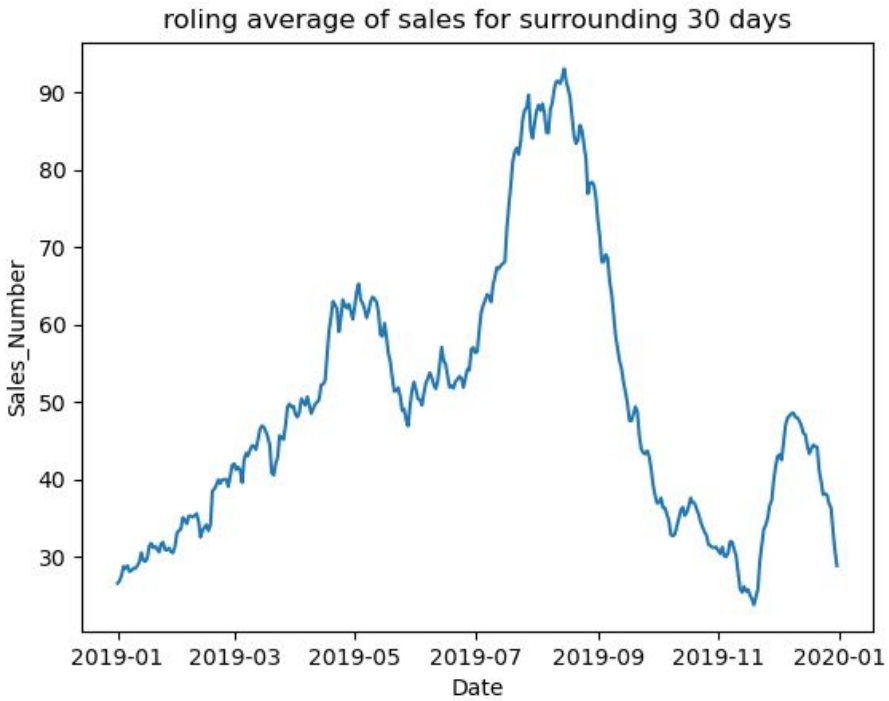
Clustering Sales Patterns

Group 6:
Arthur van Eeden, Yaqi Duan & Dion Hoogewoonink

Relevance

Sales data



Relevance

Empty shelves

Exploration

kaggle

- Start with comparing data
- Choose suitable database
- Find suitable data (age, pay method, date, etc.)

Pakistan's Largest E-Commerce Dataset

Half a Million Online Orders

📈 E-Commerce Sales Analysis

Analyze Product Performance and Customer Preferences in the E-Commerce Market

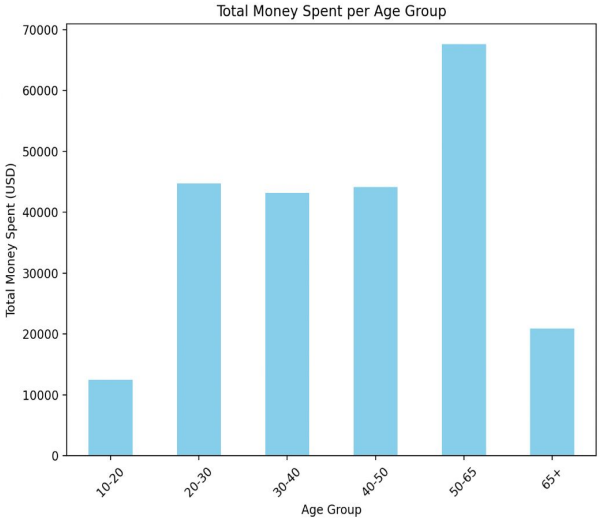
🛒 Online Shopping Dataset 📊📉📈

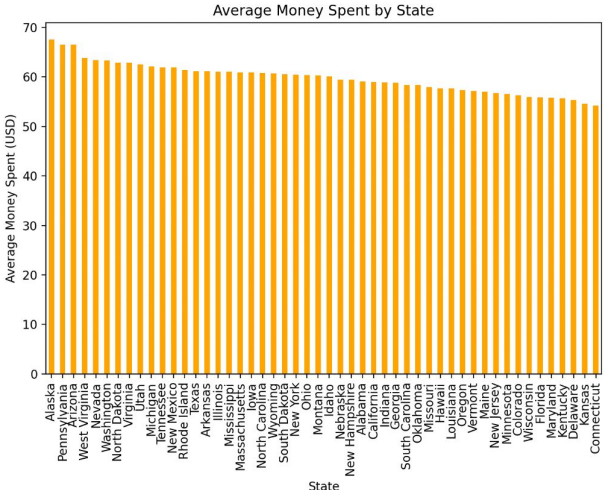
Exploring Online Shopping Trends and Patterns

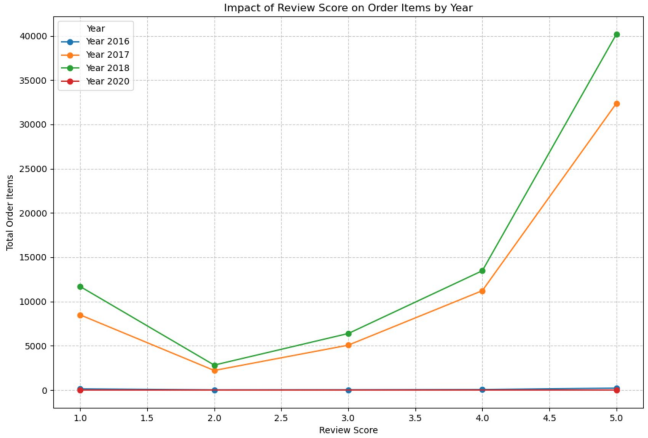
🛒 E-commerce Customer Data For Behavior Analysis

Explore Customer Shopping Habits, Churn, and Purchase Patterns 💳🛒

Some inspirations...





