Data Analyst Intern Challenge

Prateek Rawat | 02/05/2018

Contents

Mission Statement	page 2
Approach	page 2
Data Description	page 2
Data Cleaning	page 2
Data Visualization	page 2
Outliers	page 3
Data Insights	page 4
Next Step	page 10
Code for Data Cleaning	page 11

Mission Statement:

Extracting insights from transaction details of a random company.

Approach:

- Build different types of graphs to analyze sales based on placed orders and cancelled orders by countries, year, and months.
- Tools used for this approach are RStudio and Tableau.

Data Description:

- Data contains raw data about transaction of a company.
  It consists of:
  - 541910 data points.
  - 8 variables
    - InvoiceNo: Invoice number.
    - StockCode: Product (item) code.
    - Description: Product (item) name.
    - Quantity: The quantities of each product (item) per transaction.
    - InvoiceDate: Invoice Date and time.
    - UnitPrice: Unit price.
    - CustomerID: Customer number.
    - Country: Country name.

Data Cleaning:

- The data was dirty, so missing values were removed.
- Datatypes of columns such as Stock Code should be integer as per the information provided so that datatype was changed from character to integer.
- The data provided in the InvoiceDate columns has data of date, month, year and time in one column so the column InvoiceDate was divide into 4 columns of Year, Month, Date, Time respectively.(Refer R code R_cleaning)

Data Visualization:

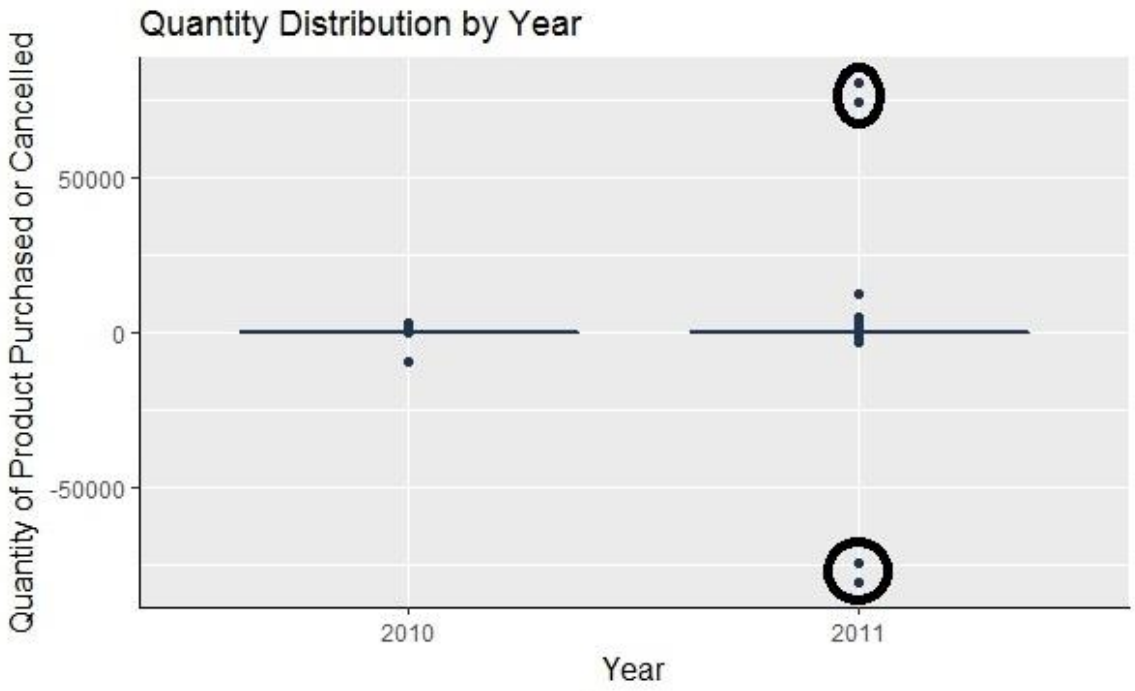
- A story of the insight has been made using Tableau and published on Tableau Public. Click the link below to access it.

- Refer to this link while going to Data Insights as the graphs are large enough to be captured in the word file.

https://public.tableau.com/profile/prateek.rawat#!/vizhome/Prateek_Rawat/DataAnalyticsChallenge?publish=yes

Outliers:

- While plotting the box plot of Quantity for outliers, some values were exceptionally high.

