

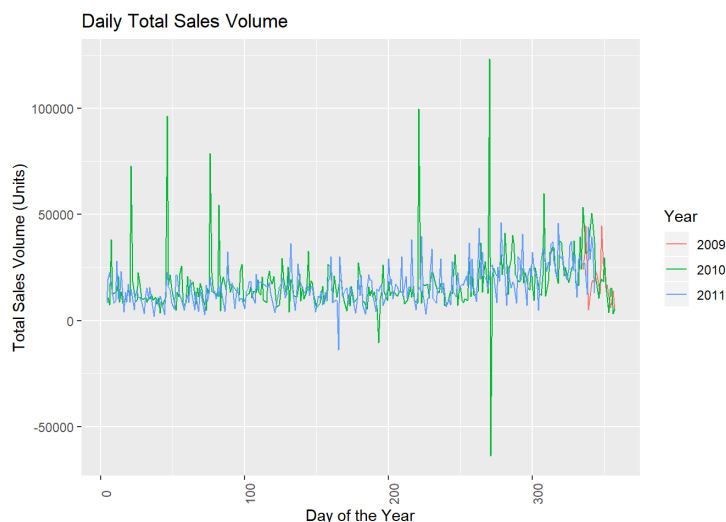
Introduction

This report details the modelling process and results found from conducting initial data exploration and fitting a simple time-series model to demand retail data from an online dataset. The dataset contains transaction data for a UK-based online retailer. Initial investigation was centred around finding patterns in the sales volume as well as patterns in product and consumer types. A range of ARIMA and seasonal time-series were considered, more robust modelling should be conducted in the future considering effects of non-temporal data to increase the model accuracy. Note all code, output and plots can be found in the Code and Plots pdf.

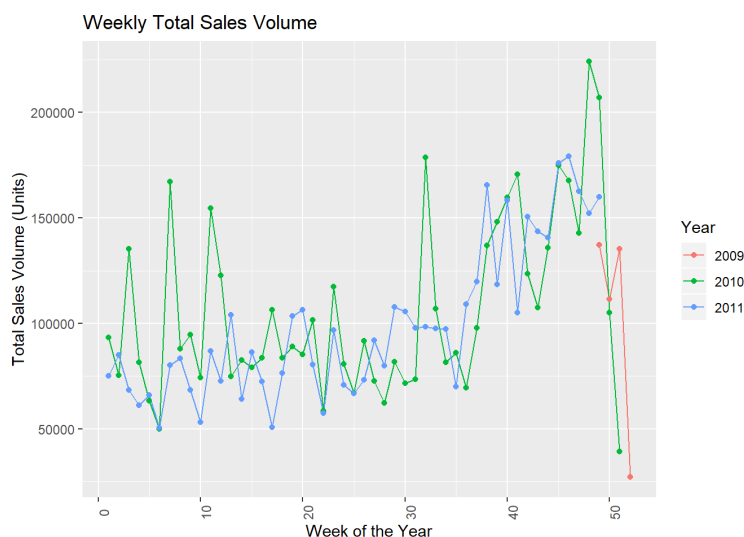
Data Exploration

After removing the entries with missing descriptions (these often corresponded to having missing quantities and IDs too so would likely cause errors in analysis) the sales volume and revenue data was analysed for any obvious trends.

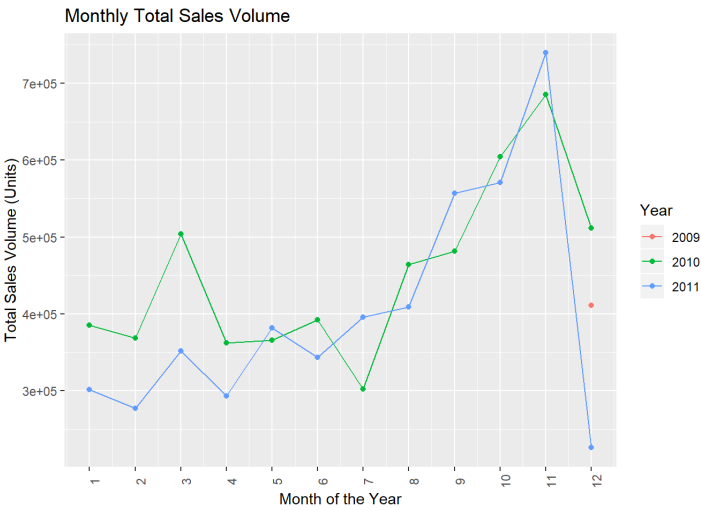
Looking at the daily total sales volumes, we can see that each of the years have similar levels of noise throughout the year, with the large spikes due to random invoices such as large postage fees, bad debts or bank costs (essentially outliers). The large spike and fall in 2010 corresponds to 2 large orders that were cancelled the next day (hence the near equal rise and fall).



