Outline

- 1. Exploratory analysis
- 2. Predicting future sales
  - Identified trends of customers and factors increasing sales using ml model
  - Dimension reduction followed by customer segmentation and proposed recommendation model
- 3. Predicting and identifying factors for churn

E-commence dataset test

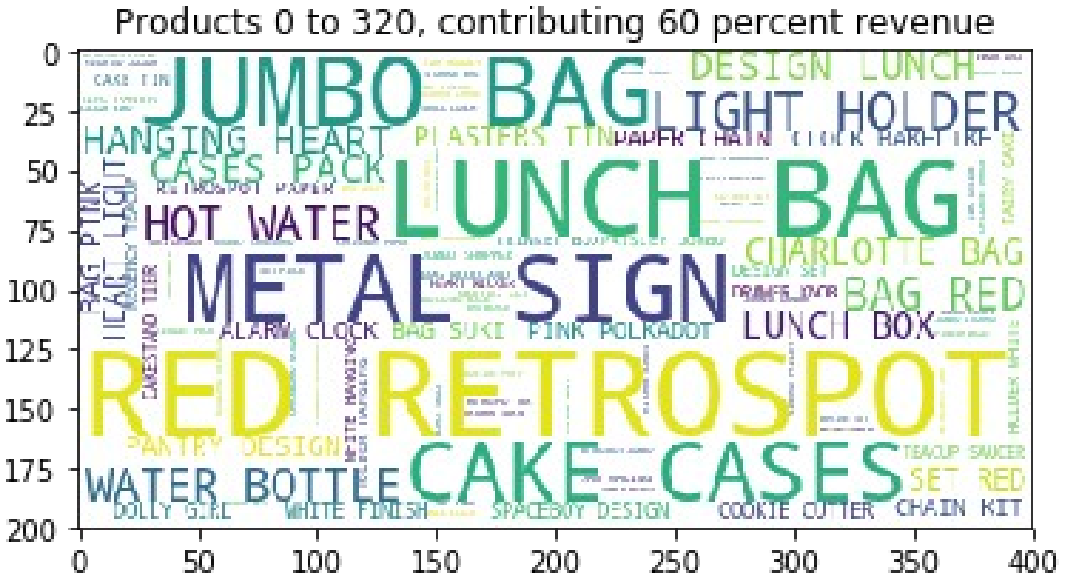
- Lee Xiong An
- https://www.kaggle.com/carrie1/ecommerce-data

Data format given

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
5000	537026	22551	PLASTERS IN TIN SPACEBOY	12	12/3/2010 16:35	1.65	12395.0	Belgium
5001	537026	85099B	JUMBO BAG RED RETROSPOT	10	12/3/2010 16:35	1.95	12395.0	Belgium
5002	537026	22355	CHARLOTTE BAG SUKI DESIGN	10	12/3/2010 16:35	0.85	12395.0	Belgium
5003	537026	84992	72 SWEETHEART FAIRY CAKE CASES	24	12/3/2010 16:35	0.55	12395.0	Belgium
5004	537026	POST	POSTAGE	2	12/3/2010 16:35	18.00	12395.0	Belgium

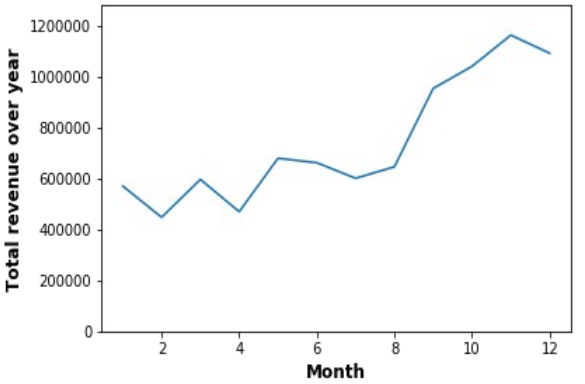
Exploratory analysis : Wholesale retail

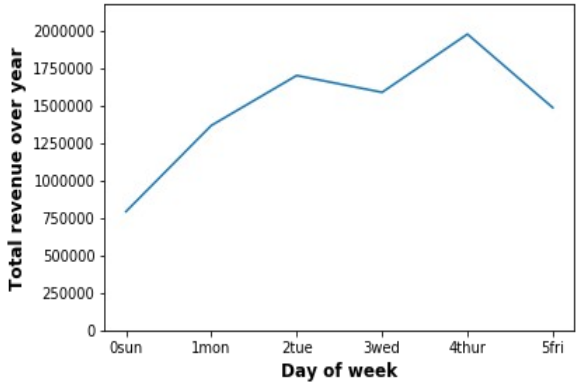
- 540K entries of purchases from Dec 2010- Dec 2011.
- 4328 Customers
- 3900 products (listed in word cloud below)
- Retro store, lifestyle store

