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.

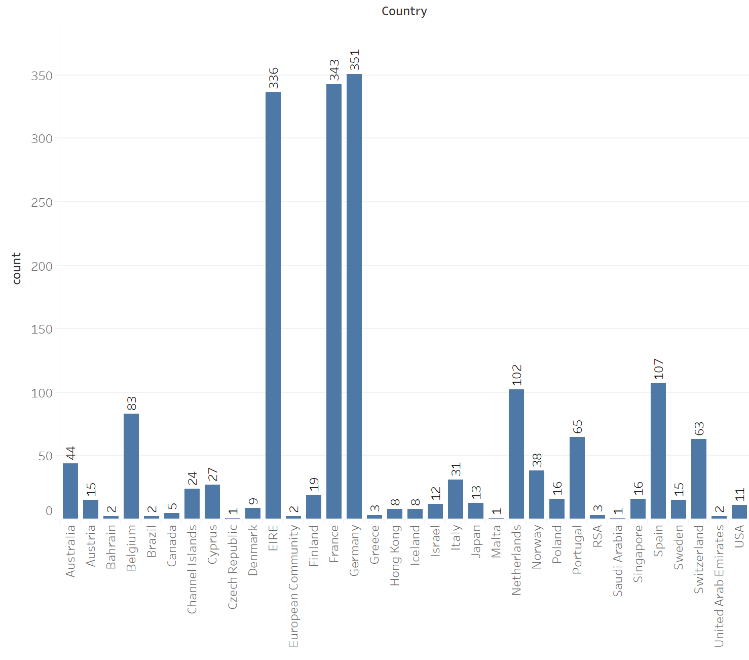
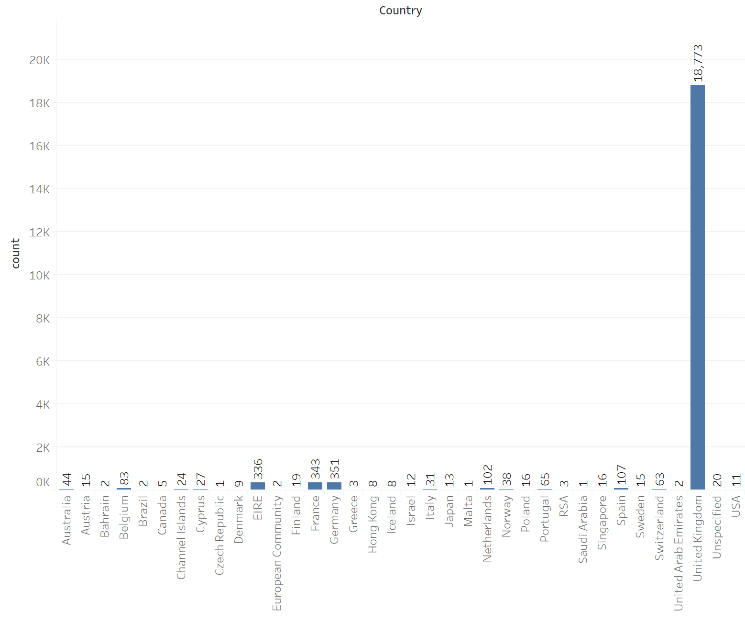


Figure 1. Bar chart of number of occurrences by country

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**  --------------------------------------   1. **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**  -----------------------------------------   1. **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**  -----------------------------   1. **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** | ⅠAverage 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. |

