University of Strathclyde Glasgow

DEPARTMENT OF

COMPUTER & INFORMATION SCIENCES

North American Bear Attacks: An Analytical Approach

by

202490978

202476189

202461546

MSc Data Analytics

Contents

List of Figures

List of Tables

1 Introduction

Rapid technological advancements in our society are contributing to an increase in the production and collection of significantly vast arrays of data and information, popularly referred to as 'big data'.

While data in this manner has existed for decades, including within the internet and the Global Position System (GPS), expanded analytical capacities have allowed us to transition into a new era of big data, where this is a non-negotiable inclusion in various areas of society (Schintler & McNeely, 2020). Big data is utilised comprehensively within business, scientific inquiry and even the Government (Schintler & McNeely, 2020).

Another discipline where big data has been used, is looking at bear attacks. Big data collection and analysis hold importance within this context for several reasons. While bear attacks are rare, they remain serious public safety concerns, especially with the rise of outdoor recreational activities in the USA and Canada. For example, camping in the USA is at an all-time high, and around 58 million Americans participated in hiking in the year 2021 (Statista, 2023). Furthermore, in 2019, approximately 25 million people visited Canadian National Parks (Statista, 2024). This information highlights how imperative it is to ensure the appropriate measures are in place to protect these people.

Moreover, this analysis can be used for conservation efforts, i.e. to also protect the bear population and to recognise trends that suggest an impact of climate change through the identification of more recent anomalies. If bears' hibernation patterns and/or habitats are altering, we can attribute climate change as a factor and adjust safety and conservation policies according to the projection of its effect in the future.

Examples of how big data has been utilised in this manner can be observed in some of the work done by the Get Bear Smart Society (GBSS), who aim to simultaneously protect bears in their natural habitat, whilst ensuring safety for people in these areas.

An example of some data-driven work within the rationale of bears and humans living in unison is that of Davis et al. (2002). This project collected and analysed vast amounts of data, intending to reduce negative human-black bear interactions in British Columbia. Following this, Davis et al. (2002) was able to make proactive recommendations regarding visitor education, as well as the distribution of natural bear food in areas of high human use, and vice versa, which was found to be influencing the negative interactions between humans and black bears. Ultimately, the safety of all parties was facilitated here using big data.

2 Dataset - Fatal Bear Attacks North America

2.1 Overview

The dataset used in this study by Umair Zia, was sourced from Kaggle. The dataset investigates fatal bear attacks in North America over the past century. It provides crucial information on each attack's date, location, and type of bear involved. This dataset, comprised of 166 rows and 16 columns, offers a valuable basis for analysing bear attack trends across the continent.

By building on localised work, such as the British Columbia-specific study by Davis et al. (2002), this broader dataset enables us to track bear attack trends up to the year 2018. We can explore if and how trends have shifted over time, potentially due to factors like climate change or increased human activity in bear habitats. The findings are expected to inform updated safety measures, benefiting both humans and bears, and aligning with conservation efforts by organisations such as the Get Bear Smart Society (GBSS).

The dataset contains the following key variables:

- Name, Age, Gender: Details on the identity and demographics of each victim involved in bear attacks.
- Date, Month, Year: Temporal information for each attack, enabling analysis over time.
- Type: Identifies whether the bear involved was wild or captive.

- Location, Latitude, Longitude: Geographic information specifying the location of each incident.
- Description: Narrative details for each attack, providing context and additional insights.
- Type of Bear: Identifies the bear species involved, such as black bear, grizzly bear, or polar bear.
- Hunter, Grizzly, Hikers, Only One Killed: Categorical variables indicating the circumstances surrounding each attack, including whether a hunter was involved, if it was a grizzly bear, if hikers were present, or if only one person was killed.

2.2 Data Pre-Processing

2.2.1 Data Cleaning

The data cleaning process involved several key steps to prepare the bears dataset for analysis by addressing issues such as inconsistent formatting, missing values, and irrelevant entries.

First, non-breaking spaces, which are common in text data were removed from all text fields. This ensured that columns like "Location" and "Gender" had a consistent format, free of unnecessary characters. Removing these spaces helps maintain uniformity, making it easier to filter and analyse the data accurately.

Next, since the focus of this study was on bear attacks occurring in natural habitats, the dataset was filtered to include only incidents involving wild bears. This narrowed down the data to attacks occurring in the wild, where conservation and public safety measures are most applicable.

Additionally, missing values were identified in a few columns, with a high concentration in "Latitude" and "Longitude." These missing geographic details could limit the scope for detailed spatial analysis. Minor missing data in the "Age" and "Gender" columns was also noted. In our analysis these rows were dropped where relevant.

2.2.2 Feature Engineering

To enhance the analytical scope of the dataset, several new features were engineered, focusing on temporal and seasonal aspects of bear attacks. These new variables provide additional insights into patterns and trends, facilitating a deeper understanding of bear-human interactions over time and across different regions.

