PDP assignment 3 – Chris Lips

Name: Christiaan Lips
Student number: 623434
Github: https://github.com/Lipsinator/Hadoop-fundamentals-Chris-Lips

Steps:

1. For this assignment I couldn't get spark to work on Hadoop so I found a workaround online that lets you install spark on windows which happened to workout for me. Link: https://phoenixnap.com/kb/install-spark-on-windows-10
2. Run spark with command: "C:\Spark\spark-2.4.5-bin-hadoop2.7\bin\spark-shell"
3. Execute the file from the terminal when spark is running.

Results:

Assignment 3A:

```
+------+------+------------------+
|   Sex|Pclass|      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|
+------+------+------------------+
```

Assignment 3B:

```
probability of survival: 41.509433962264154
```

Assignment 3C:

```
+------+-----------------+
|Pclass|        avg(Fare)|
+------+-----------------+
|     1| 84.15468752825701|
|     3|13.707707501045244|
|     2| 20.66218318109927|
+------+-----------------+
```

Code explanation:

1. First we load the data from the csv into a data frame.
2. For assignment A I select the fields survived, passenger class and Sex to group by sex and passenger class. Then the average survived is calculated with the result as show above.
3. For assignment B I select the age class and survived from the data frame. Then I filter on age under or equals 10 to then count the children that survived. Finally we calculate the probability of surviving on the titanic as a child on the third class. The results is 41,5 percent.
4. For assignment C the fare and passenger class is selected and then grouped by the passengers and class to calculate the average fare. This result is shown above aswell.