TRIMESTER 1 - 2021

Tablet Market Report

GRIFFITH UNIVERSITY
ASSIGNMENT 7130ICT DATA ANALYTICS

Prepared by
GABRIELA ALMEIDA MONTEIRO – S5198626
JULIO PIMENTEL ALBORES – S5172620

Teaching Team
DR HENRY NGUYEN - LECTURER
MR THANH CONG PHAN - TUTOR

Report Structure

The purpose of this report is to discuss and present the most relevant insights on tablet market extracted from the Amazon product review dataset. The Amazon Market contains different product categories such as Books, Video Games and Digital Music. This report only contains information from the Electronics category, especially the Tablet products. Moreover, the analysis was done using several Python libraries: Pandas, gzip, NumPy, Matplotlib, Seaborn, Scikit-Learn, PyLab, statsmodels, itertools, Calplot, pyecharts, NLTK, and Altair. The report is divided in 3 parts: Basic Analysis, Advance Analysis, and Evaluation.

The Basic Analysis describes the exploratory analysis made to the Tablet dataset. It also provides a detailed explanation of the data preparation and pre-processing of the Electronics and Tablets dataset. Finally, this section proposes several hypotheses which are the basis of the Advanced Analysis.

The Advanced Analysis is divided in 3 parts: Brand and Product Analysis, Sentiment Analysis, Time Series Analysis. First, the Brand and Product Analysis describes the most relevant brands in the market and the most popular products on sale. Second, in the Sentiment Analysis the review scores are transformed in positive and negative sentiment and a Naïve Bayes algorithm is developed to predict the sentiment of a review. Expanding on the sentiment analysis, lexical scores are generated in an attempt to predict the rating that a certain user would give to a product.

Next, in the Time Series Analysis, trend and seasonality are captured, forecasting methods are used and evaluated. Moreover, reviews by month and days of the week are further explored to identify patterns. Finally, for the whole period of the data, variation in sales rank for each brand is visualised using interactive plot.

Link to Code

https://github.com/julio-pimentel/Griffith_7130ICT_assignment

PART 1: BASIC ANALYSIS

Dataset Overview

For this analysis, the Amazon electronics dataset was considered. This dataset was extracted from the website http://jmcauley.ucsd.edu/data/amazon/ and is composed of two parts: product reviews and product metadata. Both files were in JSON format and were transformed to Pandas data frames to proceed with the analysis.

Product Reviews Dataset

	reviewerID	asin	reviewerName	helpful	reviewText	overall	summary	unixReviewTime	reviewTime
0	AO94DHGC771SJ	0528881469	amazdnu	[0, 0]	We got this GPS for my husband who is an (OTR)...	5.0	Gotta have GPS!	1370131200	06 2, 2013
1	AMO214LNFCEI4	0528881469	Amazon Customer	[12, 15]	I'm a professional OTR truck driver, and I bou...	1.0	Very Disappointed	1290643200	11 25, 2010
2	A3N7T0DY83Y4IG	0528881469	C. A. Freeman	[43, 45]	Well, what can I say. I've had this unit in m...	3.0	1st impression	1283990400	09 9, 2010
3	A1H8PY3QHMQQA0	0528881469	Dave M. Shaw "mack dave"	[9, 10]	Not going to write a long review, even thought...	2.0	Great grafics, POOR GPS	1290556800	11 24, 2010
4	A24EV6RXELQZ63	0528881469	Wayne Smith	[0, 0]	I've had mine for a year and here's what we go...	1.0	Major issues, only excuses for support	1317254400	09 29, 2011

Product Metadata Dataset

	asin	imUrl	description	categories	title	price	salesRank	related	brand
0	0132793040	http://ecx.images-amazon.com/images/I/31JIPhp%...	The Kelby Training DVD Mastering Blend Modes i...	[[Electronics, Computers & Accessories, Cables...	Kelby Training DVD: Mastering Blend Modes in A...	NaN	NaN	NaN	NaN
1	0321732944	http://ecx.images-amazon.com/images/I/31uogm6Y...	NaN	[[Electronics, Computers & Accessories, Cables...	Kelby Training DVD: Adobe Photoshop CS5 Crash ...	NaN	NaN	NaN	NaN
2	0439886341	http://ecx.images-amazon.com/images/I/51k0qa8f...	Digital Organizer and Messenger	[[Electronics, Computers & Accessories, PDAs, ...	Digital Organizer and Messenger	8.15	{'Electronics': 144944}	{'also_viewed': ['0545016266', 'B009ECM8QY', '...	NaN
3	0511189877	http://ecx.images-amazon.com/images/I/41HaAhbv...	The CLIKR-5 UR5U-8780L remote control is desig...	[[Electronics, Accessories & Supplies, Audio &...	CLIKR-5 Time Warner Cable Remote Control UR5U-...	23.36	NaN	{'also_viewed': ['B001KC08A4', 'B00KUL8O0W', '...	NaN
4	0528881469	http://ecx.images-amazon.com/images/I/51FnRkJq...	Like its award-winning predecessor, the Intell...	[[Electronics, GPS & Navigation, Vehicle GPS, ...	Rand McNally 528881469 7-inch Intelliroute TND...	299.99	NaN	{'also_viewed': ['B006ZOI9OY', 'B00C7FKT2A', '...	NaN

Columns Description

Dataset	Column name	Description
Product reviews	reviewerID	The ID of the reviewer
	asin	The ID of the product
	reviewerName	Name of the reviewer
	helpful	Helpfulness rating of the review
	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)
Product metadata	asin	The ID of the product
	imUrl	Url of the product image
	description	Description of the product
	categories	List of categories the product belongs to
	title	Name of the product
	price	Price in US dollars (at time of crawl)
	salesRank	Sales rank information
	related	Related products (also bought, also viewed, bought together, buy after viewing)
	brand	Brand name

