For office use only

T1 ________________

T2 ________________

T3 ________________

T4 ________________

Team Control Number

2022813

Problem Chosen

C

For office use only

F1 ________________

F2 ________________

F3 ________________

F4 ________________

2020

MCM/ICM

Summary Sheet

Online platform product analysis prediction model（based on NLTK and BP neural network）

Abstract

In this paper, our team analyzed the data files provided by Sunshine Company so that we can finally not only give Sunshine Company helpful suggestions after the products have been online but also provide sales forecast.

For task one, our team first used NLP algorithm to analyze the natural language of text evaluation. Secondly, in order to make the data more intuitive, we made a word cloud based on high-frequency words, so that we can see the high-frequency words of the corresponding products more clearly. In addition, frequency does not fully represent the importance, we will also use TF-IDF algorithm for text feature extraction. Thus, the trend of product sales volume and average score over time can be analyzed, and we find that the evaluation has a "driving effect".

For task two, we used BP neural network prediction method to predict the product sales volume and get the corresponding sales curve. According to the market rules and figures, our team speculated that there would be a peak of daily sales, that is, the daily sales in a certain period (from the analysis, it is the middle stage of sales) composite Gaussian distribution, so we got a preliminary sales strategy. Besides that, we noticed the influence of some factors on the sales volume. The higher the initial product is evaluated, the better the final sales volume will be. Moreover, we also considered the influence of other possible factors on the sales volume of products, that is, the durability of products, and concluded that the sales volume of durable goods is lower than that of consumables. At the same time, we analyzed the relationship between rating and evaluation and obtained that the higher the rating is, the better the emotional tendency in the text evaluation will be, which once again confirms the relationship between the text evaluation and the final star rating.

For task three, we combined the above analysis results, wrote a letter to the president of sunshine company, and put forward our own suggestions for the company's sales. In the early stage of product launch, we should pay attention to the feedback of users in time to improve the product, and finally we can achieve better sales. Of course, in order to get more text evaluation and get timely feedback from users, we can also set up some incentive policies for the customers who conduct text evaluation.

Finally, based on the problems in our analysis and the results of our analysis, we objectively evaluated the advantages and disadvantages of our model. We believe that we have built a more perfect model.

Keywords: NLP，queuing theory，BP neural network, emotional tendency

Contents

I. Introduction

1.1 Background

Online shopping has become one of our mainstream shopping methods. According to data released by the United Nations Conference on Trade and Development (UNCTAD) on March 29, 2019, the global e-commerce (EC) transaction volume increased by 13% over the previous year, reaching 29.3670 trillion US dollars. It is said that e-commerce continues to expand mainly in developed countries. There are 1.3 billion online shoppers worldwide, which is equivalent to one out of every four people in the world shopping online.

It means that, in the United States, online shopping will gradually become the mainstream shopping method. Therefore, the company should also make a transition from offline to online, adapt to the sales model of online shopping, and use the unique star ratings, reviews, helpfulness rating and commentary time, etc. to come up with online sales strategies, identify potentially important design features to enhance product appeal, and make online retailing more successful.

Therefore, more reasonable use of online and background data is required to promote online retail.

1.2 Problem Restatement

Sunshine Company plans to launch and sell three new products online: microwave, baby pacifiers and hair dryers. We will look for quantitative or qualitative relationships by analyzing the sales data of similar products that they have provided in the past, summarize the measurement standards of star ratings and text evaluations, and judge the trend of product reputation and sales over time to give them information on online sales strategies Suggest and identify potentially important design features to increase product appeal.

1.3 Symbols

The symbols needed to read this article:

m: training sample

L: the number of neural network layers (excluding the input layer)

x: input sample

i: Enter the sample number / serial number

E (i): single sample error

E: error of m training samples

x: time

y: number of reviews (sales)

W: Final review volume

M: Number of consumers

a: average star rating in initial evaluation

b: average star rating for mid-term evaluation

c: average star rating in later evaluation

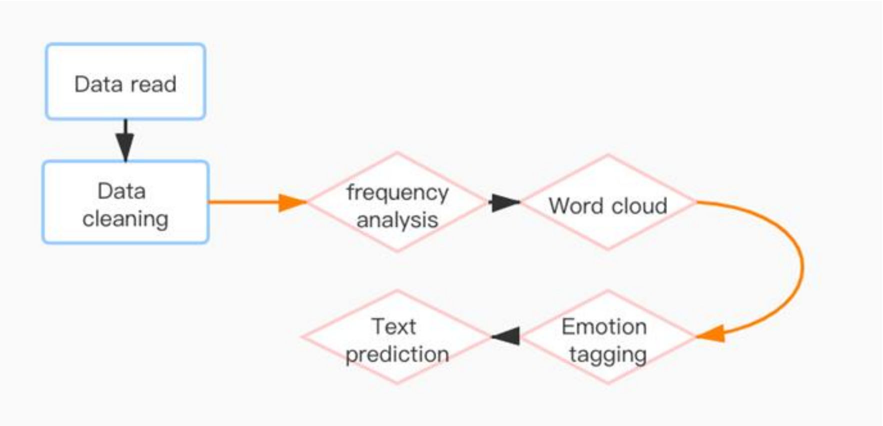
II. Models Construction

2.1 Question One

Data science is common in numerical fields, but this growing field can now also be applied to non-numeric data, such as text. At the same time, text is a common type of unstructured data that requires different mechanisms to extract insights. Text analysis or text data mining is the process of obtaining information from text using various methods. In this modeling, we will use the well-known language analysis tool called NLTK for text analysis. By processing the data of related product reviews, we can get the evaluation of product sales and consumer attention trends, so as to make sales forecasts.

The full name of NLTK is Natural Language ToolKit, a set of tools for natural language processing based on python. It was developed by Steven Bird and Edward Loper in the Department of Computer and Information Science at the University of Pennsylvania. It can implement rich natural language processing functions such as word segmentation, part-of-speech tagging, word frequency analysis, sentiment analysis and has a rich corpus. It is suitable for analysis and processing of large-scale corpora. At the same time, we will use the scikit-learn machine learning library to perform feature extraction on text.

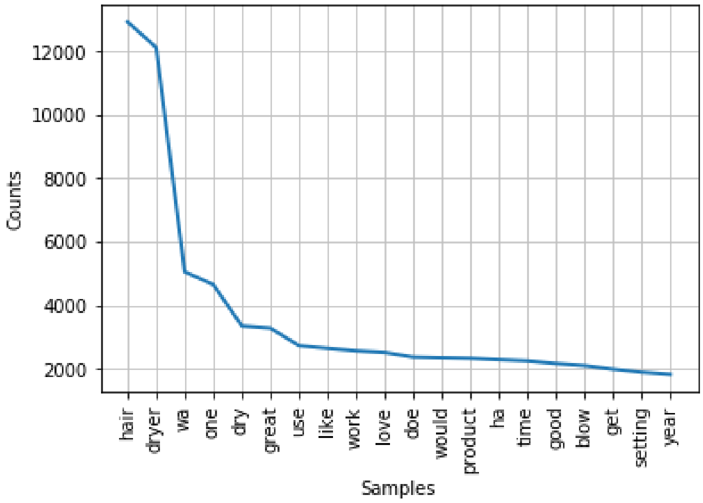
.

