Amazon Reviews Analysis

Leon Andov s5088455 | Nathan Wuiske s5058046 | Siyun Lu s5121731

Website: https://nathanwuiske.github.io

PART 1

The following assignment applies data analytics by studying and analyzing Amazon product reviews left by consumers. By using data preparation and preprocessing techniques as well as exploratory, predictive and statistical data analysis techniques, an evaluation can be made to determine appropriate future actions.

Dataset Description

The dataset that will be analyzed is the "Automotive" section of Amazon Reviews. This dataset contains approximately 20,000 lines within a JSON file, with each line representing a different review made on the website. The data of each review is enclosed into curly braces indicating that it's an object. Inside this object a number of attributes can be defined using a *"name": "value"* pairing scheme separated by commas. An example extract of one review from the file is as follows:

{"reviewerID": "AO94DHGC771SJ", "asin": "0528881469", "reviewerName": "name", "helpful": [0, 0], "reviewText": "Good product", "overall": 5.0, "summary": "Good", "unixReviewTime": 1370131200, "reviewTime": "06 2, 2013"}

By decomposing the attributes of the above review the following table can be produced:

Attribute/Feature	**Description**
reviewerID	A 13-character alphanumeric unique identifier assigned to each user/reviewer.
ASIN	ASIN (Amazon Standard Identification Number) is a 10-character alphanumeric unique identifier assigned by Amazon.com for product identification.
reviewerName	The name or username of the reviewer. Defaults to 'Amazon Customer'.
helpful	Whether or not the review was helpful. Represented by a list with 2 elements [0,0]. The first element is the number of users who thought the review was helpful and the second element is the total amount of votes.
reviewText	The main review text.
overall	A rating given between 0-5.
summary	A summary of the review

unixReviewTime	The time the review was submitted on in Unix time. Unix time is defined as the number of seconds that have elapsed since 00:00:00 Coordinated Universal Time (UTC).
reviewTime	The date the review was the submitted on in the format MM/DD/YYYY.

Table 1: Review data

A second dataset which contains the metadata of the products will also be analyzed. Its structure is similar to the review dataset but instead each line represents a given product.

Attribute	Description
ASIN	ASIN (Amazon Standard Identification Number) is a 10-character alphanumeric unique identifier assigned by Amazon.com for product identification.
title	The name of the product.
price	The price of the product in US dollars.
imurl	A url of the product’s image.
related	The related products which share a relationship such as also bought, also viewed, bought together, buy after viewing.
salesrank	Used to represent how well a product is selling; a higher number means less sales whilst lower number means product is selling well.
brand	The name of the product brand.
categories	A list of categories that the product belongs to

Table 2: Product data

An important part of the analysis process is finding the right attributes. Distinct attributes need to be selected so they can produce useful and coherent results later on. The following lists show the attributes that are going to be used for the analysis. These are attributes strictly used to help with finding trends and patterns during graphing which means it does not include attributes that may be used to help calculate them.

Reviews’ attributes	Products’ attributes
helpful	price
overall	salesRank
reviewText	categories

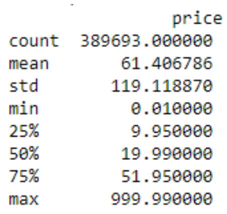
reviewTime	related

Table 3: Selected attributes

The entire dataset will be used to gather information from these attributes.

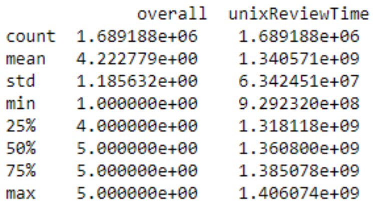
Given a numerical distribution the questions of central tendency and variation can be answered. Many different measures can be used to represent the center of data such as mean, median and mode whilst the variation can be expressed by finding the range, standard deviation or variance. Luckily pandas provides a useful function called *describe()* which uses most of these techniques on all numerical attributes within the dataset.

Using the function on each dataset we can get some useful descriptive statistics:



```
               price
count  389693.000000
mean       61.406786
std       119.118870
min         0.010000
25%         9.950000
50%        19.990000
75%        51.950000
max       999.990000
```

Product



```
            overall  unixReviewTime
count  1.689188e+06    1.689188e+06
mean   4.222779e+00    1.340571e+09
std    1.185632e+00    6.342451e+07
min    1.000000e+00    9.292320e+08
25%    4.000000e+00    1.318118e+09
50%    5.000000e+00    1.360800e+09
75%    5.000000e+00    1.385078e+09
max    5.000000e+00    1.406074e+09
```

Review

Range isn’t included but it can easily be calculated by subtracting the *max* value with the *min* value. These results can help by giving us the average rating amazon customers tend to give as well as product price averages and ranges.

<u>Data Preparation and Preprocessing</u>

Pandas was used to load the dataset using a function called *read_json(“”)* whereby a JSON file can be specified within quotes. An important parameter is used called *lines=True* which tells pandas to read the file as a JSON object per line (due to each object/review being on a different line). Any variable that is assigned to this read function is a pandas DataFrame which is a two dimensional size data structure with labeled axes. Special functions can now be applied to this variable to manipulate the contents and structure of it.

In order to be able to work with the datasets properly they need to be cleaned first. This process involves searching for missing, erroneous, inconsistent, irrelevant, or malicious data and ensuring the formatting is consistent throughout. First off, for the reviews dataset a check was made to find the sum of missing values in each column and it was found that *reviewerName* had over 24,000 missing values. There are several ways to deal with this missing data such as creating new classifications, interpolating based on existing data (although you can’t do it for strings) or omitting it. Since *reviewerID* already exists which uniquely identifies a given user, the *reviewerName* attribute is irrelevant and will therefore be removed. Secondly, the *unixReviewTime* attribute is another irrelevant piece of data because *reviewTime* already exists. *ReviewTime* is kept over *unixReviewTime* due to being much easier to handle during graphing, as the latter would require more computation due to calculating the specific date. *ReviewTime’s* data type and format were also changed. Initially it was stored as an object and followed the format of MM/DD/YYYY; this was changed to be of type datetime and the format was changed to YYYY/MM/DD. An additional parameter was specified called *errors='coerce'* which automatically sets invalid parsing to NaT (not a time).

The table below shows that there is very little missing data within the reviews dataset:

Attribute	Missing data
Asin	0%
Helpful	0%
Overall	0%
reviewTime	0%
reviewerID	0%
reviewText	0%
reviewerName	1.46%

Table 4: Missing data percentage in reviews data

For the product dataset the first piece of irrelevant data found was the image URL. This attribute will be removed as it offers no real use to the analysis process and scope. When checking for missing values it was found that the product dataset contains much more missing data:

Attribute	Missing data
Imgurl	0.035%
Description	7.77%
Title	1.40%
Price	21.77%
Sales Rank	74.16%
Related	26.34%
Brand	71.39%

Table 5: Missing data percentage in product data

Even though large portions are missing from several attributes (such as Sales Rank); the data that remains can still be used to extract useful information which is why these weren’t inherently removed. As for dealing with this missing data, it will simply be ignored and the useable data will be analyzed.

