Problem Chosen	2020	Team Control Number
C	**MCM/ICM** **Summary Sheet**	**2003148**

Summary

With the vigorous development of E-commerce industry, more and more companies like SunShine are constantly emerging, and they hope to hit the market and gain people's hearts with their products. As a result, there is a growing need to help companies make business decisions through data analysis.

In this paper, we analyzed the user evaluation behaviors of the three vertical markets (hair dryer, microwave oven and pacifier), and obtained a series of quantitative models based on reviews, rating, helpful votes and time series. We also studied the relationship between certain critical variables. Finally, we derived an evaluation model that could guide the company's decision-making and provide suggestions to make the product successful.

To be more precise, first, we preprocess the data, including the processing of abnormal values and normalization of the data. We combine the indicators that belong to the same product, and sum up or average the values according to the characteristics of different indicators. In some problems, we normalized data with *"beneficial"* character, which means the higher the value, the greater the indicator.

Second, we performed sentiment analysis using *Natural Language Processing"*. After processing, we obtain the sentiment value range from -1 to 1 for each text-reviews. The more close to 1, the more satisfaction are expressed in the review; the more close to -1, the more negative emotions are expressed. We use the newly obtained values to describe satisfaction level of the reviews.

Third, we used *"Multiple Linear Regression Analysis"* to conduct a functional relation between the success of the product and many influence factors.

Then, we used *"Principal Component Analysis" (PCA)* to find the most informative data track, and obtained the formula for the principal components which we need to track in the online marketplace. We used Time Series Model to address the problem related to review data. Combined with holiday factors, we first conducted a predicting model based on the data from the beginning to 2014, and then we used the data of 2015 to test the accuracy of the model. We conclude that the model is quite acceptable.

After that, we used *"Linear Fitting"* to discover the combination of text-reviews and ratings, and obtained a good linear regression model with high level of fitness. We draw 2-demensional Scatter Plots to study the relations between specific star rating and text-reviews. The result shows the correctness of our intuitive understanding.

Finally, we use *"Correlation Analysis"* to determine the relationship between specific descriptors and rating levels. We deduce that there is a significant relationship between those factors. We also provide several advices to the company based on our analysis.

Keywords: Market Data Analysis; NLP; Multiple Linear Regression; PCA; Time Series; Linear Fitting; Correlation Analysis

Contents

A Letter to the Marketing Director of Sunshine Company

Dear Marketing Director of Sunshine Company,

It is our honor to provide you our suggestions after performing data analysis and modeling based on the online product evaluation system.

To begin with, many companies are willing to grasp customers' feedback, but it is really annoying and time-consuming if the colleague tries to dig out quantitative useful information by looking up all the customers' reviews online manually. We solve this problem by implementing "Natural Language Processing" technology, which can transform text-reviews into digital numbers, representing the level of customers satisfaction. We also use these transformed digital numbers to do other quantitative analysis. We develop a simple linear model, which is easy to understand but provide high accuracy. We also determine a new index though a convincing mathematical analysis. This index can be considered as the most informative and important factor for the success of products.

Second, your company must be interested in how the sales number of product distributed among the year, even among the week. With the help of Time Series Theory and Python library "fbprophe", we build a reliable time-based model. We tested our model with real data of 2015, and the model great fits the data of 2015. We obtained several interesting conclusions. For instance, the products usually have the highest reputation on Thursday during the whole week, and the indicator "star_rating" has a huge fluctuate on holiday (Christmas, The Black Friday etc.).

Third, We prove that there is a relationship between the star rating and text reviews by testing their correlation, determining whether there is a bond between specific descriptors and rating points. Besides, we also studied how these two factors collectively affect the success of the product.

Based on our analysis, we can offer you the following two suggestions:

- On the one hand, the company should pay more attention on the influence of powerful reviews, especially those have more votes or helpful votes, trying to preserve and enhance exposure of positive reviews and avoid high-spotted negative reviews. On the other hand, company should give more chances to have a taste of newly-published products for loyal customers, and invite them to make great comments on the product.
- By conducting word cloud figures from the data, the main concern of customers are the quality of products. For instance, customers prefer powerful and small hair dryer with retractable body. They are interested in powerful microwave oven with quality warranty. They like small and durable baby parcifiers. Your company can pay more attention to these product characteristics.

We hope that our results would be helpful for you to improve your management and make a better online marketing strategy. Wish you could make greater success in your business. We are looking forward to hearing from you.

Best regard,

Team 2003148

1 Introduction

1.1 Background

People need to consider lots of things to make a decision of purchasing, one of the essential factors that effect peoples decision-making is the feedback from buyers. Assessments from others in a sense represent true feelings of ordinary people towards the product. On the other hand, manufacturers can incorporate and review product attributes when a product is launched and over time correct product issues, understand customer requirements, and maintain customer satisfaction.[5]

Digital networks give new opportunity for consumers by allowing them to easily share their opinions and experiences with other members of large-scale online communities instead of traditional word-of-mouth social networks. It also provides an effective way for the retailers, vendors and manufacturers to analyze the consumers' interests and demands, in order to better update and construct new products, as well as to support business decision making.

Much of the work involved in ratings and reviews analysis have reached several conclusions. Samita, Chen and Smith (2007) found that reviewer information has a stronger impact on less popular products than on more popular ones.[3] Callen (2013) used sentiment analysis to determine which features of text are indicative of the context (positive, negative, objective, subjective, etc.).[5] Cristian (2009) found that a reviews perceived helpfulness depends not just on its content, but also the relation of its score to other scores.[1] L. Jack and Tsai (2015) outlined a method to apply text mining to understand consumer feedback about purchased products.[6]

1.2 Problem Restatement

One of the biggest online selling platform Amazon provides customers with three fundamental evaluating funtions:

- Individual ratings - called **star ratings** - allow purchasers to express their level of satisfaction with a product using a scale of 1 (low rated, low satisfaction) to 5 (highly rated, high satisfaction).
- Text-based messages - called **reviews** - that express further opinions and information about the product.
- Ratings on these reviews as being helpful or not - called **helpfulness rating** - towards assisting their own product purchasing decision.

In this work, we focus on the following problems:

1. Determine a mathematical model or pattern to best describe the evaluating system and support the company to success in their three new products.
2. Identify the most informative data measures based on rating and reviews.
3. Identify and discuss time-based measures to predict the tendency of the product.

4. Determine combinations of text-based and ratings-based measures that best indicate a potentially successful or failing product.
5. Identify the relations between star ratings and reviews.
6. Identify whether certain sentiment descriptors associate with rating levels or not.

2 Assumptions and Notations

2.1 Assumptions

- The success of the product is determined by the sales volume of the product. Due to the data given in the problem, the impact of product price is not considered.
- The interactions between customers are aviliable only on the Amazon product evaluating system, without any other online platform and offline communications.
- All sales, reviews, and stars are not affected by the seller′s preferential strategy (or promotion strategy).

2.2 Notations

The primary notations used in this paper are listed in Table 1.

Table 2.1: Notations

Notations	Definition
r_{ij}	each normalized value of one indicator
a_{ij}	the original value
$\max\limits_i a_{ij}$	maximum of all values for the j-th indicator
$\min\limits_i a_{ij}$	minimum of all values for the j-th indicator
$\beta_1, \beta_2, ..., \beta_m$	regression coefficients
$F_{1-\alpha/2}, F_{\alpha/2}$	quantile values
$g(t)$	trend function
$s(t)$	periodic function
$h(t)$	function of the effects of holidays
ϵ_t	any idiosyncratic changes which are not recommodated by the model

3 Data Preprocessing

For data-analysis problem, there are usually some incomplete and abnormal values in the raw data, which may seriously affect the accuracy of analysis. So it is important to preprocess the datasets.

3.1 Data Selecting

There are quite a number of values in the given datasets of the problem.

- Some datasets have incomplete data lines, which contain only a few variable values. We delete some of the incomplete lines according to the context, and use different methods to fit the missing values which may useful in future analysis.

- Some datasets have extremely high or low values that are obviously error values. We delete the extreme values and the data lines they belong to, regarding them as abnormal values.

- Some datasets contain long list of zeros, which may affect the accuarcy of data analysis, and the indicator *"star_rating"* accepts only five discrete numbers from 1 to 5. Besides, The data is presented in the form of transaction records, which is not convenient for our analysis. Therefore, we combine the indicators that belong to the same product, and sum up or average the values according to the characteristics of different indicators. This step in a way alleviates the problems mentioned above, and also allows us to perform dimensionality reduction on the data.

3.2 Data Normalization

In order to better illustrate some conclusions from data insights, and better perform accurate and more reliable data analysis, we normalized datasets in some particular problems. The main method we use to normalize the indicators is expressed by the following formula:

$$r_{ij} = \frac{a_{ij} - \min\limits_{i} a_{ij}}{\max\limits_{i} a_{ij} - \min\limits_{i} a_{ij}} \tag{3.1}$$

where r_{ij} represents the each normalized value of one indicator, a_{ij} represents the original value, $\max\limits_{i} a_{ij}$ is the maximum of all values for the j-th indicator, $\min\limits_{i} a_{ij}$ is the minimum of all values for the j-th indicator.

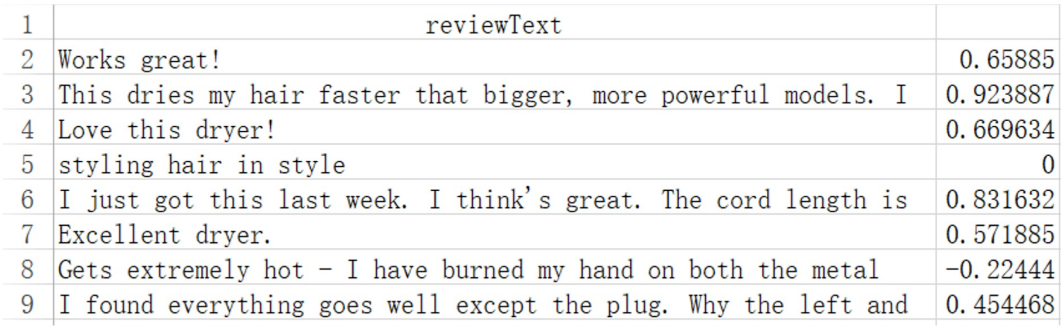
The method we apply for the normalization depends on the characteristics of different indicators. In our case, all the indicator have *"beneficial"* character, which means the higher the value, the greater the indicator. After normalization, in particular problems the indicators share the same dimension, which an avoid inaccurate results due to large differences between indicator values.

