WHERE TO PLAY, HOW TO WIN

EXPANDING E-COMMERCE SALES IN OVERSEAS MARKETS

@ 

THE UK E-GIFTING COMPANY

Prepared by :
Fred Gigou & Koustubh Jadhav Sept 20th, 2021.

Notebook link : Milestone 1 Project | Deepnote

Project motivation and data sources

Expanding e-commerce Sales in Overseas Markets

Motivation: E-commerce has grown exponentially in the last years and has continued to accelerate due to COVID, representing 13.7% of retail sales in Q1 2021. As such, capturing a share of the pie is critical for any for-profit organization. While E-commerce allows businesses to reach a broader base to maximize profit and sales, its technology and data-driven natures require modern data science techniques to harness its full potential - beyond standard business practices.

Project goal: Study current business performance and macro-economic landscape of the market to build a prediction model to increase current exports e-ecommerce sales based on country/market ranking (where to play) and product recommendations (how to win).

Questions to answer:

1- Where to play: what are our most promising international markets for our current business based on sales trends and macroeconomic indicators?

2- How to win: which products should we recommend to existing customers to maximize sales growth?

Assumptions: This study assumes that we are tackling this business question in the context of the year 2012. Lack of publicly available e-commerce data limits us to work with historical but extensive dataset (The dataset includes 25 months of transaction history). We are confident that our approach is just as much suitable if we were to apply it on recent data if it was available.
Our deliverable will provide a real, actionable business recommendation.

Primary Dataset: UK-based e-commerce retailer dataset from 2011 given its unique features: multi-country footprint, customer location availability, a significant transaction base (+541k) and a 25-month time series. The data set contains all the transactions occurring between 01/12/2009 and 01/12/2011. The company mainly sells gifts to many customers that are primarily wholesalers. The dataset contains information such as transaction ID, customer ID, country, transaction Date, Quantity and Price.

UK based E-commerce Retailer Transaction Dataset
Source: Center for Machine Learning and Intelligent Systems, University of California Irvine.
https://archive.ics.uci.edu/ml/datasets/Online+Retail+II
Size: 44MB
Records: 541909 rows
Access method: download

Secondary Dataset: Secondary dataset consists of multiple datasets containing World Bank reported macro-economic indicators – GDP, Purchasing Power Parity, Population, Merchandise Imports, Inflation CPI, Access to Internet, Life Expectancy and Health Expenditure. For each indicator, we are interested in extracting performance for the year 2011 for each country in our export market. Please refer to code to load datasets (Fred and K's Milestone 1 project | Deepnote) for more information. CERDI Sea Distance dataset is used to extract shipping distance between countries. All secondary datasets are merged using Country code as a key.

World Bank Datasets
Location: World Bank for macro indicators – https://data.worldbank.org/indicator/
Size: Each table ranges from 50 to 250KB
Format: Excel
Records: 250 to 270 rows
Access method: Download

CERDI Sea Distance Datasets
Location: CERDI distance, https://zenodo.org/record/46822
Size: 2MBs
Format: Excel
Records: 51302
Access method: download

Data manipulation

A thoughtful approach to answer our business questions

Business Performance EDA - Data manipulation approach

Step 1: We read Primary Dataset into a pandas dataframe and cleaned it up for non transactional entries. Transactional entries were identified based on their unique alphanumerical combination used in the StockCode column.

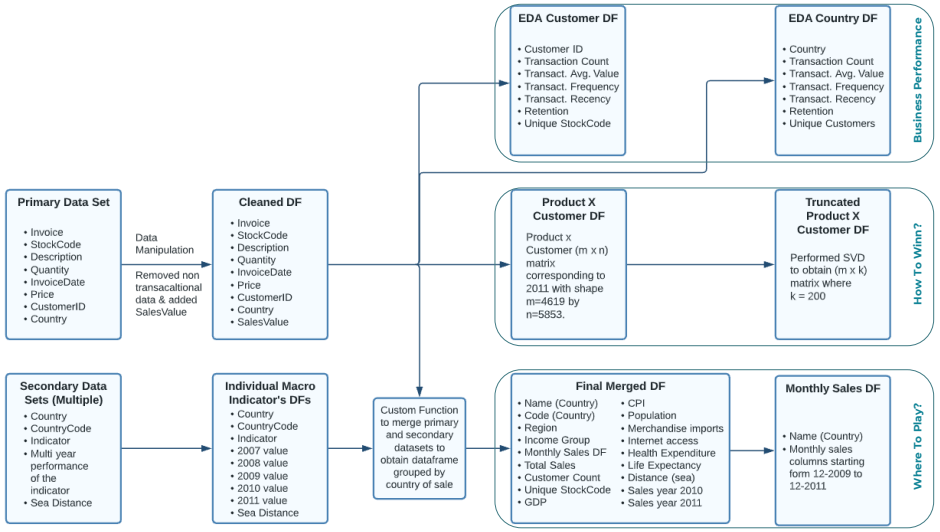
Step 2: There were other non sales related entries in the primary dataset, which were cleaned by filtering non-negative entries in the Quantity column.

