University of Nottingham Ningbo China

Faculty of Business

Academic Year 2019/20 Autumn Semester

Data at Scale: Management, Processing and Visualization (BUSI4389 UNNC) (AUC1 19-20)

Dr Zhao Cai

Student：

Mingyuan Zhu

Data at scale: Coursework

Introduction:

In this part, we preprocessed the data and cleaned the data by drop the null values.

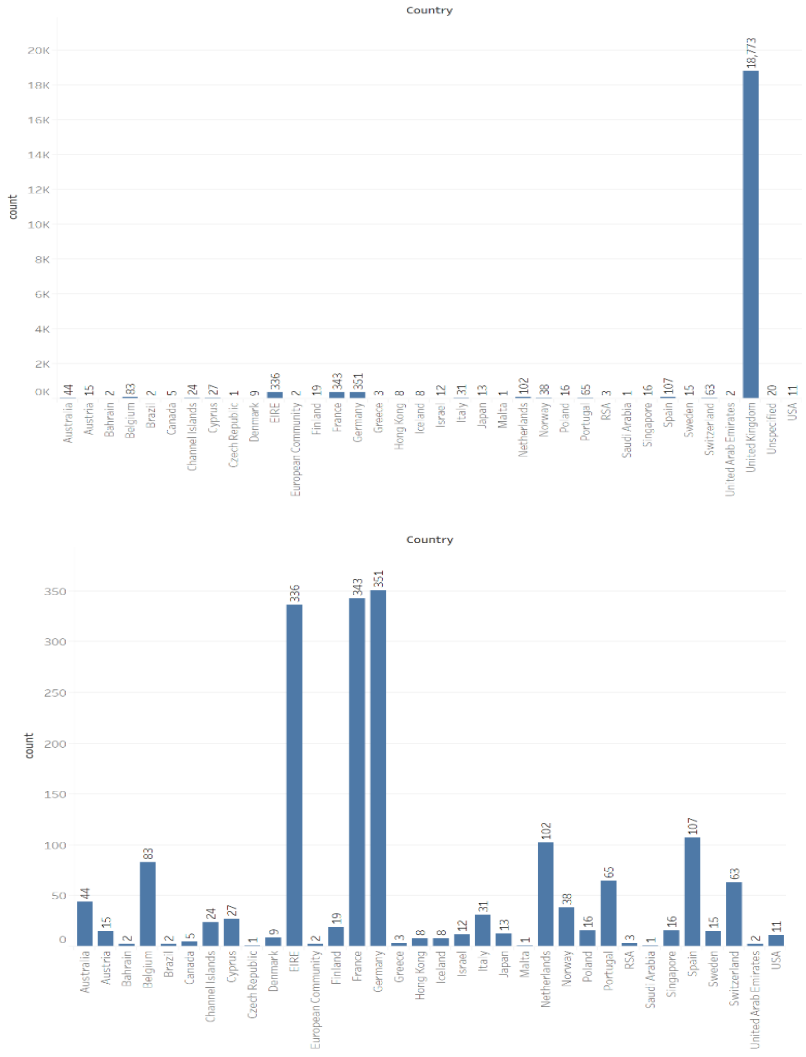

Figure 1. Bar chart of number of occurrences by country

The TOP 5 countries that place the highest number of records are: United Kingdom, Germany, France, EIRE, Spain. Then, we cleaned missing data, added months, and named non-target countries as ‘other’.

Data import

Steps:

1、

CREATE TABLE onlineretail (invoiceno TEXT, stockcode TEXT, description TEXT, quantity INT, invoicedate timestamp, unitprice NUMERIC, customerid TEXT, Country TEXT);

2、

\COPY onlineretail FROM '/home/lab test 1/coursework/20219292.csv' csv;

Data clean:

