Data Analysis Presentation

On

“Car Accidents Trends in USA”

Arnav Katyayan
November, 2022

Analytical Goals

- To identify trends in Car Accidents in USA based on records from 2016 to 2021.
- To analyze the distribution of Counts of Accidents over time and physical conditions

N.B. — Python Notebook created as a part of this analysis can be found here.

About the Dataset

- Dataset has been taken from Kaggle.
- It contains Data about car accidents in USA over the period of 5 Years from 2016 to 2021.
- It covers 49 States of US, except New York.
- Data is missing for January, 2016.
- It has about 2.8M records.

About the Dataset (Continued)

- Size of Dataset -> 133,731,074
- Shape of Dataset -> Rows =>2,845,342; Column=>47
- Description of Numerical Columns

	Severity	Start_Lat	Start_Lng	End_Lat	End_Lng	Distance(mi)	Number	Temperature(F)	Wind_Chill(F)	Humidity(%)	Pressure(in)	Visibility(mi)	Wind_Speed(mph)	Precipitation(in)
count	2.845342e+06	2.845342e+06	2.845342e+06	2.845342e+06	2.845342e+06	2.845342e+06	1.101431e+06	2.776068e+06	2.375699e+06	2.772250e+06	2.786142e+06	2.774796e+06	2.687398e+06	2.295884e+06
mean	2.137572e+00	3.624520e+01	-9.711463e+01	3.624532e+01	-9.711439e+01	7.026779e-01	8.089408e+03	6.179356e+01	5.965823e+01	6.436545e+01	2.947234e+01	9.099391e+00	7.395044e+00	7.016940e-03
std	4.787216e-01	5.363797e+00	1.831782e+01	5.363873e+00	1.831763e+01	1.560361e+00	1.836009e+04	1.862263e+01	2.116097e+01	2.287457e+01	1.045286e+00	2.717546e+00	5.527454e+00	9.348831e-02
min	1.000000e+00	2.456603e+01	-1.245481e+02	2.456601e+01	-1.245457e+02	0.000000e+00	0.000000e+00	-8.900000e+01	-8.900000e+01	1.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00
25%	2.000000e+00	3.344517e+01	-1.180331e+02	3.344628e+01	-1.180333e+02	5.200000e-02	1.270000e+03	5.000000e+01	4.600000e+01	4.800000e+01	2.931000e+01	1.000000e+01	3.500000e+00	0.000000e+00
50%	2.000000e+00	3.609861e+01	-9.241808e+01	3.609799e+01	-9.241772e+01	2.440000e-01	4.007000e+03	6.400000e+01	6.300000e+01	6.700000e+01	2.982000e+01	1.000000e+01	7.000000e+00	0.000000e+00
75%	2.000000e+00	4.016024e+01	-8.037243e+01	4.016105e+01	-8.037338e+01	7.640000e-01	9.567000e+03	7.600000e+01	7.600000e+01	8.300000e+01	3.001000e+01	1.000000e+01	1.000000e+01	0.000000e+00
max	4.000000e+00	4.900058e+01	-6.711317e+01	4.907500e+01	-6.710924e+01	1.551860e+02	9.999997e+06	1.960000e+02	1.960000e+02	1.000000e+02	5.890000e+01	1.400000e+02	1.087000e+03	2.400000e+01