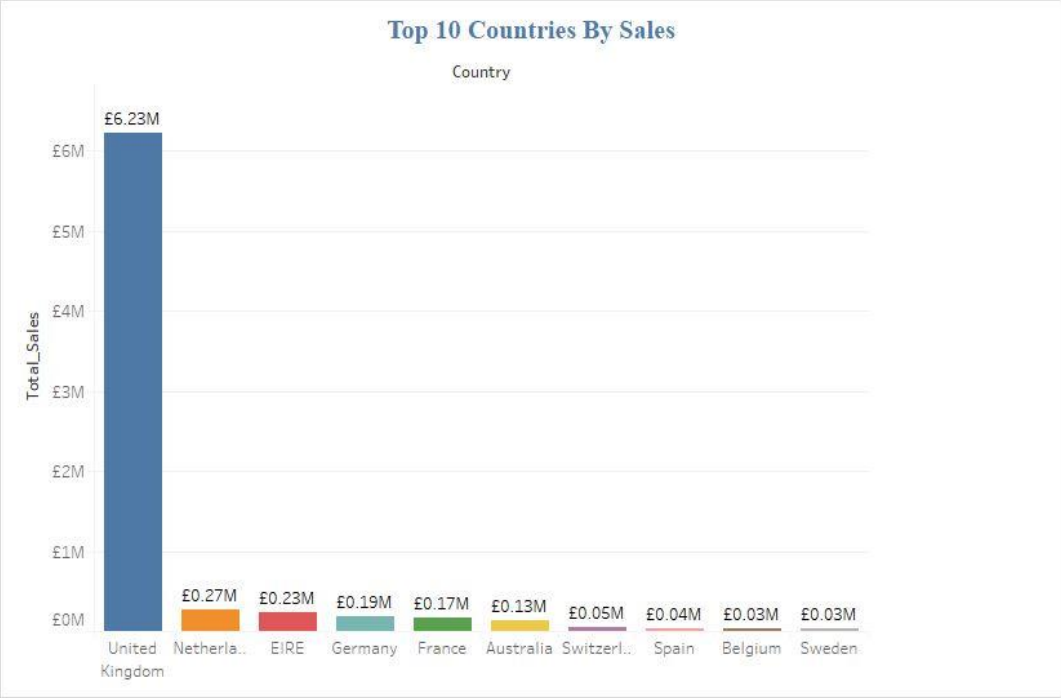


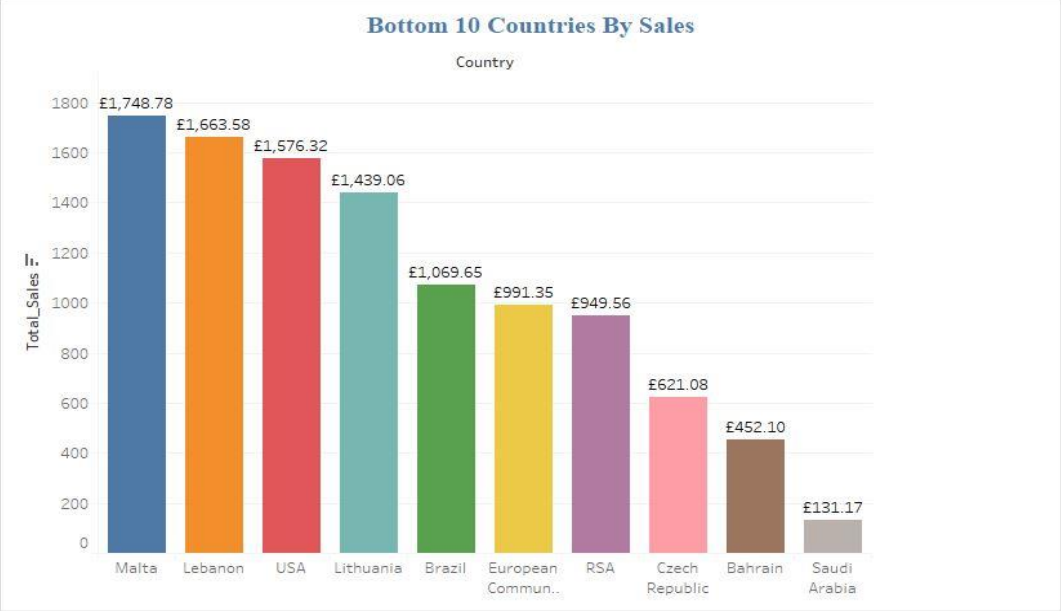
- The data which has been provided looks like it a data for retail store and the outliers which have been marked above might have been booked my mistake from a customer. This needs to be further investigated with help of Sales team.
- I have eliminated these values for analysis because this can lead to wrong results. So, I have assumed that these sales take place when,
    1. Some customer might have booked a product by mistake.
    2. Some customer might have cancelled a product after purchasing it.
    3. Some promotional offer was going on and customer might have purchased a product in a bulk.

