US Accidents & Severity

A study reverting in benefit to the authorities to better understanding traffic issues

Group 2

Research question

Which key factors are determining the level of accident severity in the US?

Sub-questions

- To what extent do weather conditions impact the severity of traffic accidents?
- To what extent does time conditions impact the severity of traffic accidents?
- To what extent do infrastructure conditions impact the severity of traffic accidents?

We also paid attention to:

- Geographical distribution
- Distance in Traffic congestion
- Duration of accidents

Data sources

- Moosavi, Sobhan. (2019). US Accidents(2016 - 2021), Version 6 (Dec 2021). Retrieved 12th of January 2023 from kaggle https://www.kaggle.com/datasets/sobhanmoosavi/us-accidents?datasetId=199387&sortBy=voteCount&search=severity.

- Moosavi, Sobhan, Mohammad Hossein Samavatian, Srinivasan Parthasarathy, and Rajiv Ramnath. “A Countrywide Traffic Accident Dataset.”, 2019.

- Moosavi, Sobhan, Mohammad Hossein Samavatian, Srinivasan Parthasarathy, Radu Teodorescu, and Rajiv Ramnath. "Accident Risk Prediction based on Heterogeneous Sparse Data: New Dataset and Insights." In proceedings of the 27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, ACM, 2019.

kaggle™

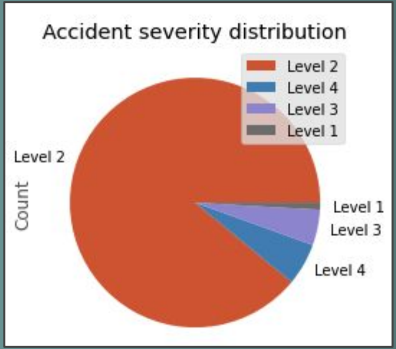
Exploring the dataset

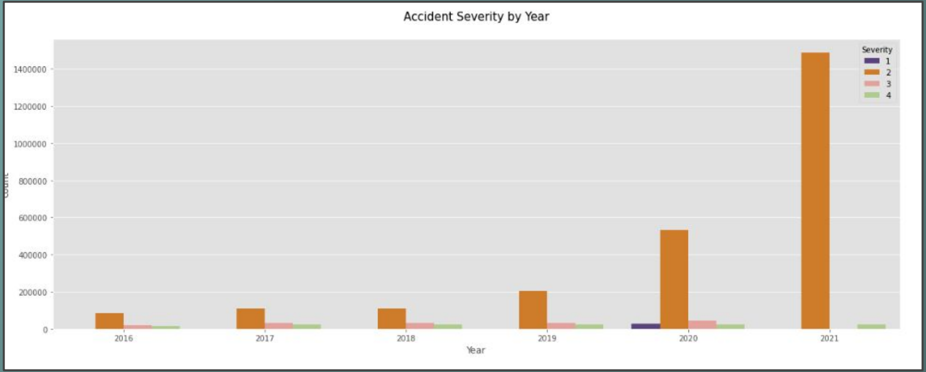
Before wrangling data, data exploration is necessary.

- 2.845.342 instances of accidents.
- 45 columns containing conditions
- Unique identifier as index
- Focus on *Severity* feature
  - Score between 1 and 4
  - 1 is short delay of 2 minutes and 30 second (low traffic impact)
  - 4 is long delay of 18 minutes or more (high traffic impact)
- Divided the remainder of the columns in 4 groups.
  - Weather conditions
  - Phase of day
  - Infrastructure conditions
  - Location attributes

Data Wrangling methods (part 1)

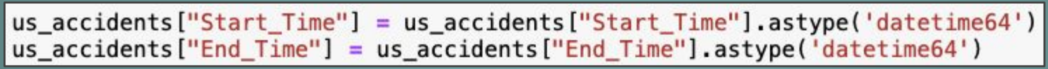
- **Severity analysis**
  - 89% of instances has score 2
  - Very unevenly distributed over timespan
  - This distribution could be problematic for our analyses
  - no clear explanation, no adjustment





