Walmart_Data_Prep...

READY

LOADING THE DATA- Walmart Retail Dataset

SPARK JOB FINISHED

```
val walmartDF = spark.read
  .option("header", "true")
  .option("inferSchema", "true")
  .csv("/user/sc10670_nyu_edu/project/Walmart-Retail-Dataset.csv")

walmartDF.show(5)
walmartDF.printSchema()
```

```
product_name|product_sub_category|     profit| region|  sales| ship_date|     ship_mode|     s
hipping_cost|    state|unit_price|zip_code|
+-------------+------------+-----------------+---------------+--------+----------+-----------
---------+--------------+--------------+-------------------+----------------+----------------
+--------------------+--------------------+-----------+-------+-------+----------+------------
--+------------------+---------+----------+--------+
|Stevens Point|          60|    Dennis Bolton|       Corporate|    0.17|2020-02-29|a42c8cff-57
57-4e9...| Not Specified|             7|               0.55|       Furniture|       Jumbo Drum
|Global Enterprise...|  Chairs & Chairmats|19559.92268|Central|  21.84|2020-03-02|Delivery Tru
ck| 3.772509354070991|Wisconsin|      3.29|   54481|
|Stevens Point|          60|    Dennis Bolton|       Corporate|    0.17|2020-02-29|1c37f301-56
4f-40f...| Not Specified|             7|               0.55|       Furniture|       Jumbo Drum
|Global Enterprise...|  Chairs & Chairmats|19559.92268|Central|1811.67|2020-03-07|Delivery Tru
ck| 816.3408935057945|Wisconsin|    258.98|   54481|
|    Grapevine|          49|Anthony Garverick|  Small Business|    0.05|2021-11-11|ec649eae-53
5d-415...|        Medium|            42|               0.69|       Furniture|        Jumbo Box
|Bevis Rectangular...|              Tables|  7535.9388|Central|6129.06|2021-11-15|Delivery Tru
ck| 4530.505983276593|    Texas|    145.98|   76051|
|        Tempe|          30|   Anne McFarland|        Consumer|    0.05|2020-08-02|efdcbace-53
20-400...| Not Specified|            30|               0.37| Office Supplies|        Small Box
|          Xerox 1923|               Paper|18860.92419|   West|  198.9|2020-08-08|   Regular A
ir|128.73150520457037|  Arizona|      6.68|   85281|
|Coconut Creek|          80|     Raymond Fair|     Home Office|    0.14|2021-08-13|8fd6c0f6-9e
28-45b...|           Low|            44|                \N| Office Supplies|        Small Box
|SAFCO Mobile Desk...|Storage & Organiz...| 24750.4921|  South|1875.28|2021-08-18|   Express A
ir| 33.60838488294798|  Florida|     42.76|   33063|
+-------------+------------+-----------------+---------------+--------+----------+-----------
---------+--------------+--------------+-------------------+----------------+----------------
+--------------------+--------------------+-----------+-------+-------+----------+------------
--+------------------+---------+----------+--------+
only showing top 5 rows

root
 |-- city: string (nullable = true)
 |-- customer_age: string (nullable = true)
 |-- customer_name: string (nullable = true)
 |-- customer_segment: string (nullable = true)
 |-- discount: string (nullable = true)
 |-- order_date: date (nullable = true)
 |-- order_id: string (nullable = true)
```

Took 17 sec. Last updated by anonymous at November 22 2024, 10:26:17 AM. (outdated)

```
walmartDF.show()
```

SPARK JOB (http://nyu-dataproc-w-0.c.hpc-dataproc-19b8.internal:39359/jobs/job?id=3) FINISHED

