PySpark LabWork – 2

- You Can find the iris.csv file in the provided zip.
- Use this datafile and answer the following questions
- Use the following link as documentation:
https://spark.apache.org/docs/latest/api/python/reference/pyspark.sql/dataframe.html

Questions:

1. What is the count of each “variety”?
2. Find the average of "sepal.length","sepal.width","petal.length","petal.width" based on the variety
3. Find the difference between the average of the above mentioned based on variety.
Eg: y is the difference between mean sepal.length between ,”Setosa” and “Versicolor”
You are expected to find the value of y for everything
4. Find the outliers for each variety.
Process:
- Find the mean, then standard deviation
- If some value is outside the mean ± standard deviation consider it as outlier

5. Out of all the examples above figure out more about transformations and actions. Use the explain and understand how the spark is executing the instructions under the hood.