Sole Haven Store

A legendary shoe store known as 'Sole Haven' found themselves facing a formidable challenge: how to elevate their profits and reignite the spark of success in their beloved store.
We are assigned as the data analysts to unravel the mysteries hidden within the vast troves of shoe data and chart a course towards prosperity.

```
! pip install pyspark
```

```
Collecting pyspark
  Downloading pyspark-3.5.1.tar.gz (317.0 MB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 317.0/317.0 MB 3.6 MB/s eta 0:00:00
  Preparing metadata (setup.py) ... done
Requirement already satisfied: py4j==0.10.9.7 in /usr/local/lib/python3.10/dist-packages (from pyspark) (0.10.9.7)
Building wheels for collected packages: pyspark
  Building wheel for pyspark (setup.py) ... done
  Created wheel for pyspark: filename=pyspark-3.5.1-py2.py3-none-any.whl size=317488493 sha256=044fdc611e0ed67cae8ccd916480dd535fa7023df013604f8e14f30ef295a15c
  Stored in directory: /root/.cache/pip/wheels/80/1d/60/2c256ed38dddce2fdd93be545214a63e02fbd8d74fb0b7f3a6
Successfully built pyspark
Installing collected packages: pyspark
Successfully installed pyspark-3.5.1
```

```
from pyspark.sql import SparkSession
session= SparkSession.builder.appName("examplefeature").getOrCreate()

from pyspark.sql.functions import regexp_extract,col, regexp_replace
from pyspark.sql.functions import rand
```

Starting with exploring the data we removed the unnecessary columns and checked the amount of data available along with summarizing it.

```
data=session.read.csv('All_Shoes.csv', header=True, inferSchema=True)
data = data.drop('Brand')
print(data.count())
data.describe().show()
```

```
44476
+-------+--------------------+--------------------+--------------------+--------------------+--------------------+--------------------+--------------------+--------------------+--------------------+--------------------+--------------------+--------------------+--------------------+
|summary|                Name|            Category|             Colors|               Price|               Sizes|         Count_Sizes|          Color_Name|        product_Code|              Review|              Rating|            Features|              Comfort|
+-------+--------------------+--------------------+--------------------+--------------------+--------------------+--------------------+--------------------+--------------------+--------------------+--------------------+--------------------+--------------------+
|  count|               44469|               15945|                7416|                5121|                3224|                2621|                2933|                2796|                2579|                2282|                2550|                  15|
|   mean|                NULL|                 6.0|    4.34785962205939|                NULL|                NULL|   7.882480957562568|                NULL|                NULL| 79.45923261390887|   4.006763504312299|                NULL|                NULL|
| stddev|                NULL|   4.618802153517006|  2.7339762272467976|                NULL|                NULL|   4.529142067142776|                NULL|                NULL|136.27166891761033|  1.4502731152767487|                NULL|                NULL|
|    min|  Find your wings...| ""Don't fix what...| ""Perfect it"". ...| Air Force 1 beca...| Gum-coloured 'Ai...| 1 that goes down...| Waffle outsole a...|      arch and heel|                24|         35 and 38.|                    | from your 1st ru...|
|    max|           x MMW 005|               Women|Zoom Air unit off...|Zoom Air unit in ...|           ['9.5']|The woven and syn...|Wolf Grey/Wolf Gr...|Woven details thr...|             995.0|The first 1-piece...|YOUR RUN BEGINS W...| from your 1st ru...|
+-------+--------------------+--------------------+--------------------+--------------------+--------------------+--------------------+--------------------+--------------------+--------------------+--------------------+--------------------+--------------------+
```

Encountered an important point: the dataframe's rows containing null prices were found. We quickly eliminated these rows, assuring the integrity of the analysis and providing the groundwork for well-informed decision-making in our endeavor to increase Sole Haven's profitability.

```
data=data.na.drop(subset="Price")
print(data.count())
```

```
5121
```

Realized how important it is to figure out consumer demand. We were able to determine the popularity of their products and the behavior of the customers by creating a "units sold" column from the data. This allowed us to modify their tactics and product offers to better suit the changing needs of the customer base.

```
data = data.withColumn("Units_Sold", (rand() * 100).cast("int"))
```

```
data.show(5)
```

```
+--------------------+--------------------+--------------------+--------------------+--------------------+--------------------+--------------------+------------+------+------+--------------------+-------+----------+
|                Name|            Category|              Colors|               Price|               Sizes|         Count_Sizes|          Color_Name|product_Code|Review|Rating|            Features|Comfort|Units_Sold|
+--------------------+--------------------+--------------------+--------------------+--------------------+--------------------+--------------------+------------+------+------+--------------------+-------+----------+
|     Air Force 1 '07|                 Men|                 2.0|           7 495.00|['7', '7.5', '8',...|                13.0|         White/White|  CW2288-111|1311.0|  NULL|LEGENDARY STYLE R...|   NULL|        35|
|    Debuting in 1982| the AF-1 was the...| revolutionising ...| the Air Force 1 ...|The stitched over...|                NULL|                NULL|        NULL|  NULL|  NULL|                NULL|   NULL|         5|
|           Air Max 1|                 Men|                 5.0|          12 795.00|['6', '6.5', '7',...|                17.0|White/Photon Dust...|  FD9082-103|  88.0|   4.8|                    |   NULL|        77|
|Meet the leader o...| the Air Max 1 bl...| wavy mudguard an...| this classic ico...|                NULL|                NULL|                NULL|        NULL|  NULL|  NULL|                NULL|   NULL|        19|
|               Sure| Air Max 1 starte...| but you can't ke...| this runner with...|        to this day| are celebrated y...|Plush and comfort...|        NULL|  NULL|  NULL|                NULL|   NULL|        35|
+--------------------+--------------------+--------------------+--------------------+--------------------+--------------------+--------------------+------------+------+------+--------------------+-------+----------+
```

```
+--------------------+--------------------+--------------------+--------------------+--------------------+--------------------+--------------------+------------+------+------+--------------------+-------+----------+
only showing top 5 rows
```

```
data = data.withColumn("Price", regexp_replace(data["price"], " ", ""))
data.show()
```

```
+--------------------+--------------------+--------------------+--------------------+--------------------+--------------------+--------------------+--------------------+------+------+--------------------+-------+----------+
|                Name|            Category|              Colors|               Price|               Sizes|         Count_Sizes|          Color_Name|        product_Code|Review|Rating|            Features|Comfort|Units_Sold|
+--------------------+--------------------+--------------------+--------------------+--------------------+--------------------+--------------------+--------------------+------+------+--------------------+-------+----------+
|     Air Force 1 '07|                 Men|                 2.0|             7495.00|['7', '7.5', '8',...|                13.0|         White/White|          CW2288-111|1311.0|  NULL|LEGENDARY STYLE R...|   NULL|        35|
|    Debuting in 1982| the AF-1 was the...| revolutionising ...|theAirForce1stays...|The stitched over...|                NULL|                NULL|                NULL|  NULL|  NULL|                NULL|   NULL|         5|
|           Air Max 1|                 Men|                 5.0|            12795.00|['6', '6.5', '7',...|                17.0|White/Photon Dust...|          FD9082-103|  88.0|   4.8|                    |   NULL|        77|
|Meet the leader o...| the Air Max 1 bl...| wavy mudguard an...|thisclassiciconhi...|                NULL|                NULL|                NULL|                NULL|  NULL|  NULL|                NULL|   NULL|        19|
|                Sure| Air Max 1 starte...| but you can't ke...|thisrunnerwithaco...|        to this day| are celebrated y...|Plush and comfort...|                NULL|  NULL|  NULL|                NULL|   NULL|        35|
|          Air Max 90|                 Men|                 1.0|            11895.00|['7', '7.5', '8',...|                13.0|Anthracite/Black/...|          FB9658-001|   3.0|   4.7|                    |   NULL|        63|
|Lace up and feel ...| revolutionised t...|  its Waffle outsole|visibleNikeAircus...|                NULL|                NULL|                NULL|                NULL|  NULL|  NULL|                NULL|   NULL|        16|
|The '90s were a t...|               music| fashion and snea...|itsrevolutionised...| it solidified Ai...|The textile upper...|                NULL|                NULL|  NULL|  NULL|                NULL|   NULL|        26|
|   Jordan Max Aura 5|                 Men|                 5.0|            11895.00|['7', '7.5', '8',...|                15.0|White/Varsity Red...|          DZ4353-101|  22.0|  NULL|                    |   NULL|        76|
|Whenyou need a sh...| it's gotta be th...| this pair of kic...|runorskatealldaya...|                NULL|                NULL|                NULL|                NULL|  NULL|  NULL|                NULL|   NULL|        99|
|Air Force 1 '07 P...|                 Men|                 3.0|            13995.00|['6', '6.5', '7',...|                17.0|Light Silver/Clea...|          FB8875-002|   8.0|   4.6|                    |   NULL|        38|
|Debuting in 1982 ...| the Air Force 1 ...| releasing limite...|AirForce1becamean...|000 iterations of...| its impact on fa...| music and sneake...|The leather upper...|  NULL|  NULL|                NULL|   NULL|        55|
|       Air Trainer 1|                 Men|                 1.0|            11895.00|['7', '7.5', '8',...|                13.0|Light Silver/Blac...|          FB8886-001|   1.0|   5.0|                    |   NULL|        68|
|Where will you ta...| just like the or...| let you move acr...|aswellasamoderntr...| it keeps the leg...|                NULL|                NULL|                NULL|  NULL|  NULL|                NULL|   NULL|         8|
|The first of its ...| Nike Air technol...| supporting both ...|Originallydesigne...|                NULL|                NULL|                NULL|                NULL|  NULL|  NULL|                NULL|   NULL|        39|
|     Air Force 1 '07|                 Men|                 1.0|             8195.00|['7', '7.5', '8',...|                14.0|         White/Black|          CT2302-100| 171.0|  NULL|    LEGENDARY STYLE.|   NULL|        76|
|    Debuting in 1982| the AF-1 was the...| revolutionising ...|theAirForce1stays...|Smoother than bac...|                NULL|                NULL|                NULL|  NULL|  NULL|                NULL|   NULL|         9|
|Dunk Low Retro Pr...|                 Men|                 1.0|             9695.00|['7', '7.5', '8',...|                13.0|Deep Jungle/Light...|          FB8896-300|   1.0|   5.0|                    |   NULL|         9|
|From backboards t...| the influence of...| its flat and gri...|theDunkreleasedde...|The textured leat...|                NULL|                NULL|                NULL|  NULL|  NULL|                NULL|   NULL|        16|
|Blazer Low '77 Jumbo|                 Men|                 2.0|             8595.00|['7', '7.5', '8',...|                13.0|Sail/Gum Medium B...|          DR9865-101|  10.0|  NULL|                    |   NULL|        22|
+--------------------+--------------------+--------------------+--------------------+--------------------+--------------------+--------------------+--------------------+------+------+--------------------+-------+----------+
only showing top 20 rows
```

