

▼ Vine Review Analysis

By Emmanuel Martinez

```
import os
# Find the latest version of spark 2.0 from http://www-us.apache.org/dist/spark/ and enter as
# For example:
# spark_version = 'spark-2.4.7'
spark_version = 'spark-2.4.7'
os.environ['SPARK_VERSION']=spark_version

# Install Spark and Java
!apt-get update
!apt-get install openjdk-11-jdk-headless -qq > /dev/null
!wget -q http://www-us.apache.org/dist/spark/\$SPARK\_VERSION/\$SPARK\_VERSION-bin-hadoop2.7.tgz
!tar xf $SPARK_VERSION-bin-hadoop2.7.tgz
!pip install -q findspark

# Set Environment Variables ",
import os
os.environ["JAVA_HOME"] = "/usr/lib/jvm/java-11-openjdk-amd64"
os.environ["SPARK_HOME"] = f"/content/{spark_version}-bin-hadoop2.7"

# Start a SparkSession
import findspark
findspark.init()
```

```
Ign:1 https://developer.download.nvidia.com/compute/cuda/repos/ubuntu1804/x86\_64 InRel
Hit:2 https://cloud.r-project.org/bin/linux/ubuntu bionic-cran40/ InRelease
Hit:3 http://ppa.launchpad.net/c2d4u.team/c2d4u4.0+/ubuntu bionic InRelease
Hit:4 http://security.ubuntu.com/ubuntu bionic-security InRelease
Ign:5 https://developer.download.nvidia.com/compute/machine-learning/repos/ubuntu1804/x86\_64
Hit:6 http://archive.ubuntu.com/ubuntu bionic InRelease
Hit:7 https://developer.download.nvidia.com/compute/cuda/repos/ubuntu1804/x86\_64 Relea
Hit:8 https://developer.download.nvidia.com/compute/machine-learning/repos/ubuntu1804/x86\_64
Hit:9 http://archive.ubuntu.com/ubuntu bionic-updates InRelease
Hit:10 http://ppa.launchpad.net/cran/libgit2/ubuntu bionic InRelease
Hit:11 http://archive.ubuntu.com/ubuntu bionic-backports InRelease
Hit:12 http://ppa.launchpad.net/deadsnakes/ppa/ubuntu bionic InRelease
Hit:13 http://ppa.launchpad.net/graphics-drivers/ppa/ubuntu bionic InRelease
Reading package lists... Done
```



```
!pip install pyspark
```

```
Collecting pyspark
  Downloading https://files.pythonhosted.org/packages/45/b0/9d6860891ab14a39d4bddf80ba2
|████████████████████████████████████████| 212.3MB 55kB/s
Collecting py4j==0.10.9
  Downloading https://files.pythonhosted.org/packages/9e/b6/6a4fb90cd235dc8e265a6a2067f
|████████████████████████████████████████| 204kB 13.5MB/s
Building wheels for collected packages: pyspark
  Building wheel for pyspark (setup.py) ... done
  Created wheel for pyspark: filename=pyspark-3.1.1-py2.py3-none-any.whl size=212767604
  Stored in directory: /root/.cache/pip/wheels/0b/90/c0/01de724414ef122bd05f056541fb6a0
Successfully built pyspark
Installing collected packages: py4j, pyspark
Successfully installed py4j-0.10.9 pyspark-3.1.1
WARNING: The following packages were previously imported in this runtime:
[py4j,pyspark]
You must restart the runtime in order to use newly installed versions.
```

```
# Download the Postgres driver that will allow Spark to interact with Postgres.
!wget https://jdbc.postgresql.org/download/postgresql-42.2.16.jar
```

```
--2021-03-11 18:34:33-- https://jdbc.postgresql.org/download/postgresql-42.2.16.jar
Resolving jdbc.postgresql.org (jdbc.postgresql.org)... 72.32.157.228, 2001:4800:3e1:1::
Connecting to jdbc.postgresql.org (jdbc.postgresql.org)|72.32.157.228|:443... connected
HTTP request sent, awaiting response... 200 OK
Length: 1002883 (979K) [application/java-archive]
Saving to: 'postgresql-42.2.16.jar.1'
```

```
postgresql-42.2.16. 100%[=====>] 979.38K --.-KB/s in 0.1s
```

```
2021-03-11 18:34:34 (6.59 MB/s) - 'postgresql-42.2.16.jar.1' saved [1002883/1002883]
```

