Visualizing Amazon Sales & Ecommerce Trends (May 2017)

Raja Harsha Chinta

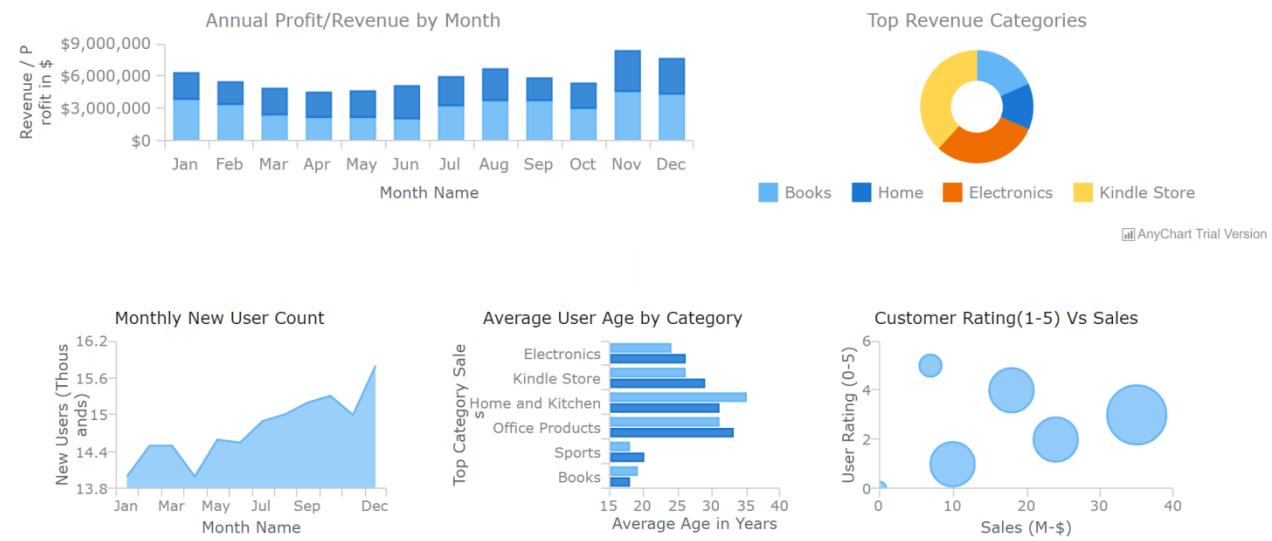

Fig. 1. A snapshot of Sales Analysis Dashboard among the two visualization screens developed.

Abstract— Aim of this paper topic is to analyze Amazon sales data over a period and create a visualization dashboard which helps business users to draw key insights. Often Ecommerce companies assess their business trends in various sales segments and make decisions with the help of results from sophisticated visualization platforms. The author analyzed the data obtained from the Amazon product metadata, transactions dataset to produce some interactive visualization using D3.js, HighCharts.js and AnyCharts.js to show summarized results over the years. SentenTree is a new novel visualization tool used for visualizing unstructured data like tweets is applied in context of user reviews and results are observed.

1 INTRODUCTION

AMAZON is an electronic commerce company and largest internet-based retailer in the world by sales. Amazon sells vast number of products in various categories with more than 100 billion users. The amount of data generated at this scale is very large and by careful analysis of data, many key insights can be drawn. The transactional (sales), feedback (user reviews) data from each geographical user holds underlying personal interest of users and common people's choices.

The key reason for analyzing sales data over several years enables the company to establish sales patterns. This aids while setting up sales budget and drives many business decisions in future. When we develop a marketing plan it is important to analyze strengths and weaknesses of your product or service range. Similarly analyzing a customer feedback is best way to understand and summarize the reach of a product/service in the target market. The usual feedback medium in Ecommerce portals like Amazon is user reviews and ratings.

Raja Harsha Chinta is pursuing masters with Department of Computer Science at Old Dominion University, Email: rchin001@odu.edu

Web or tool based visualizations serves as an effective and robust medium to represent results and perform the critical analysis. Also, they share a great ease of access and offer options to slice and dice a data based on user search criteria. In the project we have chosen to build Web based visualization. As an ecommerce enthusiast, I always wanted to build a simple dashboard which comprehends larger data and help business users to understand it from multiple facets.

