Transaction History Data Set Analysis

1. Introduction

The data set given comprised of a record of transactions of a company from 12/01/2010 to 12/09/2011 with the definition of the variables given as follows:

1) Invoice number: A 6-digit number uniquely assigned to each transaction. If it starts with letter ‘C’, it indicates a cancellation.
2) Stock code: A 5-digit number uniquely assigned to each distinct product.
3) Description: The product name.
4) Quantity: The quantities of each product per transaction.
5) Invoice date: The day and time when each transaction was generated.
6) Unit price: Product price per unit in sterling.
7) Customer ID: A 5-digit integral number uniquely assigned to each customer.
8) Country: The name of the country where each customer resides.

Given this information, initial insights and trends in customer transaction. Then, text mining was used to categorize products based on the words that make up the description of each product. Using the categorized products, customers were categorized into 9 categories via k-means clustering. Finally, a prediction model for customer category based on purchase history was made using random forests.

2. Data Cleaning

First, the data was prepared by getting rid of any entries which have any blank columns and then deleting any duplicate entries so that we have a data set fully filled data set with no duplicates. After these two processes, the data was reduced from 541,909 entries to 401,604 entries. Now the data is ready for analysis and for the remainder of this paper when referring to the data set, it will be this cleaned data set.

3. Initial Exploration

3.1 Content

This is a rather large data set but with many repeated entries for invoice numbers, product descriptions, customer IDs, and countries. I decided to see how many unique entries were in each variable and obtained the following results.

Number of Unique Entries	
Customer ID	4372
Stock Code	3684
Invoice No.	22190
Country	37

Indicating that there are 4,372 unique customers from 37 countries who purchased from a range of 3,684 products through 22,190 orders.

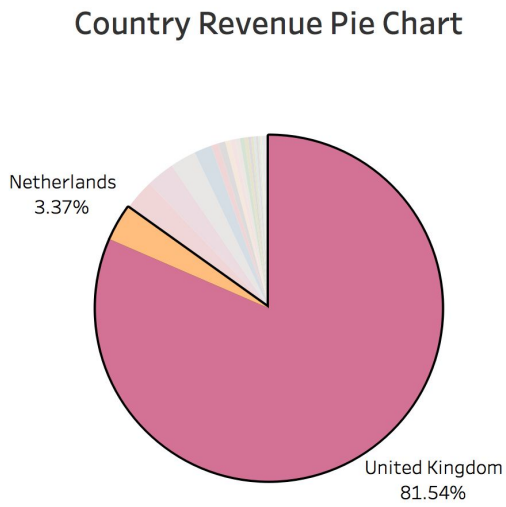
3.2 Country

The Country column of this data set indicates that there are customers from 37 countries but in which countries do the customers reside?

The colored countries in the map above are the countries in which customers reside. We can tell from a glance that the United Kingdom seems to be the most important contributor to the company's revenue. We can also see that Japan is the only Asian country to have customers and that a lot of European countries, Australia, the United States, Canada, Brazil, and Saudi Arabia are all places in which customers reside in. Overall, customers come from 37 countries.

From the color-coded map above it is easy to tell that a lot of revenue comes from the United Kingdom but exactly what percent of the revenue does it contribute? The pie chart below shows us that the United Kingdom contributes 81.54% to the company's revenue making it by far the most important

contributor to revenue, the second highest contributor being Netherlands only makes up 3.37% of the revenue.