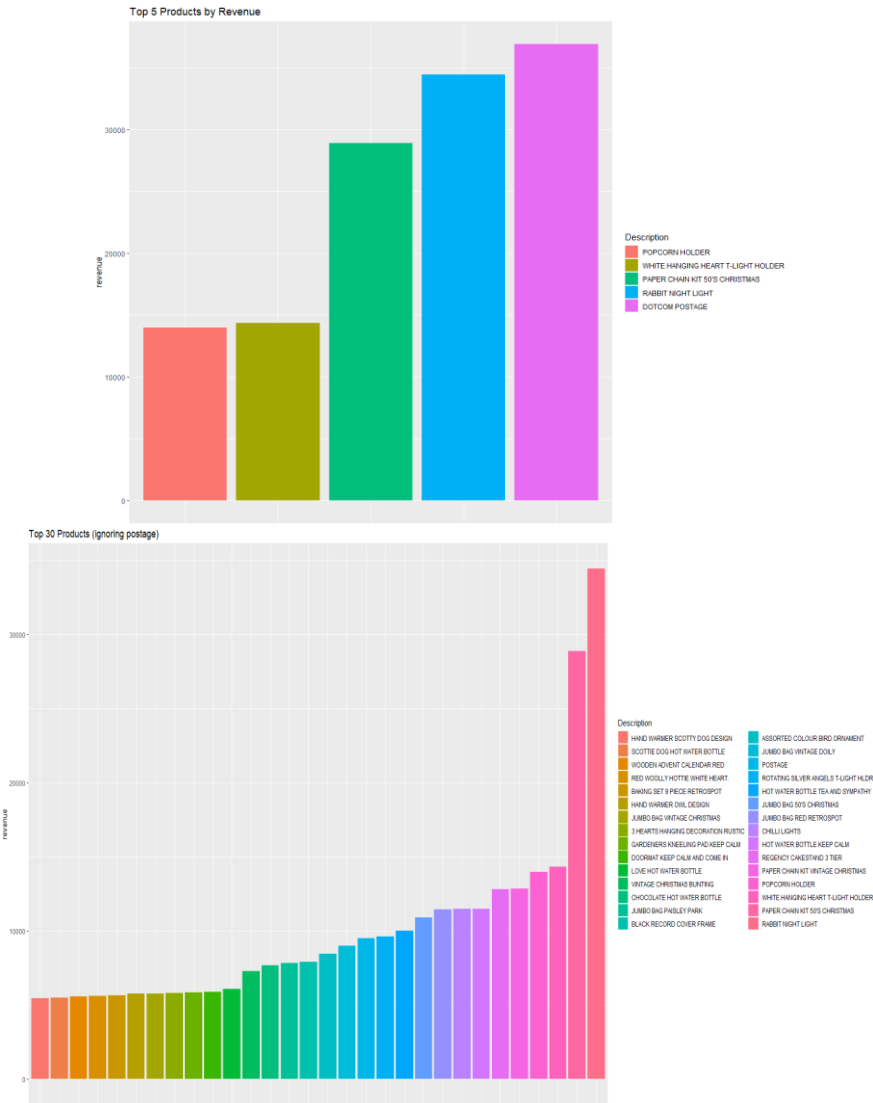
The weekly sales plot (on the following page) shows clearer trends than the daily data. It indicates that for most of the year, the weekly volume sold has approximately the same mean, with a seasonal increase occurring around the 48-week mark in 2010 and 2011 (likely the Christmas season). This is potentially due to the distribution being stationary with a small seasonal peak at the above time. It also shows that the sales data is reasonably constant across the 3 years, with 2010 seemingly having the highest average weekly sales volume but also the greatest variation.



