Applied Data Analytics (Feb 2018)
MSc in Big Data Management and Analytics
Assessment

Online Retail Dataset

Pronoy D'sa – 295742

Data Analytics and Information Dashboards for the dataset in the selected domain

Part I

Exploratory Analytics and Visualisation of Data via Information Dashboard (40% of the overall module)

E-commerce space is extremely competitive & current day consumers want personalized experiences, easy to access product information & exceptional services. To deal with high expectations from consumers, online retailers need to leverage data being collected & actively integrate analytics to improve their decision-making process. Retailers should use analytics to better understand consumer preferences, & further provide them with right product offerings. Also, online retailers should extract value from their data beyond just information gathering & monitoring.

The following dataset which I've selected is a multinational dataset which contains the history of all the transactions that have taken place between December 1, 2010 and December 9, 2011 for a UK-based and registered non-store online retail corporation (name not disclosed). The company mainly sells unique all-purpose items or inventory. Most of the customers of the company are wholesalers.

The Dataset is currently located at UCI Machine Learning Repository which is a reliable, online service for Machine learning & Intelligent systems.

Link: https://archive.ics.uci.edu/ml/datasets/Online+Retail

Source:

Dr Daqing Chen, Director: Public Analytics group. chend@isbu.ac.uk, School of Engineering, London South Bank University, London SE1 0AA, UK.

Relevant Papers:

The evolution of direct, data and digital marketing, Richard Webber, Journal of Direct, Data and Digital Marketing Practice (2013) 14, 291â€“309.

Clustering Experiments on Big Transaction Data for Market Segmentation,

Ashishkumar Singh, Grace Rumantir, Annie South, Blair Bethwaite, Proceedings of the 2014 International Conference on Big Data Science and Computing.

A decision-making framework for precision marketing, Zhen You, Yain-Whar Si, Defu Zhang, XiangXiang Zeng, Stephen C.H. Leung c, Tao Li, Expert Systems with Applications, 42 (2015) 3357â€“3367.

Pronoy D'sa 2952742

Citation Request:

Daqing Chen, Sai Liang Sain, and Kun Guo, Data mining for the online retail industry: A case study of RFM model-based customer segmentation using data mining, Journal of Database Marketing and Customer Strategy Management, Vol. 19, No. 3, pp. 197â€"208, 2012 (Published online before print: 27 August 2012. doi: 10.1057/dbm.2012.17).

Attribute Description:

Data Set Characteristics:	Attribute Characteristics:	Associated Tasks:	Number of Instances:	Number of Attributes:	Missing Values?	Area:	Date Donated
Multivariate, Sequential, Time-Series	Integer, Real	Classification, Clustering	541909	8	N/A	Business	06/11/15

The dataset consists of the following attributes: -

>**InvoiceNo:** Invoice number. It is a Nominal attribute, with a 6-digit code uniquely assigned to each transaction. If this code starts with letter 'c', it indicates a cancellation.

>**StockCode:** Product (item/service) code. It is a Nominal attribute, with a 5-digit code uniquely assigned to each distinct product.

>**Description:** Product (item/service) name. It is a Nominal attribute giving the name of the item/service in question.

>**Quantity:** The quantities of each product (item/service) per transaction. It is Numeric.

>**InvoiceDateTime:** Invoice Date & time. It is a Numeric attribute, giving the day, month, year & time when each transaction was made.

>**UnitPrice:** Unit price. It is Numeric. It is the product price per unit in sterling.

>**CustomerID:** Customer number. A Nominal attribute, with a 5-digit code uniquely assigned to each customer.

>**Country:** Country name. A Nominal attribute which provides the name of the country where each customer resides.

After observing the dataset, composite attributes were introduced in excel spreadsheets to perform & make exploratory analytics on tableau easier & understandable. The composite attributes are:

>**InvoiceDate:** Made by splitting InvoiceDateTime field. It gives us data on the Invoice Date in the order of "dd/mm/yyyy" that is the day, month or year when each transaction was generated.

Formula used: =INT(E2) …(Also, change date format)

>**Sales(Quantity*UnitPrice):** It is the product of Quantity & Unit price of an item that is purchased to give us the total sale value or revenue of an item/service.

*Formula used: =D2*G2*

>**International Shipping/Service?:** It is created by taking the country field into consideration. Since, the retail store is in UK, the attribute finds out if international shipping is required or not (Y/N?).

Formula used: =IF(J2="United Kingdom","No","Yes")

Also, calculated fields were created using tableau to make the data concise & clear.

>**InvoiceTime:** Made by splitting InvoiceDateTime field. It gives us the time of the invoice that is the hours, minutes, etc. when each transaction was generated.

Invoice Time

```
IF DATEPART('hour',[Invoice Date Time]) = 0 THEN '12'
ELSEIF DATEPART('hour',[Invoice Date Time])<10 THEN STR(DATEPART('hour',[Invoice Date Time]))
ELSEIF DATEPART('hour',[Invoice Date Time])>12 THEN STR(DATEPART('hour',[Invoice Date Time])-12)
ELSE STR(DATEPART('hour',[Invoice Date Time])) END //hour
+ ':' +
IF DATEPART('minute',[Invoice Date Time])<10 THEN '0'+STR(DATEPART('minute',[Invoice Date Time]))
ELSE STR(DATEPART('minute',[Invoice Date Time])) END //minute
+' '+
IF DATEPART('hour',[Invoice Date Time])>=12 THEN 'PM' ELSE 'AM' END //AM or PM
```

>**Profit/Loss:** We used an if condition where if quantity > 0 then it's true for profit otherwise it's false which is a loss.

Profit/Loss

```
IIF([Quantity]>0,"Profit","Loss")
```

Before undertaking the data analytics task, the questions/issues that were deemed necessary by me to be answered/explored were as follows:

1) What are the most popular that is the top items/services bought which are catalogued at this online retail establishment?

2) What are the top items/services that are bought in the most quantity as per customers or countries?

3) What are the least popular items/services that are bought at this online retail site.

4) Are there any items/services that have suffered the retail store a loss? If there are what are these items/services?

5) What are the top items/services that have the most unit price and are bought the most by the customers?

6) Who are the top buyers or customers? What do they buy?

7) Who are our customers that make the least sales purchase? What products do they buy?

8) Which are the top countries which produce the most sales profits?

9) Which are the top countries that produce the most loss of revenue?

10) What are the top years, quarters, months, days, etc when sales are high? At what days are they low?

11) What are the top hours during this period for the most sales?

12) Can we forecast, or make future predictions of profits or loss for the years to come?

These questions/issues are important from the perspective of both the consumer, that is the customer & producer which in this case is the Online Retail Store or the Retailer because e-commerce business is a very competitive market nowadays. It is essential for Retailer to know the needs of a customer and adapt to these needs as quickly as possible. Analytics could be used to improve factors such as customer retention, customer experience, customized recommendations, targeted promotions, increasing sales, profit & minimizing loss to the business.

For success, an online retailer needs to attract consumers to the site & provide them with the appropriate set of products & offerings which would lead to a purchase. Online retailers must provide customized product recommendations & promotions by better understanding consumer's purchase behaviours, past purchases & popularity of products on the site. It is also important for any online retailer to develop a good understanding about consumer demand. Thus, for the following dataset, the online retailer should:

1. Find ideal customers & build complete profiles. Also, should discover & nurture best customers.

2. Introduce much suitable prices & convenient payment schemes or options for purchasing products.

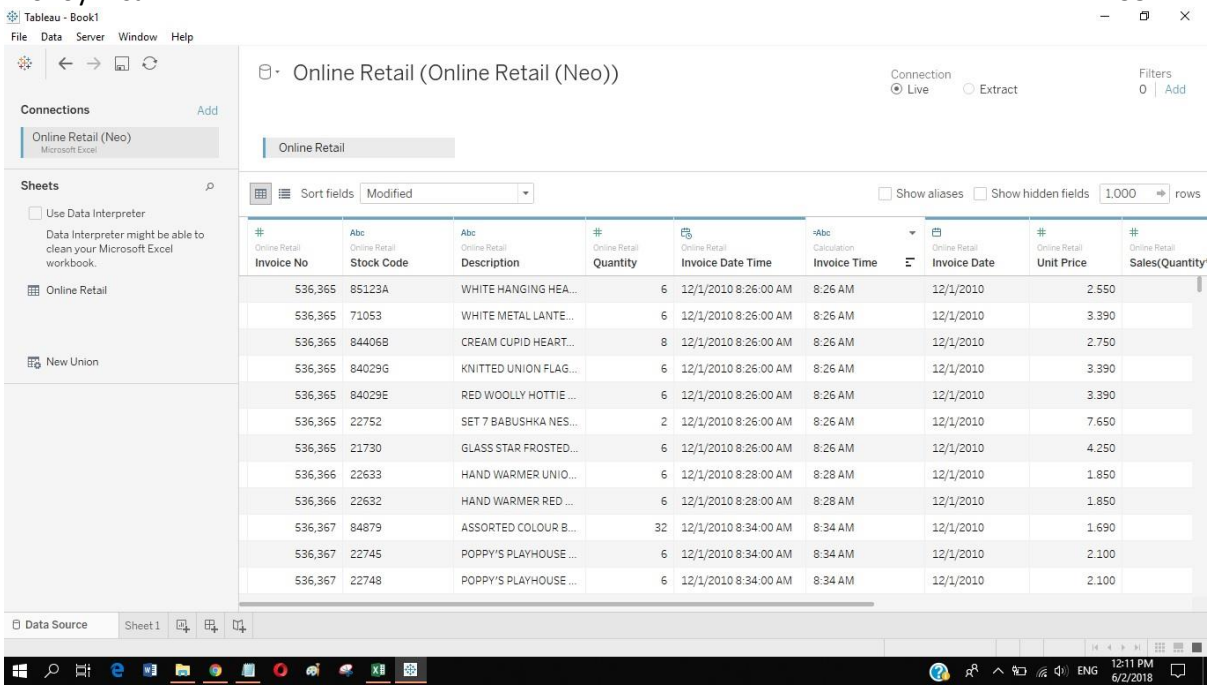
3. Check the popularity of a product in a country by its sales to increase or decrease its supply & demand, maximizing profits & minimizing losses.

4. Gradually introduce rewards or targeted promotions to keep the customers engaged.

5. Use market-basket analysis to pair items that are not frequently purchased with items that are most frequently purchased.

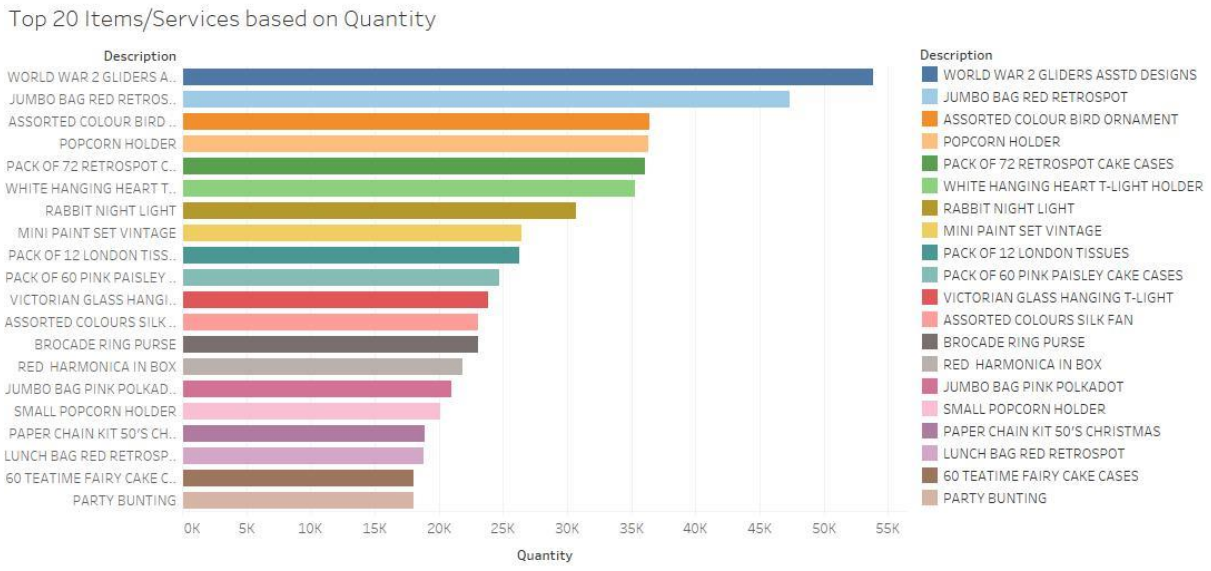
6. Assess risks by predicting future profits or loss based on historical data.

We shall now use Exploratory Data Analytics (EDA). The aim of EDA is to analyse the Online Retail Dataset to summarize its main characteristics or attributes through visual methods and to identify outliers, trends & patterns in it.

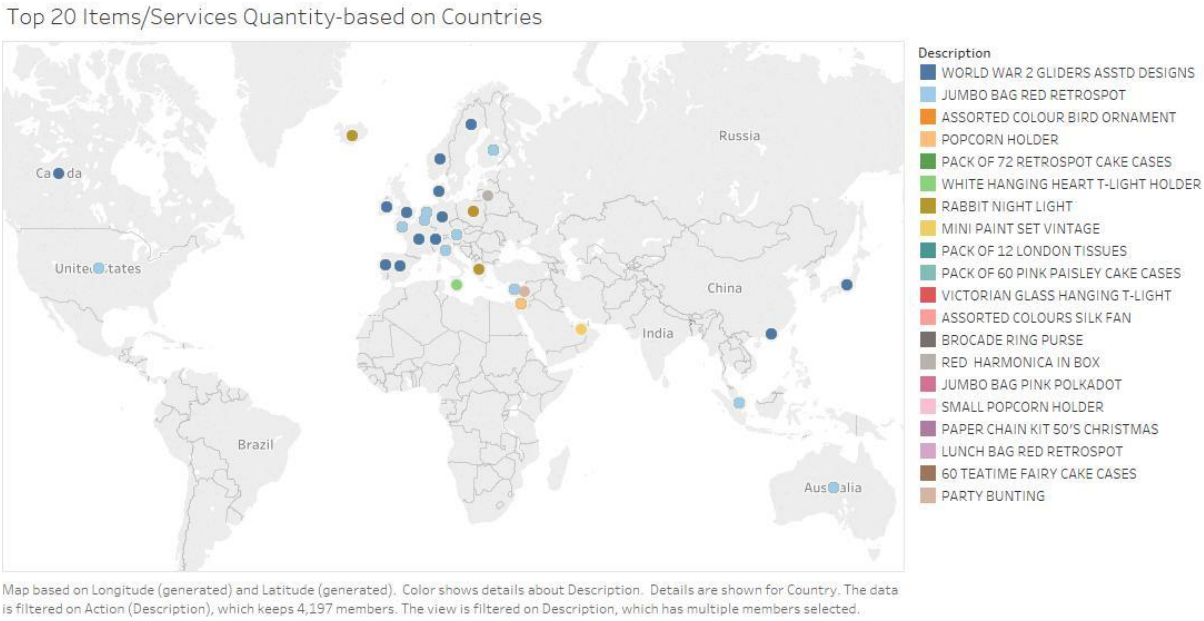


Dashboard 1: Top 20 Item/Services Quantity-based

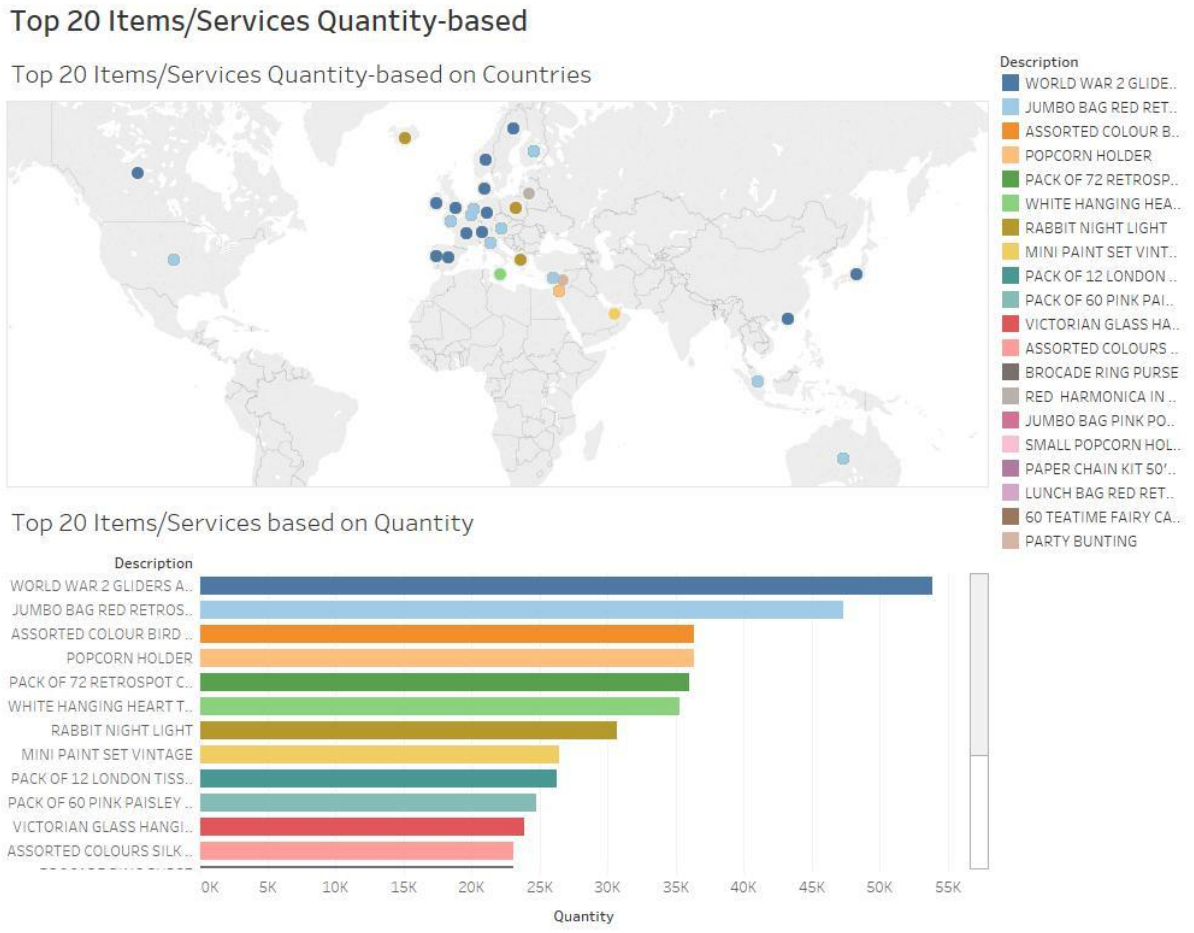
Here I have taken Quantity as the independent variable whose relation we're interested in. Quantity is aggregated to SUM(Quantity).



For Worksheet 1 – Top 20 Items/Services based on Quantity, I have taken SUM(Quantity) and Description of product as the 2 variables where Description is arranged in a descending order to get the most quantity items/services & description is filtered by Top 20 SUM(Quantity) items/services.

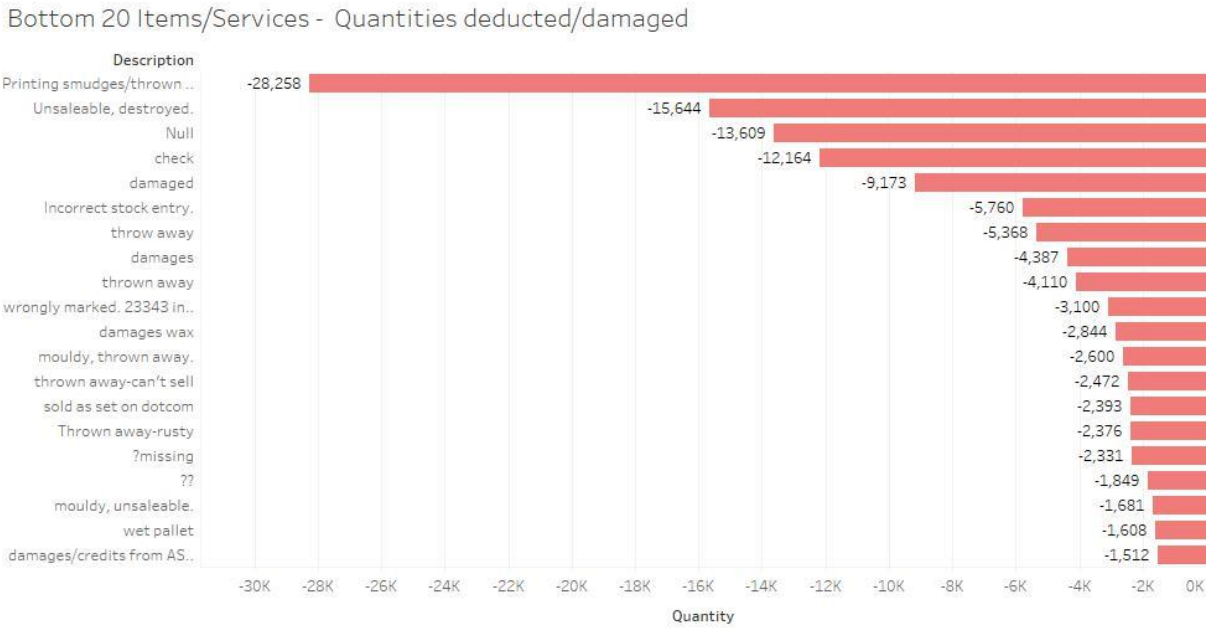


For Worksheet 2 – Top 20 Items/Services Quantity-based on Countries, I have taken countries and Description as the interested relation & filtered Description according to SUM(Quantity). The resulting dashboard gives us a view of the top 20 items/services demanded or bought in most numbers/quantities in those countries.



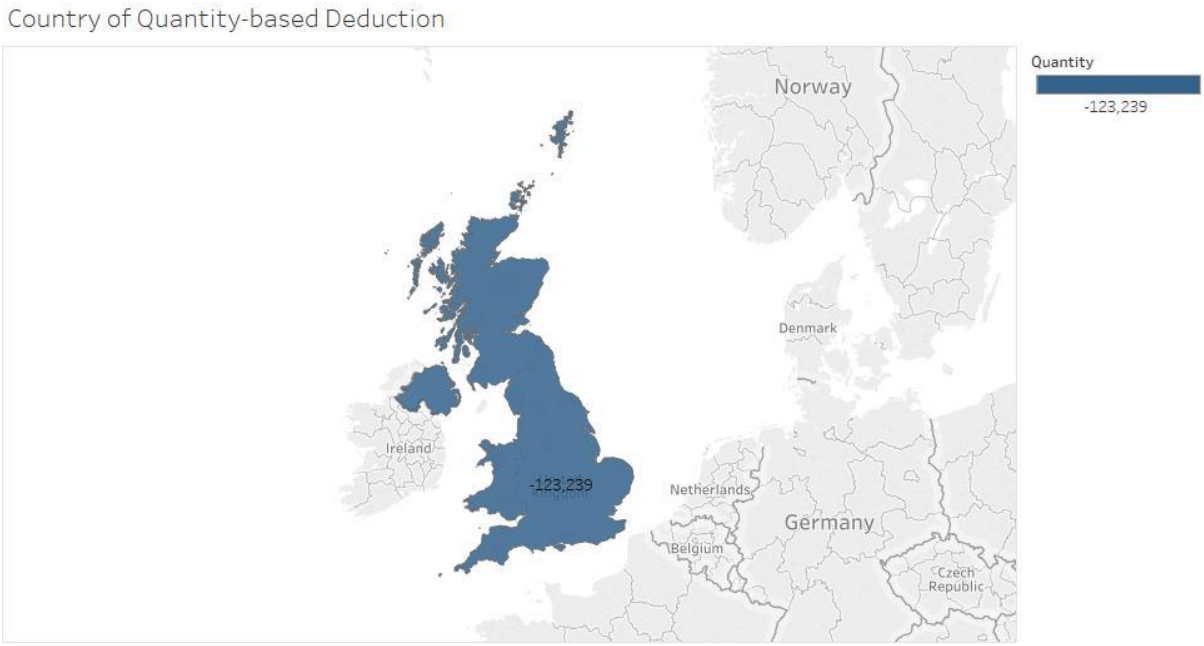
Dashboard 2: Bottom 20 Items/Services Quantity-based

Here again Quantity is the independent variable to which importance is given. We shall again use aggregate function SUM(Quantity) to discover the relations.



Sum of Quantity for each Description. The data is filtered on Action (Description) and Action (Country). The Action (Description) filter keeps 4,197 members. The Action (Country) filter keeps 38 members. The view is filtered on Description, which has multiple members selected.

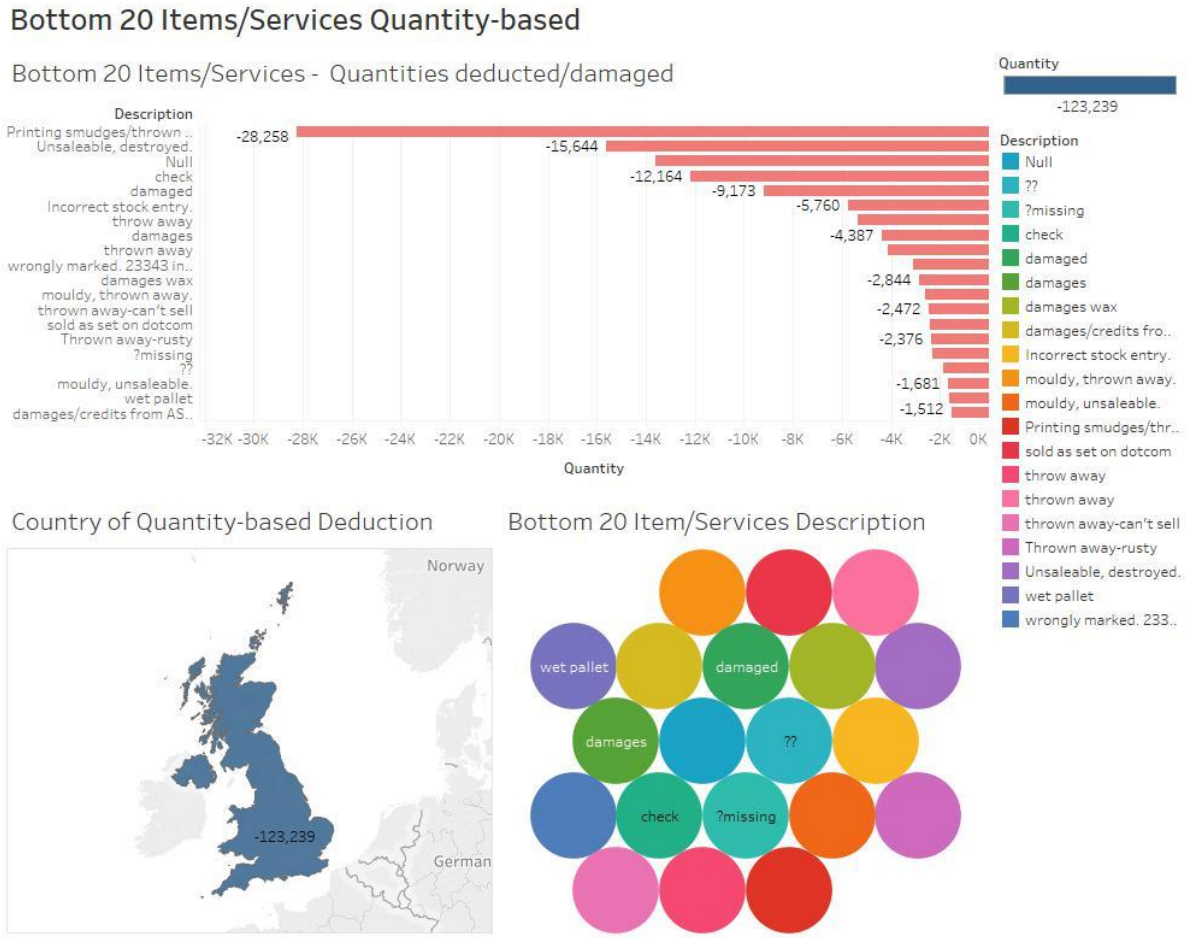
For Worksheet 1 – Bottom 20 Items/Services – Quantities deducted/damaged, we have SUM(Quantity) and Description as the 2 variables & Description is arranged ascendingly by quantity & filtered to get the bottom 20 items/services/reasons for quantity deductions.



Map based on Longitude (generated) and Latitude (generated). Color shows sum of Quantity. Details are shown for Country. The data is filtered on Description and Action (Description). The Description filter has multiple members selected. The Action (Description) filter keeps 4,197 members.

For Worksheet 2 – Country of Quantity-based Deduction, country & description is taken where description is filtered by bottom 20 SUM(Quantity) & countries are coloured according quantity deductions.

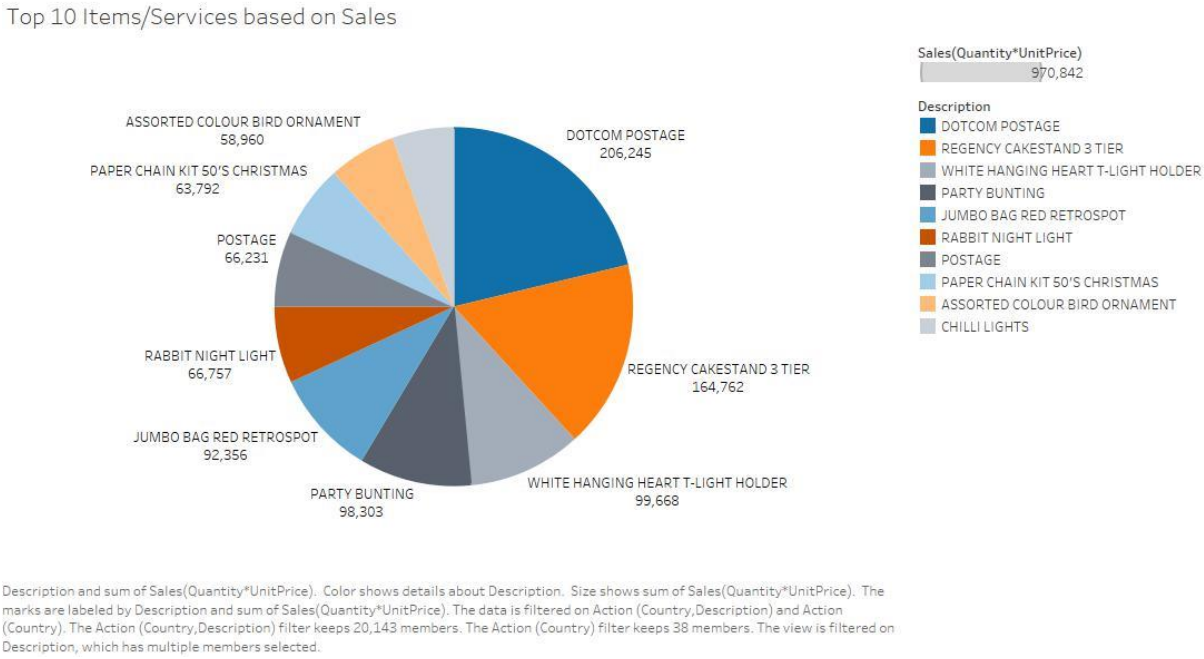
For Worksheet 3 – Bottom 20 Item/Services Description, I have arranged the bottom 20 items/services/reasons based on Quantity to give us a detail of the stock code for stock/reason emphasis. So, we get the names & stock codes of the bottom 20 items.



Additionally, we come to know that these quantity deductions are only happening in United Kingdom which means the damage/deduction is due to some human mistake or machine failure happening in UK. This means we can narrow our search to UK & focus on fixing the mistakes there.

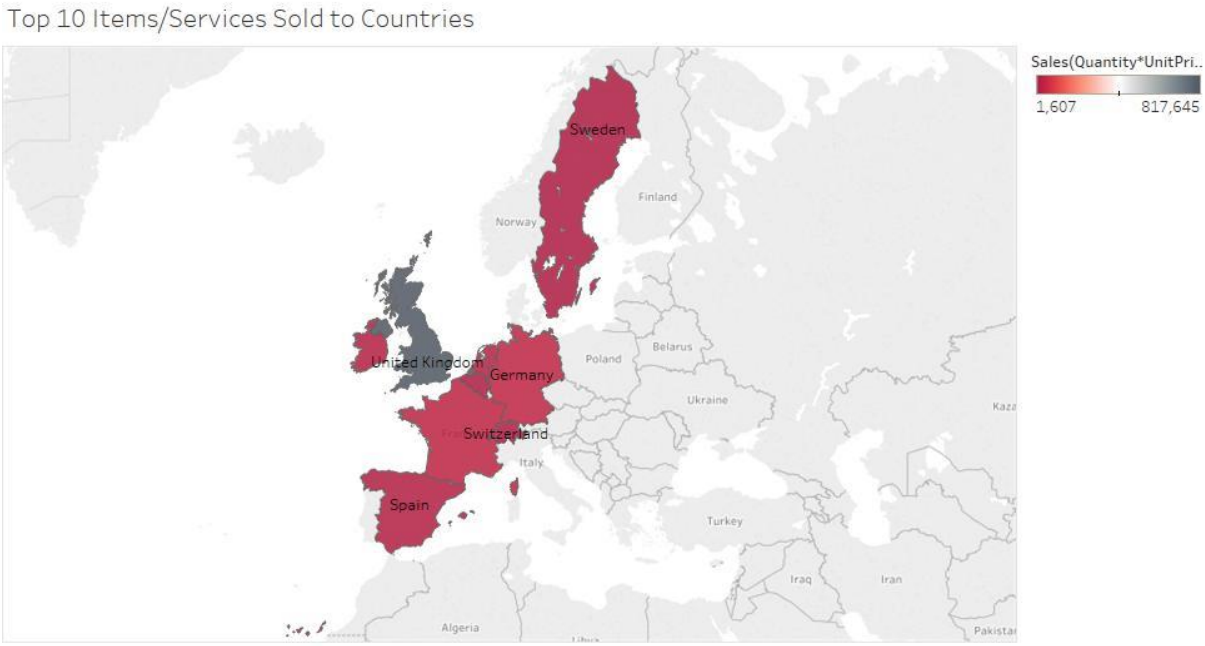
Dashboard 3: Sales Distribution of Top 10 Items/Services per Country

Here Sales is our variable of importance. Sales is aggregated to SUM(Sales).



