Capstone Project Documentation

Amazon Product Review Analysis

Aayush Goel
Abhishek Janjal
Syed Mohammed

Introduction

In recent years, there has been a significant amount of research conducted on textual data sources to extract useful information and gain insights into the opinions of customers associated with large businesses. One such application of Natural Language Processing is Sentiment Analysis, which employs machine algorithms to predict the sentiments of customers based on their reviews posted on various platforms. These sentiments can be further analyzed using Time Series Analysis to forecast the likely sentiments/reviews to be received based on trends observed in the available data. Our project aims to optimize inventory and predict demand for video games, toys, and other game products on Amazon.

Data Description

Review Data

- reviewer Id - ID of the reviewer, e.g. A2SUAM1J3GNN3B
- asin - ID of the product, e.g. 0000013714
- reviewer Name - name of the reviewer
- vote - helpful votes of the review
- style - a disctionary of the product metadata, e.g., "Format" is "Hardcover"
- reviewText - text of the review
- overall - rating of the product
- summary - summary of the review
- unixReviewTime - time of the review (unix time)
- reviewTime - time of the review (raw)

- image - images that users post after they have received the product

Meta Data

- asin - ID of the product, e.g. 0000031852
- title - name of the product
- feature - bullet-point format features of the product
- description - description of the product
- price - price in US dollars (at time of crawl)
- imageURL - url of the product image
- imageURL - url of the high-resolution product image
- related - related products (also bought, also viewed, bought together, buy after viewing)
- salesRank - sales rank information
- brand - brand name
- categories - list of categories the product belongs to
- tech1 - the first technical detail table of the product
- tech2 - the second technical detail table of the product
- similar - similar product table

We have taken Video games,Toys & games category from Amazon dataset page and merged both review and meta data as one after and have taken 1,00,000 records for the analysis.

Importing data

When we first received the data, we encountered difficulty loading the large dataset using pandas library. Therefore, we opted for the polars library, which is known for its faster data loading capabilities. Once loaded, we converted the data to pandas for further processing.

Next, we uploaded the dataset to a local SQL database for further data manipulation. We performed initial pre-processing tasks such as handling missing values and empty cells, merging review data, and creating a target column to manage the data effectively. This ensured that the data was readily available for analysis at any given time.

Missing value Treatment

To ensure data accuracy and avoid incorrect analysis, we removed unnecessary columns that contained empty cells amounting to more than 30% of the data. We decided against filling these missing values as it may lead to skewed results.

Moreover, we dropped the records with missing values as the data was in string format and imputing values using the mode was not feasible. By removing these records, we ensured that the remaining data was reliable and suitable for further analysis.

Data Cleaning

1)First we found the duplicate records from the data.

2)The "reviewtime" column was converted to date-time data type

3)A sample data was taken of 50,000 rows for the sentiment analysis

4)The data was cleaned with the help of re(RegEx) library

5)The library nlppreprocessing to remove stopwords, independent numbers, punctuation's, html-tags, replace words(doesn't --> does not)

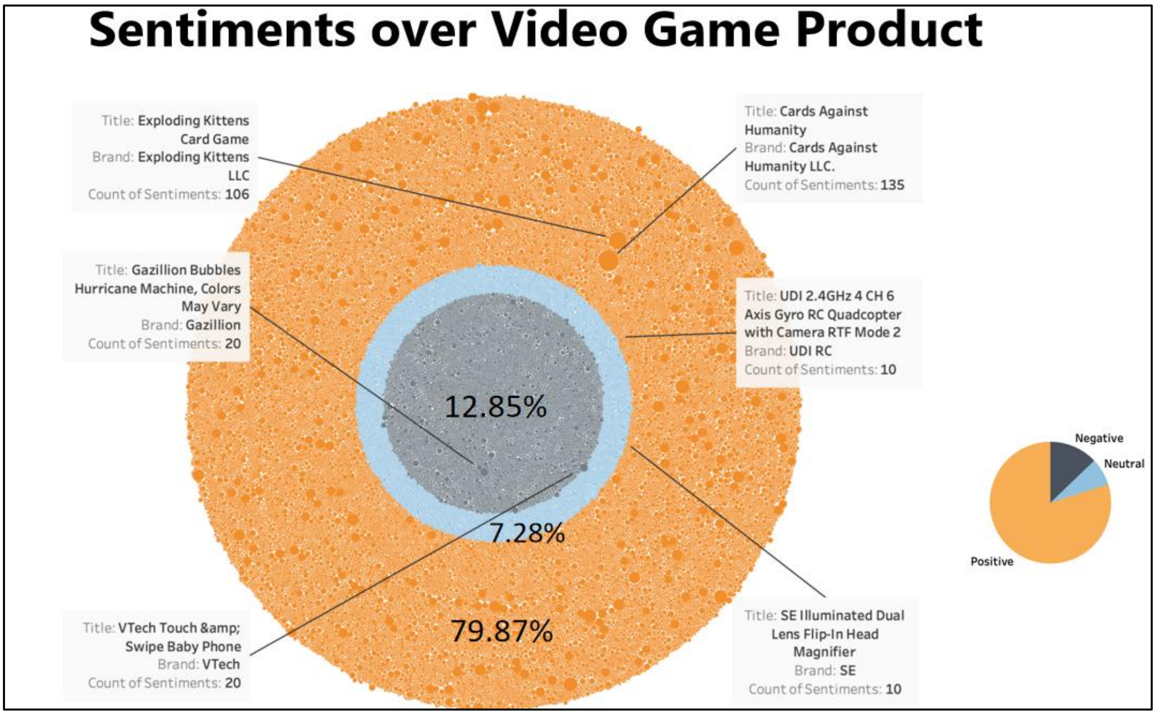
6)The nouns were removed using nltk.tag.post_tag

