Andre Schweizer
Prof. Diego Klabjan
IEMS 308 – Data Science & Analytics
February 10th, 2018

HOMEWORK ASSIGNMENT II
Association Rules

I | EXECUTIVE SUMMARY
Dillard's Inc. is a major chain of department stores with around 292 stores across 29 states in the US. It has product offerings in at least 60 different departments and 1960 brands, and is interested in redesigning its stores' planograms to group SKUs that are frequently bought together.

The formulation of association rules is a data-mining technique that establishes co-occurrence relationships among activities performed by individuals or groups. When used in market basket analysis, it reveals affinities between individual products and product groupings. In other words, when given a set of transactions, this method allows us to find rules that predict the occurrence of an item in a basket based on the occurrence of (an)other item(s).

After using this technique to mine rules from Dillard's point-of-sales (POS) data, I found that the vast majority of products either have negligible contributions to the stores' aggregate profit margin, or are present in very few baskets. Hence, given the manpower constraints for relocating SKUs, great part of the products would not be good candidates to modify the planograms. I also found that the majority of co-occurrence patterns between profitable, popular SKUs involve products in the cosmetics and clothing departments, making of them the two areas that Dillard's should focus on first when redesigning its planograms. By incorporating these insights to its planning, Dillard's could increase its customers' basket sizes, growing its volume of sales, and, ultimately, improving its financial performance.

II | PROBLEM STATEMENT
The objective was to investigate Dillard's point-of-sales (POS) data to find evidence of repeated co-occurrences of SKUs in the same baskets to support the formulation of association rules. The strongest of these rules would reveal what SKUs were often bought together, allowing Dillard to optimize its stores' planograms and increase customers' basket sizes by placing frequently co-purchased products close to each other. It is important to note that, due to manpower constraints, Dillard's would only be able to make 20 moves across its chain, and that's it's the first time it is doing this sort of analysis (i.e. it is highly unlikely that SKUs are already appropriately located). Hence, the goal was to suggest the 100 SKUs that were the best candidates to be relocated.

III | ASSUMPTIONS

1. *Frequently co-purchased SKUs are not located close to each other.*
   Due to the facts that there is no information on the current layout of the stores, and that it is the first time Dillard's is conducting this kind of analysis, it was assumed that SKUs aren't grouped appropriately. In other words, if an association rule linking SKU A and SKU B was found, it was assumed that A and B were currently placed in different sections of the planograms, and, therefore, should be relocated.

2. *Lists of transactions and SKUs are exhaustive.*
   It was assumed that all the transactions that occurred within the given time frame were recorded – there are no SKUs that are not in the provided dataset that could potentially be better candidates for relocation.

3. *There are no errors in data entry.* Since POS data is recorded through the cashier machines, I assumed there to be no typos or things of the sort, especially because I wouldn't be able to predict them. If they indeed existed, they could potentially skew the results, but taking into consideration the theoretical rarity of these errors in comparison to the size of the data, it is safe to assume that their impact would be negligible.

4. *Relocation costs are the same for all SKUs.* Recommendations for planogram redesign were based solely on the results offered by the mining of association rules – it assumed that there are no other factors that could potentially influence the decision of which SKUs to relocate.

5. *SKUs are primarily sold for positive profit margins.* It was assumed that SKUs with negative profit margins were either products that were on sale or that had been returned and were later resold at a lower price.

IV | METHODOLOGY

1. *Clean data*
   - I chose to consider only SKUs that were being purchased for the first time due to the complications that the investigation of return transactions would present. Hence, I only kept the rows in the transaction records that were listed as P's under the STYPE column. I then removed the unused columns from that same file in order to have more condensed data.
   - I rearranged the *skstinfo* file so that the SKU numbers were in an actual column rather than being used as indices for the rows.

2. *Select SKUs and transactions*
   - I decided to only keep data pertaining to transactions occurred and SKUs sold at Illinois stores, as I found it to be a state that contained a representative subset of the records in terms of the variety of SKUs sold.
   - I then identified the SKUs with the highest aggregate profit margin (unit profit margin multiplied by the quantity sold), as these were the most relevant to Dillard's underlying goal of profit maximization. Using these results, I cut down the transactions to have only

those that contained SKUs that, together, represented 99% of Dillard's total profit margin in IL.
- For each of the rows, I condensed the STORE, TRANNUM, and SALEDATE columns into a single one that served as a basket ID. This allowed me to identify what SKUs were bought together, to calculate each basket's size and to see in how many baskets each of the SKUs was present in.
- To cut the data to a size that could be managed by the functions that mine association rules, I only kept the 2,000 profitable SKUs in IL that had the highest occurrences in baskets and their corresponding transactions.