Description and sum of Sales(Quantity*UnitPrice). Color shows details about Description. Size shows sum of Sales(Quantity*UnitPrice). The marks are labeled by Description and sum of Sales(Quantity*UnitPrice). The data is filtered on Action (Country,Description) and Action (Country). The Action (Country,Description) filter keeps 20,143 members. The Action (Country) filter keeps 38 members. The view is filtered on Description, which has multiple members selected.

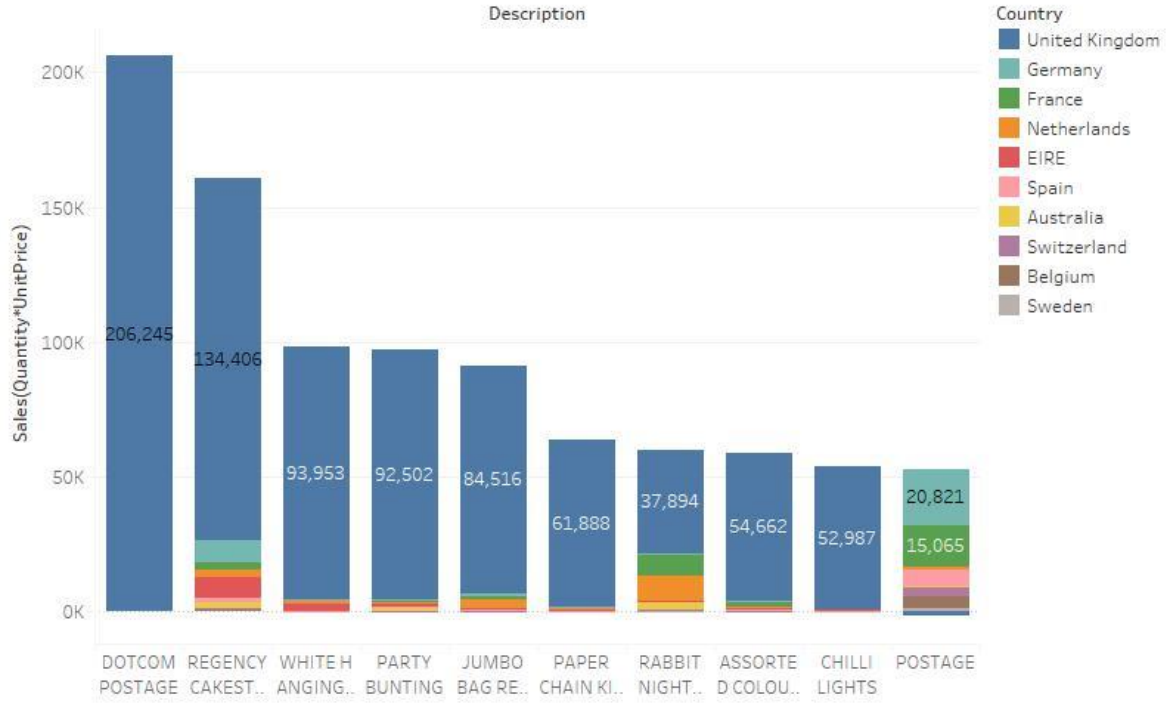
For Worksheet 1: Top 10 Items/Services based on Sales – The given Pie chart provides the top 10 Items/Services based on Sales by filtering item Description with SUM(Sales) for top 10 items & arranging it descending by Sales. So, we get the total sales on each of these items/services.



Map based on Longitude (generated) and Latitude (generated). Color shows sum of Sales(Quantity*UnitPrice). The marks are labeled by Country. Details are shown for Country. The data is filtered on Description, Action (Description) and Action (Country,Description). The Description filter has multiple members selected. The Action (Description) filter keeps 4,197 members. The Action (Country,Description) filter keeps 20,143 members. The view is filtered on Country, which has multiple members selected.

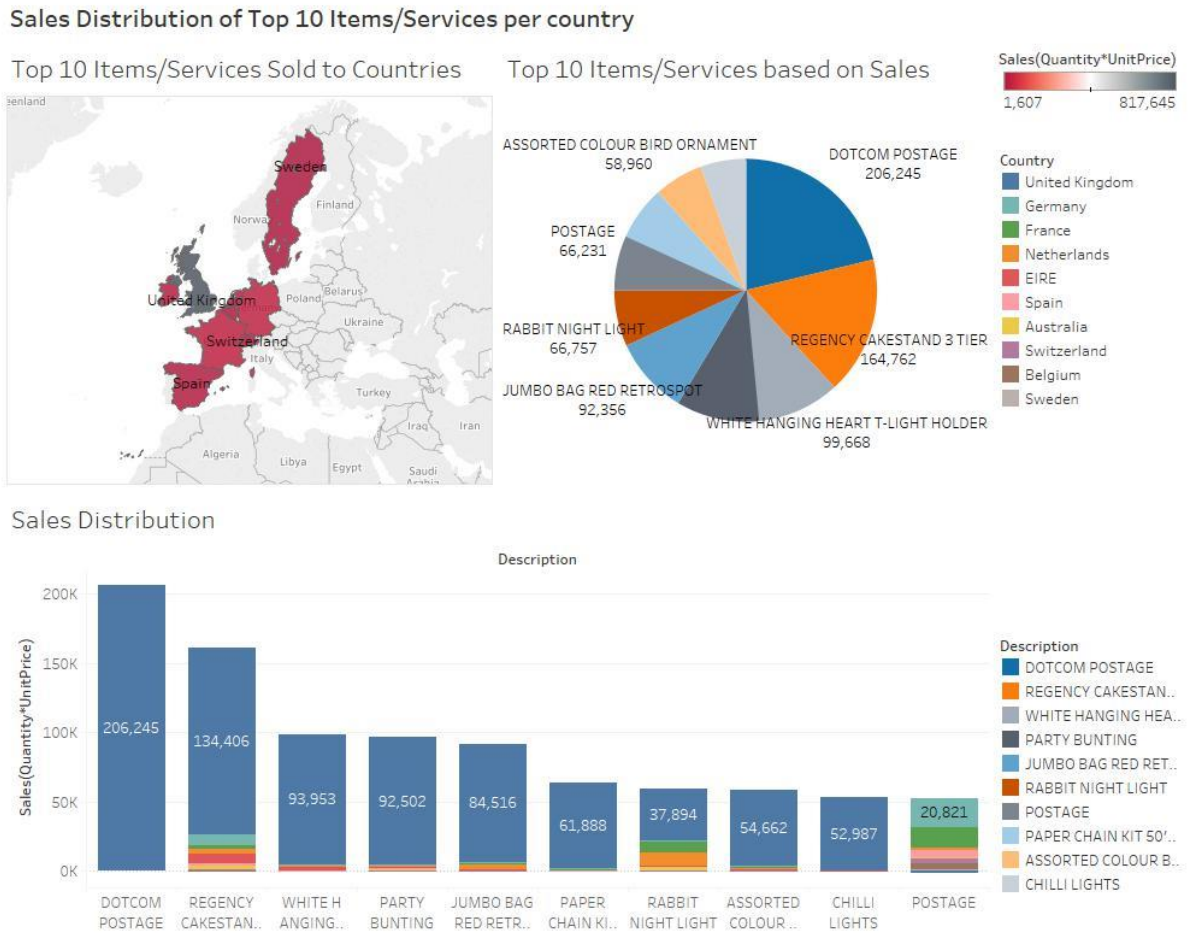
For Worksheet 2: Top 10 Items/Services Sold to Countries – It gives us the top 10 countries from where the sales of top 10 items/services have been produced. We do this by taking Description, filtering it with top 10 by SUM(Sales) & then taking Country, filtering it with top 10 by SUM(Sales). Colours are according to sales.

Sales Distribution



Sum of Sales(Quantity*UnitPrice) for each Description. Color shows details about Country. The data is filtered on Action (Description) and Action (Country). The Action (Description) filter keeps 4,197 members. The Action (Country) filter keeps 38 members. The view is filtered on Country and Description. The Country filter has multiple members selected. The Description filter has multiple members selected.

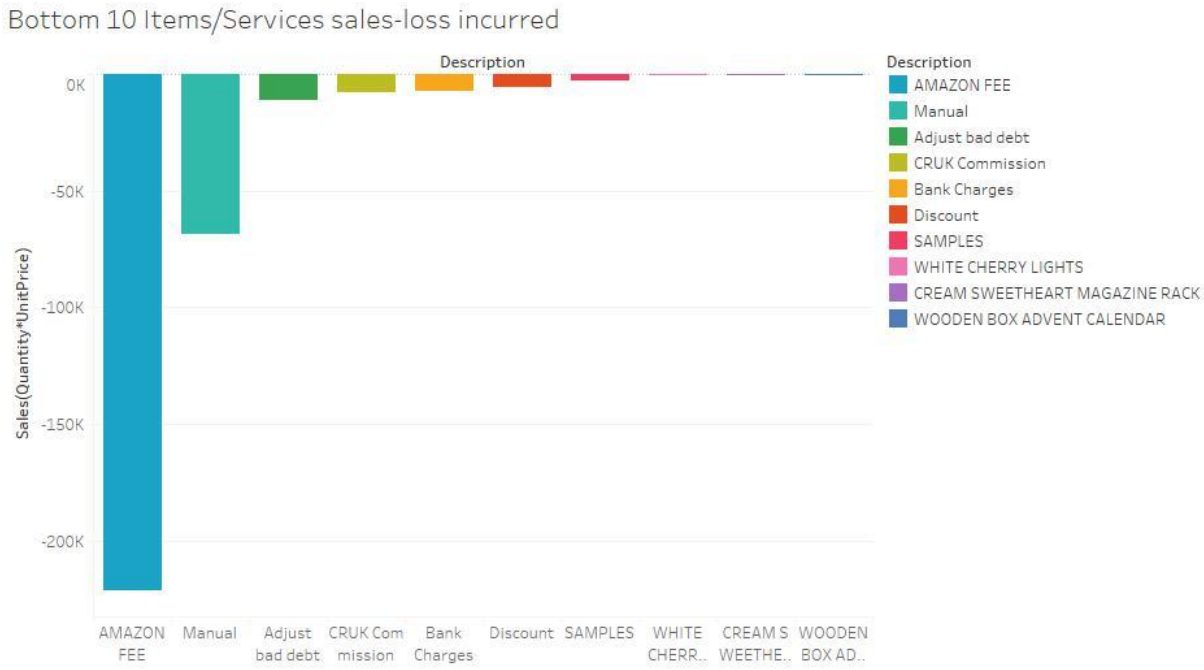
For Worksheet 3: Sales Distribution – Here we get a distribution of sales for the top 10 items/services per top 10 country based on sales. We take Description & SUM(Sales) where description is filtered by top 10 SUM(Sales). This is coloured according to top 10 SUM(Sales) – Countries.



Resulting dashboard gives us the distribution of sales of the top 10 most sold items in the top 10 countries where sales is the highest.

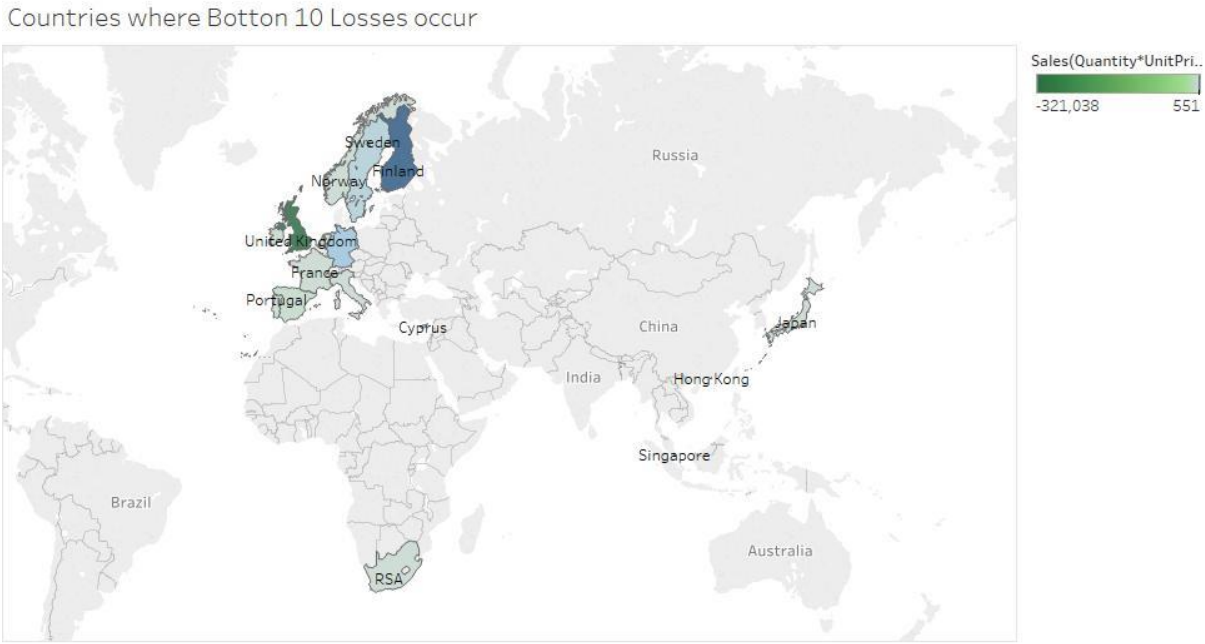
Dashboard 4: Bottom 10 Items/Services Sales Loss

Here Sales is the important factor again because now, we're interested in the Items/Services where the losses have occurred.



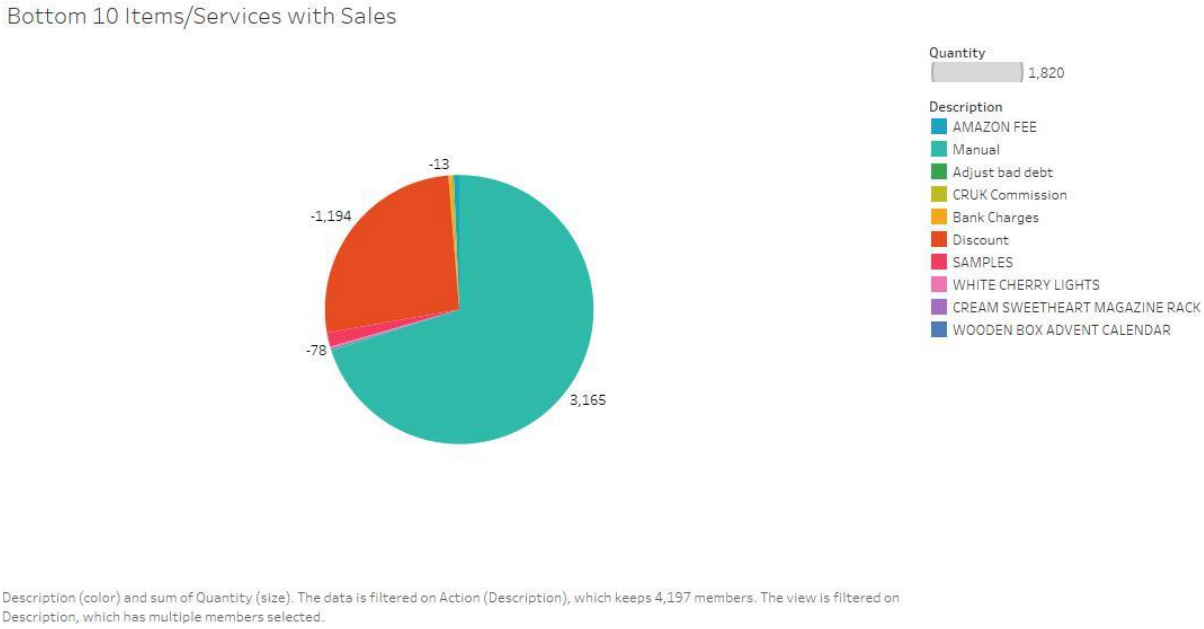
Sum of Sales(Quantity*UnitPrice) for each Description. Color shows details about Description. The data is filtered on Action (Description), which keeps 4,197 members. The view is filtered on Description, which has multiple members selected.

For Worksheet 1: Bottom 10 Items/Services sales-loss incurred – Taking SUM(Sales) & item Description as the 2 variables where Description is filtered with bottom 10 SUM(Sales) into consideration & sorted ascending by sales, the graph shows us a negative bar graph with the bottom 10 Items/Services where loss has occurred.

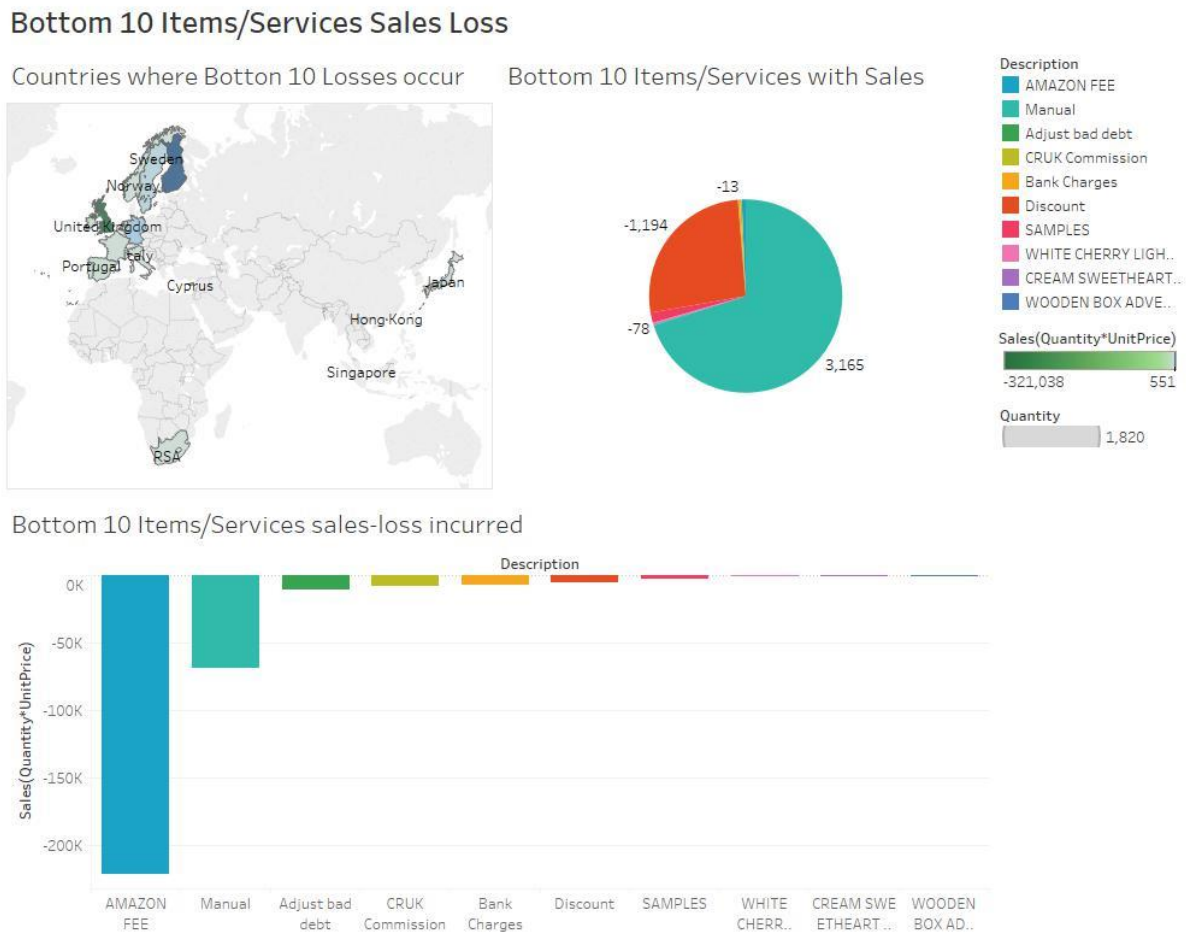


Map based on Longitude (generated) and Latitude (generated). Color shows sum of Sales(Quantity*UnitPrice). The marks are labeled by Country. Details are shown for Country. The data is filtered on Description and Action (Description). The Description filter has multiple members selected. The Action (Description) filter keeps 4,197 members.

For Worksheet 2: Countries where Bottom 10 Losses occur - We investigate the countries where this loss of revenue is taking place. This is done by considering Country & Description where Description is filtered by bottom 10 SUM(sales) giving us the bottom 10 countries with the least sales.



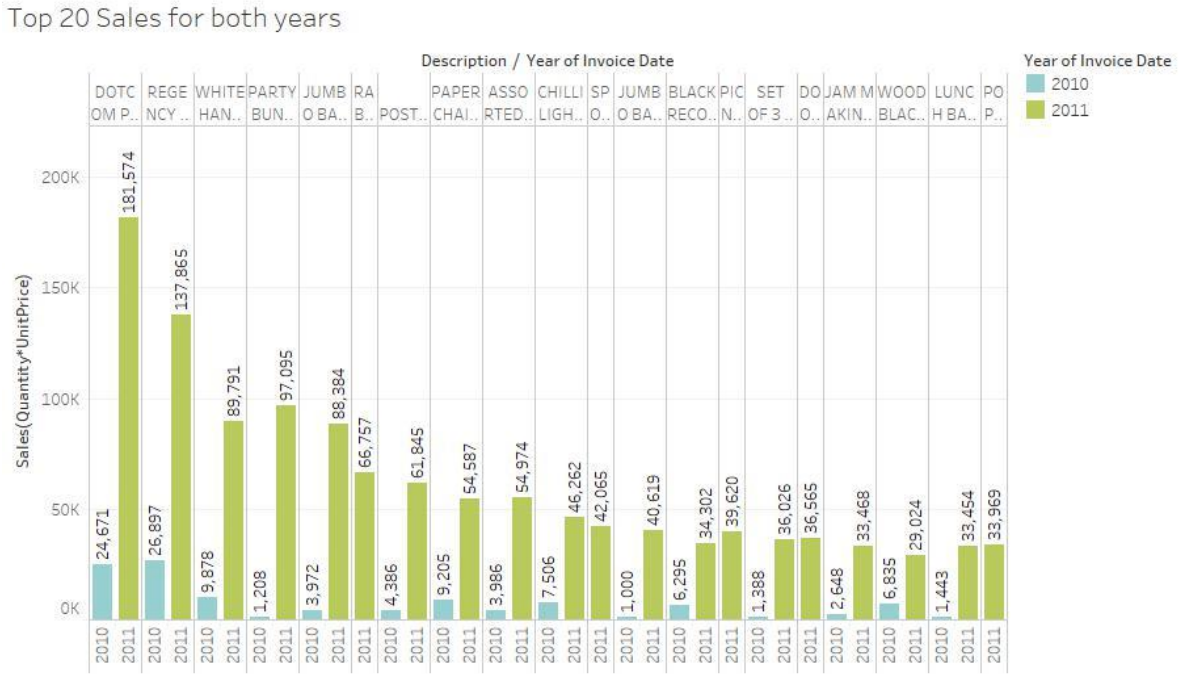
For Worksheet 3: Bottom 10 Items/Services with Quantity – since we know which items are causing the loss & where (country) they're being caused, to know the cause of the loss of revenue we consider Description which is filtered by bottom 10 & sorted ascending by SUM(Sales). Colours on the pie chart show us item's description.



The resulting dashboard gives us details about the loss of revenue in sales occurring due to the bottom 10 items that is the ones which are causing a loss with sales values going in negative.

Dashboard 5: Top Items/Services list w.r.t Sales for both years

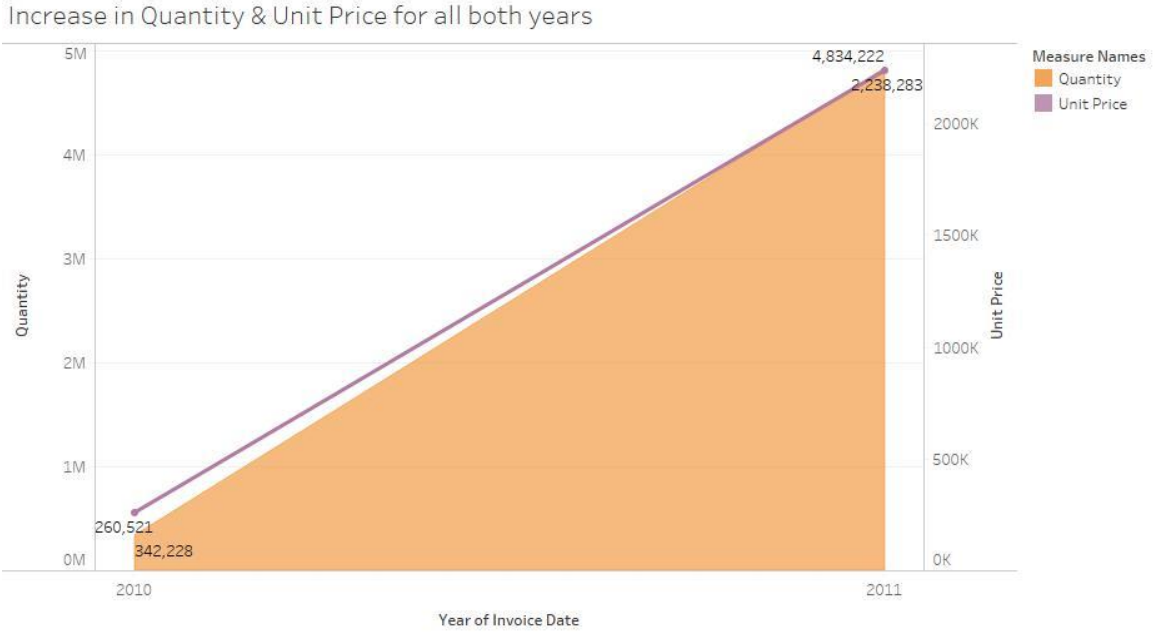
Sales & InvoiceDate are the variables to be used to discover the relationships. We aggregate Sales as SUM(Sales).



Sum of Sales(Quantity*UnitPrice) for each Invoice Date Year broken down by Description. Color shows details about Invoice Date Year. The data is filtered on Action (Description,YEAR(Invoice Date)), which keeps 6,847 members. The view is filtered on Description and Action (YEAR(Invoice Date)). The Description filter has multiple members selected. The Action (YEAR(Invoice Date)) filter keeps 2 members.

For Worksheet 1: Top 20 Sales for both years

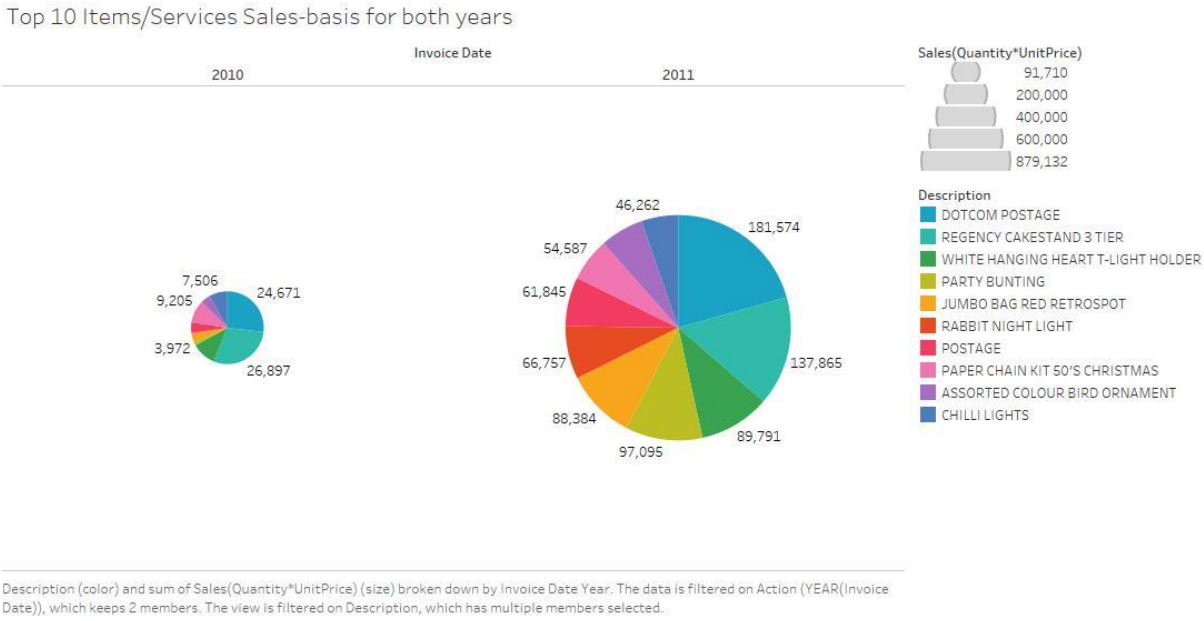
Here, we have taken item Description filtered by top 20 & sorted descending by SUM(Sales) along with YEAR(InvoiceDate) in the columns. SUM(Sales) is taken in the row. The bar graph show us the sales for these top 20 items in both years.



The trends of Quantity and Unit Price for Invoice Date Year. Color shows details about Quantity and Unit Price. The data is filtered on Action (Description,YEAR(Invoice Date)), which keeps 6,847 members.

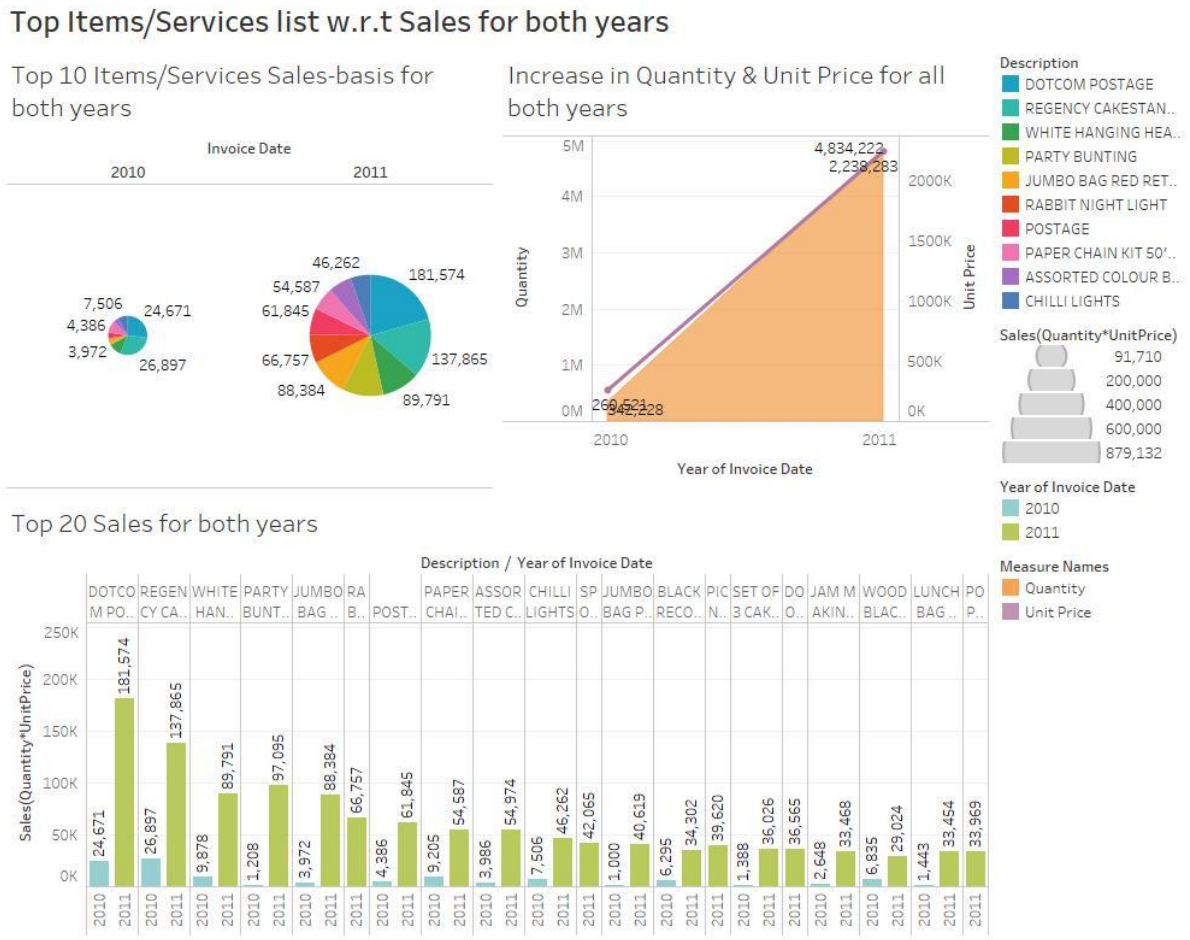
For Worksheet 2: Increase in Quantity & Unit Price for all both years

We have taken YEAR(InvoiceDate) in columns & SUM(Quantity) along with SUM(UnitPrice) in rows. Purple line is the unit price of each item & the orange area is the quantity of items purchased in both years. The area graph shows us this increase.



For Worksheet 3: Top 10 Items/Services Sales-basis for both years

The 2 pie charts depict the top 10 items which are bought the most in both years. For this, we have taken Description, filtered & sorted descending by top 10 SUM(Sales) along with YEAR(InvoiceDate). The colours are for the top 10 items.

