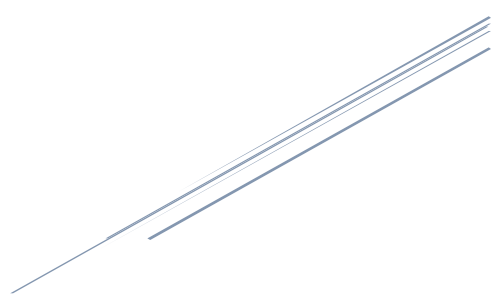


# **STATISTICAL ANALYSIS IN THE CONTEXT OF BUSINESS GROWTH**



Department of Mathematics  
Faculty of Science  
University of Peradeniya

## **GROUP VII - Members**

- **S/16/849 : W. M. L. R. Wijethunga**
  - **S/16/807 : J. K. R. Fernando**
  - **S/16/828 : R. A. M. Ranawaka**
  - **S/16/814 : R. N. Madushika**
  - **S/16/842 : S. H. N. Thapangani**
  - **S/16/821 : S. N. M. Najwan**
  - **S/16/835 : S. A. R. Senanayake**
- 

# **Introduction**

## **1.1 Background of the study**

Business logistics is a critical and important part of a successful and healthy economy in the modern world. The globalization trends and competition among the various economies has made logistics more important. The transportation, manufacturing and distribution of goods to local and international markets plays an immense role in every business aspect. With the innovative emerging global economy, a business organization's capability to compromise speedy and effectual delivery of goods and services will be critical competitive significance (Voortman, 2004). The process of production, transportation and distribution of goods to the customers, from the origin to the point of consumption is a very vital and important procedure in every multi- national business organization.

The manufacturing, transportation and distribution of goods of a non- store online retail business organization is a very critical procedure because of the absence of warehouses to store the finished goods. The non-store retailers are known by the medium they use to communicate with their customers such as direct marketing, direct selling and vending machines or e-tailing. Since there's no storing facilities, there is always a direct connection with the customer in these types of businesses. The main advantage that can be taken from these types of businesses is that non-store retailing requires customers to submit some necessary personal details to buy a product. This way, the company can keep a record of their customers, can access them with promotional offers and can use the data to analyze and make necessary business and organizational decisions.

The statistical analysis of the data is a very critical and important step that should practice in these type of business organizations to take up to date accurate decisions. The globalization and the modern economic fluctuations may cause the variations of the trends in the market which will cause significant losses if the business organization doesn't take the correct decision at the right time. The statistical analysis of data is much more important because of those purposes which decides the future of any business organization in the modern economy.

This study is based on well reputed non-store online retail company in United Kingdom namely ABC (Pvt) Limited, which has been an industry with over 12 years in the online retail business. During the past years the company has experienced a gradual increase in the attrition of existing customers which has affected the organizational performance. To identify what factors to focus in retaining the customers by understanding the gaps from organization point of view, a comprehensive study was required.

## **1.2. Project problem identification**

As stated in the background of the study statistical data analyzing plays a major role in these types of businesses. In the current challenging environment, it is important to identify the factors effecting for the growth of the business, the current trends of the business, most selling items and the geographical regions with highest sales etc. In the existing competitive scenery of these type of industries accomplish customer retention to safeguard the company's customer base and the increasing of new customers and new sales are important. The group identifies the problem of this project study as to identifying factors affecting sales in ABC (Pvt) Limited for the retention of existing customers and for new customers.

# Methodology

This analysis was conducted on a UK based non-store online retail business. The selected dataset consists of all transactions occurred between 01/12/2019 and 09/12/2020. Each row of the dataset represents a transaction and there are a total of 785,179 observations with 8 parameters.

- Invoice – Invoice number
- ProductCode – Product code
- Description – Product description
- Quantity – No of units sold
- InvoiceDate – Date of the transaction
- Price – Unit price
- CustomerID – Customer ID
- Country – Country of the customer

In this analysis R statistical software was used and mainly focused on three statistical analysis techniques.

## 2.1 Data Mining

Data mining is the process of identifying and extracting relationships and patterns within a dataset. Under the data mining techniques exploratory data analysis was used to summarize the dataset with visual methods and association rule mining was done to discover relationships between variables. It is easy to identify and compare relative performances using visualized methods in an analysis. Bar charts and line charts was used to visualize the dataset and using wordclouds text analysis was also done in this analysis. Association rule mining was done under the apriori principle with support of 2% and confidence of 70% to find out the products which are bought together by customers most frequently.

## 2.2 Cluster Analysis

To group observations which are sharing a similarity, cluster analysis was done. K-mean clustering technique under pre-defined no of clusters and WCSS Elbow Method (Within Cluster Sum of Squares) was used in this analysis. The term K-Means describes an algorithm that assigns each item to the cluster having the nearest centroid (mean). In its simplest version, the process is composed of three steps (Richard A. Johnson, Dean W. Wichern, 2012):

1. Partition the items into K initial clusters
2. Proceed through the list of items, assigning an item to the cluster whose centroid (mean) is nearest. (Distance is usually computed using Euclidean distance with either standardized or unstandardized observations.) Recalculate the centroid for the cluster receiving the new item and for the cluster losing the item.
3. Repeat step 2 until no more reassignments take place.

The final assignment of items to clusters will be, to some extent, dependent upon the initial partition or the initial selection of seed points (Richard A. Johnson, Dean W. Wichern, 2012).

The data points are clubbed together by detecting correspondences according to the attributes found in raw data (Tripathi, 2018), however the main purpose of WCSS Elbow Method is to find the suitable number of clusters which are relevant as well as insightful for analysis purposes.

It works on the principal that after a certain number of 'K' clusters the difference is SSE (Sum of Squared Errors) starts to decrease and diminish gradually. Here, the WCSS (Within-Cluster-Sum-of-Squares) metric is used as an indicator of the same. Hence the 'K' value specifies the number of clusters (Tripathi, 2018).

## 2.3 Time Series Analysis

Time Series Analysis is the way of studying the characteristics of the response variable with respect to time, as the independent variable. The objective of conducting a time series analysis in this study is to forecast the sales for next two years.

### *Model identification*

The following steps are outline approach to the model identification

Step 01: Plot the data

Step 02: Deeside the transformation is necessary to establish the variation to make seasonal effect additive. For this Box-Cox transformation can be used.

Step 03: Assess the stationary of this series using time series plot. ACF and PACF are also useful for stationary test. For stationary time series the data plotted is scattered around a constant mean. Also, ACF and PACF drop or near zero quickly.

- When the data appear non-stationary, they can be made stationary by differencing.
- For non-seasonal data take first difference of the data. If first difference data is also non-stationary take the first difference of the difference data (second difference of the original data). Most of the time maximum of two differences will transform data into a stationary data.
- For seasonal time series take seasonal difference of the data.

Step 04: When stationary has been achieved examine the ACF and PACF to see if any pattern remains.

Step 05: Diagnostic checking

$H_0$  : Residuals follows a white noise series with mean zero and constant variance.

$H_1$  : Residuals do not follow a white noise series with mean zero and constant variance.

Step 06: Forecasting

### 2.3.1 Residual Checking Methods

#### Box-Pierce Method

$$Q = n' \sum_{k=1}^h \gamma_k^2(\hat{Z}) \sim \chi_{h-m}^2$$

If  $Q$  statistic is significant the model is inadequate. Hence  $n' = (n - d)$  if differences have been taken and  $n' = (n - d - LD)$  when seasonal difference used.  $n$  is the maximum no of lags considered, choice of  $h$  is somewhat arbitrary.  $L$  Is the no of seasons.  $d$  is the non-seasonal differencing used to transform the original time series in to stationary.

#### Ljung-Box method

$$Q^* = n'(n' + 2) \sum_{k=1}^h \frac{\gamma_k^2(\hat{Z})}{(n' - k)} \sim \chi_{h-m}^2$$

If  $Q^*$  statistic is significant the model is inadequate. Hence  $n' = (n - d)$  if differences have been taken and  $n' = (n - d - LD)$  when seasonal difference used.  $n$  is the maximum no of lags considered. Choice of  $h$  is somewhat arbitrary.  $L$  is the no of seasons.  $d$  Is the non-seasonal differencing used to transform the original time series in to stationary.  $\gamma_k(\hat{Z}_t)$  is the sample residual autocorrelation at lag  $k$ .

# Results and Discussions

## 3.1 Data Wrangling

The original dataset consists of 785179 observations and 8 variables. The dataset did not contain any missing values, but the data type of some variables had to be changed for further analysis.

