By Priya Nandan Shrotri

989454108

Sentiment Analysis of Amazon Fine Food Reviews

Primary Goal

This project aims to perform sentiment analysis on the customer reviews of Fine Foods from the Amazon marketplace. The main goal is to understand the ratio of positive to negative reviews and help the business make informed market decisions.

Challenges Faced

- The volume of Amazon customer reviews across multiple product categories makes it challenging to ensure data consistency and accuracy.

- Unstructured textual data, with variations in language, tone, and grammar, complicates achieving accurate sentiment classification.

- The subjective nature of customer reviews introduces challenges in maintaining unbiased sentiment labeling.

Project Approach

Data Extraction: Collecting customer reviews from the Amazon Fine Foods dataset for analysis.

Data Preprocessing: Cleaning the data by handling missing values, removing noise (e.g., punctuation and stop words), and normalizing text for consistent processing.

Sentiment Analysis with VADER: Utilizing the VADER library in Python to analyze customer reviews and assign sentiment scores based on textual data.

Linear Regression Analysis: Implementing a linear regression model to compare the actual review scores with the sentiment scores generated from the text, assessing the model's predictive accuracy.

Key Concepts – VADER Library

Overview: VADER (Valence Aware Dictionary and sEntiment Reasoner) is a Python library used for sentiment analysis.

Specialization: It is designed to analyze text, particularly social media content, for sentiment by evaluating the positivity, negativity, or neutrality of the text.

Key Features: VADER combines a lexicon-based approach with rule-based sentiment intensity scoring.

Advantages: It is simple to use, performs well on short text, and provides normalized scores for better comparability.

Key Concepts – Linear Regression

Overview: Linear regression is a statistical method used to model the relationship between a dependent variable and one or more independent variables.

Purpose: It predicts the dependent variable's value based on the linear relationship with the independent variables.

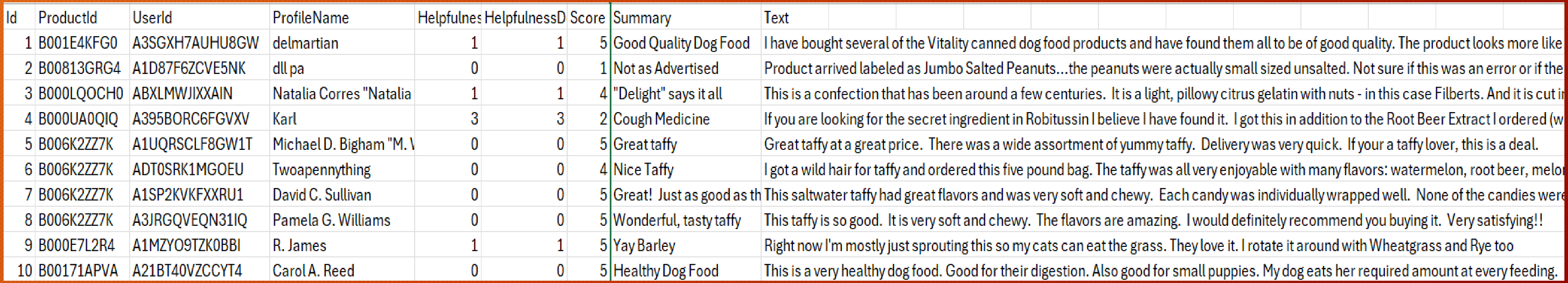
Key Elements: The model includes coefficients that measure the impact of each predictor and an intercept representing the baseline value.

Evaluation: Performance is measured using metrics like Mean Squared Error (MSE) and R-squared to assess accuracy and variability explained.

Applications: Commonly used in forecasting, trend analysis, and establishing relationships between variables.

