

Demand Forecasting

Abdullah Sayyid Ayyash

Content

1. Background : Latar belakang
2. Objective
3. Summary
 - a. Resume
 - b. Recommendation
4. Data context
5. Analysis
 - a. Data Quality Check
 - b. Metrics and Definition
 - c. EDA
6. Forecasting
 - a. Dataset Prep for Modelling
 - b. Modelling Methods
 - c. Evaluation Model
 - d. Forecasting for the next 6 month
7. Conclusion & recommendation

Background

The CEO wants to know some demand forecasting, predict how the company will do in terms of sales in the next few months, and identify what products need to be stocked up.

Need analysis based on historical transaction and demand prediction on the next 6 months.

Objective

- Analysis Demand from historical transaction
- Demand Forecasting for the next 6 months, what products that need to be stocked up for the next 6 months

Summary

MONTHLY SUMMARY

Revenue have same fluctuation with quantity on every month, when the revenue increase, the quantity also increase in certain month

Revenue & Quantity Always increase Significantly on Q4 in every year (i.e. growth revenue in Q4 2010 = 73% & growth quantity Q4 2010 = 35.2%), Possibly because of Holiday Season (New Year & Christmas) so People go shopping more often

On 2009 the transaction only happen in December, so that the trend data is not showing anything, with total revenue roughly 825,000 GBP

PRODUCT & COUNTRY SUMMARY

product_id DOT (DOTCOM POSTAGE) & 85123A (HANGING HEART T-LIGHT HOLDER) always purchased by customer in every year. Those products always contribute with high revenue (Top 5 Products). On 2009 the products gave almost 36,000 GBP and the number always increase every year up to 275,000 GBP in 2011, especially for DOTCOM POSTAGE Product with 20% Growth from 2010 to 2011

UK always significantly have the highest revenue on every year. On 2009 UK revenue 750,317 GBP, Increased on 8,800,000 GBP on 2010. 2011 UK revenue is 8,300,000 GBP

DOT (DOTCOM POSTAGE), 85123A (HANGING HEART T-LIGHT HOLDER) & 22423 (REGENCY CAKESTAND 3 TIER) always on demand every month from 2010 until 2011 in UK that contribute with High revenue (Top 5)

DEMAND CLASSIFICATION

From 27744 of Product & country combination (demand profile), there is Lumpy demand pattern that dominated the combination (72%) followed by Intermittent pattern with percentage around 22%. Erratic & Smooth pattern (ideal pattern) only have 5.8% and 0.4% from total combination

Summary

DEMAND FORECAST - MODEL

Use Model XGBoost for only Erratic & Smooth demand pattern product in certain country (1726 product-country) because in this case the MAPE have lower number (28%) and still can predict top 4 product-country that give the most quantity of all time.



RECCOMENDATION

For the next 6 months (January - June 2012) would be increasing significantly around 400,000 - 500,000 qty, but on June, 2012 the demand will decreasing (around 350,000 qty) 2 of them are product_id 84077 (WORLD WAR 2 GLIDERS ASSTD DESIGNS) & 22197 (POPCORN HOLDER) in UK that have 122,185 & 104,367 demand qty, **so that the stocks of those products should be prepared.**

Data Context

Table historical transaction from 2009-12-01 until 2011-12-01

Features :

- Order_id
- Product_id
- Product_description
- quantity
- Order_date
- unit_price
- Customer_id
- country
- Revenue = quantity * price

ANALYSIS

Data Quality Check

Quantity & Unit Price have a negative value, because of order cancel (15% order).

Exclude Order cancel from the analysis because we want to focused on demand analysis that can generating value/revenue.

:	order_status	order_id	%
0	CANCEL	8292	15.462072
1	SALE	45336	84.537928

Metrics

1. Quantity : Quantity purchased by customer per product
2. Revenue : Quantity * Price
3. Number of Product-Country : Number of combination product & country
Monthly etc. (Demand Profile for Demand forecasting)
4. Demand Classification :
 - a. Classification of Demand Pattern (in this scenario monthly) per Demand Profile
 - b. Based on value ADI (regularity of order) & CV2 (variation of demand)
 - c. $ADI = \text{Maximum Period} / \text{bucket per demand profile (product \& country)}$
 - i. In this scenario maximum period = 24 months (number of month from 2010-2011)
 - ii. Bucket = months frequency per product & country that have been purchased
 - d. $CV2 = \text{coefficient of variation}(\text{Standard Deviation of Demand} / \text{average of Demand})$ per demand Profile

Demand Classification ([source](#))

