```
%pyspark
from pyspark import SparkFiles
url = "https://s3.amazonaws.com/amazon-reviews-pds/tsv/amazon_reviews_us_Video_Games_v1_00.tsv.gz"
spark.sparkContext.addFile(url)
video_games_df = spark.read.csv(SparkFiles.get("amazon_reviews_us_Video_Games_v1_00.tsv.gz"), sep="\t", header=True)
video_games_df.show()
```

```
+-----------+-----------+--------------+----------+--------------+--------------------+----------------+-----------+------------+-----------+----+-----------------
+------------+--------------------+-----------+
|marketplace|customer_id|     review_id|product_id|product_parent|       product_title|product_category|star_rating|helpful_votes|total_votes|vine|verified_purchase|     re
view_headline|         review_body|review_date|
+-----------+-----------+--------------+----------+--------------+--------------------+----------------+-----------+------------+-----------+----+-----------------
+------------+--------------------+-----------+
|         US|   12039526| RTIS3L2M1F5SM|B001CXYMFS|     737716809|Thrustmaster T-Fl...|     Video Games|          5|           0|          0|   N|                Y|an amaz
ing joysti...|Used this for Eli...| 2015-08-31|
|         US|    9636577| R1ZV7R40OLHKD|B00M920ND6|     569686175|Tonsee 6 buttons ...|     Video Games|          5|           0|          0|   N|                Y|Definit
ely a sile...|Loved it,  I didn...| 2015-08-31|
|         US|    2331478|R3BH071QLH8QMC|B0029CSOD2|      98937668|Hidden Mysteries:...|     Video Games|          1|           0|          1|   N|                Y|
One Star|poor quality work...| 2015-08-31|
|         US|   52495923|R127K9NTSXA2YH|B00GOOSV98|      23143350|GelTabz Performan...|     Video Games|          3|           0|          0|   N|                Y|good, b
ut could b...|nice, but tend to...| 2015-08-31|
|         US|   14533949|R32ZWUXDJPW27Q|B00Y074JOM|     821342511|Zero Suit Samus a...|     Video Games|          4|           0|          0|   N|                Y|   Grea
t but flawed.|Great amiibo, gre...| 2015-08-31|
|         US|    2377552|R3AQQ4YUKJWBA6|B002UBI6W6|     328764615|Psyclone Recharge...|     Video Games|          1|           0|          0|   N|                Y|
One Star|The remote consta...| 2015-08-31|
|         US|   17521011|R2F0POU5K6F73F|B008XHCLFO|      24234603|Protection for yo...|     Video Games|          5|           0|          0|   N|                Y|
A Must|I have a 2012-201...| 2015-08-31|
|         US|   19676307|R3VNR804HYSMR6|B00BRA9R6A|     682267517|   Nerf 3DS XL Armor|     Video Games|          5|           0|          0|   N|                Y|
Five Stars|Perfect, kids lov...| 2015-08-31|
|         US|     224068| R3GZTM72WA2QH|B009EPWJLA|     435241890|One Piece: Pirate...|     Video Games|          5|           0|          0|   N|                Y|
Five Stars|            Excelent| 2015-08-31|
|         US|   48467989| RNQOY62705W1K|B0000AV7GB|     256572651|Playstation 2 Dan...|     Video Games|          4|           0|          0|   N|                Y|
Four Stars|Slippery but expe...| 2015-08-31|
|         US|     106569|R1VTIA3JTYBY02|B00008KTNN|     384411423|Metal Arms: Glitc...|     Video Games|          5|           0|          0|   N|                N|
Five Stars|Love the game. Se...| 2015-08-31|
|         US|   48269642|R29DOU8791QZL8|B000A3IA0Y|     472622859|72 Pin Connector ...|     Video Games|          1|           0|          0|   N|                Y| Game w
ill get stuck|Does not fit prop...| 2015-08-31|
```

