Purchase Prediction

Group 8

- Joyce Huang
- Adrina Garrett
- Maxwell Chang
- Israa Aljarb
- Lindong Ye
- Mengqi Zhang

Introduction

This project aims at analyzing the content of an E-commerce database that lists purchases made by ~ 4000 customers over a period of one year from 12/01/2010 to 12/09/2011. Based on this analysis, we developed a model that allows us to anticipate the purchases that will be made by a new customer, during the following year and from their first purchase.

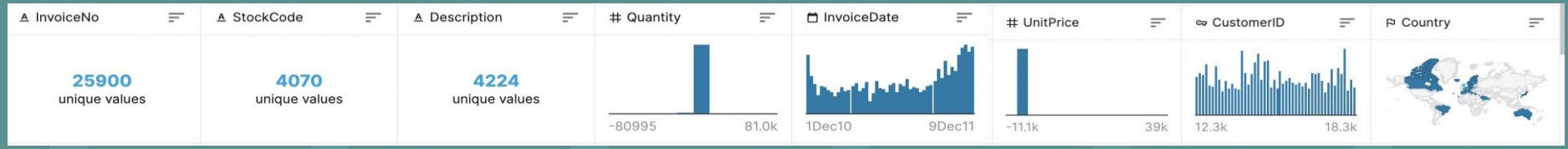


Fig1. Dataset Summary

Methodology

- ❑ Data Preparation
- ❑ Customer categories
- ❑ Classifying customers
- ❑ Testing the predictions
- ❑ Conclusion

Data cleaning

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
0	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	2010-12-01 08:26:00	2.55	17850	United Kingdom
1	536365	71053	WHITE METAL LANTERN	6	2010-12-01 08:26:00	3.39	17850	United Kingdom
2	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	2010-12-01 08:26:00	2.75	17850	United Kingdom
3	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	2010-12-01 08:26:00	3.39	17850	United Kingdom
4	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	2010-12-01 08:26:00	3.39	17850	United Kingdom
...	...	...	...	...	...	...	...	...
541904	581587	22613	PACK OF 20 SPACEBOY NAPKINS	12	2011-12-09 12:50:00	0.85	12680	France
541905	581587	22899	CHILDREN'S APRON DOLLY GIRL	6	2011-12-09 12:50:00	2.10	12680	France
541906	581587	23254	CHILDRENS CUTLERY DOLLY GIRL	4	2011-12-09 12:50:00	4.15	12680	France
541907	581587	23255	CHILDRENS CUTLERY CIRCUS PARADE	4	2011-12-09 12:50:00	4.15	12680	France
541908	581587	22138	BAKING SET 9 PIECE RETROSPOT	3	2011-12-09 12:50:00	4.95	12680	France

Fig2. Purchase data prior to cleaning

Data Cleaning: Null Values

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
column type	object	object	object	int64	datetime64[ns]	float64	object	object
null values (nb)	0	0	1454	0	0	0	135080	0
null values (%)	0.0	0.0	0.268311	0.0	0.0	0.0	24.926694	0.0