Null values: ----------

Total: 20571 records

CustomerID: 5108 NULLS

Descriptions : 58 NULLS

Country: 20 Unspecified country

Invoiceno: 361 NULLS

STEPS: ----------

```
DELETE
FROM onlineretail
WHERE customerid IS NULL
        OR description IS NULL
        OR country = 'Unspecified'
        OR invoiceno IS NULL;

DELETE 12
---------------------------------------
SELECT COUNT (*) FROM onlineretail;
 count
-------
 15451
(1 row)
3----------------------------------------
SELECT COUNT (*)
FROM
    (SELECT COUNT (DISTINCT country) AS country_number
    FROM onlineretail
    GROUP BY customerid)a
```

```
WHERE country_number >1;
-----------------------------
 count
-------
     0
(1 row)

They buy the same products in the same country, not exist buying in different country
4、----------------------------
CREATE TABLE onlineretail_zmy ASSELECT invoiceno,
            stockcode,
            description,
            quantity,
            invoicedate,
            unitprice,
            customerid,
            country,
            quantity*unitprice AS income,
            DATE_TRUNC ('month', invoicedate) AS month, DATE_TRUNC('day',
invoicedate) AS day
FROM onlineretail;
----------------------------------------
CREATE TABLE onlineretail_zmy ASSELECT invoiceno,
            stockcode,
            description,
            quantity,
            invoicedate,
            unitprice,
            customerid,
            country,
            quantity*unitprice AS income,
            DATE_TRUNC('month', invoicedate) AS month, DATE_TRUNC('day',
invoicedate) AS day
FROM onlineretail ;

SELECT 15451
-------------------------
SELECT COUNT (*) FROM onlineretail_zmy ;
 count
-------
```

```
  15451
(1 row)
5、
SELECT * FROM onlineretail_zmy;
6------------------
UPDATE onlineretail_zmy SET country= 'others'
WHERE country NOT IN ('EIRE', 'France', 'Germany', 'United Kingdom');

UPDATE 729
------------------------
SELECT COUNT (*),country    FROM onlineretail_zmy GROUP BY country;

  count |        country
-------+----------------
      729 | others
      351 | Germany
      342 | France
      310 | EIRE
  13719 | United Kingdom
(5 rows)
```

Section 1: Cohort Analysis

This section focused on the cohort analysis for top 4 countries' retention rate in Germany, France, EIRE and United Kingdom. The retention rate measures customers activity irrespective of transactions records by counting its number for each customer.

For retails sometimes is non-contractual and one-off business, using the KPI of current customer number and income will lose some information.
We interested in customer numbers during the specified recency. Here we choose the month period that works for business and customers buying period. For the online or offline selling companies, like Taobao and Amazon, we want to see our customer's retention rate month on month by counting consumption times and consumption cycle.

Retention rate we used is of vital importance, and it seems profitable selling to existing customer rather new customers. The retention rate is a deadly KPI for us to indicate which will influence future growth and focusing solely on the income (mostly contributed by new customers and new advertisements investment) will lose the perimeter of long-lasting income. Here are details steps in the follow's analysis:

Steps:

1. **Assign customers to cohorts:**

```
CREATE TABLE cohort_assignment_zmy AS SELECT customerid,
            MIN (month)::DATE AS cohort_date,
            country
FROM onlineretail_zmy
GROUP BY    1,3;
```

SELECT 3075

---

2. **Count the number of customer's active each in each subsequent period.**

```
CREATE TABLE cohort_mth_cts_zmy ASSELECT cohort_date,
            EXTRACT(month
FROM AGE (invoicedate, cohort_date)) + 12*EXTRACT( year
FROM AGE (invoicedate, cohort_date) ) AS relative_period, COUNT(DISTINCT
customerid) AS active_ct ,country
```

```
FROM onlineretail_zmy
JOIN cohort_assignment_zmy USING (customerid, country)
GROUP BY cohort_date, relative_period, country;
```

SELECT 328

```
----------------------------------
CREATE TABLE cohort_mth_cts_zmy ASSELECT cohort_date,
        EXTRACT(month
FROM AGE (invoicedate, cohort_date)) + 12*EXTRACT( year
FROM AGE (invoicedate, cohort_date) ) AS relative_period, COUNT(DISTINCT
customerid) AS active_ct ,cohort_assignment_zmy. Country
FROM onlineretail_zmy
JOIN cohort_assignment_zmy USING (customerid)
GROUP BY cohort_date, relative_period, cohort_assignment_zmy. Country;
```

SELECT 328

---

3. Convert the counts to percentage

```
CREATE TABLE cohort_totals_zmy ASSELECT cohort_date,
        COUNT (DISTINCT customerid) AS cohort total,
        country
FROM cohort_assignment_zmy
GROUP BY    1,3;
```

SELECT 54

```
----------------------------------------
CREATE TABLE cohort_mth_percent_zmy ASSELECT cohort_date,
        relative_period,
        active_ct::NUMERIC / cohort_total AS active_percent,
        country
FROM cohort_mth_cts_zmy
JOIN cohort_totals_zmy
USING (cohort_date, country);
```

SELECT 328

---

4. Count the number of users per cohort. Add column to table.

```
CREATE TABLE cohort_analysis_zmy ASSELECT cohort_date AS row_id,
```

```
        relative_period::TEXT AS col_id,
        active_percent AS val,
        country
FROM cohort_mth_percent_zmy
UNION
ALLSELECT cohort_date AS row_id,
        'total'::TEXT AS col_id, cohort_total AS val, country
FROM cohort_totals_zmy;
```

SELECT 382

Section 2: The KPIs

This part we calculate the KPIs and visualize them in the tableau.