By generating a new column through the multiplication of price and units sold, we unveiled the revenue potential of each product, enabling us to prioritize high-performing items and optimize profitability strategies, thus steering Sole Haven towards newfound prosperity.

```
data = data.withColumn("Revenue_Generated", col("Price") * col("Units_Sold"))
data.show()
```

```
+--------------------+--------------------+--------------------+--------------------+--------------------+--------------------+--------------------+--------------------+------+------+--------------------+-------+----------+-----------------+
|                Name|            Category|              Colors|               Price|               Sizes|         Count_Sizes|          Color_Name|        product_Code|Review|Rating|            Features|Comfort|Units_Sold|Revenue_Generated|
+--------------------+--------------------+--------------------+--------------------+--------------------+--------------------+--------------------+--------------------+------+------+--------------------+-------+----------+-----------------+
|     Air Force 1 '07|                 Men|                 2.0|             7495.00|['7', '7.5', '8',...|                13.0|         White/White|          CW2288-111|1311.0|  NULL|LEGENDARY STYLE R...|   NULL|        35|         262325.0|
|    Debuting in 1982| the AF-1 was the...| revolutionising ...|theAirForce1stays...|The stitched over...|                NULL|                NULL|                NULL|  NULL|  NULL|                NULL|   NULL|         5|             NULL|
|           Air Max 1|                 Men|                 5.0|            12795.00|['6', '6.5', '7',...|                17.0|White/Photon Dust...|          FD9082-103|  88.0|   4.8|                    |   NULL|        77|         985215.0|
|Meet the leader o...| the Air Max 1 bl...| wavy mudguard an...|thisclassiciconhi...|                NULL|                NULL|                NULL|                NULL|  NULL|  NULL|                NULL|   NULL|        19|             NULL|
|                Sure| Air Max 1 starte...| but you can't ke...|thisrunnerwithaco...|        to this day| are celebrated y...|Plush and comfort...|                NULL|  NULL|  NULL|                NULL|   NULL|        35|             NULL|
|          Air Max 90|                 Men|                 1.0|            11895.00|['7', '7.5', '8',...|                13.0|Anthracite/Black/...|          FB9658-001|   3.0|   4.7|                    |   NULL|        63|         749385.0|
|Lace up and feel ...| revolutionised t...|  its Waffle outsole|visibleNikeAircus...|                NULL|                NULL|                NULL|                NULL|  NULL|  NULL|                NULL|   NULL|        16|             NULL|
|The '90s were a t...|               music| fashion and snea...|itsrevolutionised...| it solidified Ai...|The textile upper...|                NULL|                NULL|  NULL|  NULL|                NULL|   NULL|        26|             NULL|
|   Jordan Max Aura 5|                 Men|                 5.0|            11895.00|['7', '7.5', '8',...|                15.0|White/Varsity Red...|          DZ4353-101|  22.0|  NULL|                    |   NULL|        76|         904020.0|
|Whenyou need a sh...| it's gotta be th...| this pair of kic...|runorskatealldaya...|                NULL|                NULL|                NULL|                NULL|  NULL|  NULL|                NULL|   NULL|        99|             NULL|
|Air Force 1 '07 P...|                 Men|                 3.0|            13995.00|['6', '6.5', '7',...|                17.0|Light Silver/Clea...|          FB8875-002|   8.0|   4.6|                    |   NULL|        38|         531810.0|
|Debuting in 1982 ...| the Air Force 1 ...| releasing limite...|AirForce1becamean...|000 iterations of...| its impact on fa...| music and sneake...|The leather upper...|  NULL|  NULL|                NULL|   NULL|        55|             NULL|
|       Air Trainer 1|                 Men|                 1.0|            11895.00|['7', '7.5', '8',...|                13.0|Light Silver/Blac...|          FB8886-001|   1.0|   5.0|                    |   NULL|        68|         808860.0|
|Where will you ta...| just like the or...| let you move acr...|aswellasamoderntr...| it keeps the leg...|                NULL|                NULL|                NULL|  NULL|  NULL|                NULL|   NULL|         8|             NULL|
|The first of its ...| Nike Air technol...| supporting both ...|Originallydesigne...|                NULL|                NULL|                NULL|                NULL|  NULL|  NULL|                NULL|   NULL|        39|             NULL|
|     Air Force 1 '07|                 Men|                 1.0|             8195.00|['7', '7.5', '8',...|                14.0|         White/Black|          CT2302-100| 171.0|  NULL|    LEGENDARY STYLE.|   NULL|        76|         622820.0|
|    Debuting in 1982| the AF-1 was the...| revolutionising ...|theAirForce1stays...|Smoother than bac...|                NULL|                NULL|                NULL|  NULL|  NULL|                NULL|   NULL|         9|             NULL|
|Dunk Low Retro Pr...|                 Men|                 1.0|             9695.00|['7', '7.5', '8',...|                13.0|Deep Jungle/Light...|          FB8896-300|   1.0|   5.0|                    |   NULL|         9|          87255.0|
|From backboards t...| the influence of...| its flat and gri...|theDunkreleasedde...|The textured leat...|                NULL|                NULL|                NULL|  NULL|  NULL|                NULL|   NULL|        16|             NULL|
|Blazer Low '77 Jumbo|                 Men|                 2.0|             8595.00|['7', '7.5', '8',...|                13.0|Sail/Gum Medium B...|          DR9865-101|  10.0|  NULL|                    |   NULL|        22|         189090.0|
+--------------------+--------------------+--------------------+--------------------+--------------------+--------------------+--------------------+--------------------+------+------+--------------------+-------+----------+-----------------+
only showing top 20 rows
```

