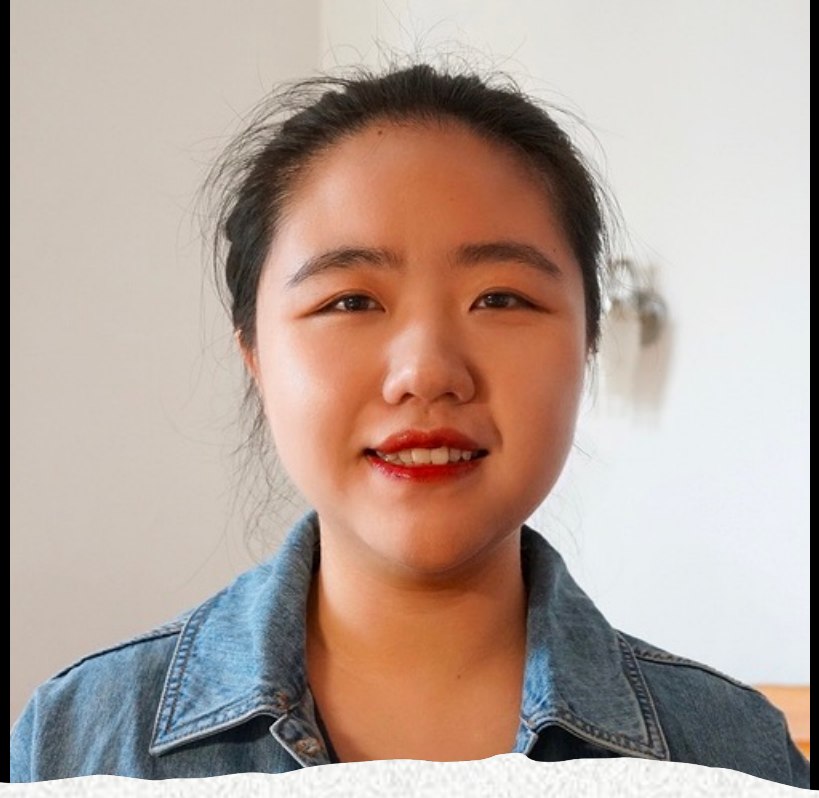


9 months | Data Science Intern | CDFW

4 years | Data Analyst | Consulting

Master in Data Science | USF

Bachelor in Applied Mathematics | BIT

XIN AI

San Jose | San Francisco

Specialized in photography

Xinnnnn.ai@gmail.com

linkedin.com/in/xinai-x/

Photographed by Xin Ai | 2022-2023 | Bay Area

CDFW BLACKBEAR

06/23/2023

SOCIAL MEDIA NLP PROJECT

INTRODUCTION

History Pain Points

1. Insufficient bear-encounter data points to analyze the frequency and geographic distribution over time.

2. Limited data hampers the ability to extract insights and update California Black bear's management plan.

Target of the practicum

1. Gather a comprehensive dataset of social media posts relevant to bear-encounters.

2. Apply machine learning algorithms and NLP techniques to detect potential patterns and analyze public sentiments associated with bear-encounters.

MILESTONES

Pipeline

- Built a data pipeline with Twitter API
- Conducted it in Airflow

Database

- Collected 450k tweets
- Put the database to GCP Storage for data retrieve

Feature Engineering

- Assigned coordinates by merging data sources
- Unified county labels with Spark

Data Refinement

- POS Tagging
- DBSCAN Clustering
- Cut ~50% irrelevant tweets

Labeling

- Identified data imbalances
- Utilized iterative semi-supervised learning

Modeling

- ML: Ensemble model with classification algorithms
- DL: Fine tune RoBERTa
- Reached 0.91 F1 score

Dashboard

- Built an interactive dashboard
- Deployed it to AWS EC2

CHALLENGES | Identify bear-encounter tweets

Challenge

- "Bear" is a versatile word.

Noun **Verb** **Adjective**

- "Bear" is employed in fixed collocations.

bear market **Big Bear Lake**

- "Bear" has diverse meanings beyond the animal species.

Pet **Doll** **People**

What I thought

- We can focus on collecting tweets where 'bear' is used as a noun.
- We can utilize patterns or collocations of 'bear' to filter out irrelevant tweets.

What I did

- POS Tagging
- DBSCAN Clustering

CHALLENGES | Handle imbalanced data

Challenge	What I thought	What I did
• The bear-encounter tweets, representing the minority class, have a class weight of less than 5%.	• Prior to modeling, we can augment the number of minority class data. • During the modeling, we can: 1. simulate class weights of the whole data; 2. enhance the penalty for misclassifying minority instances.	• Iterative semi-supervised learning: Spectral Clustering + AdaBoost • Data Augmentation + Modeling

OVERVIEW

- Data source: Twitter API
- Posted Period: 2010/01/01 – 2022/12/31
- Posted Area: California
- Entries: 6017
- Features: 11

Mainly scenarios

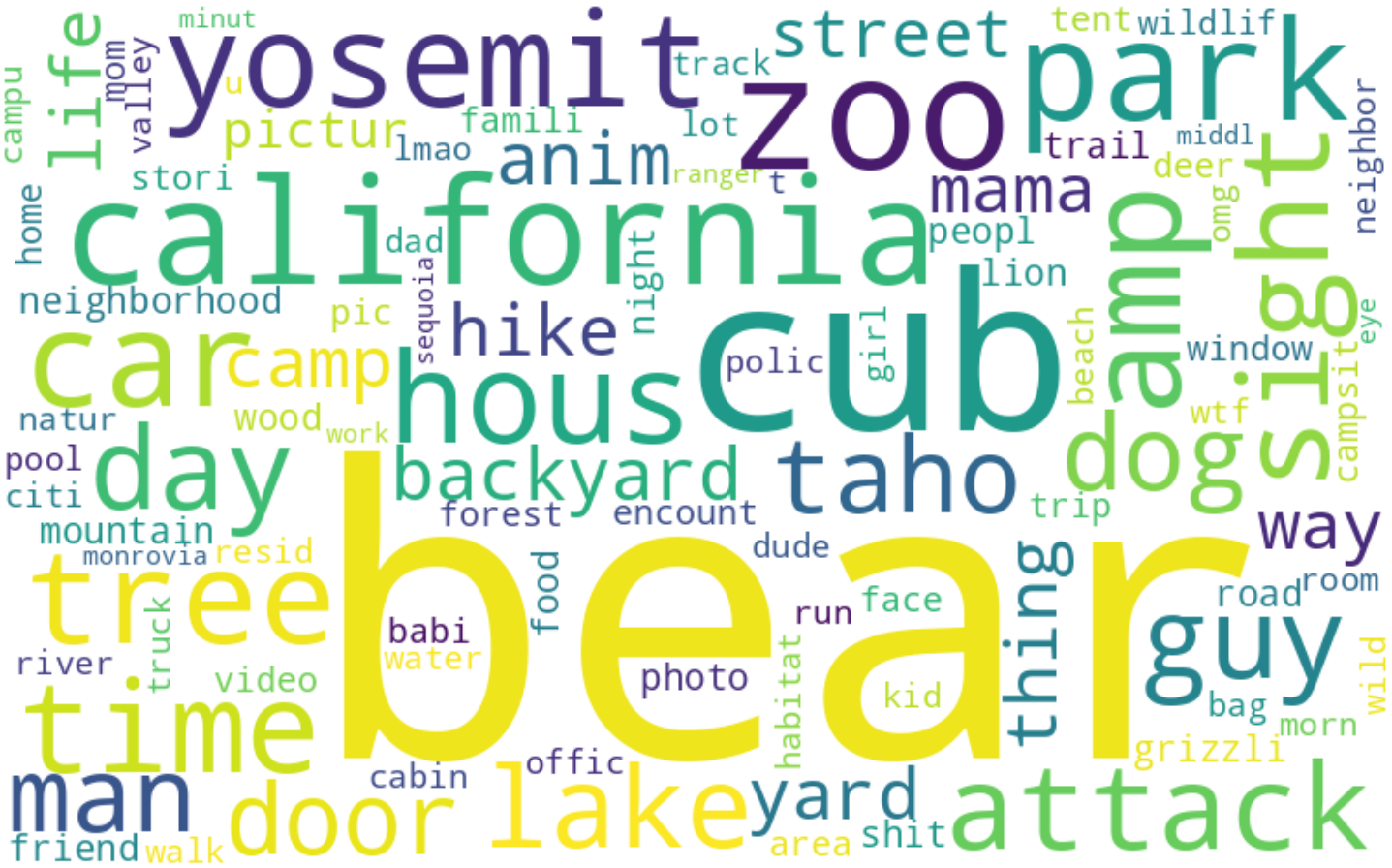
Wild - Forest
Yosemite, Sequoia, hike, trail, tree, habitat, mountain, etc.

