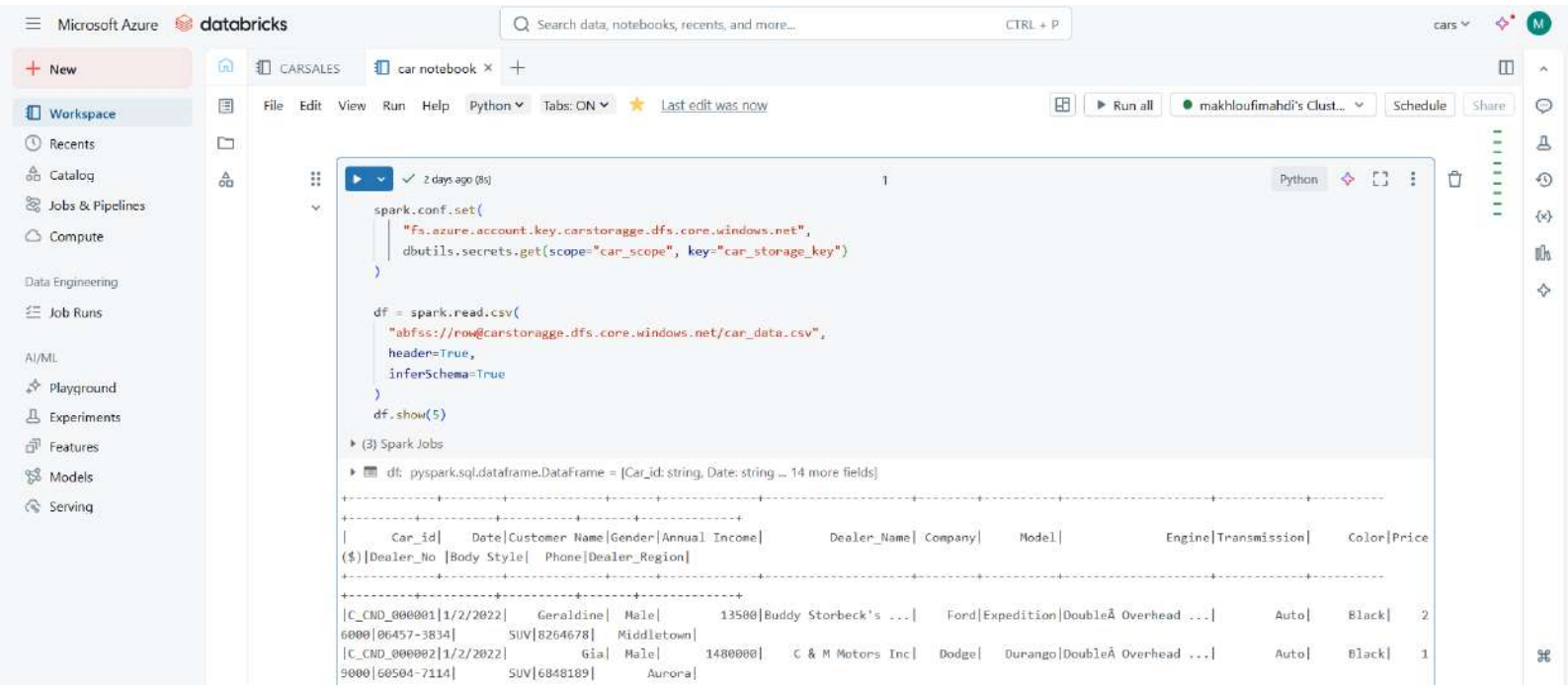
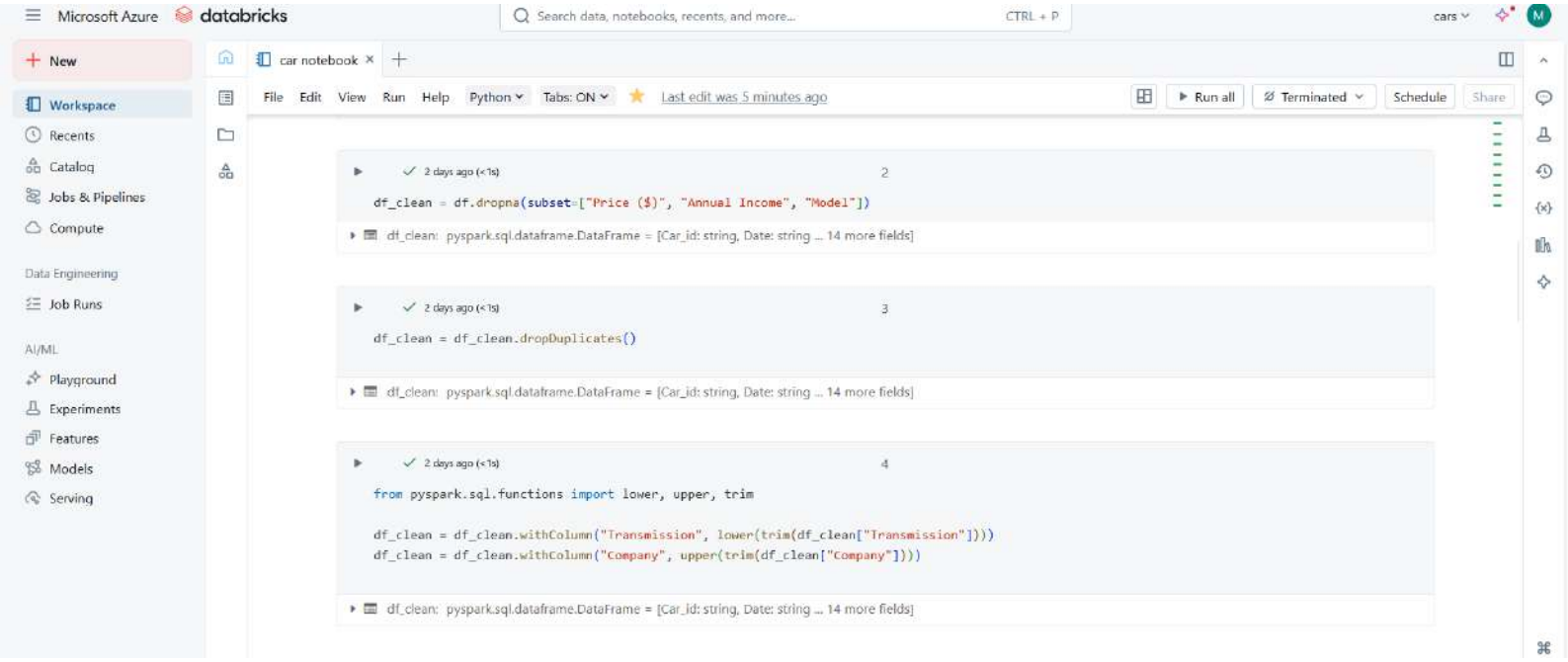