Table 1: Data Types

Variable	Original Data Type	Converted Data Type
Invoice	Integer	Factor
ProductCode	Character	Factor
Description	Character	Character
Quantity	Integer	Integer
InvoiceDate	Character	Date
Price	Numerical	Numerical
CustomerID	Integer	Factor
Country	Character	Character

As the next step, some modifications were done to some variables. The 'InvoiceDate' variable was further split in three variables as 'Month', 'Year' and 'Date'. A new variable 'InvoiceTotal' was formed to get the total value of that transaction.

Table 2: Original Dataset (Preview)

Invoice	ProductCode	Description	Quantity	InvoiceDate	Price	CustomerID	Country
536365	71053	Red Bag	6	12/1/2019	3.39	17850	UK

Table 3: Modified Dataset (Preview)

Invoice	StockCode	Description	Quantity	Price	CustomerID	Country	Month	Year	Date	InvoiceTotal
536365	71053	Red Bag	6	3.39	17850	UK	12	2019	2019-12	20.34

A secondary dataset was used to compare monthly sales with the country's temperature. Only the temperature for the necessary years were load and this was also modified for the analysis.

Table 4: Original Temperature Dataset (Preview)

Year	Period	Temp (C°)
2019	JAN	0.80
2019	FEB	1.60

Table 5: Modified Temperature Dataset (Preview)

Year	Period	Temp (C°)	Date
2019	JAN	0.80	2019-01
2019	FEB	1.60	2019-02



### 3.2 Exploratory Analysis

The performance of the business was first visualized by plotting the value of sales of each month from 2018-12 to 2020-12.

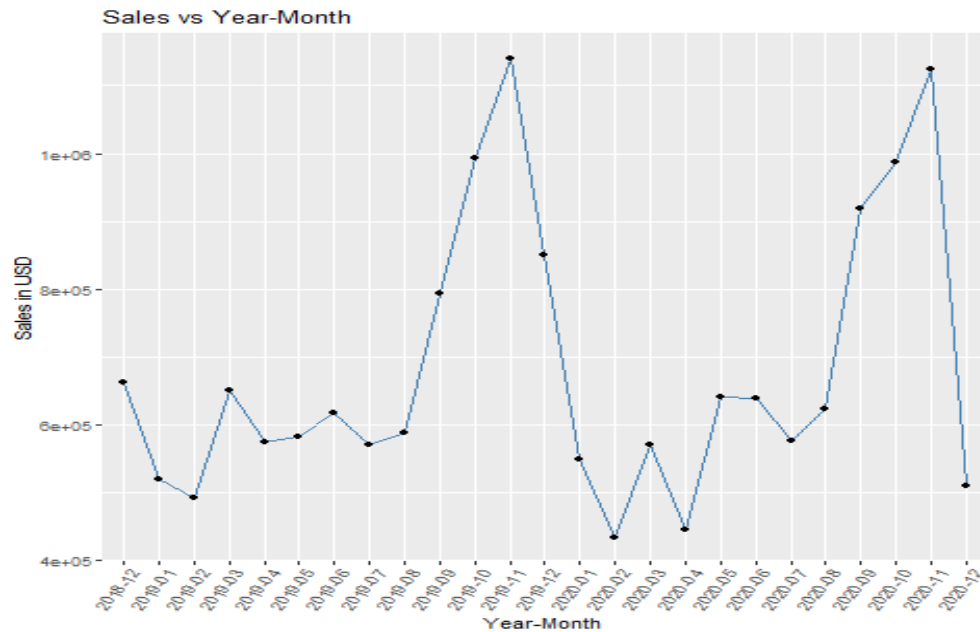


Figure 1: Monthly Sales Performance

It is evident from the above plot that in both years there is a significant increase in sales from the months September to November. This also depicts that a seasonal effect is present in the sales data. Further analyses were carried out specifically on those three months to identify the best performing products.

Top ten products of 2019 and 2020 were identified depending on their demand and sales (Revenue generated). A visualization was used to compare their relative performances to each other.

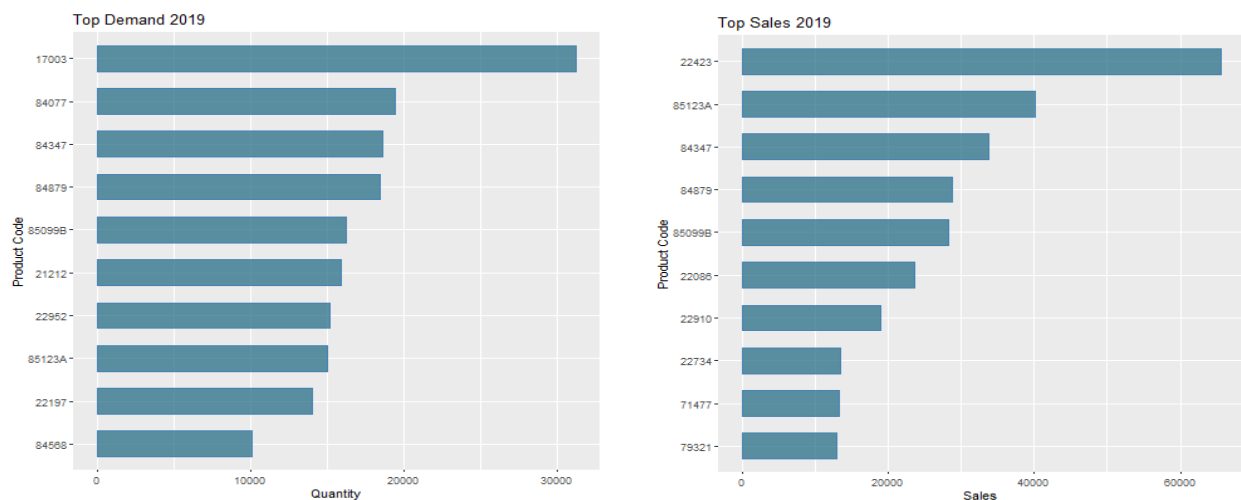


Figure 2: Best Products 2019

The product code 17003 has the highest demand among the products. Compared to the second highest demand product 84077, it has sold more than 10,000 units. But this product is not present among the highest revenue generating products. According to the above chart, the product with highest income is 22423 (product code). The products which were common in both top demand and top sales charts were considered as the best performing products.

Table 6: Best Products 2019

ProductCode	Description
85123A	White Hanging Heart T-Light Holder
84879	Assorted Color Bird Ornament
85099B	Red Retrospot Jumbo Bag
84347	Rotating Silver Angels T-Light Holder

Similar analysis was conducted on those months for 2020 and the results were as follows.

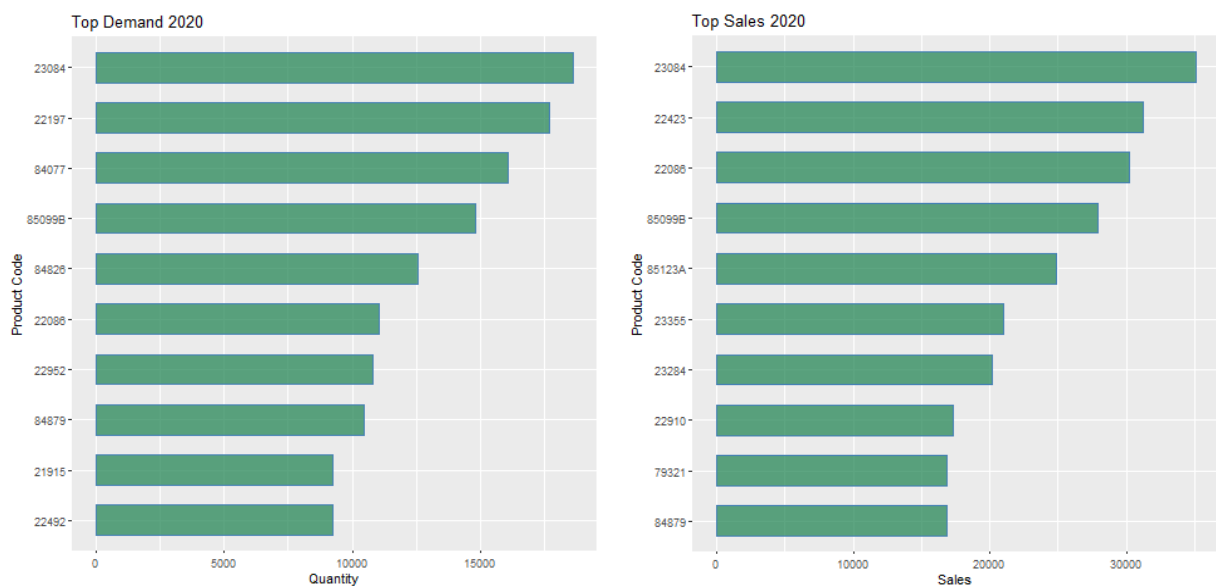


Figure 3: Best Products 2020

Table 7: Best Products 2020

ProductCode	Description
84879	Assorted Color Bird Ornament
22086	Paper Chain Kit 50's Christmas
85099B	Red Retrospot Jumbo Bag
23084	Rabbit Night Light

Next, a text analysis was conducted on the 'Descriptions' variables to identify most popular styles, colors and items among the customers during that period.