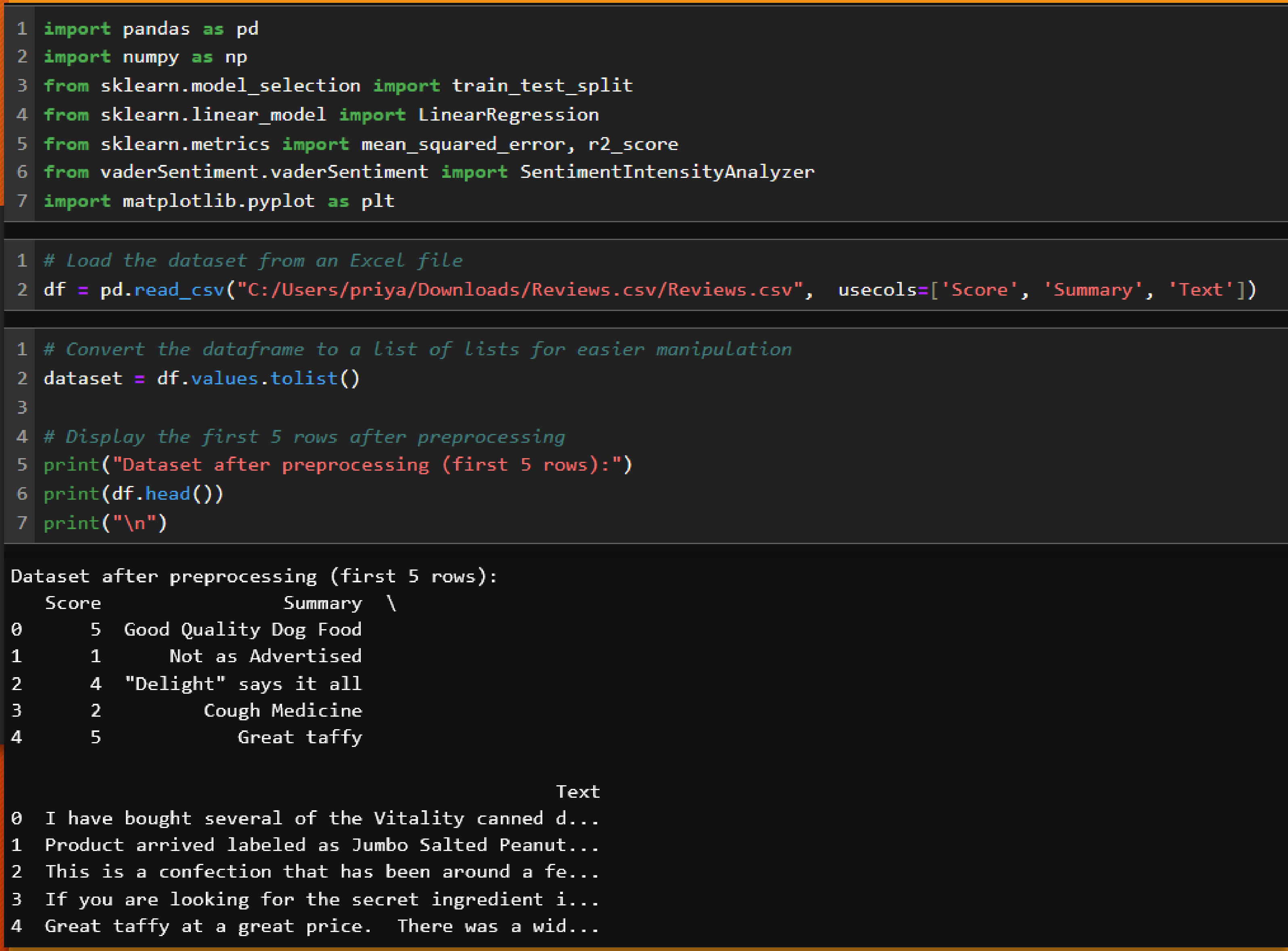
Data Source

The dataset comprises approximately 500,000 Amazon customer reviews on fine food products. It includes information such as helpfulness ratings, review scores (1 to 5), summaries, and full reviews. The important columns are Score and Text.



Id	ProductId	UserId	ProfileName	Helpfulnes	HelpfulnessD	Score	Summary	Text
1	B001E4KFG0	A3SGXH7AUHU8GW	delmartian	1	1	5	Good Quality Dog Food	I have bought several of the Vitality canned dog food products and have found them all to be of good quality. The product looks more like
2	B00813GRG4	A1D87F6ZCVE5NK	dll pa	0	0	1	Not as Advertised	Product arrived labeled as Jumbo Salted Peanuts...the peanuts were actually small sized unsalted. Not sure if this was an error or if the
3	B000LQOCH0	ABXLMWJIXXAIN	Natalia Corres "Natalia	1	1	4	"Delight" says it all	This is a confection that has been around a few centuries. It is a light, pillowy citrus gelatin with nuts - in this case Filberts. And it is cut i
4	B000UA0QIQ	A395BORC6FGVXV	Karl	3	3	2	Cough Medicine	If you are looking for the secret ingredient in Robitussin I believe I have found it. I got this in addition to the Root Beer Extract I ordered (w
5	B006K2ZZ7K	A1UQRSCLF8GW1T	Michael D. Bigham "M. \	0	0	5	Great taffy	Great taffy at a great price. There was a wide assortment of yummy taffy. Delivery was very quick. If your a taffy lover, this is a deal.
