Zena's CS251 Project 9

Abstract

This is a UCI data set from 2014 about clients' spending at a wholesale distrubutor. Based on the region names, this seems to be data collected in Portugal. Every row represents how many monetary units within a category of item the given client bought in a year. 440 clients' spendings on each category were recorded. There are 8 attributes: channel (e.g. the client type: "1" for hotel/restaurant/cafe (which they call "horeca" but which I will be referring to as restaurant/hotel) and "2" for retail such as a supermarket), region (Lisbon, Porto, or other, numbered 1, 2, and 3 respectively), fresh products, milk products, grocery products, frozen products, detergents/paper products, and delicatessen (deli products like cold cuts). The numbers under the product types represent monetary units (m.u.), which is a substitute for measuring in regular currency. Wikipedi a defines it as "the change in the utility from an increase in the consumption of that good or service". I will be working with this data by observing relationships between certain features, performing clustering, and doing a PCA analysis.

Problem Statement

As I mentioned before, this data set is from the UCI site, and it was donated in 2014. According to the data homepage, 77 clients are from Lisbon, 47 are from Porto, and 316 are from other regions. 298 are restaurants/hotels and 142 are retail. The main questions I seek to answer are, in general terms, "Is there a relationship between the kind of client and the type of goods of which the most were purchased?", "Is there a relationship between certain product types, whether it be positive or negative correlation?", and "Can we cluster the data into groups based on similar number of products purchased in certain categories?"

Methods

Linear regression

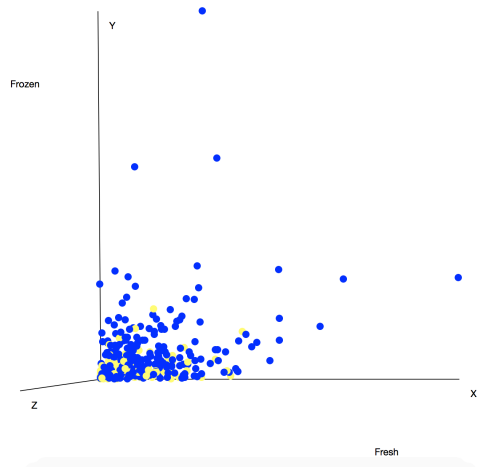
PCA

Kmeans clustering

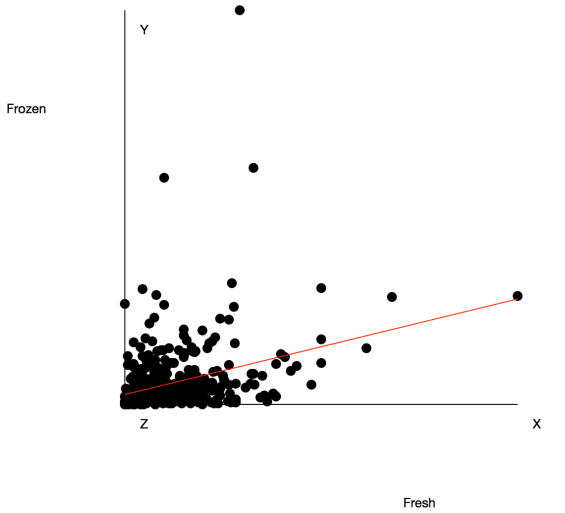
Results

Note: When colored by client type, blue is restaurant/hotel and yellow is retail

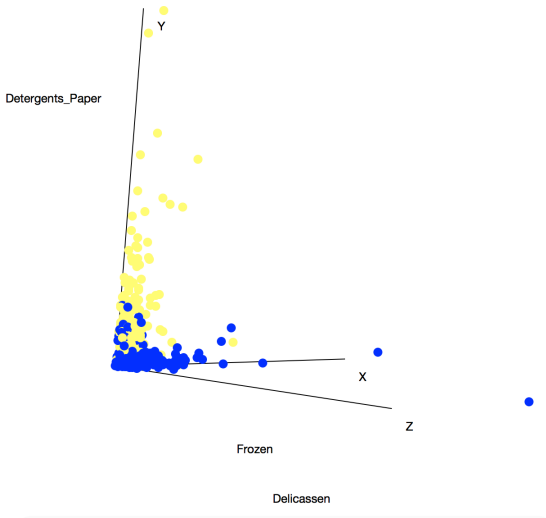
This graph is colored by client type. It seems that retail as a whole buys less while some hotel/restaurant clients buy a lot; they have more outliers and more points in general.