Wild - Places with water
Tahoe, lake, river, valley, etc.

Human Territory
zoo, house, backyard, camp, car, road, pool, campus, etc.

Top Nouns

Root-word Cloud



Mainly scenarios

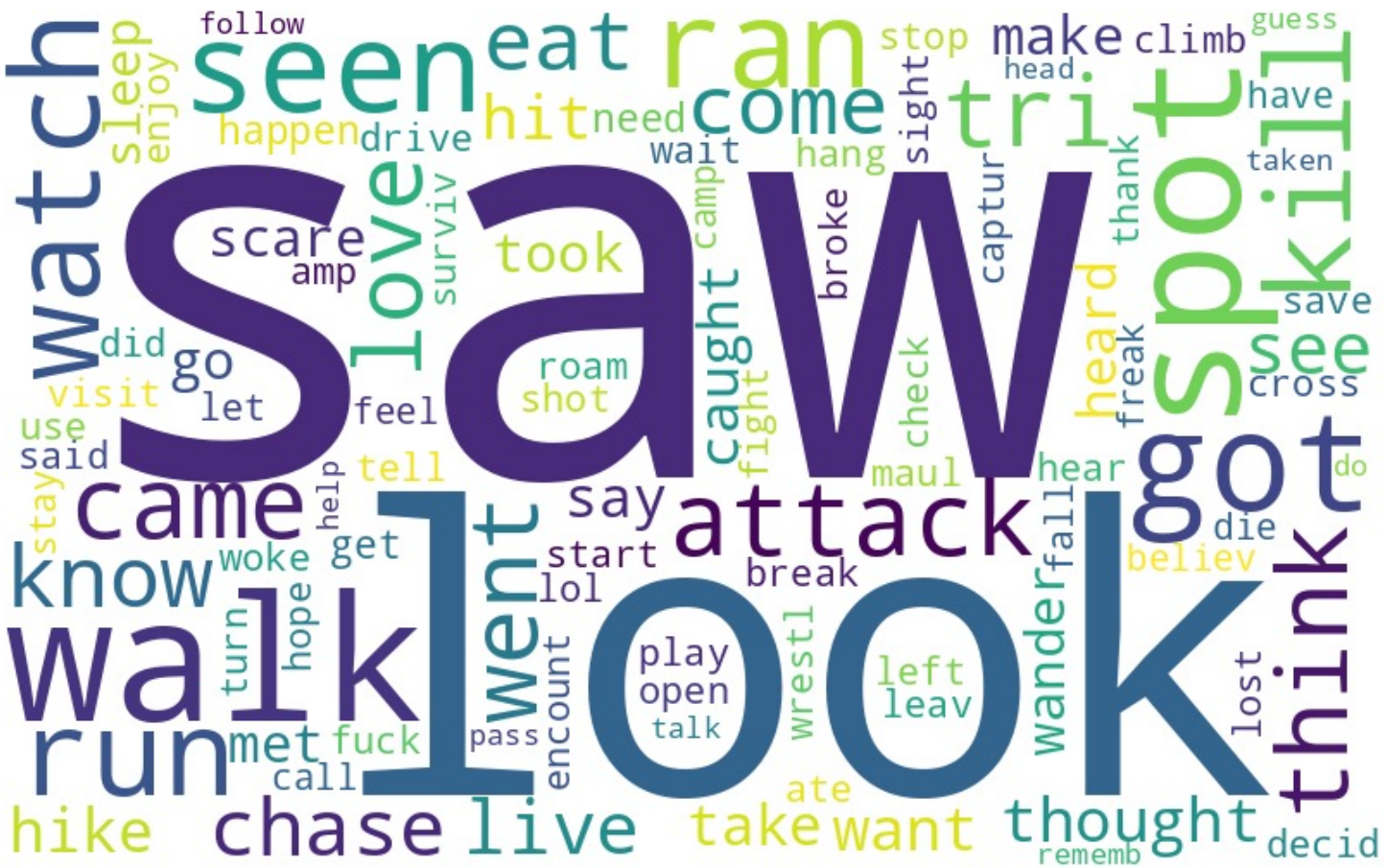
How to meet
saw, look, watch, seen, spot, sight, heard, encounter, etc.

What bears did
attack, chase, walk, play, cross, wander, hit, visit, etc.

What people did
ran, drive, leave, pass, wait, caught, fight, survive, kill, etc.

Top Verbs

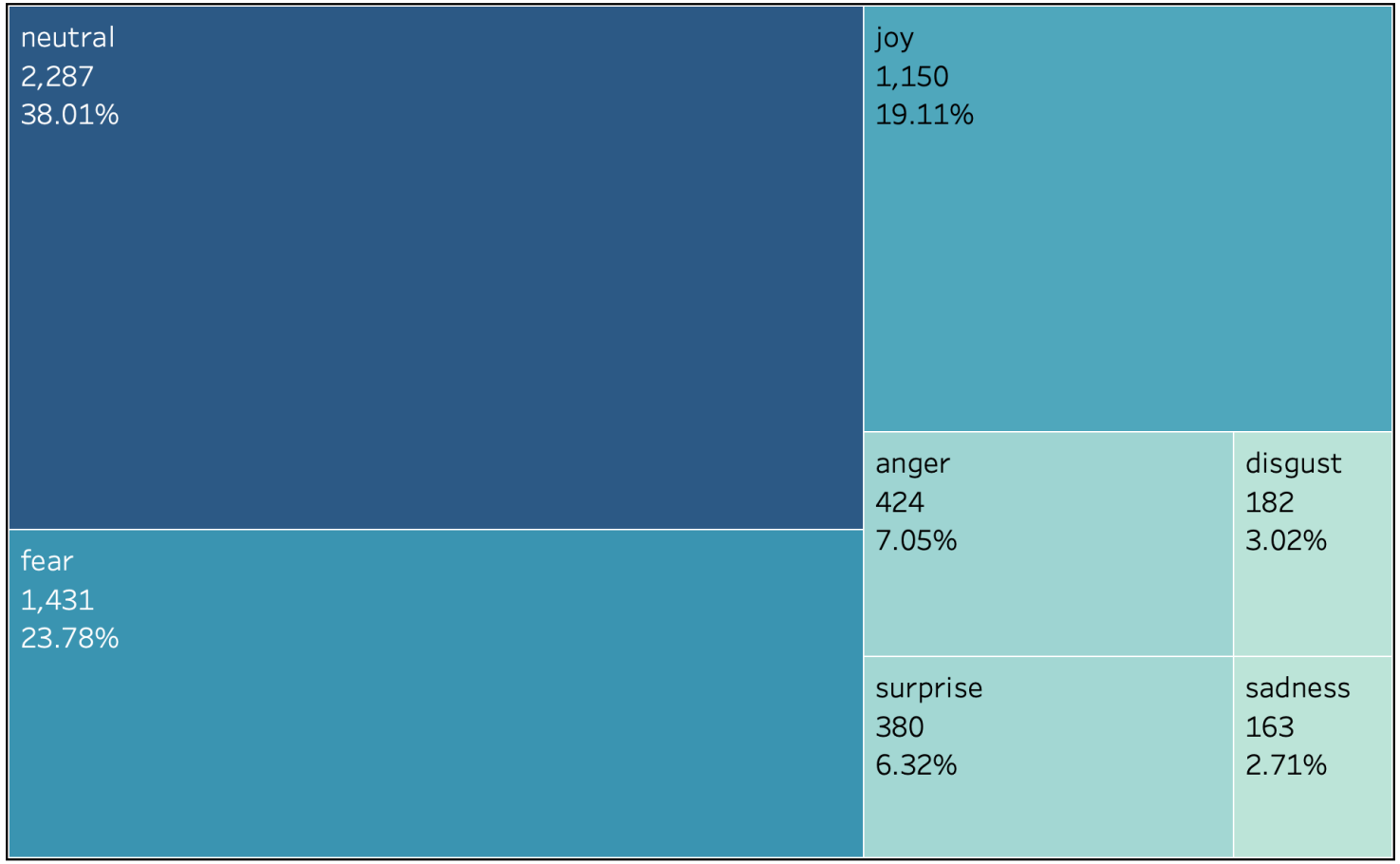
Root-word Cloud



Aside from the 38% of people who are indifferent, the predominant responses to bear encounters are a mixture of **fear (23.78%)** and **joy (19.11%).**

Public Sentiments on Bear-encounters

Public Sentiment Analysis of Bear Encounters Inferred from Tweets (2010-2022)

