FINAL DATA ANALYSIS REPORT

Exploratory Data Analysis on US Accidents Dataset (2016–2021)

By Himanshu Kushwaha

Dataset overview

Dataset Name - US Accidents (2016 - 2023)

Source - Kaggle, compiled by Sobhan Moosavi Real-time data from traffic service

Dataset Size - 7.7 millions records (46 columns)

Purpose - To explore and analyze the US Accidents Dataset in order to identify patterns, trends, and key factors contributing to road accidents across the country. This EDA helps in understanding accident severity, timing, locations, and environmental conditions for better insights and future predictions.

Data Cleaning & Removing Nulls

- **To clean the dataset, we removed several columns due to high missing values or low relevance to the analysis. These included:**

  Amenity, Bump, Crossing, Give_Way, Junction, No_Exit, Railway, Roundabout, Station, Stop,Traffic_Calming, Traffic_Signal, Turning_Loop, Sunrise_Sunset, Civil_Twilight, Nautical_Twilight, Astronomical_Twilight, Wind_Chill(F), Humidity(%), Pressure(in), Visibility(mi),Wind_Direction, Wind_Speed(mph), Precipitation(in), Weather_Condition

- **Handling Missing Values**

  After dropping unnecessary columns, we checked for null values and handled them by either filling or removing, ensuring a clean dataset for analysis.

```
End_Lng                3402762
End_Lat                3402762
Temperature(F)          163853
Weather_Timestamp       120228
Airport_Code             22635
Street                   10869
Timezone                  7808
Zipcode                   1915
City                       253
Description                  5
ID                           0
Severity                     0
Source                       0
End_Time                     0
Start_Time                   0
Start_Lng                    0
Start_Lat                    0
Distance(mi)                 0
Country                      0
County                       0
State                        0
dtype: int64
```
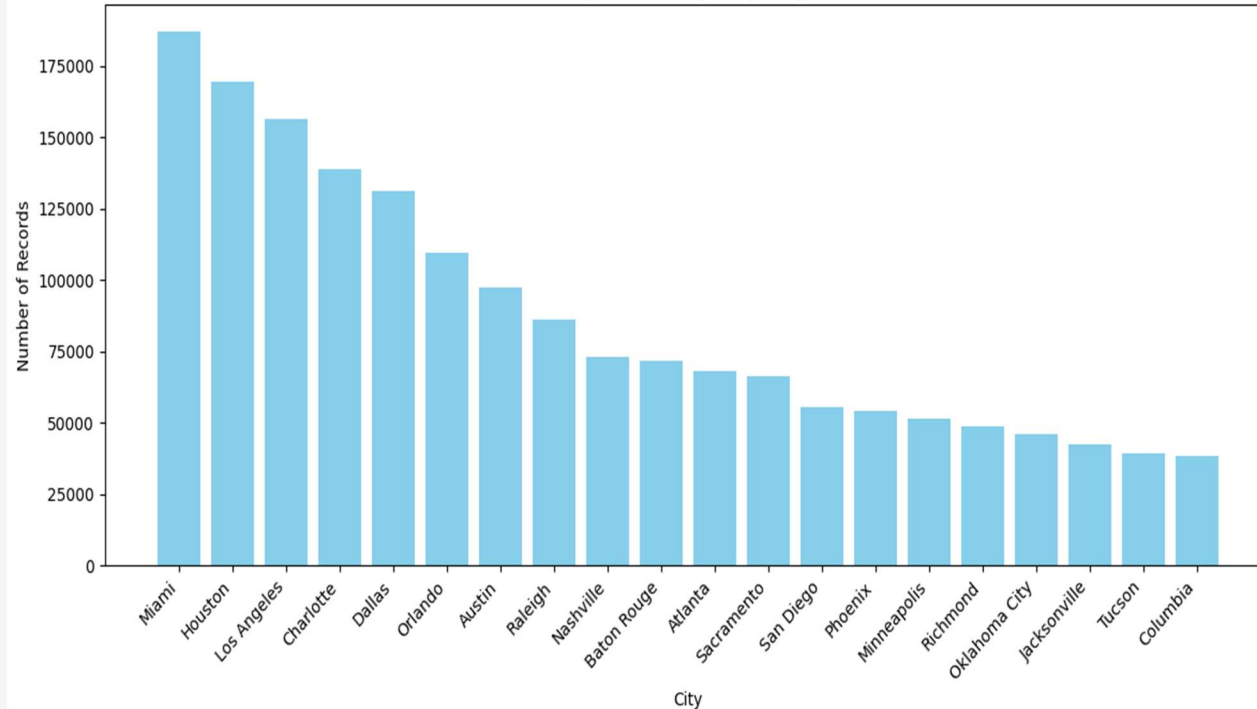
Road Accidents in Major U.S. Cities"

