CSCI 185: Final Project

Project Title: Bakery Bites: What items do Bakery Customers like to Buy Together?

Group Members: Camelia Siadat, Gloria Zhao, and Samantha Lee

Dataset: French Bakery Daily Sales

For this project, we have chosen to use a dataset from Kaggle titled *"Bakery sales.csv",* that includes data taken from a French bakery's transactions. The observed period of transactions is from January 1st, 2021 to September 30th, 2022. The dataset includes the following features: date, time, ticket_number, article (name of the product purchased in French), Quantity (of item purchased), and unit_price (in Euros).

Project Objective: The goal of our project is to use the Apriori and Association Rule Mining algorithms to generate association rules for the bakery's transactions. In our implementation, we have generated association rules through grouping the data via the feature ticket_number, as well as generated association rules through grouping each row of the table (which could be considered its own transaction in this second case) by time period (morning, noon, and evening).

Methodology: Library Importing, Implementation, Results, and Analysis

- **Importing Libraries**
  - In this initial step of our data analysis process, we lay the groundwork by importing essential libraries and loading our dataset. We imported critical libraries and modules tailored to our needs, including pandas and Mlxtend. These libraries provide comprehensive support for DataFrame creation, data manipulation, and the process of transaction encoding. One of our initial steps was to load the dataset, "*Bakery sales.csv*," utilizing pandas' read_csv() function, which facilitated the creation of a DataFrame named bakeryData.

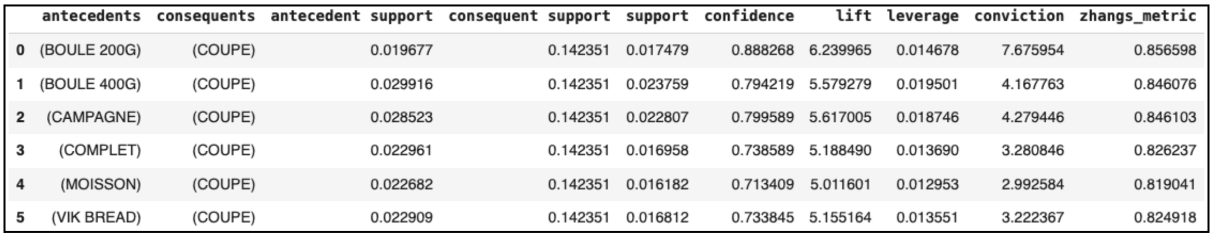
- **Implementation**
  - To begin our code implementation of our specified task, it was necessary to first load the "*Bakery sales.csv*" file into the bakeryData variable using pandas'

pd.read_csv command. Realizing that rows containing the same ticket_number implied multiple products being bought within the same transaction, we decided to group each row of the bakeryData DataFrame by the ticket_number feature. At this point, it became necessary to drop the other columns in this dataframe, since they would not affect our Association Rule generation. In order to group each row of bakeryData by ticket_number, it was necessary to create a dictionary, ticketItemGroupings. In this dictionary, each key was a ticket number, and each value was a list of the associated products purchased under that particular ticket number. A new DataFrame, called *result*, was then created in order to store each item in this dictionary. This process involved first iterating through each key in ticketItemGroupings, and storing the key and value as *ticketNum* and *items* respectively. Each row in the *result* DataFrame was generated using the pd.Series command, which generated two columns, "Ticket Number" and "Items Purchased," whose values were *ticketNum* and *items* respectively. Using the pd.concat command, each row, transposed for the purposes of generating a DataFrame, was subsequently added to the *result* DataFrame.

After combining the items purchased based on ticket number, we converged the items purchased to a list of lists and performed one hot encoding on this list. We applied the Apriori algorithm and performed Association Rule Mining in two different examples. For the first example, we set the minimum support to one percent and the minimum confidence to sixty percent. For the second example, we set the minimum support to one and a half percent and the minimum confidence to fifty percent. We chose these values because we believed they gave us a variety of results of association rules. To show some further analysis of our association rules results, we created two different graphs. First, we plotted a scatter plot to show the relationship between support, confidence, and lift of each association rule. In the scatter plot, support is represented as the x-value, confidence is represented as the y-value, and lift is represented by the varying colors of the points on the graph. Second, we made a network graph to demonstrate the relationship between the antecedents and consequents along with the confidences between antecedent and consequent shown on the edges of the graph.

