Springboard Capstone 2

Attempting to predict the next month of sales and cancellations

Topics:

- Who is this presentation for?
- About the data
- Clustering insights
- Time-series insights
- Other insights
- Conclusions and ideas for moving forward

Who is this presentation meant for?

The marketing and/or product ordering department(s) of the company

*** Of note, this project is also to satisfy the Springboard capstone 2 requirements and would need more work before it could be presented before a company. ***

About the Data

Where is the data from and what does it contain?

This project utilizes data from:

- The UCI repository containing sales and invoice data for a single company based in the UK
- A website containing a list of holidays in the UK which was made into an excel spreadsheet

More about the sales data

The sales data contains columns for:

- InvoiceNo: Invoice number. Nominal, a 6-digit integral number uniquely assigned to each transaction. If this code starts with letter 'c', it indicates a cancellation.
- StockCode: Product (item) code. Nominal, a 5-digit integral number uniquely assigned to each distinct product.
- Description: Product (item) name. Nominal.
- Quantity: The quantities of each product (item) per transaction. Numeric.
- InvoiceDate: Invoice Date and time. Numeric, the day and time when each transaction was generated.
- UnitPrice: Unit price. Numeric, Product price per unit in sterling.
- CustomerID: Customer number. Nominal, a 5-digit integral number uniquely assigned to each customer.
- Country: Country name. Nominal, the name of the country where each customer resides.

Cancelled versus purchased products

I separated the cancelled invoices (9,288 products) from the purchased products (532,621 products) and analyzed them individually.

In an attempt to predict the next month of activity in cancelled and purchased products, I utilized clustering and time-series analysis.

Clustering insights

But first, what is clustering?

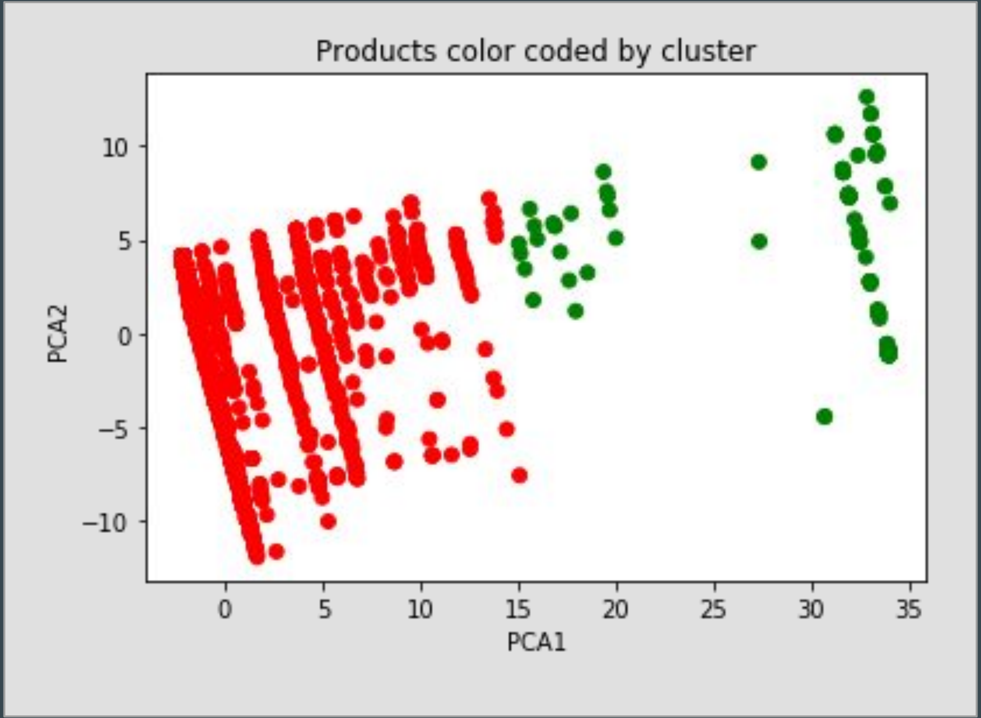
Clustering aims to take data and put it into “piles” or clusters based on similar characteristics.

The benefit is that we don’t know what the potential categories are before we start, so we avoid missing a possible common characteristic.

