Big Data Integration

GROUP 4

Submitted by
Christina Schwierling – M00626134
Mahitha Sree Tammineedi – M12789619
Rashmi Prathigadapa – M12853369
Sanjana Bhosekar – M12870994

Executive Summary

The objective of the project is to provide recommendations to improve the profitability of sales and delivery of Dual Core by harnessing the power of Big Data. In order to achieve this, we have integrated the given data with two external data sources.

The Dual Core dataset has a relational data model. It consists of data related to customers, products, suppliers and order details. We have analyzed this data to understand the trends in sales and identify any potential issues that might be affecting sales.

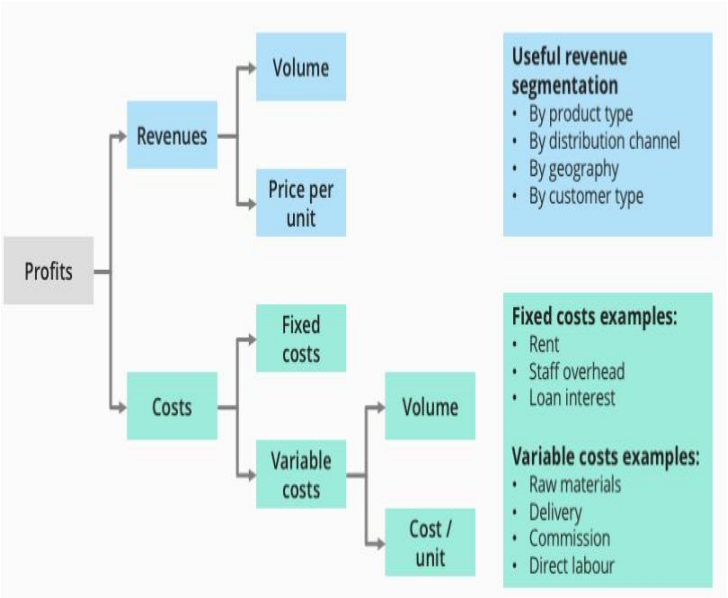
In order to build our mediated schema, we have integrated the Dual Core dataset with two external data sources - the US Zip Codes database and the US holiday calendar database. Our mediated schema will create formatting rules for the data and rules for how we intend to identify duplicate rows and consolidate those rows into a single record. We will use HDFS as our database for records and Sqoop to import the data.

The Zip Codes database consists of location details like – City, State and Zip Code of all the locations in United States. By integrating this data with our Dual Core, we are planning to identify the locations with least activity and therefore, come up with recommendations on where to set up delivery hubs to expand the business. By integrating the US holidays calendar database, the flow/trend of orders generated will be analyzed, based on which recommendations on discounts, order quantity, and product categories will be presented to improve the sales.

Introduction

The dataset used in this project is Dual Core data set which consists of information related to its products, sales, employees and customers. The Sales team leader wanted to leverage Big Data technologies to improve the profitability of sales and delivery. The main goal of this project is to present a thoughtful proposition, analysis, and recommendation to the executive team for how to improve profitability for a web-based sales and delivery company.

Let’s have a look at general profitability framework



We plan on presenting recommendations to increase the volume of the products sold by identifying locations with least activity which may be potential locations to set up delivery hubs to expand the business.
We also merged the holiday dataset with the Dual Core dataset to look into shopping patterns and trends on the products sold over the years. Based on which, recommendations on shopping discounts to increase the sales can be identified.

External Data Sources

External Data Integration with US Holiday dataset

US Holidays list from the year 2008 to 2013 is scraped from the website - https://www.timeanddate.com/holidays/us/ using Python script. It contains four columns - Date, Holiday Name, Holiday Type (whether it is Federal Holiday or State Holiday), and the state where it is observed if it is state holiday. The dataset 'Holidays' has a list of all the holidays in USA. This data is merged with the Dual Core Customer order data.

