Final Project Report

Product Popularity and Sales Analysis

Submitted by Lin Di, Matthew Berryhill

# Motivation

In this digital age, it is common for people to search for a product online and purchase it directly with an online retailer. It is prudent to assume that as products get more popular online, they are also more likely to be purchased by customers. In this project, we will first explore the relationship between popularity of a product and the ultimate buying behavior of that product by customers. Google Trends is used as a measure of product popularity online and Amazon review count as a proxy for the number of purchases made.

We chose to focus on one specific product category, the wireless headsets. This product category gets updated regularly as technology advances. Launches of more feature-rich headsets always generated heated discussion online. Therefore, peaks and troughs of a product’s Google searches can often be observed over time. On the other hand, these Google searches might lead to Amazon purchase easily because Amazon is one leading sales channel for wireless headsets.

After researching on the relationship between Google Trend and Amazon review counts, we delve into other implications of being a popular product. We made use of the review ratings and the review text to associate a product’s popularity with its customer feedback. We then pointed out the most relevant features customers are looking for in wireless headsets. These features can be used by the wireless headset companies to design their future products. By carefully calibrating product feature descriptions with customers’ expectations, they can also improve their product exposures online.

The following questions will be addressed in this project.

1. Does product popularity bring in product sales on Amazon?
2. Are popular products better than less popular ones (based on customer reviews)?
3. What are the salient words and phrases used in the reviews of popular products?

# Data Sources

We selected 9 wireless headsets launched since 2019 for evaluation. They came from the “Best wireless headphones of 2020[[1]](#footnote-1)” list from CNET.com. CNET.com is often returned in the search results when customers make inquiries about electronic gadgets. Thus, it is a good reference to scope popular wireless headsets for our project. After removing two products that were launched much earlier or later than the rest and one product with too few data points, the final list of 9 products is given in Table 1 below.

|  |  |
| --- | --- |
| **Brand** | **Product** |
| Samsung | Galaxy Buds Plus |
| Apple | AirPods Pro |
| Bose | Noise Cancelling Headphones 700 |
| Jabra | Elite 75t |
| Anker | Soundcore Liberty Air 2 |
| Sony | WF-1000XM3 |
| EarFun | Free |
| Anker | Soundcore Life Q20 |
| Anker | Soundcore Liberty 2 Pro |

*Table 1: List of Products*

## Source 1: Amazon Review

The Amazon review data comes from one of our team member’s past work. A web scraping tool is already in place to retrieve historical reviews based on Amazon URLs. The scraped reviews were stored in the Microsoft SQL Server. We used SQL to extract the reviews of our interested products. As accessing this database requires an intranet connection, we decided to export each product’s reviews as individual csv files instead of using a direct database connection. Total number of reviews ranges between 500 to 4,000 for different products. All reviews are cut off on May 2nd, 2020.

## Source 2: Google Trend Index

We typed each product’s exact names (e.g. “AirPods Pro” instead of “AirPods”) as a search term to find its search trend in Google. This minimizes the ambiguity in search results from similarly named products. The United States is selected for location to align with the domestic reviews on Amazon.com. Search category is specified as “Computer & Electronics”. The trend data is aggregated on a weekly basis with the date period specified from January 1st, 2019 to May 2nd, 2020. In total, there are 69 data points for each product as there are 69 weeks in this specified period. The trend data was downloaded as individual csv files from the website. They were then consolidated into a single Excel file with tabs named after each product.

The two data sources used in this project are summarized in Table 2 below.

|  |  |  |
| --- | --- | --- |
|  | **Source 1** | **Source 2** |
| **Location** | Microsoft SQL Server | [trends.google.com](https://trends.google.com/trends/?geo=US) |
| **Access Method** | SQL | Direct Download |
| **Format** | CSV | Excel |
| **Data Fields** | Manufacturer, Product Name, Review Rating, Verified Purchase, Review Date, Review Title | Week, Index |
| **Number of Records** | Varied between 500 to 4,000 reviews by products | 69 records per product |
| **Time Period** | From Jan 1st, 2019 to May 2nd, 2020 | From Jan 1st, 2019 to May 2nd, 2020 |

*Table 2: Summary of Data Sources*

# Data Manipulation

## Manipulating Amazon Review Data

We used Pandas to load the csv files for Amazon review data. These reviews were filtered by the “Verified Purchase” flag as we are only interested in the actual purchase actions. The review date column was converted to Pandas datetime format before we grouped them by week. Weekly average review rating and review count were calculated. To match the week definition in Google Trend, review weeks start on and are labelled by Sunday’s date.