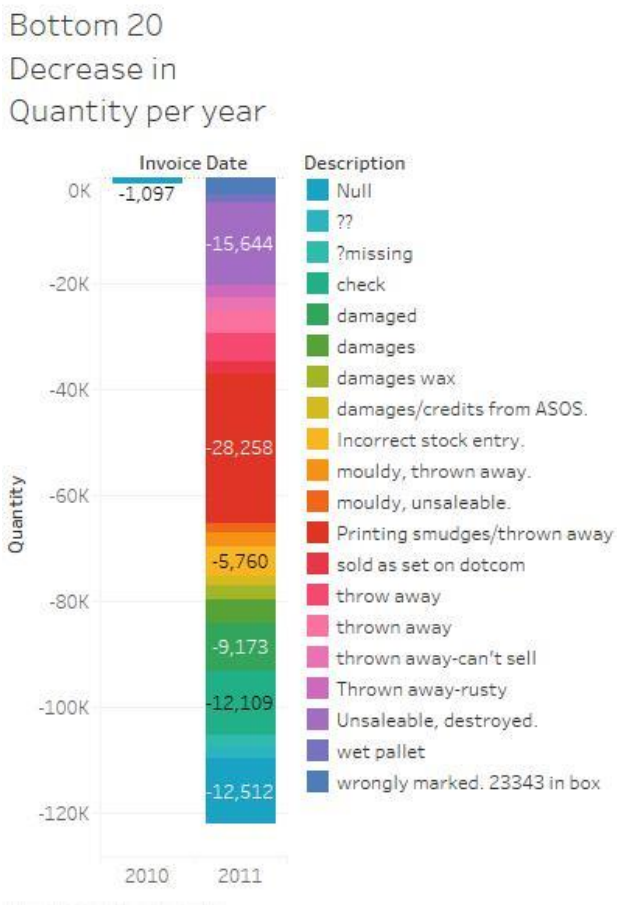


Thus, the resulting dashboard gives us the details of the top 10 & the top 20 most bought items in both years & compares it to the increase in Quantity & UnitPrice to check their popularity and price change over the years was justified or not.

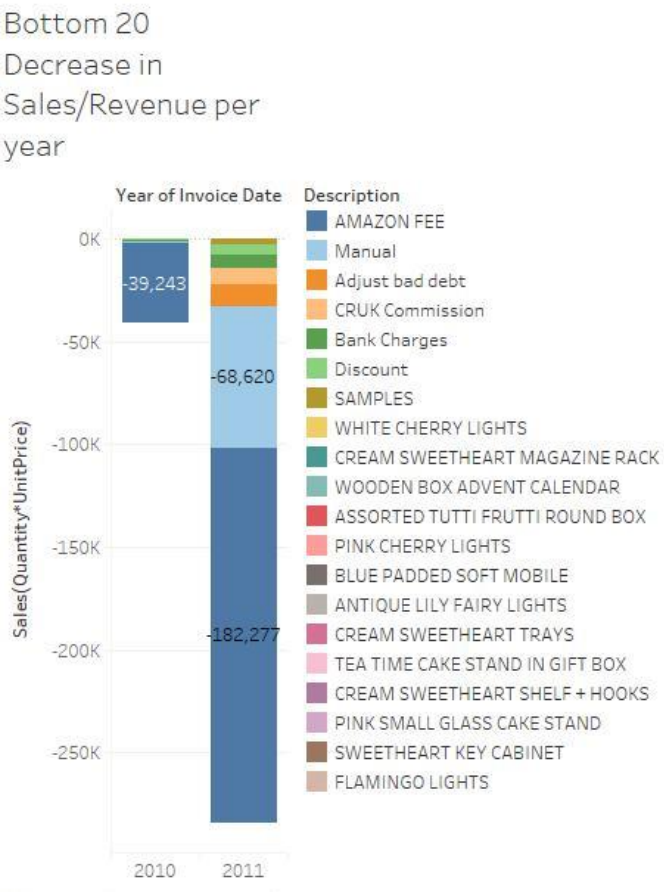
Dashboard 6: Decrease in Quantity, Sales/Revenue for Bottom 20 Items/Services

For Worksheet 1: Bottom 20 Decrease in Quantity per year

Here, Bottom 20 items are our preference and deserve our focus to get details on why there are not performing well in sales.

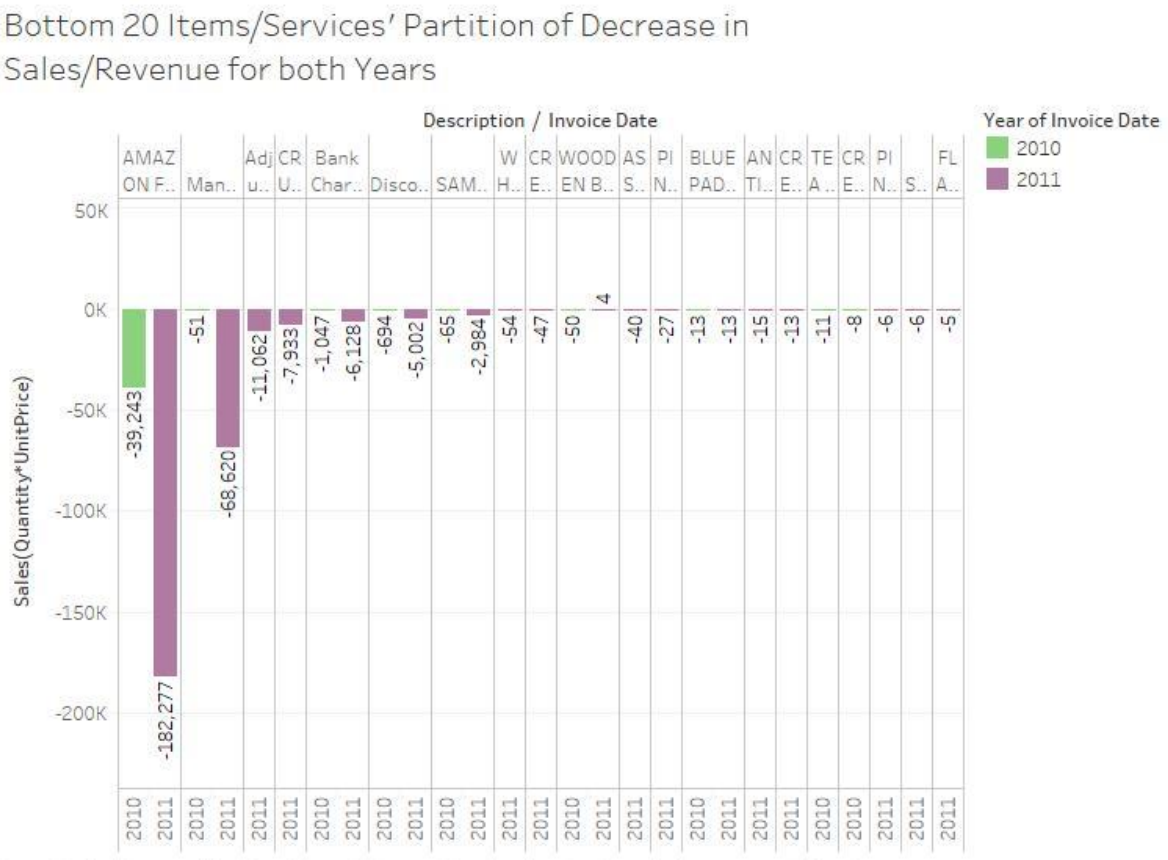


We have considered YEAR(InvoiceDate) and SUM(Quantity) where we have applied a filter of Description which is only for the Bottom 20 items by SUM(Quantity). So, we have a negative bar graph partitioned by bottom 20, less quantified items for both years



For Worksheet 2: Bottom 20 Decrease in Sales/Revenue per year

Now, we're taking YEAR(InvoiceDate) & notably, SUM(Sales) where we apply a filter of Description which is for Bottom 20 items by SUM(Sales). So, we have a negative bar graph partitioned by bottom 20, least sold items for both years.



For Worksheet 3: Bottom 20 Items/Services' Partition of Decrease in Sales/Revenue for both Years

We have taken description & YEAR(InvoiceDate) in columns and SUM(Sales) in rows where Description is filtered by bottom 20 items & sorted ascending by SUM(Sales). So, we get dual bar graphs on revenues lost for bottom 20 items in both years.

Stock Code with Description

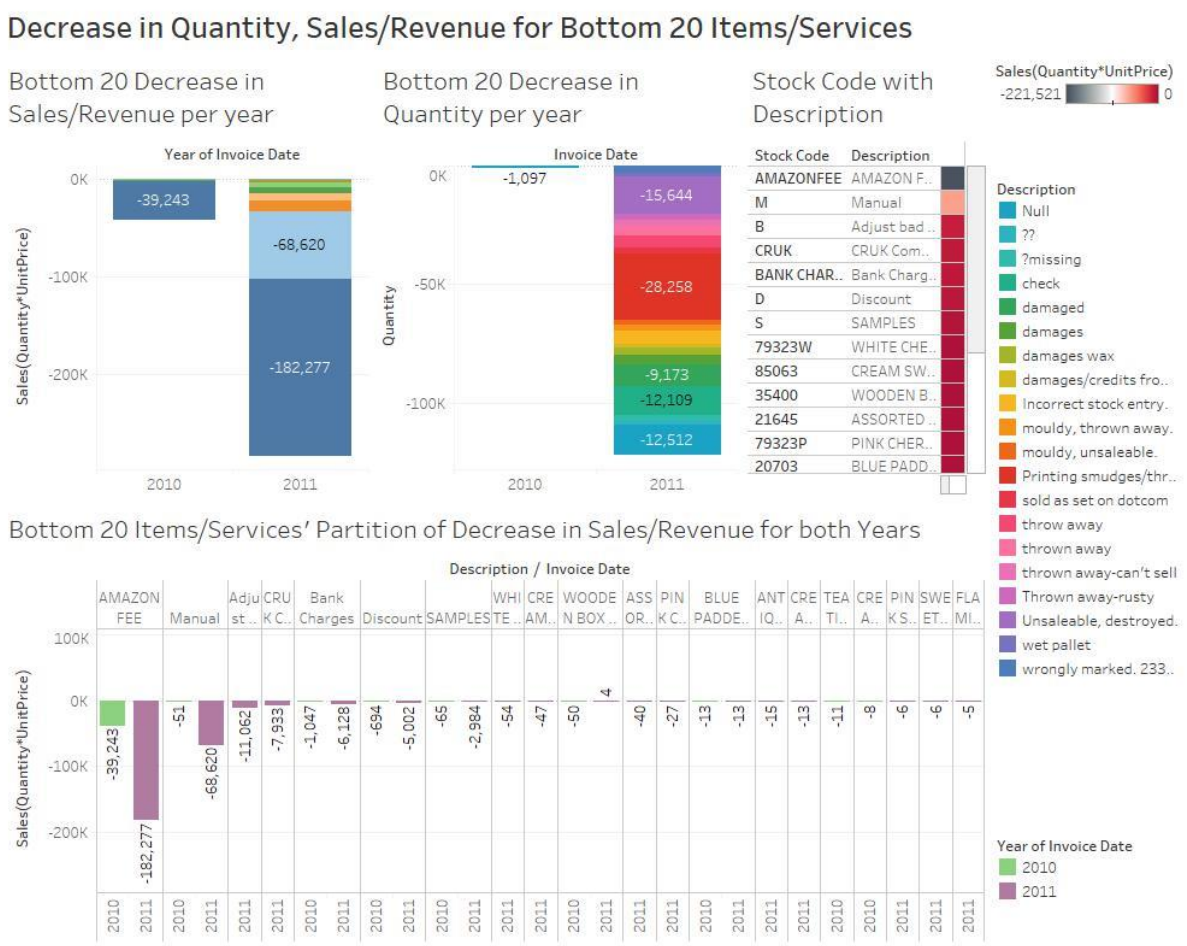
Stock Code	Description
AMAZONFEE	AMAZON FEE
M	Manual
B	Adjust bad debt
CRUK	CRUK Commission
BANK CHARGES	Bank Charges
D	Discount
S	SAMPLES
79323W	WHITE CHERRY LIGHTS
85063	CREAM SWEETHEART MA..
35400	WOODEN BOX ADVENT C..
21645	ASSORTED TUTTI FRUTTI ..
79323P	PINK CHERRY LIGHTS
20703	BLUE PADDED SOFT MOBI..
85042	ANTIQUE LILY FAIRY LIGH..
85065	CREAM SWEETHEART TR..
37503	TEA TIME CAKE STAND IN ..
85068	CREAM SWEETHEART SH..
84750A	PINK SMALL GLASS CAKE ..
84839	SWEETHEART KEY CABIN..
79320	FLAMINGO LIGHTS
21656	samples

Sales(Quantity*UnitPri..

-221,521 0

For Worksheet 4: Stock Code with Description

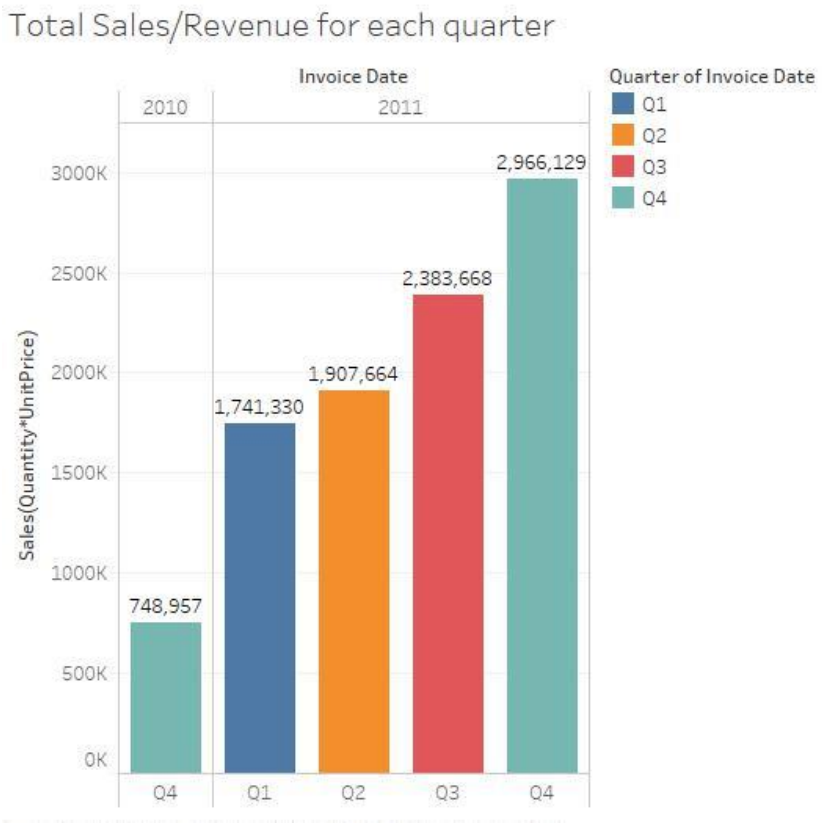
The colour table represents the stock code of each of these bottom 20 items by considering StockCode & Description where description is filtered by bottom 20 items & sorted ascending by SUM(Sales). Colours represent the SUM(Sales).



The resulting dashboard gives us information on decrease in Quantity, Sales/Revenue for each of the Bottom 20 items. Using this information, we could focus strategies to improve sales of such items.

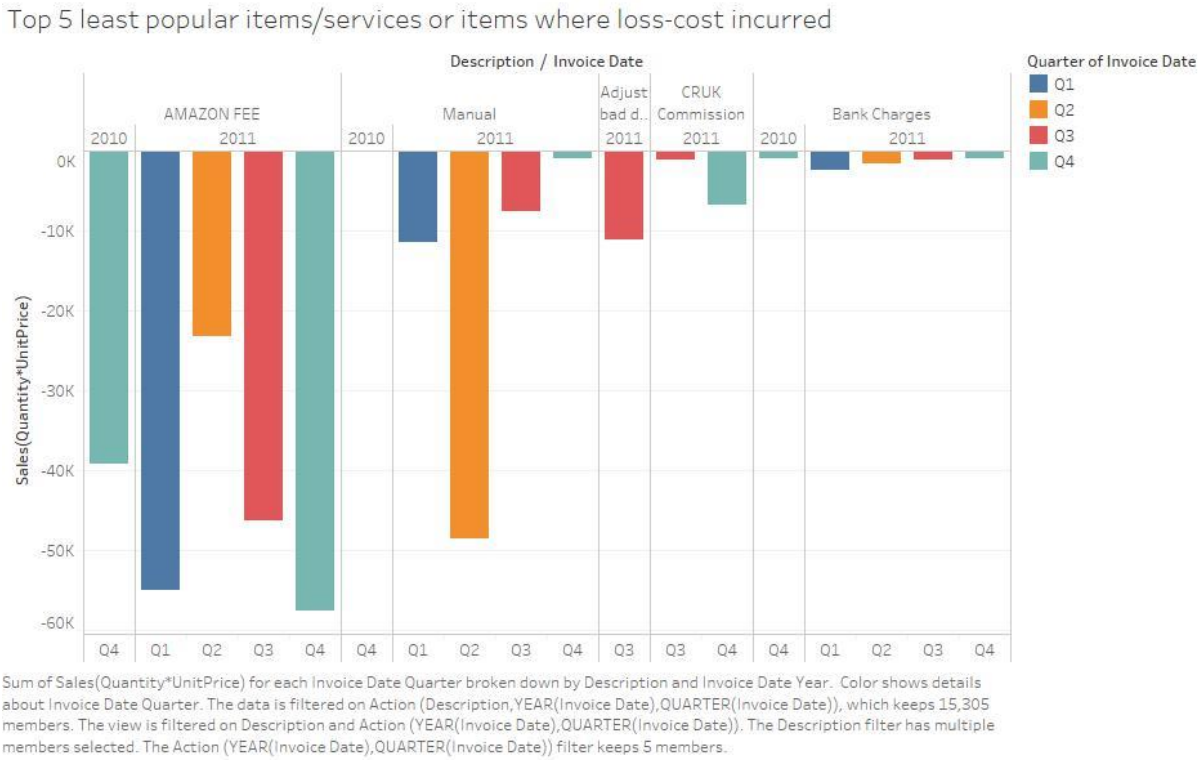
Dashboard 7: Popularity of Item/Services on Sales-basis

Here, our focus is to find out the Sales of most or least popular items according to sales over yearly quarters. We shall do this by considering QUARTER(InvoiceDate) as the class.



For Worksheet 1: Total Sales/Revenue for each quarter

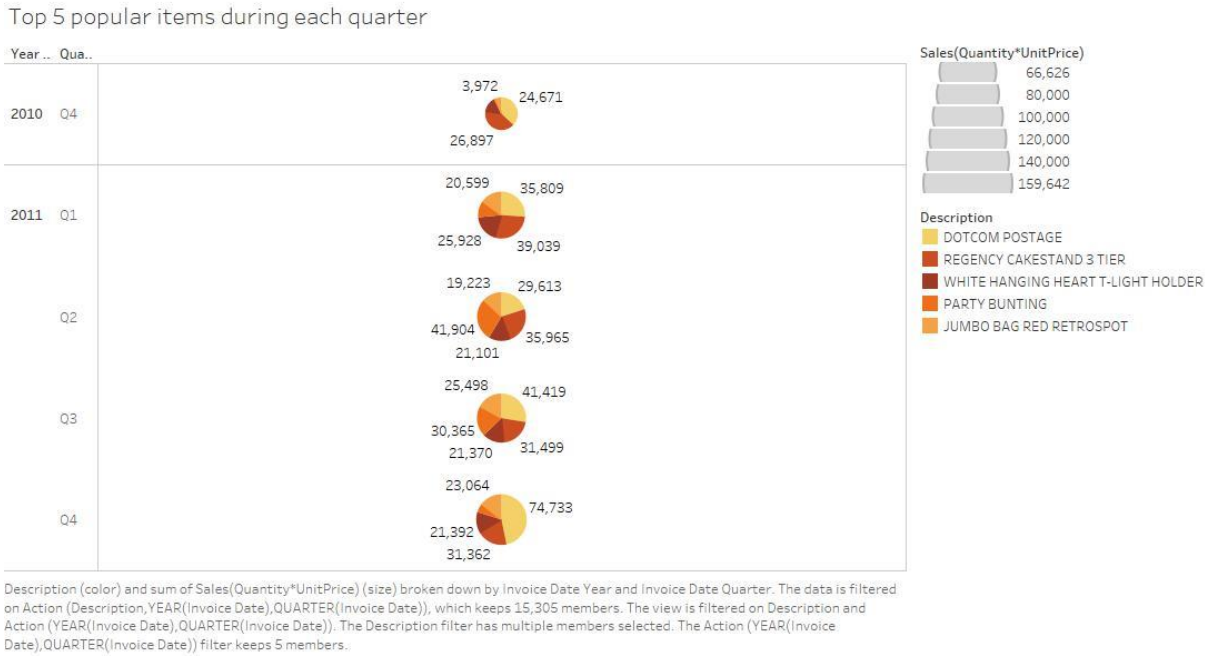
We take Quarter(InvoiceDate) under YEAR(InvoiceDate) & SUM(Sales) as our variables. This gives us a sales chart distributed over yearly quarters starting from Q4 2010 to Q4 2011.



Sum of Sales(Quantity*UnitPrice) for each Invoice Date Quarter broken down by Description and Invoice Date Year. Color shows details about Invoice Date Quarter. The data is filtered on Action (Description,YEAR(Invoice Date),QUARTER(Invoice Date)), which keeps 15,305 members. The view is filtered on Description and Action (YEAR(Invoice Date),QUARTER(Invoice Date)). The Description filter has multiple members selected. The Action (YEAR(Invoice Date),QUARTER(Invoice Date)) filter keeps 5 members.

For Worksheet 2: Top 5 least popular items/services or items where loss-cost incurred

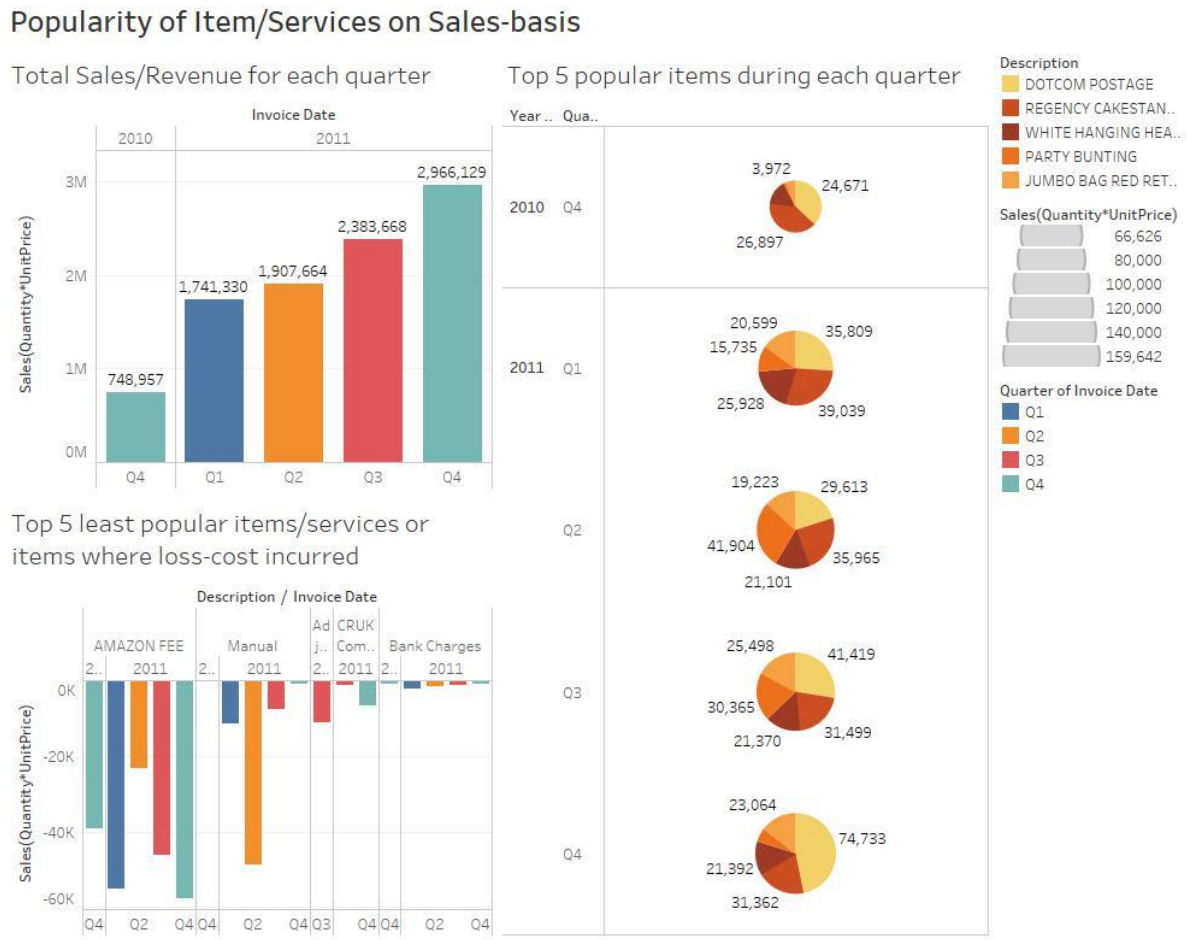
Here, we have considered Description which is filtered by bottom 5 & sorted ascending by SUM(Sales) along with Quarter(InvoiceDate), YEAR(InvoiceDate) in columns. We have SUM(Sales) in rows. This gives us the loss of sales due to the bottom 5, least purchased items over yearly quarters.



Description (color) and sum of Sales(Quantity*UnitPrice) (size) broken down by Invoice Date Year and Invoice Date Quarter. The data is filtered on Action (Description,YEAR(Invoice Date),QUARTER(Invoice Date)), which keeps 15,305 members. The view is filtered on Description and Action (YEAR(Invoice Date),QUARTER(Invoice Date)). The Description filter has multiple members selected. The Action (YEAR(Invoice Date),QUARTER(Invoice Date)) filter keeps 5 members.

For Worksheet 3: Top 5 popular items during each quarter

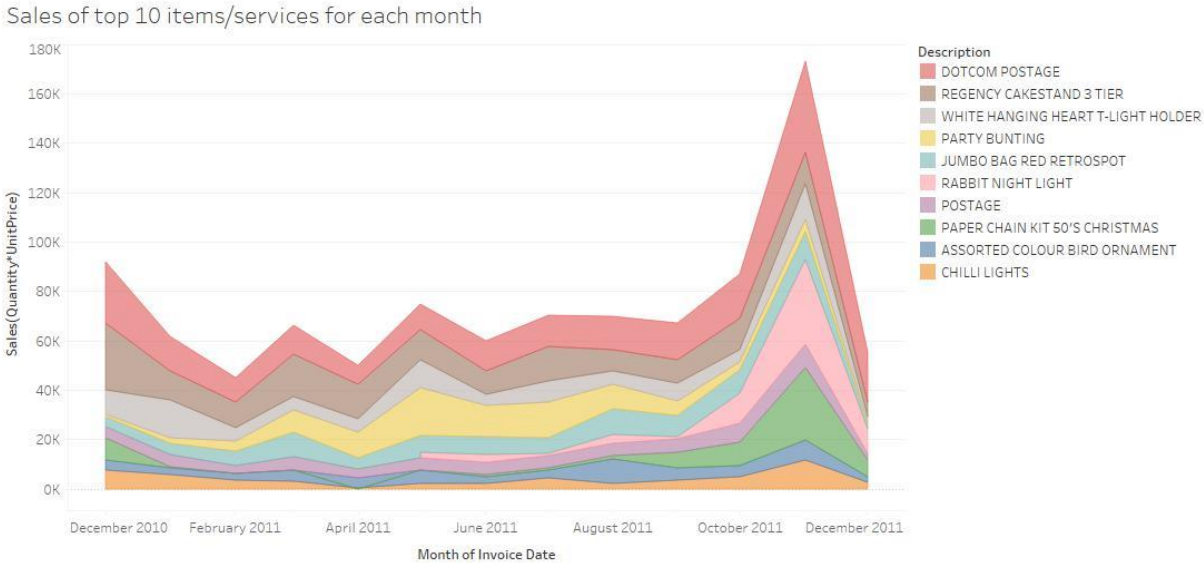
Here, we have Description which is filtered by top 5 & sorted descending by SUM(Sales) with Quarter(InvoiceDate), YEAR(InvoiceDate) in rows. We thus, have pie charts for each yearly quarter. This gives us the total sales of the top 5, most sold items over yearly quarters.



The resulting dashboard gives us a comparison of information between top 5 most bought items and bottom 5, loss incurring items based on sales for each yearly quarter.

Dashboard 8: Sales/Revenue of items/services for each month

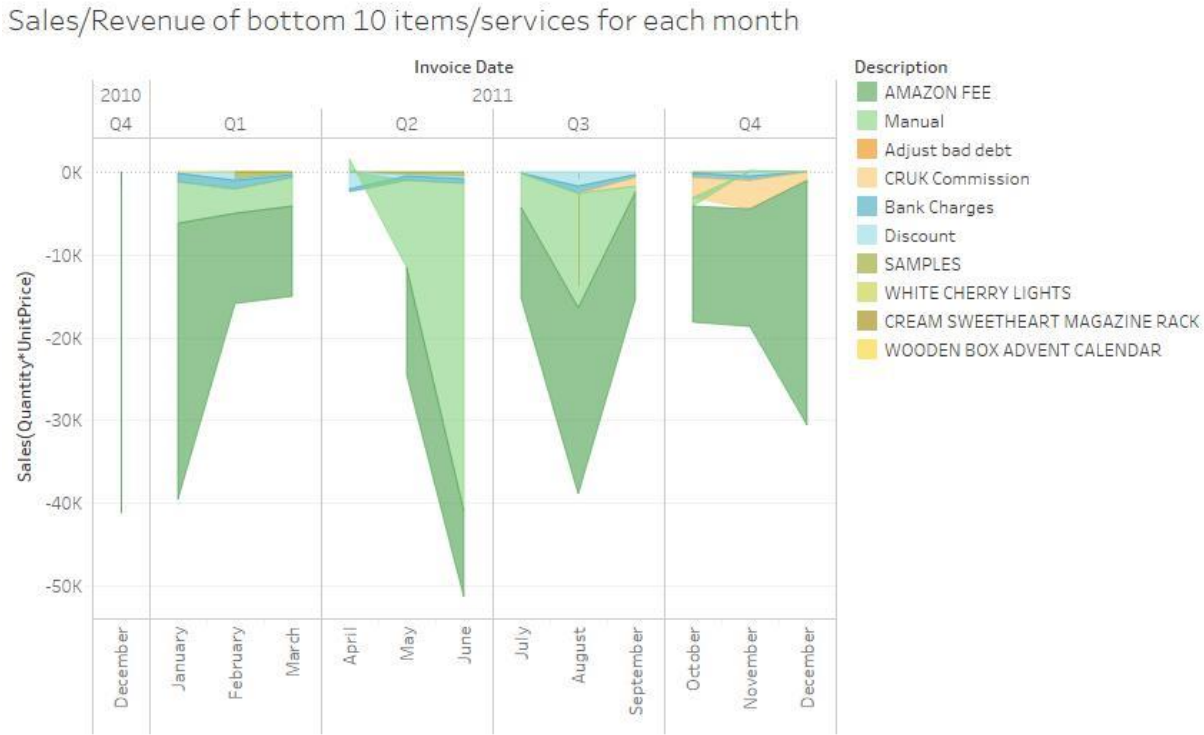
In this dashboard, we are considering the sales/revenue of items over all the months in the period.



The plot of sum of Sales(Quantity*UnitPrice) for Invoice Date Month. Color shows details about Description. The data is filtered on Action (International Shipping/Service?,YEAR(Invoice Date),QUARTER(Invoice Date),MONTH(Invoice Date)), Action (Description,YEAR(Invoice Date),MONTH(Invoice Date)) and Action (Description,YEAR(Invoice Date),QUARTER(Invoice Date),MONTH(Invoice Date)). The Action (International Shipping/Service?,YEAR(Invoice Date),QUARTER(Invoice Date),MONTH(Invoice Date)) filter keeps 26 members. The Action (Description,YEAR(Invoice Date),MONTH(Invoice Date)) filter keeps 34,137 members. The Action (Description,YEAR(Invoice Date),QUARTER(Invoice Date),MONTH(Invoice Date)) filter keeps 34,137 members. The view is filtered on Description, which has multiple members selected.

