Big Data Assignment

Data sets: Shoes and Apparel

Statistical analysis to determine whether reviews from Amazon's Vine program are trustworthy.

Findings

- The Top 5 Customers with more reviews were Non-Vine Customers. Of these 5, four can be found in the apparel data set and one in the shoes data set.

- The Top 5 Products with more reviews were reviewed by Non-Vine Customers.

- The percentage of Vine Customer Reviews (0.02%) is very small compared with the percentage of Non-Vine Customer Reviews (99.97%), so the next findings are expressed in percentages to be able to compare the results in some way.

- In the shoes data set, the percentage of reviews with 4- and 5-star rating are slightly higher for Vine Customer than for Non-Vine Customer (80% vs. 76%). Also, in the apparel data set, the percentage of reviews with 4- and 5-star rating are slightly higher for Vine Customer than for Non-Vine Customer (88% vs. 83%).

- On average, the Top 5 Customers with more reviews in the Non-Vine Program rated the 83.1% of their reviews with 4 and 5 stars, while the Top 5 Customers with more reviews in the Vine Program rated the 78.4% of their reviews with 4 and 5 stars.

- On average, the Top 5 Products with more reviews in the Non-Vine Program rated the 83.2% of their reviews with 4 and 5 stars, while the Top 5 Products with more reviews in the Vine Program rated the 93% of their reviews with 4 and 5 stars.

- On average, in five shoe products that have both types of reviews, the Non-Vine Customers rated the 71% of their reviews with 4 and 5 stars, while the Vine Customers rated the 81% of their reviews with 4 and 5 stars.

- On average, in five apparel products that have both types of reviews, the Non-Vine Customers rated the 90% of their reviews with 4 and 5 stars, while the Vine Customers rated the 86% of their reviews with 4 and 5 stars.

- On average, in the Top 5 reviews with more votes for shoe products, the 94% of the total votes, of the reviews made by Non-Vine Customers, were classified as helpful, while the 95% of the total votes, of the reviews made by Vine Customers, were classified as helpful.

- On average, in the Top 5 reviews with more votes for apparel products, the 96% of the total votes, of the reviews made by Non-Vine Customers, were classified as helpful, while the 94% of the total votes, of the reviews made by Vine Customers, were classified as helpful.

Conclusion

For the data sets chosen, based on the amount of reviews made by Vine Customers and the results listed in the findings, there is no significant differences that pointed out that the reviews from Amazon's Vine program are not trustworthy, but we always must consider that it is part of the human condition to have implicit biases and it could be included when expressing opinions of a product.

Process

1. Calculate totals

Totals	
Reviews	**10,273,249**
Shoes	4,366,916
Apparel	5,906,322

Unique customers	**5,182,889**

Unique products	**4,206,294**

Top 5 Customers	
Customer id	**# reviews**
50612720	624
33883540	428
45547332	388
37474421	362
33924372	344

Top 5 Products	
Product	**# reviews**
B004M6XUI2 - RFID Blocking Men's Leather Classic Bifold Wallet Black	1,834
B004M6UDF0 - Alpine Swiss Mens Wallet Leather Money Clip Thin Slim Front Pocket Wallet	1,762
B006PGGJOE - SHARKK® Aluminum Wallet Credit Card Holder	1,495
B0045H0L1W - LED Light	1,427
B004M6UD46 -Men's Leather Wallet Euro Traveler Extra Capacity Bifold	1,395

2. Calculate the composition of the reviews

Reviews for Shoes by Star rating				
Type of review	**Star rating**	**# reviews**	**% by type of review and star rating**	**% by type of review**
Non-Vine customer		**4,366,021**		**99.98**
	1	232,170	5	
	2	242,807	6	
	3	404,107	9	
	4	847,464	19	
	5	2,639,473	60	
Vine customer		**895**		**0.02**
	1	5	1	
	2	25	3	
	3	75	8	
	4	328	37	
	5	462	52	

Reviews for Apparel by Star rating				
Type of review	**Star rating**	**# reviews**	**% by type of review and star rating**	**% by type of review**
Non-Vine customer		**5,903,986**		**99.96**
	1	445,430	8	
	2	369,514	6	
	3	623,196	11	
	4	1,146,396	19	
	5	3,319,450	56	
Vine customer		**2,336**		**0.04**
	1	26	1	
	2	87	4	
	3	275	12	
	4	841	36	
	5	1,107	47	

3. **Calculate the composition of the star ratings**

Top 5 Customers with Vine reviews			
Data set	**Customer id**	**# reviews**	**% reviews 4- and 5-star rating**
Shoes	43698610	5	100
Shoes	40581989	4	75
Shoes	48156368	4	100
Shoes	52228204	4	100
Shoes	51070985	4	75
Apparel	52188216	34	50
Apparel	36983626	31	81
Apparel	13814078	25	76
Apparel	49620639	24	54
Apparel	49598970	15	73