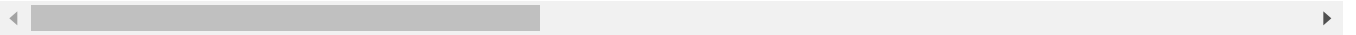
```
from pyspark.sql import SparkSession
spark = SparkSession.builder.appName("BigData-ChallengeD2").config("spark.driver.extraClassPa

from pyspark import SparkFiles
url = "https://s3.amazonaws.com/amazon-reviews-pds/tsv/amazon_reviews_us_Furniture_v1_00.tsv"
spark.sparkContext.addFile(url)
df = spark.read.option("encoding", "UTF-8").csv(SparkFiles.get("amazon_reviews_us_Furniture_v
df.show())
```

marketplace	customer_id	review_id	product_id	product_parent	product_title
US	24509695	R3VR960AHLFKDV	B004HB5E0E	488241329	Shoal Creek Compu...
US	34731776	R16LGVMFKIUT0G	B0042TNMMS	205864445	Dorel Home Produc...
US	1272331	R1AIMEEPYHMOE4	B0030MPBZ4	124663823	Bathroom Vanity T...
US	45284262	R1892CCSZWZ9SR	B005G02ESA	382367578	Sleep Master Ulti...
US	30003523	R285P679YVWKD1	B005JS8AUA	309497463	1 1/4" GashGuards...
US	18311821	RLB33HJBXHZHU	B00AVUQQGQ	574537906	Serta Bonded Leat...
US	42943632	R1VGTZ94DBAD6A	B00CFY20GQ	407473883	Prepac Shoe Stora...
US	43157304	R168KF82ICSOHD	B00FKC48QA	435120460	HomCom PU Leather...
US	51918480	R20DIYIJ0OCMOG	B00N9IAL9K	356495985	Folding Step Stool
US	14522766	RD46RNV0HNZSC	B001T4XU1C	243050228	Ace Bayou Adult V...

US	43054112	R2JDOCETTM3AXS	B002HRFLBC	93574483	4D Concepts Audio...
US	26622950	R33YMW36IDZ6LE	B006MISZOC	941823468	Zinus SC-SBBK-14N...
US	17988940	R30ZGGUHZ04C1S	B008BMGABC	460567746	Poundex Marble Di...
US	18444952	RS2EZU76IK2BT	B00C02VH5Y	829613894	Safavieh Lyndhurs...
US	16937084	R1GJC1BP028X09	B00LI4RJQ0	816478187	Sauder Boone Moun...
US	23665632	R2VKJPGXXEK5GP	B0046EC1D0	358594389	Winsome Wood Brea...
US	4110125	R17KS83G3KLT97	B00DQQPL36	312571325	HODEDAH IMPORT Me...
US	107621	R3PQL8SR4NEHWL	B003X7RWB2	402665054	Flash Furniture H...
US	2415090	R2F5WW7WNO5RRG	B001TJYPJ8	854989315	Sleep Revolution ...
US	48285966	R3UDJKVWQCFIC9	B000TMHX9A	814079288	Flash Furniture V...

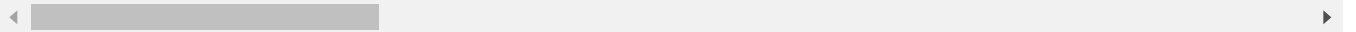
only showing top 20 rows



DELIVERABLE 2

```
# Confirm total_vote, helpful_votes, and star_rating are all integers
df
```

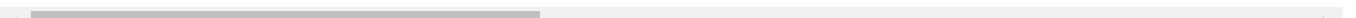
```
DataFrame[marketplace: string, customer_id: int, review_id: string, product_id: string,
```



```
# Filter furniture data frame on total_vote greater than or equal to 20
totalvotes_df = df.filter("total_votes>=20")
totalvotes_df.show()
```