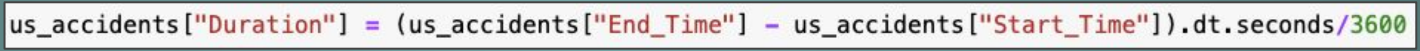
Data Wrangling methods (part 2)

- **Datetime adjustment**
  - two columns with time object: *Start_Time* and *End_Time*



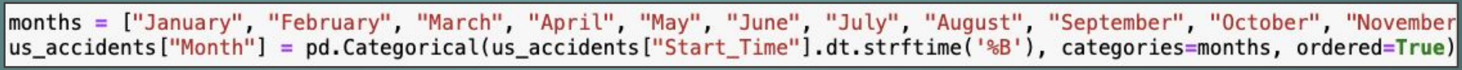
```
us_accidents["Start_Time"] = us_accidents["Start_Time"].astype('datetime64')
us_accidents["End_Time"] = us_accidents["End_Time"].astype('datetime64')
```

- **Allows us to create new columns**
  - Aggregate accidents by year, month, weekday and hour
  - Calculate the duration of an accident



```
us_accidents["Duration"] = (us_accidents["End_Time"] - us_accidents["Start_Time"]).dt.seconds/3600
```

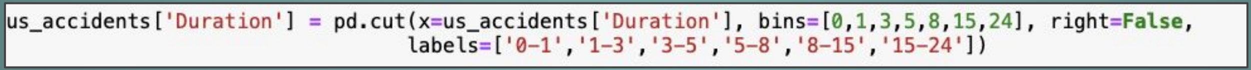
- **Made sure month were chronological for plotting**



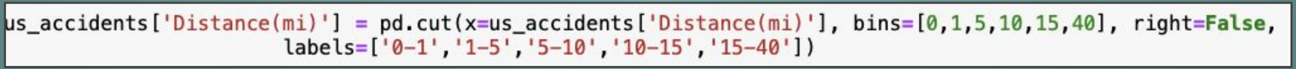
```
months = ["January", "February", "March", "April", "May", "June", "July", "August", "September", "October", "November
us_accidents["Month"] = pd.Categorical(us_accidents["Start_Time"].dt.strftime('%B'), categories=months, ordered=True)
```

Data Wrangling methods (part 3)

- **Improve new duration feature and distance feature**
    - Since Duration is a continuous variable, it is still not in its most meaningful format. It is better suited for analysis if it is converted into a categorical format using bins

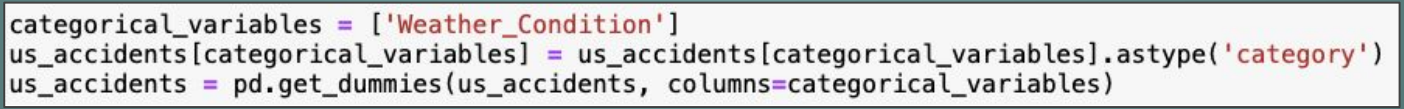


```
us_accidents['Duration'] = pd.cut(x=us_accidents['Duration'], bins=[0,1,3,5,8,15,24], right=False,
                                   labels=['0-1','1-3','3-5','5-8','8-15','15-24'])
```

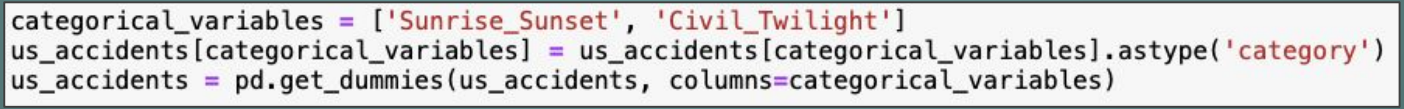


```
us_accidents['Distance(mi)'] = pd.cut(x=us_accidents['Distance(mi)'], bins=[0,1,5,10,15,40], right=False,
                            labels=['0-1','1-5','5-10','10-15','15-40'])
```

- **One-hot encoding Weather condition and phase of day condition**
    - This operation was revealed necessary because it splitted the Weather_Condition, Sunrise_Sunset and Civil_Twilight variables into multiple binary categorical variables to serve for correlation maps for the data exploration section later on