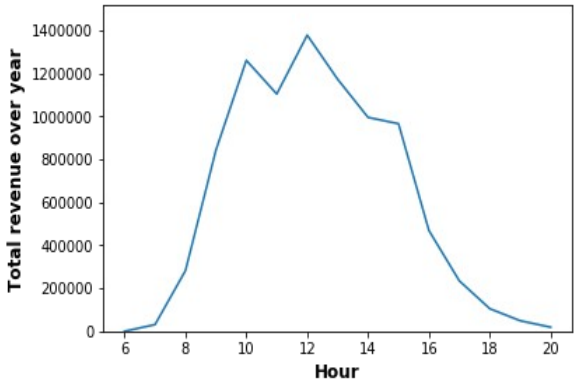


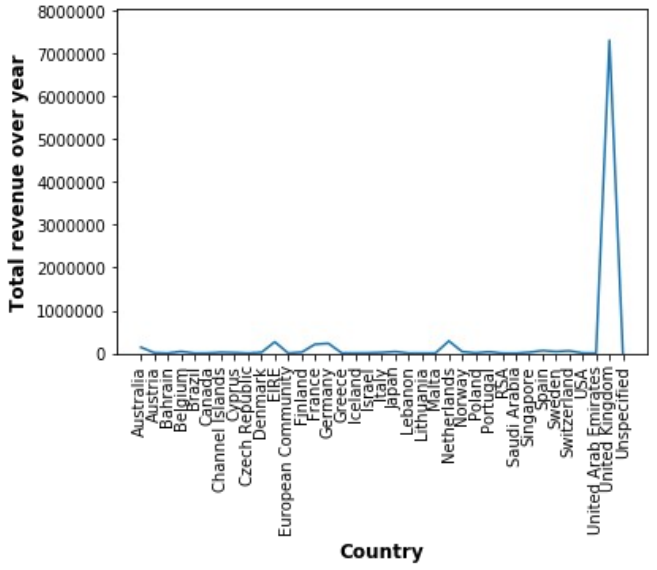
Exploratory analysis : Total revenue by month, weekday

- Unclear total revenue in first half and second half of year differ so greatly
- Sales mostly similar during weekdays, lower for Sundays
- Sales mostly from 8am-4pm
- Data from UK (mostly)





