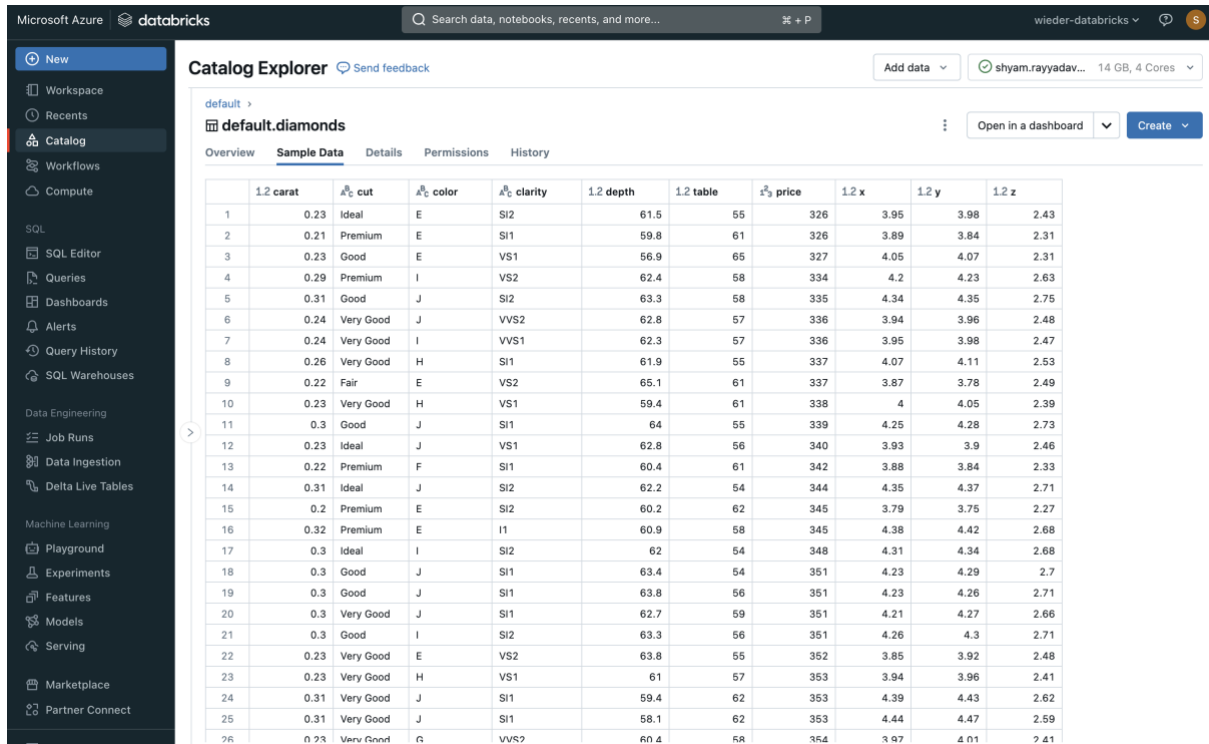


Data Science Infrastructures – Exercise 07

Assignment I – Getting data into Databricks

Answer: After uploading the .csv file into the Databricks and creating the table, this is the result that I have got into Microsoft Azure.