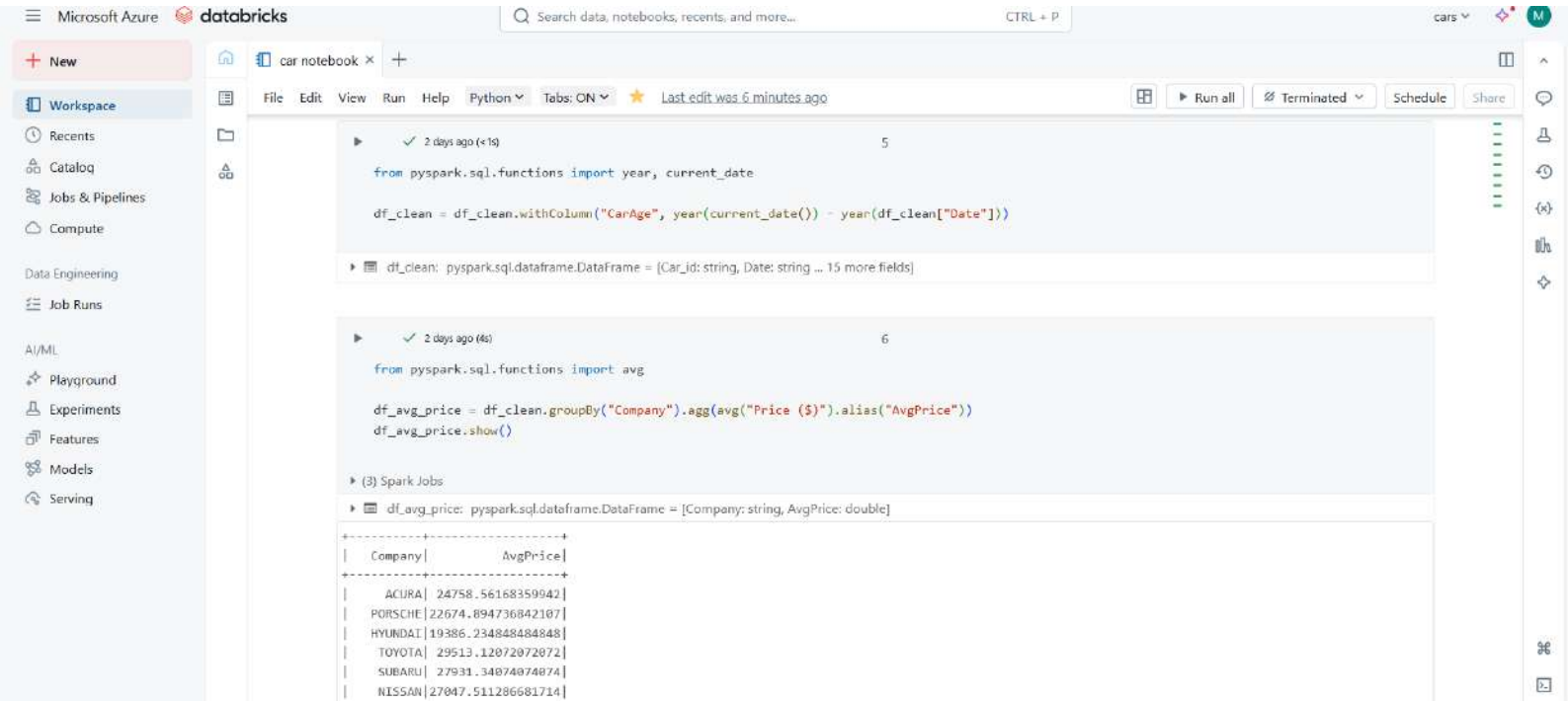