Data Processing

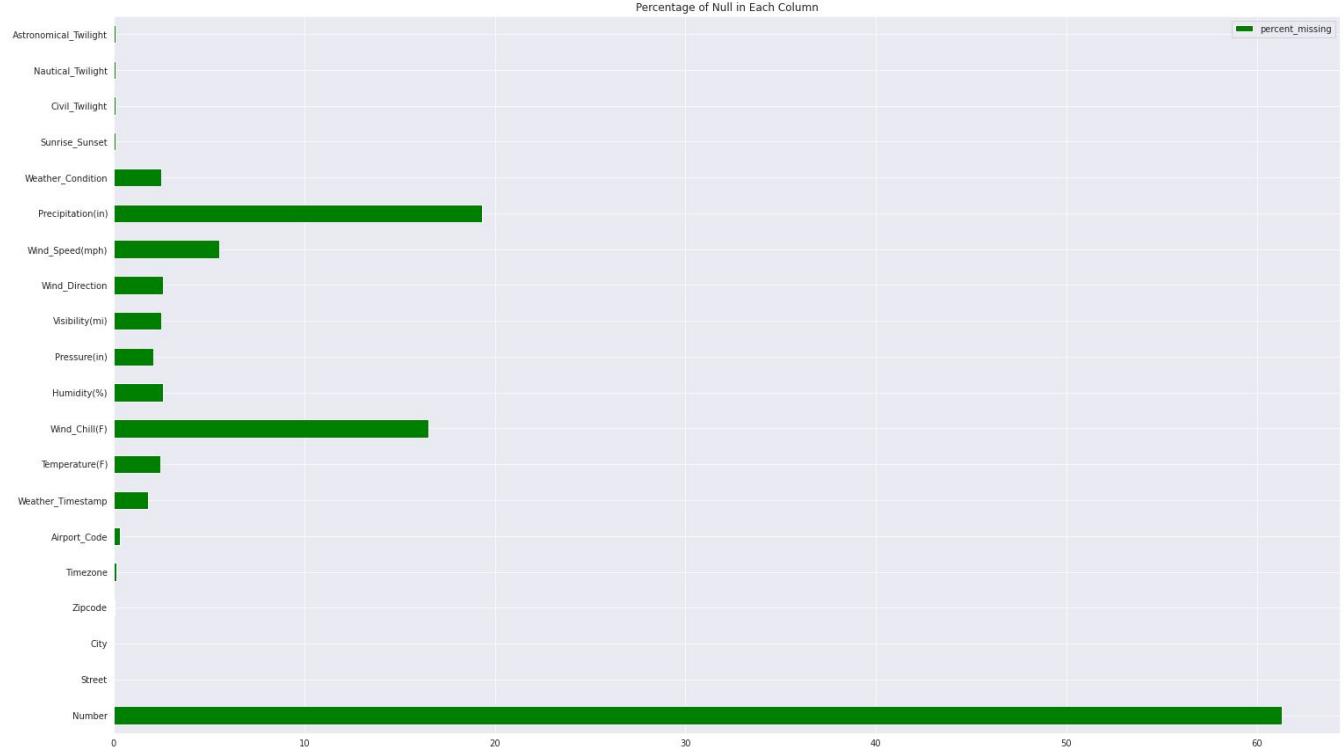
Data Processing involved the following

- Identification of columns with missing values, to exclude them from analysis.
- Identification of Duplicate values
- Correction of Data type for column containing

Data Processing

Identification of columns with missing values, to exclude them from analysis.

- Number Column (61.290%) had highest percentage of Null values followed by weather_condition(19.311%) and wind_chill(16.506%) columns.
- 20 columns out of total 47 columns contained Null Values.
- 17 columns viz. Contained Null less than 10%
    - Wind_Speed(mph) 5.551%
    - Wind_Direction 2.593%
    - Humidity(%) 2.568%
    - Weather_Condition 2.482%
    - Visibility(mi) 2.479%
    - Temperature(F) 2.434%
    - Pressure(in) 2.080%
    - Weather_Timestamp 1.783%
    - Airport_Code 0.335%
    - Timezone 0.128%
    - Sunrise_Sunset 0.100%
    - Civil_Twilight 0.100%
    - Nautical_Twilight 0.100%
    - Astronomical_Twilight 0.100%
    - Zipcode 0.046356%
    - City 0.004815%
    - Street 0.000070%


Percentage of Null in Each Column
percent_missing
Astronomical_Twilight
Nautical_Twilight
Civil_Twilight
Sunrise_Sunset
Weather_Condition
Precipitation(in)
Wind_Speed(mph)
Wind_Direction
Visibility(mi)
Pressure(in)
Humidity(%)
Wind_Chill(F)
Temperature(F)
Weather_Timestamp
Airport_Code
Timezone
Zipcode
City
Street
Number
0
10
20
30
40
50
60

Data Processing

Data Processing involved the following

- Identification of Duplicate values
  - 0 found.

```
[16]  1  temp_df0 = df0[df0.duplicated()].copy()
```

17s

```
[17]  1  temp_df0.size
```

0s

```
0
```

Data Processing

Data Processing involved the following

- Correction of Data type for column containing - ‘Dates’.

```
[98]  1  df0[['Start_Time','End_Time']].dtypes
```

```
Start_Time    object
End_Time      object
dtype: object
```

```
[99]  1  df0['Start_Time'] = pd.to_datetime(df0['Start_Time'])
```

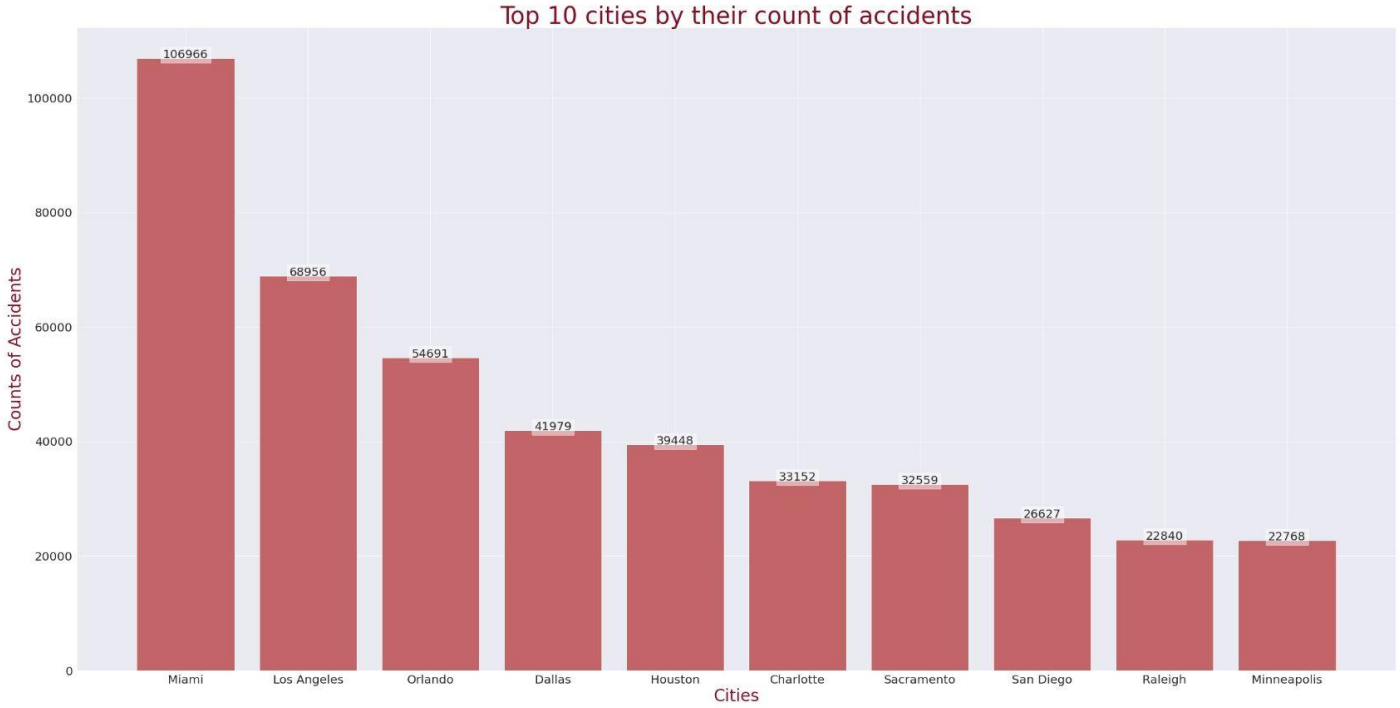
```
[100]  1  df0['End_Time'] = pd.to_datetime(df0['End_Time'])
```

```
[102]  1  df0[['Start_Time','End_Time']].dtypes
```

