Decision Making Via Data Analytics

Final Project

Topic: Supply Chain analysis with EDA visualizations

Name: Harsh Shah

Guidance: Professor Amineh Zadbood

Introduction

Supply chain is a key measure to consider for any business. A business with good supply chain network has the capability to be pioneer in the industry. Supply Chains are the Totality of processes spanning operations from supplier to end-customer, focused on material, work and information flow. Supply chain optimization directly or indirectly affect the profit for any company. A Supply Chain is a set of three or more units (organizations or individuals) directly involved in the upstream and downstream flows of products, services, finances and/or information from a source to a customer. A Supply Chain is the alignment of firms that brings products or services to the market.

Some key characteristics of a supply chain are customer is an integral part of the supply chain, Includes movement of products from suppliers to manufacturers to distributors and information, funds, and products in both directions. For a company to analyze its supply chain various factors are consider here that are origin of the product, shipping information, cost for the product etc. Here taking example of an online retailer like Amazon has to analyze various factors like origin country, time for shipping, quantity for the product. This all factors affect both the profits and customer satisfaction.

Supply chain management is a crucial process because an optimized supply chain results in lower costs and a more efficient production cycle. Companies seek to improve their supply chains so they can reduce their costs and remain competitive.

Abstract

In this project a detailed analysis is provided for a whole supply chain network consisting of various factors like product information, shipping information, profit for the products, origin country for the product, Delivery status. This data is taken from Kaggle. This dataset is used by the company DataCo Global for analysis of supply chain and making decisions. This dataset contains mainly Fitness, Clothing, Book shop etc. This dataset contains about 180000 rows and 54 columns. The orders are from all over the countries across the world with different status like complete order, still in process etc. Here we have considered only orders from USA and with status of complete.

EDA which exploratory data analysis is presented here with the help of R software and various libraries like maps, mapproj, Ggplot2, data.table etc. Live and interactive graph are also presented with the help of GoogleVis library.

This report presents and answers various questions like:

1. What are number of order distributions across states of USA?

2. What is the profit analysis for different product departments?

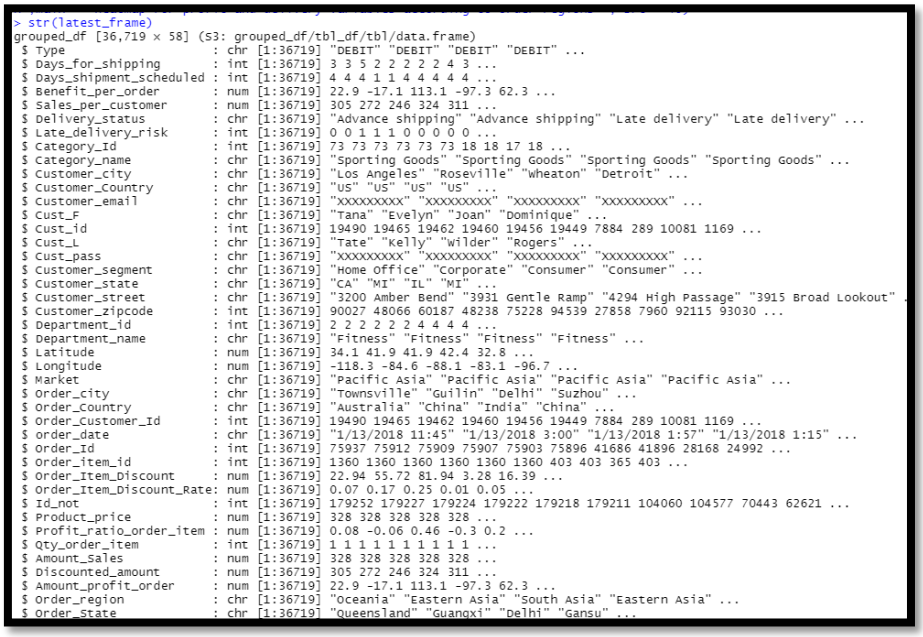
3. Does shipping have any affect on profit earned on the product?

4. Which origin area is responsible for late shipping?

Various other conclusions and observation were drawn from this report.

