UL for predictive modeling in E-commerce

Using a Kaggle dataset with shopping data from Hunter's E-Grocery

Preparation

First, the data is cleaned and normalized. SQL was used for storage and data transformation.

EDA

Then, we perform initial analysis to identify target values, outliers, and emerging themes.

Processing

Informed by the EDA, we now read in the clean data, select our model and standardize that data.

Modeling

Last, we perform K-means and PCA to classify customer data into a number of clusters.

Introduction

As e-grocery sales continue to climb, food service brands will need insights on highly valuable data to better understand shopper's needs and habits.

Using unsupervised machine learning, we hope to categorize shoppers based on their purchase behavior.

Our Team

Peter Warren

Data Cleaning and Storage

Oluwatobi Adelaja

Exploratory Data Analysis

Anna Barbera

Data Standardization, PCA

Patrick Brennan

K-Means Clustering

Daniel King-Allen

Interpretation & Final Analysis

Prepare & Process

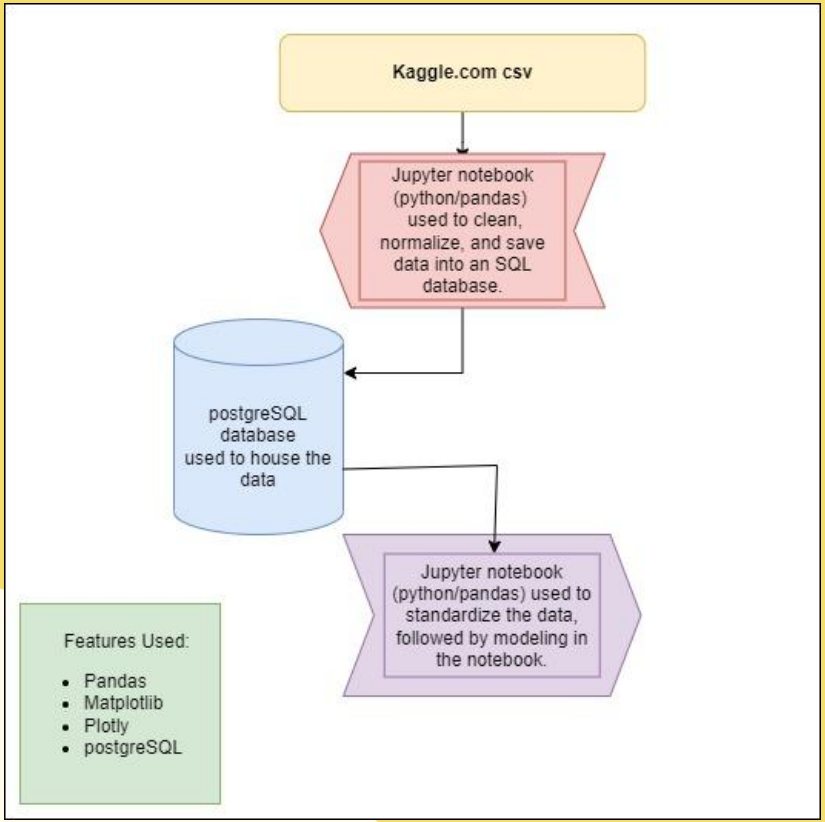
Getting the data ready for machine learning

Data Cleaning & Storage

Our Process:

- We selected a dataset from Kaggle. Pre project it was saved as a parquet file due to its size.
- It was cleaned and used to populate a postgres database table.
- Views were created to easily pull transformed data.

Processing Flowchart



Takeaways

- Leverage easily used tools for the storage of large data sets.

- Database storage and views leverage the advantages of SQL.

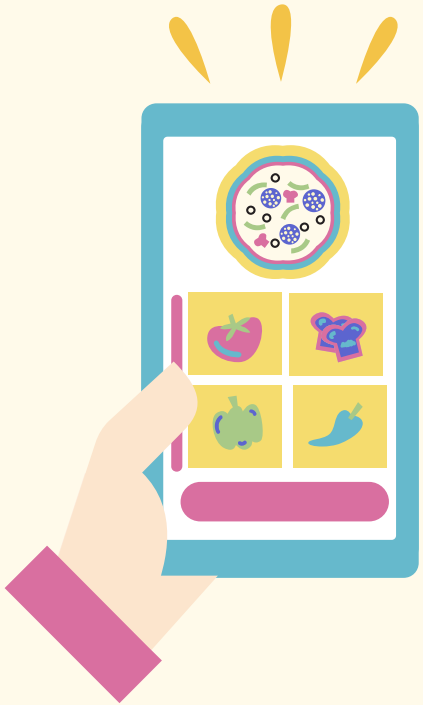
- The impact of bad data quality is magnified in these data sets.

- **Authorship matters.**

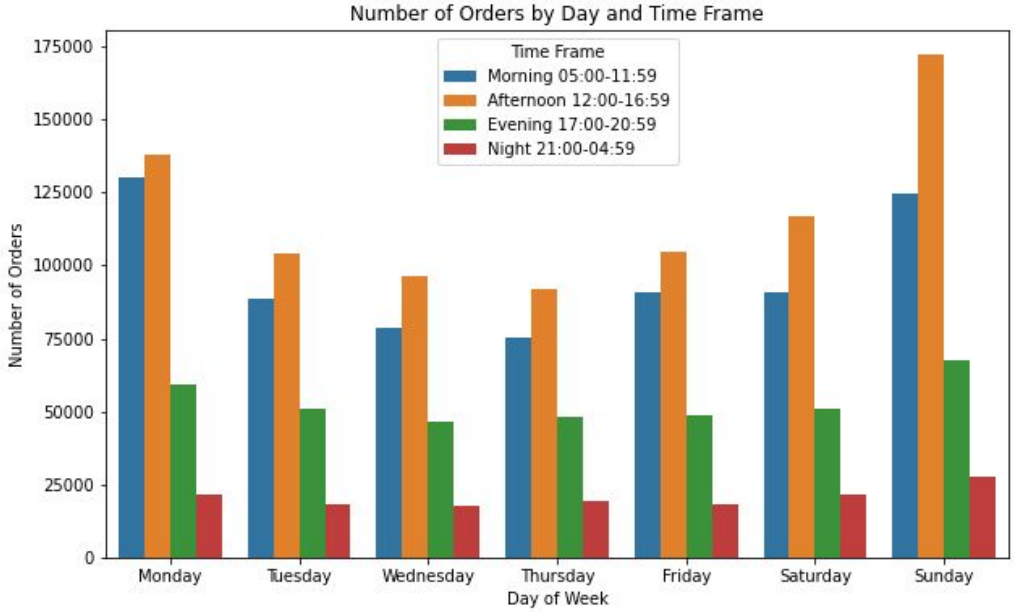
Initial EDA

- **Visualize** the **relationships** between orders and the time they are placed using a pivot table and bar graph.
- **Explore** the **peak hours** using a line graph and heat map.
- **Reveal** the **weekly top 10** most ordered items using a pie chart.

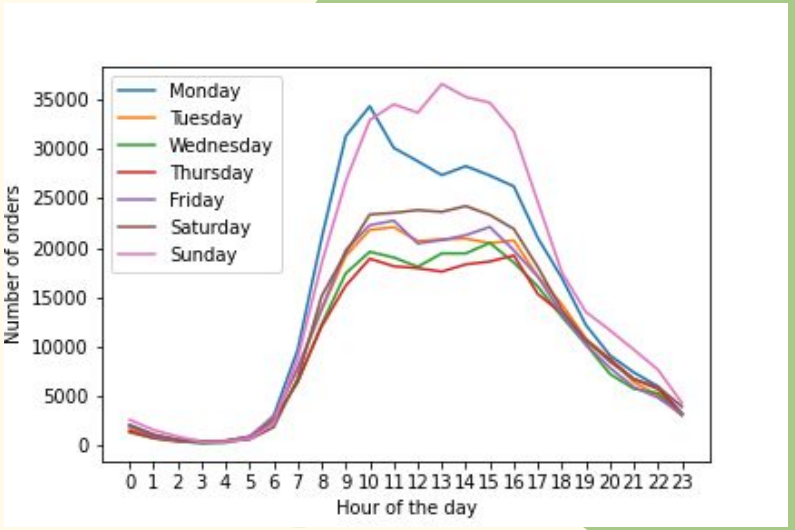
Average Weekly Orders

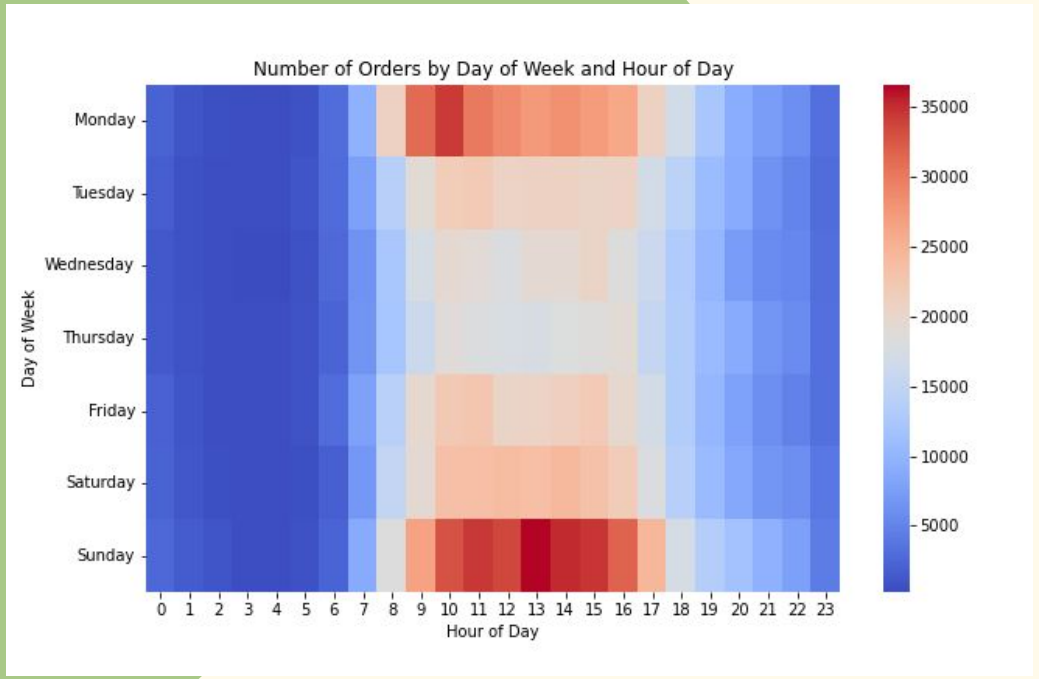


DayOfWeek	Morning 05:00-11:59	Afternoon 12:00-16:59	Evening 17:00-20:59	Night 21:00-04:59	Total
Monday	130332	137957	59382	21565	349236
Tuesday	88526	103813	51077	18496	261912
Wednesday	78385	96083	46722	17540	238730
Thursday	75266	91827	48189	19602	234884
Friday	90614	104468	48585	18490	262157
Saturday	91062	117003	51078	21608	280751
Sunday	124582	171971	67392	27886	391831
Total	678767	823122	372425	145187	2019501

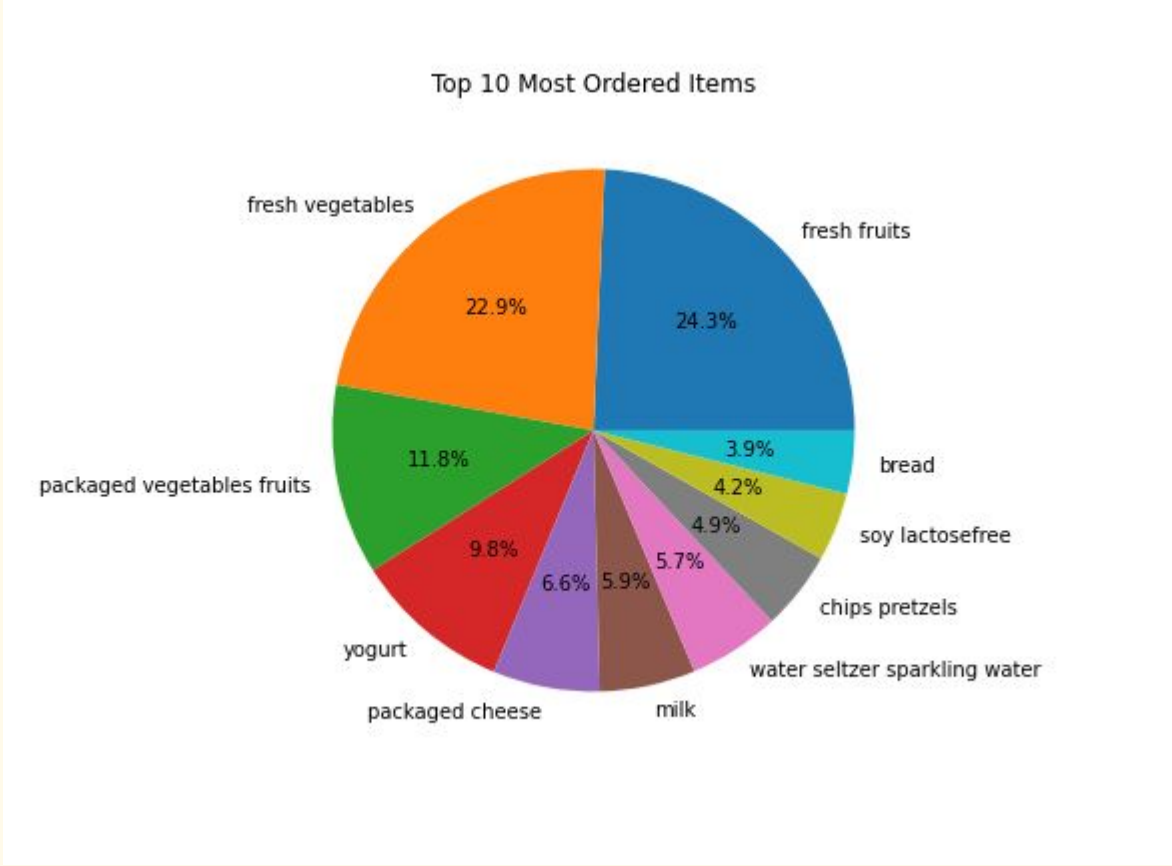


Peak Days & Hours





Top 10 Product Categories



K-MEANS & PCA

Using modeling to define clusters

Preparing the data for clustering

The dataset was initially organized such that each row was a singular item with a corresponding order ID and customer ID. For this analysis, we were interested in **clustering the customers** to learn what we could about those groupings and how we may be able to serve those customers better

com_id	order_id	user_id	order_number	order_dow	order_hour_of_day	days_since_prior_order	product_id	a
0	2425083	49125	1	2	18	0.000000	17	
1	2425083	49125	1	2	18	0.000000	91	
2	2425083	49125	1	2	18	0.000000	36	
3	2425083	49125	1	2	18	0.000000	83	
4	2425083	49125	1	2	18	0.000000	83	

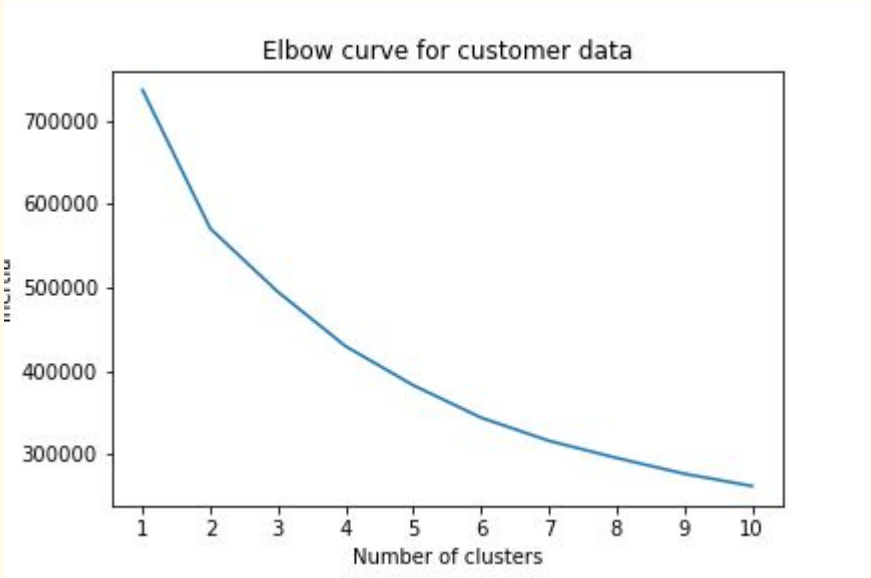
Using SQL to create views of the original database from Kaggle, we aggregated the below metrics by customer ID:

- Average day of the week ordered
- Average time of day ordered
- Average number of days between orders
- Number of total reorders
- Total orders
- Total number of products ordered
- Average order size

order_dow	order_hour_of_day	days_since_prior_order	reordered	order_number	product_id	avg_ordersize
3.181818	10.000000	6.818182	4	3	11	3.666667
1.947368	18.473684	12.421053	9	11	19	1.727273
0.000000	18.000000	30.000000	20	3	24	8.000000
3.000000	15.000000	14.000000	13	4	30	7.500000
5.000000	11.000000	30.000000	3	5	11	2.200000

4 clusters

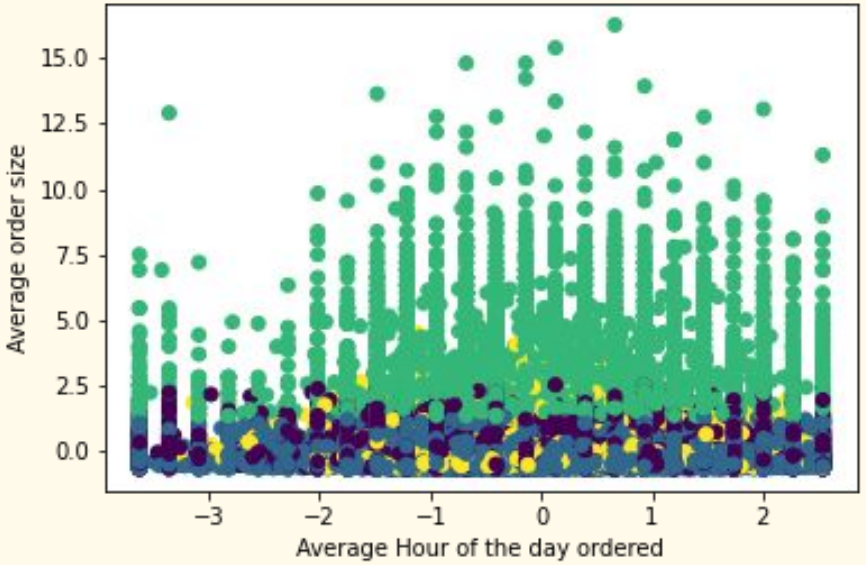
Using an elbow curve, we determined that the ideal number of clusters initially was 4.



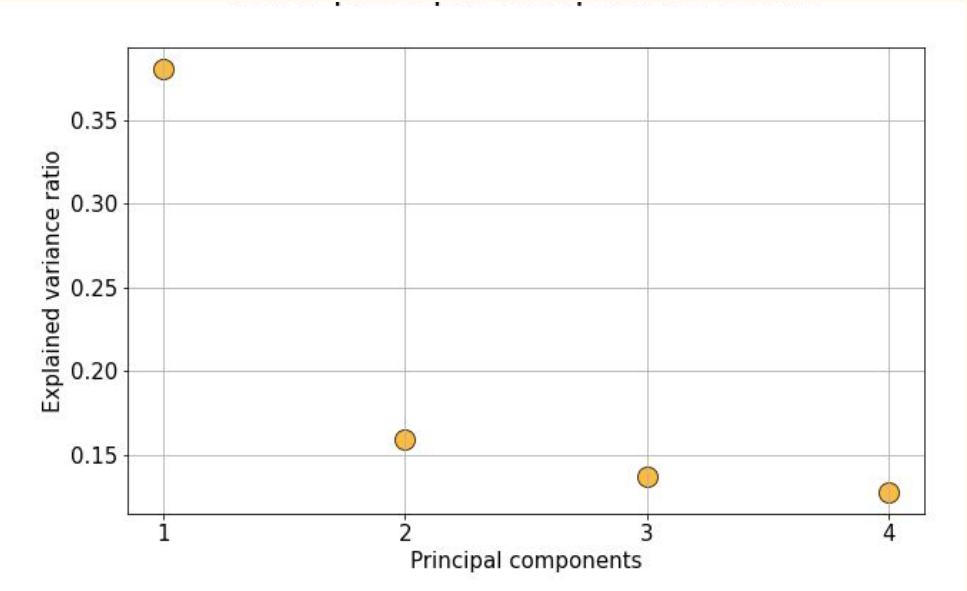
Initial cluster scatter plot

Our initial clustering showed some grouping that seemed to be roughly broken out by high and low volume shoppers, but the lower volume shoppers seem to have a lot of noise in their clusters

We could perform EDA on these clusters, but instead we decided to further refine our analysis by performing a PCA analysis to reduce the number of features we're examining to hopefully create clusters with cleaner breaks



Explained Variance Ratio



PC 1: **38%**

PC 2: **16%**

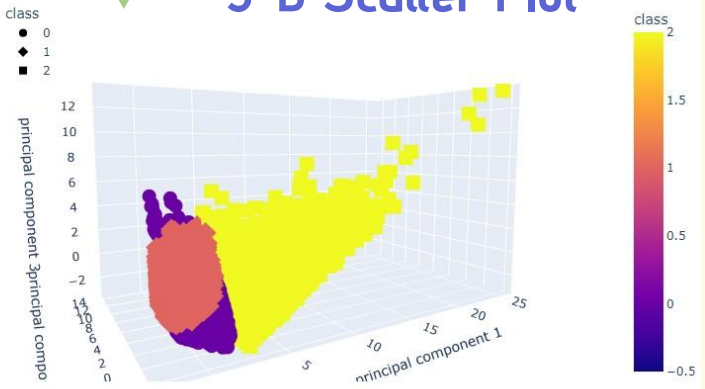
PC 3: **14%**

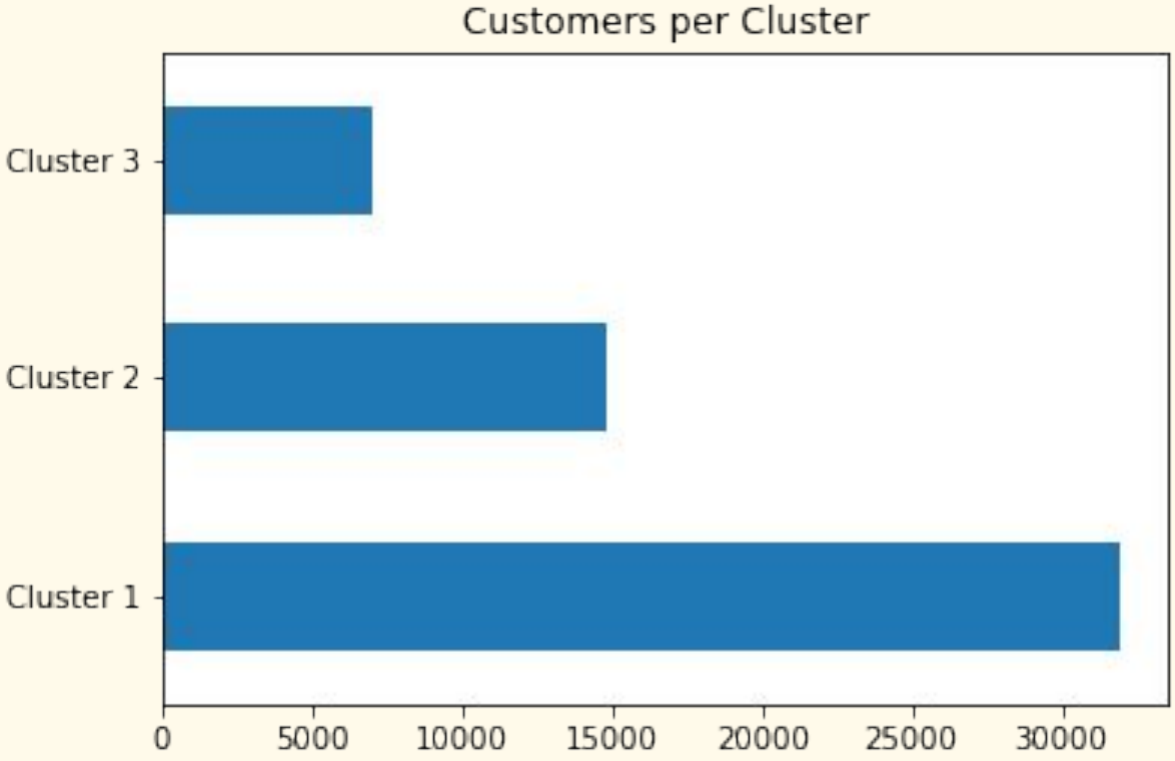
Total: **68%**

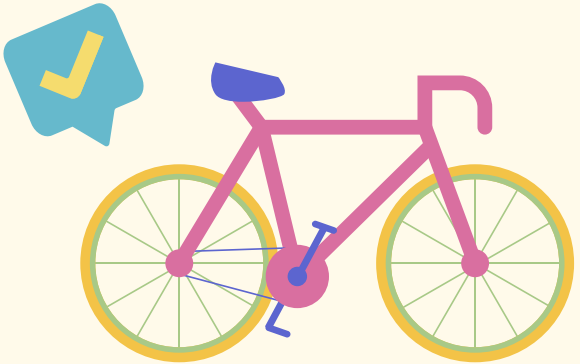
Principal Component Analysis

Visualizing the explained variance ratio for each component, we can use this to refine our clusters from 4 to 3.

3-D Scatter Plot

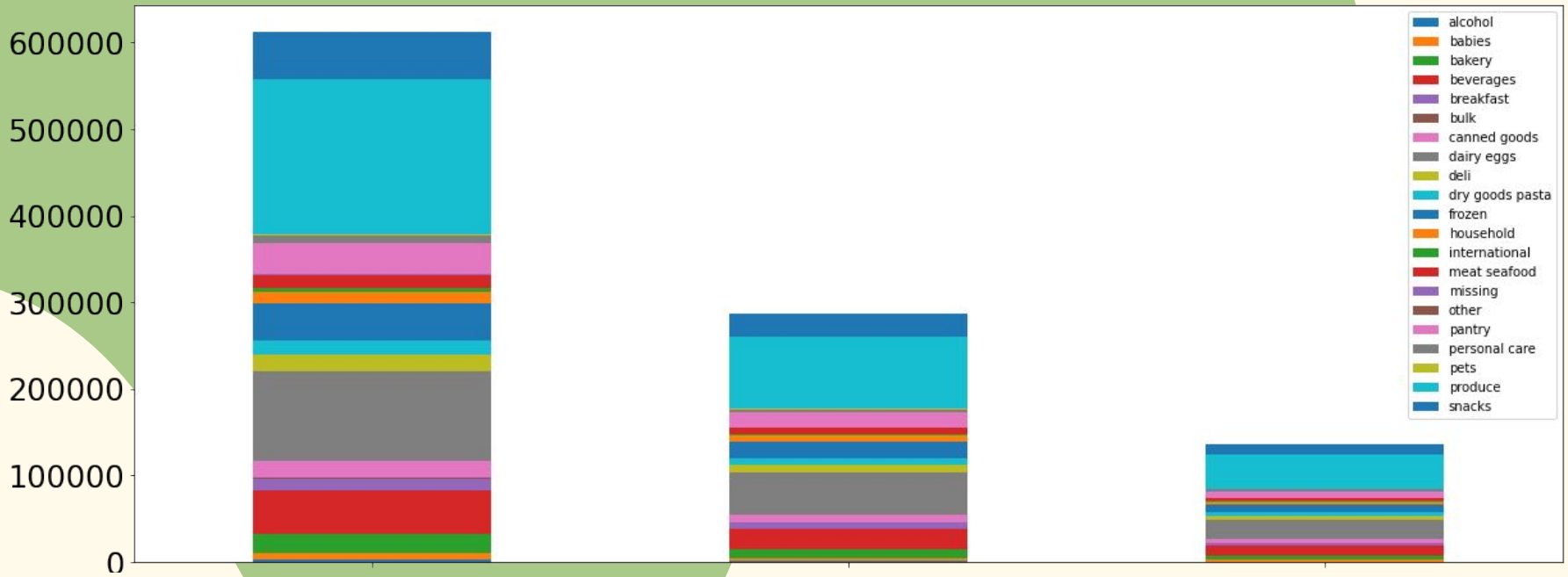



Customers per Cluster
Cluster 3
Cluster 2
Cluster 1
0
5000
10000
15000
20000
25000
30000



Final Results

class	alcohol	babies	bakery	beverages	breakfast	bulk	canned goods	dairy eggs	deli	dry goods pasta	frozen	household	international	meat seafood	missing	other	pantry	personal care	pets	produce	snacks
Cluster 1	2946	7610	22195	50466	13516	683	20001	102637	19694	16241	42247	13936	5145	13630	1440	686	35204	8528	1766	179449	54699
Cluster 2	1350	3614	10453	23710	6266	322	9508	47733	9302	7655	19770	6340	2381	6299	654	300	16712	3968	889	83877	25495
Cluster 3	562	1878	4917	11590	3034	140	4534	22485	4543	3639	9312	3210	1166	2918	322	152	7645	1930	464	39482	12095
Total	4858	13102	37565	85766	22816	1145	34043	172855	33539	27535	71329	23486	8692	22847	2416	1138	59561	14426	3119	302808	92289



Thank you!

Thanks!

Do you have any questions?
youremail@freepik.com
+91 620 421 838
yourcompany.com

CREDITS: This presentation template was created by **Slidesgo**, including icons by **Flaticon**, and infographics & images by **Freepik**

Please keep this slide for attribution.

Alternative Icons

Premium Icons

Alternative Resources