Step 3: We added Sales Value column to obtain total transaction value of each transaction based on Quantity and Unit Price columns.

Step 4: The cleaned up dataframe for primary dataset was also used in the “Where to Play?” and “How to win?” analysis.

Step 5: We decided to perform exploratory data analysis by analyzing sales trends by grouping cleaned dataframe based on CustomerID and Country of sale.

Step 6: We performed data manipulation to find frequency, recency and nature of transactions such as - quantity, average order value, unique purchased StockCode and unique customers.



WHERE TO PLAY? - Data manipulation approach

Step 1: We imported secondary datasets into its respective pandas dataframe capturing macro-economic indicators performance for years 2007-2011.

Step 2: We employed a custom function written to **merge primary and secondary datasets** to obtain a pandas dataframe that contained transaction details grouped based on country of sale and its key macro-economic indicators, and its shipping distance from the primary market i.e. UK **using country code as a key**.

Step 3: The resultant dataframe was used to develop SARIMA model

Step 4: The resultant dataframe from Step 2 was further expanded to obtain monthly sales performance of each export market country into a separate dataframe.

Step 5: The monthly sales dataframe was used to perform OLS regression analysis of monthly sales data.

HOW TO WIN? - Data manipulation approach

Step 1: We included all UK customers for the comparison as they account for 91% of the total non-nan count (5832).

Step 2: We created a product x customer (m x n) matrix corresponding to 2011 with shape m=4619 by n=5832.

Step 3: We created an X truncated matrix m products x k customers by reducing to 200 groups of customers through the singular value decomposition technique with a shape m=4619 by k=200.

Step 4: We ordered the matrix in descending order to have top-selling items and groups of customers on top.

Step 5: For our top 5 countries, we compared all their customers versus the X truncated matrix and determined their most similar group of customers based on cosine similarity.

Step 6: For each country/customer pair, we returned the top missing items number vs their similar group of customers to a table. In total we return a 509- row x 5-product matrix corresponding to all non-nan active and unique country/customer pairs.

E-commerce business plan: Where to play – How to win

Executive Summary

Background

The UK e-gifting company wants to accelerate its business overseas to take advantage of the rise of e-commerce. The newly appointed CEO, Ric Mc Gregor, has extensive international business and data experience acquired at Amazon. He wants us to use both business acumen and the power of data science techniques to answer two questions to grow our business :

Where to play

What key markets should we focus on in 2012 out of the 38 where we operate?

How to win

which products should we recommend to existing customers to maximize sales growth?

Recommended Action

In 2012, focusing on 5 countries is likely to yield the most benefit: **France, Australia, Germany, The Netherlands and Switzerland** based on their extrapolated sales using an ordinary least square regression (OLS).

Where to play: sales OLS to calculate potential

We applied the parsimony principle and literature on forecasting (1), which supports that linear regression is a simple yet effective way to identify sales trends and extrapolate on a yearly basis, the project scope. Therefore, to get to our country shortlist, we projected the sales of 14 markets with a statistically robust history, at least 20 months, to reach the British Pound (GBP) potential for 2012. The other countries had too irregular sales history (2) to provide a solid calculation.

How to win: using a customer similarity technique

For each 259 customers of key markets, **we are recommending 5 products to launch** as well as a short list of the most recommended products overall.

Our approach consisted in returning the top 5 selling missing items versus their most similar group of customers based on their 2011 product sales history.

We used cosine similarity versus the total base of 5832 customers, which we reduced to a group of 200 “typical” customers based on singular value decomposition (SVD) for efficiency reason.

Rejected Option

Macro factors regression entails too much prediction error

We used an OLS model to regress sales versus a set of 8 macro indicators. The winning combination includes three variables: GDP, expenditure on health, and merchandise imports. Although the F test, with a probability of 0.000822, indicates that the variables are jointly significant (3) with alpha=0.05, a low adjusted r-squared of 0.346 warns of inaccurate predictions (4).

