```
!pip install pyspark
```

```
Collecting pyspark
  Downloading pyspark-3.5.0.tar.gz (316.9 MB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 316.9/316.9 MB 3.7 MB/s eta 0:00:00
  Preparing metadata (setup.py) ... done
Requirement already satisfied: py4j==0.10.9.7 in /usr/local/lib/python3.10/dist-packages (from pyspark) (0.10.9.7)
Building wheels for collected packages: pyspark
  Building wheel for pyspark (setup.py) ... done
  Created wheel for pyspark: filename=pyspark-3.5.0-py2.py3-none-any.whl size=317425344 sha256=a4ccdf504ee9fcd4df3c08069a260b26cc9a
  Stored in directory: /root/.cache/pip/wheels/41/4e/10/c2cf2467f71c678cfc8a6b9ac9241e5e44a01940da8fbb17fc
Successfully built pyspark
Installing collected packages: pyspark
Successfully installed pyspark-3.5.0
```

```
import pyspark
from pyspark.sql.functions import *


from pyspark import SparkContext, SparkConf
from pyspark.sql import SparkSession
conf = pyspark.SparkConf().setAppName("app").setMaster("local")
sc = pyspark.SparkContext(conf=conf)
spark = SparkSession(sc)


data = spark.read.csv ("car_data.csv", header = True, inferSchema = True)


data.show(5)
```

```
+--------+----+-------------+-------------+----------+---------+-----------+------------+
|Car_Name|Year|Selling_Price|Present_Price|Kms_Driven|Fuel_Type|Seller_Type|Transmission|
+--------+----+-------------+-------------+----------+---------+-----------+------------+
|    ritz|2014|         3.35|         5.59|     27000|   Petrol|     Dealer|      Manual|
|     sx4|2013|         4.75|         9.54|     43000|   Diesel|     Dealer|      Manual|
|    ciaz|2017|         7.25|         9.85|      6900|   Petrol|     Dealer|      Manual|
| wagon r|2011|         2.85|         4.15|      5200|   Petrol|     Dealer|      Manual|
|   swift|2014|          4.6|         6.87|     42450|   Diesel|     Dealer|      Manual|
+--------+----+-------------+-------------+----------+---------+-----------+------------+
only showing top 5 rows
```

```
type(data)
```

```
pyspark.sql.dataframe.DataFrame
```

```
len(data.columns)
```

```
8
```

```
list(data.columns)
```

```
['Car_Name',
 'Year',
 'Selling_Price',
 'Present_Price',
 'Kms_Driven',
 'Fuel_Type',
 'Seller_Type',
 'Transmission']
```

```
data.select("Car_Name", "Year").show()
```

```
+-------------+----+
|     Car_Name|Year|
+-------------+----+
|         ritz|2014|
|          sx4|2013|
|         ciaz|2017|
|      wagon r|2011|
|        swift|2014|
|vitara brezza|2018|
|         ciaz|2015|
|      s cross|2015|
|         ciaz|2016|
|         ciaz|2015|
|     alto 800|2017|
|         ciaz|2015|
|         ciaz|2015|
|       ertiga|2015|
|        dzire|2009|
```

```
|       ertiga|2016|
|       ertiga|2015|
|       ertiga|2016|
|      wagon r|2015|
|          sx4|2010|
+-------------+----+
only showing top 20 rows
```

```
data.select("*").filter(col('Year') == 2017).show()
```

```
+--------------------+----+-------------+-------------+----------+---------+-----------+------------+
|            Car_Name|Year|Selling_Price|Present_Price|Kms_Driven|Fuel_Type|Seller_Type|Transmission|
+--------------------+----+-------------+-------------+----------+---------+-----------+------------+
|                ciaz|2017|         7.25|         9.85|      6900|   Petrol|     Dealer|      Manual|
|           alto 800|2017|         2.85|          3.6|      2135|   Petrol|     Dealer|      Manual|
|              ignis|2017|          4.9|         5.71|      2400|   Petrol|     Dealer|      Manual|
|             swift|2017|            6|         6.49|     16200|   Petrol| Individual|      Manual|
|              ciaz|2017|         7.75|         9.29|     37000|   Petrol|     Dealer|   Automatic|