marketplace	customer_id	review_id	product_id	product_parent	product_title
US	41681546	RL8D0KJ0J9L00	B00BWC1X3S	328960153	Zinus 14 Inch Eli...
US	16806846	R1BEINAIQFBRJC	B007I81A60	68465765	8" Night Therapy ...
US	17820936	R2L59KIJH302P9	B007QULII0	812343799	LexMod Stencil Di...
US	50476494	RR99CPG695T0I	B013L4ZWWW	124365349	Arctic Dreams 10"...
US	21846903	R1XQNKKUPCMWVO	B00IPN64CC	867520220	Upholstered Platf...
US	36021042	R3JUXVCT1NSK2A	B003PSNAI8	219454481	Fashion Bed Group...
US	2067832	R3GNSIFV1J2Y2B	B008RKOL9G	589861187	Big Joe Media Lou...
US	15624624	RTCRZARYY4LXX	B00E97HPLW	132456290	Flash Furniture M...
US	13214979	R30FB4P7Y8WR27	B00QBZ25R4	637686095	Tuft & Needle Mat...
US	37722720	R3MTAYGQM25N63	B007J0E7WQ	562508640	Target Marketing ...
US	31278317	RJNDSWES5ISZ7	B00HTU00BA	995797488	Chrome Metal Bar ...
US	21515203	R15R7STOZZP2A4	B002V1Q65Y	320882925	Coaster Company P...
US	48672213	R33V5WV529NK8E	B00MULXWMU	491669252	Modway Annabel Tw...
US	50914859	RIR9ZI3L80P7P	B00T3MUB4Q	76857765	WinkBeds - Mattre...
US	7691605	R10P6SDC1D6C3I	B00LZM5I0K	420170495	Modway Lily Twin ...
US	46509642	R110G9UVLI2MLS	B00OKIPS0A	415535727	Milton Greens Sta...
US	47359106	R1I4LN1WR3YVJX	B00NUS5D4C	148579574	AmazonBasics 5-Sh...
US	23511203	R1B76MPCS05UX9	B00P7RWWM0	259554464	Mega Motion Easy ...
US	6407891	R17PJIAKAZ3U6BG	B00DW7LB0G	366981061	Simpli Home Lared...
US	8113934	R2T3TLCX42RWLY	B009V5YVT6	598492986	AVE SIX Reflectio...

only showing top 20 rows



```
helpful_50_df = totalvotes_df.filter("helpful_votes/total_votes>=.50")
helpful_50_df.show()
```

marketplace	customer_id	review_id	product_id	product_parent	product_title
US	41681546	RL8D0KJ0J9L00	B00BWC1X3S	328960153	Zinus 14 Inch Eli...
US	16806846	R1BEINAIQFBRJC	B007I81A60	68465765	8" Night Therapy ...
US	17820936	R2L59KIJH302P9	B007QULII0	812343799	LexMod Stencil Di...
US	50476494	RR99CPG695T0I	B013L4ZWWW	124365349	Arctic Dreams 10"...
US	21846903	R1XQNKKUPCMWVO	B00IPN64CC	867520220	Upholstered Platf...
US	36021042	R3JUXVCT1NSK2A	B003PSNAI8	219454481	Fashion Bed Group...
US	15624624	RTCRZARYY4LXX	B00E97HPLW	132456290	Flash Furniture M...
US	13214979	R30FB4P7Y8WR27	B00QBZ25R4	637686095	Tuft & Needle Mat...
US	37722720	R3MTAYGQM25N63	B007J0E7WQ	562508640	Target Marketing ...
US	31278317	RJNDSWES5ISZ7	B00HTU00BA	995797488	Chrome Metal Bar ...
US	21515203	R15R7STOZZP2A4	B002V1Q65Y	320882925	Coaster Company P...
US	48672213	R33V5WV529NK8E	B00MULXWMU	491669252	Modway Annabel Tw...
US	50914859	RIR9ZI3L80P7P	B00T3MUB4Q	76857765	WinkBeds - Mattre...
US	7691605	R10P6SDC1D6C3I	B00LZM5I0K	420170495	Modway Lily Twin ...
US	46509642	R110G9UVLI2MLS	B00OKIPS0A	415535727	Milton Greens Sta...
US	47359106	R1I4LN1WR3YVJX	B00NUS5D4C	148579574	AmazonBasics 5-Sh...
US	23511203	R1B76MPCS05UX9	B00P7RWWM0	259554464	Mega Motion Easy ...
US	6407891	R17PJIKAZ3U6BG	B00DW7LB0G	366981061	Simpli Home Lared...
US	8113934	R2T3TLCX42RWLY	B009V5YVT6	598492986	AVE SIX Reflectio...
US	48960668	R39YWJ09ZCPW7P	B00UXRSDF4	949941628	2xhome - Black - ...