3. *One-hot encode transactions*
   - To get the data in the form taken by the *apriori* function, I one-hot encoded the transactions by first grouping them by basket ID.
   - I then made a cross-table with basket IDs as rows and SKUs as columns – it described how many of each SKU was present in each of the baskets.
   - I wasn't interested in whether 2 or 5 of a certain SKU was sold in a particular basket, only in whether that basket contained that SKU or not. Hence, I reduced all values in the cross-table that were above 1 to 1, and converted them into Booleans.

4. *Association rules*
   - I used the *apriori* function in the *mlxtend* library to find the support and frequent items set from the pre-processed data. It returned the support for each of the SKUs present in the transactions.
   - I then used the *association_rules* function from that same library twice to come up with the co-occurrence patterns based on the two different metrics: support and confidence.
   - Note that the minimum support and confidence for a rule was set to 0.001. This number was chosen because the values for these metrics were very low due to the vast number of transactions.

V | ANALYSIS

The process of determining what SKUs were the best candidates for relocation was three-fold. The first two parts were briefly described in section (IV.2): the number of potential candidates was cut down based on the products' relevance to Dillard's finances (via profit margin analysis) and customers' preferences (based on the number of occurrences in baskets). The third part was comprised of extracting insights from the information yielded by the association rules algorithm.

1. *Profit margin analysis*
   Taking a look at the data provided in the *skstinfo* file, which contains the cost and sales price for each SKU, it was possible to see that there was a range in profit margins, including negative ones (for 51,983 of the 164,241 listed). With the assumption that Dillard's would only aim at selling products for a profit, it's reasonable to assume that these SKUs were either on sale or had been returned and had to be resold at a lower price. Because regularly-sold products offer the highest margins, they should be the primary focus of stores when redesigning floorplans. Therefore, SKUs with negative profit margins are bad candidates for relocation.

Furthermore, the aggregate profit margin for each of the SKUs (i.e. the total profit made by Dillard's on that product for the given time period) revealed that 9,516 of the 47,269 SKUs sold in IL for a positive profit composed only 1% of Dillard's profits for this subgroup of products (see Figure 5.1). If the store has limited manpower and can only make a few moves to redesign its planograms, then SKUs that are almost negligible in terms of profit are also not good candidates for relocation.