6	B006K2ZZ7K	ADT0SRK1MGOEU	Twoapennything	0	0	4	Nice Taffy	I got a wild hair for taffy and ordered this five pound bag. The taffy was all very enjoyable with many flavors: watermelon, root beer, melo
7	B006K2ZZ7K	A1SP2KVKFXXRU1	David C. Sullivan	0	0	5	Great! Just as good as th	This saltwater taffy had great flavors and was very soft and chewy. Each candy was individually wrapped well. None of the candies were
8	B006K2ZZ7K	A3JRGQVEQN31IQ	Pamela G. Williams	0	0	5	Wonderful, tasty taffy	This taffy is so good. It is very soft and chewy. The flavors are amazing. I would definitely recommend you buying it. Very satisfying!!
9	B000E7L2R4	A1MZYO9TZK0BBI	R. James	1	1	5	Yay Barley	Right now I'm mostly just sprouting this so my cats can eat the grass. They love it. I rotate it around with Wheatgrass and Rye too
10	B00171APVA	A21BT40VZCCYT4	Carol A. Reed	0	0	5	Healthy Dog Food	This is a very healthy dog food. Good for their digestion. Also good for small puppies. My dog eats her required amount at every feeding.

Code explanation – Data Extraction and Preprocessing



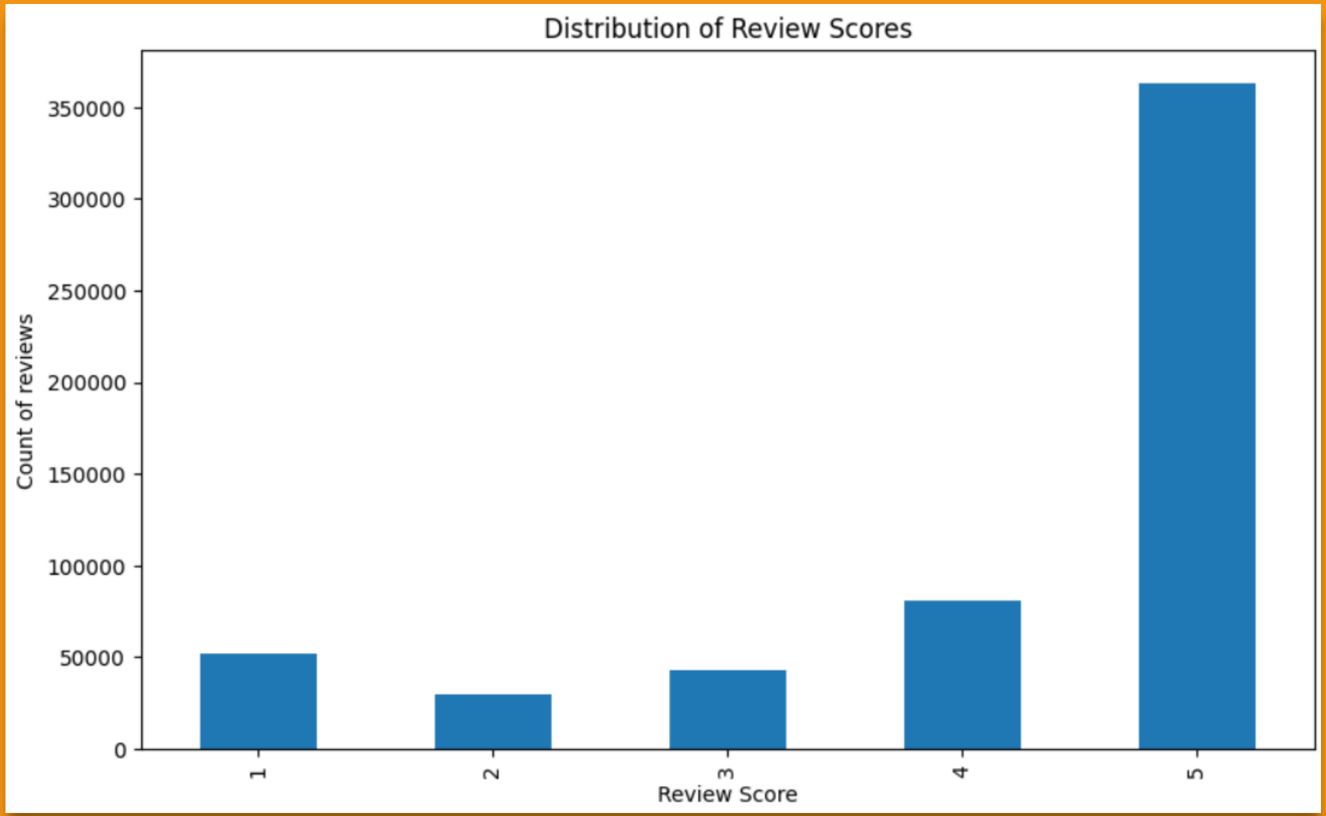
```python
import pandas as pd
import numpy as np
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_squared_error, r2_score
from vaderSentiment.vaderSentiment import SentimentIntensityAnalyzer
import matplotlib.pyplot as plt
```