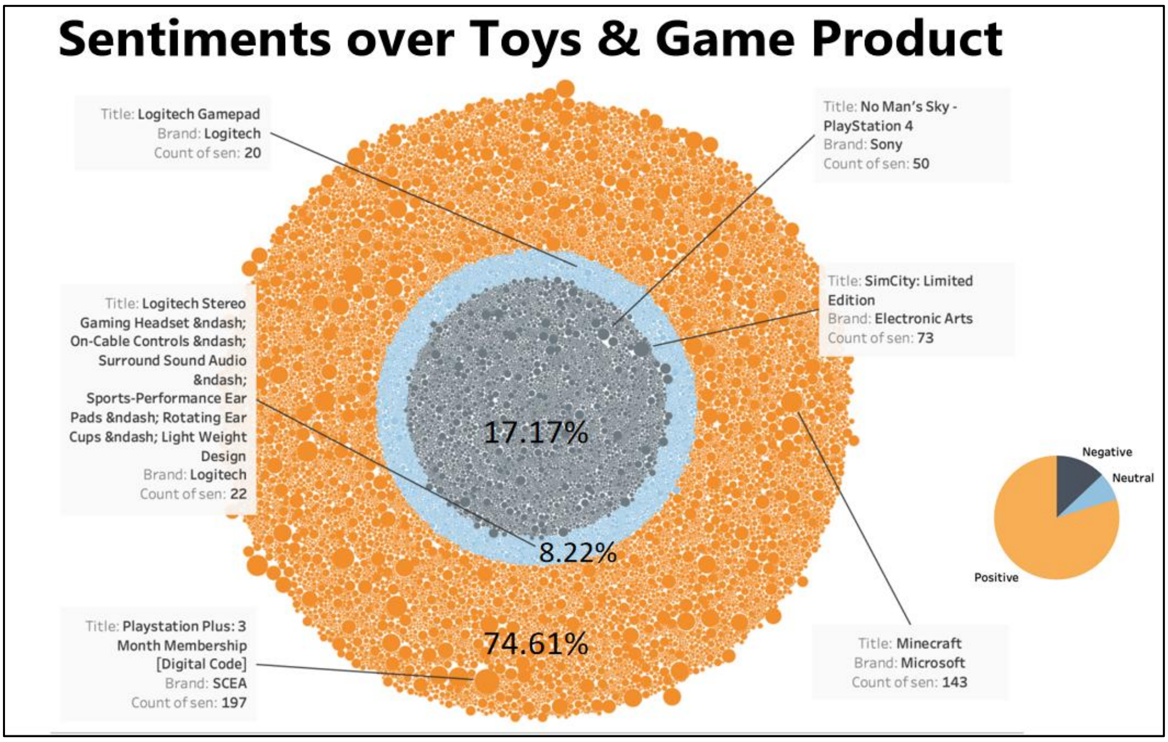
7)Text was lemmatized using spacy library

Sentiment Analysis

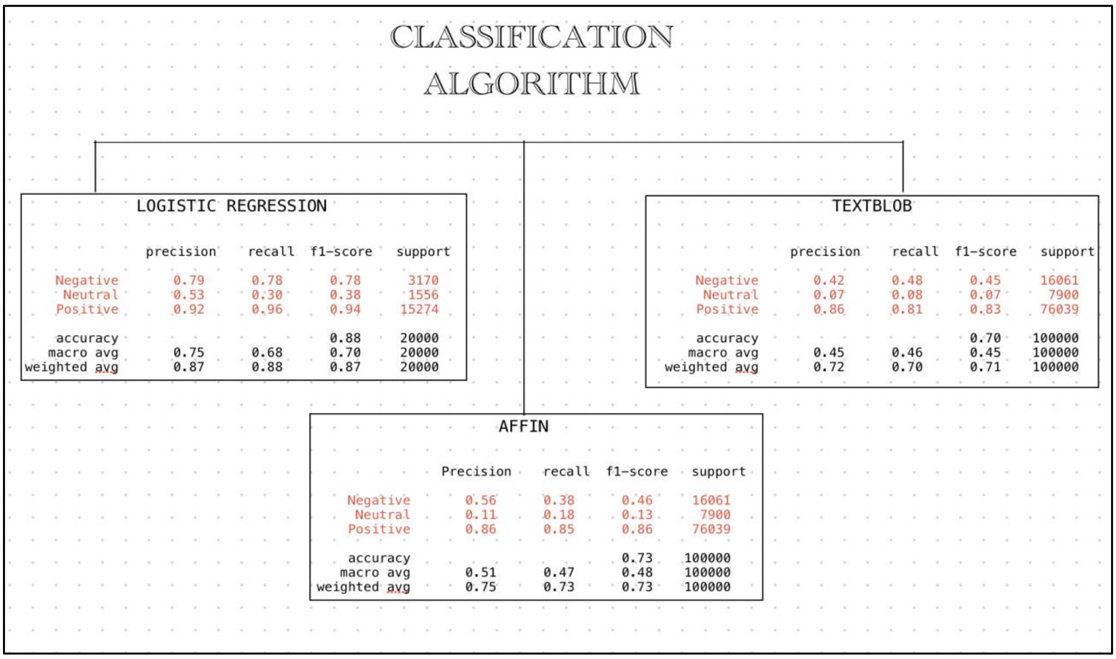
After performing data cleaning, we conducted sentiment analysis on the review text column using pretrained models such as TextBlob and Afinn. Based on the polarity scores obtained from these models, we categorized the reviews as positive, negative, or neutral.