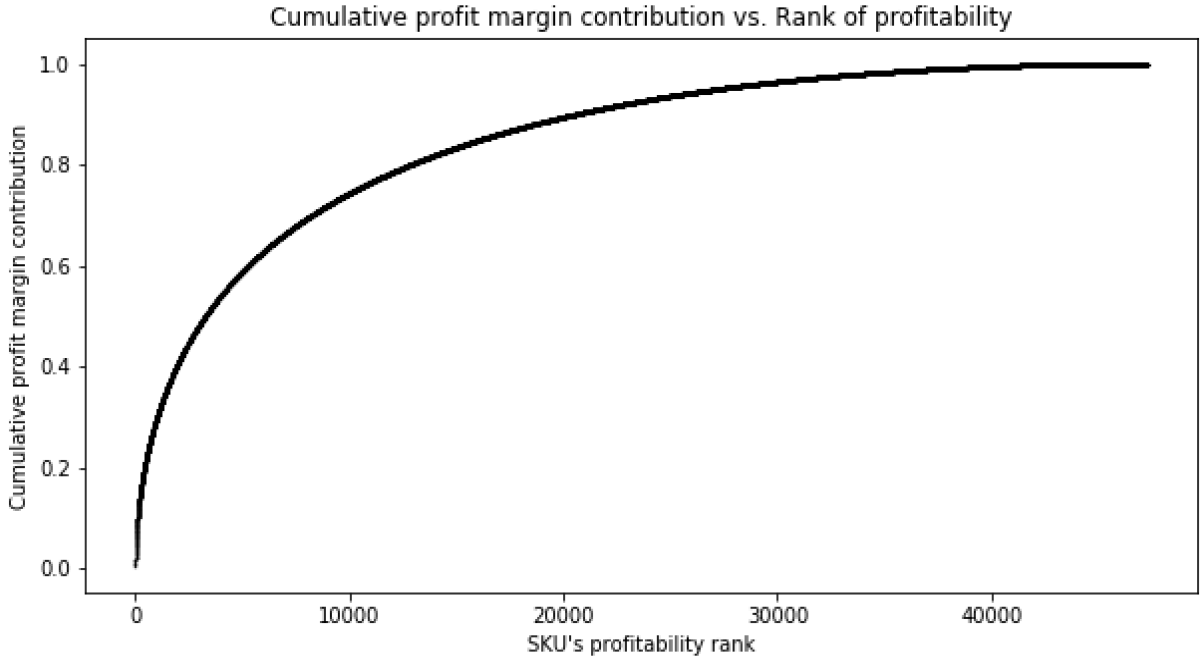

Figure 5.1: Highest-ranking SKUs in profitability contribute to the vast majority of Dillard's aggregate profit margin (e.g. 10K most profitable SKUs in IL amount to approximately 75% of the profit margin)

2. *Frequency analysis*

Something similar happens when looking at the frequency of occurrence of each SKU in baskets: the most popular products occur in many more baskets than the less popular ones (see Figure 5.2). In fact, out of the 37,753 SKUs that were potential good candidates for relocation at this point, only 153 occurred in more than 100 baskets while the most popular product was in 1,064 of them. We see here that a minority of the SKUs are "magnet" products – people buy them a lot even when store planograms aren't optimized. The frequency also tells us about the nature of these products: the most popular ones are high volume, cheaper SKUs while the least popular are more expensive. In a grocery store, the high-frequency products would be things like milk and bread, which people come to buy regardless of where they are placed. Hence, they are great candidates for relocation, as they will very probably catalyze the sales of associated but less popular products when put close to them.

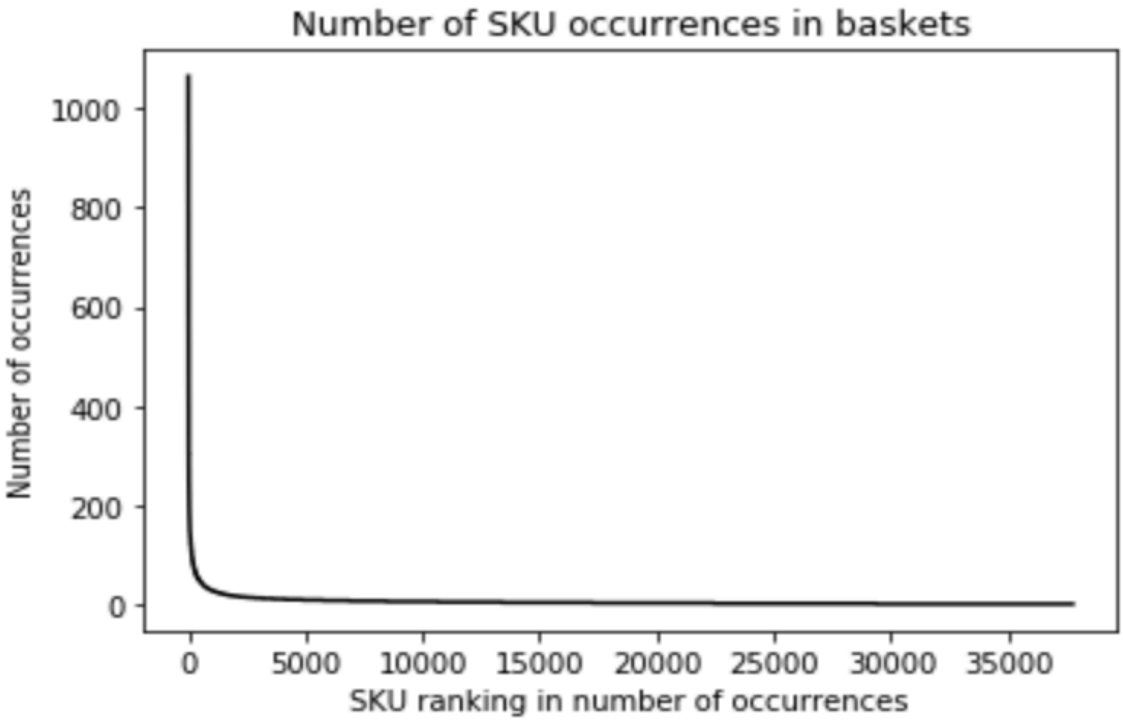

Figure 5.2: Frequency of occurrence in baskets increases exponentially with SKUs' popularity

