MARKETING AND RETAIL ANALYTICS - ASSIGNMENT1

EXPLORATORY ANALYSIS - EXECUTIVE SUMMARY

The given dataset is the 1 year (1-Apr-2010 till 31-Mar-2011) transaction data set of a Café Chain for one of its restaurants. Please find below the details of the data set.

Each row represents the transaction details of a unique item(but can have multiple quantities) per order.

- Number of rows: 1,45,830
- Number of columns(variables): 10
- Data of 580 unique Menu Items.
- Data of 69,982 transactions.(identified by Bill numbers).

***Note**: The given excel file has 3 sheets of data. As the bill numbers and dates of the second and third sheets are not matching with the first one, I have used only the first sheet(named as Sheet2) for this analysis.*

DETAILS OF VARIABLES

#	Variable	Description
1	Date	Date of transaction
2	Bill Number	Bill number of the transaction
3	Item Desc	Description of the item ordered
4	Time	Time of transaction
5	Quantity	Number of quantity ordered for the item
6	Rate	Price per one quantity of the item
7	Tax	Total tax amount for the item
8	Discount	Total discount amount
9	Total	Total price excluding the discount
10	Category	Category of the Item Values: Beverage, Food, Liquor, Liquor &Tobacco*,Merchandise, Misc, Tobacco, Wines.

**Note: Tobacco is misspelt as Tpbacco in the sheet*

MISSING VALUES, OUTLIERS

- No missing values were found in the data
- 1+1 WINE GLASS having Rate as 1 could be an outlier and needs further investigation.

SUMMARY STATISTICS

Initial summary statistics will yield us the below insights:

- **Date: The values range from 1-Apr-2010 to 31-Mar-2010**

  The maximum number(top 5) of unique items per order were ordered on the following dates:

  31-Dec-10: 834

  3-Apr-10 : 631

  18-Dec-10: 620

  29-Jan-11: 607

  5-Feb-11 : 573

- **Bill Number:** The maximum number(top 5) of unique items per order were ordered in the following transactions(identified by the bill number):

  G0490530: 23, G0518006: 23 , G0489943: 21, G0526679: 19 , G0495644: 18

- **Item Desc:** The items that have appeared in maximum number of orders(top 5):

  NIRVANA HOOKAH SINGLE : 8553 MINT FLAVOUR SINGLE : 5817

  CAPPUCCINO : 5495 GREAT LAKES SHAKE : 4895 SAMBUCA : 4425

- **Time:** The maximum number(top 5) of unique items per order were ordered on the following times:

  10:25:36 PM: 33

  11:35:33 PM: 30

  10:58:37 PM: 29

  11:02:58 PM: 26

  1:25:14 AM : 25

  This indicates that peak time for the restaurant could be roughly from 10:25PM to 1:30AM
- **Quantity:** The values range from 1 to 30
- **Rate:** The values range from 0.01(Mothers day Spl) to 2100(GOSSIPS CHARD AUS (BTL) –Wines)
- **Tax:** The values range from 0 to 2731.25.
- **Discount:** The values range from 0 to 825.
- **Total:** The values range from 0.01(Mothers Day Spl) to 14231.25(PARTY CHARGES @ 500/-)

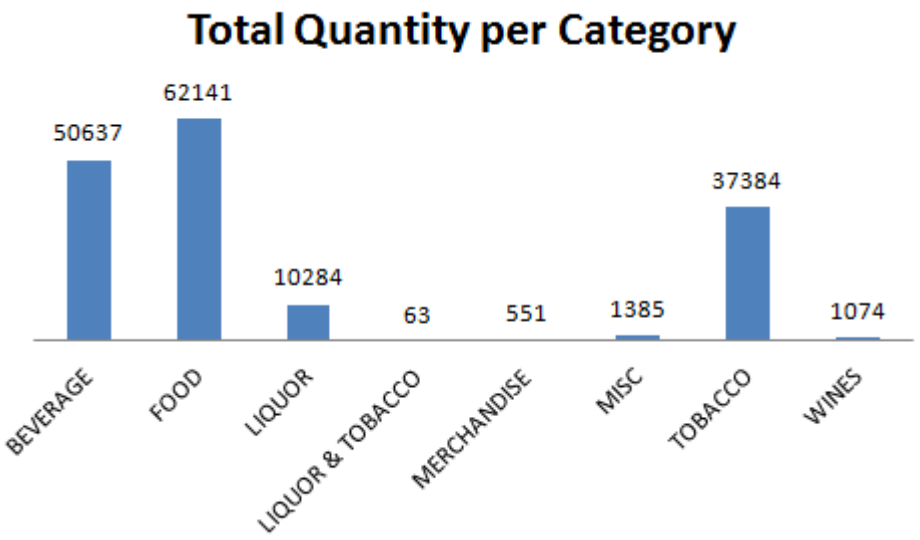
- Category:

#	Category	Count
1	FOOD	57023
2	BEVERAGE	43573
3	TOBACCO	36496
4	LIQUOR	6201
5	MISC	1187
6	WINES	809
7	MERCHANDISE	487
8	LIQUOR & TPBACCO	54

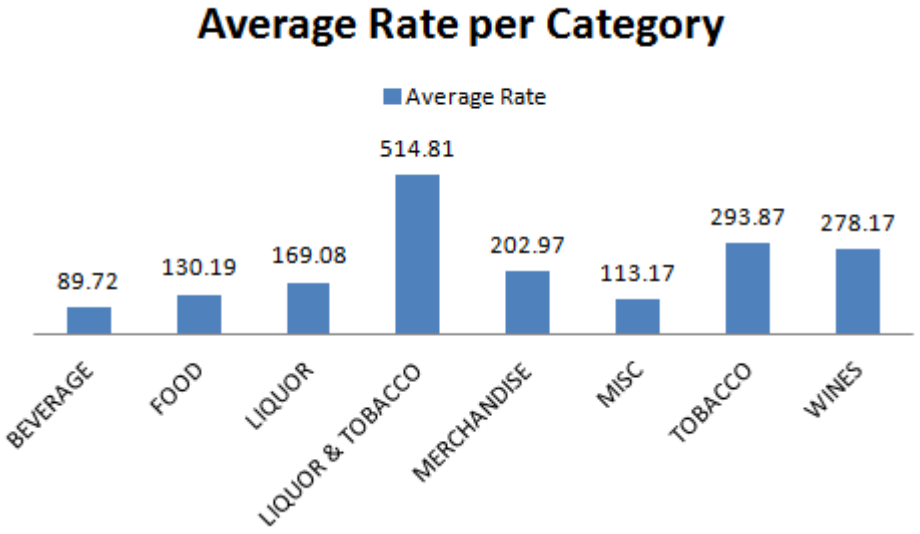
We can see that items of category “food” is ordered the maximum number of times(unique items per order – not considering multiple quantities of the same item in the same order) followed by “beverage” and “tobacco”.

