**Data Analytics broad plan**

1. US ACCIDENTS dataset is single dataset.
2. **Introduction:** This public dataset has car accidents (in USA) data since Feb, 2016 and going till Dec 2019 and covers most of the officially reported accidents in 49 states of USA (except Alaska and Hawaii but including Washington DC) categorized by severity level. This data is non-personal data used for community. This data is collected automatically through API calls to MapQuest and Bing database systems and merging them together in one file. This single dataset has approx. 3 million (though I got only 1,048,576 on my local machine) rows of car accident records but it keeps growing. It has 49 different columns all of which are scalar data including accident description in plain English etc. It begins with unique id (called ID) for each row, source of data which may be ‘MapQuest’ or ‘Bing’ or ‘MapQuest-Bing’ meaning both. It has Traffic Message Control code (TMC) type with which type of this accident can be tracked in various systems in the nation. The structure of TMC is generally complex implying the section of road, direction of traffic, zone and unique identifier together. But in our case, this appears to be like a category of accident. There is a severity column that encodes levels of severity in the accident which starts from 0 and goes upto 4 (meaning high severity). There are many columns like the address of accident site, start\_location (Start\_longitude, Start\_lattitude), end\_location (longitude, latitude) (in case it spread far), incident time (start\_time, end\_time) (date & time), distance covered within zone of accident (in case it spread far), its’ impact on traffic flow, the lane side (right or left) where accident landed, address, daytime/nighttime, weather condition, amenities, infra-structure traffic constraints (like traffic light or intersection or round-about or traffic calmer etc).
3. The web-link is : <https://www.kaggle.com/sobhanmoosavi/us-accidents>

In order to download, create a Kaggle account and download the dataset into S3 bucket.

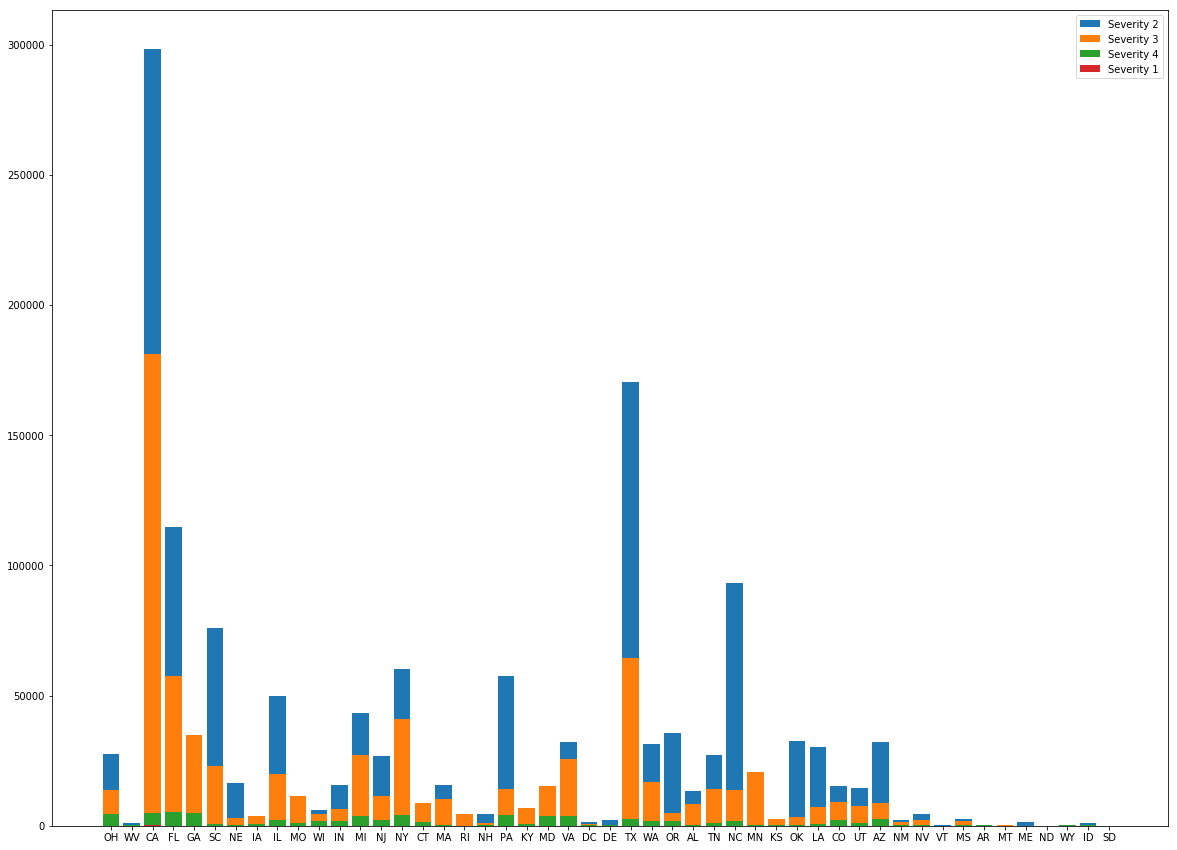
1. In Athena, check if we can query the data in bucket. Also check if you can use Quicksight.
2. The detailed description of the 49 columns of data can be found in either the web-site above or another one as <https://smoosavi.org/datasets/us_accidents> . Please learn them.
3. **Broad steps of data processing:**
4. **Initial check:** We need to check some very basic things like: State has only 49 distinct non-null values; Country is non-null and ‘US’. Start\_Time of accident occurrence is not null and a valid datetime. Zipcode is not null. Start\_Lattitude, Start\_Longitude is non-null and > 0. If there is any data in violation, delete those rows. If we have severity other than 0-5, we can delete the data. You should delete all data with severity 0.