The first new variable, Decade, was created to allow analysis over extended periods. By grouping each bear attack by the decade in which it occurred, this feature enables the identification of long-term trends, such as changes in bear attack frequency over time. Analysing data at the decade level provides a more stable view of bear behaviour and human interactions with bears, highlighting any significant shifts across generations. By carrying out the following line of code the decade could be calculated from the year field within the data:

*bear_data['Decade'] = (bear_data['Year'] // 10) * 10*

Furthermore, the Season variable was introduced to capture seasonal patterns in bear attacks. Each attack was categorised into Winter, Spring, Summer, or Autumn based on the month in which it occurred. This feature is valuable for identifying any seasonal fluctuations in bear behaviour and human activities, such as increased attacks in the summer when outdoor activities are more frequent. Seasonal trends can provide insights into when bear-human encounters are most likely to occur, helping authorities and the public better prepare for potential risks during peak seasons.

Finally, a new data frame was created for the Supervised Analysis part of this report which aims to examine bear attack trends over time. The dataset was grouped by Decade and Season, with the count of bear attacks calculated for each combination. This new feature, No. of Attacks, represents the number of bear attacks by decade and season.

2.3 Exploratory Data Analysis

2.3.1 Age and Gender

Figure 2.1 Boxplot of Bear Attacks by Age of Victim

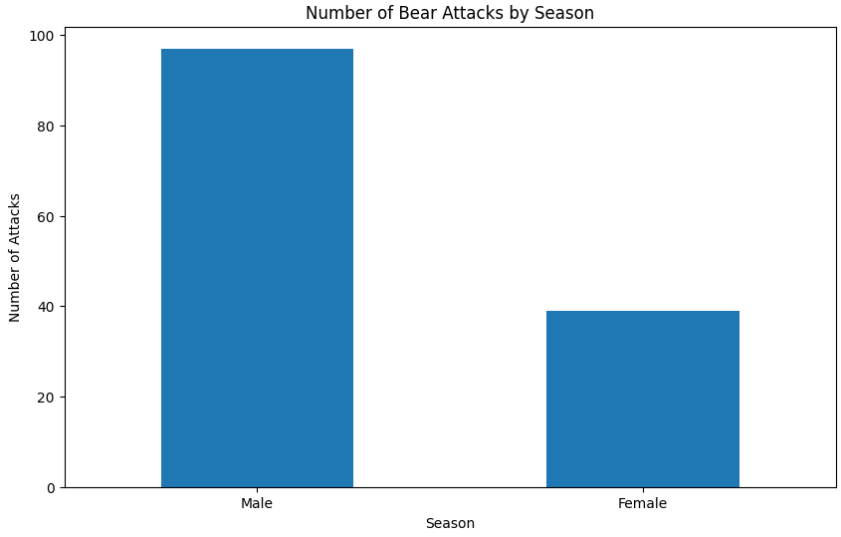

Figure 2.2 Bar Chart of Bear Attacks by Gender of Victim

When studying the demographics affected by bear attacks it tells us that it is predominately middle-aged men. This can simply be down to what individuals put themselves in in these dangerous environments. The USA sees around 60 million hikers per year, reaching over 60

million in 2023 *(Number of hiking participants in the United States from 2010 to 2023, 2024).* The demographics of these hikers show the gender to be predominately male, as our data represents. On top of this, outdoor recreational activities such as hunting and fishing are more dominated by the male gender *(Schmitt, 2013)* and positions these individuals in higher-risk areas.