KPI Description	｜Average transaction value(ATV) and the number of transaction
KPI formula	Total value of all transactions, divided by the number of transactions, per month, per country
Steps to realize KPI:	CREATE TABLE avt_zmy ASSELECT month, SUM (income)/count(DISTINCT invoiceno) AS atv, count (DISTINCT invoiceno) AS count, country FROM onlineretail_zmy GROUP BY month, country; SELECT 65 Tableau: Visualized via tableau as graph titled 'ATV'. See the Tableau file
Additional Notes:	The average transaction value is calculated by dividing the total value of all transactions by the number of transactions or sales. Here we bead on the Table onlineretail_zmy.

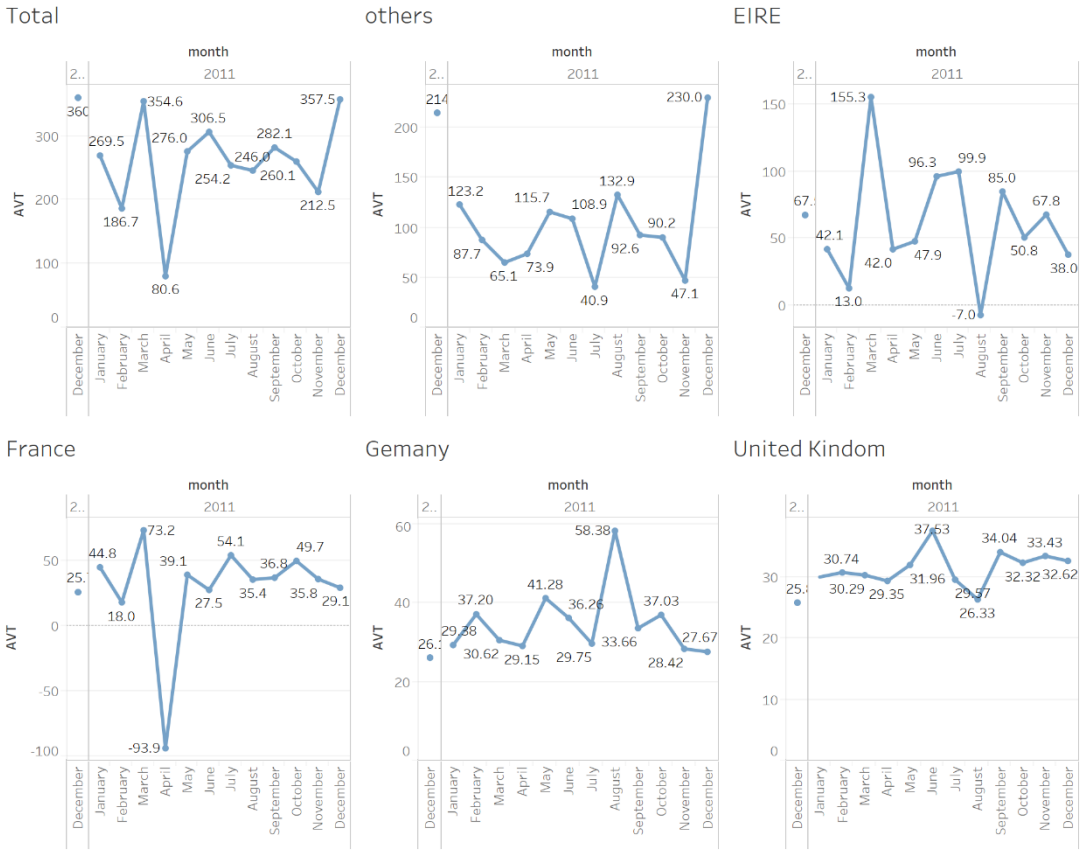

Figure 2: ATV for total and top 4 countries

KPI Description	|| Monthly Sales Income
KPI formula	Totally income, per month
Steps to realize KPI:	SQL: CREATE TABLE sales_per_month_zmy AS SELECT SUM (income) AS Sales, month, country FROM onlineretail_zmy GROUP BY 2,3; Tableau: Visualized via tableau as graph titled 'sales' and use the prediction
Additional Notes:	With prediction in tableau

