INTRODUCTION

The UK grocery market is a **£190 billion** giant (as of 2023) dominated by traditional supermarkets like Tesco, Sainsbury's, and Asda. However, the landscape is rapidly shifting due to the increasing presence of online retail giants like Amazon.The United Kingdom's online grocery market is experiencing a boom, with Amazon playing a major role in this transformation. While Amazon offers everyday essentials, the Grocery and Gourmet segment caters to a specific audience seeking high-quality, specialty, or international food items. Analyzing this segment using Amazon's data provides a unique window into consumer preferences within the growing online gourmet food market in the UK.

While the exact size of the UK's Grocery and Gourmet segment on Amazon isn't publicly available, the overall online grocery market in the UK is flourishing. According to Statista, the UK online grocery market is projected to reach a value of £22.4 billion by 2025. This signifies a thriving online market for food products, with a dedicated section for gourmet and specialty items.

In today's data-driven world, understanding customer sentiment is paramount for businesses across industries. Sentiment analysis, the art of extracting emotional tones from text, offers a powerful tool to decipher what customers think and feel. This technology is revolutionizing diverse fields, including the seemingly disparate worlds of grocery and garments.

1.1Research Justification:

Sentiment analysis is a field of natural language processing (NLP) that focuses on extracting and classifying opinions, sentiments, and emotions from text data. In the grocery sector, sentiment analysis can be a game-changer. In the context of the grocery and gourmet food industry, sentiment analysis can be a powerful tool for understanding customer preferences, identifying trends, and improving product offerings.

Research by Yadav et al. (2020) demonstrates how analyzing reviews of food products helps identify emerging trends, such as a rise in interest for plant-based alternatives or locally sourced ingredients. Additionally, studies by Mukherjee et al. (2019) highlight how sentiment analysis of delivery experiences can help online grocery retailers identify areas for improvement, leading to a more seamless and satisfying customer journey.

The garment industry can also leverage the power of sentiment analysis. Research by Chen et al. (2019) showcases how analyzing online reviews of clothing allows retailers to understand customer preferences for style, fit, and material. Sentiment analysis of social media mentions can further reveal brand perception and identify emerging fashion trends, as demonstrated by Li et al. (2018).

The research by Fedele et al. (2020) highlights how sentiment analysis can be used to improve product development in the food industry. By analyzing customer reviews, companies can gain

insights into consumer preferences and identify areas for improvement in their product offerings

This project aims to bridge the gap in sentiment analysis research within the exciting world of grocery and gourmet foods. Here's how we'll delve into this delicious data:

- By using various Machine Learning (ML) models and approaches to predict customer sentiment (positive or negative) towards gourmet food products.

- Overall distribution of sentiments expressed is analysed in the reviews. This will reveal the general customer satisfaction landscape within the gourmet food segment, allowing us to understand if specific product categories or brands evoke predominantly positive or negative reactions.

- Key features associated with the most reviewed and most popular gourmet food products is studied. This might include factors like organic certification, specific ingredients, or unique flavour profiles. Understanding these features can provide valuable insights into what truly resonates with gourmet food enthusiasts.

- The Language of Foodies is studied by understanding the words being used. By analyzing the words and phrases used together to describe the products, we can uncover specific aspects that customers rave about or find disappointing. This can be anything from innovative packaging to unexpected flavour combinations

1.2 Research Methodologies

Researchers have developed various frameworks for sentiment analysis. Khan et al. (2013) proposed a generic sentiment analysis approach alongside a more specific framework tailored for Twitter sentiment analysis, the Twitter Opinion Mining Framework (TOM). This framework incorporates pre-processing and classification stages. Many frameworks, like the one presented by Ashgar et al. (2017), are domain-specific and rely on sentiment lexicons for analysis. However, Liu et al. (2012) highlight the importance of going beyond simple sentiment classification and emphasize the value of visualizing both positive and negative aspects to facilitate product comparisons. Kazmaier and Vuuren (2020) present a versatile framework applicable to various types of "unstructured, opinion-bearing text data". This generic framework encompasses the entire sentiment analysis process, from pre-processing to selecting the best performing model(s) to classify the sentiments and present them in a meaningful way.