3. *Association rules*
Table 5.1 shows a summary of the support values that were obtained through the *apriori* algorithm for each of the 2,000 most popular SKUs:

Table 5.1: Summary of individual support values

Average	Standard deviation	Median	Minimum	Maximum
0.0007	0.0013	0.0004	0.0001	0.0305

As can be seen, the support values are very low due to the massive number of baskets and transactions – the SKU with the highest support is only in around 3% of the baskets of the subset being analyzed (i.e. no SKUs with negative profit margins or low frequency in baskets). The median being lower than the average also shows that the distribution of support is skewed: the majority of SKUs have support near the lower end of the range, which is in accordance with Figure 5.2.

Tables 5.2-5.4 show summaries of the support, confidence, and lift values that were obtained for the association rules:

Table 5.2: Summary of support values for association rules

Base metric used	Average	Standard deviation	Median	Minimum	Maximum
Support	0.0018	0.0007	0.0016	0.0010	0.0033
Confidence	0.0020	0.0002	0.0001	0.0001	0.0033

Table 5.3: Summary of confidence values for association rules					
Base metric used	Average	Standard deviation	Median	Minimum	Maximum
Support	0.2313	0.2057	0.1594	0.0332	0.9048
Confidence	0.1665	0.2179	0.0584	0.0039	1.000

Table 5.4: Summary of lift values for association rules					
Base metric used	Average	Standard deviation	Median	Minimum	Maximum
Support	26.0910	22.8081	17.0889	1.9433	70.6052
Confidence	233.2034	440.7141	11.3240	0.4374	3487.2917

From the values above, we can see that, in general, the support-based rules are somewhat stronger than the confidence-based ones. The only instance where this isn't the case is when we look at the lift metric, where confidence-based rules have incredibly high numbers. This is once again due to the low individual support values, as the formula for lift includes those in the denominator. This implies that we can't take the calculated lift results at face value (i.e. we can't expect the co-occurrence of two SKUs in a rule to happen 3,487 more often than if they were bought independently) because we're using a small subset of the data. It is also interesting to see that some of the lift values are less than 1, indicating that the SKUs involved in those rules are substitutes rather than complements – people buy one or the other, but rarely both. Despite those limitations, the values still are good relative measurements and useful for the comparison of the rules.

Looking at the SKUs involved in the support-based rules, we see that 10 out of the 12 are also present in confidence-based ones with the same combinations, and that the 2 SKUs that are left out are in the weakest of the support-based rules. Hence, confidence-based rules give us a good representation of what SKUs would be good candidates for relocation.

By sorting the confidence-based rules by the average of their ranks in the three metrics (support, confidence, and lift), we find which are the strongest overall, and are able to extract the 100 SKUs that we are the most confident to have co-occurrence relationships (see Table A.1 in *Appendix A*). It is interesting to note that the final SKUs selected are contained within only 13 of the 60 existing departments and are of the same 26 brands (out of the 1960 available), mostly in the cosmetics sector – this is evidence that customers in fact buy similar products together, and, subsequently, serves as validation for the association rule mining method used.

VI | CONCLUSIONS

The analysis in (V) reveals insights related to the purpose proposed for this investigation:

- *Dillard's should focus first on the cosmetics and clothing departments when redesigning its planograms*. Good part of SKUs found to be the best candidates for relocation were in these departments. Not only do they have co-occurrence patterns between them, but they are profitable products for the store – even small increases in sales could lead to significant rises in profit.

- *Dillard's should expand its planograms redesign efforts.* There are more than 20 SKUs that could be candidates for relocation. Hence, if the store has the opportunity, it should try to acquire more manpower to conduct those changes as well.

- *Dillard's should also see what SKUs it should not place together.* Lift values for some of the association rules mined were less than 1, indicating that some SKUs are substitutes and should, therefore not be placed close to each other.

- *Dillard's should place "magnet products" far from the entrance/exit of stores.* SKUs with high frequency in baskets act as "magnets" and should, therefore, be located far from the entrance. Since customers will buy those no matter their location, by doing so, Dillard's will force customers to walk more through the stores, exposing them to more product choices and, likely increasing their basket size.