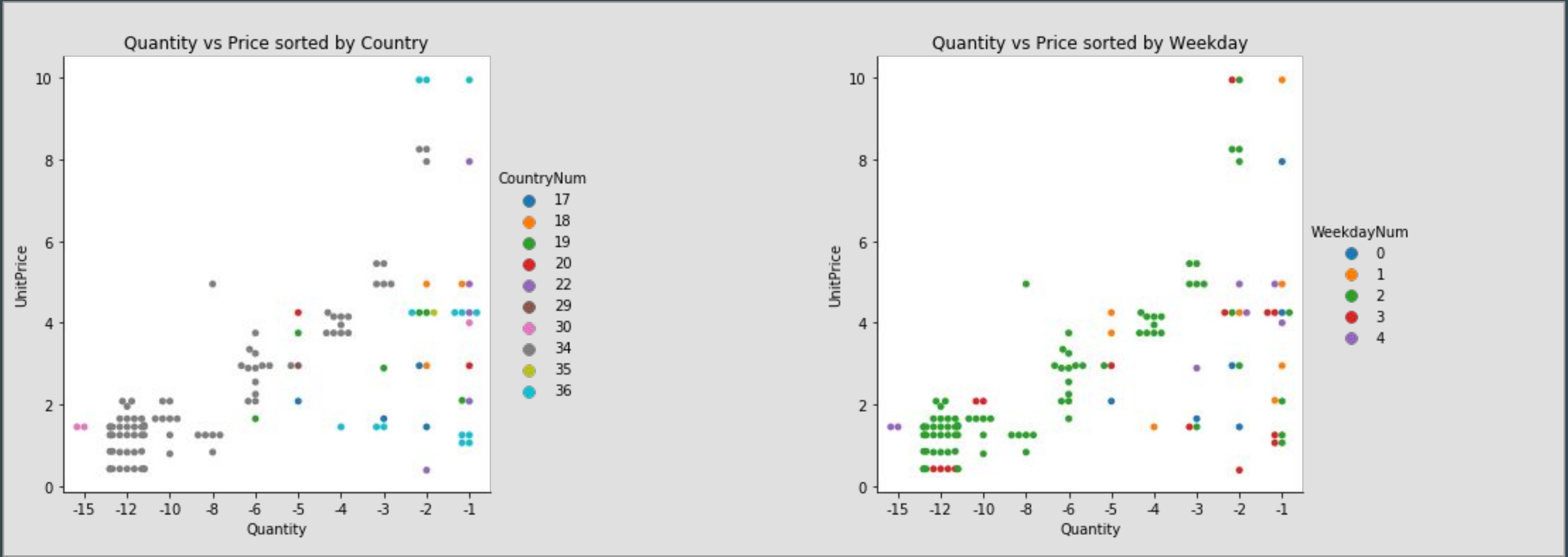
Cancelled products

Statistically, the best clustering analysis gave two clusters that appeared to be sorted by which country was doing the canceling.

The problem is that sorting into two clusters is not a particularly useful prediction method.



Insights from clustering

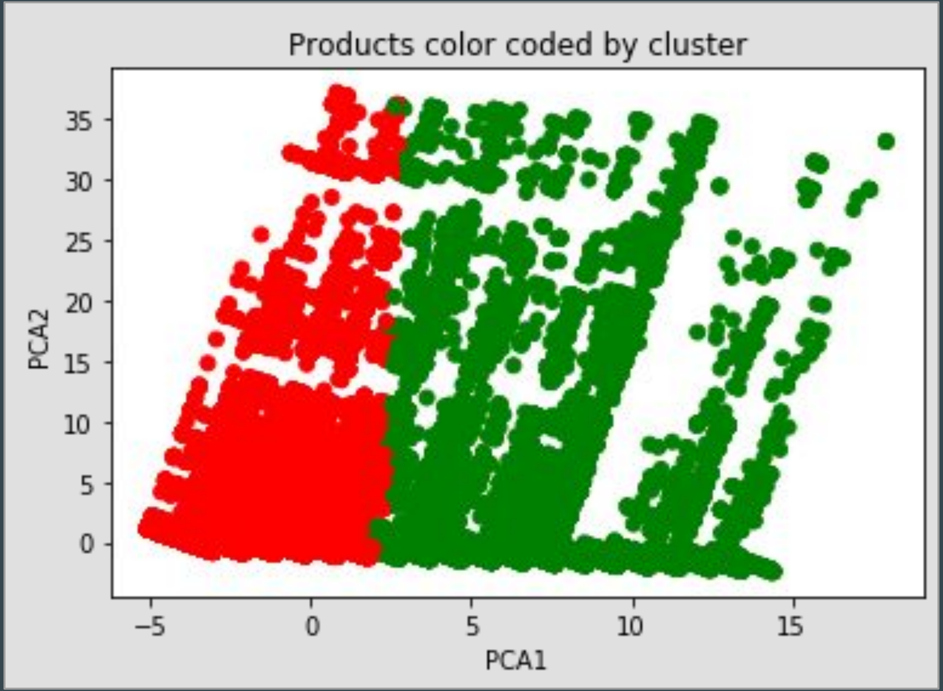


Sorting by country and by weekday did provide the insight that country 34 (USA) aligns well with weekday 2 (Wednesday) suggesting that Americans cancel most frequently on Wednesdays.