Walmart_Data_Prep...

```
+-------------+------------+-------------+----------------+--------+----------+--------------
-----+--------------+--------------+------------------+----------------+-----------------+---
----------------+--------------------+-----------+-------+-----+----------+--------------+---
--------------+---------+----------+--------+
|         city|customer_age|customer_name|customer_segment|discount|order_date|            ord
er_id|order_priority|order_quantity|product_base_margin|product_category|product_container|
product_name|product_sub_category|     profit| region|sales| ship_date|     ship_mode|    ship
ping_cost|    state|unit_price|zip_code|
+-------------+------------+-------------+----------------+--------+----------+--------------
-----+--------------+--------------+------------------+----------------+-----------------+---
----------------+--------------------+-----------+-------+-----+----------+--------------+---
--------------+---------+----------+--------+
|Stevens Point|          60|Dennis Bolton|       Corporate|    0.17|2020-02-29|a42c8cff-5757-4
e9...| Not Specified|             7|              0.55|       Furniture|       Jumbo Drum|Glo
bal Enterprise...|  Chairs & Chairmats|19559.92268|Central|21.84|2020-03-02|Delivery Truck|3.7
72509354070991|Wisconsin|      3.29|   54481|
+-------------+------------+-------------+----------------+--------+----------+--------------
```

Took 0 sec. Last updated by anonymous at November 22 2024, 10:26:50 AM. (outdated)

NULL Checks and Cleaning up of Data

SPARK JOB FINISHED

```
val columnStats = walmartDF.columns.map(colName => {
  val nullCount = walmartDF.filter(col(colName).isNull).count()
  val nonNullCount = walmartDF.filter(col(colName).isNotNull).count()
  (colName, nullCount, nonNullCount)
})

println("Column stats (Null and Non-Null counts):\n" + columnStats.map {
  case (col, nulls, nonNulls) => s"$col: Nulls = $nulls, Non-Nulls = $nonNulls"
}.mkString("\n"))
```

```
city: Nulls = 8, Non-Nulls = 1041819
customer_age: Nulls = 0, Non-Nulls = 1041827
customer_name: Nulls = 14, Non-Nulls = 1041813
customer_segment: Nulls = 2595, Non-Nulls = 1039232
discount: Nulls = 0, Non-Nulls = 1041827
order_date: Nulls = 4776, Non-Nulls = 1037051
order_id: Nulls = 4776, Non-Nulls = 1037051
order_priority: Nulls = 6925, Non-Nulls = 1034902
order_quantity: Nulls = 4776, Non-Nulls = 1037051
product_base_margin: Nulls = 4776, Non-Nulls = 1037051
product_category: Nulls = 4776, Non-Nulls = 1037051
product_container: Nulls = 4776, Non-Nulls = 1037051
product_name: Nulls = 4787, Non-Nulls = 1037040
product_sub_category: Nulls = 4776, Non-Nulls = 1037051
profit: Nulls = 4776, Non-Nulls = 1037051
region: Nulls = 4776, Non-Nulls = 1037051
sales: Nulls = 4776, Non-Nulls = 1037051
ship_date: Nulls = 4776, Non-Nulls = 1037051
```

```
shipping_cost: Nulls = 4841, Non-Nulls = 1036986
state: Nulls = 4776, Non-Nulls = 1037051
unit_price: Nulls = 4776, Non-Nulls = 1037051
zip_code: Nulls = 4776, Non-Nulls = 1037051
columnStats: Array[(String, Long, Long)] = Array((city,8,1041819), (customer_age,0,1041827),
(customer_name,14,1041813), (customer_segment,2595,1039232), (discount,0,1041827), (order_dat
```

Walmart Data Prep...

Took 1 min 22 sec. Last updated by anonymous at November 22 2024, 11:05:11 AM. (outdated)

```
val reducedWalmartDF = walmartDF.drop(
  "customer_name",
  "order_id",
  "order_priority",
  "product_base_margin",
  "product_container",
  "sales",
  "ship_date",
  "ship_mode",
  "unit_price"
)

reducedWalmartDF.show(5)
reducedWalmartDF.printSchema()
```