Beyond 2012: a complementary model

Once we tap the “low hanging fruit” that represent key markets, our macro factor model can help spot markets beyond 2012 due to their potential.

Conclusion - action Steps:

- Formalize key markets list in 2012 plan.
- Review recommended product list and propose to customers.
- Incorporate potential markets in strategic plan, beyond 2012.

(1) How to Choose the Right Forecasting Technique (hbr.org) By John C. Chambers, Satinder K. Mullick and Donald D. Smith, July 1971

(2) How Much Data is Needed to Train a (Good) Model? | DataRobot, By Ryan Sevey, August 4, 2017

(3) Correlation and Regression Analysis: A Historian’s Guide. University of Wisconsin, By Archdeacon, T. (1994).

(4) How To Interpret R-squared in Regression Analysis – Statistics By Jim Frost, 2021 https://statisticsbyjim.com/regression/interpret-r-squared-regression/

Data Cleaning, Manipulation and Exploratory Data Analysis:

Sales Trends, Customer and Country Specific Analysis of Export Business Performance

2011 Sales Total	2010 Sales Total	Export Customers	Export Markets
£ 1.43 mil.	£ 1.32 mil.	497*	36*

Top 10 Key Export Markets Performance

	FY2011 (£)	FY2010 (£)
Netherlands	275,135	259,772
Ireland	261,481	348,501
Germany	192,290	187,504
France	176,052	135,022
Australia	137,138	31,075
Spain	54,021	36,837
Switzerland	51,862	41,937
Belgium	35,327	21,707
Sweden	33,005	53,062
Japan	29,711	17,427

Retained, New and Lost - 2011

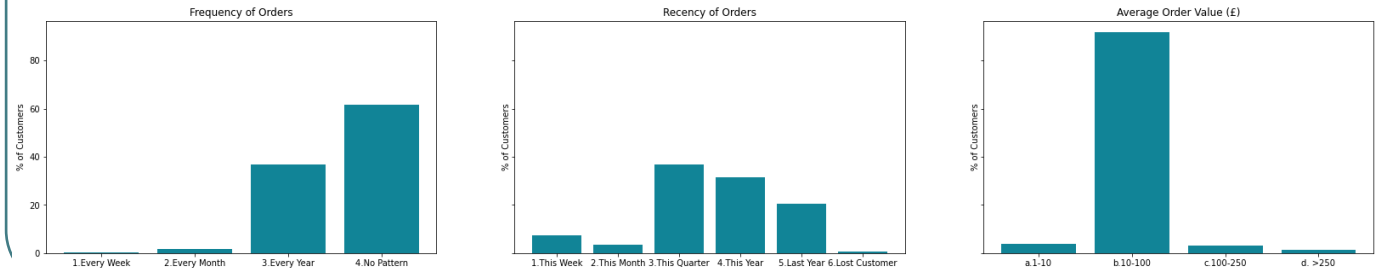
Customer Characteristics Summary

Retained	44 %
New	35 %
Lost	21 %

2011 Export Market Characteristics Summary*

Retained	29
New	3
Lost	4

Export Customers' Order Patterns – Frequency, Recency and Avg. Order Value



Handling of data anomalies:

We analyzed the sales data for 2010 and 2011, and it is presented here in a dashboard format. We cleaned up the following markets due to –

1. Mislabeled data – European Community
2. Non-official entities – Channel Island, West Indies, unspecified.

Findings:

The data shows that to maintain growth in sales, we must identify and prioritize our key markets. We have **retained sales in 29 countries**, while **expanding to 3 new countries** in the year 2011. Although in 4 countries we were not able to record any new sales in the year 2011.

For the year 2011, we have retained sales from **44% of our customers, and 35% of customers are new customers** placing order with us for the first time in 2011. While **21% of our customers did not place any new order** in the year 2011.

Conclusion:

Considering difference between numbers of new and lost customers, and number of countries where we did not record any new sales in the year 2011 – we must actively look at strategies to improve vertical expansion such as more average order value, customer retention and frequency of transactions.

1. **Focus on Key Markets. (Where to play?)**
2. **Improve stock portfolio recommendations to customers (How to Win?)**

*Non-Nan valid cumulative number of countries and customers (2009-2011). 38 is total number of countries in data set but Bermuda and Hong Kong are not active.

