Module 3 - Data Integration & Aggregation

April 19, 2025

[1]:
```
from pyspark.sql import SparkSession
spark = SparkSession.builder \
.appName('OlistData') \
.getOrCreate()
```

```
25/04/19 06:51:09 WARN SparkSession: Using an existing Spark session; only
runtime SQL configurations will take effect.
```

[2]:
```
hdfs_path = '/data/olist/'
```

[3]:
```
customers_df = spark.read.csv(hdfs_path + 'olist_customers_dataset.
 ↪csv',header=True,inferSchema=True)
category_translation_df = spark.read.csv(hdfs_path +␣
 ↪'product_category_name_translation.csv',header=True,inferSchema=True)
orders_df = spark.read.csv(hdfs_path + 'olist_orders_dataset.
 ↪csv',header=True,inferSchema=True)
order_item_df = spark.read.csv(hdfs_path + 'olist_order_items_dataset.
 ↪csv',header=True,inferSchema=True)
payments_df = spark.read.csv(hdfs_path + 'olist_order_payments_dataset.
 ↪csv',header=True,inferSchema=True)
reviews_df = spark.read.csv(hdfs_path + 'olist_order_reviews_dataset.
 ↪csv',header=True,inferSchema=True)
products_df = spark.read.csv(hdfs_path + 'olist_products_dataset.
 ↪csv',header=True,inferSchema=True)
sellers_df = spark.read.csv(hdfs_path + 'olist_sellers_dataset.
 ↪csv',header=True,inferSchema=True)
geolocation_df = spark.read.csv(hdfs_path + 'olist_geolocation_dataset.
 ↪csv',header=True,inferSchema=True)
```

[4]:
```
## Cache frequently used data for better performance

orders_df.cache()
customers_df.cache()
order_item_df.cache()
```

[4]:
```
DataFrame[order_id: string, order_item_id: int, product_id: string, seller_id:
string, shipping_limit_date: timestamp, price: double, freight_value: double]
```

[5]:
```
orders_items_joined_df = orders_df.join(order_item_df,'order_id','inner')
```

[6]:
```
orders_items_products_df = orders_items_joined_df.
 ↪join(products_df,'product_id','inner')
```

[7]:
```
orders_items_products_sellers_df = orders_items_products_df.
 ↪join(sellers_df,'seller_id','inner')
```

[8]:
```
full_orders_df = orders_items_products_sellers_df.
 ↪join(customers_df,'customer_id','inner')
```

[11]:
```
# Geolocation data

full_orders_df = full_orders_df.join(geolocation_df,full_orders_df.
 ↪customer_zip_code_prefix == geolocation_df.
 ↪geolocation_zip_code_prefix,'left')
```

[12]:
```
full_orders_df = full_orders_df.join(reviews_df,'order_id','left')
```

[13]:
```
full_orders_df = full_orders_df.join(payments_df,'order_id','left')
```

[14]:
```
full_orders_df.cache()
```

```
25/04/19 07:06:24 WARN SparkStringUtils: Truncated the string representation of
a plan since it was too large. This behavior can be adjusted by setting
'spark.sql.debug.maxToStringFields'.
```

[14]:
```
DataFrame[order_id: string, customer_id: string, seller_id: string, product_id:
string, order_status: string, order_purchase_timestamp: timestamp,
order_approved_at: timestamp, order_delivered_carrier_date: timestamp,
order_delivered_customer_date: timestamp, order_estimated_delivery_date:
timestamp, order_item_id: int, shipping_limit_date: timestamp, price: double,
freight_value: double, product_category_name: string, product_name_lenght: int,
product_description_lenght: int, product_photos_qty: int, product_weight_g: int,
product_length_cm: int, product_height_cm: int, product_width_cm: int,
seller_zip_code_prefix: int, seller_city: string, seller_state: string,
customer_unique_id: string, customer_zip_code_prefix: int, customer_city:
string, customer_state: string, geolocation_zip_code_prefix: int,
geolocation_lat: double, geolocation_lng: double, geolocation_city: string,
geolocation_state: string, review_id: string, review_score: string,
review_comment_title: string, review_comment_message: string,
review_creation_date: string, review_answer_timestamp: string,
payment_sequential: int, payment_type: string, payment_installments: int,
payment_value: double]
```

[16]:
```
from pyspark.sql.functions import *
```

[17]:
```
# Total Revenues Per Seller

seller_revenue_df = full_orders_df.groupBy('seller_id').agg(sum('price').
 ↪alias('total_revenue'))
```

[18]:
```
seller_revenue_df.show(5)
```

```
[Stage 34:=====================================================>(995 + 2) / 1000]

+--------------------+--------------------+
|           seller_id|       total_revenue|
+--------------------+--------------------+
|3d5d0dc7073a299e3…|            170639.6|
|2138ccb85b11a4ec1…|  1943866.0699999996|
|33ac3e28642ab8bda…|   615628.8499999986|
|cc419e0650a3c5ba7…|1.4751464500000061E7|
|8e6cc767478edae94…|  1145757.4000000013|
+--------------------+--------------------+
only showing top 5 rows
```

0.1 Optimized joins for Data Integration

[19]:
```
orders_items_seller_df = orders_items_products_df.
 ↪join(broadcast(sellers_df),'seller_id','inner')
```

[20]:
```
full_orders_df = orders_items_seller_df.join(customers_df,'customer_id','inner')
```

[21]:
```
full_orders_df = full_orders_df.join(broadcast(geolocation_df),full_orders_df.
 ↪customer_zip_code_prefix == geolocation_df.
 ↪geolocation_zip_code_prefix,'left')
```

[22]:
```
full_orders_df = full_orders_df.join(broadcast(reviews_df),'order_id','left')
```

[23]:
```
full_orders_df = full_orders_df.join(payments_df,'order_id','left')
```

[24]:
```
full_orders_df.cache()
```

```
25/04/19 07:18:41 WARN CacheManager: Asked to cache already cached data.
```

[24]:
```
DataFrame[order_id: string, customer_id: string, seller_id: string, product_id:
string, order_status: string, order_purchase_timestamp: timestamp,
order_approved_at: timestamp, order_delivered_carrier_date: timestamp,
order_delivered_customer_date: timestamp, order_estimated_delivery_date:
timestamp, order_item_id: int, shipping_limit_date: timestamp, price: double,
```

```
freight_value: double, product_category_name: string, product_name_lenght: int,
product_description_lenght: int, product_photos_qty: int, product_weight_g: int,
product_length_cm: int, product_height_cm: int, product_width_cm: int,
seller_zip_code_prefix: int, seller_city: string, seller_state: string,
customer_unique_id: string, customer_zip_code_prefix: int, customer_city:
string, customer_state: string, geolocation_zip_code_prefix: int,
geolocation_lat: double, geolocation_lng: double, geolocation_city: string,
geolocation_state: string, review_id: string, review_score: string,
review_comment_title: string, review_comment_message: string,
review_creation_date: string, review_answer_timestamp: string,
payment_sequential: int, payment_type: string, payment_installments: int,
payment_value: double]
```