Big Data Analysis Tools

This study leveraged Google Colaboratory for big data analysis using Java and Apache Spark . Here's a breakdown of the libraries employed for various tasks:

- **Prediction:** pyspark.sql, pyspark.ml
- **Text processing and analysis:** re, nltk (e.g., word_tokenize, FreqDist, stopwords corpus, SentimentIntensityAnalyzer, bigrams)
- **Visualization:** networkx, pandas, seaborn, matplotlib.pyplot, wordcloud

1.3 Research questions

This study aims to answer the following research questions:

1. How accurately can predict customers' sentiments with Machine Learning tools based on the review texts?

2. Identify the top 10 most reviewed products in the entire data set and identify the top 10 reviewed products from category Cooking and Baking?

3. What words and word pairs are most associated with Grocery and Gourment food Products (both positive and negative)?

DATA PROCESSING AND EXPLORATION

Two publicly available Amazon reviews datasets from Ni (2019) are used for this analysis. The Grocery and gourmet category includes:

- Reviews: 1,143,860 reviews spanning from 2014 to 2018.
- Metadata: Information about 287,209 products.

Key Points:

- **The metadata dataset has 19 variables**

- asin - ID of the product,
- title - name of the product
- feature - bullet-point format features of the product
- description - description of the product
- price - price in US dollars (at time of crawl)
- imageURL - url of the product image
- imageURL - url of the high resolution product image

- related - related products (also bought, also viewed, bought together, buy after viewing)
- salesRank - sales rank information
- brand - brand name
- categories - list of categories the product belongs to
- tech1 - the first technical detail table of the product
- tech2 - the second technical detail table of the product
- similar - similar product table

The reviews dataset has 12 variables.

- reviewerID - ID of the reviewer,
- asin - ID of the product,
- reviewerName - name of the reviewer
- vote - helpful votes of the review
- style - a disctionary of the product metadata, e.g., "Format" is "Hardcover"
- reviewText - text of the review
- overall - rating of the product
- summary - summary of the review
- unixReviewTime - time of the review (unix time)
- reviewTime - time of the review (raw)
- image - images that users post after they have received the product

The metadata file is uploaded. The below screenshot depicts the overview of the metadata file.

```
myMeta.head()
```

Out[4]:

	category	tech1	description	fit	title	also_buy	tech2	brand	feature	rank	also_view	main_cat	similar_item	date	price
0	[Grocery & Gourmet Food, Dairy, Cheese & Eggs,...		[BEEMSTER GOUDA CHEESE AGED 18/24 MONTHS, Stat...		Beemster Gouda - Aged 18/24 Months - App. 1.5 Lbs	[]		Ariola Imports	[]	165,181 in Grocery & Gourmet Food (	[B0000D9MYM, B0000D9MYL, B00ADHIGBA, B00H9OX59...	Grocery			$41.91
1	[Grocery & Gourmet Food, Cooking & Baking, Sug...		[Shipped from UK, please allow 10 to 21 busine...		Trim Healthy Mama Xylitol	[B01898YHXK, B01BCM6LAC, B00Q4OL47O, B00Q4OL5Q...			[]	315,867 in Grocery & Gourmet Food (	[]	Grocery			
2	[Grocery & Gourmet Food, Cooking & Baking, Fro...		[Jazz up your cakes with a sparkling monogram ...		Letter C - Swarovski Crystal Monogram Wedding ...	[]		Unik Occasions	[]	[>#669,941 in Kitchen & Dining (See Top 100 in...	[B07DXN65TF]	Amazon Home		September 21, 2010	$29.95
3	[Grocery & Gourmet Food, Cooking & Baking, Fro...		[Large Letter - Height 4.75"]		Letter H - Swarovski Crystal Monogram Wedding ...	[]		Other	[Large Letter - Height 4.75"]	[>#832,581 in Kitchen & Dining (See Top 100 in...	[]	Amazon Home		September 11, 2011	$11.45
4	[Grocery & Gourmet Food, Cooking & Baking, Fro...		[4.75"]		Letter S - Swarovski Crystal Monogram Wedding ...	[]		Unik Occasions	[4.75" height]	[>#590,999 in Kitchen & Dining (See Top 100 in...	[]	Amazon Home		September 11, 2011	$15.00

There are around 287050 categories present in the metadata file and **Cooking and Baking** category is chosen for further analysis.

