

## Amazon Vine Analysis

The Amazon Vine program is a service that allows manufacturers and publishers to receive reviews for their products. Companies like SellBy pay a small fee to Amazon and provide products to Amazon Vine members, who are then required to publish a review. In this project, we will study a dataset holding reviews for shoes by completing the following steps:

- Use PySpark to perform the ETL process to extract the dataset, transform the data, connect to an AWS RDS instance, and load the transformed data into pgAdmin.
- Use PySpark to determine if there is any bias toward favorable reviews from Vine members in the dataset.
- Write a summary of the analysis to submit to the SellBy stakeholders.

### Objective:

- In this project we will analyze Amazon reviews written by members of the paid Amazon Vine program to determine if there is any bias.

### Resources:

- Data Source: Amazon Shoes Reviews
- Software: spark-3.2.3 - pgAdmin 4 - SQL - Google Colab Notebook - Amazon Web Services
- Scripts: Amazon\_Reviews\_ETL.ipynb, Vine\_Review\_Analysis.ipynb, SchemaTesting.sql

### Analysis of Data:

#### Perform ETL on Amazon Product Reviews

Using cloud ETL process, we have created an AWS RDS database with tables in pgAdmin, picked the shoes reviews dataset from Amazon Review Dataset and extracted the dataset in a DataFrame. After that we have transformed this DataFrame into 4 separate DataFrames that match the table schema in pgAdmin. Finally, we have uploaded the transformed data into the appropriate tables and ran queries in pgAdmin to confirm that the data has been uploaded

Table 1 - The customers \_table DataFrame:



	<b>customer_id</b> [PK] integer 	<b>customer_count</b> integer 
1	16121565	5
2	48146680	1
3	11062912	1
4	51451778	1
5	27920838	1
6	4919528	3
7	47802851	2
8	42560427	1
9	29467780	1
10	49703087	12

Table 2 - The products\_table DataFrame:



	<b>product_id</b> [PK] text 	<b>product_title</b> text 
1	B001CJL5ES	L.B. Evans M...
2	B002MCVH...	New Balanc...
3	B00UXBFW...	Kenox Sling ...
4	B002CMM4...	Capezio Wo...
5	B00HNO32...	Timberland ...
6	B00C6BQQ...	New Balanc...
7	B00M0NRE...	New Balanc...
8	B00ZC4VIKK	Donalworld ...
9	B00291CGH8	Justin Boots...
10	B000P6GK...	Nike Men's R...

Table 3 - The review\_id\_table DataFrame:

	review_id [PK] text	customer_id integer	product_id text	product_parent integer	review_date date
1	R3P2HIOQCIN5ZU	18069663	B000XB31...	265024781	2015-08-31
2	R12VVR0WH5Q2...	16251825	B00CFYZH...	259035853	2015-08-31
3	RNCKKB6TV5EEF	20381037	B00S8JNN...	666066660	2015-08-31
4	R2NZXYIVCGB13W	108364	B00XFBPO...	448483263	2015-08-31
5	R2EQ1TG9IT30EQ	45449350	B00SW64...	7853171	2015-08-31
6	R1WXA9JSC2H1...	19324665	B011F9E6LI	14311457	2015-08-31
7	R12ENYLFGGNW...	50073594	B00HAUP...	264821602	2015-08-31
8	R2R07E5PNXEUO5	21706057	B00L1RKO...	767118055	2015-08-31
9	R27BA52AKWM...	13708216	B005WA9...	813856438	2015-08-31
10	RLF8DOID2KD5O	40542649	B00BEYQI...	661491213	2015-08-31

Table 4 - The vine\_table DataFrame:

	review_id [PK] text	star_rating integer	helpful_votes integer	total_votes integer	vine text	verified_purchase text
1	R3P2HIO...	1	0	0	N	Y
2	R12VVR0...	5	0	0	N	Y
3	RNCKKB6...	4	0	0	N	Y
4	R2NZXYI...	5	0	6	N	Y
5	R2EQ1TG...	3	0	0	N	Y
6	R1WXA9J...	5	1	1	N	Y
7	R12ENYL...	5	1	1	N	Y
8	R2R07E5...	4	0	0	N	Y
9	R27BA52...	5	0	0	N	Y
10	RLF8DOI...	3	0	0	N	Y

Determine Bias of Vine Reviews:

Using PySpark we have extracted the same dataset in a new Google Colab Notebook Vine\_Review\_Analysis.ipynb . We have recreated the vine\_table (Table 4) and we have performed the following steps:

- Created a new DataFrame to retrieve all the rows where the total\_votes count is equal to or greater than 20 to pick reviews that are more likely to be helpful and to avoid having division by zero errors.
- Filtered the new DataFrame and created a new DataFrame to retrieve all the rows where the number of helpful\_votes divided by total\_votes is equal to or greater than 50%.
- Filtered the later DataFrame and created a new DataFrame or table that retrieves all the rows where a review was written as part of the Vine program/paid (Table 5), and another one where the review was not part of the Vine program/unpaid (Table 6)

Table 5 - Paid Reviews DataFrame

review_id	star_rating	helpful_votes	total_votes	vine	verified_purchase
R2N45ZKRRZS856	5	21	22	Y	N
R5OMLMK13A8NS	5	34	38	Y	N
R2MPEQ4SPTEQNS	4	180	184	Y	N
R1R0D3KJ0CQ31	4	21	21	Y	N
R1SPWJDHWWC5E	5	88	98	Y	N
R1X6M5XA3FT98W	5	24	26	Y	N
R37VCW6HA0Z72T	5	27	28	Y	N
R2XRYNV2SY3ZKL	5	53	56	Y	N
R1Y93KWKAX1P5N	2	26	31	Y	N
R2QVTDYLYTP8SL	5	21	24	Y	N
R3KOK2SH39BZU1	3	94	96	Y	N
R2VOM73EHLPXJW	4	35	38	Y	N
R3SEZS7BZEC69Y	5	16	20	Y	N
R1MJ5J272V19O6	4	49	51	Y	N
R3A7BQX1JDKOM5	4	20	20	Y	N
R1307JMPUEQXOW	5	31	35	Y	N
RXV0SDXE5B15T	2	39	42	Y	N
RBMDYE7LUH9FI	5	26	32	Y	N
R1N4W961QV59BV	5	25	34	Y	N
R11XKHFS4KQS3Z	4	205	211	Y	N

only showing top 20 rows

Table 6 – Unpaid Reviews DataFrame

review_id	star_rating	helpful_votes	total_votes	vine	verified_purchase
R37F42INKX7L9K	5	45	49	N	Y
R2EHKYNEP8WVSR	5	25	25	N	Y
RXOS7BHID0UHL	5	16	27	N	N
R33HHGFPB403GM	5	19	21	N	Y
RY9O9XNLP464N	2	19	22	N	Y
R2VP11C28JAEZF	5	30	30	N	Y
R1TXGR1HAZM3P3	5	28	29	N	Y
R6OD85TMEHQO	5	28	28	N	Y
R1G4PAJXP3FTN7	2	43	51	N	Y
R2P2S8UGUMCOLX	5	21	22	N	Y
RBTQKXGJ9RP7C	1	33	33	N	Y
R342L93CXGDZO5	2	73	76	N	Y
R31Z97KFD4HMNF	3	23	26	N	Y
R2NDB7EQLYK87A	5	44	46	N	Y
R1C0W0OSBCVXC2	1	45	46	N	Y
R154Q41ID7FTFC	5	24	26	N	Y
R3GFX9ZNZ0T3TA	1	35	35	N	Y
R2H8MH0N2IZKMY	5	22	23	N	Y
R1MYE8WQ3812L8	5	17	20	N	Y
RYSL8MU2858GB	5	71	73	N	Y

only showing top 20 rows

- In the last part, we have calculated the total number of reviews, the number of 5-star reviews, and the percentage of 5-star reviews for paid and unpaid program

Results and Summary:

Based on our analysis, we have found the following:

- The total number of Vine reviews is 22, while the total number of reviews for unpaid program is 26987.

- The number of 5-star reviews for paid program is 13 while it's 14475 for unpaid program.
- The percentage of 5-star reviews for paid program is: 0.59% which is so close to the percentage of 5-star reviews for unpaid program: 0.53%

Even if the percentage of 5-star reviews for paid and unpaid program is close one should consider the existence of a bias for reviews in the Vine program due to the considerably big difference between the total number of Vine reviews and non-Vine reviews; In other words, the number of Vine reviews is a not even close to 1% of that of the non-Vine reviews. I would suggest performing a T-test to make sure if there's a statistical difference between the mean of the sample and population distribution.