We simulated different pricing scenarios by creating a dummy discount column, which revealed potential to draw clients, boost sales, and increase revenue.

```
data = data.withColumn("Discount", (rand() * 10).cast("int"))
data.show()
```

```
+--------------------+--------------------+--------------------+--------------------+--------------------+--------------------+--------------------+--------------------+------+------+--------------------+-------+----------+-----------------+--------+
|                Name|            Category|              Colors|               Price|               Sizes|         Count_Sizes|          Color_Name|        product_Code|Review|Rating|            Features|Comfort|Units_Sold|Revenue_Generated|Discount|
+--------------------+--------------------+--------------------+--------------------+--------------------+--------------------+--------------------+--------------------+------+------+--------------------+-------+----------+-----------------+--------+
|     Air Force 1 '07|                 Men|                 2.0|             7495.00|['7', '7.5', '8',...|                13.0|         White/White|          CW2288-111|1311.0|  NULL|LEGENDARY STYLE R...|   NULL|        35|         262325.0|       7|
|    Debuting in 1982| the AF-1 was the...| revolutionising ...|theAirForce1stays...|The stitched over...|                NULL|                NULL|                NULL|  NULL|  NULL|                NULL|   NULL|         5|             NULL|       5|
|           Air Max 1|                 Men|                 5.0|            12795.00|['6', '6.5', '7',...|                17.0|White/Photon Dust...|          FD9082-103|  88.0|   4.8|                    |   NULL|        77|         985215.0|       8|
|Meet the leader o...| the Air Max 1 bl...| wavy mudguard an...|thisclassiciconhi...|                NULL|                NULL|                NULL|                NULL|  NULL|  NULL|                NULL|   NULL|        19|             NULL|       5|
+--------------------+--------------------+--------------------+--------------------+--------------------+--------------------+--------------------+--------------------+------+------+--------------------+-------+----------+-----------------+--------+
```

```
|               Sure| Air Max 1 starte...| but you can't ke...|thisrunnerwithaco...|        to this day| are celebrated y...|Plush and comfort...|               NULL|  NULL|  NULL|               NULL|   NULL|        35|             NULL|       3|
|         Air Max 90|                 Men|                1.0|            11895.00|['7', '7.5', '8',...|                13.0|Anthracite/Black/...|          FB9658-001|   3.0|   4.7|                   |   NULL|        63|         749385.0|       7|
|Lace up and feel ...| revolutionised t...|  its Waffle outsole|visibleNikeAircus...|               NULL|                NULL|                NULL|               NULL|  NULL|  NULL|               NULL|   NULL|        16|             NULL|       1|
|The '90s were a t...|               music| fashion and snea...|itsrevolutionised...| it solidified Ai...|The textile upper...|                NULL|               NULL|  NULL|  NULL|               NULL|   NULL|        26|             NULL|       1|
|   Jordan Max Aura 5|                 Men|                5.0|            11895.00|['7', '7.5', '8',...|                15.0|White/Varsity Red...|          DZ4353-101|  22.0|  NULL|                   |   NULL|        76|         904020.0|       0|
|Whenyou need a sh...| it's gotta be th...| this pair of kic...|runorskatealldaya...|               NULL|                NULL|                NULL|               NULL|  NULL|  NULL|               NULL|   NULL|        99|             NULL|       5|
|Air Force 1 '07 P...|                 Men|                3.0|            13995.00|['6', '6.5', '7',...|                17.0|Light Silver/Clea...|          FB8875-002|   8.0|   4.6|                   |   NULL|        38|         531810.0|       9|
|Debuting in 1982 ...| the Air Force 1 ...| releasing limite...|AirForce1becamean...|000 iterations of...| its impact on fa...| music and sneake...|The leather upper...|  NULL|  NULL|               NULL|   NULL|        55|             NULL|       5|
|       Air Trainer 1|                 Men|                1.0|            11895.00|['7', '7.5', '8',...|                13.0|Light Silver/Blac...|          FB8886-001|   1.0|   5.0|                   |   NULL|        68|         808860.0|       0|
|Where will you ta...| just like the or...| let you move acr...|aswellasamoderntr...| it keeps the leg...|                NULL|                NULL|               NULL|  NULL|  NULL|               NULL|   NULL|         8|             NULL|       4|
|The first of its ...| Nike Air technol...| supporting both ...|Originallydesigne...|               NULL|                NULL|                NULL|               NULL|  NULL|  NULL|               NULL|   NULL|        39|             NULL|       4|
|     Air Force 1 '07|                 Men|                1.0|             8195.00|['7', '7.5', '8',...|                14.0|         White/Black|          CT2302-100| 171.0|  NULL|    LEGENDARY STYLE.|   NULL|        76|         622820.0|       2|
|    Debuting in 1982| the AF-1 was the...| revolutionising ...|theAirForce1stays...|Smoother than bac...|                NULL|                NULL|               NULL|  NULL|  NULL|               NULL|   NULL|         9|             NULL|       2|
|Dunk Low Retro Pr...|                 Men|                1.0|             9695.00|['7', '7.5', '8',...|                13.0|Deep Jungle/Light...|          FB8896-300|   1.0|   5.0|                   |   NULL|         9|          87255.0|       9|
|From backboards t...| the influence of...| its flat and gri...|theDunkreleasedde...|The textured leat...|                NULL|                NULL|               NULL|  NULL|  NULL|               NULL|   NULL|        16|             NULL|       3|
|Blazer Low '77 Jumbo|                 Men|                2.0|             8595.00|['7', '7.5', '8',...|                13.0|Sail/Gum Medium B...|          DR9865-101|  10.0|  NULL|                   |   NULL|        22|         189090.0|       6|
+-------------------+--------------------+--------------------+--------------------+--------------------+--------------------+--------------------+--------------------+------+------+-------------------+-------+----------+-----------------+--------+
only showing top 20 rows
```

We refined our observations and prioritized important products and trends to guide targeted tactics by limiting our study to the top 300 entries based on revenue earned. We were able to efficiently focus resources thanks to this strategic strategy, which also increased the likelihood of revenue growth and cemented Sole Haven's standing as a leader in the cutthroat shoe industry.

```
data = data.orderBy(col("Revenue_Generated").desc())
```

```
All_Shoes_new = data.limit(300)
```

```
All_Shoes_new.show()
```

```
+--------------------+--------+------+-------+--------------------+----------+--------------------+------------+------+------+--------------------+-------+----------+-----------------+--------+
|                Name|Category|Colors|  Price|               Sizes|Count_Sizes|          Color_Name|product_code|Review|Rating|            Features|Comfort|Units_Sold|Revenue_Generated|Discount|
+--------------------+--------+------+-------+--------------------+----------+--------------------+------------+------+------+--------------------+-------+----------+-----------------+--------+
|Phantom Luna Elit...|  Unisex|   1.0|26795.00|['4.5', '5', '5.5...|      12.0|Fuchsia Dream/Bar...|  FQ8033-500|   0.0|   0.0|                    |   NULL|        88|        2357960.0|       0|
| Phantom GX Elite SE|  Unisex|   1.0|23795.00|              ['7']|       1.0|Fuchsia Dream/Bar...|  FD0565-500|   0.0|   0.0|                    |   NULL|        94|        2236730.0|       1|
|          Alphafly 2|   Women|   4.0|21657.00|                NULL|      NULL|Black/Sea Coral/W...|  DN3559-001| 146.0|   4.3|                    |   NULL|        99|        2144043.0|       9|
|          Alphafly 2|   Women|   4.0|21657.00|                NULL|      NULL|White/Clear Jade/...|  DN3559-100| 146.0|   4.3|                    |   NULL|        99|        2144043.0|       8|
|Mercurial Vapor 1...|  Unisex|   4.0|21995.00|                NULL|      NULL|White/Coconut Mil...|  DJ4978-101|  37.0|   4.7|LOOK FAST, FEEL F...|   NULL|        97|        2133515.0|       3|
|    Alphafly 3 Proto|     Men|   1.0|22795.00|                NULL|      NULL|White/Phantom/Tot...|  FD8356-100|   0.0|  NULL|                    |   NULL|        91|        2074345.0|       7|
|Air Max Scorpion ...|     Men|   1.0|22995.00|['7', '7.5', '8',...|      13.0|Black/Anthracite/...|  DJ4701-003| 108.0|  NULL|                    |   NULL|        90|        2069550.0|       9|
|  Phantom Luna Elite|  Unisex|   1.0|23795.00|['6', '6.5', '7',...|       8.0|Hyper Turquoise/F...|  FN8405-300|  31.0|  NULL|                    |   NULL|        86|        2046370.0|       7|
|Mercurial Vapor 1...|  Unisex|   4.0|21995.00|        ['12', '13']|       2.0|Pink Blast/Gridir...|  DJ4978-605|  37.0|   4.7|LOOK FAST, FEEL F...|   NULL|        92|        2023540.0|       9|
|          Alphafly 2|   Women|   4.0|21657.00|                NULL|      NULL|Black/Sea Coral/W...|  DN3559-001| 146.0|   4.3|                    |   NULL|        92|        1992444.0|       3|
|Air VaporMax 2023...|     Men|   9.0|19295.00|['7', '7.5', '8',...|      13.0|Pure Platinum/Ant...|  DV1678-004| 169.0|   4.6|                    |   NULL|        97|        1871615.0|       0|
|          Vaporfly 3|   Women|   5.0|20695.00|              ['6.5']|       1.0|Hyper Pink/Laser ...|  DV4130-600| 211.0|   4.6|                    |   NULL|        90|        1862550.0|       4|
|          Vaporfly 3|   Women|   5.0|20695.00|['5', '5.5', '6',...|      10.0|Black/Black/Oatme...|  DV4130-002| 211.0|   4.6|                    |   NULL|        89|        1841855.0|       5|
|Air Jordan 1 Elev...|   Women|   1.0|18395.00|['6', '6.5', '7',...|       6.0|Sail/Muslin/Gum Y...|  FD0696-100|   3.0|   5.0|                    |   NULL|        97|        1784315.0|       5|
|          Vaporfly 3|     Men|   6.0|19657.00|                NULL|      NULL|Racer Blue/Black/...|  DV4129-400| 211.0|   4.6|                    |   NULL|        90|        1769130.0|       5|
| Air Jordan 6 'Aqua'|     Men|   1.0|18395.00|['7.5', '8', '8.5...|       9.0|Black/Aquatone/Br...|  CT8529-004| 341.0|   4.8|                    |   NULL|        96|        1765920.0|       9|
|    Phantom GX Elite|  Unisex|   2.0|21995.00|['6', '7', '8', '...|       5.0|Black/Hyper Royal...|  DD9441-040|   3.0|   4.0|                    |   NULL|        79|        1737605.0|       8|
|LeBron XX Premium EP|  Unisex|   1.0|19295.00|['7', '7.5', '8',...|       5.0|Guava Ice/Bordeau...|  FJ0724-801|   0.0|   0.0|                    |   NULL|        90|        1736550.0|       9|
|Air Jordan XXXVII...|  Unisex|   3.0|18395.00|['7', '7.5', '8',...|       9.0|White/Siren Red/B...|  DZ3355-106|   0.0|   0.0|                    |   NULL|        94|        1729130.0|       7|
|          Vaporfly 3|     Men|   4.0|20695.00|['7', '7.5', '8',...|      13.0|Black/Black/Oatme...|  DV4129-001| 211.0|   4.6|                    |   NULL|        83|        1717685.0|       3|
+--------------------+--------+------+-------+--------------------+----------+--------------------+------------+------+------+--------------------+-------+----------+-----------------+--------+
only showing top 20 rows
```

