# Wholesale Customer Data

## Clustering

By - Divya Ganjoo

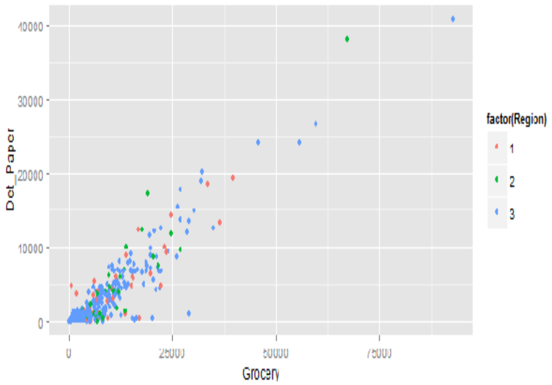
# Data

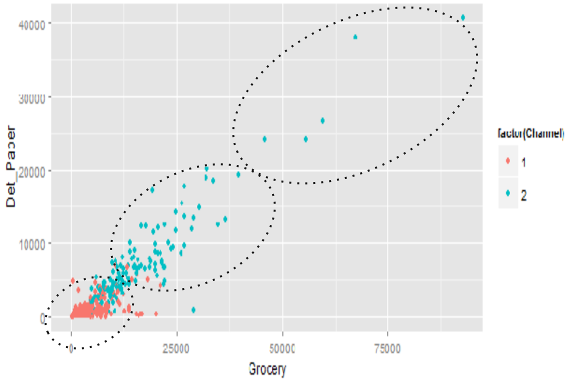
https://archive.ics.uci.edu/ml/datasets/Wholesale+customers#

Attribute Information:

1) FRESH: annual spending (m.u.) on fresh products (Continuous);
2) MILK: annual spending (m.u.) on milk products (Continuous);
3) GROCERY: annual spending (m.u.)on grocery products (Continuous);
4) FROZEN: annual spending (m.u.)on frozen products (Continuous)
5) DETERGENTS_PAPER: annual spending (m.u.) on detergents and paper products (Continuous)
6) DELICATESSEN: annual spending (m.u.)on and delicatessen products (Continuous);
7) CHANNEL: customers' Channel - Horeca (Hotel/Restaurant/Café) or Retail channel (Nominal)
8) REGION: customers' Region – Lisnon, Oporto or Other (Nominal)

# Exploring data





- Exploring the data reveals that channels contain more variability than the regions (Example graph above). Similar patterns emerge for other variables.
- Strong correlation is found in - Grocery & Detergent - 0.92, Milk & Det - 0.66 and Milk & Grocery - 0.73
- Looking at various graphs such as the one above, we can roughly estimate 3 clusters

