C.J.’s notes

* ASRP started 4/30/1975, modified in 1983, 1987, 1999 ([section 6](https://asrs.arc.nasa.gov/overview/immunity.html))
* Data are [anonymized](https://asrs.arc.nasa.gov/overview/confidentiality.html), FAA does not use data against reporters in enforcement (except for accidents and criminal offenses – cf. [FAR 91.25](https://asrs.arc.nasa.gov/overview/immunity.html))
* Description of [airspace classification](https://aspm.faa.gov/aspmhelp/index/Airspace_Classification.html)
* [Coding](https://asrs.arc.nasa.gov/docs/dbol/ASRS_CodingTaxonomy.pdf) Taxonomy, may be useful in interpreting database entries
* There will likely be some big effects from covid. (How) do we need to address that?
* Should we be looking at safety in accidents per flight or accidents per hour / mile?
  + I guess we can see what the data have to say here
  + Presumably takeoff and landing are the most dangerous segments of a flight, in which case accidents/flight makes more sense
* Jake’s proposed problems
  + NTSB – in the universe of aviation accidents, what are the predictors of severity?
  + ASRS – determine prominence of various types of incidents over time

5/27 -- Questions we have

* What is the research question?
* What data do we need to withhold for testing?

Meeting 6/10

* To share
  + Large dataset has number of uninjured passengers, so we can determine how many people were onboard
    - But there are a good handful (~2%) where this number is greater than the number of seats, in some cases by a wide margin
  + Is there a way using T-100 queries to sort by flight type? (Re: seasonality of crashes / 100k flights)
  + In ~5/6 fatal flights, everyone onboard died
  + We impute values after train/test split, right?
* From meeting
  + Number of seats issue seems to be resulting from multi-plane crashes; the reported injury counts are aggregated but the seat counts are not
  + When deciding the frequency threshold for dummy variables, do we look at the total number of entries or the number of non-na entries in the column?

Meeting 6/12

* Qs for Francesca (decreasing priority)
  + What models should we use / consider using?
  + What metrics / tests would be useful for feature selection?
  + There’s lots of data missing from July 2022 onward, should we drop this from dataset if we’re not doing a time series?
  + When deciding the frequency threshold for dummy variables, do we look at the total number of entries or the number of entries present in the column?
  + How should we impute longitude / latitude?
* To share
  + Recommendations for handling missing data
    - Restrict to accidents, USA
    - Consider column sparsity on train data only (i.e. pre-covid)
    - Drop rows that are missing values in columns that are ~99% present
  + There are a bunch of rows with injury counts on the event level but not on the aircraft level. The overwhelming majority of these are aircraft #1 so we should be able to calculate the value (only issue is if the counts are unknown for both aircraft in a two-craft-crash)
  + ‘ntsb\_no’ may actually be helpful – the first three letters contain information. E.g. CEN seemingly indicates damage discovered in an inspection
* From meeting
  + How should we handle missing a target variable (in this case, damage)?

Meeting 6/17

* Qs for Francesca
  + Should we stratify the train/test split on damage? Should we produce two separate train/test splits for predicting damage and injuries?
  + Should we clean data (e.g. category merging) before or after train/test split?
  + How should we impute aircraft damage?
* To share
  + Edits to ntsb\_table\_join.ipynb
    - Restricted to accidents in USA before train/test split
  + How much of a pain would it be to use CAROL to get BroadPhaseOfFlight? How much of a pain would it be to incorporate this into ntsb\_table\_join.ipynb?
* From meeting
  + Don’t impute aircraft damage, drop the rows instead when we’re building the model
  + What exactly is the variable for injuries? Seems like we’re using proportions and doing regression
  + Order of operations in train/validation/test split, data cleaning

Meeting 6/19

* To share
  + I did KNN model using just location to predict damage, it did not do very well. I struggled with the code for predicting probabilities rather than a single class; that may give a more interesting picture
  + I’ve started working on the presentation and executive summary so we’re not in too much of a pinch next week if we’re still finishing up the modeling
  + Nice package from most recent problem set
    - from interpret.glassbox import ExplainableBoostingRegressor
  + Separated data cleaning into its own file, cleaned up that notebook a bit
* Qs to address
  + If we know one but not both injury counts in a multi-aircraft crash, should we keep the line for the aircraft whose injury counts we do know?
* From meeting
  + To do to finish data cleaning
    - Drop total\_seats