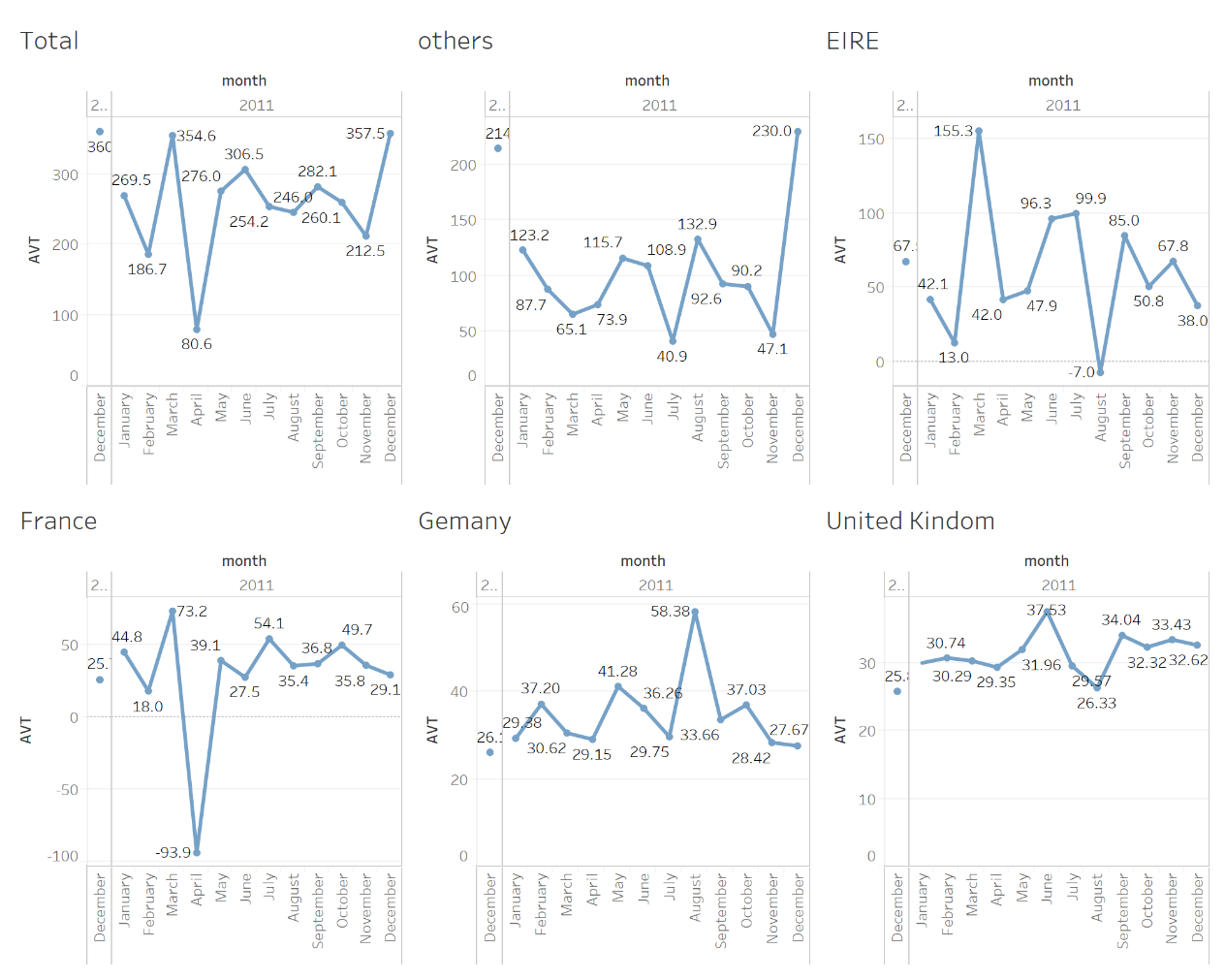


Figure 2: ATV for total and top 4 countries

|  |  |
| --- | --- |
| KPI Description | ⅡMonthly 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 |

