This exercise is aimed at comparing the running time of two equivalent solutions in PySpark, one based on RDD and another on dataframe, to determine which is faster.

- **PC Features:**

  -Windows 10 Pro 64 bits

  -Processor: Intel(R) Core(TM) i5-4460 CPU @ 3.2 GHz 3.2 GHz

  -16 Gb RAM

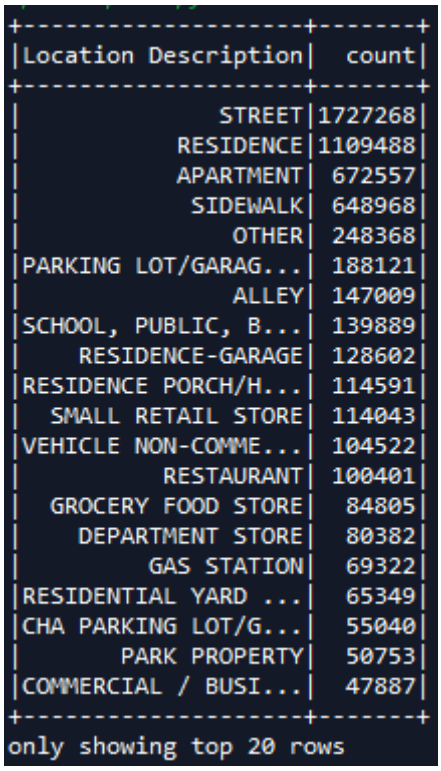
  -Hard Drive SSD 500 Gb

  -Java JDK 1.8.0_161

  -Spark 2.3.0

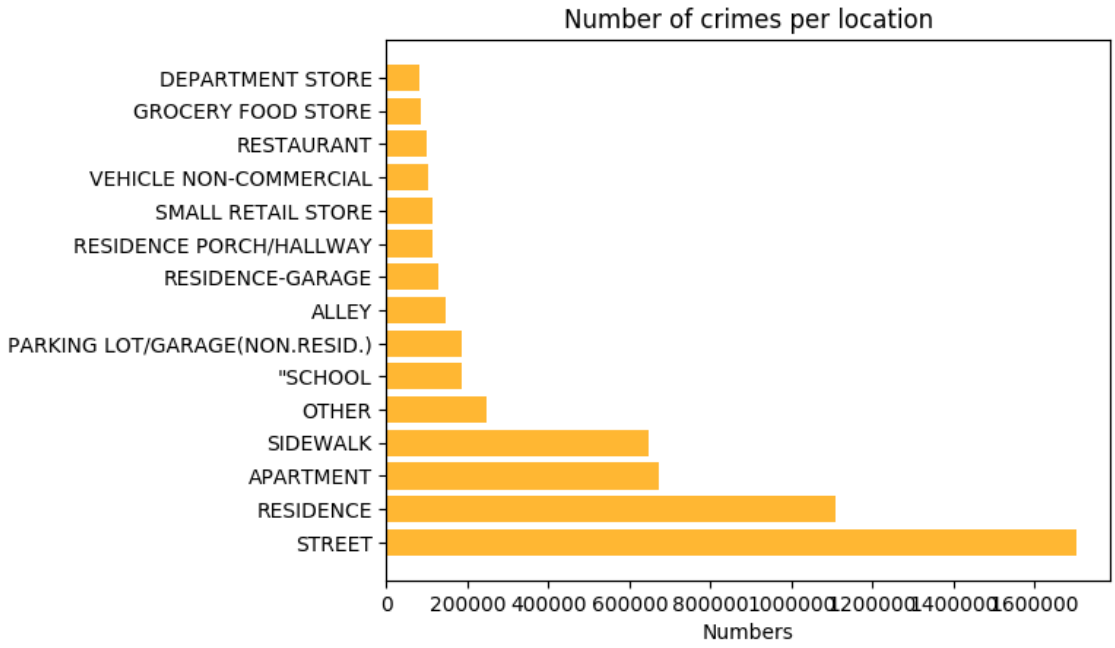
Results:

Numbers of crimes per location, first 20 items:



Location Description	count
STREET	1727268
RESIDENCE	1109488
APARTMENT	672557
SIDEWALK	648968
OTHER	248368
PARKING LOT/GARAG...	188121
ALLEY	147009
SCHOOL, PUBLIC, B...	139889
RESIDENCE-GARAGE	128602
RESIDENCE PORCH/H...	114591
SMALL RETAIL STORE	114043
VEHICLE NON-COMME...	104522
RESTAURANT	100401
GROCERY FOOD STORE	84805
DEPARTMENT STORE	80382
GAS STATION	69322
RESIDENTIAL YARD ...	65349
CHA PARKING LOT/G...	55040
PARK PROPERTY	50753
COMMERCIAL / BUSI...	47887

only showing top 20 rows



- Computing Time Based on RDD

Nº Cores: 1

Computing time: 149.79009985923767 sg. / 2’ 29”

Nº Cores: 2

Computing time: 117.60313010215759 sg. / 1’ 57”

Nº Cores: 3

Computing time: 117.3607542514801 sg. / 1’ 5726,14,10,8“

Nº Cores: 4

Computing time: 111.81673979759216 sg. / 1’ 51”

- Computing Time Based on dataframe

Nº Cores: 1

Computing time: 26.997371673583984 sg.

Nº Cores: 2

Computing time: 14.671002388000488 sg.

Nº Cores: 3

Computing time: 10.259229183197021 sg.

Nº Cores: 4

Computing time: 8.28029727935791 sg.