Mediated Schema

Mediated_Fields	DualCore	USZipCodes
prod_id	order_details.prod_id	
order_id	orders.order_id	
holdate	orders.order_date	holidays.holdate
frequency		holidays.No_holidays
weekday		holidays.weekday

External Data Integration with USZipCodes data

The dataset 'USZipCodes' has a list of all the zip codes in USA. It has 46k unique zip codes of all the states and cities of USA. This data is merged with the Dual Core Customer zip codes data.

Mediated Schema

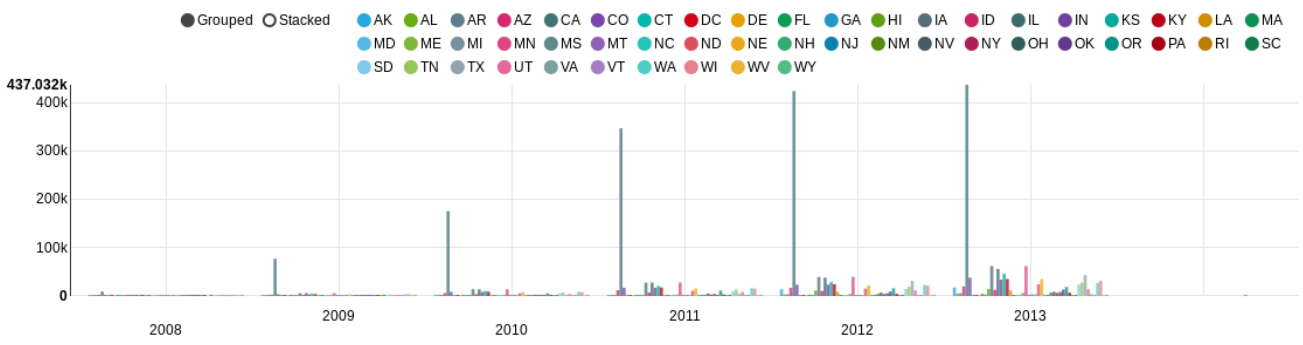
Mediated_Fields	DualCore	USZipCodes
prod_id	order_details.prod_id	
cust_id	orders.cust_id	
order_id	orders.order_id	
city	customers.city	city
state	customers.state	State
lat		lat
lon		Long
zipcode	customers.zipcode	Zipcode

There were some duplicate records in the holiday dataset, because some dates have more than one holiday. Therefore, those duplicate dates were merged and number of holidays (frequency) on a unique date is considered.

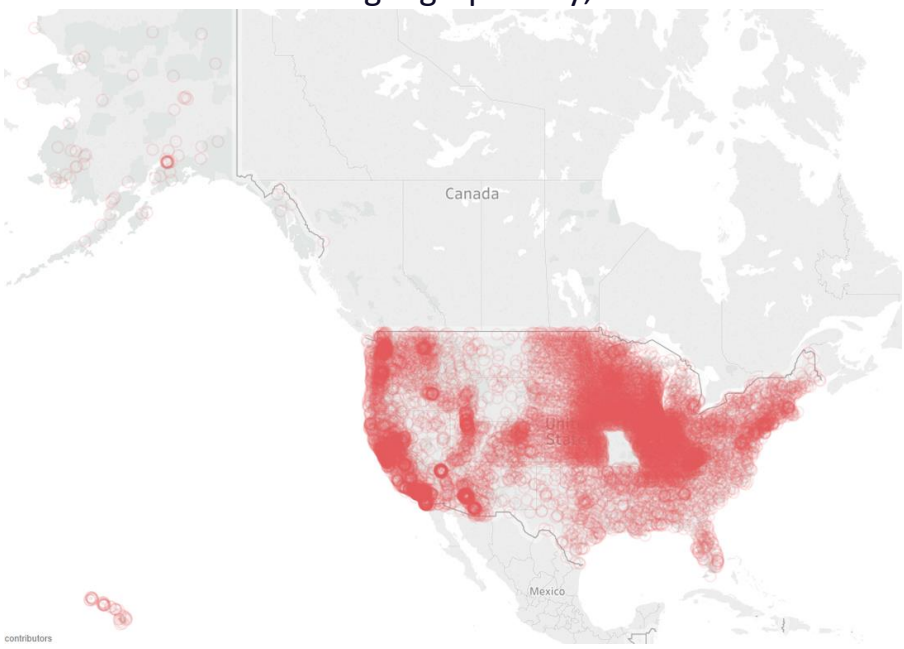
Zip code Analysis:

Analysis to identify locations

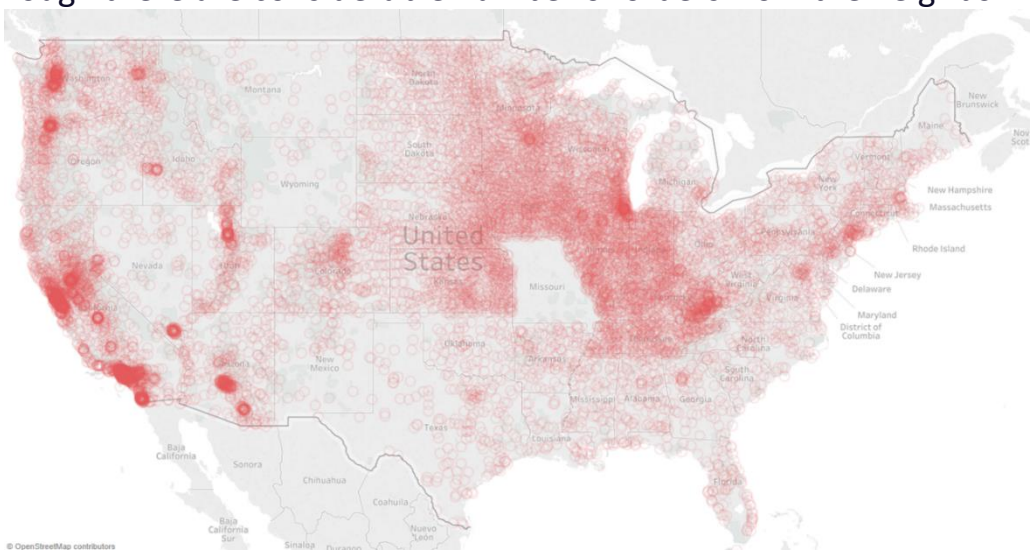
Initially, we looked at the distribution of number of orders across the United States. We observed that most of the orders came in from customers in California. Rest of the orders from other 49 states were negligible when compared to California



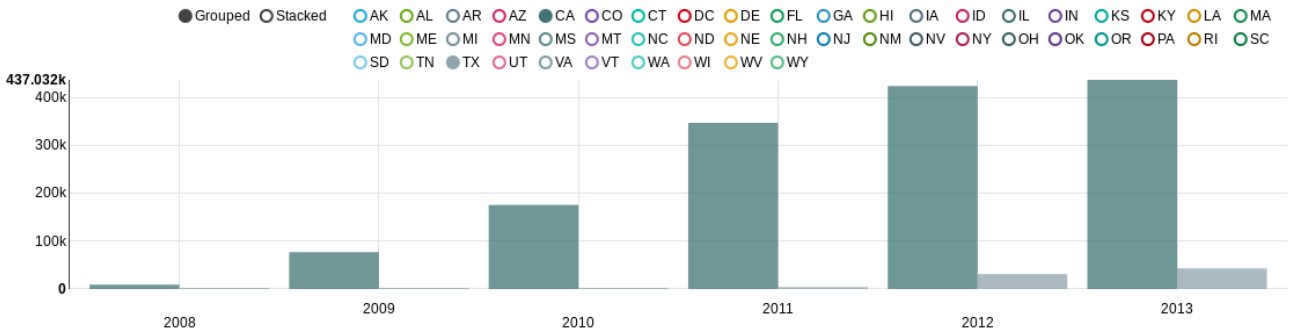
A closer look at distribution of orders geographically,



