MONASH University

ETC5147 Data Exploration Assignment

Prepared by Arpan Sarkar, 32559844, asar0035@student.monash.edu

Introduction

With the data set related to global supply chain, an analytical study is made to find out causes that lead to profit making or incurring losses by exporting nations. In order to find out so, the factors of performance in this sectors have been studied. Analysis was done for pin pointing the category of items that are better for inter country business than some other categories. Also study focused on the better performing nations in export business than others. Although data pertains to particular period, but is indicative of the performances and the trend in the global market. With data wrangling, data checking to creating related tables and visualizations, the final observations are produced as conclusion and the questions that were raised as an objective to this project are answered. Due to constraints the top 10 nations data are taken for analysis.

Data Wrangling

The data has been downloaded from Mendeley Data - DataCo SMART SUPPLY CHAIN FOR BIG DATA ANALYSIS , it has 180K rows x 53 columns and as the data set is large, I dropped few columns, filtered top 10 performing nations and year in (2015,2016,2017 and 2018) and then took 1 subset of the main data of 50238 rows, apart from this the date and time column was in character type, so changed the type of column using mutate function and lubridate package. Some snapshots of the process are given below.

Step1

```
```{r}
supply<- read.csv("Data/DataCoSupplyChainDataset.csv")

head(supply,40000)
```
```

Description: df[,53] [40,000 × 53]

	Type <chr>	Days.for.shipping..real. <int>	Days.for.shipment..scheduled. <int>
1	DEBIT	3	4
2	TRANSFER	5	4
3	CASH	4	4
4	DEBIT	3	4
5	PAYMENT	2	4
6	TRANSFER	6	4
7	DEBIT	2	1
8	TRANSFER	2	1
9	CASH	3	2
10	CASH	2	1

1-10 of 40,000 rows | 1-4 of 53 columns Previous 1 2 3 4 5 6 ... 100 Next

Step2

```
```{r}
n <- nrow(supply)
a <- split(supply,rep(1:ceiling(n/90260),each=90260)[1:n])
a
```
```

R Console data.frame 90260 × 53 data.frame 90259 × 53

Description: df[,53] [90,259 × 53]

	Type <chr>	Days.for.shipping..real. <int>	Days.for.shipment..scheduled. <int>
90261	PAYMENT	2	2
90262	PAYMENT	5	2
90263	PAYMENT	6	2
90264	PAYMENT	5	2
90265	PAYMENT	6	2
90266	PAYMENT	4	2
90267	PAYMENT	3	2
90268	PAYMENT	4	2
90269	PAYMENT	3	2
90270	PAYMENT	6	2

1-10 of 90,259 rows | 1-4 of 53 columns Previous 1 2 3 4 5 6 ... 100 Next

Step3

```r
col_remove<- c("Customer.Email", "Customer.Fname", "Customer.Id", "Customer.Lname",
"Customer.Password", "Order.Id", "Order.Item.Cardprod.Id", "Order.Customer.Id", "Late_delivery_risk"
"Category.Id", "Product.Image", "Product.Status", "Order.Zipcode", "Product.Card.Id",
"Product.Category.Id", "Product.Description")

supply1<- a$`1`%>% mutate(mdy_hm(order.date..DateOrders.))%>% filter(Order.Country  %in% c("Estados
Unidos", "Francia", "México", "Alemania", "Brasil", "Australia","Reino Unido", "China", "Italia",
"India"), year(mdy_hm(order.date..DateOrders.)) %in% c(2015,2016,2017,2018) )%>%
  select(- one_of(col_remove))
dim(supply1)
```

```
[1] 50238    38
```

Data Checking

Before doing analysis I did a bit of checking on the dataset so that I can find possible errors or any unusual behavior which could cause some error in finding answers of the questions.

Table 1:Missing Value Summary

A tibble: 37 x 3

variable <chr>	n_miss <int>	pct_miss <dbl>
Customer.Zipcode	1	0.002222222
Type	0	0.000000000
Days.for.shipping..real.	0	0.000000000
Days.for.shipment..scheduled.	0	0.000000000
Benefit.per.order	0	0.000000000
Sales.per.customer	0	0.000000000
Delivery.Status	0	0.000000000
Category.Name	0	0.000000000
Customer.City	0	0.000000000
Customer.Country	0	0.000000000

1-10 of 37 rows Previous 1 2 3 4 Next

From the above table **Table 1** we can tell that there were no as such missing values in the data-set except for 1 variable (Customer.Zipcode) which is having only 1 missing value, rest all the values are present in the data-set.

