ScholarNest

Spark Azure Databricks

Databricks Spark Certification and beyond

Instructor: Prashant Kumar Pandey

APACHE Spark

Azure Databricks

Absolute Beginner to Specialization in Apache Spark and Azure Databricks

Assignments

Assignment - 1

- Setup your Databricks Community Cloud environment

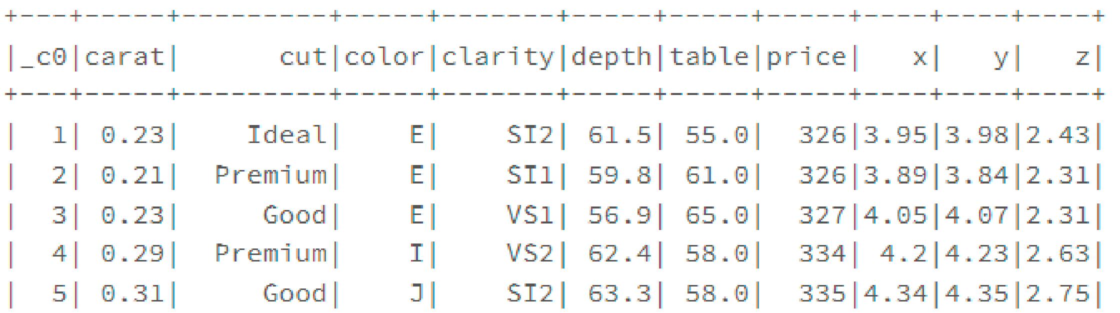
- Objectives:
  - You have access to Databricks Community Cloud
  - You can create a compute cluster in Databricks Community
  - You can create a Python Notebook
  - You can run Spark Data Frame Code (Diamonds Data Analysis)
- Solution:
  - 01-getting-started.ipynb

Diamonds Data Analysis

Given data file

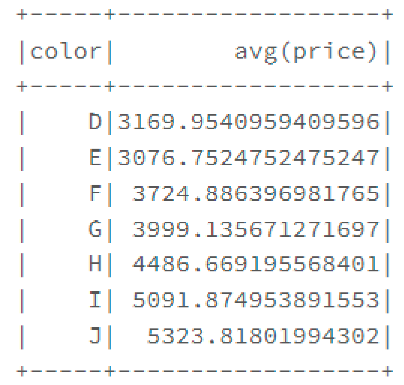
/databricks-datasets/Rdatasets/data-001/csv/ggplot2/diamonds.csv

1. Read the data into a frame and display the data frame



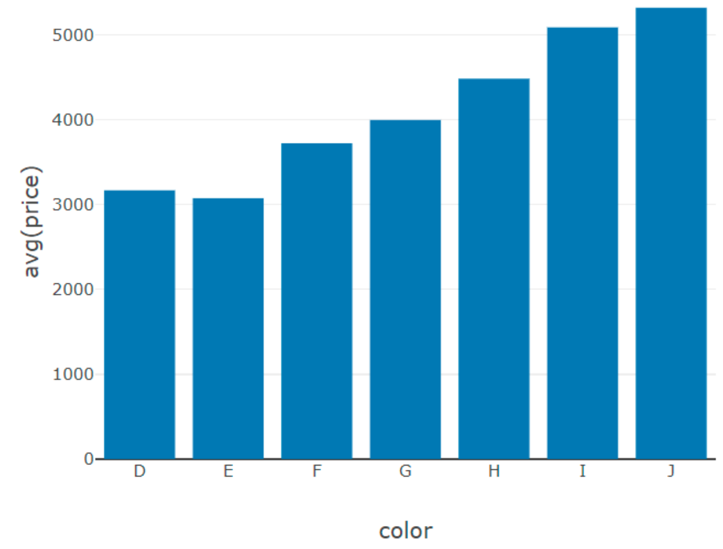
```
+---+-----+---------+-----+-------+-----+-----+-----+----+----+----+
|_c0|carat|      cut|color|clarity|depth|table|price|   x|   y|   z|
+---+-----+---------+-----+-------+-----+-----+-----+----+----+----+
|  1| 0.23|    Ideal|    E|    SI2| 61.5| 55.0|  326|3.95|3.98|2.43|
|  2| 0.21|  Premium|    E|    SI1| 59.8| 61.0|  326|3.89|3.84|2.31|
|  3| 0.23|     Good|    E|    VS1| 56.9| 65.0|  327|4.05|4.07|2.31|
|  4| 0.29|  Premium|    I|    VS2| 62.4| 58.0|  334| 4.2|4.23|2.63|
|  5| 0.31|     Good|    J|    SI2| 63.3| 58.0|  335|4.34|4.35|2.75|
```

2. Calculate Average Price by Colour



```
+-----+------------------+
|color|        avg(price)|
+-----+------------------+
|    D|3169.9540959409596|
|    E|3076.7524752475247|
|    F| 3724.886396981765|
|    G| 3999.135671271697|
|    H| 4486.669195568401|
|    I| 5091.874953891553|
|    J|  5323.81801994302|
+-----+------------------+
```