4 Models

4.1 Sentiment Analysis of Reviews with NLP

One of the biggest issue of our problem is we need to process thousands of text-reviews, extract from them valuable information, and in some way transform them into measurable numbers that we could handle and use by doing sentiment analysis. The **Natural Language Processing (NLP)** technology is a powerful tool to deal with this problem. NLP is based on statistical machine learning, automata, formal logic, formal grammar and many other knowledges from Computer Science, Statistics and Linguistics.

We use MATLAB R2019b to perform sentiment analysis. The new version of Matlab update brand new Text Analysis Toolbox in which we can use well-programmed function. After processing, we obtain the sentiment value range from -1 to 1 for each text-reviews. The more close to 1, the more satisfaction are expressed in the review; the more close to -1, the more negative emotions are expressed in the review. We add new column for each dataset named *"review_value"*, to represent the values we get. In the future analysis we will use the values in this new column instead of raw text-reviews.



1	reviewText	
2	Works great!	0.65885
3	This dries my hair faster that bigger, more powerful models. I	0.923887
4	Love this dryer!	0.669634
5	styling hair in style	0
6	I just got this last week. I think's great. The cord length is	0.831632
7	Excellent dryer.	0.571885
8	Gets extremely hot - I have burned my hand on both the metal	-0.22444
9	I found everything goes well except the plug. Why the left and	0.454468

Figure 4.1: Example of NLP results

4.2 Product Analysis Based On Online Evaluation System

For the marketing department of a company, what they want to know most is what external and internal factors are related to the success of a product, so that they can adopt more precise marketing methods, and learn some key influencing factors, such as how can the characteristics of the product better meet the needs of users, and pass it to other departments of the company to make corresponding actions. To make it clearer mathematically, we need to get the functional relationship between a product's success degree y and many influence factors x_1, ..., x_n, that is:

$$y = f(x_1, x_2, ...x_n). \tag{4.1}$$

Based on the data provided by the problem, we assume that the success of the product is equivalent to the sales number of the product. According to the datasets, we extract six basic variables: *"star_rating"*, *"helpful_votes"*, *"total_votes"*, *"vine"*, *"review_value"* and *"review_length"*,The method we use is called **Mutilple Linear Regression Analysis** , which it is widely used to solve the problem of multivariate model fitting.

The model is expressed by:

$$\begin{cases} y = \beta_0 + \beta_1 x_1 + ... + \beta_m x_m + \varepsilon, \\ \varepsilon \sim N(0, \sigma^2), \end{cases} \tag{4.2}$$

where β_0 , β_1 , ... , β_m , σ^2 are all irrelevant unknown parameters, β_0 , β_1 , ... , β_m are called regression coefficients.

The basic steps of MLRA are following:

1. Estimate parameters β_0 , β_1 , ... ,β_m in the model using least squares method.
2. Test whether there is a strong linear relationship between the dependent variable y and the independent variables x_1 , x_2 , ..., x_n : if inequality

$$F_{1-\alpha/2}(m, n-m-1) < F = \frac{U/m}{Q/(n-m-1)} < F_{\alpha/2}(m, n-m-1) \tag{4.3}$$

(where $F_{1-\alpha/2}$, $F_{\alpha/2}$ are quantile value, m is number of variables, n is number of samples, $Q = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n}(b_i - \hat{b_i})^2$, $U = \sum_{i=1}^{n}(\hat{b_i} - \bar{b})^2, \hat{b_i}$ is regression coefficients, and $\bar{b}$ is average value) is established at significance level α, then accept the assumption that there is no linear relationship, otherwise not accept.

3. Perform hypothesis testing and interval estimation of regression coefficients: if inequality

$$\left| t_j = \frac{\hat{\beta}_j/\sqrt{c_{jj}}}{\sqrt{Q/(n-m-1)}} \right| < t_{\alpha/2}(n-m-1) \tag{4.4}$$

(where c_{jj} is the (j, j)th element of $(X^T X)^{-1}$, $t_{\alpha/2}$ is upper $\alpha/2$ quantile value) is established for some compounds of statistical vector, then corresponding coefficients can be abandoned, otherwise they should be accepted.

4. Predict dependent variable with regression model.

We applied MLRA in this problem with software Matlab. Based on our analysis and calculation, we get a linear expression which illustrate the relationship between *verified_purchase* and other variables:

$$y_{dryer} = 0.616 + 0.703x_1 - 0.039x_2 + 0.306x_3 + 1.775x_4 - 0.002x_5 - 0.002x_6, \tag{4.5}$$

where x_1 , x_2 , ..., x_6 stand for *"star_rating"*, *"helpful_votes"*, *"total_votes"*, *"vine"*, *"review_value"* and *"review_length"*.

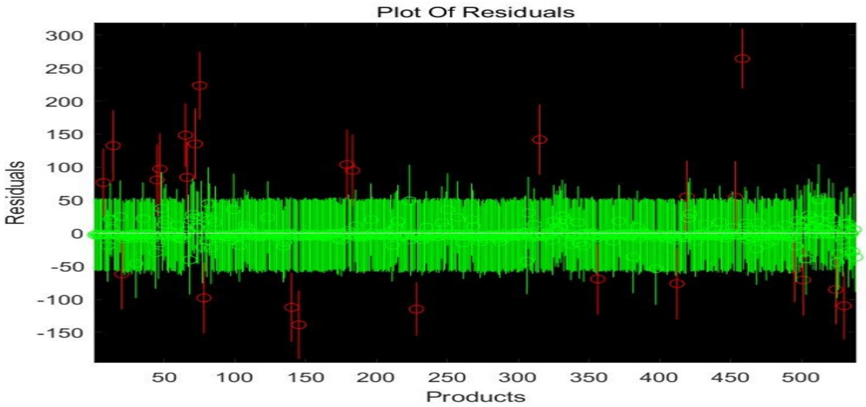


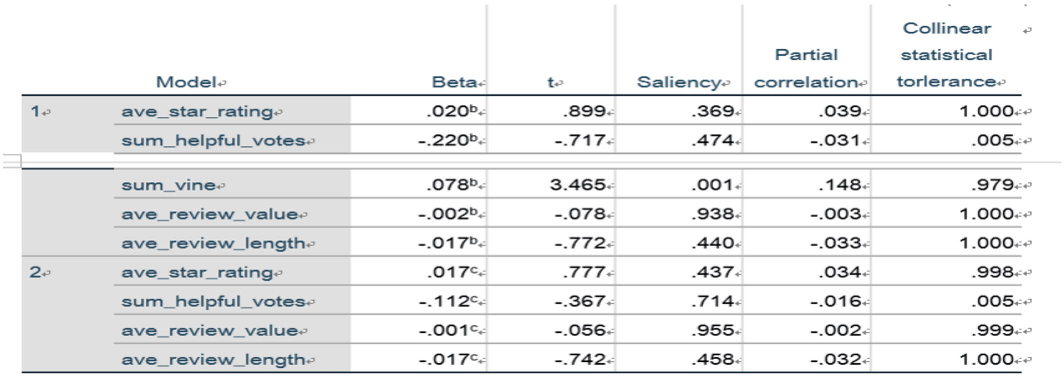
Figure 4.2: Plot of Residuals for hair dryer

The relationship between *"verified_purchase"* and independent variables can be regarded as linear relation, which is confirmed by the hypothesis testing given in the

MLRA model. We can alse see from the Figure 4.2 above that the residual is small, which means a good fit of linear model.

From the formula we obtained, its turned out that *"star_rating"*, *"total_votes"* and *"vine"* are positively related to *"verified_purchase"*, and *"helpful_votes"*, *"review_value"* and *"review_length"* are negatively related to *"verified_purchase"*.

In the estimation of regression coefficients, there are three variables which are accepted to be abandoned, leaving *"total_votes"* and *"vine"* to be the most influential factors of our model.This fact has also been verified by the implementation of MLRA in the statistical software SPSS.



Model		Beta	t	Saliency	Partial correlation	Collinear statistical torlerance
1	ave_star_rating	$.020^{b}$	.899	.369	.039	1.000
	sum_helpful_votes	$-.220^{b}$	-.717	.474	-.031	.005
	sum_vine	$.078^{b}$	3.465	.001	.148	.979
	ave_review_value	$-.002^{b}$	-.078	.938	-.003	1.000
	ave_review_length	$-.017^{b}$	-.772	.440	-.033	1.000
2	ave_star_rating	$.017^{c}$	.777	.437	.034	.998
	sum_helpful_votes	$-.112^{c}$	-.367	.714	-.016	.005
	ave_review_value	$-.001^{c}$	-.056	.955	-.002	.999
	ave_review_length	$-.017^{c}$	-.742	.458	-.032	1.000

Figure 4.3: SPSS Implementation

From the Figure 4.3 above we can see that there are 4 variables being thrown away, and leave two variables *"total_votes"* and *"vine"*.

From the discussion above, we can reach some initial conclusions: the most influential factors are the numbers of votes for the reviews and the vines. It indicates that:

- On the one hand, the company should pay more attention on the influence of powerful reviews, especially those have more votes or helpful votes (because helpful votes is the majority of all votes), trying to preserve and enhance exposure of positive reviews and avoid high-spotted negative reviews, which attract more votes and effect the products sale.
- On the other hand, company should give more chances to have a taste of newly-published products for loyal customers, and invite them to make great comments on the product, because these customers really prove how good the product is to make them be loyal to the company, and their deep insight of product using experience can really impress other customers and help new customers fully understand what are the advantages and disadvantages of the product.

The other two datasets are analyzed in the same way. Here we give the expression of those two products.

