Intro and Purpose

This analysis was done by Cassidy Bell, Neelam Prasad, Aaron Suarez, and Tyler Beringer. We analyzed the Kaggle dataset named **US Tornado Dataset 1950-2021.** Our goal was to explore historical tornado events.

Data and Data Cleaning

The dataset contains 67,558 entries with information on things like magnitude, injuries, and fatalities. To better uncover insights about trends and patterns, we narrowed in on specific parts of the data.

We noticed some problems with the dataset. Types had to be changed and columns had to be renamed. Plus, duplicates and magnitudes equal to -9 had to be removed.

Changing the type of date:

```
df['date'] = pd.to_datetime(df['date'])
```

Renaming the columns:

```
df = df.rename(columns={ 'yr': 'year', 'mo': 'month', 'dy': 'day', 'date': 'date', 'st': 'state', 'mag': 'tornado_magnitude', 'inj': 'injuries', 'fat': 'fatalities', 'slat': 'start_latitude', 'slon': 'start_longitude', 'elat': 'end_latitude', 'elon': 'end_longitude', 'len': 'tornado_length', 'wid': 'tornado_width' })
```

Removing the duplicates:

```
duplicates_to_remove = df[df.duplicated(keep='first')]
df = df.drop_duplicates(keep='first')
```

Removing the magnitudes equal to -9:

```
magnitudes_to_remove = df[df['tornado_magnitude'] == -9]
df = df.drop(magnitudes_to_remove.index)
```

By the time we concluded the data cleaning process, we had a quality dataset that was ready for use. It was time to start answering questions.

Research Questions

1. Which year has had the most number of tornadoes?
2. Which magnitude has had the most number of tornadoes?
3. Is there a correlation between the length and magnitude of the tornado?
4. Which magnitude has had the most fatalities?

Limitations and Biases

One of the limitations that our group came up with was when there were multiple tornadoes that touched down in a similar area, on the same day, with the same magnitude. This would be a limitation because it would make it difficult to show on the map since the drop points only show the information for one tornado at a time. This could cause the dataset to not be fully complete as some items would be not included on the dataset.

A bias that our group came up with was a population bias. This is because a tornado needs to be reported and in order to do so, someone needs to be aware that it occurred. There is a possibility that if a tornado touched down in a remote area, only the surrounding areas would know that it hit in the vicinity. There would be no correct report of the touchdown location so the data would not be completely accurate.

Conclusion

This was an interesting analysis which revealed unique insights about US tornadoes.