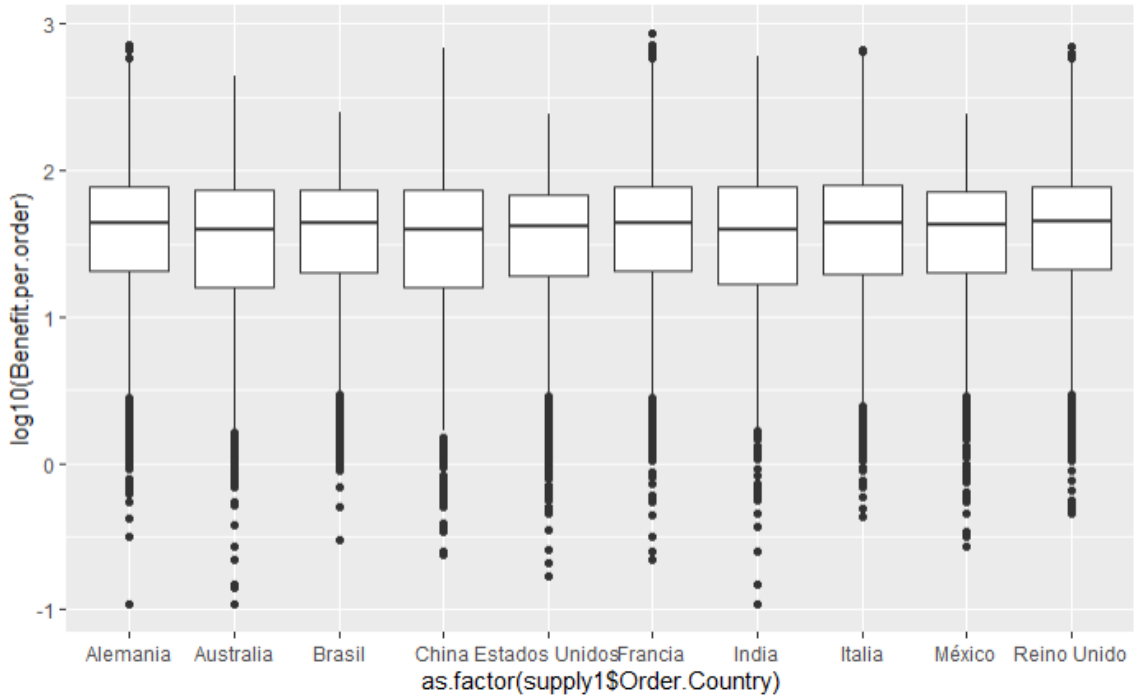


Figure 1: Box-Plot

- The above Box-Plot is created to visualize the data distribution pertaining to the most important dependent variable i.e benefits per order. In the above plot we can see that there are some outliers in all the countries in terms of their profit/loss, however in Australia, Brazil, Italy and Reindo Unido (United Kingdom) there are no outliers in terms of profit per order but there are some outliers in other countries in terms of profit per order.
- In terms loss per order we find that Alemania (Germany), Australia and India have incurred higher loss in one order item.
- For our analysis these outliers will not affect to draw the output of the results.