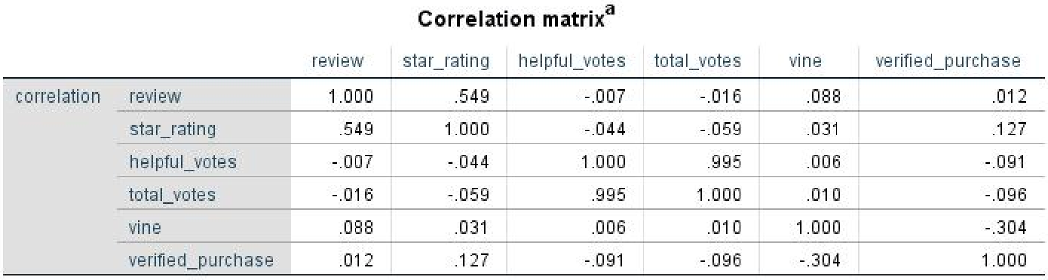
$$y_{microwave} = -1.370 + 3.739x_1 - 0.198x_2 + 0.245x_3 - 2.939x_4 + 2.926x_5 - 0.015x_6, \quad (4.6)$$

$$y_{parcifier} = -0.435 + 0.534x_1 - 0.877x_2 + 1.124x_3 + 2.393x_4 + 0.012x_5 - -0.003x_6. \quad (4.7)$$

4.3 The Most Informative Data Track

In determining the most informative tracking measure, we used **Principal Component Analysis (PCA)**. The main purpose of PCA is to use fewer variables to explain most of the original data, to convert many of the highly relevant variables into a small number of independent variables.

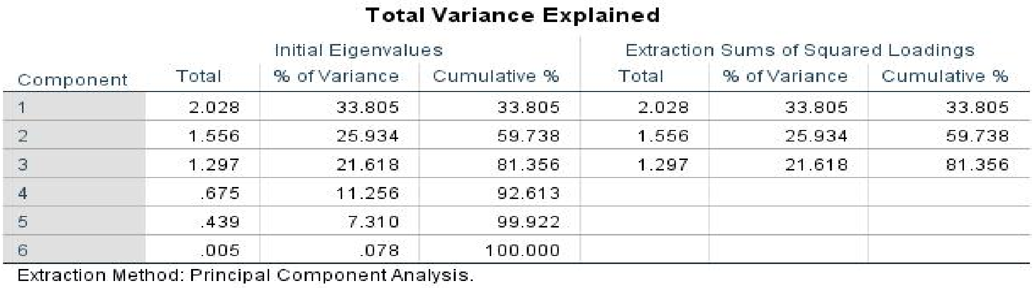
Here we use the *"hair_dryer"* data set for the demonstration of analysis. We apply the same method to the other two datasets. Quantize variables *"review_value"*, *"star_rating"*, *"helpful_votes"*, *"total_votes"*, *"vine"*, and *"verified_purchase"* and import them into statistical software SPSS for analysis. Through Factor Analyze in Data Reduction we obtain **Correlation Matrix** (Figure 4.4), **Total Variance Explained** (Figure 4.5), **Scree Plot** (Figure 4.6), and **Component Matrix** (Figure 4.7). Three principal components were finally extracted.



Correlation matrix[a]

		review	star_rating	helpful_votes	total_votes	vine	verified_purchase
correlation	review	1.000	.549	-.007	-.016	.088	.012
	star_rating	.549	1.000	-.044	-.059	.031	.127
	helpful_votes	-.007	-.044	1.000	.995	.006	-.091
	total_votes	-.016	-.059	.995	1.000	.010	-.096
	vine	.088	.031	.006	.010	1.000	-.304
	verified_purchase	.012	.127	-.091	-.096	-.304	1.000

Figure 4.4: Correlation Matricx



Total Variance Explained

	Initial Eigenvalues			Extraction Sums of Squared Loadings		
Component	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	2.028	33.805	33.805	2.028	33.805	33.805
2	1.556	25.934	59.738	1.556	25.934	59.738
3	1.297	21.618	81.356	1.297	21.618	81.356
4	.675	11.256	92.613			
5	.439	7.310	99.922			
6	.005	.078	100.000			

Extraction Method: Principal Component Analysis.

Figure 4.5: Total Variance Explained

Analyzing Figure 4.4 (Correlation Matrix), we found that there is a significant relationship between *"review_value"* and *"star_rating"* , and there is also a relationship between *"helpful_votes"* and *"total_votes"*. It turns out that many variables have a strong direct correlation, and they have overlapping information, which can be extracted by dimensionality reduction factors to get principal components.

The rule of principal component extraction is grasping the first m principal components whose corresponding eigenvalues are greater than *one*. The eigenvalue can be regarded as a strength indicator of the principal component. If the eigenvalue is less than 1, it means that the explanatory strength of the principal component is not as good as the average explanatory strength by directly introducing an original variable. So eigenvalues greater than one is widely used as inclusion criteria.

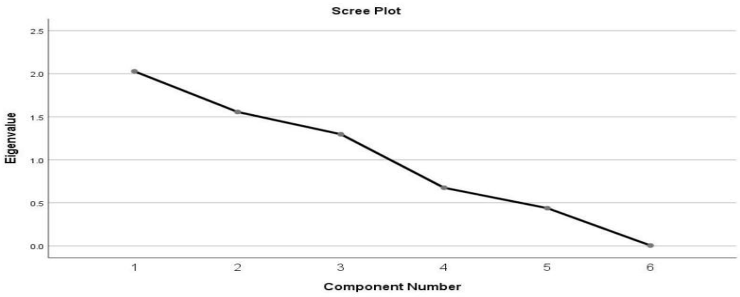
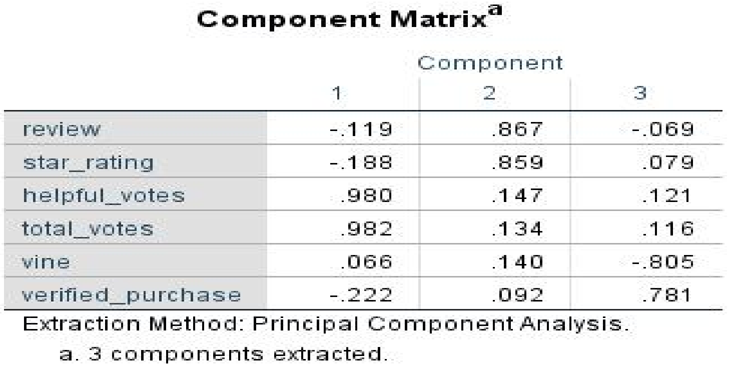


Figure 4.6: Scree Plot

From Figure 4.6 (Scree Plot), it can be seen intuitively that there are three components with eigenvalues greater than 1. At the same time, according to Figure 4.5 (Total Variance Explained), it can be known that 3 principal components are extracted, that is m = 3.



Component Matrix[a]

	Component		
	1	2	3
review	-.119	.867	-.069
star_rating	-.188	.859	.079
helpful_votes	.980	.147	.121
total_votes	.982	.134	.116
vine	.066	.140	-.805
verified_purchase	-.222	.092	.781

Extraction Method: Principal Component Analysis.

a. 3 components extracted.

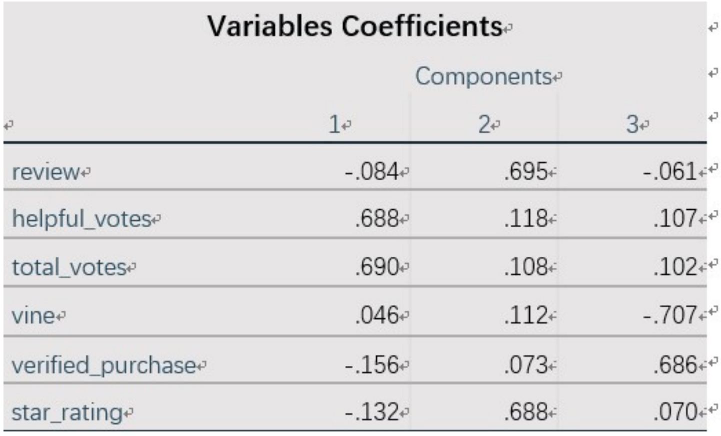
Figure 4.7: Component Matrix

From Figure 4.7 (Component Matrix), it can be seen that the *"helpful_votes"* and *"total_votes"* have a higher load on the first principal component, indicating that the first principal component basically reflects the information of these two variables; the *"review_value"* and *"star_rating"* have a higher load on the second principal component, indicating that the second principal component basically reflects the information of these two variables; and *"verified_purchase"* has a higher load on the third principal component, indicating that the third principal component basically reflects the information of this variable. So The extracted 3 principal components basically reflect all the original variables, so we can replace the original six variables with 3 new variables.

But the expressions of these three new variables cannot be directly obtained from the output window, because in "Component Matrix", each load amount represents the correlation coefficient between the principal component and the corresponding variable.

Dividing the data in Figure 4.5 (Total Variance Explained) by the square root of the eigenvalue corresponding to the principal component, we obtain the coefficients corresponding to each of the three principal components (Figure 4.8).

Then, the original six index data are normalized in SPSS and obtain variables $\tilde{x}_1$, $\tilde{x}_2$, $\tilde{x}_3$, ..., $\tilde{x}_6$, and the normalized data vector is pointwise-multiplied with the obtained



Variables Coefficients

	Components		
	1	2	3
review	-.084	.695	-.061
helpful_votes	.688	.118	.107
total_votes	.690	.108	.102
vine	.046	.112	-.707
verified_purchase	-.156	.073	.686
star_rating	-.132	.688	.070

Figure 4.8: Variables Coefficients

eigenvalues coefficients vector. Finally the principal components are expressed by:

$$F_1 = -0.084 \times \tilde{x}_1 + 0.688 \times \tilde{x}_2 + 0.69 \times \tilde{x}_3 + 0.046 \times \tilde{x}_4 - 0.156 \times \tilde{x}_5 - 0.132 \times \tilde{x}_6, \tag{4.8}$$

$$F_2 = 0.695 \times \tilde{x}_1 + 0.118 \times \tilde{x}_2 + 0.108 \times \tilde{x}_3 + 0.112 \times \tilde{x}_4 + 0.073 \times \tilde{x}_5 + 0.688 \times \tilde{x}_6, \tag{4.9}$$

$$F_3 = -0.061 \times \tilde{x}_1 + 0.107 \times \tilde{x}_2 + 0.102 \times \tilde{x}_3 - 0.707 \times \tilde{x}_4 + 0.686 \times \tilde{x}_5 + 0.07 \times \tilde{x}_6. \tag{4.10}$$

At last, considering the ontribution rate of each component to the principal component, we obtain the principal component comprehensive model, which is:

$$F = 0.416 \times F_1 + 0.319 \times F_2 + 0.266 \times F_3. \tag{4.11}$$

The obtained comprehensive evaluation value F can be used as the company's key tracking data.

