions as columns. First, we will filter out any transactions where a patient received zero or one prescription as associations can only be made when more than one item was purchased or obtained in a transaction.

data = data[data.isna().sum(axis=1) < 19].reset\_index(drop=True)

Before we can create our new data frame, we need to create the column labels which are the prescriptions present in this data set.

prescriptions = []

for i, col in enumerate(data.columns):

for j in range(len(data[col])):

val = str(data.iloc[j][i])

if val != 'nan' and val not in prescriptions:

prescriptions.append(val)

With the column names recorded, we can create an empty dataframe with column names that are the prescriptions in the data set and begin creating an encoded counts matrix by comparing the values in each transaction to the column names and updating the values in the data frame accordingly.

data\_counts = pd.DataFrame(index=range(data.shape[0]), columns=prescriptions)

rows, cols = data.shape

rows = range(rows)

cols = range(cols)

for row in rows:

for col in cols:

val = data.iloc[row, col]

for cnt\_col in data\_counts.columns:

if cnt\_col == val:

data\_counts[cnt\_col][row] = 1

data\_counts.fillna(0, inplace=True)

The resulting data frame is saved as a csv which is provided as the attachment OFM3\_basket\_clean.csv.

1. Input:

freq\_items = apriori(data\_counts, min\_support=0.03, use\_colnames=True).sort\_values('support', ascending=False).reset\_index(drop=True)

freq\_items['length'] = freq\_items['itemsets'].apply(lambda x: len(x))

ass\_rule = association\_rules(freq\_items, metric='lift', min\_threshold=1).sort\_values('lift', ascending=False).reset\_index(drop=True)

Output:

Table

Description automatically generated

Table

Description automatically generated

Table

Description automatically generated

1. The values for the support, confidence, and lift of the rules with lift greater than one are visible in the screenshots above. The values of these metrics for the remaining rules are provided below.

Table

Description automatically generated

1. The top three association rules are shown in the screenshot below. This screenshot shows six association rules, but every rule at an odd index position in the screenshot is an inverse of the preceding rule, with the antecedent and consequent swapped. By evaluating the values of support, confidence, and lift for these, we see that only the confidence is different between inverted rules. We will focus on the first of each of these pairs of rules in the following analysis; these are the rules at index position zero, two, and four. The support metric tells us that the rule at index zero occurs in 5% of the transactions available in the data set, the rule at index position two occurs in 3% of the transactions, and the rule at index position four also occurs in 3% of transactions. The confidence metric of rule zero tells us that 41% of transactions where a patient is prescribed lisinopril, they are also prescribed carvedilol. Rule one contains the same prescriptions as rule zero, but with the antecedent and consequent swapped; this changes the support metric, which now tells us that 23% of patients prescribed carvedilol were also prescribed lisinopril. Rule two shows us that 10% of patients prescribed abilify are also prescribed metformin, while rule three show us that 47% of patients prescribed metformin are prescribed abilify as well. Rule four shows us that 25% of patients prescribed metoprolol are also prescribed atorvastatin, while rule five show us that 19% of patients prescribed atorvastatin are also prescribed metoprolol. Finally, the lift metric shows us that a patient prescribed either lisinopril or carvedilol is 88% more likely to be prescribed the other than if they weren’t prescribed that medicine, a patient prescribed either abilify or metformin is 58% more likely to be prescribed the other than if they weren’t prescribed that medicine, and a patient prescribed either metoprolol or atorvastatin is 56% more likely to be prescribed the other than if they weren’t prescribed that medicine.

Table

Description automatically generated

**Part IV: Data Summary and Implications**

1. Summarize your data analysis by doing the following:
2. Higher values of support indicate that a rule is more likely to occur than a rule with low support. Our best rules exhibit a support of approximately 5% or below, meaning they are fairly uncommon. High confidence values imply higher probabilities that a rule’s consequents will be prescribed if the antecedent was prescribed and are significant if they are higher than the support of the consequent on its own. In our top rules, all of the consequents are higher than the consequent’s support, visible in the screenshot in Part C.4. Finally, lift is significant when it is above one, as a value below one implies a consequent is less likely to be prescribed if the antecedent was prescribed. Our top three rules show strong significance in their lift metrics with values of 1.88, 1.58, and 1.56 representing in an 88%, 58%, and 56% increase in likelihood of the consequent in each rule, respectively.
3. This analysis provides the most practical information to organization of the pharmacy storage. Prescriptions with high support are more likely to be prescribed and should be made more accessible while rules with high lift and confidence tell us which prescriptions should be placed near one another. This allows for an increase in efficiency that leads to shorter wait times for prescriptions, more prescriptions filled per hour, and higher customer satisfaction as a result.
4. Based on this analysis, we recommend the pharmacy be organized as follows:

* Abilify, Amphetamine Salt Combo XR, Carvedilol, Diazepam, Glyburide, Atorvastatin, Losartan, Lisinopril, Metoprolol, Doxycycline Hyclate, Citalopram, and Dextroamphetamine XR should be placed in the most accessible position in the pharmacy area for pharmacists to access.
* Prescriptions in rules with high lift should be placed near one another.

To better visualize the second point, a flow chart in provided showing a possible organization. The prescriptions shown as diamonds should be placed in the most accessible location with connected prescriptions placed near one another.

Diagram

Description automatically generated

**Part V: Attachments**

1. <https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=dec7f2b9-dc67-478a-b3d4-add00136e47a#>
2. The following source was used to guide the implementation of the Apriori algorithm and as a guide for the process of market basket analysis.

Jihargifari. (2020, August 3). How to perform Market Basket Analysis in python. Medium. Retrieved October 29, 2021, from https://medium.com/@jihargifari/how-to-perform-market-basket-analysis-in-python-bd00b745b106.

1. Acknowledge sources, using in-text citations and references, for content that is quoted, paraphrased, or summarized.

McColl, L. (2018, May 31). Market basket analysis: Understanding customer behaviour. Select Statistical Consultants. Retrieved October 29, 2021, from https://select-statistics.co.uk/blog/market-basket-analysis-understanding-customer-behaviour/.