Data Validation

Here analysing the structure of data. There are total 54 columns with unique variables and there are in total 180000 rows of data. Below is attached the type distribution for each of the variable among the 47. Our dataset is divided in types of “num”, “chr”, “logi” and “int”. Description of each variable is taken from dataset reference website. Here dataset is taken from Kaggle and company DataCo. Figure 1 shows the sample data structure.



```
> str(latest_frame)
grouped_df [36,719 × 58] (S3: grouped_df/tbl_df/tbl/data.frame)
 $ Type                    : chr [1:36719] "DEBIT" "DEBIT" "DEBIT" "DEBIT" ...
 $ Days_for_shipping       : int [1:36719] 3 3 5 2 2 2 2 2 4 3 ...
 $ Days_shipment_scheduled : int [1:36719] 4 4 4 1 1 4 4 4 4 4 ...
 $ Benefit_per_order       : num [1:36719] 22.9 -17.1 113.1 -97.3 62.3 ...
 $ Sales_per_customer      : num [1:36719] 305 272 246 324 311 ...
 $ Delivery_status         : chr [1:36719] "Advance shipping" "Advance shipping" "Late delivery" "Late delivery" ...
 $ Late_delivery_risk      : int [1:36719] 0 0 1 1 1 0 0 0 0 0 ...
 $ Category_Id             : int [1:36719] 73 73 73 73 73 73 18 18 17 18 ...
 $ Category_name           : chr [1:36719] "Sporting Goods" "Sporting Goods" "Sporting Goods" "Sporting Goods" ...
 $ Customer_city           : chr [1:36719] "Los Angeles" "Roseville" "Wheaton" "Detroit" ...
 $ Customer_Country        : chr [1:36719] "US" "US" "US" "US" ...
 $ Customer_email          : chr [1:36719] "XXXXXXXXX" "XXXXXXXXX" "XXXXXXXXX" "XXXXXXXXX" ...
 $ Cust_F                  : chr [1:36719] "Tana" "Evelyn" "Joan" "Dominique" ...
 $ Cust_id                 : int [1:36719] 19490 19465 19462 19460 19456 19449 7884 289 10081 1169 ...
 $ Cust_L                  : chr [1:36719] "Tate" "Kelly" "Wilder" "Rogers" ...
 $ Cust_pass               : chr [1:36719] "XXXXXXXXX" "XXXXXXXXX" "XXXXXXXXX" "XXXXXXXXX" ...
 $ Customer_segment        : chr [1:36719] "Home Office" "Corporate" "Consumer" "Consumer" ...
 $ Customer_state          : chr [1:36719] "CA" "MI" "IL" "MI" ...
 $ Customer_street         : chr [1:36719] "3200 Amber Bend" "3931 Gentle Ramp" "4294 High Passage" "3915 Broad Lookout" ...
 $ Customer_zipcode        : int [1:36719] 90027 48066 60187 48238 75228 94539 27858 7960 92115 93030 ...
 $ Department_id           : int [1:36719] 2 2 2 2 2 2 4 4 4 4 ...
 $ Department_name         : chr [1:36719] "Fitness" "Fitness" "Fitness" "Fitness" ...
 $ Latitude                : num [1:36719] 34.1 41.9 41.9 42.4 32.8 ...
 $ Longitude               : num [1:36719] -118.3 -84.6 -88.1 -83.1 -96.7 ...
 $ Market                  : chr [1:36719] "Pacific Asia" "Pacific Asia" "Pacific Asia" "Pacific Asia" ...
 $ Order_city              : chr [1:36719] "Townsville" "Guilin" "Delhi" "Suzhou" ...
 $ Order_Country           : chr [1:36719] "Australia" "China" "India" "China" ...
 $ Order_Customer_Id       : int [1:36719] 19490 19465 19462 19460 19456 19449 7884 289 10081 1169 ...
 $ order_date              : chr [1:36719] "1/13/2018 11:45" "1/13/2018 3:00" "1/13/2018 1:57" "1/13/2018 1:15" ...
 $ Order_Id                : int [1:36719] 75937 75912 75909 75907 75903 75896 41686 41896 28168 24992 ...
 $ Order_item_id           : int [1:36719] 1360 1360 1360 1360 1360 1360 403 403 365 403 ...
 $ Order_Item_Discount     : num [1:36719] 22.94 55.72 81.94 3.28 16.39 ...
 $ Order_Item_Discount_Rate: num [1:36719] 0.07 0.17 0.25 0.01 0.05 ...
 $ Id_not                  : int [1:36719] 179252 179227 179224 179222 179218 179211 104060 104577 70443 62621 ...
 $ Product_price           : num [1:36719] 328 328 328 328 328 ...
 $ Profit_ratio_order_item : num [1:36719] 0.08 -0.06 0.46 -0.3 0.2 ...
 $ Qty_order_item          : int [1:36719] 1 1 1 1 1 1 1 1 1 1 ...
 $ Amount_Sales            : num [1:36719] 328 328 328 328 328 ...
 $ Discounted_amount       : num [1:36719] 305 272 246 324 311 ...
 $ Amount_profit_order     : num [1:36719] 22.9 -17.1 113.1 -97.3 62.3 ...
 $ Order_region            : chr [1:36719] "Oceania" "Eastern Asia" "South Asia" "Eastern Asia" ...
 $ Order_State             : chr [1:36719] "Queensland" "Guangxi" "Delhi" "Gansu" ...
```