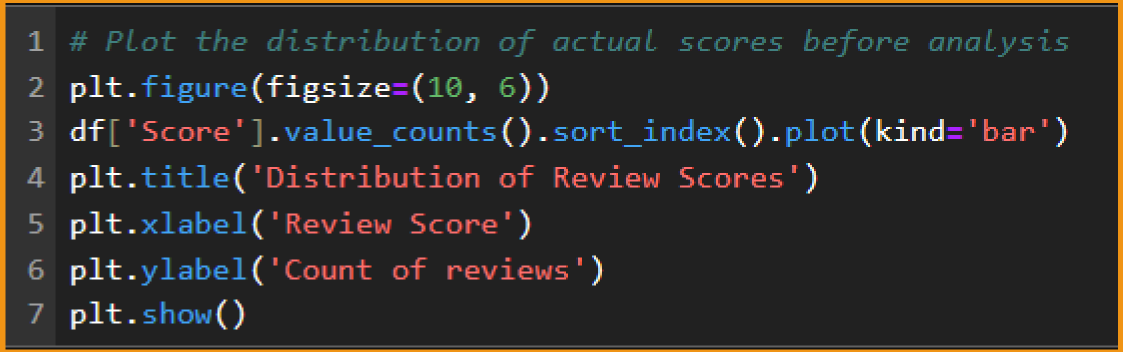
```python
# Load the dataset from an Excel file
df = pd.read_csv("C:/Users/priya/Downloads/Reviews.csv/Reviews.csv",  usecols=['Score', 'Summary', 'Text'])
```

```python
# Convert the dataframe to a list of lists for easier manipulation
dataset = df.values.tolist()

# Display the first 5 rows after preprocessing
print("Dataset after preprocessing (first 5 rows):")
print(df.head())
print("\n")
```

```
Dataset after preprocessing (first 5 rows):
   Score                Summary  \
0      5  Good Quality Dog Food
1      1      Not as Advertised
2      4  "Delight" says it all
3      2         Cough Medicine
4      5            Great taffy


                                                Text
0  I have bought several of the Vitality canned d...
1  Product arrived labeled as Jumbo Salted Peanut...
2  This is a confection that has been around a fe...
3  If you are looking for the secret ingredient i...
4  Great taffy at a great price.  There was a wid...
```

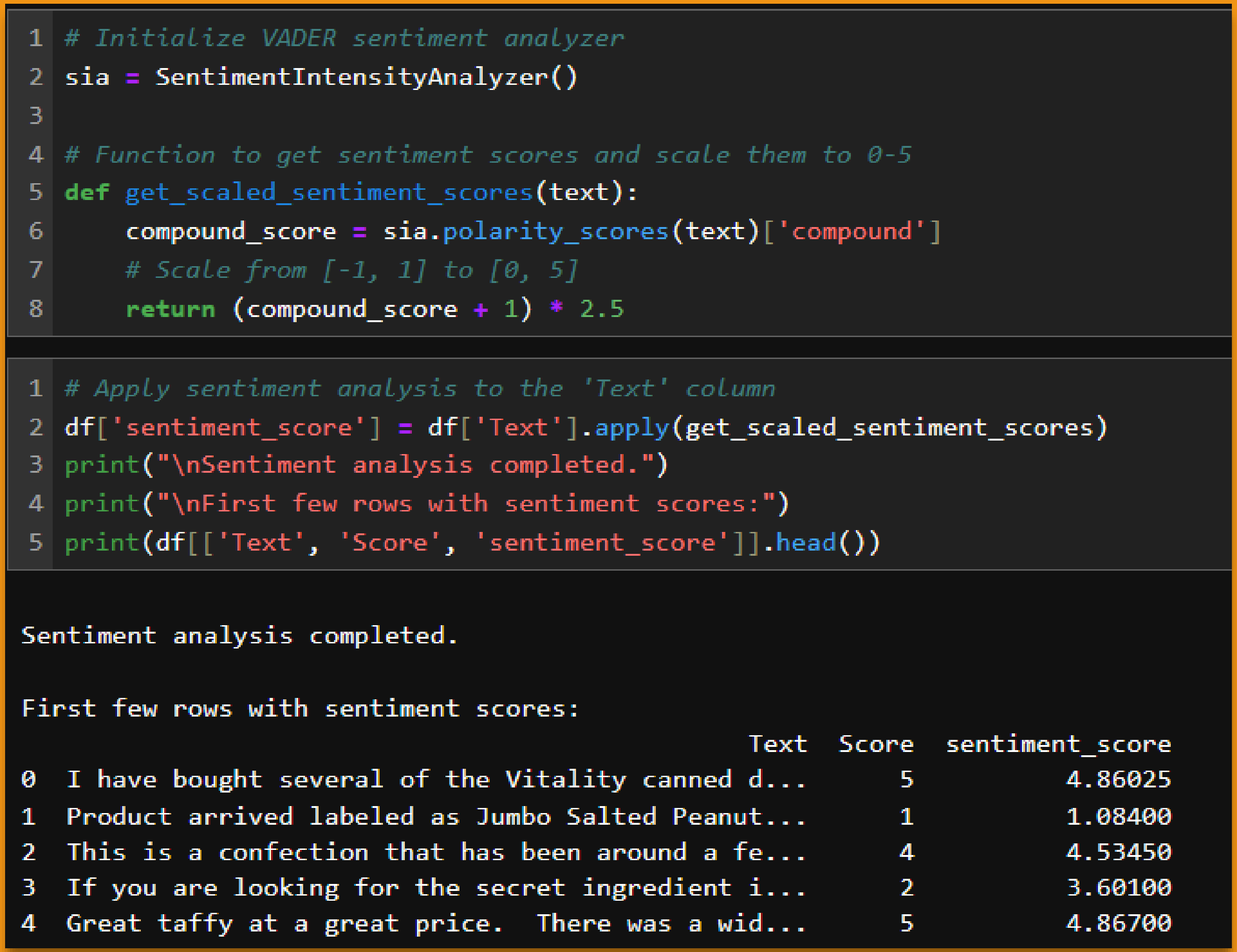
Code Explanation - Plotting Actual Review Scores





```
# Plot the distribution of actual scores before analysis
plt.figure(figsize=(10, 6))
df['Score'].value_counts().sort_index().plot(kind='bar')
plt.title('Distribution of Review Scores')
plt.xlabel('Review Score')
plt.ylabel('Count of reviews')
plt.show()
```