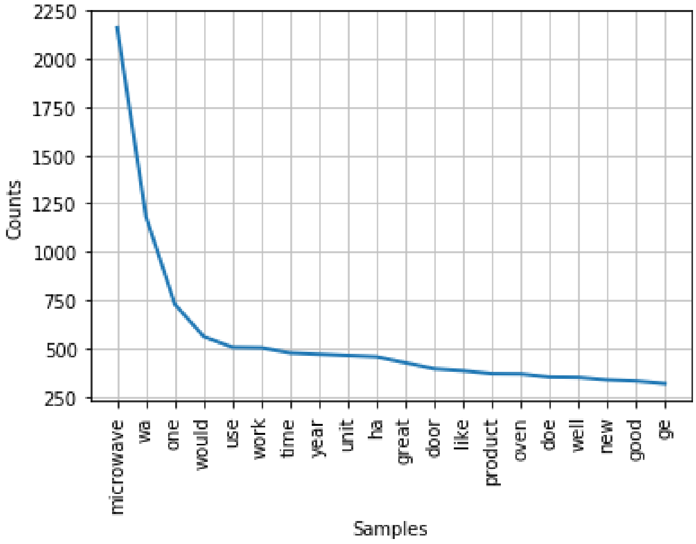


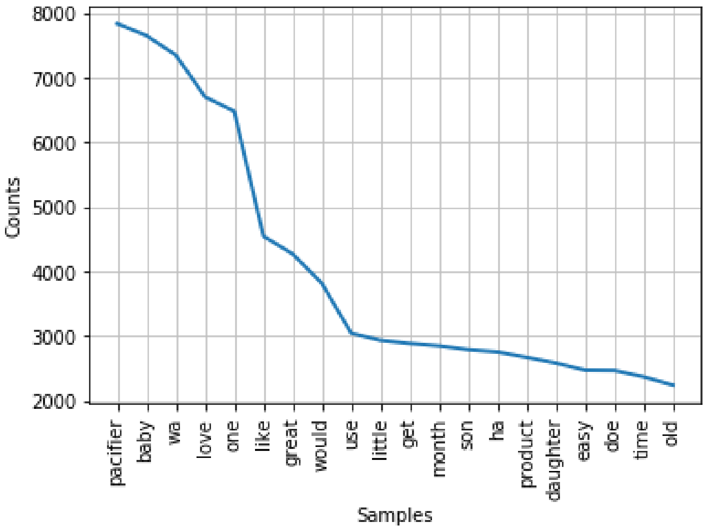
The process of NLTK text analysis

First, we export the review data of the three products and save it as a txt file. Then, we import the file into the program for text cleaning. After segmentation and word segmentation processing, we normalize the corresponding participle to morphology, remove punctuation, and remove stop words to retain the most useful data. Please see clean.py for related code.

Tokenization is usually the first step in working with a collection of text. NLTK includes a frequency distribution class called FreqDist that identifies the frequency of each marker (word or punctuation) found in the text. These tokens are stored as tuples containing the words and their occurrences in the text. We can use the most_common function to limit the output tuples to the first 20, and then output this set to get the high-frequency words and their occurrences. Finally, use matplotlib to plot the frequency distribution.







Secondly, in order to make the data more intuitive, we have created a word cloud based on high-frequency words, so that we can more clearly see the high-frequency words of the corresponding products.

hair
dryer
one
dryer
like
use
dry
hair
wa
also
great
blow

hair dryer

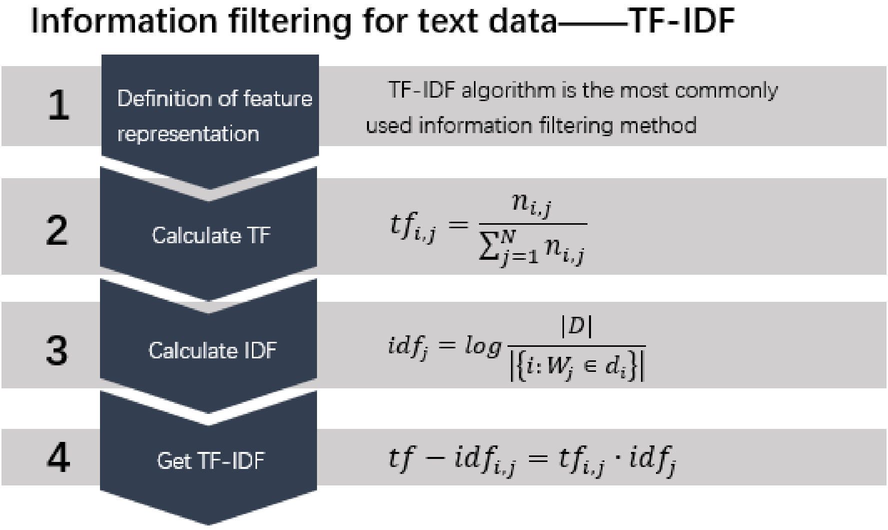
microwave
could
need
would
like
one
appliance
back
unit
wa
look
time
month
problem
doe
kitchen
model
small
product
cook
bought
samsung
oven
use
ha
fit
also
part
well
get
good

microwave

one
like
wa
baby
pacifier
ha
great
use
love

baby pacifier

In addition, frequency does not fully represent importance. We will also use TF-IDF algorithm for text feature extraction. TF-IDF is an abbreviation of Term Frequency-Inverse Document Frequency, which is also referred to as "word frequency-inverse document frequency", which includes two parts.



Information filtering for text data——TF-IDF

1 Definition of feature representation — TF-IDF algorithm is the most commonly used information filtering method

2 Calculate TF — $tf_{i,j} = \frac{n_{i,j}}{\sum_{j=1}^{N} n_{i,j}}$

3 Calculate IDF — $idf_j = log\frac{|D|}{|\{i: W_j \in d_i\}|}$

4 Get TF-IDF — $tf - idf_{i,j} = tf_{i,j} \cdot idf_j$

TF refers to word frequency, which indicates how often words appear in a piece of text. If a word appears frequently in this paragraph of text, then keywords that are more representative of this text, such as words "fried vegetables" appearing in gourmet texts. Suppose there are n words in a text $W = \{w_1, w_2, \ldots, w_n\}$, and the word x appears m times, then the frequency of the word x is

$$TF(x) = \frac{m}{n}$$

IDF refers to the inverse document frequency, which indicates the infrequent occurrence of words in all texts. If a word often appears in the corpus, such as "me", "of", "good", etc., these words are relatively important; if a word does not appear frequently in the corpus, such as "court", "attack", etc., these words are relatively more important.

Suppose N represents all the texts in the corpus, and N (x) represents the total number of texts containing the word x in the corpus, then the IDF of the word x is:

$$IDF(x) = \log \frac{N}{N(x)}$$

After getting TF and IDF, then the TF-IDF value of a word x is calculated as:

$$TF - IDF(x) = TF(x) \cdot IDF(x)$$

We use TfidfVectorizer to implement TF-IDF in scikit-learn, so as to get the corresponding text keywords.

Based on the IDF-TF, we selected 10 key words of hair dryer reviews and classified them, from which we found some problems that may exist in the products.

advertising，reputable	Consumers consider advertising and brand reputation.
caveat	Consumers care about return and exchange policies.
aging	Consideration should be given to the needs of multiple age groups.
unhealthy，allergic	The product seems to have an allergy risk.
charging，clicks，collapses，humming	The product needs to be improved in terms of charging, pressing buttons and reducing noise.

Similarly, we got ten key words of microwave oven and analyzed them.

programmable intelligent	Consumers want the microwave oven to be more intelligent.
colors yellow	Consumers want more beautiful products.
danger prevention	To ensure the safety performance of products.
flawed intermittent screeching washable	Guarantee the quality of microwave.
space saving weighs	Microwave should reduce the occupation of land and reduce the weight.