Fig3. Null value percentage

- Description - 0.268311%
- Customer ID- 24.926694%

Data Exploration: Canceled Orders

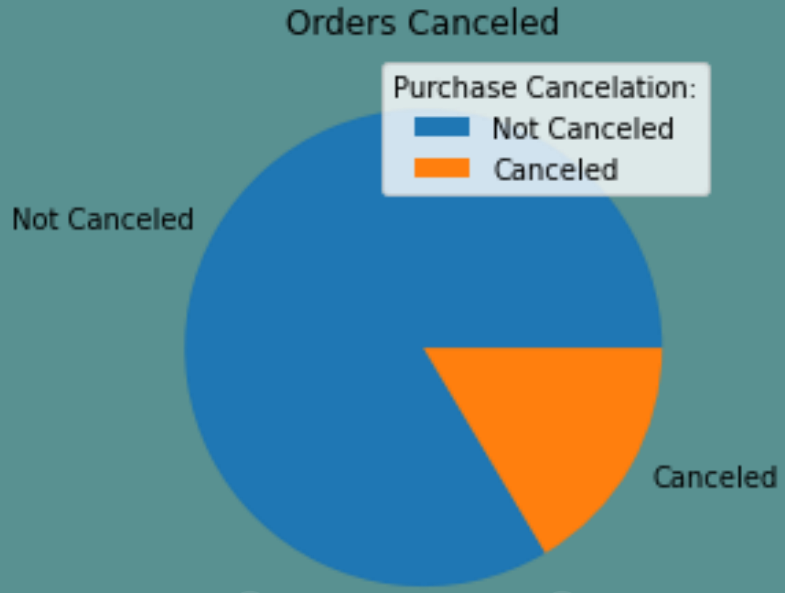


Fig4. Order canceled vs Not canceled

- Orders were canceled slightly over 16%

Data Exploration: Basket Price

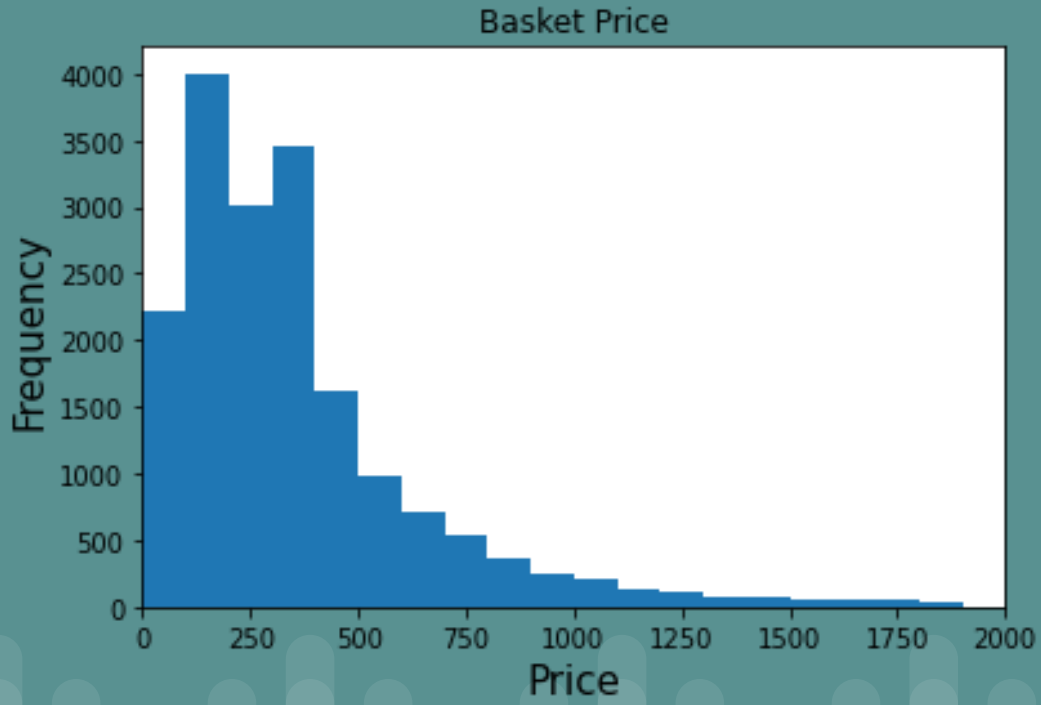


Fig5. Basket Price Frequency Distribution

Customer Categories

In this part, we aim to focus on finding characteristics related to customers , which hopefully would help predict new customer purchase or purchase recommendation.