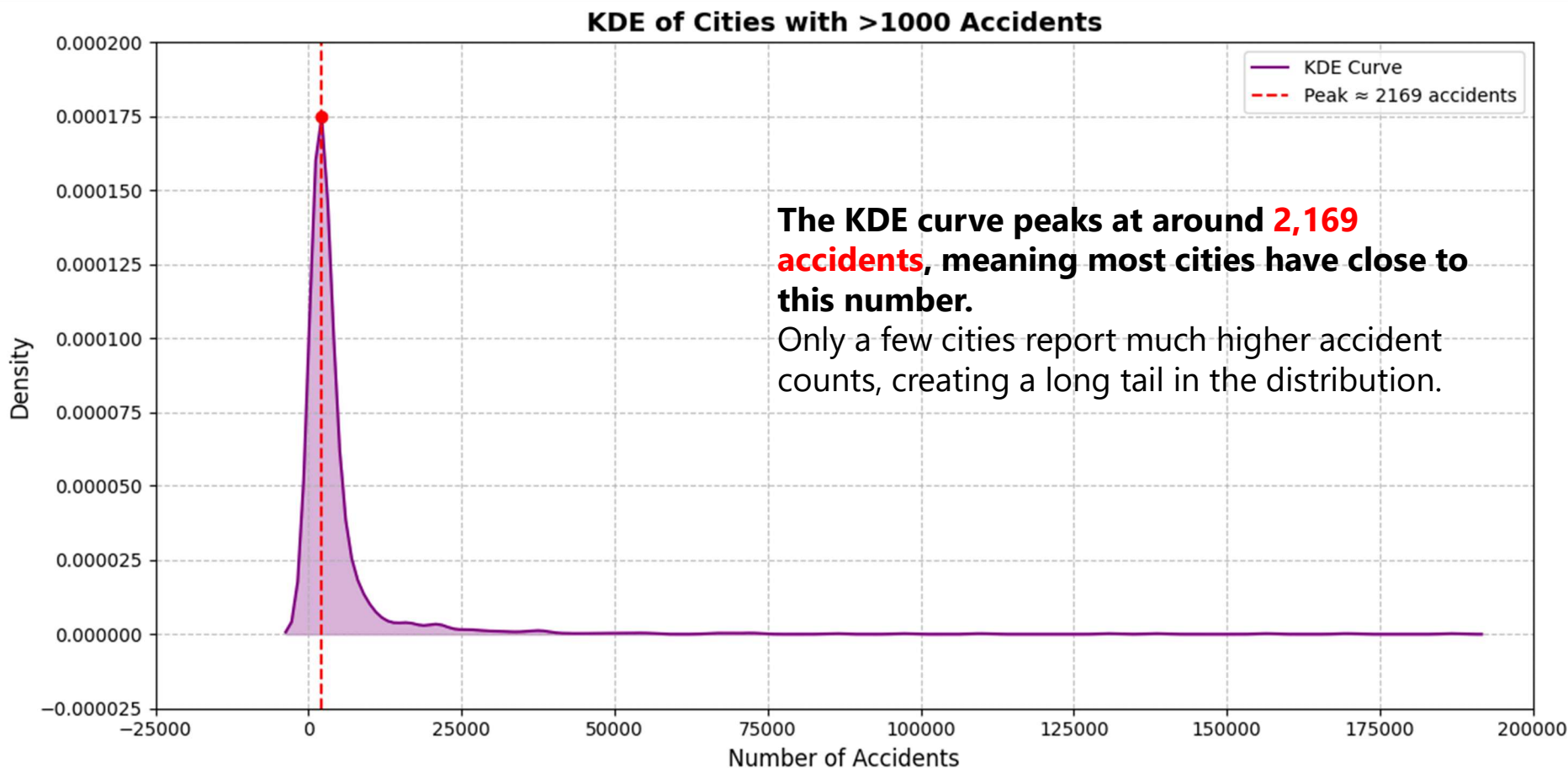**Top 5 cities with the highest number of road accidents:**

- **Miami** – 186,949 accidents
- **Houston** – 178,011 accidents
- **Los Angeles** – 166,043 accidents
- **Charlotte** – 145,031 accidents
- **Phoenix** – 112,511 accidents