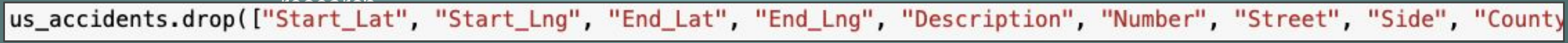
```
categorical_variables = ['Weather_Condition']
us_accidents[categorical_variables] = us_accidents[categorical_variables].astype('category')
us_accidents = pd.get_dummies(us_accidents, columns=categorical_variables)
```



```
categorical_variables = ['Sunrise_Sunset', 'Civil_Twilight']
us_accidents[categorical_variables] = us_accidents[categorical_variables].astype('category')
us_accidents = pd.get_dummies(us_accidents, columns=categorical_variables)
```

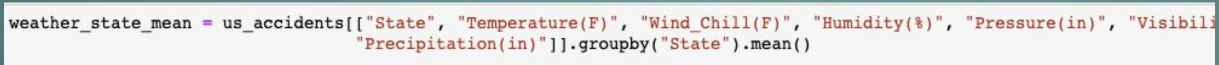
Data Wrangling methods (part 4)

- **Dropping features**
  - A long list of features was dropped because they are not relevant for our research



```
us_accidents.drop(["Start_Lat", "Start_Lng", "End_Lat", "End_Lng", "Description", "Number", "Street", "Side", "County
```

- **Handling missing values**
  - **Dropped**
    - Dropped 137 records with a missing value for the column city.
    - Dropped 329 records with a missing value for distance in miles
    - Dropped 3659 records with a missing value for time zones
  - **Filled**
    - The variables Humidity (%), Pressure (in)) & Wind_speed (mph) had nearly 950.000 NaN combined (see Table) This is roughly 33% of the dataset, hence the records are not dropped.
    - We decided to take the average of each variable per State in the US



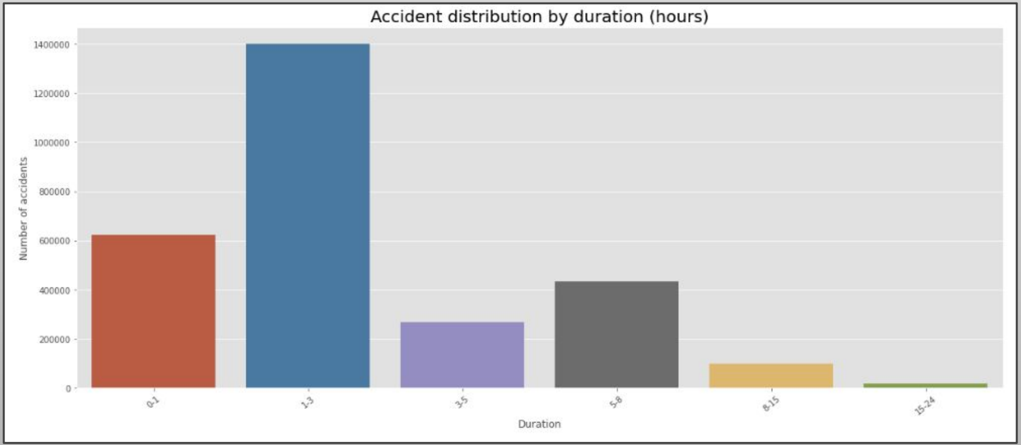
```
weather_state_mean = us_accidents[["State", "Temperature(F)", "Wind_Chill(F)", "Humidity(%)", "Pressure(in)", "Visibili
                                   "Precipitation(in)"]].groupby("State").mean()
```

Variable	Missing values
Humidity (%)	73089
Pressure (in)	59196
Wind_speed (mph)	1579333