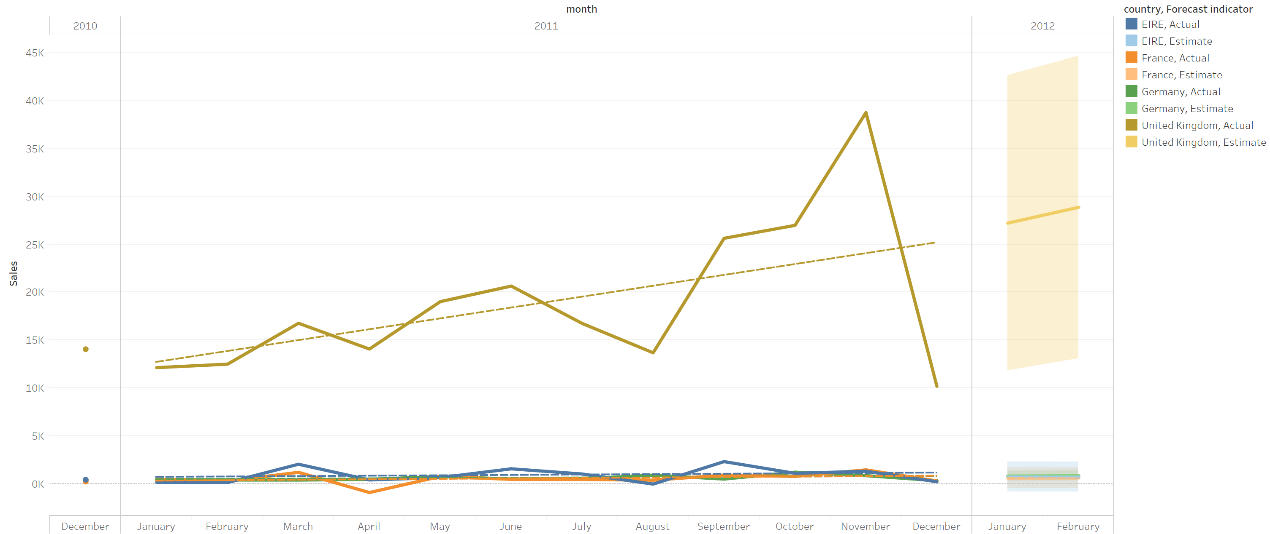


Figure 3: sales per month for each country and prediction

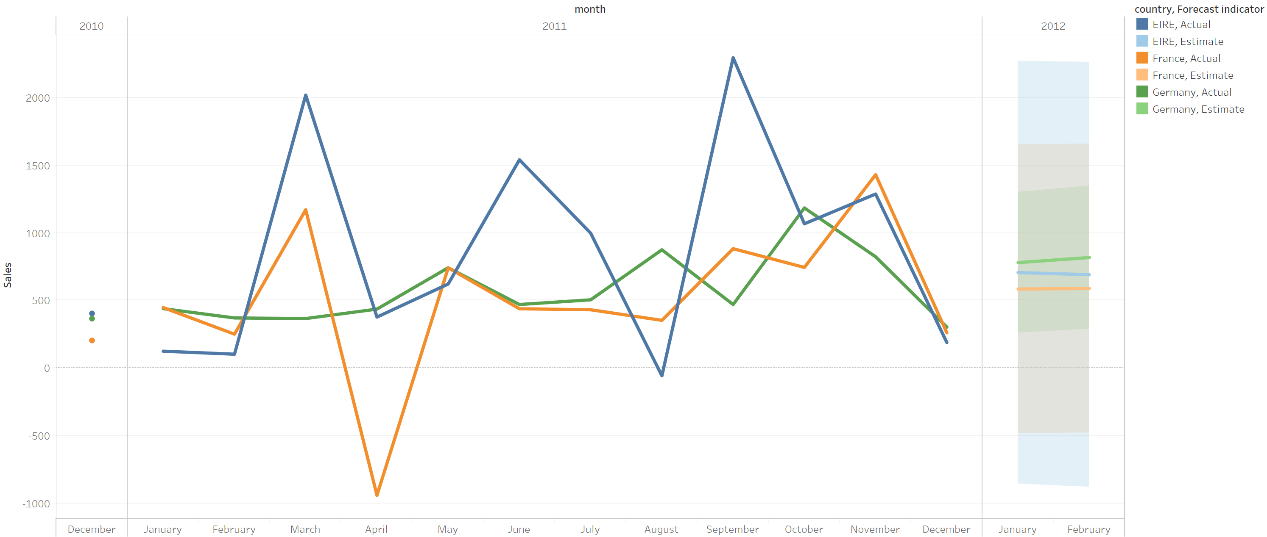


Figure 4: sales per month for France, Germany, EIRE and its prediction

|  |  |
| --- | --- |
| KPI Description | ⅢThe 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: |  |