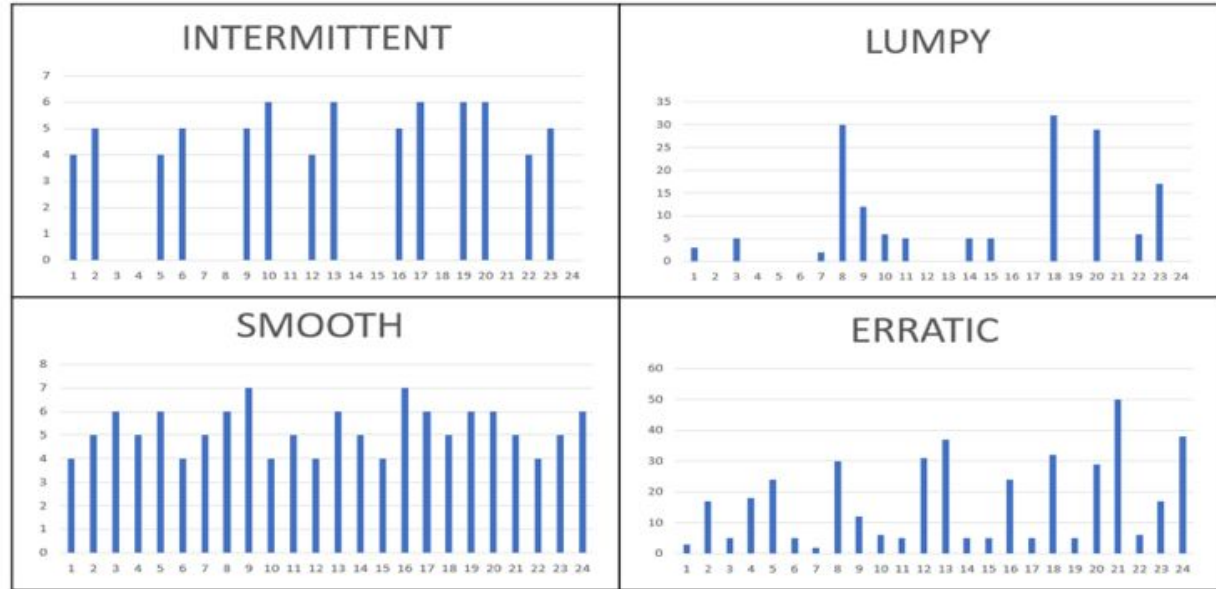
Variability in
demand timing

HIGH

ADI = 1.32

LOW

0



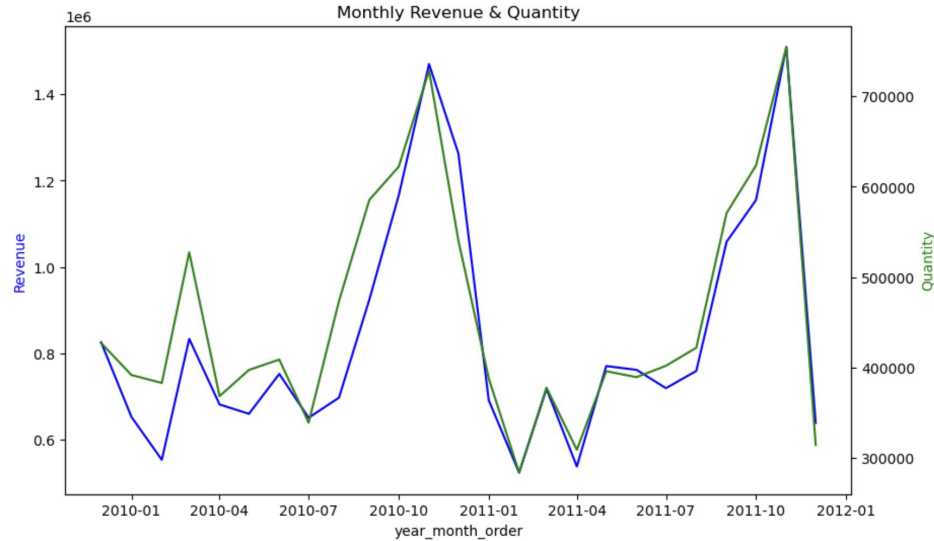
LOW

CV2 = 0.49

HIGH

Variability in
demand quantity

Monthly Revenue & Quantity

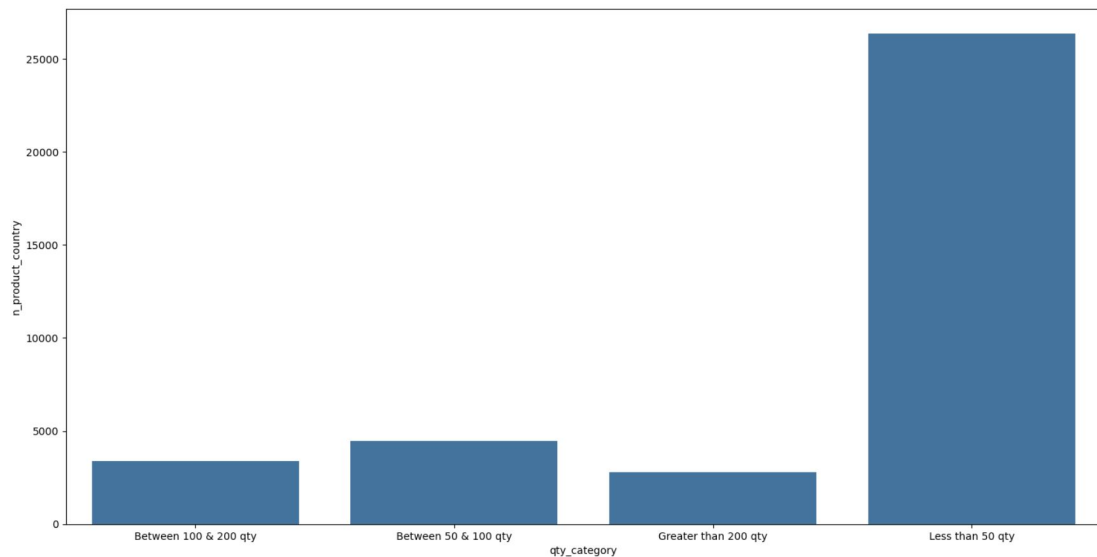


	year_quarter_order	revenue	quantity	growth_revenue	growth_quantity
3	2010-07-01	2272320.861	1397697	0.085335	0.190502
4	2010-10-01	3898355.182	1890742	0.715583	0.352755
7	2011-07-01	2536949.743	1394349	0.225530	0.274815
8	2011-10-01	3303286.310	1692157	0.302070	0.213582

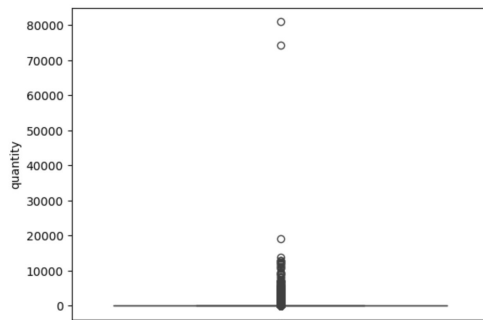
- Revenue have same fluctuation with quantity on every month, when the revenue increase, the quantity also increase in certain month

- Revenue & Quantity Always increase Significantly on Q4 in every year (i.e. growth revenue in Q4 2010 = 73% & growth quantity Q4 2010 = 35.2%), Possibly because of Holiday Season (New Year & Christmas) so People go shopping more often

Number of Product & Country per Quantity Category & Quantity Distribution

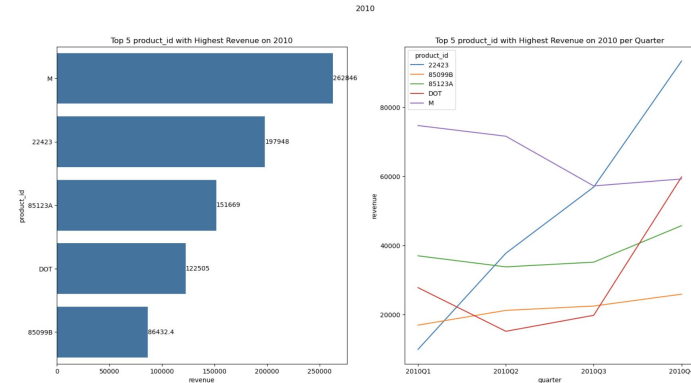
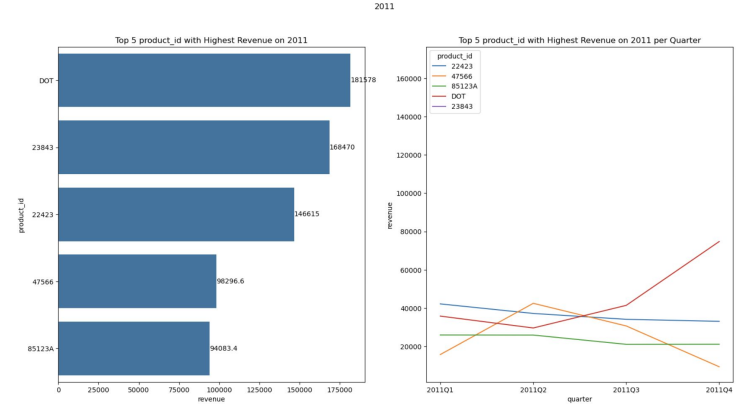
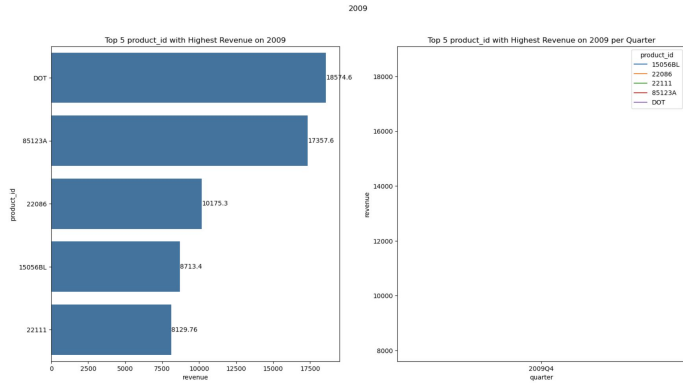