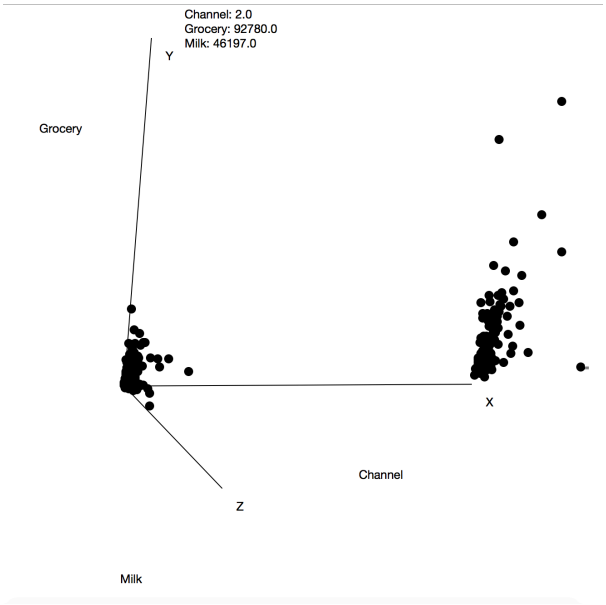
In the next picture, we can see that there is a higher rate of fresh foods than of frozen being bought by clients. Given that most clients are hotels, restaurants, or cafes, this makes sense because they are less likely to stock frozen foods than retailers, since they intend to use the food they obtain more quickly while supermarkets and the like will store food on shelves for days or weeks. There is a lot of data packed at the bottom for clients who buy mostly fresh and almost nothing frozen, which is why the line has the slope it does.



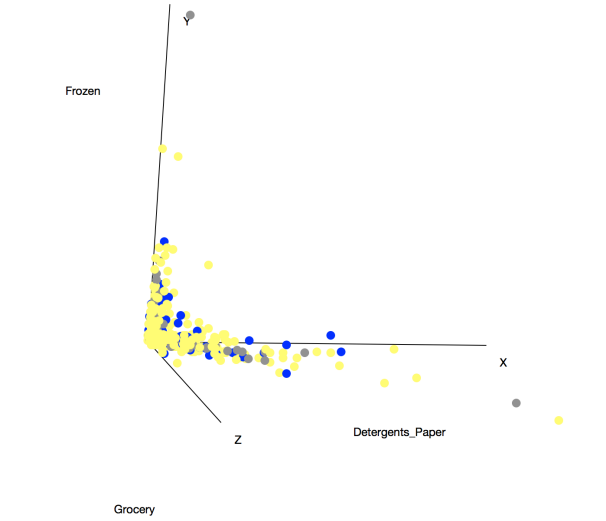
In the next picture, I've plotted the frozen foods vs the deli meats vs the toiletries-type products. Retail seems to buy a lot more toiletries than restaurants or hotels...which makes sense because the customers are not at hotels or restaurants to get those kinds of things. Meanwhile, the deli meats seem to be bought about equally, except for some outliers in restaurant/hotels that may perhaps be actual delis or sub shops and therefore need a lot.



In the next picture, we see a result where retail (the right side) actually buys more than restaurants/hotels. This may be because milk products are important and abundant in Europe, and many people want to buy them to have daily access to them, while restaurants/hotels are not somewhere a customer would go to every day to get dairy from. The information displayed is about the most upper-right point, incidentally.



Also, the regions seem to have an impact on the kinds of products bought, as well. In the picture below, blue is Lisbon, gray is Porto, and yellow is other. Lisbon is a southern coastal city (the largest in Portugal), while Porto is a northern coastal city (the second largest). Other regions seem to buy a lot more of everything in general, and they also have the most outliers, which makes sense because they have more clients than Lisbon and Porto.



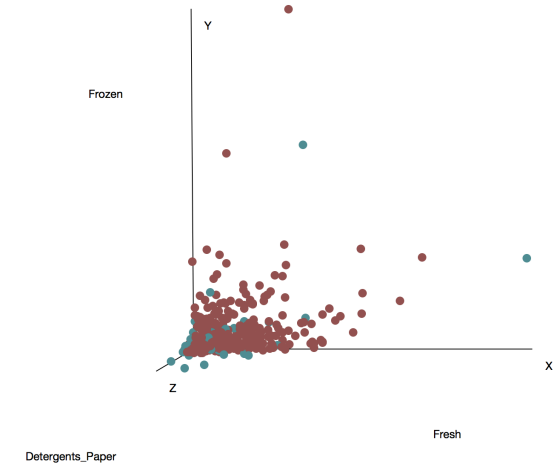
As the PCA analysis of all of the features below shows, the first two eigenvectors are the most important ones as they make up 90% of the variation in the data. The chief feature in the first one is the client type while the chief feature in the second is the region. Looking at the colors of the previous plots, the reason behind this variation is clear; we can see that these two features have a large impact on the amount of various items clients buy.

