

# Module 2 - Data Cleaning and Transformation

March 7, 2025

## 1 Module 2 - Data Cleaning and Transformation

```
[1]: from pyspark.sql import SparkSession
```

```
spark = SparkSession.builder \
    .appName('OlistData') \
    .getOrCreate()
```

25/02/26 11:44:49 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)
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)
category_translation_df = spark.read.csv(hdfs_path + '
    ↳product_category_name_translation.csv',header=True,inferSchema=True)
```

```
[4]: customers_df.show()
```

```
+-----+-----+-----+-----+
-----+-----+
```

customer_id	customer_unique_id	customer_zip_code_prefix	customer_city	customer_state
+-----+-----+-----+-----+				
-----+-----+-----+-----+				
06b8999e2fba1a1fb...	861eff4711a542e4b...	14409		
franca	SP			
18955e83d337fd6b2...	290c77bc529b7ac93...	9790	sao bernardo	
do c...	SP			
4e7b3e00288586ebd...	060e732b5b29e8181...	1151		
sao paulo	SP			
b2b6027bc5c5109e5...	259dac757896d24d7...	8775	mogi	
das cruzeiras	SP			
4f2d8ab171c80ec83...	345ecd01c38d18a90...	13056		
campinas	SP			
879864dab9bc30475...	4c93744516667ad3b...	89254		
jaragua do sul	SC			
fd826e7cf63160e53...	addec96d2e059c80c...	4534		
sao paulo	SP			
5e274e7a0c3809e14...	57b2a98a409812fe9...	35182		
timoteo	MG			
5adf08e34b2e99398...	1175e95fb47ddff9d...	81560		
curitiba	PR			
4b7139f34592b3a31...	9afe194fb833f79e3...	30575	belo	
horizonte	MG			
9fb35e4ed6f0a14a4...	2a7745e1ed516b289...	39400		
montes claros	MG			
5aa9e4fdd4dfd2095...	2a46fb94aef5cbeeb...	20231	rio de	
janeiro	RJ			
b2d1536598b73a9ab...	918dc87cd72cd9f6e...	18682	lencois	
paulista	SP			
eabebad39a88bb6f5...	295c05e81917928d7...	5704		
sao paulo	SP			
1f1c7bf1c9b041b29...	3151a81801c838636...	95110		
caxias do sul	RS			
206f3129c0e4d7d0b...	21f748a16f4e1688a...	13412		
piracicaba	SP			
a7c125a0a07b75146...	5c2991dbd08bbf3cf...	22750	rio de	
janeiro	RJ			
c5c61596a3b6bd0ce...	b6e99561fe6f34a55...	7124		
guarulhos	SP			
9b8ce803689b3562d...	7f3a72e8f988c6e73...	5416		
sao paulo	SP			
49d0ea0986edde72d...	3e6fd6b2f0d499456...	68485		
pacaja	PA			
+-----+-----+-----+-----+				
-----+-----+-----+-----+				

only showing top 20 rows

```
[5]: # Identify Missing Values
```

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

```
[8]: def missing_values(df,df_name):  
    print(f'Missing values in {df_name}:')  
    df.select([count(when(col(c).isNull(),1)).alias(c) for c in df.columns]).  
    ↪ show()
```

```
[9]: missing_values(customers_df,'customer')
```

Missing values in customer:

[Stage 19:=====> (1 + 1) / 2]

```
+-----+-----+-----+-----+-----+  
-----+  
|customer_id|customer_unique_id|customer_zip_code_prefix|customer_city|customer_  
state|  
+-----+-----+-----+-----+-----+  
-----+  
|          0|          0|          0|          0|  
0|  
+-----+-----+-----+-----+-----+  
-----+
```

```
[10]: missing_values(orders_df,'orders')
```

Missing values in orders:

[Stage 22:=====> (1 + 1) / 2]

```
+-----+-----+-----+-----+-----+-----+  
-----+  
|order_id|customer_id|order_status|order_purchase_timestamp|order_approved_at|or  
der_delivered_carrier_date|order_delivered_customer_date|order_estimated_deliver  
y_date|  
+-----+-----+-----+-----+-----+-----+  
-----+  
|          0|          0|          0|          0|          0|          160|  
1783|          2965|  
+-----+-----+-----+-----+-----+-----+  
-----+
```

```
[12]: missing_values(order_item_df, 'order_item')
```

Missing values in order\_item:

```
+-----+-----+-----+-----+-----+-----+-----+
----+
|order_id|order_item_id|product_id|seller_id|shipping_limit_date|price|freight_v
alue|
+-----+-----+-----+-----+-----+-----+-----+
----+
|      0|          0|          0|          0|              0|      0|
0|
+-----+-----+-----+-----+-----+-----+-----+
----+
```

```
[13]: missing_values(payments_df, 'payments')
```

Missing values in payments:

```
+-----+-----+-----+-----+-----+
|order_id|payment_sequential|payment_type|payment_installments|payment_value|
+-----+-----+-----+-----+-----+
|      0|          0|          0|              0|          0|
+-----+-----+-----+-----+-----+
```

```
[ ]:
```

## 2 Handle Missing Values

1. Drop missing Values ( for non - critical columns )
2. Fill missing values ( for numerical columns )
3. Impute Missing Values ( for continous data )

```
[15]: orders_df_cleaned = orders_df.na.
      ↪drop(subset=['order_id', 'customer_id', 'order_status'])
```

```
[16]: orders_df_cleaned.show()
```

```
+-----+-----+-----+-----+
+-----+-----+-----+-----+
+-----+
|          order_id|
customer_id|order_status|order_purchase_timestamp| order_approved_at|order_delivered_carrier_date|order_delivered_customer_date|order_estimated_delivery_date|
+-----+-----+-----+-----+
+-----+-----+-----+-----+
```

```

+-----+
|e481f51cbdc54678b...|9ef432eb625129730...|   delivered|   2017-10-02
10:56:33|2017-10-02 11:07:15|   2017-10-04 19:55:00|   2017-10-10
21:25:13|   2017-10-18 00:00:00|
|53cdb2fc8bc7dce0b...|b0830fb4747a6c6d2...|   delivered|   2018-07-24
20:41:37|2018-07-26 03:24:27|   2018-07-26 14:31:00|   2018-08-07
15:27:45|   2018-08-13 00:00:00|
|47770eb9100c2d0c4...|41ce2a54c0b03bf34...|   delivered|   2018-08-08
08:38:49|2018-08-08 08:55:23|   2018-08-08 13:50:00|   2018-08-17
18:06:29|   2018-09-04 00:00:00|
|949d5b44dbf5de918...|f88197465ea7920ad...|   delivered|   2017-11-18
19:28:06|2017-11-18 19:45:59|   2017-11-22 13:39:59|   2017-12-02
00:28:42|   2017-12-15 00:00:00|
|ad21c59c0840e6cb8...|8ab97904e6daea886...|   delivered|   2018-02-13
21:18:39|2018-02-13 22:20:29|   2018-02-14 19:46:34|   2018-02-16
18:17:02|   2018-02-26 00:00:00|
|a4591c265e18cb1dc...|503740e9ca751ccdd...|   delivered|   2017-07-09
21:57:05|2017-07-09 22:10:13|   2017-07-11 14:58:04|   2017-07-26
10:57:55|   2017-08-01 00:00:00|
|136cce7faa42fdb2c...|ed0271e0b7da060a3...|   invoiced|   2017-04-11
12:22:08|2017-04-13 13:25:17|   null|
null|   2017-05-09 00:00:00|
|6514b8ad8028c9f2c...|9bdf08b4b3b52b552...|   delivered|   2017-05-16
13:10:30|2017-05-16 13:22:11|   2017-05-22 10:07:46|   2017-05-26
12:55:51|   2017-06-07 00:00:00|
|76c6e866289321a7c...|f54a9f0e6b351c431...|   delivered|   2017-01-23
18:29:09|2017-01-25 02:50:47|   2017-01-26 14:16:31|   2017-02-02
14:08:10|   2017-03-06 00:00:00|
|e69bfb5eb88e0ed6a...|31ad1d1b63eb99624...|   delivered|   2017-07-29
11:55:02|2017-07-29 12:05:32|   2017-08-10 19:45:24|   2017-08-16
17:14:30|   2017-08-23 00:00:00|
|e6ce16cb79ec1d90b...|494dded5b201313c6...|   delivered|   2017-05-16
19:41:10|2017-05-16 19:50:18|   2017-05-18 11:40:40|   2017-05-29
11:18:31|   2017-06-07 00:00:00|
|34513ce0c4fab462a...|7711cf624183d843a...|   delivered|   2017-07-13
19:58:11|2017-07-13 20:10:08|   2017-07-14 18:43:29|   2017-07-19
14:04:48|   2017-08-08 00:00:00|
|82566a660a982b15f...|d3e3b74c766bc6214...|   delivered|   2018-06-07
10:06:19|2018-06-09 03:13:12|   2018-06-11 13:29:00|   2018-06-19
12:05:52|   2018-07-18 00:00:00|
|5ff96c15d0b717ac6...|19402a48fe860416a...|   delivered|   2018-07-25
17:44:10|2018-07-25 17:55:14|   2018-07-26 13:16:00|   2018-07-30
15:52:25|   2018-08-08 00:00:00|
|432aaf21d85167c2c...|3df704f53d3f1d481...|   delivered|   2018-03-01
14:14:28|2018-03-01 15:10:47|   2018-03-02 21:09:20|   2018-03-12
23:36:26|   2018-03-21 00:00:00|
|dcb36b511fcac050b...|3b6828a50ffe54694...|   delivered|   2018-06-07
19:03:12|2018-06-12 23:31:02|   2018-06-11 14:54:00|   2018-06-21

```

