[[](https://towardsdatascience.com/?source=post_page-----4c18164ef82b--------------------------------)](https://towardsdatascience.com/?source=post_page-----4c18164ef82b--------------------------------)

[Sign in](https://medium.com/m/signin?operation=login&redirect=https%3A%2F%2Ftowardsdatascience.com%2Fmba-for-breakfast-4c18164ef82b&source=post_page-----4c18164ef82b---------------------nav_reg-----------)

[Get started](https://medium.com/m/signin?operation=register&redirect=https%3A%2F%2Ftowardsdatascience.com%2Fmba-for-breakfast-4c18164ef82b&source=post_page-----4c18164ef82b---------------------nav_reg-----------)

[Follow](https://medium.com/m/signin?actionUrl=%2F_%2Fsubscribe%2Fcollection%2Ftowards-data-science&operation=register&redirect=https%3A%2F%2Ftowardsdatascience.com%2Fmba-for-breakfast-4c18164ef82b&source=post_page-----4c18164ef82b---------------------follow_header-----------)

[556K Followers](https://towardsdatascience.com/followers?source=post_page-----4c18164ef82b--------------------------------)

·

[Editors' Picks](https://towardsdatascience.com/tagged/editors-pick?fromNav=true&source=post_page-----4c18164ef82b--------------------------------)[Features](https://towardsdatascience.com/tagged/tds-features?fromNav=true&source=post_page-----4c18164ef82b--------------------------------)[Explore](https://towardsdatascience.com/tagged/tds-explore?fromNav=true&source=post_page-----4c18164ef82b--------------------------------)[Contribute](https://towardsdatascience.com/questions-96667b06af5?source=post_page-----4c18164ef82b--------------------------------)

[About](https://towardsdatascience.com/about?source=post_page-----4c18164ef82b--------------------------------)

**MBA For Breakfast — A Simple Guide to Market Basket Analysis**

[George Wong](https://medium.com/@georgewong?source=post_page-----4c18164ef82b--------------------------------)

[George Wong](https://medium.com/@georgewong?source=post_page-----4c18164ef82b--------------------------------)

[Oct 12, 2018·5 min read](https://towardsdatascience.com/mba-for-breakfast-4c18164ef82b?source=post_page-----4c18164ef82b--------------------------------)

Image for post



Photo by [Liene Vitamante](https://unsplash.com/photos/PO2GsmeN8hU?utm_source=unsplash&utm_medium=referral&utm_content=creditCopyText) on [Unsplash](https://unsplash.com/search/photos/pastry-store?utm_source=unsplash&utm_medium=referral&utm_content=creditCopyText)

So recently, I was fortunate enough to work on a project that involves doing market basket analysis but obviously I wouldn’t be able to discuss my work at [Kaodim](https://www.kaodim.com/)on Medium. So instead, I try to look for suitable datasets on [Kaggle.com](http://kaggle.com/). Which I managed to find one [here](https://www.kaggle.com/xvivancos/transactions-from-a-bakery) AND applied whatever that I know to make it happen! Also Shout-out to

[Susan Li](https://medium.com/u/731d8566944a?source=post_page-----4c18164ef82b--------------------------------)

 for her wonderful work on MBA, which can be found [here](https://towardsdatascience.com/a-gentle-introduction-on-market-basket-analysis-association-rules-fa4b986a40ce)!

So what is a Market Basket Analysis? According to the book [Database Marketing](https://www.springer.com/gp/book/9780387725789):

*Market basket analysis scrutinizes the products customers tend to buy together, and uses the information to decide which products should be cross-sold or promoted together. The term arises from the shopping carts supermarket shoppers fill up during a shopping trip.*

On a side note, **Market Basket**is also a chain of 79 supermarkets in New Hampshire, Massachusetts, and Maine in the United States, with headquarters in Tewksbury, Massachusetts (Wikipedia).

Normally MBA is done on transaction data from the point of sales system on a customer level. We can use MBA to extract interesting association between products from the data. Hence its output consists of a series of product association rules: for example, if customers buy product A they also tend to buy product B. We will follow the three most popular criteria evaluating the quality or the strength of an association rule (will get back to this later).