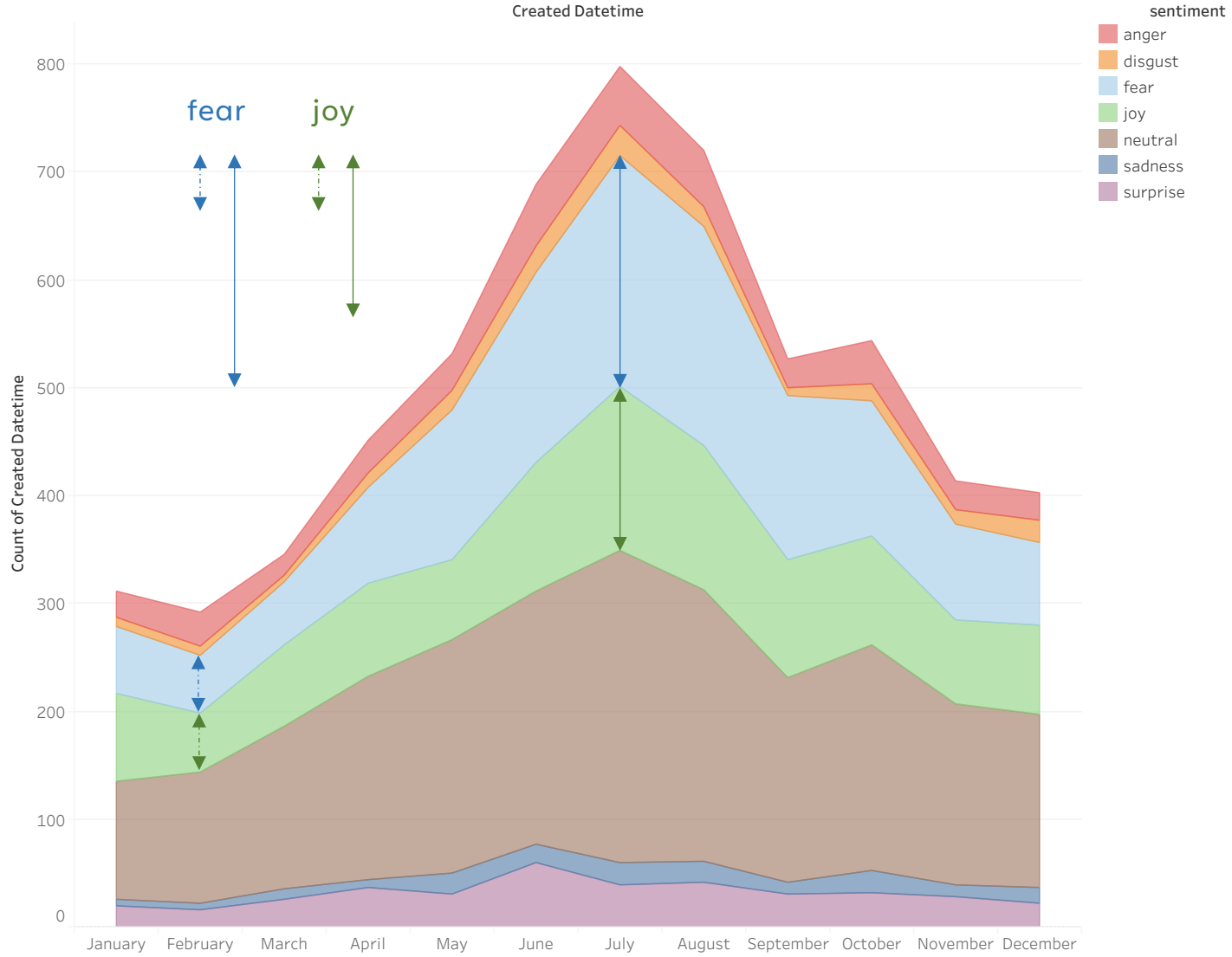


Sentiment, count of Sentiment and % of Total Count of Sentiment. Color shows count of Sentiment. Size shows count of Sentiment. The marks are labeled by Sentiment, count of Sentiment and % of Total Count of Sentiment. Details are shown for Sentiment.

Summer is the season with the highest number of bear encounters, and it's also the time when fear is most pronounced, often over-shadowing the sense of joy.

Frequency and Sentiments of Bear-encounters

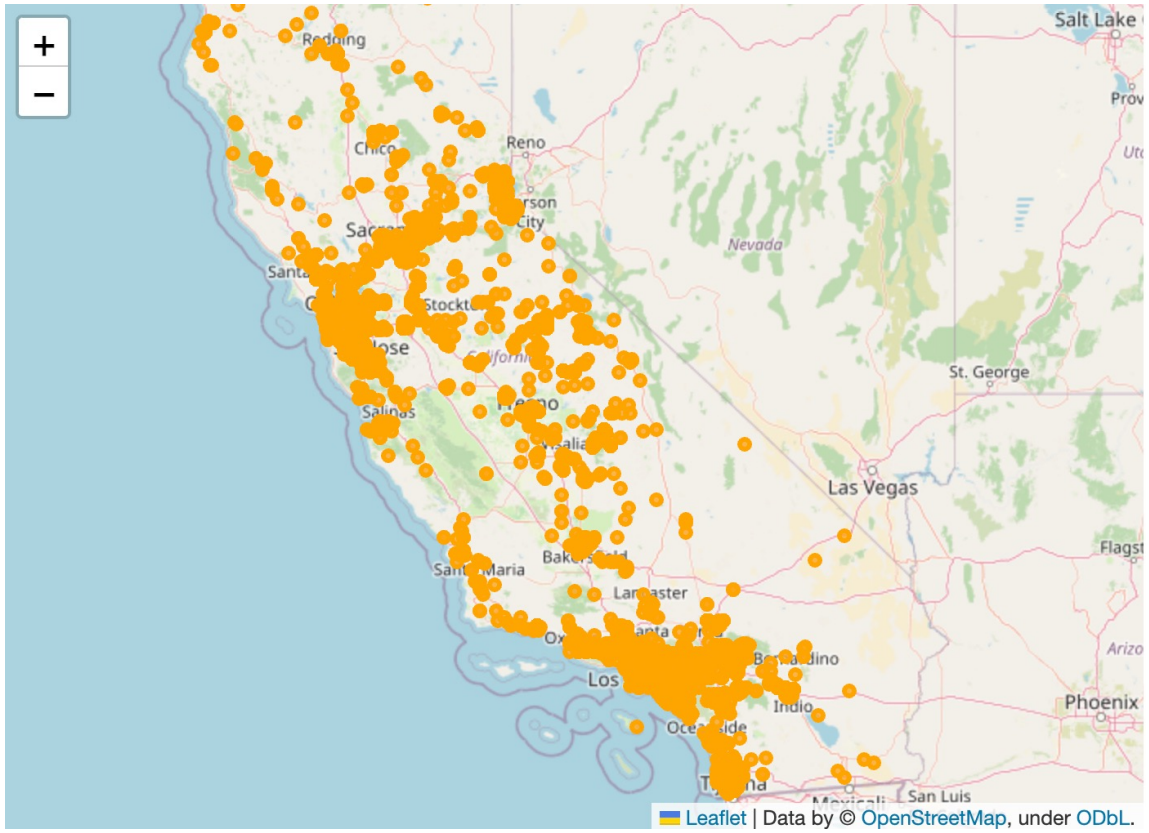
Monthly Trends in Sentiments Towards Bear Encounters Inferred from Tweets (2010-2022)