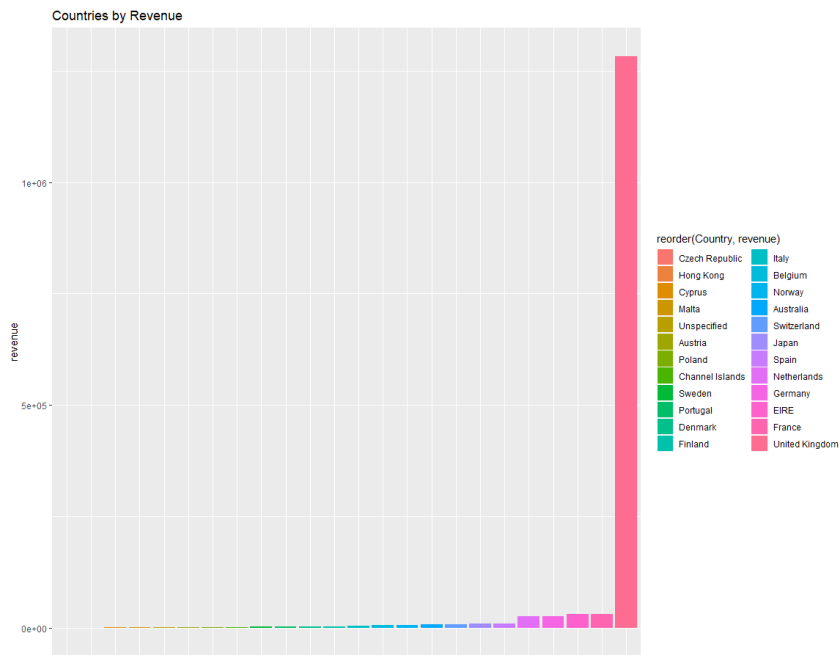
The monthly sales volumes below confirm the trends identified in the weekly sales. However, it indicates that the seasonal component may be larger than evaluated above, starting in about September or October. The large drop in December 2011 is due to the data collection period ending well before the end of the month.



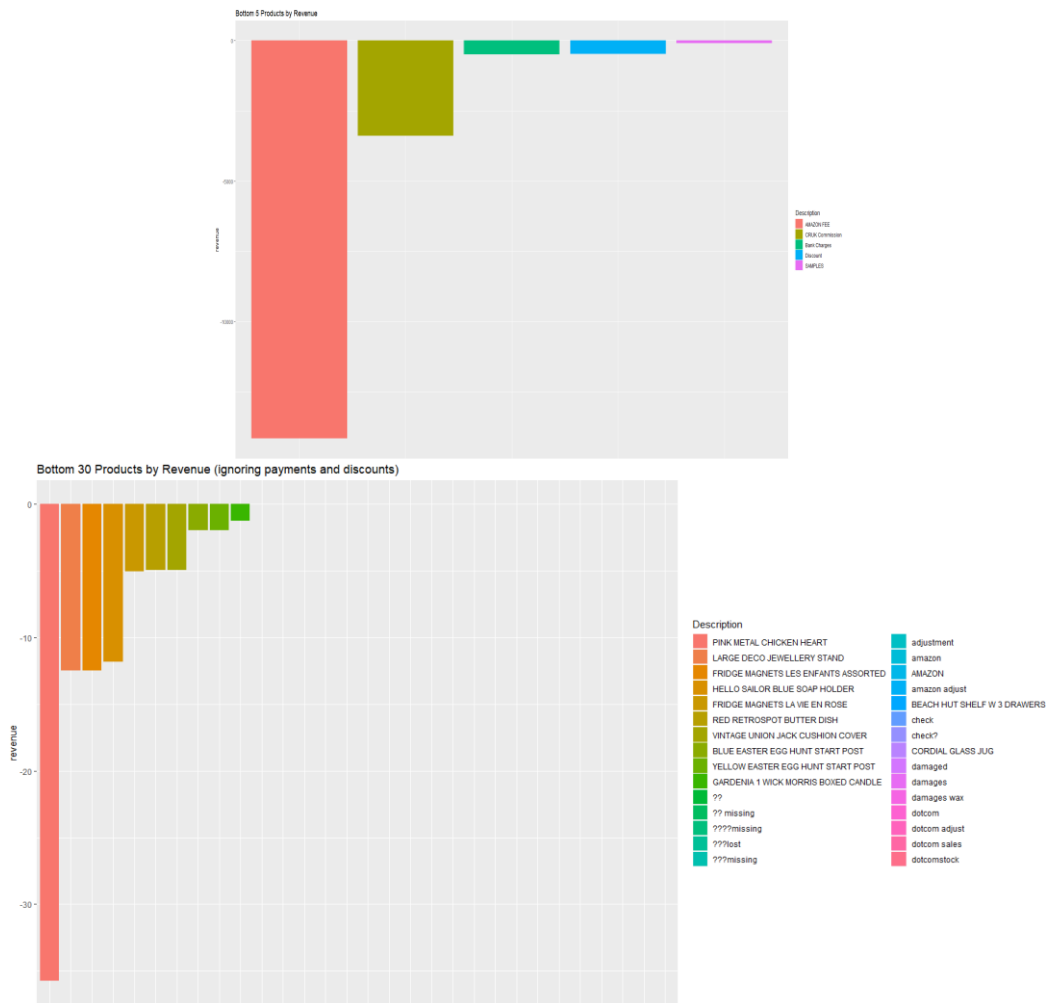
Looking at the revenue data for the last full month (November 2011) we can see that the largest revenue contributor is DOTCOM postage (note that variations in this description exist but are not included). Plotting the top 30 other contributors we can see that the primary earners are Christmas or winter-themed items. This coincides with the seasonal Christmas increase in shopping and is because the majority of the retailer’s clients are from the Northern Hemisphere, which is in winter during this month. There are only 11 product (descriptions) with monthly revenue over £20,000.