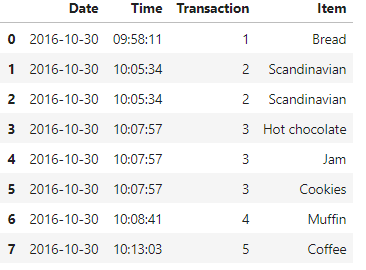
Getting the right packages (Python):

import pandas as pd  
import numpy as np   
import seaborn as sns  
import matplotlib.pyplot as plt  
from mlxtend.frequent\_patterns import apriori  
from mlxtend.frequent\_patterns import association\_rules  
import mlxtend as ml

Let’s take a peek into our BreadBasket’s data:

bread = pd.read\_csv(r"D:\Downloads\BreadBasket\_DMS.csv")  
bread.head(8)

Image for post

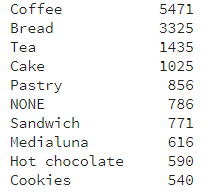


“Data set containing 15'010 observations and more than 6.000 transactions from a bakery.” More information on the variables can be found in [Kaggle](https://www.kaggle.com/xvivancos/transactions-from-a-bakery).

What are the “hot” items at BreadBasket?

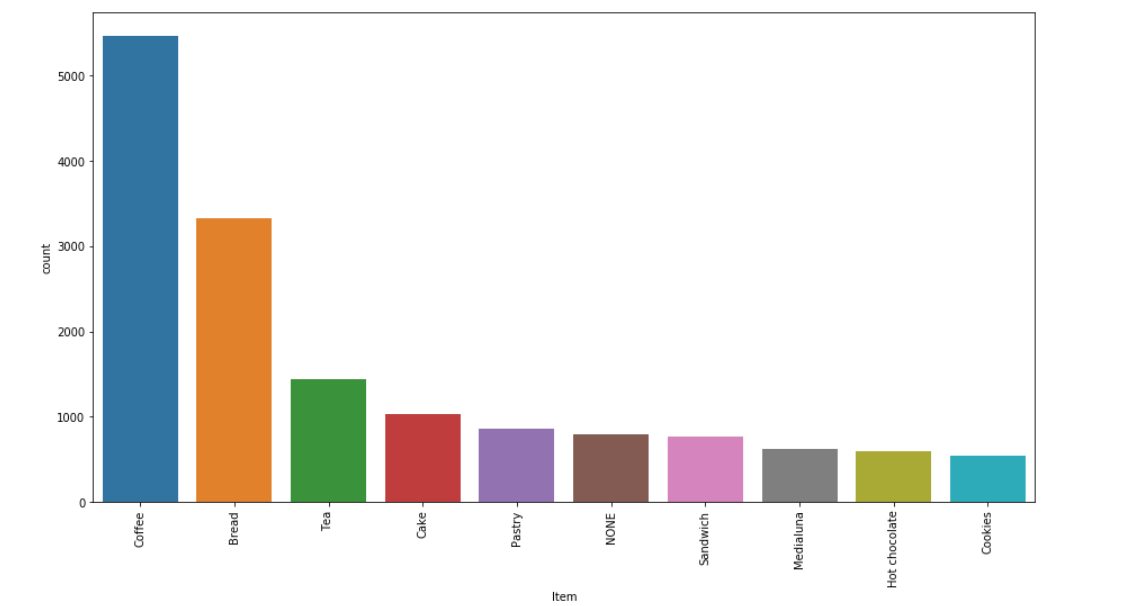
sns.countplot(x = 'Item', data = bread, order = bread['Item'].value\_counts().iloc[:10].index)  
plt.xticks(rotation=90)

Image for post



Lit items at BreadBasket

Image for post



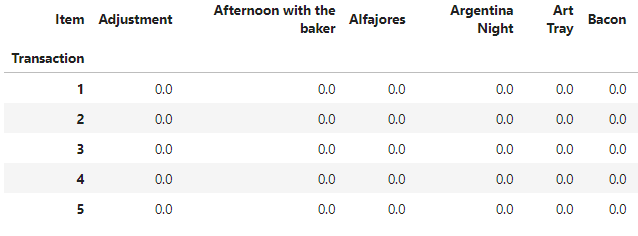
It seems like coffee are the hottest item in the dataset, I guess everybody wants a cup of hot coffee in the morning perhaps.

How many items are sold daily?

Time to get right into the MBA itself! Kudos to Chris Moffitt on the awesome guide and tutorial of MBA using [python](http://pbpython.com/market-basket-analysis.html).

