What to expect in the face of wildlife?

Analysis of Human-wildlife coexistence incidents across Canada

Image credit from https://www.rmotoday.com/banff/banff-residents-struggling-to-maintain-quality-of-life-with-growing-tourism-economy-1574334

Canada is a country known for its diverse and abundant wildlife. The country's vast landscapes, ranging from the Arctic tundra to the Pacific rainforest, provide habitats for many species of animals, including threatened and endangered species. However, as human populations grow and expand, human activities often come into conflict with wildlife. It is therefore essential to consider ways in which humans and wildlife can coexist in a mutually beneficial way.

But before that, we need to understand what is the magnitude of human and wildlife incidents across Canada using data. To this effort, I have collected the Open Data Record that documents human-wildlife coexistence incidents and response actions by Parks Canada Agency, totalling 36525 incidents from 2017 to 2021. A human-wildlife coexistence (HWC) "incident" is any potential conflict situation between people and wildlife that was assigned to Parks Canada staff to manage to help ensure the safety and wellbeing of people and wildlife.

The whole dataset is comprised of four separate CSVs , including pca-human-wildlife-coexistence-incidents-detailed-records (*incident* dataset), pca-human-wildlife-coexistence-responses-detailed-records (*response* dataset), pca-human-wildlife-coexistence-animals-involved-detailed-records (*animal* dataset), and pca-human-wildlife-coexistence-activities-detailed-records (*activity* dataset). Depending on the analysis, I joined different combinations of CSVs together. The data is collected and used under the Open Government Licence—Canada, allowing me as a University student to copy, modify, publish, translate, adapt, distribute or otherwise use the Information in any medium, mode or format for any lawful purpose.

I have split the analysis into sections, covering Google Maps distribution, visualization, time series analysis, Statistical analysis for Activity and Incident independence, and a Machine Learning model to predict hours required for incidents.

1. Google Maps distribution: Geographic distribution for different incidents

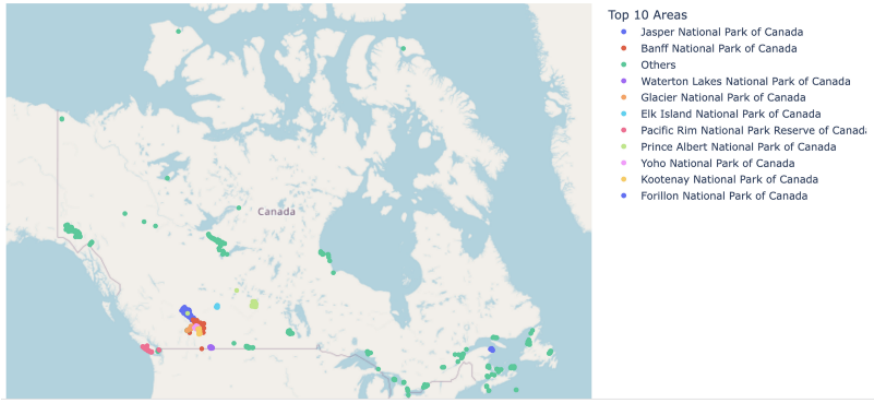
2. Visualization: Incidents vs. Average Staff Hours / Staff total required, Incidents vs. Protected Heritage Areas, Incidents vs. Animal Species, Incidents vs. Human Activity, Incidents vs. Time / Season.
3. Statistical analysis for Activity and Incident independence
4. Machine Learning model to predict hours required for incidents

Geographic distribution for different incidents

I took the *incident* dataset and filtered it down to the top 10 Protected Heritage Areas that have the most incidents. The rest areas are labelled as Others. Although incidents occurred across Canada, it is obvious that almost all top 10 areas are in the East area, with Alberta being the top hit. The reason is also very obvious. Alberta is home to two popular tourist places, Banff National Park and Jasper National Park with a diverse range of wildlife, including large carnivores such as grizzly bears and wolves, as well as herbivores such as elk and deer. As tourism and human populations continue to grow and expand into wildlife habitats, encounters between humans and wildlife are becoming increasingly common, leading to conflicts.

```
incidents_grouped = incidents.groupby(['Protecte
incidents_grouped['Total'] = incidents_grouped.s
# I will only look at the top 10 parks based on
incidents_grouped = incidents_grouped.sort_value
```