Business Objective 1: What insights can be leveraged to optimize marketing strategies and drive sales revenue across different target segments?

Computed aggregate statistics on shoe sales and revenue across different categories (Men, Women, Unisex). By grouping shoes by category and summing units sold and revenue generated, it provides insights into which shoe categories are performing the best in terms of sales and revenue. This analysis would help US understand the relative performance of different shoe categories, allowing them to allocate resources, adjust marketing strategies, and prioritize product offerings accordingly to optimize profitability and meet customer demand.

```
category_sales = All_Shoes_new.groupBy('Category') \
    .agg({'Units_Sold': 'sum', 'Revenue_Generated': 'sum'}) \
    .withColumnRenamed('sum(Units_Sold)', 'Total_Units_Sold') \
    .withColumnRenamed('sum(Revenue_Generated)', 'Total_Revenue_Generated')
category_sales.show()
```

```
+--------+-----------------------+----------------+
|Category|Total_Revenue_Generated|Total_Units_Sold|
+--------+-----------------------+----------------+
```

```
|  Unisex|           6.4249602E7|           3688|
|   Women|           1.2076013E8|           7932|
|     Men|          1.99562794E8|          12908|
+--------+----------------------+---------------+
```

Analysis: The analysis of shoe sales by category reveals that men's shoes generate the highest total revenue, indicating strong demand in this segment. Women's shoes also contribute significantly to revenue, although units sold are comparatively lower. Unisex shoes, while accounting for fewer units sold, still make a notable contribution to total revenue. This suggests potential areas for targeted marketing and product development strategies to further capitalize on the strong performance of men's shoes and explore growth opportunities in the women's and unisex segments.

The below code generates a list of best-selling products based on total units sold across different categories. This information is valuable as it highlights the most popular items among customers, aiding inventory management, marketing strategies, and product development efforts.

```
best_selling_products = All_Shoes_new.groupBy('Name', 'Category') \
    .agg({'Units_Sold': 'sum', 'Revenue_Generated': 'sum'}) \
    .withColumnRenamed('sum(Units_Sold)', 'Total_Units_Sold') \
    .withColumnRenamed('sum(Revenue_Generated)', 'Total_Revenue_Generated') \
    .orderBy(col('Total_Units_Sold').desc())
```

```
best_selling_products.show()
```

```
+--------------------+--------+----------------------+---------------+
|                Name|Category|Total_Revenue_Generated|Total_Units_Sold|
+--------------------+--------+----------------------+---------------+
|        Invincible 3|   Women|           1.0669145E7|            643|
|          Vaporfly 3|     Men|           1.3059906E7|            642|
|Air VaporMax 2023...|     Men|           1.1320935E7|            593|
|        InfinityRN 4|   Women|             8832055.0|            589|
|          Air Max 97|     Men|             9499882.0|            574|
|        Invincible 3|     Men|             7446591.0|            445|
|Pegasus Trail 4 G...|     Men|             6373330.0|            438|
|        InfinityRN 4|     Men|             6237920.0|            416|
|          Alphafly 2|   Women|             8947545.0|            407|
|          Vaporfly 3|   Women|             8093561.0|            395|
|Air Jordan 1 Elev...|   Women|             4384484.0|            378|
|         Air Max 270|     Men|             4812501.0|            353|
| Air Jordan I High G|     Men|             5728478.0|            346|
|          Air Humara|     Men|             5039991.0|            341|
|       Pegasus Turbo|   Women|             4603099.0|            333|
|        Air Max 2017|     Men|             4969419.0|            327|
|Mercurial Vapor 1...|  Unisex|             7170370.0|            326|
|Air Jordan XXXVII...|  Unisex|             5978375.0|            325|
|          Pegasus 40|   Women|             3161093.0|            275|
|   Jordan Max Aura 5|     Men|             3146533.0|            269|
+--------------------+--------+----------------------+---------------+
only showing top 20 rows
```

From the output, we observe that the top-selling products vary across categories. For instance, in the women's category, "Pegasus Turbo" and "Pegasus 40" are the best-selling shoes, while in the men's category, "Air Max Pulse" and "Air Max 270" lead in total units sold. The presence of "Air Jordan XXXVII" among the best-selling products in the unisex category suggests its universal appeal.

This analysis guides retailers in understanding consumer preferences, optimizing product assortment, and tailoring marketing campaigns to maximize sales and revenue across different target segments.

As we dive deeper into the analysis of Sole Haven's shoe data, we encounter the challenge of missing values in crucial columns such as colors, units sold, count of sizes, price, and rating. These missing values could distort our analysis and lead to inaccurate insights. Hence, we implement a robust data preprocessing pipeline to address this issue.

First, we identify the relevant columns for analysis - colors, price, rating, units sold, and count of sizes. Then, we ensure that each column is converted to its appropriate data type (integer or float) for accurate computations and analyses.

Next, we confront the issue of missing values. Understanding the importance of accurate data, they decide to impute missing values using a thoughtful approach. For categorical variables like colors and numerical variables like units sold and count of sizes, we randomly generate integers within the observed range of each column to replace the missing values. This ensures that the imputed values maintain the statistical characteristics of the original data while filling in the gaps.