Merging Dataframes

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 1689188 entries, 0 to 1689187
Data columns (total 17 columns):
 #   Column          Non-Null Count    Dtype
---  ------          --------------    -----
 0   reviewerID      1689188 non-null  object
 1   asin            1689188 non-null  object
 2   reviewerName    1664458 non-null  object
 3   helpful         1689188 non-null  object
 4   reviewText      1689188 non-null  object
 5   overall         1689188 non-null  float64
 6   summary         1689188 non-null  object
 7   unixReviewTime  1689188 non-null  int64
 8   reviewTime      1689188 non-null  object
 9   imUrl           1687975 non-null  object
 10  description     1655511 non-null  object
 11  categories      1689188 non-null  object
 12  title           1643686 non-null  object
 13  price           1639882 non-null  float64
 14  salesRank       810070 non-null   object
 15  related         1662142 non-null  object
 16  brand           954251 non-null   object
dtypes: float64(2), int64(1), object(14)
memory usage: 232.0+ MB
```

The product reviews dataset has 1,689,188 rows and 9 columns, while the product metadata dataset has 498,196 rows and 9 columns. Both data frames were merged using the product ID (asin) as a standard foreign key. As expected, the connected data frame has 1,689,188 rows and 17 columns. Most of the attributes are object type. However, four of them have a numeral structure, such as overall (float64), unixReviewTime (int64), and price (float64).

Data Preparation and Preprocessing

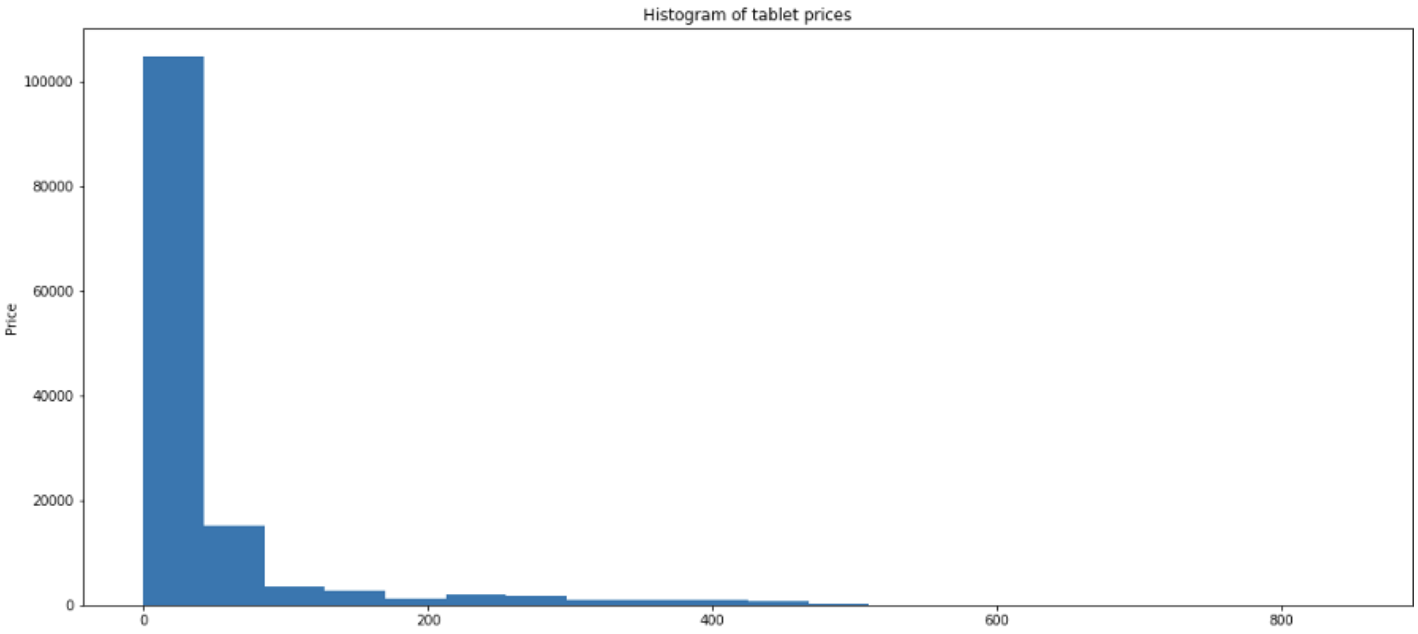
To improve the time efficiency while running the code, the analysis was only focused on tablets. The following steps were followed to filter only data for tablets: all null values of the product title column were dropped because this column was used to filter the tablets subset. Next, all rows that contained the keywords "tablet" and "iPad" in the title column were selected. After doing so, the dataset was reduced to 135.939 rows.

While exploring the dataset with the 'describe' method, it was noticed that the dataset contained many products within a low-price range. Most probably, these were accessories for tablets.