The datasets *"microwave"* and *"parcifier"* are discussed and analyzed in the some way. Here we give the formulas of principal components:

For *"microwave"*:

$$F_1 = -0.009 \times \tilde{x}_1 - 0.073 \times \tilde{x}_2 + 0.657 \times \tilde{x}_3 + 0.658 \times \tilde{x}_4 + 0.298 \times \tilde{x}_5 - 0.202 \times \tilde{x}_6, \tag{4.12}$$

$$F_2 = 0.575 \times \tilde{x}_1 + 0.635 \times \tilde{x}_2 + 0.102 \times \tilde{x}_3 + 0.095 \times \tilde{x}_4 + 0.072 \times \tilde{x}_5 + 0.492 \times \tilde{x}_6, \tag{4.13}$$

$$F = 0.523 \times F_1 + 0.477 \times F_2. \tag{4.14}$$

For *"parcifier"*:

$$F_1 = -0.103 \times \tilde{x}_1 - 0.175 \times \tilde{x}_2 + 0.68 \times \tilde{x}_3 + 0.685 \times \tilde{x}_4 + 0.048 \times \tilde{x}_5 - 0.16 \times \tilde{x}_6, \tag{4.15}$$

$$F_2 = 0.703 \times \tilde{x}_1 + 0.679 \times \tilde{x}_2 + 0.15 \times \tilde{x}_3 + 0.123 \times \tilde{x}_4 + 0.083 \times \tilde{x}_5 - 0.009 \times \tilde{x}_6, \tag{4.16}$$

$$F_3 = -0.028 \times \tilde{x}_1 + 0.079 \times \tilde{x}_2 + 0.117 \times \tilde{x}_3 + 0.109 \times \tilde{x}_4 - 0.714 \times \tilde{x}_5 + 0.676 \times \tilde{x}_6, \tag{4.17}$$

$$F = 0.432 \times F_1 + 0.315 \times F_2 + 0.253 \times F_3. \tag{4.18}$$

4.4 Time-Based Data Forecast Model

In determining time-based measures to suggest a product's reputation tendency in the online marketplace, we think that the products reputation has a positive correlation with *"star_rating"*, and with the development of technology the average of *"star_rating"* may go higher year by year, so the first thing we want to do is to find the relationship between *'review_date"* and *"star_rating"*.

What we use to find the relationship is called **"fbprophet"**, which is a time series prediction method opensourced by Facebook. According to the paper, we believe that our variables *'review_date"* and *"star_rating"* are also satisfied with the model which is called **"Decomposable Time Series Model"** (Harvey & Peters 1990). Here in the model we have three components: *trend*, *seasonality* and *holidays*, and they are combined in the following equation:

$$y(t) = g(t) + s(t) + h(t) + \epsilon_t, \tag{4.19}$$

where $g(t)$ is the trend function which models non-periodic changes in the value of the time series, $s(t)$ represents periodic changes (e.g., weekly and yearly seasonality), and $h(t)$ represents the effects of holidays which occur on potentially irregular schedules over one or more days. The error term ϵ_t represents any idiosyncratic changes which are not accommodated by the model, later we will make the parametric assumption that ϵ_t is normally distributed.

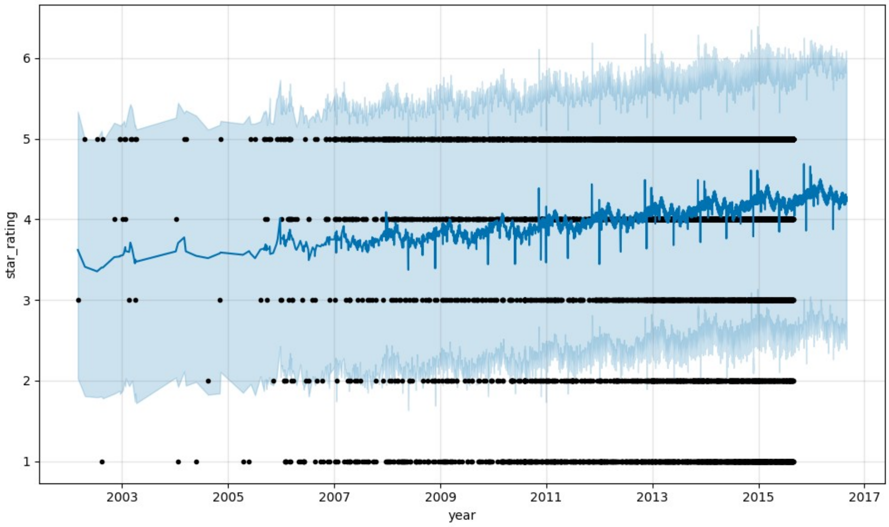


Figure 4.9: *"hair_dryer"* general trend with holiday

We reformed the data into a new form, which has the first column called *ds* and include all *"review_date"* from earliest to latest, and the second column called *y* and include all *"star_rating"* corresponding with *"review_date"*.

We first analyze *"hair_dryer"*, by programming we get two figures, which are Figure 4.9 and Figure 4.10.

In the Figure 4.9, all black dots are *"star_rating"* that we have in the datasets; the dark blue line is the prediction of *"star_rating"* (we make a prediction for next year,

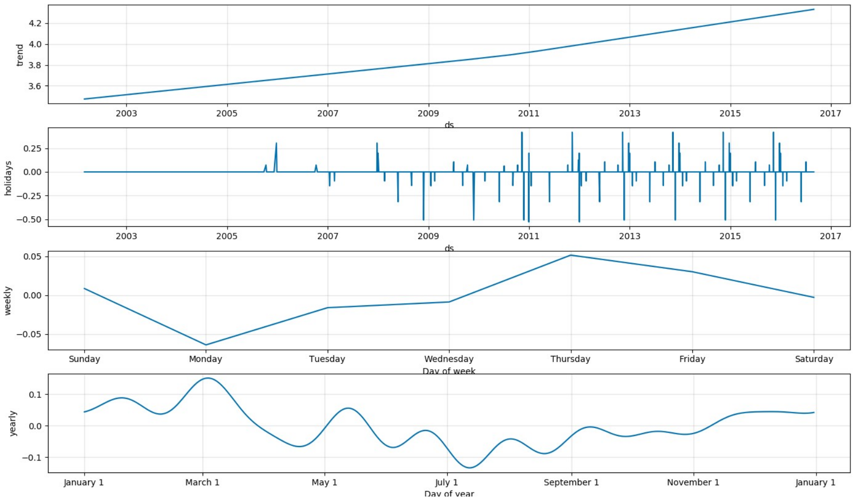


Figure 4.10: *"hair_dryer"* component analysis with holiday

which means from 2015/8/15-2016/8/15); and the light blue is flexible area (include all the possibility of *"star_rating"*). In the Figure 4.10, the first picture is called **"General Trend"**, which directly shows the change of *"star_rating"*, the second picture takes holiday into consideration, means *"star_rating"* will fluctuate as a result of holidays, and the value of *"star_rating"* will be the trend value in figure 1 plus holiday value in Figure 4.10. The third and forth picture are similar to the second one.

For the microwave and baby pacifier we implement our method in the same way.

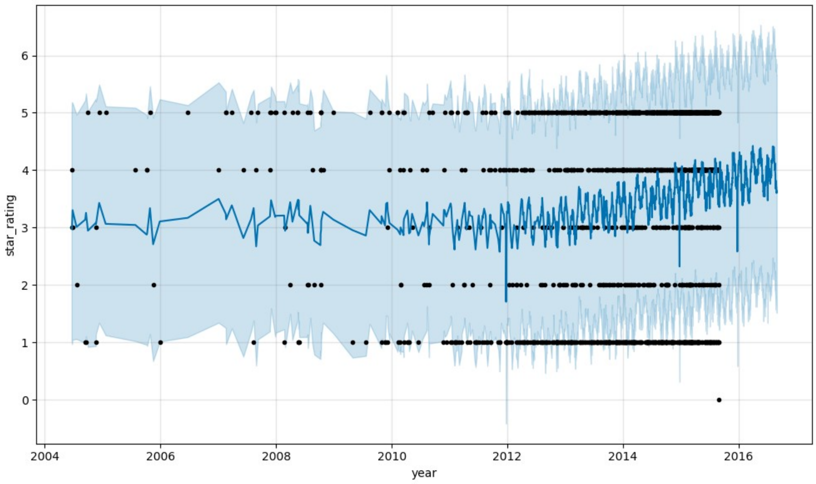


Figure 4.11: *"microwave"* general trend with holiday

From those figures, we obtain following conclusions:

1. *"star_rating"* do raise year by year;
2. *"star_rating"* has a huge fluctuate on holiday (Christmas, The Black Friday etc.);
3. *"star_rating"* usually go higher on Thursday, Friday and Saturday;
4. *"star_rating"* of *"microwave"* is more volatile than *"hair_dryer"* and *"parcifier"*;

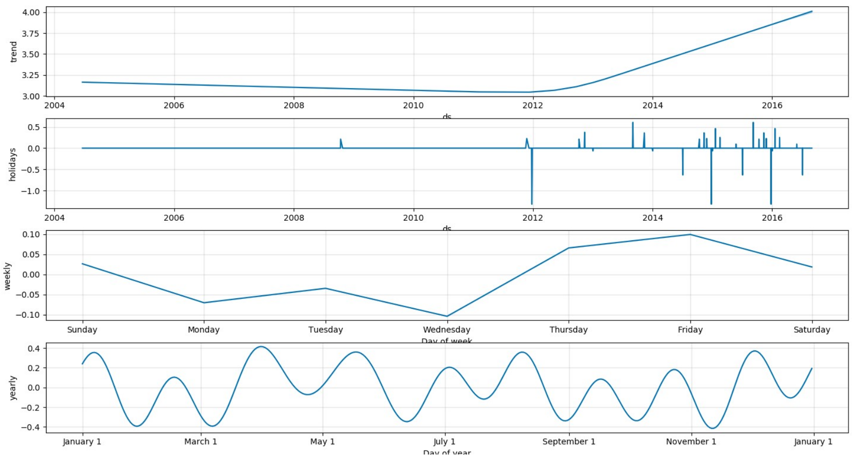


Figure 4.12: *"microwave"* component analysis with holiday

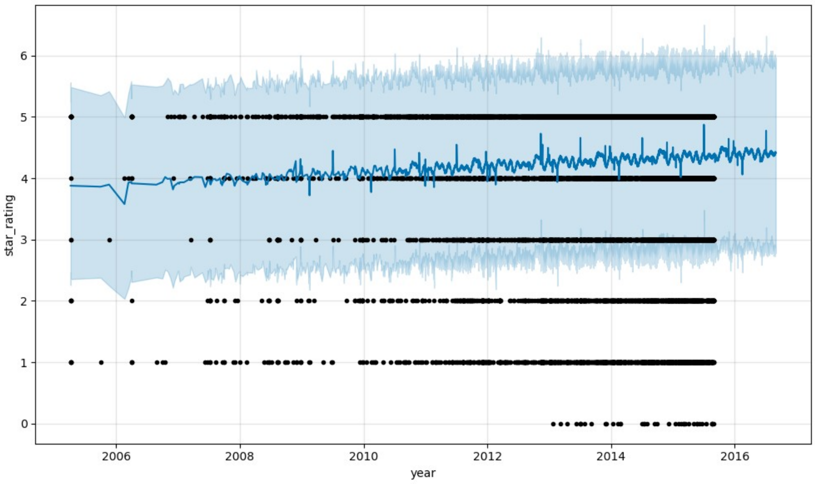


Figure 4.13: *"parcifier"* general trend with holiday

4.5 The Combination Of Texts and Rating

In determining the combination of text-based measures and ratings-based measures that best indicate a potentially successful or failing product, we use **"Linear Fitting"**, by regarding the success of the product is linear related only to the text-reviews and rating-points. The reason we use linear fitting model instead of non-linear fitting is that the goodness of fitting of linear is better than non-linear fitting. The following analysis shows that the linear expression we obtain has high-level of accuracy.

The datasets we use here are normalized, in order to better illustrate and improve the output results. The text-reviews are transformed into numbers from -1 to 1 from Natural Language Processing, in order to describe the satisfaction level of reviews. *"Zverified_purchase"* is the dependent variable representing the sales number of the product, which is the same meaning with previous sub problems. *"Zstar_rating"* and *"Zreview_value"* are the variables that we want to give weights to. We input these data into MATLAB and try to find a formula that can describe the relationship between them. The linear model we want to fit is

$$f(x,y) = p_{00} + p_{10}x + p_{01}y \tag{4.20}$$

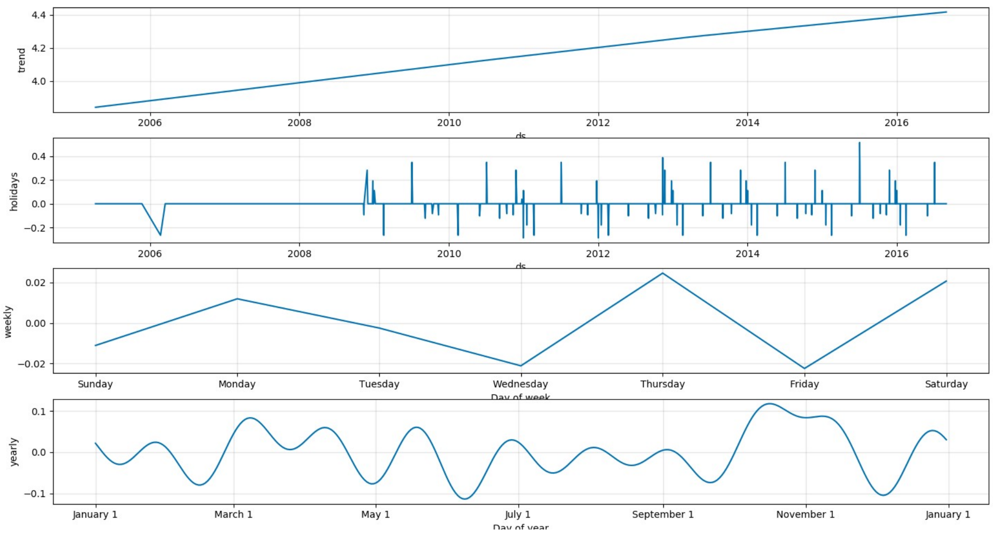


Figure 4.14: *"parcifier"* component analysis with holiday

where x represents *"Zstar_rating"*, and y represents *"Zreview_value"*.

In MATLAB we mainly use **"Curve Fitting Tool"**. After searching, we find that for *"hair_dryer"*, if we use method *"Polynomial"*, degrees of *"Zstar_rating"* and *"Zreview_value"* both equal to 1, then choose *"Bisquare"* on robust, we can get the best formula.

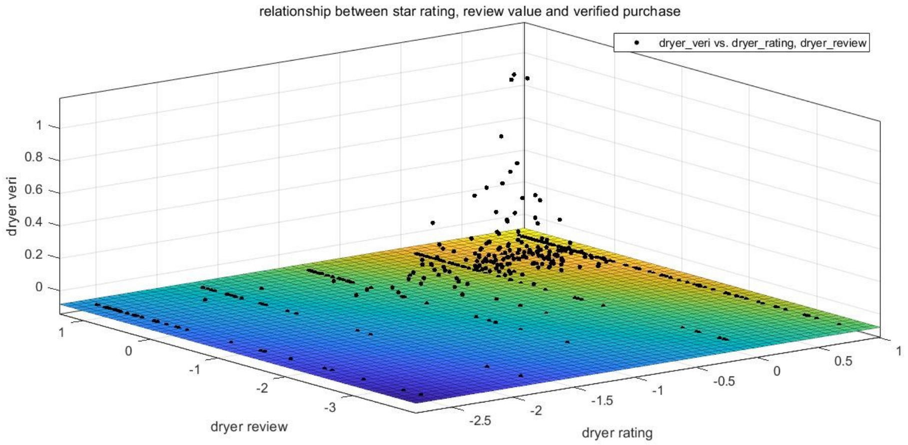


Figure 4.15: The linear-fitting 3-d graph

Finally we find an expression that best satisfies with most of observation values. We obtain coefficients p_{00} = -0.08383, p_{10} = 0.00007454 and p_{01}= 0.00003033 with 95% confidence bounds. The statistics are SSE=0.1416, R-square=0.9834, RMSE=0.01627, which show that this polynomial can describe the relationship well:

$$Z_{dryer} = -0.08383 + 0.00007 \times Z_{star\_rating} + 0.00003 \times Z_{review\_value} \tag{4.21}$$

The formula shows that *"Zstar_rating"* and *"Zreview_value"* have a positive correlation with *"Zverified_purchase"*, or simply we can say "higher star rating with more positive comment can bring more sales".

For the microwave and baby pacifier we do the same job. It has some differences between results, but generally, models are similar to each other. We omit the discussion for the next two datasets, only give the results.

$$Z_{microwave} = -0.16383 + 0.02922 \times Z_{star\_rating} + 0.00034 \times Z_{review\_value} \tag{4.22}$$

$$Z_{parcifier} = -0.02026 + 9.97324 \times 10^{-18} \times Z_{star\_rating} + -3.03876 \times 10^{-18} \times Z_{review\_value} \tag{4.23}$$

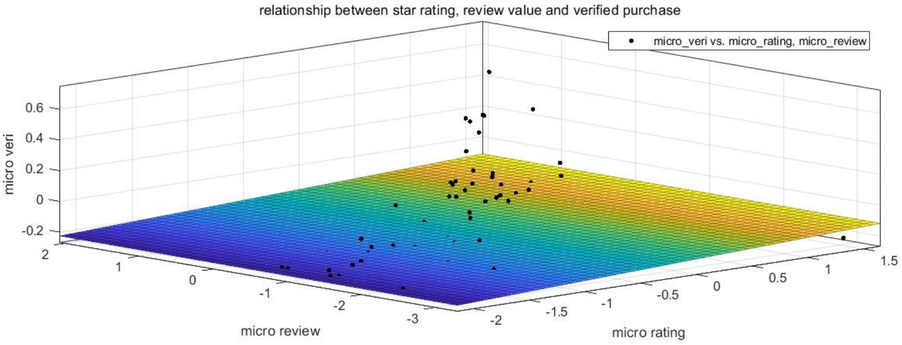


Figure 4.16: The linear-fitting 3-d graph for microwave oven

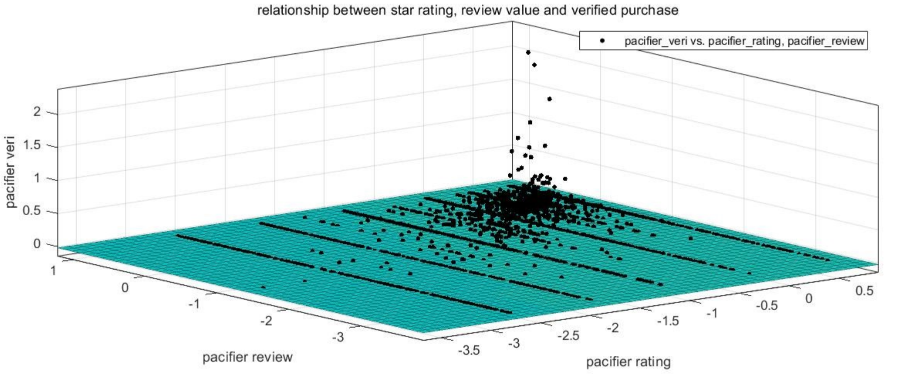


Figure 4.17: The linear-fitting 3-d graph for parcifier

4.6 Specific Star Rating And Text-Reviews

It is quite natural to reach an intuitive understanding that, the more stars voted, the more satisfied with the product. To prove that, we draw a 2-dimentional Scatter Plot with *"star_rating"* and *'review_date"*, and observe how these points are distributed in the figures. The data are normalized thus the are no corresponding relations between values and meanings.