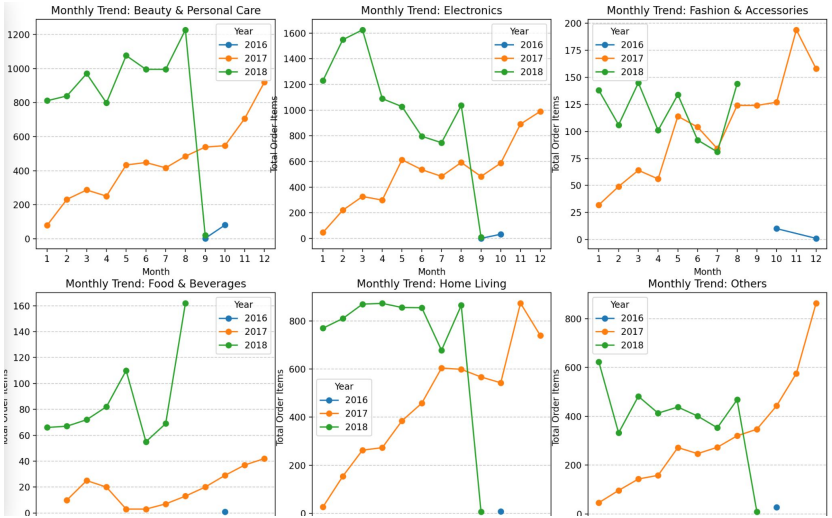


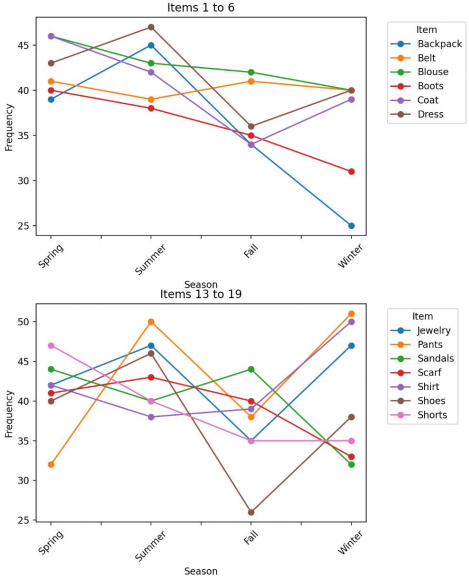
age

location

review score

Some inspirations...





category and time

Research Question

How valid are trend and seasonality modeling techniques, such as polynomial regression and Fourier analysis, in clustering and predicting daily shopping behavior?

Hypothesis

- **Ha (goodness-of-fit): Test if the residuals from the polynomial fit and Fourier fit follow a normal distribution.**
  - H0_a1: The residuals from the polynomial fit follow a standard normal distribution.
  - H0_a2: The residuals from the Fourier fit it follow a standard normal distribution.

- **Hb (compare between models)**
  - H0_b: There is no significant difference in the clustering performance of the polynomial fit and Fourier Fit.

Data Overview

0,	1,	2,	3,	4,	5,	6,	7,
rownr,	CustomerID,	Gender,	Location,	Tenure_Months,	Transaction_ID,	Transaction_Date,	Product_SKU,
int,	float,	char,	str,	float,	float	str(date),	str,
0,	17850.0,	M,	Chicago,	12.0,	16679.0,	2019-01-01,	GGOENEBJ079499,

8,	9,	10,	11,	12,	13,	14,
Product_Description,	Product_Category,	Quantity,	Avg_Price,	Delivery_Charges,	Coupon_Status,	GST,
str,	str,	float,	float,	float,	str,	float,
Nest Learning The__,	Nest-USA,	1.0,	153.71,	6.5,	Used,	0.1,

15,	16,	17,	18,	19,	20
Date,	Offline_Spend,	Online_Spend,	Month,	Coupon_Code,	Discount_pct
str(date),	float,	float,	int,	str,	float
1/1/2019,	4500.0,	2424.5,	1,	ELEC10,	10.0

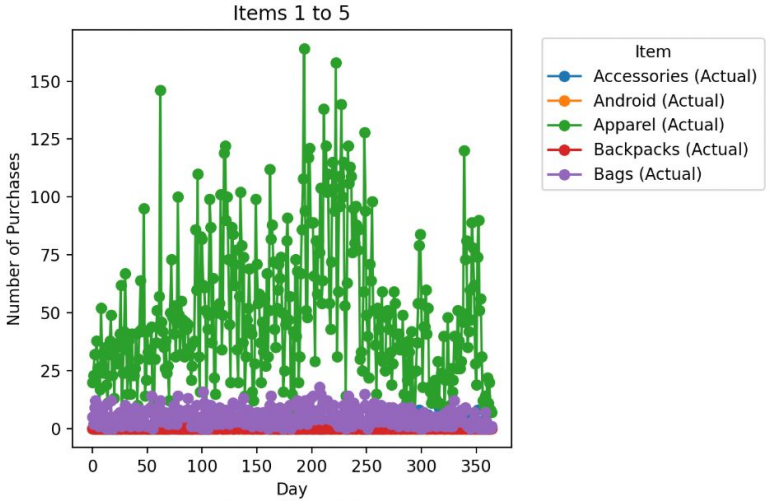
- time: 2019.01.01 - 2019.12.31
- dv: number of transactions

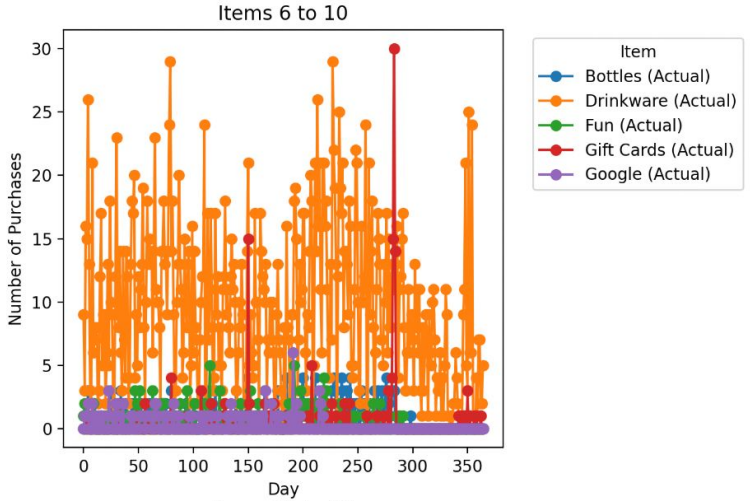
Our Analysis

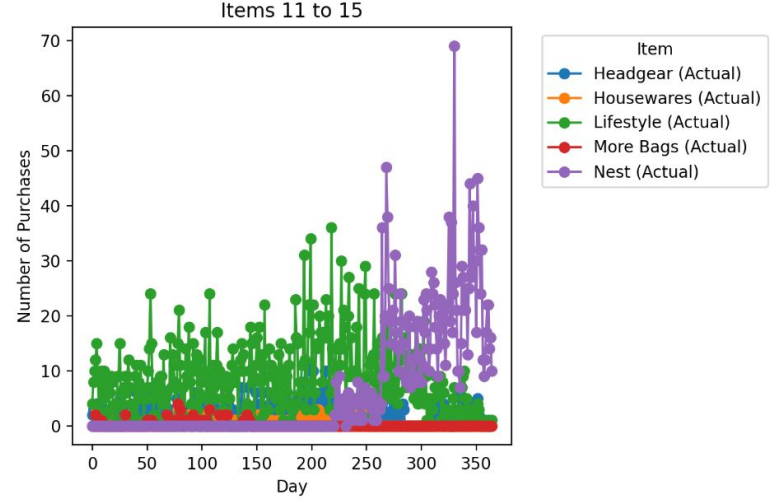
- Polynomial regression
- Fourier analysis
- K-means clustering