Where to play - Exploratory Data Analysis:

Opportunity is to max out existing markets before further expansion

Data preparation

We standardized macro economic and sales data to prepare it for comparison.

Visualization:

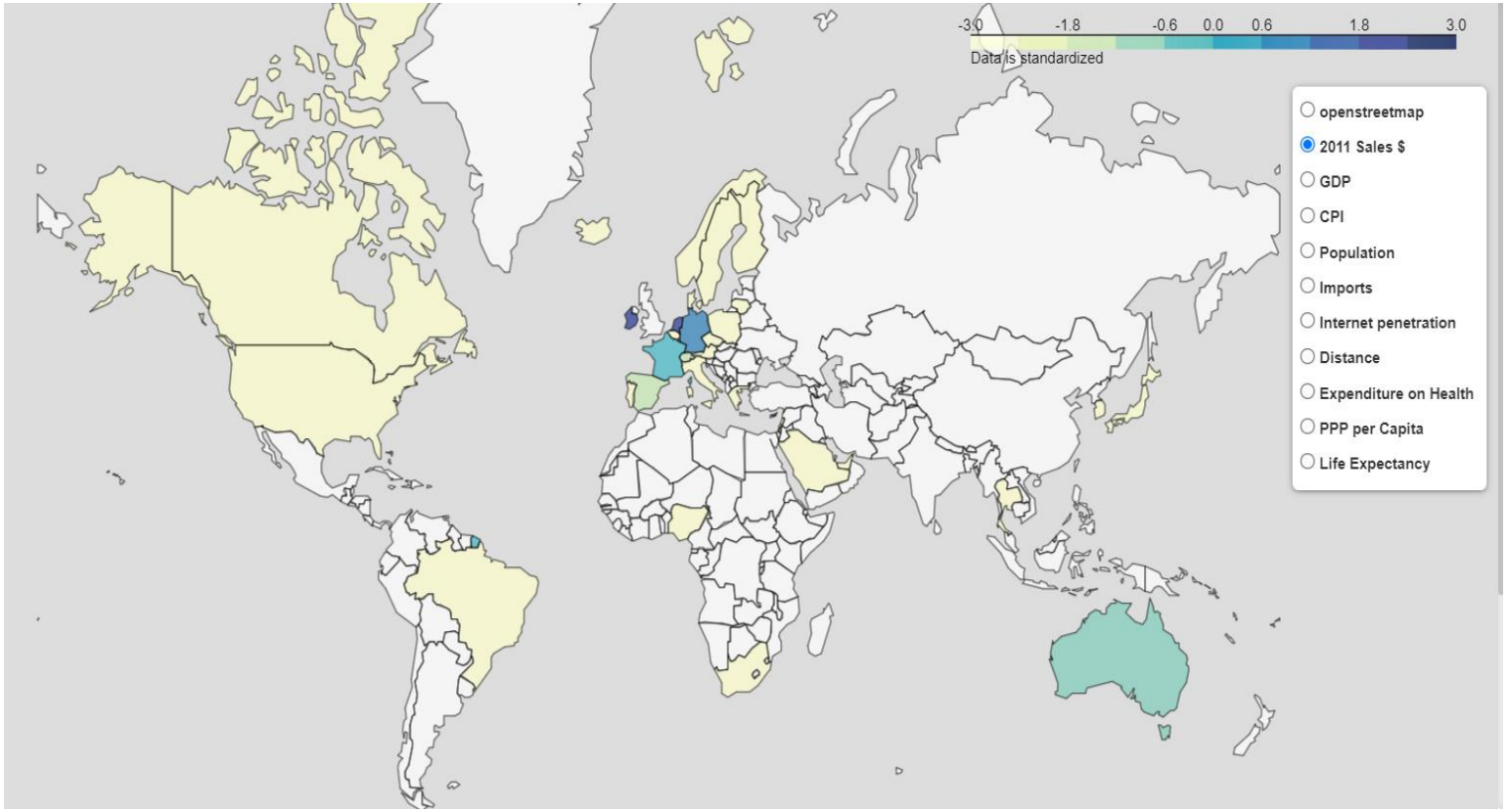
To compare standardized macro-economic performance of our sales markets, we have created this interactive visualization in folium which can be found in Section 2-iii of our deepnote file.

Findings:

We hope this visualization will provide you with geographical perspective of our operational presence, along with each market's macro economic performance for the year 2011.