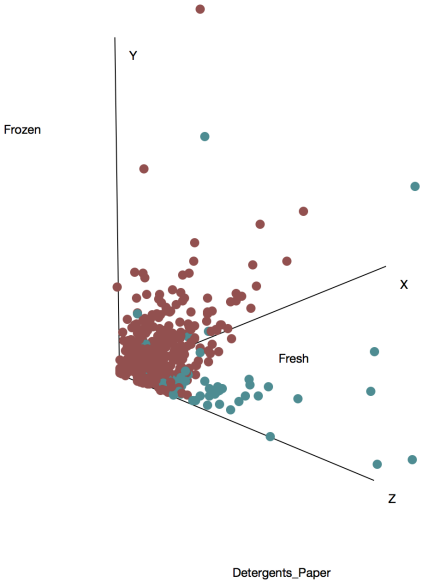
PCA Eigenvectors

Eigenvec	Eigenval	Energy	Channel	Region	Fresh	Milk	Grocery	Frozen	tergents_Pap	Delicassen
PCA00	0.23	0.55	-0.96	-0.13	0.04	-0.1	-0.14	0.03	-0.16	-0.01
PCA01	0.15	0.35	0.12	-0.99	-0.03	0.01	0.03	-0.0	0.04	-0.01
PCA02	0.02	0.04	0.18	-0.0	-0.43	-0.52	-0.47	-0.23	-0.46	-0.19
PCA03	0.01	0.03	-0.17	0.03	-0.77	0.12	0.29	-0.29	0.42	-0.13
PCA04	0.01	0.01	0.03	0.0	-0.47	0.31	-0.13	0.67	-0.27	0.39
PCA05	0.0	0.01	0.0	0.01	-0.03	-0.63	0.19	0.61	0.35	-0.27
PCA06	0.0	0.0	-0.01	-0.0	-0.01	-0.46	0.22	-0.21	0.06	0.83
PCA07	0.0	0.0	0.0	-0.0	-0.03	-0.04	0.76	-0.01	-0.63	-0.18

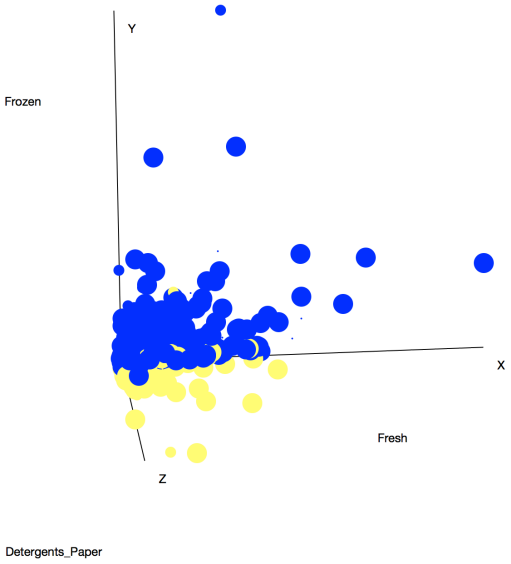
OK

The two pictures below show the kmeans clustering result for all of the features except region and client type. We can see that the result of clustering this data has produced similar results in relation to the third picture. By leaving out the two biggest sources of variation and observing the clustering, it is clear that they are still similarly divided into clusters as if the cluster colors were still based off of region/client type.





Compare these two clustering results to the plot of the data before clustering, colored by client type and sized by region:



This plot shows the original clients (blue is restaurant/hotel and yellow is retail). Those that buy the most of frozen food, for example, are usually restaurants/hotels, as we can see from this result. But the clustering result has a few of these same points colored differently (we can assume that brown is restaurant/hotel and green is retail). Therefore, the clustering results are imperfect, as they are showing some inaccuracies.

Conclusion

For this project, I have examined the relationship between types of products bought and between types of products and the clients who bought them. In general, fresh products are bought more than frozen ones. Retail clients also buy a lot more detergents/paper, groceries, and milk products, while restaurants/hotels buy a bit more delicatessen and fresh foods. Restaurant/hotels seem to produce the most outliers due to high amounts of a certain kind of product they may buy. Regions also affect the results, as the "other" regions seem to buy the most, and the retailer sells to mostly these clients, from what it seems. Lisbon buys the next most while Porto buys the least.

Acknowledgements:

Thank you to my high school buddy John for explaining some economics things