```

15:34:32|                2018-07-04 00:00:00|
|403b97836b0c04a62...|738b086814c6fcc74...|    delivered|        2018-01-02
19:00:43|2018-01-02 19:09:04|                2018-01-03 18:19:09|                2018-01-20
01:38:59|                2018-02-06 00:00:00|
|116f0b09343b49556...|3187789bec9909876...|    delivered|        2017-12-26
23:41:31|2017-12-26 23:50:22|                2017-12-28 18:33:05|                2018-01-08
22:36:36|                2018-01-29 00:00:00|
|85ce859fd6dc634de...|059f7fc5719c7da6c...|    delivered|        2017-11-21
00:03:41|2017-11-21 00:14:22|                2017-11-23 21:32:26|                2017-11-27
18:28:00|                2017-12-11 00:00:00|
|83018ec114eee8641...|7f8c8b9c2ae27bf33...|    delivered|        2017-10-26
15:54:26|2017-10-26 16:08:14|                2017-10-26 21:46:53|                2017-11-08
22:22:00|                2017-11-23 00:00:00|
+-----+-----+-----+-----+
+-----+-----+-----+-----+
+-----+
only showing top 20 rows

```

[ ]:

```

[17]: orders_df_cleaned = orders_df.fillna({'order_delivered_customer_date':
↳ '9999-12-31'})

```

```

[18]: orders_df_cleaned.show()

```

```

+-----+-----+-----+-----+
+-----+-----+-----+-----+
+-----+
|                order_id|
customer_id|order_status|order_purchase_timestamp|    order_approved_at|order_delivered_carrier_date|order_delivered_customer_date|order_estimated_delivery_date|
+-----+-----+-----+-----+-----+
+-----+-----+-----+-----+
+-----+
|e481f51cbdc54678b...|9ef432eb625129730...|    delivered|        2017-10-02
10:56:33|2017-10-02 11:07:15|                2017-10-04 19:55:00|                2017-10-10
21:25:13|                2017-10-18 00:00:00|
|53cdb2fc8bc7dce0b...|b0830fb4747a6c6d2...|    delivered|        2018-07-24
20:41:37|2018-07-26 03:24:27|                2018-07-26 14:31:00|                2018-08-07
15:27:45|                2018-08-13 00:00:00|
|47770eb9100c2d0c4...|41ce2a54c0b03bf34...|    delivered|        2018-08-08
08:38:49|2018-08-08 08:55:23|                2018-08-08 13:50:00|                2018-08-17
18:06:29|                2018-09-04 00:00:00|
|949d5b44dbf5de918...|f88197465ea7920ad...|    delivered|        2017-11-18
19:28:06|2017-11-18 19:45:59|                2017-11-22 13:39:59|                2017-12-02
00:28:42|                2017-12-15 00:00:00|
|ad21c59c0840e6cb8...|8ab97904e6daea886...|    delivered|        2018-02-13

```

21:18:39 2018-02-13 22:20:29	2018-02-14 19:46:34	2018-02-16
18:17:02	2018-02-26 00:00:00	
a4591c265e18cb1dc... 503740e9ca751ccdd...	delivered	2017-07-09
21:57:05 2017-07-09 22:10:13	2017-07-11 14:58:04	2017-07-26
10:57:55	2017-08-01 00:00:00	
136cce7faa42fdb2c... ed0271e0b7da060a3...	invoiced	2017-04-11
12:22:08 2017-04-13 13:25:17	null	9999-12-31
00:00:00	2017-05-09 00:00:00	
6514b8ad8028c9f2c... 9bdf08b4b3b52b552...	delivered	2017-05-16
13:10:30 2017-05-16 13:22:11	2017-05-22 10:07:46	2017-05-26
12:55:51	2017-06-07 00:00:00	
76c6e866289321a7c... f54a9f0e6b351c431...	delivered	2017-01-23
18:29:09 2017-01-25 02:50:47	2017-01-26 14:16:31	2017-02-02
14:08:10	2017-03-06 00:00:00	
e69bfb5eb88e0ed6a... 31ad1d1b63eb99624...	delivered	2017-07-29
11:55:02 2017-07-29 12:05:32	2017-08-10 19:45:24	2017-08-16
17:14:30	2017-08-23 00:00:00	
e6ce16cb79ec1d90b... 494dded5b201313c6...	delivered	2017-05-16
19:41:10 2017-05-16 19:50:18	2017-05-18 11:40:40	2017-05-29
11:18:31	2017-06-07 00:00:00	
34513ce0c4fab462a... 7711cf624183d843a...	delivered	2017-07-13
19:58:11 2017-07-13 20:10:08	2017-07-14 18:43:29	2017-07-19
14:04:48	2017-08-08 00:00:00	
82566a660a982b15f... d3e3b74c766bc6214...	delivered	2018-06-07
10:06:19 2018-06-09 03:13:12	2018-06-11 13:29:00	2018-06-19
12:05:52	2018-07-18 00:00:00	
5ff96c15d0b717ac6... 19402a48fe860416a...	delivered	2018-07-25
17:44:10 2018-07-25 17:55:14	2018-07-26 13:16:00	2018-07-30
15:52:25	2018-08-08 00:00:00	
432aaf21d85167c2c... 3df704f53d3f1d481...	delivered	2018-03-01
14:14:28 2018-03-01 15:10:47	2018-03-02 21:09:20	2018-03-12
23:36:26	2018-03-21 00:00:00	
dcb36b511fcac050b... 3b6828a50ffe54694...	delivered	2018-06-07
19:03:12 2018-06-12 23:31:02	2018-06-11 14:54:00	2018-06-21
15:34:32	2018-07-04 00:00:00	
403b97836b0c04a62... 738b086814c6fcc74...	delivered	2018-01-02
19:00:43 2018-01-02 19:09:04	2018-01-03 18:19:09	2018-01-20
01:38:59	2018-02-06 00:00:00	
116f0b09343b49556... 3187789bec9909876...	delivered	2017-12-26
23:41:31 2017-12-26 23:50:22	2017-12-28 18:33:05	2018-01-08
22:36:36	2018-01-29 00:00:00	
85ce859fd6dc634de... 059f7fc5719c7da6c...	delivered	2017-11-21
00:03:41 2017-11-21 00:14:22	2017-11-23 21:32:26	2017-11-27
18:28:00	2017-12-11 00:00:00	
83018ec114eee8641... 7f8c8b9c2ae27bf33...	delivered	2017-10-26
15:54:26 2017-10-26 16:08:14	2017-10-26 21:46:53	2017-11-08
22:22:00	2017-11-23 00:00:00	

+-----+-----+-----+-----+

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

### 3 Impute missing values

