**Capstone Project**

1. **Data Wrangling**

1. **Primary Exploratory Data Analysis**
   1. **Dataset**

I have chosen to work on an interesting dataset that I found on the NASA’s website (<https://data.nasa.gov/Space-Science/Meteorite-Landings/gh4g-9sfh>), encompassing the complete list of around 45000 meteorite landings on earth since the year 601. Each meteorite is further described by the 10 columns listed below:

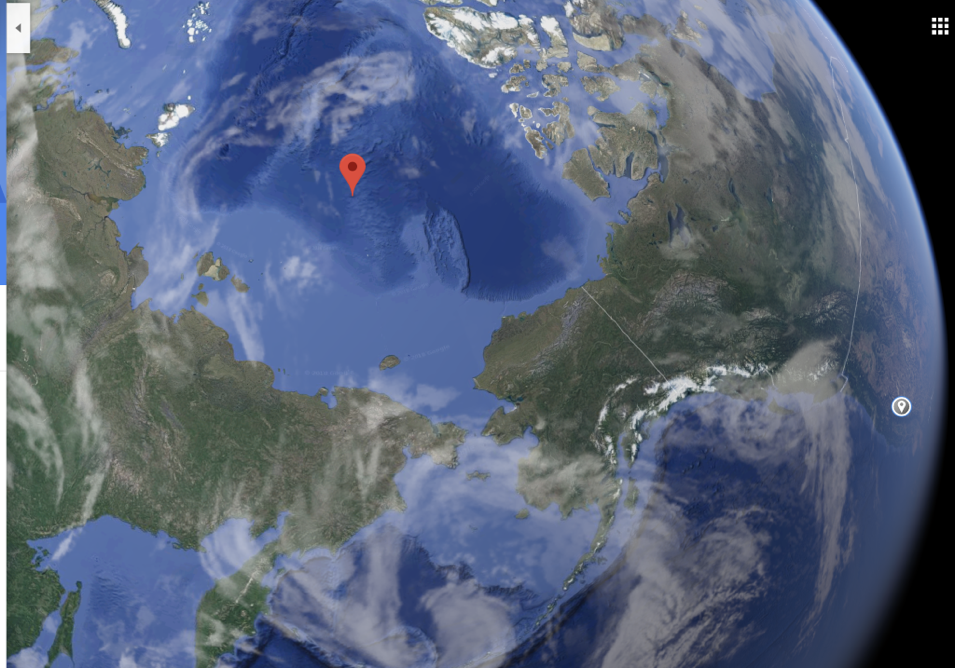
* **name**: the name of the meteorite (typically a location, often modified with a number, year, composition, etc)
* **id**: a unique identifier for the meteorite
* **nametype**: one of: -- *valid*: a typical meteorite -- *relict*: a meteorite that has been highly degraded by weather on Earth
* **recclass**: the class of the meteorite; one of a large number of classes based on physical, chemical, and other characteristics
* **mass**: the mass of the meteorite, in grams
* **fall**: whether the meteorite was seen falling, or was discovered after its impact; one of: -- *Fell*: the meteorite's fall was observed -- *Found*: the meteorite's fall was not observed
* **year**: the year the meteorite fell, or the year it was found (depending on the value of fell)
* **reclat**: the latitude of the meteorite's landing
* **reclong**: the longitude of the meteorite's landing
* **GeoLocation**: a parentheses-enclose, comma-separated tuple that combines reclat and reclong
  1. **Data Exploration**

            The row dataset is a csv file that I imported using the Jupyter Notebook. The imported file was quite clean, with, however, a lot of missing values (around 17%). The cleaning process is detailed in a separate ipynb file. Below is a preview of the steps that I took to make the data more readable:

* + - Used the head() method to briefly inspect the data
    - Looked for missing values using the isnull() method
    - Checked the unicity of the critical column ID
    - Dropped duplicated rows
    - Checked the percentage, per column, of the missing values
    - Dropped the missing values that I could not retrieve
    - Did some statistical analysis (mean, median, std...) to make sure that the data is accurate
    - Looked for outliers by fact-checking the mass, the name and the geographical locations of certain meteorites
    - Grouped the meteorites by classes

* 1. **Looking for Outliers**

In order to verify the accuracy of the data, I googled the names of random meteorites and checked if their mass or the geographical location of their landing matched the information provided by Wikipedia. For example: I checked the maximum longitude and saw whether the location existed or not. The result seemed logical.

.  

*Screenshot from Google Maps*

                 I also selected the heaviest meteorite, Hoba, and verified all the information related to it, and once again, the results were completely satisfying. I couldn’t, for obvious reasons, fact-check all the 45000 rows but for a primary analysis I think the dataset is for the most part accurate.