1. **Clean-up of data:** If there are Timezones (for incident-site; start) with more than 5 distinct values (Esatern, Central, Mountain, Pacific, Hawai), then update them to the standard 5 values. If Start\_Time is not a valid datetime then delete those data. If Side is neither null nor ‘R’ nor ‘L’, then make it null. Distance, Temperature, Wind-Chill, Humidity, Pressure, wind\_Speed, Precipitation should be >= 0 or null. Amenity, Bump, Crossing, Give\_way, Junction, No\_Exit, Railway, RoundAbout,Station, Stop, Traffic\_Calming, Traffic\_Signal, Turning\_Loop should be TRUE or FALSE or null. Change other values to null. Sunrise\_Sunset, Civil\_Twilight, Nautical\_Twilight, Astronomical\_Twilight should have allowed values (‘Night’, ‘Day’, null). Change other values to null.
2. **Queries:** This table is very simple with simple columns. Only complex columns can be Description and Weather (and may be a few more). We will have simple queries. But we will pay attention to picking majority. For example, what is the majority severity in state of Texas? Like this. In addition, we may need to calculate various co-relation coefficients and save on tables. We will have about 5-6 such problems.
3. **Data visualization:** Later we will move onto presenting result data in colored visual forms using graphs and charts for each of the saved result tables. Various visual forms will be used like histogram graphs, plots, pie-charts and if possible colored matrix squares. We will try to use Quickstart tool in AWS for this purpose.

That was just a broad plan where our objective is to draw the diagrams. Now we enumerate each problem.

