Problem Chosen	2020 MCM/ICM Summary Sheet	Team Control Number
C		2002768

Summary

In the past decade or so, interactive online shopping based on star ratings and reviews has brought great convenience to people. At the same time, data mining related online shopping has become a vital preparation before a product launch.

The purpose of this paper is to help Sunshine Company generate their online sales strategy and identify popular features of three products by analyzing the data from online marketplace. We have established different models for different indicators and problems.

Firstly, we used machine learning-based sentiment analysis algorithms [1] to reviews quantification in the data. We built a multi-valued logit model for star_rating, helpful_votes, vine and verified_purchase to calculate their correlation. Secondly, we used moving average method with the star ratings as the reputation index to analyze the tendency of data. Time series diagrams of different products reflect their reputation trends. After that, we could judge the success of products from star ratings after the reputation was stable. Thirdly we used Difference-in-difference (DID) module to observe the impact of special star ratings on reviews.

Finally, we got the following results:

1. The positive correlation between the emotional tendency of reviews and star rating is very significant at the 1% level. There is no significant relationship among the emotional tendency of reviews, vine reviews and verified_purchase

2 In general, the reputation of the product has generally approached positive evaluation over time, but there are also products with a gradually decreasing reputation such as microwave ovens B00219QFNC, B002X7669C, etc.

3. Low star ratings could increase negative reviews after appearing on a product, while high stars could not. The positive correlation between the number of emotional extreme words and star rating is significant at the level of 1%.

In addition, we also identify popular product features, which will have implications for product design.

Key words: Data Mining; Natural Language Processing; Emotional Tendency; DID Model

Table of Contents

1 Introduction

1.1 Background

For the past few years, product and customers have experienced the integration of the Internet into shopping. However, in e-commerce, the space or time separation between buyers and sellers creates information asymmetry between the two parties. For sellers, in order to accurately and effectively grasp market conditions and succeed, they have to master potential laws of the online feedback system.

In Amazon, there are two types of interactive tools, star ratings and reviews. And how to find assessment pattern from oceans of data becomes the key to the success of a commodity. As for them, we need to adopt data mining [2] and text mining [3] to determine product reputation and consumer preferences. Chen Yi (2018) built key word cloud and topic model to measure emotional tendencies of reviews. And Zhou Qiang (1995) [4] briefly introduced the basics of corpus-based and statistical-oriented natural language processing(NLP) techniques. So, we could refer to their method to quantify data supplied by Sunshine Company, and then build statistical models to achieve the following goals.

1.2 Research Purposes

- Use the data provided to build a model to explore the correlation between them.
- Perform experiment and observe product reputation changes over time series.
- Use star ratings and review to build a Moving Average Module to evaluate the success of the products.
- Identify certain popular product characteristics to find consumer needs.
- Investigate whether stars and evaluations interact with each other and how expressions affect.
- Provide helpful advice based on experimental results to Sunshine Company.

2 Related work

- We need to preprocess the data, especially the quantitative processing of reviews.

This paper employs the theory of text mining to analyze the emotional tendency of reviews. Two indicators are used to measure: emotional polarity and emotional subjectivity. Emotional polarity refers to the bias of the subject's evaluation between positive and negative emotions. For example, “like” and “good” are both commendatory term, the emotional polarity tends to be positive; while “hate” and “poor” are derogatory words, so the emotional polarity tends to be negative. Emotional subjectivity refers to the degree of emotional strength of the subject's evaluation, which is often expressed through the use of degree adverbs.

●We need to consider how to measure the reputation of a product on a time series.

Considering the time lag between commodity trading time and interaction time of Online shopping, we build a moving average model to reduce the errors caused by time or irregular data.

●We need to have a related sentiment lexicon to find specific quality descriptors in reviews.

3 Symbol description

Table 1: Main variables and definitions

Symbol	Definition
$polarity_{i,t}$	emotional polarity.
$subjectivity_{i,t}$	emotional subjectivity.
$sr_{i,t}$	star_rating.
$hv_{i,t}$	net likes,$hv_{i,t} = (helpful\_votes - (total_{votes} - helpful_{votes}))$.
vp_i	dummy variable， $verified\_purchase$ shows “Y” is 1 and “N” is 0.
v_i	dummy variable， $vine$ shows “Y” is 1 and “N” is 0.
P_i	the number of positive words
N_i	the number of negative words

4 Basic Assumptions

To build the model, we made assumptions to ignore meaningless details.

1. The sentiment indicators identified by NLP are all correct.

2. Emotional tendencies in reviews are divided into three categories: positive reviews, neutral reviews, negative reviews.

3. Extreme emotion words have only two directions: extreme positive and extreme negative.

4. The data in the three data sets provided does not have the behavior of "water army" beautifying comments.

We know that in order to make their products easy to sell, some e-commerce companies will hire some people to praise the products they operate. We assume that these misconducts do not exist in the data.

5 Model design and results analysis

5.1 Emotional polarity and subjectivity

This paper uses machine learning-based sentiment analysis algorithms to process natural language, using the Blobtext module in Python to process customer reviews in three product datasets to get the polarity and subjectivity of each review. For some ambiguous reviews, the Blobtext module automatically changes the emotional subjectivity to 0, so this paper excludes data with emotional subjectivity of 0.

Table 2: Descriptive statistics of emotional polarity and emotional subjectivity

	Variable	Obs	Mean	Std. Dev.	Min	Max
hair_dryer	$polarity_{i,t}$	10,713	0.255804	0.252178	-1	1
	$subjectivity_{i,t}$	10,713	0.562191	0.161808	0.033333	1
microwave	$polarity_{i,t}$	1,522	0.206019	0.260526	-1	1
	$subjectivity_{i,t}$	1,522	0.543288	0.168362	0.05	1
pacifier	$polarity_{i,t}$	17,036	0.272732	0.259497	-1	1
	$subjectivity_{i,t}$	17,036	0.579101	0.180069	0.025641	1

5.2 Quantitative relationship measures based on ratings and reviews

This paper uses a multi-valued logit model for regression analysis, which was first proposed by McFadden in 1973. The formula is as follows:

$$p_i = F(y_i) = F(\alpha + \beta x_i) = \frac{1}{1 + e^{-y_i}} = \frac{1}{1 + e^{-(\alpha + \beta x_i)}} \tag{1}$$

Where p_i represents probability, $F(y_i)$ represents logistic cumulative probability density function. For a given x_i, p_i represents the probability of the individual making a certain choice. y_i is called a latent variable. The value range of y_i is (-∞, ∞). y_i is converted into a probability by the logistic function. The multi-valued logit model is suitable for more than two cases.