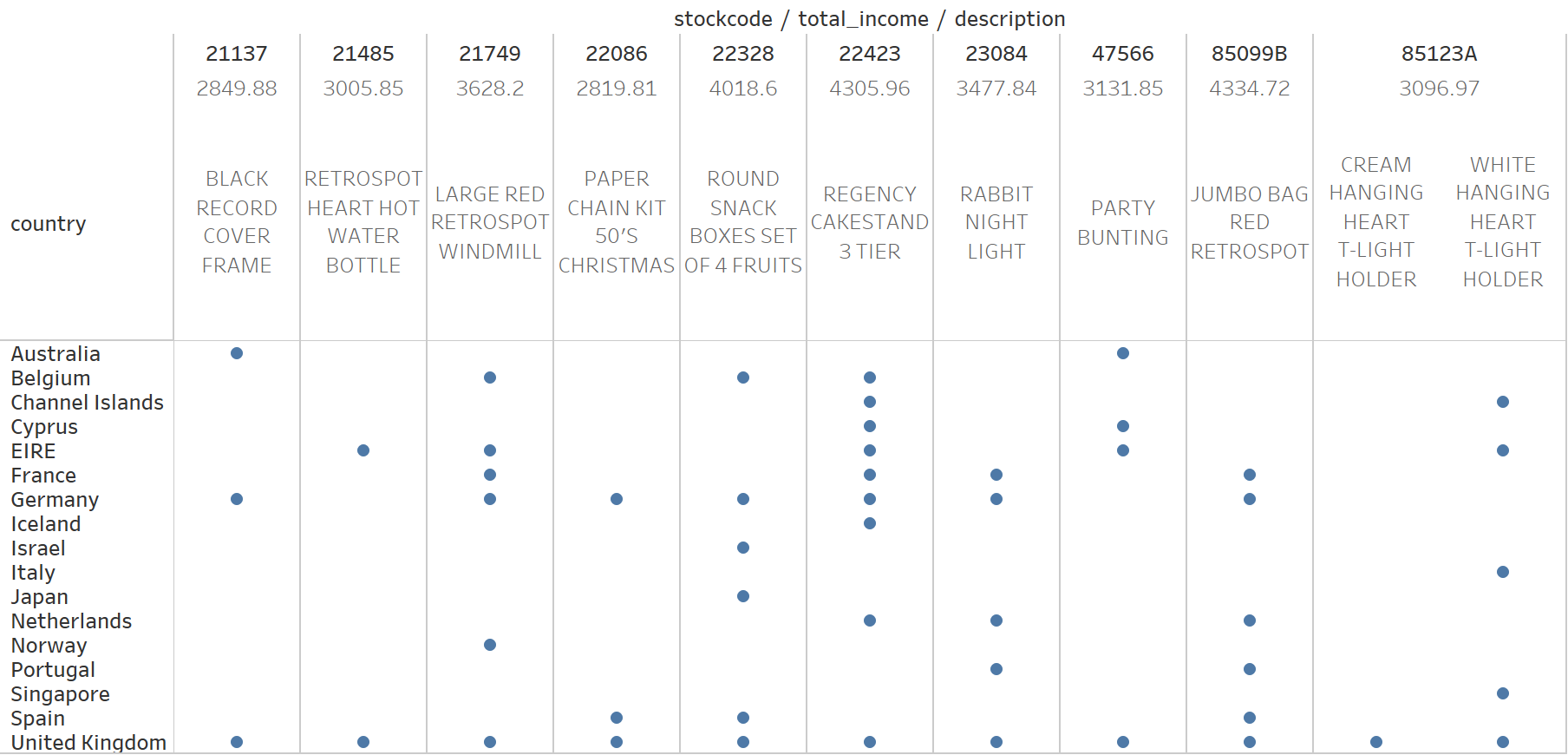


Figure 5: items with most incomes and its country

|  |  |
| --- | --- |
| KPI Description | ⅣMost 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: | ⅤTotally popular items |

Results:

|  |  |
| --- | --- |
| UK | EIRE |
| 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) | 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 | France |
| 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) | 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 | ⅦTransactions 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 | ⅧHourly 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**

* **Analysis process:**

In this report used the cohort analysis and KPIs to capture the hidden information from the retail’s dataset. We more focus on the sales income improvements and customer retention rate instead of one-time consumption. To find out the efficient location for investment, we here explore the behavior patterns with the performance statistics customers performance as well as total trend.

* **Recommendations**

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 and brand of ‘woodland’ before December and provided more free goods in December to attract more customers. 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.

* **Sales data Overview**

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

- This company is probably a UK-based international retailer and it provided retail to overseas(like Germany) using the port

- Different countries have different popular item brands, Germany more favor of ‘Woodland’ brand and gifts

- Customers are more wholesalers(like ERIE and Netherland), but for Germany they have more retailers

**- 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, like they hardly shopping on Saturday and midnight.

- **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 KPIⅡMonthly 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.