	qty_category	n_product_country	pct
0	Between 100 & 200 qty	3377	0.091297
1	Between 50 & 100 qty	4456	0.120468
2	Greater than 200 qty	2783	0.075239
3	Less than 50 qty	26373	0.712996



- Quantity have many Outliers, and after categorizing the quantity, combination of Product-Country that have qty < 50 is significantly higher than other (around 26,373 combination of product-country)
- There are 4387 Product-country (16% from all product-country combination) that have quantity > 100. 84% of the product-country already covered in other quantity category (i.e. Less than 50 qty), Meaning if we want to take out the product-country that have quantity > 100, the 84% of it still can be predicted

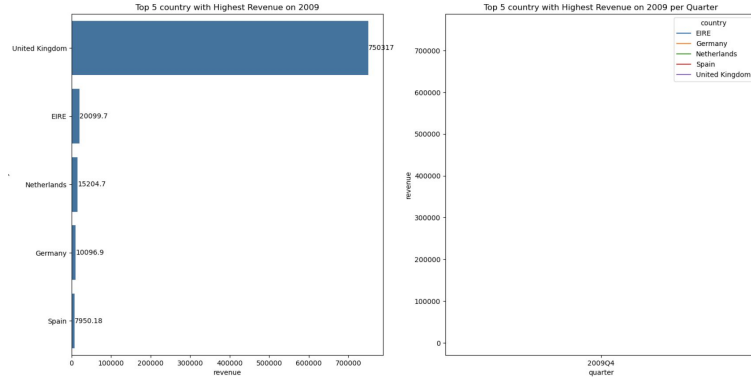
Revenue Contribution per Product in every Year & its dynamic per Quarter (Top 5 Product)



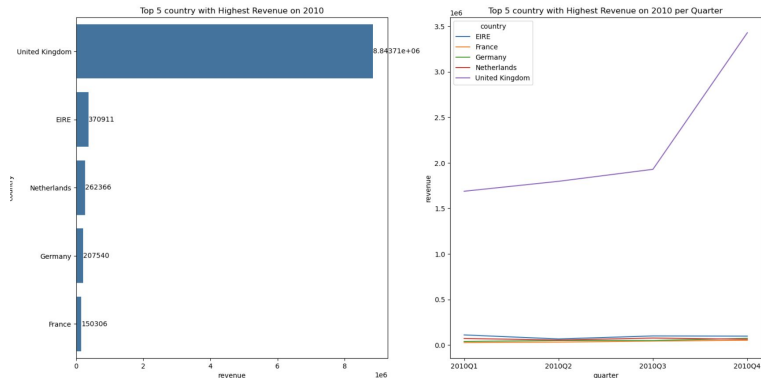
- product_id DOT (**DOTCOM POSTAGE**) & 85123A (**HANGING HEART T-LIGHT HOLDER**) always purchased by customer in every year. Those products always contribute with high revenue (Top 5 Products). On 2009 the products gave almost 36,000 GBP and the number always increase every year up to 275,000 GBP in 2011, especially for **DOTCOM POSTAGE** Product with 20% Growth from 2010 to 2011
- On 2009 the transaction only happen in December, so that the trend data is not showing anything, with total revenue roughly 825,000 GBP
- Revenue for product_id 22423 (**REGENCY CAKESTAND 3 TIER**) always increase significantly every quarter on 2010.
- All Top 5 Product in 2011 have revenue that tends to be more stable every quarter

Revenue Contribution per Country in every Year & its dynamic per Quarter (Top 5 Country)