|            innova|2017|           18|        19.77|     15000|   Diesel|     Dealer|   Automatic|
|           fortuner|2017|           33|        36.23|      6000|   Diesel|     Dealer|   Automatic|
|            innova|2017|        19.75|        23.15|     11000|   Petrol|     Dealer|   Automatic|
|            innova|2017|           23|        25.39|     15000|   Diesel|     Dealer|   Automatic|
|      corolla altis|2017|           17|        18.64|      8700|   Petrol|     Dealer|      Manual|
| UM Renegade Mojave|2017|          1.7|         1.82|      1400|   Petrol| Individual|      Manual|
|          KTM RC200|2017|         1.65|         1.78|      4000|   Petrol| Individual|      Manual|
|   Bajaj Dominar 400|2017|         1.45|          1.6|      1200|   Petrol| Individual|      Manual|
|Royal Enfield Cla...|2017|         1.35|         1.47|      4100|   Petrol| Individual|      Manual|
|Royal Enfield Cla...|2017|          1.2|         1.47|     11000|   Petrol| Individual|      Manual|
|   Bajaj Avenger 220|2017|          0.9|         0.95|      1300|   Petrol| Individual|      Manual|
|Honda CB Hornet 160R|2017|          0.8|         0.87|      3000|   Petrol| Individual|      Manual|
|   Yamaha FZ S V 2.0|2017|         0.78|         0.84|      5000|   Petrol| Individual|      Manual|
|Honda CB Hornet 160R|2017|         0.75|         0.87|     11000|   Petrol| Individual|      Manual|
|   Bajaj Avenger 220|2017|         0.75|         0.95|      3500|   Petrol| Individual|      Manual|
+--------------------+----+-------------+-------------+----------+---------+-----------+------------+
only showing top 20 rows
```

```
data.select("Fuel_Type").distinct().show()
```

```
+---------+
|Fuel_Type|
+---------+
|   Diesel|
|      CNG|
|   Petrol|
+---------+
```

```
data.select("*").filter( (col('Year') == 2017) & (col('Fuel_Type') == 'Petrol')).show(5)
```

```
+--------+----+-------------+-------------+----------+---------+-----------+------------+
|Car_Name|Year|Selling_Price|Present_Price|Kms_Driven|Fuel_Type|Seller_Type|Transmission|
+--------+----+-------------+-------------+----------+---------+-----------+------------+
|    ciaz|2017|         7.25|         9.85|      6900|   Petrol|     Dealer|      Manual|
|alto 800|2017|         2.85|          3.6|      2135|   Petrol|     Dealer|      Manual|
|   ignis|2017|          4.9|         5.71|      2400|   Petrol|     Dealer|      Manual|
|   swift|2017|            6|         6.49|     16200|   Petrol| Individual|      Manual|
|    ciaz|2017|         7.75|         9.29|     37000|   Petrol|     Dealer|   Automatic|
+--------+----+-------------+-------------+----------+---------+-----------+------------+
only showing top 5 rows
```

```
data.select("Car_Name", "Year", "Selling_Price").filter( (col('Year') == 2017) & (col('Fuel_Type') == 'Petrol')).show(5)
```

```
+--------+----+-------------+
|Car_Name|Year|Selling_Price|
+--------+----+-------------+
|    ciaz|2017|         7.25|
|alto 800|2017|         2.85|
|   ignis|2017|          4.9|
|   swift|2017|            6|
|    ciaz|2017|         7.75|
+--------+----+-------------+
only showing top 5 rows
```

```
data.select("*").orderBy("Selling_Price", ascending = False). show(10)
```

```
+-------------+----+-------------+-------------+----------+---------+-----------+------------+
|     Car_Name|Year|Selling_Price|Present_Price|Kms_Driven|Fuel_Type|Seller_Type|Transmission|
+-------------+----+-------------+-------------+----------+---------+-----------+------------+
|         city|2015|          9.7|         13.6|     21780|   Petrol|     Dealer|      Manual|
|     fortuner|2010|         9.65|        20.45|     50024|   Diesel|     Dealer|      Manual|