### 3.3 Text Analysis

To find what are the preferred characteristics of the best products, the product descriptions were visualized using wordclouds. This was done for the highest demand products during September to November of 2019 and 2020 separately.



Figure 4: Wordcloud For 2019



Figure 5: Wordcloud For 2020

According to the above figures, products with colors white, red and pink are popular among the customers. Also, vintage designs and designs with hearts can also be seen frequently. Christmas products, decorative products, bags, jumbo packs, metal decorative can also be seen as popular products. It will be better if more products with such features are introduced during those months in the future.

### 3.4 Association Rule Mining

Association rule mining was done using apriori principle to find what products are bought together by customers most frequently. For this purpose, apriori rule with a support of 2% and confidence of 70% was used.

Table 8: Association Rules

Association Rule	Supp	Conf
{Pink Regency Teacup and Saucer} => {Green Regency Teacup and Saucer}	0.02172	0.8478
{Pink Regency Teacup and Saucer} => {Roses Regency Teacup and Saucer}	0.02068	0.8074
{Poppy's Playhouse Bedroom} => {Poppy's Playhouse Kitchen}	0.02275	0.8666
{Poppy's Playhouse Kitchen} => {Poppy's Playhouse Bedroom}	0.02275	0.7750
{Green Regency Teacup and Saucer} => {Roses Regency Teacup and Saucer}	0.02482	0.7878
{Wooden Tree Christmas Scandinavian} => {Wooden Star Christmas Scandinavian}	0.02577	0.7960
{Wooden Star Scandinavian} => {Wooden Heart Christmas Scandinavian}	0.03222	0.7130

These association rules were also generated for the products during months of September to November. Association rules are important when advertising products. It is better to advertise products with association rules simultaneously .

### 3.5 Cluster Analysis

Firstly K-Means clustering was used to cluster the products based on units sold. This analysis was done for all the months except for September, October and November. The products with lowest demand during the off-season were identified using this process. The number of required clusters was identified using the WCSS elbow method.

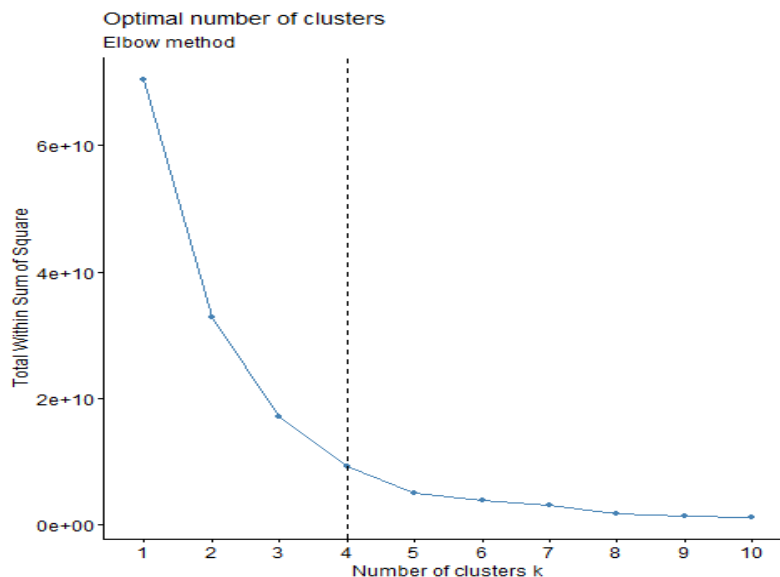


Figure 6: Number of Clusters for Products

From the above graph, the optimal number of clusters needed to cluster products on demand is 4.

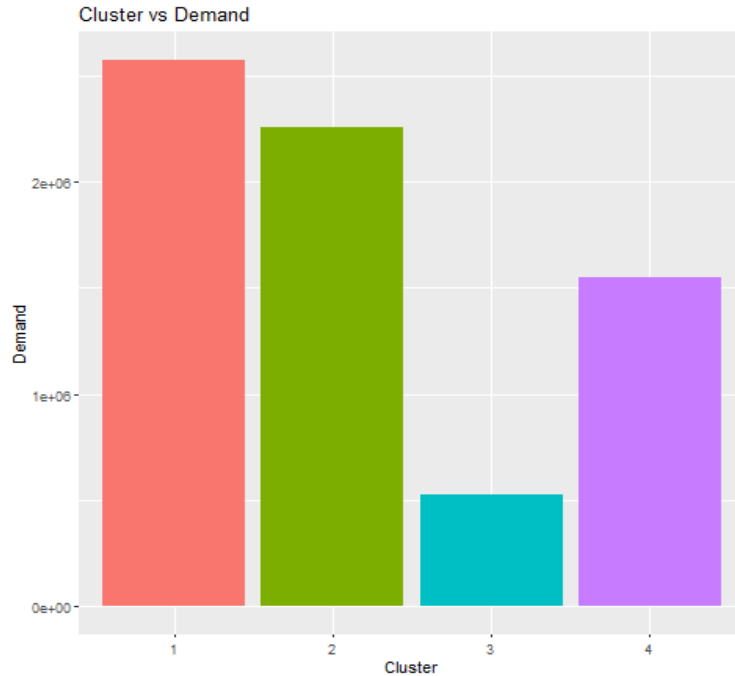


Figure 7: Cluster vs Demand

From the above column graph, it is evident that the products belonging to the cluster three has the lowest total demand. These products of cluster three were considered as the worst performing products during the off-season.

Table 9: Cluster Three Products

Description	Quantity
60 Teatime Fairy Cake Cases	39211
Assorted Color Bird Ornament	48539
Medium Ceramic Top Storage Jar	76676
Pack Of 60 Pink Paisley Cake Cases	39298
Pack Of 72 Retro Spot Cakes Cases	39298
Paper Craft, Little Birdie	80995
Red Retrosport Jumbo Bag	61333
White Hanging Heart T-Light Holder	68337
World War 2 Gliders Asstd Designs	70355

We can see that some of the products in cluster were also the best performing products during the months September to October. So, it is better to manage the inventory, by having them only during those three months. Other products can be replaced with new trendy products. This will reduce unwanted carrying cost and probably will increase sales too.

Next, the months were also clustered using K-Means algorithm to identify what months were the best performing , averagely performing and the worst performing.

The months were clustered into 3 clusters and their contribution towards the total sales are depicted in the pie chart below.

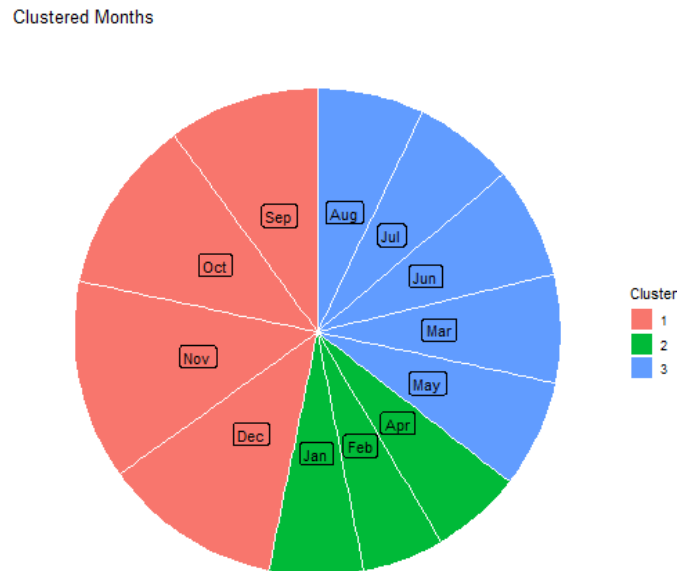


Figure 8: Clustered Months

From the above pie chart, we can see that the months September, October, November and December are responsible for about 45% of the total sales in those months are clustered as cluster 1. March, May, June, July and August are in the cluster 3, this cluster is the cluster with second highest sales. This cluster is responsible for about 35% percent of the total sales. Cluster 2 consist of the months January, February and April. It is the worst performing cluster, only responsible for 20% of the total sales. It is important to improve sales during these months, hence using proper timed marketing campaign within those months could be profitable. Analysis on the countries of customers was also done using the same K-Means clustering. The clustering was done based on the demand for products by the customers of these countries. To identify the needed number of clusters, WCSS elbow method was used.