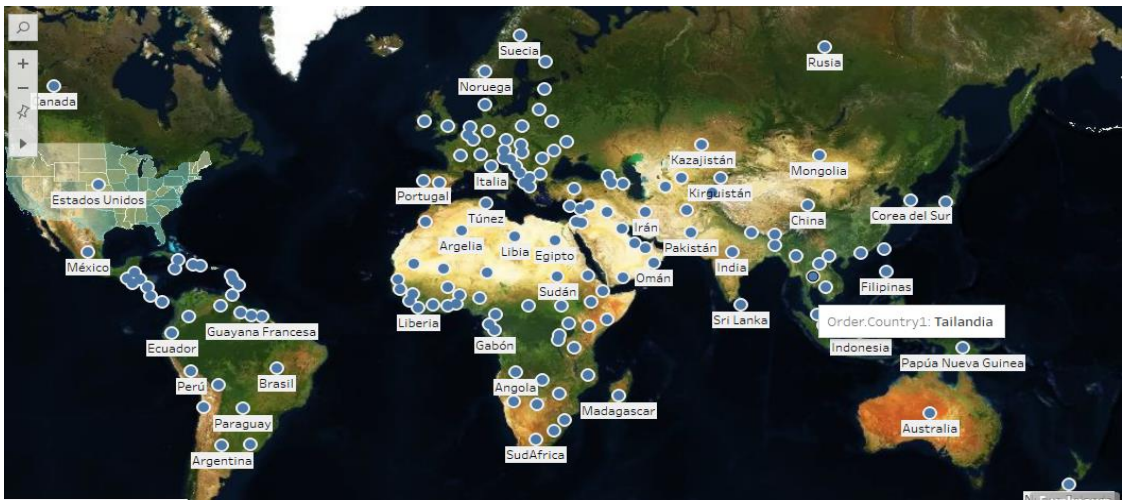


Figure 2: Map Visual

When I look to the spatial points given in above map I didn't find any error as such, it was distributed accurately on the map visual.

Data Exploration

```
supply1 %>% mutate(delay_in_shipment=  supply1$Days.for.shipping..real.-
supply1$Days.for.shipment..scheduled.)%>% group_by(Order.Country) %>%
   summarise(m1 = min(delay_in_shipment, na.rm=TRUE),
                m2 = max(delay_in_shipment, na.rm=TRUE),
                m3 = median(delay_in_shipment, na.rm=TRUE))%>%kbl(caption = "Table",
table.attr="style='width:70%;'") %>%
  kable_paper("hover",full_width = T,html_font = "Cambria", position= "left" )
```

Table

Order.Country	m1	m2	m3
Alemania	-2	4	1
Australia	-2	4	1
Brasil	-2	4	1
China	-2	4	1
Estados Unidos	-2	4	1
Francia	-2	4	1
India	-2	4	1
Italia	-2	4	1
México	-2	4	1
Reino Unido	-2	4	1

Table 2: Minimum, Maximum, Median

In the above table **Table 2** we can see that m1(minimum delay) was -2 days which was 2 days early for all the countries and m2(maximum delay) was 4 days in all the countries and median delay was 1 day in all the countries. From the table we can observe that there was no significant difference in terms of their delivery.

Figure 3:Cross-Tabs

Count OF Orders In Profit And Loss **Cross-Tab 2**

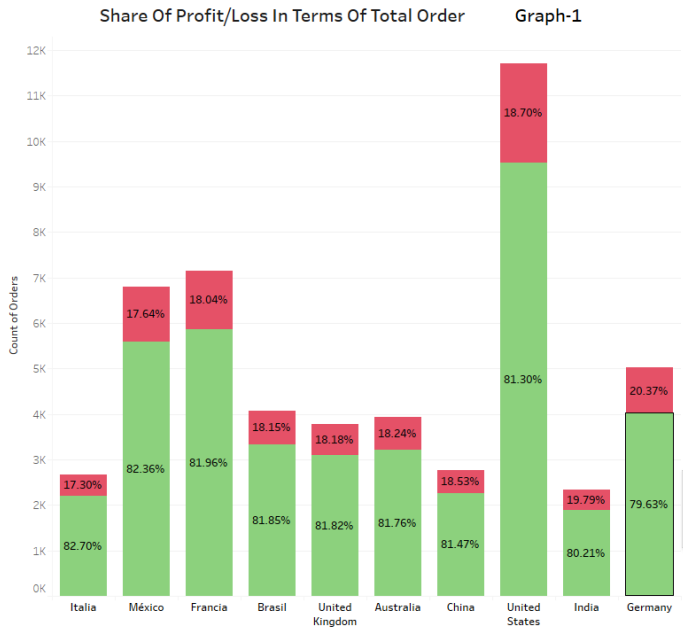
Order_Country	Profit/Loss: Loss	Profit/Loss: Profit
United States	2,189	9,515
Francia	1,291	5,866
México	1,198	5,594
Germany	1,024	4,004
Brasil	738	3,328
Australia	718	3,218
United Kingdom	687	3,091
China	513	2,255
Italia	461	2,203
India	464	1,881

