Finding the states of north america which have the most bear attack records by month. Also examining is there any relation between age, season and attack.

We can identify which state is the most dangerous place for bear attack and which month is the most dangerous time or season.

Question

1. Which state and season is most dangerous for bear attack?
2. In which month which state has the highest records of bear attack?
3. Is there any relation between age to attack?

Data Sources

1. Datasource: kaggle
   - Metadata URL: https://www.kaggle.com/datasets/danela/fatal-bear-attacks-north-america/data
   - Data URL: https://raw.githubusercontent.com/szabolcsfule/bear_attacks/master/bear_attacks.csv
   - Data Type: CSV
   - Description: This dataset contains information of deadly attacks of wild bears in North America. The original data is from Wikipedia, latitude and longitude are added. The column "Bear" contains the type of bear, i.e., Brown Bear, Polar Bear or Black Bear. Cases of several deaths in one attack are split into one row per dead person.
   - License Type: CC BY-SA 3.0
2. Datasource: kaggle
   - Metadata URL: https://www.kaggle.com/datasets/stealthtechnologies/bear-attacks-north-america
   - Data URL: https://www.kaggle.com/datasets/stealthtechnologies/bear-attacks-north-america/data
   - Data Type: csv
   - License Type: OpenData License
   - Description: This dataset shows every recorded killing by a black, brown, or polar bear from 1900-present day in North America.

Data Pipeline:

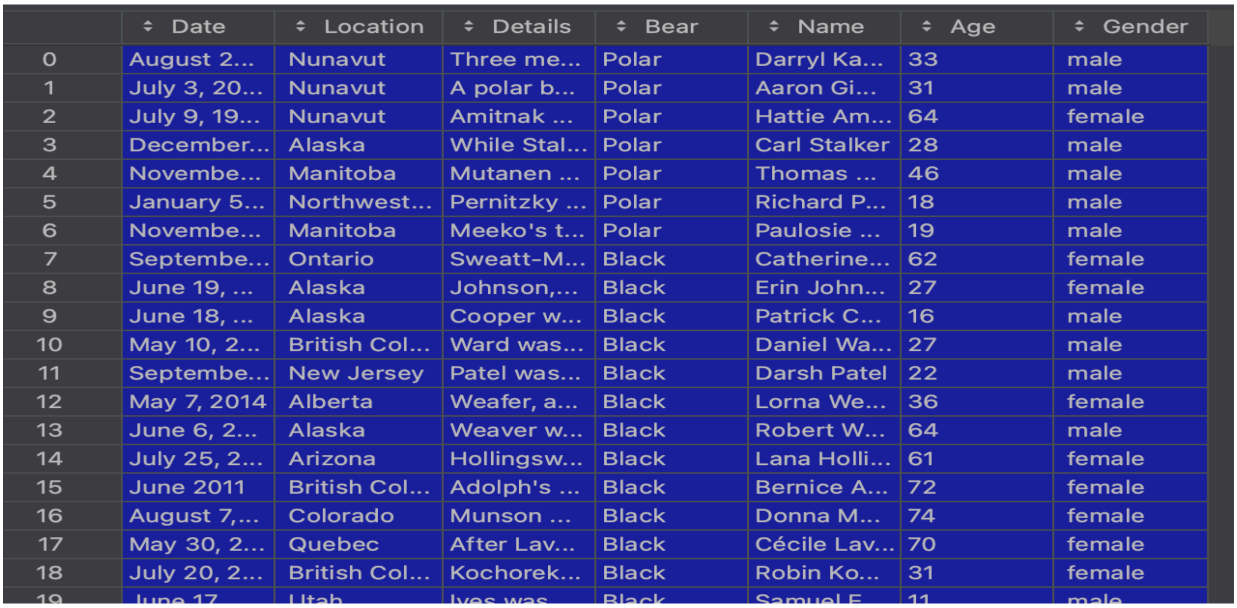
While building the data pipeline I have encountered a few crucial steps. There few columns had invalid data type and noisy data. Like age had some extra texts and decimal numbers. That was a crucial part for data validation. And also for place names , we needed to extract the perfect state name from the address.

Technologies Used: Python,Pandas,SQLite,SQLAlchemy, Opendatasets

Data Cleaning/Transformation Steps:
There were some missing values inside the datasets in the age and time. Those were filled with interpolation methods. There were some outliers which were also solved. I have normalized the data according to the necessity of good output and reports.