Purchased products

Again, the best clustering, statistically, was two clusters.

As you can see, there is not a clear line of separation between them, meaning that you would have great difficulty actually differentiating between the two, especially on the border.

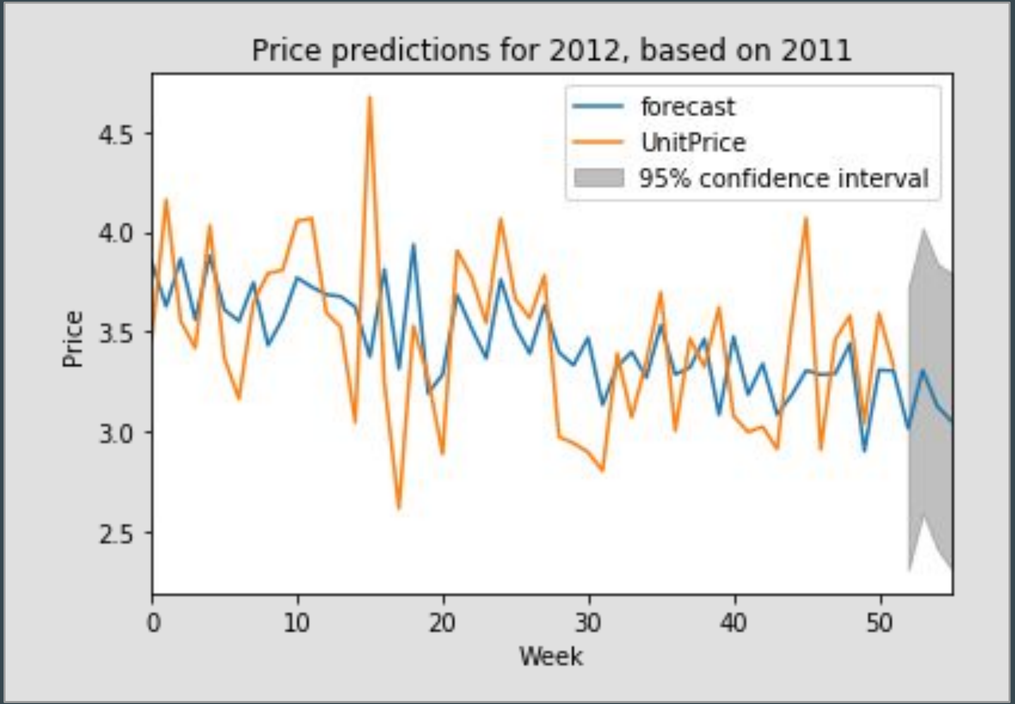


Time-series insights

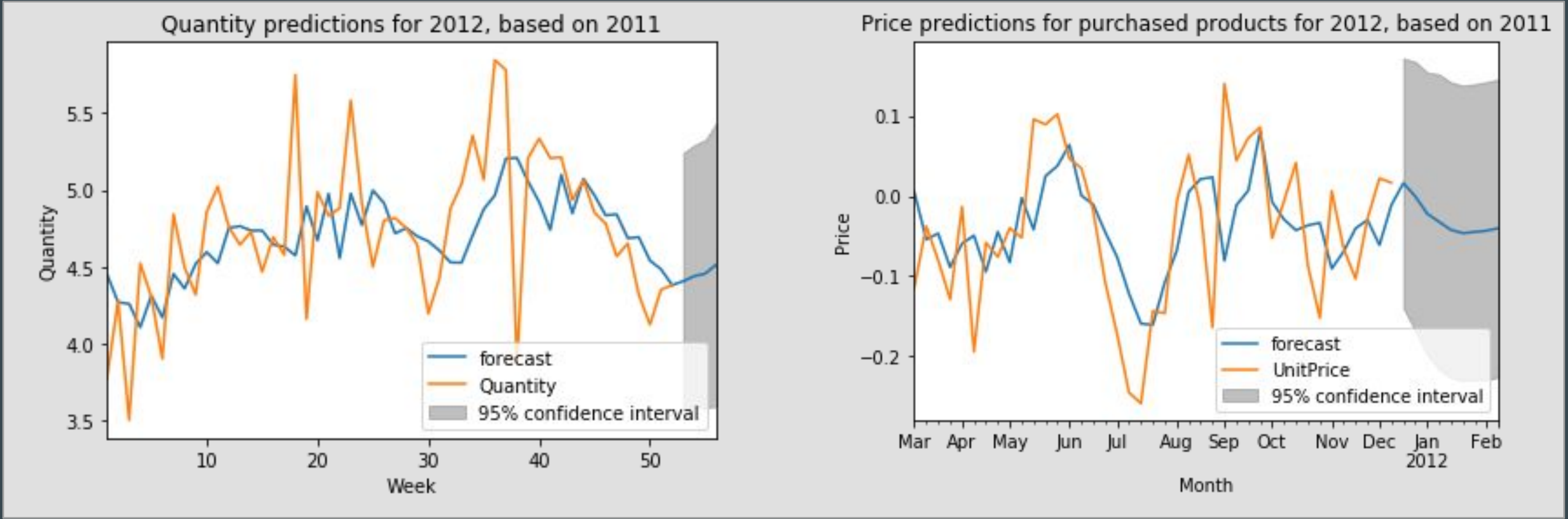
Cancelled products

The best time-series analysis concerned the price of product.