Total Benefit/Loss Per Country **Cross-Tab 3**

Order_Country	Loss	Profit
United States	-219,950	450,120
Francia	-157,472	333,269
México	-126,662	283,998
Germany	-124,513	226,199
Brasil	-78,711	172,006
United Kingdom	-89,096	177,277
Australia	-84,004	160,083
Italia	-50,930	123,241
China	-59,824	117,702
India	-48,817	97,399

From above cross-tabs we can easily say that highest orders were from United States and the lowest from India. Further, total profit & loss against orders are shown country wise in cross tab 3 (above).

Figure 4: Graph-1 & Cross-Tab 5



Average Benefit Per Order Country Wise Cross-Tab 5

Order_Country	Total_Benefit	average benefit per order	Count of Order..
Australia	76,078	19	3,936
Brasil	93,295	23	4,066
China	57,878	21	2,768
Francia	175,797	25	7,157
Germany	101,686	20	5,028
India	48,582	21	2,345
Italia	72,310	27	2,664
México	157,336	23	6,792
United Kingdom	88,181	23	3,778
United States	230,170	20	11,704

With the data in the tables and one more table, the profit and loss share in different countries are visualized in the **Graph-1**, displayed above.

Figure 5: Graph-2, Cross-Tab 6 & 7

Average Benefit Per Order In Terms Of Shipping Mode
Cross-Tab 6

	Shipping.Mode1				
Order_Country	First Class	Same Day	Second Class	Standard Cla..	Grand Total
Italia	29.89	19.98	23.86	28.08	27.14
United Kingdom	21.14	29.55	16.16	26.11	23.34
México	21.29	17.16	21.03	25.02	23.16
Francia	29.30	12.86	27.70	23.21	24.56
China	29.64	21.16	13.83	21.10	20.91
Brasil	26.20	31.99	24.79	20.74	22.95
India	23.23	29.55	17.61	20.36	20.72
United States	19.54	15.42	21.14	19.56	19.67
Australia	25.76	17.33	16.18	19.23	19.33
Germany	24.72	20.92	24.52	17.33	20.22

Total Benefit Overall In Terms Of Shipping Mode
Graph-2

Share Of Standard Class Shipment Cross-Tab 7

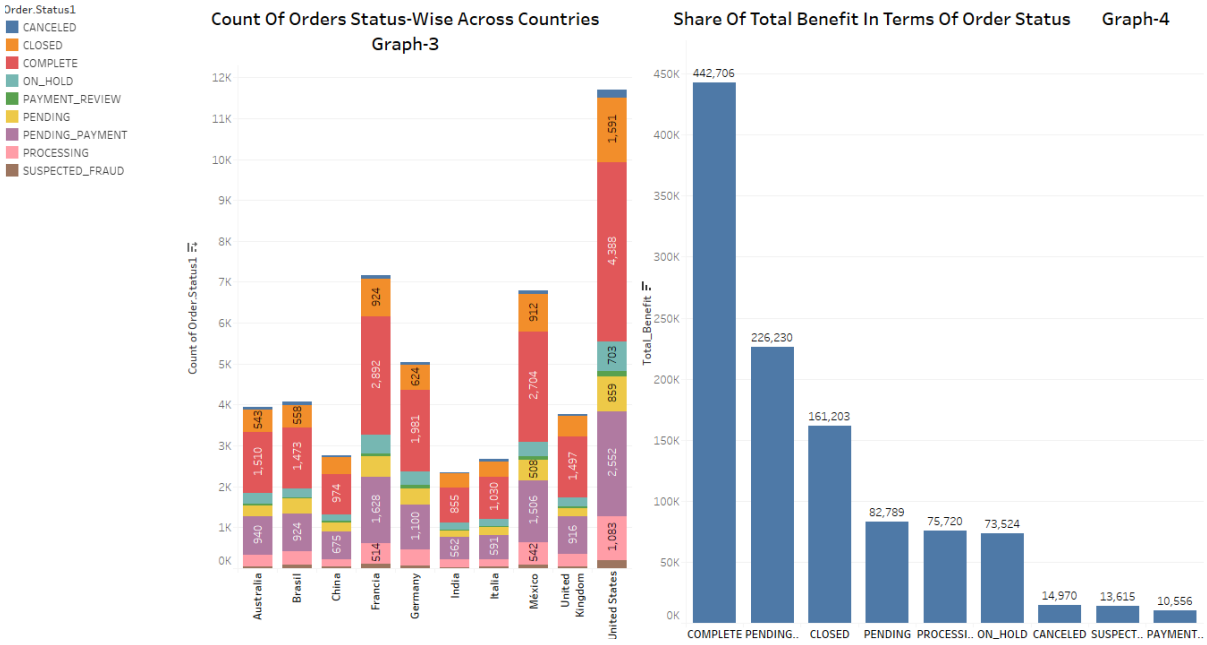
Order_Country	Percentage of standard class shipment	Count of Shipping.Mode1
Australia	55.3	3,936
Brasil	59.4	4,066
China	54.9	2,768
Francia	57.2	7,157
Germany	57.7	5,028
India	54.7	2,345
Italia	61.6	2,664
México	57.0	6,792
United Kingdom	57.1	3,778
United States	55.1	11,704