only showing top 20 rows

```
print(helpful_50_df.count())
```

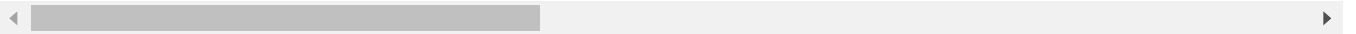
18155

```
vine_review_df = helpful_50_df.filter(helpful_50_df["vine"] == "Y")
vine_review_df.show()
```

marketplace	customer_id	review_id	product_id	product_parent	product_title
US	35119071	R2BQOD1R0228FN	B00H2RSA88	405483618	Sleep Innovations...
US	44737123	RC31RUPFOHBHQ	B0125QZ50G	350975212	Zinus Viscolatex ...
US	51369093	REN3N1WITLF1Y	B00S5EI4C4	153136932	Brentwood Home S-...
US	49598970	R71RZQ9UZVG47	B00H2RSA88	405483618	Sleep Innovations...
US	50507621	R38NMQBH88HLM6	B00H2RSA88	405483618	Sleep Innovations...
US	44737123	R33FGX9EE3QVR6	B00K7G8M34	291701473	Fashion Bed Group...
US	17957446	R1KIOIK6WEYE59	B00ZY8DOCY	858354913	Harmony Ergonomic...
US	42483100	R25X9BMOB3FD0E	B00S5EI4C4	153136932	Brentwood Home S-...
US	52464985	R3VCKFCX2377Q2	B00Y4BGT06	383077572	Sleep Innovations...
US	46266548	R1E00UG63HMSM4	B00Y4BGT06	383077572	Sleep Innovations...
US	51923814	R1V45RUW5ZB3LS	B00W00S5EM	720135802	Serta Pearce Firm...
US	52862683	RTF6DSZ1UTLHH	B00LJ7D280	576833455	Classic Brands Gr...

US	51903105	R1CZV9N2YLJAP7	B00LJ7D280	576833455	Classic Brands Gr...
US	35575415	R10F3X9W99Y300	B00LJ7D280	576833455	Classic Brands Gr...
US	51956455	R1JYKEH4CQVJ1B	B00LJ7D280	576833455	Classic Brands Gr...
US	37982975	R1093XVB0H2QOL	B00Y4BGT06	383077572	Sleep Innovations...
US	22656237	R3Q81B31F1CPGH	B00P21TAIK	298322939	Chill Sack Bean B...
US	34160155	R2P6XIZZPJF7AE	B00P21TAIK	298322939	Chill Sack Bean B...
US	33167968	R3N5S06UW5MKFE	B00UVAJUBO	584650655	Classic Brands 10...
US	52846213	R3J9EJC VKFCRW0	B00UA8TD60	78126929	Safavieh Tranquil...

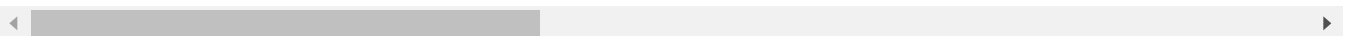
only showing top 20 rows



```
novine_review_df = helpful_50_df.filter(helpful_50_df["vine"] == "N")
novine_review_df.show()
```

marketplace	customer_id	review_id	product_id	product_parent	product_title
US	41681546	RL8D0KJ0J9L00	B00BWC1X3S	328960153	Zinus 14 Inch Eli...
US	16806846	R1BEINAIQFBRJC	B007I81A60	68465765	8" Night Therapy ...
US	17820936	R2L59KIJH302P9	B007QULII0	812343799	LexMod Stencil Di...
US	50476494	RR99CPG695T0I	B013L4ZWWW	124365349	Arctic Dreams 10"...
US	21846903	R1XQNK KUPCMWV0	B00IPN64CC	867520220	Upholstered Platf...
US	36021042	R3JUXVCT1NSK2A	B003PSNAI8	219454481	Fashion Bed Group...
US	15624624	RTCRZARYY4LXX	B00E97HPLW	132456290	Flash Furniture M...
US	13214979	R30FB4P7Y8WR27	B00QBZ25R4	637686095	Tuft & Needle Mat...
US	37722720	R3MTAYGQM25N63	B007J0E7WQ	562508640	Target Marketing ...
US	31278317	RJNDSWES5ISZ7	B00HTU00BA	995797488	Chrome Metal Bar ...
US	21515203	R15R7STOZZP2A4	B002V1Q65Y	320882925	Coaster Company P...
US	48672213	R33V5WV529NK8E	B00MULXWMU	491669252	Modway Annabel Tw...
US	50914859	RIR9ZI3L80P7P	B00T3MUB4Q	76857765	WinkBeds - Mattre...
US	7691605	R10P6SDC1D6C3I	B00LZM5I0K	420170495	Modway Lily Twin ...
US	46509642	R110G9UVLI2MLS	B00OKIPS0A	415535727	Milton Greens Sta...
US	47359106	R1I4LN1WR3YVJX	B00NUS5D4C	148579574	AmazonBasics 5-Sh...
US	23511203	R1B76MPCS05UX9	B00P7RWWM0	259554464	Mega Motion Easy ...
US	6407891	R17PJIAKAZ3U6BG	B00DW7LB0G	366981061	Simpli Home Lared...
US	8113934	R2T3TLCX42RWLY	B009V5YVT6	598492986	AVE SIX Reflectio...
US	48960668	R39YWJ09ZCPW7P	B00UXRSDF4	949941628	2xhome - Black - ...