0.2 Aggregation

[26]:
```
#Total order per customer
customer_order_count_df = full_orders_df.groupBy('customer_id')\
.agg(count('order_id').alias('total_orders'))\
.orderBy(desc('total_orders'))

customer_order_count_df.show(5)
```

```
[Stage 45:==============>                                          (1 + 3) / 4]
```

```
+--------------------+------------+
|         customer_id|total_orders|
+--------------------+------------+
|351e40989da90e704…|       11427|
|50920f8cd0681fd86…|       10752|
|9b43e2a62de9bab3a…|        8556|
|270c23a11d024a44c…|        8001|
|5c87184371002d49e…|        6876|
+--------------------+------------+
only showing top 5 rows
```

[31]:
```
#Average review score per seller
seller_review_df = full_orders_df.groupBy('seller_id')\
.agg(count('review_score').alias('avg_review_score'))\
.orderBy(desc('avg_review_score'))

seller_review_df.show(5)
```

```
[Stage 76:=====================================================>(998 + 2) / 1000]
```

```
+--------------------+----------------+
|           seller_id|avg_review_score|
+--------------------+----------------+
```

```
|4a3ca9315b744ce9f…|          328216|
|1f50f920176fa81da…|          295792|
|6560211a19b47992c…|          284533|
|da8622b14eb17ae28…|          263122|
|cc419e0650a3c5ba7…|          254431|
+--------------------+----------------+
only showing top 5 rows
```

[30]:
```
# Top 10 most sold products

top_products_df = full_orders_df.groupBy('product_id')\
.agg(count('order_id').alias('total_sold'))\
.orderBy(desc('total_sold'))

top_products_df.show(10)
```

```
[Stage 73:=============================================>             (3 + 1) / 4]
+--------------------+----------+
|          product_id|total_sold|
+--------------------+----------+
|aca2eb7d00ea1a7b8…|     86740|
|422879e10f4668299…|     81110|
|99a4788cb24856965…|     78775|
|389d119b48cf3043d…|     60248|
|d1c427060a0f73f6b…|     59274|
|368c6c730842d7801…|     58358|
|53759a2ecddad2bb8…|     52654|
|53b36df67ebb7c415…|     52105|
|154e7e31ebfa09220…|     42700|
|3dd2a17168ec895c7…|     40787|
+--------------------+----------+
only showing top 10 rows
```

0.3 Window Function And Ranking

[32]:
```
from pyspark.sql.window import Window
```

[35]:
```
window_spec = Window.partitionBy('seller_id').orderBy(desc('price'))
```