The table above is generated for understanding the role of shipping mode in analyzing performance in terms of delivery as well as cost.

The graph gives a clear picture of the shipping in terms of overall benefits. An eye opener for exporter to decide.

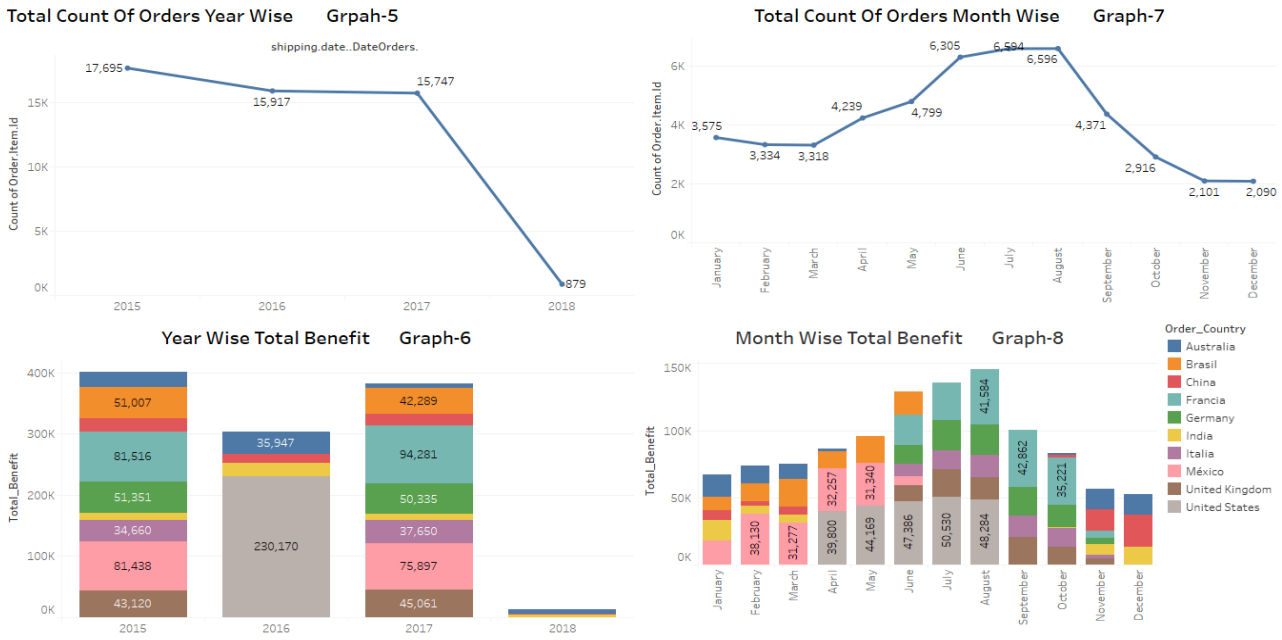
The table data table-7(above) is created for further ease of understanding the role of this factor.