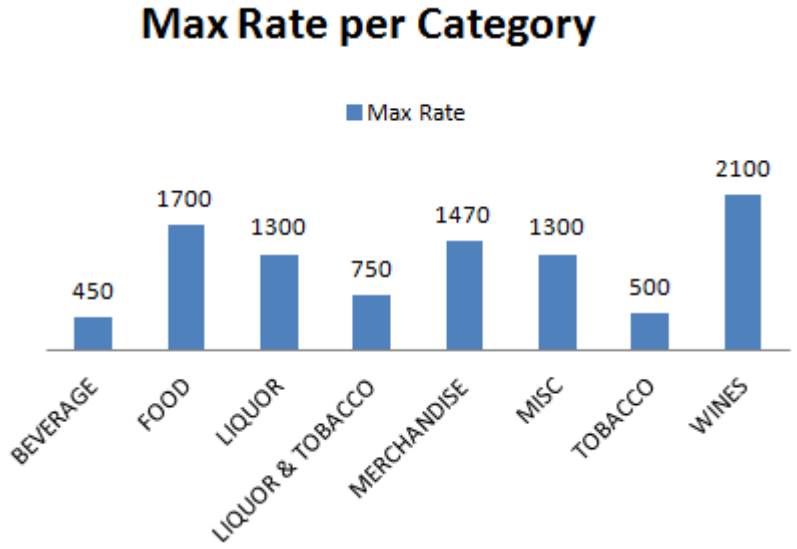
TRENDS IN THE DATASET

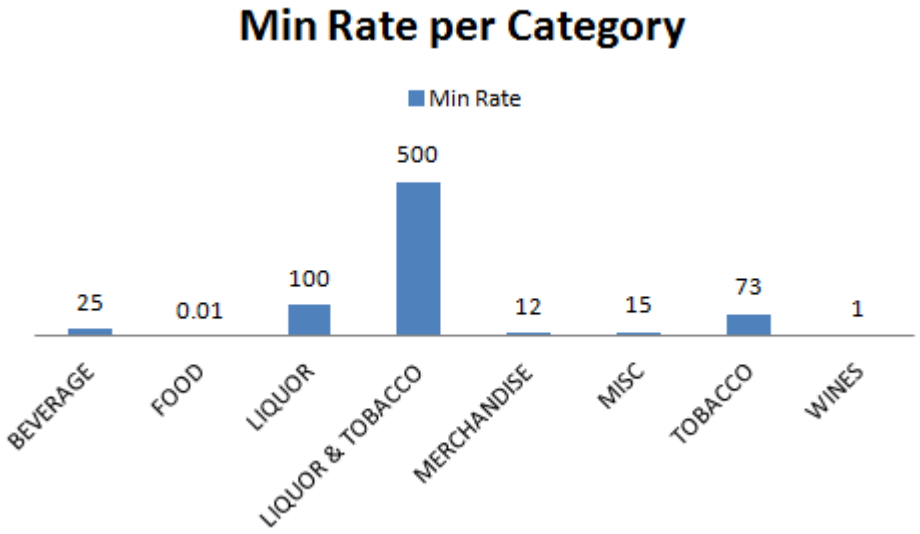
As the size of the dataset is quite large, mining the data using the Category dimension would enable us to have an easy and crisper grasp of the data than other dimensions like Item Desc, Date, Time etc.



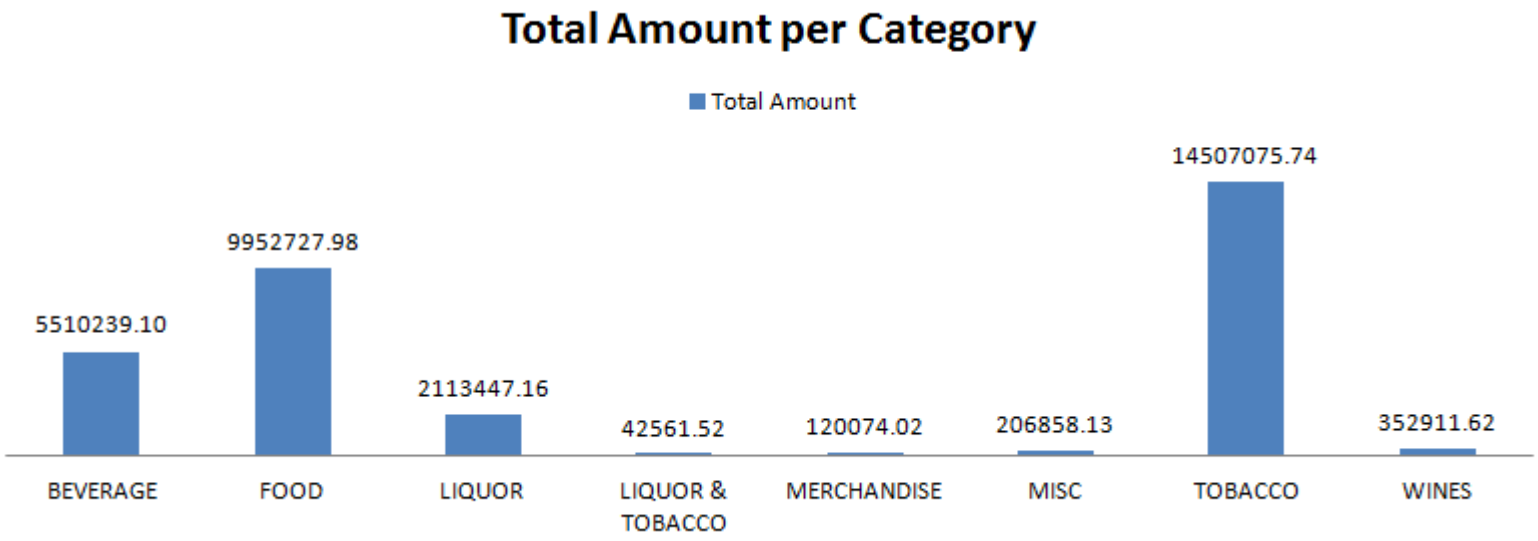
From this plot we can observe that in the 1year dataset, Food items had the maximum sales volume followed by Beverage and Tobacco.