3.3 Stock Code

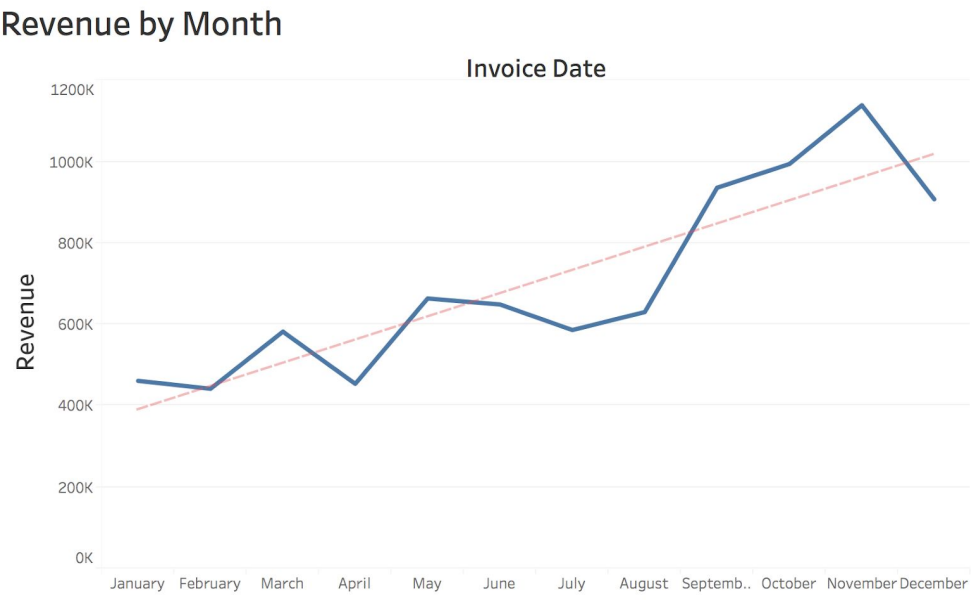
One question I had was specifically in regards to the Stock Code. Do all of them really have a 5-digit number? After searching for Stock Codes that do not adhere to this definition, I obtained the following results.

Stock Code	Description
M	Manual
POST	POSTAGE
CRUK	CRUK Commission
D	Discount
C2	CARRIAGE
BANK CHARGES	Bank Charges
DOT	DOTCOM POSTAGE
PADS	PADS TO MATCH ALL CUSHIONS

The results showed that not all Stock Code are 5-digit integers. The ones that are not a 5-digit integer seem to be special codes that are mostly not relevant to the product but rather are linked to operations or services.