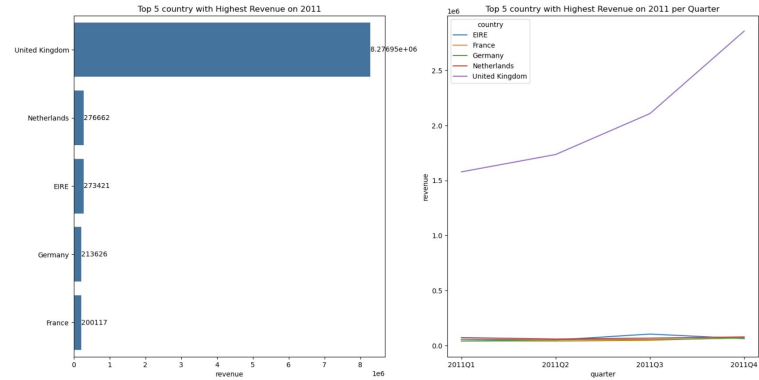
2009



2010

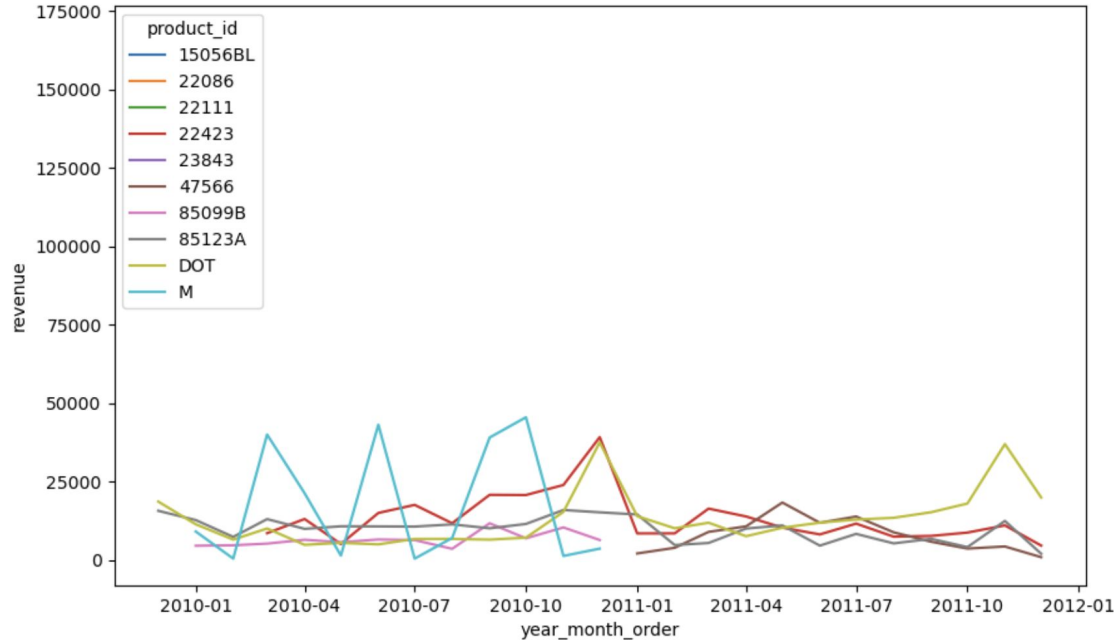


2011



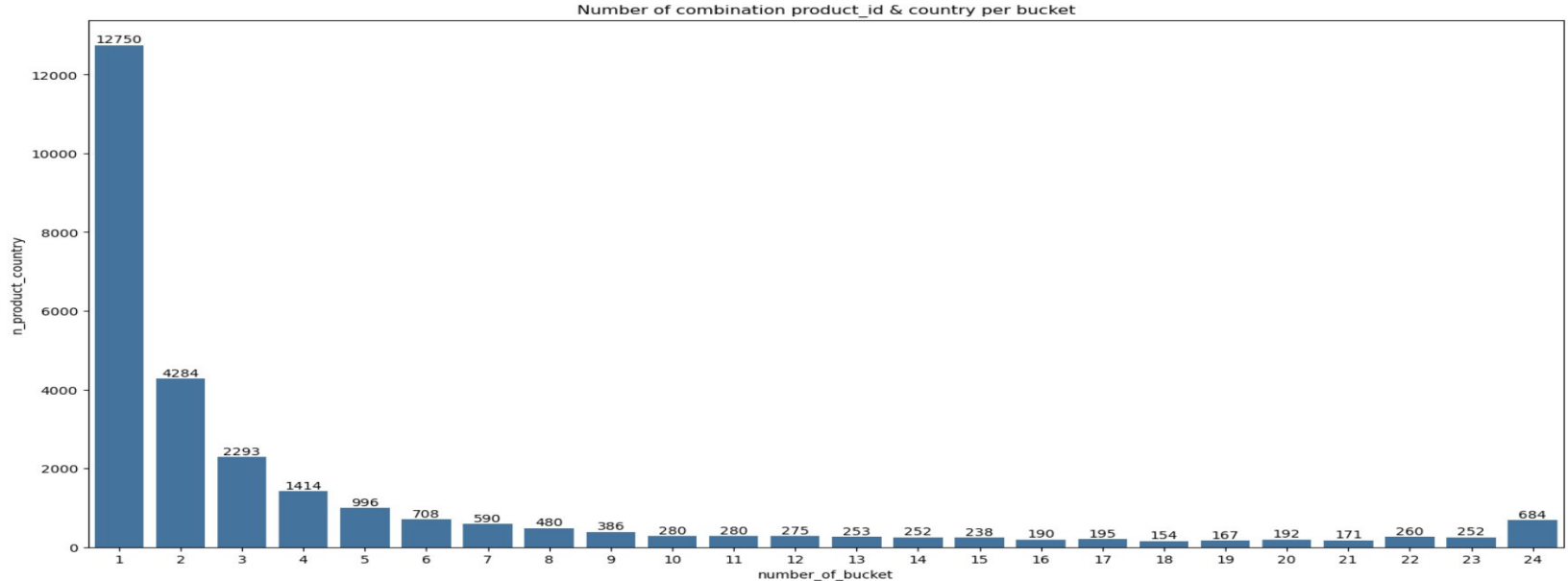
- UK always significantly have the highest revenue on every year. On 2009 UK revenue 750,317 GBP, Increased on 8,800,000 GBP on 2010. 2011 UK revenue is 8,300,000 GBP
- EIRE (Ireland), Netherlands, Germany and France always on TOP 5 country that have high contribution every year on Revenue besides UK.
- UK Revenue Always increasing on every quarter from 2010 until 2011. On the other hand, other Top 5 countries relatively stagnan on every quarter.

Monthly Revenue for Top 5 Product in UK



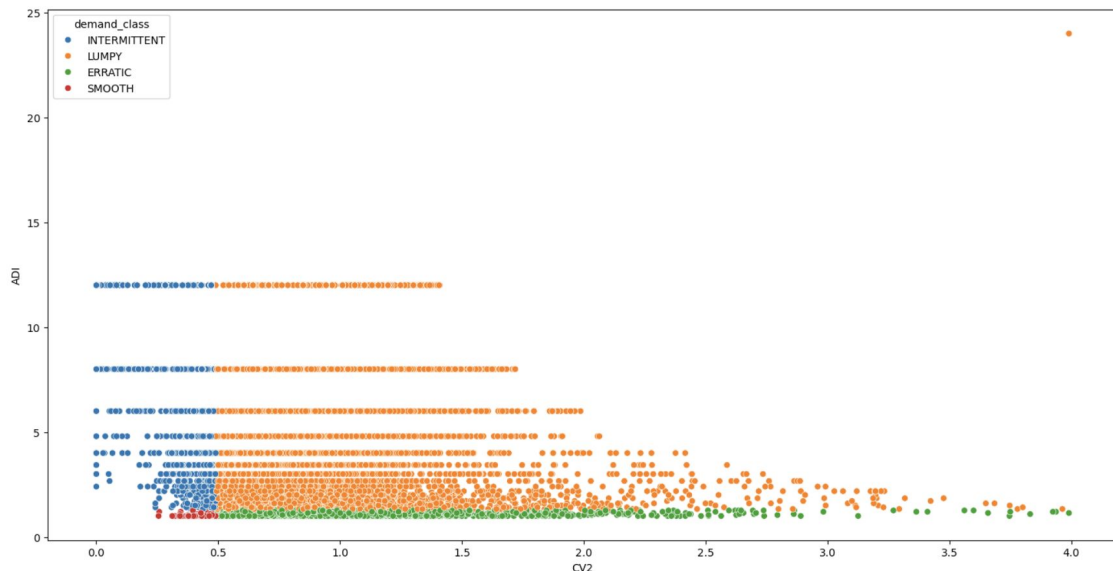
- DOT (**DOTCOM POSTAGE**), 85123A (**HANGING HEART T-LIGHT HOLDER**) & 22423 (**REGENCY CAKESTAND 3 TIER**) always on demand every month from 2010 until 2011 in UK that contribute with High revenue (Top 5)
- M (**MANUAL**) & 85099B (**JUMBO BAG**) products are contributing with high revenue in UK (Top 5) on 2010, but on 2011 those products not in Top 5
- 47566 (**PARTY BUNTING**) product not on TOP 5 product that contribute with high revenue in UK, but on 2011 these product on Top 5 that contribute with high revenue on every month
- **DOTCOM POSTAGE** Product give the Highest revenue on the end of the year from 2010 to 2011 in UK with value around 37,000 GBP each

