



Nanyang Business School
BC2407 Analytics II: Advanced Predictive Techniques
AY 20/21 Semester 2

Project Report
Seminar Group 2 Team 8
Instructor: Chew Chee Hua, Neumann

Colin Tan G-Hao (U1922153D)

Joshua Toh Sheng Jie (U1921471F)

Tsoi Kin Yi (U1910835C)

Choo Yun Sheng Ryan (U1910399C)

Liu Jia Yang (U1910152C)

Table of Contents

Executive Summary	3
Introduction	4
Business Problem	5
Measuring Success Of Business Outcomes	5
1. Sales Conversion Rate	6
2. Average Order Value	6
3. Customer Satisfaction & Returning Rate	6
4. Customer Lifetime Value	7
The Usage of Analytics in E-commerce	8
1. Customers' purchasing patterns and sales forecast	8
2. Improving Inventory Management	9
3. Other Applications of Analytics in the e-commerce Industry	9
Data Exploration	10
1. Most Purchased Items	11
2. Sales Distribution throughout the different countries	11
3. Sales distribution throughout a week/month timeframe	13
Building Our Prediction Model	13
1. Improving Inventory Management - Time Series Forecasting	14
1.1. Exploration of historical time series	15
1.2. Time-series forecasting	16
1.2.1. Holt-Winters Exponential Smoothing	16
1.2.2. Seasonal ARIMA Method	17
1.3. Model Evaluation	17
1.4. Shortfalls	18
2. Predicting Customers Purchasing Pattern - Association Rules	19
2.1. Results	20
2.2. Exploration and conclusion of the rules	20
2.3. Shortfall	21
Conclusion	21
Appendix	22
Time Series	22

Executive Summary

This report primarily focuses on the E-commerce industry and the usage of analytics behind future sales prediction. In the long run, the main aim is to reduce losses and increase the E-commerce business's profits. This can be done through strategic pricing based on the accurate predictions of future sales and identifying patterns and trends associated with customers' purchases.

Our report starts with a brief introduction to the e-commerce industry and defining the business problem. The business problem we identified is how to help E-commerce companies increase their sales prediction and better their inventory management. Following that, our team has looked into various metrics which E-commerce businesses utilize to measure the success of their business outcomes. After that, we will then explain how the current E-commerce businesses use analytics to solve their business problems.

We will then be applying what we have learned from the E-commerce industry and use various models to demonstrate how analytics can solve the business problems that companies in e-commerce are currently facing. After performing some data cleaning, we did some exploratory data analysis and examined the relationships between the variables. Then we developed some machine-learning algorithms (Time Series Analysis & Association Rules) to gain more insights on future sales, and how purchasing the different items are associated with one another.

Finally, the report discusses the insights our team obtained from our analysis and how they can be applied to our business problem. All in all, although we managed to derive certain associations and trends from our predictive models, it still lies on E-commerce to do their due diligence to make improvements to their businesses.

Introduction

Global e-commerce sales are on an ever-rising trend, with a 265% growth rate from 2014 anticipated in 2021, accounting for an estimated 17.5% of the worldwide retail market (Beeketing, 2020)¹.

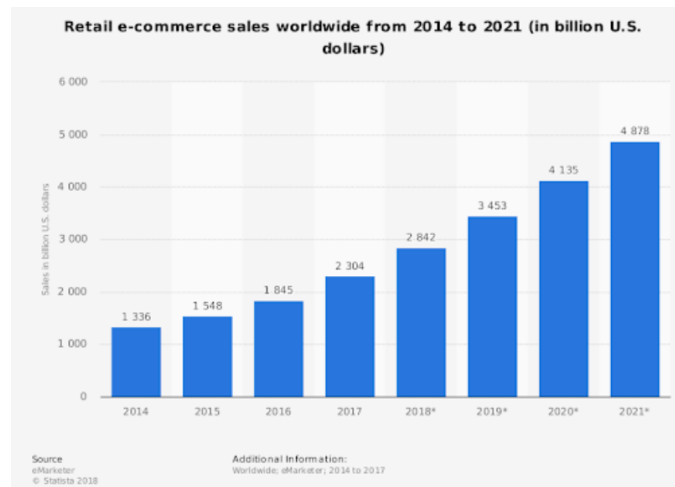


Figure 1: Growth of e-commerce sales worldwide from 2014 to 2021

This can be attributed to the rise in technological advancements and breakthroughs, making the world more interconnected than ever before. Online shopping has been made convenient for all to enjoy, and has seen a considerable rise in significance in the retail world today. Another notable factor would be due to the Covid-19 pandemic, as retail stores are shifting online to continue operating during lockdowns imposed by numerous countries, and consumers are shopping from home instead of at malls.

As such, the need for analysis tools to streamline such processes have also increased. With the importance of such an industry in today's era, our group has decided to focus on e-commerce analysis.

¹ Future of Ecommerce: 10 International Growth Trends (2020). Retrieved April 2 from [Future of Ecommerce: 10 International Growth Trends \(Updated 2020\) - Beeketing Blog](#)

Business Problem

Problem Statement: To help companies that are engaged in e-commerce to increase their predictive accuracy of what customers would likely buy to increase their sales and better manage the inventory that they have.

E-commerce companies need to be able to roughly predict the demand of the customers that they have to more optimally run their business as a whole. There are many factors that are in play for these companies with the massive amounts of data that they collect, and analytics will be able to help these companies narrow what are some of the factors that contribute most to the prediction of the customer's demands.

Additionally, being able to predict what a customer is likely to purchase in a certain month or season will provide the company with valuable information that will greatly help in inventory management as well as sales forecast. This allows them to plan the optimal amounts and timing to produce or order inventory with greater accuracy and detail, thus cutting down on any potential loss in terms of unsold goods or insufficient warehouse space.

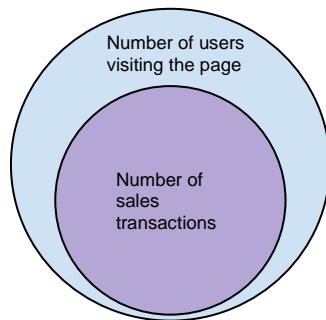
Measuring Success of Business Outcomes

There is a large advantage that E-commerce sites have over physical retail stores, and that is the existence of digital footprint. In comparison to other marketing efforts, digital offers far more visibility to behaviors and actions. You are unable to tell how many people have seen your billboard along the road, but you can see the actual number of people coming to your website, clicking on your ads, or the number of items into their shopping cart. In other words, the metrics present in an E-commerce sector are far greater and intricate than in the Retail store sector.

