Problem Chosen	2020 MCM/ICM Summary Sheet	Team Control Number
C		2001866

Data Analytics For Sunshine Company

Summary

For assisting Sunshine Company to sell the new products, models are used to analyze data including customer-supplied ratings and reviews in past which associated with other competing products.

In the first model, a model was constructed to relate the "**real star rating**". Overviewing the data, it deserves attention that the satisfaction of customers cannot be shown objectively through the star ratings of customers because of different standard. For judging the star ratings of each kind of products objectively, the model "real star rating" function is built. With the function, different standards involved in the original star ratings are corrected in the real star rating system, which can show the satisfaction of products objectively.

In the second model, a model was fit to show **how is the satisfaction of products changing over time**. The model shows the time-based measures and trends which manifest the changing of products' reputation and popularity of products' design in the online marketplace. Furthermore, based on the measures, the model also shows us the future trends of marketplace, which can assist Sunshine Company in making better preparations for future marketing.

In the third and fourth model, model was built to **combine text-based measures and numerical measures** for research purpose. With the model which combines reviews and real star ratings, it is easier to analyze and study the keywords with high frequency. Therefore, we can find out the welcome products' properties and unwelcome products' properties. With the model which is combing reviews and number of total votes, it is easy to analyze which types of specific reviews are more attractive and strongly associated with rating levels. Combining text-based measures and numerical measure, models are working in a new frontier.

In the final model, model was set to **options influence on customers**. For studying if low star ratings of a product will induce lower ranking and more negative reviews, it is necessary to research the trend of percentage changing of positive reviews and negative reviews on a specific product. We adopt the notion of Markov process and obtain results agreeing with our assumptions. As the evaluations are somewhat screened by the randomness related to considerable number of customers, it is only appropriate to introduce the modification in the first model in order to loyally reflect the satisfaction of the products.

Contents

1 Introduction

1.1 Background

"Without big data analytics, companies are blind and deaf, wandering out onto the Web like deer on a freeway."

When the author Geoffrey expressed his opinions in 2012, it may have been perceived as an overstatement. However, it is universally acknowledged that nothing is more important than data in marketing. Now, big data is universally accepted in almost every vertical, not least of all in marketing and sales. Data-driven marketing refers to strategies built on insights pulled from the analysis of big data, collected through consumer interactions and engagements, to form predictions about costumers' preference, competitiveness of products, future behaviors and more.

1.2 Issue

As the world's largest online marketplace, Amazon provides various customers with an opportunity to rate and review purchase. For each bargain, purchasers could express their level of satisfaction to the products, in term of numerical messages (star ratings) and text-based messages (reviews). Apart from them, mutual communication between customers is also promoted by setting helpfulness rating function for others' reviews. With these data, it is closely related to the properties of products, which can help to find the potential success of product design and supply better development of the online marketplace companies.

For assisting Sunshine Company to introduce and sell the three new products (microwave ovens, baby pacifier and hair dryer) in the online marketplace, models are used to analysis the data including customer-supplied ratings and reviews in past which associated with other competing products.

The five various model in this paper, are built for solving different problems and studying various situations in the online marketplace. Through combing and analyzing different types of data, we believe the studying of online marketplace will become deeper.

2 Model 1: the satisfaction of customers toward a product

2.1 Background

Upon seeing the data sets, one is very likely to stop at the star ratings, as we usually do when visiting Amazon. I believe many will wonder the same question, "what products make good sales?" This is exactly where we start from, but with a quite different aspect.

Do the star ratings honestly reflect the satisfaction of the products? Probably not, as one quickly discerns that some reviews have higher "helpful votes" than others. Still, for those who are members of Amazon Vine Voices, they are reputed for writing accurate and insightful comments, whom we have good reasons to trust in. However, as for the people who buy products with deep discount, it is not impossible that they tend to give higher ratings in return to the courtesy they've been granted. So, we set out to devise a model aiming to eliminate the possible subjective factors and recover the ***"real star rating"*** that the product deserves.

2.2 Model construction

The real star rating we're going to talk about is a derived number that we believe can represent the real, unprejudiced, satisfaction of the customers toward a product.

DEFINITION: *fidelity factor f*

The ***fidelity factor*** f of a review is a real number within [0,1]. f=1 represents that the review is entirely trustworthy, without a shadow of doubt. Contrarily, f=0 means just the opposite and has no value to be referred to.

We conclude there are four factors that affect the fidelity of reviews. They are the *total votes*, *helpful votes*, *"vine"*, and *"verified_purchase"*. Among the total votes, people who consider this review helpful contribute to the helpful votes. Moreover, people who is a member of Amazon Vine Voices will have vine=Y. Still, those who purchase the item with a deep discount will have verified_purchase=N.

The original star ratings are in the range between 1 to 5. We then decide to choose 3 as the zero point, so that ratings lower than 3 become negative, and those greater than 3 remain positive. This is called the *shifted star ratings*. We then multiplied this shifted star ratings with our judiciously determined fidelity factor f to obtain numbers which represent the reliable reputation of the product (called *shifted real star ratings*). Finally, we shift the numbers back by 3, so that the modified ratings still lie in between 1 and 5, which will become the ***real star ratings***.

We first consider the case that the reviewer is not a vine member and examine the factors of the total votes and helpful votes. We shall use k as the modification factor concerning these two factors. k again lies in between 0 and 1. A naïve guess is to use the fraction helpful votes/total votes to represent the fidelity as far as these two factors are concerned, but we soon discover that the total votes vary in a big range and should have different level of importance. In particular, there are troubles for those extreme cases of total vote=0 or 1. So, we devise the following form of k as a function of helpful and total votes.

$$k = \begin{cases} a_1 & if\ total\ vote = 0 \\ \dfrac{helpful\ votes}{total\ votes} \cdot a_2 & if\ 1 \le total\ votes < 10 \\ \dfrac{helpful\ votes}{total\ votes} \cdot a_3 & if\ 10 \le helpful\ votes \end{cases}$$

where a1, a2, and a3 are parameters within [0,1] to be determined through analyzing the data.

As those who do and do not receive deep discount may evaluate the products based on different standards, we define the modification factor g to treat this difference. Let b1 and b2 be two parameters lie with [0,1] which are determined through further analysis.

$$g = \begin{cases} b_1, if\ verified\_purchase = Y \\ b_2, if\ verified\_purchase = N \end{cases}$$

We then combine the two factors k and g by defining the fidelity factor f to be f=0.5*(k+g). In addition, if the reviewer is a vine member, we stipulate that k=g=1, and therefore f=1. We put our entire trust on the vine members who are reputed for their accurate and fair reviews.

2.3 Implementation and Results

For all hair dryers, we believe that the parameters a1, a2, a3, b1, and b2 are the same for all products and do not depend on brands or types. However, it is appropriate to assume that different products, may have different parameters as the different ways the discounts are presented have quite different effects on customers' reviews.