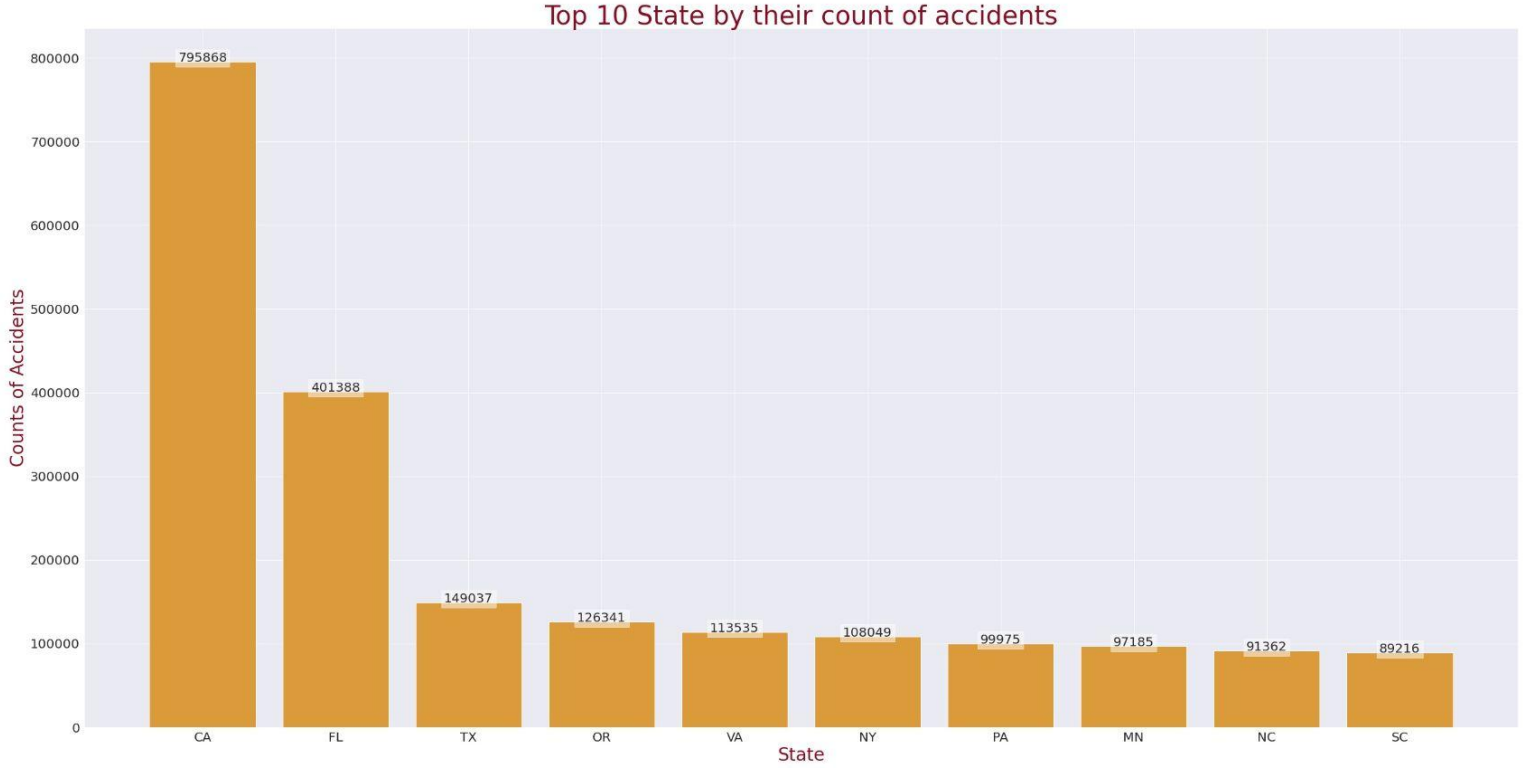
```
Start_Time    datetime64[ns]
End_Time      datetime64[ns]
dtype: object
```