Data Exploration: Country

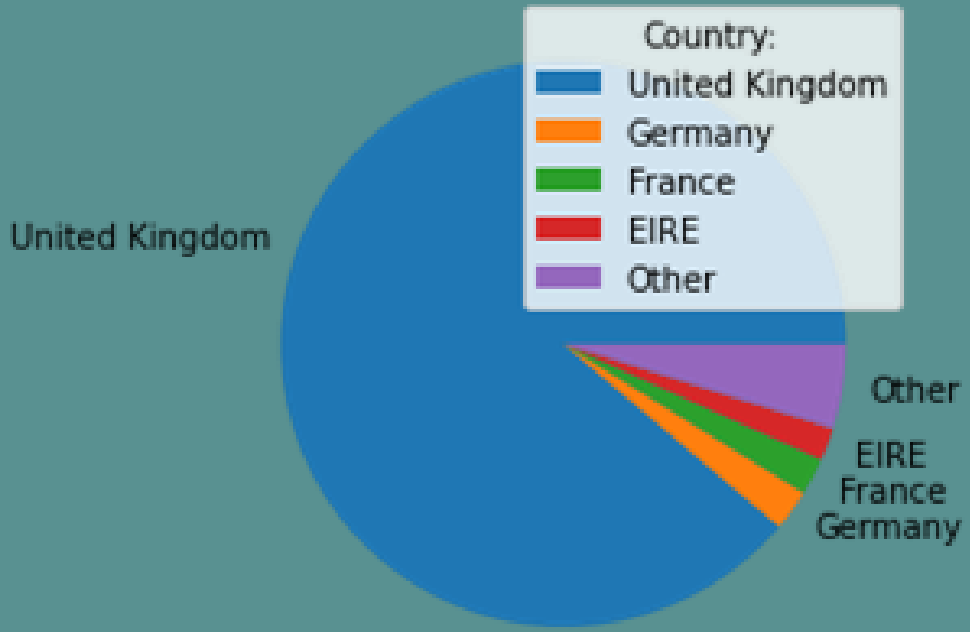


Fig6. Purchase by Country.

- Great majority of purchases are make in the U.K.
- Germany, France, and Ireland are most prevalent after the U.K.

Customer demographic

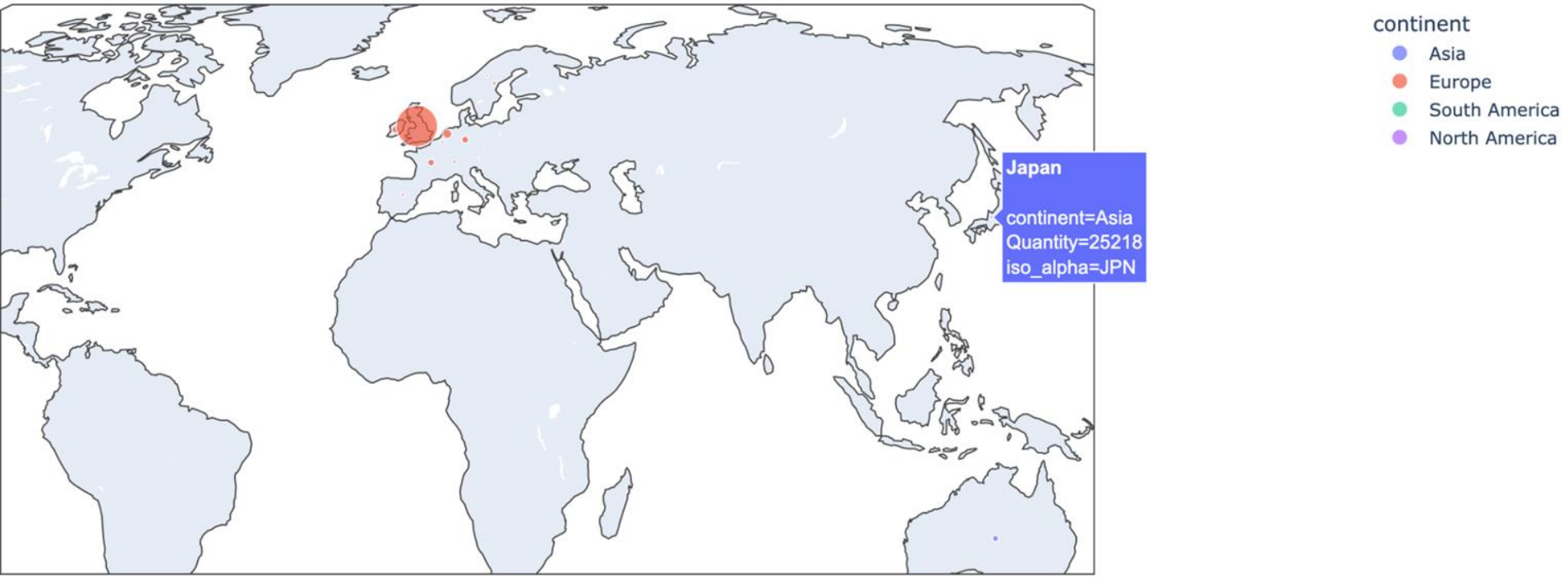


Fig7. Customer geographical distribution

Word occurence

Interesting findings

- Customers who make only one purchase.
  - This type of customer represents 1/3 of the customers.
- Business corresponding strategies
  - Target these customers in order to retain them.
  - Focusing on the sustainability of loyal customers.

