CS 6350

ASSIGNMENT \_\_2\_\_\_\_\_\_\_\_\_

Names of students in your group:

Venkata Kowsik Temididapathi – VXT200001

Veda Nandan Gandi – VXG200001

Number of free late days used: \_\_\_\_0\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_   
Note: You are allowed a **total** of 4 free late days for the **entire semester**. You can use at most 2 for each assignment. After that, there will be a penalty of 10% for each late day.

Please list clearly all the sources/references that you have used in this assignment.

* <http://lintool.github.io/UMD-courses/bigdata-2015-Spring/slides/session05.pdf>
* <https://towardsdatascience.com/databricks-how-to-save-files-in-csv-on-your-local-computer-3d0c70e6a9ab>
* <https://stackoverflow.com/questions/49019706/databricks-download-a-dbfs-filestore-file-to-my-local-machine>
* <https://community.databricks.com/s/question/0D53f00001HKHTXCA5/download-a-dbfsfilestore-file-to-my-local-machine>
* <https://spark.apache.org/docs/2.2.0/ml-classification-regression.html#logistic-regression>
* <https://spark.apache.org/docs/latest/ml-tuning.html#cross-validation>
* <https://spark.apache.org/docs/2.2.0/mllib-evaluation-metrics.html>

Part 1:

Input: s3://bigdatacs6350s22/airport.csv

Python Notebook: <https://databricks-prod-cloudfront.cloud.databricks.com/public/4027ec902e239c93eaaa8714f173bcfc/4241627352815915/1871106947173904/1617445708688506/latest.html>

Part 2:

Input: s3://bigdatacs6350s22/Tweets.csv

Python Notebook: <https://databricks-prod-cloudfront.cloud.databricks.com/public/4027ec902e239c93eaaa8714f173bcfc/4241627352815915/745801424267021/1617445708688506/latest.html>

Note:

* The inputs are given as a list called inputs in the first/second command.
* The input files are stored as a file in Amazon s3 and are made public.
* The output is printed in the databrick notebook as the last command.