Data Insights:

1. Sales

This is a higher-level analysis just to see where the company is making good sales and poor sales.





Rationale:

- We can see from the graphs mentioned above that United Kingdom is the major market for the company with Europe being its major area for sales. Based on this analysis sales and marketing should focus on these areas so that company should stand strong in these areas with respect to other competitors in the market.
- Based on the analysis of bottom 10 countries with respect to sales the company should focus on countries such as USA, RSA, Brazil, and Saudi Arabia. Economic stability, population makes these countries a good fit for company to expand their business. With the help of marketing and sales team we should do research on geographical and behavioral aspects of these countries so that it will help the company to add new products into their cart and this may lead to increase in sales from these countries.
- Promoting about the company online will be beneficial since it requires less capital and it is accessible to large percentage of population. We can start different market campaigns such Email marketing, YouTube marketing, Launch Mobile App.

2. Sales Comparison and Forecasting

Total sales for the month of December 2011 cannot be calculated as the data provided does not contain complete data for the month of December placed and cancelled ordered. So, I have forecasted sales for December 2011 and did a comparison of sales month wise

Rationale:

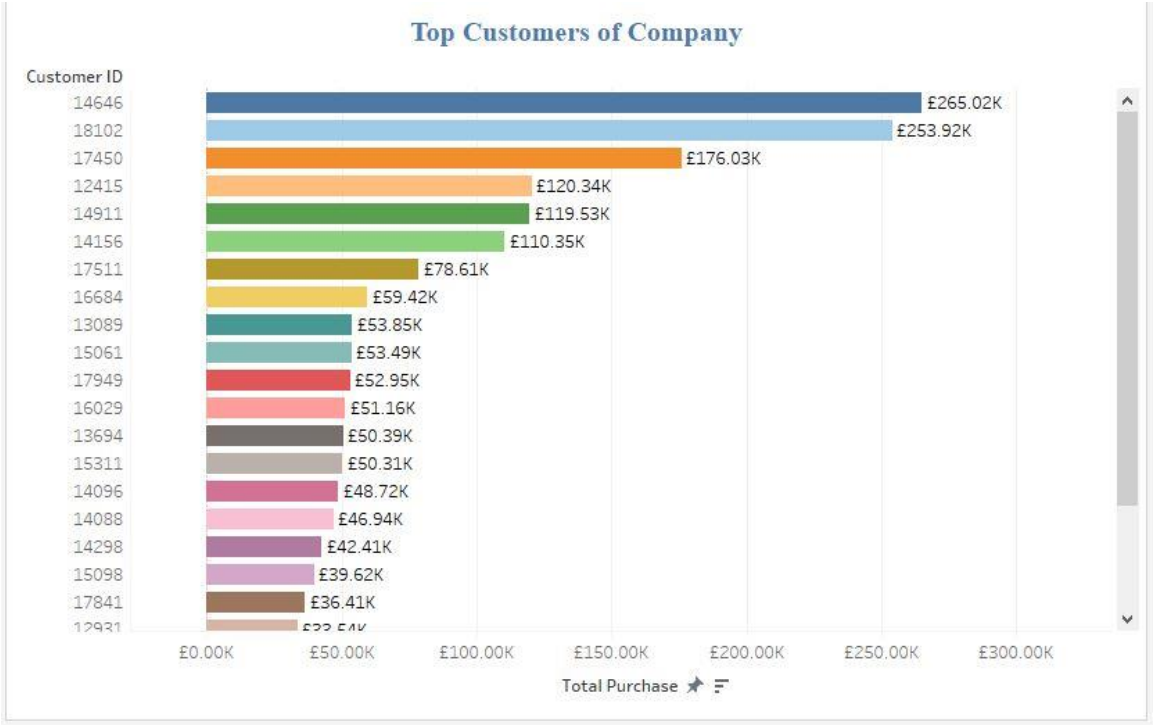
- When analyzing sales with respect to month, November has highest sales. It might be due to the black Friday sale and Thanksgiving. Based on the assumption for high sales in month of November we should do analysis on various market trends so that company can give customers various promo codes and offers in other months of the year to increase the overall sales.
- After forecasting sales for December 2011, it can be assumed that sales will increase when compared to sales in December 2010.If sales go as per the prediction than it can be concluded that marketing and sales team have done there jobs well to increase sales of the company.