[37]:
```
## Rank Top Selling Products Per Seller
```

```
top_seller_products_df = full_orders_df.withColumn('rank',rank().
  ↪over(window_spec)).filter(col('rank')<=5)
top_seller_products_df.show()
```

```
[Stage 94:>                                                         (0 + 1) / 1]
+--------------------+--------------------+--------------------+---------------
----+------------+-----------------------+--------------------+---------------
----------+----------------------------+----------------------------+--------
-----+--------------------+-----+------------+---------------------+-----------
-------+--------------------------+------------------+---------------+---------
--------+----------------+---------------+----------------------+----------+-
----------+--------------------+-----------------------+--------------------+--
------------+----------------------------+------------------+-------------------
+--------------------+----------------+--------------------+-----------+-------
------------+-----------------------+--------------------+---------------------
-+------------------+-----------+--------------------+------------+----+
|            order_id|         customer_id|           seller_id|
product_id|order_status|order_purchase_timestamp|  order_approved_at|order_deliv
ered_carrier_date|order_delivered_customer_date|order_estimated_delivery_date|or
der_item_id|shipping_limit_date|price|freight_value|product_category_name|produc
t_name_lenght|product_description_lenght|product_photos_qty|product_weight_g|pro
duct_length_cm|product_height_cm|product_width_cm|seller_zip_code_prefix|seller_
city|seller_state|  customer_unique_id|customer_zip_code_prefix|
customer_city|customer_state|geolocation_zip_code_prefix|    geolocation_lat|
geolocation_lng|   geolocation_city|geolocation_state|           review_id|revie
w_score|review_comment_title|review_comment_message|review_creation_date|review_
answer_timestamp|payment_sequential|payment_type|payment_installments|payment_va
lue|rank|
+--------------------+--------------------+--------------------+---------------
----+------------+-----------------------+--------------------+---------------
----------+----------------------------+----------------------------+--------
-----+--------------------+-----+------------+---------------------+-----------
-------+--------------------------+------------------+---------------+---------
--------+----------------+---------------+----------------------+----------+-
----------+--------------------+-----------------------+--------------------+--
------------+----------------------------+------------------+-------------------
+--------------------+----------------+--------------------+-----------+-------
------------+-----------------------+--------------------+---------------------
-+------------------+-----------+--------------------+------------+----+
|7f39ba4c9052be115…|d7fc82cbeafea77bd…|0015a82c2db000af6…|a2ff5a97bf95719e
3…|   delivered|     2017-10-18 08:16:34|2017-10-18 23:56:20|
2017-10-20 14:29:01|          2017-10-27 16:46:05|          2017-11-09 00:00:00|
1|2017-10-24 23:56:20|895.0|        21.02|      eletroportateis|
40|                        849|                 2|           11800|
40|               43|             36|                   9080|santo andre|
SP|9de5797cddb925987…|                    35490|entre rios de minas|
MG|                      35490|-20.740761932788878|-44.055832659230184|entre
```

```
rios de minas|               MG|2abb25fde5aafe9bc…|           1|
NULL|  Produto preto, re…| 2017-10-28 00:00:00|    2017-10-28 14:32:42|
1| credit_card|                   8|       916.02|   1|
|7f39ba4c9052be115…|d7fc82cbeafea77bd…|0015a82c2db000af6…|a2ff5a97bf95719e
3…|   delivered|     2017-10-18 08:16:34|2017-10-18 23:56:20|
2017-10-20 14:29:01|          2017-10-27 16:46:05|          2017-11-09 00:00:00|
1|2017-10-24 23:56:20|895.0|        21.02|      eletroportateis|
40|                       849|                2|           11800|
40|               43|             36|                  9080|santo andre|
SP|9de5797cddb925987…|                  35490|entre rios de minas|
MG|                     35490|-20.665871069633294| -44.06965958216196|entre
rios de minas|               MG|2abb25fde5aafe9bc…|           1|
NULL|  Produto preto, re…| 2017-10-28 00:00:00|    2017-10-28 14:32:42|
1| credit_card|                   8|       916.02|   1|
|7f39ba4c9052be115…|d7fc82cbeafea77bd…|0015a82c2db000af6…|a2ff5a97bf95719e
3…|   delivered|     2017-10-18 08:16:34|2017-10-18 23:56:20|
2017-10-20 14:29:01|          2017-10-27 16:46:05|          2017-11-09 00:00:00|
1|2017-10-24 23:56:20|895.0|        21.02|      eletroportateis|
40|                       849|                2|           11800|
40|               43|             36|                  9080|santo andre|
SP|9de5797cddb925987…|                  35490|entre rios de minas|
MG|                     35490|  -20.6742343929942| -44.06401792282068|entre
rios de minas|               MG|2abb25fde5aafe9bc…|           1|
NULL|  Produto preto, re…| 2017-10-28 00:00:00|    2017-10-28 14:32:42|
1| credit_card|                   8|       916.02|   1|
|7f39ba4c9052be115…|d7fc82cbeafea77bd…|0015a82c2db000af6…|a2ff5a97bf95719e
3…|   delivered|     2017-10-18 08:16:34|2017-10-18 23:56:20|
2017-10-20 14:29:01|          2017-10-27 16:46:05|          2017-11-09 00:00:00|
1|2017-10-24 23:56:20|895.0|        21.02|      eletroportateis|
40|                       849|                2|           11800|
40|               43|             36|                  9080|santo andre|
SP|9de5797cddb925987…|                  35490|entre rios de minas|
MG|                     35490|-20.669980175244774| -44.06275615289615|entre
rios de minas|               MG|2abb25fde5aafe9bc…|           1|
NULL|  Produto preto, re…| 2017-10-28 00:00:00|    2017-10-28 14:32:42|
1| credit_card|                   8|       916.02|   1|
|7f39ba4c9052be115…|d7fc82cbeafea77bd…|0015a82c2db000af6…|a2ff5a97bf95719e
3…|   delivered|     2017-10-18 08:16:34|2017-10-18 23:56:20|
2017-10-20 14:29:01|          2017-10-27 16:46:05|          2017-11-09 00:00:00|
1|2017-10-24 23:56:20|895.0|        21.02|      eletroportateis|
40|                       849|                2|           11800|
40|               43|             36|                  9080|santo andre|
SP|9de5797cddb925987…|                  35490|entre rios de minas|
MG|                     35490| -20.66742772064256|-44.064704147398096|entre
rios de minas|               MG|2abb25fde5aafe9bc…|           1|
NULL|  Produto preto, re…| 2017-10-28 00:00:00|    2017-10-28 14:32:42|
1| credit_card|                   8|       916.02|   1|
|7f39ba4c9052be115…|d7fc82cbeafea77bd…|0015a82c2db000af6…|a2ff5a97bf95719e
```

```
3…|   delivered|     2017-10-18 08:16:34|2017-10-18 23:56:20|
2017-10-20 14:29:01|          2017-10-27 16:46:05|          2017-11-09 00:00:00|
1|2017-10-24 23:56:20|895.0|        21.02|      eletroportateis|
40|                        849|                2|           11800|
40|               43|              36|                  9080|santo andre|
SP|9de5797cddb925987…|                   35490|entre rios de minas|
MG|                      35490|-20.678210850848185|-44.061081660598575|entre
rios de minas|               MG|2abb25fde5aafe9bc…|           1|
NULL|  Produto preto, re…| 2017-10-28 00:00:00|    2017-10-28 14:32:42|
1| credit_card|                    8|       916.02|   1|
|7f39ba4c9052be115…|d7fc82cbeafea77bd…|0015a82c2db000af6…|a2ff5a97bf95719e
3…|   delivered|     2017-10-18 08:16:34|2017-10-18 23:56:20|
2017-10-20 14:29:01|          2017-10-27 16:46:05|          2017-11-09 00:00:00|
1|2017-10-24 23:56:20|895.0|        21.02|      eletroportateis|
40|                        849|                2|           11800|
40|               43|              36|                  9080|santo andre|
SP|9de5797cddb925987…|                   35490|entre rios de minas|
MG|                      35490|-20.661400509606654| -44.06971920384266|entre
rios de minas|               MG|2abb25fde5aafe9bc…|           1|
NULL|  Produto preto, re…| 2017-10-28 00:00:00|    2017-10-28 14:32:42|
1| credit_card|                    8|       916.02|   1|
|7f39ba4c9052be115…|d7fc82cbeafea77bd…|0015a82c2db000af6…|a2ff5a97bf95719e
3…|   delivered|     2017-10-18 08:16:34|2017-10-18 23:56:20|
2017-10-20 14:29:01|          2017-10-27 16:46:05|          2017-11-09 00:00:00|
1|2017-10-24 23:56:20|895.0|        21.02|      eletroportateis|
40|                        849|                2|           11800|
40|               43|              36|                  9080|santo andre|
SP|9de5797cddb925987…|                   35490|entre rios de minas|
MG|                      35490| -20.66742772064256|-44.064704147398096|entre
rios de minas|               MG|2abb25fde5aafe9bc…|           1|
NULL|  Produto preto, re…| 2017-10-28 00:00:00|    2017-10-28 14:32:42|
1| credit_card|                    8|       916.02|   1|
|7f39ba4c9052be115…|d7fc82cbeafea77bd…|0015a82c2db000af6…|a2ff5a97bf95719e
3…|   delivered|     2017-10-18 08:16:34|2017-10-18 23:56:20|
2017-10-20 14:29:01|          2017-10-27 16:46:05|          2017-11-09 00:00:00|
1|2017-10-24 23:56:20|895.0|        21.02|      eletroportateis|
40|                        849|                2|           11800|
40|               43|              36|                  9080|santo andre|
SP|9de5797cddb925987…|                   35490|entre rios de minas|
MG|                      35490|-20.664249953387557| -44.07055875232306|entre
rios de minas|               MG|2abb25fde5aafe9bc…|           1|
NULL|  Produto preto, re…| 2017-10-28 00:00:00|    2017-10-28 14:32:42|
1| credit_card|                    8|       916.02|   1|
|7f39ba4c9052be115…|d7fc82cbeafea77bd…|0015a82c2db000af6…|a2ff5a97bf95719e
3…|   delivered|     2017-10-18 08:16:34|2017-10-18 23:56:20|
2017-10-20 14:29:01|          2017-10-27 16:46:05|          2017-11-09 00:00:00|
1|2017-10-24 23:56:20|895.0|        21.02|      eletroportateis|
40|                        849|                2|           11800|
```

```
40|               43|              36|                  9080|santo andre|
SP|9de5797cddb925987…|                   35490|entre rios de minas|
MG|                      35490|-20.666165502713298|-44.064587627704086|entre
rios de minas|               MG|2abb25fde5aafe9bc…|           1|
NULL|  Produto preto, re…| 2017-10-28 00:00:00|    2017-10-28 14:32:42|
1| credit_card|                   8|       916.02|   1|
|7f39ba4c9052be115…|d7fc82cbeafea77bd…|0015a82c2db000af6…|a2ff5a97bf95719e
3…|   delivered|     2017-10-18 08:16:34|2017-10-18 23:56:20|
2017-10-20 14:29:01|          2017-10-27 16:46:05|          2017-11-09 00:00:00|
1|2017-10-24 23:56:20|895.0|        21.02|      eletroportateis|
40|                       849|                2|           11800|
40|               43|              36|                  9080|santo andre|
SP|9de5797cddb925987…|                   35490|entre rios de minas|
MG|                      35490|-20.666335590361236| -44.07060940482346|entre
rios de minas|               MG|2abb25fde5aafe9bc…|           1|
NULL|  Produto preto, re…| 2017-10-28 00:00:00|    2017-10-28 14:32:42|
1| credit_card|                   8|       916.02|   1|
|7f39ba4c9052be115…|d7fc82cbeafea77bd…|0015a82c2db000af6…|a2ff5a97bf95719e
3…|   delivered|     2017-10-18 08:16:34|2017-10-18 23:56:20|
2017-10-20 14:29:01|          2017-10-27 16:46:05|          2017-11-09 00:00:00|
1|2017-10-24 23:56:20|895.0|        21.02|      eletroportateis|
40|                       849|                2|           11800|
40|               43|              36|                  9080|santo andre|
SP|9de5797cddb925987…|                   35490|entre rios de minas|
MG|                      35490| -20.65876306456585|-44.064003041859806|entre
rios de minas|               MG|2abb25fde5aafe9bc…|           1|
NULL|  Produto preto, re…| 2017-10-28 00:00:00|    2017-10-28 14:32:42|
1| credit_card|                   8|       916.02|   1|
|7f39ba4c9052be115…|d7fc82cbeafea77bd…|0015a82c2db000af6…|a2ff5a97bf95719e
3…|   delivered|     2017-10-18 08:16:34|2017-10-18 23:56:20|
2017-10-20 14:29:01|          2017-10-27 16:46:05|          2017-11-09 00:00:00|
1|2017-10-24 23:56:20|895.0|        21.02|      eletroportateis|
40|                       849|                2|           11800|
40|               43|              36|                  9080|santo andre|
SP|9de5797cddb925987…|                   35490|entre rios de minas|
MG|                      35490| -20.67834561184666| -44.06101428009934|entre
rios de minas|               MG|2abb25fde5aafe9bc…|           1|
NULL|  Produto preto, re…| 2017-10-28 00:00:00|    2017-10-28 14:32:42|
1| credit_card|                   8|       916.02|   1|
|7f39ba4c9052be115…|d7fc82cbeafea77bd…|0015a82c2db000af6…|a2ff5a97bf95719e
3…|   delivered|     2017-10-18 08:16:34|2017-10-18 23:56:20|
2017-10-20 14:29:01|          2017-10-27 16:46:05|          2017-11-09 00:00:00|
1|2017-10-24 23:56:20|895.0|        21.02|      eletroportateis|
40|                       849|                2|           11800|
40|               43|              36|                  9080|santo andre|
SP|9de5797cddb925987…|                   35490|entre rios de minas|
MG|                      35490|-20.673322674182376| -44.06039437386022|entre
rios de minas|               MG|2abb25fde5aafe9bc…|           1|
```

```
NULL|  Produto preto, re…| 2017-10-28 00:00:00|    2017-10-28 14:32:42|
1| credit_card|                   8|       916.02|   1|
|7f39ba4c9052be115…|d7fc82cbeafea77bd…|0015a82c2db000af6…|a2ff5a97bf95719e
3…|   delivered|     2017-10-18 08:16:34|2017-10-18 23:56:20|
2017-10-20 14:29:01|          2017-10-27 16:46:05|          2017-11-09 00:00:00|
1|2017-10-24 23:56:20|895.0|        21.02|      eletroportateis|
40|                        849|                2|           11800|
40|               43|             36|                  9080|santo andre|
SP|9de5797cddb925987…|                   35490|entre rios de minas|
MG|                      35490| -20.67617624228519|-44.062597094524655|entre
rios de minas|               MG|2abb25fde5aafe9bc…|           1|
NULL|  Produto preto, re…| 2017-10-28 00:00:00|    2017-10-28 14:32:42|
1| credit_card|                   8|       916.02|   1|
|7f39ba4c9052be115…|d7fc82cbeafea77bd…|0015a82c2db000af6…|a2ff5a97bf95719e
3…|   delivered|     2017-10-18 08:16:34|2017-10-18 23:56:20|
2017-10-20 14:29:01|          2017-10-27 16:46:05|          2017-11-09 00:00:00|
1|2017-10-24 23:56:20|895.0|        21.02|      eletroportateis|
40|                        849|                2|           11800|
40|               43|             36|                  9080|santo andre|
SP|9de5797cddb925987…|                   35490|entre rios de minas|
MG|                      35490|-20.667494720092538| -44.06497892111579|entre
rios de minas|               MG|2abb25fde5aafe9bc…|           1|
NULL|  Produto preto, re…| 2017-10-28 00:00:00|    2017-10-28 14:32:42|
1| credit_card|                   8|       916.02|   1|
|7f39ba4c9052be115…|d7fc82cbeafea77bd…|0015a82c2db000af6…|a2ff5a97bf95719e
3…|   delivered|     2017-10-18 08:16:34|2017-10-18 23:56:20|
2017-10-20 14:29:01|          2017-10-27 16:46:05|          2017-11-09 00:00:00|
1|2017-10-24 23:56:20|895.0|        21.02|      eletroportateis|
40|                        849|                2|           11800|
40|               43|             36|                  9080|santo andre|
SP|9de5797cddb925987…|                   35490|entre rios de minas|
MG|                      35490| -20.65876306456585|-44.064003041859806|entre
rios de minas|               MG|2abb25fde5aafe9bc…|           1|
NULL|  Produto preto, re…| 2017-10-28 00:00:00|    2017-10-28 14:32:42|
1| credit_card|                   8|       916.02|   1|
|7f39ba4c9052be115…|d7fc82cbeafea77bd…|0015a82c2db000af6…|a2ff5a97bf95719e
3…|   delivered|     2017-10-18 08:16:34|2017-10-18 23:56:20|
2017-10-20 14:29:01|          2017-10-27 16:46:05|          2017-11-09 00:00:00|
1|2017-10-24 23:56:20|895.0|        21.02|      eletroportateis|
40|                        849|                2|           11800|
40|               43|             36|                  9080|santo andre|
SP|9de5797cddb925987…|                   35490|entre rios de minas|
MG|                      35490|-20.664819453600334|-44.070168185621846|entre
rios de minas|               MG|2abb25fde5aafe9bc…|           1|
NULL|  Produto preto, re…| 2017-10-28 00:00:00|    2017-10-28 14:32:42|
1| credit_card|                   8|       916.02|   1|
|7f39ba4c9052be115…|d7fc82cbeafea77bd…|0015a82c2db000af6…|a2ff5a97bf95719e
3…|   delivered|     2017-10-18 08:16:34|2017-10-18 23:56:20|
```

```
2017-10-20 14:29:01|          2017-10-27 16:46:05|          2017-11-09 00:00:00|
1|2017-10-24 23:56:20|895.0|        21.02|      eletroportateis|
40|                       849|                2|           11800|
40|               43|              36|                  9080|santo andre|
SP|9de5797cddb925987…|                   35490|entre rios de minas|
MG|                      35490| -20.67359670349073| -44.06603636432115|entre
rios de minas|               MG|2abb25fde5aafe9bc…|           1|
NULL|  Produto preto, re…| 2017-10-28 00:00:00|    2017-10-28 14:32:42|
1| credit_card|                   8|       916.02|   1|
|7f39ba4c9052be115…|d7fc82cbeafea77bd…|0015a82c2db000af6…|a2ff5a97bf95719e
3…|   delivered|     2017-10-18 08:16:34|2017-10-18 23:56:20|
2017-10-20 14:29:01|          2017-10-27 16:46:05|          2017-11-09 00:00:00|
1|2017-10-24 23:56:20|895.0|        21.02|      eletroportateis|
40|                       849|                2|           11800|
40|               43|              36|                  9080|santo andre|
SP|9de5797cddb925987…|                   35490|entre rios de minas|
MG|                      35490| -20.67276040736089| -44.05996226520918|entre
rios de minas|               MG|2abb25fde5aafe9bc…|           1|
NULL|  Produto preto, re…| 2017-10-28 00:00:00|    2017-10-28 14:32:42|
1| credit_card|                   8|       916.02|   1|
+--------------------+--------------------+--------------------+----------------
----+------------+-----------------------+------------------+----------------
-----------+-----------------------------+-----------------------------+--------
-----+--------------------+-----+-------------+---------------------+------------
-------+-------------------------+------------------+---------------+---------
--------+-----------------+----------------+----------------------+-----------+-
-----------+--------------------+-----------------------+--------------------+--
-----------+-----------------------------+------------------+--------------------
+--------------------+----------------+--------------------+-----------+-------
------------+----------------------+--------------------+----------------------
-+--------------------+------------+--------------------+-------------+----+
only showing top 20 rows
```