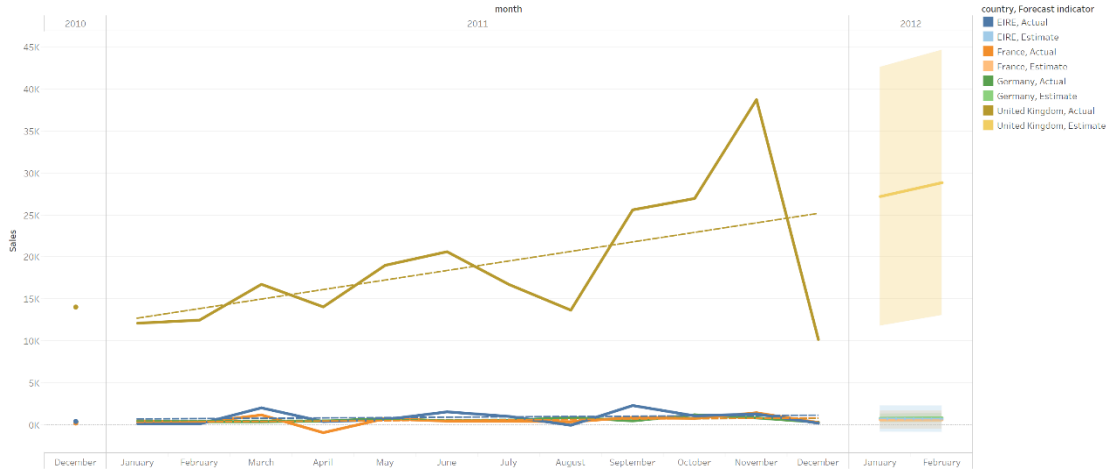

Figure 3: sales per month for each country and prediction

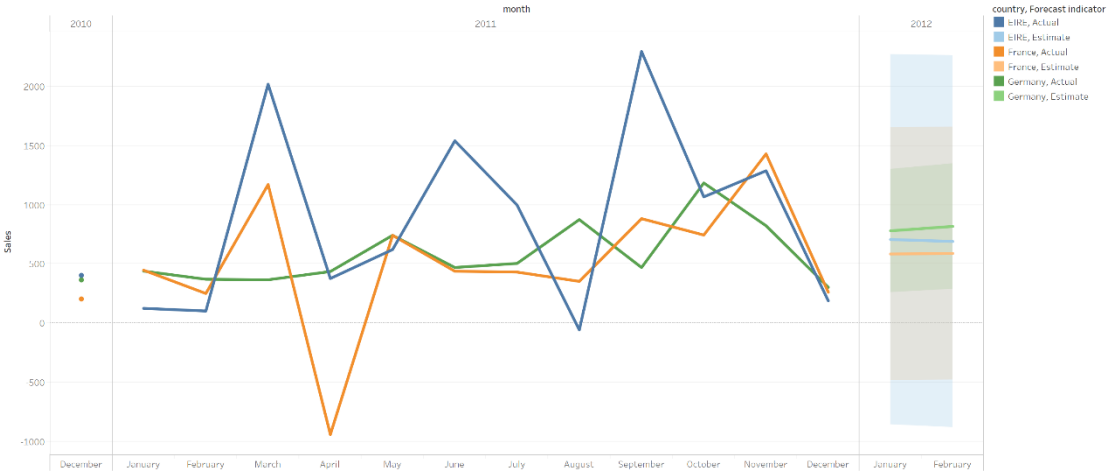

Figure 4: sales per month for France, Germany, EIRE and its prediction

KPI Description	||| The most popular items
KPI formula	Items sales rank top10
Steps to realize KPI:	COPY (SELECT stockcode, country, description, total_income FROM onlineretail_2 JOIN (SELECT stockcode, SUM (income) FROM onlineretail_zmy GROUP BY 1 ORDER BY 2 DESC LIMIT 10) a USING(stockcode) GROUP BY 1,2,3) to '/home/lab test 1/most_income_item.txt' ; COPY 49 Visualized via tableau as graph titled 'most_income_item'. See the Tableau file
Additional Notes:	