Although **New York City** is the most populated U.S. city, it ranks only **45th** with **25,580 accidents**, indicating a lower accident rate relative to its size.
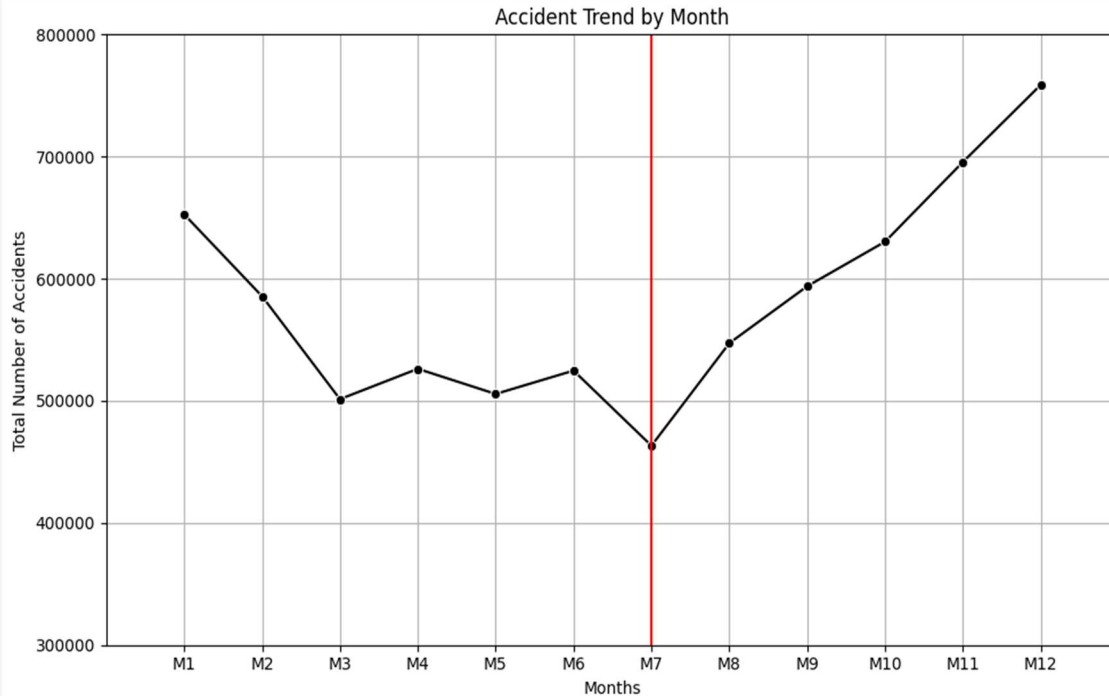

Top 20 Cities by Frequency

Distribution of Road Accident Counts Across Cities (KDE Curve)

KDE of Cities with >1000 Accidents

Legend:
- KDE Curve
- Peak ≈ 2169 accidents

The KDE curve peaks at around 2,169 accidents, meaning most cities have close to this number.
Only a few cities report much higher accident counts, creating a long tail in the distribution.

Monthly Trend of Road Accidents
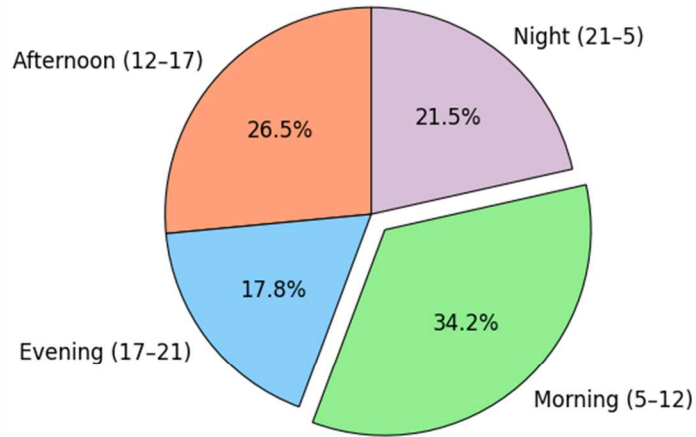

Accident Trend by Month

The line plot showing **month vs. accident count** reveals a clear **seasonal pattern** in road accidents:

January to July: A **steady decline** in accidents is observed, possibly due to favorable weather conditions or increased awareness campaigns early in the year.