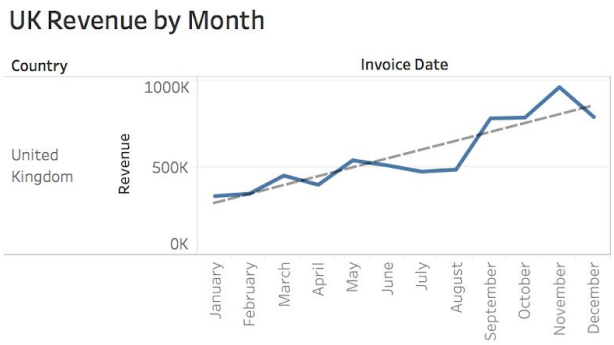
4. Trends

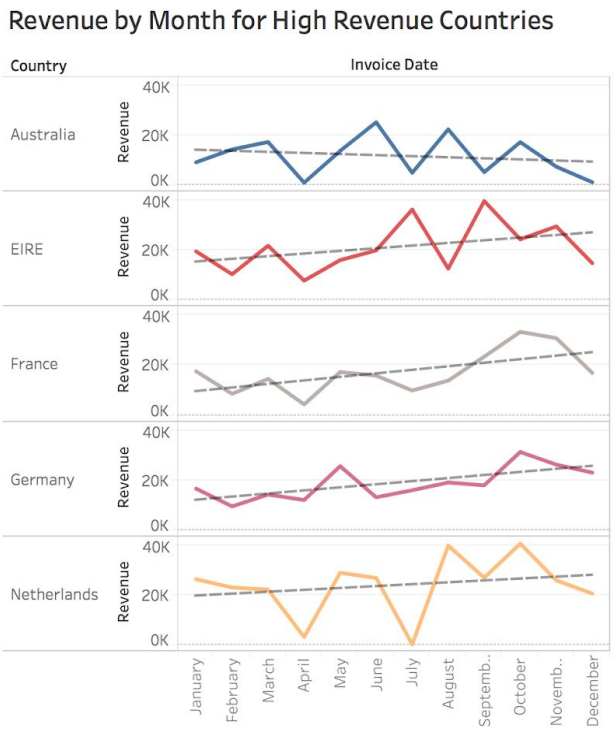
4.1 Country Trends

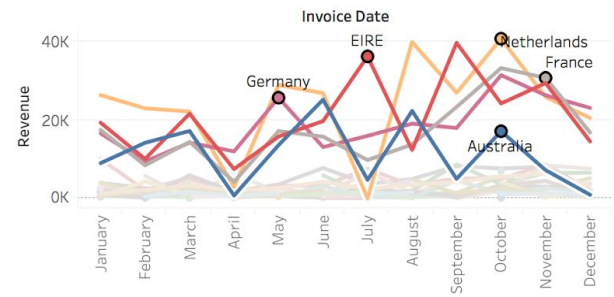


We can see that November has the highest revenue. However, transactions history in the data set is only up to 12/09/2011 and so the revenue for December is not revenue for the whole month but only for around the first third of the month. Despite this fact, December still has quite a high revenue yield and one can assume since it is holiday season that the revenue for the whole month of December will be quite high.

We can also see that the revenue for this company is increasing . The line indicates around a $57,038 increase per month and the p-value is less than 0.0001 indicating that this trend is very unlikely due to chance. Since the United Kingdom is such a big contributor of revenue, there is a possibility that the upward trend is purely due to the revenue from the United Kingdom and all other countries may be performing very badly. To see if this is the case I separated the revenue by country and found that I had to separate revenue from UK from all other countries because the revenue from UK were so high that the revenue from other countries were uninterpretable in the graph. After separating UK from all other countries, I obtained the following graphs:

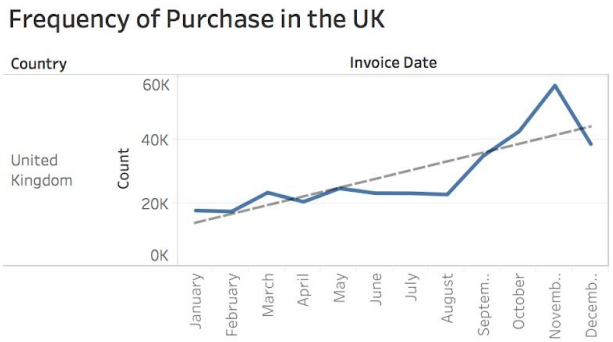


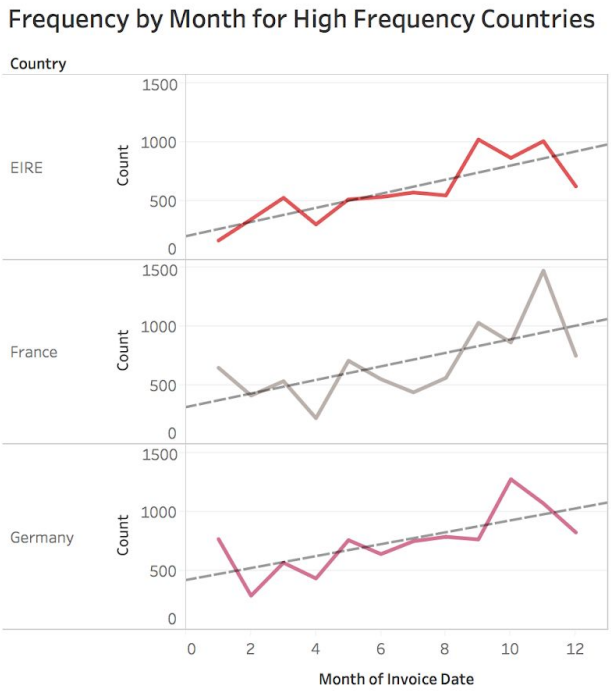


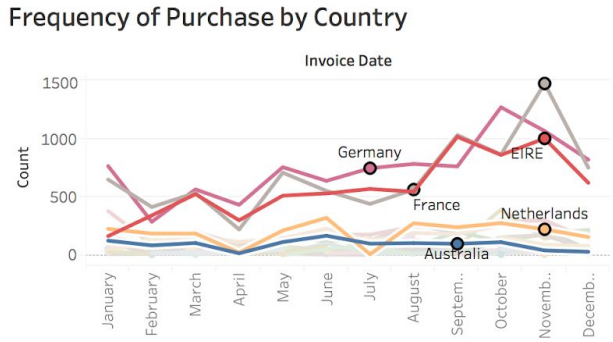


The UK has a trendline that is very similar to the overall revenue trendline that was seen earlier. However, we also see that EIRE, Netherlands, France, Germany, and Australia seem to have higher revenues than the other thirty-two countries. I decided to fit a trendline to each of the five countries to see if they show an upward trend and found that they all do with the exception of Australia. However, the p-value for Australia's trendline is 0.54 indicating that we cannot reject the null hypothesis that there is no relationship between month and revenue and this trend has a high probability of being due to random chance. From this, we can see that the United Kingdom is the driving factor behind the company's overall trend in revenue.

Revenue may not indicate how frequent purchases are in each country so I decided to analyze how frequently products were bought in each country in a similar manner as I did for the revenue and obtained the following graphs as a result.





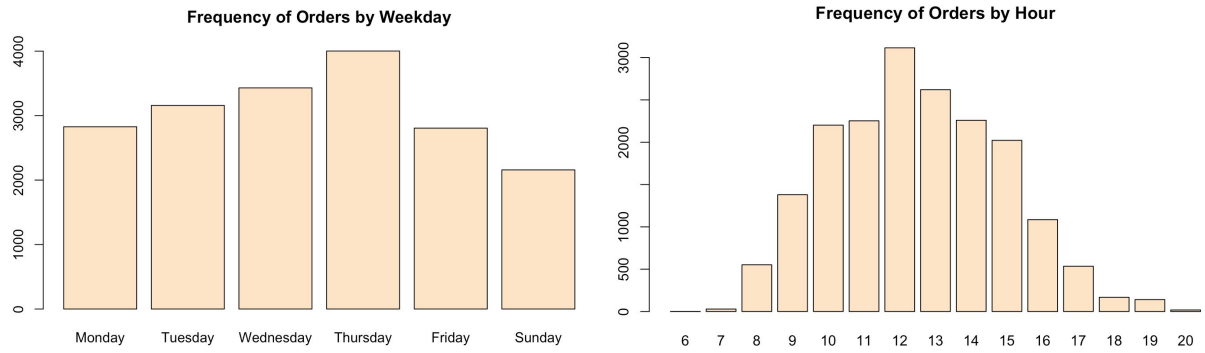


The frequency graph overall was very similar to that of the UK's but the frequency of purchase by country gave interesting results. The relatively high revenue yielding Australia and Netherlands have relatively low frequency indicating that in these two countries, the high revenue was not due to high volume of purchase but rather the high cost of these purchases. The trendlines for EIRE, France, and Germany have positive slopes and low p-value indicating that there is a significant relationship between month and revenue in these countries and more and more purchases are being made in these countries.

4.2 Customer Trends

Now that we know trends in the overall revenue and frequency of this company, an important analysis is to see the trends customers who make up these aggregate trends. I first decided to see how many cancellations make up the data set. There were 8,872 cancellations over 3,652 unique invoice numbers. The cancellation entries makeup 2.21% of all entries. I decided to remove cancellations from the data set and also removed the 1,716 transactions that had matching cancellations to conduct analysis on transactions that resulted in profit for the company.