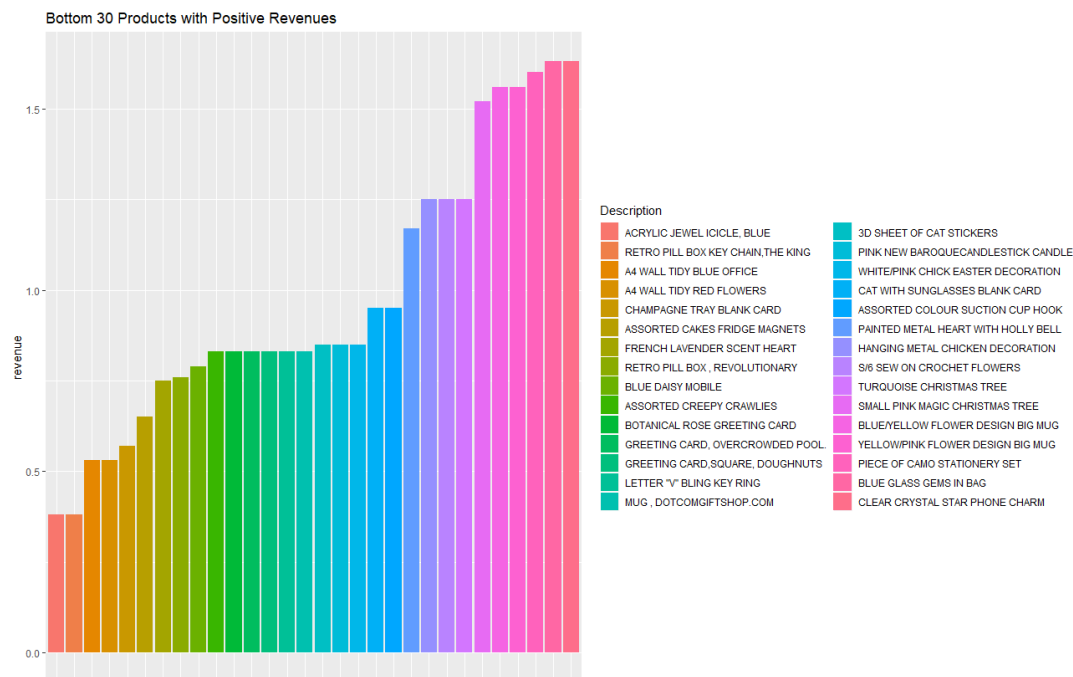
Looking at the revenue data for each country, we can see that the primary customers are based in the Northern Hemisphere (specifically in Europe). The retailer's home region of the United Kingdom earns at least 5 magnitudes more revenue than the next highest country. This may be because the business is more well-known locally, and/or locals may pay less postage and packaging fees than overseas customers.