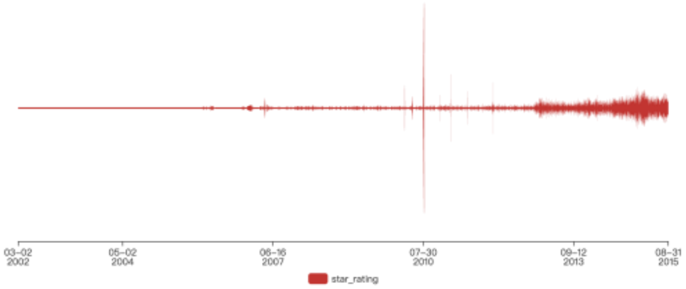
Finally, the pacifier.

activities adaptor	Suitable for infants.
alert warnings	Need to ensure product safety.
rattles	Avoid too much noise.
stainless sterilize	Ensure aseptic cleaning of products.
unscented velour bee zoo	Pay attention to the fragrance, material and style of the product

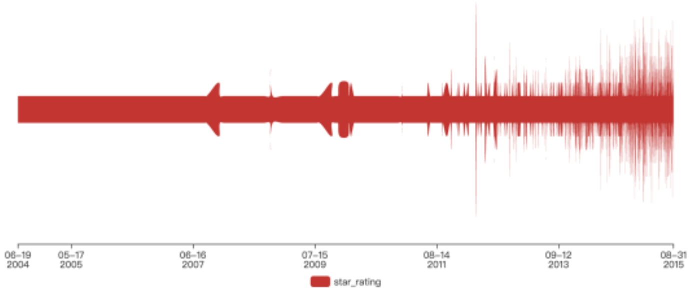
Next, we perform sentiment analysis on the text based on the clause of the first step.

Sentiment analysis or opinion mining is a process of identifying whether the author's attitude towards a certain piece of text is positive, negative, or neutral through calculation. This feedback can be useful, for example, when mining opinions about a product or service in natural language reviews. NTLK includes a simple rule-based sentiment analysis model that combines lexical features to identify sentiment intensity. We import the necessary modules (including the Vader sentiment analyzer), create a function to accept a sentence and present sentiment classification. The function first instantiates the Sentiment Intensity Analyzer and then calls the polarity_scores method with the passed sentence. The result is a set of floating-point values representing the positive or negative valence of the input text. These floating-point values are emitted for four categories (positive, neutral, negative, and compound states that represent an aggregate score). The script finally calls the passing parameters to identify emotions, and finally we get the emotion index corresponding to each comment. See sentiment.py for the code.

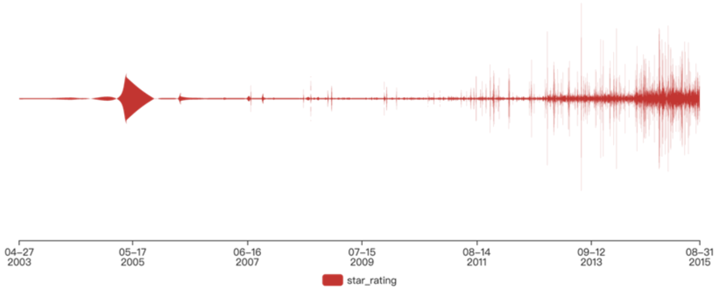
In the time dimension, we use Baidu sugar visualization product (https://cloud.baidu.com/product/sugar.html) to store the data in the cloud database, and query the corresponding data based on the sql statement to get the corresponding. The number of purchases on the date and the average rating, is to draw the trend of the number of purchases and the average rating trend of the product.



hair dryer

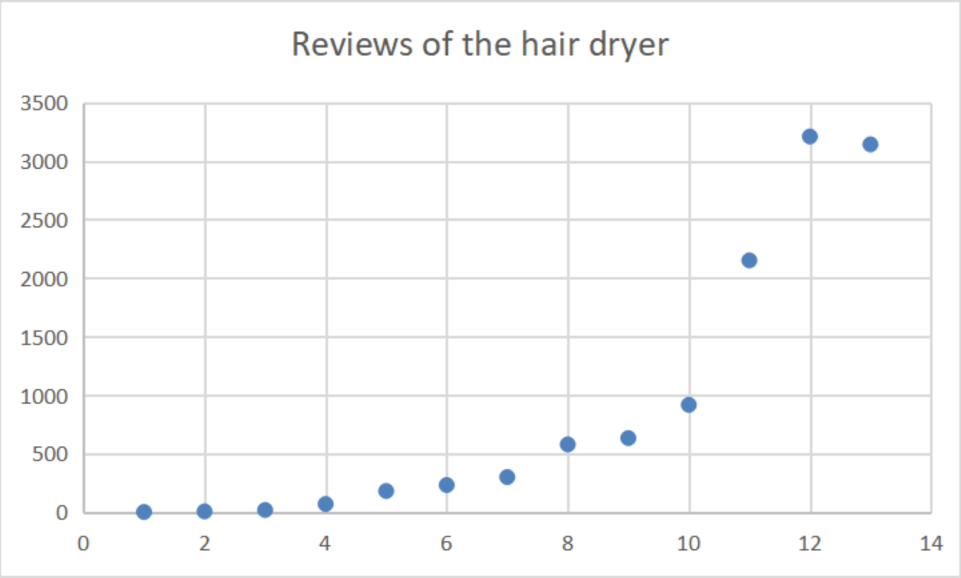


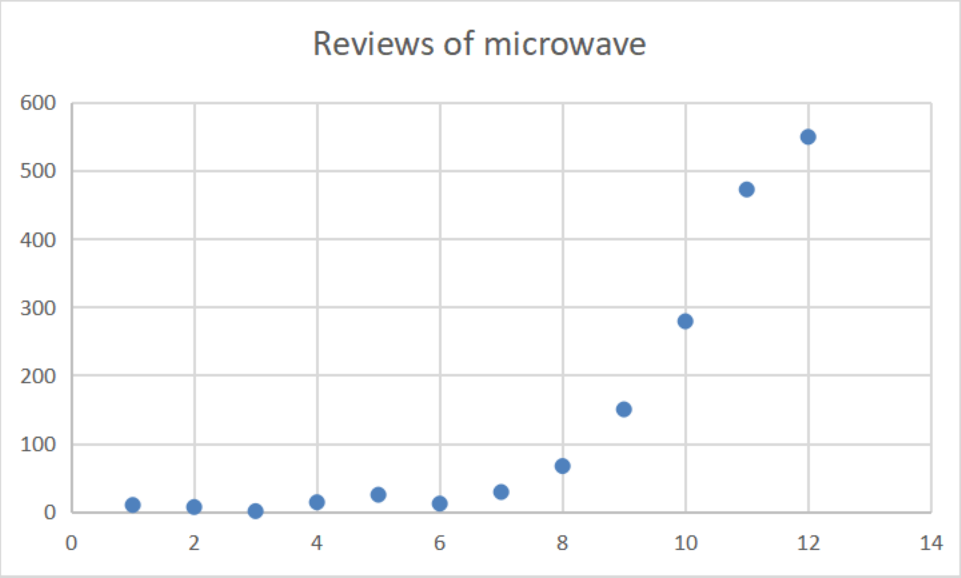
Microwave

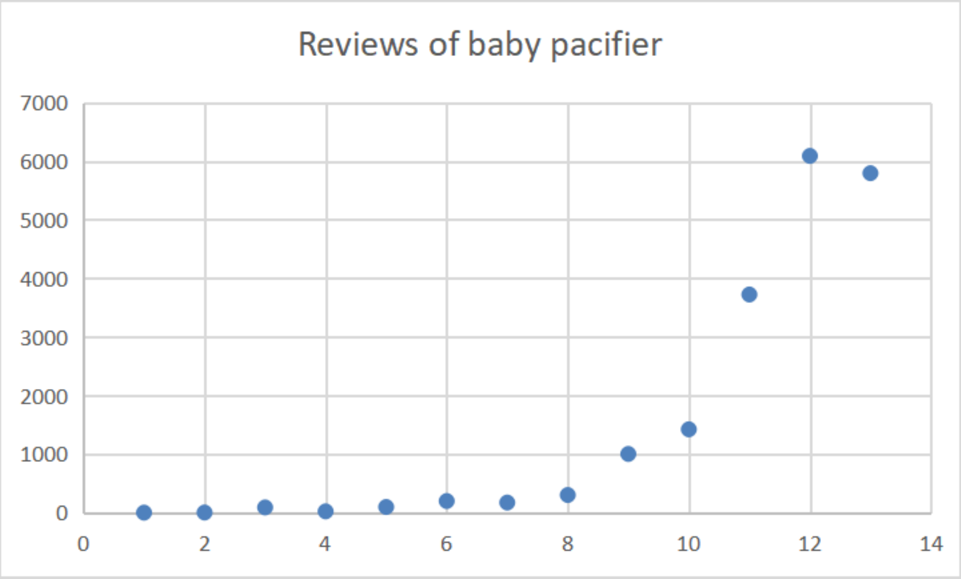


baby pacifier

After that we started to observe the following three pictures, and the three pictures show the relationship between the amount of comments and time. In order to facilitate statistics, we take one year as a unit and take the total of one year for statistics.