[38]:
```
top_seller_products_df.select('seller_id','price','rank').show()
```

```
[Stage 101:>                                                        (0 + 1) / 1]

+--------------------+-----+----+
|           seller_id|price|rank|
+--------------------+-----+----+
|0015a82c2db000af6…|895.0|   1|
|0015a82c2db000af6…|895.0|   1|
|0015a82c2db000af6…|895.0|   1|
|0015a82c2db000af6…|895.0|   1|
|0015a82c2db000af6…|895.0|   1|
|0015a82c2db000af6…|895.0|   1|
```

```
|0015a82c2db000af6…|895.0|   1|
|0015a82c2db000af6…|895.0|   1|
|0015a82c2db000af6…|895.0|   1|
|0015a82c2db000af6…|895.0|   1|
|0015a82c2db000af6…|895.0|   1|
|0015a82c2db000af6…|895.0|   1|
|0015a82c2db000af6…|895.0|   1|
|0015a82c2db000af6…|895.0|   1|
|0015a82c2db000af6…|895.0|   1|
|0015a82c2db000af6…|895.0|   1|
|0015a82c2db000af6…|895.0|   1|
|0015a82c2db000af6…|895.0|   1|
|0015a82c2db000af6…|895.0|   1|
+--------------------+-----+----+
only showing top 20 rows
```