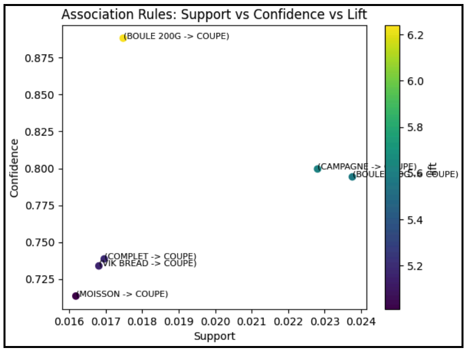
To delve deeper into consumer buying behaviors across different times of day, we segmented our dataset into distinct time periods: Morning (from opening until 11:59 AM), Noon (from 12:00 PM to 3:59 PM), and Evening (from 4:00 PM until closing). We applied Association Rule Mining and the Apriori algorithm to each segment, setting a minimum support of 0.001 to generate a broader range of itemsets for comparison. This approach allows us to discern trends in purchasing behavior specific to each time period. Additionally, to understand the relationships between different products, we conducted a link analysis. In this analysis, we constructed a network diagram where nodes represent products, and edges signify the frequency with which products are purchased together within the same order. This method helps identify commonly co-purchased items. We also calculated and visualized key network indicators, such as node degree and centrality, using Matplotlib. This visualization aids in identifying which products are central within the network, suggesting their importance in the purchasing patterns observed.

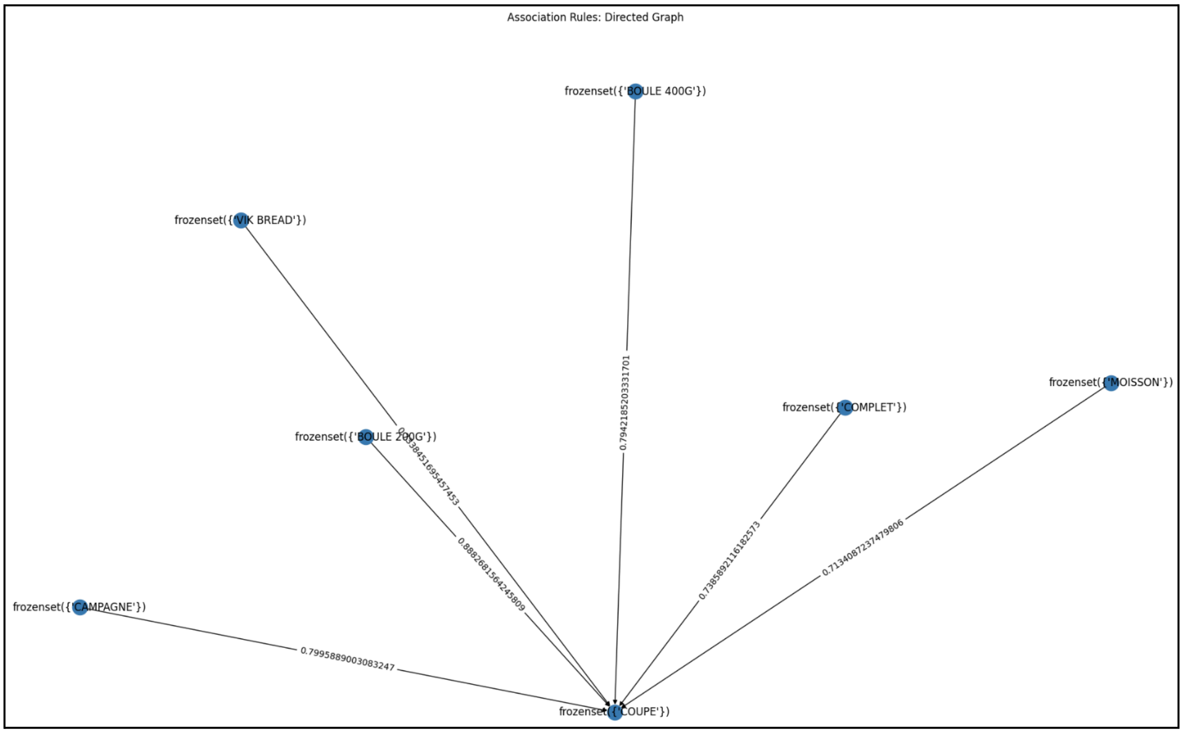
- **Results**
  - Association Rules: Minimum Support = 1% and Minimum Confidence = 60%:



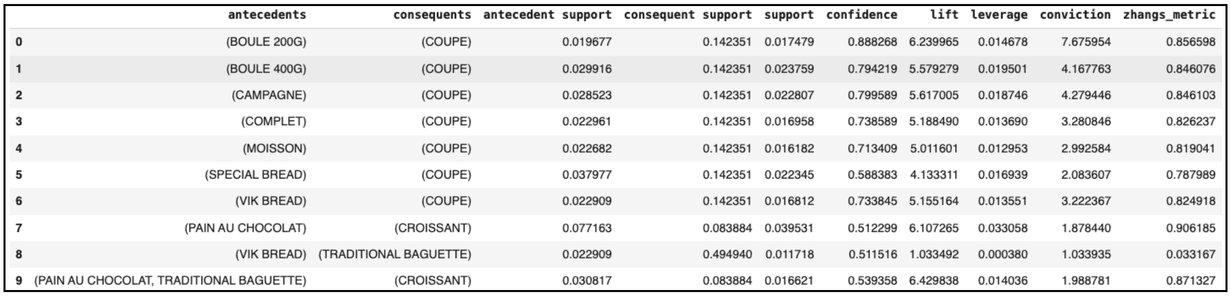
	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction	zhangs_metric
0	(BOULE 200G)	(COUPE)	0.019677	0.142351	0.017479	0.888268	6.239965	0.014678	7.675954	0.856598
1	(BOULE 400G)	(COUPE)	0.029916	0.142351	0.023759	0.794219	5.579279	0.019501	4.167763	0.846076
2	(CAMPAGNE)	(COUPE)	0.028523	0.142351	0.022807	0.799589	5.617005	0.018746	4.279446	0.846103
3	(COMPLET)	(COUPE)	0.022961	0.142351	0.016958	0.738589	5.188490	0.013690	3.280846	0.826237
4	(MOISSON)	(COUPE)	0.022682	0.142351	0.016182	0.713409	5.011601	0.012953	2.992584	0.819041
5	(VIK BREAD)	(COUPE)	0.022909	0.142351	0.016812	0.733845	5.155164	0.013551	3.222367	0.824918

  - Scatter Plot and Network Graph:



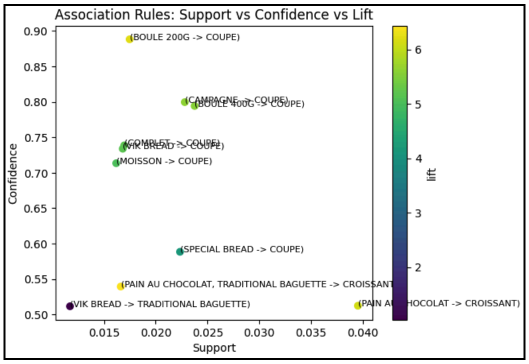


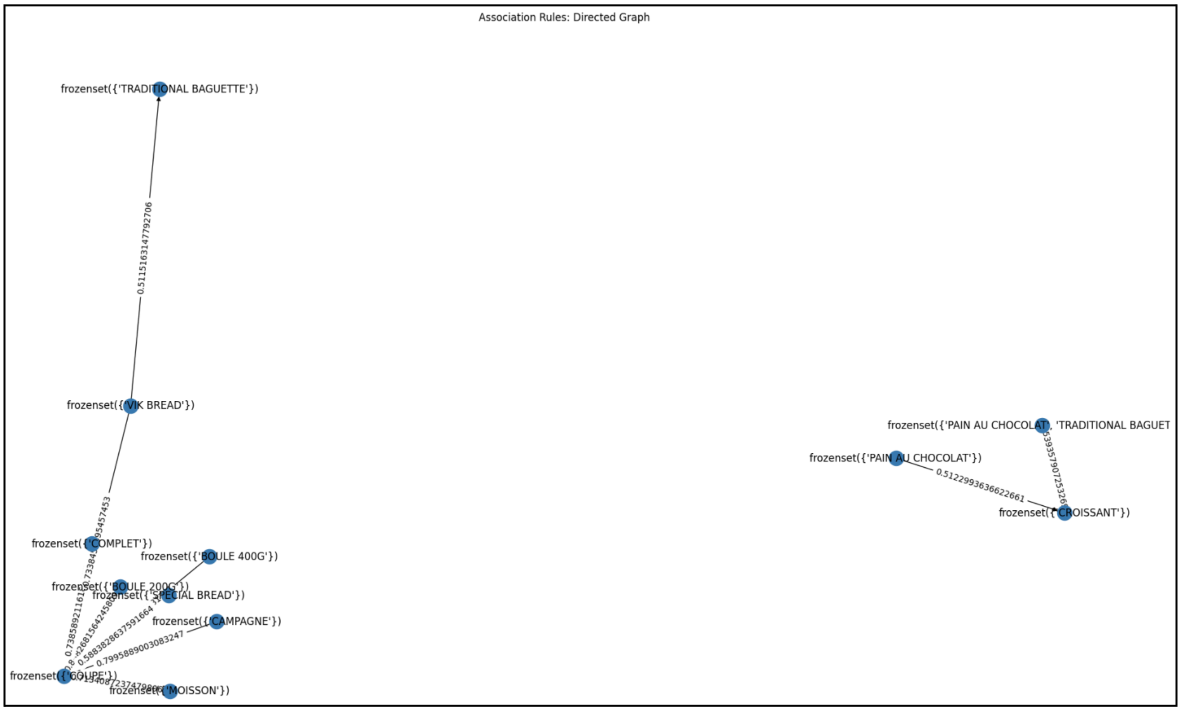
- ○ Association Rules: Minimum Support = 1.5% and Minimum Confidence = 50%:



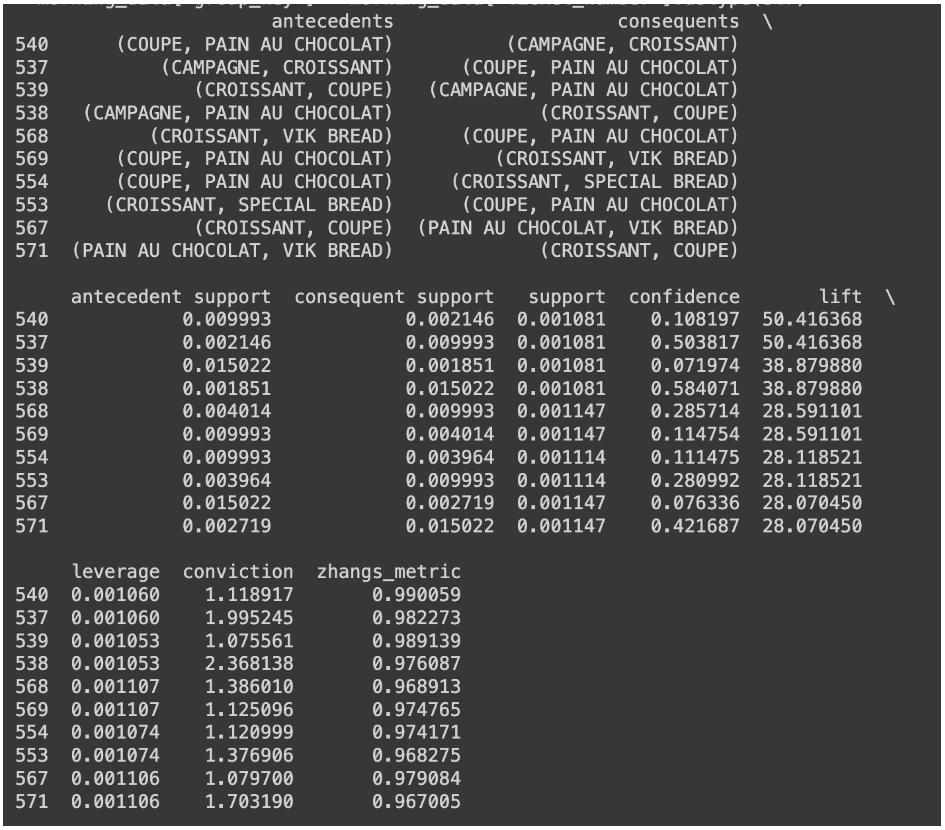
	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction	zhangs_metric
0	(BOULE 200G)	(COUPE)	0.019677	0.142351	0.017479	0.888268	6.239965	0.014678	7.675954	0.856598
1	(BOULE 400G)	(COUPE)	0.029916	0.142351	0.023759	0.794219	5.579279	0.019501	4.167763	0.846076
2	(CAMPAGNE)	(COUPE)	0.028523	0.142351	0.022807	0.799589	5.617005	0.018746	4.279446	0.846103
3	(COMPLET)	(COUPE)	0.022961	0.142351	0.016958	0.738589	5.188490	0.013690	3.280846	0.826237
4	(MOISSON)	(COUPE)	0.022682	0.142351	0.016182	0.713409	5.011601	0.012953	2.992584	0.819041
5	(SPECIAL BREAD)	(COUPE)	0.037977	0.142351	0.022345	0.588383	4.133311	0.016939	2.083607	0.787989
6	(VIK BREAD)	(COUPE)	0.022909	0.142351	0.016812	0.733845	5.155164	0.013551	3.222367	0.824918
7	(PAIN AU CHOCOLAT)	(CROISSANT)	0.077163	0.083884	0.039531	0.512299	6.107265	0.033058	1.878440	0.906185
8	(VIK BREAD)	(TRADITIONAL BAGUETTE)	0.022909	0.494940	0.011718	0.511516	1.033492	0.000380	1.033935	0.033167
9	(PAIN AU CHOCOLAT, TRADITIONAL BAGUETTE)	(CROISSANT)	0.030817	0.083884	0.016621	0.539358	6.429838	0.014036	1.988781	0.871327