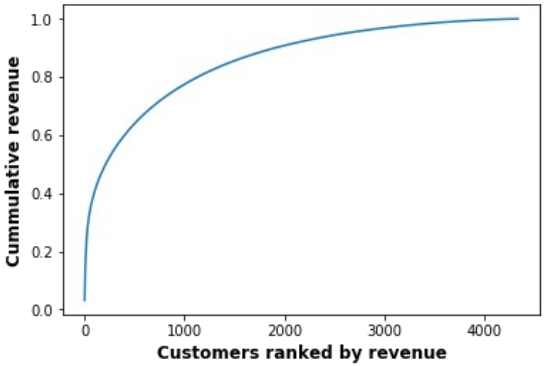


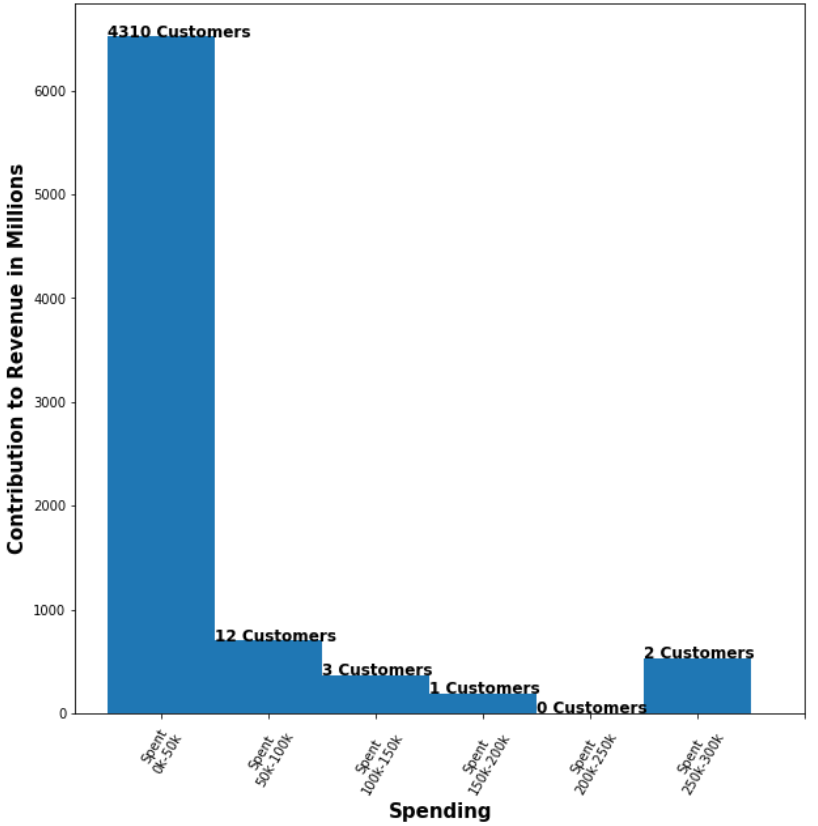


Exploratory analysis: Distribution of customers

- Largely skewed
- A small group of customers provide much of the revenue

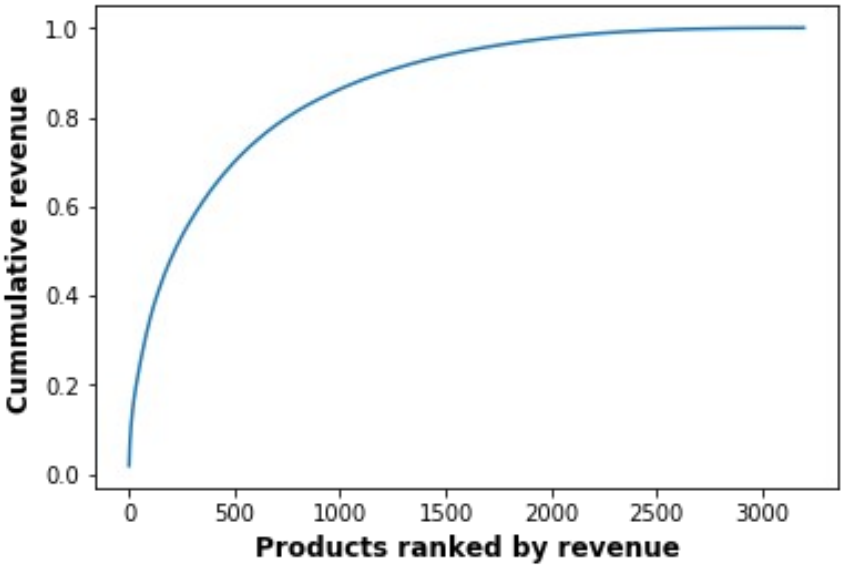
```
0 Top 0.01 percent of customers,3.0 percent of profit
4 Top 0.1 percent of customers,12.0 percent of profit
43 Top 1 percent of customers,30.0 percent of profit
87 Top 2 percent of customers,37.0 percent of profit
218 Top 5 percent of customers,49.0 percent of profit
437 Top 10 percent of customers,60.0 percent of profit
874 Top 20 percent of customers,74.0 percent of profit
2186 Top 50 percent of customers,92.0 percent of profit
```





Exploratory analysis: Distribution of revenue per product

- Largely skewed
- A small group of products provide much of the revenue




```
0 Top 0.01 percent of products,2.0 percent of profit
3 Top 0.1 percent of products,6.0 percent of profit
32 Top 1 percent of products,20.0 percent of profit
64 Top 2 percent of products,28.000000000000004 percent of profit
160 Top 5 percent of products,44.0 percent of profit
320 Top 10 percent of products,59.0 percent of profit
640 Top 20 percent of products,76.0 percent of profit
1600 Top 50 percent of products,95.0 percent of profit
1920 Top 60 percent of products,97.0 percent of profit
1853
threshold= 600 , 0.58 of products account for 0.969 of revenue
```

Predict and identifying trends for future customer behavior

- Largely skewed dataset, hence predict their log sales (magnitude) and is more qualitative than quantitative
- Training data: Revenue from months Dec 2010 – Apr 2011, products + customer embeddings[1]
- Target variable : sales on May 2011
- Revenue from Jun-Dec 2011 used to measure performance