First, I decided to see the trend in purchases during the week. I counted in terms of unique invoice number so as to not count one order multiple times for every product in that order. The barplots below represents how frequently customers made orders during the weekday and during what time of day.

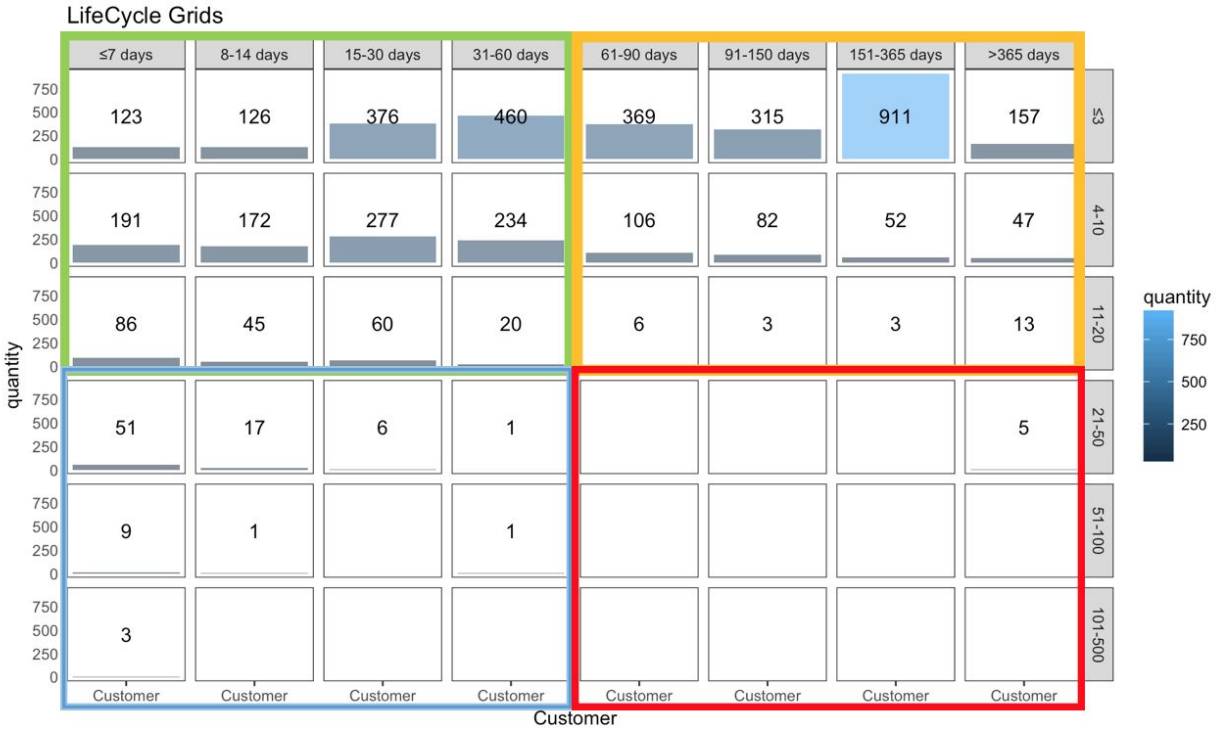


We can see that customers most frequently place orders on Thursday. One interesting thing to notice is that there were no purchases made on Saturday. It may be the case that this company does does not open for business on Saturday. We can also see that this company opens from 6 AM to 10 PM and has a peak time of 12 PM.

5. Customer Segmentation

5.1 LifeCycle Grids

One method of Customer Segmentation is through LifeCycle Grids which is based on the recency and frequency of each customer. Recency is measured in days since the last purchase and frequency is number of orders placed by the customer.



