Springboard Data Science Career Track

Capstone Project #2: Milestone Report

Sales Prediction
for an
Online Retail Store

Submitted By:
Koshika Agrawal

May 20, 2019

Table of Contents

1. Introduction

1.1. Problem Statement

Predicting sales is one of the most important business problems for any retail entity. If a business can predict the how much of each item it will sell in each month, it can manage its inventory better. Sales predictions also help in directing the marketing efforts in right direction to increase the chances of sale.

1.2. Objective

To forecast item-wise sales for an online retail store

1.3. Client

The client in this case is a UK-based and registered non-store online retail.The company mainly sells unique all-occasion gifts. Many customers of the company are wholesalers.

This model can be replicated to any similar online or physical store selling any kind of product.

2. Data Acquisition and Wrangling

2.1 Data Acquisition

The dataset has been taken from UCI Machine Learning Repository at:
https://archive.ics.uci.edu/ml/datasets/Online%20Retail

The dataset contains following attributes:

Attribute name	Type	Description
InvoiceNo	Nominal	A 6-digit integral number uniquely assigned to each transaction. If this code starts with letter 'C', it indicates a cancellation
StockCode	Nominal	A 5-digit integral number uniquely assigned to each distinct product
Description	Nominal	Product (item) name

Quantity	Numeric	The quantities of each product (item) per transaction
InvoiceDate	Numeric	The day and time when each transaction was generated
UnitPrice	Numeric	Product price per unit in sterling
CustomerID	Nominal	A 5-digit integral number uniquely assigned to each customer
Country	Nominal	Name of the country where each customer resides

2.2. Data Wrangling

Initial loading and inspection of datasets exposed some challenges in it for our study. We wrangled the data to make it fit for our analysis.

1. Description' and 'CustomerID' columns have null values
    - Each InvoiceNo should be linked to a single CustomerID. So we tried using InvoiceNo and CustomerId linkage to fill missing values. As we could not find the linkage, we dropped the null values
    - Deleting missing CustomerId removed all missing Description rows too.
2. StockCode does not identify a unique Description
    - Each StockCode should uniquely represent an item Description. But the original dataset has multiple Description for same StockCode. This is because there are data entry errors in the description as shown below

	StockCode	Description
48	20622	VIP PASSPORT COVER
49	20622	VIPPASSPORT COVER
101	20725	LUNCH BAG RED RETROSPOT
102	20725	LUNCH BAG RED SPOTTY
194	20914	SET/5 RED RETROSPOT LID GLASS BOWLS
195	20914	SET/5 RED SPOTTY LID GLASS BOWLS

The data was wrangled to contain one to many mapping between StockCode and Description.

3. Some descriptions contain incidental charges like postage/shipping charges, discounts etc.
    - As these charges are not related to our analysis, we dropped these observations

4. Some of the item quantities are negative
    - These are cancelled orders

5. Some CustomerID linked with 2 countries

    As per the data attribute description: 'Country' column is the name of the country where each customer resides. But we don't have any information on how is this data being captured. Is it through IP address of the country while creating account, or may be based on the shipping address, or may be something else.

    Logically, each CustomerID should be linked to one country only. The reason for having more than one country could be:

    a. Data entry error
    b. Customer has moved to another country, and has got the address changed in his account
    c. In case this attribute reflects the shipping address, the customer has shipped the order to an address different from his own.
    d. In case this attribute is captured through the IP address while ordering, the customer might be ordering while travelling to another country.

    Further analysis of data does not make it clear what is the reason behind 2 countries for a CustomerID, so for now, we are not making any changes in the CustomerID and country linkage.

6. Different unit price of same item for different transactions

    Unit price of an item keeps changing for different transactions. This poses problem while aggregating the data. We added another column for total price (which is quantity multiplied by unitPrice). While aggregating we added the quantity and total price, and find the unit price from the aggregated values.

2.3. Derived Features

We derived following features from the already existing ones to aid in our analysis.

