## Tentative Methodology (Project Approach)

### Figure 1: Graph of overall Methodology

### Data Collection:

* Retrieve the 4 desired datasets from Canada’s Open Government Portal: <https://open.canada.ca/data/en/dataset/cc5ea139-c628-46dc-ac55-a5b3351b7fdf>

### Data Processing:

* Data Cleaning – dealing with typos or mistaken entries. Compare unique values for Activity Type, Response, Type, etc, with the Data Dictionary provided with the Open Data record to ensure valid entries.
* Deal with duplicate Incident Number observations in each of the 4 datasets.
* Join 4 datasets into single dataframe - joining on the Unique Incident Number (include details about what was needed to join the data)

### Exploratory Data Analysis:

* Generate an Exploratory Data Analysis Report using the Panda's ProfileReport. Use report to observe variance within variables, correlation between variables, missing values, distinct values per variable, etc.
* Use histograms and various other plots to subset data and explore patterns in incident types by human activity, animals involved, location, and time of year.
* Use statsmodels Python module and Time Series Analysis (tsa) to plot and visualize trends over time. <https://towardsdatascience.com/analyzing-time-series-data-in-pandas-be3887fdd621>

### Modeling:

Before deciding on which models to use, there are some preprocessing steps that needed to be looked into and addressed. Because these steps will be dealt with in the creation of the model, I’ve included this information as part of the Modelling phase of the analysis instead of part of the Preprocessing phase.

#### Splitting Data

Splitting data into training and test data for the purpose of predictive modelling. I will use TimeSeriesSplit function of the scikit-learn package to split and cross-validate my training and test data sets. As described by Howell (2023), doing a TimeSeriesSplit “means our test [data] is always in the future compared to the data our model is fitted on.”

#### Dimensionality Reduction

When researching dimensionality reduction techniques looking for the best fit for my dataset containing mixed data (both numeric and categorical data). Mahmood (2021) indicates that “[w]hen dealing with mixed data, FAMD is a recommended way to handle the unnecessary factor and reduce the dimensions of the data”. Mahmood (2021) goes on to describe that Factor Analysis of Mixed Data (FAMD) analyzes the dataset using a combination of Principal Component Analysis (PCA) and Multiple Correspondence Analysis (MCA) techniques to determine which attributes are the most critical components for modeling the data.

#### Dealing with Missing Values

* + 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,
      1. 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>).

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>.).

* + - Variables with around 1500 missing values: Response Type,
      1. 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.
      1. 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)

#### Dealing with Imbalanced Data

My dependent variable, Incident Type, is very imbalanced. Boyle (2019) succinctly addresses the problem with imbalanced classes as “Most machine learning algorithms work best when the number of samples in each class are about equal. This is because most algorithms are designed to maximize accuracy and reduce error.”

In my research into the pros and cons of the options for dealing with imbalanced classes, both under sampling and over sampling options have some downsides. Under sampling means you are not using some of the data from the majority class in order to balance the classes. In this dataset, the majority class contains about 66% percent of the observations, with the next highest class consisting of about 18% of the observations, so there could be a lot of valuable predictive information in the majority class that would be eliminated with under sampling. Over sampling adds data points (either randomly or synthetically generated) to the minority classes to balance the classes. Stewart (2020) points out one of the concerns with using over sampling is the risk of adding false data points could lead to overfitting. In this dataset, we would need to add very large amounts of datapoints to the minority classes in order for them to balance with the majority. That many “false” data points would cause concern that the data is no longer providing an accurate depiction of the actual data.

I also researched a third method of dealing with imbalanced the cost-sensitive learning method. Brownlee (2020) describes that “while most machine learning algorithms assume that all classes have an equal number of examples… [cost-sensitive learning] modifie[s] [algorithms] to change the way learning is performed to bias towards those classes that have fewer examples in the training dataset.” Stewart (2020), Brownlee (2020), and Selvan (2020) all claim that using cost-sensitive learning to deal with imbalanced datasets - where class weights are taken into account in the machine learning algorithm – perform more effectively than under or over sampling. I’ve decided to use this method to address the imbalance in my Incident Type variable.

#### Dealing with Categorical Variables

Most decision tree classifiers (including the commonly used scikit-learn decision tree classifier and random forest classifier) cannot handle categorical data without having it encoded to numeric form. When the categorical data is ordinal, this process is quite straight forward; however, with nominal data (where there is no inherit order), the process of converting categories to numeric involves creating “a new dummy variable for each level of the original variable” (Dingwall and Potts, n.d.). There are two main methods for encoding nominal categorical data: one-hot encoding and get\_dummies. Kumar (2021) explains that the get\_dummies method can be problematic if your test data set has a different set of columns depending on whether or not some categories exist in the test set that do or do not exist in the training set. The one-hot encoding method does not have this issue. Dingwall and Potts (n.d.) describe the following two potential problems that one-hot encoding can have in tree-based models:

1. The resulting sparsity virtually ensures that continuous variables are assigned higher feature importance.
2. A single level of a categorical variable must meet a very high bar in order to be selected for splitting early in the tree building. This can degrade predictive performance.