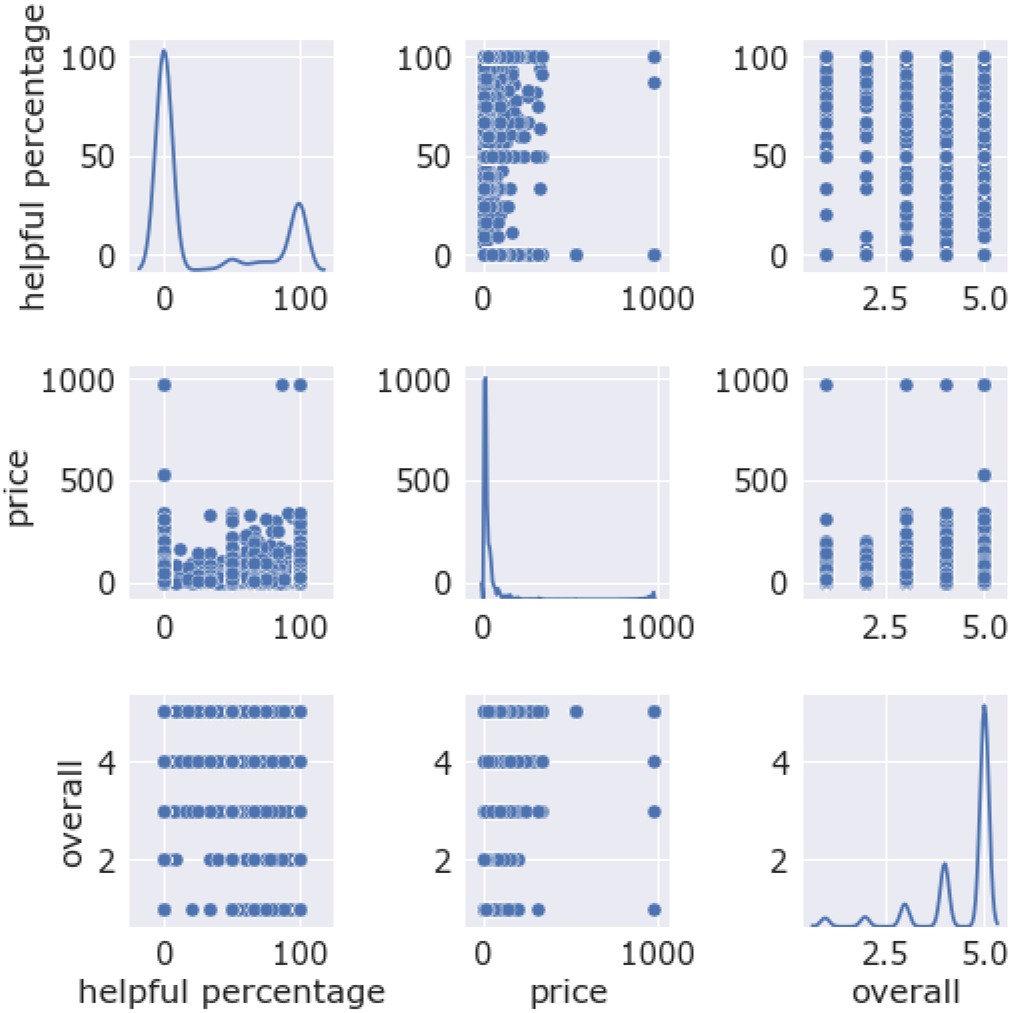
Hypothesis

The expectation of this analysis is to try to find out whether or not the trends and relationships between Amazon consumers and the purchasing of automotive products are positive. By digging deeper, problem areas can be identified and recommendations for improvements to products can be made. Implications regarding the consumer and products can also be identified and plans for improvement in marketing can be made.

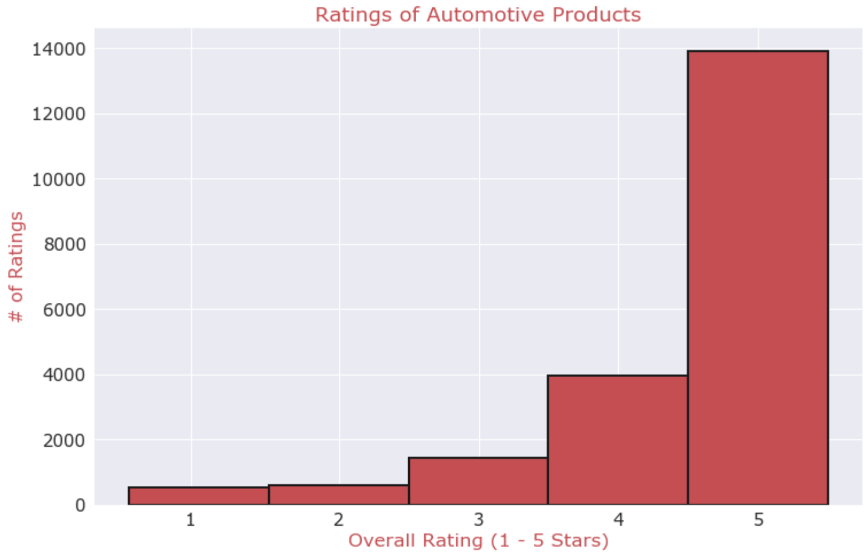
PART 2

In this part, we will do further analysis on the reviews data and perform exploratory, statistical and predictive data analysis based on the given information.

First, pair plot was used to take a look at the relationship of three numerical variables: price, overall rating and helpful percentage. Specifically, helpful percentage was a new column created to measure the helpfulness of each review.



From the distribution plots and scatter plots it’s not hard to find that these three variables had a highly skewed distribution which meant most of the values concentrated at edge values. In addition, the relationship between each variables and others seemed not strong and varied dramatically. Given the information above, we had a basic understanding of the data set. To dig into the data deeply and find the hidden pattern, more data analysis approaches need to be employed. In the following part, we will integrate and retrieve some of the data and focus on the reviews analysis.