1. CancelledOrder containing boolean values, 1 if order was cancelled, 0 otherwise
2. InternationalOrders containing boolean values, 0 if order came from UK, 1 if the order came from outside UK
3. TotalPrice containing float values = unitPrice * Quantity

2.4. Data Wrangling Summary

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
0	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	2010-12-01 08:26:00	2.55	17850.0	United Kingdom
1	536365	71053	WHITE METAL LANTERN	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom
2	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	2010-12-01 08:26:00	2.75	17850.0	United Kingdom
3	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom
4	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom

Shape: (541909,8)

	InvoiceNo	StockCode	Quantity	InvoiceDate	UnitPrice	CustomerID	Country	Description	CancelledOrder	InternationalOrders	TotalPrice
0	536365	85123A	6	2010-12-01 08:26:00	2.55	C17850	United Kingdom	CREAM HANGING HEART T-LIGHT HOLDER	0	0	15.30
1	536365	71053	6	2010-12-01 08:26:00	3.39	C17850	United Kingdom	WHITE METAL LANTERN	0	0	20.34
2	536365	84406B	8	2010-12-01 08:26:00	2.75	C17850	United Kingdom	CREAM CUPID HEARTS COAT HANGER	0	0	22.00
3	536365	84029G	6	2010-12-01 08:26:00	3.39	C17850	United Kingdom	KNITTED UNION FLAG HOT WATER BOTTLE	0	0	20.34
4	536365	84029E	6	2010-12-01 08:26:00	3.39	C17850	United Kingdom	RED WOOLLY HOTTIE WHITE HEART.	0	0	20.34

Shape: (404618,11)

The wrangled dataset is saved to a csv file.

The code for data acquisition and wrangling can be accessed in this ipython notebook.

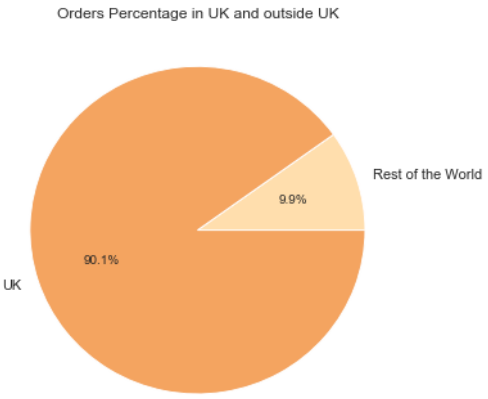
3. Data Exploration

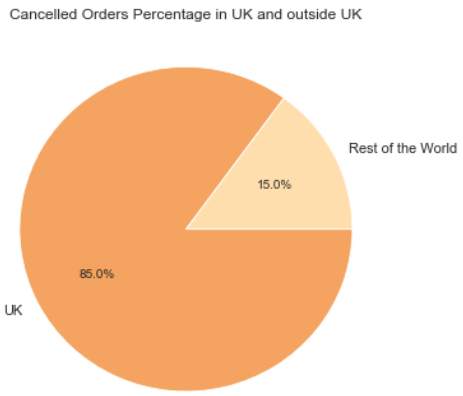
The code for EDA can be found in this ipython notebook.