1: Details next slide

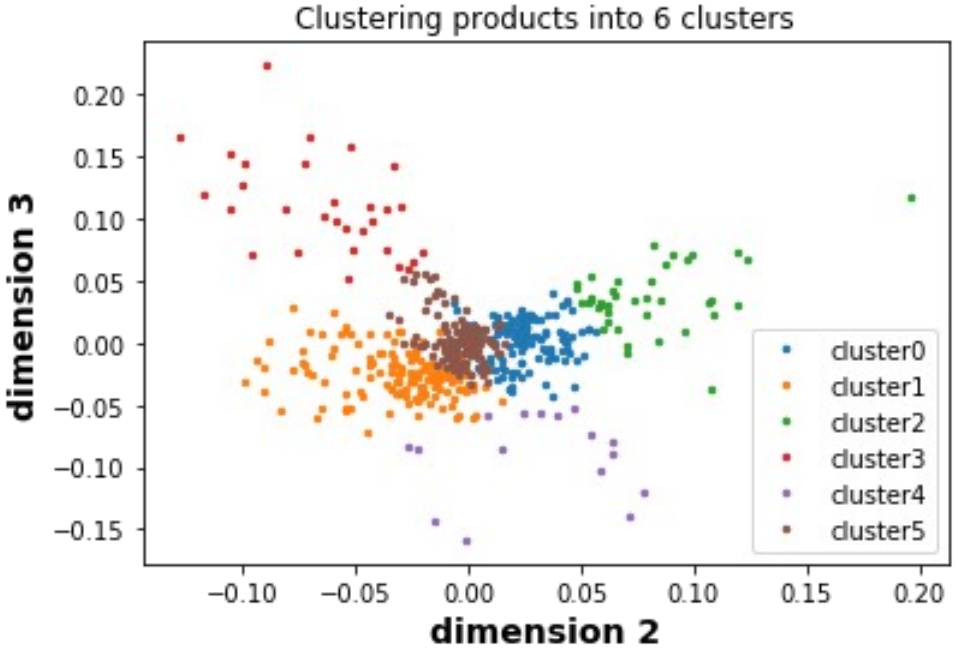
Reducing customer dimension

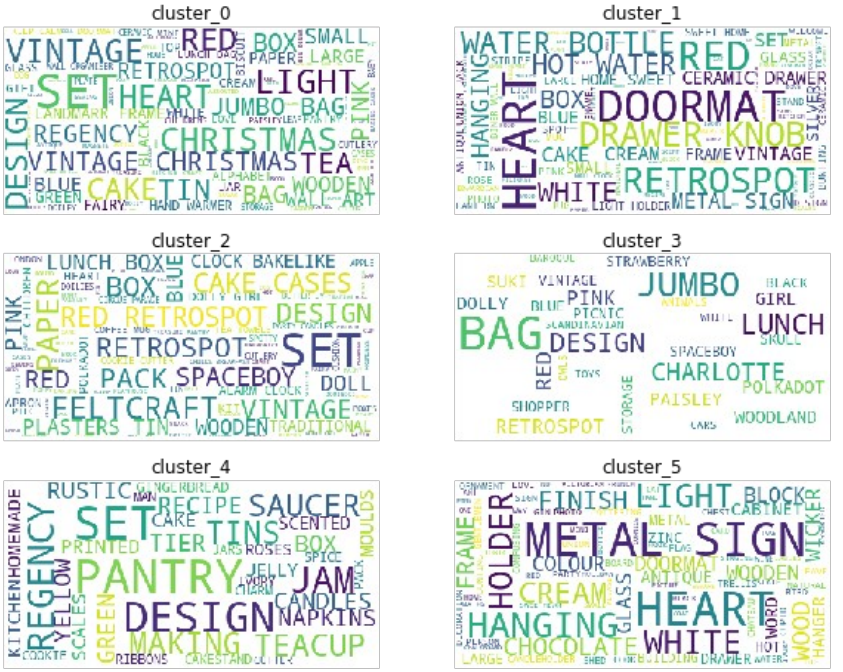
Products/ Customers	1	2	3	...	1800
1	...	...	...	...	...
2	...	...	...	...	...
3	...	...	...	...	...
...	...	...	...	...	...
	...	...	...	...	...
	...	...	...	...	...
	...	...	...	...	...
	...	...	...	...	...
5000	...	...	...	...	...

- 5000 customers
- 1800 products used comprising of 97% revenue
- Matrix contains products brought during training phase
- Matrix factorization for both customers and products using SVD or other methods
- Matrix too high dimension to be used directly

Reducing product dimension

- Dimensionality reduction to be used, I used three dimensions* out of 1800 dimensions.
- Each customer and product has three associated dimension
- Can be clustered into six clusters visualized with dimensions 2-3



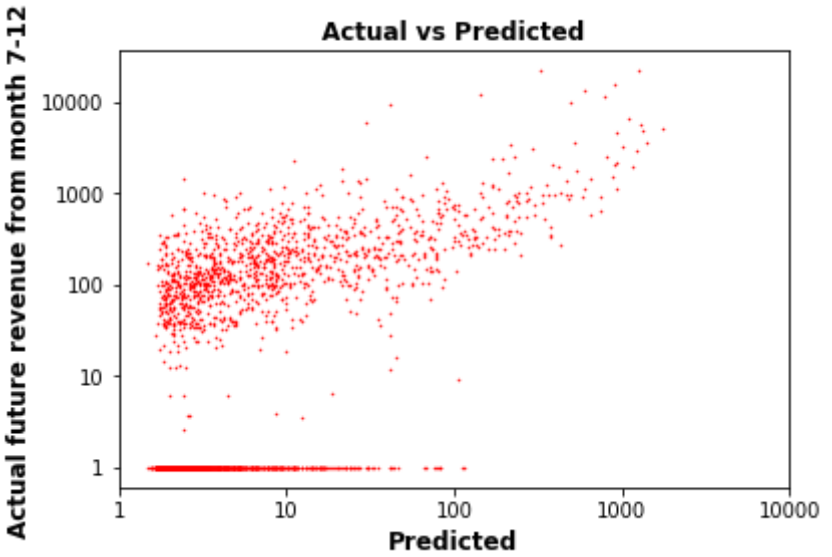


1: Only the first three dimensions are useful for the revenue prediction model in next slidef found through model tuning

Model : predicting future revenue

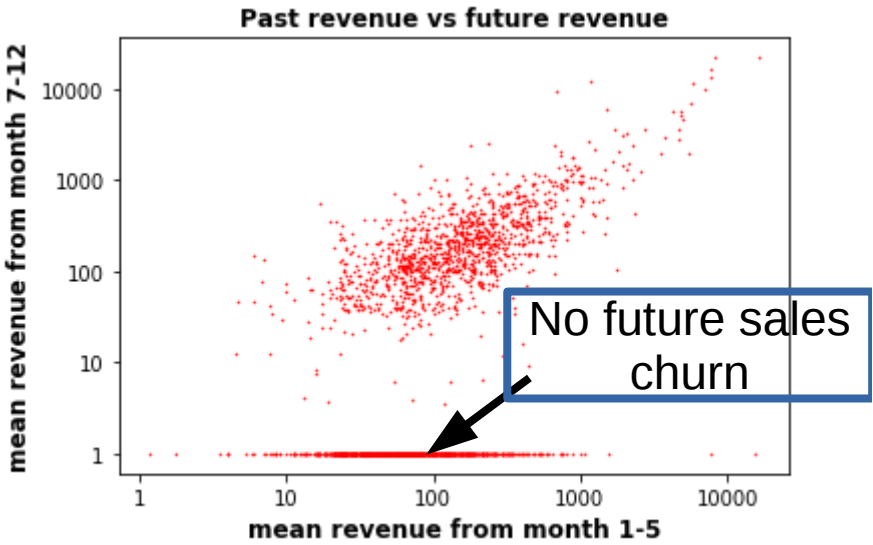
- Xgboost model has correlation ~0.5 on data from Jun-Dec 2011 training on Dec 2010 – May 2011
- Two types of features are positively correlated :
  1. previous revenue data
  2. customer embeddings (dimension reduction)