|         city|2016|          9.5|         11.6|     33988|   Diesel|     Dealer|      Manual|
```

```
|vitara brezza|2018|         9.25|         9.83|      2071|   Diesel|     Dealer|      Manual|
|        verna|2017|         9.25|          9.4|     15001|   Petrol|     Dealer|      Manual|
|     fortuner|2010|         9.25|        20.45|     59000|   Diesel|     Dealer|      Manual|
|         city|2016|         9.15|         13.6|     29223|   Petrol|     Dealer|      Manual|
|        verna|2017|          9.1|          9.4|     15141|   Petrol|     Dealer|      Manual|
|         city|2016|         8.99|         11.8|      9010|   Petrol|     Dealer|      Manual|
|         ciaz|2016|         8.75|         8.89|     20273|   Diesel|     Dealer|      Manual|
+-------------+----+-------------+-------------+----------+---------+-----------+------------+
only showing top 10 rows
```

```
data.dtypes
```

```
[('Car_Name', 'string'),
 ('Year', 'int'),
 ('Selling_Price', 'double'),
 ('Present_Price', 'double'),
 ('Kms_Driven', 'int'),
 ('Fuel_Type', 'string'),
 ('Seller_Type', 'string'),
 ('Transmission', 'string')]
```

```
# when loading data from csv everything loads as string, add inferSchema parameter while reading the file
```

```
data.select("*").orderBy("Selling_Price", ascending = False). show(10)
```

```
+------------+----+-------------+-------------+----------+---------+-----------+------------+
|    Car_Name|Year|Selling_Price|Present_Price|Kms_Driven|Fuel_Type|Seller_Type|Transmission|
+------------+----+-------------+-------------+----------+---------+-----------+------------+
|land cruiser|2010|         35.0|         92.6|     78000|   Diesel|     Dealer|      Manual|
|    fortuner|2017|         33.0|        36.23|      6000|   Diesel|     Dealer|   Automatic|
|    fortuner|2015|         23.5|        35.96|     47000|   Diesel|     Dealer|   Automatic|
|    fortuner|2015|         23.0|        30.61|     40000|   Diesel|     Dealer|   Automatic|
|    fortuner|2015|         23.0|        30.61|     40000|   Diesel|     Dealer|   Automatic|
|      innova|2017|         23.0|        25.39|     15000|   Diesel|     Dealer|   Automatic|
|      innova|2016|        20.75|        25.39|     29000|   Diesel|     Dealer|   Automatic|
|    fortuner|2014|        19.99|        35.96|     41000|   Diesel|     Dealer|   Automatic|
|      innova|2017|        19.75|        23.15|     11000|   Petrol|     Dealer|   Automatic|
|    fortuner|2014|        18.75|        35.96|     78000|   Diesel|     Dealer|   Automatic|
+------------+----+-------------+-------------+----------+---------+-----------+------------+
only showing top 10 rows
```

```
data.select("*").filter( col('Year') == 2017).count()
```

```
35
```

```
data.select("*").groupby("Year").count().show()
```

```
+----+-----+
|Year|count|
+----+-----+
|2003|    2|
|2007|    2|
|2018|    1|
|2015|   61|
|2006|    4|
|2013|   33|
|2014|   38|
|2004|    1|
|2012|   23|
|2009|    6|
|2016|   50|
|2005|    4|
|2010|   15|
|2011|   19|
|2008|    7|
|2017|   35|
+----+-----+
```

```
#writing sql in pyspark
data.createOrReplaceTempView("CarDataTb")
```

```
#notice the camel case in pyspark
```

```
spark.sql("select * from CarDataTb").show(5)
```

```
+--------+----+-------------+-------------+----------+---------+-----------+------------+
|Car_Name|Year|Selling_Price|Present_Price|Kms_Driven|Fuel_Type|Seller_Type|Transmission|
+--------+----+-------------+-------------+----------+---------+-----------+------------+
|    ritz|2014|         3.35|         5.59|     27000|   Petrol|     Dealer|      Manual|
```

```
|     sx4|2013|         4.75|         9.54|     43000|   Diesel|     Dealer|      Manual|