	overall	unixReviewTime	price
count	135939.000000	1.359390e+05	134910.000000
mean	4.190247	1.366043e+09	45.278070
std	1.170919	3.096553e+07	81.089161
min	1.000000	1.010966e+09	0.010000
25%	4.000000	1.352506e+09	9.950000
50%	5.000000	1.371773e+09	17.990000
75%	5.000000	1.389658e+09	38.440000
max	5.000000	1.406074e+09	850.000000

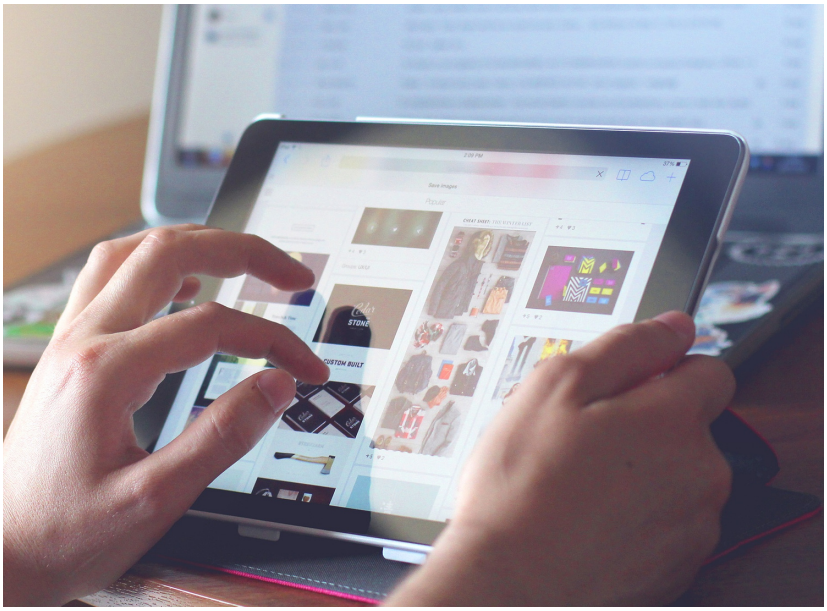
Filtering Data by Price

Since the analysis would be focused on tablets (not their accessories), they needed to be removed. A histogram was used to find the cut-off line between accessories and tablets. It was noticed that below $100, most of the products were grouped. Hence, the $100 filter was set, and the dataset ended up with 12,611 entries.



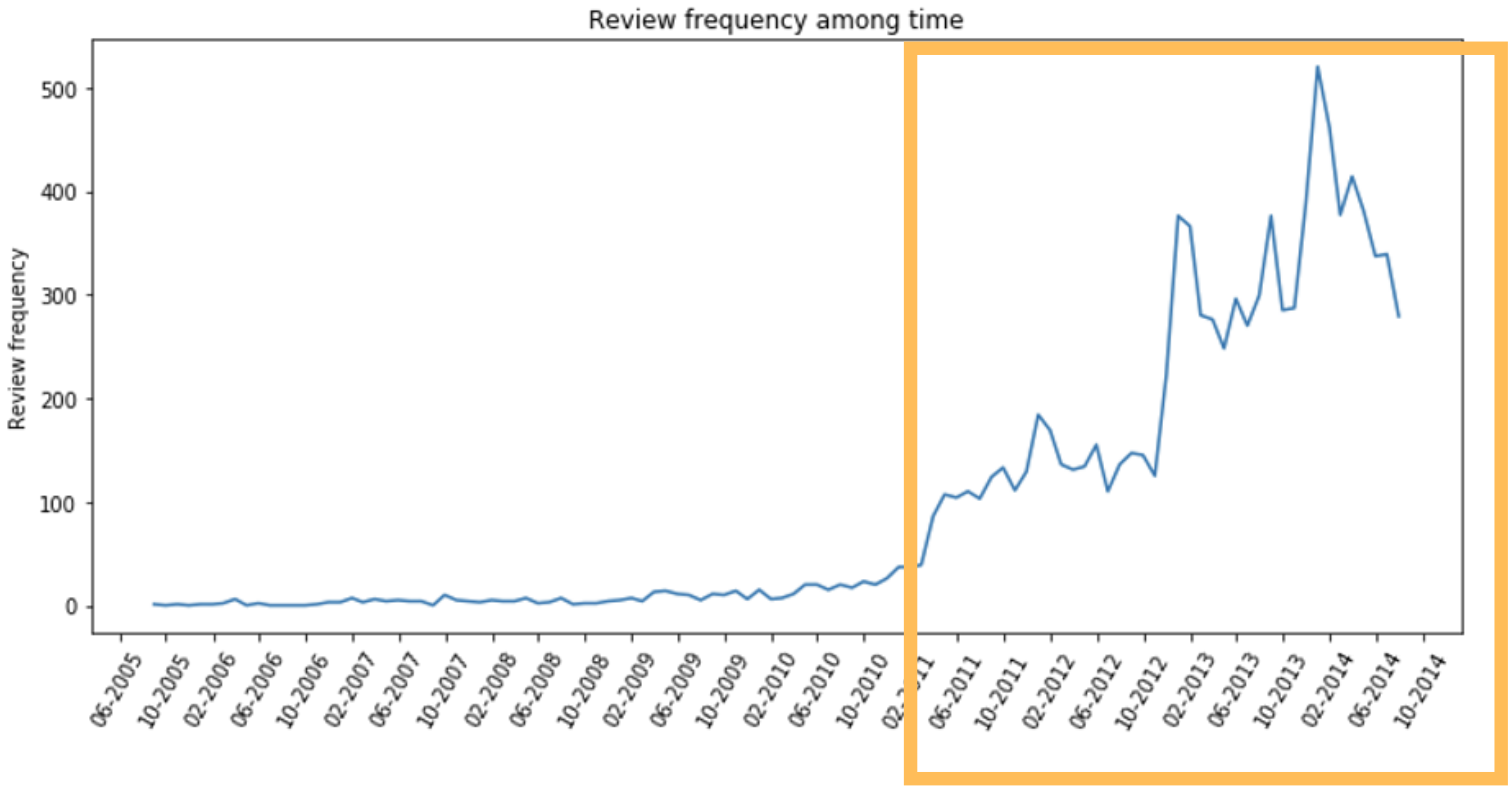
Steps for Data Normalisation

a) A new column called timeFormat was created as a date-time type based on the unixReviewTime column to perform time series-based operations.
b) There were 9227 null values in the brand column. Considering that this column would be relevant for our analysis, the first word of the title column was used to fill in the cells that had missing values in the brand column.
c) Duplicates were dropped based on specific columns (reviewerName, unixReviewTime and asin) and kept the last occurrence.
d) All observations that had null values in the columns reviewerName, price, description and related were dropped.
e) The overall column was split into negative and positive sentiment. Scores below or equal to 3 are considered negative (flagged as 0), and scores above 3 are considered positive (flagged as 1). These values were stored in the sentiment column.
f) A helpfulRatio column was created based on the helpful column.
g) The columns month_year, hour, DayofWeek and day_month were created for time series analysis.
h) The columns reviewerName, helpful, unixReviewTime, reviewTime, salesRank that would not be used in the analysis were dropped.



Filtering Data after 2010

After the data normalisation was done, it can be observed that there are some months without reviews. On the other hand, the peak of sales was after 2010. It is strongly related with the Apple iPad's release on April 3, 2010. It was decided to drop data before April 2010 to make an insight analysis of the tablets.