Number of Combination Product & country per Bucket



1. There Are **685 Product & Country Combination (around 2.5% from all combination)** that routinely purchased by customers every month from 2010 until 2011. In example **Product POST in Germany, Product POST in Spain, etc.**
2. Bucket 1 have the highest value compare to others (12750) which means there are **12750 Product & Country Combination** that only in one month purchased by customer from 2010 until 2011, i.e. **Product gift_0001_70 in UK, POST in Canada, etc.**

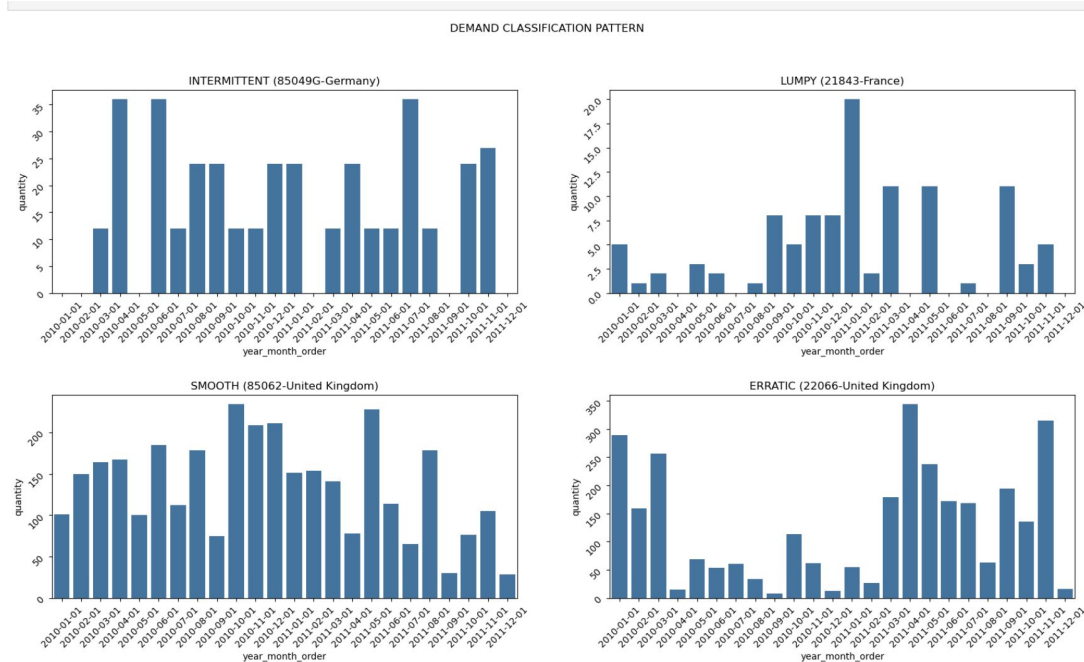
Demand Classification on every Product & Country



	demand_class	product_country	pct
0	ERRATIC	1607	0.057922
1	INTERMITTENT	6012	0.216696
2	LUMPY	20006	0.721093
3	SMOOTH	119	0.004289

- From 27744 of Product & country combination (demand profile), there is Lumpy demand pattern that dominated the combination (72%) followed by Intermittent pattern with percentage around 22%.
- Erratic & Smooth pattern (ideal pattern) only have 5.8% and 0.4% from total combination

Demand Classification on every Product & Country (Example)



- One of example of combination product & country in every demand pattern are :
 - Intermittent : **85049G-Germany**
 - Lumpy : **21843-France**
 - Smooth : **85062-United Kingdom**
 - Erratic : **22066-United Kingdom**

FORECASTING & MODELLING

Dataset Preparation for Model

Based on EDA, Here is some Data that can be include for modelling :

1. GRANULARITY : Year month order, product_id & country (because we want to predict the demand for the next 6 month)
2. FILTER : order not cancel, exclude product_country that only purchased in one month in every country (bucket = 1), order from 2010 until 2011 (because on 2009 only have 1 month (December))
3. FEATURES :
 - a. Year_month_order = Truncate Year & month `order_date`
 - b. Year_order
 - c. month_order
 - d. quarter_order
 - e. season (Winter, Fall, etc.)
 - f. product_id
 - g. country
 - h. revenue
 - i. Pre_year_quantity : previous quantity per product & country in last year for certain month
 - j. Pre_year_revenue : previous revenue per product & country in last year for certain month
 - k. Pre_year_unit_price : previous price (median) per product & country in last year for certain month
 - l. Pre_year_n_customer : previous total customer per product & country in last year for certain month
 - m. Pre_year_n_order : previous total order per product & country in last year for certain month
 - n. quantity

Dataset Preparation for Model

Use 2 type of Dataset

1. All demand pattern (df_model_1)
2. only product & country that have smooth & erratic demand pattern (df_model_2)

And after that we will evaluate which case that can be forecast from our model

Modelling Methods

1. Train & Test data split
2. Numerical Columns handling
3. Categorical Columns handling
4. Machine learning

Modelling Methods - Train & Test data split

In this case, we want to forecast for the next six month which product in certain country that have high demand or not, in conclusion

Train Data : All data transaction **except last 3 months data**

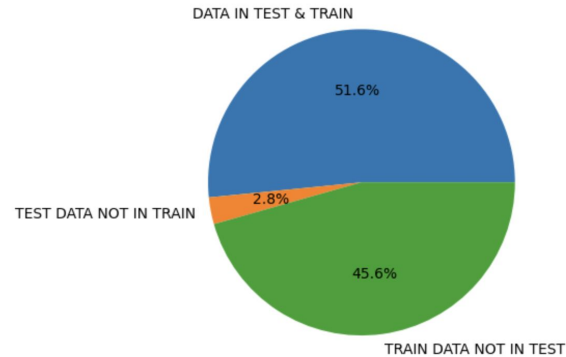
Test Data : Last 3 months transaction data

BUT

We need to see the inclusivity of product-country in test data compare to train data

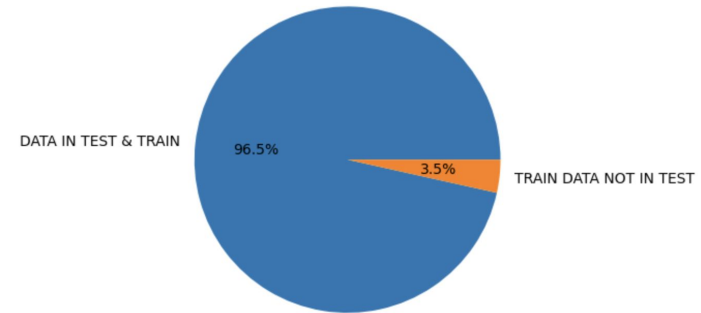
Modelling Methods - Train & Test data split - Inclusivity of product & country