VII | NEXT STEPS

This investigation fomented further questions that would also yield insights relevant to the question originally posed.

- It would be interesting to interview industry experts and come up with Minimum Item Support (MIS) thresholds to be able to identify association rules between products of different natures (i.e. more frequently bought SKUs vs. less popular ones).
- Having information on promotions would help better identify the co-occurrences between products. For example, if a SKU has an increase in sales when another one had its price reduced, that would be strong evidence that we could relate the two by some association rule.
- Different stores have different customer bases. Hence, it would be worthy to look at association rules for stores outside IL, as people in different states are very likely to have different buying patterns. This means that a planogram that is efficient in one location might not be the best for another one.
- It would be of Dillard's interest to evaluate the costs of redesigning its planograms. It was assumed that these were the same for all SKUs, which is not exactly true in reality. For instance, displays for cosmetics are very different and much more expensive than those for clothes – the store would be able to relocate more clothing products than cosmetic ones.

APPENDIX A | 100 BEST SKUs FOR RELOCATION

SKU	DEPT	STYLE	COLOR	SIZE	PACKSIZE	BRAND

39171	800	608E	42SPRING WED	PR SHADES	3	CLINIQUE
180436	6402	U6487	412NAVY	ALL	6	HUE/KAYS
184599	6402	1 NWW048	JET BLACK	ALL	6	ROYCE HO
250896	800	647J	09MATTE CRM	SUPERPOWDE	3	CLINIQUE
258065	7307	L CTU-VA	BLUE	TAIL VAL	2	NOBLE EX
348498	800	6CEG	REPAIR NIGHT	1.7 CREAM	3	CLINIQUE
376836	2105	00 F45UO7	BLACK	L	1	ROUNDTRE
477662	1202	R110	MOONLIGHT	7	1	DIM
792991	6402	U5304	NAVY	2	6	HUE
798962	1202	8 711380	BALLET PIN	8	1	VANITY F
887171	6107	808012	MOSS	BATH	6	CHRISTY
1048962	1202	8 711380	MID BLACK	8	1	VANITY F
1288422	2107	012 517SM3	ASST	BLU	1	SUMMER S
1318962	1202	8 711380	ROSE BEIGE	8	1	VANITY F
1446155	6402	U4631	ESPRESSO	ALL	6	HUE
1618015	1202	75416	BLUE	6	3	MILCO IN
1687668	6402	1505	BLACK	ALL	12	GREAT AM
1688015	1202	75416	PINK	7	3	MILCO IN
1761637	6107	121503	LINEN	BATH	1	NOBLE EX
1788015	1202	75416	BLUE	8	3	MILCO IN
1861637	6107	121503	NATURAL	HAND	1	NOBLE EX
1941658	6402	3543	NAVY	ALL	6	GREAT AM
2028015	1202	75415	BLUE	7	3	MILCO IN
2052331	6402	41 NWW021	NAVY	ALL	6	ROYCE HO
2168966	800	6F4G	MAKEUP BRUSH	CLEANER	3	CLINIQUE
2208015	1202	75415	PINK	8	3	MILCO IN
2218015	1202	75415	PINK	9	3	MILCO IN
2332876	6402	3511	BLACK	ALL	6	GREAT AM
2478302	6107	G131T1	COBBLESTON	HAND TOWEL	6	WESTPOIN
2683090	1202	R314	BLACK	6	1	CABERNET
2754854	2200	2105	09CAMEE	MAQUICOMPL	3	LANCOME
2843090	1202	R314	WHITE	6	1	CABERNET
2853090	1202	R314	WHITE	7	1	CABERNET
2947335	6107	OTTON MICROC	WHEAT	WASH	1	NOBLE EX
3113090	1202	R314	IG/MYST PINK	7	1	CABERNET
3273090	1202	R315	ZE/SAND	7	1	CABERNET
3283090	1202	R315	ZE/SAND	8	1	CABERNET
3334605	6107	73014	1203-CREAM	BATH	6	NOBILITY