We further utilized the Multiclass Logistic Regression, a supervised machine learning algorithm, to predict sentiment using the overall rating as labeled data. To do so, we split the data into train and test datasets and fitted the Multiclass Logistic Regression model. This enabled us to predict future sentiments based on the trained model.





To compare the effectiveness of these models, we evaluated them using classification evaluation metrics. The classification reports for the models are provided below.


CLASSIFICATION ALGORITHM

LOGISTIC REGRESSION

```
              precision    recall  f1-score   support

    Negative       0.79      0.78      0.78      3170
     Neutral       0.53      0.30      0.38      1556
    Positive       0.92      0.96      0.94     15274

    accuracy                           0.88     20000
   macro avg       0.75      0.68      0.70     20000
weighted avg       0.87      0.88      0.87     20000
```

TEXTBLOB

```
              precision    recall  f1-score   support

    Negative       0.42      0.48      0.45     16061
     Neutral       0.07      0.08      0.07      7900
    Positive       0.86      0.81      0.83     76039

    accuracy                           0.70    100000
   macro avg       0.45      0.46      0.45    100000
weighted avg       0.72      0.70      0.71    100000
```

AFFIN

```
              Precision    recall  f1-score   support

    Negative       0.56      0.38      0.46     16061
     Neutral       0.11      0.18      0.13      7900
    Positive       0.86      0.85      0.86     76039

    accuracy                           0.73    100000
   macro avg       0.51      0.47      0.48    100000
weighted avg       0.75      0.73      0.73    100000
```

After comparing the accuracy scores of the models, we concluded that the Multiclass Logistic Regression model had a higher accuracy compared to the TextBlob and Afinn models. Therefore, we chose the Multiclass Logistic Regression model as our primary sentiment prediction model.

Time Series Analysis

We have taken reviewtime column and the polarity column that we obtained from sentiment analysis for prediction of future sentiments.we have set reviewtime as our index column in proper datetime format and sorted the index column.Then we have resampled the data with respect to months and kept the data ready for time series analysis.Then we seasonal decompose the data to check for trends, seasonality and noise, we observed seasonality in our data and it was not stationary.To confirm the stationarity of data we then checked with adfuller test which also returned the same results.So we decided to employ SARIMA model for forecasting beacuse of the seasonality present and the lack of stationarity of data.we got the respective non seasonal parameters (p,d,q) & seasonal parameters(P,D,Q,m) from pacf and acf plots.we then got the best model which was having the lowest AIC value and the model has been used for forecasting future sentiments.

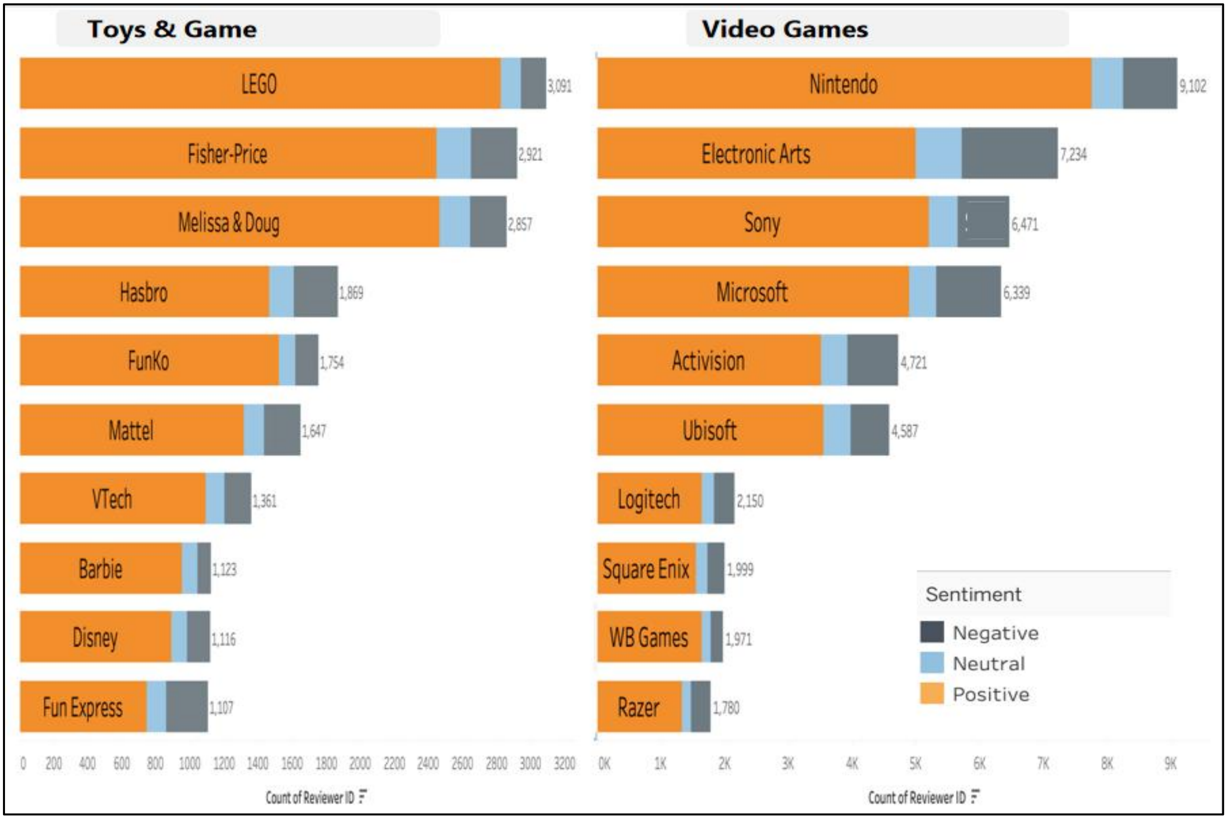
Deployment of model through web application