All demand dataset (df_model_1)



	category	n
0	DATA IN TEST & TRAIN	7733
1	TEST DATA NOT IN TRAIN	425
2	TRAIN DATA NOT IN TEST	6836

Smooth & Erratic demand dataset (df_model_2)



	category	n
0	DATA IN TEST & TRAIN	1666
1	TRAIN DATA NOT IN TEST	60

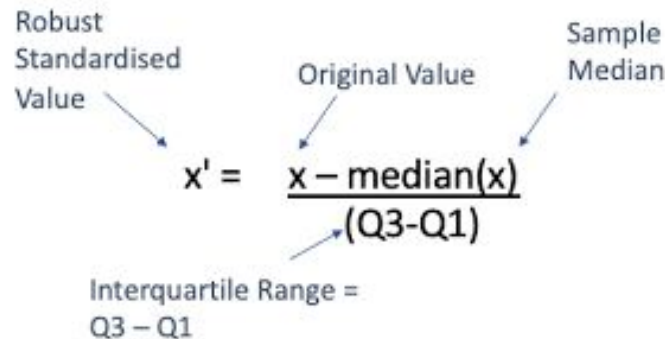
- For all demand dataset (df_model_1) there are 2.8% of product-country that in Test Data but not in Train Data. We should **exclude the data from test data** so that data not disturbs the process of evaluation model
- For Smooth & erratic demand dataset (df_model_2) all product-country in test data are in train data

Modelling Methods - Numerical columns handling

Numerical columns for model training are **revenue**, **'pre_6_quantity'**, **'pre_6_revenue'**, **'pre_6_unit_price'**, **'pre_6_n_customer'**, **'pre_6_n_order'**, **'pre_year_quantity'**, **'pre_year_revenue'**, **'pre_year_unit_price'**, **'pre_year_n_customer'**, **'pre_year_n_order'**

Use Robust Scalling, because

- most of the numerical columns data are Skewed
- Robust scaler are median based data scaling



The diagram illustrates the formula for Robust Standardised Value. It shows the equation $x' = \frac{x - \text{median}(x)}{(Q3 - Q1)}$ with arrows pointing from descriptive labels to the components of the formula. 'Robust Standardised Value' points to x' . 'Original Value' points to x . 'Sample Median' points to $\text{median}(x)$. 'Interquartile Range = $Q3 - Q1$ ' points to the denominator $(Q3 - Q1)$.

$$\text{Robust Standardised Value } x' = \frac{\text{Original Value } x - \text{Sample Median } \text{median}(x)}{\text{Interquartile Range } (Q3 - Q1)}$$

Modelling Methods - Categorical columns handling

Categorical columns for model training are 'year_order', 'quarter_order', 'month_order', 'season', 'product_id', 'country'

For categorical columns handling we can use many encoder methods, one of them are **One Hot Encoding** (pic)

But it depends on which machine learning we will use, because some of machine learning have parameters that can handling categorical columns directly (i.e. **XGBoost**)

ONE HOT
ENCODING

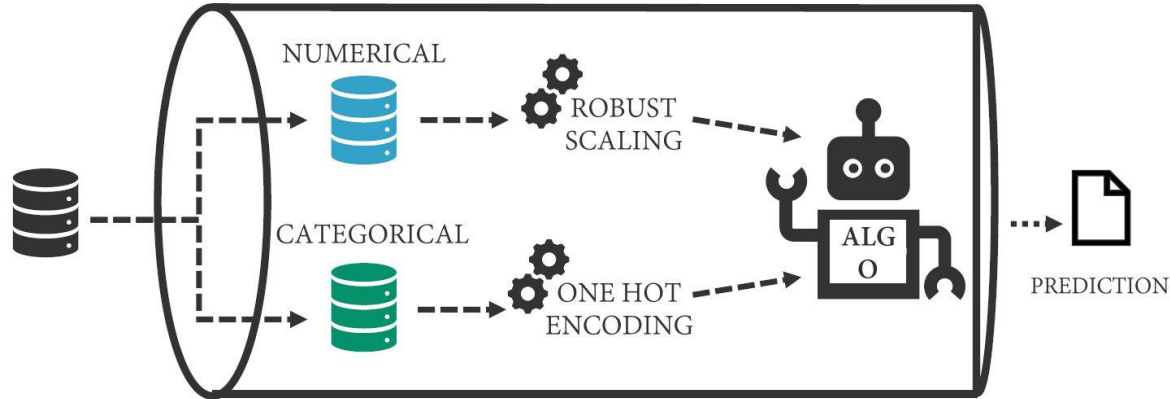
A	X	Y	Z
X	1	0	0
Y	0	1	0
Z	0	0	1

Modelling Methods - Machine Learning

For Machine Learning we use XGBoost because it is commonly use for forecasting time series data ([source](#)). This case is regression problem (predict quantity) so we use XGBRegressor()

Because XGBoost is used, then **no need to encode numerical columns**. (use parameter *enable_categorical=True*)

Also Use Pipelining Method to summarize preprocessing data & machine learning (pic below for example)



Evaluation Model

1. Metric to evaluate Model
2. Result Modeling - Evaluation Metric (decide which model we will use)
3. + & - of Model chosen
4. Feature Importance
5. Prediction Result

Evaluation Model - Metric to evaluate Model

MAPE (Mean Absolute Percentage Error)

Using MAPE because it is more easy to interpret how far the value prediction from the actual

$$MAPE = \frac{1}{n} \sum_{i=1}^n \frac{|A_i - F_i|}{A_i}$$

A_i is the actual value

F_i is the forecast value

n is total number of observations



Evaluation Model - Result Modeling

	XGB 1Testing	XGB 2Testing
R2	0.132648	0.829477
MAPE	0.839096	0.283917

- XGB 1 (model that use all demand data) have MAPE 0.83 meaning the Prediction possibly +/- 80% far from actual value, on the other hand XGB 2 (model that use **only smooth and erratic data**) have lower MAPE significantly around 0.28 meaning the Prediction possibly +/- 20% far from actual value
- **Use XGB 2**

Evaluation Model - [+ & - of Model chosen]