	Feature	Gain
0	mean_log_rev_train	43512.985171
1	Revenue_2011_3	23805.277641
2	svd_0	18979.759464
3	svd_1	15603.714923
4	Revenue_2011_2	14743.079521
5	svd_2	14398.455595
6	Revenue_2011_1	13450.233664
7	Revenue_2011_4	13304.948912
8	Revenue_2010_12	12019.267700
9	Discount_given	11727.040912
10	Country	2618.988878



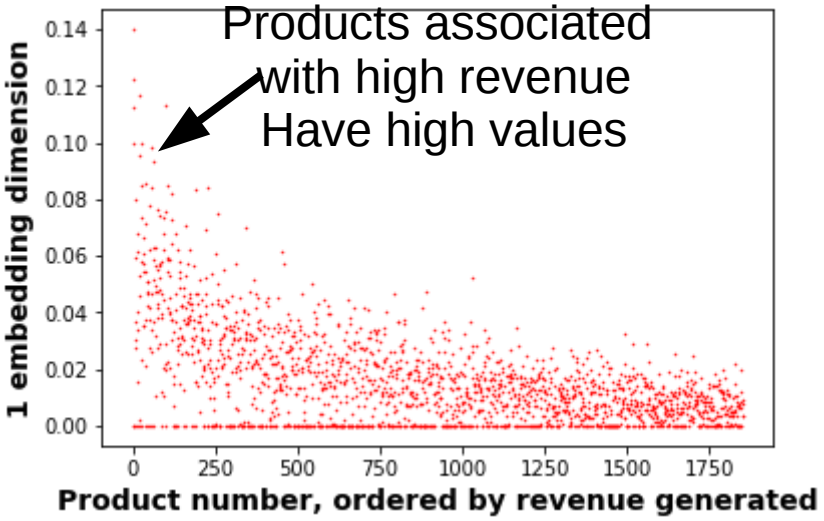
Model features : Previous Revenue Data

- There is a linear trend where large past revenue is associated with large future revenue per customer
- However there is a large group of customers who stop buying (churn) to elaborate at last section



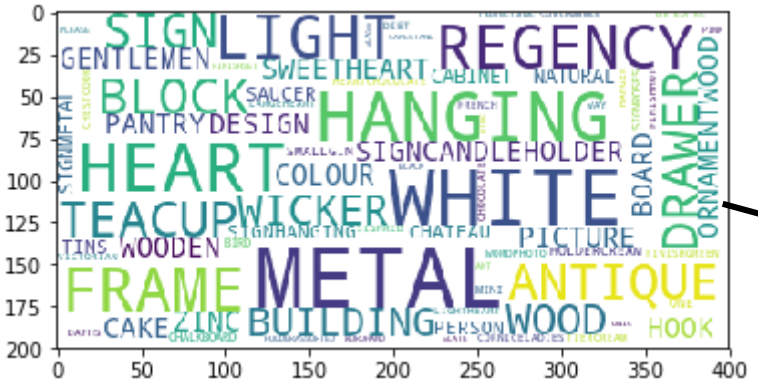
Customer/product embeddings dimension 1

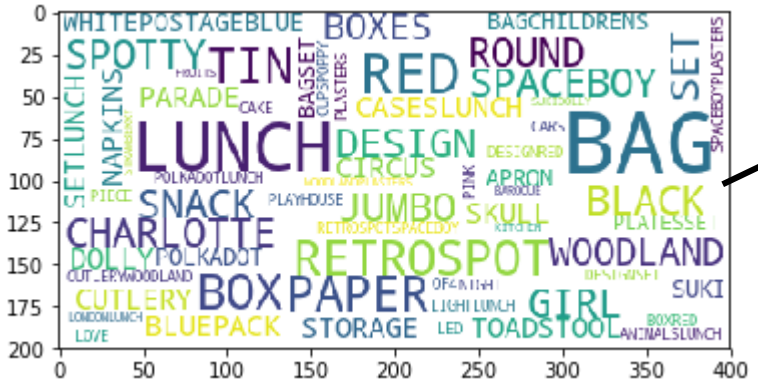
- Customer product matrix is too large to use for prediction, hence dimension-reduced with svd
- The 1st - 4th dimensions are significant features for model
- High values of 1st dimension associated with higher revenue, possibly ‘fast moving consumer goods’

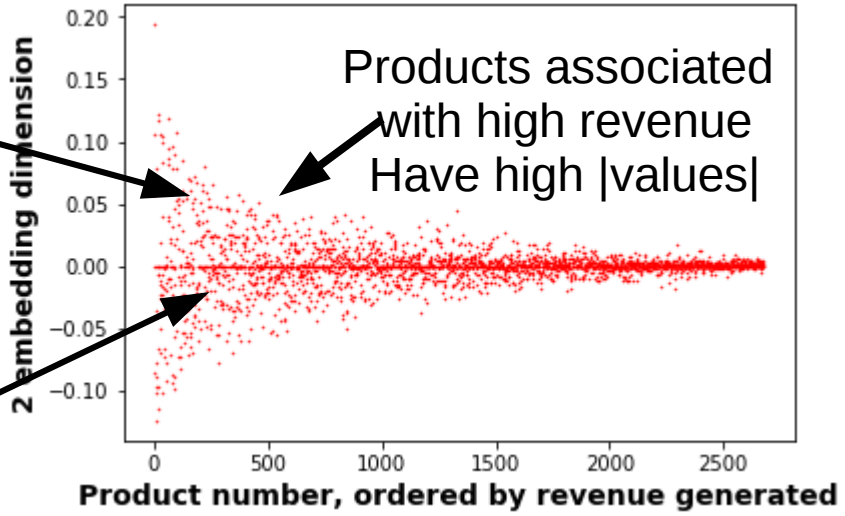


Customer/product embeddings dimension 2

- Large absolute values of 2nd dimension associated with higher revenue, and there seems to be two classes of products with positive and negative values respectively with associated word clouds