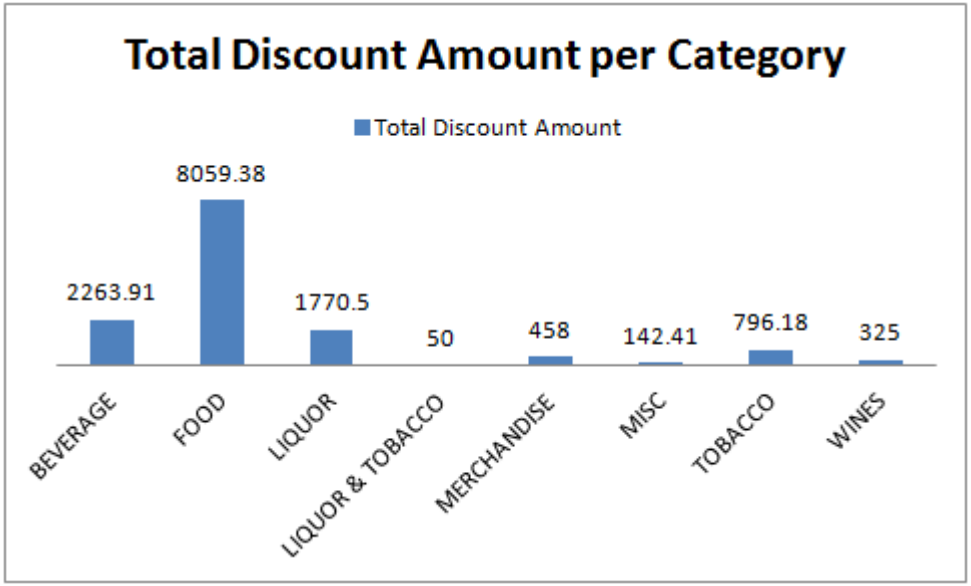




To understand the rate trend per category, I have plotted the average, min and max of rate per category. We can observe that the Liquor &Tobacco has the highest average and min rates, Wines have the highest maximum rate.

