INDIAN DATA CLUB

CODE BASICS

databricks

14 DAYS

AI CHALLENGE

DAY 11

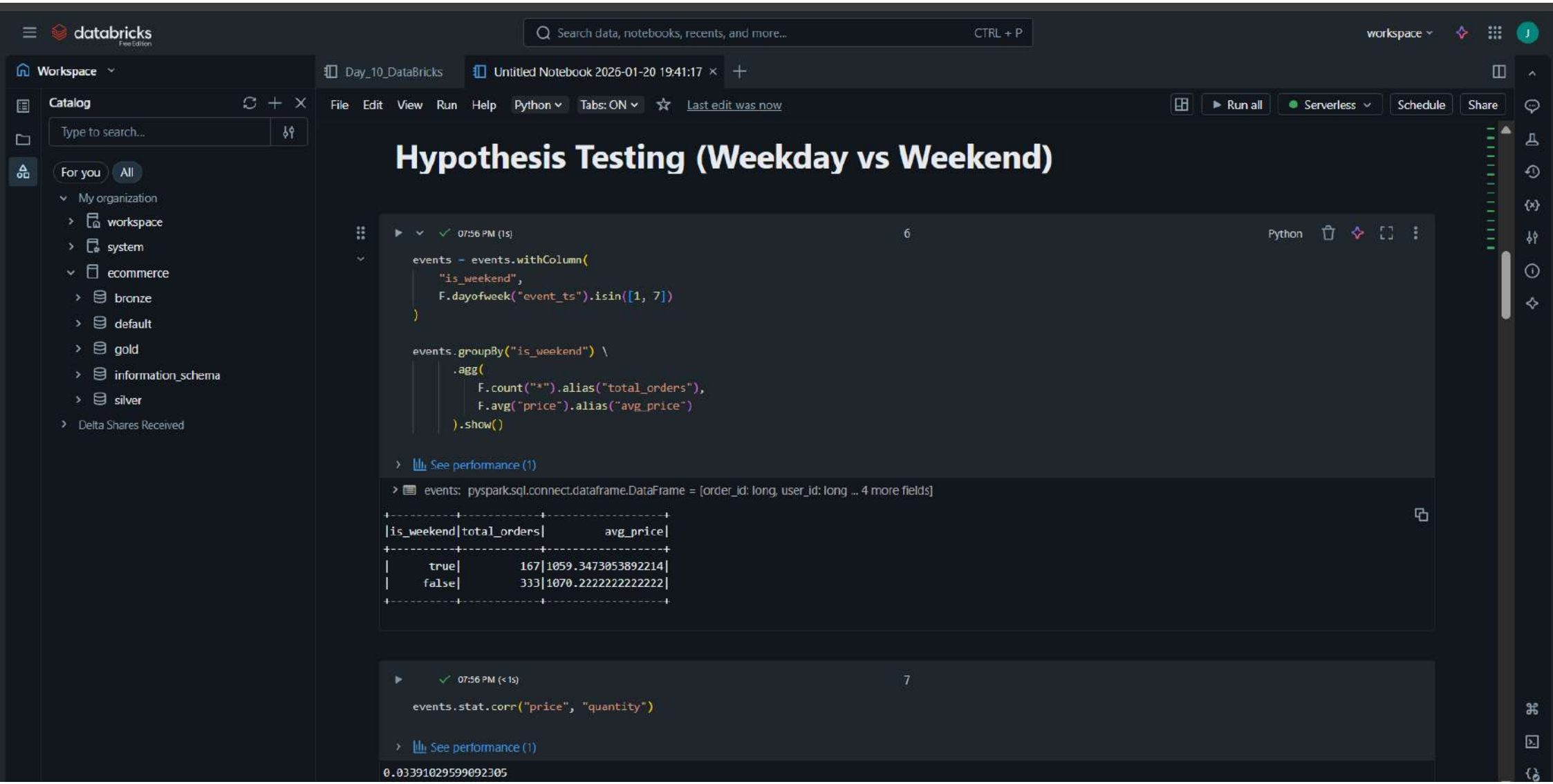
Topic:

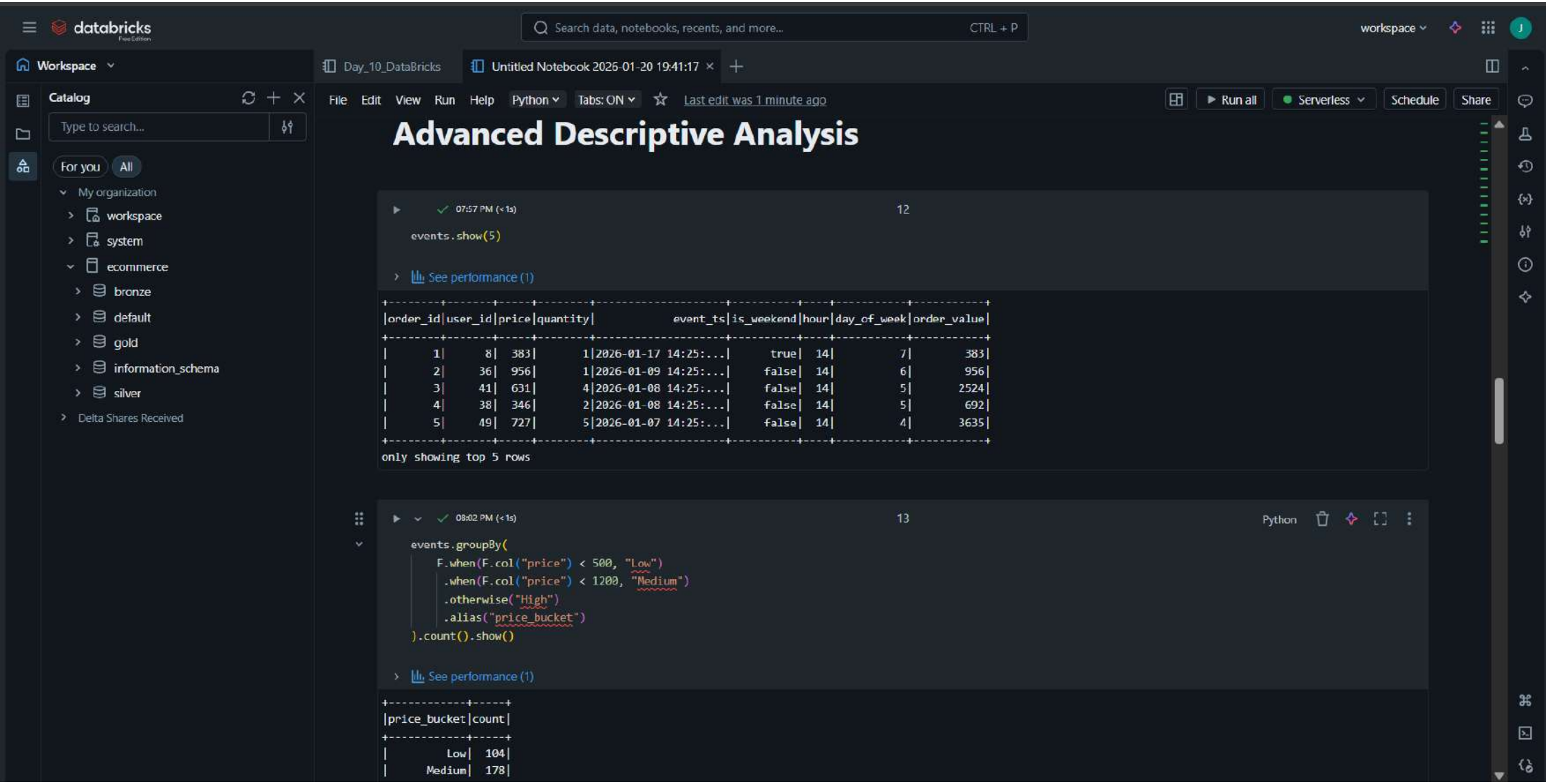
Statistical Analysis & ML Prep

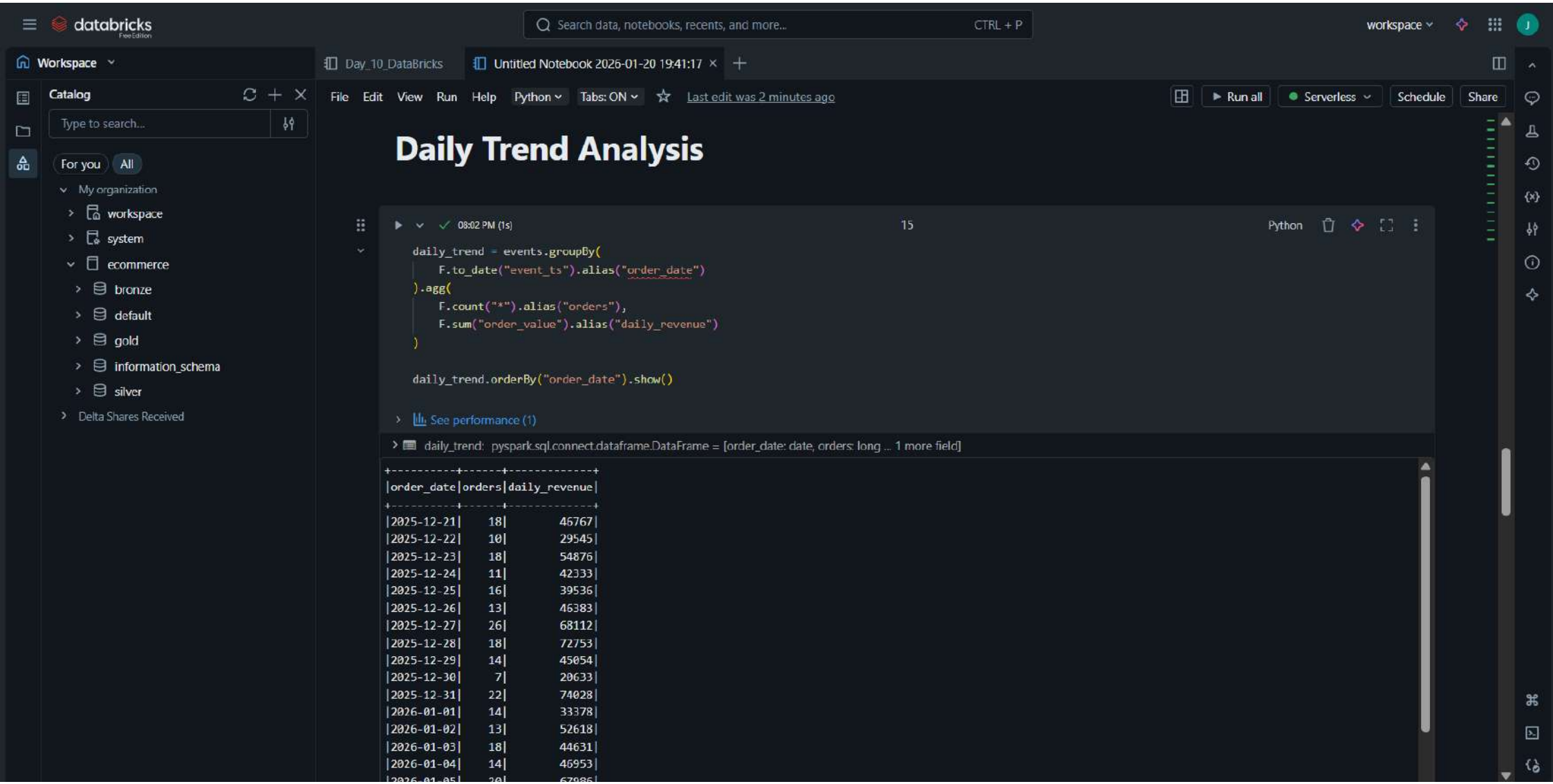
Challenge:

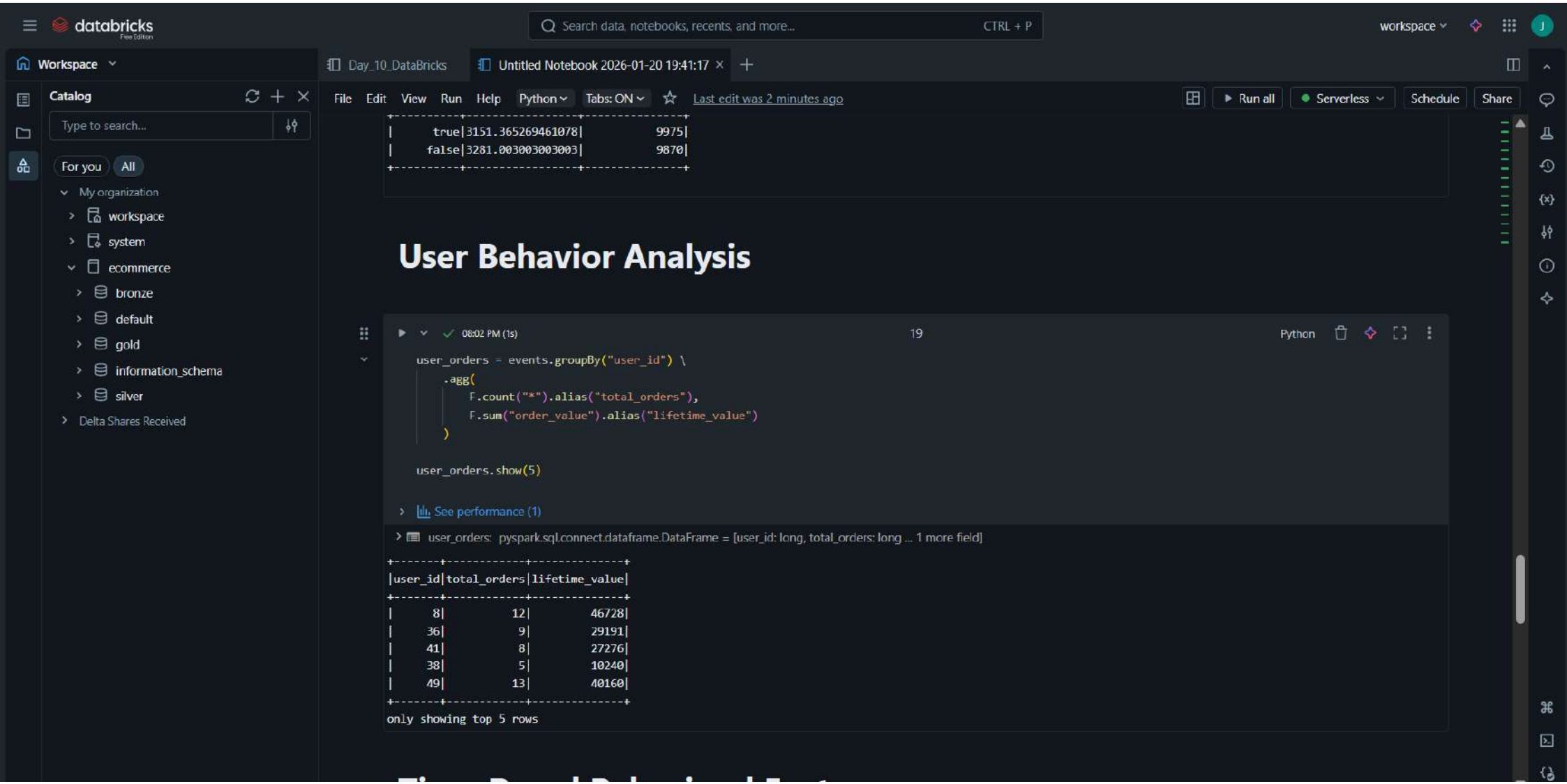
1. Calculate statistical summaries
2. Test hypotheses (weekday vs weekend)
3. Identify correlations
4. Engineer features for ML