As we can see, we have a strong European presence. In addition, we have a growing geographic footprint with only a few scaled markets with yet untapped macro potential. Our focus in 2012 should be to unleash our existing markets' potential.

Figure 1 – interactive folium map.



(13) Folium Map. Milestone 1 Project | Deepnote – *Section 2-iii*

Where to play - model 1: Sales ~ Macro Factors regression

Reaching statistical significance although with high prediction variance

The business problem, analytical answer

Our first business question was: what 5 countries offer the most sales potential out of our current exports markets? We addressed the problem by formulating the analytical question: Is there a relationship between our value sales and key macro indicators? And can we predict sales potential based on macro data?

Approach

We fitted an ordinary linear regression model (OLS) between 2011 value sales (dependent variable) and countries' key indicators. We optimized our model by running all combinations of these indicators and picking the one with the highest adjusted R squared.

This metric shows how well data points fit a curve by adding more variables (6). We used eight key indicators in our modeling optimization but could have potentially tested tens of more variables with the same code (7).

Results & interpretation

As shown in **figure 2**, The winning model includes three variables: GDP, expenditure on health and imports. The F test, which tells if a group of variables are jointly significant (3), indicates a probability of 0.000822, which means, with alpha=0.05, that there's a statistical significance.

However, although literature shows diverse views, a low adjusted r-squared of 0.346 entails more errors and warns of inaccurate predictions (4).

Conclusion

We conclude that our model can explain part of the sales potential for our key markets but is unlikely to make accurate sales predictions. Therefore, we may use it to identify markets with potential but not to predict 2012 sales.

Lastly, considering the coefficient values, the model would suggest that our company tends to do better in countries that are open to imports, developed and with an older population (high expenditure on health).

Figure 2 – OLS regression results

	coef	std err	t	P> |t|
Intercept	-5.842E+04	3.27E+04	-1.786	0.084
GDP	-3.212E-08	8.62E-09	-3.726	0.001
Merchsales (imports)	2.007E-07	5.97E-08	3.361	0.002
expenditure on health	9794.4545	4262.092	2.298	0.028

Adj. R-squared	0.346
F-statistic	7.162
Prob (F-statistic)	0.000822

(3) Correlation and Regression Analysis: A Historian's Guide. University of Wisconsin, By Archdeacon, T. (1994).

(4) How To Interpret R-squared in Regression Analysis – Statistics By Jim Frost https://statisticsbyjim.com/regression/interpret-r-squared-regression/

(6) Dodge, Y. (2008) Springer The Concise Encyclopedia of Statistics. https://www.statisticshowto.com/probability-and-statistics/statistics-definitions/adjusted-r2/

(7) OLS macro model code reference. Milestone 1 Project | Deepnote – *Section 4-i*

Where to play - model 2: Sales Regression

We recommend OLS out of 2 models assessed

Approach

As a second approach, we used past sales to forecast 2012 sales as we have 25 months of history. Although there's no golden rule, at least 12 months of continuous sales are recommended to identify trends (2). However, out of 38 countries analyzed, 24 have incomplete sales data. Therefore, we only applied the model to the 14 entities with at least 20 months of sales.

Recommended model: OLS

Figure 3 shows sales regression for countries with sales history. Analysis and observation show that the model offers a wide CI and a RMSE of 33,217. Nonetheless, it effectively captures the sales trend on a full year basis, which is our primary objective for estimating potential. Consequently, we chose this model and applied the same technique for all countries with good sales history. Please refer to code in deep note (8).

Alternative model

As shown in **figure 4**, we also fitted a SARIMA model considering the seasonal nature of our business (code reference 9). However, while visual observation shows a better monthly fit and a lower RMSE of 27,707, it lacks sufficient sales history. In fact, the statistical reference indicates that 50 to 100 observations are required for SARIMA to be valid (5).

Conclusion

Literature on forecast suggests (1) that linear regressions offer both simplicity and predictive power to identify trends, which supports our decision to use this model to estimate sales potential yearly.