2.3.2 Decade

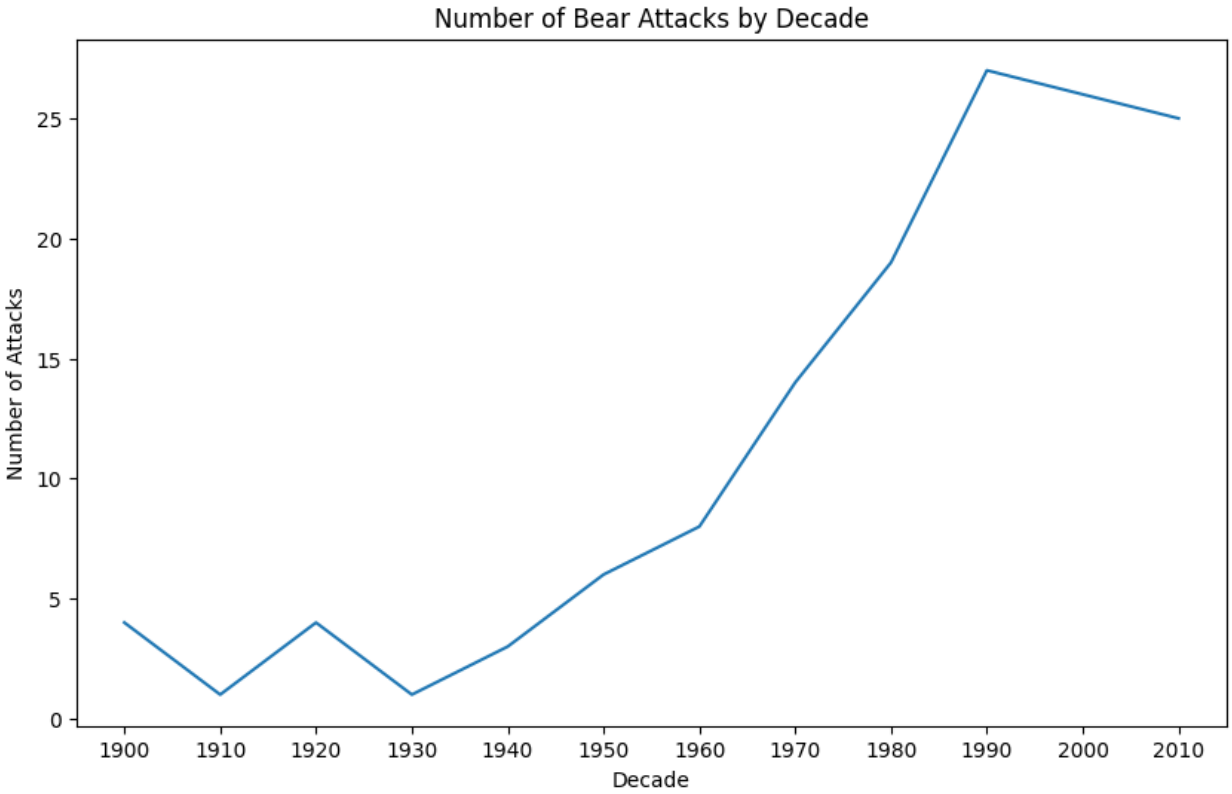

Figure 2.3 Time-Series Graph of Number of Bear Attacks

As previously mentioned, the decade column of the dataset was a result of the feature engineering we completed on the dataset. It acted as a better unit of time to better analyse the number of attacks that have been recorded within the dataset over the past century.

As shown in *Figure 2.3*, we can clearly see the upward trend of bear attacks as time progresses.