only showing top 20 rows



```
from pyspark.sql.functions import count
# Total number of reviews for each dataframe from Step 3 and Step 4:
vine_review_count = vine_review_df.count()
print("Total Number of Reviews PAID and 'helpful': %f" % vine_review_count)

novine_review_count = novine_review_df.count()
print("Total Number of Reviews NON-PAID and 'helpful': %f" % novine_review_count)
```

```
Total Number of Reviews PAID and 'helpful': 136.000000
Total Number of Reviews NON-PAID and 'helpful': 18019.000000
```

```

# Number of 5-star Reviews for above Paid and Non-Paid dataframes
star5_vine_df = vine_review_df.filter(vine_review_df["star_rating"]=="5")
print("Total Number of 5-star Reviews PAID and 'helpful': %f" % star5_vine_df.count())

star5_novine_df = novine_review_df.filter(novine_review_df["star_rating"]=="5")
print("Total Number of 5-star Reviews NON-PAID and 'helpful': %f" % star5_novine_df.count())

    Total Number of 5-star Reviews PAID and 'helpful': 74.000000
    Total Number of 5-star Reviews NON-PAID and 'helpful': 8482.000000

# Percentage of 5-star Reviews for above Paid and Non-Paid dataframes
star5_vine_pct = (star5_vine_df.count()/vine_review_count)
print("Percent of 5-Star Furniture Reviews from PAID, 'helpful' dataset: %f" % star5_vine_pct)

star5_novine_pct = (star5_novine_df.count()/novine_review_count)
print("Percent of 5-Star Furniture Reviews from NON-PAID, 'helpful' dataset: %f" % star5_novi

    Percent of 5-Star Furniture Reviews from PAID, 'helpful' dataset: 0.544118
    Percent of 5-Star Furniture Reviews from NON-PAID, 'helpful' dataset: 0.470725

# Total number of ALL reviews
total_reviews_count = df.count()
print("Total Number of Furniture Reviews: %f" % total_reviews_count)

help_reviews_count = helpful_50_df.count()
print("Total Number of 'Helpful' Furniture Reviews: %f" % help_reviews_count)

    Total Number of Furniture Reviews: 792113.000000
    Total Number of 'Helpful' Furniture Reviews: 18155.000000

#Total number of ALL 5-star reviews
star5_df = df.filter(df["star_rating"] == '5')
star5_df.show()

star5_count = star5_df.count()
print("Total Number of 5-Star Furniture Reviews: %f" % star5_count)

```

```

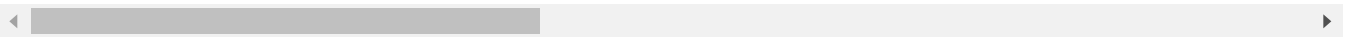
+-----+-----+-----+-----+-----+-----+
|marketplace|customer_id|    review_id|product_id|product_parent|    product_title|
+-----+-----+-----+-----+-----+-----+
|      US|    34731776|R16LGVMFKIUT0G|B0042TNMMS|    205864445|Dorel Home Produc...|
|      US|    1272331|R1AIMEEPYHMOE4|B0030MPBZ4|    124663823|Bathroom Vanity T...|
|      US|    18311821|RLB33HJBXHZHU|B00AVUQQGQ|    574537906|Serta Bonded Leat...|
|      US|    42943632|R1VGTZ94DBAD6A|B00CFY20GQ|    407473883|Prepac Shoe Stora...|
|      US|    43157304|R168KF82ICSOHD|B00FKC48QA|    435120460|HomCom PU Leather...|
|      US|    51918480|R20DIYIJ0OCMOG|B00N9IAL9K|    356495985|    Folding Step Stool|
|      US|    14522766|RD46RNV0HNZSC|B001T4XU1C|    243050228|Ace Bayou Adult V...|
|      US|    43054112|R2JDOCETTM3AXS|B002HRFLBC|    93574483|4D Concepts Audio...|
|      US|    26622950|R33YMW36IDZ6LE|B006MISZOC|    941823468|Zinus SC-SBBK-14N...|