While the trend does suggest and predict an overall negative trajectory, the grey shaded area shows that this forecast could be anywhere within that range.



Purchased products

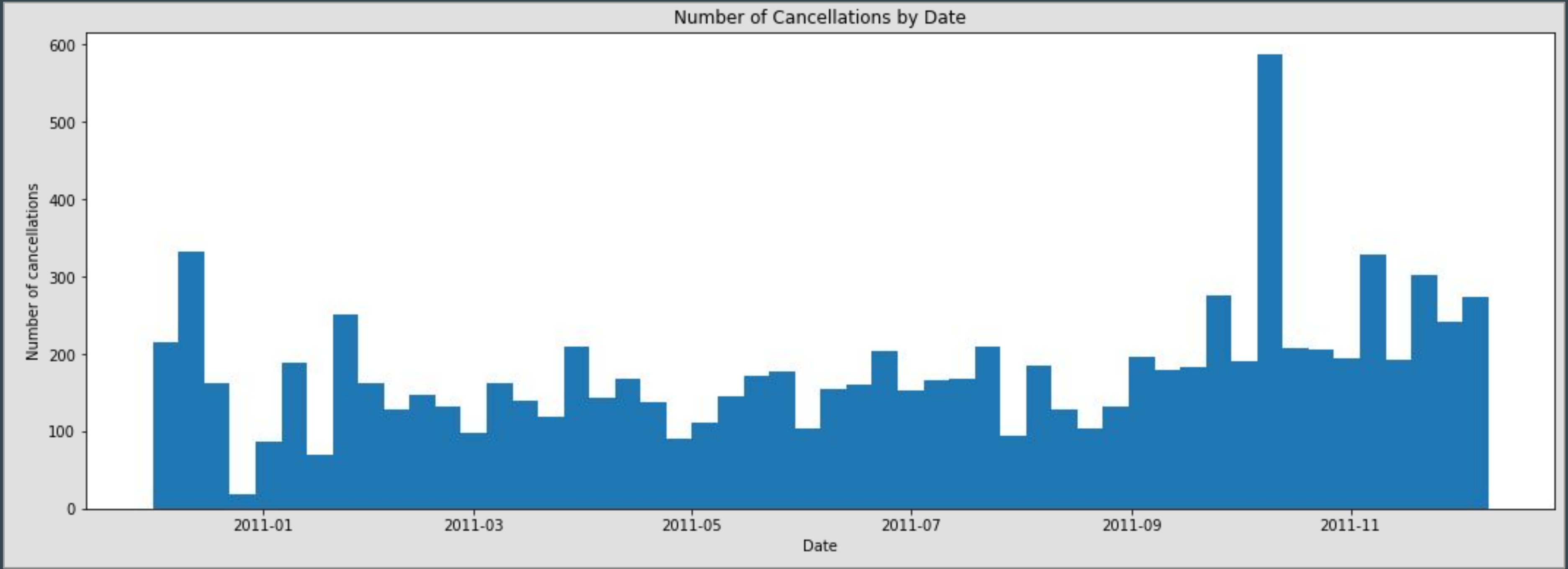


Both quantity and price predictions for purchased products have visually better fitting models, but unfortunately, the predictions for both have the same large grey area of uncertainty.

Other insights: what worked and what did not

Cancelled products

I pulled out all the cancelled invoices and plotted them by date:



Looking closer at the peak

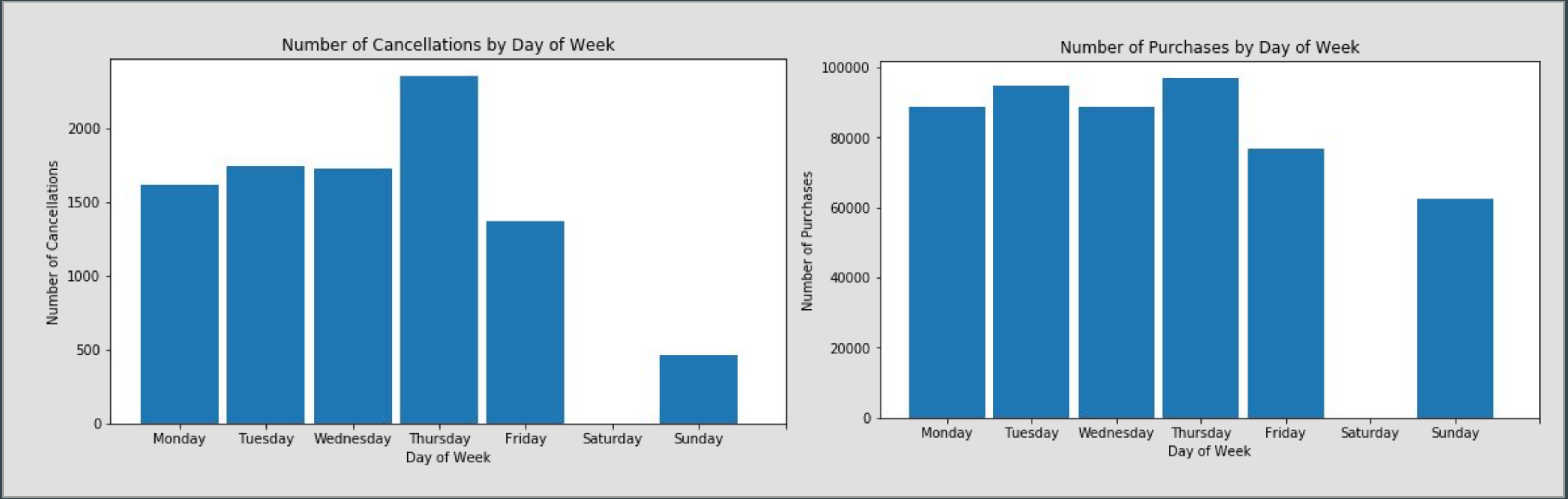
The peak in the previous slide corresponds to the time frame between 2011-09-25 to 2011-10-25.

The holidays during that time (chart to the right) don’t correspond to gift giving.

I also checked for holidays in the 30 days prior with similar results; no holidays associated with gift giving (or returning)

Date	Day of Week	Name	Type
2011-09-28	Wednesday	Navaratri	Hindu Holiday
2011-09-29	Thursday	Rosh Hashana	Jewish holiday
2011-10-04	Tuesday	Feast of St Francis of Assisi	Christian
2011-10-06	Thursday	Dussehra	Hindu Holiday
2011-10-08	Saturday	Yom Kippur	Jewish holiday
2011-10-13	Thursday	First day of Sukkot	Jewish holiday
2011-10-19	Wednesday	Hoshana Rabbah	Jewish holiday
2011-10-20	Thursday	Shemini Atzeret	Jewish holiday
2011-10-21	Friday	Simchat Torah	Jewish holiday

Cancelled and purchased products by day of week



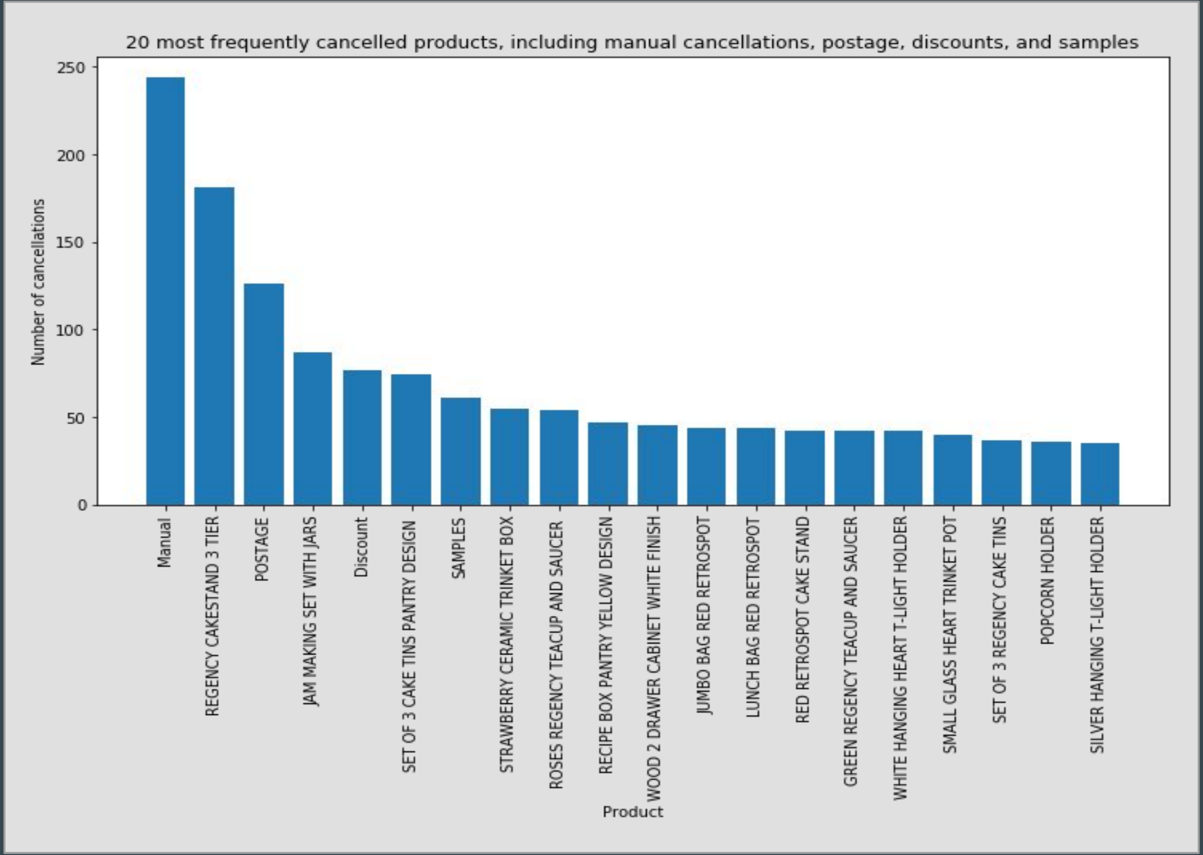
There were no products cancelled or purchased on Saturdays

