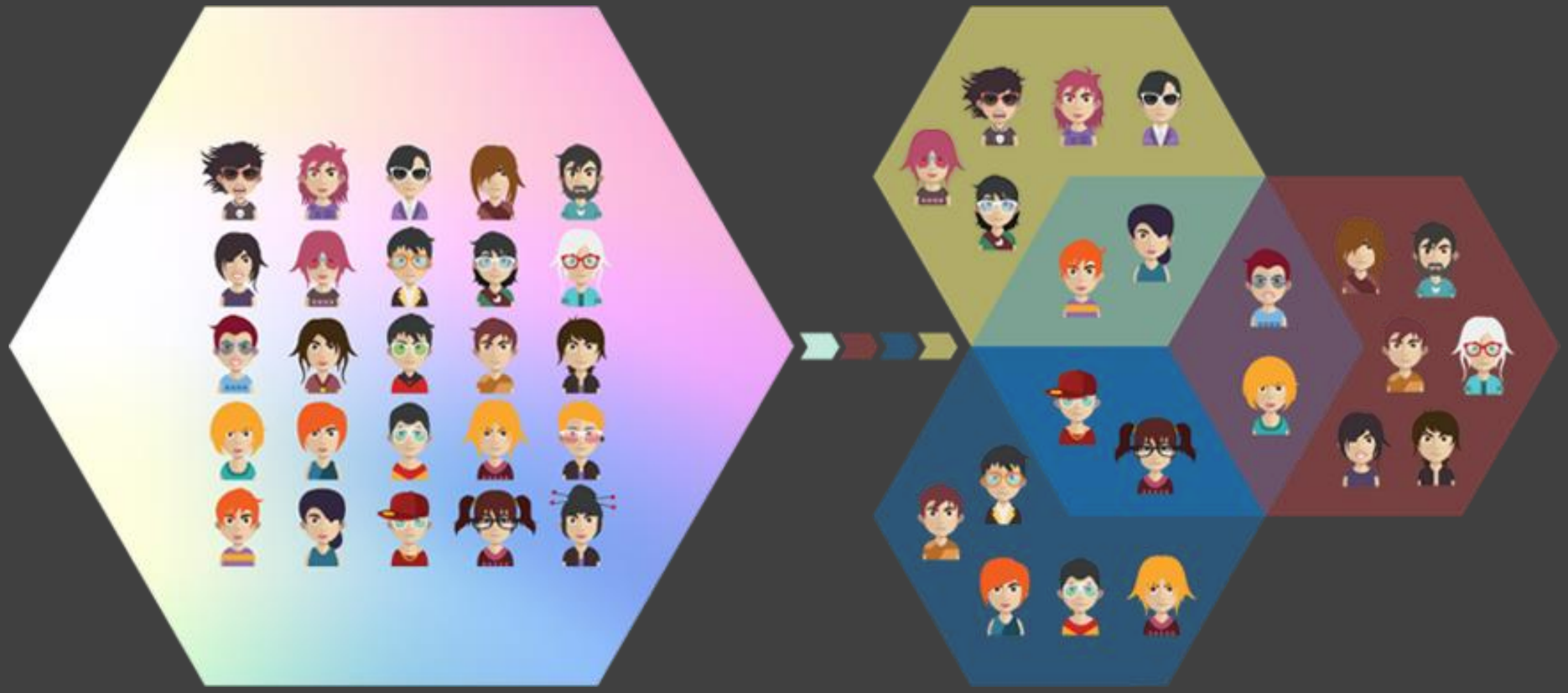
Clustering Considerations
for Customer Segmentation

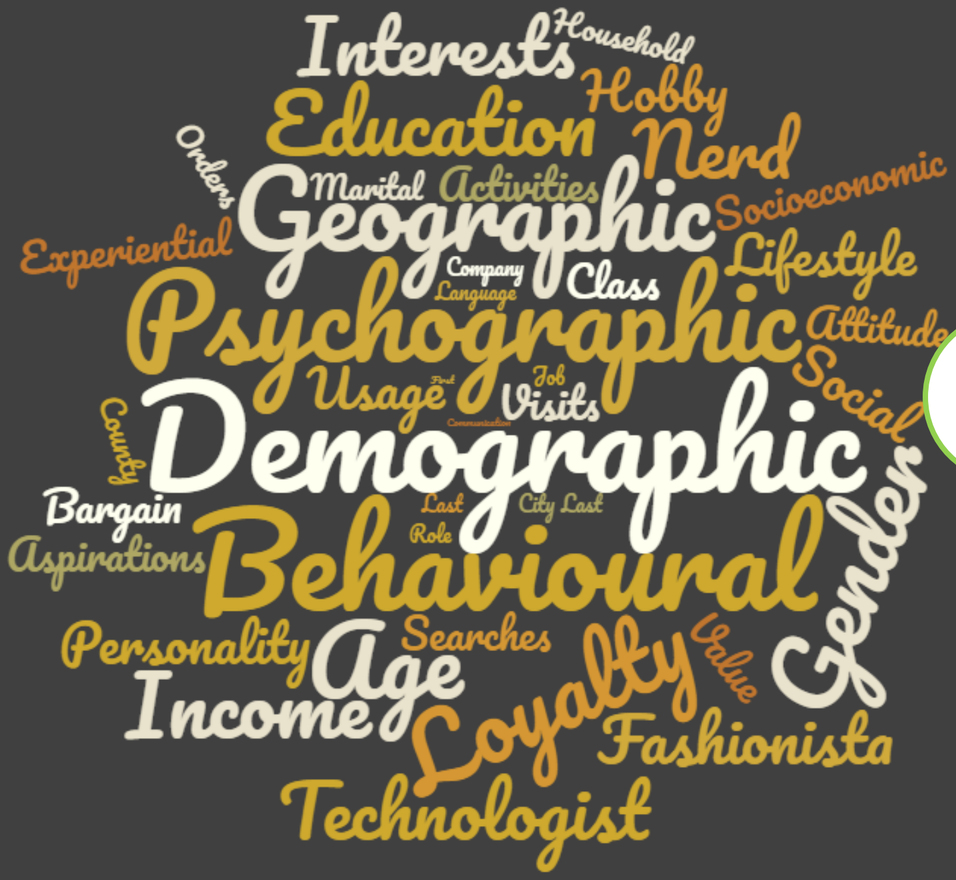
Regina Berengolts
Head of Data Science @ TVSquared

A Bit About Me

Customer Segmentation



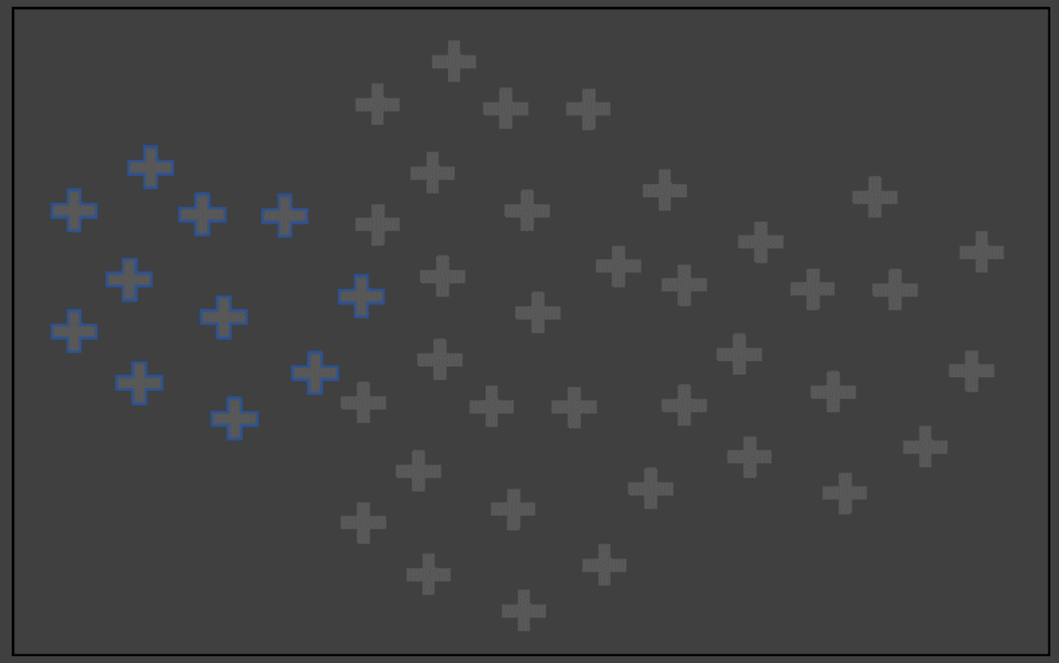
Segmentation Options

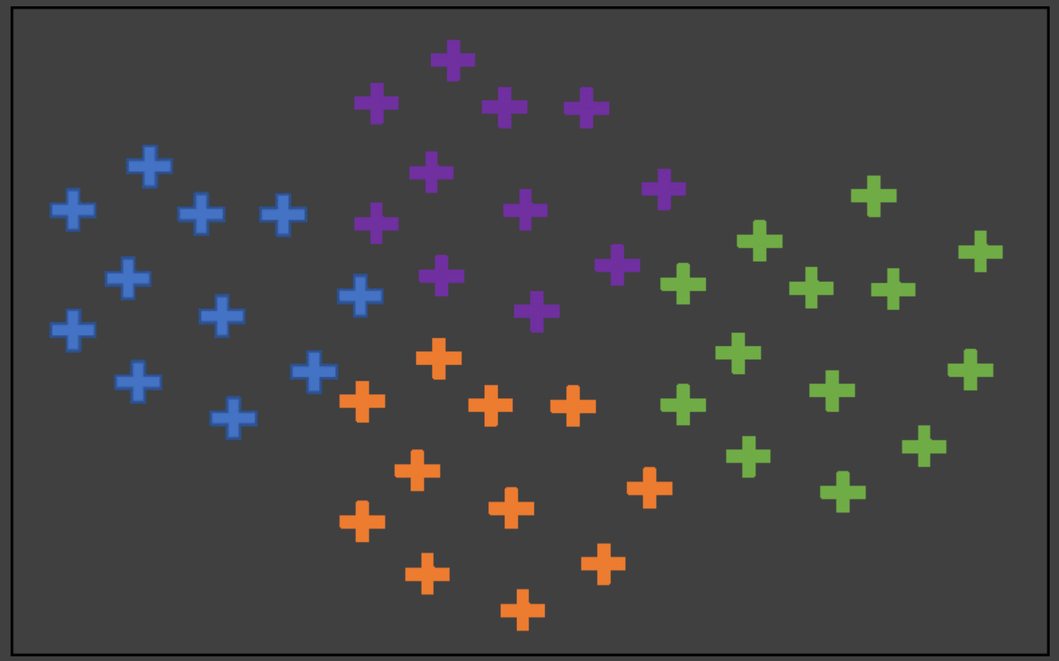


Some Techniques

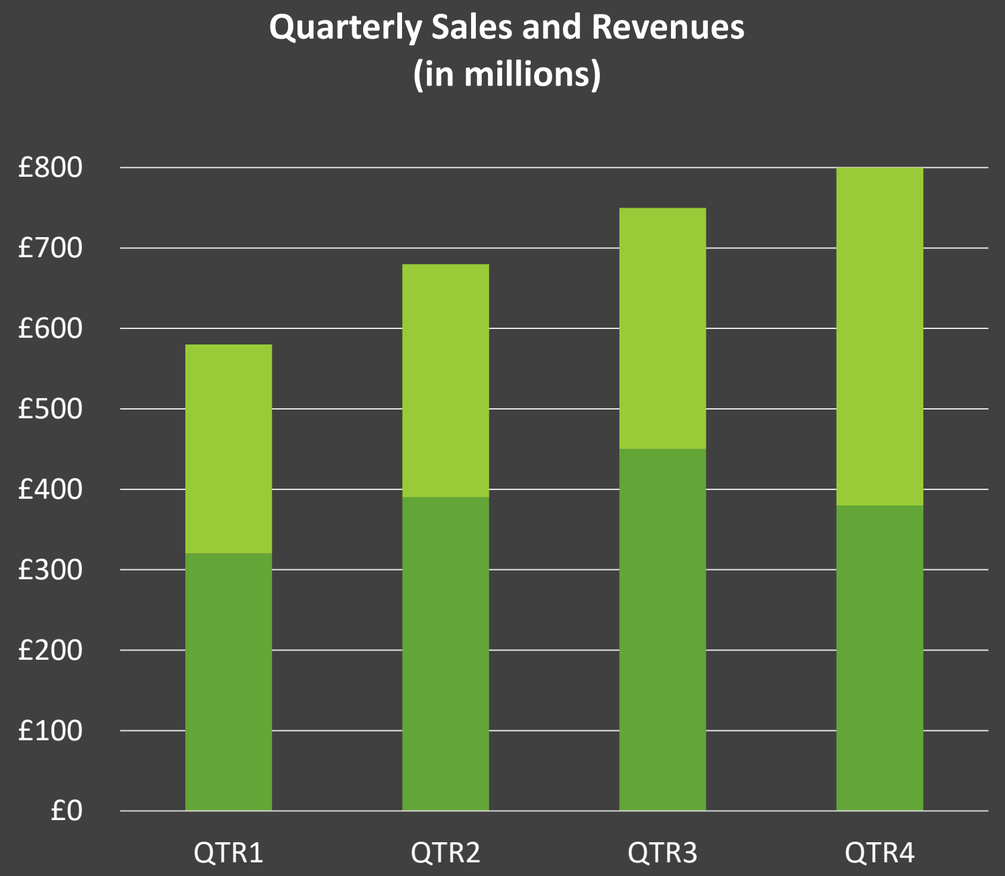
- Rules Based Policies
- Factor Segmentation
- K-Means Clustering
- TwoStep Cluster Analysis
- Latent Class Cluster Analysis
- Discriminant Analysis
- Principal Component Analysis
- Etc

Why Clustering?





Segmentation Goals



Customer Segmentations are often used as the principal basis for decisions on marketing programs, product development, service and delivery programs

Uh oh...

“We have grown gross sales and market share across both Waitrose and John Lewis, but our profits are down.”

- Sir Charlie Mayfield, Chairman of JLP

But why?

Built on being customer led

"Champion everyone's right to feel good"

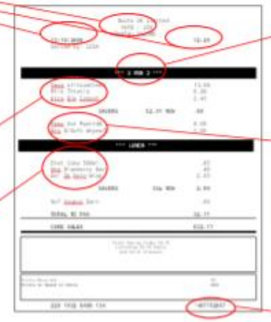
let's feel good

Advantage card at the core

Shops on weekdays at lunchtime in a local store

Buys vitamins – health conscious

Buys into meal deal offer



3 for 2 offers

3 for 2 Mix & Match

Is a parent with a young baby

Boots Advantage Card number

let's feel good

Single view of the customer

Cheryl – 33
Living alone no children
Active Ad Card Member

Online Behaviour

Likes to order online & collect in-store

Likes to browse How To..... & interested in new product launches

Signed up to weekly newsletters – high response

Rates products online

let's feel good

We have to really "Get Women"

- Women account for the majority of our sales in all major categories
- 95% of our loyalty card holders are women
- 80% of our colleagues are women

Truly Customer Led
Boots Understands Women
Through Great Insight

let's feel good

Every time our customer shops

Shops in large store Saturday mornings

Redeems coupons

Purchases self-selection cosmetics, but also premium cosmetics

Could have a partner?

Boots Advantage Card number – same as previous receipt!

let's feel good

Drives insight driven communication

Demographic Data:
We know who the customer is

Transactional Data:
We know what they bought

Contact Data:
We know who received offers

Response Data:
We know who responds to offers

Who to speak to? About what?

let's feel good

It’s not a sale until the customer decides to keep it

Faulty

Can’t be resold as new

Up to 60% returned

Returns cost retailers £435billion globally*

£221 billion preventable retail returns

& returns are growing faster than sales

**IHL Group, Retail Analysts, June 2015*

A Retailer's Dream Turned into a Nightmare

Accessories

Purchased **96** items
Returned **88**
Gross Sales = £1,144

-£45 Net after refunds & costs.

Womenswear

Purchased **104** items
Returned **100**
Gross Sales = £1,845

-£203 Net after refunds & costs.

Menswear

Purchased **28** items
Returned **25**
Gross Sales = £349

-£41 Net after refunds & costs.

Health & Beauty

Purchased **8** items
Returned **0**
Gross Sales = £252

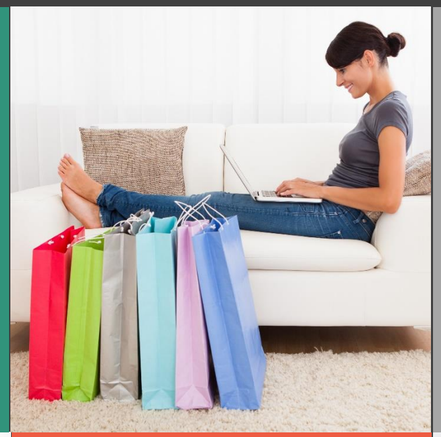
£252 Net

Total Sales = **£3,690**

Total Net after Refunds & Costs = **-£37**

A New Kind of Customer Segmentation

Not all returns are created equal



Returns and returns costs can be up to 30% of gross sales

What about customers that don’t return?

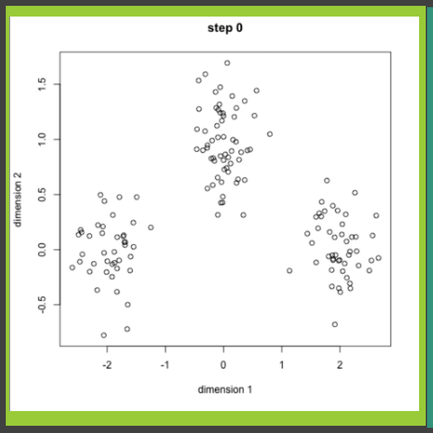
What is high value?

50% of customers return items

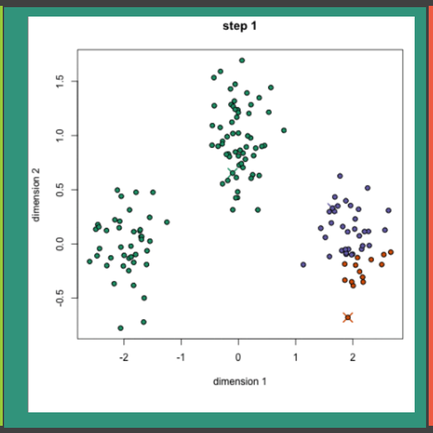
Returns sensitive customers are lost customers

$$Minimize \sum_{j=1}^{k} \sum_{i=1}^{n} (x_{ij} - c_j)^2$$

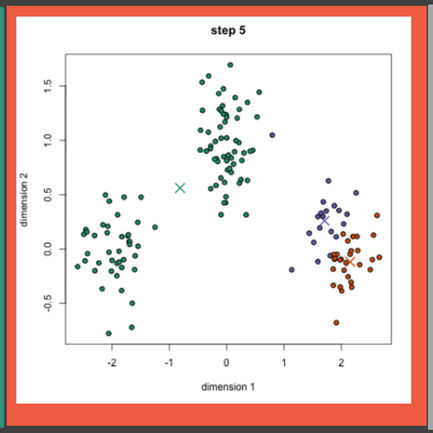
K-Means Clustering



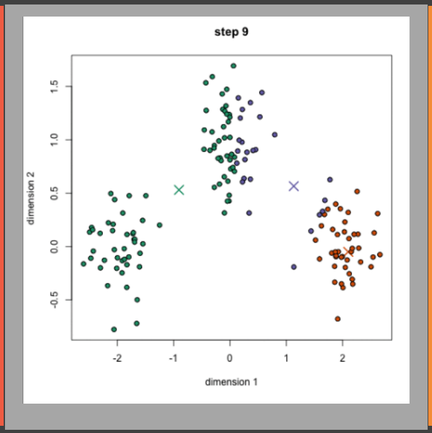
Determine and assign # of clusters (k)



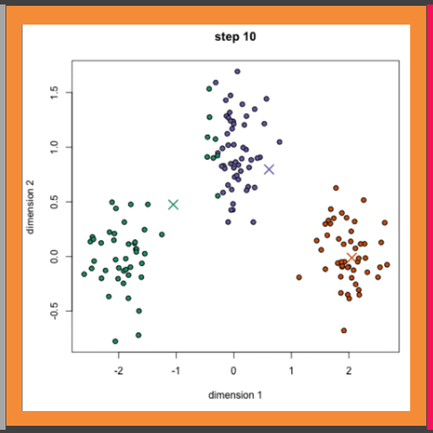
K centroids are selected from the dataset



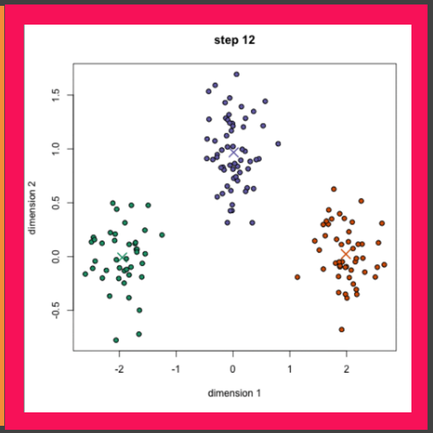
Every data point is assigned to the nearest centroid



Centroids are re-calculated



Are data points re-assigned?



Final clusters generated

K-means Tricks

Does not handle variables of different scales well

Variables needs to be standardised

Results are affected by the order of the records in the dataset

Lots of iteration on seed numbers and sorting schemes to determine robustness

Requires continuous variables

Provide information about the variable's significance to the clusters

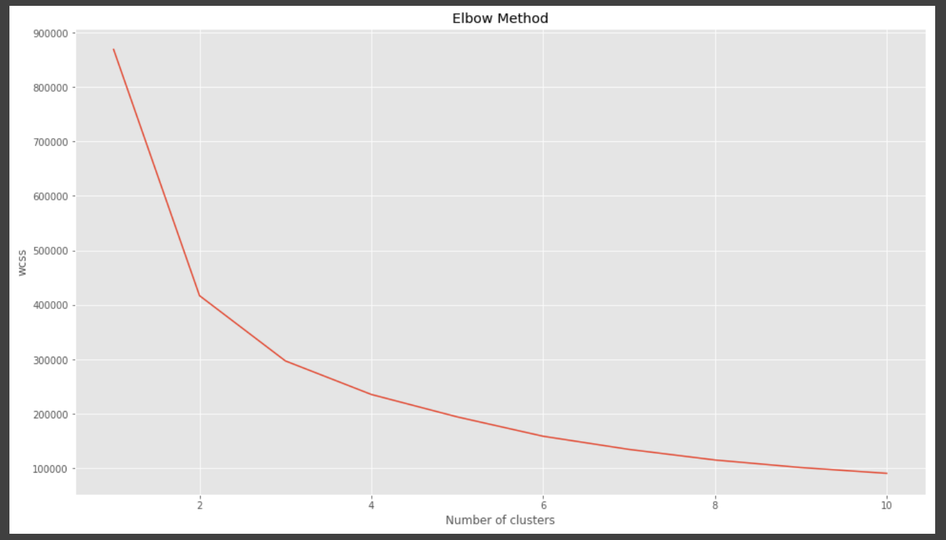
Choosing the Right Variables

Supporting the Objective

We want to understand a customer's holistic shopping habits.

Buying Behaviour	Returning Behaviour	Descriptive Variables
• Frequency • Type/value of items • Basket size	• Return rates • Type/value of items returned • Speed of returns	• Gender • Age • Cardholder • Length of time as member

Choosing K


Elbow Method
WCSS
900000
800000
700000
600000
500000
400000
300000
200000
100000
2
4
6
8
10
Number of clusters


Customer segments
Cluster1
Cluster2
Cluster3
Cluster4
Centroids
Items Purchased
35
30
25
20
15
10
5
20%
40%
60%
80%
100%
120%
Return Rate