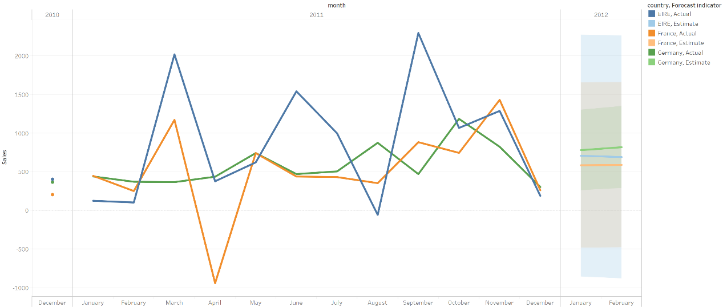


Figure 6: sales per month for France, Germany, EIRE

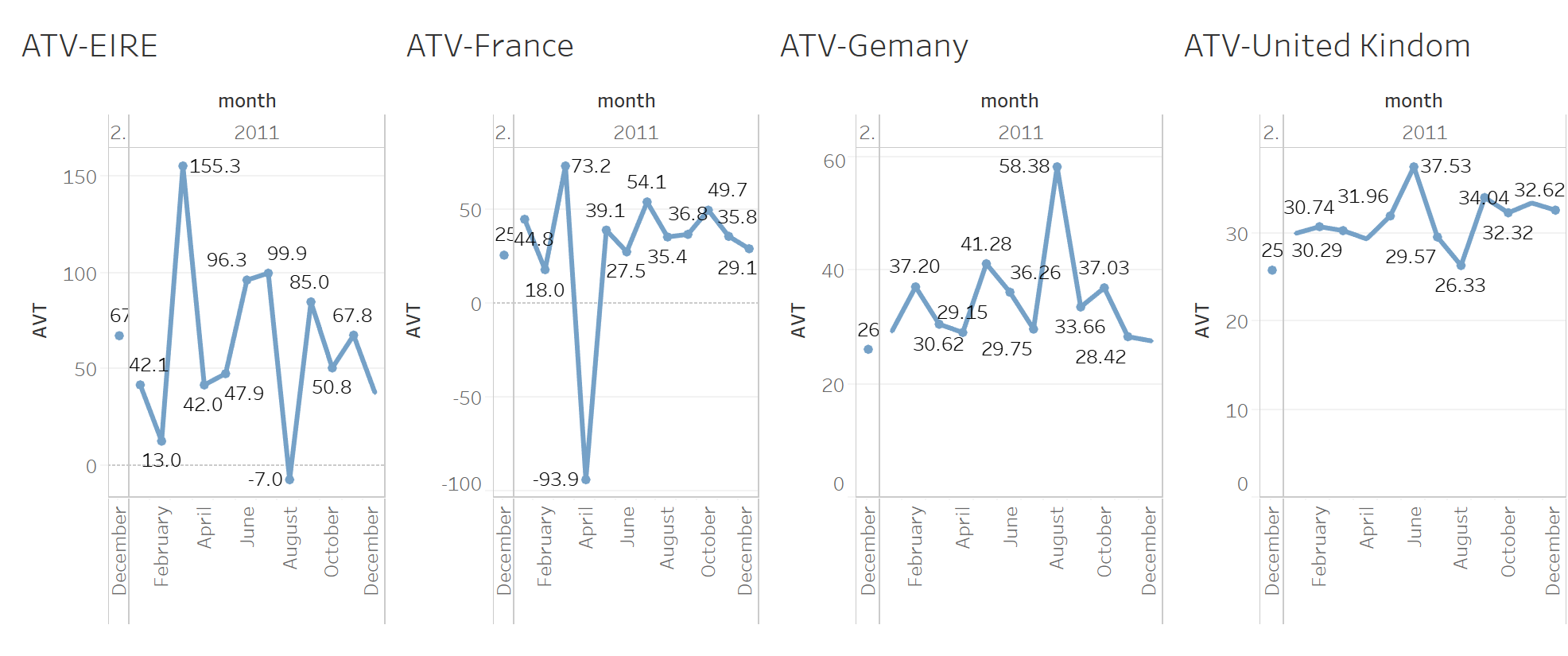
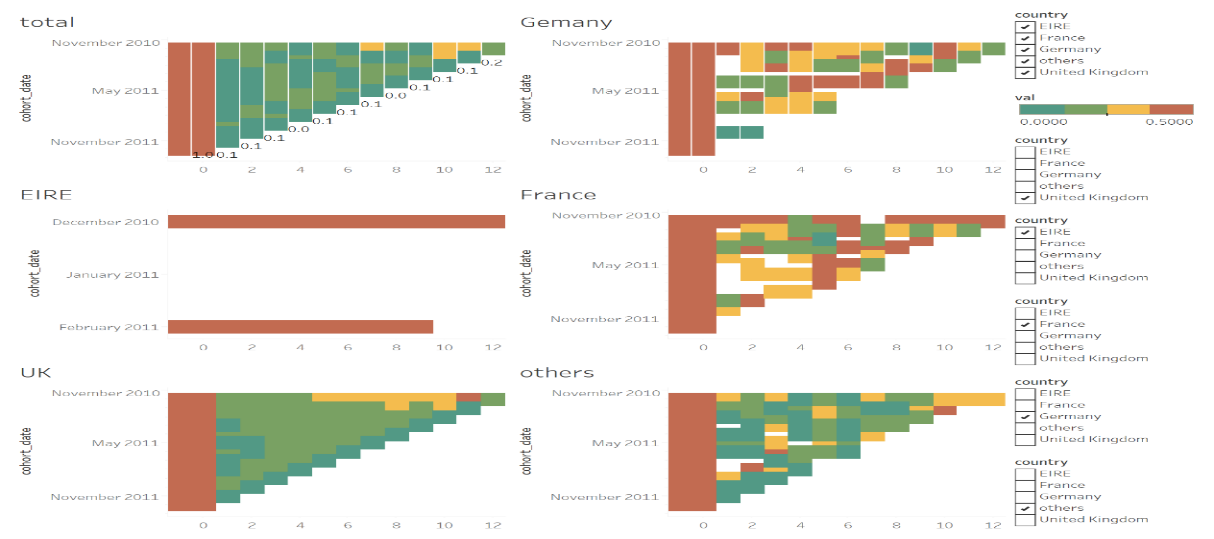
Review on the KPIⅠ Average transaction value(ATV), we can find that German ATV index is like the mature UK market and performed well for its sales amount and stability. Although ERE and French with higher ATV sometimes, but they fluctuate a lot even with higher return rate.