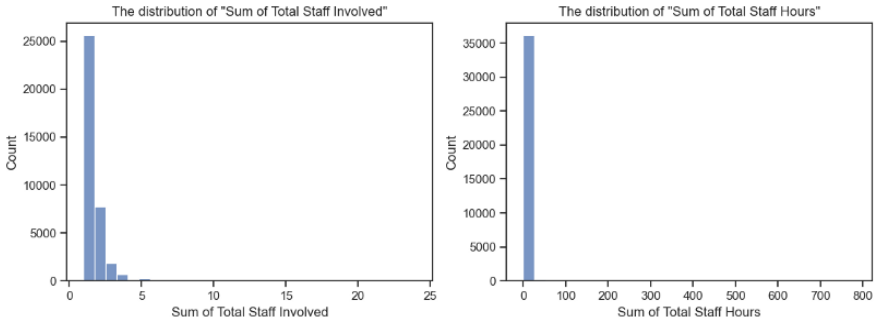
```
top_10_area = incidents_grouped.index.to_numpy()
incidents["Top 10 Areas"] = np.where(incidents['
        incidents['Protected Heritage Area'], "
fig = px.scatter_mapbox(incidents, lat="Latitude
                        hover_name="Incident Typ
                        hover_data=["Sum of Tota
                        color = "Top 10 Areas",
fig.update_layout(mapbox_style="open-street-map"
fig.update_layout(margin={"r":0,"t":0,"l":0,"b":
fig.show()
```



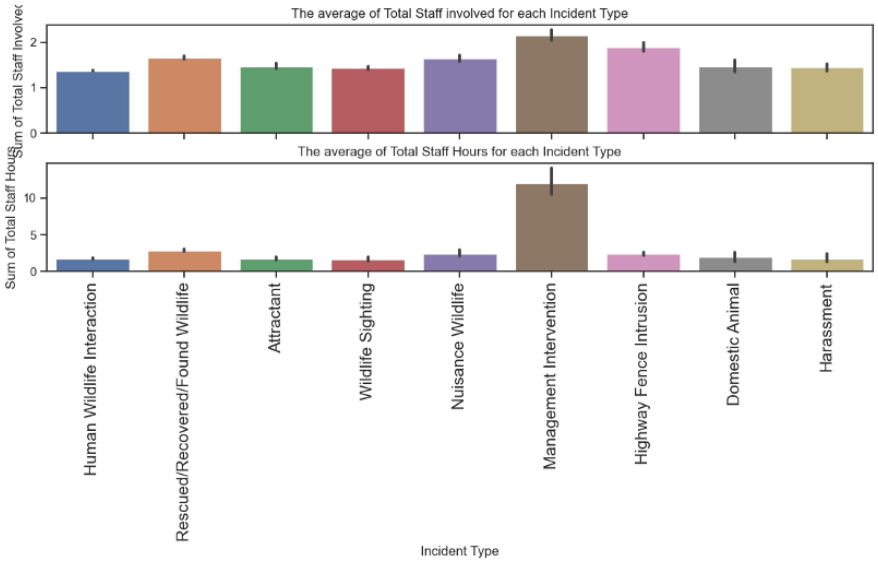
Visualization: Incidents vs. Average Staff Hours / Staff total required

Then we are curious about on average how many staff and how many hours from staff are required to handle each incident.

We first showed the distribution of Total Staff Involved and Total Staff Hours Required. Both distributions are highly left skewed, meaning most incidents only need 1 or 2 (the majority of them are less than 5) staff to address the issue, and almost exclusively a maximum of only 2 hours is required from staff members for each incident.