## Manipulating Google Trend Data

Google Trend data for each product was loaded in Pandas by the sheet names from the consolidated Excel file. The week start dates were converted to Pandas datetime format. The trend index column was renamed for consistency across products. Since product searches do not all start from Jan 1st, 2019, we filtered out records before the first search was made.



## Data Merge

The two data sources were merged based on Amazon review week and Google Trend week. Since Google Trend always has a wider date range than Amazon reviews, our merge was based on Google Trend week (right join). Product names and selling prices were added as new columns to the merged DataFrame. DataFrames for different products were then concatenated vertically into a single DataFrame for subsequent analysis.



## Additional Data Manipulation Before Analysis and Visualization

* **Correlation Calculation**

Correlation between Amazon review count and Google trend index was calculated by Pandas [*corr*](https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.corr.html) method. In addition, we iteratively shifted the Google trend index column to find the correlation at different time lags. The maximum shifting period was set to be half of the size of total data points. This ensures sufficient coverage of the correlation patterns if one exists.



* **Price Segmentation**

The concatenated DataFrame was segmented by product pricing. Four bins were created for prices in range [0, 100), [100, 150), [150, 200) and [200,).



* **Review Count Normalization**

Since review counts vary considerably from one product to the other, we normalized the review counts of each product. This also aligns with the range of Google trend index (0 to 100).



* **Extract Adjectives from Text**

We used the NLTK package to extract the part-of-speech tags for all words in reviews. We then filtered for only the adjectives (‘JJ’) for our next analysis step.



* **Unique Adjectives from Positive Reviews**

We separated positive (5-star) and negative (1-star) reviews and extracted the adjectives used in them. Taking the intersection between the two, we obtained the adjectives used in both positive and negative reviews. These common adjectives were excluded to obtain the unique adjectives used in positive reviews.



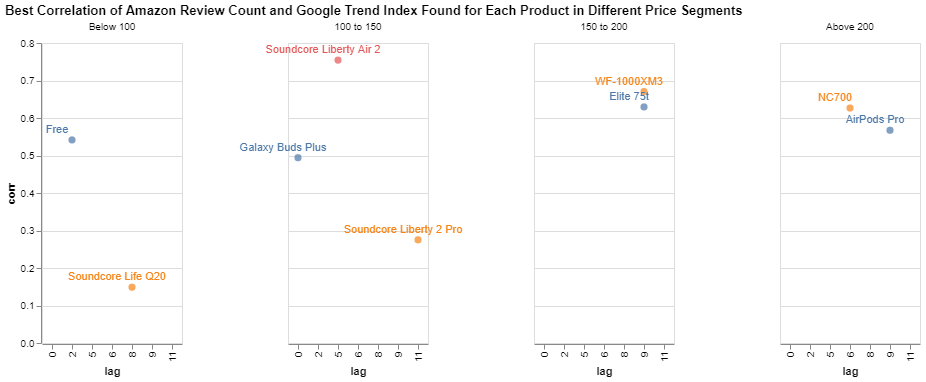
* **Word Pairs**

A word pair is defined by two consecutive words in sentences, after stemming words into root form and removing stop words. We evaluated all pairs of words by traversing the review text. The order of words does not matter, so we applied a set structure for the word pairs.



# Analysis and Visualization

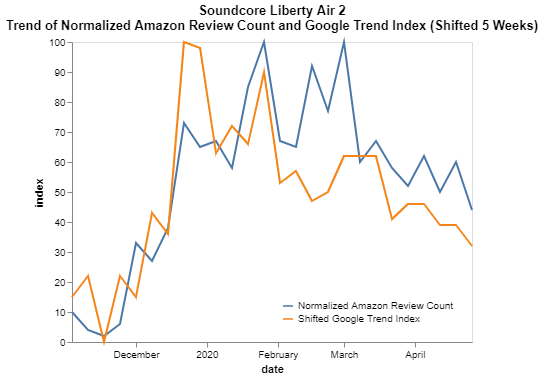
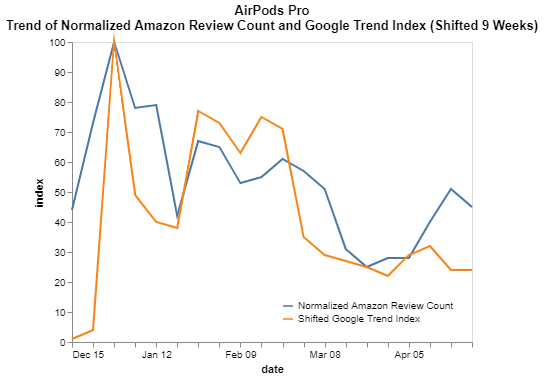
The first visualization we had created is a scatter plot in Altair. It shows the best correlation found for each product and the lag period when such correlation occurs. Furthermore, the scatter plot is faceted with price segments. For more expensive wireless headsets (typically greater than $150), a good correlation between Google trend index and Amazon review count can be found after shifting the Google trend index by 6 to 9 weeks. This time lag can be explained by the time taken for customers to complete their purchase and post reviews online after searching in Google. Apple AirPods Pro has a lower correlation coefficient than the other products in this price range. This is probably because Apple is so effective at marketing its product that customers can skip the search step before making purchases. On the other hand, correlations are not so prominent for products priced less than $150. Most of correlation coefficients are less than 0.6 at best. The only exception is Anker Soundcore Liberty Air 2 which exhibits a high correlation at 5-week time lag. One conclusion we can draw from this visualization is that customers tend to research online before purchasing if they are interested in a wireless headset priced more than $150.



The Altair code used to generate this scatter plot is referenced here.



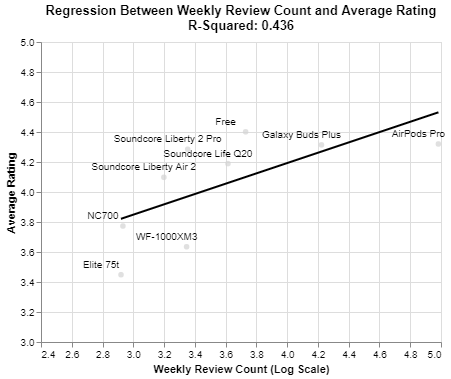
Two examples of time series graphs are drawn below to demonstrate the correlation between Google trend index and Amazon review count. Peaks and troughs of the two occur mostly in the same week (after shift). The upwards and downwards trends also match closely. It is evident that Google searches and Amazon purchases are correlated.



The Altair code used to generate the time series graph for each product is referenced here.



We further analyzed the relationship between product popularity and customer ratings. Since products were launched at different dates, directly comparing their popularities over time is not possible. However, from previous analysis, we know that Amazon review count can be considered as a measure of a product’s popularity. We plotted the weekly review counts against average ratings for each product below. The weekly review counts were converted to log scale.



An r-square value of 0.436 shows a moderate positive relationship between the review count and average rating, i.e. popular products are rated higher than less popular ones. The explanation can be two-fold. As a product gets more popular, it is also rated higher by customers. Alternatively, highly rated products are more likely to attract customers and get more popular over time.

The Altair code used to generate scatter plot and linear regression line is referenced here.

****

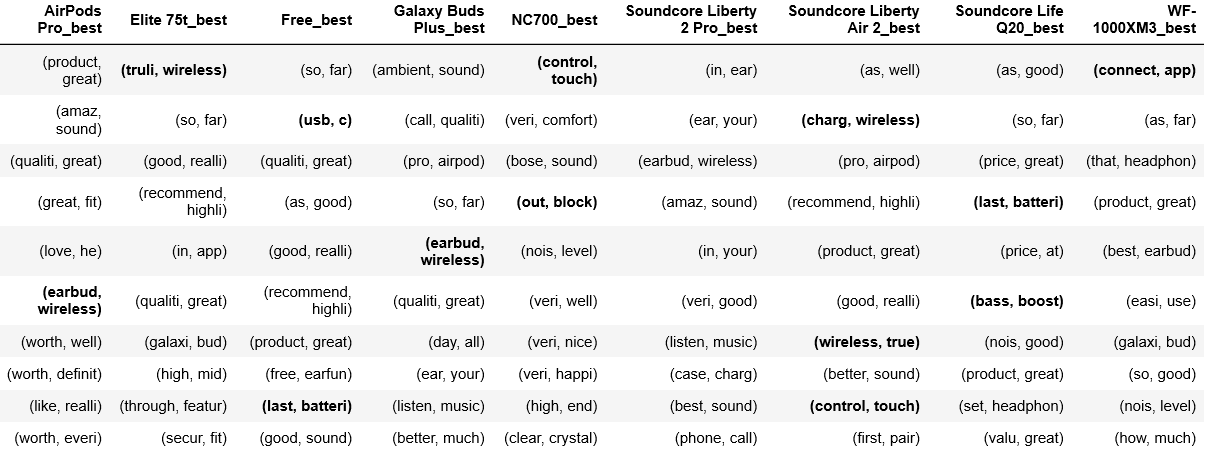
Next, we analyzed the review text to understand what is considered a good wireless headset by customers. We extracted the unique adjectives from 5-star reviews and generated a word cloud based on their frequencies. Lightweight, durable, and affordable are some of the most common words used to describe the highest rated products. They can be used as design criteria for wireless headset companies to measure in their future products.



Below code snippet demonstrates the word cloud generation with top 40 adjectives.



We moved beyond single words and looked at the word pairs from 5-star reviews. From these word pairs, we can see a few useful combinations such as "usb c", "connect app", "wireless charging", “battery life” and “touch control”. They signify product features customers are looking for in these wireless headsets. Companies can emphasize more on these “buzzwords” in the product description to make sure their products appear relevant to customers. By doing so, they can potentially boost the hit rate from Google searches because these are the words customers would be using.



This table was output from Pandas DataFrame by filtering the 10 most frequent word pairs for each product.



# Discussion and Conclusion

In this project, we explored the relationship between Google search and Amazon purchases by sampling 9 wireless headsets. The result shows that most customers will perform a Google search before purchase if the product is more than $150. The search trend and purchase trend are closely correlated after shifting Google trend index by 6 to 9 weeks. A popular product not only brings more sales, but also is rated higher. A positive regression line fits the weekly review count and average rating points with a moderate r-square value.

By analyzing the text of 5-star reviews, we generated a list of adjectives describing good wireless headsets. They should be considered for future product designs. In addition, the top word pairs for each product shows different features customers are interested in. These features can be referenced in product descriptions if companies want to boost their product exposure in search results.

Future research can extend this work to other product categories. This project relies on Pandas as the main data manipulation tool. Others can consider more efficient data processing tool (PySpark for example) if larger text corpus or longer data period is included. Lastly, our derivation of customer opinion is strictly based on review stars. More nuanced sentiment analysis can be applied to better segregate positive and negative feedback.

# Statement of Work

At the data extraction stage, Lin Di was responsible for downloading Google Trends data and exporting Amazon review data from Microsoft SQL Server.

At data manipulation and data visualization stages, Lin Di contributed to the correlation analysis while Matthew focused on the text analysis.

Project reports were written by Lin Di. Matthew performed some modifications as needed.

Matthew set up the Git repository and oversaw the version control of source codes used in this project. Other administrative tasks were carried out by Lin Di.

1. Carnoy, D. (2020, April 12). The best wireless headphones of 2020. Retrieved April 14, 2020, from https://www.cnet.com/news/best-wireless-headphones-of-2020/ [↑](#footnote-ref-1)