0.4 Advance Aggregation and Enrichment

[ ]:
```
full_orders_df
```

[41]:
```
# Total revenue & average order value(AOV) per customer

customer_spending_df = full_orders_df.groupBy('customer_id')\
.agg(
count('order_id').alias('total_orders'),
sum('price').alias('total_spent'),
round(avg('price'),2).alias('AOV')
)\
.orderBy(desc('total_spent'))

customer_spending_df.show()
```

```
[Stage 108:>                                                        (0 + 4) / 4]
+--------------------+------------+------------------+-------+
|         customer_id|total_orders|       total_spent|    AOV|
+--------------------+------------+------------------+-------+
|d3e82ccec3cb5f956…|        6876|         6662844.0|  969.0|
|df55c14d1476a9a34…|         743|         3565657.0| 4799.0|
|fe5113a38e3575c04…|        2292|         3293604.0| 1437.0|
|ec5b2ba62e5743423…|        1428|         2556120.0| 1790.0|
|63b964e79dee32a35…|        6072|         2501664.0|  412.0|
|46bb3c0b1a65c8399…|         748|         2336752.0| 3124.0|
|05455dfa7cd02f13d…|        2184| 2160194.400000087|  989.1|
|3690e975641f01bd0…|         802|         2124498.0| 2649.0|
```

```
|349509b216bd5ec11…|         743|         1923627.0| 2589.0|
|695476b5848d64ba0…|         687|1820543.1299999943|2649.99|
|73236a0796f53d60d…|         832|         1755520.0| 2110.0|
|cc803a2c412833101…|         762|         1676400.0| 2200.0|
|1ff773612ab8934db…|        5820|1658641.7999999512| 284.99|
|fced842c7dad61e8c…|         602|         1654898.0| 2749.0|
|1ecb47d23dc8203cd…|        1164|1629588.3599999903|1399.99|
|de832e8dbb1f588a4…|        2190|1584990.5999999817| 723.74|
|803cd9b04f9cd252c…|         488|         1512312.0| 3099.0|
|d72181923840c8895…|        2721|1488114.8999999566|  546.9|
|06d478ba352a27a51…|        1146|         1461150.0| 1275.0|
|0049e8442c2a3e4a8…|        1204|         1444800.0| 1200.0|
+------------------+------------+------------------+-------+
only showing top 20 rows
```

[42]:
```
# Seller Performance Metrics (Revenue, Avg Review, order Count)
seller_performance_df = full_orders_df.groupBy('seller_id')\
.agg(
count('order_id').alias('total_orders'),
sum('price').alias('total_revenue'),
round(avg('review_score'),2).alias('avg_review_score'),
round(stddev('price'),2).alias('price_variability')
)\
.orderBy(desc('total_revenue'))
```