+

Have lower error to predict the demand

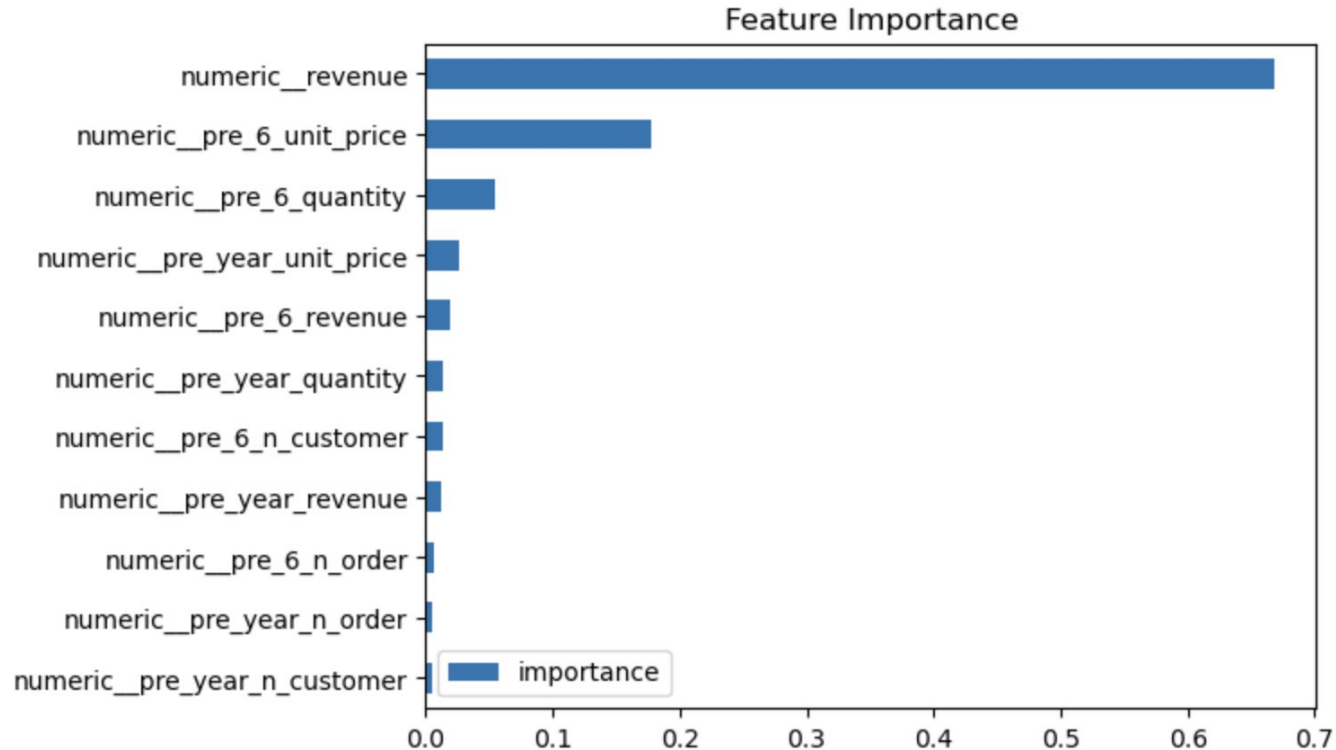
Still can predict top 4 product-country that give the most quantity of all time (pict)

-

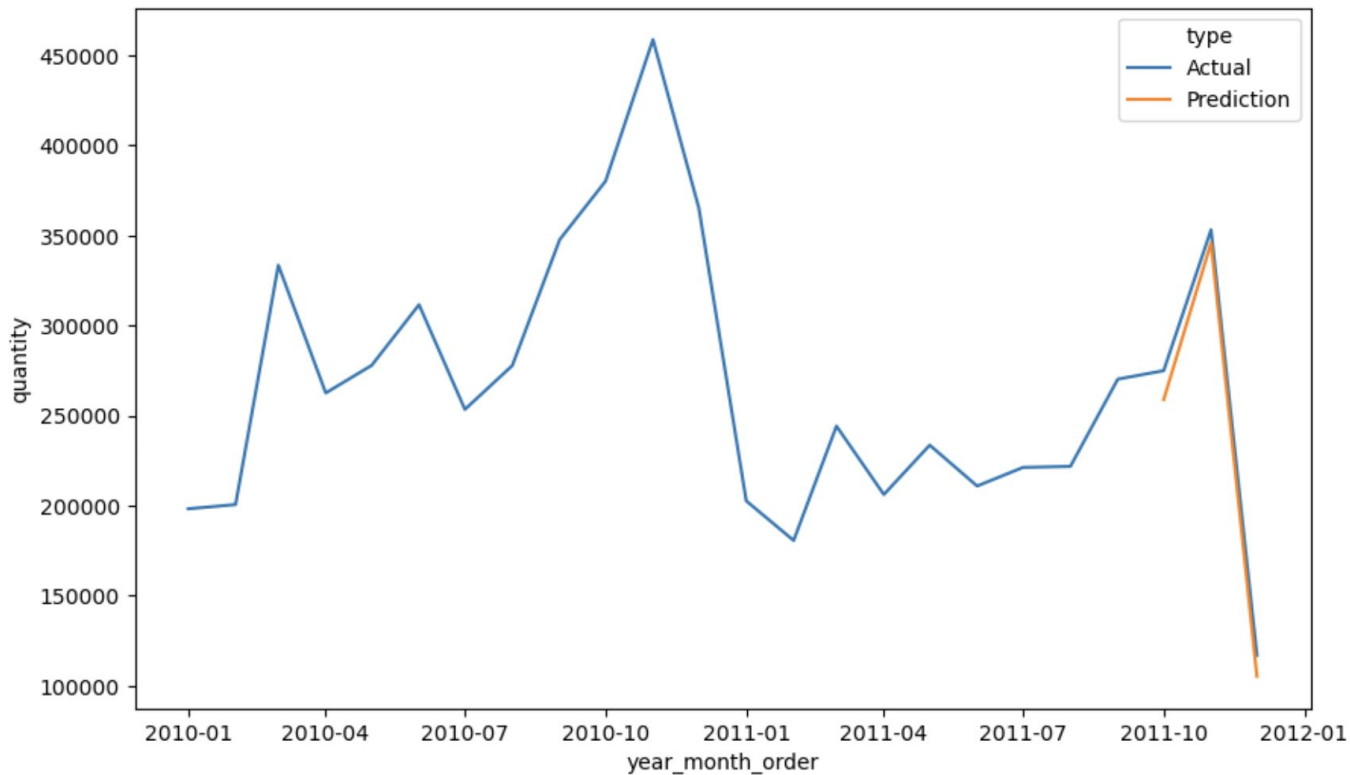
Only 1726 product-country combination than can be predicted (around 6.2% from all product-country)

	product_country	demand_class	quantity
23644	84077-United Kingdom	ERRATIC	98746
26318	85099B-United Kingdom	SMOOTH	85485
26434	85123A-United Kingdom	SMOOTH	82686
9272	22197-United Kingdom	ERRATIC	82263
21253	23843-United Kingdom	LUMPY	80995