Flask web application framework was used to deploy our model where we have created an Interface for the end users to write their reviews and get their response as positive, negative or neutral sentiments which is predicted by our model in the backend.

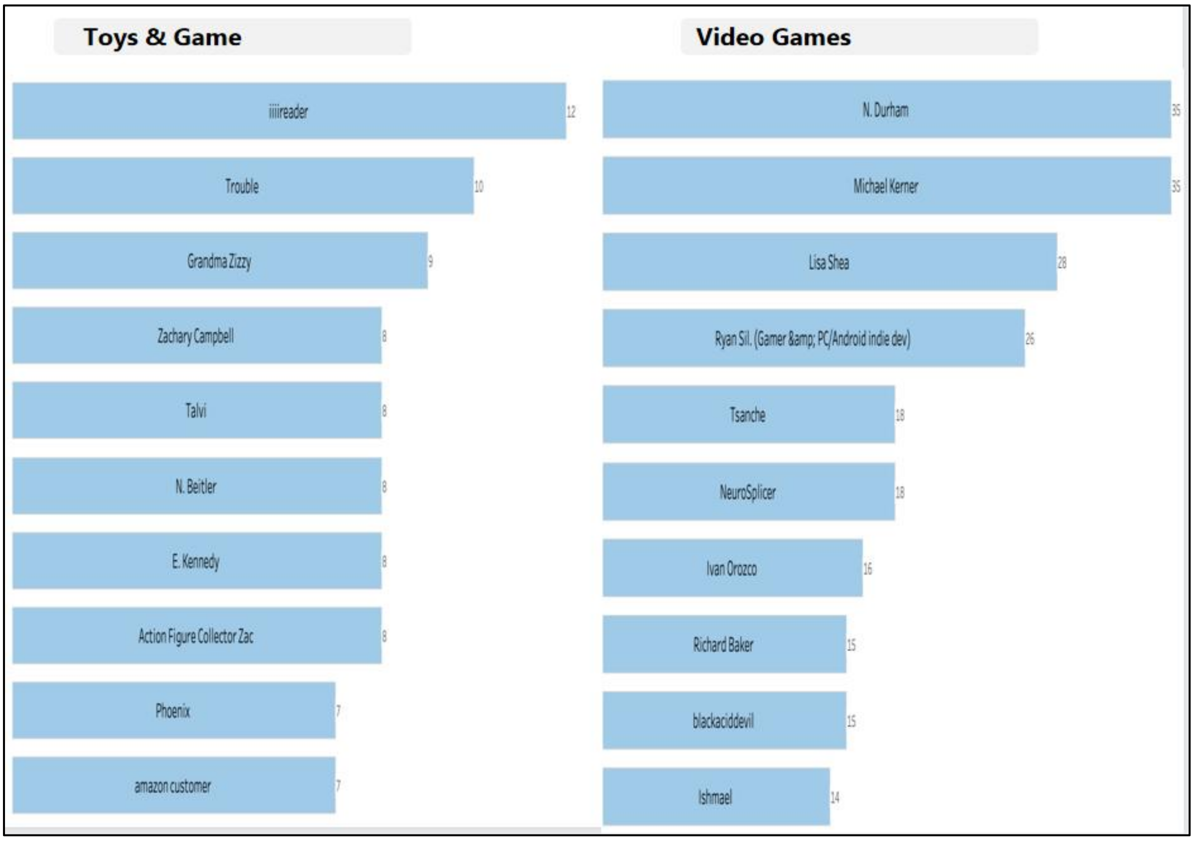
Exploratory Data Analysis

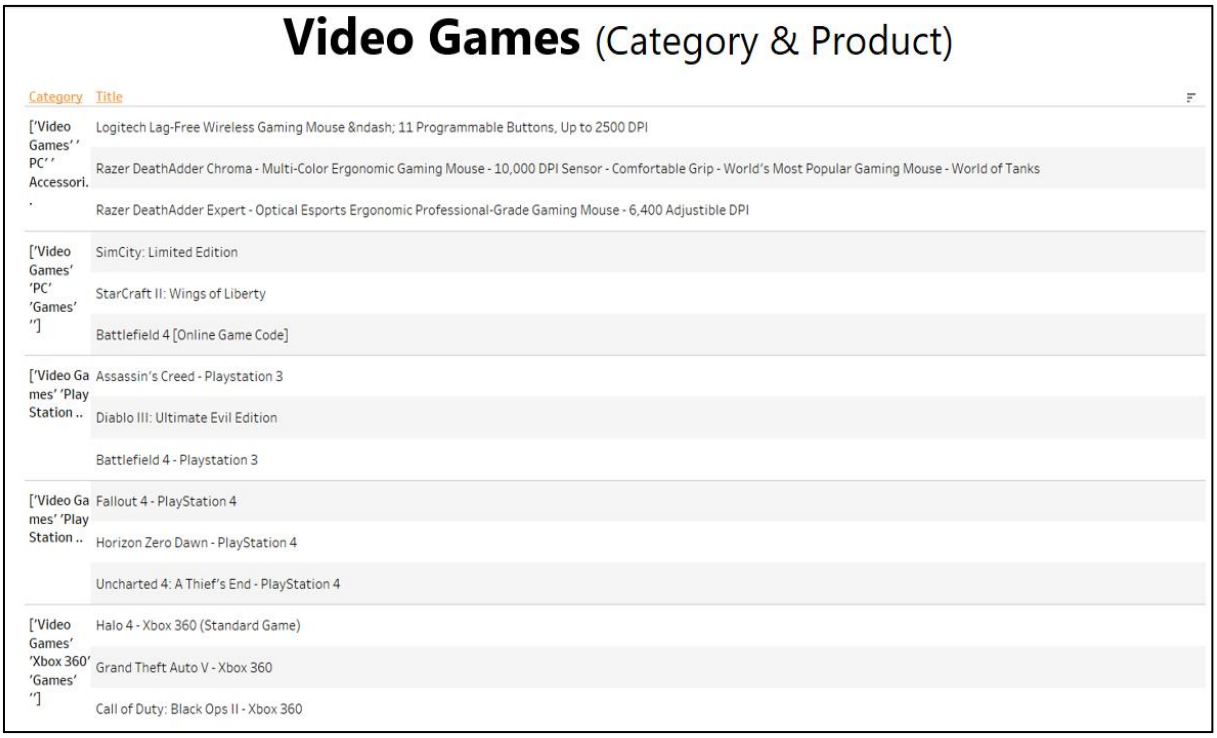