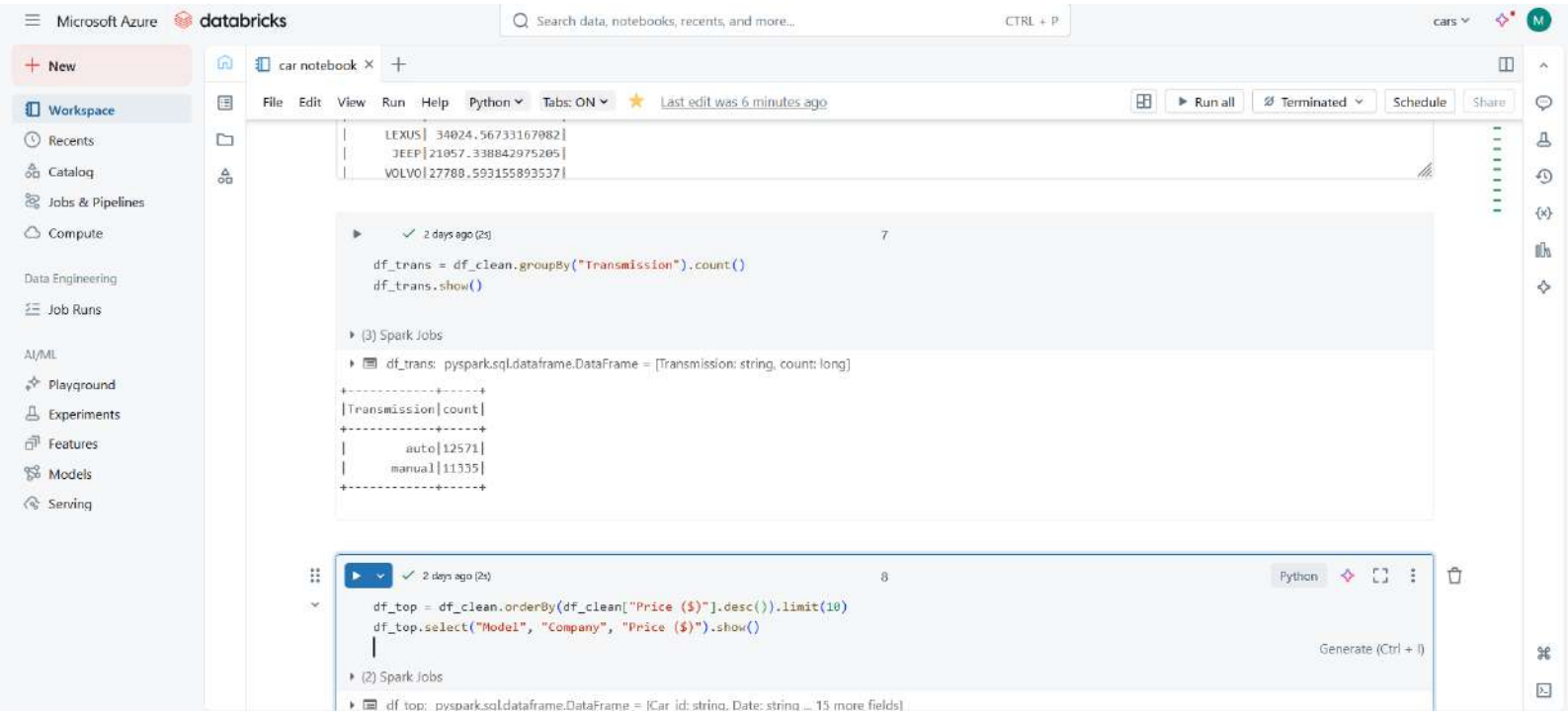
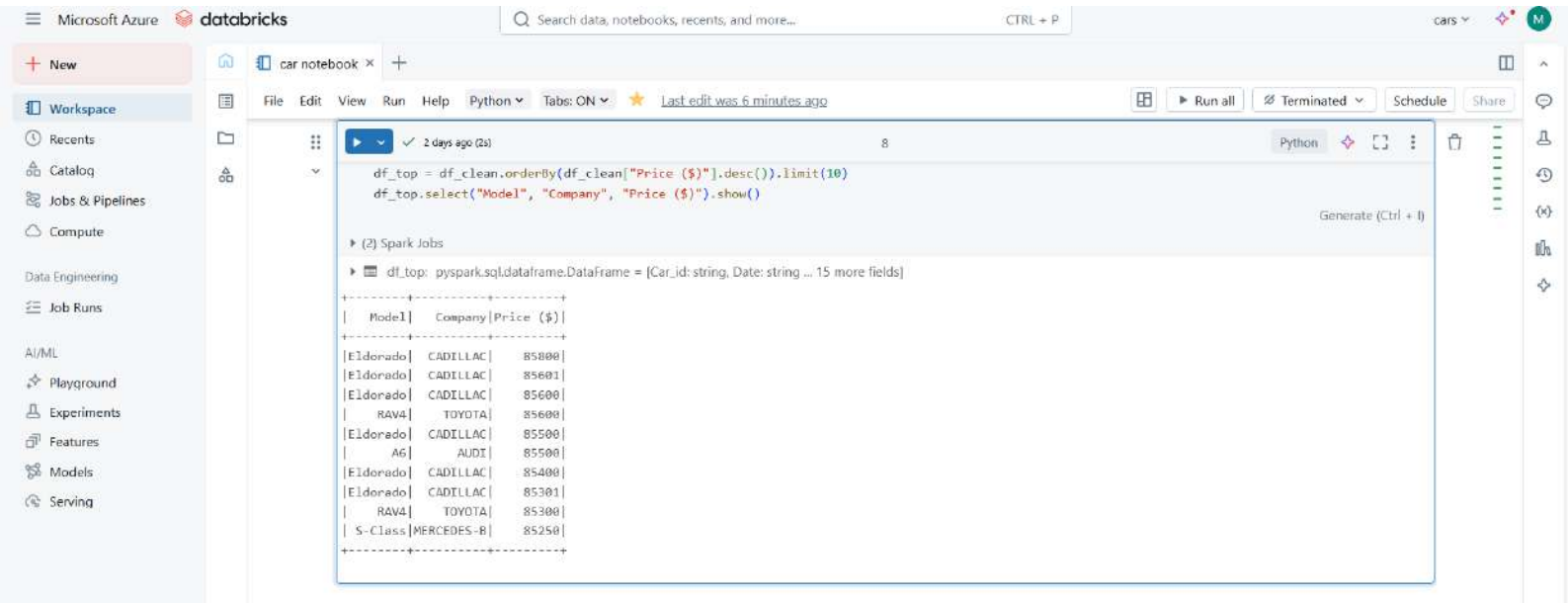