This paper divides reviews into neutral reviews, positive reviews, and negative reviews based on the emotional polarity of each review. We set neutral comments to 0, negative comments to 1, positive comments to 2, then use multi-valued logit model for regression. Because multi-valued regression needs to use one of them as the reference group, the blank column in Table 3 indicates that this column is used as the base group for regression. The regression results are as follows:

Table 3: multi-valued logit model regression results

hair_dryer									
VARIABLES	0	1	2	0	1	2	0	1	2
$sr_{i,t}$		-0.628***	0.635***	0.628***		1.263***	-0.635***	-1.263***	
		-0.0473	-0.0194	-0.0473		-0.0497	-0.0194	-0.0497	
$hv_{i,t}$		-0.0876***	-0.0121***	0.0876***		0.0755***	0.0121***	-0.0755***	
		-0.0231	-0.00252	-0.0231		-0.0232	-0.00252	-0.0232	
v_i		-13.94	-0.268	13.94		13.68	0.268	-13.68	
		-707.2	-0.166	-707.2		-707.2	-0.166	-707.2	
vp_i		0.132	0.417***	-0.132		0.285	-0.417***	-0.285	
		-0.173	-0.0643	-0.173		-0.179	-0.0643	-0.179	
Constant		-1.145***	-2.768***	1.145***		-1.623***	2.768***	1.623***	
		-0.177	-0.1	-0.177		-0.196	-0.1	-0.196	
Observations	10,713	10,713	10,713	10,713	10,713	10,713	10,713	10,713	10,713
microwave									
VARIABLES	0	1	2	0	1	2	0	1	2
$sr_{i,t}$		-0.513***	0.623***	0.513***		1.135***	-0.623***	-1.135***	
		-0.122	-0.0473	-0.122		-0.127	-0.0473	-0.127	
$hv_{i,t}$		-0.0543**	-0.0217***	0.0543**		0.0326	0.0217***	-0.0326	
		-0.0267	-0.00572	-0.0267		-0.027	-0.00572	-0.027	
v_i		-9.821	0.736	9.821		10.56	-0.736	-10.56	
		-424.9	-0.537	-424.9		-424.9	-0.537	-424.9	
vp_i		0.639*	0.646***	-0.639*		0.00716	-0.646***	-0.00716	
		-0.329	-0.156	-0.329		-0.351	-0.156	-0.351	
Constant		-1.808***	-2.826***	1.808***		-1.018***	2.826***	1.018***	
		-0.275	-0.187	-0.275		-0.318	-0.187	-0.318	
Observations	1,522	1,522	1,522	1,522	1,522	1,522	1,522	1,522	1,522
pacifier									
VARIABLES	0	1	2	0	1	2	0	1	2
$sr_{i,t}$		-0.635***	0.619***	0.635***		1.254***	-0.619***	-1.254***	
		-0.0329	-0.0164	-0.0329		-0.0352	-0.0164	-0.0352	
$hv_{i,t}$		-0.0758***	-0.00867***	0.0758***		0.0671***	0.00867***	-0.0671***	
		-0.0167	-0.00217	-0.0167		-0.0167	-0.00217	-0.0167	
v_i		-13.02	0.277	13.02		13.3	-0.277	-13.3	
		-533.2	-0.191	-533.2		-533.2	-0.191	-533.2	
vp_i		0.556***	0.197***	-0.556***		-0.359**	-0.197***	0.359**	
		-0.15	-0.0485	-0.15		-0.154	-0.0485	-0.154	
Constant		-1.062***	-2.448***	1.062***		-1.387***	2.448***	1.387***	
		-0.155	-0.0841	-0.155		-0.169	-0.0841	-0.169	

Observations	17,036	17,036	17,036	17,036	17,036	17,036	17,036	17,036	17,036
Standard errors in parentheses									
*** p<0.01, ** p<0.05, * p<0.1									

Table 3 shows that when the neutral group is used as the base group, the star ratings of the three categories of products are significantly positively correlated with the positive group, and are significantly negatively correlated with the negative group; net likes are significantly negatively correlated with both the positive and negative groups, positive and negative groups have no significant relationship with whether it is a vine review, positive group and negative group are significantly positively related to verified_purchase except for the negative group of hair dryers; when based on the negative group, the star ratings of the three categories of products are significantly positively related to the positive group and the neutral group. The positive group and the neutral group were significantly positively correlated with the number of net points except for the active group of microwave oven. Although the vine review is positively related, the relationship is not significant. The positive group and the neutral group of the pacifier are significantly negatively related to the verified_purchase. When the positive group is the basic group, The negative and neutral groups of the three types of products have a significant negative correlation with the star rating, with the exception of the negative group of the microwave oven, the other negative groups are significantly negatively correlated with the net likes, and the neutral group is significantly positively correlated with the net likes. Similarly, the relationship between the negative group and the neutral group and whether it is a vine review is not significant, the neutral group was significantly negatively correlated with verified_purchase, but only the pacifiers in the negative group were significantly positively related to verified_purchase.

In summary, it can be considered that the relationship between the emotional tendency of reviews and star ratings is very significant: the higher the star rating, the more positive the content of the review; the emotional tendency of reviews and the net likes are also significant: the higher the net likes, the more positive the content of the review ; there is no significant positive or negative relationship with whether it is a vine review; the relationship with whether it is a verified_purchase is also less clear. Therefore, the following uses stars as a measure of reputation, and explores the trend of reputation changes of various products based on time series.

