Problem Chosen	2021 MCM/ICM Summary Sheet	Team Control Number
C		2108685

A Wealth of Data

summary

In this article, relying on important tools for big data analysis and processing, we analyze and evaluate three products.

For requirement 1, we first clean up the data to remove useless data. For example, we delete samples that are not related to a given product from the data. Then, we use sentiment analysis algorithms to quantify the reviews to extract the sentiment tendency of the reviews as sentiment value. Then, we perform word frequency analysis on the comments to find keywords that reflect people's attention. We found that people were worried about the temperature of the hair dryer, the size of the microwave oven and the size of the bread. In addition, we draw a bar graph of star ratings for descriptive statistics. We found that microwave ovens have received widespread praise, but microwave ovens have relatively low ratings.

For requirement 2(a) and (c), we have established an evaluation function including numeric variables and quantified text variables to evaluate products. According to this scoring standard, we selected the sub-products with the highest comprehensive score for each product.

For Requirement 3, we provide some online sales strategies related to major markets and major directions. The letter also mentions potential design features.

contents

1. introductions

In the Internet age, online shopping has become the main way people shop. Data from online shopping platforms has become an important basis for online stores to make business decisions. Sunshine plans to launch and sell three new products on the online market: microwave ovens, baby napkins and hair dryers. Sunshine has some time-based customer online shopping data. Our team will analyze the data and solve the following problems:

i: Analyze three product data sets, and describe meaningful variables and the relationship between star ratings, reviews and useful votes.

ii: Once the three products are sold on the online market, the best data metric for Sunshine will be determined based on ratings and reviews.

iii: Determine the combination of text-based metrics and rating-based metrics that best indicate a potential success or failure product.

2. Descriptive statistics and data preprocessing

2.1 Data preprocessing

First, we need to preprocess the relevant data of the three products provided by Sunshine, including removing meaningless data and quantifying text data.

2.2 Descriptive Statistics

First, we perform descriptive statistics and data visualization on the annotation data. Data processing and cleaning are important steps before starting text mining. In this step, we remove punctuation marks, stop words, etc. to make the annotations as uniform as possible. When you are done, you can check the most frequently used words in the data. Therefore, let us define a function here that displays the n most common words in the data on a bar graph. Next, to further eliminate the interference in the text, you can use the word form in the spacy library to simplify. It can restore words to their original form and reduce the repetition of words.

The following shows the user evaluation word cloud of each product. Frequently appearing words can reflect the user's attention to certain properties of the product or the product has certain functions, and we can find some points worthy of the company's attention.

Figure 1: Word cloud of hair dryer comments

According to Figure 1, we can find that there are some keywords in the reviews, such as "good", "excellent", "strong" and "hot", etc. It can be seen that most users approve the product of hair dryer. Most hair dryers are powerful, but some hair dryers may be too hot and cause user injury. The company should conduct further research on the temperature control of the hair dryer. At the same time, we can also see some keywords such as "money", "quality", etc. It can be seen that users are also more concerned about the price and the product itself. The company can conduct further market research and research and development to reduce the price at the same time. Improve product quality.

Figure 2: Word cloud of microwave oven reviews

According to Figure 2, users may also praise the microwave oven. However, there are a lot of words about the size and color of the microwave oven. Although it is impossible to judge whether these characteristics are positive or negative by word frequency, we can find that customers are very concerned about the appearance and quality of microwave ovens.

Figure 3: Word cloud of baby pacifier comments

According to Figure 3, it can be seen that the frequency of positive words in the user's annotation keywords is very high. In addition, according to the words "cute", "pink" and other words, it can also be analyzed that the user still pays more attention to the product's cuteness and colour.

3. Establish product evaluation function

We will give the data sets "microwave.tsv", "hair_dryer.tsv", "pacifier.tsv", All were cleaned, unnecessary data was removed, and the corresponding "microwave_clean.xlsx", "microwave_clean.xlsx", and "pacifier_clean.xlsx" were obtained. Then the corresponding data was visualized.

From the perspective of the word cloud generated, users of these three products are generally quite satisfied, and they basically give five-star praise and think the products are more practical.