We will be using MLxtend library’s [Apriori Algorithm](http://rasbt.github.io/mlxtend/user_guide/frequent_patterns/apriori/) for extracting frequent item sets for further analysis. The apriorifunction expects data in a one-hot encoded pandas DataFrame. Thus your dataframe should look like this:

Image for post



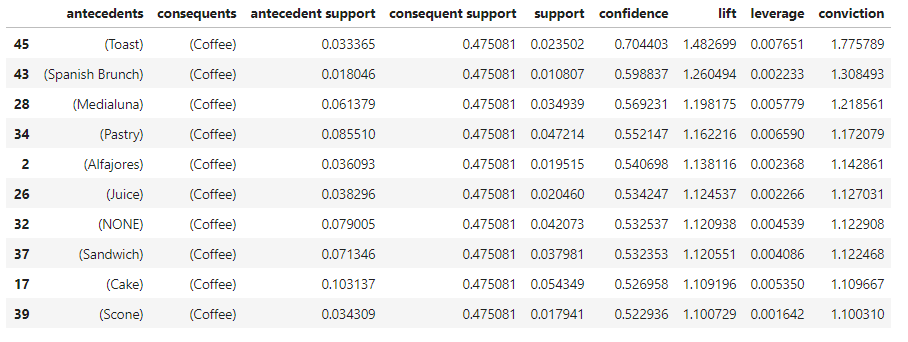
First, We’ll group the bread dataframe accordingly and display the count of items then we need to consolidate the items into 1 transaction per row with each product [1 hot encoded](http://pbpython.com/categorical-encoding.html). Which would result in the table above!

df = bread.groupby(['Transaction','Item']).size().reset\_index(name='count')basket = (df.groupby(['Transaction', 'Item'])['count']  
 .sum().unstack().reset\_index().fillna(0)  
 .set\_index('Transaction'))#The encoding function  
def encode\_units(x):  
 if x <= 0:  
 return 0  
 if x >= 1:  
 return 1basket\_sets = basket.applymap(encode\_units)

After that, we will generate frequent item sets with a minimum support of at least 1% as this a more favourable support level that could show us more results.

frequent\_itemsets = apriori(basket\_sets, min\_support=0.01, use\_colnames=True)  
rules = association\_rules(frequent\_itemsets, metric="lift")  
rules.sort\_values('confidence', ascending = False, inplace = True)  
rules.head(10)

Image for post



Final output

The end

Remember I told y’all that we’ll get back to the three most popular criteria evaluating the quality or the strength of an association rule. There are **support, confidence**and**lift**:  
1. Support is the percentage of transactions containing a particular combination of items relative to the total number of transactions in the database. The support for the combination A and B would be,

P(AB) or P(A) for Individual A

2. Confidence measures how much the consequent (item) is dependent on the  
antecedent (item). In other words, confidence is the conditional probability of the consequent given the antecedent,

P(B|A)

where P(B|A) = P(AB)/P(A)

3. Lift (also called improvement or impact) is a measure to overcome the  
problems with support and confidence. Lift is said to measure the difference — measured in ratio — between the confidence of a rule and the expected confidence. Consider an association rule “if A then B.” The lift for the rule is defined as

P(B|A)/P(B) or P(AB)/[P(A)P(B)].

As shown in the formula, lift is symmetric in that the lift for “if A then B” is the same as the lift for “if B then A.”  
4. Each criterion has its advantages and disadvantages but in general we would like association rules that have high confidence, high support, and high lift.

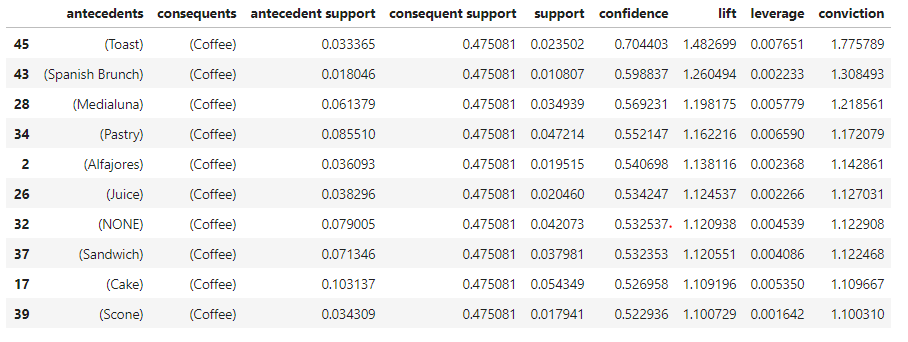
As a summary,

Confidence = P(B|A)