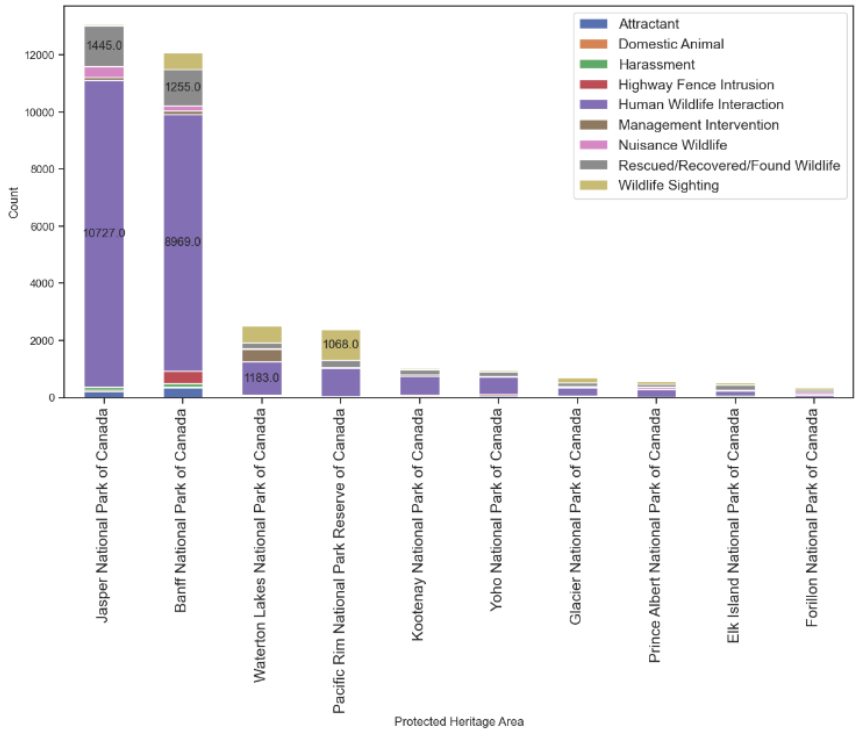


Then we broke down the distribution of Total Staff Involved and Total Staff Hours Required by the same Incident Types. Consistent with the previous distribution plot, most areas need <2 Total Staff Involved and < 5 Total Staff Hours required, except for Management Intervention. According to Parks Canada, Management Intervention refer to management action that is necessary to address multiple incidents or ecological impacts. Normally, it triggers different departments to take action. Note, a 95% confidence interval is added on each graph to show if the difference is significant.



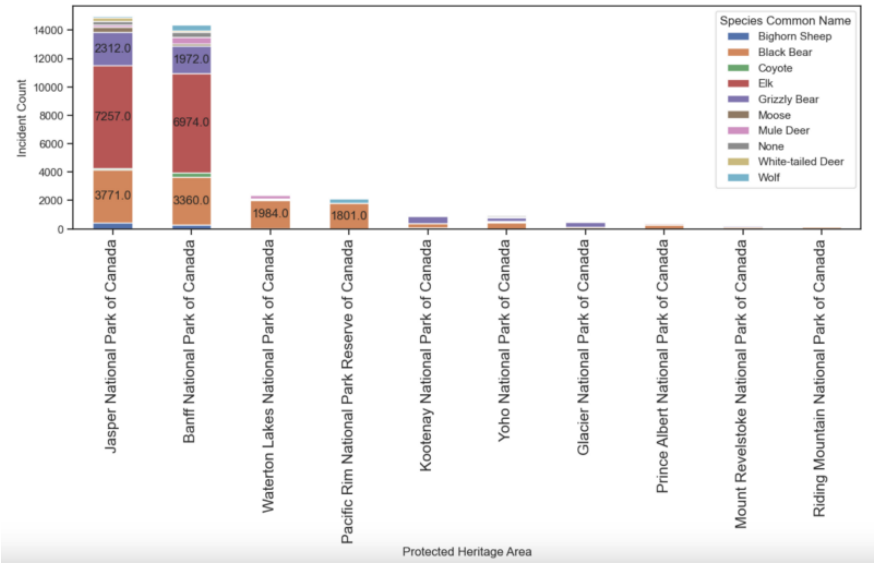
Visualization: Incidents vs. Protected Heritage Areas