Code Explanation - Sentiment Analysis

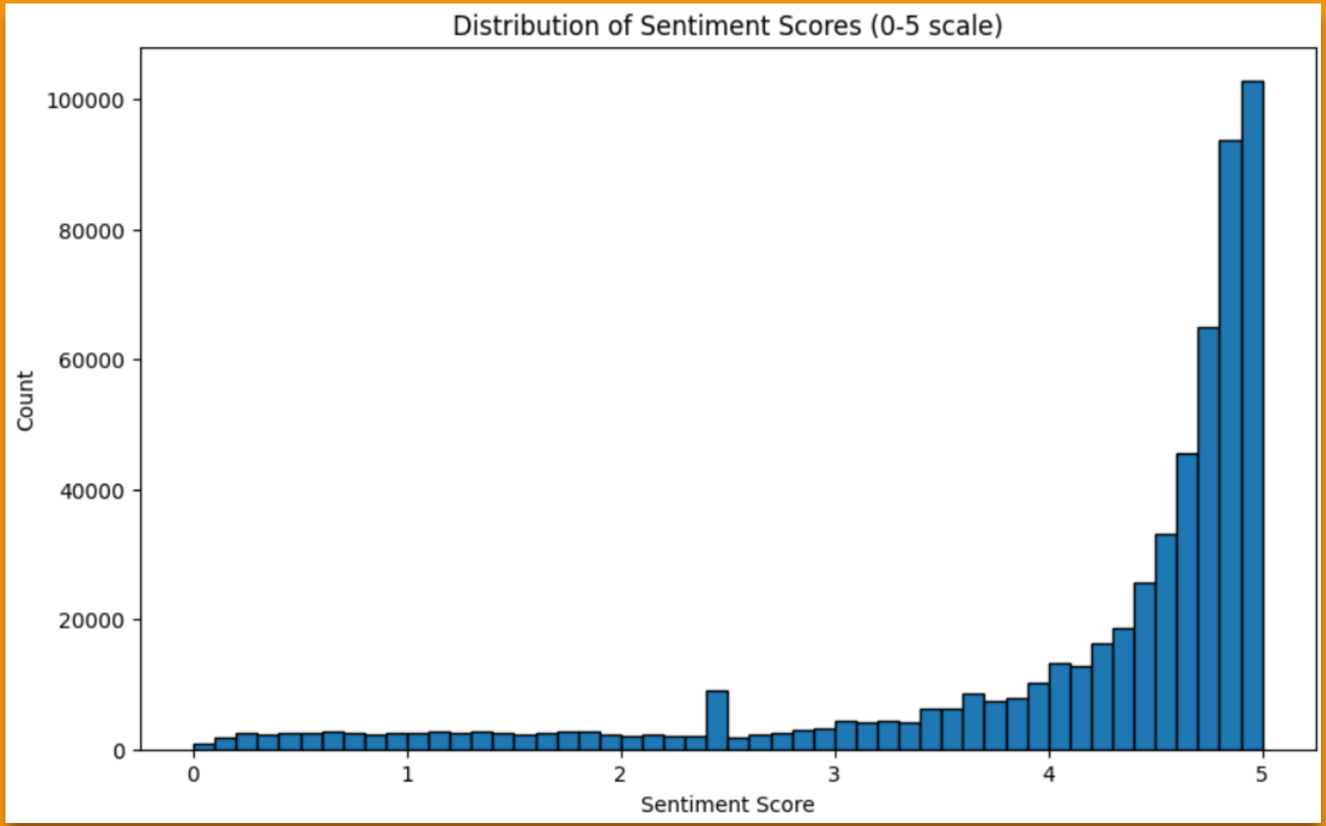


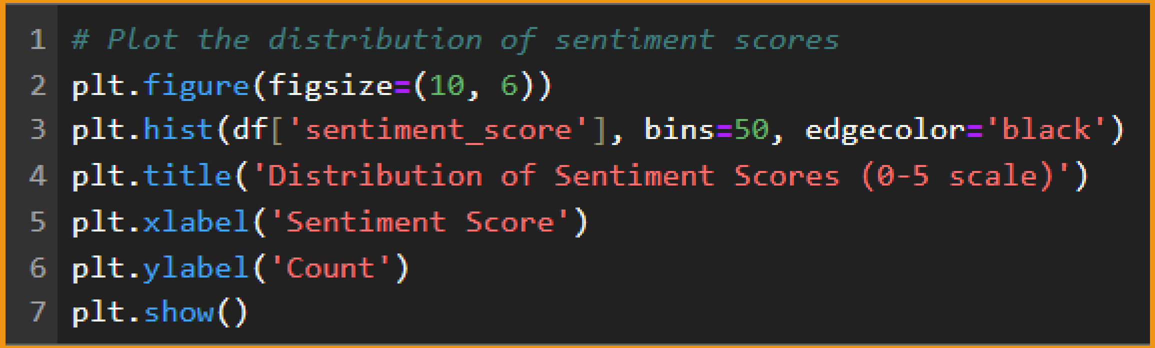
```
# Initialize VADER sentiment analyzer
sia = SentimentIntensityAnalyzer()

# Function to get sentiment scores and scale them to 0-5
def get_scaled_sentiment_scores(text):
    compound_score = sia.polarity_scores(text)['compound']
    # Scale from [-1, 1] to [0, 5]
    return (compound_score + 1) * 2.5
```

```
# Apply sentiment analysis to the 'Text' column
df['sentiment_score'] = df['Text'].apply(get_scaled_sentiment_scores)
print("\nSentiment analysis completed.")
print("\nFirst few rows with sentiment scores:")
print(df[['Text', 'Score', 'sentiment_score']].head())
```

```

Sentiment analysis completed.

First few rows with sentiment scores:
                                                Text  Score  sentiment_score
0  I have bought several of the Vitality canned d...      5          4.86025
1  Product arrived labeled as Jumbo Salted Peanut...      1          1.08400
2  This is a confection that has been around a fe...      4          4.53450
3  If you are looking for the secret ingredient i...      2          3.60100
4  Great taffy at a great price.  There was a wid...      5          4.86700
```

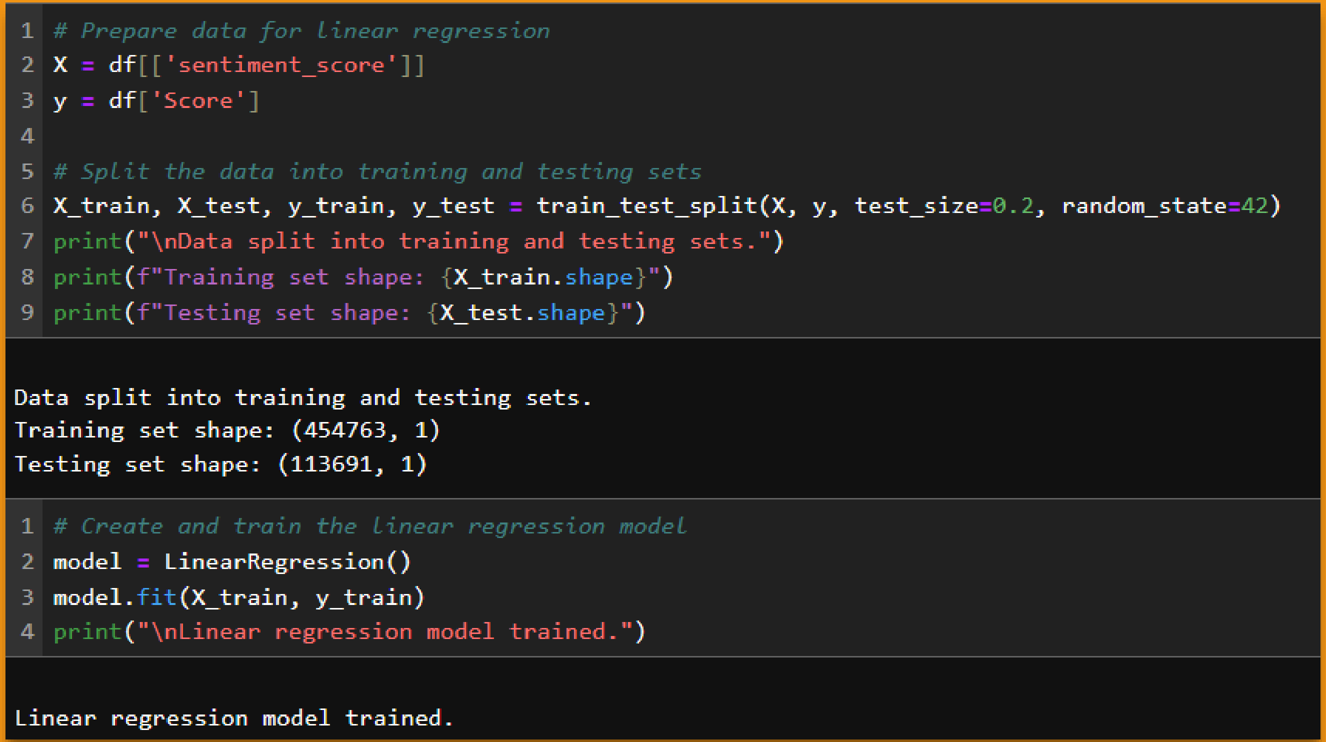
Code Explanation - Plotting New Sentiment Scores Assigned





```python
# Plot the distribution of sentiment scores
plt.figure(figsize=(10, 6))
plt.hist(df['sentiment_score'], bins=50, edgecolor='black')
plt.title('Distribution of Sentiment Scores (0-5 scale)')
plt.xlabel('Sentiment Score')
plt.ylabel('Count')
plt.show()
```

Code Explanation – Linear Regression Model Creation and Training



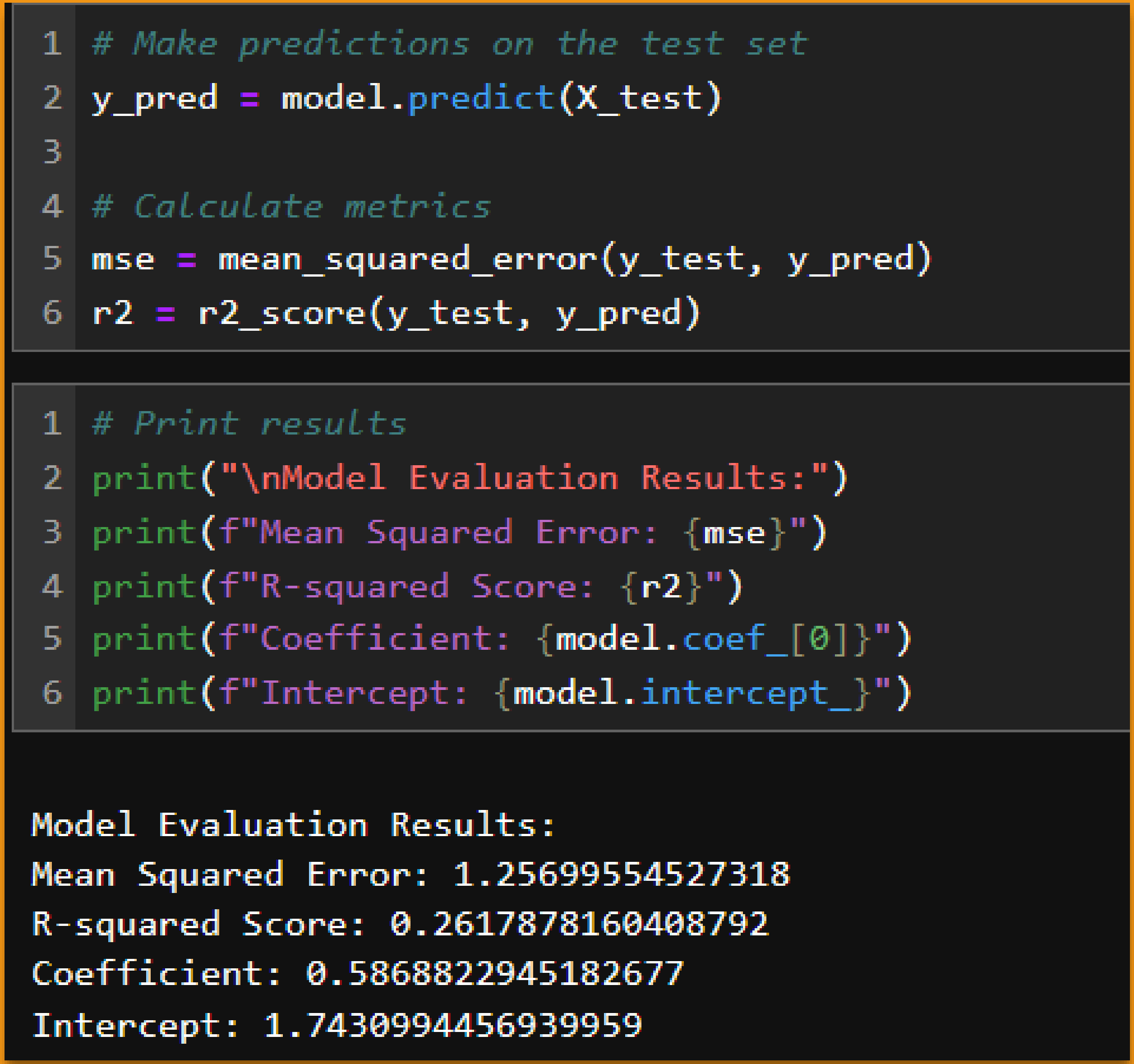
```
# Prepare data for linear regression
X = df[['sentiment_score']]
y = df['Score']

# Split the data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
print("\nData split into training and testing sets.")
print(f"Training set shape: {X_train.shape}")
print(f"Testing set shape: {X_test.shape}")
```

```

Data split into training and testing sets.
Training set shape: (454763, 1)
Testing set shape: (113691, 1)
```

```
# Create and train the linear regression model
model = LinearRegression()
model.fit(X_train, y_train)
print("\nLinear regression model trained.")
```

```

Linear regression model trained.
```

Code Explanation - Testing the Model

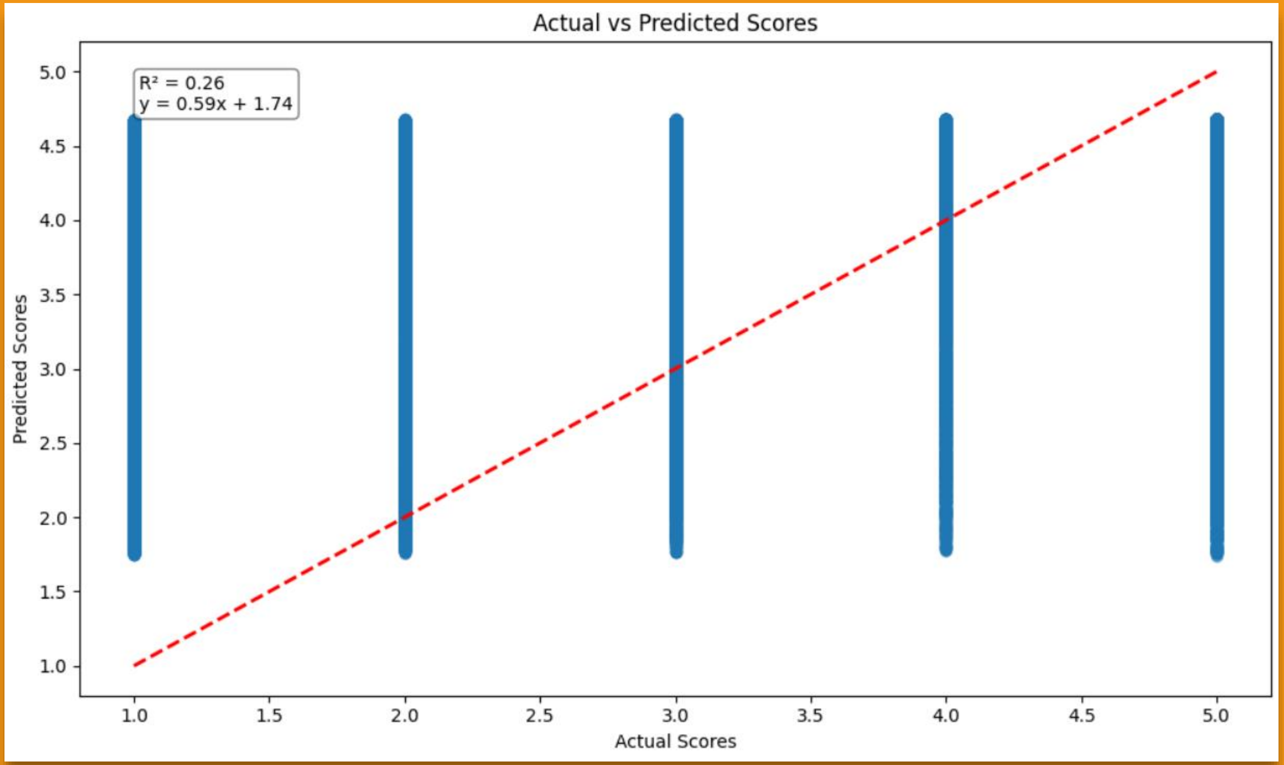


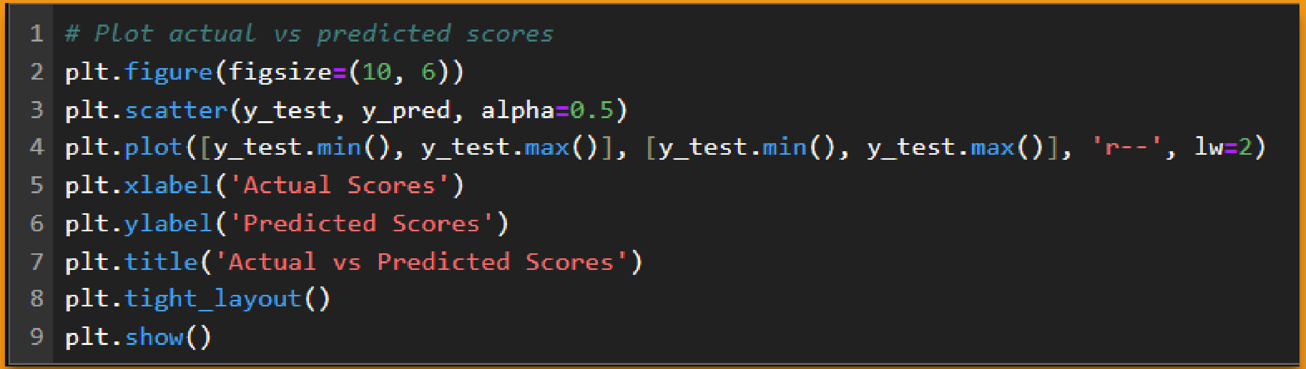
```python
# Make predictions on the test set
y_pred = model.predict(X_test)

# Calculate metrics
mse = mean_squared_error(y_test, y_pred)
r2 = r2_score(y_test, y_pred)
```

```python
# Print results
print("\nModel Evaluation Results:")
print(f"Mean Squared Error: {mse}")
print(f"R-squared Score: {r2}")
print(f"Coefficient: {model.coef_[0]}")
print(f"Intercept: {model.intercept_}")
```

```
Model Evaluation Results:
Mean Squared Error: 1.25699554527318
R-squared Score: 0.2617878160408792
Coefficient: 0.5868822945182677
Intercept: 1.7430994456939959
```

Code explanation – Plotting the Regression Model

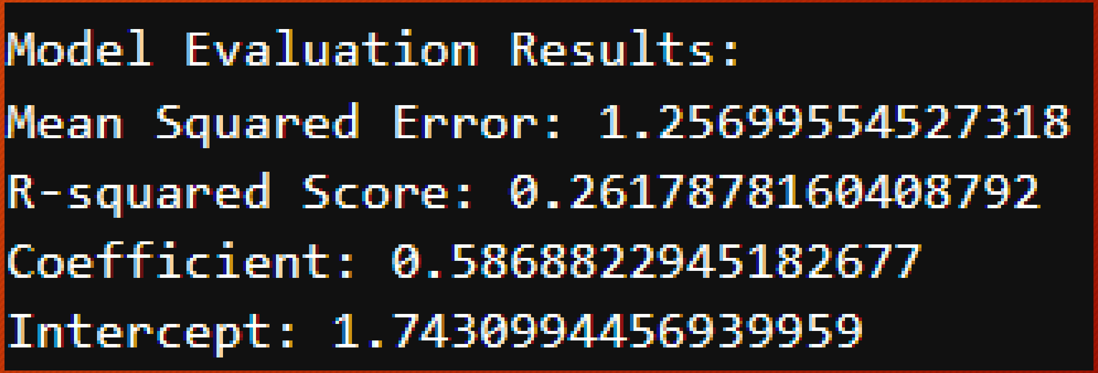




```python
# Plot actual vs predicted scores
plt.figure(figsize=(10, 6))
plt.scatter(y_test, y_pred, alpha=0.5)
plt.plot([y_test.min(), y_test.max()], [y_test.min(), y_test.max()], 'r--', lw=2)
plt.xlabel('Actual Scores')
plt.ylabel('Predicted Scores')
plt.title('Actual vs Predicted Scores')
plt.tight_layout()
plt.show()
```

Code Results – Linear Regression Results

The linear regression model produced a **Mean Squared Error (MSE)** of 1.26, indicating the average squared difference between predicted and actual review scores. An **R-squared score** of 0.26 shows that 26% of the variability in the review scores is explained by the sentiment analysis model. The **coefficient (0.59)** suggests that for each unit increase in the sentiment score, the review score increases by 0.59 on average, while the **intercept (1.74)** represents the baseline review score when the sentiment score is zero.



```
Model Evaluation Results:
Mean Squared Error: 1.25699554527318
R-squared Score: 0.2617878160408792
Coefficient: 0.5868822945182677
Intercept: 1.7430994456939959
```

Key Takeaways

Model Accuracy: The linear regression model achieved an R-squared score of 0.26, explaining 26% of the variance in review scores, and a Mean Squared Error (MSE) of 1.26.

Sentiment Prediction: Sentiment analysis using VADER provides an effective means of predicting sentiment from customer reviews, assigning scores that correlate with actual review ratings.

Sentiment Analysis: VADER successfully identifies positive, negative, and neutral sentiments in fine food reviews, offering valuable insights into customer opinions.

Model Limitations: While the model explains some variability, there is room for improvement in predicting actual review scores more accurately.

Further Developments: Enhancing the model with more advanced techniques (e.g., deep learning or additional features) could improve predictive accuracy and provide deeper insights into customer sentiment.

Business Impact: The findings can guide business decisions by helping identify key factors influencing customer satisfaction and dissatisfaction.

Challenges Faced

Sentiment Classification Accuracy: Ensuring that the VADER sentiment analysis correctly interprets nuanced language and context in reviews, as sarcasm or ambiguous language can affect sentiment prediction.