In the visualization built for the project, we have tried to project the Amazon sales data in two contexts. One is to visualize the sales and revenue trends over a time span and drill down into monthly time span to elaborate few more metrics like top selling categories and user traffic details. Second one analysis the products/brand level information and evaluates its performance from its sales initiation through average rating across a timeline. Also, a new attempt is made to analyze review data using SentenTree. SentenTree is a novel technique for visualizing the content of unstructured social media text. SentenTree gives people a high-level overview of the most common expressions in a document collection, and allows drilling down to details through interactions. This JavaScript library is integrated and an attempt is made to analyze user reviews. A provision is provided for business user to upload review/unstructured data in a specific format to visualize in the product dashboard.

2 RELATED WORK

There are very good sales dashboard representations available in Tableau, Oracle Business Intelligence Suite, SaaS Data Visualization - Looker. These share good layouts and design options to consider while creating a sales analysis dashboard. Links to these resources have been shared in reference section.

SentenTree is developed by Mengdie Hu from Georgia Institute of Technology with another 2 members. This tool is developed to visualize unstructured social media texts.

3 DATASET

The Amazon sales dataset consists of products metadata, review and ratings available in .json, .csv format to download from http://jmcauley.ucsd.edu/data/amazon/. It consists of basic yet interesting product information like price, sales rank, its category and image URL. Also, basic user information like reviewer name, review text, ratings and review helpfulness index is available.

The datasets we used here consists of files greater than 200 MB each and in nested .json format mostly. We converted the .json files into .csv files using python pandas, ijson libraries. Then we loaded the data into relational database like Hive, MySQL. Initial table data is inspected for data anomalies like null and junk characters. We have built an entity relationship among users, products, sales and reviews & ratings tables. Using SQL we have query tables and extracted data for various visualization scenarios. This has been a really helpful for avoiding calculations, aggregations in web application level and increased performance when visualizing larger datasets.

4 OVERVIEW OF THE VISUALIZATION

Visualizations are built using JavaScript libraries like Anycharts.js, Highcharts.js, D3.js and few Node.js modules. The website template is developed using PHP, HTML, CSS and Bootstrap. All these have been listed in the references section. A video demo and live visualizations are available at

https://vimeo.com/215497695
http://www.cs.odu.edu/~rchinta/InfoVizProj/az_sales.php

4.1 Sales Overview by Year

Sales Overview Dashboard gives us the sales details of Amazon ecommerce business over the pan of years. The entire dashboard is driven by a year variable which helps in drill down the data across all the charts. Various key attribute relations are made to understand different insights.

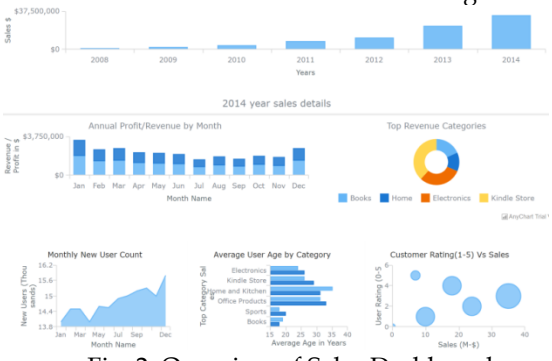

Fig. 2. Overview of Sales Dashboard

Each graph is explained in the following sections.

4.1.1 Column Chart for Yearly Sales

This graph shows the distribution of sales by year for various categories of products sold on Amazon. Years spanning from 1998 to 2014 are displayed on X axis and the total sales in a year is displayed along Y axis. There are filter buttons offered on the top right which enables choosing a smaller time frame of 6 - 8 years. Each column bar supports a drill down option to drill further and allow the user to interlink this main chart to sub charts. A star mark on top a column represents respective year being selected for drill down. This chart can be referred as a main chart for further discussion.

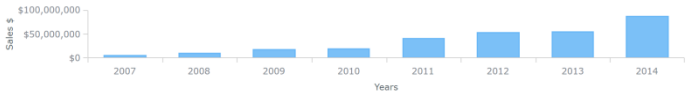

Fig 3.Column chart showing Years Vs Sales data.

