[**Assignment06: Python dataset exploration**](https://mymasonportal.gmu.edu/webapps/assignment/uploadAssignment?content_id=_8335369_1&course_id=_347779_1&group_id=&mode=view)

**Assignment06: Python data exploration [100 points]**

**Part 1: Dataset Exploration [50 points]**

**Consider the hurricane dataset at**

**https://web.archive.org/web/20080921102626/http://www.aoml.noaa.gov/hrd/hurdat/ushurrlist18512007.txt**

**It has been converted to a .csv file (using '|' as the field delimiter).**

**Explore the file using python. Answer the following:**

**a) Show appropriate summary statistics and visualizations for:**

**Months, Highest\_Category, Central\_Pressure\_mb, Max\_Winds\_kt**

**b) Show the relationship between**

**Highest\_Category and Max\_Winds\_kt**

**Central\_Pressure\_mb and Max\_Winds\_kt**

**c) Explain how you accounted for any missing data,**

**and how that may have affected your results**

**Part 2: Web/HTML Screen Scraping [50 points]**

**Go to the Top500 SuperComputer Web site at https://www.top500.org/list/2018/06/**

**View the page source and inspect the HTML table code for the entries.**

**Note the fields in the table: Rank, System, Cores, RMax, RPeak, Power.**

**a) Using Python with BeautifulSoup, extract the data from the Web site**

**and create a csv dataset containing that data.**

**b) Clean & explore the dataset, producing summary statistics**

**for Cores, RMax, RPeak, and Power.**

**c) Display and explain the relationship between**

**Cores and RPeak**

**Cores and Power**

**d) [Challenge problem, bonus 10 extra points]:**

**Display summary statistics and visualizations for the Country data.**

**Resources:**

**https://www.crummy.com/software/BeautifulSoup/**

**https://pythonprogramminglanguage.com/web-scraping-with-pandas-and-beautifulsoup/**

**https://towardsdatascience.com/the-tale-of-missing-values-in-python-c96beb0e8a9d**

**https://chrisalbon.com/python/data\_wrangling/pandas\_missing\_data/**