5.3 Product star time trend chart

This paper intends to draw the star-time trend line of all products in three categories of data sets. Since each product has a different time period for purchase, the trend chart for each product is calculated as follows:

(1) Calculation of time interval: calculate the appropriate time interval by taking the time period of customers' purchase behavior for a certain product as the numerator and the sum of the total number of days that all customers have actually purchased in that time period as the denominator;

(2) Star rating calculation: Take a moving average of star ratings of the same product based on the calculated time interval.

After excluding the data of low sales volume and too short time series, the trend charts of the three types of products are as follows (the header is the product code and sales volume).

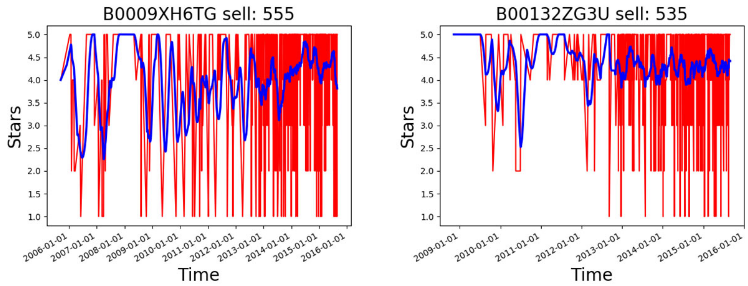


Fig 1: Time trend of hair dryer products

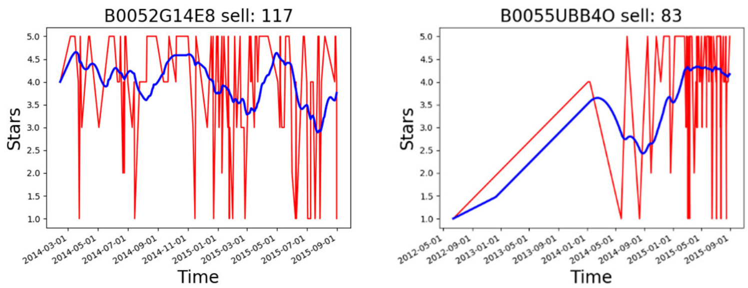


Fig 2: Time trend of microwave products

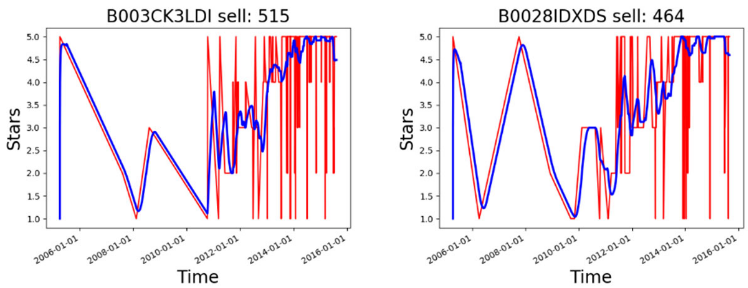


Fig 3: Time trend of pacifier products

Considering the length of this paper, the main body does not list all time trend charts. For the rest, please refer to the appendix. The blue solid line is the time trend change we are concerned about. As can be seen from the above figure, in general, the product's reputation gradually changes to a certain evaluation and the overall approach to high star rating over time, but there are also cases where the reputation gradually decreases over time, such as microwave oven products B00219QFNC, B002X7669C etc., or the reputation has not changed much, such as microwave oven products B0073YCGPI.

5.4 Best Selling Product

This paper further explores the success or failing products based on the relationship between text and stars. Our definition of successful product is: product star reputation tends to be stable and high. Taking into account the changes in the stability of product reputation when cumulative sales increase. We use the coefficient of variation within a certain sales volume to measure the stability of reputation changes. For each product, calculate the coefficient of variation curve using each 1/10 * of the total sales volume as the window, and calculate the average of the same type of product to obtain the image as follows.

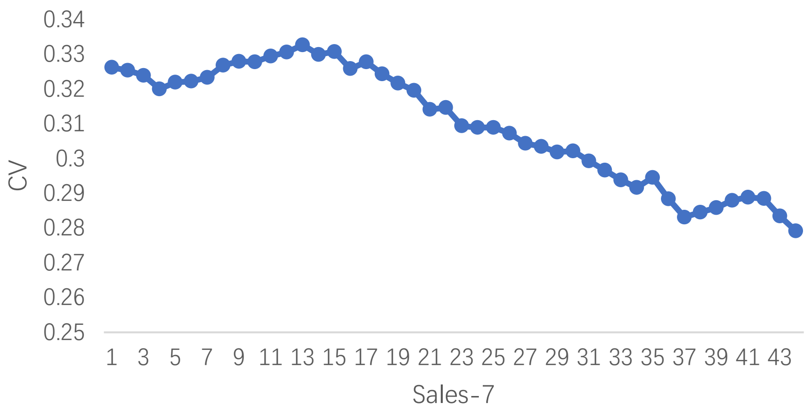


Fig 4: hair dryer coefficient of variation of Star ratings

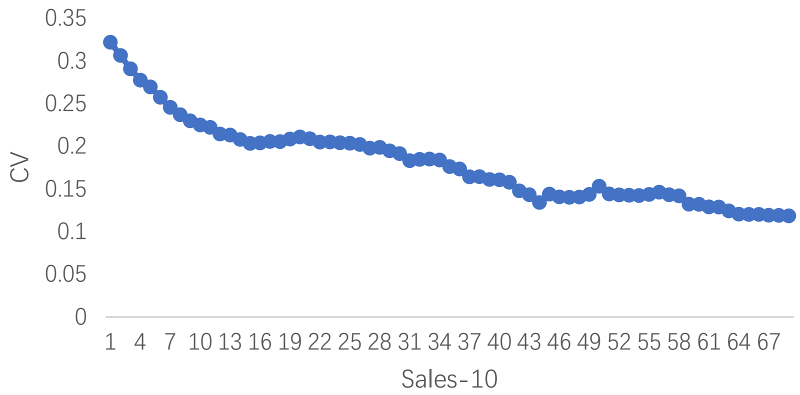


Fig 5: pacifier coefficient of variation of Star ratings

Because the sales volume is too small, the coefficient of variation of microwave oven is difficult to count. But it is easy to see from the above two figures that the coefficients of variation of both of them show a downward trend with the increase in sales, and the coefficient of variation of the hair_dryer is less obvious, which stabilizes at about 42 sales, and the pacifier decreases significantly, about 53 The situation has stabilized. Based on this, we can judge the time when the new product enters the online market to evaluate the fluctuations, so as to track the data in a timely manner.