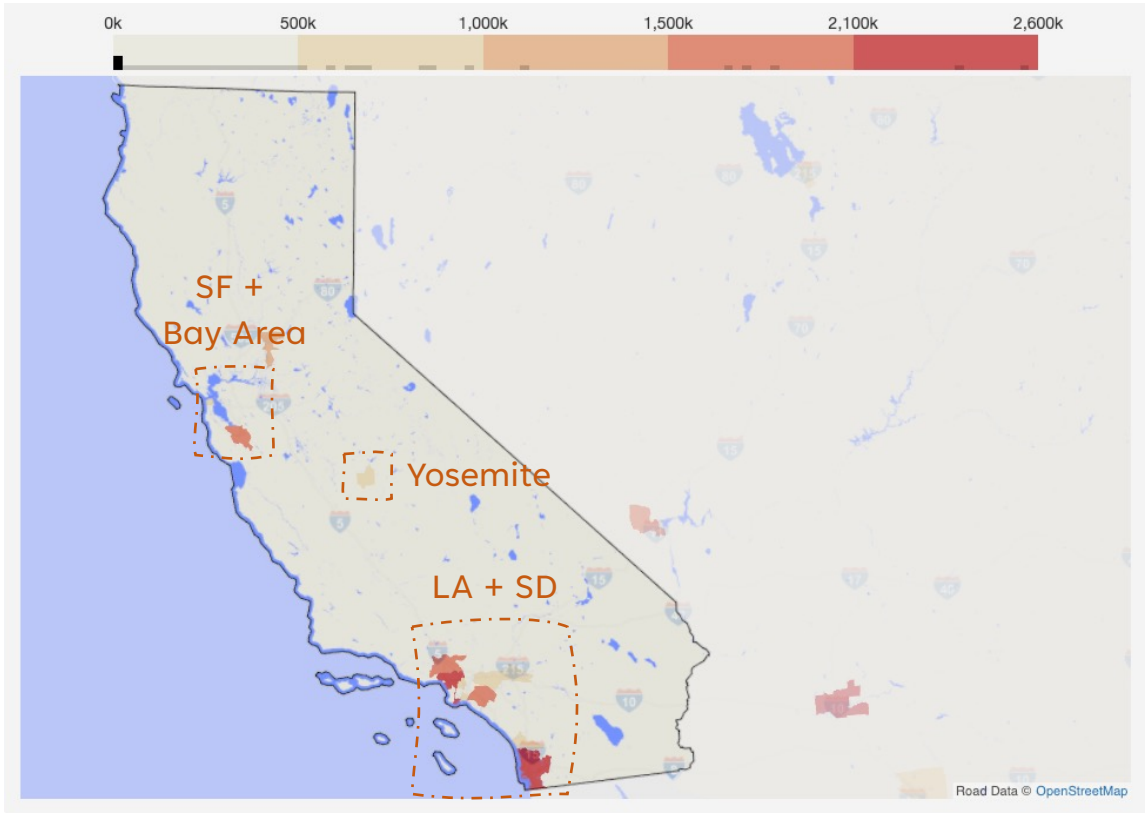
Count of Created Datetime for each Created Datetime Month. Color shows details about sentiment. The view is filtered on Created Datetime Month, which excludes Null.

Posting tweets about bear encounters is not a real-time behavior, but often occurs with a delay of several hours to days. As a result, such tweets were concentrated in areas with dense populations or popular outdoor destinations.

Geographic Distribution of Bear Encounter Tweets



Population by County Subdivision in California



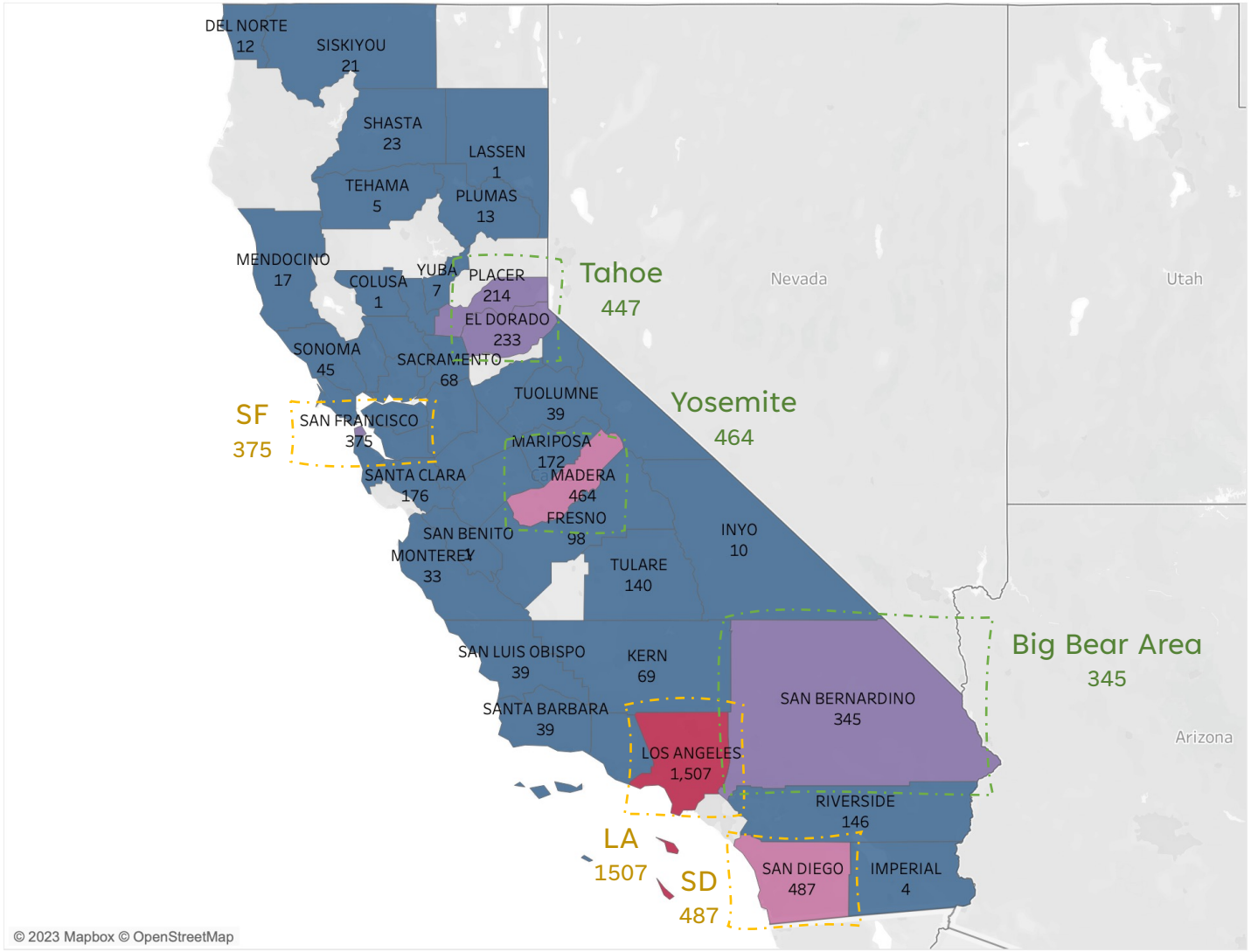
Areas with the most bear encounter tweets (Metro Area & Outdoor destinations)

- Los Angeles | 1507
- San Diego | 487
- San Francisco | 375

- Madera (Yosemite) | 464
- Placer + EL Dorado (Tahoe) | 447
- San Bernardino (Big Bear Area) | 345

Geo Distribution of Bear-encounter Tweets

Geospatial Analysis of Bear Encounters Inferred from Tweets (2010-2022)



Map based on Longitude (generated) and Latitude (generated). Color shows count of County. The marks are labeled by County and count of County. Details are shown for County. The view is filtered on County, which keeps 55 of 55 members.

According to ANOVA, environmental factors (temperature, precipitation, etc.) significantly impact bear encounters in five of the top six areas. Higher temperatures and lower precipitation tend to promote more bear encounters.

Los Angeles - 1507

Bear encounters vs. Avg temperature
p value: 0.003 | Adj.R2: 0.565 | coef: 3.540
Bear encounters vs. Precipitation
p value: 0.001 | Adj.R2: 0.621 | coef: -20.849

San Diego - 487

Bear encounters vs. Avg temperature
p value: 0.080 | Adj.R2: 0.202 | coef: 0.951
Bear encounters vs. Precipitation
p value: 0.140 | Adj.R2: 0.125 | coef: -6.383

San Diego Zoo is a significant source of bear encounters.

San Francisco - 375

Bear encounters vs. Avg temperature
p value: 0.158 | Adj.R2: 0.108 | coef: 1.419
Bear encounters vs. Precipitation
p value: 0.046 | Adj.R2: 0.276 | coef: -4.233

Madera (Yosemite) - 464

Bear encounters vs. Avg temperature
p value: 0.000 | Adj.R2: 0.812 | coef: 1.982
Bear encounters vs. Precipitation
p value: 0.001 | Adj.R2: 0.674 | coef: -32.487

Placer+EL Dorado (Tahoe) - 447

Bear encounters vs. Avg temperature
p value: 0.000 | Adj.R2: 0.812 | coef: 1.982
Bear encounters vs. Precipitation
p value: 0.001 | Adj.R2: 0.674 | coef: -32.487

San Bernardino (Big Bear) - 345

Bear encounters vs. Avg temperature
p value: 0.026 | Adj.R2: 0.346 | coef: 0.518
Bear encounters vs. Precipitation
p value: 0.322 | Adj.R2: 0.008 | coef: -1.768

Public sentiments toward bear encounters shift throughout the year. In metropolitan areas, the summer season tends to trigger heightened fear and anger towards such interactions. Residents typically seek tranquility and routine in their daily lives. Unexpected encounters with potentially threatening wildlife such as bears can dramatically disrupt this sense of normalcy, resulting in intense reactions.