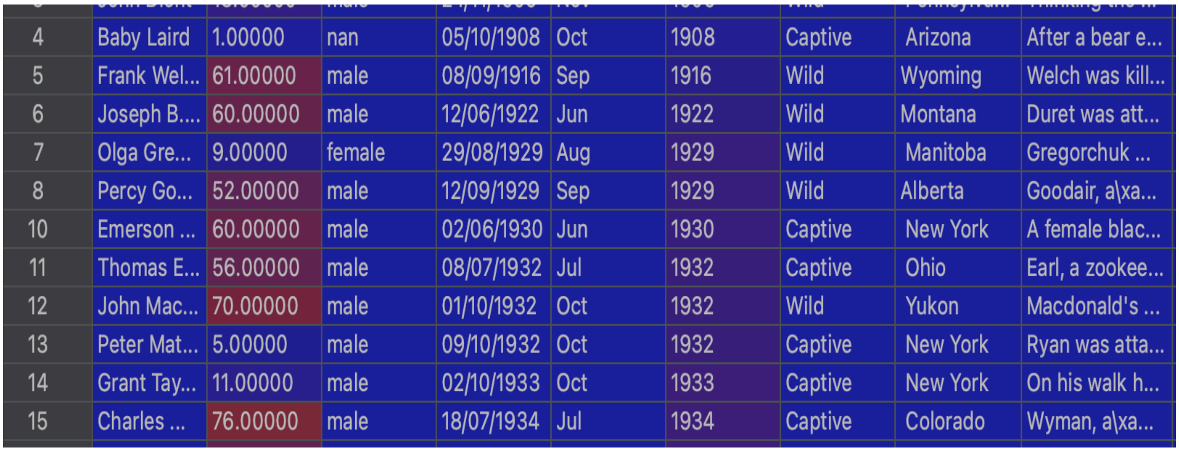
Problems Encountered: Developing the datapipeline the most complex job was to normalize the two datasets in the same manner , thus we can get the same insights. Like attack place names were in different different formats , the data type of ages were different. We need to extract the right data from those datasets in the right format. For getting meaningful insights this transformation was really necessary.

working on the data engineering process, one of the main challenges was dealing with varying accident reporting standards across different regions. Additionally, ensuring consistent time resolution in both datasets was crucial for establishing a meaningful correlation. These challenges were addressed through meticulous data validation and cleaning procedures.



	⇕ Date	⇕ Location	⇕ Details	⇕ Bear	⇕ Name	⇕ Age	⇕ Gender
0	August 2...	Nunavut	Three me...	Polar	Darryl Ka...	33	male
1	July 3, 20...	Nunavut	A polar b...	Polar	Aaron Gi...	31	male
2	July 9, 19...	Nunavut	Amitnak ...	Polar	Hattie Am...	64	female
3	December...	Alaska	While Stal...	Polar	Carl Stalker	28	male
4	Novembe...	Manitoba	Mutanen ...	Polar	Thomas ...	46	male
5	January 5...	Northwest...	Pernitzky ...	Polar	Richard P...	18	male
6	Novembe...	Manitoba	Meeko's t...	Polar	Paulosie ...	19	male
7	Septembe...	Ontario	Sweatt-M...	Black	Catherine...	62	female
8	June 19, ...	Alaska	Johnson,...	Black	Erin John...	27	female
9	June 18, ...	Alaska	Cooper w...	Black	Patrick C...	16	male
10	May 10, 2...	British Col...	Ward was...	Black	Daniel Wa...	27	male
11	Septembe...	New Jersey	Patel was...	Black	Darsh Patel	22	male
12	May 7, 2014	Alberta	Weafer, a...	Black	Lorna We...	36	female
13	June 6, 2...	Alaska	Weaver w...	Black	Robert W...	64	male
14	July 25, 2...	Arizona	Hollingsw...	Black	Lana Holli...	61	female
15	June 2011	British Col...	Adolph's ...	Black	Bernice A...	72	female
16	August 7,...	Colorado	Munson ...	Black	Donna M...	74	female
17	May 30, 2...	Quebec	After Lav...	Black	Cécile Lav...	70	female
18	July 20, 2...	British Col...	Kochorek...	Black	Robin Ko...	31	female
19	June 17	Utah	Ives was	Black	Samuel E	11	male

Fig: Attack incidents data



[illegible]	[illegible]	[illegible]	male	[illegible]	Nov	[illegible]	Wild	[illegible]	[illegible]
4	Baby Laird	1.00000	nan	05/10/1908	Oct	1908	Captive	Arizona	After a bear e...
5	Frank Wel...	61.00000	male	08/09/1916	Sep	1916	Wild	Wyoming	Welch was kill...
6	Joseph B....	60.00000	male	12/06/1922	Jun	1922	Wild	Montana	Duret was att...
7	Olga Gre...	9.00000	female	29/08/1929	Aug	1929	Wild	Manitoba	Gregorchuk ...
8	Percy Go...	52.00000	male	12/09/1929	Sep	1929	Wild	Alberta	Goodair, a\xa...
10	Emerson ...	60.00000	male	02/06/1930	Jun	1930	Captive	New York	A female blac...
11	Thomas E...	56.00000	male	08/07/1932	Jul	1932	Captive	Ohio	Earl, a zookee...
12	John Mac...	70.00000	male	01/10/1932	Oct	1932	Wild	Yukon	Macdonald's ...
13	Peter Mat...	5.00000	male	09/10/1932	Oct	1932	Captive	New York	Ryan was atta...
14	Grant Tay...	11.00000	male	02/10/1933	Oct	1933	Captive	New York	On his walk h...
15	Charles ...	76.00000	male	18/07/1934	Jul	1934	Captive	Colorado	Wyman, a\xa...