Above is the LifeCylce Grids for the customers of this data set. It has been color coded into four segments and can be interpreted as follows:

- Blue: These are the best customers who have bought a lot of items recently. For marketing purposes these customers could be considered the VIP customers
- Green: These are the customers who have bought a few items recently. The customers in this segment have the potential to move up to being a VIP customer.
- Red: The customers in this segment are former VIP customers and it is important to understand these customers for future customer retention.
- Yellow: The customers in this segment are also previous customers but these customer were not big buyers. It is still important to understand this segment for future customer retention.

We can see that there are a lot of customers in the yellow segment indicating that lots of customers chose not to buy from this company after having bought a few things. There may be many causes to this such as broken products or late delivery but one thing that could potentially retain these customers is through effective marketing strategies. This brings us to our next understanding customer preference in products to create a recommendation system.

5.2 Customer Segmentation Based on Product Description of Past Purchases

To segment customers according to their purchase preference, the first step is product categorization based on its description. In order to do this we break each description into words, remove punctuation, numbers, stop words such as “the”, “that” and also stem the words to its base form. After doing this, I created create a text document matrix which indicates which words are in each description in binary form. I decided to apply k-means clustering and decided to use 9 clusters based on the plot of the within-cluster sum of squares. The following word clouds show what kind of words were included in each cluster.

Cluster 1

Cluster 2

Cluster 3

From the word cloud there does seem to be some overlap but there does also seem to be distinct features in each cluster.

Now that the products have been categorized, customers can also be categorized with this given information and purchase history. The way in which I decided to categorize customers is by seeing what proportion of money they spent overall in each cluster and place them into 9 categories that align with the product clustering that they spent the most money on. For example, if a customer spent the most proportion of their purchase on products that are in cluster 7, they will be put into customer category 7. The number of customers in each category is as follows:

Customer Category	# of Customers
1	73
2	5
3	119
4	4007
5	54
6	33
7	34
8	1
9	1

We can see that the clusters are not evenly distributed with a vast majority of customers in cluster 4 and only one customer in cluster 8 and cluster 9.

5.3 Customer Prediction Model

In order to retain customers and improve customer satisfaction, it is a good idea to make a recommendation system especially for e-commerce to help customers find products that they like.

Up to this point, the classification has been one of unsupervised learning since we had input data but corresponding output data. Now that each customer has been given an output data in the form of a category label, I can now apply supervised learning techniques to build a prediction model.

5.3.1 Random Forests

I build a model using random forests to predict customer category based on total amount spent per product category in each invoice and how much money was spent in each invoice. I trained this model on 0.7 proportion of the data set and test it on the remaining 0.3 of the data set and by using confusion matrix found the model with the test set as the input to have an accuracy of around 98.16%.

One thing that I would like to note is that the first attempt at modeling was not with the total amount spent per product category in each invoice but with the proportion spent per purchase per invoice. Using the same method as above but with these inputs, the model had an accuracy of below 1% so it seems that using proportion of money spent on each category of product per invoice is not useful in creating a predictive model for customer category.

5.3.2 Linear Discriminant Analysis

The next method I used to make a prediction model was Linear Discriminant Analysis. I used how much money was spent per product category to predict customer category. Again, 70% of the invoice

entries were used to fit the model and 30% of the entries were used to test the model. Around 96.89% entries in the test model were given the right customer category using this model.

Conclusion

The data indicates that this company has a large customer base in the United Kingdom and has a increasing trend in revenue. Many of the customers bought a small quantity not very recently. To enhance customer experience it is a good idea to segment customers to provide product recommendations so that they can find products that are to their preference easily. Customers can be categorized into 9 categories using k-means clustering based on the description of their previously purchased products. With this categorization of customers, a prediction model can be built using random forests and linear discriminant analysis. After fitting a model based on 70% of the invoice entries, we can test the model with the remaining 30% of the invoice entries. Comparing the results two models we find that the random forests model produces slightly better results.