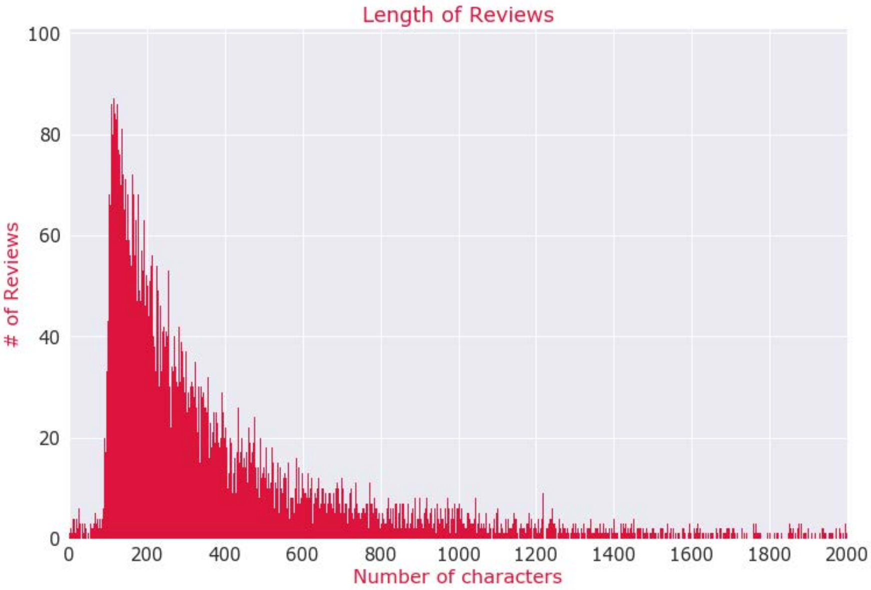
We took out the overall ratings from the data set to look carefully into the distribution of them. People who bought automotive products seem quite satisfied with what they have purchased because most of the ratings are 5-star, which leads to a left-skewed distribution.

Reviews analysis

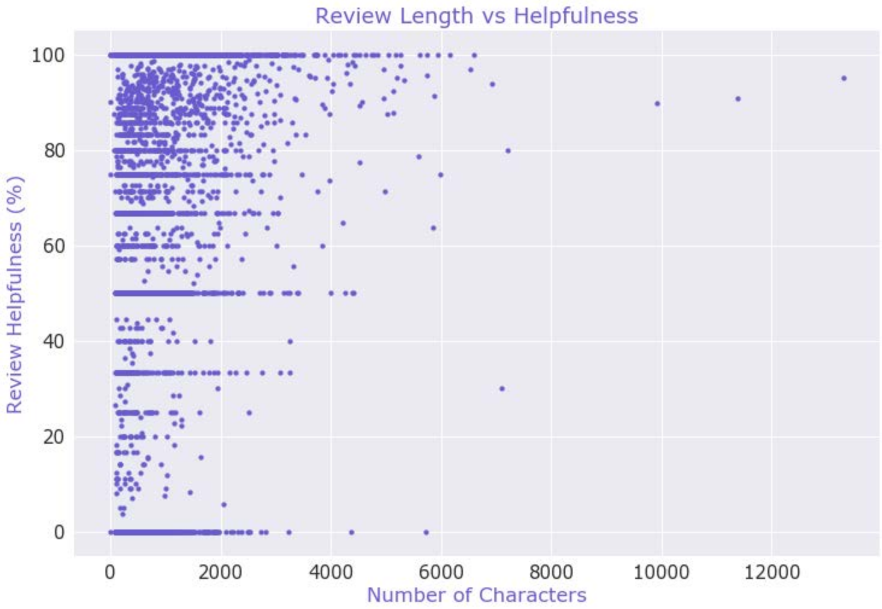
There are many ways to do a reviews analysis, here, we chose to start exploring them by obtaining the length and frequently used words of the reviews which can give us a general idea of the content.

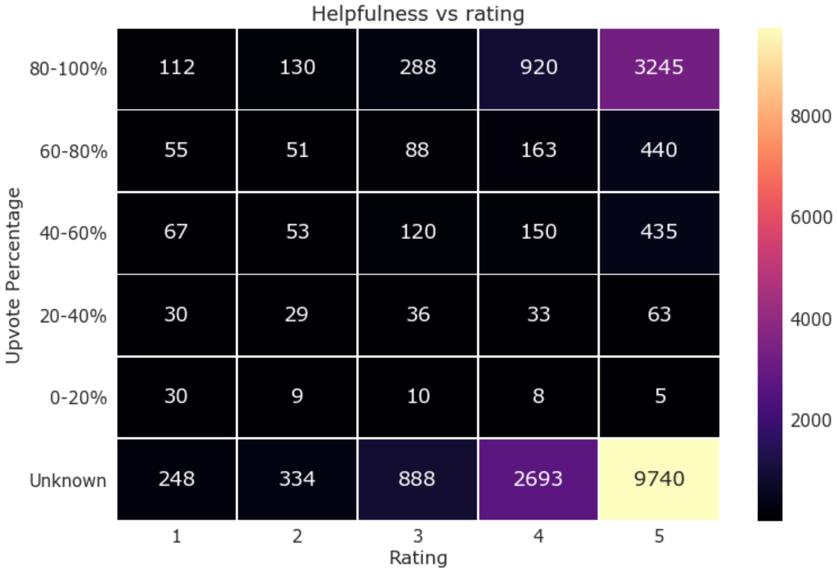
How long is every review?

Do people normally write long essays about products, or do they stick to short concise one liners?

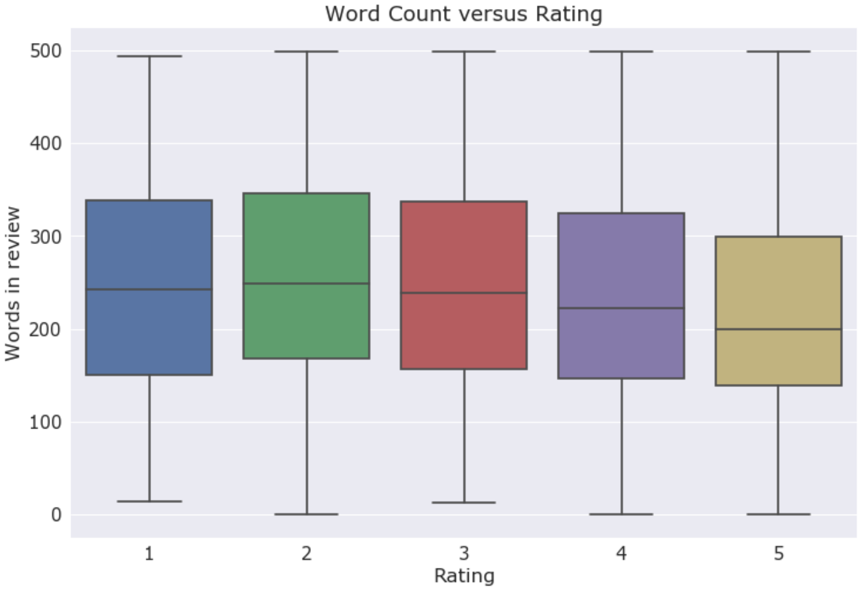


At the beginning, we used histogram to show the distribution of the number of characters in every review. From this right-skewed graph it can be inferred that a majority of reviews are within 100-600 characters which confirms people typically write a sentence or two about automotive products. However, a few people may write a review using 2000 characters which shows they are quite serious about the product comment part.





Since the length of the reviews may directly influence the quality of the reviews, which can be reflected by the helpfulness percentage. It is necessary for us to put them together to see their relationship. Thus another scatter plot was used here. It's shown in the purple plot that dots distribute in an inverted triangle shape. Generally speaking, most of the reviews are helpful regardless of their lengths. However, a trend can be clearly seen that the more characters in a review, the more helpful it may become. Higher length in reviews generally leads to higher helpfulness (positive correlation). To study the helpfulness regards to ratings, a heat map was used to show. It can be easily found that large proportion of reviews have no helpfulness. People tend to agree with 5 star ratings which implies that there is a positive relationship exists between the rating and helpfulness.