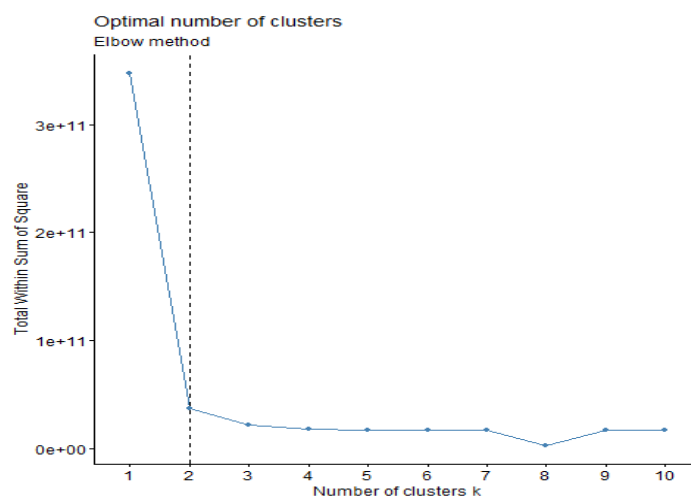


Figure 9: Number of Clusters for Countries

From the WCSS elbow method, the optimal number of clusters was obtained as 2. Countries with their clusters are depicted in the pie chart below. For this analysis United Kingdom was not considered as the shop is based in United Kingdom.

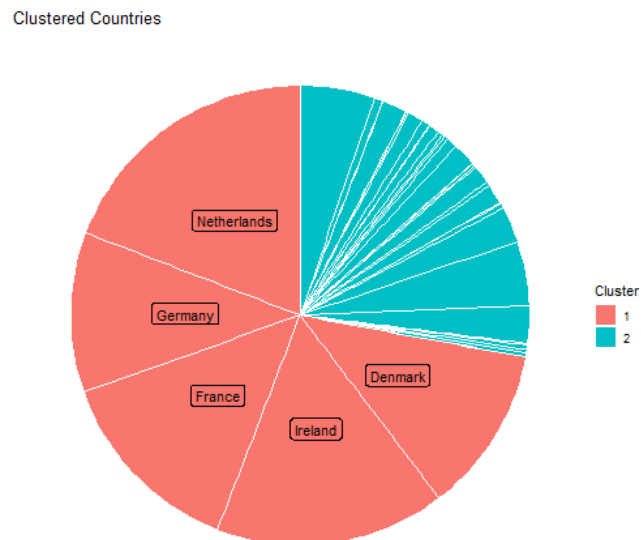


Figure 10: Clustered Countries

Out of all countries outside of United Kingdom, only 5 countries namely Netherlands, Germany, France, Ireland and Denmark are included in cluster 1 and all other countries are included in cluster 2. Interesting countries of cluster 2 is responsible for more than 70% of the total demands outside of United Kingdom. The reasons for this were analyzed using geospatial visualization.

### 3.6 Geospatial Visualization

The countries outside of United Kingdom were marked on a map using geocoding technique. The countries were marked based on the previous cluster analysis.

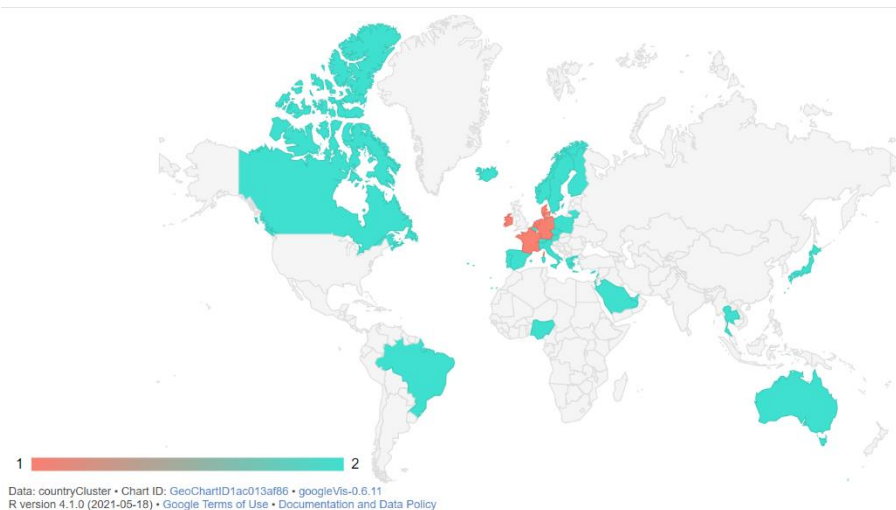
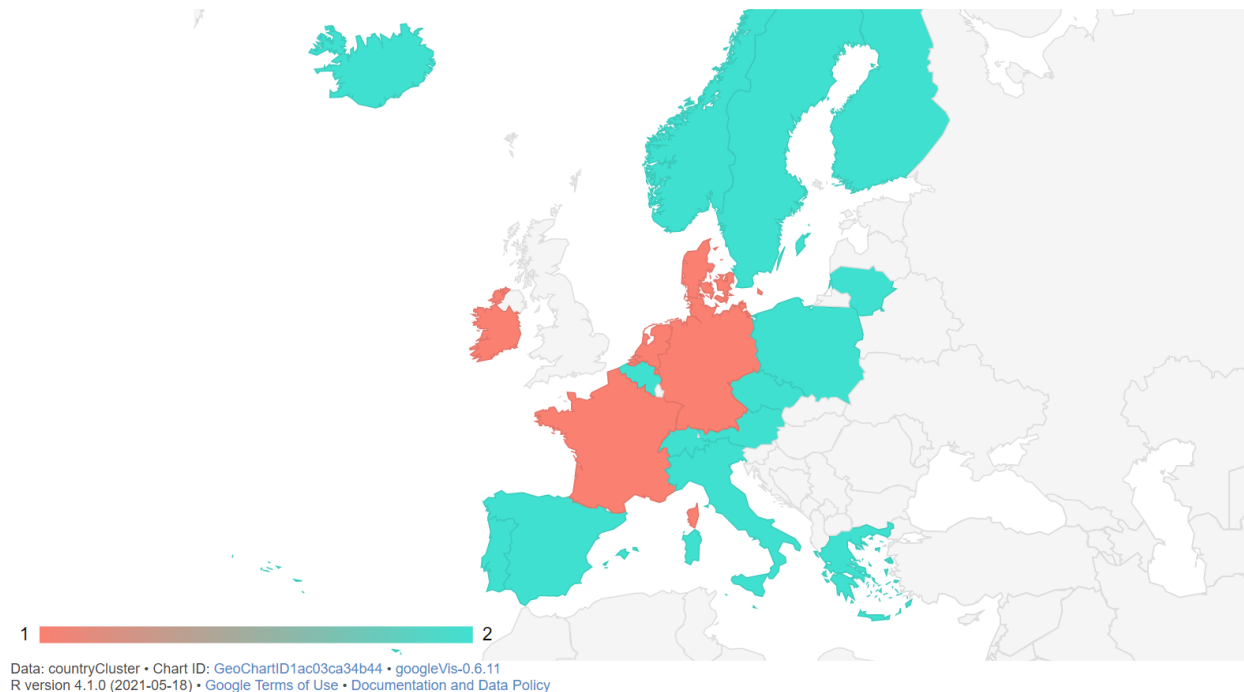


Figure 11: World Map Visualization

It is clear from the above map that majority of the countries in cluster 2 are the countries that are outside of Europe. One of the reasons for low demand among them could be because of the shipping cost. If the business owner needs to expand his business, the best locations could be Asia and the American continents. By this approach the customers within American, Asian and Australian continents will have low shipping cost hence sales could also improve massively.

There are some countries within Europe that belong to cluster 2. They are depicted in the map of Europe below.