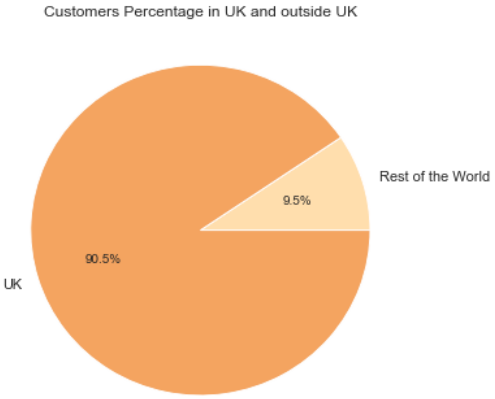
3.1. Worldwide distributions

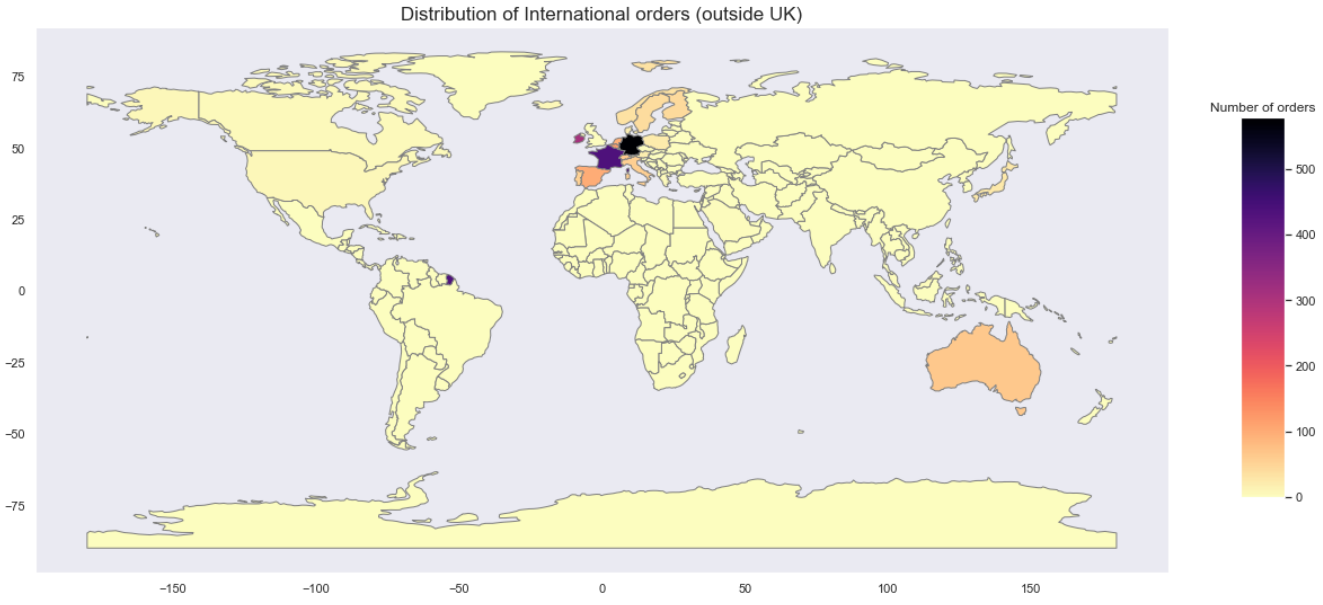
The retailer has its customers all over the world. We will take a look at following distributions on a world map:

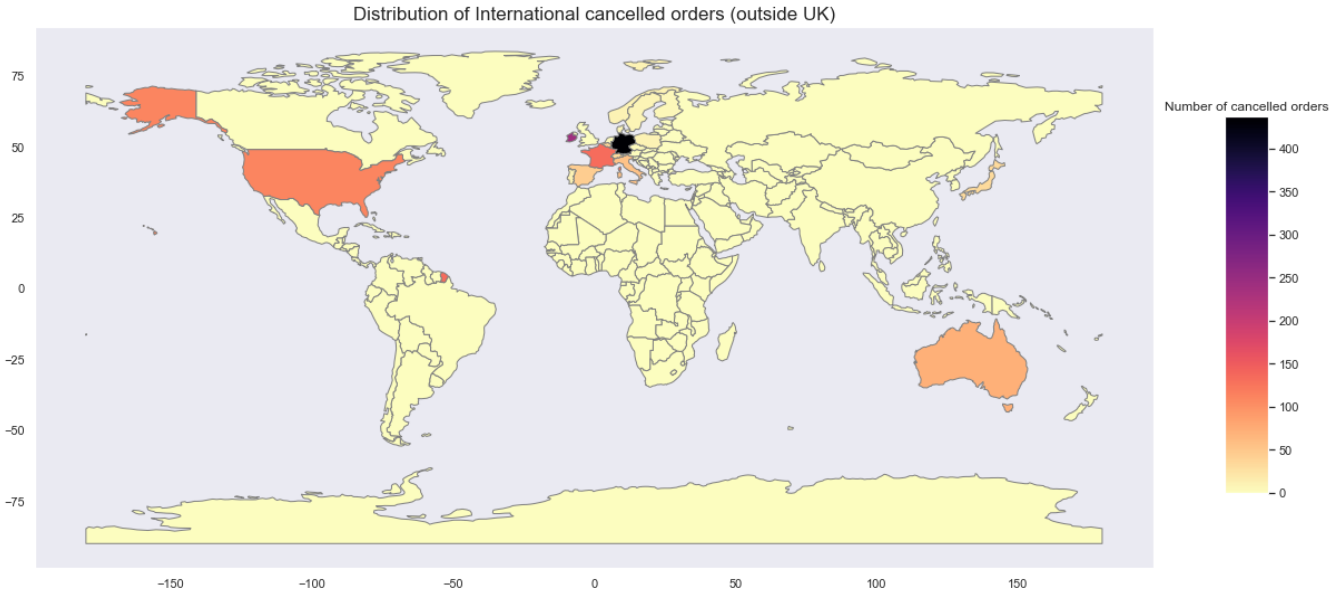
a. Orders
b. Customers
c. Cancelled orders