For Worksheet 1: Sales of top 10 items/services for each month

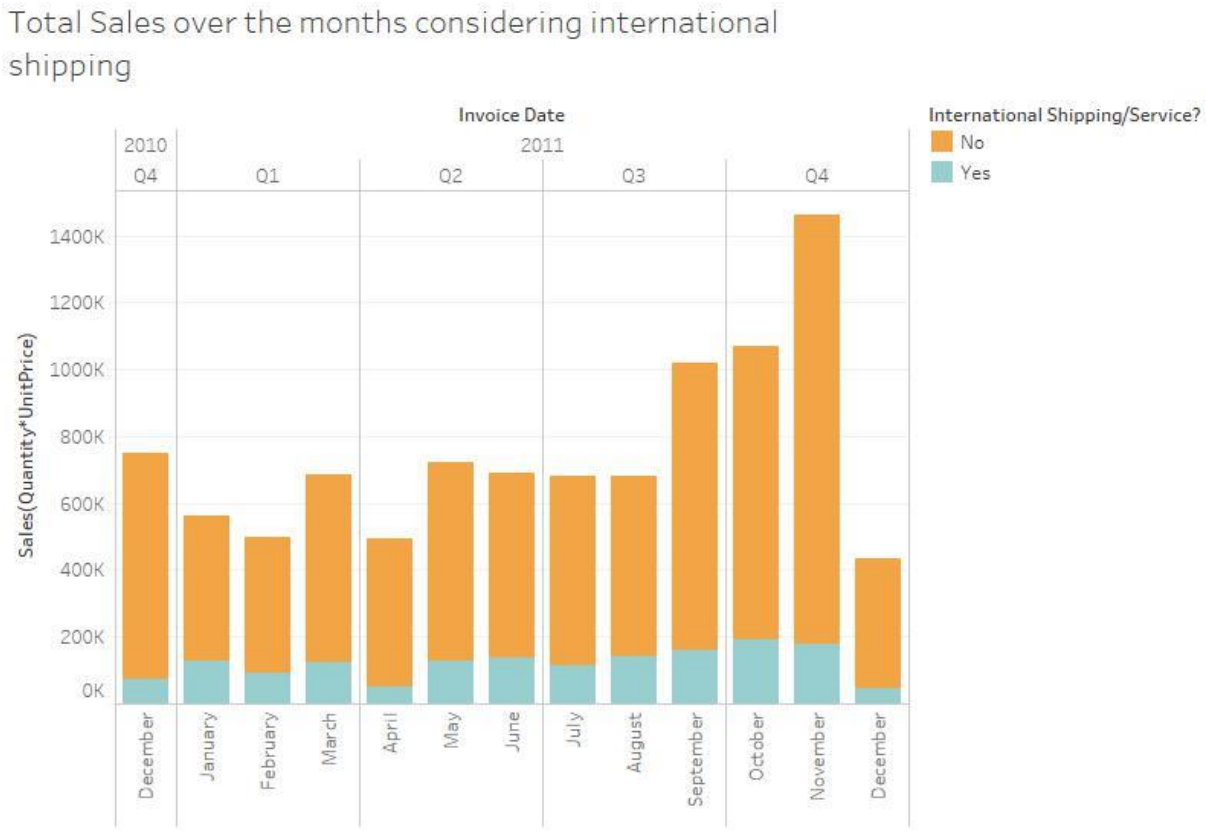
We have MONTH(InvoiceDate) in columns & SUM(Sales) in rows. We apply a filter of Description where Description is filtered by top 10 & sorted descending by SUM(Sales). The area graph thus shows us the top 10 sold items over the months.



Sum of Sales(Quantity*UnitPrice) for each Invoice Date Month broken down by Invoice Date Year and Invoice Date Quarter. Color shows details about Description. The data is filtered on Action (Description,MONTH(Invoice Date)) and Action (International Shipping/Service?,YEAR(Invoice Date),QUARTER(Invoice Date),MONTH(Invoice Date)). The Action (Description,MONTH(Invoice Date)) filter keeps 34,137 members. The Action (International Shipping/Service?,YEAR(Invoice Date),QUARTER(Invoice Date),MONTH(Invoice Date)) filter keeps 26 members. The view is filtered on Description, which has multiple members selected.

For Worksheet 2: Sales/Revenue of bottom 10 items/services for each month

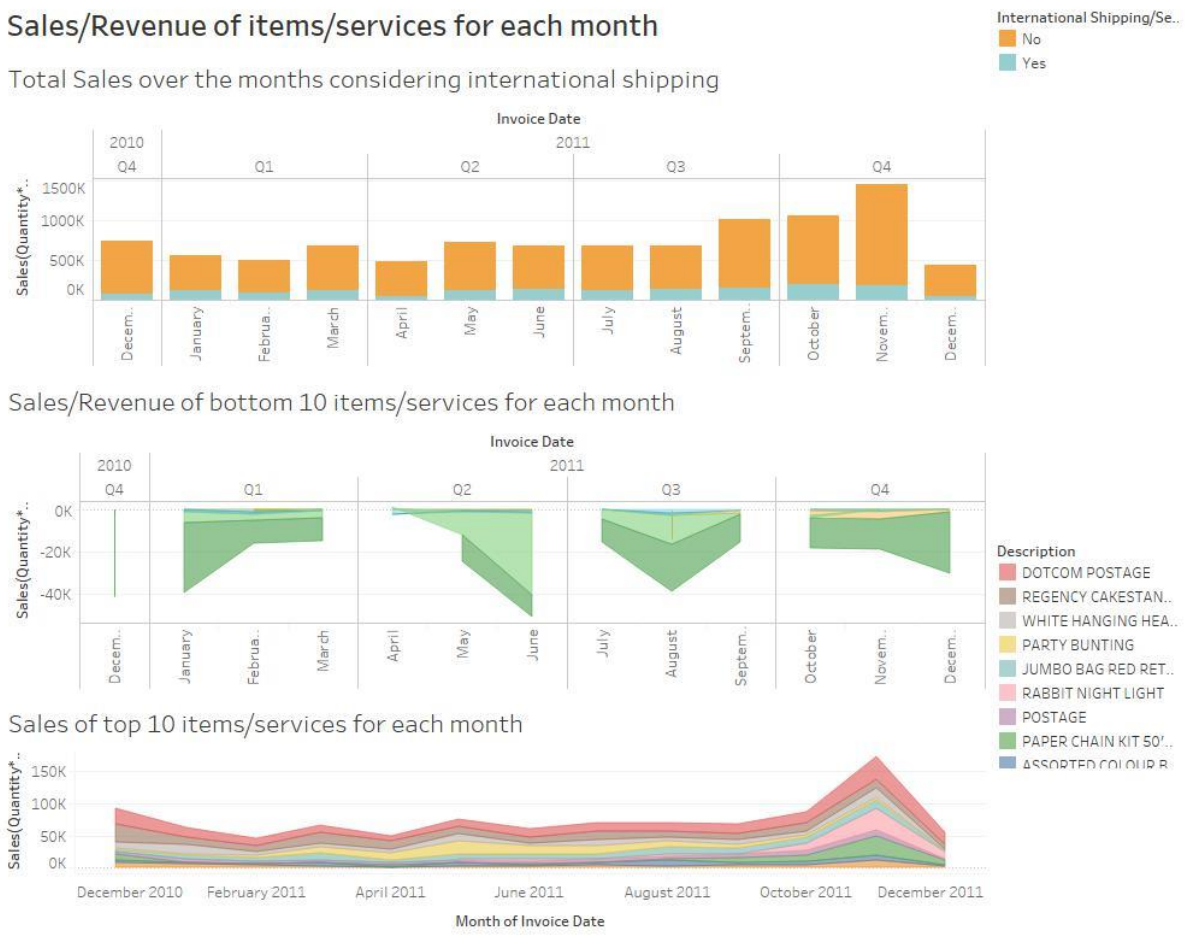
Now, we have YEAR(InvoiceDate), QUARTER(InvoiceDate), MONTH(InvoiceDate) in columns & SUM(Sales) in rows. Also, Description is filtered bottom 10 & sorted ascending by SUM(Sales). We have an area graph showing us the loss of revenue of these bottom 10 items over several months.



Sum of Sales(Quantity*UnitPrice) for each Invoice Date Month broken down by Invoice Date Year and Invoice Date Quarter. Color shows details about International Shipping/Service?. The data is filtered on Action (Description,MONTH(Invoice Date)), Action (Description,YEAR(Invoice Date),MONTH(Invoice Date)) and Action (Description,YEAR(Invoice Date),QUARTER(Invoice Date),MONTH(Invoice Date)). The Action (Description,MONTH(Invoice Date)) filter keeps 34,137 members. The Action (Description,YEAR(Invoice Date),MONTH(Invoice Date)) filter keeps 34,137 members. The Action (Description,YEAR(Invoice Date),QUARTER(Invoice Date),MONTH(Invoice Date)) filter keeps 34,137 members.

For Worksheet 3: Total Sales over the months considering international shipping

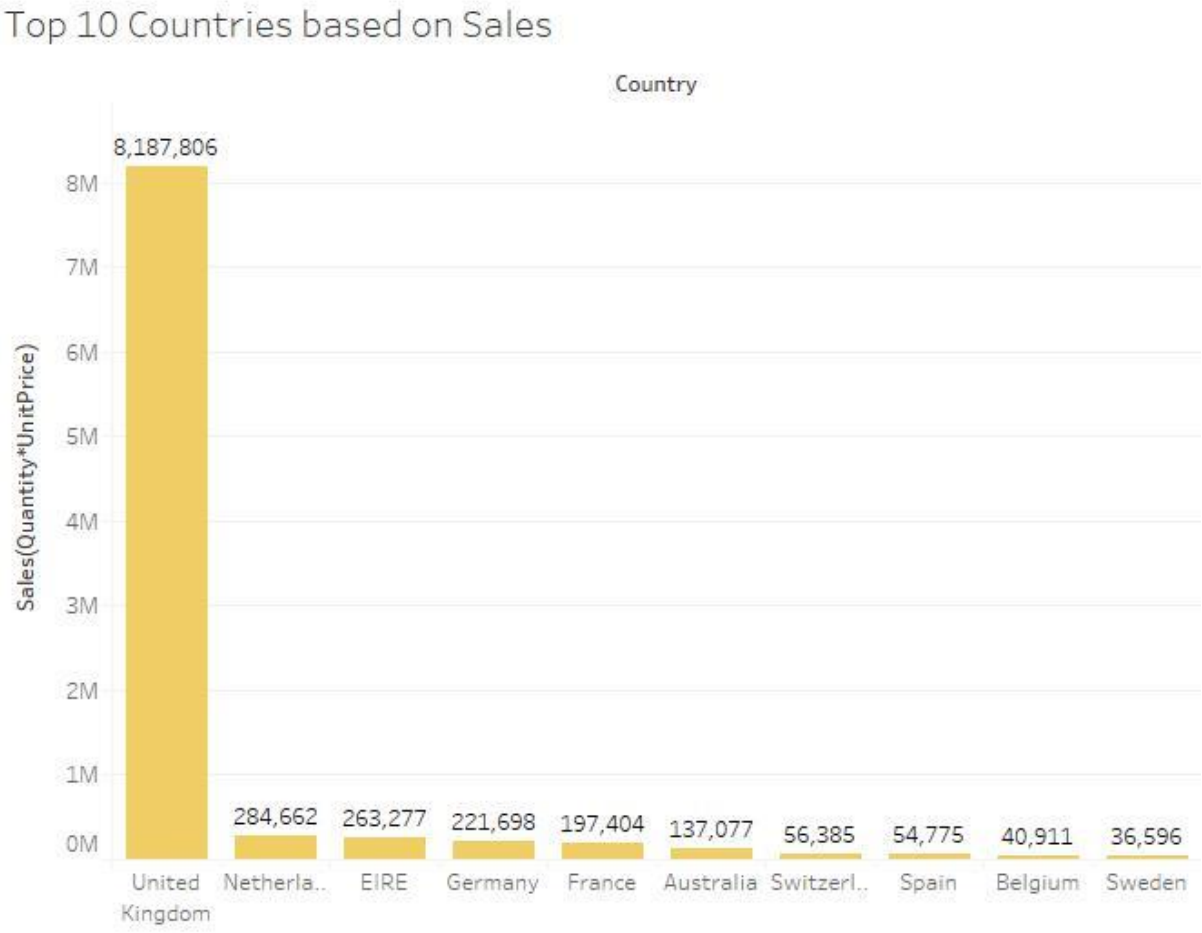
Here, we use YEAR(InvoiceDate), QUARTER(InvoiceDate), MONTH(InvoiceDate) in columns & SUM(Sales) in rows. The resulting bar graph shows us the total sales over all the months in that time period. Also, international shipping is applied as a colour filter to check for total revenue made from countries other than within UK.



Thus, the end dashboard gives notable information on the Sales/Revenue made during all the months in the period with an added observation on which sales are within UK & which are not.

Dashboard 9: Top 5 Countries Sales Distribution for Top 5 Items/Services

Now, for this dashboard we'll be considering the sales distribution with respect to the countries. For this we'll do as follows.



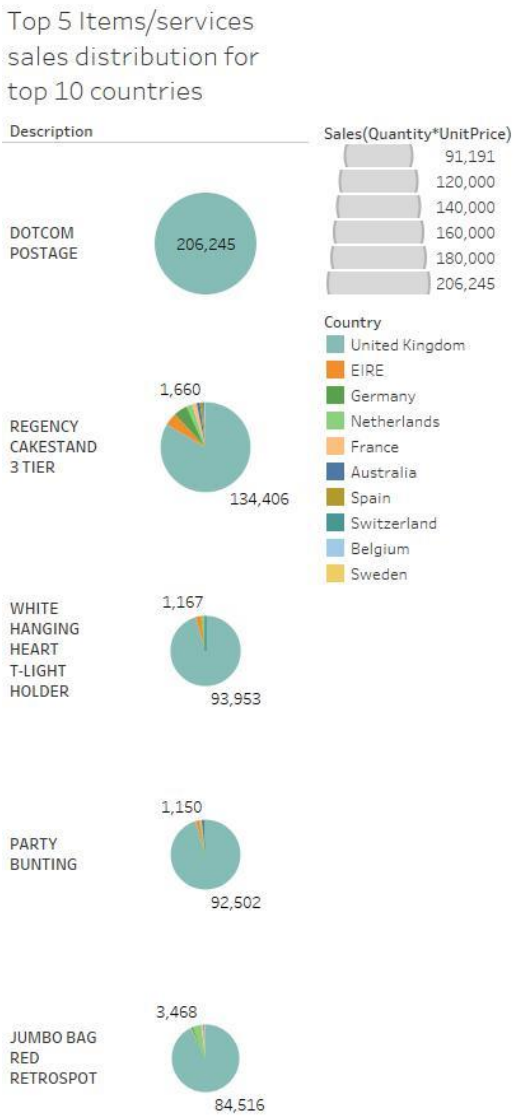
Sum of Sales(Quantity*UnitPrice) for each Country. The data is filtered on Action (Country,MONTH(Invoice Date)) and Action (Country,Description). The Action (Country,MONTH(Invoice Date)) filter keeps 314 members. The Action (Country,Description) filter keeps 20,143 members. The view is filtered on Country, which has multiple members selected.

For Worksheet 1: Top 10 Countries based on Sales

Here, we have taken Country filtered it by top 10 & sorted to descending w.r.t. SUM(Sales) and taken SUM(Sales). This gives us the top 10 countries with the most sales over several months.

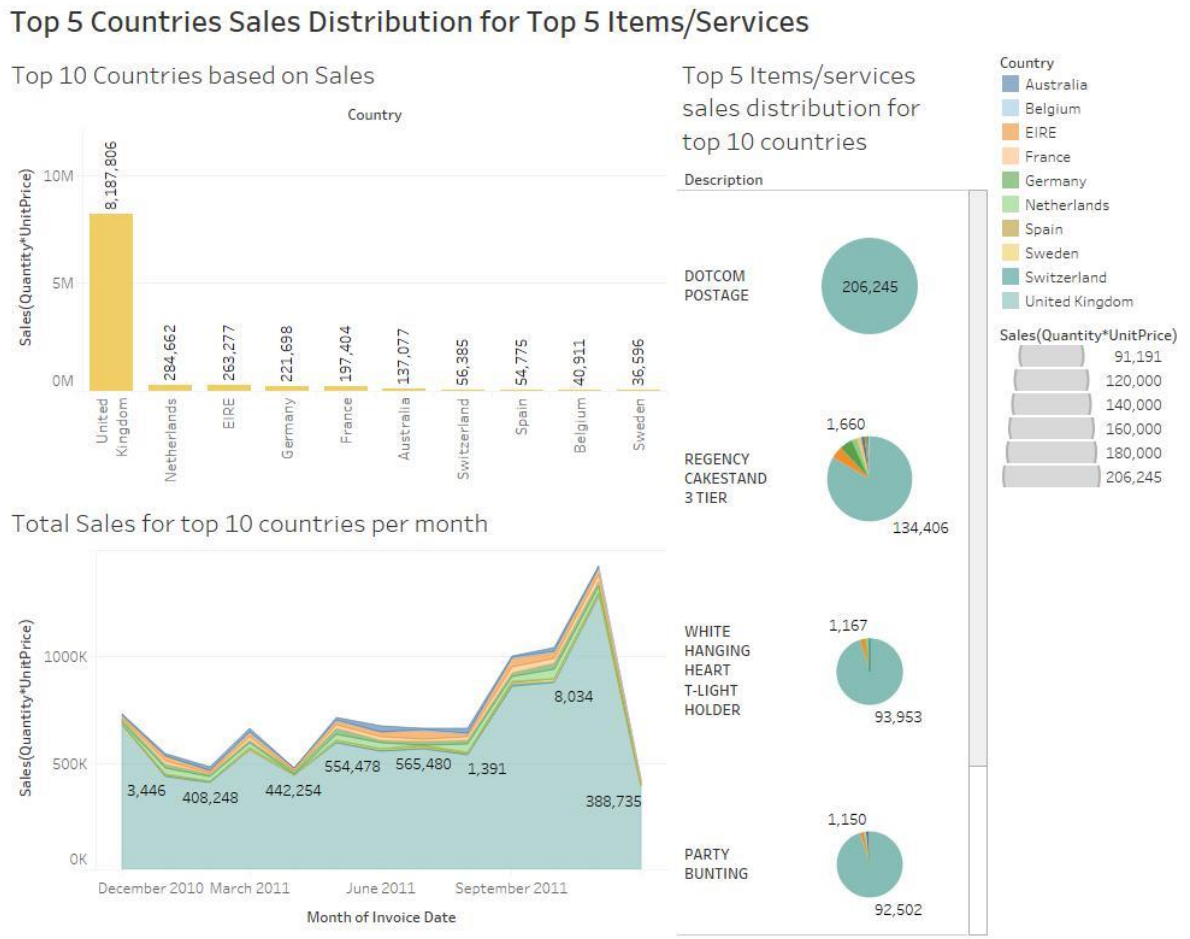
For Worksheet 2: Total Sales for top 10 countries per month

This next chart gives us the total sales for the top 10 countries and checks it for every single month. We have MONTH(InvoiceDate) in columns & SUM(Sales) in rows where we apply a filter of Description as top 10 for SUM(Sales). The graph is coloured according to country.



For Worksheet 3: Top 5 Items/services sales distribution for top 10 countries

Next, we further granularize our search by taking Description & filtering top 5 items, sorting to descending by SUM(Sales). Along with Country which is filtered by top 10 according to SUM(Sales). This gives us total sales pie charts for each of those top 5 items where the sector colours represent the countries contribution to these sales.



The resulting dashboard thus, gives us information on the top 5 items offered by the online retail which are bought in the top 10 buying countries.

Dashboard 10: Sales Distribution for Bottom 10 Countries w.r.t. top 10 items/services

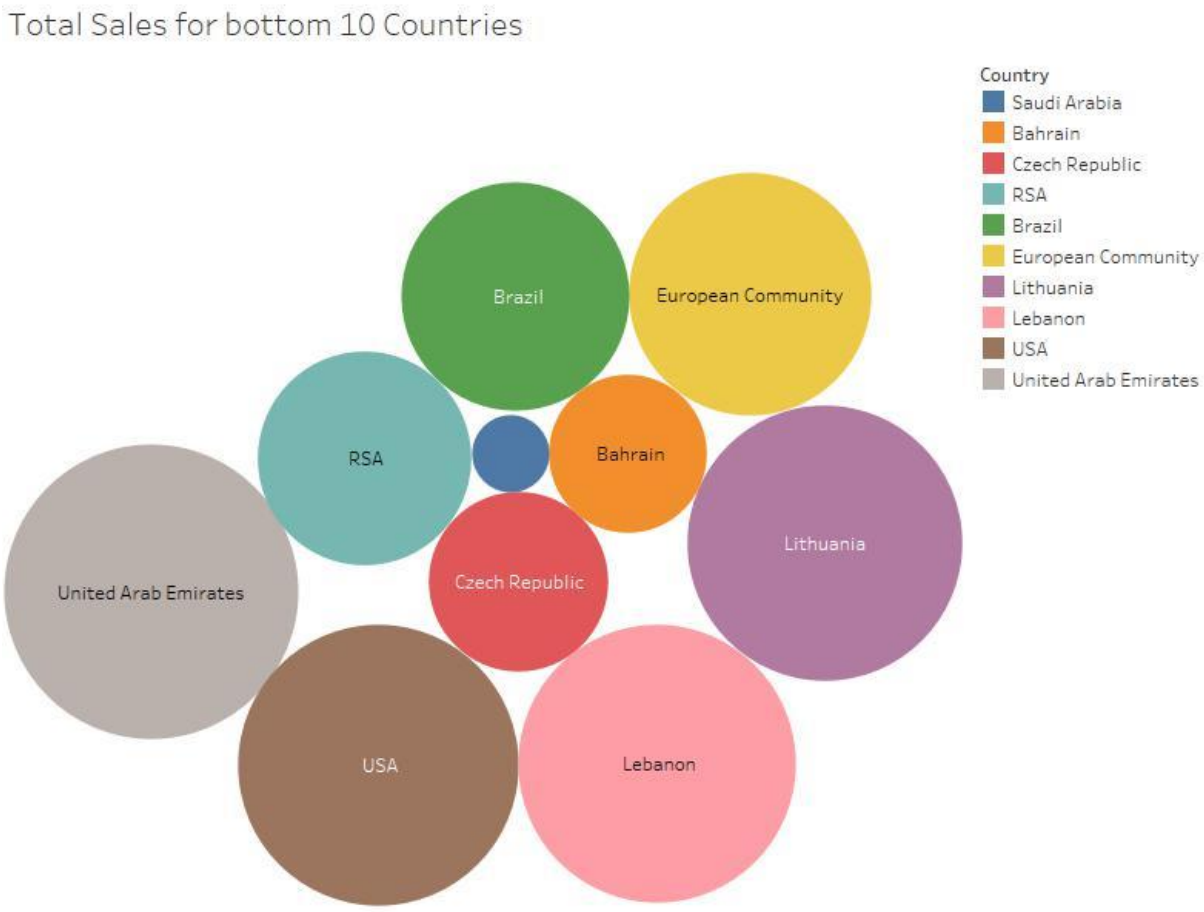
This dashboard is going to help us find out about the sales distribution of bottom 10 countries over the months & check their contribution to the sales of the top 10 most bought items.

Sales for Bottom 10 Countries per month

Sum of Sales(Quantity*UnitPrice) for each Invoice Date Month broken down by Country, Invoice Date Year and Invoice Date Quarter. Color shows details about Invoice Date Year. The data is filtered on Action (Country,Description) and Action (Country). The Action (Country,Description) filter keeps 20,143 members. The Action (Country) filter keeps 38 members. The view is filtered on Country, which keeps 10 of 38 members.

For Worksheet 1: Sales for Bottom 10 Countries per month

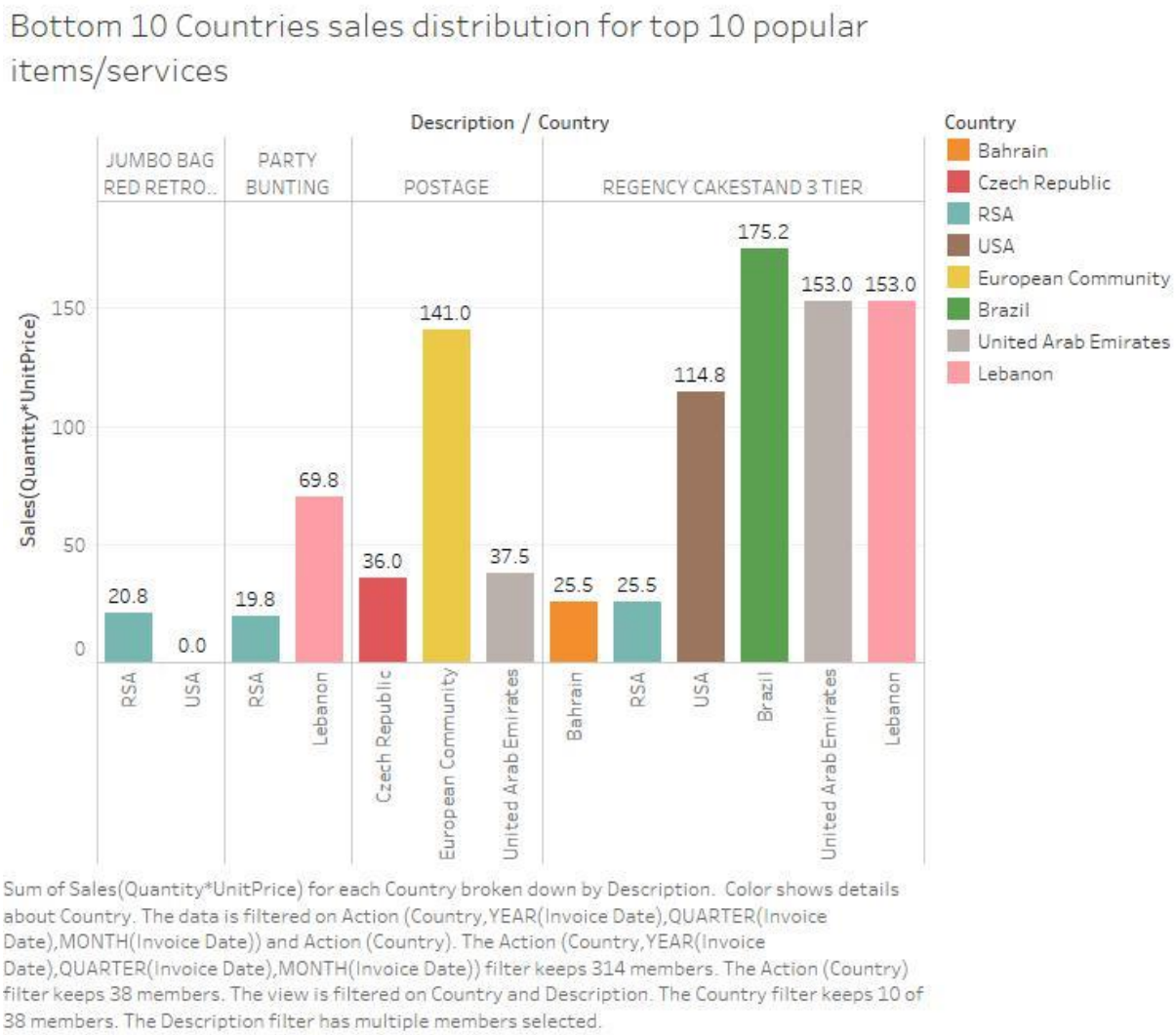
Here, we have country which is filtered as bottom 10 & sorted ascending by SUM(Sales) along with YEAR(InvoiceDate), QUARTER(InvoiceDate), MONTH(InvoiceDate) in the columns. In rows, we have our SUM(Sales). The graph gives us the revenue or sales for bottom 10 countries over all the months.



Country. Color shows details about Country. Size shows sum of Sales(Quantity*UnitPrice). The marks are labeled by Country. The data is filtered on Action (Country,YEAR(Invoice Date),QUARTER(Invoice Date),MONTH(Invoice Date)) and Action (Country,Description). The Action (Country,YEAR(Invoice Date),QUARTER(Invoice Date),MONTH(Invoice Date)) filter keeps 314 members. The Action (Country,Description) filter keeps 20,143 members. The view is filtered on Country, which keeps 10 of 38 members.

For Worksheet 2: Total Sales for bottom 10 Countries

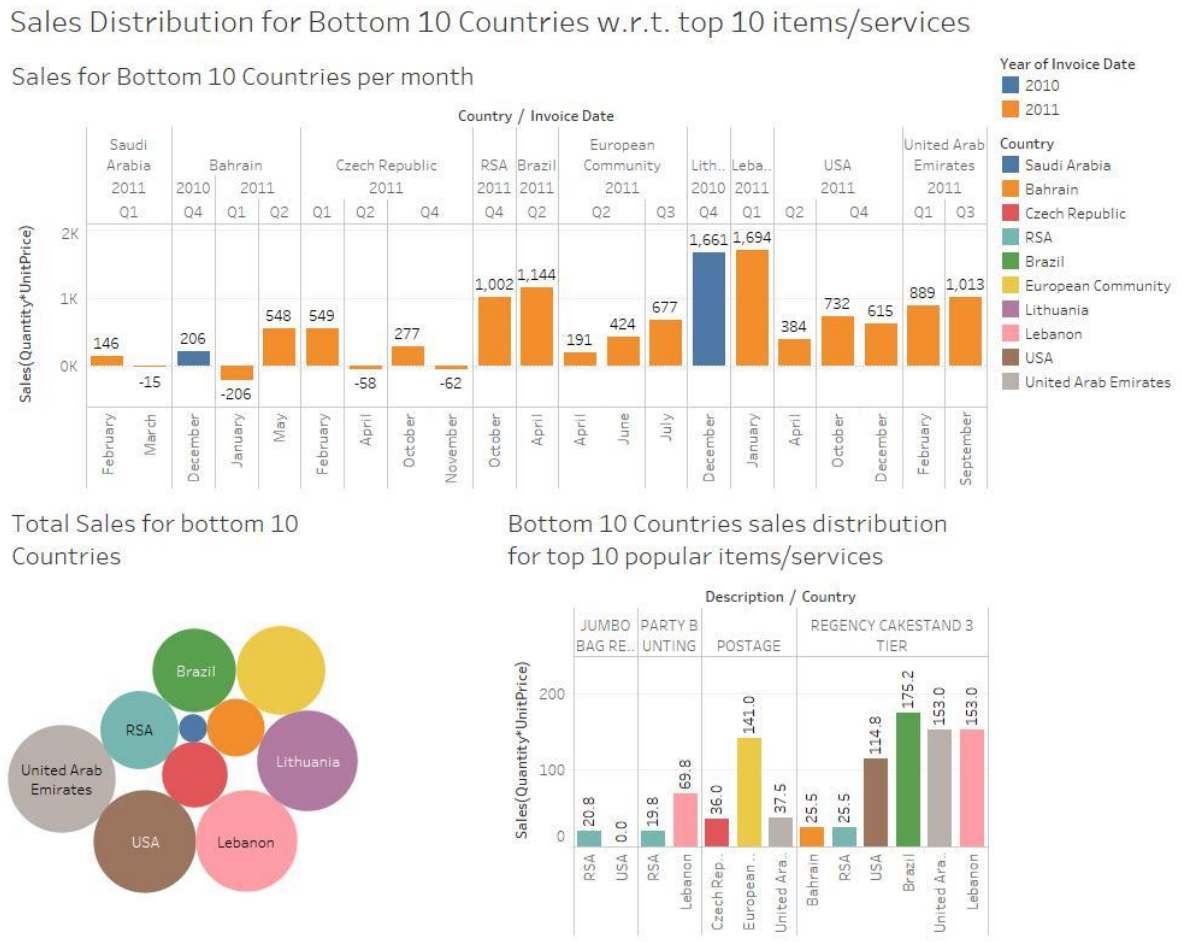
In this circle cluster graph, we have all the bottom 10 countries with the least sales where country is filtered as bottom 10 & sorted in ascending w.r.t SUM(Sales). Colour are for the different countries.



Sum of Sales(Quantity*UnitPrice) for each Country broken down by Description. Color shows details about Country. The data is filtered on Action (Country,YEAR(Invoice Date),QUARTER(Invoice Date),MONTH(Invoice Date)) and Action (Country). The Action (Country,YEAR(Invoice Date),QUARTER(Invoice Date),MONTH(Invoice Date)) filter keeps 314 members. The Action (Country) filter keeps 38 members. The view is filtered on Country and Description. The Country filter keeps 10 of 38 members. The Description filter has multiple members selected.

For Worksheet 3: Bottom 10 Countries sales distribution for top 10 popular items/services

Here, we have taken Description & Country as the columns where Description is filtered as top 10 items in terms of SUM(Sales) & Country is filtered as bottom 10 w.r.t. SUM(Sales) in ascending order. The filters applied give us only top 4 items bought by only bottom 8 countries as per sales which is an interesting intersection relation (A /\ B).



The resulting dashboard for sales distribution of top 10 items bought by bottom 10 countries gives us an intersection relation where only top 4 items are bought by only bottom 8 countries.

Dashboard 11: Top 10 Invoices by top Items/Services & top countries

In this dashboard, we shall focus our attention to InvoiceNo as the class. This is because even though Invoice number is unique id, there's a possibility that the same customer has made separate purchases resulting in different InvoiceNo's.

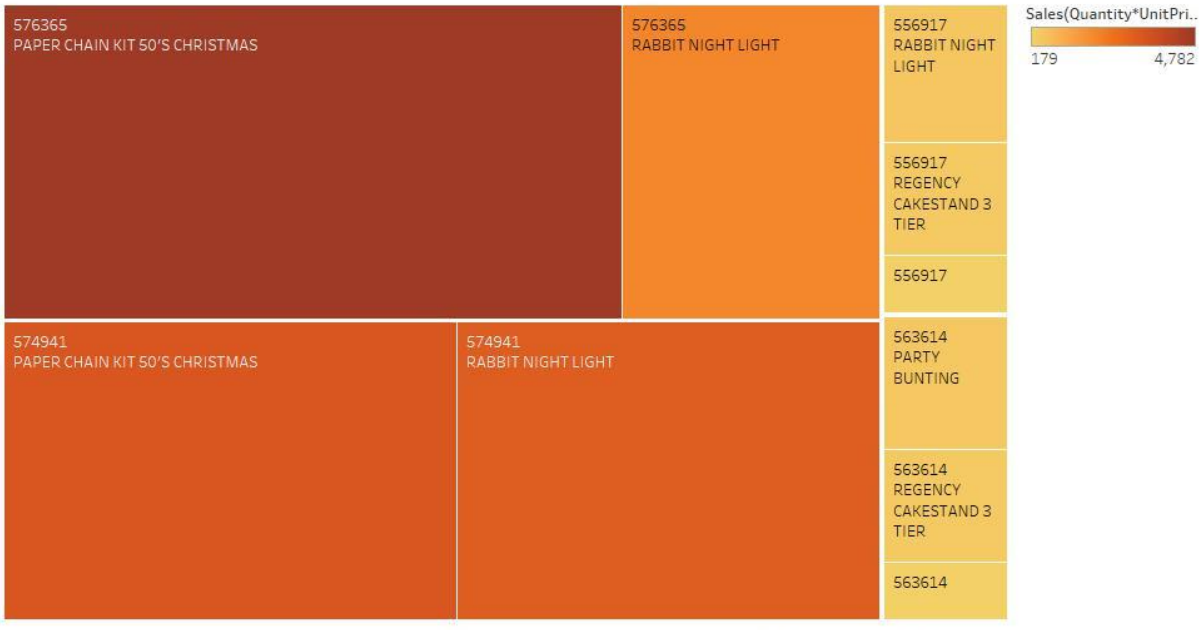
Top invoice filtered by sales

Sum of Sales(Quantity*UnitPrice) for each Invoice Date Month broken down by Invoice No and Invoice Date Year. Color shows details about Invoice Date Year. The data is filtered on Action (Country,Invoice No) and Action (Description,Invoice No). The Action (Country,Invoice No) filter keeps 25,900 members. The Action (Description,Invoice No) filter keeps 530,931 members. The view is filtered on Invoice No, which has multiple members selected.