Even with those stated potential problems with using one-hot encoding, it is still a very popular and commonly used method. The alternative to encoding categorical data is of course leaving your categorical data as-is and using a tool like H2O for Python which can build a decision tree or random forest model using categorical data.

Based on my findings, I’ve decided to use one-hot-encoding on two of the three models I am generating using scikit-learn, and will maintain the categorical data for the last model using H2O. More specifics about these three models is included in the section below.

#### Models

I will be building three different models to compare the results. All three models will be using tree classifiers, but will have differences in the type of model or treatment of categorical variables. They are summarized below:

* + - 1. Decision tree classifier: using scikit-learn in Python and Gini index as the selection measure. For this model, one hot encoding will be applied to all categorical variables and I will use the sklearn.utils.class\_weight function to balance the class weights (how: Stewart (2020) and Brownlee (2020)).
      2. Random Forest Classifier using scikit-learn: For this model, one hot encoding will be applied to all categorical variables and I will use sklearn.utils.class\_weight function to balance the class weights. (How: Brownlee (2020))
      3. Random Forest Classifier using H20: For this model, the categorical variables will remain as they are (with string values) and I will use the balance\_classes option in H20 to balance the class weights. (how: <https://docs.h2o.ai/h2o/latest-stable/h2o-docs/data-science/drf.html>).

#### Evaluation

Evaluate and compare the results of each model. The common evaluation metrics are accuracy, precision, recall, and F1 score. Accuracy is often not the best metric to use when evaluating imbalanced datasets. Precision is a good metric when your main aim is to minimize false positives and recall is a good metric when you want to maximize the true positive rate. In this situation, I see it as more important to maximize the true positive rate (successfully predict incident types) than it is to minimize false positive rates so while I will be considering all the performance metrics in my evaluation, I will use recall as my main performance metric.

### Visualization And Reporting

I will present the results of the project using various plots and graphs. Based on the research I’ve conducted, I will make recommendations for the mitigation of incident by type with intention of promoting health and safety of humans and animals in these 35 National Parks.

## Literature Review

What do I already know about the topic?

Human and wildlife coexistence is

“Parks Canada’s management efforts concentrate on improving public awareness and minimizing human-wildlife conflict, improving habitat quality, security, and connectivity where possible, and minimizing road and rail mortality” (Ellis, 2022).

“The chance to view wildlife draws millions of visitors each year to the national parks of North America. The combination of a large number of people and abundant wildlife leads to a variety of wildlife-human interactions” (Bath & Ench, 2003).

“We argue that the bridge between HWC research and the implementation of large-scale human–wildlife coexistence is good planning” (Marchini et al., 2019).

Marchini et al. (2019) proposes using strategic planning with data science and modelling to inform a framework for policy makers and citizens regarding the “conflict-to-coexistence continuum”. They propose the planning process begins with a situation assessment before proceeding to decision-making, implementation, monitoring and evaluation.

The data collected by Parks Canada regarding reported incidents of human and wildlife coexistence (HWC) in 35 national parks is the “situation assessment” phase. It is important to note that these incidents are relating to coexistence, whereas many research articles I’ve found relating to this topic refer to HWC as Human and Wildlife Conflict. Many of the coexistence incident types in my dataset could also be considered a conflict incident, for instance, “XXXXXXX”; however there are also several incidents included that would not consistent a conflict, for instance “XXXX” (animal sightings).

“Human–wildlife conflict (HWC) is a key topic in conservation and agricultural research. Decision makers need evidence-based information to design sustainable management plans and policy instruments” (König et al., 2020).

“Coexistence is defined as a dynamic but sustainable state in which humans and wildlife coadapt to living in shared landscapes, where human interactions with wildlife are governed by effective institutions that ensure long-term wildlife population persistence, social legitimacy, and tolerable levels of risk” (König et al., 2020).

“Human-wildlife conflicts is defined as interactions between wildlife humans with a negative outcome” (König et al., 2020).