Next, this paper uses text-based measure to find the relationship between customer star ratings and product titles. By separating and filtering adjectives from product titles in the three data sets, we can find the average star rating for each word. This information can be used for forecasting product evaluations. Each adjective extracted from the title can match a specific star rating, so the predicted star rating for the entire title is the average number of star ratings for each word. The prediction error rate between the actual star ratings and the predicted star ratings of the three categories of products is shown in the figure below:

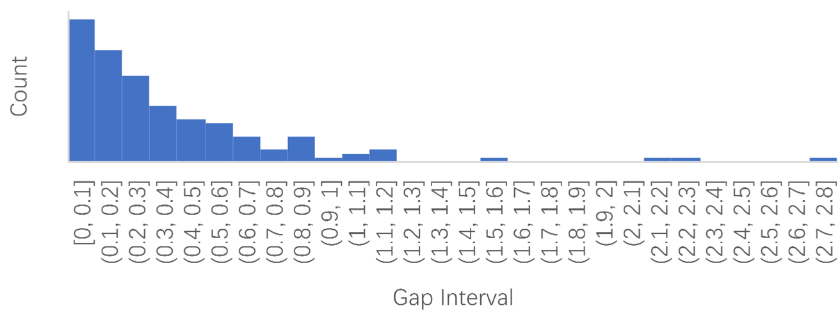


Fig 6: Gap between actual and predicted star ratings (hair dryer)

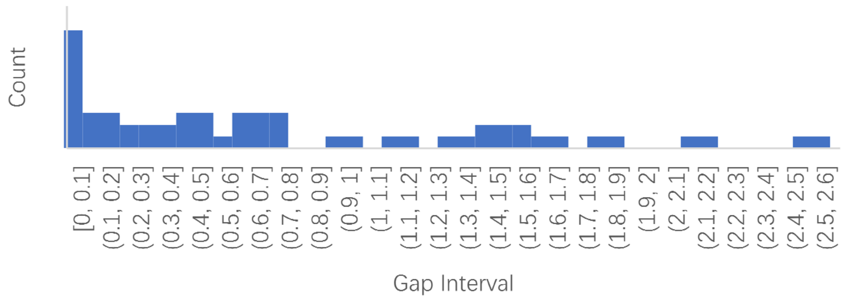


Fig 7: Gap between actual and predicted star ratings (microwave oven)

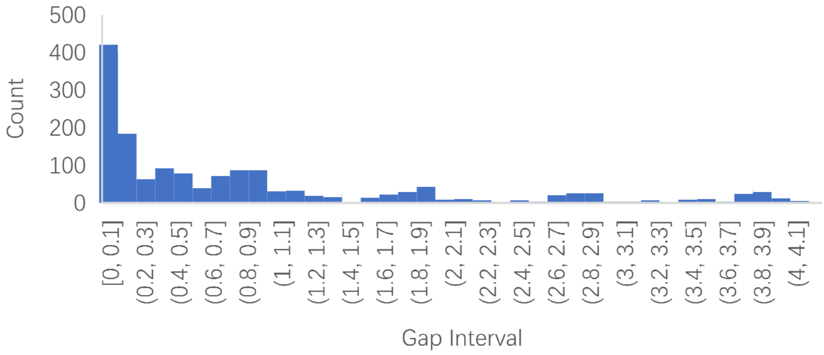


Fig 8: Gap between actual and predicted star ratings (pacifier)

Tables 4, 5, and 6 show words with a higher average star rating among the three categories of products. This data will be useful for product design and heading.

Table 4: Words with average star ratings higher than 4.5 (hair dryer)

word	1star	2star	3star	4star	5star	total	average-star
precise	10	0	7	18	124	159	4.547169811
powerful	13	2	9	21	151	196	4.505102041
ergonomic	4	4	11	19	100	138	4.5

Table 5: Words with average star ratings higher than 4.0 (microwave oven)

word	1star	2star	3star	4star	5star	total	average-star
diode	0	2	1	4	25	32	4.625
turntable	1	0	3	4	23	31	4.548387097
950-watt	1	2	8	7	27	45	4.266666667
1-2/5-cubic-foot	1	2	8	7	27	45	4.266666667
white	15	4	7	24	67	117	4.05982906

Table 6: Words with average star ratings higher than 4.5 (pacifier)

word	1star	2star	3star	4star	5star	total	average-star
limited	2	1	3	16	96	118	4.720338983
6-count	6	1	1	11	90	109	4.633027523

puppy	7	9	14	25	230	285	4.621052632
fanatic	5	5	10	18	124	162	4.549382716
2-count	29	21	41	77	564	732	4.538251366
fisher-price	6	6	10	37	155	214	4.537383178
little	4	4	12	19	114	153	4.535947712

5.5 Research on special star ratings

This paper further studies whether the negative reviews, that is, the appearance of low or high stars, will cause negative or positive reviews from customers. This paper uses multi-period double difference (DID) to explore the relationship between the two.

The basic form of multi-period double difference is double difference. This model was proposed by Ashenfeler and Card (1985) for the first time when studying the vertical change of the income structure of CETA training students. DID can better control the difference between the treat group and the control group. Systematic differences to study the changes of the processing team before and after something happened. The basic model is:

$$y_i = \beta_0 + \beta_1 treat \times post + \beta_2 treat + \beta_3 post + \sum X_{i,t} + \varepsilon_i \tag{2}$$

Where y_i represents the dependent variable, $treat$ is the dummy variable, $treat = 1$ is the treat group, $treat = 0$ is the control group, $post$ is the dummy variable, $post = 1$ is after the event, $post = 0$ is before the event, $X_{i,t}$ represents the control variable, ε_i represents the residual term , the multi-period double difference is in the form of:

$$y_i = \beta_0 + \beta_1 treated \times time + \sum X_{i,t} + \varepsilon_i \tag{3}$$

Where $treated \times time$ represents the treatment group after the event.

This paper uses this model to explore the relationship between star ratings and tendency to comment:

$$polarity_{i,t} = \beta_0 + \beta_1 treated \times time + \beta ControlVariables + \varepsilon_i \tag{4}$$

Where $polarity_{i,t}$ represents the direction of emotional polarity, $treated \times time$ is a dummy variable, and $treated \times time = 1$ represents that a low star appears again after a low star or a high star appears again after a high star ; Otherwise, $treated \times time = 0$, $ControlVariables$ represents the control variable, and ε_i is the residual.

Regression was made on three types of products, and the following regression results were obtained:

Table 7: DID regression results of three types of products

VARIABLES	(1)	(2)	(3)	(4)
hair_dryer				
$polarity_{i,t}$	0.0380***	0.0301**	0.0711***	0.0586***
	-0.00861	-0.0135	-0.0138	-0.0125
$sr_{i,t}$	0.0703***	0.0732***	0.0987***	0.0916***
	-0.00317	-0.00507	-0.00432	-0.00338
$hv_{i,t}$	-0.00405***	-0.00397***	-0.00401***	-0.00407***
	-0.000324	-0.000327	-0.000322	-0.000323
v_i	-0.0468	-0.0510*	-0.0461	-0.0487
	-0.0307	-0.0309	-0.0306	-0.0308
vp_i	0.0352***	0.0344***	0.0358***	0.0353***
	-0.00723	-0.00726	-0.00716	-0.00726
Constant	-0.0716***	-0.0838***	-0.177***	-0.142***
	-0.0118	-0.0131	-0.0195	-0.0157
Observations	9,882	9,882	9,882	9,882
Number of productID	532	532	532	532
R-squared	0.187	0.186	0.188	0.187
microwave oven				
$polarity_{i,t}$	0.0699***	0.0317	0.129***	0.0598**
	-0.0192	-0.0262	-0.0295	-0.0244
$sr_{i,t}$	0.0649***	0.0738***	0.116***	0.0954***
	-0.00662	-0.00903	-0.0115	-0.0103
$hv_{i,t}$	-0.00133***	-0.00124***	-0.00125***	-0.00131***
	-0.000379	-0.000377	-0.000383	-0.000378
v_i	-0.0705**	-0.0940**	-0.0739**	-0.0845**
	-0.0348	-0.0366	-0.0361	-0.0356
vp_i	0.022	0.018	0.0214	0.0197
	-0.0165	-0.0168	-0.0166	-0.0166
Constant	-0.0562**	-0.0743***	-0.243***	-0.146***
	-0.022	-0.0245	-0.0483	-0.0401
Observations	1,468	1,468	1,468	1,468
Number of productID	80	80	80	80
R-squared	0.224	0.218	0.226	0.22
pacifier				
$polarity_{i,t}$	0.0175*	0.0118	0.0261*	0.0302*
	-0.00943	-0.0145	-0.014	-0.0157
$sr_{i,t}$	0.0828***	0.0851***	0.0946***	0.0928***
	-0.00376	-0.00463	-0.00425	-0.00362

$hv_{i,t}$	-0.00438***	-0.00436***	-0.00438***	-0.00441***
	-0.000634	-0.000637	-0.000635	-0.000632
v_i	0.0275	0.0186	0.0251	0.0245
	-0.0464	-0.0465	-0.0466	-0.0467
vp_i	0.0139*	0.0134*	0.0135*	0.0140*
	-0.00795	-0.00794	-0.00792	-0.00794
Constant	-0.0932***	-0.101***	-0.135***	-0.126***
	-0.013	-0.0135	-0.0199	-0.0173
Observations	15,869	15,869	15,869	15,869
Number of productID	6,080	6,080	6,080	6,080
R-squared	0.148	0.148	0.148	0.148
Robust standard errors in parentheses				
*** p<0.01, ** p<0.05, * p<0.1				

In table 7, columns (1), (2), (3), and (4) indicate the relationship between emotional polarity and rating after five-star rating, four-star and five-star rating, two-star and one-star rating, and one-star rating, respectively. For the sake of robust results in this paper, high stars are defined as five stars, four stars and five stars, and low stars are defined as one star, one star and two stars. Regression is performed separately. The results obtained are similar. As can be seen from the table7, among the three types of products, the emotional polarity of reviews and the appearance of low star ratings are positively and significantly related, but not the appearance of high star ratings, indicating that customers are more concerned about low star ratings, low star ratings increase customer negative reviews after appearing on a product.

5.6 Research based on specific quality descriptors of text-based reviews

On this basis, this paper continues to explore whether extreme emotional words such as "enthusiastic" and "disappointed" are related to the star ratings. This paper uses the CNKI sentiment lexicon to first traverse each word in all comments to determine whether it appears in the positive and negative vocabularies of the CNKI, so as to count the number of positive and negative words in each comment. Positive words are defined as positive extreme emotional words such as "enthusiastic", negative words are defined as negative extreme emotional words such as "disappointed", the number of positive words is represented by P_i, the number of negative words is represented by N_i, and then use the least squares linear regression model（OLS） for regression, the results are as follows:

Table 8: Extreme emotion word regression results

	hair_dryer	microwave	pacifier
VARIABLES	star	star	star
P_i	0.0996***	0.0785**	0.0742***
	-0.0188	-0.0385	-0.0161
N_i	-0.210***	-0.142	-0.117**
	-0.0477	-0.12	-0.0471
Constant	3.884***	3.056***	3.837***
	-0.0769	-0.204	-0.0735
Observations	1,428	222	1,841
R-squared	0.025	0.02	0.012
Standard errors in parentheses			
*** p<0.01, ** p<0.05, * p<0.1			

It can be seen from Table 8 that the star ratings of hair_dryer products and pacifier products are significantly positively correlated with the number of positive extreme emotional words, and have a significant negative correlation with negative extreme emotional words, while the star ratings of microwave oven products are only significantly positive with positive extreme emotional words. In general, there is a significant relationship between extreme words and star ratings: reviews with more positive extreme emotional words have higher star ratings, reviews with more negative extreme emotional words have lower star ratings.

6 Strengths and Weaknesses

6.1 Strengths

(1) The NLP method was used to quantify the sentiment expressed by the reviews, to a certain extent, avoiding deviations caused by different scoring standards.

(2) Multi-period DID could be used in the case of different implementation time of policies. Products with different good or bad evaluation time in the data set could be well regressed.

(3) Compared with the AHP model, our model bypassed the problem of artificially weighting variables and made the results more reliable.

6.2 Weaknesses

(1) Although the MA model is suitable for long-term, periodic or irregular data, due to insufficient time, we have not continued to discuss the periodic law of potential reputation changes.

(2) Multi-valued logit model regression based on different base groups may cause bias between regression results.

(3) Statistics on emotion by NLP measures with machine may be not accurate.

7 A Letter to the Marketing Director of Sunshine Company

Dear Marketing Director of Sunshine Company:

We are very pleased to be hired by your company as an online market analysis consultant. For the historical data of the three online markets of hair dryer, microwave oven and pacifier, we have used a model to analyze the market data. The model has the following functions:

1. Being able to draw meaningful quantitative and qualitative patterns between reviews and help levels will help your company succeed in all three new online market product sales.

2. Based on qualitative or quantitative patterns between star ratings, reviews, and helpfulness ratings, determine the data measures that are best tracked for a product after its launch.

3. Being able to identify time-based measures and patterns in each market that indicate that a product's reputation in the online market is rising or falling.

4. Being able to combine text-based measures and rating-based measures to identify products that best indicate potential success or failure.

5. Being able to analyze how to design and title products for each product, so that it can better meet customers' tastes and increase sales.

The results of our analysis of the model are as follows:

1. The relationship between emotional polarity and star rating is very significant: the higher the star rating, the more positive the content of the comment; the emotional tendency of the comment and the net likes are more significant: the higher the net likes, the more positive the comment content; There is no significant relationship with the vine reviews; the relationship with verified_purchase is also less clear.

2. In general, the reputation of a product gradually changes to a certain evaluation over time and the overall approach to a positive evaluation, but there are also cases where the reputation gradually decreases over time.

3. When a product has a low star rating, it will increase the negative reviews of customers; the appearance of a high star rating is only for reference, and it will not significantly increase the positive reviews of the product.

4. There is indeed a significant relationship between extreme words and star ratings, and reviews with more positive words have higher star ratings: reviews with more negative words have lower star ratings.

Here are also some suggestions below to help your company's marketing and product design based on the frequency of these words. By mathematically modeling, we found that there is a significant relationship between words appearing in some titles and star ratings.

Therefore, for the title of the product, we recommend that your company include “precise”, “powerful” in the title of the hair dryer product. Add “diode”, “turntable” to the titles of microwave oven products. Add “fanatic”, “fisher-price”, “little”, “limited” to the title of the pacifier product. The number of pacifier products sold at a time is preferably six.

Similarly, in terms of product characteristics, we recommend that your company add ergonomic design to hair dryer products to increase power appropriately. Add power to microwave oven products and use a white case. Added pacifier products to puppy design categories.

Best wishes!

Reference

[1] Bao Jibin.Emotioal Analysis of Product Reviews of Online Shopping Platform Based on Machine Learning[D]. Harbin Institute of Technology,2019.

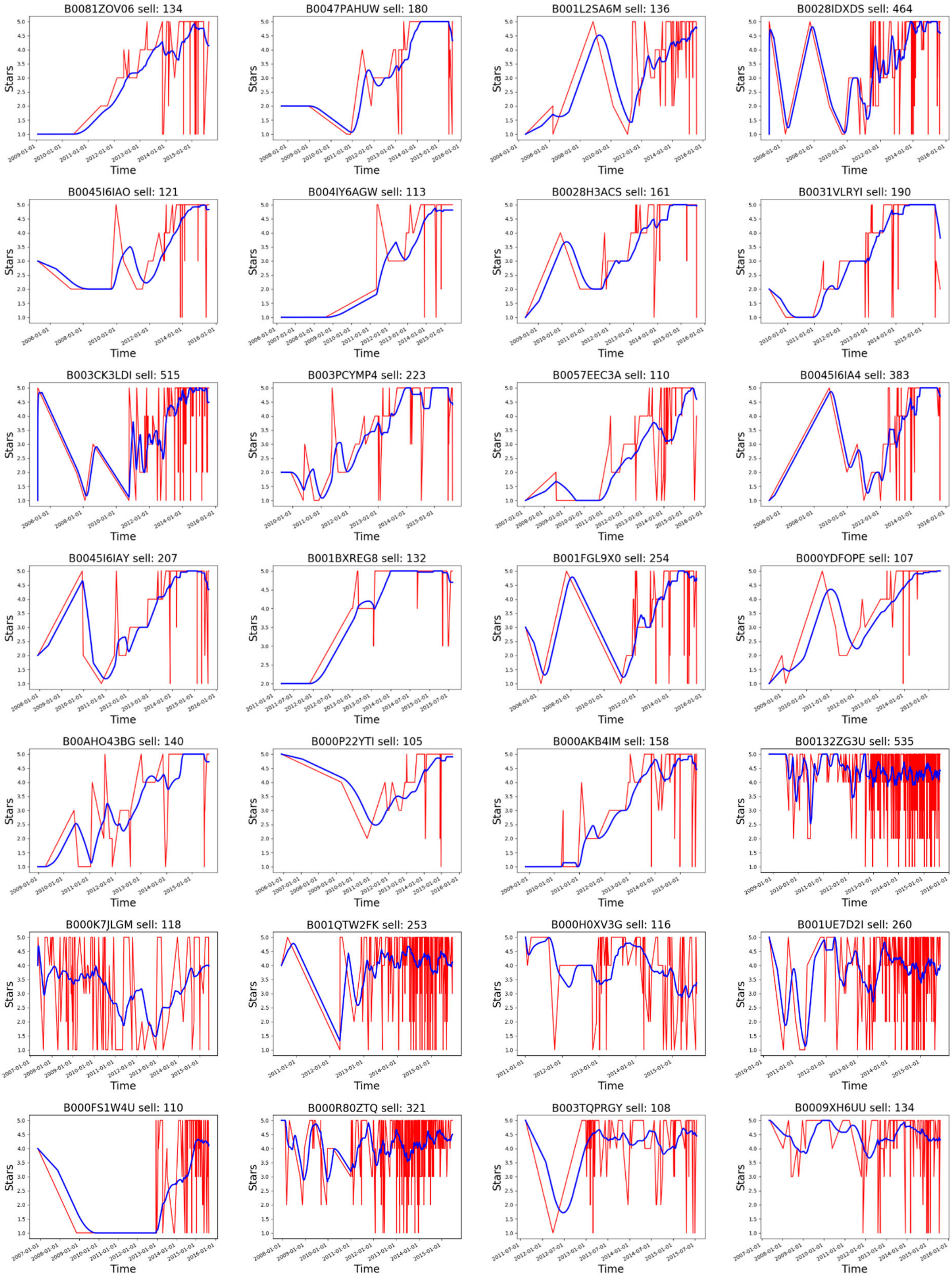
[2] Xu Qing .Empirical Research on the Impact of Online Seller Reputation on Transaction Price—Panel Data Mining Based on Taobao[C].China Academic Journal Electronic Publishing House, 2012:18-25.