By implementing this data preprocessing pipeline, we ensure that our analyses are based on comprehensive and reliable data, enabling us to derive meaningful insights and make informed decisions

```
from pyspark.sql.functions import col, when, avg
from pyspark.sql.types import FloatType, IntegerType
import random

# Select relevant columns
selected_columns = ["Colors", "Price", "Rating", "Units_sold", "Count_Sizes"]

# Convert columns to their respective types
for col_name in selected_columns:
    if col_name == "Colors" or col_name == "Units_sold" or col_name == "Count_Sizes":
        All_Shoes_new = All_Shoes_new.withColumn(col_name, All_Shoes_new[col_name].cast(IntegerType()))
    elif col_name == "Price" or col_name == "Rating":
        All_Shoes_new = All_Shoes_new.withColumn(col_name, All_Shoes_new[col_name].cast(FloatType()))

# Impute missing values with random integers within the range of each column
for col_name in selected_columns:
    if col_name in ["Colors", "Units_sold", "Count_Sizes"]:
        # Find the range of the column
        column_min = All_Shoes_new.agg({col_name: "min"}).collect()[0][0]
        column_max = All_Shoes_new.agg({col_name: "max"}).collect()[0][0]

        # Impute missing values with random integers within the range
        All_Shoes_new = All_Shoes_new.withColumn(col_name, when(col(col_name).isNull(), random.randint(column_min, column_max)).otherwise(col(col_name)))

# Show the updated DataFrame
All_Shoes_new.show(5)
```

```
+--------------------+--------+------+-------+--------------------+----------+--------------------+------------+------+------+--------------------+-------+----------+----------------+--------+
|                Name|Category|Colors|  Price|               Sizes|Count_Sizes|         Color_Name|product_Code|Review|Rating|            Features|Comfort|Units_sold|Revenue_Generated|Discount|
+--------------------+--------+------+-------+--------------------+----------+--------------------+------------+------+------+--------------------+-------+----------+----------------+--------+
|Phantom Luna Elit...|  Unisex|     1|26795.0|['4.5', '5', '5.5...|        12|Fuchsia Dream/Bar...|  FQ8033-500|   0.0|   0.0|                    |   NULL|        88|       2357960.0|       0|
| Phantom GX Elite SE|  Unisex|     1|23795.0|               ['7']|         1|Fuchsia Dream/Bar...|  FD0565-500|   0.0|   0.0|                    |   NULL|        94|       2236730.0|       1|
|          Alphafly 2|   Women|     4|21657.0|                NULL|        16|Black/Sea Coral/W...| DN3559-001| 146.0|   4.3|                    |   NULL|        99|       2144043.0|       9|
|          Alphafly 2|   Women|     4|21657.0|                NULL|        16|White/Clear Jade/...| DN3559-100| 146.0|   4.3|                    |   NULL|        99|       2144043.0|       8|
|Mercurial Vapor 1...|  Unisex|     4|21995.0|                NULL|        16|White/Coconut Mil...|  DJ4978-101|  37.0|   4.7|LOOK FAST, FEEL F...|   NULL|        97|       2133515.0|       3|
+--------------------+--------+------+-------+--------------------+----------+--------------------+------------+------+------+--------------------+-------+----------+----------------+--------+
only showing top 5 rows
```

```
from pyspark.ml.feature import VectorAssembler, OneHotEncoder, StringIndexer
from pyspark.ml.feature import StandardScaler
from pyspark.ml.clustering import KMeans
from pyspark.ml import Pipeline
```

Business Objective 2: How can we enhance customer segmentation, optimize marketing strategies, and improve inventory management in the shoe store, ultimately driving competitive advantage and customer loyalty?