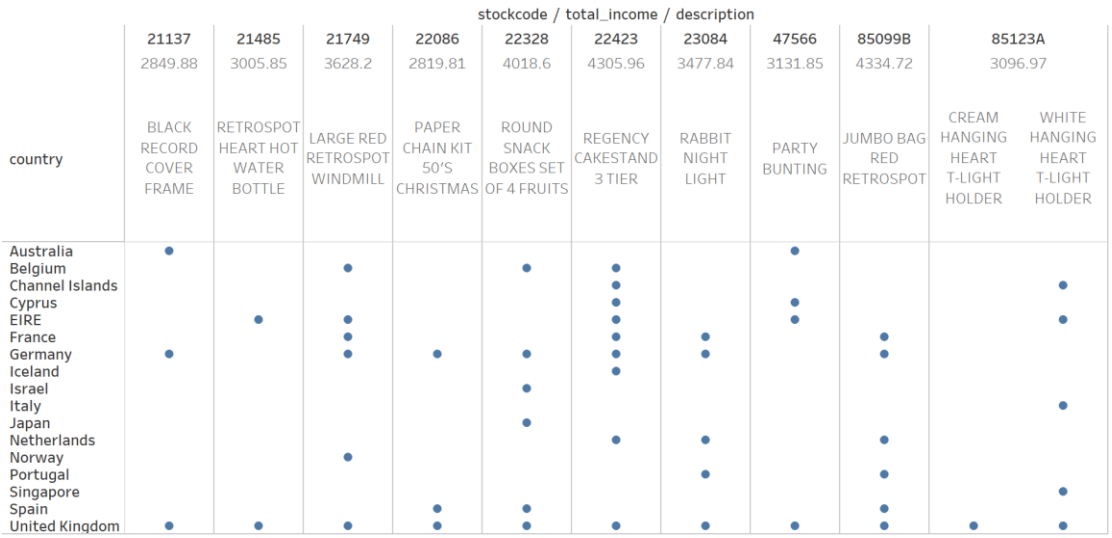

Figure 5: items with most incomes and its country

KPI Description	IV Most population items
KPI formula	Each country, each item, Count its invoice
Steps to realize KPI:	SELECT stockcode, country, COUNT (*) AS count FROM onlineretail_zmy GROUP BY 1,2 HAVING country= 'United Kingdom' ORDER BY 3 DESC LIMIT 10; ; SELECT stockcode, country, COUNT (*) AS count FROM onlineretail_zmy GROUP BY 1,2 HAVING country= 'EIRE' ORDER BY 3 DESC LIMIT 10; SELECT stockcode, country, COUNT (*) AS count FROM onlineretail_zmy GROUP BY 1,2 HAVING country= 'Germany' ORDER BY 3 DESC LIMIT 10; SELECT stockcode, country, COUNT(*) AS count FROM onlineretail_zmy GROUP BY 1,2 HAVING country= 'France' ORDER BY 3 DESC LIMIT 10; SELECT stockcode, country, COUNT(*) AS count FROM onlineretail_zmy GROUP BY 1,2 HAVING country= 'others' ORDER BY 3 DESC LIMIT 10;

Additional Notes:	∨ Totally popular items

Results:

UK

stockcode	country	count
21749	United Kingdom	98
72802C	United Kingdom	88
85123A	United Kingdom	83
85099B	United Kingdom	65
47566	United Kingdom	52
22197	United Kingdom	49
20725	United Kingdom	48
22086	United Kingdom	48
22383	United Kingdom	48
84879	United Kingdom	47

(10 rows)

EIRE

stockcode	country	count
C2	EIRE	6
22699	EIRE	5
72802C	EIRE	4
20728	EIRE	3
23245	EIRE	3
22978	EIRE	3
22727	EIRE	3
48187	EIRE	3
22961	EIRE	3
22197	EIRE	3

(10 rows)

Germany

stockcode	country	count
POST	Germany	10
22554	Germany	6
21578	Germany	4
22326	Germany	4
22551	Germany	4
22908	Germany	4
23212	Germany	3
22849	Germany	3
22728	Germany	3
20675	Germany	3

(10 rows)

France

stockcode	country	count
POST	France	15
22181	France	9
22554	France	6
21749	France	5
22630	France	4
23084	France	4
21121	France	4
22556	France	4
22551	France	4
21212	France	4

(10 rows)

Some Product Description:

VANILLA SCENT CANDLE JEWELLED BOX

WHITE HANGING HEART T-LIGHT HOLDER

JUMBO BAG RED RETROSPOT

PARTY BUNTING

SMALL POPCORN HOLDER

KPI Description	Ⅵ Total Revenue
KPI formula	Sum the income, per country, rank
Steps to realize KPI:	SELECT SUM(quantity*unitprice) AS total_income, country FROM onlineretail_2 GROUP BY 2 ORDER BY 1 DESC LIMIT 10; Top 10 countires with most revenue ---------- total_income | country ----------+---------- 241080.31 | United Kingdom 13992.35 | Netherlands 10979.87 | EIRE 7364.88 | Germany 6430.54 | France 4763.14 | Japan 3857.75 | Australia 2700.86 | Singapore 1846.01 | Sweden 1816.14 | Switzerland (10 rows)

KPI Description	Ⅶ Transactions weekly
KPI formula	Total transaction, per week
Steps to realize KPI:	SELECT EXTRACT(DOW FROM invoicedate) AS dow, SUM(income) FROM onlineretail_zmy GROUP BY 1; --- dow | sum ---+--- 0 | 26465.68 1 | 45826.98 2 | 54926.27 3 | 64369.24 4 | 68925.73 5 | 43787.96 (6 rows)
Notes:	There are no transactions on Saturday between 1st Dec 2010 - 9th Dec 2011

KPI Description	Hourly income
KPI formula	Sum income, per hour
Steps to realize KPI:	SELECT EXTRACT(hour from invoicedate) as hour, SUM(income) FROM onlineretail_zmy GROUP BY 1; ---------- hour | sum -----+---------- 6 | -14.95 7 | 1091.06 8 | 10764.66 9 | 23991.70 10 | 52657.64 11 | 37398.89 12 | 44449.18 13 | 43910.41 14 | 33694.03 15 | 30734.96 16 | 14304.49 17 | 6932.15 18 | 3062.66 19 | 1135.46 20 | 189.52
Notes:	

Section 3: Executive Summary

1. Company and data Overview

This dataset has some unique features in different countries. Here are the features:

- This company is probably a UK-based international retailer with branches in many countries, and it provided retail to overseas using the port
- The products most related to the gifts for some important festivals
- Time: only one year from December 2010 to December 2011
- Customer are mostly live in the United Kingdom
- Customers sometimes are wholesalers

2. Analysis process:

In this report, we used the cohort analysis as well as some Key Performance Indicator to capture the hidden information from the retail's dataset. If we want to increase our marking spend, we must focus on the sales income improvements brought by the new and old customers and the customer retention rate for we want to generate a return on those investments instead of one-time consumption. To find out the efficient location for investment more on marking spend, we here tried to figure out the spending behavior patterns in four countries. We use that recommend planning our distributional tactics with the performance statistics, especially those items worth more efforts by comparing its popularity, customers performance as well as total trend.

3. Target Customers

We here compared the customer's retention rate by cohort per month and compare them total purchase amount. We also compared the time distribution for their shopping and difference in consumption amount. We inferred some customer behavior pattern among the four different areas and surprisingly found that the customers' shopping time, money, devoted are varied so much.

4. Most population items

Our goal is to identify the most popular items in different countries and whole areas. We compare its sales performance and recording in different regions. We also extract its description and tried to combine it with before analysis as a new feature. We find some clues profoundly connected with the previous patterns of sales, time and countries. The pictures and tables in the third and fourth part will explain our conclusions in detail.

Section 4: Comparative Analysis

This part aimed at using cohort analysis and KPIs in SQL and tableau to figure out the performance of the different region, spending behavior patterns, perceptions, interests and customer behavior. For analysis, we express our idea by listing similarities and differences between the indicators in the four countries as well as combination of related data and features.

Performance of four regions:

- Spending behavior patterns and perceptions and interests

This figure of KPI || Monthly Sales Income shows that November 2011 is the peak with highest levels， and the overall trend is growing(from the trend line and prediction) with the fluctuating cycle. This indicate that the total income of each countries is in continues growth, and its rapid growth before December and October and dropped a lot after that period. By combined with KPI- IV Most population items, they are more willing shopping before Christmas for gifts. For German, the steadily lines show that they are more likely to shop for small pieces from time to time. And for French, they hardly shopping on April perhaps for this period between the Winter sales and summer sales(the only time that stores can have a big purchase regulated by law), and they may be returned some items for the impulse spending in the last shopping season.

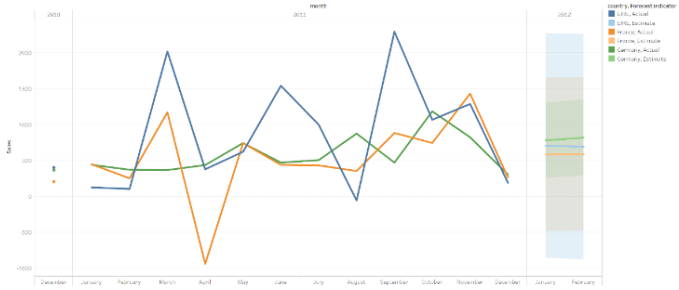

Figure 6: sales per month for France, Germany, EIRE

Review on the KPII Average transaction value(ATV), we can find that German ATV index is like the mature UK market and performed well for its sales amount and stability. Although ERE and French with higher ATV sometimes, but they fluctuate a lot even with higher return rate.

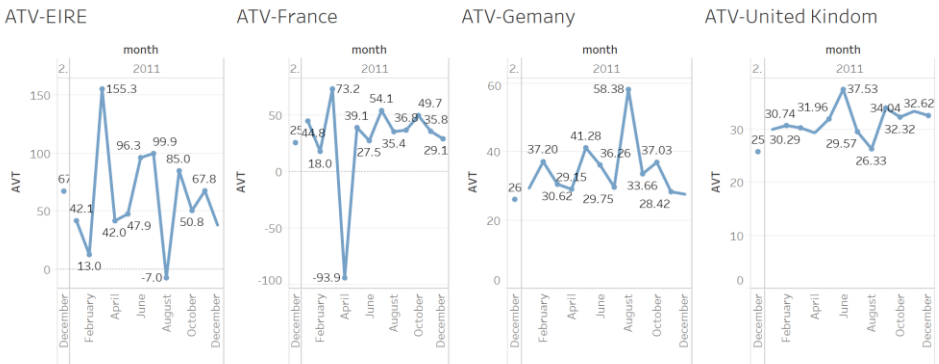

Figure 7: ATV for each country

Customer behavior

-Customer loyalty(retention rate)

Customer loyalty is measured by the retention rate. Here we compute the retention rate by cohort analysis- each customer repeat purchased by month to gauge the customers' loyalty.