Figure 1

Attributes Description:

1. Days for shipping: Actual shipping days of the purchased product
2. Days for shipment (scheduled): Days of scheduled delivery of the purchased product
3. Benefit per order: Earnings per order placed
4. Sales per customer: Total sales per customer made per customer
5. Delivery Status: Delivery status of orders: Advance shipping , Late delivery , Shipping canceled , Shipping on time
6. Late_delivery_risk: Categorical variable that indicates if sending is late (1), it is not late (0).
7. Category Name: Description of the product category
8. Customer Country: Country where the customer made the purchase
9. Customer Segment: Types of Customers: Consumer , Corporate , Home Office
10. Customer State: State to which the store where the purchase is registered belongs
11. Department Name: Department name of store

12. Market: Market to where the order is delivered : Africa , Europe , LATAM , Pacific Asia , USCA
13. Order Item Discount: Order item discount value
14. Order Item Product Price: Price of products without discount
15. Sales: Value in sales
16. Order Profit Per Order: Order Profit Per Order
17. Order Region: Region of the world where the order is delivered : Southeast Asia ,South Asia ,Oceania ,Eastern ...
18. Order State: State of the region where the order is delivered
19. Shipping Mode: The following shipping modes are presented : Standard Class , First Class , Second Class , Same Day.

Data Preparation

Here data was pretty much clean and there were no NA values. But at some point in Data the text value was in Portuguese language hence we convert it to the English as per our convenience where we require it. For scatter plot only 1000 random samples were presented for better view of the situation. Also we filter with the country to be USA and order status as complete for better analysis. Hence we filter with customer country to USA means order from USA are only considered for analysis.

```
new_data_supplychain_frame$Customer_Country[new_data_supplychain_frame$Customer_Country == "EE. UU."] <- "US"
latest_frame <- subset(new_data_supplychain_frame, Order_status == "COMPLETE")
view(head(latest_frame))
```

```
> sum(is.na(latest_frame$Type))
[1] 0
```

Figure 2

Here also the names of the columns for dataframe was changed as "_" to the name was added which will not be creating any problem while putting formula in future. Also for scatter plot and tree map random samples were generated for better analysis and smooth run of R studio. NA values were dropped for some of the columns.

Modelling

Here starting our modelling process we have first installed necessary libraries as shown in Figure 3. As our dataset is pretty large we have used "data.table" library for easing running process of R studio.

```
library(data.table)
library(maps)
library(RColorBrewer)
library(mapproj)
library(ggplot2)
library(tidyverse)
library(dplyr)
library(treemap)
library(googleVis)
```