Model Generalization: The linear regression model might overfit or underfit, as it only explains a portion (26%) of the variance in review scores, indicating the need for more robust modeling techniques.

Scalability: Processing and analyzing such a large dataset (500,000 reviews) may require significant computational resources and efficient data handling strategies.

Sentiment Bias: Bias in sentiment prediction could arise from the nature of the reviews themselves (e.g., overly positive or negative reviews), which may not represent the general customer sentiment accurately.

Future Prospects

Enhanced Sentiment Analysis: Incorporating advanced natural language processing (NLP) techniques, such as deep learning models (e.g., BERT or GPT), could improve sentiment classification accuracy and handle more complex expressions.

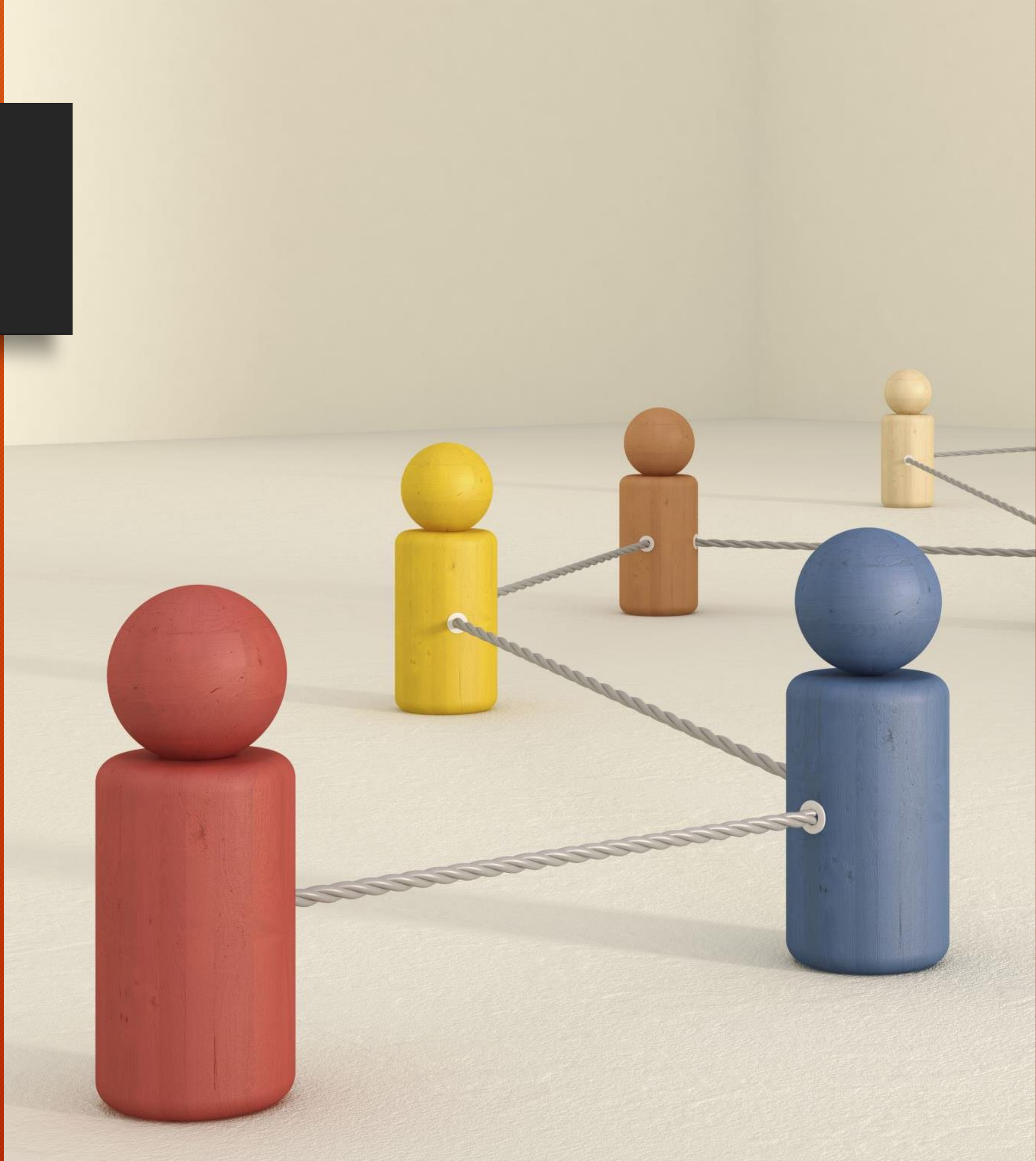
Predictive Modeling: Expanding the model to predict review scores more accurately based on additional features (e.g., user demographics, product categories) could offer more granular insights.

Multilingual Analysis: Extending sentiment analysis capabilities to multiple languages would allow businesses to analyze reviews across different regions, enabling global market insights.

Automated Customer Insights: Using sentiment data to automatically generate reports or alerts for product managers, helping them identify emerging trends or issues faster.

Conclusion

In conclusion, this project demonstrates the power of sentiment analysis and linear regression in understanding customer feedback. By leveraging VADER for sentiment classification and linear regression for model validation, we can gain valuable insights into customer sentiments and their relationship with product reviews. Future advancements in model accuracy and real-time analysis can further enhance decision-making, providing businesses with the tools to stay ahead of market trends.



References

•Liu, B. (2012). *Sentiment analysis and opinion mining*. Morgan & Claypool Publishers.

•Hutto, C. J., & Gilbert, E. E. (2014). VADER: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the Eighth International Conference on Weblogs and Social Media, ICWSM 2014* (pp. 216-225).

•Montgomery, D. C., Peck, E. A., & Vining, G. G. (2012). *Introduction to linear regression analysis* (5th ed.). Wiley.

•Choudhury, P., & Kaur, H. (2020). Role of sentiment analysis in business decision making: A review. *Journal of Business Research*, 112, 1-10. https://doi.org/10.1016/j.jbusres.2019.10.018

•Provost, F., & Fawcett, T. (2013). *Data science for business: What you need to know about data mining and data-analytic thinking*. O'Reilly Media.

•Jha, S., & Kumar, A. (2019). Sentiment analysis of product reviews: A case study on Amazon Fine Food Reviews. *International Journal of Computer Applications*, 178(16), 1-8. https://doi.org/10.5120/ijca2019918797