MIE1512 Data Analytics Project Submission Guidelines

The final course project submission must include the following deliverables (please upload all the files to portal in a single compressed folder).

1. A final report document (your choice of latex, word, pdf, html). Please name this document LastName-FirstName-ProjDesc (where ProjDesc is a short 2-3 word version of your project title).
2. A final notebook that can be reproduced on Databricks (preferred) or DSW (these requires including data file references as described below under "Reproducibility"). Please name this notebook LastName-FirstName-ProjDesc (where ProjDesc is a short 2-3 word version of your project title).
3. Any additional working notebooks (including additional work completed since your V3 notebook submission that did not make it in to the final notebook), as well as any other deliverable specific to the project (e.g., data gathering programs, tableau workbooks).

In most cases the final report document can be generated from the final notebook. Other cases where there has been additional programs (e.g., data gathering) or tools (eg, tableau) used in the project, will have sections in the final report that do not correspond to the contents of the final notebook. Students using Databricks/Jupyter notebooks can easily export their contents to create the final document (latex, html, pdf, etc.). Students using Zepellin may want to consider importing the notebook into www.zeppelinhub.com/viewer to have access to an html version (that can be used to create the final report).

The final report/notebook must

1. Introduce the data analysis objectives of your project and relate them to your selected paper.
2. Describe the datasets used (the data sources and how they were combined), the data exploration required, the data quality considerations, the data preparation steps, and the application of the data analysis method(s) selected in your project.
3. Describe what analysis you will do, why, and how, as well as the validation/assessment of the results (describe what you have obtained and what does it mean, using visualizations whenever appropriate, and identifying interesting results).
4. Describe your results over a few iterations (e.g., iterations that analyze results on different subsets and/or categories of the data, allowing you to compare and contrast results).
5.) Include a complete list of full bibliographic references (incomplete or improperly styled/formatted references will cause deductions!). Please include as your first reference the paper selected to (approximately) reproduce in your project.

Reproducibility:

Make sure all the data files that are used in your notebook are accessible directly from code in the notebook. Providing a link to a webpage which has the files is not sufficient.

You can store the files on an online public storage service (such as google drive) and access this data from your notebook (again, we will not download the data. It has to be accessed in the notebook).

See example (taken from Lab 2):

```
import org.apache.commons.io.IOUtils
import java.net.URL
import java.nio.charset.Charset

val taxiText = sc.parallelize(
    IOUtils.toString(
        new
URL("https://drive.google.com/uc?export=download&id=0B1K-
gHwDFeUENDJaS3p6bWtWN0E"),
        Charset.forName("utf8")).split("\n"))
```