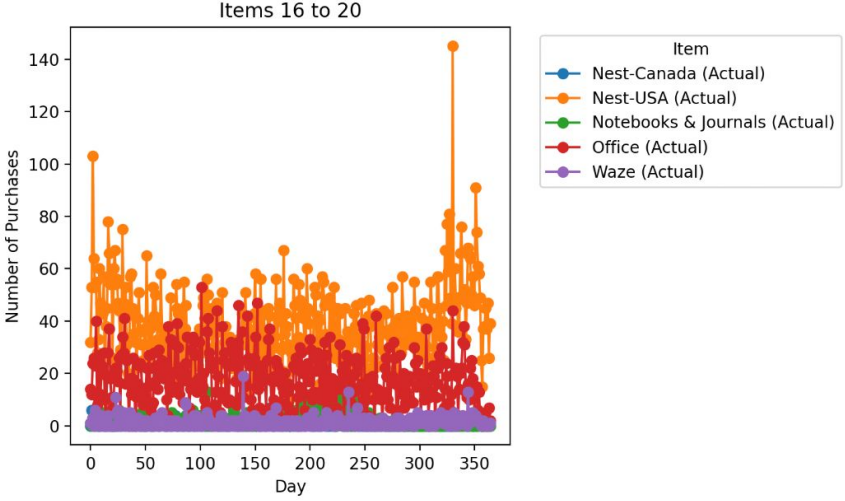
Polynomial Fit

- Start by performing the fit
- Subtract fitted values from real values
- Cosine Similarity Matrix
- Clustering

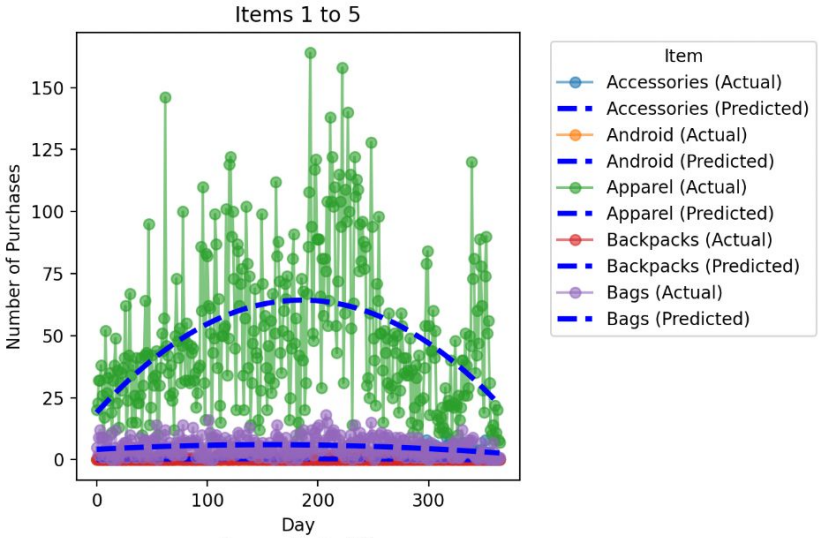

Items 1 to 5
Number of Purchases
Day
Item
Accessories (Actual)
Android (Actual)
Apparel (Actual)
Backpacks (Actual)
Bags (Actual)

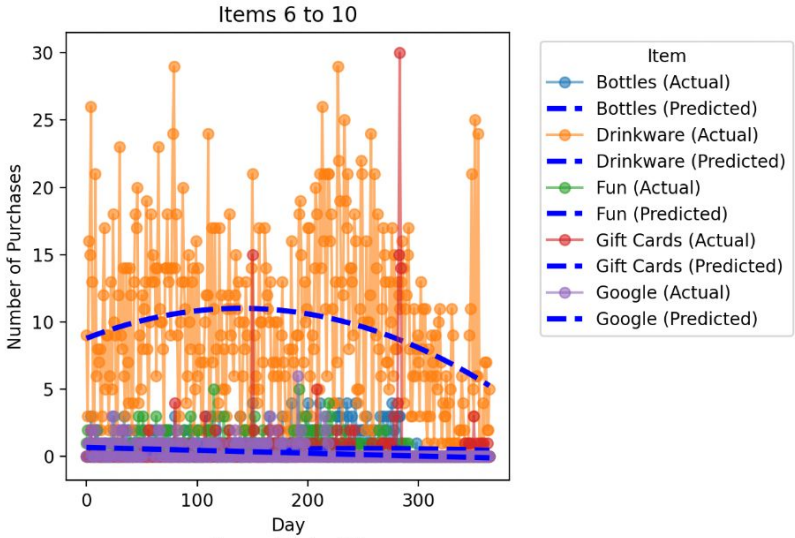

Items 6 to 10
Number of Purchases
Day
Item
Bottles (Actual)
Drinkware (Actual)
Fun (Actual)
Gift Cards (Actual)
Google (Actual)


Items 11 to 15
Number of Purchases
Day
Item
Headgear (Actual)
Housewares (Actual)
Lifestyle (Actual)
More Bags (Actual)
Nest (Actual)

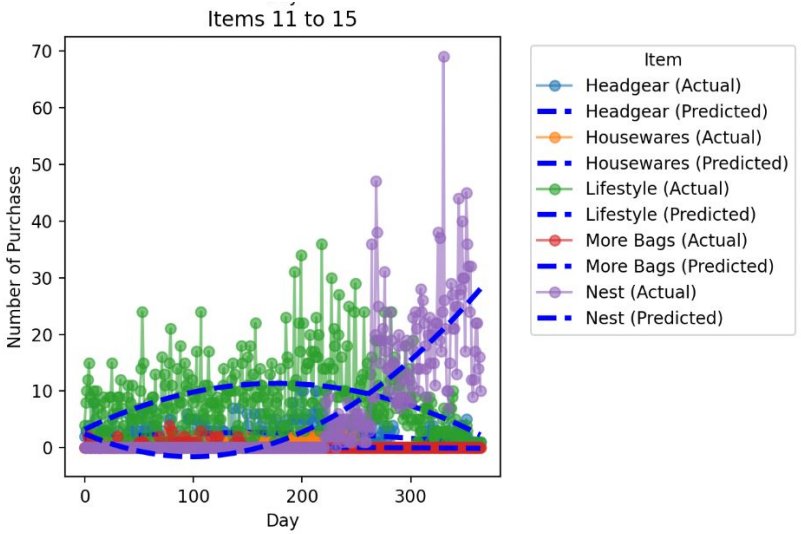

Items 16 to 20
Number of Purchases
Day
Item
Nest-Canada (Actual)
Nest-USA (Actual)
Notebooks & Journals (Actual)
Office (Actual)
Waze (Actual)