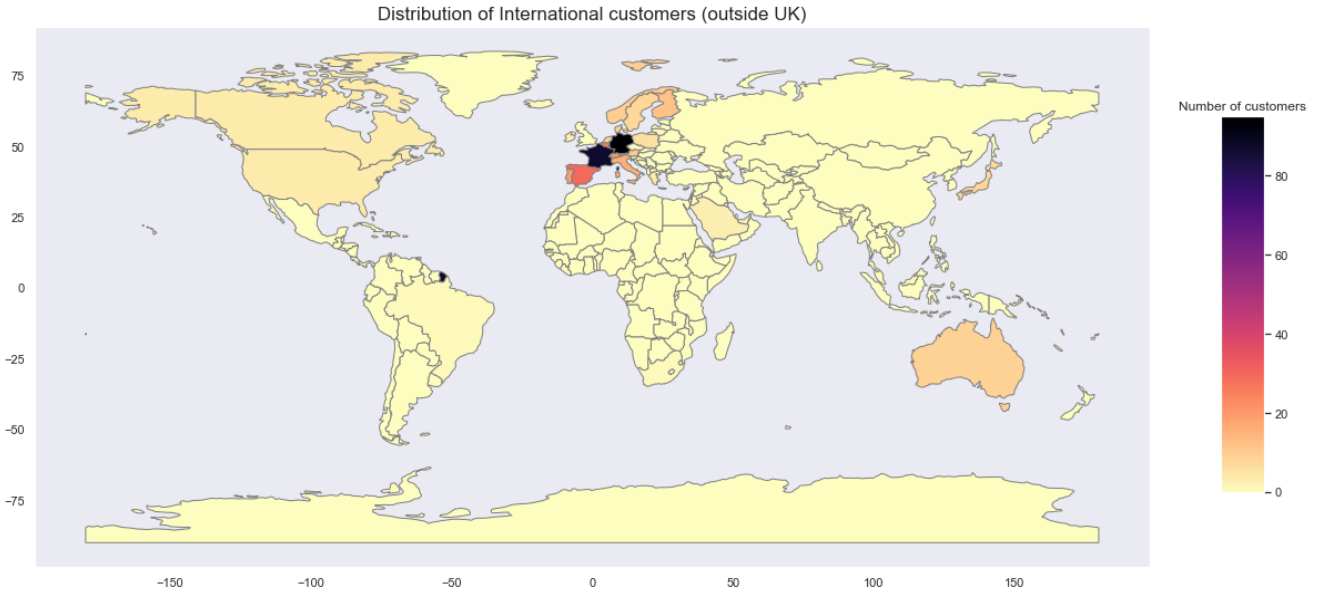

Orders Percentage in UK and outside UK
Rest of the World
9.9%
90.1%
UK


Cancelled Orders Percentage in UK and outside UK
Rest of the World
15.0%
85.0%
UK