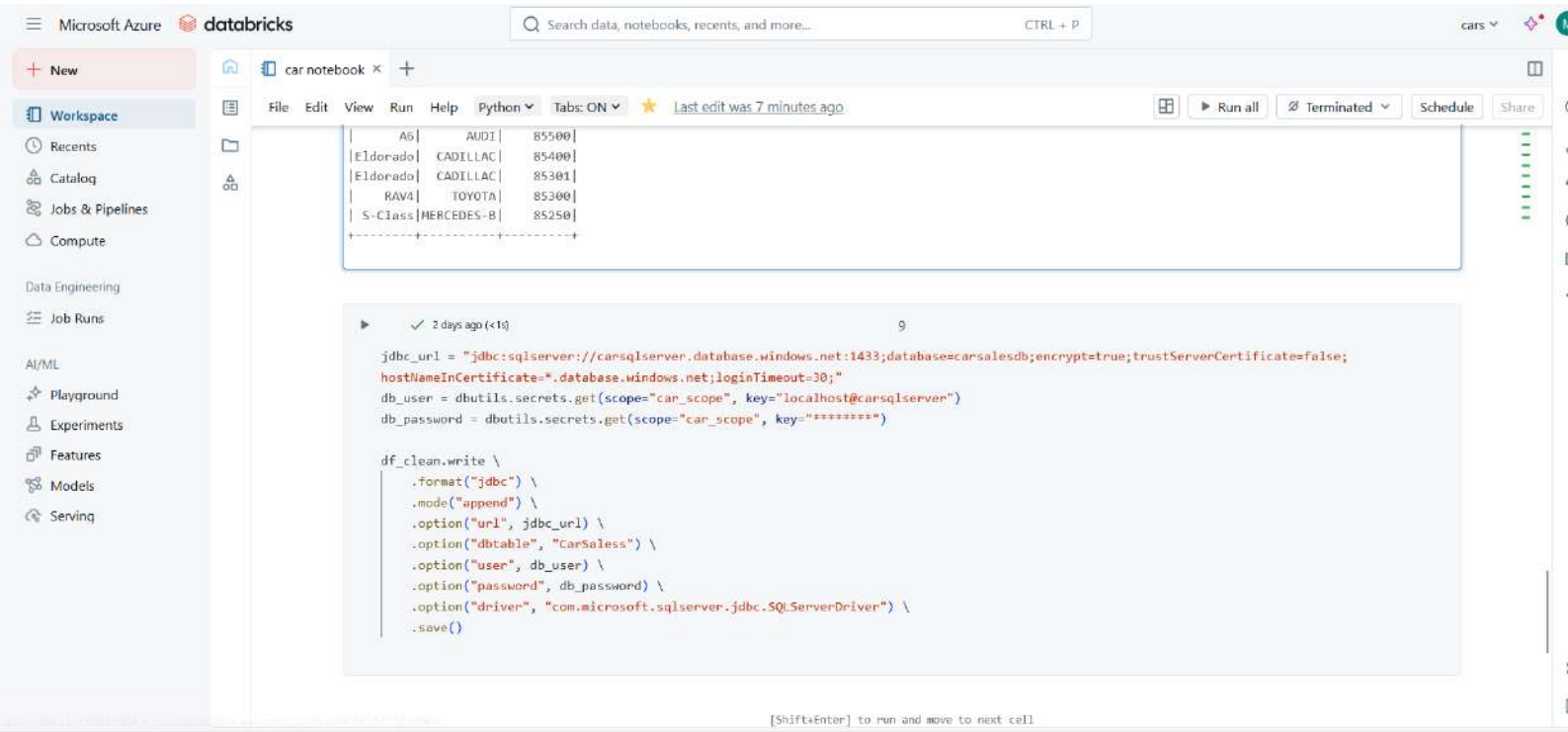

Microsoft Azure
databricks
Search data, notebooks, recents, and more...