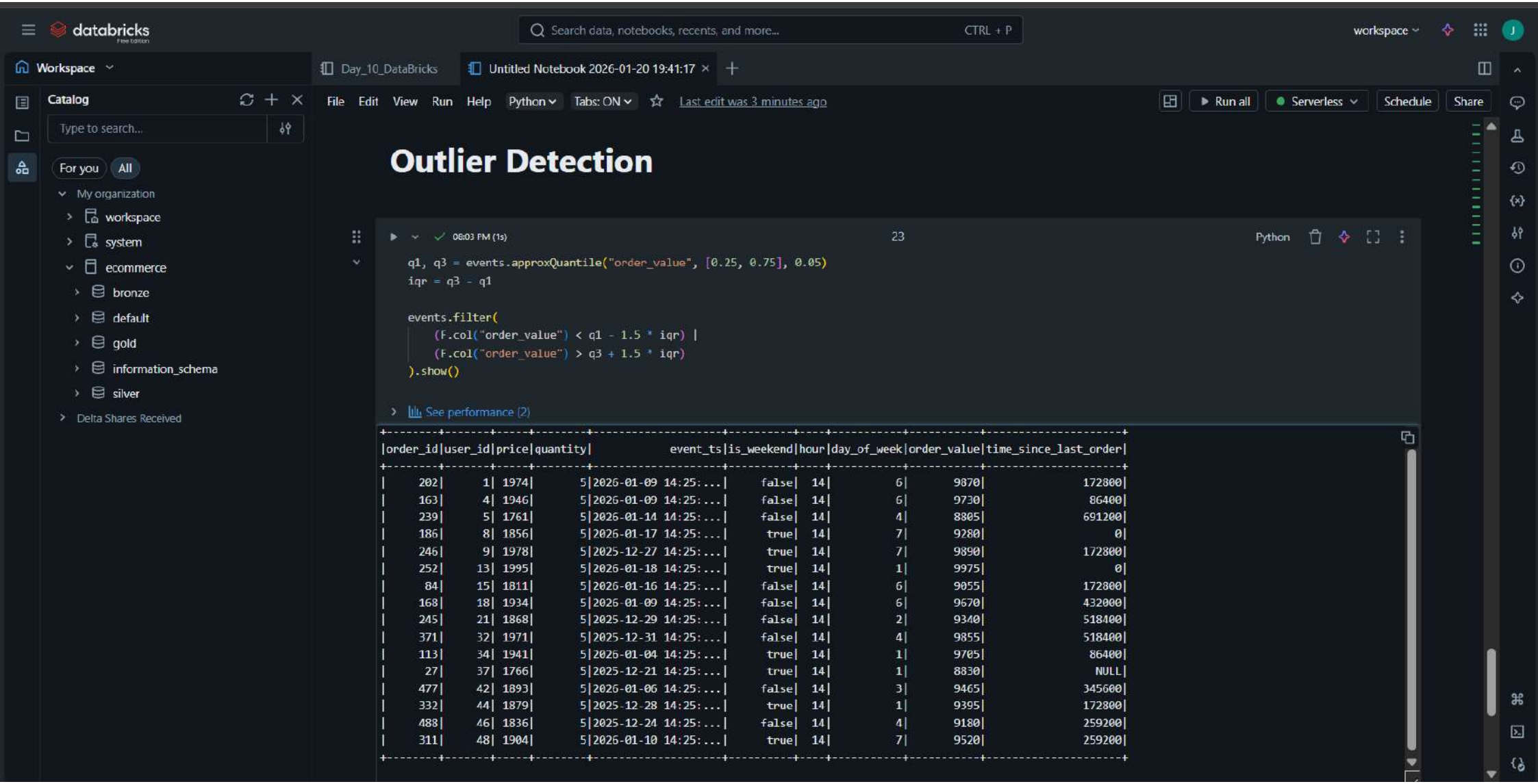
#DatabricksWithIDC


databricks
Search data, notebooks, recents, and more... CTRL + P
workspace
Workspace
Day_10_DataBricks
Untitled Notebook 2026-01-20 19:41:17
Catalog
Type to search...
For you
All
My organization
workspace
system
ecommerce
bronze
default
gold
information_schema
silver
Delta Shares Received
File Edit View Run Help Python Tabs: ON Last edit was now
Run all
Serverless
Schedule
Share
Hypothesis Testing (Weekday vs Weekend)
07:56 PM (1s)
6
Python
events = events.withColumn(
"is_weekend",
F.dayofweek("event_ts").isin([1, 7])
)
events.groupBy("is_weekend") \
.agg(
F.count("*").alias("total_orders"),
F.avg("price").alias("avg_price")
).show()
See performance (1)
events: pyspark.sql.connect.dataframe.DataFrame = [order_id: long, user_id: long ... 4 more fields]
+----------+------------+------------------+
|is_weekend|total_orders| avg_price|
+----------+------------+------------------+
| true| 167|1059.3473053892214|
| false| 333|1070.2222222222222|
+----------+------------+------------------+
07:56 PM (< 1s)
7
events.stat.corr("price", "quantity")
See performance (1)
0.03391029599092305


databricks
Search data, notebooks, recents, and more... CTRL + P
workspace
Workspace
Day_10_DataBricks
Untitled Notebook 2026-01-20 19:41:17
Catalog
Type to search...
For you
All
My organization
workspace
system
ecommerce
bronze
default
gold
information_schema
silver
Delta Shares Received
File Edit View Run Help Python Tabs: ON Last edit was 1 minute ago
Run all
Serverless
Schedule
Share
Advanced Descriptive Analysis
07:57 PM (<1s)
12
events.show(5)
See performance (1)
+--------+-------+-----+--------+-------------------+----------+----+-----------+-----------+
|order_id|user_id|price|quantity| event_ts|is_weekend|hour|day_of_week|order_value|
+--------+-------+-----+--------+-------------------+----------+----+-----------+-----------+