Madden (2010) offers another description of human-wildlife conflict as follows:

“Human-wildlife conflict occurs when the needs and behavior of wildlife impact negatively on the goals of humans or when the goals of humans negatively impact the needs of wildlife. These conflicts may result when wildlife damage crops, injure or kill domestic animals, threaten or kill people”

“Human–wildlife conflict is increasing in both frequency and severity worldwide and will likely continue to escalate. Protected areas are increasingly becoming islands of habitat surrounded by seas of cultivation and development. Wildlife and humans increasingly compete for space, resources, and places to call home. Although ecosystem-based approaches (including the development of corridors between protected areas) offer improved long-term protection for many species from a biological perspective, they also involve extensive regional opportunities for interaction and conflict between local people and wildlife. Without properly addressing HWC in the effort to conserve wildlife and their habitat, conservation efforts will lose stability and progress, as well as the support of local communities.” (Madden, 2010)

“Human–wildlife conflict, as we understand it today, is not always inevitable and has not been the norm in all cultures and communities. In some communities and cultures, evidence of human–wildlife co-evolution and cultural tolerance to wildlife may offer clues as to how coexistence can be achieved elsewhere.” (Madden, 2010)

What do I have to say critically about what is already known?

Xxx

Has anyone else ever done anything exactly the same?

**Email from David Gummer, Wildlife Management Specialist, Natural Resource Management Branch  
Parks Canada / Government of Canada:**

“On your second question, the main way that these data have been reported and currently affect policies, programs and operations is that we have built a live internal dashboard using PowerBI so that staff and management of the sites can refer to, summarize, analyze and export the (dynamic) incident data as needed to support day to day decisions for their operations (delivery of human-wildlife coexistence operational programs – teams that respond to HWC incidents and take management actions to avoid/reduce future conflict incidents). We are interested and pursuing the possibility of providing a similar interactive data tool that could be accessed externally or even promoted as a public tool, but have not gotten there yet—so in the meantime our effort was to at least publish the same underlying data for others to use until we have a better way to share a more efficient/functional interactive. I don’t know how to share a functional version of our dashboard with you currently, but if you would like more info on this, please let me know. Perhaps we could come-up with a good way to share some views with you.

These data and our ongoing analyses are also helping to inform new national policy/guidance that we are working on and many more that we propose for future, however there is not yet already one good stand alone example that I can send your way. This is why that sentence you quoted is written more generally and in the present tense. In addition, individual parks and sites may use their subsets of the data in their own local reports, policies or protocols/procedures. I do not have a ready example at hand for those either, but if there is a certain park, region or topic you are interested in, that could help me investigate if there are relevant examples or who might be a good contact I could refer you to.”

(D. Gummer, personal communication, February 3, 2023)

I have not been able to find any other usage of this exact data online.

Has anyone else done anything that is related?

Baral et al. (2021)

**What?:** “With the intent to better management human wildlife conflict (HWC) and wildlife conservation in mid-hills outside protected areas of Gandaki province, Nepal, we analyzed the patterns and drivers of HWC. Using data collected from literature, government records and questionnaire survey, we investigated temporal, seasonal and spatial distribution of human casualties caused by wildlife attacks.”

**Data Analysis Methods**:

“We conducted multivariate logistic regression with the entire independent variables [total of 11 variables: temporal (year, month, season, time), socioeconomic (gender, occupation and age group of the victims), and spatial (district, elevation, distance between point of attack and forest), and land use (forest, agricultural land, road and settlement)] in the model to understand the relationship between predictive and explanatory variable (human death and injury due to wildlife attack).”

“We conducted two different statistical approaches following Naha et al. [[52](https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0260307#pone.0260307.ref052)]. Chi-square test of independence was used to understand the association between temporal (year, month, season and time), and socio-demographic variables (gender, age and occupation) with the wildlife attacks. We classified victims into four age groups, < 20, 21–40, 41–60 and > 61 years. The association between socio-demographic and temporal variables and the attacks were analyzed using Pearson chi-square test ([Table 1](https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0260307#pone-0260307-t001)). For the spatial dataset, we conducted a generalized linear model with binomial distribution to predict the effect of variables on the wildlife attacks following Acharya et al. [[6](https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0260307#pone.0260307.ref006)]. We used a priori candidate model and ranked them based on Akaike Information Criterion (AIC) values. Those models with lowest AIC values were considered the appropriate for explaining the wildlife attacks ([Table 2](https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0260307#pone-0260307-t002)). A model with a ΔAIC (the difference between the two AIC values being compared) of ≤ 2 is considered significantly better than the model it is being compared to. The number of attacks per year was summarized in terms of mean (M) and number of events with regard to different spatial, temporal and socio economic variables were analyzed by using percentage. The variability was recorded in terms of SE and at confidence interval of 95%. Statistical analyses were undertaken in R statistical software (R Core Team 2016; Version 1.0.44).”