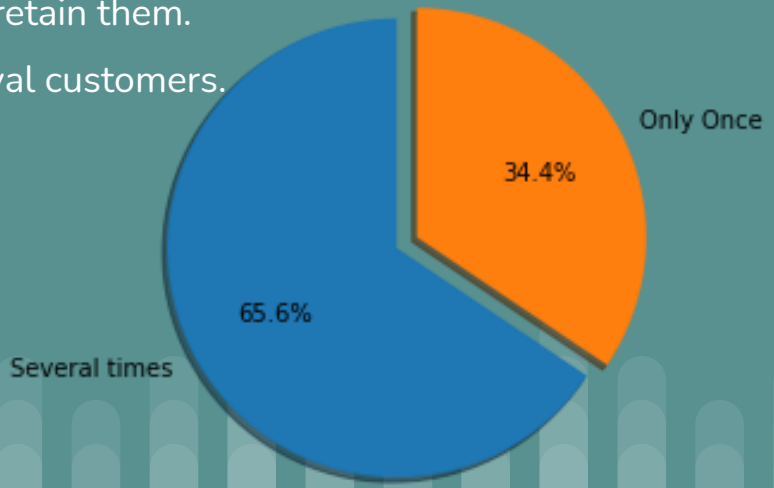


Fig8. Purchase once percentage.

Radar charts

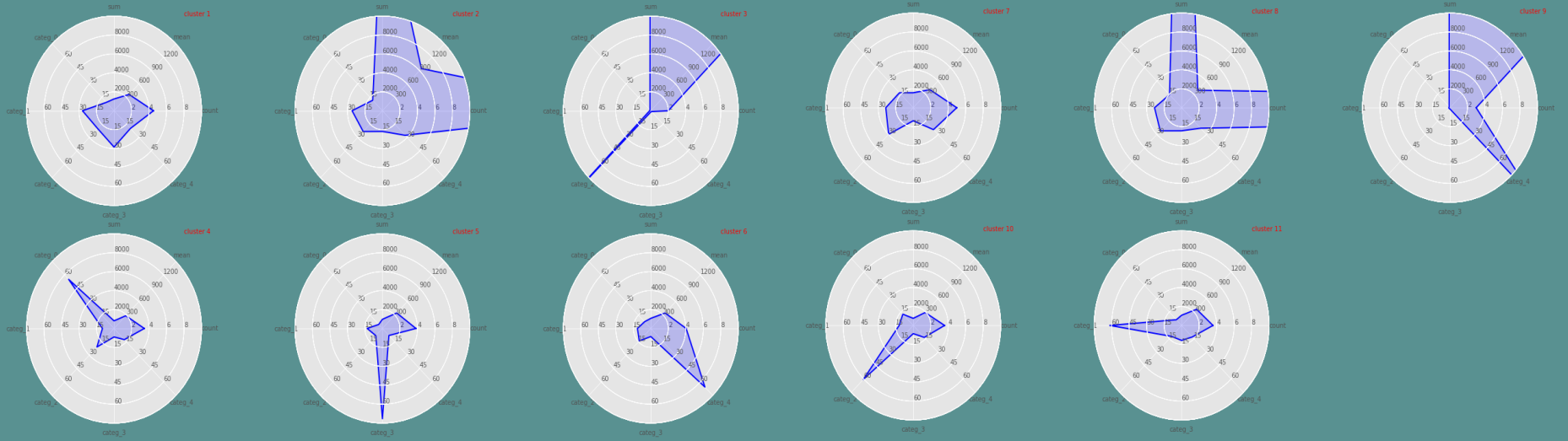
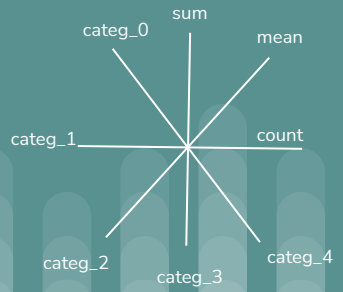


Fig9. Customer Radar chart

- Classify customers into 11 categories (using k-means)
- Each category has unique and specific consuming behaviors and habits