1	8	383	1	2026-01-17 14:25:...	true	14	7	383
2	36	956	1	2026-01-09 14:25:...	false	14	6	956
3	41	631	4	2026-01-08 14:25:...	false	14	5	2524
4	38	346	2	2026-01-08 14:25:...	false	14	5	692
5	49	727	5	2026-01-07 14:25:...	false	14	4	3635
+--------+-------+-----+--------+-------------------+----------+----+-----------+-----------+
only showing top 5 rows
08:02 PM (<1s)
13
Python
events.groupBy(
F.when(F.col("price") < 500, "Low")
.when(F.col("price") < 1200, "Medium")
.otherwise("High")
.alias("price_bucket")
).count().show()
See performance (1)
+------------+-----+
|price_bucket|count|
+------------+-----+
| Low| 104|
| Medium| 178|


databricks
Free Edition
Search data, notebooks, recents, and more...
CTRL + P
workspace
Workspace
Day_10_DataBricks
Untitled Notebook 2026-01-20 19:41:17
Catalog
Type to search...
For you
All
My organization
workspace
system
ecommerce
bronze
default
gold
information_schema
silver
Delta Shares Received
File
Edit
View
Run
Help
Python
Tabs: ON
Last edit was 2 minutes ago
Run all
Serverless
Schedule
Share
Daily Trend Analysis
08:02 PM (1s)
15
Python
daily_trend = events.groupBy(
F.to_date("event_ts").alias("order_date")
).agg(
F.count("*").alias("orders"),
F.sum("order_value").alias("daily_revenue")
)
daily_trend.orderBy("order_date").show()
See performance (1)
daily_trend: pyspark.sql.connect.dataframe.DataFrame = [order_date: date, orders: long ... 1 more field]
+----------+------+-------------+
|order_date|orders|daily_revenue|
+----------+------+-------------+
|2025-12-21| 18| 46767|
|2025-12-22| 10| 29545|
|2025-12-23| 18| 54876|
|2025-12-24| 11| 42333|
|2025-12-25| 16| 39536|
|2025-12-26| 13| 46383|
|2025-12-27| 26| 68112|
|2025-12-28| 18| 72753|
|2025-12-29| 14| 45054|
|2025-12-30| 7| 20633|
|2025-12-31| 22| 74028|
|2026-01-01| 14| 33378|
|2026-01-02| 13| 52618|
|2026-01-03| 18| 44631|
|2026-01-04| 14| 46953|