Data Visualization

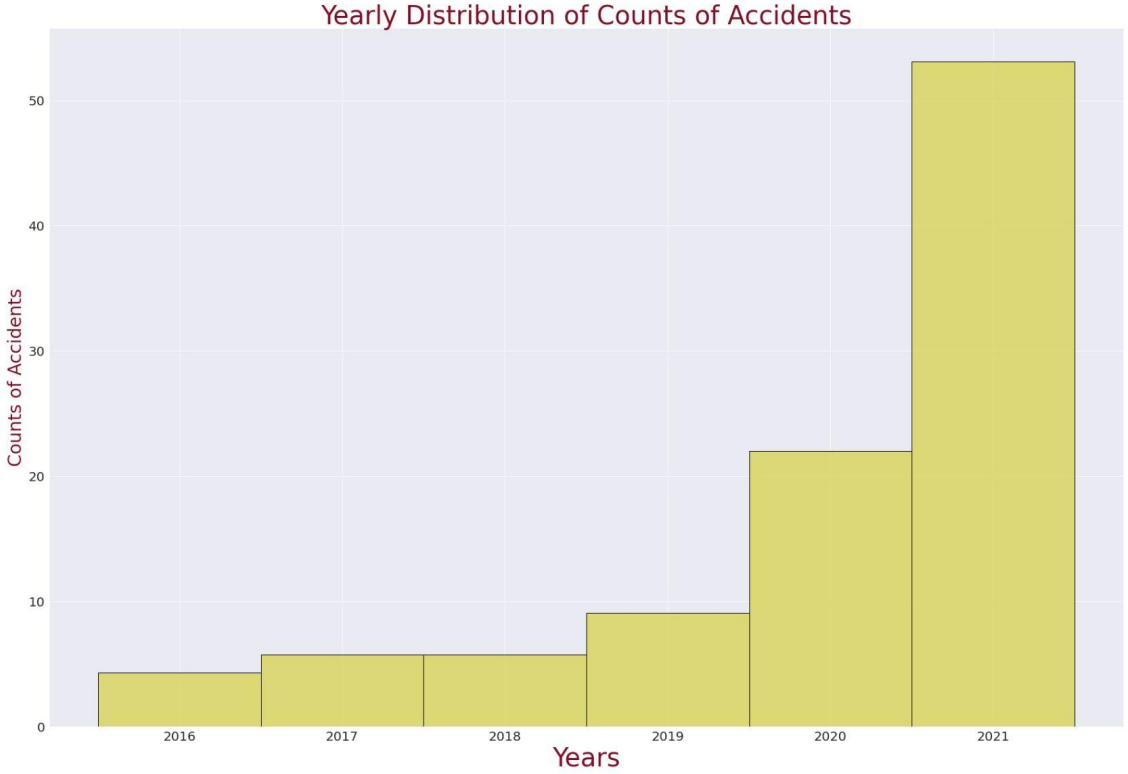
- Geographical distribution of Accidents
  - Top 10 cities by count of accidents.
  - Top 10 states by count of accidents.
- Time Dependence of Accidents
  - Yearly Distribution of Counts of Accidents
  - Monthly Distribution of Counts of Accidents
  - Weekly Distribution of Counts of Accidents
  - Hourly Distribution of Counts of Accidents (Overall)
  - Hourly Distribution of Counts of Accidents (Weekdays)
  - Hourly Distribution of Counts of Accidents (Weekends)
  - Friday's Analysis.
- Physical Conditions dependence of accidents.
  - Temperature Distribution during the accidents
  - Natural Lighting Conditions during the accidents



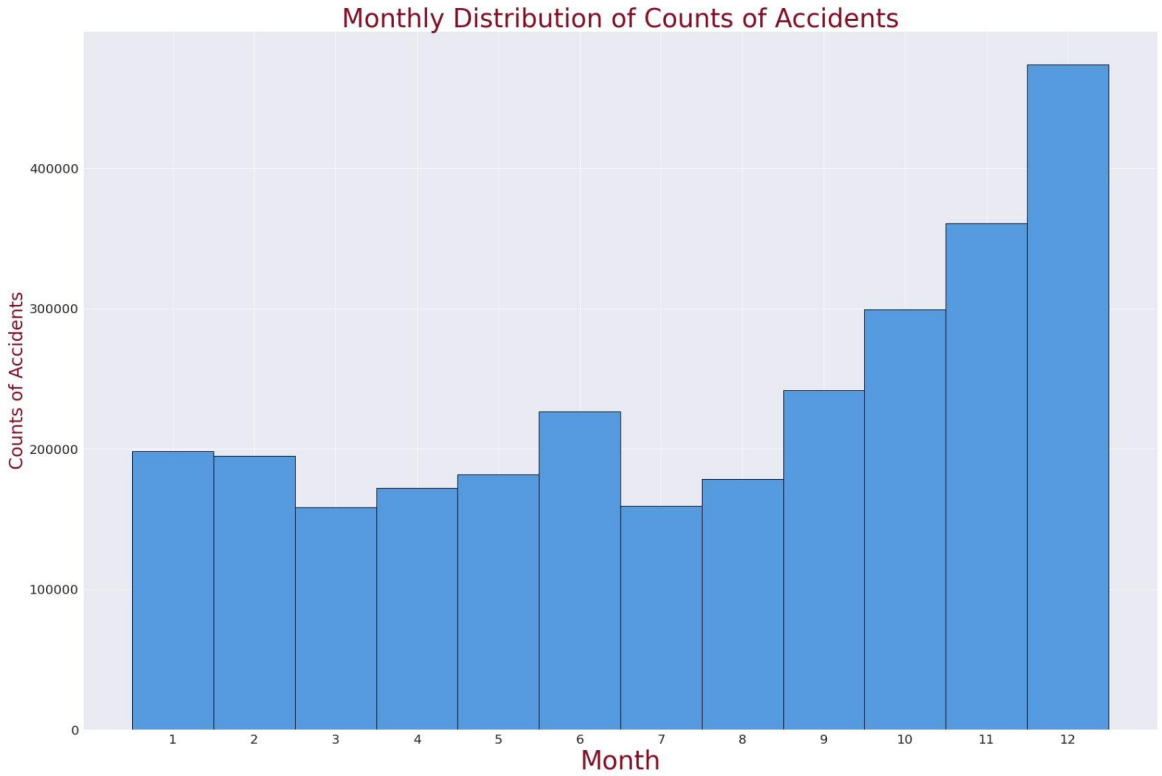
- Miami had highest accidents in the period of 2016 to 2021
- Top 10 count of accidents by cities varies from 1,06,966 to 22,768
- Difference between the highest and second highest accidents count by cities is nearly 38,000