Taking into account our business problem, we have come to a conclusion that the desired outcomes we hope to achieve will be increasing E-commerce sales and reducing losses over time through accurate prediction of future sales. We will be looking at different business metrics that can give an accurate representation of whether the E-commerce business is a success and also, some that are more relevant to the context of achieving our business outcomes through accurate predictions.

1. Sales Conversion Rate

By taking the total sales divided by total number of visitors, it measures how successful the business is in converting their website visitors into paying customers.



Through having a clearer understanding of what percentage of the e-commerce platform users are making successful purchases, will be able to allow business owners to gauge the success of their e-commerce platforms and identify areas for improvement.

Figure 2: Sales conversion rate depicted in Venn-diagram

2. Average Order Value

Average Order Value is calculated as the average value of each purchase made online. This can be derived from division of the total quantity sold with the total sales generated. As a for-profit business, it is natural for business owners to want their customers to purchase more within a single receipt.



Figure 3: Average Order Value

Thus, average order value can be a significant and clear indicator as to how well a business is doing in ensuring their customers are purchasing more instead of less. With this indicator on hand, business owners can then think of different ways to

drive up this metric, such as bundling products together and selling them at a discounted price. Customers will then be incentivized to make more purchases as opposed to buying the items individually.

3. Customer Satisfaction & Returning Rate



Figure 4: Customer Satisfaction Feedback

Customer Satisfaction & Returning Rate are important indicators of how satisfied the customers are with the company's products. Customer Satisfaction can be measured through various different channels such as surveys and the Customer Satisfaction (CSAT) score card². Through conducting this post purchase interaction to gather customers' feedback, the business owners will be able to attain a general understanding of the level of satisfaction of consumers regarding the service and products they are providing. This will not only provide a firm foundation of the current customer success, but also help to boost engagement with their customers.

4. Customer Lifetime Value

The customer lifetime value metric will paint a clear picture for business owners with regards to how far they should go to retain their customers. As this metric will show how much they can earn from a single customer over the course of their life. Knowing your customer lifetime value is important as a means to lower the cost of retaining customers that has a much lower lifetime. By understanding this metric, business owners can then derive ways to increase it, such as increasing their average order value and also building a loyalty program, to see an increase in repeating customers.

² Customer Effort Score, NPS & Customer Satisfaction Score (Mar 22, 2019). Retrieved April 02 from [Customer Satisfaction Metrics Explained: Effort Score, NPS & CSAT \(datapine.com\)](https://datapine.com/blog/customer-satisfaction-metrics-explained-effort-score-nps-csat)

The Usage of Analytics in E-commerce

With E-commerce forecasted to experience an exponential growth upwards, it is inevitable for E-commerce businesses to utilize analytics to anticipate changes in the market through reliable data insights. With analytics, big data coming from e-commerce can be effectively and efficiently assessed for business owners to make informed decisions.

1. Customers' purchasing patterns and sales forecast

As mentioned in the Business Statement, one of the largest problems faced by e-commerce companies is the inability to predict and identify customers' purchasing patterns. As such, they are unable to maximize their sales revenue as there will always be a gap between what the customer could buy and what the customer actually buys.



Figure 5: 10 Powerful Behavioral Segmentation Methods

By analyzing historical data in the shop, e-commerce companies are able to sieve out a ton of information regarding the purchasing patterns of their customers. This enables the companies to improve their sales number by utilizing the information to cross-sell and up-sell to their customers. In doing so, customers will be prompted to purchase more products from the shop, thus increasing their sales revenue.

2. Improving Inventory Management

Inventory management may seem like a problem that only physical retail stores need to address, however, e-commerce shops also need to anticipate their upcoming sales so that they are able to accurately time their purchase of stock and supplies. A lot of potential revenue can be lost just due to out-of-stock products. Being able to anticipate their upcoming sales would minimize these situations to a minimum and increase revenue significantly.

Having out of stock items is a big deal for e-commerce businesses. Once a customer sees that a product is out of stock, the customer will most likely look for the next best alternative at another retailer. This means that the company is not only losing out to their competitor, but also losing any potential future sales that they might gain and leaving a negative impression on the customer.

The analytics solution can help the company to predict future sales numbers, and consequently, the quantity of goods to be sold. In doing so, companies can more accurately and easily forecast and plan inventory needed for the upcoming period.³ Thus, this will reduce the occurrence of inventory mismanagement.

³ Ecommerce analytics 101: How to drive more online sales with data. (2021, March 12). Retrieved April 04, 2021, from [Ecommerce analytics 101: How to drive more online sales with data \(supermetrics.com\)](https://supermetrics.com/blog/ecommerce-analytics-101-how-to-drive-more-online-sales-with-data/)

3. Other Applications of Analytics in the e-commerce Industry

There also exists many other benefits that enhancements that analytics can bring as well that we will not be focusing on. Below are some of the other ways companies use analytics.

- 1.1. Creating a personalized shopping experience.⁴ In a research by Epsilon, it mentions that 80% of consumers are more likely to buy an item from that company if they offer a personalized experience.⁵ Analytics can make use of customer demographics in order to personalize the shopping experience for their customers. This not only includes the ability to personalize products that you are buying, but also personalized in the sense of recommending the right kind of products to the right customer.
- 1.2. Using analytics to design new products for customers. Analytics can be used to see the trends that customers have with the evidence of their customers' preferences and dislikes as well. Companies can use that to their advantage by designing and selling products that customers will want to buy.
- 1.3. Optimizing the pricing of products.⁶ Analytics can be used to analyze pricing trends and determine what the optimum price of a product is in order to maximize the profitability of that product. Amazon is one company that uses this method to price their products.