3373090	1202	R315	IO/BLOND	7	1	CABERNET
3533090	1202	R316	WHITE	7	1	CABERNET
3573090	1202	R316	ZE/SAND	7	1	CABERNET
3613090	1202	R316	LQ/PORC BLUE	7	1	CABERNET
3672270	2200	3687	MOUSSE	CLARTE	3	LANCOME
3690654	800	68MG	DDML TUBE	DDML TUBE	3	CLINIQUE
3711062	1202	R110	BLACK	7	1	DIM
3772798	2105	11 Y25UR1	MULTI	36	2	MAIN KNI
3903562	6402	2630	BLACK	ALL	6	GREAT AM
3998011	800	6126	CLARIFY	#3 12 OZ	3	CLINIQUE
4376618	1202	12D028	WHITE CABE	XL	1	CABERNET
4438239	1202	3008	NUDE	7	3	MILCO IN
4751496	6402	2816	NAVY	ALL	6	GREAT AM
4771991	6402	3843	KHAKI	ALL	6	GREAT AM
4931662	6402	3543	BLACK	ALL	6	GREAT AM
5102064	6402	U6113	BLACK	ALL	6	HUE
5189905	1202	75416	BEIGE	7	3	MILCO IN
5199905	1202	75416	BEIGE	10	3	MILCO IN
5369905	1202	75415	BEIGE	8	3	MILCO IN
5500969	1202	R111	WHITE	9	1	CABERNET
5709904	1202	75415	BLACK	6	3	MILCO IN
5739904	1202	75415	WHITE	6	3	MILCO IN
5779904	1202	75415	CANDLEGLOW	7	3	MILCO IN
6060179	800	65FX	QUICK BRONZE	TINTED	3	CLINIQUE
6139962	1202	8 711310	BLACK	7	3	CABERNET
6189446	4505	7BRG 581972	UNFR KHAKI	40	1	POLO FAS
6236635	1202	R132	MOCHA	8	1	HENSON-K
6309962	1202	8 711310	WHITE	6	3	CABERNET
6343721	6402	U4631	NAVY	ALL	6	HUE
6359231	7200	150 F23729	EDT SPRAY	1.7 OZ	3	LANCASTE
6402521	6107	11 F11G75	RATTAN	HAND	6	NOBLE EX
6412521	6107	12 F11G75	RATTAN	WASH	6	NOBLE EX
6490353	6107	22 F11G71	LINEN	WASH CLOTH	6	NOBLE CH
6762521	6107	15 F11G75	SLATE	BATH	6	NOBLE EX
6776974	1301	16 P4102E	NAVY/WHT	2X	1	FU DA IN
6832521	6107	12 F11G75	SOFT PINK	WASH	6	NOBLE EX
6979904	1202	75416	CANDLEGLOW	6	3	MILCO IN
6989904	1202	75416	CANDLEGLOW	7	3	MILCO IN

7063806	6107	26028 HAND60	NATURAL	HAND TOWEL	6	CROSCILL
7147624	1202	12D028	CARNIVALSTR	L	1	CABERNET
7248011	800	651302	SCRUF 2 1/2	6OZ.	3	CLINIQUE
7262679	2200	2014	03-LIGHT B	PHOTO CONC	1	LANCOME
7579904	1202	75416	WHITE	11	3	MILCO IN
7602182	1202	8 711310	ROSE BEIGE	6	3	CABERNET
7756167	6402	U6420	488DENIM H	ALL	6	HUE/KAYS
7839904	1202	1007	WHITE	7	3	MILCO IN
7856618	1202	14D027	WHITE CABE	M	1	CABERNET
8122644	6107	0 F30G87	COBBLESTONE	BATH	6	WESTPOIN
8132644	6107	0 F30G87	COBBLESTONE	HAND	6	WESTPOIN
8288357	6402	91 NWW004	BLACK	ALL	6	ROYCE HO
8538445	6402	U5653	BLACK	2	6	HUE
8567094	6107	52589W	BLUE	WASH	6	NOBILITY
8837386	6402	3516	BLACK	ALL	6	ANGEL IN
8939872	2107	13 S57HMX	MULTI	ALL	1	MAIN ING
8976618	1202	11D028	WHITE CABE	M	1	CABERNET
9087401	2301	16D S4101E	RED/WHT	L	1	FU DA IN
9427369	5207	001 TR417L	ORANGE	26	1	MURANO
9709941	1202	R111	ROSEWOOD	7	1	CABERNET
9719941	1202	R111	ROSEWOOD	8	1	CABERNET
9744811	1202	R165	CAMEO	9	1	CABERNET
9789904	1202	3008	WHITE	8	3	MILCO IN
9799904	1202	3008	WHITE	9	3	MILCO IN