SPARK JOB (http://hyl-dataproc-w-0.c.hpc-dataproc-19b8.internal:39359/jobs/job?id=142) FINISHED

```
+-------------+-----------+---------------+--------+----------+--------------+--------------
--+-------------------+--------------------+-----------+-------+------------------+---------+
--------+
|         city|customer_age|customer_segment|discount|order_date|order_quantity|product_catego
ry|        product_name|product_sub_category|     profit| region|     shipping_cost|    state|
zip_code|
+-------------+-----------+---------------+--------+----------+--------------+--------------
--+-------------------+--------------------+-----------+-------+------------------+---------+
--------+
|Stevens Point|          60|       Corporate|    0.17|2020-02-29|             7|       Furnitu
re|Global Enterprise...|  Chairs & Chairmats|19559.92268|Central| 3.772509354070991|Wisconsin|
54481|
|Stevens Point|          60|       Corporate|    0.17|2020-02-29|             7|       Furnitu
re|Global Enterprise...|  Chairs & Chairmats|19559.92268|Central| 816.3408935057945|Wisconsin|
54481|
|    Grapevine|          49|  Small Business|    0.05|2021-11-11|            42|       Furnitu
re|Bevis Rectangular...|              Tables|  7535.9388|Central| 4530.505983276593|    Texas|
```

Took 1 sec. Last updated by anonymous at November 22 2024, 12:03:57 PM. (outdated)

```
val columnStats = reducedWalmartDF.columns.map(colName => {
  val nullCount = reducedWalmartDF.filter(col(colName).isNull).count()
  val nonNullCount = reducedWalmartDF.filter(col(colName).isNotNull).count()
  (colName, nullCount, nonNullCount)
})

println("Updated Column stats (Null and Non-Null counts):\n" + columnStats.map {
  case (col, nulls, nonNulls) => s"$col: Nulls = $nulls, Non-Nulls = $nonNulls"
}.mkString("\n"))
```

SPARK JOB FINISHED

```
Updated Column stats (Null and Non-Null counts):
city: Nulls = 8, Non-Nulls = 1041819
```

```
customer_age: Nulls = 0, Non-Nulls = 1041827
customer_segment: Nulls = 2595, Non-Nulls = 1039232
discount: Nulls = 0, Non-Nulls = 1041827
order_date: Nulls = 4776, Non-Nulls = 1037051
order_quantity: Nulls = 4776, Non-Nulls = 1037051
product_category: Nulls = 4776, Non-Nulls = 1037051
product_name: Nulls = 4787, Non-Nulls = 1037040
product_sub_category: Nulls = 4776, Non-Nulls = 1037051
profit: Nulls = 4776, Non-Nulls = 1037051
region: Nulls = 4776, Non-Nulls = 1037051
shipping_cost: Nulls = 4841, Non-Nulls = 1036986
state: Nulls = 4776, Non-Nulls = 1037051
zip_code: Nulls = 4776, Non-Nulls = 1037051
columnStats: Array[(String, Long, Long)] = Array((city,8,1041819), (customer_age,0,1041827),
(customer_segment,2595,1039232), (discount,0,1041827), (order_date,4776,1037051), (order_quant
```

Took 42 sec. Last updated by anonymous at November 22 2024, 12:05:09 PM. (outdated)

Walmart_Data_Prep...

PreProcessing and Cleaning up Data

SPARK JOB (http://hjd-dataproc-w-0.c.hpc-dataproc-19b8.internal:39359/jobs/job?id=199) FINISHED

```
val cleanedReducedWalmartDF = reducedWalmartDF.na.drop()

cleanedReducedWalmartDF.show(5)
cleanedReducedWalmartDF.printSchema()
```

```
+-------------+------------+----------------+--------+----------+--------------+--------------
--+--------------------+--------------------+-----------+-------+------------------+---------+
--------+
|         city|customer_age|customer_segment|discount|order_date|order_quantity|product_catego
ry|        product_name|product_sub_category|     profit| region|     shipping_cost|    state|
zip_code|
+-------------+------------+----------------+--------+----------+--------------+--------------
--+--------------------+--------------------+-----------+-------+------------------+---------+
--------+
|Stevens Point|          60|       Corporate|    0.17|2020-02-29|             7|       Furnitu
re|Global Enterprise...|  Chairs & Chairmats|19559.92268|Central| 3.772509354070991|Wisconsin|
54481|
|Stevens Point|          60|       Corporate|    0.17|2020-02-29|             7|       Furnitu
re|Global Enterprise...|  Chairs & Chairmats|19559.92268|Central| 816.3408935057945|Wisconsin|
54481|
|    Grapevine|          49|  Small Business|    0.05|2021-11-11|            42|       Furnitu
re|Bevis Rectangular...|              Tables|  7535.9388|Central| 4530.505983276593|    Texas|
76051|
```