Fig: Bear attack data

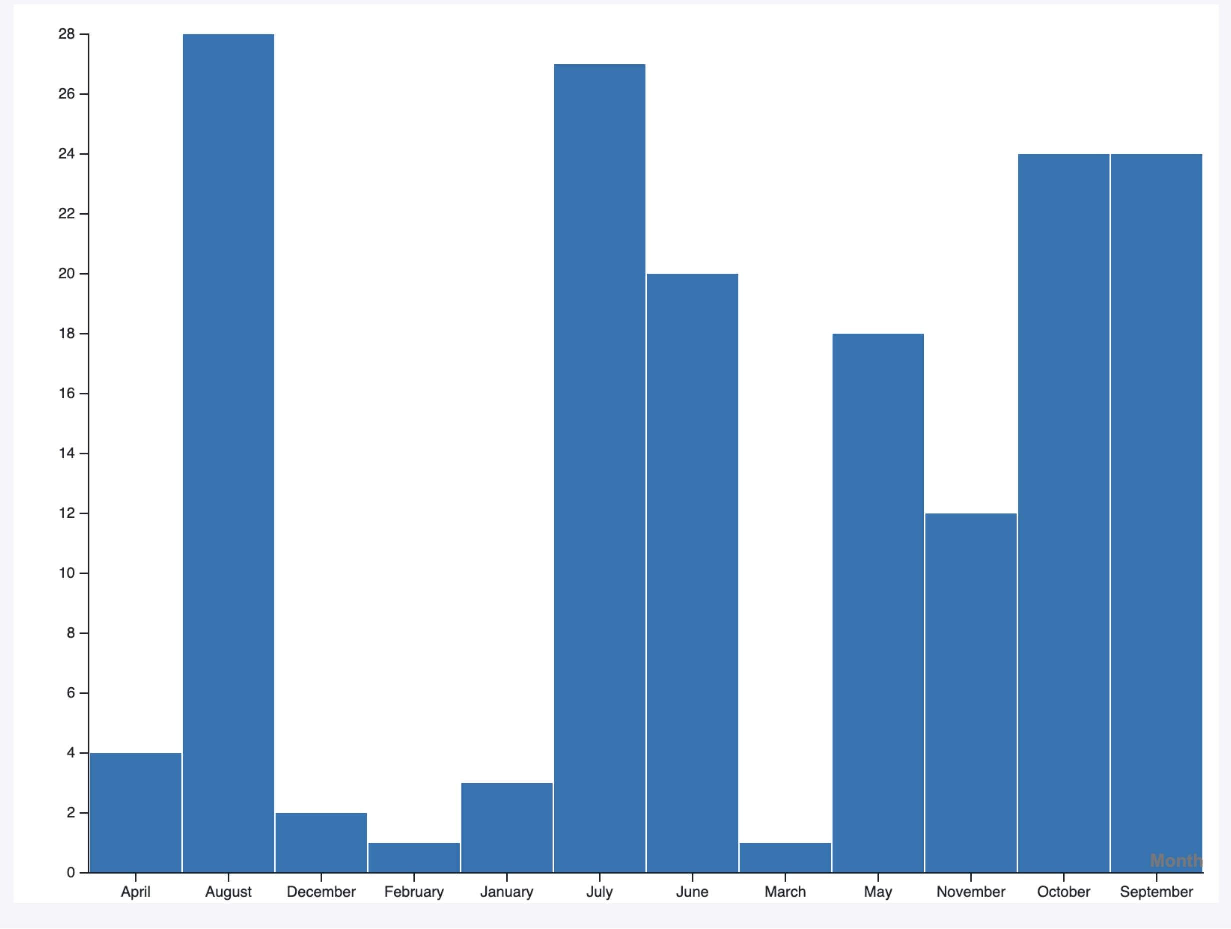


Fig: Bear attack per month

Result and Limitations

This plot clearly shows that the bear attacks increase in Summer season and also it depends on the State. Their notable time increment after April of bear attacks. In August bear attacks had its highest peak. And in February bear attacks had its lowest peak. Which indicates that seasonal temperature is correlated.

This dataset contains information of deadly attacks of wild bears in North America. The original data is from Wikipedia, latitude and longitude are added. The column "Bear" contains the type of bear, i.e., Brown Bear, Polar Bear or Black Bear. Cases of several deaths in one attack are split into one row per dead person. Then we shaped data into common format. Like date, month, year, state name etc. Thus we can get the correlation.

Resolution: Monthly incidents show us a correlation between incidents and the season. Next we will try to find out what is the relation between states and the attacks. Next we will try to find out if there is any relation between age and number of incidents.