[43]:
```
seller_performance_df.show()
```

```
[Stage 111:=====================================================>(995 + 1) / 1000]

+------------------+------------+------------------+---------------+-------
----------+
|           seller_id|total_orders|
total_revenue|avg_review_score|price_variability|
+------------------+------------+------------------+---------------+-------
----------+
|4869f7a5dfa277a7d…|      184587| 3.613871731999997E7|           4.09|
111.65|
|53243585a1d6dc264…|       54514| 3.429159294999997E7|           4.12|
499.65|
|4a3ca9315b744ce9f…|      330661| 3.375957084000011E7|           3.77|
59.37|
|7c67e1448b00f6e96…|      233306|3.2282321789999746E7|           3.42|
50.39|
|fa1c13f2614d7b5c4…|       87686|3.0139386310000006E7|           4.38|
307.7|
|da8622b14eb17ae28…|      264433|2.9857669730000038E7|           3.98|
```

```
72.92|
|7e93a43ef30c4f03f…|       50226|2.6315706299999956E7|           4.15|
377.24|
|1025f0e2d44d7041d…|      229587| 2.293751851999998E7|           3.89|
84.3|
|46dc3b2cc0980fb8e…|       90426| 2.179177328999997E7|           4.18|
187.49|
|955fee9216a65b617…|      232364| 2.096441067000004E7|           4.04|
84.94|
|7a67c85e85bb2ce85…|      167231| 2.031279489000004E7|           4.26|
56.23|
|620c87c171fb2a6dd…|      142232| 2.011983960000001E7|           4.36|
100.45|
|7d13fca1522535862…|       88807|1.8156881909999996E7|           4.07|
151.18|
|a1043bafd471dff53…|      132672|1.7662675979999974E7|           4.25|
37.19|
|6560211a19b47992c…|      286539|1.7315932900000002E7|           3.86|
35.04|
|edb1ef5e36e0c8cd8…|       38945|1.6624835149999965E7|           4.43|
460.85|
|1f50f920176fa81da…|      297292|1.6497454440000003E7|           4.04|
7.39|
|5dceca129747e92ff…|       50420|1.4910548339999983E7|           4.17|
299.84|
|cc419e0650a3c5ba7…|      256032|1.4751464500000065E7|           4.07|
22.67|
|3d871de0142ce09b7…|      175876|1.4184525299999991E7|           4.15|
38.14|
+--------------------+------------+--------------------+---------------+-------
----------+
only showing top 20 rows
```

[44]:
```
## Product Popularity Metrics

product_metrics_df = full_orders_df.groupBy('product_id')\
.agg(
count('order_id').alias('total_sales'),
sum('price').alias('total_revenue'),
round(avg('price'),2).alias('avg_price'),
round(stddev('price'),2).alias('price_volatility'),\
collect_list('seller_id').alias('unique_sellers')
)\
.orderBy(desc('total_sales'))
```

[45]:
```
product_metrics_df.show()
```

```
+--------------------+-----------+------------------+---------+----------------
+--------------------+
|          product_id|total_sales|     total_revenue|avg_price|price_volatility|
unique_sellers|
+--------------------+-----------+------------------+---------+----------------
+--------------------+
|aca2eb7d00ea1a7b8…|      86740| 6164630.300000013|    71.07|
3.17|[955fee9216a65b61…|
|422879e10f4668299…|      81110| 4442791.510000013|    54.77|
4.46|[1f50f920176fa81d…|
|99a4788cb24856965…|      78775| 6921762.710000019|    87.87|
4.08|[4a3ca9315b744ce9…|
|389d119b48cf3043d…|      60248|3280533.1300000125|    54.45|
4.37|[1f50f920176fa81d…|
|d1c427060a0f73f6b…|      59274| 8220103.329999987|   138.68|
16.58|[a1043bafd471dff5…|
|368c6c730842d7801…|      58358| 3181698.899999993|    54.52|
4.59|[1f50f920176fa81d…|
|53759a2ecddad2bb8…|      52654|2893017.4999999995|    54.94|
4.52|[1f50f920176fa81d…|
|53b36df67ebb7c415…|      52105| 6159887.409999995|   118.22|
20.13|[7d13fca152253586…|
|154e7e31ebfa09220…|      42700| 962160.9999999977|    22.53|
1.92|[cc419e0650a3c5ba…|
|3dd2a17168ec895c7…|      40787| 6116941.300000008|   149.97|
0.85|[de722cd6dad950a9…|
|e53e557d5a159f5aa…|      39516|3329353.9499999867|    84.25|
11.32|[88460e8ebdecbfec…|
|2b4609f8948be1887…|      36179|3171618.7700000047|    87.66|
4.22|[cc419e0650a3c5ba…|
|35afc973633aaeb6b…|      31206|2735669.0000000666|    87.66|
3.32|[4a3ca9315b744ce9…|
|e0d64dcfaa3b6db5c…|      31153| 5226407.629999994|   167.77|
30.9|[4869f7a5dfa277a7…|
|42a2c92a0979a949c…|      30486|1810926.0000000098|     59.4|
0.64|[813348c996469b40…|
|7c1bd920dbdf22470…|      29018|1739338.8200000012|    59.94|
2.77|[cc419e0650a3c5ba…|
|a62e25e09e05e6faf…|      28898|         3079869.0|   106.58|
1.5|[634964b17796e643…|
|5a848e4ab52fd5445…|      28737| 3534363.630000008|   122.99|
0.0|[c826c40d7b19f62a…|
|c4baedd846ed09b85…|      28166| 2802044.650000006|    99.48|
11.9|[a1043bafd471dff5…|
```

```
|b532349fe46b38fbc…|      27176| 993089.5699999988|    36.54|
1.92|[1025f0e2d44d7041…|
+--------------------+----------+-----------------+---------+---------------
+--------------------+
only showing top 20 rows
```

[49]:
```
# Customer Retention Analysis

customer_retention_df = full_orders_df.groupBy('customer_id')\
.agg(
first('order_purchase_timestamp').alias('first_order_date'),
last('order_purchase_timestamp').alias('last_order_date'),
count('order_id').alias('total_orders'),
round(avg('price'),2).alias('aov')
)\
.orderBy(desc('total_orders'))
```