3. Customer Analysis

I have visualized top 25 customers for the company. Customers plays a vital role for any company. So, company should invest time to improve customer experience. Company should focus on to make loyal customers by giving them special privilege and offers. This will help the company in the long run as loyal customers are the customers who trusts you and you can cross sell your new products among them.

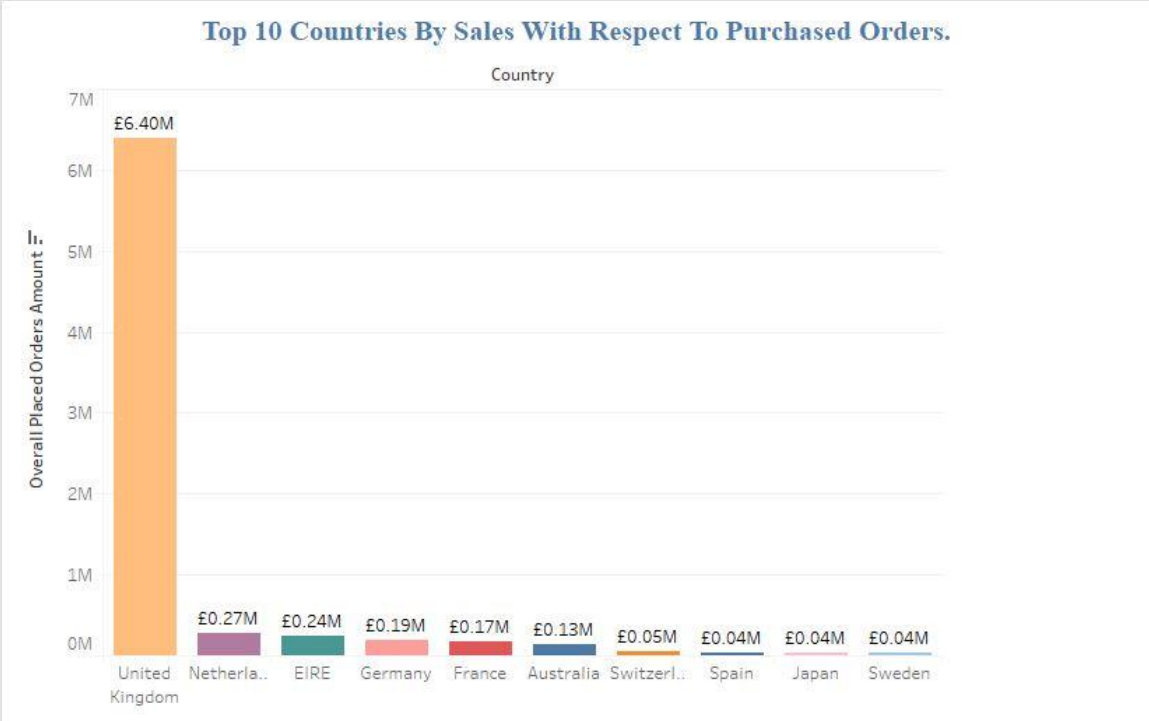
Let's take an analogy over here, If I have Starbucks and Dunkin Donuts in front of me, then I would love to go to Starbucks. The reason is my cup has my name on it. It's a connection or attachment and because of that I feel like going in Starbucks. So, company should focus on improving customer experience.