```
[19]: from pyspark.ml.feature import Imputer

imputer = Imputer(inputCols=['payment_value'], outputCols=['payment_value_imputed']).
         setStrategy('mean')

payments_df_cleaned = imputer.fit(payments_df).transform(payments_df)
```

```
[20]: payments_df_cleaned.show()
```

```
+-----+-----+-----+-----+-----+
+-----+
|          order_id|payment_sequential|payment_type|payment_installments|payme
nt_value|payment_value_imputed|
+-----+-----+-----+-----+-----+
+-----+
|b81ef226f3fe1789b...|          1| credit_card|          8|
99.33|          99.33|
|a9810da82917af2d9...|          1| credit_card|          1|
24.39|          24.39|
|25e8ea4e93396b6fa...|          1| credit_card|          1|
65.71|          65.71|
|ba78997921bbcdc13...|          1| credit_card|          8|
107.78|         107.78|
|42fdf880ba16b47b5...|          1| credit_card|          2|
128.45|         128.45|
|298fcd1f73eb413e...|          1| credit_card|          2|
96.12|          96.12|
|771ee386b001f0620...|          1| credit_card|          1|
81.16|          81.16|
|3d7239c394a212faa...|          1| credit_card|          3|
51.84|          51.84|
|1f78449c87a54faf9...|          1| credit_card|          6|
341.09|         341.09|
|0573b5e23cbd79800...|          1|      boleto|          1|
51.95|          51.95|
|d88e0d5fa41661ce0...|          1| credit_card|          8|
188.73|         188.73|
|2480f727e869fdeb3...|          1| credit_card|          1|
```



141.9	141.9		
616105c9352a9668c...		1  credit_card	1
75.78	75.78		
cf95215a722f3ebf2...		1  credit_card	5
102.66	102.66		
769214176682788a9...		1  credit_card	4
105.28	105.28		
12e5cfe0e4716b59a...		1  credit_card	10
157.45	157.45		
61059985a6fc0ad64...		1  credit_card	1
132.04	132.04		
79da3f5fe31ad1e45...		1  credit_card	1
98.94	98.94		
8ac09207f415d55ac...		1  credit_card	4
244.15	244.15		
b2349a3f20dfbeef6...		1  credit_card	3
136.71	136.71		

+-----+-----+-----+-----+-----+  
-----+-----+  
only showing top 20 rows

```
[29]: payments_df_with_null = payments_df.  
      ↪withColumn('payment_value',when(col('payment_value')!= 99.  
      ↪33,col('payment_value')).otherwise(lit(None)))
```

```
[30]: payments_df_with_null.show()
```

order_id	payment_sequential	payment_type	payment_installments	payment_value
b81ef226f3fe1789b...		1  credit_card	8	
null				
a9810da82917af2d9...		1  credit_card	1	
24.39				
25e8ea4e93396b6fa...		1  credit_card	1	
65.71				
ba78997921bbcdc13...		1  credit_card	8	
107.78				
42fdf880ba16b47b5...		1  credit_card	2	
128.45				
298fcdf1f73eb413e...		1  credit_card	2	
96.12				
771ee386b001f0620...		1  credit_card	1	
81.16				

3d7239c394a212faa...	1  credit_card	3
51.84		
1f78449c87a54faf9...	1  credit_card	6
341.09		
0573b5e23cbd79800...	1  boleto	1
51.95		
d88e0d5fa41661ce0...	1  credit_card	8
188.73		
2480f727e869fdeb3...	1  credit_card	1
141.9		
616105c9352a9668c...	1  credit_card	1
75.78		
cf95215a722f3ebf2...	1  credit_card	5
102.66		
769214176682788a9...	1  credit_card	4
105.28		
12e5cfe0e4716b59a...	1  credit_card	10
157.45		
61059985a6fc0ad64...	1  credit_card	1
132.04		
79da3f5fe31ad1e45...	1  credit_card	1
98.94		
8ac09207f415d55ac...	1  credit_card	4
244.15		
b2349a3f20dfbeef6...	1  credit_card	3
136.71		
+-----+-----+-----+-----+-----		
-----+		

only showing top 20 rows

```
[58]: from pyspark.ml.feature import Imputer

imputer = Imputer(
    inputCols=['payment_value'], outputCols=['payment_value_imputed']).
    setStrategy('median')

payments_df_cleaned = imputer.fit(payments_df_with_null).
    transform(payments_df_with_null)

payments_df_cleaned.show()
```

	order_id	payment_sequential	payment_type	payment_installments	payment_value	payment_value_imputed
+-----+-----+-----+-----+-----+-----+						

-----+-----			
b81ef226f3fe1789b...		1  credit_card	8
null	100.0		
a9810da82917af2d9...		1  credit_card	1
24.39	24.39		
25e8ea4e93396b6fa...		1  credit_card	1
65.71	65.71		
ba78997921bbcdc13...		1  credit_card	8
107.78	107.78		
42fdf880ba16b47b5...		1  credit_card	2
128.45	128.45		
298fcd1f73eb413e...		1  credit_card	2
96.12	96.12		
771ee386b001f0620...		1  credit_card	1
81.16	81.16		
3d7239c394a212faa...		1  credit_card	3
51.84	51.84		
1f78449c87a54faf9...		1  credit_card	6
341.09	341.09		
0573b5e23cbd79800...		1  boleto	1
51.95	51.95		
d88e0d5fa41661ce0...		1  credit_card	8
188.73	188.73		
2480f727e869fdeb3...		1  credit_card	1
141.9	141.9		
616105c9352a9668c...		1  credit_card	1
75.78	75.78		
cf95215a722f3ebf2...		1  credit_card	5
102.66	102.66		
769214176682788a9...		1  credit_card	4
105.28	105.28		
12e5cfe0e4716b59a...		1  credit_card	10
157.45	157.45		
61059985a6fc0ad64...		1  credit_card	1
132.04	132.04		
79da3f5fe31ad1e45...		1  credit_card	1
98.94	98.94		
8ac09207f415d55ac...		1  credit_card	4
244.15	244.15		
b2349a3f20dfbeef6...		1  credit_card	3
136.71	136.71		
+-----+-----+-----+-----+-----			
-----+-----			
only showing top 20 rows			

## 4 Standardizing the format

```
[36]: def print_schema(df, df_name):  
      print(f'schema of {df_name}:')  
      df.printSchema()
```

```
[37]: print_schema(orders_df, 'orders')
```

```
schema of orders:  
root  
  |-- order_id: string (nullable = true)  
  |-- customer_id: string (nullable = true)  
  |-- order_status: string (nullable = true)  
  |-- order_purchase_timestamp: timestamp (nullable = true)  
  |-- order_approved_at: timestamp (nullable = true)  
  |-- order_delivered_carrier_date: timestamp (nullable = true)  
  |-- order_delivered_customer_date: timestamp (nullable = true)  
  |-- order_estimated_delivery_date: timestamp (nullable = true)
```

```
[38]: print_schema(customers_df, 'customers')
```

```
schema of customers:  
root  
  |-- customer_id: string (nullable = true)  
  |-- customer_unique_id: string (nullable = true)  
  |-- customer_zip_code_prefix: integer (nullable = true)  
  |-- customer_city: string (nullable = true)  
  |-- customer_state: string (nullable = true)
```

```
[39]: print_schema(payments_df, 'payments')
```

```
schema of payments:  
root  
  |-- order_id: string (nullable = true)  
  |-- payment_sequential: integer (nullable = true)  
  |-- payment_type: string (nullable = true)  
  |-- payment_installments: integer (nullable = true)  
  |-- payment_value: double (nullable = true)
```