To look into the review lengths in every rating, we used a box plot to show the result. 5-star reviews have lowest median word count of approximately 200 words. People who give higher rating tend to write a shorter review on the product. Thus, an initial finding of the relationship between review length and review helpfulness can be post here.

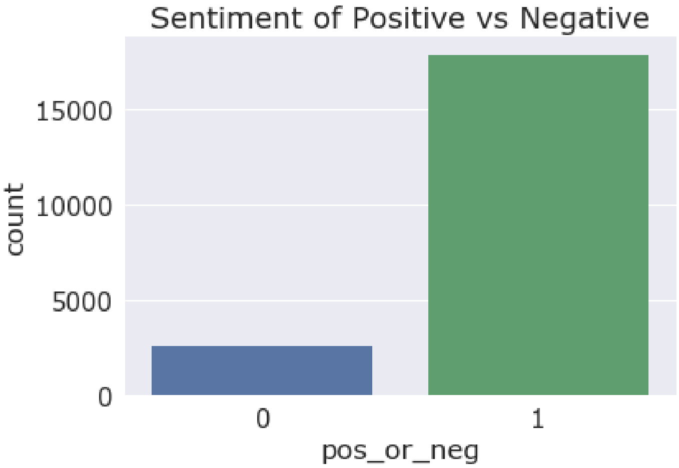
Sentiment Analysis: Frequently Used Words

Using the knowledge obtained from the exploratory data analysis, we can classify a review to be either positive or negative based on its overall rating. A decision needs to be made whether to include neutral reviews or to exclude them. Due to the overall ratings distribution being left skewed, the neutral reviews will therefore be classified under negative reviews.

Using the rules stated above, a binary prediction (which is the sentiment score) can be made:

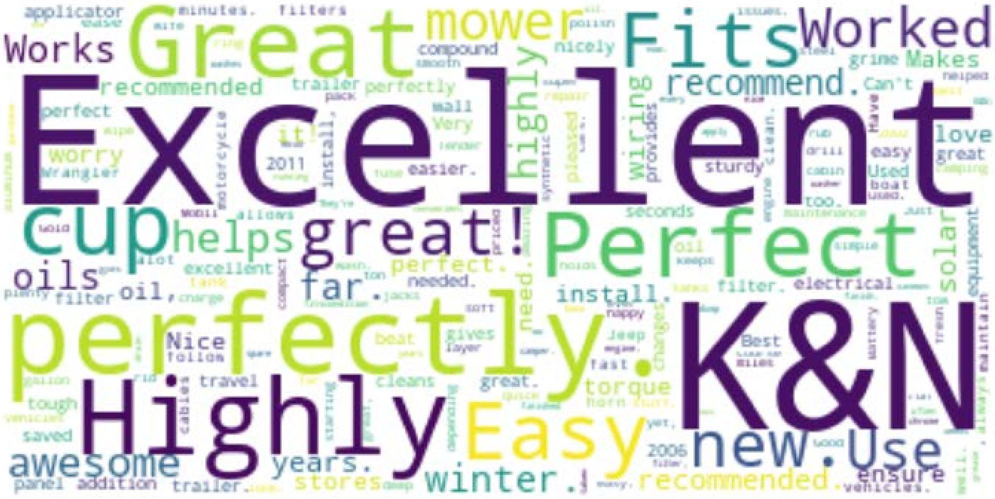
- 1-3 stars is negative ∴ = 0
- 4-5 stars is positive ∴ = 1

A new column is created within the dataset called “*pos_or_neg*” which stores this binary value.
By applying this to every review in the dataset, a general idea of sentiment of products can be graphed:

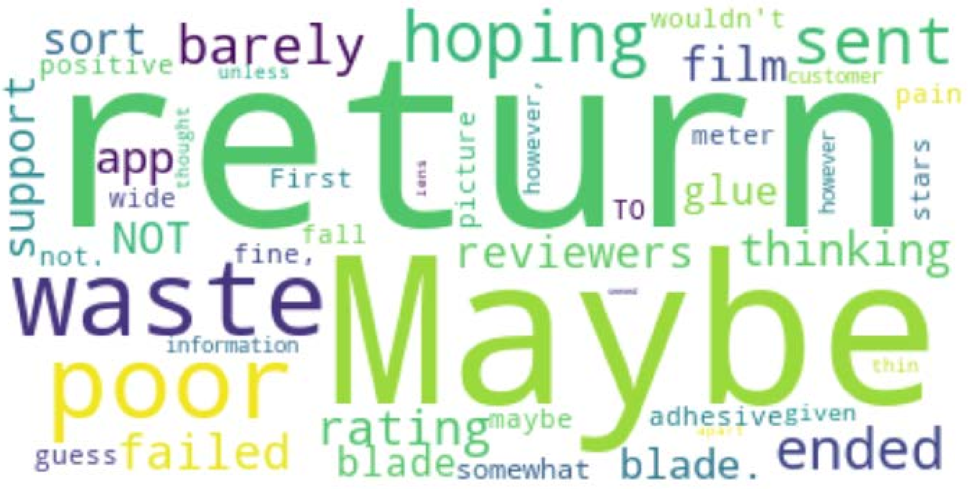


It was found that 80.2% of electronic reviews are positive in comparison to 19.8% negative reviews. Now that the reviews are generalized into positive and negative categories the review text can be extrapolated to find commonly used words in positive and negative reviews. In doing so, a dictionary can be created with commonly used words for both review types. A score is given to each word. This is calculated by getting the frequency of positive word counts and dividing it by the frequency of negative word counts. If this score is high for a given word, then it is likely that it has occurred within a positive review. Therefore words with the highest score are graphed for the positive word cloud whereas the lowest scores are graphed for the negative word cloud. Below are the generated word clouds:

Positive Word Cloud



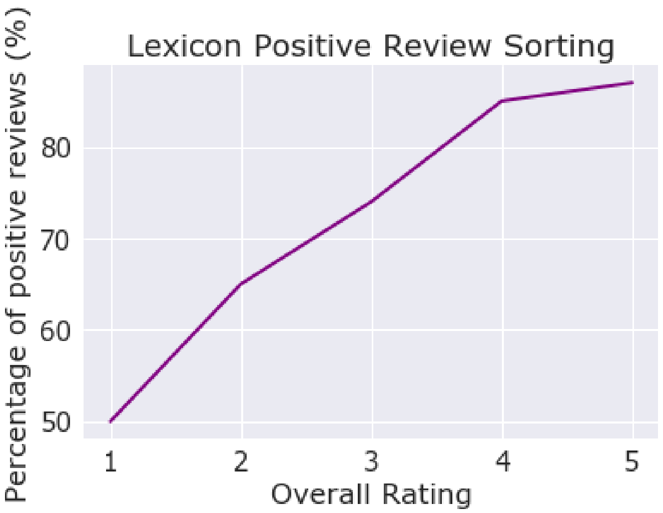
Negative Word Cloud



It can be seen directly from the pictures above that the word cloud includes most of the words which are supposed to occur in a positive review or a negative one. For example, ‘Excellent’, ‘perfectly’, and ‘Great’ can often be seen in a 5-star review whilst ‘return’, ‘poor’ and ‘waste’ are most likely to be used in a 1-star review. The word clouds can retrieve the exact information accurately.