Customers Percentage in UK and outside UK
Rest of the World
9.5%
90.5%
UK


Distribution of International orders (outside UK)
75
50
25
0
−25
−50
−75
−150
−100
−50
0
50
100
150
Number of orders
500
400
300
200
100
0

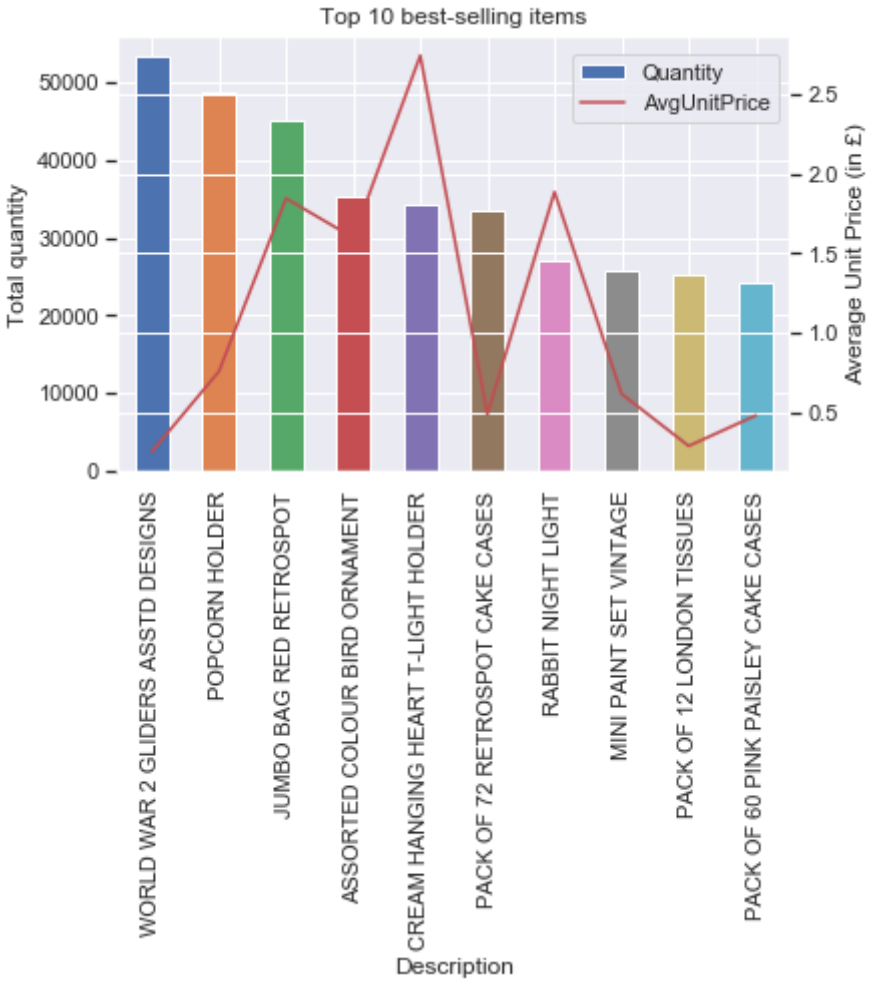


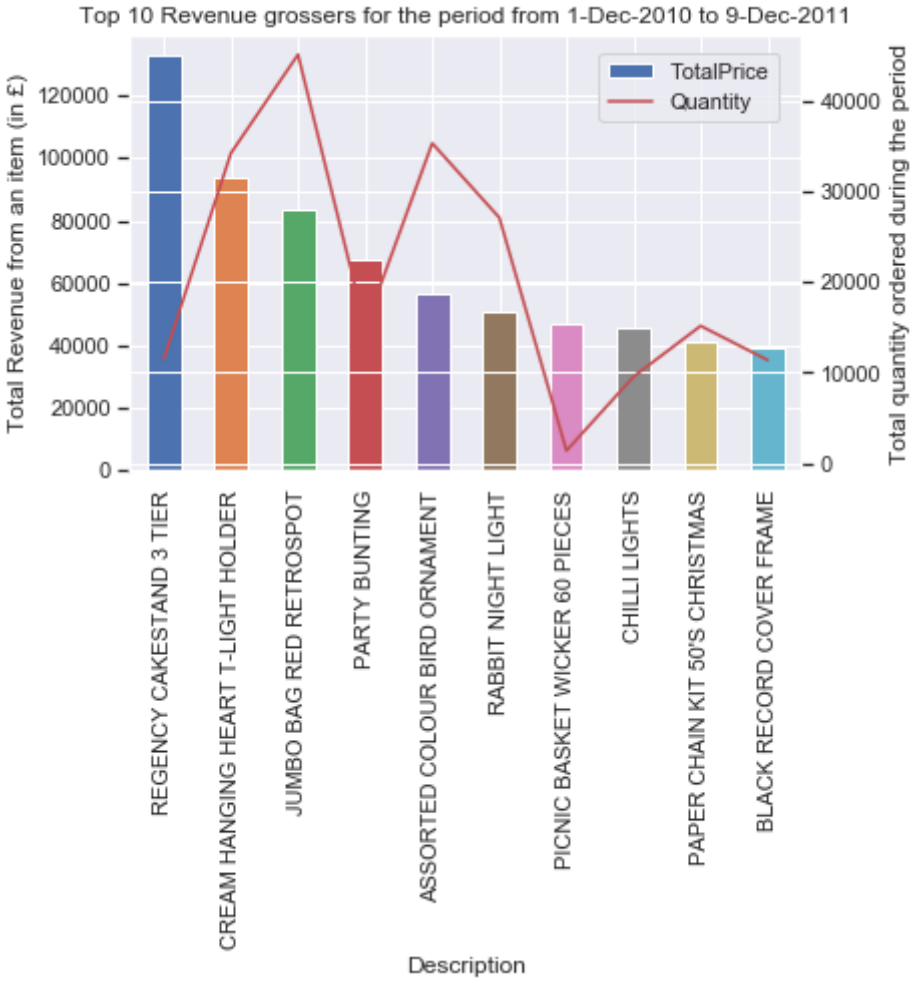


Observations

1. 90% sales comes from UK and 90% customers also are from UK
2. Outside UK, most of the sales is from Europe
3. In international sales, Germany, France and Ireland are among the highest
4. Outside UK, most customers are from Germany, France and Spain
5. Outside UK, Germany, Ireland, France and US show highest number of cancelled orders
6. Outside Europe, highest sales comes from Australia, while the highest number of cancellations come from United States.

3.2. Best Sellers





Observations

1. The best selling products vary in their average unit price. So it doesn't seem to have any relation with its price.
2. The No. 1 best selling product sells almost double the quantity of the 10th best seller.
3. The no. 1 revenue grosser leads its immediate follower by 26%.