```
[40]: orders_df.show()
```

```
+-----+-----+-----+-----+  
+-----+-----+-----+-----+  
+-----+  
|          order_id|  
customer_id|order_status|order_purchase_timestamp|  order_approved_at|order_deli
```

vered_carrier_date	order_delivered_customer_date	order_estimated_delivery_date
e481f51cbdc54678b... 9ef432eb625129730...	delivered	2017-10-02
10:56:33 2017-10-02 11:07:15	2017-10-04 19:55:00	2017-10-10
21:25:13	2017-10-18 00:00:00	
53cdb2fc8bc7dce0b... b0830fb4747a6c6d2...	delivered	2018-07-24
20:41:37 2018-07-26 03:24:27	2018-07-26 14:31:00	2018-08-07
15:27:45	2018-08-13 00:00:00	
47770eb9100c2d0c4... 41ce2a54c0b03bf34...	delivered	2018-08-08
08:38:49 2018-08-08 08:55:23	2018-08-08 13:50:00	2018-08-17
18:06:29	2018-09-04 00:00:00	
949d5b44dbf5de918... f88197465ea7920ad...	delivered	2017-11-18
19:28:06 2017-11-18 19:45:59	2017-11-22 13:39:59	2017-12-02
00:28:42	2017-12-15 00:00:00	
ad21c59c0840e6cb8... 8ab97904e6daea886...	delivered	2018-02-13
21:18:39 2018-02-13 22:20:29	2018-02-14 19:46:34	2018-02-16
18:17:02	2018-02-26 00:00:00	
a4591c265e18cb1dc... 503740e9ca751ccdd...	delivered	2017-07-09
21:57:05 2017-07-09 22:10:13	2017-07-11 14:58:04	2017-07-26
10:57:55	2017-08-01 00:00:00	
136cce7faa42fdb2c... ed0271e0b7da060a3...	invoiced	2017-04-11
12:22:08 2017-04-13 13:25:17	null	
null	2017-05-09 00:00:00	
6514b8ad8028c9f2c... 9bdf08b4b3b52b552...	delivered	2017-05-16
13:10:30 2017-05-16 13:22:11	2017-05-22 10:07:46	2017-05-26
12:55:51	2017-06-07 00:00:00	
76c6e866289321a7c... f54a9f0e6b351c431...	delivered	2017-01-23
18:29:09 2017-01-25 02:50:47	2017-01-26 14:16:31	2017-02-02
14:08:10	2017-03-06 00:00:00	
e69bfb5eb88e0ed6a... 31ad1d1b63eb99624...	delivered	2017-07-29
11:55:02 2017-07-29 12:05:32	2017-08-10 19:45:24	2017-08-16
17:14:30	2017-08-23 00:00:00	
e6ce16cb79ec1d90b... 494dded5b201313c6...	delivered	2017-05-16
19:41:10 2017-05-16 19:50:18	2017-05-18 11:40:40	2017-05-29
11:18:31	2017-06-07 00:00:00	
34513ce0c4fab462a... 7711cf624183d843a...	delivered	2017-07-13
19:58:11 2017-07-13 20:10:08	2017-07-14 18:43:29	2017-07-19
14:04:48	2017-08-08 00:00:00	
82566a660a982b15f... d3e3b74c766bc6214...	delivered	2018-06-07
10:06:19 2018-06-09 03:13:12	2018-06-11 13:29:00	2018-06-19
12:05:52	2018-07-18 00:00:00	
5ff96c15d0b717ac6... 19402a48fe860416a...	delivered	2018-07-25
17:44:10 2018-07-25 17:55:14	2018-07-26 13:16:00	2018-07-30
15:52:25	2018-08-08 00:00:00	
432aaf21d85167c2c... 3df704f53d3f1d481...	delivered	2018-03-01
14:14:28 2018-03-01 15:10:47	2018-03-02 21:09:20	2018-03-12

```

23:36:26|                2018-03-21 00:00:00|
|dcb36b511fcac050b...|3b6828a50ffe54694...|    delivered|        2018-06-07
19:03:12|2018-06-12 23:31:02|                2018-06-11 14:54:00|                2018-06-21
15:34:32|                2018-07-04 00:00:00|
|403b97836b0c04a62...|738b086814c6fcc74...|    delivered|        2018-01-02
19:00:43|2018-01-02 19:09:04|                2018-01-03 18:19:09|                2018-01-20
01:38:59|                2018-02-06 00:00:00|
|116f0b09343b49556...|3187789bec9909876...|    delivered|        2017-12-26
23:41:31|2017-12-26 23:50:22|                2017-12-28 18:33:05|                2018-01-08
22:36:36|                2018-01-29 00:00:00|
|85ce859fd6dc634de...|059f7fc5719c7da6c...|    delivered|        2017-11-21
00:03:41|2017-11-21 00:14:22|                2017-11-23 21:32:26|                2017-11-27
18:28:00|                2017-12-11 00:00:00|
|83018ec114eee8641...|7f8c8b9c2ae27bf33...|    delivered|        2017-10-26
15:54:26|2017-10-26 16:08:14|                2017-10-26 21:46:53|                2017-11-08
22:22:00|                2017-11-23 00:00:00|
+-----+-----+-----+-----+
+-----+-----+-----+-----+
+-----+
only showing top 20 rows

```

```

[41]: orders_df_cleaned = orders_df_cleaned.
      ↪withColumn('order_purchase_timestamp',to_date(col('order_purchase_timestamp')))

```

```

[42]: orders_df_cleaned.show()

```

```

+-----+-----+-----+-----+
+-----+-----+-----+-----+
+-----+
|                order_id|
customer_id|order_status|order_purchase_timestamp|  order_approved_at|order_delivered_carrier_date|order_delivered_customer_date|order_estimated_delivery_date|
+-----+-----+-----+-----+
+-----+-----+-----+-----+
+-----+
|e481f51cbdc54678b...|9ef432eb625129730...|    delivered|
2017-10-02|2017-10-02 11:07:15|                2017-10-04 19:55:00|                2017-10-10
21:25:13|                2017-10-18 00:00:00|
|53cdb2fc8bc7dce0b...|b0830fb4747a6c6d2...|    delivered|
2018-07-24|2018-07-26 03:24:27|                2018-07-26 14:31:00|                2018-08-07
15:27:45|                2018-08-13 00:00:00|
|47770eb9100c2d0c4...|41ce2a54c0b03bf34...|    delivered|
2018-08-08|2018-08-08 08:55:23|                2018-08-08 13:50:00|                2018-08-17
18:06:29|                2018-09-04 00:00:00|
|949d5b44dbf5de918...|f88197465ea7920ad...|    delivered|
2017-11-18|2017-11-18 19:45:59|                2017-11-22 13:39:59|                2017-12-02
00:28:42|                2017-12-15 00:00:00|

```

ad21c59c0840e6cb8... 8ab97904e6daea886...	delivered	
2018-02-13 2018-02-13 22:20:29	2018-02-14 19:46:34	2018-02-16
18:17:02 2018-02-26 00:00:00		
a4591c265e18cb1dc... 503740e9ca751ccdd...	delivered	
2017-07-09 2017-07-09 22:10:13	2017-07-11 14:58:04	2017-07-26
10:57:55 2017-08-01 00:00:00		
136cce7faa42fdb2c... ed0271e0b7da060a3...	invoiced	
2017-04-11 2017-04-13 13:25:17	null	9999-12-31
00:00:00 2017-05-09 00:00:00		
6514b8ad8028c9f2c... 9bdf08b4b3b52b552...	delivered	
2017-05-16 2017-05-16 13:22:11	2017-05-22 10:07:46	2017-05-26
12:55:51 2017-06-07 00:00:00		
76c6e866289321a7c... f54a9f0e6b351c431...	delivered	
2017-01-23 2017-01-25 02:50:47	2017-01-26 14:16:31	2017-02-02
14:08:10 2017-03-06 00:00:00		
e69bfb5eb88e0ed6a... 31ad1d1b63eb99624...	delivered	
2017-07-29 2017-07-29 12:05:32	2017-08-10 19:45:24	2017-08-16
17:14:30 2017-08-23 00:00:00		
e6ce16cb79ec1d90b... 494dded5b201313c6...	delivered	
2017-05-16 2017-05-16 19:50:18	2017-05-18 11:40:40	2017-05-29
11:18:31 2017-06-07 00:00:00		
34513ce0c4fab462a... 7711cf624183d843a...	delivered	
2017-07-13 2017-07-13 20:10:08	2017-07-14 18:43:29	2017-07-19
14:04:48 2017-08-08 00:00:00		
82566a660a982b15f... d3e3b74c766bc6214...	delivered	
2018-06-07 2018-06-09 03:13:12	2018-06-11 13:29:00	2018-06-19
12:05:52 2018-07-18 00:00:00		
5ff96c15d0b717ac6... 19402a48fe860416a...	delivered	
2018-07-25 2018-07-25 17:55:14	2018-07-26 13:16:00	2018-07-30
15:52:25 2018-08-08 00:00:00		
432aaf21d85167c2c... 3df704f53d3f1d481...	delivered	
2018-03-01 2018-03-01 15:10:47	2018-03-02 21:09:20	2018-03-12
23:36:26 2018-03-21 00:00:00		
dcb36b511fcac050b... 3b6828a50ffe54694...	delivered	
2018-06-07 2018-06-12 23:31:02	2018-06-11 14:54:00	2018-06-21
15:34:32 2018-07-04 00:00:00		
403b97836b0c04a62... 738b086814c6fcc74...	delivered	
2018-01-02 2018-01-02 19:09:04	2018-01-03 18:19:09	2018-01-20
01:38:59 2018-02-06 00:00:00		
116f0b09343b49556... 3187789bec9909876...	delivered	
2017-12-26 2017-12-26 23:50:22	2017-12-28 18:33:05	2018-01-08
22:36:36 2018-01-29 00:00:00		
85ce859fd6dc634de... 059f7fc5719c7da6c...	delivered	
2017-11-21 2017-11-21 00:14:22	2017-11-23 21:32:26	2017-11-27
18:28:00 2017-12-11 00:00:00		
83018ec114eee8641... 7f8c8b9c2ae27bf33...	delivered	
2017-10-26 2017-10-26 16:08:14	2017-10-26 21:46:53	2017-11-08
22:22:00 2017-11-23 00:00:00		