From the Geographic distribution, we identified the top 10 Protected Heritage Areas that have the highest incidents count. And we also know different incident types vary from the last visualization plot. In this section, we broke down the top 10 areas by incident types. *Jasper National Park* and *Banff National Park* are two dominant places in terms of incident count. Within them, the Human-Wildlife Interaction type prevails. According to Parks Canada, Human-Wildlife Interaction is defined as a negative interaction between wildlife and people and/or their property; whether major or minor; with or without physical contact

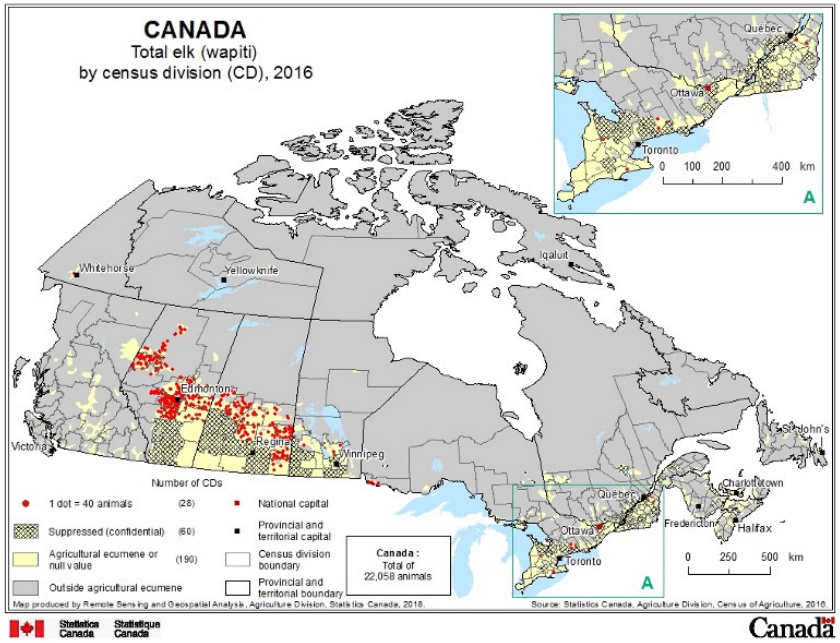


Visualization: Incidents vs. Animal Species

We also want to know what animal species are most common to come across during human activity. Using the same top 10 Protected Heritage Areas that have the highest incidents count, this time we break down the incident counts by Animal Species. There are small diversities, like Jasper and Banff national parks have 3 dominant species, Elk, Blake Bear, and Grizzly Bear in descending order; while Waterton Lakes and Pacific Rim national parks have 1 dominant species Black Bear.