|    ciaz|2017|         7.25|         9.85|      6900|   Petrol|     Dealer|      Manual|
| wagon r|2011|         2.85|         4.15|      5200|   Petrol|     Dealer|      Manual|
|   swift|2014|          4.6|         6.87|     42450|   Diesel|     Dealer|      Manual|
+--------+----+--------------+-------------+----------+---------+-----------+------------+
only showing top 5 rows
```

```
spark.sql("select Car_Name, Year from CarDataTb").show(5)
```

```
+--------+----+
|Car_Name|Year|
+--------+----+
|    ritz|2014|
|     sx4|2013|
|    ciaz|2017|
| wagon r|2011|
|   swift|2014|
+--------+----+
only showing top 5 rows
```

```
spark.sql("select Year, count(*) from CarDataTb group by Year").show(5)
```

```
+----+--------+
|Year|count(1)|
+----+--------+
|2003|       2|
|2007|       2|
|2018|       1|
|2015|      61|
|2006|       4|
+----+--------+
only showing top 5 rows
```

```
spark.sql("select Year, Fuel_Type, count(*) from CarDataTb group by Year, Fuel_Type order by Year, Fuel_Type").show(5)
```

```
+----+---------+--------+
|Year|Fuel_Type|count(1)|
+----+---------+--------+
|2003|   Petrol|       2|
|2004|   Petrol|       1|
|2005|   Diesel|       1|
|2005|   Petrol|       3|
|2006|   Petrol|       4|
+----+---------+--------+
only showing top 5 rows
```

```
#old method
data.select("*").filter( (col('Year') == 2017) & (col('Fuel_Type') == 'Petrol')).show(5)
```

```
+--------+----+-------------+-------------+----------+---------+-----------+------------+
|Car_Name|Year|Selling_Price|Present_Price|Kms_Driven|Fuel_Type|Seller_Type|Transmission|
+--------+----+-------------+-------------+----------+---------+-----------+------------+
|    ciaz|2017|         7.25|         9.85|      6900|   Petrol|     Dealer|      Manual|
|alto 800|2017|         2.85|          3.6|      2135|   Petrol|     Dealer|      Manual|
|   ignis|2017|          4.9|         5.71|      2400|   Petrol|     Dealer|      Manual|
|   swift|2017|          6.0|         6.49|     16200|   Petrol| Individual|      Manual|
|    ciaz|2017|         7.75|         9.29|     37000|   Petrol|     Dealer|   Automatic|
+--------+----+-------------+-------------+----------+---------+-----------+------------+
only showing top 5 rows
```

```
#spark sql method
spark.sql("Select * from CarDataTb where year = 2017 and fuel_type = 'Petrol'").show(5)
```

```
+--------+----+-------------+-------------+----------+---------+-----------+------------+
|Car_Name|Year|Selling_Price|Present_Price|Kms_Driven|Fuel_Type|Seller_Type|Transmission|
+--------+----+-------------+-------------+----------+---------+-----------+------------+
|    ciaz|2017|         7.25|         9.85|      6900|   Petrol|     Dealer|      Manual|
|alto 800|2017|         2.85|          3.6|      2135|   Petrol|     Dealer|      Manual|
|   ignis|2017|          4.9|         5.71|      2400|   Petrol|     Dealer|      Manual|
|   swift|2017|          6.0|         6.49|     16200|   Petrol| Individual|      Manual|
|    ciaz|2017|         7.75|         9.29|     37000|   Petrol|     Dealer|   Automatic|
+--------+----+-------------+-------------+----------+---------+-----------+------------+
only showing top 5 rows
```

```
#VIFI
# Any question practive in 3 formats - Sql, Pandas Datframe, Spark dataframe, (Spark sql is a given)
```

```
spark.sql("select Year, count(*) as total_cars from CarDataTb group by Year, Fuel_Type order by count(*) desc ").show(5)
```

```
spark.sql("select Year, count(*) as total_cars from carDataTb group by Year, Fuel_Type order by count(*) desc").show(5)
```

```
+----+----------+
|Year|total_cars|
+----+----------+
|2015|        45|
|2016|        42|
|2017|        31|
|2013|        28|
|2014|        24|
+----+----------+
only showing top 5 rows
```