```

+-----+-----+-----+-----+
+-----+-----+-----+-----+
+-----+
only showing top 20 rows

```

[ ]:

[59]: payments\_df\_cleaned.show()

```

+-----+-----+-----+-----+-----+
+-----+
|          order_id|payment_sequential|payment_type|payment_installments|payme
nt_value|payment_value_imputed|
+-----+-----+-----+-----+-----+
+-----+
|b81ef226f3fe1789b...|          1| credit_card|          8|
null|          100.0|
|a9810da82917af2d9...|          1| credit_card|          1|
24.39|          24.39|
|25e8ea4e93396b6fa...|          1| credit_card|          1|
65.71|          65.71|
|ba78997921bbcdc13...|          1| credit_card|          8|
107.78|          107.78|
|42fdf880ba16b47b5...|          1| credit_card|          2|
128.45|          128.45|
|298fcdf1f73eb413e...|          1| credit_card|          2|
96.12|          96.12|
|771ee386b001f0620...|          1| credit_card|          1|
81.16|          81.16|
|3d7239c394a212faa...|          1| credit_card|          3|
51.84|          51.84|
|1f78449c87a54faf9...|          1| credit_card|          6|
341.09|          341.09|
|0573b5e23cbd79800...|          1| boleto|          1|
51.95|          51.95|
|d88e0d5fa41661ce0...|          1| credit_card|          8|
188.73|          188.73|
|2480f727e869fdeb3...|          1| credit_card|          1|
141.9|          141.9|
|616105c9352a9668c...|          1| credit_card|          1|
75.78|          75.78|
|cf95215a722f3ebf2...|          1| credit_card|          5|
102.66|          102.66|
|769214176682788a9...|          1| credit_card|          4|
105.28|          105.28|
|12e5cfe0e4716b59a...|          1| credit_card|         10|
157.45|          157.45|

```



61059985a6fc0ad64...	132.04	132.04	1  credit_card	1
79da3f5fe31ad1e45...	98.94	98.94	1  credit_card	1
8ac09207f415d55ac...	244.15	244.15	1  credit_card	4
b2349a3f20dfbeef6...	136.71	136.71	1  credit_card	3

only showing top 20 rows

```
[60]: payments_df_cleaned = payments_df_cleaned.
      ↪withColumn('payment_type',when(col('payment_type')=='boleto','Bank Transfer')
      ↪
      ↪when(col('payment_type')=='credit_card','Credit Card')
      ↪
      ↪when(col('payment_type')=='debit_card','Debit Card')
      ↪
      ↪.otherwise('other'))
```

```
[61]: payments_df_cleaned.show()
```

order_id payment_sequential	payment_type payment_installments payment_value payment_value_imputed
b81ef226f3fe1789b...	1  Credit Card  8
null  100.0	
a9810da82917af2d9...	1  Credit Card  1
24.39  24.39	
25e8ea4e93396b6fa...	1  Credit Card  1
65.71  65.71	
ba78997921bbcdc13...	1  Credit Card  8
107.78  107.78	
42fdf880ba16b47b5...	1  Credit Card  2
128.45  128.45	
298fcdf1f73eb413e...	1  Credit Card  2
96.12  96.12	
771ee386b001f0620...	1  Credit Card  1
81.16  81.16	
3d7239c394a212faa...	1  Credit Card  3
51.84  51.84	
1f78449c87a54faf9...	1  Credit Card  6
341.09  341.09	
0573b5e23cbd79800...	1 Bank Transfer  1

51.95	51.95		
d88e0d5fa41661ce0...		1  Credit Card	8
188.73	188.73		
2480f727e869fdeb3...		1  Credit Card	1
141.9	141.9		
616105c9352a9668c...		1  Credit Card	1
75.78	75.78		
cf95215a722f3ebf2...		1  Credit Card	5
102.66	102.66		
769214176682788a9...		1  Credit Card	4
105.28	105.28		
12e5cfe0e4716b59a...		1  Credit Card	10
157.45	157.45		
61059985a6fc0ad64...		1  Credit Card	1
132.04	132.04		
79da3f5fe31ad1e45...		1  Credit Card	1
98.94	98.94		
8ac09207f415d55ac...		1  Credit Card	4
244.15	244.15		
b2349a3f20dfbeef6...		1  Credit Card	3
136.71	136.71		

```
+-----+-----+-----+-----+-----+
-----+-----+
```

only showing top 20 rows

```
[64]: customers_df.printSchema()
```

```
root
 |-- customer_id: string (nullable = true)
 |-- customer_unique_id: string (nullable = true)
 |-- customer_zip_code_prefix: integer (nullable = true)
 |-- customer_city: string (nullable = true)
 |-- customer_state: string (nullable = true)
```

```
[63]: customers_df.show()
```

```
+-----+-----+-----+-----+
-----+-----+
|      customer_id| customer_unique_id|customer_zip_code_prefix|
customer_city|customer_state|
+-----+-----+-----+-----+
-----+-----+
|06b8999e2fba1a1fb...|861eff4711a542e4b...|      14409|
franca|          SP|
|18955e83d337fd6b2...|290c77bc529b7ac93...|      9790|sao bernardo
do c...|          SP|
|4e7b3e00288586ebd...|060e732b5b29e8181...|      1151|
```

sao paulo	SP		
b2b6027bc5c5109e5...	259dac757896d24d7...	8775	mogi
das cruzeiras	SP		
4f2d8ab171c80ec83...	345ecd01c38d18a90...	13056	
campinas	SP		
879864dab9bc30475...	4c93744516667ad3b...	89254	
jaragua do sul	SC		
fd826e7cf63160e53...	addec96d2e059c80c...	4534	
sao paulo	SP		
5e274e7a0c3809e14...	57b2a98a409812fe9...	35182	
timoteo	MG		
5adf08e34b2e99398...	1175e95fb47ddff9d...	81560	
curitiba	PR		
4b7139f34592b3a31...	9afe194fb833f79e3...	30575	belo
horizonte	MG		
9fb35e4ed6f0a14a4...	2a7745e1ed516b289...	39400	
montes claros	MG		
5aa9e4fdd4dfd2095...	2a46fb94aef5cbeeb...	20231	rio de
janeiro	RJ		
b2d1536598b73a9ab...	918dc87cd72cd9f6e...	18682	lencois
paulista	SP		
eabebad39a88bb6f5...	295c05e81917928d7...	5704	
sao paulo	SP		
1f1c7bf1c9b041b29...	3151a81801c838636...	95110	
caxias do sul	RS		
206f3129c0e4d7d0b...	21f748a16f4e1688a...	13412	
piracicaba	SP		
a7c125a0a07b75146...	5c2991dbd08bbf3cf...	22750	rio de
janeiro	RJ		
c5c61596a3b6bd0ce...	b6e99561fe6f34a55...	7124	
guarulhos	SP		
9b8ce803689b3562d...	7f3a72e8f988c6e73...	5416	
sao paulo	SP		
49d0ea0986edde72d...	3e6fd6b2f0d499456...	68485	
pacaja	PA		

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

```
[65]: customers_df_cleaned = customers_df.
      ↪withColumn('customer_zip_code_prefix',col('customer_zip_code_prefix')).
      ↪cast('string'))
```

```
[66]: customers_df_cleaned.printSchema()
```

```
root
 |-- customer_id: string (nullable = true)
```

```

|-- customer_unique_id: string (nullable = true)
|-- customer_zip_code_prefix: string (nullable = true)
|-- customer_city: string (nullable = true)
|-- customer_state: string (nullable = true)

```

```
[ ]:
```

```
[ ]:
```

## 5 Remove Duplicate Records

```
[ ]:
```

```
[68]: customers_df_cleaned = customers_df_cleaned.dropDuplicates(['customer_id'])
```