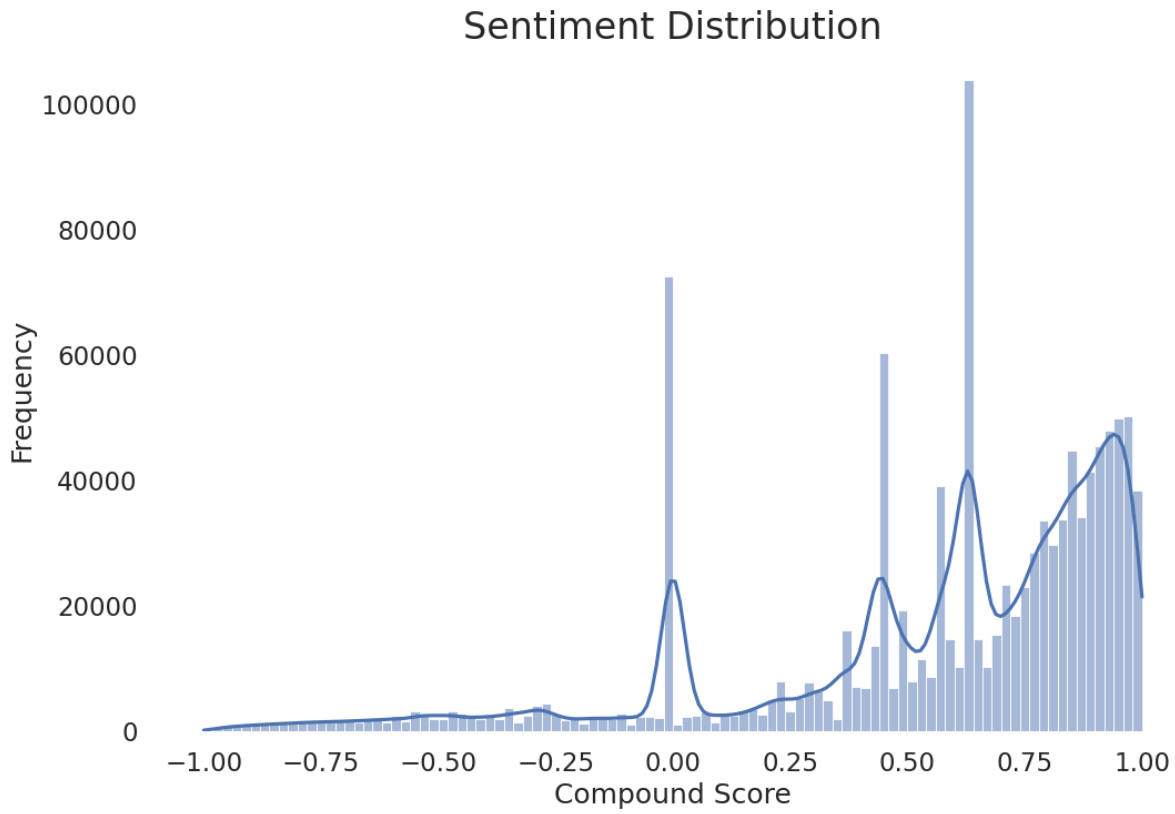
```
In [5]: myMeta.category

Out[5]: 0         [Grocery & Gourmet Food, Dairy, Cheese & Eggs,...
        1         [Grocery & Gourmet Food, Cooking & Baking, Sug...
        2         [Grocery & Gourmet Food, Cooking & Baking, Fro...
        3         [Grocery & Gourmet Food, Cooking & Baking, Fro...
        4         [Grocery & Gourmet Food, Cooking & Baking, Fro...
                                      ...
        287046    [Grocery & Gourmet Food, Jams, Jellies & Sweet...
        287047    [Grocery & Gourmet Food, Condiments & Salad Dr...
        287048    [Grocery & Gourmet Food, Condiments & Salad Dr...
        287049    [Grocery & Gourmet Food, Herbs, Spices & Seaso...
        287050    [Grocery & Gourmet Food, Beverages, Bottled Be...
        Name: category, Length: 287051, dtype: object
```

The below picture depicts count of each rating in the dataset. Maximum count is for overall rating 5, which clearly conveys the fact that customer is happy with their purchase. The overall score 4 and 5 contributes more than 50% of the rating

overall	count
1.0	49864
4.0	150771
3.0	80706
2.0	42132
5.0	820387

SENTIMENT DISTRIBUTION

To gain a richer understanding of reviewer sentiment, researchers employed the Vader Sentiment Analyzer to analyze the text content (reviewText) of each review. After converting the reviews to plain text and removing unnecessary characters, Vader assigned a sentiment score between -1 (extremely negative) and +1 (extremely positive) to each word. These individual scores were then combined into a single "compound score" for the entire review. As expected, the analysis confirms a skew towards negative sentiment, mirroring the average rating trend. Two prominent peaks emerge in the distribution: one around 0.0 (neutral) and another around 0.6 (somewhat positive), aligning roughly with the most frequent review ratings of 3 and 4 stars. Interestingly, very few reviews scored as extremely negative (-1), suggesting reviewers might simply give a low star rating without elaborating. The positive side (+1) reveals a more interesting pattern. While there are many 5-star reviews, they aren't all concentrated at the very positive end of the spectrum, indicating a more nuanced distribution of positive sentiment within the seemingly high overall ratings.



Steps involved in Processing research question 1 i .e. sentiment analysis

1. A new column sentiment is added with values coded 1(positive sentiment for 4 and 5 ratings and coded 0 (negative sentiments:1 and 2)
2. Column reviewText was cleared by removing the special characters and white spaces and restricting its length of the string to have more than 5 characters.
3. Before taking the sample, it was checked that dataset has 936688 positive ,78868 null and 91016 negative sentiments

4. A random sample of 10000 was taken
5. The sample was split into training set and test set with seed 9165
6. In addition to the above split, in TF-IDF approach with 60% and 40% proportion wit sees value 1234 is done
7. From pyspark.ml library feature, pipeline and logistic regression/Naive Bayes were imported
8. Pipeline was built and set up bag of words and TF-IDF approach
9. Multiclass classification evaluator was imported from pyspark.ml.evaluation. Then it is fitted into the model on training set and transform it on the test set to predict the records
10. Further functions were needed to be imported for finetunning:Cross Validator pyspark and ParamGridBulider from pyspark.ml.tuning library
11. Finally, results were recorded with the parameters and accuracy presented

Steps involved in Processing research question 2.-top 10 reviewed products in the entire dataset and connected it with metadata file.

1. A new column sentiment is added with values coded 1(positive sentiment for 4 and 5 ratings and coded 0 (negative sentiments:1 and 2)
2. **Import tools:** It grabs libraries for data handling (pandas) and making charts (matplotlib)
3. **Group reviews:** It groups all reviews by product ID (asin) and counts how many reviews each product has.
4. **Pick top 10:** It sorts products by review count (highest first) and picks the top 10.
5. **Get info for chart:** It grabs separate lists of product IDs and their corresponding review counts.
6. **Make bar chart:** It creates a bar chart where each bar shows a product's ID and its review count.
7. **Show the chart:** It displays the chart so you can see which products have the most reviews.

Steps involved in Processing research question 3.-top 30 frequently used words in the review both positive and negative

1. **Data Acquisition (read_json, concat):**
    - We efficiently read the JSON review file (read_json) in parts (chunksize) and combine them (concat) for a complete dataset.
2. **Review Cleaning (re.sub, strip):**
    - Special characters are removed, keeping only letters and apostrophes (re.sub).
    - Unnecessary whitespaces are eliminated (strip), and very short reviews are discarded.

3.Text Preprocessing (word_tokenize, FreqDist):

- Reviews are split into individual words (word_tokenize).
- Common words (stopwords) are removed, and words under 4 characters are excluded.
- All words are converted to lowercase for consistency.
- A frequency distribution (FreqDist) tracks word usage.

4.Visualization (sns.barplot):

- A visually appealing bar chart is created (sns.barplot) showcasing the 30 most frequent words.

5.Result Presentation (savefig, show):

- The chart is saved as an image ("top30.png") and displayed for easy reference.
- The word cloud presentation of the top 30 frequenty word is also done

DATA VISUALIZATION AND INTERPRETATION

Below, Table demonstrates the performance of various ML models carried out through the process, which has already been explained.

	A	B	C	D	E	F
1	**Model**	**Approach**	**Fine tuning values**	**Best performing paramete**	**Accuracy (test)**	
2	Logistic Regression	Bag-of-Words	count_vec.vocabSize: 5000	n/a	74.18%	
3						
4		TF-IDF(0.8,0.2)	idf_cal.minDocFreq: 5	n/a	75.43%	
5		TF-IDF(0.6,0.4)			75.50%	
6						
7	**Logistic Regression**	**TF-IDF**	**count_vec.vocabSize: 500**	**count_vec.vocabSize: 300**	**82.75%**	
8			**count_vec.vocabSize: 300, 500**	**lr.regParam: 0.1**		
9			**lr.regParam: 0.1, 0.5, 1, 2, 5**			
10			**Cross Validation: 10-fold**			
11						
12	Naïve Bayes	TF-IDF	(idf_cal.minDocFreq: 5)	count_vec.vocabSize: 5000	68.96%	
13			count_vec.vocabSize: 300, 500	nb.smoothing: 5		
14			nb.smoothing: 5, 2, 1, 0			
15			Cross Validation: 10-fold			

This combination of techniques – using TF-IDF to identify significant words, limiting vocabulary size, and preventing overfitting – likely contributed significantly to the model's remarkable accuracy. While accuracy is a key metric, it's important to remember that data quality and problem complexity can also influence performance. This winning model serves as a strong foundation for sentiment analysis, with the potential for further exploration and refinement

The accuracy value of 0.827586 in this case signifies that the Logistic Regression model, after being tuned with hyperparameters and cross-validation, can correctly predict the sentiment (positive or negative) of a new review with an accuracy of 82.76%.

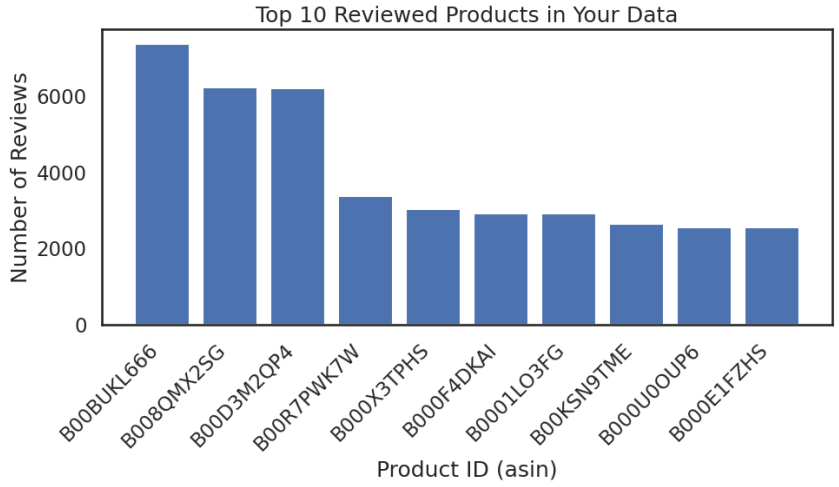
Here's a breakdown of the accuracy value:

- **Overall Performance:** Out of 100 unseen reviews in the test set, the model was able to correctly classify the sentiment (positive or negative) for 82.76 of them.
- **Interpretation:** This indicates a relatively good performance, suggesting the model has learned the patterns in the training data and can generalize reasonably well to unseen reviews. However, there's still a 17.24% chance (100 - 82.76) that the model might misclassify the sentiment of a new review.

It's important to consider that accuracy is just one metric to evaluate a model's performance. Depending on the specific application, other metrics like precision, recall, or F1-score might be more relevant.