*Figure 12: Map of Europe*

Compared to countries outside of Europe, countries within Europe must pay only a small shipping cost. Therefore, it is important to improve sales among countries within Europe before expanding the business in other continents. Countries such as Poland, Spain, Italy, Austria, Belgium, Switzerland, Sweden, Norway Finland and Greece are included in the cluster 2. New customers from these countries can be attracted with proper usage of digital marketing targeting those specified countries to improve sales.



### 3.7 Impact of Country's Temperature on Sales

It was important to check whether there is any significant effect on sales by united kingdom's temperature. For this purpose, average monthly temperature of United Kingdom from 2019-01 to 2020-12 was used.

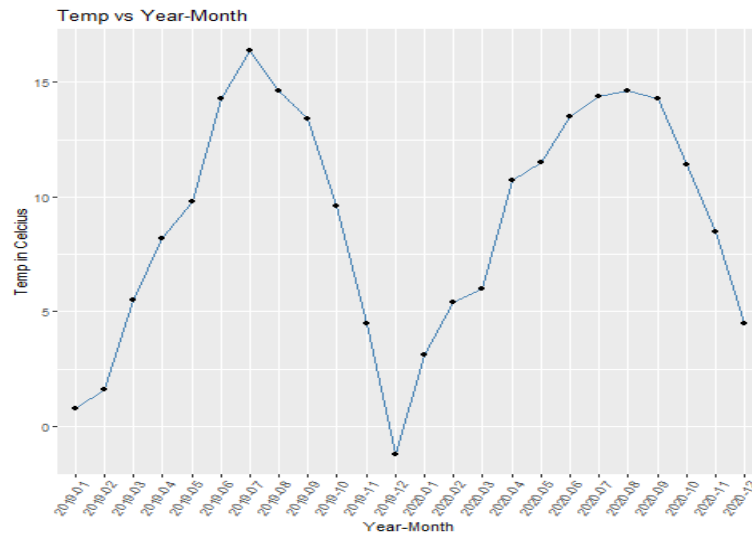


Figure 13: Average Monthly Temperature

To compare the monthly temperature and monthly sales pattern, both variables were normalized and plotted on the same graph.

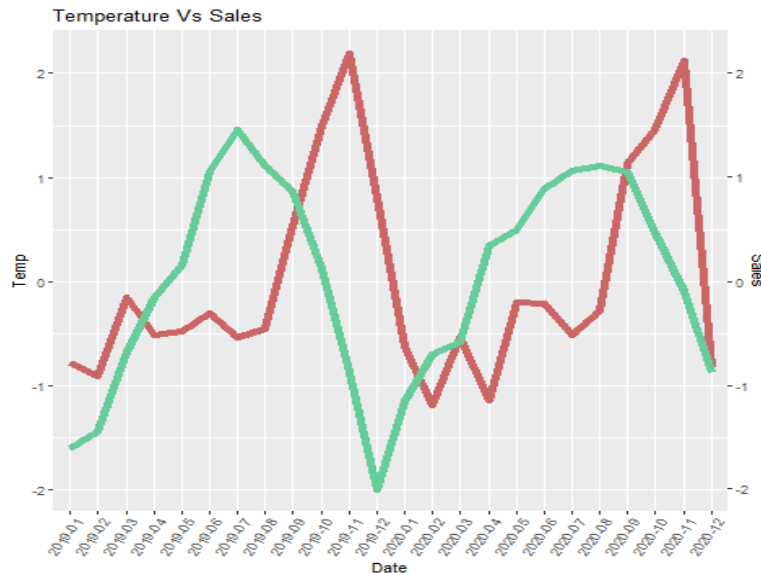


Figure 14: Monthly Sales vs Temperature

From the above line plot, it seemed that a similar pattern is shared between them. Hence a Pearson's correlation was conducted to test whether the correlation between monthly sales and monthly is significant. The correlation test results were as follows:

Table 10: Pearson's Correlation Test Result

Correlation Estimate	t- Value	Degrees of Freedom	p-Value	95% C.I
0.04121	0.19347	22	0.8484	(-0.368,0.437)

Even though the line plots of two variables seems to have some pattern, the Pearson's correlation between the variables was estimated as 0.04121. The p-Value of the test is above 0.05, hence the null hypothesis of 'true correlation is zero' at 5% significance level cannot be rejected. This concludes that there is no significant correlation present between the monthly sales and monthly average temperature of United Kingdom.

### 3.8 Time Series Analysis

Time series analysis on monthly sales was performed to fit a model and hence to forecast sales for the next 2 years. Initially, monthly sales were plotted to identify any patterns within the data.

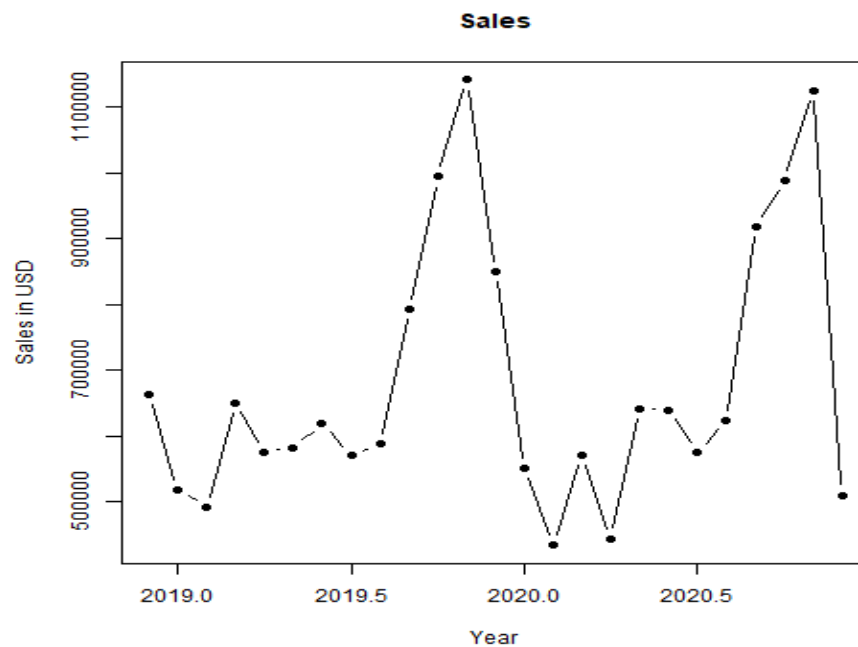


Figure 15: Monthly Sales

From the above plot, there seems to be increase in sales from September to November in both years. This could imply that a season effect is present in the sales series. Further it does not seem to contain any trend in both years. These components can be further analyzed using the auto correlation plot and the partial auto correlation plots.

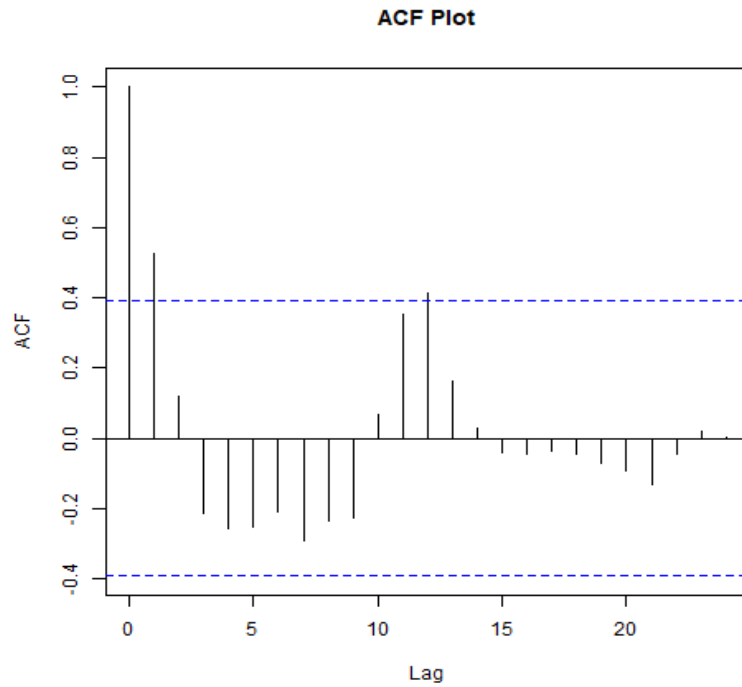


Figure 16: ACF Plot for Sales

The auto correlation function plot has significant spikes at lag 1 and lag 12. This further assures that a seasonal effect is present in the data. The significant spike is only at the first seasonal lag and it is insignificant at the second seasonal lag.