Took 0 sec. Last updated by anonymous at November 22 2024, 12:06:30 PM. (outdated)

```
val columnStats = cleanedReducedWalmartDF.columns.map(colName => {
  val nullCount = cleanedReducedWalmartDF.filter(col(colName).isNull).count()
  val nonNullCount = cleanedReducedWalmartDF.filter(col(colName).isNotNull).count()
  (colName, nullCount, nonNullCount)
})

println("Updated Column stats (Null and Non-Null counts):\n" + columnStats.map {
  case (col, nulls, nonNulls) => s"$col: Nulls = $nulls, Non-Nulls = $nonNulls"
}.mkString("\n"))
```

SPARK JOB FINISHED

```
Updated Column stats (Null and Non-Null counts):
city: Nulls = 0, Non-Nulls = 1034372
customer_age: Nulls = 0, Non-Nulls = 1034372
customer_segment: Nulls = 0, Non-Nulls = 1034372
discount: Nulls = 0, Non-Nulls = 1034372
order_date: Nulls = 0, Non-Nulls = 1034372
order_quantity: Nulls = 0, Non-Nulls = 1034372
product_category: Nulls = 0, Non-Nulls = 1034372
product_name: Nulls = 0, Non-Nulls = 1034372
product_sub_category: Nulls = 0, Non-Nulls = 1034372
profit: Nulls = 0, Non-Nulls = 1034372
region: Nulls = 0, Non-Nulls = 1034372
shipping_cost: Nulls = 0, Non-Nulls = 1034372
state: Nulls = 0, Non-Nulls = 1034372
zip_code: Nulls = 0, Non-Nulls = 1034372
columnStats: Array[(String, Long, Long)] = Array((city,0,1034372), (customer_age,0,1034372),
(customer_segment,0,1034372), (discount,0,1034372), (order_date,0,1034372), (order_quantity,0,
1034372), (product_category,0,1034372), (product_name,0,1034372), (product_sub_category,0,1034
```


Took 1 min 7 sec. Last updated by anonymous at November 22 2024, 12:10:23 PM. (outdated)

Performing Analysis on Data

SPARK JOB FINISHED

```
val uniqueCategoriesCount = cleanedReducedWalmartDF.select("product_category").distinct().coun
println(s"Total Number of Unique Product Categories: $uniqueCategoriesCount")
```

```
Total Number of Unique Product Categories: 4
uniqueCategoriesCount: Long = 4
```

Took 4 sec. Last updated by anonymous at November 22 2024, 12:46:21 PM. (outdated)

```
val topCategoriesDF = cleanedReducedWalmartDF.groupBy("product_category"
  .agg(
    count("product_name").alias("number_of_products"),
    sum("profit").alias("total_profit")
  )
  .orderBy(desc("total_profit"))
  .limit(20)

println("Top Most Profitable Categories:")
topCategoriesDF.show()
```

SPARK JOB FINISHED

```
Top Most Profitable Categories:
+---------------+-----------------+--------------------+
|product_category|number_of_products|        total_profit|
+---------------+-----------------+--------------------+
| Office Supplies|           558819| 3.547570299275302E9|
|      Technology|           239838|1.5734510402173965E9|
|       Furniture|           231793|  1.52340048983276E9|
|              \N|             3922|2.6731939334000006E7|
+---------------+-----------------+--------------------+

topCategoriesDF: org.apache.spark.sql.Dataset[org.apache.spark.sql.Row] = [product_category: s
tring, number_of_products: bigint ... 1 more field]
```

