# **INFO 208: Big Data Technologies**

# Mini Project: Amazon Toy Products Dataset Analysis

Divya Puraswani

San Jose State University

Table of Contents

OVERVIEW 3

DATASET STUDIED 5

DESCRIPTION AND STRUCTURE OF THE DATA 6

DATA EXPLORATION/EXPLORATORY Analysis 7

QUESTIONS TO ASK/EXPLORE ABOUT DATASET AFTER EXPLORATION 10

sUMMARY 10

REFERENCES 14

**Overview**

Says Carly Fiorina (former executive, president, and chair of Hewlett-Packard Co.), “The goal is to turn data into information, and information into insight.” We are entering in a world where data is increasing tremendously in all forms like structured, videos, audios etc. and reflects as an integral part in all areas like retail, medical science, banking and financial, automobile, telecommunication etc. Talking about retail industry, as per Aretove’s blog, by the end of 2022 the analytics market for retail is expected to grow globally from $5.1 billion to $8.64 billion. “Success comes from listening to your customers,” Ms. Steiger said. “Your customers are telling you something every time they shop at your store either online or offline, and the data helps tell the customer story. You can discover what they want or don’t want, what products you’re missing that they can get at other places, and more. When you trend and track your customers, you can meet their needs better and in turn grow their loyalty.” [1] Thus, behind every successful organization or business in retail sector, there is a happy and satisfied customer whose purchase life cycle or behavior plays a major role in framing companies’ sales and marketing strategies. Data Science has several use cases like Customer profiling, customer segmentation, market- basket analysis, recommendation systems etc. to predict what customer wants.

Coming from a business class family background, having elementary knowledge in business and managerial field, interest in subjects like Data mining, data warehousing and after exploring the web about data science use-cases in retail; my interest inclined towards working in retail industry and exploring Amazon which is one of the giants in this sector.

Amazon.com, online retailer, manufacturer of [electronic book](https://www.britannica.com/technology/e-book) readers, and Web services provider is an [iconic](https://www.merriam-webster.com/dictionary/iconic) example of [electronic commerce](https://www.britannica.com/technology/e-commerce) and vast [Internet](https://www.britannica.com/technology/Internet) based enterprise that sells books, music, movies, housewares, electronics, toys, and many other goods, either directly or as the middleman between other retailers having its headquarters in Seattle, Washington.[2]

I choose a pre-crawled dataset providing information about toy products, taken as a subset of a bigger [dataset (more than 115k products)](https://www.promptcloud.com/datastock-access-ready-to-use-datasets/?utm_source=kaggle&utm_medium=referral&utm_campaign=amazon) that was created by extracting data from Amazon.com through Kaggle to know more about the customer and getting closer to them by

knowing their preferences or choice for product and also their reviews about products and can those reviews be used for sentiment analysis or content analysis which can help in building a better recommendation engine in future.

The lifecycle of a data science project is an iterative process which can be divided into three phases, as explained in Brian Godsey’s Think like a Data Scientist-

|  |  |  |
| --- | --- | --- |
| **Prepare** | **Build** | **Finish** |
| Set goals | Plan | Deliver |
| Explore | Analyze | Revise |
| Wrangle | Engineer | Wrap-up |
| Assess | Optimize |  |
|  | Execute |  |

(Figure-1 Showing Data science project lifecycle/process in tabular form)

Setting goals is the first step in Data Science Lifecycle-Preparation stage, where the problem statement arises and questions which can be asked from Machine Learning and Business Perspective on playing role of both a data scientist and customer as explained in Brain Godsey’s Think like a Data scientist.

Problem Statement: To find following insights from the data:

1.Machine Learning Perspective/Research

* Is the data easily accessed, extracted and can be trusted?
* Is there anything to know about the data that could be important?
* Do I have enough data?
* Do I have too much data? Will it take too long to process?
* Is there missing data?

2.Business Perspective

* What types of toys are bought most by the customer on Amazon?
* Who is the leading manufacturer?
* Which product has got good reviews and bad reviews?
* Can break down of reviews be done for performing sentiment and content analysis?

**Dataset Studied**

The data set I studied is a pre-crawled dataset, taken as subset of a bigger [dataset (more than 7 million products)](https://www.promptcloud.com/datastock-access-ready-to-use-datasets/?utm_source=am-kaggle&utm_medium=referral) that was created by extracting data from Amazon.com.

Source: URL-https://www.kaggle.com/PromptCloudHQ/toy-products-on-amazon

This dataset was created by PromptCloud's in-house web-crawling service.

**Description and Structure of the Data**

Dataset is in CSV file format and has 10000 rows and following 17 columns-

* uniq\_id-gives unique id of product
* product\_name-name of toy products like Hornby Santa’s Express train Set, Funko POP! Harry Potter Sword etc.
* manufacturer - The item manufacturer, as reported on Amazon. Some common "manufacturers", like Disney, actually outsource their assembly line.
* price-gives price of product.
* number\_available\_in\_stock-tells the quantity available in stock and its status new etc.
* number\_of\_reviews-number of reviews given by customer on the product.
* number\_of\_answered\_questions - Amazon includes a Question and Answer service on all or most of its products. This field is a count of how many questions that were asked actually got answered.
* average\_review\_rating-gives average rating out of 5
* amazon\_category\_and\_sub\_category - A tree-based, >>-delimited categorization for the item in question.
* customers\_who\_bought\_this\_item\_also\_bought - References to other items that similar users bought. This is a recommendation engine component that played a big role in making Amazon popular initially.
* description-gives description of the product.
* product\_information-gives technical details like weight, recommended age etc. about the toy product
* product\_description-gives description of the product.
* items\_customers\_buy\_after\_viewing\_this\_item-it is similar to column customers\_who\_bought\_this\_item\_also\_bought
* customer\_questions\_and\_answers - A string entry with all of the product's JSON question and answer pairs.
* customer\_reviews - A string entry with all of the product's JSON reviews.
* sellers - A string entry with all of the product's JSON seller information (many products on Amazon are sold by third parties).

**Data Exploration/Exploratory Analysis**

After setting goals and plan, exploring datasets forms an important pillar of data science process which helps in detailed understanding of data and allows to stay close with data as well by cleaning, wrangling/munging (transforming data) and snooping (poking around data). There are several tools available for the same. In this project, I used extensive Python libraries like Pandas, NumPy, Matplotlib and Seaborn.

There are different formats in which datasets are available (XLS, TXT, CSV, JSON). Python has several libraries which makes easy to load data from any source. Amazon toy products data is in CSV format which is loaded using read\_csv function. After having read the dataset using functions- head() and tail() following inferences were drawn-

* Missing values are present in few columns like number\_available\_in\_stock.
* Column product\_description and description have same values, so description can be dropped and product\_description can also be merged with product\_information.
* amazon\_category\_and\_sub\_category is tree like, can be divided into subcategories.

On gaining more information about dtypes, non-null values using info() function following Inferences were drawn-

* Dataset has 1 numerical column and 16 columns are stored as object.
* uniq\_id and product\_name does not have any null/missing values and rest all other columns have null/missing values.
* price column should be float but is stored as object.
* uniq\_id can be renamed to unique\_id
* Memory storage is 1.3+ MB

Checking missing values in the dataset using isnull().any() and also count of null values present in particular column using isnull().sum() inferred that only uniq\_id and product  
\_name does not contain null values. Maximum null values about 90.86% are present in customer\_questions\_and\_answers column which is dropped. Columns product\_description and description are merged, duplicates are removed which results in no null values and results in new column description\_of\_product. Null values in amazon\_category\_and\_sub\_category are replaced with “ ”.

Splitting of column “number\_available\_in\_stock” into quantity\_available and quantity\_available\_status and Changing datatype of quantity\_available from object to float, changing datatype of column number\_of\_reviews from object to numerical using helper function(those functions which makes programs easy to compute and read) and map function, Changing datatype of price from string/object to numerical and filling null values using mean price of data, Changing datatype of column average\_review\_rating from object to numerical using helper function and map function. Renaming column name from uniq\_id to unique\_id.

On analyzing for price column which is numerical, using describe() method gives descriptive statistics like minimum, maximum and mean price. And than plotting histogram gives the visualization.

A screenshot of a cell phone

Description automatically generated

(Figure-2 Showing descriptive statistics for price column in dataset)

A screenshot of a social media post

Description automatically generated

(Figure-3 Showing histogram for price column in dataset)

On analyzing for product\_name, manufacturer which is the non-numerical value and calculating descriptive statistics it gives top, frequency of the column.