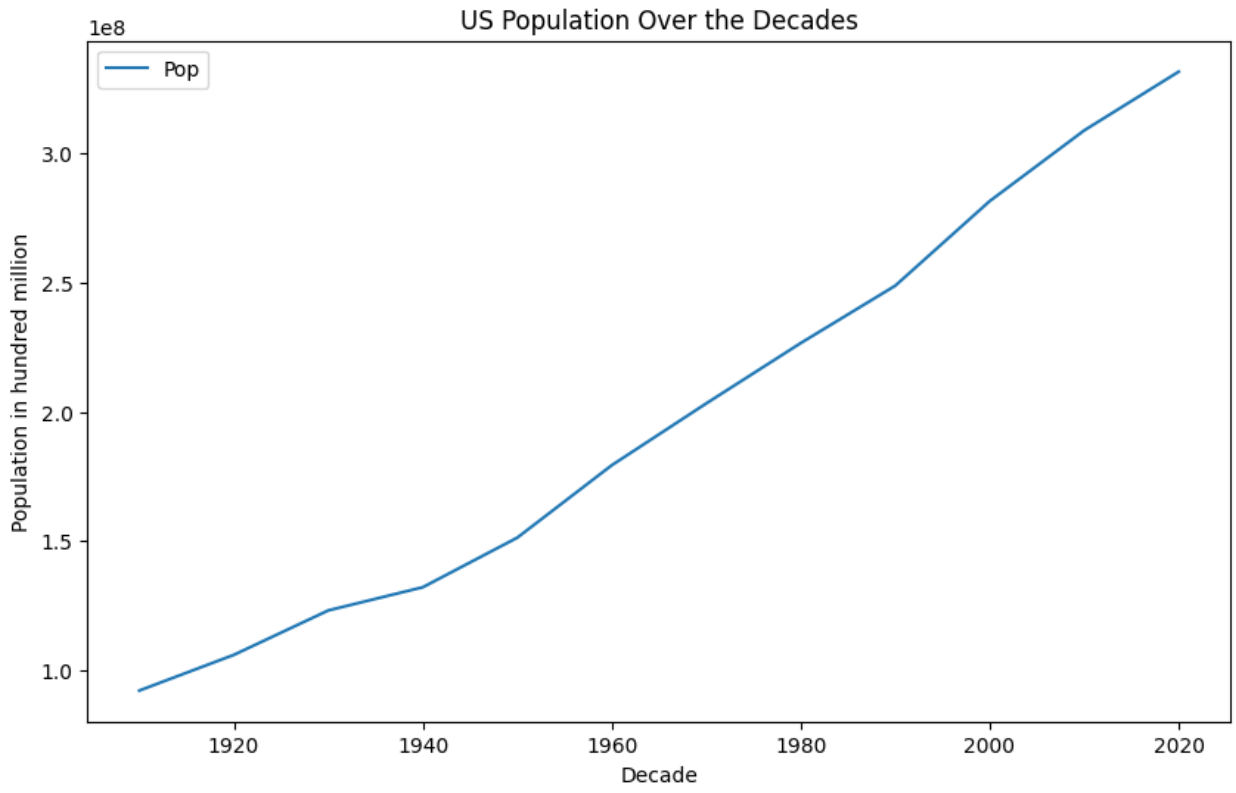

Figure 2.4 Time-Series Graph of US Population

This may be in relation to *Figure 2.4* where the population has been on the rise exponentially. With a rising population, there leads to more people up taking outdoor recreational activities and more threat of bear attacks. Furthermore, a rising population leads to a natural expansion of built-up areas, some of which may conflict with typical bear habitats. This population hike correlates to the trends in bear attacks over the same timeframe, as shown in *Figure 2.3.*

2.3.3 Season

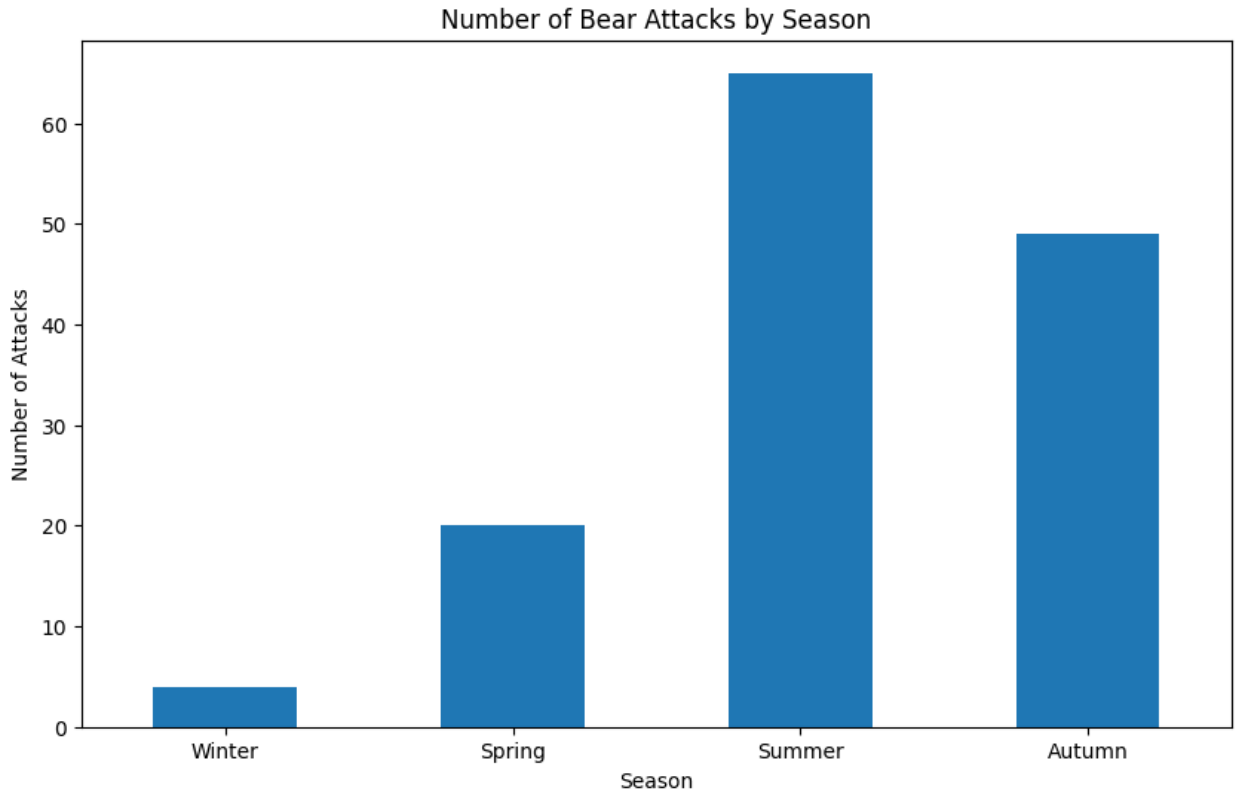

Figure 2.5 Bar Chart of Bear Attacks by Season

Season was an important data feature that we implemented into the dataset, the seasonal data allowed us to understand the patterns in annual bear attacks and how the climate may affect their nature. To understand the seasonal importance of bear attacks, it is important to understand the hibernation periods of bears.