- ○ Scatter Plot and Network Graph:





- Morning Association Rule Mining and Apriori Algorithm:

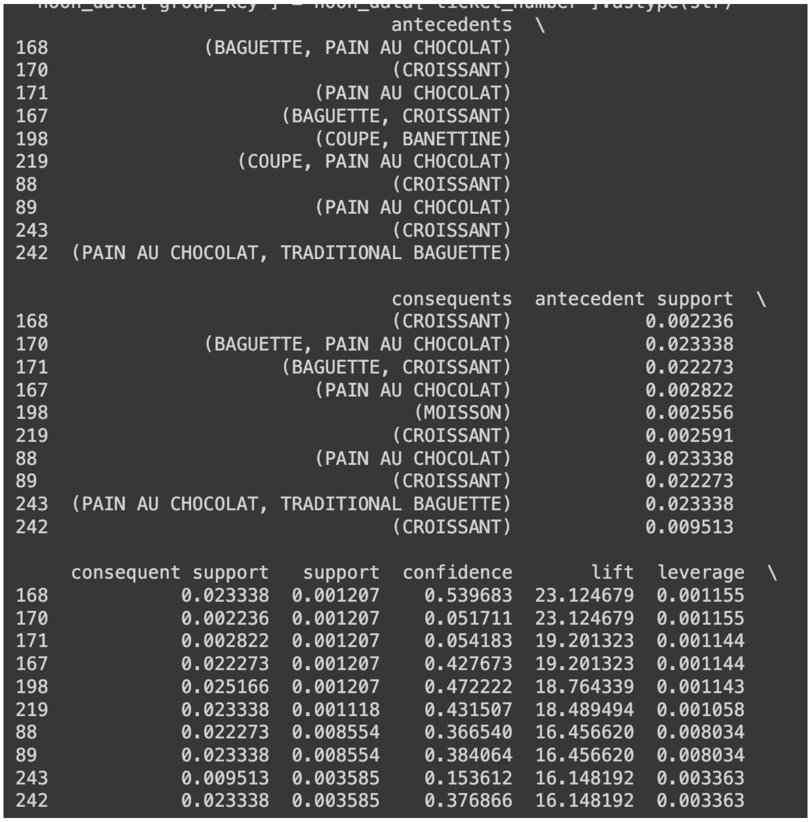


```
                        antecedents                         consequents  \
540      (COUPE, PAIN AU CHOCOLAT)          (CAMPAGNE, CROISSANT)
537          (CAMPAGNE, CROISSANT)      (COUPE, PAIN AU CHOCOLAT)
539             (CROISSANT, COUPE)   (CAMPAGNE, PAIN AU CHOCOLAT)
538   (CAMPAGNE, PAIN AU CHOCOLAT)             (CROISSANT, COUPE)
568         (CROISSANT, VIK BREAD)      (COUPE, PAIN AU CHOCOLAT)
569      (COUPE, PAIN AU CHOCOLAT)         (CROISSANT, VIK BREAD)
554      (COUPE, PAIN AU CHOCOLAT)     (CROISSANT, SPECIAL BREAD)
553     (CROISSANT, SPECIAL BREAD)      (COUPE, PAIN AU CHOCOLAT)
567             (CROISSANT, COUPE)  (PAIN AU CHOCOLAT, VIK BREAD)
571  (PAIN AU CHOCOLAT, VIK BREAD)             (CROISSANT, COUPE)

     antecedent support  consequent support   support  confidence       lift  \
540            0.009993            0.002146  0.001081    0.108197  50.416368
537            0.002146            0.009993  0.001081    0.503817  50.416368
539            0.015022            0.001851  0.001081    0.071974  38.879880
538            0.001851            0.015022  0.001081    0.584071  38.879880
568            0.004014            0.009993  0.001147    0.285714  28.591101
569            0.009993            0.004014  0.001147    0.114754  28.591101
554            0.009993            0.003964  0.001114    0.111475  28.118521
553            0.003964            0.009993  0.001114    0.280992  28.118521
567            0.015022            0.002719  0.001147    0.076336  28.070450
571            0.002719            0.015022  0.001147    0.421687  28.070450

     leverage  conviction  zhangs_metric
540  0.001060    1.118917       0.990059
537  0.001060    1.995245       0.982273
539  0.001053    1.075561       0.989139
538  0.001053    2.368138       0.976087
568  0.001107    1.386010       0.968913
569  0.001107    1.125096       0.974765
554  0.001074    1.120999       0.974171
553  0.001074    1.376906       0.968275
567  0.001106    1.079700       0.979084
571  0.001106    1.703190       0.967005
```