A screenshot of a cell phone

Description automatically generated

(Figure-4 Showing descriptive statistics for product\_name column in dataset)

On analyzing amazon\_category\_and\_sub\_category and splitting it gives the primary category and secondary categories under which toys are popular.

**Questions to ask/explore about dataset after exploration**

Business Perspective

* What are major categories and sub categories under which toy products are popular?
* Does number\_of\_reviews relate to selling of products?
* What is the relation between number\_of\_reviews given and average\_review\_ratings?

**Summary**

After exploring dataset, I was able to get closer to the data and answer questions-

1. Machine Learning Perspective/Research

* Is the data easily accessed, extracted and can be trusted?

Answer-Yes, the data is easily accessible, extractable and can be trusted as it was extracted from amzon.com.

* Is there anything to know about the data that could be important?

Answer-No, sufficient information was available.

* Do I have enough data?

Answer-There were 10000 rows and 17 columns which are just 0.14% ((10000/7000000) multiply by 100)) of the data which is not sufficient and difficult to infer accurate results.

* Do I have too much data? Will it take too long to process?

Answer-No data is not too much it is small and took less time to process.

* Is there missing data?

Answer-Yes, missing values or null values were present in several columns.

2.Business Perspective

* What types of toys are bought most by the customer on Amazon?

Answer-Popular Toy brands in Amazon are- Zoo Animal Hand Sock Glove Finger Puppets Sack Plush Toy Cow, Polyhedral Dice which have high frequency and than Melissa & Doug Stacking Construction Vehicles, The Trash Pack Sewer Truck

* Who is the leading manufacturer?

Answer-Lego is the leading manufacturer for toy products in amazon.

A screenshot of a social media post

Description automatically generated

(Figure-5 Showing leading manufacturer of toys)

* What are major categories and sub categories under which toy products are popular?

Answer-Hobbies being Major Category under which toys are bought. Toy, Vehicles & Accessories being top among sub category for purchasing toys. Scaled Models are preferred more to buy. Under Scaled Models, Vehicles are popular over Trains, Locomotives etc. to be bought.

A screenshot of a cell phone

Description automatically generated

(Figure-6 Showing major category under which toys are bought)

* Does number\_of\_reviews relate to selling of products?

Answer-Yes, generally products with higher reviews are sold more.

* What is the relation between number\_of\_reviews given and average\_review\_ratings?

Answer-Scatter plot between number\_of\_reviews and average\_review\_rating having ratings maximum between 4-5 out of 5 which indicates that products sold are good in quality and preferred for purchase by customers.

* Can break down of reviews be done for performing sentiment and content analysis?

Answer-Yes, reviews can be broken down for sentiment and content analysis purpose and can also be used in future models for recommendation engines.

This project helped me to understand that data plays a very crucial part in company’s growth and to extract meaningful information from data by cleaning, exploring and applying several techniques is yet another important part which leads to forming the pillar for company’s survival and future growth. It helped me in understanding how to keep data ready by using python libraries and functions so that next process of building can be executed.

Extracting true business value from cleaned data requires unique combination of technical skills, mathematical know-how, storytelling, and intuition. But making that data meaningful requires lot of time.

It is imperative that the data scientist understand what is happening to the data from inception to model to business decision. Thus, A data scientist needs to stay in front of the curve in research, as well as understand what technology to apply when. [3]

Brian Godsey’s Think like a Data Scientist talks about Priorities which can set as gaining knowledge or awareness of the domain working in. Technical Skills and then forming a strong opinion on the basis of knowledge and technical acumen.

As technology continues to advance, one thing is certain: Data Science still has a lot to offer in the world of retail! [4]

# **Works Cited**

(n.d.)1. Retrieved from https://www.retaildive.com/ex/mobilecommercedaily/5-businesses-that-benefit-from-data-science

(n.d.)2. Retrieved from https://www.britannica.com/topic/Amazoncom

(n.d.)3. Retrieved from https://www.cio.com/article/3263790/data-science/the-essential-skills-and-traits-of-an- expert-data-scientist.html

(n.d.)4. Retrieved from https://www.margo-group.com/en/news/data-science-applied-to-the-retail-industry-10-essential-use-cases/

Godsey, B. (2017). Think Like a Data Scientist: Tackle the Data Science Process Step-By-Step. Manning Publications.