August to December: A **gradual rise** starts in August, peaking in **December**, which has the **highest number of accidents**—likely due to increased travel, festivals, and foggy weather.

The line plot of month vs. accident count shows a clear seasonal trend. Accidents are lowest in **July (Month 7)** and gradually increase towards the end of the year. **December records the highest number of accidents**, indicating a possible link to winter and holiday travel.

Monthly Accident Patterns Across Different Years

Heatmap of Events by Year and Month

Year	January	February	March	April	May	June	July	August	September	October	November	December
2016	7	985	6330	18088	17610	30538	45780	56457	54698	55669	65040	59619
2017	55713	51679	57315	48115	41136	46272	42459	80955	76902	74975	70127	71642
2018	75093	71535	74484	72657	76529	67831	65802	76436	73354	87270	82156	70279
2019	79368	74631	69437	73020	73898	65525	68257	75269	87699	106914	81920	98364
2020	92037	85848	86054	87845	87700	100470	35227	37297	82912	127266	166574	172368
2021	139848	141624	83121	82962	92959	118555	102010	108795	121633	120728	141409	158789
2022	101801	121876	104312	143523	115782	95559	103728	111997	96617	57503	88386	127722
2023	108815	37428	20309	0	0	0	0	0	0	0	0	0

Distribution of Accidents by Temperature Category

- **7 AM – 9 AM** and **4 PM – 6 PM**, with **over 500,000 accidents** in each time block.
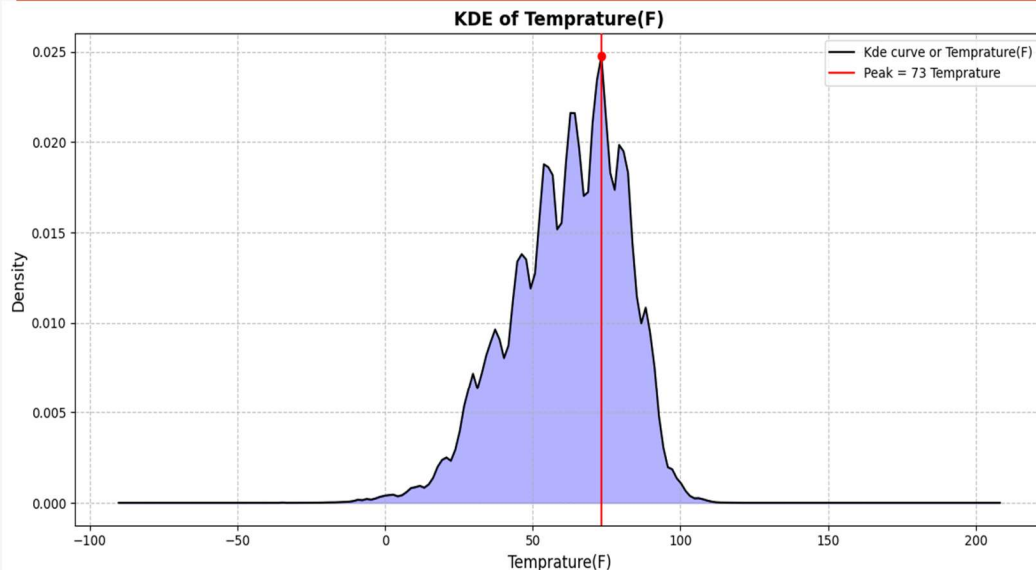- **Fewer accidents** occur during **noon to 4 PM** (less than 100000) with the **lowest around 3 PM – 4 PM**.

- Most accidents occur during **daylight hours (5–17)** — totaling over **60%**.