K-means clustering empowers with insights into customer preferences and product segmentation, fostering tailored marketing strategies and informed inventory management. By identifying distinct clusters within the shoe data, we can optimize product placement, personalize customer experiences, and swiftly adapt to evolving market trends, ultimately enhancing its competitive edge and customer loyalty.

```
data = All_Shoes_new["Name","Category","Colors","Price","Color_Name","Rating","Features","Comfort","Units_Sold","Revenue_Generated","Discount", "Count_Sizes"]
print(data.count())

for col in ["Colors", "Price", "Rating", "Units_sold", "Count_Sizes"]:
    data = data.withColumn(col, data[col].cast("float"))

category_indexer = StringIndexer(inputCol="Category", outputCol="CategoryIndex")
encoder = OneHotEncoder(inputCols=["CategoryIndex"], outputCols=["New_Category"])
pipeline = Pipeline(stages=[category_indexer, encoder])
pipeline_model = pipeline.fit(data)
data = pipeline_model.transform(data)

# Assemble features into a single column
feature_columns = ["Colors", "Price", "Rating", "New_Category", "Units_sold", "Count_Sizes"]


assembler = VectorAssembler(inputCols=feature_columns, outputCol="assembled_features", handleInvalid="skip")
data = assembler.transform(data)

# Scale features
scaler = StandardScaler(inputCol="assembled_features", outputCol="scaled_features", withStd=True, withMean=False)
scaler_model = scaler.fit(data)
data = scaler_model.transform(data)

# Perform k-means clustering
kmeans = KMeans(featuresCol="scaled_features").setK(3)
results = kmeans.fit(data).transform(data)

# Show the results
results.show(10)
```

```
300
+--------------------+--------+------+-------+--------------------+------+--------------------+-------+----------+-----------------+--------+-----------+-------------+------------+--------------------+--------------------+----------+
|                Name|Category|Colors|  Price|          Color_Name|Rating|            Features|Comfort|Units_sold|Revenue_Generated|Discount|Count_Sizes|CategoryIndex| New_Category|  assembled_features|     scaled_features|prediction|
+--------------------+--------+------+-------+--------------------+------+--------------------+-------+----------+-----------------+--------+-----------+-------------+------------+--------------------+--------------------+----------+
|Phantom Luna Elit...|  Unisex|   1.0|26795.0|Fuchsia Dream/Bar...|   0.0|                    |   NULL|      88.0|        2357960.0|       0|       12.0|          2.0|    (2,[],[])|[1.0,26795.0,0.0,...|[0.36895869487533...|         1|
| Phantom GX Elite SE|  Unisex|   1.0|23795.0|Fuchsia Dream/Bar...|   0.0|                    |   NULL|      94.0|        2236730.0|       1|        1.0|          2.0|    (2,[],[])|[1.0,23795.0,0.0,...|[0.36895869487533...|         1|
|          Alphafly 2|   Women|   4.0|21657.0|Black/Sea Coral/W...|   4.3|                    |   NULL|      99.0|        2144043.0|       9|       16.0|          1.0|(2,[1],[1.0])|[4.0,21657.0,4.30...|[1.47583477950135...|         2|
|          Alphafly 2|   Women|   4.0|21657.0|White/Clear Jade/...|   4.3|                    |   NULL|      99.0|        2144043.0|       8|       16.0|          1.0|(2,[1],[1.0])|[4.0,21657.0,4.30...|[1.47583477950135...|         2|
|Mercurial Vapor 1...|  Unisex|   4.0|21995.0|White/Coconut Mil...|   4.7|LOOK FAST, FEEL F...|   NULL|      97.0|        2133515.0|       3|       16.0|          2.0|    (2,[],[])|[4.0,21995.0,4.69...|[1.47583477950135...|         0|
|Mercurial Vapor 1...|  Unisex|   4.0|21995.0|Pink Blast/Gridir...|   4.7|LOOK FAST, FEEL F...|   NULL|      92.0|        2023540.0|       9|        2.0|          2.0|    (2,[],[])|[4.0,21995.0,4.69...|[1.47583477950135...|         0|
|          Alphafly 2|   Women|   4.0|21657.0|Black/Sea Coral/W...|   4.3|                    |   NULL|      92.0|        1992444.0|       3|       16.0|          1.0|(2,[1],[1.0])|[4.0,21657.0,4.30...|[1.47583477950135...|         2|
|Air VaporMax 2023...|     Men|   9.0|19295.0|Pure Platinum/Ant...|   4.6|                    |   NULL|      97.0|        1871615.0|       0|       13.0|          0.0|(2,[0],[1.0])|[9.0,19295.0,4.59...|[3.32062825387805...|         0|
|          Vaporfly 3|   Women|   5.0|20695.0|Hyper Pink/Laser ...|   4.6|                    |   NULL|      90.0|        1862550.0|       4|        1.0|          1.0|(2,[1],[1.0])|[5.0,20695.0,4.59...|[1.84479347437669...|         2|
|          Vaporfly 3|   Women|   5.0|20695.0|Black/Black/Oatme...|   4.6|                    |   NULL|      89.0|        1841855.0|       5|       10.0|          1.0|(2,[1],[1.0])|[5.0,20695.0,4.59...|[1.84479347437669...|         2|
+--------------------+--------+------+-------+--------------------+------+--------------------+-------+----------+-----------------+--------+-----------+-------------+------------+--------------------+--------------------+----------+
only showing top 10 rows
```

```
cluster1=results.filter(results['prediction']==0)
cluster1.select(["Name", "Colors", "Price", "Revenue_Generated", "New_Category", "Units_sold", "Count_Sizes"]).show()
cluster1.count()
print(cluster1.agg(avg('Price')).collect()[0][0])
print(int(cluster1.agg(avg('Count_Sizes')).collect()[0][0]))
```

```
+--------------------+------+-------+-----------------+-------------+----------+-----------+
|                Name|Colors|  Price|Revenue_Generated| New_Category|Units_sold|Count_Sizes|
+--------------------+------+-------+-----------------+-------------+----------+-----------+
|Mercurial Vapor 1...|   4.0|21995.0|        2133515.0|    (2,[],[])|      97.0|       16.0|
|Mercurial Vapor 1...|   4.0|21995.0|        2023540.0|    (2,[],[])|      92.0|        2.0|
|Air VaporMax 2023...|   9.0|19295.0|        1871615.0|(2,[0],[1.0])|      97.0|       13.0|
|          Vaporfly 3|   6.0|19657.0|        1769130.0|(2,[0],[1.0])|      90.0|       16.0|
| Air Jordan 6 'Aqua'|   1.0|18395.0|        1765920.0|(2,[0],[1.0])|      96.0|        9.0|
|          Vaporfly 3|   4.0|20695.0|        1717685.0|(2,[0],[1.0])|      83.0|       13.0|
|Mercurial Vapor 1...|   4.0|21995.0|        1693615.0|    (2,[],[])|      77.0|        2.0|
|       Invincible 3|   7.0|16995.0|        1665510.0|(2,[0],[1.0])|      98.0|       13.0|
| Air Jordan I High G|   3.0|16995.0|        1665510.0|(2,[0],[1.0])|      98.0|       16.0|
|Air VaporMax 2023...|   9.0|19295.0|        1640075.0|(2,[0],[1.0])|      85.0|        3.0|
| Air Jordan I High G|   2.0|16147.0|        1566259.0|(2,[0],[1.0])|      97.0|        2.0|
|       Invincible 3|   7.0|16995.0|        1563540.0|(2,[0],[1.0])|      92.0|       13.0|
| Air Jordan 13 Retro|   2.0|19295.0|        1562895.0|    (2,[],[])|      81.0|        1.0|
|  Air Jordan 3 Retro|   3.0|15995.0|        1551515.0|(2,[0],[1.0])|      97.0|       16.0|
|         Air Max 97|   9.0|16147.0|        1550112.0|(2,[0],[1.0])|      96.0|        1.0|
|       Invincible 3|   7.0|16147.0|        1550112.0|(2,[0],[1.0])|      96.0|       16.0|
|       Air Jordan 1|   2.0|17595.0|        1530765.0|    (2,[],[])|      87.0|        1.0|
|       Invincible 3|   1.0|16617.0|        1528764.0|(2,[0],[1.0])|      92.0|        2.0|
|         Vaporfly 3|   6.0|20695.0|        1510735.0|(2,[0],[1.0])|      73.0|       16.0|
|         Vaporfly 3|   6.0|20695.0|        1510735.0|(2,[0],[1.0])|      73.0|        1.0|
+--------------------+------+-------+-----------------+-------------+----------+-----------+
only showing top 20 rows
```

```
15559.612244897959
10
```

```
cluster2=results.filter(results['prediction']==1)
cluster2.select(["Name", "Colors", "Price", "Revenue_Generated", "New_Category","Units_sold", "Count_Sizes"]).show()
cluster2.count()
print(cluster2.agg(avg('Price')).collect()[0][0])
print(int(cluster2.agg(avg('Count_Sizes')).collect()[0][0]))
```

```
+--------------------+------+-------+-----------------+-------------+----------+-----------+
|                Name|Colors|  Price|Revenue_Generated| New_Category|Units_sold|Count_Sizes|
+--------------------+------+-------+-----------------+-------------+----------+-----------+
|Phantom Luna Elit...|   1.0|26795.0|        2357960.0|    (2,[],[])|      88.0|       12.0|
| Phantom GX Elite SE|   1.0|23795.0|        2236730.0|    (2,[],[])|      94.0|        1.0|
|    Phantom GX Elite|   2.0|21995.0|        1737605.0|    (2,[],[])|      79.0|        5.0|
|LeBron XX Premium EP|   1.0|19295.0|        1736550.0|    (2,[],[])|      90.0|        5.0|
|Air Jordan XXXVII...|   3.0|18395.0|        1729130.0|    (2,[],[])|      94.0|        9.0|
|    Phantom GX Elite|   2.0|21995.0|        1605635.0|    (2,[],[])|      73.0|       12.0|
|Superfly 9 Elite ...|   1.0|25095.0|        1555890.0|    (2,[],[])|      62.0|       11.0|
|    G.T. Hustle 2 EP|   1.0|15995.0|        1551515.0|(2,[0],[1.0])|      97.0|       11.0|
|          Air Max 97|   1.0|16147.0|        1550112.0|(2,[1],[1.0])|      96.0|        3.0|
|Air Jordan XXXVII...|   3.0|18395.0|        1508390.0|    (2,[],[])|      82.0|       12.0|
|     Tiger Woods '13|   2.0|21295.0|        1490650.0|(2,[0],[1.0])|      70.0|       16.0|
|Air Max 1 '87 Safari|   1.0|16995.0|        1478565.0|(2,[1],[1.0])|      87.0|        8.0|
|     Tiger Woods '13|   2.0|21295.0|        1448060.0|(2,[0],[1.0])|      68.0|        2.0|
|      Air Penny 2 QS|   1.0|18395.0|        1434810.0|(2,[0],[1.0])|      78.0|        6.0|
|Vapor 15 Elite Me...|   1.0|22995.0|        1425690.0|    (2,[],[])|      62.0|        9.0|
|LeBron XXI 'Akoya...|   2.0|18395.0|        1416415.0|    (2,[],[])|      77.0|        7.0|
|    Air Adjust Force|   1.0|15995.0|        1391565.0|(2,[1],[1.0])|      87.0|        8.0|
|Air Jordan XXXVII...|   3.0|18395.0|        1379625.0|    (2,[],[])|      75.0|       12.0|
|Air Force 1 High ...|   1.0|13995.0|        1371510.0|(2,[0],[1.0])|      98.0|        9.0|
|         Air Penny 2|   1.0|19295.0|        1369945.0|(2,[0],[1.0])|      71.0|       11.0|
+--------------------+------+-------+-----------------+-------------+----------+-----------+
only showing top 20 rows

17213.0
9
```

```
cluster3=results.filter(results['prediction']==2)
cluster3.select(["Name", "Colors", "Price", "Revenue_Generated", "New_Category","Units_sold", "Count_Sizes"]).show()
cluster3.count()
print(cluster3.agg(avg('Price')).collect()[0][0])
print(int(cluster3.agg(avg('Count_Sizes')).collect()[0][0]))
```

```
+--------------------+------+-------+-----------------+-------------+----------+-----------+
|                Name|Colors|  Price|Revenue_Generated| New_Category|Units_sold|Count_Sizes|
+--------------------+------+-------+-----------------+-------------+----------+-----------+
|          Alphafly 2|   4.0|21657.0|        2144043.0|(2,[1],[1.0])|      99.0|       16.0|
|          Alphafly 2|   4.0|21657.0|        2144043.0|(2,[1],[1.0])|      99.0|       16.0|
|          Alphafly 2|   4.0|21657.0|        1992444.0|(2,[1],[1.0])|      92.0|       16.0|
|          Vaporfly 3|   5.0|20695.0|        1862550.0|(2,[1],[1.0])|      90.0|        1.0|
|          Vaporfly 3|   5.0|20695.0|        1841855.0|(2,[1],[1.0])|      89.0|       10.0|
|Air Jordan 1 Elev...|   1.0|18395.0|        1784315.0|(2,[1],[1.0])|      97.0|        6.0|
|   Air Max 97 Futura|   2.0|17495.0|        1679520.0|(2,[1],[1.0])|      96.0|        3.0|
|        Invincible 3|  10.0|16995.0|        1631520.0|(2,[1],[1.0])|      96.0|        8.0|
|        Invincible 3|  10.0|16995.0|        1631520.0|(2,[1],[1.0])|      96.0|        7.0|
|          Vaporfly 3|   5.0|20695.0|        1552125.0|(2,[1],[1.0])|      75.0|       10.0|
|          Vaporfly 3|   5.0|19657.0|        1533246.0|(2,[1],[1.0])|      78.0|       16.0|
|  Air Jordan 3 Retro|   3.0|16595.0|        1526740.0|(2,[1],[1.0])|      92.0|       16.0|
|           Vomero 17|   2.0|14995.0|        1439520.0|(2,[1],[1.0])|      96.0|       10.0|
|Air VaporMax 2023...|   4.0|19295.0|        1427830.0|(2,[1],[1.0])|      74.0|       10.0|
|        Air Max 1 LX|   2.0|14995.0|        1424525.0|(2,[1],[1.0])|      95.0|        9.0|
|        InfinityRN 4|   8.0|14995.0|        1409530.0|(2,[1],[1.0])|      94.0|        8.0|
|        Invincible 3|  10.0|16147.0|        1404789.0|(2,[1],[1.0])|      87.0|        8.0|
|Air Adjust Force ...|   2.0|15197.0|        1382927.0|(2,[1],[1.0])|      91.0|        1.0|
|        InfinityRN 4|   8.0|14995.0|        1364545.0|(2,[1],[1.0])|      91.0|        9.0|
|         Air Max 270|   1.0|14995.0|        1349550.0|(2,[1],[1.0])|      90.0|       10.0|
+--------------------+------+-------+-----------------+-------------+----------+-----------+
only showing top 20 rows

14754.076923076924
8
```

In the provided clusters:

Cluster 1: Shows higher-priced shoes with moderate to high units sold, indicating demand for premium products.

Cluster 2: Exhibits a mix of mid to high-priced shoes with varied colors and sizes, suggesting diversity in consumer preferences.