⁴ 9 benefits of analytics in ECOMMERCE INDUSTRY. (2019, August 13). Retrieved April 04, 2021, from [9 Benefits Of Analytics In Ecommerce Industry - ComTec Info RPA](#)

⁵ New Epsilon research indicates 80% of consumers are more likely to make a purchase when brands offer personalized experiences. (n.d.). Retrieved April 04, 2021, from [New Epsilon research indicates 80% of consumers are more likely to make a purchase when brands offer personalized experiences](#)

⁶ 9 benefits of analytics in ECOMMERCE INDUSTRY. (2019, August 13). Retrieved April 04, 2021, from [9 Benefits Of Analytics In Ecommerce Industry - ComTec Info RPA](#)

Data Exploration

1. Most Purchased Items

In general, the top 3 items that generated most sales are the 'White Hanging Heart T-Light Holder', 'Regency Cakestand 3 Tier' and the 'Jumbo Bag Red Retrospot' and the other items all took up a smaller portion of items sold. Thus the business can put more emphasis on promotion of other items that are not generating as much sales, or they can stop the sales of those products and come up with new products as a form of innovation to generate more revenue.

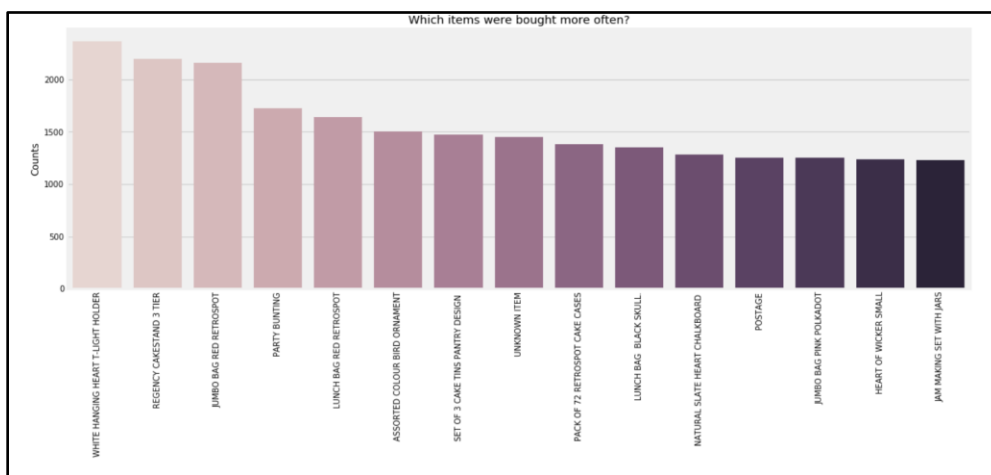


Figure 6: Python exploration - List of items bought the most often

2. Sales Distribution throughout the different countries

As seen in Figure 7 below, France, Germany and Ireland are a few of the countries with the highest amount of transactions occurring. However, the figure does not include the United Kingdom, which in our exploration was discovered to have the largest number of transactions. In fact, up to 88.83% (Figure 8) of the dataset were transactions occurring from the United Kingdom.

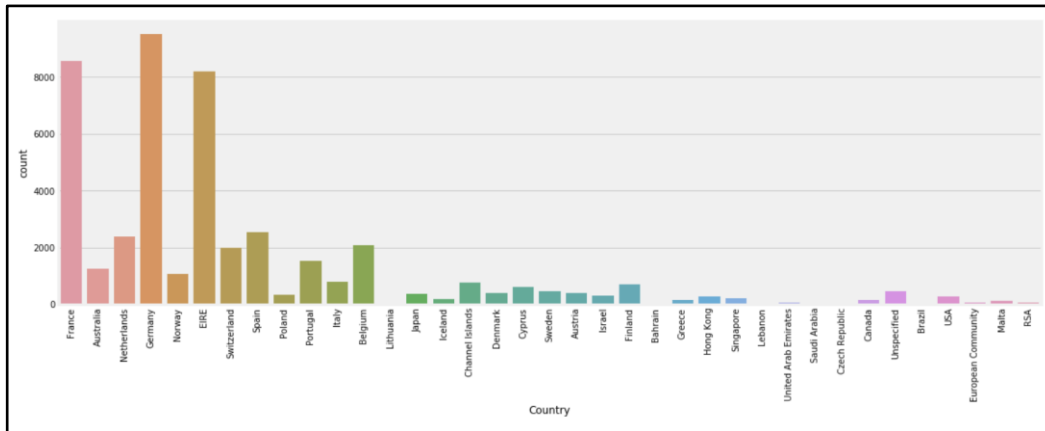


Figure 7: Python exploration - Number of transactions/invoices grouped by countries

```
print(countries.head(), '\n')

uk_count = df[df['Country'] == 'United Kingdom']['Country'].count()
all_count = df['Country'].count()
uk_perc = uk_count/all_count
print(str('{0:.2f}%').format(uk_perc*100))

United Kingdom    19857
Germany           603
France            458
EIRE              319
Belgium           119
Name: Country, dtype: int64

88.83%
```

Figure 8: Python exploration - United Kingdom's transaction as a percentage of the dataset

3. Sales distribution throughout a week/month timeframe

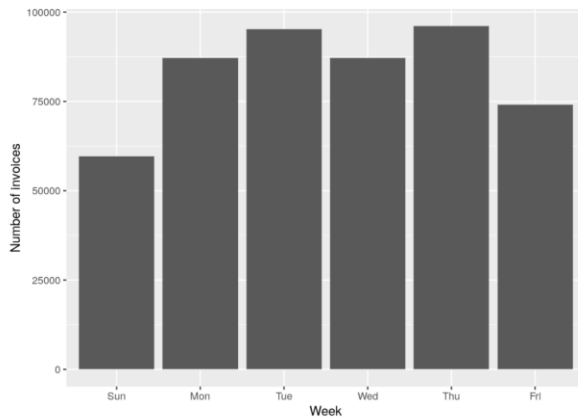


Figure 9: Python Exploration – No. of invoices by the days of a week.