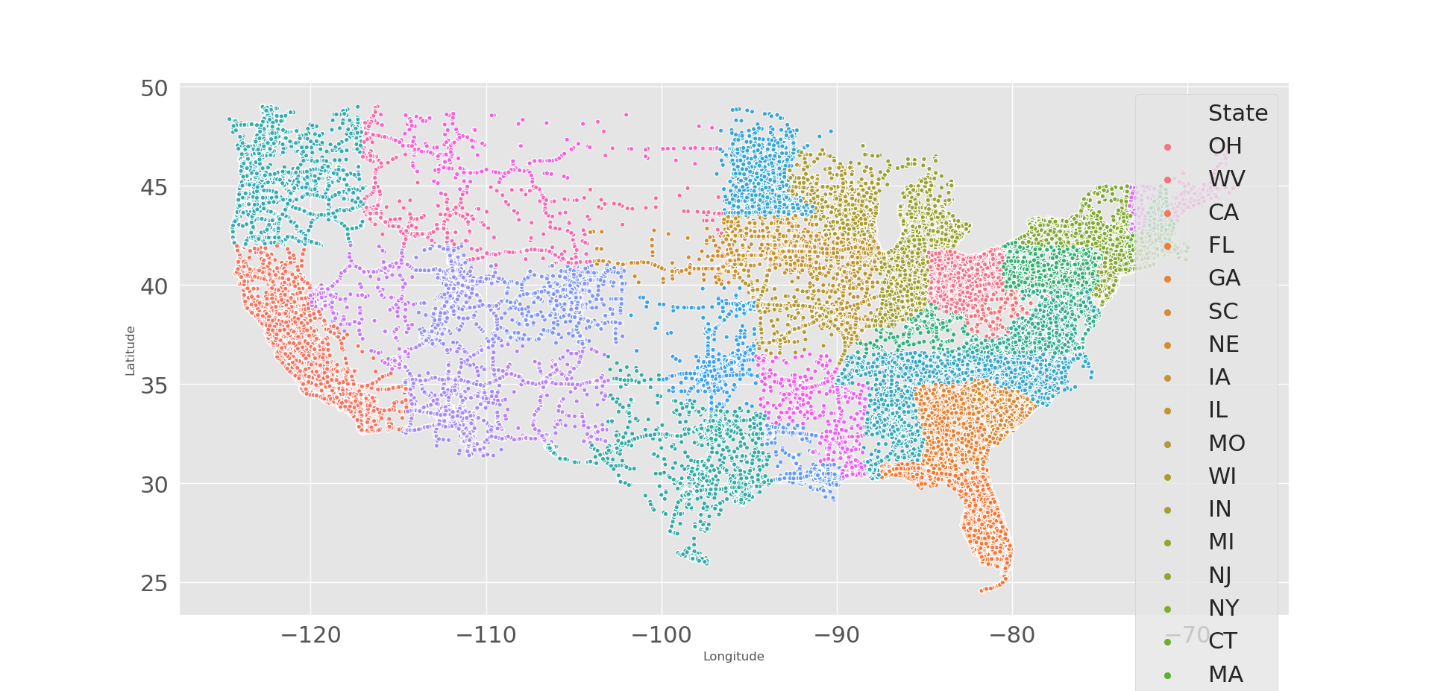
**Query & Visual 1.** We assume there is no severity 0 data. Using vertical histogram bars, draw vertical rectangles to show total no. of accidents for each state (in sequence of total no. of accidents). Use size 16 (horizontal) X 10 (vertical) aspect ratio to printable on 8.5 X 11 inches paper. In X-dimension plot states (showing state code) and in Y-dimension, draw the rectangle (say each fixed width like 0.15 inches or so) with height as total no. of accidents in the respective state. Inside each vertical rectangle, use four colors to show the strength of each severity (or percentage) in this sequence: Green (on the base) for severity 4 (most severe), Orange (above base) for severity 3 (less severe), Blue (above severity 3) for severity 2 (less severe), Red (on top) for severity 1 (least severe). The histogram plot should look like fig 1. The sequence of states on X -direction is dictated by y value the state for highest y-value should show first and then the state with next highest y-value etc.



49 States in sequential order of accidents

Fig 1

**Query & Visual 2.** Simple Scatterplot: First take a map of USA with gray shade and gray border lines and with bold state border lines. Generate a scatterplot of accidents (ignore severity 0) as points with longitude and latitude upon this map as shown in figure 2. Break the map into rectangles each of 10 degrees longitude (-130 to -70) and 5 degrees latitude (+25 to 50). So we separate these 42 rectangles with gray lines. This is second level. Now take 20 different colors and assign one to these rectangles such that no neighboring rectangles have same color. Now generate the scatterplots of accidents within these rectangles with their colors. The plot should look like fig 2.