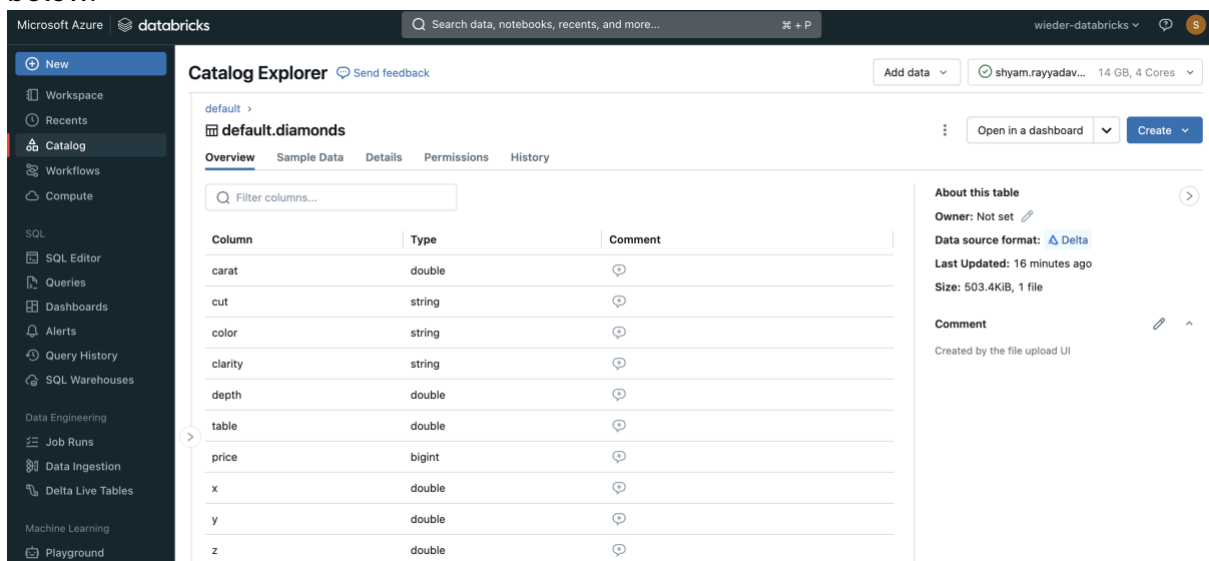


The screenshot shows the Databricks Catalog Explorer interface. The table 'default.diamonds' is displayed with 26 rows of data. The columns are: 1.2 carat, 1.2 cut, 1.2 color, 1.2 clarity, 1.2 depth, 1.2 table, 1.2 price, 1.2 x, 1.2 y, and 1.2 z. The data includes various diamond specifications such as carat weight, cut quality, color grade, clarity grade, depth, table, price, and dimensions (x, y, z).

	1.2 carat	1.2 cut	1.2 color	1.2 clarity	1.2 depth	1.2 table	1.2 price	1.2 x	1.2 y	1.2 z
1	0.23	Ideal	E	SI2	61.5	55	326	3.95	3.98	2.43
2	0.21	Premium	E	SI1	59.8	61	326	3.89	3.84	2.31
3	0.23	Good	E	VS1	56.9	65	327	4.05	4.07	2.31
4	0.29	Premium	I	VS2	62.4	58	334	4.2	4.23	2.63
5	0.31	Good	J	SI2	63.3	58	335	4.34	4.35	2.75
6	0.24	Very Good	J	VVS2	62.8	57	336	3.94	3.96	2.48
7	0.24	Very Good	I	VVS1	62.3	57	336	3.95	3.98	2.47
8	0.26	Very Good	H	SI1	61.9	55	337	4.07	4.11	2.53
9	0.22	Fair	E	VS2	65.1	61	337	3.87	3.78	2.49
10	0.23	Very Good	H	VS1	59.4	61	338	4	4.05	2.39
11	0.3	Good	J	SI1	64	55	339	4.25	4.28	2.73
12	0.23	Ideal	J	VS1	62.8	56	340	3.93	3.9	2.46
13	0.22	Premium	F	SI1	60.4	61	342	3.88	3.84	2.33
14	0.31	Ideal	J	SI2	62.2	54	344	4.35	4.37	2.71
15	0.2	Premium	E	SI2	60.2	62	345	3.79	3.75	2.27
16	0.32	Premium	E	I1	60.9	58	345	4.38	4.42	2.68
17	0.3	Ideal	I	SI2	62	54	348	4.31	4.34	2.68
18	0.3	Good	J	SI1	63.4	54	351	4.23	4.29	2.7
19	0.3	Good	J	SI1	63.8	56	351	4.23	4.26	2.71
20	0.3	Very Good	J	SI1	62.7	59	351	4.21	4.27	2.66
21	0.3	Good	I	SI2	63.3	56	351	4.26	4.3	2.71
22	0.23	Very Good	E	VS2	63.8	55	352	3.85	3.92	2.48
23	0.23	Very Good	H	VS1	61	57	353	3.94	3.96	2.41
24	0.31	Very Good	J	SI1	59.4	62	353	4.39	4.43	2.62
25	0.31	Very Good	J	SI1	58.1	62	353	4.44	4.47	2.59
26	0.23	Very Good	G	VVS2	60.4	58	354	3.97	4.01	2.41

Figure 1 Table Created from 'diamonds.csv'

Also, the screen shots for the Data Explorer representing the 'Columns' are shown below.



The screenshot shows the Databricks Catalog Explorer interface with the 'Columns' tab selected. The table 'default.diamonds' is displayed with 10 columns: carat, cut, color, clarity, depth, table, price, x, y, and z. The columns are listed with their types and comments. The 'About this table' section on the right provides additional information: Owner: Not set, Data source format: Delta, Last Updated: 16 minutes ago, Size: 503.4KiB, 1 file.