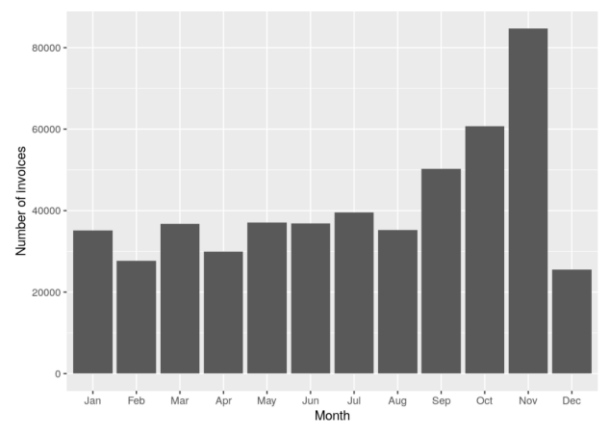


Figure 10: Python Exploration – No. of invoices by month

Both figures 9 and 10 represent the distribution of sales throughout their respective time frame. In figure 9, we can see that sales are relatively much higher from Monday to Thursday compared to the other days. In Figure 10, the sales are also higher in September to November compared to the other months. Thus from these visualizations, the business owners can have a better grasp of which months and which days their sales are performing better and derive solutions to improve their sales in other days.

Building Our Prediction Model

Data exploration, cleaning and visualization has been done under the notebook 'Data Exploration.ipynb'. Now, the next step lies in building our prediction model. As mentioned previously, the Analytics solution consists of two main components - Predicting customers' purchasing patterns and improving inventory management.

1. Improving Inventory Management - Time Series Forecasting

The technique used to explore this component of the Analytic Solution will be Time Series Forecasting. Time Series Forecasting is a model used to predict future values based on past values. This is done by extracting important insights and statistics for predictive purposes, and can identify any regularity or systematic variation in the dataset due to seasonality. At the same time, it can explain the growth rate of the trends in the data.

In supply-chain management, time-series is often used to get a forecast of future business activity. For example, in e-commerce, companies can predict the amount of goods that will be sold in subsequent weeks. With that in mind, they are able to tune their production level such that it matches the level of sales. As a result, they are able to minimize the risk of having excess and shortage of stock.

Under Time Series Forecasting, there are numerous methods that we have utilized in our notebook, but in the report we will be exploring the two best methods, namely the Exponential Smoothing Methods and ARIMA.

1.1. Exploration of historical time series

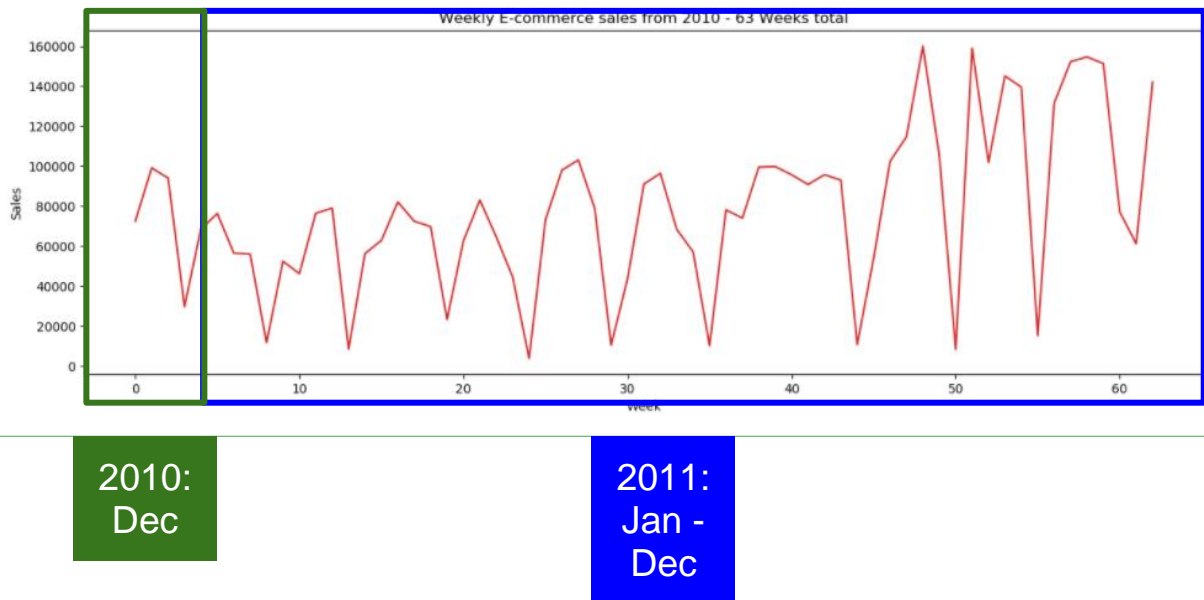
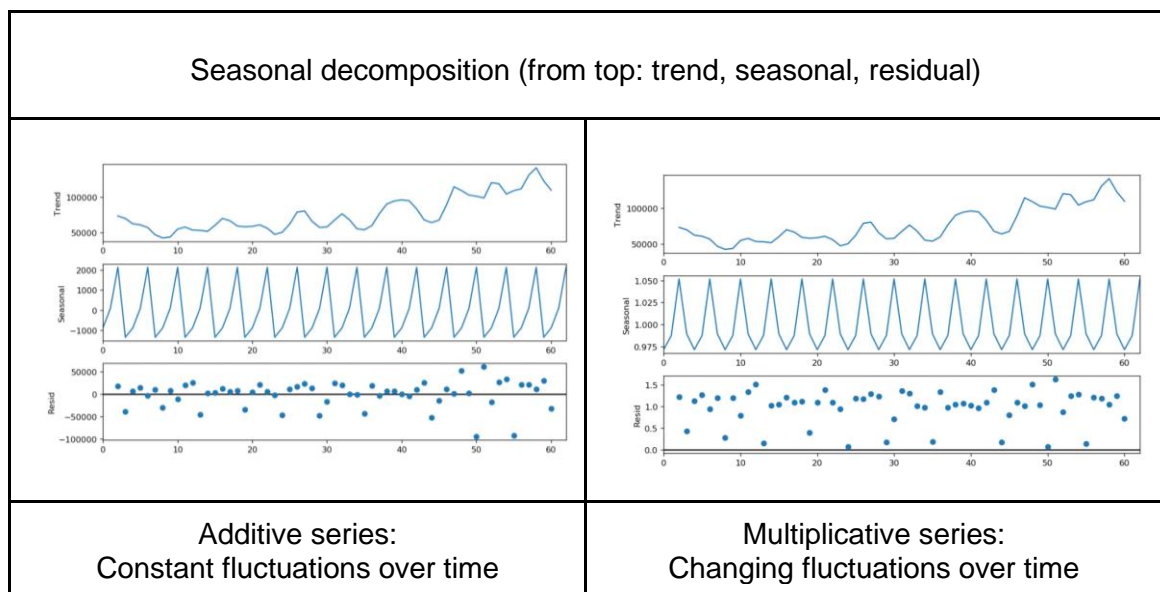


Figure 11: Historical Time Series Plot

The time series plot of total quantity sold of goods sold online can be seen above. Each unit of time is defined as a week to ensure that we have sufficient data to build an accurate model, since the dataset only has slightly over a year worth of data.