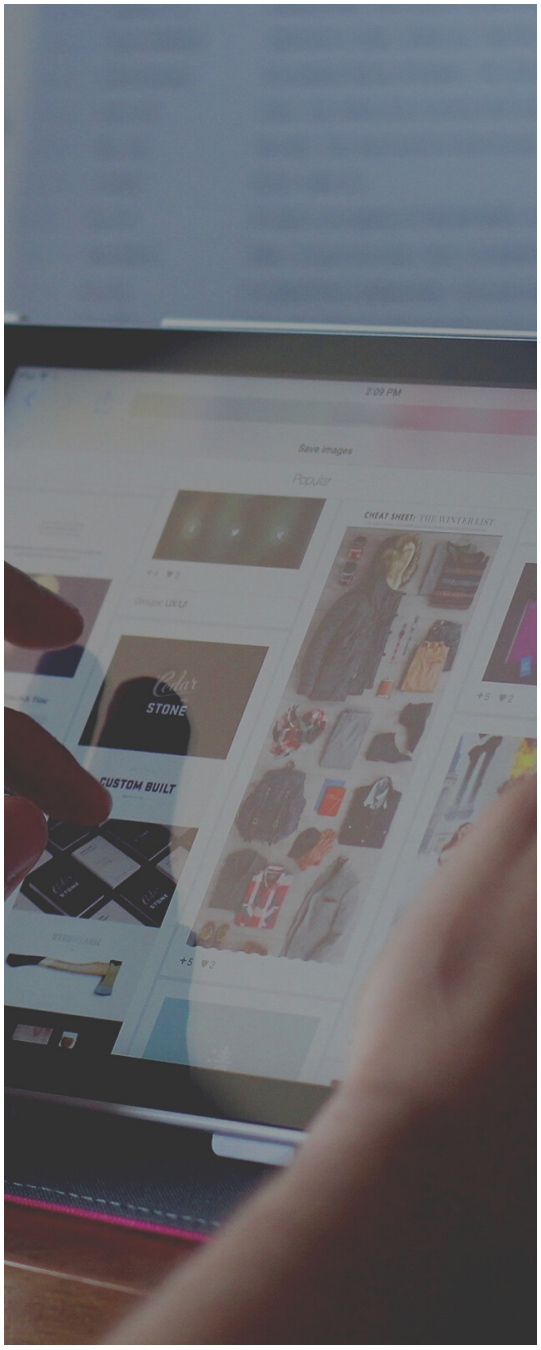


Clean Data Descriptive Statistics

The shape of the dataset is 12,276 rows and 19 columns.
The oldest post is from: 04/04/10
The newest post is from: 23/07/14
There are 10,688 unique users ID who left a review.
There are 312 unique products.
There are 125 unique brands.
There are 9,410 good reviews.
There are 2,866 bad reviews.
The average rating given by customers is 4.10.

Hypothesis

To orient the analysis, the following hypotheses are proposed to make the Advanced Analysis:

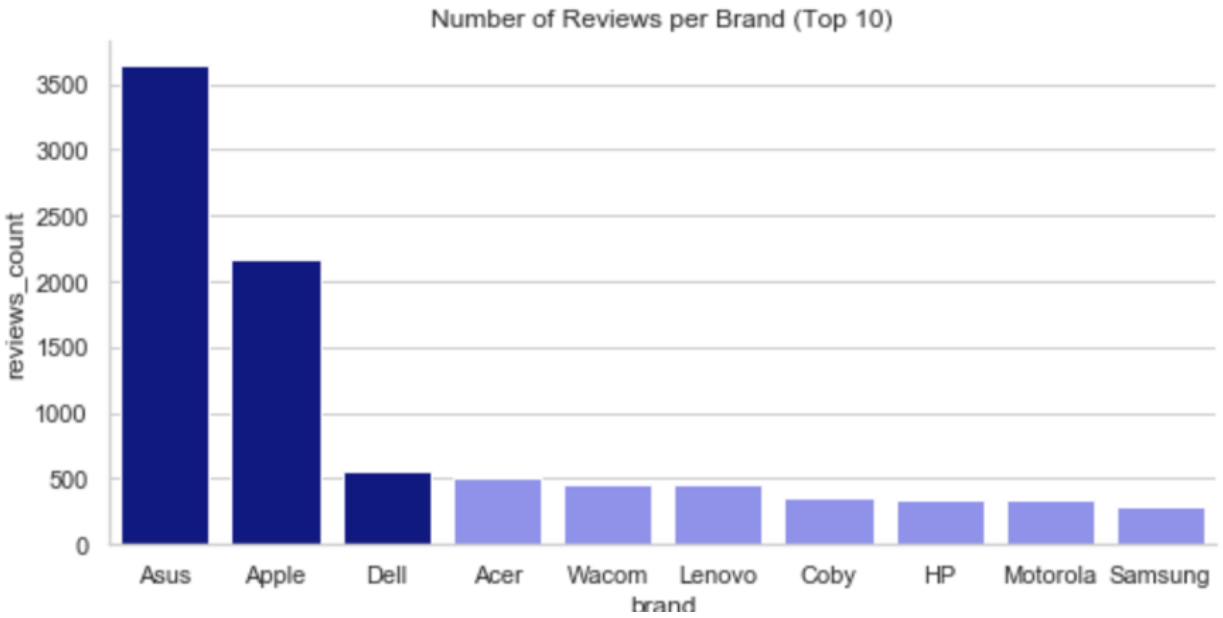


1. Which brands have more reviews and how are their products perceived by customers?
2. How is the evolution of sales per brand over time?
3. Which tablet is more popular in sales amongst customers?
4. What are the top 5 most reviewed products?
5. Is there a correlation between price and ratings?
6. Does rating affect the number of reviews?
7. What are the positive and negative words associated with the reviews?
8. Can we predict ratings based on reviews text analysis?
9. Can we detect temporal buying/reviewing trends?
10. Is there a correlation between specific times of the day and purchasing behaviour?
11. Have large events influenced the evolution of brands during the years?
12. How have the brands' sales rank changed over time?

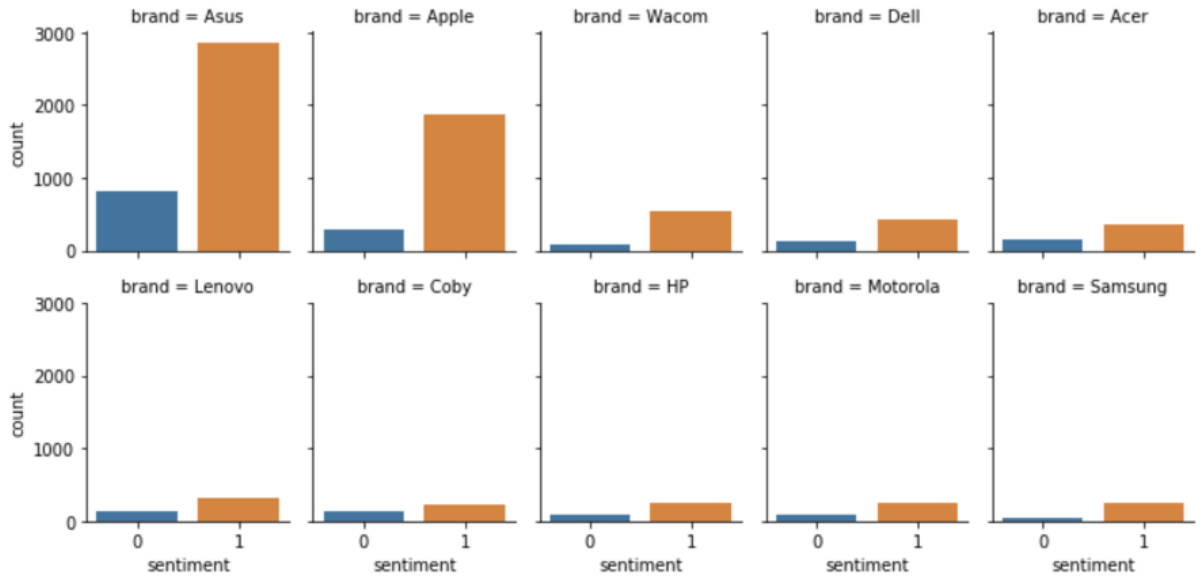
PART 2: ADVANCED ANALYSIS

Top 10 Most Reviewed Brands

As per the plot below, Apple and Asus are way ahead of their concurrence in terms of reviews.

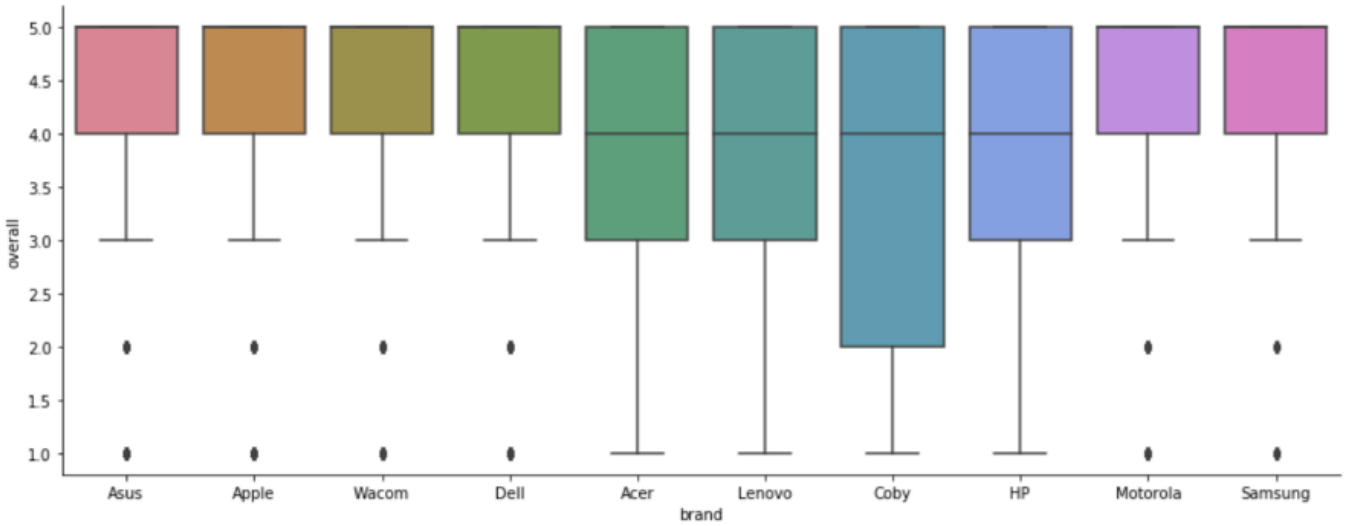


Further analysis also showed that all top 10 brands have more positive than negative reviews. However, Apple and Asus show a significant number of positive reviews compared to negative. In fact, from the plot, the more reviews you have, the more likely you are to have positive reviews and, therefore, be at the top.



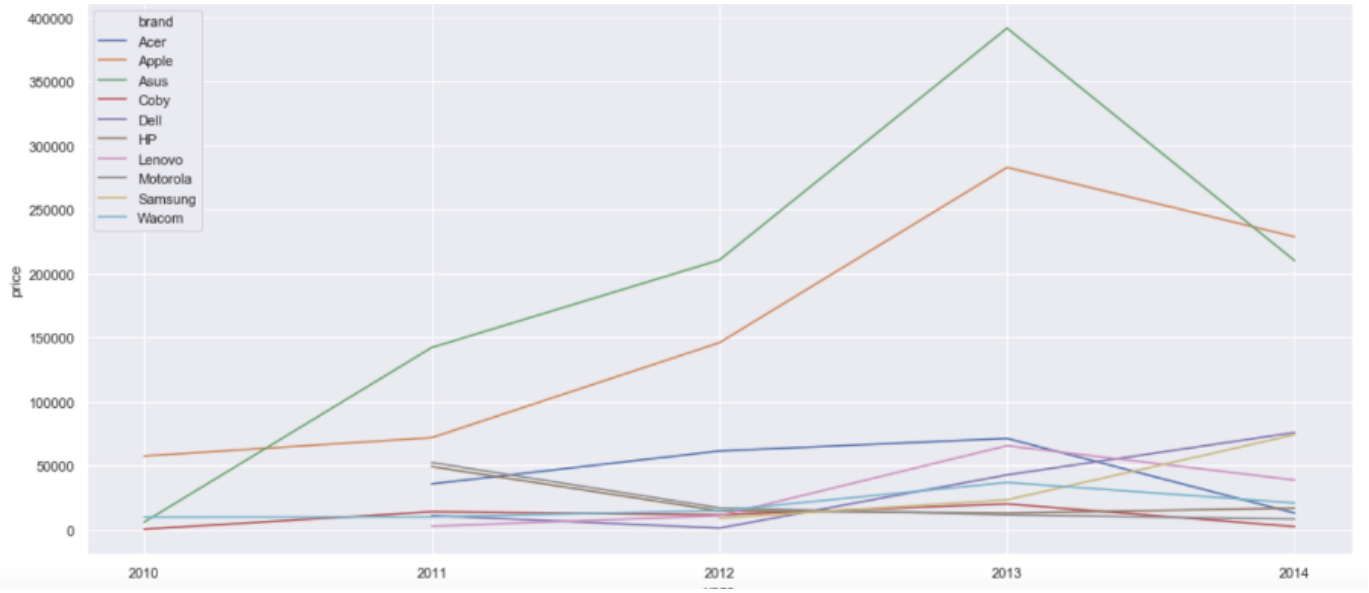
Top 10 Brands Rating Distribution

