DRAFT Methodology:

1. Preprocessing:
   * Convert which specific columns to categorical (factor equivalent)? Field Unit, Protected Heritage Area, Incident Type, Species Common Name, Animal Health Status, Cause of Animal Health Status, Animal Behaviour, Reason for Animal Behaviour, Animal Attractant, Deterrents Used, Animal Response to Deterrents, Activity Type (some values are lists…), Response Type (some values are lists…),

Referring to <https://towardsdatascience.com/dealing-with-list-values-in-pandas-dataframes-a177e534f173> for advice...

Does Professor Abdou have advice around how I merged “response type” and “activity type” variables into lists prior to joining datasets? Will working with those variables still be possible, if I want to see all rows that have a “black bear” involved, for example, can I have that include rows that have black bear in the list? Would it be better to split them into columns for “response type1” “response type 2”, etc? if that was the format, would I be able to search for all rows that have “black bear” in any of X columns?

* + Feature scaling/data normalization? What should be done for my categorical variables (issue with some being lists…) – referring to: <https://www.analyticsvidhya.com/blog/2015/11/easy-methods-deal-categorical-variables-predictive-modeling/>
    - In order to use kNNImputer, we have to encode categorical values into numerical values using mapping (<https://www.analyticsvidhya.com/blog/2020/07/knnimputer-a-robust-way-to-impute-missing-values-using-scikit-learn/#:~:text=For%20imputing%20missing%20values%20in,of%20categories%20to%20numeric%20variables>.).
  + How will I handle missing values and or special characters? Some columns can likely be deleted?
    - Variables with only a few (between 3 – 40) missing values: Incident Type, Species common name, Sum of Number of Animals, Activity Type, Latitude Public, Longitude Public, Within Park,
      * Impute this data using k-th Nearest Neighbour Imputation (kNNImputer) (referred to this source to decide: <https://arthurarchiproj.medium.com/classification-missing-data-imputation-2932166e1000>).
    - Variables with around 1500 missing values: Response Type,
      * 1500 missing values is still only about 2% data for this column so I will impute this data (using k-th nearest neighbour imputation) Note that the “Response Type” variable contains some values that are lists (from merging data for Incident ID’s prior to dataset join) which may need to be considered when imputing.
    - Variables with several missing values (over 30,000): Animal Health Status, Cause of Animal Health Status, Animal Behaviour, Reason for Animal Behaviour, Animal Attractant, Deterrents Used, Animal response to Deterrents.
      * There are too many missing values to consider any kind of imputing. But I don’t want to completely lose this data as it would be good to find patterns for when specific animal health status’ occur, for example when an animal is dead. I would like to subset the data (all rows that have complete values for these variables) variables and analyze it separately from the rest of the data (all rows but delete the specified columns with missing values)
  + Data reduction? – use dimensionalty reduction via random forest to select which variables are most important or least important for the dataset. Need to look into this more.

1. Exploratory analysis:
   * Univariate: Use frequency distribution tables/histograms to look at variance within each of the variables. Can look at some summary statistics for each variable (depending on data type, these will vary but some possible are : mean, mode, median, range, variance, maximum, minimum, quartiles, and standard deviation.)
   * Look for outliers in categorical data using the histogram (<https://analyticsindiamag.com/how-to-detect-and-treat-outliers-in-categorical-data/#:~:text=or%20box%20plot.-,Detecting%20outliers%20in%20the%20categorical%20data%20is%20something%20about%20the,the%20bar%20chart%20or%20histogram>.)
   * multivariate – what method? Can I use Kruskal Wallis Test to look at correlation between variables???
   * Look at skewness – what method?
2. Modelling:
   * Decision tree – baseline (able to compare random forest to it).
   * Random forest – (how many trees)
   * Add another model
   * What evaluation method will I use? (Cross-fold, k-nearest, etc).
   * What performance measure will I use primarily and why? What is the difference between them? (recall, precision, accuracy)

As part of methodology, look up to see who else used this dataset before and what techniques and tools did they use. Check the efficiency of their classifiers. Check conclusions/recommendations that Parks Canada may have developed based on this data.

Add details about WHY I am choosing each method I choose.