Figure 7: cohort analysis for retention rate for each country

This figure shows the total trend of the cohort analysis for retention rate. Retention rate, as the opposite of customer churn, show the 'good customers' rate' in our sailing process. Tracking customers retention rate will help us spot the leaving early signs for a competitive company. The dips and improvements of this KPI will bring insight for us to investigate the causes. Above figure shows the cohort analysis for the retention rate. The varies of color from deep green to deep red, shows the percentage of retention rate from 0 to 0.5.

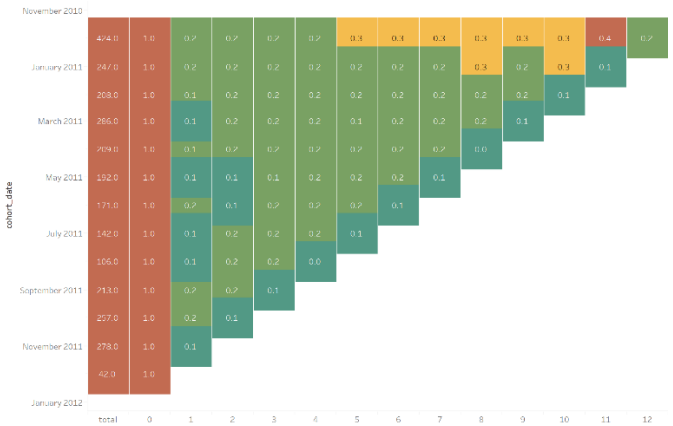

Figure 8. customer retention rate in UK

For the figure of total customers and the United Kingdom, they show all the block in deep and light blue, meaning the retention rate keep steadily to about 10% and 20%. And near the cohort-December 2010 with month 11 block become turn into deep red and yellow indicating

that some people more likely buy the product once a month. It perhaps for the consumer behavior of purchasing product near the Christmas to celebrate the festivals. Here we identify a sub-group- UK customers with a feature like to buying gifts only once a year in December.

Figure 9. customer retention rate in EIRE

The trend of seasonal buying evident in EIRE. Its figure only contains the cohort of December 2010 and February in 2010, which means that new customers are buying the new customer buying the product first only occur in that period perhaps for its advertisements or the attractive sales discount. Except for the null values, all the block with the dark red shows that higher customer retention rate more than 0.5 and higher loyalty in this country. So here we define the EIRE customer with the feature of higher loyalty and seasonal buying. Once establish the contact in customer and company, it will be a long-lasting customer life cycle throughout the entire relationship. It will also allow the new customer with the same pattern.

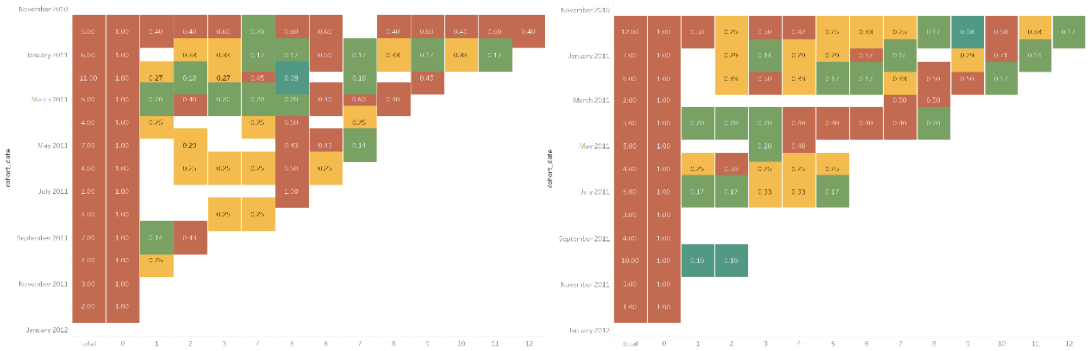

Figure 10. customer retention rate in France and Germany

For France and Germany, most of their block is yellow, or the dark red color shows that the retention rate of the customer is mostly over 25% percentage. For the line of cohort December in 2010, France shows the higher price for all months, and we think that new customer in French aroused in that cohort will be more devoted. For Germany, loyal customers are the new buyer in June or December.

Reconditions and conclusions(from KPIs):

Different country has different culture, income, education and consumer behavior, so we must decide the most effective areas for investment.

Here are some features of this retails data:

- UK with the highest numbers of orders far away from other countries, perhaps for its UK based company(Figure 1)
- The invoice with the highest money comes from EIRE(KPI Ⅰ ATV)
- November 2011 has the highest sales， and the overall trend is up with fluctuation(KPI Ⅱ Monthly Sales Income)
- By count and rank the total income of orders received by the company, there is peak in Thursday , and there are no transactions on Saturday in this data(perhaps they just close on Saturday)(KPI Ⅶ)
- The company receives the highest number of orders at 12:00pm, possibly most customers made purchases during lunch hour between 11:00am - 1:00pm(KPI Ⅷ)

UK	Germany	French	EIRE
Mature, most people, highest income	High user loyalty, ATV is close to UK	High user loyalty, large fluctuations, and high return rates	High user loyalty, moderate fluctuations, high ATV but fluctuating

Table 2: features of each countries

Recommendations for future investment:

Here we recommend we invest more during the lunchtime advertisement, especially for Wednesday to Friday in Germany focused on the gifts. Germany, with the features of loyalty, middle size marking and the sales of festivals items performed well. For its steady purchasing, we can store more gifts related well like Christmas gifts before December and provided more details in December will attract more customers. We also should more be focused on the items of description in ‘red’, ‘bag’, etc. For Germany, the company sometimes lacks the new customer in August; it would be better to provided discount items during that time.

All in all, Germany has huge market potential and performed well in almost KPIs, so we selected it as our target area.

APPENDIX:

- Invoice No: code starts with ‘c’ means cancellation with the quantity is negative
- Stock Code: distinct product
- Description: Product (item) name (infer its category)
- Quantity: The quantities of each product per transaction (negative means the returns)
- Invoice Date: Invoice Date and time, the day and time when each transaction was generated
- Unit Price: Unit price. Product price per unit
- Customer: Customer number uniquely assigned to each customer

Figure 11: Country name and map

Figure 12: Sales gradient count per invoices

Most orders concern relatively small purchases given that over 14605 purchases give prizes in less than £ 100.

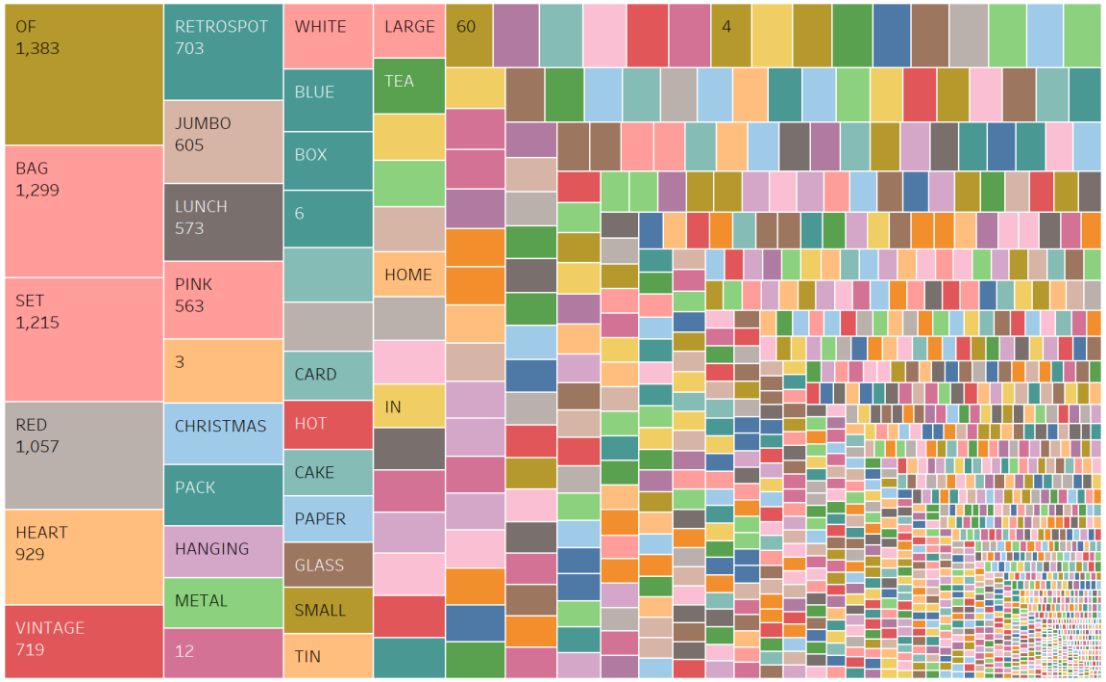

Figure 13: total buying word extract

We divided the description into different part using the tableau, and then we gather them into one line and draw a word map. As we can see, the word

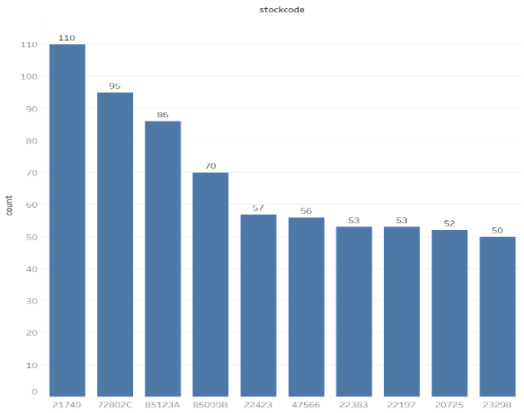

Figure 14. Totally popular items