- **Night + Evening** still contribute a significant **39%**.
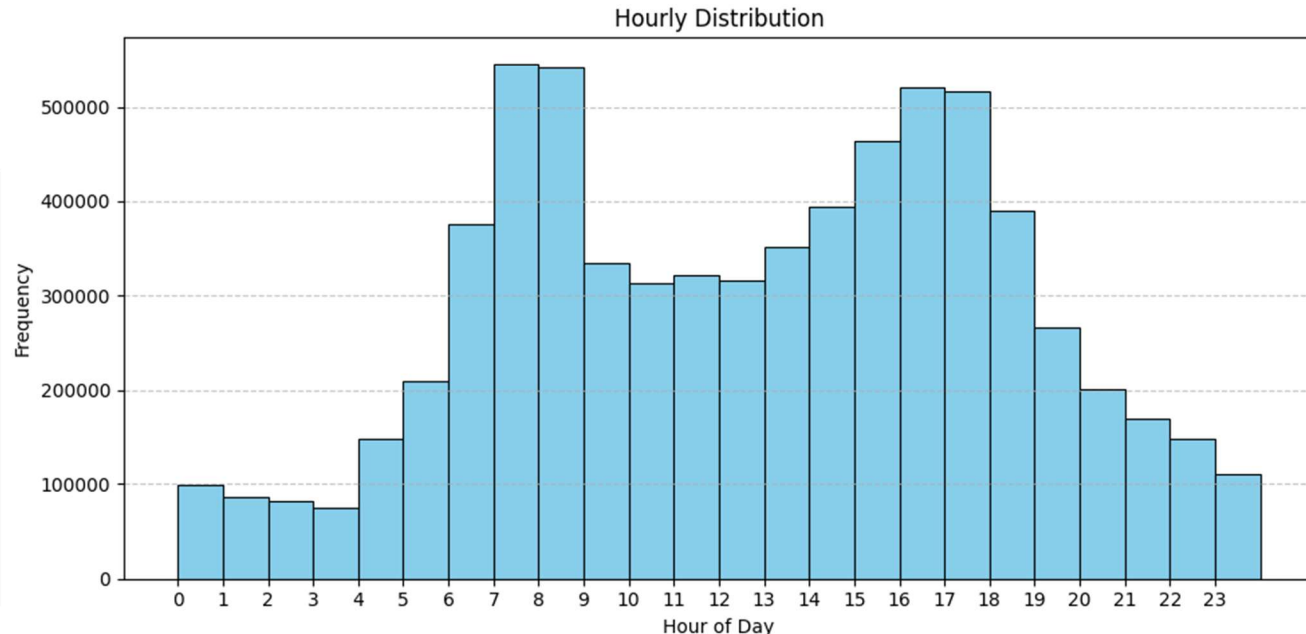
Geographic Distribution of Accidents in the US

- The map visualizes accident density across the United States.

- **Darker or denser** regions indicate areas with higher accident frequency.

- Urban centers and major highways show intense clustering, revealing traffic-heavy zones with more incidents.
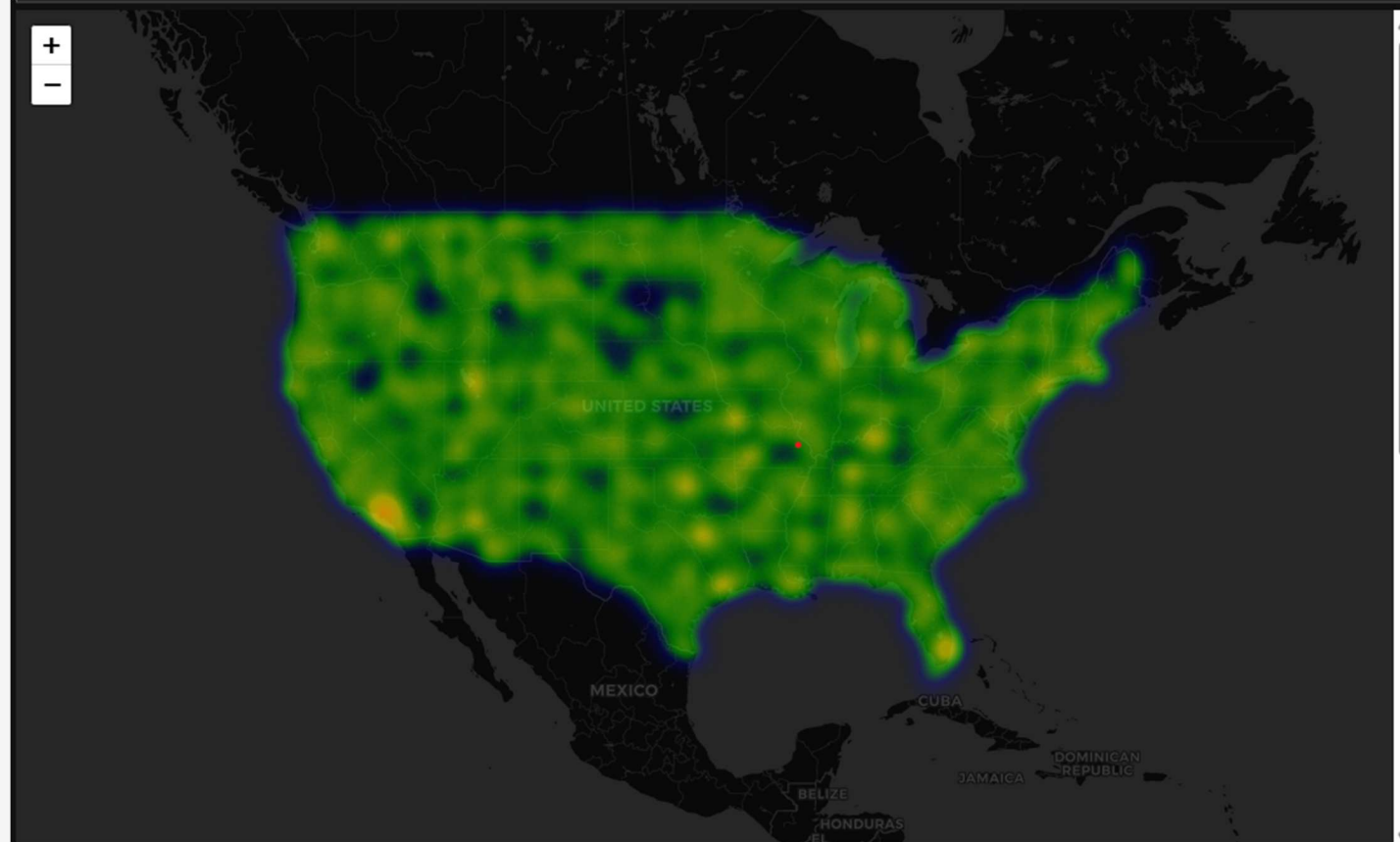
Impact of Temperature on Accident Frequency