Took 4 sec. Last updated by anonymous at November 22 2024, 12:47:22 PM. (outdated)

```
val uniqueSubCategoriesCount = cleanedReducedWalmartDF.select("product_sub_category").distinc
println(s"Total Number of Unique Product Sub-Categories: $uniqueSubCategoriesCount")
```

SPARK JOB FINISHED

Total Number of Unique Product Sub-Categories: 18
uniqueSubCategoriesCount: Long = 18

Took 3 sec. Last updated by anonymous at November 22 2024, 12:48:36 PM. (outdated)

```
val topSubCategoriesDF = cleanedReducedWalmartDF.groupBy("product_sub_category")
  .agg(
    count("product_name").alias("number_of_products"),
    sum("profit").alias("total_profit")
  )
  .orderBy(desc("total_profit"))
  .limit(18)

println("Top Most Profitable Product Sub-Categories:")
topSubCategoriesDF.show()
```

SPARK JOB FINISHED

```
+--------------------+------------------+--------------------+
|  product_sub_category|number_of_products|        total_profit|
+--------------------+------------------+--------------------+
|               Paper|            151669| 9.896964046144567E8|
|  Office Furnishings|            115242| 7.574051457315828E8|
|Binders and Binde...|            104640| 6.903699427676636E8|
|Telephones and Co...|            100239| 6.584879016238387E8|
|Computer Peripherals|             84492| 5.572830663536093E8|
| Pens & Art Supplies|             80075| 5.266428264286682E8|
|Storage & Organiz...|             68434| 4.445711705517403E8|
|          Appliances|             56074| 3.741551990876106E8|
|              Tables|             45936| 3.025028584271475E8|
|  Chairs & Chairmats|             45142| 2.965727626884409E8|
|     Office Machines|             44935| 2.922585337183212E8|
|              Labels|             31765|2.0388485414089942E8|
|           Bookcases|             25473|1.6691972298558986E8|
|           Envelopes|             24753|1.6058079390182975E8|
|        Rubber Bands|             23792| 1.576691077824092E8|
|     Copiers and Fax|             10172|6.5421538521610096E7|
|                  \N|              3922|2.6731939334000006E7|
|            Scissors|             17617|                NULL|
+--------------------+------------------+--------------------+

topSubCategoriesDF: org.apache.spark.sql.Dataset[org.apache.spark.sql.Row] = [product_sub_cate
```

Took 3 sec. Last updated by anonymous at November 22 2024, 12:48:57 PM. (outdated)

Discovered something incorrect in Values of State columns for some records

SPARK JOB FINISHED

```
val uniqueStates = cleanedReducedWalmartDF.select("state").distinct().collect().map(_.getStri
println("Unique States:\n" + uniqueStates.mkString(", "))
```

```
Unique States:
54.37866686998688, 29.250377658667592, 8.417438625969595, 211.59433436091797, 10.1969600869018
75, 40.04875073175344, 69.15304636052825, 1643.374046936125, 16.635938264566196, 28.6706463749
06976, 20.218135841280663, 18.234152934223605, 434.3673038914632, 112.86145933512076, 18.50510
```

```
979266656, 2.154502350663946, 127.13367348839809, 67.96828324687095, 58.87918154149107, 209.32
635675083043, 156.2760690799422, 1.2826614043803568, 240.42258497131175, 22900.954011226735,
8.327501361300362, 580.6854635960007, 8.198042181302613, 113.76507382912637, 79.6570938297475
1, 82.74661958619517, 11.871241377910001, 116.28864516716229, 3.7864432884456827, 1.3059278670
30061, 50.28783754195646, 55.60190501499354, 11.926422251527582, 105.79742470922444, 64.858512
33887719, 68.32949418120704, 307.2306305630433, 8054.728666784167, 3.251157728276703, 17747.69
217288078, 229.1456597220893, 27.99998417620775, 181.42786977952474, 640.9759550485813, 152.88
89938208326, 3.220755943905283, 316.0314246084376, 30.318440190002924, 43.8687860862696, 38.70
889089406849, 77.32307924978107, 40.7516719816448, Minnesota, 55.89693467062718, 107.844236506
46079, 427.32152845618094, 127.71361767730139, 6.596783853120197, 2580.4841140127614, 28683.61
102821284, 47.655160283060816, 59.34246135418783, 5.273441167054422, 4.059580520992427, 145.86
603008592522, 11862.121263838853, 34.43700314278913, 11.723141105754049, 77.28308790893179, 19
3.47797540727822, 95.17759775279838, 85.52778254262152, 19704.54118694135, 4558.625487478426,
```