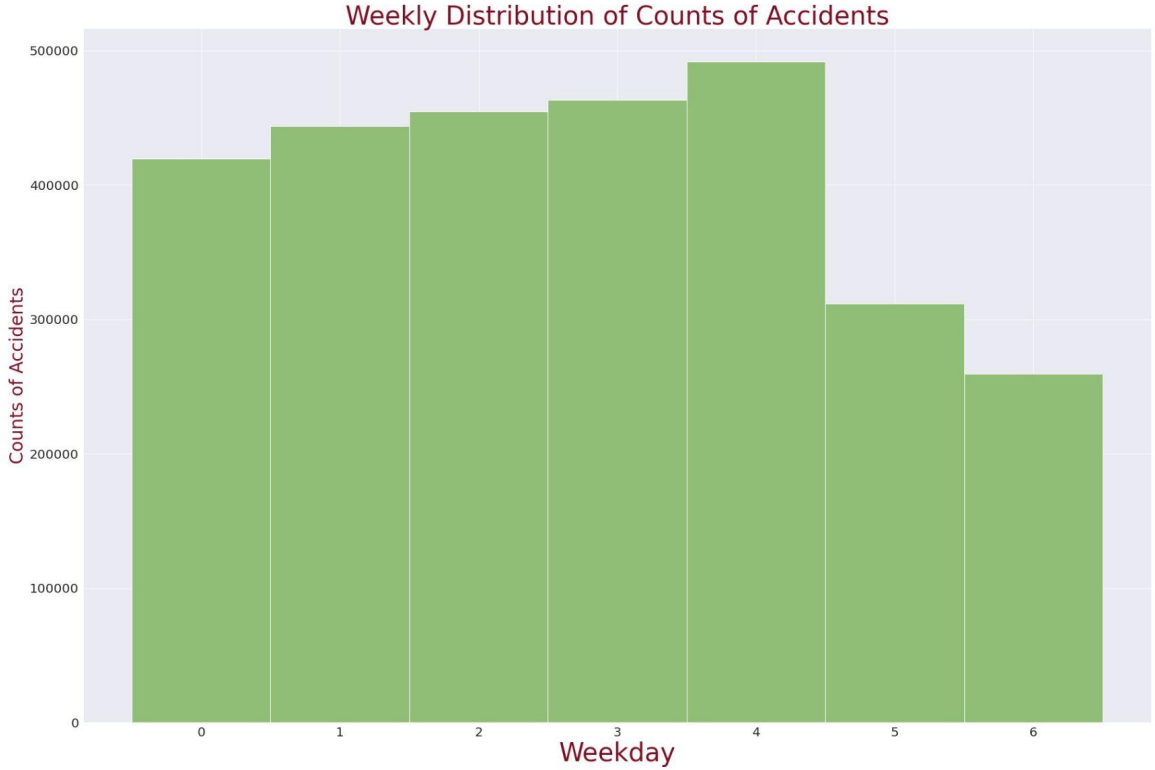
- Highest Accident count in California (CA)
- Difference between highest and second highest accidents is nearly 3,00,000
- Difference between second and third highest accidents is nearly 2,60,000



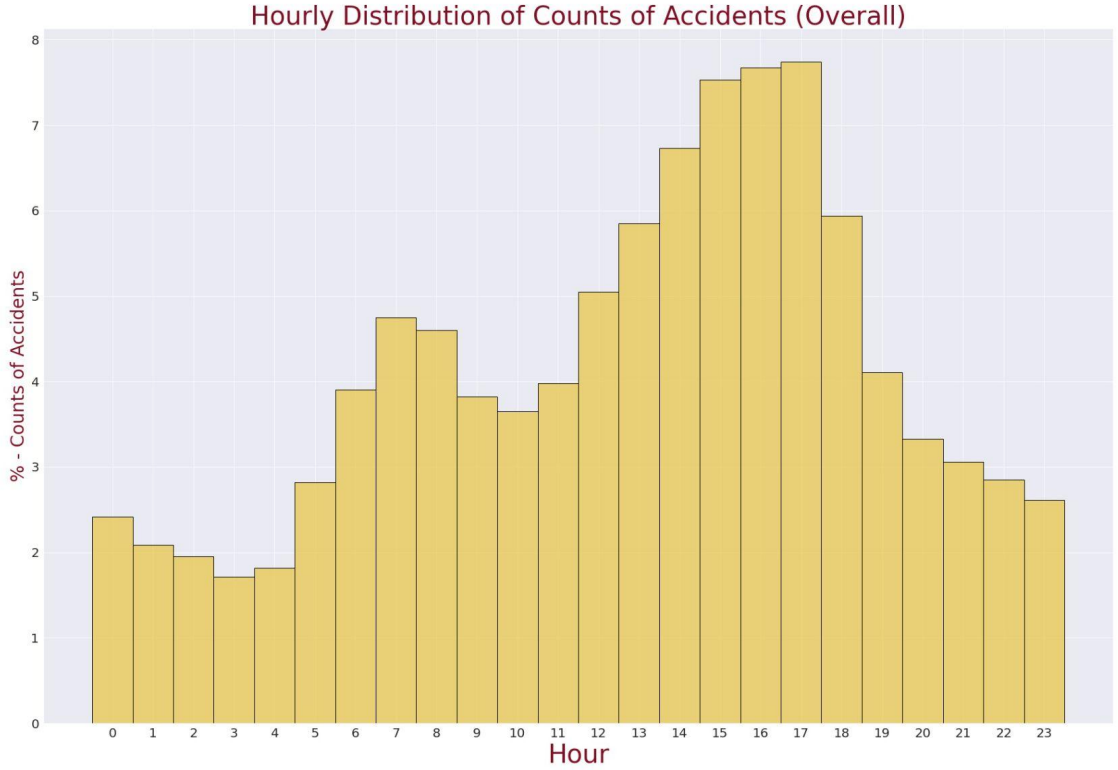
- Highest Accidents in 2021
- Massive reporting in 2021 as compared to previous years.
- Maybe the data has not been entirely reported for the previous years.