20 most commonly cancelled products

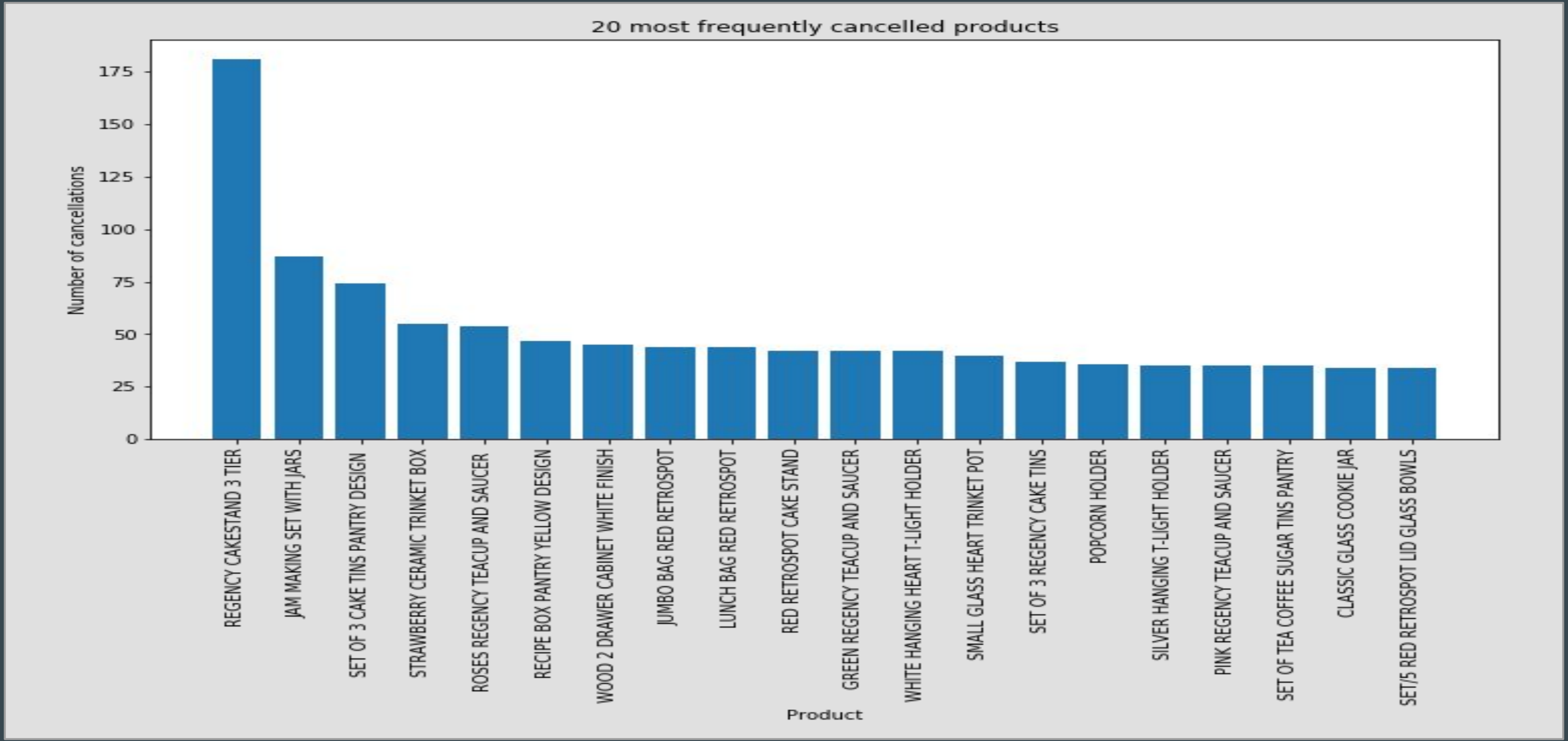
I noticed that the top 20 cancelled products included “Manual”, postage, discounts, and samples.