Table 1. Idiom Summary Table

Idiom	Column Chart
What: Data	1 Quantitative - Sales in $ 1 Categorical - Year
How: Encode	Column Bar express Sales value vertically, separated by year horizontally
Why: Task	Lookup and compare values

4.1.2 Stacked Bar Charts for Monthly Revenue/Profit

The following stacked bar chart graph shows revenue/profit incurred every month in a year selected in the main chart. Two quantitative values Revenue (light blue) and Profit (dark blue) are indicated in each bar. A tool tip is added for each segment of bar to indicate the values of Revenue, Profit separately in $.

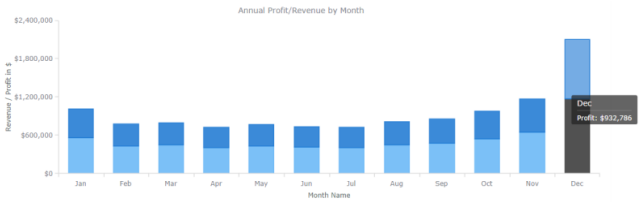

Fig. 4. Stacked bar char with Revenue, Profit again Months of an year.

Also, we can right click on the graph to use Exclude, Include options to remove/add the filters across the data easily.

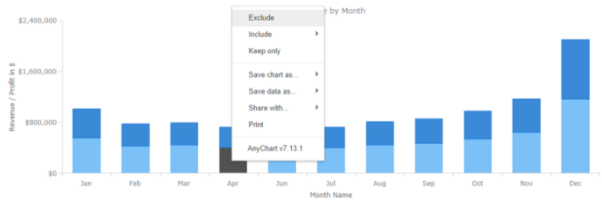

Fig. 5. Data filtering options available in the graph is explained.

Table 2. Idiom Summary Table

Idiom	Stacked Bar Charts
What: Data	1 Quantitative - $ value 2 Categorical key attributes
How: Encode	Bars are encoded with length, separated with different colors.
Why: Task	Observer relationship between revenue and profit over all months in a year.

4.1.3 Donut Chart for Yearly Top Revenue Products

A donut chart is used to represent the top 4 revenue generating product categories. For each year selected in the main chart we are displaying top performing product categories based on their revenue in each different color to differentiate. The share among top 4 is displayed using a % value and a hover on the category will display a tooltip with the revenue values. The names and color coding of product category is displayed below the Donut chart.

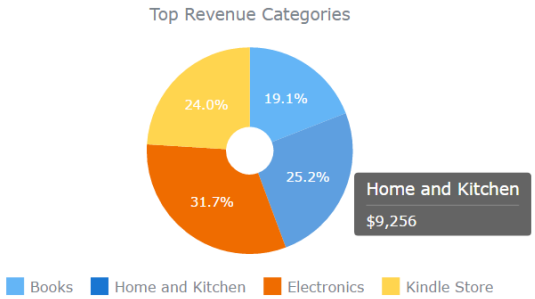

Fig 6. A Donut chart showing top 4 revenue generating products.

Table 3. Idiom Summary Table

Idiom	Donut Chart
What: Data	1 Quantitative - Revenue in $ 1 Categorical - Product Category
How: Encode	Area is marked radically, color encoding is done for various products
Why: Task	Draw comparison among products sales, identify top selling category products.

4.1.4 Area Chart for Monthly New User Tracking

This graph displays number of new users registered in each month after selecting a year from the main chart. X axis is plotted with month names and Y axis denotes the count of users on a scale of 1000 for one unit. An easy tool tip is included to represent the month name and count values on mouse hover. From a common observation across multiple years of data, we see a lot of new users are being added during months of November and December. Also few quarter ends is observed to have low new user registrations. Understanding more deeper into the data from a business perspective the users can draw more conclusion for reduction/increase of a user sales activity can help companies to make active decisions like promotions and deals in that particular months.

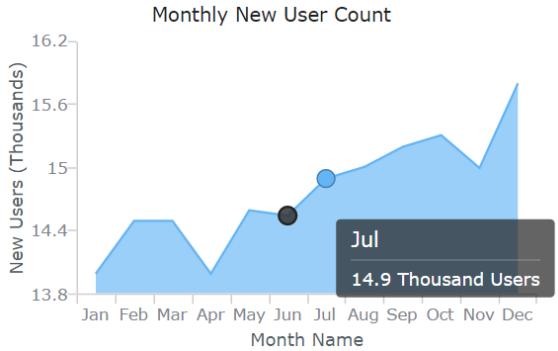

Fig. 7. An Area chart showing new users count over months in a year.

Table 4. Idiom Summary Table

Idiom	Area Chart
What: Data	1 Quantitative – Number of users 1 Ordered Attribute - Year/Month
How: Encode	Co-ordinates for data are marked to form a trend line.
Why: Task	Understand new user/customers joining trend montly.

4.1.5 Horizontal Bar Chart for Age-Gender wise Sales

This graph helps us in understanding the gender wise average age of users who purchased products. This graph indicates clearly which age group people usually purchase a category product and are targeted mostly. Sales wise top 5 product categories are displayed on Y axis and the average age of male or female is displayed on a horizontal bar chart. Two different colors are assigned to differentiate between male and female sales bars. A tool tip is provided to display the value of average age on a mouse hover.

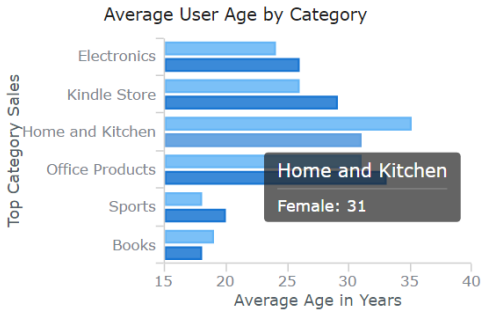

Fig. 8. A Horizontal Bar Chart showing average age of users among the purchased made in top 5 category sales.

Table 5. Idiom Summary Table

Idiom	Bar Chart
What: Data	1 Quantitative – Average Age (Male/Female) 1 Categorical - Year
How: Encode	Bar marks express Sales value vertically, separated by year horizontally
Why: Task	Lookup and compare values

4.1.6 Scatter Plot for Ratings wise Sales, Orders

Orders, sales and ratings are three attributes created for each purchase or transaction at ecommerce web site.

To understand the number of purchases happening at different people satisfactory levels (ratings), we can try to visualize these three entities in a scatter plot. There are two quantitative attributes (Sales, Orders) and a categorical attribute (Ratings, 1-5). Sales value is plotted on X axis on a scale of Million $ for each unit and User rating is plotted on Y axis. Total number of orders represents the bubble size. The combination of data is displayed using a tool tip.

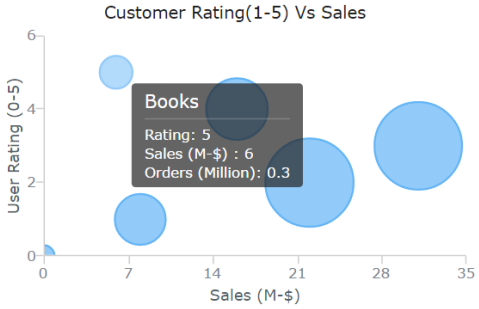

Figure 6. Scatter-Bubble Plot to show Sales, User Rating, Orders.

Table 6. Idiom Summary Table

Idiom	Scatter plots - (Bubble Chart in this case)
What: Data	3 Quantitative - Sales, Orders, Ratings
How: Encode	Express the horizontal, vertical co-ordinates with Sales, Rating Values Orders magnitude is denoted with bubble size.
Why: Task	Find trends of sales, orders for various ratings. To get outliers, distribution, correlation

4.2 Product Search Dashboard

Business users are interested to understand a product or brand performance on a day to day basis and take business decision in investment and marketing. E-commerce portal allows users to rate and review a product. These ratings and reviews can be analyzed for getting an overall user opinion on a product or brand.

Average rating (Quantitative) is calculated for each brand and visualized using a line chart across a time (Ordered attribute) frame in which the product is sold. A time slider accompanies the line chart to help in navigating across date range.

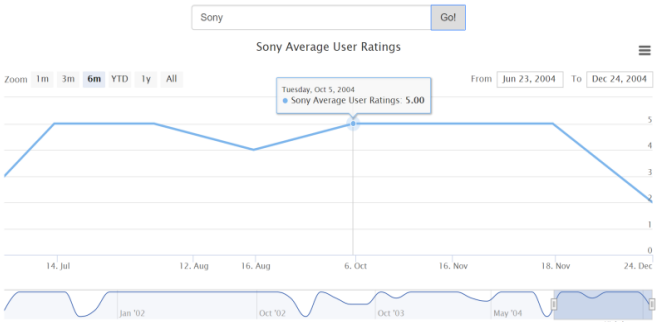

Fig. 7. Line chart to show Average user rating over time.