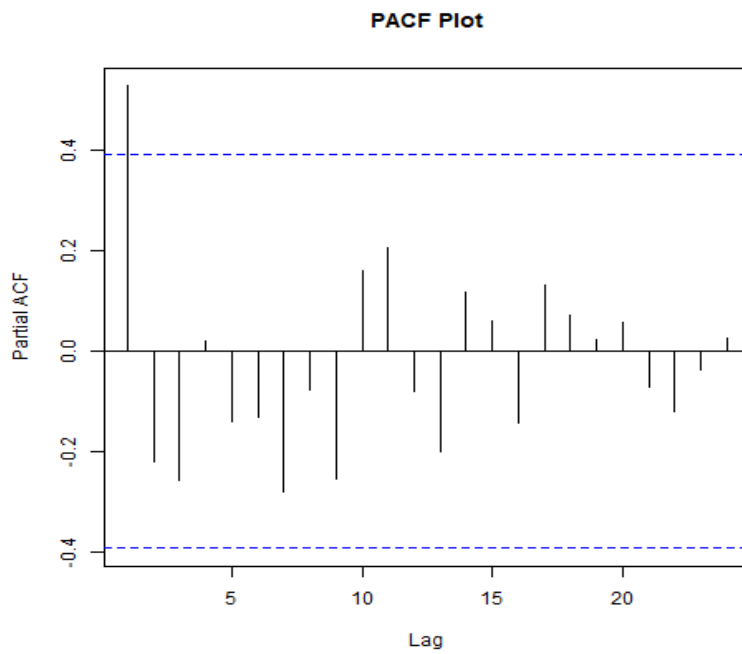


Figure 17: PACF Plot for Sales

The partial auto correlation plot does not consist of any significant lags. Hence using the above two plots we can hypothesize that the fitted model will consist of a seasonal element, moving average component of order 0 or 1 and the auto regressive component with an order of 0.

The model was fitted using `auto.arima()` function and the results showed that the fitted model is a seasonal random walk.

*Table 11: Auto Arima Results*

<b>Model</b>	<b>AIC</b>	<b>BIC</b>
ARIMA(0,0,0)(0,1,0)[12]	343.81	344.38

For further validation, a diagnostic checking was performed using Box-Pierce test and Ljung-Box test. The tested hypothesis :

$H_0$  : Residuals follow a white noise series

$H_1$  : Residuals do not follow a white noise series

The test results were as follows:

*Table 12: Box-Pierce Test Results*

<b>X-Squared</b>	<b>Degrees of Freedom</b>	<b>p-Value</b>
0.36	1	0.5485

*Table 13: Box-Ljung Test Results*

<b>X-Squared</b>	<b>Degrees of Freedom</b>	<b>p-Value</b>
0.405	1	0.5245

In both tests, the resulted p-values were higher than 0.05, indicating that there is not enough evidence to reject  $H_0$ , so the fitted model is adequate. Since the fitted model is a seasonal random walk, the monthly sales are considered as independent from each other and it suggests that future movement in the series cannot be predicted using past movement. Therefore, the forecasting step was not carried out in this analysis.

## Conclusions and Recommendations

From the analysis performed on the products, it was evident that some products with the highest demand during the months of September to November had the least demand during the other months. (E.g. : Assorted Color Bird Ornament, White Hanging Heart Light T-Holder, etc.) It is recommended to have proper inventory management within the business to allocate the only necessary quantity depending on their demands for such products to avoid extra carrying costs. The product features that are popular among customers are also important identify new trends and patterns, hence it is recommended to look into the wordsclds formed using product descriptions of the years 2019 and 2020 to identify product features that are trendy. This will also help to introduce new products. From the association rule mining, products that have a higher chance of being chosen together by customers were identified. It would be better for sales if such products are advertised or promoted together, so it will urge the customers to buy them together rather than buying a single product. Another recommendation would be to use digital marketing within Europe to increase sales, specially targeting countries such as Poland, Spain, Italy, Austria, Belgium, Switzerland, Sweden, Norway Finland and Greece. These countries have lower demand compared to other countries within Europe, therefore by proper targeted marketing, the demand from these countries could increase because even the shipping costs would not be that expensive within Europe. For countries outside of Europe, it is better to have separate warehouses (expand the business) in Asia and American continents according to the visualized map, since it could reduce the shipping cost hence increasing sales in those countries. These are the conclusions that could be arrived from the performed analyses and the recommendations provided in this chapter could have a significant effect in the increase of sales, if implemented properly.

## References

- Richard A. Johnson, Dean W. Wichern. (2012). *Applied Multivariate Statistical Analysis* (6th ed.). Pearson Prentice Hall.
- Voortman, C. (2004). Global Logistics: Management. In C. Voortman. Juta.

# Appendix

## Statistical Analysis in the Context of Business Growth

MT-325

9/14/2021

```
library(anytime)
library(factoextra)
library(dplyr)
library(ggplot2)
library(gridExtra)
library(ggfortify)
library(forecast)
library(googleVis)
library(ggvis)
library(tm)
library(htmlwidgets)
library(ggpubr)
library(wordcloud2)
library(webshot)
library(fitdistrplus)
library(lubridate)
library(arules)

df <- read.csv('data.csv')
head(df)

str(df)
```

### Modification

```
df$InvoiceDate <- as.Date(df$InvoiceDate, '%m/%d/%Y')
df$Month <- as.numeric(format(df$InvoiceDate, '%m'))
df$Year <- as.numeric(format(df$InvoiceDate, '%Y'))
df$Date <- as.factor(format(df$InvoiceDate, '%Y-%m'))
df$InvoiceTotal <- df$Quantity*df$Price
df <- subset(df, select = -c(InvoiceDate))
```

### Total Sales Series

```
s1 <- subset(df, select = c(Date, InvoiceTotal))
s1 <- s1 %>% group_by(Date) %>% summarise_at(vars(InvoiceTotal), list(Sales = sum))
head(s1)
```

### Visualize Sales Performance

```
#png("Plot1.png")
ggplot(data=s1, aes(x=Date, y=Sales))+geom_line(aes(group=1), color="steelblue")+
  geom_point()+theme(axis.text.x=element_text(angle=60, hjust=1))+
  labs(x='Year-Month', y='Sales in USD')+
  ggtitle('Sales vs Year-Month')
#dev.off()
```

- It is evident from the above plot that there is a significant increase in sales from months of September to November.

- That is, there is a seasonal effect present in this time series.

### Comparing Sales with UK Monthly Average Temperature

```
df2 <- read.csv("rainfalltemp.csv")
```

#### Modifications

```
df2 <- df2 %>% dplyr::select(Year,Period,Temp)
df2$Period <- sapply(df2$Period,function(x) grep(paste("(?i)",x,sep=""),month.abb))
df2 <- df2 %>% filter(Period >= 1 & Period <=12)
df2$Period <- as.numeric(df2$Period)
df2 <- df2%>% mutate(Date = make_date(Year, Period))
df2$Date<- format(df2$Date,"%Y-%m")
```

#### Visualize Monthly Avg Temperature

```
#png('pLot18.png')
ggplot(data = df2, aes(x = Date, y = Temp), color = "red")+
  geom_line(aes(group=1),color="steelblue")+ geom_point()+
  theme(axis.text.x=element_text(angle=60, hjust=1))+
  labs(x='Year-Month',y='Temp in Celcius')+
  ggtitle('Temp vs Year-Month')
#dev.off()
```

#### Merge Monthly Sales and Month Temperature

```
merge <- merge(s1,df2,by="Date")

#png('pLot19.png')
ggplot(merge, aes(x=Date)) +
  geom_line( aes(y=scale(Sales),group=1), size=2, color="#CC6666") +
  geom_line( aes(y=scale(Temp), group=1), size=2, color="#66CC99") +
  scale_y_continuous(
    # Features of the first axis
    name = "Temp",
    # Add a second axis and specify its features
    sec.axis = sec_axis(~1*.,name="Sales")
  ) + theme(axis.text.x=element_text(angle=60, hjust=1)) +
  ggtitle('Temperature Vs Sales')
#dev.off()
```

#### Correlation Between Sales and Temperature

```
res <- cor.test(merge$Sales, merge$Temp,
               method = "pearson")
res
```

- The correlation between temperature is not significant
- The p-value for pearson correlation test is larger than 0.05
- We cannot reject the null hypothesis where it says that true correlation is equal to zero.