ASSUMPTION

Suitable sets of parameters a1, a2, a3, b1, and b2 reduce difference between the average *shifted real star ratings* of non-vine members and the average shifted star ratings for vine

Based on this assumption, we still impose some other constraints on the parameters. We tried different constraints and obtain the fittest parameter set, which enables us to calculate the average real star ratings for all items, and compare them on an equal footing.

$$\frac{\sum(star-3)_{vine}}{|vine\ review's\ number|} \approx \frac{\sum[(star_{not\ vine}-3)\times f]}{|total\ review - vine's|}$$

(for each type of product)

For example, we list pacifiers with the top and bottom 5 real star ratings.

product_parent	Real star rating	product_title
The top		
800875796	5.0	lorex bb2411 2.4 sweet peek video baby monitor
758355808	4.8246	god kept his promise and brought you home lifelike newborn baby doll: so truly real -10 by ashton drake
155035926	4.6656	baby's quiet sounds - video monitor
266017252	4.6656	fareskind the easy liner, 0-36 months
229222458	4.6656	silicone teething necklaces with baby-safe jewelry by favefemme - bpa-free, best soothing method, better than baltic amber, teething necklace for mom
The bottom		
61825188	1.3344	Fisher45;Price Aquarium Monitor
135570229	1.3344	lullaby light cube soothing star projector- portable travel soother and musical night light
800824853	1.3344	gerber prefold premium 6-ply cloth diapers, 5-pack
929565278	1.3344	safety 1st boost air protect booster car seat, dixie
699910045	1.3344	luvable friends 7 piece drooler bibs with waterproof backing

Table: 1

For the other two products, please refer to Appendix A.

3 Model 2: Time analysis combined with real star ratings

3.1 Background

Armed with the previously derived real star rating, we are able to process the reviews on the same footing. As what we often hear, people learn lesson from the history. One of the most interesting things is to observe the change of real star ratings over time and try to extract useful information that would bring us benefits.

3.2 Overview

- For those items that enjoy high (or low) real star ratings, we compute the annual average real star ratings and plot them against time.
- Observe the said graphs and identify those with critical turns or special features.
- Investigate the text-based comments associated with those features.

3.3 Implementation and Results

As an example, we examine the time evolution of top 20 high-real-star-rating hair dryers, which are shown in the following figure.

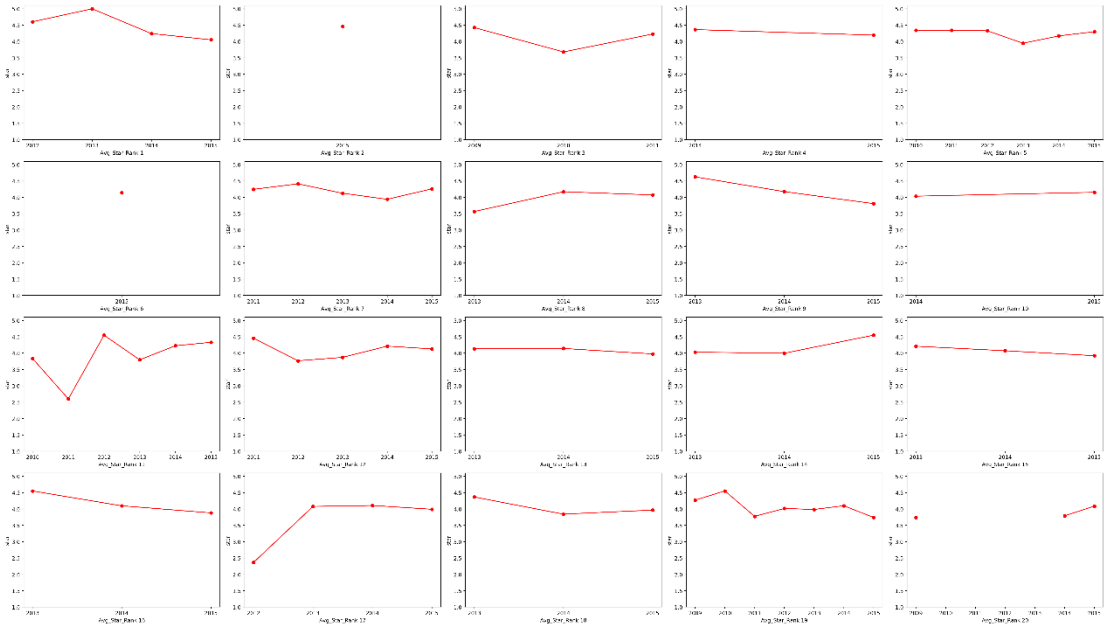


Fig1: 20 best reputed hair dryers (judging from their real star ratings).

Similarly, the graphs of the 20 hair dryers that have the least real star ratings are readily obtained.

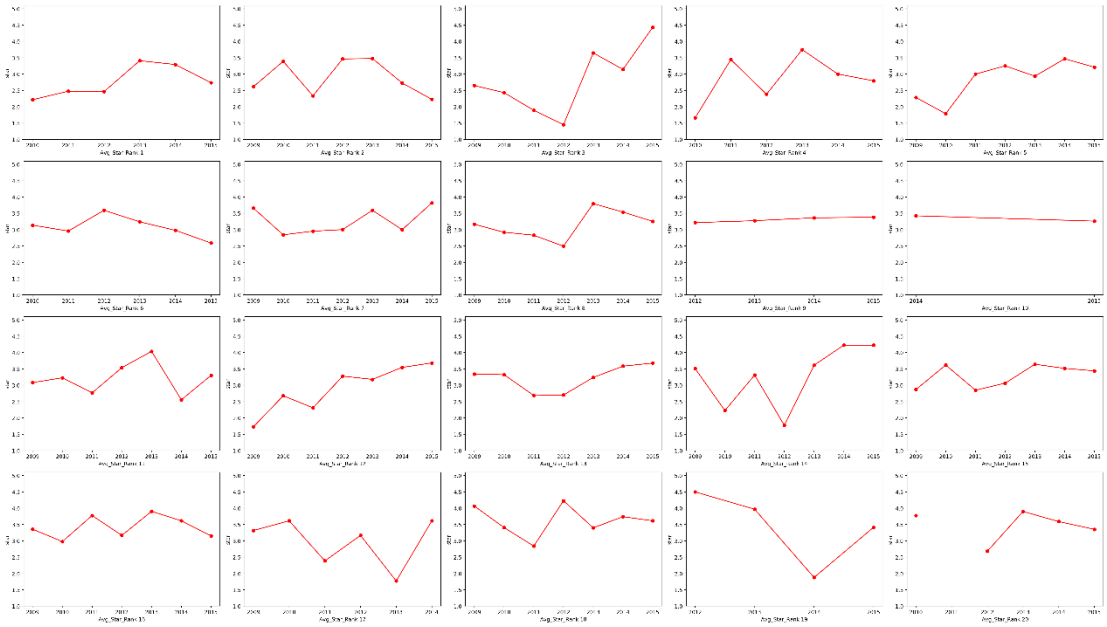


Fig2: 20 most unwelcome hair dryers (judging from their real star ratings). Note that the one on the fourth row and fourth column has a sudden plunge in 2014 and a later recover.

For the other two products, please refer to Appendix B.

Here, we point out a certain product of special interest, which is on the fourth row and fourth column in the figure above. It has fairly high real star ratings before suffering a sudden fall in 2014, but soon recover from its depression in the next year. This intriguing incidence provokes our interest in doing the following case study.