Data Analysis Duration

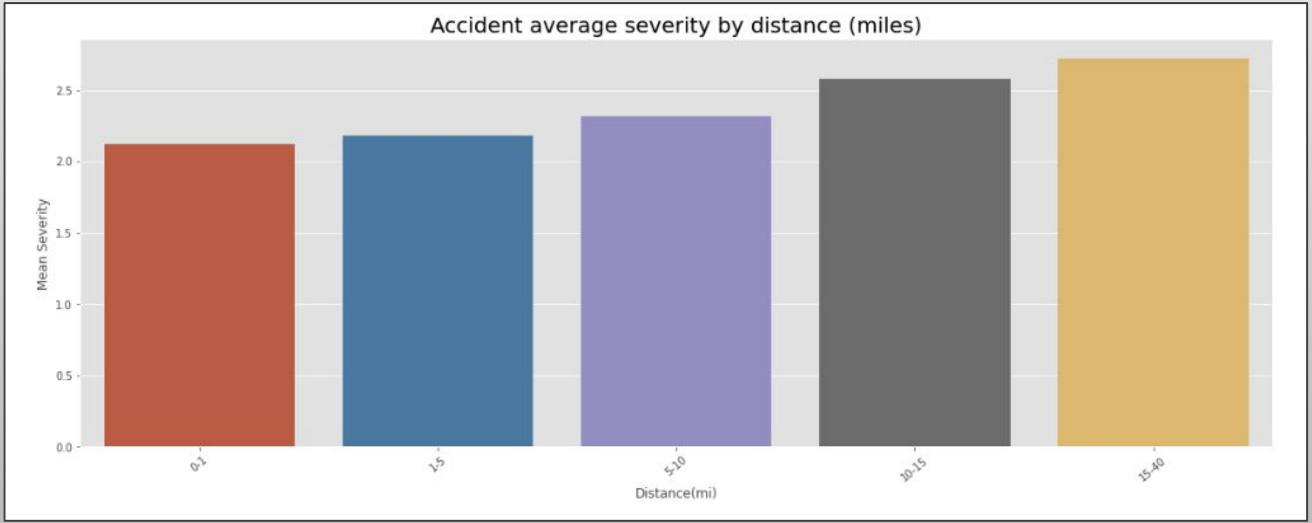
Taking into account the negative impacts of congestion caused by accidents in road traffic, it is important to understand which accidents last longer to end. In that regard, hour intervals were created to aggregate the data. The intervals are the following: [0-1]; [1-3]; [3-5]; [5-8]; [8-15]; [15-24].

We conclude that, the majority of the accident take between 1 and 3 hours to be solved. On the other hand, accidents that take up more than 15 hours to be solved are residuals



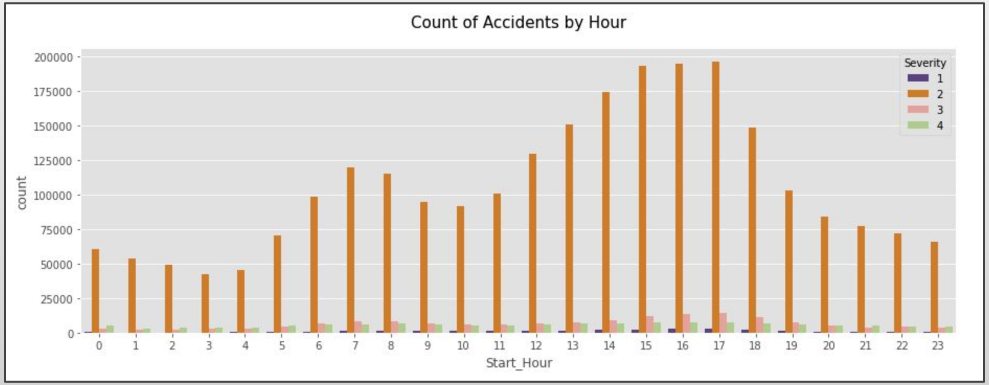
Data Analysis Distance

We can conclude, as expected, that as the distance caused by the accident increases the average severity caused by the accident itself increases as well. The data are, therefore, coherent.