This made me wonder what “Manual” was. Hand entered cancellation? Override of price?

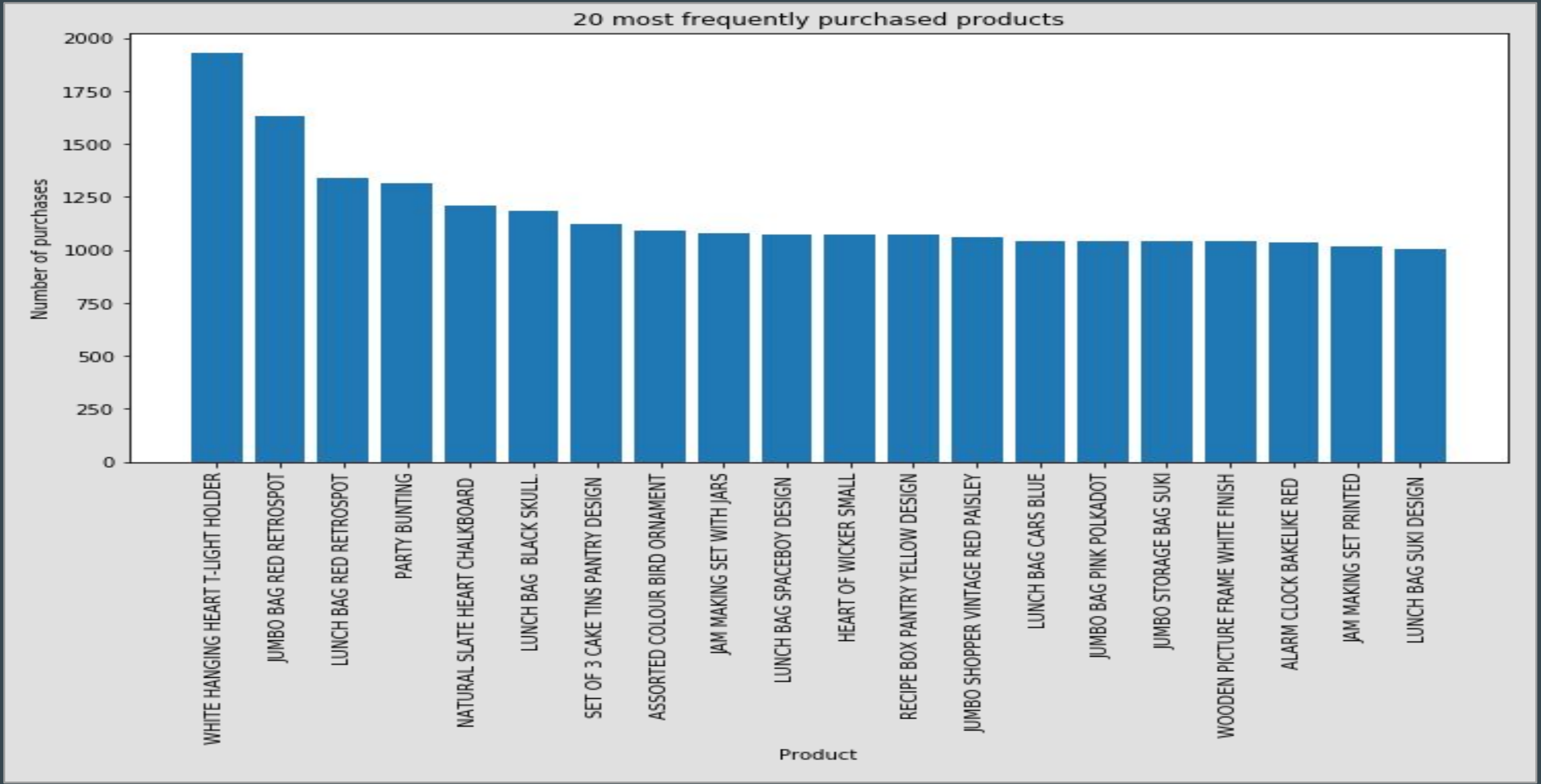
I would ask this question of the company to get better understanding.