# Hierarchical

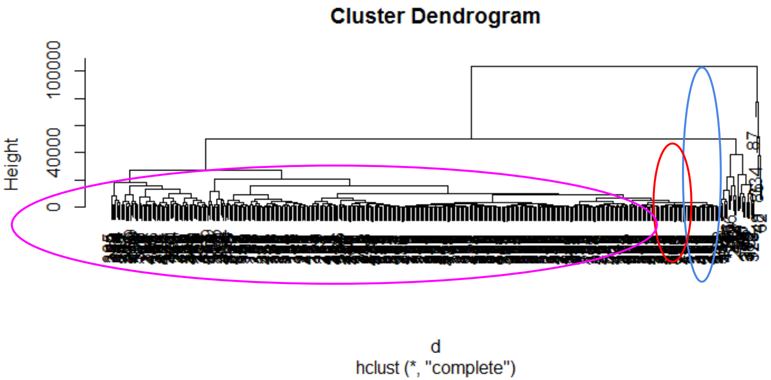


*Fig: Dendrogram: Milk and Grocery*

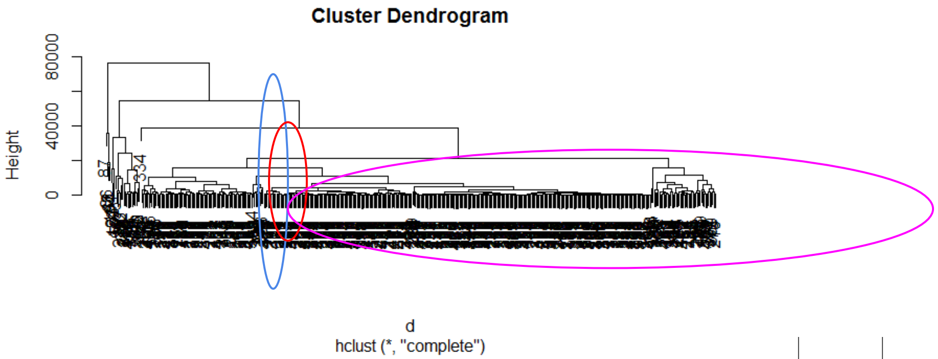


*Fig: Dendrogram: Milk and Detergent/Paper*

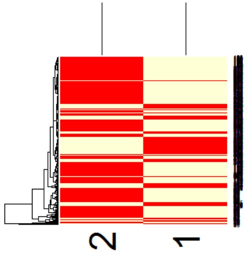
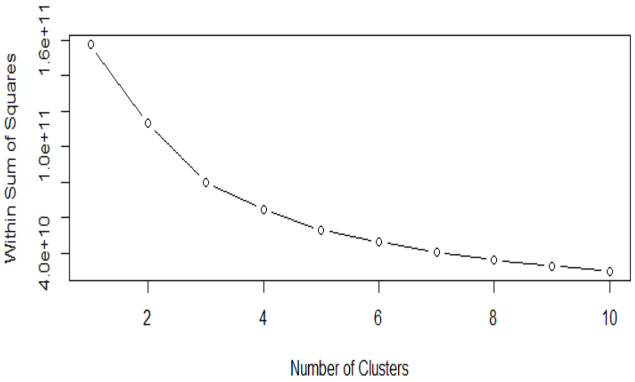
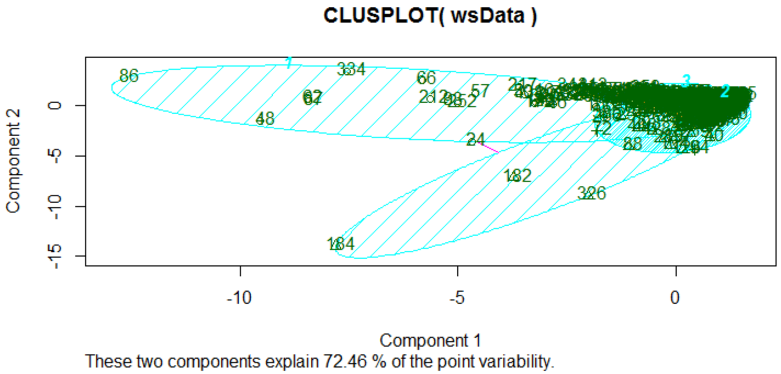


*Fig: Heatmap: Milk and Grocery*

- Different hierarchical clusterings dendrograms roughly categorize data into 3 or 4 groups
- Some observations seem to be separated out consistently (even though its hard to read here)
- Above dendrogram based on Milk and Grocery  AND Milk and Det_Paper

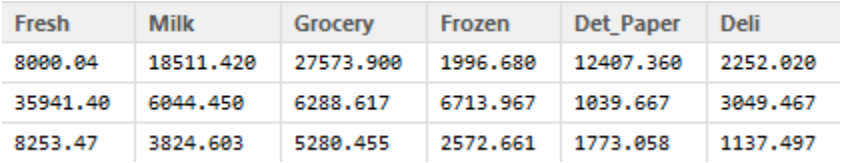
# K-means





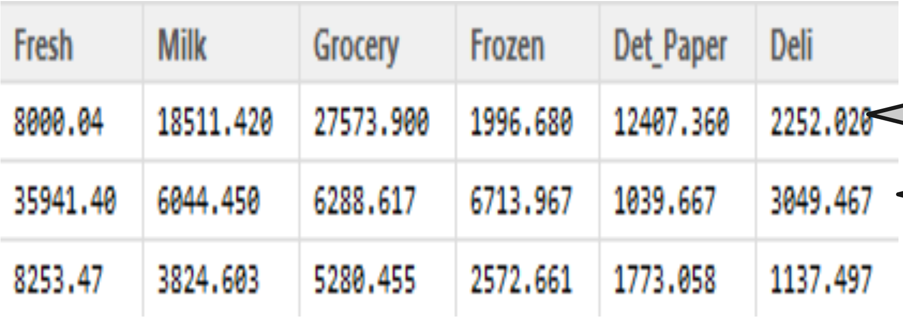
These two components explain 72.46 % of the point variability.

- 3 seems to be the optimum value for no. of clusters as seen from the plot of Within SS against no. of clusters
- Running kmeans with cluster size = 3, we get the following centers >>>



| Fresh | Milk | Grocery | Frozen | Det_Paper | Deli |
|---|---|---|---|---|---|
| 8000.04 | 18511.420 | 27573.900 | 1996.680 | 12407.360 | 2252.020 |
| 35941.40 | 6044.450 | 6288.617 | 6713.967 | 1039.667 | 3049.467 |
| 8253.47 | 3824.603 | 5280.455 | 2572.661 | 1773.058 | 1137.497 |

# Conclusion



| Fresh | Milk | Grocery | Frozen | Det_Paper | Deli |
|---|---|---|---|---|---|
| 8000.04 | 18511.420 | 27573.900 | 1996.680 | 12407.360 | 2252.020 |
| 35941.40 | 6044.450 | 6288.617 | 6713.967 | 1039.667 | 3049.467 |
| 8253.47 | 3824.603 | 5280.455 | 2572.661 | 1773.058 | 1137.497 |

**Cluster1**: High Milk, Grocery, Det_ Paper

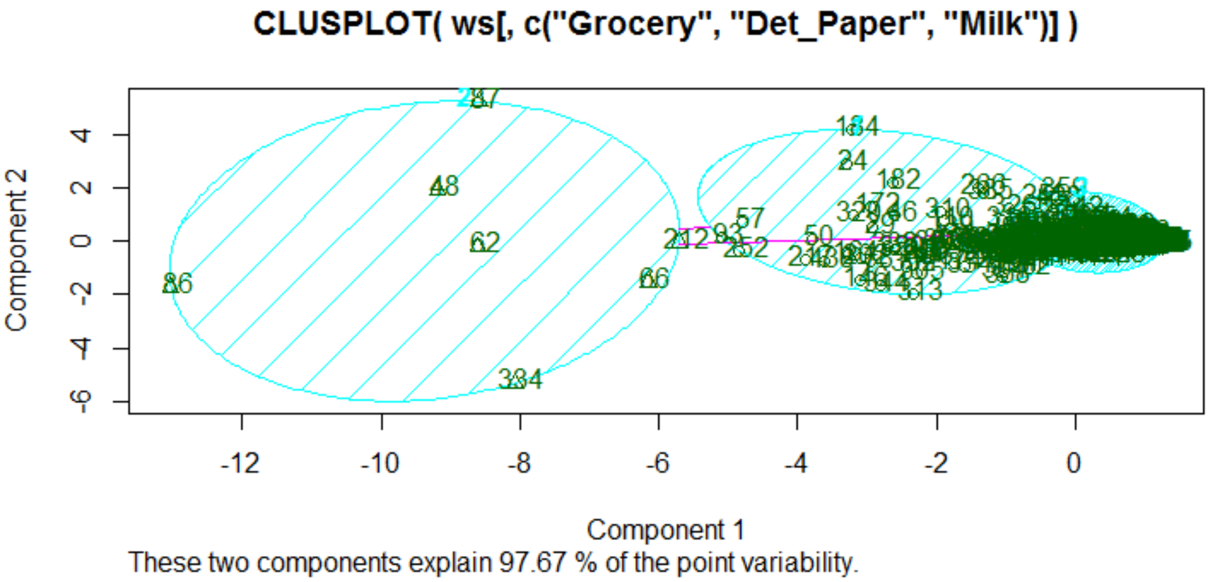
**Cluster2**: High Fresh, Frozen, Deli

**Cluster 3**: Low Spenders

| | Channel 1 | Channel 2 |
|---|---|---|
| Cluster 1 | 2 | **48** |
| Cluster 2 | **52** | 8 |
| Cluster 3 | 244 | 86 |

***Cluster 1:*** *High spenders in Retail channel (Channel 2) tend to spend on Grocery, Milk and Det_Paper categories*
***Cluster 2:*** *High spenders in Horeca channel (Channel 1) tend to spend higher on Fresh product category*
***Cluster 3:*** *Low spenders*

# Appendix



- Observation: If we run Kmeans with Grocery, Detergent_Paper and Milk, we can capture 97% variability in data