Figure 6: Graph-3 & 4



Graph-3, above is created to show case data pertaining to one more very important factor of performance, i.e. order status. Ideally all orders should have only one status, i.e. COMPLETE, but in reality it is not. We found 9 distinct class of order-status. Country wise total count of orders showing status wise share is very appealing to understand the business in this sector. In **Graph-4** above, share of total benefits among various order status is displayed for understanding the trend as well.

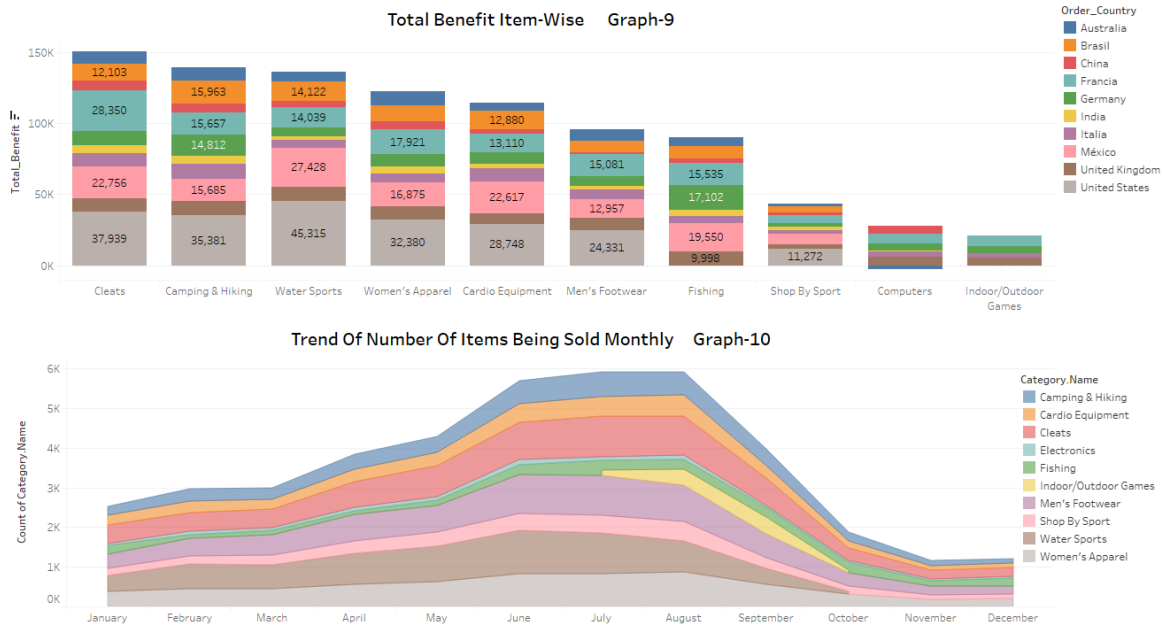
Figure 7:Graph-5,6,7 & 8



Graph-5 and **Graph-6** displayed above are created for understanding of another most vital factor's role in this business sector. This factor is the universal independent variable i.e. date variable(shipping date). Year wise total benefit for the top 10 countries in export business are placed through bar charts **Graph-6** and the line diagram **Graph-5** is drawn with total order data of 10 countries (year wise). It may be noted that for the year 2018, there is hardly any data, which could be for non-update of data set etc.

Similarly, **Graph-7** & **Graph-8** are also displayed above depicting the data trend on the basis of month of year. This is another big revelation to understand the right period of doing business of export through the faster mode. We find that mouth of June, July, August is right and most favorable time.

Figure 8: Graph-9 & 10



Conclusion

The in depth observations of Global Supply Chain System (GSCS) data and tables, graphs as created and produced give enough insight to conclude on the queries proposed as the objective of this project.