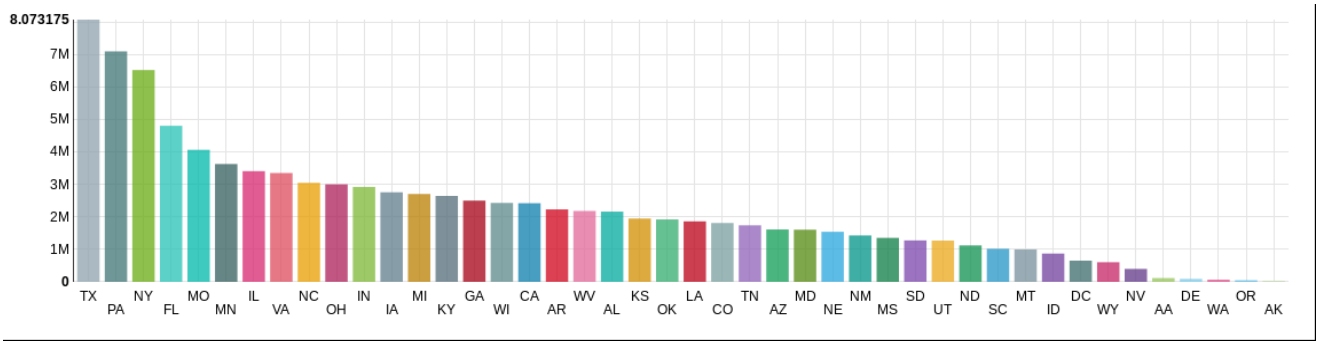
We observed that, number of orders decreased as we move from west coast to east coast. It is also surprising to know that there is not even one order from Missouri in the dataset, though there are considerable number of orders from the neighboring states.



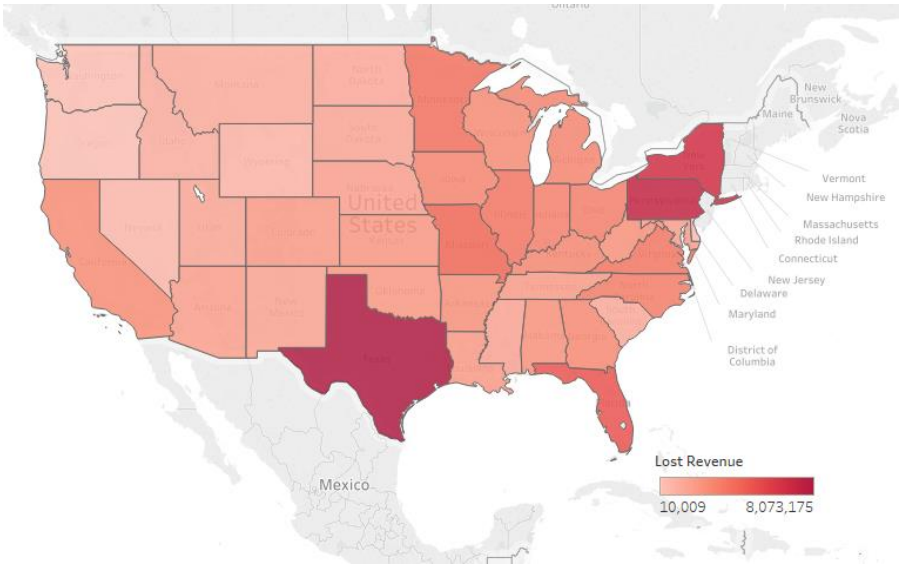
It is also observed that Texas, which is the second largest state in USA has negligible number of orders when compared to California.



This leads to two questions – Do people in other states not aware of DualCore or is there any potential issue? Further investigation at the carts data revealed that people are abandoning their carts after looking at shipping cost. We further dig down and observed that highest lost revenue is from Texas state. We also observed Missouri in the Top 5 states in terms of lost revenue.



It is also interesting to note that most lost revenue is in the eastern side of USA



We also observed that profit is continuously decreasing over the years from 2008 - 2012. Year 2013 shouldn't be considered as the sales are given until May only.