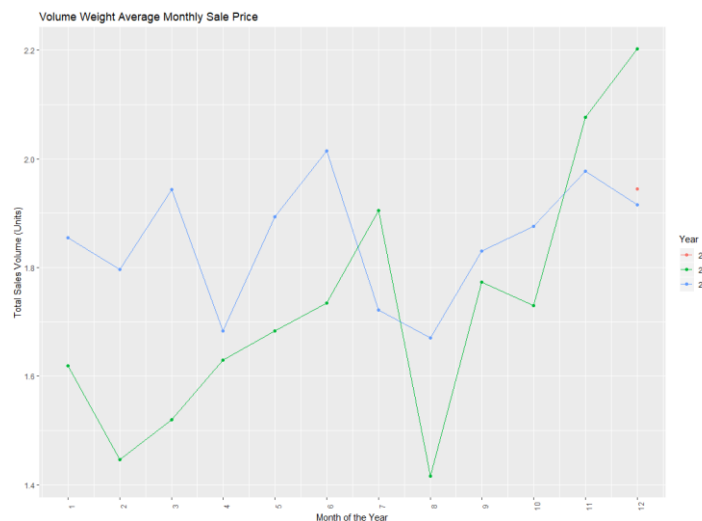
Looking at the lowest-earning products, we can see that business expenses and discounts make up the bulk of the business costs. The next 30 bottom products show that the business suffers from some human or shipping errors as indicated by the large number of different missing and damaged descriptions recorded. It is also worth noting that there were few returns during the month, costing a total of just over £100.



Looking at the bottom 30 products with positive revenues, all of them earned the company less than £2 in the last month. Most of them are also small-ticket items, so this low revenue directly correlates to a lower sales volume for these items. It is worth noting that some of these are more associated with warmer climates, and so are unlikely to be purchased during the latter half of the year in the Northern Hemisphere.



The VWAM plot shows a similar trend to the monthly sales volume, with a peak in the later months and a random noise like pattern throughout the rest of the year. Interestingly, although 2010 had similar or higher sale volumes than 2011, it has a lower VWAM with far more variability, indicating that it was primarily large-volume, low-cost goods sold in comparison to higher price per unit goods sold in 2011.



Missing Cancellation Data

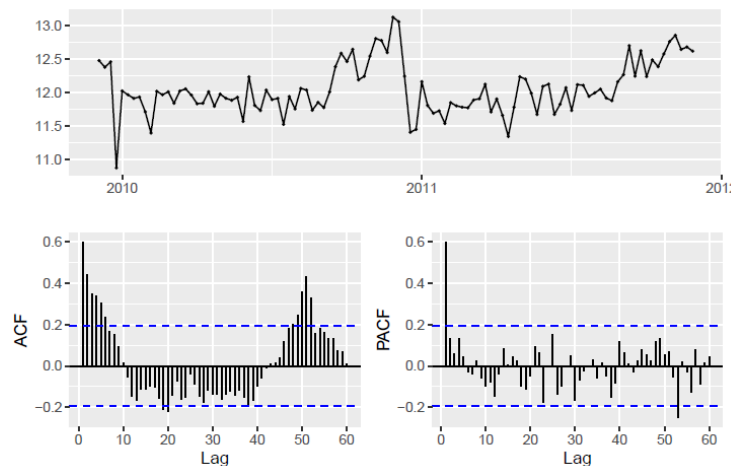
Upon examination of the returned/cancelled invoice data, it is apparent that some of the cancellations relate to invoices that were filed before the data collection period. This would lead to analysis errors (as they do not 'cancel' out any positive revenue) and so should be removed. It was noted that all cancellations from 2010 onwards seemingly match another invoice in the collection period, so only the 2009 cancellation data was inspected for errors. The customer IDs of these cancellations were compared to the rest of the 2009 invoices and if they did not match the ID of a previous purchase, they related to invoices from before the data period and were thus removed. The cancellations in the first 4 days of December (proper cancellations started around this time) with missing IDs were also removed for the above reasons.