#### Identifying Best Selling Products From Sept to Nov 2019

```
one <- df %>% filter(Year==2019 & 8<Month & Month<12)
one <- one %>% group_by(StockCode) %>% summarise_at(vars(Quantity), list(Quantity = sum))
one <- one[order(-one$Quantity),]
one <- one[1:10,]

data1 <- df %>% filter(Year==2019 & 8<Month & Month<12)
data1 <- data1 %>% group_by(StockCode) %>% summarise_at(vars(InvoiceTotal), list(Sales = sum))
data1 <- data1[order(-data1$Sales),]
data1 <- data1[1:10,]
```

```

plot1 <- ggplot(one,aes(x=reorder(StockCode,Quantity),y=Quantity))+
  geom_bar(stat = "identity",color="steelblue", fill=rgb(0.1,0.4,0.5,0.7), width = 0.6)+
  ggtitle('Top Demand 2019')+
  coord_flip()+
  labs(x='Product Code')

plot2 <- ggplot(data1, aes(x=reorder(StockCode,Sales), y=Sales)) +
  geom_bar(stat = "identity",color="steelblue", fill=rgb(0.1,0.4,0.5,0.7), width = 0.6) +
  ggtitle('Top Sales 2019')+
  coord_flip() +
  labs(x='Product Code')

##png("plot2.png")
#plot1
##dev.off()

#png('plot3.png')
#plot2
#dev.off()

grid.arrange(plot1, plot2, ncol=2, top = '2019')

```

#### Highest Demand Products Between Sep to Nov 2019

```

df %>% filter(Year==2019 & 8<Month & Month<12) %>%
  group_by(StockCode,Description) %>%
  summarise_at(vars(Quantity), list(Quantity = sum)) %>%
  arrange(desc(Quantity)) %>%
  head(10)

```

#### Top Products with Highest Sales

```

df %>% filter(Year==2019 & 8<Month & Month<12) %>%
  group_by(StockCode,Description) %>%
  summarise_at(vars(InvoiceTotal), list(Sales = sum)) %>%
  arrange(desc(Sales)) %>%
  head(10)

```

#### Best Selling Products with the Most Revenue From Sep to Nov 2019

```

products2019 <- intersect(one$StockCode,data1$StockCode)
Products1 <- unique(subset(df, select = c(StockCode, Description))) %>%
  filter(if_any(StockCode, `%in%`, products2019 ))
Products1

```

#### Identifying Best Selling Products From Sept to Nov 2020

```

two <- df %>% filter(Year==2020 & 8<Month & Month<12)
two <- two %>% group_by(StockCode) %>% summarise_at(vars(Quantity), list(Quantity = sum))
two <- two[order(-two$Quantity),]
two <- two[1:10,]

data2 <- df %>% filter(Year==2020 & 8<Month & Month<12)
data2 <- data2 %>% group_by(StockCode) %>% summarise_at(vars(InvoiceTotal), list(Sales = sum))
data2 <- data2[order(-data2$Sales),]
data2 <- data2[1:10,]

plot3 <- ggplot(two, aes(x=reorder(StockCode,Quantity), y=Quantity)) +
  geom_bar(stat = "identity",color="steelblue", fill=rgb(0.1,0.5,0.3,0.7), width = 0.6) +
  ggtitle('Top Demand 2020') +

```



```

coord_flip() +
labs(x='Product Code')

plot4 <- ggplot(data2, aes(x=reorder(StockCode,Sales), y=Sales)) +
  geom_bar(stat = "identity",color="steelblue", fill=rgb(0.1,0.5,0.3,0.7), width = 0.6) +
  ggtitle('Top Sales 2020') +
  coord_flip() +
  labs(x='Product Code')

png('plot4.png')
plot3
dev.off()

png('plot5.png')
plot4
dev.off()

grid.arrange(plot3, plot4, ncol=2, top = '2020')

```

### Highest Demand Products Between Sep to Nov 2020

```

df %>% filter(Year==2020 & 8<Month & Month<12) %>%
  group_by(StockCode,Description) %>%
  summarise_at(vars(Quantity), list(Quantity = sum)) %>%
  arrange(desc(Quantity)) %>%
  head(10)

```

### Top Products with Highest Sales

```

df %>% filter(Year==2020 & 8<Month & Month<12) %>%
  group_by(StockCode,Description) %>%
  summarise_at(vars(InvoiceTotal), list(Sales = sum)) %>%
  arrange(desc(Sales)) %>%
  head(10)

```

### Best Selling Products with the Most Revenue From Sep to Nov 2020

```

products2020 <- intersect(two$StockCode,data2$StockCode)
Products2 <- unique(subset(df, select = c(StockCode, Description))) %>%
  filter(if_any(StockCode, `%in%`, products2020 ))
Products2

```

- The best products for the seller during this season is products related to Christmas decorations.
- Jumbo bags have been one of the best products during both years 2019 and 2020 during the Christmas season, hence there is an opportunity for the seller to increase sales if new variants of Jumbo bags are introduced in 2021.

### Text Analysis to Popular Products in 2019 From Sep to Nov

```

des2019 <- df %>% filter(Year==2019 & 8<Month & Month<12) %>% pull(Description)

docs1 <- Corpus(VectorSource(des2019))
docs1 <- docs1 %>%
  tm_map(removeNumbers) %>%
  tm_map(removePunctuation) %>%
  tm_map(stripWhitespace)
docs1 <- tm_map(docs1, content_transformer(tolower))
docs1 <- tm_map(docs1, removeWords, stopwords("english"))

dtm <- TermDocumentMatrix(docs1)
matrix <- as.matrix(dtm)

```

```

words <- sort(rowSums(matrix),decreasing=TRUE)
des_2019 <- data.frame(word = names(words),freq=words)

set.seed(1234)
cloud1 <- wordcloud2(data = des_2019, size = 0.7, color = 'random-dark', shape = 'pentagon')
#saveWidget(cloud1,"1.html",selfcontained = F)
#webshot::webshot("1.html","plot15.png",vwidth = 1992, vheight = 1744, delay =10)
cloud1

```

### Text Analysis to Popular Products in 2020 From Sep to Nov

```

des2020 <- df %>% filter(Year==2020 & 8<Month & Month<12) %>% pull(Description)

docs2 <- Corpus(VectorSource(des2020))
docs2 <- docs2 %>%
  tm_map(removeNumbers) %>%
  tm_map(removePunctuation) %>%
  tm_map(stripWhitespace)
docs2 <- tm_map(docs2, content_transformer(tolower))
docs2 <- tm_map(docs2, removeWords, stopwords("english"))

dtm <- TermDocumentMatrix(docs2)
matrix <- as.matrix(dtm)
words <- sort(rowSums(matrix),decreasing=TRUE)
des_2020 <- data.frame(word = names(words),freq=words)

set.seed(1234)
cloud2 <- wordcloud2(data = des_2020, size = 0.7, color = 'random-dark', shape = 'pentagon')
#saveWidget(cloud2,"2.html",selfcontained = F)
#webshot::webshot("2.html","plot16.png",vwidth = 1992, vheight = 1744, delay =10, zoom = 2.5)
cloud2

```

- From the text analysis, we can see the preferred types of product designs and colors that are popular among the customers in this season.

### Clustering of Products

```

dis <- df %>% filter(Month <= 8 | Month >=12) %>% group_by(Description) %>%
  summarise(Quantity=sum(Quantity))

grp <- dis[,2]

#png('plot20.png')
fviz_nbclust(grp, kmeans, method = "wss") +
  geom_vline(xintercept = 4, linetype = 2)+
  labs(subtitle = "Elbow method")
#dev.off()

set.seed(100)
kmean <- kmeans(grp, 4)
dis <- cbind(dis, data.frame(Cluster = kmean$cluster))

cls <- dis %>% group_by(Cluster) %>% summarise_at(vars(Quantity), list(Demand = sum))

cls <- cls %>% arrange(desc(Demand))
cls$Cluster <- as.factor(cls$Cluster)
png('Plot17.png')
ggplot(cls,aes(Cluster,Demand,fill=Cluster)) +
  geom_bar(stat="identity") +
  theme(legend.position="none") +
  ggtitle('Cluster vs Demand')
dev.off()

```

- Products belonging to cluster 3 has the lowest demand.
- It is better to understand what these products are

#### Lowest Demanding Products

```
des <- dis %>% filter(Cluster==3)
des
```

#### Association rule mining

```
ruledf <- df %>% filter(8<Month & Month<12)

trans <- as(split(ruledf[, "Description"], ruledf["Invoice"]),
            "transactions")

rules <- apriori(trans, parameter = list(supp=0.02, conf=0.7,
                                         maxlen=3,
                                         target="rules"))

inspect(rules)
```