```
[ ]:
```

```
[ ]:
```

```
[ ]:
```

```
[ ]:
```

## 6 Data Transformation

```
[70]: order_with_details = orders_df_cleaned.join(order_item_df, 'order_id', 'left')\
      .join(payments_df_cleaned, 'order_id', 'left')\
      .join(customers_df_cleaned, 'customer_id', 'left')
```

```
[ ]:
```

```
[74]: order_with_details.show()
```

```

+-----+-----+-----+-----+
+-----+-----+-----+-----+
+-----+-----+-----+-----+
+-----+-----+-----+-----+
+-----+-----+-----+-----+
+-----+-----+-----+-----+
+-----+-----+-----+-----+
+-----+-----+-----+-----+
|      customer_id|
order_id|order_status|order_purchase_timestamp|  order_approved_at|order_deliver
ed_carrier_date|order_delivered_customer_date|order_estimated_delivery_date|orde
r_item_id|      product_id|      seller_id|shipping_limit_date|
price|freight_value|payment_sequential|
payment_type|payment_installments|payment_value|payment_value_imputed|

```

customer_unique_id	customer_zip_code_prefix	customer_city	customer_state
9ef432eb625129730... e481f51cbdc54678b...	delivered		
2017-10-02 2017-10-02 11:07:15	2017-10-04 19:55:00	2017-10-10	
21:25:13	2017-10-18 00:00:00		
1 87285b34884572647... 3504c0cb71d7fa48d...	2017-10-06 11:07:15	29.99	
8.72	2  other	1	18.59
18.59 7c396fd4830fd0422...	3149	sao paulo	
SP			
9ef432eb625129730... e481f51cbdc54678b...	delivered		
2017-10-02 2017-10-02 11:07:15	2017-10-04 19:55:00	2017-10-10	
21:25:13	2017-10-18 00:00:00		
1 87285b34884572647... 3504c0cb71d7fa48d...	2017-10-06 11:07:15	29.99	
8.72	3  other	1	2.0
2.0 7c396fd4830fd0422...	3149	sao paulo	
SP			
9ef432eb625129730... e481f51cbdc54678b...	delivered		
2017-10-02 2017-10-02 11:07:15	2017-10-04 19:55:00	2017-10-10	
21:25:13	2017-10-18 00:00:00		
1 87285b34884572647... 3504c0cb71d7fa48d...	2017-10-06 11:07:15	29.99	
8.72	1  Credit Card	1	18.12
18.12 7c396fd4830fd0422...	3149	sao paulo	
SP			
b0830fb4747a6c6d2... 53cdb2fc8bc7dce0b...	delivered		
2018-07-24 2018-07-26 03:24:27	2018-07-26 14:31:00	2018-08-07	
15:27:45	2018-08-13 00:00:00		
1 595fac2a385ac33a8... 289cdb325fb7e7f89...	2018-07-30 03:24:27	118.7	
22.76	1  Bank Transfer	1	141.46
141.46 af07308b275d755c9...	47813	barreiras	
BA			
41ce2a54c0b03bf34... 47770eb9100c2d0c4...	delivered		
2018-08-08 2018-08-08 08:55:23	2018-08-08 13:50:00	2018-08-17	
18:06:29	2018-09-04 00:00:00		
1 aa4383b373c6aca5d... 4869f7a5dfa277a7d...	2018-08-13 08:55:23	159.9	
19.22	1  Credit Card	3	179.12
179.12 3a653a41f6f9fc3d2...	75265	vianopolis	
GO			
f88197465ea7920ad... 949d5b44dbf5de918...	delivered		
2017-11-18 2017-11-18 19:45:59	2017-11-22 13:39:59	2017-12-02	
00:28:42	2017-12-15 00:00:00		
1 d0b61bfb1de832b15... 66922902710d126a0...	2017-11-23 19:45:59	45.0	
27.2	1  Credit Card	1	72.2
72.2 7c142cf63193a1473...	59296	sao goncalo do am...	

RN|  
|8ab97904e6daea886...|ad21c59c0840e6cb8...| delivered|  
2018-02-13|2018-02-13 22:20:29| 2018-02-14 19:46:34| 2018-02-16  
18:17:02| 2018-02-26 00:00:00|  
1|65266b2da20d04dbe...|2c9e548be18521d1c...|2018-02-19 20:31:37| 19.9|  
8.72| 1| Credit Card| 1| 28.62|  
28.62|72632f0f9dd73dfee...| 9195| santo andre|  
SP|  
|503740e9ca751ccdd...|a4591c265e18cb1dc...| delivered|  
2017-07-09|2017-07-09 22:10:13| 2017-07-11 14:58:04| 2017-07-26  
10:57:55| 2017-08-01 00:00:00|  
1|060cb19345d90064d...|8581055ce74af1dab...|2017-07-13 22:10:13| 147.9|  
27.36| 1| Credit Card| 6| 175.26|  
175.26|80bb27c7c16e8f973...| 86320| congonhinhas|  
PR|  
|ed0271e0b7da060a3...|136cce7faa42fdb2c...| invoiced|  
2017-04-11|2017-04-13 13:25:17| null| 9999-12-31  
00:00:00| 2017-05-09 00:00:00|  
1|a1804276d9941ac07...|dc8798cbf453b7e0f...|2017-04-19 13:25:17| 49.9|  
16.05| 1| Credit Card| 1| 65.95|  
65.95|36edbb3fb164b1f16...| 98900| santa rosa|  
RS|  
|9bdf08b4b3b52b552...|6514b8ad8028c9f2c...| delivered|  
2017-05-16|2017-05-16 13:22:11| 2017-05-22 10:07:46| 2017-05-26  
12:55:51| 2017-06-07 00:00:00|  
1|4520766ec412348b8...|16090f2ca825584b5...|2017-05-22 13:22:11| 59.99|  
15.17| 1| Credit Card| 3| 75.16|  
75.16|932afa1e708222e58...| 26525| nilopolis|  
RJ|  
|f54a9f0e6b351c431...|76c6e866289321a7c...| delivered|  
2017-01-23|2017-01-25 02:50:47| 2017-01-26 14:16:31| 2017-02-02  
14:08:10| 2017-03-06 00:00:00|  
1|ac1789e492dcd698c...|63b9ae557efed31d1...|2017-01-27 18:29:09| 19.9|  
16.05| 1|Bank Transfer| 1| 35.95|  
35.95|39382392765b6dc74...| 99655| faxinalzinho|  
RS|  
|31ad1d1b63eb99624...|e69bfb5eb88e0ed6a...| delivered|  
2017-07-29|2017-07-29 12:05:32| 2017-08-10 19:45:24| 2017-08-16  
17:14:30| 2017-08-23 00:00:00|  
1|9a78fb9862b10749a...|7c67e1448b00f6e96...|2017-08-11 12:05:32|149.99|  
19.77| 1| Credit Card| 1| 8.34|  
8.34|299905e3934e9e181...| 18075| sorocaba|  
SP|  
|31ad1d1b63eb99624...|e69bfb5eb88e0ed6a...| delivered|  
2017-07-29|2017-07-29 12:05:32| 2017-08-10 19:45:24| 2017-08-16  
17:14:30| 2017-08-23 00:00:00|  
1|9a78fb9862b10749a...|7c67e1448b00f6e96...|2017-08-11 12:05:32|149.99|  
19.77| 2| other| 1| 161.42|

161.42 299905e3934e9e181...	18075	sorocaba
SP		
494dded5b201313c6... e6ce16cb79ec1d90b...	delivered	
2017-05-16 2017-05-16 19:50:18	2017-05-18 11:40:40	2017-05-29
11:18:31	2017-06-07 00:00:00	
2 08574b074924071f4... 001cca7ae9ae17fb1...	2017-05-22 19:50:18	99.0
30.53	1  Credit Card	1  259.06
259.06 f2a85dec752b8517b...	20780	rio de janeiro
RJ		
494dded5b201313c6... e6ce16cb79ec1d90b...	delivered	
2017-05-16 2017-05-16 19:50:18	2017-05-18 11:40:40	2017-05-29
11:18:31	2017-06-07 00:00:00	
1 08574b074924071f4... 001cca7ae9ae17fb1...	2017-05-22 19:50:18	99.0
30.53	1  Credit Card	1  259.06
259.06 f2a85dec752b8517b...	20780	rio de janeiro
RJ		
7711cf624183d843a... 34513ce0c4fab462a...	delivered	
2017-07-13 2017-07-13 20:10:08	2017-07-14 18:43:29	2017-07-19
14:04:48	2017-08-08 00:00:00	
1 f7e0fa615b386bc9a... 87142160b41353c4e...	2017-07-19 20:10:08	98.0
16.13	1  Credit Card	1  114.13
114.13 782987b81c92239d9...	4278	sao paulo
SP		
d3e3b74c766bc6214... 82566a660a982b15f...	delivered	
2018-06-07 2018-06-09 03:13:12	2018-06-11 13:29:00	2018-06-19
12:05:52	2018-07-18 00:00:00	
1 72a97c271b2e42997... 094ced053e257ae8c...	2018-06-18 03:13:12	31.9
18.23	1 Bank Transfer	1  50.13
50.13 e97109680b052ee85...	35400	ouro preto
MG		
19402a48fe860416a... 5ff96c15d0b717ac6...	delivered	
2018-07-25 2018-07-25 17:55:14	2018-07-26 13:16:00	2018-07-30
15:52:25	2018-08-08 00:00:00	
1 10adb53d8faa890ca... 1900267e848cee8a8...	2018-07-27 17:55:14	19.9
12.8	1  Credit Card	3  32.7
32.7 e2dfa3127fedbbca9...	4812	sao paulo
SP		
3df704f53d3f1d481... 432aaf21d85167c2c...	delivered	
2018-03-01 2018-03-01 15:10:47	2018-03-02 21:09:20	2018-03-12
23:36:26	2018-03-21 00:00:00	
1 72d3bf1d3a790f887... 0bae85eb84b9fb3bd...	2018-03-07 15:10:47	38.25
16.11	1  Credit Card	1  54.36
54.36 04cf8185c71090d28...	5271	sao paulo
SP		
3b6828a50ffe54694... dcb36b511fcac050b...	delivered	
2018-06-07 2018-06-12 23:31:02	2018-06-11 14:54:00	2018-06-21
15:34:32	2018-07-04 00:00:00	
1 009c09f439988bc06... 89a51f50b8095ea78...	2018-06-18 18:59:02	132.4