Conversely, people visiting popular outdoor destinations often exhibit more intense reactions in summer, but these emotions encompass both fear and joy. These contrasting emotions can be attributed to the mindset of outdoor enthusiasts. Engaging in activities like hiking, camping, fishing, and skiing, these individuals generally anticipate and prepare for potential wildlife encounters. These experiences offer opportunities to immerse themselves in the wilderness, fostering a deeper connection with nature, although sometimes daunting, can also elicit feelings of exhilaration and joy.

LOS ANGELES

Sentiments & Months One-way ANOVA:

F value: 66.579 | p value: 0.000

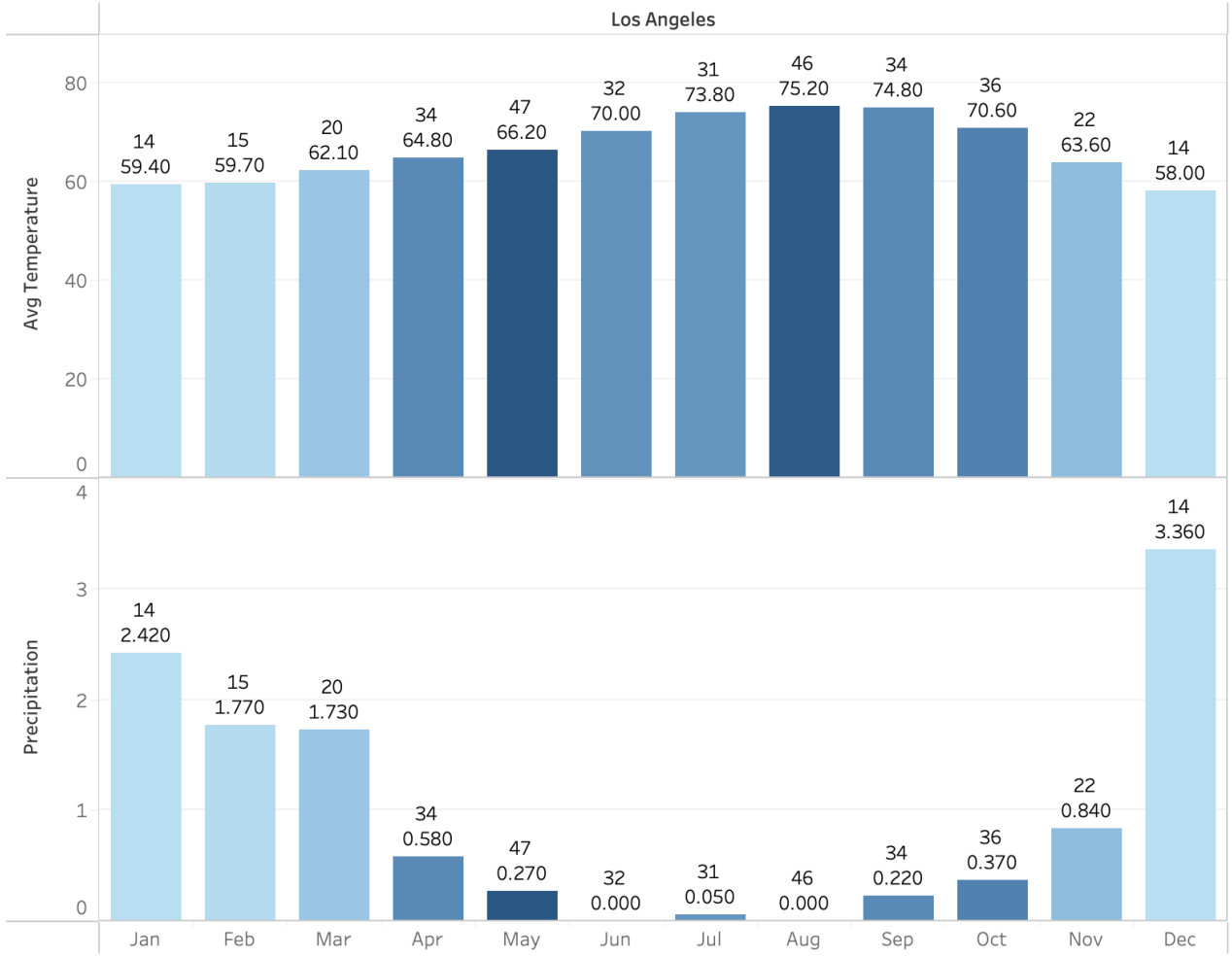
Correlation between sentiment (N>30) and Avg Temp:

fear: 0.773 | anger: 0.652 | neutral: 0.610 | sadness: 0.542 | joy: 0.337 | surprise: 0.200 | disgust: -0.060

Correlation between sentiment(N>30) and Precipitation:

fear: -0.837 | neutral: -0.707 | anger: -0.655 | sadness: -0.460 | surprise: -0.340 | disgust: 0.083 | joy: -0.020

Interplay of Climate Factors and Public Sentiment (Fear) of Bear Encounters - Los Angeles



Data Source: Twitter - tweets (N=1507), NOAA - weather data (weather station: LOS ANGELES DOWNTOWN/USC, CA), 2010-2022. This visualization illustrates the cumulative Average Temperature and Precipitation recorded in Los Angeles. The color gradient represents the collective sentiment of fear inferred from tweets. Each data point is labeled with the respective cumulative fear sentiment. For the Average Temperature pane, labels also show the corresponding temperature. Similarly, for the Precipitation pane, labels show the corresponding precipitation sum.

SAN DIEGO

Sentiments & Months One-way ANOVA:

F value: 70.477 | p value: 0.000

Correlation between sentiment (N>30) and Avg Temp:

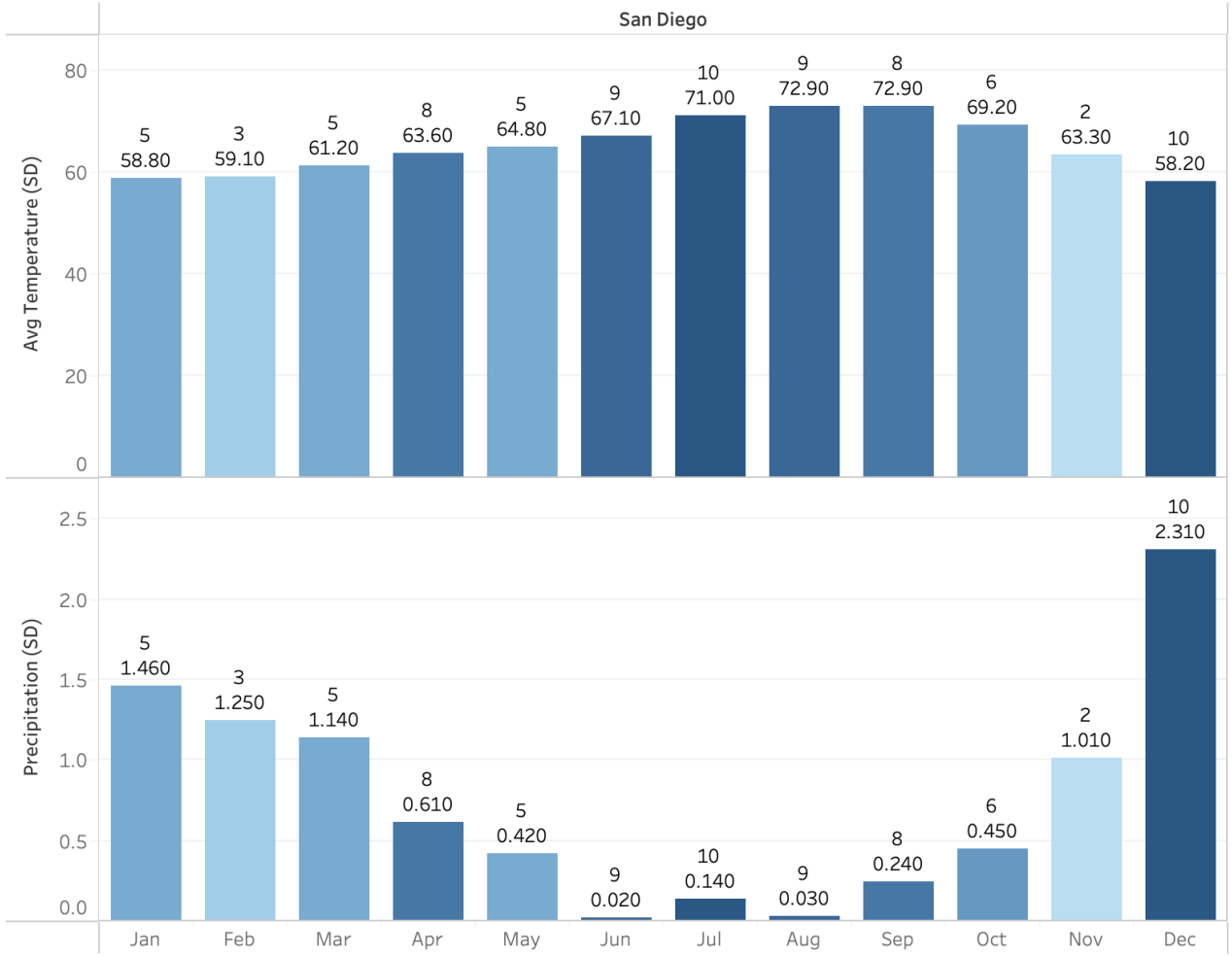
fear: 0.454 | neutral: 0.406 | joy: 0.284

Correlation between sentiment(N>30) and Precipitation:

neutral: -0.338 | fear: -0.252| joy: -0.109

The San Diego Zoo stands as a major source of bear encounters, particularly prominent in December. This notable volume during this month aligns with vacations such as the Christmas holiday, New Year's Eve, and school winter breaks.