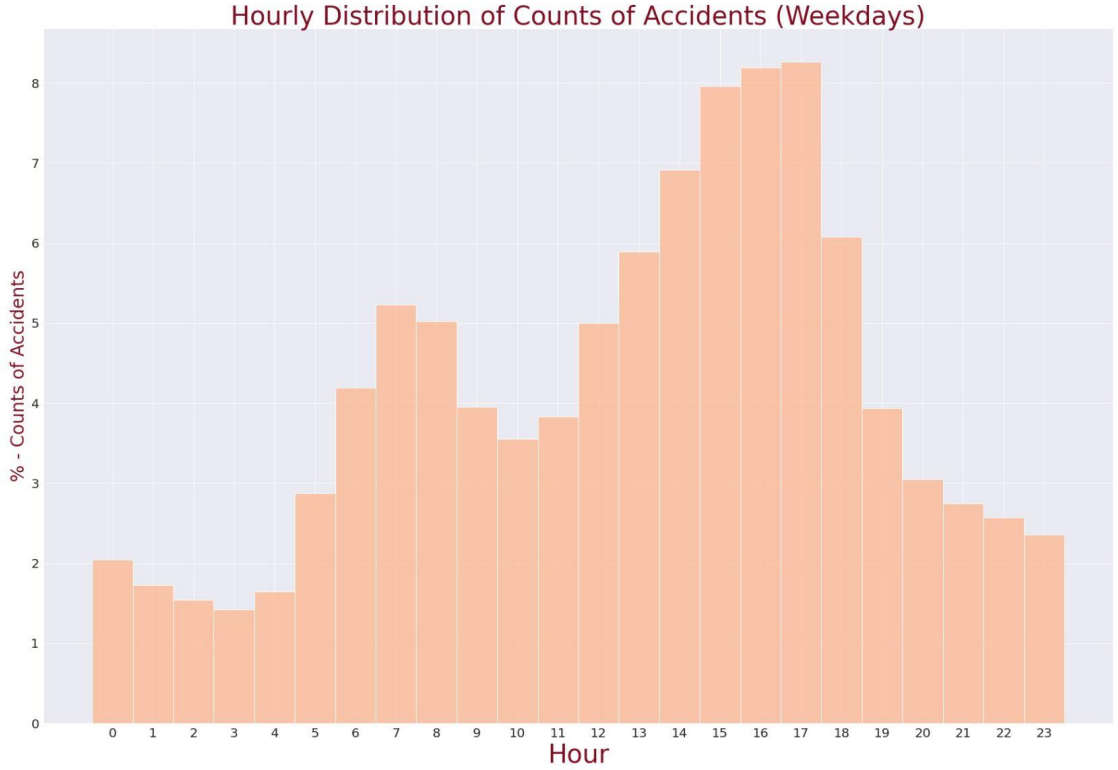
- Highest Accidents in the month of December
- Increasing Accidents in the winter months



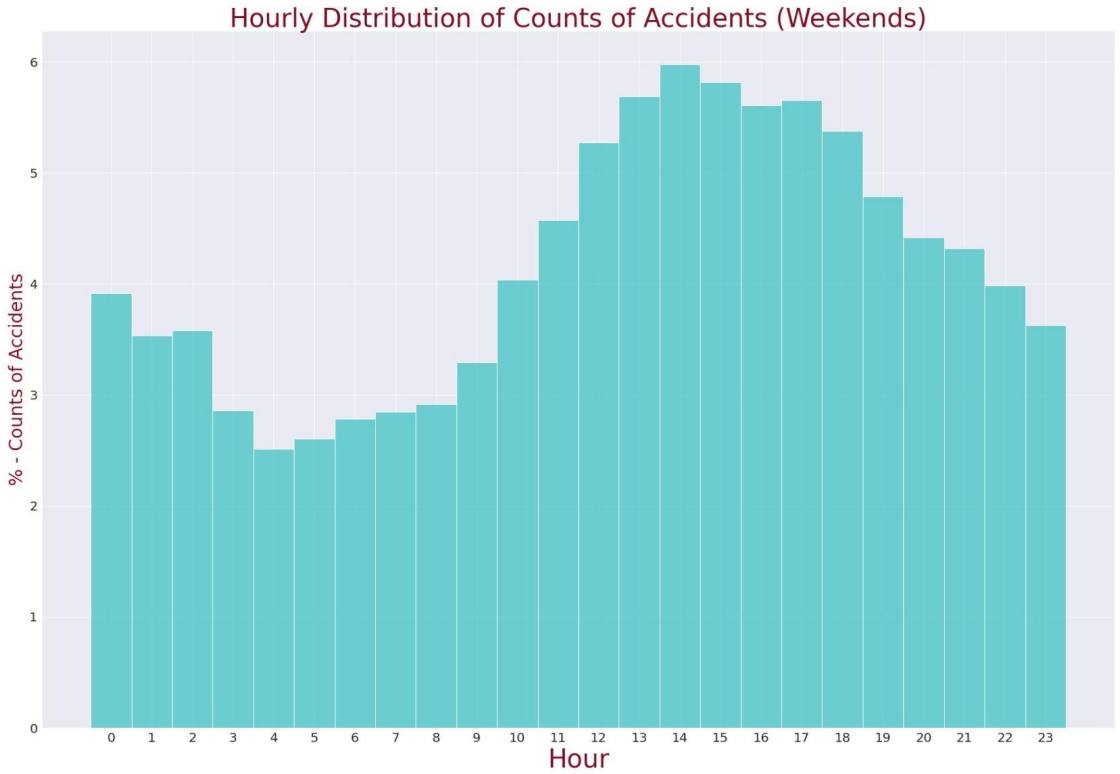
- Starting of the week is marked by 0 and represents Monday
- Most accidents on Friday(4)
- Higher Accidents on Weekdays than on Weekends



