Lucas Monnin, Luca Rizzello, Bojan Petrovski
29 January 2017

Analysis of Swiss products based on Amazon reviews data

Are Swiss products better than the average product?

Introduction

The goal of this project is using data analysis tools on a large dataset of Amazon reviews and examine if Swiss products are better than the average product on Amazon. It is important to note several assumptions and biases that are present in our data. Amazon does not reveal the country of Origin for products, so this limits us to using the brand name. While for some Swiss brands like the Watch companies we can be fairly certain that the products were made in Switzerland, for others like Logitech that is rarely the case. So when we say Swiss product we mean a product by a Swiss brand. Another bias we need to be aware of is that Amazon is not an official Swiss Watch retailer, but is what is called an "Gary Market". Watches sold on Amazon are real but in general, don't have a manufacturers warranty. This may lead to very unfavourable reviews that are not indicative of the Product, so we will try some techniques to deal with this.

Exploratory analysis

The exploratory analysis of the Swiss products is where we got our first surprise. The distribution of reviews does not follow a normal distribution. The histogram shows, in fact, a very large bias towards 5-star reviews. More than 50% of reviews are 5-star reviews, a fact which is consistent in all samples. This could indicate that either consumer is very satisfied with Amazon Products, or it could be some Bias in the data due to Human psychology, or possibly paid fake reviews. The good thing is that the trend is present in all data samples that we made so it should no interfere with our tests.

Unsurprisingly, the most exported goods are watches, knives and electronics (note that Wenger does not exist anymore, bought by Victorinox). This latter category of products is

mostly shared between Logitech and Garmin (US company based in Switzerland...), but interestingly they seem to have a slightly lower rating mean compared to watches companies.

The difference is not significant but one may conclude that the rating is correlated to the manufacturing location.

Another observation is that electronics products tend to be more reviewed than the others (fancy watches, but one may assume that people who buy these kinds of watches are more likely not to buy them online).

And finally some numbers: in these 289'878 reviews, there are 4718 different products for 48 Swiss companies, 3 of them having 51% of the Swiss products: Victorinox with 1056 products, Garmin with 780 and Logitech with 595.

Comparison of Reviews

In the first comparison, we directly compare the set of Swiss reviews with 4 other sets: a random control set, French set, German set and US set. All datasets we generated using the same spark script and Using the Wikipedia listings for the respective country brand. The random control set was generated using the spark Bernoulli sampling.

We use the Mann–Whitney U test instead of the more frequently used t-test since we can not say with certainty that the samples adhere to a normal distribution. Furthermore, the Mann–Whitney U test for large samples asymptotically approaches the t-test if the distribution is normal so it's a better choice in our case.

We can not find a statistically significant difference between the Swiss products and the control, but we can reject the null hypothesis in the other tests. For German and US products it implies that there is a significant chance that a Swiss product chosen at random will have a higher rating than a German or US product chosen at random. For French and Swiss products the reverse is true.

Comparison after grouping by product

On of the possible problems with our previous tests is that some very popular products with lot's of great reviews can skew the results so it is very difficult to generalise results while using individual results. That is why we aggregated the reviews by product.

After aggregation, we now have a sample of some continuous distributions so we can use the KS test to check if the distributions are different. In all cases, we discovered that the distributions are different. After that we performed the Mann–Whitney U test and this time

we could not find any statistical differences between the Swiss, German and US products, the only test that showed a statistically significant result was between Swiss and French products, the result being that a Swiss product chosen at random will have a smaller rating than a French product chosen at random.

Comparison after grouping by product using weight

The idea behind this experiment is to try to eliminate some effects of fake reviews, or reviews that do not address the product, so we use the helpfulness as a weight when we do the average of each product.

Afterwards, we again perform our tests and this time we find two significant results, both German and French appear better than Swiss, while there is no difference between US and Swiss Products

Comparison after grouping by product using weight

The idea behind this experiment is to try to eliminate some effects of fake reviews, or reviews that do not address the product, so we use the helpfulness as a weight when we do the average of each product.

Afterwards, we again perform our tests and this time we find two significant results, both German and French appear better than Swiss, while there is no difference between US and Swiss Products

Comparison after grouping by category

One note about the mean by category: it is actually computed from the reviews and not the products (very small difference). An interesting observation is that clothing, shoes & jewellery and Accessories are dominant in terms of products but are not as reviewed as Electronics for example.

One may conclude that people buying 'tech stuff' are more likely to review their purchase, or fashion stuff is less likely to be reviewed as most of the time the product is not far from what is being depicted online (which is not true for most electronic devices).

Overall, it seems like Switzerland is mostly above when it comes to producing electronic devices, equipment for outdoor activities and watches.

Comparison of Sentiment Data

One additional metric that we can use is Sentiment Analysis on the Text of the reviews. We Analysed all reviews and even found out that there is some correlation between the Sentiment score of the review and the number of Starts awarded. While this is not a hard metric, it may give us an indication of the attitude of Consumers towards the different products.

After analysing the data we found out the sentiment score of Swiss Reviews is greater that the scores of US, French and German Reviews. The Mann–Whitney U test showed statistical significances in all three tests. The mean of the Swiss reviews was 64.7 while the US, French and German Reviews had means of 59.1, 61.8 and 61.3 respectively.

Additionally we grouped the data by product and in that case, we can only show that Swiss products have a greater score than German and US products but a worse score then French Product.

Conclusion

Or goal was to try multiple different methodologies and to see if there is some general pattern that would be present in all the experiments. While we did find some statistically significant results there was no pattern and in fact, sometimes the results flip deepening on how you group the data. Furthermore, even when there is a statistically significant difference the difference is generally small and only 1-3%, and can easily be due to our sampling.