## Cluster the Months According to the Sales

```
s2 <- subset(df, select = c(Month, Quantity, InvoiceTotal))
s2 <- s2 %>% group_by(Month) %>% summarise_at(vars(Quantity, InvoiceTotal), list(Final = sum))

s2$Month <- as.factor(month.abb[s2$Month])
grp <- s2[, 2:3]
set.seed(100)
kmean <- kmeans(grp, 3)
s2 <- cbind(s2, data.frame(Cluster = kmean$cluster))
s2
```

## Clustering of Months

```
s2 <- s2 %>% group_by(Cluster, Month) %>% summarise_at(vars(InvoiceTotal_Final), list(Sales = sum))

s2$Cluster <- as.factor(s2$Cluster)

#png('plot6.png')

ggplot(s2, aes(x = "", y = Sales, fill = Cluster)) +
  geom_col(color = "white") +
  geom_label(aes(label = Month),
             position = position_stack(vjust = 0.5),
             show.legend = FALSE) + ggtitle('Clustered Months') +
  coord_polar(theta = "y") +
  theme_void()

#dev.off()
```

- Months clustered as cluster 1 are the best performing months while the months clustered as cluster 3 are moderately performing months.
- January, February and April are the months clustered as cluster 2, they are worst performing months according to the clustering.
- It is better if the seller use better targeted marketing strategies during the months of January, February and April to increase sales.

## Sales Forecasting Model

```
Market <- ts(s1$Sales, frequency = 12, start = c(2018, 12))
n = length(Market)
#png('plot7.png')
plot.ts(Market, type='b', xlab='Year', ylab='Sales in USD', main='Sales')
points(Market, pch=16)
#dev.off()
```

- It is evident from the plot that there is seasonal effect present in the data

## ACF Plot

```
acf <- acf(Market, lag = n-1, plot = FALSE)
acf$lag <- acf$lag*12
par(mar=c(4, 4, 3, 2))
#png('plot8.png')
plot(acf, main='ACF Plot')
#dev.off()
```

- Only the 1st and the 12th lags is significant further implying that there is a seasonal effect

#### PACF Plot

```
pacf <- pacf(Market, lag = n-1, plot = FALSE)
pacf$lag <- pacf$lag*12
par(mar=c(4,4,3,2))
#png('plot9.png')
plot(pacf, main = 'PACF Plot')
#dev.off()
```

#### Model Identification and Parameter Estimation

```
fit.auto <- auto.arima(Market)
fit.auto
```

#### Diagnostic Checking

```
res.auto <- residuals(fit.auto)
Box.test(res.auto, type = 'Box-Pierce')
Box.test(res.auto, type = 'Ljung-Box')
```

- H0:Residuals follow a white noise series VS H1:Residuals do not follow a white noise series.
- Both test have non significant p-values, hence we cannot reject H0 at 5% significance level.

#### Forecasting

```
#png('plot10.png')
plot(forecast(fit.auto))
#dev.off()
```

- Sales forecasts for next 3 years are shown as a blue line while the shaded areas are the 80% and 95% prediction intervals

#### Clustering of Countries by Demand

```
a <- subset(df, select = c(Quantity, Country))
a <- a %>% group_by(Country) %>%
  summarise_at(vars(Quantity), list(Final = sum)) %>%
  filter(Country != 'United Kingdom')
grp <- a[,2]

#png('plot11.png')
fviz_nbclust(grp, kmeans, method = "wss") +
  geom_vline(xintercept = 2, linetype = 2) +
  labs(subtitle = "Elbow method")
#dev.off()

set.seed(100)
kmean <- kmeans(grp, 2)
a <- cbind(a, data.frame(Cluster = kmean$cluster))

a <- a %>% group_by(Cluster, Country) %>% summarise_at(vars(Final), list(Demand = sum))

a$Country[a$Country == 'EIRE'] <- 'Ireland'
countryCluster = a
a$Cluster <- as.factor(a$Cluster)
a$Country[a$Cluster == 2] <- NA
```

```
#png('plot12.png')
ggplot(a, aes(x = "", y = Demand, fill = Cluster))+
  geom_col(color = "white") +
  geom_label(aes(label = Country),
                                                    posi
tion = position_stack(vjust = 0.5),
                                                    show
.legend = FALSE)+
  ggtitle('Clustered Countries') +
  coord_polar(theta = "y") +
  theme_void()
#dev.off()
```

\*The countries in cluster 1, namely Netherlands, Germany, France, Ireland and Denmark have the highest demands out of United Kingdom. The demand from cluster 1 countries is 72% of the total demand.

\*The lowest demand countries are in 2nd cluster. They only contribute 28% for the total demand.

### Geospatial Visualization

```
G1 <- gvisGeoChart(data=countryCluster, locationvar = 'Country',
                  colorvar = 'Cluster',
                  sizevar = 'Demand',
                  options = list(width="1100px", height="500px", colorAxis="{values:[1,2], colors : ['salmon','turquoise']}"))
plot(G1)
```

- According to the plotted map, it can be seen that highest demand for products is among countries in Europe. The reason for this could be the low shipping cost.
- Other main countries outside of Europe are Canada, Brazil, Thailand, Saudi Arabia, Australia and Japan.
- So the best locations for the seller to open warehouses could be the American and the Asian regions.
- We could use digital marketing to increase the demand among the countries in Europe that are among low demand countries.

### Visualization of Countries in Europe Based on Demand

```
G2 <- gvisGeoChart(data=countryCluster, locationvar = 'Country',
                  colorvar = 'Cluster',
                  sizevar = 'Demand',
                  options = list(region='150', width="1200px", height="550px", colorAxis="{values:[1,2], colors : ['salmon','turquoise']}"))
plot(G2)
```

- According to this map, it is possible to set up 3 separate digital marketing campaigns.
- One campaign to target countries Norway, Sweden and Finland
- Second campaign to cover countries Poland, Czech Republic, Lithuania and Italy.

- Finally another campaign to promote product in countries Spain and Portugal

```
df4=subset(df, Country=="United Kingdom")

uk<- subset(df4, select = c(Description, Quantity))
uk<- uk %>% group_by(Description) %>% summarise_at(vars(Quantity), list(Final = sum))
uk

uk_1=subset(uk, Final>10000)
grp <- uk_1[,2]
grp
set.seed(100)
kmean <- kmeans(grp, 5)
uk_1<- cbind(uk_1, data.frame(Cluster = kmean$cluster))
head(uk_1)

uk_2 <- uk_1 %>% group_by(Cluster) %>% summarise_at(vars(Final), list(Total_Demand = sum))
uk_2$Cluster <- as.factor(uk_2$Cluster)
```

*cluster 1 includes the items which have demands between 20,000 and 30,000 cluster 2 includes the items which have demands between 14,500 and 20,000 cluster 3 includes the items which have demands between 70,000 and 85,000 cluster 4 includes the items which have demands between 30,000 and 40,000 \*cluster 5 includes the items which have demands between 10,000 and 14,500*

\*But according to the cluster contribution highest total demand is from cluster 5 and lowest total demand is from cluster 4.

```
uk_3<- subset(df4, select = c(Description, Quantity, InvoiceTotal))
uk_3<- uk_3 %>% group_by(Description) %>% summarise_at(vars(InvoiceTotal), list(Final = sum))

uk_4=subset(uk_3, Final>10000)
grp <- uk_4[,2]
grp
set.seed(100)
kmean <- kmeans(grp, 5)
uk_5<- cbind(uk_4, data.frame(Cluster = kmean$cluster))

uk_6 <- uk_5 %>% group_by(Cluster) %>%
  summarise_at(vars(Final), list(InvoiceTotal = sum))

uk_6$Cluster <- as.factor(uk_6$Cluster)

ggplot(uk_6, aes(x = "", y = InvoiceTotal, fill = Cluster))+geom_col(color = "white") +
  geom_label(aes(label = Cluster), position = position_stack(vjust = 0.5), show.legend = FALSE)+
  ggtitle('Cluster Contribution') +
  coord_polar(theta = "y") +
  theme_void()
```

*cluster 1 includes the items which have Invoice Total between 10,000 and 17,000 cluster 2 includes the items which have Invoice Total between 17,000 and 30,000 cluster 3 includes the items which have Invoice Total higher than 120,000 cluster 4 includes the items which have Invoice Total between 50,000 and 120,000 \*cluster 5 includes the items which have Invoice Total between 30,000 and 50,000*

\*Cluster 1 and 2 contribute more to the total income and cluster 5 also contributes significantly to the total income.

\*At least contribute cluster to the total income is cluster 4.