3.4 Case study

We discover that the said product has not only a low real star rating in 2014 but also receives few reviews, part of which are listed below.

Blew up within 4 months I bought it in May and it blew up in August. It was used gently on shoulder length hair for short period of time. Up until the point when it stopped working it was a really good hair dryer. Relatively quiet, high velocity, and effective. 8/27/2014

Was pretty good for 13 months but now dead UPDATED Sept 2015:

This dryer died today after almost catching on fire. I sniffed a burning smell and 2 seconds heavy sparks came out and it died. I only used it twice a month to dry my dogs for about 30 minutes each time. Always cleaned lint out of the air vent. Very pricy to last just 13 months with limited use. Does anyone make a quality product?? his is a terrific dryer for the price here on Amazon. It seems very well built and has a lot of power without being too loud. It has three heat settings and two power settings. I love it. Great for blow drying my two dogs on a cooler setting. 8/20/2014

With regard to a sea of good reviews in the previous and succeeding year, we extract the words that frequently exist in the comments, which are manifested in the following pictures.

(a) (b)

Fig3: Frequently existing words in the reviews of the product in question (a) in 2013 and (b) in 2015.

The mystery of the plunge suddenly become clear. The comments suggest that there are some mistakes or flaws in manufacture which lead to malfunctions and safety incidents. This has severe impact on its quality and account for the negative reviews. Fortunately, the problem is identified and fixed and we can see that praise and compliments do come back in the 2015 comments.

Studying the change in the course of history always shed light on our understanding. So, the general approach is to keep on analyzing all other cases to discover the underlying causes and effects. For example, how a certain product remains high rating over long period of time or why the reputation of some certain item is declining, etc. However, we will stop at this example and move on to the focus of text-based analysis, the idea of which we have already demonstrated in the previous case study and will be further developed.

4 Model 3: Text-based analysis combined with real star ratings

4.1 Overview

Combine all reviews posted in different years and analyze the keywords existing in the titles and comments of the favorable and unwelcome products.

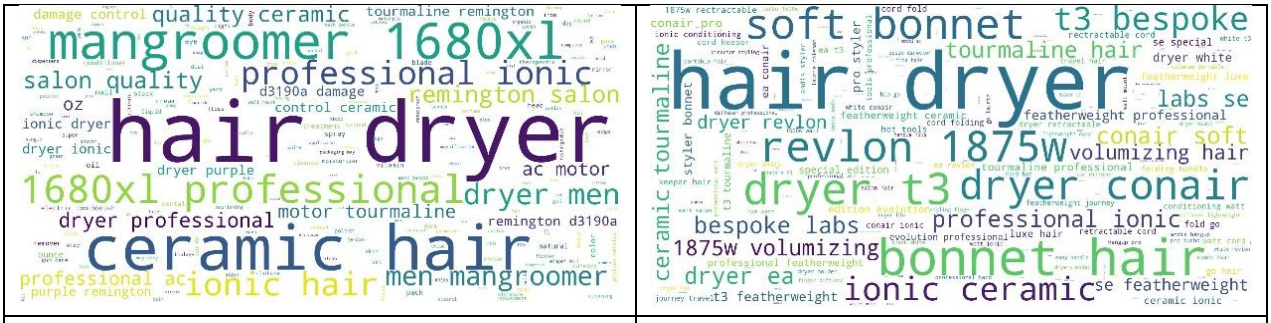
4.2 Implementation and Results

We single out the top and bottom 25% products in terms of its real star ratings before surveying the 100 most frequently appearing words in their product titles and review bodies. After filtering the meaningless words, we summarize the keywords that appear in the product titles and review bodies of the favorable and unwelcome products. The example of hair dryer is given below for example.

Keywords that appears in the **product titles** of the **top 25%** products
brand: Remington, mangroomer, Philips, **model**: 1680xl-6, d-2012, d3190a, **other descriptors in the title**: men,ac, motor, damage, control, purple, Oil, Professional, Eye, Treatment, Care, Natural, Spray, Electric, 100%, Color, Dental, Eau, de, Mirror
Keywords that appears in the **product titles** of the **bottom 25%** products
brand: Conair, t3, tourmaline, Revlon, andis, **model**: 83808-se, 83808, 83888-se, **other descriptors in the title**: bonnet, featherweight, 1875w, soft, bespoke, labs, volumizing, styler, pro, white, watt, handle, evolution, luxe, special, edition, rectractable, cord,, fold, hot, tools, conditioning, 1875, cord-keeper, w/ , retractable, cord, folding, journey, travel
Keywords that appears in the **review bodies** of the **top 25%** products
Heat, dry, long, little
Keywords that appears in the **review bodies** of the **bottom 25%** products
Dry, back, T3, heat, long, cord, years, bonnet

Table: 2

For the other two products, please refer to Appendix A.



(a) (b)

Fig4: Frequently appearing words in the comments of products with (a) top 25% and (b) bottom 25% star ratings. Larger size indicates higher frequency.

4.3 Inferences

We observe some favorable brands and models through examining the list. We find that brands such as Remington, Mangroomer, and Philips are performing well, while T3, Tourmaline, Revlon, and andis receive low ratings. Also, some certain models earn themselves good and bad reputation as listed in the previous section.

It is not difficult to discover the fact that words like “purple” and “men” appear frequently in the top 25% list and others like “white” and “retractable” are commonly seen in the bottom 25% items. This provides us some directions. For example, purple hair dryers rather than white ones are attracting more customers. Also, the products aiming to male customers may make promising sales. Furthermore, it is suggestive that “cord” seems to incur some complaints among customers, which should be treated with more care when designing and marketing.

5 Model 4: Association between total votes and descriptors

5.1 Overview

Based on the number of total votes, ranking the reviews and analyzing the keywords existing in the most popular reviews are carried out in this section.

5.2 Implementation and Results

We single out the top 25% reviews in terms of its total votes number before surveying the most frequently appearing words in review bodies of all products. After filtering the meaningless words, we summarize the keywords that appear in the review’s bodies of the popular reviews. The keywords in reviews are given below:

Fig5

Keyword in microwave ovens (appearing times)	Keyword in baby pacifier (appearing times)	Keyword in hair dryer (appearing times)
Convection (55) Stainless (49) Fit (46) Replacement (38)	Plastic (40) Regular (25) Safety (18) Quality (17) Natural (15) Silicone (15)	Plastic (38) Quality (28) Attachment (27) Frizz (27) Smooth (26) Expensive (26)

Table: 3

It is easy to see that the most frequently appearing keyword in the reviews with largest number of votes are almost objective words. For example, the keywords “plastic” and “stainless” in reviews are describing the materials of products. The keywords “convection”, “replacement” and “attachment” are describing the phenomenon or behaviors while using products. Therefore, it shows that the type of reviews with subjective commenting are more attractive, which can get largest number of votes.

On the other hand, there aren’t any keywords relating to emotion, appearing in the list. It is easy to conclude that the emotional type of reviews aren’t attractive among all types of reviews.

6 Model 5: Consequences of the customers' reviews

6.1 Background

In this model, we want to know how the early buyers' opinions influence the upcoming buyers. Do low star rankings induce more negative reviews or low rankings? Or the opposite? We start from looking into the proportion of positive and negative evaluation on products of every single year, and treat this problem by employing the so-called Markov process.

In this case, we want to observe how the public opinions influence the public themselves. This problem relates only to the reaction of customers upon seeing a high or low rating on the website, and has nothing to do with the real, objective customer satisfaction discussed in other parts of this paper.

6.2 Model Overview

Suppose we perform a certain experiment of which outcomes may turn out to be E1, E2. We now carry out a series of identical experiments. If the outcome of the succeeding experiment depends only on the outcome of the previous one and not on the history of other previous experiments, such process is called a ***Markov process***.

Now, as this process is Markovian, the probability of the occurrence of E1 (or E2) in the next experiment is completely determined by the current outcome along with the probability of the transition from current outcome to E1 (or E2). I shall use pij to denote the probability of transition from a current Ej state to a succeeding Ei state. Now, instead of knowing exactly the present state the experimented system, we consider a more general case that we only know its probability of being in each state, which is denoted by a ***probability vector***:

$$\begin{pmatrix} p \\ 1-p \end{pmatrix}$$

Where p is the probability that the system's in E1 state, 1-p, of course the probability that it's in E2 state.

So, the probability vector (q, 1-q)T of the outcome in the next experiment is given by

$$\begin{pmatrix} q \\ 1-q \end{pmatrix} = \begin{pmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \end{pmatrix} \begin{pmatrix} p \\ 1-p \end{pmatrix}$$

Where pij's are defined in the previous paragraph, and the matrix called ***transition matrix***. Note that it's evident that the matrix elements must satisfy conditions p11+p21=1 and p12+p22=1.

It is common in Markov processes that the evolution of probability vector leads to a limit of the probability vector after sufficiently large number of experiments are performed on this system. We will therefore yield a stationary probability vector determined by the following equation:

$$\begin{pmatrix} p_s \\ 1-p_s \end{pmatrix} = \begin{pmatrix} p_{11} & 1-p_{22} \\ 1-p_{11} & p_{22} \end{pmatrix} \begin{pmatrix} p_s \\ 1-p_s \end{pmatrix}$$

In which the transition matrix effects no change on the probability vector and subscript s stands for stationary. p_s is given by $p_s = \frac{1-p_{22}}{1-p_{22}+1-p_{11}}$.

6.3 Setup

Our rule is that any review with star rating greater than three is called a positive review, while those that are less than or equal to three are called negative reviews. We collect all data of one product in one year, and obtain the proportion of the positive and negative reviews.

If we consider the reviews to be the collective reaction or representation of all customers, in other words, considering the customers as one entity. The proportions can be interpreted as the probabilities of the entity to give a positive or negative review. Therefore, the experiments in our case are just observations of the decision of the customers, the outcomes being the decision of the customers (positive review or negative review).

6.4 Justification

We now provide our justification of regarding the evolution as a Markovian one. Only if some special and significant events should happen, such as the revelation of potential safety concerns or some other extensive outbreaks of distrust to certain products, would the effect manifest itself on the customer reviews on such a long period up to several years. Therefore, it's reasonable to assume that the review of the next year depends only on that of the previous year and some underlying pattern related to customers' behavior.

Under this mechanism, there is a certain probability that positive review in the current year induces another positive or negative one in the next year. In the long term, the pattern gives rise to an ultimate stable probability, which will be our main focus.

6.5 Implementation and Results

In order to obtain the stationary probability, we need to find out the matrix elements first. This is realized through solving the equation

$$\begin{pmatrix} q \\ 1-q \end{pmatrix} = \begin{pmatrix} p_{11} & 1-p_{22} \\ 1-p_{11} & p_{22} \end{pmatrix} \begin{pmatrix} p \\ 1-p \end{pmatrix}$$

for the elements p_{11} and p_{22}, given the data of p and q of two successive years. It requires two equations to identify the elements, that is, by analyzing the evolution of three successive years.

For every three successive years, one determines a specific Markov process, and a stationary probability based on that. We collect all the stationary probabilities for hair dryer and obtain that the mean value is 0.7647, the standard deviation being 0.1545. 83.33% of all data fall within the one standard deviation interval. It shows the fact that the data are highly concentrated and confirms our model.

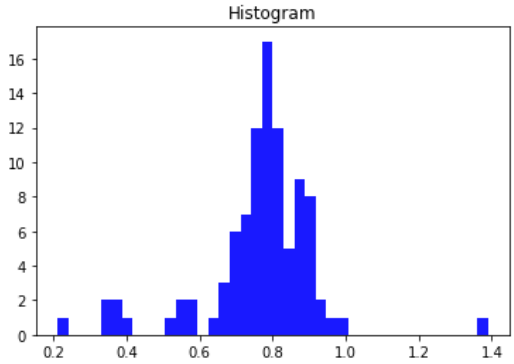


Fig6: The statistical distribution of ps for hair dryer. The vertical axis represents the number of data and the horizontal axis represents ps. The mean value is 0.7647, and the standard deviation 0.1545, 83.33% of all data fall within the one standard deviation interval.

For the other two products, please refer to Appendix B.

The concentration of data within one standard deviation interval about mean value gives the expected scenario of the Markov process assumption. This suggests that the probability of receiving positive or negative reviews from customers depends only on the proportions of the previous year and not on its history. Such result implies that the public opinions develop and change in a somewhat random way regardless of the nature of the products. And, we may only obtain useful information through deliberate consideration of its fidelity as we presented in the previous sections.

7 Conclusions

7.1 Advantages

- Based on the reasonable assumption, the models still impose other constraints on the parameters. This brings us the more objective data through the whole paper.
- Our term of "real star rating" function with five different parameters, which are including all situation during online shopping, shows that our model are reasonable and well-considered.
- For model 2, through researching the data in past changing with time escaping, it predicts the future trends for a reason.
- For model 3 and 4, with combing text-based measures and numerical measures, the research studying will be more diversity, which is different from traditional data analysis.
- Including the mathematical knowledge, the models are built more convincing and reasonable. For example, the Markov process are using in our model 5 to predict the influence of specific star ratings and reviews.

7.2 Disadvantages

- By Markov process, we assume that the star ratings pf products will trend into one specific value for finding influence between customers. In fact, there are two specific star rating occurred in microwave oven which is totally different from our prediction.
- The simplification and approximation taken in the process of going through the models may introduce errors.
- The data number of microwave oven is relative small which may take error in our models.

7.3 Improvements

- In order to simplify the solution steps of the model, we assume some parameters are fixed. In order to consider the model of this problem more comprehensively, a model for parameters are needed for self-correcting the value of parameters.
- With the part finding keywords of products, distractors are appearing in a high frequency like words "the" and "a" and we have to kick them off hand by hand. For better effectiveness, a model is needed for identify the distractors and eliminate them.

8 Letter to Sunshine Company

Team #2001866

March 09, 2020

Mr. X

Marketing Director

Sunshine Company

Dear Mr. X

It is our supreme pleasure to be invited as consultants to analyze the data of three new products in the online marketplace!