```

US	17988940	R30ZGGUHZ04C1S	B008BMGABC	460567746	Poundex Marble Di...
US	18444952	RS2EZU76IK2BT	B00CO2VH5Y	829613894	Safavieh Lyndhurs...
US	16937084	R1GJC1BP028X09	B00LI4RJQ0	816478187	Sauder Boone Moun...
US	2415090	R2F5WW7WNO5RRG	B001TJYPJ8	854989315	Sleep Revolution ...
US	48285966	R3UDJKVWQCFIC9	B000TMHX9A	814079288	Flash Furniture V...
US	33228559	R1MLGVJH3J5W6N	B005MZBB20	523675277	Amelia Floral Mut...
US	11600823	R38JYICG7ESA2P	B00JITHL9C	403803724	Classic Brands Lo...
US	12546512	R20M7LJ7CY0840	B007EEG7M0	336731383	Range Kleen SS2 D...
US	5395378	R1ZG4UWGK123C3	B004LQ1RJ2	999119538	Signature Sleep C...
US	34857314	R1ZOE0SEMAITN6	B000RPVKPC	269111919	Coaster Home Furn...
US	26511421	R38HR07J1KN5TE	B009QDA1LA	119886213	Universal Bed Sla...

only showing top 20 rows

Total Number of 5-Star Furniture Reviews: 447716.000000



#Total number of 5-star HELPFUL reviews

```
star5_help_df = helpful_50_df.filter(helpful_50_df["star_rating"] == '5')
```

```
star5_help_df.show()
```

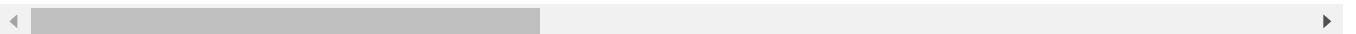
```
star5_help_count = star5_help_df.count()
```

```
print("Total Number of 5-Star 'Helpful' Furniture Reviews: %f" % star5_help_count)
```

marketplace	customer_id	review_id	product_id	product_parent	product_title
US	41681546	RL8D0KJ0J9L00	B00BWC1X3S	328960153	Zinus 14 Inch Eli...
US	16806846	R1BEINAIQFBRJC	B007I81A60	68465765	8" Night Therapy ...
US	50476494	RR99CPG695T0I	B013L4ZWWW	124365349	Arctic Dreams 10"...
US	21846903	R1XQNKKUPCMWVO	B00IPN64CC	867520220	Upholstered Platf...
US	15624624	RTCRZARYY4LXX	B00E97HPLW	132456290	Flash Furniture M...
US	31278317	RJNDSWES5ISZ7	B00HTU00BA	995797488	Chrome Metal Bar ...
US	21515203	R15R7STOZZP2A4	B002V1Q65Y	320882925	Coaster Company P...
US	7691605	R10P6SDC1D6C3I	B00LZM5I0K	420170495	Modway Lily Twin ...
US	46509642	R110G9UVLI2MLS	B00OKIPS0A	415535727	Milton Greens Sta...
US	23511203	R1B76MPCS05UX9	B00P7RWWM0	259554464	Mega Motion Easy ...
US	6407891	R17PJIAKAZ3U6BG	B00DW7LB0G	366981061	Simpli Home Lared...
US	8113934	R2T3TLCX42RWLY	B009V5YVT6	598492986	AVE SIX Reflectio...
US	48960668	R39YWJ09ZCPW7P	B00UXRSDF4	949941628	2xhome - Black - ...
US	27730443	R3JM7HESVLH0G9	B000SBJIFY	595491245	TMS Tiffany Natur...
US	19440640	R2Q53X3HCQNVML	B00CXPJG3C	680349249	Reverie 7S Adjust...
US	434516	R25KMMHJ2YX23G	B00HFQJFVW	518045729	Dorel Living King...
US	43374324	R2R34BM9Y1LQTQ	B009YQH03W	299279247	Convenience Conce...
US	47591171	R136U4VZOMTEBP	B0034JXLGY	419832011	Sauder Beginnings...
US	42860527	R2EGVZYZPED2SN	B00VKK3RCW	476069686	Naomi Home Venice...
US	38848281	R3UCKDORUOX8DR	B004JWVYBY	197794057	Safavieh Heritage...