For Worksheet 1: Top invoice filtered by sales

Here, we have taken Invoice No & Invoice month & year into columns where InvoiceNo is filtered by Top 10 invoices by Sales & sorted ascending by Sales along with SUM(Sales) in rows. The following graph gives us top 10 invoices only for 2011 as the invoice didn't exist for December 2010 & we don't have enough information if it existed before the time-period.

Top 10 Invoices by top 10 Items/Services filtered by Sales

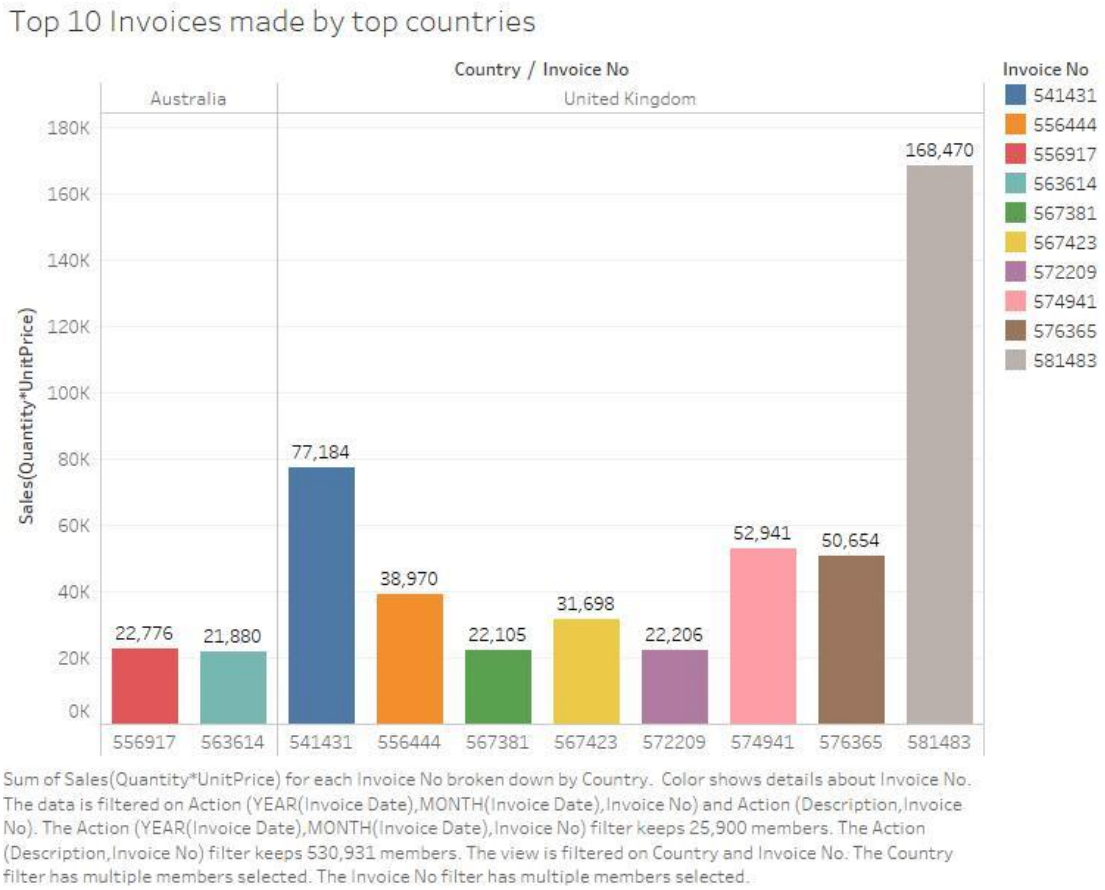


Invoice No and Description. Color shows sum of Sales(Quantity*UnitPrice). Size shows sum of Sales(Quantity*UnitPrice). The marks are labeled by Invoice No and Description. The data is filtered on Action (YEAR(Invoice Date),MONTH(Invoice Date),Invoice No) and Action (Country,Invoice No). The Action (YEAR(Invoice Date),MONTH(Invoice Date),Invoice No) filter keeps 25,900 members. The Action (Country,Invoice No) filter keeps 25,900 members. The view is filtered on Invoice No and Description. The Invoice No filter has multiple members selected. The Description filter has multiple members selected.

Pronoy D'sa 2952742

For Worksheet 2: Top 10 Invoices by top 10 Items/Services filtered by Sales

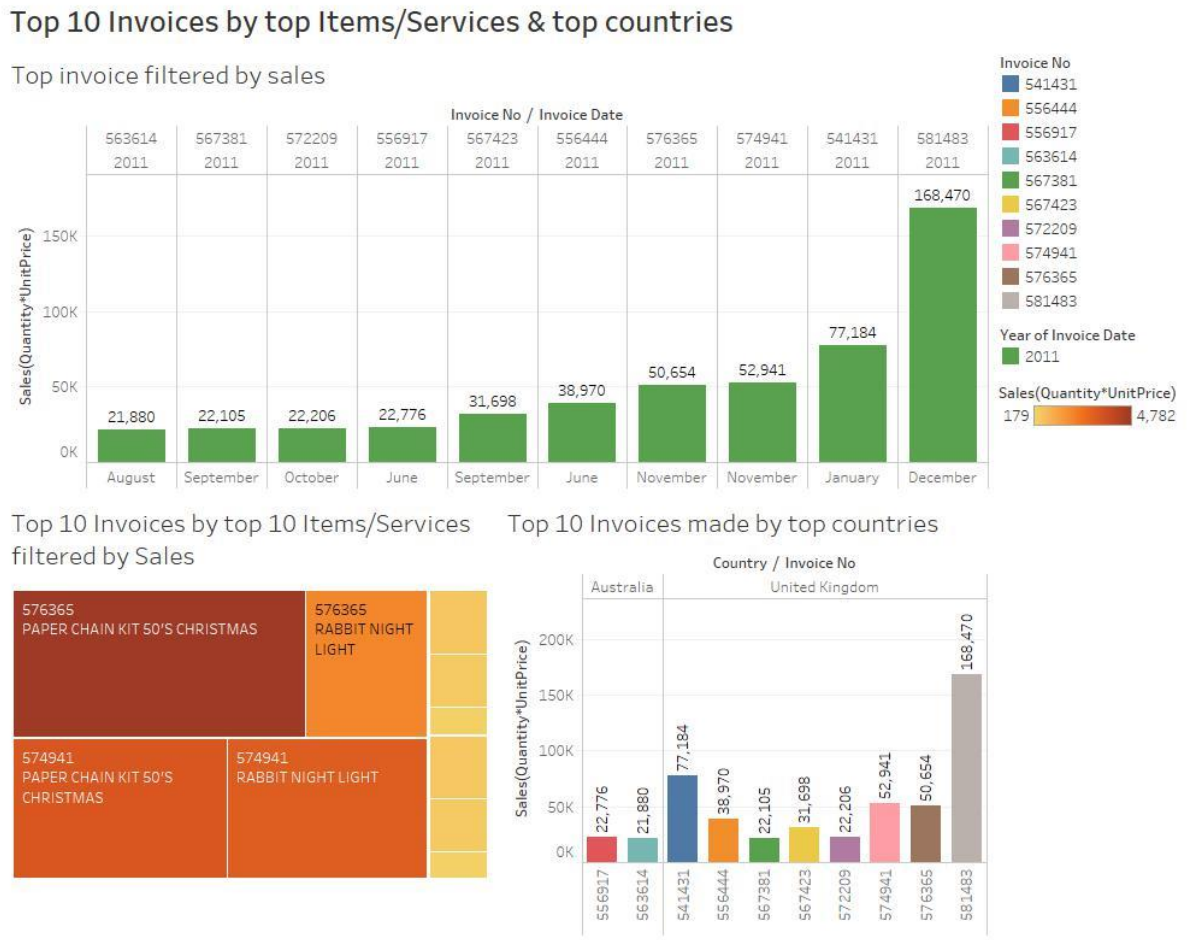
Here, we take item Description & InvoiceNo where Description is filtered as top 10 items by SUM(Sales) & InvoiceNo is filtered by top 10 invoices by SUM(Sales). The result graph gives us mainly 4 items bought by top 10 invoices.



Sum of Sales(Quantity*UnitPrice) for each Invoice No broken down by Country. Color shows details about Invoice No. The data is filtered on Action (YEAR(Invoice Date),MONTH(Invoice Date),Invoice No) and Action (Description,Invoice No). The Action (YEAR(Invoice Date),MONTH(Invoice Date),Invoice No) filter keeps 25,900 members. The Action (Description,Invoice No) filter keeps 530,931 members. The view is filtered on Country and Invoice No. The Country filter has multiple members selected. The Invoice No filter has multiple members selected.

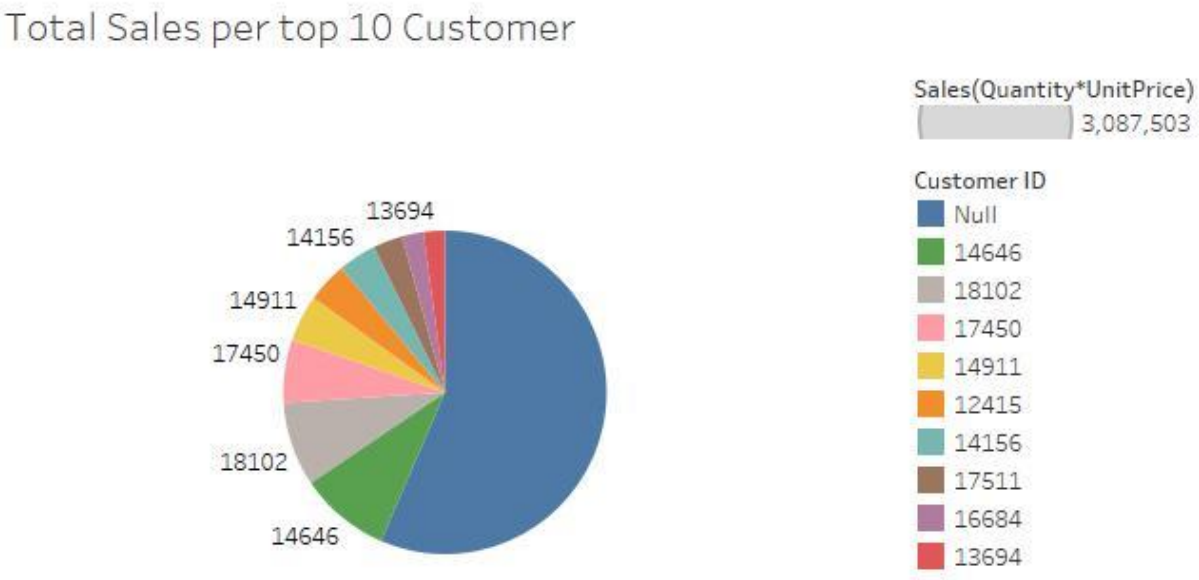
For Worksheet 3: Top 10 Invoices made by top countries

Here, we take Country and InvoiceNo under columns and filter both according to top 10 countries or invoices respectively w.r.t. SUM(Sales). Thus, the result graph gives us the two 2 top countries from where these top 10 invoices were generated.



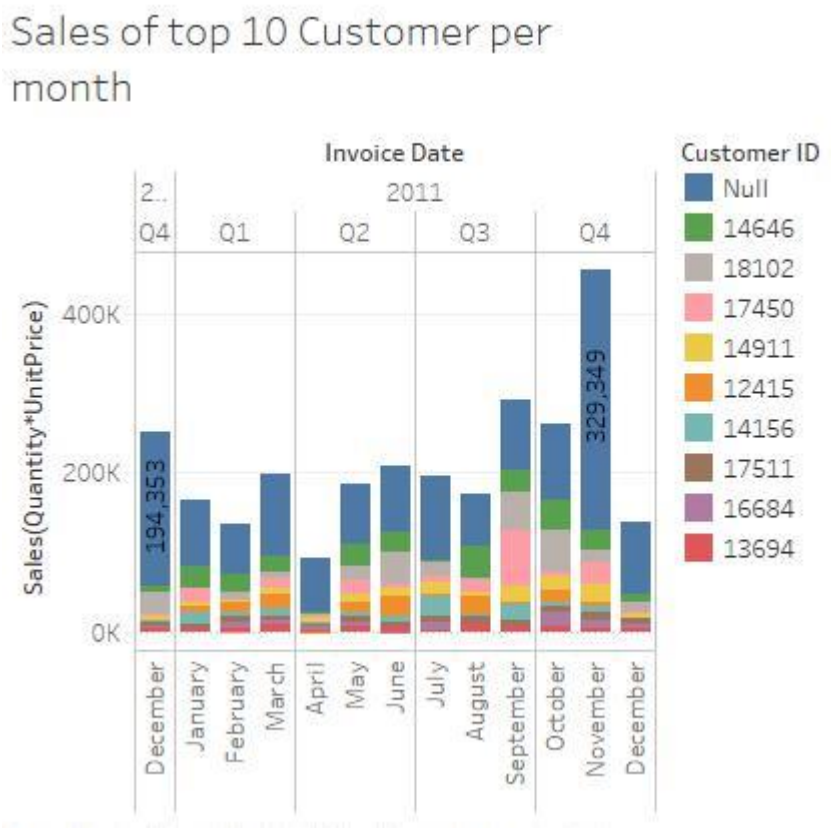
The resulting dashboard thus gives us information on the top 10 invoices according to the top 10 items/services purchased and in the top countries regarding sales.

Dashboard 12: Sales Distribution of top Customers



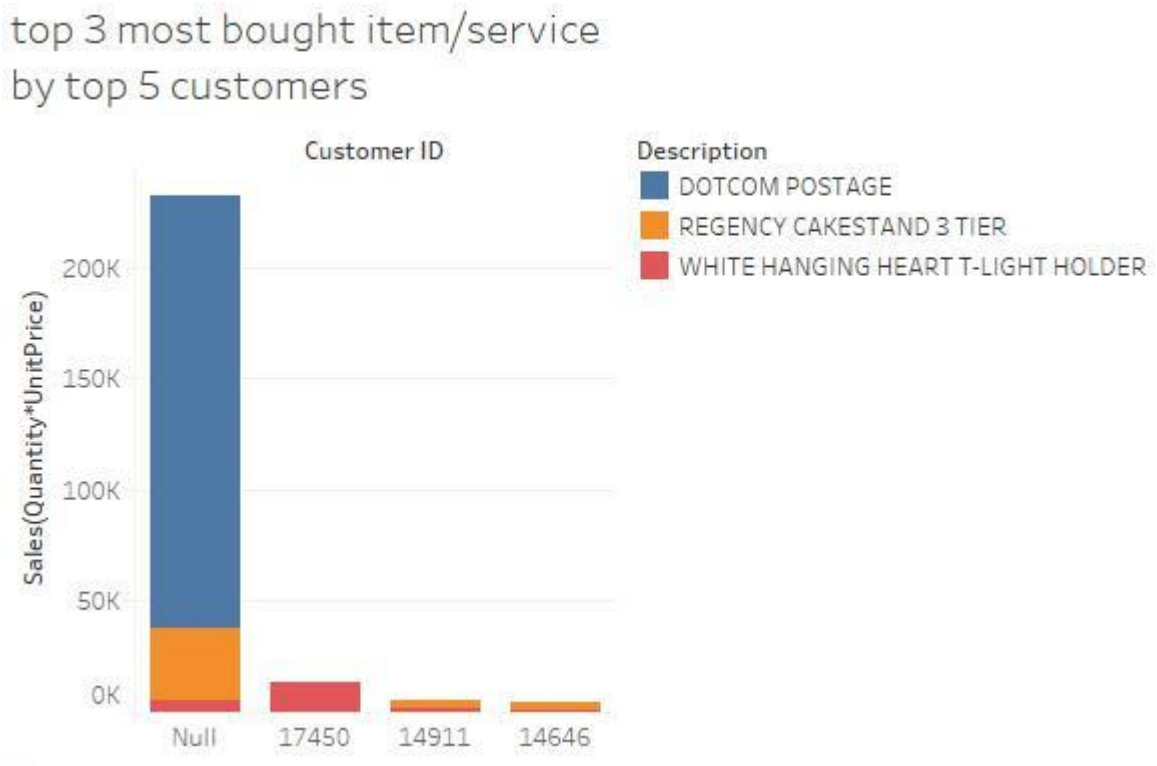
For Worksheet 1: Total Sales per top 10 Customer

In the pie chart, we are showing the total sales of the top 10 customers. This is done by taking CustomerID and filtering it by top 10 by SUM(Sales) sorted in descending order along with InvoiceNo which is filtered by top 10 invoices by SUM(Sales).



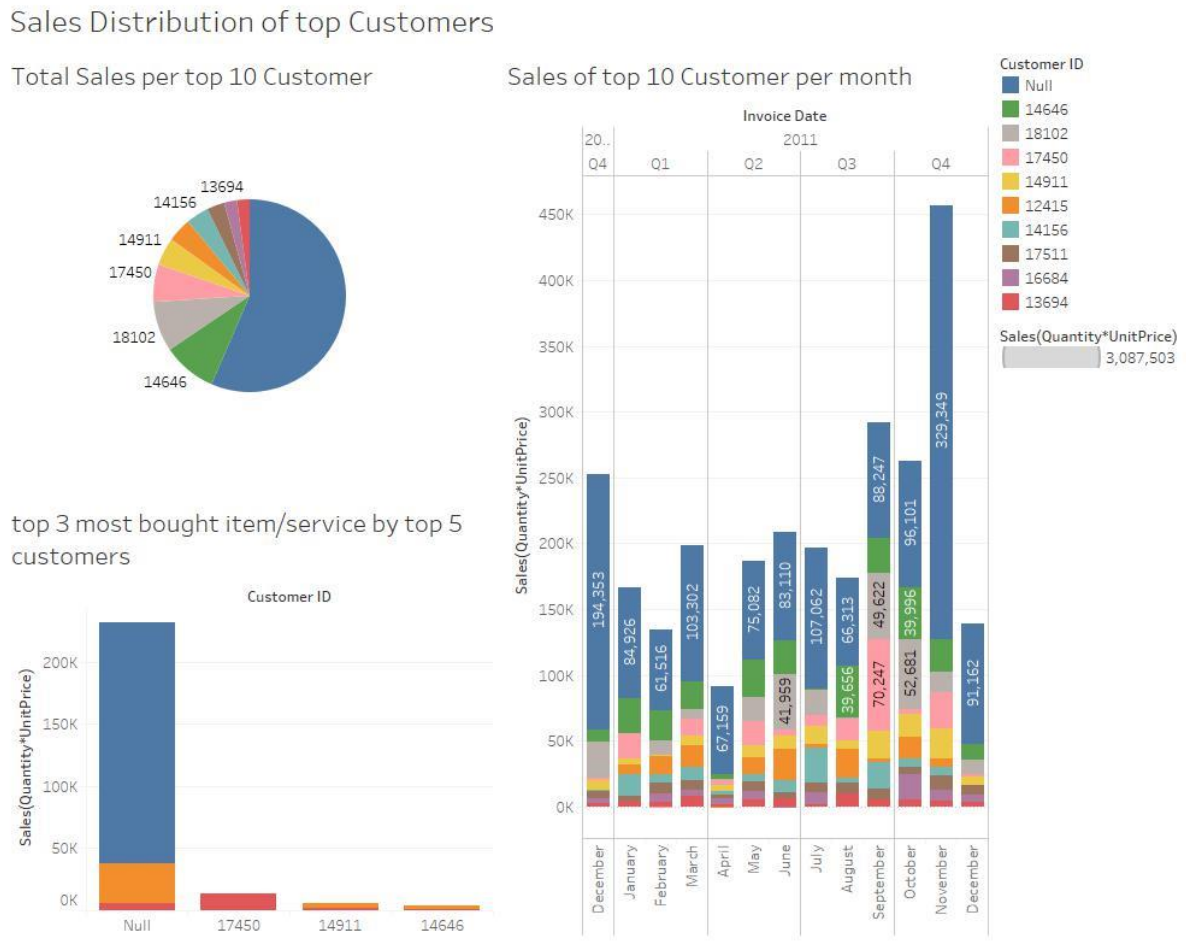
For Worksheet 2:

We have taken Invoice month under yearly quarters in columns and SUM(Sales) in rows. This results in a bar graph with total sales coloured by top 10 buyers or customers over several months.



For Worksheet 3:

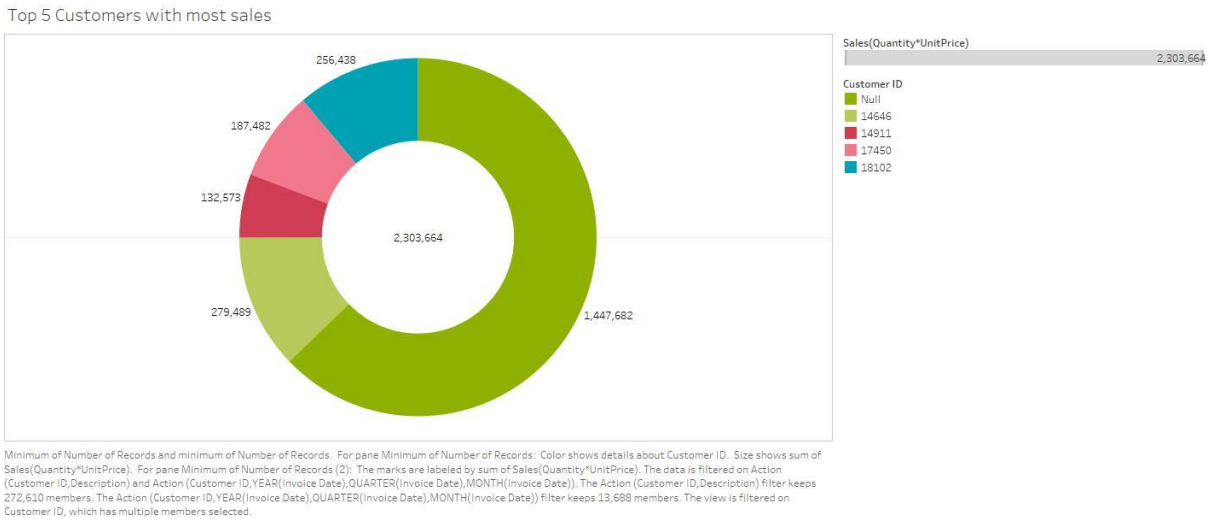
Here, we form a bar graph by taking a relation of the top 3 most bought items against the top 5 customers. This is done by taking CustomerID filtering it by top 5 customer as per SUM(Sales) against SUM(Sales) for item Description filtered by top 3 items by SUM(Sales). Colours depict the top 5 items.



For the resulting dashboard we get the following information regarding the top Customers by CustomerID as per sales. When we check the relationship of Top 5 customers against top 3 items bought, we see an intersection relation where only 4 out of 5 were possible in top customers, out of which maximum customer id's filled by null remain unknown.

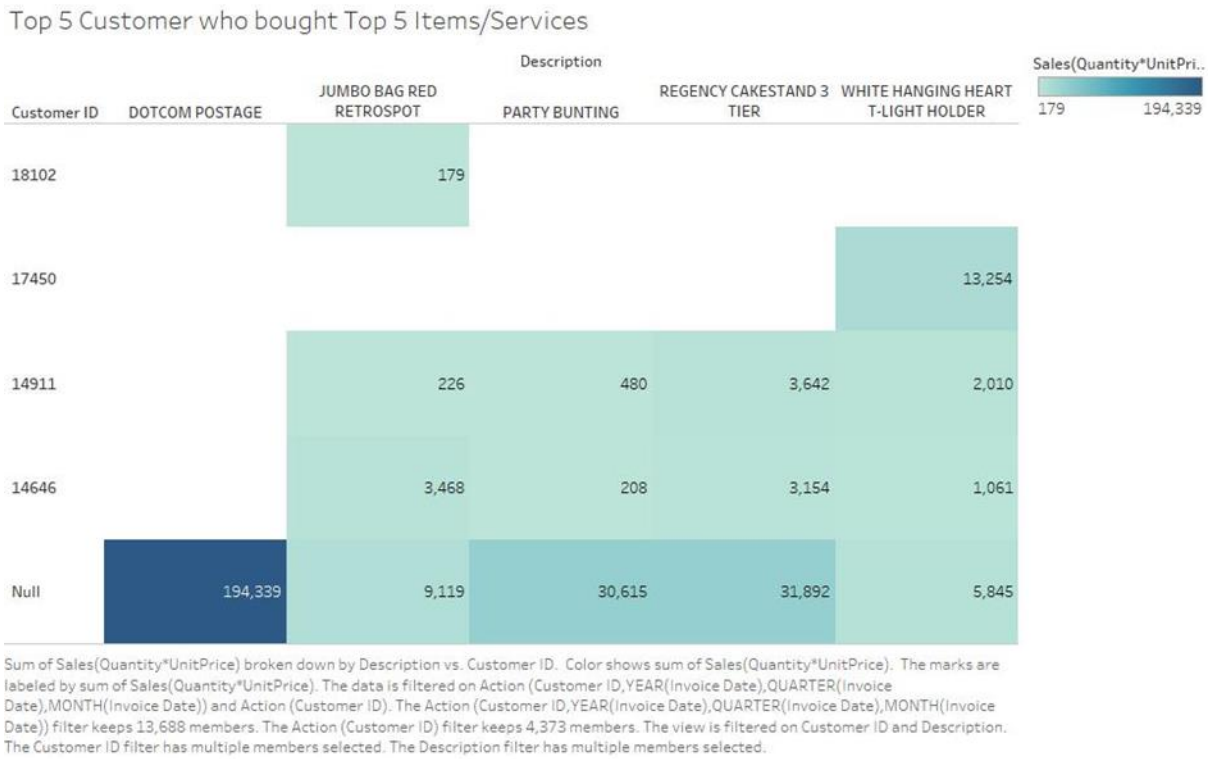
Dashboard 13: Top 5 Customers Sales & Quantities during time period

In this dashboard, we shall notice some details about the sales and quantities of items purchased by the top 5 customers.



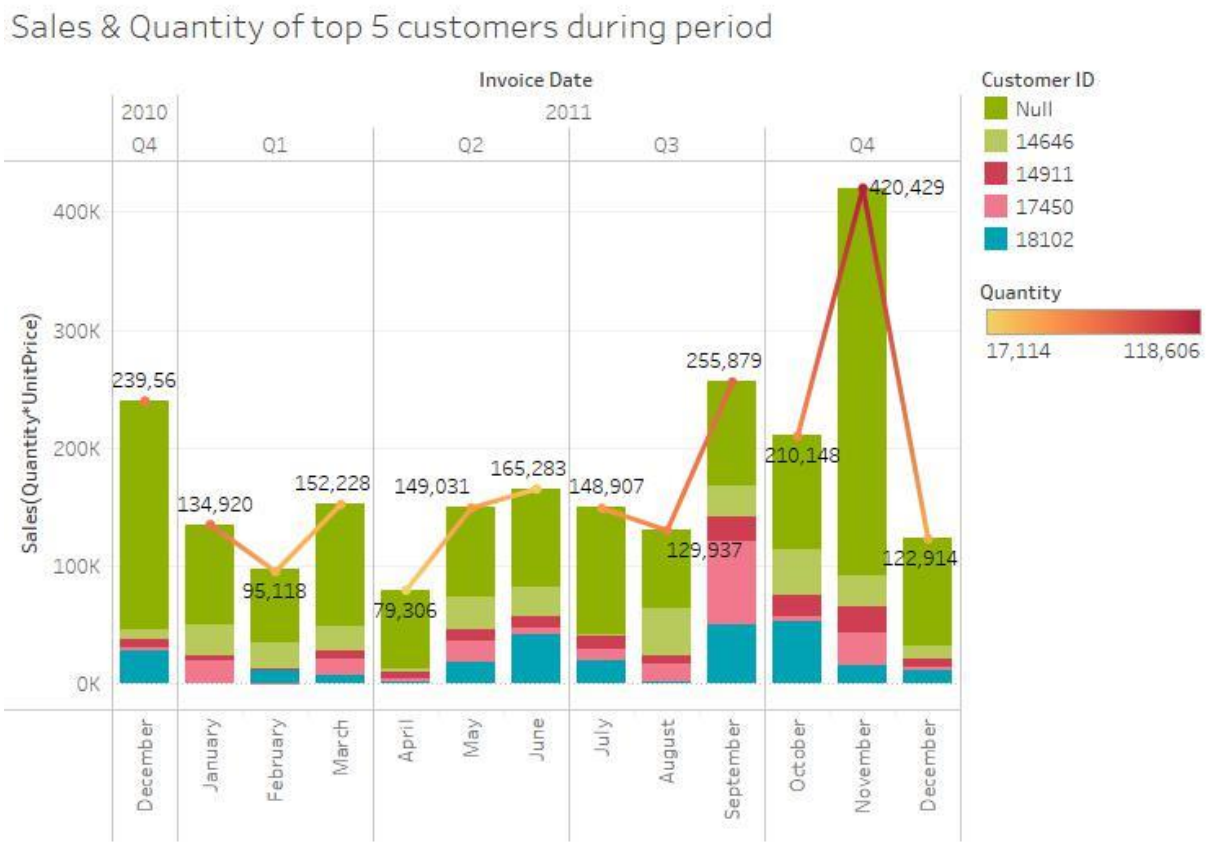
For Worksheet 1: Top 5 Customers with most sales

Here, we have created a donut chart which gives us information on the total number of sales made by the top 5 customers. This is done by taking CustomerID & SUM(Sales) into consideration where CustomerID is filtered by top 5 Customers as per SUM(Sales).



For Worksheet 2: Top 5 Customer who bought Top 5 Items/Services

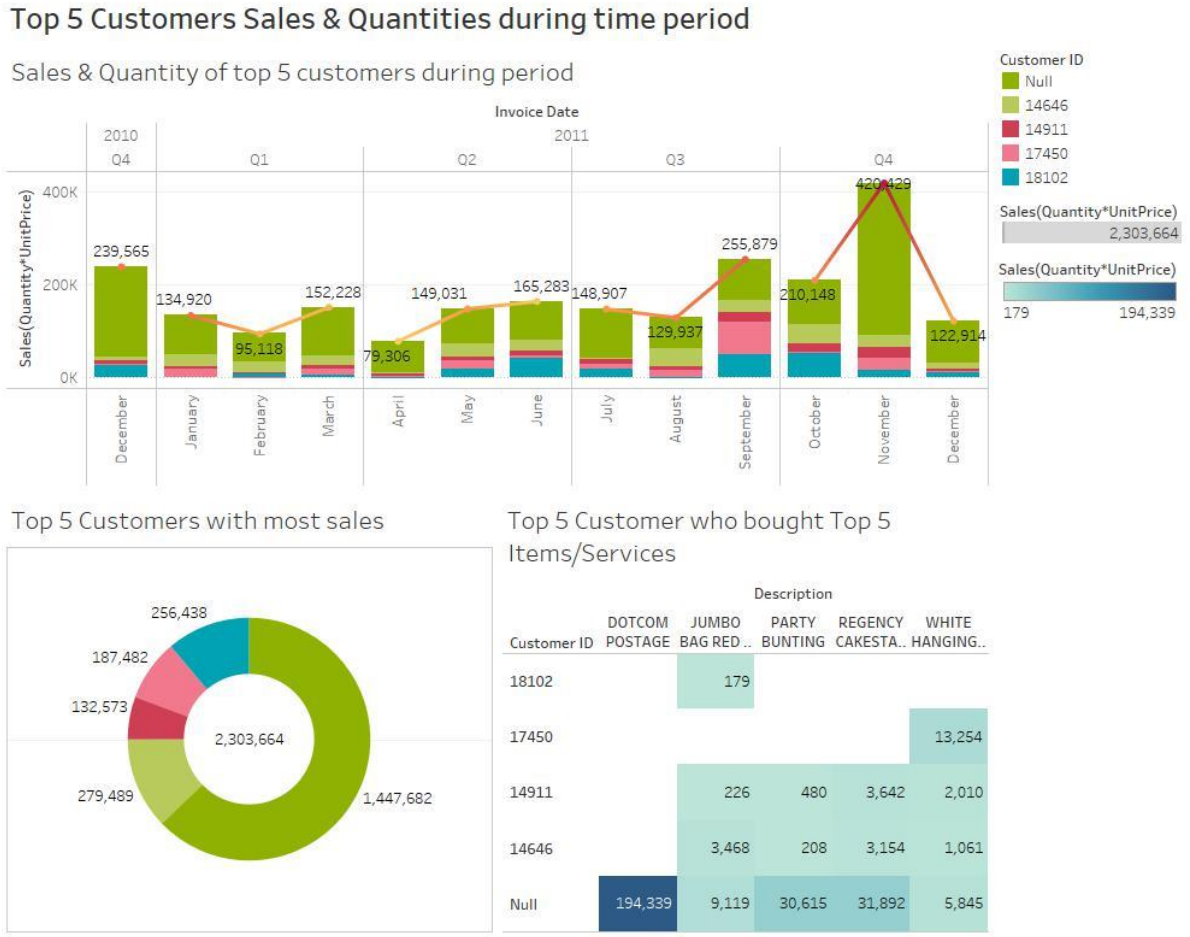
The following graph gives us a coloured matrix which is designed according to CustomerID & Description where CustomerID is filtered by top 5 customers & sorted in descending by SUM(Sales) and where Description is filtered as top 5 items by SUM(Sales). Colours in the graph represent the extent of SUM(Sales).