Interplay of Climate Factors and Public Sentiment (Fear) of Bear Encounters - San Diego



Data Source: Twitter - tweets (N=487), NOAA - weather data (weather station: San Diego Area, CA), 2010-2022.
This visualization illustrates the cumulative Average Temperature and Precipitation recorded in San Diego. The color gradient represents the collective sentiment of fear inferred from tweets. Each data point is labeled with the respective cumulative fear sentiment. For the Average Temperature pane, labels also show the corresponding temperature. Similarly, for the Precipitation pane, labels show the corresponding precipitation sum.

SAN FRANCISCO

Sentiments & Months One-way ANOVA:

F value: 23.024 | p value: 0.000

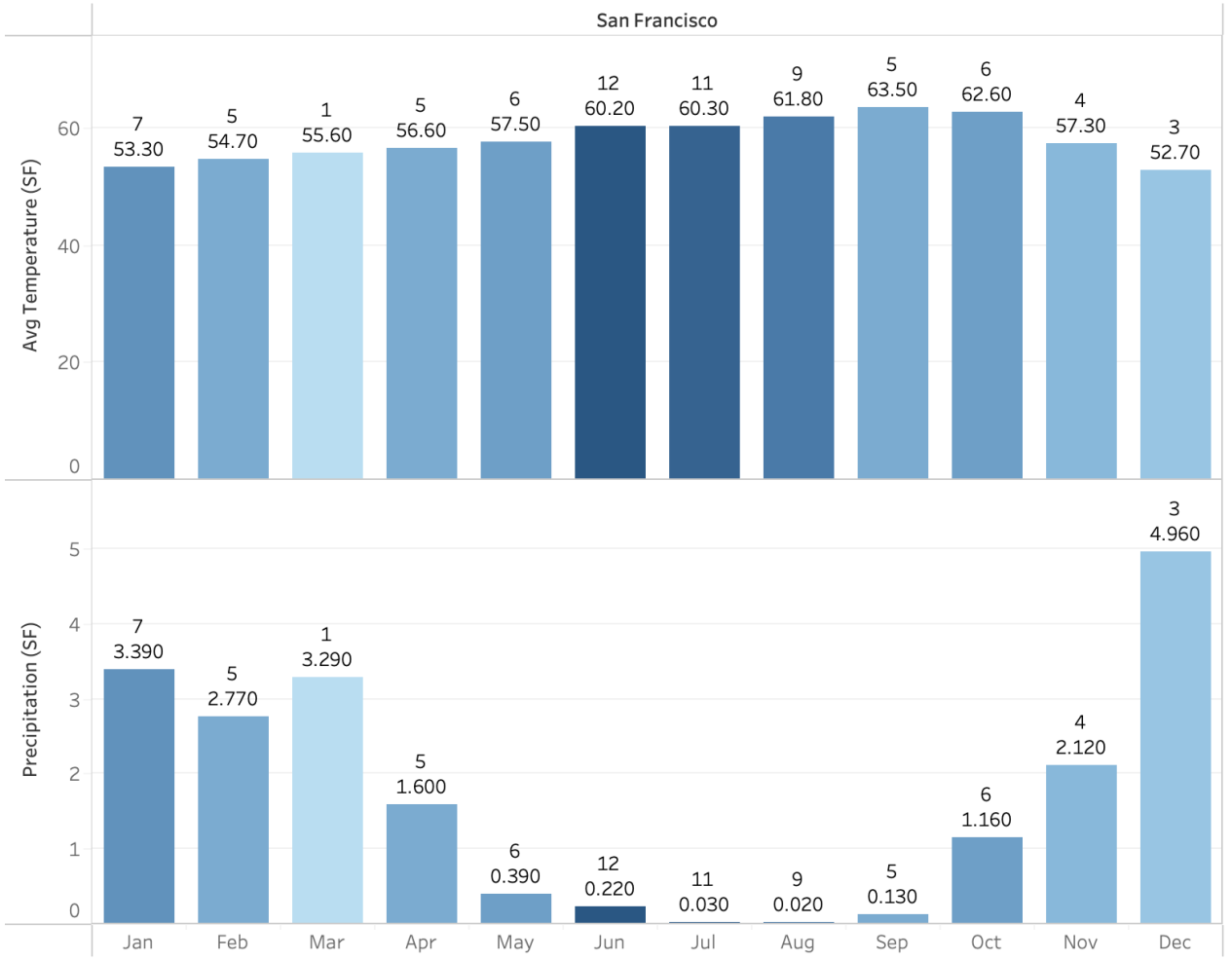
Correlation between sentiment (N>30) and Avg Temp:

fear: 0.470 | neutral: 0.396 | joy: 0.161

Correlation between sentiment(N>30) and Precipitation:

fear: -0.669 | neutral: -0.505 | joy: -0.164

Interplay of Climate Factors and Public Sentiment (Fear) of Bear Encounters - San Francisco



Data Source: Twitter - tweets (N=375), NOAA - weather data (weather station: SAN FRANCISCO DOWNTOWN, CA), 2010-2022.
This visualization illustrates the cumulative Average Temperature and Precipitation recorded in San Francisco. The color gradient represents the collective sentiment of fear inferred from tweets. Each data point is labeled with the respective cumulative fear sentiment. For the Average Temperature pane, labels also show the corresponding temperature. Similarly, for the Precipitation pane, labels show the corresponding precipitation sum.

YOSEMITE

Sentiments & Months One-way ANOVA:

F value: 10.230 | p value: 0.000

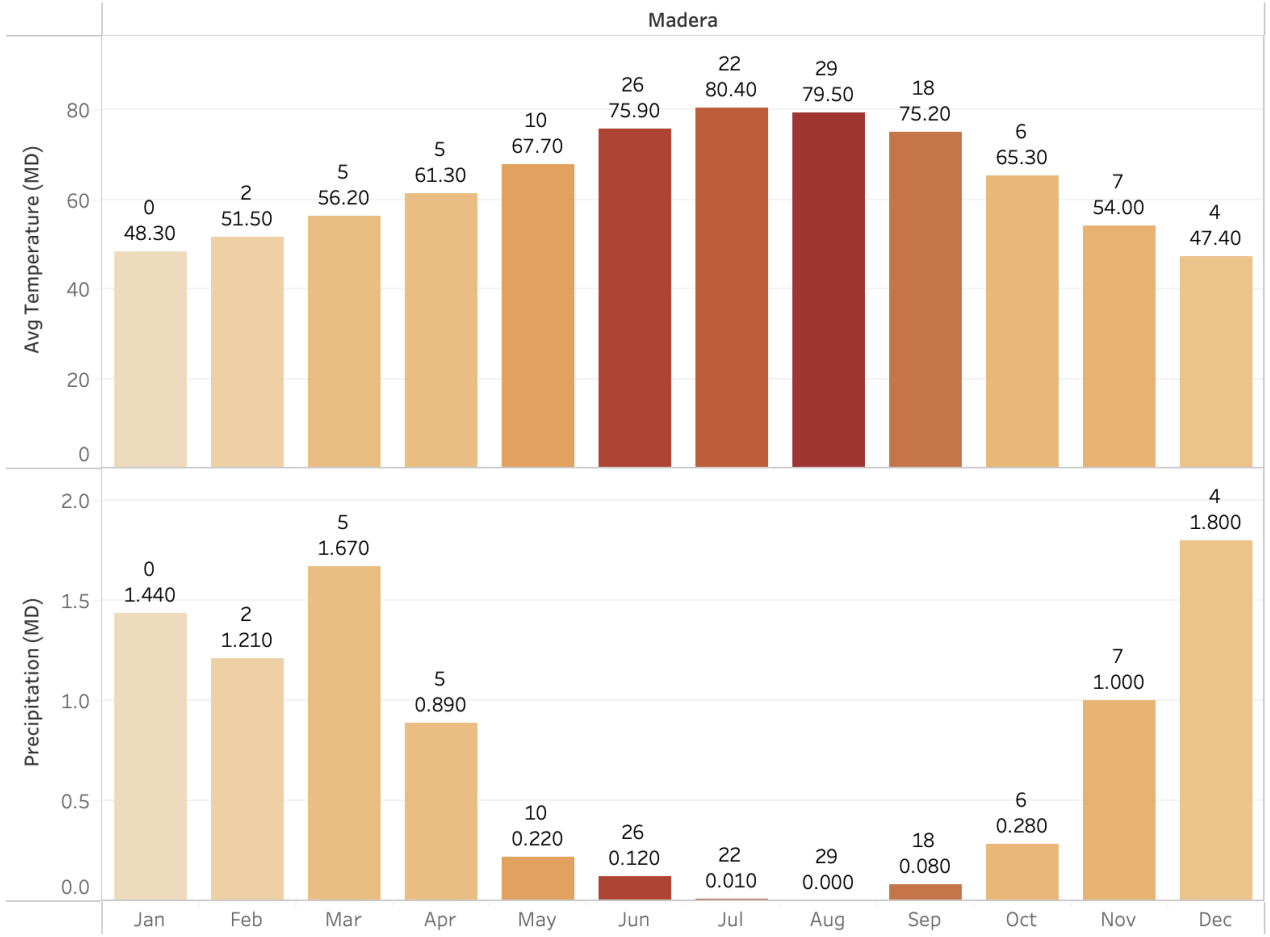
Correlation between sentiment (N>30) and Avg Temp:

fear: 0.910 | anger: 0.866 | joy: 0.859 | neutral: 0.709

Correlation between sentiment(N>30) and Precipitation:

anger: -0.801 | fear: -0.792 | joy: -0.752 | neutral: -0.729

Interplay of Climate Factors and Public Sentiment (Fear) of Bear Encounters - Madera (Yosemite)



Data Source: Twitter - tweets (N=464), NOAA - weather data (weather station: Madera Area, CA), 2010-2022.
This visualization illustrates the cumulative Average Temperature and Precipitation recorded in Madera. The color gradient represents the collective sentiment of fear inferred from tweets. Each data point is labeled with the respective cumulative fear sentiment. For the Average Temperature pane, labels also show the corresponding temperature. Similarly, for the Precipitation pane, labels show the corresponding precipitation sum.

TAHOE

Sentiments & Months One-way ANOVA:

F value: 8.112 | p value: 0.000

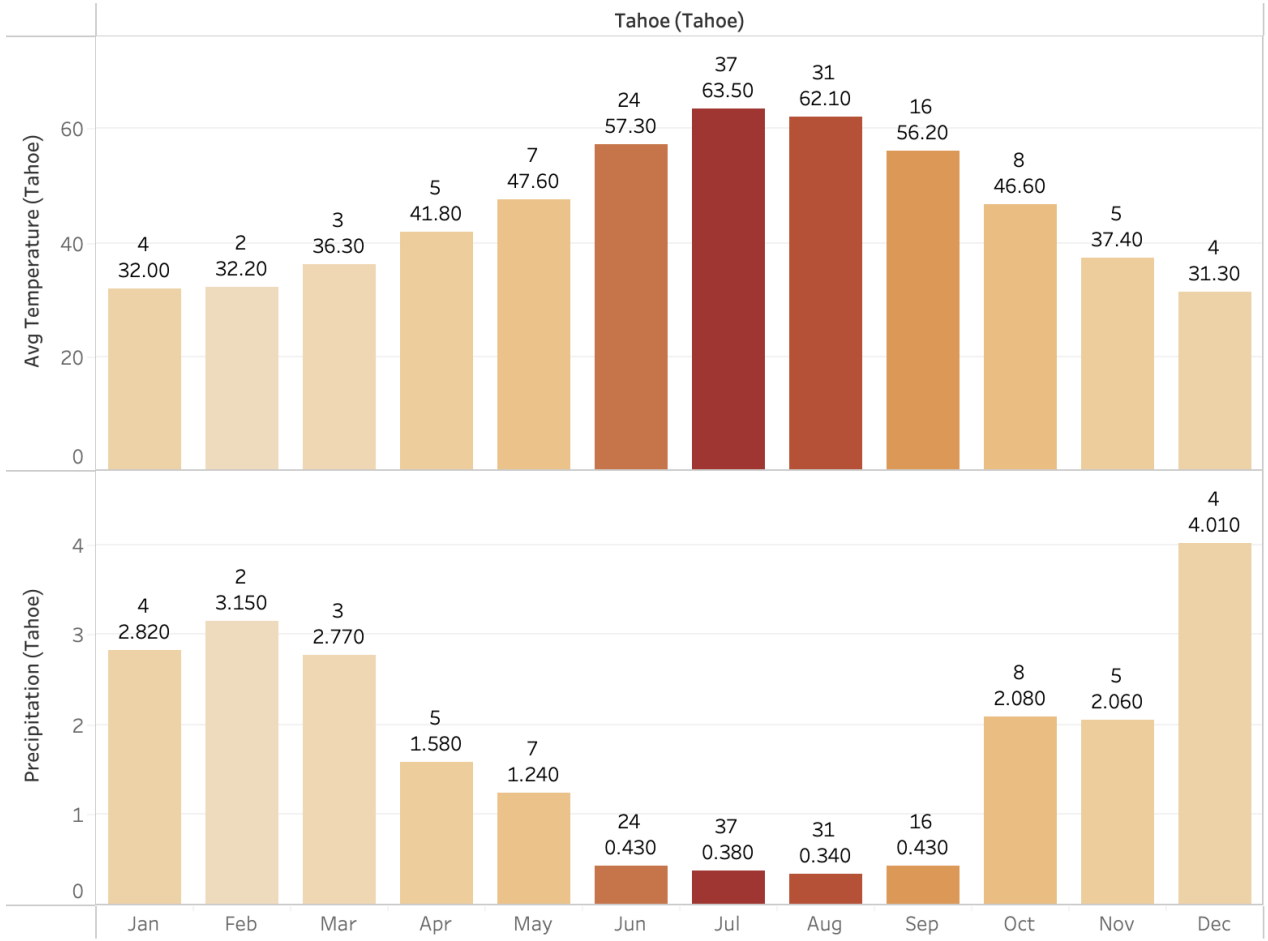
Correlation between sentiment (N>30) and Avg Temp:

fear: 0.916 | joy: 0.911 | neutral: 0.854

Correlation between sentiment(N>30) and Precipitation:

fear: -0.799 | joy: -0.770| neutral: -0.725

Interplay of Climate Factors and Public Sentiment (Fear) of Bear Encounters - Placer + EL Dorado (Tahoe)



Data Source: Twitter - tweets (N=447), NOAA - weather data (weather station: SOUTH LAKE TAHOE AP, CA), 2010-2022.
This visualization illustrates the cumulative Average Temperature and Precipitation recorded in South Lake Tahoe. The color gradient represents the collective sentiment of fear inferred from tweets. Each data point is labeled with the respective cumulative fear sentiment. For the Average Temperature pane, labels also show the corresponding temperature. Similarly, for the Precipitation pane, labels show the corresponding precipitation sum.