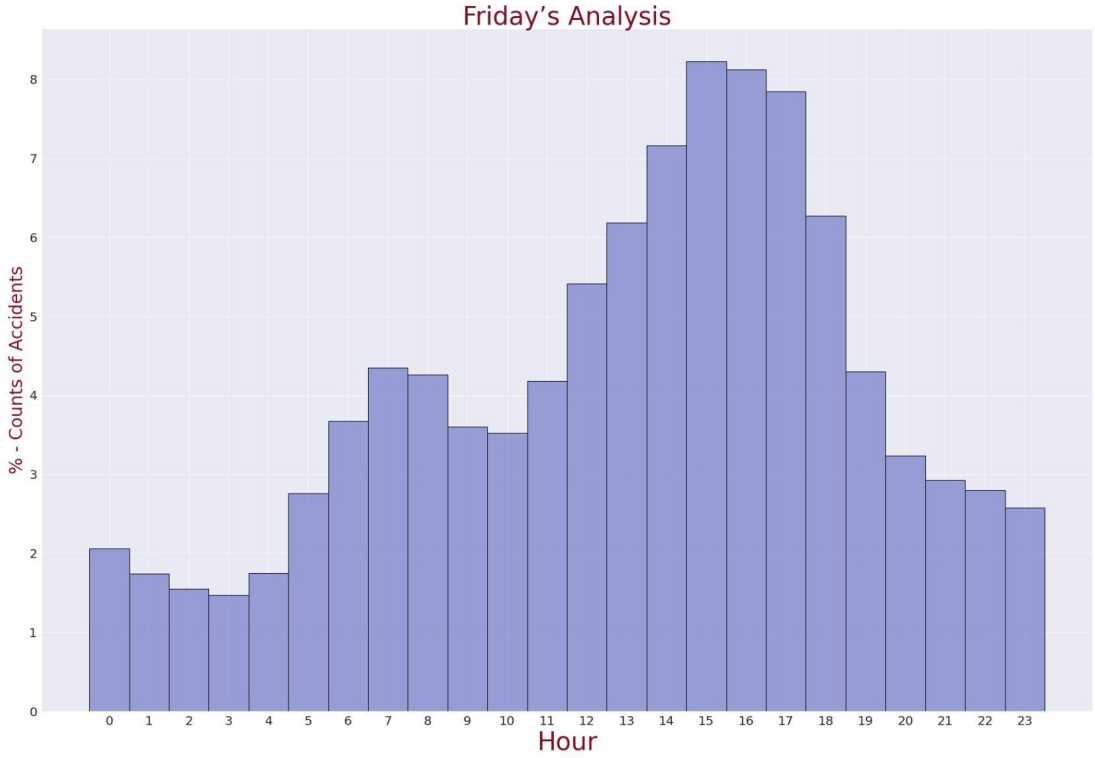
- Highest percentage of accidents occur during 14:00 to 18:00 Hrs
- Slightly higher accidents during morning 06:00 to 09:00 Hrs