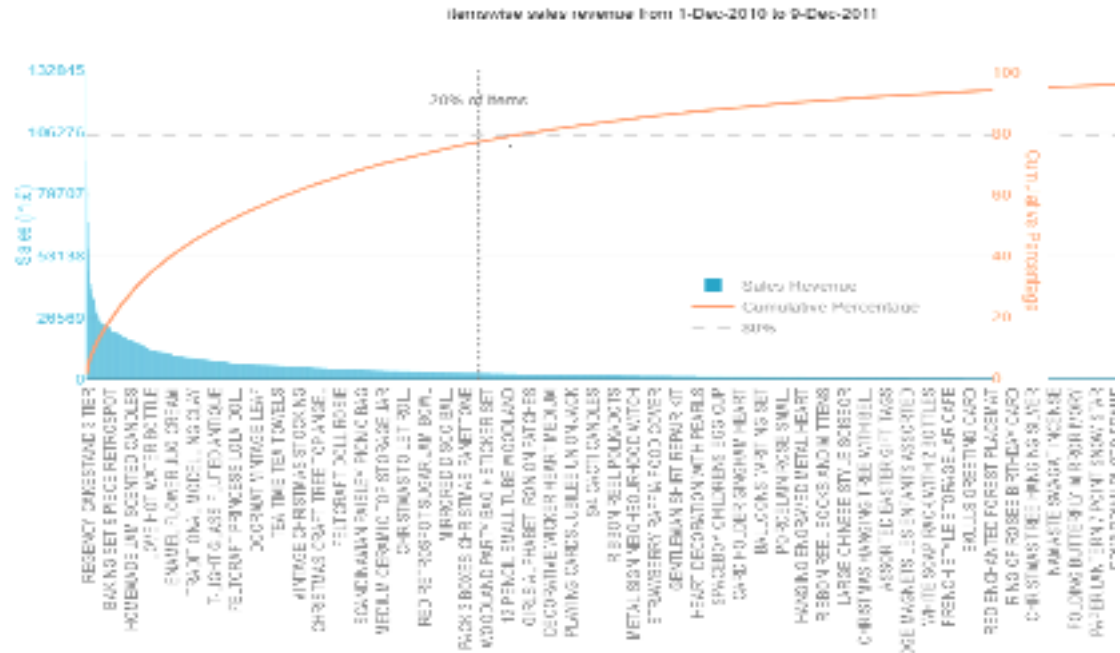
3.3 Pareto Principle (80-20 rule)

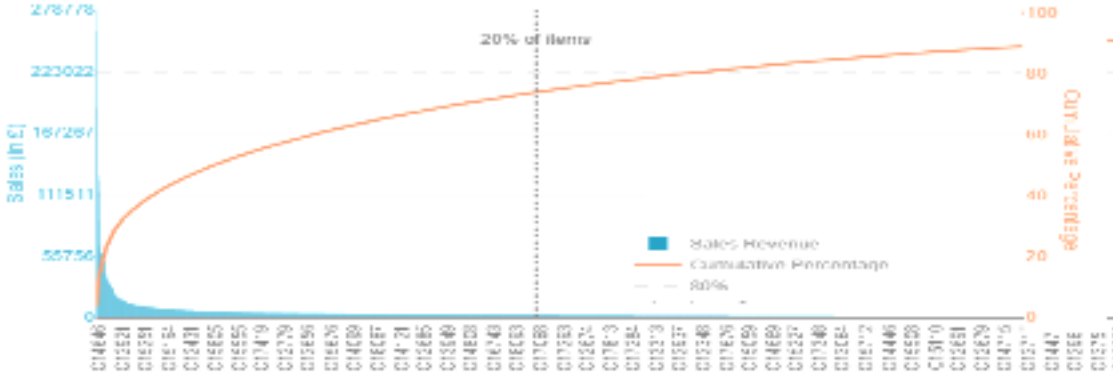
The Pareto principle (also known as the 80/20 rule) states that, for many events, roughly 80% of the effects come from 20% of the causes.(Source: wikipedia)

For a sales entity, Pareto principle could suggest that 80% sales of a company comes from 20% of its products and/or 80% of its sales comes from 20% of its customers.

Our online retail store has 3652 unique items for sale, and 4357 unique customers for the period from Dec 1 2010 to Dec 9 2011. Here, Pareto principle is of value because instead of focussing on such a huge number of items and customers, the company can just focus on 20% of these in order to effect 80% of its sales.

We can see the applicability of Pareto principle to our dataset.