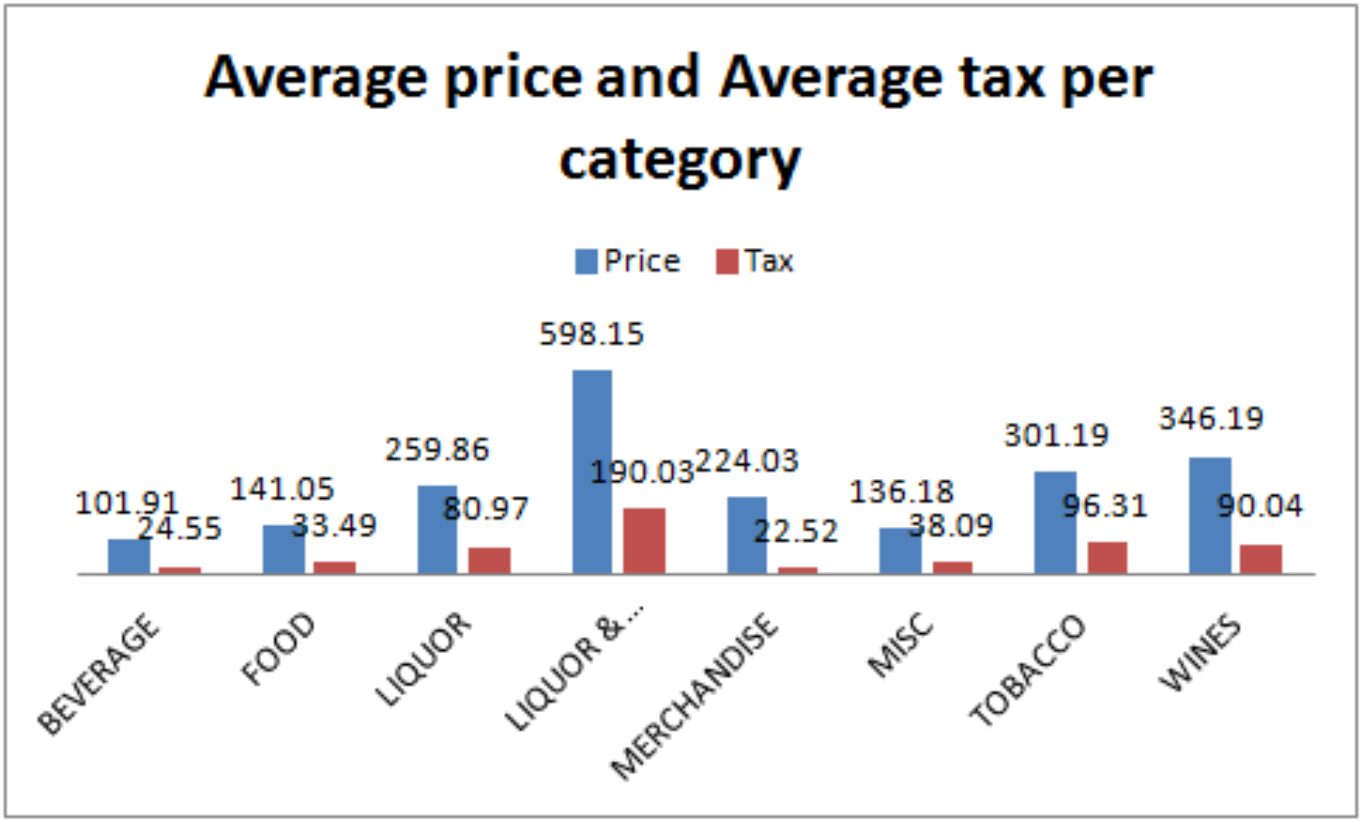


From this plot, we can infer that the maximum sales revenue is generated from the category “Tobacco” followed by Food and Beverage.



From this plot, we would be able to understand the total discount amount given by the restaurant in each of the categories. Food items were given maximum amount of discount followed by Beverage and Liquor .

- Tobacco, Liquor & Tobacco, Liquor are having the highest tax rates.



MENU ANALYSIS

Identify the most popular combos that can be suggested to the restaurant chain after a thorough analysis of the most commonly occurring sets of menu items in the customer orders.

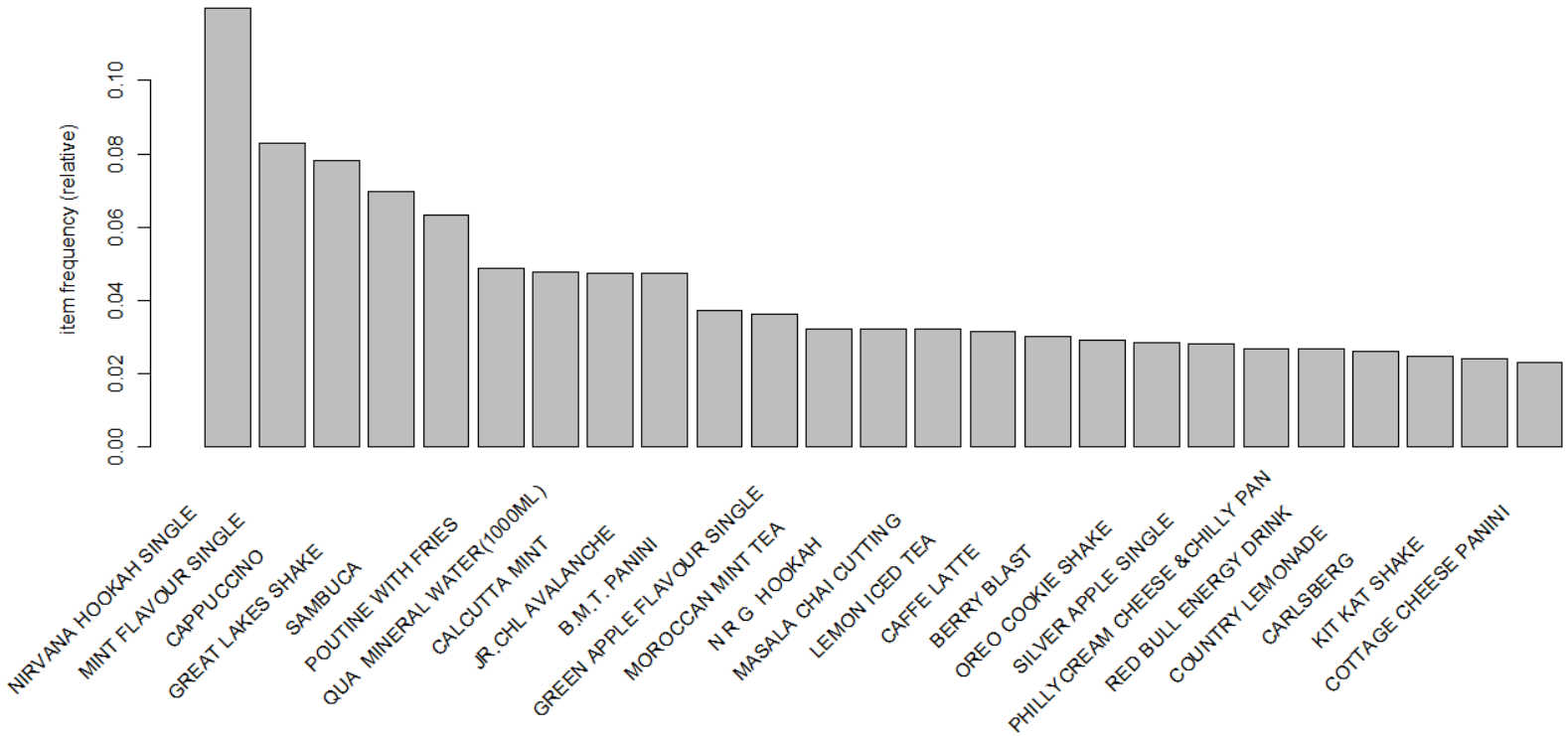
As per my understanding, this analysis can be considered similar to a market basket analysis to find the combinations of items that are ordered most frequently and hence the most popular. This info can be used to design the menu(catalogue), plan discounts/offers etc.