Output is truncated to 102400 bytes. Learn more about **ZEPPELIN_INTERPRETER_OUTPUT_LIMIT** ×

Took 3 sec. Last updated by anonymous at November 22 2024, 12:50:26 PM. (outdated)

Performing Analysis on this

SPARK JOB (http://nyu-dataproc-w-0.c.hpc-dataproc-19b8.internal:39359/jobs/job?id=303) FINISHED

```
cleanedReducedWalmartDF.select("state").show(50, truncate = false)
```

```
+--------------+
|state         |
+--------------+
|Wisconsin     |
|Wisconsin     |
|Texas         |
|Arizona       |
|Florida       |
|Ohio          |
|Florida       |
|Washington    |
|Georgia       |
|California    |
|New York      |
|Texas         |
|North Carolina|
|Ohio          |
|Florida       |
```

Took 1 sec. Last updated by anonymous at November 22 2024, 12:57:23 PM. (outdated)

SPARK JOB FINISHED

```
val validStatesDF = cleanedReducedWalmartDF.filter(col("state").rlike("^[a-zA-Z\\s]+$"))
validStatesDF.select("state").distinct().show(100, truncate = false)
```

```
+--------------+
|state         |
+--------------+
|Minnesota     |
|Nebraska      |
|Oklahoma      |
|MO            |
|Maryland      |
```

```
|Kentucky      |
|New York      |
|Connecticut   |
|Kansas        |
|Alabama       |
|Iowa          |
|Florida       |
|Indiana       |
|Arkansas      |
```

Took 3 sec. Last updated by anonymous at November 22 2024, 1:01:25 PM. (outdated)


```
val stateRecordCountsDF = validStatesDF.groupBy("state")
  .count()
  .orderBy(desc("count"))
```

FINISHED

```
stateRecordCountsDF: org.apache.spark.sql.Dataset[org.apache.spark.sql.Row] = [state: string, count: bigint]
```

Took 0 sec. Last updated by anonymous at November 22 2024, 1:01:50 PM. (outdated)

```
stateRecordCountsDF.show(100, truncate = false)

// Calculate the total number of valid records
val totalValidRecords = validStatesDF.count()
println(s"Total number of valid records: $totalValidRecords")
```

SPARK JOB FINISHED

```
|Maine         |7911 |
|Mississippi   |7827 |
|Idaho         |7796 |
|New Mexico    |7145 |
|Nebraska      |6407 |
|West Virginia |5823 |
|New Hampshire |5667 |
|Montana       |5076 |
|Nevada        |5073 |
|Kentucky      |5010 |
|North Dakota  |4360 |
|Vermont       |4348 |
|Wyoming       |3594 |
|South Dakota  |2868 |
+--------------+-----+

Total number of valid records: 1016102
totalValidRecords: Long = 1016102
```

Took 7 sec. Last updated by anonymous at November 22 2024, 1:02:31 PM. (outdated)

```
// Identify records with invalid or missing states
val invalidStatesDF = cleanedReducedWalmartDF.filter(!col("state").rlike("^[a-zA-Z\\s]+$"))

// Show the distinct invalid state values
invalidStatesDF.select("state").show(100, truncate = false)

// Count the number of records with invalid states
val totalInvalidRecords = invalidStatesDF.count()
println(s"Total number of records with invalid or missing states: $totalInvalidRecords")
```

SPARK JOB FINISHED

```
|132.45767969838714 |
|133.95069909097137 |
|460.33871954472744 |
|3[illegible]22408  |
|10962.011061470992 |
|19.198167058807538 |
|3.153306478699905  |
|65.88772863984471  |
|250.74865994561415 |
|159.6771237197717  |
|281.48454244344885 |
|80.29012480810394  |
+-------------------+
only showing top 100 rows

Total number of records with invalid or missing states: 18270
invalidStatesDF: org.apache.spark.sql.Dataset[org.apache.spark.sql.Row] = [city: string, custo
mer_age: string ... 12 more fields]
```

Took 4 sec. Last updated by anonymous at November 22 2024, 1:05:47 PM. (outdated)

Cleaning Up Data again

SPARK JOB FINISHED

```
// Filtering out records with invalid or missing state values
val cleanedValidStatesDF = cleanedReducedWalmartDF.filter(col("state").rlike("^[a-zA-Z\\s]+$")
val totalCleanedRecords = cleanedValidStatesDF.count()
println(s"Total number of records after dropping invalid states: $totalCleanedRecords")
```

```
Total number of records after dropping invalid states: 1016102
cleanedValidStatesDF: org.apache.spark.sql.Dataset[org.apache.spark.sql.Row] = [city: string,
customer_age: string ... 12 more fields]
totalCleanedRecords: Long = 1016102
```

Took 4 sec. Last updated by anonymous at November 22 2024, 1:10:01 PM. (outdated)

Continuing with our State wise Analysis

SPARK JOB FINISHED

```
// 1. Unique States
val uniqueStates = cleanedValidStatesDF.select("state").distinct().collect().map(_.getString(0
println("Unique States:\n" + uniqueStates.mkString(", "))
```

```
Unique States:
Minnesota, Nebraska, Oklahoma, MO, Maryland, Kentucky, New York, Connecticut, Kansas, Alabama,
Iowa, Florida, Indiana, Arkansas, North Dakota, Pennsylvania, Illinois, North Carolina, Arizon
a, South Dakota, Colorado, Maine, Ohio, Oregon, Texas, Vermont, Nevada, New Mexico, Montana, V
irginia, South Carolina, Utah, Washington, West Virginia, Rhode Island, Georgia, Michigan, Wyo
ming, New Jersey, Louisiana, Mississippi, Tennessee, New Hampshire, MA, Idaho, California, Wis
consin
uniqueStates: Array[String] = Array(Minnesota, Nebraska, Oklahoma, MO, Maryland, Kentucky, New
York, Connecticut, Kansas, Alabama, Iowa, Florida, Indiana, Arkansas, North Dakota, Pennsylvan
ia, Illinois, North Carolina, Arizona, South Dakota, Colorado, Maine, Ohio, Oregon, Texas, Ver
mont, Nevada, New Mexico, Montana, Virginia, South Carolina, Utah, Washington, West Virginia,
Rhode Island, Georgia, Michigan, Wyoming, New Jersey, Louisiana, Mississippi, Tennessee, New H
ampshire, MA, Idaho, California, Wisconsin)
```

Took 3 sec. Last updated by anonymous at November 22 2024, 1:11:27 PM. (outdated)

```
// 2. Number of Unique Cities per State
val stateCitiesDF = cleanedValidStatesDF.groupBy("state")
  .agg(countDistinct("city").alias("number_of_unique_cities"))
  .orderBy(desc("number_of_unique_cities"))
```

FINISHED

Walmart_Data_Prep...

```
stateCitiesDF: org.apache.spark.sql.Dataset[org.apache.spark.sql.Row] = [state: string, number
_of_unique_cities: bigint]
```

Took 1 sec. Last updated by anonymous at November 22 2024, 1:11:47 PM. (outdated)

```
println("Number of Unique Cities per State:")
stateCitiesDF.show(100, truncate = false)
```

