Amazon_Vine_Analysis.ipynb for Outdoor data

```
import os
# Find the latest version of spark 3.0  from http://www.apache.org/dist/spark/ and enter as t
# For example:
# spark_version = 'spark-3.0.3'
spark_version = 'spark-3.1.3'
os.environ['SPARK_VERSION']=spark_version



# Install Spark and Java
!apt-get install openjdk-8-jdk-headless -qq > /dev/null
!wget https://downloads.apache.org/spark/spark-3.1.3/spark-3.1.3-bin-hadoop2.7.tgz
!tar -xvf spark-3.1.3-bin-hadoop2.7.tgz
!pip install -q findspark

import os
os.environ["JAVA_HOME"] = "/usr/lib/jvm/java-8-openjdk-amd64"
os.environ["SPARK_HOME"] = "/content//spark-3.1.3-bin-hadoop2.7"

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

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

```
--2022-09-16 15:05:12--  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::2
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’

postgresql-42.2.16. 100%[===================>] 979.38K  4.57MB/s    in 0.2s

2022-09-16 15:05:12 (4.57 MB/s) - ‘postgresql-42.2.16.jar’ saved [1002883/1002883]
```

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

Load Amazon Data into Spark DataFrame

```
from pyspark import SparkFiles
url = "https://s3.amazonaws.com/amazon-reviews-pds/tsv/amazon_reviews_us_Outdoors_v1_00.tsv.g
spark.sparkContext.addFile(url)
df = spark.read.option("encoding", "UTF-8").csv(SparkFiles.get("amazon_reviews_us_Outdoors_v1_
df.show()
```

```
+-----------+-----------+--------------+----------+--------------+--------------------+-
|marketplace|customer_id|     review_id|product_id|product_parent|       product_title|
+-----------+-----------+--------------+----------+--------------+--------------------+-
|         US|   18446823|R35T75OLUGHL5C|B000NV6H94|     110804376|Stearns Youth Boa...|
|         US|   13724367|R2BV735O46BN33|B000IN0W3Y|     624096774|Primal Wear Men's...|
|         US|   51001958|R2NBEUGPQQGXP1|B008RBJXFM|     278970944|Osprey Hydraulics...|
|         US|   32866903|R17LLAOJ8ITK0S|B00FK8WUQY|     312877650|CamelBak eddy .75...|
|         US|   30907790|R39PEQBT5ISEF4|B00EZA3VW0|     305567912|Children Black Re...|
|         US|   20232229|R3GNM3SU9VHJFT|B006JA8WEG|     842306035|Ibera Bicycle Tri...|
|         US|   17698862| R2Y81OP0EK467|B002PWFSEO|     451480122|Therm-a-Rest Comp...|
|         US|   38486114|R2LFGSI6HAYH5F|B002DZGKHW|     124386306|Sawyer Products P...|
|         US|   26319572|R297G6ED1IQO7W|B00ABA08F6|     991442421|Zippo Hand Warmer...|
|         US|   27152337| RE27RFC6101N6|B003Z8WIHC|     886483892|Camp Chef Dutch O...|
|         US|   12516845|R3BPDME6E94W8Z|B007CP6UK0|     150224054|3CERA Portable Wi...|
|         US|    3225242|R2P08O1RILUOX3|B003V3U9JK|     343847969|Texsport King Kot...|
|         US|     961839|R37CVAB03PTDVI|B00Y846HN8|     858088629|Wallygadgets 2 Wh...|
|         US|   47796452| RAWNWOGXPCPMD|B00IYQ84VY|     474493517|RainStoppers 34-I...|
|         US|   32004835| R5DYGP6ASX77M|B002MYCKLY|     920014456|Alpha Deluxe Port...|
|         US|   23972939|R1O0SAOOGF2KG7|B00EZV69JG|     128489321|Speedfil Z4 BTA B...|
|         US|   40889047|R35NJUT0U3MU3V|B00AWOT3T8|     571303876|O'Brien Kids Plat...|
|         US|   11244387|R242C08MF9D1AH|B0000AXTID|     739769424|Kwik-Tek F-5R Pla...|
|         US|   20121211| R3RYG8TJTO4E2|B00IFHFJXI|     984009972|Ivation Portable ...|
|         US|   25657249|R3IKH1DNY0CP9F|B00KFILTWU|     405521681|GreenInsync Repla...|
+-----------+-----------+--------------+----------+--------------+--------------------+-
only showing top 20 rows
```

```
# filter DataFrame for total_votes above or equal to 20
df1 = df.filter(df.total_votes >= 20)
df1.show(5)
```