3.1 Analysis of three products

hair dryer:

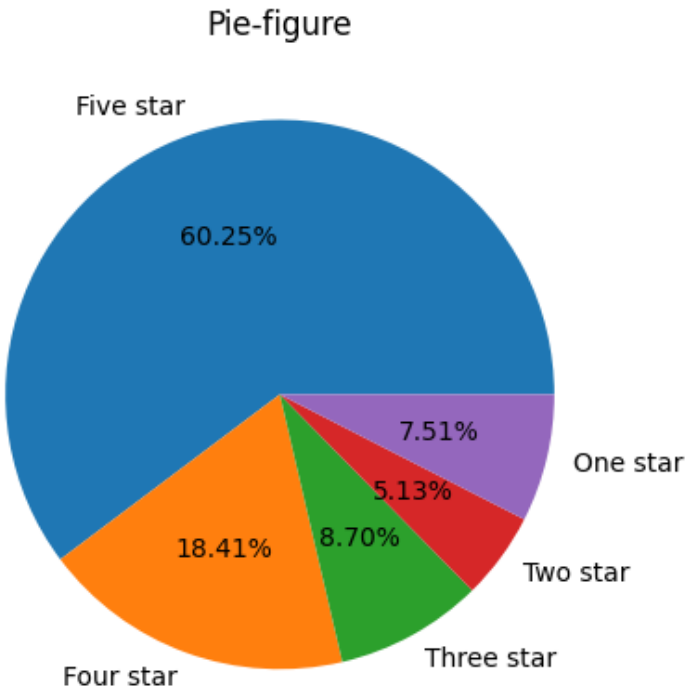

Figure 4: Pie chart of star rating for hair dryers

According to the generated word cloud and star rating pie chart, users still have a high evaluation of this product. They think this product is more practical and are willing to give five-star praise. At the same time, it can be seen from the word cloud that users are very concerned about power , Price, shape, easy to carry, etc. So for hair dryer products, the user reviews that Sunshine Company should focus on are price, power, appearance, drying effect, whether it is easy to carry, etc.

microwave

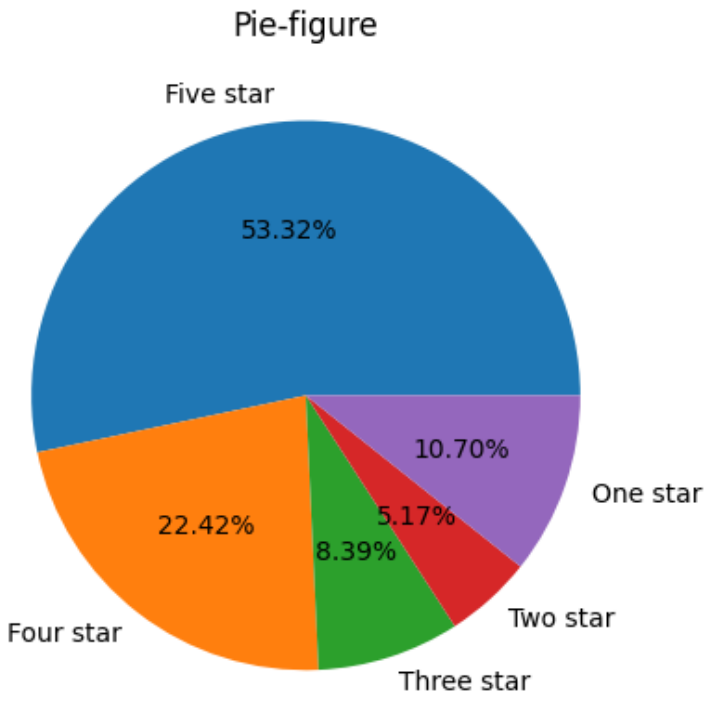

Figure 5: Pie chart of microwave oven star rating

Similarly, most users also expressed their love for microwave oven products. In the word cloud and rating pie chart generated, it was found that most of them were positive feedback from users. The high frequency of the product in the word cloud is simple operation, practicality, price, function, etc. Customers generally worry about the advantages of microwave ovens, namely simplicity, small space, cleanliness and low price. These are the comments that Sunshine should focus on.

pacifier:

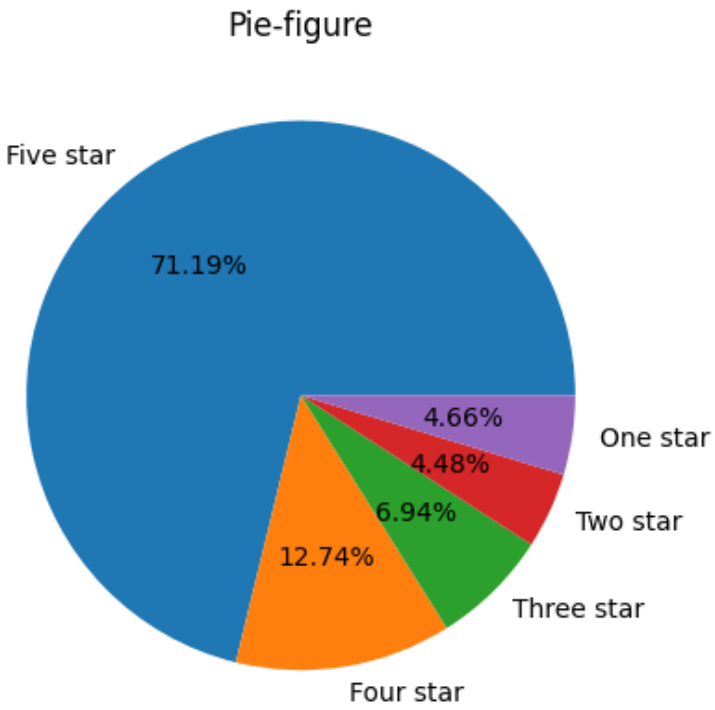

Figure 6: Star rating pie chart of baby pacifiers

For this product, users are mainly concerned about the baby experience. Most users think this product is better. However, because this product is for babies, users have higher expectations of the product, so some users expressed Disappointed. The most frequent occurrence of word cloud is whether the product is soft, cute, size, and color. So Sunshine can focus on the comments with these words in order to better receive user feedback

3.2 Evaluation results

In summary, we combine word clouds and line graphs, a combination of text-based metrics and rating-based metrics. We predict that hair dryers and baby pacifiers will maintain the current good development trend in the future, but if microwave ovens do not keep up with user expectations, they will Will become a potential failure product.

4. Letter to Sunshine Company

Dear Sunshine Company:

After filtering, categorizing and quantifying the data, our team established a scoring model to review. A higher review score indicates a higher degree of acceptance of the product, and higher product quality and service. Our team analyzed the microwave oven, hair dryer and baby pacifier market based on this model, and provided some online sales strategies and some potential design features that may enhance product demand.

Comparing microwave ovens, hair dryers and baby pacifiers, the baby pacifier market usually provides higher quality products and better services. Therefore, in the subsequent fierce competition, the baby pacifier market has won a higher reputation, which has a higher reputation for the quality of baby pacifiers , Functions and services put forward higher requirements. Products provided by the merchant. If Sunshine enters the paper market, it will face more intense competition, but the demand will be greater. Relatively speaking, in the microwave oven market,

there are more and more complaints about product quality, function and service. This is both a challenge and an opportunity for Sunshine to enter this market.
Through analysis, we can obtain the functions of different products that consumers care about. The design of the product should meet the needs of consumers as much as possible. Therefore, here are some product functional designs that companies should pay attention to:

People usually care about heat setting in hair dryers, and some hair dryers have the advantages of being fast, light, and low in sound. Among them, heat setting is the most attractive. However, consumers still complain about some disadvantages, such as sparks, unpleasant smell, too hot, too heavy, too loud, etc. It can be seen that customers are very concerned about the temperature and weight of the hair dryer.

Customers generally worry about the advantages of microwave ovens, namely simplicity, small space, cleanliness and low price, but the whirlpool microwave oven has received a lot of praise, and users hate constant maintenance and worry about its quality.