The trends of sum of Sales(Quantity*UnitPrice) and sum of Sales(Quantity*UnitPrice) for Invoice Date Month broken down by Invoice Date Year and Invoice Date Quarter. For pane Sum of Sales(Quantity*UnitPrice): Color shows details about Customer ID. For pane Sum of Sales(Quantity*UnitPrice) (2): Color shows sum of Quantity. The data is filtered on Action (Customer ID,Description) and Action (Customer ID). The Action (Customer ID,Description) filter keeps 272,610 members. The Action (Customer ID) filter keeps 4,373 members. The view is filtered on Customer ID, which has multiple members selected.

For Worksheet 3: Sales & Quantity of top 5 customers during period

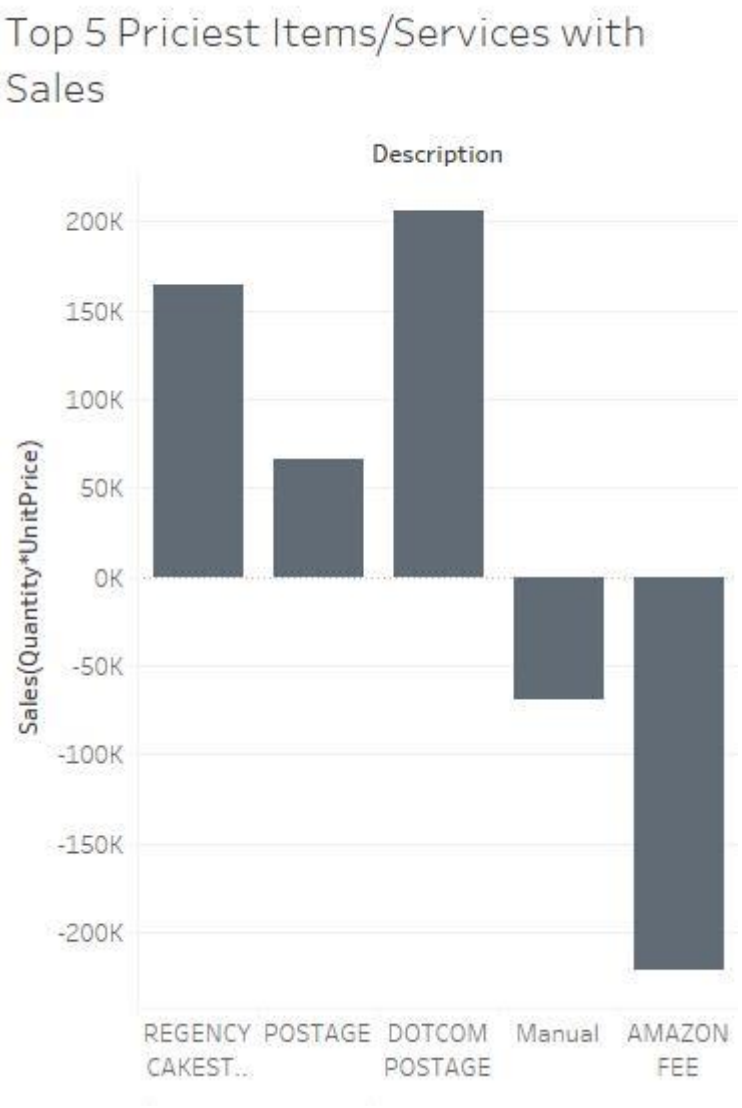
The following bar-line graph depicts sales of top 5 customers among total sales where the lines line chart is coloured according to the quantity field. This is done by taking Invoice month under yearly quarters against SUM(Sales). Also, we've applied a filter of top 5 customers as per SUM(Sales).



The resulting dashboard thus, gives us information on the top 5 customers using CustomerID with regards to their contribution to the total sales over the given time-period coupled with their contribution to sales of top 5 customers.

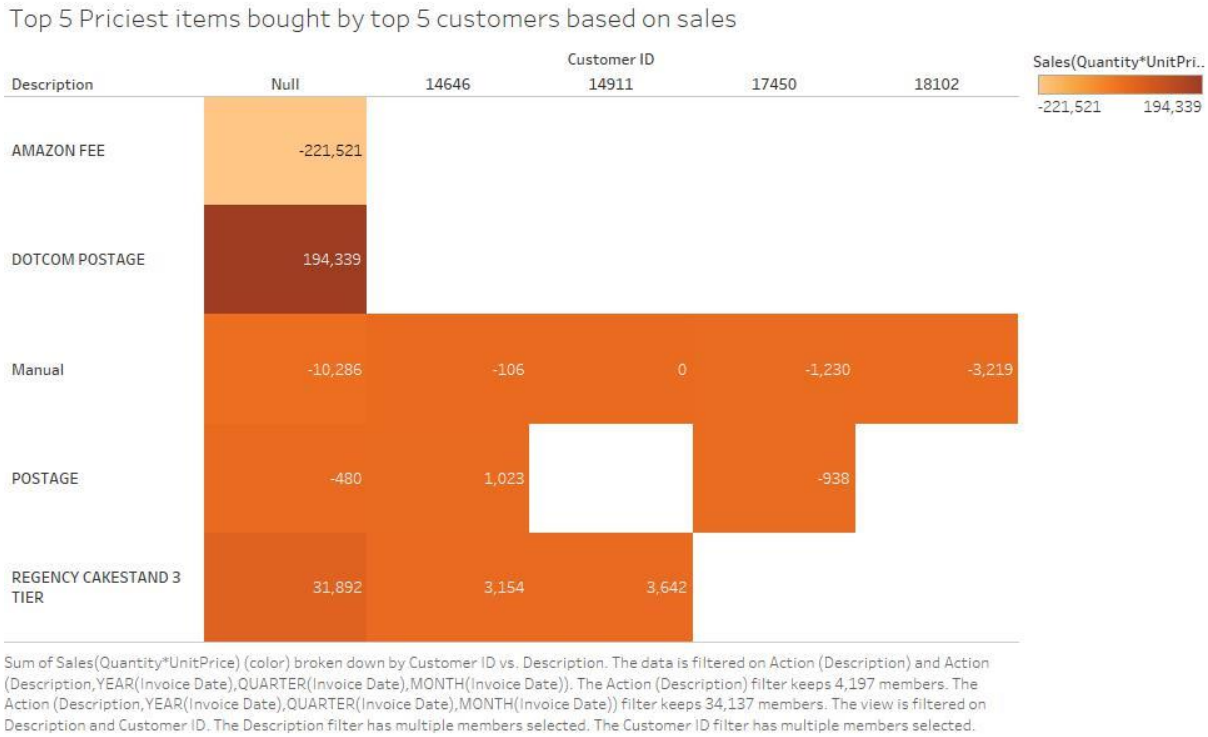
Dashboard 14: Top 5 priciest items bought by top customers w.r.t. sales & date

In this dashboard, we shall discover the relationship of unit price with attributes such as sales and invoice date.



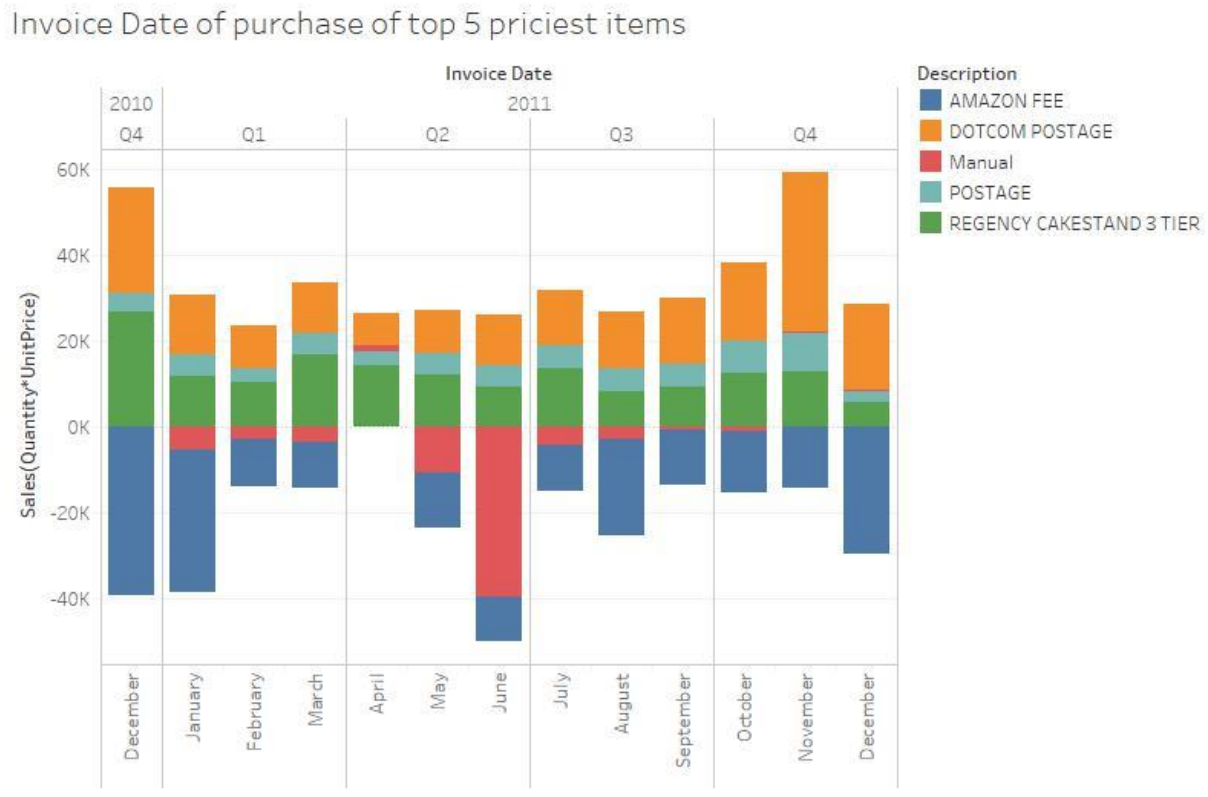
For Worksheet 1: Top 5 Priciest Items/Services with Sales

Here, we have taken item description and SUM(Sales) where Description is filtered as top 5 items by SUM(UnitPrice) which is sorted in ascending order by SUM(UnitPrice). Thus, we get a bar graph with total sales of the most priciest items.



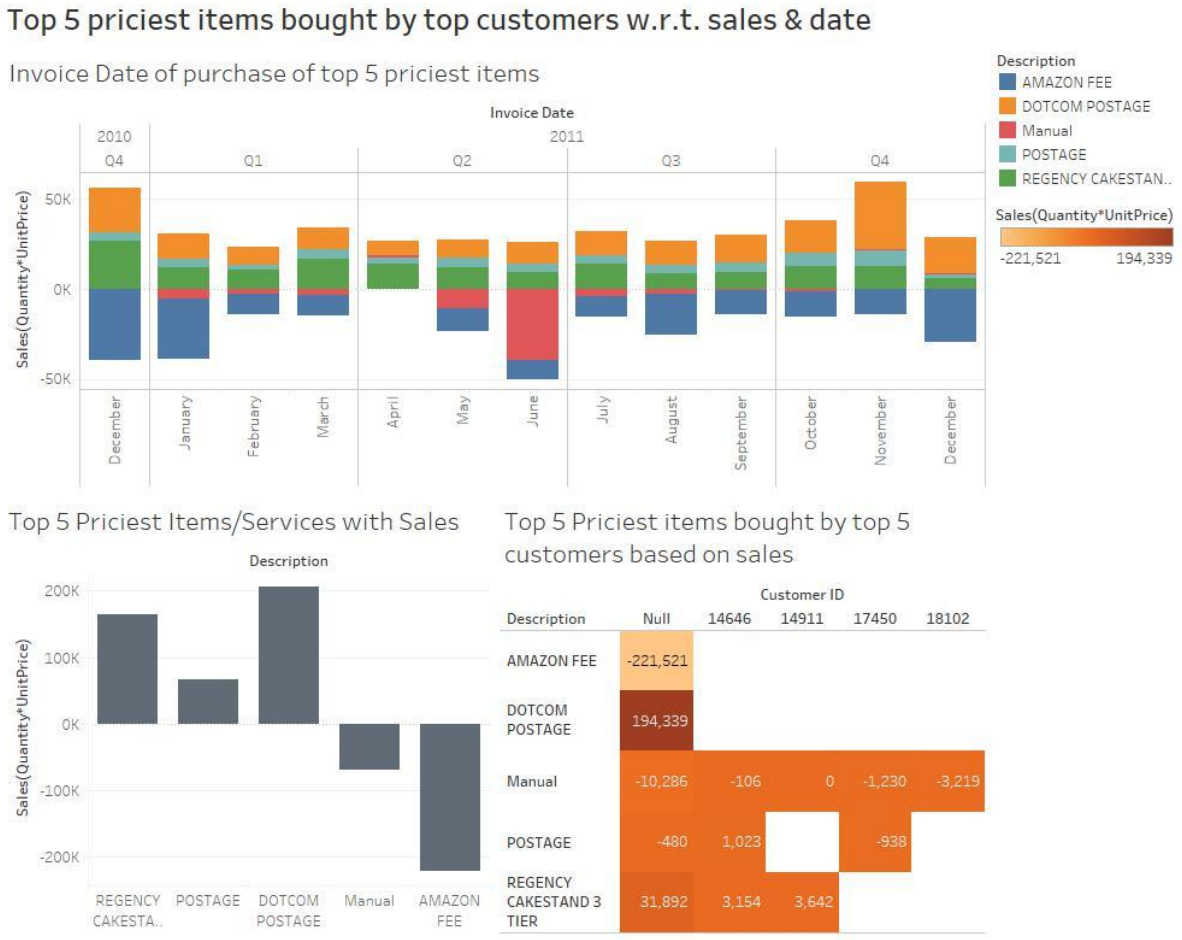
For Worksheet 2: Top 5 Priciest items bought by top 5 customers based on sales

The following colour matrix chart depicts the sales of top 5 items with the most unit price against the top 5 highest-buyers or top 5 customers. This is done by filtering description as top 5 items by SUM(UnitPrice) also, filtering CustomerID as top 5 customers by SUM(Sales). The colour is filled with the SUM(Sales)



For Worksheet 3: Invoice Date of purchase of top 5 priciest items

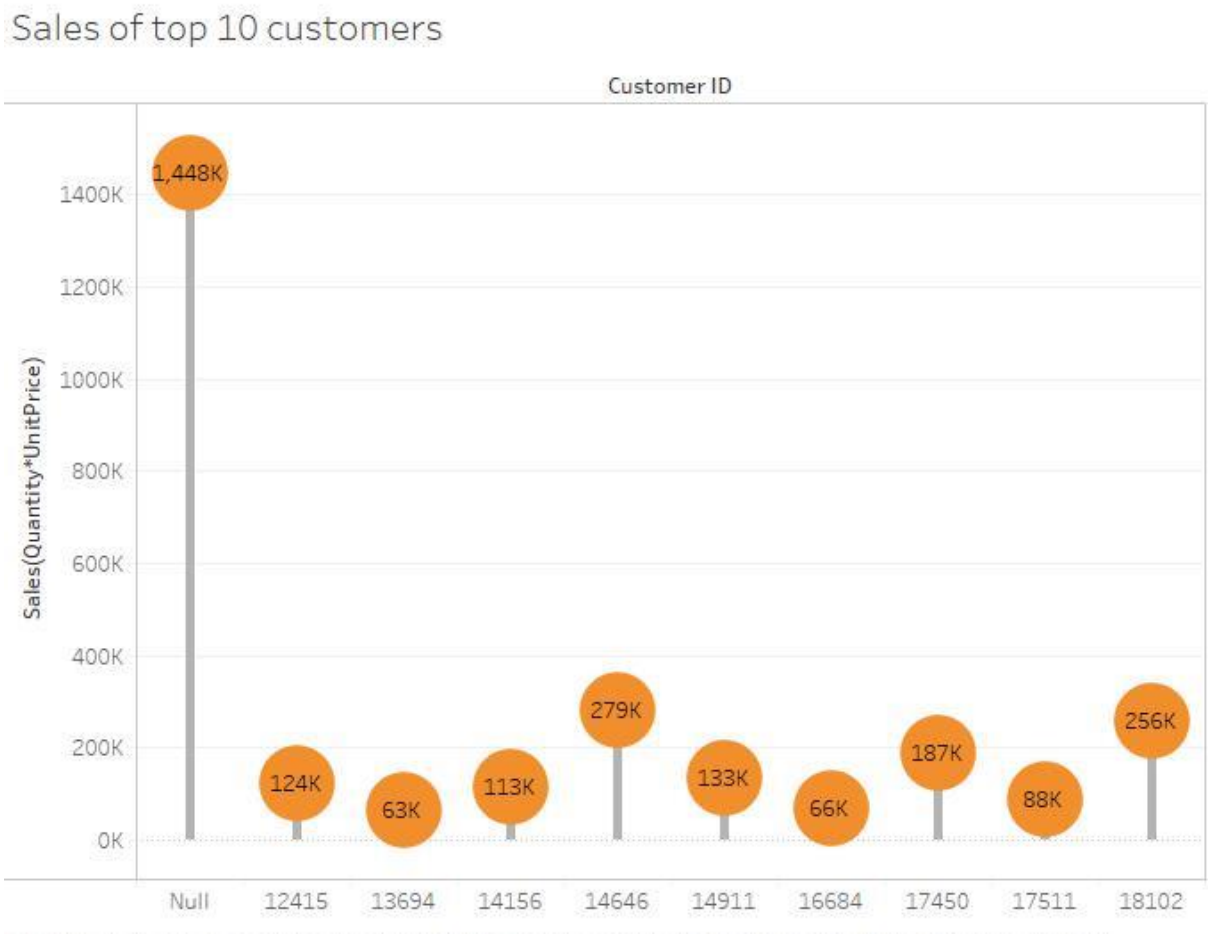
Here, we have taken Month under yearly quarters of InvoiceDate against SUM(Sales) where we have applied a filter of description of the top 5 items by SUM(UnitPrice). This gives us a bar graph of sales over all months which is viewed by contribution of top 5 priciest items in total sales.



The resulting dashboard thus gives us variable relationships like the contribution of UnitPrice with attributes such as total sales and InvoiceDate. Also, which top 5 customers are buying the top 5 priciest items.

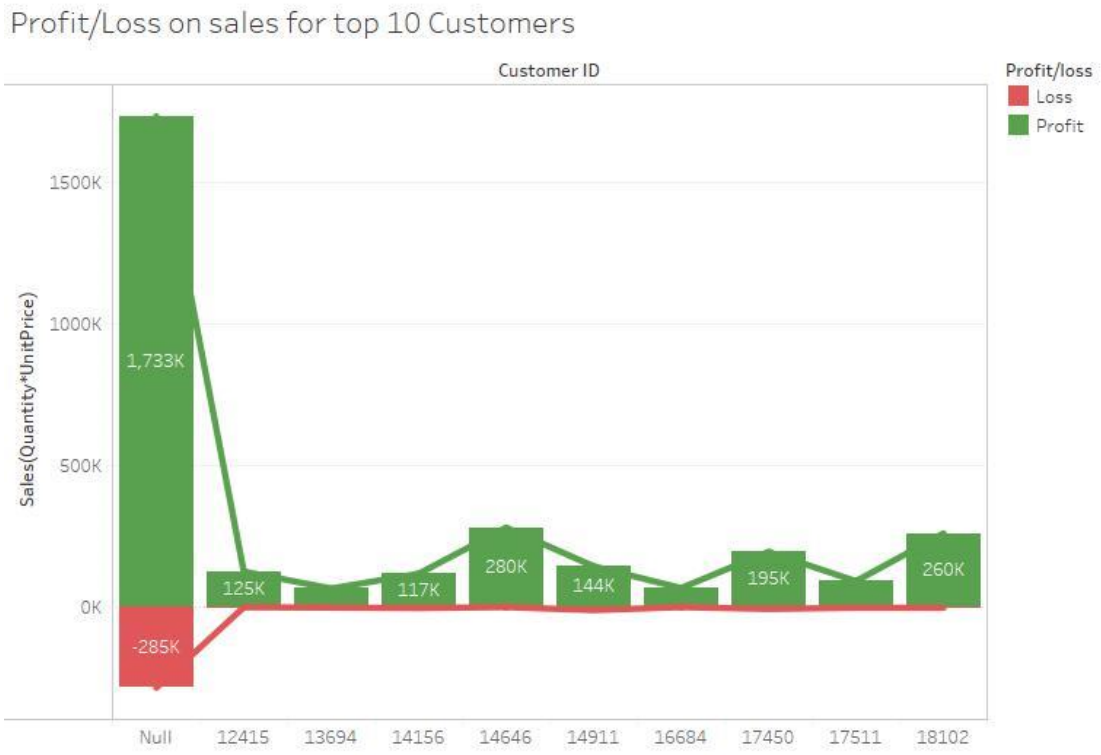
Dashboard 15: Loss/Profit on sales of top 10 customers

In the following dashboard, we shall discuss about sales for top 10 customer and over all the months by considering the calculated field called Loss/Profit created in tableau which says that if quantity goes under zero and becomes negative, it means there's a loss in sales/revenue else if greater than zero there's a profit in sales.



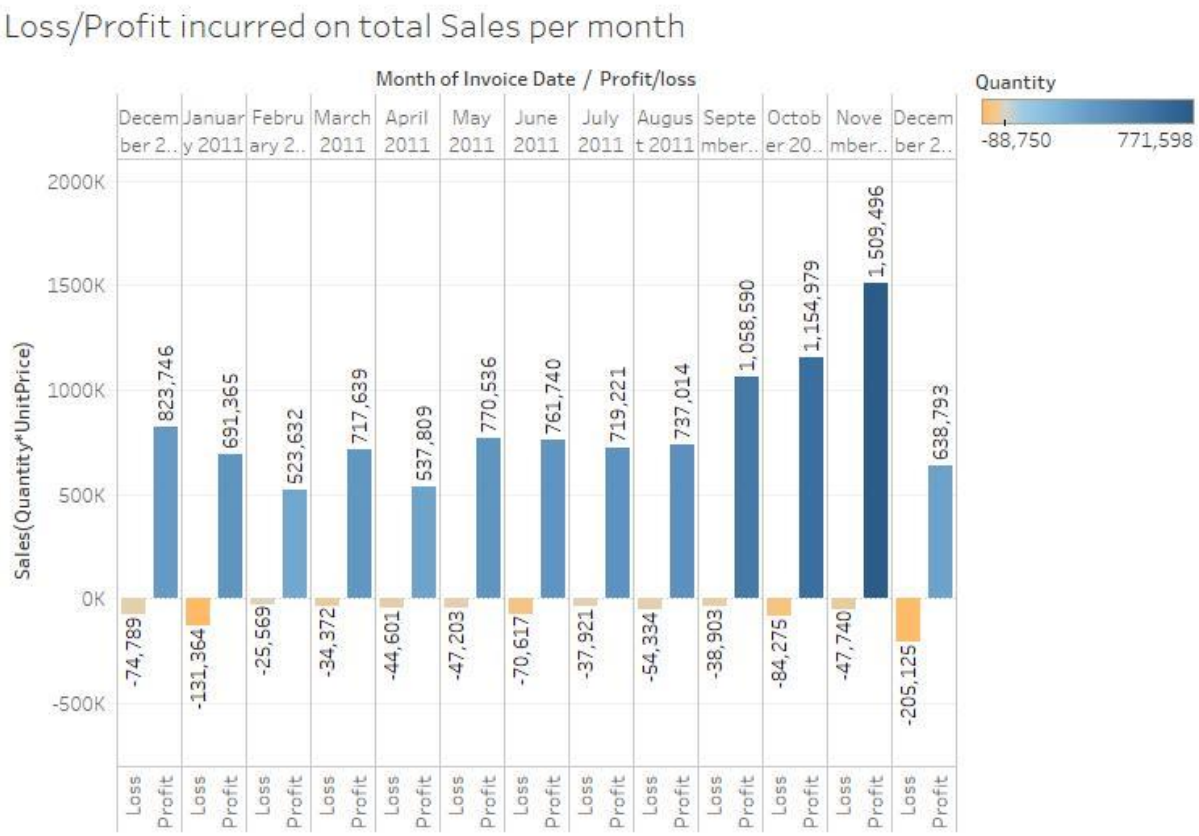
For Worksheet 1: Sales of top 10 customers

Here, we have created a lollipop chart based on the sales of the top 10 customers according to sales. This is done by taking CustomerID which is filtered as top 10 customers by SUM(Sales) coupled against SUM(Sales) using dual axes.



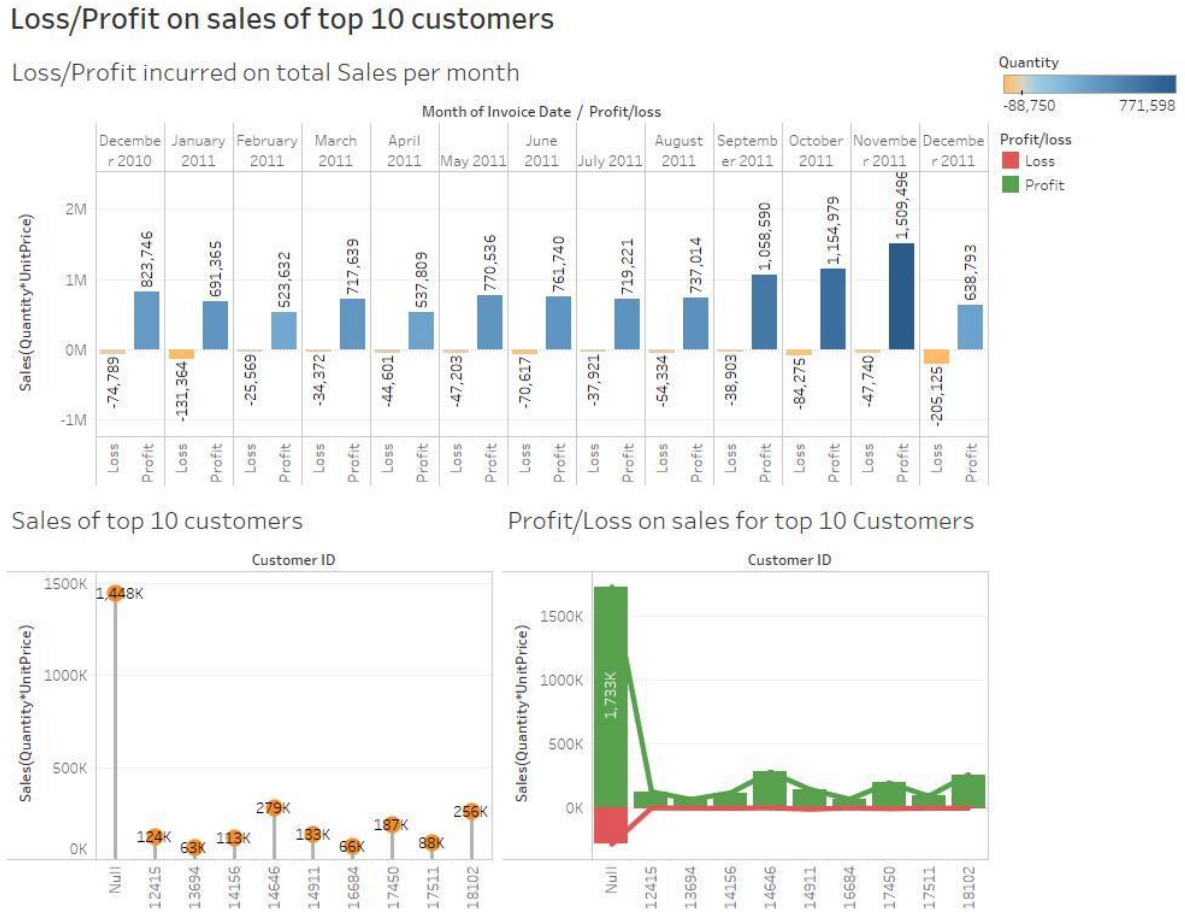
For Worksheet 2: Profit/Loss on sales for top 10 Customers

In this bar-line chart we see two types of sales bar graphs, one when quantity is less than 0, thus causing loss and another when quantity is greater than 0, thus causing profit. We do this, by keeping CustomerID which is filtered as top 10 by SUM(Sales) against SUM(Sales). Red colour signifies loss and green is for profit.



For Worksheet 3: Loss/Profit incurred on total Sales per month

We have taken month from InvoiceDate along with Profit/Loss in columns and SUM(Sales) is taken in rows. The following bar graph provides us the details of profit or loss on sales distributed over all the months in the given time frame.

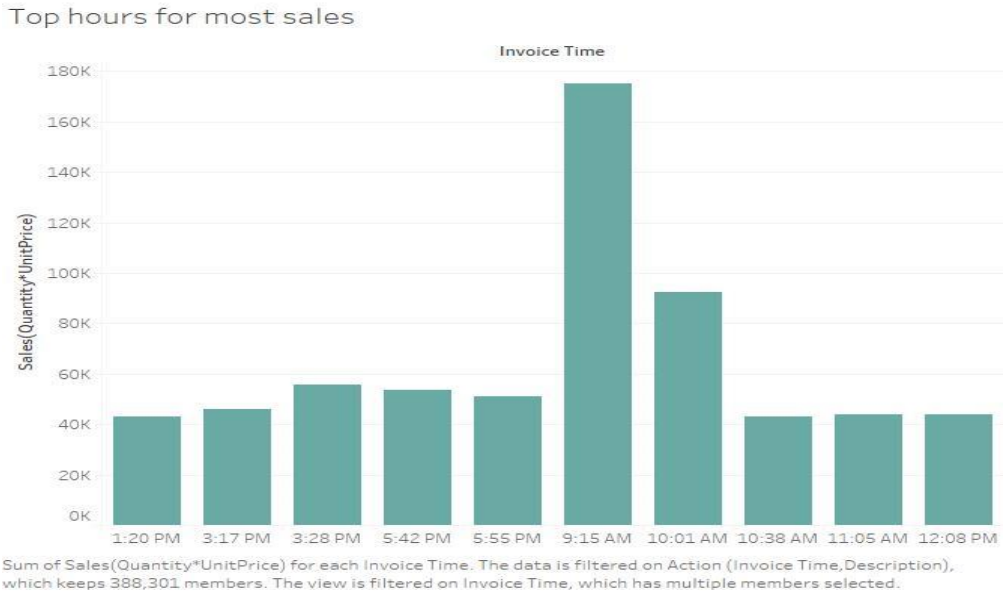


The resulting dashboard provides us the details on how profit or loss could affect the sales distribution over all the months in the time frame by considering the top 10 customers sales profits/loss against corresponding total sales profits/loss.

*Additional Dashboard

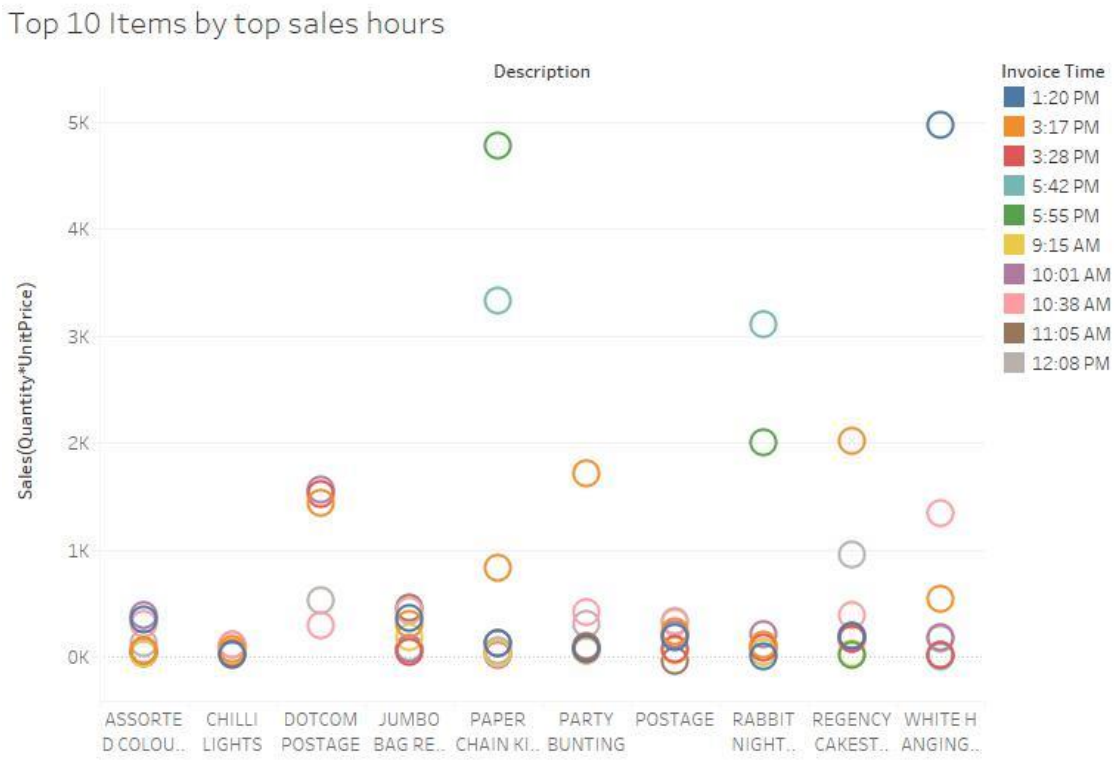
Dashboard 16: Top Hours of sales Information

The following dashboard provides insights into the top hours of purchase of items and services



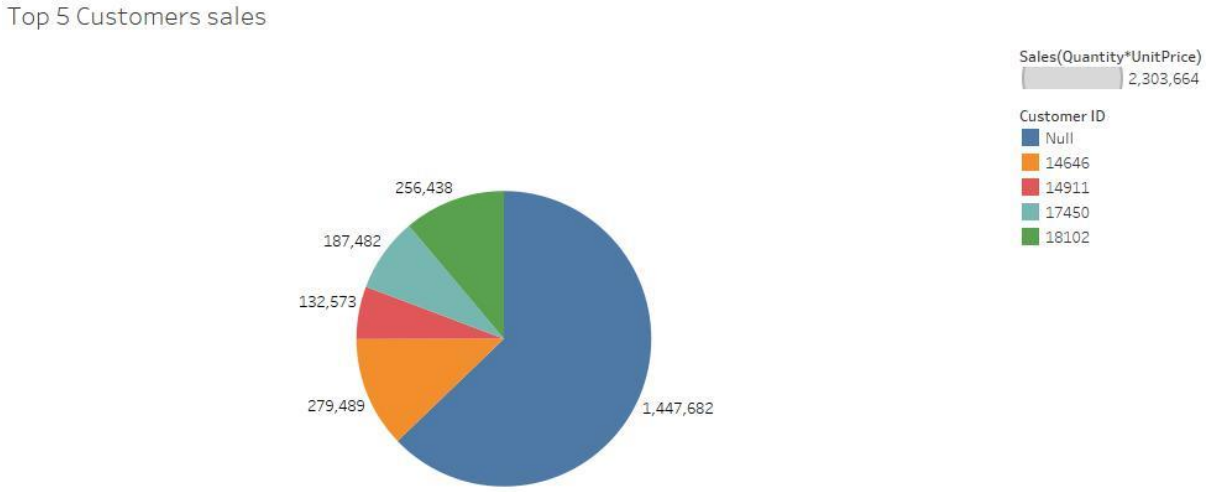
For Worksheet 1: Top hours for most sales

Here we have taken InvoiceTime and SUM(Sales) as the variables whose relations we are interested in discovering where InvoiceTime is filtered as top 10 timings by SUM(Sales). The bar graph reveals the top hours-minutes when the sales were high.