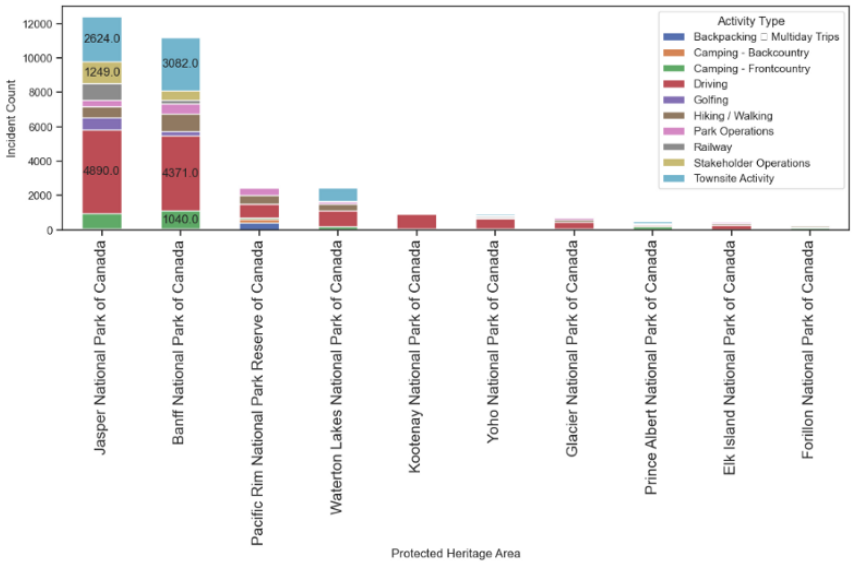


The reason for the high Elk occurrence in Jasper and Banff national park is because of its geographic distribution. A total Elk census in 2016 from the Census of Agriculture Canada shows that Elk are most found in middle Alberta and lower Saskatchewan.



Visualization: Incidents vs. Human Activity

Another interesting view is to look at what Human Activity is more likely to have incidents. We break down incident counts by Human Activity Type. Not too surprisingly, *Driving*, *Townsite activity*, and *Camping (Frontcountry)* are 3 major ones.



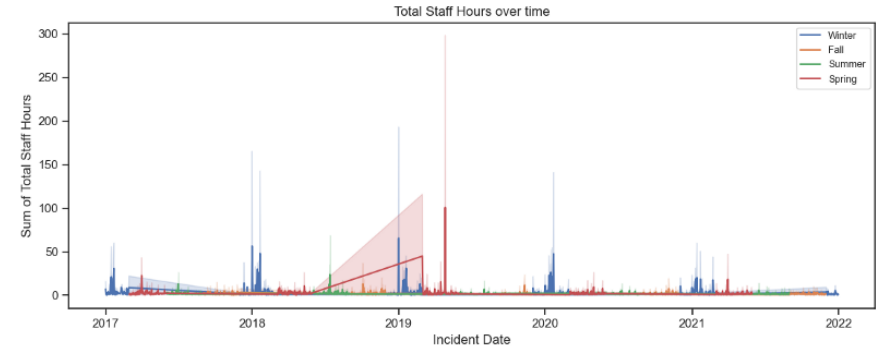
Visualization: Incidents vs. Time / Season

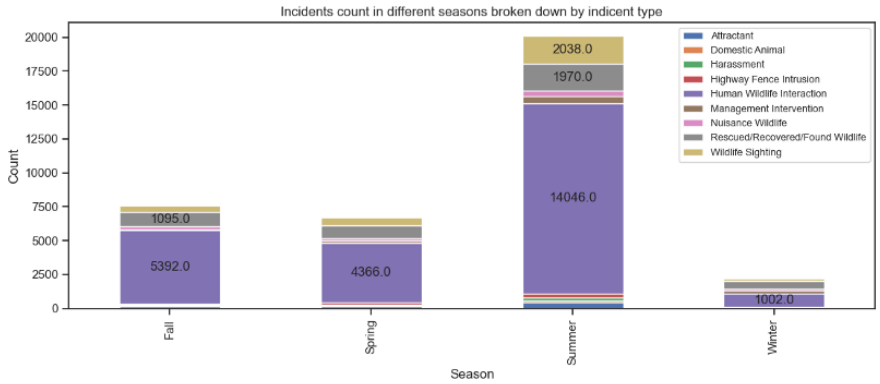
Different times and seasons are known to play a role in incidents. The incidents can come from animals' seasonal activity and also humans'. To do this, I extracted the month information from the raw time format, then mapped that into seasons based on: Spring begins on March 1, Summer on June 1, Autumn on September 1, and winter on December 1. Some very interesting findings are:

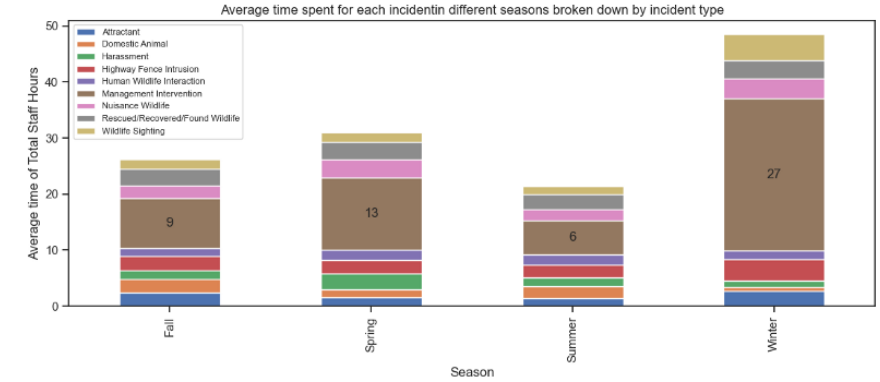
1. `Total Staff Hours over time` shows that there are consistently huge spikes in the winter season, with an exception in the 2019 Spring. This is counter-intuitive to our impression that Summer is normally busier.

2. The accompanied second plot `Incidents count in different seasons broken down by incident type` is telling us a different story where Summer is clearly having more incidents (specifically Human-Wildlife Interaction) than other seasons. This spurs an interesting question that does higher incident counts always lead to higher total staff hours? Well, the total staff hours is a sure thing, but not the average time for each incident.

3. The third plot `Average time spent for each incident in different seasons broken down by incident type` verifies our suspect. As you can see, although there are more incident counts in the Summer, the average time required for each incident is much higher in Winter. This can be argued by the fact that Winter has a more harsh environment to get things around.


Total Staff Hours over time
Sum of Total Staff Hours
Incident Date
Winter
Fall
Summer
Spring
2017
2018
2019
2020
2021
2022


Incidents count in different seasons broken down by indicent type
Count
Season
Attractant
Domestic Animal
Harassment
Highway Fence Intrusion
Human Wildlife Interaction
Management Intervention
Nuisance Wildlife
Rescued/Recovered/Found Wildlife
Wildlife Sighting
1095.0
5392.0
4366.0
2038.0
1970.0
14046.0
1002.0
Fall
Spring
Summer
Winter


Average time spent for each incidentin different seasons broken down by incident type
Average time of Total Staff Hours
Season
Attractant
Domestic Animal
Harassment
Highway Fence Intrusion
Human Wildlife Interaction
Management Intervention
Nuisance Wildlife
Rescued/Recovered/Found Wildlife
Wildlife Sighting
9
13
6
27
Fall
Spring
Summer
Winter

Statistical analysis for Activity and Incident independence

Based on the visualization for `Incidents vs. Human Activity`, we do see incidents occurrence among different human activities are different. As an experimental analysis, we want to use statistical analysis to show the proof. To do this, I first created the contingency table between Incident Type and Human Activity Type. The first plot is showing a sample of all combinations. Then we make the hypothesis for Chi-square test that:

Null hypothesis: Incident Type and Human Activity Type are independent of each other.

Alternative Hypothesis: Incident Type and Human Activity Type are not independent of each other.

	Camping - Huts and Lodges	Camping - Winter Frontcountry
Attractant	2	1
Domestic Animal	0	0
Harassment	2	0
Highway Fence Intrusion	0	0
Human Wildlife Interaction	123	6
Management Intervention	2	1
Nuisance Wildlife	16	2
Rescued/Recovered/Found Wildlife	13	2
Wildlife Sighting	6	1

After we ran the Chi square test, we obtained Pvalue < 0.05. We reject the null hypothesis and accept the alternative hypothesis, meaning Incident Type and Human Activity Type are not independent of each other. For example, there is a higher chance for Driving to have more incidents than other activities.

```
        Pearson's Chi-squared test

data:  table(Incident_Activity$Incident.Type, Incident_Activity$Activity.Type)
X-squared = 11116, df = 712, p-value < 2.2e-16
```