I have applied this analysis to our dataset using **the apriori algorithm using R and the arules package.**

Steps:

- Extracting bills which have more than 1 item. Count is 38,280
- Extracting data of bills which have more than 1 item . Count is 1,14,128
- We need to arrange records to a single record per Bill number, so that the individual items that belong to each bill are aggregated across columns into a single record as an array of products using the **split** function.
- Now we convert this array data into a "**Transaction**" object optimized for running the **apriori** algorithm. Finally, we run our algorithm setting minimum support and confidence thresholds, below which R ignores any rules.

- Plot of the relative frequency of each item (i.e. the fraction of transactions) for the top 25 items by item frequency (i.e. the fraction of transactions that each item appears in). The most frequent item appears in less than 15% of the transactions.



- Inspect the actual rules generated by the algorithm -

  In our case, the algorithm has identified 60 rules. Out of the 60, the first 50 are not helpful as there are no items on the LHS. (For these rules, there are no items on the LHS, the support = the confidence and the lift = 1.)

- The last 10 rules(51-60) are giving the combo rules(ItemX and ItemY combo). Based on the support, confidence and lift values, the **most popular combos** are identified(in blue).

- Hence the most popular combos are:

- **Poutine with fries, Nirvana Hookah Single**
- **Cappuccino, Great Lakes Shake**

followed by

- **Qua Mineral Water(1000ML), Nirvana Hookah Single**
- **Mint Flavour Single, Cappuccino**
- **Great Lakes Shake, Nirvana Hookah Single**

ItemX		ItemY	Support	Confidence	Lift
{MINT FLAVOUR SINGLE }	=>	{CAPPUCCINO }	0.0060	0.0721	0.9227
{CAPPUCCINO }	=>	{MINT FLAVOUR SINGLE }	0.0060	0.0766	0.9227
{QUA MINERAL WATER(1000ML) }	=>	{NIRVANA HOOKAH SINGLE }	0.0055	0.1151	0.9600
{NIRVANA HOOKAH SINGLE }	=>	{QUA MINERAL WATER(1000ML)	0.0055	0.0457	0.9600
{POUTINE WITH FRIES }	=>	{NIRVANA HOOKAH SINGLE }	0.0063	0.1298	1.0825
{NIRVANA HOOKAH SINGLE }	=>	{POUTINE WITH FRIES }	0.0063	0.0528	1.0825
{CAPPUCCINO }	=>	{GREAT LAKES SHAKE }	0.0055	0.0710	1.0195
{GREAT LAKES SHAKE }	=>	{CAPPUCCINO }	0.0055	0.0797	1.0195
{GREAT LAKES SHAKE }	=>	{NIRVANA HOOKAH SINGLE }	0.0051	0.0737	0.6148
{NIRVANA HOOKAH SINGLE }	=>	{GREAT LAKES SHAKE }	0.0051	0.0428	0.6148