Figure 3

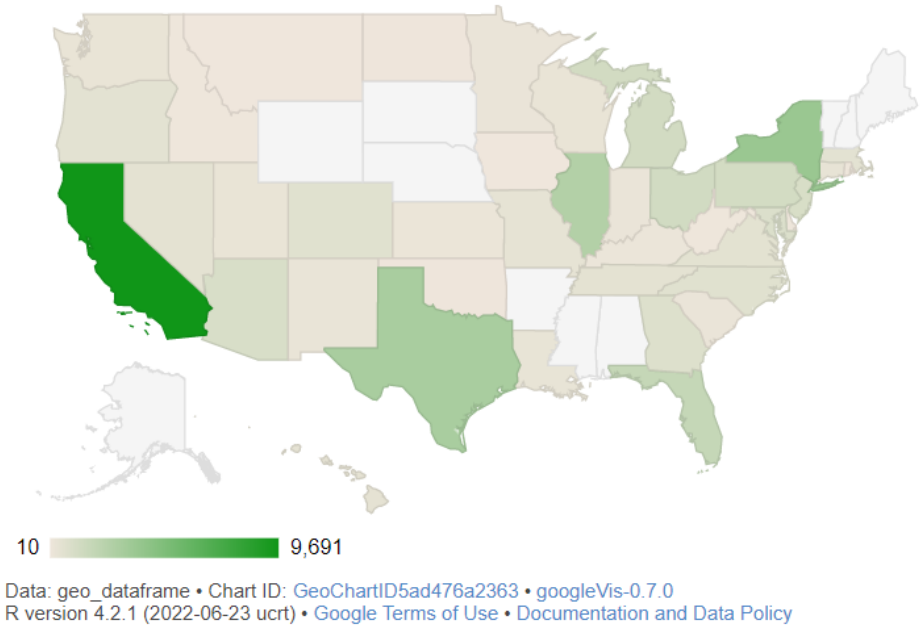


Figure 1 Order purchases across US States

Here Figure 1 shows the overall picture for the number of orders purchased from each state of the US for better understanding the context. Here From the figure it is clear that California has the highest sales of all other states followed by New york, Texas, Illinois etc. Further we stated how distribution of products is across states with the department category. Figure 1A shows the distributions according to the department name across various states. It seems that Fan shop has the highest sales among all departments across all states followed by apparel. Lowest sales were of Technology, Pet shop, Discs shop etc.

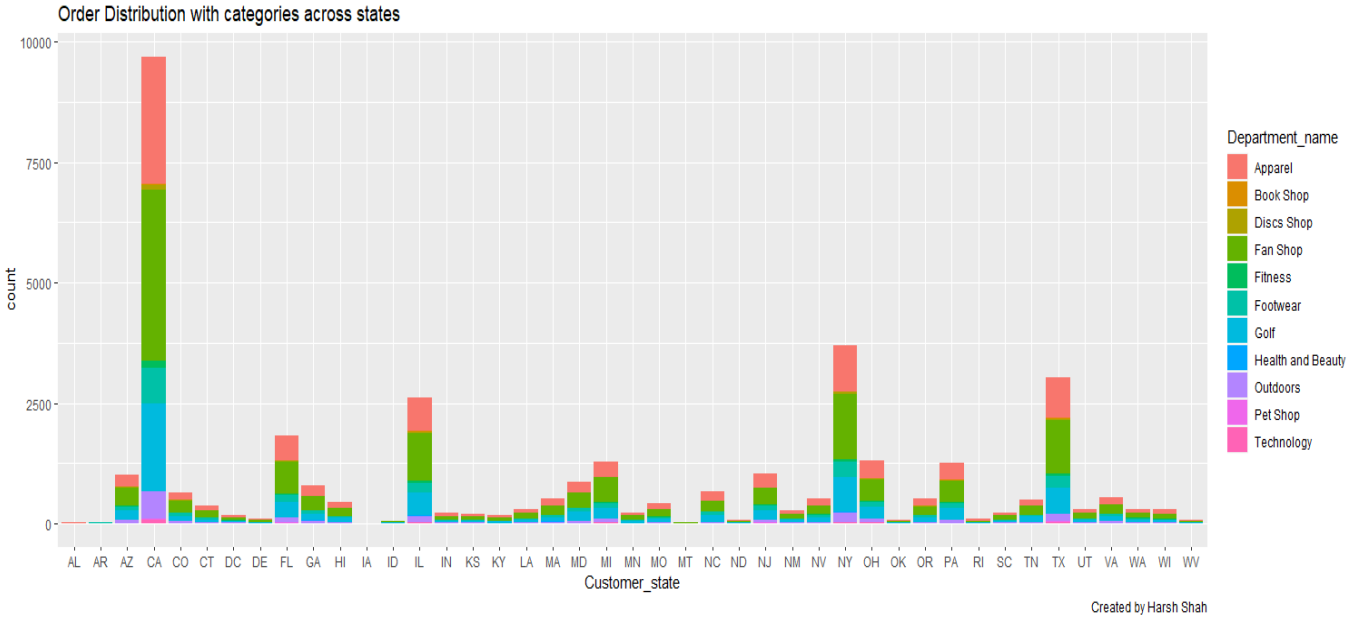


Figure 1A

From the figure it clearly seems that California has highest late shipping risk followed by New york, Texas and more. Here late shipping days is the sum of late in delivery days per order, hence higher the score more the state is late in shipping the items. Here we can observe that out of all states. And for California this is almost 6000 days of late shipping. Further we will also do analysis on profits and how late shipping may affect the benefit per order.