```
+-----------+-----------+--------------+----------+--------------+--------------------+-
|marketplace|customer_id|     review_id|product_id|product_parent|       product_title|
+-----------+-----------+--------------+----------+--------------+--------------------+-
|         US|   30222858|R2FP3U4NHNFNL2|B00YPISPNC|     468413405|Stanley Classic V...|
|         US|   35677754|R1UUK1977O38MU|B00T8NEI3A|      32341693|Camping Wood Burn...|
|         US|   45781324| RXO216LWUDV6O|B00FLTZ2ZS|     361297724|DNM Mountain Bike...|
|         US|   16699467|R3NMJF7EBMM19V|B007SZ4XJ4|     188745514|VeloChampion MLT1...|
|         US|   16299390|R2ZY0ZBDUO0XUY|B00NAINBM8|     754535833|Earthtrek [illegible] Fol...|
+-----------+-----------+--------------+----------+--------------+--------------------+-
only showing top 5 rows
```

```
# Filter DataFrame for helpful_votes ratio above or equal to 50%
```

```
df2 = df1.filter((df1.helpful_votes / df1.total_votes) >= 0.5)
df2.show(5)
```

```
+-----------+-----------+--------------+----------+--------------+--------------------+-
|marketplace|customer_id|     review_id|product_id|product_parent|       product_title|
+-----------+-----------+--------------+----------+--------------+--------------------+-
|         US|   30222858|R2FP3U4NHNFNL2|B00YPISPNC|     468413405|Stanley Classic V...|
|         US|   35677754|R1UUK1977O38MU|B00T8NEI3A|      32341693|Camping Wood Burn...|
|         US|   45781324| RXO216LWUDV6O|B00FLTZ2ZS|     361297724|DNM Mountain Bike...|
|         US|   16699467|R3NMJF7EBMM19V|B007SZ4XJ4|     188745514|VeloChampion MLT1...|
|         US|   16299390|R2ZY0ZBDUO0XUY|B00NAINBM8|     754535833|Earthtrekgear Fol...|
+-----------+-----------+--------------+----------+--------------+--------------------+-
only showing top 5 rows
```

```
#  DataFrame of paid vine
paid_df = df2.filter(df2.vine == 'Y')
paid_df.show(5)
```

```
+-----------+-----------+--------------+----------+--------------+--------------------+-
|marketplace|customer_id|     review_id|product_id|product_parent|       product_title|
+-----------+-----------+--------------+----------+--------------+--------------------+-
|         US|   43335941|R3KPC0NBUDASX3|B00R8KC02Q|     872035750|Thule EnRoute Tri...|
|         US|   36470546|R119P2A95GGXX4|B00NOYKVSK|     165748383|Wenzel Temp Contr...|
|         US|   50794278|R1HKIRME8AJ89Z|B00GK4LUXQ|     737005436|Klymit Inertia O ...|
|         US|   44173076|R3FY3GMBGOBR22|B00NFCFDRA|     585483297|Kelty Dualist 6 D...|
|         US|   43856165|R15KH3FBSVYGBU|B00NFCFIR0|     167877886|Kelty Tuck 22 Deg...|
+-----------+-----------+--------------+----------+--------------+--------------------+-
only showing top 5 rows
```

```
# DataFrame of unpaid vine
unpaid_df = df2.filter(df2.vine == 'N')
unpaid_df.show(5)
```

```
+-----------+-----------+--------------+----------+--------------+--------------------+-
|marketplace|customer_id|     review_id|product_id|product_parent|       product_title|
+-----------+-----------+--------------+----------+--------------+--------------------+-
|         US|   30222858|R2FP3U4NHNFNL2|B00YPISPNC|     468413405|Stanley Classic V...|
|         US|   35677754|R1UUK1977O38MU|B00T8NEI3A|      32341693|Camping Wood Burn...|
|         US|   45781324| RXO216LWUDV6O|B00FLTZ2ZS|     361297724|DNM Mountain Bike...|
|         US|   16699467|R3NMJF7EBMM19V|B007SZ4XJ4|     188745514|VeloChampion MLT1...|
|         US|   16299390|R2ZY0ZBDUO0XUY|B00NAINBM8|     754535833|Earthtrekgear Fol...|
+-----------+-----------+--------------+----------+--------------+--------------------+-
only showing top 5 rows
```

```
# Total number of paid reviews
total_paid_reviews = paid_df.count()
```

```
total_paid_reviews
```

```
107
```

```
# Paid 5-star reviews
paid_five_star_reviews = paid_df.filter(paid_df.star_rating == 5).count()
paid_five_star_reviews
```

```
56
```

```
# Paid 5-star reviews percentage
paid_five_star_percent = (paid_five_star_reviews / total_paid_reviews) * 100
paid_five_star_percent
```

```
52.336448598130836
```

```
# Unpaid total number of reviews
total_unpaid_reviews = unpaid_df.count()
total_unpaid_reviews
```

```
39869
```

```
# Unpaid 5-star reviews
unpaid_five_star_reviews = unpaid_df.filter(unpaid_df.star_rating == 5).count()
unpaid_five_star_reviews
```

```
21005
```

```
# Percentage Unpaid 5-star reviews
unpaid_five_star_percent = (unpaid_five_star_reviews / total_unpaid_reviews) * 100
unpaid_five_star_percent
```

```
52.68504351751988
```

✓ 0s completed at 11:21 AM ● ×