Requirements:

1. 1 folder containing only FAA report files
2. 1 folder containing only ASRS report files
3. 4 scripts
   * Historical\_data\_merge\_FAA.py
   * FAA\_data\_processing.py
   * Historical\_data\_merge\_ASRS.py
   * Cross Val\_Similarity.R
4. 2 look up files saved in the same folder as the py scripts. These are used to standardize some of the data in the reports.
   * DATA\_FAA\_aircrafts.xlsx
   * DATA\_standardization\_lookups.xlsx

Steps:

**ASRS file merging and processing**

1. Save any ASRS reports needed for processing
2. Update file path inputs in the Historical\_data\_merge\_ASRS.py to locate the folder containing the ASRS raw reports. Run the script
   1. Output is 2 files.
      1. DATA\_ASRS\_merged.csv is just the raw files combined into as single csv
      2. DATA\_ASRS\_filtered.csv is reports that mention any type of UAS that will be used for cross val
   2. ASRS filtered output will be used for cross val

**FAA file merging and processing**

1. Save any FAA reports needed for processing
2. Update file path inputs in the Historical\_data\_merge\_FAA.py to locate the folder containing raw FAA reports. Run script
   * Output is 1 file
     + DATA\_FAA\_merged.csv that will be used in FAA\_data\_processing.py. It is just the raw FAA reports joined together with some text cleaning done on the reports
3. Update file path in FAA\_data\_processing.py to locate the file out put from step 2. Run script
   * Output is
     + DATA\_FAA\_split.csv - a processed and clean csv that will have extracted data from the raw FAA records to be used for visuals, summary stats, etc…

**Cross Validation for finding actual UAS sightings**

Compares ASRS reports (synopsis and Narrative) against FAA Summaries. It compares reports that are from the same year, month, and state. The comparison is done using Cosine Similarity testing.

1. Update the file path in Cross Val\_Similarity.R to locate the DATA\_FAA\_split.csv and DATA\_ASRS\_filtered.csv. Run script.
   1. Output is 2 files
      1. DATA\_crossVal\_hits.csv = file that returns NA for non-hits and returns data for any FAA record that is considered similar to the ASRS data (hit)
      2. DATA\_FAA\_split\_validated.csv = Extra column added to the output from FAA step 4 that designates a true report.
         1. True = 1
         2. False = 0

**Next steps:**

Use the updated cross val data set, DATA\_FAA\_split\_validated.csv, a to train predictive models.