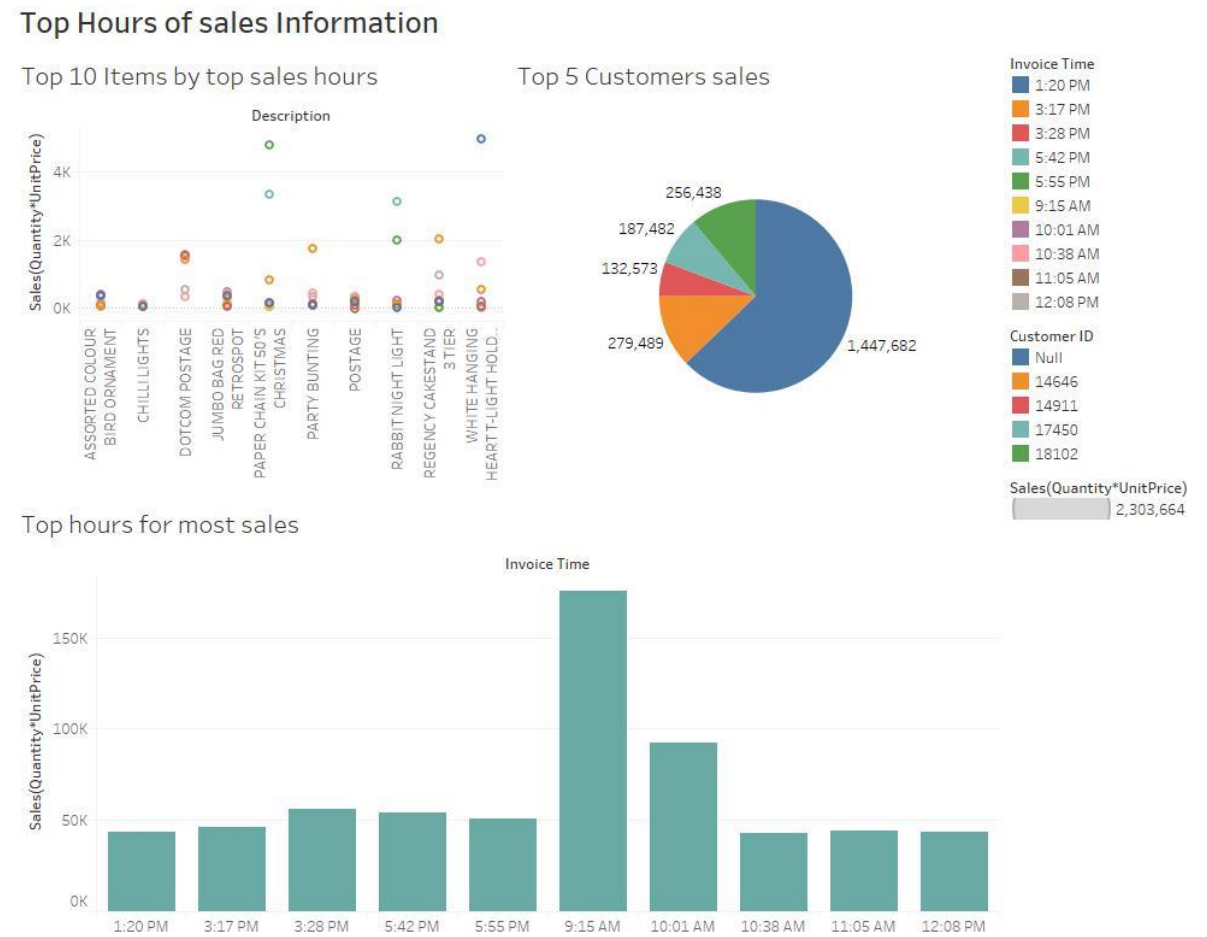
For Worksheet 2: Top 10 Items by top sales hours

The following ring plotted graph gives us the top 10 items filtered by top sales hours/timings. We take Description which is filtered as top 10 items by SUM(Sales) against SUM(Sales). We also apply filter of InvoiceTime as top 10 hours by SUM(Sales). The colours are for the different timings.



For Worksheet 3: Top 5 Customers sales

Here, we have a pie chart describing the total sales of the top 5 customers. We take CustomerID which is filtered as top 5 by SUM(Sales) against SUM(Sales). Colours are for the top 5 customers.



The final dashboard gives us details on the top hours for sales to spike up on the charts. It informs us visually that these are the times that top items & others will be bought by top customers.

Pronoy D'sa 2952742

Highlights/Insights:

1) We got the top 20 items/services based on quantity with their respective purchaser's countries.

2) We got the bottom 20 items/services that were damaged/deducted due to certain reasons & found it was mostly in the UK.

3) We found the top 10 most sold items in the top 10 countries where sales are highest.

4) We found the bottom 10 items/services where loss of revenue in sales is occurring with their respective countries.

5) Next, we found the top items/services based on sales with their contribution to total sales for both years in the time frame.

6) We found the decrease in Quantity, Sales/Revenue for each of the Bottom 20 items

7) We compared sales between top 5 most bought items and bottom 5, loss incurring items based on sales for each yearly quarter

8) We found the Sales/Revenue made during all the months in the period with an added observation on which sales are within UK & which are not

9) We found the top 5 items most bought items in the top 10 highest buying countries.

10) We found the sales distribution of top 10 items bought by bottom 10 countries which gives us an intersection relation where only top 4 items are bought by only bottom 8 countries

11) Here we get only 4 of the top 10 most bought items are bought by the top 10 invoices. Also we found that 2 of the top 10 buying countries are among the top 10 invoices.

12) We found that 4 of the top 5 customers in terms of sales buy the top 3 most bought items.

13) We found the top 5 customers regarding their contribution to the total sales over the given time frame.

14) We found the top 5 customers who buy the top 5 most priced items.

15) We found the top 10 customers sales profits/loss against corresponding total sales profits/loss.

16) We found the top hours/times for high sales of items/services which will be bought by top customers.

Actionable Intelligence:

1) Using the relations, we can find the ideal current customers & provide certain offers/schemes to boost sales. Also, we can nurture relations with new customers too.

2) We could introduce lower price offers or discounts on items that are not bought during the top hours for sales.

3) Use market-basket analysis to pair items that are not frequently purchased with items that are most frequently purchased through schemes such as 'buy one get one free' or 'buy 2 for 50% off'.

4) Check the popularity of a product in a country by its sales to increase or decrease its supply, maximizing profits & minimizing losses.

Part II

Predictive Analytics and Visualisation of Insights (60% of the overall module)

Predictive Analytics helps connecting the data to effective actions by drawing reliable conclusions about current conditions & future events.

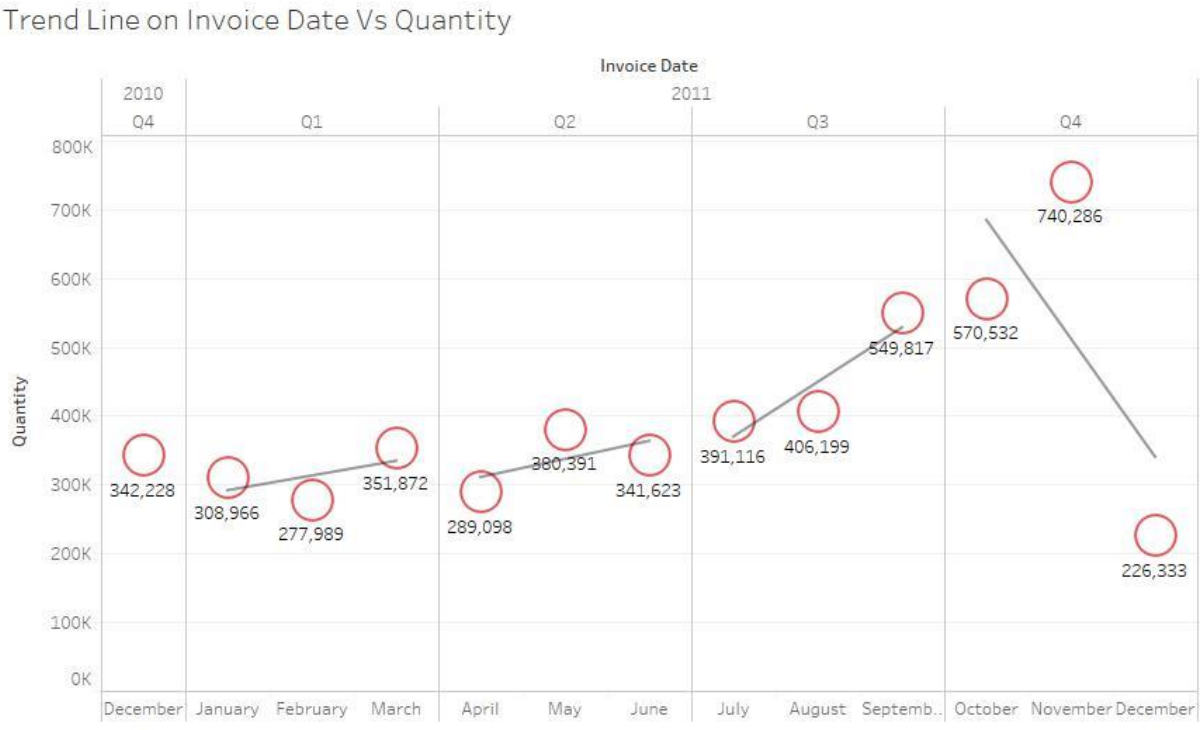
The business – Online Retail Store will be able to use predictive models to exploit patterns found in historical data to identify potential risks & opportunities before they occur.

Based on the Insights gained after performing exploratory analytics on the online retail dataset, we can say that there are strong relationships between attributes like InvoiceDate, Quantity, UnitPrice, Sales. In order to proceed to predictive modelling we must apply InvoiceDate against the attributes Quanitity, UnitPrice & Sales by splitting analysis of Invoice date into days, months, years accordingly.

We shall use these relationships to perform predictive analytics and develop an understanding of such relationships. Also, for forecasting models, it is necessary that one of the attributes to be a date attribute – InvoiceDate in this case, to perform time series forecasting and make future predictions.

Dashboard 1: Predictive Analytics on Invoice Date Vs Quantity

Trend Line:



Trend Lines Model

A linear trend model is computed for sum of Quantity given Invoice Date Month.

Model formula:	Quarter of Invoice Date*Year of Invoice Date*(Month of Invoice Date + intercept)
Number of modeled observations:	13
Number of filtered observations:	0
Model degrees of freedom:	9
Residual degrees of freedom (DF):	4
SSE (sum squared error):	8.53147e+10
MSE (mean squared error):	2.13287e+10
R-Squared:	0.648151
Standard error:	146043
p-value (significance):	0.575134

Analysis of Variance:

Field	DF	SSE	MSE	F	p-value
Quarter of Invoice Date	6	8.6286517e+10	1.43811e+10	0.674261	0.68324
Year of Invoice Date	1	2060993.5	2.06099e+06	9.663e-05	0.992628

Individual trend lines:

Panes		Line		Coefficients				
Row	Column	p-value	DF	Term	Value	StdErr	t-value	p-value
Quantity	2010, Q4	N/A	0	Month of Invoice Date	0	Because the trend line model response variable is constant, there is no information to estimate model statistics.		
				intercept	342228			
Quantity	2011, Q1	0.607493	1	Month of Invoice Date	21453	30270.5	0.70871	0.607493
				intercept	270036	65391.7	4.12952	0.151251
Quantity	2011, Q2	0.611419	1	Month of Invoice Date	26262.5	37545.4	0.699487	0.611419
				intercept	205725	190213	1.08155	0.475072
Quantity	2011, Q3	0.278457	1	Month of Invoice Date	79350.5	37104.9	2.13855	0.278457
				intercept	-185760	298381	-0.62256	0.645503
Quantity	2011, Q4	0.543474	1	Month of Invoice Date	-172100	197369	-0.871967	0.543474
				intercept	2.40548e+06	2.17703e+06	1.10493	0.46829

According to the model the trend lines formed by taking InvoiceDate against Quantity is a best-fit straight line. The data is linear since the pattern in the datapoints resembles a line. The linear trendline shows that it's increasing steadily but then sharply decreases around Quarter 4 of 2011.

Forecasting:

Forecast Summary:

Options Used to Create Forecasts

Time series:	Month of Invoice Date
Measures:	Sum of Quantity
Forecast forward:	6 months (Dec 2011 – May 2012)
Forecast based on:	Dec 2010 – Nov 2011
Ignore last:	1 month (Dec 2011)
Seasonal pattern:	None (Not enough data to search for a seasonal pattern recurring every 12 Months)

Sum of Quantity

Initial Dec 2011	Change From Initial Dec 2011 – May 2012	Seasonal Effect High	Seasonal Effect Low	Contribution Trend	Contribution Season	Quality
823.428 ± 257.460	1,136,256	None		100.0%	0.0%	Ok

Forecast Model:

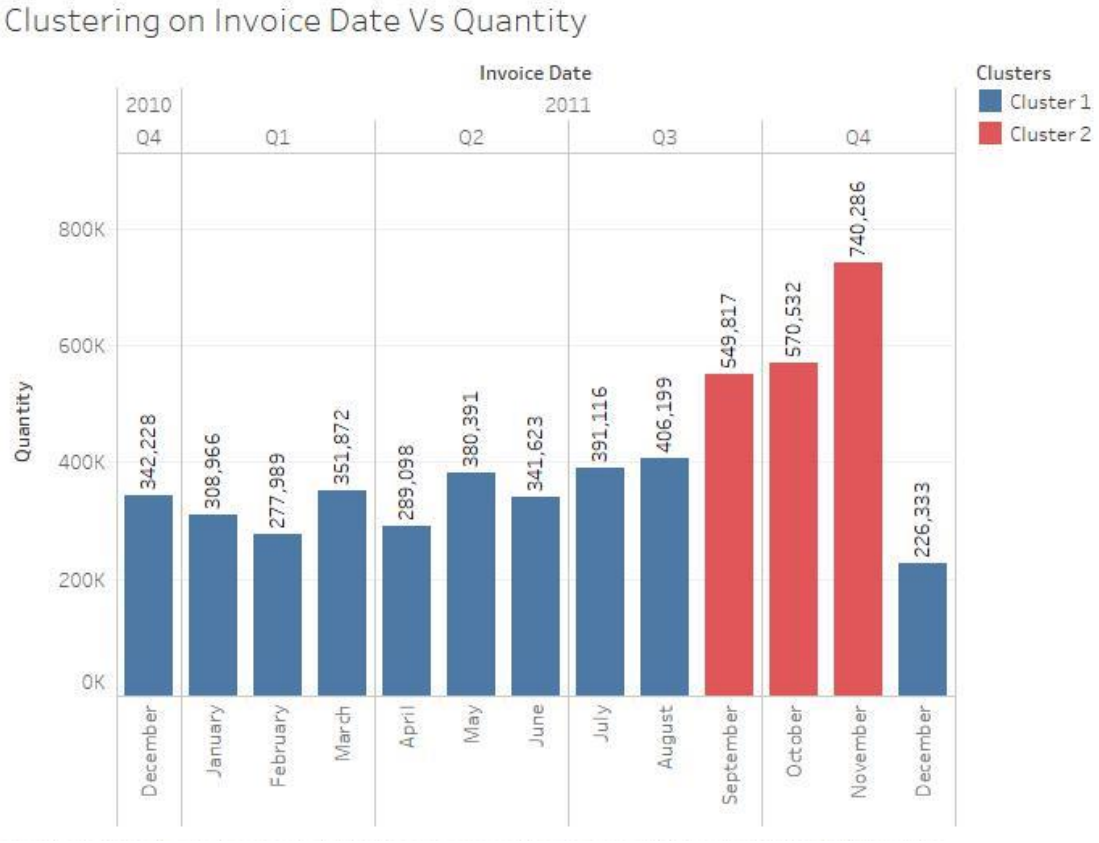
All forecasts were computed using exponential smoothing.

Sum of Quantity

Model Level	Model Trend	Model Season	Quality Metrics RMSE	MAE	MASE	MAPE	AIC	Smoothing Coefficients Alpha	Beta	Gamma
Multiplicative	Multiplicative	None	64.283	52.991	0.80	13.4%	276	0.500	0.500	0.000

The forecast model for InvoiceDate Vs Quantity thus predicts that by the end of Quarter four of 2011 there will be an increase in demand for products and to meet these demands there will be an increase in quantity of items for 2012.

Clustering:



Inputs for Clustering

Variables:	Sum of Quantity
Level of Detail:	Month of Invoice Date, Quarter of Invoice Date, Year of Invoice Date
Scaling:	Normalized

Summary Diagnostics

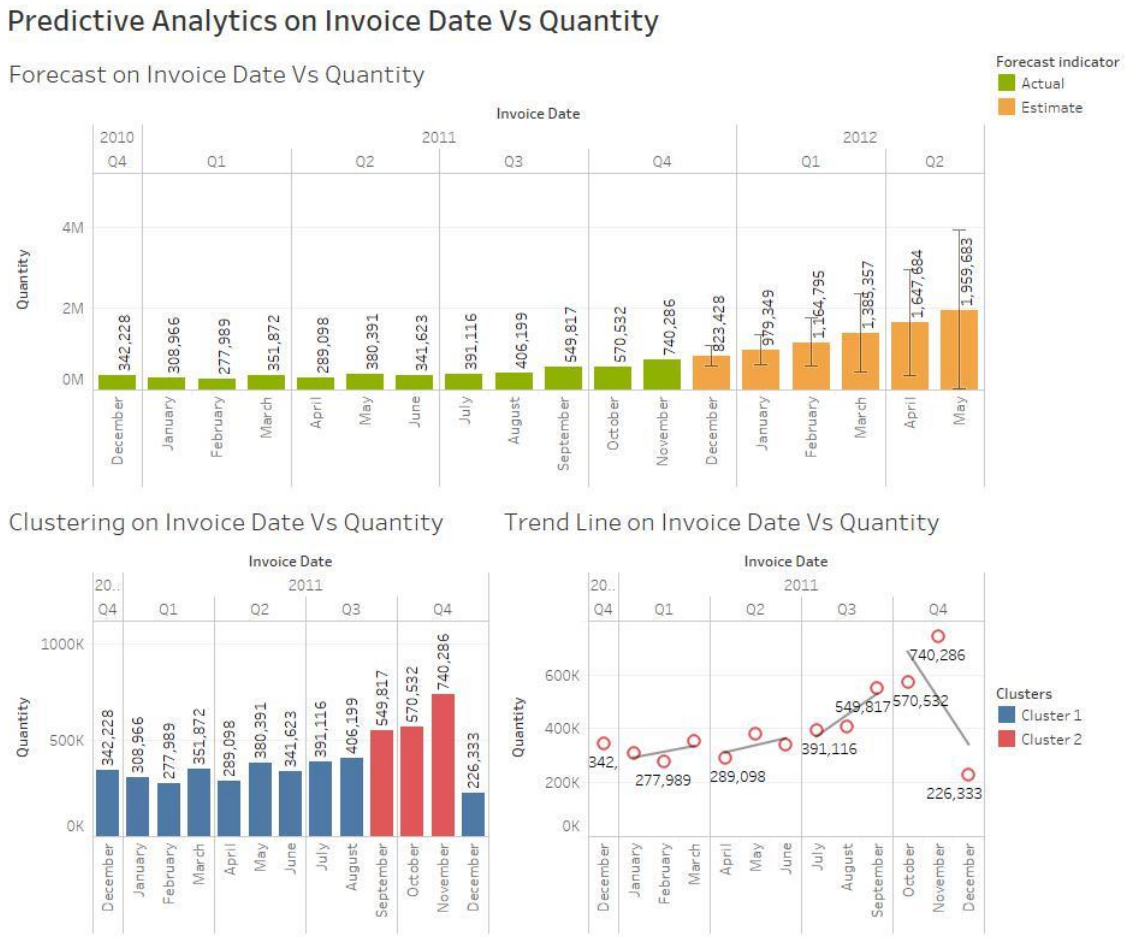
Number of Clusters:	2
Number of Points:	13
Between-group Sum of Squares:	0.7278
Within-group Sum of Squares:	0.19015
Total Sum of Squares:	0.91795

		Centers
Clusters	**Number of Items**	**Sum of Quantity**
Cluster 1	10	3.3158e+05
Cluster 2	3	6.2021e+05
Not Clustered	0	

Analysis of Variance:

			Model		Error	
Variable	**F-statistic**	**p-value**	**Sum of Squares**	**DF**	**Sum of Squares**	**DF**
Sum of Quantity	8.721	0.01313	0.7278	1	0.918	11

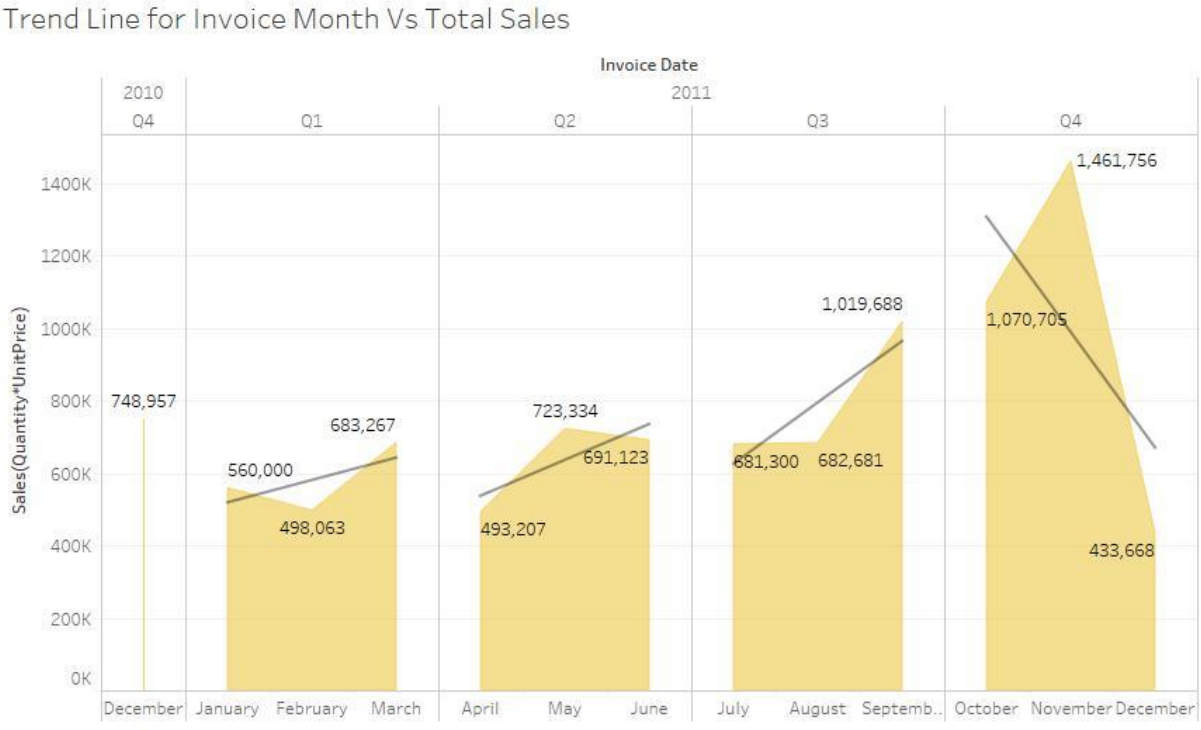
The Cluster Model reveals that after taking Sum(Quantity) & MONTH(InvoiceDate) as the parameters we obtain 2 clusters based on 13 points. Thus, from the bar graph, red bars can form one cluster and blue ones form the other. This is due to the similar properties exhibited within each cluster.



Thus, the resulting graph presents us with the information of predictive analytics performed on InvoiceDate Vs Quantity, predicting that Quantity will increase over the next few years to meet the supply and demand of items by the customers.

Dashboard 2: Predictive Analysis for Invoice Month Vs Total Sales

Trend Line:



Trend Lines Model

A linear trend model is computed for sum of Sales(Quantity*UnitPrice) given Invoice Date Month.

Model formula:	Quarter of Invoice Date*Year of Invoice Date*(Month of Invoice Date + intercept)
Number of modeled observations:	13
Number of filtered observations:	0
Model degrees of freedom:	9
Residual degrees of freedom (DF):	4
SSE (sum squared error):	3.76084e+11
MSE (mean squared error):	9.4021e+10
R-Squared:	0.610536
Standard error:	306628
p-value (significance):	0.644597

Analysis of Variance:

Field	DF	SSE	MSE	F	p-value
Quarter of Invoice Date	6	3.1655783e+11	5.27596e+10	0.561148	0.749037
Year of Invoice Date	1	3.3840199e+09	3.38402e+09	0.0359922	0.85877

Individual trend lines:

Panes Row	Column	Line p-value	DF	Coefficients Term	Value	StdErr	t-value	p-value
Sales(Quantity*UnitPrice)	2010, Q4	N/A	0	Month of Invoice Date	0	Because the trend line model response variable is constant, there is no information to estimate model statistics.		
				intercept	748957			
Sales(Quantity*UnitPrice)	2011, Q1	0.546405	1	Month of Invoice Date	61633.4	71343.8	0.863893	0.546405
				intercept	457177	154120	2.96636	0.206996
Sales(Quantity*UnitPrice)	2011, Q2	0.415844	1	Month of Invoice Date	98958	75730.1	1.30672	0.415844
				intercept	141098	383666	0.367762	0.775649
Sales(Quantity*UnitPrice)	2011, Q3	0.33108	1	Month of Invoice Date	169194	96887.1	1.7463	0.33108
				intercept	-558994	779123	-0.717465	0.603799
Sales(Quantity*UnitPrice)	2011, Q4	0.579277	1	Month of Invoice Date	-318518	409670	-0.777499	0.579277
				intercept	4.49241e+06	4.51877e+06	0.994167	0.501862

It is seen that the trend model formed by taking InvoiceDate against Sales resembles the trend model formed by taking InvoiceDate against Quantity. Since, Sales is a composite of Quantity. Thus, it forms a best fit straight line. The trendline shows it was increasing nicely but then steeply decreases around October, November & December of 2011.

Forecasting:

Forecast Summary:

Options Used to Create Forecasts

Time series:	Month of Invoice Date
Measures:	Sum of Sales(Quantity*UnitPrice)
Forecast forward:	6 months (Dec 2011 – May 2012)
Forecast based on:	Dec 2010 – Nov 2011
Ignore last:	1 month (Dec 2011)
Seasonal pattern:	None (Not enough data to search for a seasonal pattern recurring every 12 Months)

Sum of Sales(Quantity*UnitPrice)

Color	Initial	Change From Initial	Seasonal Effect		Contribution		
Profit/Loss	Dec 2011	Dec 2011 – May 2012	High	Low	Trend	Season	Quality
Profit	1,682,775 ± 714,252	2,754,551	None		100.0%	0.0%	Poor
Loss	-64,521 ± 57,257	0	None		100.0%	0.0%	Ok

Forecast Models:

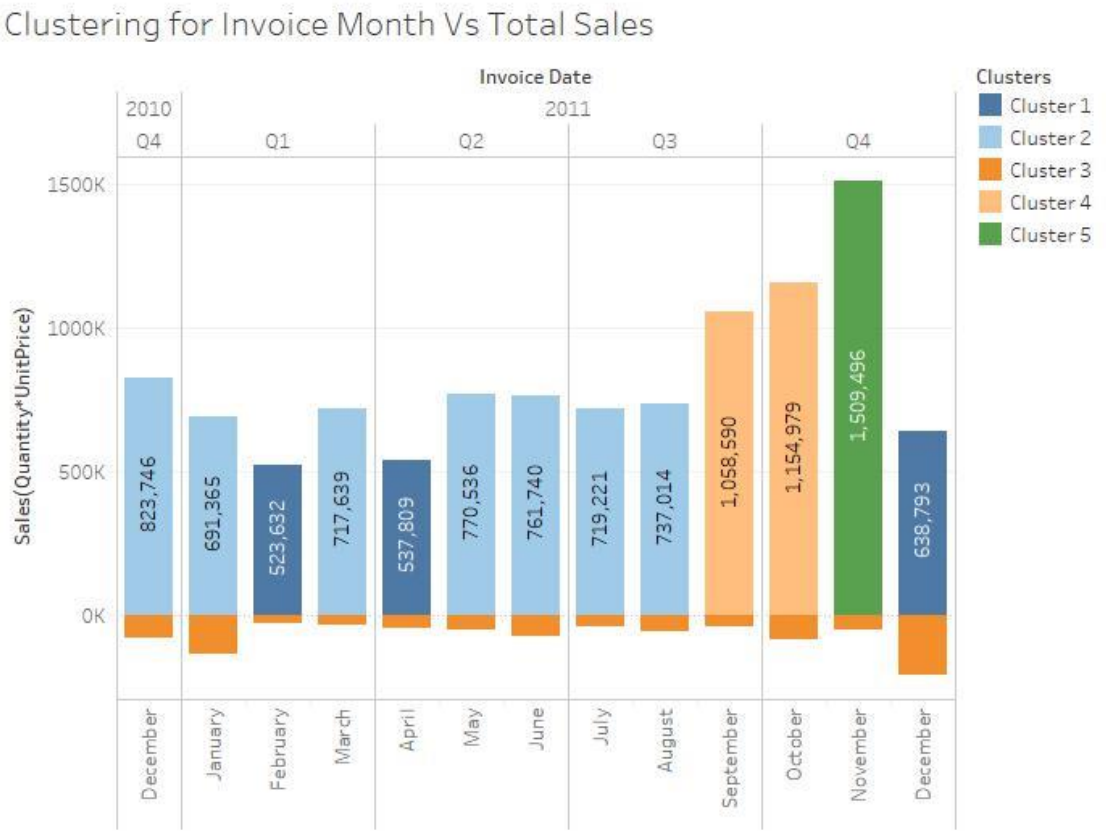
All forecasts were computed using exponential smoothing.

Sum of Sales(Quantity*UnitPrice)

Color	Model			Quality Metrics					Smoothing Coefficients		
Profit/Loss	Level	Trend	Season	RMSE	MAE	MASE	MAPE	AIC	Alpha	Beta	Gamma
Profit	Multiplicative	Multiplicative	None	175,076	148,621	0.94	19.2%	300	0.500	0.500	0.000
Loss	Additive	None	None	29,213	24,626	0.77	53.1%	253	0.020	0.000	0.000

According to the forecasting model there is bound to be an increase in Sales after December 2011. This is because is Quantity increases, it meets the demand & supply, thus demand supply increases even more. And since more quantity, there will be more sales and more profits.