After decomposing the original time series, we are able to see the standalone effects of trend and seasonality. Although both plots are similar, the multiplicative series is more applicable as the original dataset has changing fluctuations over time.



1.2. Time-series forecasting

We will be using exponential smoothing and ARIMA to forecast the time series. Simple exponential and Holt's method fits poorly due to the presence of seasonality in our dataset. However, we will still build the model to get their RMSE values for comparison. Their time series plot can be found in the appendix below.

1.2.1. Holt-Winters Exponential Smoothing

Holt-Winter's method is to apply exponential smoothing to the seasonal components in addition to level and trend. After performing the Holt-Winter's method in python, the following results were achieved.

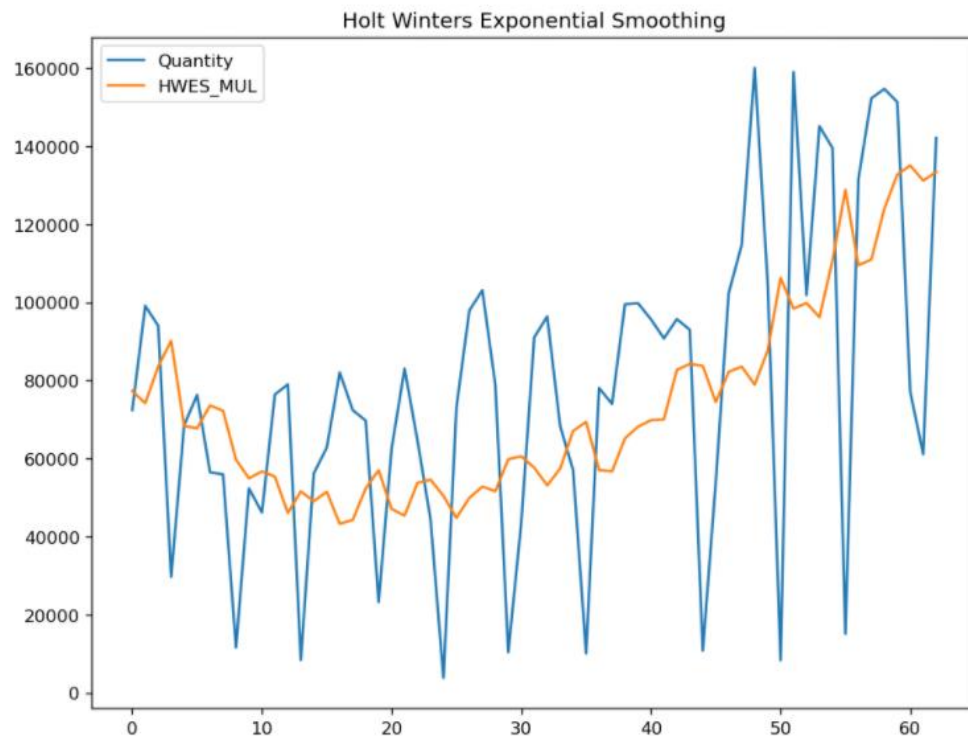


Figure 12: Holt-Winters Exponential Smoothing Plot (Actual vs Predicted)

1.2.2. Seasonal ARIMA Method⁷

The Auto-Arima algorithm in python was used to build a series of models with different parameters (p, d, q). The algorithm also compares these models based on their AIC score, using the model with the best results. After using seasonal auto ARIMA in python, the following results were achieved.

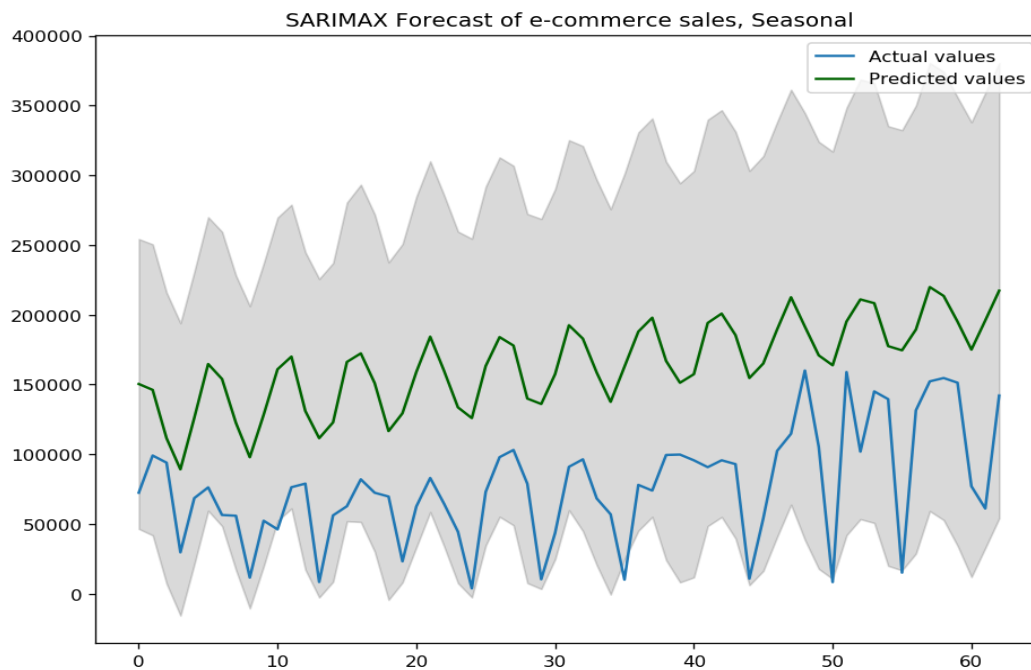


Figure 13: Seasonal ARIMA Plot. (Actual vs Predicted)

1.3. Model Evaluation

After we derive the prediction results, we will analyze the accuracy of the model using the following metrics: Root Mean Square Error (RMSE), Mean Absolute Error (MAE) and Akaike Information Criterion (AIC).

RMSE and MAE measure the differences between the actual and predicted value while the AIC quantifies the goodness of fit and the parsimony of the model in a single statistic⁸. By comparing between different models, the one with the lower AIC, RMSE and MAE is better.

⁷ ARIMA Model by Selva Prabhakaran. Retrieved April 2 from [ARIMA Model - Complete Guide to Time Series Forecasting in Python | ML+ \(machinelearningplus.com\)](#)

⁸ Using AIC to Test ARIMA models (August 14, 2013). Retrieved April 2 from [Using AIC to Test ARIMA Models – CoolStatsBlog](#)

However, we should not decide on the best model based on these metrics alone. Since every dataset have different nature, some with or without seasonality and trend, we need to consider the strengths and weaknesses of each model. Since Holt-winter's method has very good AIC, RMSE, MAE scores and considers both seasonality and trend, we decide that it is the most relevant model for this dataset.

Model	AIC	RMSE	MAE
Simple Exponential Smoothing	1337	39239	30250
Holt's method	1334	37172	29498
Holt-Winters	1347	38669	31005
ARIMA	1512	40438	31794
Seasonal ARIMA	1431	91619	86474

1.4. Shortfalls




Time series forecasting is an extrapolation of the historical data and does not use other factors or predictors unlike other techniques. Thus this statistical model is based on the assumption⁹ that the current existing pattern will continue as such into the future. Such an assumption is more likely to be accurate in the short term rather than long term. Thus this model will be able to provide us with accurate forecasts into the immediate future rather than further into the future. Additionally, the model has to be adjusted accordingly based on new arrivals of data and events that may have a significant impact on sales. This is to ensure our models are able to produce a much more accurate prediction using the past datasets.

⁹ How to choose the right forecasting technique. Retrieved April 2 from [How to Choose the Right Forecasting Technique \(hbr.org\)](https://hbr.org/forecasting-technique)

2. Predicting Customers Purchasing Pattern - Association Rules

The technique used to explore this component of the Analytic Solution will be Association Rules. Association rules, also known as Market Basket Analysis, is a model used to analyze transactions for relationships. In the case of an e-commerce store, products of transactions are analyzed to find any association between the purchase of a set of products with another set of products. The rules generated can then be used for recommending products to customers. To better understand the association rules, we can look at them in terms of probability of occurrence. The typical method of generating association rules is the Apriori algorithm, which analyses various combinations of products to find the most relevant rules. Specifically for our dataset, we will focus on data from France, as it has one of the largest total sales, to determine association with any other product or set of products in the gift shop, which can anticipate the customers' next purchase. The shop can then recommend the associated combination of products to these customers. This method aiming to improve sales revenue is actually very popular amongst e-commerce sites, and has been known to be widely adopted by many companies, such as Amazon as seen in the Figure below.

Frequently bought together




Total Price: **\$357.84**
[Add all three to Cart](#)


1 These items are shipped from and sold by different sellers. [Show details](#)

- ☒ **This item:** Scotts Whirl Hand-Powered Spreader **\$520.02**
- ☒ Scotts Turf Builder Weed and Feed 3, 5,000 Sq. Ft. **\$5147.98**
- ☒ Scotts Turf Builder Triple Action, 5M **\$5189.84**


Customers who viewed this item also viewed




Scotts Turf Builder EdgeGuard Mini Broadcast Spreader - Spreads Grass Seed, Fertilizer and Ice Melt - Holds up to 5,000 sq....
★★★★☆ 12,968
\$549.54




Precision Products HHBS-125 Handheld Broadcast Spreader
★★★★☆ 98
\$5207.69




TIPU Hand Spreader Shaker, Handheld Spreader for Ice & Snow Melt, Salt, Fertilizer, Grass Seed and Earth Food - Multiple...
★★★★☆ 524
\$527.14




Chapin 8706A 1.6L/.4 Gal All season Poly Hand Crank Spreader For Seeds, Fertilizer and Ice Melt. 1.6L (1 Spreader/Package)
★★★★☆ 82
\$527.21



Hand Spreader Shaker for Seed, Salt, De-Icer, Ice Melt, Earth Food and Fertilizer - Multiple Opening Sizes for Any Need - Up to 80 Oz...
★★★★☆ 1,316
\$537.48



Stalwart 82-YJ487 Spreader, Salt and Seed Spreader, Also for Fertilizer, Deicing/Ice Melt By Stalwart (Multipurpose Tool for...
★★★★☆ 918
\$513.59



Home-X Ice Melt Salt Dispenser, Grass Seed Spreader, Plastic Garden Container
★★★★☆ 258
\$510.85

Page 1 of 2

Figure 14: Amazon's Website depicting usages of Association Rules

2.1. Results

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
21	(SET/6 RED SPOTTY PAPER CUPS, SET/20 RED RETRO...	(SET/6 RED SPOTTY PAPER PLATES)	0.102828	0.128535	0.100257	0.975000	7.585500	0.087040	34.858612
22	(SET/6 RED SPOTTY PAPER PLATES, SET/20 RED RET...	(SET/6 RED SPOTTY PAPER CUPS)	0.102828	0.138817	0.100257	0.975000	7.023611	0.085983	34.447301
19	(SET/6 RED SPOTTY PAPER PLATES)	(SET/6 RED SPOTTY PAPER CUPS)	0.128535	0.138817	0.123393	0.960000	6.915556	0.105550	21.529563
18	(SET/6 RED SPOTTY PAPER CUPS)	(SET/6 RED SPOTTY PAPER PLATES)	0.138817	0.128535	0.123393	0.888889	6.915556	0.105550	7.843188
3	(ALARM CLOCK BAKELIKE RED)	(ALARM CLOCK BAKELIKE GREEN)	0.095116	0.097686	0.079692	0.837838	8.576814	0.070400	5.564267
2	(ALARM CLOCK BAKELIKE GREEN)	(ALARM CLOCK BAKELIKE RED)	0.097686	0.095116	0.079692	0.815789	8.576814	0.070400	4.912229
20	(SET/6 RED SPOTTY PAPER CUPS, SET/6 RED SPOTTY...	(SET/20 RED RETROSPOT PAPER NAPKINS)	0.123393	0.133676	0.100257	0.812500	6.078125	0.083762	4.620394

Figure 15: Association Rules generated with specified constraints (i.e France, Min Supp 7%)

After applying the apriori algorithm at a minimum support of 7%, and filtering the data frame to obtain rules with Lift > 6 and Confidence > 80%, the rules obtained were as seen in Figure above.