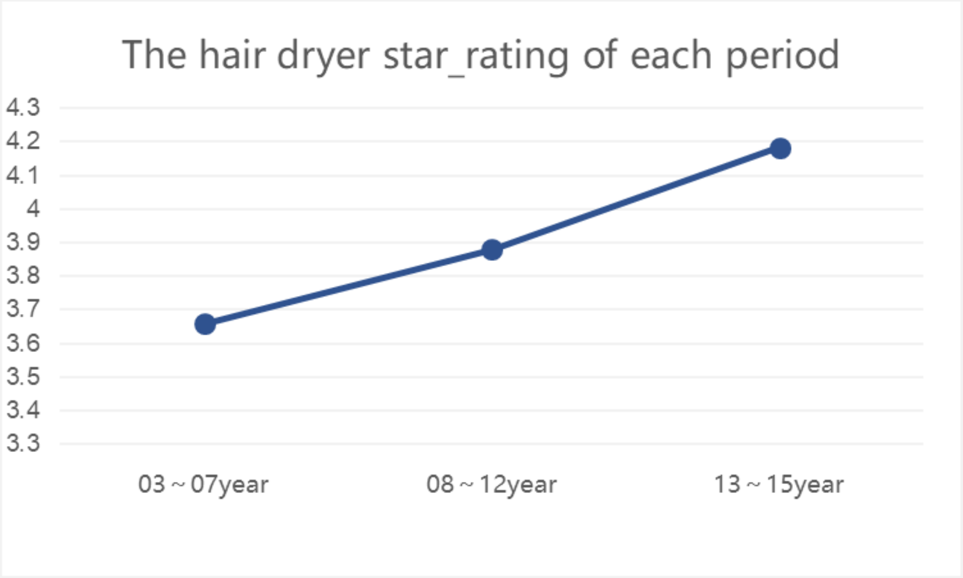


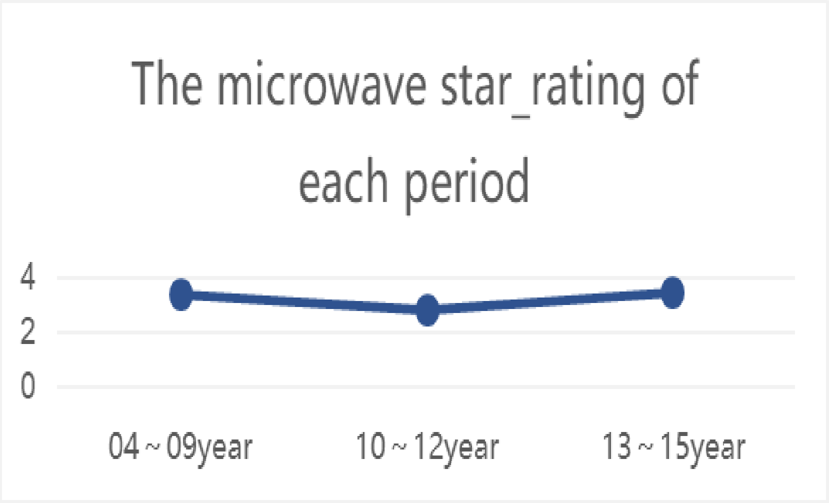


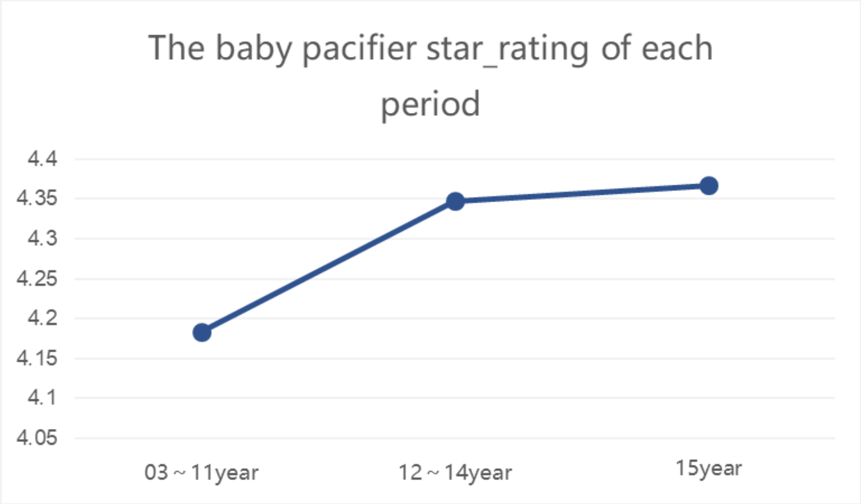


From the above pictures, we can observe that with the accumulation of time, the sales volume will gradually increase.

The following three graphs are our comparison of the scoring trends of the three products at various time periods.







According to the above diagram, we can also see that if the product is rated better in the early stage, the evaluation obtained in the middle and late stage will be better.

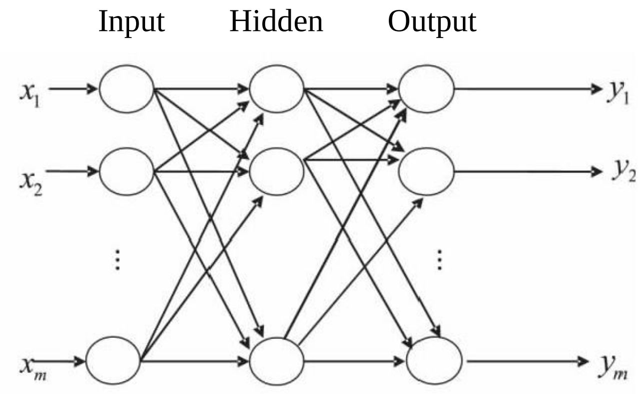
Therefore, we can speculate that reviews have a certain driving effect.

2.2 Question Two

2.2.1 *Question Two-(a)*

BP neural network has arbitrarily complex pattern classification capabilities and excellent multi-dimensional function mapping capabilities, and solves exclusive OR，XOR and some other problems that simple perceptrons cannot. Structurally, the BP network has an input layer, a hidden layer, and an output layer. In essence, the BP algorithm uses the squared error of the network as the objective function and uses the gradient descent method to calculate the minimum value of the objective function.

The basic principle of artificial neural network is an abstract mathematical model that simulates the human nervous system and processes information according to some basic characteristics of the human brain. In an artificial neural network, it consists of a group of neurons. There are connections and connection weights between any two neurons, and they are different. First enter the relevant index data into the model and train the input data through the software. During the network training process, through the process of continuous training and iteration, the weights are continuously revised and adjusted until the error is reduced to the set range. After the training, the corresponding prediction value in the training sample is finally obtained.



Suppose there are m samples, the neural network has L layers (ignoring the input layer), and the i-th input sample x (i) corresponds to the expected output H (i).