Naha et al. (2020)

“Conflict with humans is a signifcant source of mortality for large carnivores globally.With rapid loss of forest cover and anthropogenic impacts on their habitats, large carnivores are forced to occupy multiuse landscapes outside protected areas.We investigated 857 attacks on livestock in eastern Himalaya and 375 attacks in western Himalaya by leopards between 2015 and 2018. Multivariate analyses were conducted to identify the landscape features which increased the probability of livestock depredation by leopards.The risk of a leopard killing livestock increased within a heterogeneous landscape matrix comprising of both closed and open habitats (very dense forests, moderate dense forests, open forests, scrubland and non-forests).We used the results to map potential human–leopard confict hotspots across parts of the Indian Himalayan region. Our spatial risk maps indicate pockets in the eastern, central and western part of eastern Himalaya and the central, northern part of western Himalaya as hotspots of human–leopard conficts. Most of the attacks occurred when livestock were grazing freely within multi-use areas without supervision of a herder. Our results suggest that awareness about high risk areas, supervised grazing, and removing vegetation cover around human settlements should be initiated to reduce predation by leopards.”

Data analysis:

“We were interested in examining broader seasonal patterns of depredation (summer, monsoon and winter) and not just for individual months, hence each year was divided into 3 seasons of 4 months each (winter—November–February, summer i.e. February–June, monsoon i.e. June–November). We examined seasonal and temporal variation of attacks and diference in habitat types within the vicinity of predation sites using the chi-square test in R version 3.4.0. Statistical signifcance was p≤0.05 for all analyses. All spatial analyses were performed with Arc GIS 10.3.3 and R 3.4.0.”

“To model the spatial spread and extent of livestock depredation we prepared both binary (presence-1 cells with at least one or multiple attacks, absence-0-no attacks) and count data (presence—exact number of attacks recorded and absence 0-no attacks).”

“Data preparation for spatial risk analysis. We **identified a total of 5 major landscape features** (Habitat, Water, Human presence and infrastructure, Distance to Protected Reserves and Altitude) for North Bengal and Pauri Garhwal (Table 3) based on their ecological importance to model predation risk.”

“We used 4 analytical approaches to model probability of livestock depredation by leopard. In the 1st step we evaluated spatial autocorrelation among livestock kills within the cells using function moran.test (Moran’s I) in package (spdep)75 in R 3.4.0. In the 2nd and 3rd approach we used generalized linear models (GLMs) with binomial and poisson structures to quantify the efect of landscape features (area of habitat types, availability of water, human presence and infrastructure, distance to protected areas and altitude) on livestock predation. All the predictor variables considered for the analysis were continuous in nature. We used a priori candidate models and ranked them based on AIC, AICc values76. Models with the lowest AICc values were considered the best or dominant model and the output (coefcients and estimates) explained the probability of livestock predation by leopards within IHR. Based on the results of the dominant model or the model averaged coefcients, we generated confict hotspots for both the study sites. We used coefcients of the best model (binomial structure) or averaged all candidate models (GLM with binomial structure) to estimate probability of livestock depredation for each cell (25 km2 ) using the equation p (x)=exp (z)/ (1+exp (z)) and generated human–leopard confict hotspots in Arc GIS 10.377. We generated ROC curve and AUC values to predict reliability of the dominant models using package ROCR78 in R 3.4.0. Since predictor variables between the two study sites were not normally distributed, we compared the identical landscape features using nonparametric Wilcoxon Signed-Rank Test in R 3.4.0. In the 4th step, we used the predictor variables of the dominant models to calculate conditional inference (CTREEs), as prescribed in the R-package “partykit”79. Tis method was adopted to obtain threshold values for the signifcant variables for confict mitigation recommendations. Trees based on maximally selected rank statistics were ftted using the Bonferroni correction for multiple testing and a minimum sum of weights. In addition, univariate trees were ftted for variables with a signifcant split in the multivariate tree. Te results of our two analytical approaches (regression and ctree) are similar and provide an overall understanding of landscape features prone to livestock predation in accordance with the behavior of common leopards. Both analytical methods are based on a maximum likelihood approach and when interpreted together provide meaningful results. Te GLM models computes probabilities of an event based on a logistic regression framework while the CTREE uses a machine learning classifcation approach and assigns values to predefned categories. Te decision tree approach is a non-parametric approach which helps simplify complex relationships between dependent and predictor variables.”

Where does my work fit in with what has gone before?

Xxx

Why is my research worth doing in the light of what has already been done?

Xxx

## Data Description

Once I’ve corrected dataset (removed duplicate incident type column and have activity type and response type encoded), run descripting statistics and include a graph or visual of that (dataframe.describe(include=’all’))