Q1) The factors that affect the performance in a GSCS

- 'Delay in Delivery' is a factor. From **Tab-1** (**above**)data we find that almost all nations are on right path as there is no extraordinary delay by any.
- Existence of 'Order' indicates existence of Demand. From **Cross-Tab-2 (above)**, we find loss in 2189 orders and profit in 9515 orders by USA and the same for India is 464 and 1881. This factor is indicative of business volume.
- **Cross-Tab-3(above)** indicates the volume of business by nations. USA makes a loss of $ 21950 and profit of $ 450120 while India's figure is $48817 & $97399 respectively. It shows that the loss is almost 50% of profit. 'Benefit per Order' is a factor.
- In contrast to above, in **Graph-1(above)**we find share of loss is 18.7% & profit is 81.3% for USA in terms of Total Order. The same for India is 19.79% and 80.21%. This indicates that more high valued orders are executed on loss in comparison to profit.
- 'Shipping Mode' is the other factor. From **Graph-2 (above)**shows that Overall Benefit is much higher for 'Standard Class' than any other class lie 'Second Class' or 'First Class' or 'Same Day'.
- 'Order Status' is another factor. From **Graph-3 (above)**, we find, share of 'Complete' orders is more than any other Status. For USA, France, Mexico the share is pretty more in comparison to other countries. Although 'Cancelled' Status exists, but its volume is nominal. **Graph-4** (**above**)shows the total benefit figure for different Order-Status with 'Complete' having a share of 442706 & 'Payment Review' has 10556 as the least. And the second best 'Pending Payment' is little more than 50% of 'Complete' status.
- **Graph-5**, **Graph-6**, **Graph-7** and **Graph-8** (**Figure 7:Graph-5,6,7 & 8**)are produced for indicating the role of the 'Order Date' factor. This variable is independent. In **Graph-5** / **Graph-6** we find an abrupt slump /fall for the year 2018, while it reveals almost similar count

of orders for the year 2015, 2017. Year 2016 is fully dominated by USA. Ear 2018 data seems to be missing or not updated. **Graph-7** and **Graph-8** show month wise order count trend. Data says, most business happens during June to August with least in December.

Q2) Why some items or countries fare better:

- **Graph-9 (above)** displays the item wise total benefit data. Here it is fond that 'Cleats' as item has the highest business, which is almost 150K.On the other hand 'Indoor and Outdoor games' have low business. Reason for the discrepancy could be for lack of demand from importing nations.
- In year wise benefit data (**Graph-6**), we find that business by countries are not proportionate year wise. In 2015 and 2017, USA data not available and hence a huge discrepancy.
- **Graph-8** gives an indicative favorable period of the year for doing business in a GSCS. June, July, August fond more favorable. But here also, it is found that not all countries follow it. Australia, China, India were not active during that period.

Q3) What is the current trend in the sector:

- Current trend can be similar to the level of Year **2016** after adding the average of **2015** & **2017** data for countries like UK, Mexico, Brazil, France, Germany, Italia, since for **2016** data for these countries found to be missing.
- **Standard Class** is likely to prevail as shipping mode.
- **June, July, August t**o be preferable period for business for a GSCS.

Reflection

Apart from usage of tools for data wrangling, checking and plotting of graphs etc with 'R' / Tablaeu, the need to do something new was always felt. The project itself is a reflection as right from beginning till this page, I am attached not as a solver but also as a creator of task. This is a new kind of task for me and provides a lot of insight and desire to do more.

Kind of data handled in this project is interesting. All types of data like Nominal, Cardinal, Interval, Ratio are handled here. Map plotting with spatial data, provided confidence for future and Graphical presentation with temporal data also very interesting that produced intelligent results.

Bibliography

Software Used	Links	Other Links And References
R Language	R: The R Project for Statistical Computing (r-project.org)	https://stackoverflow.com/
Tableau	https://public.tableau.com/en-us/s/	https://community.tableau.com/
Excel	https://www.microsoft.com/en-in/microsoft-365/excel	https://www.myexcelonline.com/blog/50-things-you-can-do-with-excel-power-query/

Guides	Links
Academic Report Writing Guidance	https://www.monash.edu/__data/assets/pdf_file/0005/506345/qmanual.pdf
Table And Graph Interpretation Guide.	Phrases and 6 Analysis Steps to interpret a graph (explainwell.org)
Report Structure Gudie	https://lms.monash.edu/mod/resource/view.php?id=8315222

- **The assignment work done is with kind support and guidance of my guide Sarah Goodwin (Lecturer and Chief Examiner).**
- **My R programming skills is with kind support and guidance of my tutors Bruno Luis Mendivez Vasquez and Angel Das.**