For a given m training samples, E (i) is the error of a single sample, and E is the error of m training samples. The error function is defined as:

$$E = \frac{1}{m}\sum_{i=1}^{m} E(i)$$

$$E(i) = \frac{1}{2}\sum_{j=1}^{p} (H_j^{(i)} - y_j^{(i)})^2$$

$$E = \frac{1}{2m}\sum_{i=1}^{m}\sum_{j=1}^{p} (H_j^{(i)} - y_j^{(i)})^2$$

The key of BP algorithm is to find the partial derivatives of the weight parameter of each layer. The weighted and biased partial derivatives of the first layer (2⩽l⩽ L−1) can be expressed as:

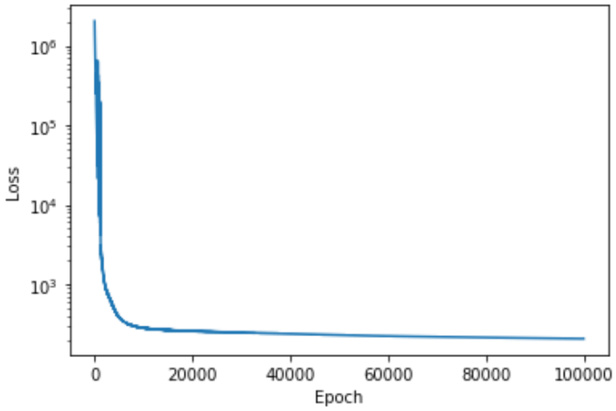
$$\frac{\partial E(i)}{\partial W_{kj}^{(l)}} = \delta_k^{(l)} h_j^{(l-1)}$$

$$\frac{\partial E(i)}{\partial b_k^{(l)}} = \delta_k^{(l)}$$

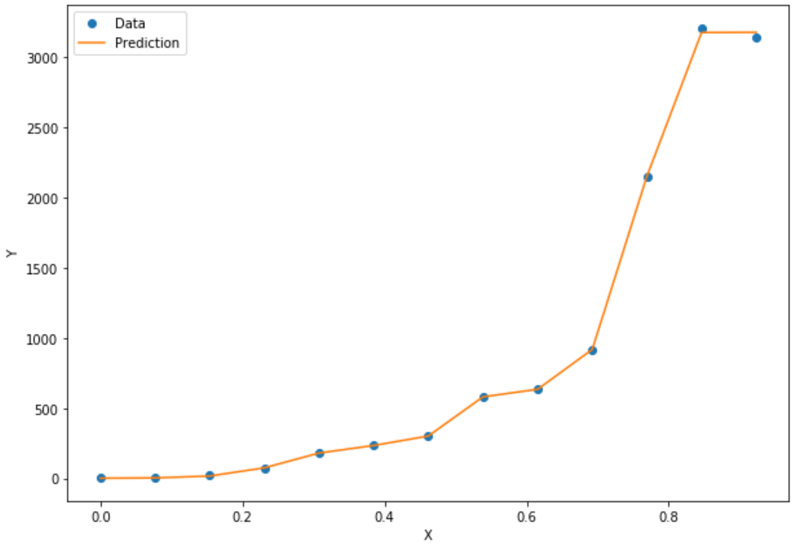
$$\delta_k^{(l)} = \sum_{r=1}^{S_{l+1}} W_{rk}^{(l+1)} \delta_r^{(l+1)} f(x)'|_{x=z^{(l)}}$$

In the actual operation process, we use the matrix to perform the operation to obtain the parameter values of the neural network.

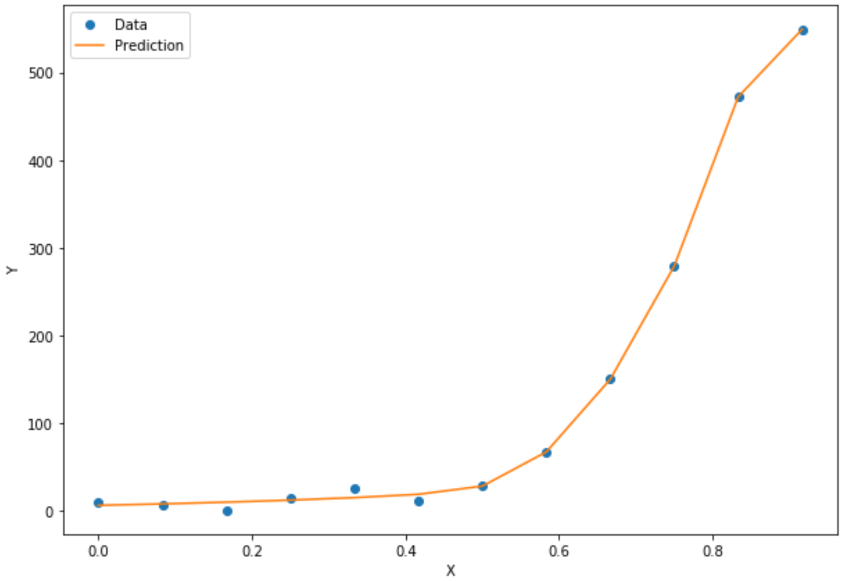
We consider the number of evaluations of hair dryers per year as sales, and then use neural networks to fit and get forecasts for future sales. Here we will use python to specifically build our neural network. First, we import dependent libraries such as numpy, pandas, import data, and initialize the weights and biases of all neural networks. Here we build a three-layer neural network, learning 100,000 times, setting the learning rate to 0.001, and performing gradient back propagation on the loss function. Get the corresponding loss curve.

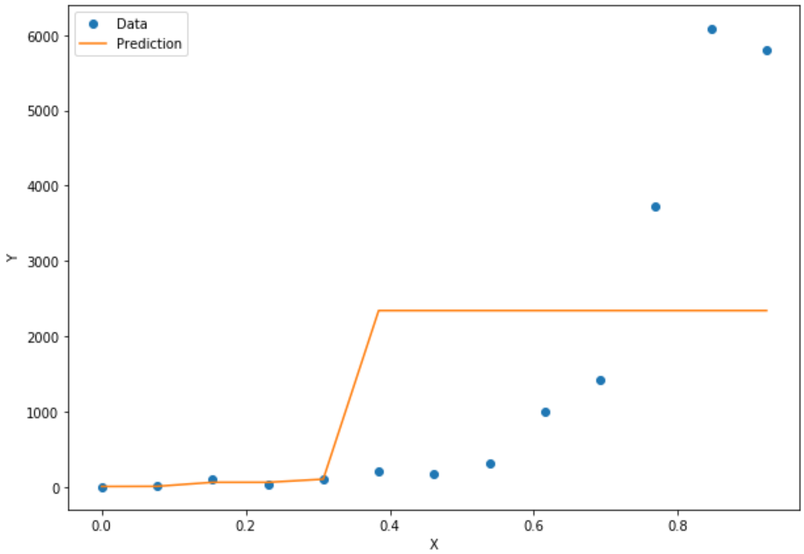


Finally, the corresponding fitted curve is obtained.



In the same way, we get the sales curve of microwave ovens and pacifiers.





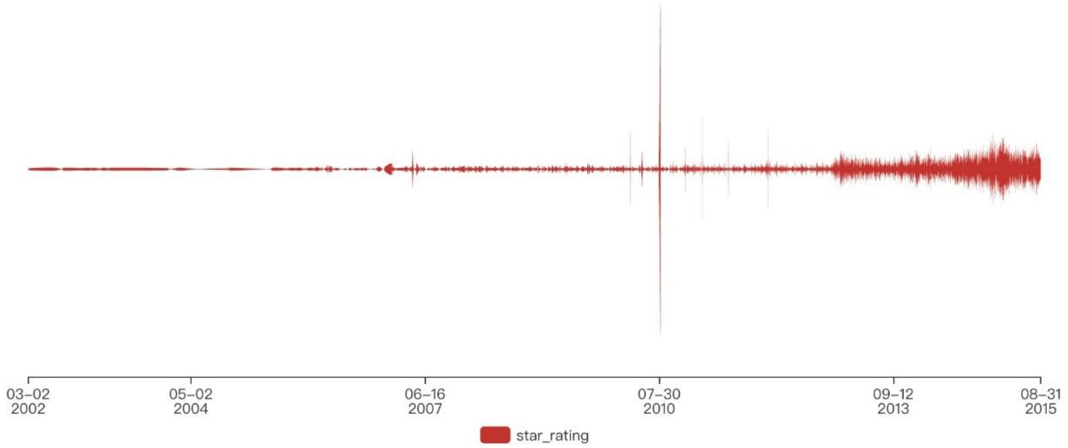
Due to some mechanism, the forecast curve of the sales volume of pacifiers has an over-fitting problem in the later stage.