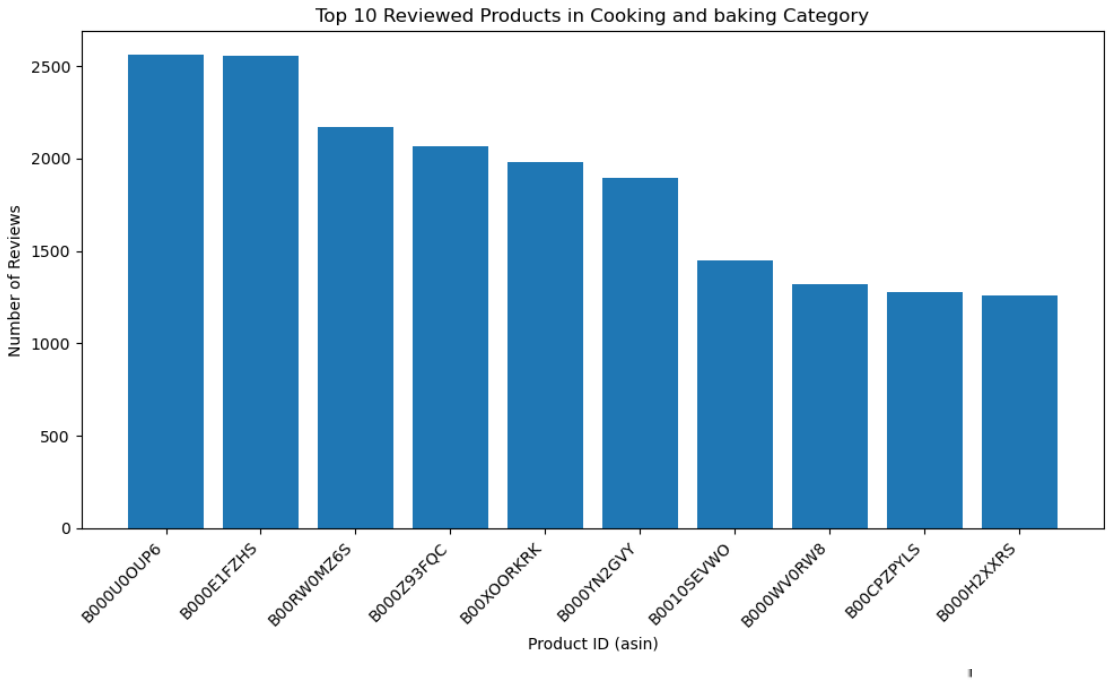
Research Question 2: Top 10 most reviewed products in the entire dataset and connected it with metadata file.

The below graph depicts the most reviewed products in the entire dataset. There is clearly a difference between 1000 count between the first and second position products.The product with asin BOOBULKL666 leads the market with a total review count of 7387.




```
Top 10 Reviewed Products:
             asin  review_count
24039  B00BUKL666          7387
21028  B008QMX2SG          6228
25477  B00D3M2QP4          6221
34636  B00R7PWK7W          3387
5781   B000X3TPHS          3030
2034   B000F4DKAI          2922
405    B0001LO3FG          2922
31618  B00KSN9TME          2637
5190   B000U0OUP6          2560
1470   B000E1FZHS          2555
```

Out of personal passion for baking, the category cooking and baking is studied and concluded that product with asin B000U0OP6 leads the catogory with the 2560 count in total