2.2. Exploration and conclusion of the rules

<pre>print("Red Retrosport Paper Napkins (X) : " ,basket['SET/20 RED RETROSPOT PAPER NAPKINS'].sum()) print("Red Spotty Paper Cups (Y) : " ,basket['SET/6 RED SPOTTY PAPER CUPS'].sum()) print("Red Spotty Paper Plates (Z) : " ,basket['SET/6 RED SPOTTY PAPER PLATES'].sum())</pre>
<pre>Red Retrosport Paper Napkins (X) : 960.0 Red Spotty Paper Cups (Y) : 1272.0 Red Spotty Paper Plates (Z) : 1116.0</pre>

Figure 16: Total Sales of items in itemset for Rule 1

We will begin by taking the first row of rules as an example. From Figure above, we can see that the items in the rules are of different values. In particular, the quantity sold for the item in the consequent is lower than the antecedents. Drawing from this observation, and the confidence of the rule, we can tell that there is up to 97.5% chance that a customer will buy the consequent item given that they purchase the antecedent item. Hence, the e-commerce site can use this result to display in the 'Frequently Bought Together' section as seen in the Amazon example, or they could even offer packages for these items to be sold together at a lower price, thus prompting the customers to buy the package and as a result increase the overall sales for all three items.

2.3. Shortfall

While relevant association rules can be helpful, not all rules are useful in analysing the probability of occurrence. For large datasets with numerous items, filtering the useful rules out of all generated rules might be time-consuming and costly.

Conclusion

In conclusion, there are many ways for an e-commerce site to improve their business by using Analytics. In this digital era, we can easily find many solutions to the problems that e-commerce shops are facing. In fact, as COVID-19 has disallowed the physical act of shopping at retail stores, e-commerce sites are more relevant now than ever before.

By making use of the Analytics Solution that we have provided in this report, there can be great improvements made to the business strategies adopted by many e-commerce sites, and aim to improve their sales revenue and customer retention.

Appendix

Time Series

Simple exponential smoothing: no seasonality and no trend

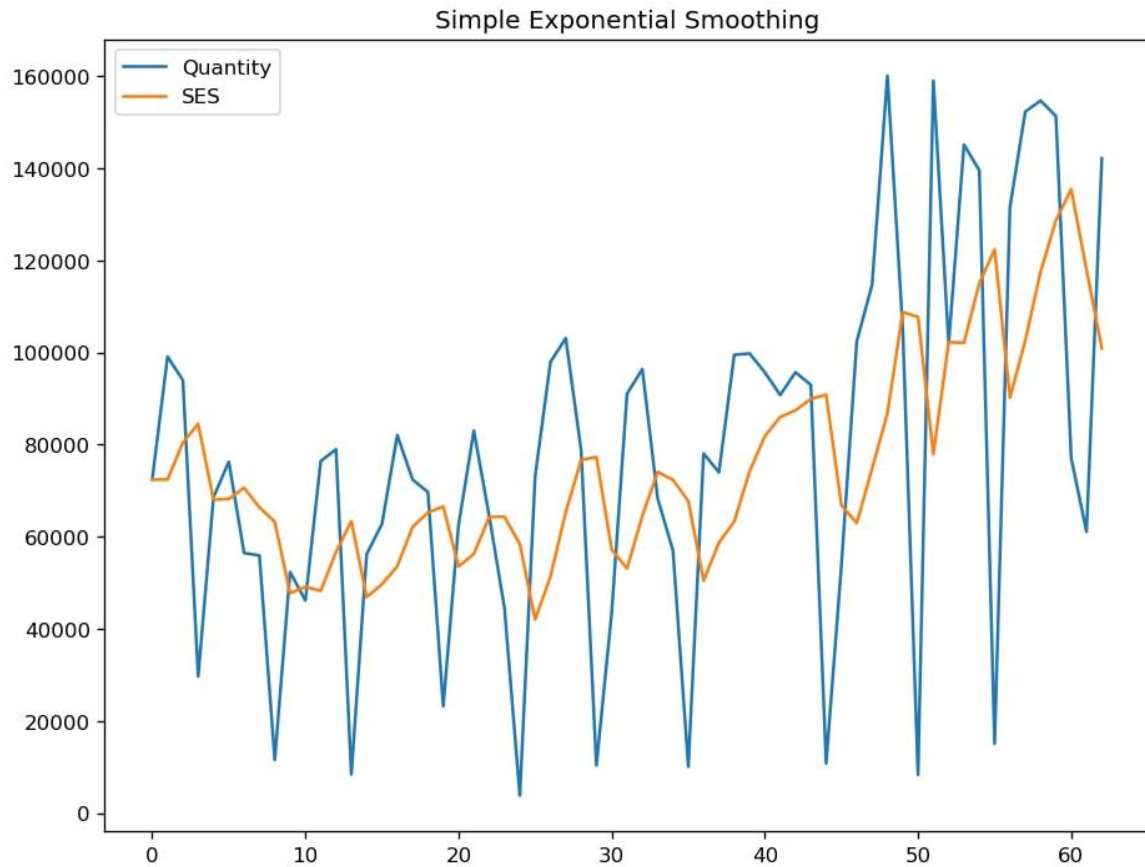


Figure 17: Simple Exponential Smoothing Plot (Actual vs Predicted) – No Seasonality and No Trend

Holt's method: trend and no seasonality

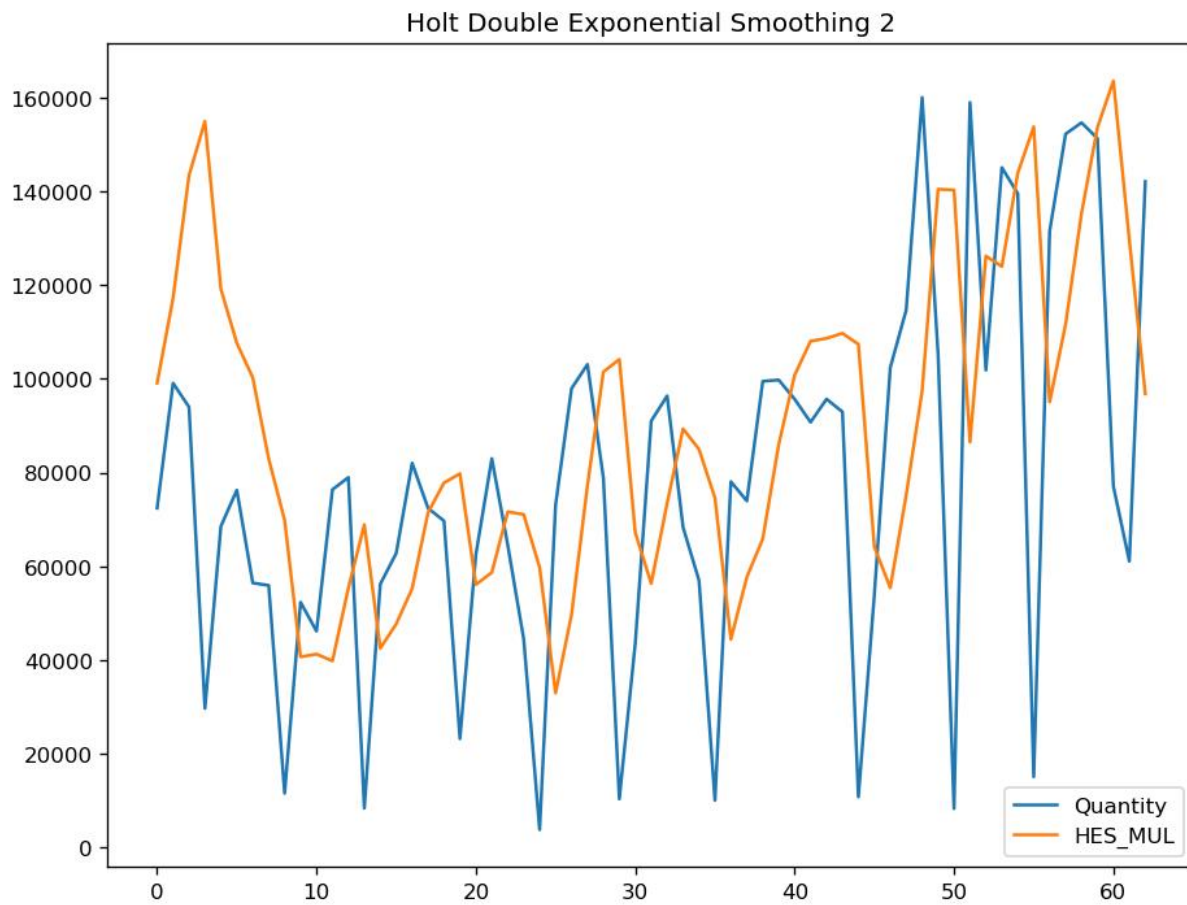


Figure 18: Holt's Method Plot (Actual vs Predicted) - Trend but No Seasonality

Non-seasonal ARIMA:

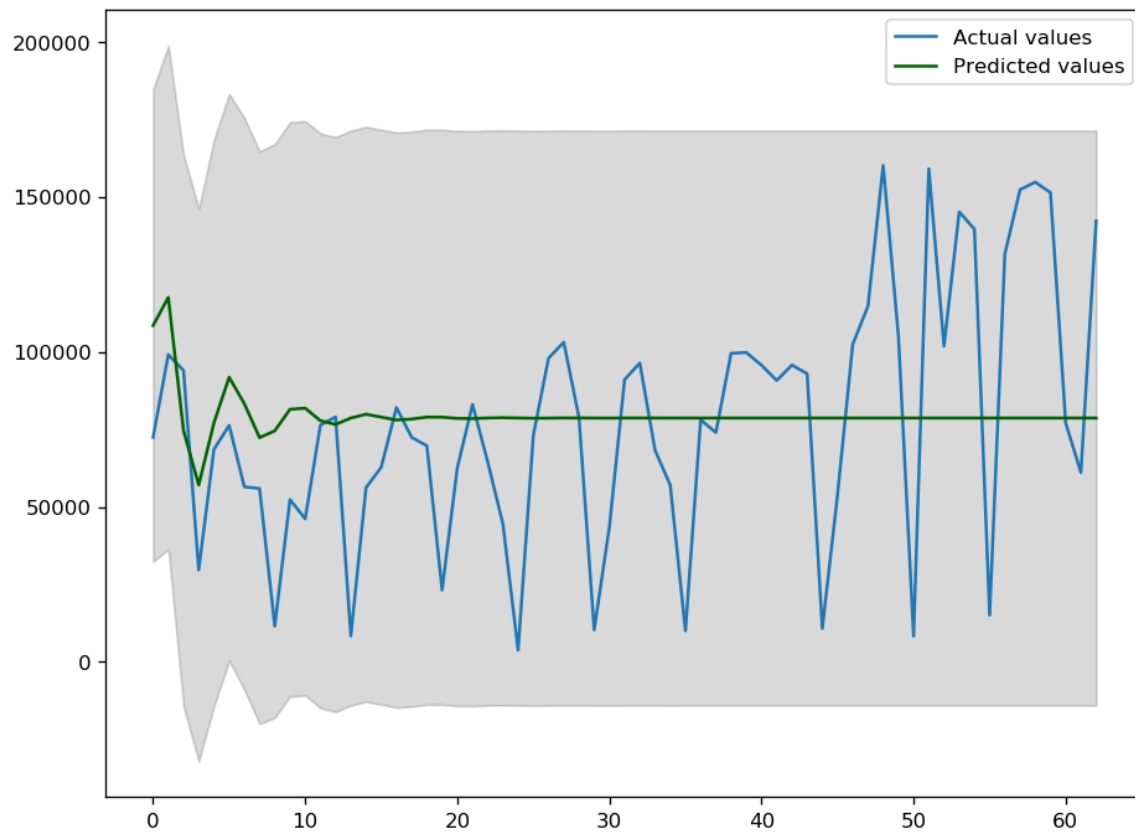


Figure 19: ARIMA Model without seasonality effects (Actual vs Predicted)