2.2.2 *Question Two-(b)*

Analysis of the relationship between sales volume and time by date:
According to the general market rules and the graphs obtained through data analysis, we can infer that the number of reviews and sales will reach the peak in the middle of product sales, that is, from 7 to 9 years after the product goes online.

According to the image, we may wish to transform this distribution into a standard normal distribution, let y be the number of reviews (sales) and x be the time. Each time to reach the peak is assumed to be the 0 coordinate on the horizontal axis.

Let's take the hair dryer as an example. The following figure shows the relationship between the review volume (sales volume) and the date of the hair dryer:



At the time of 2010/07/30, we can see that the number of reviews (sales) of microwave ovens has peaked, and x = 0 is the date.

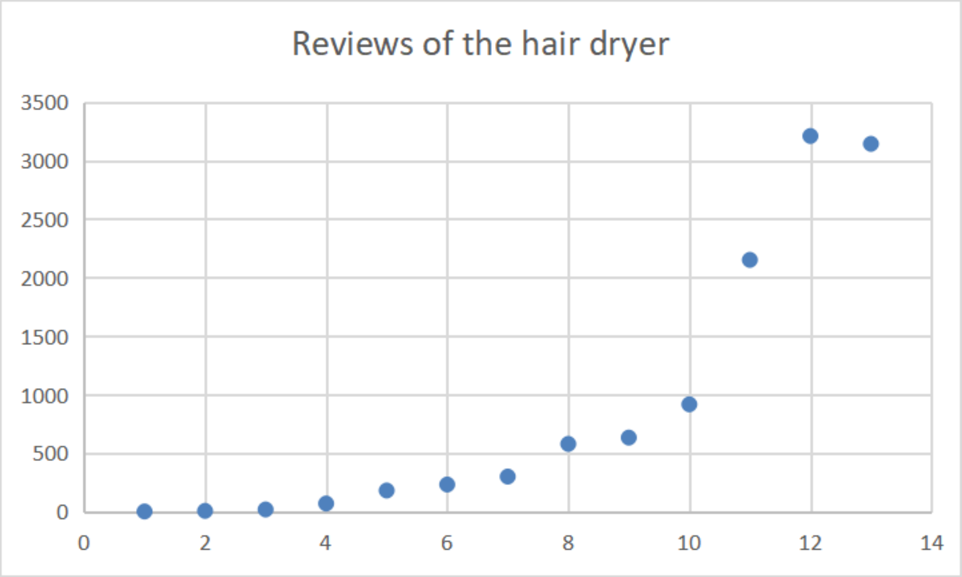
The time we currently have is from 2002/03/02 to 2015/08/31. We set the distance between two adjacent dates on the number axis to be 0.01 units (2010/07/31 is 0.01), and provide reviews The peak value that appears is one digit on the normal distribution chart, then the proportionality coefficient K between the peak value of sales predicted by the normal distribution chart and the peak value displayed on the chart is

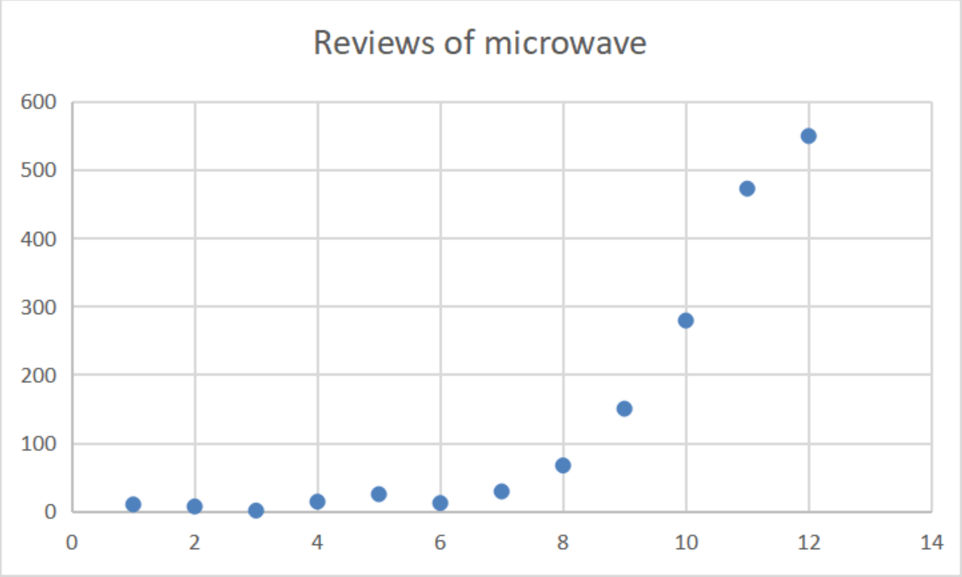
$$K = \frac{Review\ volume/Sales\ volume}{1}$$

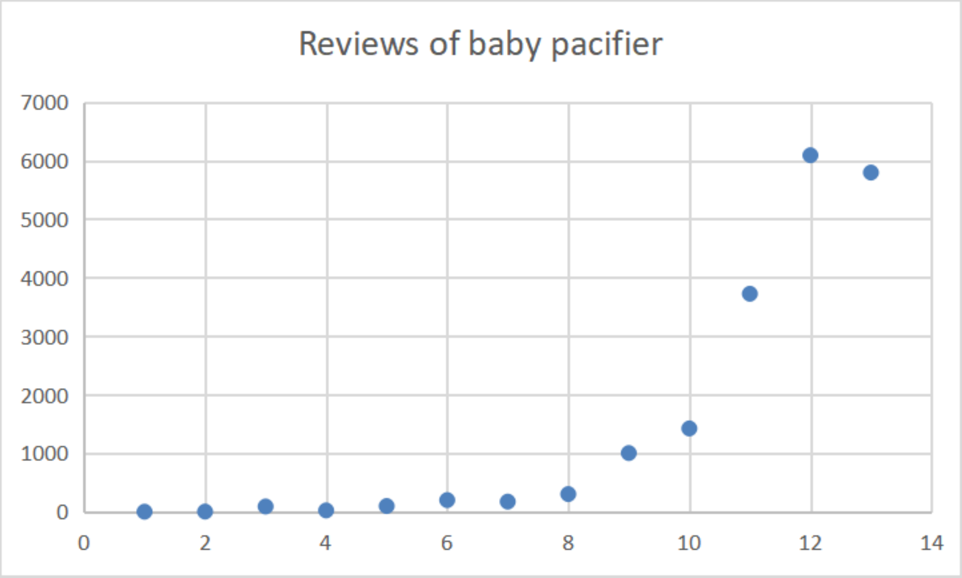
Finally, using the formula, $y = f(x) = \frac{1}{\sqrt{2\pi}\sigma}\exp(-\frac{(x-\mu)^2}{2\sigma^2})$（$\mu$ is 0，σ is1）, the final sales volume W, according to the formula W = Ky, can predict the daily sales volume.

It is not difficult for us to know that after the product is launched, the sales volume will reach a peak on a certain date or period of time. If you want to achieve better sales, then in the early stage of product launch, we must timely Feedback to improve the product, in order to achieve a higher peak value, a hair dryer this product as an example, we are from the product launch (2002/03/02) to 2010/7/30 in a short period of time, we must timely improve based on user feedback Products to meet user needs.