Sentiment Analysis: Lexicon Approach

Troll detection: to classify reviews as honest or not, lexicon sentiment analysis was performed on each review, grouped by the overall rating.



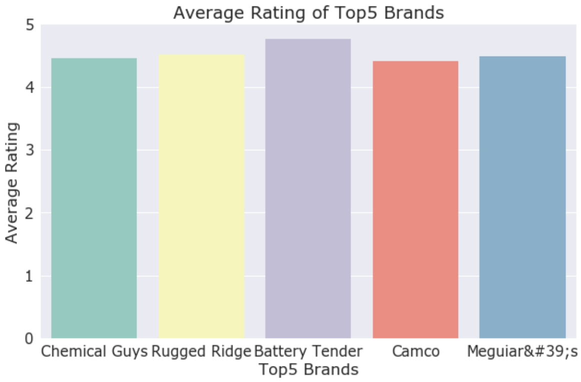
Reviews that gave a high overall score had a cumulative positive text rating of 87%, so it can be assumed they were genuine. What is really surprising is the reviews that gave a very low score as their cumulative positive text rating was 50%. Perhaps not all of those reviews were genuine.

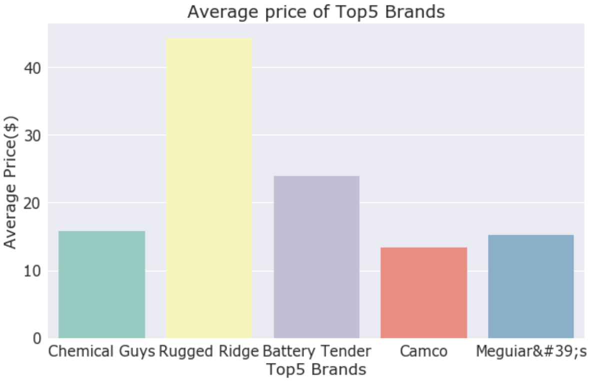
Most Reviewed Brands

When it comes to the frequency of reviews, we can get the top 5 most reviewed brands by grouping the review data by brands and counting the number of reviews on each brand. Here is the table which shows the top 5 reviewed brands.

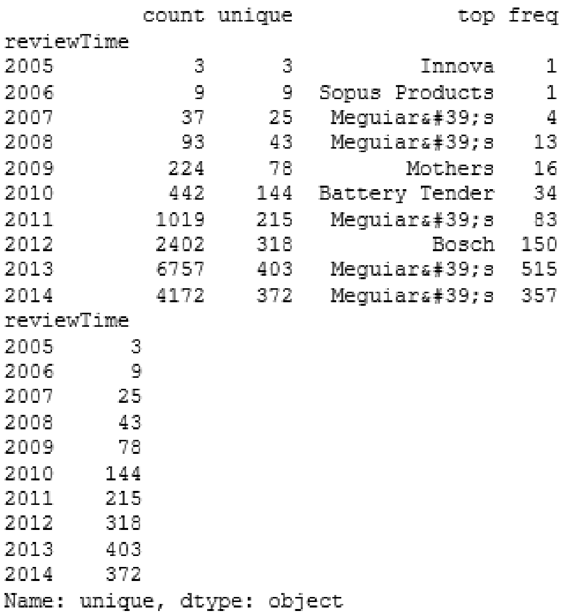
Top 5 Most Reviewed Brands				
5	**4**	**3**	**2**	**1**
Chemical Guys	**Rugged Ridge**	**Battery Tender**	**Camco**	**Meguiars**

Table 6: Top 5 Most Reviewed Brands





The above figures show the top 5 most reviewed brands have high average ratings with Rugged Ridge having over twice the average product price in comparison to Meguiars.



```
           count unique            top freq
reviewTime
2005           3      3         Innova    1
2006           9      9 Sopus Products    1
2007          37     25  Meguiar&#39;s    4
2008          93     43  Meguiar&#39;s   13
2009         224     78        Mothers   16
2010         442    144 Battery Tender   34
2011        1019    215  Meguiar&#39;s   83
2012        2402    318          Bosch  150
2013        6757    403  Meguiar&#39;s  515
2014        4172    372  Meguiar&#39;s  357
reviewTime
2005      3
2006      9
2007     25
2008     43
2009     78
2010    144
2011    215
2012    318
2013    403
2014    372
Name: unique, dtype: object
```

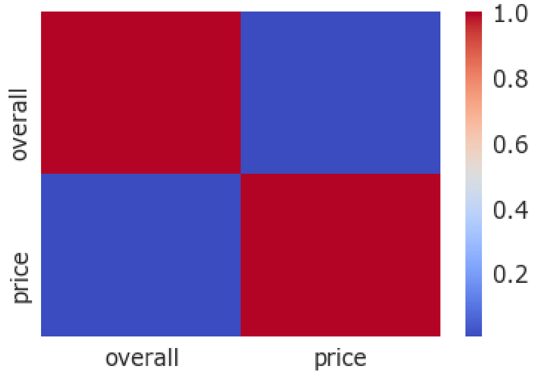
The above table shows the most reviewed brand in each year. It can be seen that Meguiars is always the most reviewed brand, somehow reflecting its popularity among consumers.

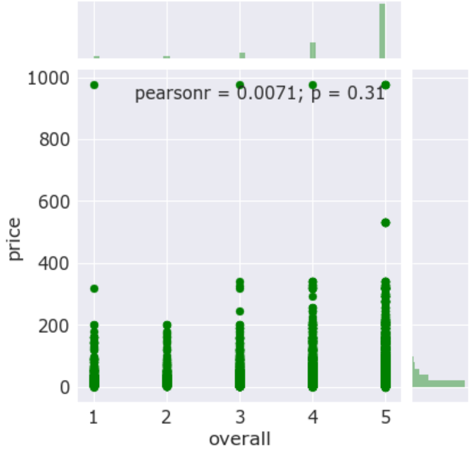
<u>Prices and Ratings analysis</u>

In last part, we dug into the reviews data from different angles. In this part, we will focus on price and ratings to see whether there are some meaningful insights. Also, we will predict the prices in the future using Moving average method and Exponential smooth method.