Scatterplots of accidents by start\_location

**Query & Visual 3.** Simple Pie-chart: Let us define duration of accidents as end\_time – start\_time in minutes & create 5 categories of durations as: 1) minutes (<= 59 minutes), 2) hours (between 60 minutes and 239 minutes), 3) 4hours\_16hours (between 240 minutes and 959 minutes), 4) 16hours\_24hours (between 960 minutes and 1,439 minutes), 5) days (>= 1440 minutes). Ignore severity 0 and count total\_no\_accidents by duration and draw a full-size pie-chart with parts of pies proportional to total\_no\_accidents for 5 different categories with 5 different contrasting colors. The figure should somewhat look like fig 3 (no years).

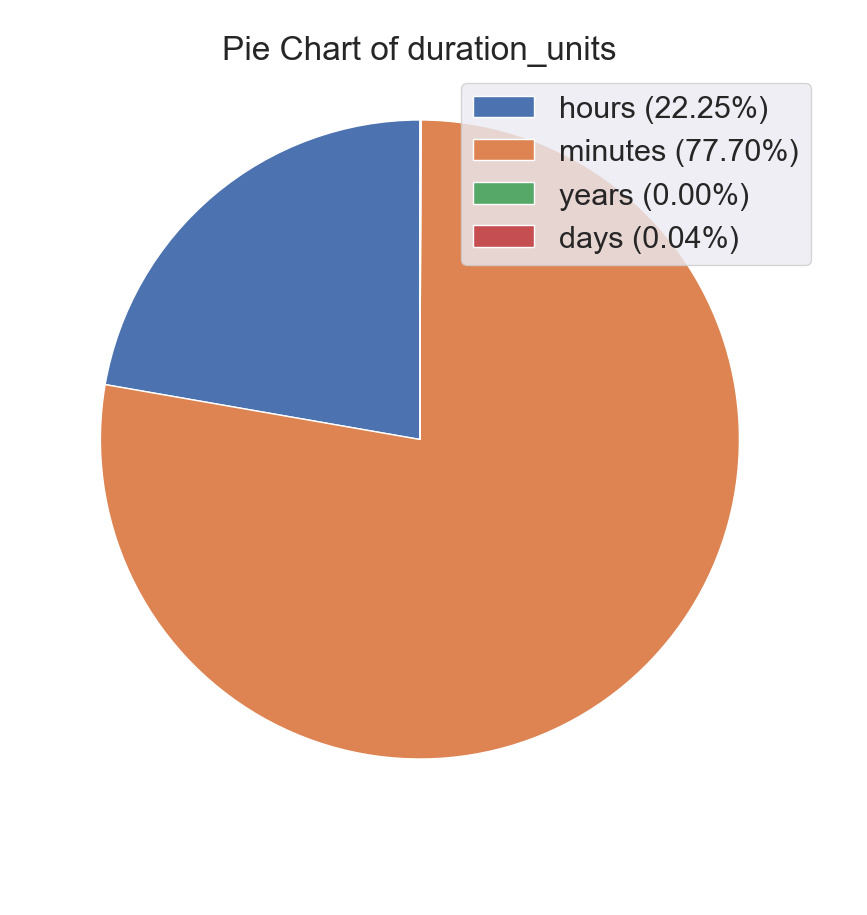


Fig 3: Pie-chart of no\_of\_accidents with duration.

**Query & Visual 4.** Relations: If we look for number of accidents against various weather conditions, we notice most accidents happen in clear or good weather. But suppose we look at breakdown of accidents within each weather condition and then we acquire further insights. Say we count total no of accidents within each weather condition and call it weather\_accidents[i]. Then we calculate percentage\_of\_accidents within each weather condition by calculating percentage\_of\_severity\_1\_accidents[i] as no\_of\_severity\_1\_accidents[i]/weather\_accidents[i] and like that for severity 2, severity 3 and severity 4. Now if we plot percentage\_of\_severity\_1\_accidents[i] , percentage\_of\_severity\_2\_accidents[i], percentage\_of\_severity\_3\_accidents[i], percentage\_of\_severity\_4\_accidents[i] (together) against I then we look for contribution of weather on severity of weather. The bar plot should look somewhat like fig 4 .



Fig 4: On x direction each weather condition and on y direction total percentage of severity\_of\_accidents

**Query & Visual 5.**