3. Show Bar chart of Avg. Price by Colour

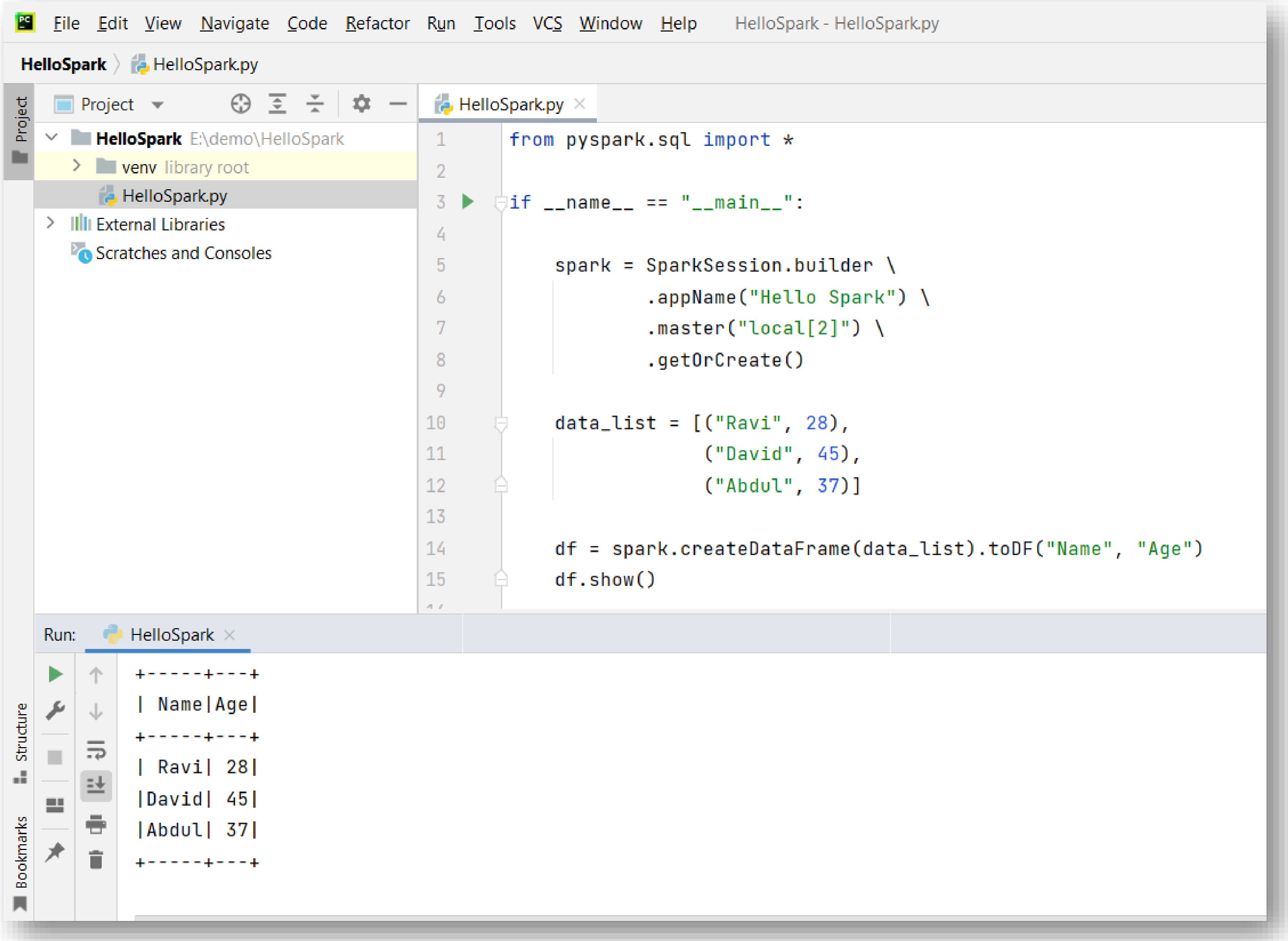


Assignment - 2

- Setup your Local Spark Development IDE

- Objectives:
  - You have access to local IDE
  - You can create and run Spark code locally (HelloSpark example)
- Solution:
  - HelloSpark.py

HelloSpark Application


File Edit View Navigate Code Refactor Run Tools VCS Window Help
HelloSpark - HelloSpark.py
HelloSpark 〉 HelloSpark.py
Project
HelloSpark E:\demo\HelloSpark
venv library root
HelloSpark.py
External Libraries
Scratches and Consoles
HelloSpark.py
Run: HelloSpark
Structure
Bookmarks

```python
from pyspark.sql import *

if __name__ == "__main__":

    spark = SparkSession.builder \
            .appName("Hello Spark") \
            .master("local[2]") \
            .getOrCreate()

    data_list = [("Ravi", 28),
                 ("David", 45),
                 ("Abdul", 37)]

    df = spark.createDataFrame(data_list).toDF("Name", "Age")
    df.show()
```

```
+-----+---+
| Name|Age|
+-----+---+
| Ravi| 28|
|David| 45|
|Abdul| 37|
+-----+---+
```

Thank You

ScholarNest Technologies Pvt Ltd.

www.scholarnest.com