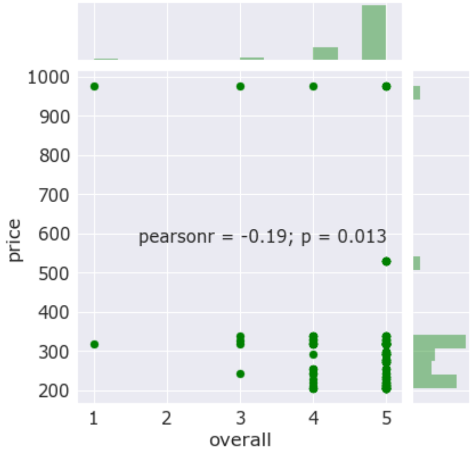
Price vs Overall Satisfaction

Correlation map and joint scatter plot were used here to explore the potential relationship between prices and ratings.





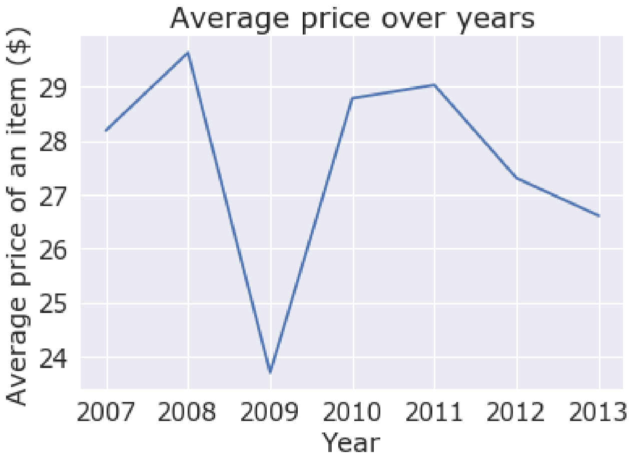
At first glance, the heat map shows no immediate correlation between the price of an item and the overall satisfaction of a customer. The joint scatter plot shows the same with a p value > 0.05 and Pearson's relationship value far from 1, which indicates there's little correlation between the two variables. Product price is not the vital factor to its rating. To look at it carefully, we picked the products with whose prices more than $200, then used the joint scatter plot once again. Here's the result.



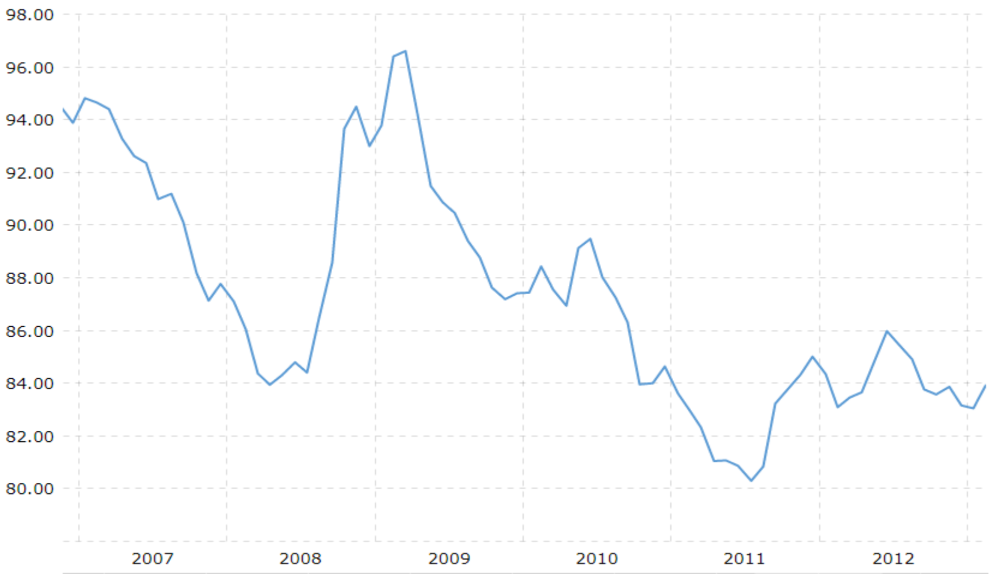
Very few items over $200 are given low ratings, with the 1, 2 and 3 rating columns being almost empty of expensive products. The above scatter plot proves a correlation exists between the two variables with a p-value < 0.05, although the Pearson's r value is only -0.19. From this we can conclude that expensive items leave customers more satisfied than inexpensive items. That being said, purchases over $200 are rarely rated as bad.

The Great Recession

When plotting the average price of the items over the years, we noticed that there's a sharply drop in 2009, which may be related to the Great Recession. The Great Recession was a period of general economic decline observed in world markets in 2009, the effects of this economic decline can be observed in the below plot.



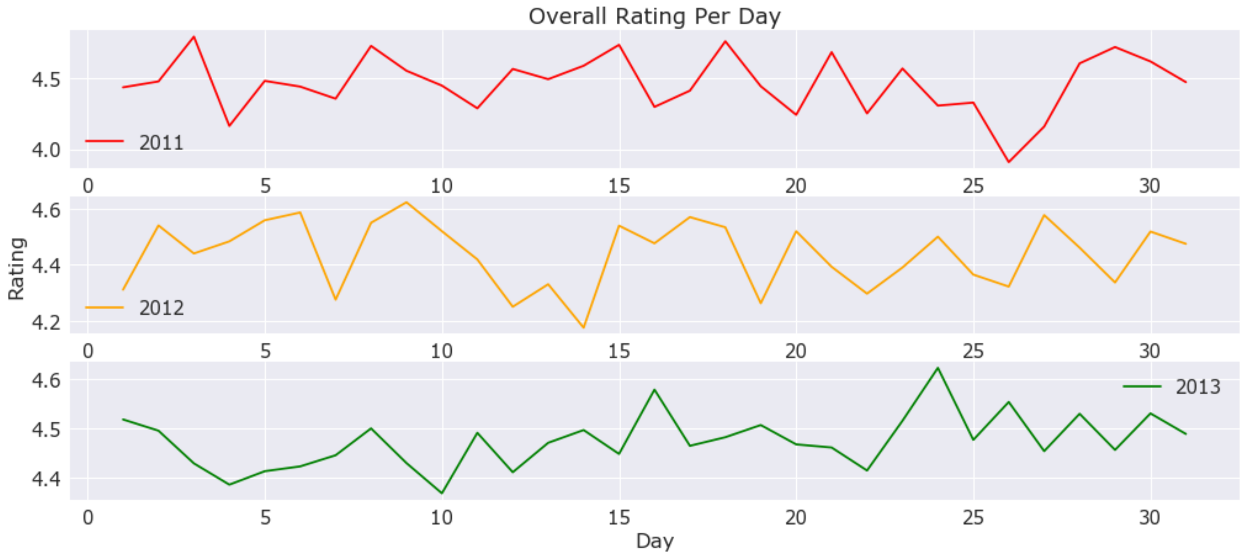
Clearly the financial crisis affected the prices in 2009, but maybe there's something else made all the prices go down, perhaps the US dollar index changed in 2009?

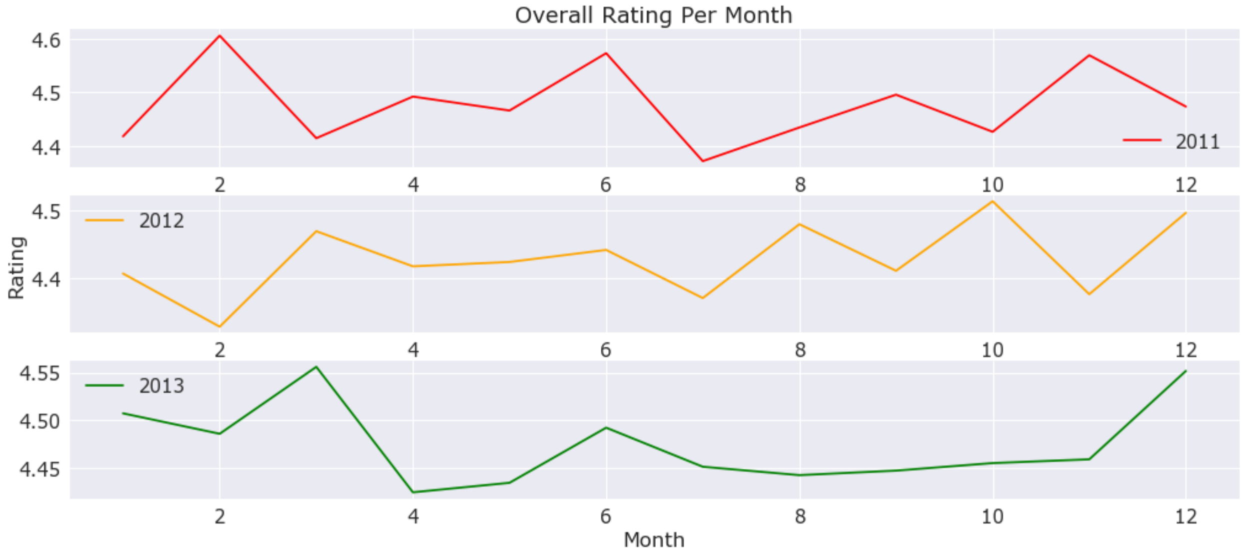


It did! The US dollar index peaked in 2009 from the given year range of the amazon reviews, this is another explanation for the sudden drop in prices.

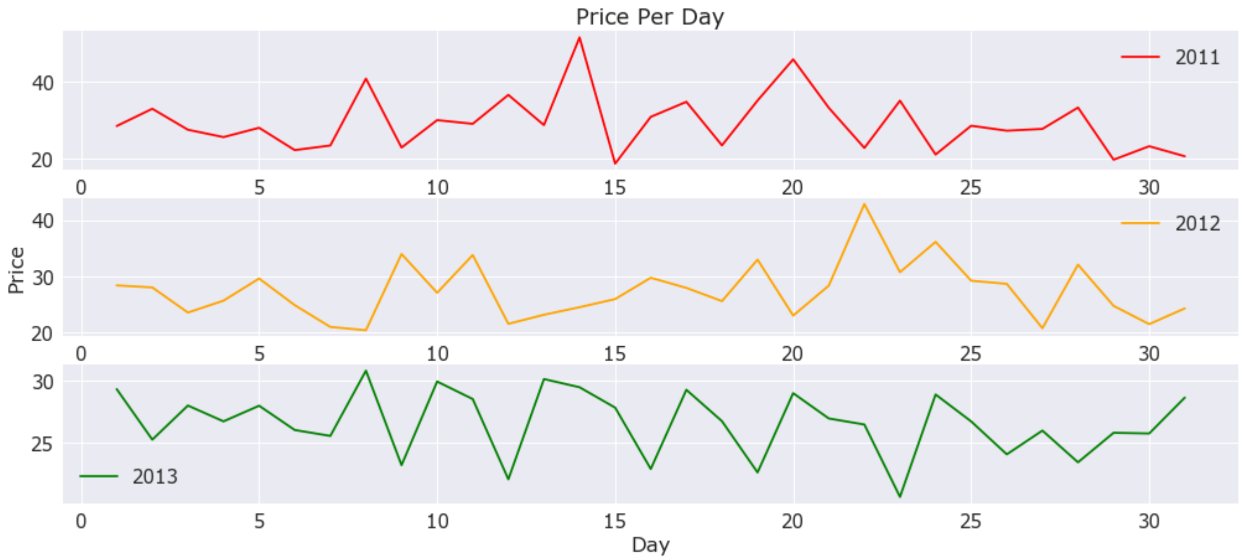
Prediction

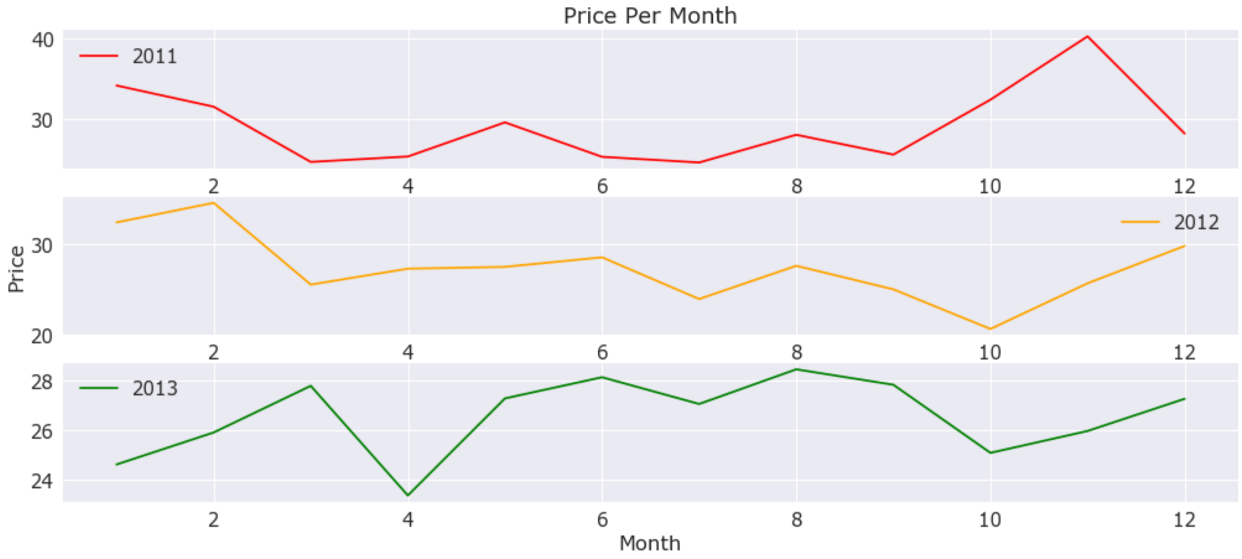
In the following part, we will predict the prices based on the given data and compare the accuracy of different approaches such as the Moving average and the Exponential smoothing. To do the prediction, first we need to find out the exact pattern of the prices. We simply plotted them with line charts.





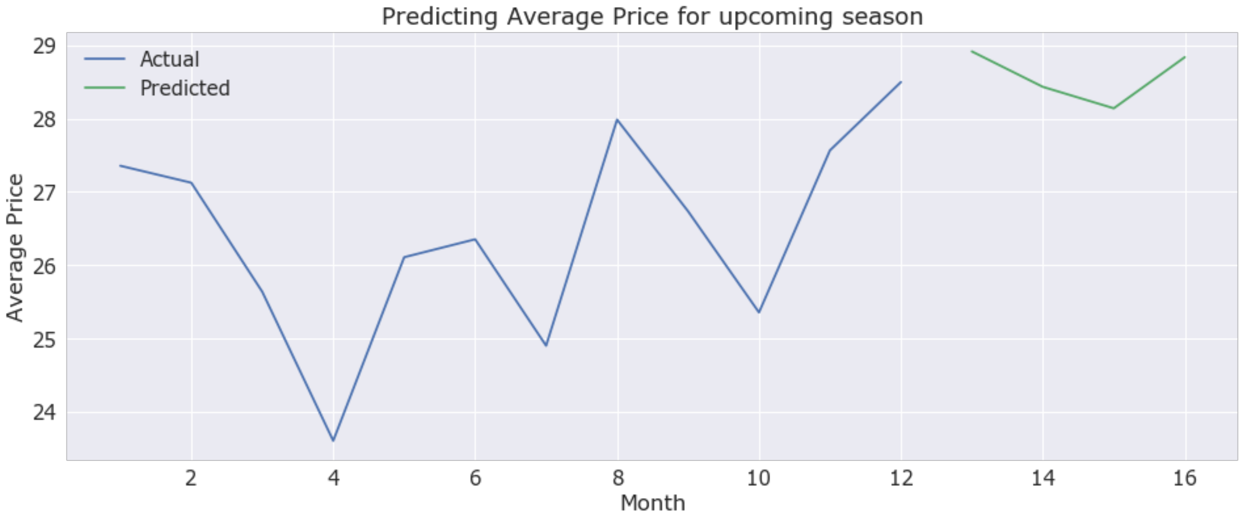
We used subplots to show the trends in 2011, 2012 and 2013. The average ratings per day and per month are quite similar which fluctuate around 4.5. Knowing that the 2013 data is the most credible, it is clear that most customers who have purchased and reviewed their products (assuming the reviews are done within 1 month of purchase) in March, July and December are the most satisfied. Need further proof? The 2012 data shows the same customer satisfaction peaks, along with an added few peaks that can be explained as random variance due to the nature of the size (fewer 2012 data than 2013 data).

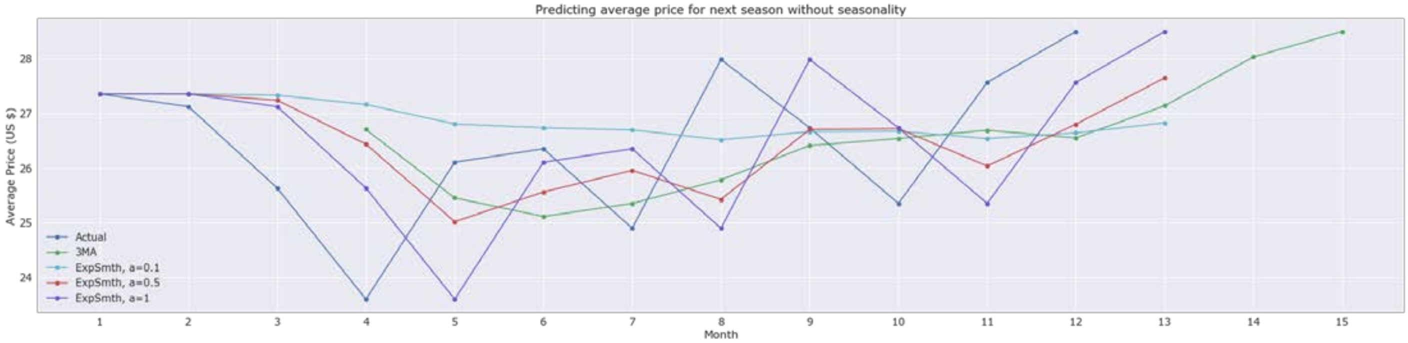




When it comes to the price, the figure above shows the average reviewed price per day and per month for a given year, the 2013 plot holds the most reviews and it is the most recent year reviewed thus it's curves are the most reliable. It can be concluded that between the 5th day and the 25th day of every month, the price fluctuates every 3 to 4 days.

Based on the above information, we can do the price prediction as follows:





The first plot is the prediction with the seasonality, it predicts the average price for upcoming season. It can be seen that the predicted trend line here is reasonably going down and down. The second plot shows the results of the predictions using different methods and different values of parameter. The best one we can see is the exponential smoothing with a=1, which captures trends perfectly but with delay.

PART 3

Findings

The most insightful findings of the data were related to the price and overall quality of products. To be able to understand if or when a product should be bought is very beneficial knowledge to everyday consumers. Despite the data repeatedly showing no mathematical correlation between the price of a product and the customer satisfaction, some predictable and some surprising patterns were found. Predictably, the Christmas holiday period showed great discounts year after year. Unpredictably, days within the 5th and 25th day of every month showed seasonality in average prices (recorded prices from reviews), with prices fluctuating every 4 to 5 days in the same pattern.

An important event impacted 2009's average product price and this sparked the need for further investigation of the year's major events. A further link was found between the USD Index of 2009 and the automotive product prices. With the prices explained, visual correlations were uncovered from a rating per month line chart that described the average monthly rating of each high performing years (years with the most reviews). A relationship in spikes between two recent years of data proved that March, July and December periods were best for high quality buys, assuming all the reviews were completed within a month of purchase. These overall rating numbers were further increased with lexicon review sorting that proved (assuming accuracy) not all negative reviews were negative.

Improving Products

To improve the quality of the data that describes products both in terms of quality (customer satisfaction) and price, more data is needed. This will ensure future questions about when a product should be bought and if it should be bought in the first place are answered with higher credibility. The best time of the month to buy products (as shown in the story document) would be further improved if there were more years recorded, only 2 years have a substantial amount of reviews.

Refining Data Analytics

The relevance of the data is extremely important for optimization, productivity and building relationships with consumers. One way more relevant data can be obtained is by introducing customer surveys which can cover how they feel about their products. However this approach can be cost ineffective and slow in receiving feedback. Perhaps a better way is by harvesting real-time customer behaviour to identify trends and relationships, that way predictions or improvements can be made for consumers very quickly.

A question of verifying the uniqueness of the automotive data arises: how can we ensure a trend is part of the norm or is problematic? We can confirm this by comparing the automotive section with other product sections. By comparing trends from multiple sources it can enable underlying problem areas to be easily distinguishable.

Implications for consumers

The main implication for consumers of the products is the lack of data for automotive parts, as stated in the improving products section a larger dataset would (likely) have larger insights to extract. The lexicon review sorting showed there can be an implication for the consumers as some of the reviews they read may be misleading and completely dishonest.