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.

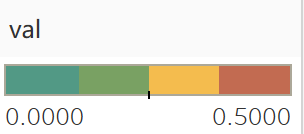


Figure 7: cohort analysis for retention rate for each country

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.

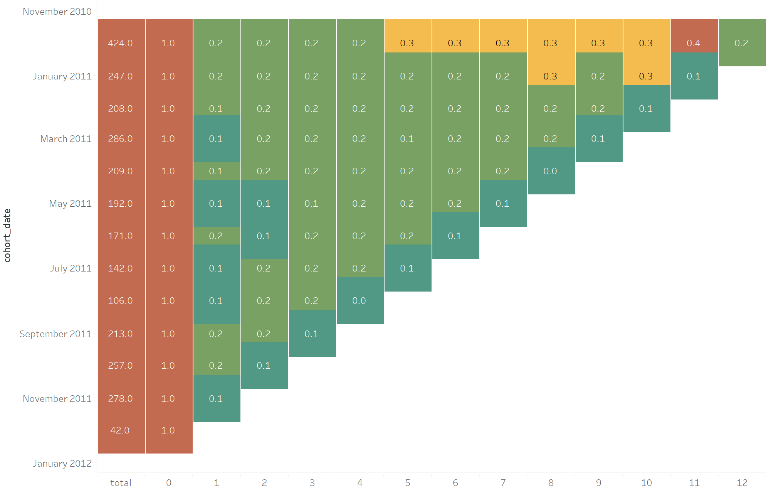


Figure 8. customer retention rate in UK

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.

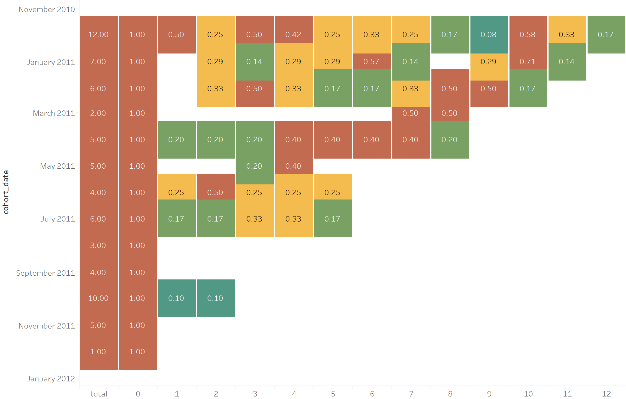
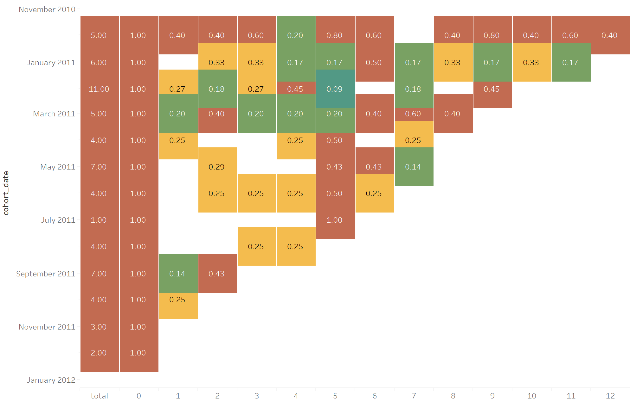


Figure 10. customer retention rate in France and Germany

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.

**Other features of each countries and comparative analysis of popular items:**

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(KPIⅠATV)
* November 2011 has the highest sales，and the overall trend is up with fluctuation(KPIⅡMonthly 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

For the analysis across popular items’ customers and sales performance:

* As for KPIⅤand descriptions, 72802C, 21759 is popular all the four regions. We also should more be focused on the items of description in ‘red’, ‘bag’, etc. Germany and French seem to like online retails using the post for its higher postage. The brand of ‘woodland’ is very popular in Germany. UK customers will buy more gifts for festivals.
* For its most popular purchased orders, ERIE have more wholesalers for its large income of each orders and per customers. For Germany there exists more personal customers for its long-lasting purchased orders and lower income each invoices.