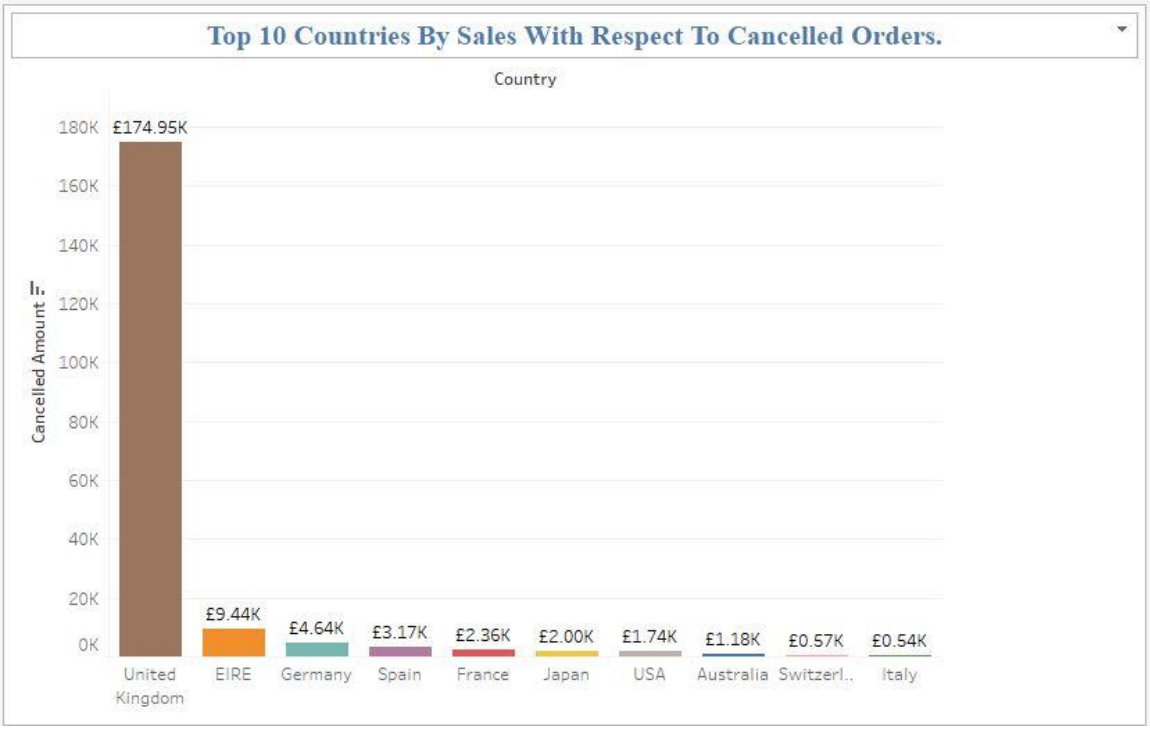
It has been seen that in negative economy also customer experience is of high priority to maintain overall sales of a company. Therefore, in positive economy customer experience can do wonders for the company.



4. Analysis for Placed and Cancelled Orders

Going deeper into analysis, we will examine pattern of Top 10 countries with respect to placed and cancelled orders. (Refer to story published on Tableau Public.)