Cluster 3: Displays higher-priced shoes with comparatively lower units sold, indicating niche or specialized products.

Analyzing these clusters can help in tailoring marketing strategies, optimizing inventory, and identifying trends. For instance, Cluster 1 may be targeted towards high-end consumers, while Cluster 2 might represent products appealing to a broader audience.

```
import pandas as pd

# Convert PySpark DataFrame to Pandas DataFrame
pandas_df = All_Shoes_new.toPandas()

# Export Pandas DataFrame to a CSV file
pandas_df.to_csv('All_Shoes_new.csv', index=False)


#Reading the data
shoesdata=session.read.csv('All_Shoes_new.csv', header=True, inferSchema=True)

#Reading the stream of data by specifying the schema and directory #read continuous data
shoesdata_stream=session.readStream.schema(shoesdata.schema).csv('shoes_stream/')

#Writing the stream of data to a table
#The output mode is append which means that the data will be added always to existing table
shoesquery=shoesdata_stream.filter("Name != 'Name'").writeStream.queryName("shoestable").format("memory").outputMode("append").start()


#Copy the Nike_new.csv file to the streaming folder #shutl lib will help to cp file from one place to another.
import shutil
src=r"All_Shoes_new.csv"
dest = r"shoes_stream"
shutil.copy(src,dest)
```

```
'shoes_stream/All_Shoes_new.csv'
```

```
session.sql("select * from shoestable limit 5").show()
```

```
+--------------------+--------+------+-------+--------------------+----------+--------------------+------------+------+------+--------------------+-------+----------+-----------------+--------+
|                Name|Category|Colors|  Price|               Sizes|Count_Sizes|         Color_Name|product_Code|Review|Rating|            Features|Comfort|Units_sold|Revenue_Generated|Discount|
+--------------------+--------+------+-------+--------------------+----------+--------------------+------------+------+------+--------------------+-------+----------+-----------------+--------+
|Phantom Luna Elit...|  Unisex|     1|26795.0|['4.5', '5', '5.5...|        12|Fuchsia Dream/Bar...|  FQ8033-500|   0.0|   0.0|                    |   NULL|        88|        2357960.0|       0|
| Phantom GX Elite SE|  Unisex|     1|23795.0|              ['7']|         1|Fuchsia Dream/Bar...|  FD0565-500|   0.0|   0.0|                    |   NULL|        94|        2236730.0|       1|
|          Alphafly 2|   Women|     4|21657.0|                NULL|        16|Black/Sea Coral/W...|  DN3559-001| 146.0|   4.3|                    |   NULL|        99|        2144043.0|       9|
|          Alphafly 2|   Women|     4|21657.0|                NULL|        16|White/Clear Jade/...|  DN3559-100| 146.0|   4.3|                    |   NULL|        99|        2144043.0|       8|
|Mercurial Vapor 1...|  Unisex|     4|21995.0|                NULL|        16|White/Coconut Mil...|  DJ4978-101|  37.0|   4.7|LOOK FAST, FEEL F...|   NULL|        97|        2133515.0|       3|
+--------------------+--------+------+-------+--------------------+----------+--------------------+------------+------+------+--------------------+-------+----------+-----------------+--------+
```

Business Objective 3: How does implementing streaming analytics in our shoe store affect profitability and fame? Can real-time insights optimize inventory, pricing, and marketing to drive profitability and enhance customer satisfaction, thereby boosting the store's reputation?

Implementing streaming in the shoe store significantly impacts both profitability and fame. Real-time insights enable the store to optimize its inventory, ensuring that popular products are readily available while minimizing excess stock. This agility enhances profitability by reducing inventory costs and maximizing sales opportunities. Additionally, dynamic pricing strategies based on real-time data analysis help to capture maximum value from each sale, further boosting profitability. Moreover, by leveraging real-time data to tailor marketing campaigns and enhance customer experiences, the store gains fame and recognition in the market. The ability to offer personalized recommendations and timely promotions increases customer satisfaction and loyalty, ultimately elevating the store's reputation and fame within the industry. Overall, streaming drives profitability through efficient operations and enhances fame by delivering exceptional customer experiences.

```
import time
for i in range (10):
    session.sql("SELECT Category, sum(Units_Sold) as Total_Units_Sold, sum(Revenue_Generated) as Total_Revenue_Generated, sum(Discount) as \
     Total_Discount_Given FROM shoestable GROUP BY Category").show()
    newfile="shoes_stream/Shoes" + str(i) +  ".csv"
    shutil.copy(src,newfile)
    time.sleep(5)
```

```
+--------+---------------+----------------------+-------------------+
|Category|Total_Units_Sold|Total_Revenue_Generated|Total_Discount_Given|
+--------+---------------+----------------------+-------------------+
|     Men|          12908|          1.99562794E8|                755|
|   Women|           7932|           1.2076013E8|                452|
```

```
|  Unisex|            3688|           6.4249602E7|                196|
+--------+----------------+----------------------+--------------------+

+--------+----------------+----------------------+--------------------+
|Category|Total_Units_Sold|Total_Revenue_Generated|Total_Discount_Given|
+--------+----------------+----------------------+--------------------+
|     Men|           25816|          3.99125588E8|                1510|
|   Women|           15864|           2.4152026E8|                 904|
|  Unisex|            7376|          1.28499204E8|                 392|
+--------+----------------+----------------------+--------------------+

+--------+----------------+----------------------+--------------------+
|Category|Total_Units_Sold|Total_Revenue_Generated|Total_Discount_Given|
+--------+----------------+----------------------+--------------------+
|     Men|           38724|          5.98688382E8|                2265|
|   Women|           23796|           3.6228039E8|                1356|
|  Unisex|           11064|          1.92748806E8|                 588|
+--------+----------------+----------------------+--------------------+

+--------+----------------+----------------------+--------------------+
|Category|Total_Units_Sold|Total_Revenue_Generated|Total_Discount_Given|
+--------+----------------+----------------------+--------------------+
|     Men|           51632|          7.98251176E8|                3020|
|   Women|           31728|           4.8304052E8|                1808|
|  Unisex|           14752|          2.56998408E8|                 784|
+--------+----------------+----------------------+--------------------+

+--------+----------------+----------------------+--------------------+
|Category|Total_Units_Sold|Total_Revenue_Generated|Total_Discount_Given|
+--------+----------------+----------------------+--------------------+
|     Men|           64540|           9.9781397E8|                3775|
|   Women|           39660|           6.0380065E8|                2260|
|  Unisex|           18440|           3.2124801E8|                 980|
+--------+----------------+----------------------+--------------------+

+--------+----------------+----------------------+--------------------+
|Category|Total_Units_Sold|Total_Revenue_Generated|Total_Discount_Given|
+--------+----------------+----------------------+--------------------+
|     Men|           77448|         1.197376764E9|                4530|
|   Women|           47592|           7.2456078E8|                2712|
|  Unisex|           22128|          3.85497612E8|                1176|
+--------+----------------+----------------------+--------------------+

+--------+----------------+----------------------+--------------------+
|Category|Total_Units_Sold|Total_Revenue_Generated|Total_Discount_Given|
+--------+----------------+----------------------+--------------------+
|     Men|           90356|         1.396939558E9|                5285|
|   Women|           55524|           8.4532091E8|                3164|
|  Unisex|           25816|          4.49747214E8|                1372|
+--------+----------------+----------------------+--------------------+

+--------+----------------+----------------------+--------------------+
|Category|Total_Units_Sold|Total_Revenue_Generated|Total_Discount_Given|
```

Business Objective 4: How are different shoe models related to each other based on factors like color availability and category similarity, and how can these relationships inform personalized product recommendations and marketing strategies?

By analyzing the graph of shoe relationships, the store can identify similar products based on attributes like colors available, category, and other features. This allows the store to offer personalized recommendations to customers, increasing the likelihood of sales and enhancing customer satisfaction.

