Summary

Business Case:
Amazon reviews are important to seller because they offer insights about who is buying the product, and what is important for their customers. This information can be used to segment the population of possible customers and better focus the development of future products and sales efforts.

Data:
The data used for this analysis was:

- A collection of reviews written in the Amazon.com marketplace and associated metadata from 1995 until 2015 for the sports category
- Prices from a selected topic: Scraped using Requests in python

What are people talking about?
The identification of topics was done using Latent Dirichlet Allocation (LDA) in python. The data was pre-processed using keras text_to_word_sequence (lower case, filtering out punctuation, tokenization), and NLTK to identify stop words to be filtered out.

The LDA model requires as an input the number of topics to be identified. I tried 20, 10 and 5. Five topics seemed to work well.

The topics evaluation was done by examining the content of the products and reviews grouped together. The goal is to have reviews and products that have communalities within the same topic.

I determined that 5 topics was a good number. After examining the texts, I labeled the topics based on the perceived commonalties as shown below:

- Topic 0: Teams and Gifts
- Topic 1: Airsoft Guns
- Topic 2: Clothes
- Topic 3: Fishing and knives
- Topic 4: Workout routines

I also looked into potential relationships between the reviews content and price of the product reviewed. I didn't find any evidence of a relationship.

I performed this evaluation with the reviews in topic one, airsoft, because the reviews included more information about the product purchased. The pipelines and models evaluated were:

- Speech: Identified part of speech and creating liner model including: average rating, number of helpful votes, total votes, number of verified purchases, percentage of verbs in text, percentage of adjectives, percentage of numeral and cardinal components in the text, percentage of nouns, number of words. I used polynomial features grade 2,

standard scaler and Lasso with cross validation. I evaluated the model using grid search for several alpha values. The r2 for the test and predicting samples was very low <0.05.

- LSA regression: transformed the reviews using TF-IDF vectorizer, and used LSA for dimensionality reduction, the resulting vectors were input in a linear regression model with CV Lasso. The alpha hyperparameter was determined through GridSearch.
- LSA classification: transformed the reviews using TF-IDF vectorizer, and used LSA for dimensionality reduction, the resulting vectors were input into Naïve Bayes, Logistic Regression and Gradient Boosting Classifier to predict if the price was within the following categories: >$100, between $100 and $20, or <$20. The AUC values for the models were similar to randomly guessing.
- Neural Networks: input sequences of tokens to a couple of NN models. The architecture of the first model was:
    - model1 = Sequential()
    - embedding_size = 100
    - model1.add(Embedding(input_dim=num_words,
      output_dim=embedding_size,
      input_length=max_tokens,
      name='layer_embedding'))
    - model1.add(LSTM(32, return_sequences=True))
    - model1.add(LSTM(32))
    - model1.add(Dropout(0.5))
    - model1.add(Dense(1,activation='linear'))
    - model1.compile(loss='mean_squared_error',
      optimizer='Adam',
      metrics=['mae'])

  The architecture for the second NN model was:
    - model2 = Sequential()
    - embedding_size = 100
    - model2.add(Embedding(input_dim=num_words,
      output_dim=embedding_size,
      input_length=max_tokens,
      name='layer_embedding'))
    - model2.add(Dropout(0.2))
    - model2.add(Conv1D(64, 5, activation='relu'))
    - model2.add(MaxPooling1D(pool_size=4))
    - model2.add(LSTM(100))
    - model2.add(Dense(1,activation='linear'))
    - model2.compile(loss='mean_squared_error',
      optimizer='Adam',
      metrics=['mae'])

  The r2 for both models was very low.

- Converted words to Word2Vect. Estimated average vector for each entry, fit linear model a NN model. The R2 square was very low.

I noticed that by looking at the most frequent words in each topic one could infer some characteristics of the review writers. For example, Topic 0: Teams and Gifts, subtopic 1 most frequently used words included Christmas, gift, fan, husband, nfl, football, arrived. Thus, one could infer that a lot of the reviewers writing under this topic bought Christmas gifts in Amazon.com and these gifts were NFL or a team related apparel. The word husband may indicate that a lot of the writers were married and likely females. The mentioning of 'arrived' may indicates that for these reviewers arrival time was important. Therefore, the topics could also be used as a segmentation tool.

Conclusions:
Five distinct topics were identified in the Amazon reviews category sports.

The language of each reviewer was used to identify commonalities and potential reviewers characteristics.

There is no evidence of relationship between product price and reviews.