Bears use hibernation to combat harsh winter temperatures and food scarcity that proceeds with colder winter months (Regad, 2024). Hibernation is a method bears use to slow down metabolic rates, conserve energy and endure the food scarcity that the winter months bring. These hibernation periods, that coincide with winter, usually begin in September or October and end in the spring months of March or April. From *Figure 2.3.3.*, it is evident that majority of bear activity will take place in the summer months.

2.3.4 Type of Bear

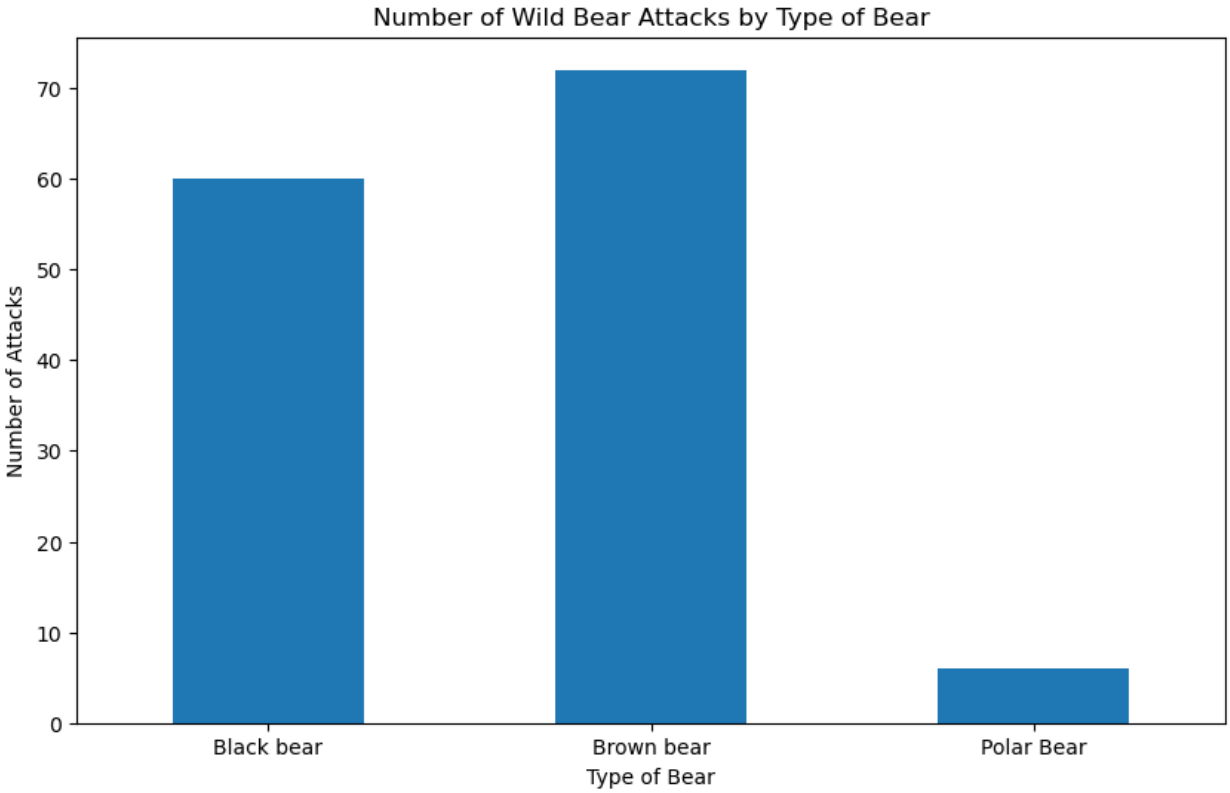

Figure 2.6 Bar Chart of Bear Attacks by Type of Bear

It is clear to see that Brown (Grizzly) bears pose the greatest threat to human life within this region. This is especially significant to consider given the population of this type of bear in comparison to their cousins – Black bears. Brown bear population in North America (around 55,000) is significantly inferior to Black bears (estimated between 600,000 – 900,000) (Smith, 2024). The aggression shown by Brown bears can be attributed to their strict carnivorous diet and sometimes hostile nature in comparison to that of the black bear, who can often be sufficiently nourished by berries, and will only attack upon provocation (Smith, 2024).

Therefore, this data shows that policy should be tailored to prioritise areas with a high Brown bear population, and this is represented in *figure 2.3.5.*

Polar bears – as represented in *figure 2.3.4*. - are an anomaly here. With their habitats being positioned in the Arctic Circle, the only attacks that have taken place involving polar bears are found to be in parts of northern Canada and Alaska that are largely uninhabitable for humans.