Customer segments
Cluster1
Cluster2
Cluster3
Cluster4
Cluster5
Cluster6
Centroids
Items Purchased
35
30
25
20
15
10
5
20%
40%
60%
80%
100%
120%
Return Rate

Putting It All Together

Accidentals

Explorers

High Risk Explorers

Overbuyers

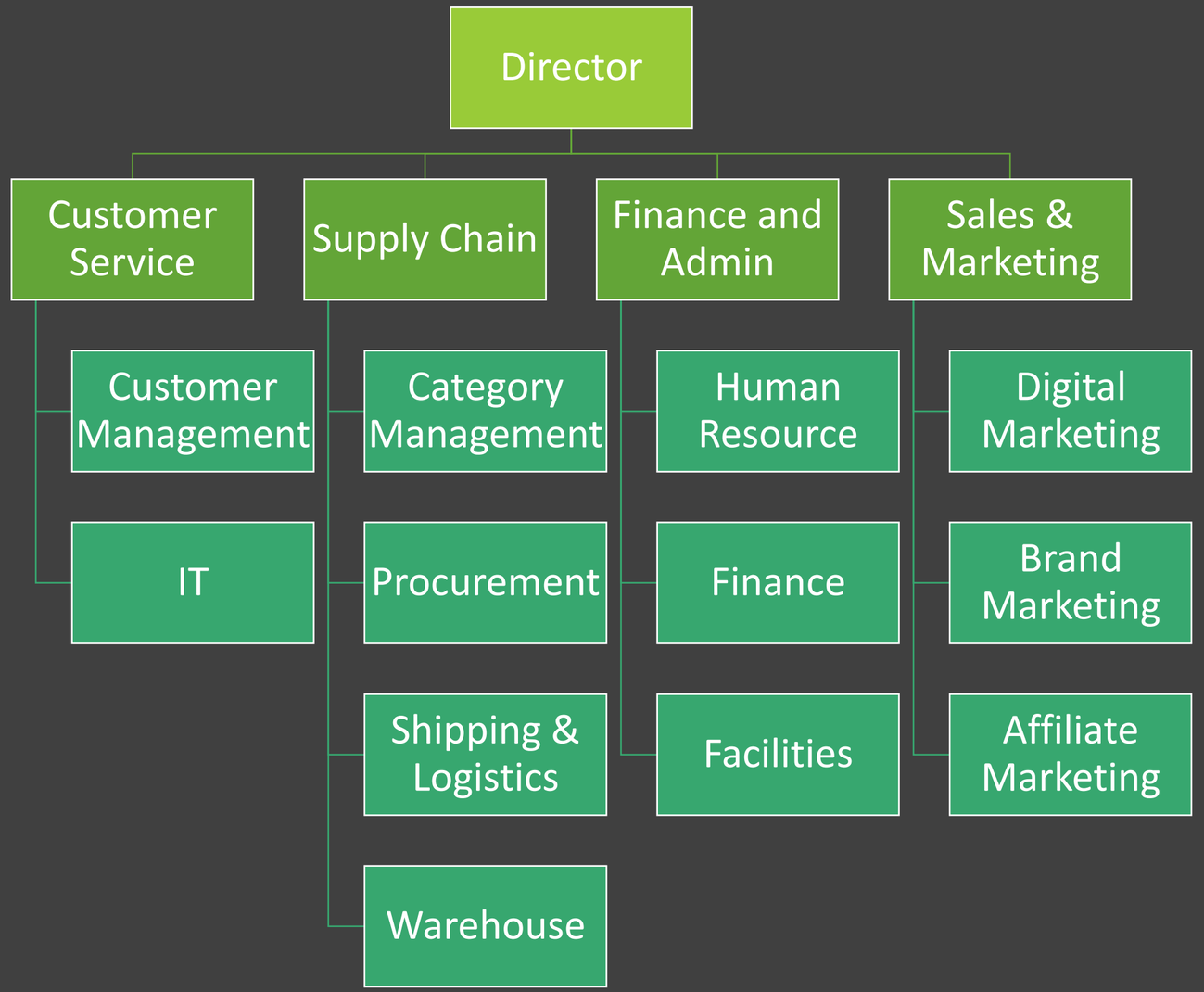
Serious Overbuyers

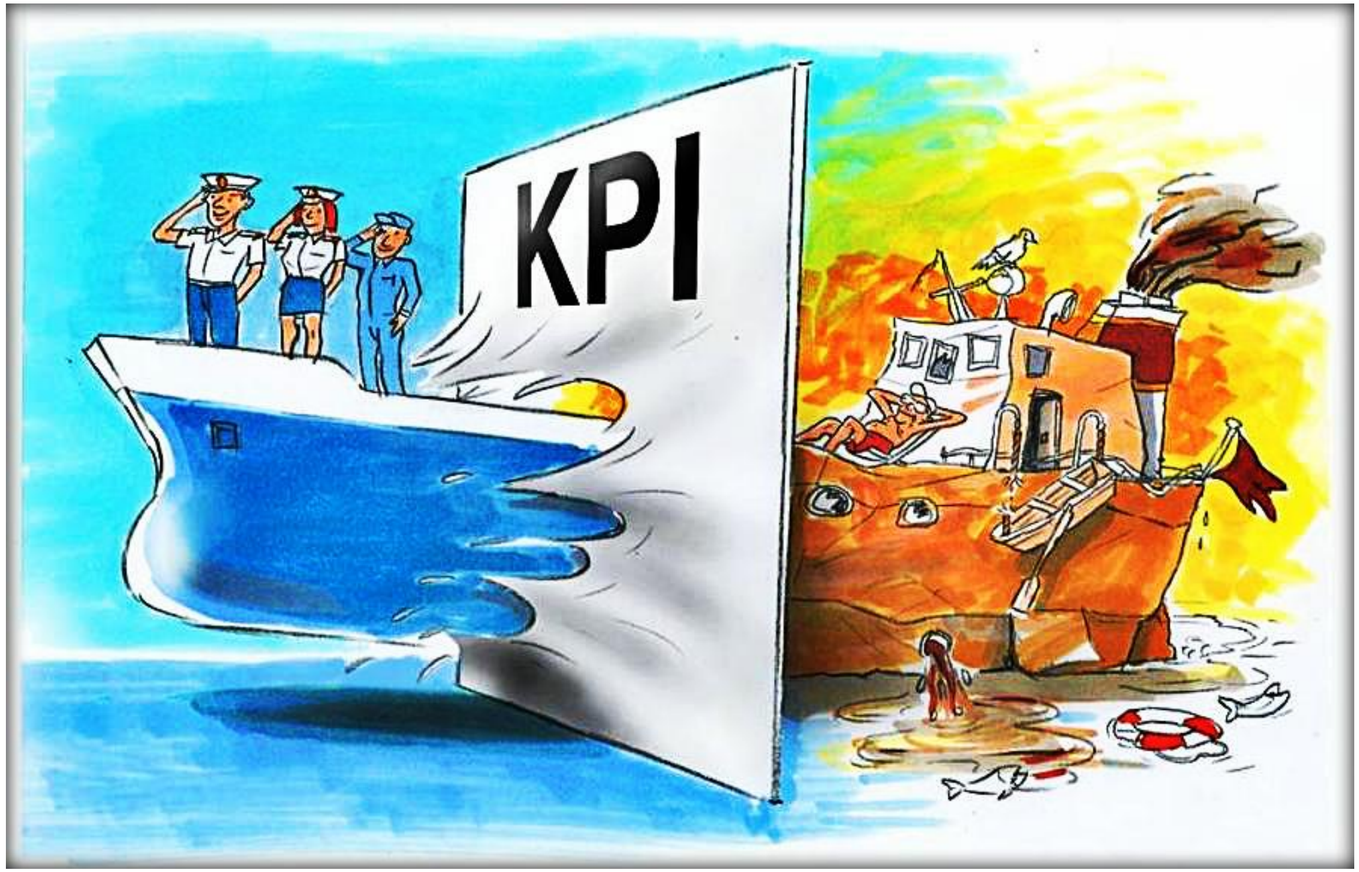
Wear and Returners

How does they compare to traditional segments?

Improving the Bottom Line

Why Did the Company Fail?





Thank You!