_c0	year1
0.12034106848429127	2008
0.11978806162901845	2009
0.11904880070613469	2010
0.11900036922495279	2011
0.11867678386809524	2012
0.10272072534621063	2013

Our recommendations:

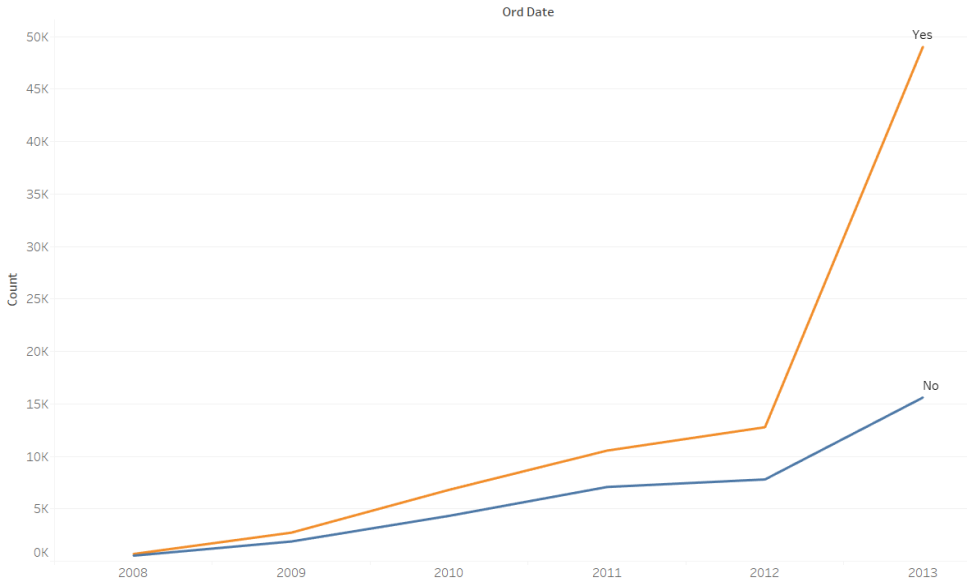
From the analysis of geographic distribution of orders, we conclude that there is market for DualCore products in middle and eastern states and DualCore should focus on increasing its market share by setting up warehouse/delivery hubs in Texas, Pennsylvania, and Florida. By doing this, shipping costs can be reduced, they are accessible to a larger market and thus can increase their revenues. There are costs associated with setting up hubs/warehouses but we observe that sales have been increasing continuously over the years and average profit is 11.9%, so we can expect profitability in the long run

Holiday Orders Analysis:

SAVING MONEY AND TIME are two important factors that affect the shopping trends in the present world. People are looking for cheaper and convenient options available. Timing is everything in the retail world. It's crucial to provide sales and discounts on seasonal items which allows customers to buy everything they want and pay significantly less. This would benefit both the customer and the seller.
In our Dual Core data, let us look at the possible trends with respect to the holidays observed.

Holidays vs Orders

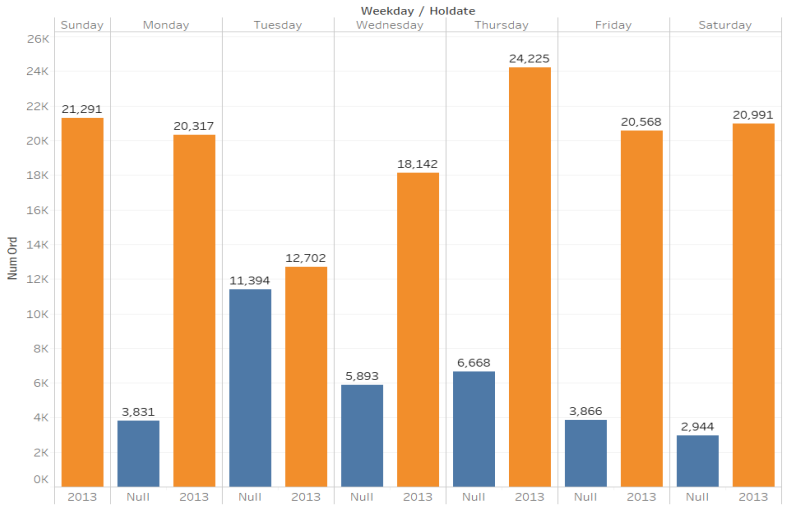
There is a total of 1.6M orders placed in our Dual Core data, out of which, 1M+ orders are placed on holidays!
Let us try understanding the pattern observed in this shopping trends. Below is a graph representing the count of sales on a holiday vs the count of sales on a normal day.