Exploratory Data Analysis(EDA) was done throughout the making of the project(constantly updated) using Tableau and we have also created an interactive dashboard which solves the problem of Inventory optimisation.

Top 10 Brands



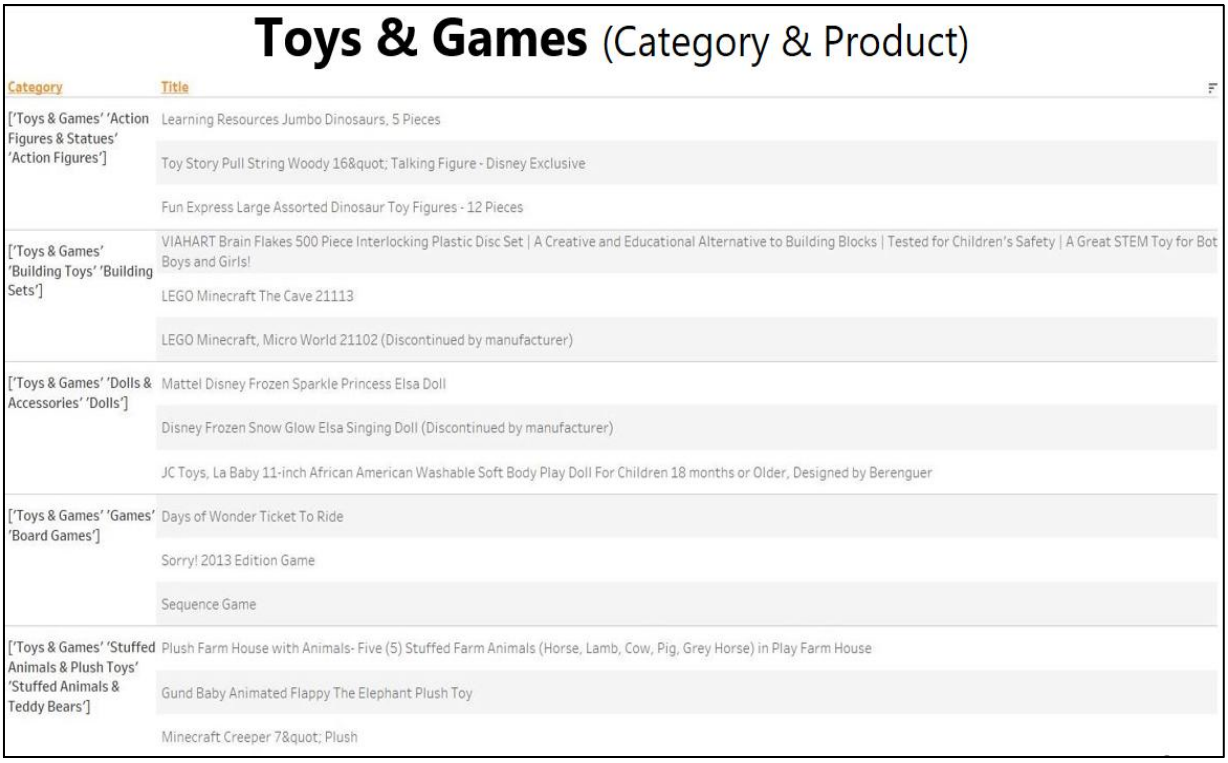
Regular customers





Video Games (Category & Product)