Figure 3: Time Series Sales OLS regression

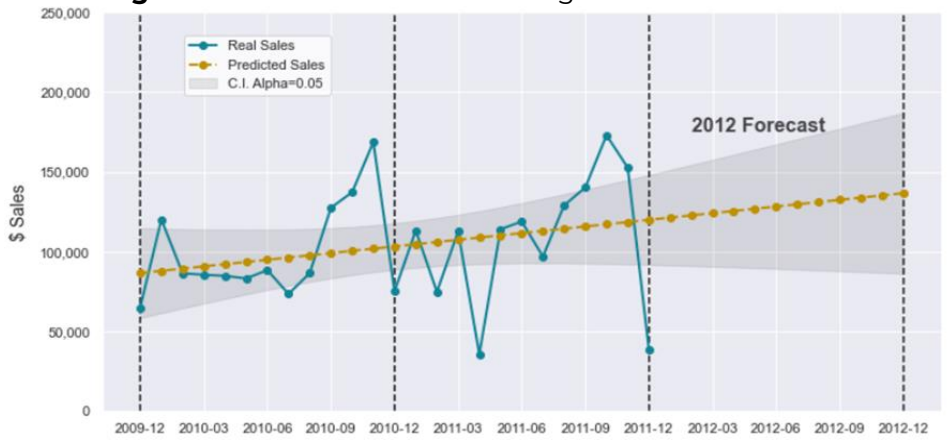
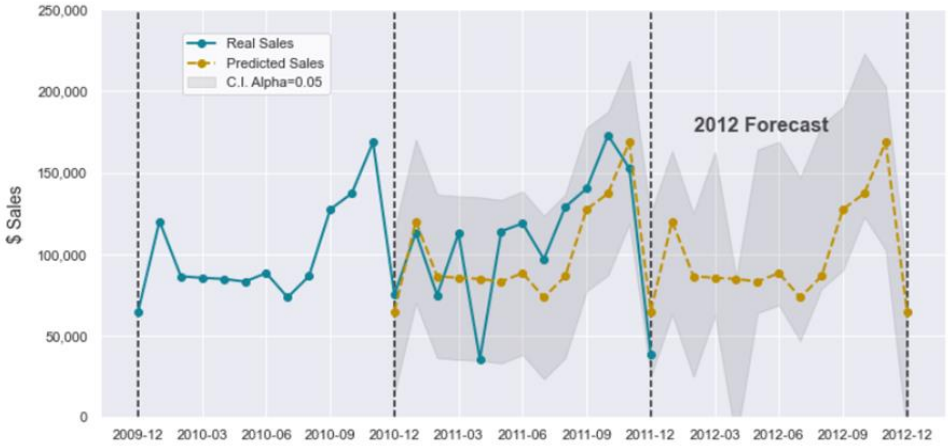


Figure 4: SARIMA time series model



(1) How to Choose the Right Forecasting Technique (hbr.org) By John C. Chambers, Satinder K. Mullick and Donald D. Smith, July 1971.

(2) How Much Data is Needed to Train a (Good) Model? | DataRobot, By Ryan Sevey, August 4, 2017

(5) Box and Tiao. 1975; Abraham 1980

(8) OLS sales regression code reference Milestone 1 Project | Deepnote – *Section 5-iv-c*

(9) SARIMA code reference Milestone 1 Project | Deepnote – *Section 5-iv-a*

Where to Play - results visualization

Selecting top 5 key markets based on sales regression potential

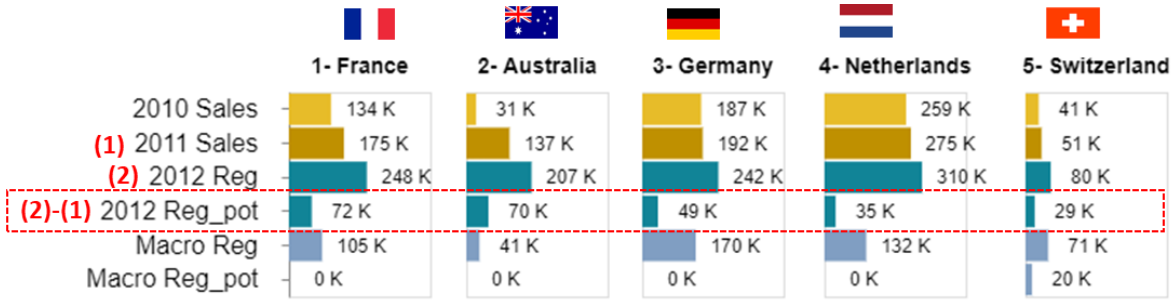
Approach & results

As mentioned in the last slide, we used the linear regression model on countries with robust time series to project sales potential. We then ranked the top 5 according to the absolute sales potential measured as 2012 projected sales minus 2011 actual sales – as shown in **figure 5**. Macro potential ranking ("Macro_pot) is also available in our notebook (10), being Canada, Belgium, Austria, Korea and Italy in the top 5.