We can notice that there is an unusually higher number of orders placed on day that is a holiday in comparison to a day that is not.

Effect of weekday on Orders

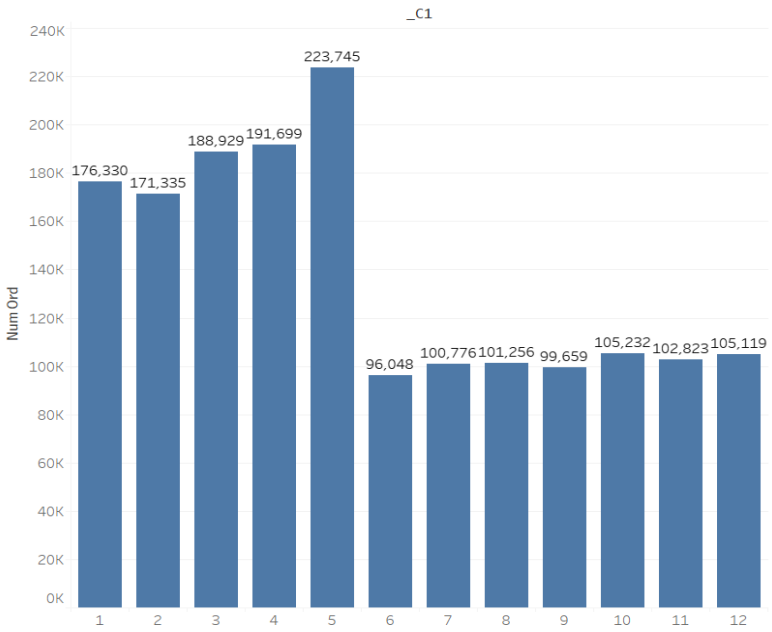
Now, let us further understand this orders trends. The graph below visualizes the top 50 orders placed on a certain day of the week. The blue bars represent the days when there was no holiday. The orange bars which show the sales on weekdays that were a holiday show a significantly high number of sales. Also, we can notice that all the top 50 orders are from the year 2013. This is not surprising as we can observe a sudden spike in the sales data from 2012 to 2013. This could possibly be because this is an online shopping sales database and e-commerce was catching up during that period. Once the trend started catching up among people, it is not surprising that a lot of orders were placed on a holiday.



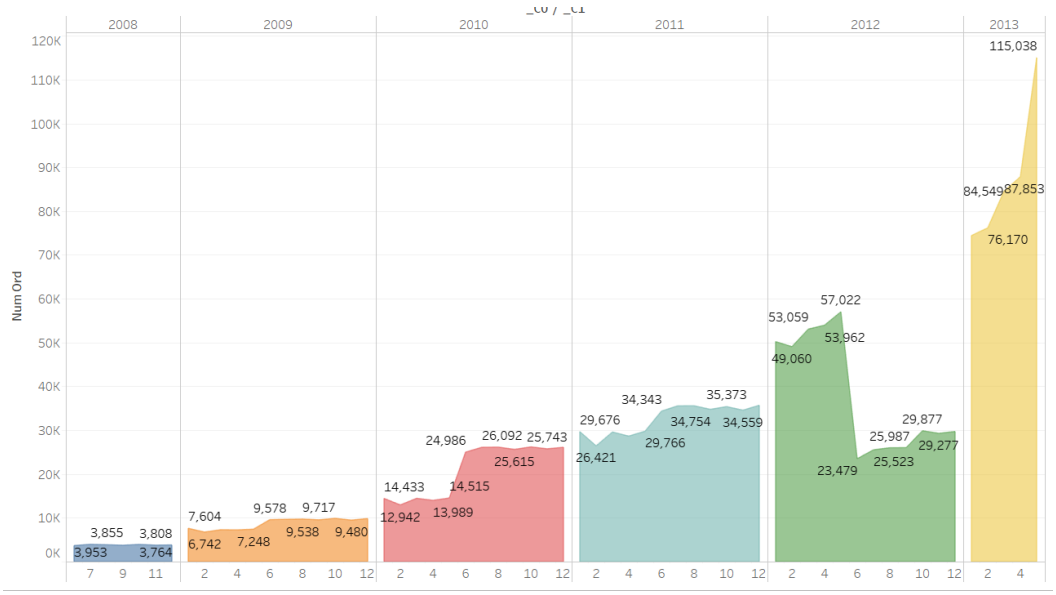
From the above bar graph, we can say that, irrespective of the day, if it's a holiday the frequency of orders is high.