Time series analysis using plotly

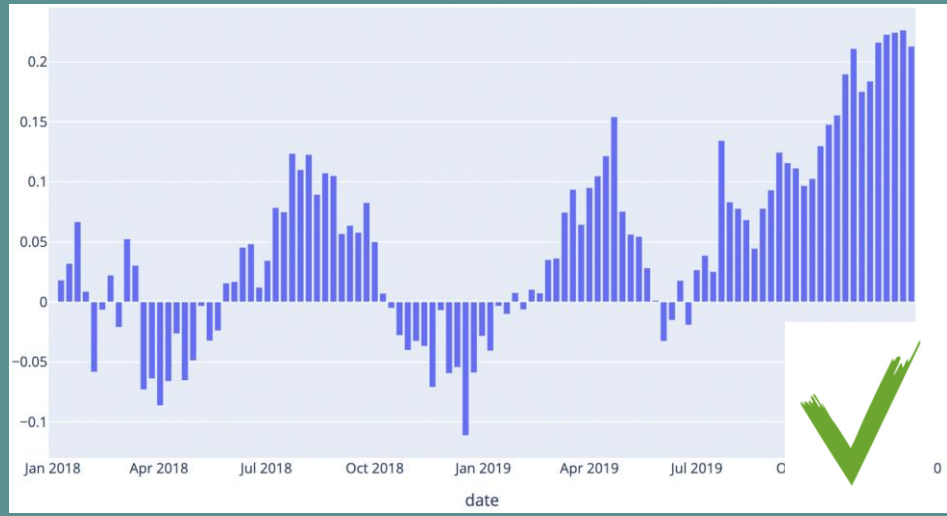


Fig10. Time series analysis comparison

United Kingdom v.s. Germany (total_price)

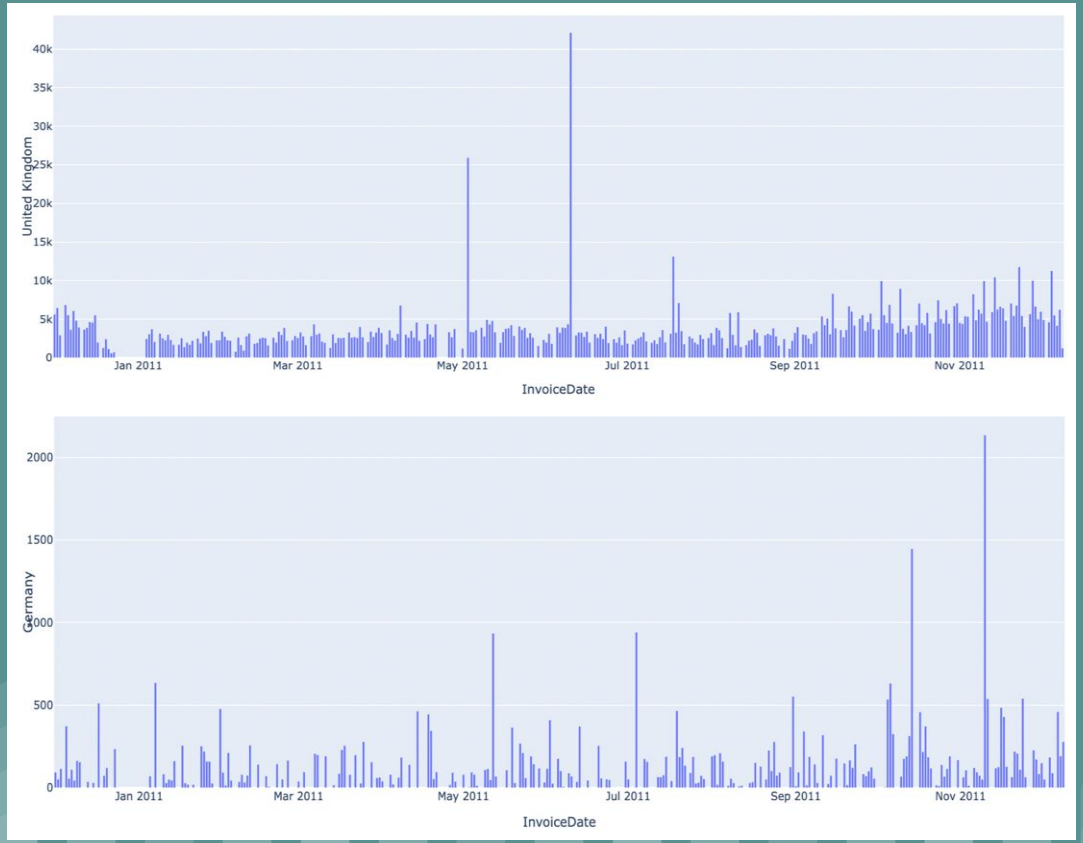


Fig11. Total price comparison between UK and Germany

Customer Type Prediction

- Dataset X: customer behavior variables
- Dataset Y: customer categories
- Dataset Split: 70% For train, 30% for validation
- Model: Logistic Regression
- Average Accuracy: 89.89%

Fig12. Logistic Regression learning curve.

Conclusion

Thank you for listening.
We are happy to answer any questions you may have.