Column	Type	Comment
carat	double	
cut	string	
color	string	
clarity	string	
depth	double	
table	double	
price	bigint	
x	double	
y	double	
z	double	

About this table

Owner: Not set

Data source format: Delta

Last Updated: 16 minutes ago

Size: 503.4KiB, 1 file

Comment

Created by the file upload UI

Figure 2 Columns of table created from 'diamonds.csv'

Assignment II: Working with Data

Answer: This is the output of the feature 'cut' instead of 'color' and corresponding visualization of 'diamonds.csv'.

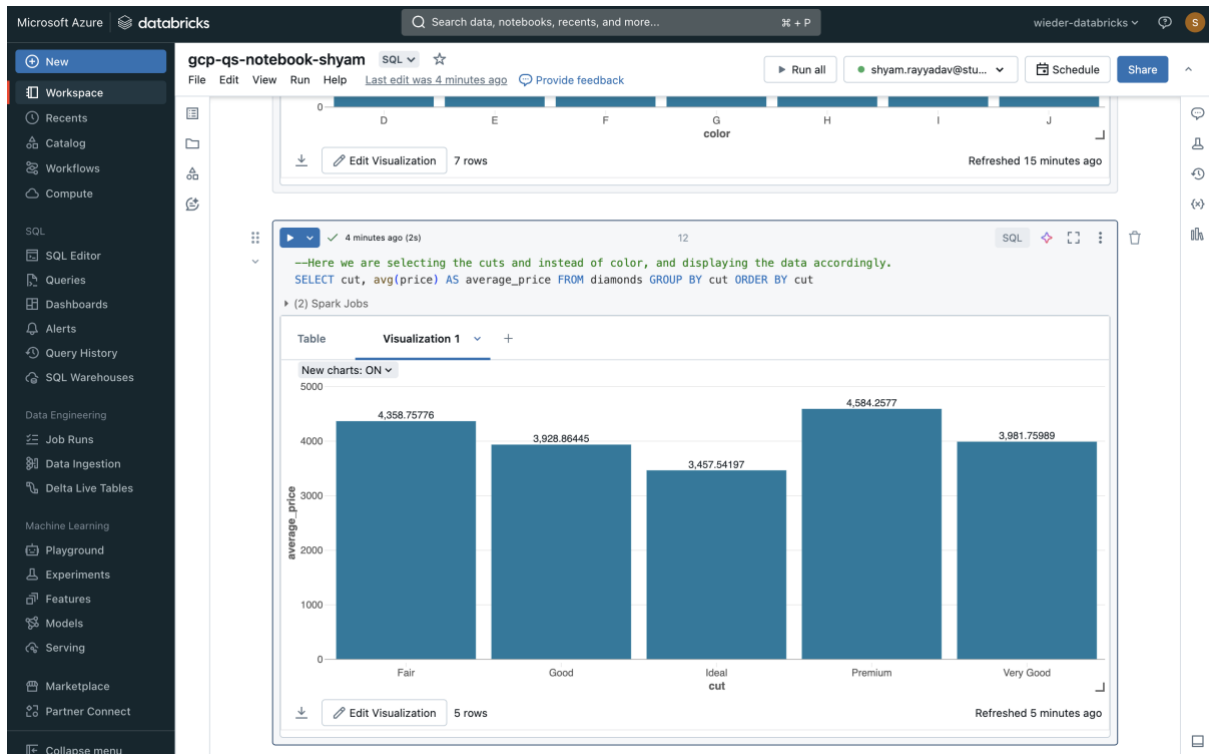


Figure 3 Changed code for visualizing average price of different cuts

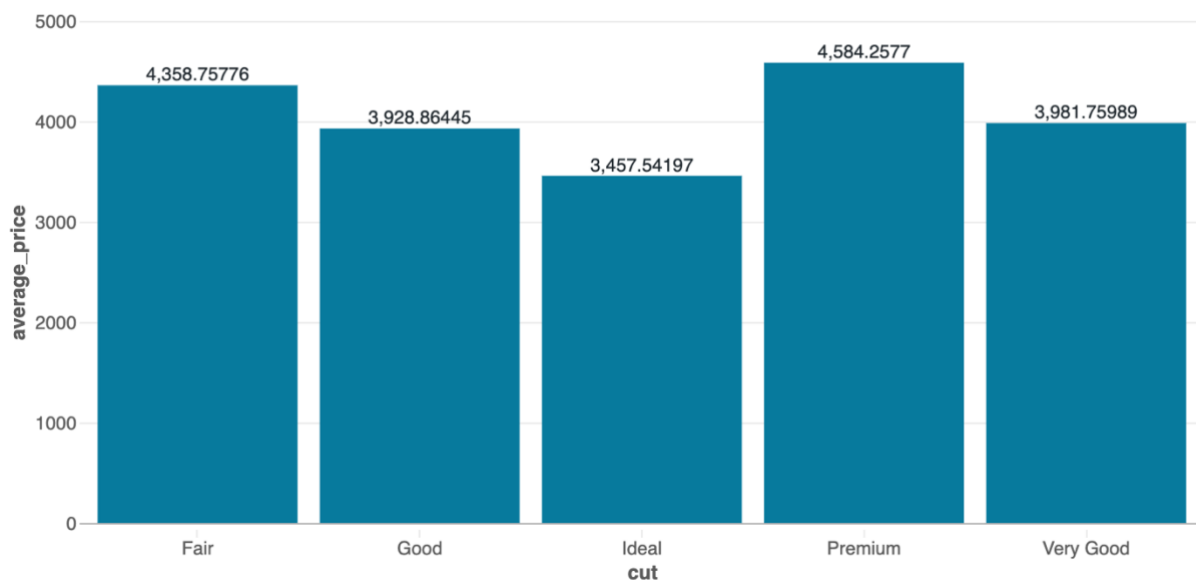


Figure 4 Correspoing output of average price of different cuts of diamonds

After having a look at the Spark UI in my cluster, I obtained the following outputs. It clearly shows the timeline and the last 8 jobs that has been executed.

JobsStagesStorageEnvironmentExecutorsSQL / DataFrameJDBC/ODBC ServerStructured StreamingConnect

Spark Jobs (?)

User: root
Started At: 2024/06/12 15:59:59
Total Uptime: 1.9 h
Scheduling Mode: FAIR
Completed Jobs: 112

Event Timeline

Completed Jobs (112)

Page: <<<1112131414 Pages. Jump to 14. Show 8 items in a page. Go

Job Id (Job Group)	Description	Submitted	Duration	Stages: Succeeded/Total	Tasks (for all stages): Succeeded/Total
7 (332458121465072756_5179057955133816262_1a4d77e3-afd7-4bbf-a6f5-43470f00a88b)	-- This is a system generated query from catalo... collectResult at OutputAggregator.scala:322	2024/06/12 16:04:15	4 s	1/1	1/1