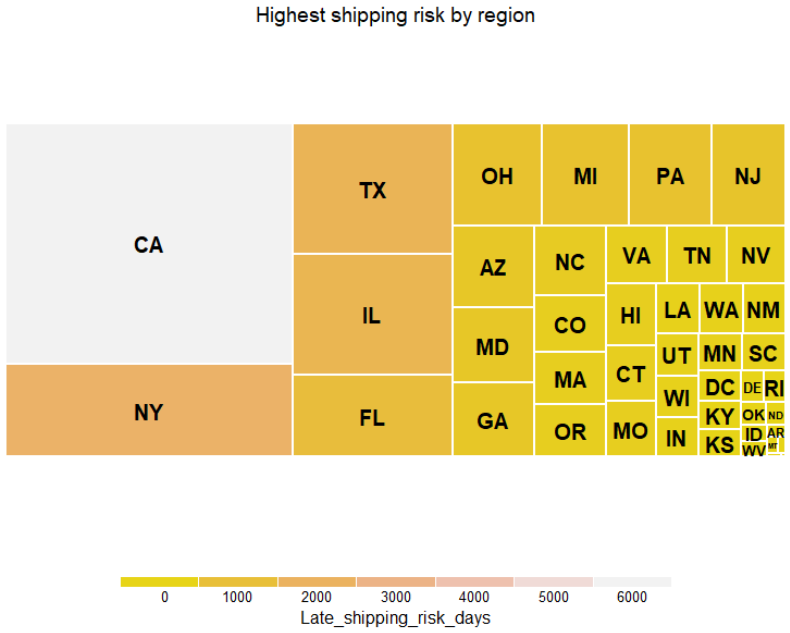


Figure 5

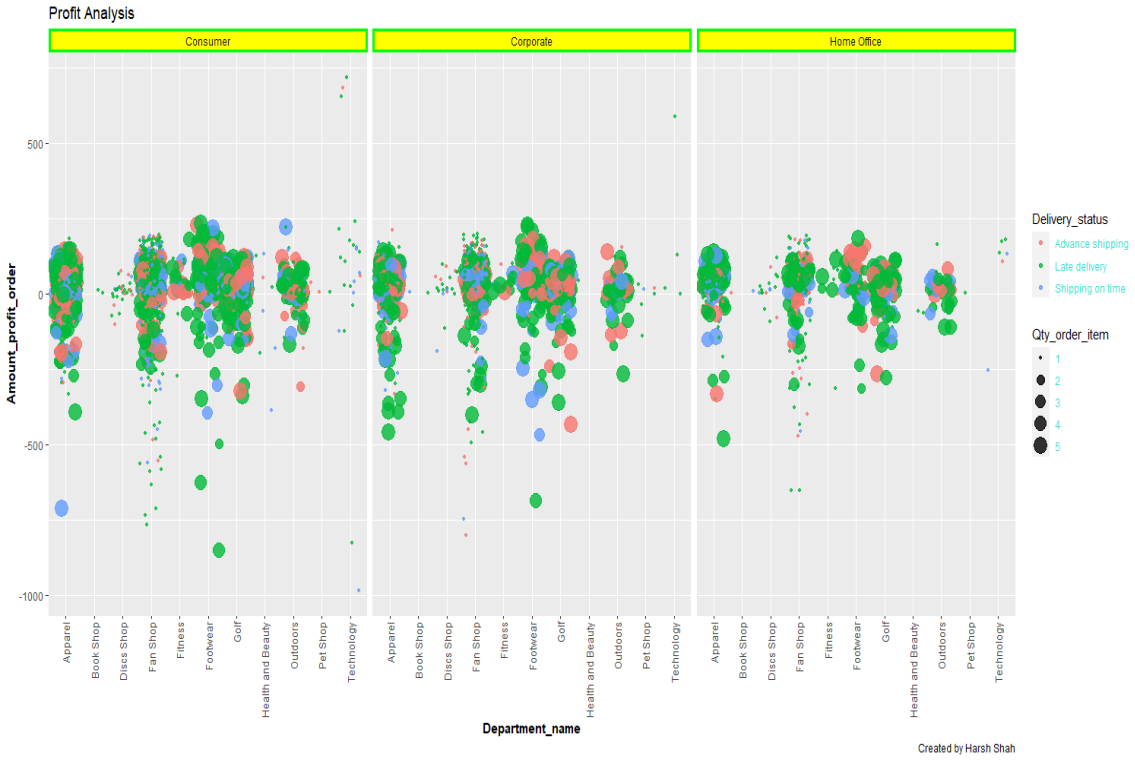


Figure 6

Here in Figure 6 Profit analysis is done based on the department categories and connected it with the category of orders, delivery status and even quantity as they may be interconnected with each other. Here we plot the graph with benefit per order and took random observations of only 1000 orders for better analysis and conclusions. This are divided according to the order category of Consumer, Corporate and Home office. Further they are divided with colours according to their shipping status which are Advance shipping, Late delivery or shipping on time. Here the size of the variables is given to the quantity of the product per order as that might affect the shipping method or benefit per order.