Its turned out from these figures that most of the points are located at up-right corner, which means high star rating and high review satisfaction, in the Figure 4.19 we see a clear positive correlation between two variables. Although its tough to fit the points with a line, especially in the case of *"hair_dryer"* and *"percifier"*, but the Scatter Plots show the correctness of our basic intuitive understanding.

Many points stand in 5 straight lines, thats because *"star_rating"* only accept five discrete values from 1 to 5.

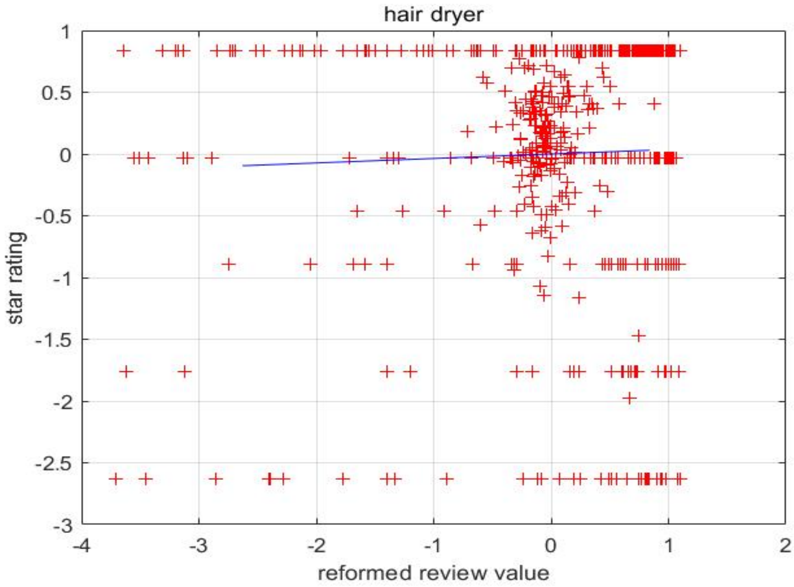


Figure 4.18: Scatter Plot of dryer

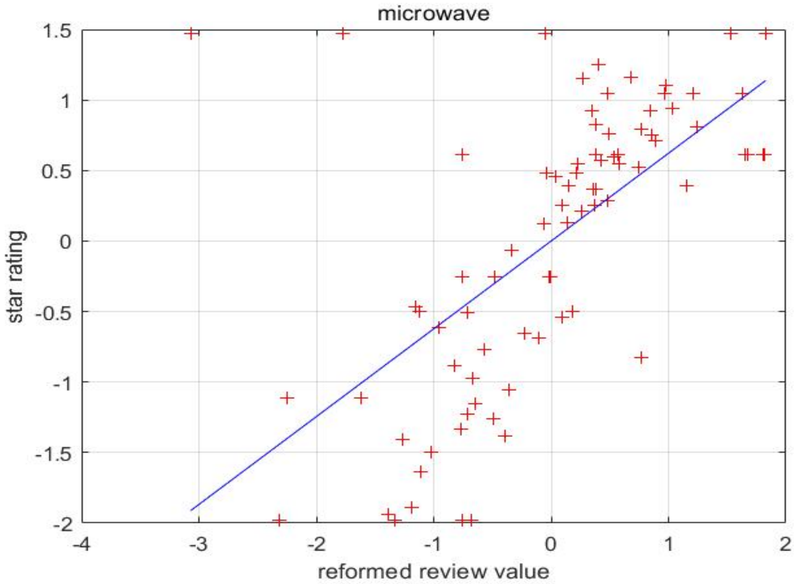


Figure 4.19: Scatter Plot of microwave

In some specific situation, where customers give 5 star to the product, but with negative and short reviews, those customers dont want to make a trouble and its beneficial in their mind for both sides. However the number of these kind of circumstances still remain low.

After reading many text-reviews, we also find an interesting fact, that many long reviews have relatively high or low star rating points, which also fits our daily experience. We make long reviews mainly in two situations: To praise the good points of the product and recommend it to others, or to criticize and complain about the bad using experience.

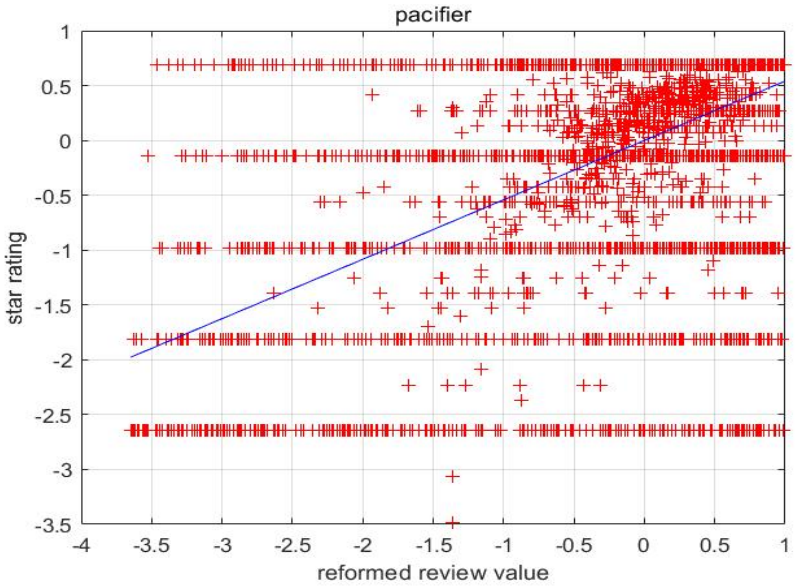


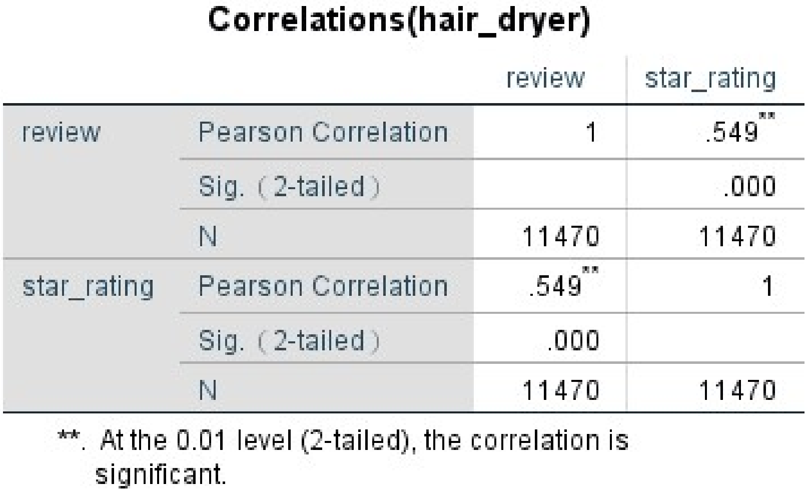
Figure 4.20: Scatter Plot of parcifier

4.7 Relationship Between Specific Descriptors And Rating Levels

In determining whether there is a certain relationship between text reviews and rating (that is, whether specific words are related to *"star_rating"*, because we have obtain *"review_value"* through NLP, we assume that certain descriptors represent levels of *"review_value"*, so we can study on the relationship between these two variables), we use SPSS software to qualitatively perform correlation tests and regression analysis on these two variables.

"Correlation Analysis" is a commonly used statistical method to study the closeness between variables. *"Linear Correlation Analysis"* studies the strength and direction of the linear relationship between two variables. The *"Correlation Coefficient"* is a statistic value describing this linear relationship. In the correlation test, the *"Pearson Correlation Analysis Method"* was used. Since it was not possible to confirm whether the correlation was positive or negative, the two-tailed t-test was selected as the significance test options. For each data set, we can obtain Correlations (Figure 4.21, Figure 4.22, Figure 4.23). In the table, Sig is the P value, and less than 0.05 is significant, a N is the sample size.

Analyzing the three correlation coefficient matrices, three values are displayed vertically at the intersection of the variable row and the variable column: the value in the first row is the correlation coefficient matrix of the row variable and the column variable. The row and column variables are the same, and the correlation coefficient is 1. For *"hair_dryer"*, the correlation between *"review_value"* and *"star_rating"* is 0.549, microcave is 0.595, and percifier is 0.501. The values in the second row are Sig. (2-tailed) two-tailed test results (probability that the hypothesis test that makes the correlation coefficient 0 is true), and the results are all 0.000 or less than 0.001. The comment line states that the significance level of the correlation coefficient marked "**" is 0.01. The results in the table show that for the three product datasets, both *"review_value"* and *"star_rating"* are significantly positive correlated.

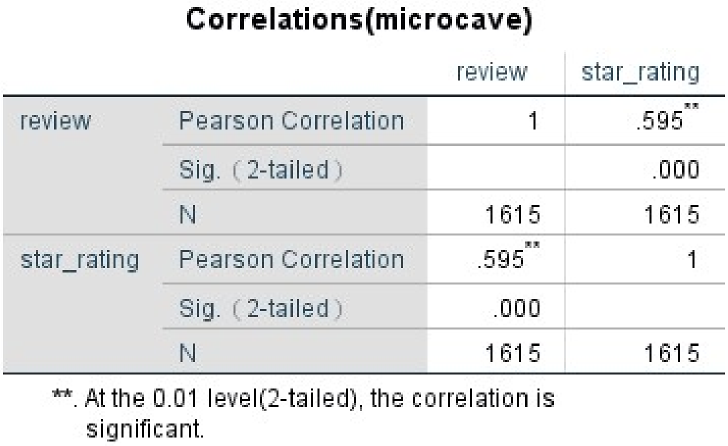


Correlations(hair_dryer)

		review	star_rating
review	Pearson Correlation	1	.549**
	Sig. (2-tailed)		.000
	N	11470	11470
star_rating	Pearson Correlation	.549**	1
	Sig. (2-tailed)	.000	
	N	11470	11470

**. At the 0.01 level (2-tailed), the correlation is significant.

Figure 4.21: Correlations("hair_dryer")



Correlations(microcave)

		review	star_rating
review	Pearson Correlation	1	.595**
	Sig. (2-tailed)		.000
	N	1615	1615
star_rating	Pearson Correlation	.595**	1
	Sig. (2-tailed)	.000	
	N	1615	1615

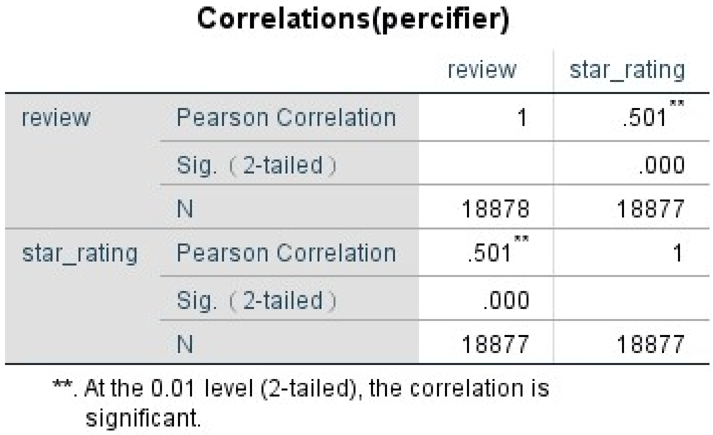
**. At the 0.01 level(2-tailed), the correlation is significant.

