Homework 1

The purpose of this homework is to get used to using the tools.

Part 1: Setup class github

Setup an account with github.com and checkout out the class repository. The repository contains an exported Databricks Notebook.

https://github.com

The class repository is at

https://github.com/dosirrah/CSCI443_AdvancedDataScience

Figure out how to clone a repository from the command-line, from a git client, or using PyCharm. If you like to develop Python locally rather than fully in a notebook, I suggest using PyCharm from jetbrains. If you have used IntelliJ then PyCharm should be familar, although PyCharm is optimized for use with Python.

If you do not already have PyCharm installed, go to

https://www.jetbrains.com/community/education/#students

and apply for a "Free Education License." Do not download the trial. You can get a free full license for educational purposes if you apply using a .edu email address to setup your account.

On the specific page, scroll down to "Apply Now" and fill out the form. You will get an email almost immediately with instrutions on what to do.

You can link PyCharm to your GitHub account allowing you to clone repositories to your local system.

Once you have cloned the repository, locate the file

hw1/Hello World Notebook.dbc

There is nothing to turn in for this problem. It is just a step toward completion of the next problems.

Part 2: "Hello World"

Setup a Databricks account.

https://community.cloud.databricks.com/login.html

As with PyCharm, avoid using the trial of the full version and instead use the free community edition. There are some slides outlining how to sign up in the Lecture 1 slides.

One you have an account, upload the "Hello World Notebook.dbc" obtained from github to Databricks.

From the notebook interface, select "Run all". This will request you attach to a cluster. You will have to create new cluster.

Part 3: Access Kaggle and Upload data to Databricks

Create an account with kaggle. Kaggle is a great source for small datasets used in data science competitions. Many of the datasets have associated forums with discussion on how to work with the data.

Download from kaggle the training titanic dataset `train.csv` from

https://www.kaggle.com/competitions/titanic/data#

From the Databricks Notebook, select "File -> Upload data to DBFS…" DBFS stands for DataBricks File System and is an abstraction on top of other file systems.

When I perform this upload, by default the files are placed in

`/FileStore/shared_uploads/harrison@cs.olemiss.edu/`

You can confirm that an upload is successful from within the Notebook by creating a python cell and running the following:

```
display(dbutils.fs.ls("/FileStore/shared_uploads/harrison@cs.olemiss.edu/"))
```

Replace `harrisonb@cs.olemiss.edu` with the path you used. In my Databricks Notebook I see

```
dbfs:/FileStore/shared_uploads/harrison@cs.olemiss.edu/train.csv
train.csv
61194
1706306094000
```

Part 4: Use a DataFrame

Extend the "Hello World" notebook from within Databricks to load `train.csv` into a DataFrame. Output the first 10 rows of the DataFrame.

Part 5: Use matplotlib

Starting from the "Hello World" notebook from within Databricks, plot a histogram of the ages of passengers on the Titanic using matplotlib using bins each spanning 5 years of age.

Submission Instructions

Export the workbook after running it and submit the notebook to blackboard to assignement "HW 1".