only showing top 20 rows

Total Number of 5-Star 'Helpful' Furniture Reviews: 8556.000000



Percentage of ALL 5-star reviews paid vs non-paid

5-star Paid reviews

```

vine_star5_all_df = star5_df.filter(star5_df["vine"] == "Y")
vine_star5_all_count = vine_star5_all_df.count()

vine_star5_all_pct = (vine_star5_all_count/star5_count)
print("Number of 5-Star Furniture Reviews PAID: %f" % vine_star5_all_count)
print("Percent of 5-Star Furniture Reviews PAID: %f" % vine_star5_all_pct)

#5-star Non-Paid reviews
novine_star5_all_df = star5_df.filter(star5_df["vine"] == "N")
novine_star5_all_count = novine_star5_all_df.count()

novine_star5_all_pct = novine_star5_all_count/star5_count
print("Number of 5-Star Furniture Reviews NON-PAID: %f" % novine_star5_all_count)
print("Percent of 5-Star Furniture Reviews NON-PAID: %f" % novine_star5_all_pct)

    Number of 5-Star Furniture Reviews PAID: 1356.000000
    Percent of 5-Star Furniture Reviews PAID: 0.003029
    Number of 5-Star Furniture Reviews NON-PAID: 446360.000000
    Percent of 5-Star Furniture Reviews NON-PAID: 0.996971

# Percentage of HELPFUL 5-star reviews paid vs non-paid (% of 5 star reviews in Step 3 datafr
# 5-star Paid "helpful" reviews
vine_star5_help_df = star5_help_df.filter(star5_help_df["vine"] == "Y")
vine_star5_help_count = vine_star5_help_df.count()

vine_star5_help_pct = (vine_star5_help_count/star5_help_count)
print("Number of 5-Star, 'Helpful' Furniture Reviews PAID: %f" % vine_star5_help_count)
print("Percent of 5-Star, 'Helpful' Furniture Reviews PAID: %f" % vine_star5_help_pct)

#5-star Non-Paid "helpful" reviews
novine_star5_help_df = star5_help_df.filter(star5_help_df["vine"] == "N")
novine_star5_help_count = novine_star5_help_df.count()

novine_star5_help_pct = novine_star5_help_count/star5_help_count
print("Number of 5-Star, 'Helpful' Furniture Reviews NON-PAID: %f" % novine_star5_help_count)
print("Percent of 5-Star, 'Helpful' Furniture Reviews NON-PAID: %f" % novine_star5_help_pct)

    Number of 5-Star, 'Helpful' Furniture Reviews PAID: 74.000000
    Percent of 5-Star, 'Helpful' Furniture Reviews PAID: 0.008649
    Number of 5-Star, 'Helpful' Furniture Reviews NON-PAID: 8482.000000
    Percent of 5-Star, 'Helpful' Furniture Reviews NON-PAID: 0.991351

```

Additional Views for Write

```

#Percent of Vine vs non-Vine all reviews
vine_all_df = df.filter(df["vine"] == "Y")
novine_all_df = df.filter(df["vine"] == "N")
vine_pct_all = (vine_all_df.count()/df.count())
novine_pct_all = (novine_all_df.count()/df.count())

```



```

print(df.count())
print(vine_all_df.count())
print(novine_all_df.count())
print(vine_pct_all)
print(novine_pct_all)

```

```

792113
2775
789338
0.003503288040973952
0.996496711959026

```

```
# Percent of All Vine that is 5-Star
```

```

vine_all_star5_df = vine_all_df.filter(vine_all_df["star_rating"] == '5')
vine_all_star_df = vine_all_df.filter(vine_all_df["star_rating"] != '5')

```

```

novine_all_star5_df = novine_all_df.filter(novine_all_df["star_rating"] == '5')
novine_all_star_df = novine_all_df.filter(novine_all_df["star_rating"] != '5')

```

```

vine_pct_star5_all = (vine_all_star5_df.count()/vine_all_df.count())
novine_pct_star5_all = (novine_all_star5_df.count()/novine_all_df.count())

```

```

print(vine_all_star5_df.count())
print(vine_all_star_df.count())

```

```

print(novine_all_star5_df.count())
print(novine_all_star_df.count())

```

```

print(vine_pct_star5_all)
print(novine_pct_star5_all)

```

```

1356
1419
446360
342978
0.48864864864864865
0.5654865216168485

```

Analysis Recommendation

```
# Pull out all Vine reviewed products
```

```

vine_prod_df = vine_all_df.select(vine_all_df["product_id"]).dropDuplicates()
vine_prod_df.show()
vine_prod_df.count()