[50]:
```
customer_retention_df.show()
```

```
[Stage 132:=================================================>(993 + 2) / 1000]
+--------------------+------------------+------------------+------------+-----
-+
|         customer_id|   first_order_date|    last_order_date|total_orders|
aov|
+--------------------+------------------+------------------+------------+-----
-+
|351e40989da90e704…|2017-07-13 10:42:37|2017-07-13 10:42:37|       11427|
85.99|
|50920f8cd0681fd86…|2018-01-27 11:28:32|2018-01-27 11:28:32|       10752|
43.82|
|9b43e2a62de9bab3a…|2017-05-25 22:27:50|2017-05-25 22:27:50|        8556|
26.4|
|270c23a11d024a44c…|2017-08-08 20:26:31|2017-08-08 20:26:31|        8001|
36.59|
|5c87184371002d49e…|2018-01-05 19:15:37|2018-01-05 19:15:37|        6876|
12.49|
|d3e82ccec3cb5f956…|2017-03-18 14:28:34|2017-03-18 14:28:34|        6876|
969.0|
|d5f2b3f597c7ccafb…|2017-12-13 14:21:15|2017-12-13 14:21:15|        6706|
59.0|
|c2f18647725395af4…|2018-03-06 19:21:47|2018-03-06 19:21:47|        6612|
34.9|
|24e7dc2ff8c071263…|2017-11-24 16:16:45|2017-11-24 16:16:45|        6597|
59.2|
|7bb57d182bdc11653…|2018-04-02 17:11:30|2018-04-02 17:11:30|        6258|
86.9|
|d22f25a9fadfb1abb…|2018-05-12 12:28:58|2018-05-12 12:28:58|        6072|
```

```
14.99|
|63b964e79dee32a35…|2018-02-14 16:34:27|2018-02-14 16:34:27|        6072|
412.0|
|1ff773612ab8934db…|2018-04-19 13:54:06|2018-04-19 13:54:06|
5820|284.99|
|13aa59158da63ba0e…|2017-09-23 14:56:45|2017-09-23 14:56:45|        5206|
79.99|
|78fc46047c4a639e8…|2017-11-28 22:24:18|2017-11-28 22:24:18|
5200|109.97|
|dd3f1762eb601f41c…|2018-07-18 12:59:21|2018-07-18 12:59:21|
4992|179.99|
|a193aa8d905b8e246…|2018-02-12 18:04:28|2018-02-12 18:04:28|        4896|
9.99|
|9eb3d566e87289dcb…|2018-06-08 16:42:11|2018-06-08 16:42:11|        4872|
5.11|
|2ba91e12e5e4c9f56…|2017-11-25 13:54:39|2017-11-25 13:54:39|        4752|
99.9|
|1b2ab6eda1946a6ff…|2017-11-24 10:41:43|2017-11-24 10:41:43|        4728|
32.99|
+--------------------+-------------------+-------------------+------------+-----
-+
only showing top 20 rows
```

0.5 Extended Enrichment

[51]:
```
# Order Status Flags
full_orders_df.select('order_status').show()
```

```
+------------+
|order_status|
+------------+
|   delivered|
|   delivered|
|   delivered|
|   delivered|
|   delivered|
|   delivered|
|   delivered|
|   delivered|
|   delivered|
|   delivered|
|   delivered|
|   delivered|
|   delivered|
|   delivered|
|   delivered|
|   delivered|
|   delivered|
|   delivered|
```

```
|   delivered|
|   delivered|
|   delivered|
|   delivered|
|   delivered|
+------------+
only showing top 20 rows
```

[54]:
```
full_orders_df = full_orders_df.
 ↪withColumn('is_delivered',when(col('order_status') == 'delivered',lit(1)).
 ↪otherwise(lit(0)))\
.withColumn('is_canceled',when(col('order_status') == 'cancelled', lit(1)).
 ↪otherwise(lit(0)))
```

[55]:
```
full_orders_df.select('order_status','is_delivered','is_canceled').show(100)
```

```
+------------+------------+-----------+
|order_status|is_delivered|is_canceled|
+------------+------------+-----------+
|   delivered|           1|          0|
|   delivered|           1|          0|
|   delivered|           1|          0|
|   delivered|           1|          0|
|   delivered|           1|          0|
|   delivered|           1|          0|
|   delivered|           1|          0|
|   delivered|           1|          0|
|   delivered|           1|          0|
|   delivered|           1|          0|
|   delivered|           1|          0|
|   delivered|           1|          0|
|   delivered|           1|          0|
|   delivered|           1|          0|
|   delivered|           1|          0|
|   delivered|           1|          0|
|   delivered|           1|          0|
|   delivered|           1|          0|
|   delivered|           1|          0|
|   delivered|           1|          0|
|   delivered|           1|          0|
|   delivered|           1|          0|
|   delivered|           1|          0|
|   delivered|           1|          0|
|   delivered|           1|          0|
|   delivered|           1|          0|
|   delivered|           1|          0|
|   delivered|           1|          0|
```

```
|   delivered|            1|          0|
|   delivered|            1|          0|
|   delivered|            1|          0|
|   delivered|            1|          0|
|   delivered|            1|          0|
|   delivered|            1|          0|
|   delivered|            1|          0|
|   delivered|            1|          0|
|   delivered|            1|          0|
|   delivered|            1|          0|
|   delivered|            1|          0|
|   delivered|            1|          0|
|   delivered|            1|          0|
|   delivered|            1|          0|
|   delivered|            1|          0|
|   delivered|            1|          0|
|   delivered|            1|          0|
|   delivered|            1|          0|
|   delivered|            1|          0|
|   delivered|            1|          0|
|   delivered|            1|          0|
|   delivered|            1|          0|
|   delivered|            1|          0|
|   delivered|            1|          0|
|   delivered|            1|          0|
|   delivered|            1|          0|
|   delivered|            1|          0|
|   delivered|            1|          0|
|   delivered|            1|          0|
|   delivered|            1|          0|
|   delivered|            1|          0|
|   delivered|            1|          0|
|   delivered|            1|          0|
|   delivered|            1|          0|
|   delivered|            1|          0|
|   delivered|            1|          0|
|   delivered|            1|          0|
|   delivered|            1|          0|
|   delivered|            1|          0|
|   delivered|            1|          0|
|   delivered|            1|          0|
|   delivered|            1|          0|
|   delivered|            1|          0|
|   delivered|            1|          0|
|   delivered|            1|          0|
|   delivered|            1|          0|
|   delivered|            1|          0|
|   delivered|            1|          0|
```

```
|   delivered|           1|          0|
|   delivered|           1|          0|
|   delivered|           1|          0|
|   delivered|           1|          0|
|   delivered|           1|          0|
|   delivered|           1|          0|
|   delivered|           1|          0|
|   delivered|           1|          0|
|   delivered|           1|          0|
|   delivered|           1|          0|
|   delivered|           1|          0|
|   delivered|           1|          0|
|   delivered|           1|          0|
|   delivered|           1|          0|
|   delivered|           1|          0|
|   delivered|           1|          0|
|   delivered|           1|          0|
|   delivered|           1|          0|
|   delivered|           1|          0|
|   delivered|           1|          0|
|   delivered|           1|          0|
|   delivered|           1|          0|
|   delivered|           1|          0|
|   delivered|           1|          0|
+------------+------------+-----------+
only showing top 100 rows
```