Support = P(AB)

Lift = P(B|A)/P(B)

Image for post



From the output, the lift of an association rule “if Toast then Coffee” is 1.48 because the confidence is 70%. This means that consumers who purchase Toast are 1.48 times more likely to purchase Coffee than randomly chosen customers. Larger lift means more interesting rules. Association rules with high support are potentially interesting rules. Similarly, rules with high confidence would be interesting rules as well.

*Reference:*

*Kaggle Dataset:*[*https://www.kaggle.com/xvivancos/transactions-from-a-bakery*](https://www.kaggle.com/xvivancos/transactions-from-a-bakery)

*Susan Li’s Market Basket Analysis:*<https://towardsdatascience.com/a-gentle-introduction-on-market-basket-analysis-association-rules-fa4b986a40ce>

*Database Marketing book:*[*https://www.springer.com/gp/book/9780387725789*](https://www.springer.com/gp/book/9780387725789)

*Apriori Algorithm:*[*http://rasbt.github.io/mlxtend/user\_guide/frequent\_patterns/apriori/*](http://rasbt.github.io/mlxtend/user_guide/frequent_patterns/apriori/)

*Market Basket Analysis using Python by Chris Moffitt:*[*http://pbpython.com/market-basket-analysis.html*](http://pbpython.com/market-basket-analysis.html)

*Kaodim.com:*[*https://www.kaodim.com/*](https://www.kaodim.com/)

Source code in **Jupyter Notebook**[here](https://github.com/georgeblu1/Data/blob/master/Market%20Basket%20Analysis.ipynb)!

[George Wong](https://medium.com/@georgewong?source=post_sidebar--------------------------post_sidebar-----------)

Data Scientist at AirAsia, former data analyst intern at [kaodim.com](http://kaodim.com/) and [carlist.my](http://carlist.my/) <https://georgewongsinhong.me/>

[Follow](https://medium.com/m/signin?actionUrl=%2F_%2Fsubscribe%2Fuser%2Fa06d00753b5b%2F4c18164ef82b&operation=register&redirect=https%3A%2F%2Ftowardsdatascience.com%2Fmba-for-breakfast-4c18164ef82b&source=post_sidebar-a06d00753b5b-------------------------follow_card-----------)

133

1

Sign up for The Daily Pick

By Towards Data Science

Hands-on real-world examples, research, tutorials, and cutting-edge techniques delivered Monday to Thursday. Make learning your daily ritual. [Take a look.](https://medium.com/towards-data-science/newsletters/the-daily-pick?source=newsletter_v3_promo--------------------------newsletter_v3_promo-----------)



Get this newsletter

By signing up, you will create a Medium account if you don’t already have one. Review our [Privacy Policy](https://policy.medium.com/medium-privacy-policy-f03bf92035c9?source=newsletter_v3_promo--------------------------newsletter_v3_promo-----------) for more information about our privacy practices.

133

1

* [Data Science](https://towardsdatascience.com/tagged/data-science)
* [Statistics](https://towardsdatascience.com/tagged/statistics)
* [Tutorial](https://towardsdatascience.com/tagged/tutorial)
* [Python](https://towardsdatascience.com/tagged/python)
* [Internships](https://towardsdatascience.com/tagged/internships)

[More from Towards Data Science](https://towardsdatascience.com/?source=follow_footer-------------------------------------)

[Follow](https://medium.com/m/signin?actionUrl=%2F_%2Fsubscribe%2Fcollection%2Ftowards-data-science%2F4c18164ef82b&operation=register&redirect=https%3A%2F%2Ftowardsdatascience.com%2Fmba-for-breakfast-4c18164ef82b&source=follow_footer--------------------------follow_footer-----------)

Your home for data science. A Medium publication sharing concepts, ideas and codes.

[Susan Li](https://actsusanli.medium.com/?source=follow_footer---------0----------------------------)

[·Oct 12, 2018](https://towardsdatascience.com/natural-language-processing-for-fuzzy-string-matching-with-python-6632b7824c49?source=follow_footer---------0----------------------------)