Clustering:



Inputs for Clustering

Variables:	Sum of Sales(Quantity*UnitPrice)
Level of Detail:	Month of Invoice Date, Profit/Loss, Quarter of Invoice Date, Year of Invoice Date
Scaling:	Normalized

Summary Diagnostics

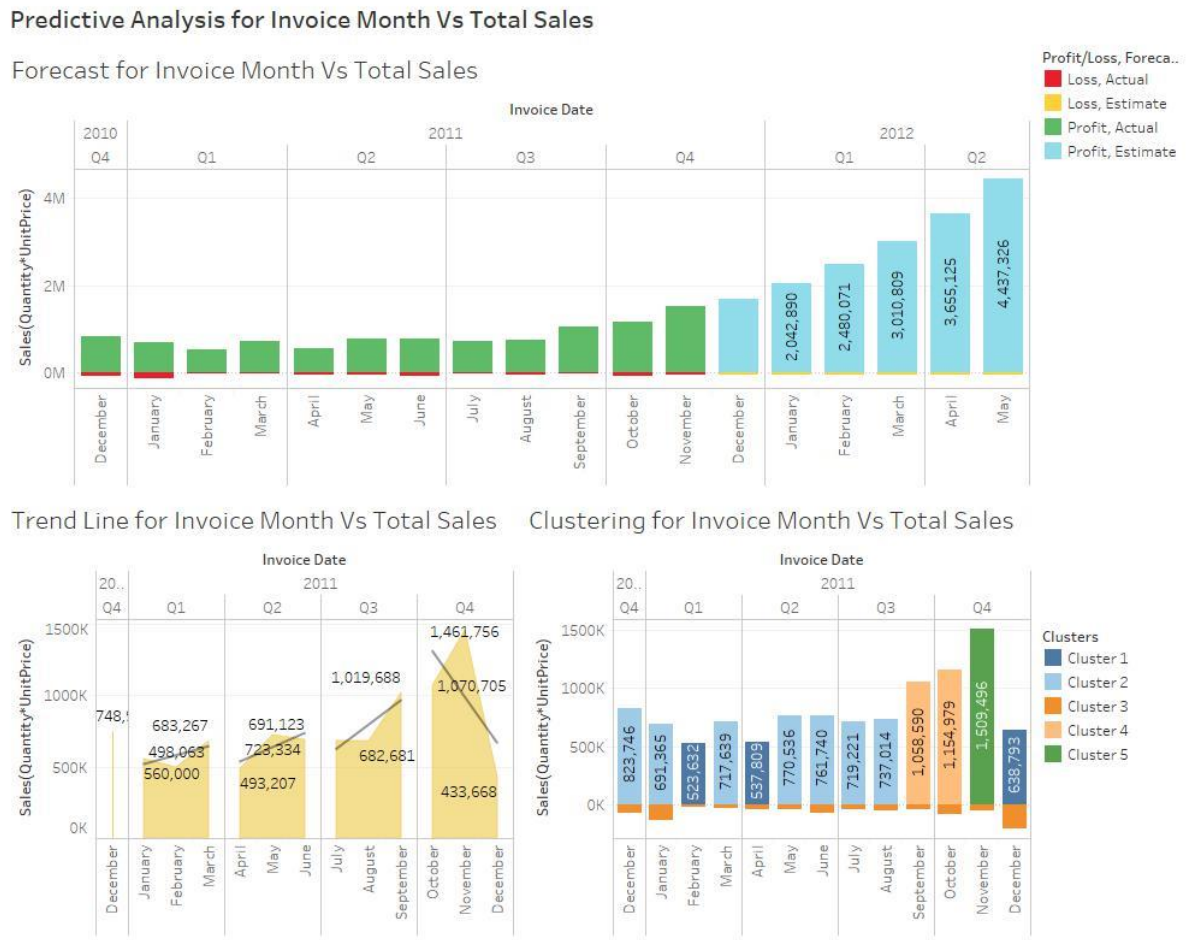
Number of Clusters:	5
Number of Points:	26
Between-group Sum of Squares:	2.0388
Within-group Sum of Squares:	0.018163
Total Sum of Squares:	2.057

		Centers
Clusters	**Number of Items**	**Sum of Sales(Quantity*UnitPrice)**
Cluster 1	3	5.6674e+05
Cluster 2	7	7.4589e+05
Cluster 3	13	-68986.0
Cluster 4	2	1.1068e+06
Cluster 5	1	1.5095e+06
Not Clustered	0	

Analysis of Variance:

			Model		Error	
Variable	**F-statistic**	**p-value**	**Sum of Squares**	**DF**	**Sum of Squares**	**DF**
Sum of Sales(Quantity*UnitPrice)	5.204	0.004504	2.039	4	2.057	21

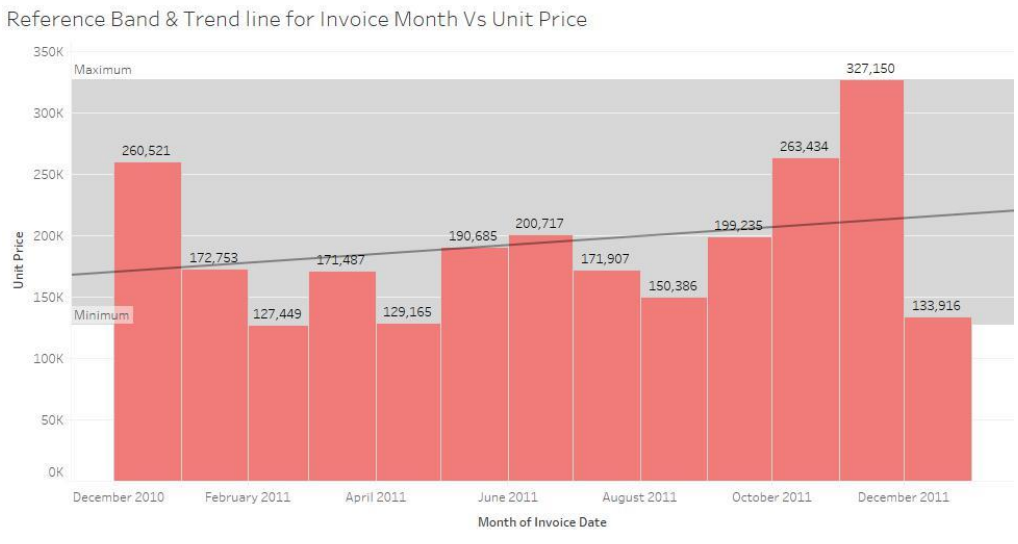
Thus, taking SUM(Sales) & MONTH(InvoiceDate) as the input parameters, we get 5 clusters based on 26 number of points. This can be used to study the clusters of the instances that lead to good sales clusters, so that we can use similar conditions to replicate such characteristics.



The resulting dashboard provides us with the trend line, forecast model & clustering of Invoice Month vs SUM(Sales) which shows a good promise in sales for the next few years based on Quantity increase.

Dashboard 3: Predictive Analysis for Invoice Month Vs UnitPrice

Trend Line:



Trend Lines Model

A linear trend model is computed for sum of Unit Price given Invoice Date Month.

Model formula:	(Month of Invoice Date + intercept)
Number of modeled observations:	13
Number of filtered observations:	0
Model degrees of freedom:	2
Residual degrees of freedom (DF):	11
SSE (sum squared error):	4.01775e+10
MSE (mean squared error):	3.6525e+09
R-Squared:	0.0570555
Standard error:	60435.9
p-value (significance):	0.431906

Individual trend lines:

Panes		Line		Coefficients				
Row	**Column**	**p-value**	**DF**	**Term**	**Value**	**StdErr**	**t-value**	**p-value**
Unit Price	Month of Invoice Date	0.431906	11	Month of Invoice Date	120.173	147.301	0.815835	0.431906
				intercept	-4.69827e+06	5.99448e+06	-0.783766	0.449725

The trend line calculated for Invoice Month Vs UnitPrice shows us a single best fit line which is formed using linear regression. This line is a straight line because of the linearity of dataset & the 13 datapoints used for this observation. Thus, displaying a steady growth in unit price which could keep increasing.

Forecasting:

The trend of sum of Unit Price (actual & forecast) for Invoice Date Month. Color shows details about Forecast indicator. The data is filtered on Action (MONTH(Invoice Date)), which keeps 13 members.

Forecast Summary:

Options Used to Create Forecasts

Time series: Month of Invoice Date
Measures: Sum of Unit Price
Forecast forward: 6 months (Dec 2011 – May 2012)
Forecast based on: Dec 2010 – Nov 2011
Ignore last: 1 month (Dec 2011)
Seasonal pattern: None (Not enough data to search for a seasonal pattern recurring every 12 Months)

Sum of Unit Price

Initial Dec 2011	Change From Initial Dec 2011 – May 2012	Seasonal Effect High	Seasonal Effect Low	Contribution Trend	Contribution Season	Quality
274,892 ± 109,341	0	None		0.0%	0.0%	Poor

Forecast Model:

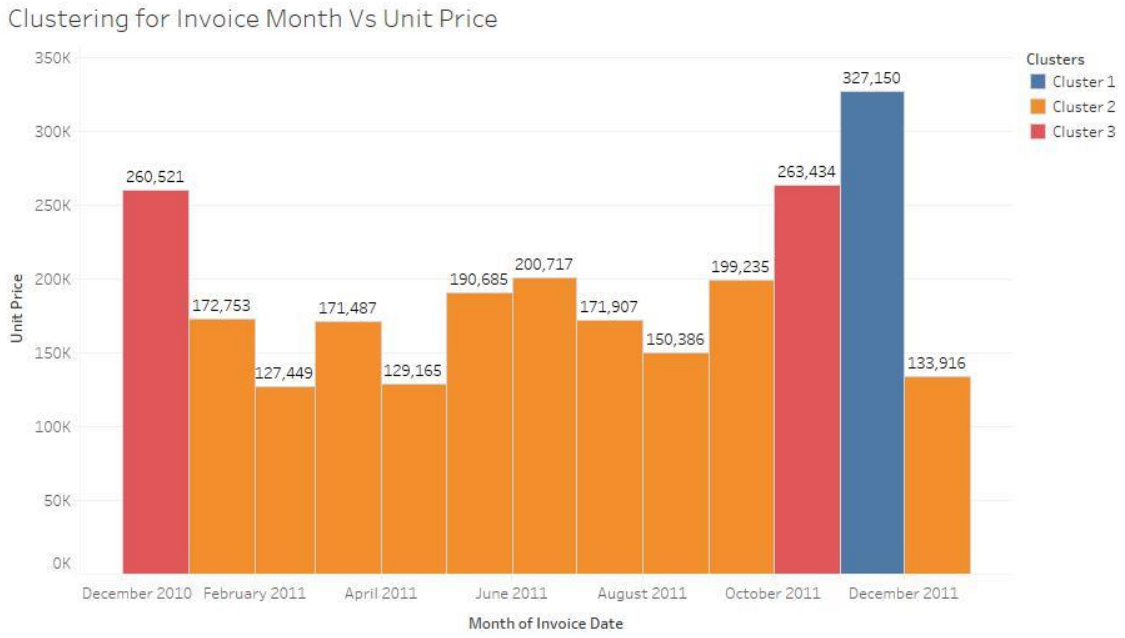
All forecasts were computed using exponential smoothing.

Sum of Unit Price

Model Level	Model Trend	Model Season	Quality Metrics RMSE	MAE	MASE	MAPE	AIC	Smoothing Coefficients Alpha	Beta	Gamma
Additive	None	None	55,787	48,279	1.03	24.3%	268	0.500	0.000	0.000

The results of the forecasting model revealed that the data available does not provide enough information to make a predicted guess of UnitPrice in the next few years. Unit price may increase or decrease depending on the market, thus there are fluctuations seen in the actual trend line for unit price whereas the estimated trend line remains constant.

Clustering:



The plot of sum of Unit Price for Invoice Date Month. Color shows details about Clusters. The data is filtered on Action (MONTH(Invoice Date)), which keeps 13 members.

Inputs for Clustering

Variables:	Sum of Unit Price
Level of Detail:	Month of Invoice Date
Scaling:	Normalized

Summary Diagnostics

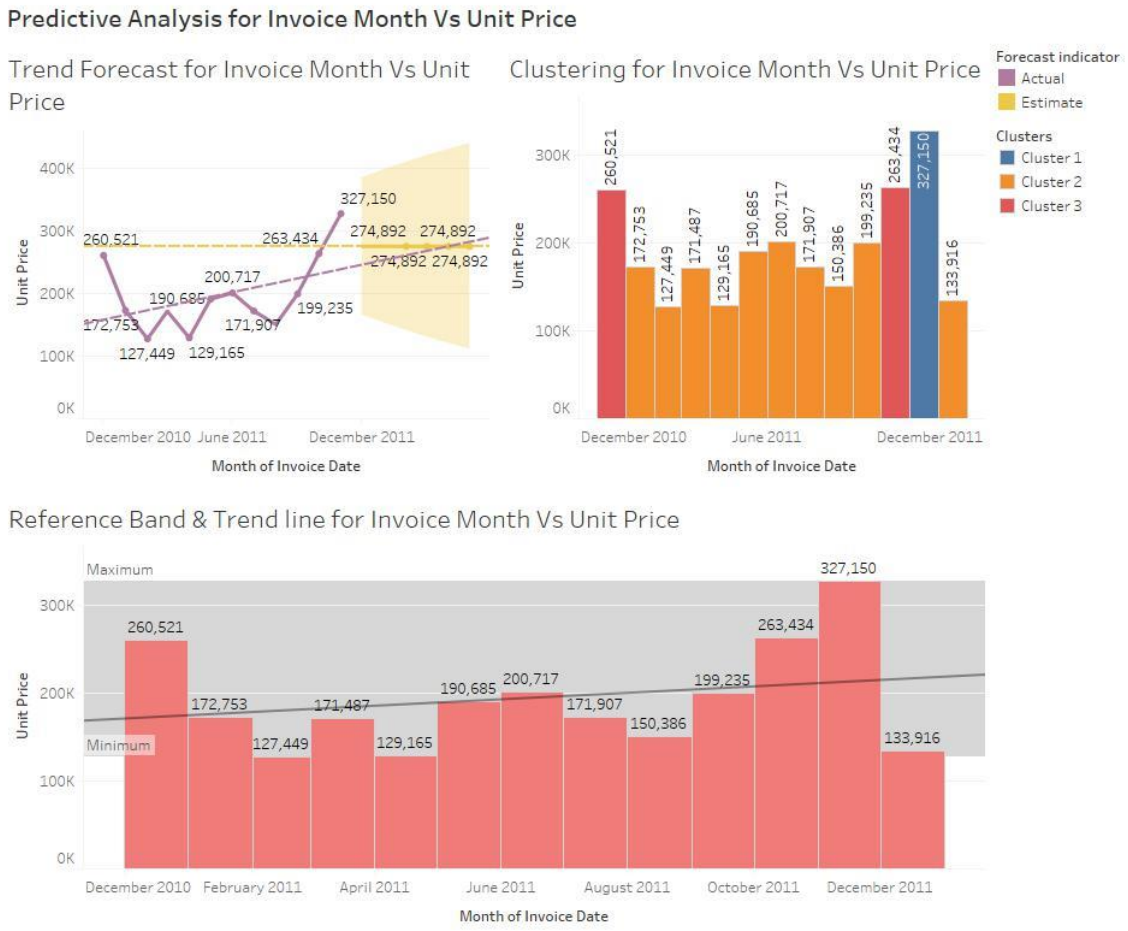
Number of Clusters:	3
Number of Points:	13
Between-group Sum of Squares:	0.88949
Within-group Sum of Squares:	0.17891
Total Sum of Squares:	1.0684

		Centers
Clusters	**Number of Items**	**Sum of Unit Price**
Cluster 1	1	3.2715e+05
Cluster 2	10	1.6477e+05
Cluster 3	2	2.6198e+05
Not Clustered	0	

Analysis of Variance:

			Model		Error	
Variable	**F-statistic**	**p-value**	**Sum of Squares**	**DF**	**Sum of Squares**	**DF**
Sum of Unit Price	4.163	0.04839	0.8895	2	1.068	10

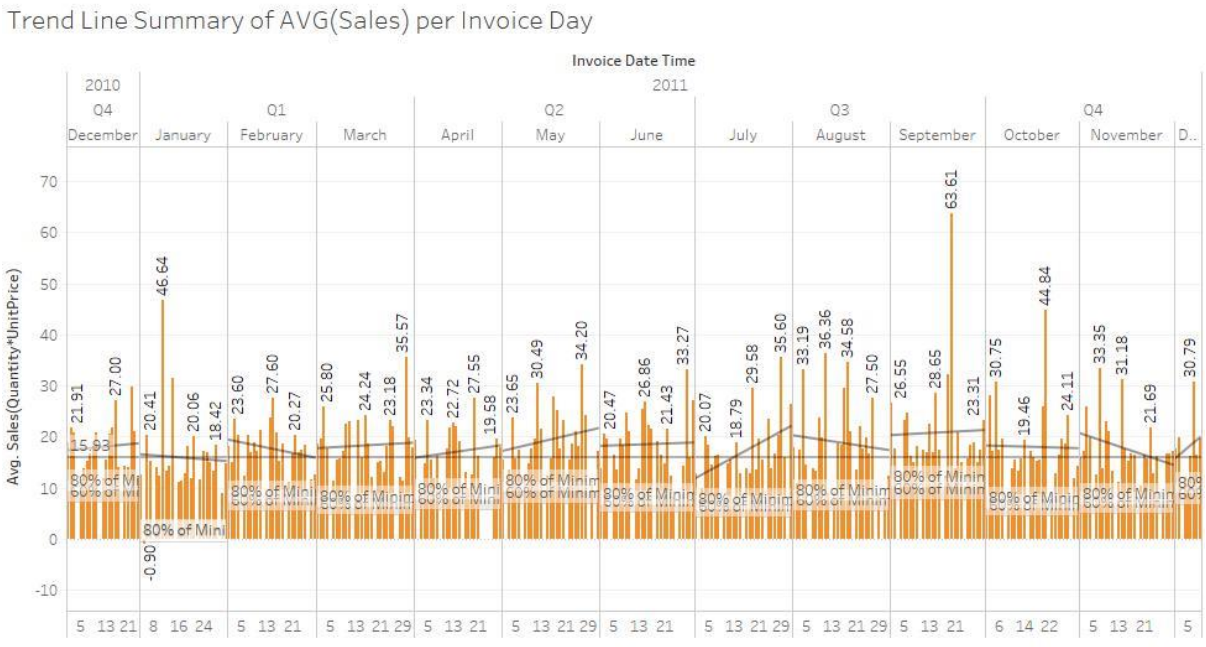
The results of clustering reveal that 3 clusters are formed out of the 13 points used observation inputs. 1 bar/point form one blue cluster, 2 points form red cluster & rest remaining form orange cluster. This reveals that in the future if the unit price needs to be grouped under blue cluster then it must possess or exhibit characteristics like it.



Thus, the resulting dashboard reveals the relationship and also provides forecasting information on the variables like Invoice Month and UnitPrice through trend lines, forecasting & clustering.

Dashboard 4: Predictive Analytics of AVG(Sales) per Invoice Day

Trend Lines:



Trend Lines Model

A linear trend model is computed for average of Sales(Quantity*UnitPrice) given Invoice Date Time Day.

Model formula:	Month of Invoice Date Time*Quarter of Invoice Date Time*Year of Invoice Date Time*(Day of Invoice Date Time + intercept)
Number of modeled observations:	305
Number of filtered observations:	0
Model degrees of freedom:	26
Residual degrees of freedom (DF):	279
SSE (sum squared error):	12941.9
MSE (mean squared error):	46.3869
R-Squared:	0.0632728
Standard error:	6.81079
p-value (significance):	0.798171

Analysis of Variance:

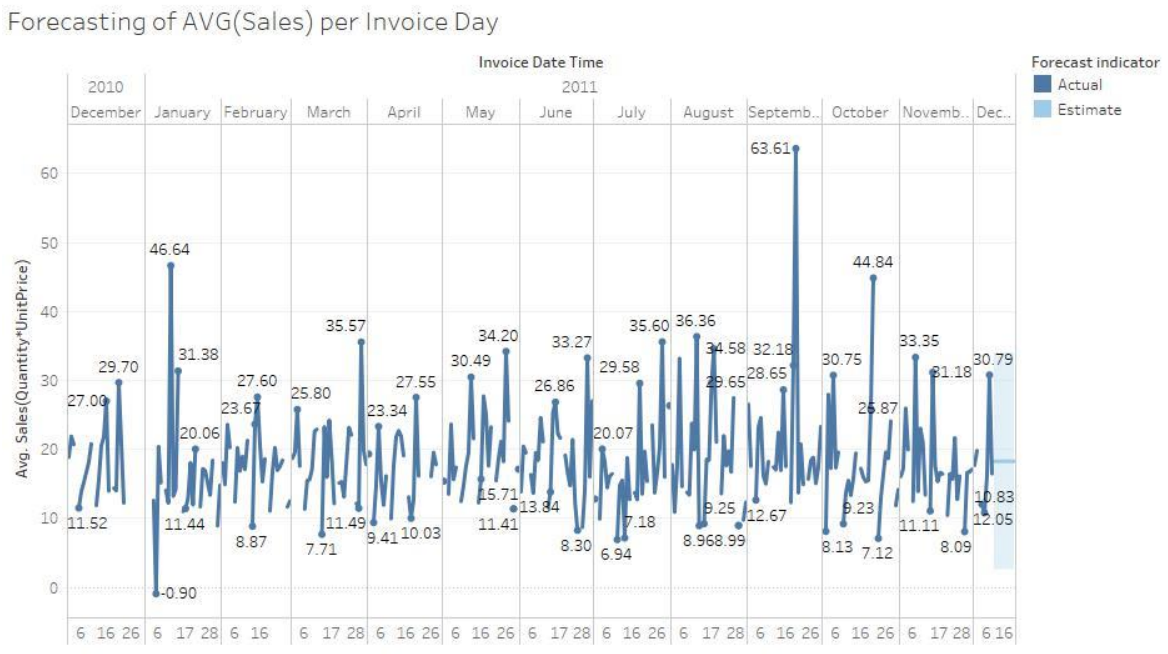
Field	DF	SSE	MSE	F	p-value
Month of Invoice Date Time	16	622.37369	38.8984	0.838563	0.641084
Quarter of Invoice Date Time	0	-1.8189894e-12	N/A	N/A	N/A
Year of Invoice Date Time	2	11.243161	5.62158	0.121189	0.885913

Individual trend lines:

Panes		Line		Coefficients				
Row	Column	p-value	DF	Term	Value	StdErr	t-value	p-value
Sales(Quantity*UnitPrice)	2010, Q4, December	0.720444	18	Day of Invoice Date Time	0.0613233	0.168689	0.363529	0.720444
				intercept	17.2529	2.34441	7.35915	< 0.0001
Sales(Quantity*UnitPrice)	2011, Q1, January	0.830731	22	Day of Invoice Date Time	-0.0478032	0.220983	-0.216321	0.830731
				intercept	16.7078	4.22862	3.95112	0.0006794
Sales(Quantity*UnitPrice)	2011, Q1, February	0.253629	22	Day of Invoice Date Time	-0.128423	0.109551	-1.17226	0.253629
				intercept	19.5021	1.804	10.8105	< 0.0001
Sales(Quantity*UnitPrice)	2011, Q1, March	0.778809	25	Day of Invoice Date Time	0.0349125	0.122965	0.283922	0.778809
				intercept	17.6972	2.27107	7.79245	< 0.0001
Sales(Quantity*UnitPrice)	2011, Q2, April	0.524655	19	Day of Invoice Date Time	0.0891886	0.137608	0.648133	0.524655
				intercept	15.7601	2.15684	7.30701	< 0.0001
Sales(Quantity*UnitPrice)	2011, Q2, May	0.289089	23	Day of Invoice Date Time	0.146036	0.134574	1.08517	0.289089
				intercept	17.0719	2.41364	7.07308	< 0.0001
Sales(Quantity*UnitPrice)	2011, Q2, June	0.871614	24	Day of Invoice Date Time	0.0222395	0.136149	0.163346	0.871614
				intercept	18.1523	2.44268	7.43132	< 0.0001
Sales(Quantity*UnitPrice)	2011, Q3, July	0.0154054	24	Day of Invoice Date Time	0.336406	0.128965	2.6085	0.0154054
				intercept	11.5825	2.35177	4.92504	< 0.0001
Sales(Quantity*UnitPrice)	2011, Q3, August	0.598316	24	Day of Invoice Date Time	-0.0948764	0.177702	-0.533908	0.598316
				intercept	20.3148	3.1637	6.42122	< 0.0001
Sales(Quantity*UnitPrice)	2011, Q3, September	0.885313	24	Day of Invoice Date Time	0.033467	0.229573	0.14578	0.885313
				intercept	20.2566	4.14626	4.8855	< 0.0001
Sales(Quantity*UnitPrice)	2011, Q4, October	0.923596	24	Day of Invoice Date Time	-0.0173853	0.179381	-0.096918	0.923596
				intercept	18.2707	3.30032	5.53602	< 0.0001
Sales(Quantity*UnitPrice)	2011, Q4, November	0.107141	24	Day of Invoice Date Time	-0.210505	0.125759	-1.67387	0.107141
				intercept	20.8271	2.24003	9.29767	< 0.0001
Sales(Quantity*UnitPrice)	2011, Q4, December	0.585699	6	Day of Invoice Date Time	0.497628	0.864277	0.575774	0.585699
				intercept	15.3262	5.07648	3.01906	0.0234256

The trend line model for Average(Sales) against Invoice of Day reveals as the data gets spread out the trend lines cannot be formed into a single straight line, instead there are multiple straight lines showing the fluctuation of change that is increase & decrease in the Avg(Sales) for each day in that time frame.

Forecasting:



Forecast Summary:

Options Used to Create Forecasts

Time series: Day of Invoice Date Time
Measures: Avg. Sales(Quantity*UnitPrice)
Forecast forward: 9 days (Dec 9, 2011 – Dec 17, 2011)
Forecast based on: Aug 11, 2011 – Dec 8, 2011
Ignore last: 1 day (Dec 9, 2011)
Seasonal pattern: None (Searched for a seasonal pattern recurring every 7 Days)

Avg. Sales(Quantity*UnitPrice)

Initial Dec 9, 2011	Change From Initial Dec 9, 2011 – Dec 17, 2011	Seasonal Effect High	Seasonal Effect Low	Contribution Trend	Contribution Season	Quality
18.29 ± 15.71	0.00	None		0.0%	0.0%	Ok

Forecast Model:

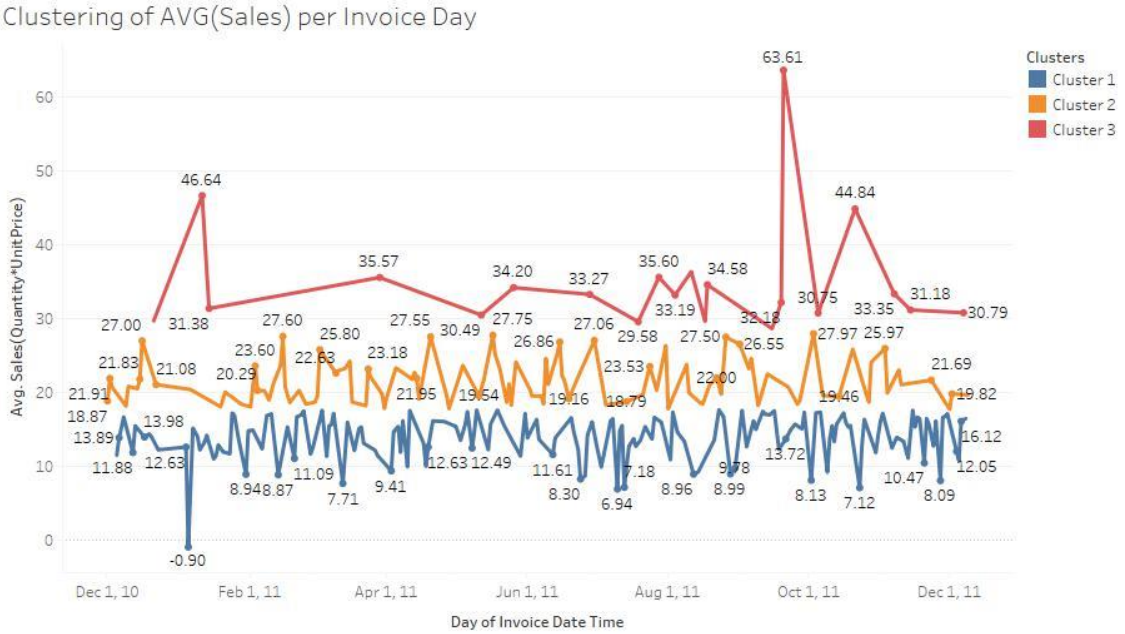
All forecasts were computed using exponential smoothing.

Avg. Sales(Quantity*UnitPrice)

Model Level	Model Trend	Model Season	Quality Metrics RMSE	MAE	MASE	MAPE	AIC	Smoothing Coefficients Alpha	Beta	Gamma
Additive	None	None	8.01	5.28	0.70	30.0%	431	0.000	0.000	0.000

The forecasting model for Invoice of Day against Avg(Sales) reveals that model has to account for so many various instances and cases as the data is too spread out. Thus, the forecast model predicts a constant line with no change throughout 2012. Too many sudden ups & downs can cause such constancy.

Clustering:



Inputs for Clustering

Variables:	Avg. Sales(Quantity*UnitPrice)
Level of Detail:	Day of Invoice Date Time
Scaling:	Normalized

Summary Diagnostics

Number of Clusters:	3
Number of Points:	305
Between-group Sum of Squares:	2.4609
Within-group Sum of Squares:	0.85892
Total Sum of Squares:	3.3198

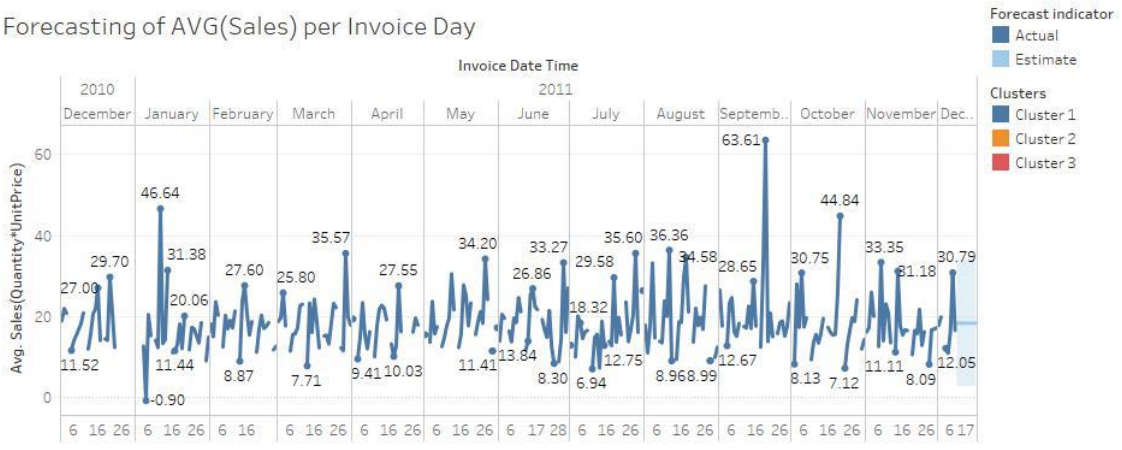
		Centers
Clusters	**Number of Items**	**Avg. Sales(Quantity*UnitPrice)**
Cluster 1	174	13.944
Cluster 2	110	21.43
Cluster 3	21	35.027
Not Clustered	0	

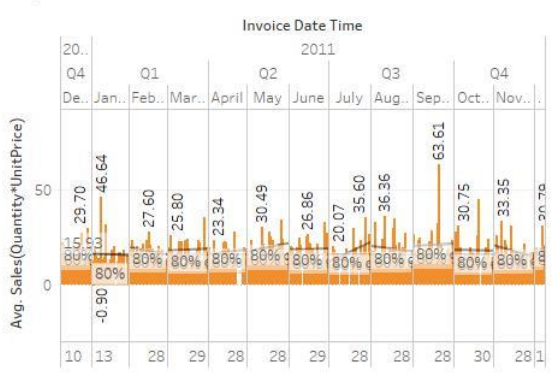
Analysis of Variance:

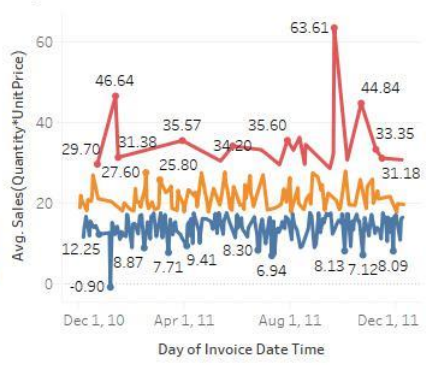
			Model		Error	
Variable	**F-statistic**	**p-value**	**Sum of Squares**	**DF**	**Sum of Squares**	**DF**
Avg. Sales(Quantity*UnitPrice)	111.9	0.0	2.461	2	3.32	302

The clustered model reveals that by taking 305 observation points as inputs we were able to generate 3 clusters without any unclustered instances remaining. The graph shows three separate line graphs with multiple fluctuations. However in order for an instance to be clustered in the future under the red cluster or any other one it must exhibit the similar corresponding cluster's avg(sales).

Predictive Analytics of AVG(Sales) per Invoice Day







The resulting dashboard gives us detailed information on the predictive analytical models used for finding trends, predicting future average(sales) & clustering similar characterised instances according to average(sales).

Highlights/Insights:

1) When taking Invoice Month Vs Quantity, we found that as demand increases, supply will too, thus there will be an increase in quantity by the end of Q4 2011.

2) When taking Invoice Month Vs Total Sales, we found that increase in supply-demand is like increase for quantity, thus there will be an increase in sales from December onwards.

3) When taking Invoice Month Vs Unit Price, we found that unit price has seen sharp fluctuations, but the highest unit price has ever gone was during December 2011. Predictions after that reveals that unit price goes constant because there's not enough data for it.

4) When taking Invoice Day Vs Average Sales, we notice predictions are difficult when the data is too spread over many days, which becomes a constant line of prediction. Thus, average sales can't really be predicted well over several days. For clustering we found three separate possibilities for future predictions of average sales.

Actionable Intelligence:

1) Online Retail should set up measures to provide more sales at a faster and efficient manner. This can be done by larger distribution of warehouses over a country where sales are high. This would increase sales directly too.

2) Also, should promote targeted advertising to keep customers engaged on products with high or low unit price. This can be done by Social or other forms of advertisement.

3) Assess risks by predicting future profits or loss based on historical data or transactions. This can be done reviewing all prices and gradually increasing prices of most bought items by a certain percentage every year. Also, products that don't sell at all should be stopped or discontinued.