- Noon Association Rule Mining and Apriori Algorithm:

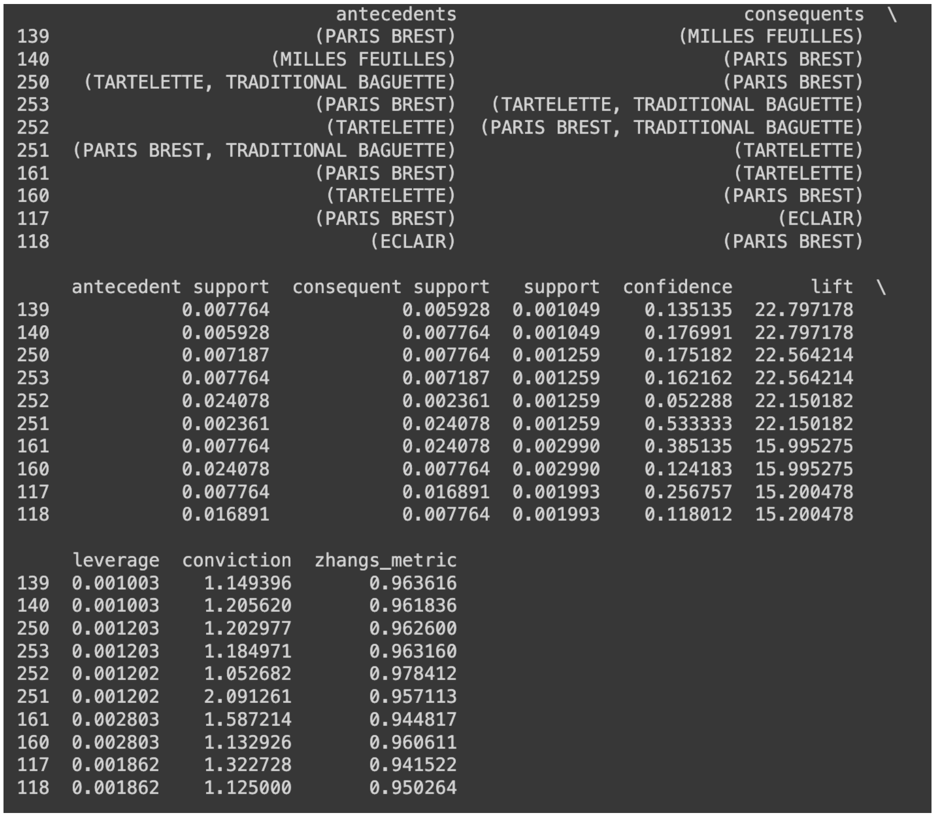


```
  noon_data['group_key'] = noon_data['ticket_number'].astype(str)
                                  antecedents  \
168              (BAGUETTE, PAIN AU CHOCOLAT)
170                               (CROISSANT)
171                        (PAIN AU CHOCOLAT)
167                     (BAGUETTE, CROISSANT)
198                        (COUPE, BANETTINE)
219                 (COUPE, PAIN AU CHOCOLAT)
88                                (CROISSANT)
89                         (PAIN AU CHOCOLAT)
243                               (CROISSANT)
242  (PAIN AU CHOCOLAT, TRADITIONAL BAGUETTE)

                                  consequents  antecedent support  \
168                               (CROISSANT)            0.002236
170              (BAGUETTE, PAIN AU CHOCOLAT)            0.023338
171                     (BAGUETTE, CROISSANT)            0.022273
167                        (PAIN AU CHOCOLAT)            0.002822
198                                 (MOISSON)            0.002556
219                               (CROISSANT)            0.002591
88                         (PAIN AU CHOCOLAT)            0.023338
89                                (CROISSANT)            0.022273
243  (PAIN AU CHOCOLAT, TRADITIONAL BAGUETTE)            0.023338
242                               (CROISSANT)            0.009513

     consequent support   support  confidence       lift  leverage  \
168            0.023338  0.001207    0.539683  23.124679  0.001155
170            0.002236  0.001207    0.051711  23.124679  0.001155
171            0.002822  0.001207    0.054183  19.201323  0.001144
167            0.022273  0.001207    0.427673  19.201323  0.001144
198            0.025166  0.001207    0.472222  18.764339  0.001143
219            0.023338  0.001118    0.431507  18.489494  0.001058
88             0.022273  0.008554    0.366540  16.456620  0.008034
89             0.023338  0.008554    0.384064  16.456620  0.008034
243            0.009513  0.003585    0.153612  16.148192  0.003363
242            0.023338  0.003585    0.376866  16.148192  0.003363
```