Figure 4.22: Correlations("microwave")

The relationship between variables is divided into two types: *"deterministic"* and *"non-deterministic"*. Functions express deterministic relationships. Mathematical statistics methods that study the non-deterministic relationship between variables and construct empirical formulas between variables are called *"Regression Analysis"*. When there is a linear relationship between the independent and dependent variables, we can construct a linear regression equation. Using SPSS software, we performed regression analysis on the scores and quantified text reviews of the three data sets, obtained three regression analysis results tables (Figure 4.24, Figure 4.25, Figure 4.26).

The table is from left to right. Mldel is the regression equation model number. Here we choose number 1 for analysis, where:

- *Unstandardized Coefficients* is the non-standardized regression coefficient.
- *Standardized Coefficients* is standardized regression coefficient.
- *t* is the t value of the hypothesis test with a partial regression coefficient of 0 (and

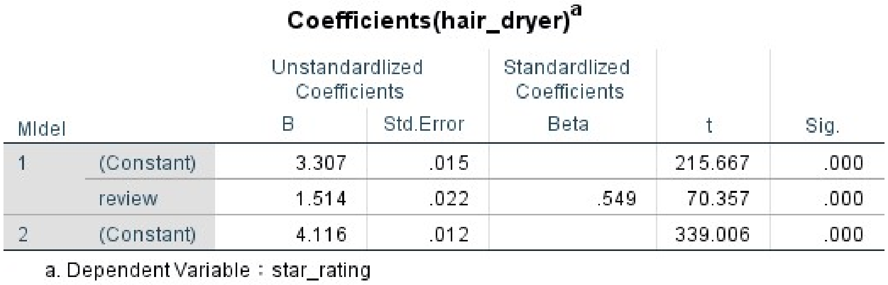


Correlations(percifier)

		review	star_rating
review	Pearson Correlation	1	.501**
	Sig. (2-tailed)		.000
	N	18878	18877
star_rating	Pearson Correlation	.501**	1
	Sig. (2-tailed)	.000	
	N	18877	18877

**. At the 0.01 level (2-tailed), the correlation is significant.

Figure 4.23: Correlations(“parcifier”)



Coefficients(hair_dryer)[a]

		Unstandardlized Coefficients		Standardlized Coefficients		
MIdel		B	Std.Error	Beta	t	Sig.
1	(Constant)	3.307	.015		215.667	.000
	review	1.514	.022	.549	70.357	.000
2	(Constant)	4.116	.012		339.006	.000

a. Dependent Variable : star_rating

Figure 4.24: Coefficient(“hair_dryer”)

a constant of 0)

- *Sig.* is the significance level value for a hypothesis test with a partial regression coefficient of 0 (and a constant of 0).
- *B* is partial regression coefficient. It is obtained after controlling other variables.
- *Beta* is a standardized regression coefficient.

Analyze the regression coefficient table of *“hair_dryer”*. From the table, we can see the estimated value and its test result. The constant term b_0 = 3.307, the regression coefficient b_1 = 1.514, the regression coefficient test statistic t = 70.357, and the significant level value of the equation constant is 0.000. The correlation probability value p < 0.001 indicates that the regression coefficient is significantly different from 0. The regression equation can be expressed as:

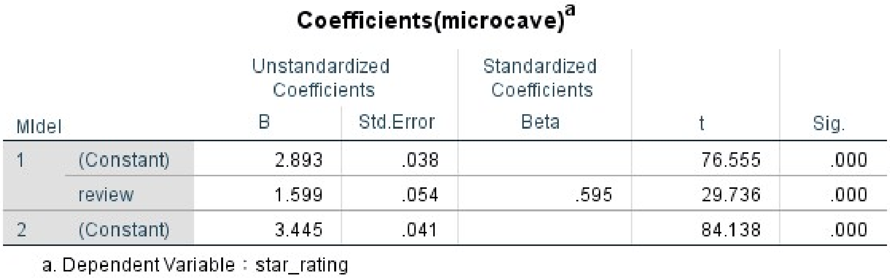
$$y_{dryer} = 3.307 + 1.514x. \tag{4.24}$$

which has statistical significance. Similarly, we can get the regression equation of microcave:

$$y_{microwave} = 2.893 + 1.599x \tag{4.25}$$

and the regression equation of parcifier:

$$y_{parcifier} = 3.516 + 1.359x. \tag{4.26}$$

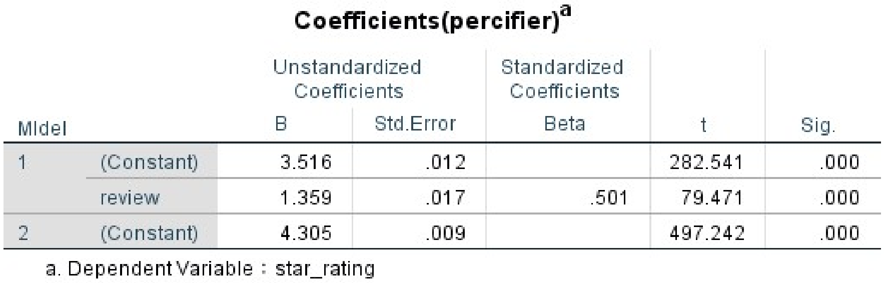


Coefficients(microcave)^{a}

		Unstandardized Coefficients		Standardized Coefficients		
Mldel		B	Std.Error	Beta	t	Sig.
1	(Constant)	2.893	.038		76.555	.000
	review	1.599	.054	.595	29.736	.000
2	(Constant)	3.445	.041		84.138	.000

a. Dependent Variable : star_rating

Figure 4.25: Coefficient("microwave")



Coefficients(percifier)^{a}

		Unstandardized Coefficients		Standardized Coefficients		
Mldel		B	Std.Error	Beta	t	Sig.
1	(Constant)	3.516	.012		282.541	.000
	review	1.359	.017	.501	79.471	.000
2	(Constant)	4.305	.009		497.242	.000

a. Dependent Variable : star_rating

Figure 4.26: Coefficient("parcifier")

From the results of correlation analysis and regression analysis, it can be known that for the three product data sets, under the condition that the other remains unchanged, when the *"review_value"* is high, *"star_rating"* will also be high, and when *"review_value"* is increased by 1, *"star_rating"* increased by 1.514, 1.599 and 1.359 respectively. Through the correlation test and regression analysis of the scoring and quantified text reviews in the three data sets, it can be concluded that there is a significant relationship between text reviews and ratings.

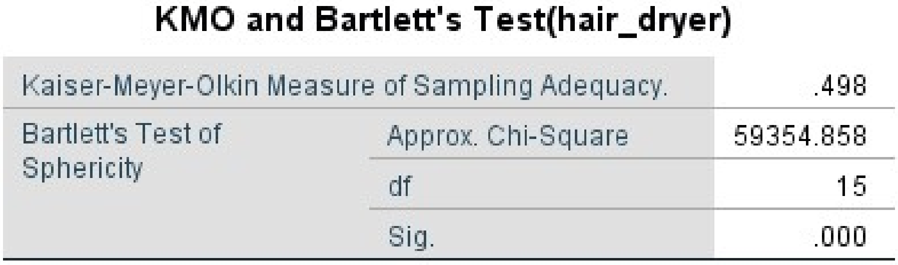
5 Strength and Weakness

5.1 Strength

- Based on the conditions of missing data and abnormal values, we take different measures to maintain the authenticity of the data as much as possible.
- We perform sentiment analysis on the text-reviews with NLP, which transforms the texts to measurable numbers, allowing us to do further research.
- We tried different ways to confirm the accuracy of our model, for example, we perform hypothesis testing and interval estimation of regression coefficients in the first problem, and conduct the formula in two platform: Matlab and SPSS.
- We use different types of mathematical and statistical methods to strength our model and analysis.
- We obtained relatively reliable results for each problems.

5.2 Weakness

- The datasets given in the problem only have transaction records with text-reviews. In the ordinary situation, there are other types of data, which may provide more information and support to the analysis.
- We can only obtain some conclusions from the aspect of analyzing online evaluating system. It will be beneficial if we could analyze the product combined with other data.
- We didnt seriously consider the impact of the product price(but lots of customers mentioned and complained about that).
- Some parameters remained to be improved. For example, when using principal component analysis, some statistic like KMO in the first dataset is not quite acceptable, as well as the statistic R^2 with relatively low correlation in MLRA.