Customers from the top 4 brands (Asus, Apple, Wacom and Dell) tend to rate the products positively (with a few outliers), which means that they are generally satisfied with their purchases. Similarly, customers of brands in position 9 and 10 are generally happy with their products. Brands ranked between the 5th and 8th position (Acer, Lenovo, Coby and HP) have the worst performance amongst the top 10 brands. Special attention to Coby, where customers ratings are more distributed than the other brands. It is not surprising that Coby closed its doors in June 2013 (CRN, 2014).



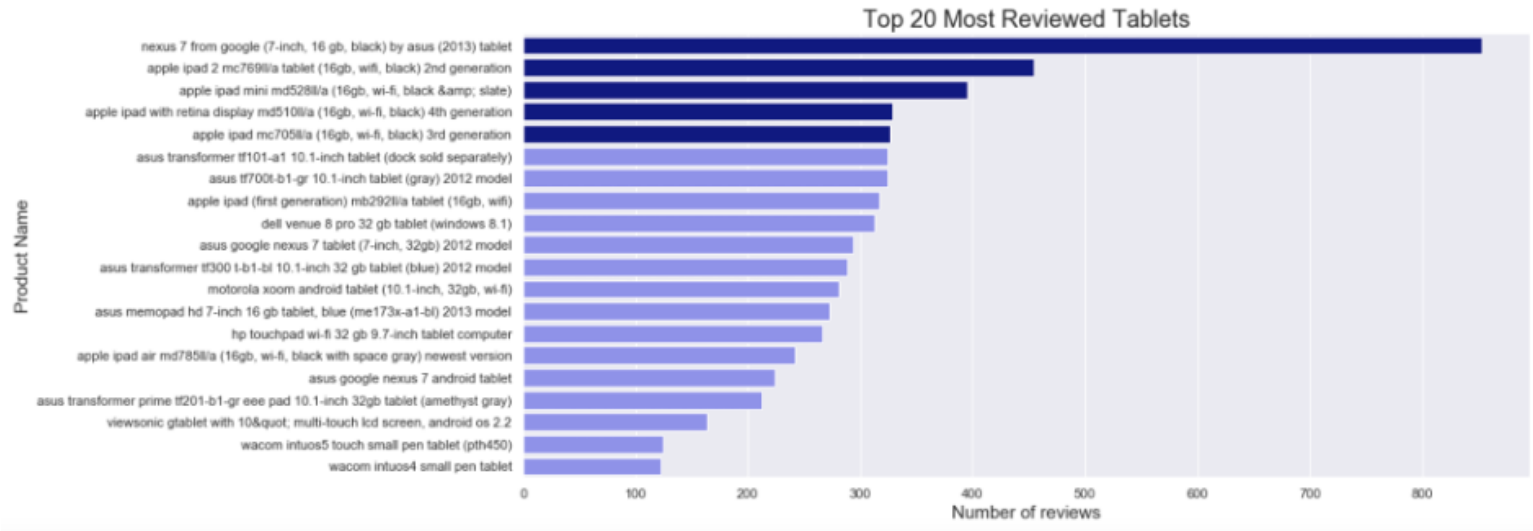
Top 10 Brands Sales

in the following graph it is possible to spot a few interesting facts. First, Apple was at the top of tablets sales until the middle of 2010. Then, Asus took the lead, and Apple stayed in the second position until the end of 2013, when it took over the first position again. The year of 2013 seems to have impacted the sales of most brands, which see their demand dwingle, except for Samsung and Dell. Samsung shows impressive performance and keeps an upward trend during the whole period. This is certainly a brand that the top companies should keep an eye on. Secondary research also shows that 2013 was a year when the tablet market was invaded by Android's tablets, which were of high quality and cheaper than iPads (CRN, 2014). Apparently, customers started to buy less tablets from Apple and Asus and more from other companies, especially Samsung.



Top 20 Products

Asus and Apple clearly lead the market of iPads in the analysed period. From the top 20 most reviewed products 8 are from Asus and 6 are from Apple. Asus Nexus 7 occupies the first place in number of reviews, followed by 4 Apple's iPads.

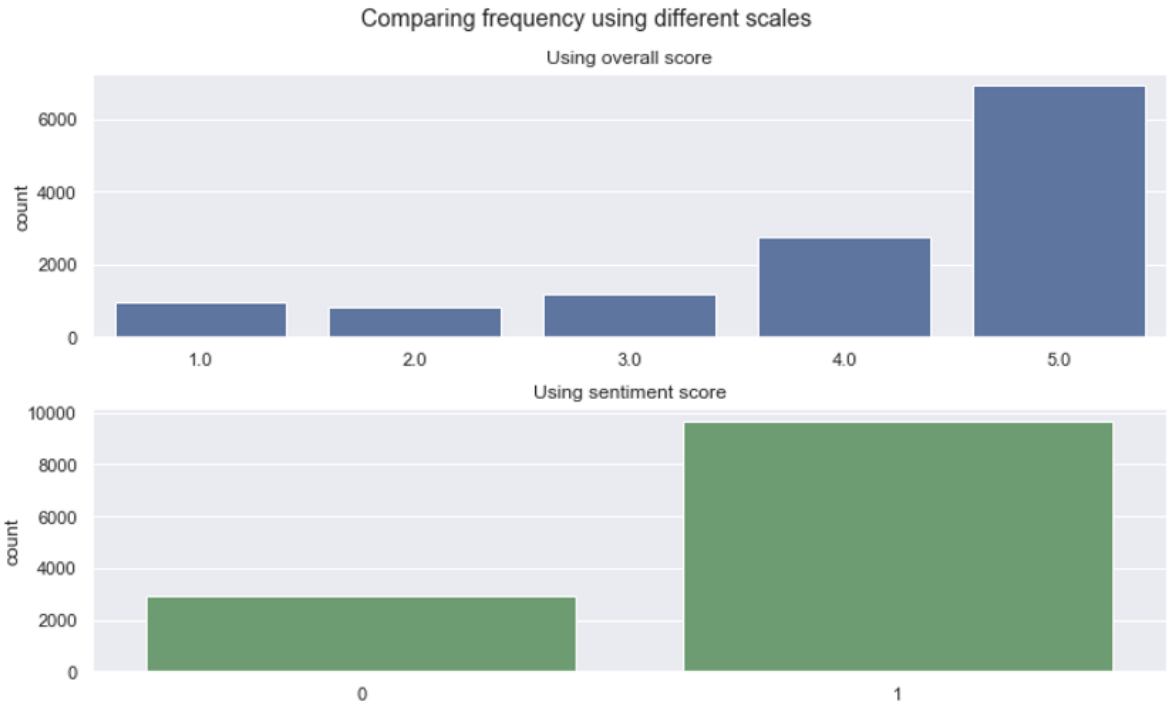


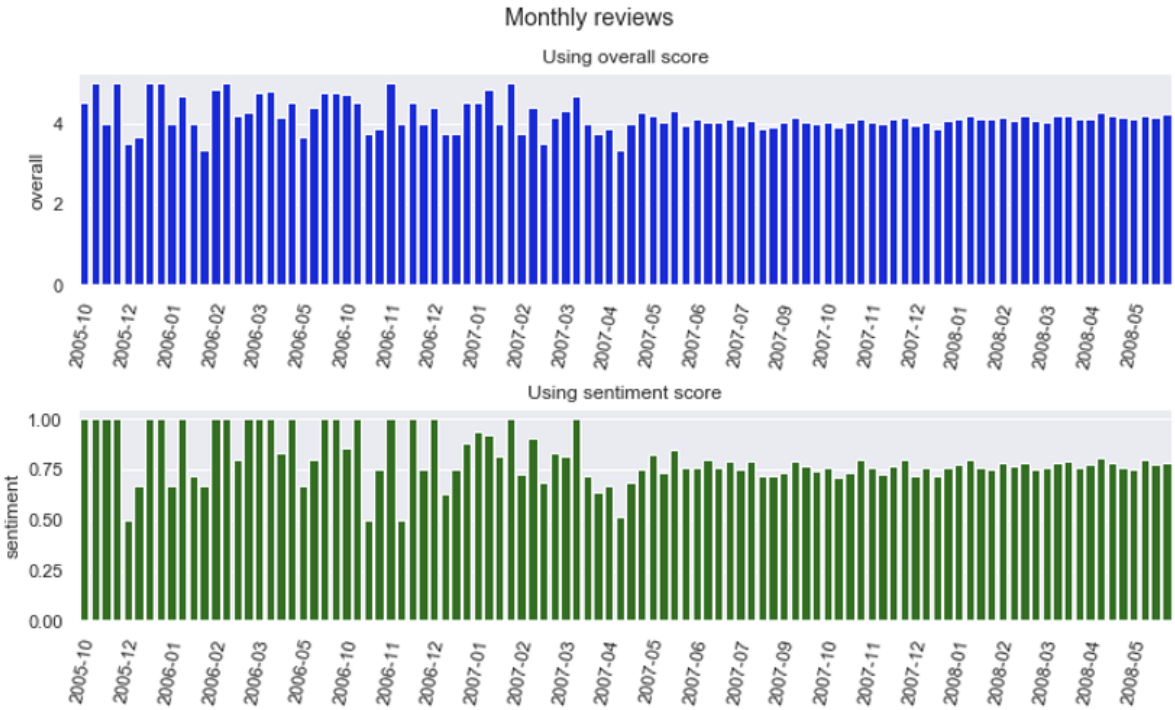
First place
Nexus 7 by Asus

Second place
iPad2

Sentiment Analysis

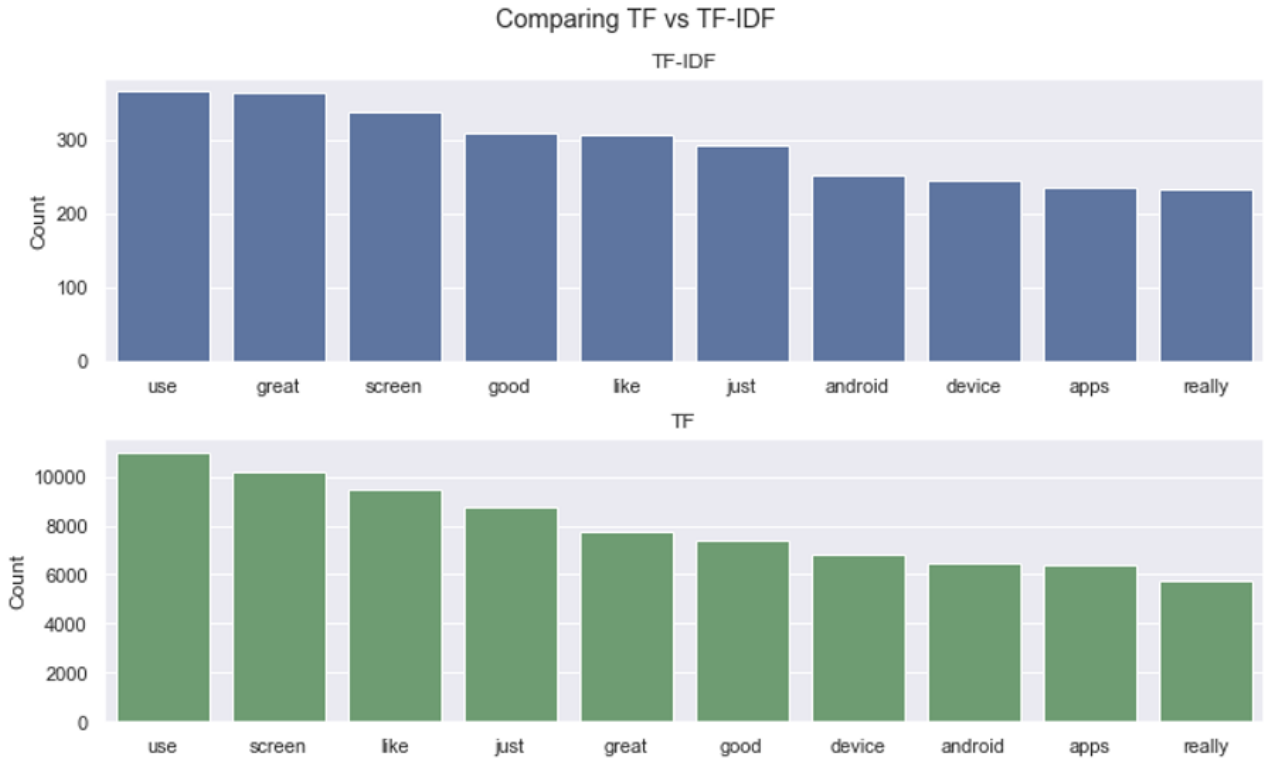
In the following histogram, it can be observed that most of the reviews made by customers were positive. The proportion of reviews were concentrated in the top score (5). This insight indicates that the overall tablets have positive feedback from customers. The monthly average behaviour of scores and sentiment of reviews suggests that the trend is consistent across time.





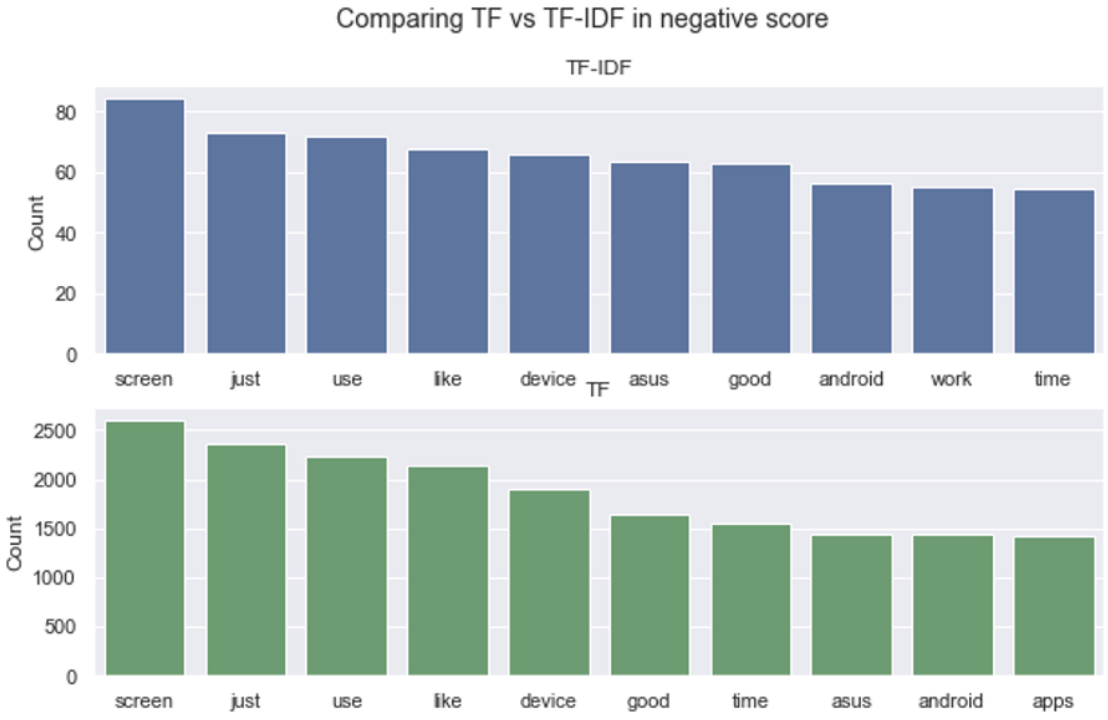
Top 10 words of TF and TF-IDF analysis of the whole dataset