By visualizing the graph starting from the category it seems Consumer and Corporate have equally distributed orders followed by the Home Office. With respect to the delivery status it seems many order are included in late shipping and it clearly reflects on the profit of the products as it seems that majority of the orders are late shipped it seems we can observe shipping may have affected the profit per order. From the point of quantity also majority of orders which have loss belongs to high quantity. Coming to the department side it seems that in Consumer category it seems that apparel, footwear, golf and fan shop have majority sales and have almost equal ratio of profit and loss and on some products loss is very large. Fan shop have highest amount of loss followed by the apparel section. Coming to the corporate section it has the similar trend to the consumer section and apparels have losses with the higher quantity. Hence from the graph we can also observe that loss is not related to categories whether it is consumer, corporate or office. Hence it might be related to shipping and major losses it has in section Apparel, Fitness, Fan shop, Foot wear and golf. Hence further we will look with region and shipping mode to further analyse.

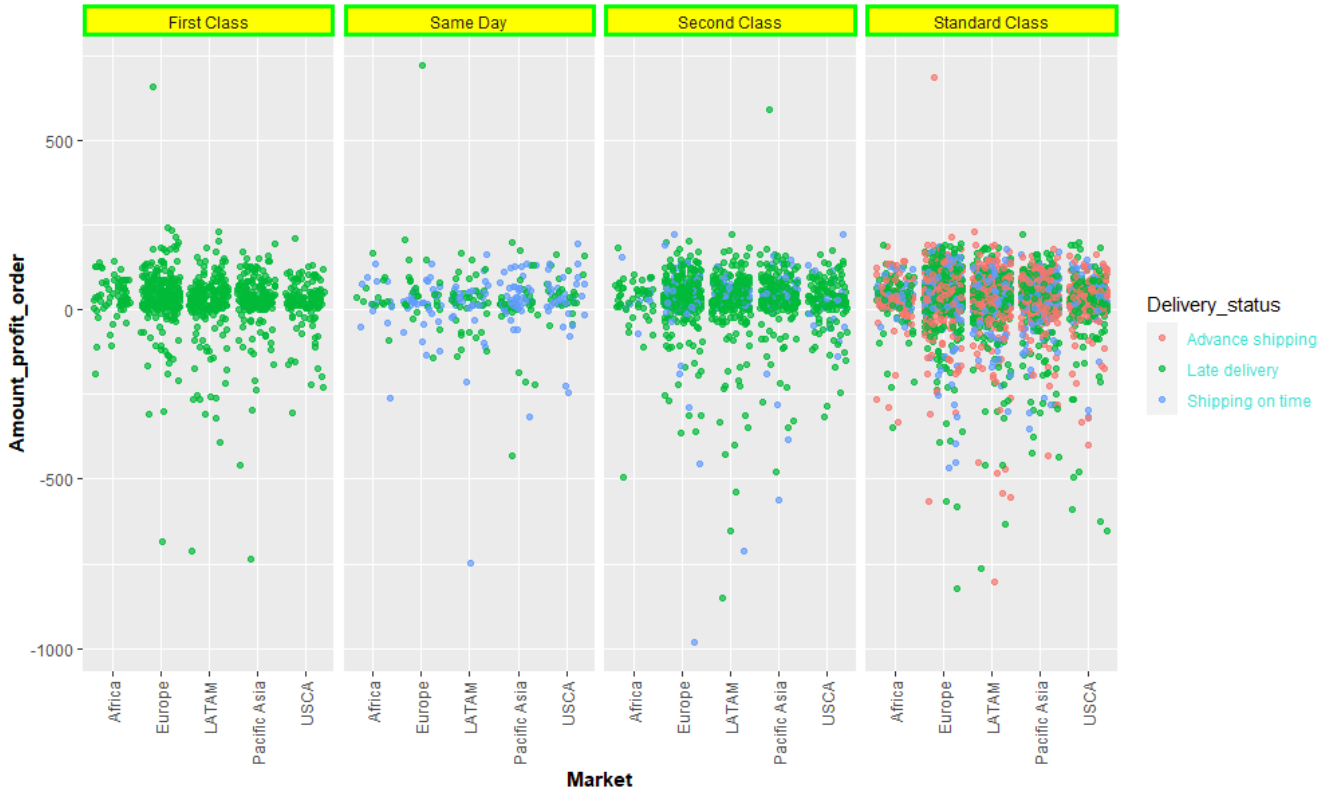


Figure 7

Here Figure 7 shows the plot for shipping mode, amount profit order, delivery status to further understand the trend. First starting with shipping mode it is divided in four categories which are First Class, Same Day, Second Class and Standard class. From the figure it is clear that all late deliveries are shipped by First class shipping but loss is not that in many products hence on the other hand standard class have highest loss in the products but with very less late delivery. It is clear from figure that late deliveries were shipped with First class or second class or it can also be interpret that shipping by first and second class all have late deliveries. With respect to the shipping region it seems that every region has equally distributed. With respect to advance shipping it seems it was shipped with standard class but even with advance shipping and standard class many orders have loss. Hence from this figure we can't conclude that shipping method or shipping status is responsible for the loss on the orders. Further we will analyse in detail and see overall trend with respect to various variables like sales, orders etc. with the help of heat to have an overall idea for the situation.