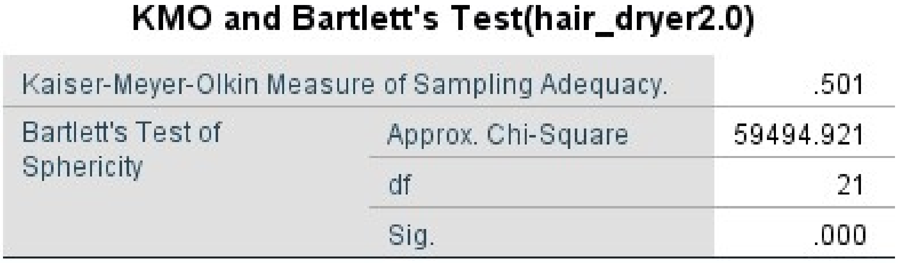
KMO and Bartlett's Test(hair_dryer)

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.498
Bartlett's Test of Sphericity	Approx. Chi-Square	59354.858
	df	15
	Sig.	.000

Figure 5.1: Inacceptable KMO

5.3 Promotion

To improve the accuracy of statistic KMO, we try to add another set of data: the number of account purchases, that is, the number of repeats of the buyer ID. After adding it to the factor analysis, it was found that the KMO value reached 0.5, which met the requirements.



KMO and Bartlett's Test(hair_dryer2.0)

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.501
Bartlett's Test of Sphericity	Approx. Chi-Square	59494.921
	df	21
	Sig.	.000

Figure 5.2: Acceptable KMO

6 Conclusions

In this work, based on three markets datum (hair dryer, microwave oven and pacifier) we have developed different mathematical models. By performing rigorous mathematical analysis, we obtain the following conclusions.

- The relationship success of product and independent variables can be regarded as linear relation with high level of fitness.
- The more stars voted, the more satisfied with the product.
- There is a significant positive correlation between text reviews and ratings.
- Higher star rating with more positive comment can bring more sales.
- *"star_rating"* raises year by year.
- *"star_rating"* has a huge fluctuate on holiday (Christmas, The Black Friday etc.).
- *"star_rating"* usually go higher on Thursday, Friday and Saturday.
- *"star_rating"* of *"microwave"* is more volatile than *"hair_dryer"* and *"parcifier"*.
- The company should pay more attention on the influence of powerful reviews, especially those have more votes or helpful votes, trying to preserve and enhance exposure of positive reviews and avoid high-spotted negative reviews, which attract more votes and effect the products sale.
- The company should also give more chances to have a taste of newly-published products for loyal customers, and invite them to make great comments on the product, because these customers really prove how good the product is to make them be loyal to the company, and their deep insight of product using experience can really impress other customers and help new customers fully understand what are the advantages and disadvantages of the product.

References

[1] C. Danescu-Niculescu-Mizil, G. Kossinets, J. Kleinberg, and L. Lee. "How opinions are received by online communities: a case study on amazon.com helpfulness votes. " In Proceedings of the 18th international conference on World wide web (2009): 141-150.

[2] A. Bhatt, A. Patel, H. Chheda and K. Gawande, "Amazon Review Classification and Sentiment Analysis. " International Journal of Computer Science and Information Technologies, Vol. 6, No. 6 (2015): 5107-5110.

[3] S. Dhanasobhon, P-Y-. Chen, and M. Smith. "An Analysis of the Differential Impact of Reviews and Reviewers at Amazon.com" ICIS 2007 Proceedings (2007): 94.

[4] A. S. Rathor, A. Agarwal, and P. Dimri. "Comparative study of machine learning approaches for Amazon reviews." Computer Science 132 (2018): 1552-1561.

[5] C. Rain. "Sentiment Analysis in Amazon Reviews Using Probabilistic Machine Learning", Swarthmore College, 2013.

[6] L. Jack, Y.D. Tsai. "Using Text Mining of Amazon Reviews to explore User-Defined Product Highlights and Issues" In Proceedings of International Conference on Data Mining, 2015(01).

[7] X-L-. Hui,L. Qiang, F. Hui. Car following safe distance model based on braking process of leading vehicle f. Journal of Guangxi Normal University(Natural Science Edition),2010,28(1):1-5.

[8] S.J. Taylor, B. Letham. "Forecasting at scale" The American Statistician, 2017.

[9] H.M. Lin, W.l. Zhang. "Similarities and Differences of Principal Component Analysis and Factor Analysis and SPSS Software". Statistical Research, 2005(03).

[10] J.X. Liu. "Factor Analysis of Comprehensive Performance of Students with SPSS Statistical Software". Statistical Education, 2006, 16 (1): 53-56.

[11] S.H. Yu "On the Application of SPSS in Educational Information Processing". Computer and Telecommunications, 2006, 15 (10): 55-58.

[12] S.K. Si, Z.L. Sun. Mathematical Modeling Algorithms and Applications .Second Edition. Beijing: National Defense Industry Press, 2019.150 158

[13] https://ww2.mathworks.cn/help/stats/regress.html

[14] https://www.mathworks.com/help/textanalytics/ug/analyze-sentiment-in-text.html

Appendices

Codes

Here are some programmes we used in our model as follow.

Input Matlab source for Sentiment analysis of reviews with NLP

```
filename = "parcifier-1.xlsx";
%filename = "dryer-1.xlsx";
%filename = "microwave-1.xlsx";
tbl = readtable(filename,’TextType’,’string’);
head(tbl)%head of xlsx
str = tbl.reviewText;%the reviews text
documents = tokenizedDocument(str);
documents(1:5)
%compoundScores is the result vector
compoundScores=vaderSentimentScores(documents)
compoundScores(1:5)
for i = 1:18939
    compoundLength(i) = length(char(str(i)))
end
compoundLengthTran = transpose(compoundLength)
```

Input Matlab source for Product Analysis Based On Online Evaluation System

```
Y = hairdryerS2(:,[6]); % Extract observed value y
% Extract observed value x
```

```
x123456 = [hairdryerS2(:,[2:5]),hairdryerS2(:,[7:8])];
% Y = microwaveS1(:,[6]);
% x123456 = [microwaveS1(:,[2:5]),microwaveS1(:,[7:8])];
% Y = percifierS1(:,[6]);
%x123456 = [percifierS1(:,[2:5]),percifierS1(:,[7:8])];
Size = size(x123456)
% Construct data matrix of multiple linear regression analysis
X=[ones(Size(1),1),x123456]
% Calculate regression coefficients and stats values
[beta,betarint,r,rint,st] = regress(Y,X)
q = sum(r.^2) % Sum of squared residuals
ybar = mean(Y) % Average value of observed y and estimated value of y
yhat = X * beta
u = sum((yhat-ybar).^2) % sum of square
m = 6 % number of variables and samples
n = length(Y)
F = u/m/(q/(n-m-1)) % stats value F
% Quantiles,if fw1<F<fw2 then model is not allowed
fw1 = finv(0.025,m,n-m-1)
fw2 = finv(0.975,m,n-m-1)
c = diag(inv(X'*X)) % Calculate c_jj and stats value t
t = beta./sqrt(c)/sqrt(q/(n-m-1))
% Quantiles,if t_i<tfw<t_j then t_0...t_i can be abandoned
tfw = tinv(0.975,n-m-1)
rcoplot(r,rint)
title("Plot Of Residuals")
xlabel("Products")
ylabel("Residuals")
```

Input Matlab source for Specific Star Rating And Text-Reviews

```
X1=[ones(size(dryer_review)),dryer_review];
y=dryer_rating;
[b,bint,r,rint,stats]=regress(y,X1);
z1=b(1)+b(2)*dryer_rating;
figure;
plot(dryer_review,y,'r+',dryer_rating,z1,'b');
xlabel('reformed review value'),ylabel('star rating');
title('hair dryer')
grid on;

X2=[ones(size(micro_review)),micro_review];
y=micro_rating;
[b,bint,r,rint,stats]=regress(y,X2);
z2=b(1)+b(2)*micro_review;
figure;
plot(micro_review,y,'r+',micro_review,z2,'b');
xlabel('reformed review value'),ylabel('star rating');
title('microwave')
grid on;

X3=[ones(size(pacifier_review)),pacifier_review];
y=pacifier_rating;
[b,bint,r,rint,stats]=regress(y,X3);
z3=b(1)+b(2)*pacifier_review;
figure
plot(pacifier_review,y,'r+',pacifier_review,z3,'b');
title('pacifier')
```

```
xlabel('reformed review value'),ylabel('star rating');
grid on;
```

Input Python source for Product Analysis Based On Online Evaluation System

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from fbprophet import Prophet

df1 = pd.read_csv(r'C:\Desktop\2020COMAP\reformed_hair_dryer.csv')
df2 = pd.read_csv(r'C:\Desktop\2020COMAP\reformed_microwave.csv')
df3 = pd.read_csv(r'C:\Desktop\2020COMAP\reformed_pacifier.csv')

m1 = Prophet()
m2 = Prophet()
m3 = Prophet()

m1.add_country_holidays(country_name='US')
m2.add_country_holidays(country_name='US')
m3.add_country_holidays(country_name='US')

m1.fit(df1)
m2.fit(df2)
m3.fit(df3)

future1 = m1.make_future_dataframe(periods=365)
future2 = m2.make_future_dataframe(periods=365)
future3 = m3.make_future_dataframe(periods=365)

future1.tail()
future2.tail()
future3.tail()

forecast1 = m1.predict(future1)
forecast1[['ds','yhat','yhat_lower','yhat_upper']].tail()
forecast2 = m2.predict(future2)
forecast2[['ds','yhat','yhat_lower','yhat_upper']].tail()
forecast3 = m3.predict(future3)
forecast3[['ds','yhat','yhat_lower','yhat_upper']].tail()

m1.plot(forecast1,xlabel="year",ylabel="star_rating").show()
m1.plot_components(forecast1).show()
m2.plot(forecast2,xlabel="year",ylabel="star_rating").show()
m2.plot_components(forecast2).show()
m3.plot(forecast3,xlabel="year",ylabel="star_rating").show()
m3.plot_components(forecast3).show()
plt.show()
```

Word Cloud images

Here are Word Cloud images we created, using them to analysis the demands and focus of the customers.

retractable
like
love
recommend
great
LIKE
nice
quality
HOT
hot
powerful
quickly
SMALL
like
small
love
RECOMMEND
PURCHASED
POWERFUL
RETRACTABLE
QUALITY

Figure .1: Word Cloud Image for hair dryer

RECOMMEND
GREAT
oven
EASY
works
oven
warranty
works
OVEN
POWER
great
good
NICE
SMALL
PERFECT
NICE
good
power
NICE
easy
MICROWAVE
great
warranty
PERFECT
nice
SMALL
good
nice
good

Figure .2: Word Cloud Image for microwave oven

Figure .3: Word Cloud Image for parcifier