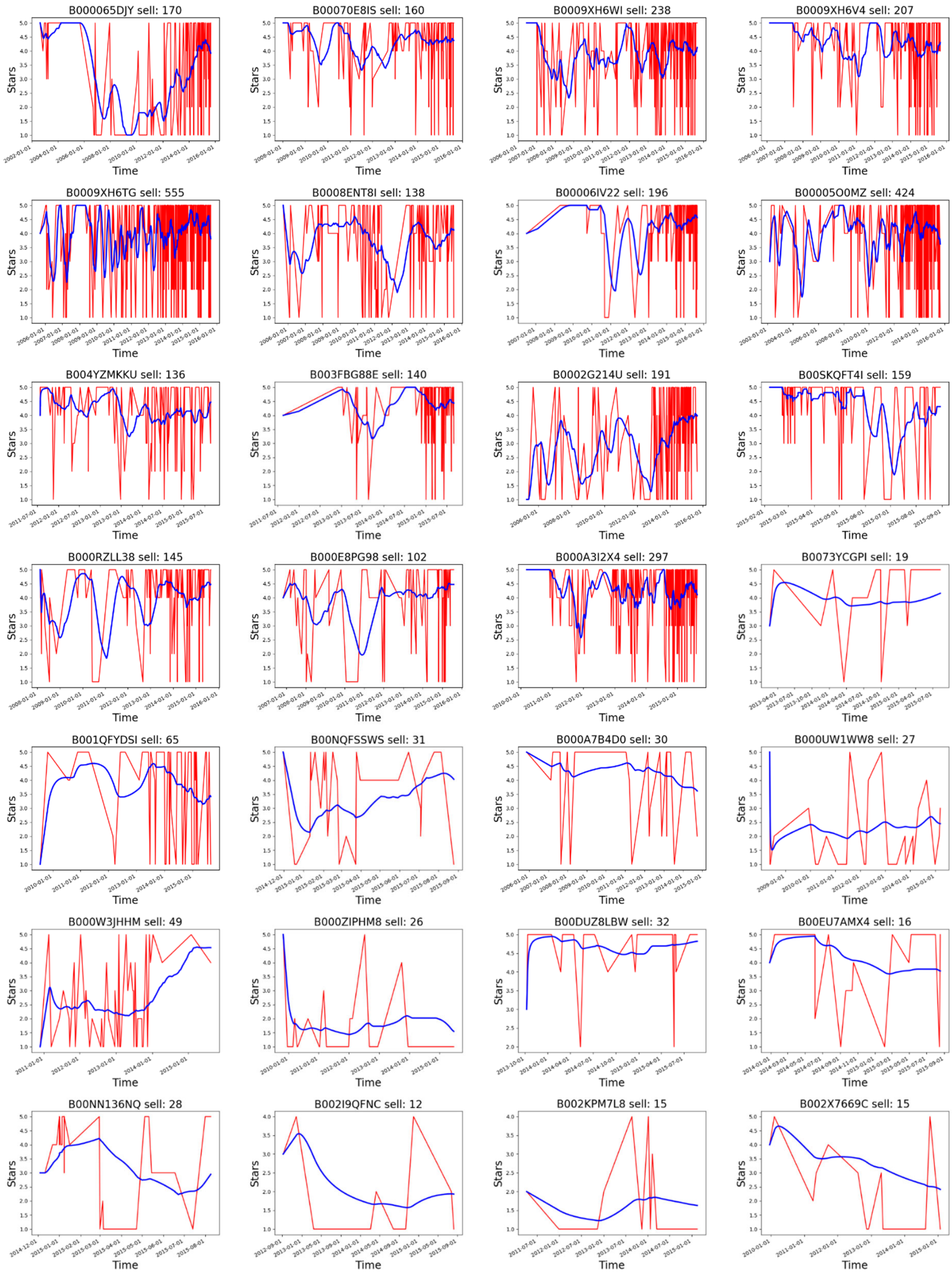
[3] Chen Yi. Application Research of Text Mining in Online Shopping User Reviews[D]. Zhejiang Gongshang University,2018.

