**Project Design: Finding Datasets**

**Assignment Specification**

**Introduction**: For the final project, you will design and carry out a data analysis of your choice. In order to do this, you will need an appropriate dataset. A good approach is to start with a problem of interest and then look for data. There are dataset links posted on Canvas.

When you go to a site, you must look for data that are available in a format that you can read and use with a python program. For example, a .pdf summary report is not useful, but there may be a dataset of the detailed data available in .txt or .csv format (or .xlsx which you can convert to .csv).

**Description**:

1. Find a dataset online that would be appropriate for a data analysis project. You will not be required to use this for your project: you may decide later on to use different data.
2. In a word/pdf file, post the following info:

**Your name**

**URL** – where dataset can be retrieved (see Note #1)

**Description 1** – brief description of the dataset

**Description 2** – brief description of the analysis you plan to perform on the dataset

**Source of the data** – organization, e.g. Department of Transportation

**Types of data items** – List or description of the types of data items in the dataset

**Number of records and sample**: run the checkfile.py program on your dataset. This will count the lines and print the first 5 lines. Take a screenshot of your results, rename the image file with your name, and submit it in the “Project Design” assignment box.

**Additional requirements**:

1. Assignments are individual and have to be unique. Different students can use the same dataset but they need to perform different analyses on it.
2. Dataset/analysis will be approved by teacher/TA based on first come/first serve principle: if 2 students propose the same analysis on the same dataset, the first one who submitted it will be accepted, the second will be refused. In the unlikely situation of same time submission, the choice will be random.
3. Students who will have their proposal refused have to submit a different one.

**Notes**:

1. The URL you post and include in your report must be the url of the specific page where you can download the data. For example, **not just data.gov**, but **https://explore.data.gov/Education/Post-Secondary-Universe-Survey-2010-Awards-degrees/fjce-ze3t**
2. Remember that it is easiest to run a python program that reads in a file, if you put the program and the file into the same folder.
3. Note that when saving a file in .csv format, it is sometimes necessary to save it as Windows .csv format to avoid the problem where it loses the line feeds and interprets the file as one long line.