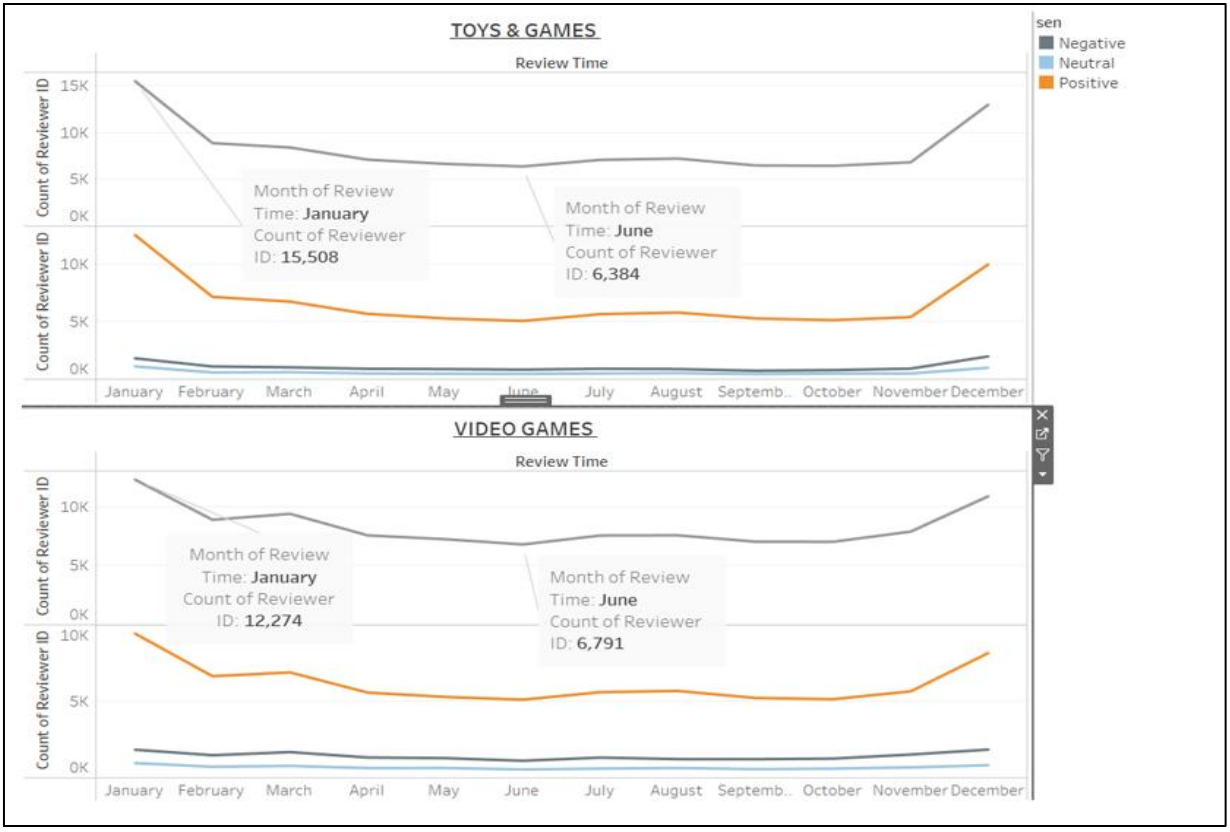
Category	Title
['Video Games'' PC'' Accessori. .	Logitech Lag-Free Wireless Gaming Mouse &ndash; 11 Programmable Buttons, Up to 2500 DPI
	Razer DeathAdder Chroma - Multi-Color Ergonomic Gaming Mouse - 10,000 DPI Sensor - Comfortable Grip - World's Most Popular Gaming Mouse - World of Tanks
	Razer DeathAdder Expert - Optical Esports Ergonomic Professional-Grade Gaming Mouse - 6,400 Adjustible DPI
['Video Games' 'PC' 'Games' '']	SimCity: Limited Edition
	StarCraft II: Wings of Liberty
	Battlefield 4 [Online Game Code]
['Video Games' 'Play Station ..	Assassin's Creed - Playstation 3
	Diablo III: Ultimate Evil Edition
	Battlefield 4 - Playstation 3
['Video Games' 'Play Station ..	Fallout 4 - PlayStation 4
	Horizon Zero Dawn - PlayStation 4
	Uncharted 4: A Thief's End - PlayStation 4
['Video Games' 'Xbox 360' 'Games' '']	Halo 4 - Xbox 360 (Standard Game)
	Grand Theft Auto V - Xbox 360
	Call of Duty: Black Ops II - Xbox 360