Interpreter: spark.pyspark. FINISHED Took 1 min 12 sec 20 millisec. Updated by ZGrinacoff on November 08 2019, 5:22:28 PM (PST)

```
%pyspark
# Filter by voters.
df2 = video_games_df.select(["star_rating", "helpful_votes", "total_votes", "vine", "verified_purchase"])
df2.show(10)

df3 = df2.filter(df2['total_votes'] >= 20)
df4 = df3.filter(df3["helpful_votes"]/df3["total_votes"] >= 0.5)
```

```
+-----------+-------------+-----------+----+-----------------+
|star_rating|helpful_votes|total_votes|vine|verified_purchase|
+-----------+-------------+-----------+----+-----------------+
|          5|            0|          0|   N|                Y|
|          5|            0|          0|   N|                Y|
|          1|            0|          1|   N|                Y|
|          3|            0|          0|   N|                Y|
|          4|            0|          0|   N|                Y|
|          1|            0|          0|   N|                Y|
|          5|            0|          0|   N|                Y|
|          5|            0|          0|   N|                Y|
|          5|            0|          0|   N|                Y|
|          4|            0|          0|   N|                Y|
+-----------+-------------+-----------+----+-----------------+
only showing top 10 rows
```

Interpreter: spark.pyspark. FINISHED Took 211 millisec. Updated by ZGrinacoff on November 08 2019, 5:27:49 PM (PST)

```
%pyspark
# Describe stats.
from pyspark.sql.functions import col, avg
paid_df = df4.filter(df4['vine']== 'Y')
unpaid_df = df4.filter(df4['vine']== 'N')

paid_df.describe().show()
unpaid_df.describe().show()
```

```
+-------+------------------+-----------------+------------------+----+-----------------+
|summary|       star_rating|     helpful_votes|       total_votes|vine|verified_purchase|
+-------+------------------+-----------------+------------------+----+-----------------+
|  count|                94|               94|                94|  94|               94|
|   mean| 4.202127659574468|54.59574468085106|61.787234042553195|null|             null|
| stddev|0.9791348741656414|65.26098459822538| 68.90976994895392|null|             null|
|    min|                 1|              111|               102|   Y|                N|
|    max|                 5|               97|                88|   Y|                N|
+-------+------------------+-----------------+------------------+----+-----------------+

+-------+------------------+------------------+------------------+-----+-----------------+
|summary|       star_rating|      helpful_votes|       total_votes| vine|verified_purchase|
+-------+------------------+------------------+------------------+-----+-----------------+
|  count|             40471|             40471|             40471|40471|            40471|
|   mean|  3.34765634651973|47.428405524943784|55.891057794470115| null|             null|
| stddev|1.6418850112078023|117.53763370687005|127.40280622961905| null|             null|
|    min|                 1|                10|               100|    N|                N|
|    max|                 5|               999|               999|    N|                Y|
+-------+------------------+------------------+------------------+-----+-----------------+
```

Interpreter: spark.pyspark. FINISHED Took 24 sec 662 millisec. Updated by ZGrinacoff on November 08 2019, 5:28:45 PM (PST)

Interpreter: spark.pyspark. FINISHED Took 34 sec 662 millisec. Updated by ZGrinacoff on November 08 2019, 5:28:45 PM (PST)

```
%pyspark
#  Determine the percentage of five-star reviews among Vine reviews
paid_five_star_number = paid_df[paid_df['star_rating']== 5].count()
paid_number = paid_df.count()
percentage_five_star_vine = float(paid_five_star_number) / float(paid_number)
print(paid_number)
print(paid_five_star_number)
print(percentage_five_star_vine)
```

```
94
48
0.510638297872
```

Interpreter: spark.pyspark. FINISHED Took 31 sec 656 millisec. Updated by ZGrinacoff on November 08 2019, 5:33:09 PM (PST)

```
%pyspark
#  Determine the percentage of five-star reviews among non-Vine reviews.
unpaid_five_star_number = unpaid_df[unpaid_df['star_rating']== 5].count()
unpaid_number = unpaid_df.count()
percentage_five_star_non_vine = float(unpaid_five_star_number) / float(unpaid_number)
print(unpaid_number)
print(unpaid_five_star_number)
print(float(unpaid_five_star_number) / float(unpaid_number))
```

```
40471
15663
0.387017864644
```

Interpreter: spark.pyspark. FINISHED Took 31 sec 605 millisec. Updated by ZGrinacoff on November 08 2019, 5:33:44 PM (PST)

Interpreter: spark.