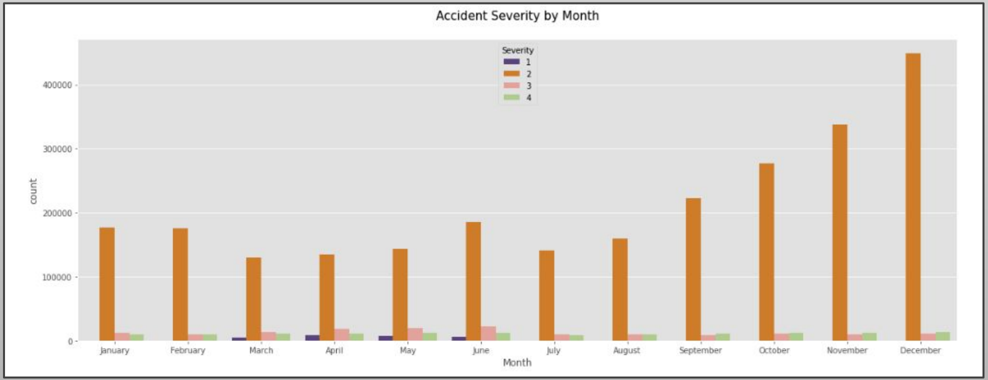


Data Analysis Time

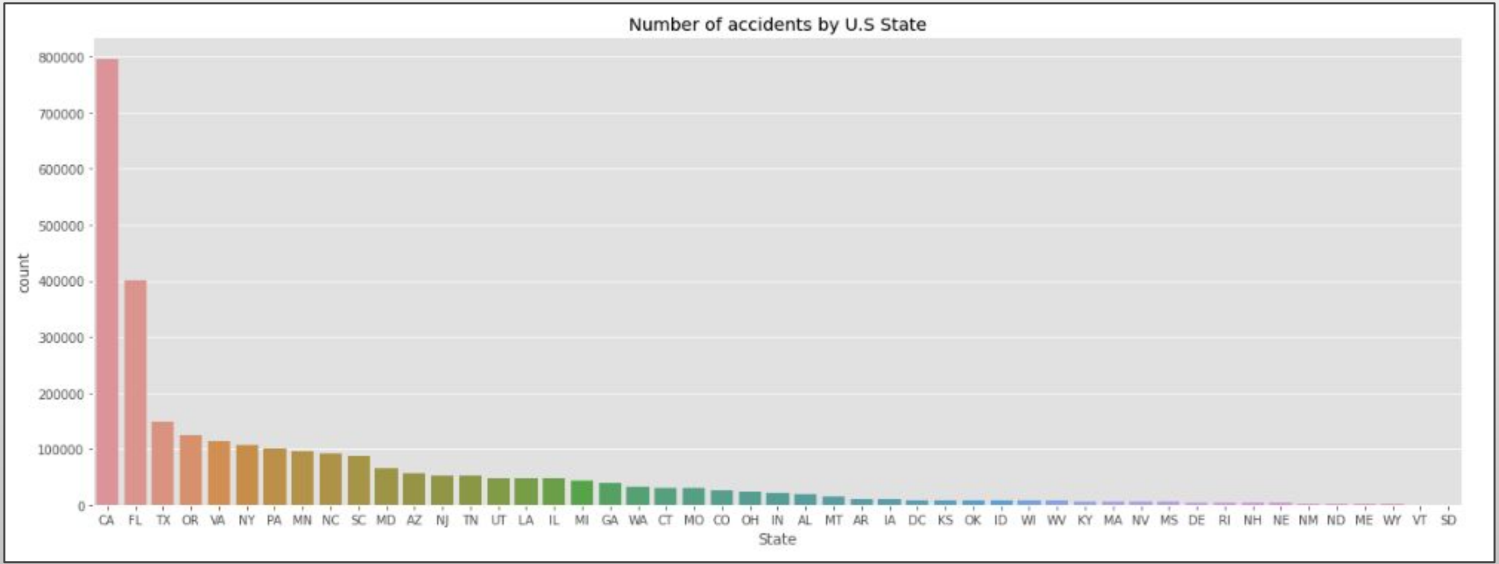
The day hour where most of the accidents happen are in the morning from 6 a.m to 8 a.m, and, in the evening from 2 p.m to 5 p.m.



The months pertaining the highest road accident rates are the latest months in year, from September to December.



Data Analysis Geographical distribution



When it comes to states, California and Florida are responsible for nearly half of the total accidents. This insight will be confirmed by the interactive United States state division map offered below.