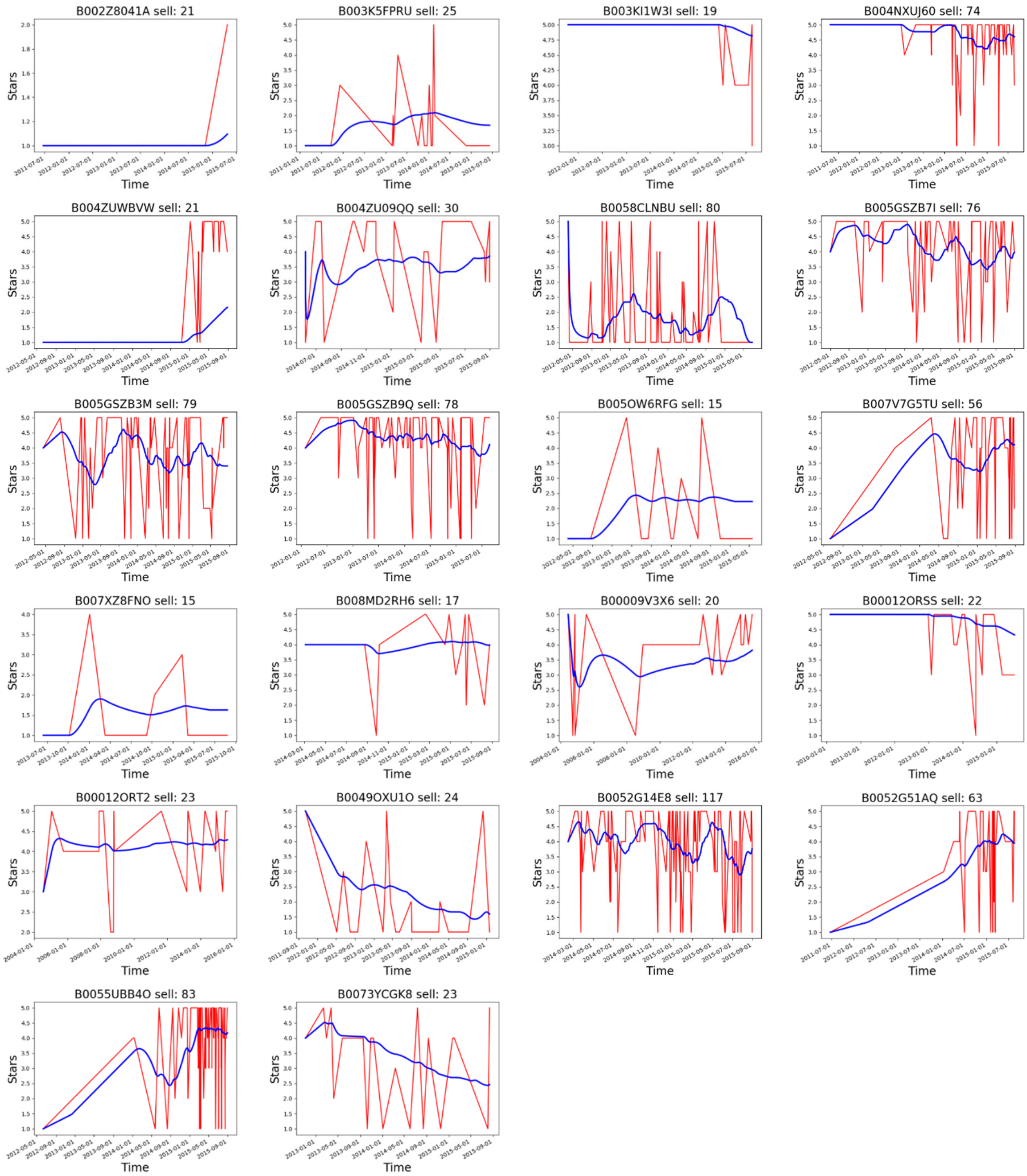
[4] Zhou Qiang. Corpus-based and Statistical-oriented Natural Language Processing technology[J] Computer Science,1995(04):36-40.

[5] Chen Qiang. Advanced Econometrics and Stata Applications (Second Edition) [M]. Higher Education Press, 2014.

Appendix: Images of star trend of products




B000065DJY sell: 170
B00070E8IS sell: 160
B0009XH6WI sell: 238
B0009XH6V4 sell: 207
B0009XH6TG sell: 555
B0008ENT8I sell: 138
B00006IV22 sell: 196
B00005O0MZ sell: 424
B004YZMKKU sell: 136
B003FBG88E sell: 140
B0002G214U sell: 191
B00SKQFT4I sell: 159
B000RZLL38 sell: 145
B000E8PG98 sell: 102
B000A3I2X4 sell: 297
B0073YCGPI sell: 19
B001QFYDSI sell: 65
B00NQFSSWS sell: 31
B000A7B4D0 sell: 30
B000UW1WW8 sell: 27
B000W3JHHM sell: 49
B000ZIPHM8 sell: 26
B00DUZ8LBW sell: 32
B00EU7AMX4 sell: 16
B00NN136NQ sell: 28
B002I9QFNC sell: 12
B002KPM7L8 sell: 15
B002X7669C sell: 15
Stars
Time


B002Z8041A sell: 21
B003K5FPRU sell: 25
B003KI1W3I sell: 19
B004NXUJ60 sell: 74
B004ZUWBVW sell: 21
B004ZU09QQ sell: 30
B0058CLNBU sell: 80
B005GSZB7I sell: 76
B005GSZB3M sell: 79
B005GSZB9Q sell: 78
B005OW6RFG sell: 15
B007V7G5TU sell: 56
B007XZ8FNO sell: 15
B008MD2RH6 sell: 17
B00009V3X6 sell: 20
B00012ORSS sell: 22
B00012ORT2 sell: 23
B0049OXU1O sell: 24
B0052G14E8 sell: 117
B0052G51AQ sell: 63
B0055UBB4O sell: 83
B0073YCGK8 sell: 23
Stars
Time