Time Series Methods

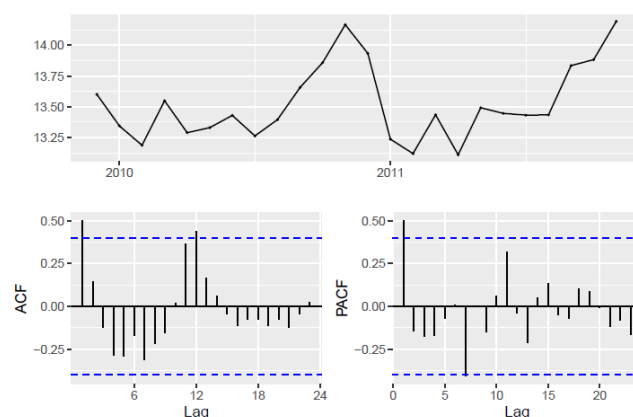
Our final task was considering the likely revenue for the latest month (12/2011) to help the owner plan for future personal expenditure. Multiple methods were considered for this:

- A two-way ANOVA was considered between the different years and months to consider their effects on sales. Revenue data would be aggregated by month and year, and then the difference in means considered for each of these. The Tukey intervals would be used to quantify temporal effects (considered as factors), with the final prediction model built around the p-values of these intervals. This approach was not used as there is only a small sample for each month factor, and the random noise in most of the year is likely to cause too many overlapping intervals for useful prediction. Note that other forms of multi-linear regression were also considered, but due to a lack of other data (such as global demand or local GDPs) these approaches were also not taken.
- The final model considered, and the one that was used, was to apply time-series analysis in the form of fitting HoltWinters and SARIMA to the data. Revenue data could be aggregated by week or month and using the AIC values of each model to test for goodness of fit before forecasts are found. This model was used as it is likely to have the best trade-off between accuracy and time required as well as allowing for the best manual adjustment for the seasonal trends.

The revenue data was converted into two time-series plots for analysis, one by week and one by month. As we are looking at monetary data, and because the variation in revenue seems to increase with time, the revenue was logged. Looking at the weekly plots, there seems to be spikes at the end of 2010 and 2011. Looking at the ACF and PACF plots to check for lag correlations in the data, we can see spikes in the ACF and PACF plots at around lag 52 (corresponding to a seasonal peak). The ACF and PACF also show a spike at lag 1, indicating a non-seasonal lag in the trend too.



Now looking at the monthly time series plots, the seasonal increase at the end of the year seems more obvious, especially at the end of 2010. Although the ACF and PACF do not have as large a spike due to this potential seasonality as the weekly data, the plot seems to suggest otherwise. There is also an initial spike at lag 1 indicating a non-seasonal component as well.



A $(1,0,0)(0,0,1)$ SARIMA model was fitted to the weekly data to account for the lag 1 and seasonal trend. MLE fitting found that the seasonality component was at lag 52 (i.e. final model is $(1,0,0)(0,0,1)[52]$). Similarly, to check for accuracy in predictions, a $(1,0,0)(0,0,1)$ model was fitted to the monthly data, which had seasonal lag 12, $(1,0,0)(0,0,1)[12]$, using MLE fitting. Checking the residuals for normality and low-auto-correlation for both models (to check prediction suitability), we see that for both models the Ljung-Box test suggests little correlation in residuals and that the residuals are approximately normal.



Thus, we proceeded to forecast the following next few months of revenue. The monthly model suggests an expected revenue of between £790,000 – 1,620,000 at the 95th percentile, with the weekly model having a lower range of about £480,000 – 1,600,000. This difference is likely due to the difference in aggregating over a longer period; calculating the month's data involved combining multiple weekly confidence intervals which likely has confounding uncertainties. It is worth noting that an auto-fitted HoltWinters model has a much narrower interval of about £1,050,000 – 1,350,000. We conclude that it is likely that the expected revenue is close to £1,000,000 and so the retail owner is likely to be able to purchase his wife a new sports car.

Dynamic regression models could be considered in the future to model the effect that other factors, such as GDP, as well as considering splitting the time-series down into smaller components (such as by country) to see if more insights can be found. The other methods outlined above may also provide further insights.