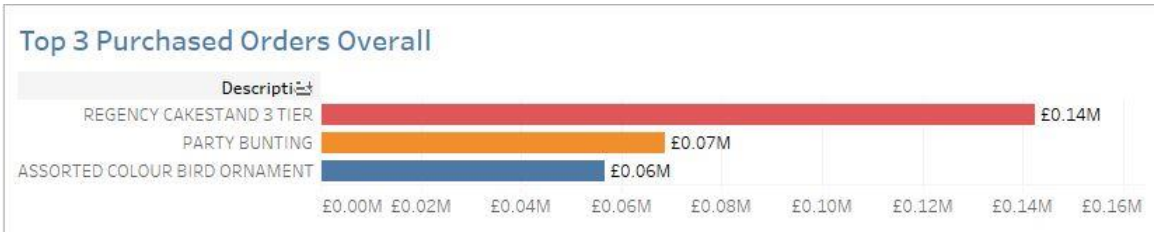


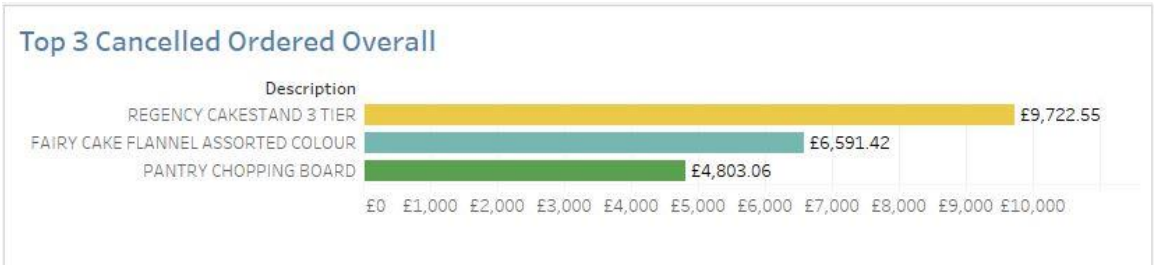
Rationale:

- We can see from graphs above that United Kingdom is on top when it comes to both orders Placed and Cancelled. As United Kingdom is the major market for the company, they should rectify reasons because of which customers cancel product after ordering them. One major reason to cancel a product may be that customer is not satisfied with the product or a customer have placed the product by mistake.
- We can see that USA is on 7th place when it comes to cancelled order and it ranks in bottom 10 countries based on overall sales which indicates that customers in USA are not getting their products as they have expected. I would suggest conducting a survey among potential customers so that the company can improve their services.

5. Analysis for Placed and Cancelled Products

We will investigate top 3 and bottom 3 products with respect to Placed and Cancelled orders across all countries.





Rationale:

- As mentioned above we can get information regarding top and least ordered products with respect to countries. This will give an idea to company to make their services better.
- Information regarding highest cancelled products can help company to dig deep into the issues the customers are facing with products and can provide them better services in future by giving good quality products and discounts.

Next Step:

- By making data collection more analytics centric will help in better analysis
  - It would be good if we can get monthly seasonality related data so that we can do trend and seasonality analysis. This will help company have products in their cart according to customer needs.
  - We need to have more Customer related data because it will help in Customer segmentation which could help company in deciding marketing strategies.
- If we can gather data state-wise or city-wise, then it would help to create strong base for our analysis.
- We need demographic information of the people to make strategies to open stores and include products as the customer needs.
- Decide on pricing strategies.
- In Future, we could use data for modelling to predict sales of different products. This could help in better inventory management.