```

```

+-----+
|product_id|
+-----+
|B00DL1AG4A|
|B00IMV6Q3Y|

```

```
|B00CPIV9BE|
|B00R465FZS|
|B007SB0554|
|B00HAR2A1I|
|B0030GJAHK|
|B00J38BRQG|
|B00FU0D2KK|
|B00W00SQ4Q|
|B007T58QPU|
|B00H4SSFLM|
|B00CPIVB1W|
|B00CP2II9Q|
|B00T2XF8B2|
|B003PJ650C|
|B00EPRCAVG|
|B005ETZIMM|
|B00032ZX4Y|
|B00FL0239U|
```

```
+-----+
```

only showing top 20 rows

421

```
# create an inner join between all the Vine Reviews df and the full df
inner_join = vine_prod_df.join(df, vine_all_df.product_id == df.product_id, how="inner").drop
inner_join.show()
inner_join.count()
```

```
+-----+-----+-----+-----+-----+-----+-----+
|product_id|marketplace|customer_id|review_id|product_id|product_parent|product_name|
+-----+-----+-----+-----+-----+-----+-----+
|B00DL1AG4A|US|21287734|R278IHKE94IZFT|B00DL1AG4A|987680523|Bush Furni
|B00DL1AG4A|US|42602916|R2MFYCTRF82JSN|B00DL1AG4A|987680523|Bush Furni
|B00DL1AG4A|US|30566260|R3V5OVE5VRD1BE|B00DL1AG4A|987680523|Bush Furni
|B00DL1AG4A|US|22896992|RFEGCCMNTP06L|B00DL1AG4A|987680523|Bush Furni
|B00DL1AG4A|US|43897601|R2VZAWFVZS8FFP|B00DL1AG4A|987680523|Bush Furni
|B00DL1AG4A|US|15699015|R2RX4XSD7I724L|B00DL1AG4A|987680523|Bush Furni
|B00DL1AG4A|US|36944012|R14XXOXUATSFXP|B00DL1AG4A|987680523|Bush Furni
|B00DL1AG4A|US|50413437|R1YFSV4EY2N9TF|B00DL1AG4A|987680523|Bush Furni
|B00IMV6Q3Y|US|39526020|R3ARTJCTBUFNUV|B00IMV6Q3Y|36874631|Serta Perf
|B00IMV6Q3Y|US|50564860|R3R2057XEHTG1C|B00IMV6Q3Y|36874631|Serta Perf
|B00IMV6Q3Y|US|14461943|R17DJ1FCVFOVRK|B00IMV6Q3Y|36874631|Serta Perf
|B00IMV6Q3Y|US|29744123|RNAKZ8DL3EPX5|B00IMV6Q3Y|36874631|Serta Perf
|B00CPIV9BE|US|17132648|R20RMSLJWPF7VH|B00CPIV9BE|58070008|Beautyrest
|B007SB0554|US|23705182|REPS0XQNM90M|B007SB0554|534624126|South Shor
|B007SB0554|US|44249079|R3UJ3WQ7ZS7YGT|B007SB0554|534624126|South Shor
|B007SB0554|US|25591162|R2TZIBKTG92WLD|B007SB0554|534624126|South Shor
|B007SB0554|US|49005234|R2YAR7CY2X05U7|B007SB0554|534624126|South Shor
|B007SB0554|US|14231385|R12I1W15ILXWTR|B007SB0554|534624126|South Shor
|B007SB0554|US|14392655|R2PI7YMXDYIHRC|B007SB0554|534624126|South Shor
|B007SB0554|US|6176221|R8VCQ744IRY5N|B007SB0554|534624126|South Shor
+-----+-----+-----+-----+-----+-----+-----+
```

only showing top 20 rows

10689

```
# Rating breakdown
new_df = inner_join.groupby("star_rating").count()
new_df.show()
```

star_rating	count
1	663
3	933
5	6093
4	2471
2	529

```
# Total Vine Reviews for Vine Products
vine_new_df = inner_join.filter(inner_join["vine"] == "Y")
vine_new_df.count()
```

2775

```
# Total Vine Reviews for Vine Products
novine_new_df = inner_join.filter(inner_join["vine"] == "N")
novine_new_df.count()
```

7914

```
# % of 5-star reviews are Vine or non-Vine Reviews
vine_star5_new = vine_new_df.groupby("star_rating").count()
vine_star5_new.show()
```

star_rating	count
1	50
3	305
5	1356
4	961
2	103

```
# % of 5-star reviews are Vine or non-Vine Reviews
novine_star5_new = novine_new_df.groupby("star_rating").count()
novine_star5_new.show()
```

star_rating	count
1	613

	3		628	
	5		4737	
	4		1510	
	2		426	
+-----+				+

Vine Review Analysis completion by **Emmanuel Martinez**