Machine Learning model to predict hours required for incidents

In the final chapter of this article, I used 2 machine learning models to do the prediction: **Tree decision regression** and **Random forest regression**. The target is `Sum of Total Staff Hour`, and a list of variables is used in the model:

1. 'Incident Type' (categorical variable)
2. 'Season' (categorical variable)
3. 'Species Common Name' (categorical variable)
4. 'Animal Behaviour' (categorical variable)
5. 'Sum of Number of Animals' (numerical variable)
6. 'Activity Type' (categorical variable)

One challenge for modelling is how to encode the categorical variables. To avoid introducing ordering for different levels, I end up using pandas dummy encoder for all categorical variables. Then the dataset is randomly split into 66% as training and 33% as testing. For both models, I used MAE as the training metric simply because it is easy to interpret. The result R square is used to indicate how good the model is.

```
#1. Tree dicision regression
#Step 1: join several tables together based on I
combined_table = pd.merge(incidents[['Incident N
                                     'Sum of Tot
                          animals_involved[['Inci
                                           'Anima
                          on = 'Incident Number')
combined_table = pd.merge(combined_table, activi
combined_table = combined_table.dropna()
#Use one hot encoder (pandas get_dummies can do
encoded_combined_table = pd.get_dummies(combined
                                        'Anima
#Step 2: scale the feature
from sklearn.preprocessing import StandardScaler
scale = StandardScaler()
encoded_combined_table_scaled_X = scale.fit_tran
encoded_combined_table_scaled_X = pd.DataFrame(e
encoded_combined_table_scaled_X.columns = encode


#split the dataset into training dataset and tes
from sklearn.model_selection import train_test_s
from sklearn.tree import DecisionTreeRegressor
from sklearn.metrics import mean_absolute_error
X_train, X_test, y_train, y_test = train_test_sp


#fit the model
model = DecisionTreeRegressor(criterion = 'mae',
model.fit(X_train, y_train)
y_pred = model.predict(X_test)
#This returns the R square of the prediction
```

```
score = model.score(X_test, y_test)
#Calcualte the MAE
mae = mean_absolute_error(y_test, y_pred)
print("The MAE is:" + str(mae) + " for tree dici
print("The R^2 is:" + str(score) + " for tree di

#2. Random forest regression
from sklearn.ensemble import RandomForestRegress
model = RandomForestRegressor(n_estimators = 25,
model.fit(X_train, y_train)
y_pred = model.predict(X_test)
#This returns the R square of the prediction
score = model.score(X_test, y_test)
#Calcualte the MAE
mae = mean_absolute_error(y_test, y_pred)
print("The MSE is:" + str(mae) + " for tree dici
print("The R^2 is:" + str(score) + " for tree di
```

Both models give relatively high MAE, 1.88 for **Tree decision regression** and 1.89 for **Random forest regression.** Neither models give a high R square value**, 2.7%** for **Tree decision regression** and 8.25% for **Random forest regression.** It is not surprising that **Random forest regression** is doing slightly better given it has a number of small **Tree decision regressions** running under the hood.

Conclusion

Circling back to the start of this article, we covered Google Maps distribution, visualization, time series analysis, Statistical analysis and a Machine Learning model. Like many other analyses, there

there are still plenty more things we could study in the dataset. For example, normally a certain deterrent is used in the face of animals, so what deterrents are most effective to expel different wide animals? Also, some animals were killed by collisions with vehicles or trains, so what are the statistics about that? Can we do better in terms of management to avoid this tragedy? Last but not the least, there are a range of other machine learning models worth trying if we want to get a better R square. Some feature engineering work would be very beneficial as well.

GitHub

The comprehensive analysis and scripts are hosted in Human-Wildlife-Coexistence-Incidents-Canada. The csvs of the raw data are retrieved from Human-wildlife coexistence incidents managed by Parks Canada ranging from 2017 to 2021.