Top 5 Customers with Non-vine reviews

Data set	Customer id	# reviews	% reviews 4- and 5-star rating
Shoes	45547332	210	43
Shoes	2761437	196	100
Shoes	52433525	171	57
Shoes	12228192	161	100
Shoes	20872710	152	88
Apparel	50612720	559	93
Apparel	33883540	351	100
Apparel	33924372	344	100
Apparel	37474421	282	85
Apparel	15006109	263	65

Top 5 Products with Vine reviews

Data set	Product id	# reviews	% reviews 4- and 5-star rating
Shoes	B00SM2LSQ8	27	89
Shoes	B0018KYMNW	23	100
Shoes	B0018KYOVW	23	78
Shoes	B00LV4D1X2	13	85
Shoes	B00M42W4XS	11	91
Apparel	B002BFLJ70	30	87
Apparel	B00FXPRJWO	22	100
Apparel	B004OA7QVI	22	100
Apparel	B004OA7QYA	21	100
Apparel	B004OA7QT0	21	100

Top 5 Products with Non-vine reviews

Data set	Product id	# reviews	% reviews 4- and 5-star rating
Shoes	B00H9RZDRM	1,250	96
Shoes	B002L9AL84	1,113	85
Shoes	B004M6W4FW	891	87
Shoes	B004RR0N8Q	801	95
Shoes	B001UQ71G4	786	85
Apparel	B004M6XUI2	1,834	84
Apparel	B004M6UDF0	1,762	82
Apparel	B006PGGJOE	1,495	66
Apparel	B0045H0L1W	1,427	65
Apparel	B004M6UD46	1,395	87

4. **Calculate the composition of star ratings for products that have both types of reviews**

Top 5 Products with both reviews for Shoes				
Product id	**Type of review**	**% reviews 4- and 5-star rating**	**Type of review**	**% reviews 4- and 5-star rating**
B00NHUW1UW	Non-vine	78	Vine	80
B0018KYMNW	Non-vine	75	Vine	100
B00NHUVF18	Non-vine	61	Vine	60
B00NHUVFGS	Non-vine	73	Vine	83
B00NHUVTRS	Non-vine	67	Vine	80

Top 5 Products with both reviews for Apparel				
Product id	**Type of review**	**% reviews 4- and 5-star rating**	**Type of review**	**% reviews 4- and 5-star rating**
B002BFLJ70	Non-vine	84	Vine	87
B00BGIQPSG	Non-vine	87	Vine	64
B00BGIQR3E	Non-vine	83	Vine	80
B004OA7QVI	Non-vine	98	Vine	100
B004OA7QT0	Non-vine	100	Vine	100

5. **Calculate the composition of votes for both types of reviews**

Top 5 Reviews by total votes for Shoes					
Review id	**Type of review**	**Star rating**	**Helpful votes**	**Total votes**	**% Helpful**
R11XKHFS4KQS3Z	Vine	4	205	211	97%
R2MPEQ4SPTEQNS	Vine	4	180	184	98%
R1SPWJDHUWWC5E	Vine	5	88	98	90%
R3KOK2SH39BZU1	Vine	3	94	96	98%
R2XRYNV2SY3ZKL	Vine	5	53	56	95%
R3DSCOKAHD7WIT	Non-vine	5	5,070	5,329	95%
R236QGQ8RZO1WC	Non-vine	5	4,646	4,815	96%
RD6DLEJLLTOSK	Non-vine	5	3,828	4,150	92%
RRO1L0B8YB0ZP	Non-vine	5	2,726	2,927	93%
R18UGASH7JSUFF	Non-vine	5	2,571	2,718	95%

Top 5 Reviews by total votes for Apparel					
Review id	**Type of review**	**Star rating**	**Helpful votes**	**Total votes**	**% Helpful**
R30QE1QK86LPYL	Vine	5	240	248	97%
R3TKG664L9MTXJ	Vine	4	164	175	94%
R6U9701C3BGO6	Vine	3	139	147	95%
R17EPR3LT1T6OW	Vine	5	95	102	93%
R2VUXJT91MXOQJ	Vine	3	72	79	91%
R2XKMDXZHQ26YX	Non-vine	5	41,278	41,889	99%
R16DWLI3AVB432	Non-vine	5	11,350	11,728	97%
RYRFJTR97GGJV	Non-vine	5	11,219	11,555	97%
R29Z83O4AK10UD	Non-vine	5	6,894	7,615	91%
R2XKMDXZHQ26YX	Non-vine	5	5,341	5,402	99%