Conclusion

Based on our approach, we can answer our where to play question. In 2012, we recommend to focus on **France, Australia, Germany, the Netherlands, and Switzerland** that are already established and posting positive growth. In our next step, "how to win", we will recommend products to launch to grow these markets.

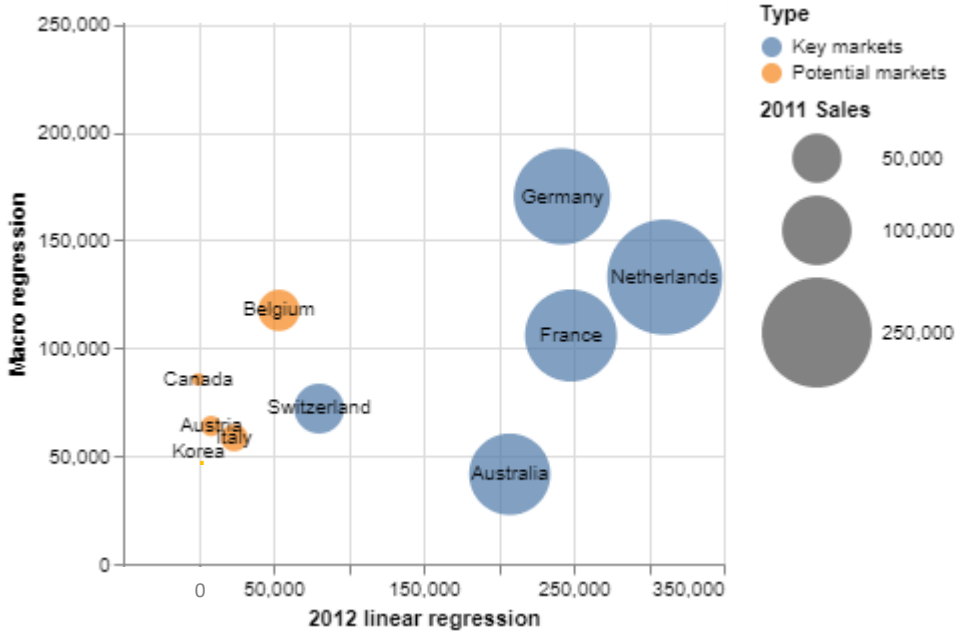
Figure 5 – Top 5 key markets based on sales regression. British Pounds.



Beyond 2012

Lastly, it is worth noting that the macro approach is complementary as it pinpoints underdeveloped markets with likely potential (potential markets). In general, business sense commands to focus first on large, growing markets to get a faster return on investment, which would be our first category (Key markets) while considering potential markets for the long run. Please refer to **Figure 6** to appreciate both categories of markets.

Figure 6 – Key markets vs top potential markets



(10) Potential markets: Altair visualization code reference.
Milestone 1 Project | Deepnote – *Section 6-iv*

How to win – analysis & visualizations

Using customers similarity to determine top-selling missing items

The business problem, analytical answer

Our second business question is related to how to win: What products should we launch in each customers key markets?
We addressed the problem by formulating the analytical question the following way: To what other customers are these customers most similar to? And, what top-selling products are they currently missing vs. their respective similar customers?

Sample results & interpretation

For key countries, Australia, France, Germany, The Netherlands, and Switzerland, our similarity model finally returns each customer's top 5 product recommendations (a total of 259 rows). In **figure 7**, we display a sample of the output corresponding to some countries' first country/customer ID pair.

Figure 7: Sample recommendation matrix

	Customer ID	Product 1	Product 2	Product 3	Product 4	Product 5
Australia	12386	35001W hand open shape	22656 vintage blue kitchen	85099C jumbo bag baroque	22113 grey heart hot	20747 piccadilly tea set
France	12413	85123A cream hanging heart	84879 assorted colour bird	22197 popcorn holder	22629 spaceboy lunch box	22630 dolly girl lunch
Germany	12426	22423 regency cakestand	85123A cream hanging heart	85099B jumbo bag red	84879 assorted colour bird	22197 popcorn holder

(11) Code reference for recommendation matrix.
Milestone 1 Project | Deepnote – *Section 7-iv*

A visualization of most recommended products

Lastly, to go along our recommendation matrix that includes 259 pairs of countries, we are offering a visualization of most recommended products for key markets.