```

14.05|          1| Credit Card|          10|          146.45|
146.45|ccaafc1c3f27041052...|          74820|          goiania|
G0|
+-----+-----+-----+-----+
+-----+-----+-----+-----+
+-----+-----+-----+-----+
+-----+-----+-----+-----+
+-----+-----+-----+-----+
+-----+-----+-----+-----+
only showing top 20 rows

```

```
[ ]:
```

```
[76]: order_with_total_value = order_with_details.groupBy('order_id')\
      .agg(sum('payment_value').alias('total_order_value'))
```

```
[77]: order_with_total_value.show(5)
```

```

[Stage 93:=====>                                (1 + 1) / 2]
+-----+-----+
|          order_id|total_order_value|
+-----+-----+
|f373335aac9a659de...|          59.18|
|118045506e1c1dda0...|         1802.0|
|cc66dee6fbc18bb79...|          136.4|
|f44cb69655f8e4d13...|          164.32|
|edcc6b79e8394346b...|          162.63|
+-----+-----+
only showing top 5 rows

```

```
[ ]:
```

```
[ ]:
```

```
[ ]:
```

```
[ ]: # Delivery Time Calculation from prev notebook
```

```
[ ]:
```

```
[ ]:
```

```
[ ]:
```



[ ]:

[ ]:

## 7 Advance Transformation

[80]: `order_item_df.show()`

```
+-----+-----+-----+-----+
|          order_id|order_item_id|          product_id|
seller_id|shipping_limit_date| price|freight_value|
+-----+-----+-----+-----+
|00010242fe8c5a6d1...|
1|4244733e06e7ecb49...|48436dade18ac8b2b...|2017-09-19 09:45:35| 58.9|
13.29|
|00018f77f2f0320c5...|
1|e5f2d52b802189ee6...|dd7ddc04e1b6c2c61...|2017-05-03 11:05:13| 239.9|
19.93|
|000229ec398224ef6...|
1|c777355d18b72b67a...|5b51032eddd242adc...|2018-01-18 14:48:30| 199.0|
17.87|
|00024acbcd0a6daa...|
1|7634da152a4610f15...|9d7a1d34a50524090...|2018-08-15 10:10:18| 12.99|
12.79|
|00042b26cf59d7ce6...|
1|ac6c3623068f30de0...|df560393f3a51e745...|2017-02-13 13:57:51| 199.9|
18.14|
|00048cc3ae777c65d...|
1|ef92defde845ab845...|6426d21aca402a131...|2017-05-23 03:55:27| 21.9|
12.69|
|00054e8431b9d7675...|
1|8d4f2bb7e93e6710a...|7040e82f899a04d1b...|2017-12-14 12:10:31| 19.9|
11.85|
|000576fe39319847c...|
1|557d850972a7d6f79...|5996cddab893a4652...|2018-07-10 12:30:45| 810.0|
70.75|
|0005a1a1728c9d785...|
1|310ae3c140ff94b03...|a416b6a846a117243...|2018-03-26 18:31:29|145.95|
11.65|
|0005f50442cb953dc...|
1|4535b0e1091c278df...|ba143b05f0110f0dc...|2018-07-06 14:10:56| 53.99|
11.4|
|00061f2a7bc09da83...|
1|d63c1011f49d98b97...|cc419e0650a3c5ba7...|2018-03-29 22:28:09| 59.99|
8.88|
```

```
|00063b381e2406b52...|
1|f177554ea93259a5b...|8602a61d680a10a82...|2018-07-31 17:30:39| 45.0|
12.98|
|0006ec9db01a64e59...|
1|99a4788cb24856965...|4a3ca9315b744ce9f...|2018-07-26 17:24:20| 74.0|
23.32|
|0008288aa423d2a3f...|
1|368c6c730842d7801...|1f50f920176fa81da...|2018-02-21 02:55:52| 49.9|
13.37|
|0008288aa423d2a3f...|
2|368c6c730842d7801...|1f50f920176fa81da...|2018-02-21 02:55:52| 49.9|
13.37|
|0009792311464db53...|
1|8cab8abac59158715...|530ec6109d11eaaf8...|2018-08-17 12:15:10| 99.9|
27.65|
|0009c9a17f916a706...|
1|3f27ac8e699df3d30...|fcb5ace8bcc92f757...|2018-05-02 09:31:53| 639.0|
11.34|
|000aed2e25dbad2f9...|
1|4fa33915031a8cde0...|fe2032dab1a61af87...|2018-05-16 20:57:03| 144.0|
8.77|
|000c3e6612759851c...|
1|b50c950aba0dcead2...|218d46b86c1881d02...|2017-08-21 03:33:13| 99.0|
13.71|
|000e562887b1f2006...|
1|5ed9eaf534f6936b5...|8cbac7e12637ed9cf...|2018-02-28 12:08:37| 25.0|
16.11|
+-----+-----+-----+-----+-----+
-----+-----+-----+
only showing top 20 rows
```

```
[82]: quantiles = order_item_df.approxQuantile('price',[0.01,0.99],0.0)
low_cutoff,high_cutoff = quantiles[0],quantiles[1]
```

```
[84]: # order_item_df.select('price').summary().show()
```

```
[Stage 98:=====> (1 + 1) / 2]

+-----+-----+
|summary|           price|
+-----+-----+
|  count|          112650|
|   mean|120.65373901471354|
| stddev|183.63392805026012|
|    min|              0.85|
|   25%|              39.9|
```

	50%	74.99
	75%	134.9
	max	6735.0
+-----+-----+		

```
[83]: low_cutoff,high_cutoff
```

```
[83]: (9.99, 890.0)
```

```
[88]: order_item_df_cleaned = order_item_df.filter((col('price') >=low_cutoff) &
↳ (col('price') <=high_cutoff))
```

```
[ ]:
```

```
[ ]:
```

```
[91]: payments_df_cleaned.select('payment_installments').summary().show()
```

+-----+-----+	
summary	payment_installments
+-----+-----+	
count	103886
mean	2.853348863176944
stddev	2.6870506738564925
min	0
25%	1
50%	1
75%	4
max	24
+-----+-----+	