**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

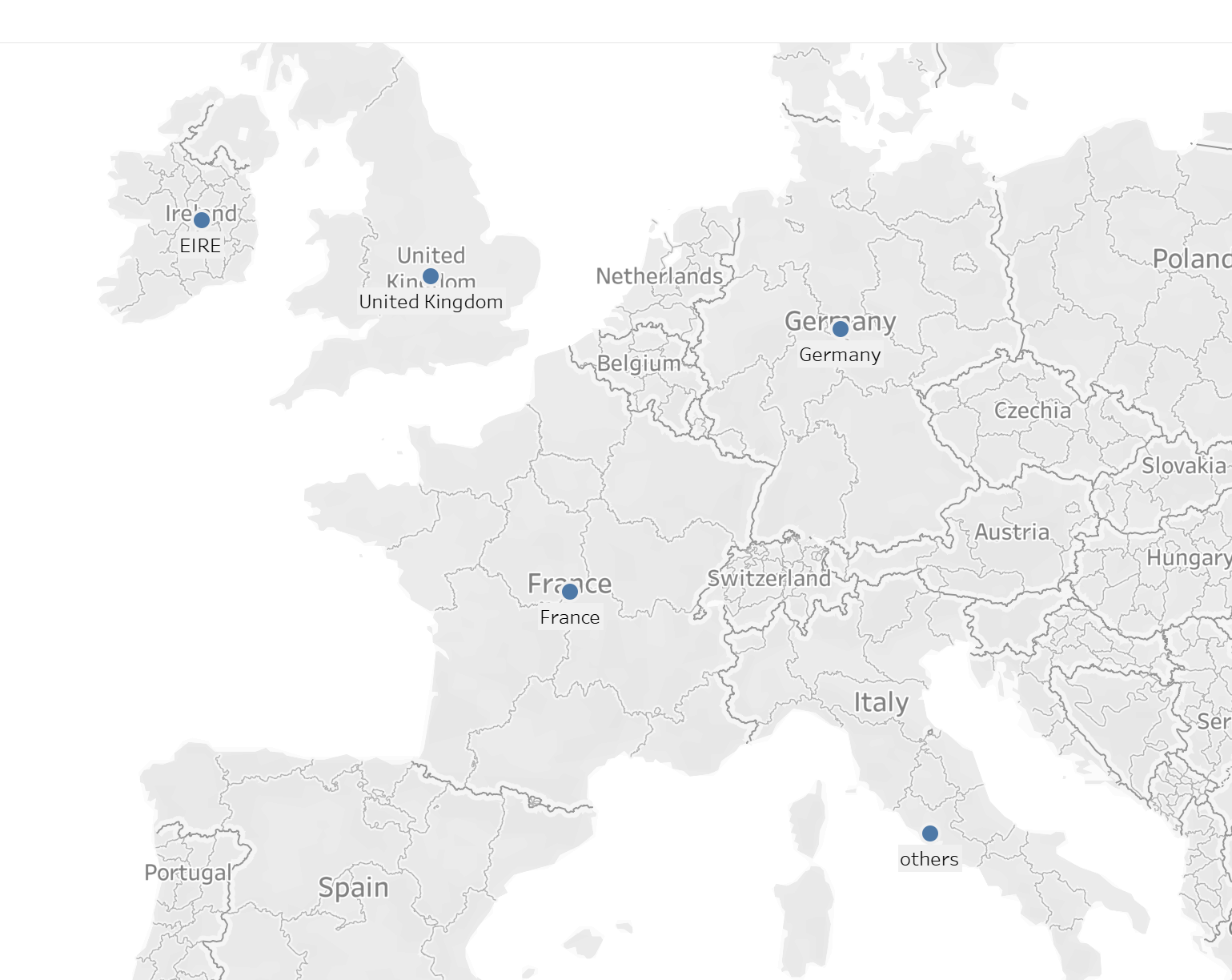


Figure 11: Country name and map



Figure 12: Sales gradient count per invoices

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

