# Databricks PySpark Homework

## Step 1: Create a Cluster

1. Go to the Compute tab in Databricks.  
2. Click 'Create Compute'.  
3. Name it 'Sakina’s Cluster'.  
4. Choose the smallest available node (e.g., e2-highmem-2).  
5. Set auto-termination to 15 minutes.  
6. Click 'Create'. Wait for it to show 'Running'.

## Step 2: Create a Notebook

1. Go to the Workspace tab.  
2. Click your username folder > Click 'Create' > Notebook.  
3. Name it 'Homework\_Transformations'.  
4. Set language to Python.  
5. Attach it to your running cluster.

## Step 3: Create a Sample DataFrame

Paste and run this code:  
  
from pyspark.sql import SparkSession  
from pyspark.sql.types import \*  
  
spark = SparkSession.builder.getOrCreate()  
  
data = [  
 ("Toyota", "Corolla", 2010, 30),  
 ("Honda", "Civic", 2012, 28),  
 ("Ford", "Focus", 2015, 25),  
 ("Nissan", "Sentra", 2018, 32),  
 ("Hyundai", "Elantra", 2020, 35)  
]  
  
schema = StructType([  
 StructField("Make", StringType(), True),  
 StructField("Model", StringType(), True),  
 StructField("Year", IntegerType(), True),  
 StructField("MPG", IntegerType(), True)  
])  
  
df = spark.createDataFrame(data, schema)  
df.show()

## Step 4: Apply Basic Transformations

Filter cars with MPG > 30:  
df.filter(df.MPG > 30).show()  
  
Group by Make and average MPG:  
df.groupBy("Make").avg("MPG").show()

## Step 5: SQL on DataFrame

Create temp view:  
df.createOrReplaceTempView("cars")  
  
Run SQL query:  
spark.sql("SELECT Make, COUNT(\*) as total\_models FROM cars GROUP BY Make").show()

## Step 6: Save Work (Optional)

Click File > Save to ensure your notebook is stored in Databricks.  
You may download it later as HTML or IPython notebook.

A screenshot of a computer

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.