After the brief data's analysis of customer-supplied rating and reviews over the time periods, it forms the patterns, relationships, measures, and parameters within and between star ratings, reviews and helpfulness ratings, which will help Sunshine Company succeed in online marketplace hair dryer, microwave and pacifier offerings.

The analysis of data and online sales strategies are given below:

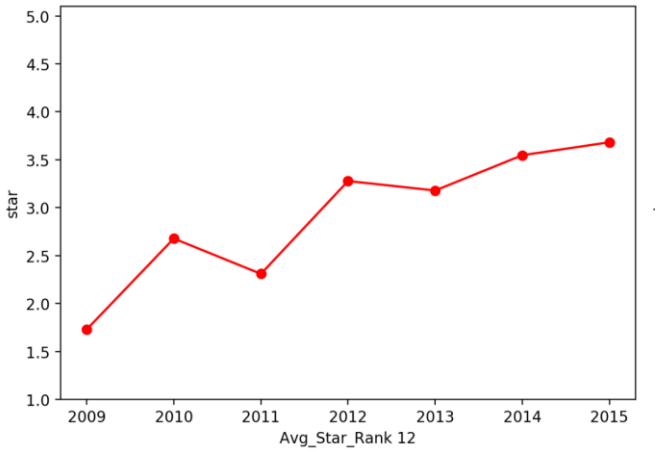
- DON'T think of star ratings of products as the real satisfaction of products. Not only satisfaction of products is affected by star rating from customers but also reviews, helpfulness vote, influence of other customers and deep discount. Therefore, instead of star ratings, "real star rating" function:

  $$\mathbf{f} = \mathbf{0.5}\times(\boldsymbol{k} + \boldsymbol{g})$$

  should be used to estimate the subjective rating of products. More details about the function will be provided in the following report.

- Do follow the properties and design of products with highest real star rating. By combing the text-based measures and numerical measures, our team analyze the reviews with higher real star rating and conclude a group of keywords which is the popular and high-rating products' "key points". For example, the words "purple" appear frequently in the top 25% rating list and this provides us some directions which purple-colored hair dryers maybe more popular in online marketplace. In the same way, the keywords of properties appear frequently in the last 25% list should be prevented. The whole keywords list by products ranking will be provided in the following report.

- Do follow the trends of online marketplace. By analyzing time-based measures and patterns, it can calculate each kinds of products' trends by rating in each single year. For example, the graph below



  show the steady increasing of real star rating and this type of products will be more welcome in the next year. Therefore, we can learn from its properties in designing our products so that it may become more popular. The whole time-based analysis graphs are provided in the following report.

- DO introduce the products in objective wordings. By combing the text-based measures and numerical measures, our team analyzes the relationship between total votes number and type of specific reviews in a comment. We found that it is more preferable to customers when seeing objective wordings. On the other hand, the wordings with emotional descriptions are less attractive to customers. The whole analysis and measure are provided in the following report.

Our team hopes that the analysis and strategies we provided can help Sunshine Company achieve marvelous success to the three new products in the future. If there are any questions about the analysis and strategies or more information is needed, feel free to contact us and we will help Sunshine Company with all our efforts as much as we can.

Yours Sincerely,

Team#2001866

9 References

[1] Markov chain. http://episte.math.ntu.edu.tw/articles/mm/mm_09_3_08/page2.html

[2] Markov chain. https://math.ntnu.edu.tw/~horng/letter/hpm17078.pdf

[3] Wikipedia. https://en.wikipedia.org/wiki/Markov_chain

10 Appendices

Appendix A Tables {top/worst 5 and 25% product}

Hair dryer

product_parent	Real star rating	product_title
The top		
872781320	5.0	Eye Revival (.5 Oz) Age Deying Eye Serum - Reduces Wrinkles, Dark Circles & Under Eye
244516305	5.0	Volupte Perfume by Oscar de la Renta for women Personal Fragrances
856753131	5.0	Delon 100% Cotton Rounds - New and Improved Premium Quality Softer Edges - 800 Rounds
190867408	5.0	OPIS Crystal Nail File
454690299	5.0	Clairol Natural Instincts Loving Care Color Pack of 3
The bottom		
247782192	1.0	Bumble and Bumble Conditioner, Seaweed, 33.8 fl oz (1 lt)
409998686	1.0	Remington SP-62 Foils and Cutters, Black
783438686	1.2572	Conair Instant Heat Compact Hot Hair Rollers
311969364	1.3930	PHERAZONE for MEN 36mg per ounce Ultra Powerful Pheromone Cologne Guaranteed to Boost Your Sex Appeal, Scented
580704943	1.4263	Braun Series 3 370 Men's Shaver

Microwave Oven

product_parent	Real star rating	product_title
The top		
664466484	5.0	ge jgb870sefss 30 stainless steel gas sealed burner double oven range - convection
149559260	4.2154	samsung basket adjuster - dd97-00119b
984005611	4.1270	sharp rmotda252wrzz microwave turntable motor

309267414	4.0634	jx7227sfss - deluxe built-in trim kit for 2.2 microwave ovens/ compatible with peb7226sf models/ stainless steel finish
523301568	4.0549	microwave cavity paint 98qbp0302
The bottom		
550562680	1.7363	Samsung SMH1713S 1.7 Cu. Ft. Stainless Steel Over-the-Range Microwave
311592014	1.7846	koolatron coca-cola indoor/outdoor party fridge
147401377	1.8645	ge jvm1540smss spacemaker 1.5 cu. ft. stainless steel over-the-range microwave
921964554	2.0455	whirlpool gmh5205xvs 2.0 cu. ft. over-the-range microwave oven 300 cfm - stainless steel
392967251	2.0613	frigidaire ffmv164l 1.6 cubic foot over-the-range microwave with fits-more capacity, 1,550 watts and,

Microwave oven

Keywords that appears in the **product titles** of the **top 25%** products
brand: sharp, amana **model:** wmc20005yw, wmc20005yd, 98qbp0302, om75p, wb27x10017, rmotda252wrzz, amv1150vaw, pem31dmww%2d, ntnt-a117wrez, jx7227sfss **other descriptors in the titles:** countertop, 0.5, white, look, feet,, cavity, paint, 950-watt, 1-2/5-cubic-foot, magnetron, diode, part, turntable, motor, ii%2dcountertop, %2d
Keywords that appears in the **product titles** of the **bottom 25%** products
brand: samsung, frigidaire, **model:** smh1816s, gh7208xrs, ffmv164l, pvm1790, gmh5205xvs, smh2117s **other descriptors in the titles:** over-the-range, steel, 1.8, 1.6, 2.0, cfm, 300, range, 1.7, gold, fits-more, capacity,, 1,550, 2.1, ceramic, enamel, interior, large, capacity, led, cook, top, light, 400, bottom
Keywords that appears in the **review bodies** of the **top 25%** products
Paint, old, little, works, space, great, fit, counter, fits, well, looks, perfect, good, easy, size, plate
Keywords that appears in the **review bodies** of the **bottom 25%** products
Service, Samsung, door, years, unit, repair, months, warranty, replace, Whirlpool, fan, problem, handle

Remark:
We identify that brands such as Sharp and Amana are performing well while Samsung, and Frigidaire receive low ratings. Certain models win themselves good or bad reputation are shown above. In addition, it is found that the after-sales service such as warranty and repair is of great concerns to the customers as reflected in the complaints.

Pacifier

Keywords that appears in the **product titles** of the **top 25%** products
brand: wubbanub, mary meyer, philips **model:**0-3month, 0-6month, 6+ **other descriptors in the titles:** giraffe, monkey, soft, plush, elephant, bottle, bpa-free, avent, brown, pink
Keywords that appears in the **product titles** of the **bottom 25%** products
brand: munchkin, gerber, razbaby, motorola **model:** silicone, keep-it-kleen, wireless, nuk **other descriptors in the titles:** natural, medicine, summer, crib, lcd
Keywords that appears in the **review bodies** of the **top 25%** products
out, daughter, find, mouth, hold, keep, cute, back, months
Keywords that appears in the **review bodies** of the **bottom 25%** products
Little, used

Remark:
This analysis reveals that brans such as wubbanub, mary meyer, and Philips earn high ratings, while others like munchkin, gerber, razbaby, and Motorola are less welcome. Moreover, as many animals appear in the high-rating reviews, one may speculate that pacifiers that come with shapes of animals appeal to the customers.

Incidentally, we would like to express our doubt on seeing items such as “motorola digital audio baby monitor” in our analysis. We wonder if this is pertinent to our subject, pacifier. However, as we do not have external information other than the data set itself, we are not capable of verifying the relevance of our subject and therefore keep it as it is.

Appendix B Plots

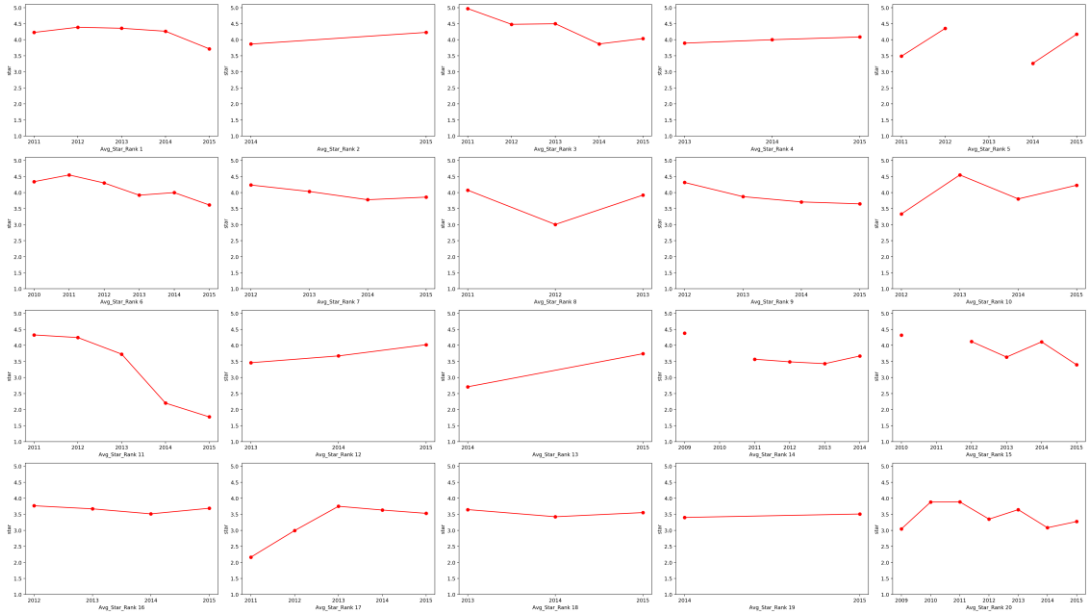


Fig7: 20 best reputed microwaves (judging from their real star ratings)

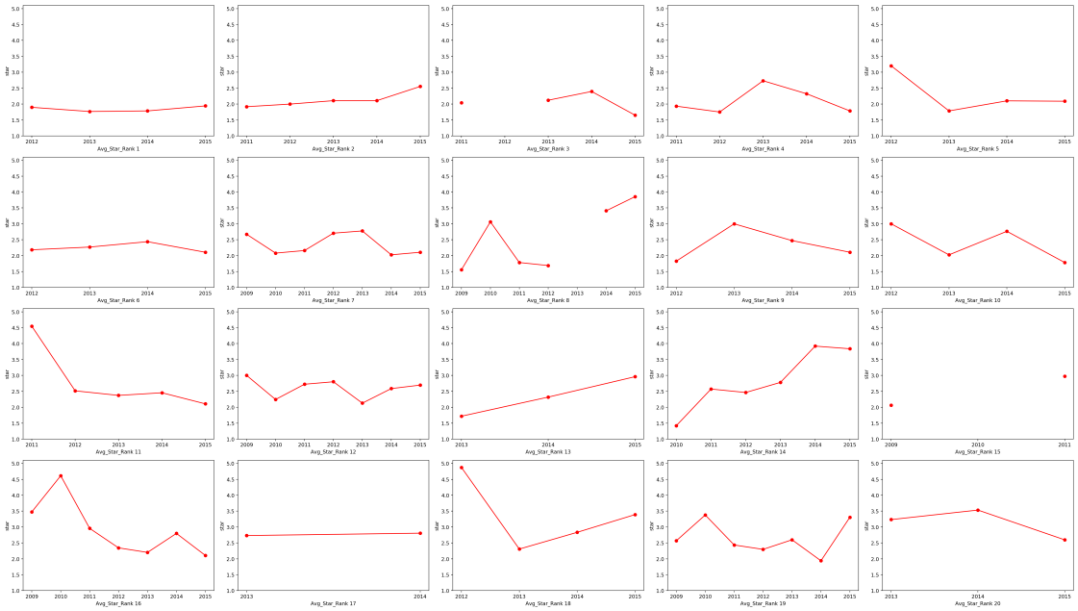


Fig8: 20 most unwelcome microwaves (judging from their real star ratings)

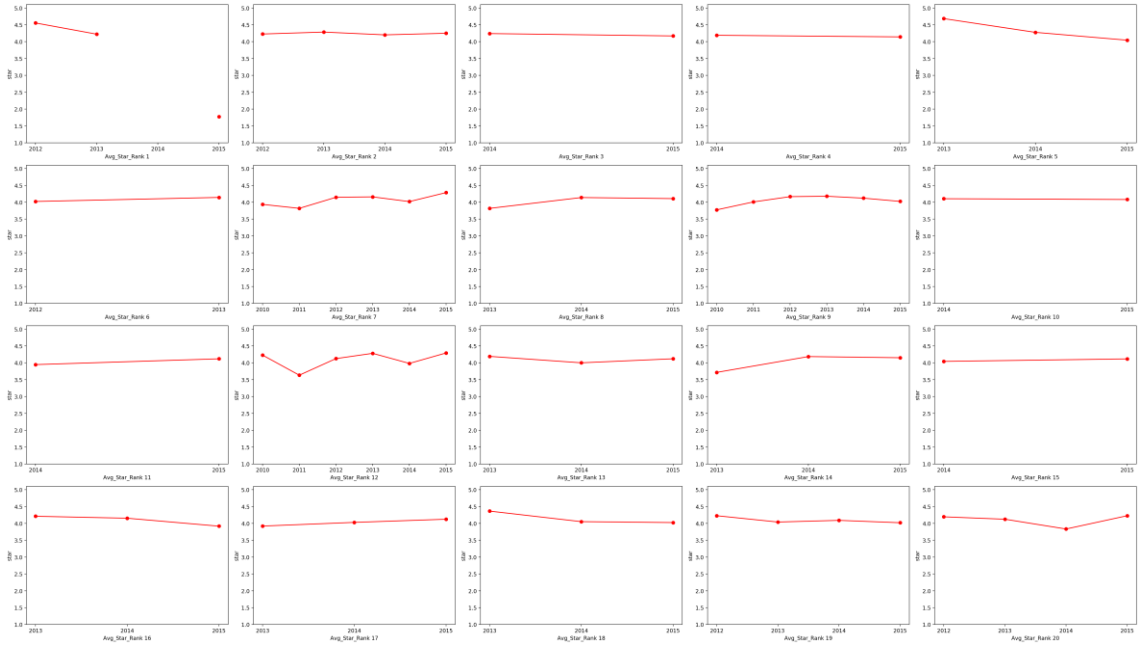


Fig9: 20 best reputed pacifiers (judging from their real star ratings)

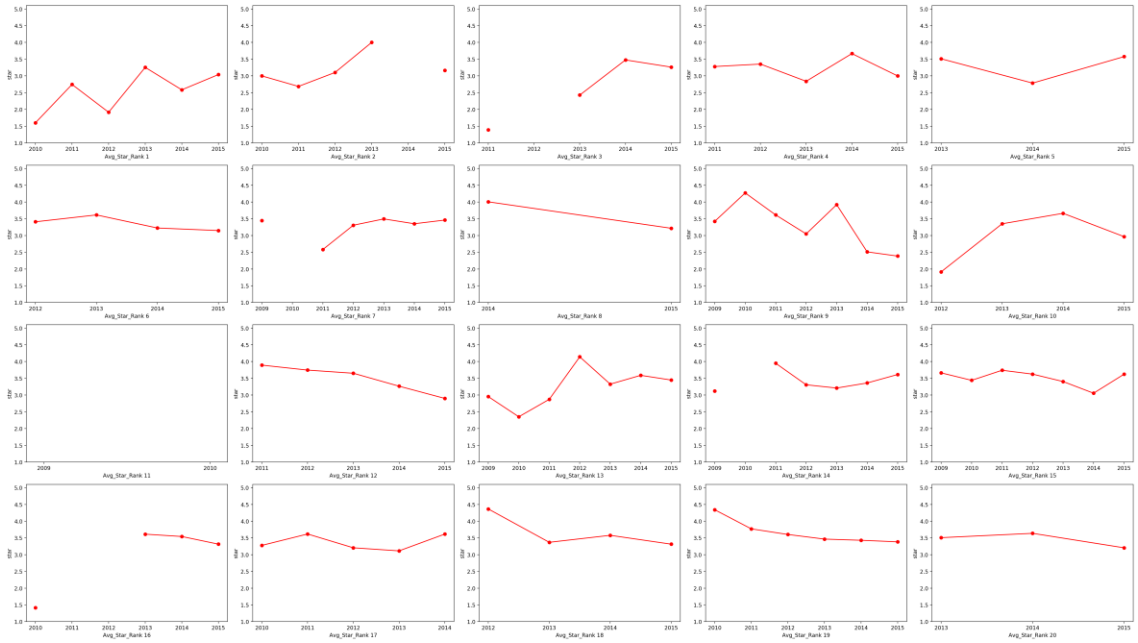


Fig10: 20 most unwelcome pacifiers (judging from their real star ratings)

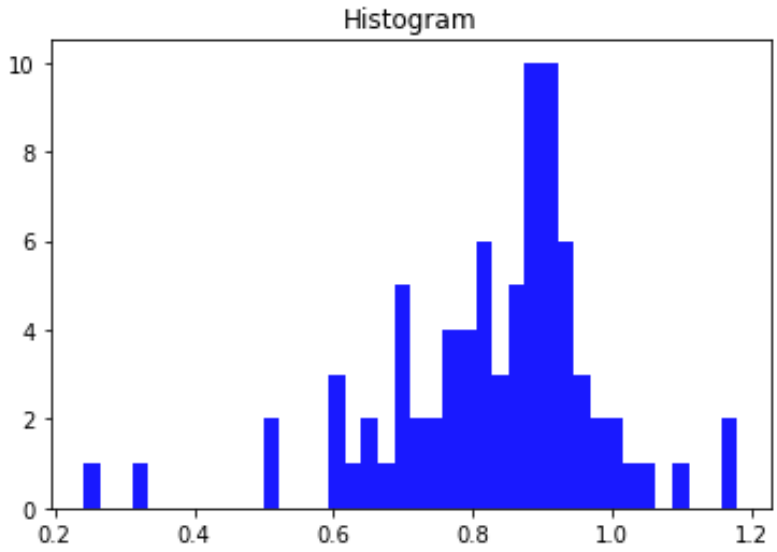


Fig11: The statistical distribution of ps for pacifier. The vertical axis represents the number of data and the horizontal axis represents ps. The mean value is 0.8276, and the standard deviation 0.1542, 78.75% of all data fall within the one standard deviation interval.

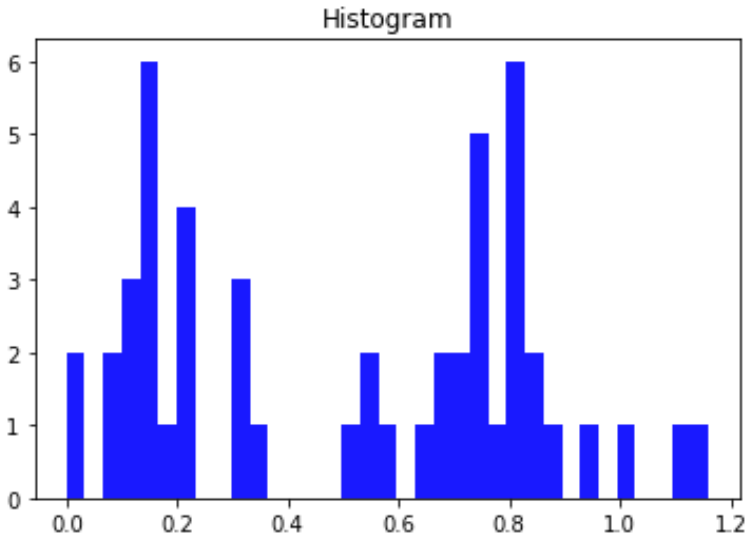


Fig12: The statistical distribution of ps for microwave. The vertical axis represents the number of data and the horizontal axis represents ps. The mean value is 0.5069, and the standard deviation 0.326, 62.5% of all data fall within the one standard deviation interval.