databricks
Search data, notebooks, recents, and more... CTRL + P
workspace
Workspace
Day_10_DataBricks
Untitled Notebook 2026-01-20 19:41:17
File Edit View Run Help Python Tabs: ON Last edit was 2 minutes ago
Run all Serverless Schedule Share
Catalog
Type to search...
For you All
My organization
workspace
system
ecommerce
bronze
default
gold
information_schema
silver
Delta Shares Received
+----------+-----------------+---------------+
| true|3151.365269461078| 9975|
| false|3281.003003003003| 9870|
+----------+-----------------+---------------+
User Behavior Analysis
08:02 PM (1s)
19
Python
user_orders = events.groupBy("user_id") \
.agg(
F.count("*").alias("total_orders"),
F.sum("order_value").alias("lifetime_value")
)
user_orders.show(5)
See performance (1)
user_orders: pyspark.sql.connect.dataframe.DataFrame = [user_id: long, total_orders: long ... 1 more field]
+-------+------------+--------------+
|user_id|total_orders|lifetime_value|
+-------+------------+--------------+
8	12	46728
36	9	29191
41	8	27276
38	5	10240
49	13	40160
+-------+------------+--------------+
only showing top 5 rows


databricks
Search data, notebooks, recents, and more... CTRL + P
workspace
Workspace
Day_10_DataBricks
Untitled Notebook 2026-01-20 19:41:17
Catalog
Type to search...
For you
All
My organization
workspace
system
ecommerce
bronze
default
gold
information_schema
silver
Delta Shares Received
File Edit View Run Help Python Tabs: ON Last edit was 3 minutes ago
Run all
Serverless
Schedule
Share
Outlier Detection
08:03 PM (1s)
23
Python
q1, q3 = events.approxQuantile("order_value", [0.25, 0.75], 0.05)
iqr = q3 - q1
events.filter(
(F.col("order_value") < q1 - 1.5 * iqr) |
(F.col("order_value") > q3 + 1.5 * iqr)
).show()
See performance (2)
+--------+-------+-----+--------+-------------------+----------+----+-----------+-----------+--------------------+
|order_id|user_id|price|quantity| event_ts|is_weekend|hour|day_of_week|order_value|time_since_last_order|
+--------+-------+-----+--------+-------------------+----------+----+-----------+-----------+--------------------+
202	1	1974	5	2026-01-09 14:25:...	false	14	6	9870	172800
163	4	1946	5	2026-01-09 14:25:...	false	14	6	9730	86400
239	5	1761	5	2026-01-14 14:25:...	false	14	4	8805	691200
186	8	1856	5	2026-01-17 14:25:...	true	14	7	9280	0
246	9	1978	5	2025-12-27 14:25:...	true	14	7	9890	172800
252	13	1995	5	2026-01-18 14:25:...	true	14	1	9975	0
84	15	1811	5	2026-01-16 14:25:...	false	14	6	9055	172800
168	18	1934	5	2026-01-09 14:25:...	false	14	6	9670	432000
245	21	1868	5	2025-12-29 14:25:...	false	14	2	9340	518400
371	32	1971	5	2025-12-31 14:25:...	false	14	4	9855	518400
113	34	1941	5	2026-01-04 14:25:...	true	14	1	9705	86400
27	37	1766	5	2025-12-21 14:25:...	true	14	1	8830	NULL
477	42	1893	5	2026-01-06 14:25:...	false	14	3	9465	345600
332	44	1879	5	2025-12-28 14:25:...	true	14	1	9395	172800
488	46	1836	5	2025-12-24 14:25:...	false	14	4	9180	259200
311	48	1904	5	2026-01-10 14:25:...	true	14	7	9520	259200
+--------+-------+-----+--------+-------------------+----------+----+-----------+-----------+--------------------+