Table 7. Idiom Summary

Idiom	Line Chart
What: Data	1 Quantitative – Number of users 1 Ordered Attribute - Year/Month
How: Encode	Co-ordinates for data are marked to form a trend line.
Why: Task	Understand average rating for a product and check for any fluctuations in product sales performance in market.

4.3 SentenTree

Review text is generally considered as unstructured data and often cannot be analyzed by regulate tools. Often analysis is done by identifying the common key words and common sentence pattern. SentenTree is a new visualization technique build to visualize unstructured data. It is originally build to analyze social media data like tweets etc. We have tried to apply the same context to analyze review text.

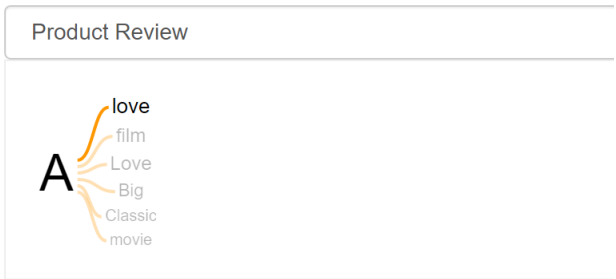

Fig. 8. Visualization of Review Summary Text for a product.

5 WHAT-WHY-HOW FRAMEWORK

The what-why-how framework described by Munzer acts as a basic guideline and framework for visualization any form of data and is discussed in Table 7.

5.1.1 What: Data

The original dataset consists of multiple files in .json, .csv format. These initial files are loaded into relational system and extracted as per the visualization requirements. The final input files used are in .json format and JavaScript objects. There are close to 12 attributes of data essentially used for various charts and is a combination of quantitative and qualitative attributes.

5.1.2 What: Derived

Since we are showing summarized and sales overview information in the dashboard we are having aggregated values are displayed in many charts.

Below are the derived or computed attributes:

5.1.1 Yearly sales, aggregated revenue and profit data
5.1.2 Average age is calculated for one of the chart.
5.1.3 Average ratings on a day are calculated.

5.2 Why: Abstract Tasks

A user can perform actions like Analyze, Search and Query using the visualization. Overview of the entire sales dataset is calculated and visualized. This allows the user to get an overview and helps him in identifying various trends across multiple dimensions in the data. Also, it gives a hint to start analyzing from a point. Sales overview dashboard allows the user to select a year and understand the various metrics across the year. In the product search dashboard, we can search for a brand and see it rating across a time line. Multiple time filters, navigation options are added.

5.3 How: Encode

Our visualizations are grouped into two main categories:

5.3.1 Sales overview dashboard has a main column chart which is aligned with vertical column in blue color and the color changes to grey once a mouse click is made. Also, a start symbol is added to identify the selection. This graph is accompanied with series of graphs beneath which always takes input variables as year.

5.3.2 Monthly Revenue and profit is mapped monthly for a year in a stacked bar chart. A lighter and darker color encoding is chosen for two different bars. A bar selected changes its color to identify the selection. No further data drill down occurs. A tool tips appears over a hover on the bars with respective values.

5.3.3 Similarly, a common color schema and template is chosen to identify the selection of elements in a graph and an identical tooltip is attached to show the attribute values.

5.3.4 User response and ratings are captured using a line chart which is attached closely to a time line slider is encoded with same color schema (light blue and white).

5.4 How: Manipulate, Facet, Reduce

Sales overview charts are interconnected with the main chart with year as a variable. A selection in the main chart level which displays the overall sales in year will automatically change the other 5 visualizations. Data filtering can be done for all the graphs under this section. User can right click and click on include/exclude options to filter data. Coming to the Product overview charts, we have the time slider which controls the visualization time span. Also, additional time interval selection buttons help in selecting exact time frames like 3, 6, 12 months or all at once.

For facet, we have used an overview and detail; juxtapose multiple views and share navigation: main column chart, overview bar chart and pie chart; and popup detail view when selecting or hovering over a certain area in the chart. The user can filter the dataset of interest by selecting a certain year on main chart and drill down data.