Team 6039

Appendix

Appendix A Data cleansing code

```
import pandas as pd
import xlsxwriter
from nltk.sentiment.vader import SentimentIntensityAnalyzer



#read the data
hairdryer=pd.read_csv('../dataset/hair_dryer.tsv',sep='\t')
microwave=pd.read_csv('../dataset/microwave.tsv',sep='\t')
pacifier=pd.read_csv('../dataset/pacifier.tsv',sep='\t')

#remove unrelated data
microwave=microwave[microwave.product_title.str.contains('microwave')]
hairdryer=hairdryer[hairdryer.product_title.str.contains('dryer')]
pacifier=pacifier[pacifier.product_title.str.contains('pacifier')]

#function: convert review to sentiment value
sid = SentimentIntensityAnalyzer()
def sentiquantify(sen):
    score = sid.polarity_scores(sen)
    score=score['compound']
    return score

#dataprocess
def dataprocess(df):
    df['review_headline']=df['review_headline'].apply(str)
    df['review_body']=df['review_body'].apply(str)
    df['review']=df['review_headline']+'. '+df['review_body']
    df['total_sentiscore']=df['review'].apply(sentiquantify)
    df['review']=df['review'].str.replace("[^a-zA-Z#]", " ")

dataprocess(hairdryer)
dataprocess(microwave)
dataprocess(pacifier)

```

```
hairdryer['sentiment']=hairdryer['total_sentiscore'].apply(lambda x: 'positive' if x>=0 else 'negtive')
microwave['sentiment']=microwave['total_sentiscore'].apply(lambda x: 'positive' if x>=0 else 'negtive')
pacifier['sentiment']=pacifier['total_sentiscore'].apply(lambda x: 'positive' if x>=0 else 'negtive')


pacifier.to_excel('../dataset/pacifier_clean.xlsx', engine='xlsxwriter')
hairdryer.to_excel('../dataset/hairdryer_clean.xlsx', engine='xlsxwriter')
microwave.to_excel('../dataset/microwave_clean.xlsx', engine='xlsxwriter')
```

Appendix B Visualization Code

```
from datetime import datetime

import matplotlib.pyplot as plt
import pandas as pd
from wordcloud import WordCloud, STOPWORDS


def get_data(list):
    simple_data = []
    for single in list:
        product_name = single[6]
        star = single[8]
        review = str(single[13])
        date = datetime.strptime(single[15], '%m/%d/%Y')
        detail = {'name': product_name, 'star': star, 'date': date, 'review': review}
        if (single[12] == 'Y'):  # 过滤未交易的订单
            simple_data.append(detail)
    return simple_data


hairdryer_data = get_data(pd.read_excel("../dataset/hairdryer_clean.xlsx", sheet_name="Sheet1").values)
pacifier_data = get_data(pd.read_excel("../dataset/pacifier_clean.xlsx", sheet_name="Sheet1").values)
microwave_data = get_data(pd.read_excel("../dataset/microwave_clean.xlsx", sheet_name="Sheet1").values)


# 词云
def ciyun(data, path, name):
    stopwords = set(STOPWORDS)
    useless_words = ['hair', 'dryer']
    for i in useless_words:
        stopwords.add(i)
    string = ''
    for i in data:
        string += i['review'] + ' '
    wordcloud = WordCloud(background_color="white", stopwords=stopwords, width=1000, height=860, margin=2).generate(string)
    plt.imshow(wordcloud)
    plt.axis("off")
    plt.savefig(path + name)
```

```
    plt.show()


# 生成对应的词云
path = '../images/'
ciyun(hairdryer_data, path, 'hairdryer_data')
ciyun(pacifier_data, path, 'pacifier_data')
ciyun(microwave_data, path, 'microwave_data')


def calc_star(data, path, name):
    labels = ['Five star', 'Four star', 'Three star', 'Two star', 'One star']
    res = [0, 0, 0, 0, 0]
    for i in data:
        star = i['star']
        if (star == 5):
            res[0] += 1
        elif (star == 4):
            res[1] += 1
        elif (star == 3):
            res[2] += 1
        elif (star == 2):
            res[3] += 1
        else:
            res[4] += 1
    fig = plt.figure()
    plt.pie(res, labels=labels, autopct='%1.2f%%')  # 画饼图（数据，数据对应的标签，百分数保留两位小数点）
    plt.title("Pie-figure")
    plt.savefig(path + name + '_pie')
    plt.show()


calc_star(hairdryer_data, path, 'hairdryer_data')
calc_star(pacifier_data, path, 'pacifier_data')
calc_star(microwave_data, path, 'microwave_data')

```