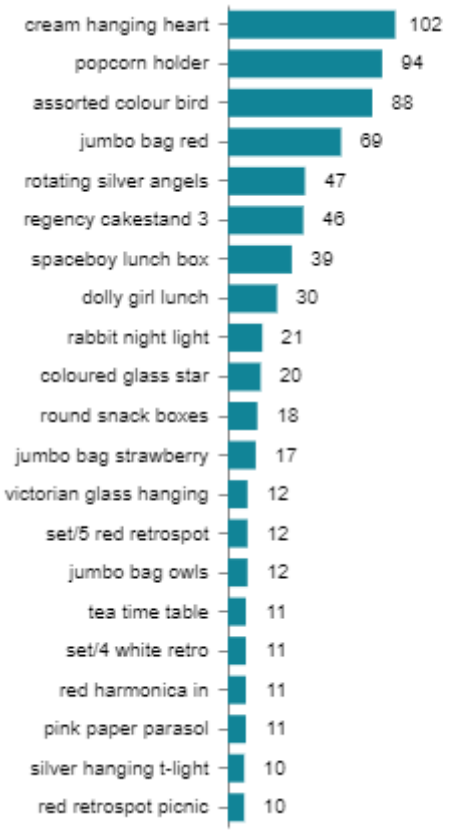
For this purpose, we used matplotlib word cloud in **figure 8** and a bar chart in **figure 9** by extracting the first three words of each product description, converted them to lower case and counted them. Please see code in deepnote as a reference (10). These 24 products account for 54% of all recommendations.

Now we are ready to win!

Figure 8: word cloud of most recommended item

(12) Code refence for Altair visualization of recommendations.
Milestone 1 Project | Deepnote – *Section 7-v*

Figure 9: count of most recommended items



Conclusion

We are ready to unleash growth in key exports market

Where to play in 2012

Sales regression indicates a strong pool of 5 markets to focus on: France, Australia, Germany, The Netherlands and Switzerland. These countries are already scaled, and the OLS regression is a simple technique to capture their sales trend. Therefore, we recommend including these markets as a priority for our 2012 business plan.

How to win

Our similarity algorithm recommends 5 products for each of 259 key markets customers. We can recommend these products automatically online to our customers, or we could propose a shortlist of 24 items that represent 54% of all recommended products (710 recommendations out of a total of 1295). On-line market research along to customer validation could also complement this recommendation approach.

Beyond 2012

Once we maximize our key markets, our macro model suggests considering five countries as part of our strategic plan: Canada, Belgium, Austria, Korea and Italy . These countries are posting very limited sales today, but their macro data suggest potential.
This project is not static; as we gain more sales history and experience internationally, we will be able to improve and update our models to reflect market and dynamics for additional business recommendations.

Statement of work

Writing and reporting: Fred and Koustubh

Cleaning: Koustubh

Manipulation: Fred and Koustubh

Dashboard: Koustubh

Analysis and visualizations: Fred

References

(1) How to Choose the Right Forecasting Technique (hbr.org) By John C. Chambers, Satinder K. Mullick and Donald D. Smith, July 1971.

(2) How Much Data is Needed to Train a (Good) Model? | DataRobot, By Ryan Sevey, August 4, 2017

(3) Correlation and Regression Analysis: A Historian's Guide. University of Wisconsin, By Archdeacon, T. (1994).

(4) How To Interpret R-squared in Regression Analysis – Statistics By Jim Frost https://statisticsbyjim.com/regression/interpret-r-squared-regression/

(5) Box and Tiao. 1975; Abraham 1980

(6) Dodge, Y. (2008) Springer The Concise Encyclopedia of Statistics. https://www.statisticshowto.com/

Code

(7) OLS macro model code reference. Milestone 1 Project | Deepnote – *Section 4-i*

(8) OLS sales regression code reference Milestone 1 Project | Deepnote – *Section 5-iv-c*

(9) SARIMA code reference Milestone 1 Project | Deepnote – *Section 5-iv-a*

(10) Potential markets: Altair visualization code reference. Milestone 1 Project | Deepnote – *Section 6-iv*

(11) Code reference for recommendation matrix. Milestone 1 Project | Deepnote – *Section 7-iv*

(12) Code refence for Altair visualization of recommendations. Milestone 1 Project | Deepnote – *Section 7-v*

(13) Folium Map. Milestone 1 Project | Deepnote – *Section 2-iii*