Table 7. What-Why-How framework

System	Summary of Visual Idiom Used
What: Data	Flat File: .csv, .json (static) - Multiple quantitative attributes Ex: Price, Orders, User Count - Multiple qualitative attributes Ex: Gender, Rating, Product Category, Year,
What: Derived	- Sales, Revenue, Profit Aggregates - Average User age
Why: Tasks	- **Present** the overview of the sales, reviews dataset. - **Locate** a particular year sales, top performing product - **Compare** yearly, monthly sales and ratings values. - **Compare** the sales, revenue, avg age of users. - **Summarize sales** data distribution across a year, also various other metrics.
How: Encode	- Column chart - Bar chart - Stacked Bar chart - Horizontal chart - Area chart - Line chart with time slider
How: Manipulate	- Select year value - Search for a product - Select product ratings by time
How: Facet	- Overview - Detail - Juxtapose multiple views - Share navigation: synchronize
How: Reduce	Filter all the quantitative attributes in bar, area charts, year filters.

6 CONCLUSION

An interactive visualization is developed to analyze Amazon sales and review dataset. The key insights developed are to display the sales and revenue data over 14 years span. A drilldown feature is created in the year level charts to analyze data further to month level on different charts. Various top performing product sales and user activity is analyzed. All the ratings defined for a brand or a product is displayed on a single line chart for a wider time. A search function is created to choose the product/brand for the ratings time line analysis over years. All the visualizations discussed here are developed using JavaScript libraries like Anycharts.js, Highcharts.js, D3.js and few node.js modules. Relational database are used to initial host the raw data and used for data corrections and filtering. To improve the performance on computing derived column inbuilt aggregating functions with the charting libraries are used effectively. This really increased the performance of the code and generated interactive visualizations very quickly for better user experience.

FINAL THOUGHTS

Visualization becomes a lengthy and complicated task if the data preparation is a part of the project. Preparing semi-structured data or unstructured data requires some scripting ability to covert the data into delimited or object notation for further processing. Also, when we are trying to visualize data entities from different files then we likely need to form a logical relationship like a foreign key. For this we need to load the data into a relational platform. From here, we can actually draw smaller or aggregated result sets which are required for our visualization. With this project experience I have gained good knowledge in handling larger datasets and build a channel to analyze and visualize using big data framework. As part of achieving efficiency and building a refined tool, a design choice is made to use the advanced JavaScript libraries. I have good familiarity with Highcharts, Anycharts JavaScript libraries. Also, I got a good chance to practically apply what I have learnt in the Paper Presentation Topic as part of course i.e. SentenTree tool. Overall, I have gained good confidence in making design choices for selecting the visual idioms to represent particular scenario and data in this course and also, advanced ways to build an interactive visualizations.

ACKNOWLEDGEMENT

The authors wish to thank Julian MacAuley from Univeristy of California San Diago for sharing the dataset links for downloading and initial guidance. Mainly, the author would like to thank Dr. Michele Weigle, Associate Professor, Computer Science, Old Dominion University for teaching us Information Visualization and sharing us many valuable ideas and resources as part of the course.

REFERENCES

[1] http://jmcauley.ucsd.edu/data/amazon/ - Dataset Source.
[2] Introduction to Amazon Customer Product Review Analysis https://www.slideshare.net/sdg31582/amazon-1-47833977/
[3] Anycharts API reference guide: https://api.anychart.com/
[4] Highcharts API reference guide: http://api.highcharts.com/highcharts
[5] SentenTree GitHub URL: https://github.com/twitter/SentenTree
[6] SentenTree IEEE paper: http://ieeexplore.ieee.org/document/7536200/
[7] Bootstrap source: http://getbootstrap.com/
[8] Jquery Auto complete: https://jqueryui.com/autocomplete/

Raja Harsha Chinta got his Bachelors in Engineering from Jawaharlal Nehru Technological University, Kakinada in 2012 and has been working in the Information Technology field for more than 4 years of experience in Databases, Data Warehousing and data Integrations in Ecommerce, Telecom Domains. Currently, he is pursuing a master's degree in Computer Science at Old Dominion University. His current course interests are data visualization, analytics on a larger data scales.