Monthly Shopping Trends

We observed that majority of the orders were placed in the month of May, where "Back to School Offers" and "Summer Sales" are predominant. And the most popular orders placed were Laptops.
Let us look at the complete sales data for these 2 years. Below is the graph showing the trend:



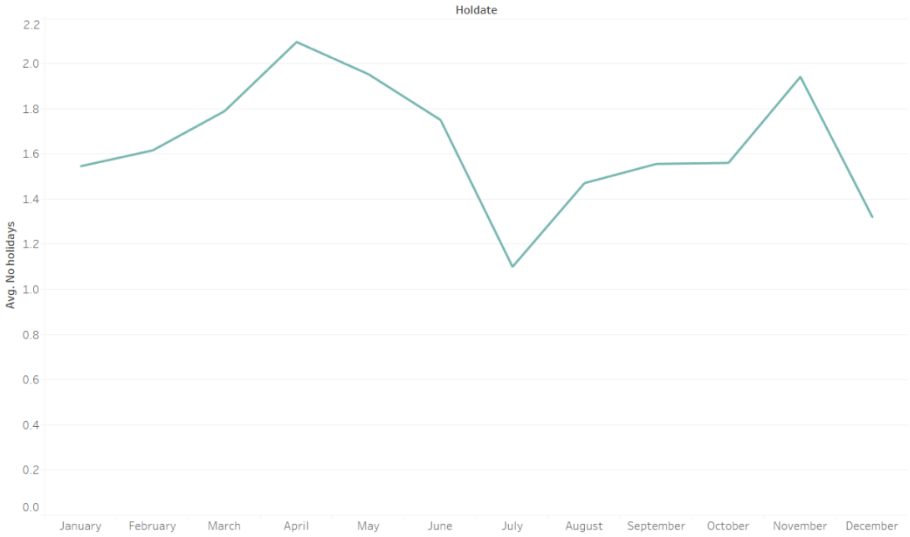
The reason for the pattern observed is possibly that our data has values ranging from June 2008 to May 2013. If we consider monthly and yearly orders, it would give us a clear picture:



We can observe that, as the month of the year increases, sales have increased till 2011, but post that, there is a peak in orders received in May and April.

Peak seasons

Below is the list of total number of holidays per month.



We can observe that, the usual holiday trend peaks in the months of May-April and November-December. The peak in the orders observed in the months of May and April above is evident but the sales in the months of November and December is not as expected.

Potential holidays to increase sales

We can observe that, the number of orders placed during Thanksgiving and Christmas, are less than the average number of orders during other holidays. This implies that the Dual Core is not providing any discounts or promos during these holidays to attract the shoppers. By coming up with strategies on improving sales on such days will increase the profitability.

Therefore, we can come up with the below recommendations:

1. Providing discounts on Christmas, thanksgiving and Black Friday can improve yearend sales.
2. Flash sales on Sundays and Saturdays to improve weekly sales (Improve weekend sales).
3. Winter sales and summer sales can improve seasonal sales trends.