Understanding the relationships between products can also inform the development of engaging customer experiences, such as curated collections, thematic product displays, and interactive online features. By creating compelling narratives around product relationships, the store can captivate customers and foster brand loyalty.

```
!apt-get install openjdk-11-jdk-headless -qq > /dev/null
!wget -q https://bitbucket.org/habedi/datasets/raw/b6769c4664e7ff68b001e2f43bc517888cbe3642/spark/spark-3.0.2-bin-hadoop2.7.tgz
!tar xf spark-3.0.2-bin-hadoop2.7.tgz
!rm -rf spark-3.0.2-bin-hadoop2.7.tgz*
!pip -q install findspark pyspark graphframes


import os
os.environ["PYSPARK_DRIVER_PYTHON"] = "jupyter"
os.environ["PYSPARK_DRIVER_PYTHON_OPTS"] = "notebook"
os.environ["PYSPARK_SUBMIT_ARGS"] = "--packages graphframes:graphframes:0.8.1-spark3.0-s_2.12 pyspark-shell"
```

```
from graphframes import *
from pyspark import *
from pyspark.sql import *
spark = SparkSession.builder.appName('function').getOrCreate()


vertices = spark.read.option('header', 'true').csv('Shoes_Nodes.csv')
edges = spark.read.option('header', 'true').csv('Shoes_Edges.csv')


vertices.show()
edges.show()
```

```
+--------------------+--------------------+--------------------+--------------------+-----------+--------------------+--------------------+--------------------+-----------+------+------+--------+-------+
|                  id|                Name|            Category|              Colors|      Price|               Sizes|         Count_Sizes|          Color_Name|product_Code|Review|Rating|Features|Comfort|
+--------------------+--------------------+--------------------+--------------------+-----------+--------------------+--------------------+--------------------+-----------+------+------+--------+-------+
|                   1|           Air Max 1|                 Men|                   5|  12 795.00|['6', '6.5', '7',...|                  17|White/Photon Dust...|  FD9082-103|    88|   4.8|        |   NULL|
|Meet the leader o...| the Air Max 1 bl...| wavy mudguard an...| this classic ico...|       NULL|                NULL|                NULL|                NULL|       NULL|  NULL|  NULL|    NULL|   NULL|
|            Benefits|                NULL|                NULL|                NULL|       NULL|                NULL|                NULL|                NULL|       NULL|  NULL|  NULL|    NULL|   NULL|
|Plush and comfort...| the Max Air cush...|                NULL|                NULL|       NULL|                NULL|                NULL|                NULL|       NULL|  NULL|  NULL|    NULL|   NULL|
|The Waffle outsol...|                NULL|                NULL|                NULL|       NULL|                NULL|                NULL|                NULL|       NULL|  NULL|  NULL|    NULL|   NULL|
|Colour Shown: Whi...|                NULL|                NULL|                NULL|       NULL|                NULL|                NULL|                NULL|       NULL|  NULL|  NULL|    NULL|   NULL|
|   Style: FD9082-103|                NULL|                NULL|                NULL|       NULL|                NULL|                NULL|                NULL|       NULL|  NULL|  NULL|    NULL|   NULL|
|Country/Region of...|                NULL|                NULL|                NULL|       NULL|                NULL|                NULL|                NULL|       NULL|  NULL|  NULL|    NULL|   NULL|
|           Air Max 1|                NULL|                NULL|                NULL|       NULL|                NULL|                NULL|                NULL|       NULL|  NULL|  NULL|    NULL|   NULL|
|                Sure| Air Max 1 starte...| but you can't ke...| this runner with...| to this day| are celebrated y...|Plush and comfort...|                NULL|       NULL|  NULL|  NULL|    NULL|   NULL|
|The Waffle outsol...|                NULL|                NULL|                NULL|       NULL|                NULL|                NULL|                NULL|       NULL|  NULL|  NULL|    NULL|   NULL|
|Colour Shown: Whi...|                NULL|                NULL|                NULL|       NULL|                NULL|                NULL|                NULL|       NULL|  NULL|  NULL|    NULL|   NULL|
|   Style: FD9082-103|                NULL|                NULL|                NULL|       NULL|                NULL|                NULL|                NULL|       NULL|  NULL|  NULL|    NULL|   NULL|
|Country/Region of...|                NULL|                NULL|                NULL|       NULL|                NULL|                NULL|                NULL|       NULL|  NULL|  NULL|    NULL|   NULL|
|                   2|          Air Max 90|                 Men|                   1|  11 895.00|['7', '7.5', '8',...|                  13|Anthracite/Black/...|  FB9658-001|     3|   4.7|        |   NULL|
|Lace up and feel ...| revolutionised t...|  its Waffle outsole| visible Nike Air...|       NULL|                NULL|                NULL|                NULL|       NULL|  NULL|  NULL|    NULL|   NULL|
|            Benefits|                NULL|                NULL|                NULL|       NULL|                NULL|                NULL|                NULL|       NULL|  NULL|  NULL|    NULL|   NULL|
|The textile upper...|                NULL|                NULL|                NULL|       NULL|                NULL|                NULL|                NULL|       NULL|  NULL|  NULL|    NULL|   NULL|
|Originally design...| its foam midsole...|                NULL|                NULL|       NULL|                NULL|                NULL|                NULL|       NULL|  NULL|  NULL|    NULL|   NULL|
|Rubber Waffle out...|                NULL|                NULL|                NULL|       NULL|                NULL|                NULL|                NULL|       NULL|  NULL|  NULL|    NULL|   NULL|
+--------------------+--------------------+--------------------+--------------------+-----------+--------------------+--------------------+--------------------+-----------+------+------+--------+-------+
only showing top 20 rows
```

```
+---+---+------------------+
|src|dst|          relation|
+---+---+------------------+
|  4|  5|similar_high_rated|
|  5| 11|similar_high_rated|
| 11| 18|similar_high_rated|
| 18| 25|similar_high_rated|
| 25| 28|similar_high_rated|
| 28| 57|similar_high_rated|
| 57| 62|similar_high_rated|
| 62| 82|similar_high_rated|
| 82| 97|similar_high_rated|
| 97|134|similar_high_rated|
|134|  5|similar_high_rated|
|151| 11|similar_high_rated|
|167| 18|similar_high_rated|
|173| 25|similar_high_rated|
|178| 28|similar_high_rated|
|184| 57|similar_high_rated|
|187| 62|similar_high_rated|
|191| 82|similar_high_rated|
|194| 97|similar_high_rated|
|198|134|similar_high_rated|
+---+---+------------------+
only showing top 20 rows
```

```
mygraph = GraphFrame(vertices, edges)
```

```
/usr/local/lib/python3.10/dist-packages/pyspark/sql/dataframe.py:168: UserWarning: DataFrame.sql_ctx is an internal property, and will be removed in future releases. Use DataFrame.sparkSession instead.
  warnings.warn(
```

```
mygraph.degrees.show(4)
```

```
/usr/local/lib/python3.10/dist-packages/pyspark/sql/dataframe.py:147: UserWarning: DataFrame constructor is internal. Do not directly use it.
  warnings.warn("DataFrame constructor is internal. Do not directly use it.")
+---+------+
| id|degree|
+---+------+
| 51|     2|
|307|     2|
|205|     2|
| 54|     4|
+---+------+
only showing top 4 rows
```

Double-click (or enter) to edit

```
result=mygraph.filterVertices("Name=='Air Jordan 1 Mid' and Category=='Unisex'");
result.vertices.show()
```

```
+---+---------------+--------+------+---------+-------------------+-----------+----------------+------------+------+------+--------------------+-------+
| id|           Name|Category|Colors|    Price|              Sizes|Count_Sizes|      Color_Name|product_Code|Review|Rating|            Features|Comfort|
+---+---------------+--------+------+---------+-------------------+-----------+----------------+------------+------+------+--------------------+-------+
| 37|Air Jordan 1 Mid|  Unisex|     1|11 495.00|['8', '8.5', '9',...|          9|White/White/White|  554724-136|   995|   4.8|FRESH COLOUR, FAM...|   NULL|
+---+---------------+--------+------+---------+-------------------+-----------+----------------+------------+------+------+--------------------+-------+
```

```
result3=mygraph.filterEdges("relation='same_category_unisex'");
result3.edges.show()
```

```
+---+---+--------------------+
|src|dst|            relation|
+---+---+--------------------+
| 37|255|same_category_unisex|
| 47|276|same_category_unisex|
| 54|293|same_category_unisex|
| 65|300|same_category_unisex|
| 75|307|same_category_unisex|
| 90|321|same_category_unisex|
|138|322|same_category_unisex|
|139|338|same_category_unisex|
|155| 37|same_category_unisex|
|156| 47|same_category_unisex|
|164| 54|same_category_unisex|
|171| 65|same_category_unisex|
|175| 75|same_category_unisex|
|182| 90|same_category_unisex|
|255|138|same_category_unisex|
|276|139|same_category_unisex|
|293|155|same_category_unisex|
|300|156|same_category_unisex|
|307|164|same_category_unisex|
|321|171|same_category_unisex|
+---+---+--------------------+
only showing top 20 rows
```

```
mygraph.triangleCount().show()
```

```
import networkx as nx
import matplotlib.pyplot as plt
# the function will plot the source and destination nodes and connect them by meand of undirected line
def plot_undirected_graph(edge_list):
    plt.figure(figsize=(9,9))
    gplot=nx.Graph()
    for row in edge_list.select("src", "dst").take(1000):
        gplot.add_edge(row["src"], row["dst"])
    nx.draw(gplot, with_labels=True, font_weight="bold", node_size=3500)
plot_undirected_graph(mygraph.edges)
```