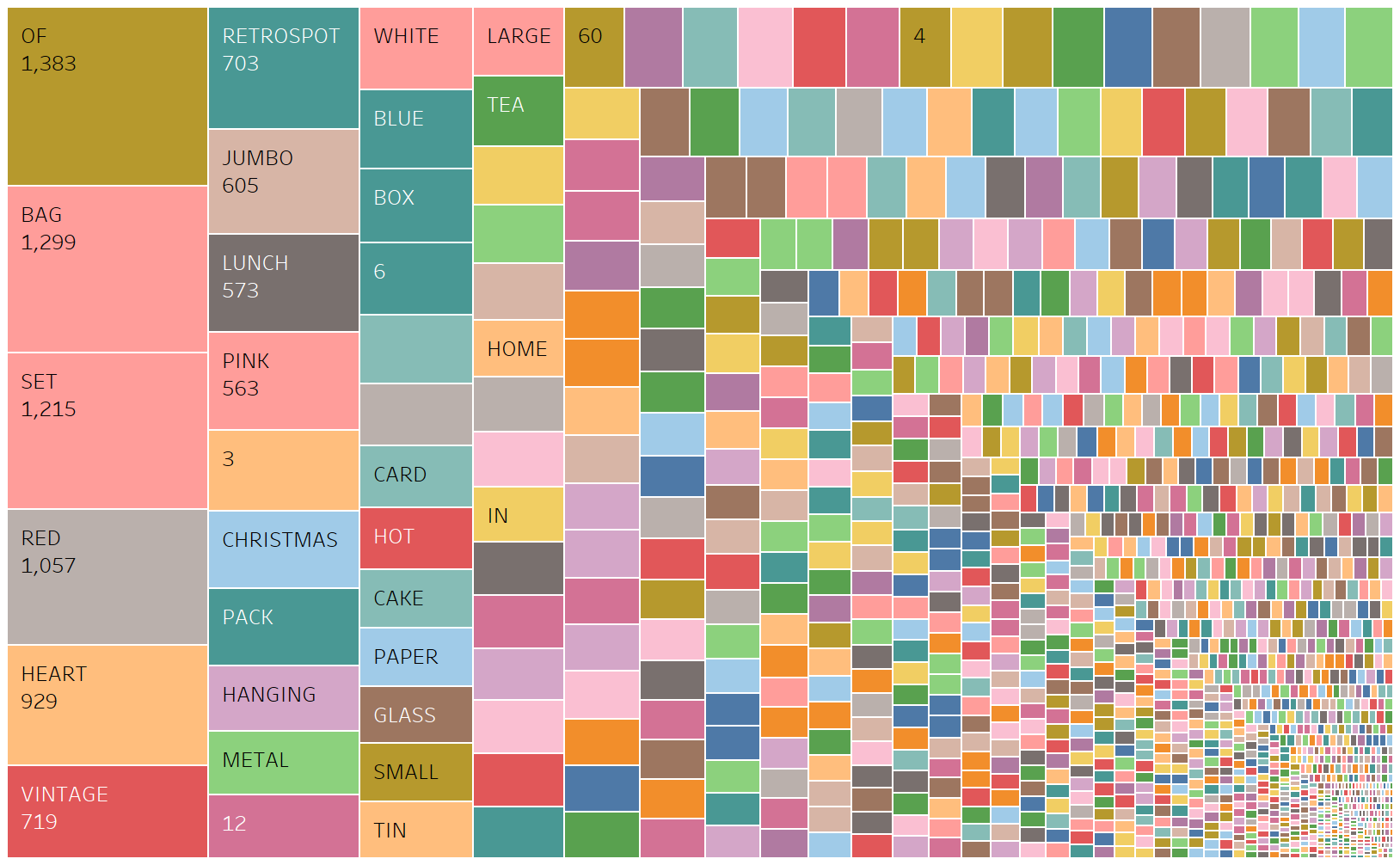


Figure 13: total buying word extract

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

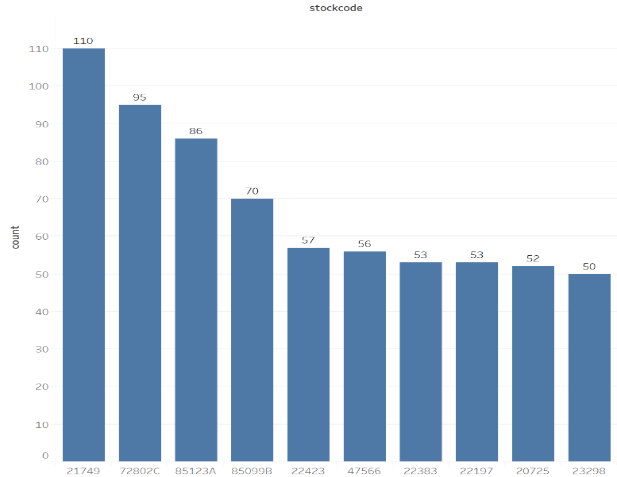


Figure 14. Totally popular items