Code for Data Cleaning

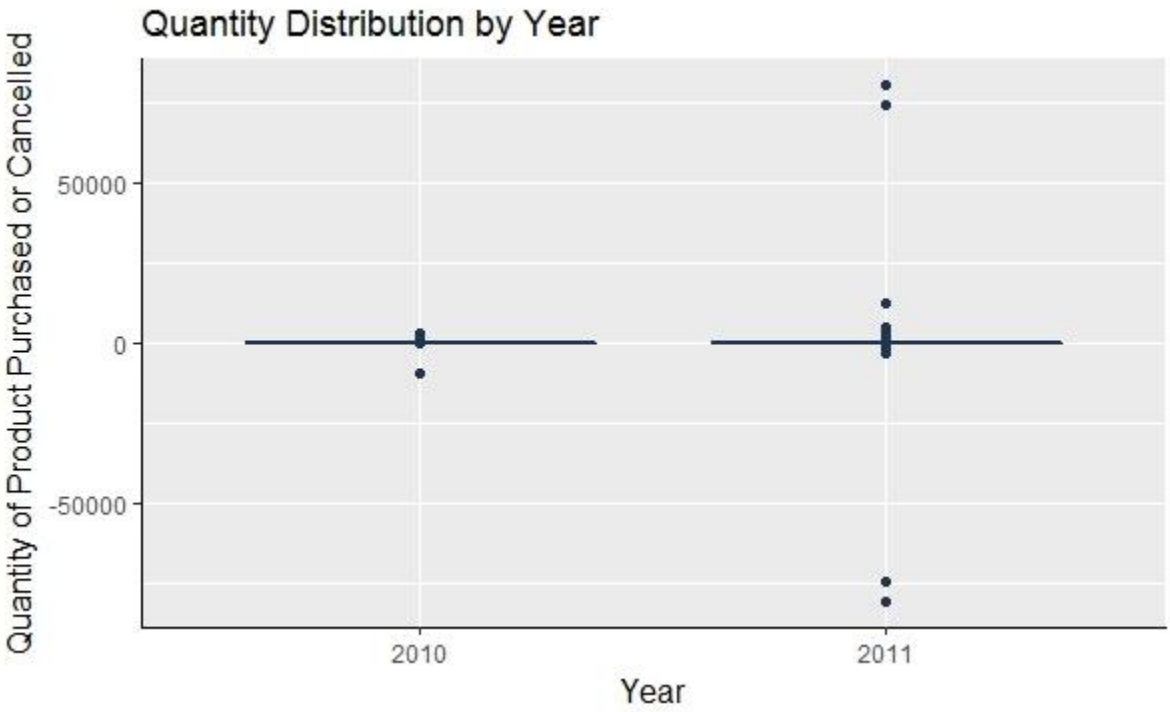
R_cleaning.R

By Prateek Rawat

Mon Feb 5 13:45:30 2018

```r
#Loading the file for cleaning
setwd("E:/zappos challenge")

#Imporitng Libraries
library(readxl)
library(outliers)
library(dplyr)
library(ggplot2)
library(extrafont)

#Imporing Data
Sales <- read_excel("challenge.xlsx")
class(Sales)
View(Sales)


#Checking NA values
any(is.na.data.frame(Sales))
sum(is.na.data.frame(Sales))

Sales<-na.omit(Sales)
any(is.na.data.frame(Sales))

#Checking Datatype for each column
sapply(Sales, mode)

#Changing Stock Code to Integer from character as mentioned in attributes details
Sales<-transform(Sales,  StockCode= as.numeric(StockCode))
any(is.na.data.frame(Sales))
Sales<-na.omit(Sales)
```

```
31
32 #Dividing  Invoice Date into two columns of date and time
33 Sales<- separate(Sales, InvoiceDate, into = c("Year", "Month", "Day"), sep = "-", remove = FALSE)
34 Sales<- separate(Sales, Day, into = c("Date", "Time"), sep = " ", remove = TRUE)
35 Sales<-Sales[c(1:4, 6:12)]
36
37 #Checking outliers and boxplot for sales
38 out <- outlier(can['Quantity'], opposite = FALSE, logical= FALSE)
39 out
40
41 fill<- "#4271AE"
42 line<- "#1F3552"
43 outlier_Sales <- ggplot(Sales, aes(x =Sales$Year, y =Sales$Quantity))+
44   geom_boxplot(fill = fill, colour = line)+
45   scale_y_continuous(name = "Quantity of Product Purchased or Cancelled")+
46   scale_x_discrete(name ="Year")+
47   ggtitle("Quantity Distribution by Year")+
48   theme(axis.line.x = element_line(size = 0.5, colour = "black"),
49         axis.line.y = element_line(size = 0.5, colour = "black"),
50         legend.position = "bottom", legend.direction = "horizontal",
51         legend.box = "horizontal", legend.key.size = unit(1, "cm"),
52         plot.title = element_text(family = "Tahoma"),
53         text = element_text(family = "Tahoma"),
54         axis.title = element_text(size = 12),
55         legend.text = element_text(size = 9),
56         legend.title = element_text(face = "bold", size = 9))
57 outlier_Sales
58
59 cleaned_sales<-filter(Sales, Sales$Quantity < 10000 & Sales$Quantity > -10000)
60 View(cleaned_sales)
```




```
61
62 #Dividing Sales data frame into subsets of placed and cancelled orders
63 can_order <-filter(cleaned_sales, Quantity < 0 )
64 View(can_order)
65 placed_order <-filter(cleaned_sales, Quantity > 0 )
66 View(placed_order)
67
68 #Importing dataframes to csv
69 write.csv(placed_order,"E:/zappos challenge/placed_order.csv",row.names=F)
70 write.csv(can_order,"E:/zappos challenge/cancelled_order.csv",row.names=F)
71 write.csv(cleaned_sales,"E:/zappos challenge/cleaned_sales.csv",row.names=F)
72
```