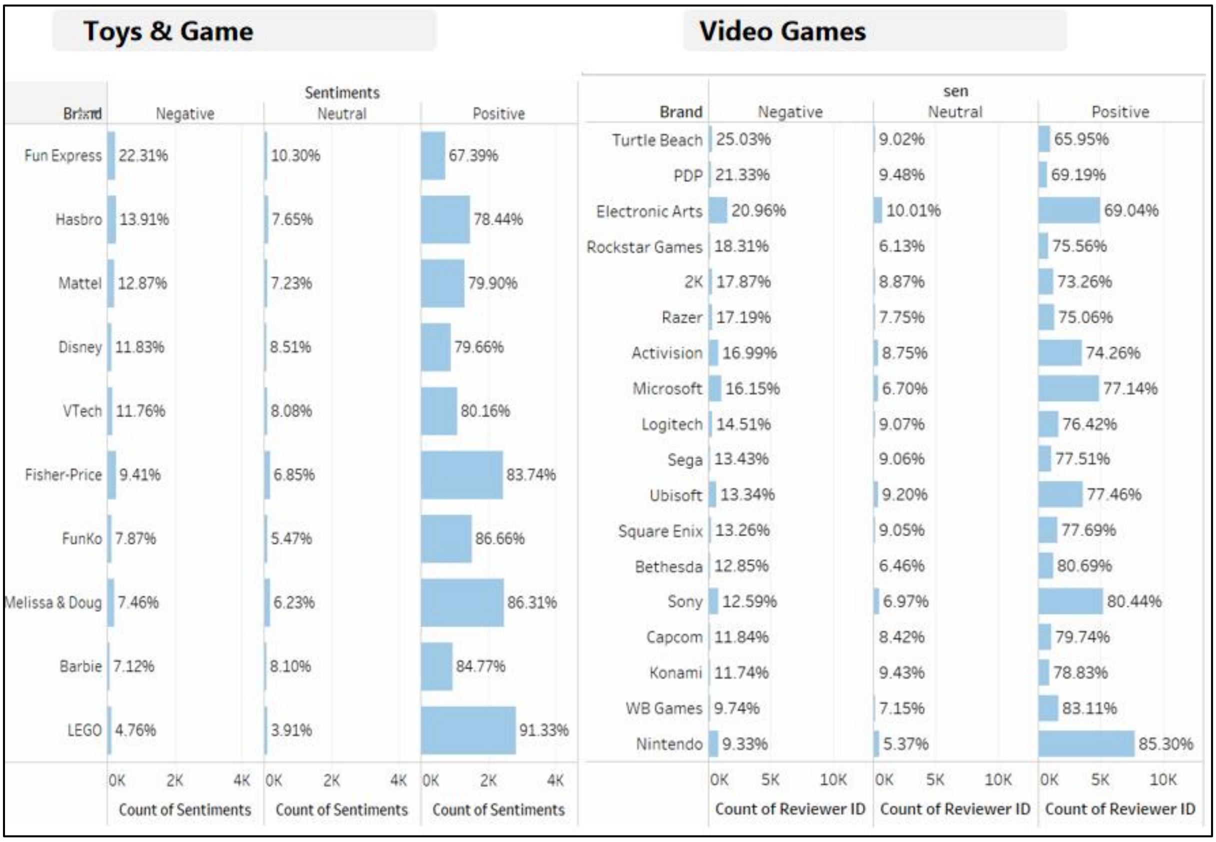
Toys & Games (Category & Product)

Category	Title
['Toys & Games' 'Action Figures & Statues' 'Action Figures']	Learning Resources Jumbo Dinosaurs, 5 Pieces
	Toy Story Pull String Woody 16&quot; Talking Figure - Disney Exclusive
	Fun Express Large Assorted Dinosaur Toy Figures - 12 Pieces
['Toys & Games' 'Building Toys' 'Building Sets']	VIAHART Brain Flakes 500 Piece Interlocking Plastic Disc Set | A Creative and Educational Alternative to Building Blocks | Tested for Children's Safety | A Great STEM Toy for Bot Boys and Girls!
	LEGO Minecraft The Cave 21113
	LEGO Minecraft, Micro World 21102 (Discontinued by manufacturer)
['Toys & Games' 'Dolls & Accessories' 'Dolls']	Mattel Disney Frozen Sparkle Princess Elsa Doll
	Disney Frozen Snow Glow Elsa Singing Doll (Discontinued by manufacturer)
	JC Toys, La Baby 11-inch African American Washable Soft Body Play Doll For Children 18 months or Older, Designed by Berenguer
['Toys & Games' 'Games' 'Board Games']	Days of Wonder Ticket To Ride
	Sorry! 2013 Edition Game
	Sequence Game
['Toys & Games' 'Stuffed Animals & Plush Toys' 'Stuffed Animals & Teddy Bears']	Plush Farm House with Animals- Five (5) Stuffed Farm Animals (Horse, Lamb, Cow, Pig, Grey Horse) in Play Farm House
	Gund Baby Animated Flappy The Elephant Plush Toy
	Minecraft Creeper 7&quot; Plush