Data Analysis Correlation

Weather conditions & Severity

Variables that have the strongest correlations in a **positive direction**:

- **Humidity(%)**
- **Pressure(in)**
- **Wind Speed(mph)**
- **Weather Condition Clear**

In a **negative direction**:

- **Temperature(F)**
- **Wind Chill(F)**
- **Weather Condition Fair**
- **Weather Condition Cloudy**

Phase of day & Severity

Despite no strong correlation is the strongest correlations in a **positive direction**: **Civil Twilight Night**

In a **negative direction**: **Civil Twilight Day**

Infrastructure conditions & Severity

When it comes to the infrastructure variables, the ones with the strongest correlations towards the Severity in a **positive direction is: Junction**

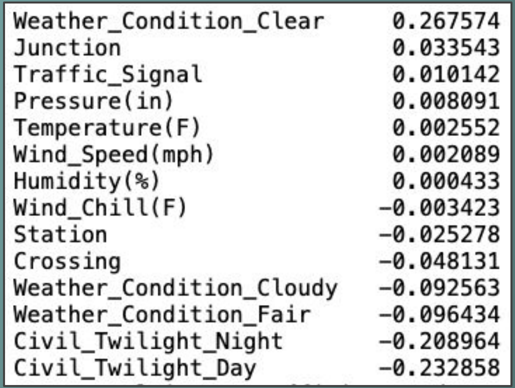
In a **negative direction**:

- Crossing
- Station
- Traffic Signal

Linear regression Result

- **Coefficient series in relation to the target (Severity) variable (see series).**

- After fitting this model on the data, the variables that have the s**trongest explanation power** over the Severity (hence the higher absolute coefficient) are:
    - Weather_Condition_Clear
    - Junction
    - Civil_Twilight_Day
    - Civil_Twilight_Nigh
    - Weather_Condition_Fair
    - Weather_Condition_Cloudy
    - Crossing



```
Weather_Condition_Clear      0.267574
Junction                     0.033543
Traffic_Signal               0.010142
Pressure(in)                 0.008091
Temperature(F)               0.002552
Wind_Speed(mph)              0.002089
Humidity(%)                  0.000433
Wind_Chill(F)               -0.003423
Station                     -0.025278
Crossing                    -0.048131
Weather_Condition_Cloudy    -0.092563
Weather_Condition_Fair      -0.096434
Civil_Twilight_Night        -0.208964
Civil_Twilight_Day          -0.232858
```

Main findings

- The great majority of the accidents in the dataset lie on a severity value of 2 (~ 89%). And the accident severity mean is approximately 2,137
- Most accidents take between 1 and 3 hours to be solved. On the other hand, accidents that take up more than 15 hours to be solved are residuals
- The great majority of the accidents do not provoke huge problems in terms of queue length, as the 0-1 distance interval is the most frequent situation
- The distance caused by the accident increases the average severity caused by the accident itself increases as well. The data are, therefore, coherent
- The year with greatest number of accidents in the dataset is, by far, 2021, and the year with less accidents is 2016
- The months pertaining to the highest road accident rates are the latest months in year, from September to December
- The highest accident rates take place during the workdays, as Friday topping. The accident numbers decline on the weekend
- The day hour where most of the accidents happen are in the morning from 6 to 8 hours, and, in the evening from 14 to 17 hours
- The states of California and Florida are responsible for nearly half of the total accidents
- The cities of Miami (Florida), Los Angeles (California) and Orlando (Florida) are top-3 most dangerous U.S. cities to drive

Conclusion

In view of the results achieved, important information can be drawn on the main predictors of accidents in the USA, as well as their impact in terms of severity and consequences for road traffic.

Thus answering the research question, we can say that the severity of accidents recorded in the USA between 2016 and 2021 is influenced by factors related to the weather situation (mainly cloudiness and wind), road infrastructure (mainly junctions, crossings and traffic signs) and with the phase of day (day or night).

In addition, the present work also references the occurrence of accidents by geographic location, identifying the states and cities with the highest accident rates.

In this sense, **this work will be able to provide important data to the authorities** to deepen the diagnosis produced and, consequently, define strategies aimed at solving the identified problems.