BIG BEAR AREA

Sentiments & Months One-way ANOVA:

F value: 26.736 | p value: 0.000

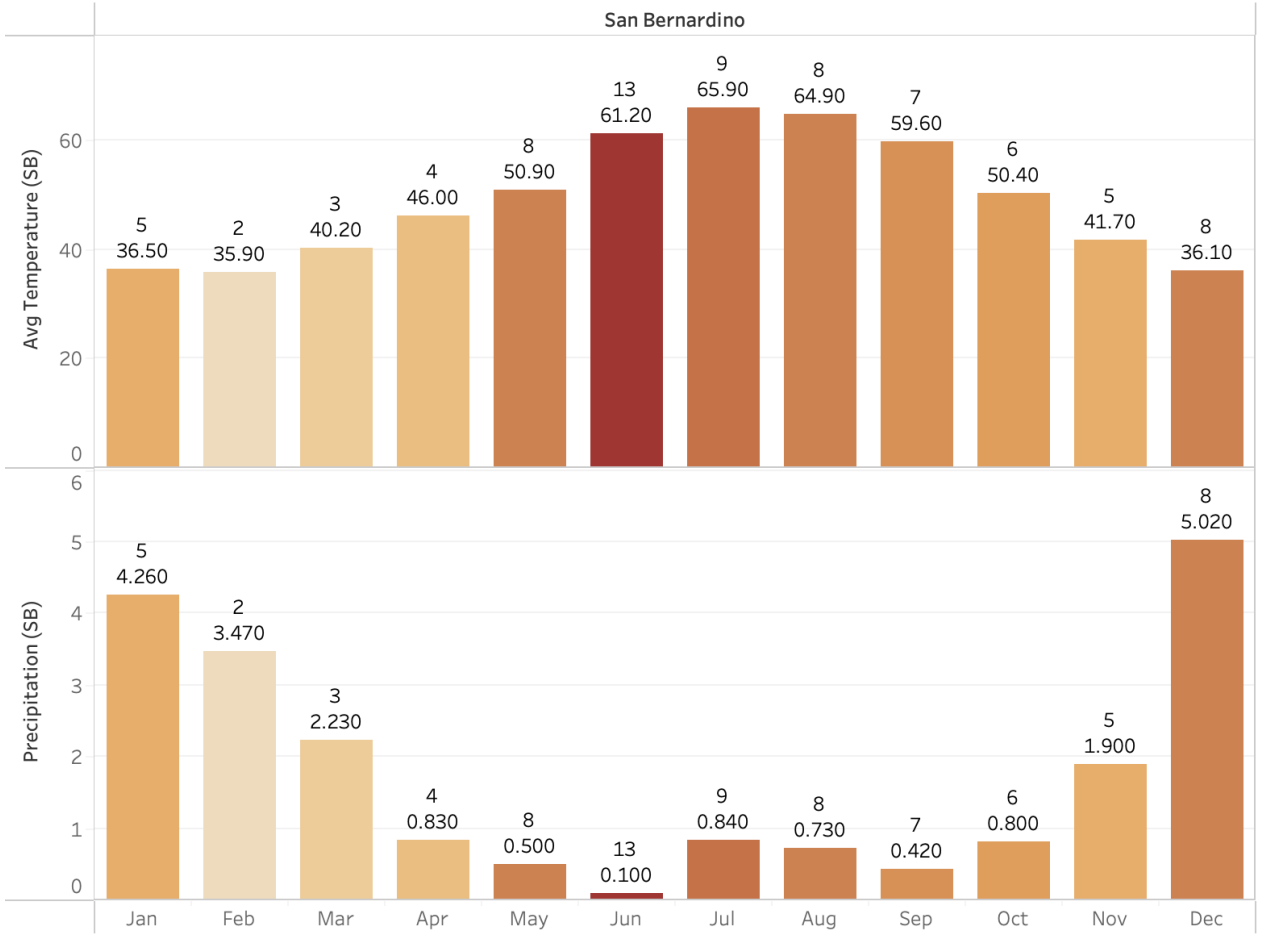
Correlation between sentiment (N>30) and Avg Temp:

fear: 0.690 | joy: 0.543 | neutral: 0.348 | anger: -0.312

Correlation between sentiment(N>30) and Precipitation:

joy: -0.576 | anger: 0.558 | fear: -0.419 | neutral: -0.027

Interplay of Climate Factors and Public Sentiment (Fear) of Bear Encounters - San Bernardino (Big Bear Area)



Data Source: Twitter - tweets (N=345), NOAA - weather data (weather station: BIG BEAR LAKE, CA), 2010-2022.
This visualization illustrates the cumulative Average Temperature and Precipitation recorded in San Bernardino. The color gradient represents the collective sentiment of fear inferred from tweets. Each data point is labeled with the respective cumulative fear sentiment. For the Average Temperature pane, labels also show the corresponding temperature. Similarly, for the Precipitation pane, labels show the corresponding precipitation sum.

LIMITATIONS

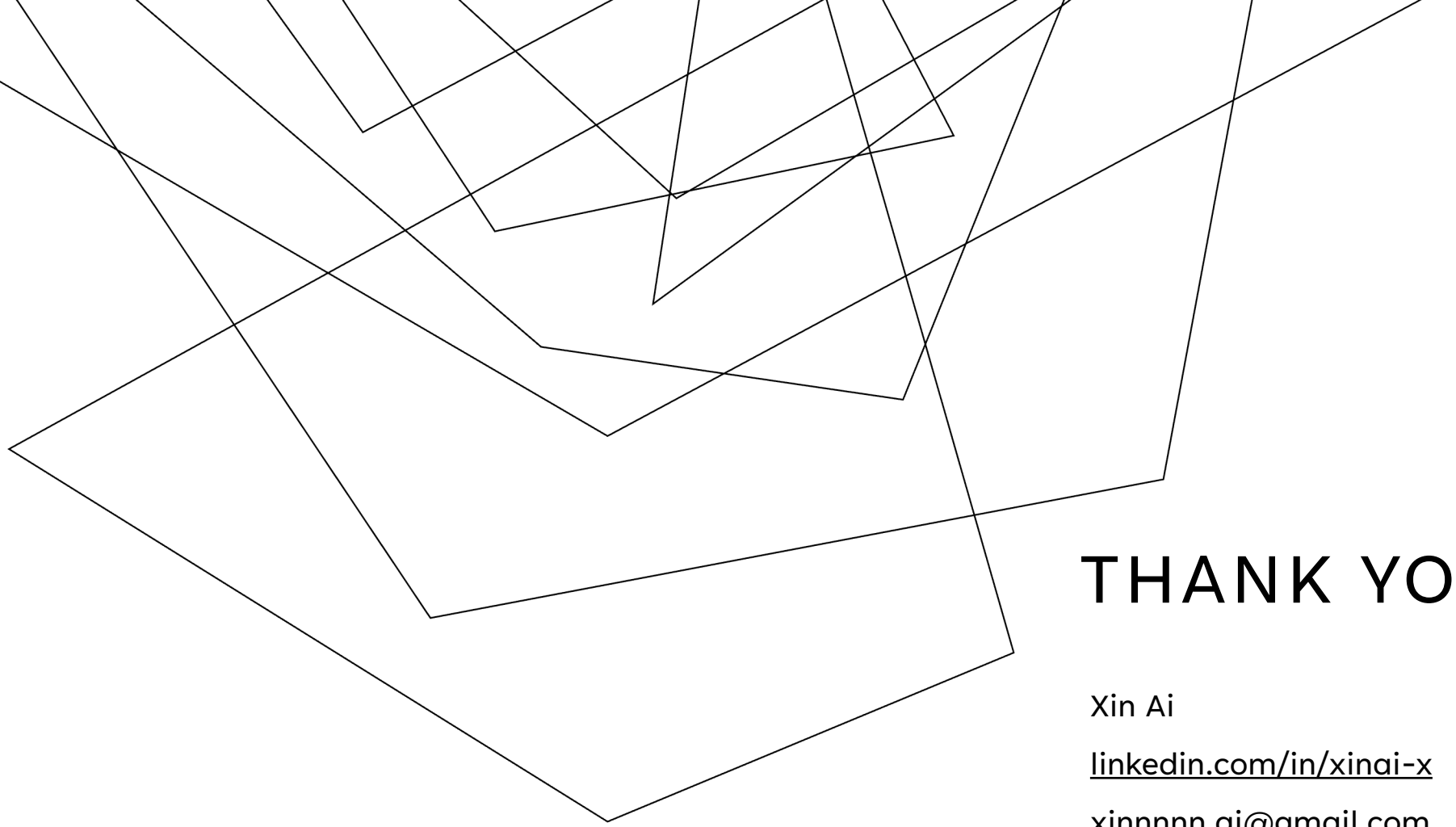
1. The geographic location of tweets does not immediately represent the specific location of bear sightings.
2. Tweets cannot filter out non-California bear encounters, as many of them are people's retrospective accounts.
3. The data points displayed in tweets are more than actual bear encounters, due to multiple people reporting the same bear encounter or discussing news events.
4. In some contexts, it is impossible to completely distinguish between a real bear and bear plush toys, certain individuals, or bears in game scenarios like #TheRevenant (game).

FURTHER DIRECTIONS

1. **Time Series Analysis:** Conduct predictive analysis on potential bear encounter frequencies and locations to inform preventive measures, aiding in the reduction of public panic and economic fallout.
2. **Exploratory Analysis:** Expand the scope to encompass other potential factors influencing bear encounters, such as incidences related to food attraction when humans are outdoors.
3. **Comparative Analysis:** Compare the characteristics of bear encounters with those of other wildlife encounters, derive unique insights and create comprehensive strategies for managing human-wildlife interactions.

DELIVERABLES

- Dashboard
- Database
- Tech blog
- Report (In progress)



THANK YOU

Xin Ai

linkedin.com/in/xinai-x

xinnnnn.ai@gmail.com

669-260-3982