CTRL + P
cars
New
Workspace
Recents
Catalog
Jobs & Pipelines
Compute
Data Engineering
Job Runs
AI/ML
Playground
Experiments
Features
Models
Serving
CARSALES
car notebook
File Edit View Run Help Python Tabs: ON
Last edit was now
Run all
makhloufimahdi's Clust...
Schedule
Share
2 days ago (8s)
1
Python
spark.conf.set(
"fs.azure.account.key.carstoragge.dfs.core.windows.net",
dbutils.secrets.get(scope="car_scope", key="car_storage_key")
)
df = spark.read.csv(
"abfss://row@carstoragge.dfs.core.windows.net/car_data.csv",
header=True,
inferSchema=True
)
df.show(5)
(3) Spark Jobs
df: pyspark.sql.dataframe.DataFrame = [Car_id: string, Date: string ... 14 more fields]
+------------+--------+-------------+------+-------------+--------------------+--------+----------+--------------------+-----------+----------
+---------+----------+----------+-------+-------------+
| Car_id| Date|Customer Name|Gender|Annual Income| Dealer_Name| Company| Model| Engine|Transmission| Color|Price
($)|Dealer_No |Body Style| Phone|Dealer_Region|
+------------+--------+-------------+------+-------------+--------------------+--------+----------+--------------------+-----------+----------
+---------+----------+----------+-------+-------------+
|C_CND_000001|1/2/2022| Geraldine| Male| 13500|Buddy Storbeck's ...| Ford|Expedition|DoubleÃ Overhead ...| Auto| Black| 2
6000|06457-3834| SUV|8264678| Middletown|
|C_CND_000002|1/2/2022| Gia| Male| 1480000| C & M Motors Inc| Dodge| Durango|DoubleÃ Overhead ...| Auto| Black| 1
9000|60504-7114| SUV|6848189| Aurora|


Microsoft Azure
databricks
Search data, notebooks, recents, and more...
CTRL + P
cars
New
Workspace
Recents
Catalog
Jobs & Pipelines
Compute
Data Engineering
Job Runs
AI/ML
Playground
Experiments
Features
Models
Serving
car notebook
File Edit View Run Help Python Tabs: ON
Last edit was 5 minutes ago
Run all
Terminated
Schedule
Share
2 days ago (<1s) 2
df_clean = df.dropna(subset=["Price ($)", "Annual Income", "Model"])
df_clean: pyspark.sql.dataframe.DataFrame = [Car_id: string, Date: string ... 14 more fields]
2 days ago (<1s) 3
df_clean = df_clean.dropDuplicates()
df_clean: pyspark.sql.dataframe.DataFrame = [Car_id: string, Date: string ... 14 more fields]
2 days ago (<1s) 4
from pyspark.sql.functions import lower, upper, trim
df_clean = df_clean.withColumn("Transmission", lower(trim(df_clean["Transmission"])))
df_clean = df_clean.withColumn("Company", upper(trim(df_clean["Company"])))
df_clean: pyspark.sql.dataframe.DataFrame = [Car_id: string, Date: string ... 14 more fields]


Microsoft Azure
databricks
Search data, notebooks, recents, and more...
CTRL + P
cars
New
Workspace
Recents
Catalog
Jobs & Pipelines
Compute
Data Engineering
Job Runs
AI/ML
Playground
Experiments
Features
Models
Serving
car notebook
File Edit View Run Help Python Tabs: ON
Last edit was 6 minutes ago
Run all
Terminated
Schedule
Share
2 days ago (< 1s)
5
from pyspark.sql.functions import year, current_date
df_clean = df_clean.withColumn("CarAge", year(current_date()) - year(df_clean["Date"]))
df_clean: pyspark.sql.dataframe.DataFrame = [Car_id: string, Date: string ... 15 more fields]
2 days ago (4s)
6
from pyspark.sql.functions import avg
df_avg_price = df_clean.groupBy("Company").agg(avg("Price ($)").alias("AvgPrice"))
df_avg_price.show()
(3) Spark Jobs
df_avg_price: pyspark.sql.dataframe.DataFrame = [Company: string, AvgPrice: double]
+----------+------------------+
| Company| AvgPrice|
+----------+------------------+
ACURA	24758.56168359942
PORSCHE	22674.894736842107
HYUNDAI	19386.234848484848
TOYOTA	29513.12072072072
SUBARU	27931.34074074074
NISSAN	27047.511286681714


Microsoft Azure
databricks
Search data, notebooks, recents, and more...