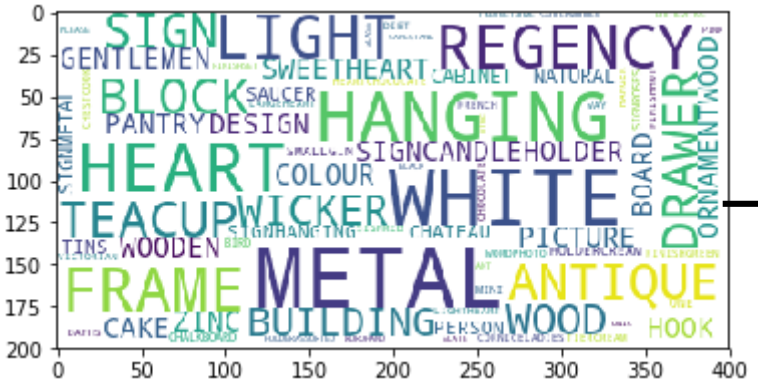






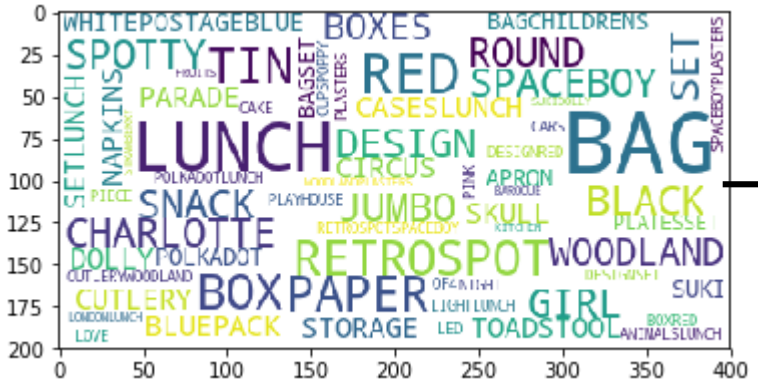
Customer/product embeddings dimension 2

- Positive dimension 2 values associated with decorative household objects, while negative dimension values associated with ‘standard’ products of the company



Top 20 products

['REGENCY CAKESTAND 3 TIER', 'CREAM HANGING HEART T-LIGHT HOLDER', 'ASSORTED COLOUR BIRD ORNAMENT', 'WOOD BLACK BOARD ANT WHITE FINISH', 'SET OF 3 CAKE TINS PANTRY DESIGN ', 'VICTORIAN GLASS HANGING T-LIGHT', 'HEART OF WICKER LARGE', 'HEART OF WICKER SMALL', 'GIN + TONIC DIET METAL SIGN', 'ROSES REGENCY TEACUP AND SAUCER ', 'NATURAL SLATE HEART CHALKBOARD ', 'PLEASE ONE PERSON METAL SIGN', 'RED HANGING HEART T-LIGHT HOLDER', 'CREAM SWEETHEART MINI CHEST', 'COOK WITH WINE METAL SIGN ', 'WOODEN PICTURE FRAME WHITE FINISH', 'GREEN REGENCY TEACUP AND SAUCER', 'WOODEN FRAME ANTIQUE WHITE ', '3 DRAWER ANTIQUE WHITE WOOD CABINET', 'HOME BUILDING BLOCK WORD']



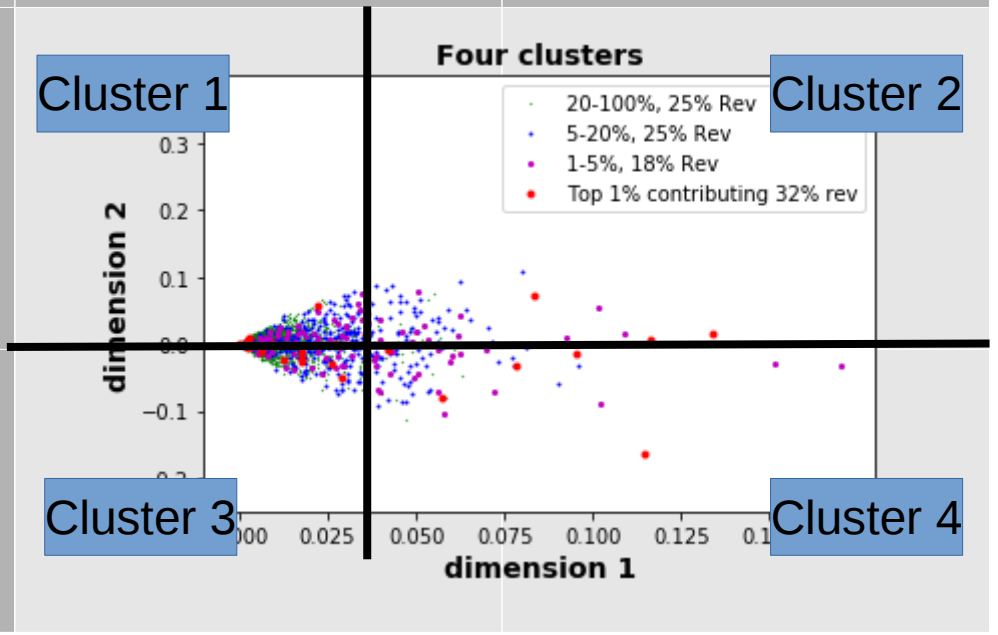
Top 20 products

['JUMBO BAG BAROQUE BLACK WHITE', 'POSTAGE', 'BLUE 3 PIECE POLKADOT CUTLERY SET', 'LUNCH BAG RED RETROSPOT', 'SPACEBOY LUNCH BOX ', 'ROUND SNACK BOXES SET OF4 WOODLAND ', 'JUMBO STORAGE BAG SUKI', 'DOLLY GIRL LUNCH BOX', 'RED TOADSTOOL LED NIGHT LIGHT', 'LUNCH BOX I LOVE LONDON', 'LUNCH BAG SUKI DESIGN ', 'LUNCH BAG BLACK SKULL.', 'ROUND SNACK BOXES SET OF 4 FRUITS ', 'LUNCH BAG CARS BLUE', 'PACK OF 72 RETROSPOT CAKE CASES', 'LUNCH BAG PINK POLKADOT', 'LUNCH BAG SPACEBOY DESIGN ', 'LUNCH BAG WOODLAND', 'PLASTERS IN TIN SPACEBOY', 'PLASTERS IN TIN WOODLAND ANIMALS']

Customer segmentation from dimensions 1 and 2

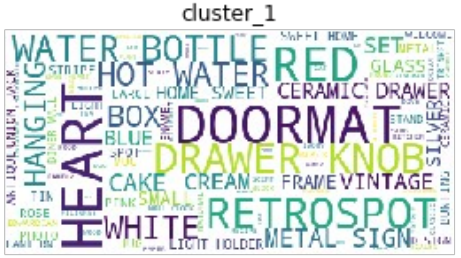
- Dimensions 1 and 2 are most associated with future sales segmenting customers by sales, and large customers are found in all four clusters

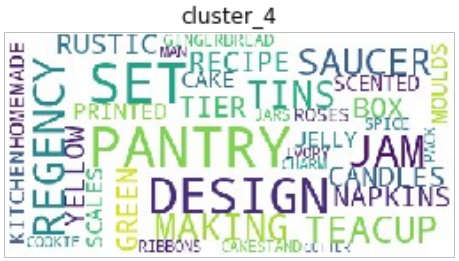
	Customers who buy less fast moving consumer goods	Customers who buy mostly fast moving consumer goods
Customers who buy mostly decorative household objects	Cluster 1	Cluster 2
Customers who buy mostly standard objects of company	Cluster 3	Cluster 4

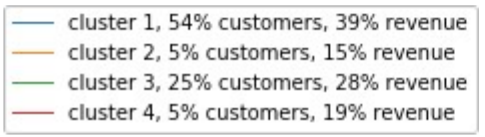


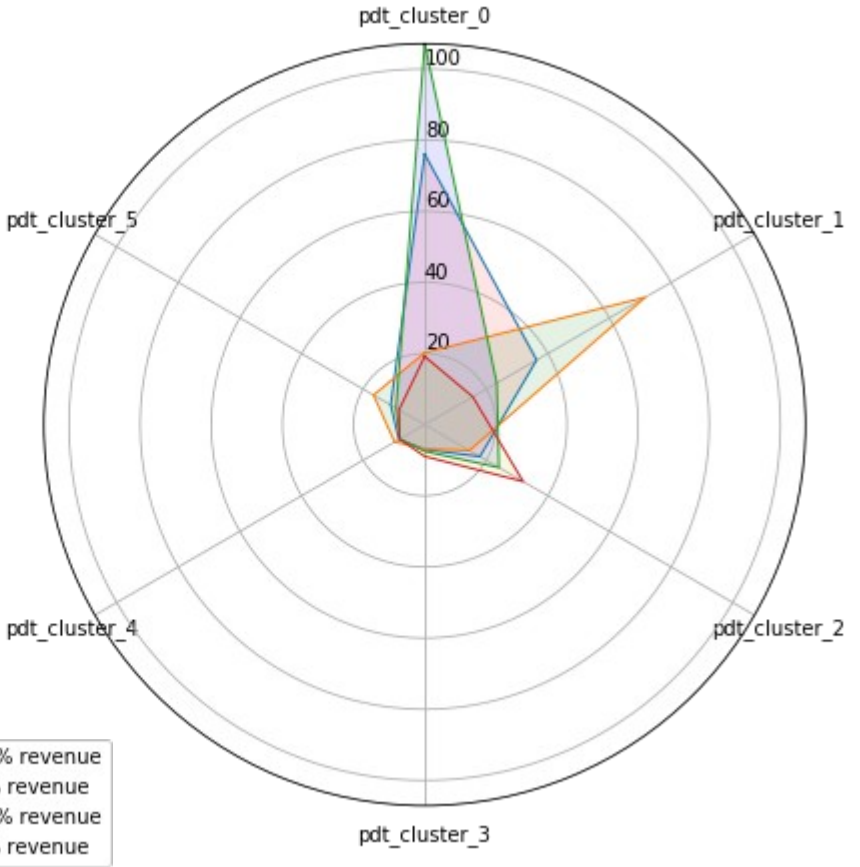
What do each segment of customers like?