- Evening Association Rule Mining and Apriori Algorithm

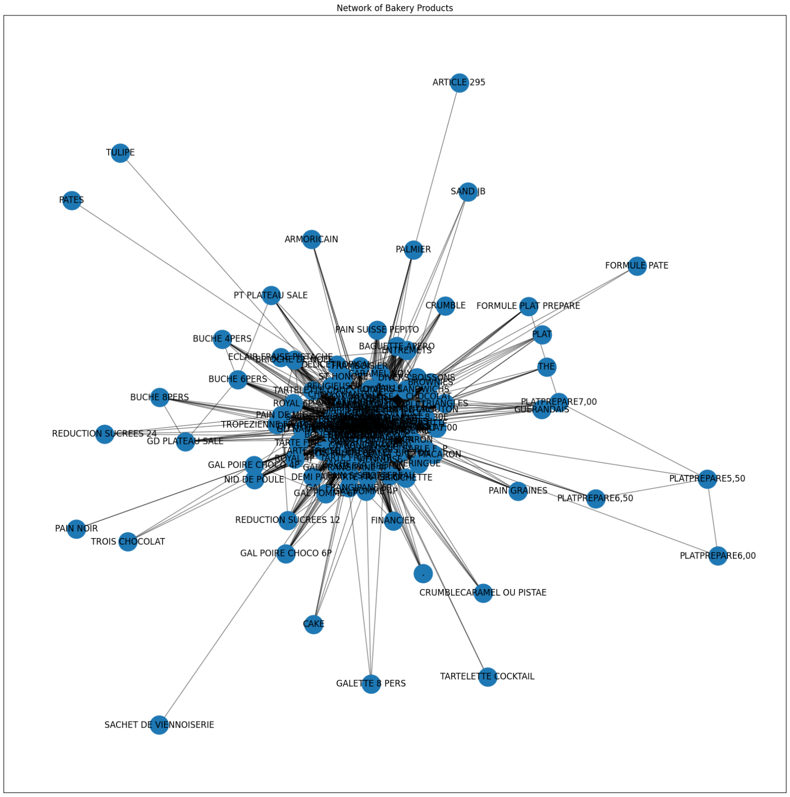


```
                             antecedents                          consequents  \
139                        (PARIS BREST)                    (MILLES FEUILLES)
140                    (MILLES FEUILLES)                        (PARIS BREST)
250   (TARTELETTE, TRADITIONAL BAGUETTE)                        (PARIS BREST)
253                        (PARIS BREST)   (TARTELETTE, TRADITIONAL BAGUETTE)
252                         (TARTELETTE)  (PARIS BREST, TRADITIONAL BAGUETTE)
251  (PARIS BREST, TRADITIONAL BAGUETTE)                         (TARTELETTE)
161                        (PARIS BREST)                         (TARTELETTE)
160                         (TARTELETTE)                        (PARIS BREST)
117                        (PARIS BREST)                             (ECLAIR)
118                             (ECLAIR)                        (PARIS BREST)

     antecedent support  consequent support   support  confidence       lift  \
139            0.007764            0.005928  0.001049    0.135135  22.797178
140            0.005928            0.007764  0.001049    0.176991  22.797178
250            0.007187            0.007764  0.001259    0.175182  22.564214
253            0.007764            0.007187  0.001259    0.162162  22.564214
252            0.024078            0.002361  0.001259    0.052288  22.150182
251            0.002361            0.024078  0.001259    0.533333  22.150182
161            0.007764            0.024078  0.002990    0.385135  15.995275
160            0.024078            0.007764  0.002990    0.124183  15.995275
117            0.007764            0.016891  0.001993    0.256757  15.200478
118            0.016891            0.007764  0.001993    0.118012  15.200478

     leverage  conviction  zhangs_metric
139  0.001003    1.149396       0.963616
140  0.001003    1.205620       0.961836
250  0.001203    1.202977       0.962600
253  0.001203    1.184971       0.963160
252  0.001202    1.052682       0.978412
251  0.001202    2.091261       0.957113
161  0.002803    1.587214       0.944817
160  0.002803    1.132926       0.960611
117  0.001862    1.322728       0.941522
118  0.001862    1.125000       0.950264
```

- Link Analysis



Centrality and degree:

Top degrees: [('TRADITIONAL BAGUETTE', 46126), ('COUPE', 38507), ('CROISSANT', 21855), ('PAIN AU CHOCOLAT', 20086), ('BAGUETTE', 12797), ('BANETTE', 10863), ('SPECIAL BREAD', 7644), ('BOULE 400G', 6888), ('VIK BREAD', 6797), ('CAMPAGNE', 6535)]

Top centrality: [('TRADITIONAL BAGUETTE', 0.04578434095827403), ('SANDWICH COMPLET', 0.03201814366755598), ('SEIGLE', 0.02988392878742063), ('VIENNOISE', 0.027044052491469447), ('DIVERS SANDWICHS', 0.024187301084497074), ('TRAITEUR', 0.023373021401109008), ('TARTELETTE', 0.022813215344960728), ('PAIN CHOCO AMANDES', 0.022561806385766804), ('VIK BREAD', 0.02212225430157864), ('MOISSON', 0.02160439288445034)]

Analysis + Conclusion

Based on the Apriori algorithm’s frequent itemset generation, with minimum support of 1%, it is clear that Baguette is a frequently purchased item, as are multiple variants of Baguette including Baguette Graine and Cereal Baguette. Generally, among the frequent itemset items, many of the products with high confidence were bread-like items, such as Baguette, Boule, Banette, and Banettine. Furthermore, based on the association rules generated from this itemset (with a minimum confidence of 60%), it is clear that “Coupe,” indicating when a customer asked for their bread/pastry to be sliced, was a consequence of every antecedent. These association rules, thus, suggest that bread/pastry items were requested to be sliced with high frequency in all of the transactions recorded during this period. Perhaps, to increase sales and improve efficiency, the bakery could offer customers the option to purchase breads or pastries whole *or* by the slice. If pricing for each slice was effective and maintained good profit margins, the sale of breads and pastries by the slice could increase overall profit for the bakery.

Based on the Apriori algorithm’s frequent itemset generation, with minimum support of 1.5%, we see again that Baguette is a frequently purchased item. One other common type of bread that we see is Croissant. Among the frequent itemset items, the products with higher confidence were Boule and Campagne. After performing association rules mining on this itemset with a minimum confidence of 50%, we see that “Coupe,” again, was a common consequence of the antecedents. In two out of three of the other association rules generated, Croissant was the consequent and Pain au Chocolat was part of the antecedent. Similar to the first example, these rules suggest that the bread items were often asked to be sliced. Additionally, customers who purchased Pain au Chocolat were more likely to purchase other types of croissants. To increase sales, the bakery could try grouping Pain au Chocolat with other croissants in their store to encourage customers to try and purchase multiple croissant types. If this grouping strategy is effective, the bakery could try grouping other similar breads to increase their sales.

Looking at the different Association Rules generated from the morning, noon, and evening time periods, we can identify several purchasing trends. In the morning, the combination of "Coupe, Pain au Chocolat" and "Campagne, Croissant" (and vice versa) exhibit significantly large lifts, indicating a strong preference for these pairs during breakfast time. These high lift values suggest that customers are looking for fulfilling and quick breakfasts composed of a variety of breads and chocolates. Therefore, bundling these items into a ‘Breakfast Special’ could increase sales and customer satisfaction during the bakery’s morning rush hours. In the afternoon (Noon) time period, rules associating "Baguette, Pain au Chocolat" and "Croissant" (and vice versa) show high support and confidence, reflecting a lunchtime preference for substantial, easy-to-eat options that can serve as a quick lunch. Therefore, creating lunch combos featuring these popular items, and possibly offering a drink or a small dessert to round out the meal, could attract more midday customers. In the evening time period, pairings like "Paris Brest, Eclair" and "Tartelette, Traditional Baguette" have high lifts and confidence, indicating that customers prefer to indulge in desserts and specialty breads in the evening. Therefore, introducing an "Evening Delight" promotion that features dessert specials could capitalize on the demand for sweet treats after dinner.

From performing link analysis, we were able to retrieve bakery products with high degree and high centrality. In high degree analysis, Traditional Baguette, Coupe, Croissant, Pain au Chocolat, and Baguette have the highest degrees. This indicates that these items are the most connected within the network, suggesting they are frequently purchased in combination with many other items. These are staple products that likely drive the bulk of traffic and sales in the bakery. Given their high degree of connectivity, placing these core products strategically within the store can facilitate easier access for customers and potentially drive sales of less popular items placed nearby. In high centrality analysis, items like Sandwich Complet and Seigle show high centrality but are not necessarily the top-selling items. High centrality for these items indicates they play a crucial role within the network, likely serving specific customer segments very effectively. Those items can be used to create cross-selling opportunities — for instance, promoting lesser-known products that connect well with high-centrality items might boost overall sales.