CountVectorizer and TfidfVectorizer methods from the Scikit-learn library were used to analyse the most relevant words from the dataset. One of the most surprising insights from the TF-IDF analysis was that the most common words were "tablet" and "ipad". Considering that these were the keywords to filter the dataset, it was assumed that they would not score high in this analysis. Therefore, these words were removed from the top 10 words.
The most relevant words in the whole dataset are: "use", "great", and "screen". The most appropriate words for the positive reviews were "use", "screen", and "like". Finally, the most relevant terms from the negative reviews were "screen", "just", and "use". In conclusion, the usability and screen aspect appear relevant features for tablet companies to consider.

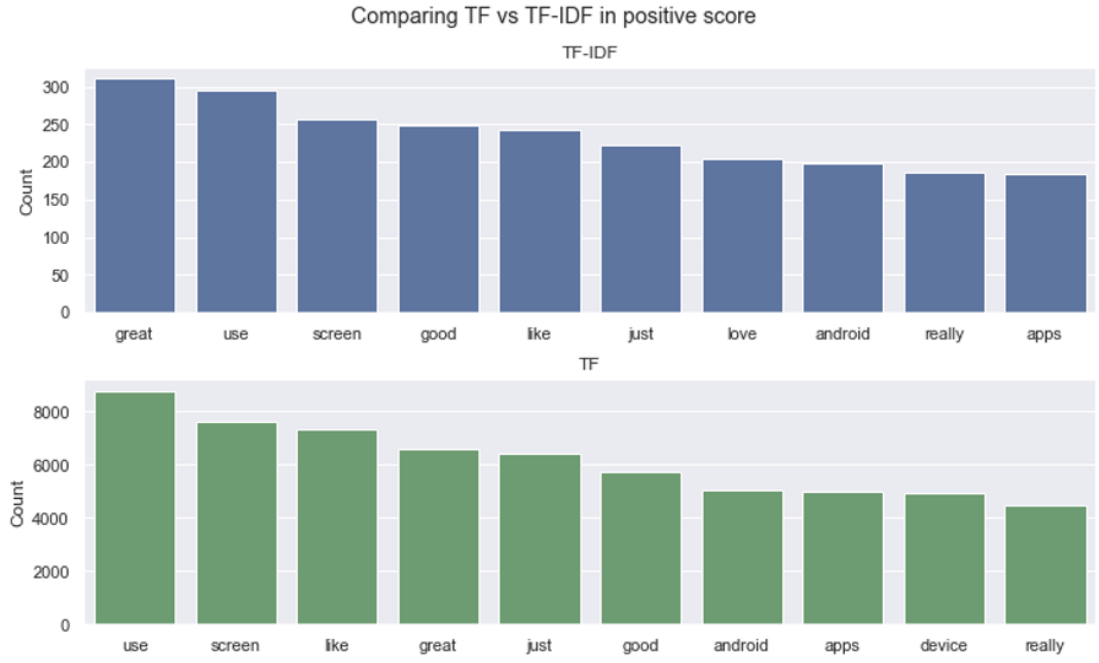


Naive Bayes algorithm was implemented using the SentimentAnalyzer method from the NLTK library. The dataset was divided into positive and negative reviews. Next, both datasets were split into training and test datasets. Finally, respective training and test datasets of the positive and negative reviews were joined. This sampling method was selected to have the same proportion of negative and positive reviews in training and test data. As a result, the accuracy of the Naive Bayes algorithm was 72.62%. However, the precision for negative reviews was 44.50%. In contrast, the precision for positive reviews was 83.51%.

Top 10 words of TF and TF-IDF analysis of negative reviews dataset



Top 10 words of TF and TF-IDF analysis of positive reviews dataset



Using Lexical Score to Predict Product Rating

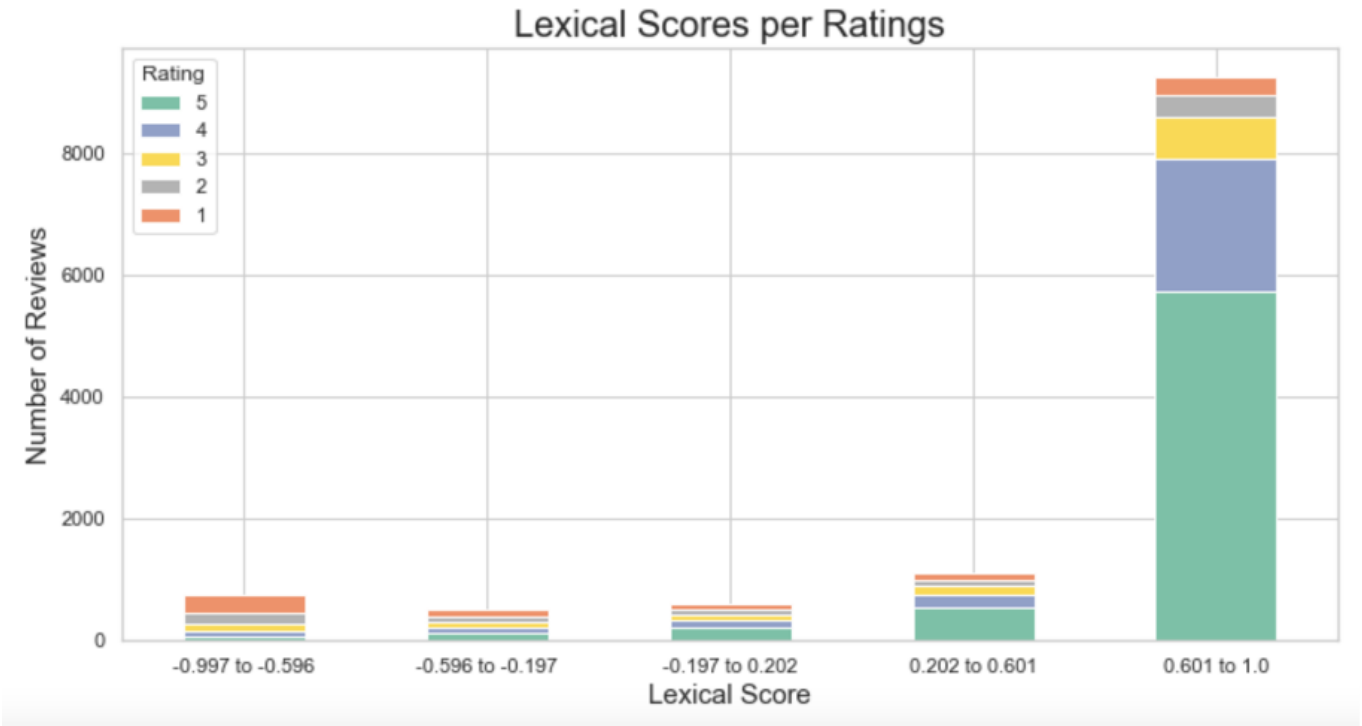
Extending the Sentiment Analysis, in this section, a Lexical Score was generated for each text review in order to predict the rating that a user would give to a product. This analysis was not successful and had a very low accuracy score (Figure 18). This might be explained because the Lexical Scores generated were not matching the real sentiment of the user. This could have happened because of ambiguous words that were used in the reviews, which ended up generating inaccurate lexical scores. Additionally, trying to predict many different categories (ratings from 1 to 5) may have impacted the results.

```
from sklearn.linear_model import LinearRegression
my_model = LinearRegression()
my_model.fit(X_train, y_train)

print(f"The accuracy score of the model is {my_model.score(X_test, y_test)}")
```

```
The accuracy score of the model is 0.20895098364991227
```

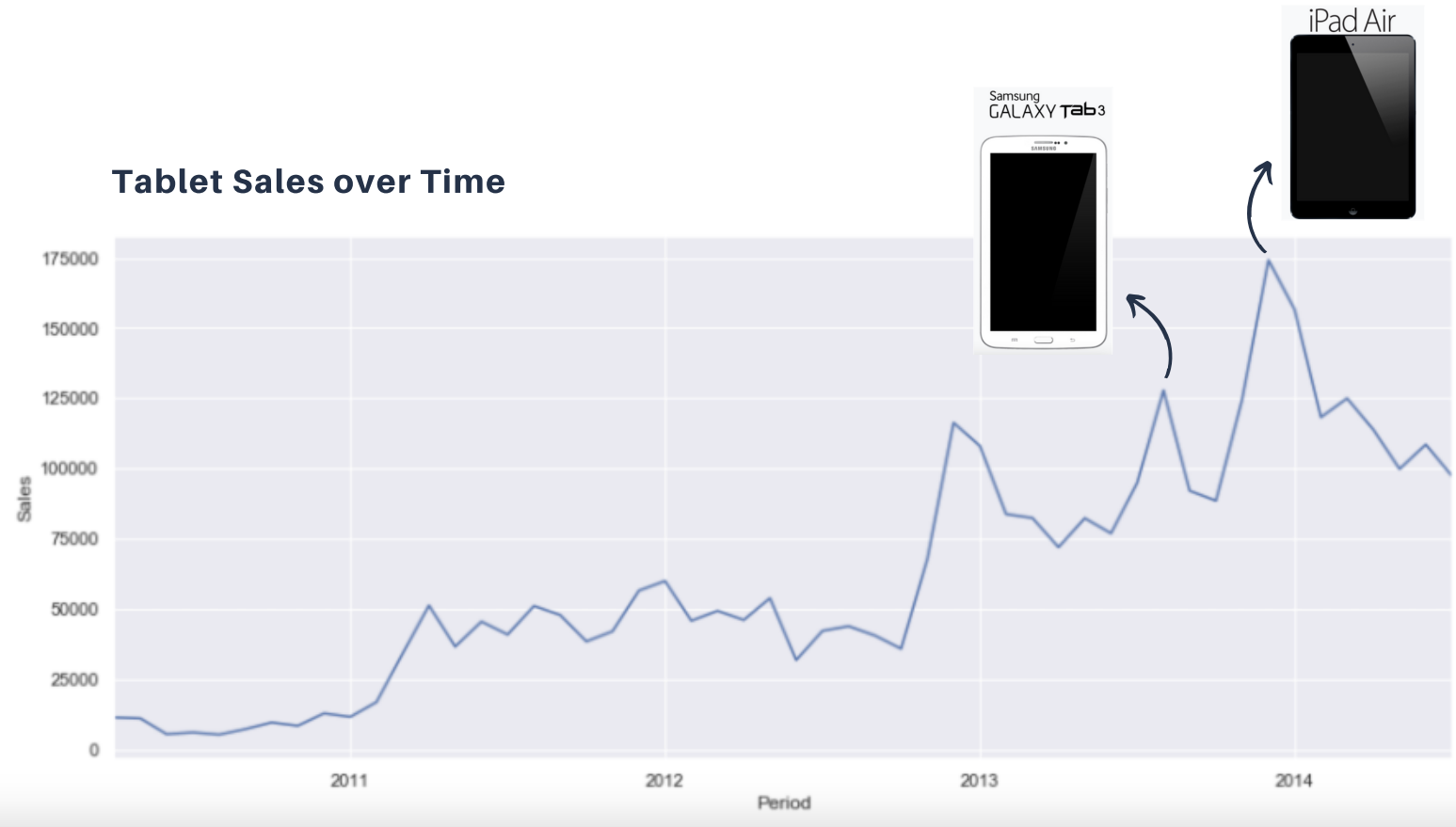
As it is possible to see in Figure 19, the bin with the highest lexical score (above 0.601) includes a great number of reviews with rating 5. Similarly, the bin with the lowest lexical score (below -0.596) has a great number of reviews with ratings 1. However, the bins in between are ambiguous. So, it is possible to conclude that the lexical scores align well with extreme ratings, but cannot accurately capture intermediate ratings.



Trend and Seasonality

The plot shows that the tablet sales have an overall upward trend. It also suggests a seasonality pattern, mostly associated with the months that Apple has new releases. This pattern started after Apple launched its first iPad in April 2010. In March 2011, there is a remarkable peak after Apple launches iPad 2. In 2012, again, the peak in sales happen in November, after Apple launches its first iPad Mini. In 2013, for the first time, there are two significant peaks. One happens in the middle of the year, and the other one, in the end of the year. The first peak is most likely due to Samsung releasing Samsung Galaxy Tab 3, which was a strong rival for Apple. The second peak is most likely associated with Apple releasing 2 new iPads (iPad Air and iPad Mini 2).

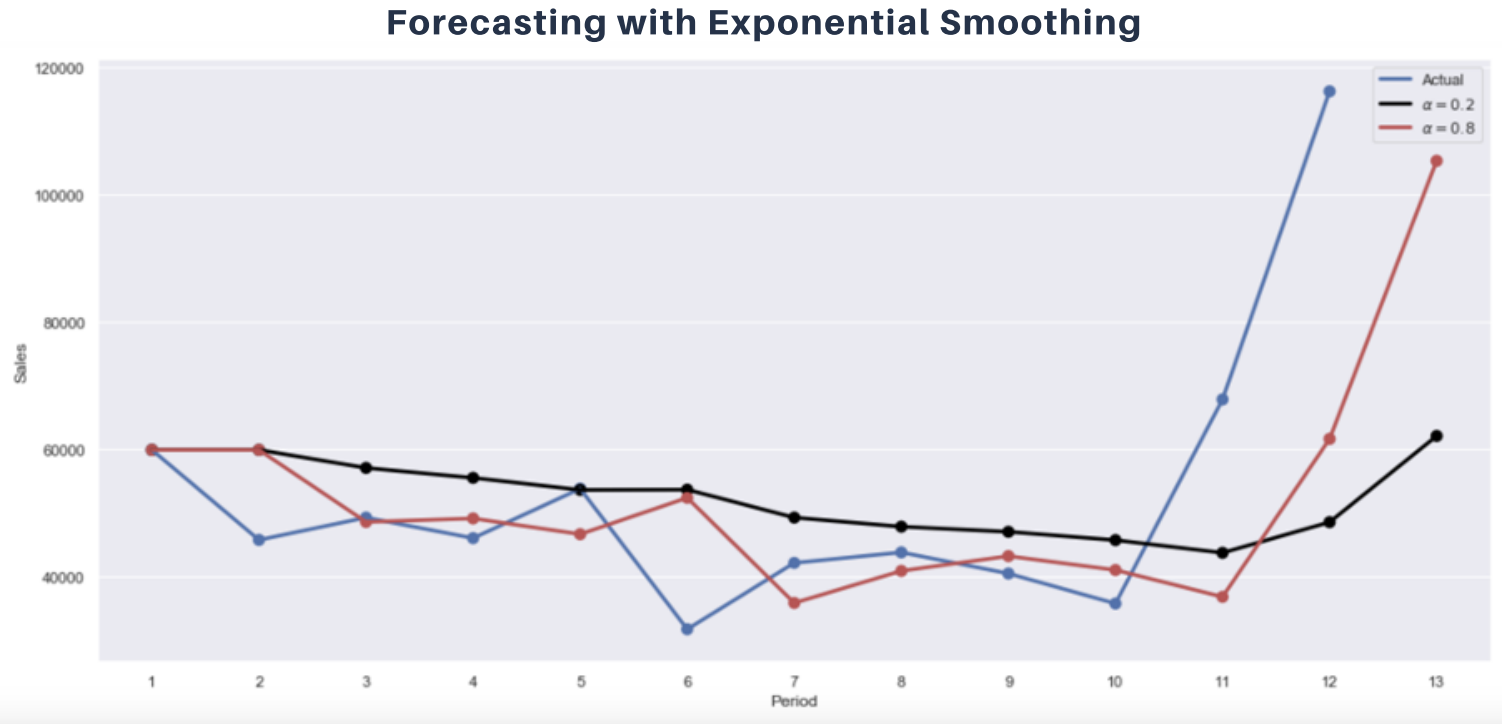
From this plot it is possible to infer that Apple releases play a tremendous impact in the overall sales of tablets, and that 2013 was a year of significant expansion in the tablet market.



Forecasting

Various methods of forecasting were used, and after calculating their Mean Absolute Deviation (MAD), it was verified that Exponential Smoothing with alpha equals to 0.8 was giving the best accuracy as seen from the outputs below. Therefore, this forecasting method should be the one used for making tablet sales predictions for future months.

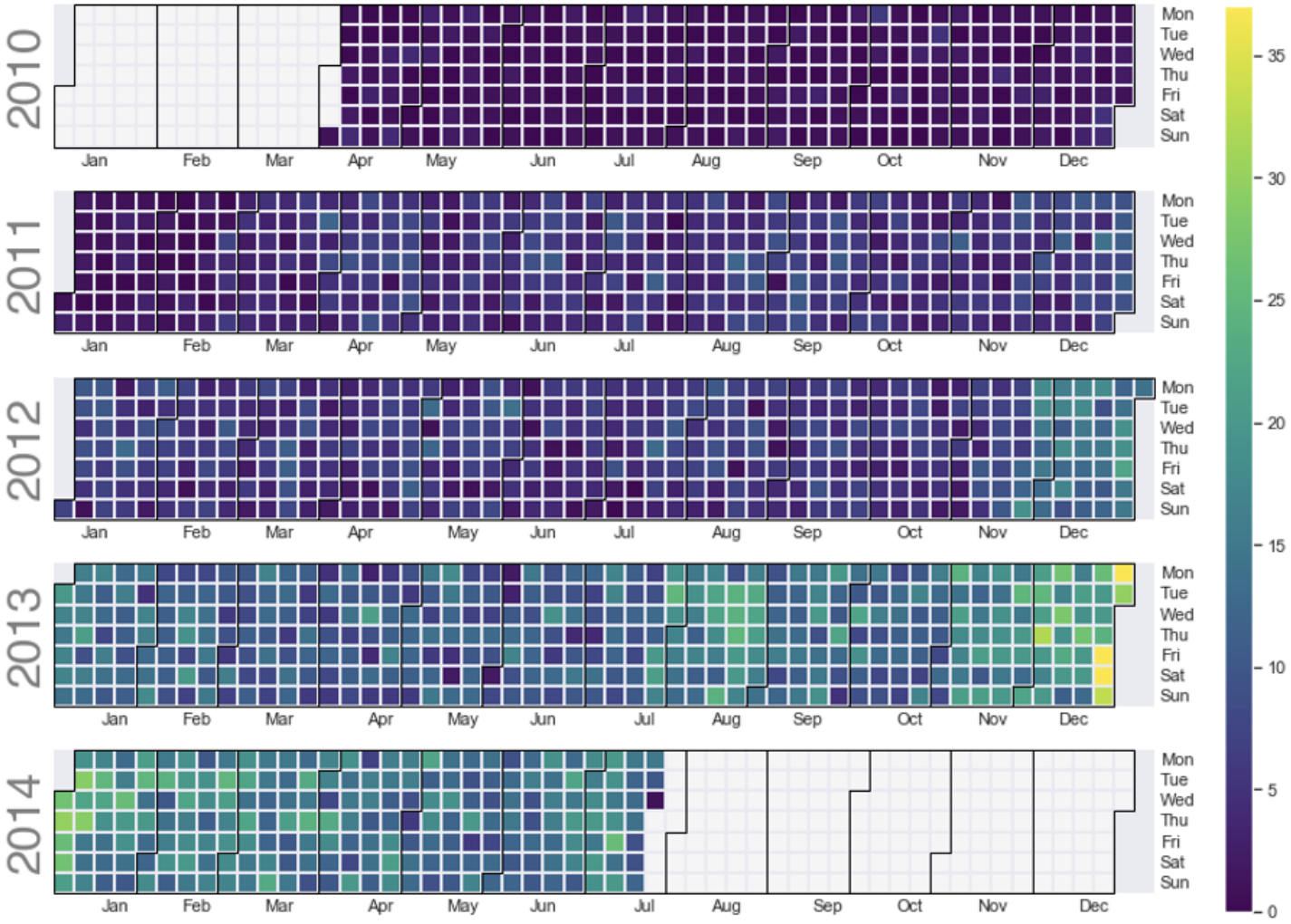
```
MAD of 3-MA: 22027.243333333576
MAD of Exponential smoothing with alpha=0.2: 17634.142799997233
MAD of Exponential smoothing with alpha=0.8: 11056.836581531104
MAD of Linear Regression: 16492.490530303316
MAD of Seasonality Method: 48217.985079197075
```



Reviews by Month

When analysing how the number of reviews changed by month it was verified that during Christmas and New Year people tend to write more reviews. This is most likely because it is a period of the year when people buy more tablets and have more free time due to holidays.
Another highlight from this analysis was that starting from the last week of July 2013, there was a noticeable increase in the number of reviews. This was the period when Samsung released its new tablet, Samsung Galaxy Tab. That year Samsung doubled its presence in the tablet market share and became a strong competitor for Apple.

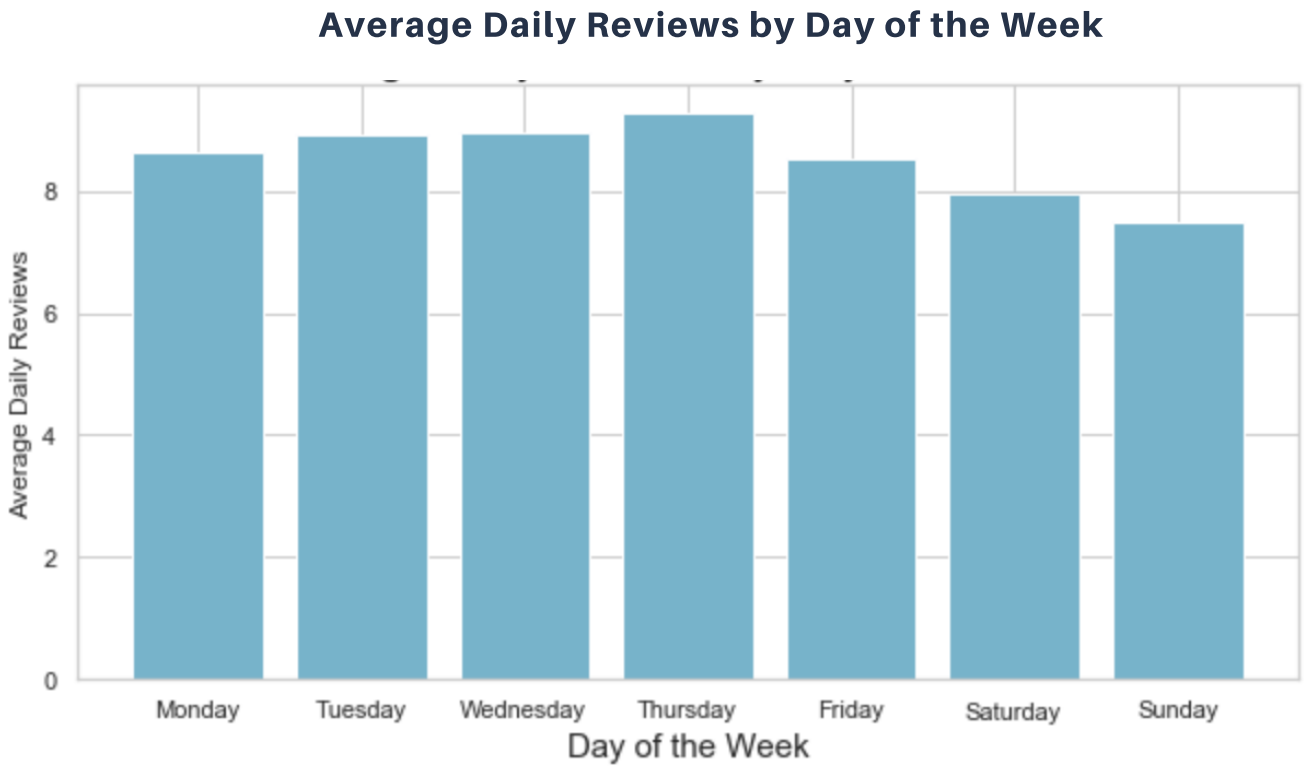
Number of Reviews per Day



Reviews by Day of the Week

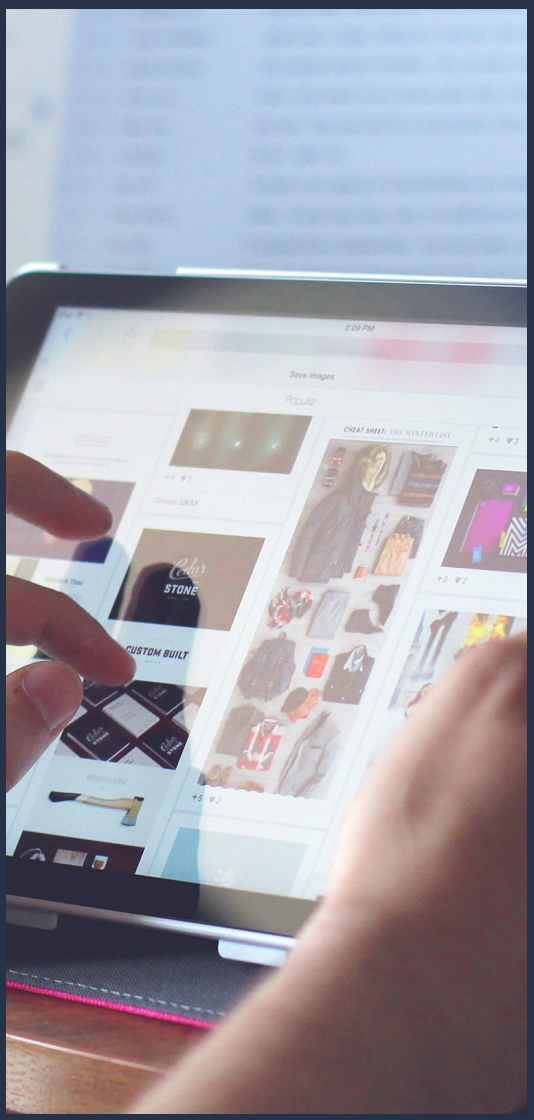
From the plot it is remarkable that weekends have the lowest number of reviews, however, starting from Monday, there is an upward trend, with the peak of the number of reviews being on Thursday.

"When asking your customers for a review, it's always a good idea to put yourself in their shoes. Always try to ask for reviews at the right time - failing to do so could cause confusion and even irritation." (Reviews.io, 2019)



PART 3: EVALUATION

Findings of Data Analysis



The following insights are the most relevant findings of this report:

1.The brands with more reviews are Asus and Apple and they are positively perceived by the customers.

2.Tablet sales follow an upward trend, after 2010, when Apple released its first iPad. Over time, some companies (like Samsung) showed a vigorous growth and doubled the market presence, while others (like Asus) had a steady drop.

3. The most popular tablet in the dataset is Nexus 7 from Asus.

4. The top 5 most reviewed products are: Asus Nexus 7, Apple iPad 2 MC769II, Apple iPad Mini MD528II, Apple iPad MD5101II, and Apple iPad MC705II.

5.There is no correlation between price and ratings. Therefore, more expensive tablets do not tend to have a more positive reviews or vice versa.

6.Does rating affect the number of reviews? There is a general tendency for tablet users to write more positive than negative reviews. Therefore, encouraging more users to write a review can help the product to be more visible and place the company at the top of the sales rank.

7.The terms "use" and "screen" are the most relevant in negative and positive reviews, meaning that usability and screen aspect should be taken into consideration by tablet designers.

Findings of Data Analysis (continued)

8.It is not possible to predict a rating that users would give to a product just based on sentiment analysis of their text reviews. A possible reason for this is that the generated lexical score for reviews was not accurate enough to predict multiple ratings (from 1 to 5). To overcome this challenge with textual analysis, Amazon should have check boxes for users to tick when evaluating an electronic product. Check boxes could include features such as speed, storage, weight, usability, responsivity, screen aspect, etc. This kind of more structured data could be more meaningful for this type of analysis.

9.People are most likely to write reviews on Thursday, therefore this would be a good day of the week for companies to send an email asking for a product review. Another recommendation is avoid asking for reviews during the weekend.

10. Due to lack of data for the review time, it was not possible to explore any correlation between specific times of the day and purchasing behaviour.

11. It is remarkable the relation between Apple tablet releases and the amount of tablet sales in the subsequent month. Hence, it is possible to conclude that Apple releases is a large event that influences the evolution of tablet sales during the years.

12. In the analysed period (2010-2014) there is a significant evolution of Samsung in the tablet market. This report would recommend Apple to closely monitor Samsung as a potential strong competitor.

Improvements for the Project

- One of the main restrictions for making significant analysis was the hardware limitations. The laptops used to analyse the original datasets were not powerful enough to process large datasets efficiently. Cloud Computing could solve this limitation by processing the original data, after this step a smaller suitable dataset to work on a computer could be created.
- The filtering method could be improved by using the categories column in the original product dataset. The category was stored in a nested list with different inner list size. It was possible to split this nested list using 1,000 rows. However, the computing time was excessive when the whole dataset was utilised. As a result, the most efficient filtering alternative was using the title column.
- On Amazon, users can review products that they have bought and used. It could be interesting to find about the reasons a user does not buy a particular product. The analysis could be done by collecting comments from YouTube and Twitter to get the impressions of a wider public.
- When trying to predict the product rating based on the lexical score generated for each review, the model had a very low accuracy most probably because words used in some reviews were ambiguous. As further explored in the code, one user received a very high lexical score for their review, when in fact the review was very negative. What happened was that the user was complimenting another product in comparison with the one purchased. The positive words for the rival product contributed to a high lexical score. A possible improvement for this type of analysis is to train the model with ambiguous situations like this.
- Also, during the time series analysis, it was assumed that the Amazon monthly sales were equal to the sum of the prices of the products reviewed in a particular month. Therefore, the seasonality that was captured might apply only to this specific market. Having access to the real monthly sales of tablets could better understand the seasonality patterns for this product. It could be obtained by government tax information of this sector.
- When looking at the boxplot to check the distribution of ratings for each brand, it was noted that there were some outliers that gave very low ratings for the products they purchased. It could be interesting for the brands to look at what the outliers are saying because they can be an essential source for insights for future improvements on the product.

REFERENCES

Altair Developers. (2019). Altair Developers. Retrieved from Customizing Visualizations: https://altair-viz.github.io/user_guide/customization.html

Altair Developers. (2019). Altair Developers. Retrieved from TimeUnit Transform: https://altair-viz.github.io/user_guide/transform/timeunit.html#user-guide-timeunit-transform

CRN. (2014). Top 10 Best-Selling Tablet Brands of 2013. Retrieved from https://www.crn.com/slide-shows/mobility/300072670/top-10-best-selling-tablet-brands-of-2013.htm/1

Geeks for Geeks. (2020, November 26). Geeks for Geeks. Retrieved from Python Seaborn – Catplot: https://www.geeksforgeeks.org/python-seaborn-catplot/

Han, D. (2020, June 8). Python Plain English. Retrieved from Make beautiful and interactive bar charts in Python: https://python.plainenglish.io/make-a-beautiful-bar-chart-in-just-few-lines-in-python-5625ebc71c49

He, R., & McAuley, J. (2016). Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. WWW.

Li, S. (2018, July 9). Towards Data Science. Retrieved from An End-to-End Project on Time Series Analysis and Forecasting with Python: https://towardsdatascience.com/an-end-to-end-project-on-time-series-analysis-and-forecasting-with-python-4835e6bf050b

Lynn, S. (n.d.). shanelynn.ie. Retrieved from Data science, Startups, Analytics, and Data visualisation: https://www.shanelynn.ie/pandas-drop-delete-dataframe-rows-columns/

McAuley, J., Targett, C., Shi, J., & Hengel, A. v. (2015). Image-based recommendations on styles and substitutes. SIGIR.

Reviews.io. (2019). Timing Is Everything: Ask For Reviews At The Right Time. Retrieved from: https://blog.reviews.io/post/timing-is-everything-ask-for-reviews-at-the-right-time

Stack Overflow. (n.d.). Stack Overflow. Retrieved from How to create a stacked bar chart for my DataFrame using seaborn? [duplicate]: https://stackoverflow.com/questions/47138271/how-to-create-a-stacked-bar-chart-for-my-dataframe-using-seaborn

Tom. (2021, March 7). Github. Retrieved from Calendar heatmaps from Pandas time series data: https://github.com/tomkwok/calplot

Yıldırım, S. (2015, May). Towards Data Science. Retrieved from Making Interactive Line Plots with Python Pandas and Altair: Making Interactive Line Plots with Python Pandas and Altair