INF 553 Homework 1

My three scala codes, were saved in a package called Georgios.
As a result in front of the Task1, I have to specify the package.
I will show the code required to run it sucessfully:

spark-submit --class Georgios.Task1 Georgios_Iliadis.jar
spark-submit --class Georgios.Task2 Georgios_Iliadis.jar
spark-submit --class Georgios.Task3 Georgios_Iliadis.jar

First you have to cd where the jar file is and then use the three commands provided to run Task1,2 and 3. I have the data needed in the directory where the .jar file is.
I did not include any arguments for the command line, so the data to be loaded has to be where the .jar file is, as I have it in my submission.
By using the commands provided, it will load the file necessary and it will create the output file with the name required.

Output Files:
Once I run the command lines provided, the output files will be created where the .jar files are.
The output files were created as folders.Tha folders have the name that the .csv file should have.
In the folder you can find the .csv file required with the results.

Task 2:
For Task 2, I do not print out the number of partitions for the RDD in Task1.
I have commented the code to do that,(it's where it says NUMBER OF PARTITIONS) but I did not write it as a number in the csv file created.
In the csv file though, this can be understood, by looking at the second column.
The second column is an array which has the number of items per partition.
You can see how many values there are in the array, and so that's the number of partitions.
For example -> partition [17294,32163] 422.
Here we can see that in the second column the length of the array is two, so the number of partitions is two.

Note:
When I run the spark-sumbit command, the output is generated correctly as a csv file, but I get this error:
"ERROR ShutdownHookManager: Exception while deleting Spark temp dir:C\Users\gelia\AppData\Local\Temp\spark-26523f80-a3da-494c-894e-8f7943774f32"...
The output files though are created correctly and they are the ones required.
Lastly, I have all the printed results (if you want to see them on the command lines) as comments. You can simply remove the comments and see the results in the command line as well instead of only having the .csv files.

Name: Georgios Iliadis
3668057286