2.2.3 *Question Two-(c)*







The above three graphs show the trend of the review volume in time. At this time, we take one year as a unit, and the review volume reflects the sales volume. We can observe that among these three, the number of reviews W1 (hair dryer), W2 (microwave), W3 (baby pacifier),

$$W_2<W_1<W_3$$

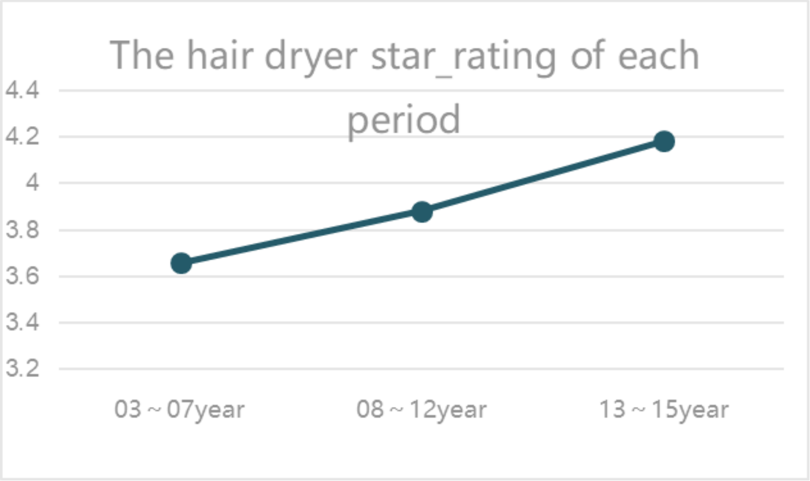
In terms of the durability of the three, microwave oven> hair dryer> baby pacifier, so we can infer that the sales volume of the product is inversely proportional to the durability of the product itself, that is, the number of consumers who will make a second purchase of this product M1 (Hair dryer), M2 (microwave), M3 (baby pacifier),、

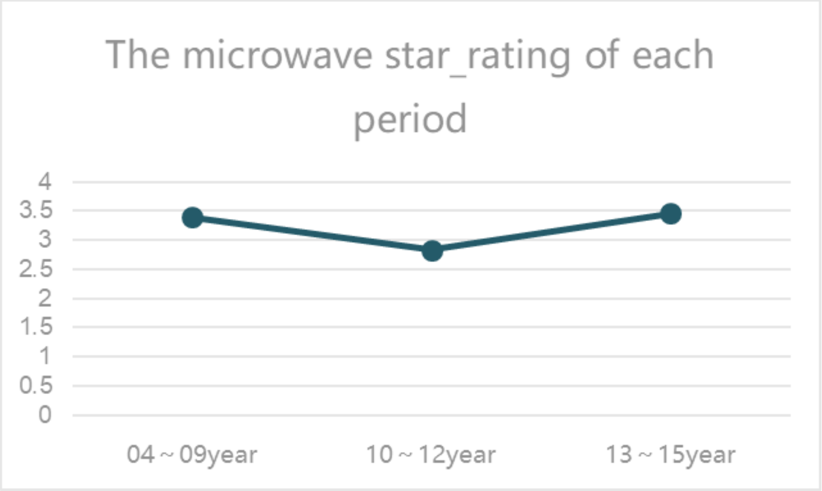
$$M_2 <M_1<M_3$$

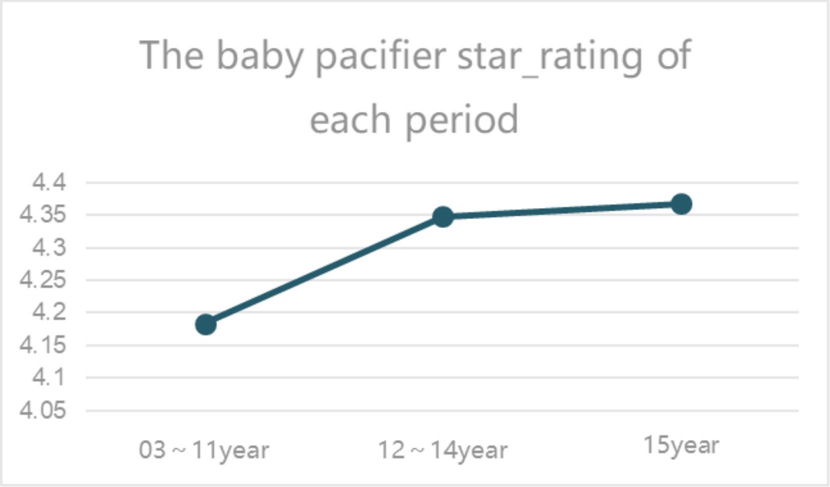
For products with high durability, more attention should be paid to the quality of the product when it is initially launched, because consumers who buy it rarely make a second purchase, which means that the evaluation will basically not change, so once the initial evaluation goes down, Consumers in the mid-to-late period may rarely choose this product, which will greatly affect sales.

2.2.4 *Question Two-(d)*

Since we are currently considering the "driving effect", then we have to consider time, so we start with the earliest reviews and see if the average of the final total reviews of all products is positively related to the earliest reviews.







Let a1, a2, and a3 represent the star rating of the consumer (hair dryer, microwave oven, and pacifier), b1, b2, and b3 (hair dryer, microwave oven, pacifier) respectively, and c1, c2, c3 (hair dryer) , Microwave oven, pacifier) respectively represent the stars of the later period.

Comparing the sizes of a1 (3.658803828), a2 (3.39952381), and a3

(4.183545879), we get

a2<a1<a3

Then compare the sizes of b1 (3.880690877), b2 (2.829263158), and b3 (4.347573386), we know that

b2<b1<b3

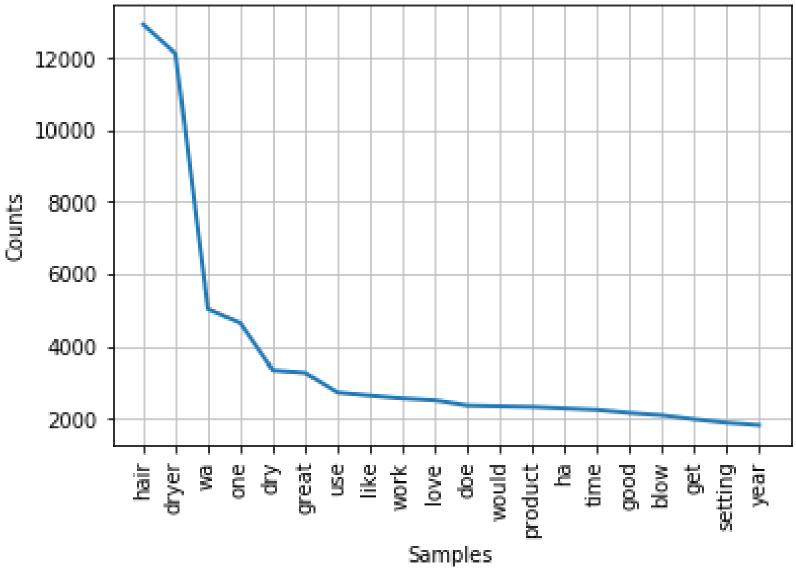
Finally, comparing c1 (4.183353973), c2 (3.4656049), and c3 (4.367325103), we get