Seasonality of data



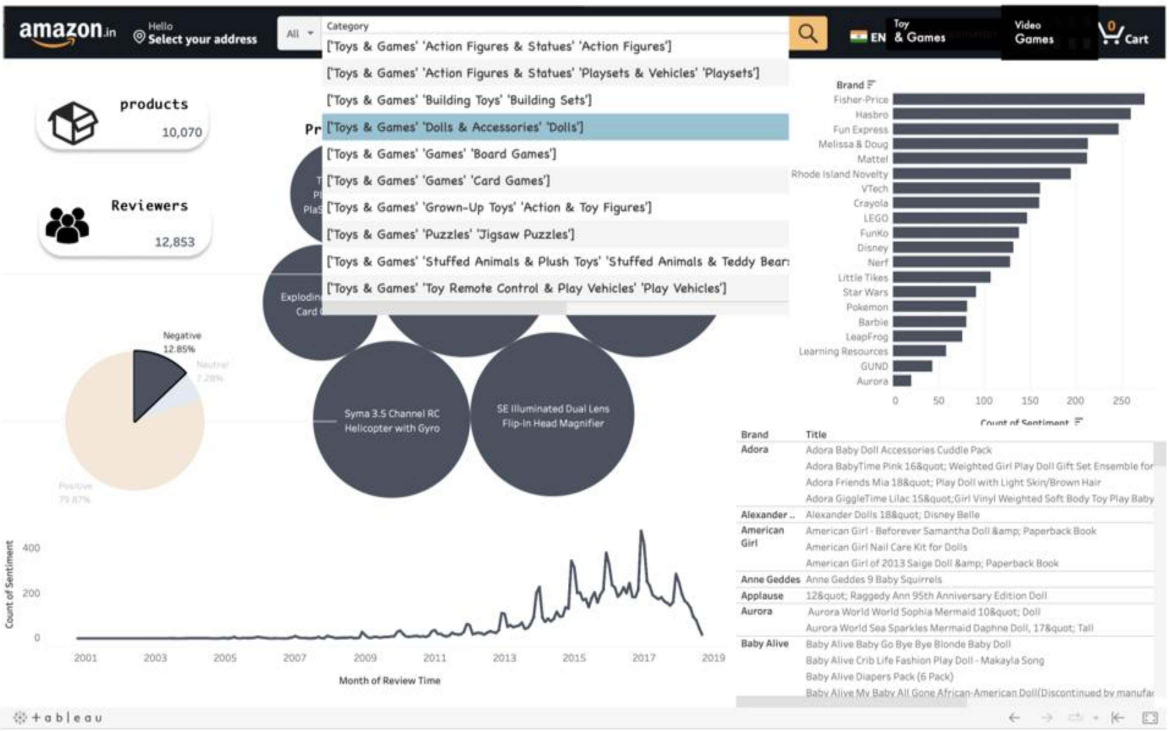
Bad Influencers



Toys & Game

Video Games

Optimised Inventory DashBoard



amazon.in
Hello
Select your address
All
Category
['Toys & Games' 'Action Figures & Statues' 'Action Figures']
['Toys & Games' 'Action Figures & Statues' 'Playsets & Vehicles' 'Playsets']
['Toys & Games' 'Building Toys' 'Building Sets']
['Toys & Games' 'Dolls & Accessories' 'Dolls']
['Toys & Games' 'Games' 'Board Games']
['Toys & Games' 'Games' 'Card Games']
['Toys & Games' 'Grown-Up Toys' 'Action & Toy Figures']
['Toys & Games' 'Puzzles' 'Jigsaw Puzzles']
['Toys & Games' 'Stuffed Animals & Plush Toys' 'Stuffed Animals & Teddy Bear
['Toys & Games' 'Toy Remote Control & Play Vehicles' 'Play Vehicles']
EN
Toy & Games
Video Games
0 Cart
products
10,070
Reviewers
12,853
Pr
Explodin Card
Syma 3.5 Channel RC Helicopter with Gyro
SE Illuminated Dual Lens Flip-In Head Magnifier
Negative 12.85%
Neutral 7.28%
Positive 79.87%
Brand
Fisher-Price
Hasbro
Fun Express
Melissa & Doug
Mattel
Rhode Island Novelty
VTech
Crayola
LEGO
FunKo
Disney
Nerf
Little Tikes
Star Wars
Pokemon
Barbie
LeapFrog
Learning Resources
GUND
Aurora
0 50 100 150 200 250
Count of Sentiment
Count of Sentiment
400
200
0
2001 2003 2005 2007 2009 2011 2013 2015 2017 2019
Month of Review Time
+ableau

Brand	Title
Adora	Adora Baby Doll Accessories Cuddle Pack Adora BabyTime Pink 16&quot; Weighted Girl Play Doll Gift Set Ensemble for Adora Friends Mia 18&quot; Play Doll with Light Skin/Brown Hair Adora GiggleTime Lilac 15&quot;Girl Vinyl Weighted Soft Body Toy Play Baby
Alexander ..	Alexander Dolls 18&quot; Disney Belle
American Girl	American Girl - Beforever Samantha Doll &amp; Paperback Book American Girl Nail Care Kit for Dolls American Girl of 2013 Saige Doll &amp; Paperback Book
Anne Geddes	Anne Geddes 9 Baby Squirrels
Applause	12&quot; Raggedy Ann 95th Anniversary Edition Doll
Aurora	Aurora World World Sophia Mermaid 10&quot; Doll Aurora World Sea Sparkles Mermaid Daphne Doll, 17&quot; Tall
Baby Alive	Baby Alive Baby Go Bye Bye Blonde Baby Doll Baby Alive Crib Life Fashion Play Doll - Makayla Song Baby Alive Diapers Pack (6 Pack) Baby Alive My Baby All Gone African-American Doll(Discontinued by manufa

Technologies used

We utilized SQL for data management, Tableau for visualization, and Python for data processing and analysis.

Libraries used

- pandas
- mysql
- sqlalchemy
- spacy
- nltk
- scikit-learn
- matplotlib
- seaborn
- statsmodels
- Itertools

New Innovative libraries used

- polars
- Textblob
- Affinn
- nlppreprocess
- Pipeline
- Column Transformers
- pickle

- word cloud
- Flask

Conclusion

In our analysis we discovered that that less than 20 % of the customers are dissatisfied so only minor change are required to improve the user satisfaction towards our products

Suggestions

- Improve Sales : Improve the sales on summer months by organizing a few sales to pique the interest of the users in the products of the two main categories (February- June) since the sales are uniform throughout the period.
- Restock : Before winter season the products(especially the popular products) should be restocked as the sales in winters increase dramatically
- Improve user Satisfaction:-As per our analysis of the brands with respect to dissatisfied customers compared to the overall per of dissatisfied customers.(brand should have more than 1000 reviews)
    - We discovered that TURTLE BEACH is receiving 25% dissatisfied customers which is allot higher than 12% of overall dissatisfied customers.
    - We discovered that FUN EXPRESS is receiving 22% dissatisfied customers which is allot higher than 17% of overall dissatisfied customers.
    - The sellers associated with these brands need to be investigated to find root cause of the dissatisfaction of the customers with respect to sellers