6 (332458121465072756_5179057955133816262_1a4d77e3-afd7-4bbf-a6f5-43470f00a88b)	-- This is a system generated query from catalo... executeCollect at DatasetRefCache.scala:73	2024/06/12 16:04:13	0.7 s	1/1	4/4
5 (332458121465072756_7833744315458573157_845b590f-3907-468e-8230-b4c624932f17)	-- This is a system generated query from catalo... tail at StatisticsCollection.scala:1240	2024/06/12 16:04:06	0.3 s	1/1 (1 skipped)	1/1 (4 skipped)
4 (332458121465072756_7833744315458573157_845b590f-3907-468e-8230-b4c624932f17)	-- This is a system generated query from catalo... isEmpty at StatisticsCollection.scala:1237	2024/06/12 16:04:05	0.7 s	1/1 (1 skipped)	1/1 (4 skipped)
3 (332458121465072756_7833744315458573157_845b590f-3907-468e-8230-b4c624932f17)	-- This is a system generated query from catalo... \$anonfun\$withThreadLocalCaptured\$7 at LexicalThreadLocal.scala:63	2024/06/12 16:04:04	0.1 s	1/1	4/4
2 (332458121465072756_4637531669405663616_bcb62325-7681-4bda-a2cf-19bfc4ab50be)	-- This is a system generated query from catalo... tail at StatisticsCollection.scala:1240	2024/06/12 16:03:22	0.6 s	1/1 (1 skipped)	1/1 (4 skipped)
1 (332458121465072756_4637531669405663616_bcb62325-7681-4bda-a2cf-19bfc4ab50be)	-- This is a system generated query from catalo... isEmpty at StatisticsCollection.scala:1237	2024/06/12 16:03:20	2 s	1/1 (1 skipped)	1/1 (4 skipped)
0 (332458121465072756_4637531669405663616_bcb62325-7681-4bda-a2cf-19bfc4ab50be)	-- This is a system generated query from catalo... \$anonfun\$withThreadLocalCaptured\$7 at LexicalThreadLocal.scala:63	2024/06/12 16:03:16	2 s	1/1	4/4

Page: <<<1112131414 Pages. Jump to 14. Show 8 items in a page. Go