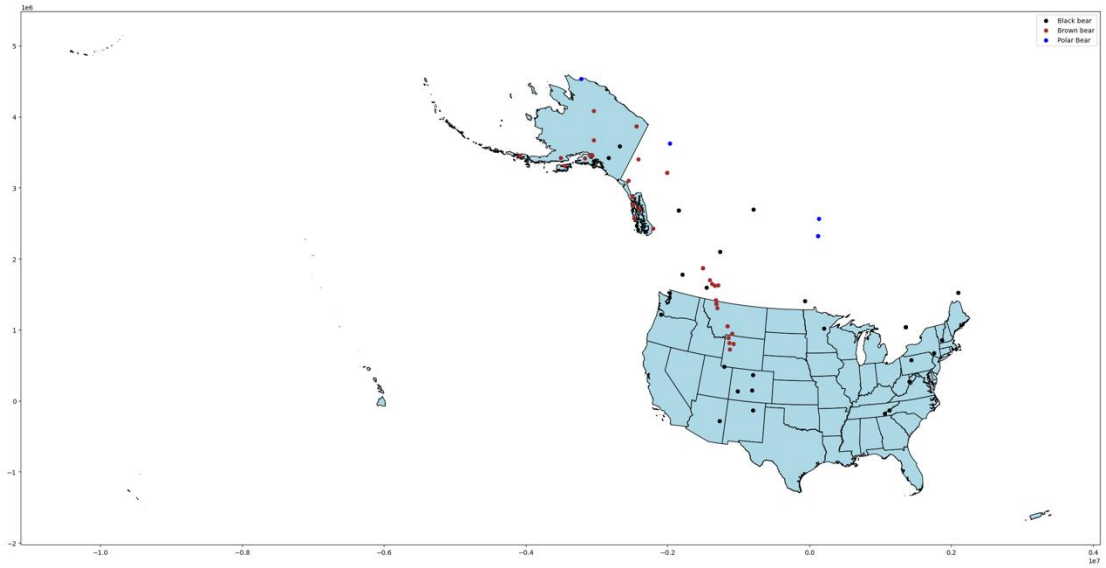

Figure 2.37.5 Location of Bear Attacks by Bear Type

3 Unsupervised Analysis

3.1 K-means Clustering

For the government and local authorities, identifying 'hotspots' where bear attacks occur is a priority to protect both the human and bear populations in such areas. Ensuring these locations are correctly identified and then appropriately managed is imperative, and this cannot be done sufficiently without recognising the geospatial trends of fatal bear attacks.

It is important to understand such trends in bear attacks per location as comprehensively as possible, therefore a k-means clustering, renowned for being well suited for identifying spatial trends, will be implemented for such analysis.

This analysis used the silhouette score to determine the optimal number of clusters for identifying bear attack hotspots. The silhouette score evaluates how well each data point fits

within its assigned cluster relative to others, with scores closer to 1 indicating well-defined and distinct clusters.

To implement this approach, latitude and longitude variables were selected to form clusters, which were then scaled to ensure equal weighting of latitude and longitude in the clustering process.

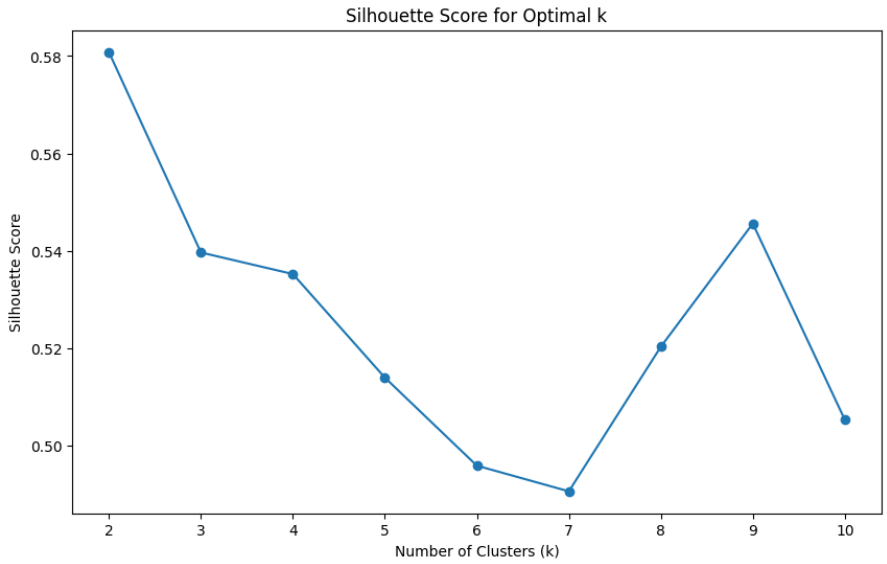

Figure 3.1 Silhouette Scores to find Optimal Number of Clusters

A range of potential numbers of geographical clusters from 2 to 10 was tested to find the quantity with the highest silhouette score. The silhouette scores for each value were calculated and plotted (see *Figure 3.1*). This plot shows that the silhouette score peaks at k=2, with a score of approximately 0.58. With k=2 identified as the optimal number of clusters; k-means clustering was applied to create two geographic clusters representing bear attack hotspots.

