Assignment 3 –Spark script

Name: Tim Gras
Studentnumber: 630259
Github URL =>   https://github.com/dantim1997/PDP-Assignments-
630259/tree/main/Assignment%203

Setup

I could not run this script on Hadoop so I runed it locally in my pc with this tutorial:
https://phoenixnap.com/kb/install-spark-on-windows-10
To run the code, install a following this tutorial.
Make sure to have spark running.
Run the file Assignment_3.py.

execute

Run the file Assignment_3.py to execute the code.

Information

The script searches gets the information about the titanic.
In assignment A it will look which group on average survived based on the class they were in.
In assignment B it will look for the probability a child survived based on age and class.
In assignment C it will look for the average fare the different classes payed
The script itself has the code explanation.
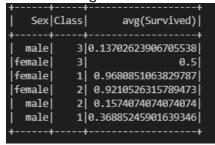
GitHub

Assignment_3.py   This is the script that will be used to look at all the assignments A trough C.
Assignment 3.docx   This is the documentation.
Result.jpg   the results of the script in a jpg format.

Result:

Result for assignment A:

```
+------+-----+-------------------+
|   Sex|Class|      avg(Survived)|
+------+-----+-------------------+
|  male|    3|0.13702623906705538|
|female|    3|                0.5|
|female|    1| 0.9680851063829787|
|female|    2| 0.9210526315789473|
|  male|    2| 0.1574074074074074|
|  male|    1|0.36885245901639346|
+------+-----+-------------------+
```

Result for assignment B:

```
probability that a child in the third class will survive: 41.509433962264154
```

Result for assignment C:

```
+-----+-----------------+
|Class|   avg(PayedFare)|
+-----+-----------------+
|    1| 84.15468752825701|
|    3|13.707707501045244|
|    2| 20.66218318109927|
+-----+-----------------+
```