Figure 8 shows the heat map with rows of regions and columns with profit sum, sum sales, sum delivery risk, average delivery time, delivery scheduled, and number of orders. From the figure it seems that with terms of sum profit, sum sales, and delivery risk, Western Europe and Central America have the highest profit, sales and delivery risk. But with respect to average delivery time southern Africa has the highest time of delivery. And lowest have for Canada. With comparison to average delivery time and average delivery scheduled delivery time is bigger than delivery scheduled for regions like Southern Africa and Central Asia are among the highest. Whereas for Canada average delivery time is less than delivery scheduled. Whereas for East Africa also the average delivery time is less than delivery scheduled.

As from above figures and observations it seems we need a more clear picture regarding how profits and orders are distributed across various states. Figure 9 shows the profit analysis by shipping mode and divided by delivery status. Here sum of profit is calculated with states, shipping mode and delivery status. Backing up our previous observations for advance shipping and Standard shipping it is clear from the figure that advances shipping and choosing standard shipping as service is very beneficial from all the situations. We can say by this shipping status does contribute in some amount to profit. This can be backed by evidence of the state Kentucky as in Advance shipping there is profit whereas in Late shipping there is heavy loss from state of Kentucky hence that problem of late shipping should be taken care. Whereas with the state of Indiana there is Overall loss for all the shipping options hence we will discuss further for this by analysing the graph. Also seen that many state bear loss when there is late shipping. Hence from here it is reasonable to say that shipping might affect the profit for the states. Whereas for shipping courier it is seen that Standard class is best from the graph and loss for various states have second class or first class shipping. Whereas it is also seen states like Arizona, California are leading in every situation as also numbers of orders are also more from all other states. From this plot is observed that with states with less orders and delivery late has loss or very less profit. Further we will try to analyse how late shipping has affected the state Kentucky which have the loss and try to analyse the situation. Figure 10 shows the profit analysis for Kentucky state and to which region it sells for better analysis of loss of state.