Appendix C Codes

Python Source Code: Plotting the statistical distribution of ps for pacifier (Fig6、11、12)

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import math

tsv=pd.read_csv('microwave.tsv', sep='\t', header=0)

star=tsv['star_rating']
hvote=tsv['helpful_votes']
tvote=tsv['total_votes']
#cnt=tsv['star_rating'].argmin()
pparent=tsv['product_parent']
ptitle=tsv['product_title']
total=star.count()
comment=tsv['review_body']
year=tsv['review_date']
for i in range(total):
    year[i]=year[i][-4:]   #year ready

pplist=[]      #***important***
ptlist=[]

for i in range(total):
    if pparent[i] not in pplist:
        pplist.append(pparent[i])
        ptlist.append(ptitle[i])
series=pd.Series(ptlist, index=pplist)
staravg=[0 for _ in range(series.count())]
number=[0 for _ in range(series.count())]
sumnumber=[0 for _ in range(series.count())]
```

```
dt={#'0': series,
   '1': pd.Series(staravg, index=pplist),
   '2': pd.Series(number, index=pplist),
   '3': pd.Series(sumnumber, index=pplist)}
nianfen={}
for i in ['2009','2010','2011','2012','2013','2014','2015',
        '2009b','2010b','2011b','2012b','2013b','2014b','2015b']:  #years
   nianfen[i]=pd.Series(staravg, index=pplist)

d=pd.DataFrame(dt)
f=pd.DataFrame(nianfen)

for i in range(total):
   if hvote[i]>=0:    #helpful_votes>2
      d['1'][pparent[i]]+=star[i]
      d['2'][pparent[i]]+=1
      if int(year[i])>=2010:    #consider the year
         if star[i]>=4:
            f[year[i]][pparent[i]]+=1
         else:
            f[year[i]+'b'][pparent[i]]+=1
   d['3'][pparent[i]]+=1
s=d['1']/d['2']

d1=d.drop(['1'], axis=1)
d1.insert(0, '1', s)
d1=d1.join(f)
d1=d1.sort_values('2', axis=0, ascending=False)
for i in ['2009','2010','2011','2012','2013','2014','2015']:
   d1[i]=d1[i]/(d1[i]+d1[i+'b'])
d1=d1.drop(['2009b','2010b','2011b','2012b','2013b','2014b','2015b'], axis=1)

def solve(p, q, r):  #p=last year, q=this year, r=next year
   x=((p-1)*r-q**2+q)/(p-q)
   y=(p*r-q**2)/(p-q)
   return [round(x, 2),round(y,2)]
```

```
nian={}
basic=[0,0]
empty=[0 for _ in range(30)]
for i in range(30):
    empty[i]=basic

for i in ['2012','2013','2014','2015']:
    nian[i]=pd.Series(empty, index=d1.index[:30])
maerkefu=pd.DataFrame(nian)

maerkefu_solve=pd.DataFrame(index=d1.index[:30], columns=['2012','2013','2014','2015'])

for i in maerkefu.index:
    for j in maerkefu.columns:
        maerkefu[j][i]=solve(d1[str(int(j)-2)][i], d1[str(int(j)-1)][i], d1[j][i])
        ss=maerkefu[j][i][1]/(1+maerkefu[j][i][1]-maerkefu[j][i][0])
        if ss>=0:
            maerkefu_solve[j][i]=ss

#%%matplotlib.pyplot%%
data=np.array(maerkefu_solve).flatten()     #data
plt.hist(data, 35, normed=0, histtype='stepfilled', facecolor='b', alpha=0.9)
plt.title('Histogram')
plt.show()
```

Python Source Code: Plotting the 20 best/worst thing (judging from their real star ratings)

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt

tsv=pd.read_csv('hair_dryer.tsv', sep='\t', header=0)

star=tsv['star_rating']
hvote=tsv['helpful_votes']
tvote=tsv['total_votes']
#cnt=tsv['star_rating'].argmin()
pparent=tsv['product_parent']
ptitle=tsv['product_title']
total=star.count()
comment=tsv['review_body']
year=tsv['review_date']
vine=tsv['vine']
purchase=tsv['verified_purchase']
for i in range(total):
    year[i]=year[i][-4:]   #year ready

pplist=[]       #***important***
ptlist=[]

for i in range(total):
    if pparent[i] not in pplist:
        pplist.append(pparent[i])
        ptlist.append(ptitle[i])
series=pd.Series(ptlist, index=pplist)
staravg=[float(0) for _ in range(series.count())]
number=[0 for _ in range(series.count())]
sumnumber=[0 for _ in range(series.count())]

dt={#'0': series,
    '1': pd.Series(staravg, index=pplist),
    '2': pd.Series(number, index=pplist),
    '3': pd.Series(sumnumber, index=pplist),
```

```
    'vine': pd.Series(sumnumber, index=pplist)}
nianfen={}
for i in ['2009','2010','2011','2012','2013','2014','2015',
          '2009b','2010b','2011b','2012b','2013b','2014b','2015b']:
    nianfen[i]=pd.Series(staravg, index=pplist)

d=pd.DataFrame(dt)
f=pd.DataFrame(nianfen)

def realstar(s, vine, helpful, total, purchase):
    s-=3
    if vine=='Y':
        return s+3
    else:
        k,g=float(0),float(0)
        if helpful==0:
            k=0.8
        elif 1<=helpful and helpful<10:
            k=(float(helpful)/total)*1.3
        else:
            k=(float(helpful)/total)*2
        if purchase=='Y' :
            g=1.11
        else:
            g=0.59
        s=s*0.5*(k+g)*2/3.11+3
        return s

for i in range(total):
    if hvote[i]>=0:      #helpful_votes>=0
        d['1'][pparent[i]]+=realstar(star[i], vine[i], hvote[i], tvote[i], purchase[i])
        d['2'][pparent[i]]+=1
        if vine[i]=="Y":
            d['vine'][pparent[i]]+=1
        if int(year[i])>=2009:      #consider the year
            f[year[i]][pparent[i]]+=realstar(star[i], vine[i], hvote[i], tvote[i], purchase[i])
            f[year[i]+'b'][pparent[i]]+=1
```

```
    d['3'][pparent[i]]+=1
s=round(d['1']/d['2'], 4)

d1=d.drop(['1'], axis=1)
d1.insert(0, '1', s)
d1=d1.join(f)

for i in ['2009','2010','2011','2012','2013','2014','2015']:
    d1[i]=round(d1[i]/d1[i+'b'], 4)
d1=d1.drop(['2009b','2010b','2011b','2012b','2013b','2014b','2015b'], axis=1)

d1=d1.sort_values('2', axis=0, ascending=False)
okd1=d1.head(124).sort_values('1', axis=0, ascending=True)

yearindex=['2009','2010','2011','2012','2013','2014','2015']
newd1=okd1.loc[:, yearindex].head(20)

#%%pyplot%%
fig = plt.figure(figsize=(30, 17), dpi=200)
for i in range(20):
    plt.subplot(4,5,i+1)
    plt.axis(ymin=1, ymax=5.1)
    plt.ylabel('star')
    plt.xlabel('Avg_Star_Rank '+str(i+1))
    plt.plot(newd1.iloc[i], 'r-o')
    plt.tight_layout()
```