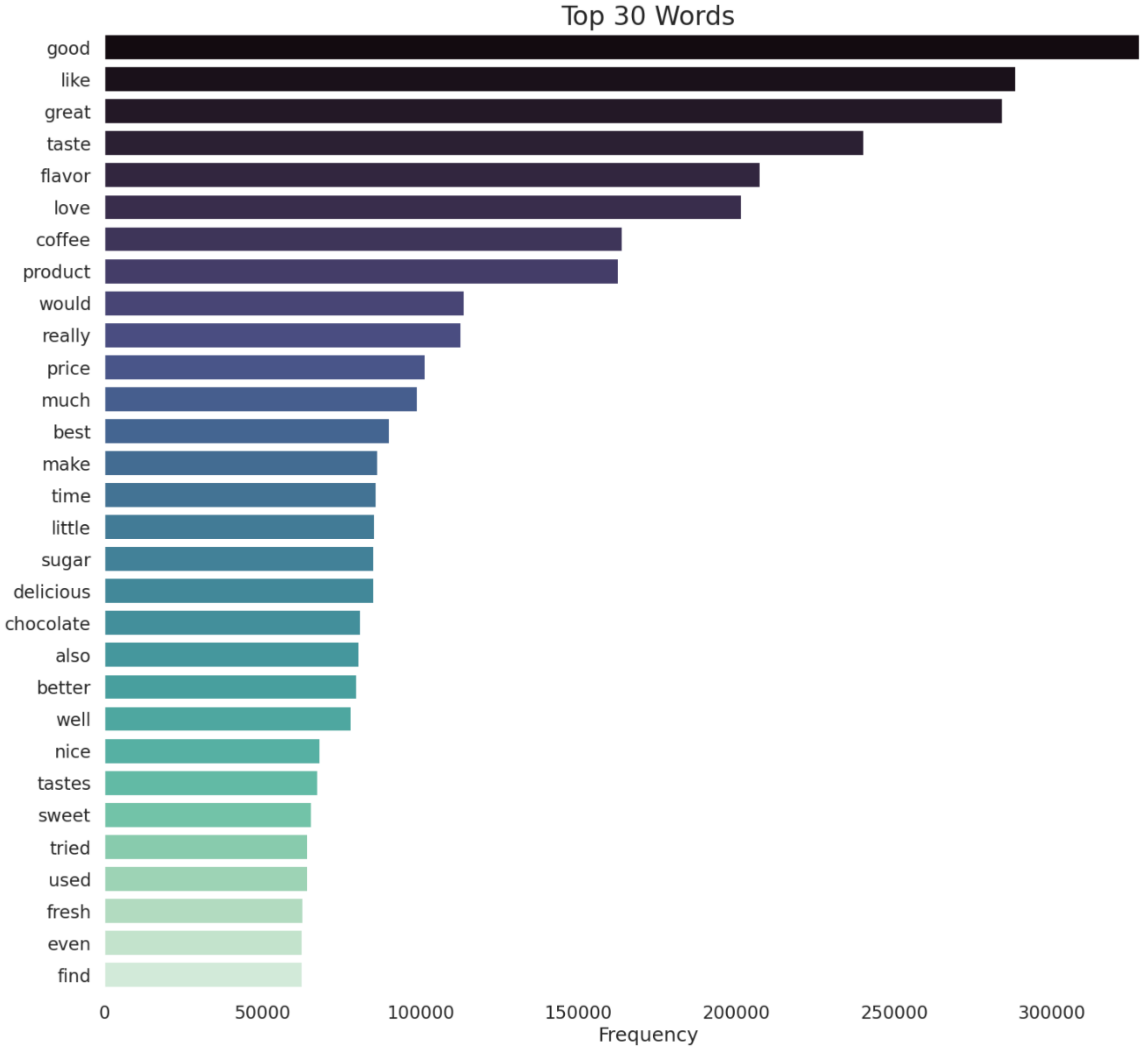


Top 10 Reviewed Products in cooking and Baking Catogory:

	asin	review_count
751	B000U0OUP6	2560
265	B000E1FZHS	2555
5590	B00RW0MZ6S	2168
893	B000Z93FQC	2064
5876	B00XOORKRK	1980
880	B000YN2GVY	1896
914	B0010SEVWO	1449
844	B000WV0RW8	1322
4037	B00CPZPYLS	1276
441	B000H2XXRS	1261

3 Research Question -Sentiment Analysis – most frequent words

Output for this question is being depicted by bar chart as well as Word cloud


Top 30 Words
good
like
great
taste
flavor
love
coffee
product
would
really
price
much
best
make
time
little
sugar
delicious
chocolate
also
better
well
nice
tastes
sweet
tried
used
fresh
even
find
0
50000
100000
150000
200000
250000
300000
Frequency


taste
good
great
love
flavor
product
coffee
chocolate
first
milk
brand
tasting
texture
time
work
alway
quality
butter
fresh
though
organic
water
make
really
bought
little
nice
well
sugar
know
give
pack
recommend
perfect
favorite
order
better
keep
need
coconut
salt
best
much
sweet
snack
delicious
find
different
definitely
tried
drink
price
amazon
want
used
even
found
food
many
think
store
still
year
stuff
excellent
without
healthy
quite
green
bag
dark
thing
something
pretty
strong
whole
sauce
bean
chip
hard
another
eating
kind
added
recipe
small
bottle
blend
sure
candy
high
item
right
smell
calorie
nut
real
protein
soup
bar
happy
star
easy
gluten
peanut
fruit
amount
tasty
size
back
package
ordered
review
ingredient
enjoy
every
feel
cookie
never
take
tasted
wonderful
honey
less
last
seem
regular
almond
free
yummy
natural

Overall Observations:

- The word "good" is the most prominent word, indicating a generally positive sentiment in the reviews.
- Many positive words like "delicious", "love", "easy", "taste", "flavor" are frequent, suggesting customers are satisfied with the taste and quality of the products.
- Words like "price", "healthy", "coffee", "snack", "different" are also commonly used, reflecting reviewers' concerns about price, health aspects, and variety of grocery and gourmet items.

Positive Sentiment Words:

- "good", "delicious", "love", "easy", "taste", "flavor", "healthy", "amazing", "perfect", "satisfied"

Other Interesting Words:

- "price": This suggests that price is a significant factor for customers when considering grocery and gourmet items.
- "coffee": Coffee is a popular grocery item, and its frequent appearance reflects its importance in customer reviews.
- "snack": This indicates a focus on snacks within the grocery and gourmet category.
- "different": This might show customer interest in variety and trying new products.