Figure 8

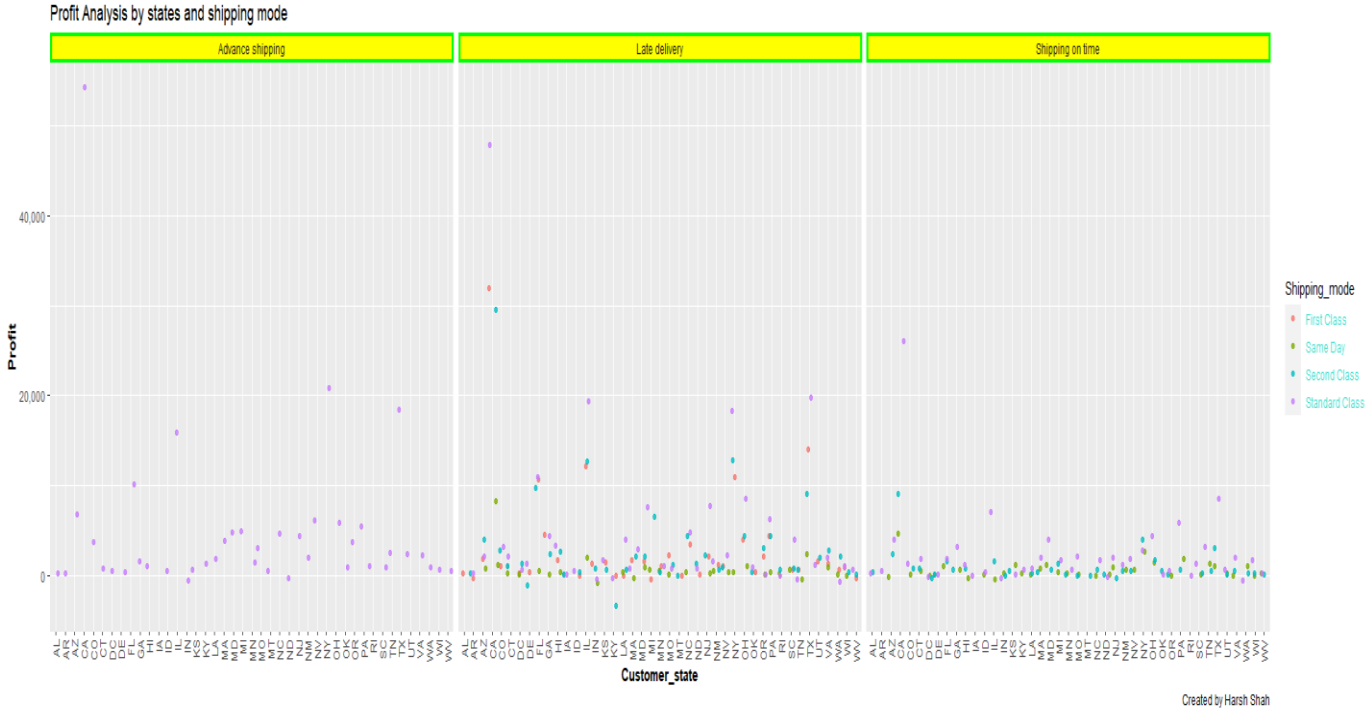


Figure 9

From the Figure 10 it seems that orders from Kentucky faces loss when products bare shipped to any region out of 5. And majority of loss have been shipped late. Hence it can be reasonable to conclude that shipping might affect the profit for the orders. For Africa region losses are in department of Fishing, Cardio equipment and Camping and Hiking that too with late shipped. Hence this problem should be addressed. For Europe region also cardio equipment is a loss product and also golf products. For Latin America, Camping and Hiking equipment faces loss and even indoor and outdoor games. There is not much loss in ASIA region and USCA region.

Figure 10

Observations and Conclusions

- From the initial graph it is clear that California has the highest number of orders followed by New york, Texas, Illinois etc.
- Coming further to our analysis highest sales were Western Europe followed by Central America followed by Southern Europe and South America.
- For our analysis on department sales it was observed that Fan shop has the highest sales followed by Apparel and Golf.
- California, New York, Illinois, Florida has highest risk in shipping. Higher the number of orders higher is the risk of shipping it seems.
- From profit analysis it seems there is loss with many products which mainly were delivered late and higher quantities. Some of departments with highest loss were apparel, Fitness, Golf and Fan shop.
- Coming to our shipping analysis it clearly seems shipping done by First class were all late deliveries and majority of shipping done by second class were also late. Whereas loss is highest on products shipped with standard shipping.
- From analysis it seems that with terms of sum profit, sum sales, and delivery risk, Western Europe and Central America have the highest profit, sales and delivery risk. But with respect to average delivery time southern Africa has the highest time of

delivery. And lowest have for Canada. With comparison to average delivery time and average delivery scheduled delivery time is bigger than delivery scheduled for regions like Southern Africa and Central Asia are among the highest. Whereas for Canada average delivery time is less than delivery scheduled. Whereas for East Africa also the average delivery time is less than delivery scheduled.

- From the figure it seems that for advance shipping only standard shipping service and there was no loss in advance shipping which seems that shipping might affect the profits. And there is loss when delivery is late delivery. And when there is loss it was shipped with second class. And we took example of Kentucky for further analysis.
- And majority of loss have been shipped late. Hence it can be reasonable to conclude that shipping might affect the profit for the orders. For Africa region losses are in department of Fishing, Cardio equipment and Camping and Hiking that too with late shipped. Hence this problem should be addressed. For Europe region also cardio equipment is a loss product and also golf products. For Latin America, Camping and Hiking equipment faces loss and even indoor and outdoor games. There is not much loss in ASIA region and USCA region.