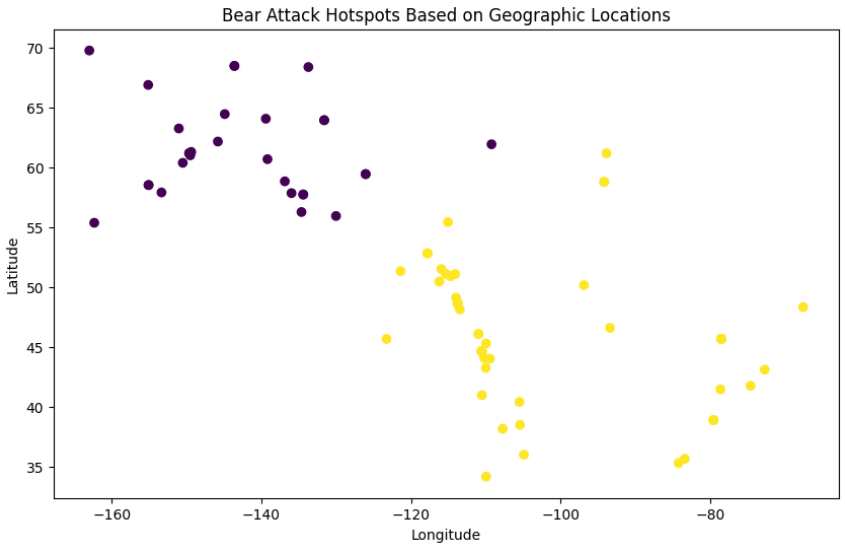

Figure 3.2 K-Means Clustering Results (k=2)

Figure 3.2 shows the spatial clustering of bear attacks across North America, with two broad clusters identified. Cluster 1 covers a large area in the northwest, while Cluster 2 spans a wide region across the midwestern and northeastern parts. To understand this within context and to identify more specific hotspots, *Figure 3.3* presents these clusters points overlayed onto a map of North America.

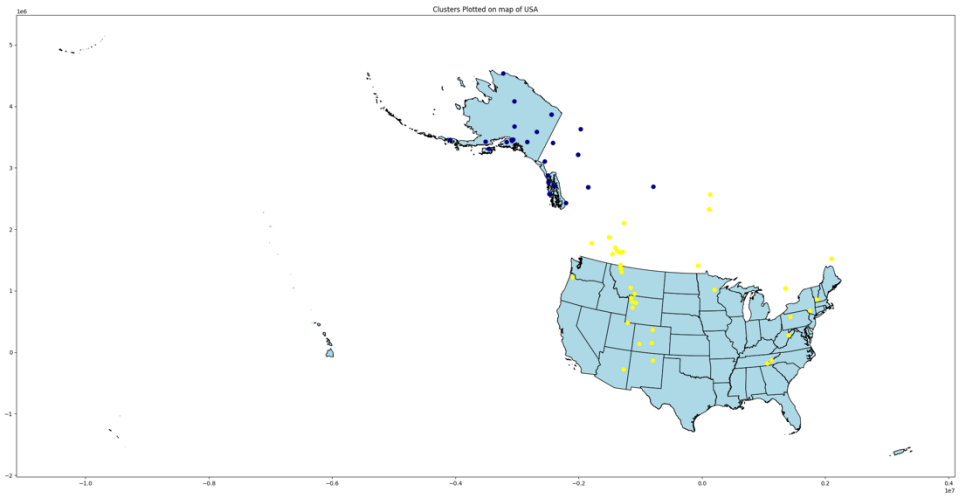

Figure 3.3 K-Means Clustering Points Plotted on Map of USA / North America

While this clustering clearly highlights general areas with a higher frequency of bear attacks, the clusters are broad and lack finer geographic detail. Despite the inconclusive nature of these results, they can be seen as a foundation for future research; as more bear attacks occur, patterns both between, and within clusters may become more visible.

4 Supervised Analysis

4.1 Linear Regression

As mentioned previously, we are interested in exploring a potential causal relationship between bear attacks by season over time. Establishing the extent of this relationship will allow us to observe if bears' hibernation periods are reducing due to extraneous variables such as a change in climate. This analysis will be approached as a multiple linear regression, as there are two independent variables: season and decade. These variables directly relate to our research question based on finding their influence on the dependent variable: the number of bear attacks.