```
[93]: products_df.show()
```

+-----+-----+-----+-----+				
-----+-----+-----+-----+				
--+-----+				
	product_id	product_category_name	product_name_lenght	product_descript
	ion_lenght	product_photos_qty	product_weight_g	product_length_cm
	product_height_cm	product_width_cm		
+-----+-----+-----+-----+				
-----+-----+-----+-----+				
--+-----+				
1e9e8ef04dbcff454...		perfumaria	40	
287	1	225	16	10
14				

3aa071139cb16b67c...		artes	44	
276	1	1000	30	18
20				
96bd76ec8810374ed...		esporte_lazer	46	
250	1	154	18	9
15				
cef67bcfe19066a93...		bebes	27	
261	1	371	26	4
26				
9dc1a7de274444849...		utilidades_domest...	37	
402	4	625	20	17
13				
41d3672d4792049fa...		instrumentos_musi...	60	
745	1	200	38	5
11				
732bd381ad09e530f...		cool_stuff	56	
1272	4	18350	70	24
44				
2548af3e6e77a690c...		moveis_decoracao	56	
184	2	900	40	8
40				
37cc742be07708b53...		eletrodomesticos	57	
163	1	400	27	13
17				
8c92109888e8cdf9d...		brinquedos	36	
1156	1	600	17	10
12				
14aa47b7fe5c25522...		cama_mesa_banho	54	
630	1	1100	16	10
16				
03b63c5fc16691530...		bebes	49	
728	4	7150	50	19
45				
cf55509ea8edaaac1...		instrumentos_musi...	43	
1827	3	250	17	7
17				
7bb6f29c2be577161...		moveis_decoracao	51	
2083	2	600	68	11
13				
eb31436580a610f20...		construcao_ferram...	59	
1602	4	200	17	7
17				
3bb7f144022e67327...		esporte_lazer	22	
3021	1	800	16	2
11				
6a2fb4dd53d2cdb88...		perfumaria	39	
346	2	400	27	5
20				

```
|a1b71017a84f92fd8...| informatica_acess...| 59|
636| 1| 900| 40| 15|
20|
|a0736b92e52f6cead...| moveis_decoracao| 56|
296| 2| 1700| 100| 7|
15|
|f53103a77d9cf245e...| cama_mesa_banho| 52|
206| 1| 500| 16| 10|
16|
+-----+-----+-----+-----+
-----+-----+-----+-----+
--+-----+
```

only showing top 20 rows

```
[94]: products_df_cleaned = products_df.withColumn(
      'product_size_category',
      when(col('product_weight_g') < 500, 'Small')
        .when(col('product_weight_g').between(500, 2000), 'Medium')
        .otherwise('Large')
    )
```

```
[ ]:
```

```
[95]: products_df_cleaned.show()
```

```
+-----+-----+-----+-----+
-----+-----+-----+-----+
--+-----+
| product_id|product_category_name|product_name_lenght|product_descript
ion_lenght|product_photos_qty|product_weight_g|product_length_cm|product_height_
cm|product_width_cm|product_size_category|
+-----+-----+-----+-----+
-----+-----+-----+-----+
--+-----+
|1e9e8ef04dbcff454...| perfumaria| 40|
287| 1| 225| 16| 10|
14| Small|
|3aa071139cb16b67c...| artes| 44|
276| 1| 1000| 30| 18|
20| Medium|
|96bd76ec8810374ed...| esporte_lazer| 46|
250| 1| 154| 18| 9|
15| Small|
|cef67bcfe19066a93...| bebes| 27|
261| 1| 371| 26| 4|
26| Small|
|9dc1a7de274444849...| utilidades_domest...| 37|
```

402	4	625	20	17
13	Medium			
41d3672d4792049fa...	instrumentos_musi...		60	
745	1	200	38	5
11	Small			
732bd381ad09e530f...	cool_stuff		56	
1272	4	18350	70	24
44	Large			
2548af3e6e77a690c...	moveis_decoracao		56	
184	2	900	40	8
40	Medium			
37cc742be07708b53...	eletrodomesticos		57	
163	1	400	27	13
17	Small			
8c92109888e8cdf9d...	brinquedos		36	
1156	1	600	17	10
12	Medium			
14aa47b7fe5c25522...	cama_mesa_banho		54	
630	1	1100	16	10
16	Medium			
03b63c5fc16691530...	bebes		49	
728	4	7150	50	19
45	Large			
cf55509ea8edaaac1...	instrumentos_musi...		43	
1827	3	250	17	7
17	Small			
7bb6f29c2be577161...	moveis_decoracao		51	
2083	2	600	68	11
13	Medium			
eb31436580a610f20...	construcao_ferram...		59	
1602	4	200	17	7
17	Small			
3bb7f144022e67327...	esporte_lazer		22	
3021	1	800	16	2
11	Medium			
6a2fb4dd53d2cdb88...	perfumaria		39	
346	2	400	27	5
20	Small			
a1b71017a84f92fd8...	informatica_acess...		59	
636	1	900	40	15
20	Medium			
a0736b92e52f6cead...	moveis_decoracao		56	
296	2	1700	100	7
15	Medium			
f53103a77d9cf245e...	cama_mesa_banho		52	
206	1	500	16	10
16	Medium			
+-----+-----+-----+-----+				

```

-----+-----+-----+-----+-----+
--+-----+-----+
only showing top 20 rows

```

```
[ ]:
```

```
[ ]: # Total Revenue per Seller
```

```
[ ]:
```

```
[99]: !hadoop fs -mkdir /data/olist_proc
```

```
[101]: order_with_details.write.mode('overwrite').parquet('/data/olist_proc/
      ↪cleaned_data.parquet')
```

```
[104]: products_df_cleaned.write.mode('overwrite').parquet('/data/olist_proc/
      ↪product_df_cleaned.parquet')
```

```
[105]: !hadoop fs -ls /data/olist_proc
```

```

Found 2 items
drwxr-xr-x  - root hadoop          0 2025-02-26 13:14
/data/olist_proc/cleaned_data.parquet
drwxr-xr-x  - root hadoop          0 2025-02-26 13:16
/data/olist_proc/product_df_cleaned.parquet

```

```
[103]: order_with_details.printSchema()
```

```

root
 |-- customer_id: string (nullable = true)
 |-- order_id: string (nullable = true)
 |-- order_status: string (nullable = true)
 |-- order_purchase_timestamp: date (nullable = true)
 |-- order_approved_at: timestamp (nullable = true)
 |-- order_delivered_carrier_date: timestamp (nullable = true)
 |-- order_delivered_customer_date: timestamp (nullable = true)
 |-- order_estimated_delivery_date: timestamp (nullable = true)
 |-- order_item_id: integer (nullable = true)
 |-- product_id: string (nullable = true)
 |-- seller_id: string (nullable = true)
 |-- shipping_limit_date: timestamp (nullable = true)
 |-- price: double (nullable = true)
 |-- freight_value: double (nullable = true)
 |-- payment_sequential: integer (nullable = true)
 |-- payment_type: string (nullable = true)
 |-- payment_installments: integer (nullable = true)

```

```

|-- payment_value: double (nullable = true)
|-- payment_value_imputed: double (nullable = true)
|-- customer_unique_id: string (nullable = true)
|-- customer_zip_code_prefix: string (nullable = true)
|-- customer_city: string (nullable = true)
|-- customer_state: string (nullable = true)

```

[106]: `products_df_cleaned.printSchema()`

```

root
 |-- product_id: string (nullable = true)
 |-- product_category_name: string (nullable = true)
 |-- product_name_lenght: integer (nullable = true)
 |-- product_description_lenght: integer (nullable = true)
 |-- product_photos_qty: integer (nullable = true)
 |-- product_weight_g: integer (nullable = true)
 |-- product_length_cm: integer (nullable = true)
 |-- product_height_cm: integer (nullable = true)
 |-- product_width_cm: integer (nullable = true)
 |-- product_size_category: string (nullable = false)

```

[109]: `CREATE EXTERNAL TABLE cleaned_payments(`

```

    product_id STRING,
    product_category_name STRING,
    product_name_lenght STRING,
    product_description_lenght INT,
    product_photos_qty INT,
    product_weight_g INT,
    product_length_cm INT,
    product_height_cm INT,
    product_width_cm INT,
    product_size_category STRING
)
STORED AS PARQUET
LOCATION '/data/olist_proc/product_df_cleaned.parquet';

```

```

File "/tmp/ipykernel_29064/1652699355.py", line 1
    CREATE EXTERNAL TABLE cleaned_payments (
    ~
SyntaxError: invalid syntax

```

[ ]: