In this module, we had the opportunity to analyze customers reviews based on the Amazon product of choice. Amazon has Simple Storage Service (S3) buckets with information that is extractable and downloadable to end users. This link has over 50 review types to choose from <https://s3.amazonaws.com/amazon-reviews-pds/tsv/index.txt>. In addition, this site provides a breakdown of the data columns, links to each review type and the file format (tsv/gz). The tsv is a tab-separated values file and gz is GNU zipped file. The columns in these reviews are as follows:

|  |  |
| --- | --- |
| **Data Columns** | **Description** |
| marketplace | 2 letter country code of the marketplace where the review was written. |
| customer\_id | Random identifier that can be used to aggregate reviews written by a single author. |
| review\_id | The unique ID of the review. |
| product\_id | The unique Product ID the review pertains to. In the multilingual dataset the reviews for the same product in different countries can be grouped by the same product\_id. |
| product\_parent | Random identifier that can be used to aggregate reviews for the same product. |
| product\_title | Title of the product. |
| product\_category | Broad product category that can be used to group reviews (also used to group the dataset into coherent parts). |
| star\_rating | The 1-5 star rating of the review. |
| helpful\_votes | Number of helpful votes. |
| total\_votes | Number of total votes the review received. |
| vine | Review was written as part of the Vine program. |
| verified\_purchase | The review is on a verified purchase. |
| review\_headline | The title of the review. |
| review\_body | The review text. |
| review\_date | The date the review was written. |

In addition to the listed columns, we created a customer\_count column for the analysis. This report will include information about this Big Data analysis project.

**Project Tools**

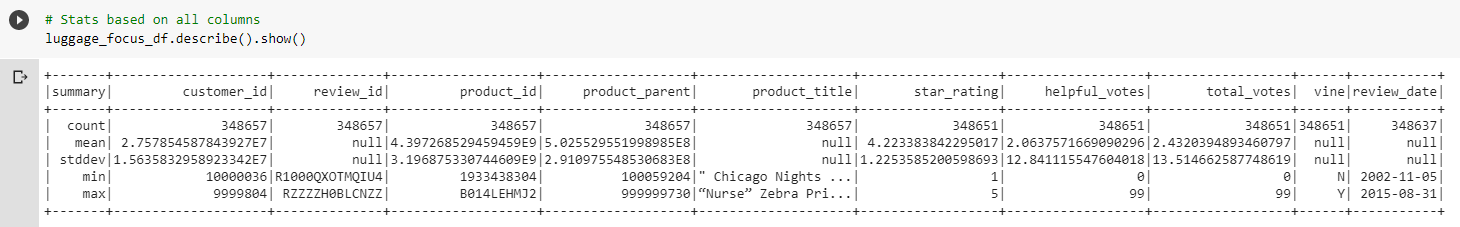
* Colab
* PySpark
* PgAdmin
* AWS S3 & RDS
* ETL

In addition to these tools, the following module lessons were revisited to shape and complete the challenge; 16.4.2, 16.4.3 and 16.9.1. These lessons have a significant amount of information, useful for understanding the process.

Data Analysis

For my project, I have selected the reviews on Luggage <https://s3.amazonaws.com/amazon-reviews-pds/tsv/amazon_reviews_us_Luggage_v1_00.tsv.gz>. This is a topic of interest for me, as I have replaced my luggage 5 times over the past 3 years due to work travel. Since travel has come to a halt, I may get a little more use from my current set. However, I was curious to see what the market had to say about luggage.

Time Frame

The reviews in the dataset are from November 5, 2002 – August 31, 2015. During that 13-year time period, there were 348,657 reviews conducted. Below is a statistical summary of this information, along with a link to the Colab workbook <https://colab.research.google.com/drive/1DDJtsr-pEwqkLpGsm9o42p5Z3N-69j2Z?usp=sharing>.

Vine & Star Reviews/Ratings

There were 904 vine reviews with “Y”, 6 that were “null” and 347,747 that are “N”. This is less than 1% of the dataset population, amounting to .003%.

By contrast, there are more 5-star ratings in this population. The total is 216,495 which is 62% of the total review.

|  |  |  |
| --- | --- | --- |
| **Star Rating** | **Percentage of 348,657** | **Total Reviews** |
| 5 | 62% | 216,495 |
| 4 | 18% | 61,431 |
| 3 | 8% | 27,846 |
| 2 | 5% | 17,871 |
| 1 | 7% | 25,008 |
| null | 0% | 6 |

As listed in the Colab notebook, the mean for the star ratings is 4.22 with a standard deviation of 1.23. The vine reviews were non-calculable due to the low number. The number of times the helpful votes were selected was 99.

Conclusion

Vine reviews are not completely trustworthy. The amount of information we obtain from all the review is not large enough to glean a significant amount of information from the review. In addition, reviews are difficult to base our decisions. Even though a lot of us use review data to determine whether we will buy a product, there are biases in the review process. Those biases included but are not limited to;

* Product satisfaction
* Valuation of the product
* Customer service
* Durability
* Type of reviewer (stern, rigid, kind, generous etc.)
* Warranty
* Ease of product use
* Return and refund policy

Also, additional analysis should be done on the product type and the reviews for that product. Since the subject is luggage, we’re looking at data that supports all brands and types of luggage. This does not factor in the size, sets of luggage, style or costs. It’s possible that we may have a different understanding of these reviews if we analyzed the data using some of the variables.

I also looked at the helpful votes data (see table below). The greater number of customers had no “helpful\_votes listed by their id. Then, the number of times someone said a review was helpful, fell in the 1-36x category.

Recommendations

I would recommend that Amazon has a column that provides specific info on the brand of the luggage. If that were the case, we could analyze this data a lot more uniformly. The current breakdown seem to be specific to an item type and is difficult to just see brand.

I enjoyed looking at the data, I hope it is updated to provide information for the past five years. As much has happened with brands, styles, airport requirements and airport technology altogether. It would be good to look at more recent data and see what we can learn from the reviews.

Finally, we could better analyze the data if the S3 bucket provided a more clearly defined explanation of the contents of the column name.