From this model we can observe the following:

Variable	*Coefficient*
Decade	*0.079180*
Spring	*-2.333603*
Summer	*2.039164*
Winter	*-4.062838*

Table 4.1 Regression Output

- Decade: The positive coefficient of 0.079 suggests an increase in bear attacks over time, potentially linked to rising human populations settling on bear habitats, leading to more interactions.

- Spring: The coefficient of -2.334 indicates fewer bear attacks in Spring compared to the baseline season (Autumn).

- Summer: The positive coefficient for Summer (2.039) indicates an increase in bear attacks, likely due to increased bear activity and human outdoor activities in warmer months.

- Winter: The coefficient of -4.063 suggests significantly fewer bear attacks in Winter, likely due to hibernation, which reduces bear activity during colder months.

The model's Mean Absolute Error (MAE) is 1.96, indicating that the average difference between the predicted and actual number of bear attacks is approximately 1.96. Moreover, the model achieved an R-squared value of 0.62, meaning it explains about 62% of the variation in bear attacks. This suggests a moderately strong fit, although other factors may also play a role in bear attack frequency, such as a growing population and the urbanisation of bear habitats. Future research could apply similar regression models as seen in this study looking into the influence of these independent variables, ultimately giving us a more holistic picture of bear attack trends.

The results indicate that bear attacks – while increasing - vary significantly by season, with higher incidents in Summer and lower in Winter, aligning with expected typical bear activity patterns. Although these results do not necessarily indicate that there is a present increase in bear attacks in typically colder seasons, the model can act as a strong foundation for future research and should be referred to in the future as our climate continues to change.

5 Reflections

While perhaps the results of this analysis did not go how exactly we would have liked, we still believe that utilising both the k-means clustering approach and the multiple linear regression were the most appropriate for our dataset and research questions.

K-means clustering approaches are renowned for their suitability to geospatial analysis, which is what we were aiming to investigate. Moreover, regression analysis is an ideal method for establishing causal relationships between variables, we were looking to investigate the extent to which the progression of time predicts the increase in bear attacks in 'unusual' months. Albeit insignificant, the regression analysis provided us with this exact metric.

We are aware of the likes of publication bias and the 'file-drawer problem', so we do not want the non-statistically significant results dictating how publishable our report is. Rather, we would like the reader to focus on the rationale and methodology of our analysis and consider how it may be implemented further down the line. As the likes of climate and urbanisation continue to change, we have a foundation of a model to analyse how bear attacks will change as a result.

6 Conclusion

In conclusion, to identify key trends and issues in the best interest of safety for both humans and bears, this dataset by Umar Zia was comprehensively analysed. We recognised both the changes in bear attack location over time and the number of bear attacks per season over time to be of particular interest, thus a k-means clustering and multiple regression analyses were implemented respectively.

While results did not perhaps indicate that trends were significant enough to merit any present action to be taken, the analysis techniques are a strong foundation for future research, which can be implemented with any new data in this field in the future if required.

References

Davis, H., Wellwood, D.W. and Ciarniello, L.M., 2002. *" Bear Smart" Community Program: Background Report* (p. 108). Ministry of Water, Land and Air Protection.

Regad (2024). *Bear Hibernation Behavior: What you Need to Know.* Animal Behavior Corner. Available at: https://animalbehaviorcorner.com/bear-hibernation-behavior-what-you-need-to-know/?utm_content=cmp-true (Accessed 27/10/2024)

Schintler, L.A. and McNeely, C.L. (2022) *Encyclopedia of big data / [internet resource]*. Cham: Springer.

Schmitt (2013). *More Women Give Hunting a Shot*. National Geographic. Available at: https://www.nationalgeographic.com/culture/article/131103-women-hunters-local-meat-food-outdoor-sports#:~:text=Although%20men%20still%20account%20for%20the%20majority%20of,of%20women%20actively%20hunting%20is%20on%20the%20rise. (Accessed 25/10/2024)

Statista (2023). *Number of hiking participants in the United States from 2010 to 2023*. Statista. Available at: https://www.statista.com/statistics/191240/participants-in-hiking-in-the-us-since-2006/ (Accessed 28/10/2024)

United States Census Bureau (2021). *Historical Population Change Data (1910-2010).* United States Census Bureau. Available at: https://www.census.gov/data/tables/time-series/dec/popchange-data-text.html (Accessed 25/10/2024)

Wildlife Informer (2024). *Black Bear Population by State (Current Estimates)*. Wildlife Informer. Available at: https://wildlifeinformer.com/black-bear-population-by-state/ (Accessed 31/10/11)

Zia, U. (2021). *Bear Attacks North America*. Kaggle. Available at: https://www.kaggle.com/datasets/stealthtechnologies/bear-attacks-north-america (Accessed: 2024).

Appendix

Development Details - Graphs were produced on *VS Studio Code,* coded in a jupyter environment with the following packages:

- Matplotlib
- Pandas
- Numpy
- Geopandas
- Sklearn