Exercise – SQL and Plotting with Pandas

Objective

The objectives for this exercise are:

1. Gain familiarity with creating a SparkSession with pyspark
2. Use the Spark DataFrame API to query a dataset
3. Gain familiarity with converting a Spark DataFrame to a Pandas DataFrame

Steps

1. Create an empty Jupyter notebook
2. Create a SparkSession object with master set to “local[*]”
3. Read the csv file “countries.csv” into a Spark DataFrame
4. Use the Spark DataFrame API to answer the following questions about the data that was just loaded:
    a. Display the first 15 rows
    b. Display all rows that have the region “Antarctica” (hint: use the ‘filter’ function)
    c. How many unique entries are in the Code field? (hint: use the ‘select’ function)
    d. Create a tempView so that you can query the data with SQL statements
    e. Using the tempView display the unique values of the “GovernmentForm” column
    f. Convert the data type of the LifeExpectancy, Population & GNP columns to the type “pyspark.sql.types.DoubleType”
    g. What is the average population per continent (hint: use the “groupby” and “avg” functions)
    h. Convert the Spark DataFrame to a pandas DataFrame and display the first 5 rows
    i. Optional: use the pandas DataFrame to create a line chart of LifeExpectancy Vs GNP