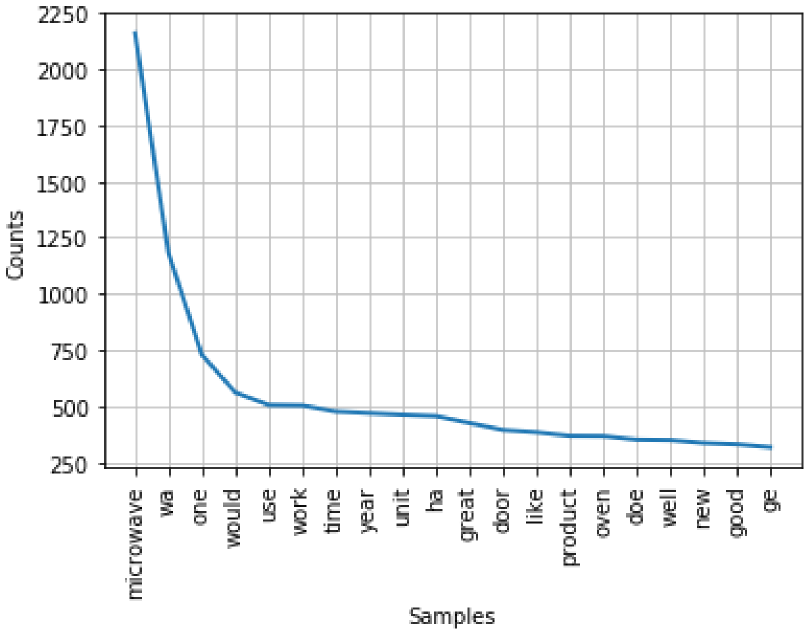
c2<c1<c3

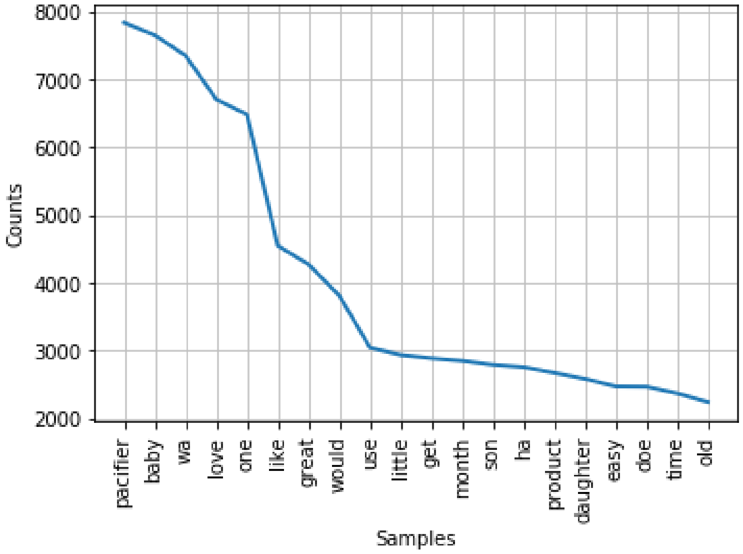
It is not difficult to speculate that the higher the evaluation obtained in the early stage, the better the corresponding evaluation in the middle and late stages, so we can conclude that the customer sees the existing reviews, and subsequent evaluations will be subject to The impact of existing evaluations has a positive correlation, that is, if the initial evaluation is better, it will drive the evaluation of the middle and late stages so that the evaluation tendency of the middle and late stages will still be better. In the initial stage of the product, we can still continue to improve the product, in order to get the user's praise, so that the user's reviews will be driven by the initial reviews, and in the middle and late stages of product sales, we can build an excellent product image.

2.2.5 *Question Two-(e)*

Analysis of high-frequency vocabulary







First of all, we can see that for each product, the most frequently appearing word in the review is the "name" of this product, followed by the "function" related words of this product (for example: in the evaluation of the product "dryer", this "dry" Words appear more frequently), and some quantifiers (for example, one). After that, let's focus on the emotional tendency of the product text evaluation. From our previous analysis, we know that the three star ratings are generally

microwave <hair dryer <baby pacifier

We observe from the above three pictures that some words with obvious positive emotional tendencies (such as great) still appear in the text evaluation with the same frequency microwave <hair dryer <baby pacifier

the same comparison result.

Therefore, we can infer that the higher the star rating product, the higher the frequency of keywords with good emotional tendency in the text evaluation, the same reason, the text with good emotional tendency in the evaluation The higher the frequency of keywords, the better the final product star rating will be.

III. Question Three

3.1 Model A

Dear marketing director of sunshine company:

After analyzing the sales of these three products of other companies, our team has come to some expectations for the future sales of these three products of your company. I'd like to give you some suggestions, which I hope will be helpful to your company's product sales.

First of all, we get the star rating and text rating from the product. After analysis, we can see that the evaluation of the text is closely related to the star level of the final product, and the evaluation of the text will have a significant impact on the star level. Star rating is the most visible and intuitive evaluation for consumers. The level of star rating plays an extremely important role in whether consumers will finally buy the product. At the initial stage of product launch, we suggest that your company can offer some corresponding preferential policies to users who have given text evaluation, so as to win users' favor and play a positive role in the final product star rating evaluation. For example, you can give some corresponding discounts to users who have made text evaluation.

Secondly, we get from the analysis that the more keywords with good emotional tendency appear, the higher the final star rating of the product will be, so we can set some text options to provide users in the early stage, just to provide the key of the evaluation (key words can refer to the thesaurus established by our team). This gives the option of text evaluation, we can not only facilitate users, get more feedback about the product, but also more quickly and intuitively analyze the user experience for the product, so that we can improve the product.

After that, we also found that if we look at the daily sales volume, the daily sales volume will reach the peak after the product goes online for a period of time (our statistical chart reflects the sales volume based on the quantity of evaluation). Before reaching the peak, your company has the time to improve the product. The higher the peak, the better the final sales volume, so at the peak In the time before value, we can establish a more rapid response mechanism, timely improve the product according to the feedback of consumers, in order to achieve the best sales volume.

Finally, we found that the durability of the product itself has a great relationship with the final sales volume. The sales volume of durable goods is far less than that of consumables. Therefore, for durable goods, before the online sales, we need to have a more strict control on the quality, and at the beginning, we need to pay attention to making users more satisfied, otherwise it will have a greater negative impact on the sales in the middle and later stages.

Sincerely yours
Team 2022813

IV. Strengths and Weaknesses

4.1 Strengths

1. All of the algorithms we used are more accurate and scientific.
2. We not only consider the information reflected in the data, but also pay attention to the impact of the actual market rules on the sales situation.
3. For the data analysis is also more sufficient, relying on a variety of tools for analysis.

4.2 Weaknesses

1. For example, in some places, the average value is used to represent the overall level of this period, which may lead to more errors in the final prediction.
2. In some places, the analysis does not give a clear relationship, but only gives a correlation, which is not accurate enough.

V. References

[1] Frank R. Giordano, William P. Fox and Steven B. Horton. A First Course in Mathematical Modeling (Fifth Edition) , 2015.

[2] Po-Han Chen and Feng Feng. Fire Safety Journal. 2009.

[3] Yang Bingrong. Passenger car market prediction model based on multiple linear regression and BP neural network [D]. Huazhong University of Science and Technology, 2017.

VI. Appendix

BP neural network

```
import numpy as np

import pandas as pd

import matplotlib.pyplot as plt

import torch

#from torch.autograd import Variable

import torch.optim as optim



counts = [num]

x = np.arange(len(counts))

y = np.array(counts)

plt.figure(figsize = (10, 7))

plt.plot(x, y, 'o-')

plt.xlabel('X')
plt.ylabel('Y')
plt.show()
sz=10
x = torch.tensor(np.arange(len(counts), dtype = float) / len(counts), requires_grad = True)
y = torch.tensor(np.array(counts, dtype = float), requires_grad = True)
weights = torch.randn((1, sz), dtype = torch.double, requires_grad = True)
biases = torch.randn(sz, dtype = torch.double, requires_grad = True)
weights2 = torch.randn((sz, 1), dtype = torch.double, requires_grad = True)
learning_rate = 0.001
losses = []
x = x.view(13, -1)
y = y.view(13, -1)

for i in range(100000):
    hidden = x * weights + biases
    hidden = torch.sigmoid(hidden)
    predictions = hidden.mm(weights2)
    loss = torch.mean((predictions - y) ** 2)
    losses.append(loss.data.numpy())
    if i % 10000 == 0:
    print('loss:', loss)
    loss.backward()
    weights.data.add_(- learning_rate * weights.grad.data)
    biases.data.add_(- learning_rate * biases.grad.data)
    weights2.data.add_(- learning_rate * weights2.grad.data)
    weights.grad.data.zero_()
    biases.grad.data.zero_()
    weights2.grad.data.zero_()


plt.semilogy(losses)
plt.xlabel('Epoch')
plt.ylabel('Loss')
plt.show()


x_data = x.data.numpy()
plt.figure(figsize = (10, 7))
xplot, = plt.plot(x_data, y.data.numpy(), 'o')
yplot, = plt.plot(x_data, predictions.data.numpy())
plt.xlabel('X')
plt.ylabel('Y')
plt.legend([xplot, yplot],['Data', 'Prediction'])
plt.show()
```