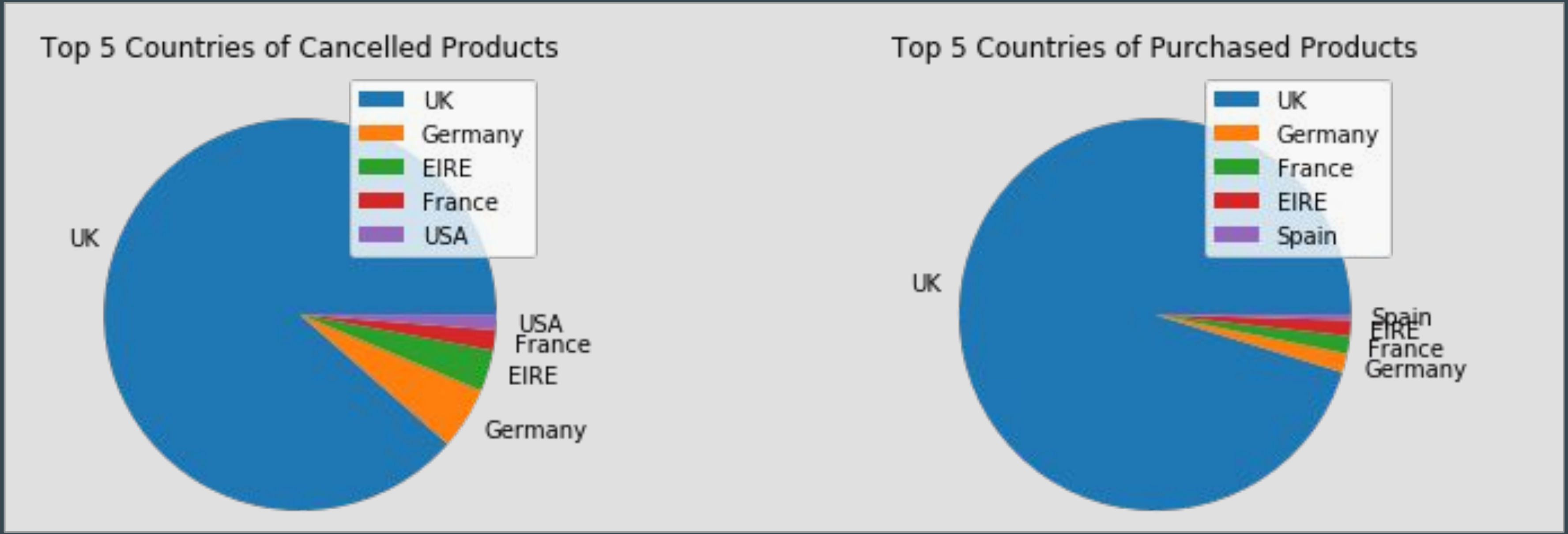
Top 20 cancelled products, without postage, discounts, manual, and samples



Top 20 purchased products



Which countries are canceling and purchasing the most?



Top 5 Countries of Cancelled Products

Top 5 Countries of Purchased Products

Since this is a UK based company, it’s not surprising that the UK has the most transactions on both sides. In terms of the top 5, the USA is a more frequent canceller whereas Spain is more commonly a purchaser.

Products in both top 20 cancelled and purchased

Description	Number cancelled	Number Purchased
JAM MAKING SET WITH JARS	87	1076
SET OF 3 CAKE TINS PANTRY DESIGN	74	1125
RECIPE BOX PANTRY YELLOW DESIGN	47	1070
JUMBO BAG RED RETROSPOT	44	1632
LUNCH BAG RED RETROSPOT	44	1340
WHITE HANGING HEART T-LIGHT HOLDER	42	1929

Conclusions and ideas for moving forward

Conclusions

- For this business, by far a majority of their cancellations are “manual”.
  - This may warrant a more in-depth look. What does manual mean and what are the implications for their system?
- The most commonly cancelled product is a 3-tier Regency Cakestand
  - Is there a common reason for this?
  - Possibly a product to consider removing from the inventory?
- For purchased products, almost half of the top 20 are bags of some type
  - May be a direction to expand sales and products
- More cancellations happen on Thursdays, and the fewest purchases occur on Fridays and Saturdays
  - Consider a sale to increase purchases on these days and possibly offset cancellations?
- Attempting to predict the next month of cancelled or purchased products gives very broad trends with wide ranges of possible outcomes
- 6 products made the top 20 list for purchased and cancelled. Of those, the Jam making set with jars had 7% of its sales cancelled.
  - May be another area to investigate.

Ideas for moving forward

- This analysis would benefit from more data. Even another year to make comparisons between the two would be helpful.
- Looking at a different type of machine learning, like a linear regression to look at trends, may be more informative for the company.
- Taking a deeper dive into a specific product or group of products (example: all the bags) may yield more meaningful insights.
- Looking at other compounding factors (are there multiple stores? Are there other economic things happening in this same time frame, like a recession, that may be driving data?, does the weather have an impact?) to attempt to better explain the data.