SPARK JOB FINISHED

```
Number of Unique Cities per State:
+--------------+----------------------+
|state         |number_of_unique_cities|
+--------------+----------------------+
|California    |108                   |
|Texas         |102                   |
|MA            |101                   |
|New Jersey    |92                    |
|Florida       |80                    |
|Illinois      |71                    |
|Ohio          |58                    |
|New York      |54                    |
|Michigan      |44                    |
|Washington    |40                    |
|Minnesota     |37                    |
|Pennsylvania  |37                    |
|North Carolina|35                    |
|Connecticut   |33                    |
```

Took 3 sec. Last updated by anonymous at November 22 2024, 1:12:01 PM. (outdated)

```
cleanedValidStatesDF.describe().show()
```

SPARK JOB FINISHED

```
+-------+--------+------------------+---------------+------------------+------------------+-
--------------+--------------------+--------------------+------------------+-------+---------
----------+-------+------------------+
|summary|    city|      customer_age|customer_segment|          discount|    order_quantity|p
roduct_category|        product_name|product_sub_category|            profit| region|       sh
ipping_cost|  state|          zip_code|
+-------+--------+------------------+---------------+------------------+------------------+-
--------------+--------------------+--------------------+------------------+-------+---------
----------+-------+------------------+
|  count| 1016102|           1016102|        1016102|           1016102|           1016102|
1016102|             1016102|             1016102|           1016102|1016102|             1016
102|1016102|           1016102|
|   mean|    NULL| 55.00850406750504|           NULL| 0.1250008070055965|25.490501937797582|
NULL|   4612.694820017559|                NULL| 6561.502601757924|   NULL|  1167.7360835099732
|   NULL| 47035.51055405855|
| stddev|    NULL|20.221224787697658|           NULL|0.07224230086886456| 14.15356501828778|
NULL|   2628.850817250653|                NULL|11941.983613210834|   NULL|    4988.65435239251
|   NULL|30958.734201602382|
```

Took 21 sec. Last updated by anonymous at November 22 2024, 1:20:19 PM. (outdated)

```
cleanedValidStatesDF.summary().show(truncate = false)
```

SPARK JOB FINISHED

Walmart_Data_Prep...

```
+-------+-------+------------------+----------------+------------------+------------------+-----------------+-------
---------------+------------------------------+--------------------+------------------+-------+-----------------+
|summary|city   |customer_age      |customer_segment|discount          |order_quantity    |p
roduct_category|product_name                  |product_sub_category|profit            |region
|shipping_cost       |state  |zip_code         |
+-------+-------+------------------+----------------+------------------+------------------+-----------------+-
--------------+------------------------------+--------------------+------------------+-------
+--------------------+-------+-----------------+
|count  |1016102 |1016102           |1016102         |1016102           |1016102           |1
016102         |1016102                       |1016102             |1016102           |1016102
|1016102             |1016102|1016102          |
|mean   |NULL    |55.00850406750504 |NULL            |0.1250008070055965 |25.490501937797582|N
ULL            |4612.694820017559             |NULL                |6561.502601757924 |NULL
|1167.7360835099732  |NULL   |47035.51055405855|
|stddev |NULL    |20.221224787697654|NULL            |0.07224230086886456|14.15356501828778 |N
ULL            |2628.850817250653             |NULL                |11941.983613210834|NULL
```

Took 49 sec. Last updated by anonymous at November 22 2024, 1:21:08 PM. (outdated)

Saving the DataFrame as Parquet

SPARK JOB (http://nyu-dataproc-w-0.c.hpc-dataproc-19b8.internal:39359/jobs/job?id=337) FINISHED

```
val outputPath = "/user/sc10670_nyu_edu/project/cleaned-walmart-data.parquet"
cleanedValidStatesDF.write.mode("overwrite").parquet(outputPath)
println(s"DataFrame saved to Parquet at: $outputPath")
```

```
DataFrame saved to Parquet at: /user/sc10670_nyu_edu/project/cleaned-walmart-data.parquet
outputPath: String = /user/sc10670_nyu_edu/project/cleaned-walmart-data.parquet
```

Took 8 sec. Last updated by anonymous at November 22 2024, 1:22:35 PM. (outdated)

READY