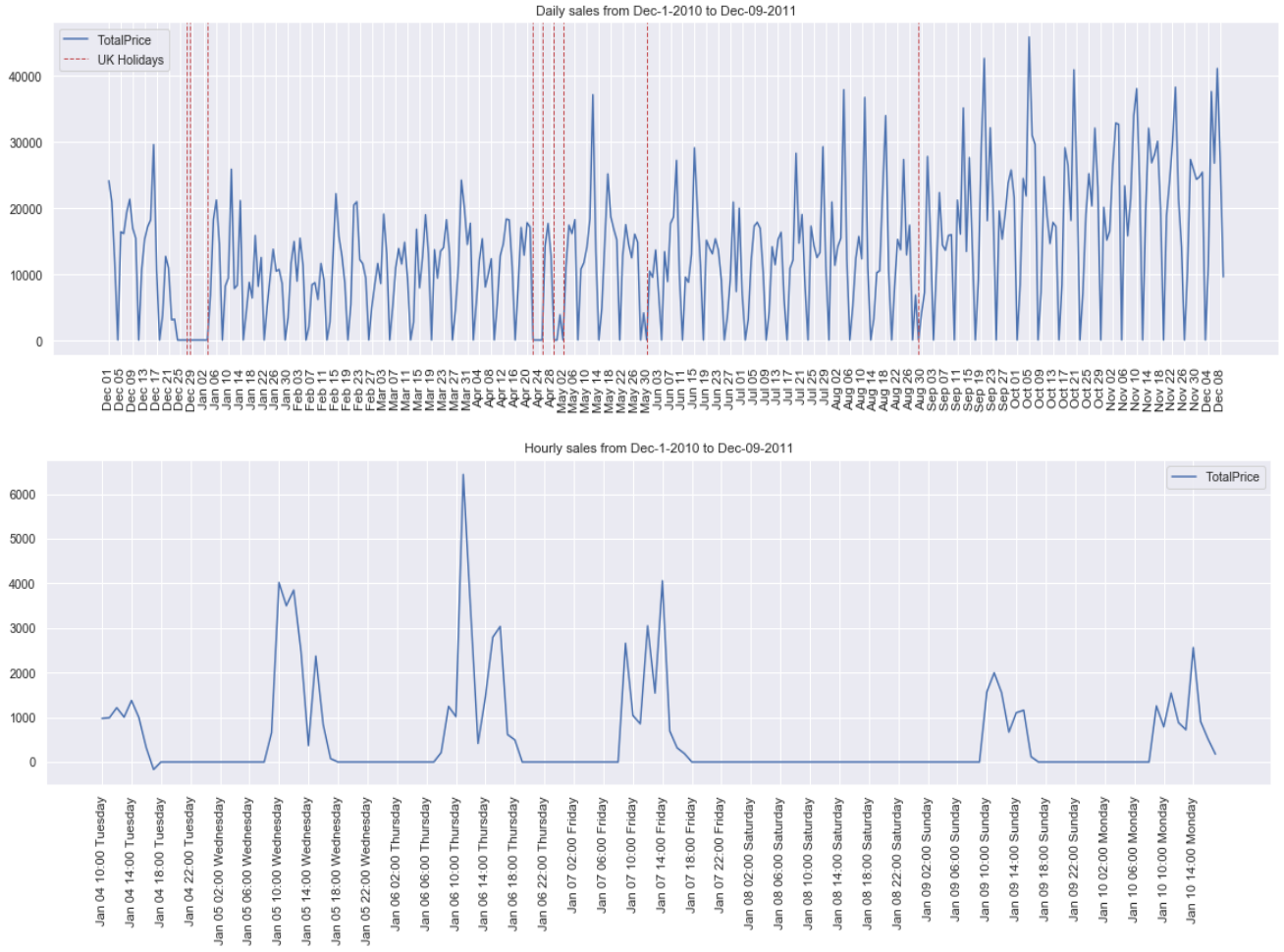




Observations

1. Pareto principle holds true for items in our dataset, as 22% of all the items are contributing to 80% of the sales revenue. 22% of all items means 803 items.
2. 27% of all the customers are contributing to 80% of the sales revenue. 27% of all items means 1176 items

3.4 Time series plots



Observations

1. There is no sales happening on Saturdays.
2. Sales falls down during the holidays
3. There is zero sales happening on Boxing day (Dec 26th) which probably means that the company does not offer any promotions during the holidays/special days.
4. Sales happens only during the working hours - between 8am and 6pm.
5. There is no particular trend seen in monthly and weekly sales. The overall sales has increased through the months.

4. Scope of further study

1. Predict customer churn and suggest ways to prevent the churn
2. Customer segmentation and buying behavior analysis