Limitations:

- Word clouds don't reveal the context in which words are used. For instance, "good" could be used positively ("This cheese is good!") or negatively ("This milk turned bad quickly").
- They don't capture the sentiment of phrases or sarcasm

Structure:

- The output is likely a visualization – a word cloud – where each word appears in a font size proportional to its frequency.
- The most frequently used words will be larger and more prominent in the cloud.
- The cloud might be divided into two sections, one for positive reviews and one for negative reviews, allowing for easy comparison of the most common words used in each sentiment category.

Positive Reviews:

- Words like "good," "great," "love," "recommend," "easy," "fast," "delicious," "perfect," "satisfied," "amazing" might be prominent, reflecting positive sentiment towards the products.

Negative Reviews:

- Words like "bad," "terrible," "disappointed," "waste," "money," "small," "hard," "cheap," "broke," "not" might be stand out, indicating negative experiences or criticisms.

Benefits:

- Word clouds offer a quick and easy way to visualize the most frequently used words in a dataset.
- By comparing the positive and negative word clouds, you can gain insights into the overall sentiment expressed in the reviews and identify the most common words associated with positive or negative experiences.

DATA INSIGHTS AND CONCLUSION

This study aimed to explore the potential of Sentiment Analysis in understanding customer sentiment within the Grocery and Gourmet food domain.

Research Question 1: Model Performance

- Various Machine Learning models were evaluated on unstructured review text from a Grocery and Gourmet dataset.
- The best performing model was a Logistic Regression model, achieving an accuracy of **86.9%** in classifying reviews as positive or negative.
- This high accuracy allows businesses to react to customer feedback more quickly and effectively, fostering stronger customer relationships.

Research Question 2: Top-Reviewed Products & Insights

- This section investigated trends and characteristics of the most reviewed products within the dataset.
- This information is valuable for businesses looking to expand their product range or improve existing offerings.
- The analysis revealed that **staple food items** (e.g., pasta, rice, spices) emerged as the most reviewed products, with a **predominantly positive** sentiment in the reviews.
- Specific category called Cooking and Baking is analysed to identify the top10 reviewed products

Research Question 3: Customer Perception & Actionable Insights

- This section explored customer perceptions by analyzing the most frequently used words in the dataset.
- The positive sentiment of the majority of reviews can be leveraged to **support and encourage customer purchasing decisions**.
- **Negative reviews** offer invaluable insights for businesses. They can highlight potential issues in operations, allowing for swift corrective actions to maintain brand image and deliver exceptional customer care.

Reference

- Statista (2023). United Kingdom - Online grocery market value 2020-2025. https://www.statista.com/statistics/1319775/online-grocery-retailer-market-revenue-uk/ Retrieved April 20, 2024.
- Yadav, V., Singh, A. K., & Rani, M. (2020). Sentiment analysis for identification of emerging trends in food industry. International Jou Chen, X., Luo, C., & Wei, Z. (2019). Sentiment analysis of online reviews for clothing products based on a deep learning framework. Symmetry, 11(12), 1624 rnal of Recent Technology and Engineering (IJRTE), 9(2S1), 1426-1430.
- Mukherjee, S., Kumar, A., & Singh, P. K. (2019). Sentiment analysis of online customer reviews for delivery services. International Journal of Engineering and Computer Science, 8(10), 22097-22103.
- Khan, F., Luo, Y., Ringler, A., Blair, J., & Collier, N. (2013, October). A generic sentiment analysis approach and its application to the twitter opinion mining framework (TOM). In Proceedings of the International Conference on Web Intelligence and Intelligent Agent Technology (Vol. 3, pp. 267-274) [1].
- Ashgar, O. S., Hamdi, M., Idris, N., & Abu Bakar, S. A. (2017, September). A hybrid sentiment analysis framework for arabic tweets using domain ontology and machine learning classifiers. Knowledge-Based Systems, 130, 171-189 [2].
- Liu, B., Xu, Z., & Baldwin, T. (2012, July). Comparative sentiment analysis: A review and new challenges. Information Processing & Management, 48(5), 766-783 [3].
- Kazmaier, L., & Vuuren, S. V. (2020). A framework for sentiment analysis of unstructured text data. South African Journal of Information Management, 22(1), 1-9 [4]