PySpark RDD Operations:

- RDD is core data Structure of PySpark.
- RDD's are ~~be~~ a low level obj and are highly efficient in performing distributed tasks.

Set of Operations:-

① Transformations:- Operations that take RDD as input and produce another RDD as output.

- After Transformation is applied to an RDD, it returns a new RDD, the original RDD remains the same and thus are immutable.
- After Applying Transformation a DAG is created for computations. & ends after applying any actions on it.
- This is called the lazy evaluation process.

② Actions.

- Applied on RDD to produce a single value.
- Applied on resultan RDDs producing a non-RDD result, thus removing the laziness of the transformations of RDD.

for eg :- Collect()

The collect() action gives a list of all elements of the RDD.

```
collect_rdd = sc.parallelize([1,2,3,4,5])
print(collect_rdd.collect())
```

O/P will be :- [1, 2, 3, 4, 5].

Selecting, Renaming, filtering Data in Pandas DataFrame.

1. Creating DataFrame
2. Using withColumnRenamed()
3. Using selectExpr()
4. Using select()
5. Using toDF()