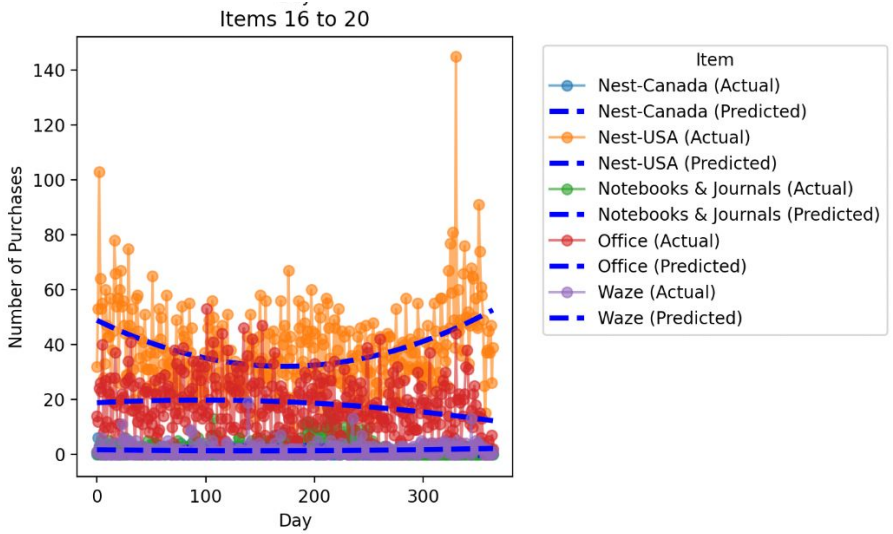
Polynomial Fit: Performing the Fit



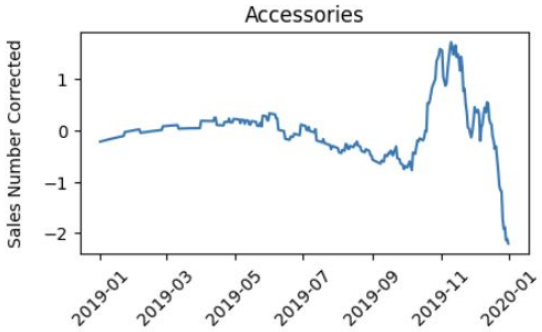


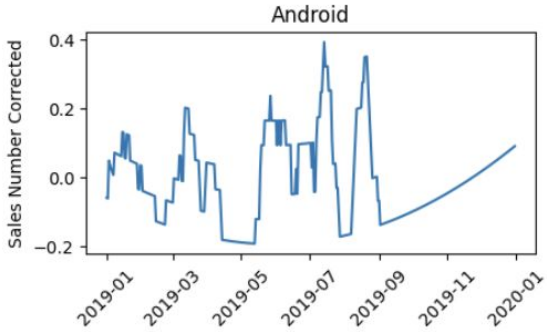
Polynomial Fit: Performing the Fit

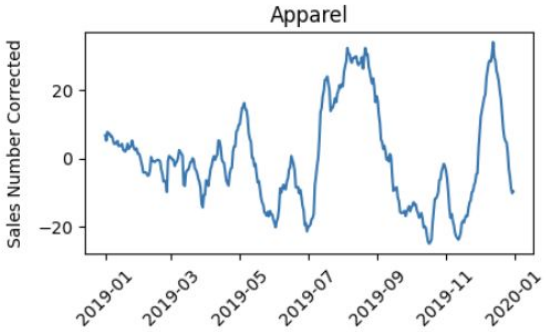


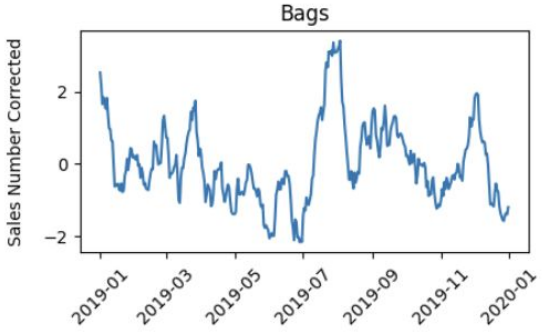


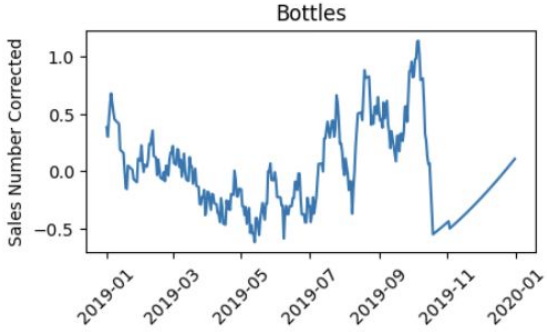
Polynomial Fit: Subtract Values

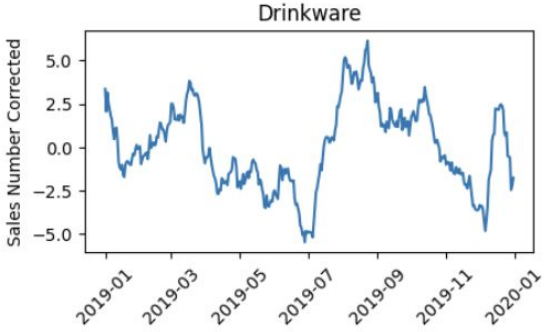