KDE of Temprature(F)

- **Most accidents occur around 73°F**, a moderate and comfortable temperature when more people are likely to travel.
- **Fewer accidents happen during extreme temperatures** (very hot or very cold), likely because fewer vehicles are on the road during such conditions.

Temperature Verdict Distribution

Temperature Ranges
- Very Cold (≤ 32°F)
- Cold (33–50°F)
- Mid (51–68°F)
- Hot (69–85°F)
- Very Hot (> 85°F)

- **Hot and Mid-temperature categories** together account for the majority of accidents, each with over **2.4 million** incidents
- **Very Cold** and **Very Hot** categories have the fewest, with around **0.6M–0.8M** cases, indicating reduced road activity in extreme weather.
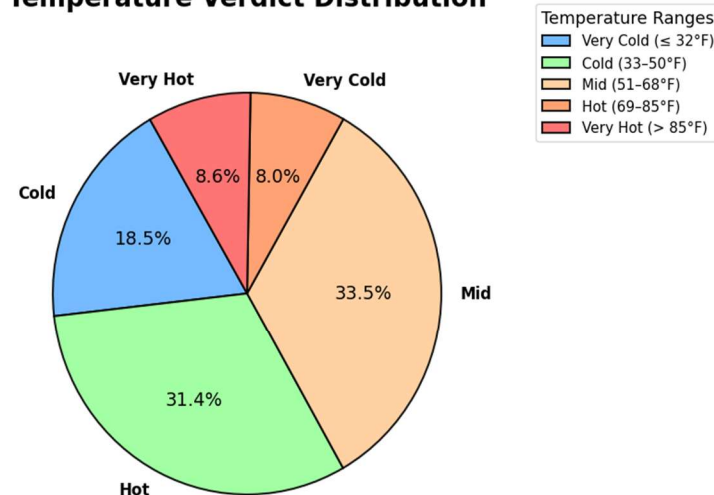
Key Findings & Trends

- **Top Cities:** Miami, Houston, Los Angeles, Charlotte And Pheonix report the most accidents, with City Miami recording the highest number that is 186,949 accidents.

- **Monthly Pattern:** Accidents peak in December and are lowest in July, first the accident decrease from January to July then from July to December it increases Linearly.

- **Yearly Trend:** 2021 recorded the highest number of accidents, crossing 1.2 million cases.

- **Peak Hours:** More than 500,000 accidents occurred during 7–9 AM and 4–6 PM, aligning with rush hours.

- **Low Noon Risk:** Fewer accidents occur between 12–4 PM, with the minimum around 3–4 PM.

- **Temperature Effect:** KDE reveals most accidents happen around 73°F, the most frequent temperature.

- **Temp Category Impact:** Combined Mid and Hot categories account for over 1.2 million accidents.

- **Regional Clusters:** Eastern and southern U.S. show denser accident concentrations on the geographical map

Thank you