[57]:

```
full_orders_df.where(full_orders_df['order_status'] != 'delivered').
 ↪select('order_status','is_delivered','is_canceled').show(100)
```

```
+------------+------------+-----------+
|order_status|is_delivered|is_canceled|
+------------+------------+-----------+
|    canceled|           0|          0|
|    canceled|           0|          0|
|    canceled|           0|          0|
|    canceled|           0|          0|
|    canceled|           0|          0|
|    canceled|           0|          0|
|    canceled|           0|          0|
|    canceled|           0|          0|
|    canceled|           0|          0|
|    canceled|           0|          0|
|    canceled|           0|          0|
|    canceled|           0|          0|
|    canceled|           0|          0|
|    canceled|           0|          0|
```

```
|    canceled|           0|           0|
|    canceled|           0|           0|
|    canceled|           0|           0|
|    canceled|           0|           0|
|    canceled|           0|           0|
|    canceled|           0|           0|
|    canceled|           0|           0|
|    canceled|           0|           0|
|    canceled|           0|           0|
|    canceled|           0|           0|
|    canceled|           0|           0|
|    canceled|           0|           0|
|    canceled|           0|           0|
|    canceled|           0|           0|
|    canceled|           0|           0|
|    canceled|           0|           0|
|    canceled|           0|           0|
|    canceled|           0|           0|
|    canceled|           0|           0|
|    canceled|           0|           0|
|    canceled|           0|           0|
|    canceled|           0|           0|
|    canceled|           0|           0|
|    canceled|           0|           0|
|    canceled|           0|           0|
|    canceled|           0|           0|
|    canceled|           0|           0|
|    canceled|           0|           0|
|    canceled|           0|           0|
|    canceled|           0|           0|
|    canceled|           0|           0|
|    canceled|           0|           0|
|    canceled|           0|           0|
|    canceled|           0|           0|
|    canceled|           0|           0|
|    canceled|           0|           0|
|    canceled|           0|           0|
|    canceled|           0|           0|
|    canceled|           0|           0|
|    canceled|           0|           0|
|    canceled|           0|           0|
|    canceled|           0|           0|
|    canceled|           0|           0|
|    canceled|           0|           0|
|    canceled|           0|           0|
|    canceled|           0|           0|
|    canceled|           0|           0|
|    canceled|           0|           0|
```

```
|    canceled|           0|          0|
|    canceled|           0|          0|
|    canceled|           0|          0|
|    canceled|           0|          0|
|    canceled|           0|          0|
|    canceled|           0|          0|
|    canceled|           0|          0|
|    canceled|           0|          0|
|    canceled|           0|          0|
|    canceled|           0|          0|
|    canceled|           0|          0|
|    canceled|           0|          0|
|    canceled|           0|          0|
|    canceled|           0|          0|
|    canceled|           0|          0|
|    canceled|           0|          0|
|    canceled|           0|          0|
|    canceled|           0|          0|
|    canceled|           0|          0|
|    canceled|           0|          0|
|    canceled|           0|          0|
|    canceled|           0|          0|
|    canceled|           0|          0|
|    canceled|           0|          0|
|    canceled|           0|          0|
|    canceled|           0|          0|
|    canceled|           0|          0|
|    canceled|           0|          0|
|    canceled|           0|          0|
|    canceled|           0|          0|
|    canceled|           0|          0|
|    canceled|           0|          0|
|    canceled|           0|          0|
|    canceled|           0|          0|
|    canceled|           0|          0|
|    canceled|           0|          0|
|    canceled|           0|          0|
|    canceled|           0|          0|
+------------+------------+-----------+
only showing top 100 rows
```

[58]:
```
# order revenue calculation

full_orders_df = full_orders_df.withColumn('order_revenue',col('price') +␣
 ↪col('freight_value'))
```

[59]:
```
full_orders_df.select('price','freight_value','order_revenue').show()
```

```
+-----+-------------+------------------+
|price|freight_value|     order_revenue|
+-----+-------------+------------------+
|28.99|         7.46|36.449999999999996|
|28.99|         7.46|36.449999999999996|
|28.99|         7.46|36.449999999999996|
|28.99|         7.46|36.449999999999996|
|28.99|         7.46|36.449999999999996|
|28.99|         7.46|36.449999999999996|
|28.99|         7.46|36.449999999999996|
|28.99|         7.46|36.449999999999996|
|28.99|         7.46|36.449999999999996|
|28.99|         7.46|36.449999999999996|
|28.99|         7.46|36.449999999999996|
|28.99|         7.46|36.449999999999996|
|28.99|         7.46|36.449999999999996|
|28.99|         7.46|36.449999999999996|
|28.99|         7.46|36.449999999999996|
|28.99|         7.46|36.449999999999996|
|28.99|         7.46|36.449999999999996|
|28.99|         7.46|36.449999999999996|
|28.99|         7.46|36.449999999999996|
|28.99|         7.46|36.449999999999996|
+-----+-------------+------------------+
only showing top 20 rows
```

[67]:
```
# Customer Segmentation based on spending
customer_spending_df = customer_spending_df.withColumn(
    'customer_segment',
    when(col('AOV') >= 1200 , "High-Value")
    .when((col('AOV')<1200) & (col('AOV') >= 500),'Medium-Value')
    .otherwise("Low-Level"))
```

[68]:
```
customer_spending_df.show()
```

```
[Stage 155:============================>                          (2 + 2) / 4]

+--------------------+------------+------------------+-------+----------------+
|         customer_id|total_orders|       total_spent|    AOV|customer_segment|
+--------------------+------------+------------------+-------+----------------+
|d3e82ccec3cb5f956…|        6876|         6662844.0|  969.0|    Medium-Value|
|df55c14d1476a9a34…|         743|         3565657.0| 4799.0|      High-Value|
|fe5113a38e3575c04…|        2292|         3293604.0| 1437.0|      High-Value|
|ec5b2ba62e5743423…|        1428|         2556120.0| 1790.0|      High-Value|
|63b964e79dee32a35…|        6072|         2501664.0|  412.0|       Low-Level|
|46bb3c0b1a65c8399…|         748|         2336752.0| 3124.0|      High-Value|
```

```
|05455dfa7cd02f13d…|        2184| 2160194.400000087|  989.1|    Medium-Value|
|3690e975641f01bd0…|         802|         2124498.0| 2649.0|      High-Value|
|349509b216bd5ec11…|         743|         1923627.0| 2589.0|      High-Value|
|695476b5848d64ba0…|         687|1820543.1299999943|2649.99|      High-Value|
|73236a0796f53d60d…|         832|         1755520.0| 2110.0|      High-Value|
|cc803a2c412833101…|         762|         1676400.0| 2200.0|      High-Value|
|1ff773612ab8934db…|        5820|1658641.7999999512| 284.99|       Low-Level|
|fced842c7dad61e8c…|         602|         1654898.0| 2749.0|      High-Value|
|1ecb47d23dc8203cd…|        1164|1629588.3599999903|1399.99|      High-Value|
|de832e8dbb1f588a4…|        2190|1584990.5999999817| 723.74|    Medium-Value|
|803cd9b04f9cd252c…|         488|         1512312.0| 3099.0|      High-Value|
|d72181923840c8895…|        2721|1488114.8999999566|  546.9|    Medium-Value|
|06d478ba352a27a51…|        1146|         1461150.0| 1275.0|      High-Value|
|0049e8442c2a3e4a8…|        1204|         1444800.0| 1200.0|      High-Value|
+--------------------+------------+------------------+-------+----------------+
only showing top 20 rows
```

[ ]: