NAME: Sarah Grobe
DATE: 5/10/2022
Final Project – Technical review
STAT/CS 387

Code and Methods

The first step in this investigation was to conduct an exploratory data analysis, or EDA, on the data. This is available in the file "eda.py," and was conducted on the original data subset of ten thousand reviews. Upon conducting this exploratory analysis, however, I found that it would be necessary to only use a six year subset of that data set due to the relatively low volume of reviews early on in the data. However, subsetting the data even further ultimately resulted in very low sample sizes, particularly since the reviews were further divided into 24 distinct product categories.

I then explored a few different methods to combat this, including grouping the 24 product categories into fewer, broader categories. This is shown in "eda_newData.py." Upon examining the EDA for this approach, though, I felt that it masked too many of the interesting relationships which had seemed to present themselves in the original EDA, and as a result decided that the 24 product categories as were given by the data's metadata should remain. The method which I ended up pursuing was instead using a larger sample of one hundred thousand reviews. This increase in the total number of reviews helped reduce the effects that were originally seen from taking a six year subset of the data, while still allowing enough reviews within each category to show the time series relationships within them without the need for further consolidation into fewer categories. Additionally, while the increase in the total number of reviews from ten thousand to one hundred thousand did obviously have an impact on time and space complexity, I determined it was not significant enough to warrant a smaller data set, and was worth the benefit of the increased sample size. Additionally, data structures were frequently written to and read from files in order to help facilitate a lower time complexity. Note that much of the process to create the subsets, including sampling from the raw data of 83 million reviews, joining the metadata to the reviews themselves, and performing the sentiment analysis, was conducted during the first part of this project during Stat 287, and therefore is outside the scope of this particular project. Instead, the data sets obtained here essentially pick up where that project left off, which is why code for obtaining the subsets and performing the sentiment analysis is not explicitly included.

The file "decomposition.py" is then where the bulk of the analysis occurs, using the statsmodels library to split the number of reviews and compound sentiment into its trend, seasonality, and remainder (or residual). Using functions from this library, these different pieces could be plotted to help with visualization, including evaluating the seasonality for product categories both with and without the trend component. Dozens of graphs were created throughout this process in order to help gain a full appreciation for the relationships within the data and reduce the risk of important relationships being overlooked. The most interesting and/or noteworthy of these plots and results were included in the final scientific article write up.

Finally, a bit of forecasting was attempted in the file “forecasting.py.” This is incomplete and not explicitly mentioned in the article write up. Additionally, as I was decomposing and further analyzing the trends and seasonalities within decomposition.py, I found that to be more interesting and have far more to investigate than I had originally expected. As a result, forecasting became less of a focus than was originally intended. Similarly, due to the amount of information coming from the decomposition, I found that to be a really interesting part of the project in terms of final results, making the forecasting ultimately, in my opinion, less important than was originally expected. For this reason, the forecasting portion of the project was not deeply explored.

Challenges and Project Scope

The main challenge I faced throughout the course of this project was that nearly every piece of it ultimately took far longer for me to do than I had been anticipating; similarly, many parts of the project in terms of results did not end up matching what I had been expecting or hypothesizing, which made it important for me to frequently reevaluate next steps, as well as the project scope as a whole.

In particular, I had not anticipated the initial issue of sample size which arose in the EDA, resulting in the need to investigate a six year subset of the data and ultimately use a new, larger subset altogether. This led to several extra steps evaluating the efficacy of different methods to combat this issue, including the need to go back and investigate code written for the group project last semester, which is where this data set comes from.

Additionally, I needed to do work with the seasonality and decomposition code a good bit in order to familiarize myself with it, since I had never used it before. This resulted in a learning curve that took a bit longer than I expected, especially since I wanted to be sure the results that I was getting from the code were the ones I was looking for and expecting.

As a result of all of this, some changes also occurred to the scope of the project, compared to what I had initially envisioned and proposed. In particular, I had initially expected the decomposition of the time series to have little to know interesting data, since I did not expect the number of reviews or the sentiment of those reviews to have any relationship with time, other than the overall effect of the increased number of reviews as time went on. Therefore, I had anticipated the decomposition portion of the project to be rather lackluster, with more of the focus falling on forecasting, but the end result was actually the opposite. I instead became very interested in the decomposition, which also yielded some surprising results as described in the article. For this reason, the forecasting ended up taking a backseat as the decomposition was investigated further.