CTRL + P
cars
New
Workspace
Recents
Catalog
Jobs & Pipelines
Compute
Data Engineering
Job Runs
AI/ML
Playground
Experiments
Features
Models
Serving
car notebook
File Edit View Run Help Python Tabs: ON
Last edit was 6 minutes ago
Run all
Terminated
Schedule
Share
LEXUS	34024.56733167082
JEEP	21057.338842975205
VOLVO	27788.593155893537
2 days ago (2s)
7
df_trans = df_clean.groupBy("Transmission").count()
df_trans.show()
(3) Spark Jobs
df_trans: pyspark.sql.dataframe.DataFrame = [Transmission: string, count: long]
+------------+-----+
|Transmission|count|
+------------+-----+
| auto|12571|
| manual|11335|
+------------+-----+
2 days ago (2s)
8
Python
df_top = df_clean.orderBy(df_clean["Price ($)"].desc()).limit(10)
df_top.select("Model", "Company", "Price ($)").show()
Generate (Ctrl + I)
(2) Spark Jobs
df_top: pyspark.sql.dataframe.DataFrame = [Car_id: string, Date: string ... 15 more fields]


Microsoft Azure
databricks
Search data, notebooks, recents, and more...
CTRL + P
cars
New
Workspace
Recents
Catalog
Jobs & Pipelines
Compute
Data Engineering
Job Runs
AI/ML
Playground
Experiments
Features
Models
Serving
car notebook
File Edit View Run Help Python Tabs: ON
Last edit was 6 minutes ago
Run all
Terminated
Schedule
Share
2 days ago (2s)
8
Python
df_top = df_clean.orderBy(df_clean["Price ($)"].desc()).limit(10)
df_top.select("Model", "Company", "Price ($)").show()
Generate (Ctrl + I)
(2) Spark Jobs
df_top: pyspark.sql.dataframe.DataFrame = [Car_id: string, Date: string ... 15 more fields]
+--------+----------+---------+
| Model| Company|Price ($)|
+--------+----------+---------+
|Eldorado| CADILLAC| 85800|
|Eldorado| CADILLAC| 85601|
|Eldorado| CADILLAC| 85600|
| RAV4| TOYOTA| 85600|
|Eldorado| CADILLAC| 85500|
| A6| AUDI| 85500|
|Eldorado| CADILLAC| 85400|
|Eldorado| CADILLAC| 85301|
| RAV4| TOYOTA| 85300|
| S-Class|MERCEDES-B| 85250|
+--------+----------+---------+


Microsoft Azure
databricks
Search data, notebooks, recents, and more...
CTRL + P
cars
New
Workspace
Recents
Catalog
Jobs & Pipelines
Compute
Data Engineering
Job Runs
AI/ML
Playground
Experiments
Features
Models
Serving
car notebook
File Edit View Run Help Python Tabs: ON
Last edit was 7 minutes ago
Run all
Terminated
Schedule
Share
| A6| AUDI| 85500|
|Eldorado| CADILLAC| 85400|
|Eldorado| CADILLAC| 85301|
| RAV4| TOYOTA| 85300|
| S-Class|MERCEDES-B| 85250|
2 days ago (<1s)
9
jdbc_url = "jdbc:sqlserver://carsqlserver.database.windows.net:1433;database=carsalesdb;encrypt=true;trustServerCertificate=false;
hostNameInCertificate=*.database.windows.net;loginTimeout=30;"
db_user = dbutils.secrets.get(scope="car_scope", key="localhost@carsqlserver")
db_password = dbutils.secrets.get(scope="car_scope", key="********")
df_clean.write \
.format("jdbc") \
.mode("append") \
.option("url", jdbc_url) \
.option("dbtable", "CarSaless") \
.option("user", db_user) \
.option("password", db_password) \
.option("driver", "com.microsoft.sqlserver.jdbc.SQLServerDriver") \
.save()
[Shift+Enter] to run and move to next cell