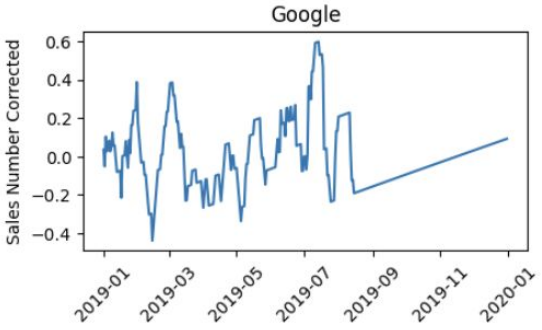


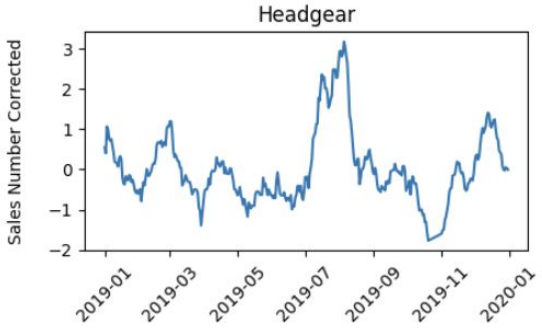


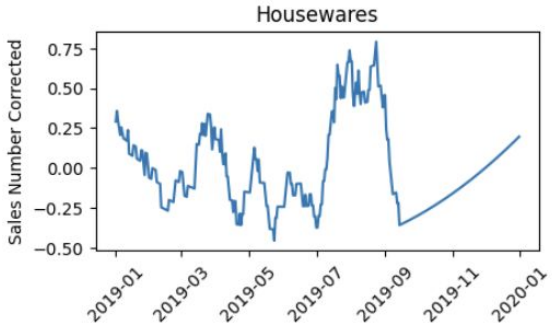


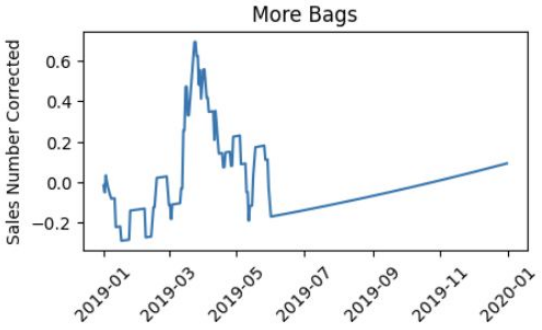


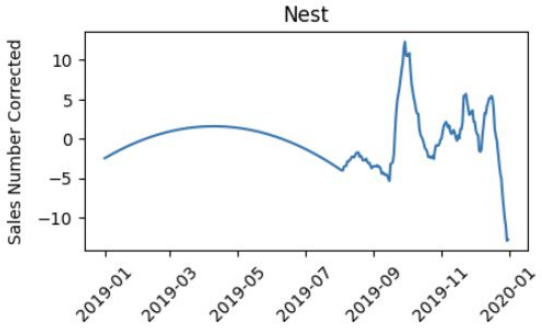
Polynomial Fit: Subtract Values

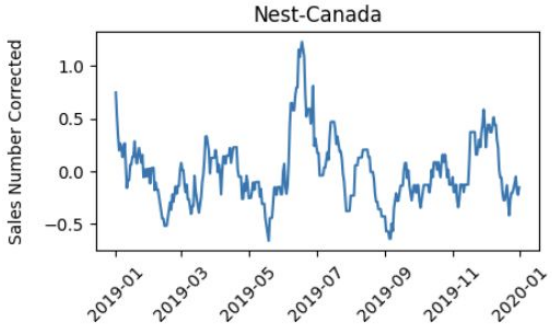




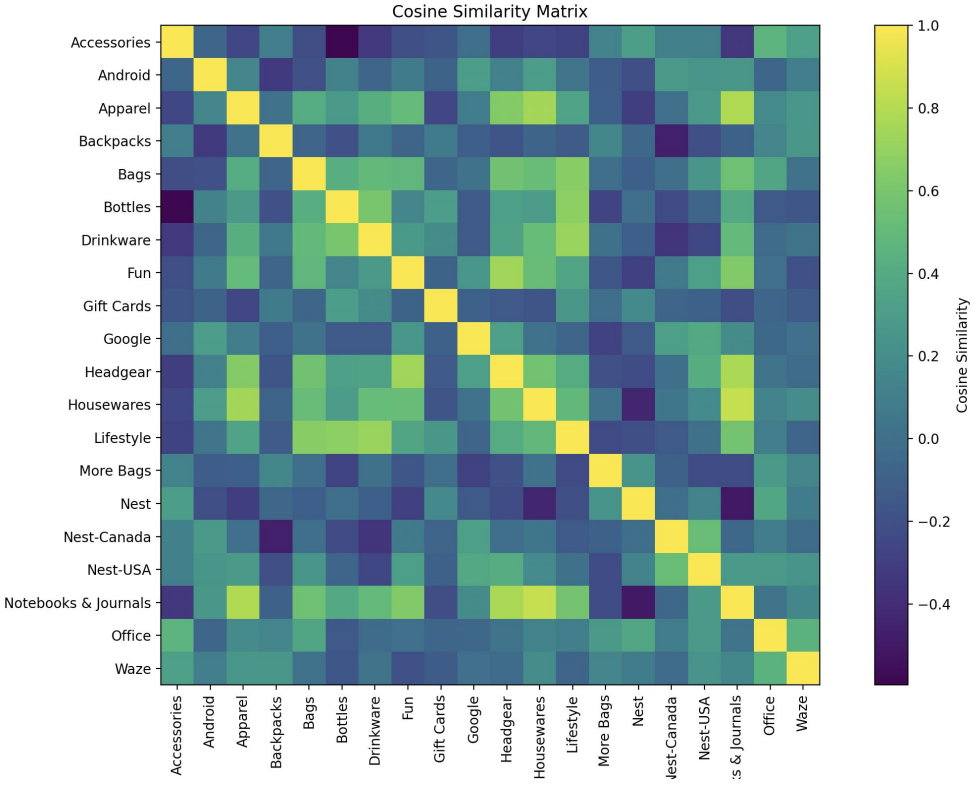




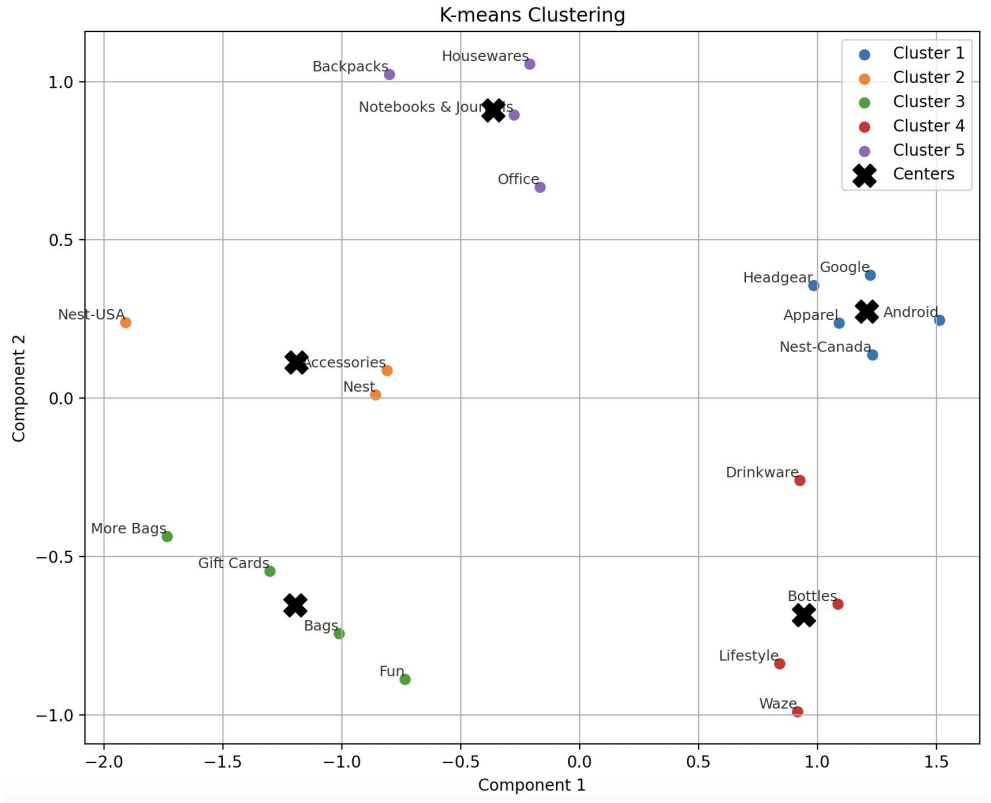




Polynomial Fit:Cosine Similarity Matrix



Polynomial Fit: Clustering



Fourier Fit: Formula

$$a_0 + a_1 x^{b_1} + \cdots + a_n x^{b_n}$$

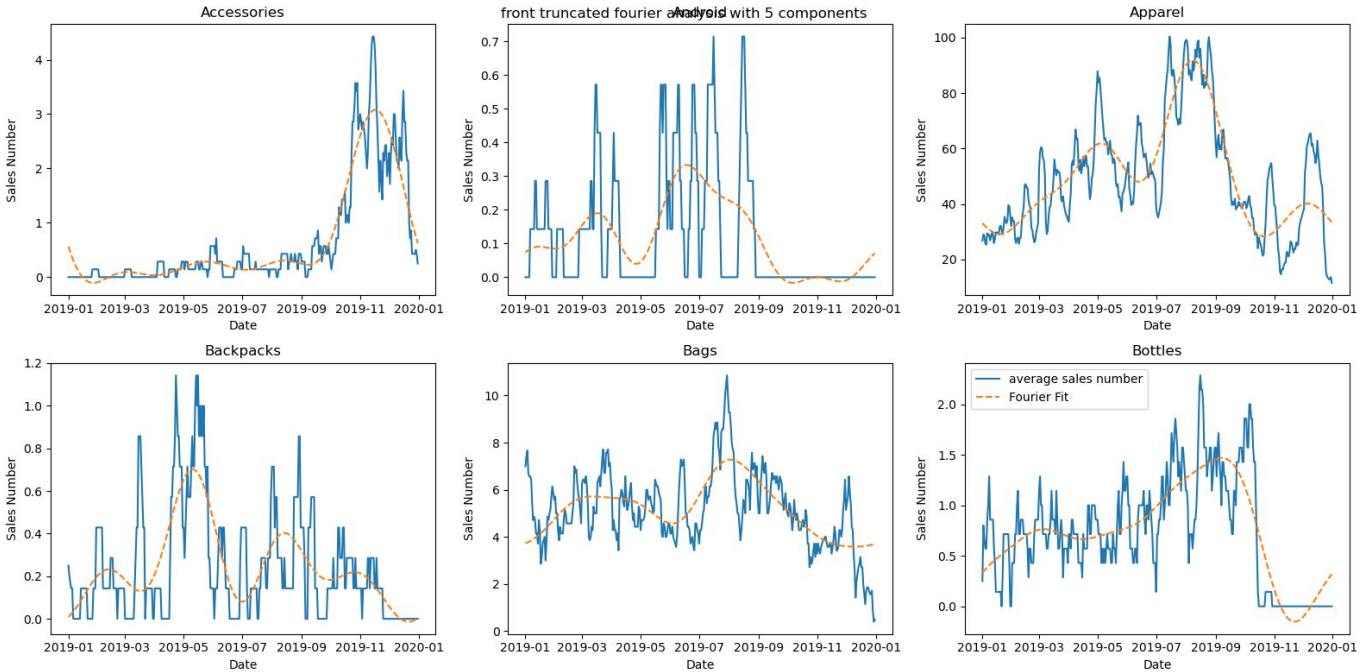
$$\downarrow$$

$$a_0 + a_1 \cdot \sin(\frac{x}{b_1} + c_1) + \cdots + a_n \cdot \sin(\frac{x}{b_n} + c_n)$$

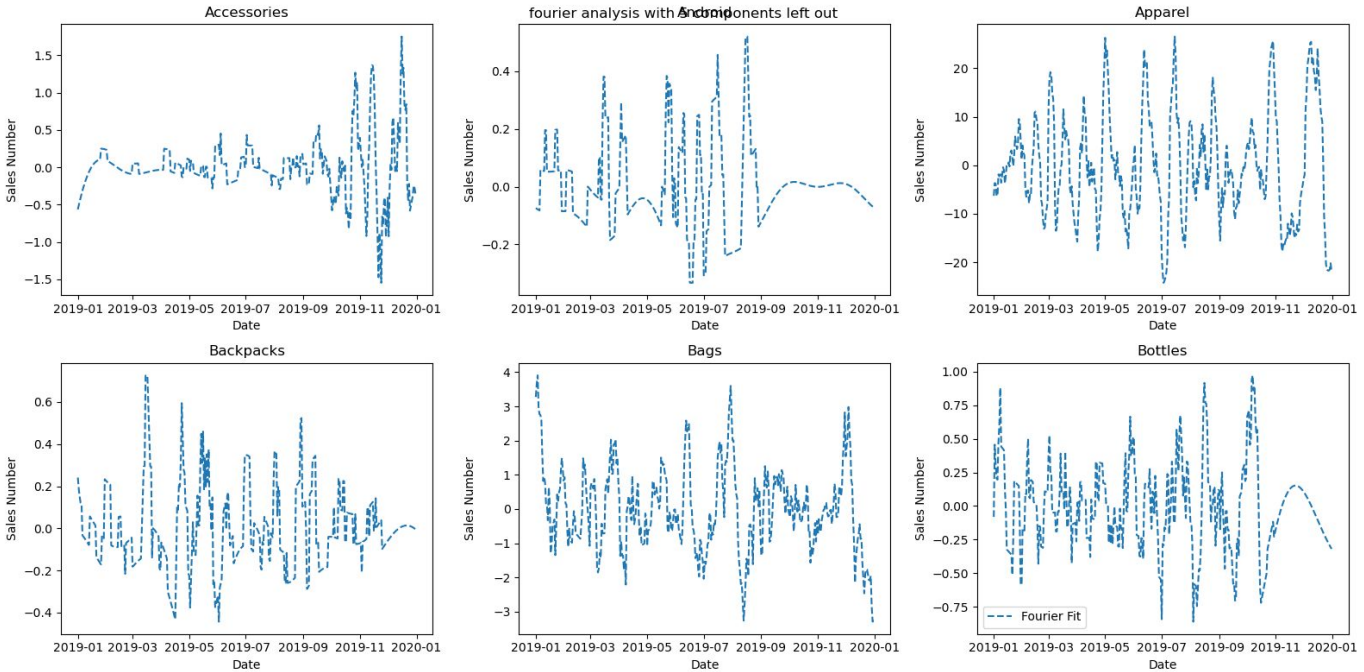
Fourier Fit

- Start by performing the fit
- **remove large trend components**
- Cosine Similarity Matrix
- Clustering

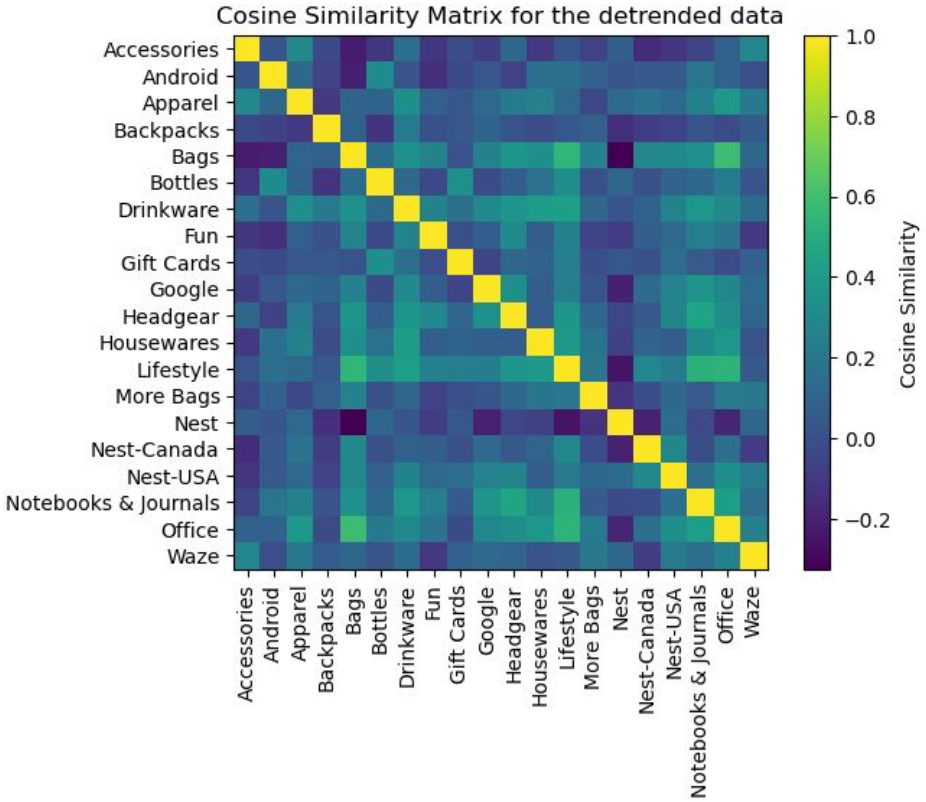
Fourier Fit: Performing the Fit



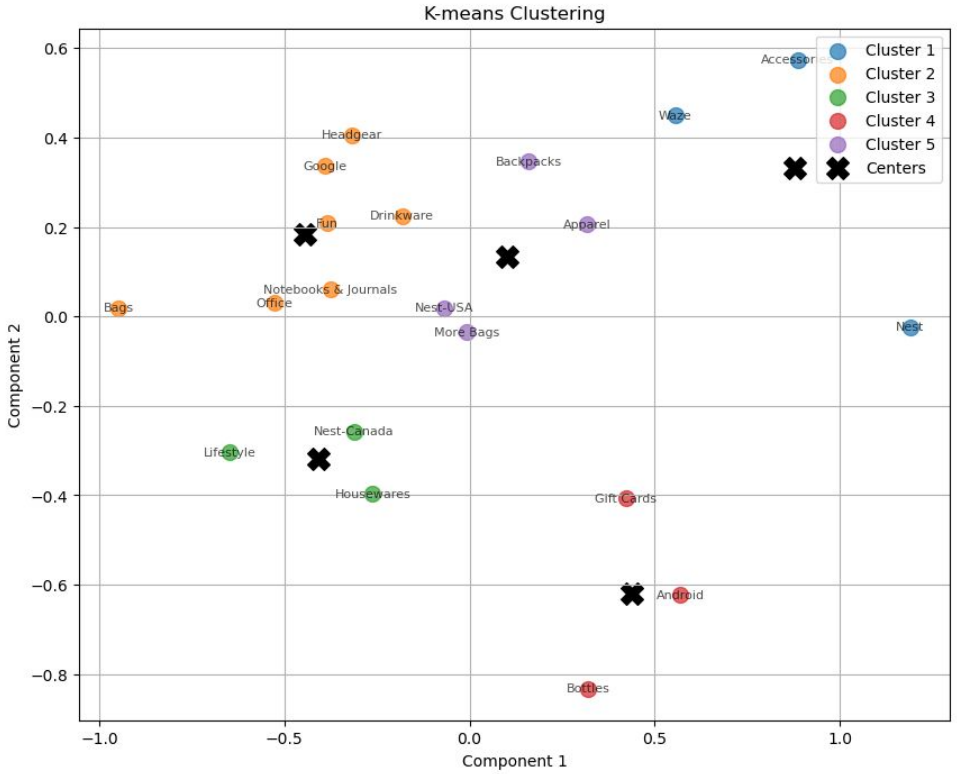
Fourier Fit: Remove Large Trends



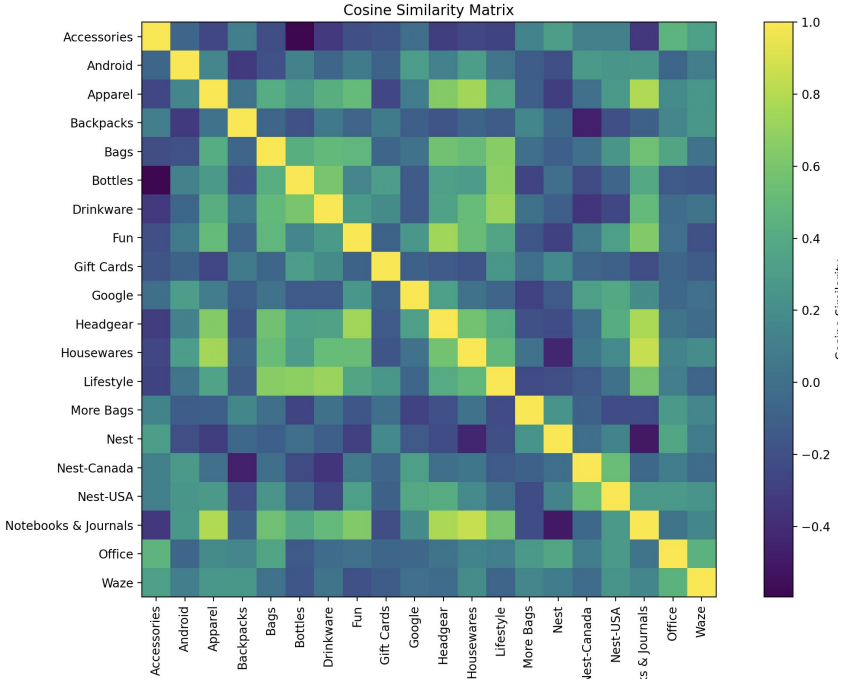
Fourier Fit: Cosine Similarity Matrix



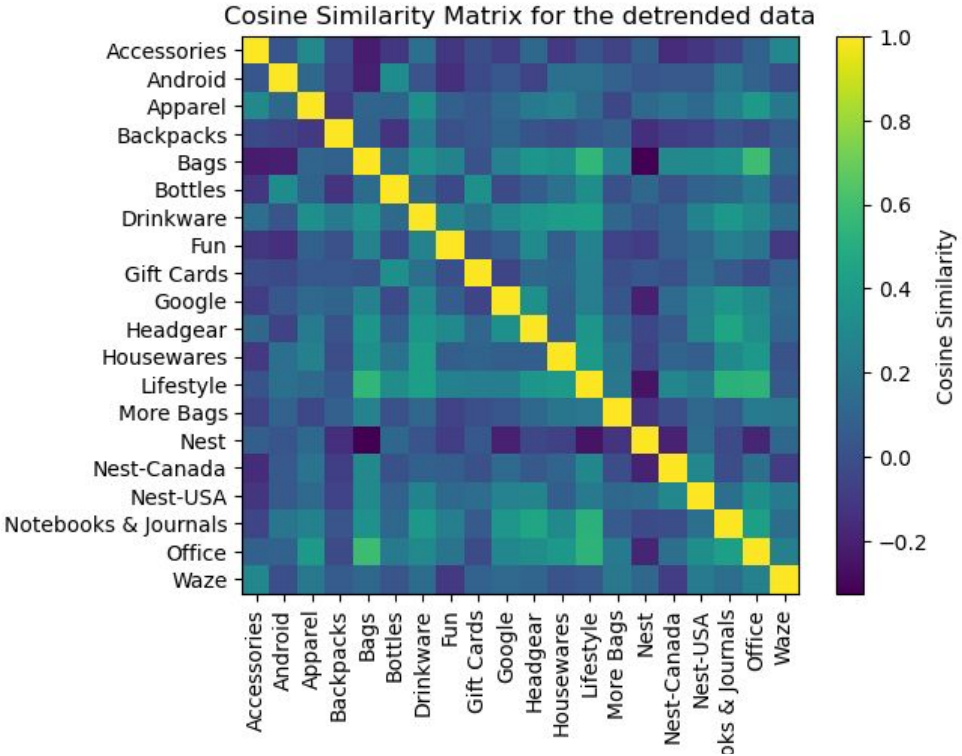
Fourier Fit: Clustering



Comparison: Similarity Matrix

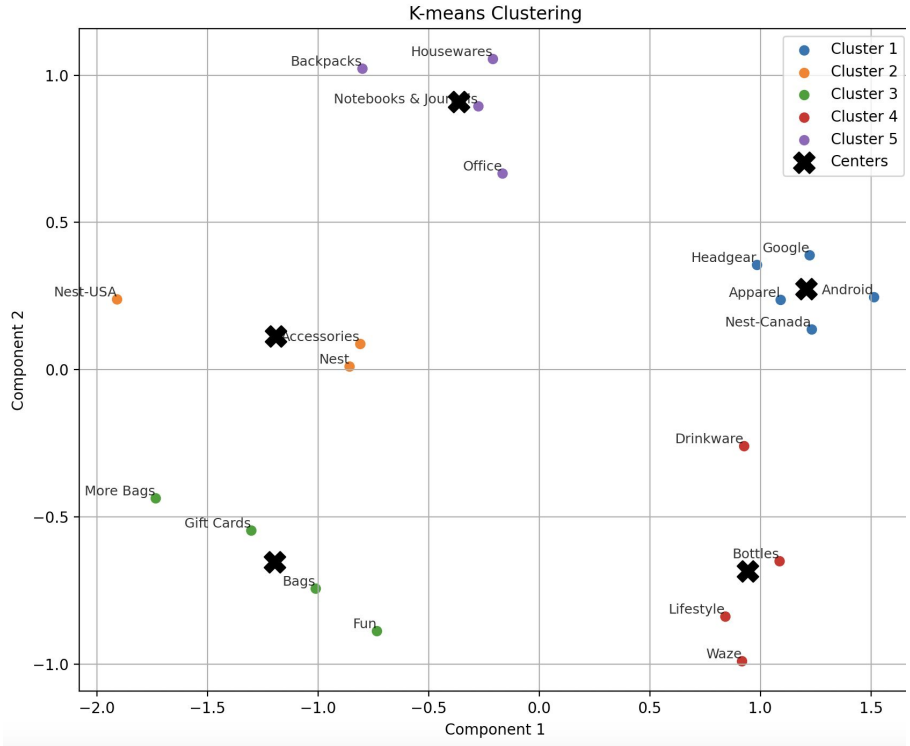


Polynomial Regression

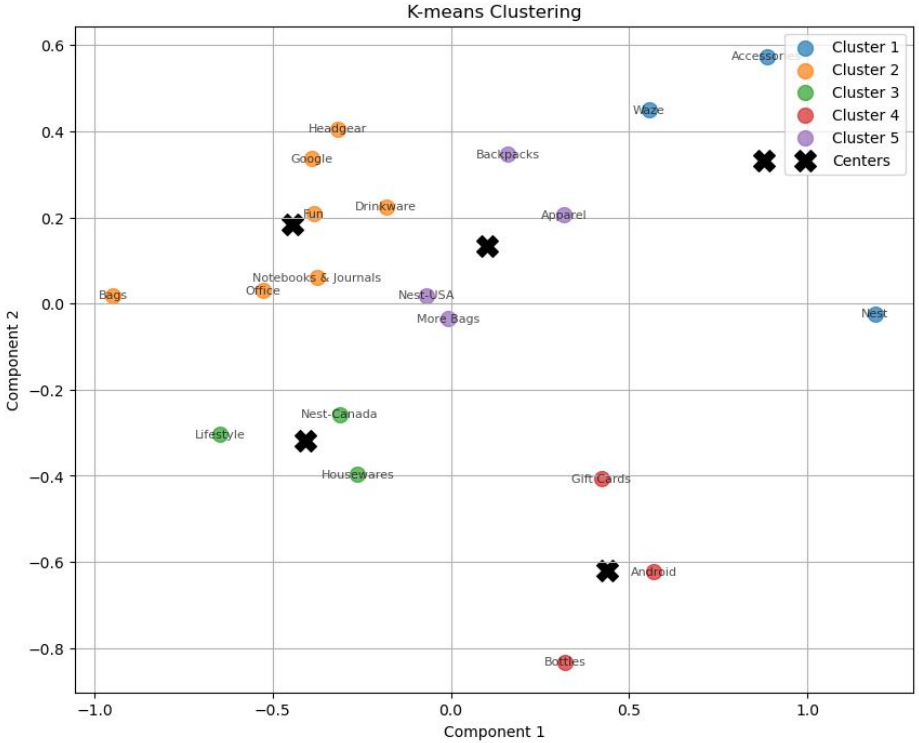


Fourier Fit

Comparison: Clustering



Polynomial Regression



Fourier Fit

Clustering

- objective: put points in the same cluster close together, separate the clusters
- how? **k-means**

  initialize random centers → assign points to clusters → update centroids → repeat → final clusters

- choosing *k* value
- how do we know if the fit is good? **k-s test**

Conclusion: Can we reject H0?

- H0_a1: The residuals from the polynomial fit follow a standard normal distribution.
  - **Reject since there is not enough evidence found**

```
KS test for polynomial residuals:
Total categories: 20
categories passing KS test (p > 0.05): 3
categories failing KS test (p ≤ 0.05): 17
passed Categories: ['Android', 'Drinkware', 'Office']
failed Categories: ['Accessories', 'Apparel', 'Backpacks'
```

- H0_a2: The residuals from the Fourier fit it follow a standard normal distribution.
  - **Reject since there is not enough evidence found**

```
KS test for fourier residuals:
total categories: 20
categories passing KS test (p > 0.05): 3
categories failing KS test (p ≤ 0.05): 17
passed Categories: ['Drinkware', 'Nest-Canada', 'Waze']
failed Categories: ['Accessories', 'Android', 'Apparel', '
```

- H0_b: There is no significant difference in the clustering performance of the polynomial fit and Fourier Fit.
  - **Reject since the ARI is around zero**

```
[0, 0, 0, 0, 0, 1, 1, 1, 2, 2, 2, 2, 3, 3, 3, 3, 4, 4, 4, 4]
[4, 1, 1, 2, 3, 4, 0, 0, 1, 1, 3, 4, 1, 2, 0, 3, 1, 1, 4, 2]
Adjusted Rand Index (ARI): -0.06782244236997523
```

Conclusion: Answer to Research Question

How valid are the two methods?

- the methods don't agree
- one or both are **wrong**?
- both captured different aspects?
- hybrid technique

Q&A