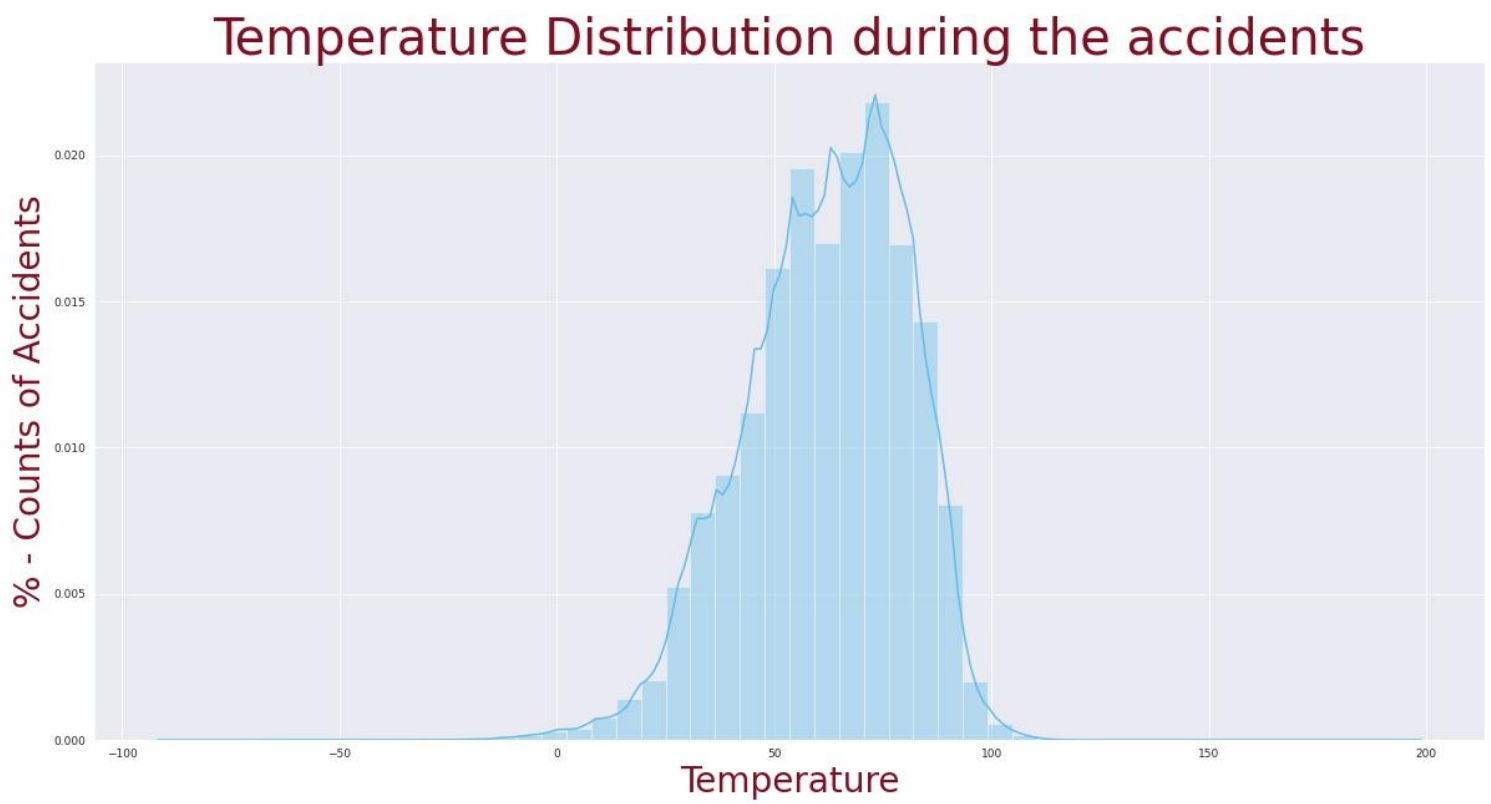
- Hourly accidents on Weekdays is higher at around 14:00 to 18:00 Hrs.
- Slightly higher accident percentage in the morning from 06:00 to 09:00 Hrs



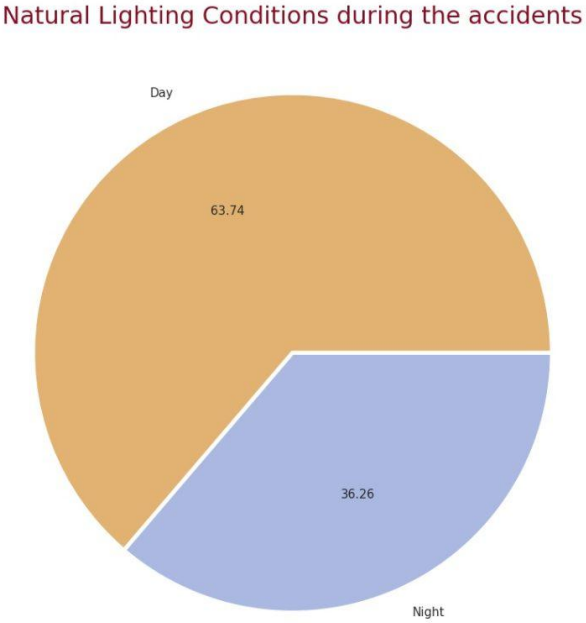
- Hourly Distribution still higher at at around 14:00 to 18:00 Hrs.
- Higher accidents in the evening as compared to Weekdays.



- Similar to weekday’s trends
- Higher percentage during the 15:00 to 17:00 Hrs as compared to entire week.



- Roughly between 50 to 90 Fahrenheit most accidents occur which is the normal temperature.
- No significant impact of temperature observed on accidents.



- 63.7% of accidents were reported during daytime, and rest 36.3% during night
- Lack of natural lighting seems to be a significant factor for accidents.

Final Notes

- There is a significant difference in accident reporting in Miami vs Other Cities
  - the data for Miami is relatively biasing the Analysis.
- The accident reported in California is highest, and has significant difference from Florida and Texas
  - creates a bias in the analysis towards California.
- The Accident reported in 2021 and that in 2020 has a major difference.
  - The data has not been recorded correctly for the previous years.
  - The analysis is biased for 2021 majorly.

Final Notes

- Accidents were highest reported in December
    - The observation is in sync with the observation for temperatures which reported that most accidents occur above 50 degree Fahrenheit and below 90 degree fahrenheit
    - The accidents increase over the months from October to December
- Weekdays have significantly higher accident rates than weekends.
- The accidents reporting nearly follows similar pattern everyday i.e. occurs in the evening from 15:00 to 18:00 Hrs.
    - Weekends have higher accident rates in the evening even after 17:00 Hrs.

Final Notes

- Most accidents have been reported on Fridays which followed similar pattern as any other weekday with slightly higher count percentage.
- Temperature varied from 50 to 90 degree Fahrenheit.
- Lighting conditions impact the accidents significantly

This analysis is open for queries and discussions.

You may refer below to contact the author.

Contact Details

Arnav Katyayan

arnavkatyayan121@gmail.com

November , 2022

Thank You for your Patience!