Evaluation Model - Feature Importance



Evaluation Model - Prediction Result



Forecasting for the next 6 month

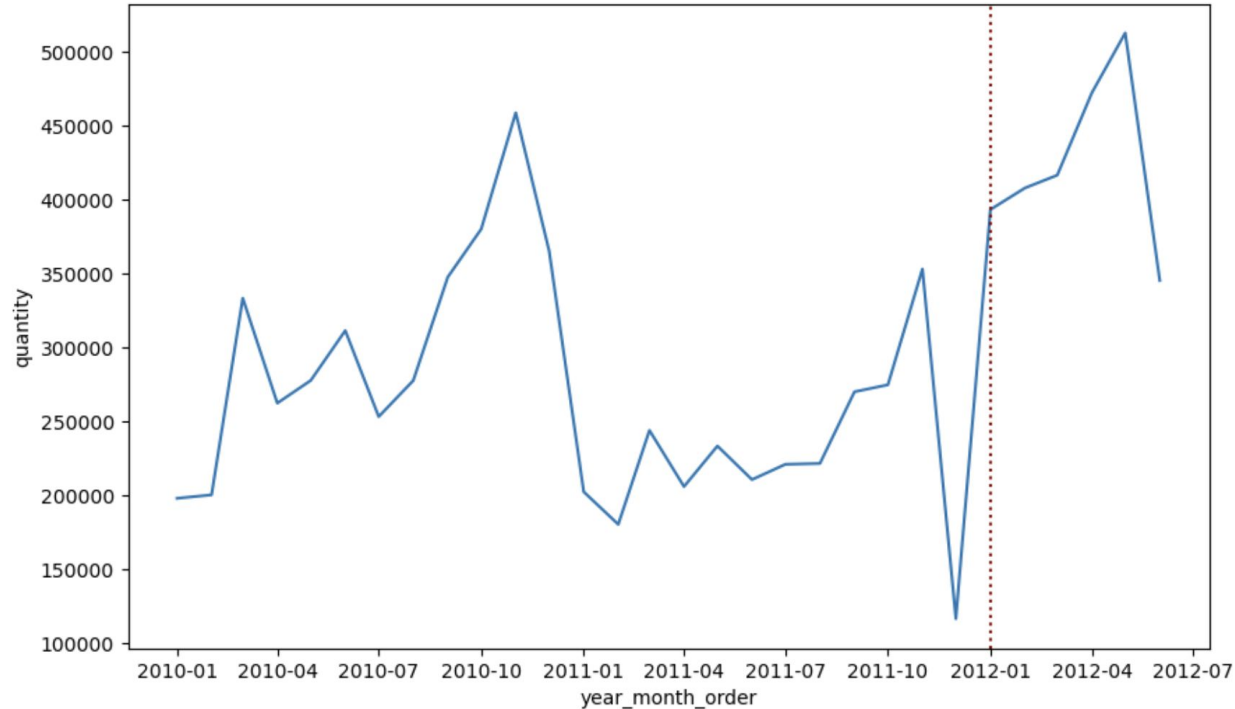
First RE-TRAIN MODEL with all data (merging the train & test dataset)

After that create dataset criteria for the next 6 month (January until June 2012)

1. for the next 6 month (2012-01-01 until 2012-06-01) (monthly)
2. only for product & country that in **SMOOTH & ERRATIC** demand pattern
3. Revenue = Target Revenue per month, product & country for the next 6 month
 - a. Assume that the target **at least same as last year in each month + growth last year**
 - b. i.e on Target Revenue 2012-01-01 for product A in Country A = (Revenue 2011-01-01 for product A in Country A) + (Growth Revenue for product A in Country A between 2011-01-01 & 2010-01-01)
 - c. if the growth < 0 then growth = 0
4. last 6 months of quantity, revenue, unit_price, n_customer & n_order in every month, product & country
5. last year of quantity, revenue, unit_price, n_customer & n_order in every month, product & country

Forecasting for the next 6 month

Monthly Demand projection for the next 6 month



the demand for the next 5 month would be increasing significantly around 400,000 - 500,00 qty, but on June, 2012 the demand will decreasing (around 350,000 qty)

Conclusion & recommendation

EDA

1. Revenue have same fluctuation with quantity on every month, when the revenue increase, the quantity also increase in certain month
2. Revenue & Quantity Always increase Significantly on Q4 in every year (i.e. growth revenue in Q4 2010 = 73% & growth quantity Q4 2010 = 35.2%), Possibly because of Holiday Season (New Year & Christmas) so People go shopping more often
3. Quantity have many Outliers, and after categorizing the quantity, combination of Product-Country that have qty < 50 is significantly higher than other (around 26,373 combination of product-country)
4. product_id DOT (DOTCOM POSTAGE) & 85123A (HANGING HEART T-LIGHT HOLDER) always purchased by customer in every year. Those products always contribute with high revenue (Top 5 Products). On 2009 the products gave almost 36,000 GBP and the number always increase every year up to 275,000 GBP in 2011, especially for DOTCOM POSTAGE Product with 20% Growth from 2010 to 2011
5. On 2009 the transaction only happen in December, so that the trend data is not showing anything, with total revenue roughly 825,000 GBP
6. UK always significantly have the highest revenue on every year. On 2009 UK revenue 750,317 GBP, Increased on 8,800,000 GBP on 2010. 2011 UK revenue is 8,300,000 GBP
7. DOT (DOTCOM POSTAGE), 85123A (HANGING HEART T-LIGHT HOLDER) & 22423 (REGENCY CAKESTAND 3 TIER) always on demand every month from 2010 until 2011 in UK that contribute with High revenue (Top 5)
8. There Are 685 Product & Country Combination (around 2.5% from all combination) that routinely purchased by customers every month from 2010 until 2011. In example Product POST in Germany, Product POST in Spain, etc., on the other hand there are 12750 Product & Country Combination that only in one month purchased by customer from 2010 until 2011, i.e. Product gift_0001_70 in UK, POST in Canada, etc.
9. From 27744 of Product & country combination (demand profile), there is Lumpy demand pattern that dominated the combination (72%) followed by Intermittent pattern with percentage around 22%. Erratic & Smooth pattern (ideal pattern) only have 5.8% and 0.4% from total combination

Conclusion & recommendation

Forecasting

Use Model XGBoost for only **Erratic & Smooth demand pattern** product in certain country (1726 product-country) because in this case the MAPE have lower number (28%) and still can predict top 4 product-country that give the most quantity of all time.

Recommendation

For the next 6 months (January - June 2012) would be increasing significantly around 400,000 - 500,000 qty, but on June, 2012 the demand will decreasing (around 350,000 qty) 2 of them are product_id **84077 (WORLD WAR 2 GLIDERS ASSTD DESIGNS) & 22197 (POPCORN HOLDER)** in UK that have **122,185 & 104,367** demand qty, so that the stocks of those products should be prepared.

Other than those 2 products, here is TOP 10 product in certain country that have the highest demand and should be preparing the stocks :

	product_id	country	quantity
1410	84077	United Kingdom	122185.937988
860	22197	United Kingdom	104367.280518
1593	85099B	United Kingdom	104171.015869
1604	85123A	United Kingdom	100670.480957
1493	84879	United Kingdom	83169.278076
41	17003	United Kingdom	81554.885284
308	21212	United Kingdom	79680.479797
698	21977	United Kingdom	54616.470276
6	15036	United Kingdom	53629.971008
842	22178	United Kingdom	52069.897339