- What products do customers segments prefer?









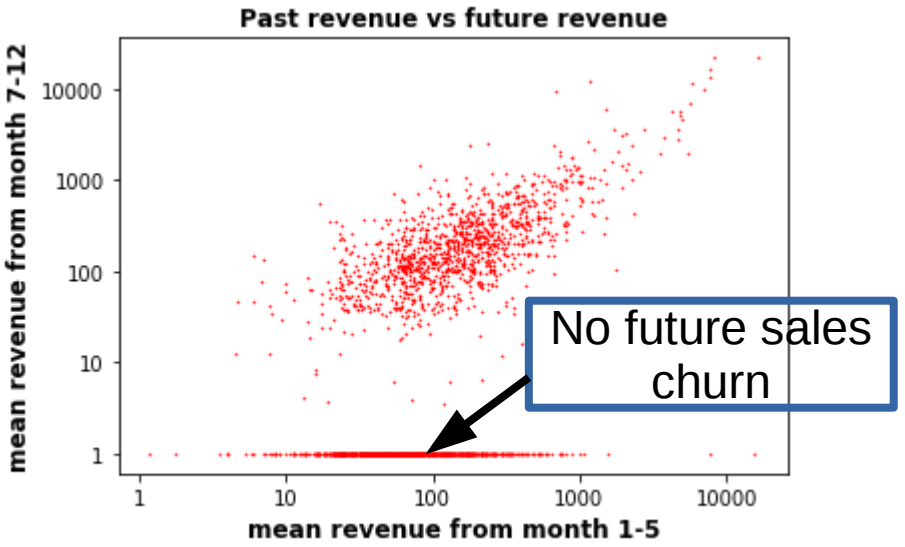
Recommendation system for each customer

- Recommendation example , customer 1333, largest customer from cluster 2. (#7 largest customer total)

	StockCode_NR_int	Description	labels	affinity_to_customer1333	brought_customer1333
576	576	[METAL SIGN HER DINNER IS SERVED]	1	0.012934	1
468	468	[HANGING METAL HEART LANTERN]	1	0.012386	0
532	532	[LAUNDRY 15C METAL SIGN]	1	0.012270	1
227	227	[AREA PATROLLED METAL SIGN]	1	0.012186	0
389	389	[LANTERN CREAM GAZEBO]	1	0.011772	0
244	244	[3 HEARTS HANGING DECORATION RUSTIC]	1	0.011368	0
189	189	[DOORMAT BLACK FLOCK]	1	0.011300	0
136	136	[3 HOOK PHOTO SHELF ANTIQUE WHITE]	1	0.011201	1
137	137	[BLACK HEART CARD HOLDER]	1	0.011182	0

Predicting Churn

- Predicting churn
- Unbiased regularized logistic regression, using first 6 months of data as training, and labels are whether customer has churned in next 6 months



Predicting Churn : model

- Model has test AUC : 0.72
- Most important factors are sales in earliest preceding months – higher log(sales) in most recent months decreased $e^{-0.65} = 0.5$ log odds ratio compared to second earliest preceding month $(0.7)^1$.

Generalized Linear Model Regression Results

Dep. Variable:	Churn	No. Observations:	2710
Model:	GLM	Df Residuals:	2702
Model Family:	Binomial	Df Model:	7
Link Function:	logit	Scale:	1.0
Method:	IRLS	Log-Likelihood:	-1403.5
Date:	Sun, 25 Aug 2019	Deviance:	2807.0
Time:	22:07:56	Pearson chi2:	2.60e+03
No. Iterations:	6		

	coef	std err	z	P>|z|	[0.025	0.975]
log_revenue_2010_12	-0.2891	0.059	-4.878	0.000	-0.405	-0.173
log_revenue_2011_2	-0.2867	0.060	-4.787	0.000	-0.404	-0.169
log_revenue_2011_3	-0.3275	0.060	-5.450	0.000	-0.445	-0.210
log_revenue_2011_4	-0.3835	0.061	-6.246	0.000	-0.504	-0.263
log_revenue_2011_5	-0.6465	0.064	-10.078	0.000	-0.772	-0.521
mean_log_rev_firsthalf	-0.0690	0.078	-0.882	0.378	-0.222	0.084
constant	-1.2697	0.061	-20.872	0.000	-1.389	-1.151
svd_0	-0.5008	0.103	-4.865	0.000	-0.703	-0.299

1. aware that the representation is problematic, will improve on it

Predicting Churn : model for top 20% customers

- Model has test AUC : 0.66
- Only important factor are average sales. Larger sales
- Need to keep customers engaged over a longer period of time, and preceding month is less important compared to overall time frame.

1. aware that the representation is problematic, will improve on it

Conclusions of analysis

- Customers/products mostly obey the 80/20 rule.
- Customers and products segmented into 4 and 6 categories respectively, where strategies can be applied to each 6 customer segments.
- Predicting churn, largely predicted by continual and recent engagement of customer.

Areas not worked on

- Data integrity not checked, minimal cleaning done due to time factors
- Negative sales – returns or discount not considered in this part of analysis
- Assumed time frame of 6 months is reasonable frame of time and customer behaviour changes over large time scales
- Curse of dimensionality, distances work poorly if dimensions > 5 which I did not really investigate further
- High dimension visualization is complex